cloudera
Ask Bigger Questions

Big Data and Data Science 101

Todd Lipcon | Software Engineer
(with much credit due our Data Science team, in particular Josh Wills)
December 2013

Introductions

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

- **No background in survey research!**
  - Some background in statistics and machine learning
  - Unlike last year's PAPOR, this year I'm not going to fake it.

What’s a ‘Data Scientist?’

Josh Wills
@josh_wills

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Delete Favorite

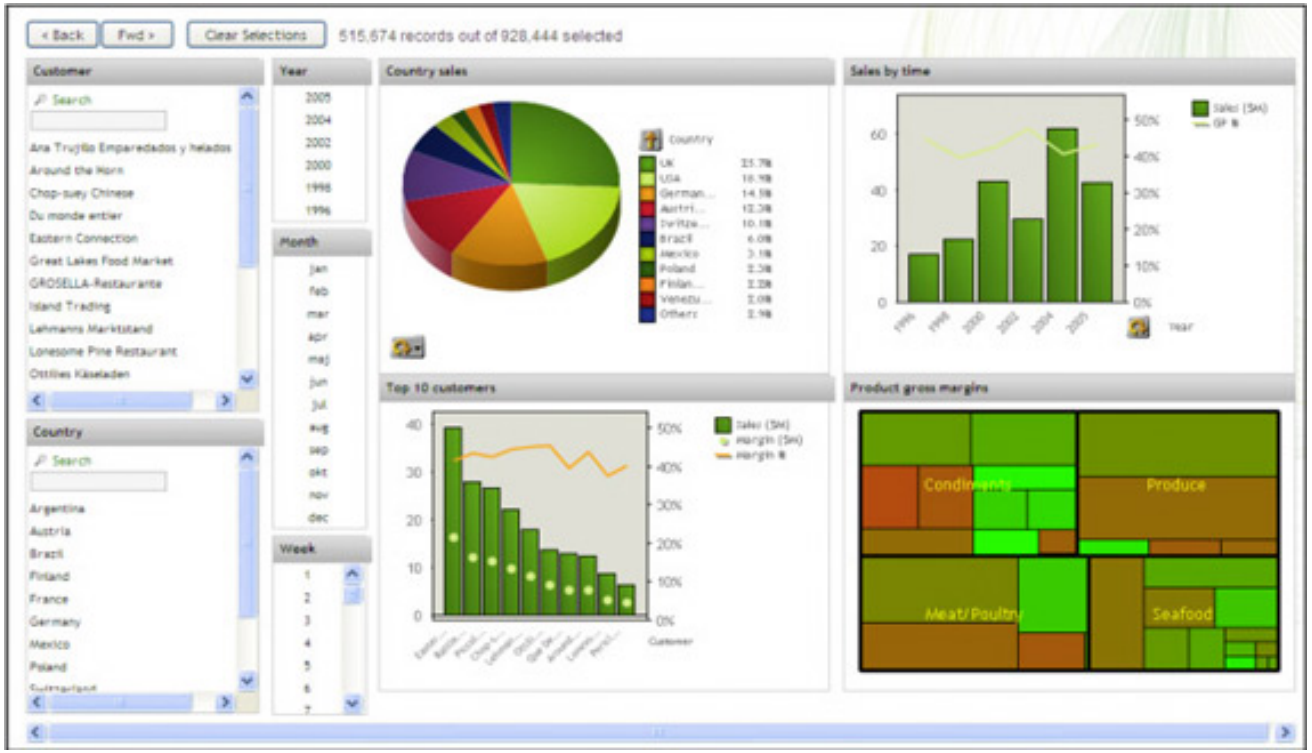
501 RETWEETS | **183** FAVORITES

9:55 AM - 3 May 12 · Embed this Tweet

Another definition of a data scientist

- A person who mixes computer science, statistics, and data visualization to **analyze** sets of data
  - Often “funny looking” sets of data
  - More complex analyses than “slice and dice” summary statistics (eg machine learning)

The Humble Sales Dashboard



(an aside on pie charts)

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

Another definition of a data scientist

- A person who mixes computer science, statistics, and data visualization to **build analytical applications**
  - Rich visualizations
  - Interactive analysis that lets the consumer explore the data themselves
  - Things which make our lives better

