Welcome to Data Bootcamp

Joseph Adler, Drew Conway, Jake Hofman, Hilary Mason

February 1, 2011

Getting the code

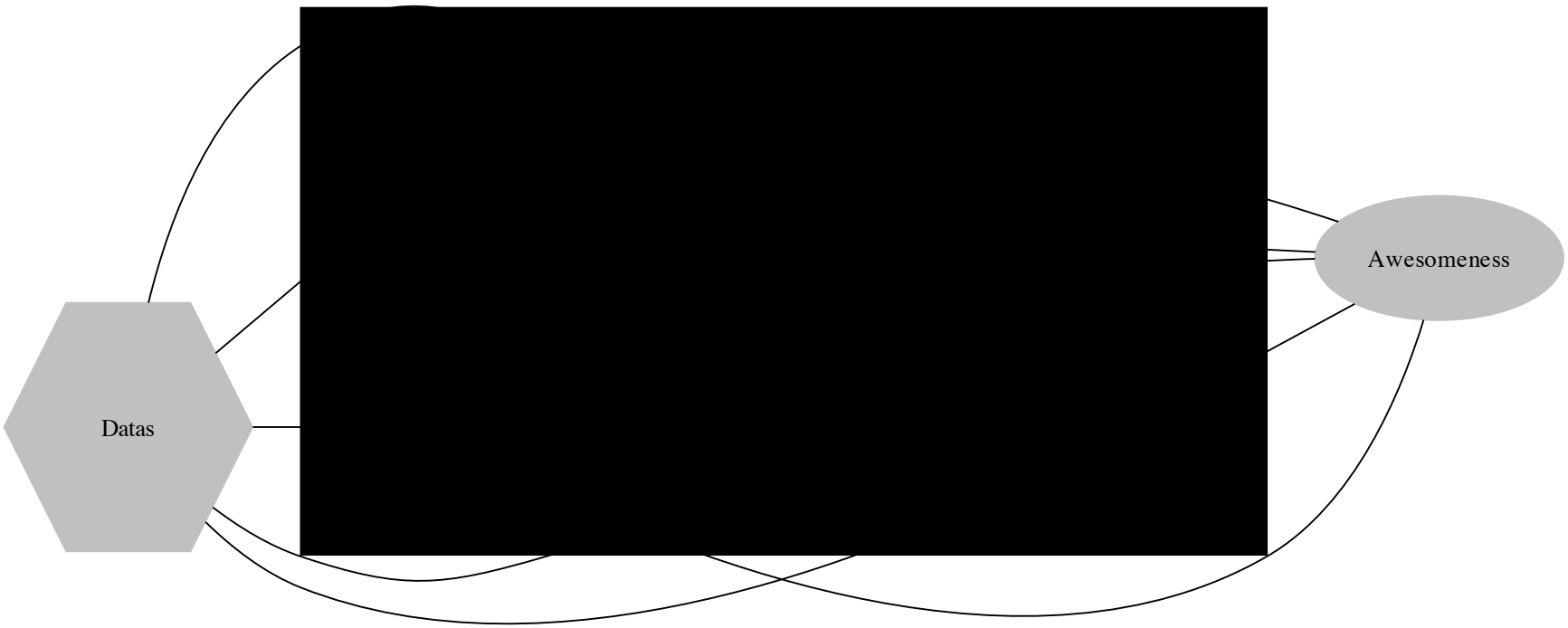
All of the slides, code and images from today's tutorial are available on Github:

