Federal University of São Carlos

How and Why Computers Read the Web

Estevam R. Hruschka Jr.

Federal University of São Carlos and Carnegie Mellon University

Federal University of São Carlos

Carnegie Mellon University

FAPESP

Disclaimers

- Previous versions of this tutorial were presented at:
  - IBERAMIA2012 http://iberamia2012.dsic.upv.es/tutorials/;
  - WWW2013 http://www2013.org/program/machine-reading-the-web/
  - ECMLPKDD2015 http://www.ecmlpkdd2015.org
  - AAAI2016 http://www.aaai.org/Conferences/AAAI/2016/aaai16tutorials.php#SA4
- Feel free to e-mail me (estevam.hruschka@gmail.com) with questions about this tutorial or any feedback/suggestions/criticisms. Your feedback can help improving the quality of these slides, thus, they are very welcome.

Disclaimers

- Due to time constraints, I do not intend to cover all the algorithms and publications related to DBPedia, YAGO, KnowItAll and NELL. What I do intend, instead, is to give an overview of all four projects and what is the main approach used in each project.
- DBPedia, YAGO, KnowItAll and NELL are not the only research efforts focusing on “Reading the Web”. They were selected, to be presented in this tutorial, because they represent four different and very relevant approaches to this problem, but it **does not** mean they are the best (or the only relevant) ones.

Outline

- Machine Learning
- Machine Reading
- Reading the Web
  - YAGO
  - KnowItAll
  - NELL
  - DBPedia

Outline

- **Machine Learning**
- Machine Reading
- Reading the Web
  - YAGO
  - KnowItAll
  - NELL
  - DBPedia

Outline

- Machine Learning
- Machine Reading
- Reading the Web
  - YAGO
  - KnowItAll
  - NELL
  - DBPedia

Outline

- Machine Learning
- **Machine Reading**
- Reading the Web
  - YAGO
  - KnowItAll
  - NELL
  - DBPedia

Auto-Text to Knowledge

Picture taken from [DARPA, 2012]

Machine Reading

Auto-Text to Knowledge

Picture taken from [DARPA, 2012]

William Shakespeare | Smart People | Intelligence | Learning New Things | +4

If you can't understand Shakespeare when you read it, are you dumb?

Answer Wiki

No. Answers below note:

- Professional actors and directors need reference works to help them understand Shakespeare.
- It's easier to follow a Shakespeare play by watching it rather than reading it.
- Vocabulary, idiom and culture have changed since the 17th century. Seventeeth century British English is like a foreign language to many 21st century English speakers.
- Many people who say they understand Shakespeare are faking it because they don't want to appear dumb.
- Being able to read and understand Shakespeare is a particular kind of intelligence that you can develop.
- Reading the synopsis of a Shakespeare play can help you understand it better when you read the full play.

Quora | Ask or Search Quora | Ask Question | Read

Reading Comprehension | Understanding and Comprehension | Psychology | +1

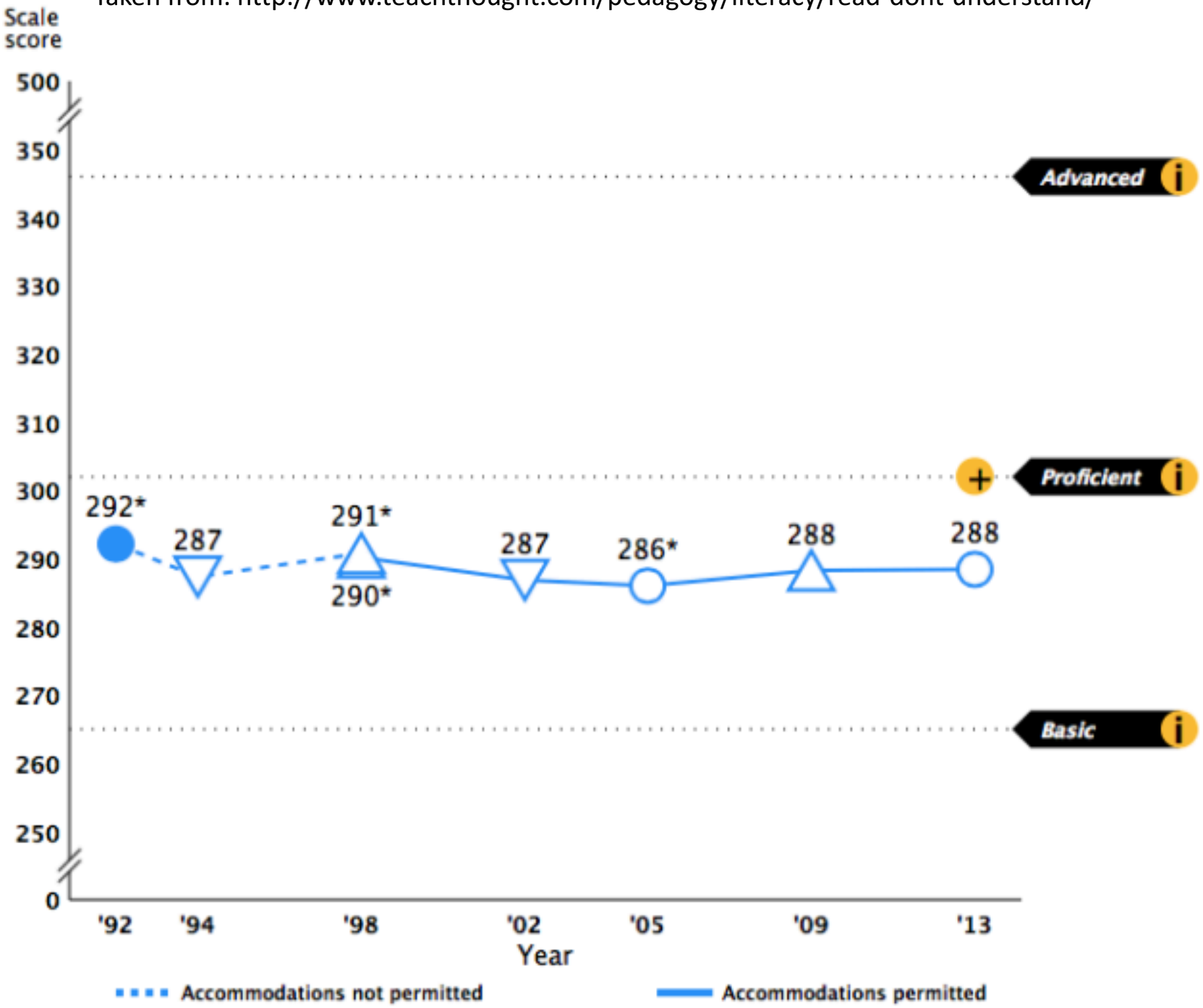
Why don't I understand what I read?

I don't have dyslexia, and my English teacher says I am the best writer in my class (I am in the top stream). Yet still when I read something, I don't take it in.

I normally have to read a paragraph and then go back and read it again.
One reading to hear it in my head, and the other to get the meaning. Why can't I simultaneously read and comprehend? Why must every sentence be read and then understood?

Trend in twelfth-grade NAEP reading average scores

Taken from: http://www.teachthought.com/pedagogy/literacy/read-dont-understand/



Outline

- Machine Learning
- Machine Reading
- Reading the Web
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

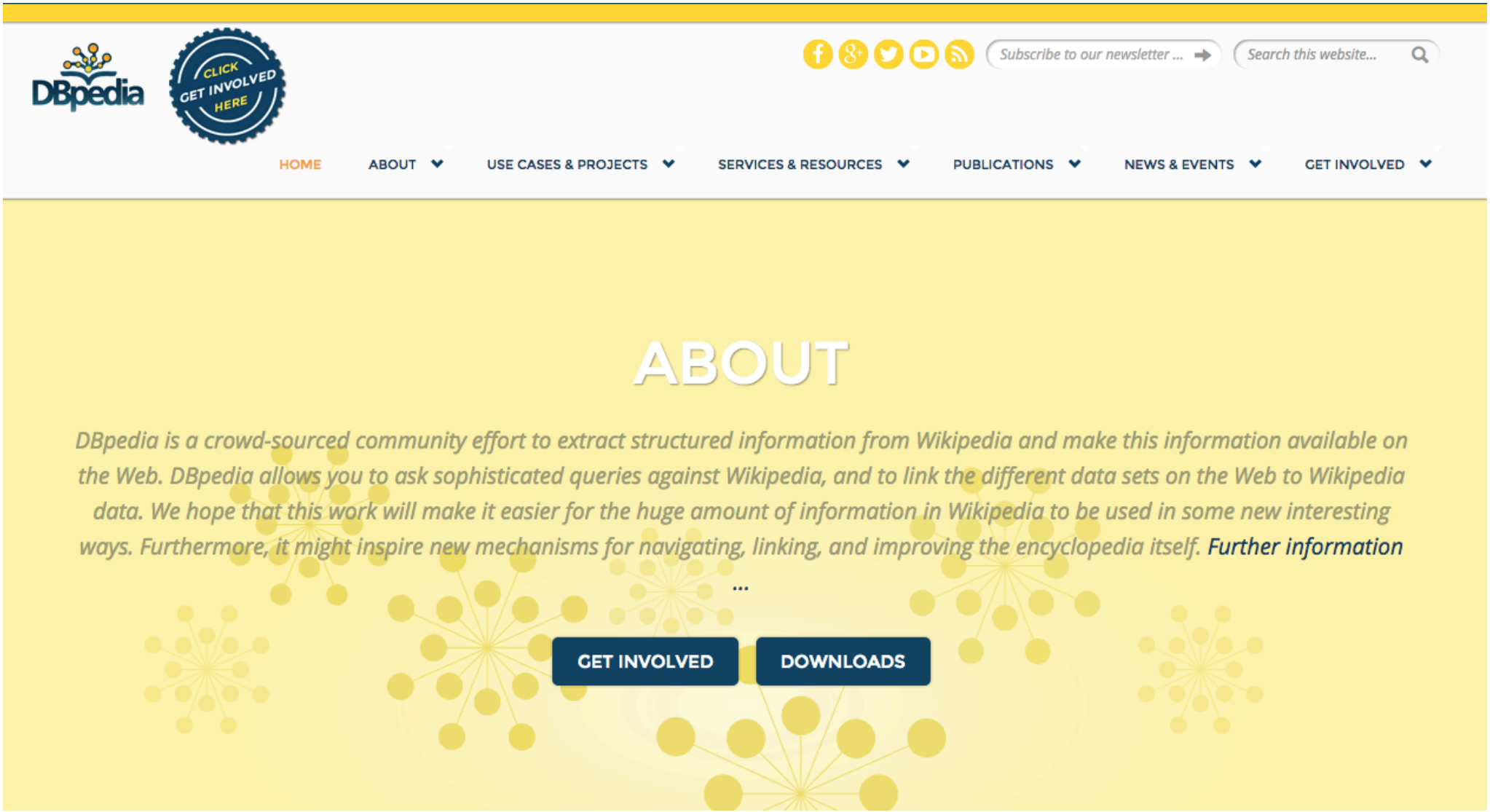
Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - **DBPedia**
  - YAGO
  - KnowItAll
  - NELL

DBPedia


DBpedia
CLICK GET INVOLVED HERE
Subscribe to our newsletter ...
Search this website...
HOME
ABOUT
USE CASES & PROJECTS
SERVICES & RESOURCES
PUBLICATIONS
NEWS & EVENTS
GET INVOLVED
ABOUT
DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia, and to link the different data sets on the Web to Wikipedia data. We hope that this work will make it easier for the huge amount of information in Wikipedia to be used in some new interesting ways. Furthermore, it might inspire new mechanisms for navigating, linking, and improving the encyclopedia itself. Further information ...
GET INVOLVED
DOWNLOADS

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - **YAGO**
  - KnowItAll
  - NELL

Federal University of São Carlos

The YAGO-NAGA Project:
Harvesting, Searching, and Ranking Knowledge from the Web

Deutsch

max planck institut informatik

- Homepage
- About the Institute
- Departments:
  - Algorithms & Complexity
  - Computer Vision and Multimodal Computing
  - Computational Biology & Applied Algorithmics
  - Computer Graphics
  - Databases and Information Systems
    - People
    - Research
      - YAGO-NAGA
        - AIDA
        - ANGIE
        - HYENA
        - JAVATOOLS
        - K2
        - NAGA
        - PATTY

The YAGO-NAGA Project: Harvesting, Searching, and Ranking Knowledge from the Web

Research | **Publications** | **People**

The YAGO-NAGA project started in 2006 with the goal of building a conveniently searchable, large-scale, highly accurate knowledge base of common facts in a machine-processible representation.

We have already harvested knowledge about millions of entities and facts about their relationships, from Wikipedia and WordNet with careful integration of these two sources. The resulting knowledge base, coined YAGO, has very high precision and is freely available. The facts are represented as RDF triples, and we have developed methods and prototype systems for querying, ranking, and exploring knowledge. Our search engine NAGA provides ranked answers to queries based on statistical models.

Several interlinked sub-projects are growing on the YAGO-NAGA basis. Our vision is a confluence of Semantic Web (Ontologies), Social Web (Web 2.0), and Statistical Web (Information Extraction) assets towards a comprehensive repository of human knowledge. Our methodologies combine concepts, models, and algorithms from several fields, including database systems, information retrieval, statistical learning, and logical reasoning.

AIDA

AIDA is a method, implemented in an online tool, for disambiguating mentions of named entities that occur in natural-language text or Web tables.

FAPESP

Carnegie Mellon University

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - **KnowItAll**
  - NELL

Federal University of São Carlos

KnowItAll

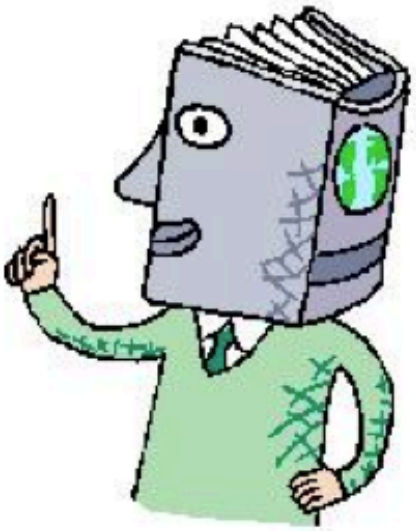
UNIVERSITY OF WASHINGTON | ABOUT US | CONTACT US | MY CSE | INTERNAL

Computer Science & Engineering

UNIVERSITY of WASHINGTON

News & Events | People | Education | Research | Current Students | Prospective Students | Faculty Candidates | Alumni | Industry Affiliates | Support CSE

KnowItAll



1. How can a computer accumulate a massive body of knowledge?
2. What will Web search engines look like in ten years?

To address the questions above, the KnowItAll project has been developing a variety of domain-independent systems that extract information from the Web in an autonomous, scalable manner.

FAPESP

Carnegie Mellon University

KnowItAll: Open Information Extraction

Open Information Extraction

How can a computer accumulate a massive body of knowledge? What will Web search engines look like in ten years?

To address these questions, the Open IE project has been developing a Web-scale information extraction system that reads arbitrary text from any domain on the Web, extracts meaningful information, and stores it in a unified knowledge base for efficient querying. In contrast to traditional information extraction, the Open Information Extraction paradigm attempts to overcome the knowledge acquisition bottleneck by extracting a large number of relations at once.

Demo: TextRunner extracted over 500,000,000 assertions from 100 million Web pages.
Software: ReVerb Open Information Extraction Software and additional information.
Data: Horn-clause inference rules learned by the Sherlock system.
Demo: Selectional Preferences from Web Text compute admissible argument values for a relation.
Data: 10,000 Functional Relations learned from Web Text predict the functionality of a phrase.

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - **NELL**

NELL

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 15 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 1,888,984 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

Outline

- **Machine Learning**
- Machine Reading
- Reading the Web
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Machine Learning

- **What is Machine Learning?**

The field of Machine Learning seeks to answer the question

"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" [Mitchell, 2006]

Machine Learning

- **What is Machine Learning?**

a machine *learns* with respect to a particular:

- task T
- performance metric P
- type of experience E

if the system reliably improves its performance P at task T, following experience E. [Mitchell, 1997]

Machine Learning

- **Examples of Machine Learning approaches for different tasks (T), performance metrics (P) an experiences (E)**

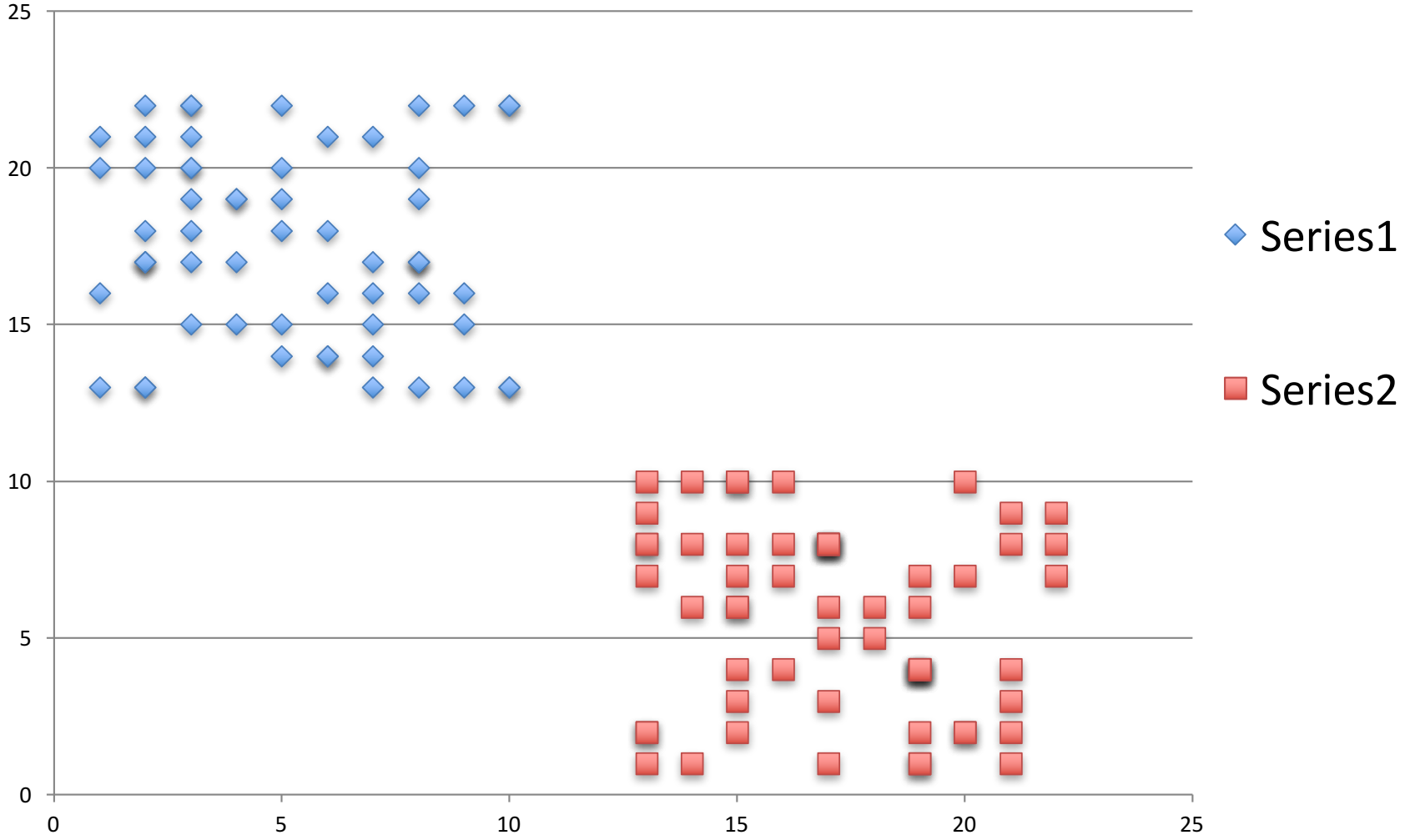
- data mining
- autonomous discovery
- database updating
- programming by example
- Pattern recognition

Machine Learning

- **Supervised Learning;**
- **Unsupervised Learning**
- **Semi-Supervised Learning**

Supervised Learning

(one simple anecdotal approach)



Supervised Learning

How supervised learning algorithms work [edit]

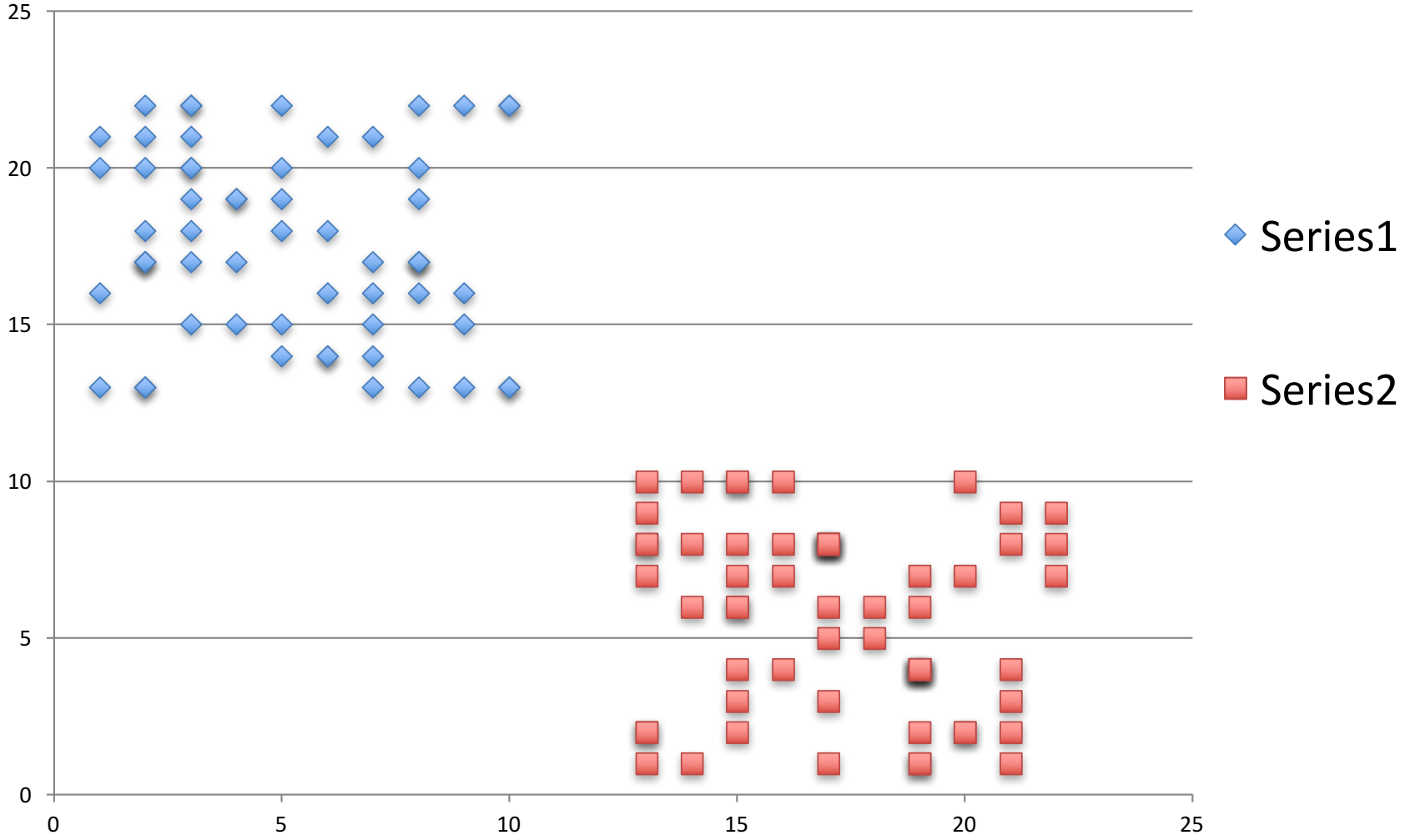
Wikipedia definition

Given a set of N training examples of the form $\{(x_1, y_1), ..., (x_N, y_N)\}$ such that x_i is the feature vector of the i-th example and y_i is its label (i.e., class), a learning algorithm seeks a function $g : X \to Y$, where X is the input space and Y is the output space. The function g is an element of some space of possible functions G, usually called the *hypothesis space*. It is sometimes convenient to represent g using a scoring function $f : X \times Y \to \mathbb{R}$ such that g is defined as returning the y value that gives the highest score: $g(x) = \arg\max_y f(x, y)$. Let F denote the space of scoring functions.

0.5	1	blue
2.9	1.9	blue
1.2	3.1	blue
0.8	4.7	blue
2.7	5.4	blue
8.1	4.7	red
8.3	6.6	red
6.3	6.7	red
8	9.1	red
5.4	8.4	red

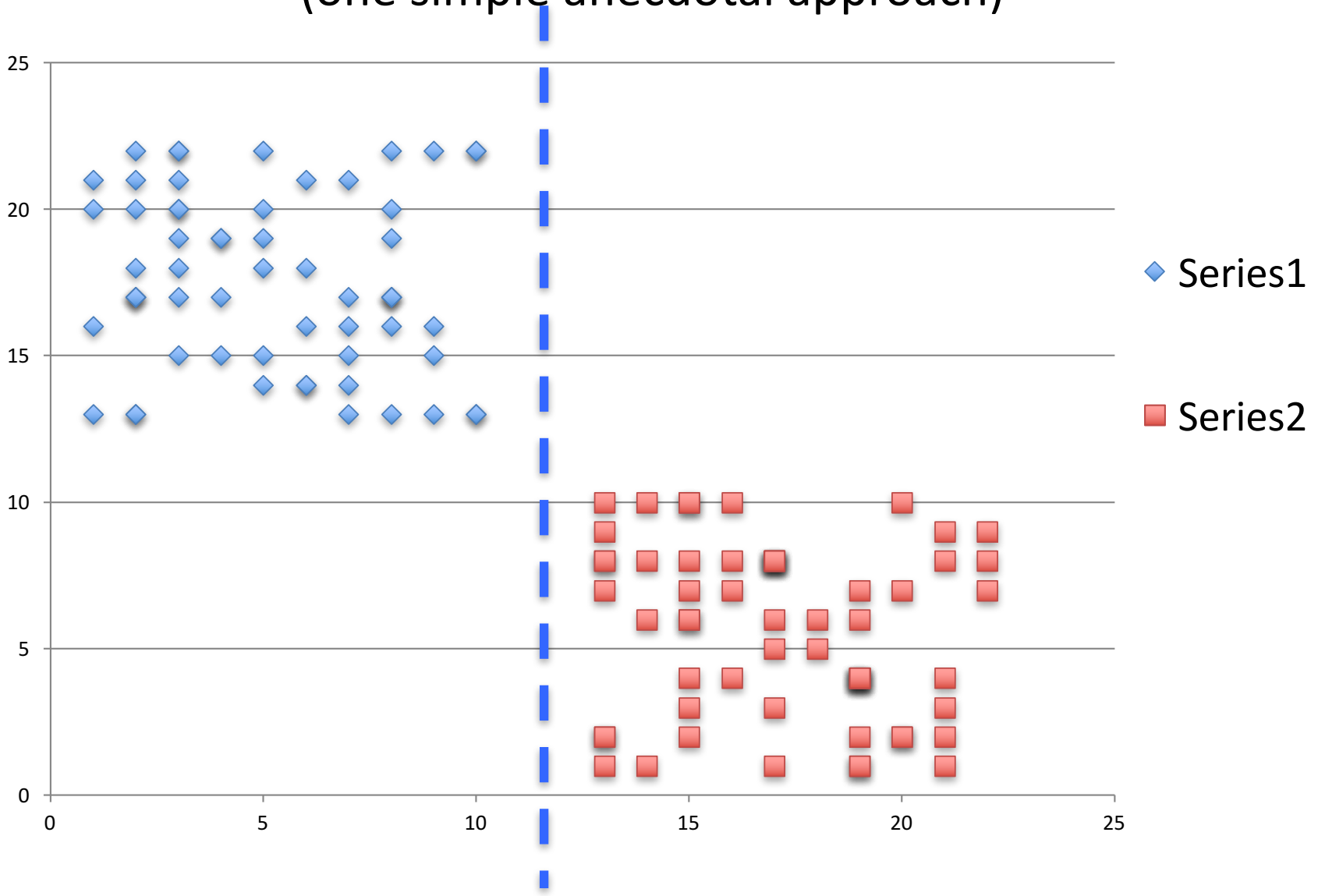
Supervised Learning

(one simple anecdotal approach)



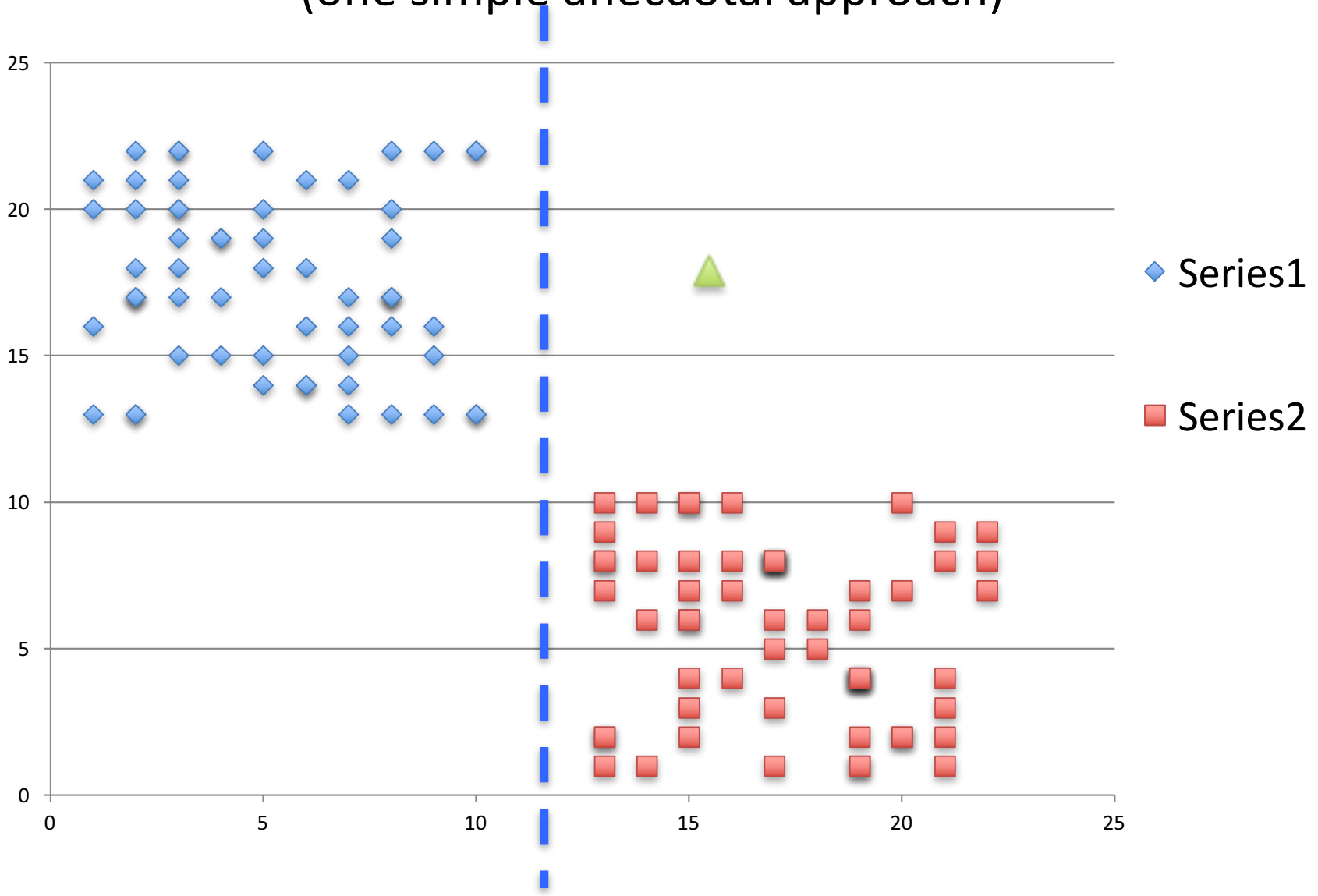
Supervised Learning

(one simple anecdotal approach)



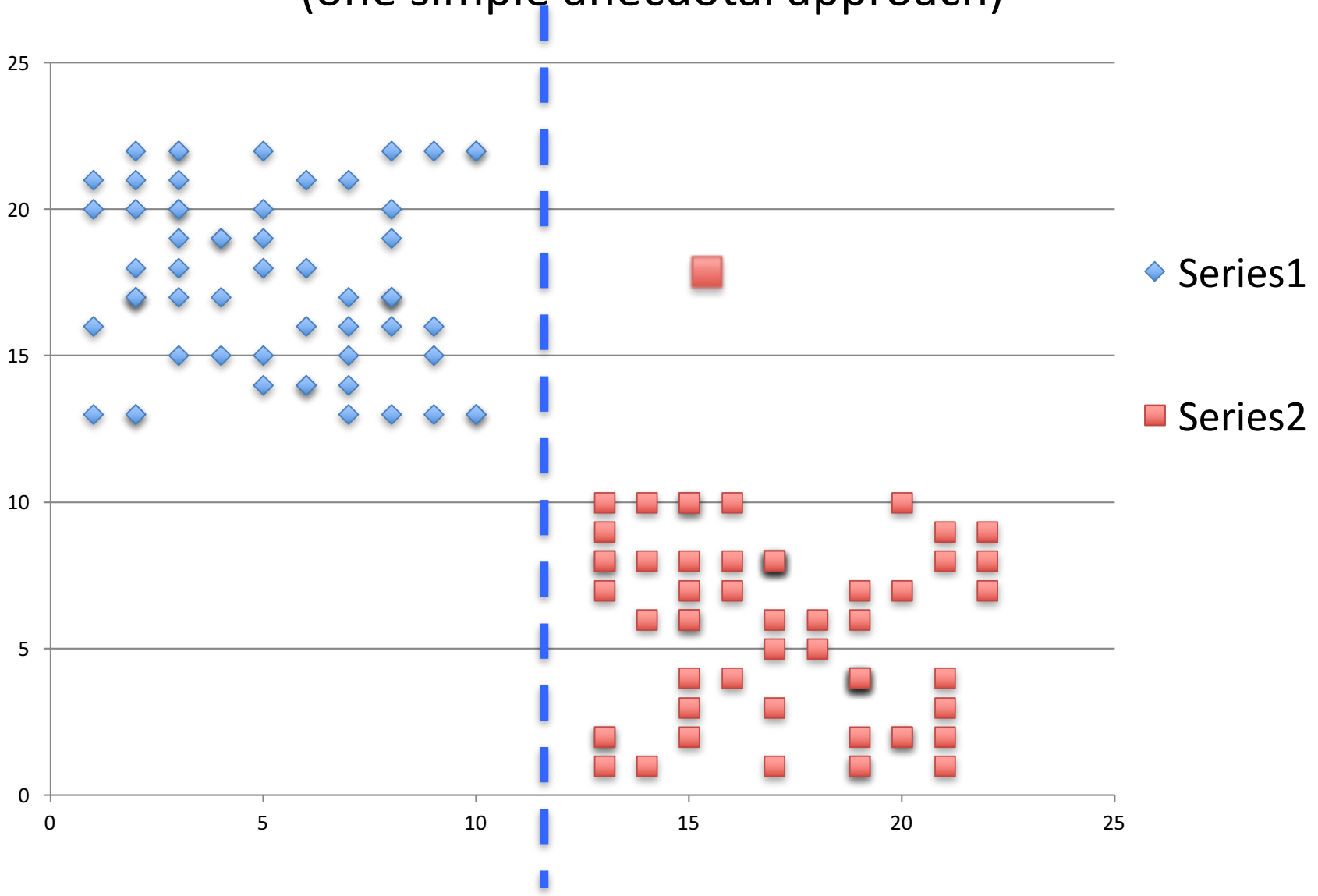
Supervised Learning

(one simple anecdotal approach)



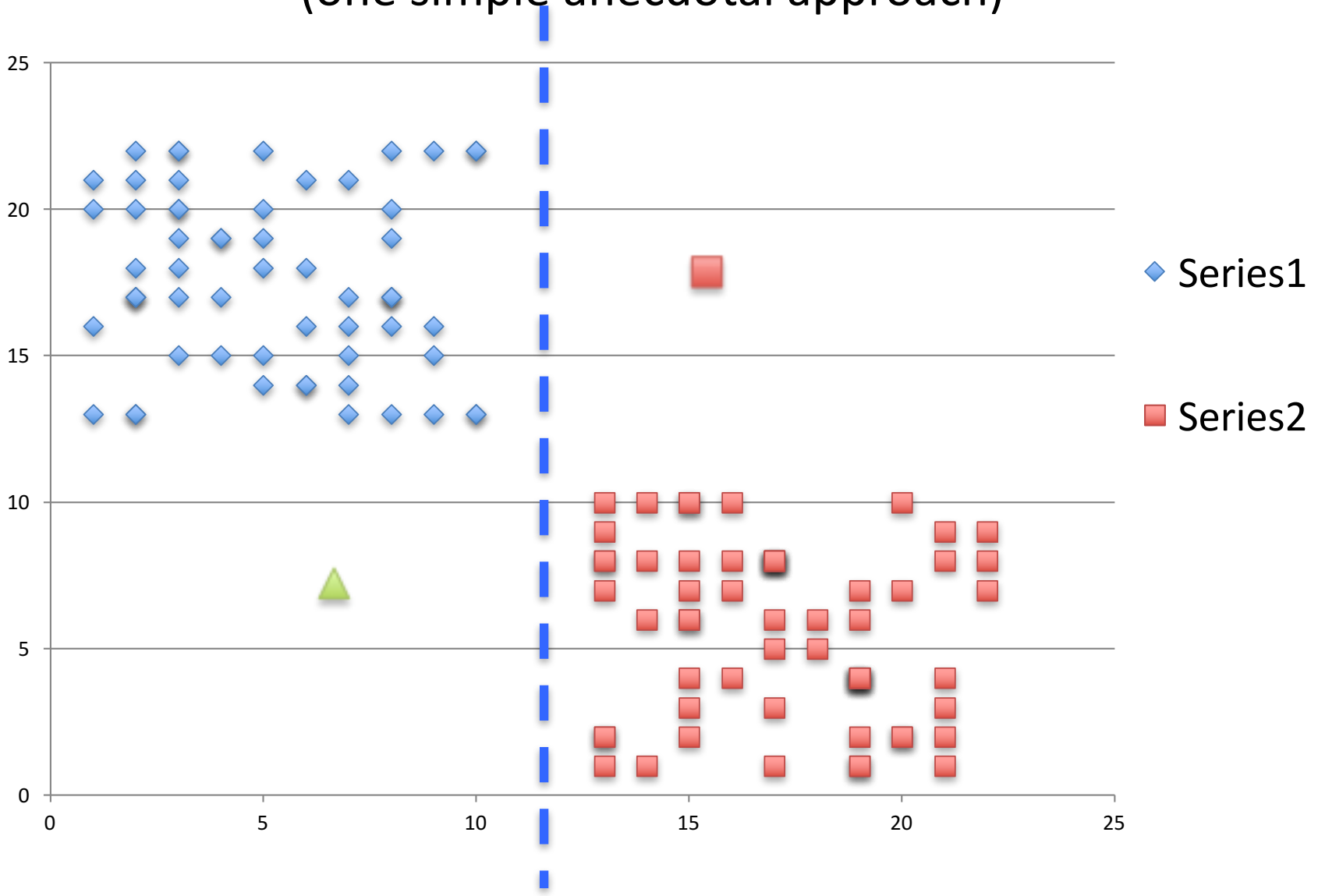
Supervised Learning

(one simple anecdotal approach)



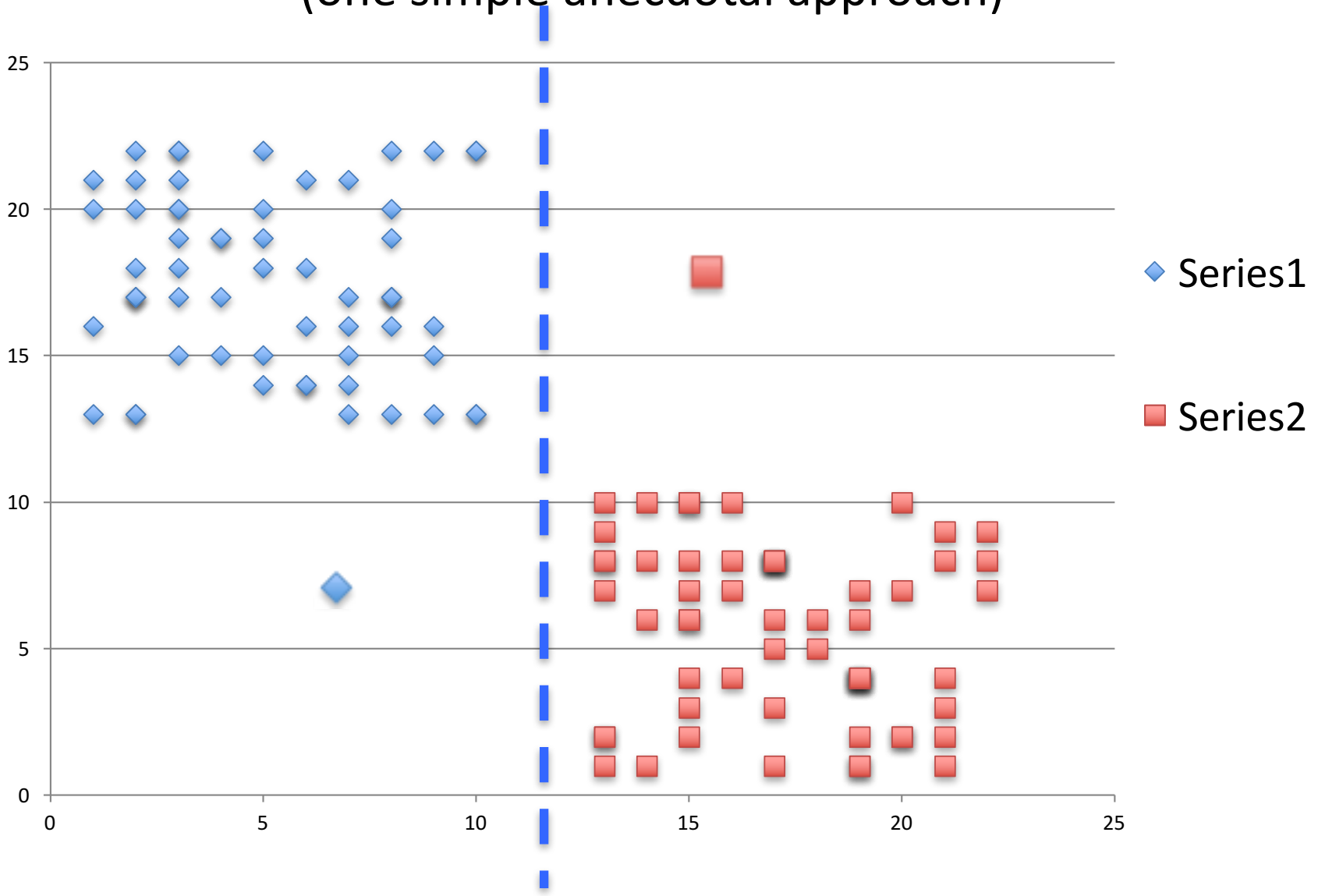
Supervised Learning

(one simple anecdotal approach)



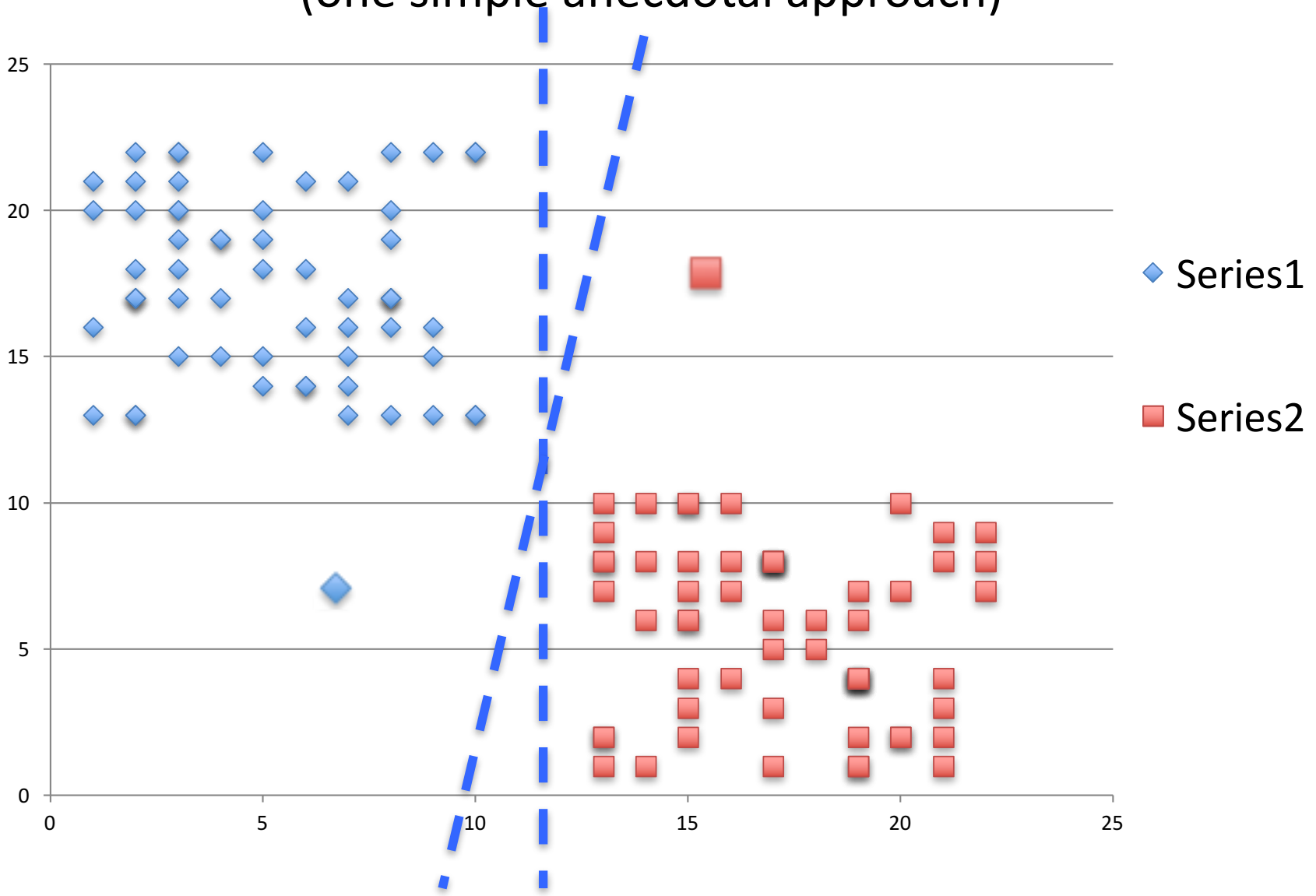
Supervised Learning

(one simple anecdotal approach)



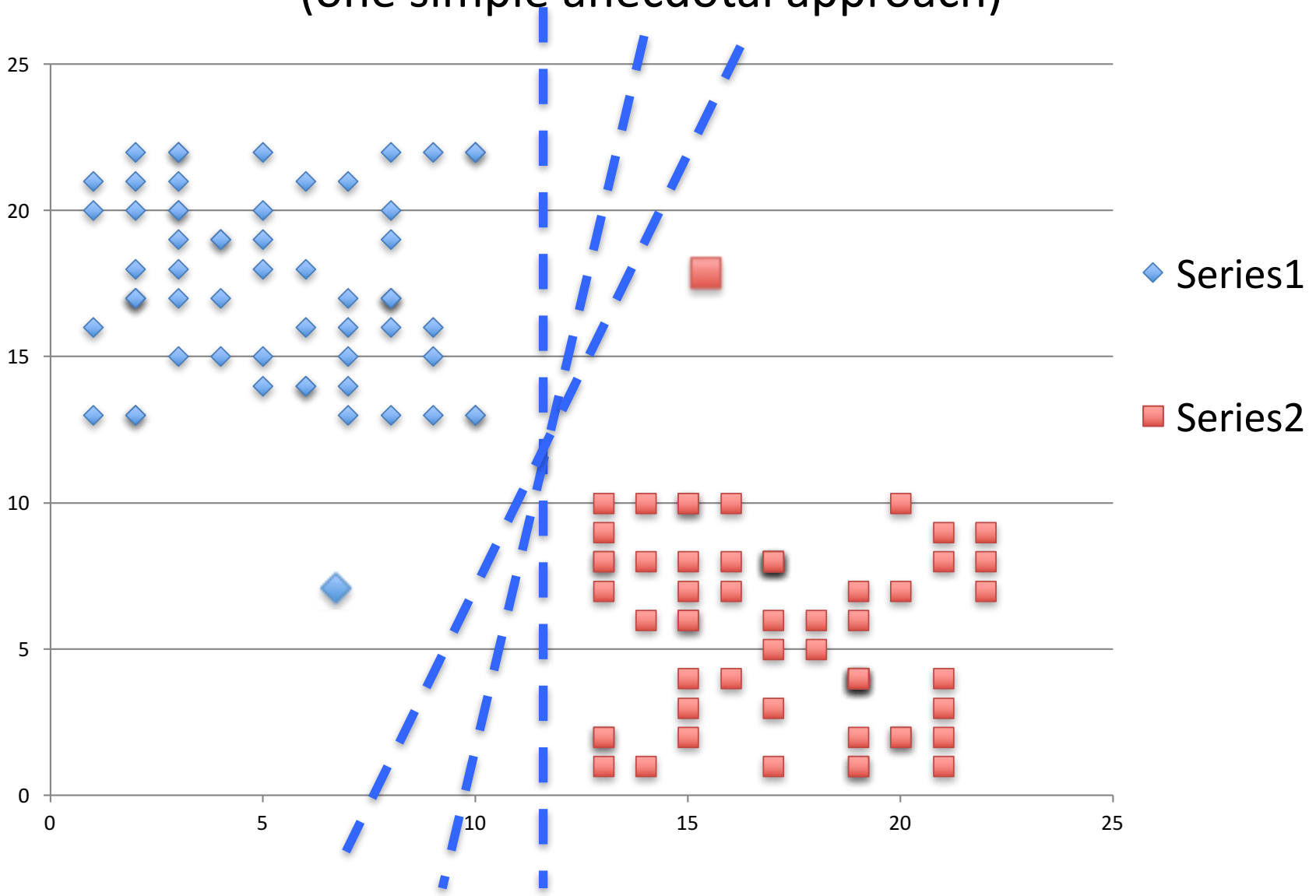
Supervised Learning

(one simple anecdotal approach)



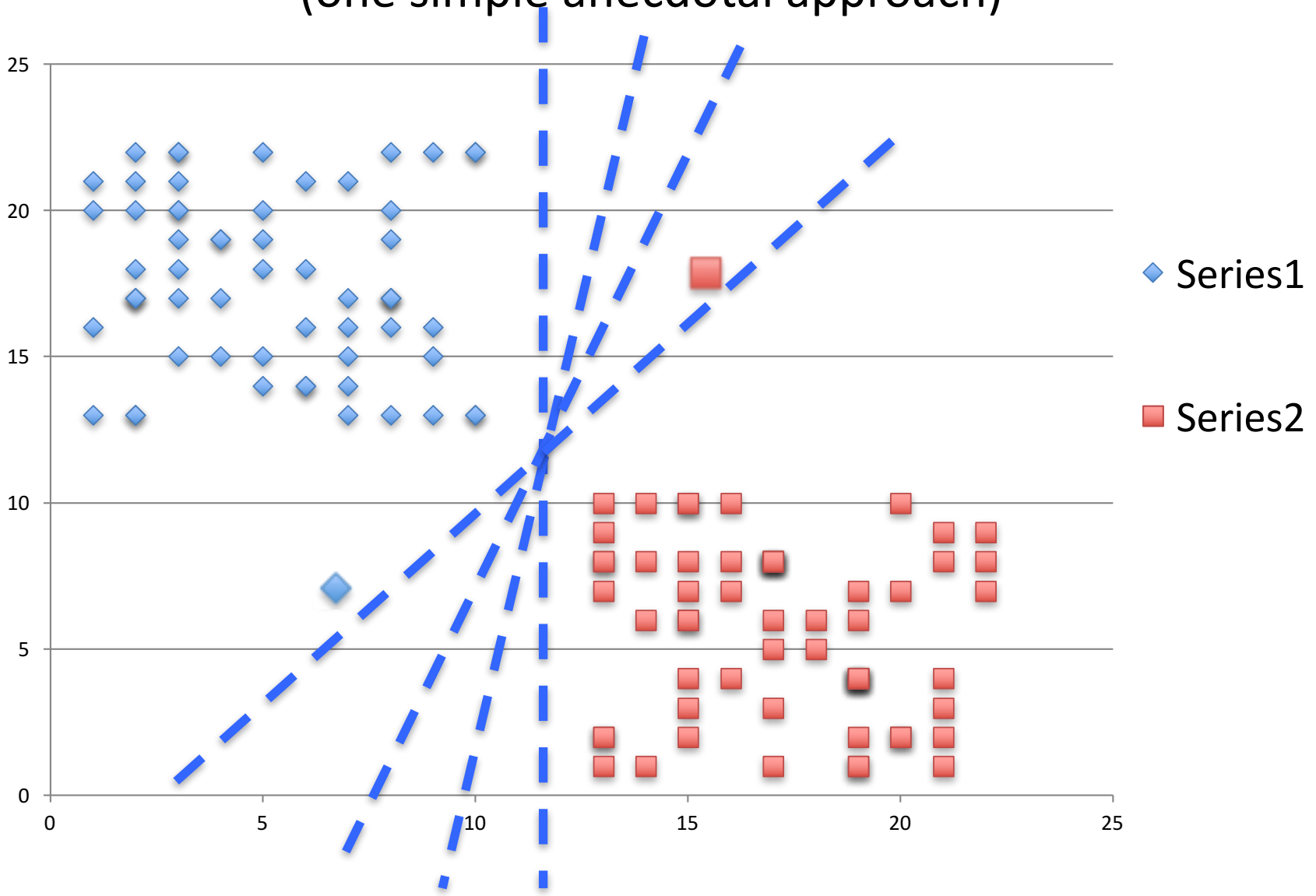
Supervised Learning

(one simple anecdotal approach)



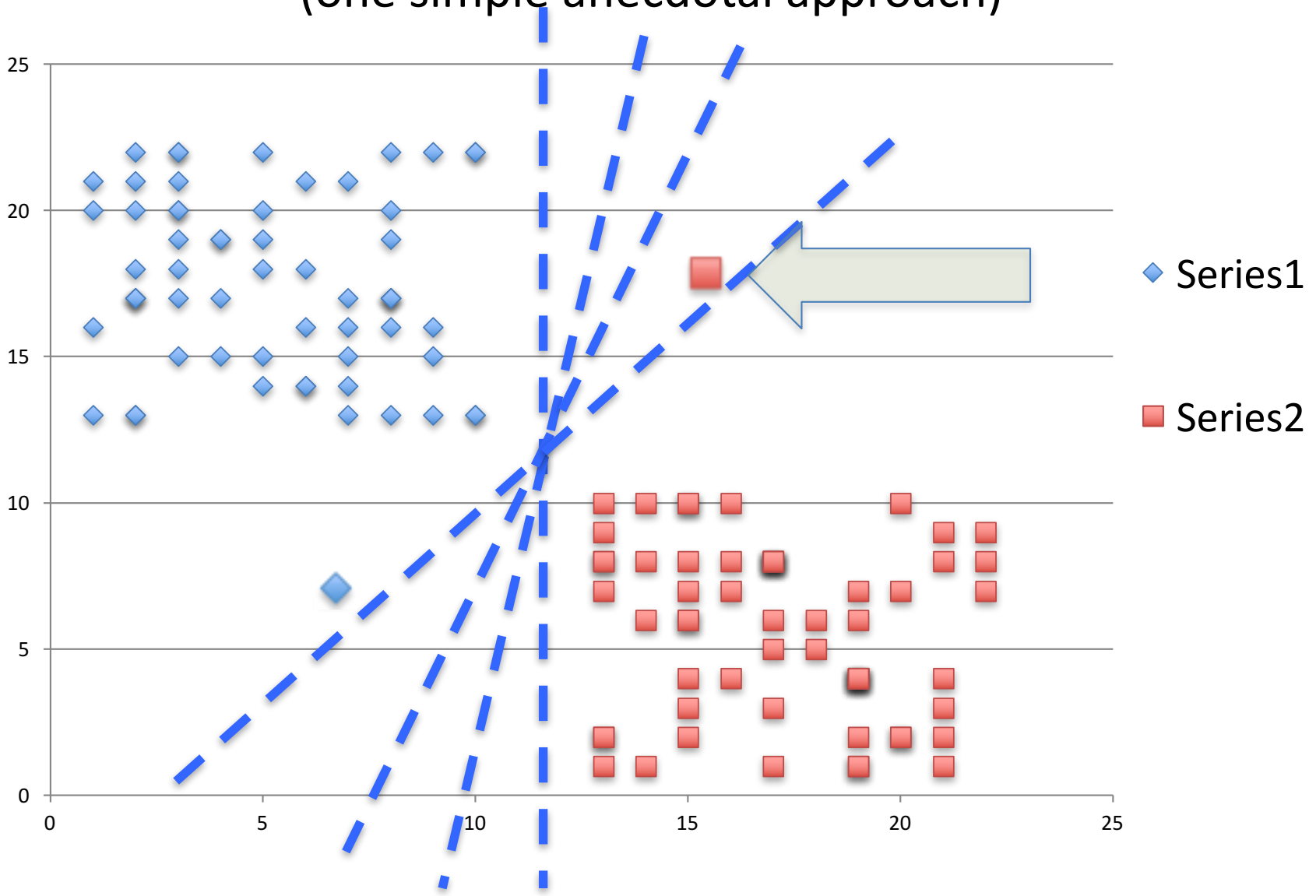
Supervised Learning

(one simple anecdotal approach)



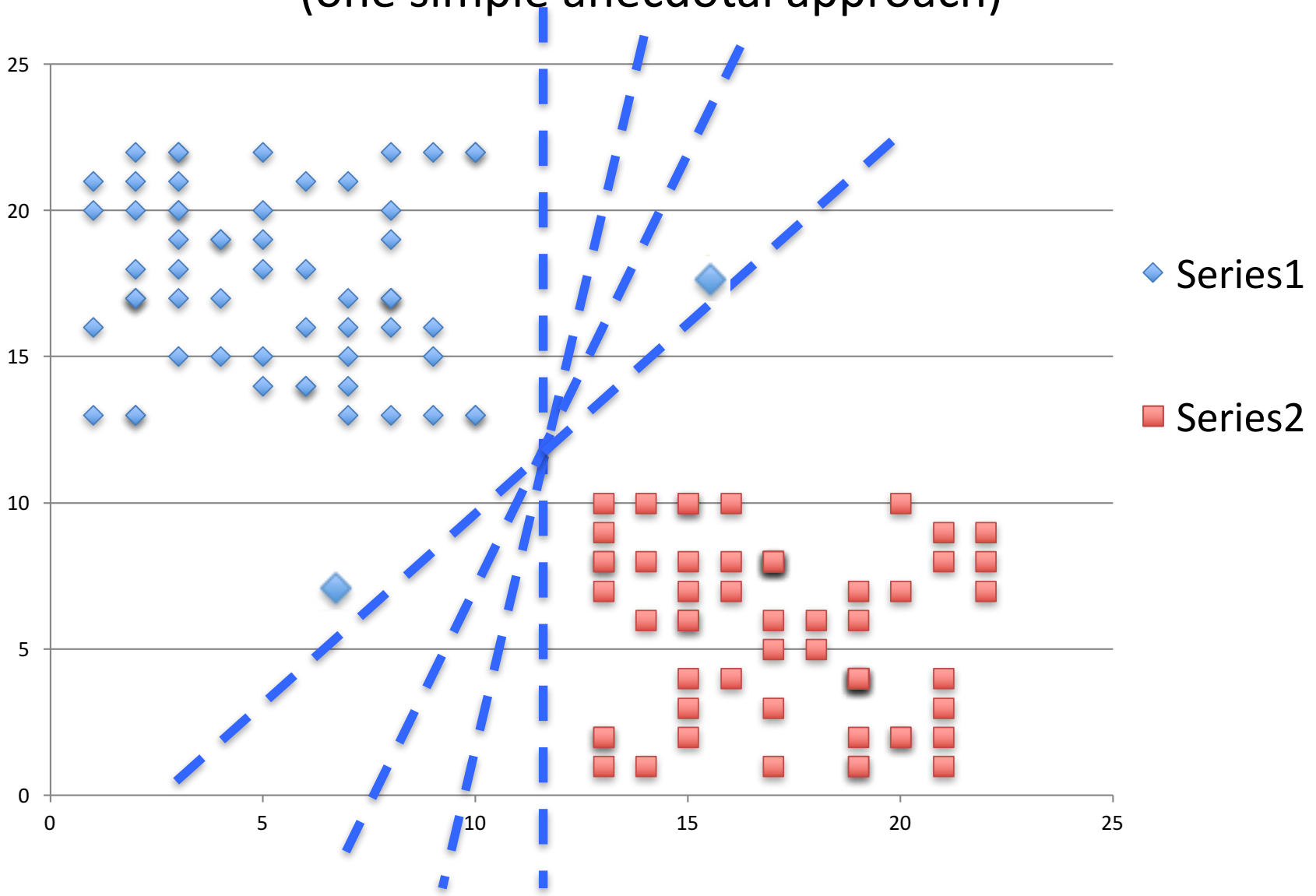
Supervised Learning

(one simple anecdotal approach)



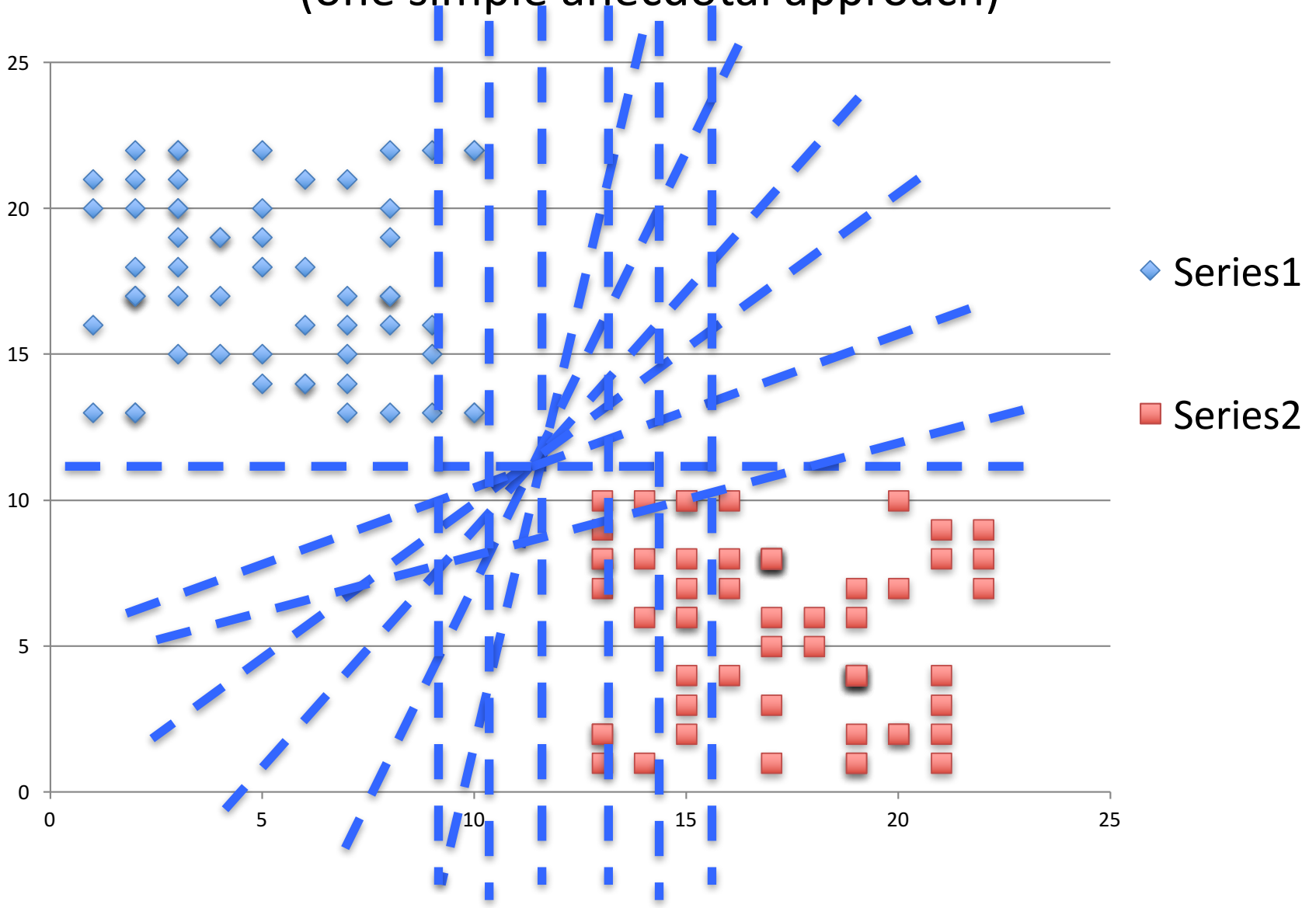
Supervised Learning

(one simple anecdotal approach)



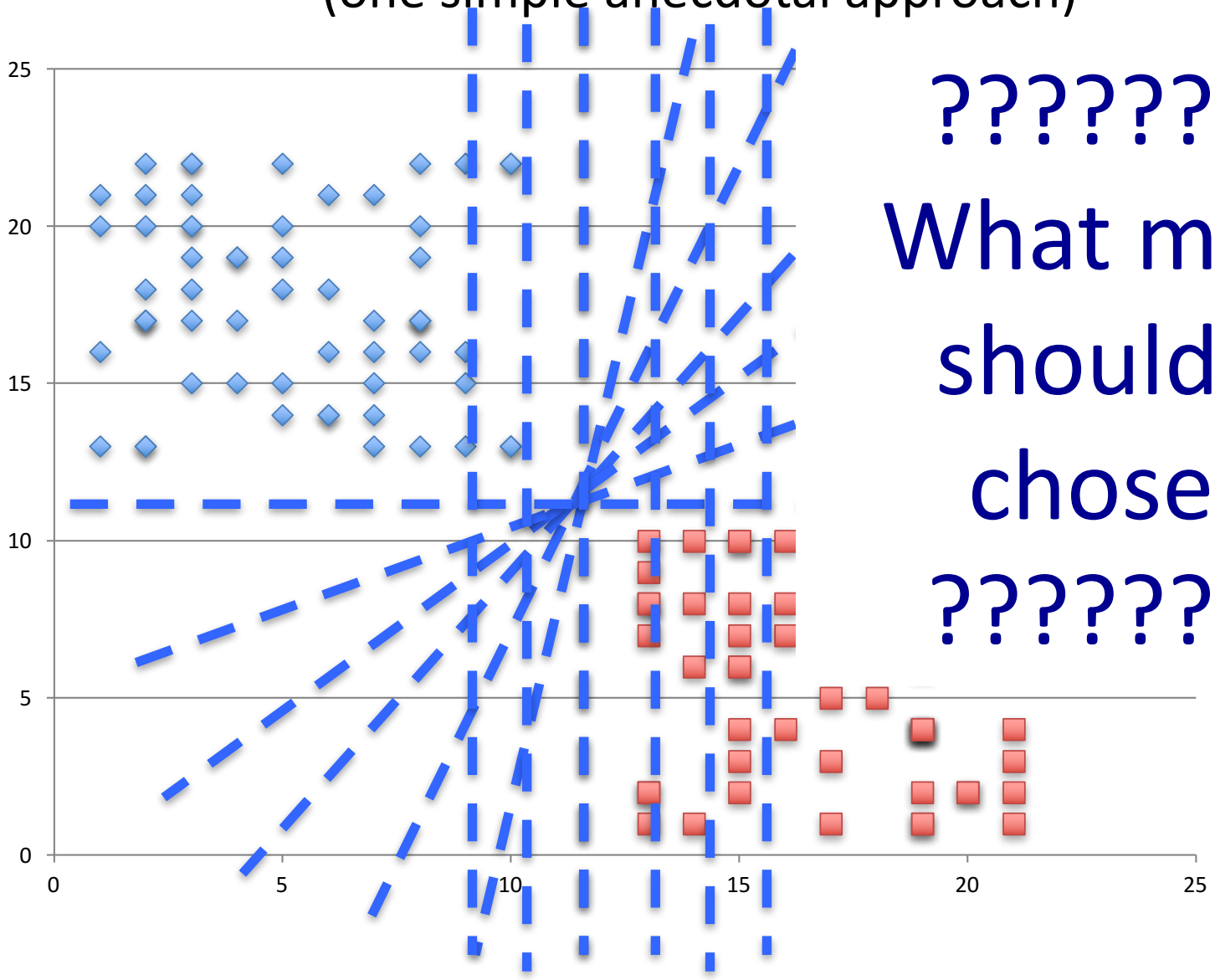
Supervised Learning

(one simple anecdotal approach)



Supervised Learning

(one simple anecdotal approach)



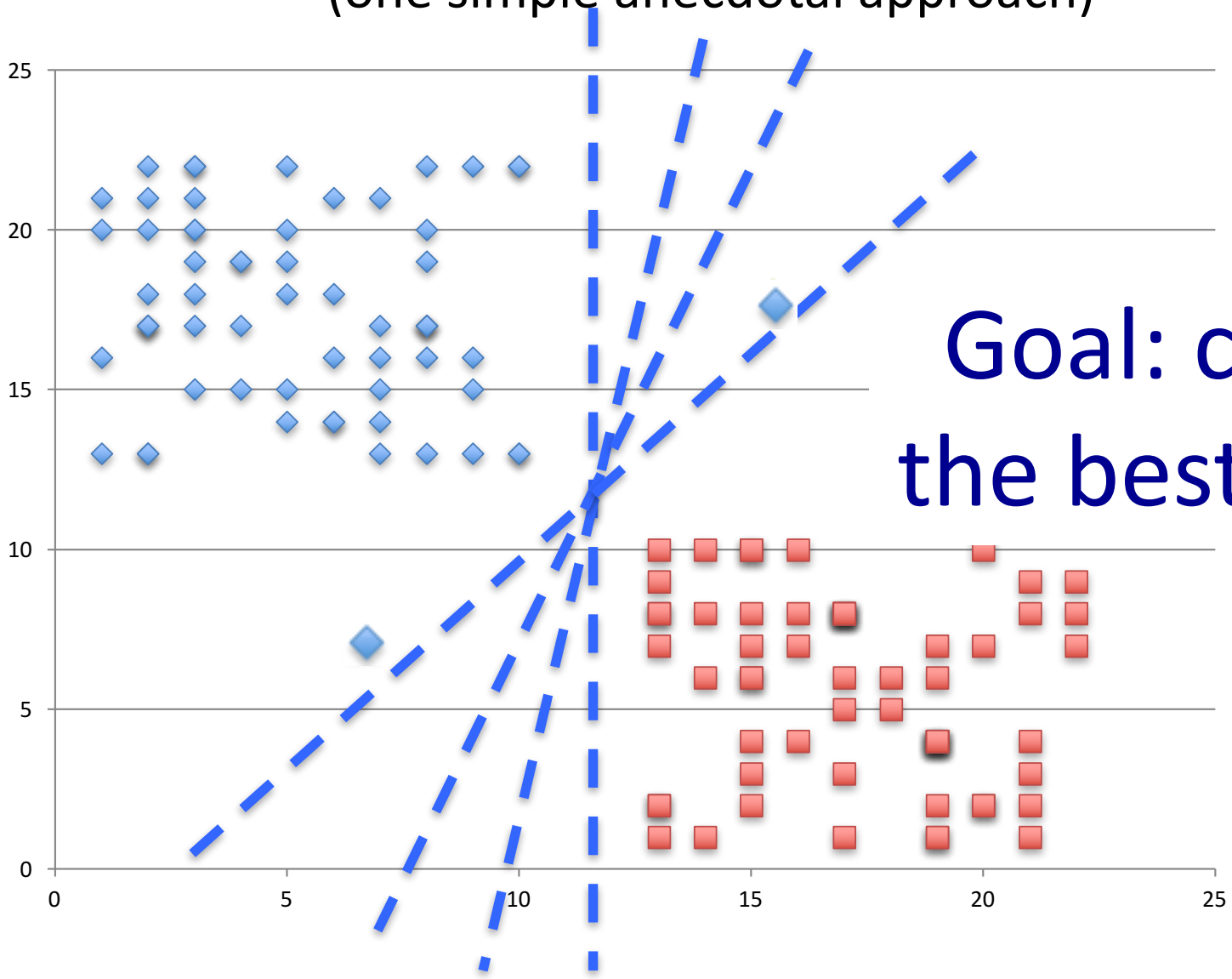
?????????

What model should be chosen?

?????????

Supervised Learning

(one simple anecdotal approach)



Goal: choose the best model

Unsupervised Learning

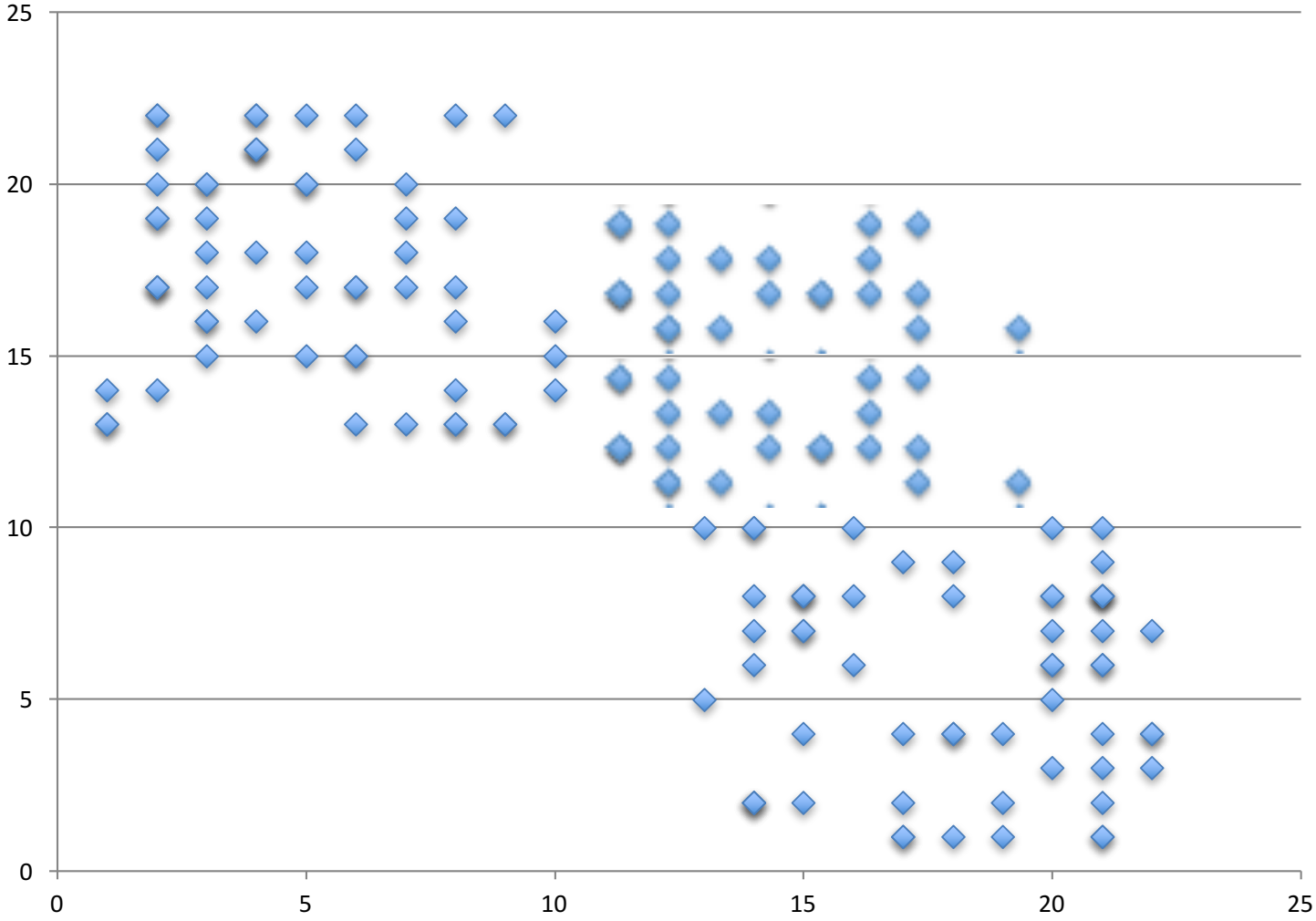
0.5	1	blue
2.9	1.9	blue
1.2	3.1	blue
0.8	4.7	blue
2.7	5.4	blue
8.1	4.7	red
8.3	6.6	red
6.3	6.7	red
8	9.1	red
5.4	8.4	red

Unsupervised Learning

0.5	1
2.9	1.9
1.2	3.1
0.8	4.7
2.7	5.4
8.1	4.7
8.3	6.6
6.3	6.7
8	9.1
5.4	8.4

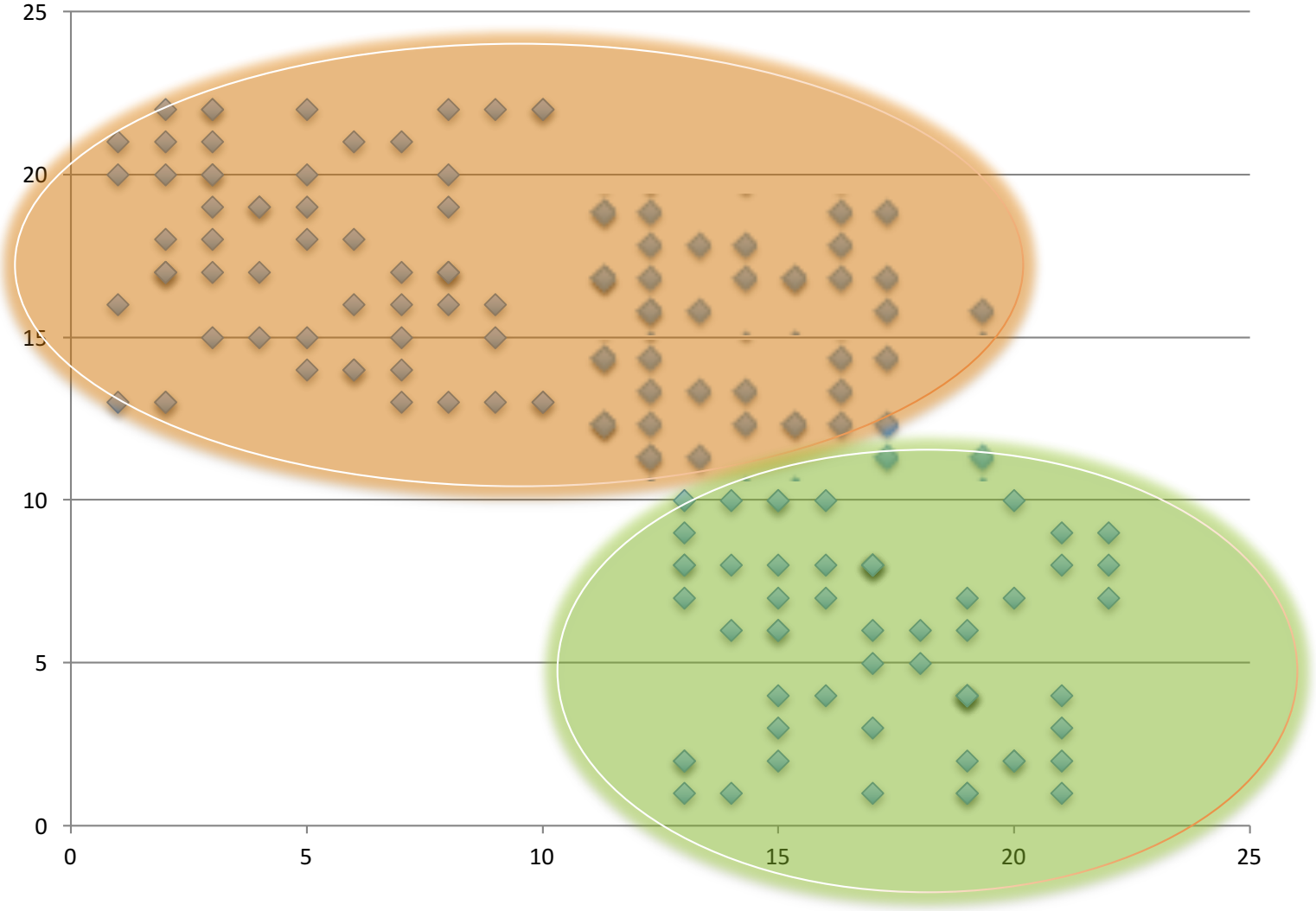
Unsupervised Learning

(one simple anecdotal approach)



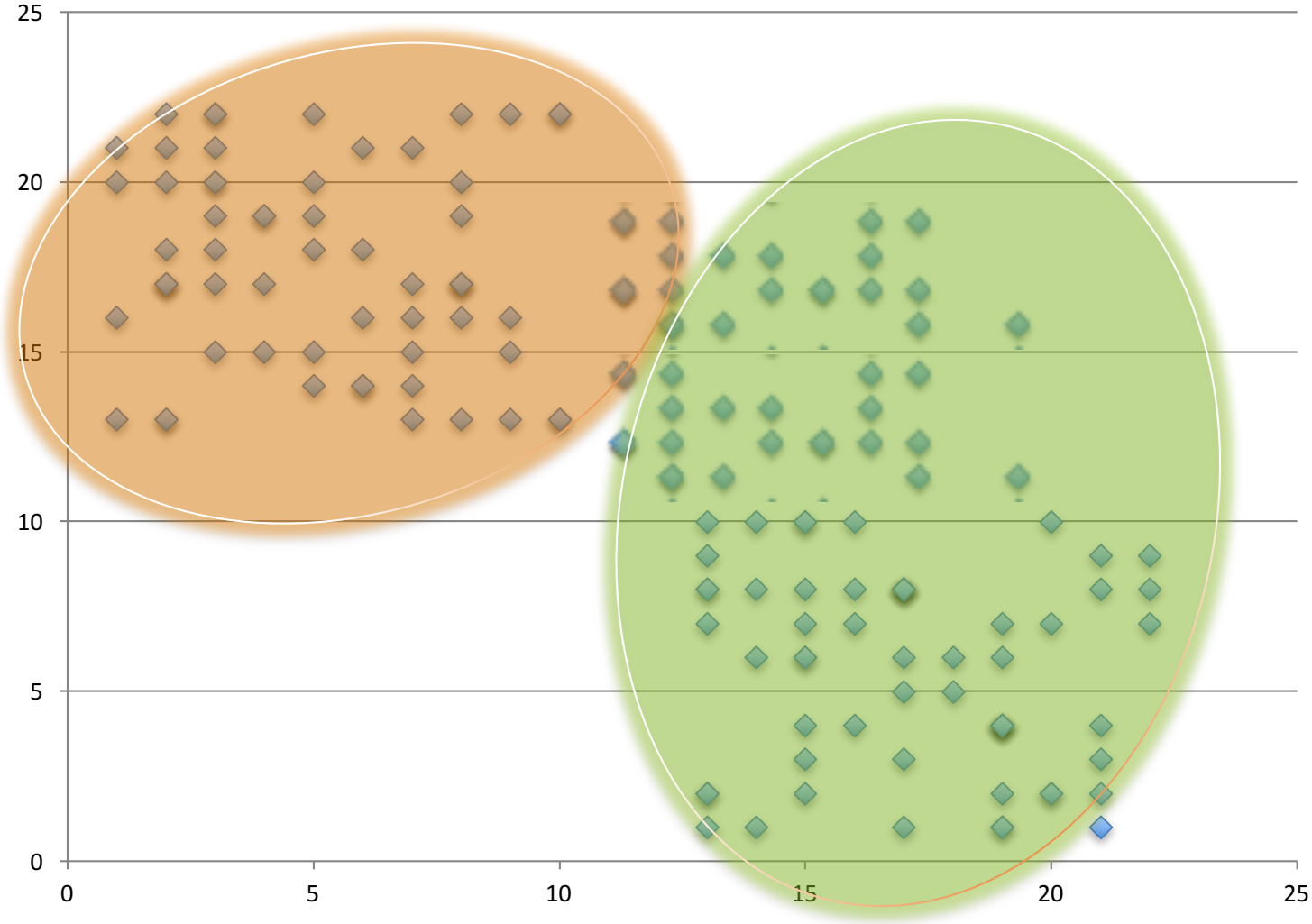
Unsupervised Learning

(one simple anecdotal approach)



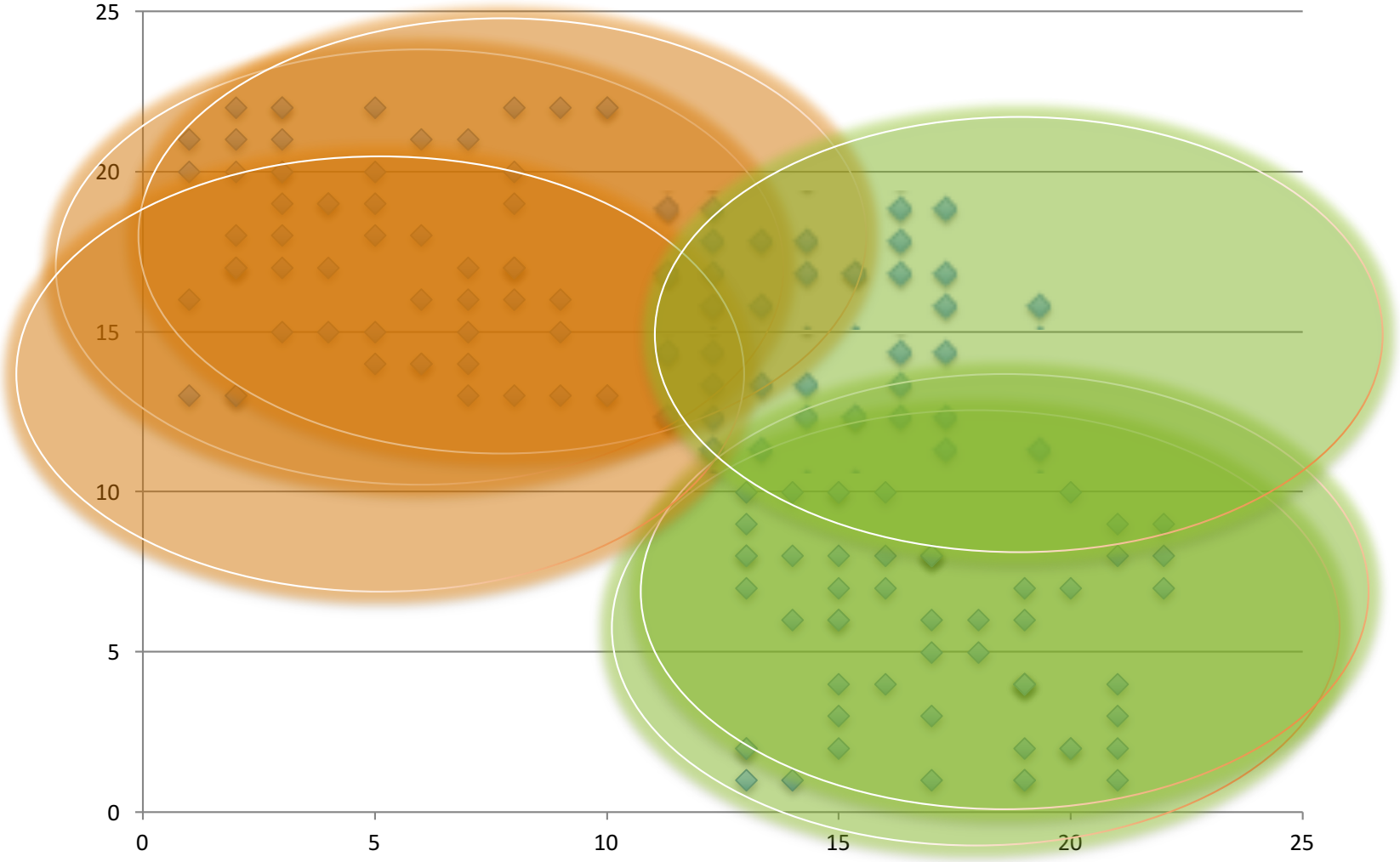
Unsupervised Learning

(one simple anecdotal approach)



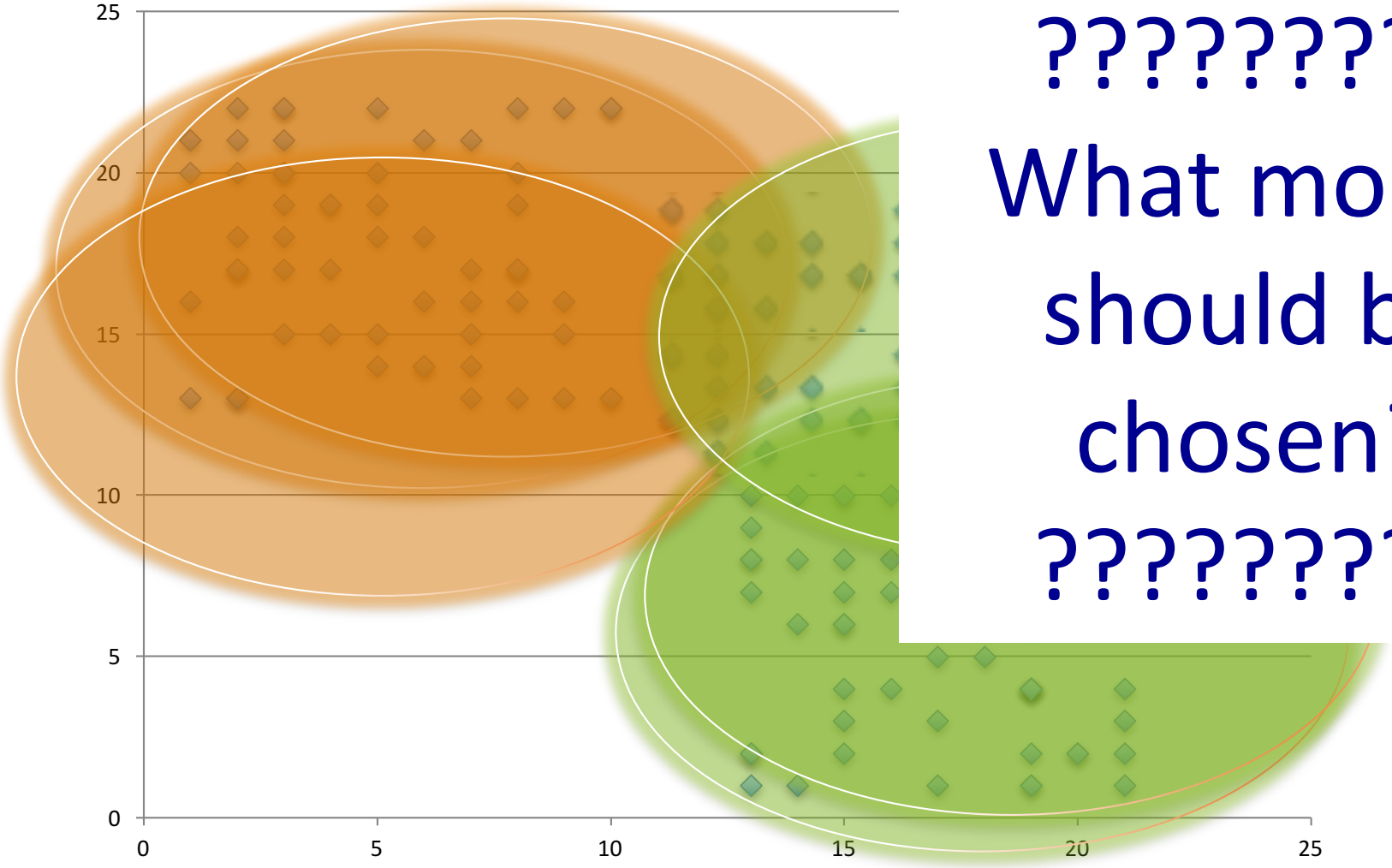
Unsupervised Learning

(one simple anecdotal approach)



Unsupervised Learning

(one simple anecdotal approach)



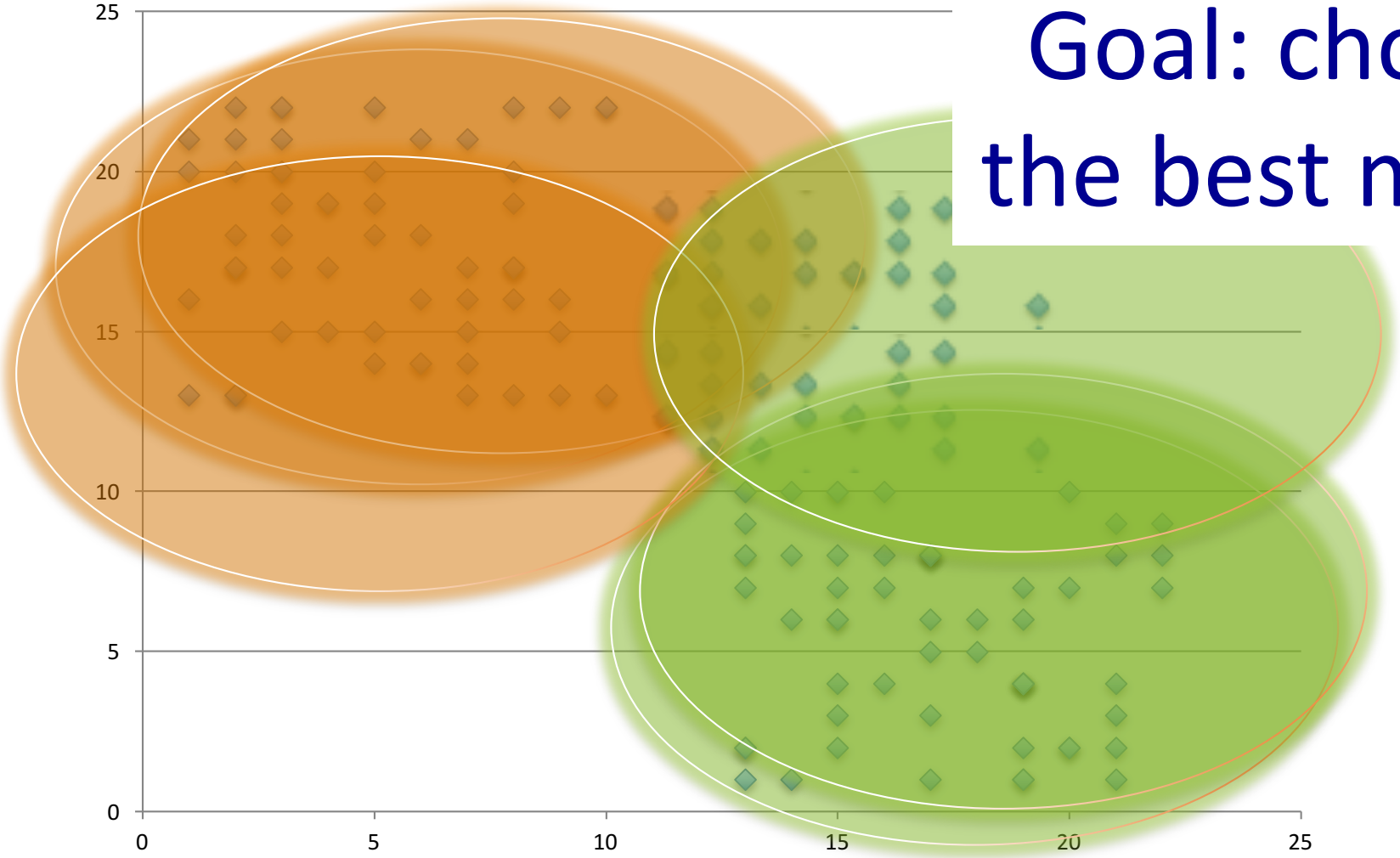
?????????

What model should be chosen?

?????????

