Welcome to Data Bootcamp

Joseph Adler, Drew Conway, Jake Hofman, Hilary Mason

February 1, 2011

Creative Commons Attribution-Share Alike 3.0

Speakers

@jadler, @hmason, @drewconway, @jakehofman

Joseph Adler

LinkedIn

Joseph Adler has many years of experience in data mining and data analysis at companies including DoubleClick, American Express, and VeriSign. He graduated from MIT with an B.Sc. and M.Eng in Computer Science and Electrical Engineering. He is the inventor of several patents for computer security and cryptography, and the author of "Baseball Hacks" and "R in a Nutshell". Currently, he is a senior data scientist at LinkedIn.

Hilary Mason

bit.ly

Hilary is the lead scientist at bit.ly, where she is finding sense in vast data sets. She is a former computer science professor with a background in machine learning and data mining, has published numerous academic papers, and regularly releases code on her personal site, www.hilarymason.com. She has discovered two new species, loves to bake cookies, and asks way too many questions.

► Web site

Drew Conway

New York University

Drew Conway is a PhD student in political science at New York University. Drew studies terrorism and armed conflict; using tools from mathematics and computer science to gain a deeper understanding of these phenomena.

► Web site

Jake Hofman

Yahoo!

Jake Hofman is a member of the Human Social Dynamics group at Yahoo! Research. His work involves data-driven modeling of social data, focusing on applications of machine learning and statistical inference to large-scale data. He holds a B.S. in Electrical Engineering from Boston University and a Ph.D. in Physics from Columbia University.

Play along at home

All of the materials from today's tutorial are available on Github:

Clone the repository for data/code/slides

```
git clone https://github.com/drewconway/strata_bootcamp
```

Data-dependent products

★★★★★

Not Interested

Our best guess for Jake: **5 stars**

Average of 4,275,920 ratings: **3.8 stars**

The Big Lebowski

1998 R 117 minutes

Slacker Jeff "The Dude" Lebowski (Jeff Bridges) gets involved in a gargantuan mess of events when he's mistaken for another man named Lebowski, whose wife has been kidnapped and is being held for ransom. All the while, Dude's friend, Walter (John Goodman), stirs the pot. Brothers Joel Coen and Ethan Coen write and direct this cult comedy classic that also stars Steve Buscemi, Philip Seymour Hoffman, Julianne Moore and John Turturro.

Cast: Jeff Bridges, John Goodman, Philip Seymour Hoffman, Steve Buscemi, Julianne Moore, Tara Reid, Peter Stormare, David Huddleston, Philip Moon, Mark Pellegrino, Flea, Torsten Voges, Jimmie Dale Gilmore, Jack Kehler, John Turturro, James G. Hoosier, Richard Gant, Christian Clemenson, David Thewlis, Peter Siragusa, Sam Elliott, Ben Gazzara, Jon Polito, Asia Carrera, Paris Themmen

Director: Joel Coen

Genres: Comedy, Cult Comedies, Universal Studios Home Entertainment, Blu-ray

This movie is: Quirky, Witty

Format: DVD and streaming (Blu-ray availability date unknown) (HD available)

Play

Add to Instant Queue

Add to DVD Queue

Play Trailer

Recommended based on your interest in: Fargo, O Brother, Where Art Thou? and No Country for Old Men

Data-dependent products

- Effective/practical systems that learn from experience impact our daily lives, e.g.:
  - Recommendation systems
  - Spam detection
  - Optical character recognition
  - Face recognition
  - Fraud detection
  - Machine translation
  - ...

Black[1]-boxified?

Google code labs e.g. "ajax apis" or "open source" Search

Watch Google I/O keynotes live on May 19 and 20!

Google Prediction API (Labs) Home Docs FAQ Forum Terms

What is the Google Prediction API?

The Prediction API enables access to Google's machine learning algorithms to analyze your historic data and predict likely future outcomes. Upload your data to Google Storage for Developers, then use the Prediction API to make real-time decisions in your applications. The Prediction API implements supervised learning algorithms as a RESTful web service to let you leverage patterns in your data, providing more relevant information to your users. Run your predictions on Google's infrastructure and scale effortlessly as your data grows in size and complexity.

How do I start?

- Learn more about Google Prediction API.
- Request access.
- Try out the sample code.

