College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Data Mining

Session 17

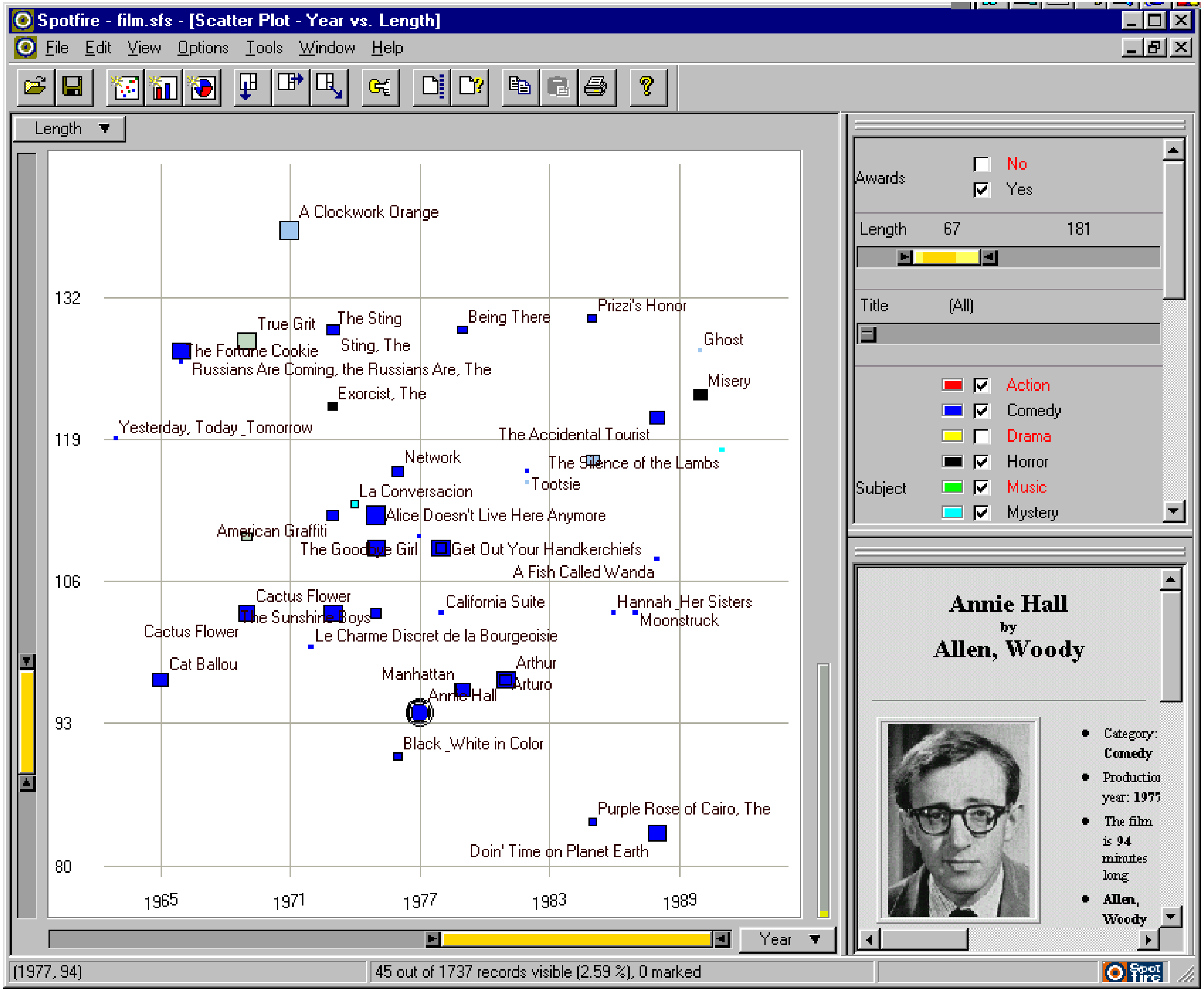
INST 301

Introduction to Information Science

Agenda

- Visualization
- Data mining
- Supervised machine learning

Starfield Visualization



Spotfire - film.sfs - [Scatter Plot - Year vs. Length]

Awards ☐ No ☑ Yes

Length 67 181

Title (All)

