cloudera
Ask Bigger Questions

Big Data 101 for Public Opinion Research

Todd Lipcon

Introductions

- Engineer at Cloudera
- I build software (Hadoop) for big data storage and analysis

- Some background in statistics and machine learning
- **No background in survey research!**
  - Looking forward to being told I’m way off base during Q&A and panel!
  - Or tweet disparagingly about @tlipcon

Cloudera

What We Do

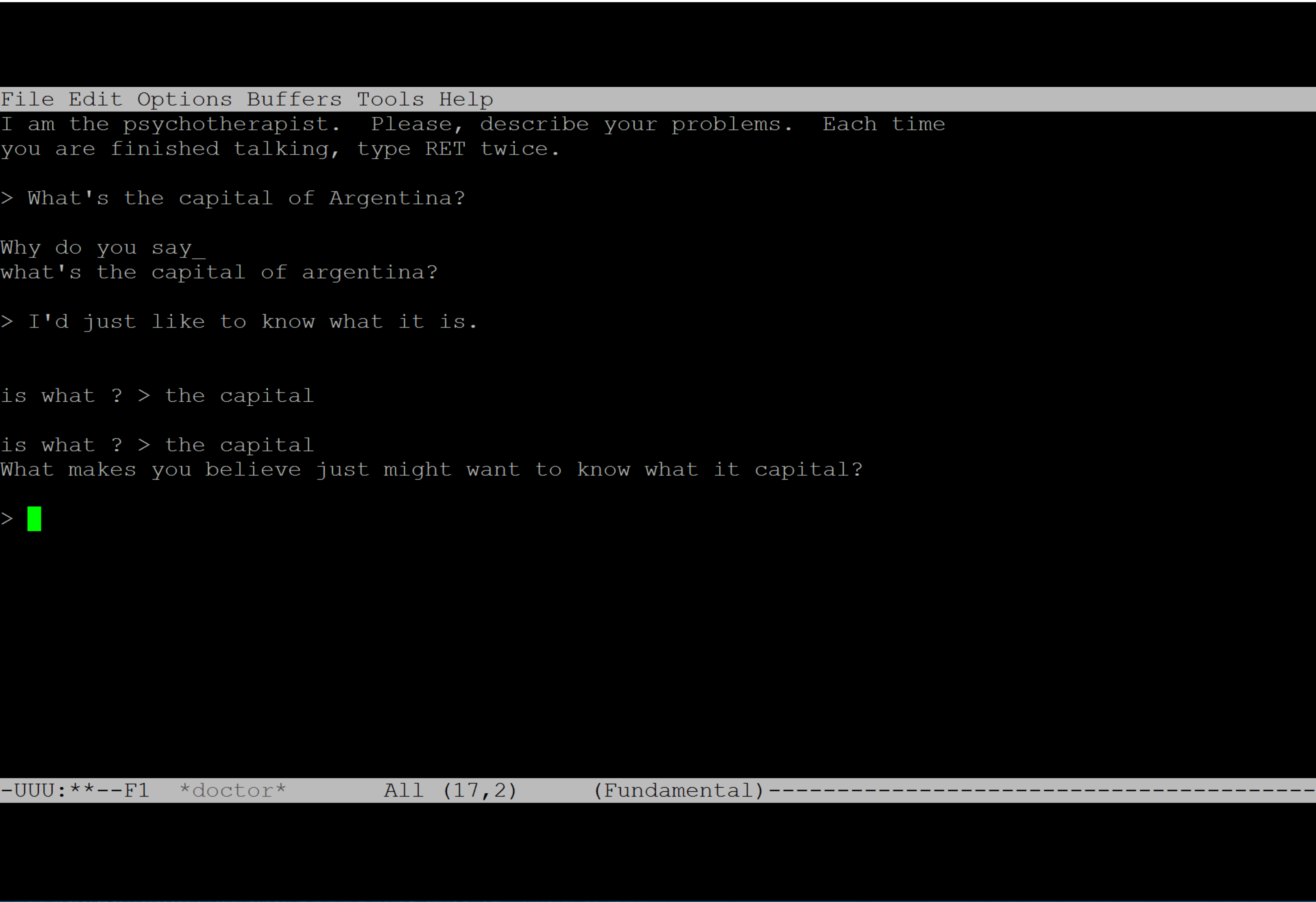
We enable enterprises and other organizations, including governments, foundations, charities and more to “Ask Bigger Questions” and become truly data-driven.

Our Vision

A world where more is accomplished faster and significant problems are resolved sooner and with less cost because people are acting on what they know, rather than hunches and guesses due to lack of viable insight.