Features

- Lightweight RESTful API
- Asynchronous training
- Automatically selects from several available machine learning techniques
- Supported inputs: numeric data or unstructured text
- Outputs hundreds of discrete categories
- Accessible from many platforms: Google App Engine, Apps Script (Google Spreadsheets), web & desktop apps, and command line

Uses

- Language identification
- Customer sentiment analysis
- Product recommendations & upsell opportunities
- Message routing decisions
- Diagnostics
- Document and email classification
- Suspicious activity identification
- Churn analysis
- And many more...

[1] s/black/blue/g

Roadmap?

Step 1: Have data

Step 2: ???

Step 3: Profit

Learning by example

Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O

non urgent - whoops! yes that's what i meant, thanks for decoding my questi

SourceForge.net: variational bayes for network modularity - can i get admin

Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery a

Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.

Re: JAFOS 2008, Applied Math Session - yes. the listening post dude. On N

Access to over 5,000 Health Plan Choices! - Affordable health insurance. Ins

More effective - If you are having trouble viewing this email click here. Thurs

Special Offer! Cialis, Viagra, VicodinES! - Order all your Favorite Rx~Medica

Financial Aid Available: Find Funding for Your Education - Get the financial a

Find The Perfect School and Financal Aid for your College Degree - HI ! It ha

PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

Learning by example

Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O
non urgent - whoops! yes that's what i meant, thanks for decoding my questi
SourceForge.net: variational bayes for network modularity - can i get admin
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery a
Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.
Re: JAFOS 2008, Applied Math Session - yes. the listening post dude. On N
Access to over 5,000 Health Plan Choices! - Affordable health insurance. Ins
More effective - If you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, VicodinES! - Order all your Favorite Rx~Medica
Financial Aid Available: Find Funding for Your Education - Get the financial a
Find The Perfect School and Financal Aid for your College Degree - HI ! It ha
PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

- How did you solve this problem?
- Can you make this process explicit (e.g. write code to do so)?

Learning by example

Learning by example

- We learn quickly from few, relatively unstructured examples ... but we don't understand *how* we accomplish this
- Can we develop algorithms that enable machines to learn by example from large data sets?

Goals

- Many fields ...
  - Statistics
  - Pattern recognition
  - Data mining
  - Machine learning
- ... similar goals
  - Extract and recognize patterns in data
  - Interpret or explain observations
  - Test validity of hypotheses
  - Efficiently search the space of hypotheses
  - Design efficient algorithms enabling machines to learn from data

Statistics vs. machine learning[2]

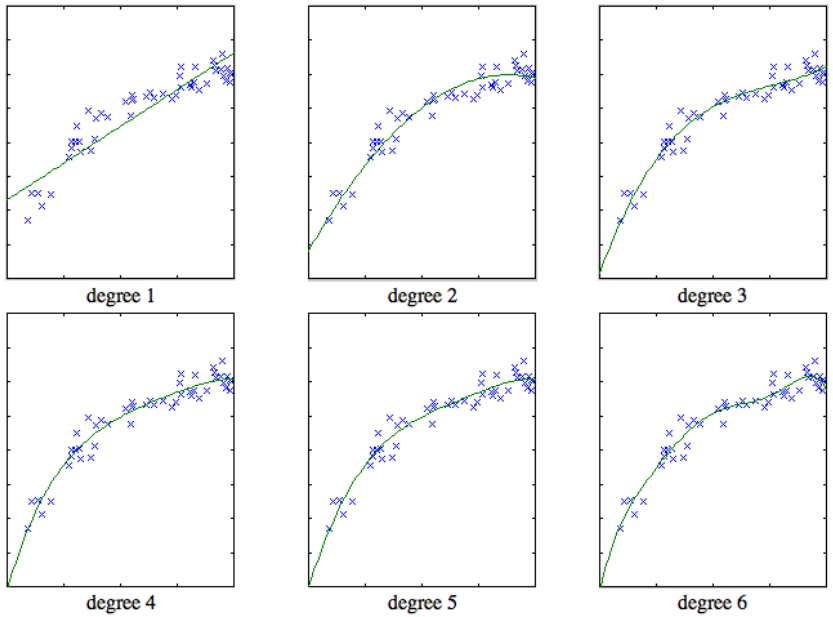
Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

