Machine-Assisted Indexing

Week 12

LBSC 671

Creating Information Infrastructures

Machine-Assisted Indexing

- Goal: Automatically suggest descriptors
  - Better consistency with lower cost

- Approach: Rule-based expert system
  - Design thesaurus by hand in the usual way
  - Design an expert system to process text
    - String matching, proximity operators, …
  - Write rules for each thesaurus/collection/language
  - Try it out and fine tune the rules by hand

Machine-Assisted Indexing Example
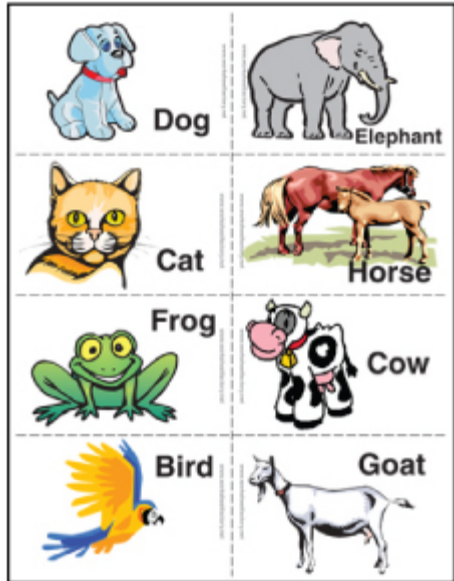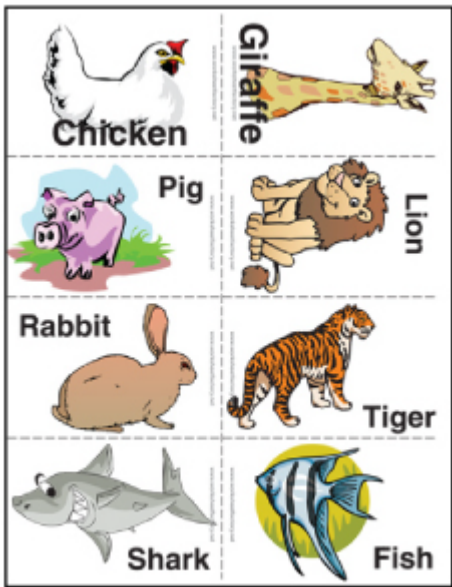
Access Innovations system:

```
//TEXT: science
IF (all caps)
  USE research policy
  USE community program
ENDIF
IF (near "Technology" AND with "Development")
  USE community development
  USE development aid
ENDIF
```

---

near: within 250 words
with: in the same sentence

Modeling Use of Language

- Normative
  - Observe how people **<u>do</u>** talk or write
    - Somehow, come to understand what they mean each time
  - Create a **<u>theory</u>** that associates language and meaning
  - Interpret language use based on that theory

- Descriptive
  - Observe how people **<u>do</u>** talk or write
    - Someone "trains" us on what they mean each time
  - Use **<u>statistics</u>** to learn how those are associated
  - Reverse the model to guess meaning from what's said

Chicken | Giraffe
Pig | Lion
Rabbit | Tiger
Shark | Fish

Dog | Elephant
Cat | Horse
Frog | Cow
Bird | Goat

Cute Mynah Bird Tricks

- Make scanned documents into e-text
- Make speech into e-text
- Make English e-text into Hindi e-text
- Make long e-text into short e-text
- Make e-text into hypertext
- Make e-text into metadata
- Make email into org charts
- Make pictures into captions
- …

Waikato researcher 'wikifies' the web

Software detects topics in documents and links to appropriate Wikipedia articles

By Computerworld staff, Auckland | Thursday, 27 November, 2008

A PhD student in computer science at the University of Waikato is working on ways to automatically "wikify" documents, by detecting the topics in the document and creating links to the appropriate Wikipedia articles.

David Milne, and his supervisor Professor Ian Witten, recently won an award for their paper at the Computers in Knowledge Management conference, held in California's Napa Valley. Close to 800 papers were submitted to the conference, which is organised annually by the Association of Computing Machinery (ACM).

Milne has used existing Wikipedia articles to "train" his software to make the same decisions as humans regarding what is important in any document, he says.

"Every single Wikipedia article is an example of how to cross-reference a document with Wikipedia," he says. "That means we have millions of examples for how to do the job."

Automatic systems face several hurdles, he says. They have to resolve ambiguity; to decide, for example, if the word "kiwi" refers to the bird, the fruit or to New Zealanders. They must also allow for polysemy — where there are different terms with similar

Wikification Demo Results

The Wikification system has identified the following entities with Wikipedia articles.
Click on an entity to visit the corresponding Wikipedia page.
Hover over links to view the categories associated with each entity.

Disambiguating **concepts** and **entities** in a context
sensitive way is a fundamental **problem**
in natural language processing. The comprehensiveness
of **Wikipedia** has made the online
encyclopedia an increasingly popular **target**
for **disambiguation**. **Disambiguation** to
Wikipedia is similar to a traditional Word
Sense **Disambiguation** task, but distinct in that
the **Wikipedia** link **structure** provides additional
information about which **disambiguations**
are compatible. In this work we analyze
approaches that utilize this information to arrive
at coherent **sets** of **disambiguations** for a
given **document** (which we call "global" **approaches**),
and compare them to more traditional
(local) **approaches**. We show that previous
approaches for global **disambiguation** can
be improved, but even then the local **disambiguation**
provides a baseline which is very
hard to beat.

http://cogcomp.cs.illinois.edu/demo/wikify/?id=25

Abraham Lincoln's Watch, around 1858

DESCRIPTION

Lincoln's English gold watch was purchased in the 1850s from George Chatterton, a Springfield, Illinois, jeweler. Lincoln was not considered to be outwardly vain, but the fine gold watch was a conspicuous symbol of his success as a lawyer.

The watch movement and case, as was often typical of the time, were produced separately. The movement was made in Liverpool, where a large watch industry manufactured watches of all grades. An unidentified American shop made the case. The Lincoln watch has one of the best grade movements made in England and can, if in good order, keep time to within a few seconds a day. The 18K case is of the best quality made in the US.

A Hidden Message

Just as news reached Washington that Confederate forces had fired on Fort Sumter on April 12, 1861, watchmaker Jonathan Dillon was repairing Abraham Lincoln's timepiece. Caught up in the moment, Dillon unscrewed the dial and engraved: "April 13, 1861 Fort Sumpter was attacked by the rebels on the above date J Dillon April 13, 1861 Washington" and "thank God we have a government Jonth Dillon."

In 1864 a second watchmaker, L. E. Gross, signed his name. Also, at some point someone etched "Jeff Davis" inside the watch, either as a joke or a statement of support for the Confederacy.

Lincoln never knew about the messages he carried in his watch. The inscription remained hidden behind the dial for over a century. After hearing from Jonathan Dillon's great-great-grandson, the Museum removed the dial on March 10, 2009, to reveal the watchmakers' declarations.

Gift of Lincoln Isham, great-grandson of Abraham Lincoln, 1958

PHYSICAL DESCRIPTION
gold; glass (watch material)

metal (key material)

gold (watch chain material)

wood; metal; fabric (box material)

MEASUREMENTS
watch: 2 in x 3 in x 1/2 in; 5.08 cm x 7.62 cm x 1.27 cm

watch chain: 13 in; 33.02 cm

ID NUMBER
PL*219098.01

CATALOG NUMBER
219098.01

ACCESSION NUMBER
219098

SUBJECT
Clothing & Accessories

Government, Politics, and Reform

Selections from the Abraham Lincoln Collection

EVENT
Battle of Fort Sumter, 1861

SEE MORE ITEMS IN
Political History: Political History, Presidential History Collection

Selections from the Abraham Lincoln Collection

DATA SOURCE
National Museum of American History, Kenneth E. Behring Center

CREDIT LINE
gift of Lincoln Isham, great-grandson of Abraham Lincoln, 1958

RELATED PUBLICATION
Rubenstein, Harry R.. Abraham Lincoln: An Extraordinary Life

http://americanhistory.si.edu/collections/search/object/nmah_516567

Lincoln's English gold watch was purchased in the 1850s from George Chatterton, a Springfield, Illinois, jeweler. Lincoln was not considered to be outwardly vain, but the fine gold watch was a conspicuous **symbol** of his success as a lawyer.

The watch movement and **case**, as was often typical of the time, were produced separately. The **movement** was made in Liverpool, where a large watch industry manufactured watches of all grades. An unidentified American **shop** made the **case**. The Lincoln watch has one of the best grade **movements** made in England and can, if in **good order**, keep time to within a few **seconds** a day. The 18K **case** is of the best **quality** made in the US.

A Hidden **Message**

Just as news reached Washington that Confederate **forces** had fired on Fort Sumter on April 12, 1861, watchmaker Jonathan Dillon was repairing Abraham Lincoln's timepiece. Caught up in …

NEIL A. ARMSTRONG 🔊

INTERVIEWED BY DR. STEPHEN E. AMBROSE AND DR. DOUGLAS BRINKLEY
HOUSTON, TEXAS – 19 SEPTEMBER 2001

ARMSTRONG: I'd always said to colleagues and friends that one day I'd go back to the university. I've done a little teaching before. There were a lot of opportunities, but the University of Cincinnati invited me to go there as a faculty member and pretty much gave me carte blanche to do what I wanted to do. I spent nearly a decade there teaching engineering. I really enjoyed it. I love to teach. I love the kids, only they were smarter than I was, which made it a challenge. But I found the governance unexpectedly difficult, and I was poorly prepared and trained to handle some of the aspects, not the teaching, but just the—universities operate differently than the world I came from, and after doing it—and actually, I stayed in that job longer than any job I'd ever had up to that point, but I decided it was time for me to go on and try some other things.