This talk: *what if*, not *how to*

Let’s start with a story…



```
I am the psychotherapist.  Please, describe your problems.  Each time
you are finished talking, type RET twice.

> What's the capital of Argentina?

Why do you say_
what's the capital of argentina?

> I'd just like to know what it is.


is what ? > the capital

is what ? > the capital
What makes you believe just might want to know what it capital?

> 
```

```
-UUU:**--F1  *doctor*       All (17,2)      (Fundamental)----------------------------------------
```

PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
सोचिए
$3,400
$4,400
$1,200
KEN
WATSON
BRAD

Data is the difference

The New York Times

Thomas Hannich for The New York Times

Customers who used their credit cards at a bar were **four times more likely** to miss a payment than those who used their cards at the dentist.

Prepping for Hurricane Charlie at Wal-Mart

Image credit: Sam Dundon

Prepping for Hurricane Charlie at Wal-Mart

Prepping for Hurricane Charlie at Wal-Mart

What’s this got to do with public opinion research?

Survey Research Trends 1997-2012
(AAPOR, Pew Research)

	1997	2012
CON2: Contact rate (% reached)	**90%**	**62%**
COOP3: Cooperation rate (% contacted that yielded data)	**43%**	**14%**
RR3: Response rate (% sampled that yielded data)	**36%**	**9%**

Survey Research Trends 1997-2012
(AAPOR, Pew Research)

- **Harder to contact** survey respondents
- **Faster** (24 hour) news cycle
- More surveys demanded from **less responsive** public

- Survey research using traditional methods getting **increasingly expensive**!

A potential solution: web surveying?

- Email blasts?
- Pop-up surveys?
- Pay-per-answer?
- Answer for access to content?

- Traditional survey methods on a new medium **doesn't solve the problems**!
  - **Expensive**, more **biased**, and even **lower response rate**!

And yet…

- 80% of Americans are online. (Pew Internet)
- 69% of online adults use social networking sites.
- Users regularly share intimate details of their life with anyone who will listen.

- **Prediction:** public opinion research will rely more and more on data collected through these new media.

cloudera®
Ask Bigger Questions

Just how much do we know about people?

Geographical Location

- “Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity” (Facebook Data Team @WWW2010)

- Authors infer user’s location accurately based on location of friends, *even if the user does not volunteer the info!*

Age

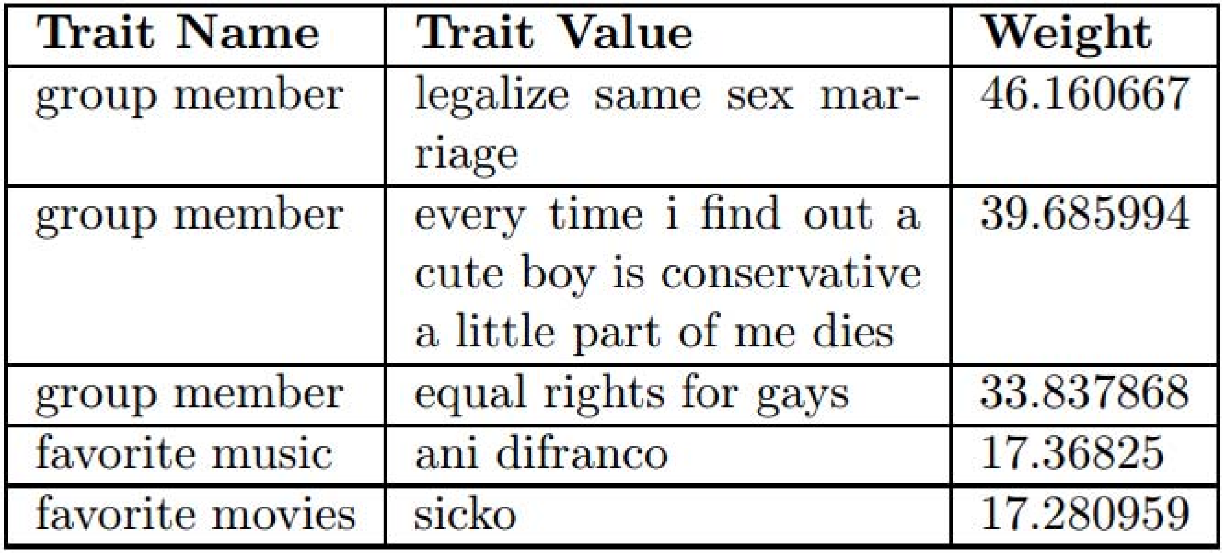
- “Estimating Age Privacy Leakage in Online Social Networks” (Dey, Tang, Ross, Saxena @INFOCOM 2012)

- Authors infer age within 4 years for 84% of users using friend network and educational history
  - Even though only 1.5% specify age!
  - Even though only 20% specify educational history!