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

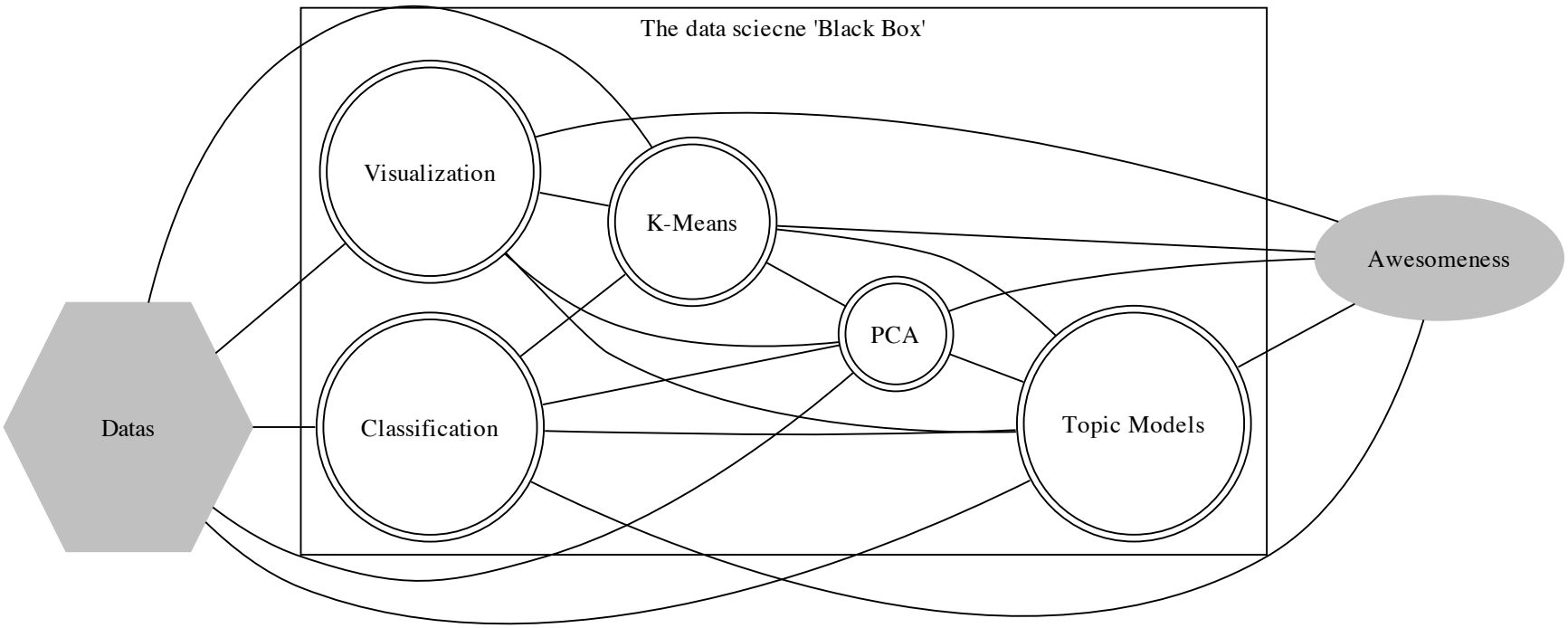
The play the home game

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

The data science "black box"



The data science "black box"



