We Don’t Care Who You Are - We Care Who You Are Right Now

Jon Roberts, SVP Data Science and Audience Dev, about.com

The Promise:

If we collect all the data, we can provide the most relevant recommendations *for you*.

The Fears:

- “Google and Facebook know everything about me”
- “Amazon follows me around the internet”
- “Ad networks are taking my data and selling it on around the web”

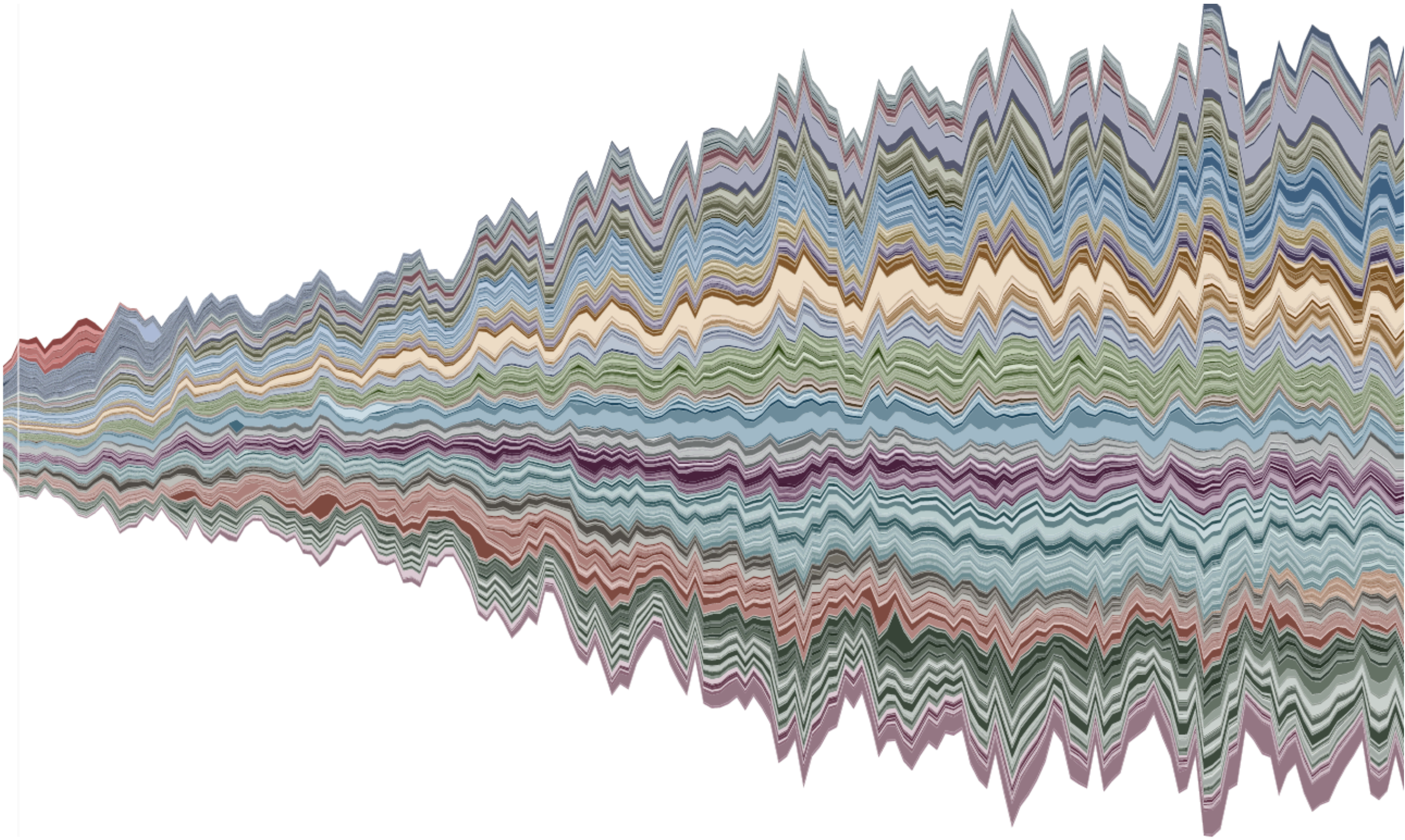
The Utopian Outline

- Record everything you look at, click, read, on what device, where you were at the time, what networks you use (Fb, Google, Slack)
- Infer:
  - age
  - gender
  - household income
  - interest profile
- Use this to provide more relevant recommendations