Political Leanings

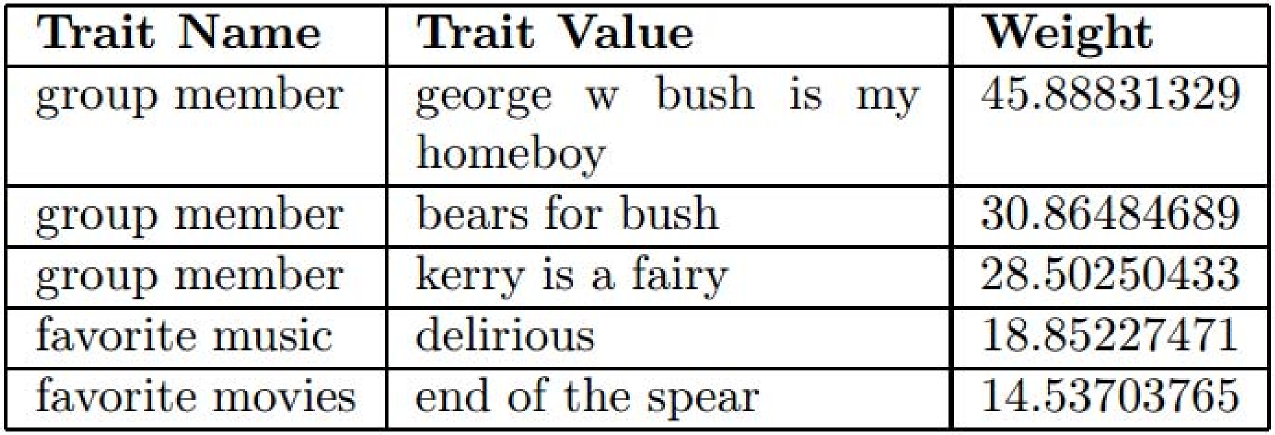
- “Inferring Private Information Using Social Network Data” (Lindamood and Kantarcioglu, UT Dallas Tech Report 2008)

- Authors infer political party with 80% accuracy based on actions and friend network



Trait Name	Trait Value	Weight
group member	legalize same sex marriage	46.160667
group member	every time i find out a cute boy is conservative a little part of me dies	39.685994
group member	equal rights for gays	33.837868
favorite music	ani difranco	17.36825
favorite movies	sicko	17.280959

Table 4: A sample of the most liberal traits



Trait Name	Trait Value	Weight
group member	george w bush is my homeboy	45.88831329
group member	bears for bush	30.86484689
group member	kerry is a fairy	28.50250433
favorite music	delirious	18.85227471
favorite movies	end of the spear	14.53703765

Table 5: A sample of the most conservative traits

If we can learn location, age, politics, what else can we learn?

Likely voter modeling online?

- Suggested that likely voter modeling has become more important than opinion tracking
- Strong, easy-to-detect signals:
  - Google search for “polling place <zipcode>”
  - Comments about voting or candidates
  - Membership in political groups
- Combine with historical “ground truth” data
  - Posts about voting from previous elections
  - Facebook online vote counter

Longitudinal studies

- Track the opinion of an individual over time
- How does their engagement shift as a political campaign’s message changes?
- Correlate engagement shifts with demographics

Obama’s “Big Data” Victory

> Who got which email depended on “big data”—information about each fundraising prospect and how different people react to different messages. In this year's election, it looks as if the Obama team's use of such data was one of its biggest edges over the Romney effort.

THE WALL STREET JOURNAL.

November 19, 2012

cloudera®
Ask Bigger Questions

Big Data Opinion Research

Stop asking questions.

Start analyzing what the public is already saying and doing.

Big Data Public Opinion Research

- **Use passive subjects**
  - If they won't *participate*, just *observe* them.
  - aka *observational* studies
- **Collect all data all the time**
  - Not targeted around a specific question.
- **Accept "messy data"**
  - Will require new analysis techniques (natural language, etc)
  - Will require software and algorithms to add structure

Big Data Public Opinion Research

- **Trust the machine**
  - Methodologies might not be easy to interpret.
  - That's OK!
- **Work with new data collectors**
  - Facebook, Twitter, Google may collect the aggregate data.

cloudera®
Ask Bigger Questions

The Role of the Big Data Opinion Researcher

Role of the Big Data Opinion Researcher

> “The *practical statistician* first needs a sound understanding of how baseball, poker, elections or other uncertain processes work, what measures are reliable and which not, and what scales of aggregation are useful, and then to utilize the tool kit as well as possible.”

— Nate Silver

Role of the Big Data Opinion Researcher

> “A *data scientist* is a person who is better at statistics than any software engineer and better at software engineering than any statistician.”

— Josh Wills
Director of Data Science
Cloudera

Role of the Big Data Opinion Researcher

- **Data Curation**
  - Know what data is available and how to collect it
- **Data Munging**
  - Convert “messy data” to “structured data”
  - Automated coding algorithms
- **Data Insight**
  - *Computers* are dumb.
  - *People* provide **insight**.
- ***Augment* traditional methods**

Summary

Summary

- Big Data enables new insights about people’s behaviors and likely actions.
- The public may be less cooperative with traditional research, but we can learn a lot about people online.
- New methods and ways of thinking will be necessary to work with these new technologies.
- Big data research will likely *augment* rather than *replace* existing methods.

cloudera®

Ask Bigger Questions