[2] http://anyall.org/blog/2008/12/statistics-vs-machine-learning-fight/

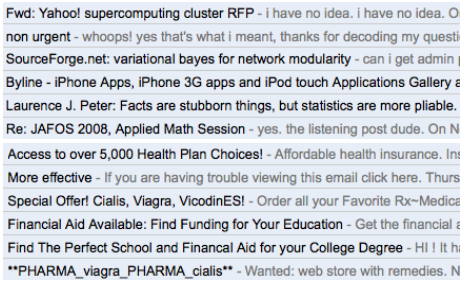
Philosophy

- We would like models that:
  - Provide predictive and explanatory power
  - Are complex enough to describe observed phenomena
  - Are simple enough to generalize to future observations



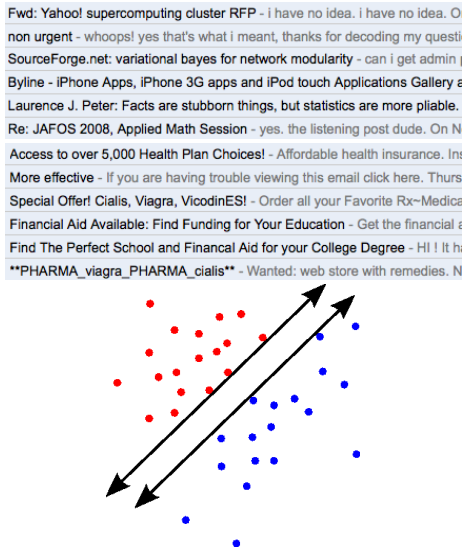
Framework

1. Get data



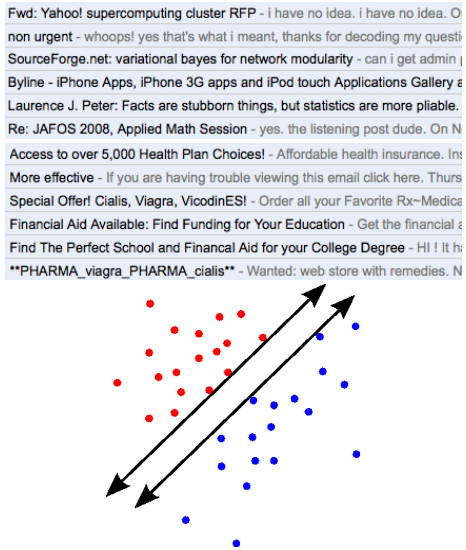
Framework

1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data



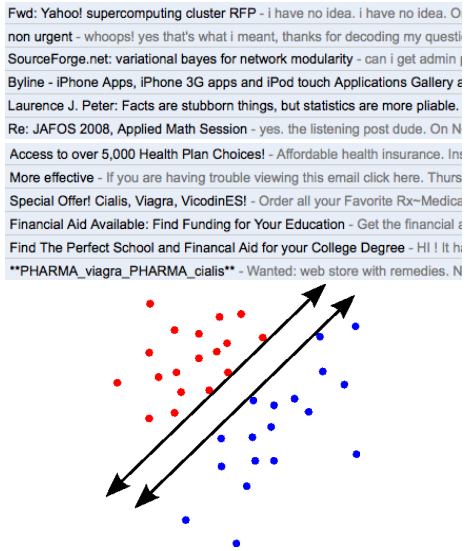
Framework

1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data
5. Specify model
6. Specify loss function



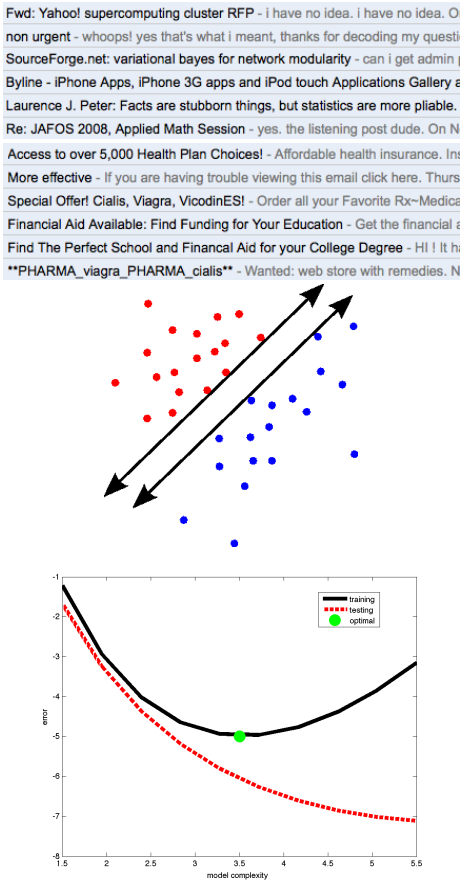
Framework

1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data
5. Specify model
6. Specify loss function
7. Develop algorithm to minimize loss



