CS544: Semantic Class Learning

February 1, 2011

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

Semantic Class Learning: Objectives

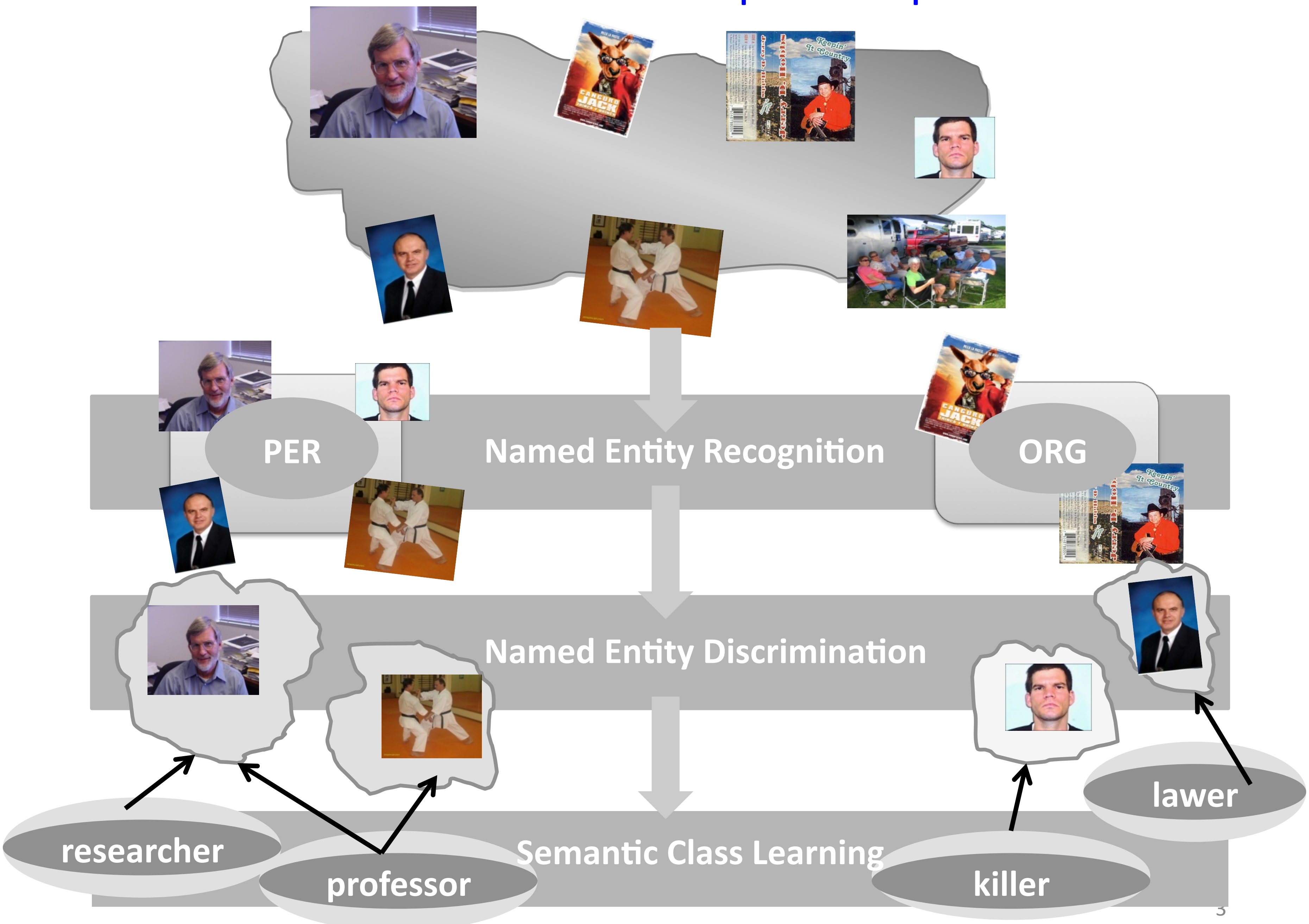
- Given a class and an instance, learn automatically with minimum supervision new <u>instances</u>, classes and the ISA relations among them.

- Examples:
  - *class_name*: Nobel prize winners
  - *instances*: Albert Einstein, Max Plank …

  - *class_name*: former Russian federation states
  - *instances*: Georgia, Urkaine, Lithuania …

Why do we need Semantic Class Learners?

Helps solve the puzzle from the first lecture



The problem with automated Question Answering

- Where do lobsters like to live?

The problem with automated Question Answering

- Where do lobsters like to live?
  - *on the table*

The problem with automated Question Answering

- Where do lobsters like to live?
  — *on the table*
- Where are zebras most likely found?

The problem with automated Question Answering

- Where do lobsters like to live?
  — *on the table*

- Where are zebras most likely found?
  — *in the dictionary*

The problem with automated Question Answering

- Where do lobsters like to live?
  — *on the table*
- Where are zebras most likely found?
  — *in the dictionary*
- What is an invertebrate?

The problem with automated Question Answering

Helps NLP systems

- Where do lobsters like to live?
  - *on the table*
- Where are zebras most likely found?
  - *in the dictionary*
- What is an invertebrate?
  - *Dukakis*

Michael Dukakis is a member of the **Democratic Party**, I have long suspected that elected **officials from the Democratic Party** are some previously unclassified form of **invertebrate**, a totally spineless creature capable of great noise but no real movement or action

Helps humans retrieve information faster

How are Max Planck, Angela Merkel and Dalai Lama related?

All have doctoral degrees from German universities

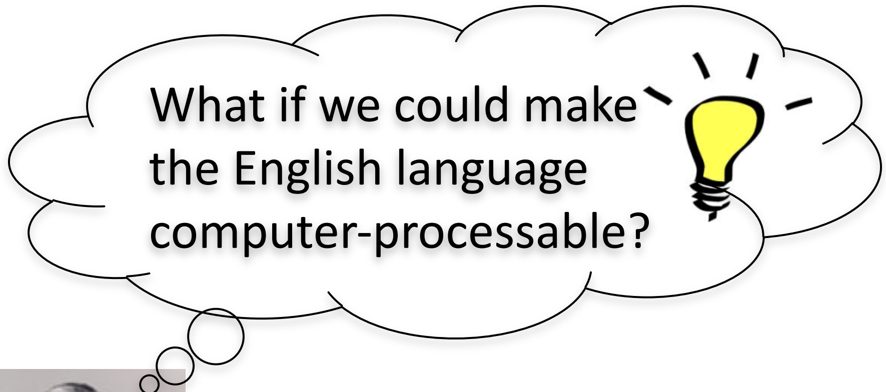
platypus

echidna

Tell me mammals that lay eggs

WordNet Semantic Classes

WordNet



George Miller

- started in 1985
- Cognitive Science Laboratory, Princeton University
- written by lexicographers
- goal: support automatic text analysis and AI applications

[Miller, CACM 1995]

WordNet

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: dog Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

