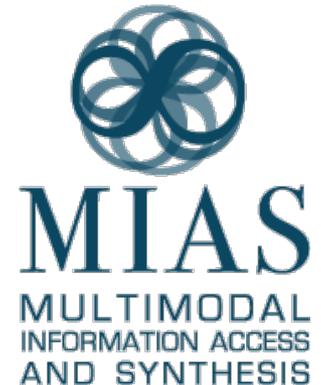




An Introduction to Machine Learning and Natural Language Processing Tools



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

Advisor: Dan Roth

Motivation: the News



Motivation: the News



Motivation: the News



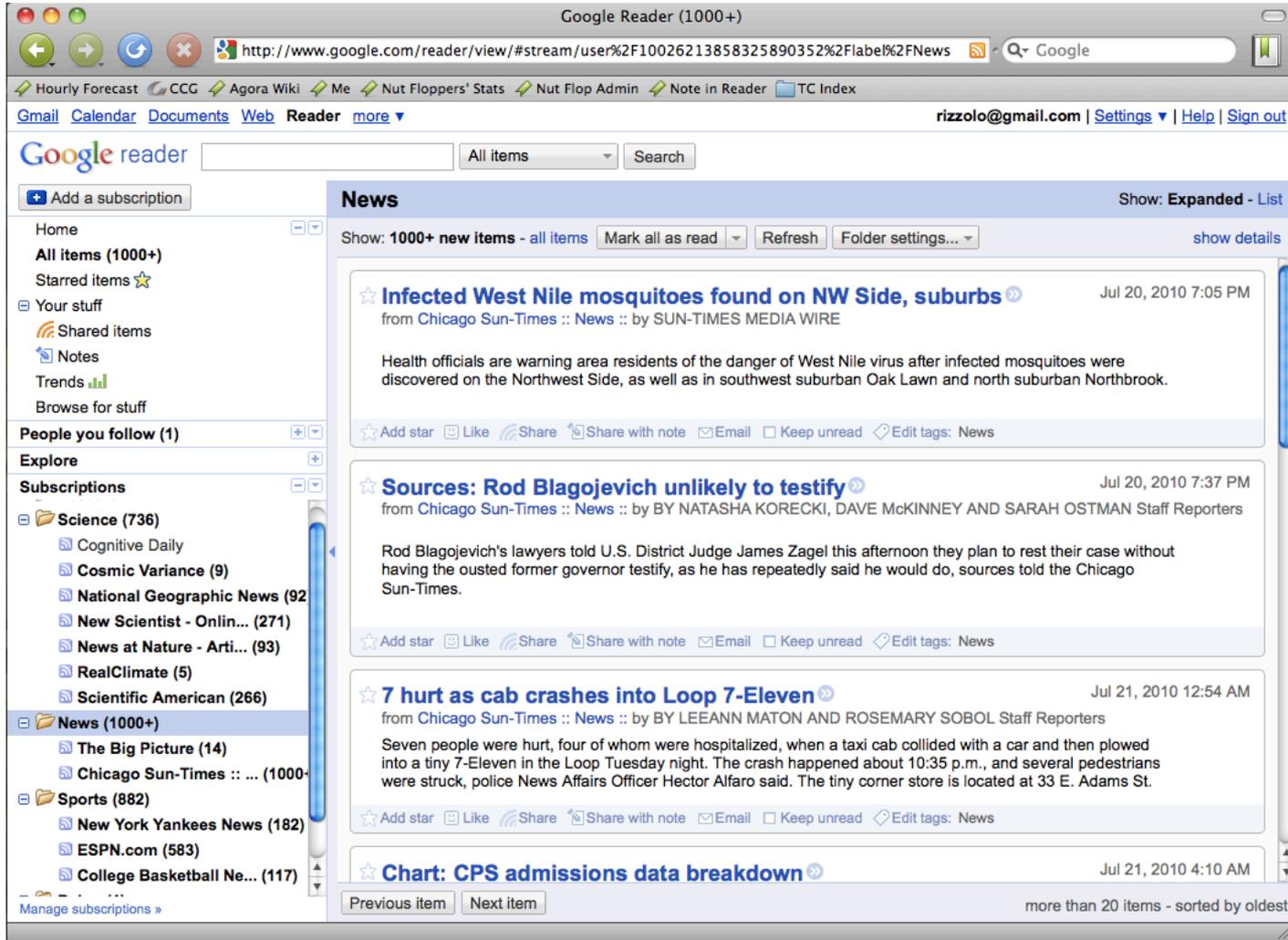
Motivation: the News



Motivation: the News



News readers help with organization



The screenshot shows the Google Reader web interface. The browser address bar displays the URL: <http://www.google.com/reader/view/#stream/user%2F10026213858325890352%2Flabel%2FNews>. The page title is "Google Reader (1000+)". The user is logged in as "rizzolo@gmail.com".

