COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# An Introduction to Machine Learning and Natural Language Processing Tools

MIAS
MULTIMODAL
INFORMATION ACCESS
AND SYNTHESIS

Presented by:
Nick Rizzolo, Mark Sammons, Vivek Srikumar, and James Clarke

Advisor: Dan Roth

08/23/10

# Motivation: the News

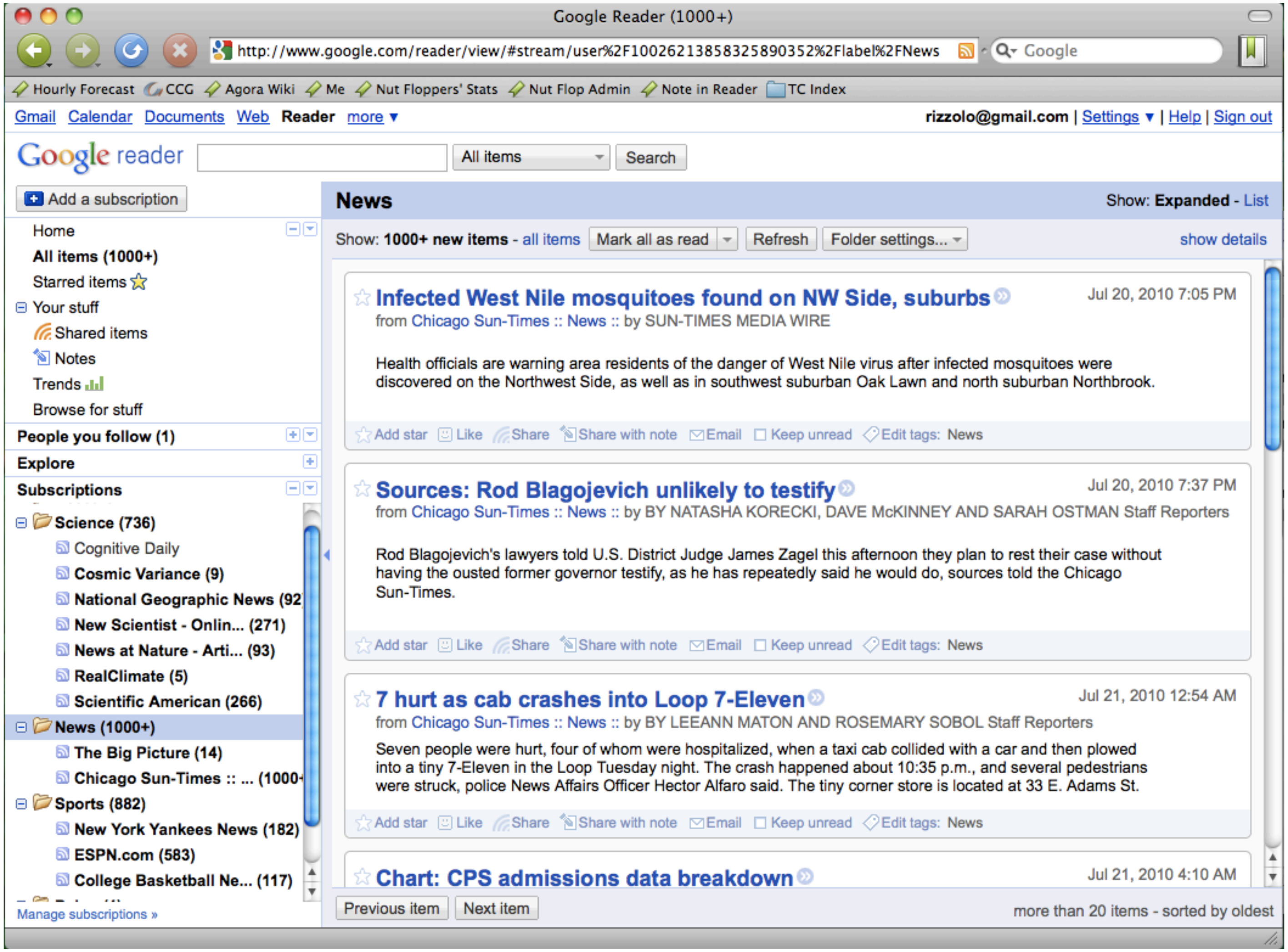
# Motivation: the News

# Motivation: the News

# Motivation: the News

# Motivation: the News

# News readers help with organization



**Infected West Nile mosquitoes found on NW Side, suburbs** — Jul 20, 2010 7:05 PM

from Chicago Sun-Times :: News :: by SUN-TIMES MEDIA WIRE