AMBROSE: Well, dealing with administrators and then dealing with your colleagues, I know—but Dwight Eisenhower was convinced to take the presidency of Columbia [University, New York, New York] by Tom Watson when he retired as chief of staff in 1948, and he once told me, he said, "You know, I thought there was a lot of red tape in the army, then I became a college president." He said, "I thought we used to have awful arguments in there about who to put into what position." Have you ever been with a bunch of deans when they're talking about—

ARMSTRONG: Yes. And, you know, there's a lot of constituencies, all with different perspectives, and it's quite a challenge.

Supervised Machine Learning

Steven Bird et al., *Natural Language Processing*, 2006
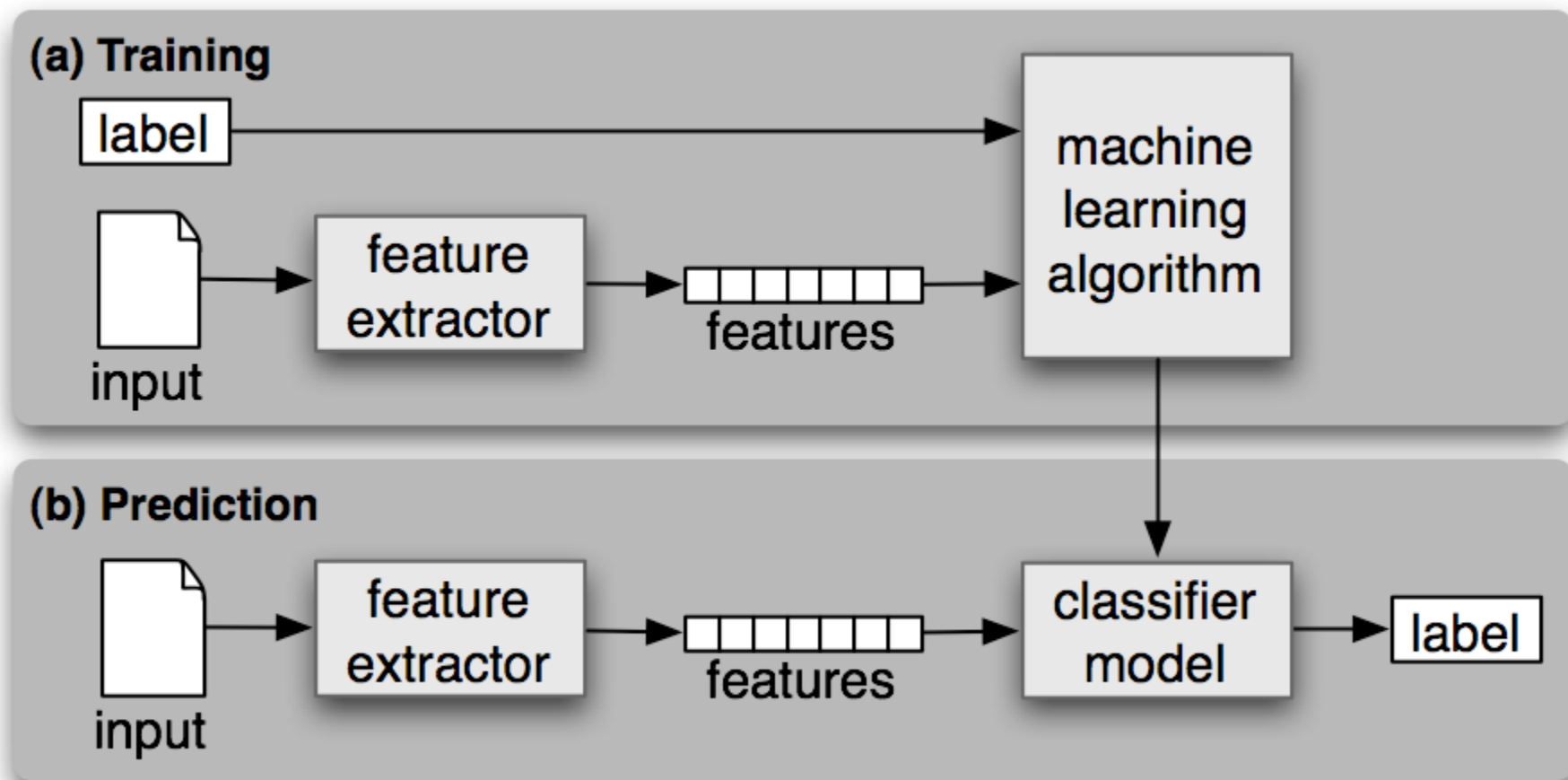
Rule Induction

- Automatically derived Boolean profiles
  - (Hopefully) effective <u>and</u> easily explained

- <u>Specificity</u> from the "perfect query"
  - AND terms in a document, OR the documents

- <u>Generality</u> from a bias favoring short profiles
  - e.g., penalize rules with more Boolean operators
  - Balanced by rewards for precision, recall, …

Statistical Classification

- Represent documents as vectors
  - e.g., based on TF, IDF, Length

- Build a statistical model for each label
  - e.g., a "vector space"

- Use that model to label new instances
  - e.g., by largest inner product

Machine Learning for Classification: The k-Nearest-Neighbor Classifier
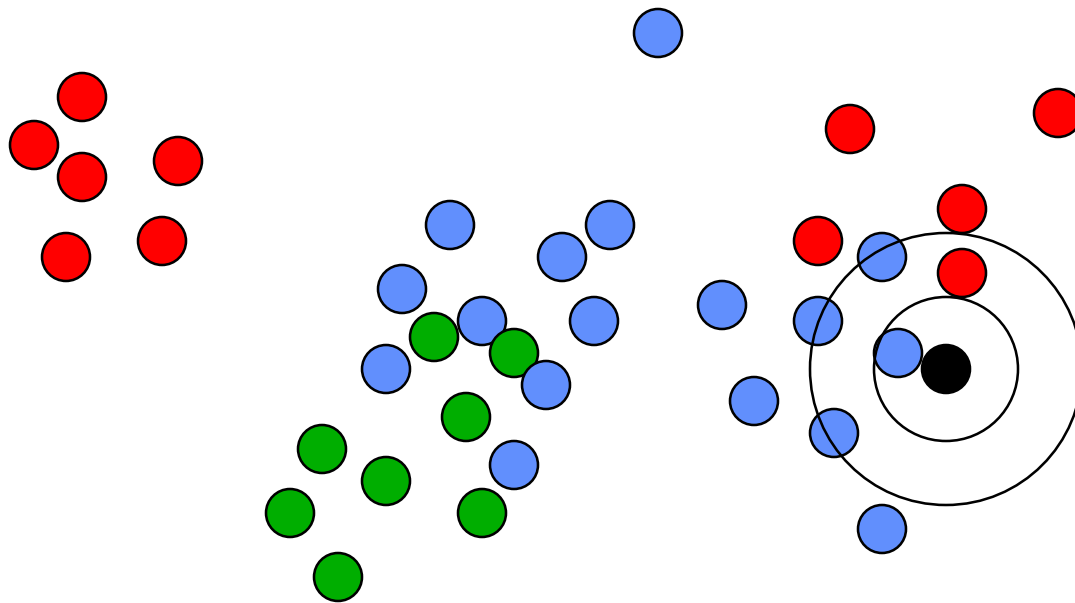
Machine Learning Techniques

- Hill climbing (Rocchio)
- Instance-based learning (kNN)
- Rule induction
- Statistical classification
- Regression
- Neural networks
- Genetic algorithms

