

# Data-driven modeling

## APAM E4990

Jake Hofman

Columbia University

January 23, 2012

# 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
  - ...

# 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 Financial Aid for your College Degree - HI ! It h  
\*\*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 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
- We'd like to develop algorithms that enable machines to learn by example from large data sets

# Got data?

- Web service APIs expose lots of data

 programmableweb

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

Sort by: Name Date Popularity Category

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, WAP, and email messaging	Messaging	120
Last.fm	Online radio service	Music	120
Facebook	Social networking service	Social	107

Filter APIs  
Keywords:  
Category:  
Company:  
Protocols / Styles:  
Data Format:  
Managed By:  
Date:  
All  
Reset  
View by Category

# Got data?

- Many free, public data sets available online

The screenshot shows the Infochimps homepage. At the top, there's a navigation bar with links for Sign up, Home, About, Help, Blog, and Gallery. Below the navigation is the Infochimps logo with the tagline "Find any dataset in the world". A yellow banner at the top of the main content area says: "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!"

The main content area has three main sections: "Search for Data", "Browse Data", and "Share Data".

- Search for Data:** A search bar with placeholder text "search for data" and a magnifying glass icon.
- Browse Data:** Buttons for "Datasets", "Categories", "Tags", and "Sources".
- Share Data:** A green button labeled "Sign up" and a link to "feedback".

Below these sections are two lists of datasets and tags:

- 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 Income
- Top Tags:** A list of tags including government, census, population, america, demographics, state, selected, olympics, and type.

# Black-boxified?

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

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...

# Black-boxified?

The screenshot shows a GitHub repository page for 'JohnLangford/vowpal\_wabbit'. The top navigation bar includes links for 'Explore', 'Gist', 'Blog', and 'Help'. On the right, there's a user profile for 'jhofman' with a star icon indicating 0 stars. Below the header, the repository name 'JohnLangford / vowpal\_wabbit' is displayed, along with a note that it was forked from 'aparker/vowpal\_wabbit'. The main navigation tabs are 'Code', 'Network', 'Pull Requests (0)', 'Wiki (18)', and 'Stats & Graphs'. The 'Wiki' tab is currently selected. A secondary navigation bar below shows 'Home', 'Pages', 'Wiki History', and 'Git Access', with 'Home' being the active tab.

## Home

### Vowpal Wabbit

[New Page](#)[Edit Page](#)[Page History](#)

The [Vowpal Wabbit](#) (VW) project is a fast out-of-core learning system sponsored by [Yahoo Research](#). Support is available through the [mailing list](#).

There are two ways to have a fast learning algorithm: (a) start with a slow algorithm and speed it up, or (b) build an intrinsically fast learning algorithm. This project is about approach (b), and it's reached a state where it may be useful to others as a platform for research and experimentation.

There are several optimization algorithms available with the baseline being sparse gradient descent (GD) on a loss function (several are available). The code should be easily usable. Its only external dependence is on the [boost library](#), which is often installed by default.

# Roadmap?

Step 1: Have data

Step 2: ???

Step 3: Profit

# Roadmap, take two

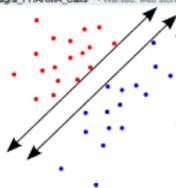
## ① Get data

Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O  
non urgent - whoopie! yes that's what i meant, thanks for decoding my questi  
SourceForge.net: variational bayes for network modularity - can i get admin i  
Byline - iPhone Apps, iPhone 3D apps and iPod touch Applications Gallery i  
Laurenca J. Peter: Facts are stubborn things, but statistics are more pliable.  
Re: JAFOIS 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, VicerolESI - Order all your Favorite Rx-Media  
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 th  
\*\*PHARMA\_viagra\_PHARMA\_calls\*\* - Wanted: web store with remedies. N

# Roadmap, take two

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

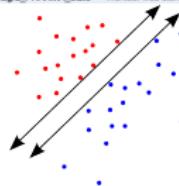
Fwd: Yahoo! supercomputing cluster RFP - i have no idea. i have no idea. O non urgent - whoopie! yes that's what i meant, thanks for decoding my questi  
SourceForge.net: variational bayes for network modularity - can i get admin i Byline - iPhone Apps, iPhone 3D apps and iPod touch Applications Gallery i Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.  
Re: JAFOIS 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, VicksinEST - Order all your Favorite Rx-Media 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 th  
\*\*PHARMA\_vigra\_PHARMA\_calls\*\* - Wanted: web store with remedies. N



# Roadmap, take two

- 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

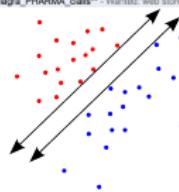
Fwd: Yahoo supercomputing cluster RFP - i have no idea. i have no idea. O non urgent - whoopie! yes that's what i meant, thanks for decoding my quasi SourceForge.net: variational bayes for network modularity - can i get admin | Byline - iPhone Apps, iPhone 3D apps and iPod touch Applications Gallery | Laurence J. Peter: Facts are stubborn things, but statistics are more pliable. Re: JAFOIS 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, VicksinEST - Order all your Favorite Rx-Media 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 th \*\*PHARMA\_vigra\_PHARMA\_calls\*\* - Wanted: web store with remedies. N



# Roadmap, take two

- 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

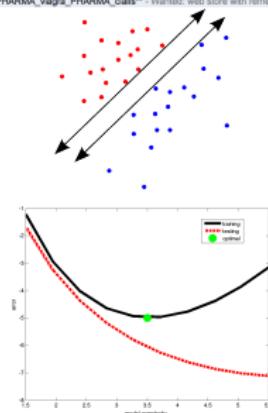
Fwd: Yahoo supercomputing cluster RFP - i have no idea. i have no idea. O non urgent - whoopie! yes that's what i meant, thanks for decoding my quasi SourceForge.net: variational bayes for network modularity - can i get admin i Byline - iPhone Apps, iPhone 3G apps and iPod touch Applications Gallery i Laurence J. Peter: Facts are stubborn things, but statistics are more pliable. Re: JAFOIS 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, VicksDEESI - Order all your Favorite Rx-Media 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 th \*\*PHARMA\_vietnam\_PHARMA\_calls\*\* - Wanted: web store with remedies. N



# Roadmap, take two

- 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

Fwd: Yahoo supercomputing cluster RFP - i have no idea. I have no idea. O non urgent - whoopie yes that's what I meant, thanks for decoding my questi  
SourceForge.net: variational bayes for network modularity - can i get admin i Byline - iPhone Apps, iPhone 3D apps and iPod touch Applications Gallery i Laurence J. Peter: Facts are stubborn things, but statistics are more pliable.  
Re: JAFOIS 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, VicksinEST - 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 th  
\*\*PHARMA\_viagra\_PHARMA\_calls\*\* - Wanted: web store with remedies.



# Topics

- Supervised
  - k-nearest neighbors
  - Naive Bayes
  - Linear regression
  - Logistic regression
  - Support vector machines
  - Collaborative filtering
  - Matrix factorization
- Unsupervised
  - K-means
  - Mixture models
  - Principal components analysis
  - Topic models
- Data representation: feature space, selection, normalization
- Model assessment: complexity control, cross-validation, ROC curve, Bayesian Occam's razor
- Large-scale learning

# Everything old is new again<sup>1</sup>

- 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

---

<sup>1</sup><http://cbcl.mit.edu/publications/theses/thesis-rifkin.pdf>



# Statistics vs. machine learning<sup>2</sup>

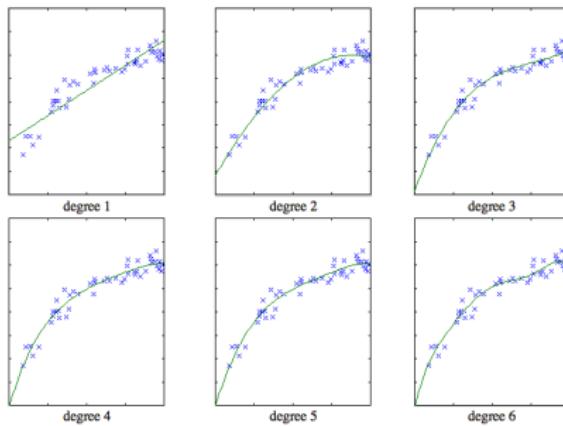
## Glossary

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

<sup>2</sup>[http:](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



# Example: Netflix Prize



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

# Example: Netflix Prize

The screenshot shows the official Netflix Prize website. At the top, the Netflix logo is visible. Below it, a large yellow banner features the text "Netflix Prize" and a red "COMPLETED" stamp. A navigation bar below the banner includes links for "Home", "Rules", "Leaderboard", and "Update". The main content area is titled "Frequently Asked Questions" in large blue text. Under this title, a question "How does Cinematch do it?" is listed, followed by a detailed answer about the Cinematch algorithm's complexity and the real-world system's requirements.