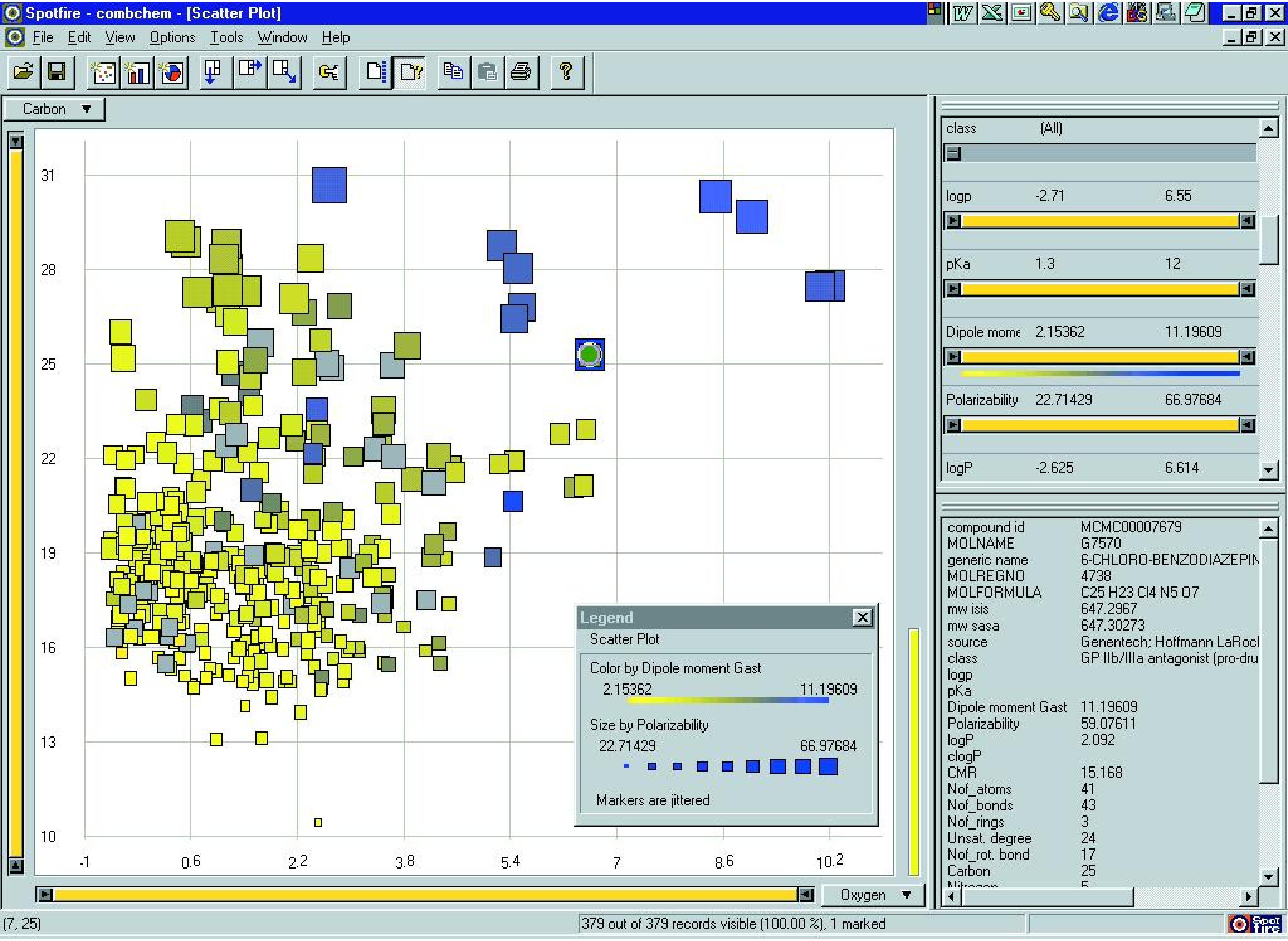
Subject

🟥	☑	Action
🟦	☑	Comedy
🟨	☐	Drama
⬛	☑	Horror
🟩	☑	Music
▨	☑	Mystery

Annie Hall
by
Allen, Woody

- Category: **Comedy**
- Production year: 1977
- The film is 94 minutes long
- **Allen, Woody**

(1977, 94) 45 out of 1737 records visible (2.59 %), 0 marked



Color by Dipole moment Gast
2.15362 11.19609

Size by Polarizability
22.71429 66.97684

class	(All)	
logp	-2.71	6.55
pKa	1.3	12
Dipole mome	2.15362	11.19609
Polarizability	22.71429	66.97684
logP	-2.625	6.614

compound id	MCMC00007679
MOLNAME	G7570
generic name	6-CHLORO-BENZODIAZEPIN
MOLREGNO	4738
MOLFORMULA	C25 H23 Cl4 N5 O7
mw isis	647.2967
mw sasa	647.30273
source	Genentech; Hoffmann LaRocl
class	GP IIb/IIIa antagonist (pro-dru
logp	
pKa	
Dipole moment Gast	11.19609
Polarizability	59.07611
logP	2.092
clogP	
CMR	15.168
Nof_atoms	41
Nof_bonds	43
Nof_rings	3
Unsat. degree	24
Nof_rot. bond	17
Carbon	25
Nitrogen	5

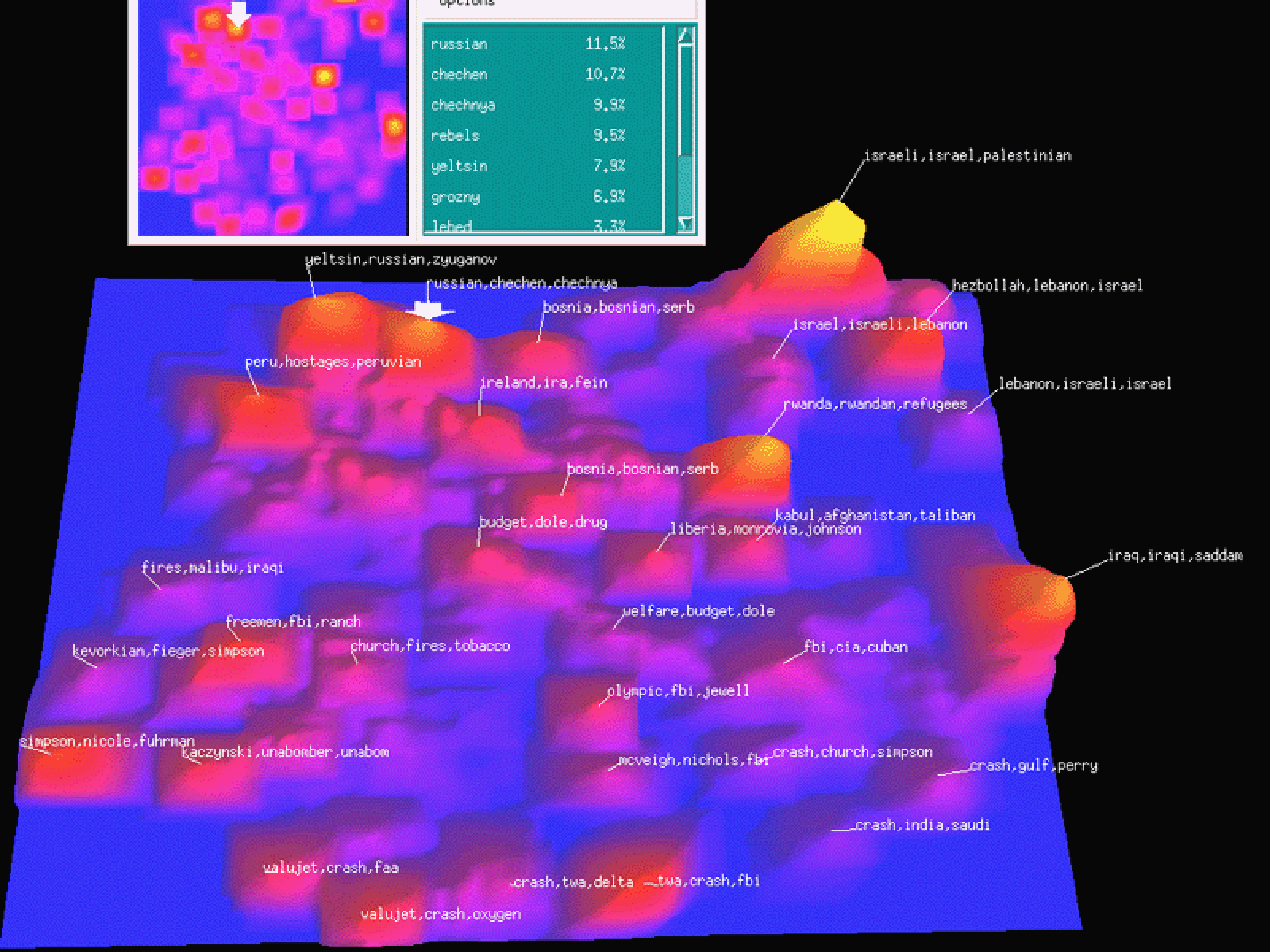
(7, 25)   379 out of 379 records visible (100.00 %), 1 marked