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

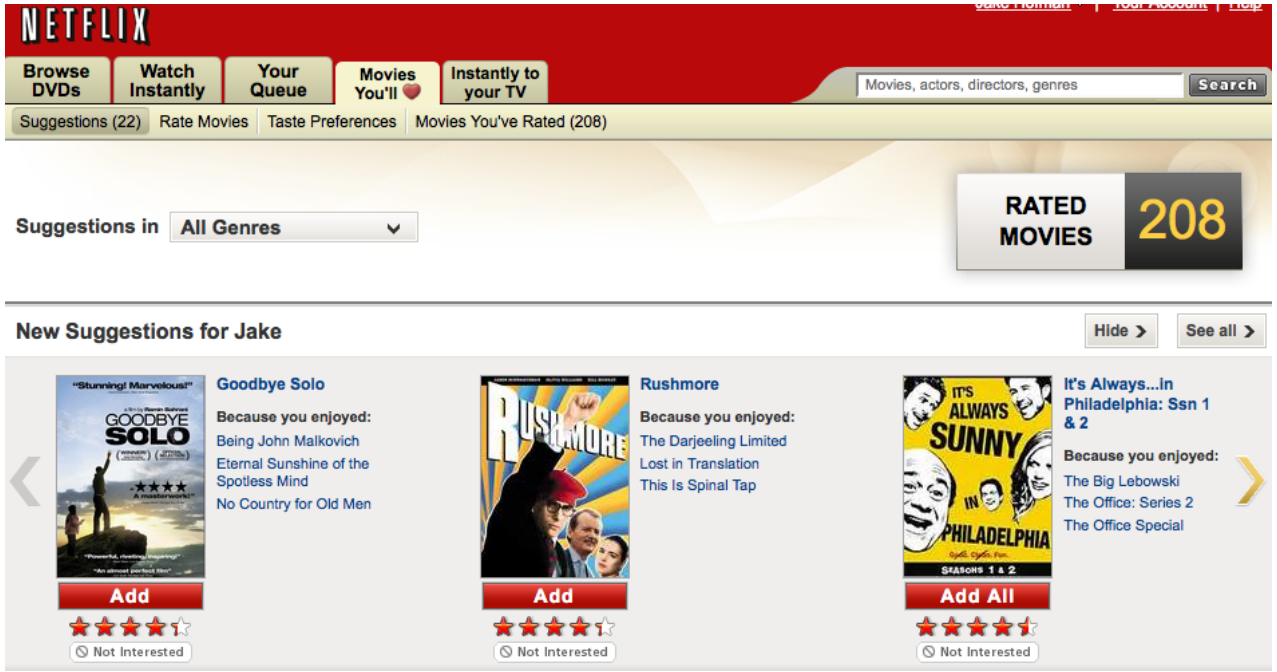
Learning by example

- 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?

Common applications

- Effective/practical algorithms exist, and impact our daily lives
- Entire industries built around these techniques, e.g.:
  - Spam detection (Email)
  - Information retrieval (Search)
  - Recommendation Systems ("You might also like ...")
  - Fraud detection (Identity theft)
  - Face recognition (Camera auto-focus)
  - Optical character recognition (Mail routing via ZIP codes)

Netflix prize



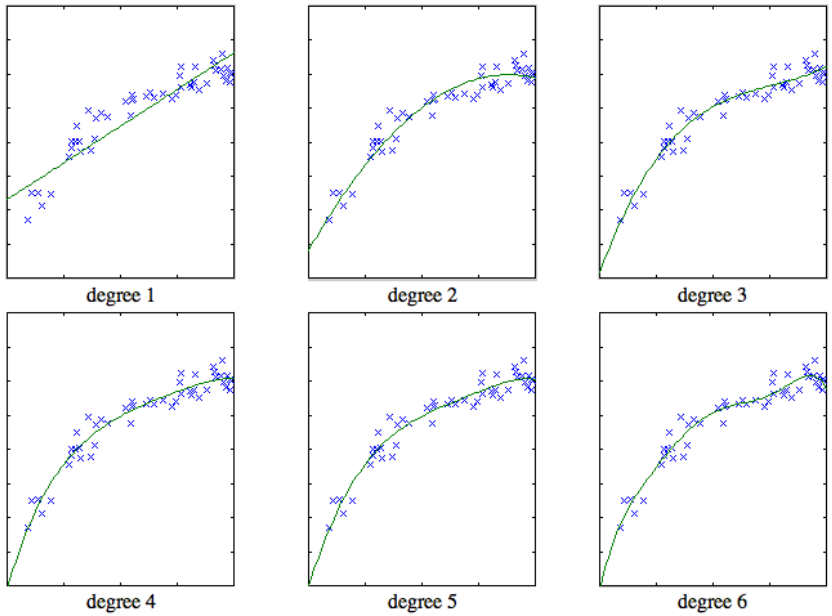
- $1M for a 10% improvement in predicted rating
- More than 1000 submissions over 2.5 years
- Top two teams within 0.01% of each other (winners announced soon)