Framework

1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data
5. Specify model
6. Specify loss function
7. Develop algorithm to minimize loss
8. Choose performance measure
9. "Train" to minimize loss
10. "Test" to evaluate generalization



Topics

- Supervised
  - Linear regression
  - Classification / regression trees
  - Logistic regression
  - Naive Bayes
  - k-nearest neighbors
  - Support vector machines
  - Boosting
- Unsupervised
  - K-means
  - Mixture models
  - Principal components analysis
  - Factor analysis
  - Topic models
  - Collaborative filtering

Topics

- Supervised
  - Linear regression
  - Classification / regression trees
  - Logistic regression
  - Naive Bayes
  - k-nearest neighbors
  - Support vector machines
  - Boosting
- Unsupervised
  - K-means
  - Mixture models
  - Principal components analysis
  - Factor analysis
  - Topic models
  - Collaborative filtering
- Data representation: feature space, selection, normalization
- Model assessment: complexity control, cross-validation, ROC curve, Bayesian Occam's razor, information-theoretic measures

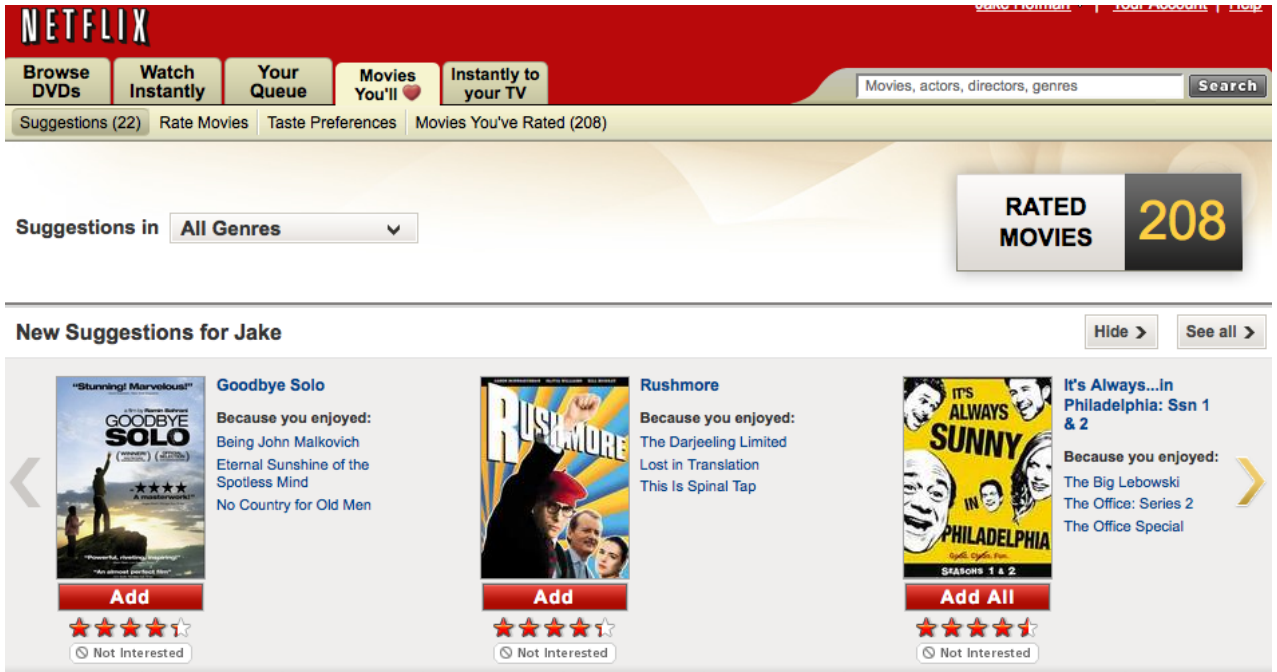
Topics

- Supervised
  - Linear regression
  - Classification / regression trees
  - Logistic regression
  - Naive Bayes
  - k-nearest neighbors
  - Support vector machines
  - Boosting
- Unsupervised
  - K-means
  - Mixture models
  - Principal components analysis
  - Factor analysis
  - Topic models
  - Collaborative filtering
- Data representation: feature space, selection, normalization
- Model assessment: complexity control, cross-validation, ROC curve, Bayesian Occam's razor, information-theoretic measures
- Probabilistic inference: graphical models, variational methods, sampling
- Large-scale learning (?)