This raises two questions

- Can you infer demographics from internet behavior?
- Are your demographic features predictive of what you want to see?
- We can test! (and did)

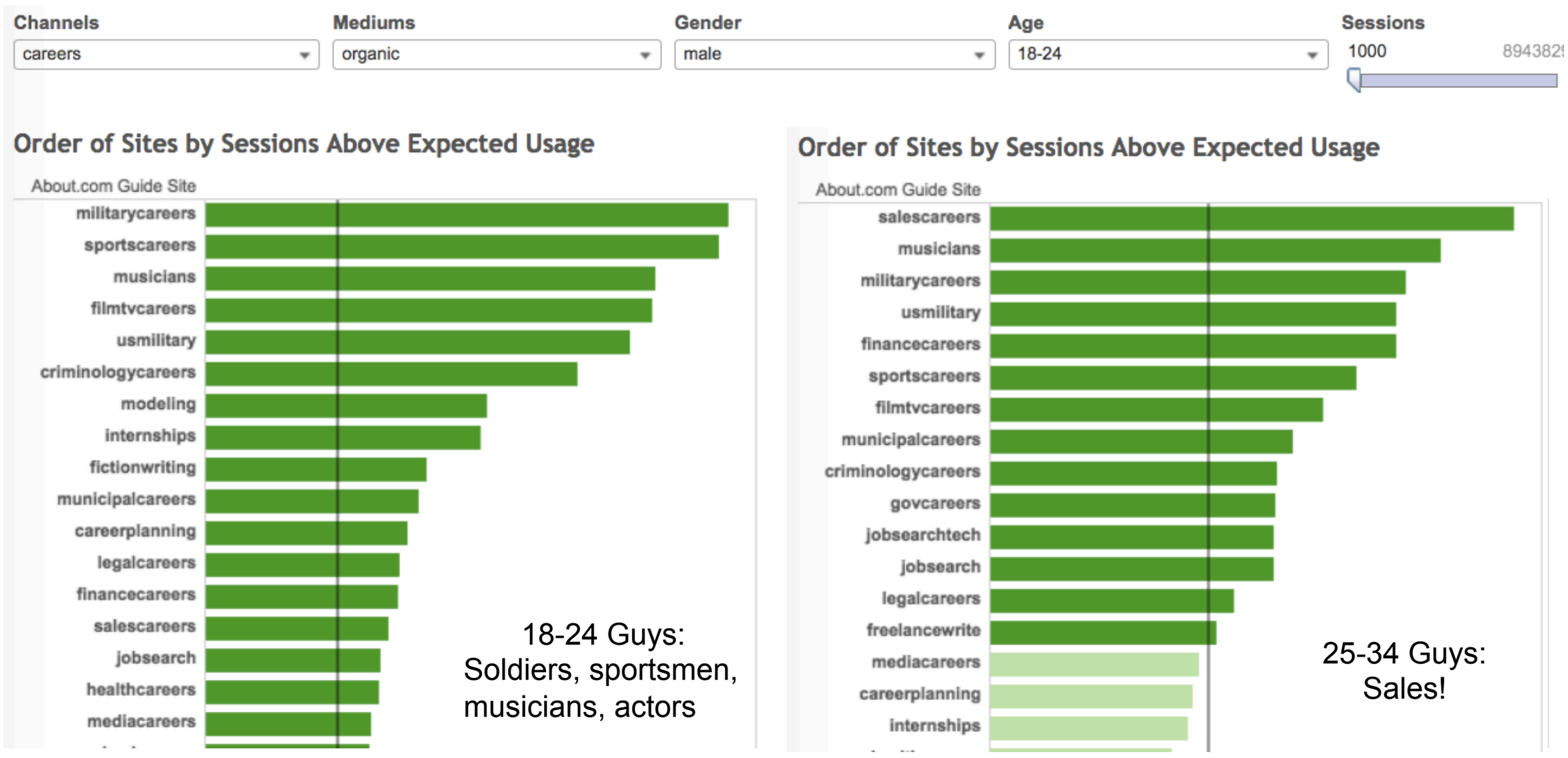
My Lab: About



Can we use the content you read to infer your demographics?