Constructing Starfield Displays

- Two attributes determine the position
  - Can be dynamically selected from a list
- Numeric position attributes work best
  - Date, length, rating, …
- Other attributes can affect the display
  - Displayed as color, size, shape, orientation, …
    - Each point can represent a cluster
  - Interactively specified using "dynamic queries"

Projection

- Depict many numeric attributes in 2 dimensions
  - While preserving important spatial relationships
- Typically based on the vector space model
  - Which has about 100,000 numeric attributes!
- Approximates multidimensional scaling
  - Heuristics approaches are reasonably fast
- Often visualized as a starfield
  - But the dimensions lack any particular meaning



russian	11.5%
chechen	10.7%
chechnya	9.9%
rebels	9.5%
yeltsin	7.9%
grozny	6.9%
lebed	3.3%

Contour Map Display

- Display a cluster density as terrain elevation
  - Fit a smooth opaque surface to the data
- Visualize in three dimensions
  - Project two 2-D and allow manipulation
  - Use stereo glasses to create a virtual "fishtank"
  - Create an immersive virtual reality experience
    - Mead mounted stereo monitors and head tracking
    - "Cave" with wall projection and body tracking

Cluster Formation

- Based on inter-document similarity
  - Computed using the cosine measure, for example
- Heuristic methods can be fairly efficient
  - Pick any document as the first cluster “seed”
  - Add the most similar document to each cluster
    - Adding the same document will join two clusters
  - Check to see if each cluster should be split
    - Does it contain two or more fairly coherent groups?
- Lots of variations on this have been tried