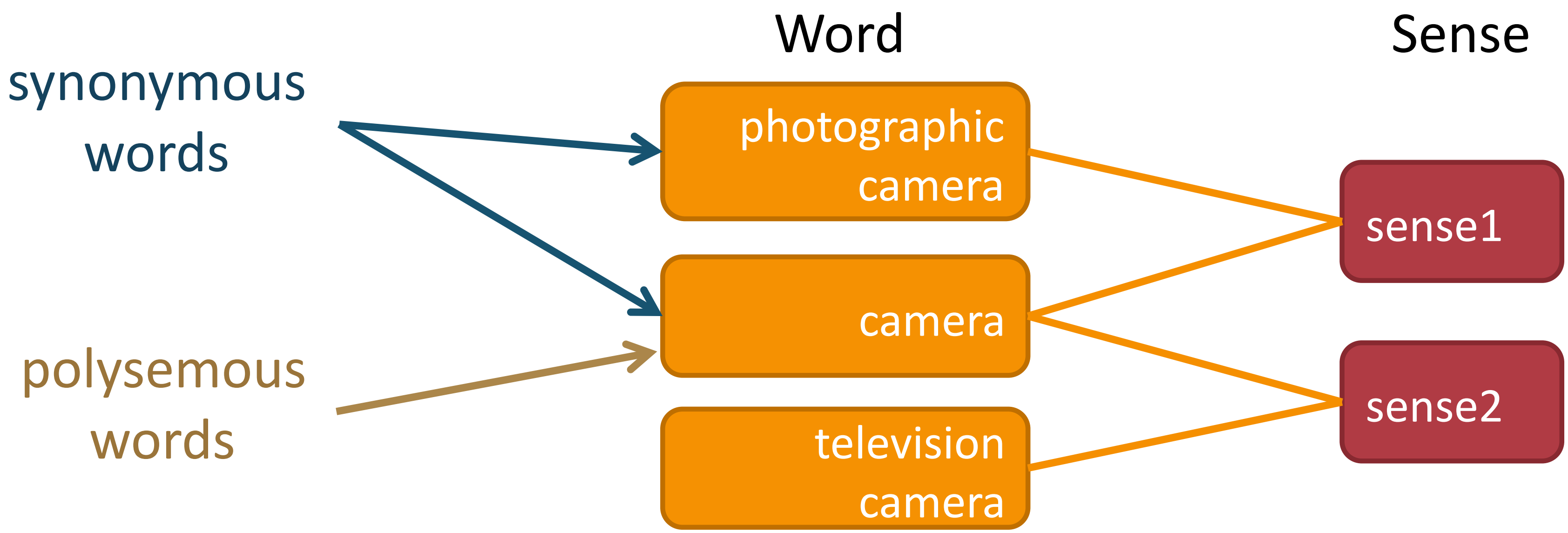
Noun

- S: (n) **dog**, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
- S: (n) frump, **dog** (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"*
- S: (n) **dog** (informal term for a man) *"you lucky dog"*
- S: (n) cad, bounder, blackguard, **dog**, hound, heel (someone who is morally reprehensible) *"you dirty dog"*
- S: (n) frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) pawl, detent, click, **dog** (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) andiron, firedog, **dog**, dog-iron (metal supports for logs in a fireplace) *"the andirons were too hot to touch"*

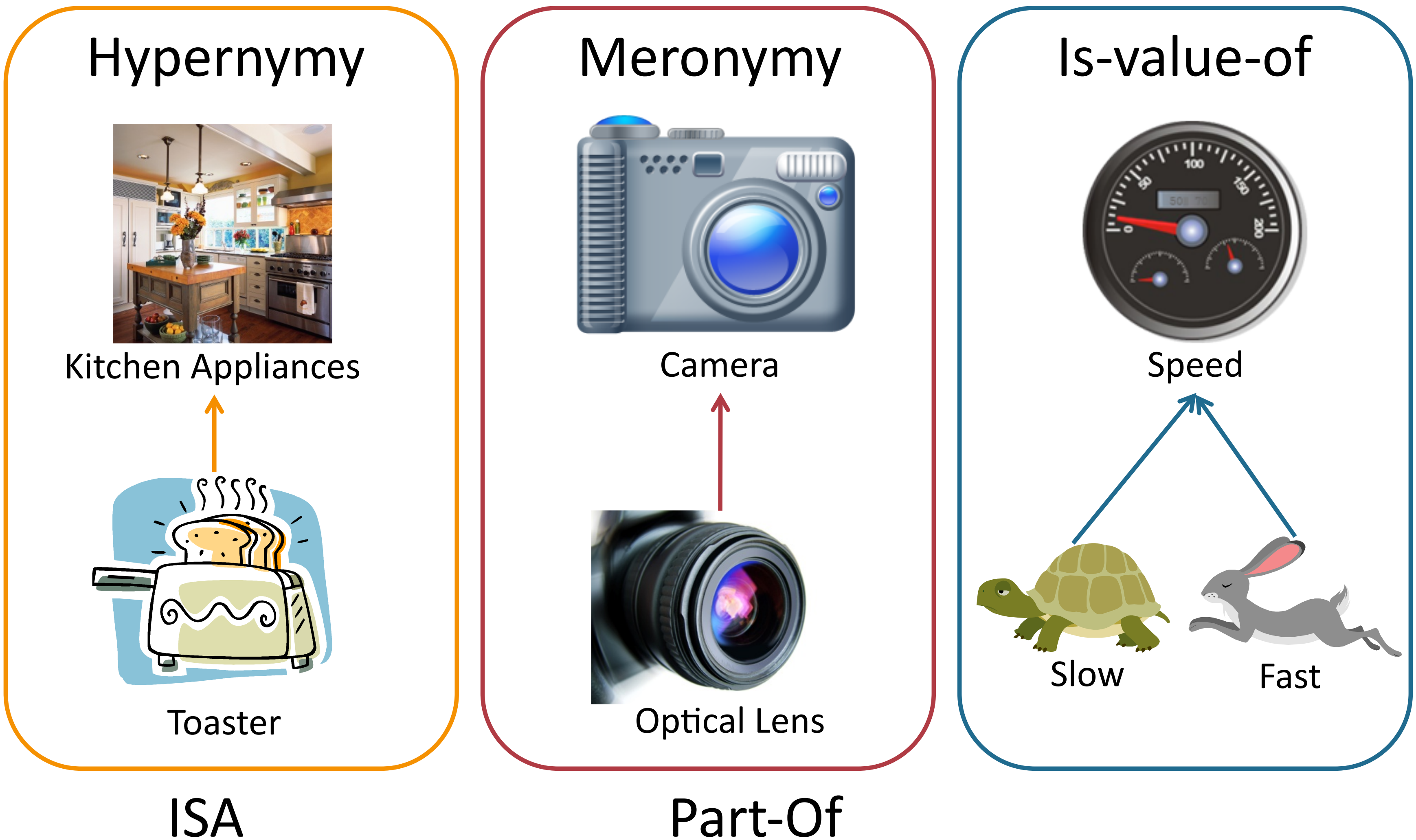
Verb

- S: (v) chase, chase after, trail, tail, tag, give chase, **dog**, go after, track (go after with the intent to catch) *"The policeman chased the mugger down the alley"; "the dog chased the rabbit"*