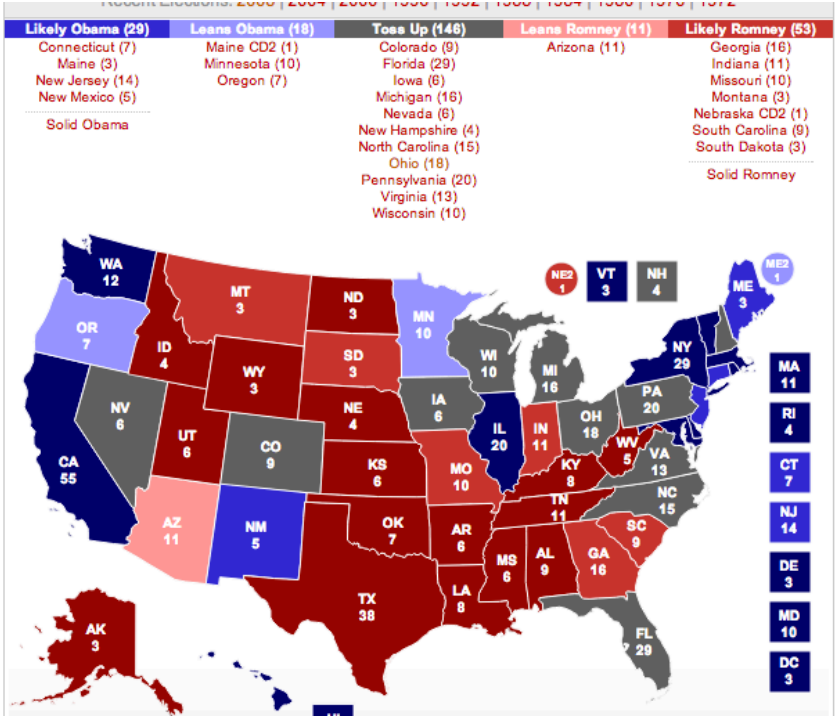
Developing Analytical Applications

A Case Study

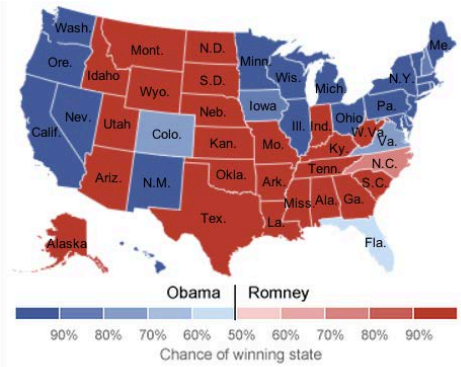
2012: The Predicting of the President

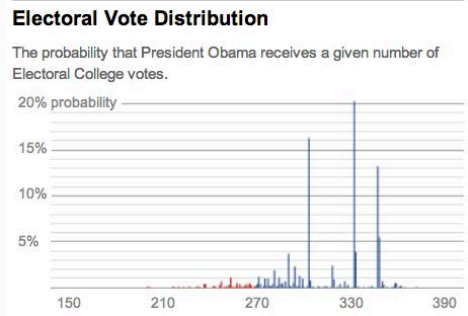
RealClearPolitics

- Simple Average of Polls
- Transparent
- Simple Interactions
  - “what if” analysis on state output