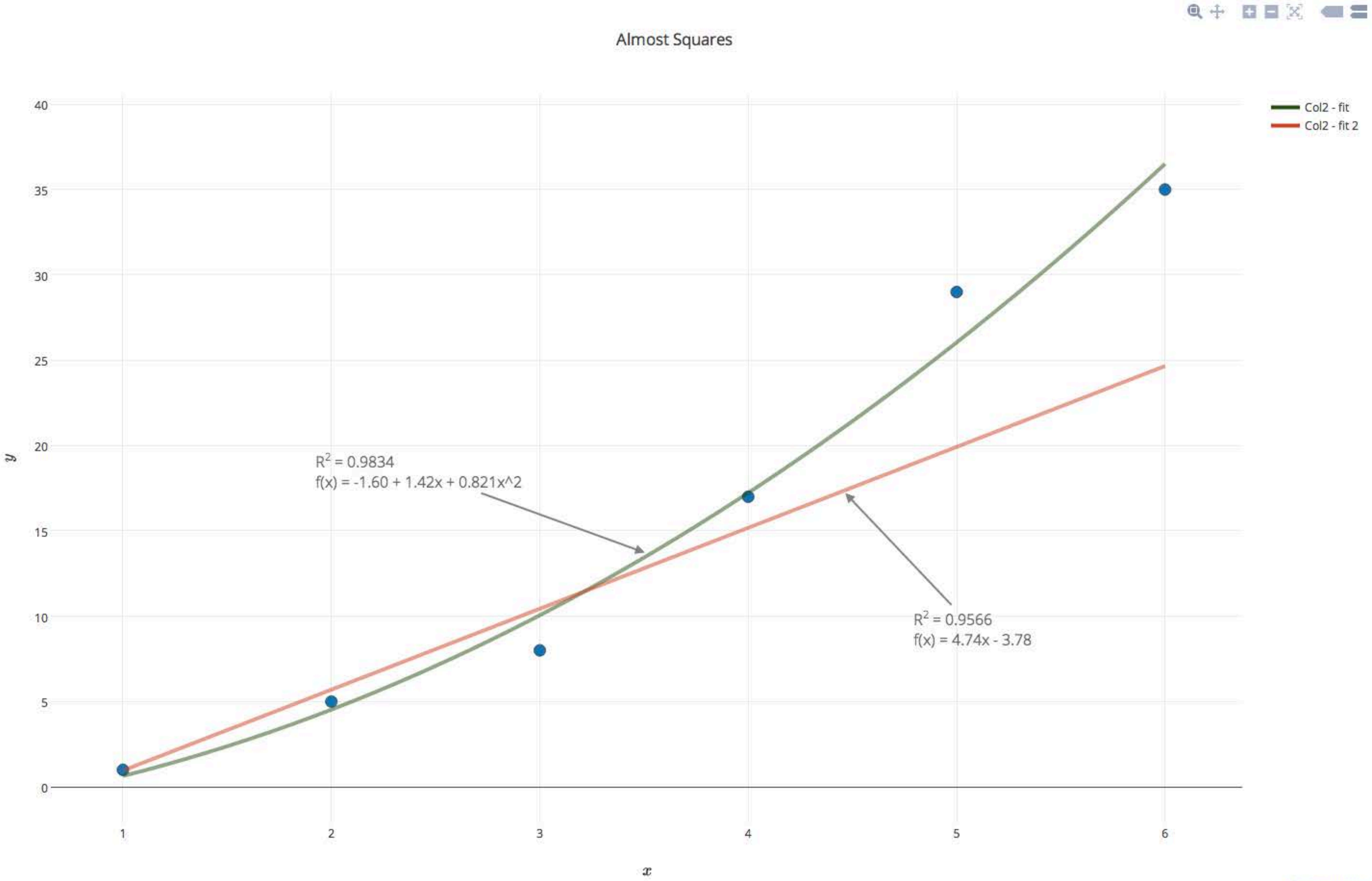
Agenda

- Visualization

➢ Data mining

- Supervised machine learning

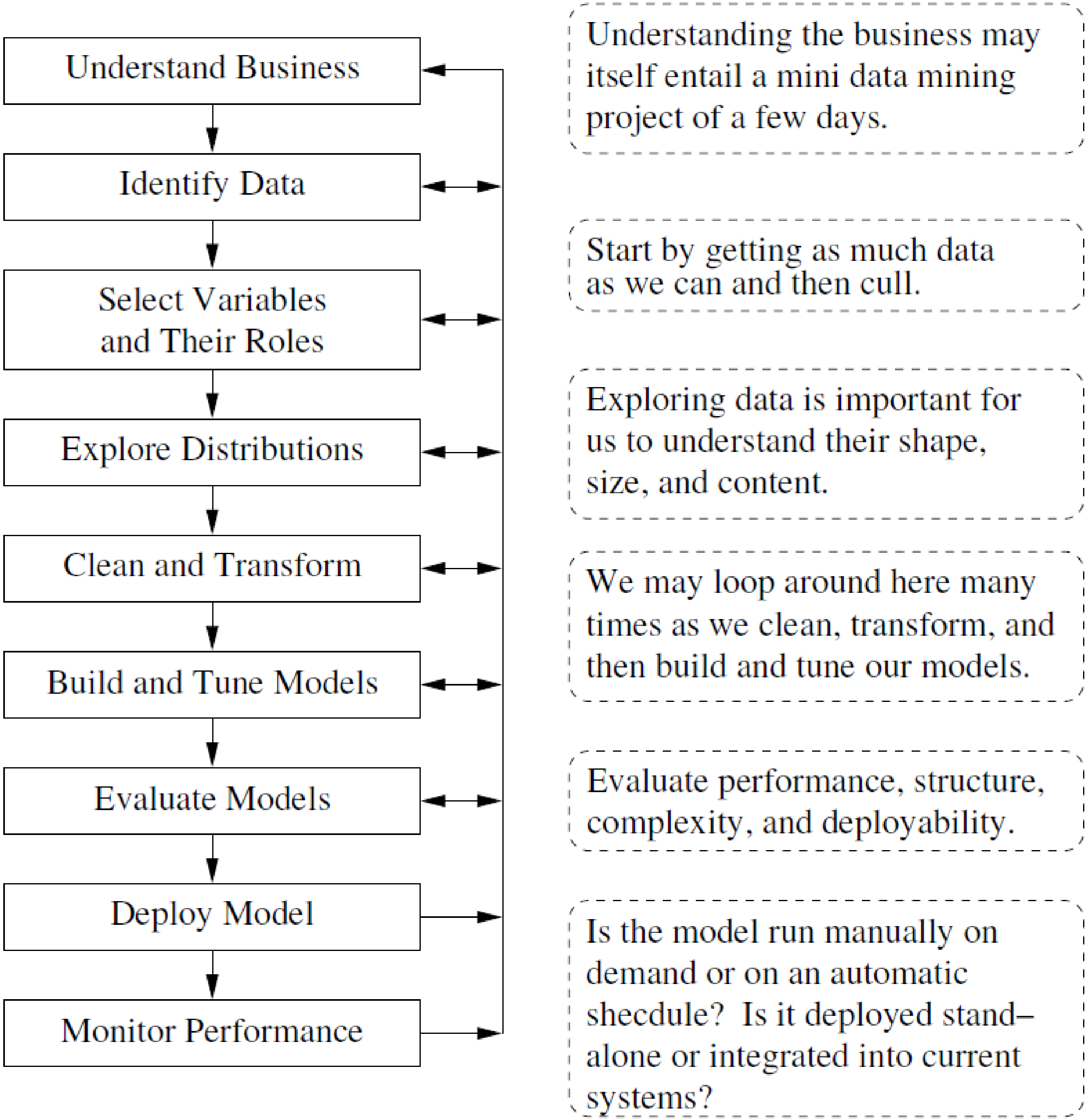
Curve Fitting



$R^2 = 0.9834$
f(x) = -1.60 + 1.42x + 0.821x^2

$R^2 = 0.9566$
f(x) = 4.74x - 3.78

Source: almost squares



Understanding the business may itself entail a mini data mining project of a few days.

Start by getting as much data as we can and then cull.

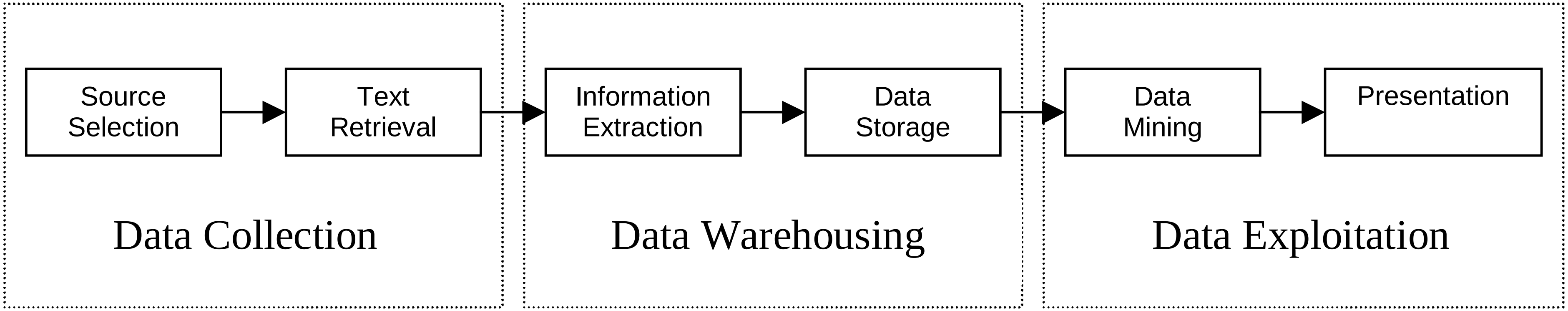
Exploring data is important for us to understand their shape, size, and content.

We may loop around here many times as we clean, transform, and then build and tune our models.

Evaluate performance, structure, complexity, and deployability.