WordNet: Lexical Database



WordNet: Semantic Relations



But WordNet is not enough …

But WordNet is not enough …

- Has limited coverage
  – many instances and classes are missing
  – not all relations are listed
  – knowledge does not cover all domains

Ex. if you are interested in extracting:

- *all names of US presidents, you will notice that the name of Barack Obama is not present*
- *Chinese, French, Italian presidents, you will notice that these classes and their instances are not listed at all*
- *the is more information present for **animals** than **people***

Towards Automated Semantic Class Learning

Necessity for Automated Methods

- Even the biggest knowledge repository must be constantly updated, over time instances of a class may change

Ex. Presidents of a Country

- Barack Obama (2009-present)
- George Bush (2001-2009)

Country Names

- Czechoslovakia (1918-1992)
- Spain

General Characteristics

- Semantic classes are diverse:
  - closed
    - small (names of countries, states, planets)
    - large (names of diseases, cities)
  - open

    Ex. singers, movie titles

- Users might not know sample instance of a class
- An instance can belong to multiple classes

  Ex. orange the *fruit* vs. orange the *color*

The Challenge

- Relevant information is scattered across multiple Web pages
- Can we create an automated procedure that will acquire the necessary knowledge ?
- How does one evaluate precision and recall for the harvested information?
  - currently no repository contains all the information

Lexico-Syntactic Patterns (Hearst 92)

(S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

Lexico-Syntactic Patterns (Hearst 92)

(S1) `Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.`

(1a) NP_0 *such as* NP_1 $\{, NP_2 \ldots$, *(and | or)* $NP_i\}$ $\quad i \geq 1$

are such that they imply

(1b) *for all* NP_i, $i \geq 1$, *hyponym*(NP_i, NP_0)

Thus from sentence (S1) we conclude

hyponym(*"Gelidium"*, *"red algae"*).

Examples are adapted from Marti Hearst

Lexico-Syntactic Patterns (Hearst 92)

(2) *such NP as* $\{NP\ ,\}^*$ $\{$*(or* | *and)*$\}$ *NP*

`... works by such authors as Herrick, Goldsmith, and Shakespeare.`

Lexico-Syntactic Patterns (Hearst 92)

(2) *such NP as* $\{NP\ ,\}^*$ $\{(or \mid and)\}$ *NP*

`... works by such authors as Herrick, Goldsmith, and Shakespeare.`

$\Longrightarrow$ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

Lexico-Syntactic Patterns (Hearst 92)

(2) *such NP as {NP ,}* {(or | and)} NP*

```
... works by such authors as Herrick, Goldsmith, and Shakespeare.
```

$\Longrightarrow$ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

(3) *NP {, NP}* {,} or other NP*

```
Bruises, ..., broken bones or other injuries ...
```

$\Longrightarrow$ *hyponym("bruise", "injury"),*
hyponym("broken bone", "injury")

Examples are adapted from Marti Hearst

Estimating Pattern Reliability

- What is a good pattern?

Estimating Pattern Reliability

- What is a good pattern?
  - should occur frequently in text
  - should (nearly) always suggest the relation of interest
  - should be recognizable with little pre-encoded knowledge

Examples

- **Cities** such as **Boston**, **Los Angeles**, and **Seattle**…"

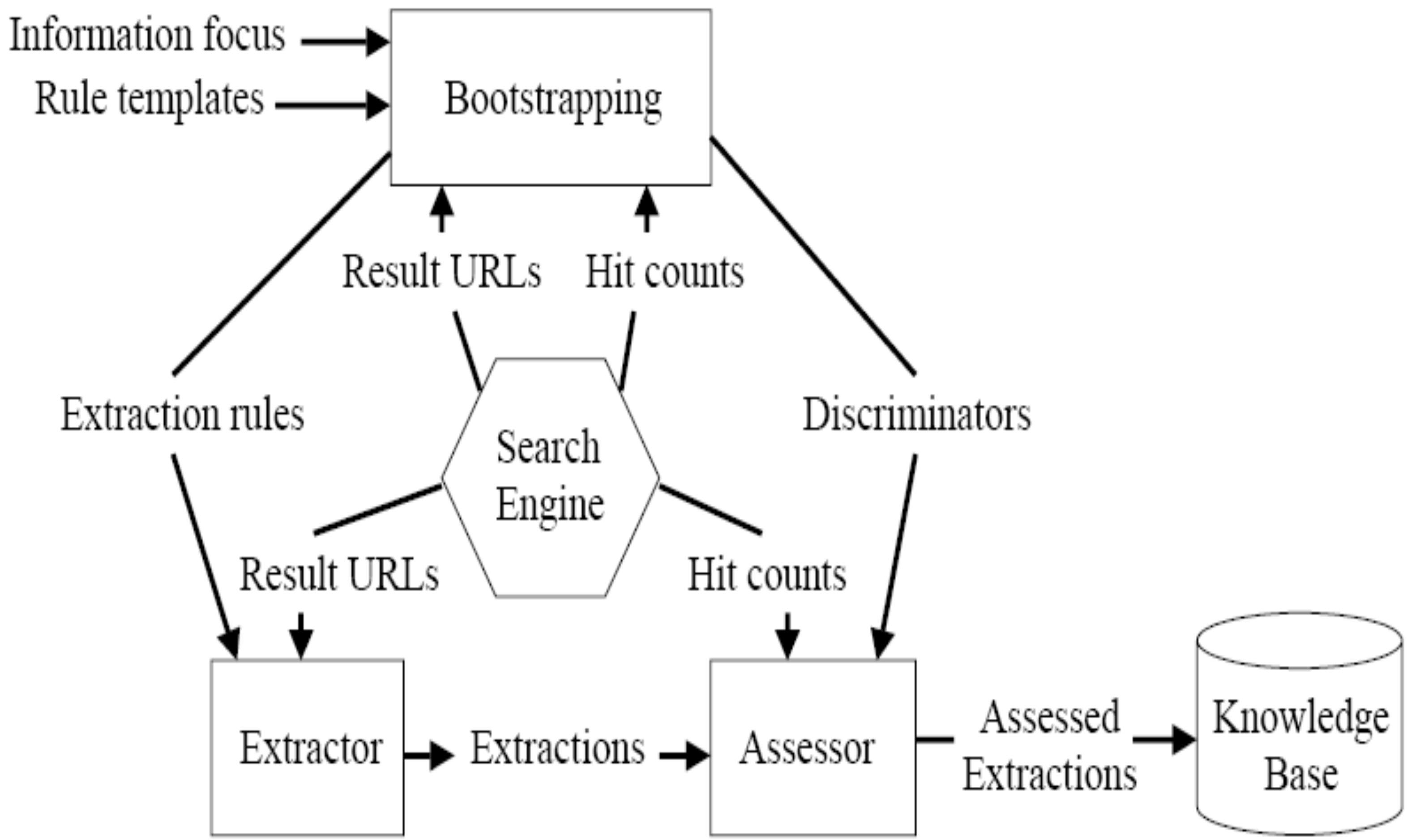
("**C** such as **NP1**, **NP2**, and **NP3**") => IS-A(each(head(**NP**)), **C**)

