

Welcome to Data Bootcamp

Joseph Adler, Drew Conway, Jake Hofman, Hilary Mason

February 1, 2011



Creative Commons Attribution-Share Alike 3.0

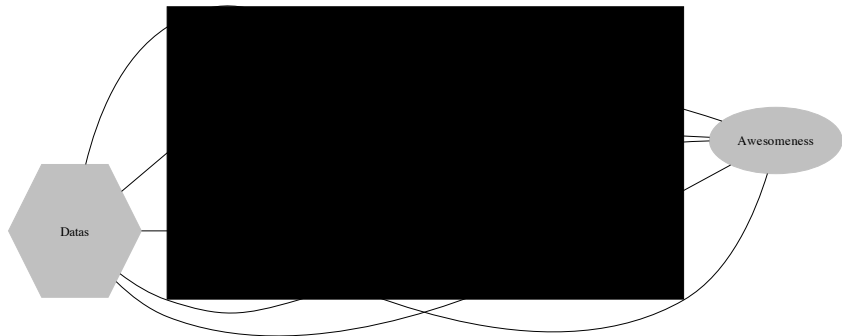
All of the slides, code and images from today's tutorial are available on Github:

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

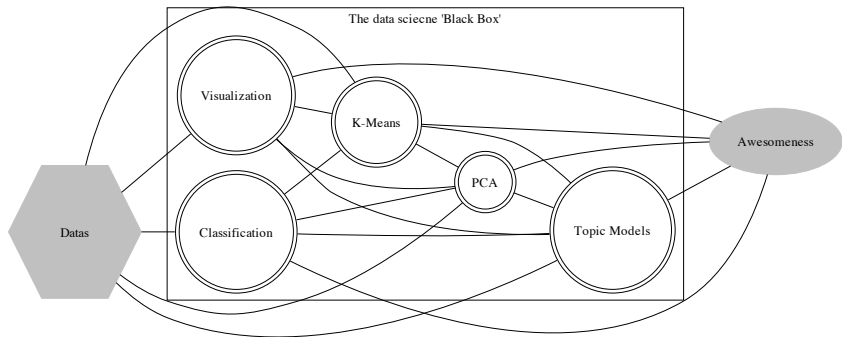
The play the home game

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

The data science “black box”



The data science “black box”



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 Financial Aid for your College Degree - HI ! It h
PHARMA_viagra_PHARMA_cialis - Wanted: web store with remedies. N

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 Financial Aid for your College Degree - HI ! It h
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

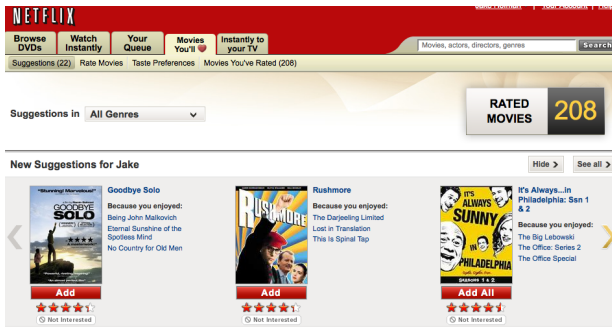




- ▶ We learn quickly from few, relatively unstructured examples ... but we don't understand *how* we accomplish this
- ▶ Can we develop algorithms that enable machines to learn by example from large data sets?

- ▶ 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)

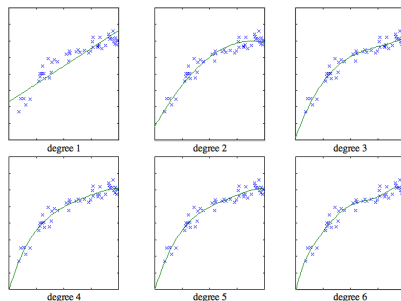
Netflix prize



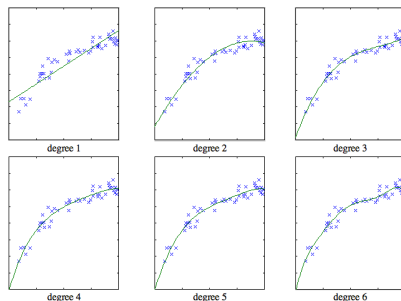
- ▶ \$1M for a 10% improvement in predicted rating
- ▶ More than 1000 submissions over 2.5 years
- ▶ Top two teams within 0.01% of each other (winners announced soon)

- ▶ 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

- ▶ 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



- ▶ 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



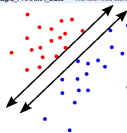
- ▶ How can we quantify an “optimal” model
 - ▶ What to optimize?
 - ▶ How to optimize it?