Topics

- Simple approaches often do surprisingly well for large problems

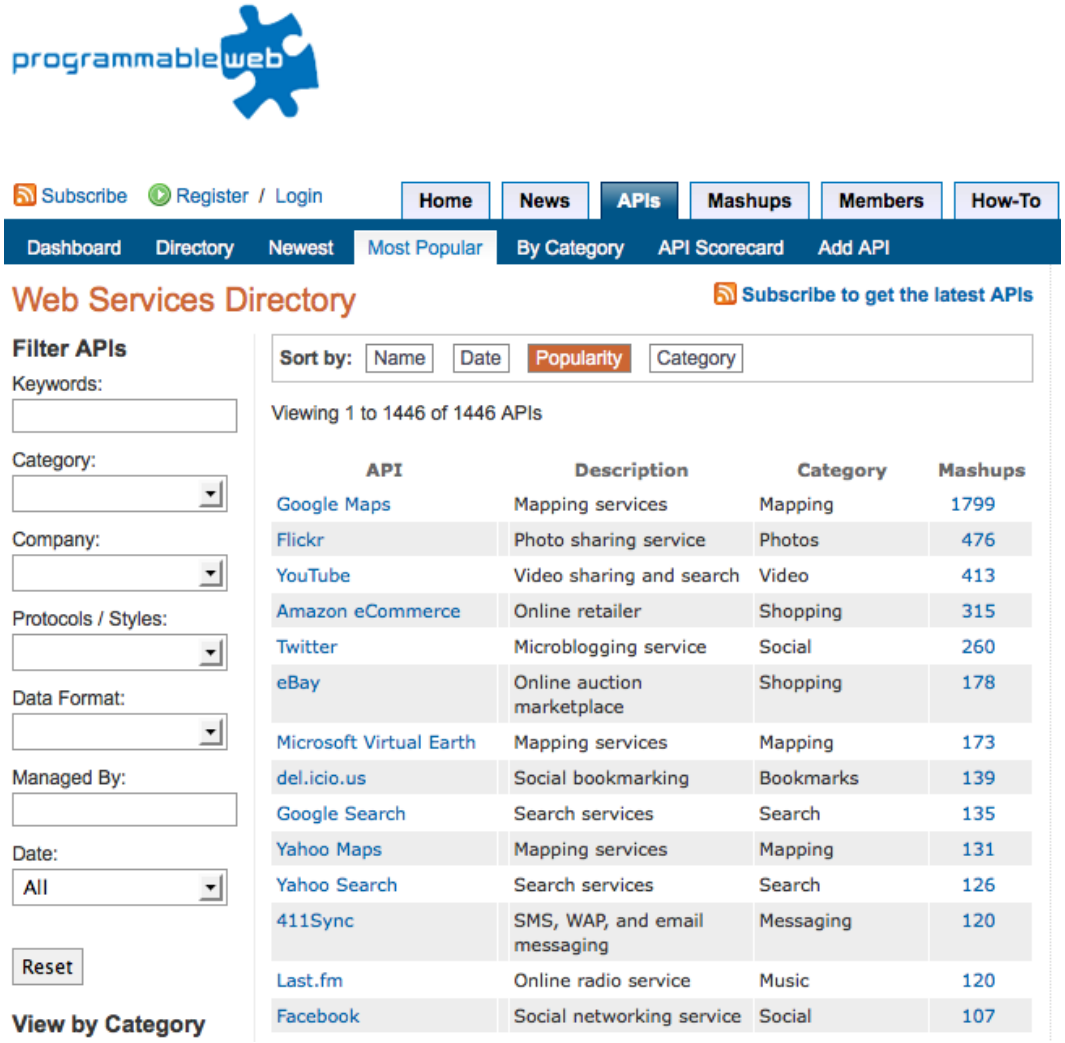
Netflix prize



- $1M for a 10% improvement in predicted rating
- More than 1000 submissions over 2.5 years
- Top two teams within 0.01% of each other