Vector space example: query "canine" (1)
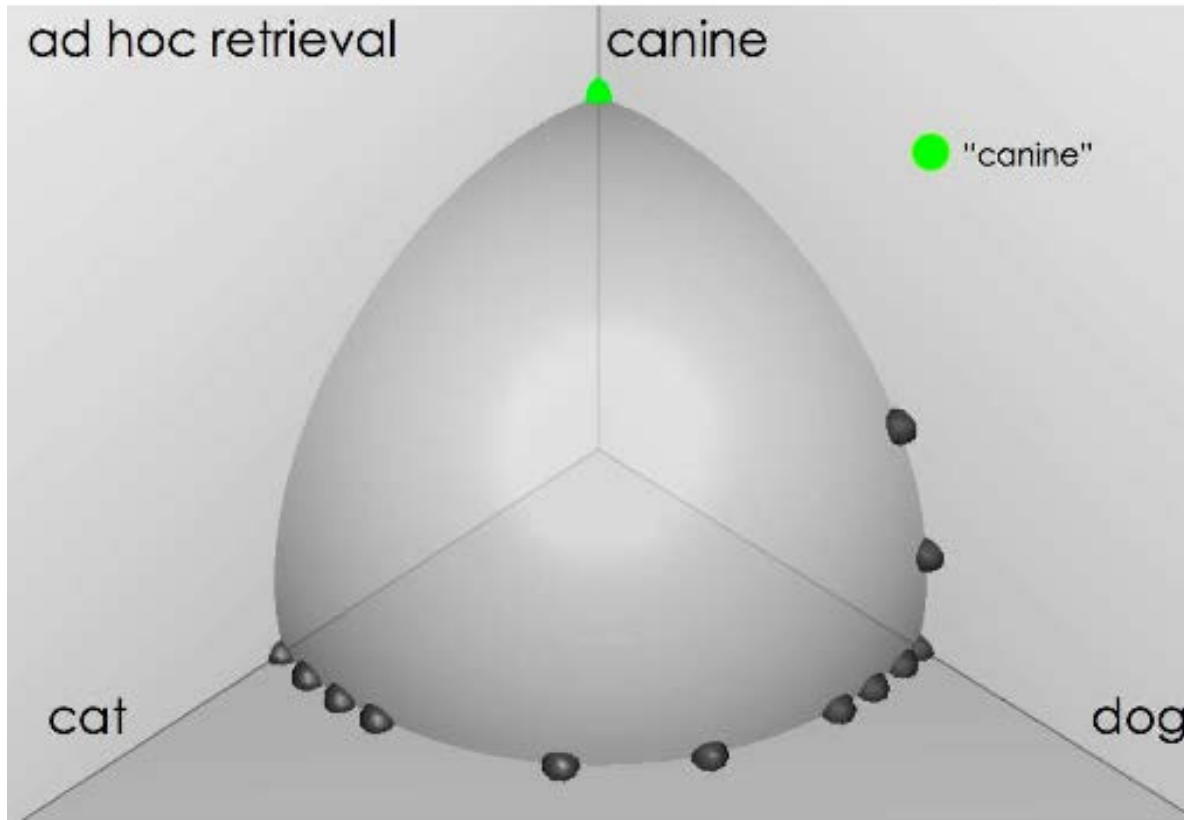
Source:

Fernando Díaz

Similarity of docs to query "canine"
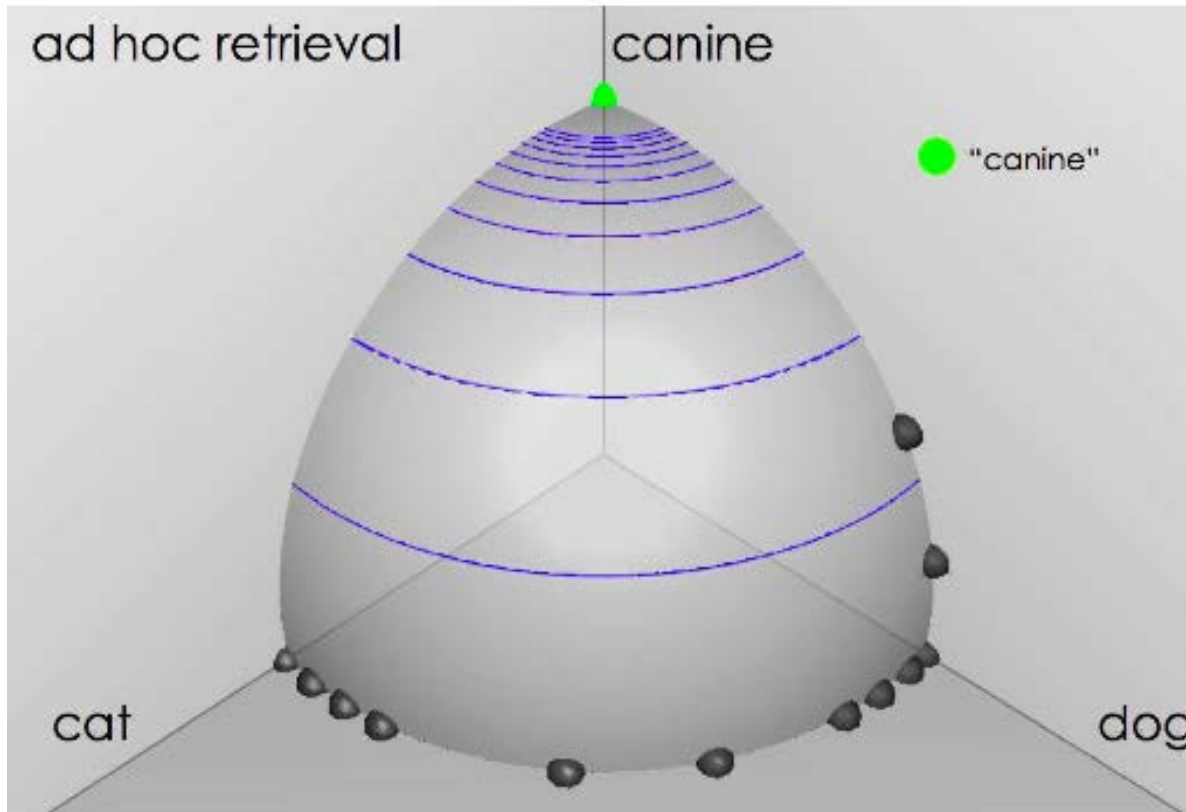
Source:

Fernando Díaz

User feedback: Select relevant documents
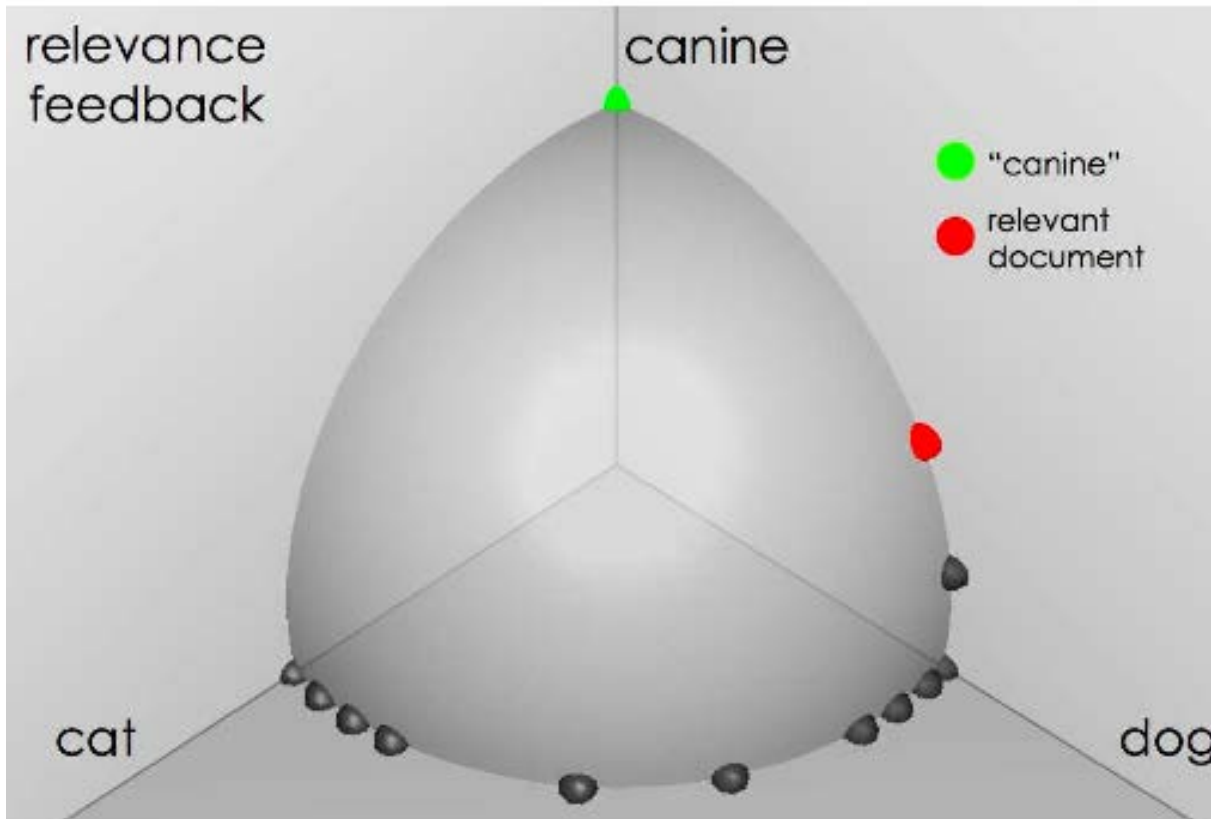
Source:

Fernando Díaz

Results after relevance feedback
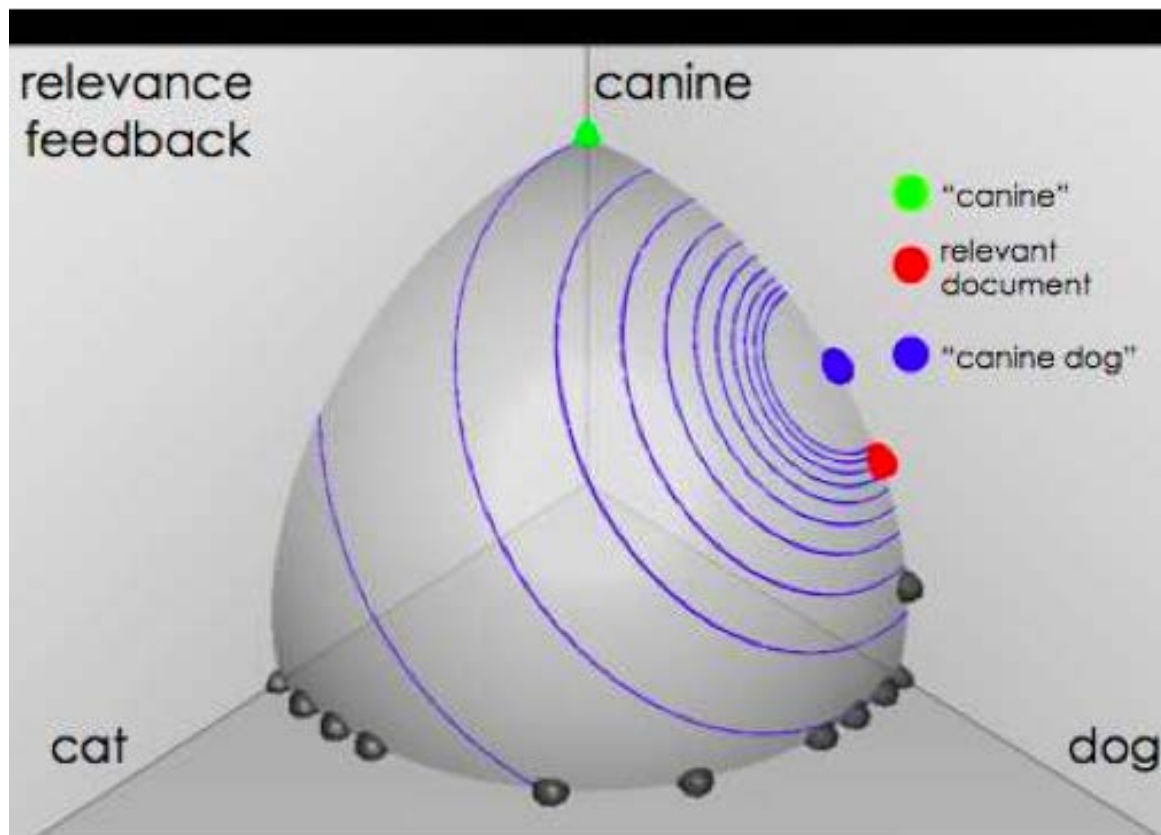
Source:

Fernando Díaz

Rocchio' illustrated
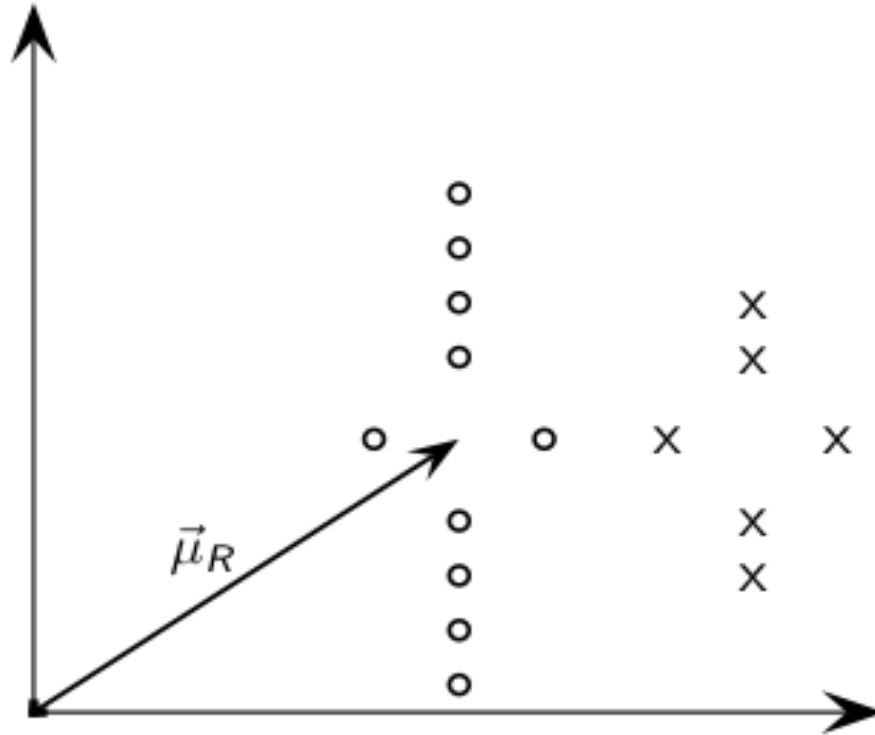
$\vec{\mu}_R$: centroid of relevant documents

Rocchio' illustrated
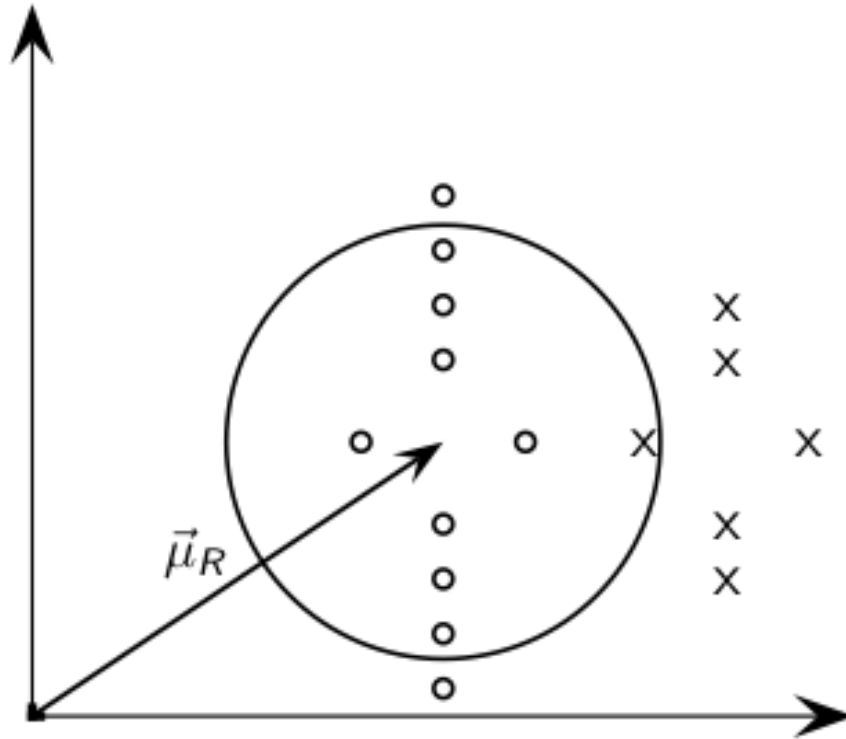
$\vec{\mu}_R$ does not separate relevant / nonrelevant.

Rocchio' illustrated
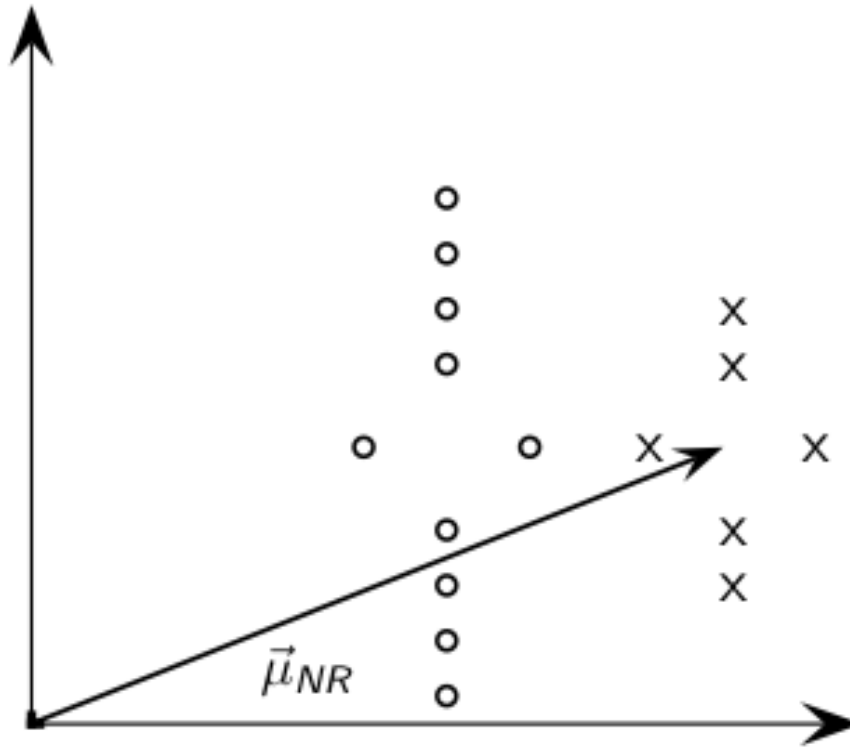
$\vec{\mu}_{NR}$: centroid of nonrelevant documents.