1. Get data

Fed: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
non urgent - whoopee! yes that's what I meant, thanks for decoding my quest
SourceForge.net: variational bayes for network modularity - can i get admin |
Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery |
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. In
More effective - if you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, Vioxx/ESI - Order all your Favorite Rx-Medica
Financial Aid Available: Find Funding for Your Education - Get the financial i
Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
PHARMA_vagra_PHARMA_cialis - Wanted: web store with remedies. N

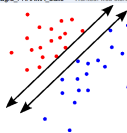
1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data

Fed: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
non urgent - whopel yes that's what I meant, thanks for decoding my quest
SourceForge.net: variational bayes for network modularity - can i get admin |
Ilyline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery |
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. Im
More effective - if you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, Vioxx/ESI - Order all your Favorite Rx-Medica
Financial Aid Available: Find Funding for Your Education - Get the financial i
Find The Perfect School and Financial Aid for your College Degree - Hi I'll h
PHARMA_vagra_PHARMA_cialis - Wanted: web store with remedies. N



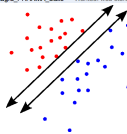
1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data
5. Specify model
6. Specify loss function

Fed: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
non urgent - whopel yes that's what I meant, thanks for decoding my quest
SourceForge.net: variational bayes for network modularity - can i get admin
llyline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery
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. In
More effective - if you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, Vioxx/ESI - Order all your Favorite Rx-Medica
Financial Aid Available: Find Funding for Your Education - Get the financial
Find The Perfect School and Financial Aid for your College Degree - H I I t h
PHARMA_vagra_PHARMA_cialis - Wanted: web store with remedies. N



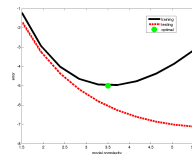
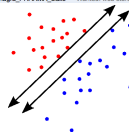
1. Get data
2. Visualize/perform sanity checks
3. Clean/filter observations
4. Choose features to represent data
5. Specify model
6. Specify loss function
7. Develop algorithm to minimize loss

Fed: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
non urgent - whoopee! yes that's what I meant, thanks for decoding my quest
SourceForge.net: variational bayes for network modularity - can i get admin
Ilyline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery
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. Im
More effective - if you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, Vioxx/ESI! - Order all your Favorite Rx-Medica
Financial Aid Available: Find Funding for Your Education - Get the financial
Find The Perfect School and Financial Aid for your College Degree - H I I 8 h
PHARMA_vagra_PHARMA_cialis - Wanted: web store with remedies. N



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

Fed: Yahoo! supercomputing cluster RFP - I have no idea. I have no idea. O
non urgent - whoopee! yes that's what I meant, thanks for decoding my quest!
SourceForge.net: variational bayes for network modularity - can I get admin |
Blythe - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery |
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. Im
More effective - if you are having trouble viewing this email click here. Thurs
Special Offer! Cialis, Viagra, Vioxx/ESI - Order all your Favorite Rx-Medica
Financial Aid Available: Find Funding for Your Education - Get the financial i
Find The Perfect School and Financial Aid for your College Degree - H I I 8 h
PHARMA_vioxx_PHARMA_cialis - Wanted: web store with remedies. N



▶ 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

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

- ▶ 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 (?)

- ▶ Simple approaches often do surprisingly well for large problems

- Web service APIs expose vast amounts of data



Subscribe Register / Login

Home News APIs Mashups Members How-To

Dashboard Directory Newest Most Popular By Category API Scorecard Add API

Web Services Directory

Subscribe to get the latest APIs

Filter APIs

Keywords:

Category:

Company:

Protocols / Styles:

Data Format:

Managed By:

Date:

View by Category

Sort by:

Viewing 1 to 1446 of 1446 APIs

API	Description	Category	Mashups
Google Maps	Mapping services	Mapping	1799
Flickr	Photo sharing service	Photos	476
YouTube	Video sharing and search	Video	413
Amazon eCommerce	Online retailer	Shopping	315
Twitter	Microblogging service	Social	260
eBay	Online auction marketplace	Shopping	178
Microsoft Virtual Earth	Mapping services	Mapping	173
del.icio.us	Social bookmarking	Bookmarks	139
Google Search	Search services	Search	135
Yahoo Maps	Mapping services	Mapping	131
Yahoo Search	Search services	Search	126
411Sync	SMS, WAR, and email messaging	Messaging	120
Last.fm	Online radio service	Music	120
Facebook	Social networking service	Social	107

- ▶ Many free, public data sets available online

Infochimps



Find any dataset in the world

[Sign up](#) [Home](#) [About](#) [Help](#) [Blog](#) [Gallery](#)

Infochimps.org is still in beta testing. Anyone can browse and download data, but to upload, edit or add datasets you need an invite code. [Request your beta invite now](#), and follow [@infochimps](#) on twitter!

Search for Data

Start your quest for knowledge here: enter a search term or just start browsing.



Browse Data

Once you find an interesting dataset, see what connects to it: by topic, source, format, whatever.

Datasets

Tags

Categories

Sources

Share Data

Share data of any size, any shape. If it's interesting and has an open license, we'll handle the storage and distribution.

Sign up

feedback

Some Interesting Datasets

- Stock Symbols & Metadata for all three US Stock Exchanges
- Word List - 1000 Most Frequent Words from an Internet Corpus
- Household Debt-Service Payments and Financial Obligations as a Percentage of Disposable Personal Inc

Top Tags

government census population america
demographics state selected olympics type

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

- ▶ 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
```

- ▶ 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}'
```

- ▶ 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
```