Unsupervised Learning

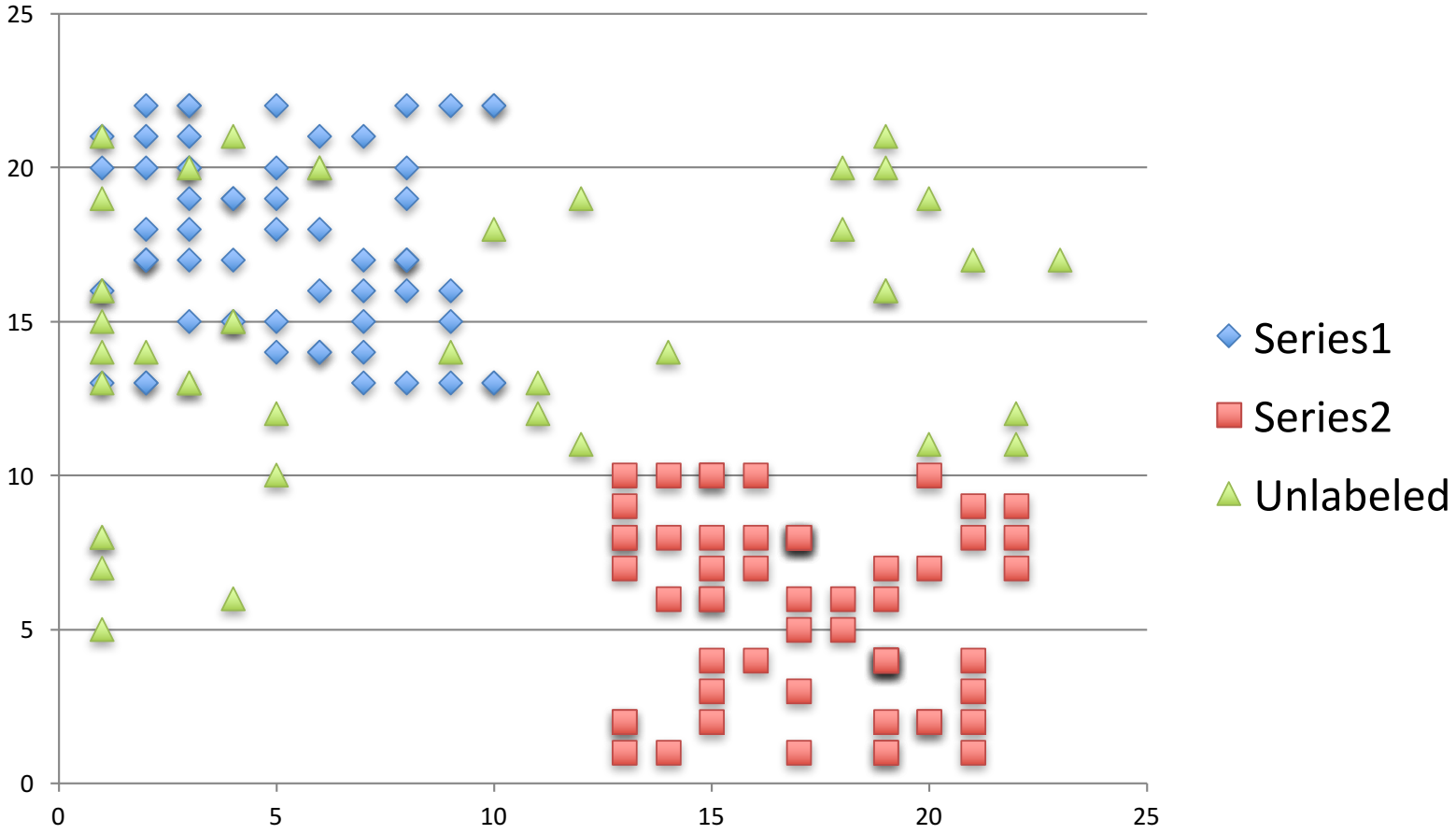
(one simple anecdotal approach)



Goal: choose the best model

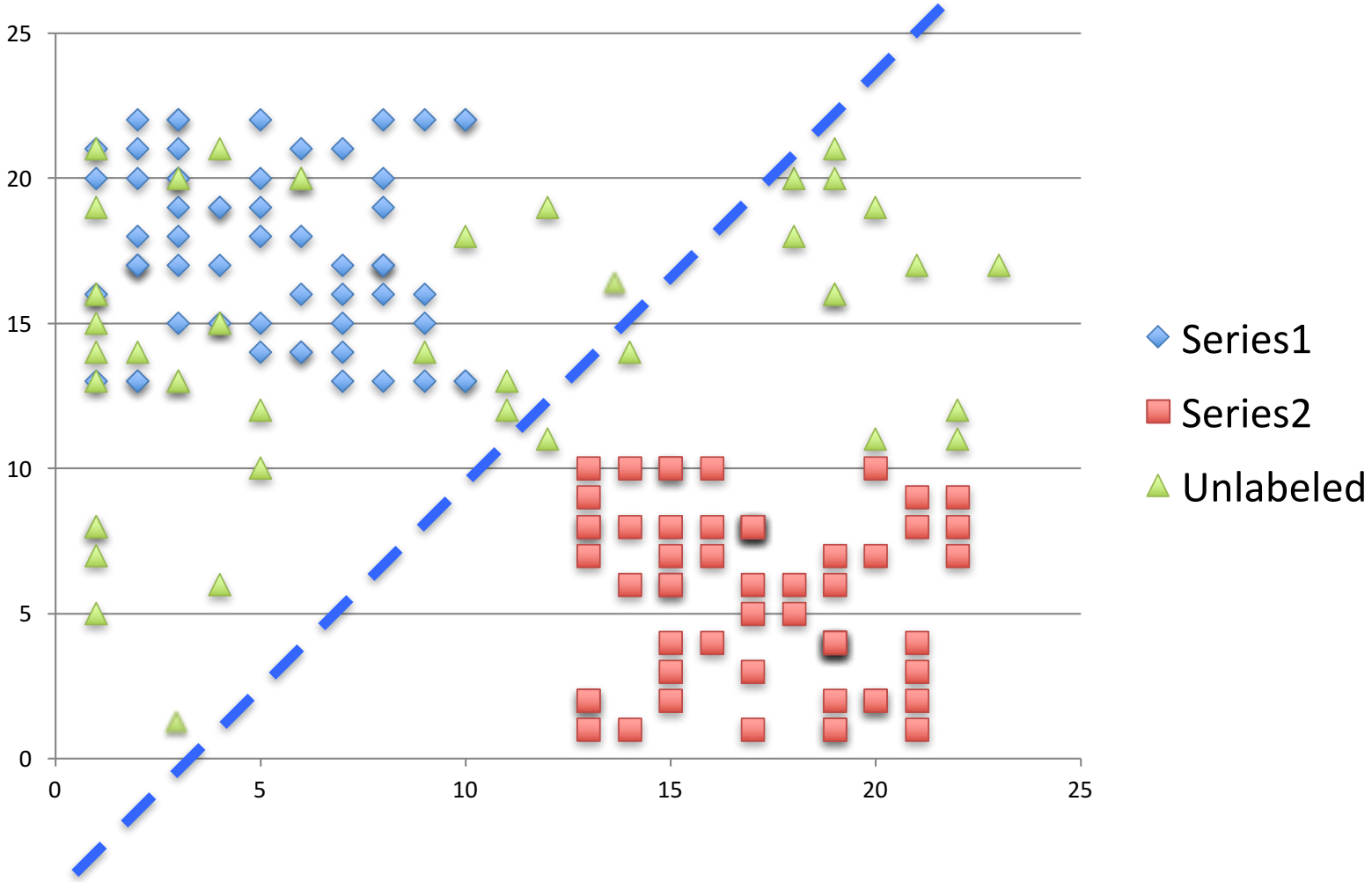
Semi-supervised Learning

(one simple anecdotal approach)



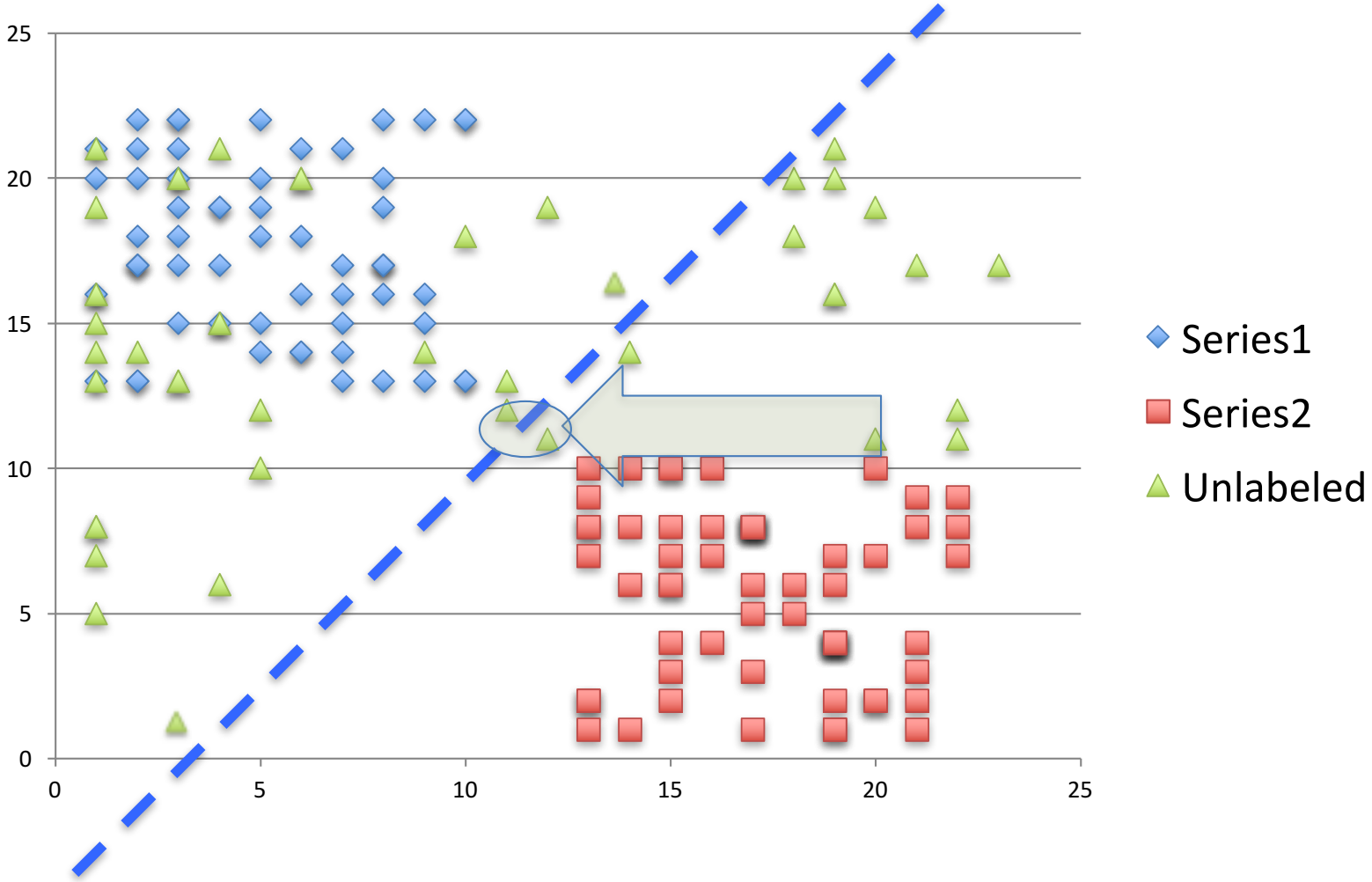
Semi-supervised Learning

(one simple anecdotal approach)



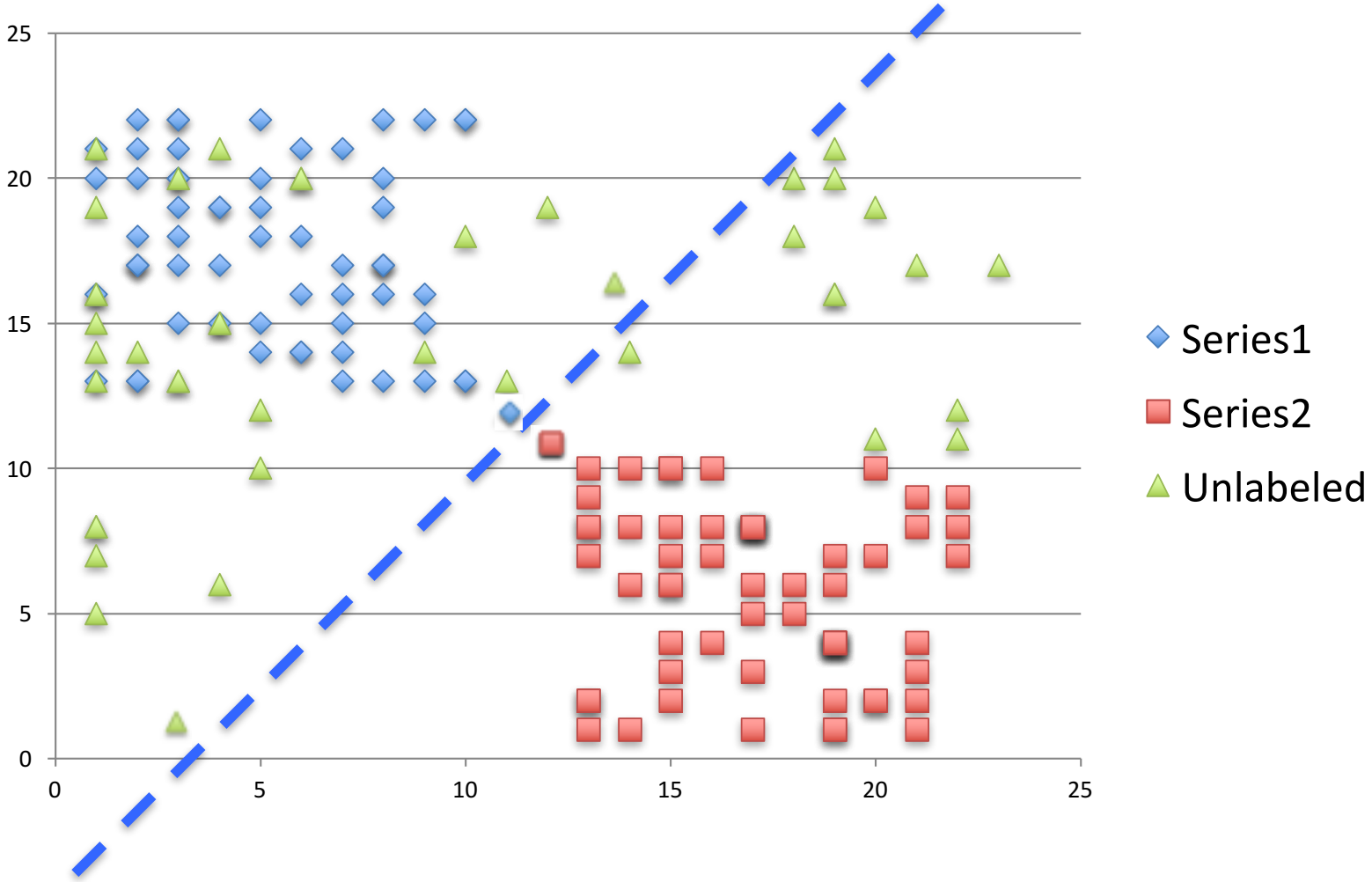
Semi-supervised Learning

(one simple anecdotal approach)



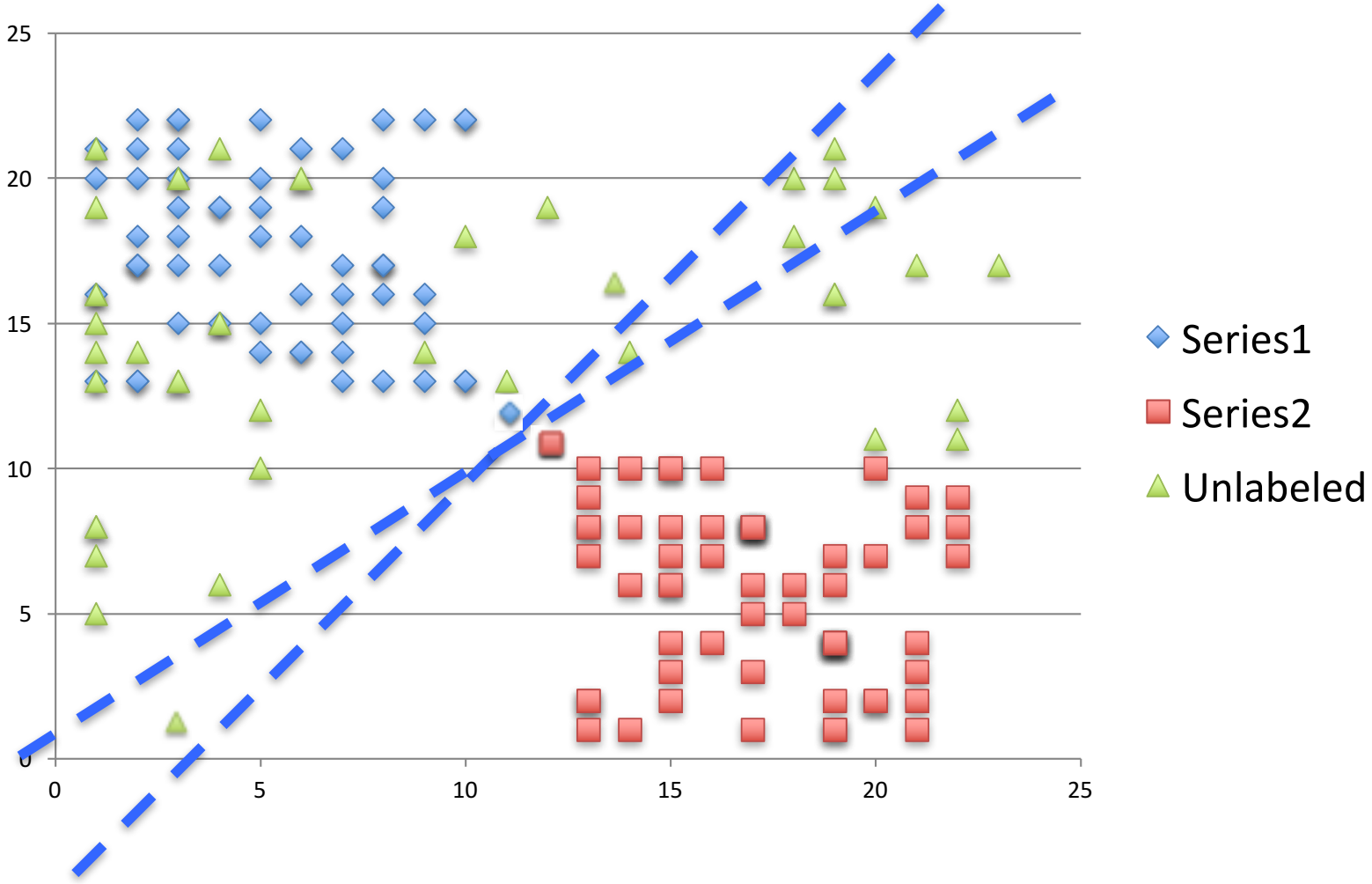
Semi-supervised Learning

(one simple anecdotal approach)



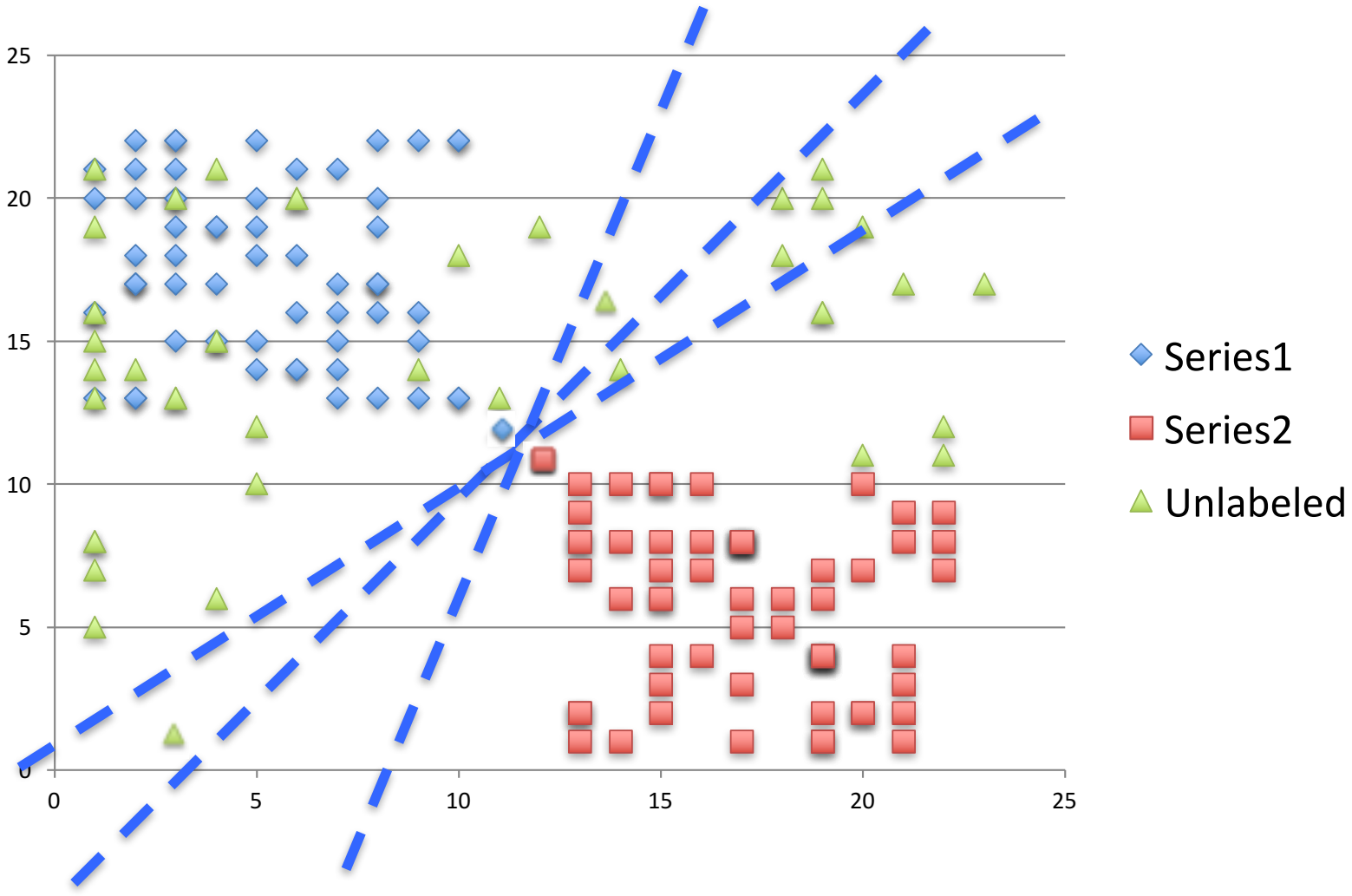
Semi-supervised Learning

(one simple anecdotal approach)



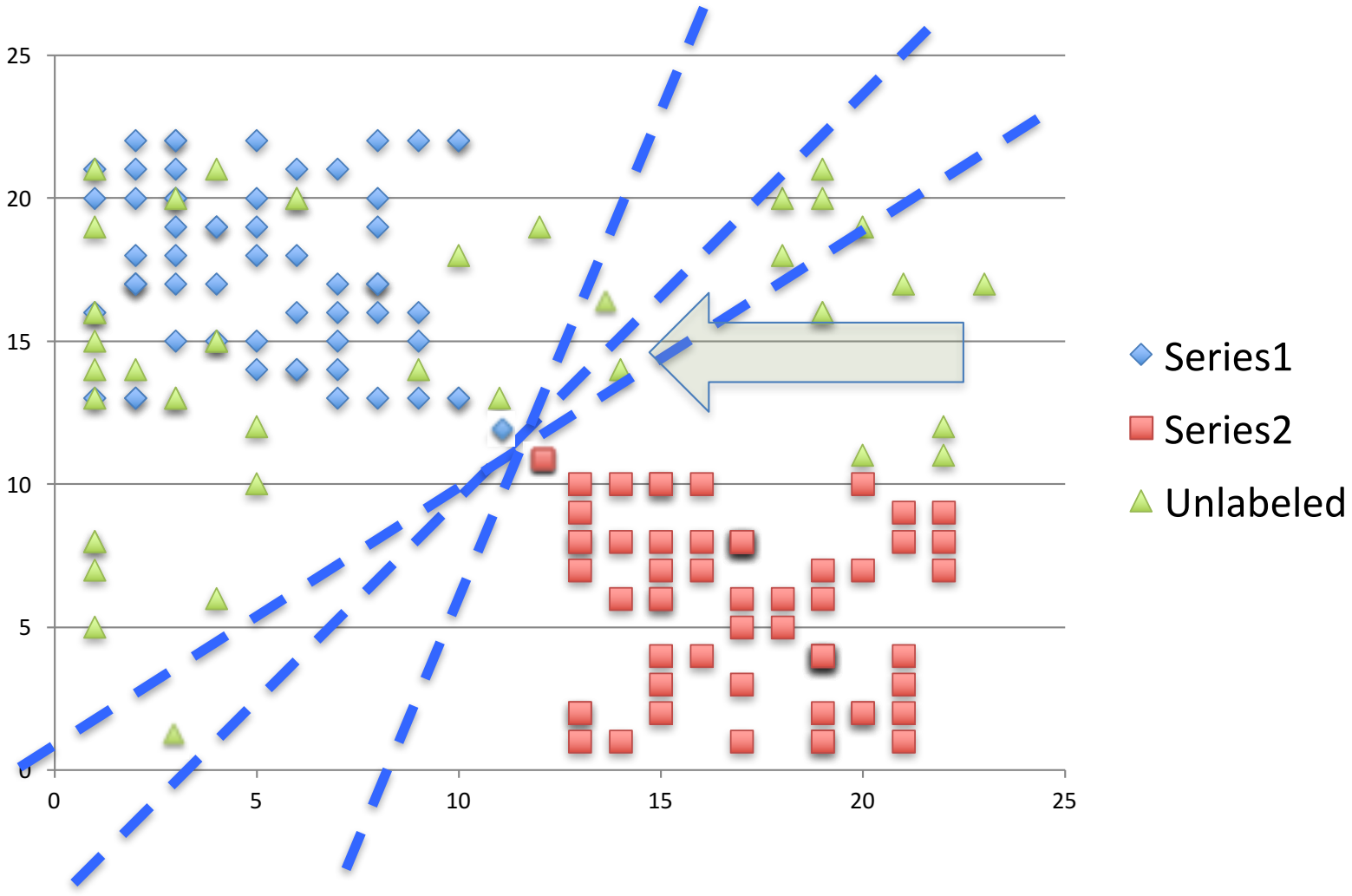
Semi-supervised Learning

(one simple anecdotal approach)



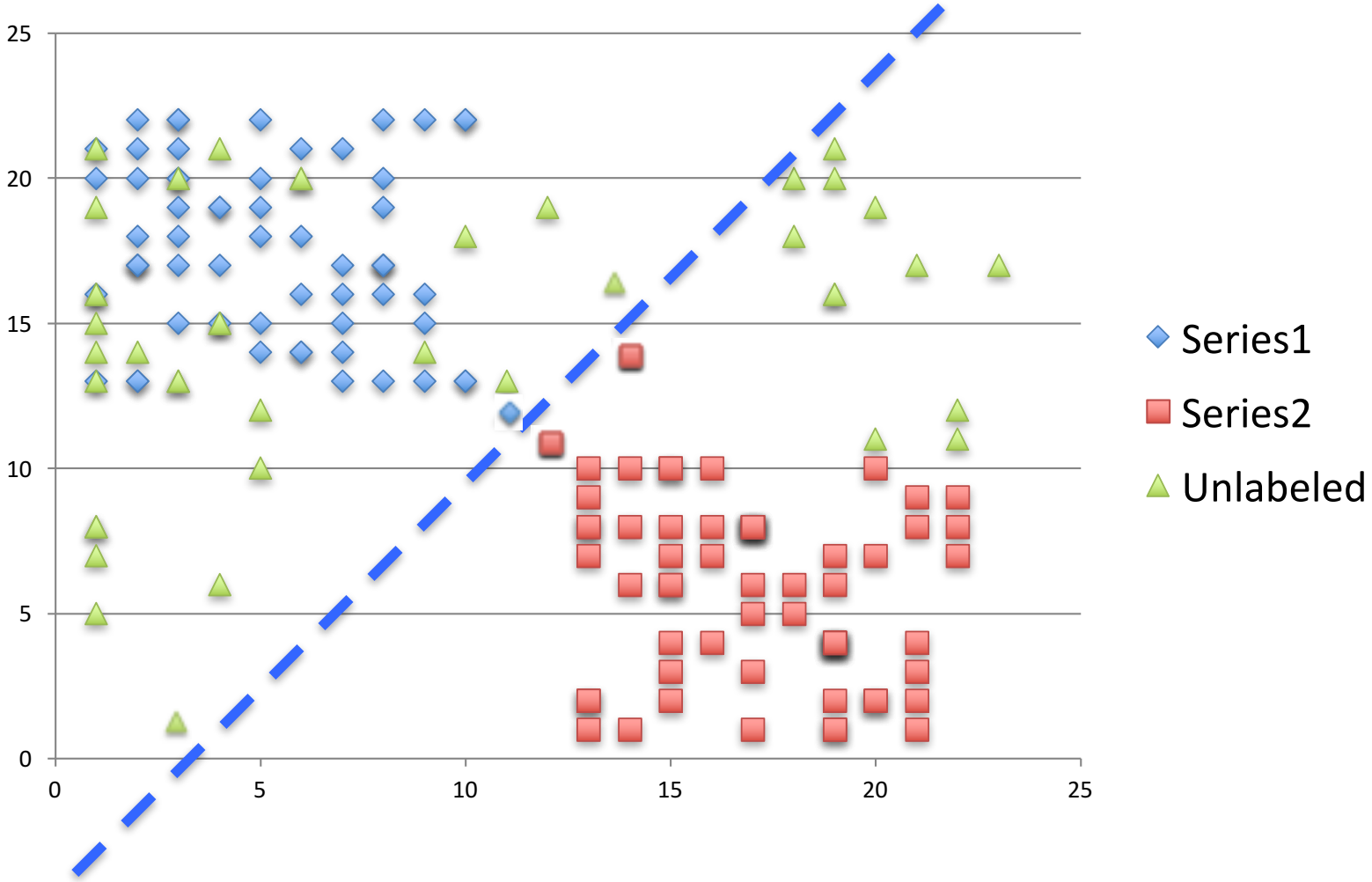
Semi-supervised Learning

(one simple anecdotal approach)



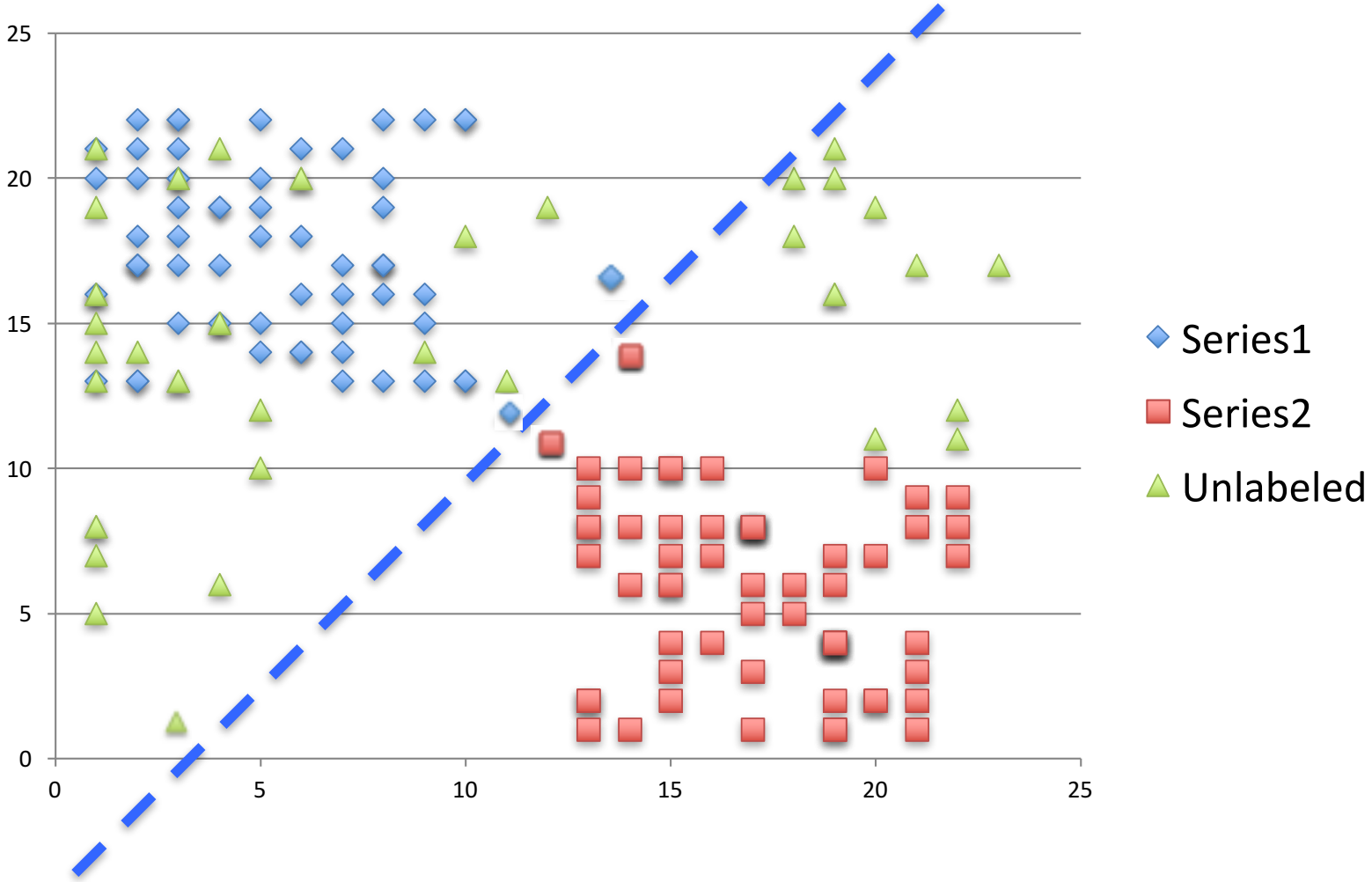
Semi-supervised Learning

(one simple anecdotal approach)



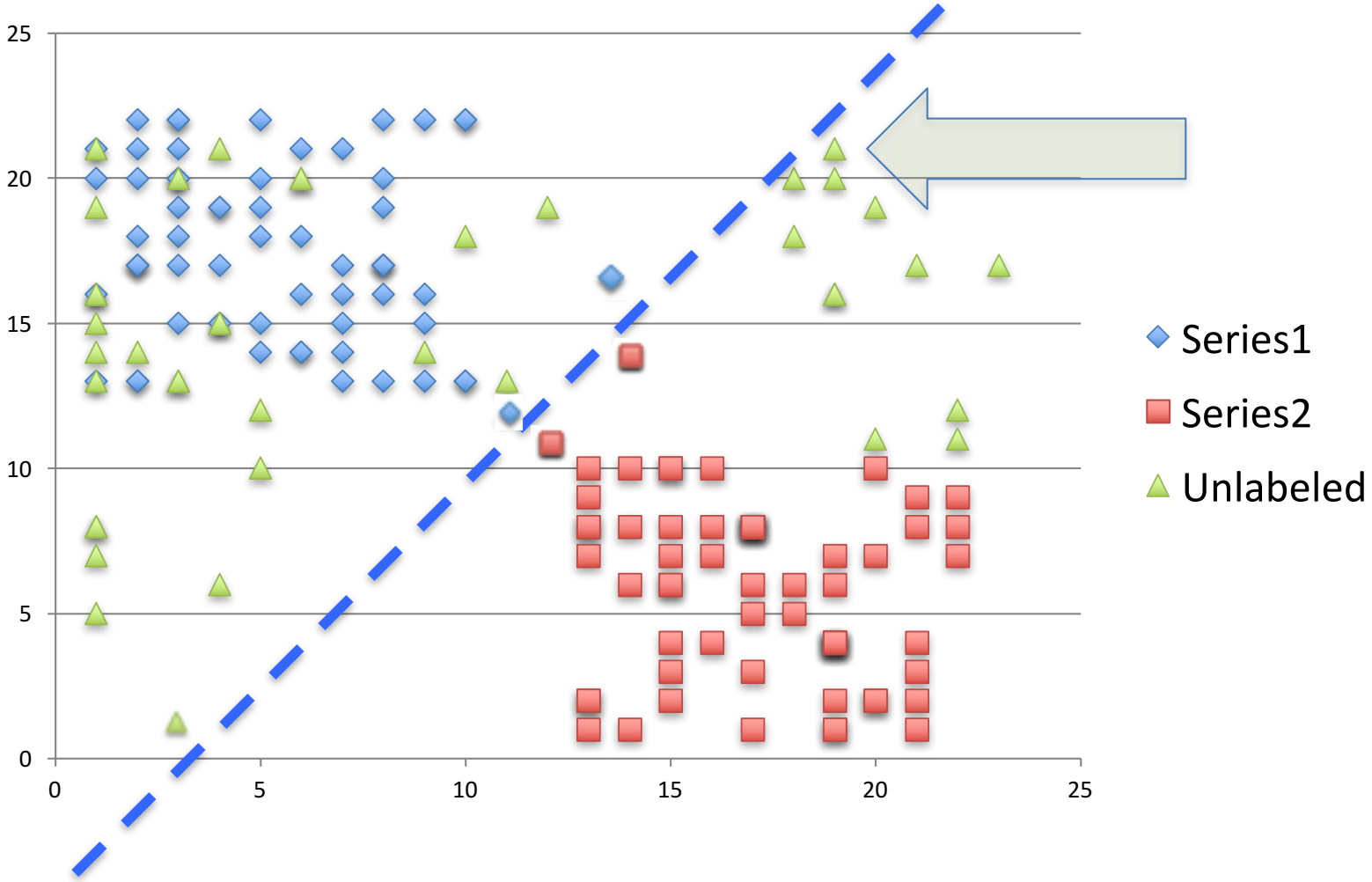
Semi-supervised Learning

(one simple anecdotal approach)



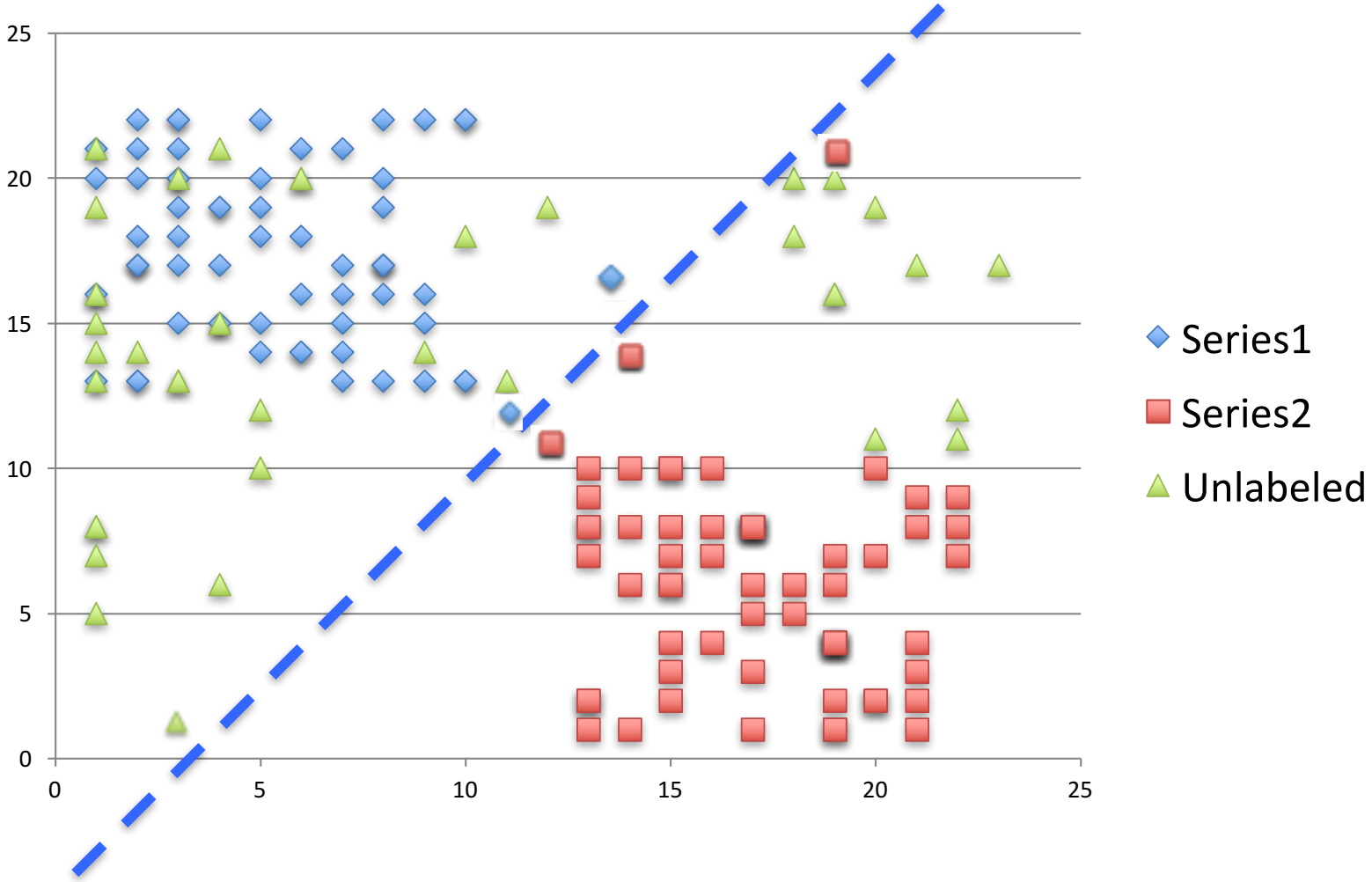
Semi-supervised Learning

(one simple anecdotal approach)



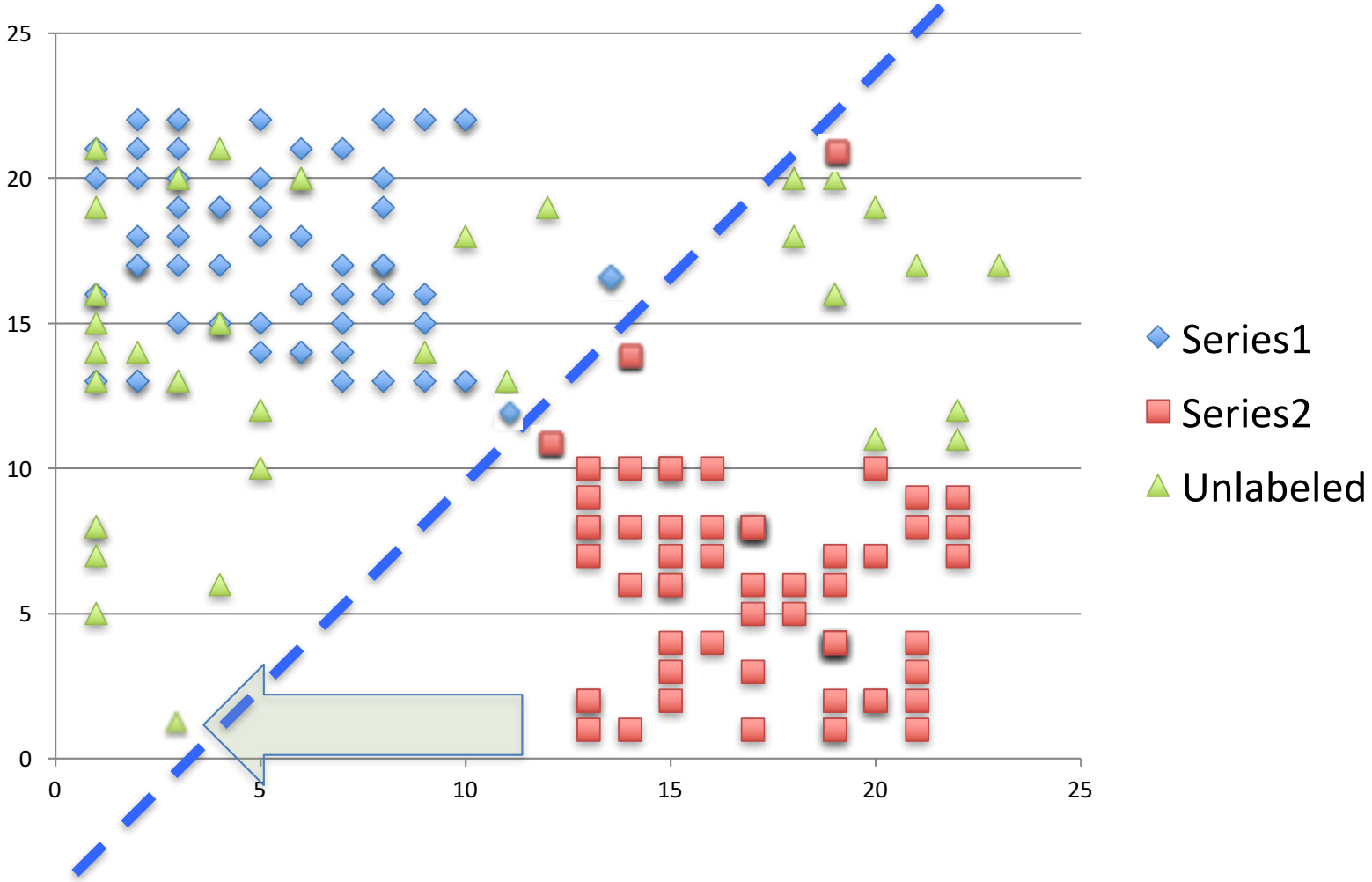
Semi-supervised Learning

(one simple anecdotal approach)



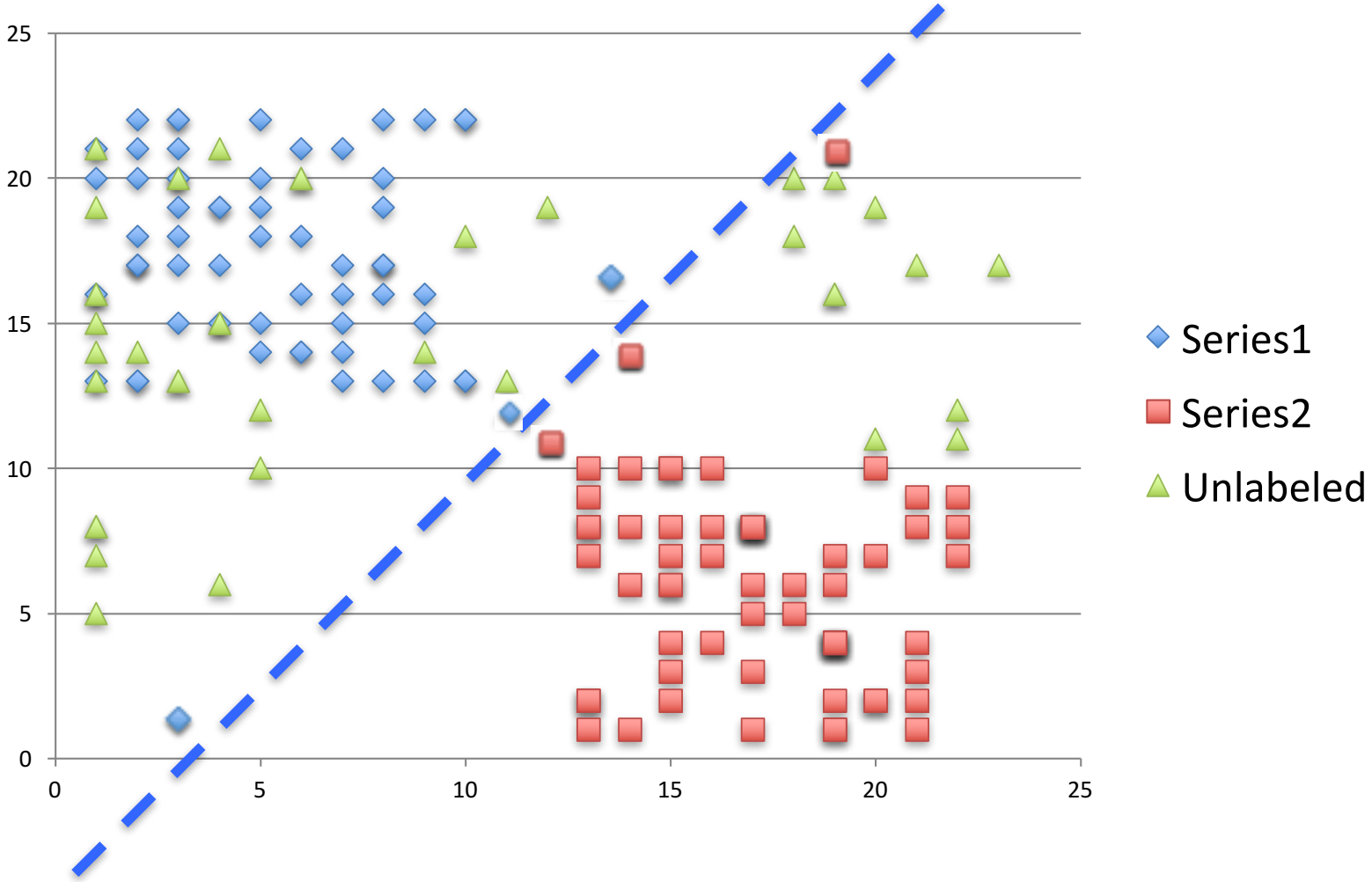
Semi-supervised Learning

(one simple anecdotal approach)



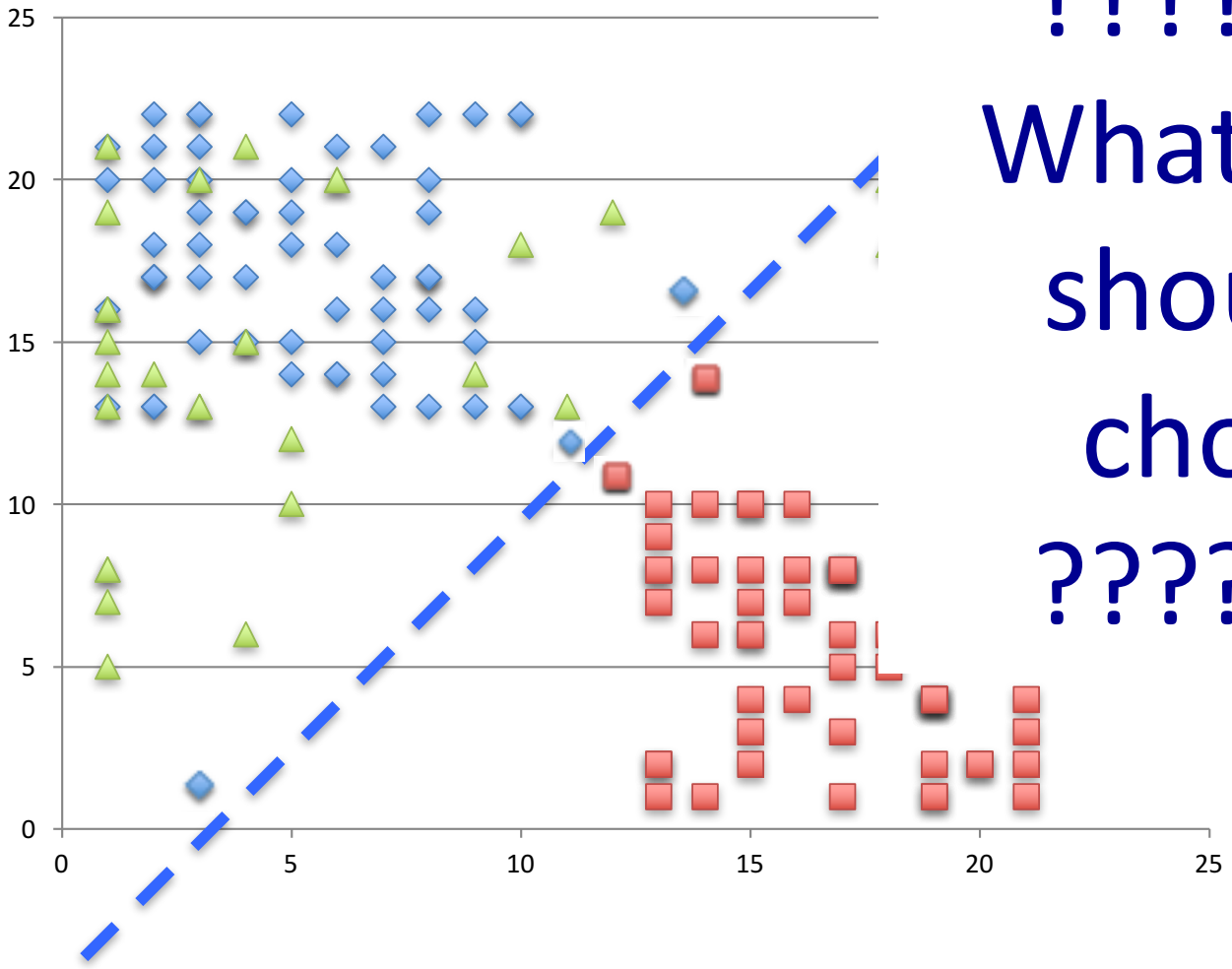
Semi-supervised Learning

(one simple anecdotal approach)



Semi-supervised Learning

(one simple anecdotal approach)



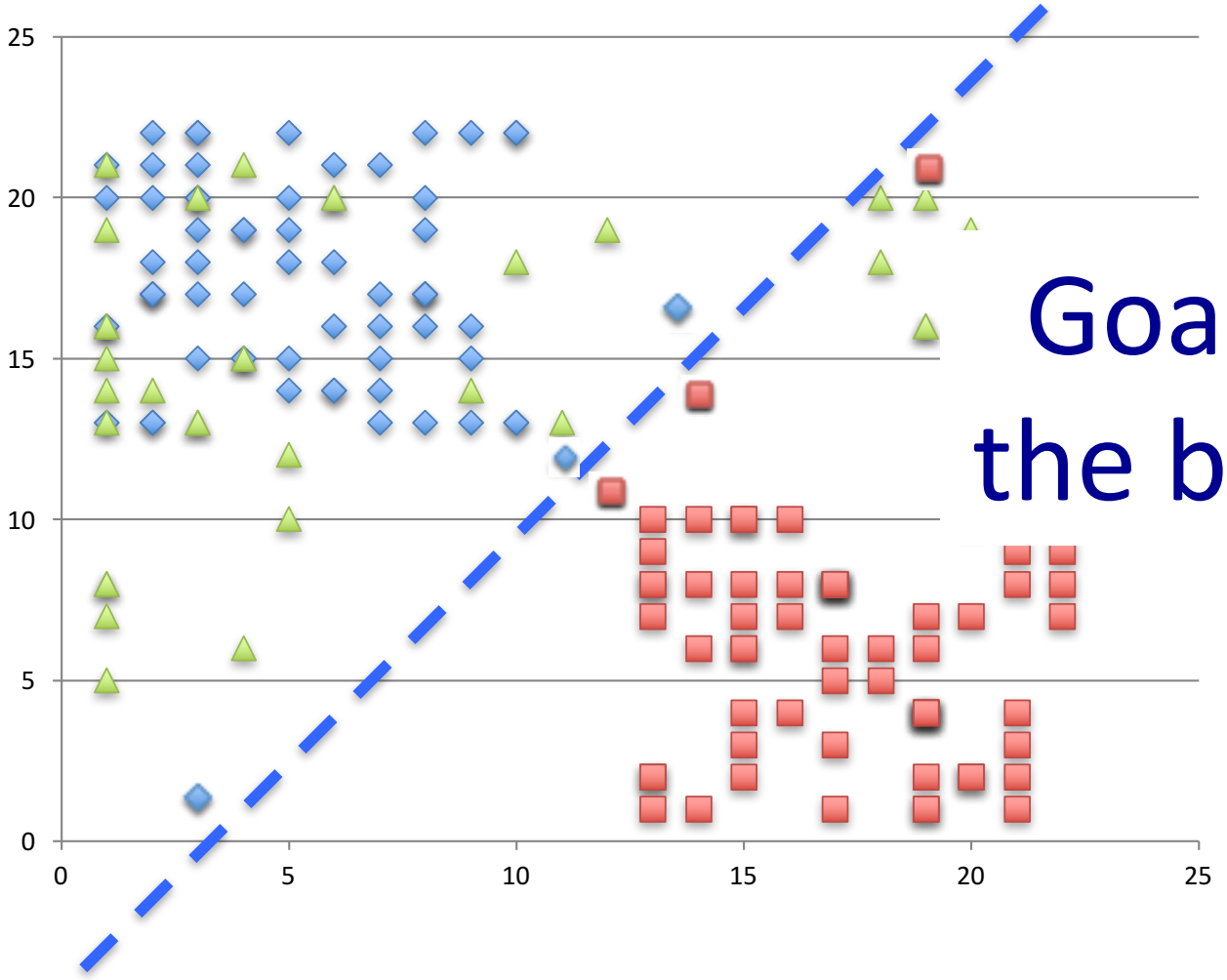
?????????

What model should be chosen?

?????????

Semi-supervised Learning

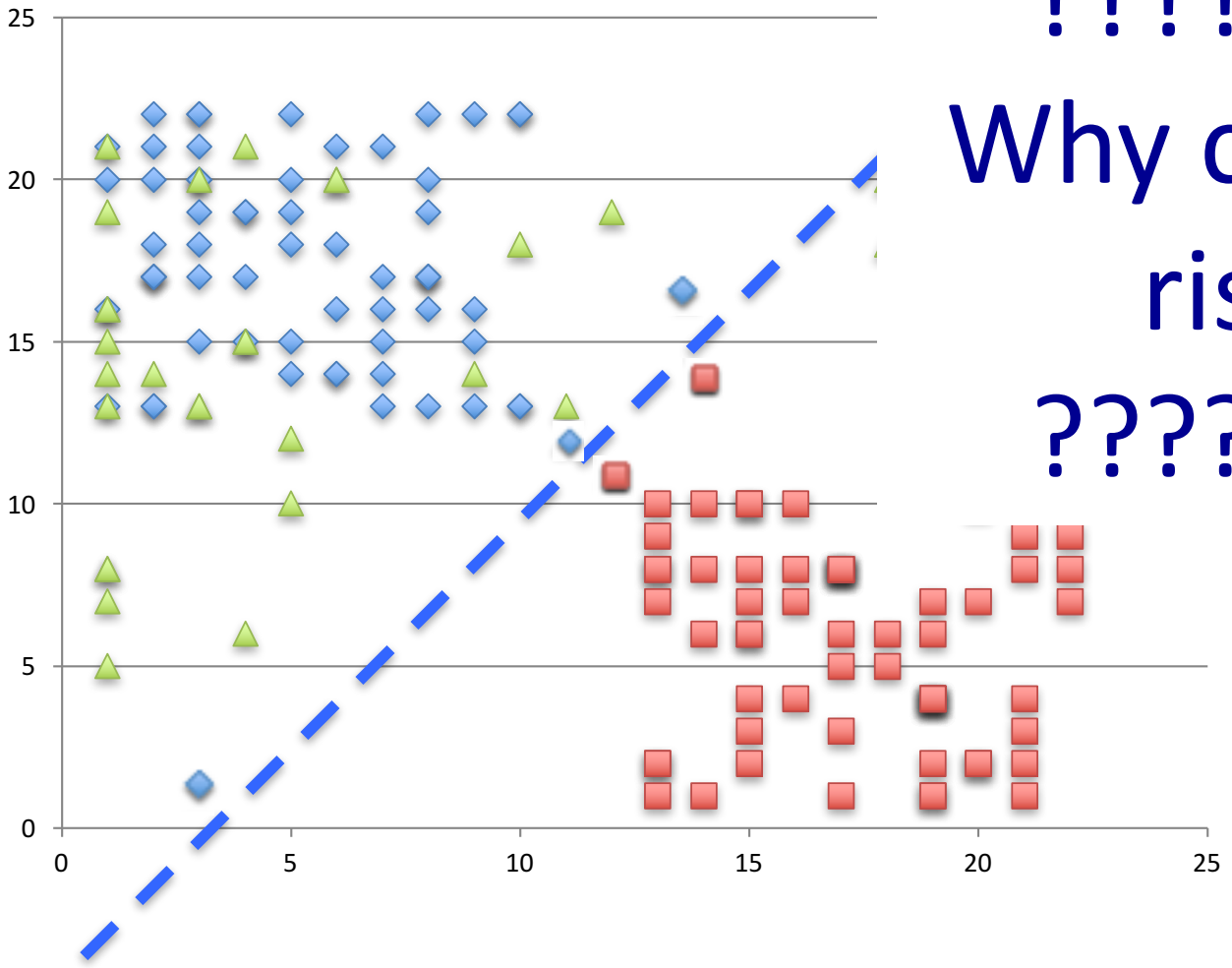
(one simple anecdotal approach)



Goal: choose the best model

Semi-supervised Learning

(one simple anecdotal approach)



??????????

Why can it be risky?

??????????

Outline

- Machine Learning
- **Machine Reading**
- Reading the Web
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Machine Reading

- "The autonomous understanding of text" [Etzioni et al., 2007]
- "One of the most important methods by which human beings learn is by reading" [Clark et al., 2007], thus why not building machines capable of learning by reading?

Machine Reading

- "The problem of deciding what was implied by a written text, of reading between the lines is the problem of inference." [Norvig, 2007]

- Typically, Machine Reading is different from Natural Language Processing alone

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, specializes that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same same same same same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible]inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[illegible]overing that Salander has hacked into his computer, he persuade[illegible] assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible]gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
same
same
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] te on th[illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist ag[illegible] spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible] inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.
Af[illegible] overing that Salander has hacked into his c[illegible] persuade[illegible] assist him with research. They even[illegible] lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.
A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible] gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
affairWith
affairWith
enemyOf
same
same
headOf

This slide was adapted from [Hady et al., 2011]

Federal University of São Carlos

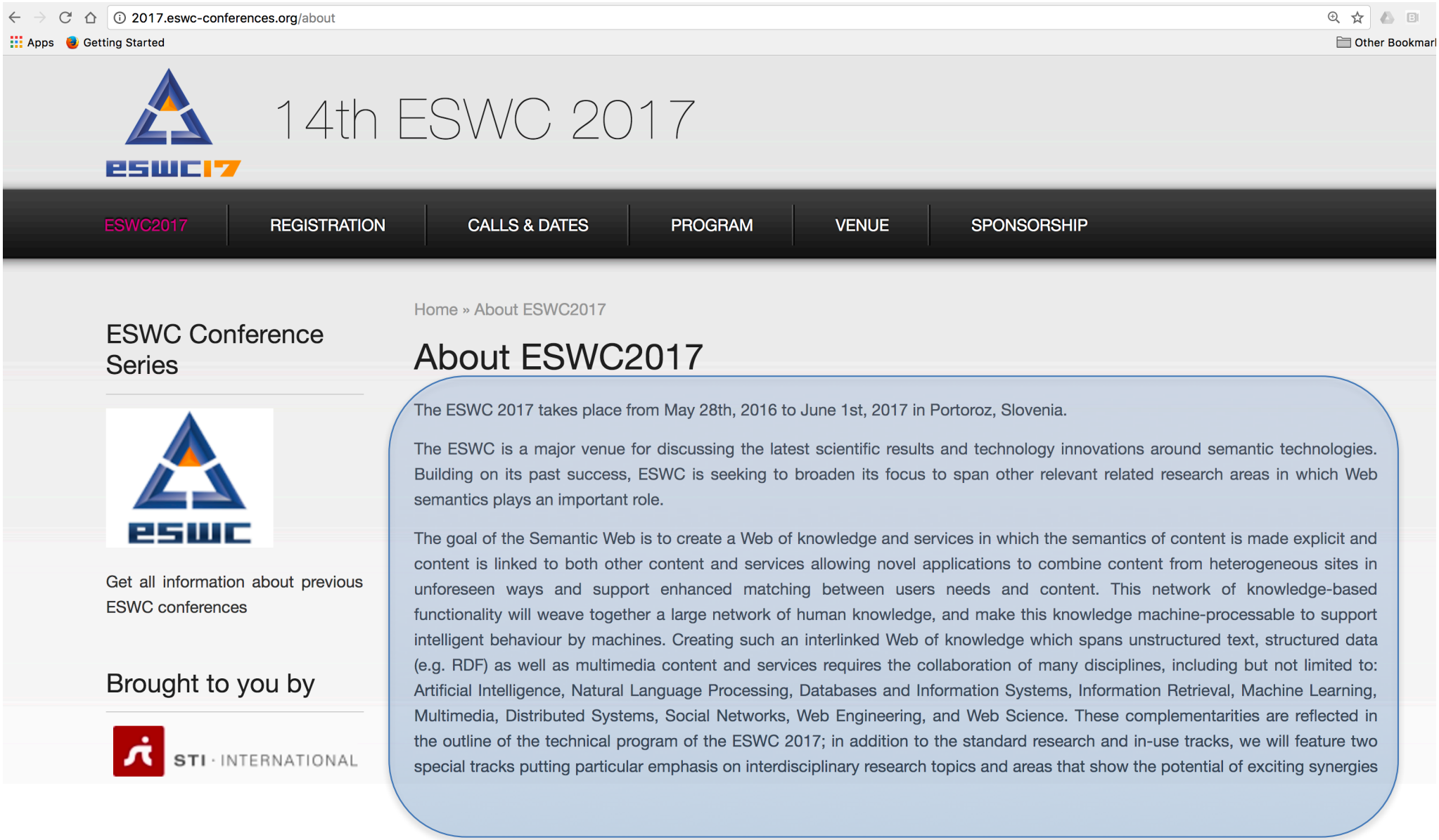
Machine Reading

FAPESP

Carnegie Mellon University

Federal University of São Carlos

Machine Reading



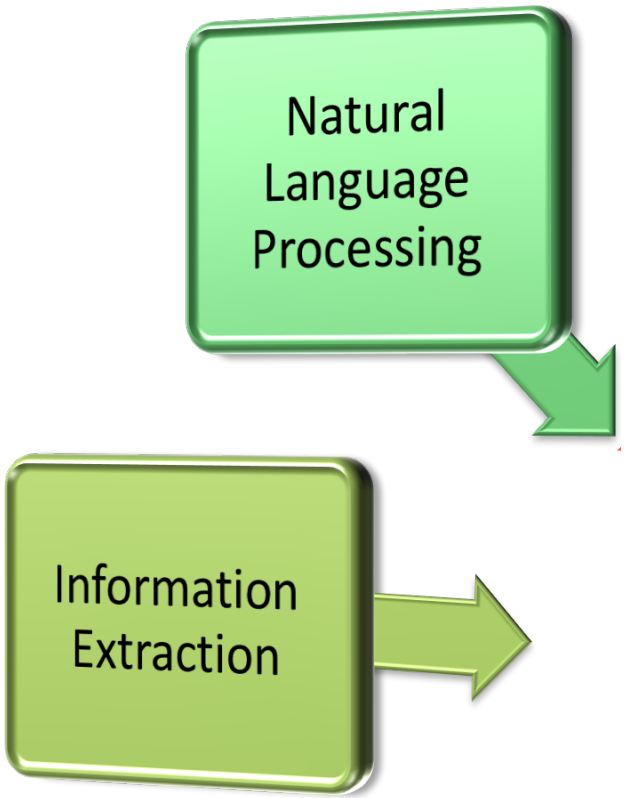
FAPESP

Carnegie Mellon University

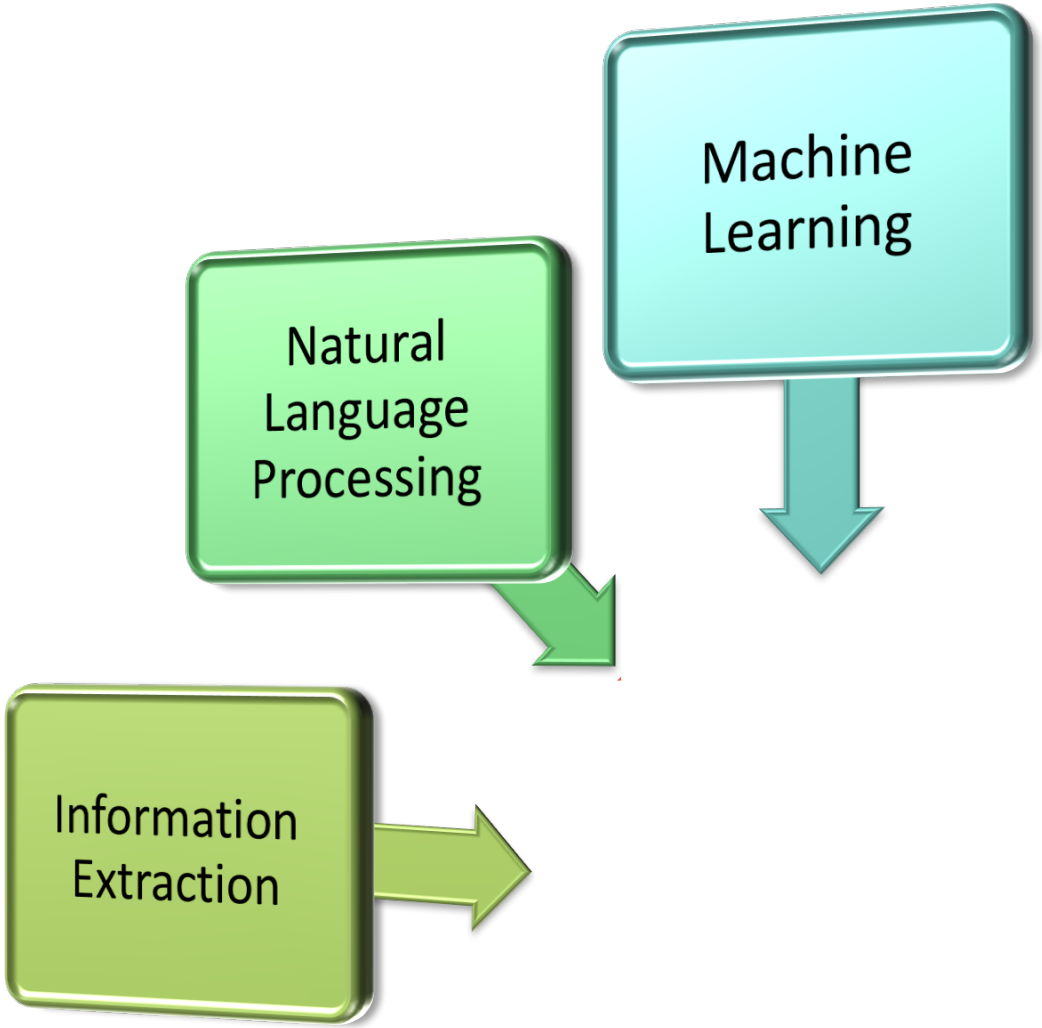
Machine Reading

Machine Reading

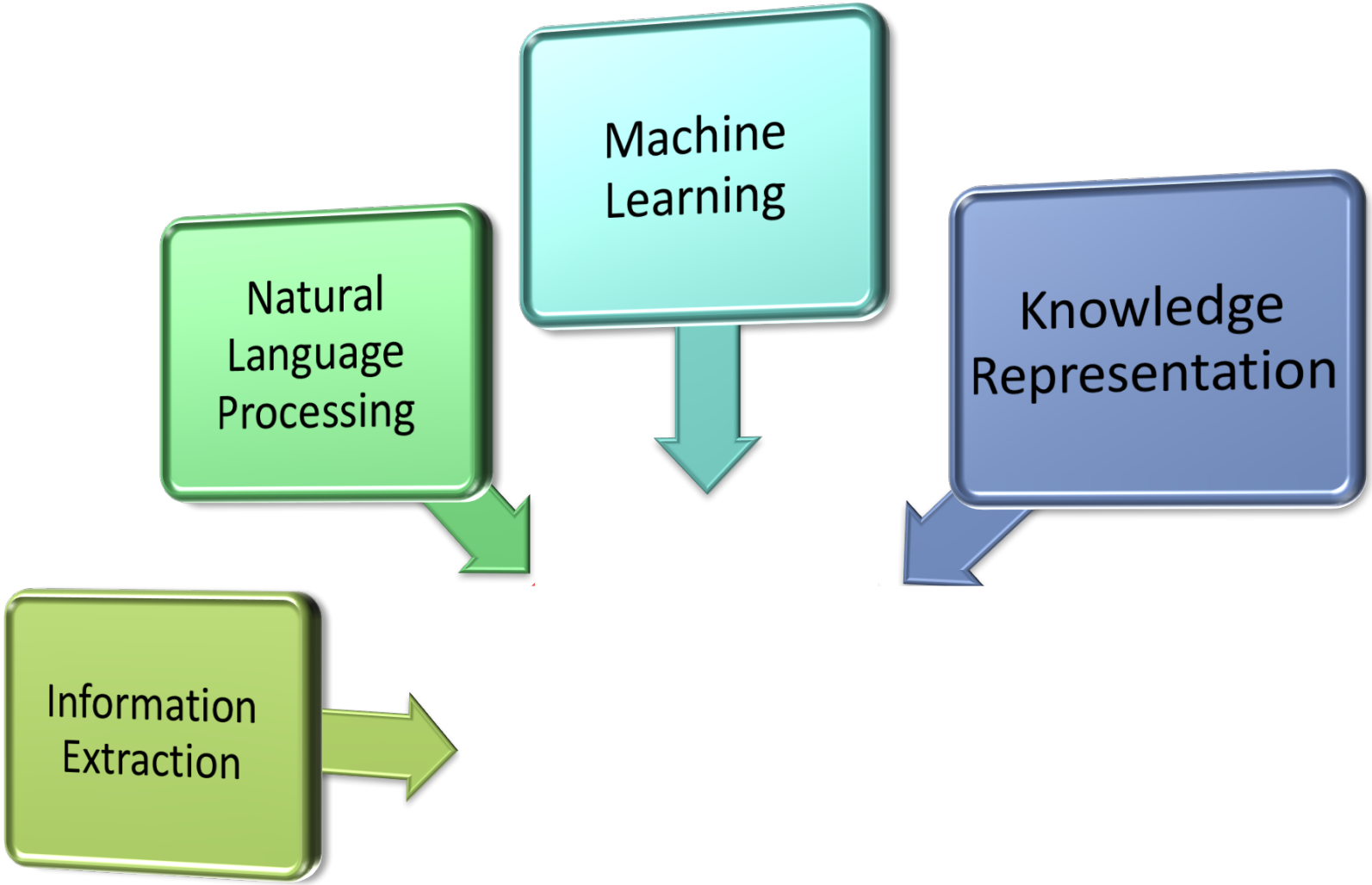
Machine Reading



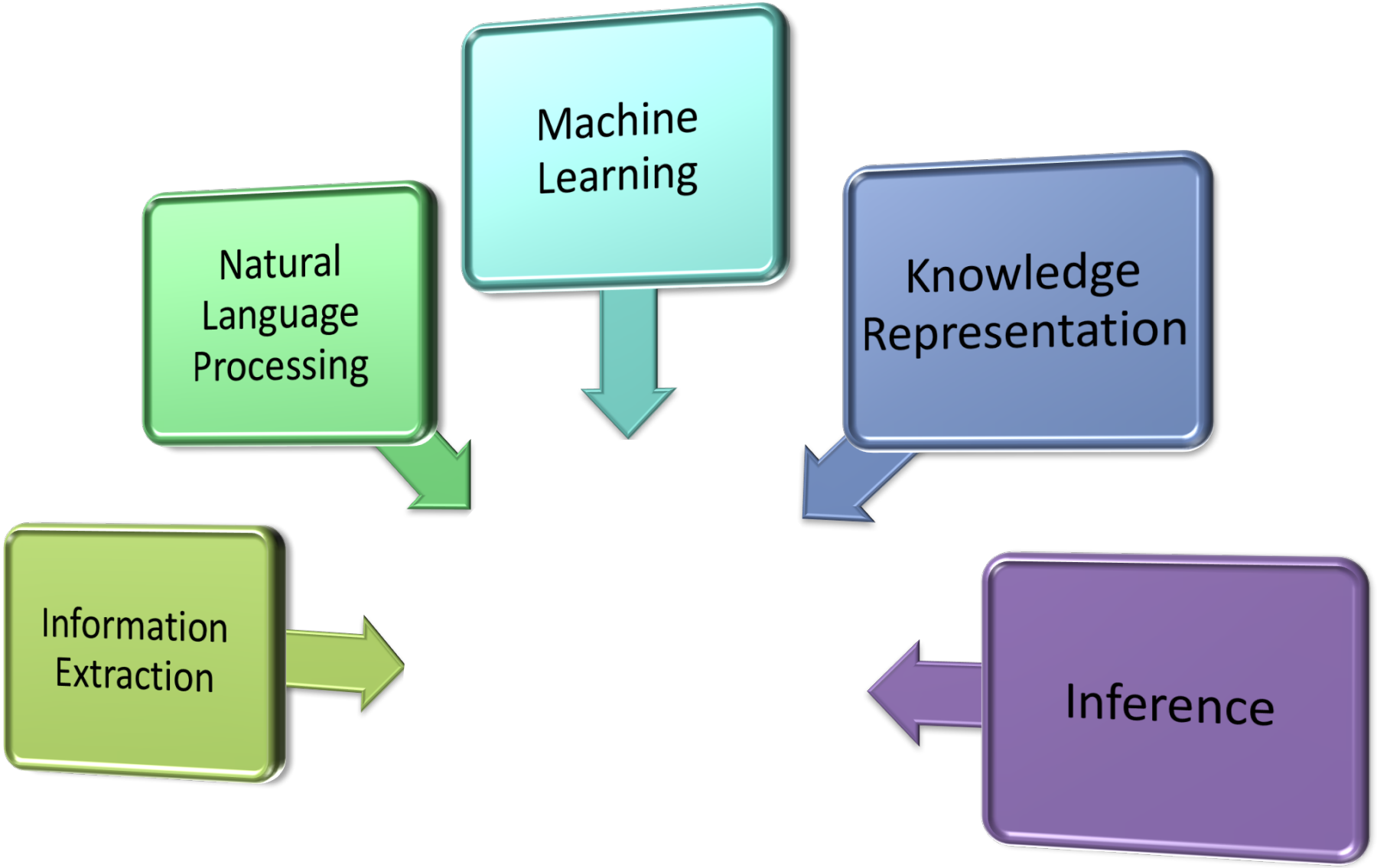
Machine Reading



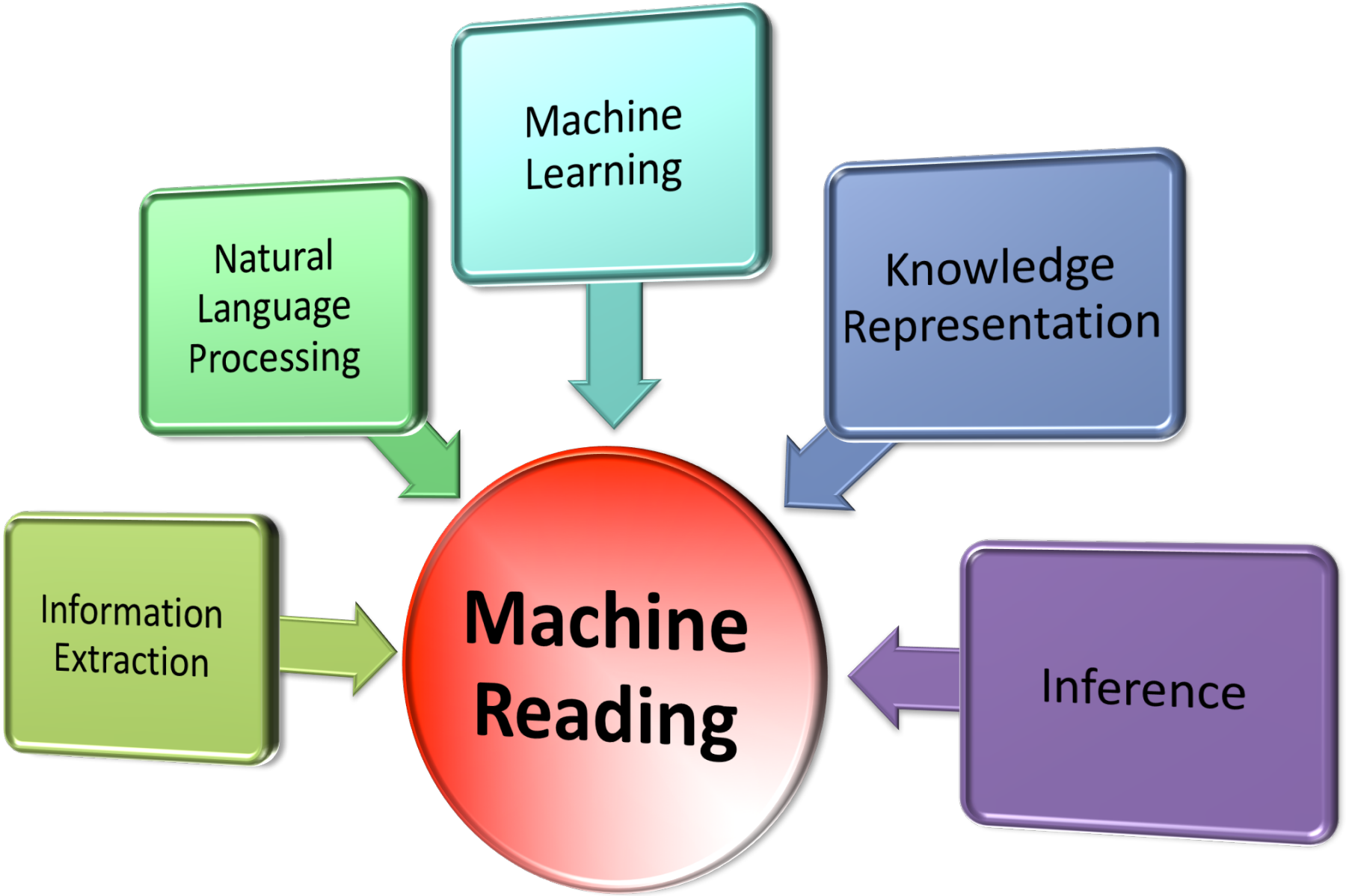
Machine Reading



Machine Reading



Machine Reading



Machine Reading

- One important *(initial)* approach to machine reading is to extract facts from text and store them in a structured form.
- Facts can be seen as entities and their relations
- Ontology is one of the most common representation for the extracted facts

Machine Reading

Some possible subtasks

- Named Entity Resolution/Recognition
- Relation Extraction
- Co-reference and Polysemy Resolution
- Relation Discovery
- Inference
- Knowledge Base
- Document/Sentence Understanding (Micro-Reading)

Machine Reading

- Named Entity Resolution/Recognition
- Relation Extraction
- Co-reference and Polysemy Resolution
- Relation Discovery
- Inference
- Knowledge Base
- Document/Sentence Understanding (Micro-Reading)

Machine Reading

- Named Entity Resolution/Recognition
  - **Semi-structured data**

    The "Low-Hanging Fruit"
    - Wikipedia infoboxes & categories
    - HMTL lists & tables, etc.

  - **Free text**
    - Hearst-patterns; clustering by verbal phrases
    - Natural-language processing
    - Advanced patterns & iterative bootstrapping

Named Entity Recognition

- Named Entity Recognition [Nadeau & Sekine, 2007]
  - term “Named Entity” coined for the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996).
  - important sub-tasks of IE called “Named Entity Recognition and Classification (NERC)”.

Named Entity Recognition

[Nadeau & Sekine, 2007]

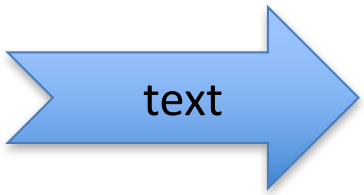
- recognize information units like names, including **person**, **organization** and **location** names, and numeric expressions including **time**, **date**, **money** and **percent expressions**.
- In Machine Reading, many other entities: **product**, **kitchen item**, **sport**, etc.

Named Entity Resolution

- Named Entity Resolution [Theobald & Weikum, 2012]
  - Which individual entities belong to which classes?
    - *instanceOf (Surajit Chaudhuri, computer scientists),*
    - *instanceOf (BarbaraLiskov, computer scientists),*
    - *instanceOf (Barbara Liskov, female humans), …*

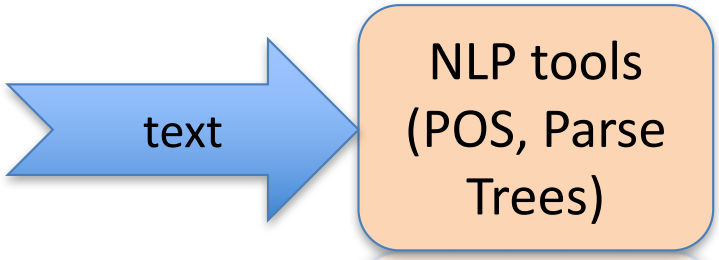
Named Entity Recognition

- Named Entities Recognition as a machine learning task.
  - Supervised Learning



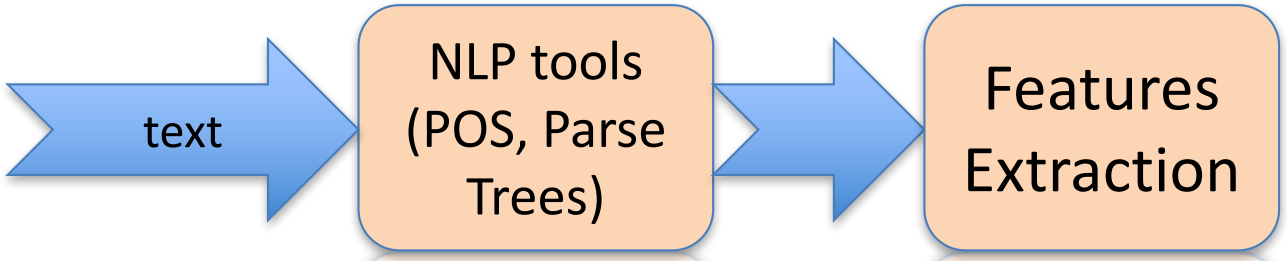
Named Entity Recognition

- Named Entities Recognition as a machine learning task.
  - Supervised Learning



Named Entity Recognition

- Named Entities Recognition as a machine learning task.
  - Supervised Learning

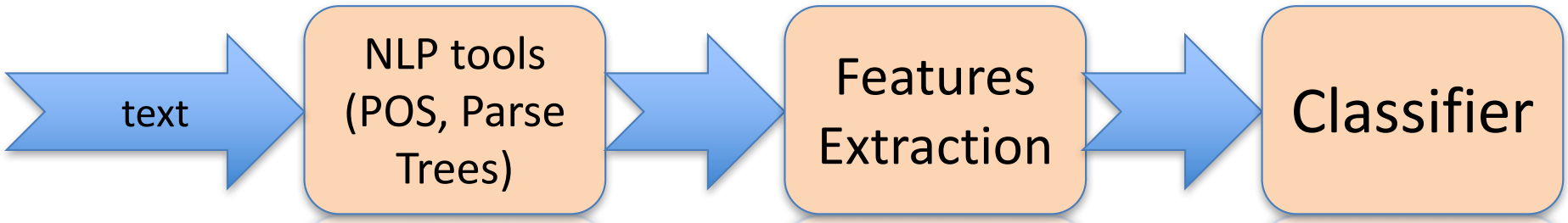


Named Entity Recognition

- Named Entity Recognition as a Machine Learning task.
  - Supervised Learning
  - Possible features [Ratinov & Roth, 2009], [Khambhatla, 2004], [Zhou et. al. 2005]
    - Words “around” and including entities
    - POS (Part-Of-Speech)
    - Prefixes and suffixes
    - Capitalization
    - Number of words
    - Number of characters
    - First word, last word
    - gazetteer matches

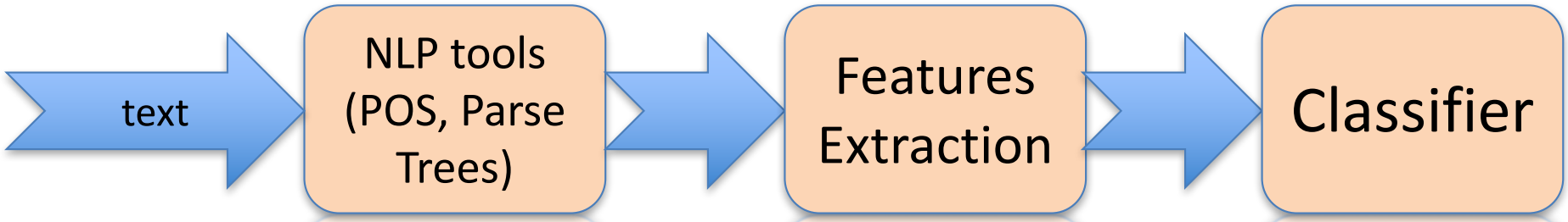
Named Entity Recognition

- Named Entities Recognition as a machine learning task.
  - Supervised Learning



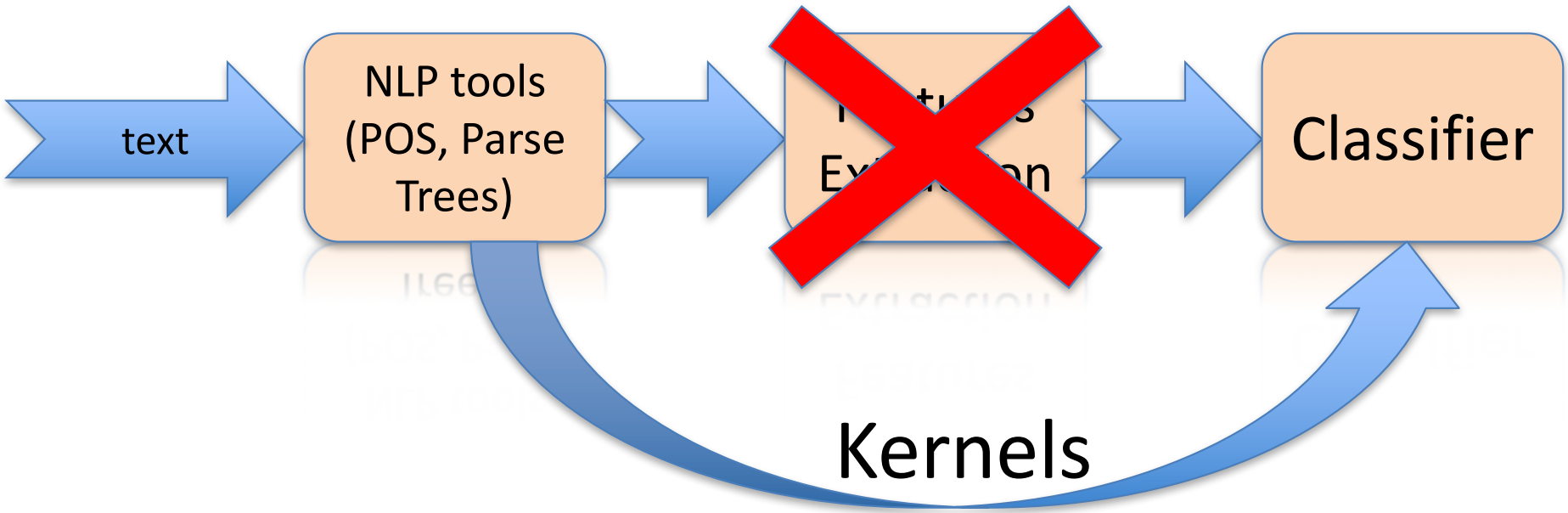
Named Entity Recognition

- Supervised Learning



Named Entity Recognition

- Supervised Learning



Named Entity Recognition

[Bach & Badaskar, 2007]

- Supervised Learning using Kernels
  - A Kernel defines similarity implicitly in a higher dimensional space
  - Can be based on Strings, Word Sequences, Parse Trees, etc.
    - For strings similarity  number of common substrings (or subsequences)
    - Recommended reading on string kernels [Lodhi et. al., 2002]

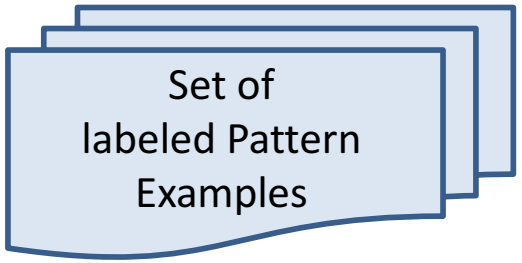
Named Entity Recognition

[Bach & Badaskar, 2007]

	Feature set Definition	Computational Complexity
Feature based Methods	Required to define a feature-set to be extracted after textual analysis. Good features arrived at by experimentation	Relatively lower
Kernel Methods	No need to define a feature-set. Similarity computed over a much larger feature space implicitly.	Relatively higher

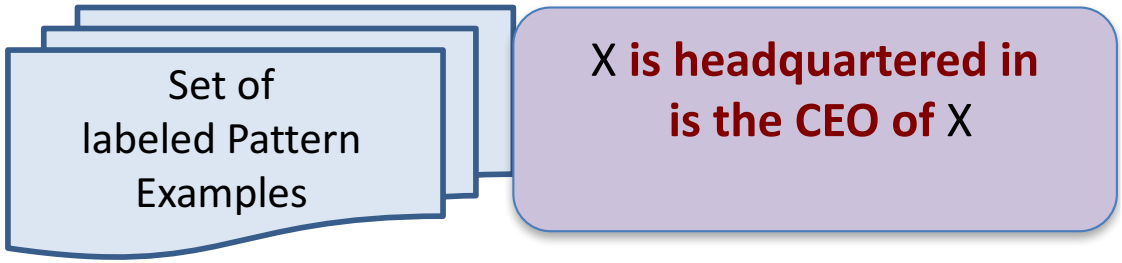
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



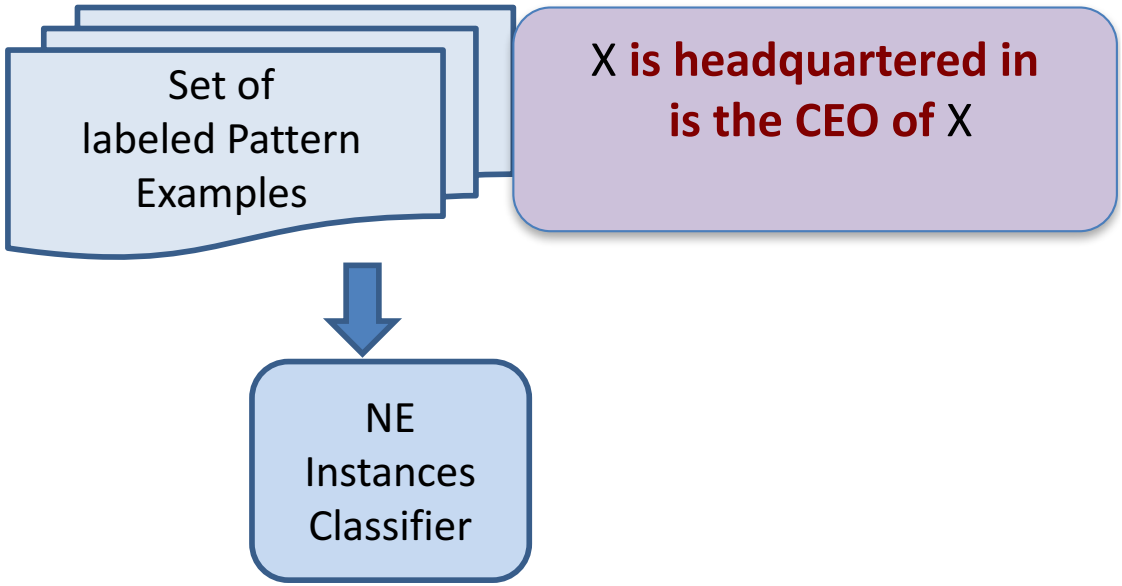
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



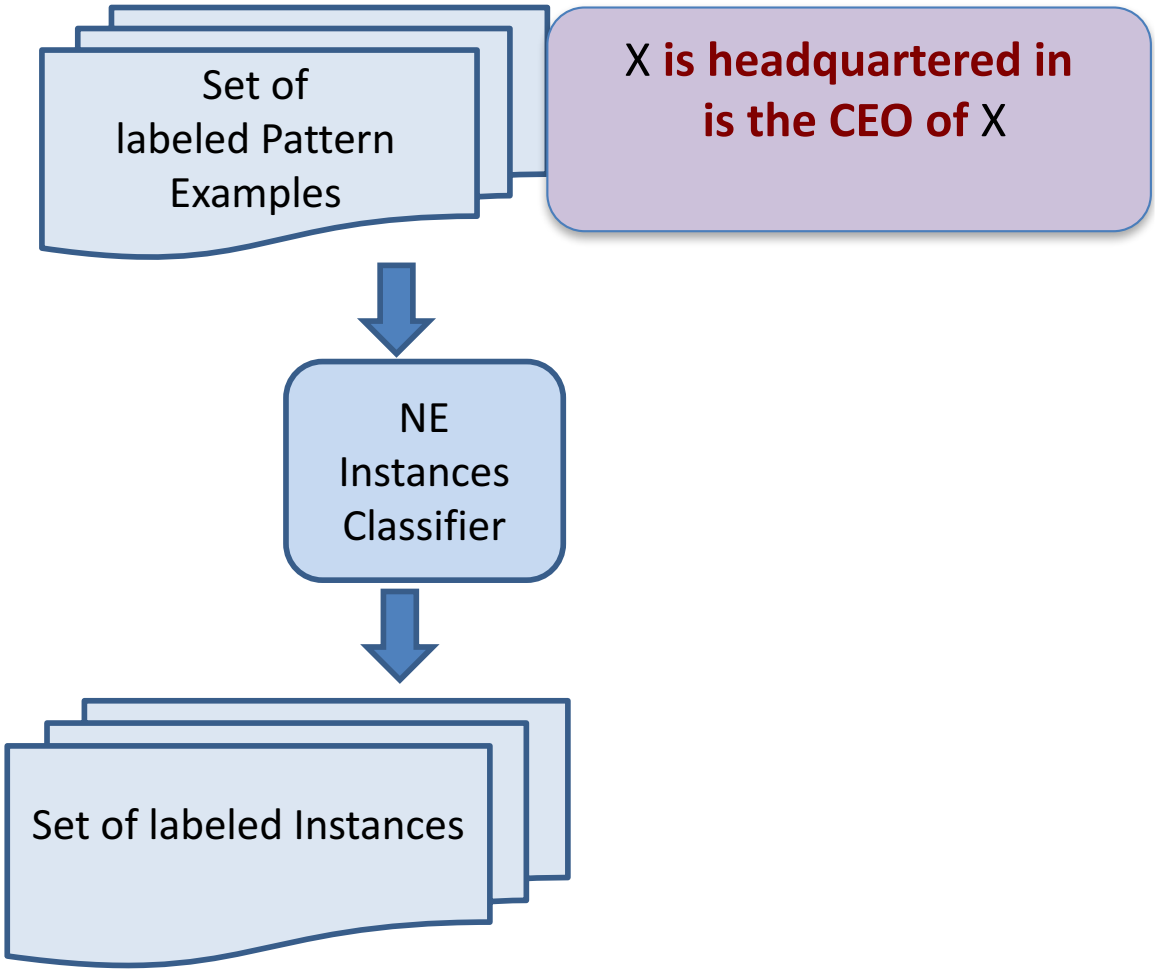
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



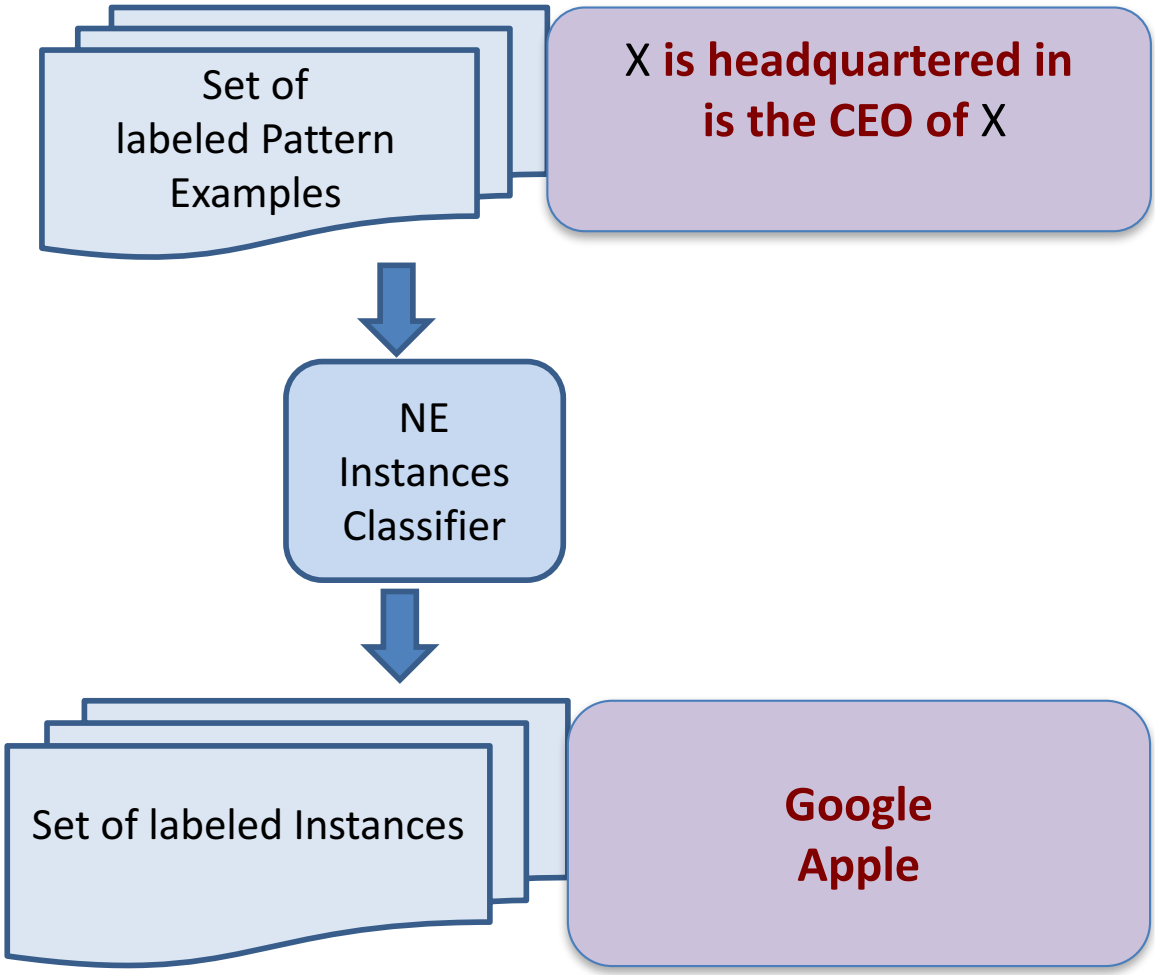
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



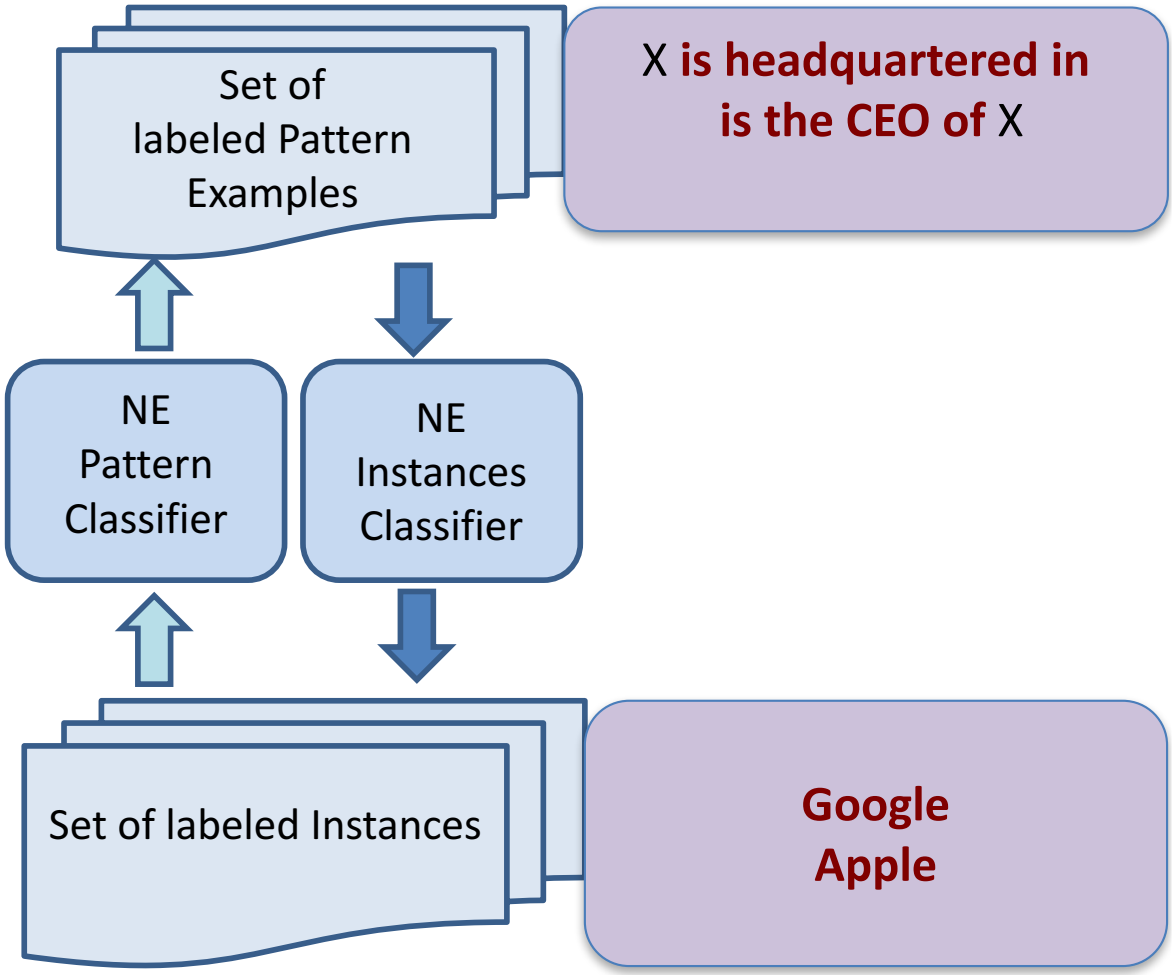
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



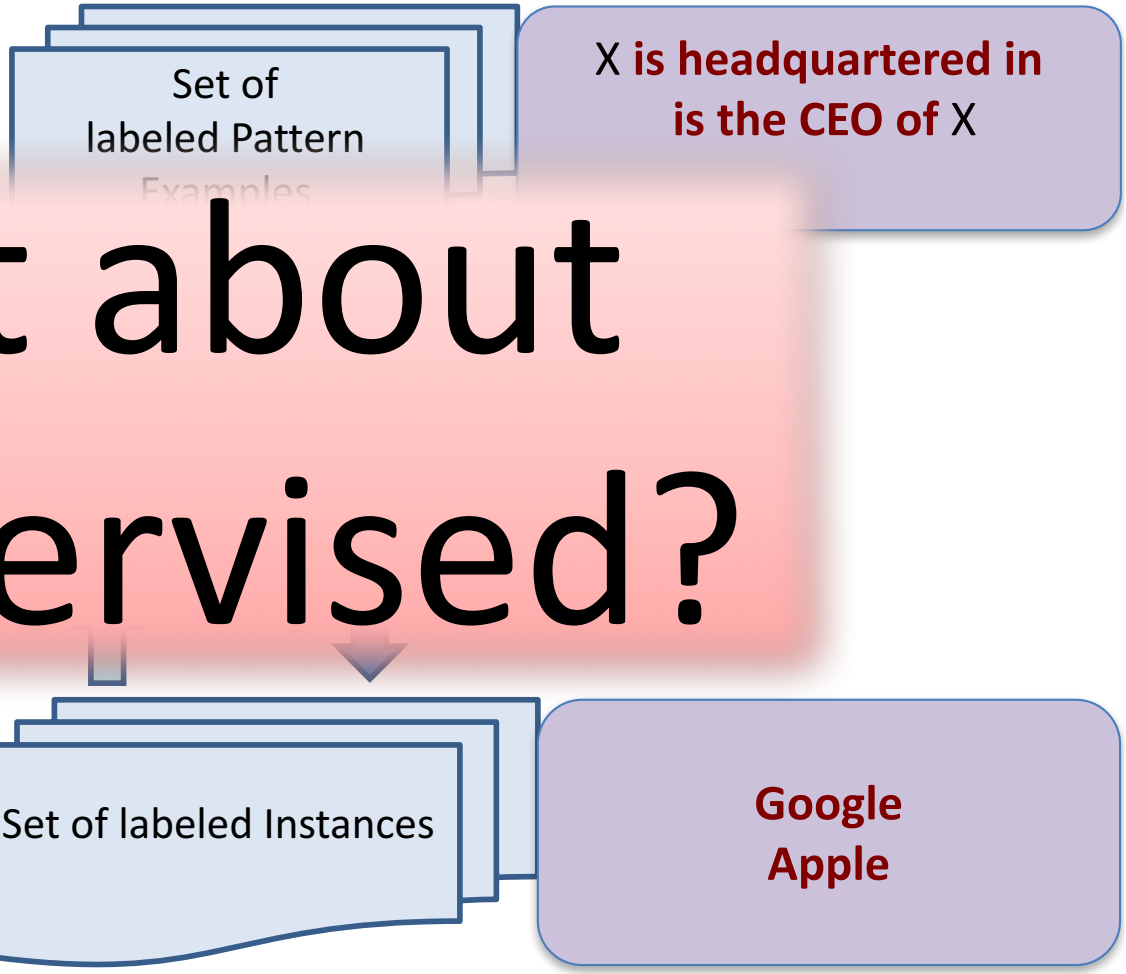
Named Entity Recognition

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



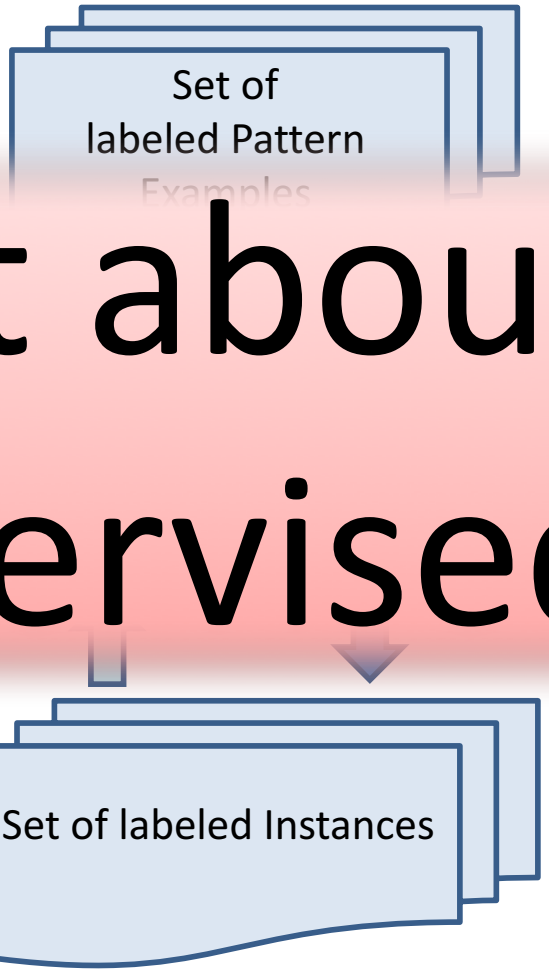
Named Entity Recognition

- Semi-supervised
  - Bo
    ge
    nu
    an



Named Entity Recognition

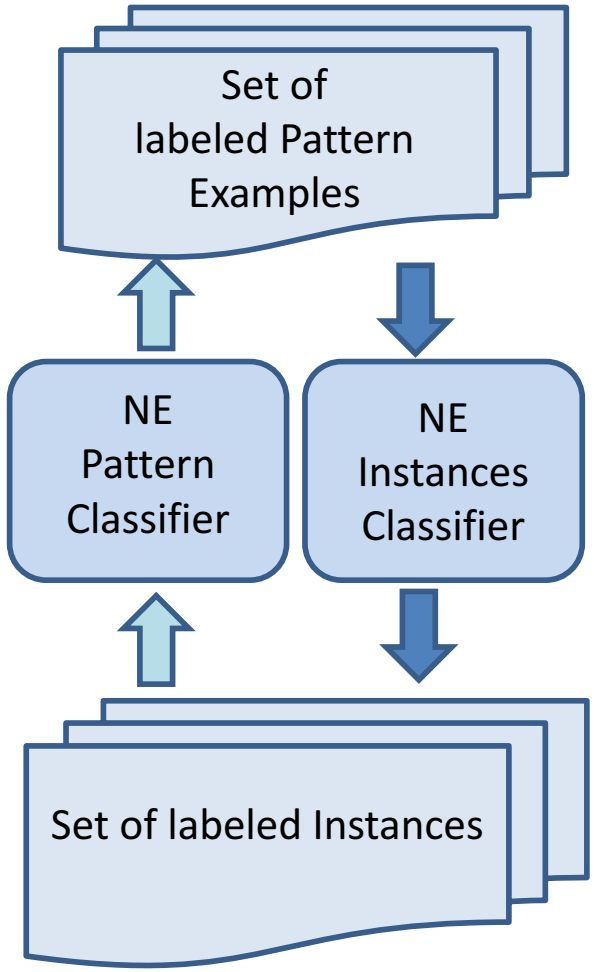
- Semi-supervised
  - Bo
    ge
    nu
    and NE instances.