The interface includes a navigation menu on the left with sections: "Home", "All items (1000+)", "Starred items", "Your stuff", "Shared items", "Notes", "Trends", "Browse for stuff", "People you follow (1)", "Explore", and "Subscriptions". The "Subscriptions" section is expanded to show "Science (736)" and "News (1000+)". Under "News (1000+)", several sources are listed, including "The Big Picture (14)", "Chicago Sun-Times :: ... (1000-)", "Sports (882)", "New York Yankees News (182)", "ESPN.com (583)", and "College Basketball Ne... (117)".

The main content area displays a list of news items under the heading "News". The items are sorted by "Show: Expanded - List". The first item is "Infected West Nile mosquitoes found on NW Side, suburbs" from "Chicago Sun-Times :: News" by "SUN-TIMES MEDIA WIRE", dated "Jul 20, 2010 7:05 PM". The second item is "Sources: Rod Blagojevich unlikely to testify" from "Chicago Sun-Times :: News" by "BY NATASHA KORECKI, DAVE McKINNEY AND SARAH OSTMAN Staff Reporters", dated "Jul 20, 2010 7:37 PM". The third item is "7 hurt as cab crashes into Loop 7-Eleven" from "Chicago Sun-Times :: News" by "BY LEEANN MATON AND ROSEMARY SOBOL Staff Reporters", dated "Jul 21, 2010 12:54 AM". The fourth item is "Chart: CPS admissions data breakdown" dated "Jul 21, 2010 4:10 AM".

Each news item includes a star icon, a "Like" button, a "Share" button, a "Share with note" button, an "Email" button, a "Keep unread" checkbox, and an "Edit tags" button. The interface also features a search bar, a "Add a subscription" button, and a "Manage subscriptions" link at the bottom left.



News readers help with organization

The screenshot shows the Google Reader web interface. The browser address bar displays the URL: <http://www.google.com/reader/view/#stream/user%2F10026213858325890352%2Flabel%2FNews>. The page title is "Google Reader (1000+)". The user is logged in as "rizzolo@gmail.com". The interface includes a search bar, a list of subscriptions on the left, and a main news feed. A callout box with a black border and white background contains the text "I can search for Rod Blagojevich." in black font. The news feed shows several articles, including one titled "Sources: Rod Blagojevich unlikely to testify" and another titled "7 hurt as cab crashes into Loop 7-Eleven".

Google Reader (1000+)

http://www.google.com/reader/view/#stream/user%2F10026213858325890352%2Flabel%2FNews

Hourly Forecast CCG Agora Wiki Me Nut Floppers' Stats Nut Flop Admin Note in Reader TC Index

Gmail Calendar Documents Web Reader more ▾ rizzolo@gmail.com | Settings ▾ | Help | Sign out

Google reader All items Search

+ Add a subscription

Home

All items (1000+)

Starred items ☆

Your stuff

Shared items

Notes

Trends

Browse for stuff

People you follow (1)

Explore

Subscriptions

Science (736)

- Cognitive Daily
- Cosmic Variance (9)
- National Geographic News (92)
- New Scientist - Onlin... (271)
- News at Nature - Arti... (93)
- RealClimate (5)
- Scientific American (266)

News (1000+)

- The Big Picture (14)
- Chicago Sun-Times :: ... (1000+)

Sports (882)

- New York Yankees News (182)
- ESPN.com (583)
- College Basketball Ne... (117)

Manage subscriptions »

News Show: Expanded - List

Show: 1000+ new items - all items Mark all as read Refresh Folder settings... show details

Jul 20, 2010 7:05 PM

mosquitoes were suburban Northbrook.

Add star Like Share Share with note Email Keep unread Edit tags: News

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.

Add star Like Share Share with note Email Keep unread Edit tags: News

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.