Is the model run manually on demand or on an automatic shecdule? Is it deployed stand–alone or integrated into current systems?

Text Data Mining

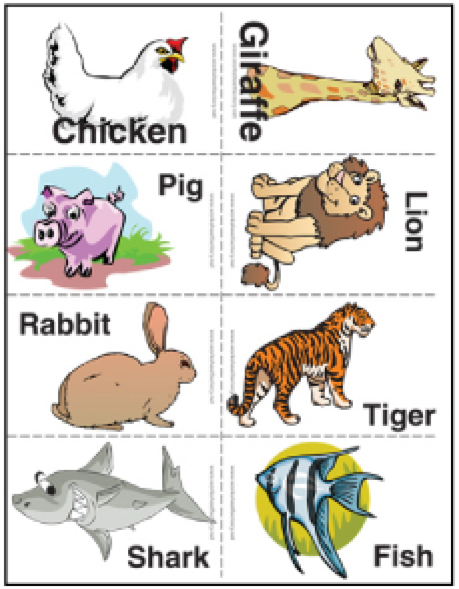


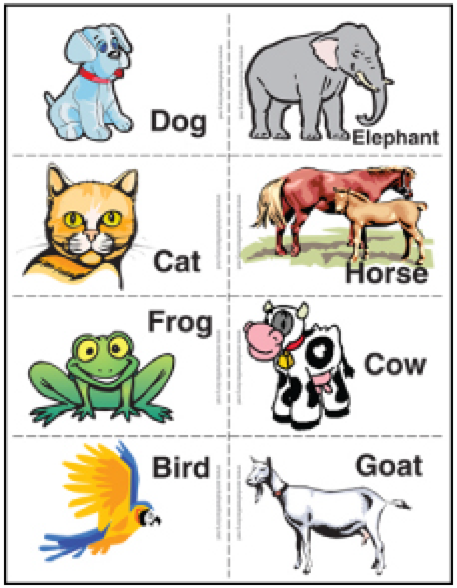
Agenda

- Visualization

- Data mining

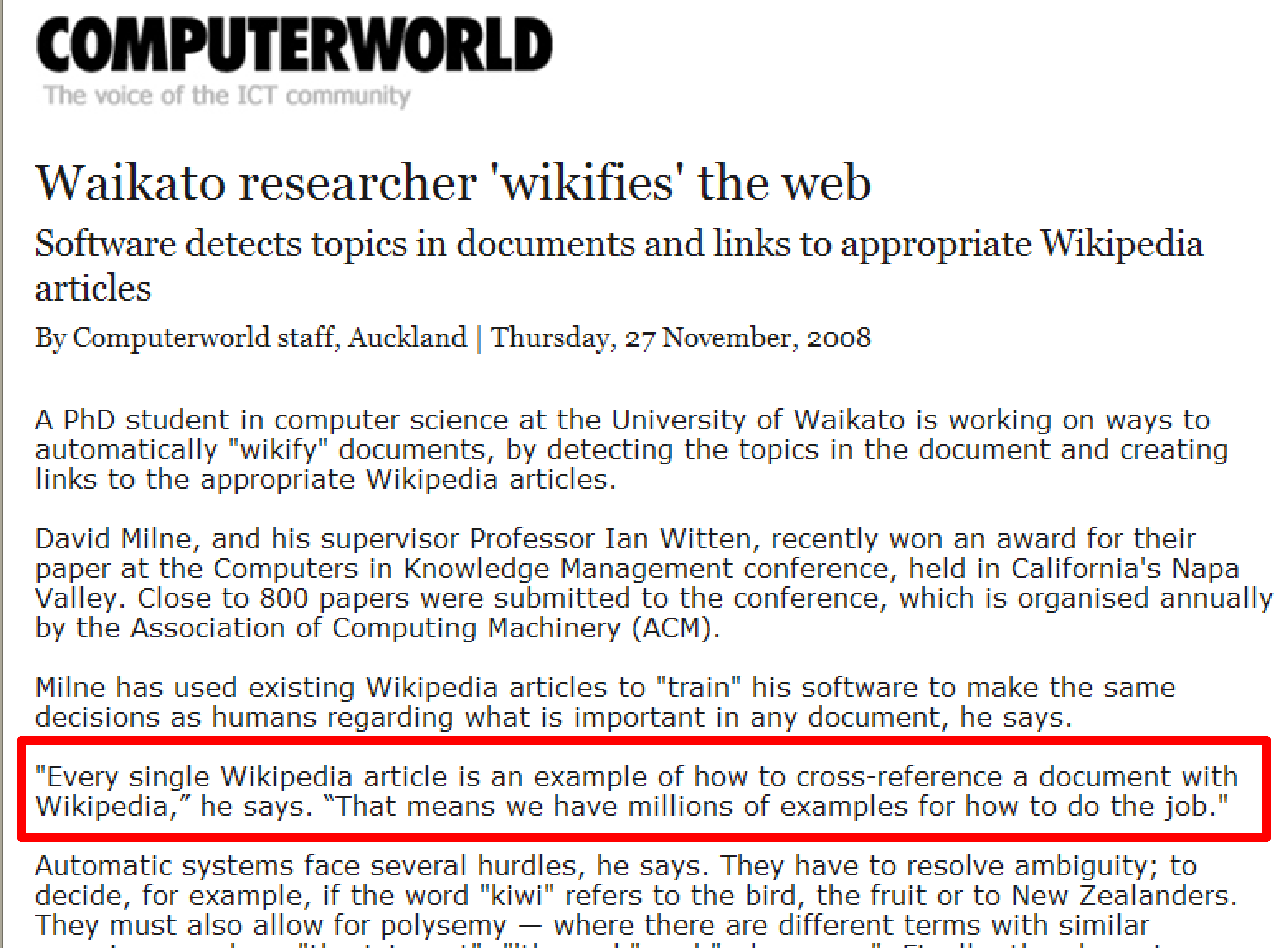
➢ Supervised machine learning


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
- …



COMPUTERWORLD
The voice of the ICT community