## FAQ

### How does Cinematch do it?

Straightforward statistical linear models with a lot of data conditioning. But a real-world system is much more than an algorithm, and Cinematch does a lot more than just optimize for RMSE. After all, we have a website to support. In production we have to worry about system scaling and performance, and we have additional sources of data we can use to guide our recommendations. But, as mentioned in the [Rules](#) and just to be perfectly clear, for the purposes of the Prize the RMSE values we report here do not use any of this extra data.

# Shipping = Feature

Add an asymmetric frequency feature  $\mathbf{y}_{j,f_{ut}}^{(3)}$ : **SBRMF-UTB-UTF-MTF-ATF-MFF-AFF**

$$\widehat{r_{uit}} = \mu_i + \mu_u + \mu_{u,t} + \mu_{i,\text{bin}(t)} + \left( \mathbf{p}_i^{(1)} + \mathbf{p}_{i,\text{bin}(t)}^{(2)} + \mathbf{p}_{i,f_{ut}}^{(3)} \right)^T \left( \mathbf{q}_{u,t}^{(1)} + \mathbf{q}_{u,t}^{(2)} + \frac{1}{\sqrt{|N(u)|}} \sum_{j \in N(u)} \left( \mathbf{y}_j^{(1)} + \mathbf{y}_{j,\text{bin}(t)}^{(2)} + \mathbf{y}_{j,f_{ut}}^{(3)} \right) \right) \quad (34)$$

Model extension (+)	epoch time	#epochs	probeRMSE, $k = 50$ features
<b>SBRMF</b> - SVD with biases	17[s]	69	0.9054
<b>SBRMF</b> - asymmetric part	50[s]	30	0.8974
+ <b>UTB</b> - user time bias	61[s]	50	0.8919
+ <b>UTF</b> - user time feature	62[s]	38	0.8911
+ <b>MTF</b> - movie time feature	74[s]	37	0.8908
+ <b>ATF</b> - asymmetric time feature	74[s]	44	0.8905
+ <b>MFF</b> - movie frequency feature	149[s]	46	0.8900
+ <b>AFF</b> - asymmetric frequency feature	206[s]	45	0.8886 (0.8846 with $k = 1000$ )

# References



# Disclaimer

You may be bored if you already know how to ...

- Acquire data from APIs
- Clean/explore/visualize data
- Classify and cluster various types of data (e.g., images, text)
- Code in Python, R, SciPy/NumPy, etc.
- Scale solutions to large data sets (e.g. Hadoop, SGD)
- Script with unix tools on the command line, e.g.

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

# Themes

## Data jeopardy

Regardless of scale, it's difficult to find the right questions to ask  
of the data

# Themes

## Data hacking

Cleaning and normalizing data is a substantial amount of the work  
(and likely impacts results)

# Themes

## Data hacking

The ability to iterate quickly, asking and answering many questions, is crucial

# Themes

## Data hacking

Hacks happen: sed/awk/grep are useful, and scale

# Themes

“Data science”

Simple methods (e.g., linear models) work surprisingly well,  
especially with lots of data

# Themes

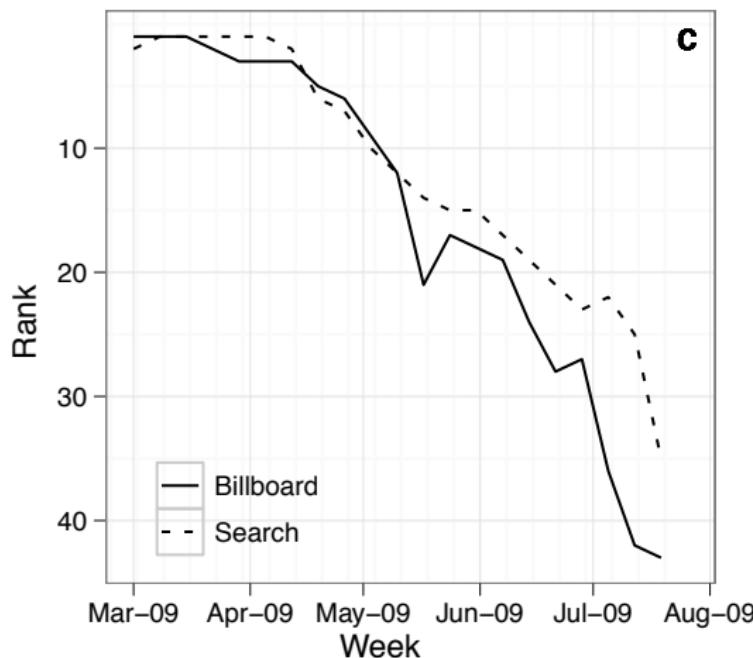
“Data science”

It's easy to cover your tracks—things are often much more complicated than they appear



# Predicting consumer activity with Web search

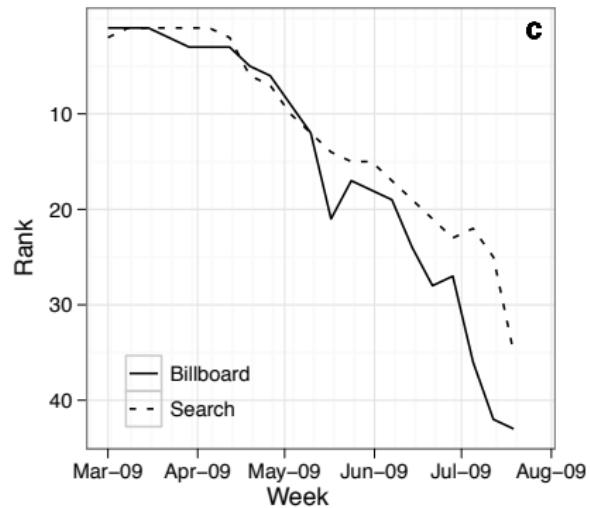
with Sharad Goel, Sébastien Lahaie, David Pennock, Duncan Watts



# Search predictions

## Motivation

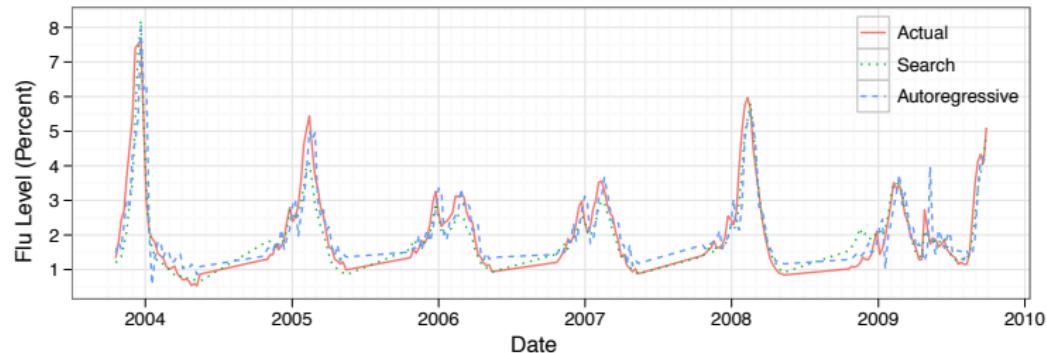
Does collective search activity provide useful predictive signal about real-world outcomes?