Rocchio' illustrated
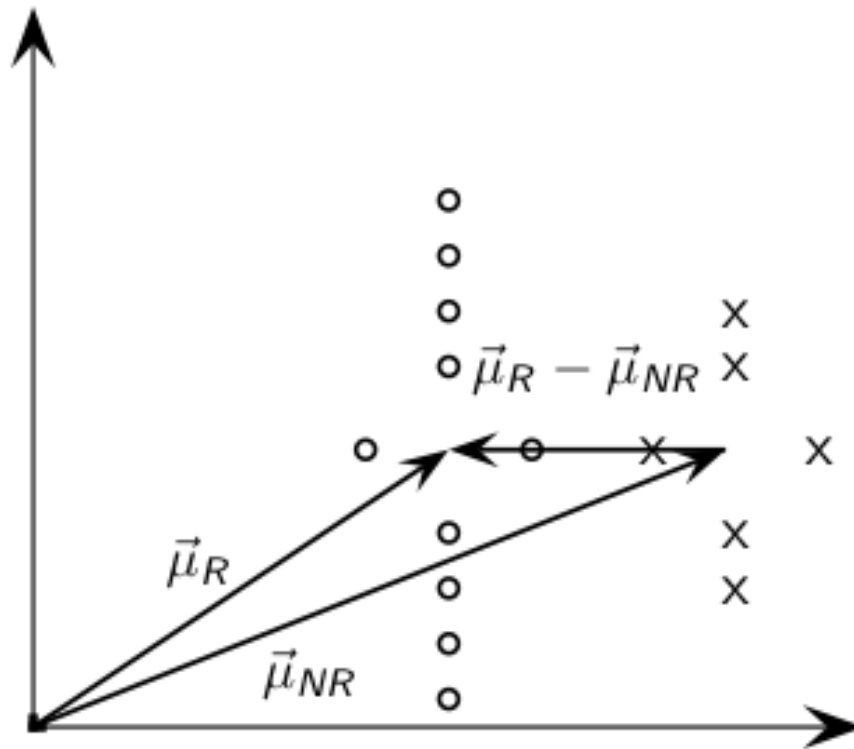
$$\vec{\mu}_R \quad \vec{\mu}_{NR}: \quad \text{difference vector}$$

Rocchio' illustrated

$\vec{\mu}_R - \vec{\mu}_{NR}$

$\vec{\mu}_R$

$\vec{\mu}_{NR}$
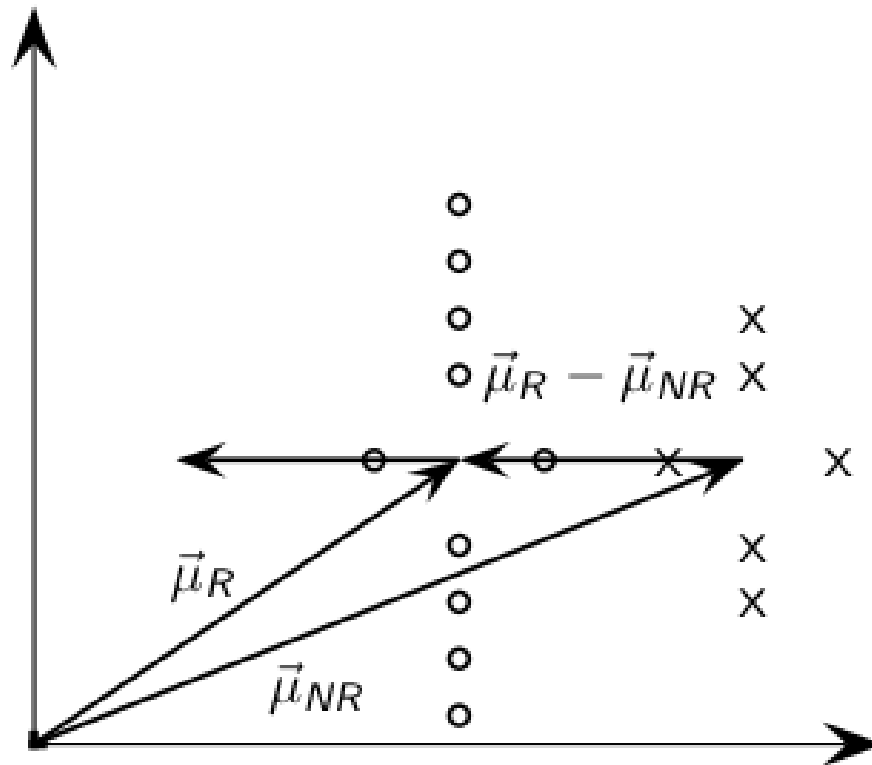
Add difference vector to $\vec{\mu}_R$ …

Rocchio' illustrated
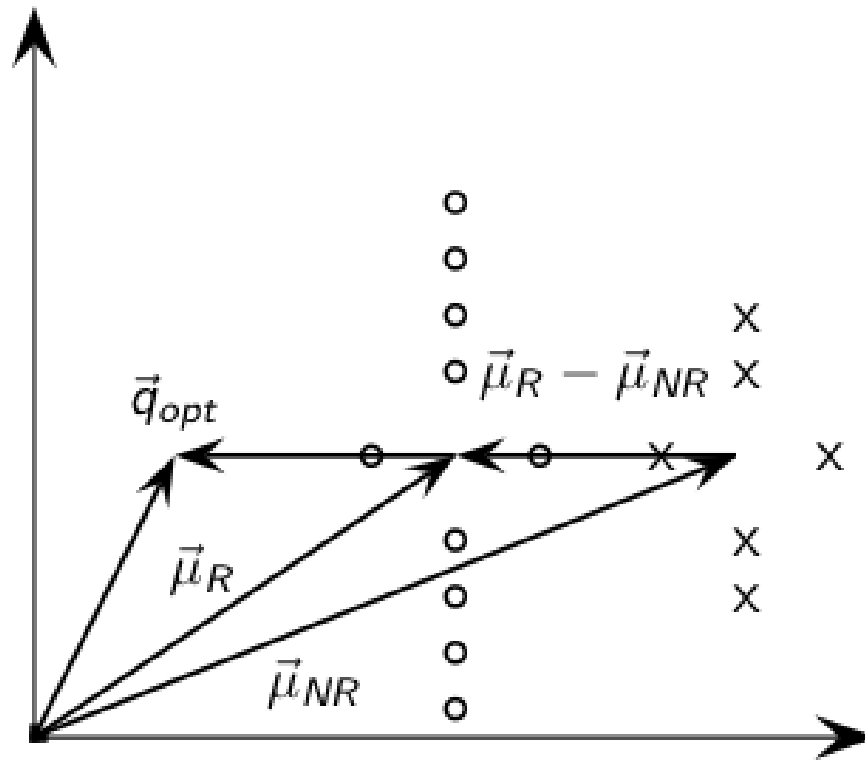
… to get $\vec{q}_{opt}$

Rocchio' illustrated
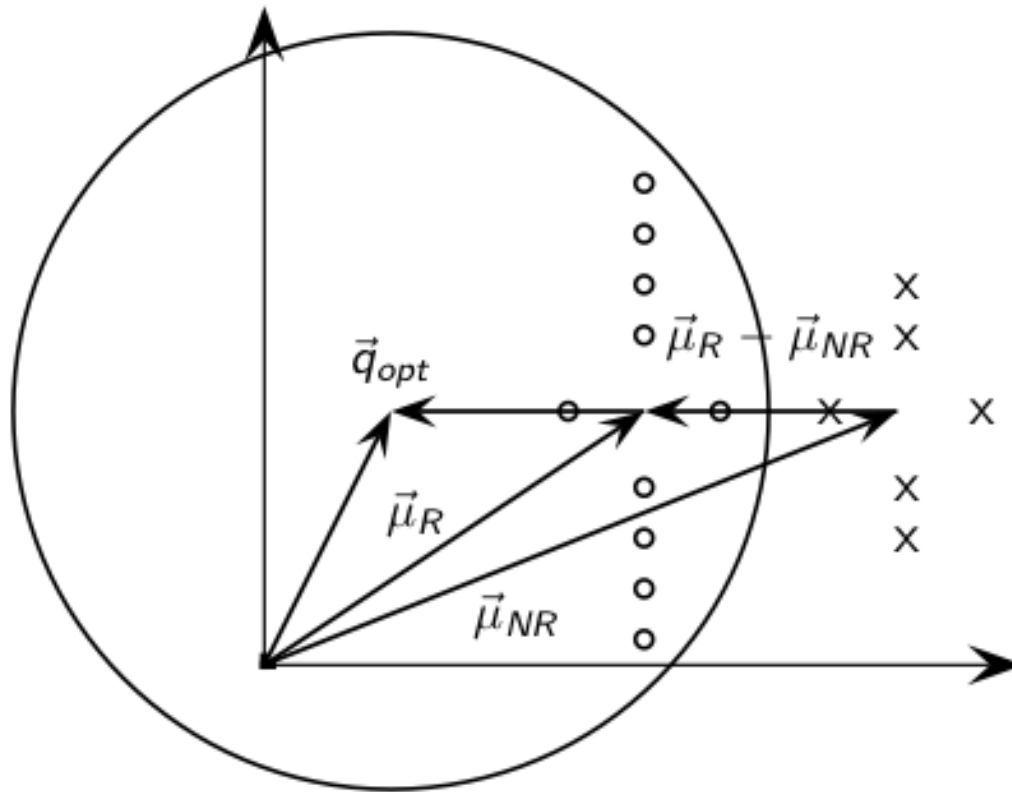
$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Rocchio' illustrated
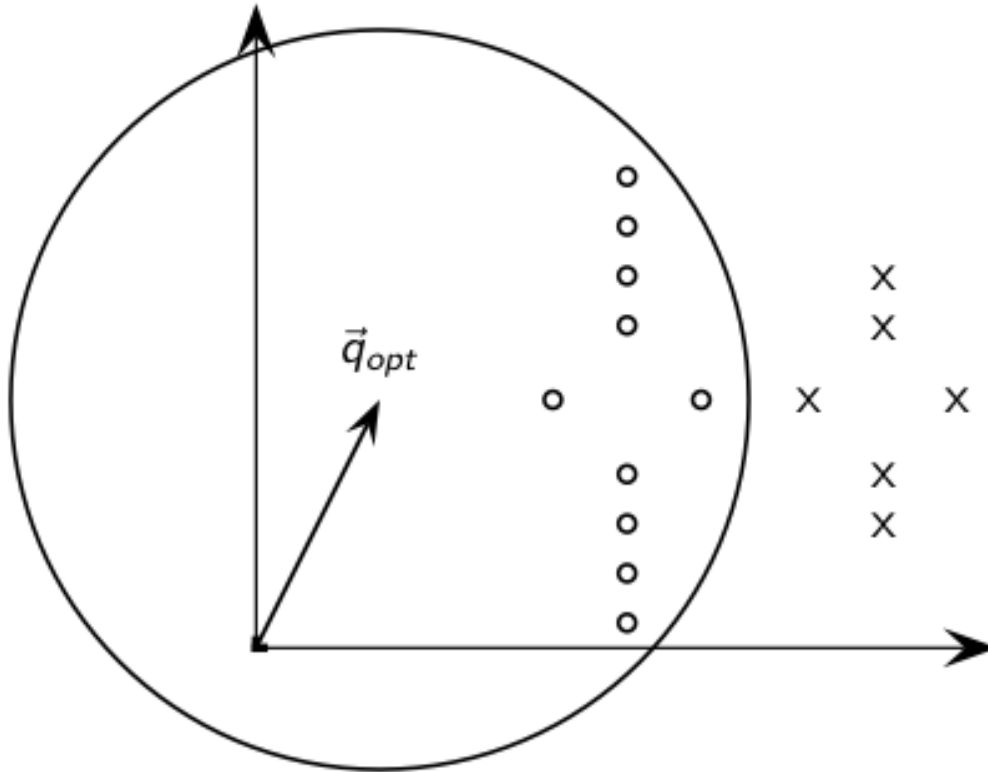
$\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Linear Separators