Waikato researcher 'wikifies' the web

Software detects topics in documents and links to appropriate Wikipedia articles

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

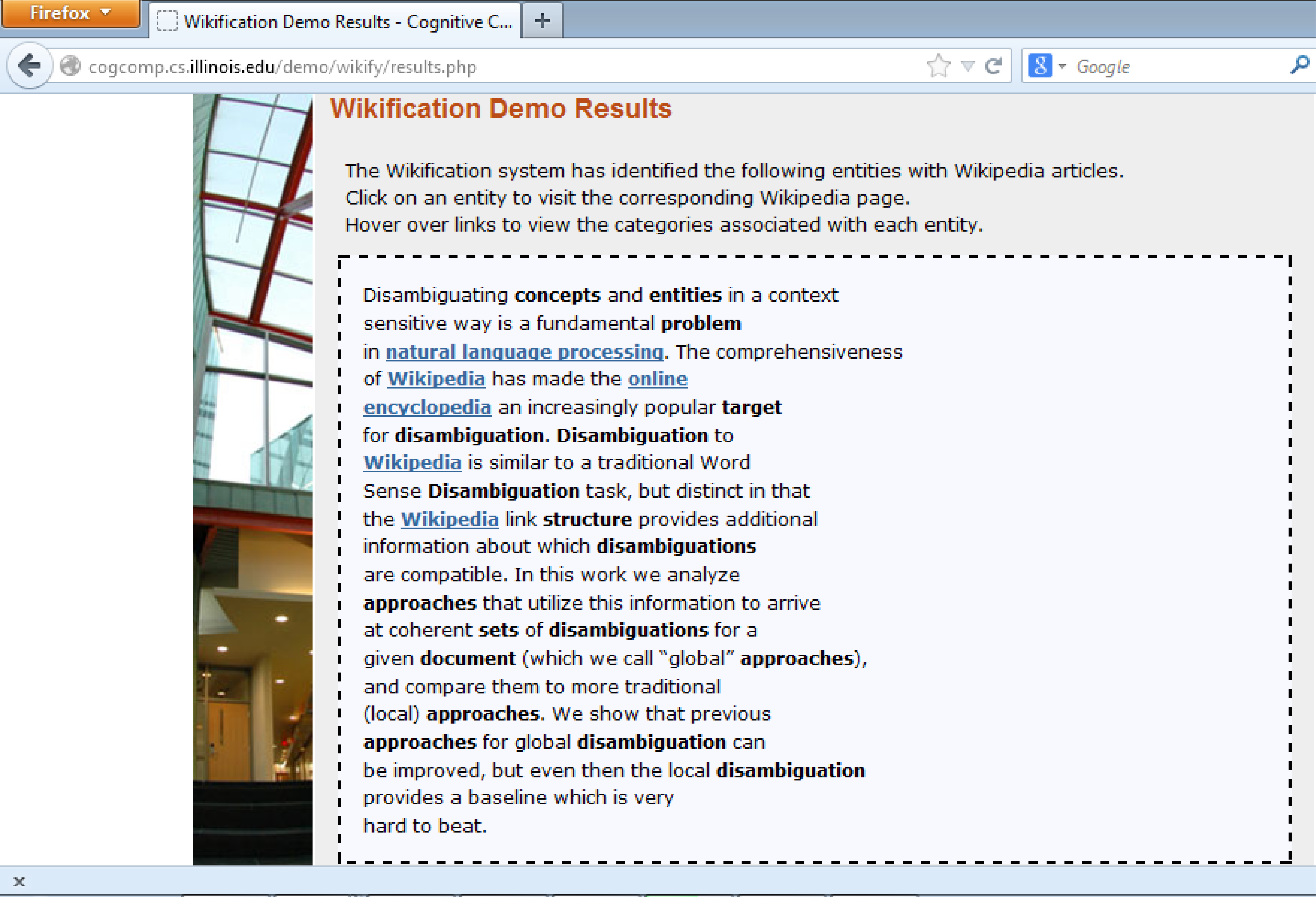
A PhD student in computer science at the University of Waikato is working on ways to automatically "wikify" documents, by detecting the topics in the document and creating links to the appropriate Wikipedia articles.