- Detailed information for several **countries** such as **maps**

- I listen to pretty much all music but prefer **country** such as **Garth Brooks**

Semantic Class Learning from the Web

KnowItAll Architecture (Etzioni et al.05)



Lets Learn City Names

- Input:
  – search query:
    - “*city; town*”, “*cities; towns*”
  – use all extraction rules following (Hearst 92):
    - <*class2*> such as <NPList>
    - <NP> is a <*class1*>
    - <*class2*> including <NPList>
- Generate extraction queries for any search engine:
  – “*cities* such as”
  – “is a *town*”
  – “*towns* including”

Learning City Names

- Submit extraction queries to Google and collect the returned snippets:

Central Highlands Council - Welcome - Enjoy the historic buildings ...
Enjoy historic buildings and friendly **towns including** Bothwell, Hamilton, Gretna and Ellendale to name a few. Fish at great fishing spots.
www.centralhighlands.tas.gov.au/ - Cached - Similar

Wichita, Kansas RE/MAX Agent serving Wichita and surrounding **towns** ...
Wichita, Kansas RE/MAX realtor serving Wichita, Goddard, Maize, Bentley, Halstead, Sedgwick, Park City, Valley Center, Bel Aire, Andover, Derby, Rose Hill, ...
www.wichitarealestate4you.net/ - Cached

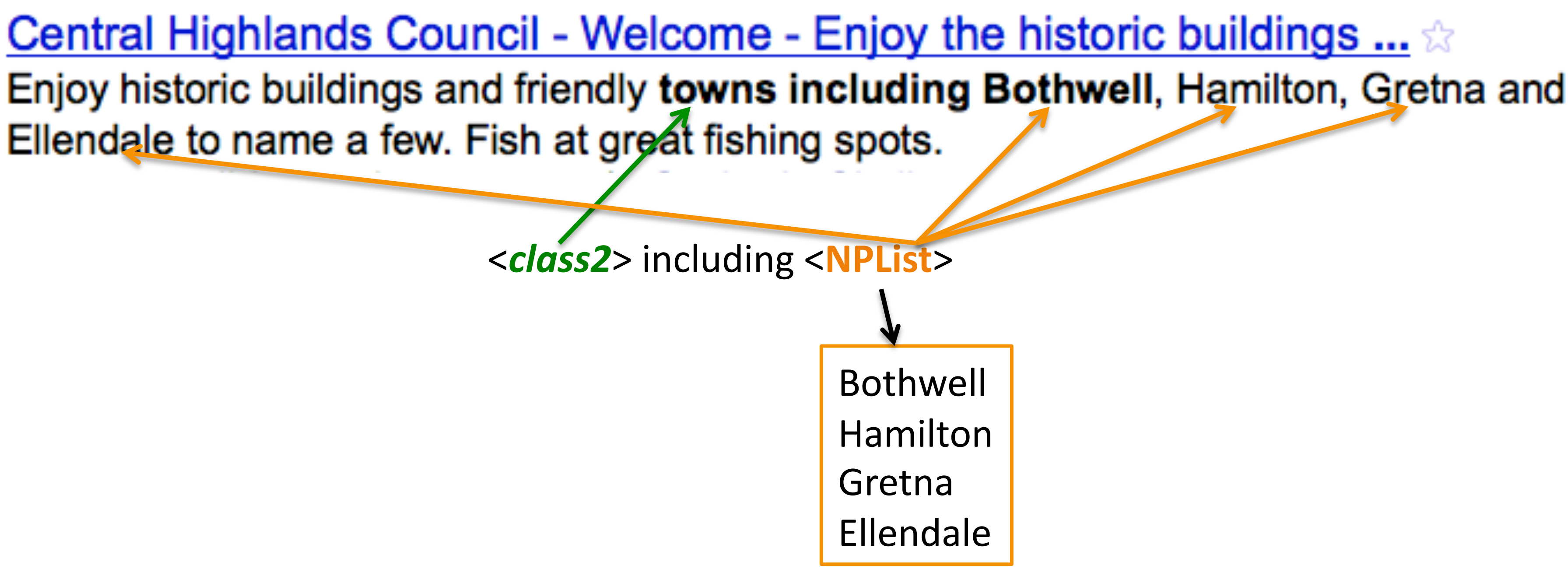
Public Health And Poor-Law Medical Services
towns, including London,. 6,I44 births and 5,I67 deaths were registered during the week ending Saturday,. July 25th. The annual rate of mortality ...
www.jstor.org/stable/20236873

John D. Williams, M.D., B.Sc.Edin., Honorary Gynæcologist To The ...
by JWB - 1901
towns, including London,. 6561 births and 3674 deaths were registered during the week ending Saturday last, May 25th. The annual rate of mortality ...
www.jstor.org/stable/20268562

Sanitary and meteorological notes
annually of 21"2 in twenty-eight large English **towns (including** London, in which the rate was 19"7), 30"8 in the sixteen chief towns of Ireland, ...
www.springerlink.com/index/30401P77HV34488X.pdf

Extracting City Names

- Pull all ***candidate*** city names from the snippets using extraction rules



Assessing Candidates

- **STEP1:** Generate *discriminators* from the rules and the user input
  - cities such as <*Candidate*>
  - <*Candidate*> is a town
  - <*Candidate*> is a city
  - towns including <*Candidate*>

Assessing Candidates

- **STEP2:** Generate *discriminator queries* from the discriminators and the extracted candidates
  - cities such as *London*
  - *London* is a town
  - *London* is a city
  - towns including *London*

Assessing Candidates

- **STEP3:** Evaluate each *candidate* with each *discriminator query* and compute PMI *as:*

$$PMI(Cnd, Disc) = \frac{|Hits(Disc + Cnd)|}{|Hits(Cnd)}$$

$$PMI(London, city) = \frac{Hits(city \quad London)}{Hits(London)} = \frac{8{,}590{,}000}{533{,}000{,}000} = 0.0161$$