- Which of the linear separators is optimal?
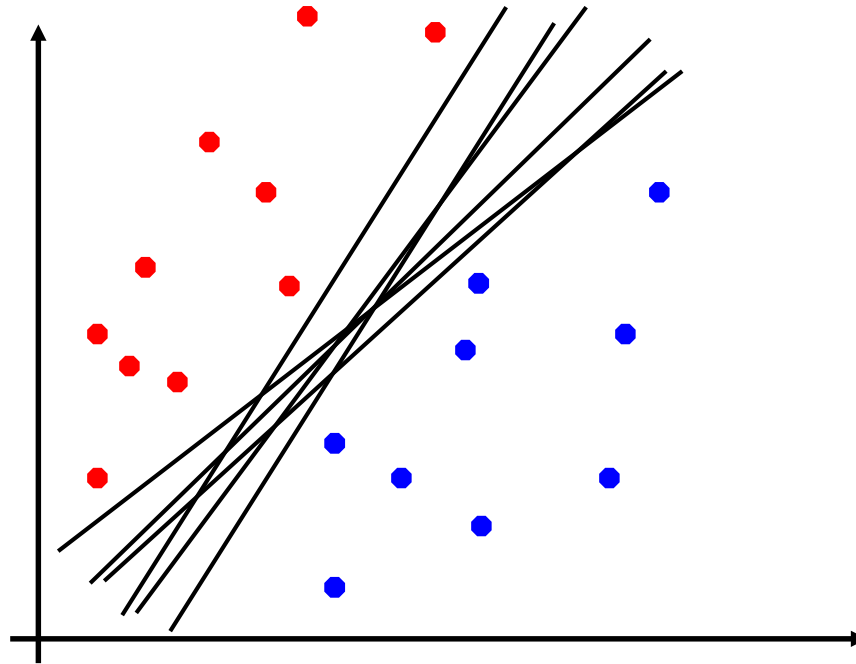
Maximum Margin Classification

- Implies that only "support vectors" matter; other training examples are ignorable.

Original from Ray Mooney
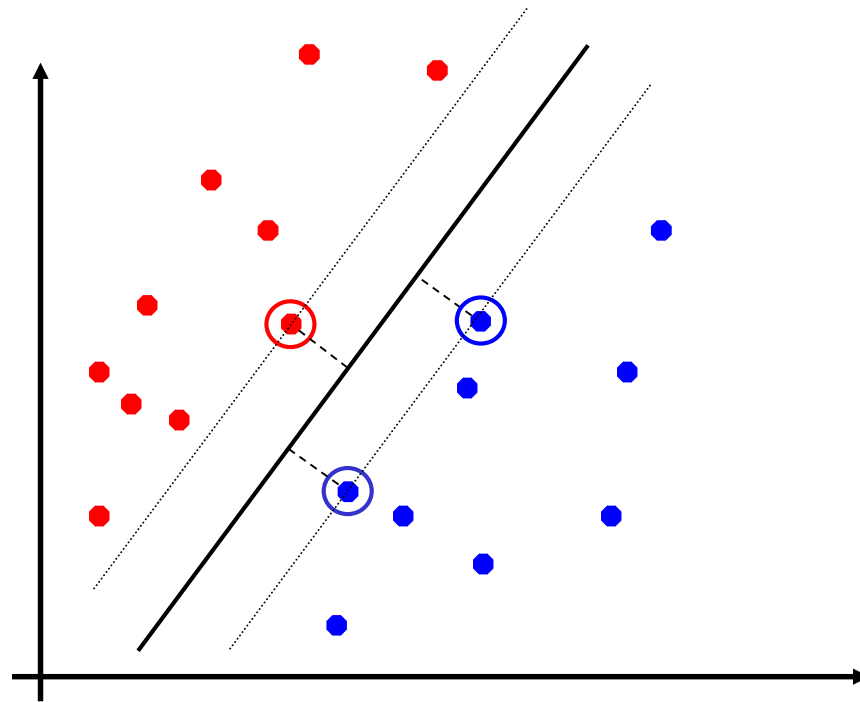
Soft-Margin
Support Vector Machine