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

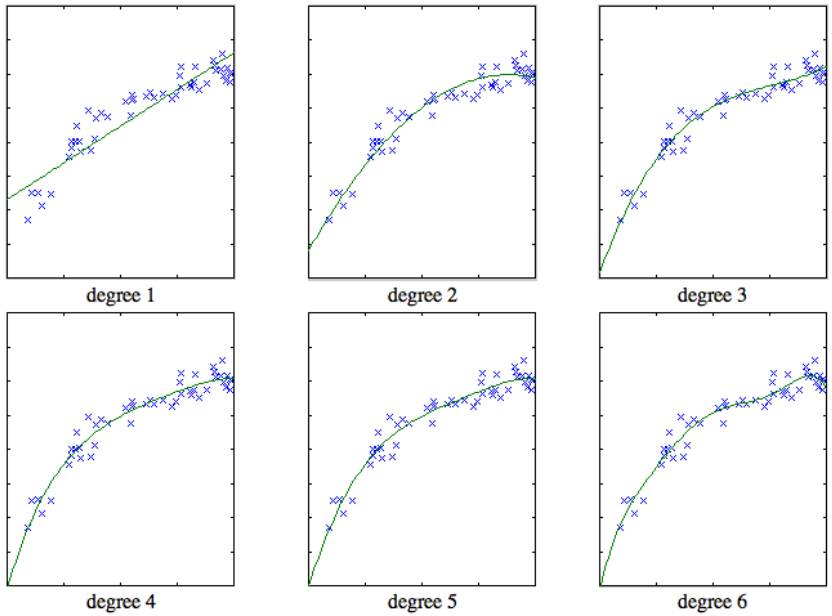
Philosophy

- 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



Philosophy

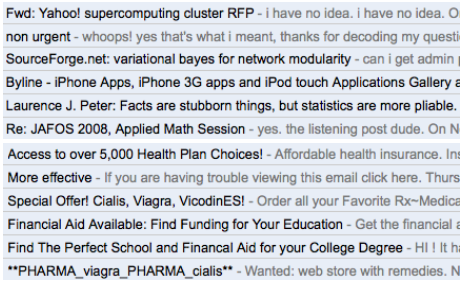
- We would like models that:
  - Provide predictive and explanatory power
  - Are complex enough to describe observed phenomena
  - Are simple enough to generalize to future observations



- How can we quantify an "optimal" model
  - What to optimize?
  - How to optimize it?