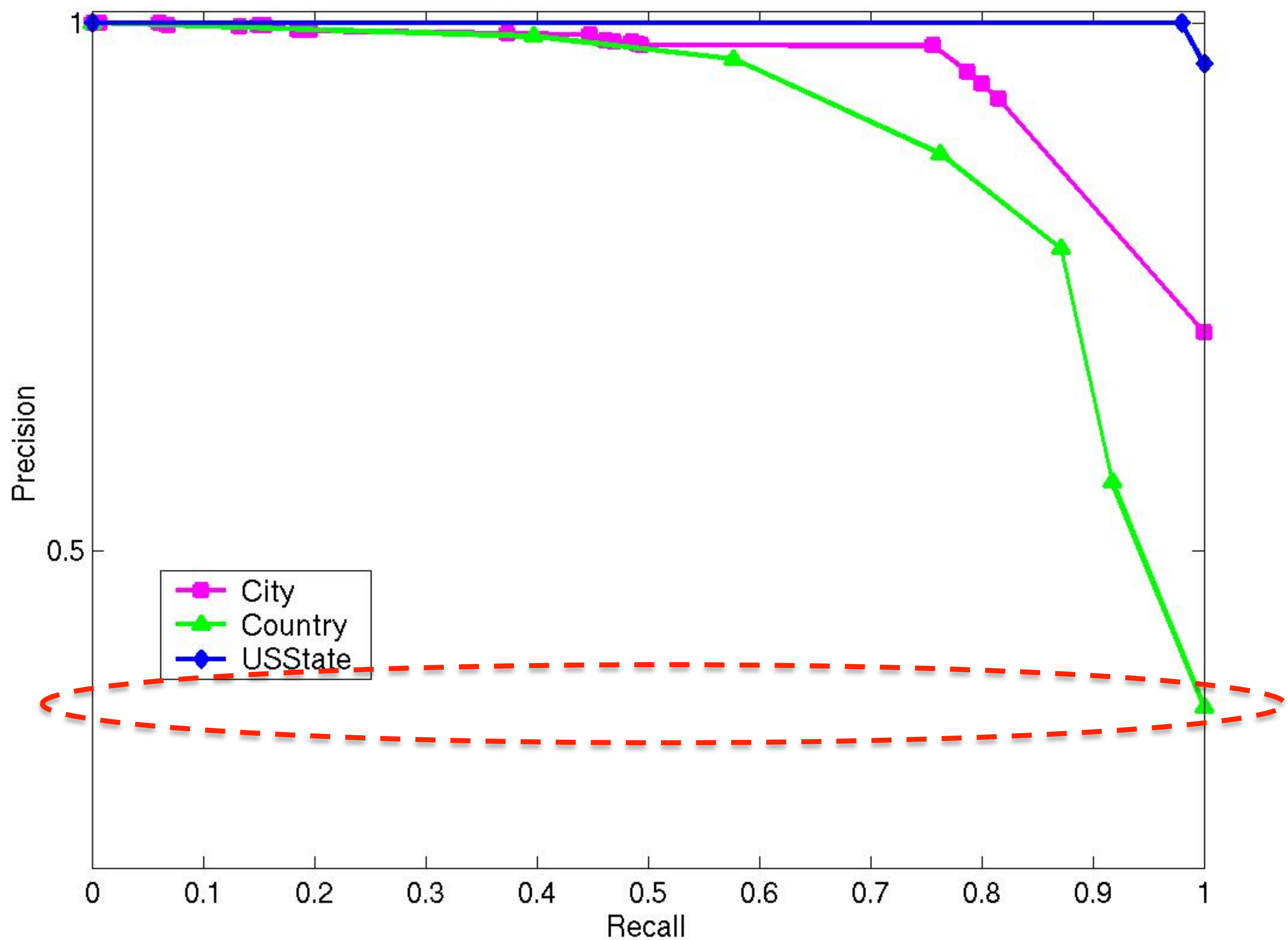
$$PMI(Avocado, city) = \frac{Hits(city \quad Avocado)}{Hits(Avocado)} = \frac{5{,}980}{8{,}320{,}000} = 0.000718$$

$$PMI(London, city) >> PMI(Avocado, city)$$

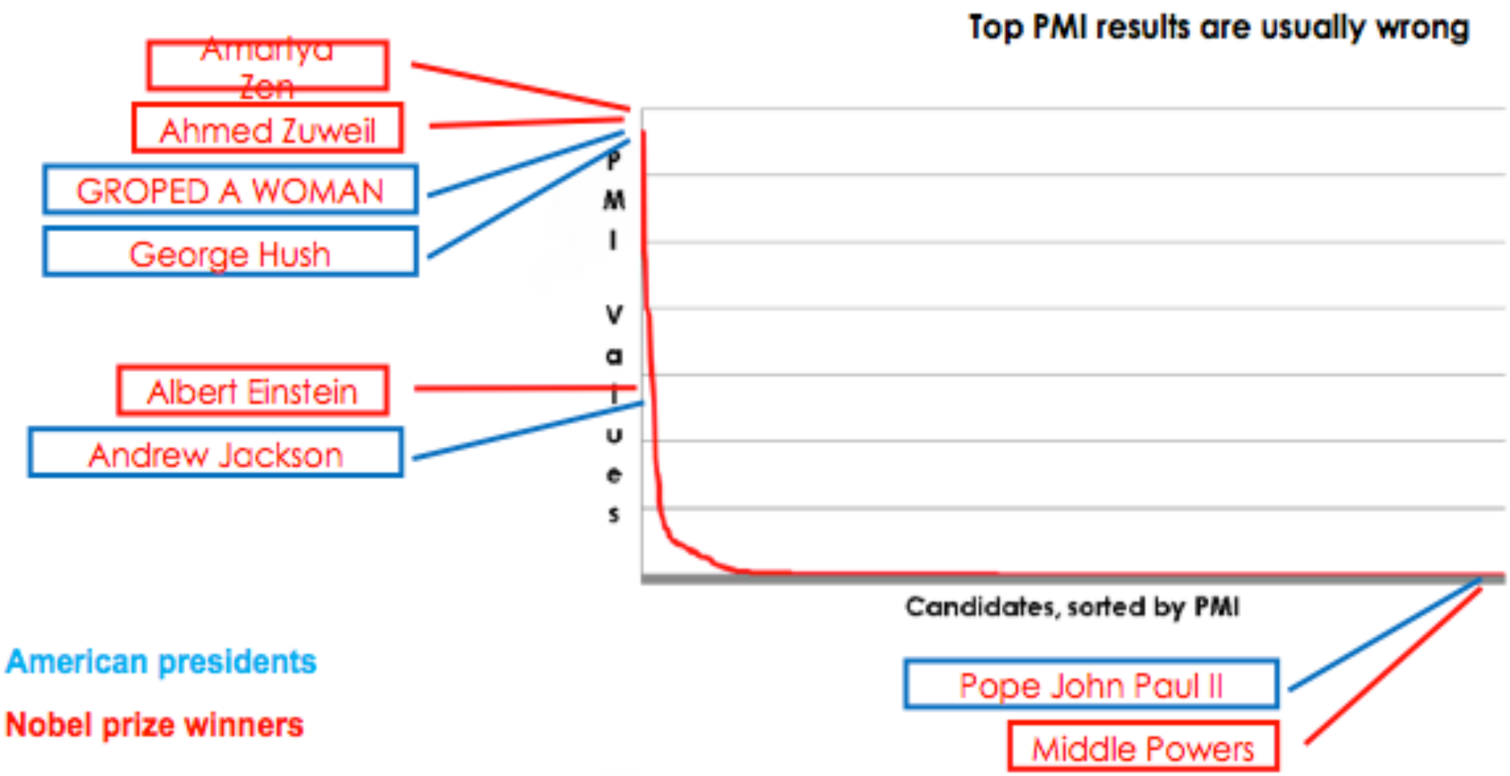
Assessing Candidates

- Train NaïveBayes classifier using PMI as features
- Training set contains positive and negative instances of the class
  - choose ***n*** candidates
  - compute average PMI, take ***m*** candidates with highest average PMI as positive examples and ***m*** candidates with lowest average PMI as negative examples
  - select ***k*** best discriminators tested on ***m***
- Evaluate all candidates on ***k*** discriminators

Results for City, Country and US State extraction



Errors due to Mutual Information



- Top PMI features are not always useful
- An extractor with high PMI can harvest wrong candidate examples

Open Questions

- How to choose synonyms for class expansion? (this can be tricky even for humans)
- How many seed examples are necessary to learn the instances of a class?
- How to eliminate ambiguous examples?
- Can we improve precision/recall?
- How well does the method scale?