# Search predictions

## Motivation

Past work mainly focuses on predicting the present<sup>1</sup> and ignores baseline models trained on publicly available data

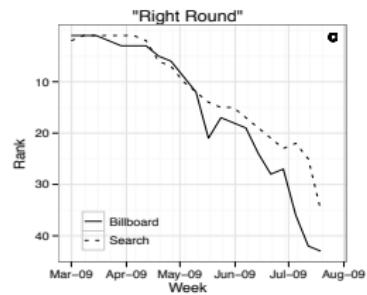
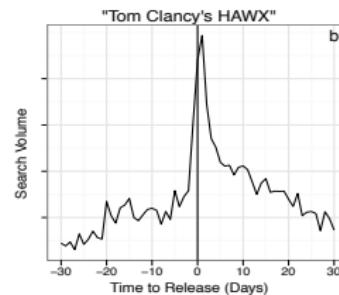
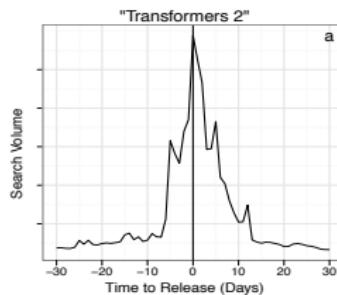


<sup>1</sup>Varian, 2009

# Search predictions

## Motivation

We predict future sales for movies, video games, and music



# Search predictions

## Search models

For movies and video games, predict opening weekend box office and first month sales, respectively:

$$\log(\text{revenue}) = \beta_0 + \beta_1 \log(\text{search}) + \epsilon$$

For music, predict following week's Billboard Hot 100 rank:

$$\text{billboard}_{t+1} = \beta_0 + \beta_1 \text{search}_t + \beta_2 \text{search}_{t-1} + \epsilon$$

# Search predictions

## Search volume

YAHOO!®

Web Images Video Local Shopping News More ▾

no country

Search Options ▾

QuickApps

SafeSearch - On

509,000,000 results for no country:

Show All

W Wikipedia

IMDb

MySpace

NY Daily News

GameSpot

Sponsored Results

Also try: [no country for old men](#), [no country for old men ending](#), [more...](#)

**No Country for Old Men (film) - Wikipedia, the ...**  
[Plot](#) | [Cast and characters](#) | [Themes and style](#) | [Production](#)  
No Country for Old Men is a 2007 American crime thriller directed by Joel Coen and Ethan Coen, and starring Tommy Lee Jones, Javier Bardem, and Josh Brolin. The film was adapted from...  
[en.wikipedia.org/wiki/No\\_Country\\_for\\_Old\\_Men\\_\(film\)](http://en.wikipedia.org/wiki/No_Country_for_Old_Men_(film)) - [Cached](#)

**No Country for Old Men (2007) - IMDb**  
Violence and mayhem ensue after a hunter stumbles upon some dead bodies, a stash of heroin and more than \$2 million in cash near the Rio Grande. With Tommy Lee Jones ...  
[www.imdb.com/title/tt0477348](http://www.imdb.com/title/tt0477348) - [Cached](#)

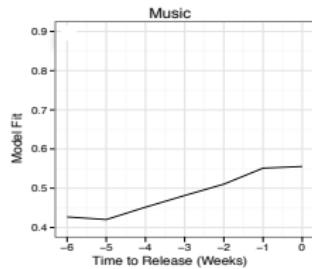
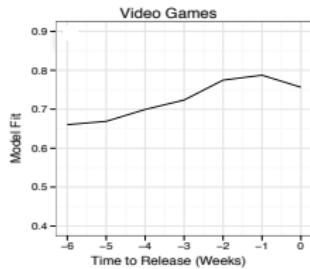
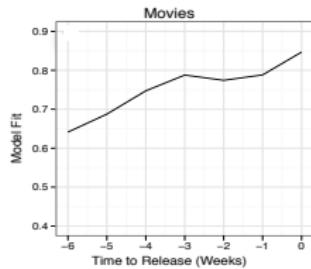
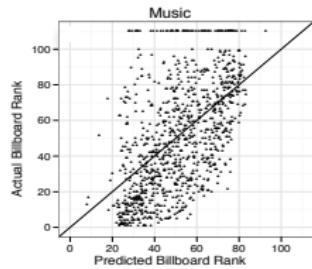
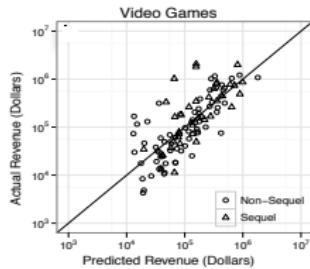
**no country | Free Music, Tour Dates, Photos, Videos**  
no country's official profile including the latest music, albums, songs, music videos and more updates.  
[www.myspace.com/nocountrytheband](http://www.myspace.com/nocountrytheband) - [Cached](#)

**No Country - Video Results**

# Search predictions

## Search models

Search activity is **predictive** for movies, video games, and music weeks to months in advance



# Search predictions

## Baseline models

For movies, use **budget**, number of **opening screens** and **Hollywood Stock Exchange**:

$$\log(\text{revenue}) = \beta_0 + \beta_1 \log(\text{budget}) + \beta_2 \log(\text{screens}) + \beta_3 \log(\text{hsx}) + \epsilon$$

# Search predictions

## Baseline models

For video games, use critic ratings and predecessor sales (sequels only):

$$\log(\text{revenue}) = \beta_0 + \beta_1 \text{rating} + \beta_2 \log(\text{predecessor}) + \epsilon$$

# Search predictions

## Baseline models

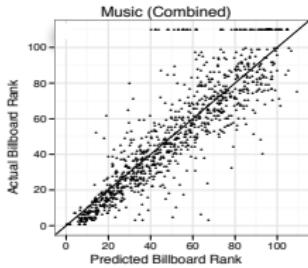
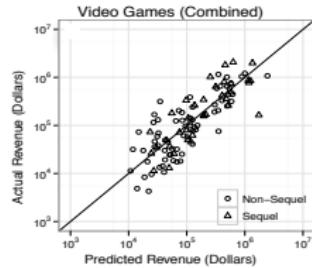
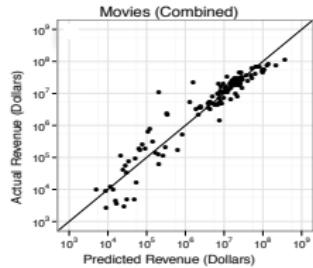
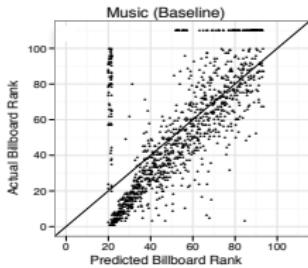
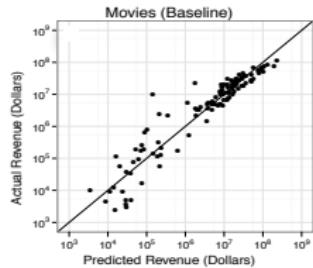
For **music**, use an autoregressive model with the previously available rank:

$$\text{billboard}_{t+1} = \beta_0 + \beta_1 \text{billboard}_{t-1} + \epsilon$$

# Search predictions

Baseline + combined models

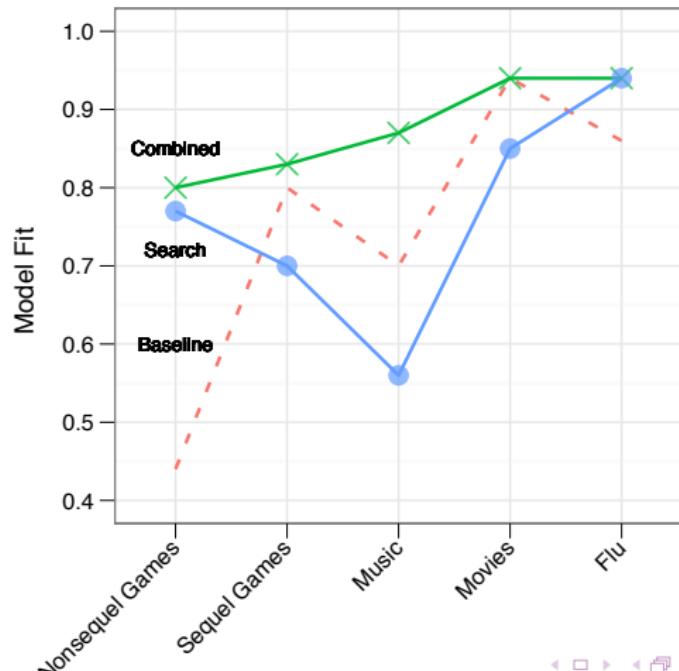
Baseline models are often surprisingly good