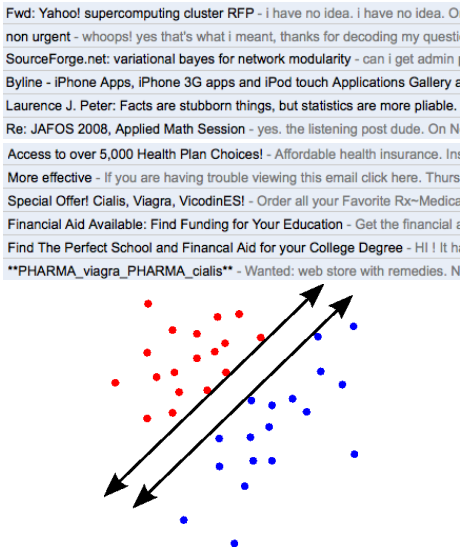
Framework

1. Get data



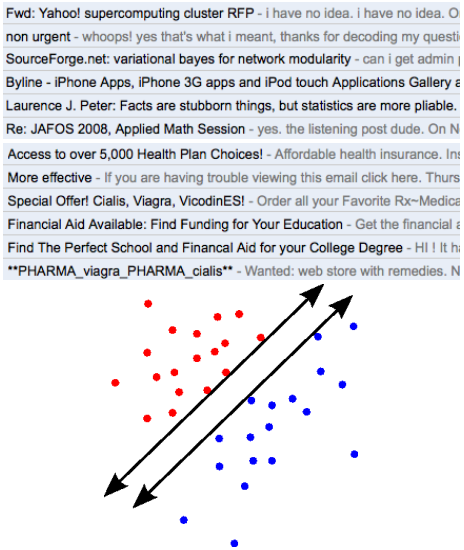
Framework

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



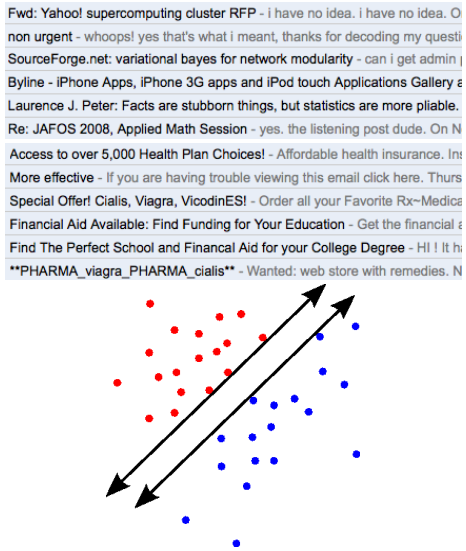
Framework

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

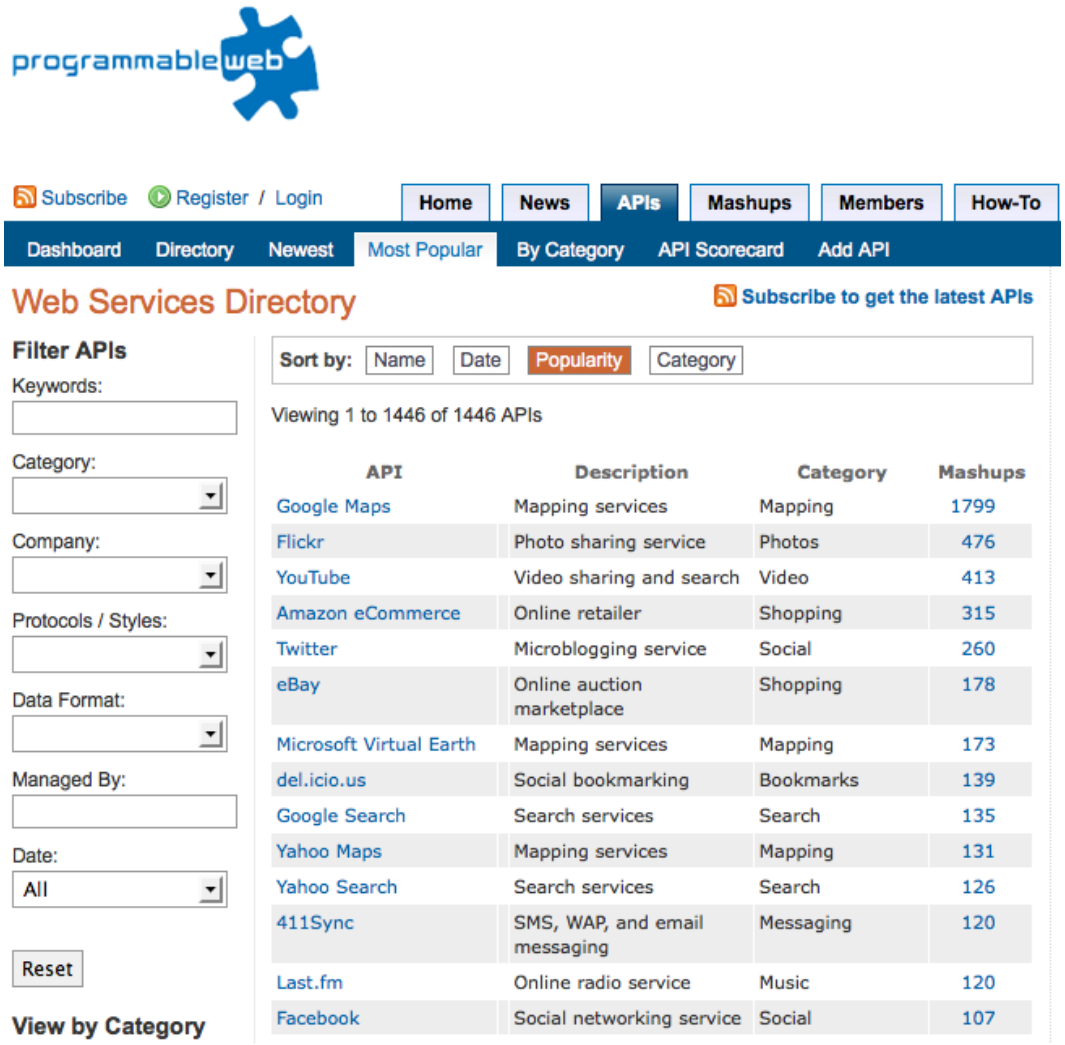
- Supervised
  - Linear regression
  - Classification / regression trees
  - Logistic regression
  - Naive Bayes
  - k-nearest neighbors
  - Support vector machines
  - Boosting
- Unsupervised
  - K-means
  - Mixture models
  - Principal components analysis
  - Factor analysis
  - Topic models
  - Collaborative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