Health officials are warning area residents of the danger of West Nile virus after infected mosquitoes were discovered on the Northwest Side, as well as in southwest suburban Oak Lawn and north suburban Northbrook.

**Sources: Rod Blagojevich unlikely to testify** — Jul 20, 2010 7:37 PM

from Chicago Sun-Times :: News :: by BY NATASHA KORECKI, DAVE McKINNEY AND SARAH OSTMAN Staff Reporters

Rod Blagojevich's lawyers told U.S. District Judge James Zagel this afternoon they plan to rest their case without having the ousted former governor testify, as he has repeatedly said he would do, sources told the Chicago Sun-Times.

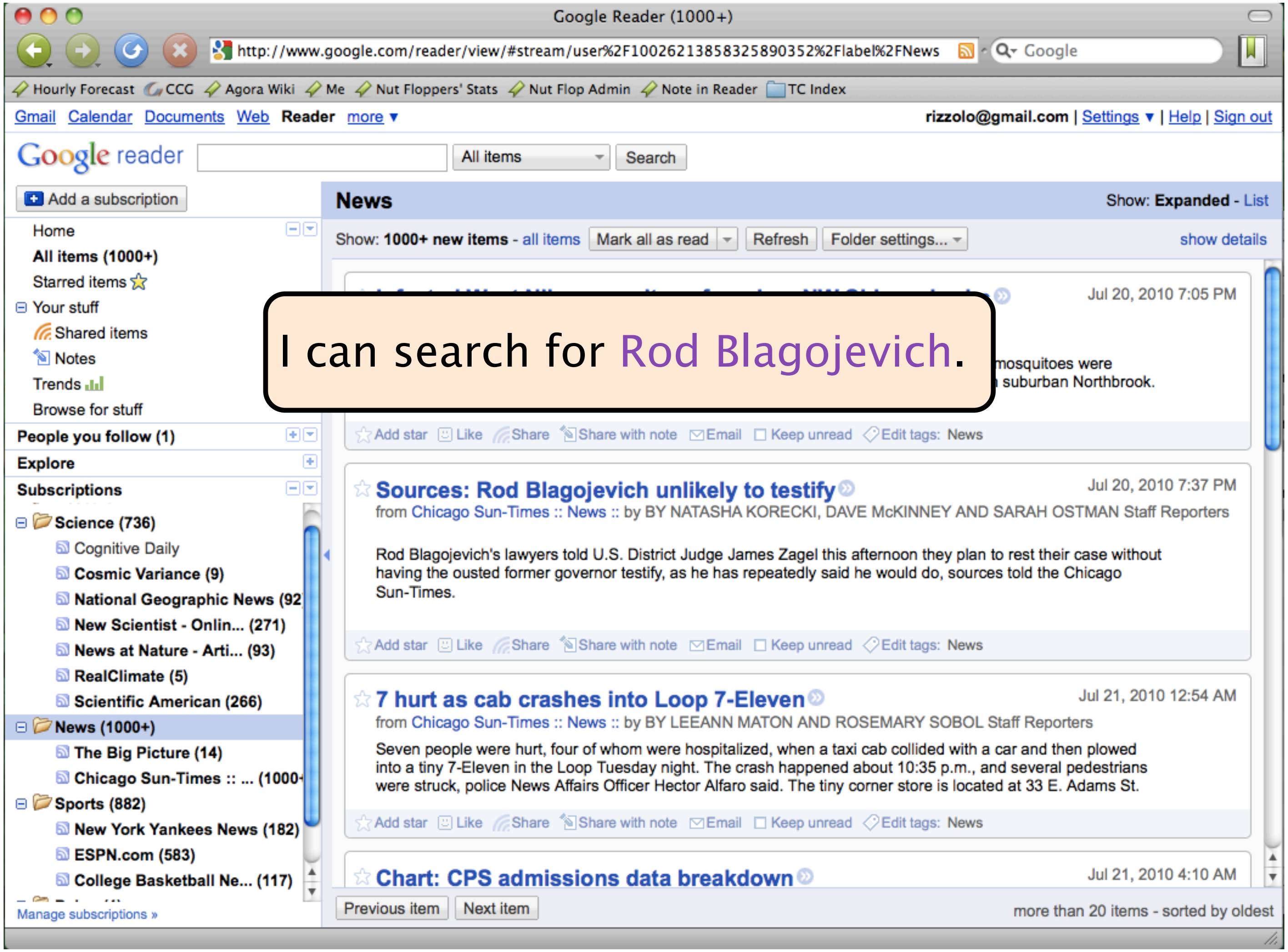
**7 hurt as cab crashes into Loop 7-Eleven** — Jul 21, 2010 12:54 AM