# Search predictions

## Model comparison

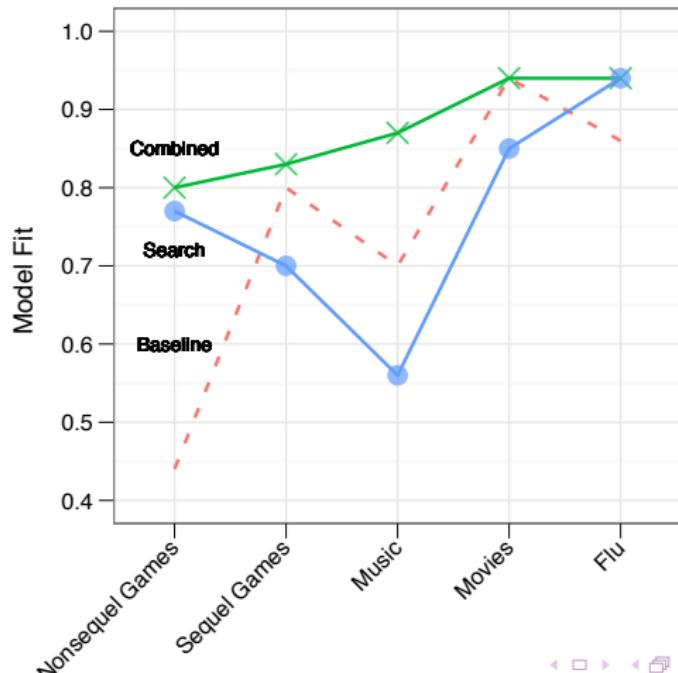
For movies, search is outperformed by the baseline and of little marginal value



# Search predictions

## Model comparison

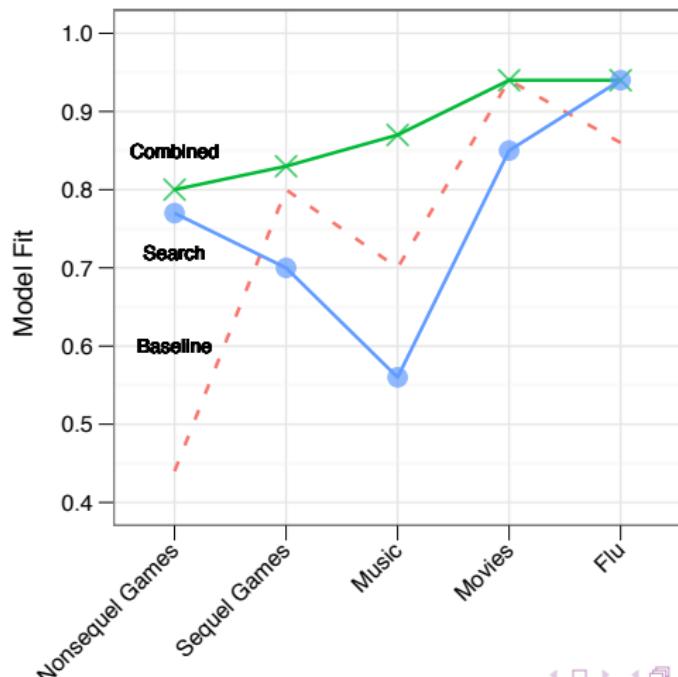
For video games, search helps substantially for non-sequels, less so for sequels



# Search predictions

## Model comparison

For music, the addition of search yields a substantially better combined model



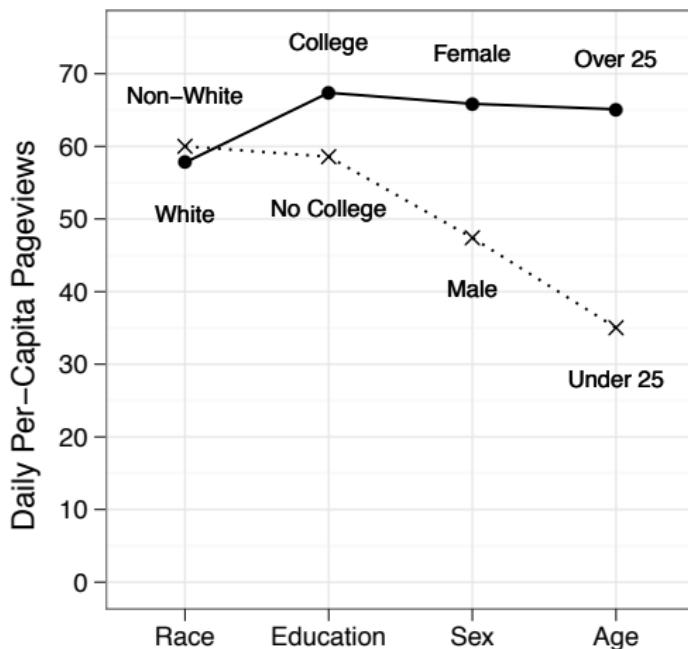
# Search predictions

## Summary

- Relative performance and **value** of search varies across domains
- Search provides a **fast**, **convenient**, and **flexible signal** across domains
- “Predicting consumer activity with Web search”  
Goel, Hofman, Lahaie, Pennock & Watts, PNAS 2010

# Demographic diversity on the Web

with Irmak Sirer and Sharad Goel



# Motivation

Science 17 April 1998:  
Vol. 280 no. 5362 pp. 390-391  
DOI: 10.1126/science.280.5362.390

< Prev | Table of Contents | Next >

POLICY

INFORMATION ACCESS

## Bridging the Racial Divide on the Internet

Donna L. Hoffman and Thomas P. Novak

 Author Affiliations

The Internet is expected to do no less than transform society (1); its use has been increasing exponentially since 1994 (2). But are all members of our society equally likely to have access to the Internet and thus participate in the rewards of this transformation? Here we present findings both obvious and surprising from a recent survey of Internet access and discuss their implications for social science research and public policy.

Previous work is largely **survey-based** and focuses and group-level differences in online **access**

# Motivation

*"As of January 1997, we estimate that 5.2 million African Americans and 40.8 million whites have ever used the Web, and that 1.4 million African Americans and 20.3 million whites used the Web in the past week."*

-Hoffman & Novak (1998)

# Motivation

Focus on activity instead of access



How diverse is the Web?

To what extent do online experiences vary across demographic groups?

## nielsen MegaPanel

- Representative sample of **265,000 individuals** in the US, paid via the Nielsen MegaPanel<sup>2</sup>
- Log of **anonymized, complete browsing activity** from June 2009 through May 2010 (URLs viewed, timestamps, etc.)
- Detailed individual and household **demographic information** (age, education, income, race, sex, etc.)