Doubly-anchored pattern (DAP)

- doubly-anchored pattern

  “ *ClassName* such as ***<u>ClassInstance</u>*** and * ”

  - *ClassName* is the name of the semantic class to be learned
  - *ClassInstance* is an instance of the semantic class
  - (*) indicates the location of the extracted terms

Power of DAP

- virtually eliminates ambiguity, because the *ClassName* and the *ClassInstance* mutually disambiguate each other

languages such as *

English

C++

Java

Spanish

Power of DAP

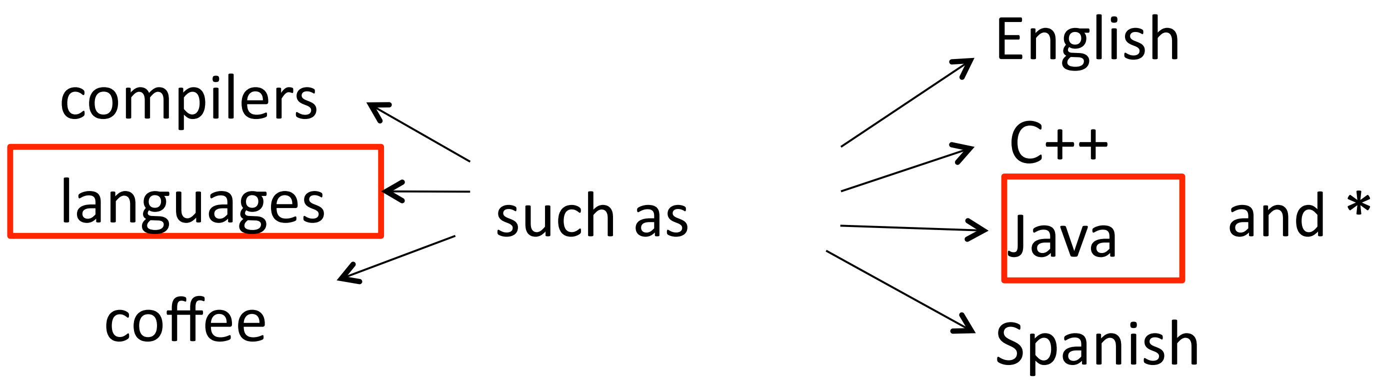
- virtually eliminates ambiguity, because the *ClassName* and the *ClassInstance* mutually disambiguate each other

compilers
languages
coffee

* such as Java

Power of DAP

- virtually eliminates ambiguity, because the *ClassName* and the *ClassInsance* mutually disambiguate each other



- it is more likely to generate instances of the desired list type

Power of DAP

- virtually eliminates ambiguity, because the *ClassName* and the *ClassInstance* mutually disambiguate each other
- it is more likely to generate instances of the desired list type
- increases the likelihood of finding true list construction

states such as Alabama and * { Mississippi, California, Texas, Arizona }

DAP characteristics

- Limitation(1): sparse data hurts recall
- Solution(1): collect evidence from the web

- Limitation(2): single class instance hurts recall
- Solution(2): incorporate bootstrapping

Bootstrapping

- Instantiate DAP with *ClassName* and one *<seed>* instance
- Feed the newly learned terms on *<seed>* position
- Conduct a breadth-first search