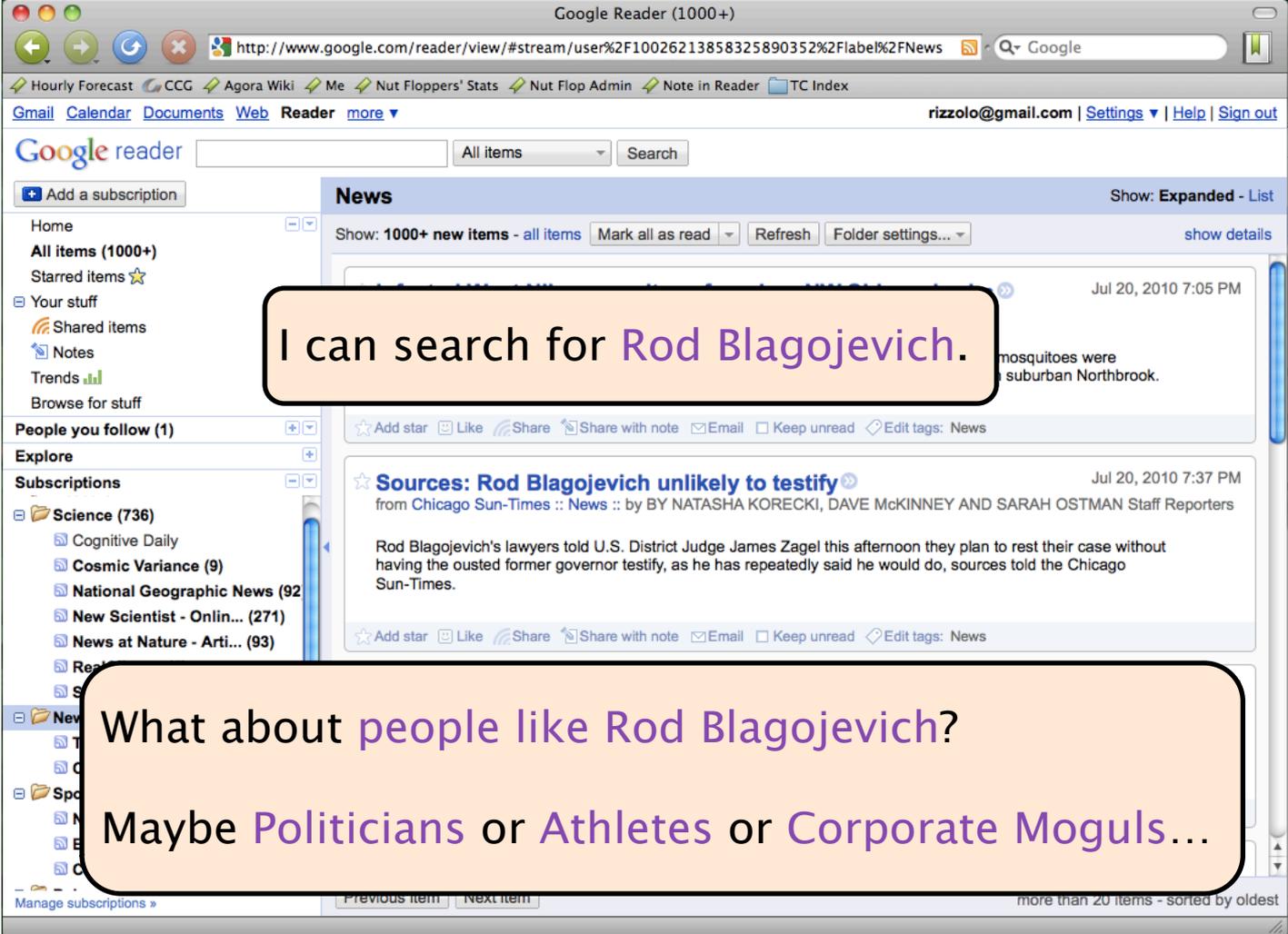
Add star Like Share Share with note Email Keep unread Edit tags: News

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

Previous item Next item more than 20 items - sorted by oldest



News readers help with organization



The screenshot shows the Google Reader web interface. The browser address bar displays the URL: <http://www.google.com/reader/view/#stream/user%2F10026213858325890352%2Flabel%2FNews>. The page title is "Google Reader (1000+)". The user is logged in as "rizzolo@gmail.com". The interface includes a search bar, a "News" section with "Show: 1000+ new items - all items", and a list of news items. A specific article is highlighted with the title "Sources: Rod Blagojevich unlikely to testify" from the Chicago Sun-Times, dated July 20, 2010, 7:37 PM. The article text states: "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." The interface also shows a sidebar with "All items (1000+)", "Starred items", "Your stuff", "People you follow (1)", "Explore", and "Subscriptions" (including Science with 736 items).

I can search for Rod Blagojevich.

What about people like Rod Blagojevich?
Maybe Politicians or Athletes or Corporate Moguls...