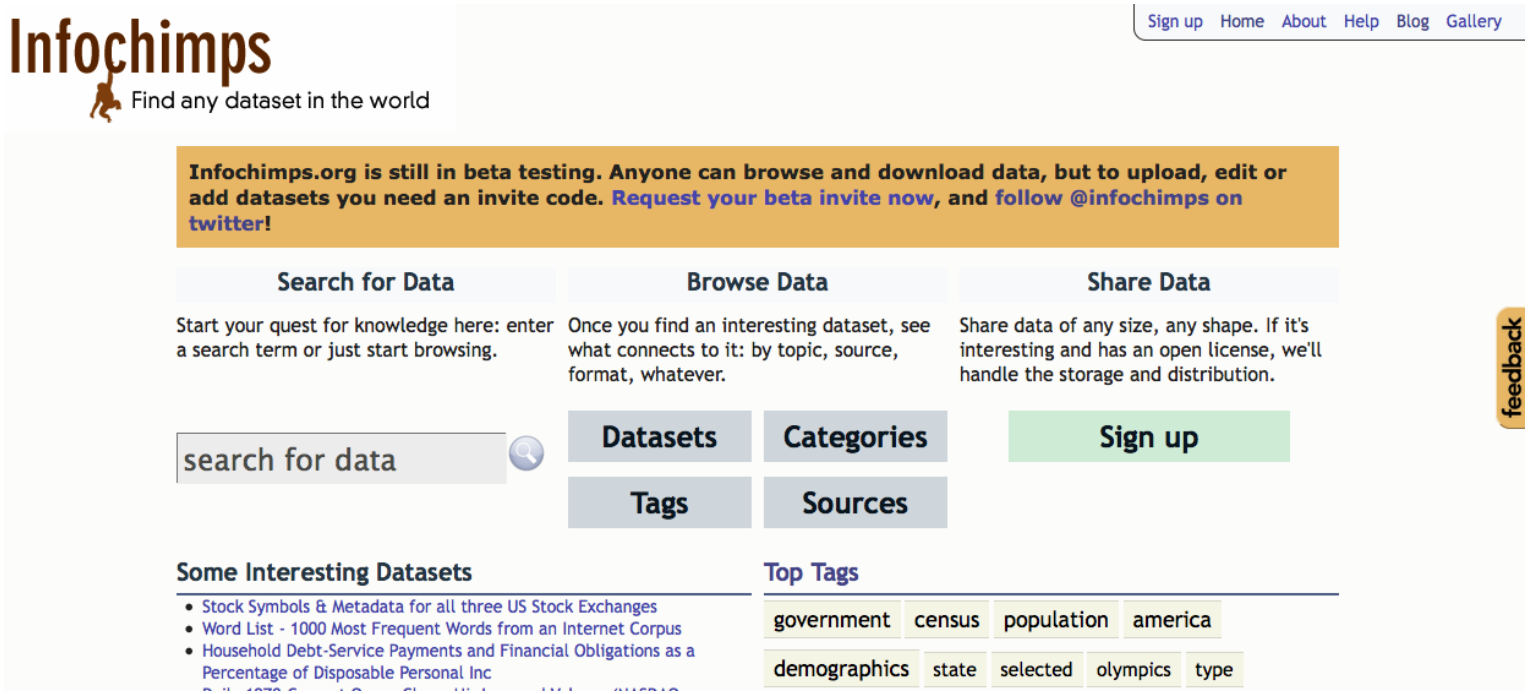
Got data?

- Web service APIs expose vast amounts of data



Got data?

- Many free, public data sets available online



Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq, ...

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq, ...

```
$ tr , '\t' < data.csv > data.tsv
```

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq, ...

```
$ tr , '\t' < data.csv > data.tsv
```

```
$ bzcat data.tsv.bz2 | awk -F'\t' 'NF != 16 {print}'
```

Coding

- Scripting: Python, Ruby, Perl, bash, ...
- Computing: R, SciPy/NumPy, MATLAB, ...
- Wrangling: sed, awk, grep, tr, wc, cut, sort, uniq, ...

```
$ tr , '\t' < data.csv > data.tsv
```

```
$ bzcat data.tsv.bz2 | awk -F'\t' 'NF != 16 {print}'
```

```
$ sed -e 's/<[^>]*>//g' < page.html > page.txt
```