Texas

states such as *Mississippi* *Alabama* and *

Arkansas

Florida

Texas

California

Mississippi

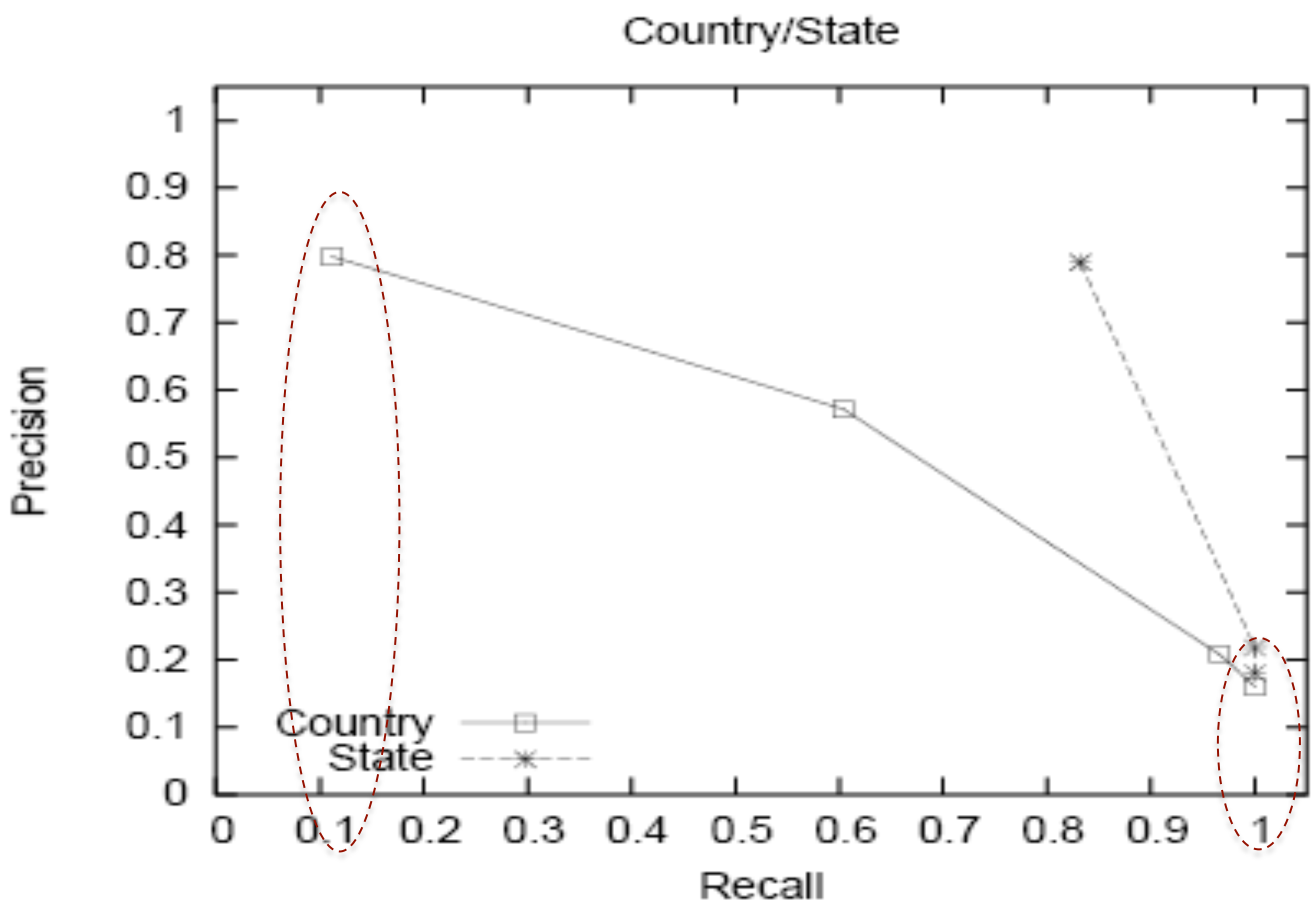
…

Arkansas

Tennessee

Nevada

Performance of Bootstrapping



Problem: search needs guidance

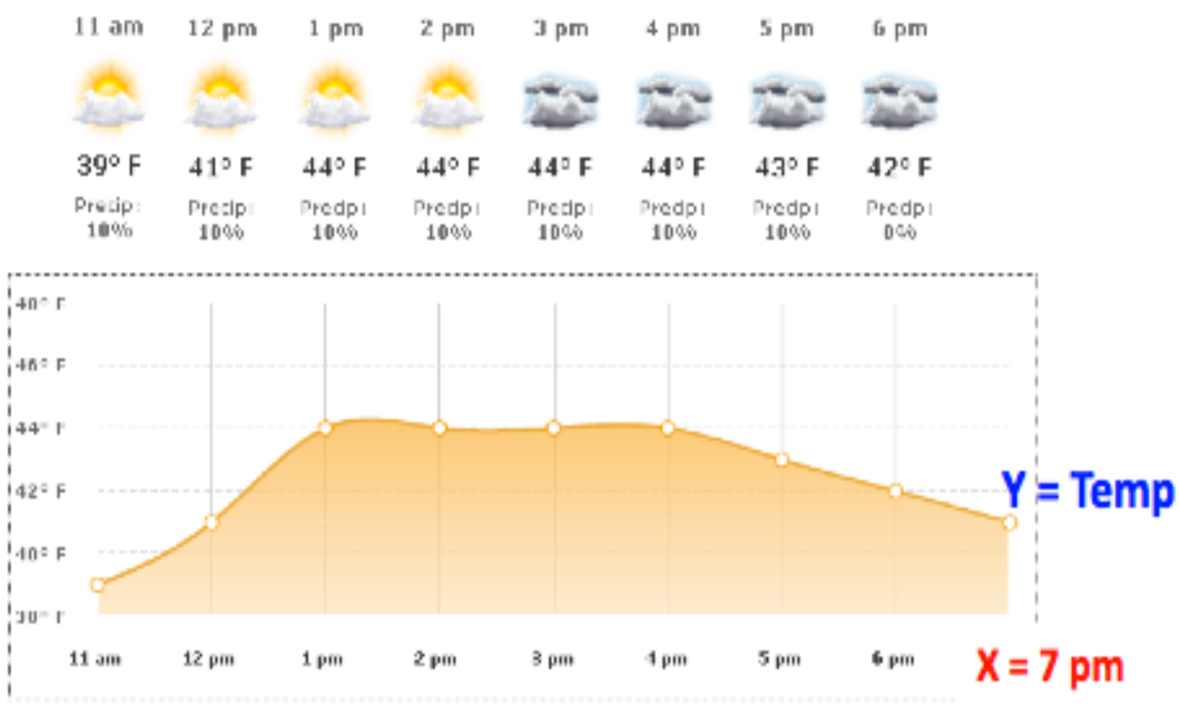
Solution: rank the learned instances

NEXT TIME WHEN WE SEE EACH OTHER

Regression

- We will talk about weather, flight or stock market predication systems

Weather Prediction

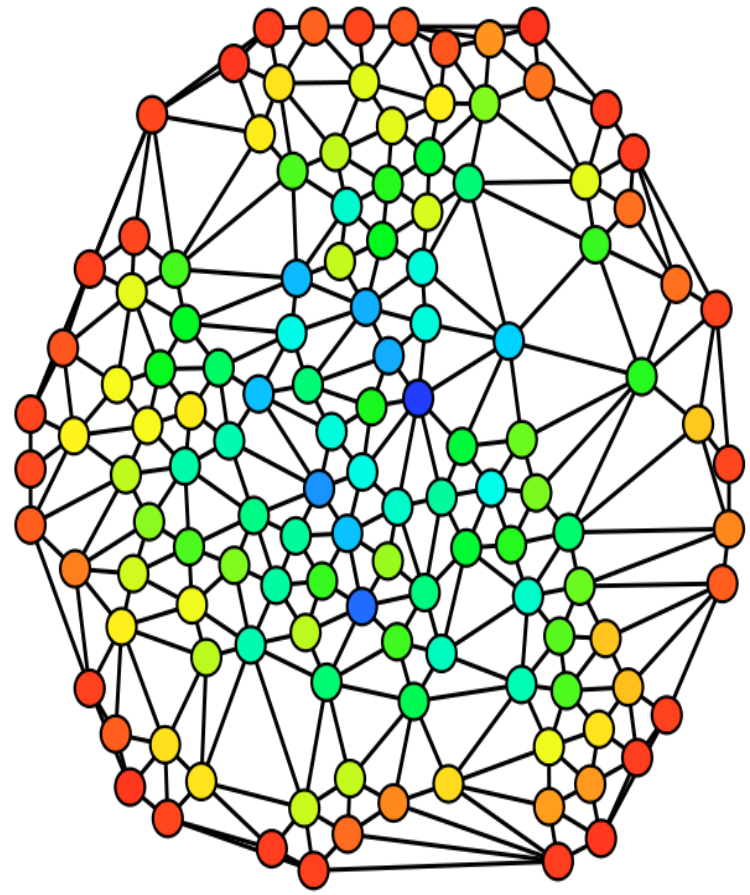


Graph Theory

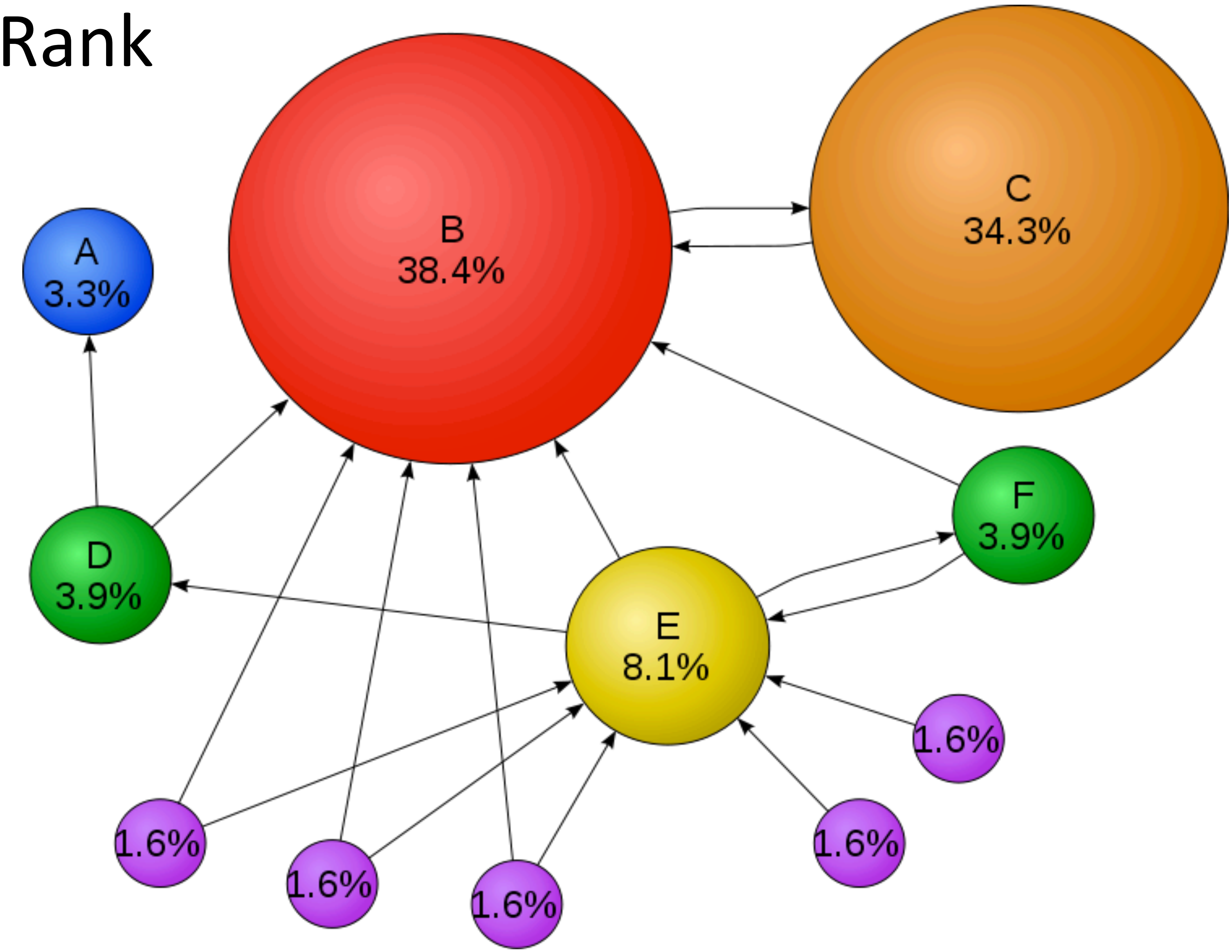