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

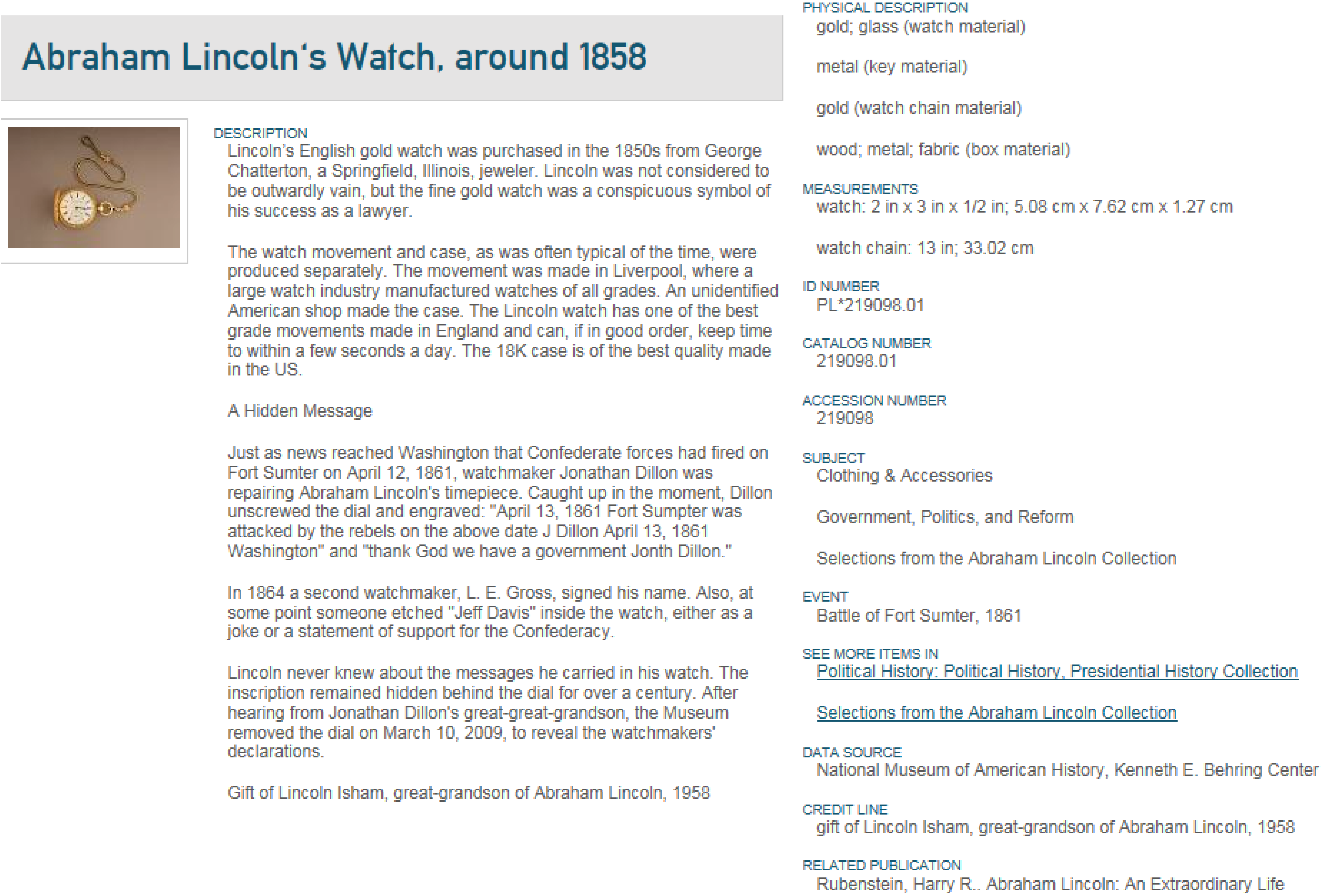

Firefox
Wikification Demo Results - Cognitive C...
cogcomp.cs.illinois.edu/demo/wikify/results.php
Google

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

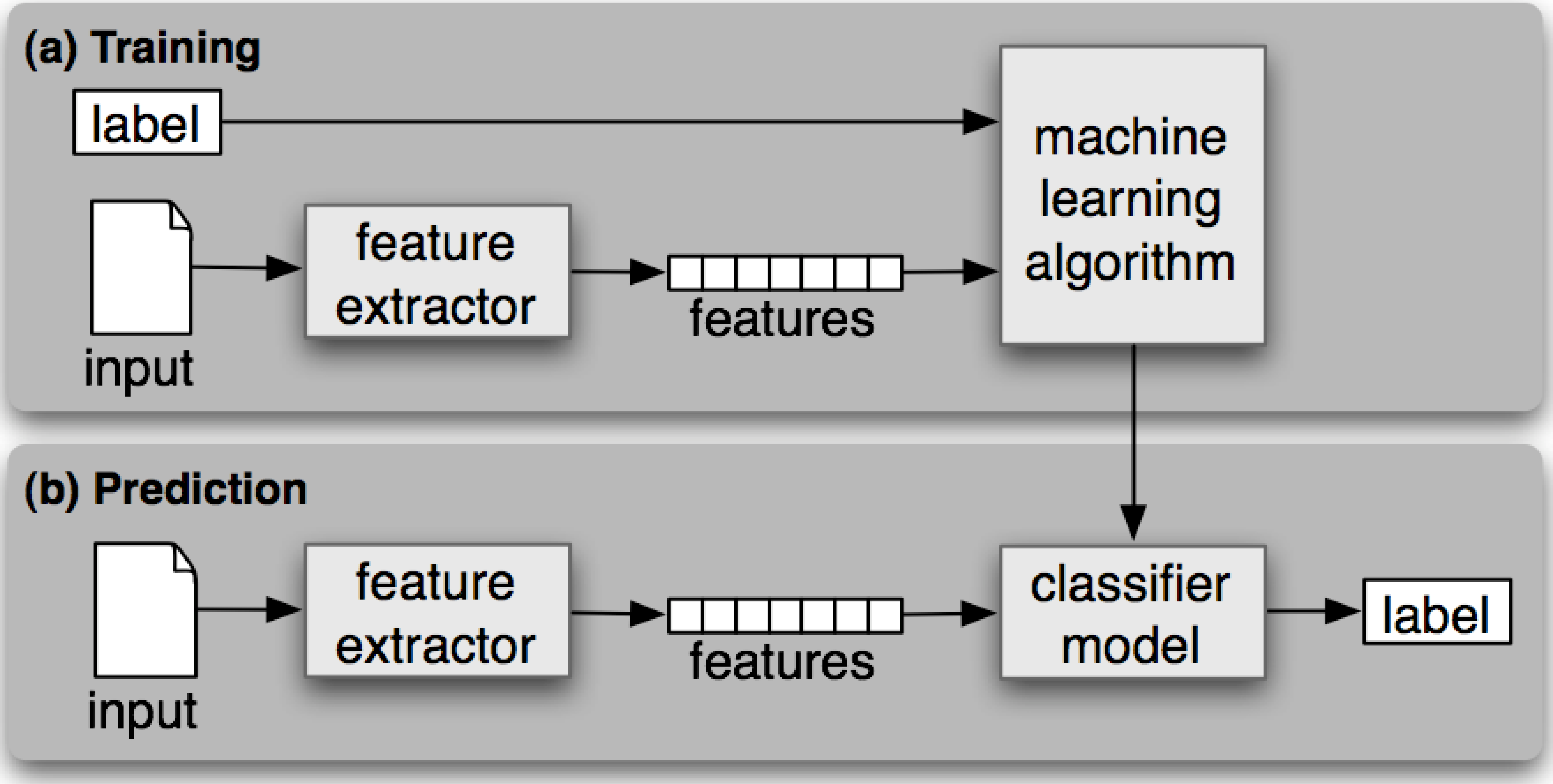
INTERVIEWED BY DR. STEPHEN E. AMBROSE AND DR. DOUGLAS BRINKLEY
HOUSTON, TEXAS – 19 SEPTEMBER 2001

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. http://wikipedia-miner.cms.waikato.ac.nz/demos/annotate/