from Chicago Sun-Times :: News :: by BY LEEANN MATON AND ROSEMARY SOBOL Staff Reporters

Seven people were hurt, four of whom were hospitalized, when a taxi cab collided with a car and then plowed into a tiny 7-Eleven in the Loop Tuesday night. The crash happened about 10:35 p.m., and several pedestrians were struck, police News Affairs Officer Hector Alfaro said. The tiny corner store is located at 33 E. Adams St.

**Chart: CPS admissions data breakdown** — Jul 21, 2010 4:10 AM

# News readers help with organization



Jul 20, 2010 7:05 PM

mosquitoes were suburban Northbrook.

I can search for Rod Blagojevich.

**Sources: Rod Blagojevich unlikely to testify**

Jul 20, 2010 7:37 PM

from Chicago Sun-Times :: News :: by BY NATASHA KORECKI, DAVE McKINNEY AND SARAH OSTMAN Staff Reporters

Rod Blagojevich's lawyers told U.S. District Judge James Zagel this afternoon they plan to rest their case without having the ousted former governor testify, as he has repeatedly said he would do, sources told the Chicago Sun-Times.

**7 hurt as cab crashes into Loop 7-Eleven**

Jul 21, 2010 12:54 AM

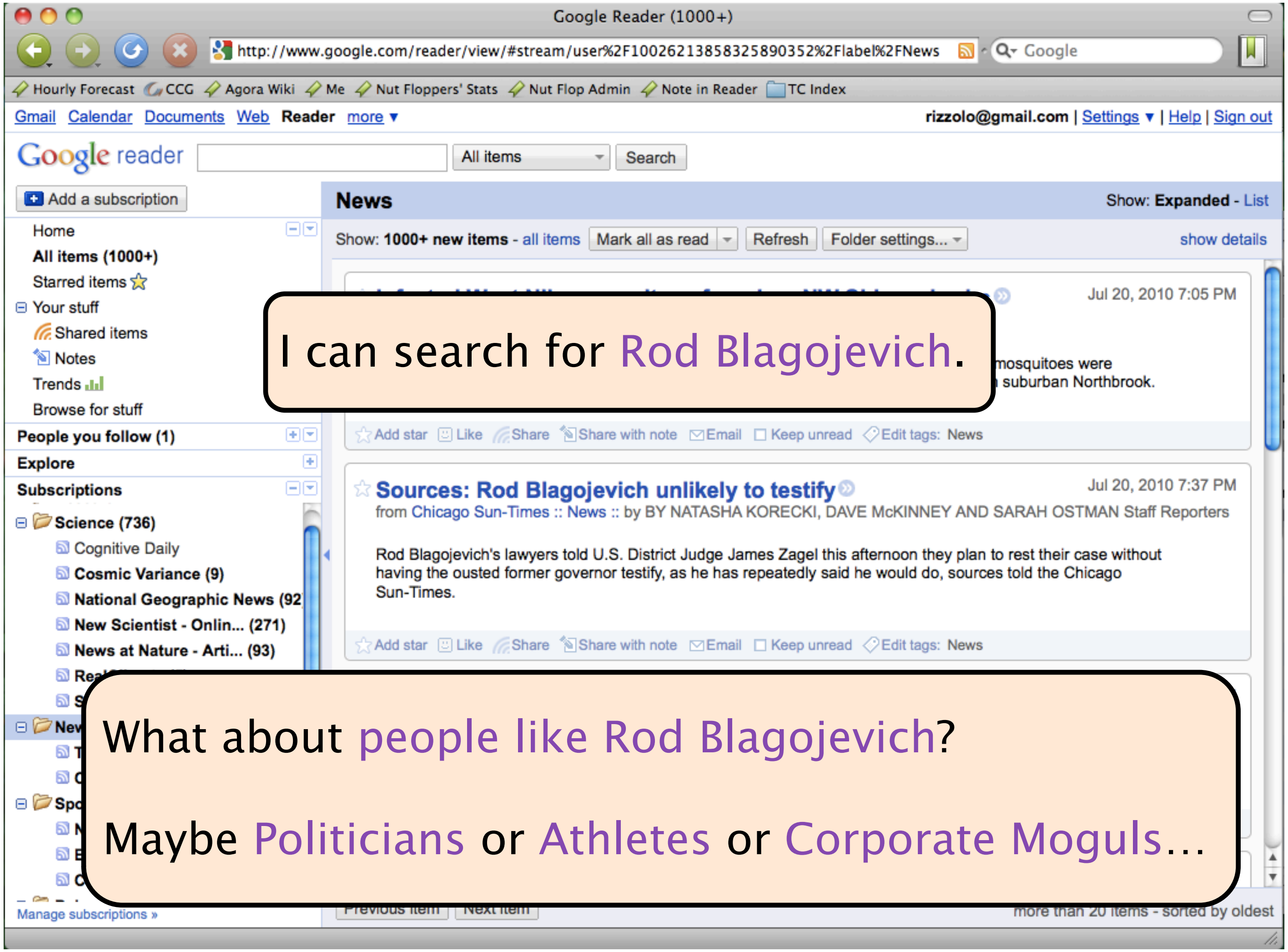
from Chicago Sun-Times :: News :: by BY LEEANN MATON AND ROSEMARY SOBOL Staff Reporters

Seven people were hurt, four of whom were hospitalized, when a taxi cab collided with a car and then plowed into a tiny 7-Eleven in the Loop Tuesday night. The crash happened about 10:35 p.m., and several pedestrians were struck, police News Affairs Officer Hector Alfaro said. The tiny corner store is located at 33 E. Adams St.

**Chart: CPS admissions data breakdown**

Jul 21, 2010 4:10 AM

# News readers help with organization



Jul 20, 2010 7:05 PM

mosquitoes were … suburban Northbrook.

**Sources: Rod Blagojevich unlikely to testify**

Jul 20, 2010 7:37 PM

from Chicago Sun-Times :: News :: by BY NATASHA KORECKI, DAVE McKINNEY AND SARAH OSTMAN Staff Reporters

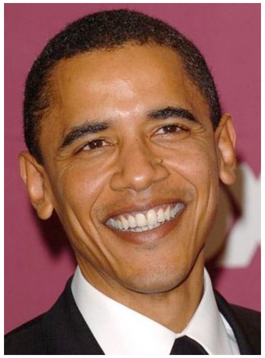
Rod Blagojevich's lawyers told U.S. District Judge James Zagel this afternoon they plan to rest their case without having the ousted former governor testify, as he has repeatedly said he would do, sources told the Chicago Sun-Times.