---

<sup>2</sup>Special thanks to Mainak Mazumdar

# Data

```
# ls -alh nielsen_megapanel.tar  
-rw-r--r-- 100G Jul 17 13:00 nielsen_megapanel.tar
```

# Data

- Transform all demographic attributes to binary variables  
e.g., Age → Over/Under 25, Race → White/Non-White,  
Sex → Female/Male

# Data

- **Transform** all demographic attributes to **binary variables**  
e.g., Age → Over/Under 25, Race → White/Non-White,  
Sex → Female/Male
- **Normalize** pageviews to at most **three domain levels**, sans www  
e.g. `www.yahoo.com` → `yahoo.com`,  
`us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`

# Data

- **Transform** all demographic attributes to **binary variables**  
e.g., Age → Over/Under 25, Race → White/Non-White,  
Sex → Female/Male
- **Normalize** pageviews to at most **three domain levels**, sans www  
e.g. `www.yahoo.com` → `yahoo.com`,  
`us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- **Restrict** to top 100k (out of 9M+ total) **most popular** sites  
(by unique visitors)

# Data

- **Transform** all demographic attributes to **binary variables**  
e.g., Age → Over/Under 25, Race → White/Non-White,  
Sex → Female/Male
- **Normalize** pageviews to at most **three domain levels**, sans www  
e.g. `www.yahoo.com` → `yahoo.com`,  
`us.mg2.mail.yahoo.com/neo/launch` → `mail.yahoo.com`
- **Restrict** to top 100k (out of 9M+ total) **most popular** sites  
(by unique visitors)
- **Aggregate** activity at the **site**, **group**, and **user** levels

# Hadoop + Pig (+ awk)

100GB → ~1GB

```
-- define streaming command for normalizing urls
DEFINE add_star_domains `awk 'n=split($2,a,"."); (n==3) {print $1"\t*."a[2]".a[3]"'\t"$3}'`;

-- flatten user histories
user_pageviews = FOREACH users GENERATE
    uid,
    FLATTEN(history) AS (did, pageviews, weight);

-- join user pageviews against top domains
user_pageviews = JOIN user_pageviews BY did, top_domains BY did USING 'replicated';
user_pageviews = FOREACH user_pageviews GENERATE
    user_pageviews::uid AS uid,
    top_domains::domain AS domain,
    user_pageviews::pageviews AS pageviews;

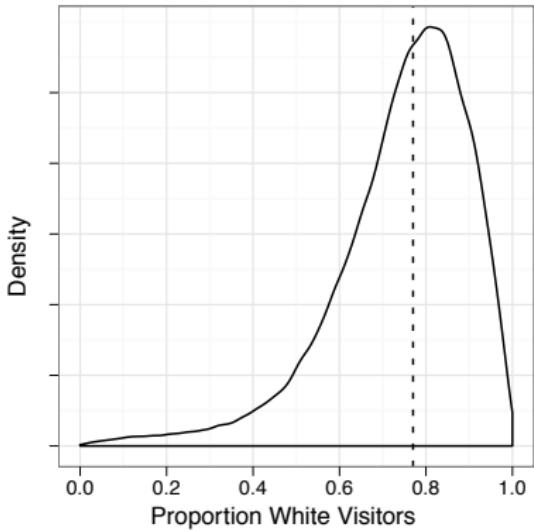
-- stream through awk to extract two-level domains (e.g. *.yahoo.com)
user_pageviews = STREAM user_pageviews THROUGH add_star_domains AS (uid:long, domain:chararray, pageviews:int);

-- regroup and count pageviews by normalized domains
-- (userid, normalized_domain, num_pageviews)
user_pageviews = GROUP user_pageviews BY (uid, domain) PARALLEL 10;
user_pageviews = FOREACH user_pageviews GENERATE
    group.uid AS uid,
    group.domain AS domain,
    SUM(user_pageviews.pageviews) AS pageviews;
```

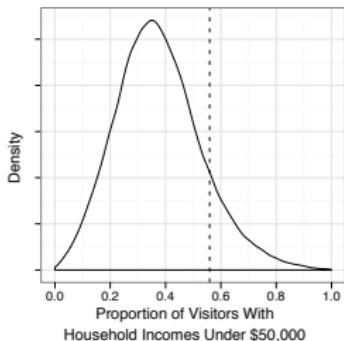
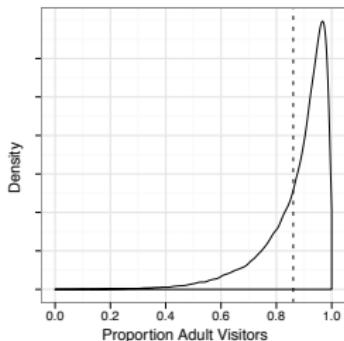
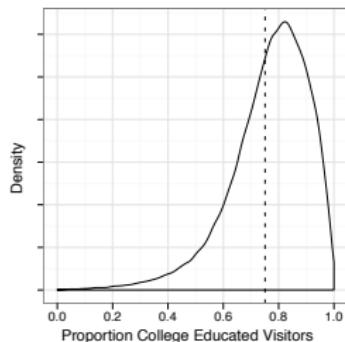
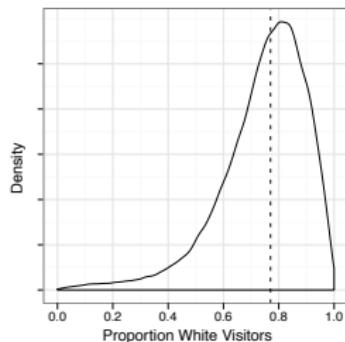
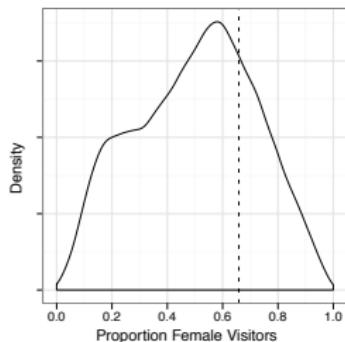
# Site-level skew

How diverse are site audiences?

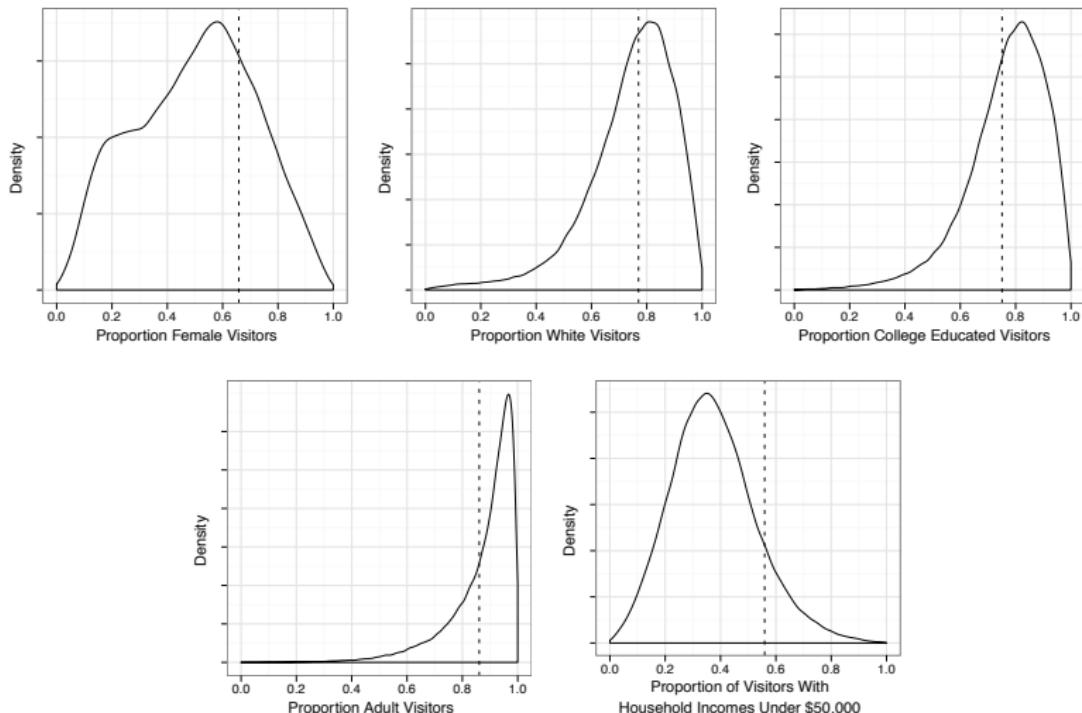
- For each **site** and **attribute**, calculate the **skew** in visitors (e.g., 93% of pageviews on foxnews.com are by White users)
- For each attribute, plot the **distribution of visitor skew** across all sites