Supervised Machine Learning



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

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 ...
```

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

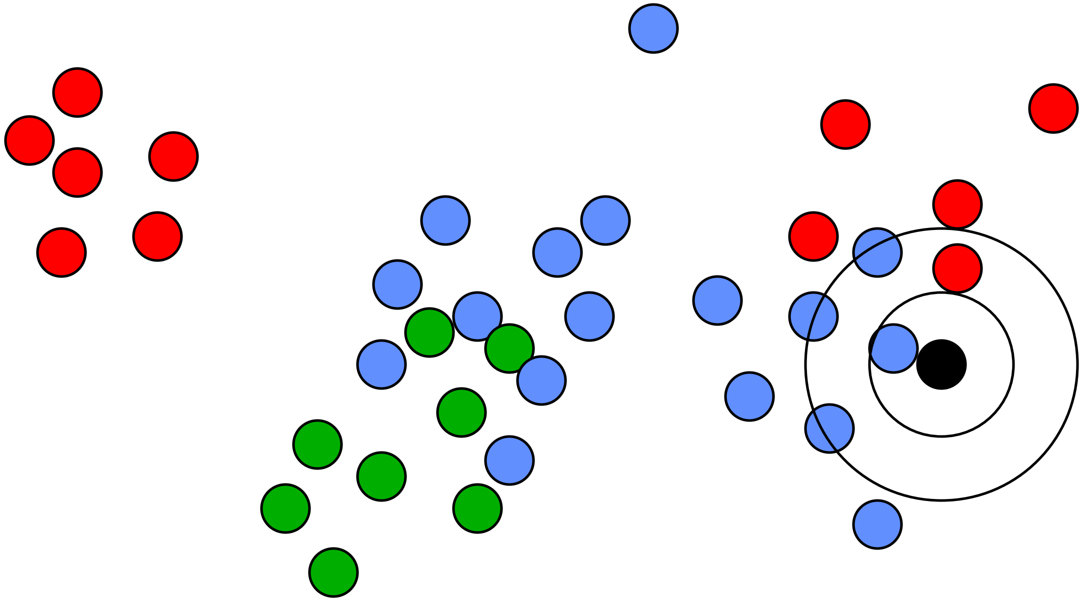
Rule Induction

- Automatically derived Boolean profiles
  - (Hopefully) effective and easily explained
- Specificity from the "perfect query"
  - AND terms in a document, OR the documents
- Generality from a bias favoring short profiles
  - e.g., penalize rules with more Boolean operators
  - Balanced by rewards for precision, recall, …

Statistical Classification

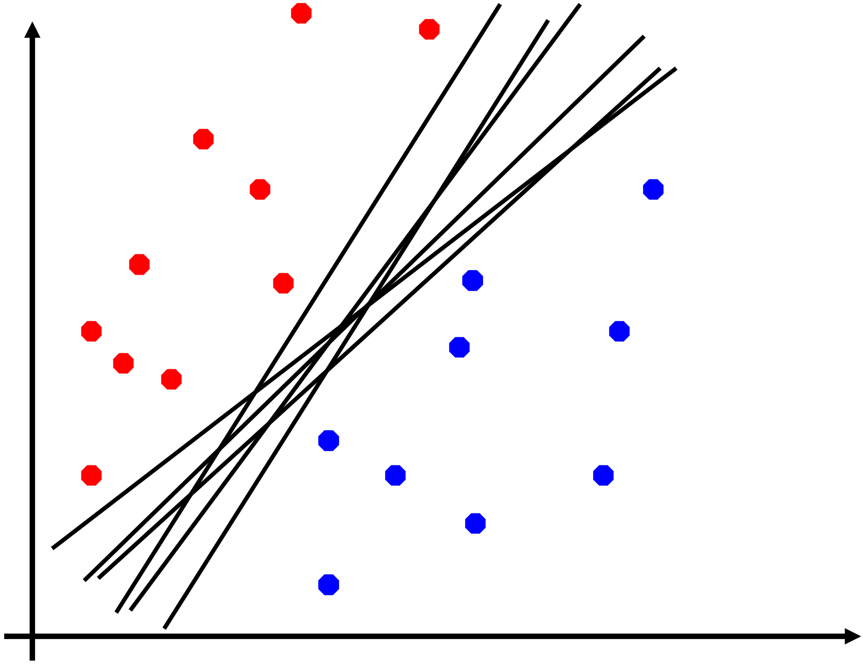
- Represent documents as vectors
  - e.g., based on TF, IDF, Length
- Build a statistical model for each label
  - e.g., a “vector space”
- Use that model to label new instances
  - e.g., by largest inner product

The *k*-Nearest-Neighbor Classifier



Linear Separators

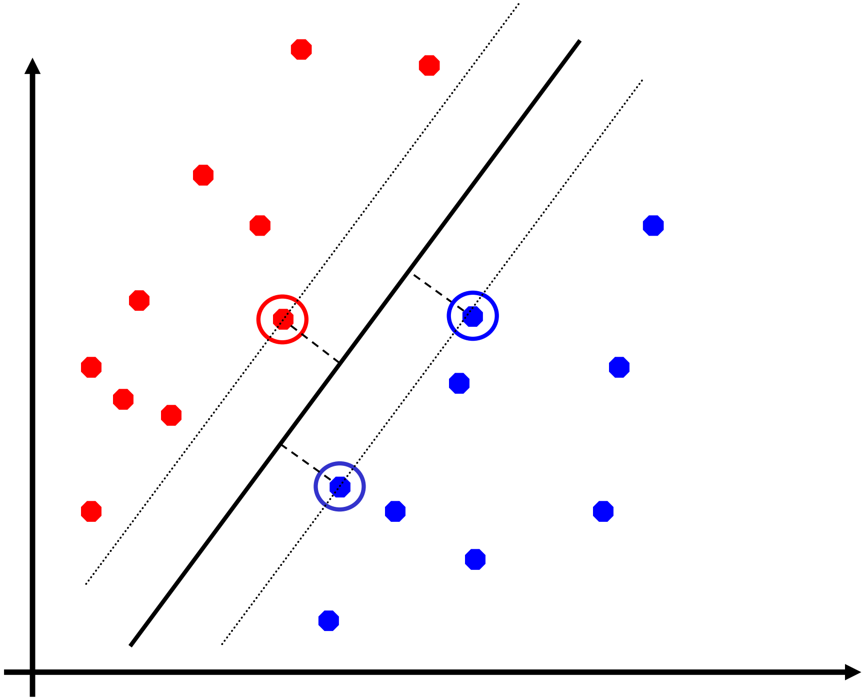
- Which of the linear separators is optimal?



Original from Ray Mooney

Maximum Margin Classification

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



Original from Ray Mooney

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