- General introduction (terminology)

- Directed Graphs
- Undirected Graphs

- Refresh shortest path algorithm

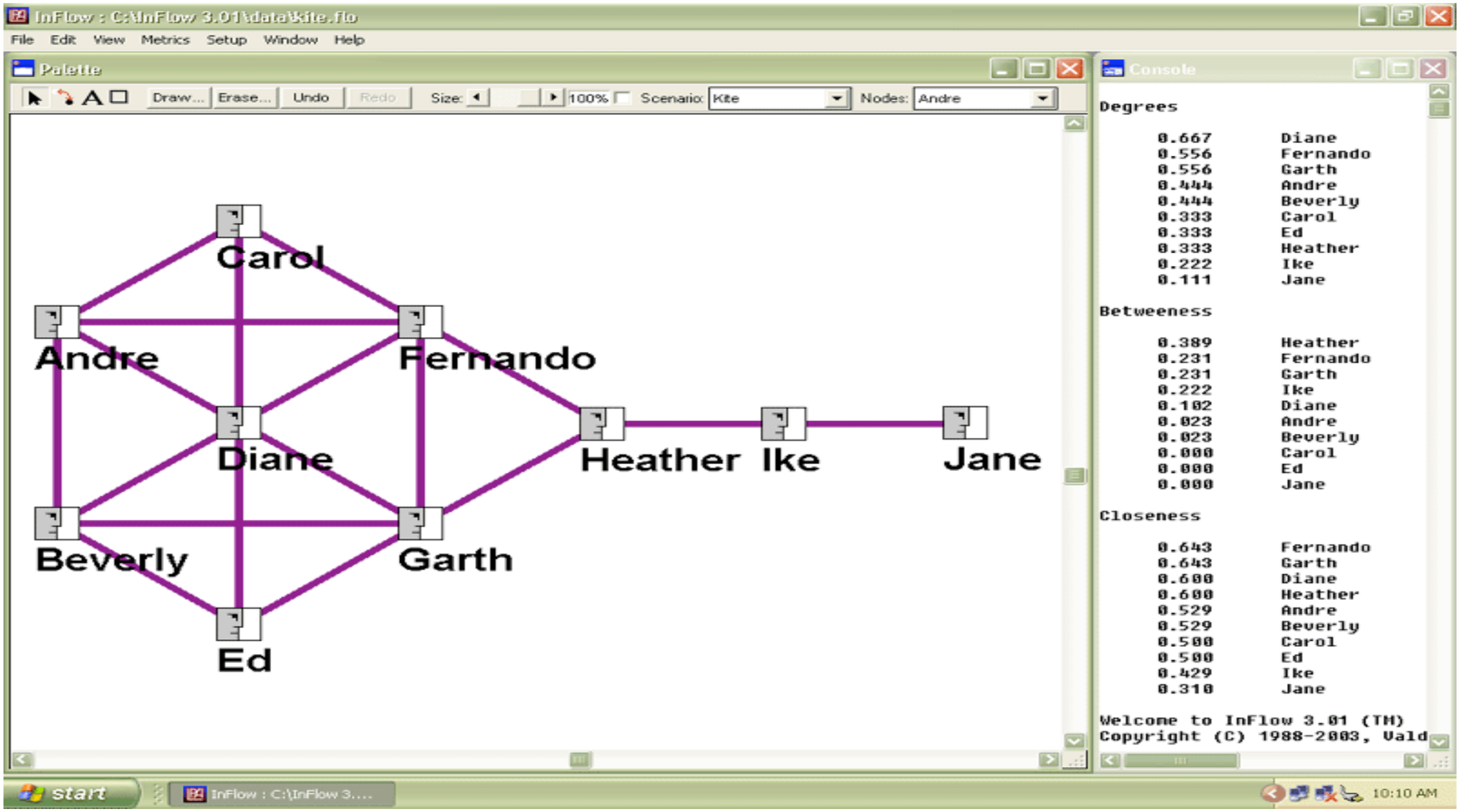


PageRank



created by Page and Brin

Centrality Measures



- Ever wondered how to eliminate gossip spreaders?
- Who is the most influential person in your friend circle?

What would you do with this knowledge?

- Identify influence of people on Facebook or any social network
- Trace e-mail topic exchange between people
- Learn how to rank Web pages or any information
- …

QUESTIONS ON HOMEWORK