What about unsupervised?

Named Entity Recognition

- Unsupervised Approaches
  - Bootstrap can generate a large number of patterns and NE instances.



Named Entity Recognition

- [Ratinov & Roth, 2009]

	System	Resources Used	F_1
+	LBJ-NER	Wikipedia, Nonlocal Features, Word-class Model	90.80
-	(Suzuki and Isozaki, 2008)	Semi-supervised on 1G-word unlabeled data	89.92
-	(Ando and Zhang, 2005)	Semi-supervised on 27M-word unlabeled data	89.31
-	(Kazama and Torisawa, 2007a)	Wikipedia	88.02
-	(Krishnan and Manning, 2006)	Non-local Features	87.24
-	(Kazama and Torisawa, 2007b)	Non-local Features	87.17
+	(Finkel et al., 2005)	Non-local Features	86.86

Named Entity Recognition

- [Pennington & Socher & Manning, 2014]

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Polysemy Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Machine Reading

- Relation Extraction
  - **Semi-structured data**

    The "Low-Hanging Fruit"
    - Wikipedia infoboxes & categories
    - HMTL lists & tables, etc.

  - **Free text**
    - Hearst-patterns; clustering by verbal phrases
    - Natural-language processing
    - Advanced patterns & iterative bootstrapping

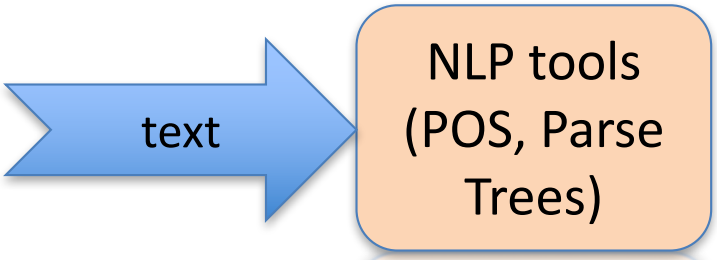
Machine Reading

- Relation Extraction [Theobald & Weikum, 2012]
  - Which instances (pairs of individual entities) are there for given binary relations with specific type signatures?
    - hasAdvisor (JimGray, MikeHarrison)
    - hasAdvisor (HectorGarcia-Molina, Gio Wiederhold)
    - hasAdvisor (Susan Davidson, Hector Garcia-Molina)
    - graduatedAt (JimGray, Berkeley)
    - graduatedAt (HectorGarcia-Molina, Stanford)
    - hasWonPrize (JimGray, TuringAward)
    - bornOn (JohnLennon, 9Oct1940)
    - diedOn (JohnLennon, 8Dec1980)
    - marriedTo (JohnLennon, YokoOno)

Relation Extraction

[Bach & Badaskar, 2007]

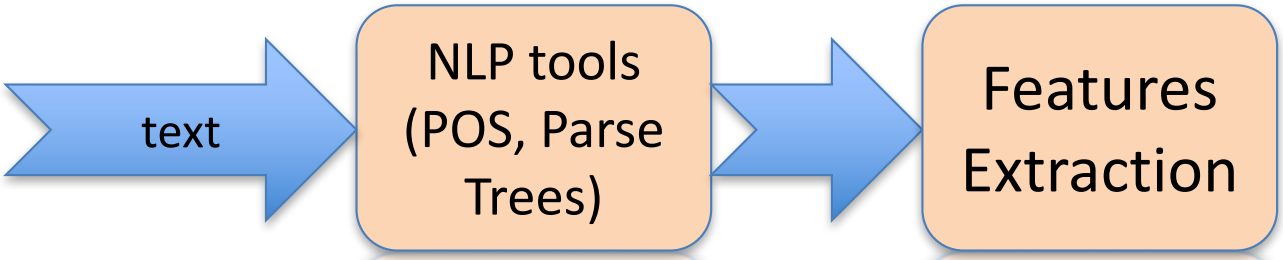
- Extracting semantic relations between entities in text
- Relation extraction as a Machine Learning task.
  - Supervised Learning



Relation Extraction

[Bach & Badaskar, 2007]

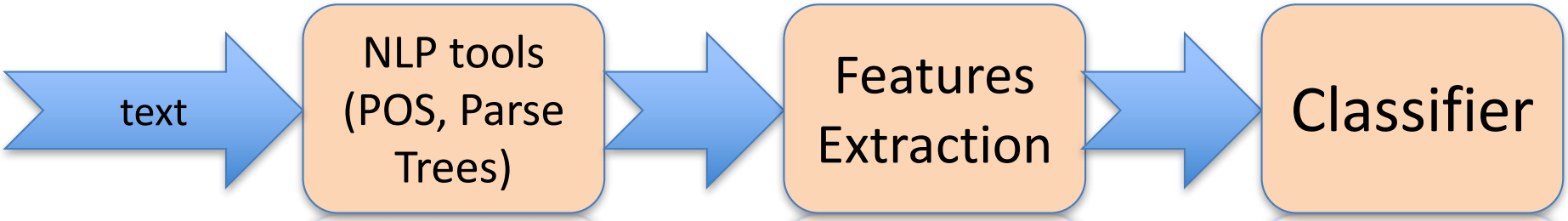
- Extracting semantic relations between entities in text
- Relation extraction as a Machine Learning task.
  - Supervised Learning



Relation Extraction

[Bach & Badaskar, 2007]

- Extracting semantic relations between entities in text
- Relation extraction as a Machine Learning task.
  - Supervised Learning



Relation Extraction

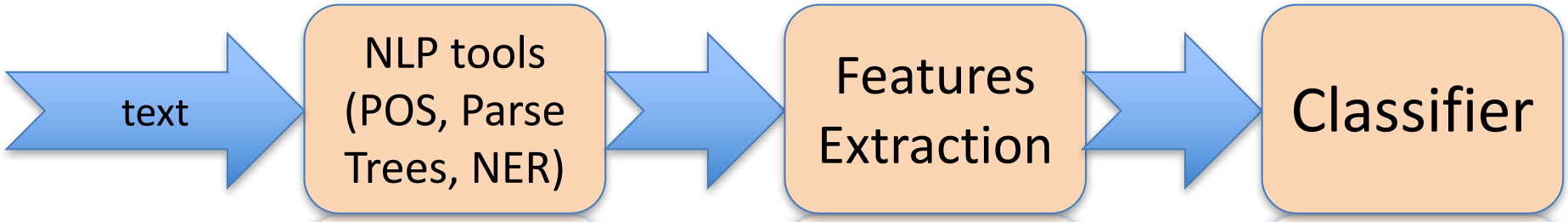
[Bach & Badaskar, 2007]

- Relation extraction as a Machine Learning task.
  - Supervised Learning
  - Possible features [Khambhatla, 2004], [Zhou et. al. 2005]
    - Words between and including entities
    - Types of entities (person, location, etc)
    - Number of entities between the two entities, whether both entities belong to same chunk
    - # words separating the two entities
    - Path between the two entities in a parse tree

Relation Extraction

[Bach & Badaskar, 2007]

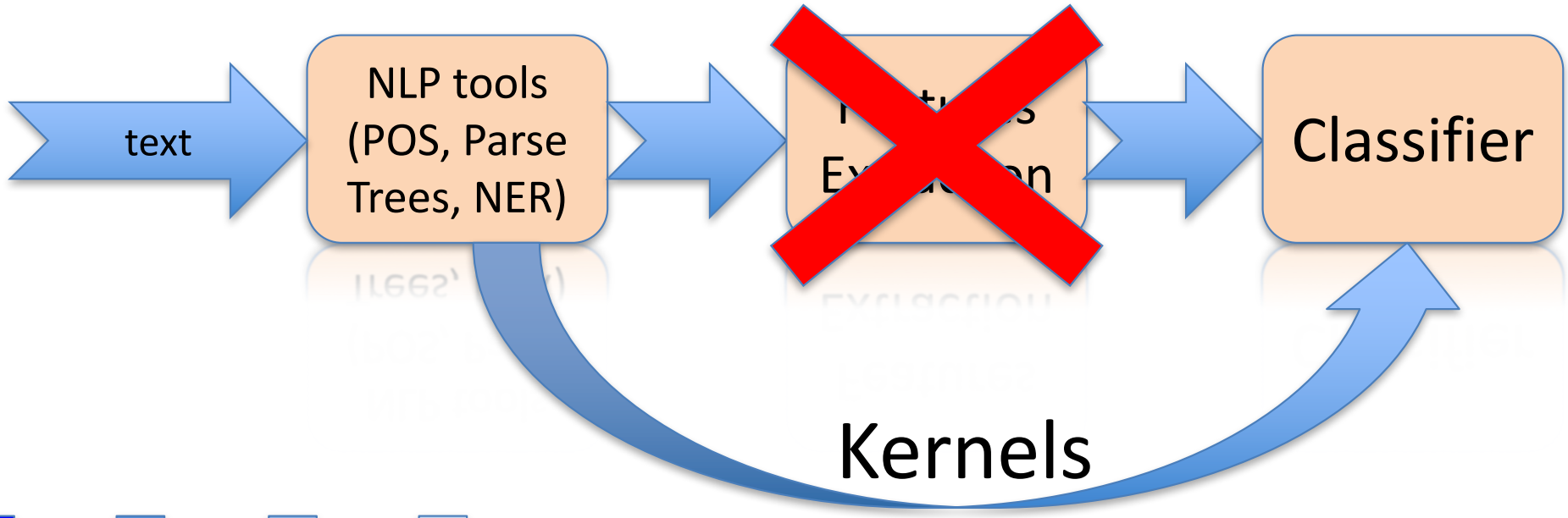
- Extracting semantic relations between entities in text
- Relation extraction as a classification task.
  - Supervised Learning



Relation Extraction

[Bach & Badaskar, 2007]

- Extracting semantic relations between entities in text
- Relation extraction as a classification task.
  - Supervised Learning



Relation Extraction

[Bach & Badaskar, 2007]

- Supervised Learning using Kernels
  - A Kernel defines similarity implicitly in a higher dimensional space
  - Can be based on Strings, Word Sequences, Parse Trees, etc.
    - For strings, similarity  number of common substrings (or subsequences)
    - Recommended reading on string kernels [Lodhi et. al., 2002]

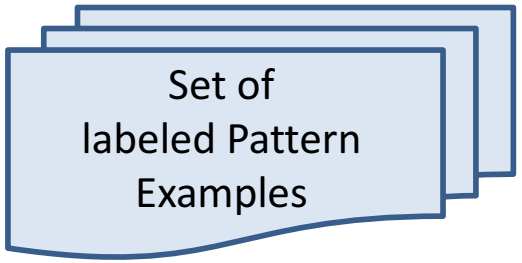
Relation Extraction

[Bach & Badaskar, 2007]

	Feature set Definition	Computational Complexity
Feature based Methods	Required to define a feature-set to be extracted after textual analysis. Good features arrived at by experimentation	Relatively lower
Kernel Methods	No need to define a feature-set. Similarity computed over a much larger feature space implicitly.	Relatively higher

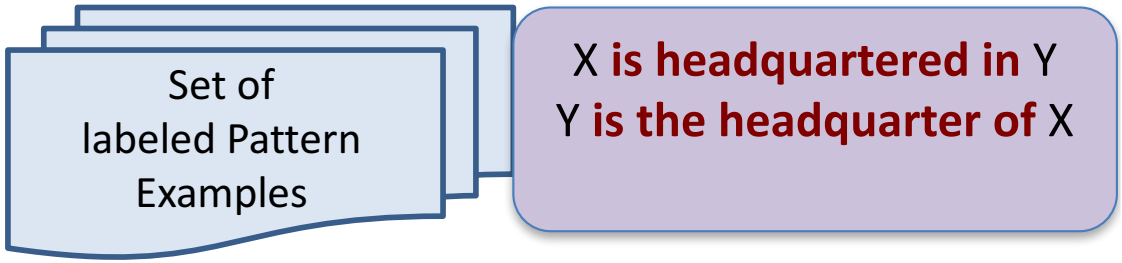
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



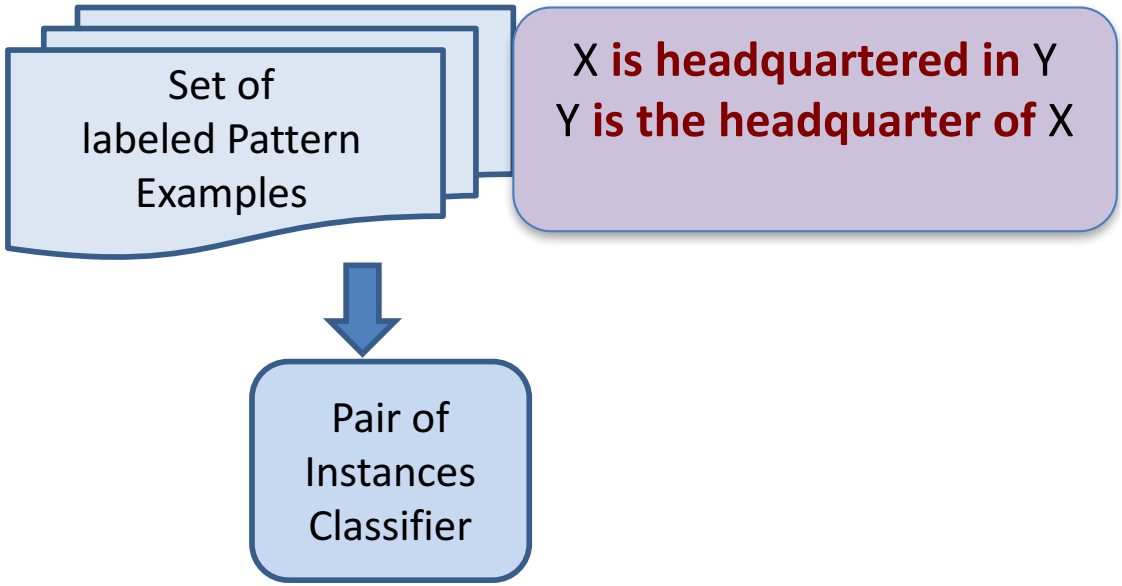
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



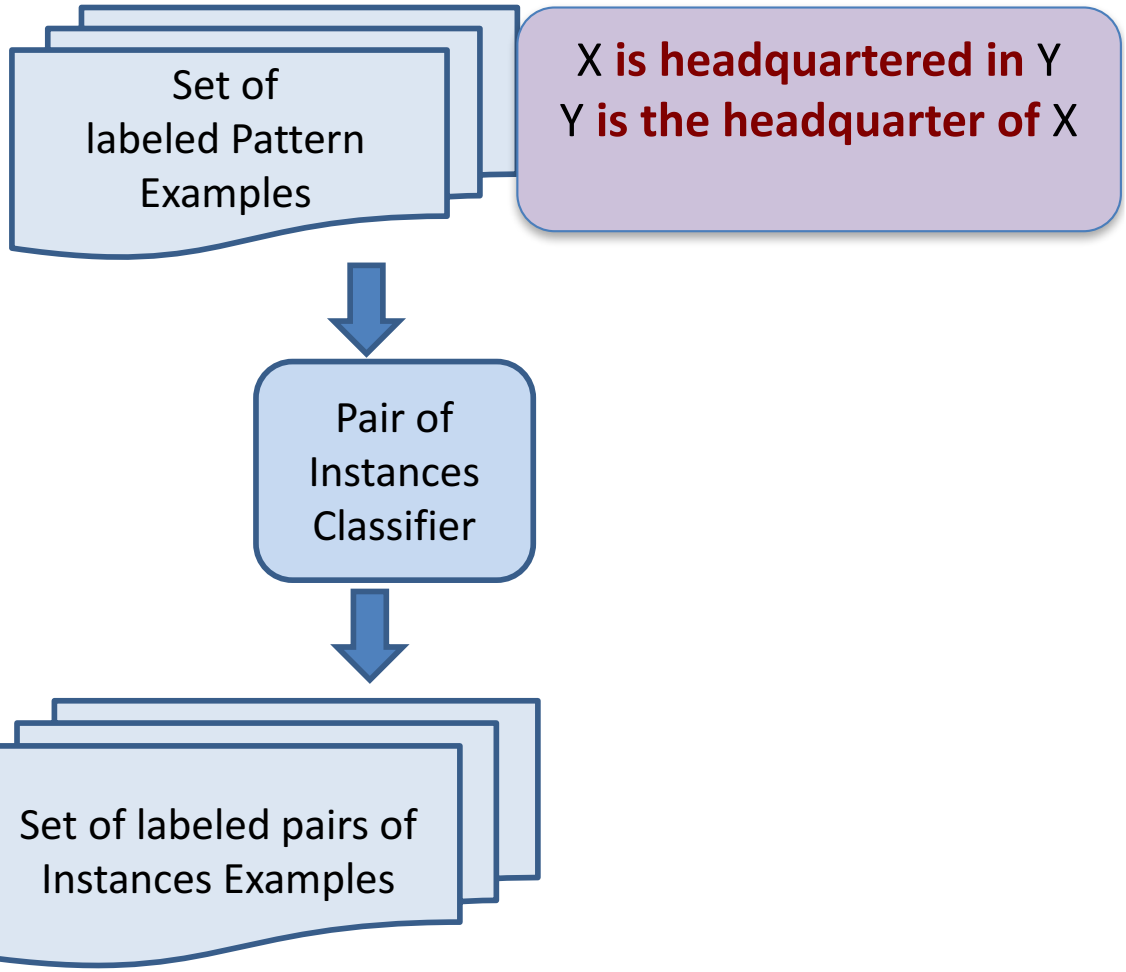
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



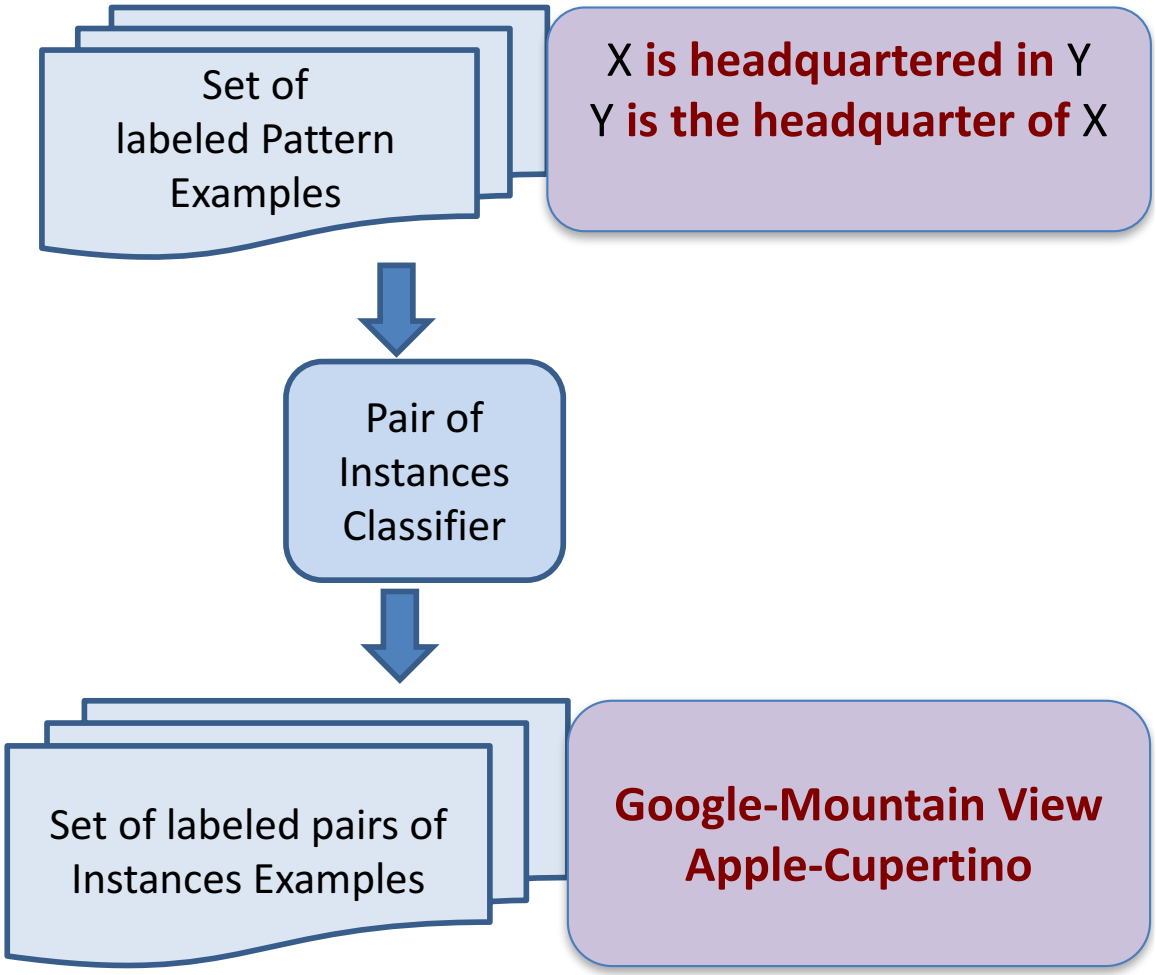
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



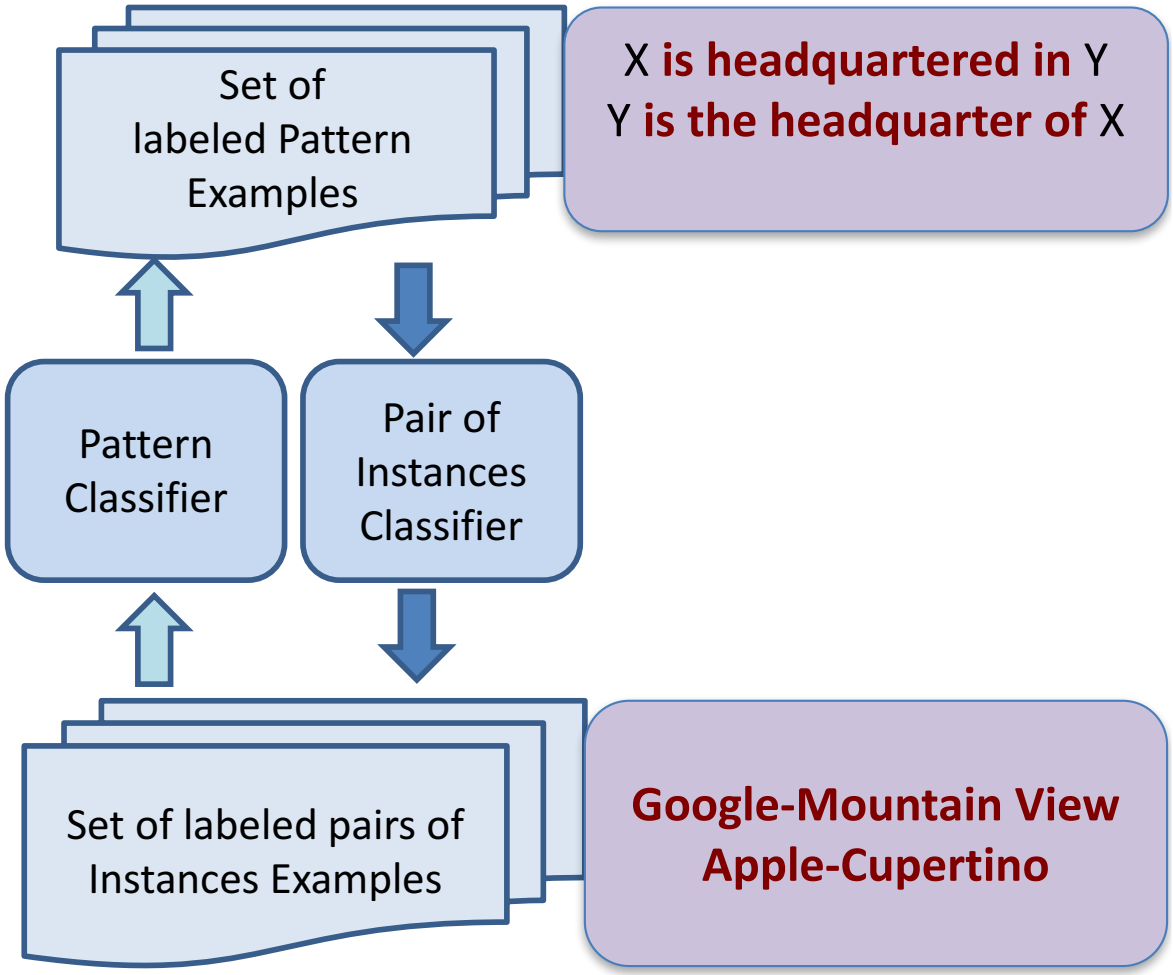
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



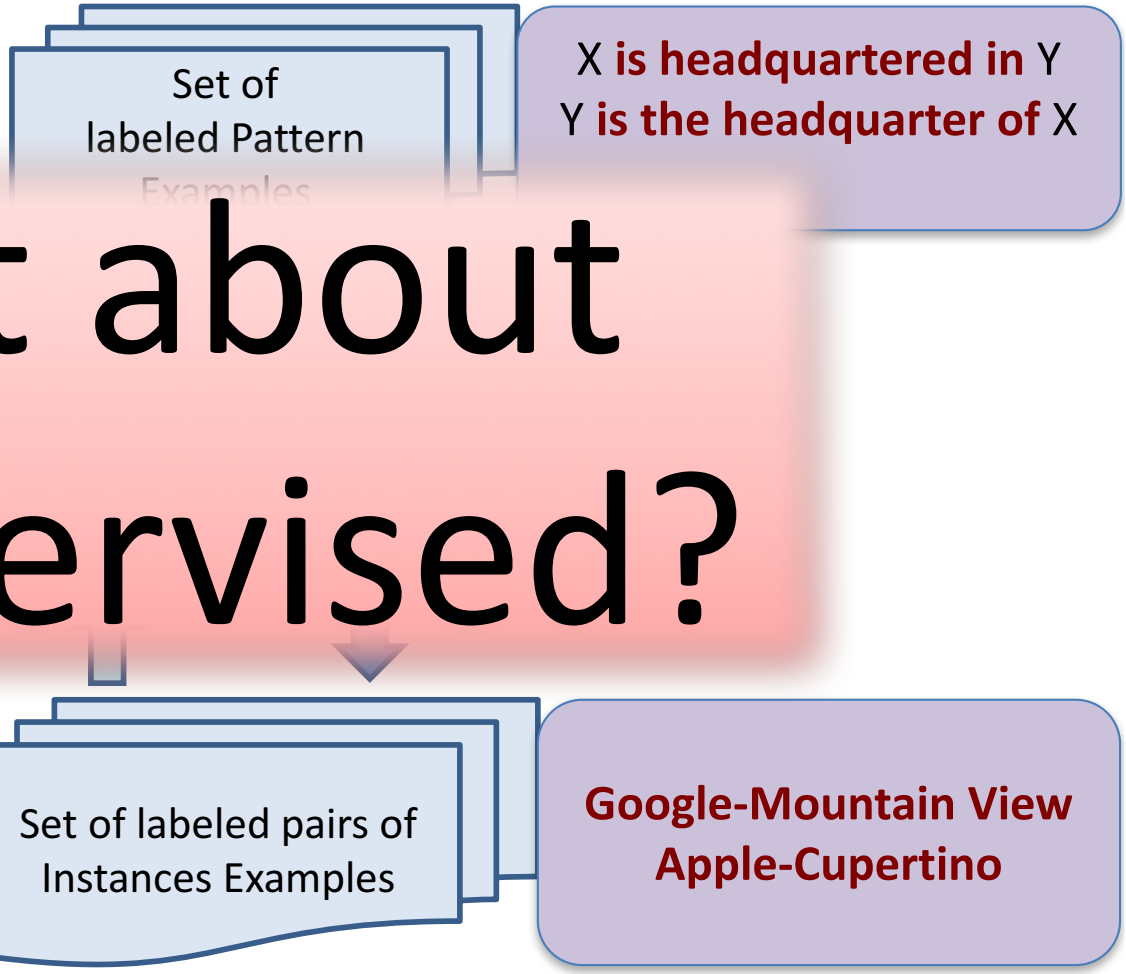
Relation Extraction

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



Relation Extraction

- Semi-supervised Approaches
  - Bo… ge… nu… an… in…

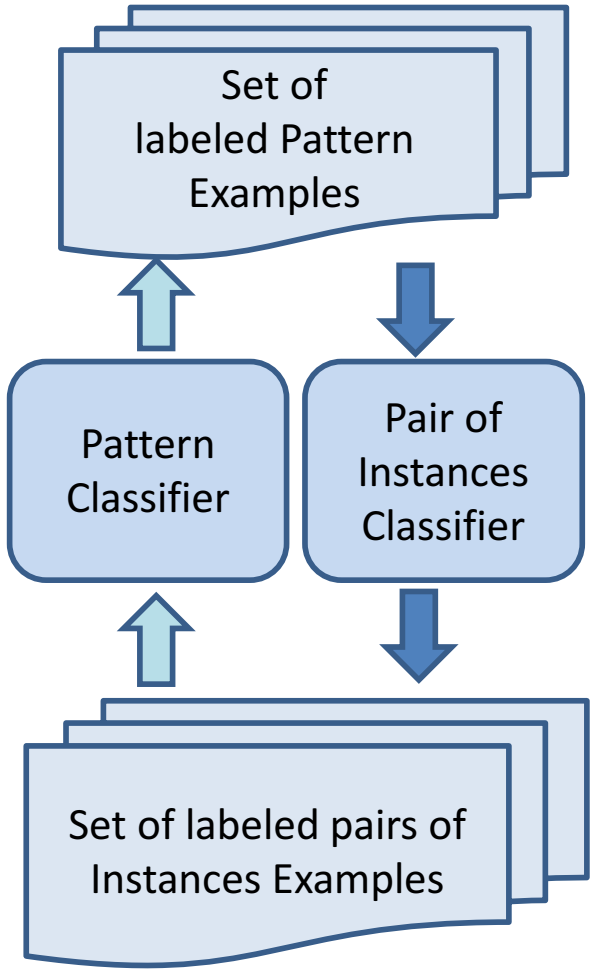


Relation Extraction

- Semi-supervised Approaches
  - Bo… ge… nu… an… in…

Relation Extraction

- Unsupervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



Relation Extraction

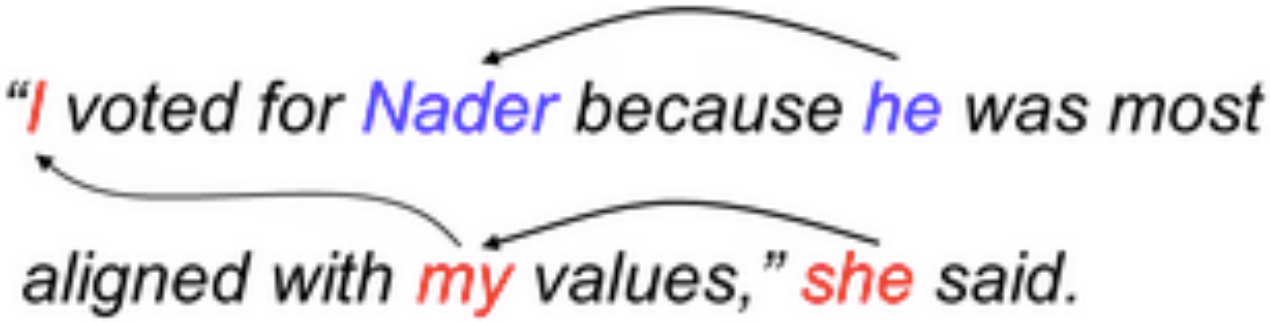
- Supervised learning [Bunescu & Mooney, 2005]
- Distant and Partial Supervised [Angeli & Tibshirani & Wu & Manning, 2014]

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Polysemy Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Co-Reference and Polysemy Resolution

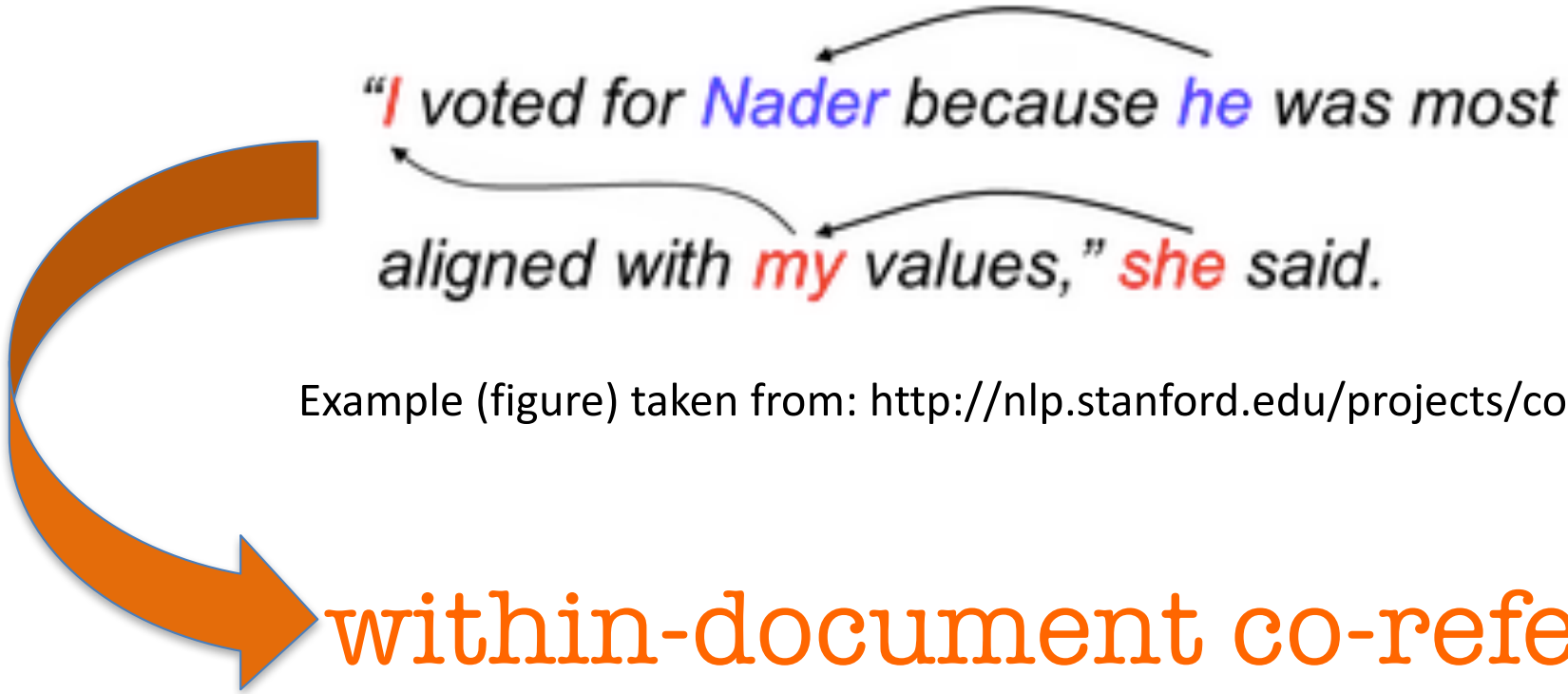
- Co-reference: expressions that refer to the same entity



Example (figure) taken from: http://nlp.stanford.edu/projects/coref.shtml

Co-Reference and Polysemy Resolution

- Co-reference: expressions that refer to the same entity



Example (figure) taken from: http://nlp.stanford.edu/projects/coref.shtml

Co-Reference and Polysemy Resolution

- Co-reference: expressions that refer to the same entity

Example (figu… …jects/coref.shtml

with… document co… eference

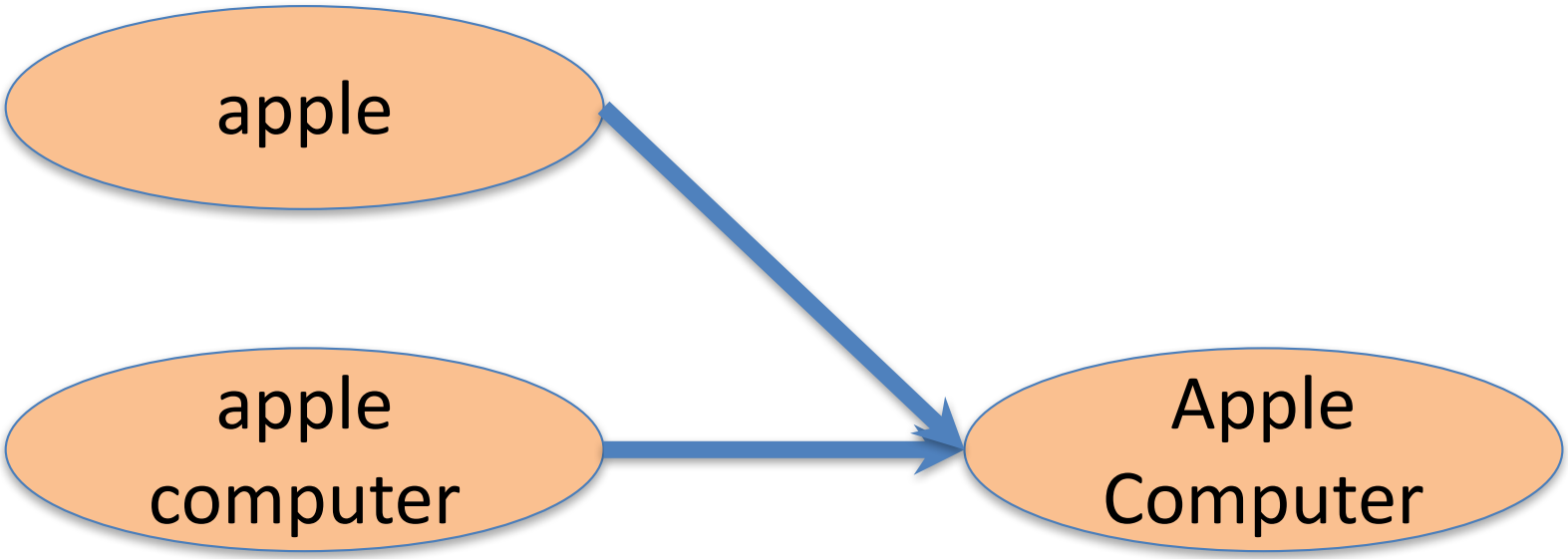
Co-Reference and Polysemy Resolution

- Co-reference: expressions that refer to the same entity

Example (figure) adapted from [Krishnamurthy & Mitchell, 2011]

Co-Reference and Polysemy Resolution

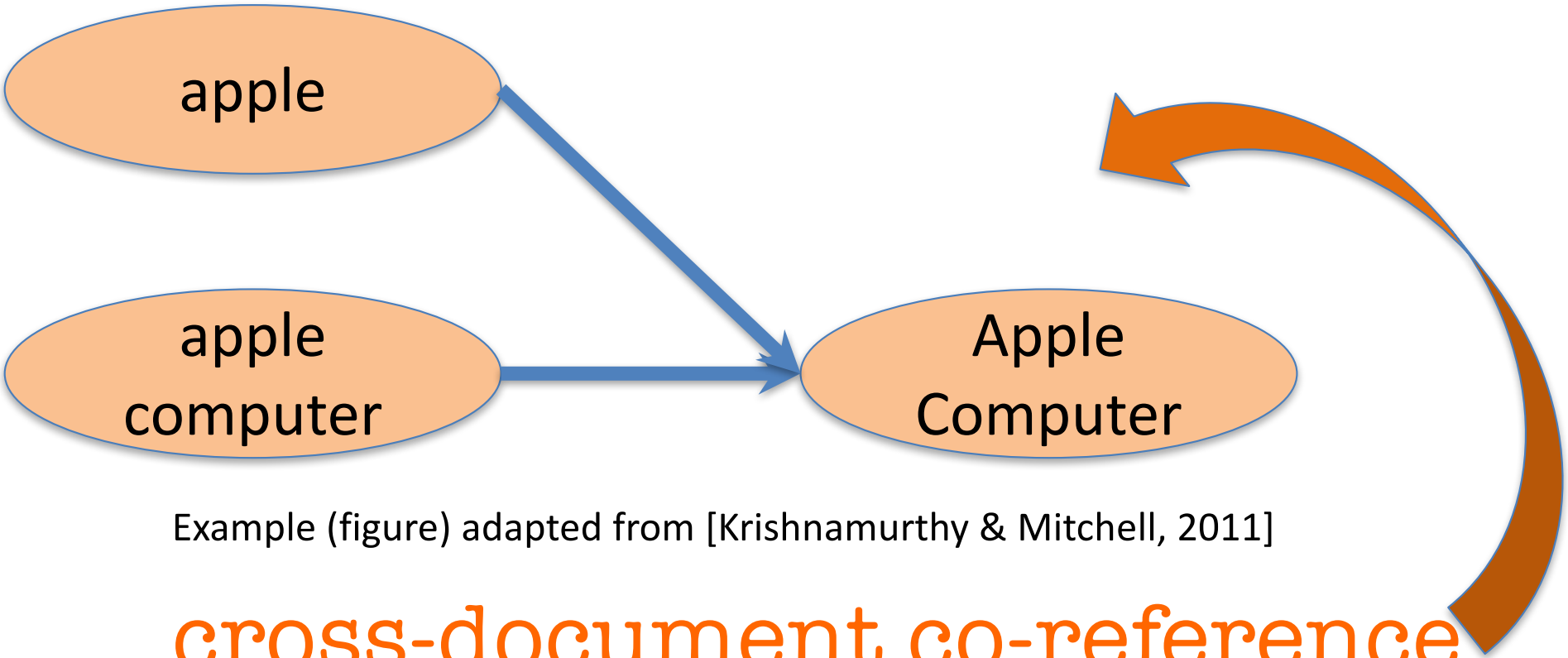
- Co-reference: expressions that refer to the same entity



Example (figure) adapted from [Krishnamurthy & Mitchell, 2011]

Co-Reference and Polysemy Resolution

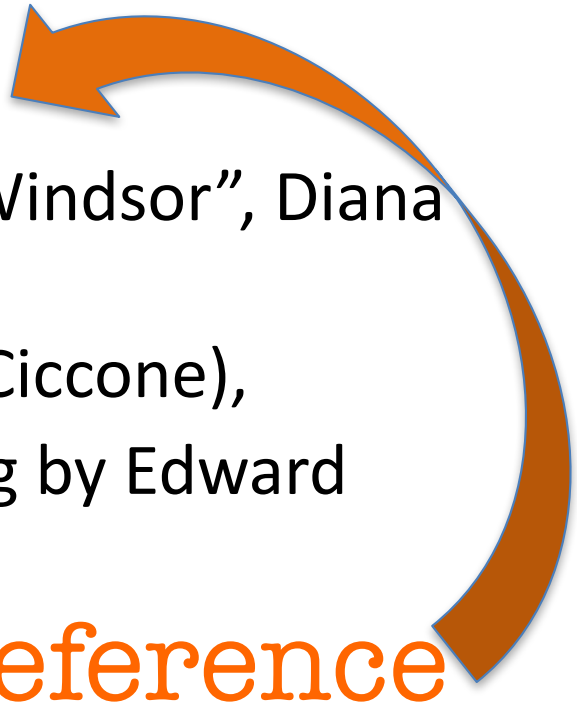
- Co-reference: expressions that refer to the same entity



Example (figure) adapted from [Krishnamurthy & Mitchell, 2011]

Co-Reference and Polysemy Resolution

- Co-reference: expressions that refer to the same entity
- Which names denote which entities? [Theobald & Weikum, 2012]
  - means ("Lady Di", Diana Spencer),
  - means ("Diana Frances Mountbatten-Windsor", Diana Spencer), …
  - means ("Madonna", Madonna Louise Ciccone),
  - means ("Madonna", Madonna(painting by Edward Munch)), …



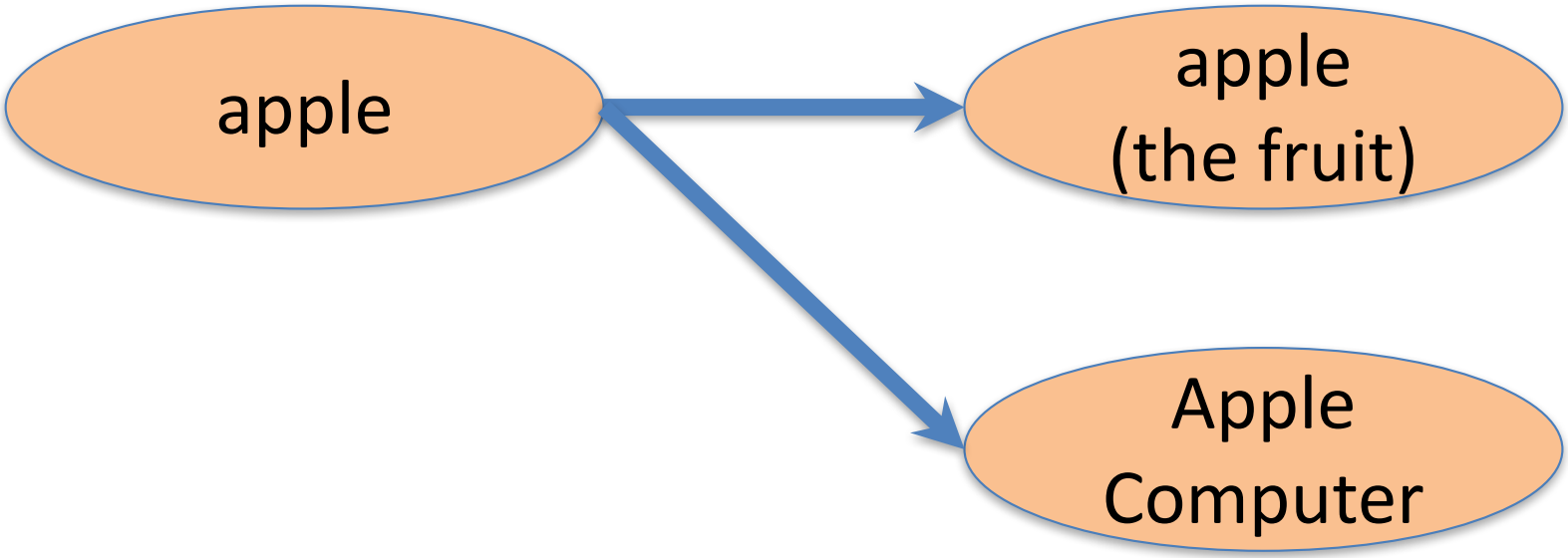
cross-document co-reference

Co-Reference and Polysemy Resolution

- Polysemy: is the capacity for a sign (such as a word, phrase, or symbol) to have multiple meanings [Wikipedia]

Co-Reference and Polysemy Resolution

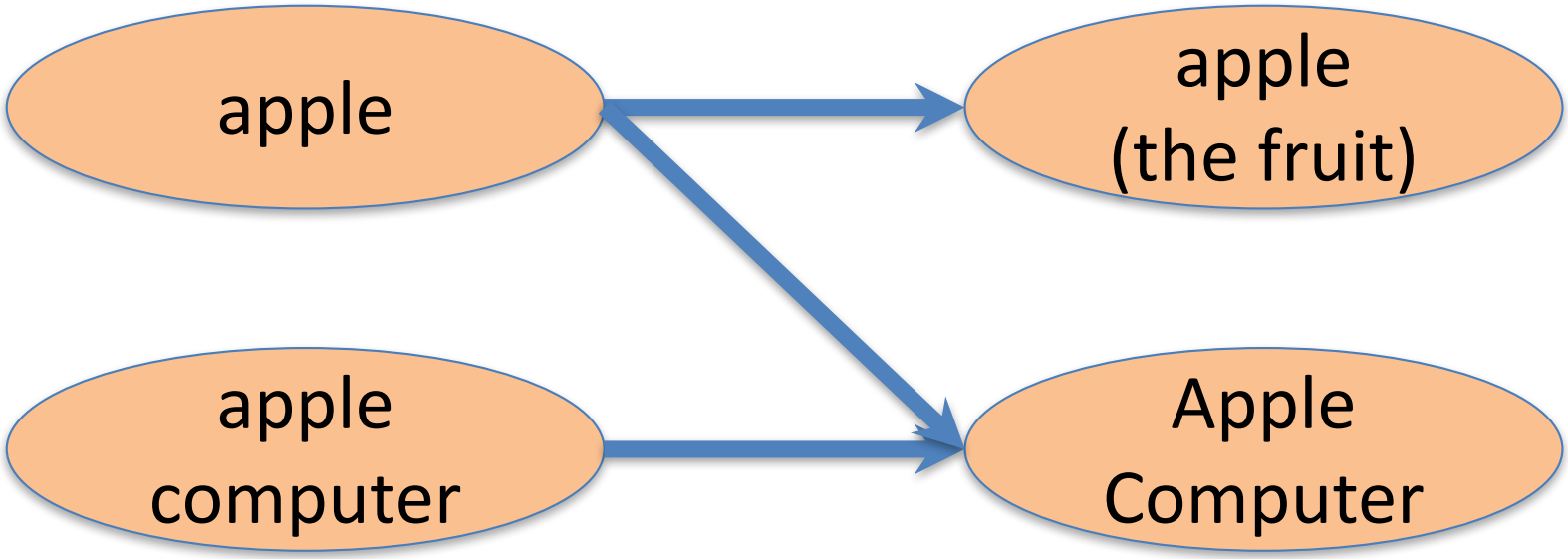
- Polysemy: is the capacity for a sign (such as a word, phrase, or symbol) to have multiple meanings [Wikipedia]



Example (figure) adapted from [Krishnamurthy & Mitchell, 2011]

Co-Reference and Polysemy Resolution

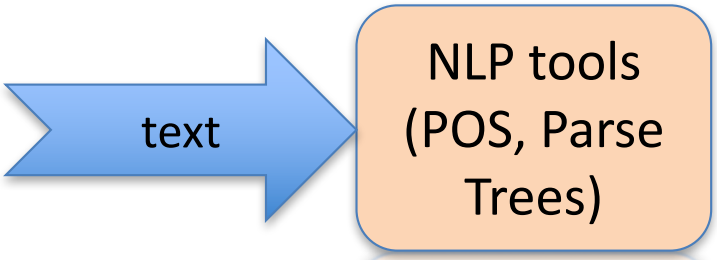
- Co-Reference and Polysemy



Example (figure) adapted from [Krishnamurthy & Mitchell, 2011]

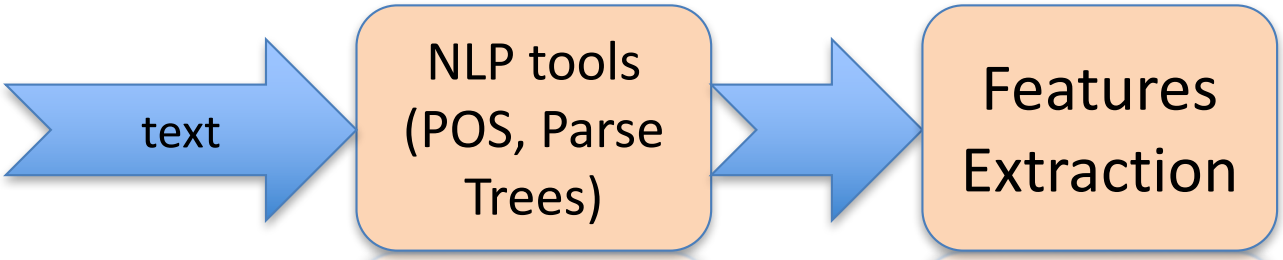
Co-Reference and Polysemy Resolution

- Co-reference and Polysemy:
  - Supervised Learning



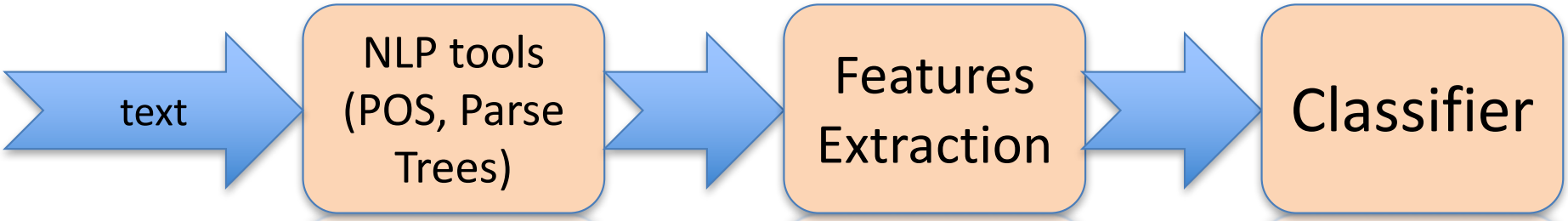
Co-Reference and Polysemy Resolution

- Co-reference and Polysemy:
  - Supervised Learning



Co-Reference and Polysemy Resolution

- Co-reference and Polysemy:
  - Supervised Learning



Co-Reference and Polysemy Resolution

- Co-Reference Resolution.
  - Supervised Learning
  - Possible features [Bengtson & Roth, 2008]

Category	Feature	Source
Mention Types	Mention Type Pair	Annotation and tokens
String Relations	Head Match	Tokens
	Extent Match	Tokens
	Substring	Tokens
	Modifiers Match	Tokens
	Alias	Tokens and lists
Semantic	Gender Match	WordNet and lists
	Number Match	WordNet and lists
	Synonyms	WordNet
	Antonyms	WordNet
	Hypernyms	WordNet
	Both Speak	Context
Relative Location	Apposition	Positions and context
	Relative Pronoun	Positions and tokens
	Distances	Positions
Learned	Anaphoricity	Learned
	Name Modifiers Predicted Match	Learned
Aligned Modifiers	Aligned Modifiers Relation	WordNet and lists
Memorization	Last Words	Tokens
Predicted Entity Types	Entity Types Match	Annotation and tokens
	Entity Type Pair	WordNet and tokens

Co-Reference and Polysemy Resolution

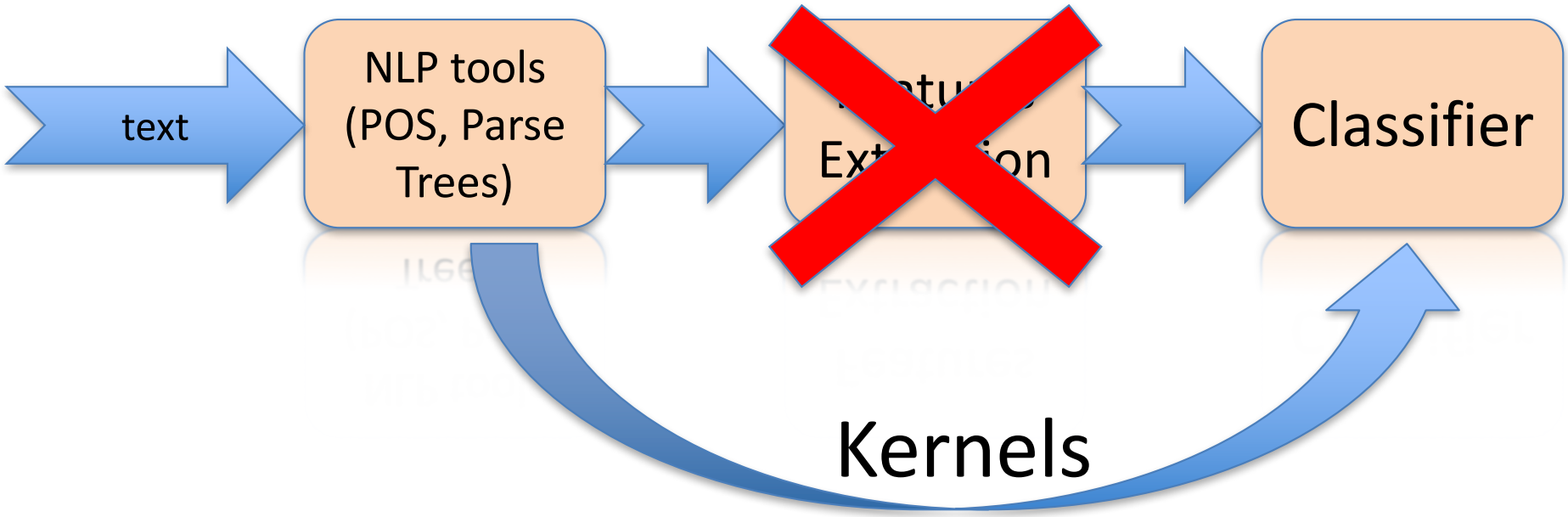
- Co-Reference Resolution.
  - Supervised Learning
  - Possible features [Bengtson & Roth, 2008]

Category	Feature	Source
Mention Types	Mention Type Pair	Annotation
String Relations	Head Match	Token
	Extent Match	T
	Substring	
	Modifiers Match	
	Alias	ns and lists
Semantic	Gender Match	WordNet and lists
	Number Ma	WordNet and lists
	Synony	WordNet
	An	WordNet
		WordNet
	k	Context
Relative Loca	sition	Positions and context
	elative Pronoun	Positions and tokens
	Distances	Positions
	Anaphoricity	Learned
	Name Modifiers Predicted Match	Learned
ed Modifiers	Aligned Modifiers Relation	WordNet and lists
Memorization	Last Words	Tokens
Predicted Entity Types	Entity Types Match	Annotation and tokens
	Entity Type Pair	WordNet and tokens

within-document co-reference

Co-Reference and Polysemy Resolution

- Co-reference and Polysemy:
  - Supervised Learning

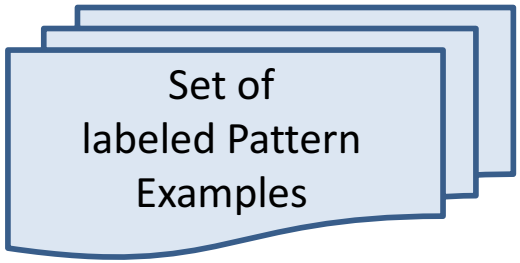


Co-Reference and Polysemy Resolution

- Supervised Learning using Kernels
  - A Kernel defines similarity implicitly in a higher dimensional space
  - Can be based on Strings, Word Sequences, Parse Trees, etc.
    - For strings similarity  number of common substrings (or subsequences)
    - Recommended reading on string kernels [Lodhi et. al., 2002]

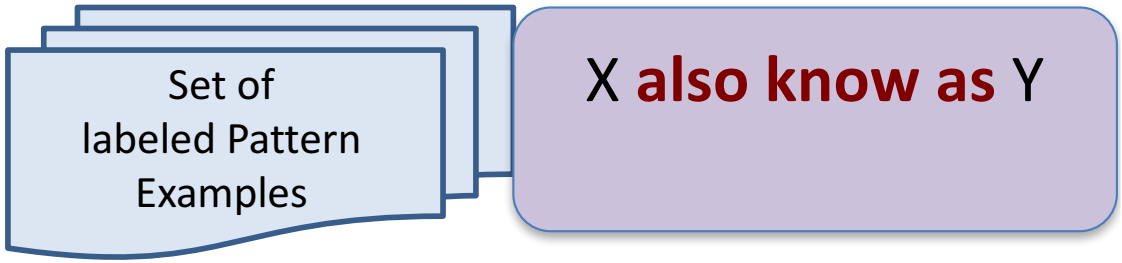
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



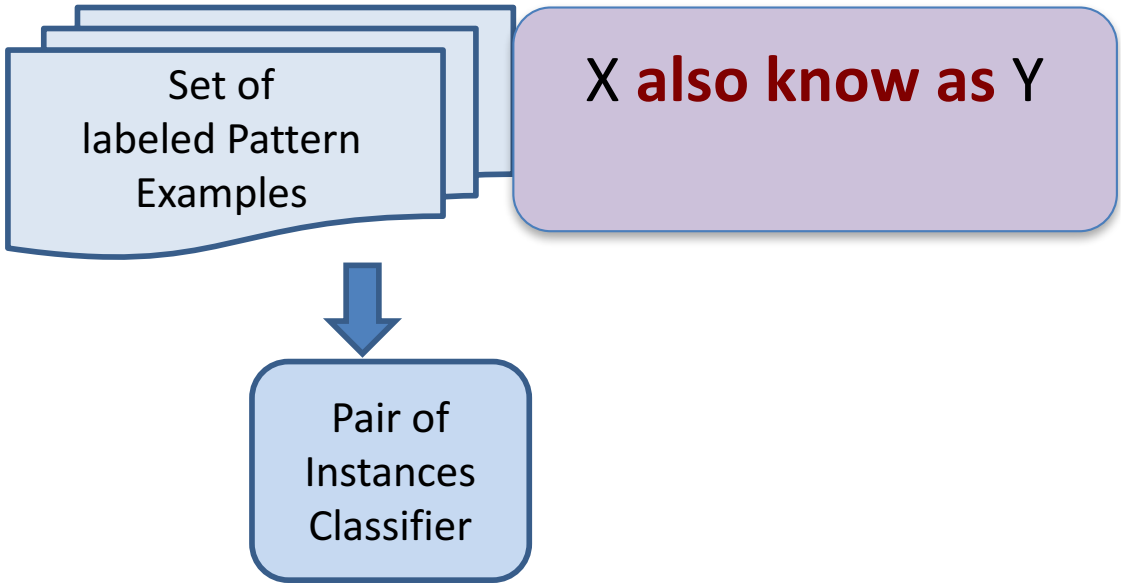
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



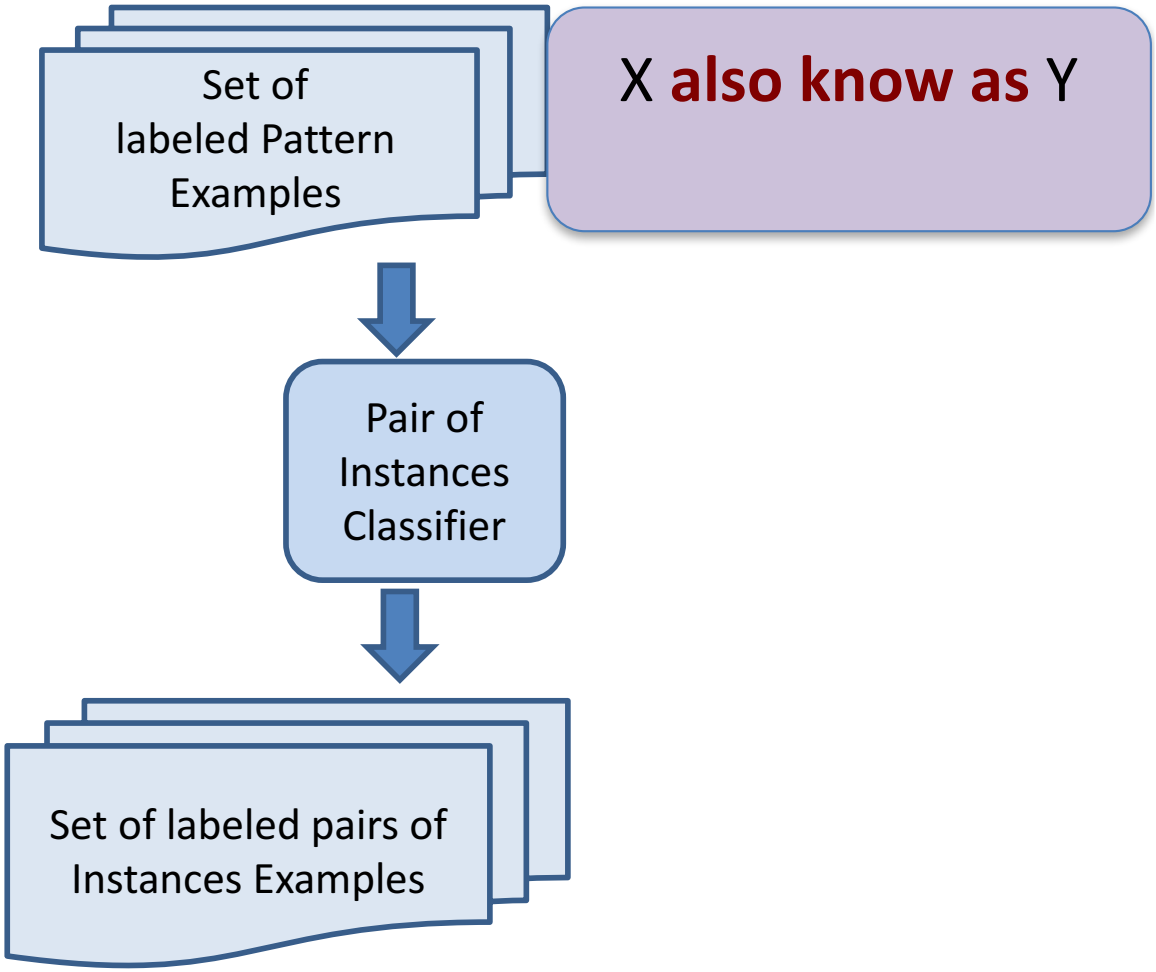
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



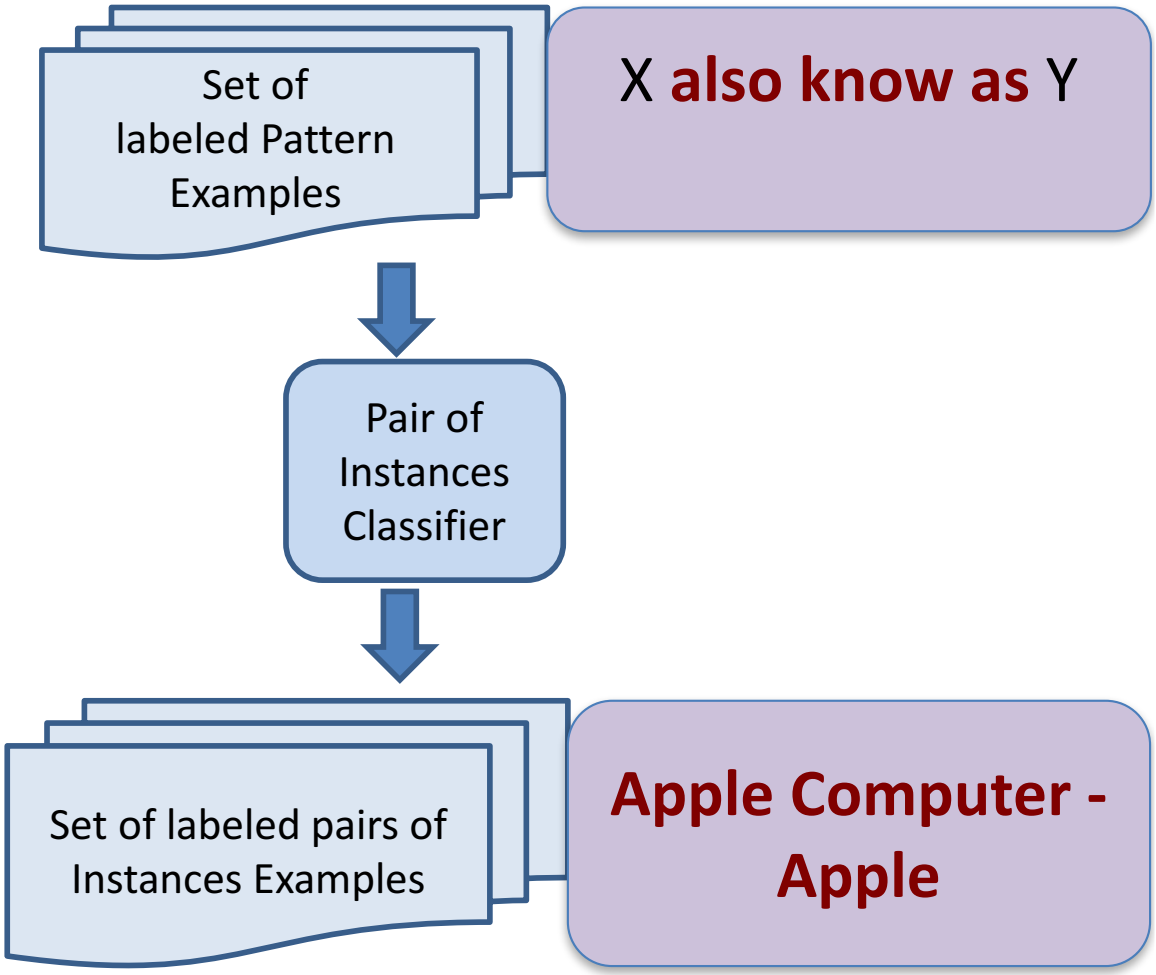
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



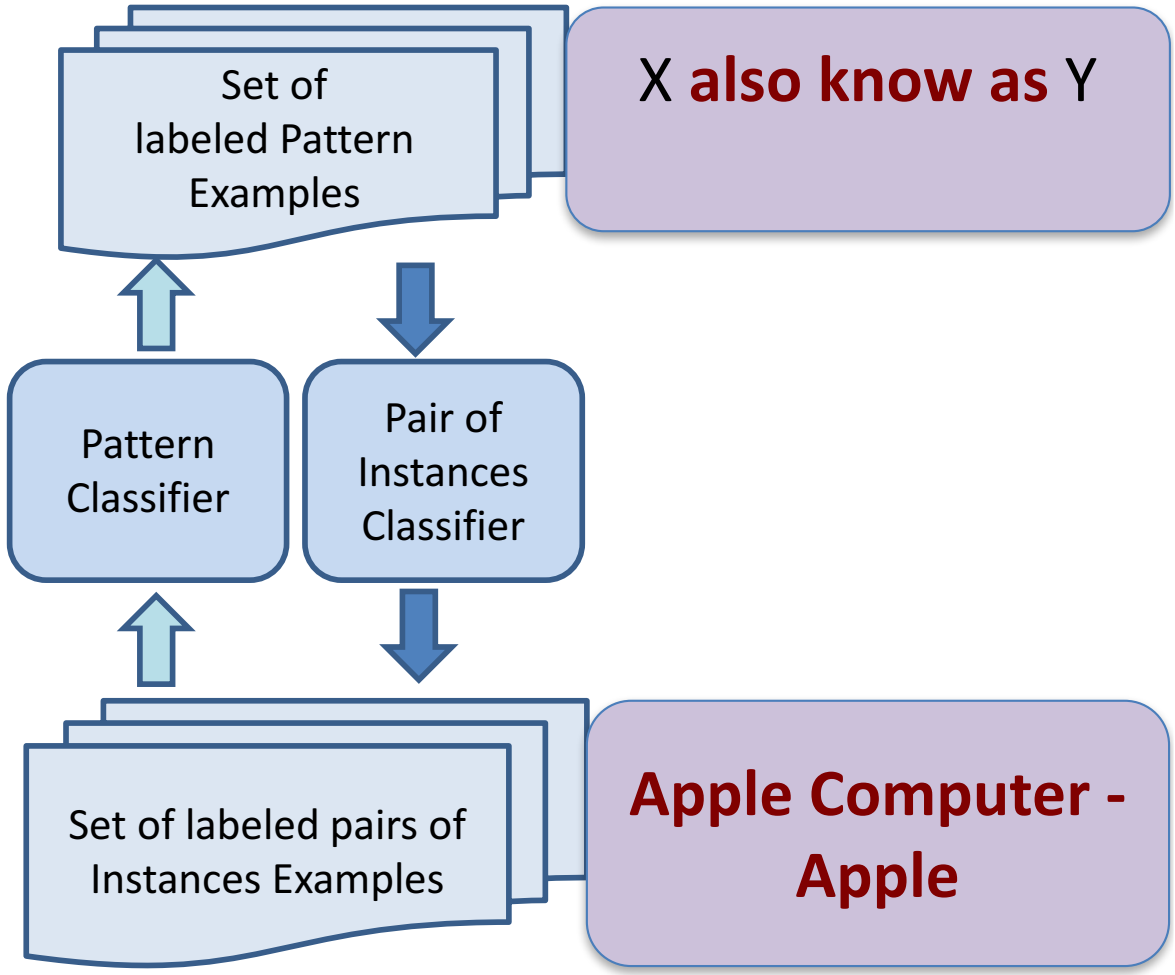
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



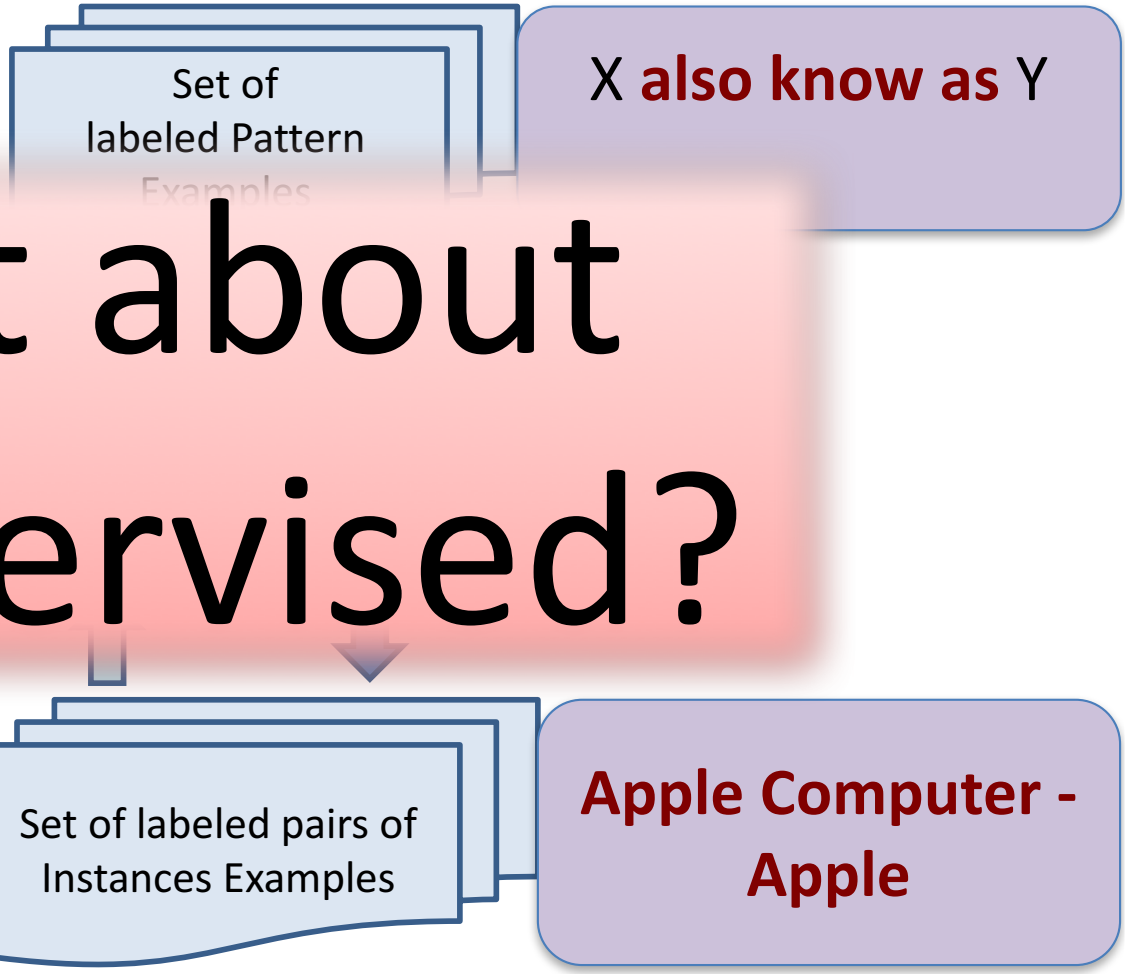
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



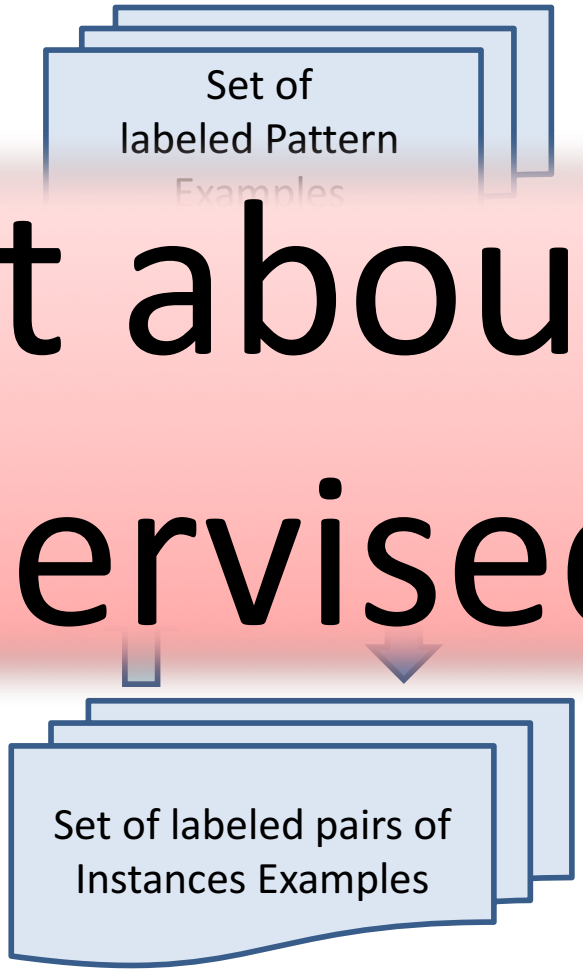
Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bo… ge… nu… an… instances.



Co-Reference and Polysemy Resolution

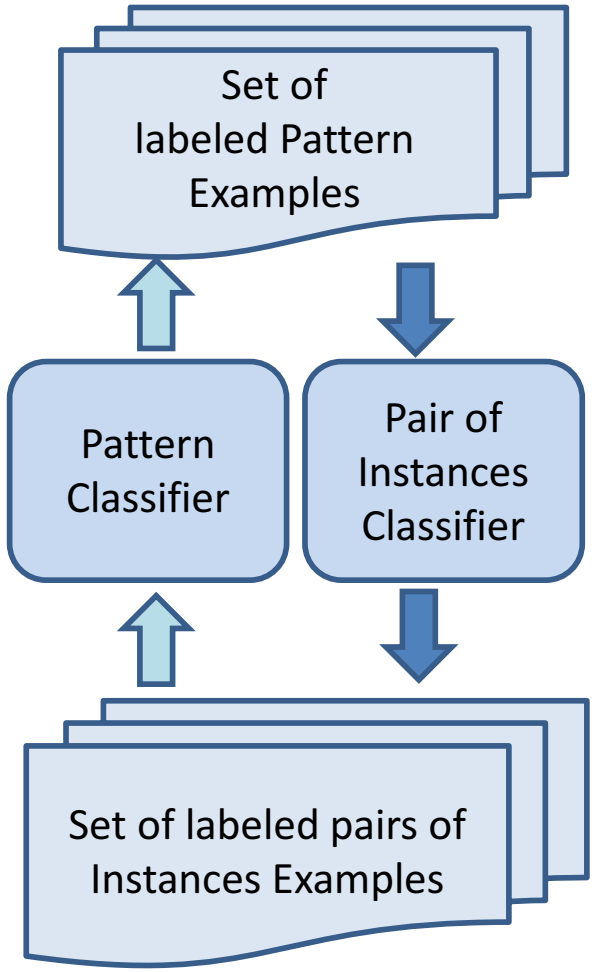
- Semi-supervised Approaches
  - Bo… ge… nu… an… instances.



What about unsupervised?

Co-Reference and Polysemy Resolution

- Semi-supervised Approaches
  - Bootstrap can generate a large number of patterns and relation instances.



Co-Reference and Polysemy Resolution

- Co-Reference Resolution:

[Singh et al., 2011], [Krishnamurthy & Mitchell, 2011],[Dutta & Weikum, 2015]

- Polysemy Resolution:

[Krishnamurthy & Mitchell, 2011], [Galárraga et al., 2014]

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Machine Reading

- Relation Discovery
  - Which new relations are there for given pair of entities?
    - hasAdvisor (JimGray, MikeHarrison)

Machine Reading

- Relation Discovery
  - Which new relations are there for given pair of entities?
    - hasAdvisor (JimGray, MikeHarrison)
    - hasCoAuthor(HectorGarcia-Molina, Gio Wiederhold)

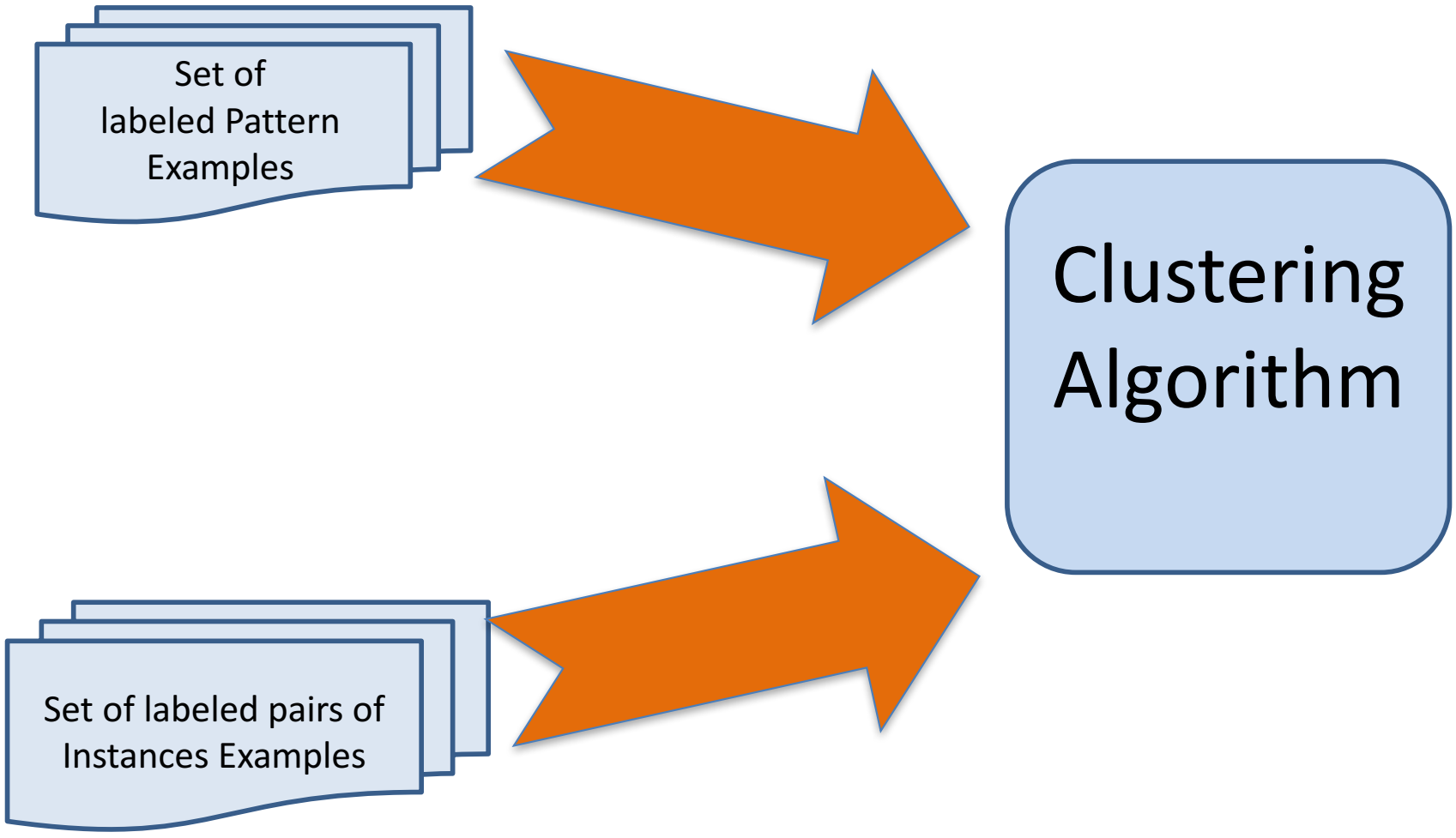
Machine Reading

- Relation Discovery
  - Which new relations are there for given pair of entities?
    - hasAdvisor (JimGray, MikeHarrison)
    - hasCoAuthor(HectorGarcia-Molina, Gio Wiederhold)
    - graduatedAt (JimGray, Berkeley)

Machine Reading

- Relation Discovery
  - Which new relations are there for given pair of entities?
    - hasAdvisor (JimGray, MikeHarrison)
    - hasCoAuthor(HectorGarcia-Molina, Gio Wiederhold)
    - graduatedAt (JimGray, Berkeley)
    - studiedAt (HectorGarcia-Molina, Stanford)
    - bornOn (JohnLennon, 9Oct1940)
    - releasedAlbum (JohnLennon, 10Dec1965)

Relation Discovery



Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Inference

- **Inference** is the act or process of deriving logical conclusions from premises known or assumed to be true [Wikipedia]

Inference

- Manually crafted inference rules
- Automatically learned inference rules
- Data mining the Knowledge Base

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Machine Reading

- Ontology Representation

Facts (RDF triples)

1: (Jim, hasAdvisor, Mike)

2: (Surajit, hasAdvisor, Jeff)

3: (Madonna, marriedTo, GuyRitchie)

4: (Nicolas, marriedTo, Carla)

5: (ManchesterU, wonCup, ChampionsLeague)

Reification:
"Facts about Facts":

6: (1, inYear, 1968)

7: (2, inYear, 2006)

8: (3, validFrom, 22-Dec-2000)

9: (3, validUntil, Nov-2008)

10: (4, validFrom, 2-Feb-2008)

11: (2, source, SigmodRecord)

12: (5, inYear, 1999)

13: (5, location, CampNou)

14: (5, source, Wikipedia)

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Document/Sentence Understanting (Micro-Read)

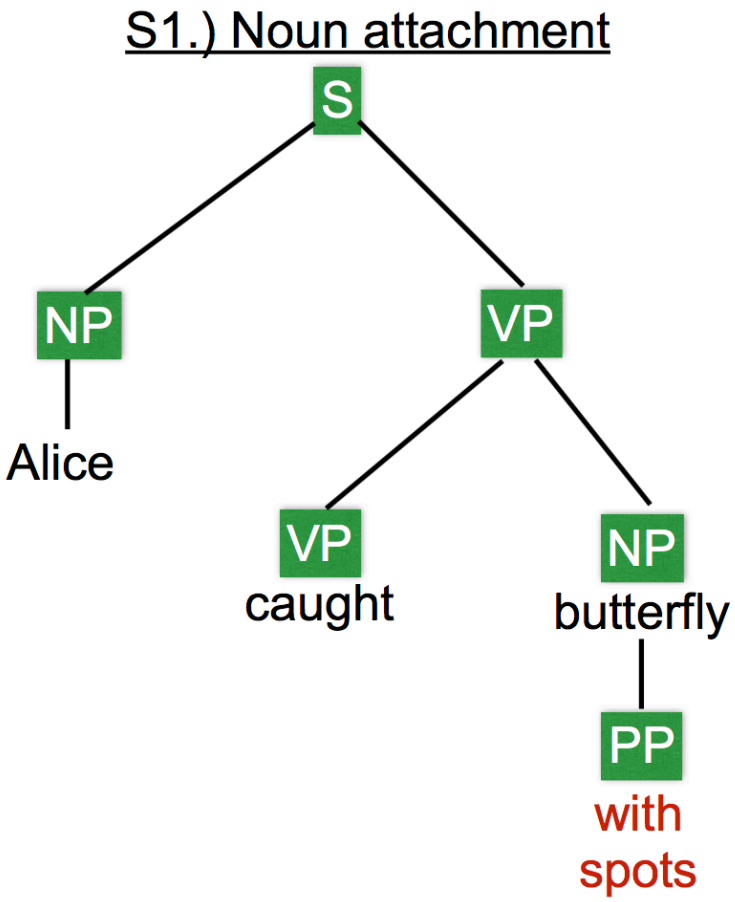
Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

Document/Sentence Understanting (Micro-Read)

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

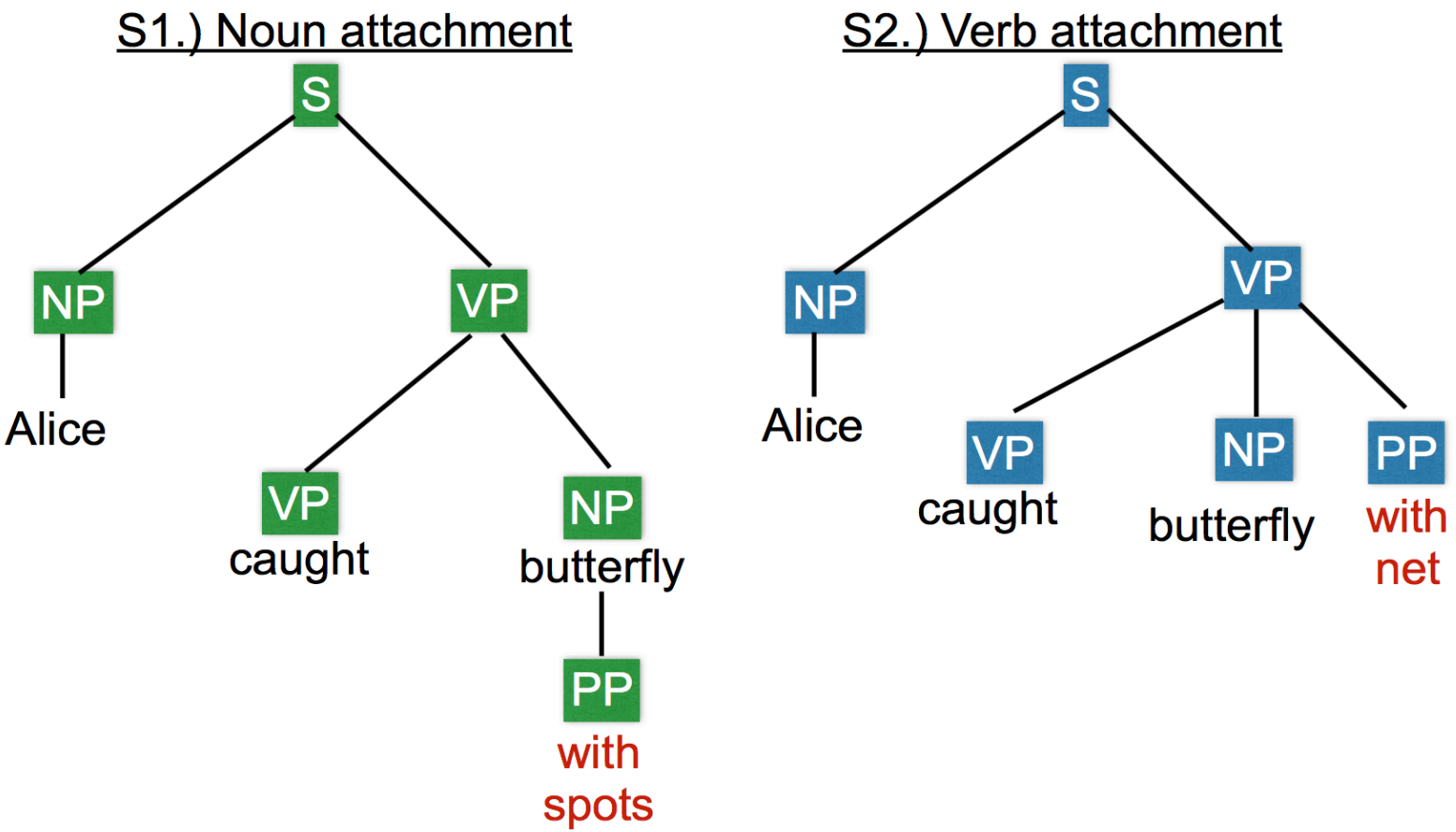


Document/Sentence Understanting (Micro-Read)

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.