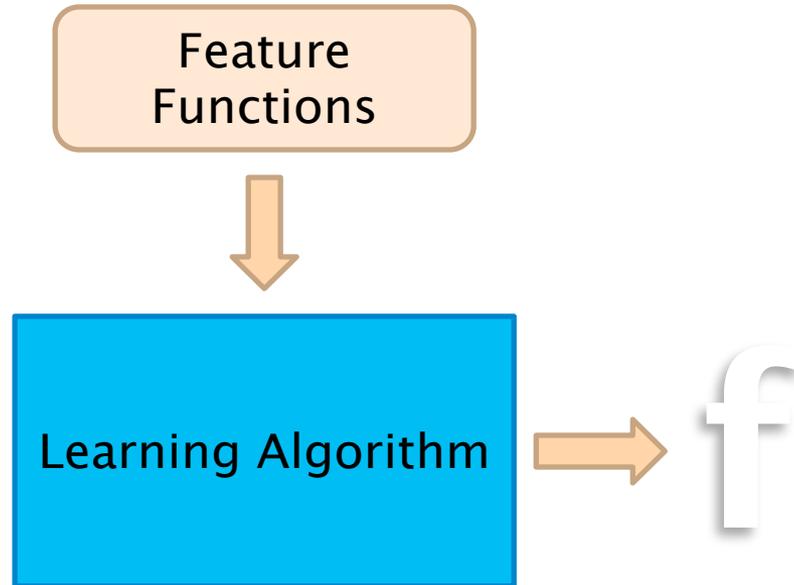
What we need:

$f(\text{img}_{\text{Obama}}) = \text{“politician”}$

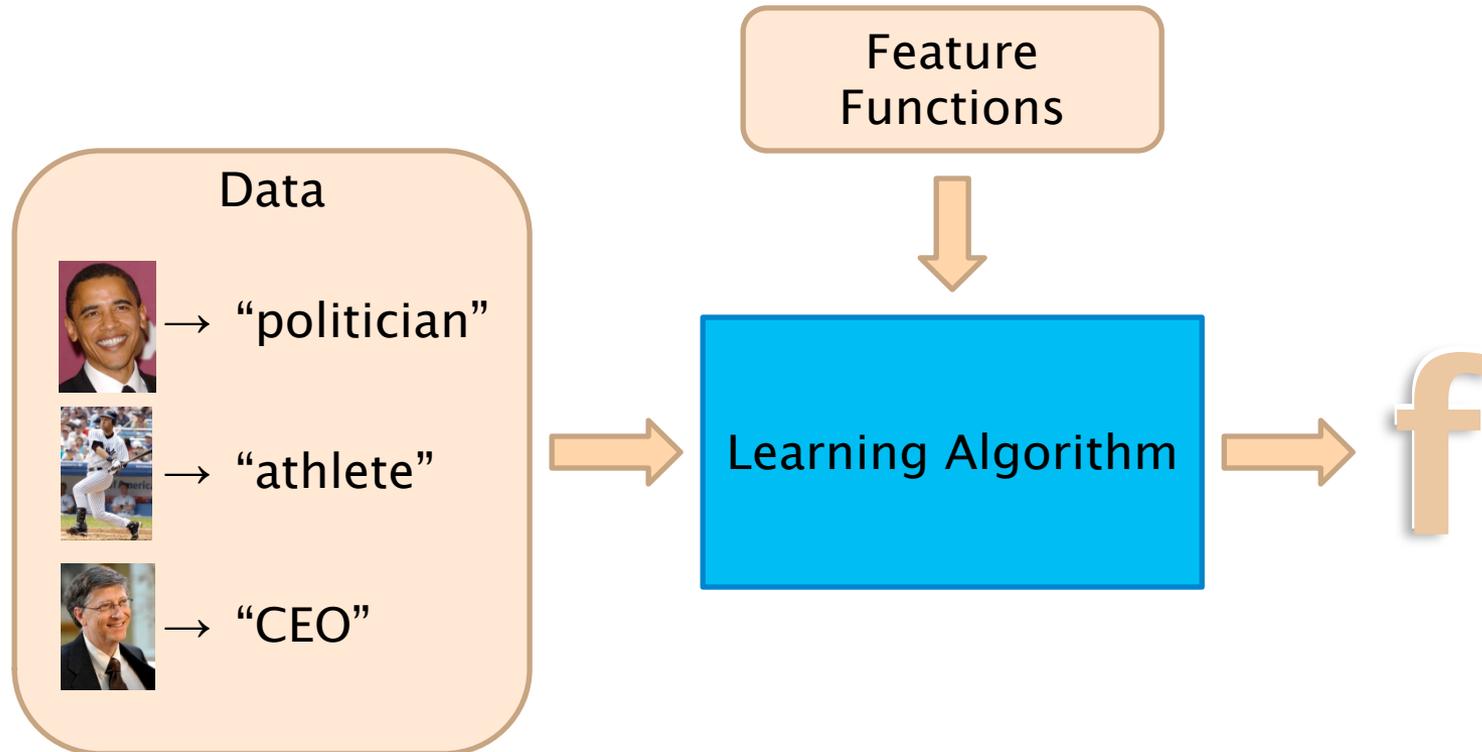
$f(\text{img}_{\text{baseball}}) = \text{“athlete”}$

$f(\text{img}_{\text{Bill Gates}}) = \text{“CEO”}$

Where to get it: Machine Learning



Where to get it: Machine Learning



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