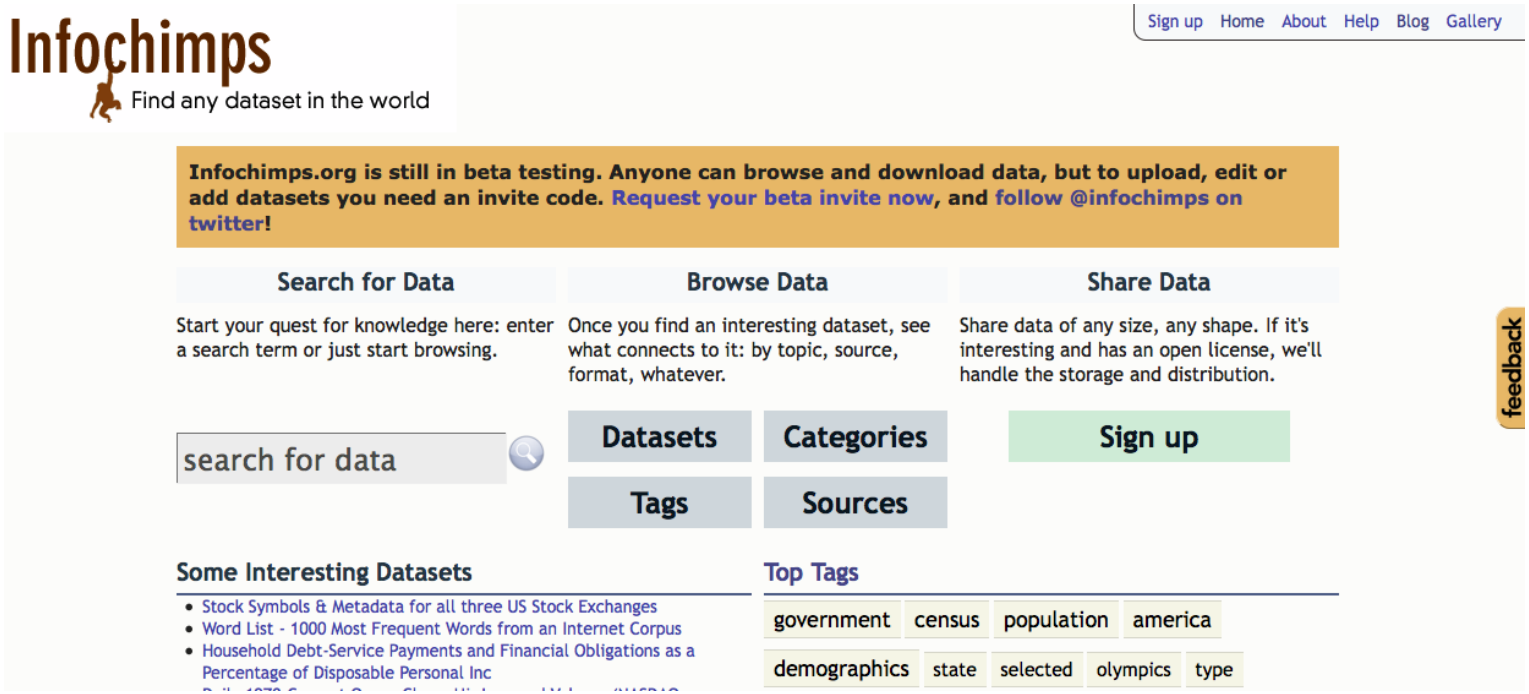
Got data?

- 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
```