Document/Sentence Understanting (Micro-Read)

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),
animalHasCoatPattern(leopard, spots),
animalHasCoatPattern(butterfly, spots), …

caught(*instrument*, *animal*):
caught(net, butterfly),
caught(fish, fishing rod), …

Categories:

city(*city*):
city(Saint Etienne), …

animal(*animal*):
animal(zebra),
animal(butterfly), …

Document/Sentence Understanting (Micro-Read)

Nakashole and Mitchell, 2015

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

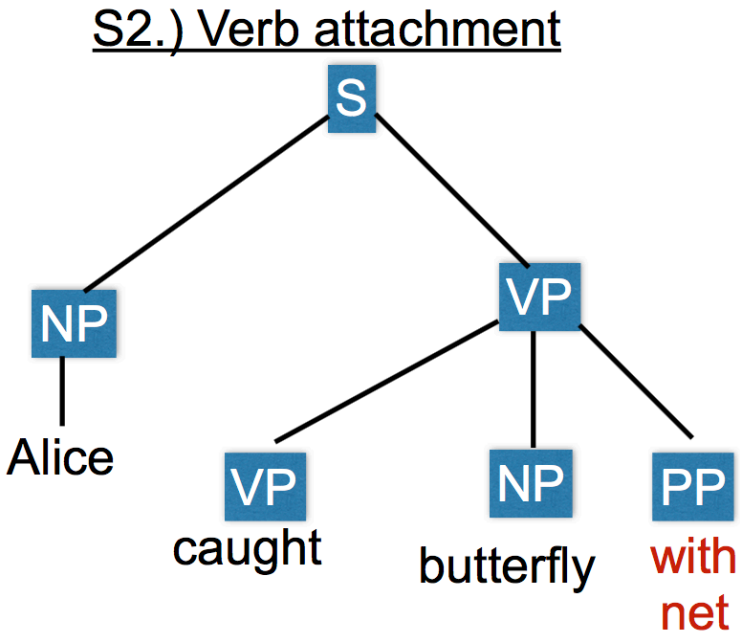
animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),

animalHasCoatPattern(leopard, *spots*),

animalHasCoatPattern(butterfly, *spots*), …

caught(*instrument*, *animal*):

caught(*net*, butterfly),
caught(fish, fishing rod), …



Categories:

city(*city*):
city(Saint Etienne), …

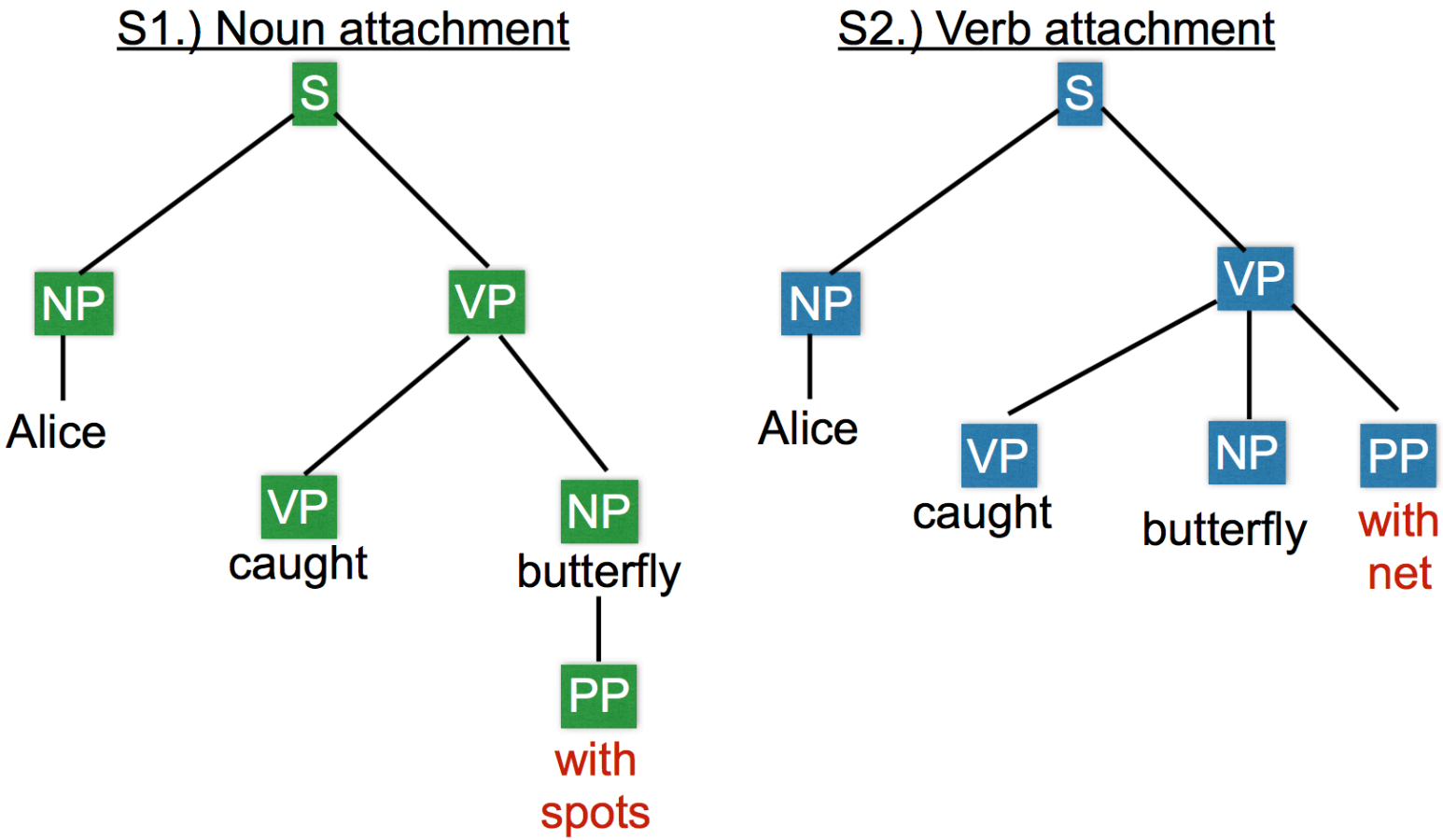
animal(*animal*):
animal(zebra),
animal(butterfly), …

Document/Sentence Understanting (Micro-Read)

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.



Outline

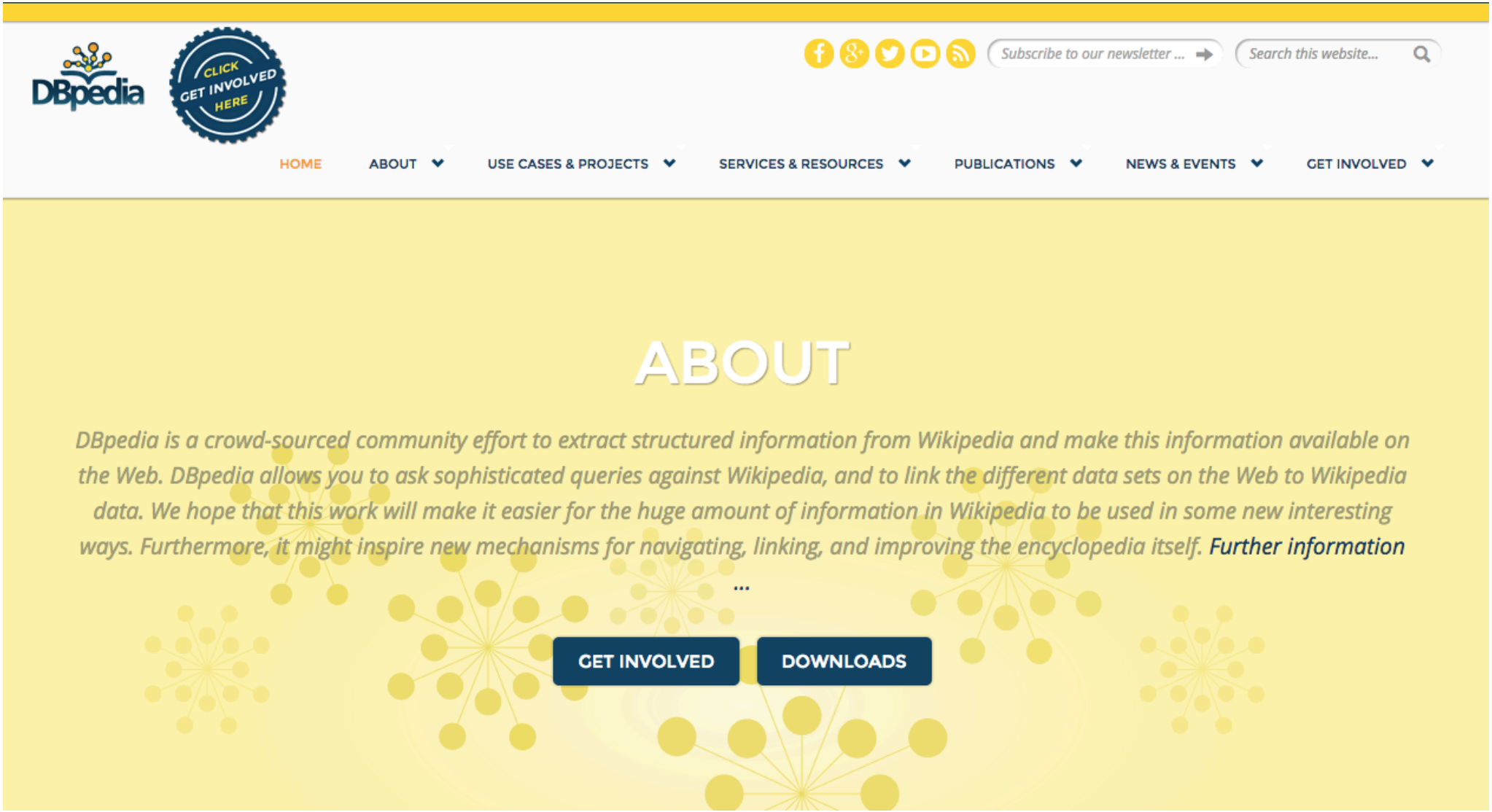
- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - **DBPedia**
  - YAGO
  - KnowItAll
  - NELL

DBPedia

http://wiki.dbpedia.org/



DBPedia

Mapping Wikipedia semi-structured data into RDF triples

DBPedia

Mapping Wikipedia semi-structured data into RDF triples

Semi-structured data

The “Low-Hanging Fruit”

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, specializes that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same same same same same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible]inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[illegible]overing that Salander has hacked into his computer, he persuade[illegible] assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible]gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
same
same
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

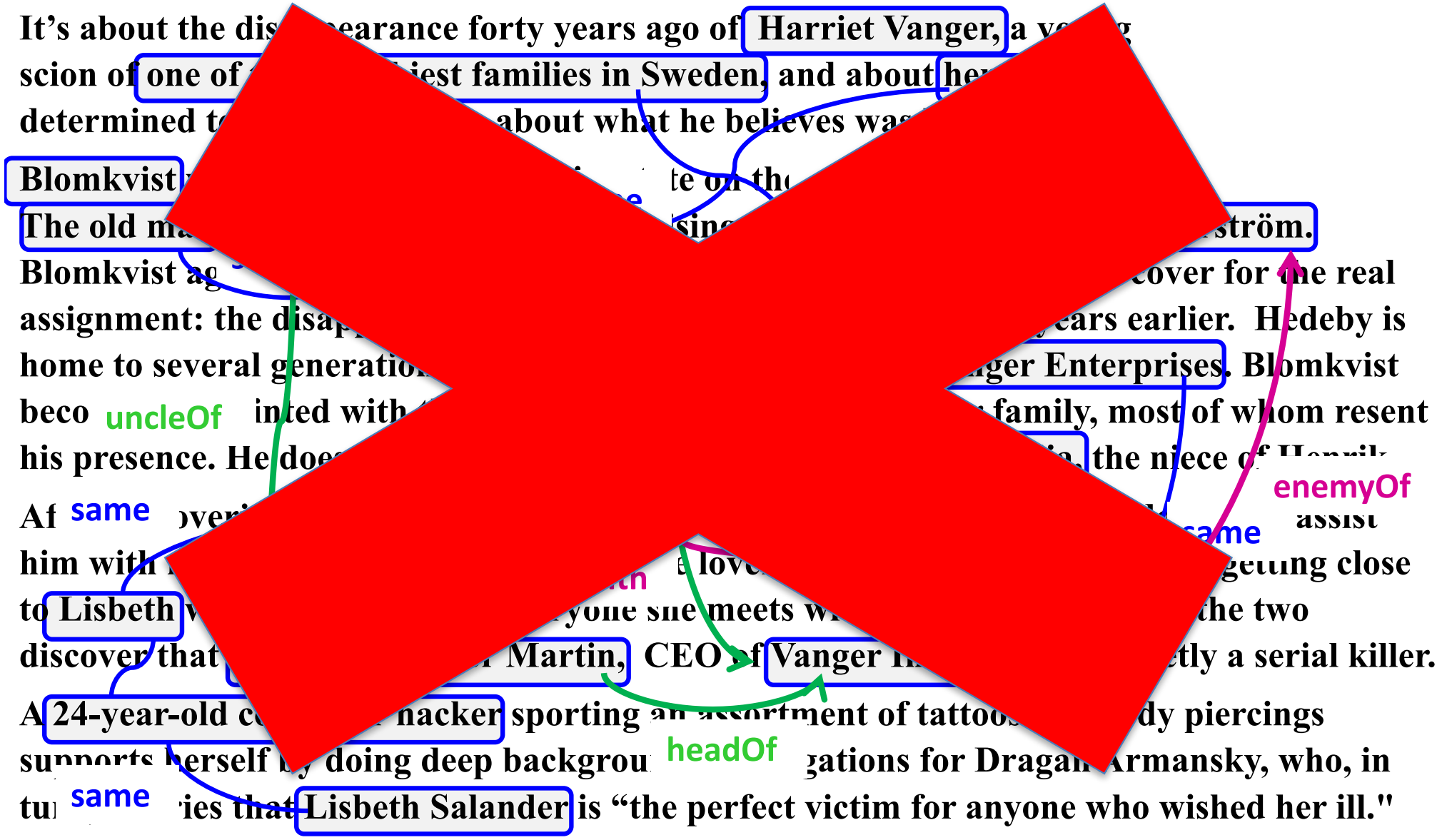
It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] te on th[illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist ag[illegible] spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible] inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.
Af[illegible] overing that Salander has hacked into his c[illegible] persuade[illegible] assist him with research. They even[illegible] lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.
A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible] gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
affairWith
affairWith
enemyOf
same
same
headOf

This slide was adapted from [Hady et al., 2011]

Machine Reading



This slide was adapted from [Hady et al., 2011]

DBPedia

- How to Read Wikipedia Semi-structured data?
  [Lehmann et al., 2014]
  - Parse Wikipedia Markup language
  - Overcome the lack of standard problem
    - Same properties might have different names
    - “Datebirth” and “Birth_date”
    - “Birthplace” and “Birth_place”
  - Instead of “Modeling the World”, try to structure the available information

DBPedia

About: Porto

An Entity of Type : Municipality, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Porto (Portuguese pronunciation: [ˈportu]), also known as Oporto in English, is the second-largest city in Portugal, after Lisbon, and one of the major urban areas in Southwestern Europe. The urban area of Porto, which extends beyond the administrative limits of the city, has a population of 1.4 million (2011) in an area of 389 km2 (150 sq mi), making it the second-largest urban area in Portugal. Porto Metropolitan Area, on the other hand, includes an estimated 1.8 million people.

Property	Value
dbo:abstract	▪ Porto (Portuguese pronunciation: [ˈportu]), also known as Oporto in English, is the second-largest city in Portugal, after Lisbon, and one of the major urban areas in Southwestern Europe. The urban area of Porto, which extends beyond the administrative limits of the city, has a population of 1.4 million (2011) in an area of 389 km2 (150 sq mi), making it the second-largest urban area in Portugal. Porto Metropolitan Area, on the other hand, includes an estimated 1.8 million people. It is recognized as a Gamma-level global city by the Globalization and World Cities (GaWC) Study Group, the only Portuguese city besides Lisbon with such recognition.Located along the Douro river estuary in Northern Portugal, Porto is one of the oldest European centres, and its historical core was proclaimed a World Heritage Site by UNESCO in 1996. The western part of its urban area extends to the coastline of the Atlantic Ocean. Its settlement dates back many centuries, when it was an outpost of the Roman Empire. Its Latin name, Portus Cale, has been referred to as the origin of the name "Portugal", based on transliteration and oral evolution from Latin. In Portuguese the name of the city is spelled with a definite article ("o Porto"; English: the port). Consequently, its English name evolved from a misinterpretation of the oral pronunciation and referred to as Oporto in modern literature and by many speakers.One of Portugal's internationally famous exports, port wine, is named for Porto, since the metropolitan area, and in particular the adegas of Vila Nova de Gaia, were responsible for the production and export of the fortified wine. In 2014, Porto was elected The Best European Destination by the Best European Destinations Agency.
dbo:areaCode	▪ (+351) 22
dbo:country	▪ dbr:Portugal
dbo:daylightSavingTimeZone	▪ dbr:Western_European_Summer_Time
dbo:district	▪ dbr:Porto_District
dbo:elevation	▪ 104.000000 (xsd:double)
dbo:government	▪ dbr:Concelho ▪ dbr:Câmara_municipal
dbo:governmentElevation	▪ 120.000000 (xsd:double)
dbo:governmentPosition	▪ dbr:Porto__governmentPosition__1
dbo:governmentRegion	▪ dbr:Santo_Ildefonso_(Porto)
dbo:highestPosition	▪ dbr:Porto__highestPosition__1
dbo:highestRegion	▪ dbr:Bonfim_(Porto)
dbo:lowest	▪ Sea level
dbo:lowestMountain	▪ dbr:Atlantic_Ocean
dbo:lowestPlace	▪ dbr:Atlantic_Ocean
dbo:maximumElevation	▪ 149.000000 (xsd:double)
dbo:mayor	▪ dbr:Rui_Moreira

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - **YAGO**
  - KnowItAll
  - NELL

Federal University of São Carlos

The YAGO-NAGA Project:
Harvesting, Searching, and Ranking Knowledge from the Web

Location | Press | Deutsch

max planck institut informatik

25ANNIVERSARY | HOME | INSTITUTE | NEWS | DEPARTMENTS | PUBLICATIONS | PEOPLE | SOFTWARE | SERVICES

Departments ▸ Databases and Information Systems ▸ Research ▸ YAGO-NAGA

- Algorithms & Complexity
- Computer Vision and Multimodal Computing
- Computational Biology & Applied Algorithmics
- Computer Graphics
- Databases and Information Systems
  - People
  - Research
    - YAGO-NAGA
      - AIDA
      - AMIE
      - ANGIE
      - DEANNA
      - EVIN
      - HIGGINS
      - HYENA
      - IBEX
      - Javatools
      - K2

YAGO-NAGA

Harvesting, Searching, and Ranking Knowledge from the Web

The YAGO-NAGA project started in 2006 with the goal of building a conveniently searchable, large-scale, highly accurate knowledge base of common facts in a machine-processible representation.

We have already harvested knowledge about millions of entities and facts about their relationships, from Wikipedia and WordNet with careful integration of these two sources. The resulting knowledge base, coined YAGO, has very high precision and is freely available. The facts are represented as RDF triples, and we have developed methods and prototype systems for querying, ranking, and exploring knowledge. Our search engine NAGA provides ranked answers to queries based on statistical models.

Several interlinked sub-projects are growing on the YAGO-NAGA basis. Our vision is a confluence of Semantic Web (Ontologies), Social Web (Web 2.0), and Statistical Web (Information Extraction) assets towards a comprehensive repository of human knowledge. Our methodologies combine concepts, models, and algorithms from several fields, including database systems, information retrieval, statistical learning, and logical reasoning.

People

Weikum, Gerhard
Hoffart, Johannes
Suchanek, Fabian
Berberich, Klaus
Kuzey, Erdal

Yahya, Mohamed
Elbassuoni, Shady
Kasneci, Gjergji
Lewis-Kelham, Edwin
de Melo, Gerard

Nakashole, Ndapandula
Neumann 1, Thomas
Preda, Nicoleta
Sozio, Mauro
Taneva, Bilyana

Wang, Yafang
Yosef, Mohamed Amir

Federal University of São Carlos

The YAGO-NAGA Project:
Harvesting, Searching, and Ranking Knowledge from the Web

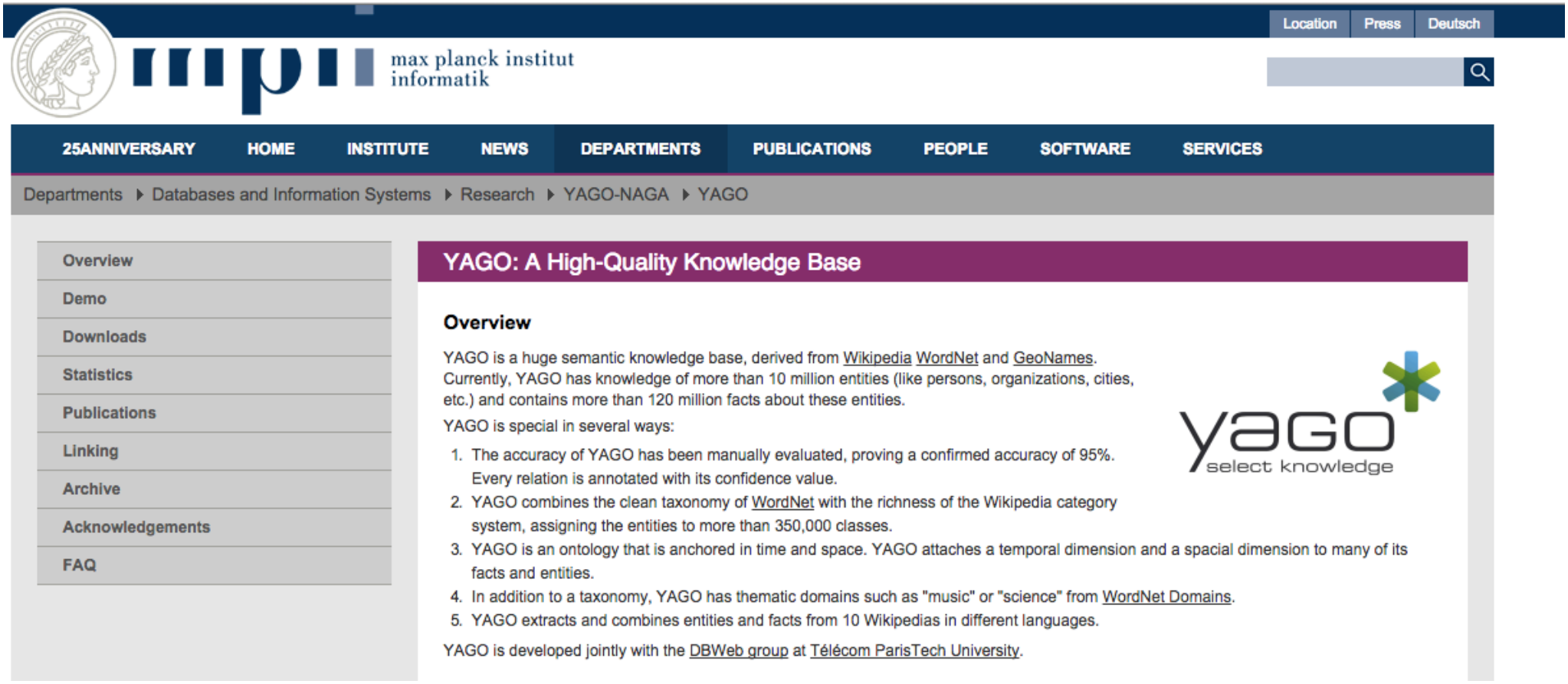
FAPESP

Carnegie Mellon University

YAGO

- **Y**et **A**nother **G**reat **O**ntology - YAGO
- Main Goal: building a conveniently searchable, large-scale, highly accurate knowledge base of common facts in a machine-processable representation

YAGO


Location
Press
Deutsch
max planck institut informatik
25ANNIVERSARY
HOME
INSTITUTE
NEWS
DEPARTMENTS
PUBLICATIONS
PEOPLE
SOFTWARE
SERVICES
Departments ▸ Databases and Information Systems ▸ Research ▸ YAGO-NAGA ▸ YAGO
Overview
Demo
Downloads
Statistics
Publications
Linking
Archive
Acknowledgements
FAQ
YAGO: A High-Quality Knowledge Base
Overview
YAGO is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities.
YAGO is special in several ways:
1. The accuracy of YAGO has been manually evaluated, proving a confirmed accuracy of 95%. Every relation is annotated with its confidence value.
2. YAGO combines the clean taxonomy of WordNet with the richness of the Wikipedia category system, assigning the entities to more than 350,000 classes.
3. YAGO is an ontology that is anchored in time and space. YAGO attaches a temporal dimension and a spacial dimension to many of its facts and entities.
4. In addition to a taxonomy, YAGO has thematic domains such as "music" or "science" from WordNet Domains.
5. YAGO extracts and combines entities and facts from 10 Wikipedias in different languages.
YAGO is developed jointly with the DBWeb group at Télécom ParisTech University.
yago
select knowledge

YAGO

- Turn Web into Knowledge Base [Weikum et al., 2009]
  - Building a comprehensive Knowledge Base of human knowledge
  - knowledge from Wikipedia and WordNet
  - the ontology check itself for precision

YAGO

- The knowledge base is automatically constructed from Wikipedia
- Each article in Wikipedia becomes an entity in the kb (e.g., since Leonard Cohen has an article in Wikipedia, LeonardCohen becomes an entity in YAGO).

YAGO

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
العربية
Asturianu
Български
Bosanski
Català
Česky
Dansk
Deutsch
Eesti
Ελληνικά
Español

Leonard Cohen

From Wikipedia, the free encyclopedia

Leonard Norman Cohen, CC GOQ (born 21 September 1934) is a Canadian singer-songwriter, musician, poet, and novelist. His work often explores religion, isolation, sexuality, and interpersonal relationships.[1] Cohen has been inducted into the American Rock and Roll Hall of Fame and both the Canadian Music Hall of Fame and the Canadian Songwriters Hall of Fame. He is also a Companion of the Order of Canada, the nation's highest civilian honour. In 2011 Cohen received Prince of Asturias Award for literature.

In a speech at Cohen's induction into the Rock and Roll Hall of Fame on March 10, 2008, Lou Reed described Cohen as belonging to the "highest and most influential echelon of songwriters."[2]

The critic Bruce Eder wrote an assessment of Cohen's overall career in popular music, writing, "[Cohen is] one of the most fascinating and enigmatic. . .singer/songwriters of the late '60s. . . [and] has retained an audience across four decades of music-making. . . Second only to Bob Dylan (and perhaps Paul Simon) [in terms of influence], he commands the attention of critics and younger musicians more firmly than any other musical figure from the 1960s who is still working at the outset of the 21st century."[3]

The Academy of American Poets has commented more broadly on Cohen's overall career in the arts, including his work as a poet, novelist, and songwriter, stating that "[Cohen's] successful blending of poetry, fiction, and music is made most clear in *Stranger Music: Selected Poems and Songs*, published in 1993, which gathered more than two hundred of Cohen's poems . . .several novel excerpts, and almost sixty song lyrics. . .While it may seem to some that Leonard Cohen departed from the literary in pursuit of the musical, his fans continue to embrace him as a Renaissance man who straddles the elusive artistic borderlines." [4]

Contents [hide]

Leonard Cohen

Cohen in 2008

Background information

Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

YAGO

Free Text

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
العربية
Asturianu
Български
Bosanski
Català
Česky
Dansk
Deutsch
Eesti
Ελληνικά
Español

Leonard Cohen

From Wikipedia, the free encyclopedia

Leonard Norman Cohen, CC GOQ (born 21 September 1934) is a Canadian singer-so poet, and novelist. His work often explores religion, isolation, sexuality, and interpersonal relationships.[1] Cohen ha d into the American Rock and Roll Hall of Fame and both the Canadian Music Hall of Fame and the Canadian Songwriters le is also a Companion of the Order of Canada, the nation's highest civilian honour. In 2011 Cohen received Prince of Asturias Award for literature.

In a speech at Cohen's induction into the Rock and Roll Hall of Fame on March 10, 2008, Lou Reed described Cohen as belonging to the "highest and most influential echelon of songwriters."[2]

The critic Bruce Eder wrote an assessment of Cohen's overall career in popular music, writing, "[Cohen is] one of the most fascinating and enigmatic. . .singer/songwriters of the late '60s. . . [and] has retained an audience across four decades of music-making. . . Second only to Bob Dylan (and perhaps Paul Simon) [in terms of influence], he commands the attention of critics and younger musicians more firmly than any other musical figure from the 1960s who is still working at the outset of the 21st century."[3]

The Academy of American Poets has commented more broadly on Cohen's overall career in the arts, including his work as a poet, novelist, and songwriter, stating that "[Cohen's] successful blending of poetry, fiction, and music is made most clear in *Stranger Music: Selected Poems and Songs*, published in 1993, which gathered more than two hundred of Cohen's poems . . .several novel excerpts, and almost sixty song lyrics. . .While it may seem to some that Leonard Cohen departed from the literary in pursuit of the musical, his fans continue to embrace him as a Renaissance man who straddles the elusive artistic borderlines." [4]

Contents [hide]

Leonard Cohen

Cohen in 2008

Background information

Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

YAGO

Free Text

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
العربية
Asturianu
Български
Bosanski
Català
Česky
Dansk
Deutsch
Eesti
Ελληνικά
Español

Leonard Cohen

From Wikipedia, the free en

Leonard Norman Co 1934) is a Can et, and novelist. His work often explores religion, is l relatio to the American Rock and Roll Hall of Fame and both the C e is also a Companion of the Order of Canada, the nation's hi ward for literature.

In a speech at Cohen's induction into Reed described Cohen as belonging to the "highest and most influential echelon of

The critic Bruce Eder wrote an assessment o writing, "[Cohen is] one of the most fascinating and enigmatic. . .singer/songwriters of the late '60 cross four decades of music-making. . . Second only to Bob Dylan (and perhaps Paul Simon) [in t ion of critics and younger musicians more firmly than any other musical figure from the 1960 ry."[3]

The Academy of American Poets h he arts, including his work as a poet, novelist, and songwriter, statin made most clear in *Stranger Music: Selected Poems and Song* oems . . .several novel excerpts, and almost sixty song lyrics nard Co ursuit of the musical, his fans continue to embrace h dles the elusive

e]

1 Early life
2 Poetry and novels
3 Recording career
3.1 1960s and 1970s
3.2 1980s
3.3 1990s
3.4 2000s
3.4.1 Post-monastery records: *Ten New Songs*, *Dear Heather* and Anjani's *Blue Alert*
3.4.2 Lawsuits and financial troubles
3.4.3 *Book of Longing*
3.5 2008-2010 World Tour
3.5.1 2008 tour

Leonard Cohen

Cohen in 2008

Background information

Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

Federal University of São Carlos

YAGO

Free Text

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
العربية
Asturianu
Български
Bosanski
Català
Česky
Dansk
Deutsch
Eesti
Ελληνικά
Español

Leonard Cohen

From Wikipedia, the free enc

Leonard Norman Co 1934) is a Cana poet, and novelist. His work often explores religion, is al relation to the American Rock and Roll Hall of Fame and both the C he is also a Companion of the Order of Canada, the nation's hi ard for literature.

In a speech at Cohen's induction int Reed described Cohen as belonging to the "highest and most influential echelon of

The critic Bruce Eder wrote an assessment writing, "[Cohen is] one of the most fascinating and enigmatic. . .singer/songwriters of the late '60s cross four decades of music-making. . . Second only to Bob Dylan (and perhaps Paul Simon) [in te ion of critics and younger musicians more firmly than any other musical figure from the 1960s ry."[3]

The Academy of American Poets h the arts, including his work as a poet, novelist, and songwriter, statin made most clear in *Stranger Music: Selected Poems and Songs* oems . . .several novel excerpts, and almost sixty song lyrics. nard Co ursuit of the musical, his fans continue to embrace h dles the elusiv

1 Early life
2 Poetry and novels
3 Recording career
- 3.1 1960s and 1970s
- 3.2 1980s
- 3.3 1990s
- 3.4 2000s
  - 3.4.1 Post-monastery records: *Ten New Songs*, *Dear Heather* and Anjani's *Blue Alert*
  - 3.4.2 Lawsuits and financial troubles
  - 3.4.3 *Book of Longing*
- 3.5 2008-2010 World Tour
  - 3.5.1 2008 tour

Leonard Cohen

Cohen in 2008

Background information	
Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

InfoBox

FAPESP

Carnegie Mellon University

YAGO

Wikipedia InfoBox

Background information	
Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

YAGO

Semi-structured data

The "Low-Hanging Fruit"

Wikipedia InfoBox

Background information	
Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

YAGO

Semi-structured data

The "Low-Hanging Fruit"

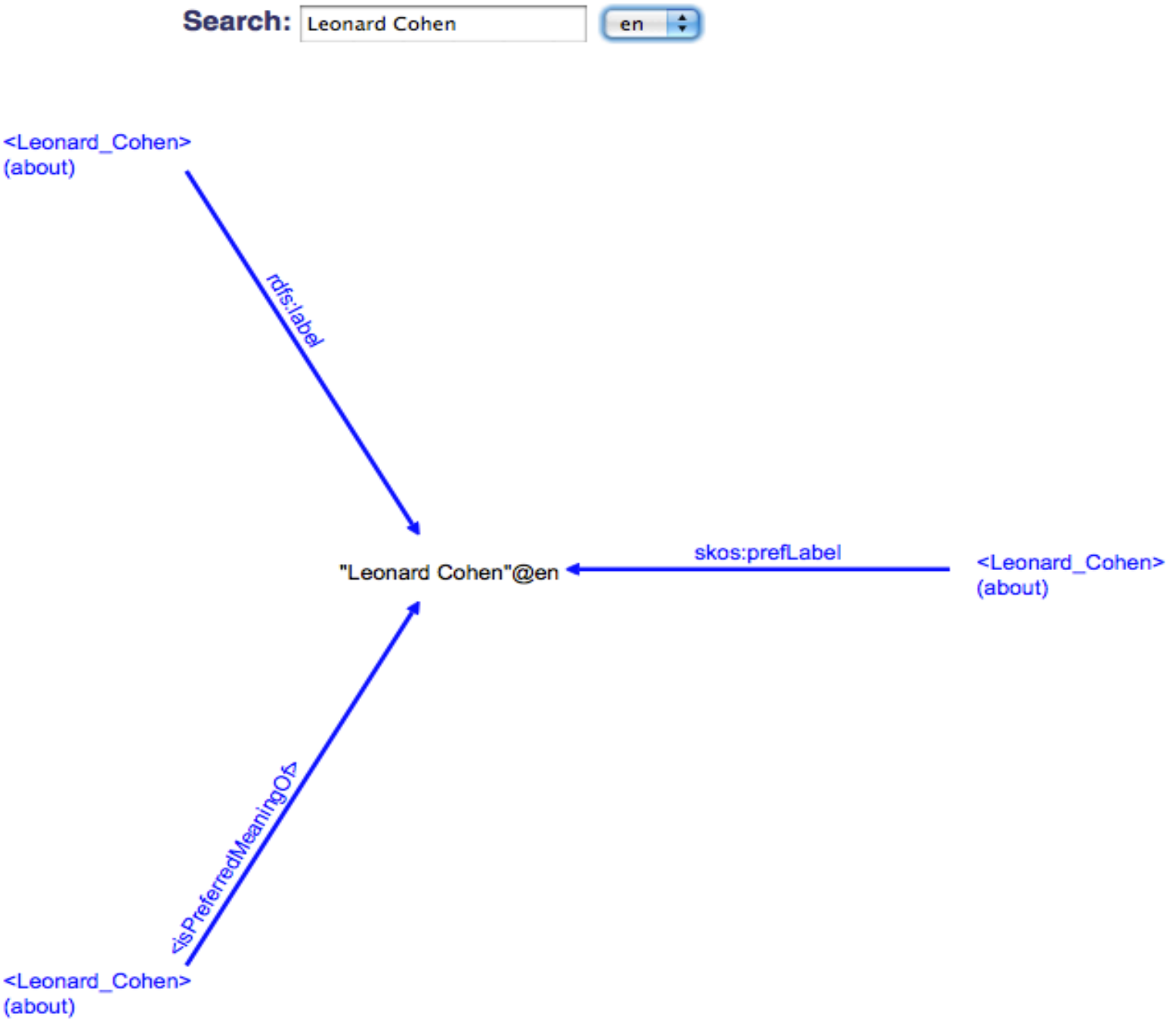
Wikipedia InfoBox

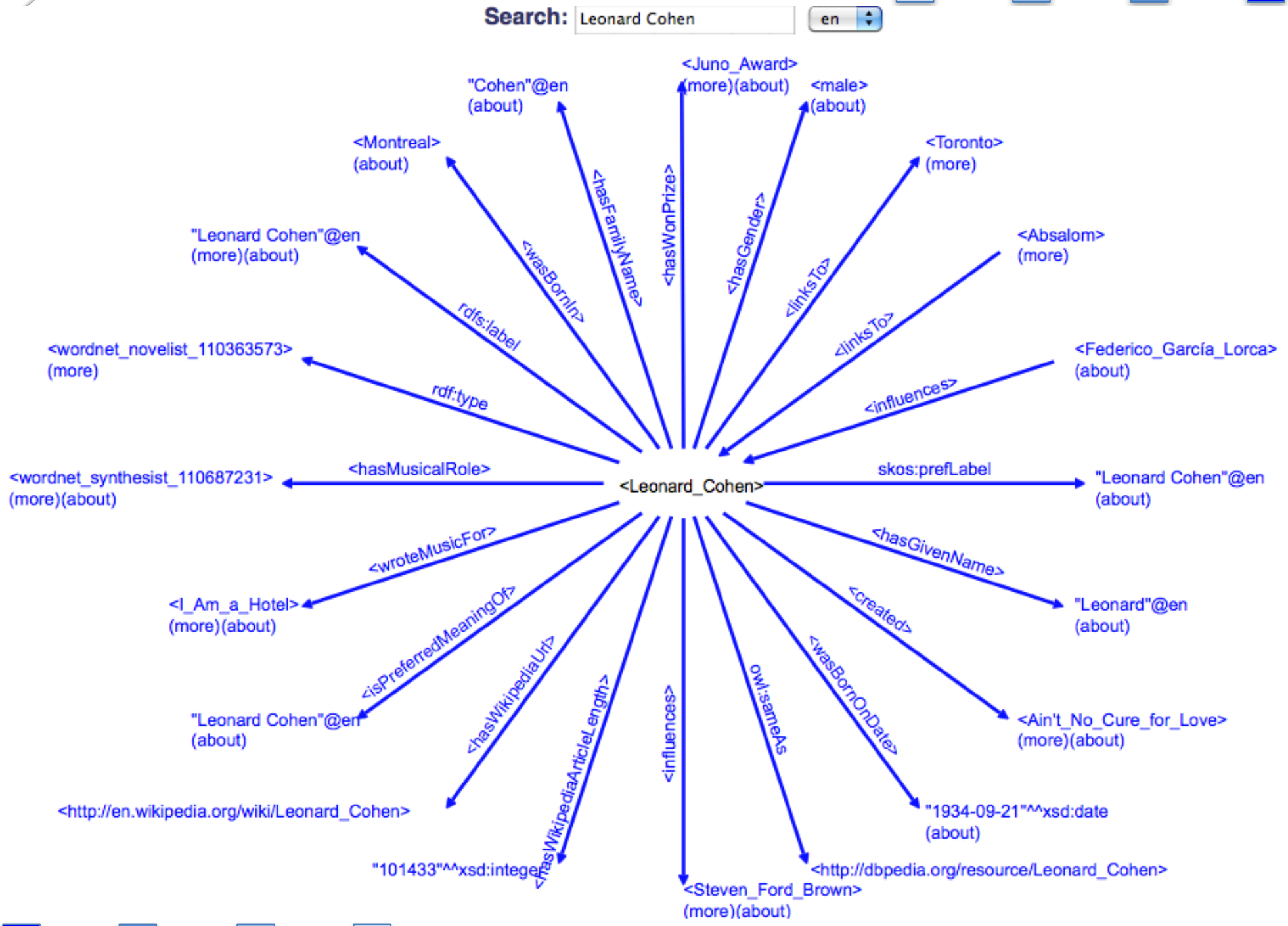
Background information	
Birth name	Leonard Norman Cohen
Born	21 September 1934 (age 78) Montreal, Quebec, Canada
Genres	Folk, folk rock, rock, pop rock, spoken word, synthpop
Occupations	Musician, singer-songwriter, poet, novelist
Instruments	Vocals, guitar, piano, keyboards, synthesizer

YAGO

- Certain categories are exploited to deliver type information (e.g., the article about Leonard Cohen is in the category Canadian male poets, so he becomes a Canadian poet).

Categories: Leonard Cohen | 1934 births | Anglophone Quebec people | Ashkenazi Jews | Canadian buskers | Canadian folk singers | Canadian male singers | Canadian Music Hall of Fame inductees | Canadian male novelists | Canadian people of Lithuanian-Jewish descent | Canadian people of Polish-Jewish descent | Canadian singer-songwriters | Canadian soft rock musicians | Canadian Zen Buddhists | Companions of the Order of Canada | Buddhists of Jewish descent | Genie and Canadian Screen Award winning people | Governor General's Award winning poets | Grammy Award winners | Grand Officers of the National Order of Quebec | Jewish Canadian musicians | Jewish songwriters | Jewish poets | Jewish singers | Jewish Canadian writers | Juno Award winners | Governor General's Performing Arts Award winners | Kohanim | Living people | McGill University alumni | Musicians from Montreal | Mystics | Rinzai Buddhists | Rock and Roll Hall of Fame inductees | People from Westmount, Quebec | Writers from Montreal | Fellows of the American Academy of Arts and Sciences | Converts to Buddhism | Canadian Buddhists | 20th-century Canadian novelists | 20th-century Canadian singers | 21st-century Canadian singers | 20th-century Canadian poets | Canadian male poets | Columbia Records artists | Genie Award winners for Best Achievement in Music - Original Song | Canadian screenwriters | Male screenwriters


Search:
Leonard Cohen
en
<Leonard_Cohen>
(about)
rdfs:label
"Leonard Cohen"@en
skos:prefLabel
<Leonard_Cohen>
(about)
<isPreferredMeaningOf>
<Leonard_Cohen>
(about)


Search:
Leonard Cohen
en
<Leonard_Cohen>
"Cohen"@en (about)
<hasFamilyName>
<Juno_Award> (more)(about)
<hasWonPrize>
<male> (about)
<hasGender>
<Montreal> (about)
<wasBornIn>
<Toronto> (more)
<linksTo>
"Leonard Cohen"@en (more)(about)
rdfs:label
<Absalom> (more)
<linksTo>
<wordnet_novelist_110363573> (more)
rdf:type
<Federico_García_Lorca> (about)
<influences>
<wordnet_synthesist_110687231> (more)(about)
<hasMusicalRole>
skos:prefLabel
"Leonard Cohen"@en (about)
<wroteMusicFor>
<hasGivenName>
<I_Am_a_Hotel> (more)(about)
"Leonard"@en (about)
<isPreferredMeaningOf>
<created>
"Leonard Cohen"@en (about)
<hasWikipediaUrl>
<Ain't_No_Cure_for_Love> (more)(about)
<http://en.wikipedia.org/wiki/Leonard_Cohen>
<hasWikipediaArticleLength>
<influences>
owl:sameAs
<wasBornOnDate>
"1934-09-21"^^xsd:date (about)
"101433"^^xsd:integer
<http://dbpedia.org/resource/Leonard_Cohen>
<Steven_Ford_Brown> (more)(about)

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes
- thematic categories (such as Canadian poetry) are different from conceptual Categories

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes
- thematic categories (such as Canadian poetry) are different from conceptual Categories

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes
- thematic categories (such as Canadian poetry) are different from conceptual Categories

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes
- thematic categories (such as Canadian poetry) are different from conceptual Categories

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes
- thematic categories (such as Canadian poetry) are different from conceptual Categories

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes

YAGO

- For each category of a page [Hoffart et al., 2012]
  - Using shallow parsing, determine the head word of the category name. In the example of Canadian poets, the head word is poets.
  - If the head word is in plural, then proposes the category as a class and the article entity as an instance
  - Link the class to the WordNet taxonomy (most frequent sense of the head word in WordNet)
- only countable nouns can appear in plural form
- only countable nouns can be ontological classes

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

YAGO

- About 100 manually defined relations
  - wasBornOnDate
  - locatedIn
  - hasPopulation
- Categories and infoboxes are exploited to deliver facts (instances of relations).

YAGO

- Manually defined patterns that map categories and infobox attributes to fact templates
  - infobox attribute born=Montreal, thus wasBornIn(LeonardCohen, Montreal)
- Pattern-based extractions resulted in 2 million extracted entities and 20 million facts

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

YAGO

- Based on declarative rules (stored in text files)
- The rules take the form of subject-predicate-object triples, so that they are basically additional facts
- There are different types of rules

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the kn… …lorn cla…
- **Re…** cle…
- **Ext…** …the Wi…s, article titles, and even other regular elements in the source such as headings, links, or references.

Knowledge Representation

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Re**
  cle
- **Ext** the
  Wi s,
  art source
  suc

Inference

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wik… s, arti… source suc…

Knowledge Representation

YAGO

- **Factual rules:** definition of all relations, their domains and ranges, and the definition of the classes that make up the YAGO hierarchy of literal types.
- **Implication rules:** express that if certain facts appear in the knowledge base, then another fact shall be added. Horn clause rules.
- **Replacement rules:** for interpreting micro-formats, cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

YAGO

- **Factual rules:** definition of all relations, their domains and ran the YAG
- **Imp** in the kno rn clau
- **Rep** cleaning up HTML tags, and normalizing numbers.
- **Extraction rules:** apply primarily to patterns found in the Wikipedia infoboxes, but also to Wikipedia categories, article titles, and even other regular elements in the source such as headings, links, or references.

Information Extraction

YAGO

- Automatically verifies consistency
  - Check uniqueness of functional arguments
    - spouse(x,y) $\wedge$ diff(y,z) $\Rightarrow \neg$spouse(x,z)
  - Check domains and ranges of relations
    - spouse(x,y) $\Rightarrow$ female(x)
    - spouse(x,y) $\Rightarrow$ male(y)
    - spouse(x,y) $\Rightarrow$ (f(x)$\wedge$m(y)) $\vee$ (m(x)$\wedge$f(y))

YAGO

- Automatically verifies consistency
  - Check uniqueness of functional arguments
    - spouse(x,y) $\wedge$ diff(y,z) $\Rightarrow \neg$spouse(x,z)
  - Check domains and ranges of relations
    - spouse(x,y) $\Rightarrow$ female(x)
    - spouse(x,y) $\Rightarrow$ male(y)
    - spouse(x,y) $\Rightarrow$ (f(x)$\wedge$m(y)) $\vee$ (m(x)$\wedge$f(y))

Inference

YAGO

- Automatically verifies consistency
  - Hard Constraint
- $hasAdvisor(x,y) \wedge graduatedInYear(x,t) \wedge graduatedInYear(y,s) \Rightarrow s < t$
  - Soft Constraint
- $firstPaper(x,p) \wedge firstPaper(y,q) \wedge author(p,x) \wedge author(p,y)) \wedge inYear(p) > inYear(q) + 5years \Rightarrow hasAdvisor(x,y)$ [0.6]

YAGO

- Automatically verifies consistency
  - Hard Constraint
- $hasAdvisor(x,y) \wedge graduatedInYear(x,t) \wedge graduatedInYear(y,s) \Rightarrow s < t$
  - Soft Constraint
- $firstPaper(x,p) \wedge firstPaper(y,q) \wedge author(p,x) \wedge author(p,y)) \wedge inYear(p) > inYear(q) + 5years \Rightarrow hasAdvisor(x,y)$ [0.6]

Inference

YAGO

- Ontology Representation
  - Entities and Relations of public interest
  - Format: TSV, RDF, XML, N3, Web Interface
  - Learns
    - Instances and patterns from Wikipedia;
    - Taxonomy from WordNet;
    - Geotags information from Geonames.

YAGO

- Named Entity Resolution/Extraction [Theobald & Weikum, 2012]
  - Based on rules and patterns extracted from Wikipedia
  - Disambiguation is a relevant issue
  - **Semi-structured data**

    The "Low-Hanging Fruit"
    - Wikipedia infoboxes & categories
    - HMTL lists & tables, etc.

YAGO

- Named Entity Resolution/Extraction [Theobald & Weikum, 2012]
  - Based on rules and patterns extracted from Wikipedia
  - Disambiguation is a relevant issue
  - Natural Language Processing
    - Wikipedia infoboxes & categories
    - HMTL lists & tables, etc.

YAGO

- Nam eobald & Weiku
  - Ba m Wikipedia
  - Disambiguation is a relevant issue
  - Semi-structured data

Natural Language Processing

Machine Learning

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, specializes that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same same same same same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible]inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[illegible]overing that Salander has hacked into his computer, he persuade[illegible] assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible]gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
same
same
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

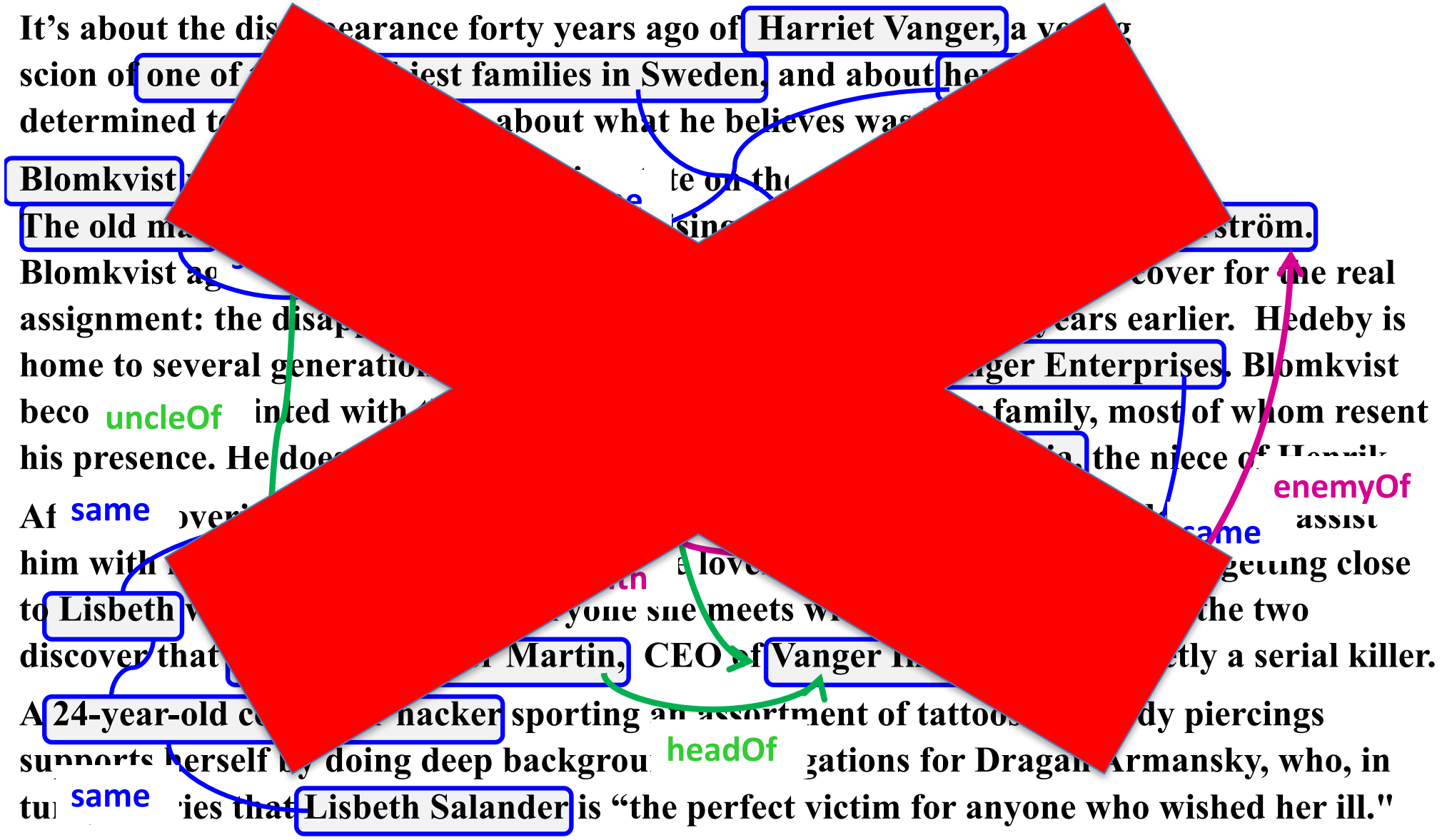
It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] te on th[illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist ag[illegible] spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible] inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.
Af[illegible] overing that Salander has hacked into his c[illegible] persuade[illegible] assist him with research. They even[illegible] lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.
A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible] gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
affairWith
affairWith
enemyOf
same
same
headOf

This slide was adapted from [Hady et al., 2011]

Machine Reading



This slide was adapted from [Hady et al., 2011]

YAGO

- YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages
  - New relations specifically designed to cover time, space and context
  - Wikipedia translated pages as sources for other languages

YAGO

- YAGO3 [Mahdisoltani & Biega & Suchanek, 2015]
  - an extension of the YAGO knowledge base;
  - built from the Wikipedias in multiple languages.
  - fuses the multilingual information with the English WordNet
  - categories, infoboxes, and Wikidata, to learn the meaning of infobox attributes across languages
  - 10 different languages
  - precision of 95%-100% in the attribute mapping
  - enlarges YAGO by 1m new entities and 7m new facts.

YAGO

- More on YAGO:
  - Very nice tutorials:
    - "Knowledge Bases for Web Content Analytics" at WWW 2015, Florence, May 2015.
    - "Semantic Knowledge Bases from Web Sources" at IJCAI 2011, Barcelona, July 2011 "Harvesting Knowledge from Web Data and Text" at CIKM 2010, Toronto, October 2010 "From Information to Knowledge: Harvesting Entities and Relationships from Web Sources" at PODS 2010, Indianapolis, June 2010
  - Project Website:
    - **http://www.mpi-inf.mpg.de/yago-naga/**

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)

The YAGO-NAGA Project: Harvesting, Searching, and Ranking Knowledge from the Web

Research | **Publications** | **People**

The YAGO-NAGA project started in 2006 with the goal of building a conveniently searchable, large-scale, highly accurate knowledge base of common facts in a machine-processible representation.

We have already harvested knowledge about millions of entities and facts about their relationships, from Wikipedia and WordNet with careful integration of these two sources. The resulting knowledge base, coined YAGO, has very high precision and is freely available. The facts are represented as RDF triples, and we have developed methods and prototype systems for querying, ranking, and exploring knowledge. Our search engine NAGA provides ranked answers to queries based on statistical models.

Several interlinked sub-projects are growing on the YAGO-NAGA basis. Our vision is a confluence of Semantic Web (Ontologies), Social Web (Web 2.0), and Statistical Web (Information Extraction) assets towards a comprehensive repository of human knowledge. Our methodologies combine concepts, models, and algorithms from several fields, including database systems, information retrieval, statistical learning, and logical reasoning.

AIDA

AIDA is a method, implemented in an online tool, for disambiguating mentions of named entities that occur in natural-language text or Web tables.

AMIE ↗

AMIE (Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases) is a joint project with the Ontologies group.

ANGIE

ANGIE is an active knowledge system for interactive exploration.

DEANNA

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)

Homepage
About the Institute
Departments:
Algorithms & Complexity
Computer Vision and Multimodal Computing
Computational Biology & Applied Algorithmics
Computer Graphics
Databases and Information Systems
People
Research
YAGO-NAGA
AIDA
AMIE ↗
ANGIE
DEANNA
EVIN
HYENA
JAVATOOLS
K2
NAGA
PATTY
PRAVDA
PROSPERA
RDF3X
SOFIE
UWN
YAGO

The YAGO-NAGA Project: Harvesting, Searching, and Ranking Knowledge from the Web

Research **Publications** **People**

The YAGO-NAGA project started in 2006 with the goal of building a conveniently searchable, large-scale, highly accurate knowledge base of common facts in a machine-processible representation.

We have already harvested knowledge about millions of entities and facts about their relationships, from Wikipedia and WordNet with careful integration of these two sources. The resulting knowledge base, coined YAGO, has very high precision and is freely available. The facts are represented as RDF triples, and we have developed methods and prototype systems for querying, ranking, and exploring knowledge. Our search engine NAGA provides ranked answers to queries based on statistical models.

Several interlinked sub-projects are growing on the YAGO-NAGA basis. Our vision is a confluence of Semantic Web (Ontologies), Social Web (Web 2.0), and Statistical Web (Information Extraction) assets towards a comprehensive repository of human knowledge. Our methodologies combine concepts, models, and algorithms from several fields, including database systems, information retrieval, statistical learning, and logical reasoning.

AIDA

AIDA is a method, implemented in an online tool, for disambiguating mentions of named entities that occur in natural-language text or Web tables.

AMIE ↗

AMIE (Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases) is a joint project with the Ontologies group.

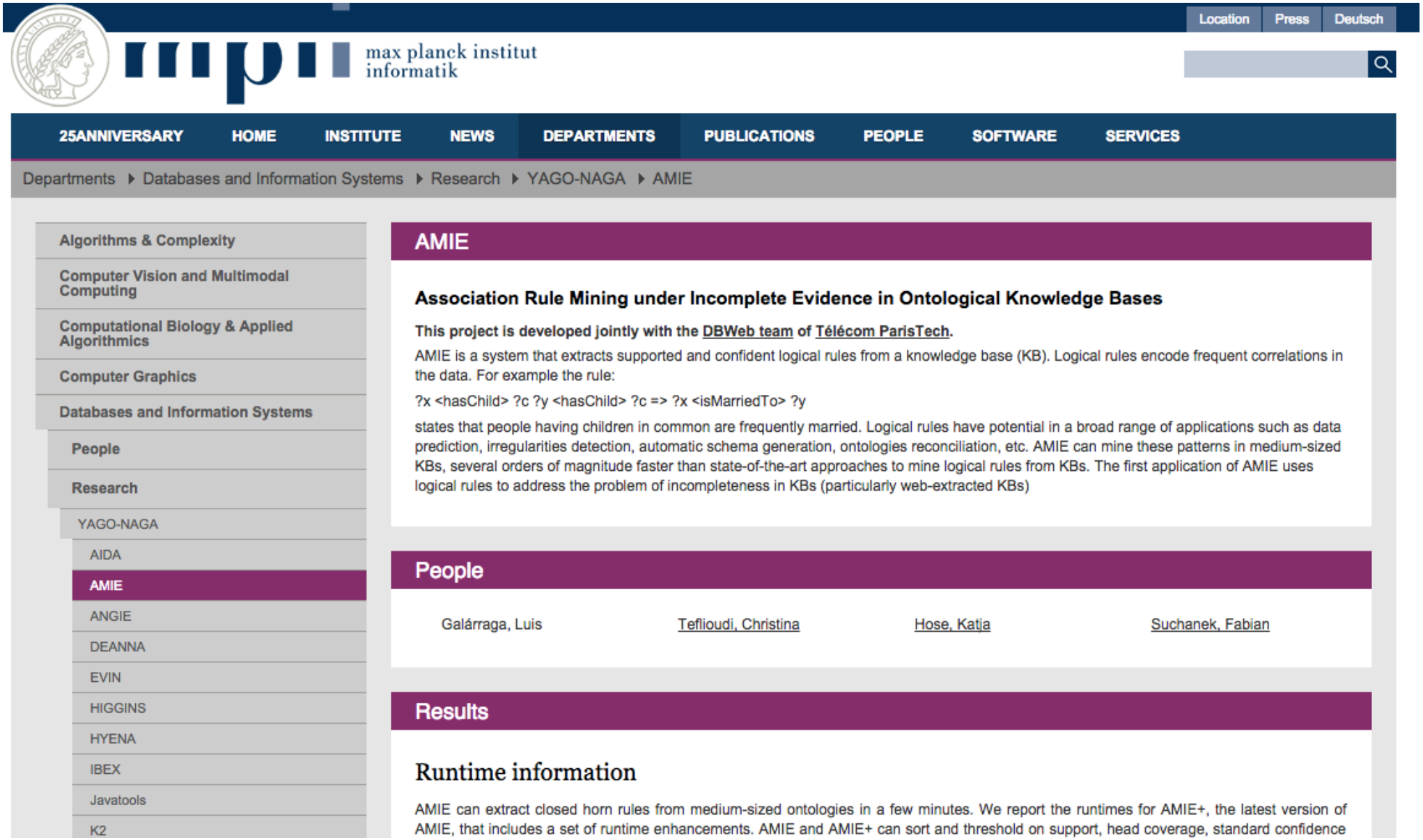
ANGIE

ANGIE is an active knowledge system for interactive exploration.

DEANNA

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)



YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)

AMIE

Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases

This project is developed jointly with the DBWeb team of Télécom ParisTech.

AMIE is a system that extracts supported and confident logical rules from a knowledge base (KB). Logical rules encode frequent correlations in the data. For example the rule:

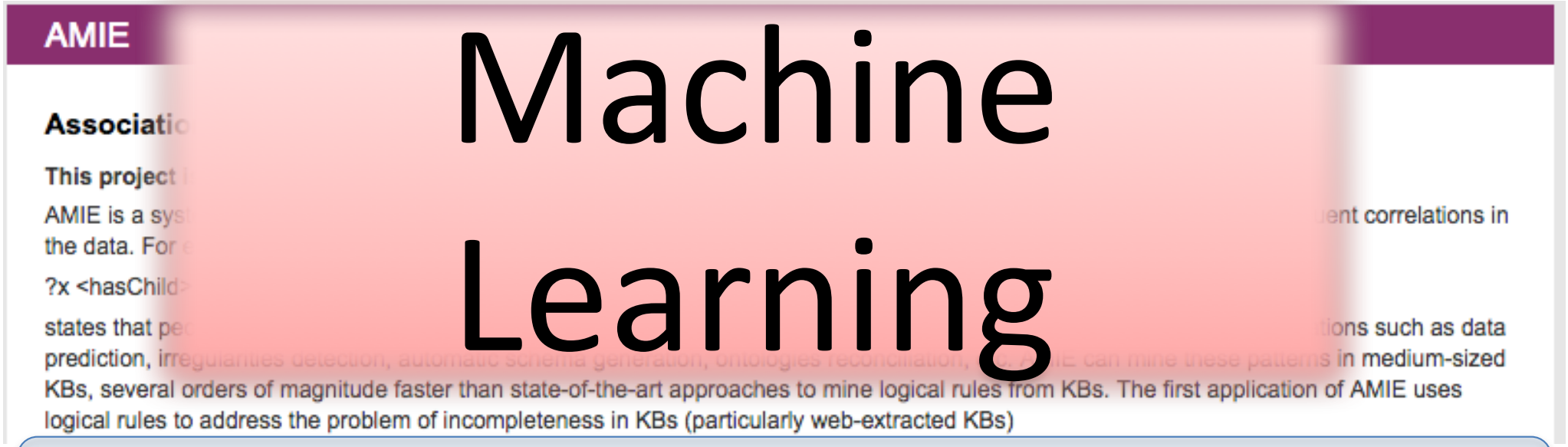
?x <hasChild> ?c ?y <hasChild> ?c => ?x <isMarriedTo> ?y

states that people having children in common are frequently married. Logical rules have potential in a broad range of applications such as data prediction, irregularities detection, automatic schema generation, ontologies reconciliation, etc. AMIE can mine these patterns in medium-sized KBs, several orders of magnitude faster than state-of-the-art approaches to mine logical rules from KBs. The first application of AMIE uses logical rules to address the problem of incompleteness in KBs (particularly web-extracted KBs)

?X <hasChild> ?C ?Y <hasChild> ?C => ?X <isMarriedTo> ?Y

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)

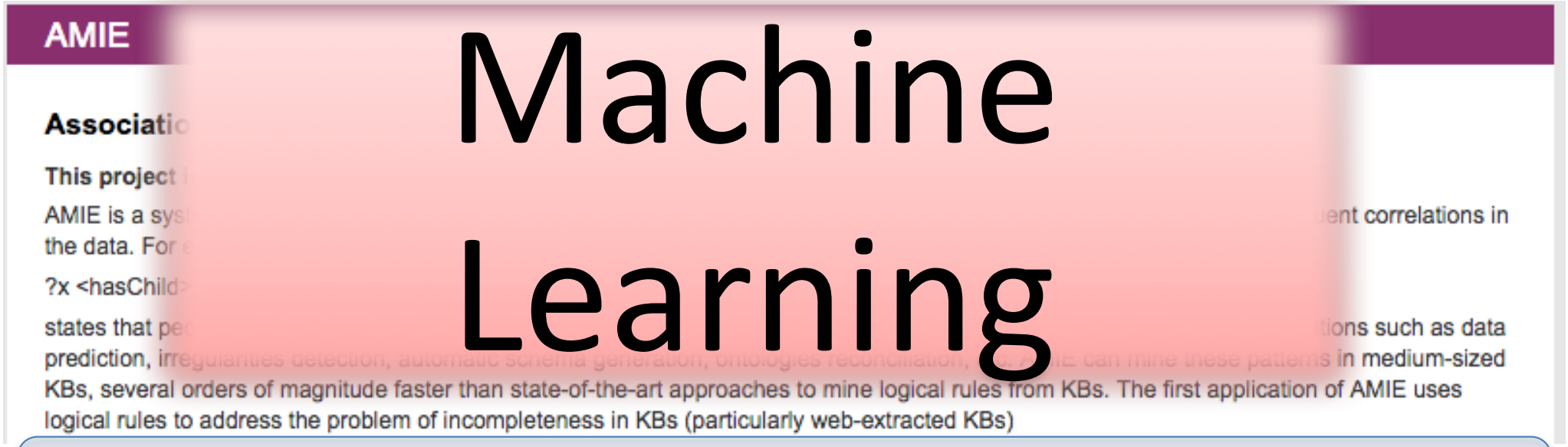


Machine Learning

?X <hasChild> ?C ?Y <hasChild> ?C => ?X <isMarriedTo> ?Y

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)



Machine Learning

?X <hasChild> ?C ?Y <hasChild> ?C => ?X <isMarriedTo> ?Y

Inference

YAGO

- More on YAGO (**http://www.mpi-inf.mpg.de/yago-naga/**)

* YAGO3 is out, and combines information from Wikipedias in 10 different languages:
https://suchanek.name/work/publications/cidr2015.pdf

* Automatically find rules such as “if you are married then you live in the same city as your spouse“ in YAGO:
http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/amie/

* Plans to add commercial products to YAGO:
http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/ibex/

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - **KnowItAll**
  - NELL

Federal University of São Carlos

KnowItAll

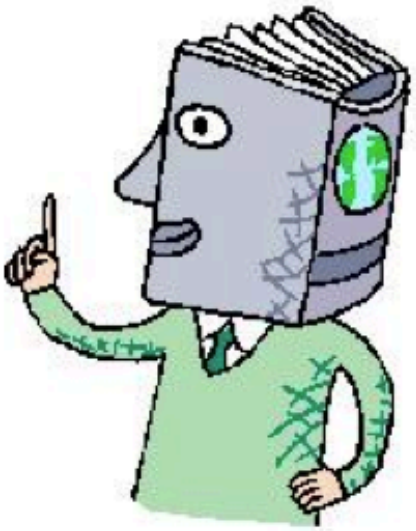
UNIVERSITY OF WASHINGTON | ABOUT US | CONTACT US | MY CSE | INTERNAL

Computer Science & Engineering

UNIVERSITY of WASHINGTON

News & Events | People | Education | Research | Current Students | Prospective Students | Faculty Candidates | Alumni | Industry Affiliates | Support CSE

KnowItAll



1. How can a computer accumulate a massive body of knowledge?
2. What will Web search engines look like in ten years?

To address the questions above, the KnowItAll project has been developing a variety of domain-independent systems that extract information from the Web in an autonomous, scalable manner.

FAPESP

Carnegie Mellon University

KnowItAll: Open Information Extraction

Open Information Extraction

How can a computer accumulate a massive body of knowledge? What will Web search engines look like in ten years?

To address these questions, the Open IE project has been developing a Web-scale information extraction system that reads arbitrary text from any domain on the Web, extracts meaningful information, and stores it in a unified knowledge base for efficient querying. In contrast to traditional information extraction, the Open Information Extraction paradigm attempts to overcome the knowledge acquisition bottleneck by extracting a large number of relations at once.

Demo: TextRunner extracted over 500,000,000 assertions from 100 million Web pages.
Software: ReVerb Open Information Extraction Software and additional information.
Data: Horn-clause inference rules learned by the Sherlock system.
Demo: Selectional Preferences from Web Text compute admissible argument values for a relation.
Data: 10,000 Functional Relations learned from Web Text predict the functionality of a phrase.

KnowItAll: Open Information Extraction

Turing Center

Investigating problems at the crossroads of natural language processing, data mining, Web search, and the Semantic Web.

UNIVERSITY OF WASHINGTON

HOME | PEOPLE | PUBLICATIONS | PRESS | EVENTS | JOBS | CONTACT

People

Director:
Prof. Oren Etzioni

Advisory Board:
Prof. Emily Bender
Prof. David Notkin
Dr. Jonathan Pool

Assistant Director:
Patrick Allen

Events Committee:
Prof. Emily Bender
Christie Evans
Dr. Jonathan Pool

Research Scientists:
Niranjan Balasubramanian
Prof. Mausam
Michael Schmitz
Dr. Stephen Soderland

Affiliated Faculty:
Prof. Emily Bender
Prof. Jeff Bilmes
Prof. Pedro Domingos
Prof. Katrin Kirchhoff
Prof. Mari Ostendorf
Prof. Dan Weld

Students:
Janara Christensen
Tony Fader
Jeff Huang
Tom Lin
Alan Ritter
Stefan Schoenmackers
Fei Wu

Alumni:
Michele Banko (Microsoft, 2009)
Mike Cafarella (University of Michigan, 2009)
Prof. Doug Downey (Northwestern University, 2008)
Michael Skinner (Google, 2008)
Dr. Ana-Maria Popescu (Yahoo, 2007)
Kobi Reiter (Google, 2007)
Prof. Alex Yates (Temple University, 2007)

KnowItAll

- Motivation: New Paradigm for Search [Etzioni, 2008]
  - The future of Web Search
  - Read the Web instead of retrieving Web pages to perform Web Search

KnowItAll

- Information Extraction (IE) + tractable inference
  - IE(sentence) = who did what?
    - speaker(Sheila McIlraith, IJCAI-16)
  - Inference = uncover implicit information
    - Will Pittsburgh Steelers be champions again?

- Open Information Extraction [Banko et al., 2007]

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

Open Information Extraction

[Banko et al., 2007]

- Open IE systems avoid specific nouns and verbs
- Extractors are unlexicalized—formulated only in terms of:
  - syntactic tokens (e.g., part-of-speech tags)
  - closed-word classes (e.g., of, in, such as).
- Open IE extractors focus on generic ways in which relationships are expressed in English
  - naturally generalizing across domains.

Open Information Extraction

[Banko et al., 2007]

- Open IE extractors focus on generic ways in which relationships are expressed in English
  - naturally generalizing across domains.

Relation Discovery

Open Information Extraction

- Open IE systems are traditionally based on three steps [Etzioni et al., 2011]:
  - **1. Label:** Sentences are automatically labeled with extractions using heuristics or distant supervision.

Unsupervised Learning

Open Information Extraction

- Open IE systems are traditionally based on three steps [Etzioni et al., 2011]:
  - **1. Label:** Sentences are automatically labeled with extractions using heuristics or distant supervision.
  - **2. Learn:** A relation phrase extractor is learned using a sequence-labeling graphical model (e.g., CRF).

Supervised Learning

Open Information Extraction

- Open IE systems are traditionally based on three steps [Etzioni et al., 2011]:
  - **1. Label:** Sentences are automatically labeled with ext
  - **2.** L sing a sequence-labeling graphical model (e.g., CRF).
  - **3. Extract:** given a sentence as input, identifies a candidate pair of NP arguments (Arg1, Arg2) from the sentence, and then uses the learned extractor to label each word between the two arguments as part of the relation phrase or not.

Inference

Open Information Extraction

- TextRunner [Banko et al., 2007] was the first OIE system;
- OIE became the main focus of the KnowItAll project;
- Two main problems:
  - incoherent extractions;
  - uninformative relations

Open Information Extraction

- incoherent extractions

Sentence	Incoherent Relation
The guide *contains* dead links and *omits* sites.	contains omits
The Mark 14 *was central* to the *torpedo* scandal of the fleet.	was central torpedo
They *recalled* that Nungesser *began* his career as a precinct leader.	recalled began

Open Information Extraction

- uninformative relations

is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. in, has a cameo in
made	made a deal with, made a promise to
took	took place in, took control over, took advantage of
gave	gave birth to, gave a talk at, gave new meaning to
got	got tickets to see, got a deal on, got funding from

Open Information Extraction

- TextRunner was based on

Relative Frequency	Category	Simplified Lexico-Syntactic Pattern
37.8	Verb	E_1 Verb E_2 *X established Y*
22.8	Noun+Prep	E_1 NP Prep E_2 *X settlement with Y*
16.0	Verb+Prep	E_1 Verb Prep E_2 *X moved to Y*
9.4	Infinitive	E_1 to Verb E_2 *X plans to acquire Y*
5.2	Modifier	E_1 Verb E_2 Noun *X is Y winner*
1.8	Coordinate$_n$	E_1 (and|,|-|:) E_2 NP *X-Y deal*
1.0	Coordinate$_v$	E_1 (and|,) E_2 Verb *X , Y merge*
0.8	Appositive	E_1 NP (:|,)? E_2 *X hometown : Y*

OIE: the second generation

- New syntactic constraint based on POS tag patterns

$$V \mid VP \mid VW^*P$$
$$V = \text{verb particle? adv?}$$
$$W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det})$$
$$P = (\text{prep} \mid \text{particle} \mid \text{inf. marker})$$

- simple verb phrase (e.g., invented)
- verb phrase followed immediately by a preposition or particle (e.g., located in)
- verb phrase followed by a simple noun phrase and ending in a preposition or particle (e.g., has atomic weight of)
- multiple possible matches, then the longest possible match is chosen.

OIE: the second generation

- New lexical constraint to separate valid relation phrases from over-specified relation phrases
- The lexical constraint is based on the intuition that a valid relation phrase should take many distinct arguments in a large corpus.

OIE: the second generation

- New OIE System: ReVerb [Fader et al., 2011]
  - Input: a POS-tagged and NP-chunked sentence
  - Output: a set of (x,r,y) extraction triples
  - Based on two extraction algorithm:
    - 1. Relation Extraction: based on the new constraints
    - 2. Argument Extraction: For each relation phrase r identified in Step 1, find the nearest noun phrase x to the left and the nearest noun phrase y to the right of r in s.

OIE: the second generation

- New OIE System: ReVerb [Fader et al., 2011]

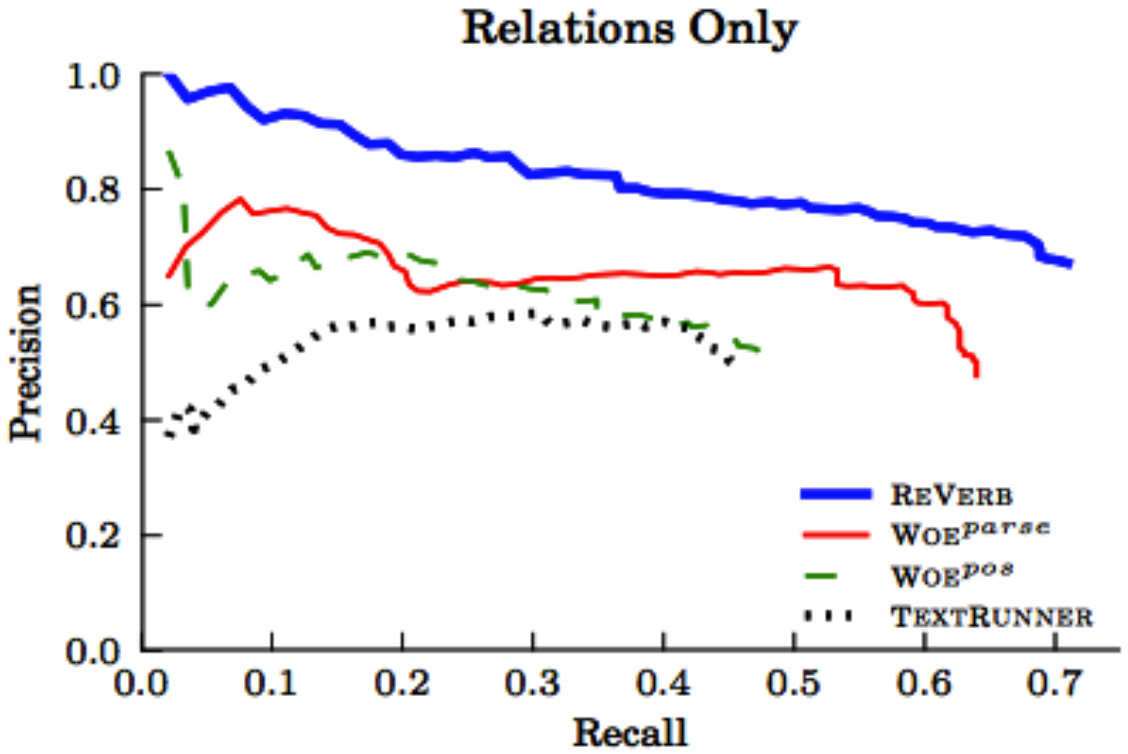


Figure 2: REVERB identifies correct relation phrases with substantially higher precision and recall than state-of-the-art open extractors, including WOE^{parse} that uses patterns learned over dependency parse paths.

OIE: the second generation

Open Information Extraction

Hosted by AI2 Created at

Argument 1: porto Relation: Argument 2: All Search

286 answers from **886 sentences** (cached)

Do you mean...

Porto (78)

Porto (Portuguese pronunciation: ['portu]), also known as Oporto in English, is the second largest city in Portugal, after Lisbon, and one of the major urban areas in Southern Europe. Its administrative limits (an area of 41.66 km²/16 sq.mi) include a population of 237,584 (2011) inhabitants distributed within 15 civil parishes. The urbanized area of Porto, which extends beyond the administrative limits of the city, has a population of 1.3 million (2011) in an area of 389 km (150 sq mi),...read more

F.C. Porto (14)

Futebol Clube do Porto MH IH MH OM (Portuguese pronunciation: [futebɔɫ 'klube du 'portu]) (Euronext: FCP), commonly known as FC Porto, Porto, or FCP, is a Portuguese multi-sports club from the city of Porto, in the northern region of the country. Although they successfully compete in a number of different sports, FC Porto is mostly known for its association football team. Founded in Porto in September 28, 1893, it is one of the "Três Grandes" (Big Three, in English), football clubs in...read more

OIE: the second generation

Table extracted from [Etzioni et al., 2011]

Category	Patterns	Frequency Arg1	Frequency Arg2
Basic NP	NN, JJ NN, etc	65% **Chicago** *was founded in* 1833.	60% Calcium *prevents* **osteoporosis**.
Prepositional Attachments	NP PP^+	19% **The forest in Brazil** *is threatened by* ranching.	18% Lake Michigan *is one of* **the five Great Lakes of North America**.
List	NP (,NP)* ,? and/or NP	15% **Google and Apple** *are headquartered in* Silicon Valley.	15% A galaxy *consists of* **stars and stellar remnants**.
Independent Clause	(that|WP|WDT)? NP VP NP	0% **Google will acquire YouTube,** *announced* the New York Times.	8% Scientists *estimate* **that 80% of oil remains a threat**.
Relative Clause	NP (that|WP|WDT) VP NP?	<1% **Chicago, which is located in Illinois,** *has* three million residents.	6% Most galaxies *appear to be* **dwarf galaxies, which are small**.

Table 4: Taxonomy of arguments for binary relationships. In each sentence, the argument is bolded and the relational phrase is italicized. Multiple patterns can appear in a single argument so percentages do not need to add to 100. In the interest of space, we omit argument structures that appear in less than 5% of extractions. Upper case abbreviations represent noun phrase chunk abbreviations and part-of-speech abbreviations.

OIE: the second generation

- New OIE System: ArgLearner [Etzioni et al., 2011]

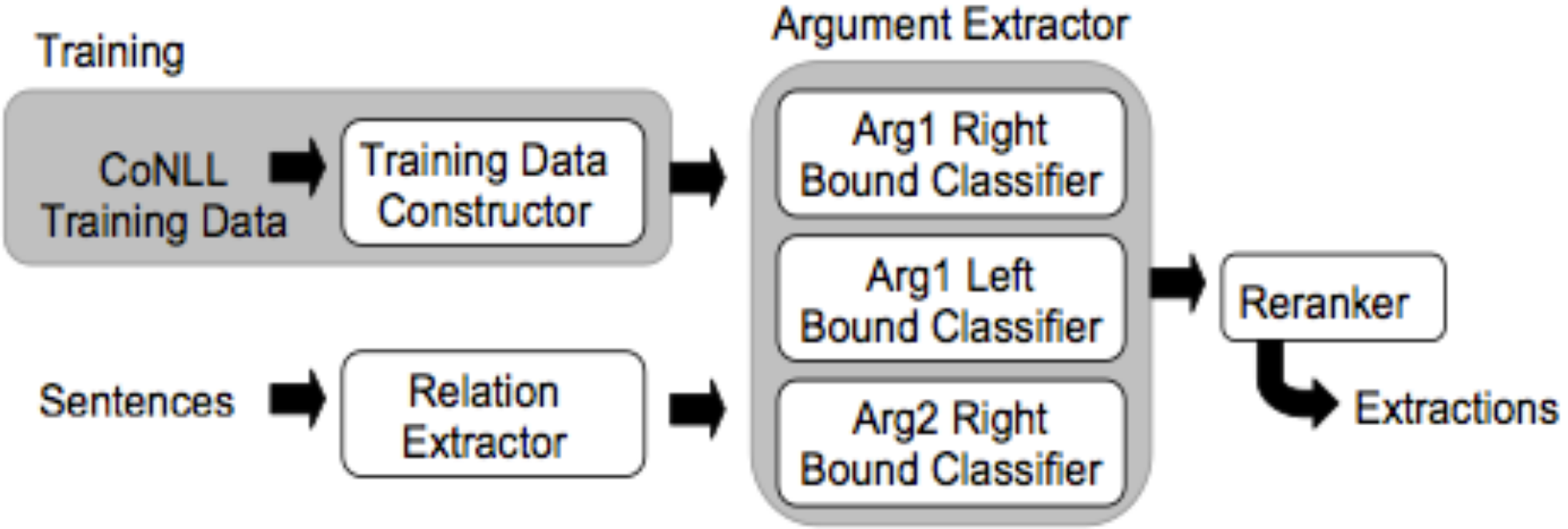


Figure 4: ARGLEARNER's system architecture.

OIE: the second generation

- New OIE System:
- ReVerb + ArgLearner = R2A2 [Etzioni et al., 2011]

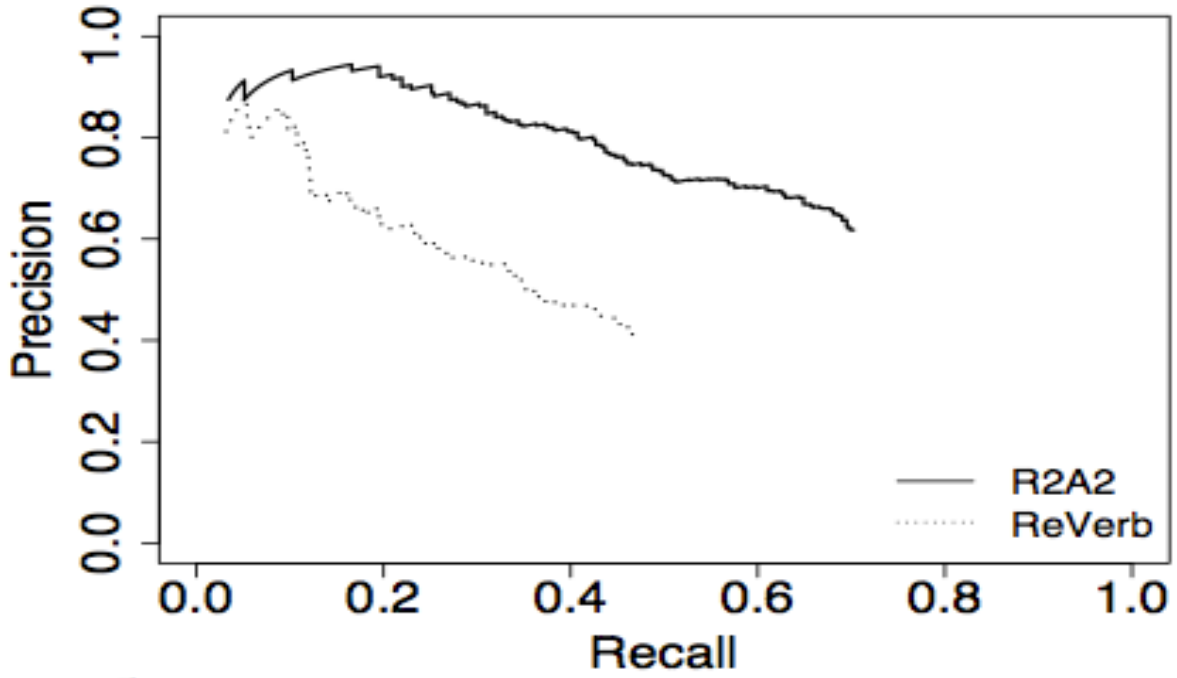


Figure 5: R2A2 has substantially higher recall and precision than REVERB.

OIE: the second generation

- New OIE System:
- ReVerb + ArgLearner = R2A2 [Etzioni et al., 2011]

Free text

Hearst-patterns; clustering by verbal phrases

Natural-language processing

Advanced patterns & iterative bootstrapping

(“Dual Iterative Pattern Relation Extraction”)

Fi
ReVerb.

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, specializes that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same same same same same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible]inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[illegible]overing that Salander has hacked into his computer, he persuade[illegible] assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible]gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
same
same
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

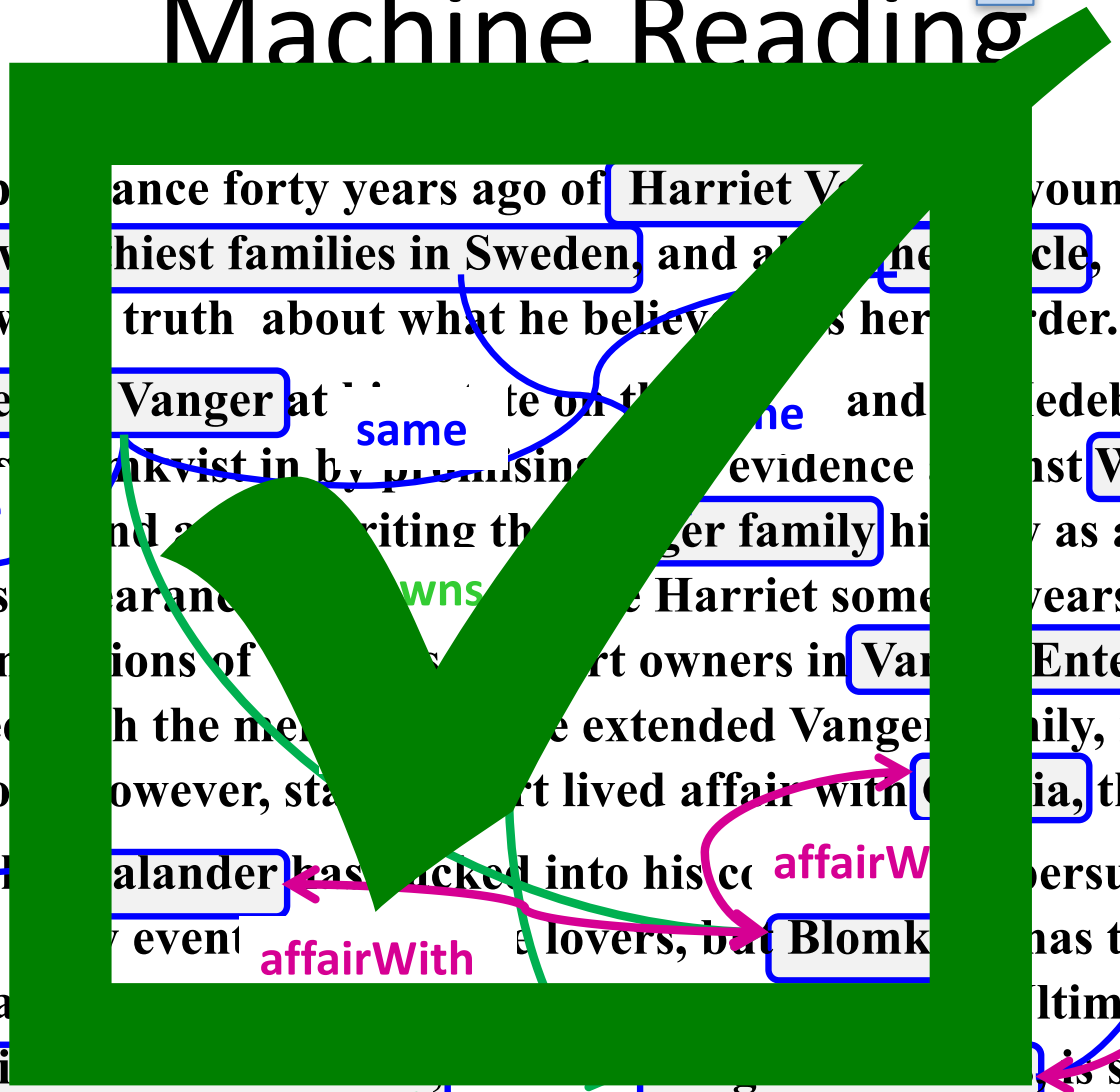
It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] te on th[illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist ag[illegible] spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible] inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.
Af[illegible] overing that Salander has hacked into his c[illegible] persuade[illegible] assist him with research. They even[illegible] lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.
A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible] gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
affairWith
affairWith
enemyOf
same
same
headOf

This slide was adapted from [Hady et al., 2011]

Machine Reading



It's about the disap ance forty years ago of Harriet V young scion of one of the v hiest families in Sweden, and a ne cle, determined to know truth about what he believ s her der.