- Take >1M documents with evergreen interest
- Take demographic data from Google and Facebook
  - (cross-check - they both have 1st party demo data, and they agree pretty well)
- Correlate topic interest with demographics
- Example: What interests over-index for millennials?

Example: Career Interest by Age



18-24 Guys:
Soldiers, sportsmen, musicians, actors

25-34 Guys:
Sales!

Millennial women are 3x more interested in going to Paris than non-millennial women

Millennial women are 3x more interested in going to Paris than non-millennial women

Millennial men are just as un-interested in going to Paris as non-millennial men

Health: topics that over-index for Millennials

- 18-24:
    - teen health, eating disorders, addictions, schizophrenia, acne, stds, phobias, contraception, social anxiety disorder
    - all greater than 3x average interest
- 24-35:
    - breastfeeding, pregnancy, miscarriage, preemies, infertility, multiples.

Millennial health interests are the same as any previous generation (with a little stress and anxiety added in)

Entertainment: topics that over-index for Millennials

- 18-24:
    - women: shortstories, *manga*, tv dramas, young adult books, contemporary literature, performing arts, R&B, Punk music, romance novels
    - men: *anime*, rap, manga, celebs, sci-fi, prowrestling, punk music, dance music, war movies, horror, animated TV

Yes, these are in order

@jproberts00 jroberts@about.com

Guys should read a novel or two, and talk to a girl about it.

News & Issues: topics that over-index for Millennials

- 18-24:
    - foreign policy, journalism, liberal politics, terrorism, race relations, civil liberties, animal rights, the economy, womens' issues, environmental issues, the Middle East
- Younger millennials are cause driven, women's health driven (both men and women), but **also** very interested in the economy, and foreign policy

Being interested in women's rights is a better predictor of age than gender.

We can also look inside a topic

- The most predictive topic within our Christianity site, and our Islam site for guys 18-24?
- Whether or not their religion allows them to **——————?**

We can also look inside a topic

- The most predictive topic within our Christianity site, and our Islam site for guys 18-24?

- Whether or not their religion allows them to **get a tattoo**

We've connected demographics and interests. Great!

- If I know your demographics, I can create a probability distribution over topics, and update it over time

- If I see you on a topic I can create a probability distribution over age and gender

- Success!

@jproberts00 jroberts@about.com

But when we look more widely:
One broad trend stands out

- The most predictive topics of a woman's age are specific to the people for whom she cares

- The most predictive topics of a man's age are related to himself

- Women are quantitatively described as *many people*

What about other signals?
Correlate other signals to site use:

- People who click through to content from email skew older
- People who get to the site via a search ad, skew older
- People who reach the site from Facebook skew 35-45

What about other signals?
Correlate other signals to site use:

- People who click through to content from email skew older
- People who get to the site via a search ad, skew older
- People who reach the site from Facebook skew 35-45
- At the weekend, people browse more, and favour different content to the weekday.
- When people come from Facebook they are 10-20x more likely to share to Facebook

@jproberts00 jroberts@about.com

An interesting pattern emerges

- Users who come in from Facebook share to Facebook **10-20x** more than other users

- People who come to the site read more content and browse more

- “Who you are” is a small perturbation vs “who you are right now”

- This is great! Every website can learn who you are right now.

The utopian outline: fixed

- Measure what you're looking at *right now*, clicked *right now*, where you are *now*, what network you came from *right now*