FiveThirtyEight

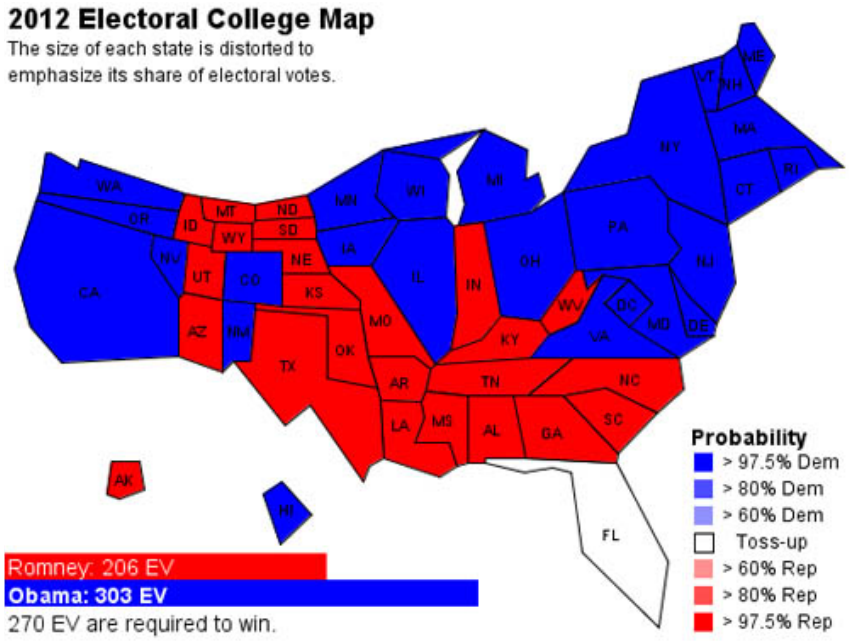




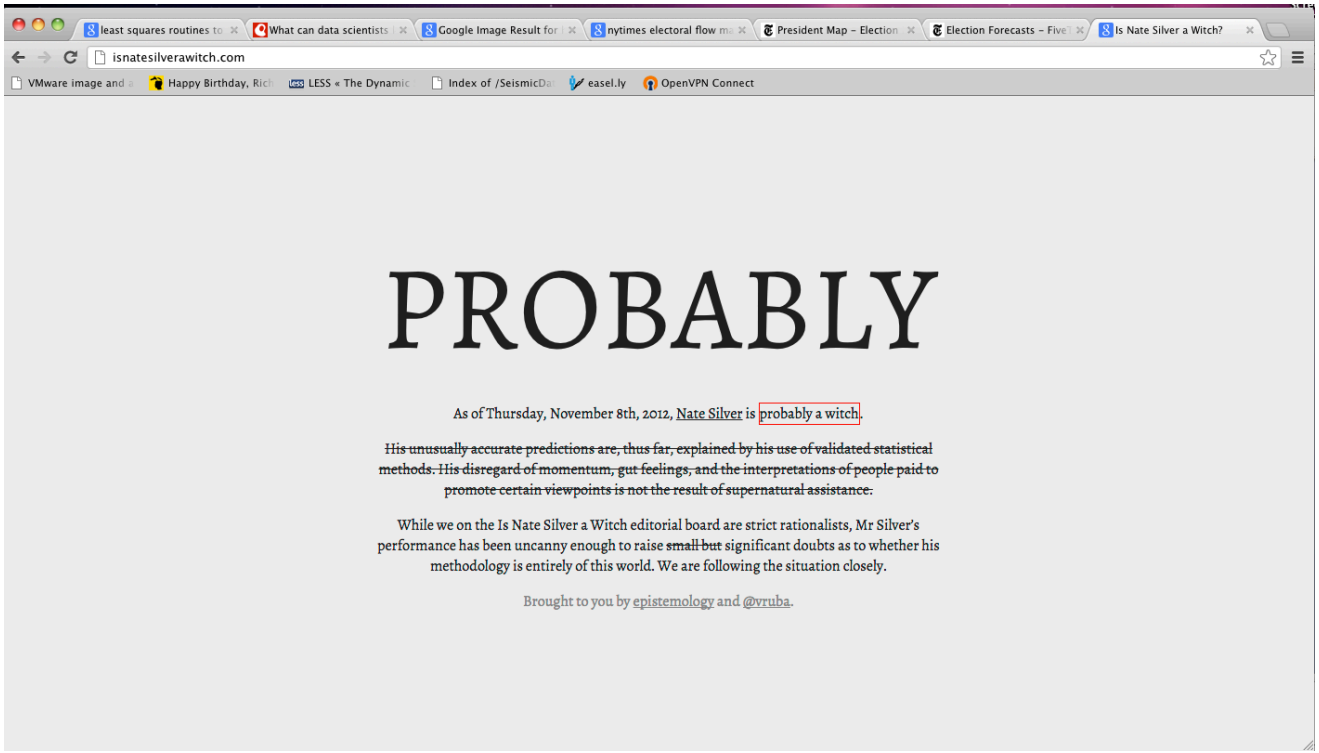
- Complex Model
  - Many factors (economic, correlations, etc)
- Opaque
  - Secret sauce
- Simple Interactions with a richer UI