# Site-level skew



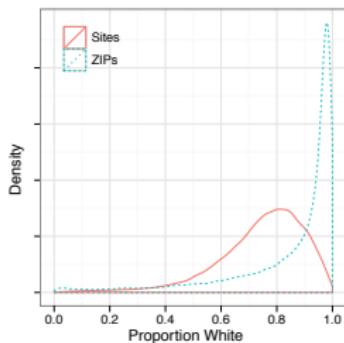
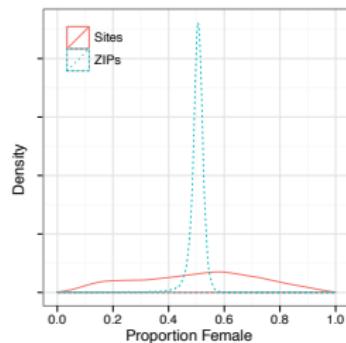
# Site-level skew



Many sites have skew close the overall mean, but there also  
popular, highly-skewed sites

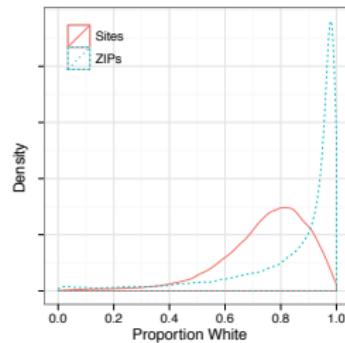
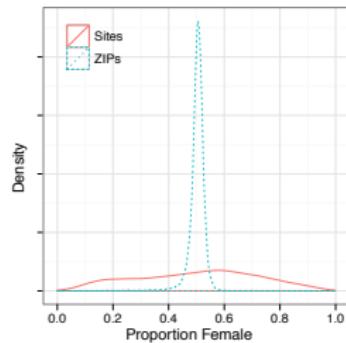
# Sites vs. ZIPs

How do diversity of the online and offline worlds compare?



# Sites vs. ZIPs

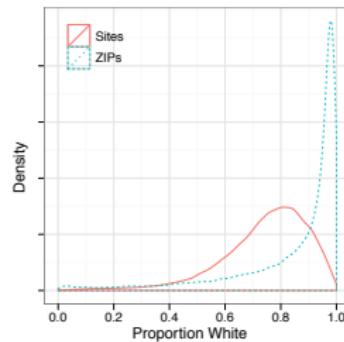
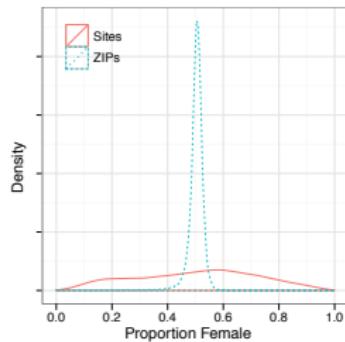
How do diversity of the online and offline worlds compare?



As expected, neighborhoods are more gender-balanced than sites

# Sites vs. ZIPs

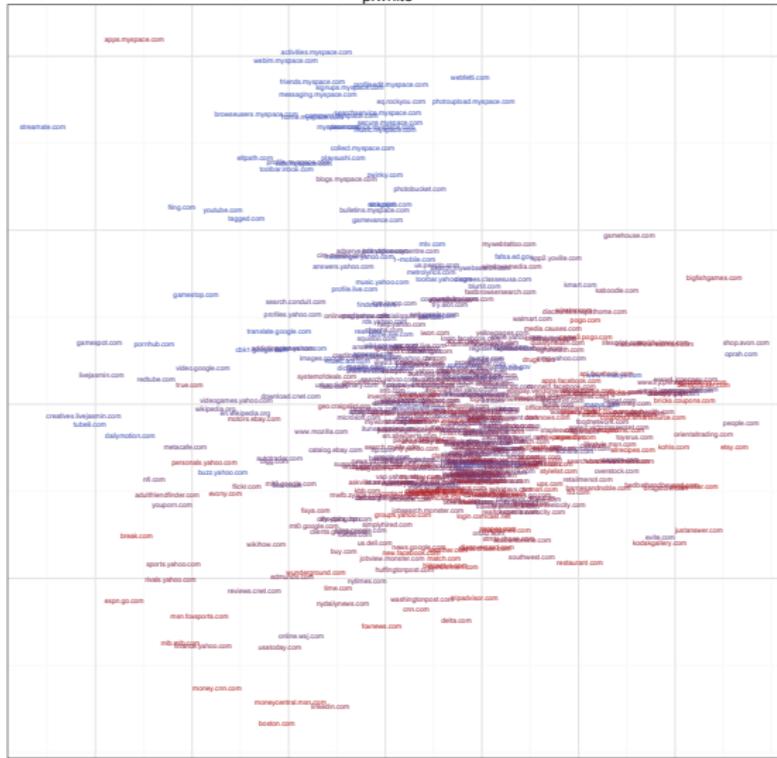
How do diversity of the online and offline worlds compare?



But sites typically have more racially diverse audiences than neighborhoods have residents

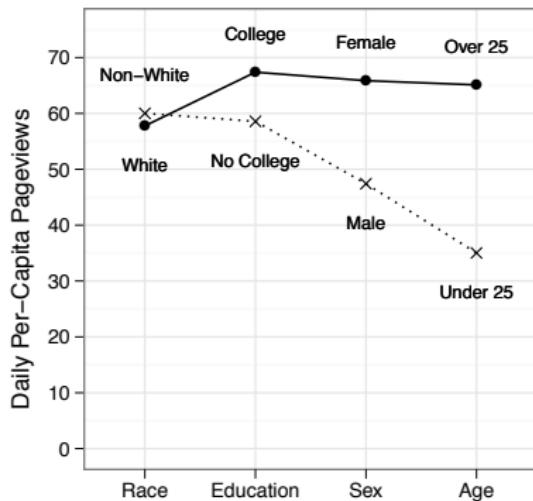
## Site-level skew

p.white



# Group-level activity

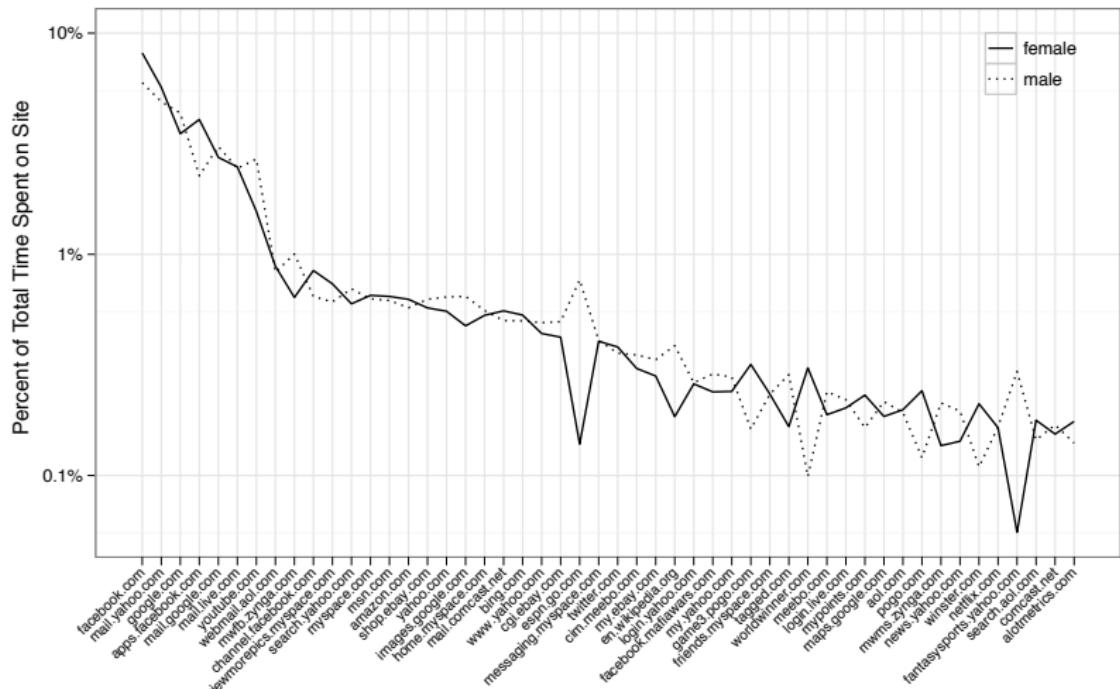
How does browsing activity vary at the group level?