- Infer:
  - ~~age~~
  - ~~gender~~
  - ~~household income~~
  - interest areas

- Use this to provide more relevant recommendations

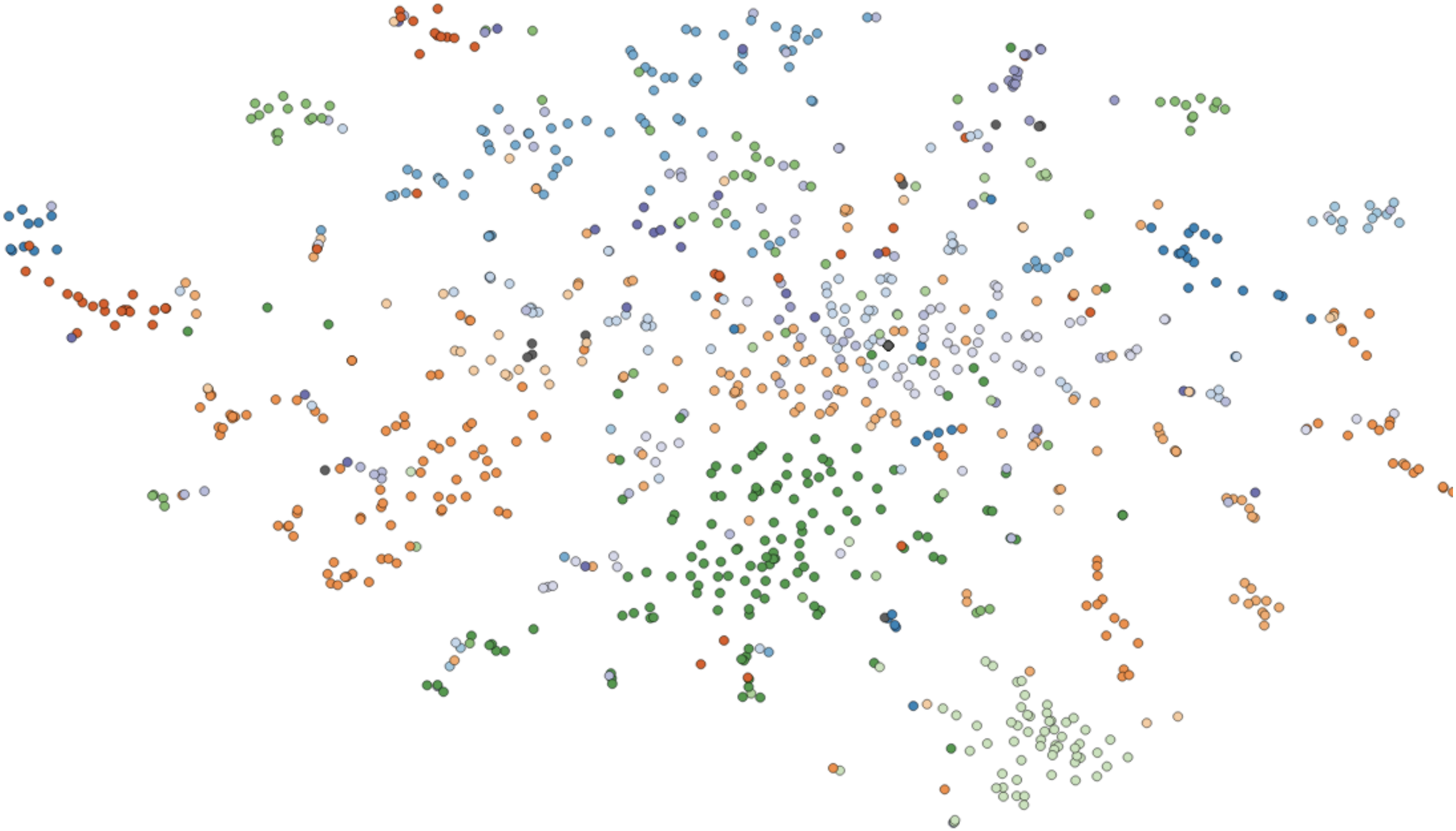
So What Do We Do?

- Treat people right now as unique people
  - “The past is a foreign country: they do things differently there”
- Rank signals that are predictive of behavior, and build a perturbative recommendation system:
  - Content first
  - Then source of traffic (Facebook vs Google vs Direct…)
  - Then day of week/time of day

Building a content based recommendation system

- Add meta-data to all documents:
  - Vertical specific TF-IDF and stop words
  - Entity extraction
  - Grade level
- Use these features to build a text-distance between documents
- Layer on separate distance measures for documents:
  - user co-occurence
  - SERP distance overlap