Princeton Election Consortium

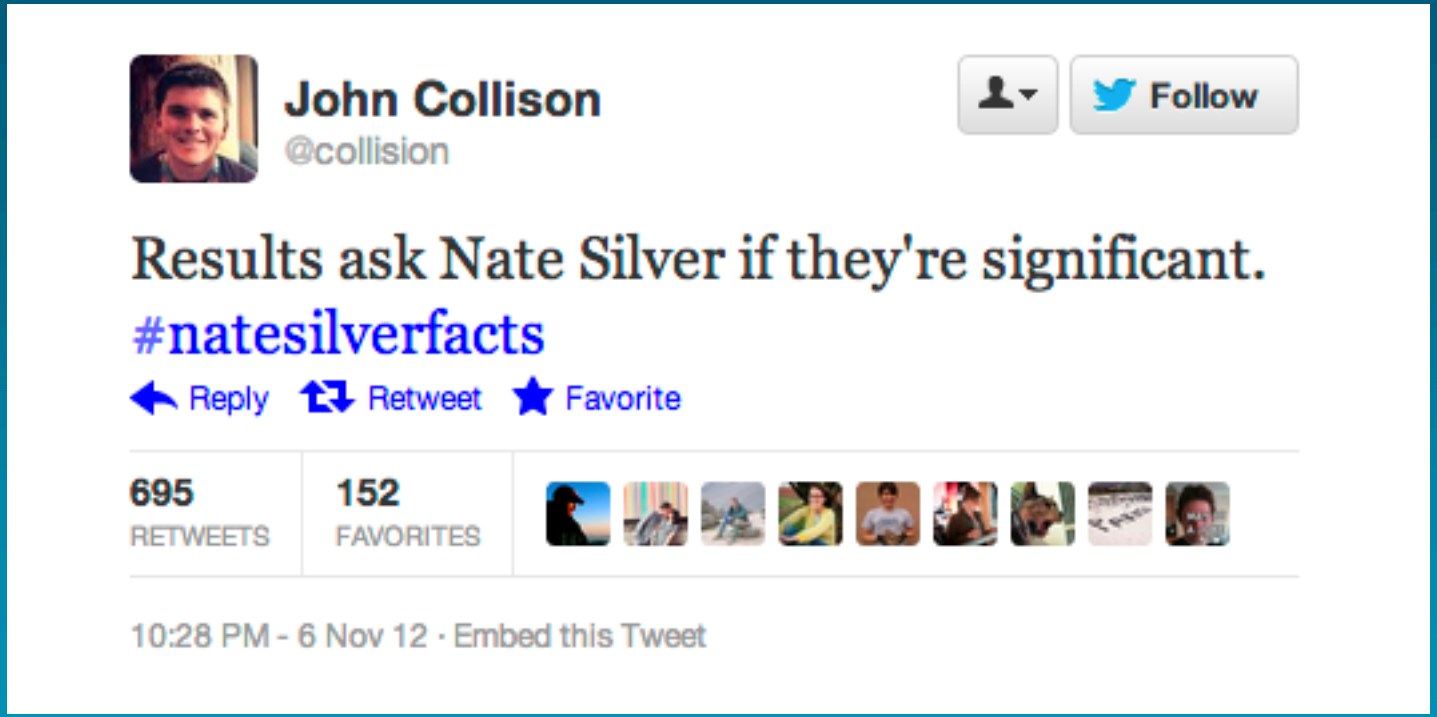
- Medians and Polynomials
- Transparent
- Rich Interactions
  - “What if” a given poll has bias?



How Did They Do?