Non-linear SVMs

$\Phi: \ \mathbf{x} \rightarrow \varphi(\mathbf{x})$

Original from Ray Mooney
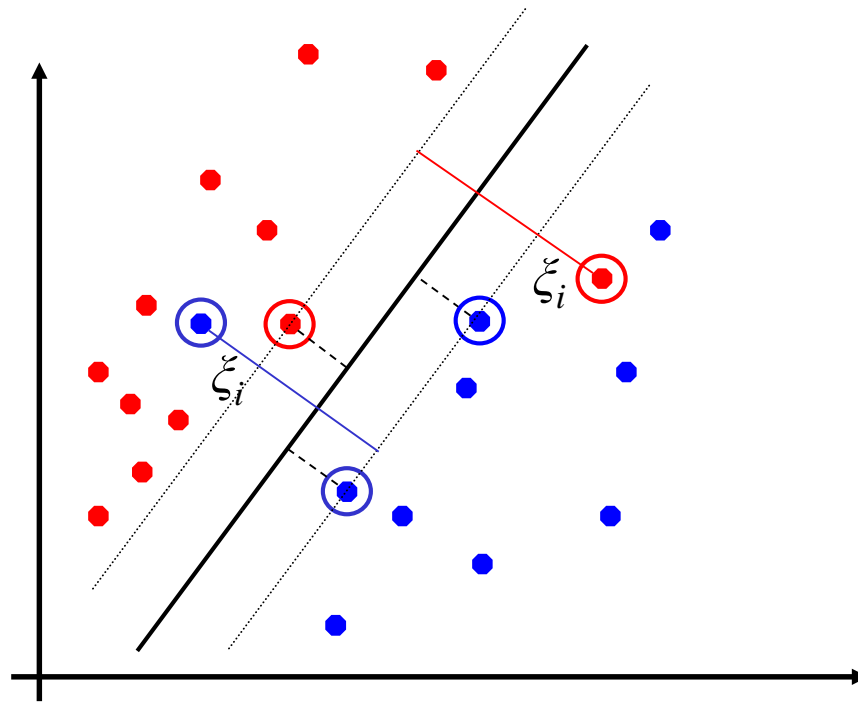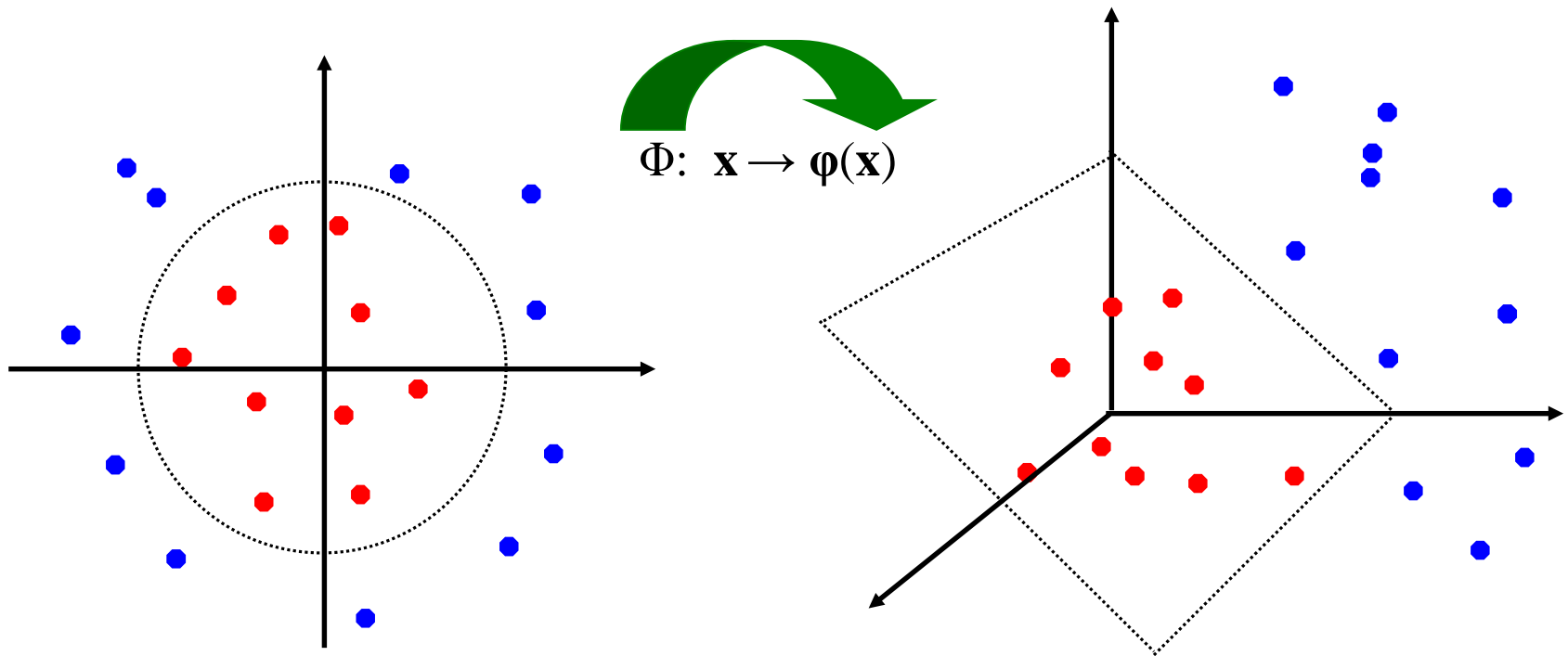
Gender Classification Example

```
>>> classifier.show_most_informative_features(5)
Most Informative Features
last_letter = 'a' female : male = 38.3 : 1.0
last_letter = 'k' male : female = 31.4 : 1.0
last_letter = 'f' male : female = 15.3 : 1.0
last_letter = 'p' male : female = 10.6 : 1.0
last_letter = 'w' male : female = 10.6 : 1.0
```

```
>>> for (tag, guess, name) in sorted(errors):
print 'correct=%-8s guess=%-8s name=%-30s'
correct=female guess=male name=Cindelyn ...
correct=female guess=male name=Katheryn
correct=female guess=male name=Kathryn ...
correct=male guess=female name=Aldrich ...
correct=male guess=female name=Mitch ...
correct=male guess=female name=Rich ...
```

NLTK Naïve Bayes

Sentiment Classification Example

```
>>> classifier.show_most_informative_features(5)
Most Informative Features
contains(outstanding) =     True pos : neg = 11.1 : 1.0
contains(seagal) =          True neg : pos =  7.7 : 1.0
contains(wonderfully) =     True pos : neg =  6.8 : 1.0
contains(damon) =           True pos : neg =  5.9 : 1.0
contains(wasted) =          True neg : pos =  5.8 : 1.0
```

Some Supervised Learning Methods

- ## Support Vector Machine

  - ### High accuracy

- ## k-Nearest-Neighbor

  - ### Naturally accommodates multi-class problems

- ## Decision Tree (a form of Rule Induction)

  - ### Explainable (at least near the top of the tree)

- ## Maximum Entropy

  - ### Accommodates correlated features

Supervised Learning Limitations

- Rare events

  – It can't learn what it has never seen!

- Overfitting

  – Too much memorization, not enough generalization

- Unrepresentative training data

  – Reported evaluations are often very optimistic

- It doesn't know what it doesn't know

  – So it always guesses some answer

- Unbalanced "class frequency"

  – Consider this when deciding what's good enough

Metadata <u>Extraction</u>: Named Entity "Tagging"

- Machine learning techniques can find:
  - Location
  - Extent
  - Type

- Two types of features are useful
  - Orthography
    - e.g., Paired or non-initial capitalization
  - Trigger words
    - e.g., Mr., Professor, said, …

File Edit View Go Favorites Help Address D:\RnRDemo.htm

Search | Clear

OR
AND

Topic | Person
Organization | Location
Speaker | Text
Story | Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terroris

5 stories about: Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terr
- Jewish-Arab relations : Politics and government : Palestinian Arabs : Middle East : Israel : Terrorism : Pale
- Jewish-Arab relations : Israel : Middle East : Middle East peace negotiations : Politics and government : P

male 5

Well as all work during president Clinton's trip to New York tonight and he enjoys the performance of the opera Carmen at Lincoln Center and see the scene there is a lot of Broadway. Now earlier today Mr. Clinton announced that the UN united Nations general assembly that he plans to send a nuclear test ban treaty to the Senate the treaty bans all nuclear test explosions and is regarded as a milestone in the arms control. Two israeli security guards were wounded in an early morning shooting in Jordan a government official says three men and a car opened fire on the guard's car wounding both before Skipping guards were treated at a hospital and released it is real several West Bank villages were sealed by israeli soldiers who search for the islamic militants behind two recent suicide bombings in Jerusalem palestinian leader Yasser Arafat says that he believes those was counsel for the bombing case and abroad.

Jewish-Arab relations

Middle East peace negotiations

Middle East

Palestinian self-rule areas

Israel

Politics and government

Arafat, Yasir

Palestinian Arabs

Features Engineering

- Topic
  - Counts for each word
- Sentiment
  - Counts for each word
- Human values
  - Counts for each word

- Sentence splitting
  - Ends in one of .!?
  - Next word capitalized
- Part of speech tagging
  - Word ends in –ed, -ing, …
  - Previous word is a, to, …
- Named entity recognition
  - All+only first letters caps
  - Next word is said, went, …

Normalization

- Variant forms of names ("name authority")
  - Pseudonyms, partial names, citation styles

- Acronyms and abbreviations

- Co-reference resolution
  - References to roles, objects, names
  - Anaphoric pronouns

- Entity Linking

Entity Linking

Example: Bibliographic References

15. Faloutsos, C., Oard, D. (1995). "A Survey of Information Retrieval and Filtering Methods," avail. As UMIACS-TR-95-33, College Park: U of MD.

[11] Faloutsos, C. and Oard, D. W., "A Survey of Information Retrieval and Filtering Methods", University of Maryland, Technical Report CS-TR-3514, August 1995.

[47] Christos Faloutsos and Douglas W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, University of Maryland, August 1995. http://www.enee.umd.edu/medlab/filter/papers/survey.ps.

[Faloutsos] Christos Faloutsos and Dougtas Oard, A Survey of Information Retrieval and Filtering Method, <URL:http://www.glue.umd.edu/enee/medlab/filter/papers/survey.ps>

When Lisa's mother Marge Simpson went to a weekend getaway at Rancho Relaxo, …

After two years in the academic quagmire of Springfield Elementary, Lisa finally has a teacher that she connects with. But she soon learns that the problem with being middle-class is that