We have a distance measure between every piece of content



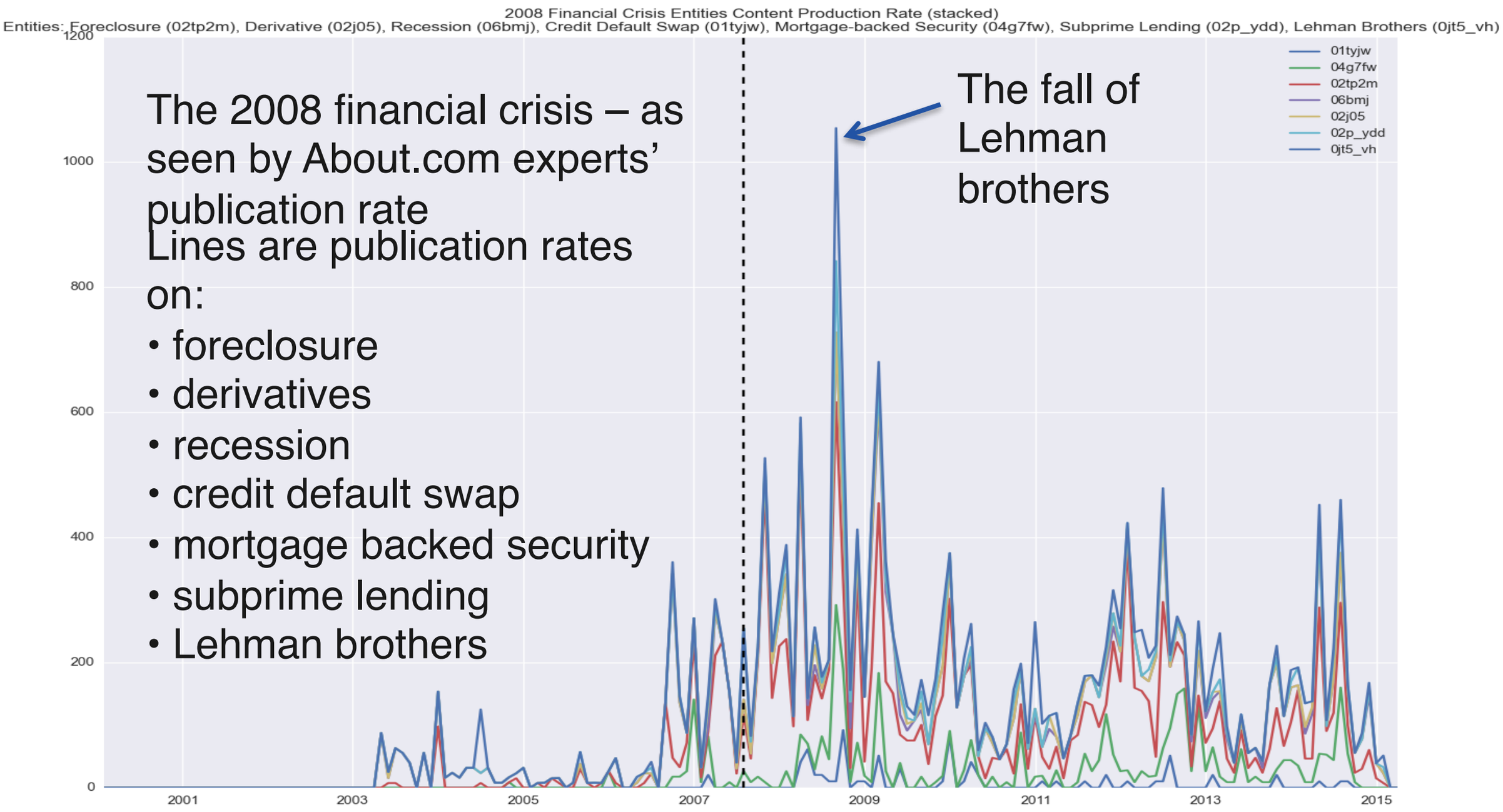
What happened?

- When we rolled out a content first recommendation engine:
    - Engagement on recommendations **doubled**
    - We consistently beat editorial recommendation
    - We've beaten every personalized recommendation engine we've tested against

Conclusions

- We can understand the behavior of large populations
- We can understand the behavior of individuals right now
- It’s massively arrogant to assume that a handful of data points over months will allow us to “understand a user”
- you will never know what they just heard on the phone before opening your site

When we listen to all our users - we can learn a lot



And we can build predictive measures

We’ve proved that an IAI trading strategy outperforms standard VIX and S&P 500 based trading strategies