A Few of These, Because They're Fun

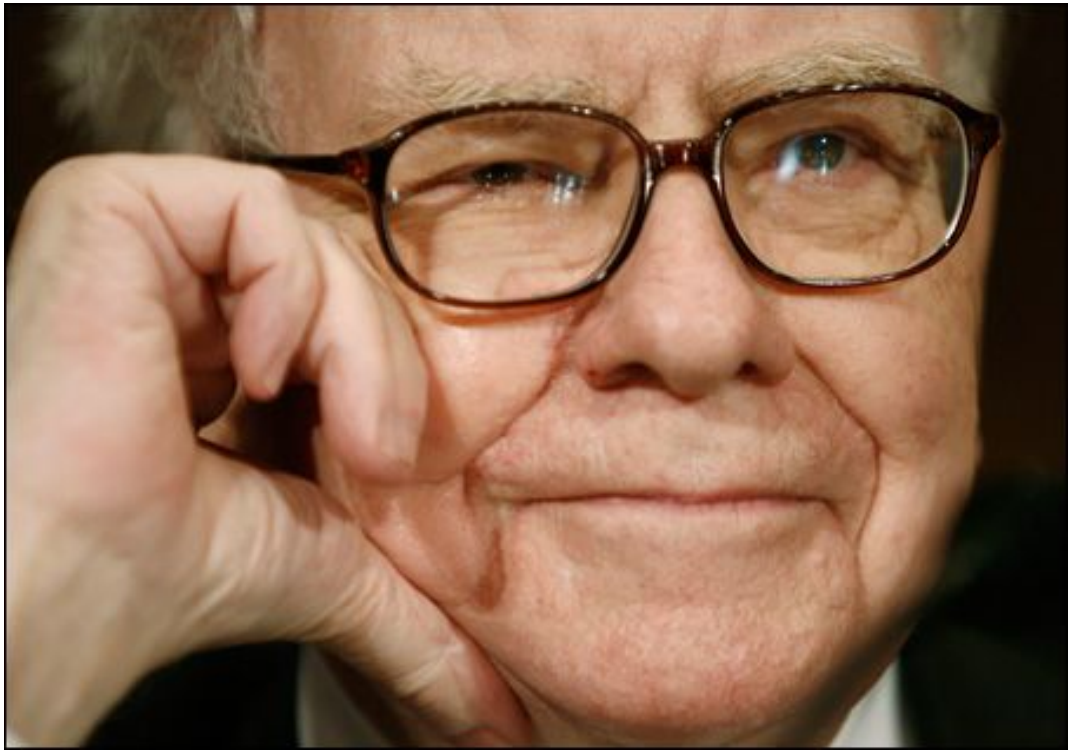


A Few of These, Because They're Fun

Here's the Rub: One Expert Beat Nate

(Markos Moulitsas at DailyKos)