Large differences exist even at the aggregate level  
(e.g. women on average generate 40% more pageviews than men)

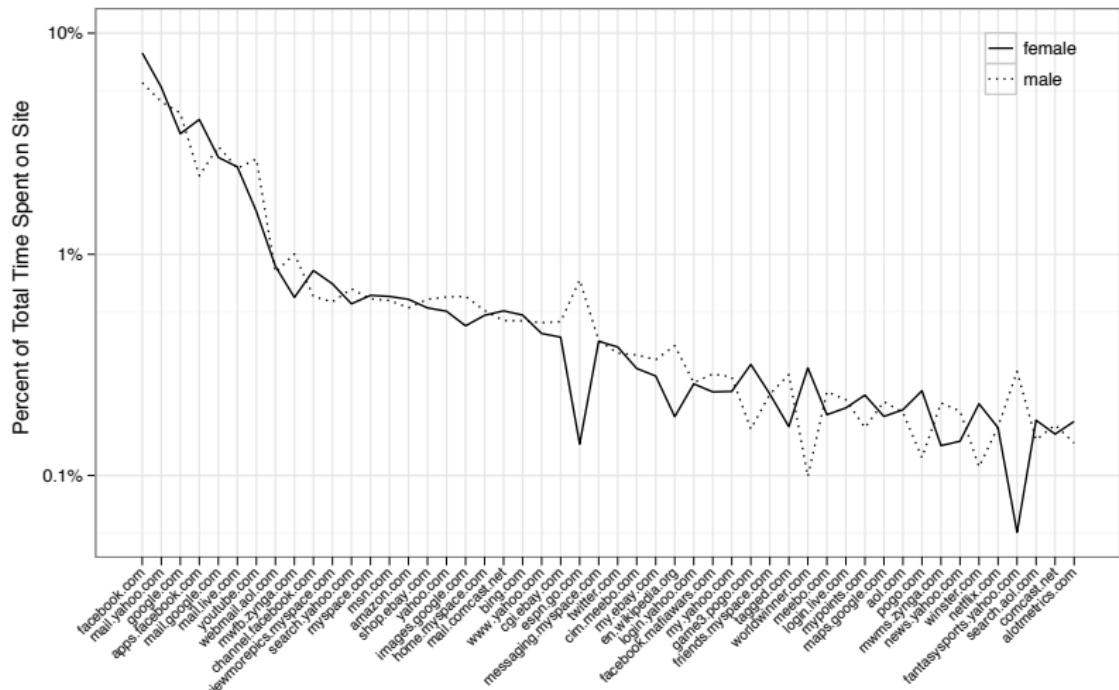
# Group-level activity

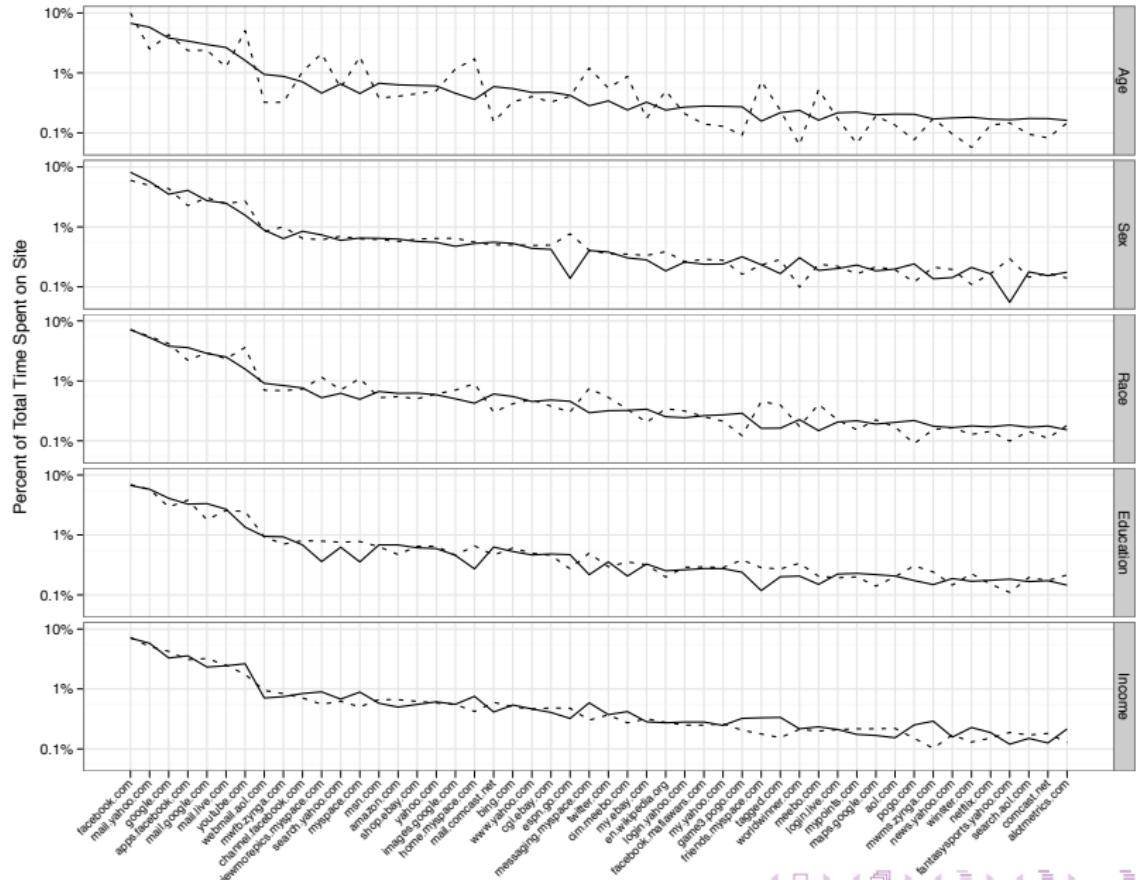
All groups spend more than a third of their time on a handful of email, search, and social networking sites



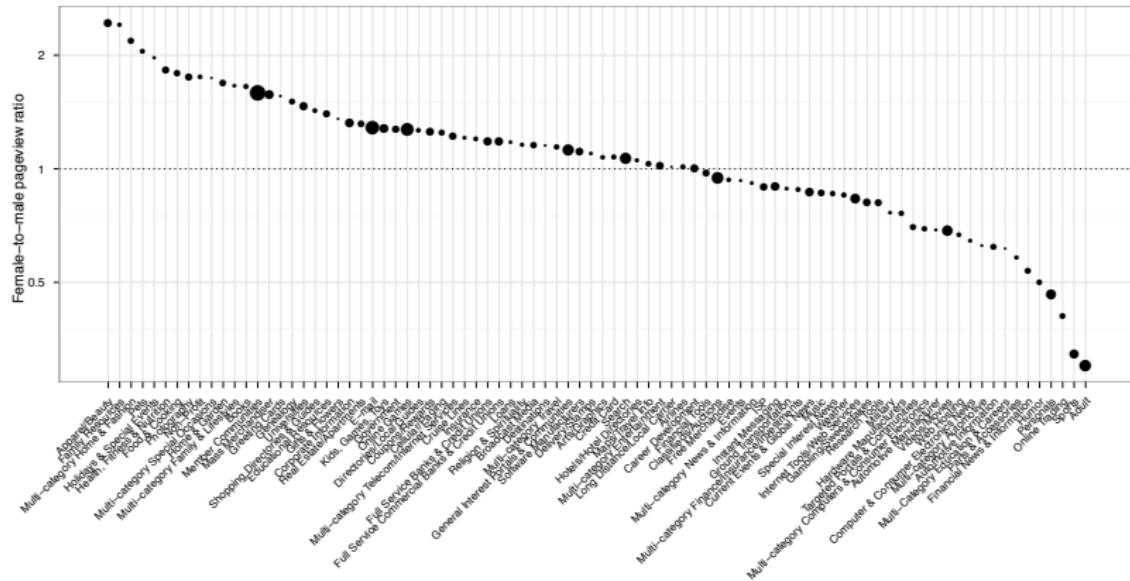
# Group-level activity

But different groups **distribute their time differently**, both on universally popular and on more niche sites