Blomkvist visits He Vanger at te on th ne and edeby. The old man draws nkvist in by promising evidence nst Wennerström. Blomkvist ag nd a riting th er family hi as a cover for the real assignment: the dis arane wns Harriet some years earlier. Hedeby is home to several ger ions of rt owners in Van Enterprises. Blomkvist beco uncleOf inte h the me e extended Vange ily, most of whom resent his presence. He do owever, sta rt lived affair with ia, the niece of Henrik Af overing th alander has icked into his co persuade assist him with research. even lovers, bu Blomk has trouble getting close to Lisbeth who trea ltimately the two discover that Harri is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou gations for Dragan Armansky, who, in tu ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
affairWith
enemyOf
same
affairWith
headOf
same

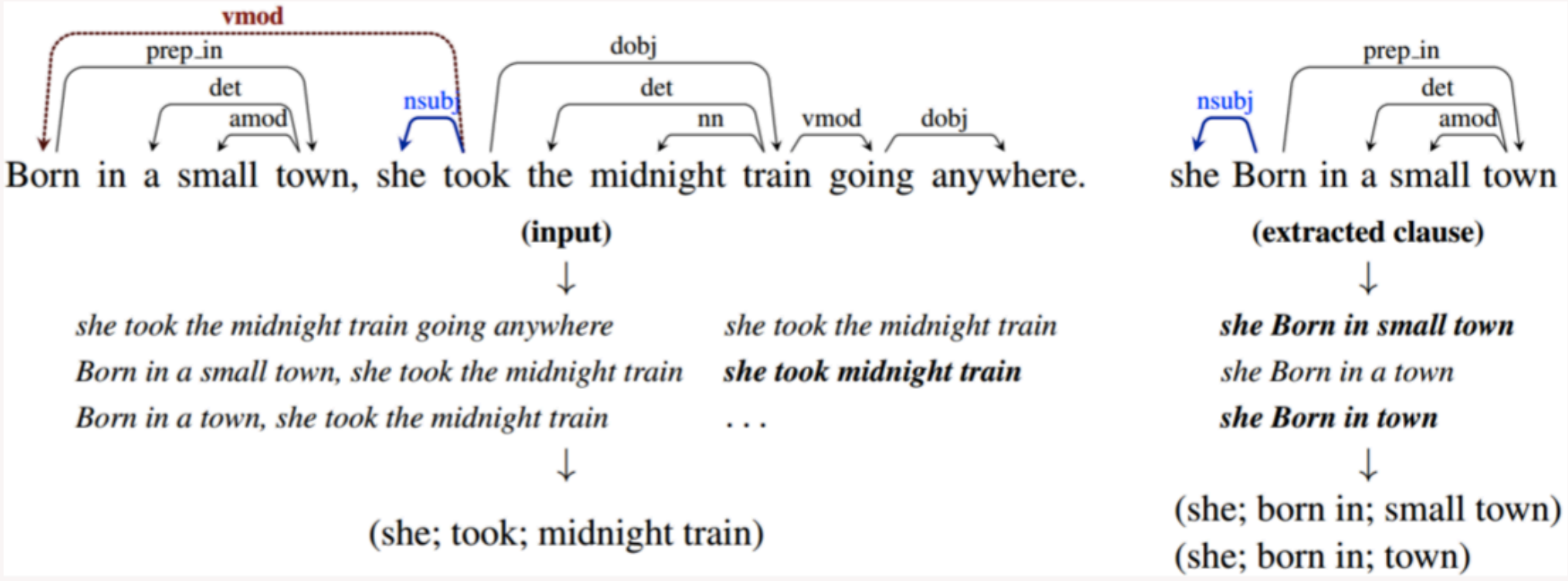
This slide was adapted from [Hady et al., 2011]

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

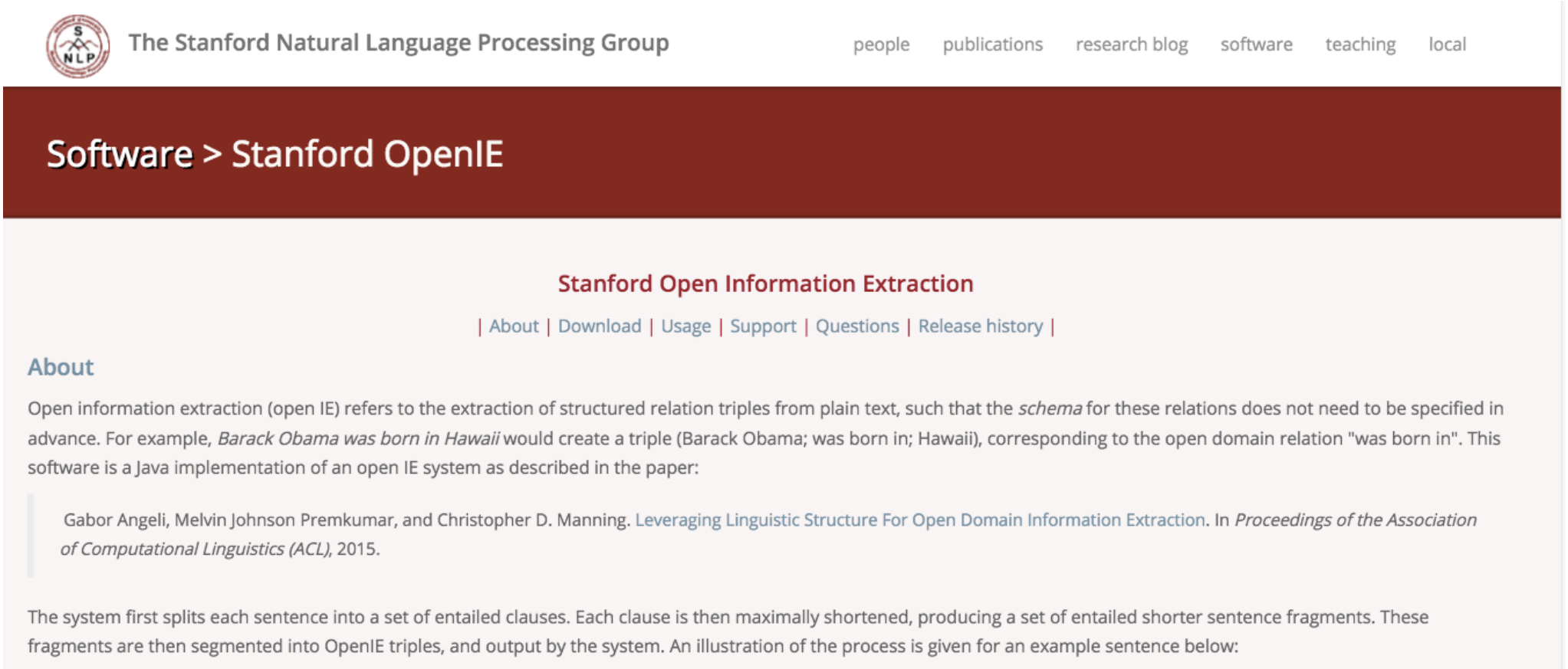
More on KnowItAll

- Open IE is now included in the Stanford CoreNLP package: http://nlp.stanford.edu/software/openie.shtml (ACL 2015 paper is linked to from this page).



More on KnowItAll

- Open IE is now included in the Stanford CoreNLP package: http://nlp.stanford.edu/software/openie.shtml (ACL 2015 paper is linked to from this page).



The Stanford Natural Language Processing Group

people publications research blog software teaching local

Software > Stanford OpenIE

Stanford Open Information Extraction

| About | Download | Usage | Support | Questions | Release history |

About

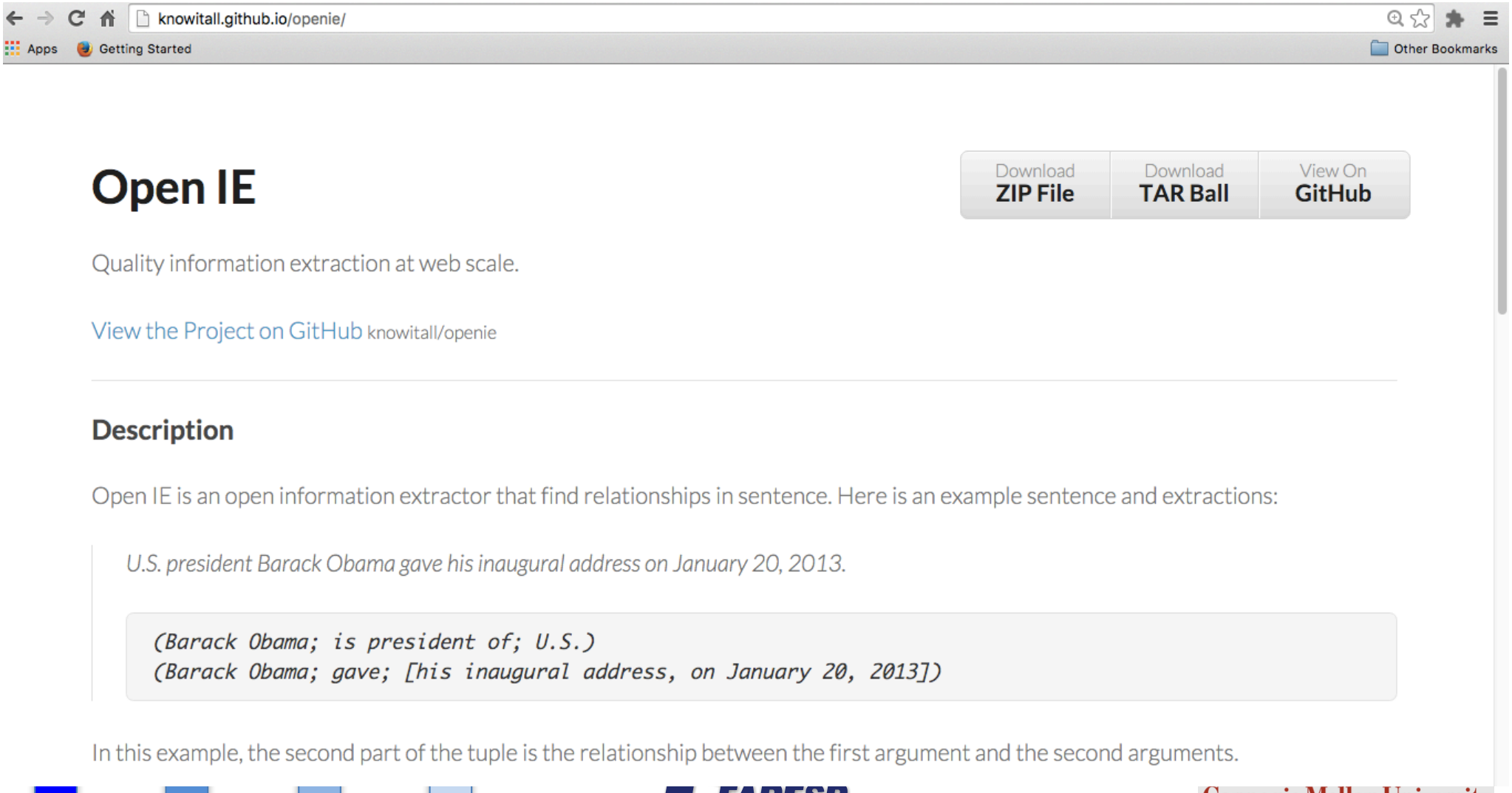
Open information extraction (open IE) refers to the extraction of structured relation triples from plain text, such that the *schema* for these relations does not need to be specified in advance. For example, *Barack Obama was born in Hawaii* would create a triple (Barack Obama; was born in; Hawaii), corresponding to the open domain relation "was born in". This software is a Java implementation of an open IE system as described in the paper:

> Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the Association of Computational Linguistics (ACL)*, 2015.

The system first splits each sentence into a set of entailed clauses. Each clause is then maximally shortened, producing a set of entailed shorter sentence fragments. These fragments are then segmented into OpenIE triples, and output by the system. An illustration of the process is given for an example sentence below:

More on KnowItAll

- Open IE 4.0: http://knowitall.github.io/openie/



More on KnowItAll

- Open IE in Chinese: https://www.semanticscholar.org/paper/19f83e125aa6384197

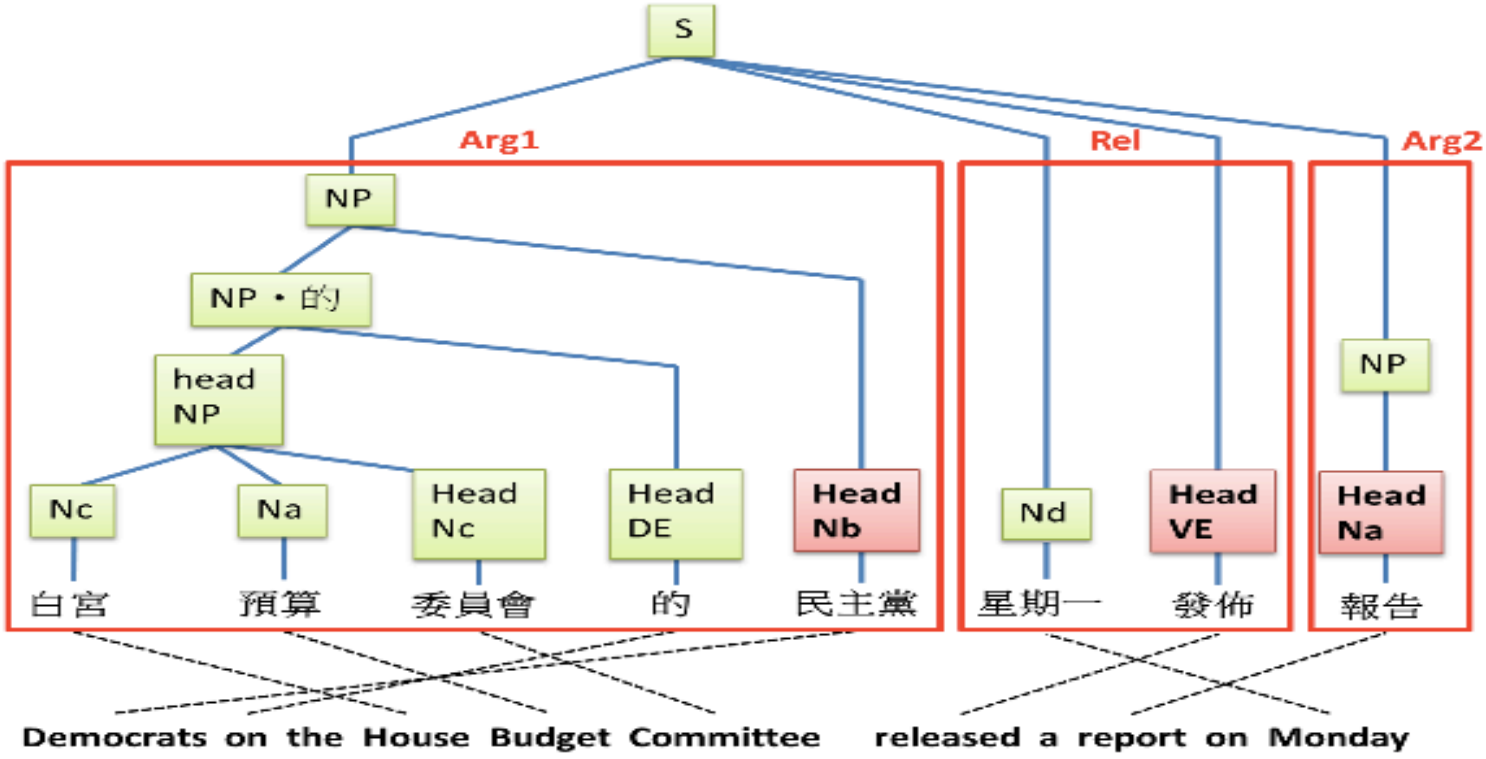


Figure 1: The parsed tree of a Chinese sentence.

More on KnowItAll

- Open IE in Chinese: https://www.semanticscholar.org/paper/19f83e125aa6384197b7a1fb138f0de531ab73eb

Chinese Open IE		Precision	Recall	F1
Exact Match	ReVerb	**0.5820**	0.0987	0.1688
	CORE	0.5579	**0.3291**	**0.4140**
Relation Only	ReVerb	0.8361	0.1425	0.2435
	CORE	**0.8463**	**0.5000**	**0.6286**

Table 1: Performance evaluation on Chinese Open IE.

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - NELL

Outline

- Machine Learning
- Machine Reading
- **Reading the Web**
  - DBPedia
  - YAGO
  - KnowItAll
  - **NELL**

Federal University of São Carlos

Never-Ending Language Learner

Joint work with Carnegie Mellon Read The Web Project Group
(http://rtw.ml.cmu.edu/rtw/)
and MaLL (Machine Learning Lab) from Federal University of São Carlos
(http://www.dc.ufscar.br/MaLL/MaLL.html)

FAPESP

Carnegie Mellon University

Humans learn many things, for years, and become better learners over time

Why not machines?

Never-Ending Learning

We'll never really understand learning until we build machines that

- learn many different things,
- over years,
- and become better learners over time.

Never-Ending Learning

We'll never produce natural language understanding systems until we have systems that react to arbitrary sentences by saying one of:

- I understand, and already knew that
- I understand, and didn't know, but accept it
- I understand, and disagree because …

NELL: Never-Ending Language Learner

- Inputs:
  - initial ontology
  - handful of examples of each predicate in ontology
  - the web
  - occasional interaction with human trainers

- The task:
  - run 24x7, forever
  - each day:
    1. extract more facts from the web to populate the initial ontology
    2. learn to read (perform #1) better than yesterday

NELL: Never-Ending Language Learner

- Goal:
- • run 24x7, forever
- • each day:
- 1. extract more facts from the web to populate given ontology
- 2. learn to read better than yesterday

- Today...
- Running 24 x 7, since January, 2010
- Input:
- • ontology defining ~800 categories and relations
- • 10-20 seed examples of each
- • 1 billion web pages (ClueWeb – Jamie Callan)

- Result:
- • continuously growing KB with +90.000,000 extracted beliefs (different levels of confidence)

http://rtw.ml.cmu.edu

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 15 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 1,471,011 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

E
Se

NELL: Never-Ending Language Learner

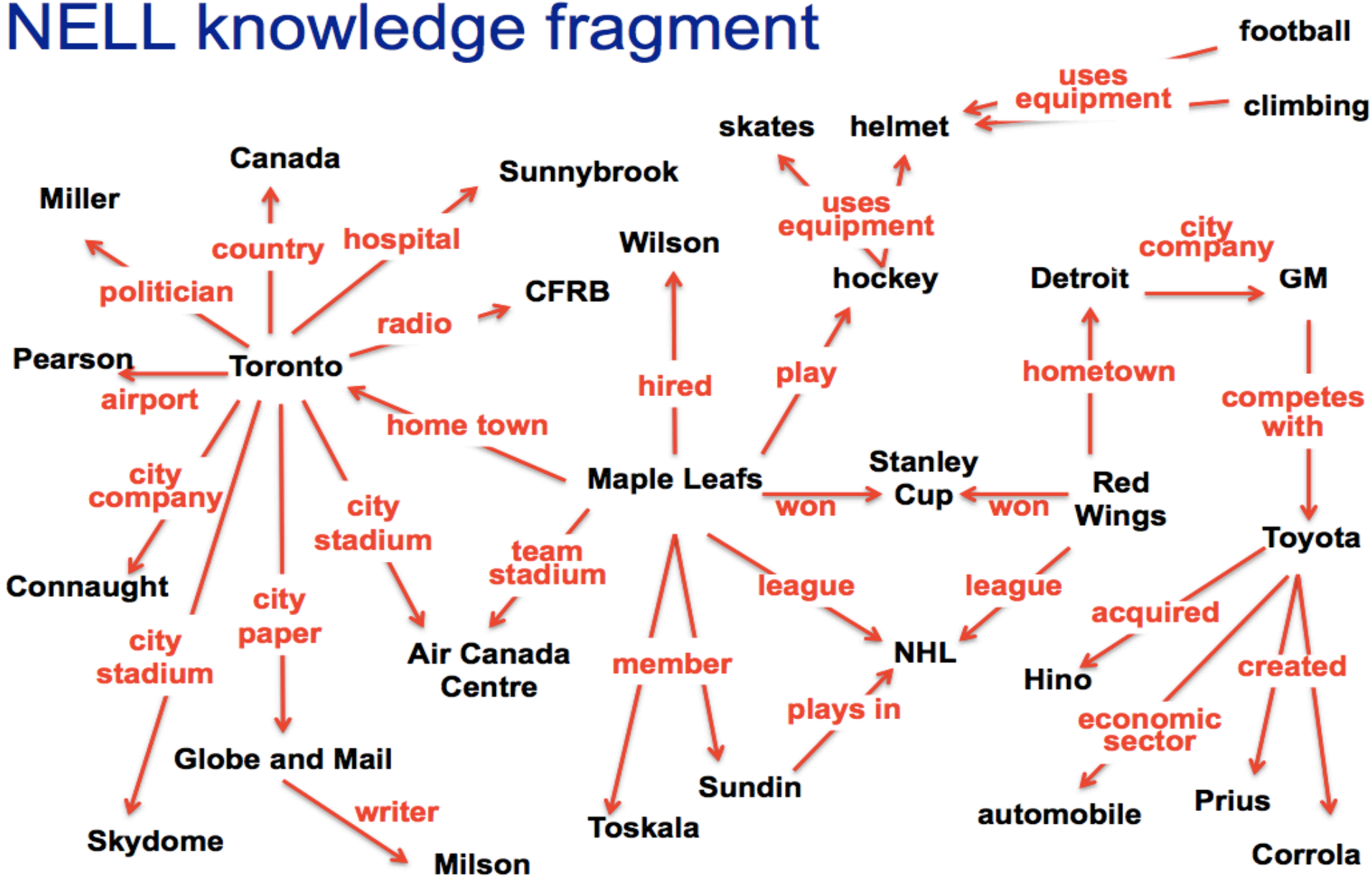
- http://rtw.ml.cmu.edu

Recently-Learned Facts

Refresh

instance	iteration	date learned	confidence
chase_bank is a credit union	989	10-apr-2016	100.0
unnamed_subpeak_of_long_valley_peak is a mountain	991	27-apr-2016	91.4
currency_funding is an economic sector	989	10-apr-2016	98.9
snow_mountains_grassland_melomys is a mammal	989	10-apr-2016	90.5
the_paramus_post is a blog	989	10-apr-2016	91.3
aashto is headquartered in the country u_s_	994	15-may-2016	100.0
division is an organization also known as us_department	992	02-may-2016	93.8
dome is an attraction made of bronze	994	15-may-2016	100.0
jeff_bezos is the CEO of amazon	994	15-may-2016	100.0
tomato is an agricultural product that is usually cooked with salt	991	27-apr-2016	99.6

NELL knowledge fragment



Learning to Read and Learning by Reading

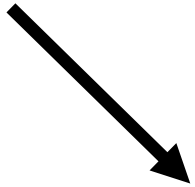
1. Classify noun phrases (NP's) by category

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino



mayor of arg1
live in arg1

The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

San Francisco
Austin
denial

mayor of arg1
live in arg1

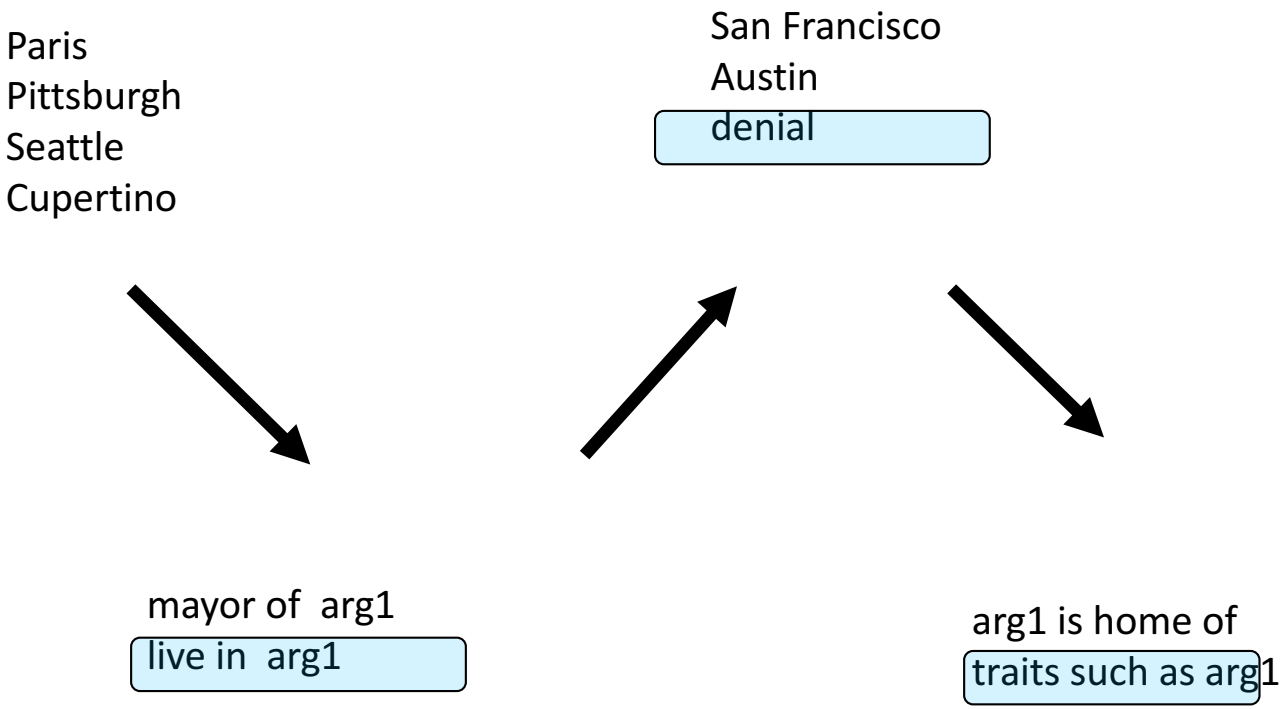
The Problem with Semi-Supervised Bootstrap Learning

Paris
Pittsburgh
Seattle
Cupertino

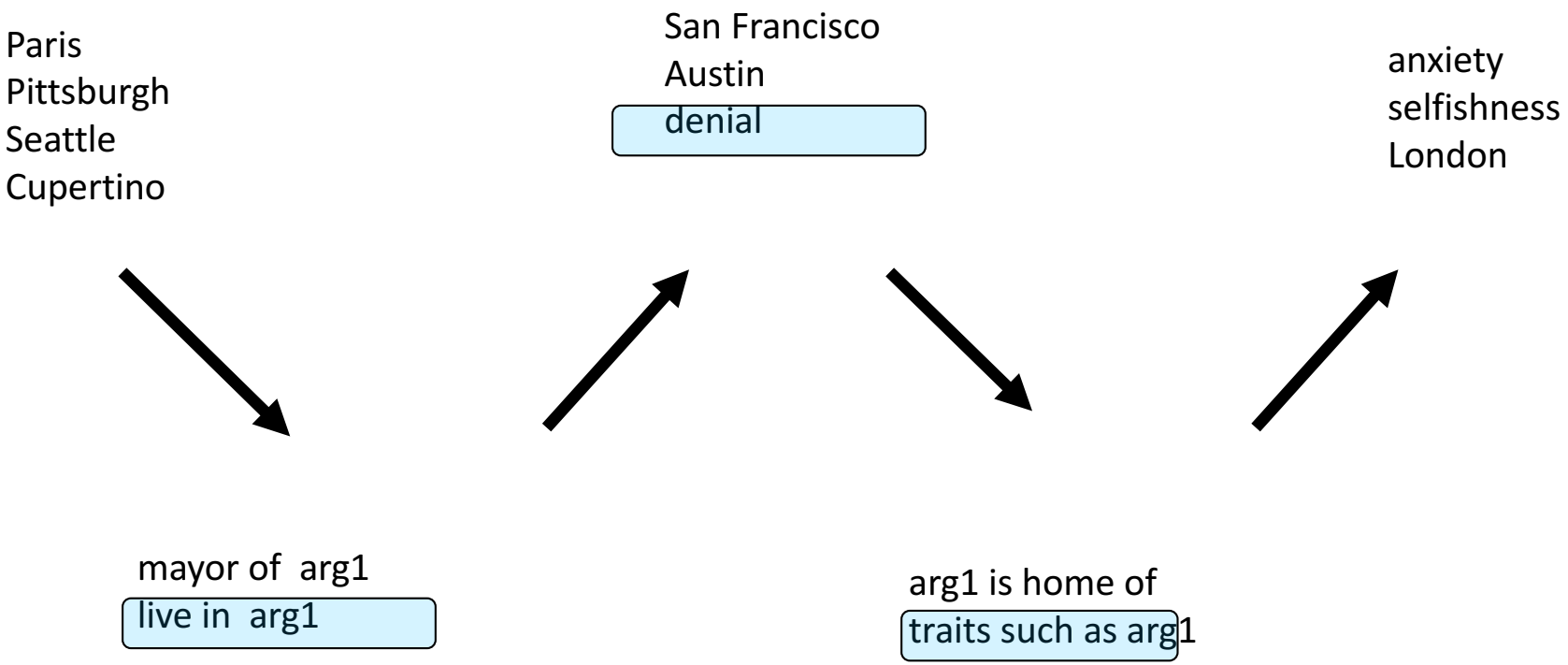
San Francisco
Austin
denial

mayor of arg1
live in arg1

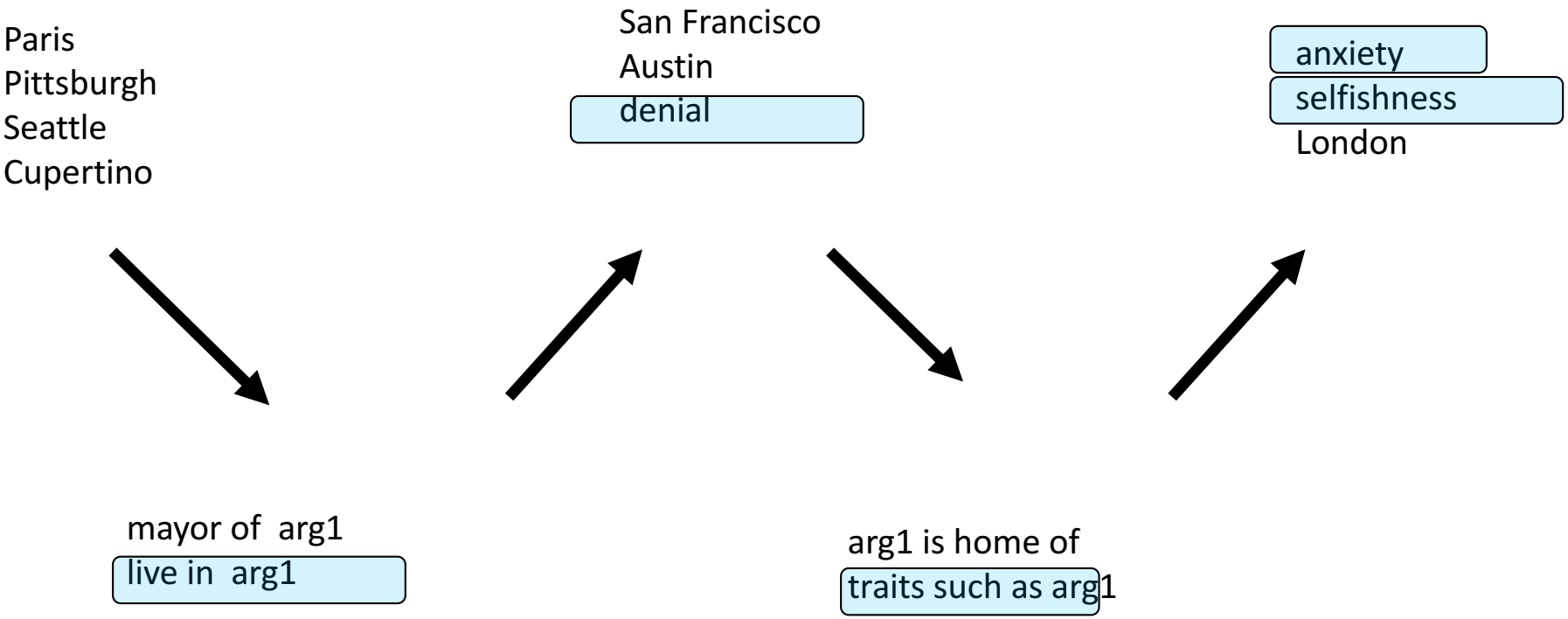
The Problem with Semi-Supervised Bootstrap Learning



The Problem with Semi-Supervised Bootstrap Learning



The Problem with Semi-Supervised Bootstrap Learning



The Problem with Semi-Supervised Bootstrap Learning

it's underconstrained!!

Paris
Pittsburgh
Seattle
Cupertino

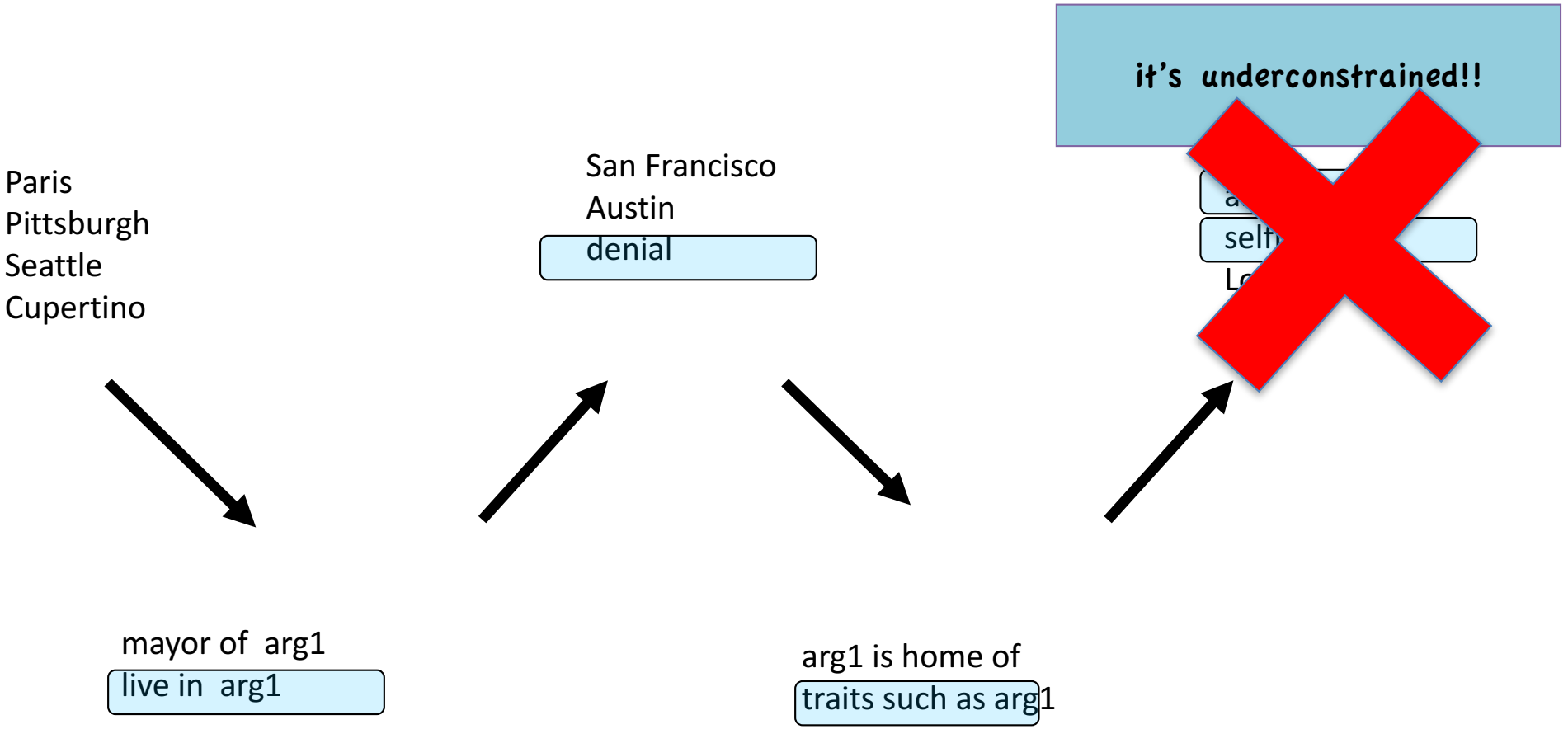
San Francisco
Austin
denial

anxiety
selfishness
London

mayor of arg1
live in arg1

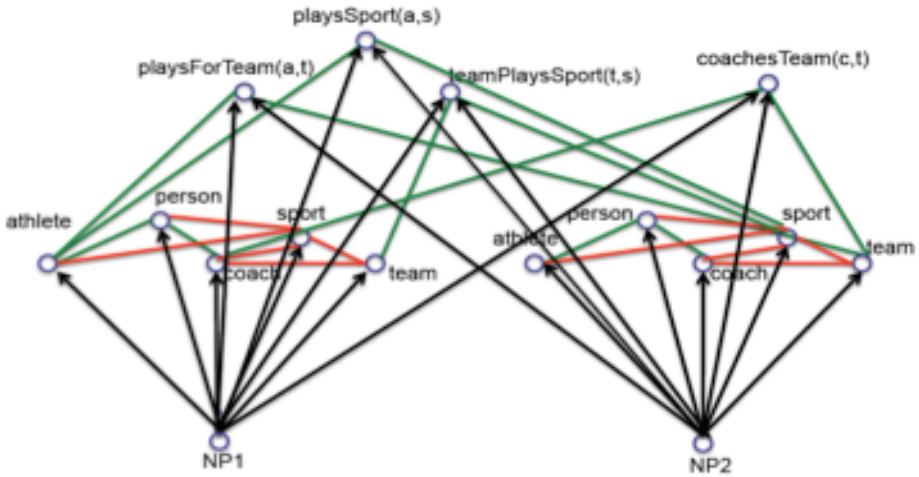
arg1 is home of
traits such as arg1

The Problem with Semi-Supervised Bootstrap Learning



Key Idea 1: Coupled semi-supervised training of many functions

hard
(underconstrained)
semi-supervised
learning problem



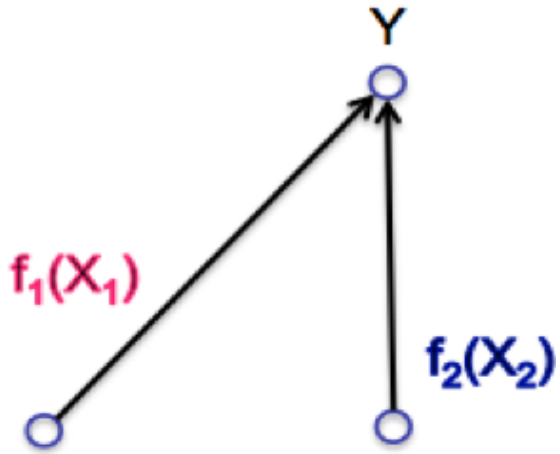
much easier (more constrained)
semi-supervised learning problem

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]

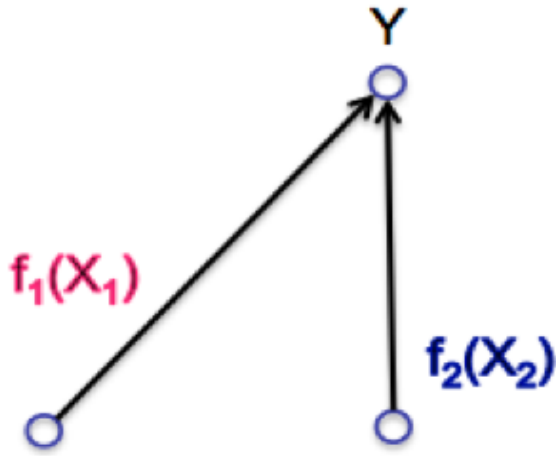


$\mathbf{X} = < X_1 , X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

Coupled Training Type 1: Co-training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]



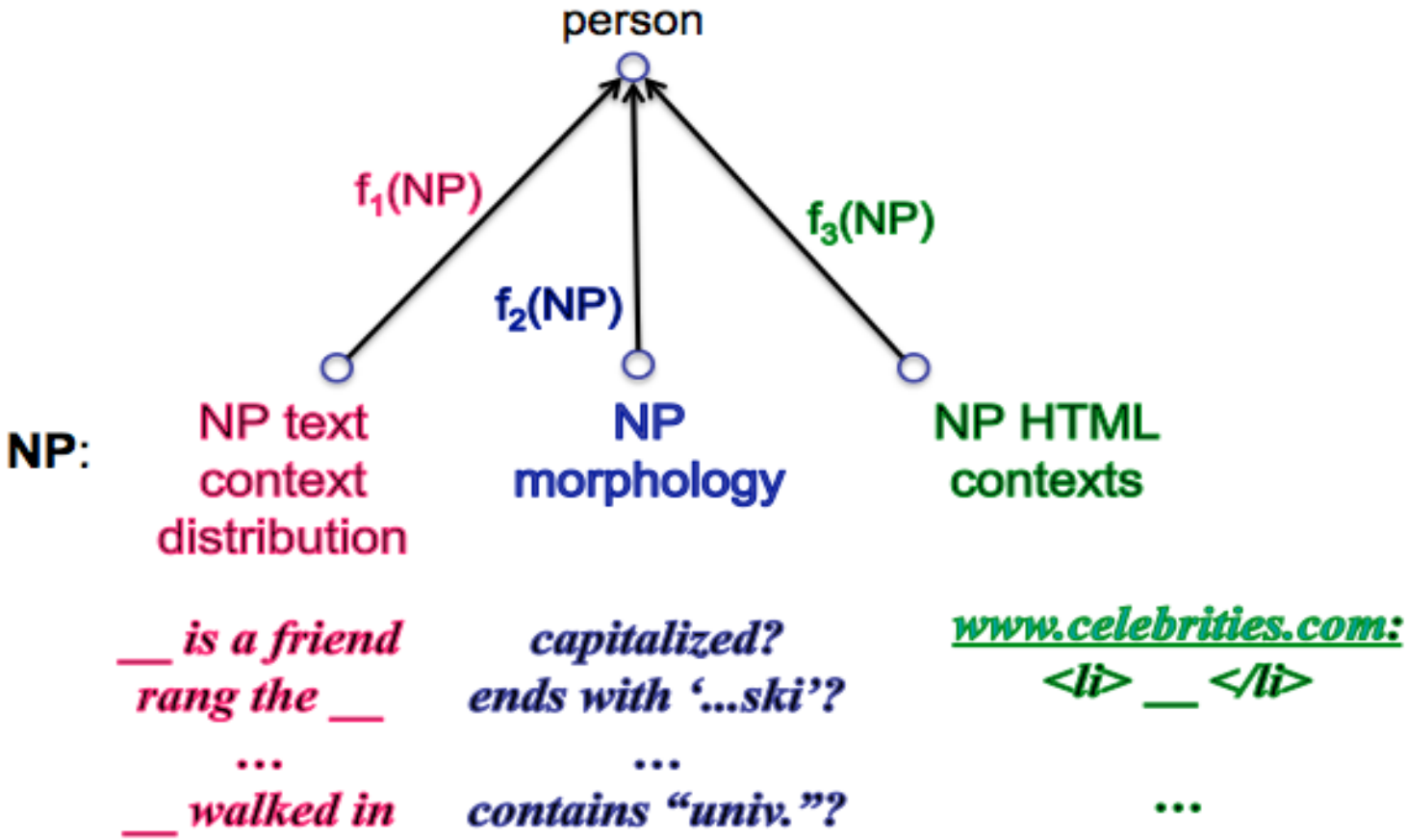
$\mathbf{X} = < X_1 , X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

If f_1, f_2 PAC learnable,
X_1, X_2 conditionally indep
Then PAC learnable from
unlabeled data and
weak initial learner

and disagreement between
f_1, f_2 bounds error of each

Type 1 Coupling Constraints in NELL



Coupled Training Type 2:

Structured Outputs, Multitask, Posterior Regularization, Multilabel

Learn functions with the same input, different outputs, where we know some constraint

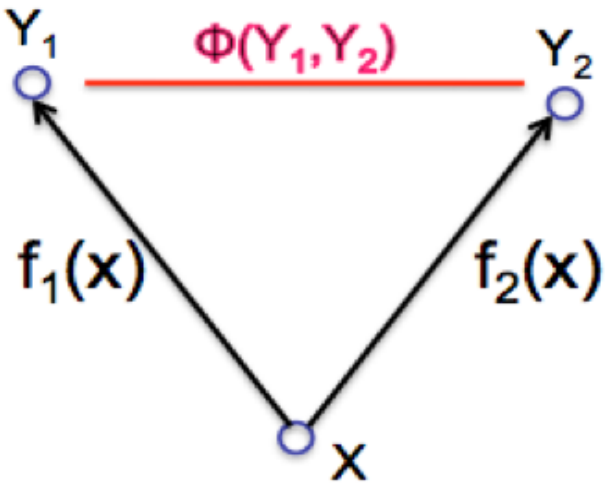
[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]

Coupled Training Type 2:

Structured Outputs, Multitask, Posterior Regularization, Multilabel

Learn functions with the same input, different outputs, where we know some constraint

[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]



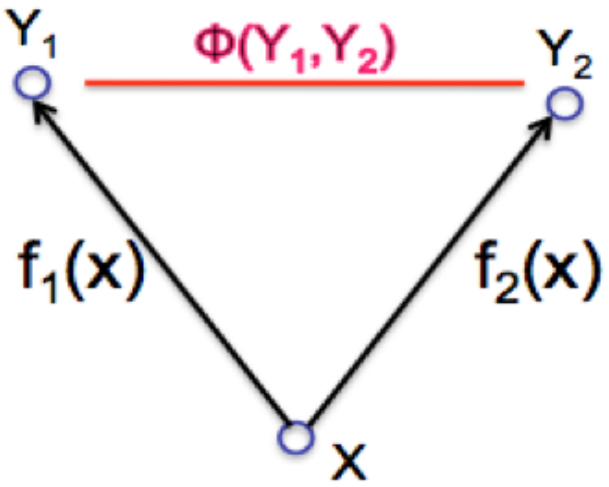
Constraint: $\Phi(f_1(x), f_2(x))$

Coupled Training Type 2:

Structured Outputs, Multitask, Posterior Regularization, Multilabel

Learn functions with the same input, different outputs, where we know some constraint

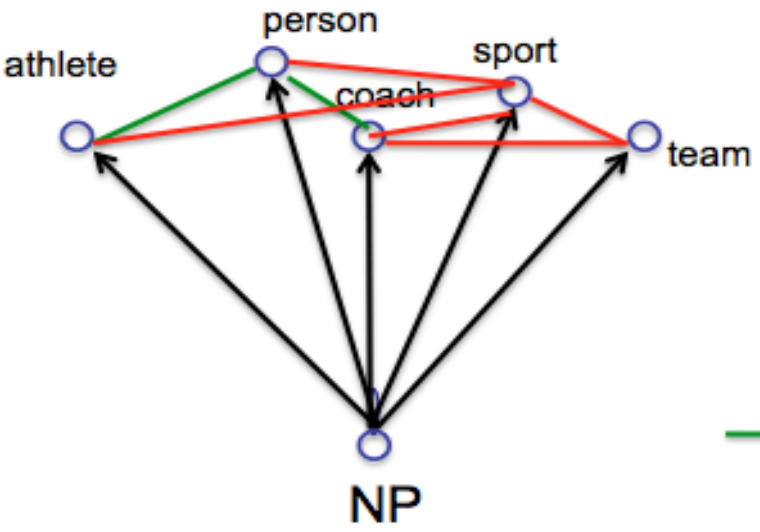
[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]



Effectiveness ~ probability that $\Phi(Y_1,Y_2)$ will be violated by incorrect f_j and f_k

Constraint: $\Phi(f_1(x), f_2(x))$

Type 2 Coupling Constraints in NELL

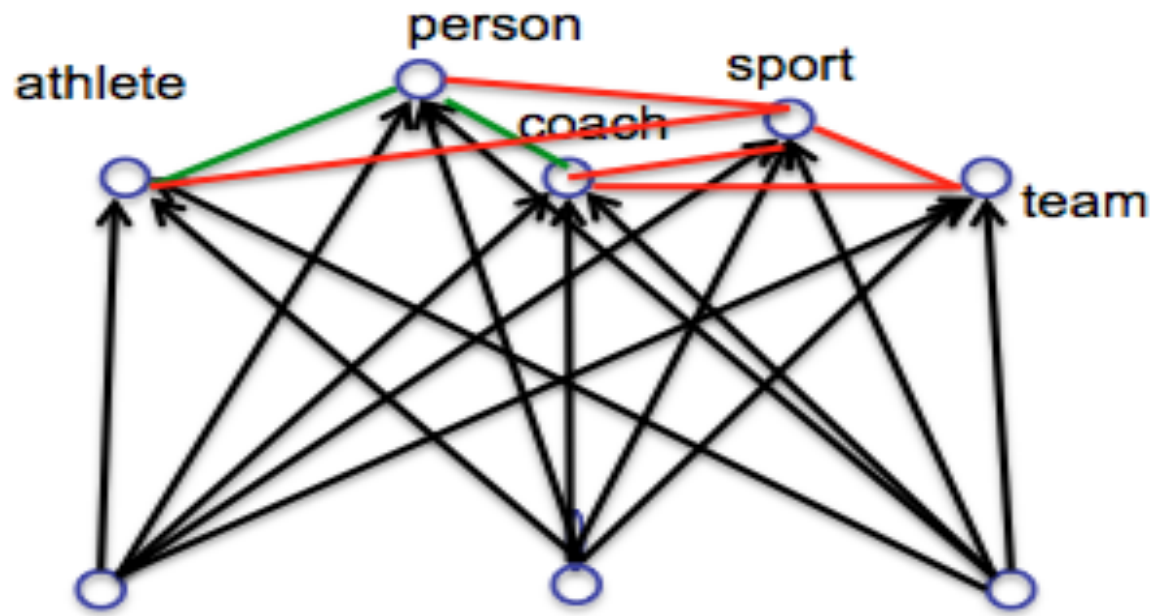


athlete(NP) → person(NP)

athlete(NP) → NOT sport(NP)

NOT athlete(NP) ← sport(NP)

Multi-view, Multi-Task Coupling



NP:

NP text context distribution

NP morphology

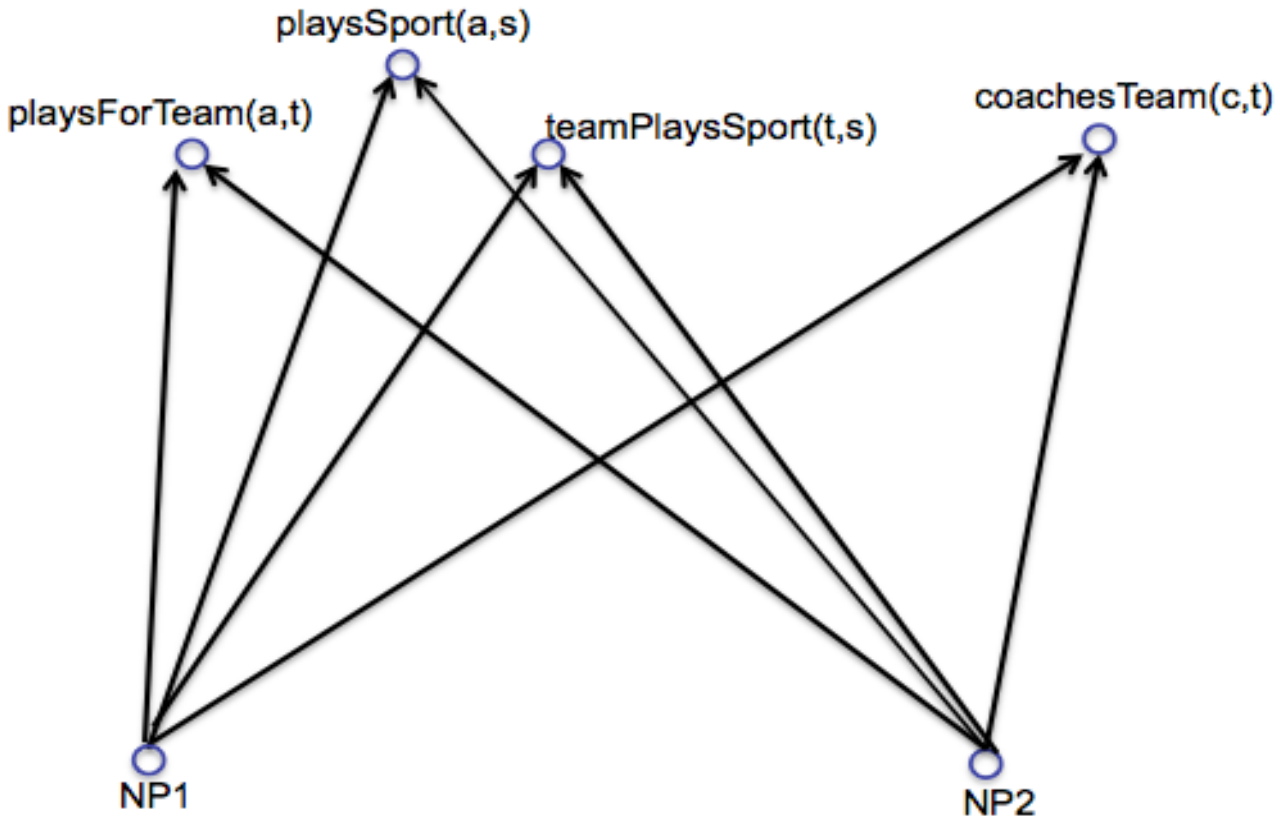
NP HTML contexts

over 2500 coupled functions in NELL

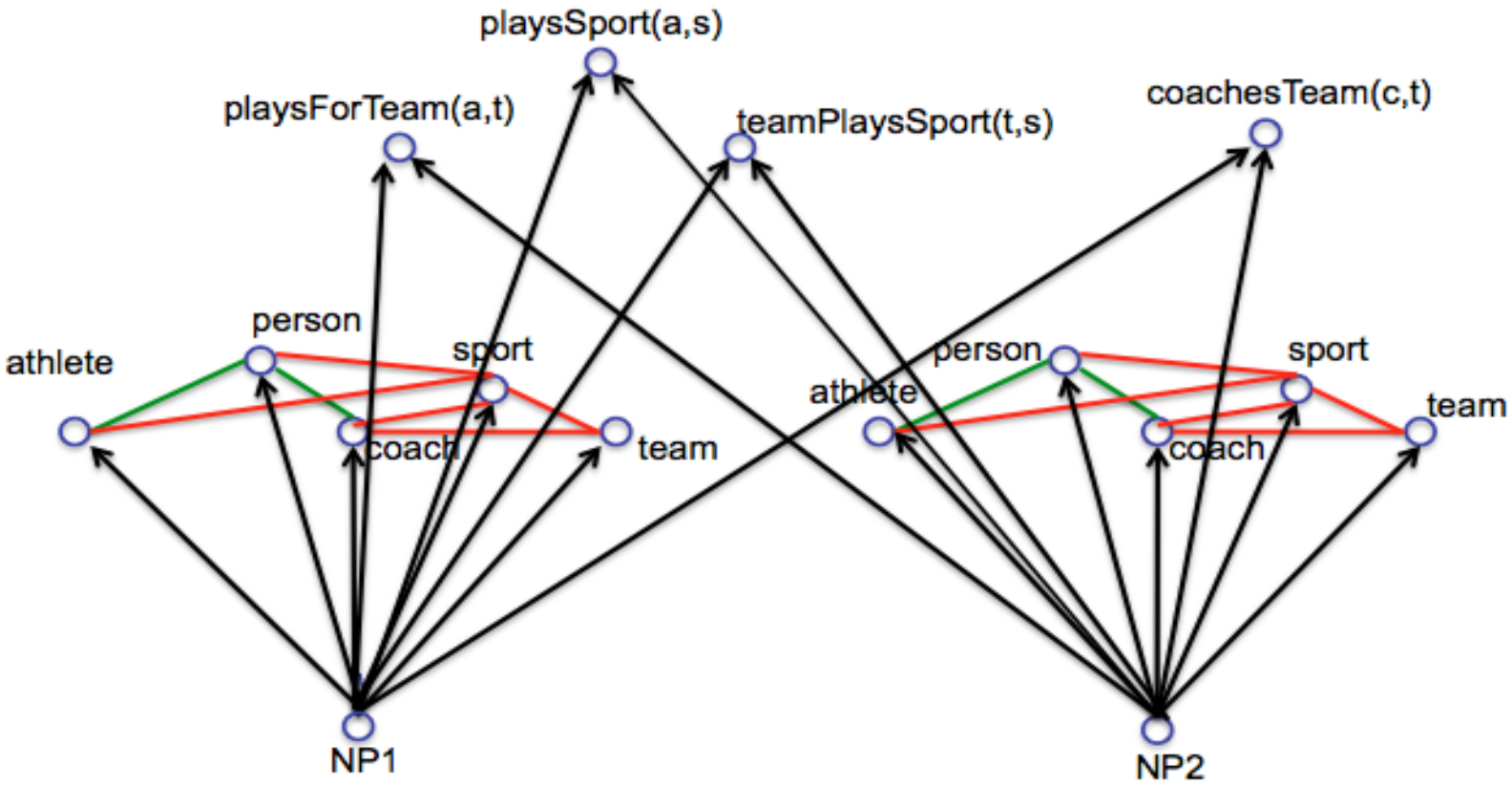
Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation

Learning Relations between NP's

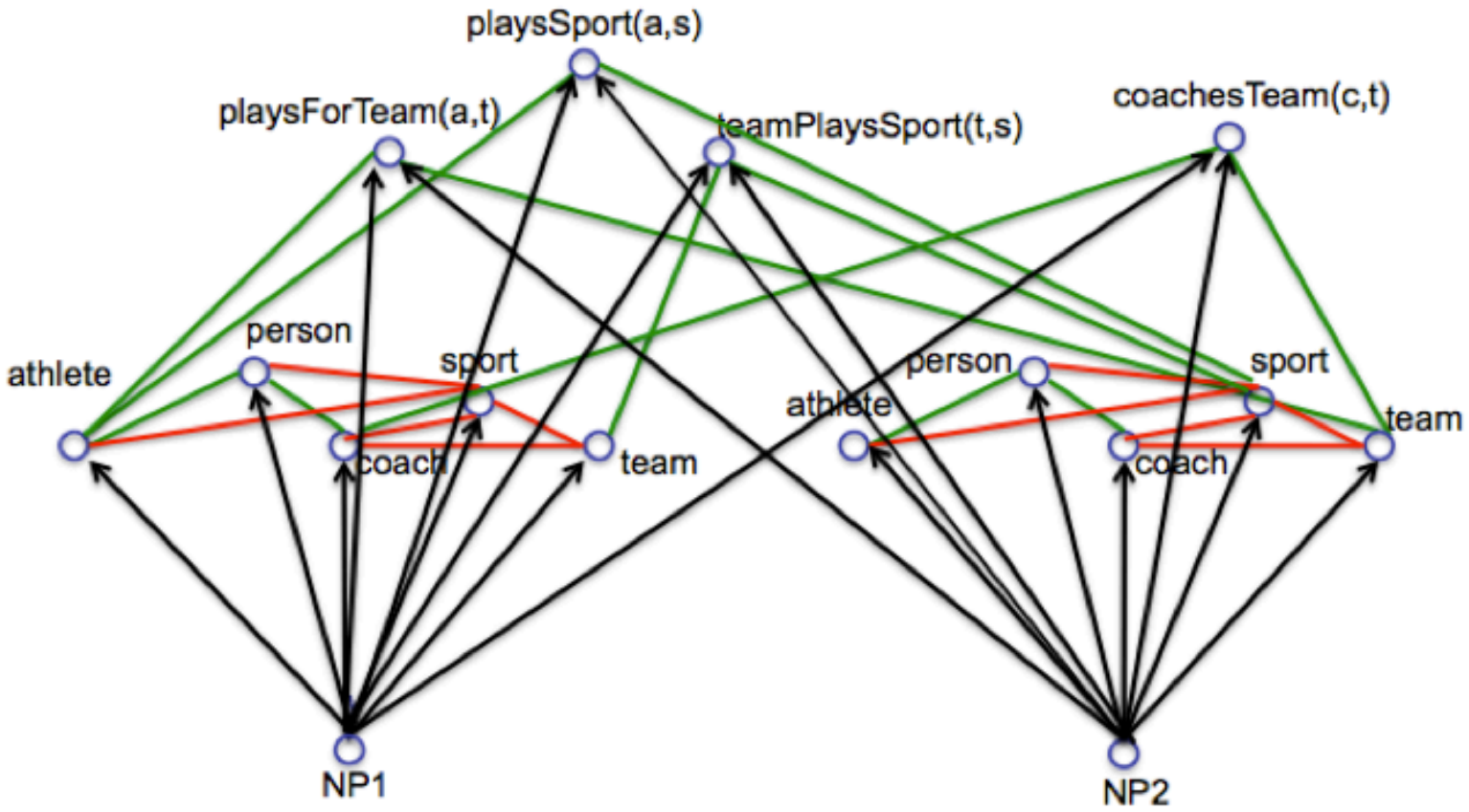


Learning Relations between NP's



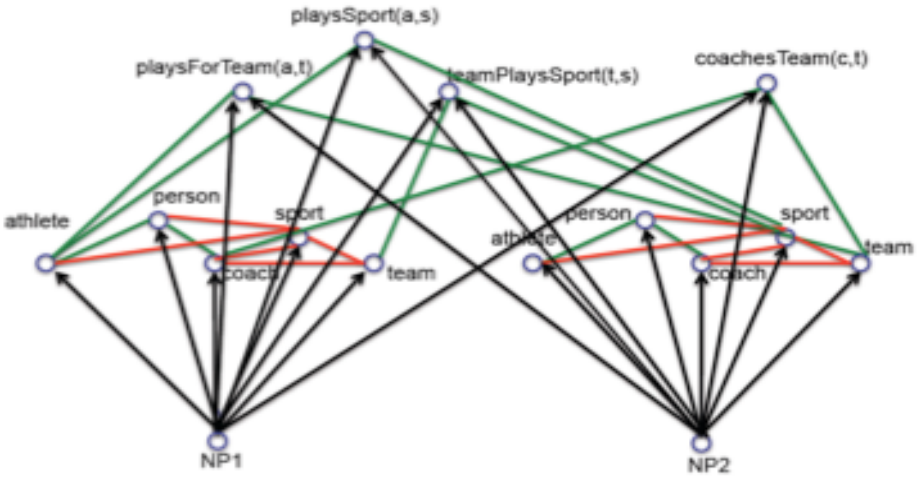
Type 3 Coupling: Argument Types

Constraint: f3(x1,x2) → (f1(x1) AND f2(x2))



playsSport(NP1,NP2) → athlete(NP1), sport(NP2)

Pure EM Approach to Coupled Training



- **E:** jointly estimate latent labels for each function of each unlabeled example
- **M:** retrain all functions, based on these probabilistic labels

Scaling problem:

- **E** step: 20M NP's, 10^{14} NP pairs to label
- **M** step: 50M text contexts to consider for each function → 10^{10} parameters to retrain
- even more URL-HTML contexts..

NELL's Approximation to EM

E' step:

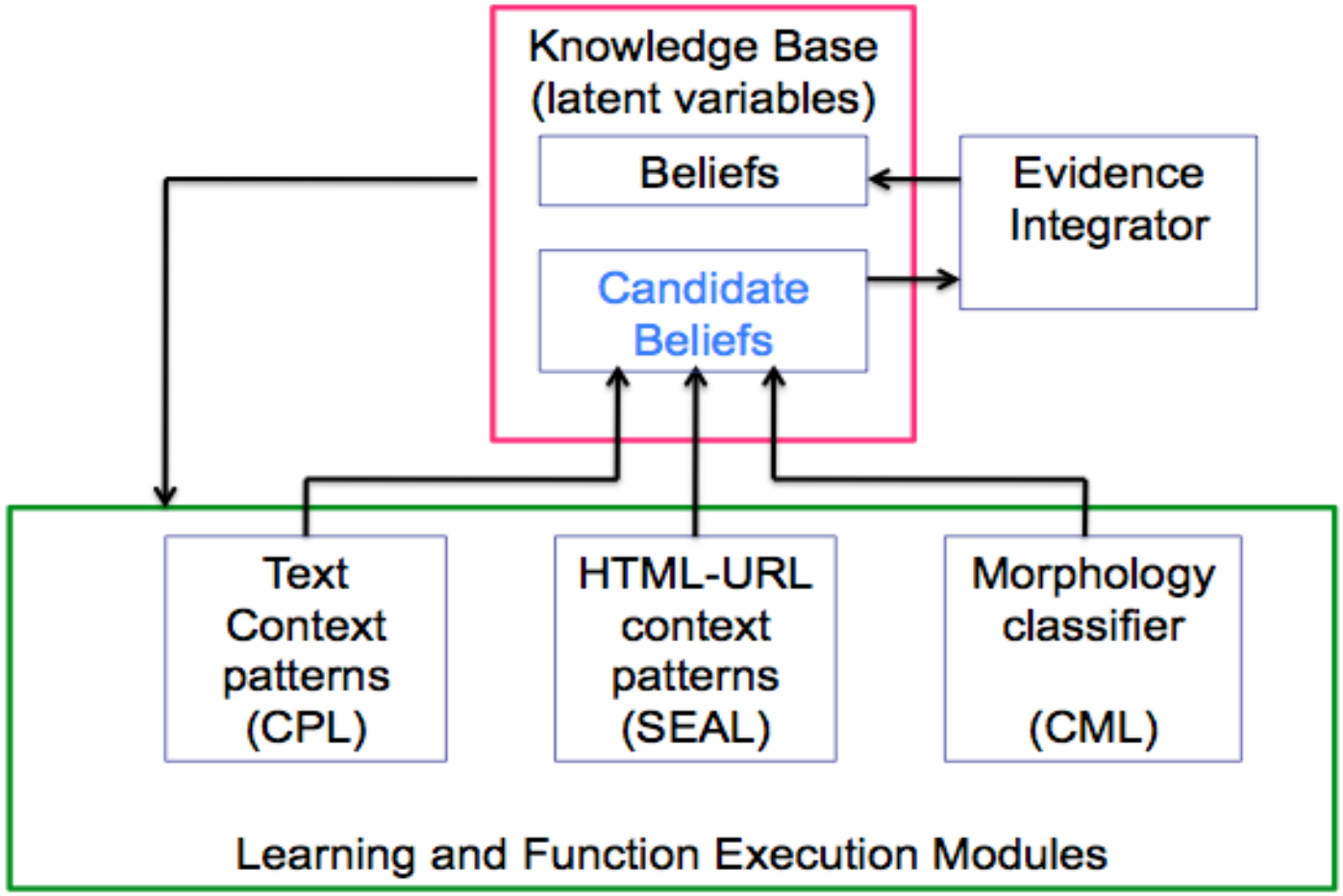
- Consider only a growing subset of the latent variable assignments

– category variables: up to 250 NP's per category per iteration
– relation variables: add only if confident and args of correct type
– this set of explicit latent assignments ***IS*** the knowledge base

M' step:

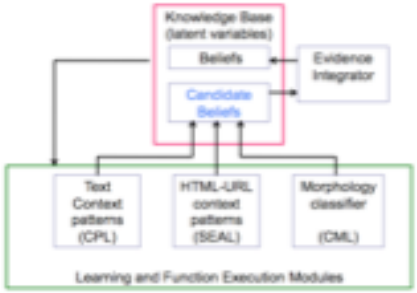
- Each view-based learner retrains itself from the updated KB
- "context" methods create growing subsets of contexts

NELL Architecture



Never-Ending Language Learning

arg1_was_playing_arg2 arg2_megastar_arg1 arg2_icons_arg1
arg2_player_named_arg1 arg2_prodigy_arg1
arg1_is_the_tiger_woods_of_arg2 arg2_career_of_arg1
arg2_greats_as_arg1 arg1_plays_arg2 arg2_player_is_arg1
arg2_legends_arg1 arg1_announced_his_retirement_from_arg2
arg2_operations_chief_arg1 arg2_player_like_arg1
arg2_and_golfing_personalities_including_arg1 arg2_players_like_arg1
arg2_greats_like_arg1 arg2_players_are_steffi_graf_and_arg1
arg2_great_arg1 arg2_champ_arg1 arg2_greats_such_as_arg1
arg2_professionals_such_as_arg1 arg2_hit_by_arg1 arg2_greats_arg1
arg2_icon_arg1 arg2_stars_like_arg1 arg2_pros_like_arg1
arg1_retires_from_arg2 arg2_phenom_arg1 arg2_lesson_from_arg1
arg2_architects_robert_trent_jones_and_arg1 arg2_sensation_arg1
arg2_pros_arg1 arg2_stars_venus_and_arg1 arg2_hall_of_famer_arg1
arg2_superstar_arg1 arg2_legend_arg1 arg2_legends_such_as_arg1
arg2_players_is_arg1 arg2_pro_arg1 arg2_player_was_arg1
arg2_god_arg1 arg2_idol_arg1 arg1_was_born_to_play_arg2
arg2_star_arg1 arg2_hero_arg1 arg2_players_are_arg1
arg1_retired_from_professional_arg2 arg2_legends_as_arg1
arg2_autographed_by_arg1 arg2_champion_arg1

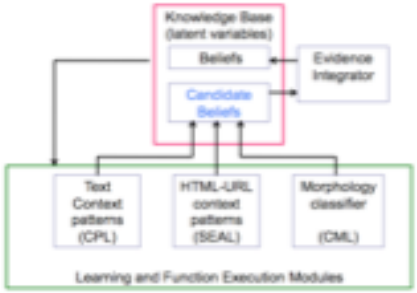


Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/ScholSearch.Asp	` [X] -`
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	`<a href='d_author.aspx?a=[X]'>-`
bird	http://www.michaelforsberg.com/stock.html	`<option>[X]</option>`
bookAuthor	http://lifebehindthecurve.com/	`</li> <li>[X] by [Y] &#8211;`

Never-Ending Language Learning

arg1_was_playing_arg2 arg2_megastar_arg1 arg2_icons_arg1
arg2_player_named_arg1 arg2_prodigy_arg1
arg1_is_the_tiger_woods_of_arg2 arg2_career_of_arg1
arg2_greats_as_arg1 arg1_plays_arg2 arg2_player_is_arg1
arg2_legends_arg1 arg1_announced_his_retirement_from_arg2
arg2_operations_chief_arg1 arg2_player_like_arg1
arg2_and_golfing_personalities_including_arg1 arg2_players_like_arg1
arg2_greats_like_arg1 arg2_players_are_steffi_graf_and_arg1
arg2_great_arg1 arg2_champ_arg1 arg2_greats_such_as_arg1
arg2_professionals_such_as_arg1 arg2_hit_by_arg1 arg2_greats_arg1
arg2_icon_arg1 arg2_stars_like_arg1 arg2_pros_like_arg1
arg1_retires_from_arg2 arg2_phenom_arg1 arg2_lesson_from_arg1
arg2_architects_robert_trent_jones_and_arg1 arg2_sensation_arg1
arg2_pros_arg1 arg2_stars_venus_and_arg1 arg2_hall_of_famer_arg1
arg2_superstar_arg1 arg2_legend_arg1 arg2_legends_such_as_arg1
arg2_players_is_arg1 arg2_pro_arg1 arg2_player_was_arg1
arg2_god_arg1 arg2_idol_arg1 arg1_was_born_to_play_arg2
arg2_star_arg1 arg2_hero_arg1 arg2_players_are_arg1
arg1_retired_from_professional_arg2 arg2_legends_as_arg1
arg2_autographed_by_arg1 arg2_champion_arg1

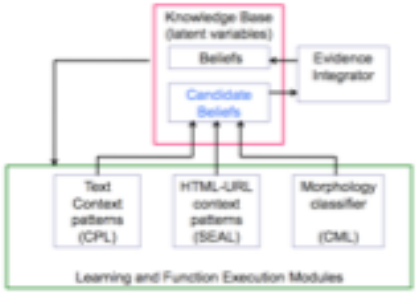


Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/ScholSearch.Asp	` [X] -`
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	`<a href='d_author.aspx?a=[X]'>-`
bird	http://www.michaelforsberg.com/stock.html	`<option>[X]</option>`
bookAuthor	http://lifebehindthecurve.com/	`</li> <li>[X] by [Y] &#8211;`

Never-Ending Language Learning

arg1_was_playing_arg2 arg2_megastar_arg1 arg2_icons_arg1
arg2_player_named_arg1 arg2_prodigy_arg1
arg1_is_the_tiger_woods_of_arg2 arg2_career_of_arg1
arg2_greats_as_arg1 arg1_plays_arg2 arg2_player_is_arg1
arg2_legends_arg1 arg1_announced_his_retirement_from_arg2
arg2_operations_chief_arg1 arg2_player_like_arg1
arg2_and_golfing_personalities_including_arg1 arg2_players_like_arg1
arg2_greats_like_arg1 arg2_players_are_steffi_graf_and_arg1
arg2_great_arg1 arg2_champ_arg1 arg2_greats_such_as_arg1
arg2_professionals_such_as_arg1 arg2_hit_by_arg1 arg2_greats_arg1
arg2_icon_arg1 arg2_stars_like_arg1 arg2_pros_like_arg1
arg1_retires_from_arg2 arg2_phenom_arg1 arg2_lesson_from_arg1
arg2_architects_robert_trent_jones_and_arg1 arg2_sensation_arg1
arg2_pros_arg1 arg2_stars_venus_and_arg1 arg2_hall_of_famer_arg1
arg2_superstar_arg1 arg2_legend_arg1 arg2_legends_such_as_arg1
arg2_players_is_arg1 arg2_pro_arg1 arg2_player_was_arg1
arg2_god_arg1 arg2_idol_arg1 arg1_was_born_to_play_arg2
arg2_star_arg1 arg2_hero_arg1 arg2_players_are_arg1
arg1_retired_from_professional_arg2 arg2_legends_as_arg1
arg2_autographed_by_arg1 arg2_champion_arg1



Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/ScholSearch.Asp	` [X] -`
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	`<a href='d_author.aspx?a=[X]'>-`
bird	http://www.michaelforsberg.com/stock.html	`<option>[X]</option>`
bookAuthor	http://lifebehindthecurve.com/	`</li> <li>[X] by [Y] &#8211;`

If coupled learning is the key idea, how can we get new coupling constraints?

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances

Key Idea 2: Discover New Coupling Constraints

- • first order, probabilistic horn clause constraints

- 0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z), teamPlaysSport(?z,?y)

- – connects previously uncoupled relation predicates
- – infers new beliefs for KB

Example Learned Horn Clauses

- 0.95 athletePlaysSport(?x,basketball) :- athleteInLeague(?x,NBA)

- 0.93 athletePlaysSport(?x,?y) :- athletePlaysForTeam(?x,?z)
- teamPlaysSport(?z,?y)

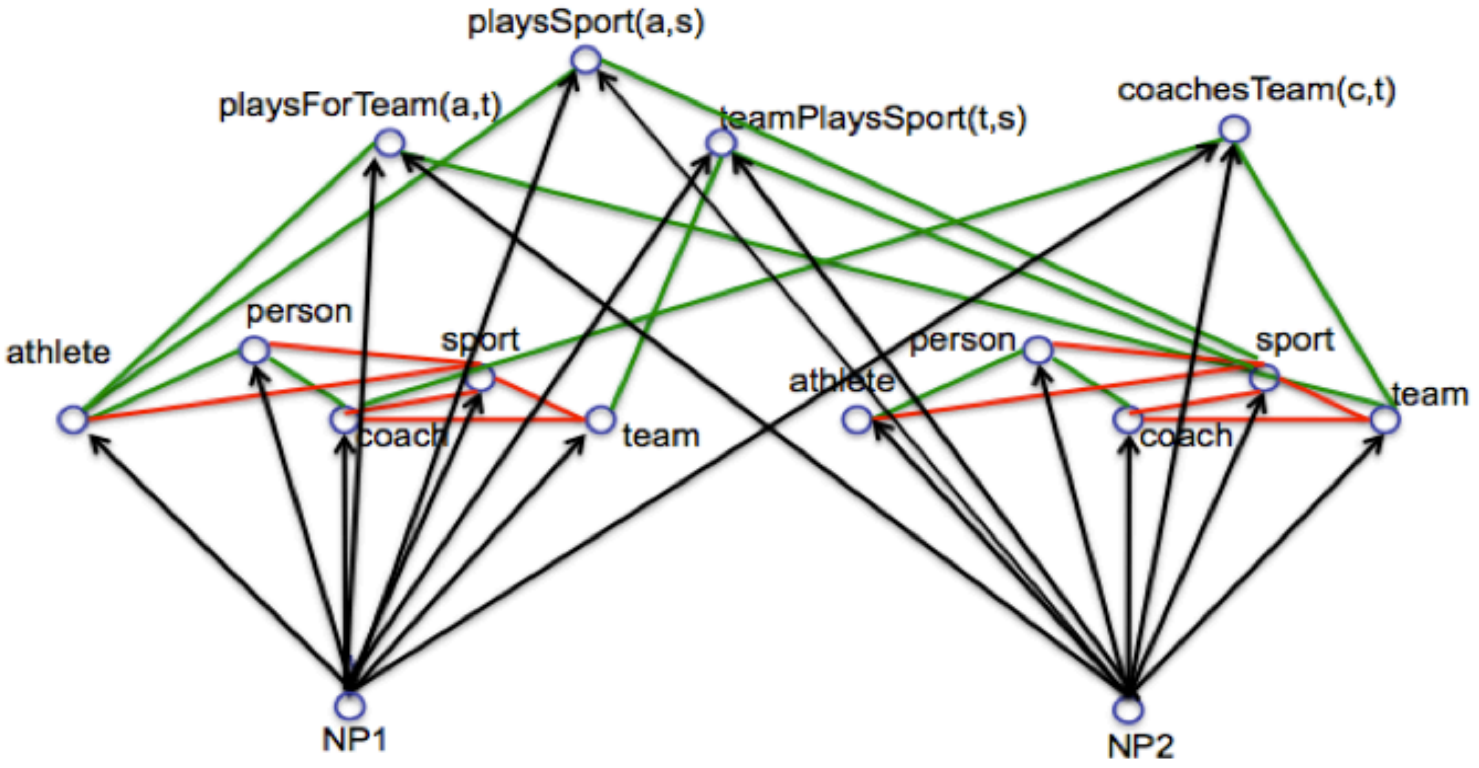
- 0.91 teamPlaysInLeague(?x,NHL) :- teamWonTrophy(?x,Stanley_Cup)

- 0.90 athleteInLeague(?x,?y):- athletePlaysForTeam(?x,?z),
  teamPlaysInLeague(?z,?y)

- 0.88 cityInState(?x,?y) :- cityCapitalOfState(?x,?y),
- cityInCountry(?y,USA)

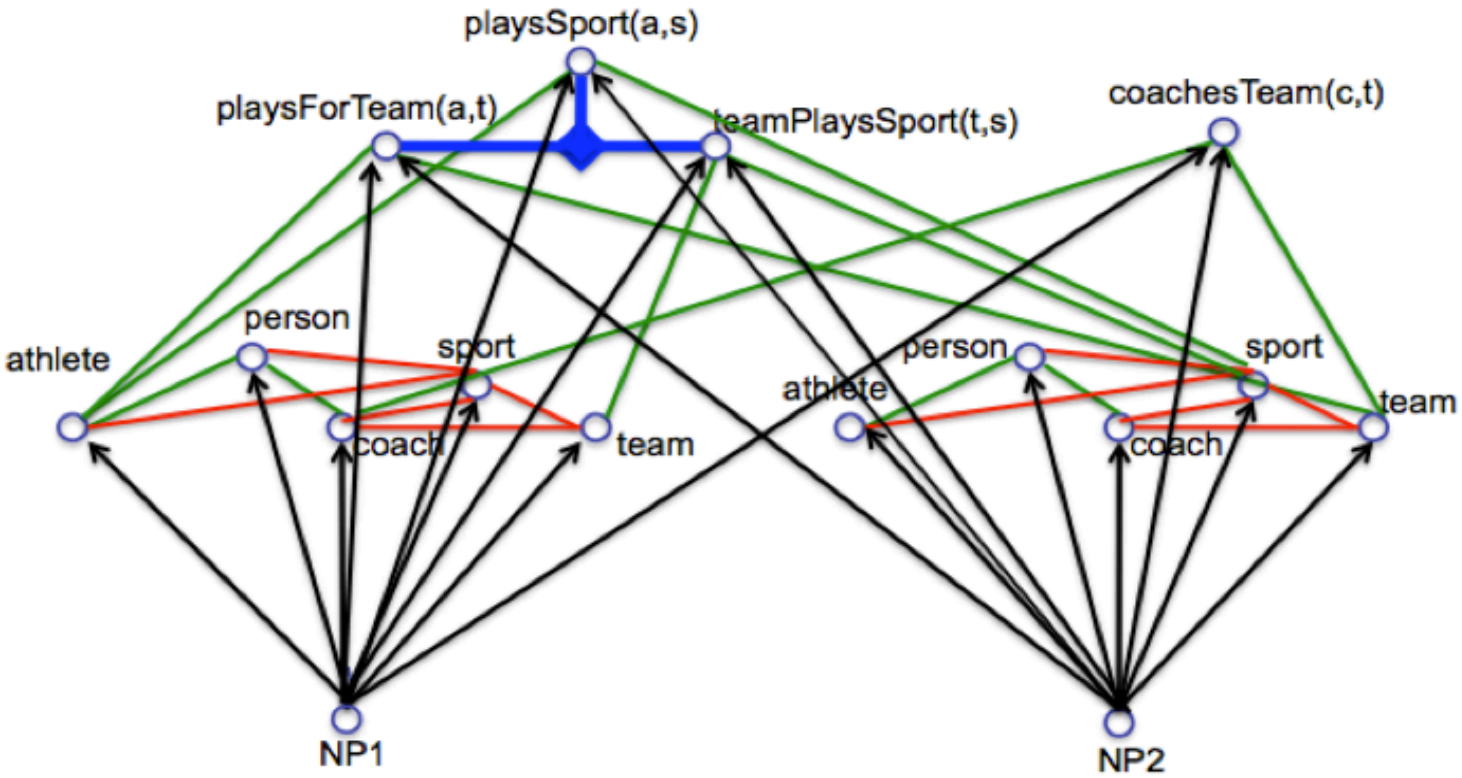
- 0.62* newspaperInCity(?x,New_York) :- companyEconomicSector(?x,media),
  generalizations(?x,blog)

Learned Probabilistic Horn Clause Rules



Learned Probabilistic Horn Clause Rules

0.93 playsSport(?x,?y) ← playsForTeam(?x,?z), teamPlaysSport(?z,?y)



NELL Architecture

Knowledge Base
(latent variables)

Beliefs

Candidate Beliefs

Evidence Integrator

Text Context patterns (CPL)

HTML-URL context patterns (SEAL)

Morphology classifier (CML)

Rule Learner (RL)

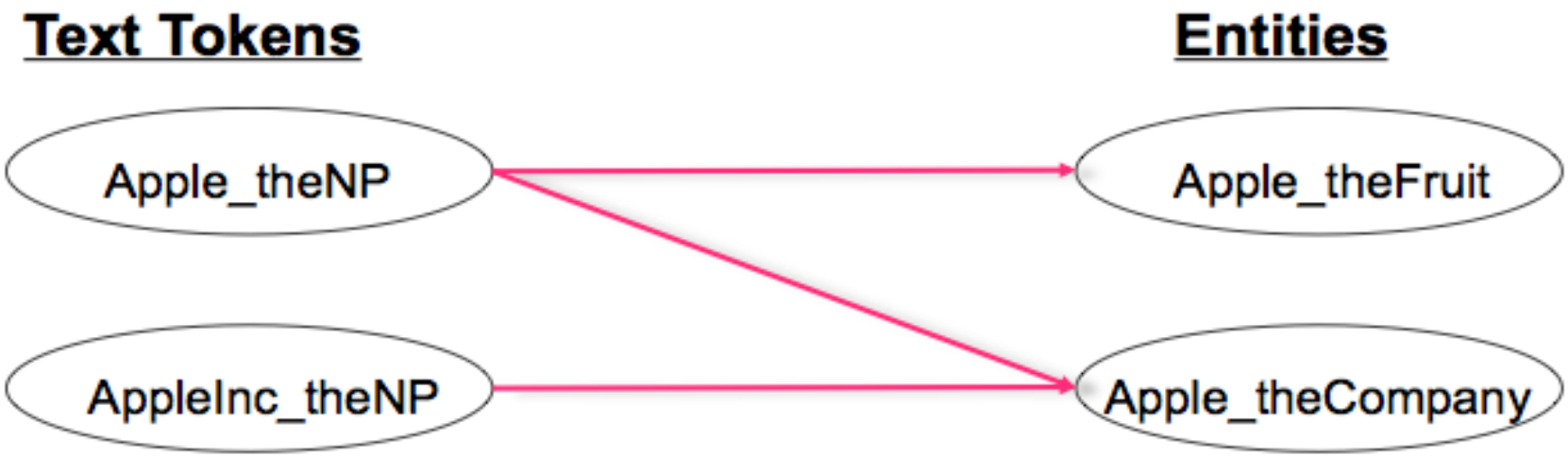
Learning and Function Execution Modules

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts

Distinguish Text Tokens from Entities

[Jayant Krishnamurthy]



Coreference Resolution:

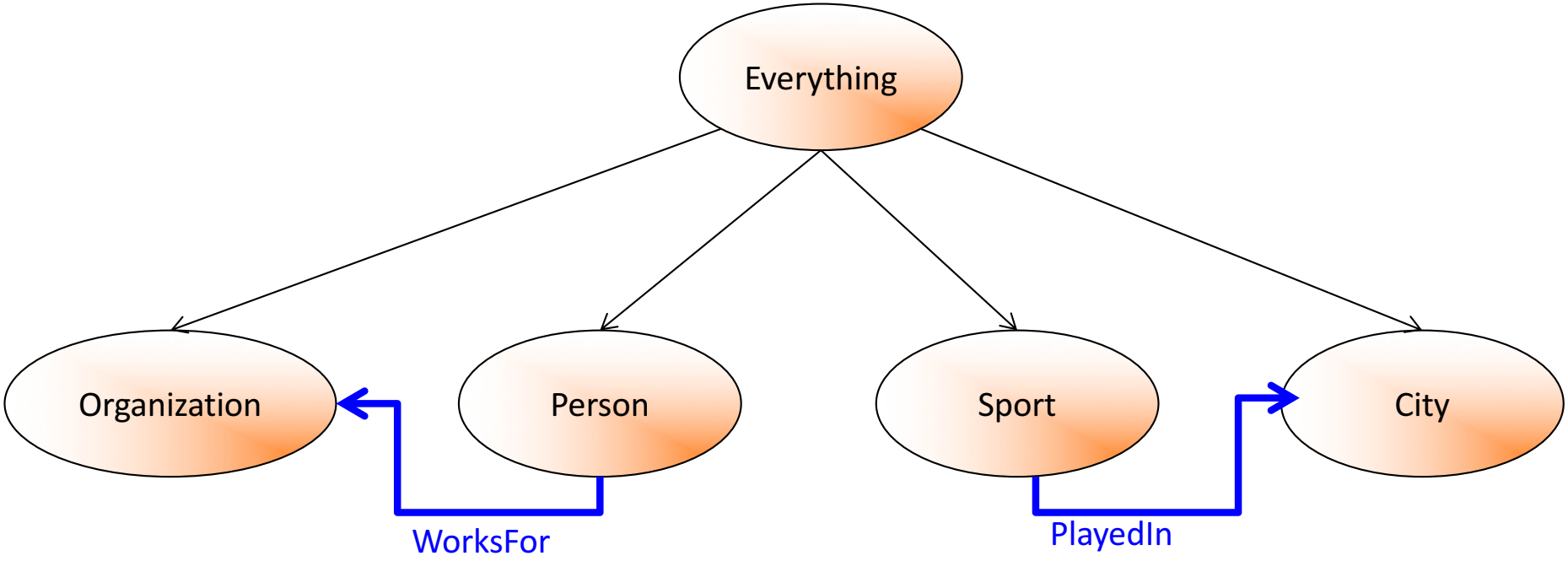
- Co-train classifier to predict coreference as f(string similarity, extracted beliefs)
- Small amount of supervision: ~10 labeled coreference decisions
- Cluster tokens using f as similarity measure

Building the Knowledge Graph by Reading

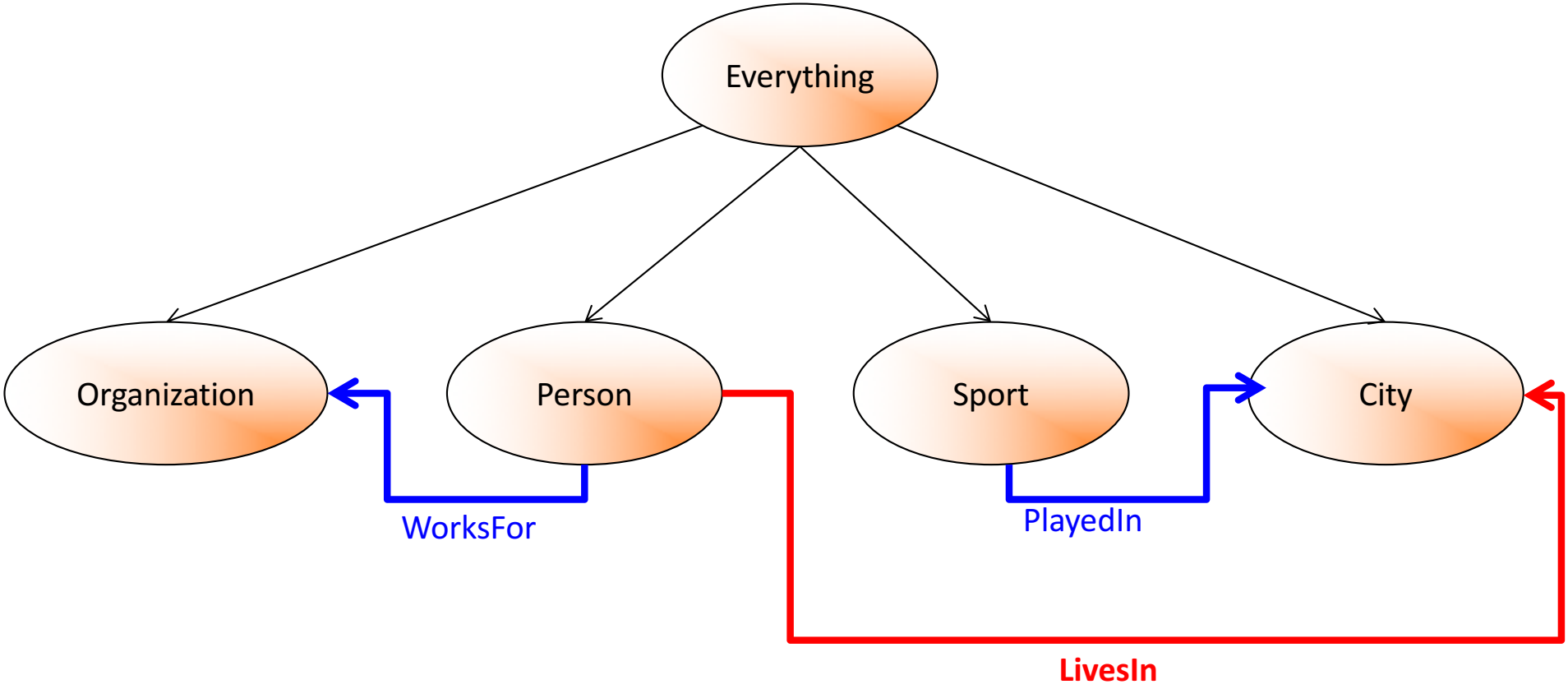
1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology

Key Idea 3: Automatically Extending the Ontology

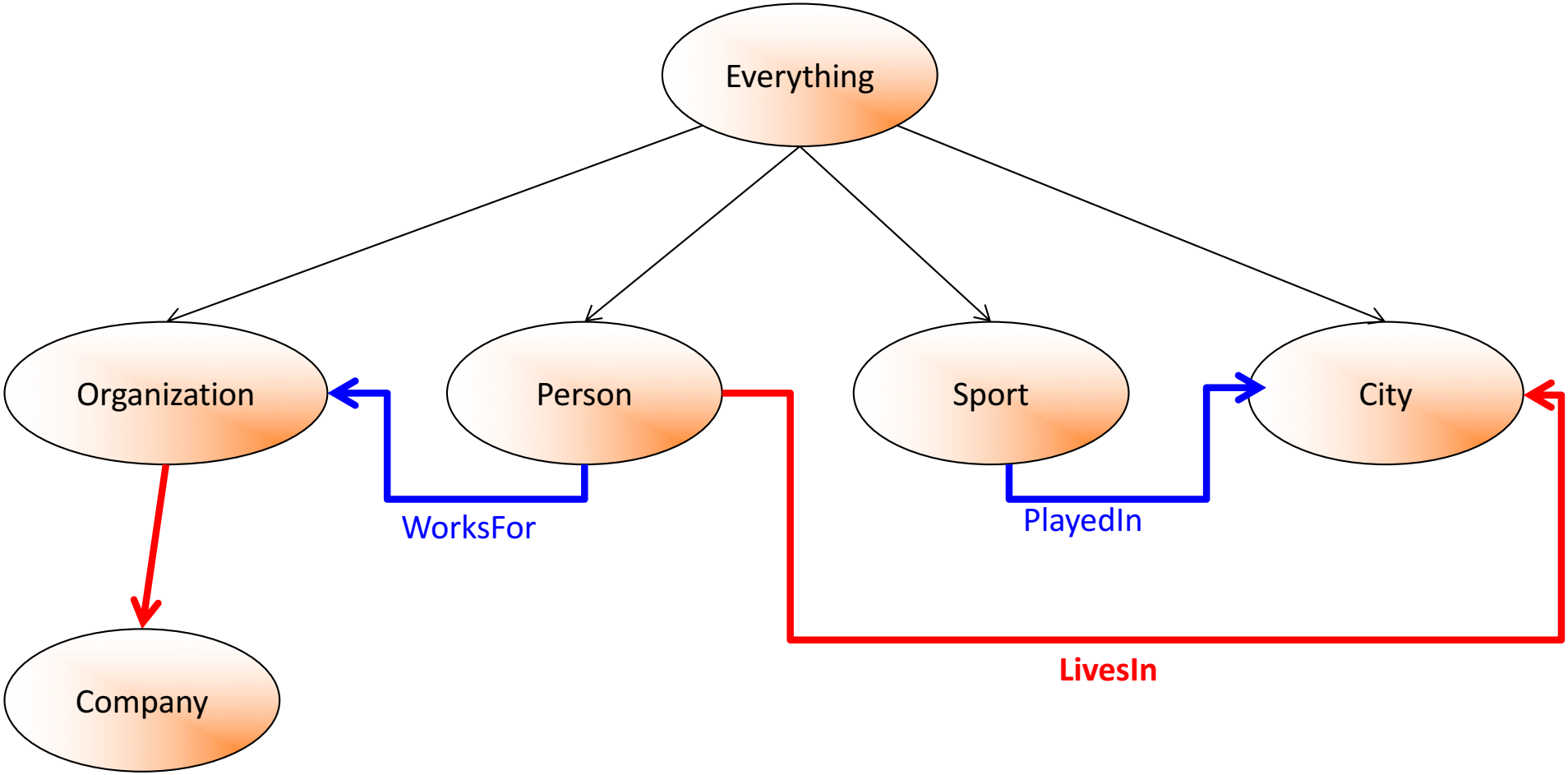
OntExt (Ontology Extension)



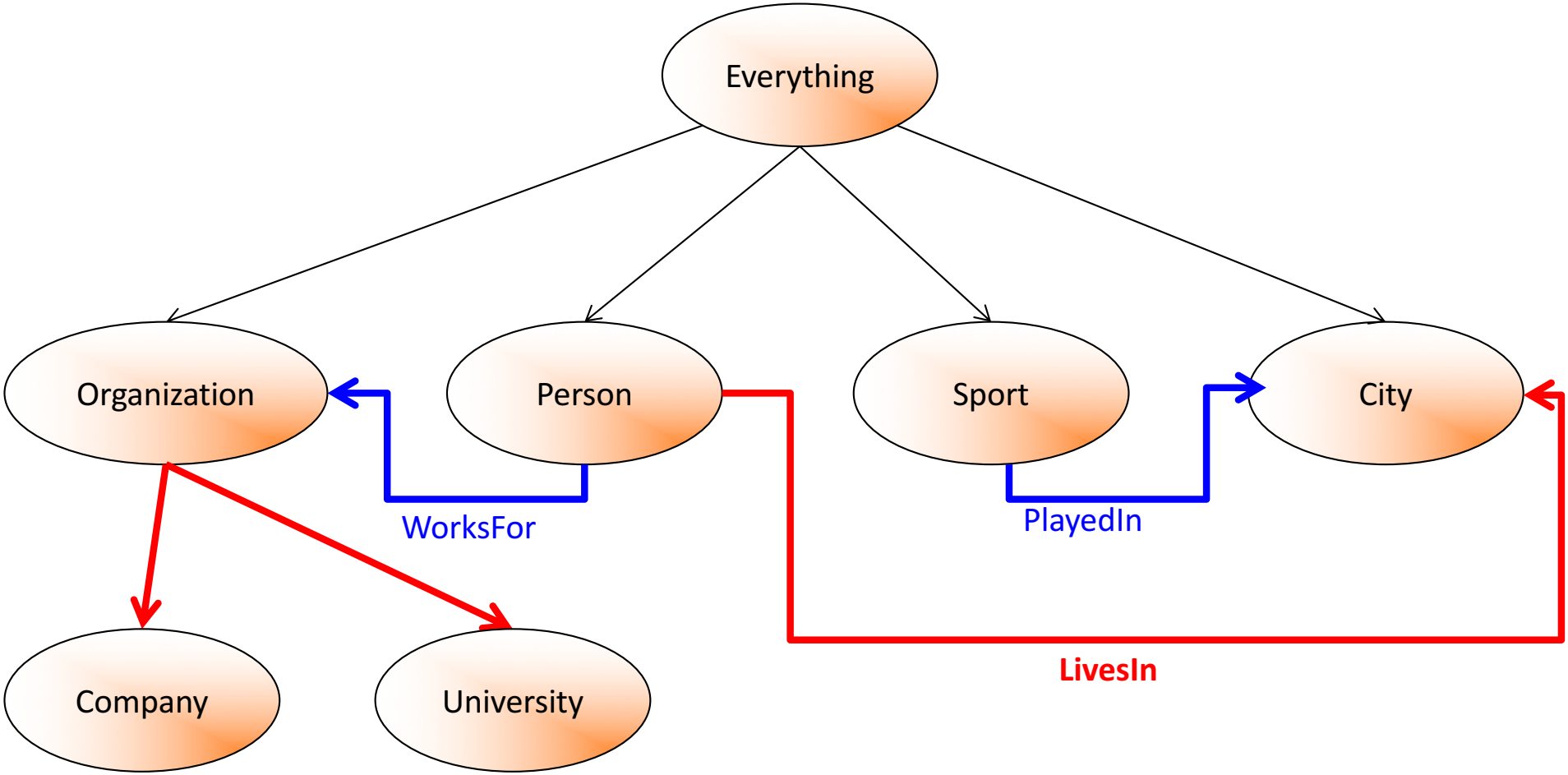
OntExt (Ontology Extension)



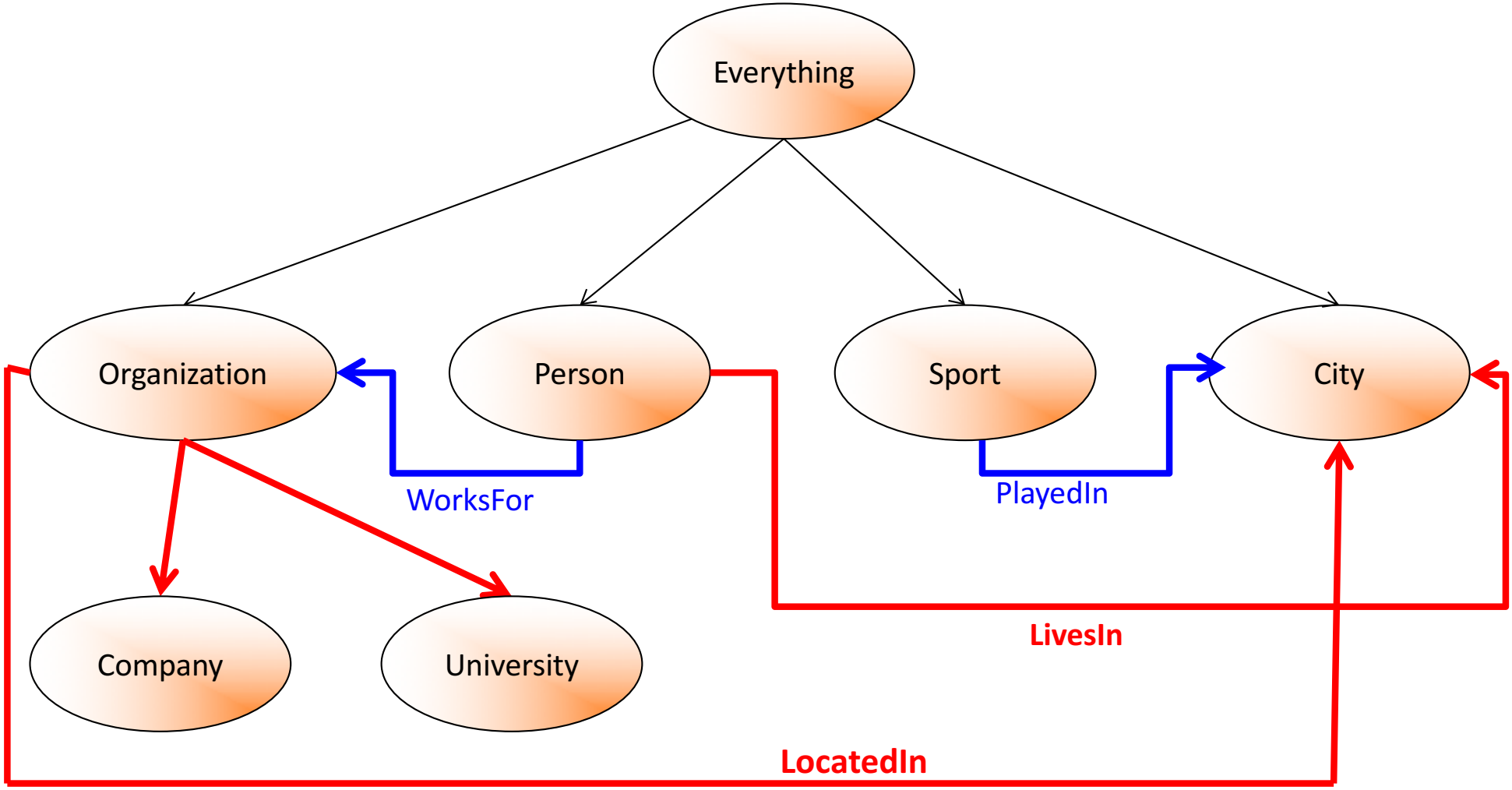
OntExt (Ontology Extension)