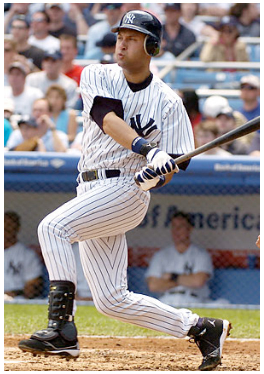
I can search for Rod Blagojevich.

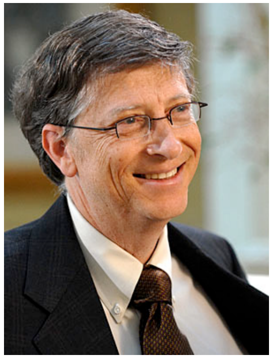
What about people like Rod Blagojevich?

Maybe Politicians or Athletes or Corporate Moguls…

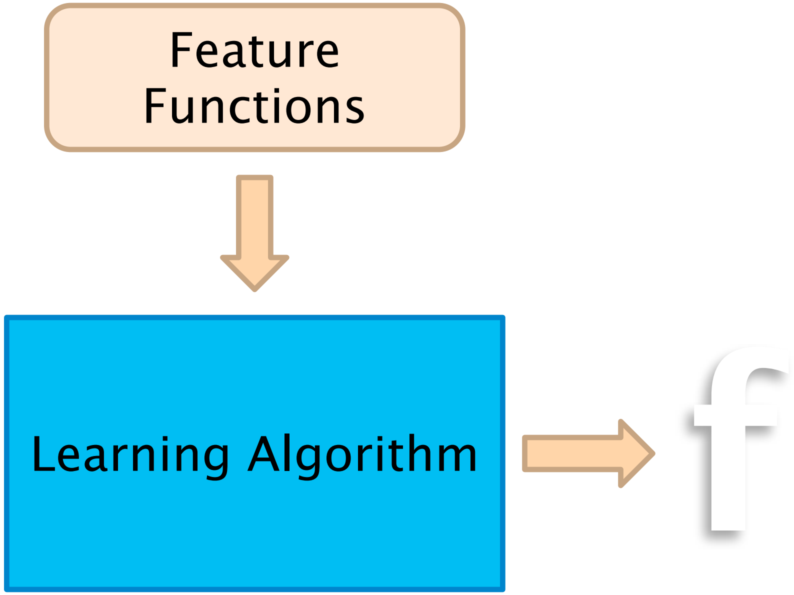
## What we need:

f(  ) = “politician”