Index Funds, Hedge Funds, and Warren Buffett

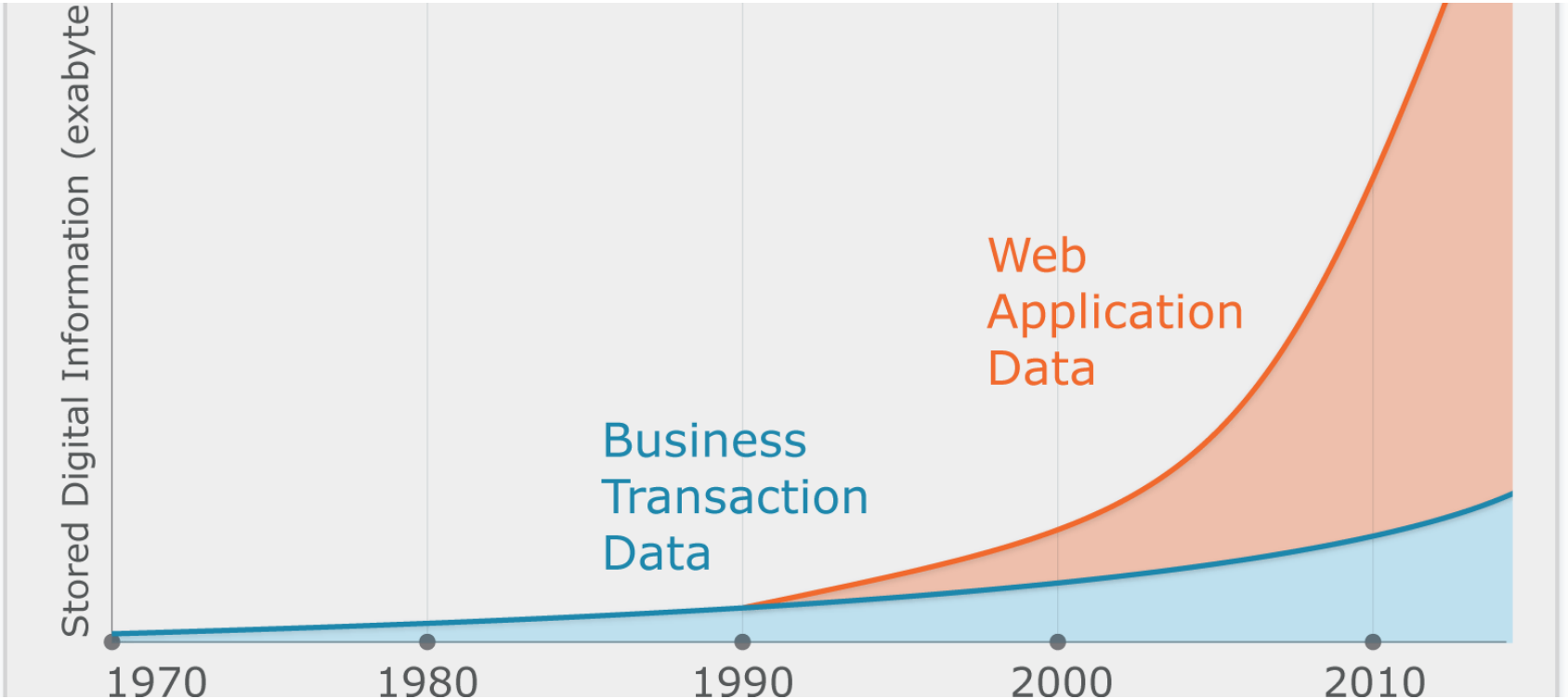


A Brief Introduction to Big Data and Hadoop

Data Storage in 2001: Databases

- Structured (tabular) data sets
- Intensive processing done where data is stored (SQL)
- Somewhat reliable
- **Expensive at scale**

And Then, This Happened



Big Data Economics

- No individual record is particularly valuable
- Having *every* record is incredibly valuable
  - Web index
  - Recommendation systems
  - Market basket analysis
  - Online advertising

“In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers.”