OntExt (Ontology Extension)

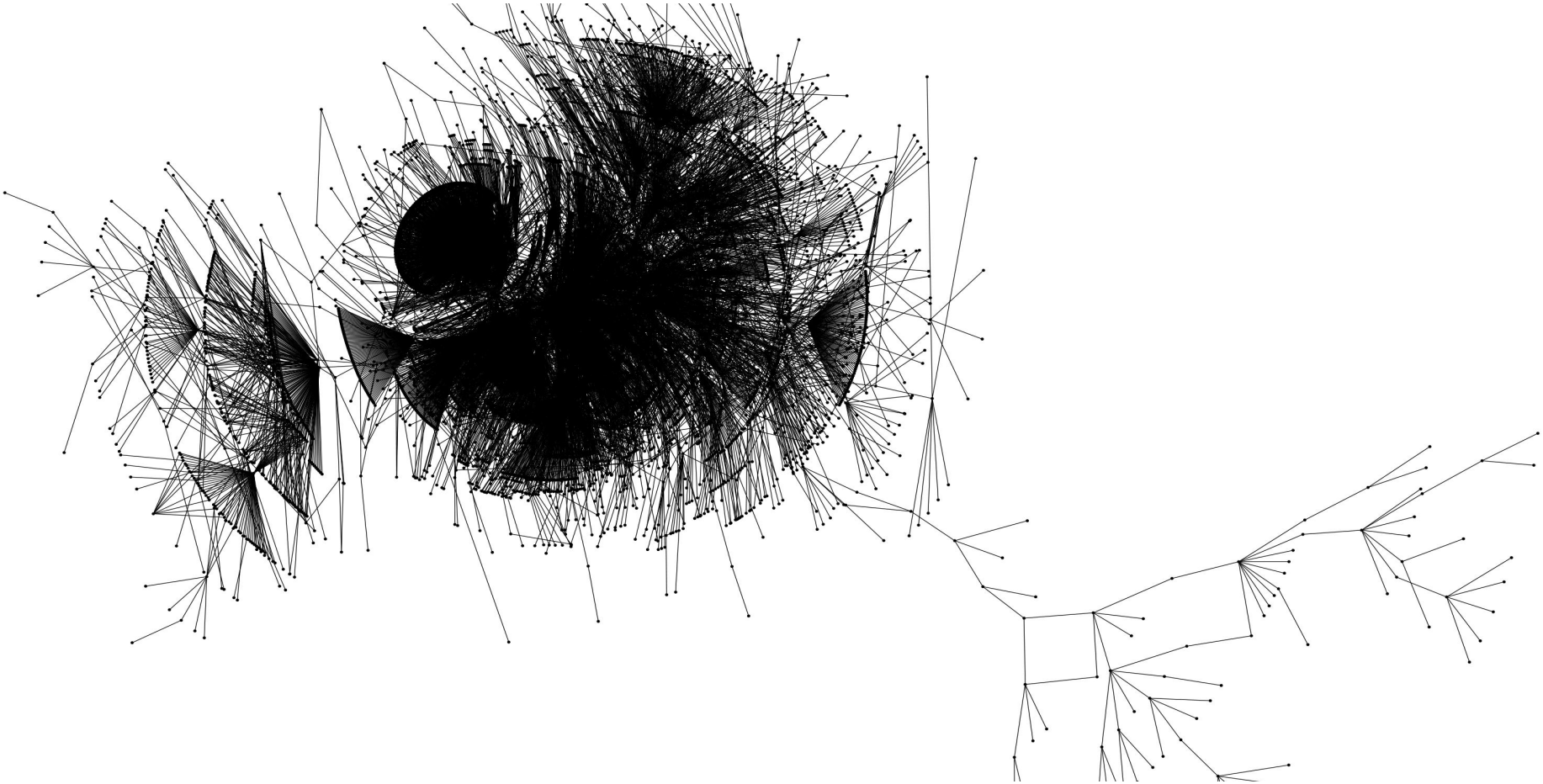


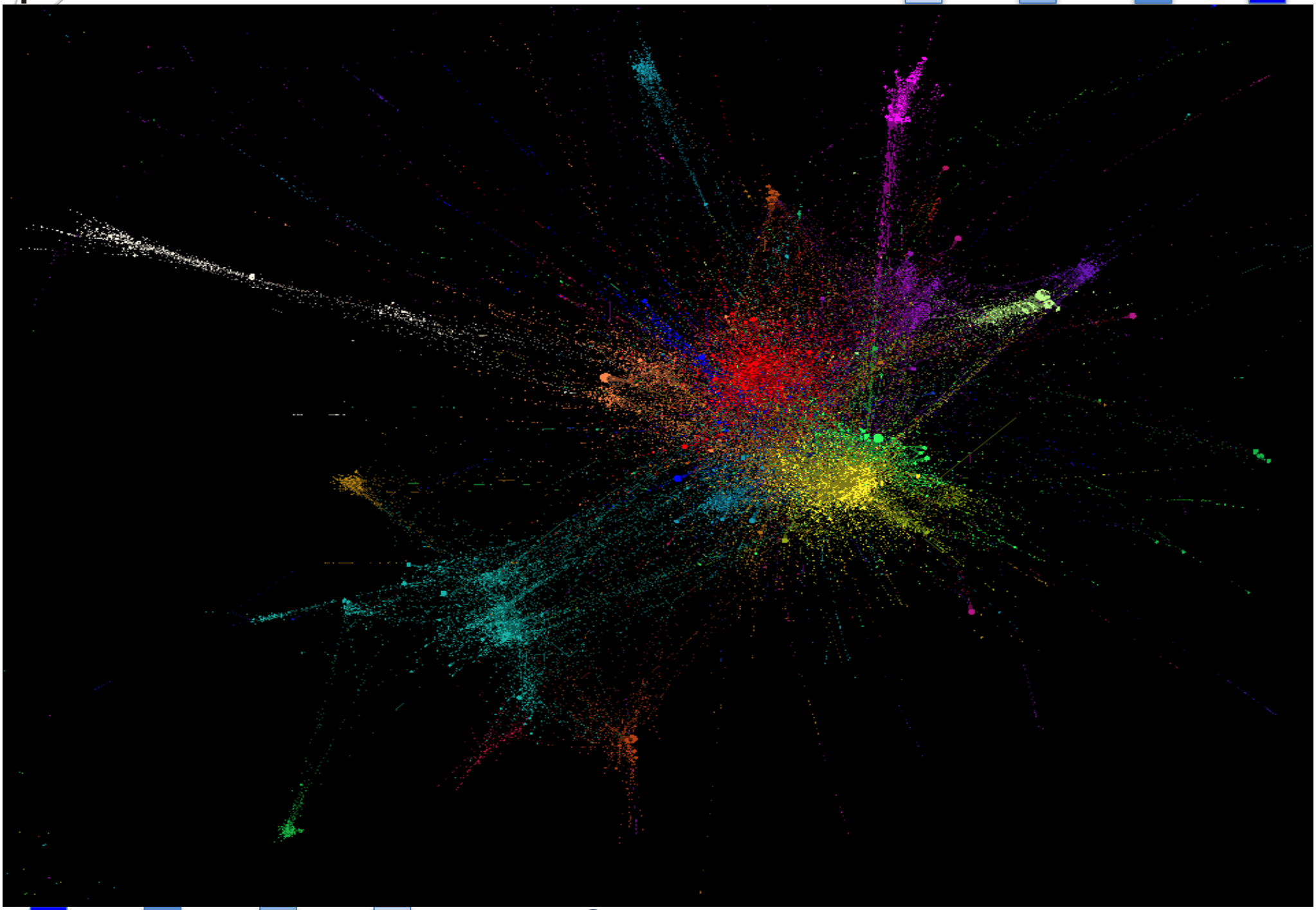
OntExt (Ontology Extension)

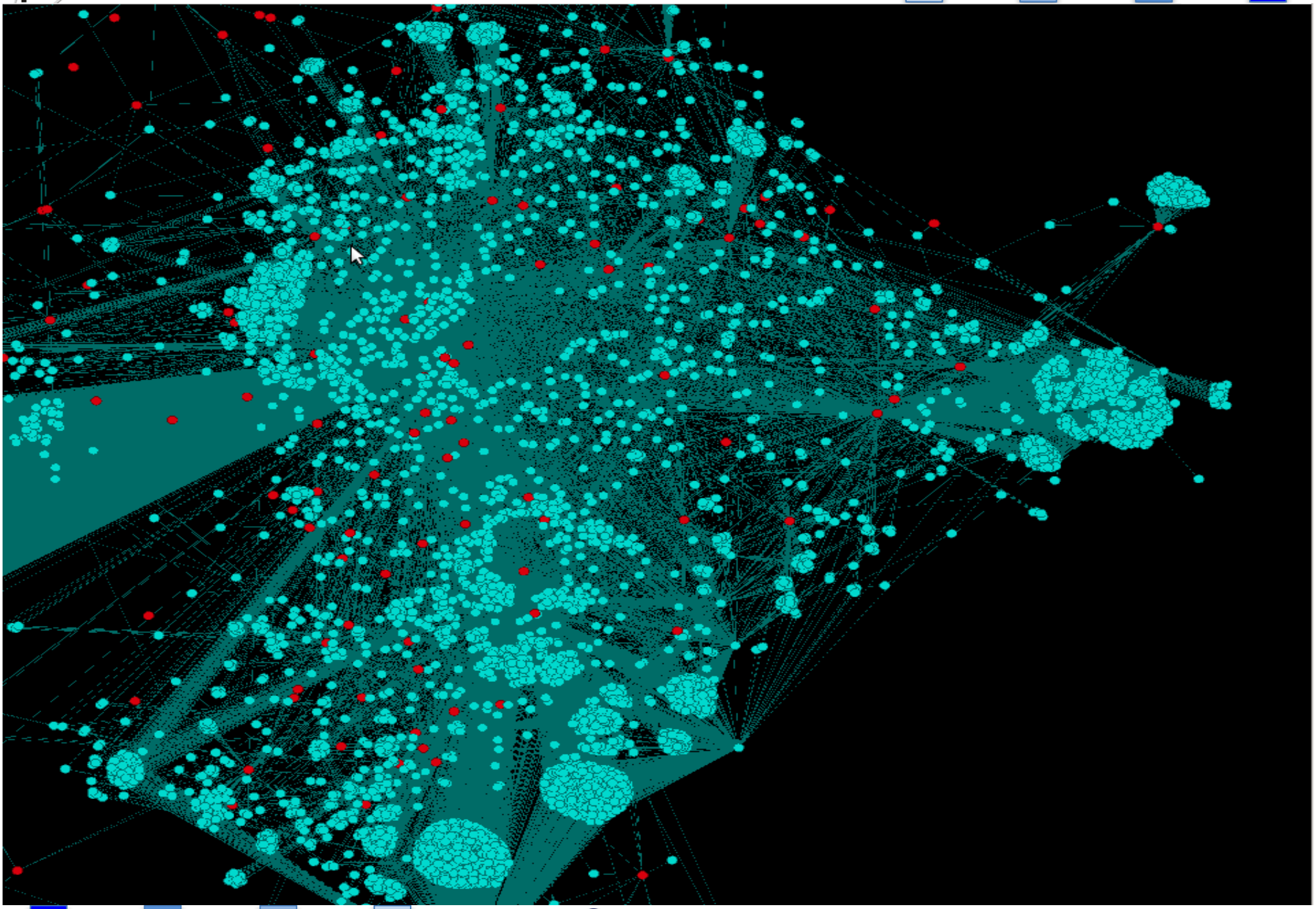


Prophet

- Find open triangles in the Graph

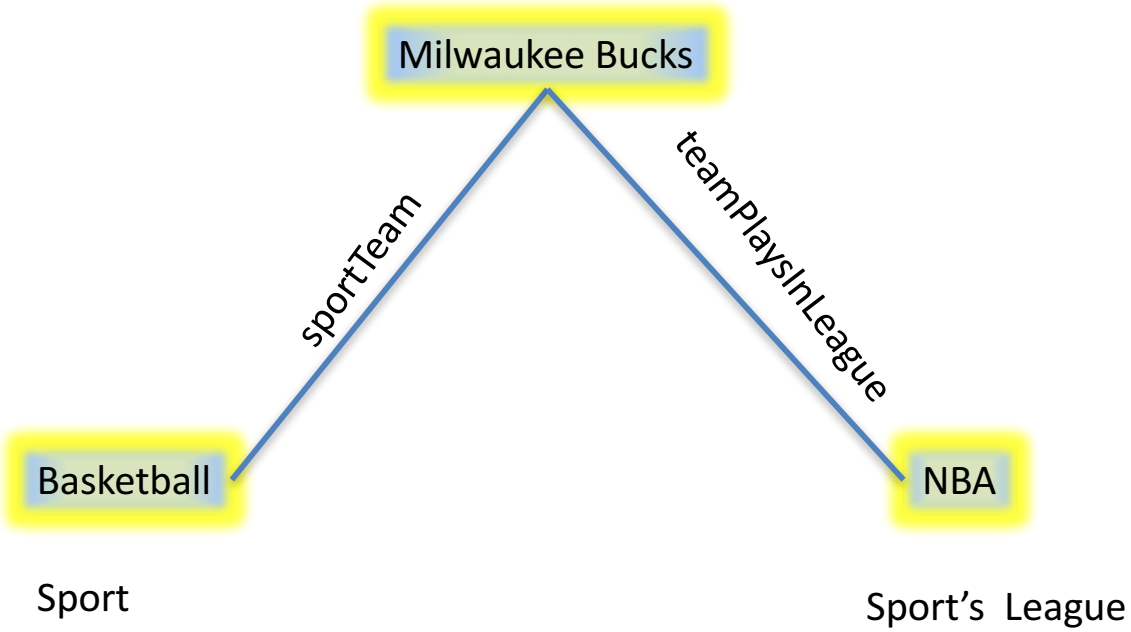






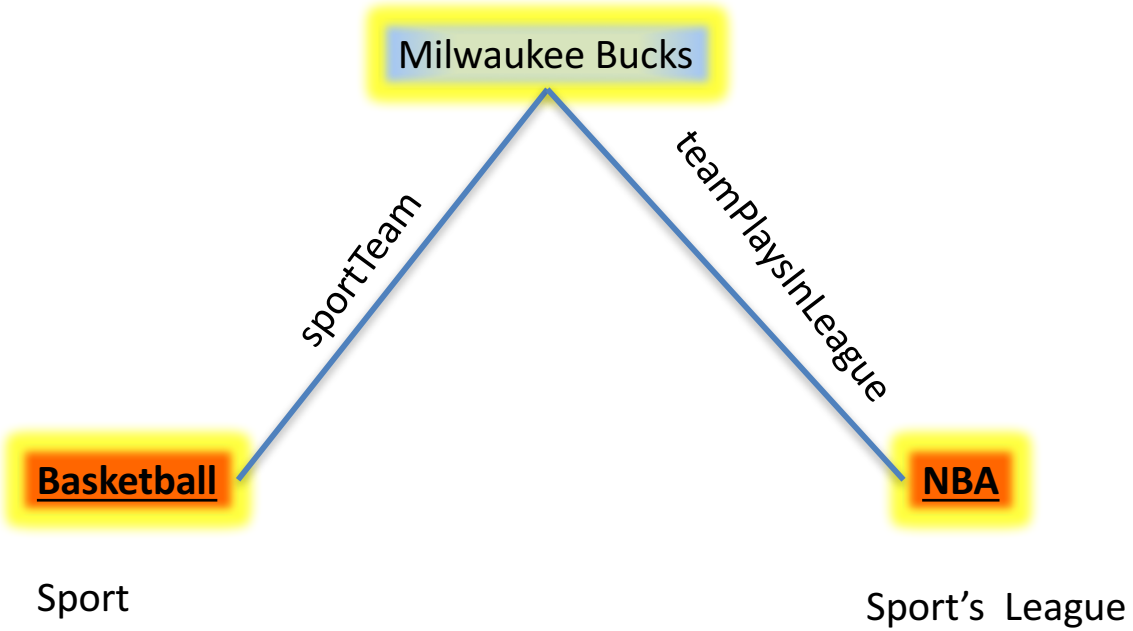
Prophet

- open triangles



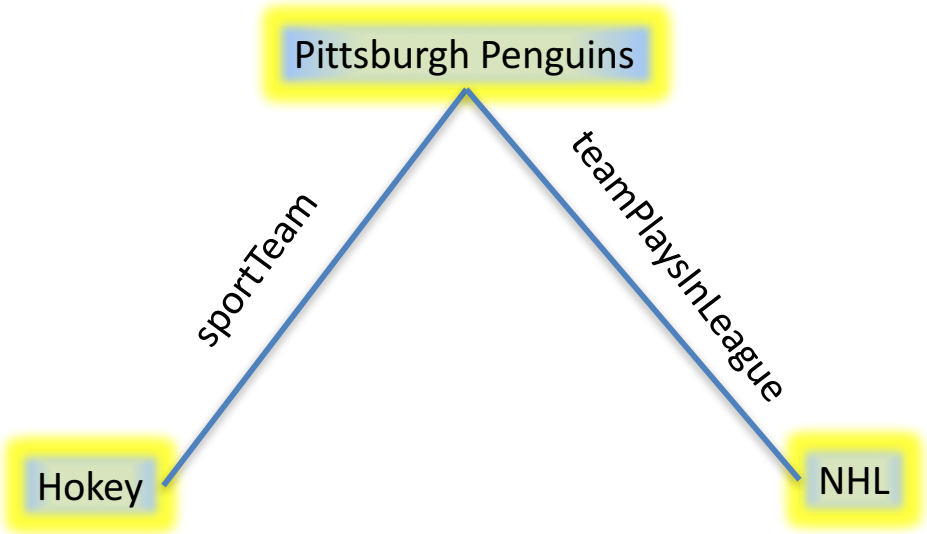
Prophet

- open triangles



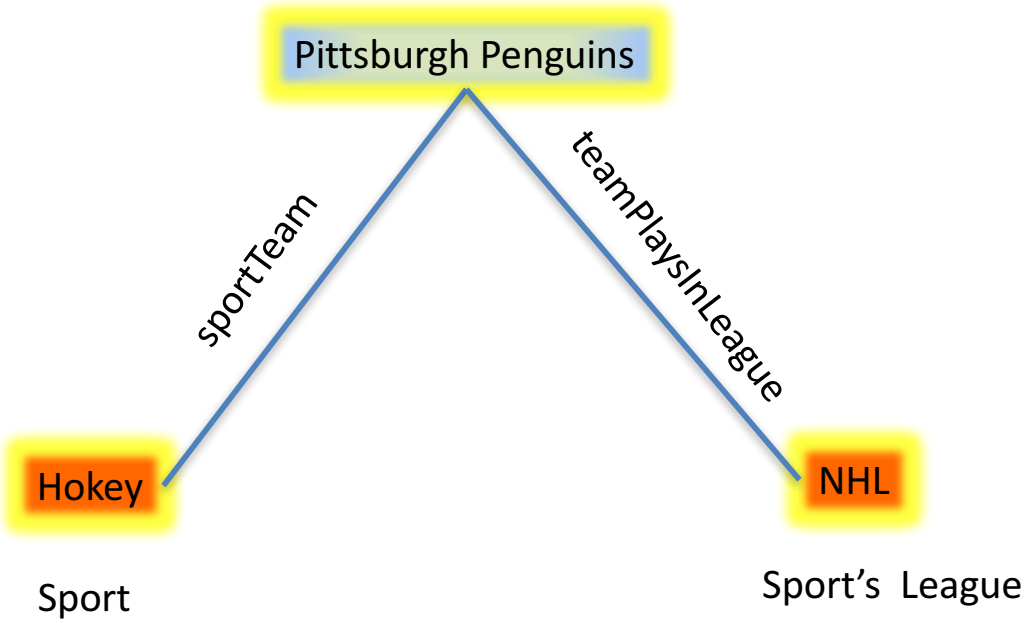
Prophet

- open triangles



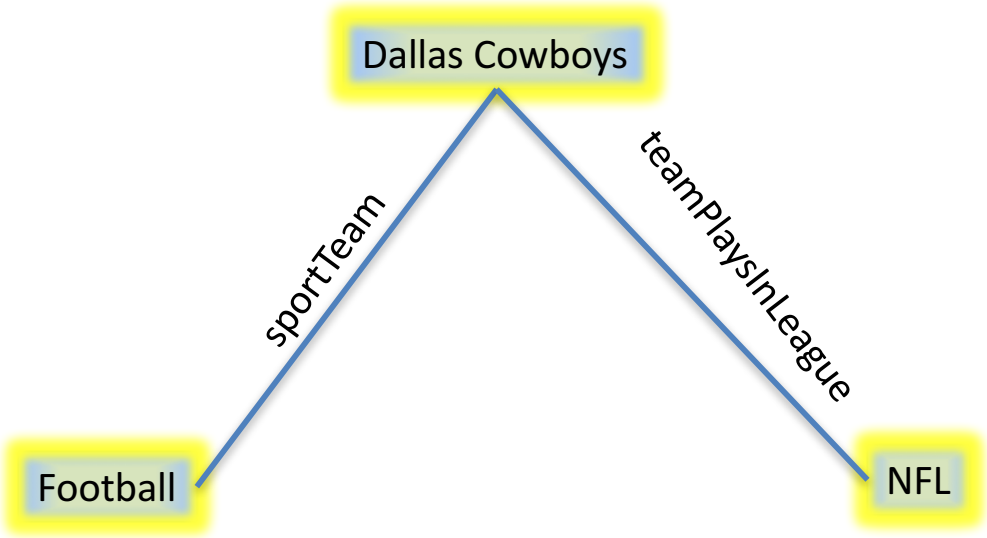
Prophet

- open triangles



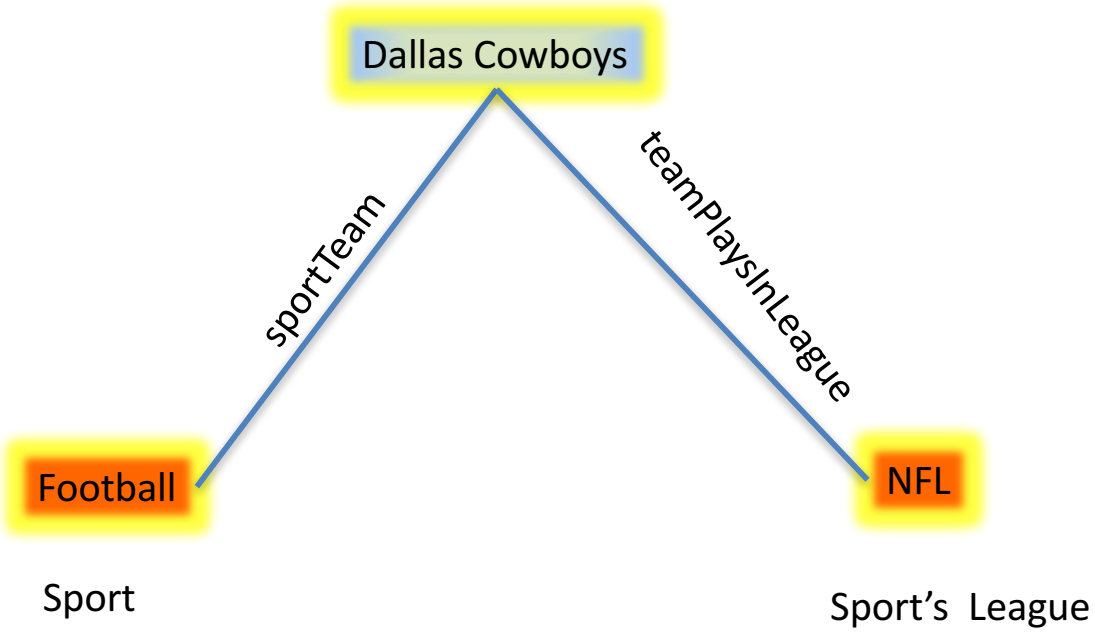
Prophet

- open triangles



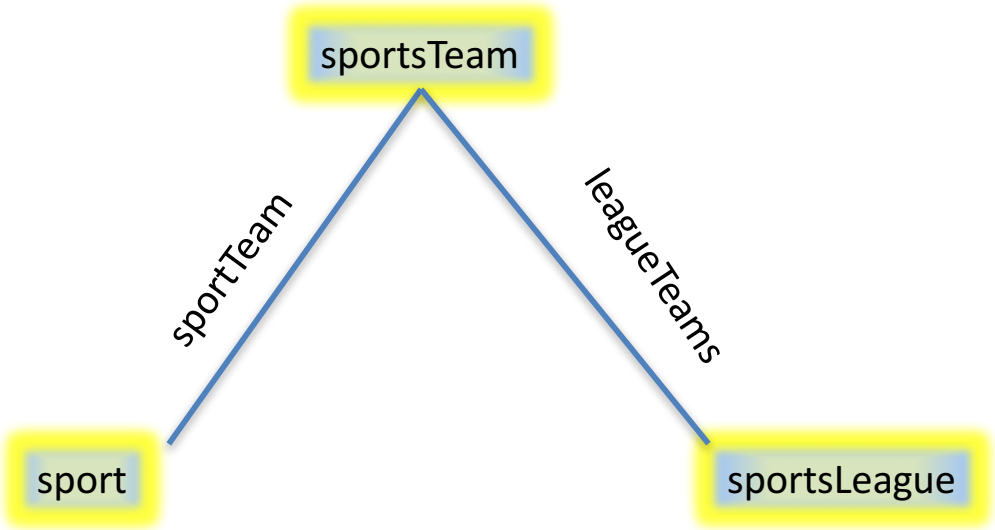
Prophet

- open triangles



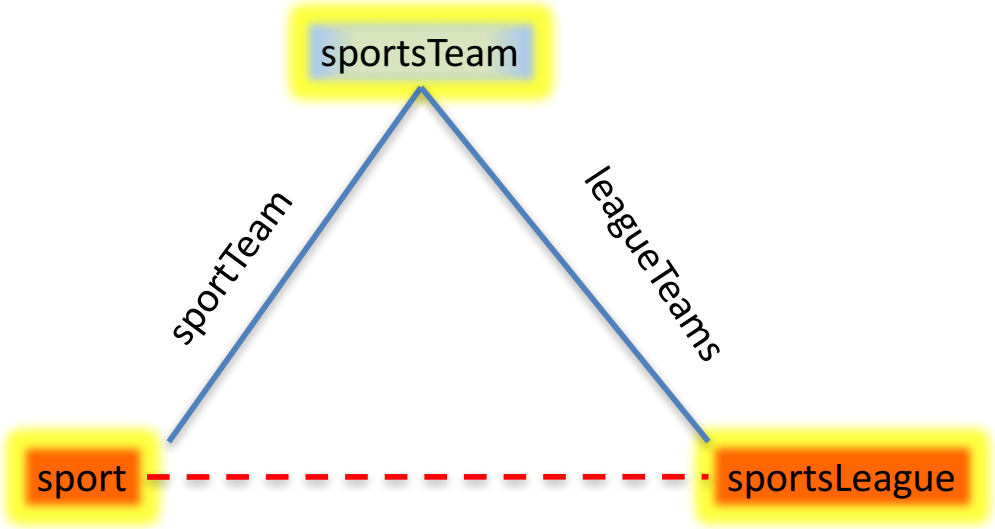
Prophet

- open triangles



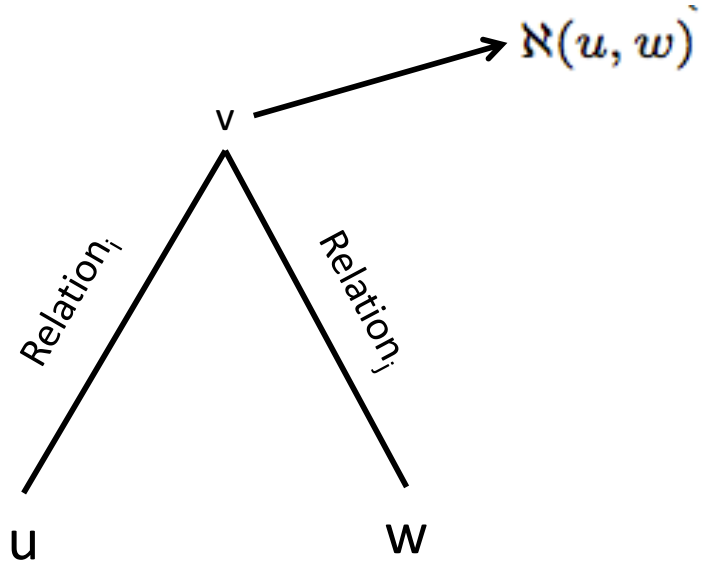
Prophet

- open triangles



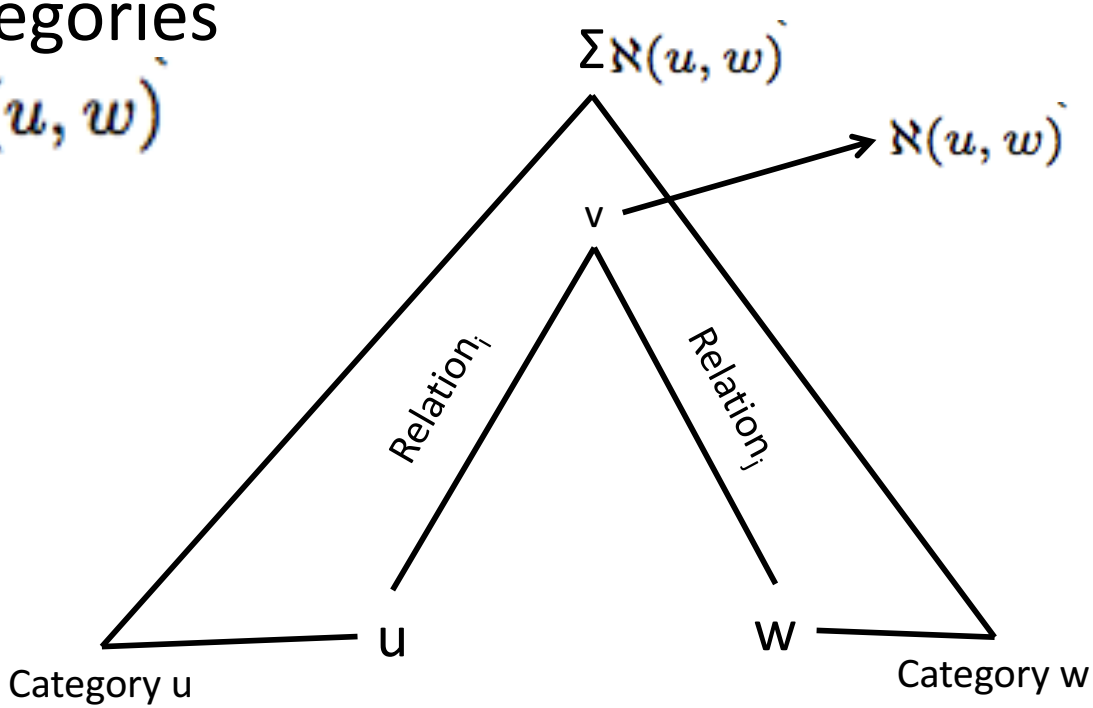
Prophet

- Compute the number of common neighbors $\aleph(u, w)$



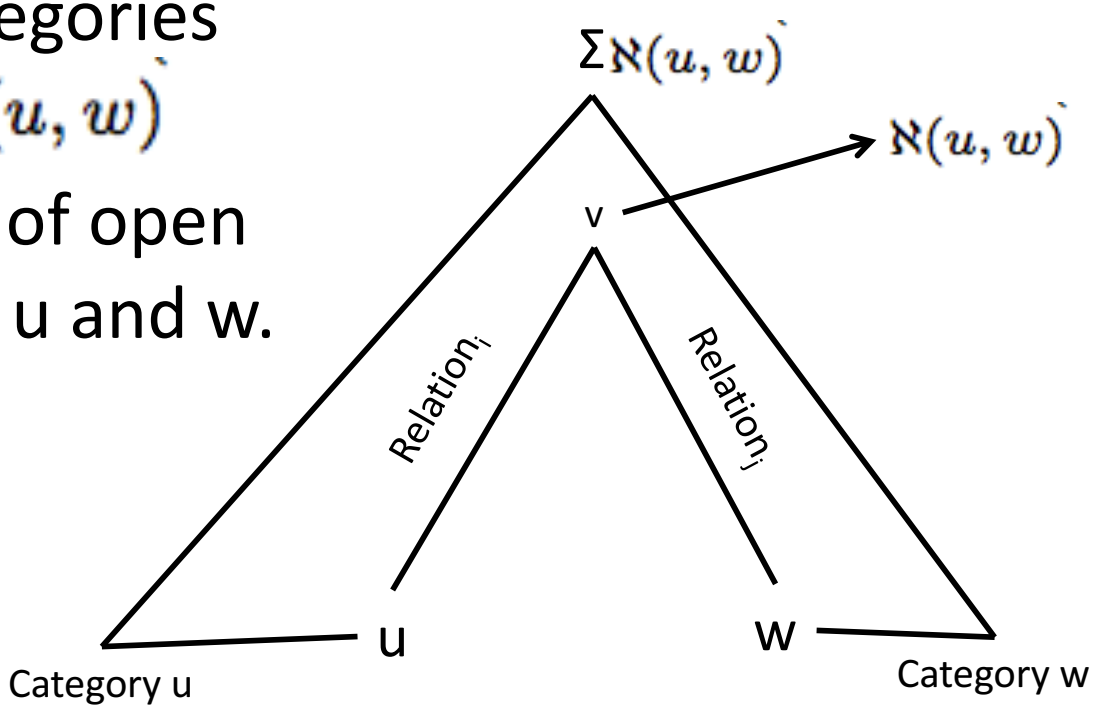
Prophet

- Compute the number of common neighbors $\aleph(u, w)$
- Compute the cumulative number of instances for the categories nodes Σ $\aleph(u, w)$



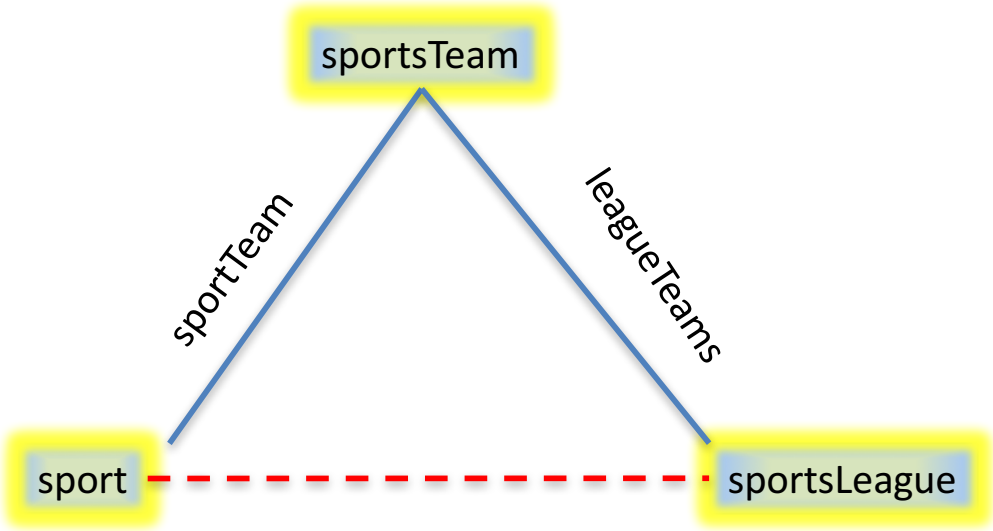
Prophet

- Compute the number of common neighbors $\aleph(u, w)$
- Compute the cumulative number of instances for the categories nodes Σ $\aleph(u, w)$
- $N_{\Lambda_c(u_c,w_c)}$ is the number of open triangles for categories u and w.



Prophet

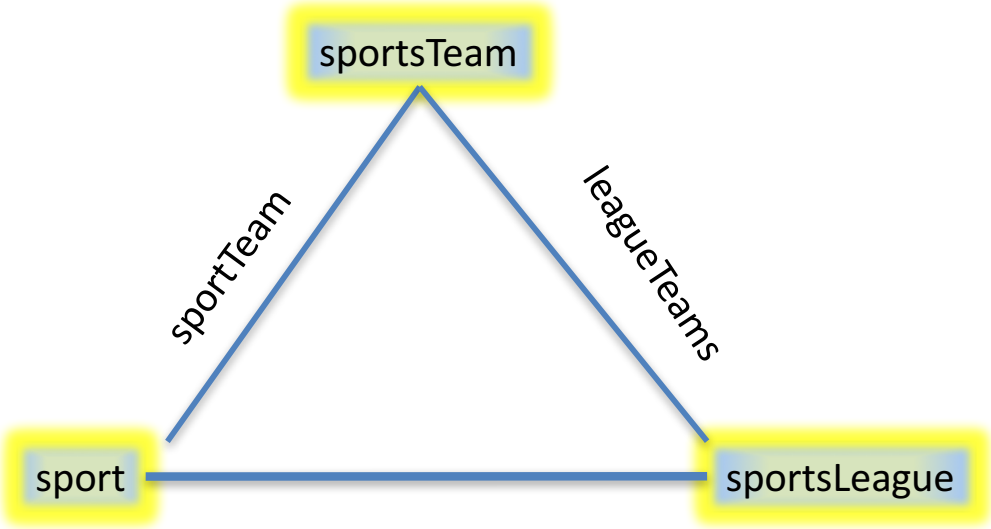
$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$



Prophet

$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

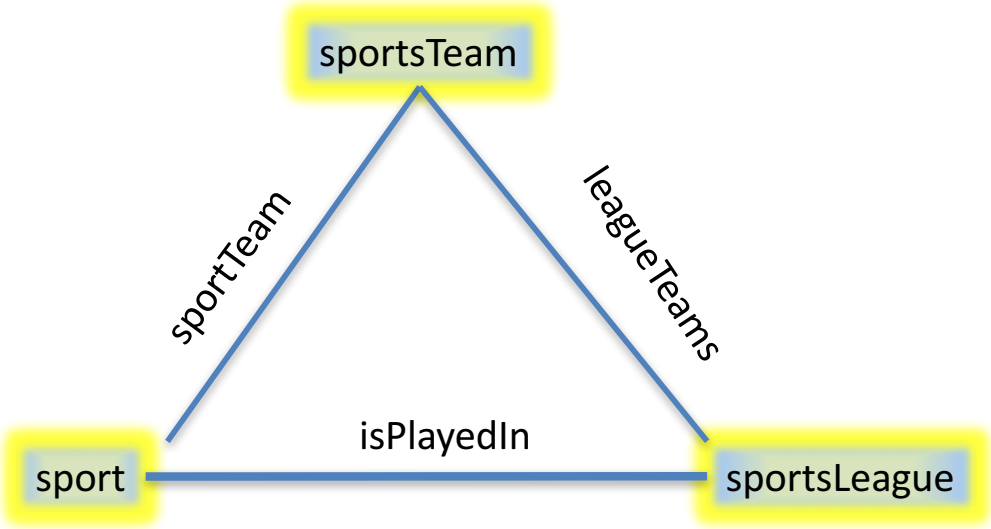
- If $\aleph_c(u_c, w_c)$ > ξ then create the new relation
- ξ = 10 (empirically)



Prophet

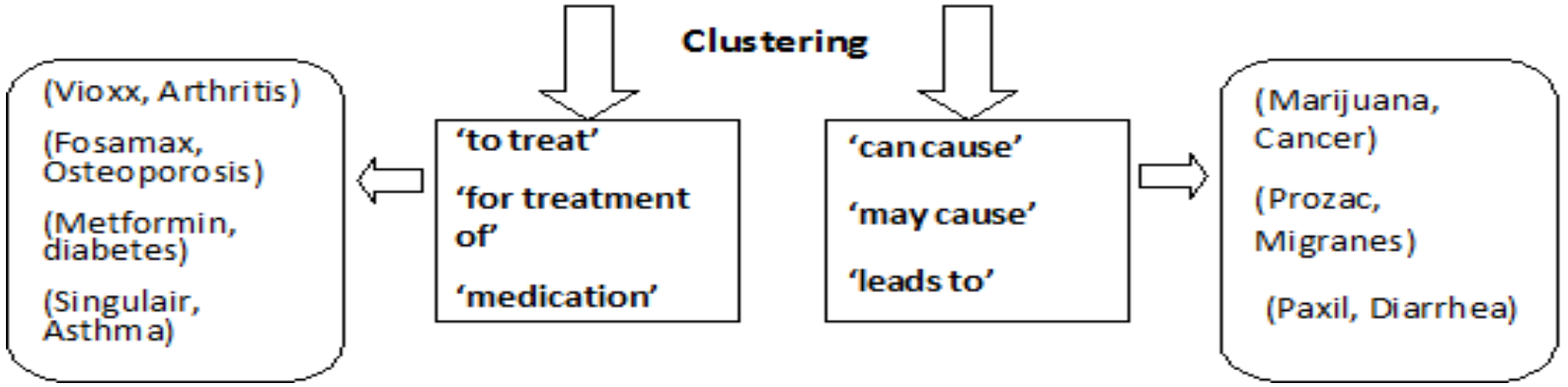
$$\aleph_c(u_c, w_c) = \sum \aleph(u, w) - N_{\Lambda_c(u_c, w_c)}$$

- If $\aleph_c(u_c, w_c)$ > ξ then create the new relation
- ξ = 10 (empirically)
- Name the new relation based on text



OntExt

Contexts/ Contexts	may cause	can cause	can lead to	to treat	for treatment of	medication
may cause	0.176	0.074	0.030	0.015	0.011	0.000
can cause	0.051	0.150	0.039	0.018	0.013	0.010
can lead to	0.034	0.064	0.189	0.019	0.021	0.018
to treat	0.006	0.011	0.007	0.109	0.043	0.015
for treatment of	0.005	0.008	0.008	0.045	0.086	0.023
medication	0.000	0.011	0.009	0.030	0.036	0.111



NELL: sample of self-added relations

- athleteWonAward
- animalEatsFood
- languageTaughtInCity
- clothingMadeFromPlant
- beverageServedWithFood
- fishServedWithFood
- athleteBeatAthlete
- athleteInjuredBodyPart
- arthropodFeedsOnInsect
- animalEatsVegetable
- plantRepresentsEmotion
- foodDecreasesRiskOfDisease
- clothingGoesWithClothing
- bacteriaCausesPhysCondition
- buildingMadeOfMaterial
- emotionAssociatedWithDisease
- foodCanCauseDisease
- agriculturalProductAttractsInsect
- arteryArisesFromArtery
- countryHasSportsFans
- bakedGoodServedWithBeverage
- beverageContainsProtein
- animalCanDevelopDisease
- beverageMadeFromBeverage

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks

CityLocatedInCountry(Pittsburgh) = ?

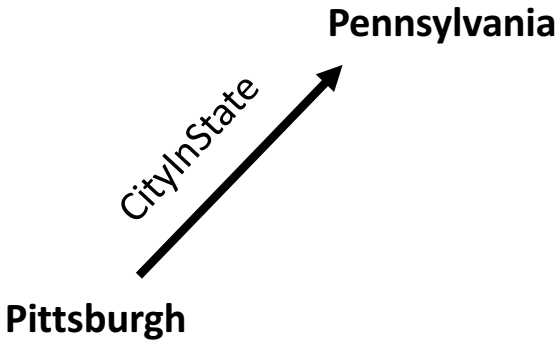
[Lao, Mitchell, Cohen, *EMNLP* 2011]

Pittsburgh

Feature = Typed Path	**Feature Value**	**Logistic Regresssion Weight**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.32	

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path

CityInState, CityInstate^{-1}, CityLocatedInCountry

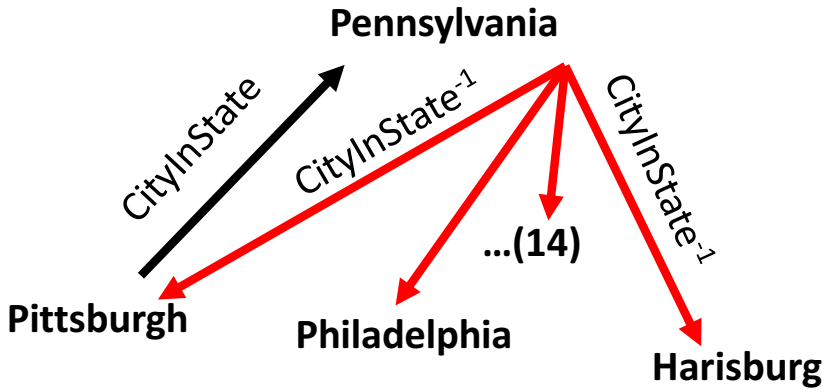
Feature Value

0.32

Logistic Regresssion Weight

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



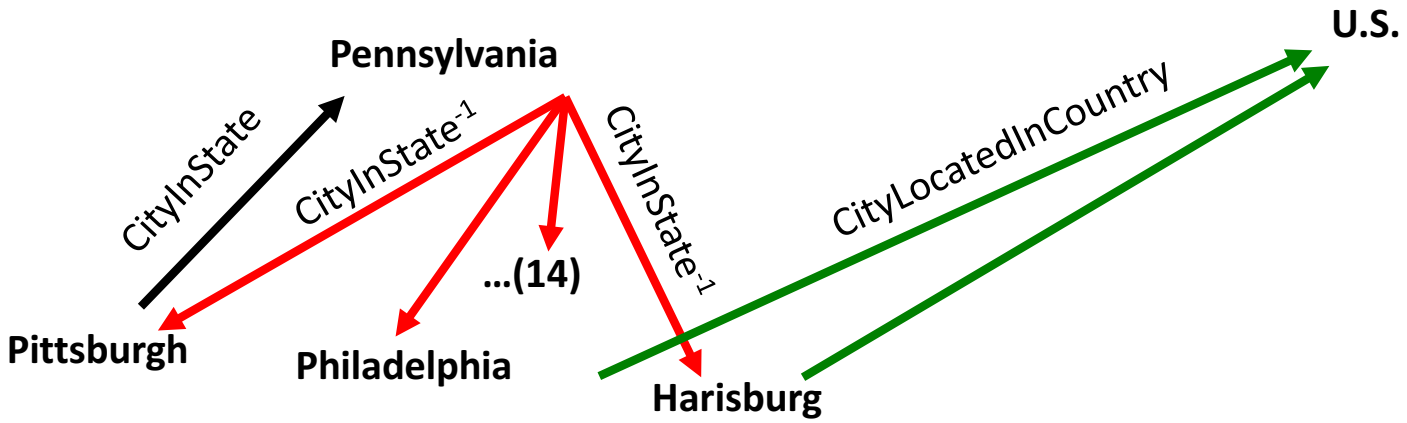
Feature = Typed Path
CityInState, CityInstate^{-1}, CityLocatedInCountry

Feature Value
0.32

Logistic Regresssion Weight

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path

CityInState, CityInstate^{-1}, CityLocatedInCountry

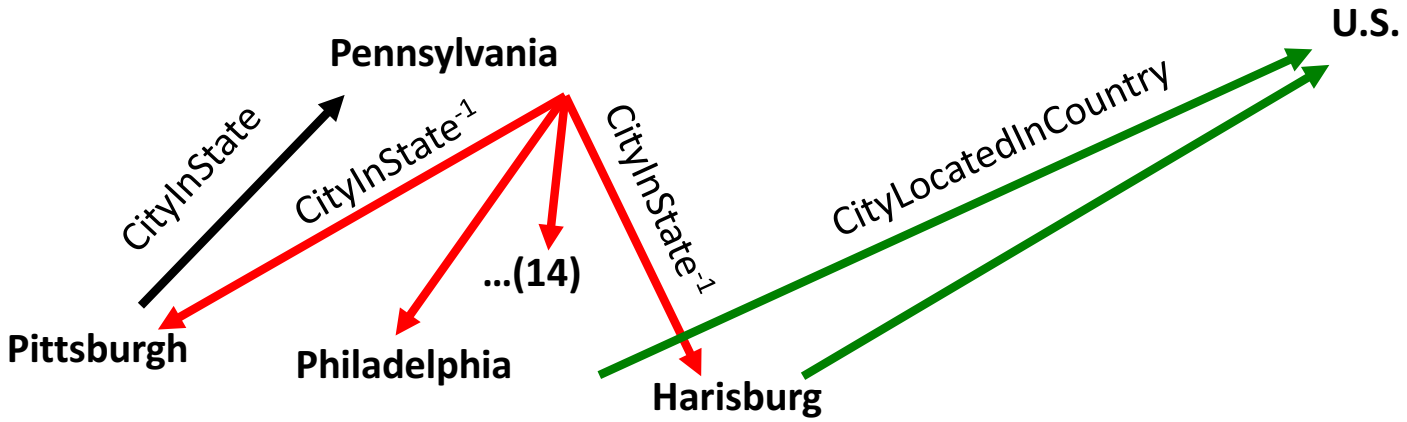
Feature Value

0.32

Logistic Regresssion Weight

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



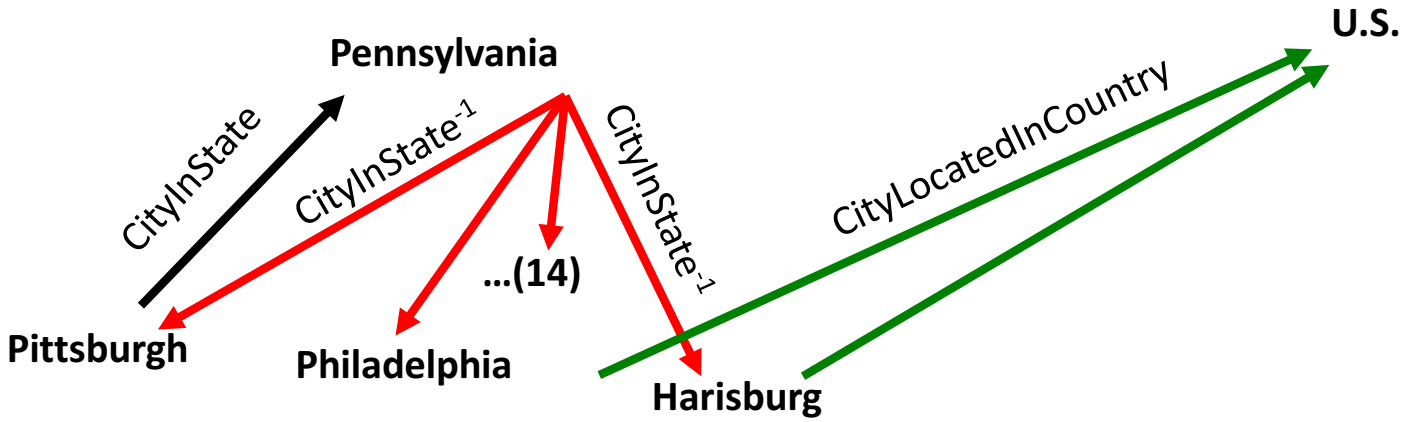
Pr(U.S. | Pittsburgh, TypedPath)

Logistic Regresssion Weight

Feature = Typed Path		**Feature Value**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]

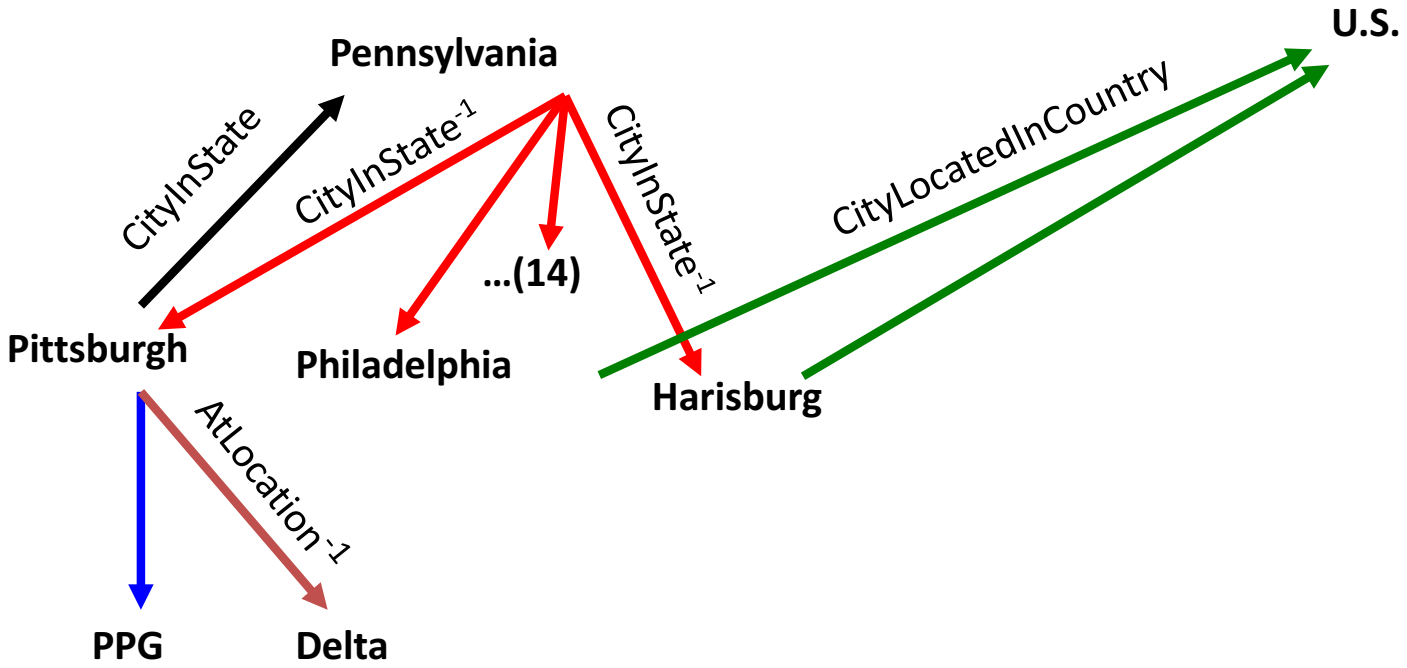


Feature = Typed Path		Feature Value
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

Logistic Regresssion Weight

CityLocatedInCountry(Pittsburgh) = ?

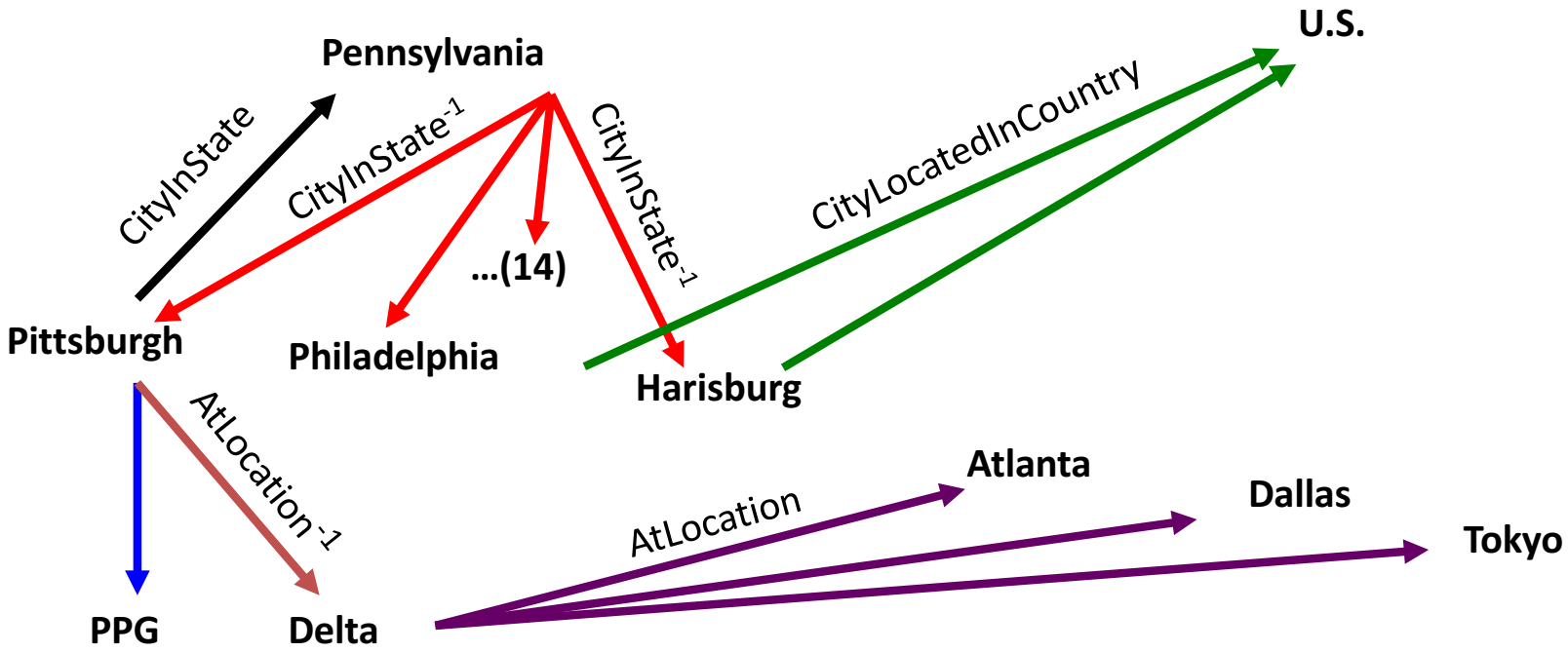
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	**Logistic Regresssion Weight**	**Feature Value**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

CityLocatedInCountry(Pittsburgh) = ?

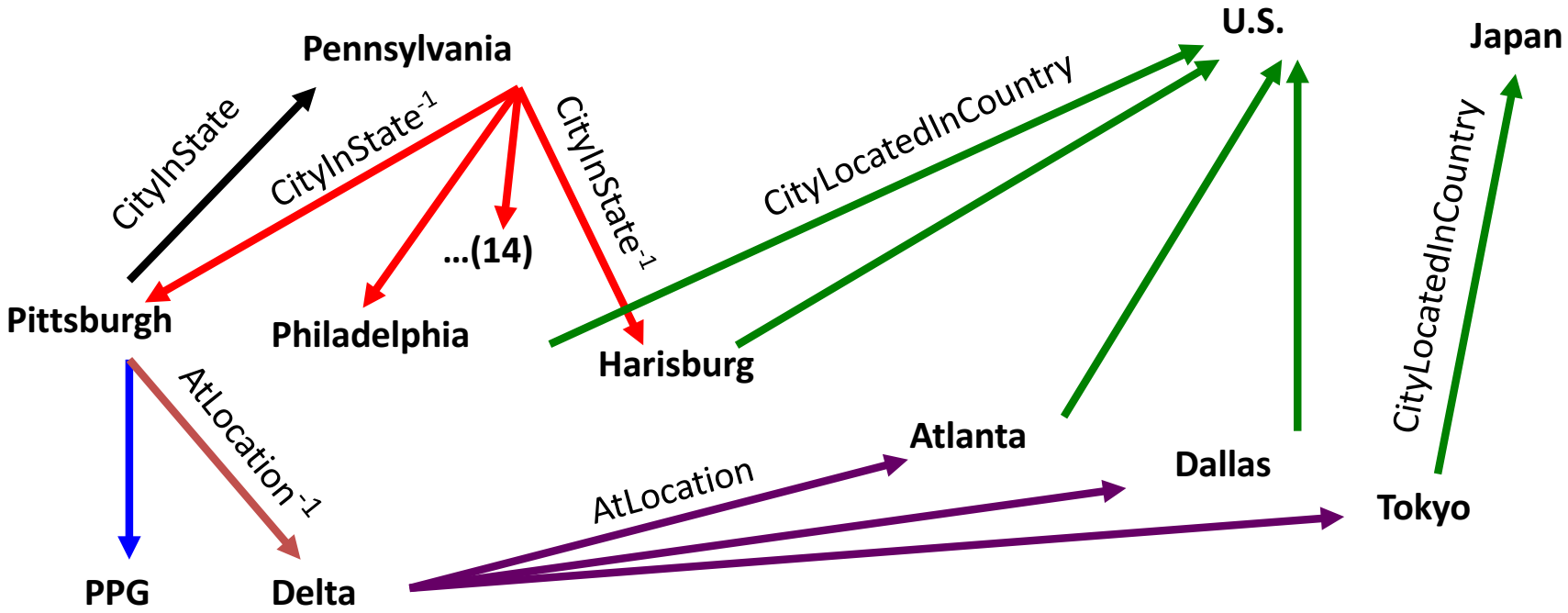
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	Logistic Regresssion Weight	Feature Value
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry		0.20

CityLocatedInCountry(Pittsburgh) = ?

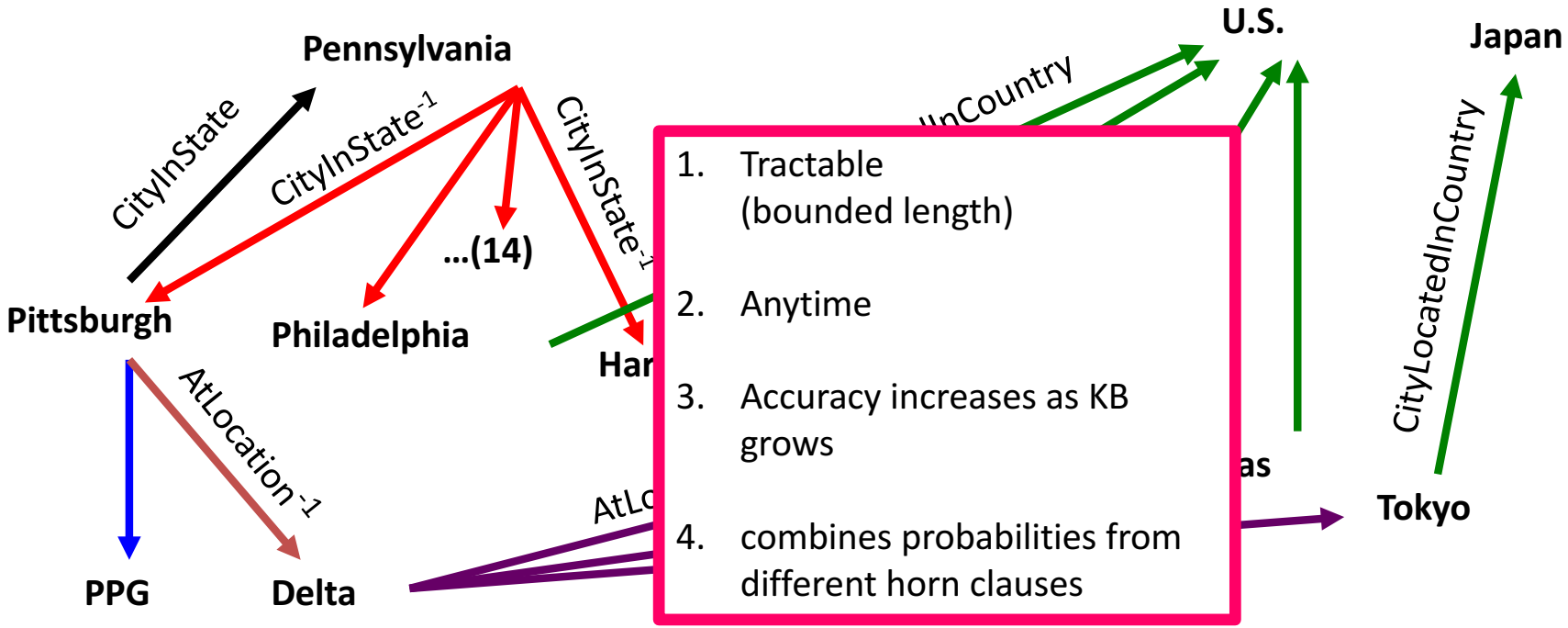
[Lao, Mitchell, Cohen, *EMNLP* 2011]



Feature = Typed Path	Logistic Regresssion Weight	Feature Value
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry	0.6	0.20

CityLocatedInCountry(Pittsburgh) = ?

[Lao, Mitchell, Cohen, *EMNLP* 2011]



Logistic Regresssion Weight

Feature = Typed Path		**Feature Value**
CityInState, CityInstate^{-1}, CityLocatedInCountry	0.8	0.32
AtLocation^{-1}, AtLocation, CityLocatedInCountry	0.6	0.20
...	...	...

CityLocatedInCountry(Pittsburgh) = U.S. p=0.58

Random walk inference: learned rules

- CityLocatedInCountry(*city, country*):

- 8.04 cityliesonriver, cityliesonriver^{-1}, citylocatedincountry
- 5.42 hasofficeincity^{-1}, hasofficeincity, citylocatedincountry
- 4.98 cityalsoknownas, cityalsoknownas, citylocatedincountry
- 2.85 citycapitalofcountry,citylocatedincountry^{-1},citylocatedincountry
- 2.29 agentactsinlocation^{-1}, agentactsinlocation, citylocatedincountry
- 1.22 statehascapital^{-1}, statelocatedincountry
- 0.66 citycapitalofcountry
- .
- .
- .

7 of the 2985 learned rules for CityLocatedInCountry

Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL

New Direction: Integrate Vision with Text

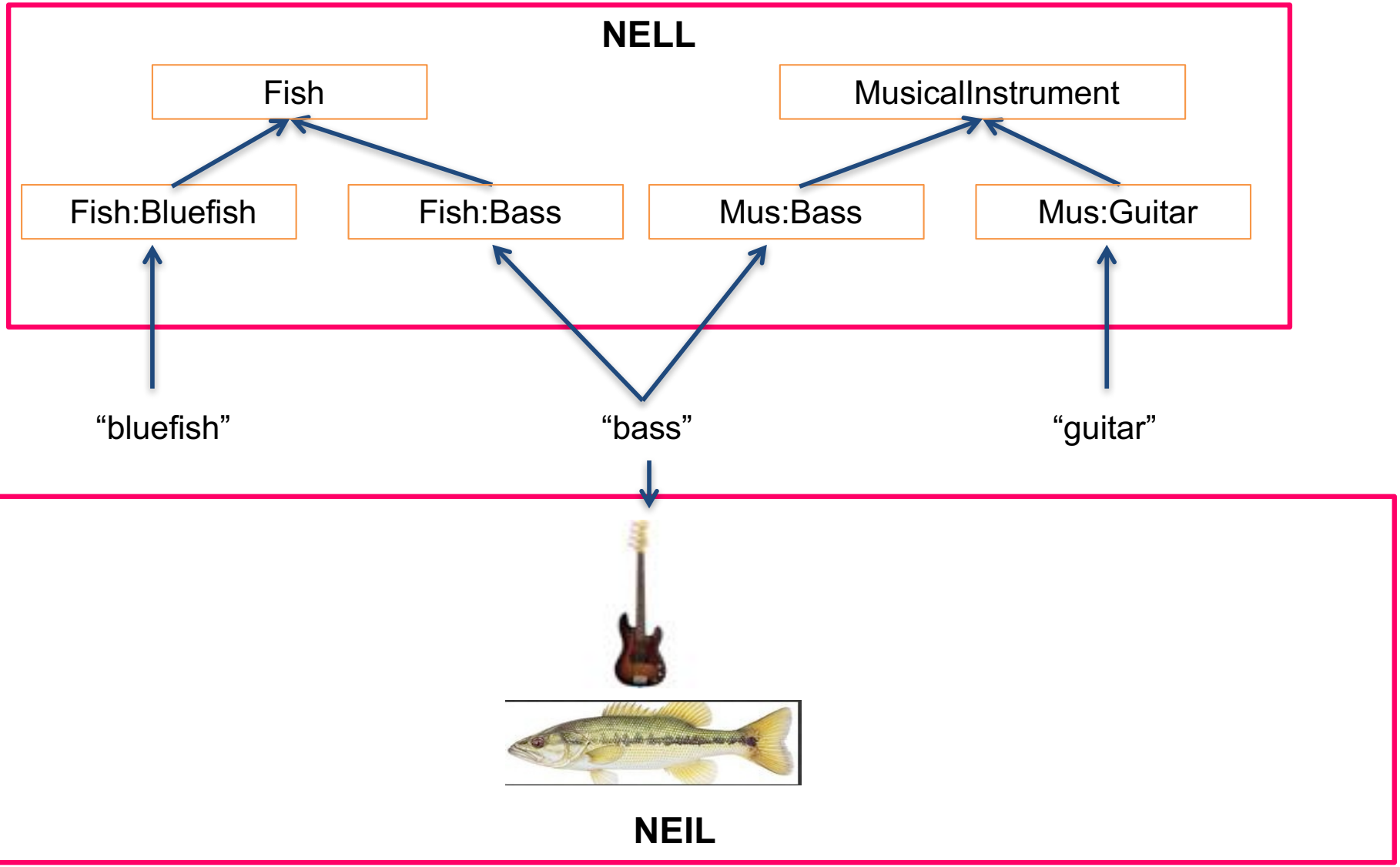
•The problem:

- Many things not learnable from text

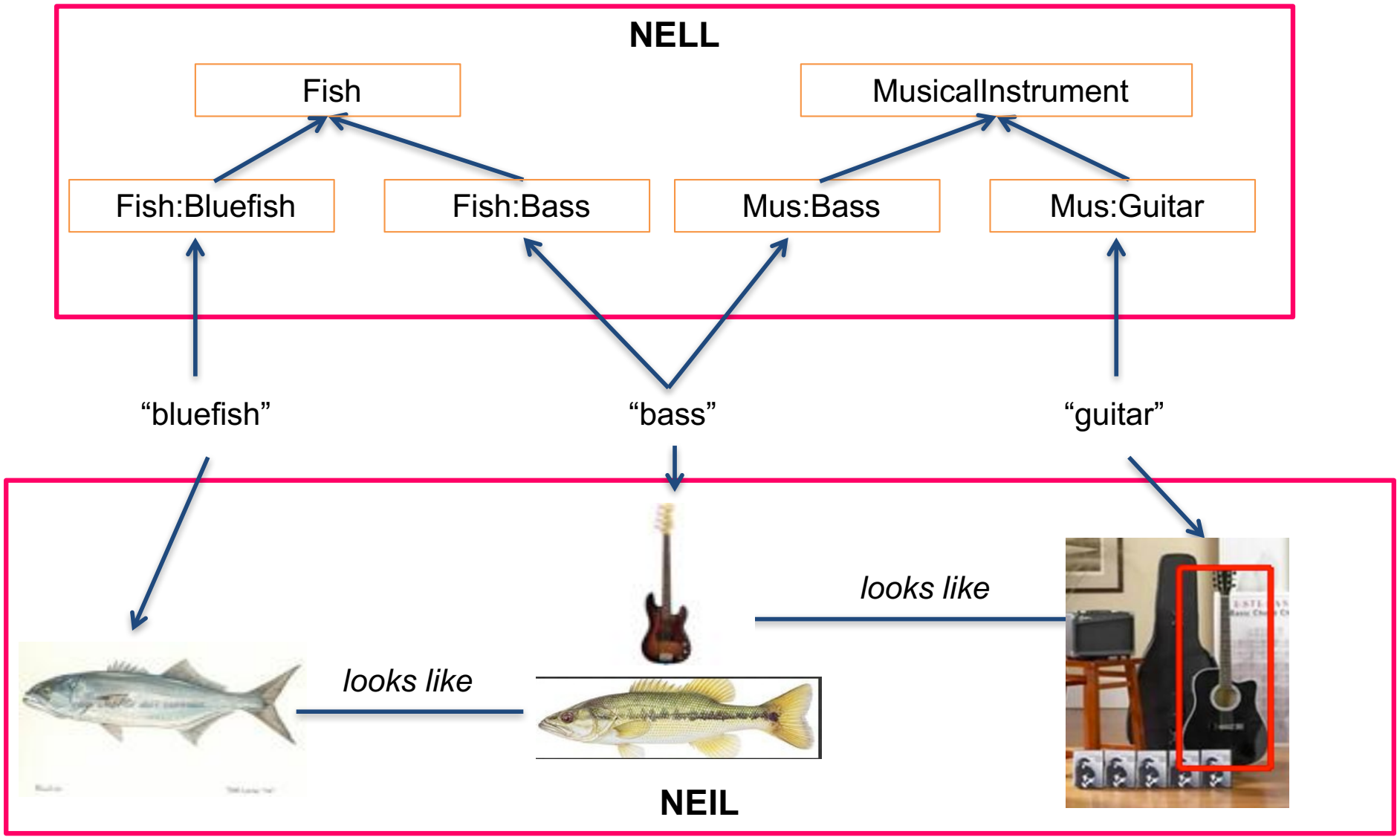
•New direction:

- integrate NELL with NEIL (Never Ending Image Learner) [Gupta, Chen, 2013]
- NELL gives noun phrases it understands to NEIL
- NEIL collects images associated with these, and analyzes
- NELL, NEIL cotraining

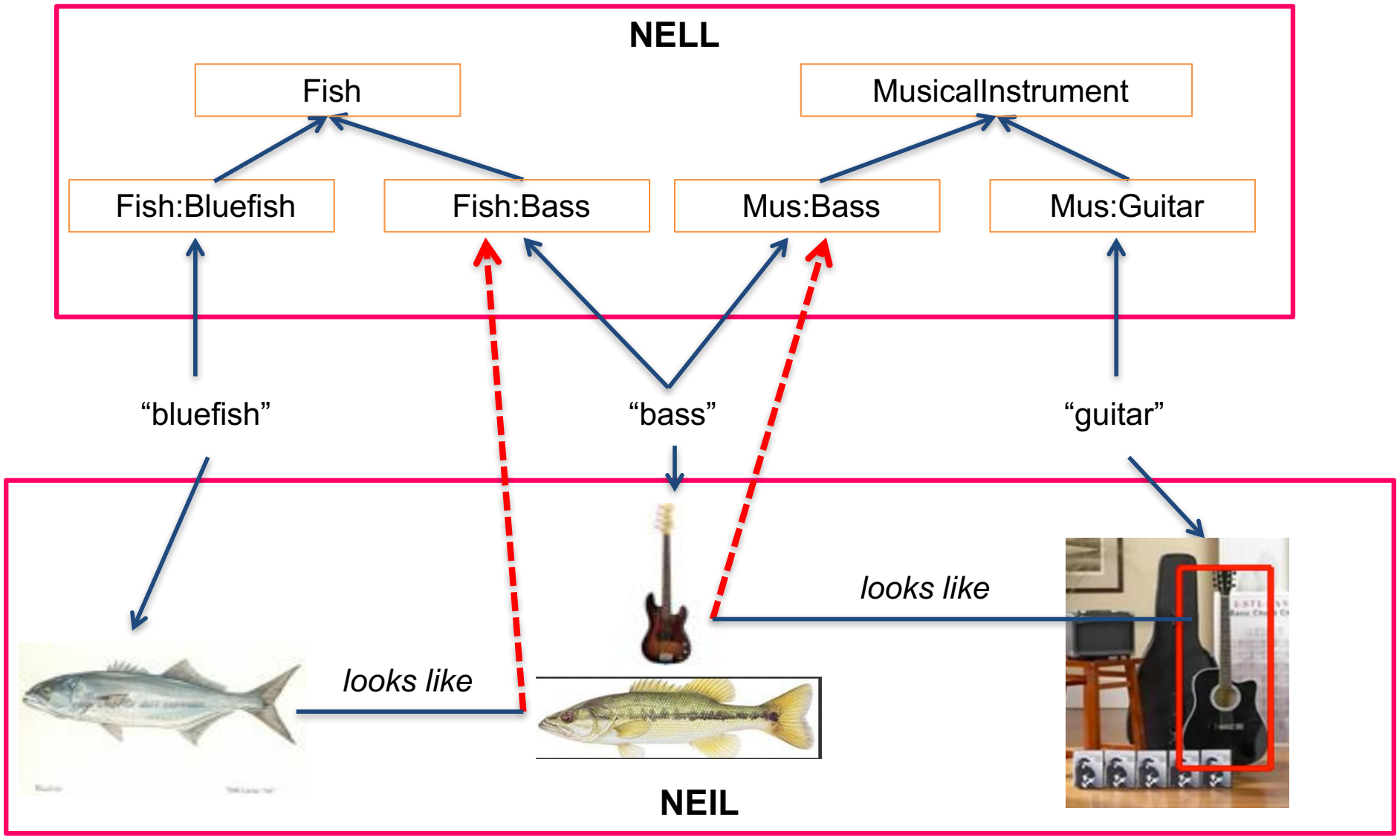
NEIL / NELL Polysemy: Bass



NEIL / NELL Polysemy: Bass



NEIL / NELL Polysemy: Bass



Building the Knowledge Graph by Reading

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL
8. Mutilingual NELL (English, Portuguese, Spanish, ...)

Recently-Learned Facts

twitter

Refresh

Recently learned beliefs (from English text)

instance	iteration	date learned	confidence
actimmune is a product	890	11-dec-2014	100.0
dogwood_drive is a street	892	30-dec-2014	100.0
the_news_progress is a newspaper	892	30-dec-2014	100.0
university_of_washington is a train station	892	30-dec-2014	100.0
iranian_real is a currency	892	30-dec-2014	91.5
lotronex is a drug worked on by glaxosmithkline	892	30-dec-2014	93.8
peter_finch starred in the movie network	892	30-dec-2014	100.0
bmw is a specific automobile maker dealer in tampa_bay	893	02-jan-2015	100.0
jeremy is a person who died at the age of 5	895	22-jan-2015	98.4
johannes_brahms is a person born on the date n1833	895	22-jan-2015	100.0

Recently learned beliefs (from Portuguese text)

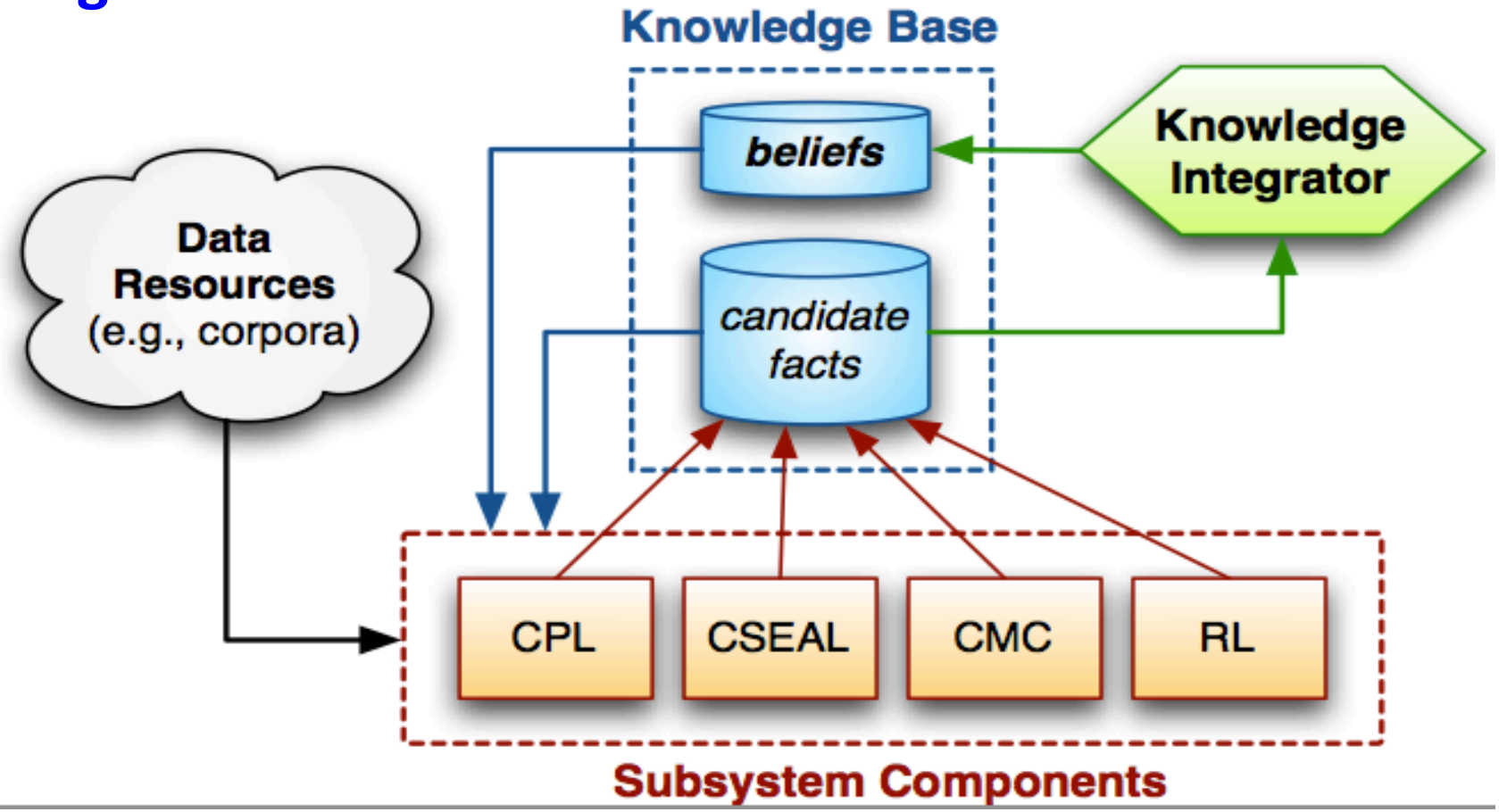
instance	iteration	date learned	confidence
william_maxwell is a South American person	105	17-oct-2016	99.4
emily_procter is an actor	105	17-oct-2016	97.7
five_petalled_flowers is a plant	105	17-oct-2016	99.3
detergent_compatible_protein_assay is a chemical	107	19-oct-2016	99.5
southern_us_ is a geopolitical entity that is an organization	106	18-oct-2016	98.3
edgar_degas is part of the visual art movement realismo	110	21-oct-2016	100.0
suzano is a sports team that plays in the league superliga_de_volei_volei	110	21-oct-2016	100.0
adobe has acquired macromedia	110	21-oct-2016	99.9
ryan_giggs plays for the team manchester_united	110	21-oct-2016	100.0
ponta_grossa is a city located in the state or province paran	110	21-oct-2016	99.2

Carnegie Mellon University

How to Read the Web in Many Languages?

NELL: Never-Ending Language Learner

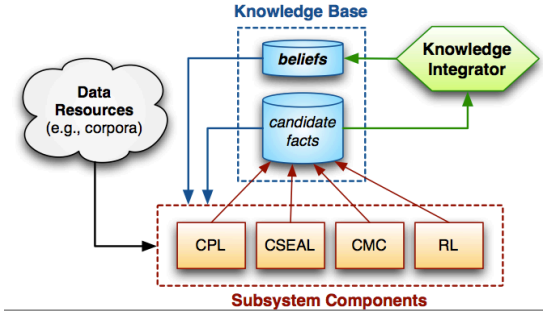
English Version



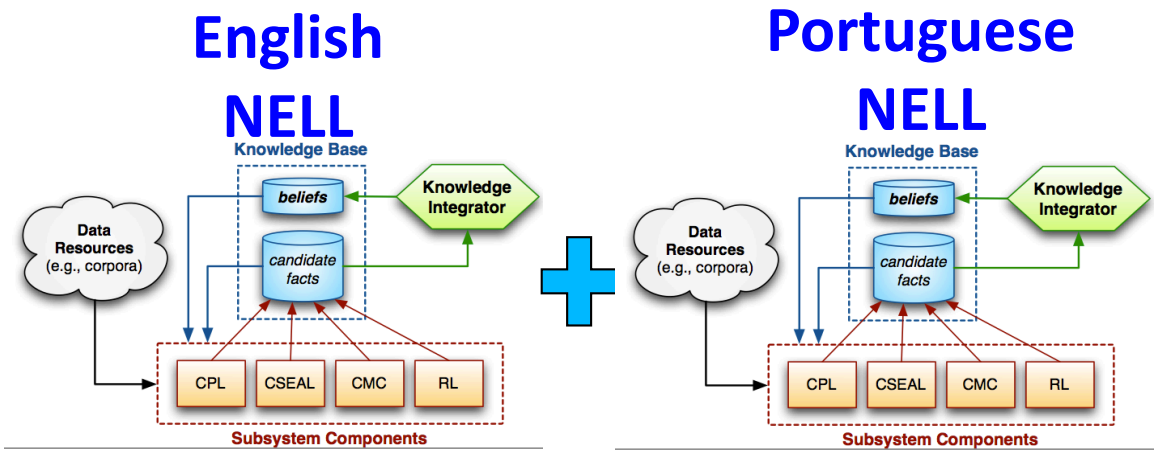
NELL: Never-Ending Language Learner

English

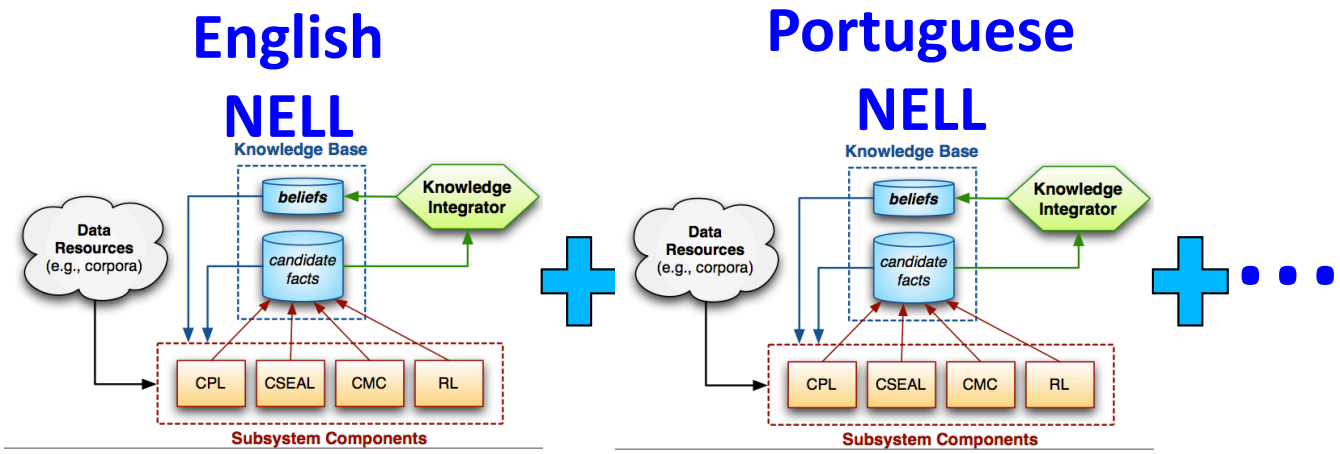
NELL



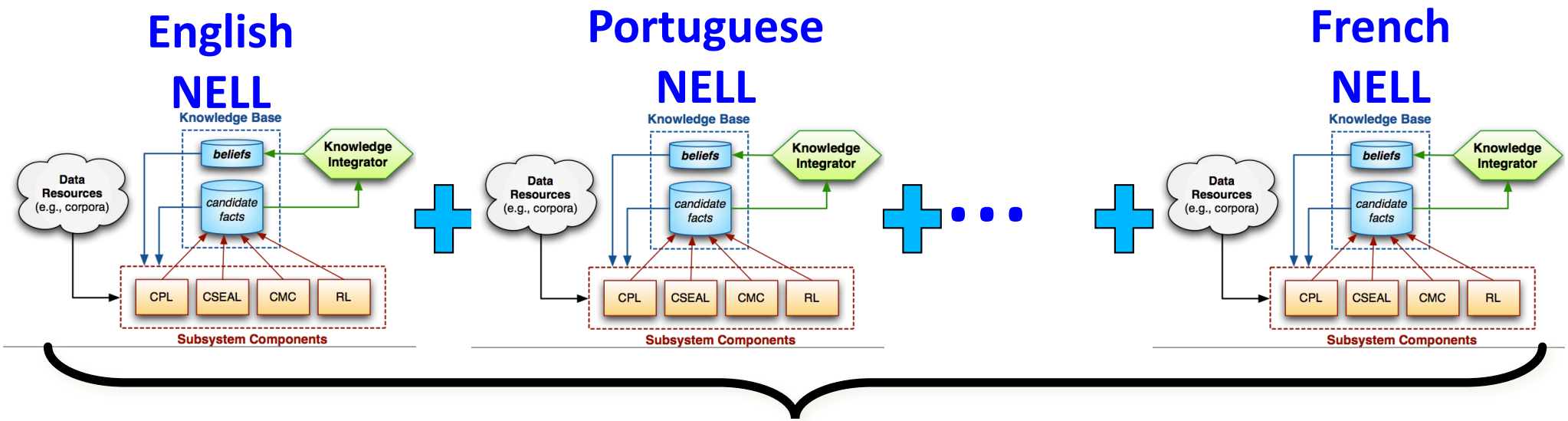
NELL: Never-Ending Language Learner



NELL: Never-Ending Language Learner

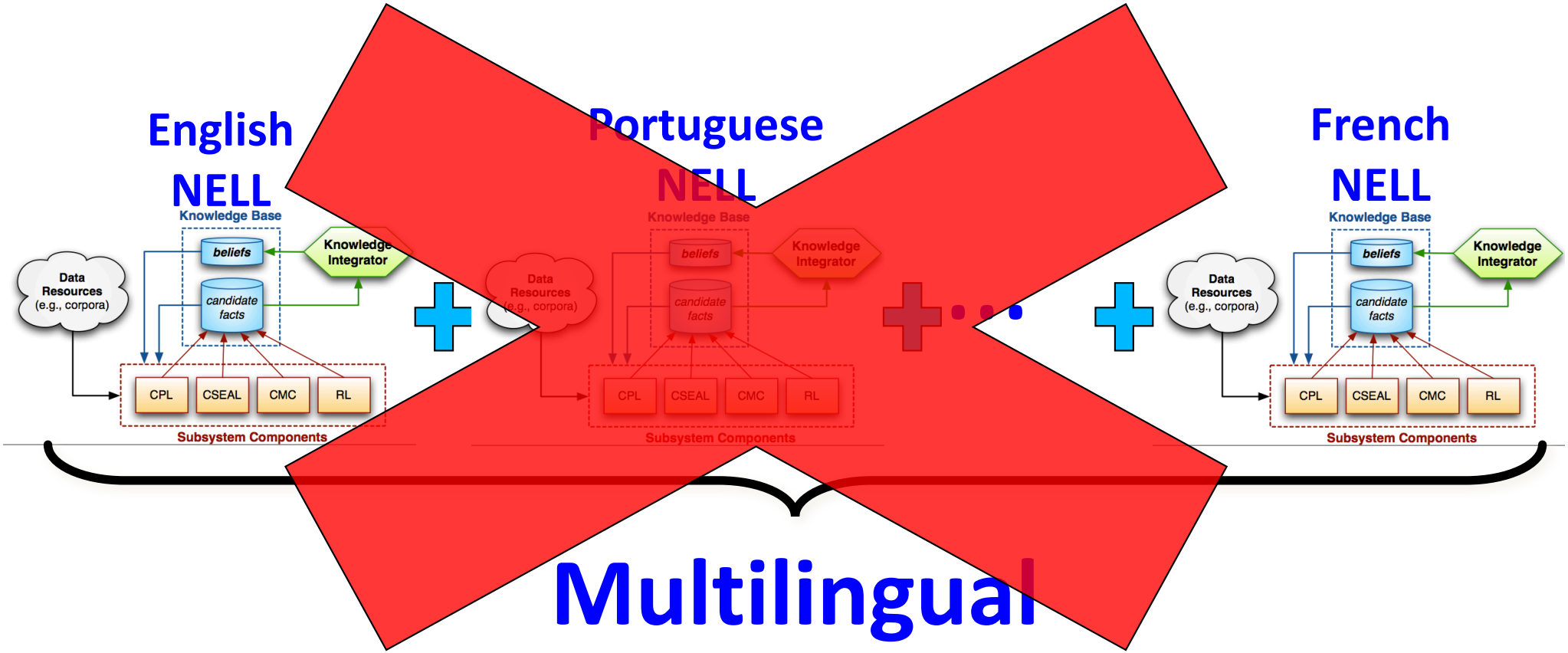


NELL: Never-Ending Language Learner



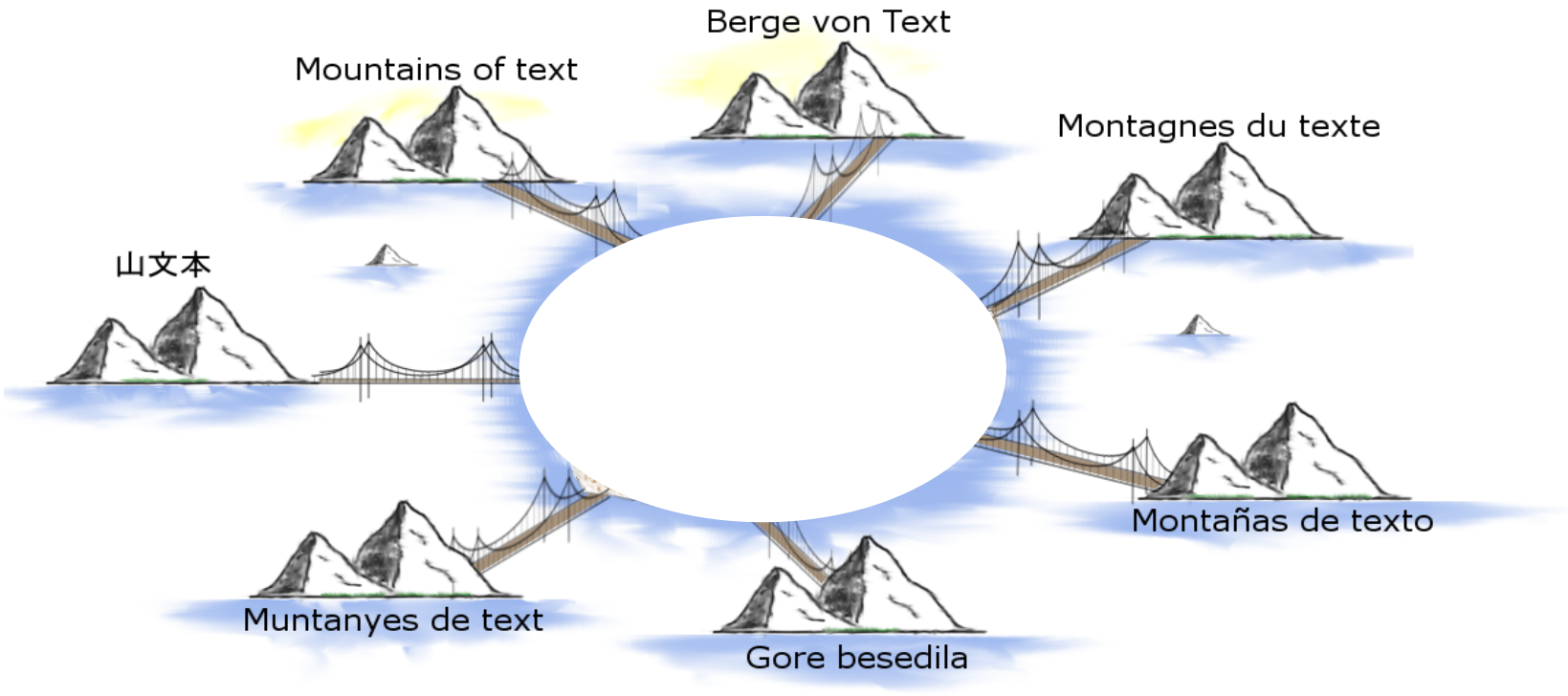
Multilingual NELL

NELL: Never-Ending Language Learner



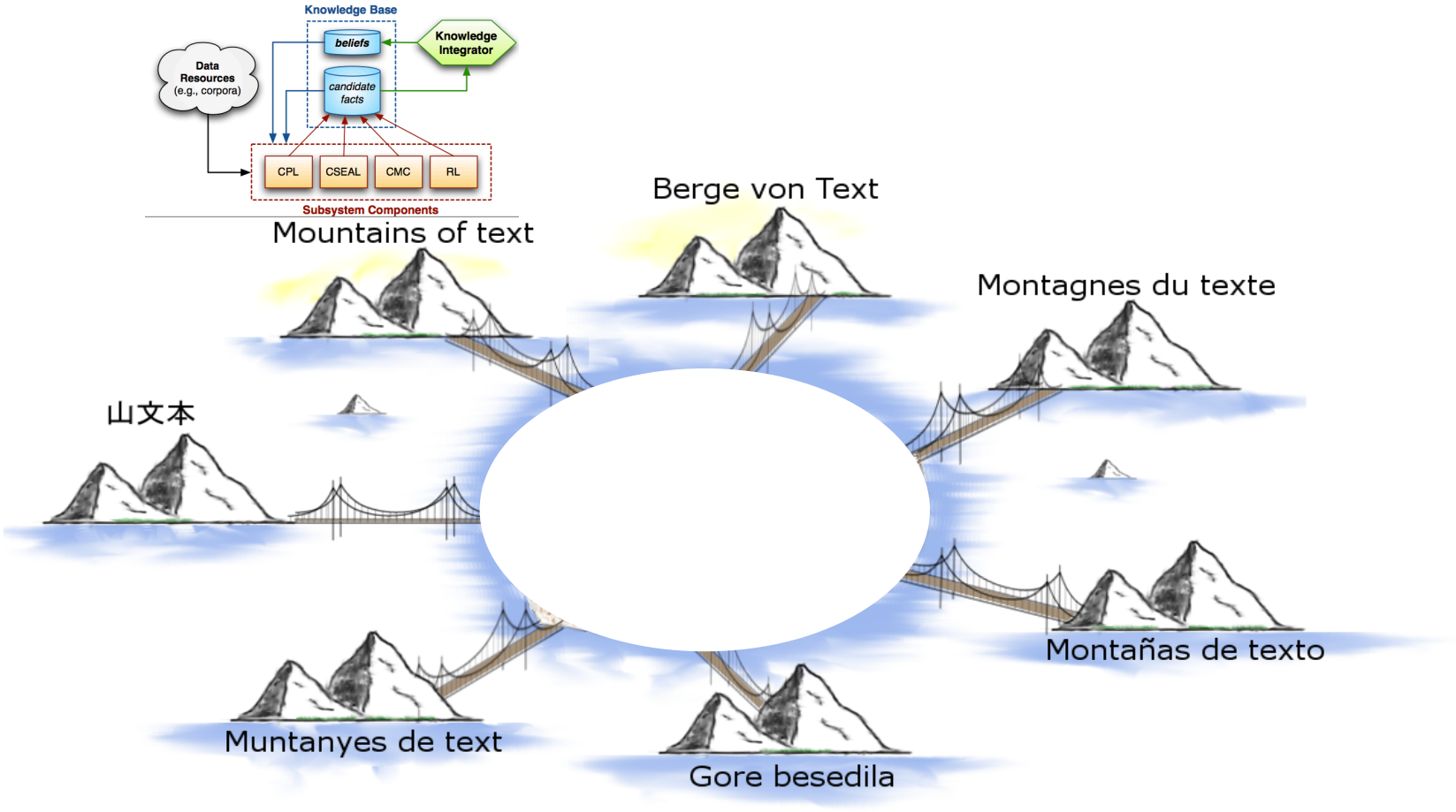
Multilingual NELL

Multilingual Reading The Web



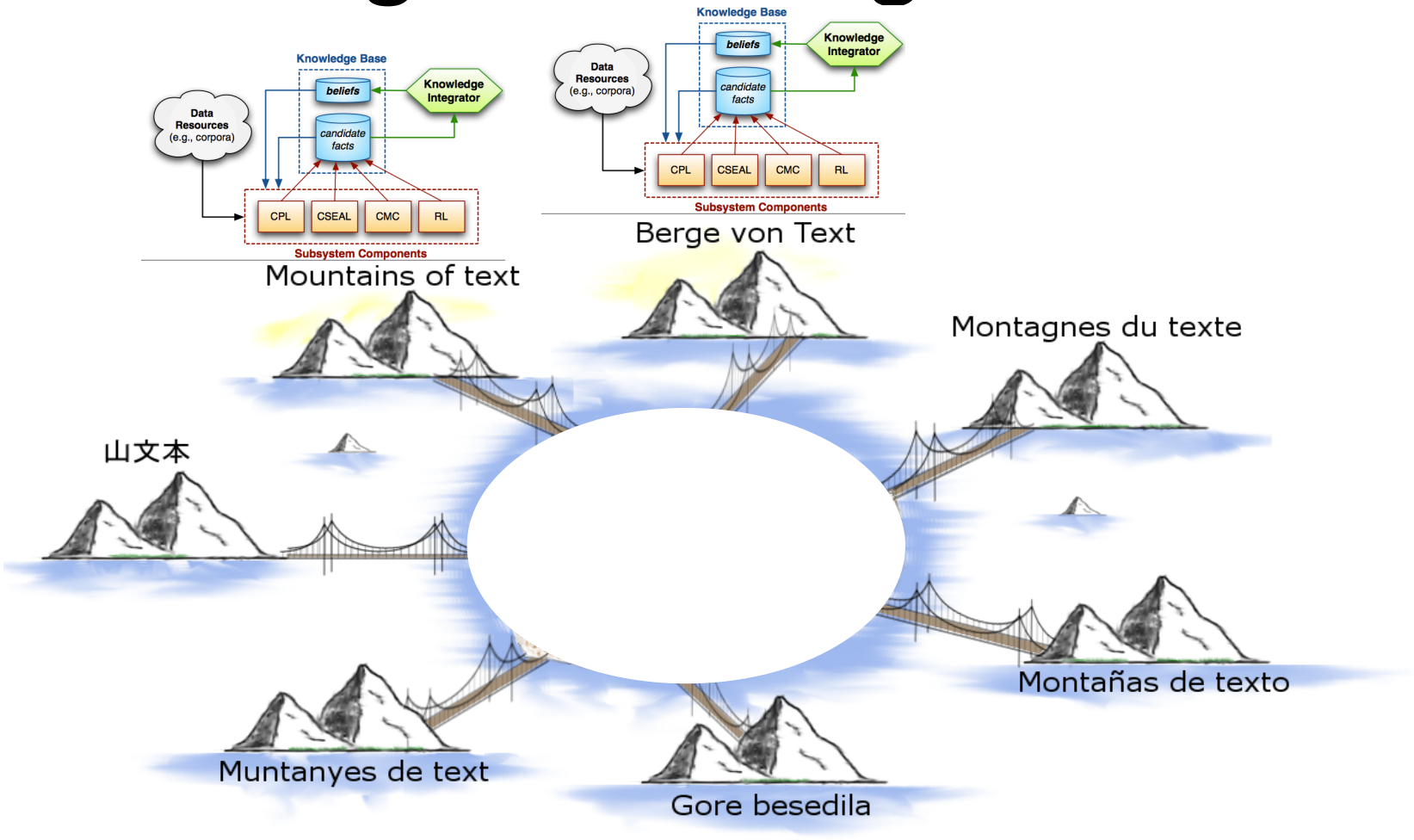
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