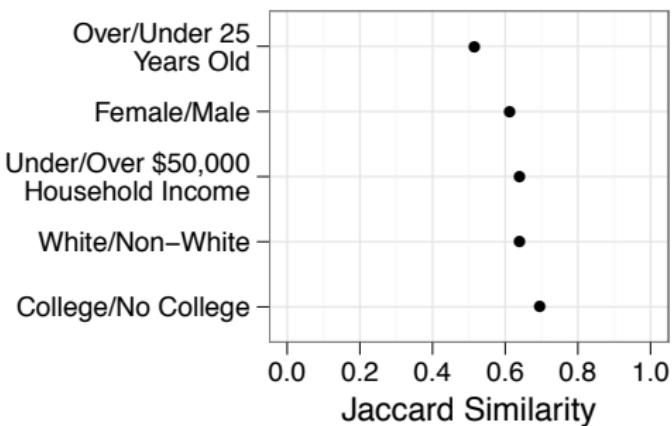


# Group-level activity



# Group-level activity

There is both **reasonable overlap** and **variation** amongst the most popular sites within groups



# Individual-level prediction

How well can one predict an individual's demographics from their browsing activity?

- Represent each user by the set of sites visited
- Fit linear models to predict majority/minority for each attribute on 80% of users
- Tune model parameters using a 10% validation set
- Evaluate final performance on held-out 10% test set

# Individual-level prediction

```
model <- glm(is.female ~ ., data=nielsen, family=binomial)
```

???

# Individual-level prediction

$$\hat{y}(x_i) = w \cdot x_i + b$$

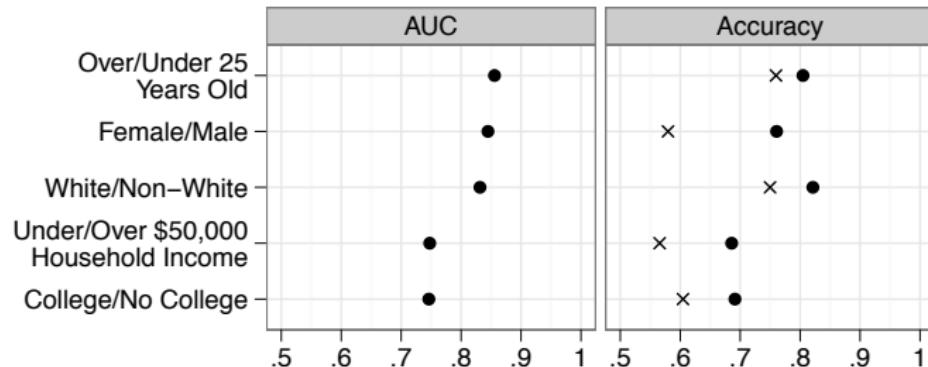
Support Vector Machine<sup>3</sup> :  $L(y, \hat{y}) = C \sum_i [1 - y_i \hat{y}(x_i)]_+ + \|w\|^2$

---

<sup>3</sup><http://bit.ly/svmpref>

# Individual-level prediction

Reasonable (~70-85%) accuracy and AUC across all attributes



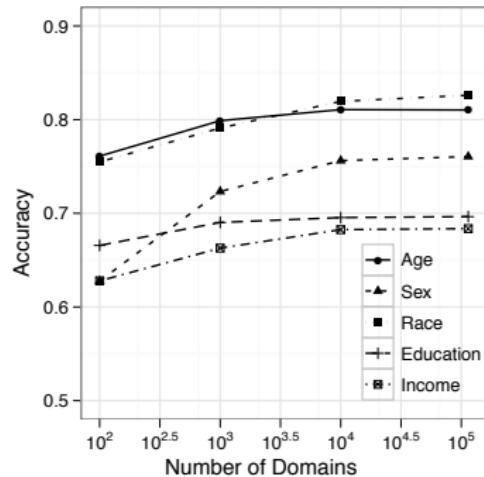
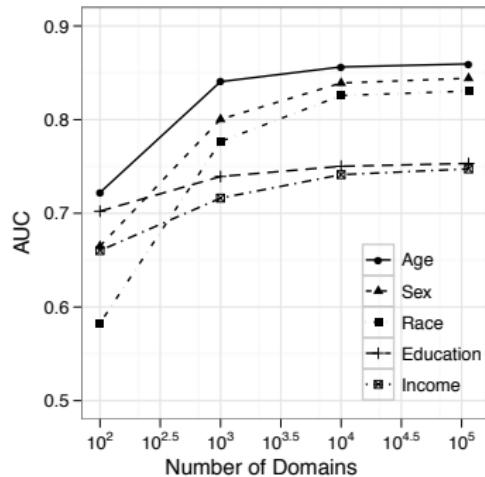
# Individual-level prediction

## Highly-weighted sites under the fitted models

	Large positive weight	Large negative weight
Female	winsters.com lancome-usa.com	sports.yahoo.com espn.go.com
White	marlboro.com cmt.com	mediatakeout.com bet.com
College Educated	news.yahoo.com linkedin.com	youtube.com myspace.com
Over 25 Years Old	evite.com classmates.com	addictinggames.com youtube.com
Household Income Under \$50,000	eharmony.com tracfone.com	rownine.com matrixdirect.com

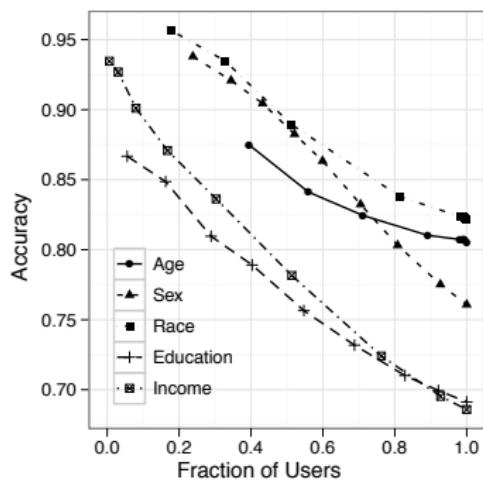
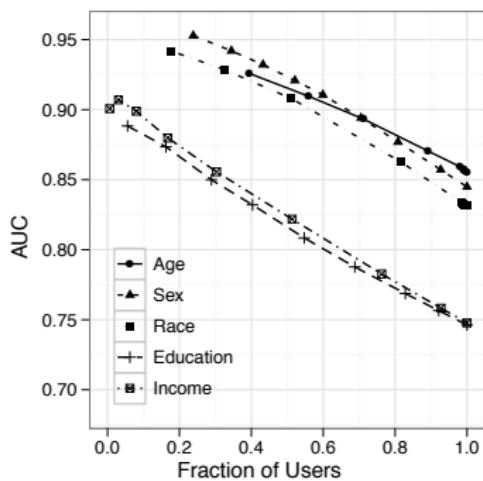
# Individual-level prediction

Similar performance even when restricted to top 1k sites



# Individual-level prediction

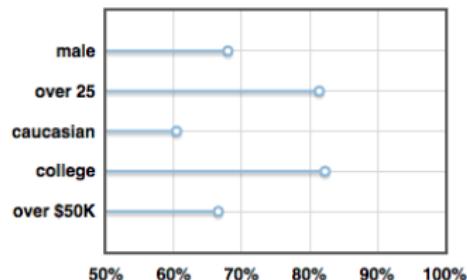
Substantially better performance when restricted to “stereotypical” users ( $\sim 80\text{-}90\%$ )



# Individual-level prediction

## Proof of concept browser demo

From the 28 sites we found in your browser history, it appears that you're a **caucasian male** who is **over 25** years old with a **college** education earning **over \$50K** per year.



<http://bit.ly/surfpreds>

# Summary

- Notable differences in **online** and **offline** diversity
- Large **group-level** differences in how users **distribute** their time
- User demographics can be **inferred** from **browsing activity** with reasonable accuracy