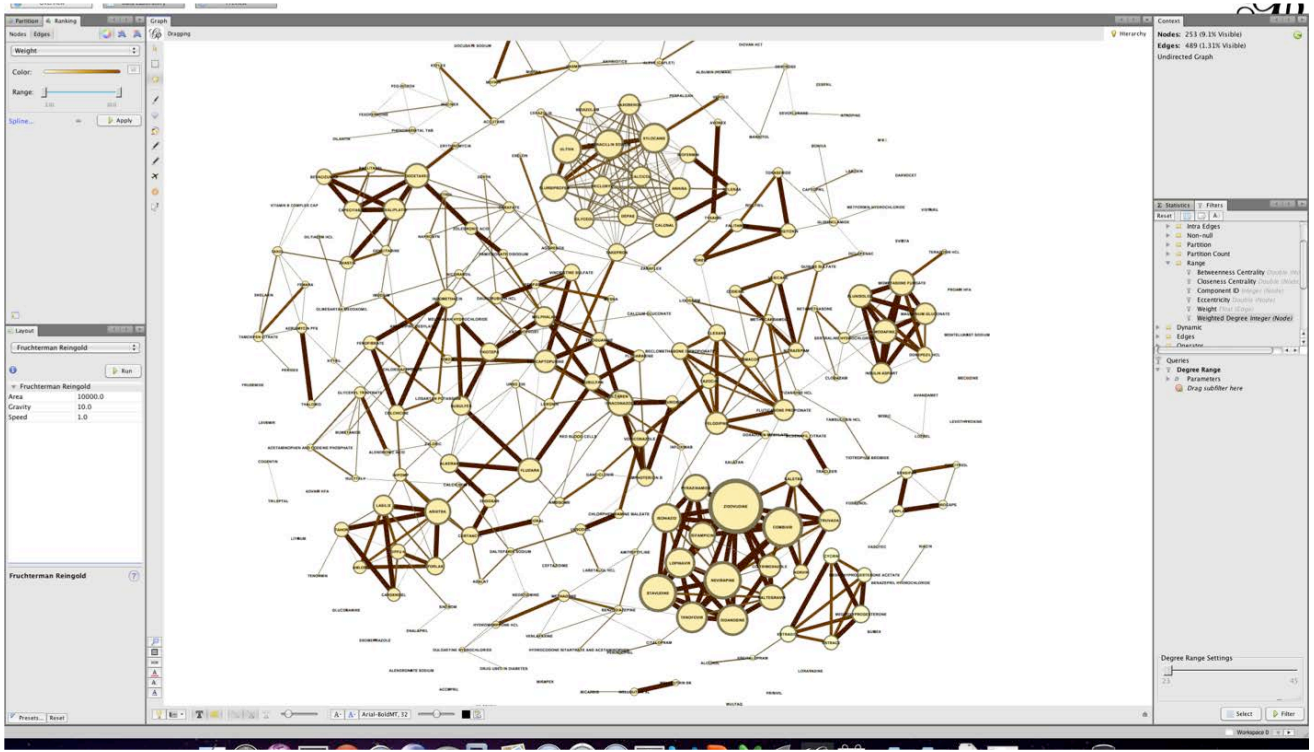
- Grace Hopper

Data Storage in 2013: Hadoop

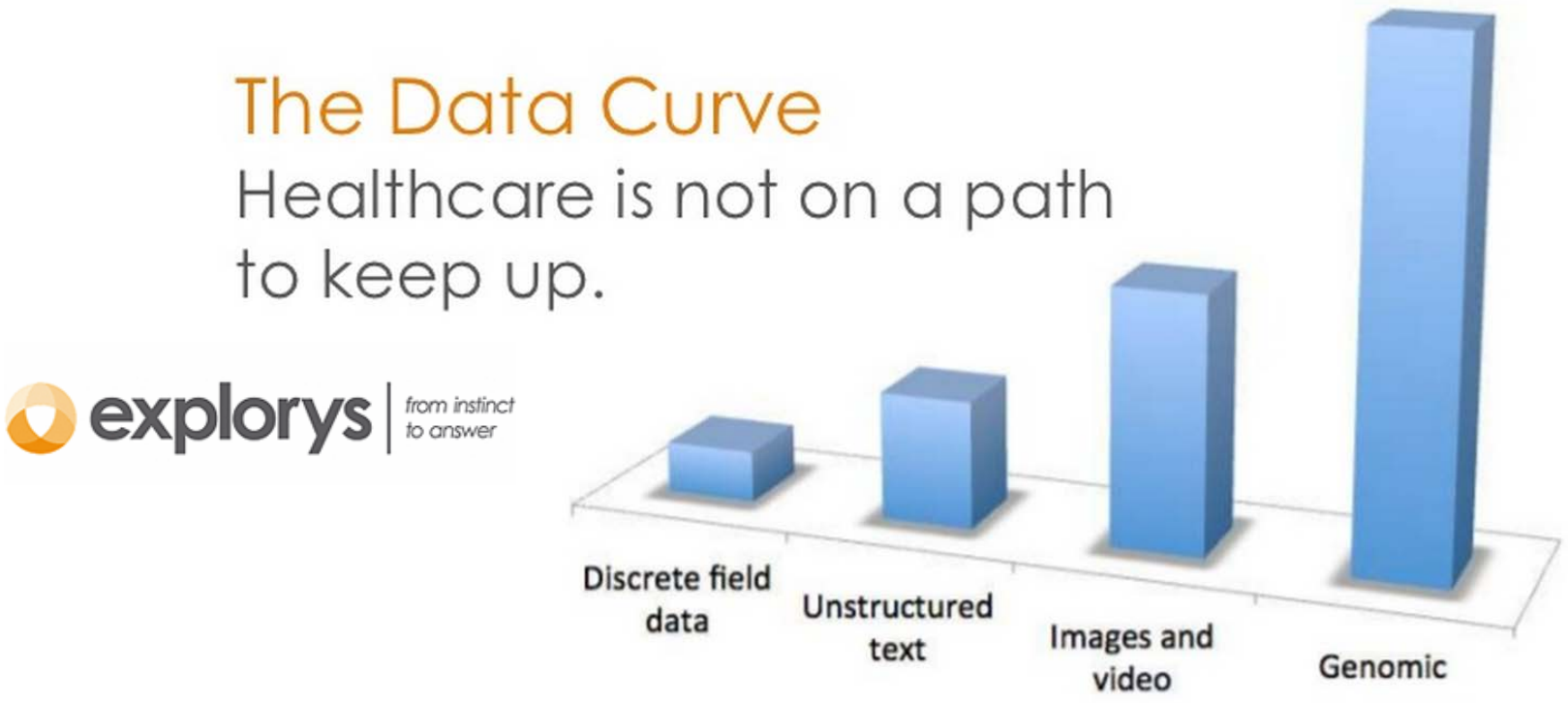
- Stores any kind of data
- Many different in-situ processing engines (R, SAS, SQL, Search, etc)
- Reliable
- **Cheap, even at scale**

What can you build with big data?

Adverse Drug Events



Medical record analytics



Durkheim Project

A Couple of Themes

1. Interactive applications, not just static “reports”
2. Integrate data from many sources, not just one.
3. Some amount of programming usually necessary, but you don’t always need a CS degree!


cloudera®
Ask Bigger Questions