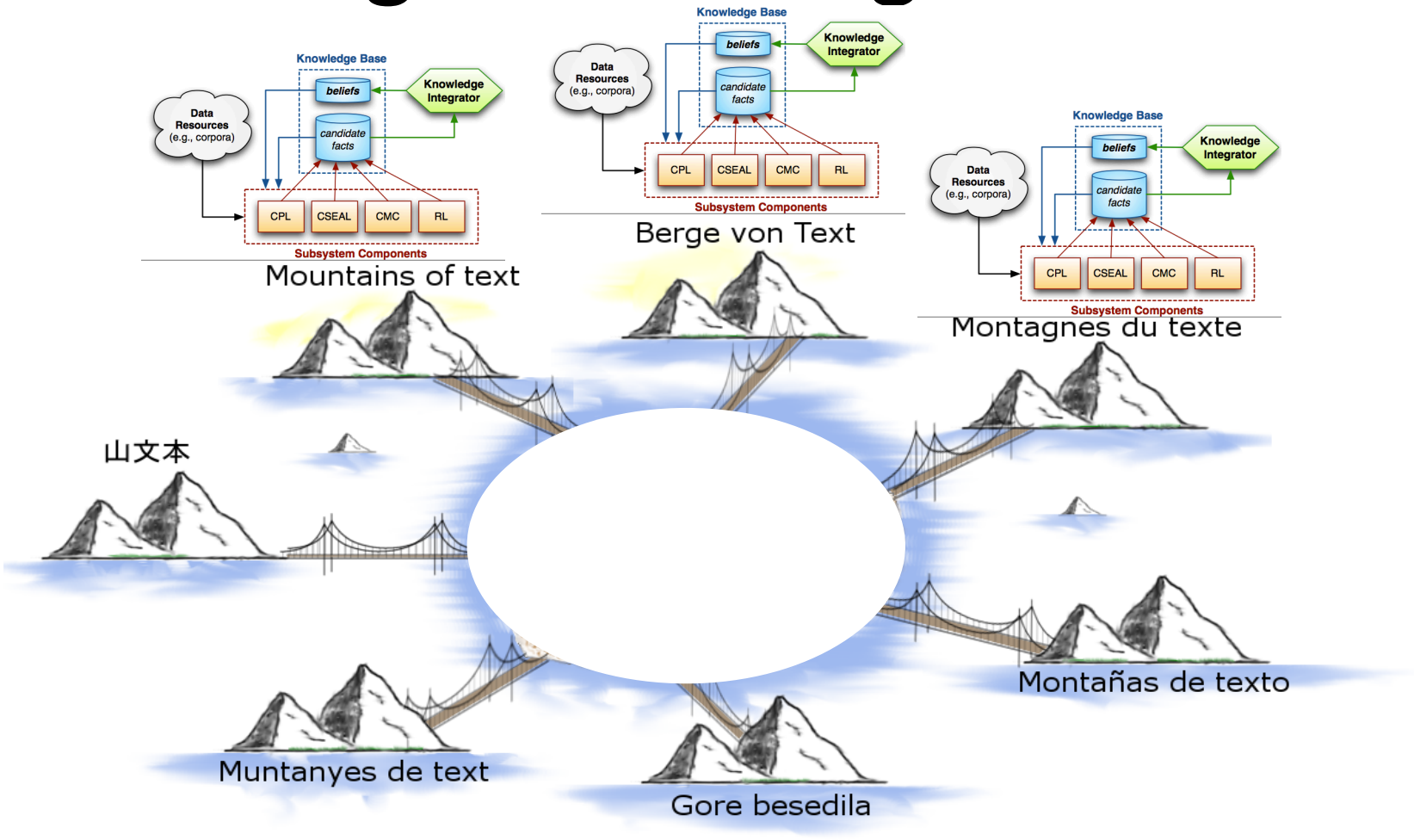
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



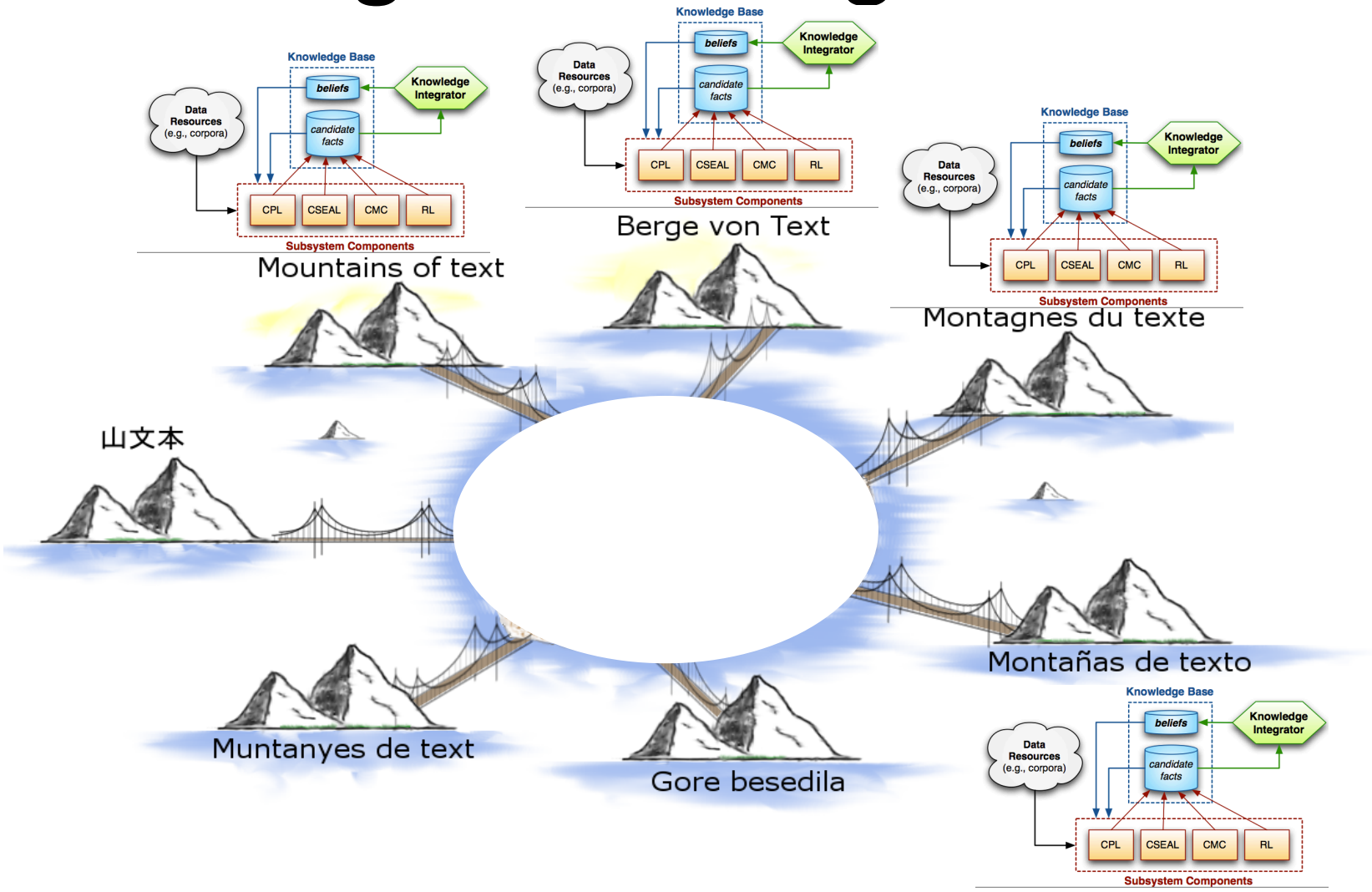
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



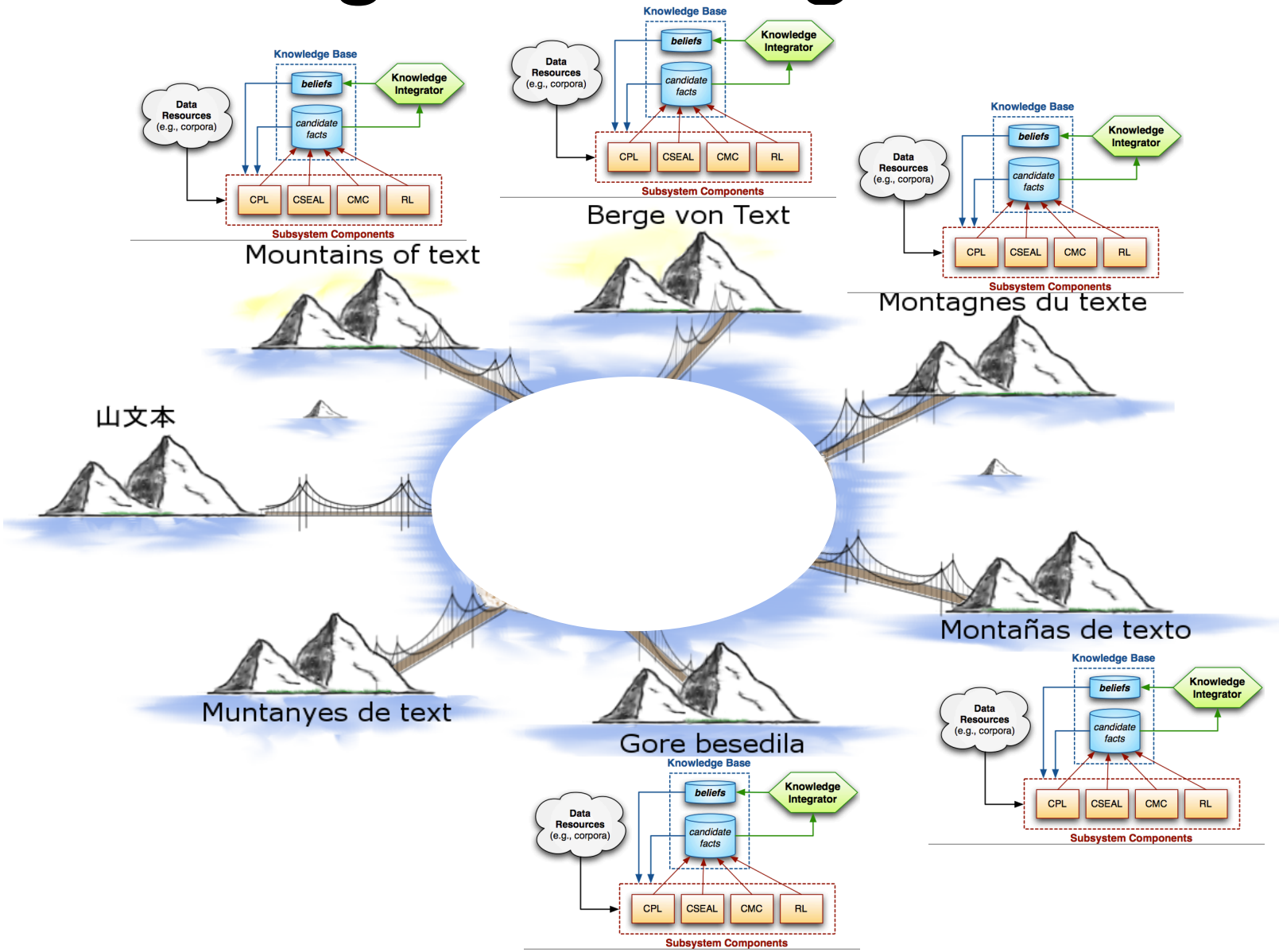
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



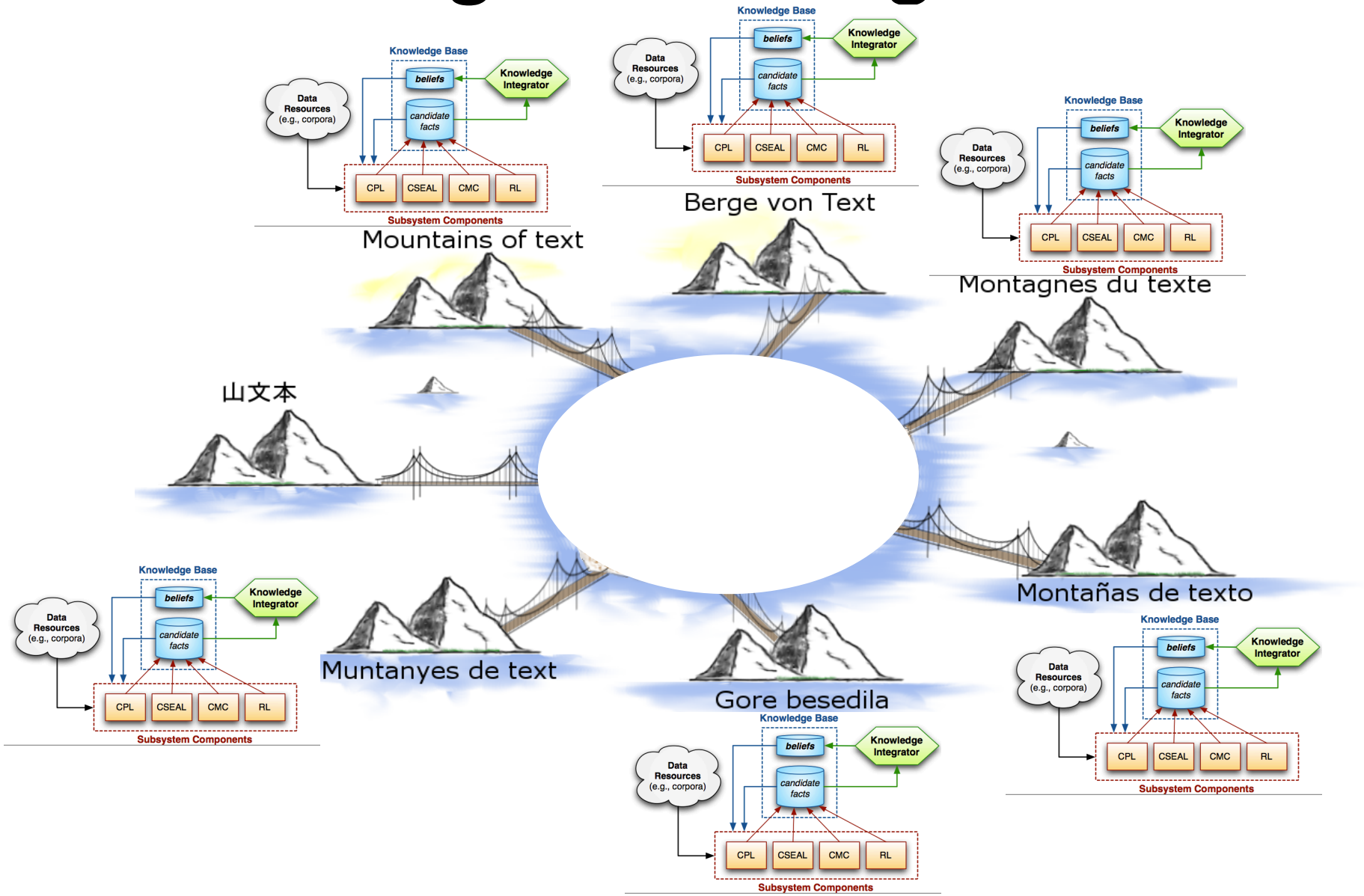
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



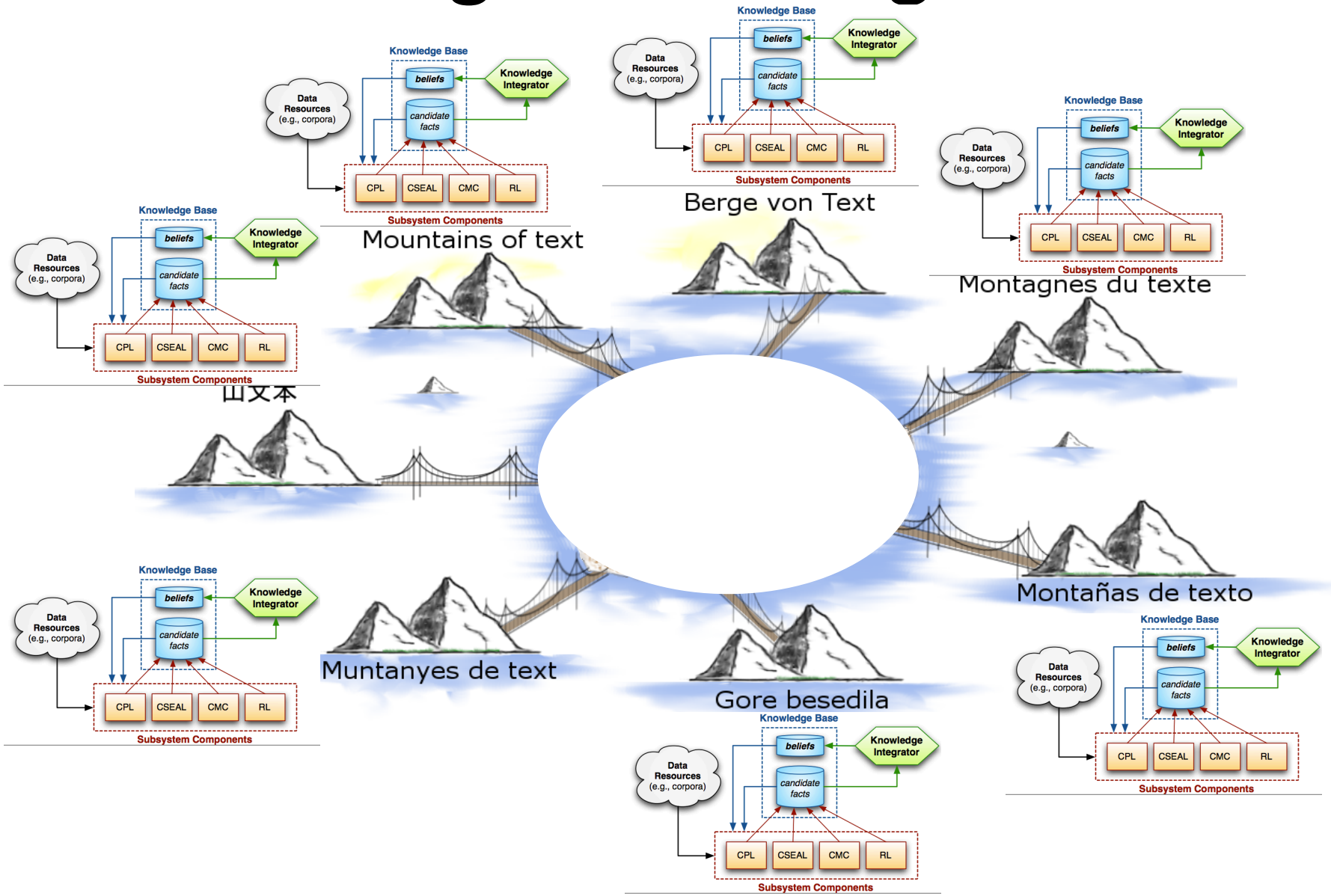
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



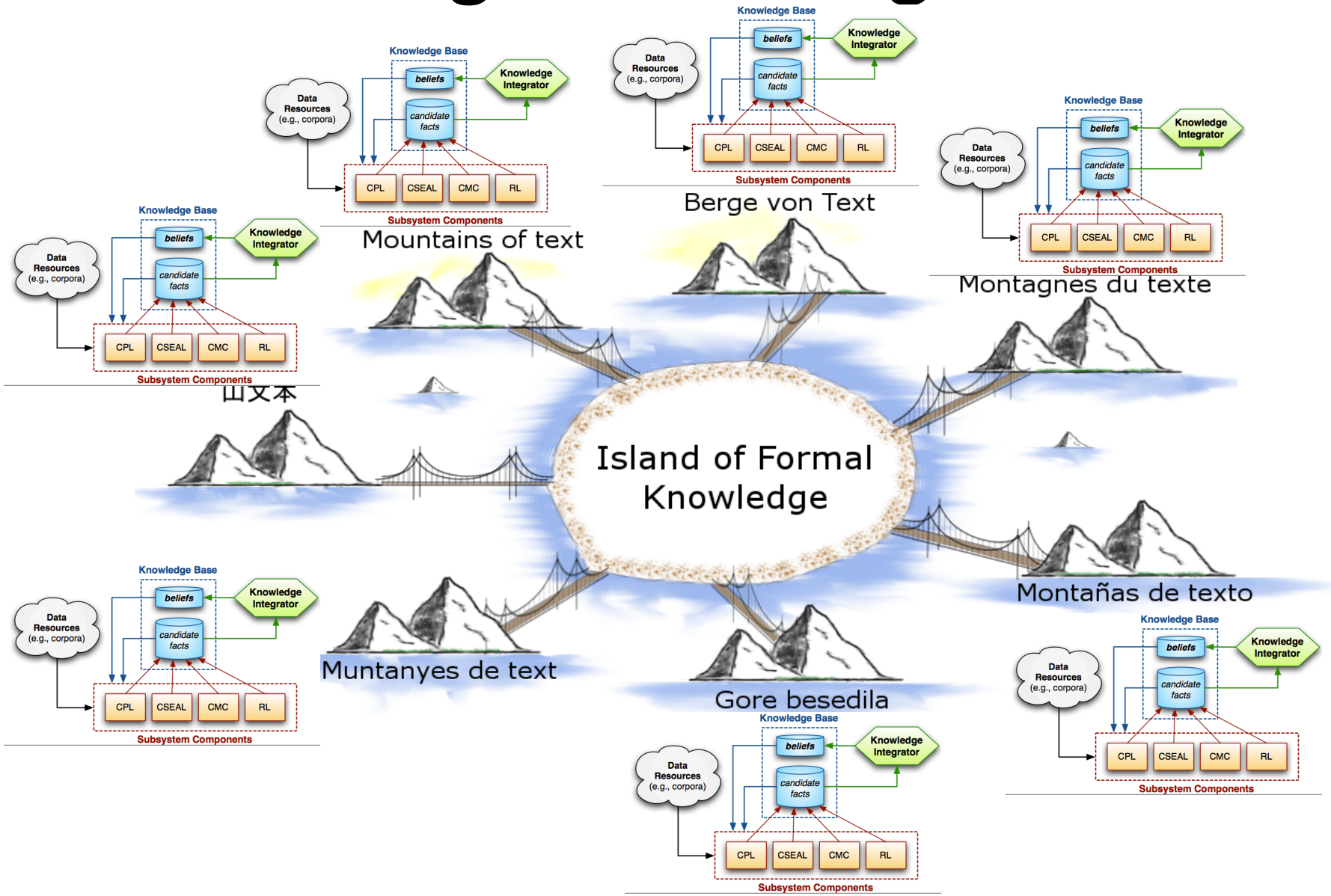
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



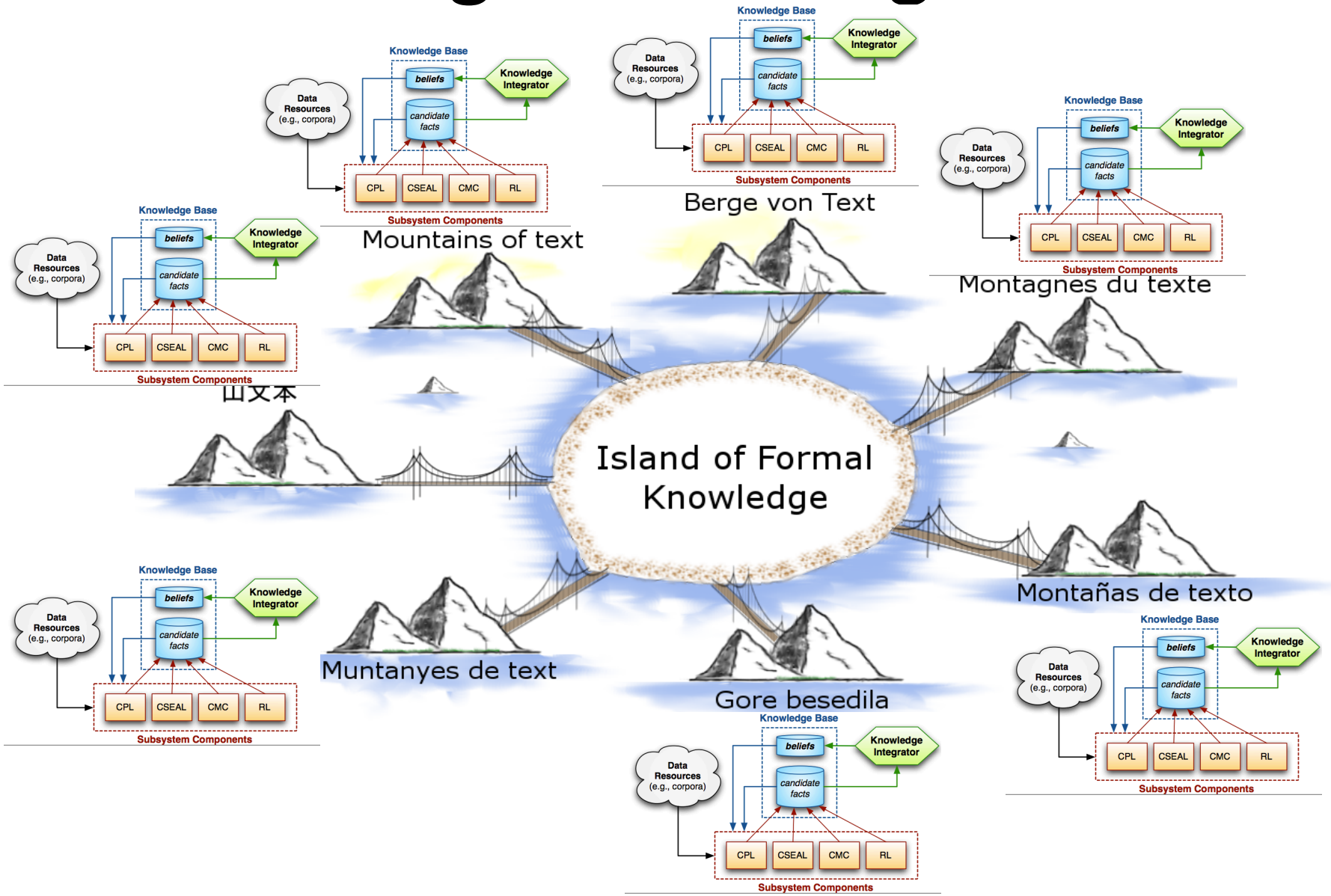
Picture Adapted from http://www.xlike.org

Multilingual Reading The Web



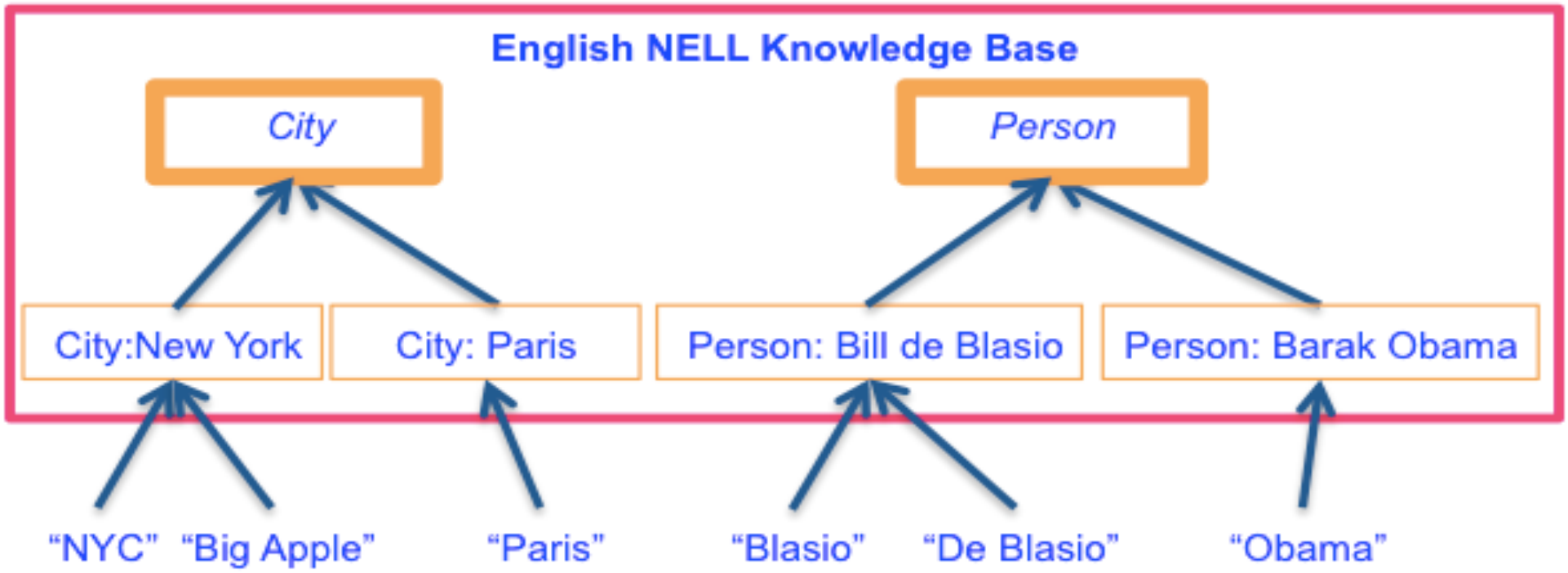
Picture Adapted from http://www.xlike.org

Crosslingual Reading The Web

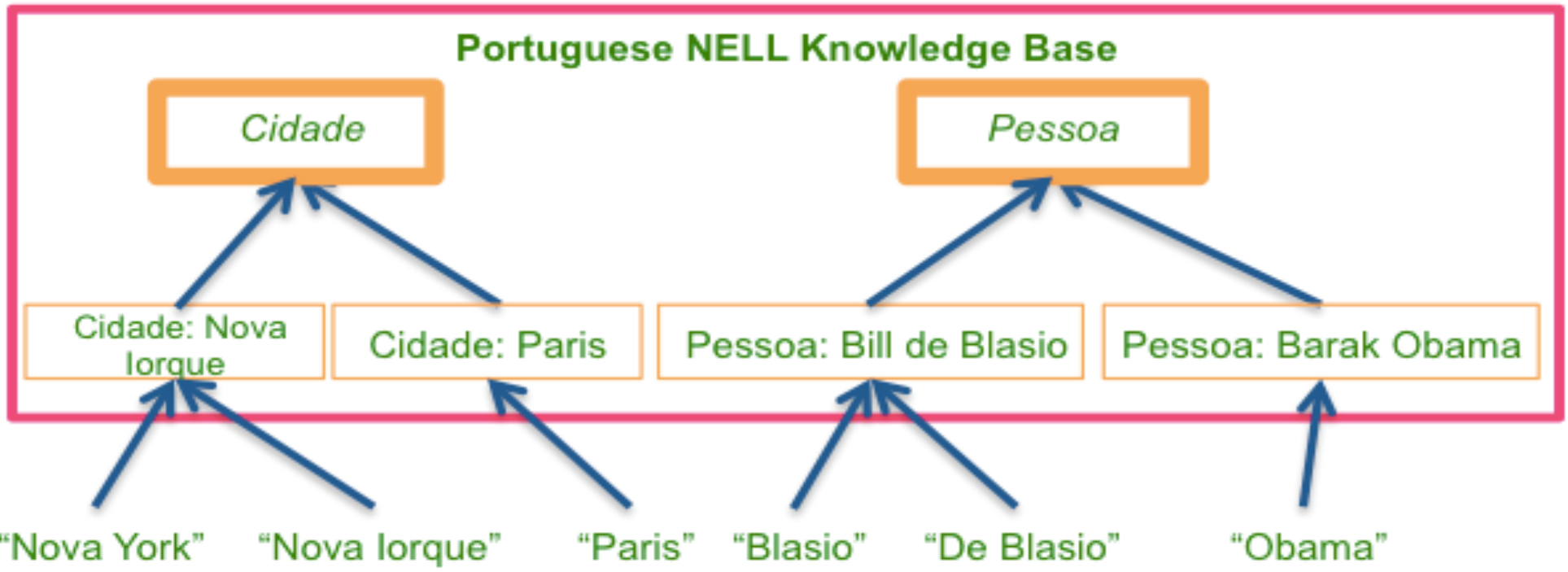


Picture Adapted from http://www.xlike.org

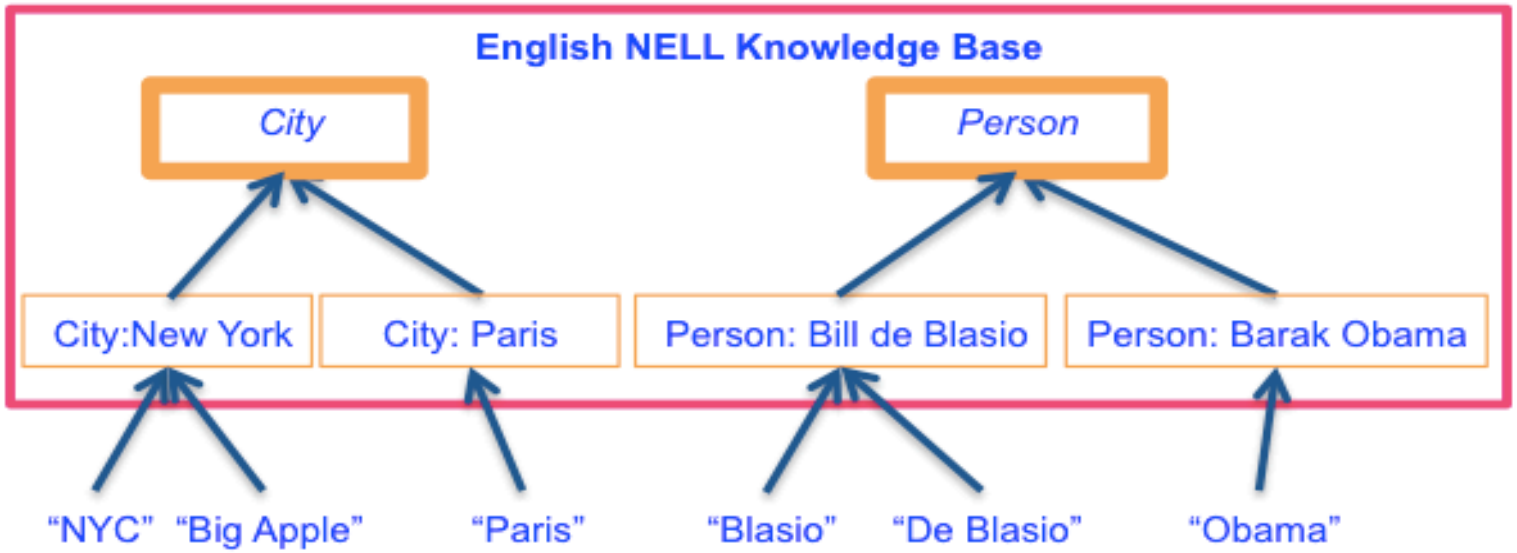
Crosslingual Reading The Web

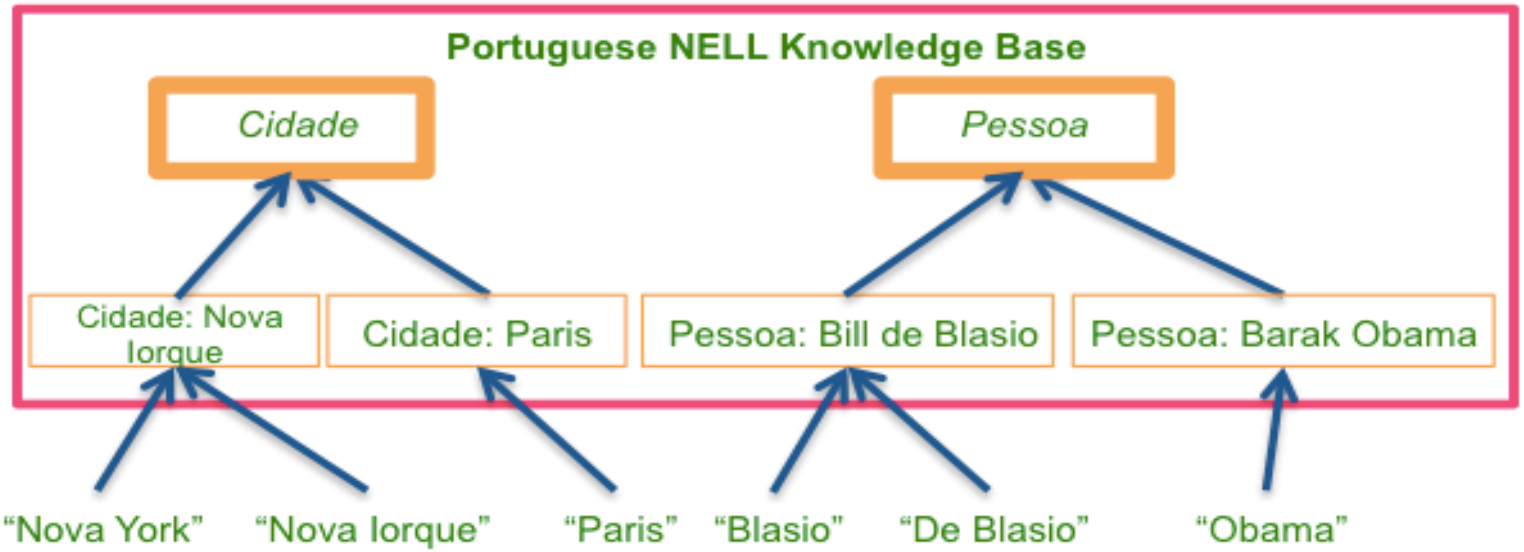


Crosslingual Reading The Web

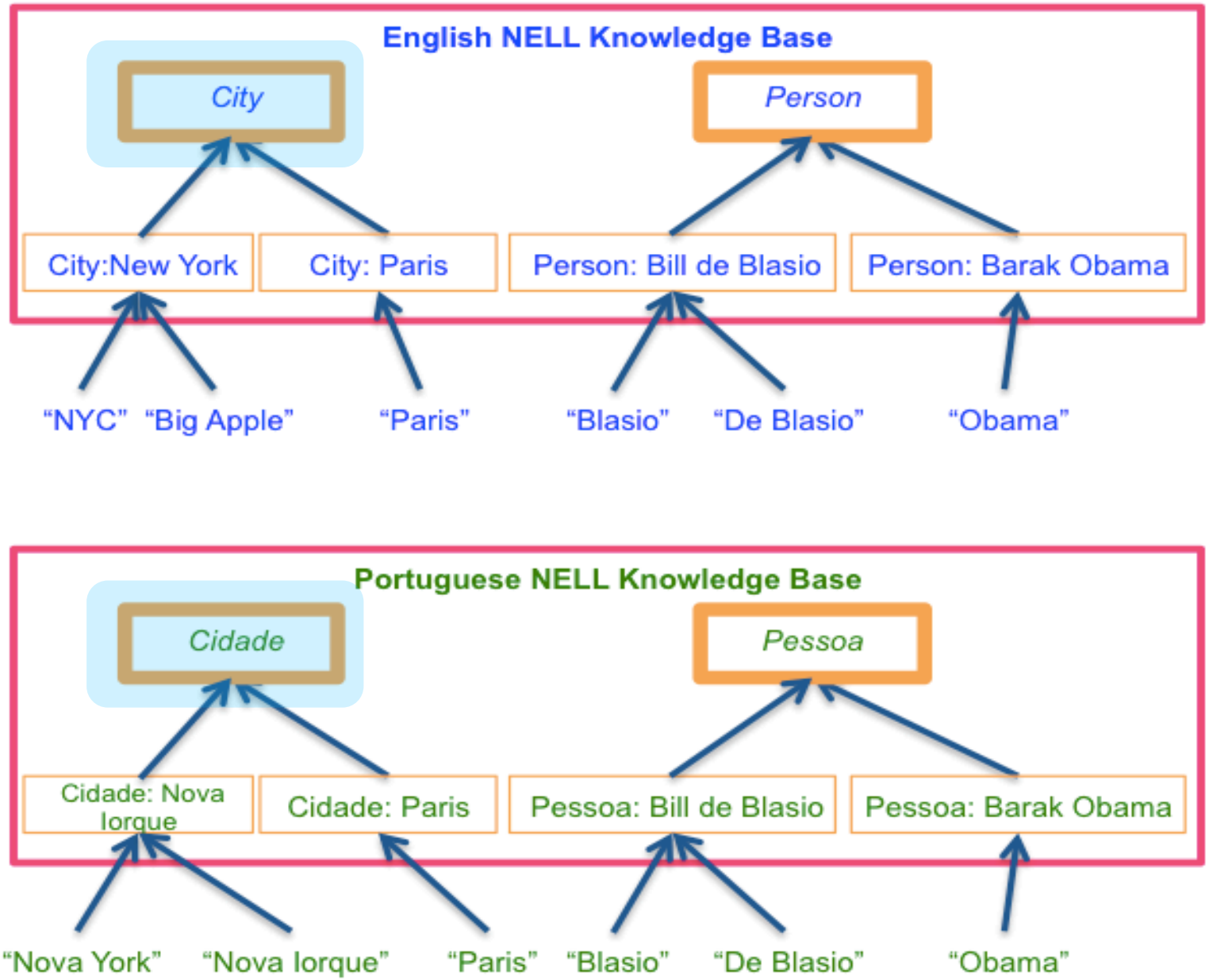


Crosslingual Reading The Web

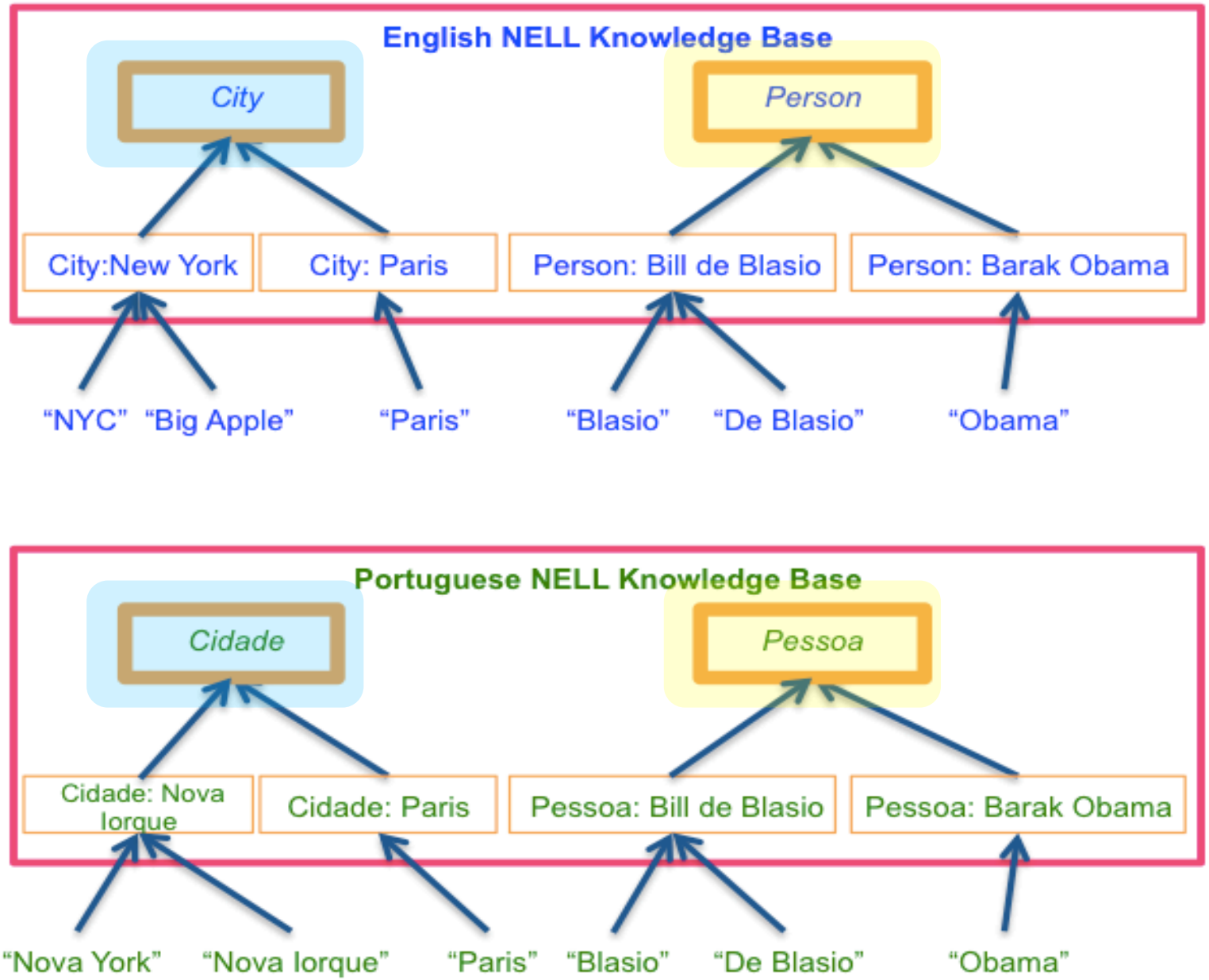




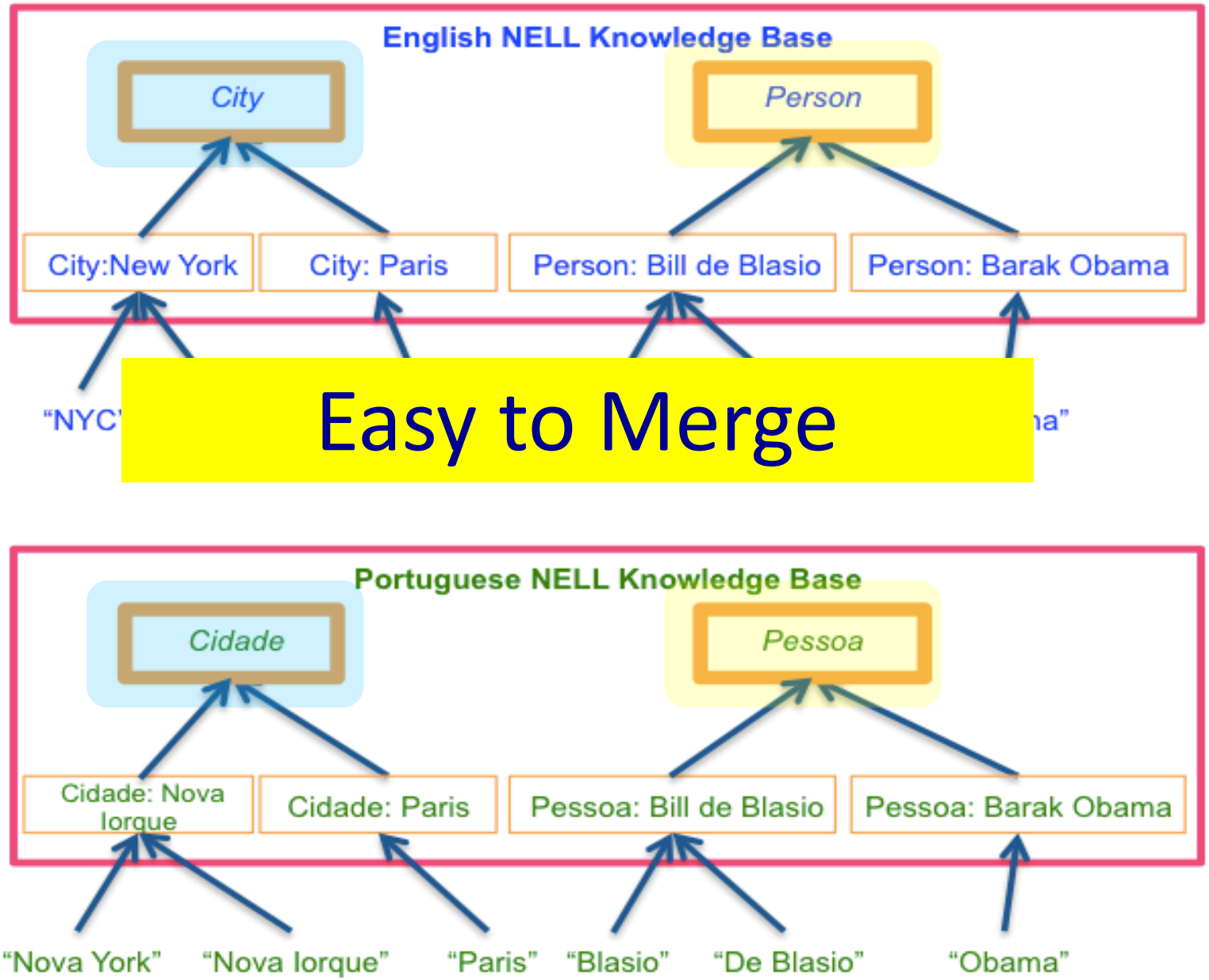
Crosslingual Reading The Web



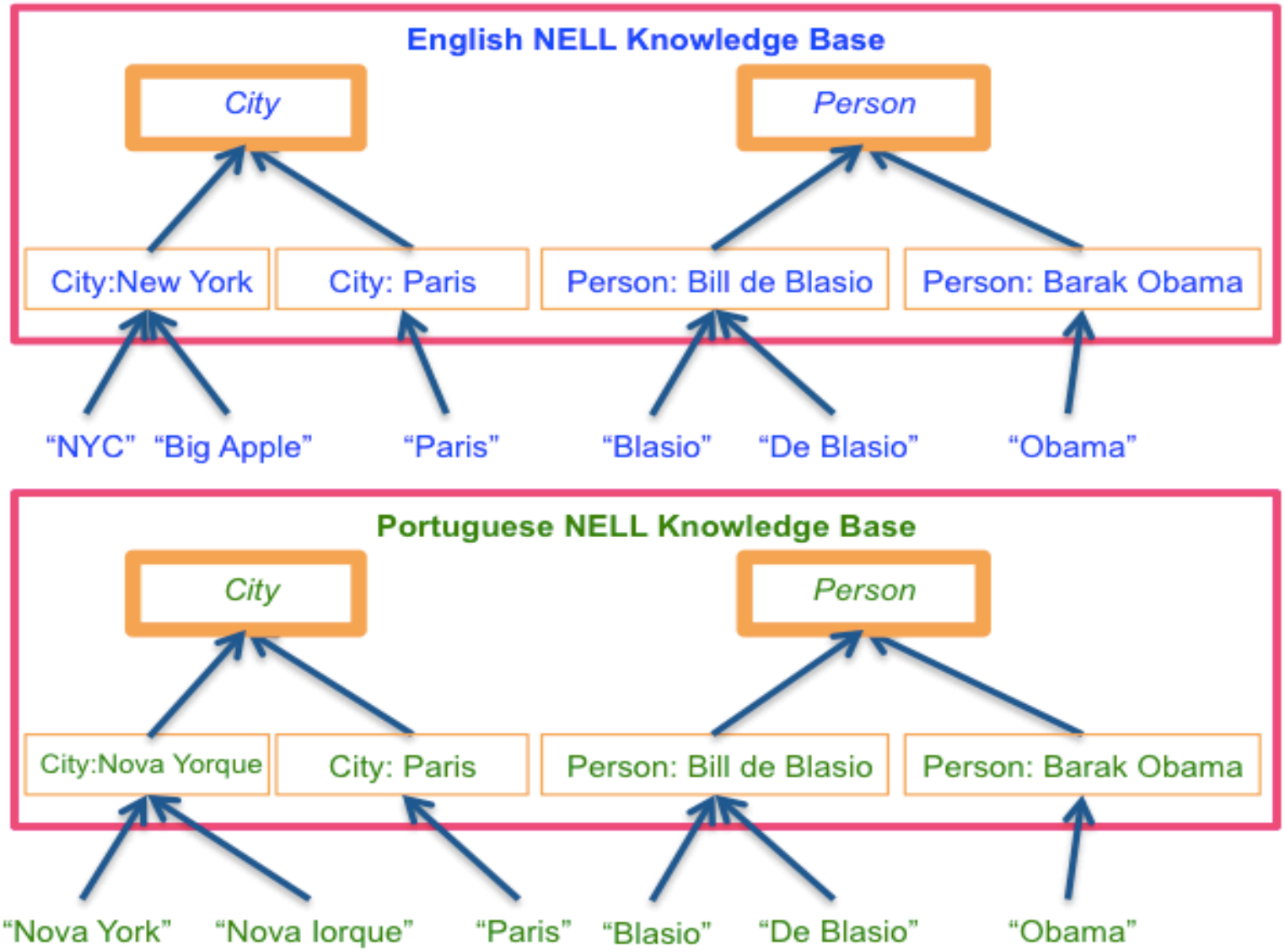
Crosslingual Reading The Web



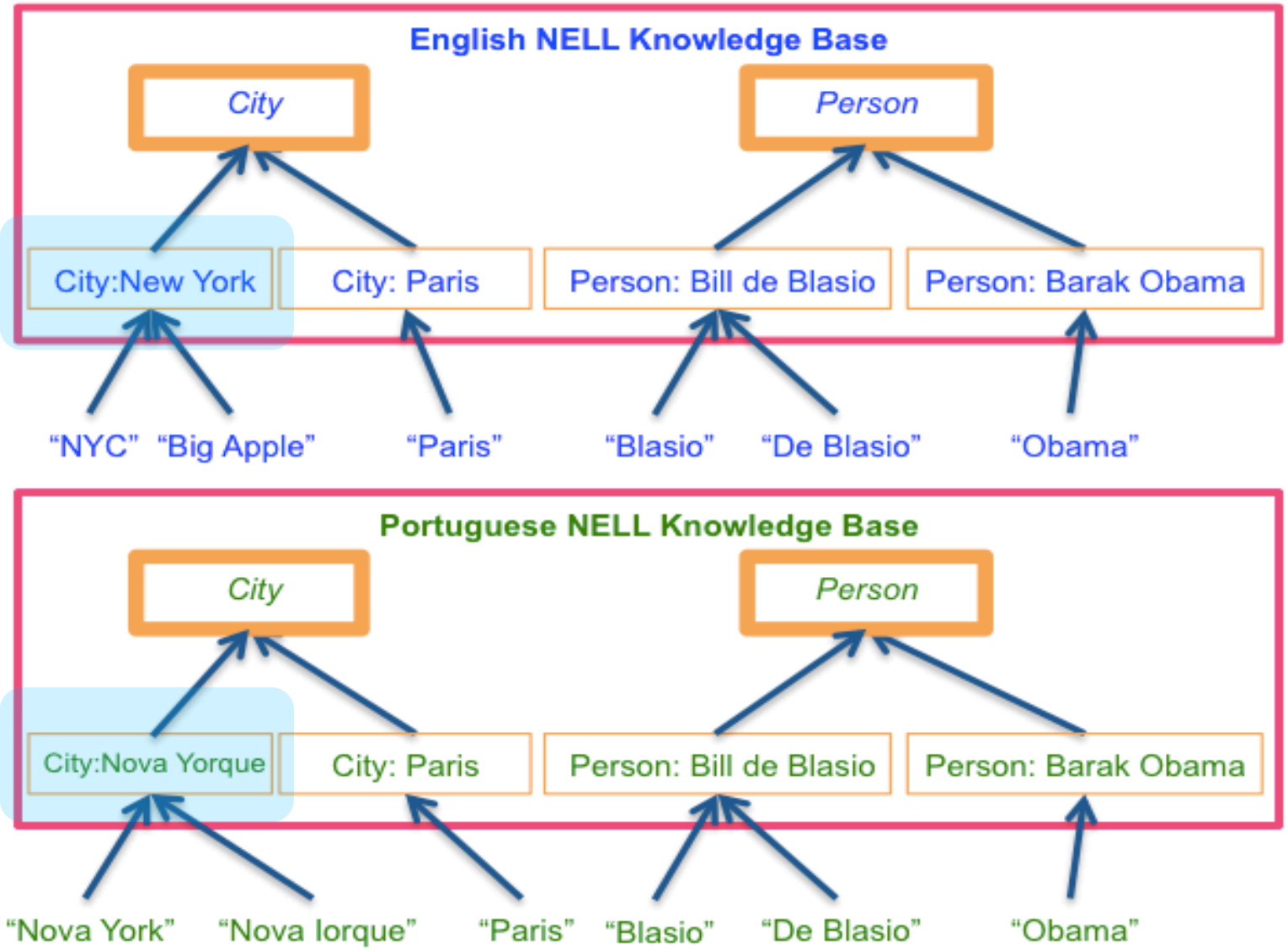
Crosslingual Reading The Web



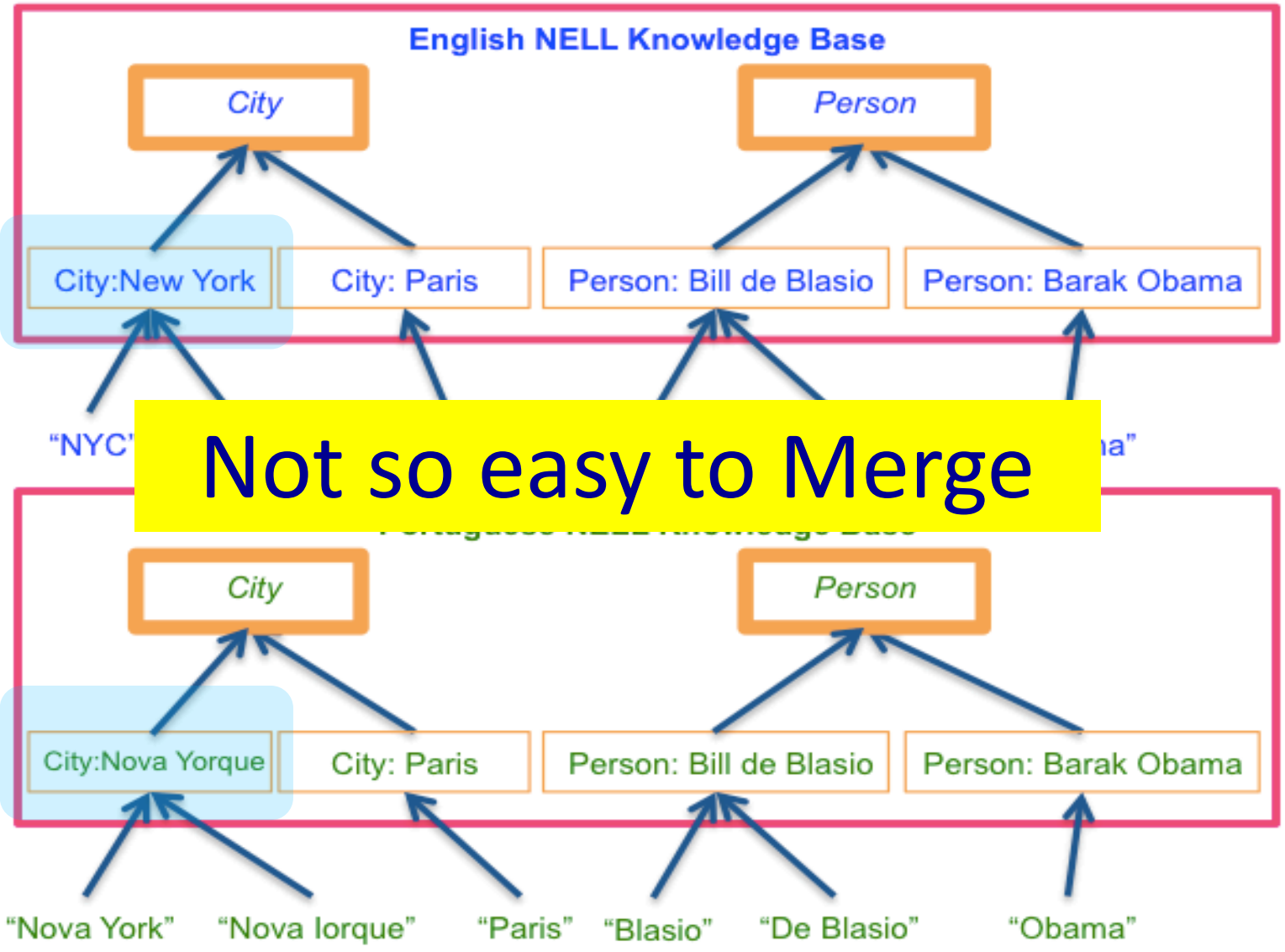
Crosslingual Reading The Web



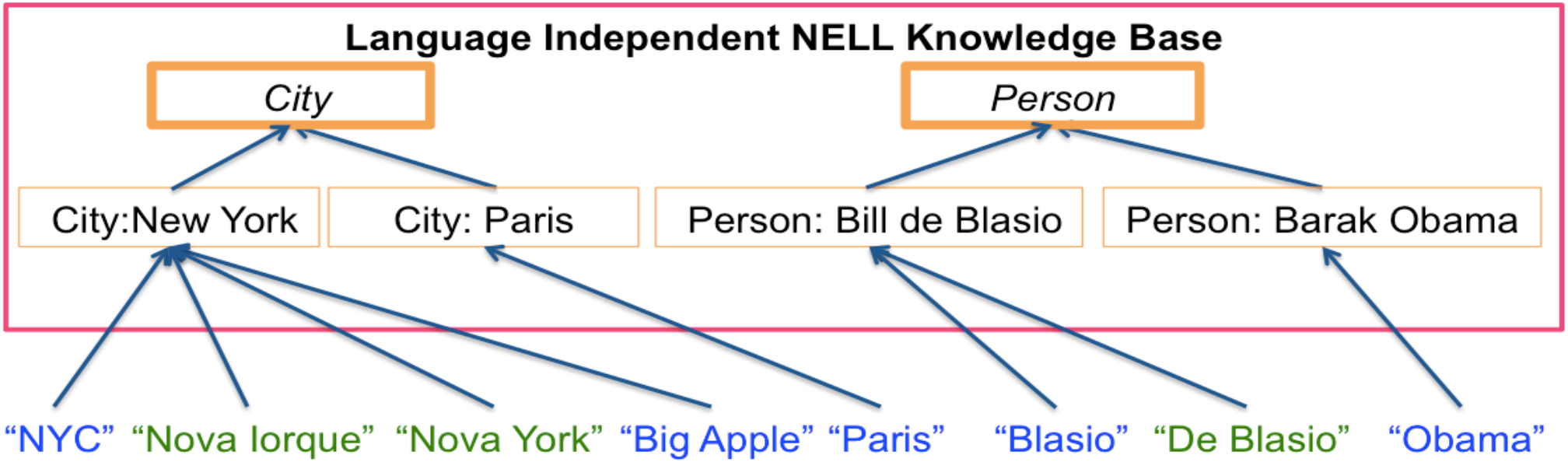
Crosslingual Reading The Web



Crosslingual Reading The Web



Crosslingual Reading The Web



Federal University of São Carlos

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

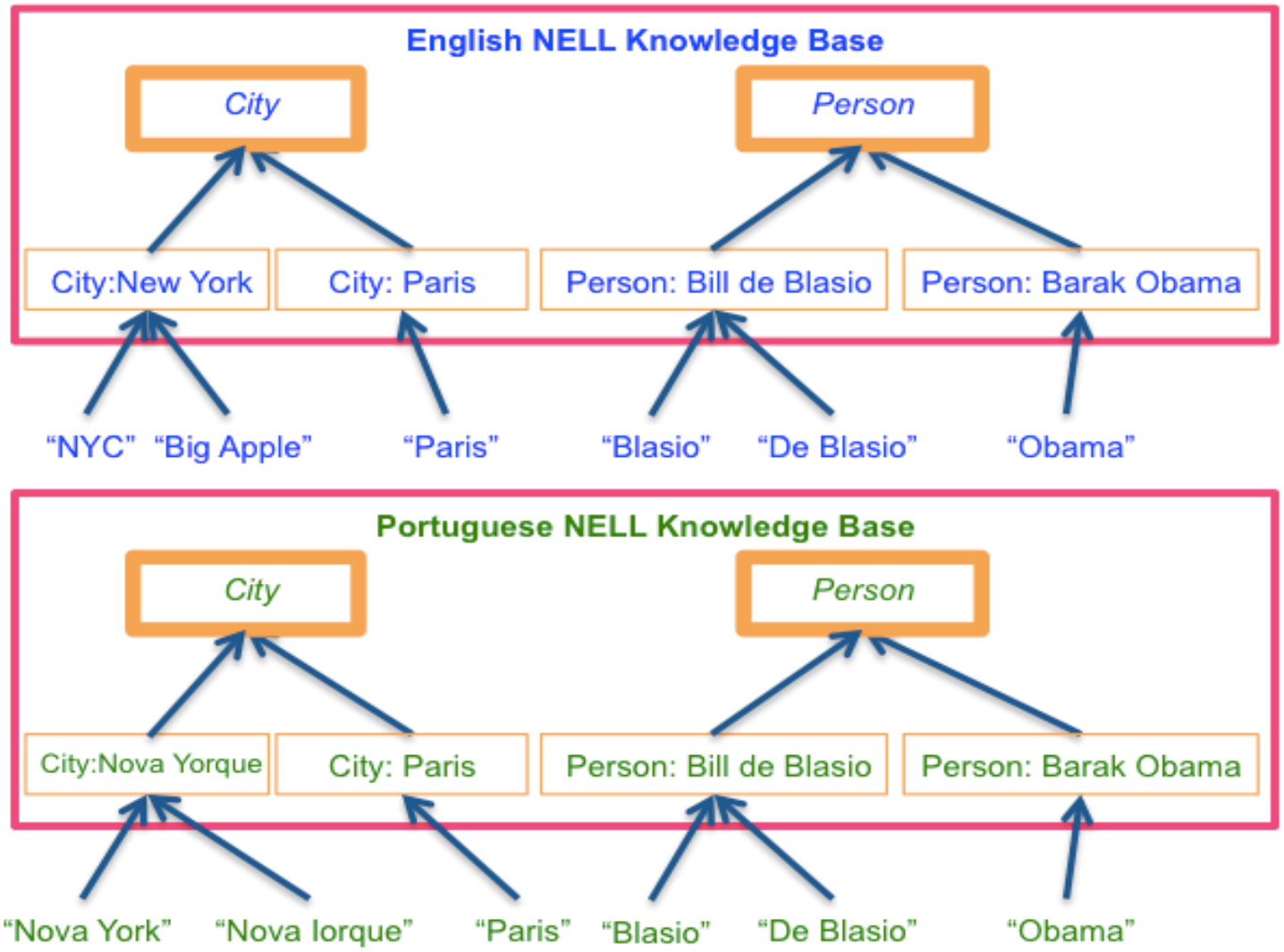
Input:

Carnegie Mellon University

Merging Concepts Learned in Different Languages

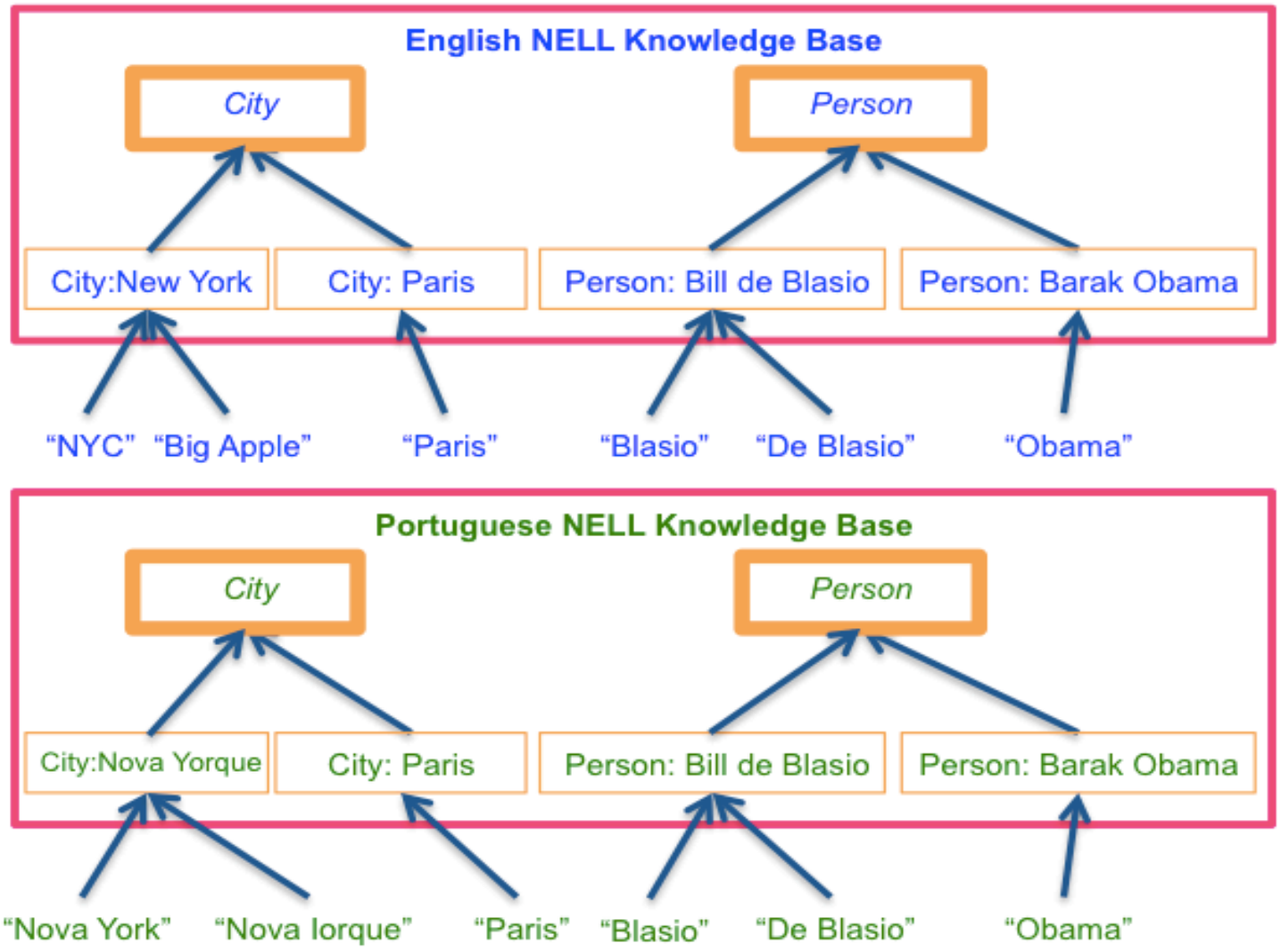
Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

Input:



Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

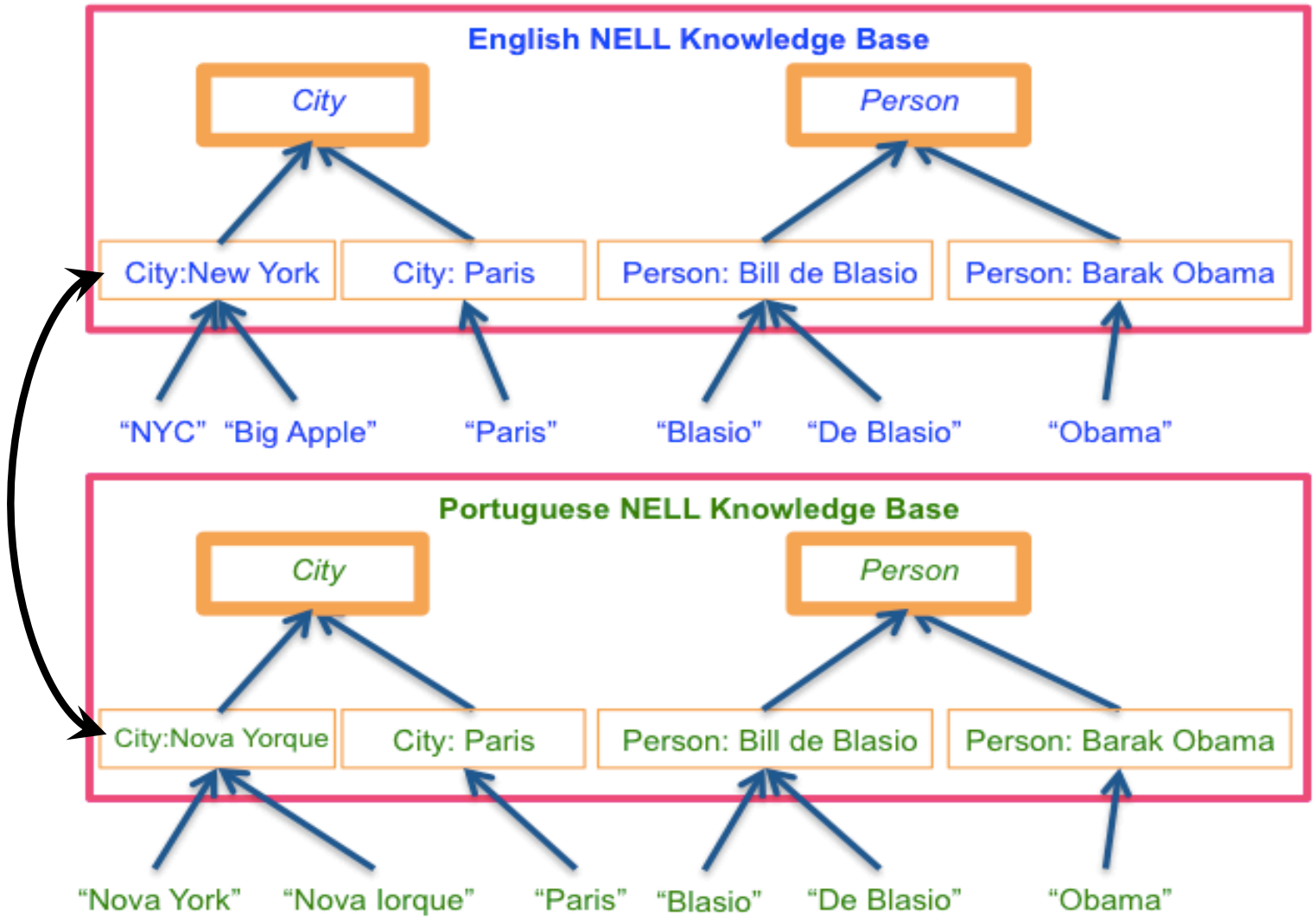


Input:

Positive Examples:
About 4000 manually
Defined equivalence relations

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016



Input:

EquivalentTo

Positive Examples:
About 4000 manually Defined equivalence relations

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

Input: Negative Examples: 8000 automatically defined non-equivalence relations

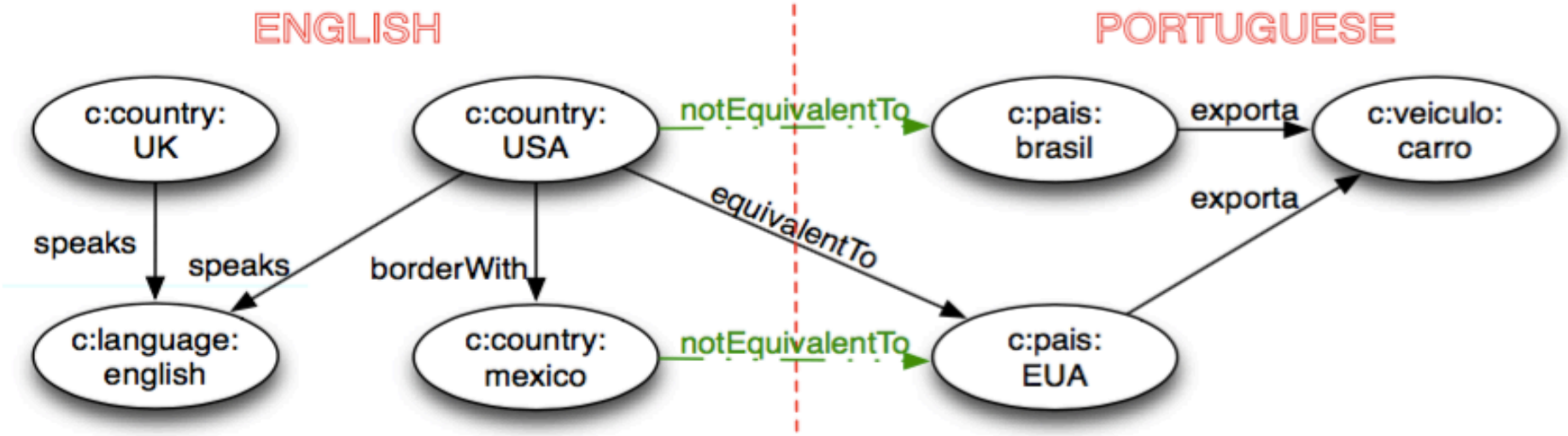
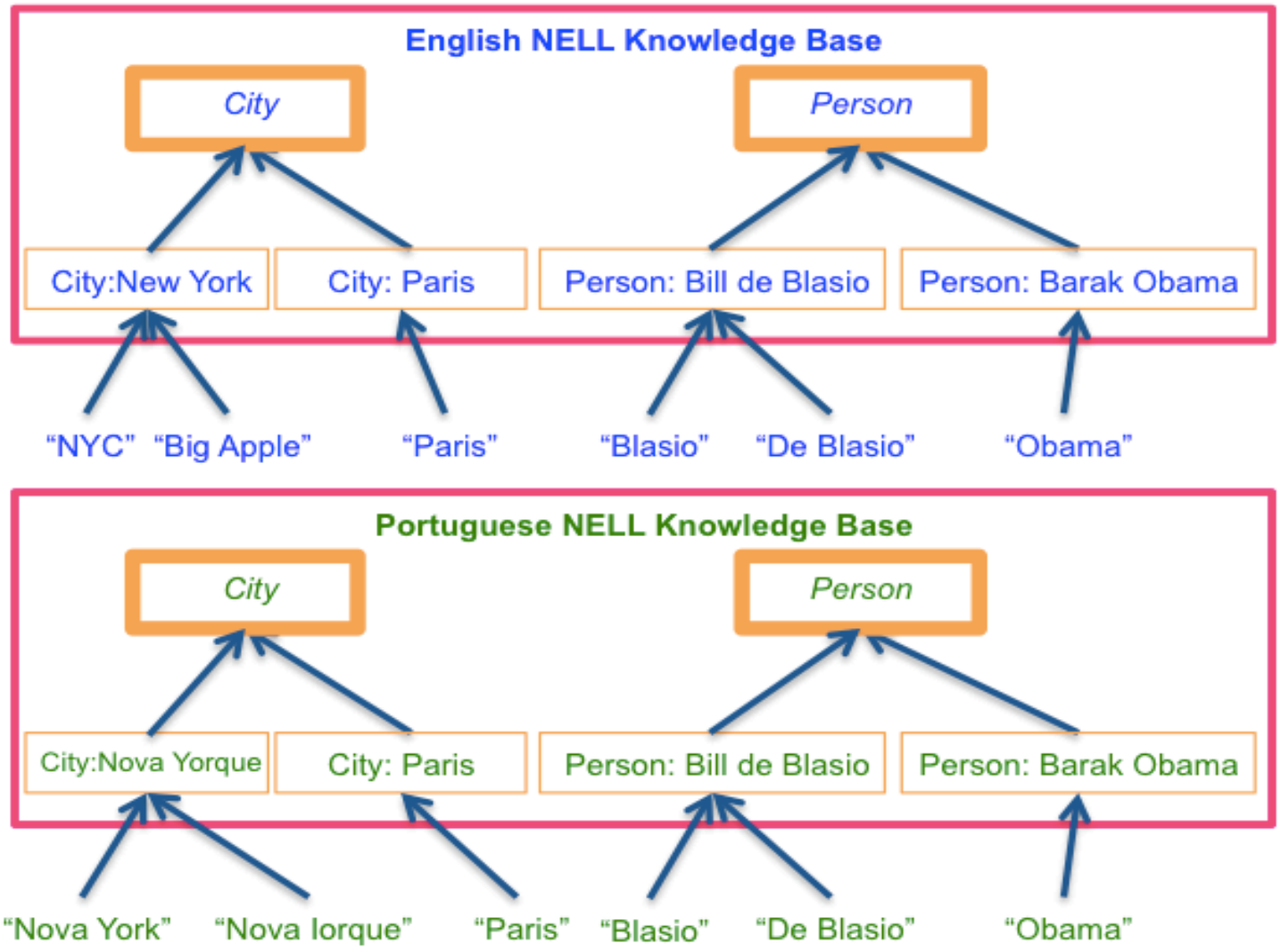


Figure 1: Graphical representation of the generation of negative examples. Given a positive pair (*USA*, *equivalentTo*, *EUA*), the relationship (*USA*, *borderWith*, *Mexico*) is followed to reach an entity with the same category (*Mexico*, *Country*), generating the negative example (*Mexico*, *notEquivalentTo*, *EUA*).

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

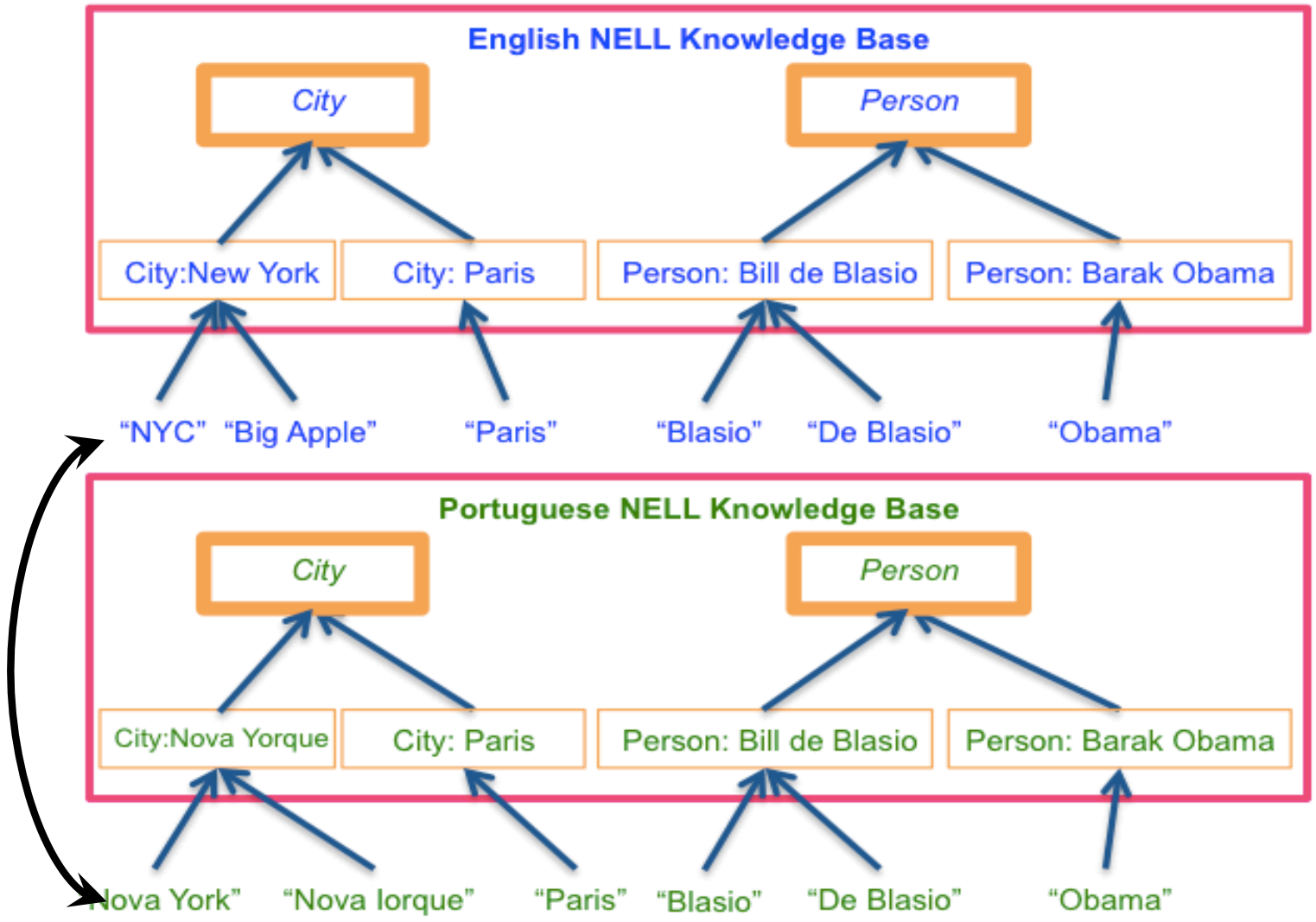


Input:

1.4M automatically translated literal strings pairs

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016



Input:

1.4M automatically translated literal strings pairs

CanBeTranslatedAs

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

ML Task:

binary classifier (yes/no) for a given "equivalentTo" Relation candidate

The Classifier:

If

sameCategory(concept1,concept2), and
differentLang(concept1,concept2), and
higherThanTreshold(score)

then, yes

Federal University of São Carlos

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

3 Approaches:

Personalized Page Rank – PPR

Path Ranking Algorithm – PRA

Dictionary-based – Baseline

Carnegie Mellon University

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

PPR configuration:

- 2,000 random walks
- walk length: 5
- probability of restart: 0.01

PRA configuration:

- breadth-first search
- depth: 2

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

	English						Portuguese					
Category	In graph			In positive pairs			In graph			In positive pairs		
animal	12436	591	746	36	23	27	101	63	97	36	14	35
country	6031	443	460	106	93	93	153	94	136	106	87	103
city	18893	4437	5311	460	237	263	5767	483	1404	460	138	282
movie	7008	712	831	42	38	38	368	64	132	42	25	40
person	6693	2898	3050	403	395	399	621	611	614	403	304	376
writer	18911	1707	2143	61	39	40	114	26	63	61	23	37
actor	28361	794	1421	512	139	167	1870	129	793	512	36	208
sport	5022	205	381	109	66	75	153	34	125	109	29	94
global	1909339	66239	96086	4126	2112	2331	30401	5119	12930	4126	1827	2565

Table 1: For both language subgraphs, the total number of nodes in a graph and the number of nodes included in positive pairs are shown. The three columns in each vertical subdivision show counts for the original graph, the pruned graph and the SVO+pruned graph, respectively from left to right. Each line shows the counts for the considered categories (in the last line, all the categories).

Merging Concepts Learned in Different

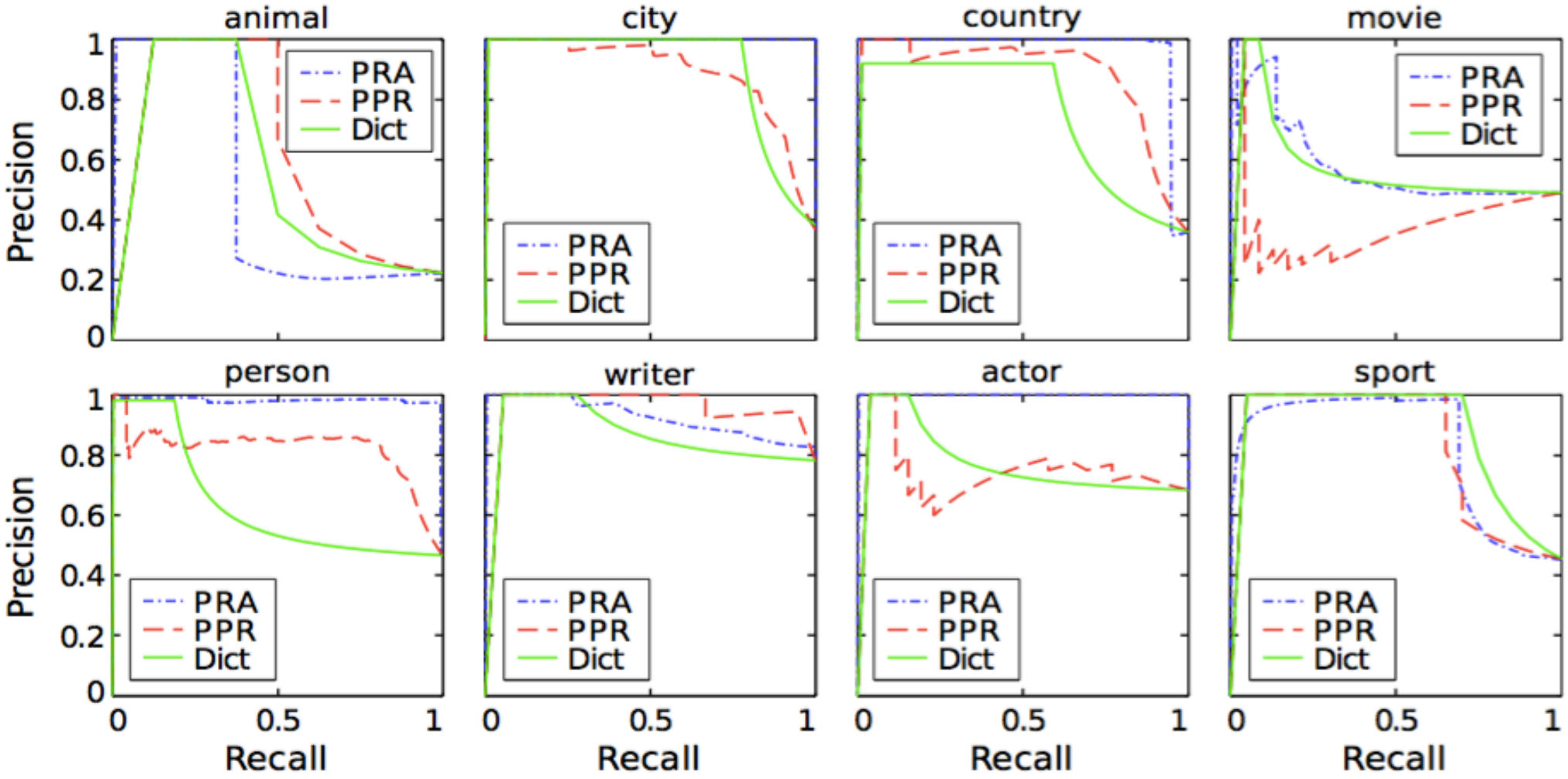


Figure 2: PR curves comparing both proposals with the dictionary as a classifier. Each figure displays the results of the three approaches with the examples of a specific category of the ontology using the pruned graph.

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

Originally Sparse Graphs

	English		Portuguese	
Category	Pruned graph	SVO+pruned graph	Pruned graph	SVO+pruned graph
animal	13.32 ± 47.12	392.02 ± 1859.43	1.57 ± 1.17	27.23 ± 58.58
country	22.65 ± 68.60	289.97 ± 970.55	5.04 ± 15.99	180.79 ± 820.71
city	5.00 ± 22.31	134.72 ± 1176.05	1.71 ± 4.33	32.03 ± 135.13
movie	1.60 ± 1.69	88.95 ± 891.37	1.33 ± 0.79	34.77 ± 177.64
person	2.99 ± 6.76	244.10 ± 1394.07	1.18 ± 0.66	16.62 ± 101.67
writer	2.72 ± 4.75	43.50 ± 419.41	1.15 ± 0.46	6.40 ± 12.04
actor	2.19 ± 2.12	77.21 ± 819.10	1.67 ± 1.41	6.96 ± 11.40
sport	19.94 ± 132.71	256.49 ± 1430.06	3.88 ± 6.31	26.98 ± 83.30
global	4.48 ± 28.95	275.49 ± 1715.53	1.81 ± 3.93	51.26 ± 199.03

Table 2: Mean out-degree value and associate standard deviation of the nodes in the pruned and SVO+pruned graphs of both languages. Each line shows the values for the considered categories (in the last line, all the categories).

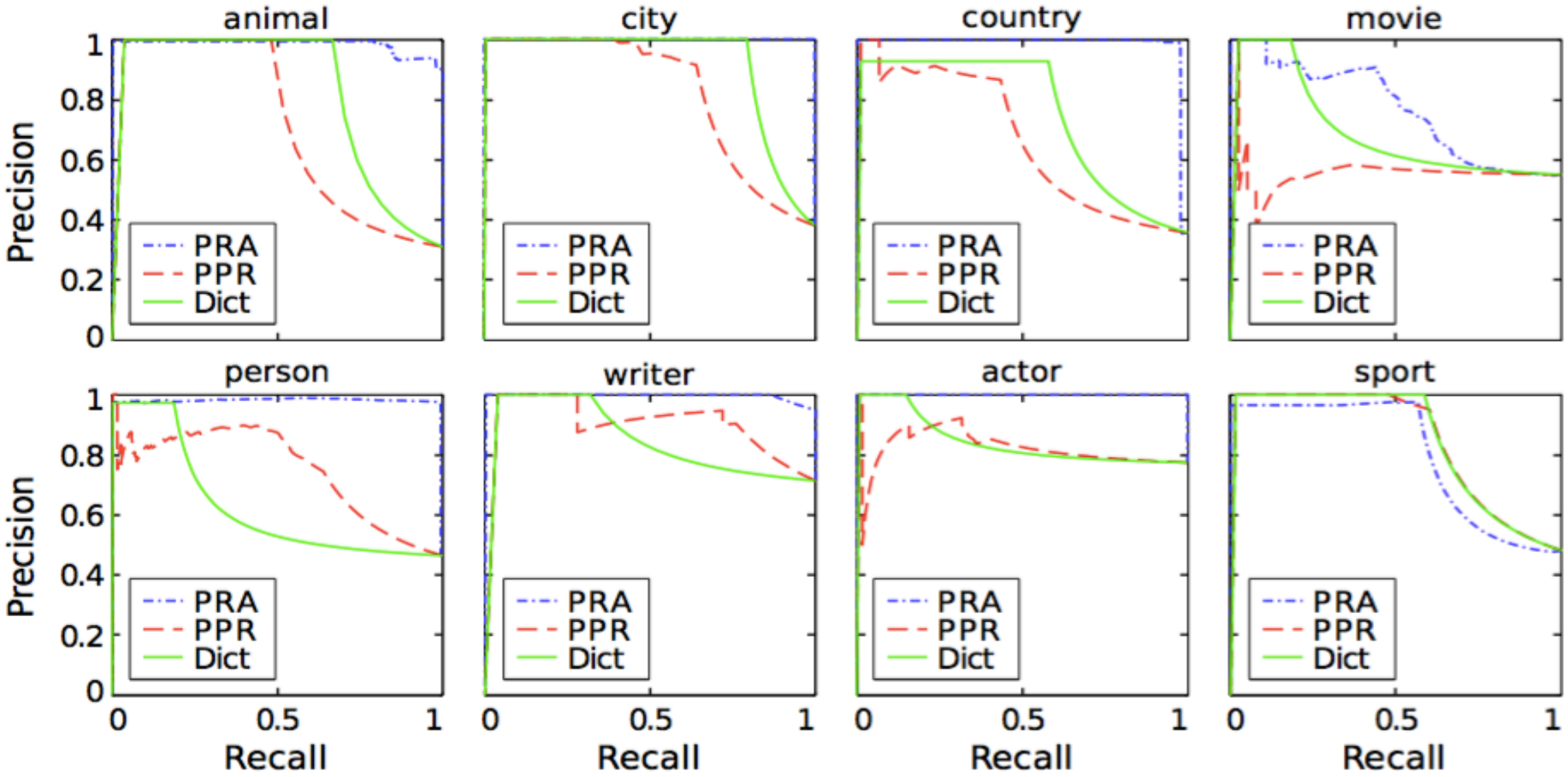


Figure 3: PR curves comparing both proposals with the dictionary as a classifier. Each figure displays the results of the three approaches with the examples of a specific category of the ontology using the graph pruned after populating it with new relationships inferred from a SVO corpus.