Springfield

Springfield Elementary

Marge Simpson

Homer Simpson

Lisa Simpson

Bart Simpson
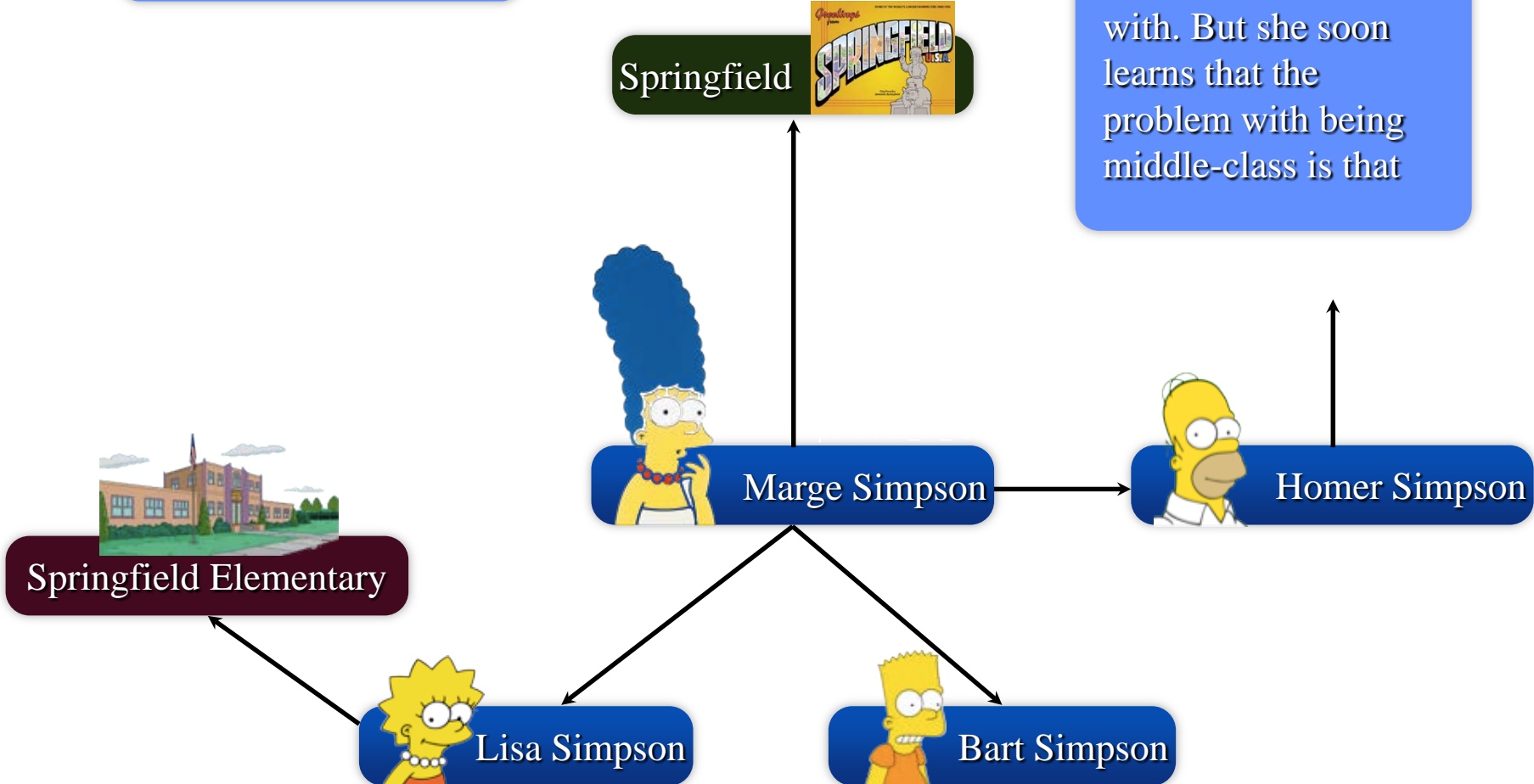
Knowledge-Base Population
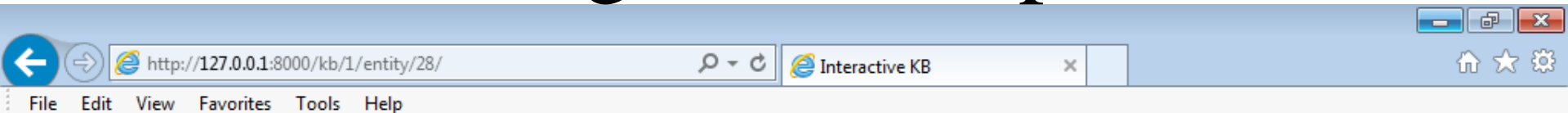
- **Entity ID:** e_XIN_ENG_20100312_0139_27
- **Type:** PER
- **Canonical mention strings:**
  - Vijitha Herath (x1)
- **Mention strings:**
  - Herath (x2)
  - Vijitha Herath (x1)
- **Mentioned in Documents:**
  - Roundup: Sri Lanka's war hero to face trial before key polls (x3)

	Subject	Predicate	Object	Sentence
Annotate	Vijitha Herath	per:employee_or_member_of	Sri Lanka	Herath also said as he was the senior most military officer of the country and the junior officer could not be in the bench.

Roundup: Sri Lanka's war hero to face trial before key polls (x3)

by Amanda Sri

COLOMBO, March 12 (Xinhua) -- The glorious days where he plied on Colombo roads escorted by hundreds of elite Commandos were over.

KB summary:

- **Run ID**: Protest over arrest of Sri Lanka reporter linked to Fonseka [AFP_ENG_20100318.0623]
- Annotations
- Entities 34
    - 14 people
    - 17 orgs
    - 3 GPEs
- Facts 32
    - gpe:employees_or_members 3
    - gpe:residents_of_country 1
    - org:alternate_names 3
    - org:employees_or_members 4
    - org:top_members_employees 1
    - per:age 1
    - per:charges 1
    - per:countries_of_residence 1
    - per:employee_or_member_of 7
    - per:siblings 2
    - per:title 7
    - per:top_member_employee_of 1
- Source Documents 3
    - Protest over arrest of Sri Lanka reporter linked to Fonseka
    - 'No democracy' in S.Lanka, says top Fonseka supporter
    - Roundup: Sri Lanka's war hero to face trial before key polls

CLiMB: Metadata from Description

Web Ontology Language (OWL)

```xml
<owl:Class rdf:about="http://dbpedia.org/ontology/Astronaut">
        <rdfs:label xml:lang="en">astronaut</rdfs:label>
        <rdfs:label xml:lang="de">Astronaut</rdfs:label>
        <rdfs:label xml:lang="fr">astronaute</rdfs:label>
        <rdfs:subClassOf
                rdf:resource="http://dbpedia.org/ontology/Person">
        </rdfs:subClassOf>
</owl:Class>
```

Linked Open Data
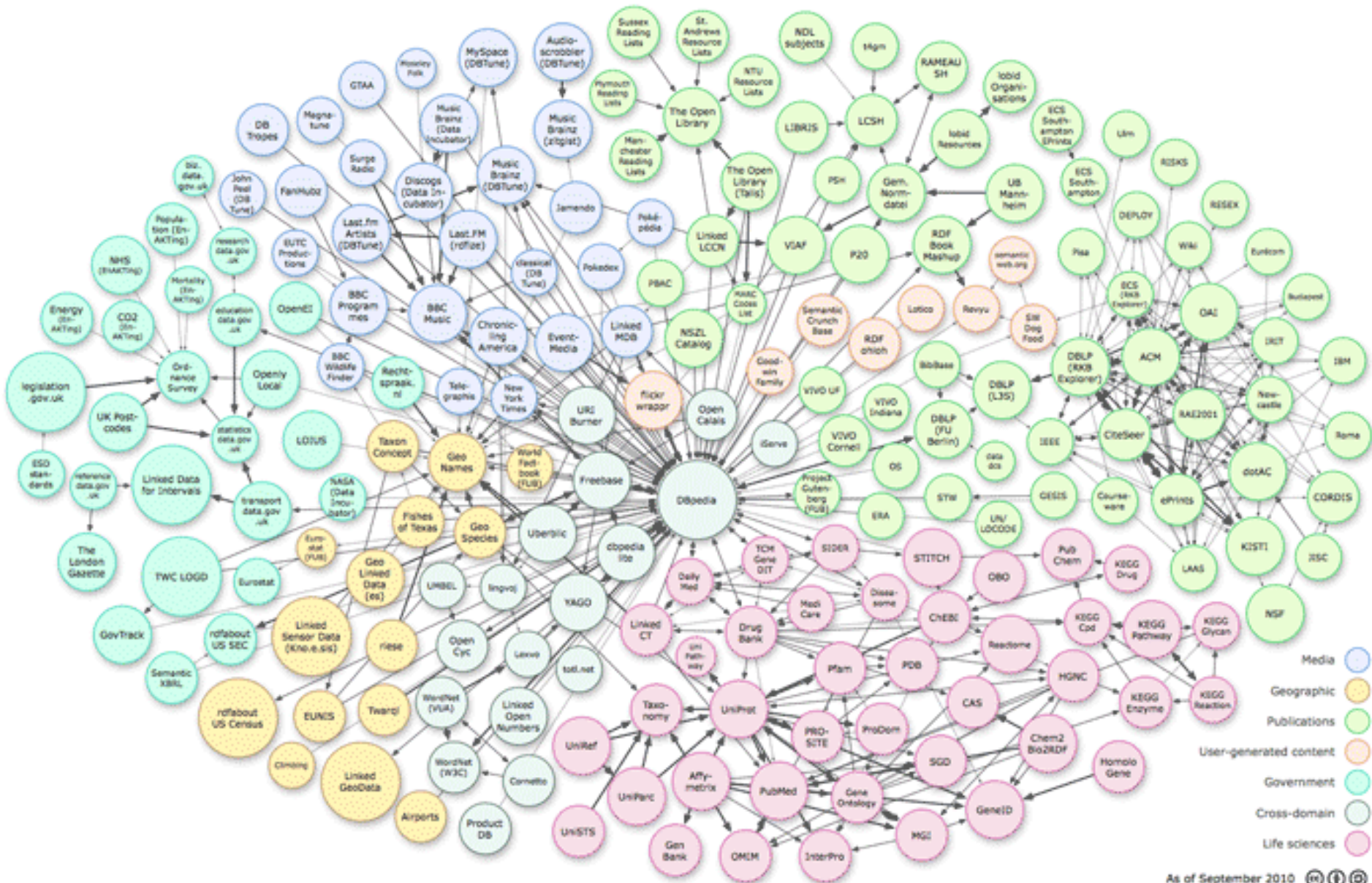
As of September 2010

Semantic Web Search

Before You Go!

- On a sheet of paper (no names), answer the following question:

  What was the muddiest point in today's class?