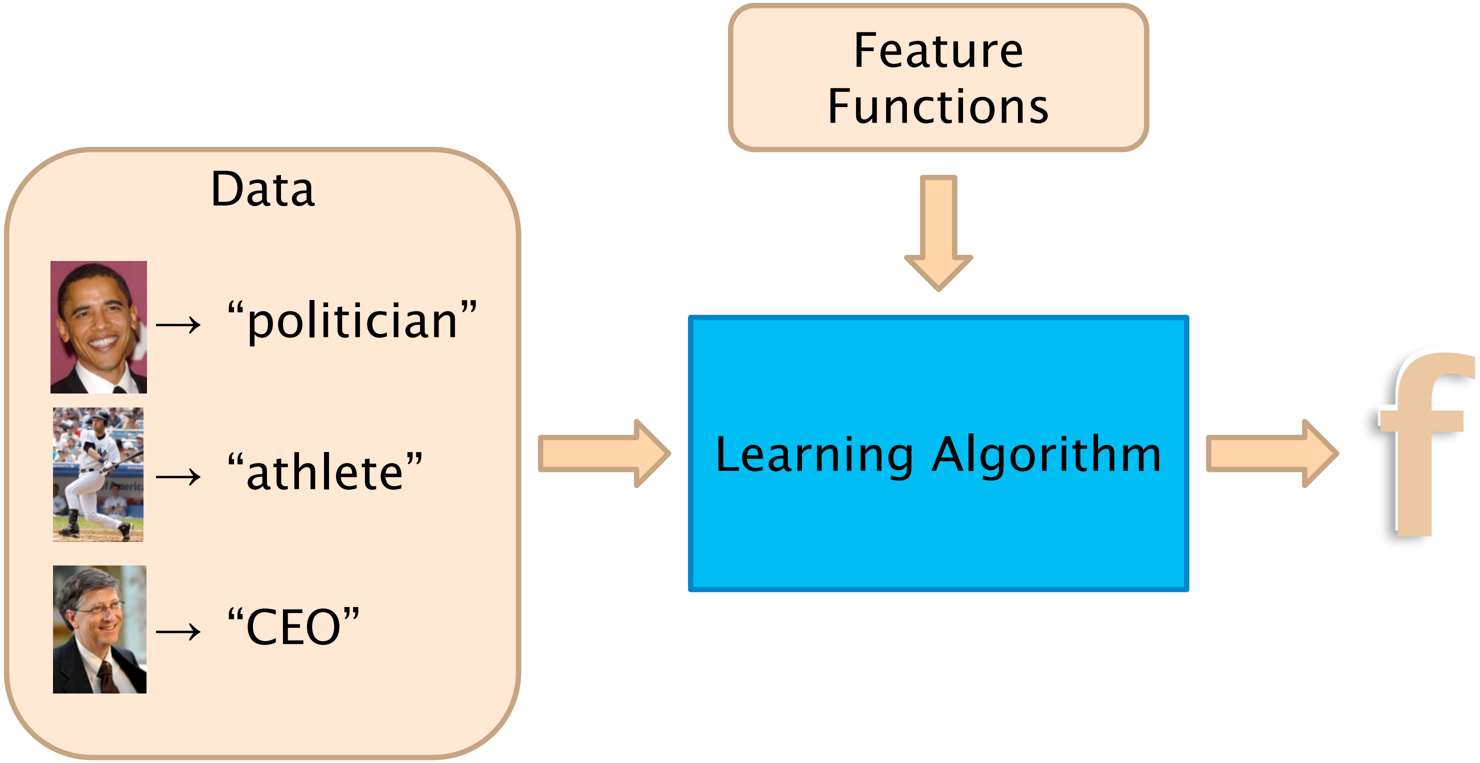
f(  ) = “athlete”

f(  ) = “CEO”

# Where to get it:  Machine Learning



# Where to get it:  Machine Learning



Feature Functions

Data

→ “politician”

→ “athlete”

→ “CEO”

Learning Algorithm

f

# So, what are “feature functions”?

- Take same input as f
- Indicate some property of the input a.k.a., a feature

# So, what are “feature functions”?

- Take same input as f
- Indicate some property of the input a.k.a., a feature

- **Typical NLP feature functions**
  - **Binary**
    - Appearance of a given word
    - Appearance of two words consecutively a.k.a., a bigram
    - Appearance of a word with a given part of speech
    - Appearance of a named entity (e.g. “Barack Obama”)
  - **Real**
    - Counts of binary features
    - TFIDF (a statistical measure of a document)

# In This Tutorial, We Will…

- Introduce **Learning Based Java (LBJ)**
  - A modeling language for learning and inference
- Introduce our state-of-the-art **NLP tools**
  - Describe their functionality
  - Demonstrate their use in new classifiers

# In This Tutorial, We Will…

- Day 1:
  - Simple Classifiers
    - 20 Newsgroups
    - Spam detection
    - Language identification
- Day 2:
  - The "Fame" Classifier
  - Write a program that:
    - Takes a collection of news articles
    - Outputs lists of people organized by what they're famous for