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

- denser SVO+pruned graph enhances the performance of the PRA-based proposal

- PPR is less regular and changes considerably among categories

Merging Concepts Learned in Different Languages

Hernandez-Gonzalez, Hruschka Jr. and Mitchell, 2016

- Dictionary is very competitive in categories we can find translations for their instances. In category sport, the performance was very good, but not in category Movie

- PRA often considers the “canBeTranslatedAs” path as strong feature, thus it can be seen as internally using the dictionary approach. But, can help when the Dictionary brings bad results (see country and person).

Key Idea 4: Cumulative, Staged Learning

Learning X improves ability to learn Y

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL
8. Mutilingual NELL (Portuguese)
9. CrossLingual NELL
10. Learn to microread single sentences and single documents
11. Self reflection, self-directed learning
12. Goal-driven reading: predict, then read to corroborate/correct
13. Make NELL learn by conversation (e.g, Twitter)
14. Add a robot body, or mobile phone body, to NELL

Key Idea 4: Cumulative, Staged Learning

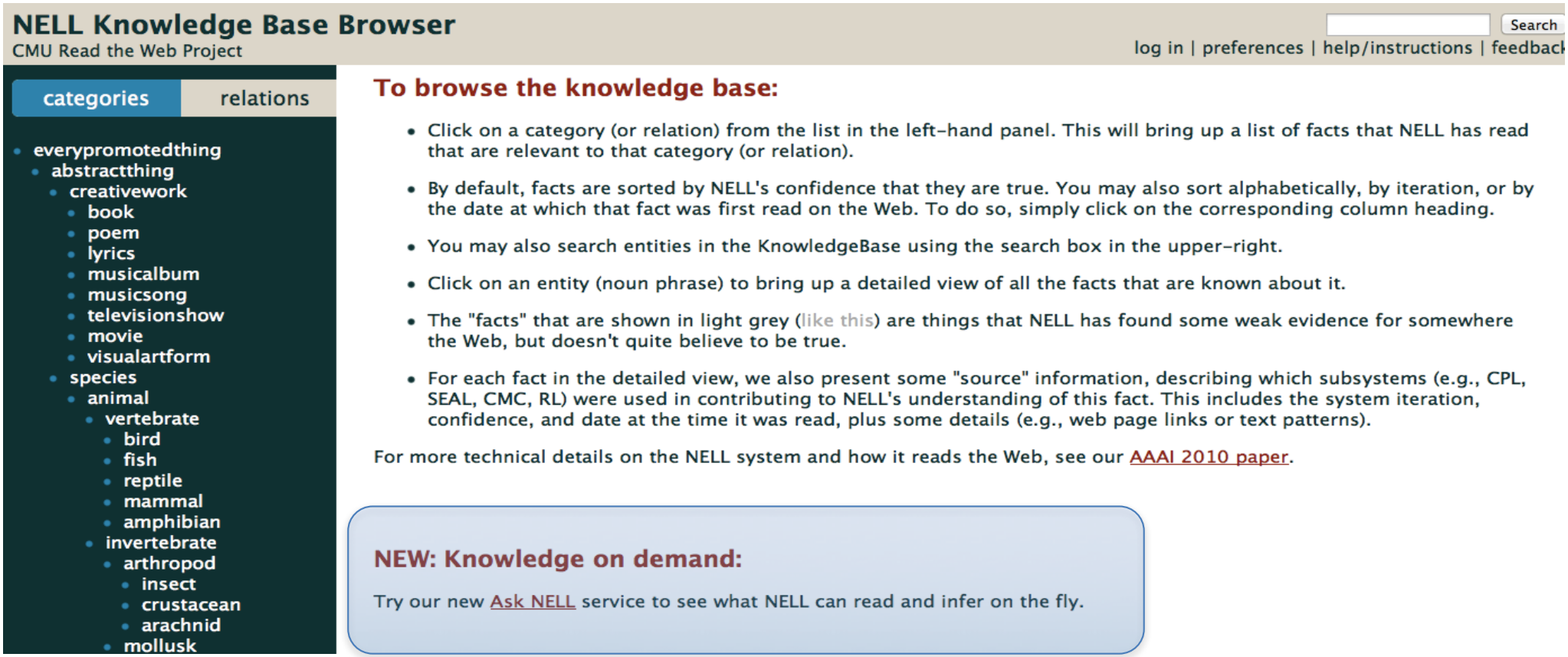
Learning X improves ability to learn Y

1. Classify noun phrases (NP's) by category
2. Classify NP pairs by relation
3. Discover rules to predict new relation instances
4. Learn which NP's (co)refer to which latent concepts
5. Discover new relations to extend ontology
6. Learn to infer relation instances via targeted random walks
7. Vision: connect NELL and NEIL
8. Mutilingual NELL (Portuguese)
9. CrossLingual NELL
10. Learn to microread single sentences and single documents
11. Self reflection, self-directed learning
12. Goal-driven reading: predict, then read to corroborate/correct
13. Make NELL learn by conversation (e.g, Twitter)
14. Add a robot body, or mobile phone body, to NELL

NELL is here

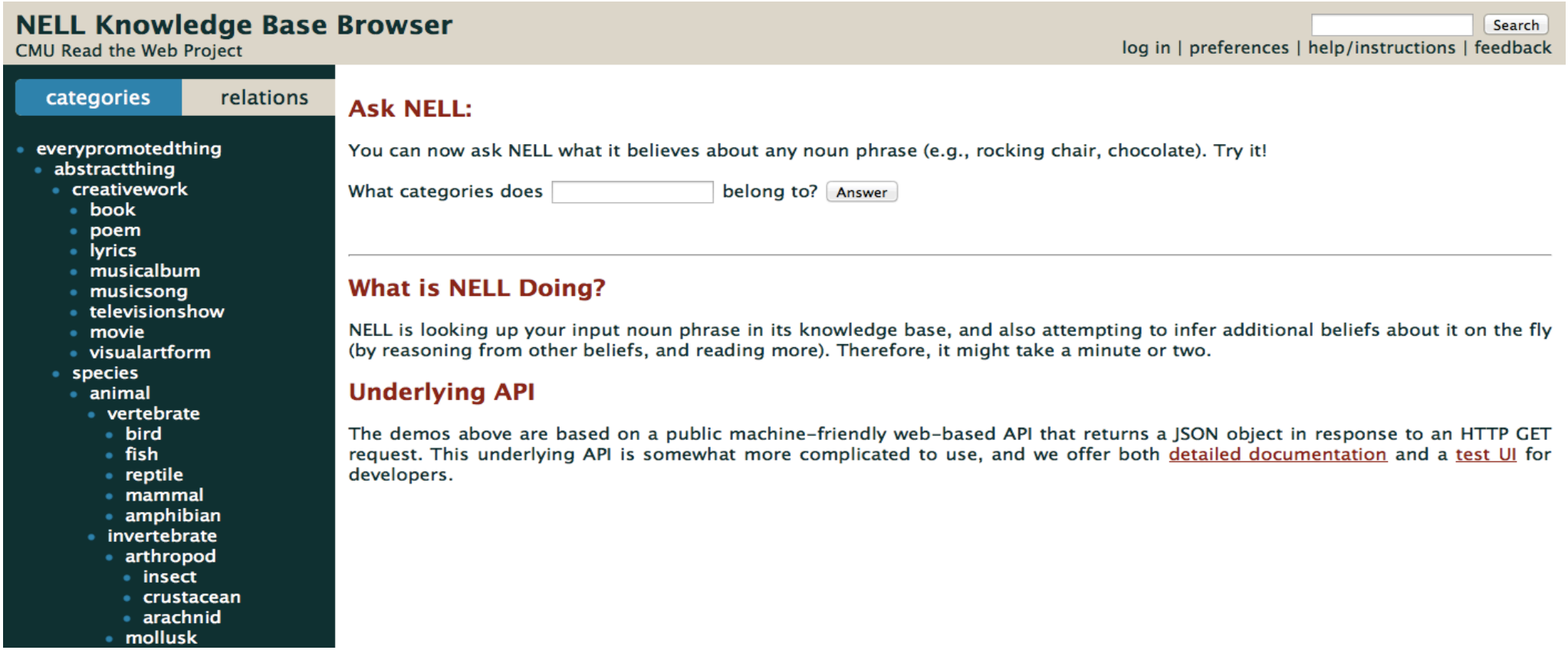
NELL: Never-Ending Language Learner

- **NELL is grown enough for new steps**
- **Knowledge on Demand**

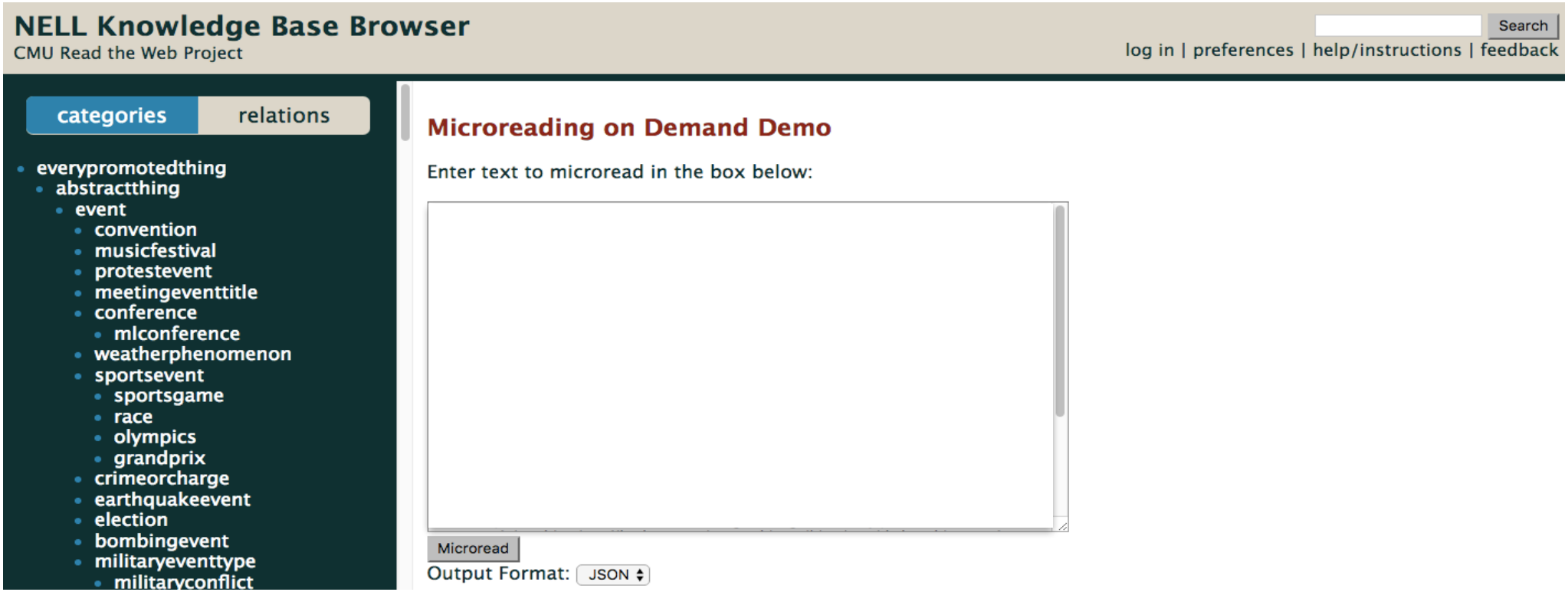


NELL: Never-Ending Language Learner

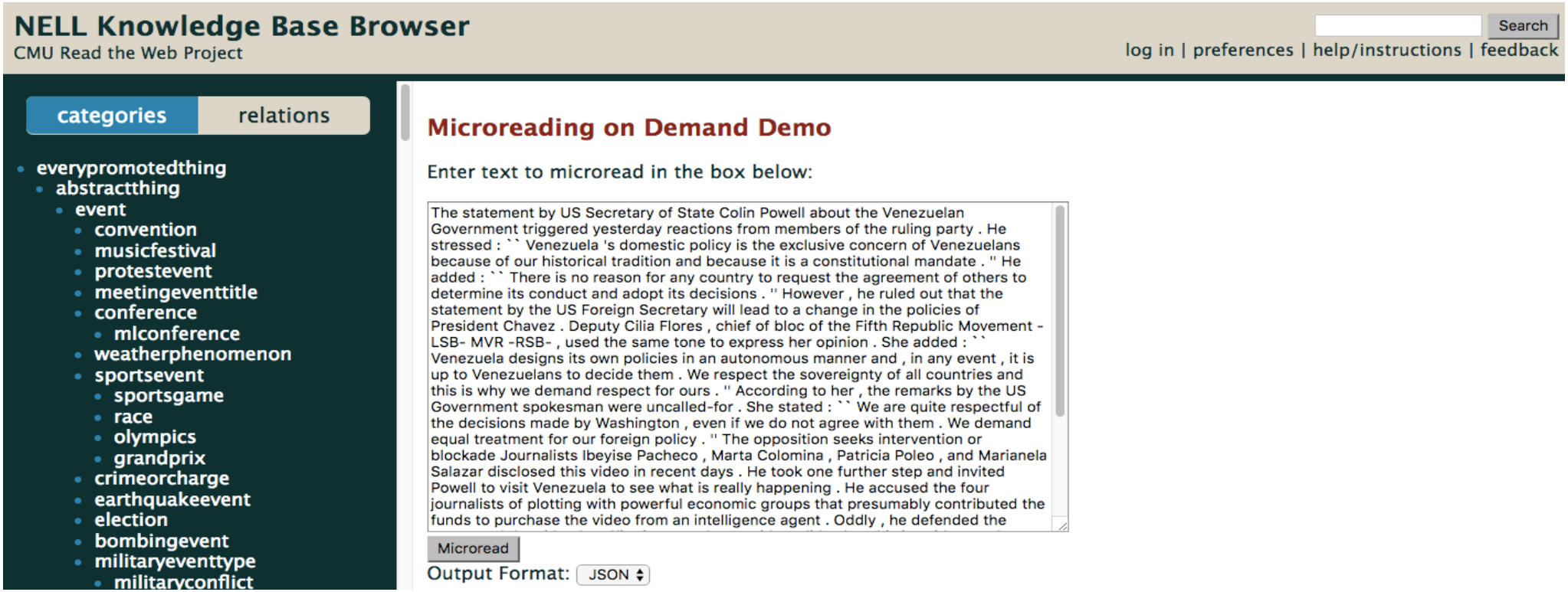
- **NELL is grown enough for new steps**
- **Knowledge on Demand – Ask NELL**



Micro-Reading



Micro-Reading



Micro-Reading

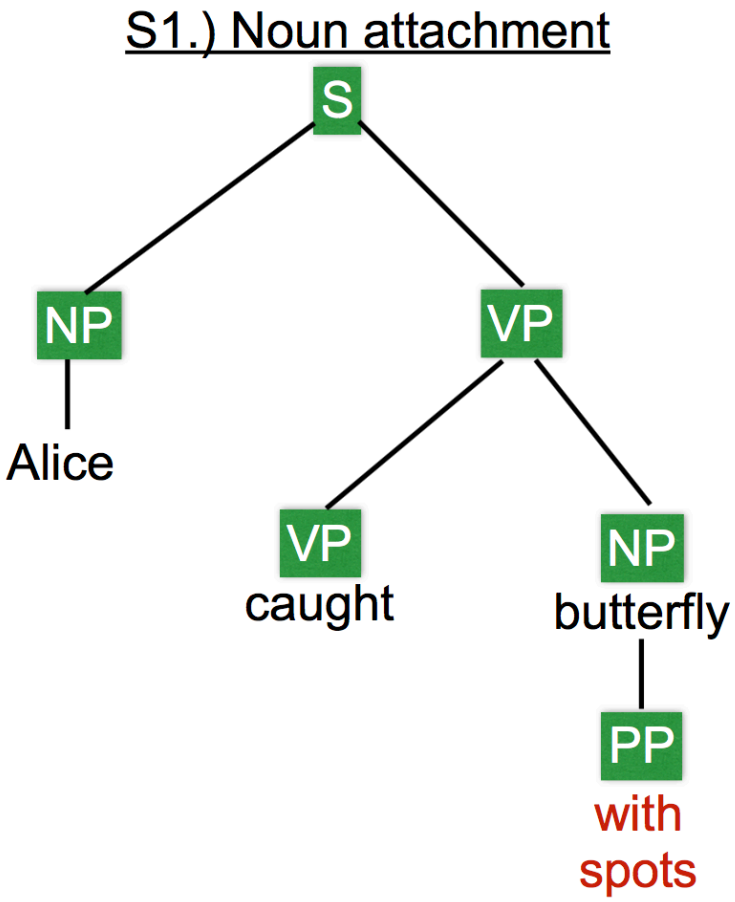
Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

Micro-Reading

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.



Micro-Reading

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.

Micro-Reading

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),
animalHasCoatPattern(leopard, spots),
animalHasCoatPattern(butterfly, spots), …

caught(*instrument*, *animal*):
caught(net, butterfly),
caught(fish, fishing rod), …

Categories:

city(*city*):
city(Saint Etienne), …

animal(*animal*):
animal(zebra),
animal(butterfly), …

Micro-Reading

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),

animalHasCoatPattern(leopard, *spots*),

animalHasCoatPattern(butterfly, *spots*), …

caught(*instrument*, *animal*):
caught(net, butterfly),
caught(fish, fishing rod), …

Categories:

city(*city*):
city(Saint Etienne), …

animal(*animal*):
animal(zebra),
animal(butterfly), …

Micro-Reading

Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),

animalHasCoatPattern(leopard, *spots*),

animalHasCoatPattern(butterfly, *spots*), …

caught(*instrument*, *animal*):

caught(*net*, butterfly),
caught(fish, fishing rod), …

Categories:

city(*city*):
city(Saint Etienne), …

animal(*animal*):
animal(zebra),
animal(butterfly), …

Micro-Reading

Nakashole and Mitchell, 2015

S2.) Alice caught the butterfly with the net.

Relations:

locatedIn(*city*, *country*):
locatedIn(Saint Etienne, France), …

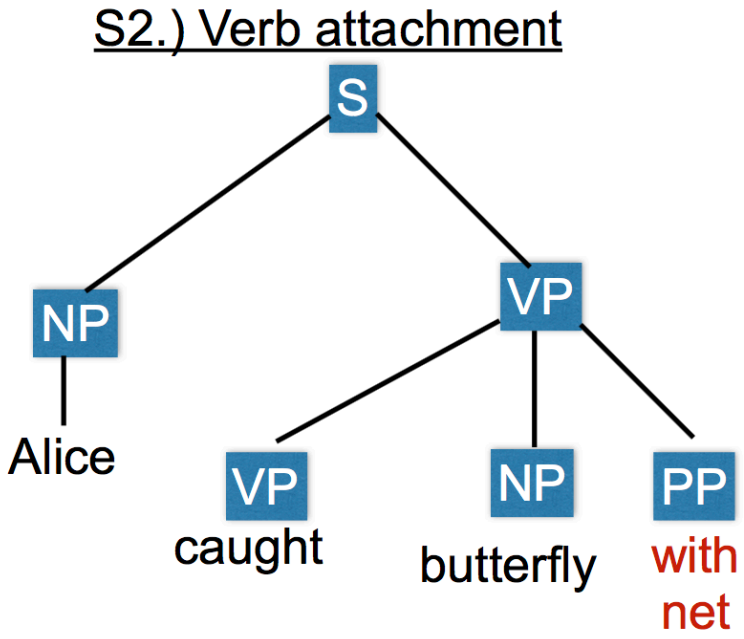
animalHasCoatPattern(*animal*, *coatPattern*):
animalHasCoatPattern(zebra, stripes),

animalHasCoatPattern(leopard, *spots*),

animalHasCoatPattern(butterfly, *spots*), …

caught(*instrument*, *animal*):

caught(*net*, butterfly),
caught(fish, fishing rod), …



Categories:

city(*city*):
city(Saint Etienne), …

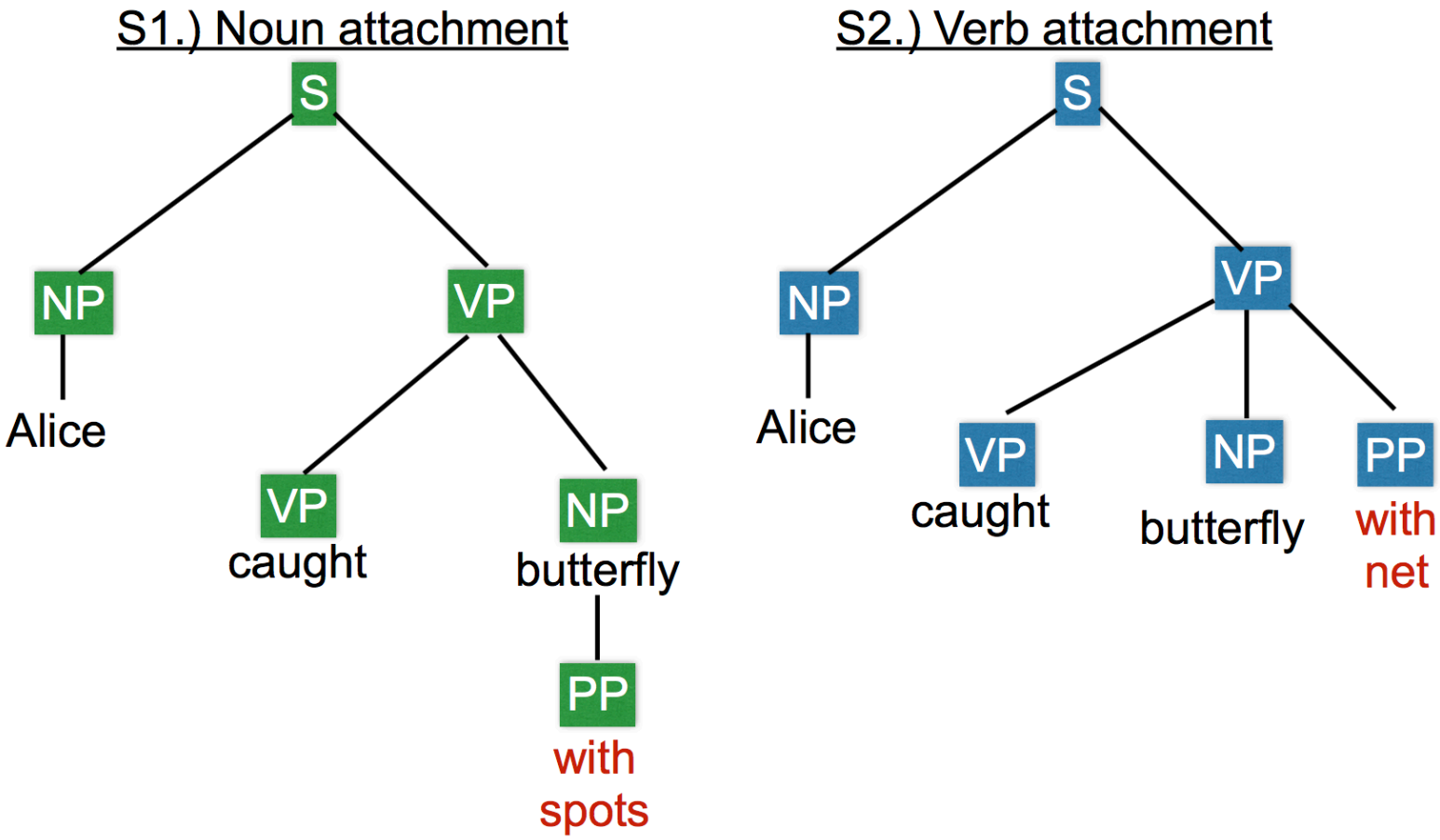
animal(*animal*):
animal(zebra),
animal(butterfly), …

Micro-Reading

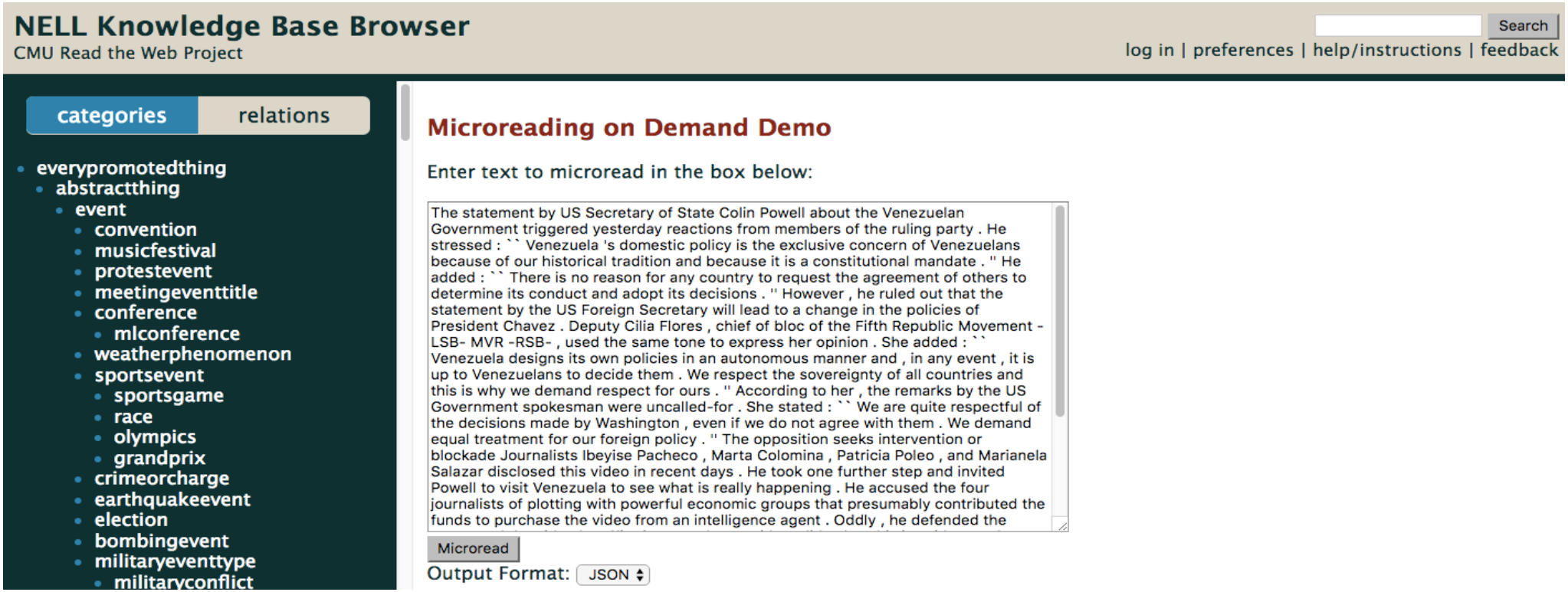
Nakashole and Mitchell, 2015

S1.) Alice caught the butterfly with the spots.

S2.) Alice caught the butterfly with the net.



Micro-Reading



Micro-Reading

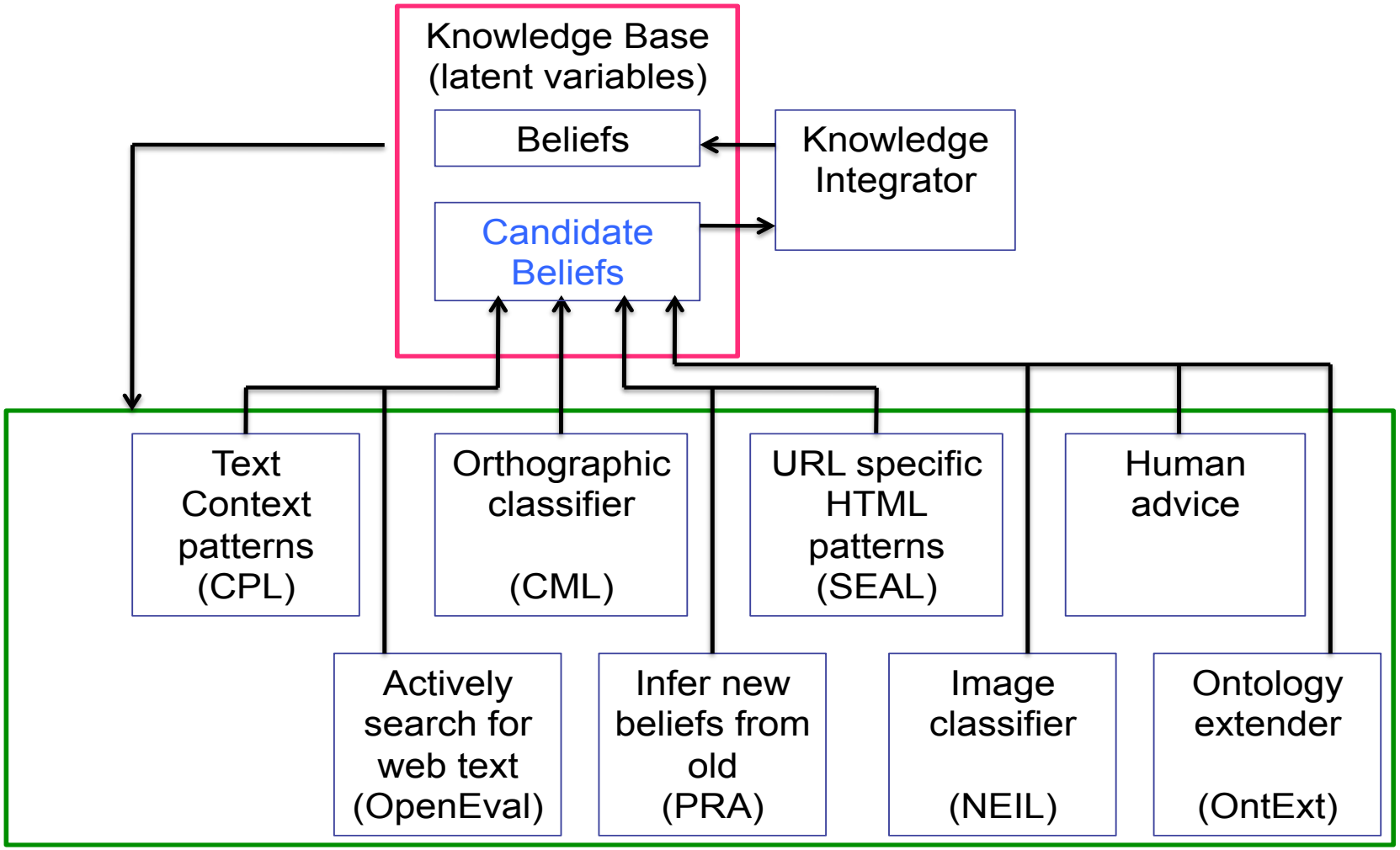
The statement by US Secretary of State Colin Powell about the Venezuelan Government triggered yesterday reactions from members of the ruling party . He stressed : `` Venezuela 's domestic policy is the exclusive concern of Venezuelans because of our historical tradition and because it is a constitutional mandate . '' He added : `` There is no reason for any country to request the agreement of others to determine its conduct and adopt its decisions . '' However , he ruled out that the statement by the US Foreign Secretary will lead to a change in the policies of President Chavez . Deputy Cilia Flores , chief of bloc of the Fifth Republic Movement -LSB- MVR -RSB- , used the same tone to express her opinion . She added : `` Venezuela designs its own policies in an autonomous manner and , in any event , it is up to Venezuelans to decide them . We respect the sovereignty of all countries and this is why we demand respect for ours . '' According to her , the remarks by the US Government spokesman were uncalled-for .

Montoya)))) (VP (VBD stated) (: :) (`` ``) (S (NP (DT The) (JJ national) (NN government)) (VP (VBZ has) (VP (VBN been) (ADJP (RB extremely) (JJ incompetent) (PP (IN in) (S (VP (VBG handling) (NP (NNS relations)) (PP (IN with) (NP (DT the) (NNP United) (NNPS States))))))))))) (. .)))

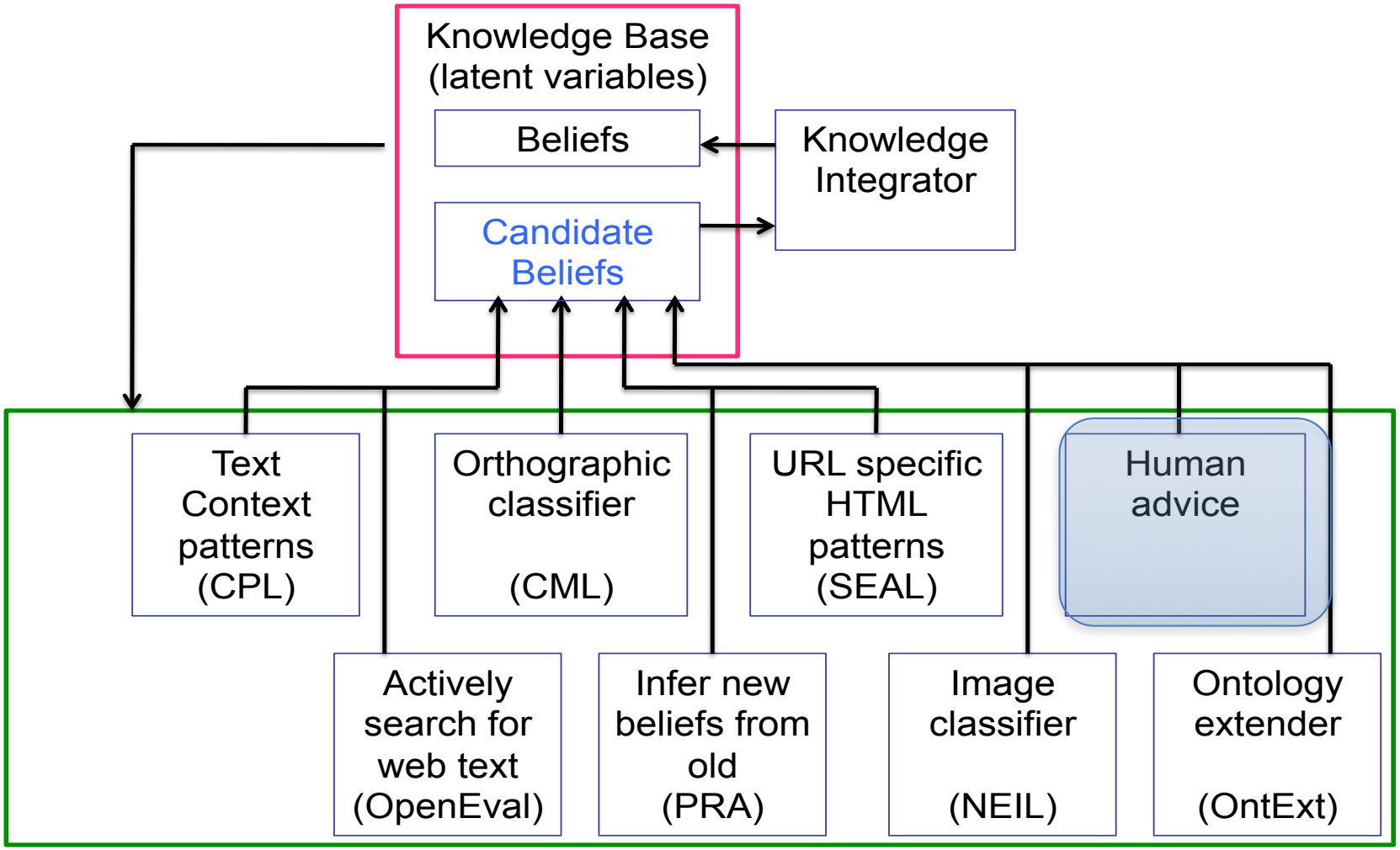
Type: con-parse Score: 0.0000 Value: (ROOT (S (NP (PRP It)) (VP (VBZ has) (VP (VBN been) (ADJP (JJ vague) (PP (IN about) (NP (NP (NP (NNP Venezuela) (POS 's)) (NNS intentions)) (PP (VBG regarding) (NP (NP (NP (DT the) (NN protection)) (PP (IN of) (NP (NN crude) (NNS prices)))) (CC and) (NP (NP (DT the) (NN war)) (PP (IN on) (NP (NN terrorism))))))))))) (. .)))

Type: dep-parse Value: det(statement-2, The-1) nsubj(triggered-14, statement-2) case(Secretary-5, by-3) compound(Secretary-5, US-4) nmod(statement-2, Secretary-5) case(Powell-9, of-6) compound(Powell-9, State-7) compound(Powell-9, Colin-8) nmod(Secretary-5, Powell-9) case(Government-13, about-10) det(Government-13, the-11) amod(Government-13, Venezuelan-12) nmod(Powell-9, Government-13) root(ROOT-0, triggered-14) compound(reactions-16, yesterday-15) dobj(triggered-14, reactions-16) case(members-18, from-17) nmod(triggered-14, members-18) case(party-22, of-19) det(party-22, the-20) amod(party-22, ruling-21) nmod(members-18, party-22) punct(triggered-14, .-23)

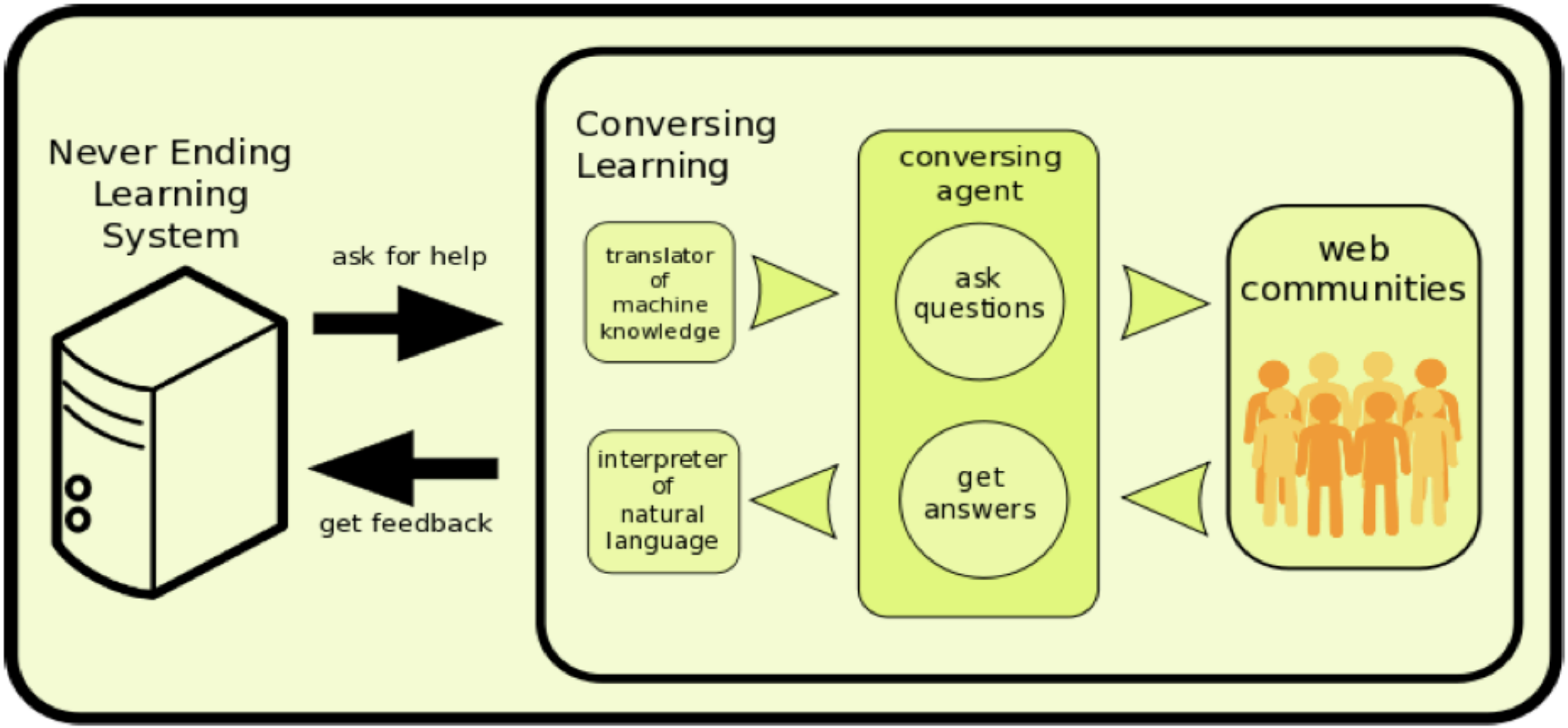
NELL Architecture



NELL Architecture



Conversing Learning



Conversing Learning

- Basic Steps:

- Decide which task is going to be asked
- Determine who are the oracles the ML system is going to consult
- Propose a method of conversation with oracles, often humans
- Determine how to feedback the ML system with the community inputs

Conversing Learning

- Basic Steps:

- **Decide which task is going to be asked**
- Determine who are the oracles the ML system is going to consult
- Propose a method of conversation with oracles, often humans
- Determine how to feedback the ML system with the community inputs

Conversing Learning

- **Decide which task is going to be asked**

- Learned facts
- Learned Inference Rules
- Metadata (mainly for automatically extending the ontology)

Conversing Learning

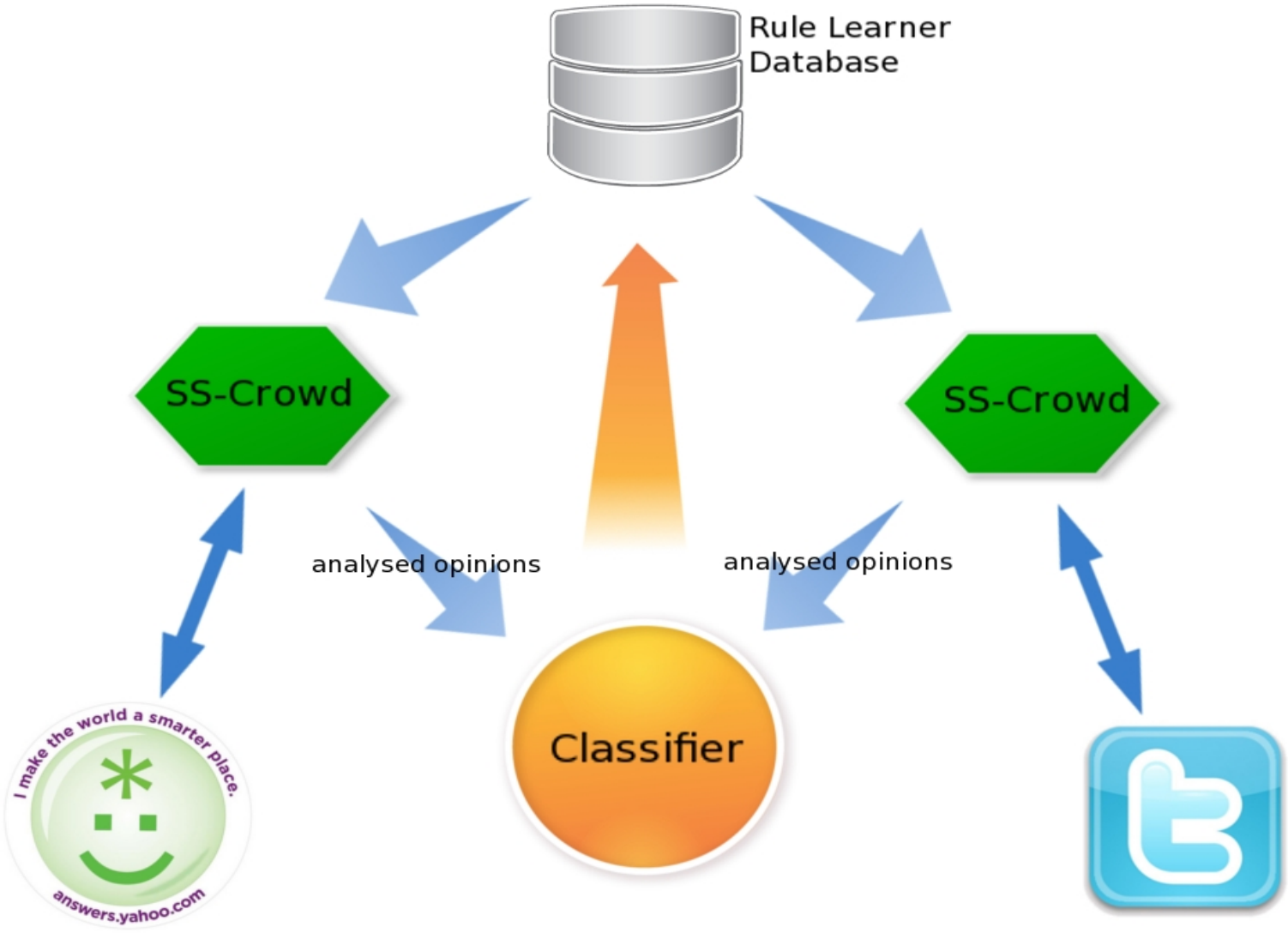
- Basic Steps:

- Decide which task is going to be asked
- **Determine who are the oracles the ML system is going to consult**
- Propose a method of conversation with oracles, often humans
- Determine how to feedback the ML system with the community inputs

Conversing Learning

- **who are the oracles the ML system is going to consult**
- Yahoo! Answers
  - very popular on the Web
  - a lot of metadata to harvest
- Twitter
  - millions of users worldwide
  - a system that was not designed to work as a QA environment
- Both web communities have API to connect to their database

Conversing Learning



Conversing Learning

- Basic Steps:

  - Decide which task is going to be asked
  - Determine who are the oracles the ML system is going to consult
  - **Propose a method of conversation with oracles, often humans**
  - Determine how to feedback the ML system with the community inputs

Conversing Learning

- **Propose a method of conversation with oracles, often humans**
- Macro Question-Answering
- For each posted question:
  - Ask for yes/no simple answers
  - Try to understand every answer
  - Discard answers too difficult to understand
  - Conclude based only on fully understood answers

Conversing Learning

- Basic Steps:

- Decide which task is going to be asked
- Determine who are the oracles the ML system is going to consult
- Propose a method of conversation with oracles, often humans
- **Determine how to feedback the ML system with the community inputs**

Conversing Learning

- **how to feedback the ML system with the community inputs?**
- Suggested actions to NELL:
  - Synonym/co-reference resolution
  - Automatically update the Knowledge Base
  - Feature Selection/Weighting

Conversing Learning

- Some Initial Results with First Order Rules:
- Take top 10% of rules from Rule Learner
- 60 rules were converted into questions and asked with both the regular and the Yes/No question approach
- The 120 questions received a total of 350 answers.

Conversing Learning

- Some Initial Results with First Order Rules:
- Rule extracted from NELL in PROLOG format
- stateLocatedInCountry(x,y):- statehascapital(x,z), citylocatedincoutry(z,y)

- converted into question:
- Is this statement always true? If state X has capital Z and city Z is located in country Y then state X is located in country Y.

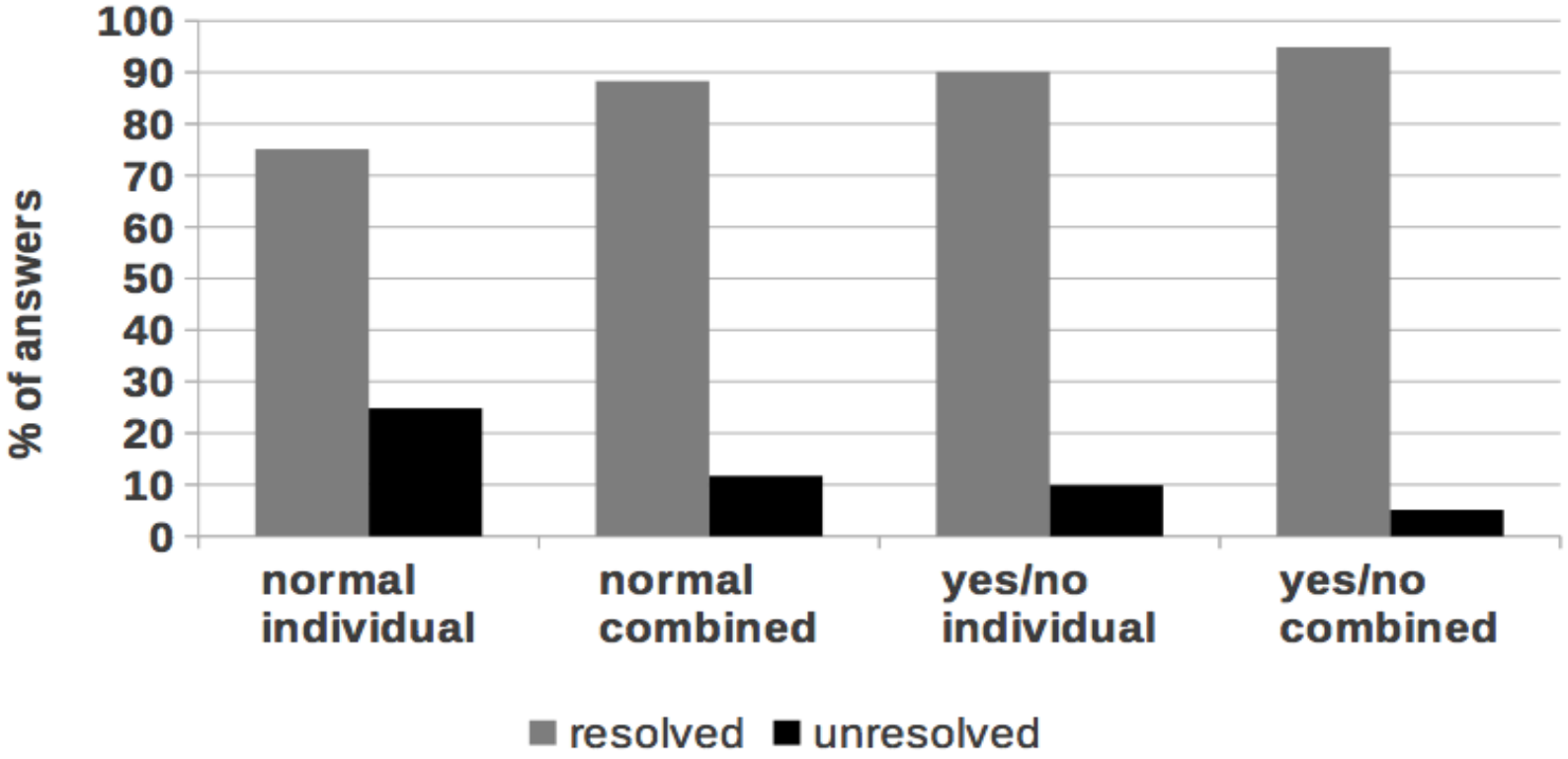
Conversing Learning - FOR

•Question: (Yes or No?) If athlete Z is member of team X and athlete Z plays in league Y, then team X plays in league Y.

- Twitter answers sample:
- **No. (Z in X) ∧ (Z in Y) → (X in Y)**

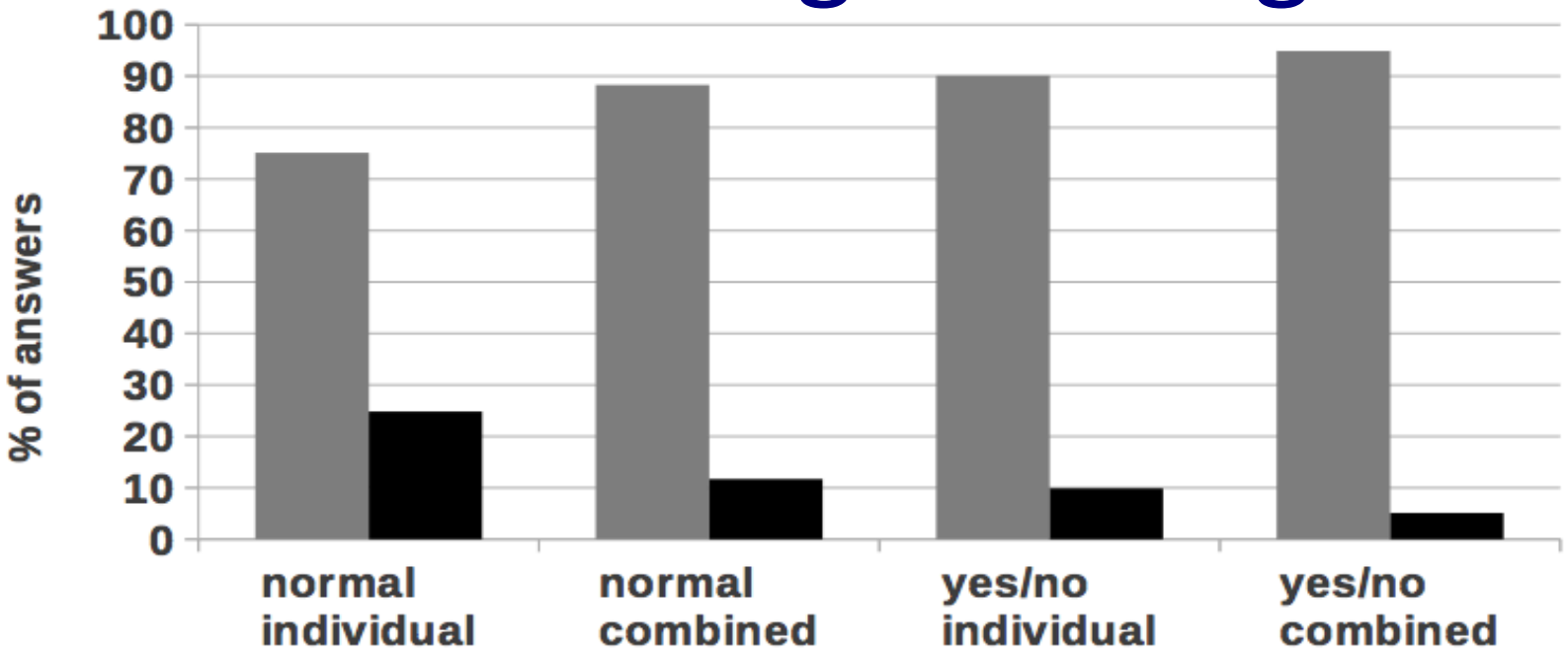
- Yahoo! Answers sample:
- NO, Not in EVERY case. Athlete Z could be a member of football team X and he could also play in his pub's Friday nights dart team. The Dart team could play in league Y (and Z therefore by definition plays in league Y). This does not mean that the football team plays in the darts league!

Conversing Learning



	Approved	Rejected	Unresolved
Twitter	51	17	4
Yahoo! Answers	124	168	58

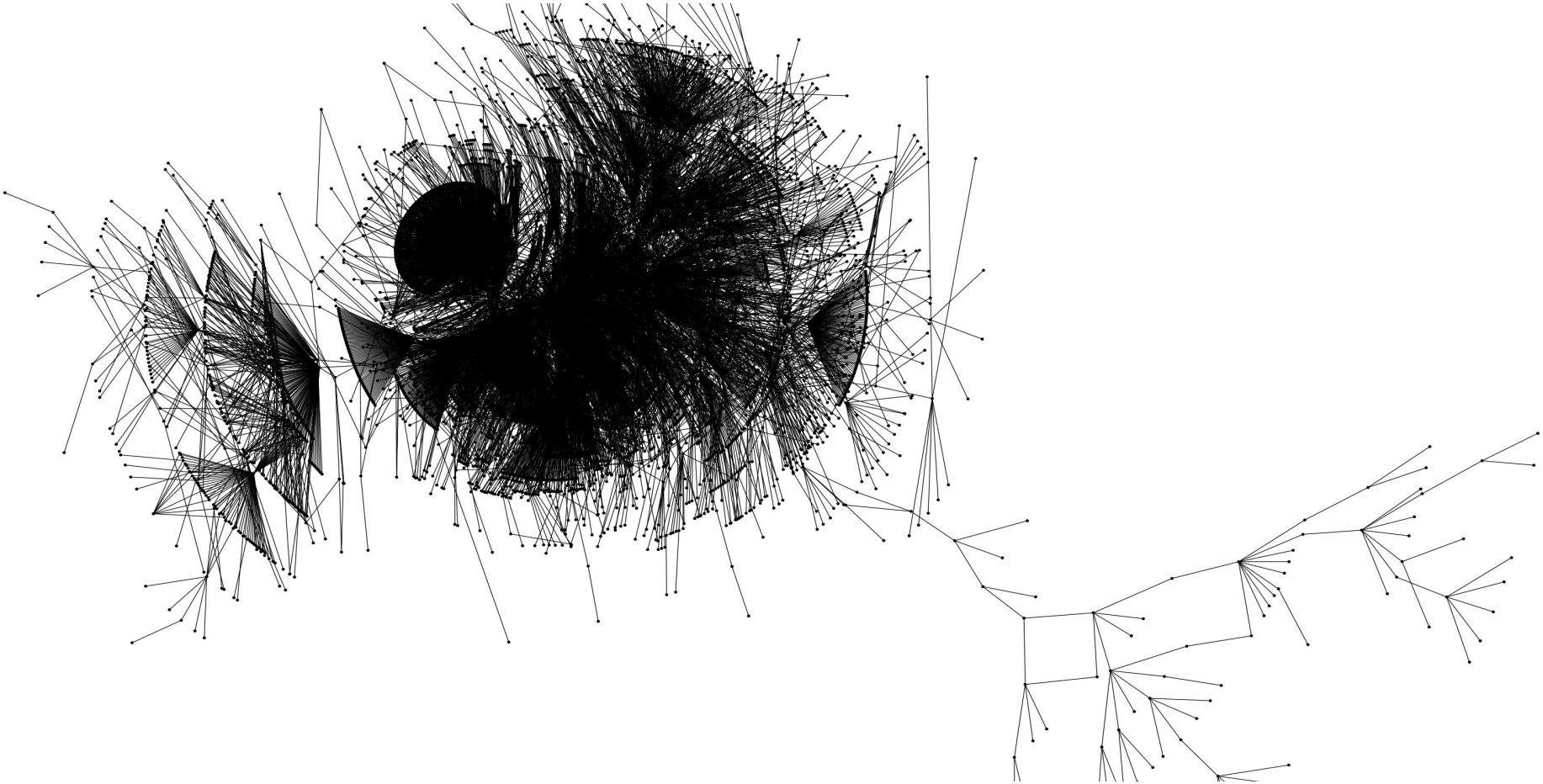
Conversing Learning



Question Type	TP	TN	FP	FN	Prec.	Recall	Acc.	F1
Regular Individual	41	31	7	39	0.85	0.51	0.61	0.64
Regular Combined	27	2	17	7	0.61	0.79	0.54	0.69
Yes/No Individual	62	53	14	45	0.81	0.57	0.66	0.67
Yes/No Combined	27	7	11	11	0.71	0.71	0.60	0.71
Best Answers	19	28	3	29	0.86	0.39	0.59	0.54

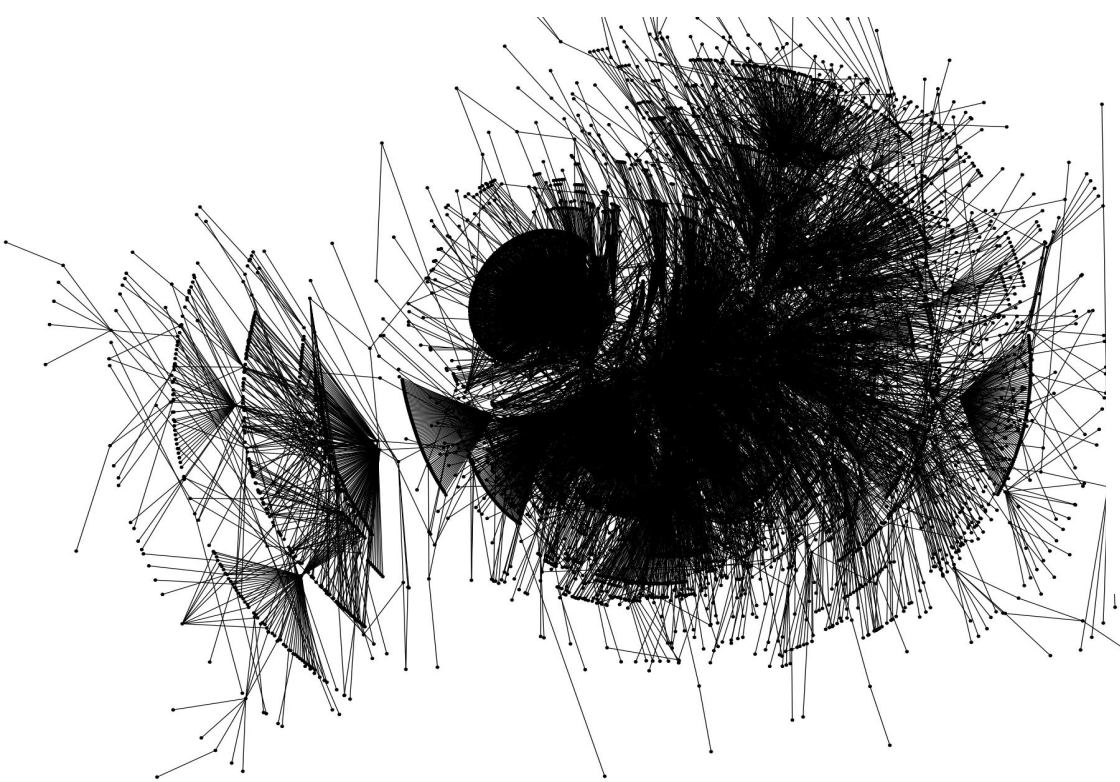
Conversing Learning

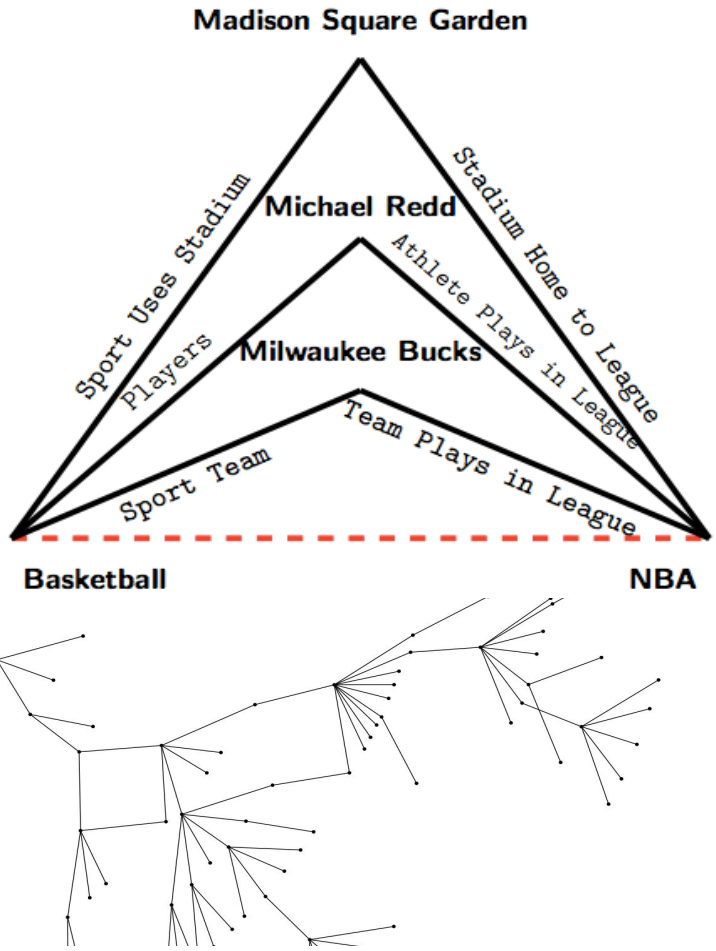
- Some Initial Results with Facts Validation:



Conversing Learning

- Some Initial Results with Facts Validation:





Conversing Learning

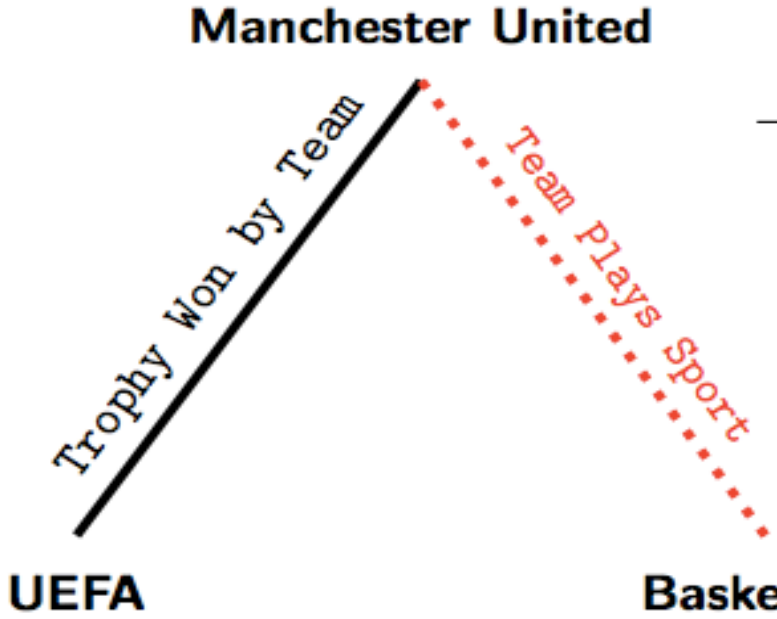
- Some Initial Results with Facts Validation:

Relations	outliers	answers
AthletePlaysInLeague & Players	9	72
TeamPlaysSport & TeamPlaysInLeague	20	144
TeamPlaysSport & TeamWonTrophy	53	386

Conversing Learning

- Some Initial Results with Facts Validation:

Relations	outliers	answers
AthletePlaysInLeague & Players	9	72
TeamPlaysSport & TeamPlaysInLeague	20	144
TeamPlaysSport & TeamWonTrophy	53	386



Outliers	
at least one wrong edge	39 (47.56%)
both edges correct	40 (48.19%)
unresolved edges	3 (03.65%)

Conversing Learning

- Some Initial Results with Metadata:
  - Question: Could you please give me some examples of clothing?
  - Answer 01: Snowshoes, rain ponchos, galoshes, sunhats, visors, scarves, mittens, and wellies are all examples of weather specific clothing!
  - Answer 02: pants
  - Answer 03: Training shoes can be worn by anyone for any purpose, but the term means to train in sports

Conversing Learning

- Some Initial Results with Metadata:

- Users replied with 552 seeds for 129 categories
  Total of 5900 promotions with seeds created by NELL's developers
- Total of 5300 promotions with seeds extracted from answers of Twitter users (similar precision)

Conversing Learning

- Some Initial Results with Metadata:
- For Relation Discovery Components
  - Symmetry: Is it always true that if a person P1 is neighbor of a person P2, then P2 is neighbor of P1?
  - Anti-symmetry: Is it always true that if a person P1 is the coach of a person P2, then P2 is not coach of P1?

Conversing Learning

- Some Initial Results with Metadata:
- Feature Weighting/Selection for CMC
  - Logistic Regression features are based on noun phrase morphology
  - (true or false) hotel names tend to be compound noun phrases having “hotel” as last the word.
  - (true or false) a word having “burgh” as sufix (ex. Pittsburgh) tend to be a city name.

Conversing Learning

- On going and future work

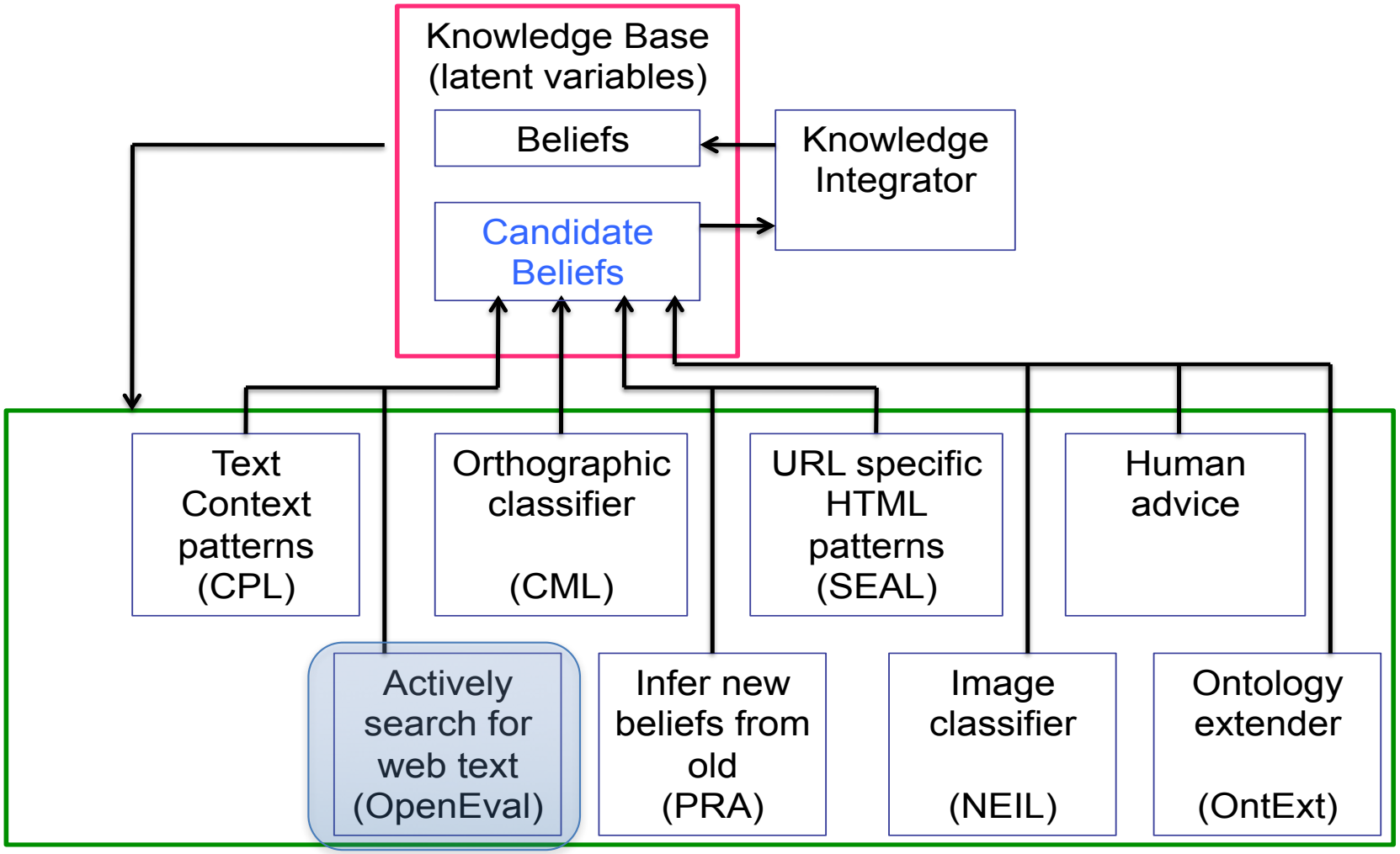
- Asking to the right community and to the right person
- Asking the right thing to maximize the results with minimum questions (**multi-view** Active Learning)
- Better Question-Answering methods
- Asking in different languages and explore time zones.

NELL: Never-Ending Language Learner

- Knowledge Base Validation in NELL

  - **Human Supervision:** RTW group members;
  - **Conversing Learning:** NELL can autonomously talk to people in web communities and ask for help
  - **Web Querying:** NELL can query the Web on specific facts to verify correctness, or to predict the validity of a new fact;
  - **Hiring Labelers:** NELL can autonomously hire people (using web services such as Mechanical Turk) to label data and help the system to validate acquired knowledge.

NELL Architecture



Motivation

Samadi, **Veloso** and **Blum, 2013**

I can wait more…

Shrimp is healthy

healthyFood **(shrimp)**

healthyFood **(shrimp)** 0.72

healthyFood **(apple)** 0.88

- Querying by human or agent
- Information validation

- Open Web
- Online/Anytime

- Scalable
- Few seed examples for training

- Small ontology

FAPESP

Carnegie Mellon University

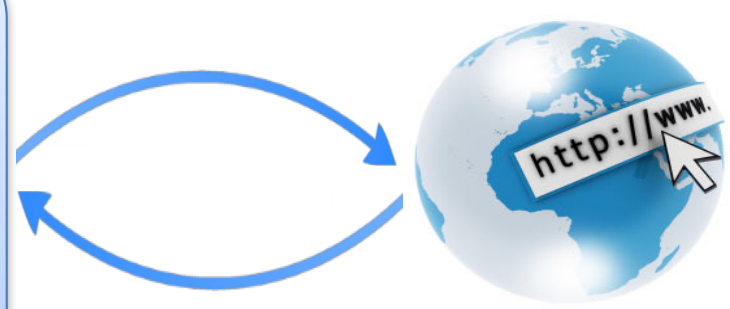
Predicate Instance Evaluator

Samadi, **Veloso** and **Blum, 2013**

*healthyFood(**shrimp**)?*

keywords:

Vitamin	0.88
Calories	0.83
Grow	0.69
Tree	0.66
Amount	0.59
Minerals	0.49
.	
.	
.	



Given the input time, which CBIs should be extracted?

NELL: Never-Ending Language Learner

Samadi, **Veloso** and **Blum, 2013**

- OpenEval in the last iteration:
  - –academicfield 0.8976357986206526 Environmental Anthropology.
  - –Several excellent textbooks and readers in environmental anthropology have now appeared, establishing a basic survey of the field.

NELL: Never-Ending Language Learner

Samadi, **Veloso** and **Blum, 2013**

- OpenEval in the last iteration:
  - –academicfield 0.912473775634353 Anesthesiology.
  - –The Department of Anesthesiology is committed to excellence in clinical service, education, research and faculty development.

NELL: Never-Ending Language Learner

Samadi, **Veloso** and **Blum, 2013**

- OpenEval in the last iteration:
  - –worksfor 0.9845774661303888 (charles osgood, cbs).
  - –Charles Osgood, often referred to as CBS News' poet-in-residence, has been anchor of "CBS News Sunday Morning" since 1994.

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at his estate on the tiny island of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, worries that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby. The old man draws Blomkvist in by promising solid evidence against Wennerström. Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of Vanger's niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist becomes acquainted with the members of the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

After discovering that Salander has hacked into his computer, he persuades her to assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep background investigations for Dragan Armansky, who, in turn, specializes that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same same same same same

This slide was adapted from [Hady et al., 2011]

Machine Reading

It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist agrees to spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible]inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.

Af[illegible]overing that Salander has hacked into his computer, he persuade[illegible] assist him with research. They eventually become lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible]gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
same
same
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

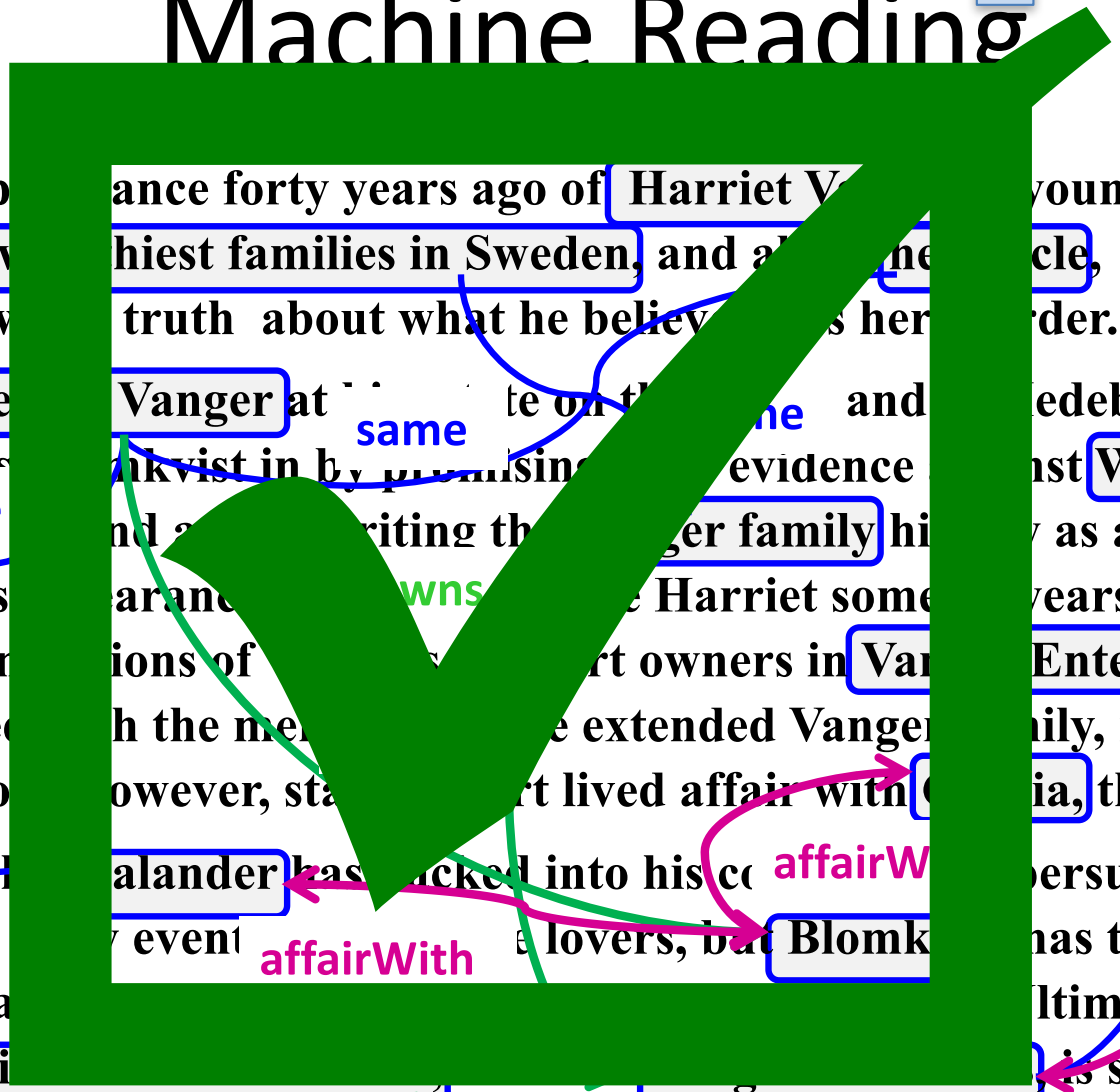
It's about the disappearance forty years ago of Harriet Vanger, a young scion of one of the wealthiest families in Sweden, and about her uncle, determined to know the truth about what he believes was her murder.

Blomkvist visits Henrik Vanger at [illegible] te on th[illegible] and of Hedeby.
The old man draws Blomkvist in by promising solid evidence against Wennerström.
Blomkvist ag[illegible] spend a year writing the Vanger family history as a cover for the real assignment: the disappearance of V[illegible] niece Harriet some 40 years earlier. Hedeby is home to several generations of Vangers, all part owners in Vanger Enterprises. Blomkvist beco[illegible] inted with the men[illegible] the extended Vanger family, most of whom resent his presence. He does, however, start a short lived affair with Cecilia, the niece of Henrik.
Af[illegible] overing that Salander has hacked into his c[illegible] persuade[illegible] assist him with research. They even[illegible] lovers, but Blomkvist has trouble getting close to Lisbeth who treats virtually everyone she meets with hostility. Ultimately the two discover that Harriet's brother Martin, CEO of Vanger Industries, is secretly a serial killer.
A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou[illegible] gations for Dragan Armansky, who, in tu[illegible]ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
same
owns
uncleOf
hires
affairWith
affairWith
enemyOf
same
same
headOf

This slide was adapted from [Hady et al., 2011]

Machine Reading



It's about the disap ance forty years ago of Harriet V young scion of one of the v hiest families in Sweden, and a ne cle, determined to know truth about what he believ s her der.

Blomkvist visits He Vanger at te on th ne and edeby. The old man draws nkvist in by promising evidence nst Wennerström. Blomkvist ag nd a riting th er family hi as a cover for the real assignment: the dis arane wns Harriet some years earlier. Hedeby is home to several ger ions of rt owners in Van Enterprises. Blomkvist beco uncleOf inte h the me e extended Vange ily, most of whom resent his presence. He do owever, sta rt lived affair with ia, the niece of Henrik Af overing th alander has icked into his co persuade assist him with research. even lovers, bu Blomk has trouble getting close to Lisbeth who trea ltimately the two discover that Harri is secretly a serial killer.

A 24-year-old computer hacker sporting an assortment of tattoos and body piercings supports herself by doing deep backgrou gations for Dragan Armansky, who, in tu ies that Lisbeth Salander is "the perfect victim for anyone who wished her ill."

same
same
affairWith
enemyOf
same
affairWith
headOf
same

This slide was adapted from [Hady et al., 2011]

Machine Reading

- Named Entity Resolution/Extraction
- Relation Extraction
- Co-reference and Synonym Resolution
- Relation Discovery
- Inference
- Knowledge Base Representation
- Document/Sentence Understanding (Micro-Reading)

More on NELL

- http://rtw.ml.cmu.edu/rtw/publications

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL in the News

- Aiming to Learn as We Do, a Machine Teaches Itself, *New York Times*, by Steve Lohr, October 5, 2010.
- Read the Internet, Speak English, *Universe*, by Clair Evans, republished at *io9*, October 8, 2010.
  Includes interview with Burr Settles and Tom Mitchell.
- Meet NELL. See NELL Run, Teach NELL How To Run, *TechCrunch*, by Evelyn Rusli, October 9, 2010.
  Includes video interview and NELL demonstration.
- Is NELL The New HAL?, *Word of Mouth* (NH Public Radio), by Virginia Prescott, October 27, 2010.
  Five-minute live interview segment with Burr Settles.
- Tech of the Future, Today: Breakthroughs in Artificial Intelligence, *PCWorld*, by Ian Paul, February 27, 2011.
  NELL is featured on slide 5.
- NELL atesta evolução da inteligência artificial no País, *Jornal do Comércio*, by Patricia Knebel, May 3, 2012.

NELL Overview: Video lectures

- Machine Learning and NELL, *Princeton Centenial Celebration of Alan Turing*, Tom Mitchell, May 2012. [this is more generally accessible than the following lecture]
- Never Ending Language Learning, *NAACL Workshop on Knowledge Extraction*, Tom Mitchell, June 2012. [contains more technical detail than above lecture]

Scientific Publications

- **Coupled Bayesian Sets Algorithm for Semi-supervised Learning and Information Extraction.**
  S. Verma and E. R. Hruschka Jr. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2012. [PDF] [abstract] [bib] [slides]
- **Acquiring Temporal Constraints between Relations.**
  P.P. Talukdar, D.T. Wijaya and T.M. Mitchell. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, 2012. [PDF] [abstract] [bib]
- **Weakly Supervised Training of Semantic Parsers.**

Industrial Approaches

- Google Knowledge Vault:
- “The Knowledge Vault is substantially bigger than any previously published structured knowledge repository, and features a probabilistic inference system that computes calibrated probabilities of fact correctness”.
- 271M facts with probability of being true at (or above) 0.9.
- It uses 16 different information extraction systems to extract (subject, predicate, object) knowledge triples from webpages.

Industrial Approaches

- IBM Watson:

- From Wikipedia: “Watson is a question answering computer system capable of answering questions posed in natural language, (...) The computer system was specifically developed to answer questions on the quiz show Jeopardy! (...) Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia, but was not connected to the Internet during the game. (...) Watson consistently outperformed its human opponents on the game's signaling device, but had trouble in a few categories, notably those having short clues containing only a few words.”

estevam.hruschka@gmail.com

Thank you very much!

and thanks to all people from NELL, KnowItAll, YAGO and DBPedia projects for very nice discussions and suggestions. And thanks to FAPESP (Fundação de Amparo a Pesquisa do Estado de Sao Paulo), and thanks to Federal University of São Carlos and thanks to Carnegie Mellon University.

References and Bibliography

- [Angeli et al., 2014] Angeli, Gabor, Julie Tibshirani, Jean Y. Wu, and Christopher D. Manning. "Combining distant and partial supervision for relation extraction." In*Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
- [Appel & Hruschka Jr., 2011] Appel, A.P., Hruschka Jr., E.R.: Prophet – a link-predictor to learn new rules on Nell. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 917–924. ICDMW '11, IEEE Computer Society, Washington, DC, USA (2011)
- Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.
- [Bach & Badaskar, 2007] Bach, Nguyen, and Sameer Badaskar. "A survey on relation extraction."*Language Technologies Institute, Carnegie Mellon University* (2007).
- [Banko, et al., 2007] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, Oren Etzioni: Open Information Extraction from the Web. IJCAI 2007: 2670-2676
- Banko, M., and Etzioni, O. 2007. Strategies for lifelong knowledge extraction from the web. In *Proc. of K-CAP*.
- Biega, J.; Kuzey, E.; and Suchanek, F. M. 2013. In- side yago2s: A transparent information extraction architecture. In *Proceedings of the 22nd international conference on World Wide Web companion*, 325–328. International World Wide Web Conferences Steering Committee.
- Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; and Hellmann, S. 2009. Dbpedia-a crystal-lization point for the web of data. *Web Semantics: science, services and agents on the world wide web* 7(3):154–165.
- [Bengtson & Roth, 2008] Bengtson, E., & Roth, D. (2008, October). Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 294-303). Association for Computational Linguistics.

References and Bibliography

- [Carlson et al., 2010b] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010).
- Carlson, A.; Betteridge, J.; Wang, R. C.; Hruschka Jr., E. R.; and Mitchell, T. M. 2010b. Coupled semi-supervised learn- ing for information extraction. In *Proc. of WSDM*.
- [Clark et al., 2007] Peter Clark, Phil Harrison, John Thompson, Rick Wojcik, Tom Jenkins, David Israel, Reading to Learn: An Investigation into Language Understanding. The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.
- [DARPA, 2012] DARPA Machine Reading Program, http://www.darpa.mil/Our_Work/I2O/Programs/Machine_Reading.aspx.
- Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge vault: A web-scale approach to proba- bilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 601–610. ACM.
- Dong, X. L.; Gabrilovich, E.; Murphy, K.; Dang, V.; Horn, W.; Lugaresi, C.; Sun, S.; and Zhang, W. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment* 8(9):938–949.
- [Dutta & Weikum, 2015]Dutta, S., & Weikum, G. (2015). Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association for Computational Linguistics*, *3*, 15-28.
- [Etzioni et al., 2007] Oren Etzioni, Michele Banko, and Michael J. Cafarella, Machine Reading.The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.
- [Etzioni et al., 2011] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam "Open Information Extraction: the Second Generation“. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011).*

References and Bibliography

- [Etzioni, 2008] Oren Etzioni. 2008. Machine reading at web scale. In *Proceedings of the international conference on Web search and web data mining* (WSDM '08). ACM, New York, NY, USA, 2-2.
- [Fader et al., 2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. "Identifying Relations for Open Information Extraction". *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*
- [Fern, 2008] Xiaoli Z. Fern, CS 434: Machine Learning and Data Mining, School of Electrical Engineering and Computer Science, Oregon State University, Fall 2008.
- Ferrucci, D. A.; Levas, A.; Bagchi, S.; Gondek, D.; and Mueller, E. T. 2013. Watson: beyond jeopardy! *Artif. Intell.* 199:93–105.
- Ferrucci, D. 2010. Build watson: an overview of deepqa for the jeopardy! challenge. In *Proceedings of the 19th interna- tional conference on Parallel architectures and compilation techniques*, 1–2. ACM.
- Ferrucci, D. A. 2012. Introduction to this is watson. *IBM Journal of Research and Development* 56(3.4):1–1.
- [Galárraga et al., 2014] Luis Galárraga, Geremy Heitz, Kevin Murphy, and Fabian M. Suchanek. 2014. Canonicalizing Open Knowledge Bases. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (CIKM '14). ACM, New York, NY, USA, 1679-1688.
- Gardner, M.; Talukdar, P. P.; Kisiel, B.; and Mitchell, T. 2013. Improving learning and inference in a large knowledge-base using latent syntactic cues. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.
- Gardner, M.; Talukdar, P.; Krishnamurthy, J.; and Mitchell, T. 2014. Incorporating vector space similarity in random walk inference over knowledge bases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Lan- guage Processing (EMNLP)*. Doha, Qatar.: Association for Computational Linguistics.

References and Bibliography

- Gerber, D.; Esteves, D.; Lehmann, J.; Bu¨hmann, L.; Usbeck, R.; Ngonga Ngomo, A.-C.; and Speck, R. 2015. Defacto - temporal and multilingual deep fact validation. *Web Seman- tics: Science, Services and Agents on the World Wide Web*.
- [Grishman & Sundheim, 1996] Grishman, Ralph; Sundheim, B. 1996. Message Understanding Conference - 6: A Brief History. In *Proc. International Conference on Computational Linguistics*.
- [Hady et al., 2011] Hady W. Lauw, Ralf Schenkel, Fabian Suchanek, Martin Theobald, and Gerhard Weikum, "Semantic Knowledge Bases from Web Sources" at IJCAI 2011, Barcelona, July 2011
- Hoffart, J.; Suchanek, F. M.; Berberich, K.; Lewis-Kelham, E.; de Melo, G.; and Weikum, G. 2011. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, 229– 232. New York, NY, USA: ACM.
- [Hoffart et al., 2012] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Special issue of the Artificial Intelligence Journal, 2012
- [Hoffart et al., 2015] Johannes Hoffart, Nicoleta Preda, Fabian M. Suchanek, and Gerhard Weikum. Knowledge Bases for Web Content Analytics. WWW 2015 Tutorial.
- [Hruschka et al., 2012] E. R. Hruschka Jr. and M. C. Duarte and M. C. Nicoletti. Coupling as Strategy for Reducing Concept-Drift in Never-ending Learning Environments. Fundamenta Informaticae, IOS Press, 2012.
- [Krishnamurthy & Mitchell, 2011] Krishnamurthy, J., Mitchell, T.M.: Which noun phrases denote which concepts. In: Proceedings of the Forty Ninth Annual Meeting of the Association for Computational Linguistics (2011)

References and Bibliography

- [Lao et al., 2011] Lao, N., Mitchell, T., Cohen, W.W.: Random walk inference and learning in a large scale knowledge base. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 529–539. Associa- tion for Computational Linguistics, Edinburgh, Scotland, UK. (July 2011), http://www.aclweb.org/anthology/D11-1049
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... & Bizer, C. (2014). DBpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, *5*, 1-29.
- Lehmann, J.; Schu¨ppel, J.; and Auer, S. 2007. Discovering unknown connections-the dbpedia relationship finder. *CSSW* 113:99–110.
- Lin, T.; Mausam; and Etzioni, O. 2012a. Entity link- ing at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, 84–88. Montre´al, Canada: Association for Computational Linguistics.
- Lin, T.; Mausam; and Etzioni, O. 2012b. No noun phrase left behind: Detecting and typing unlinkable entities. In *Pro- ceedings of the 2012 Joint Conference on Empirical Meth- ods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, 893– 903. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Lin, T. 2013. *Leveraging Knowledge Bases in Web Text Processing*. Ph.D. Dissertation, University of Washington.
- Lukovnikov, D.; Kontokostas, D.; Stadler, C.; Hellmann, S.; and Lehmann, J. 2014. Dbpedia viewer-an integrative inter- face for dbpedia leveraging the dbpedia service eco system. In *Proceedings of the 7th Workshop on Linked Data on the Web*.

References and Bibliography

- [Lodhi et al., 2002] Lodhi, H., Saunders, C., Shawe-Taylor, J., & Cristianini, N. (2002). Text classification using string kernels. Journal of Machine Learning Research (pp. 419–444).
- [Mahdisoltani et al., 2015] Mahdisoltani, F., Biega, J., & Suchanek, F. M. YAGO3: A Knowledge Base from Multilingual Wikipedias. 7th Biennial Conference on Innovative Data Systems Research, CIDR 2015
- [Mitchell, 1997] Tom M. Mitchell, Machine Learning. McGraw-Hill, 1997.
- [Mitchell, 2006] Tom M. Mitchell, The Discipline of Machine Learning, my perspective on this research field, July 2006 (http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf).
- Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Bet- teridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2015. Never-ending learn- ing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.
- [Mohamed et al., 2011] Mohamed, T.P., Hruschka, Jr., E.R., Mitchell, T.M.: Discovering relations between noun categories. In: Proceedings of the Conference on Empirical Methods in Nat- ural Language Processing. pp. 1447–1455. EMNLP '11, Association for Computa- tional Linguistics, Stroudsburg, PA, USA (2011)
- Morsey, M.; Lehmann, J.; Auer, S.; and Ngomo, A.-C. N. 2011. Dbpedia sparql benchmark–performance assessment with real queries on real data. In *The Semantic Web–ISWC 2011*. Springer. 454–469.
- Morsey, M.; Lehmann, J.; Auer, S.; Stadler, C.; and Hell- mann, S. 2012. Dbpedia and the live extraction of structured data from wikipedia. *Program* 46(2):157–181.
- [Nadeau & Sekine, 2007] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30.1 (2007): 3-26.

References and Bibliography

- [Navarro et al., 2013] Navarro, L. F. and Appel, A. P. and Hruschka Jr., E. R., GraphDB – Storing Large Graphs on Secondary Memory. In New Trends in Databases and Information. Advances in Intelligent Systems and Computing, Springer, 177-186, 2013.
- Nakashole, N.; Weikum, G.; and Suchanek, F. M. 2012. Discovering and Exploring Relations on the Web. In *Demo at VLDB*.
- Nickel, M.; Murphy, K.; Tresp, V.; and Gabrilovich, E. 2015. A review of relational machine learning for knowl- edge graphs: From multi-relational link prediction to au-
- tomated knowledge graph construction. *arXiv:1503.00759*.
- [Norvig, 2007] Peter Norvig, Inference in Text Understanding. The 2007 AAAI Spring Symposium. Published by The AAAI Press, Menlo Park, California, 2007.
- [Pedro & Hruschka Jr., 2012a] Saulo D.S. Pedro and Estevam R. Hruschka Jr., Conversing Learning: active learning and active social interaction for human supervision in never-ending learning systems. Xiii Ibero-american Conference On Artificial Intelligence, IBERAMIA 2012, 2012.
- [Pedro & Hruschka Jr., 2012b] Pedro, S.D.S., Hruschka Jr., E.R.: Collective intelligence as a source for machine learning self-supervision. In: Proc. of the 4th International Workshop on Web Intelligence and Communities. WIC12, NY, USA, ACM (2012) 5:1–5:9
- [Pedro et al., 2013] Saulo D.S. Pedro, Ana Paula Appel, and Estevam R. Hruschka, Jr. Autonomously reviewing and validating the knowledge base of a never-ending learning system. In *Proceedings of the 22nd international conference on World Wide Web companion* (WWW '13 Companion), 1195-120, 2013.

References and Bibliography

- [Pennington et al., 2014] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12 (2014): 1532-1543.
- [Raghunathan et al., 2010] Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010, October). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 492-501). Association for Computational Linguistics.
- [Ratinov & Ruth, 2009] Ratinov, Lev, and Dan Roth. "Design challenges and misconceptions in named entity recognition." *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2009.
- [Recasens et al., 2013] Recasens, M., Can, M., & Jurafsky, D. (2013). Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. In *Proceedings of NAACL-HLT* (pp. 897-906).
- Samadi, M.; Veloso, M. M.; and, M. 2013. Openeval: Web information query evaluation. In *AAAI*. Citeseer.
- [Settles, 2011] Settles, B.: Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In: Proc. of the EMNLP'11, Edinburgh, ACL (2011) 1467–1478 5.

References and Bibliography

- [Singh et al., 2011] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 793-803.
- Suchanek, F., and Weikum, G. 2013. Knowledge harvesting in the big-data era. In *Proceedings of the 2013 international conference on Management of data*, 933–938. ACM.
- Suchanek, F. M.; Hoffart, J.; Kuzey, E.; and Lewis-Kelham, E. 2013. Yago2s: Modular high-quality information extrac- tion with an application to flight planning. In *BTW*, 515– 518.
- Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*. New York, NY, USA: ACM Press.
- Talukdar, P. P.; Wijaya, D.; and Mitchell, T. 2012. Cou- pled temporal scoping of relational facts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM)*.

References and Bibliography

- [Theobald & Weikum, 2012] Martin Theobald and Gerhard Weikum. From Information to Knowledge: Harvesting Entities and Relationships from Web Sources. Tutorial at PODS 2012
- Tseng, Y.-H.; Lee, L.-H.; Lin, S.-Y.; Liao, B.-S.; Liu, M.-J.; Chen, H.-H.; Etzioni, O.; and Fader, A. 2014. Chinese open relation extraction for knowledge acquisition. In *EACL*.
- [Verma & Hruschka Jr., 2012] S. Verma and E. R. Hruschka Jr. Coupled Bayesian Sets Algorithm for Semi-supervised Learning and Information Extraction. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2012.
- [Wang & Cohen, 2007] Richard C. Wang and William W. Cohen: Language-Independent Set Expansion of Named Entities using the Web. In *Proceedings of IEEE International Conference on Data Mining* (ICDM 2007), Omaha, NE, USA. 2007.
- [Weikum et al., 2009] G. Weikum, G., Kasneci, M. Ramanath, F. Suchanek. DB & IR methods for knowledge discovery. Communications of the ACM 52(4), 2009.
- Wijaya, D.; Talukdar, P. P.; and Mitchell, T. 2013. Pidgin: Ontology alignment using web text as interlingua. In *Pro- ceedings of the Conference on Information and Knowledge Management (CIKM 2013)*. San Francisco, USA: Associa- tion for Computing Machinery.