Data Science in Action

Kamal Maheshwari, Boulder, CO

[On loan from - Sean Owen| Director, Data Science]

Data Science Deluge..

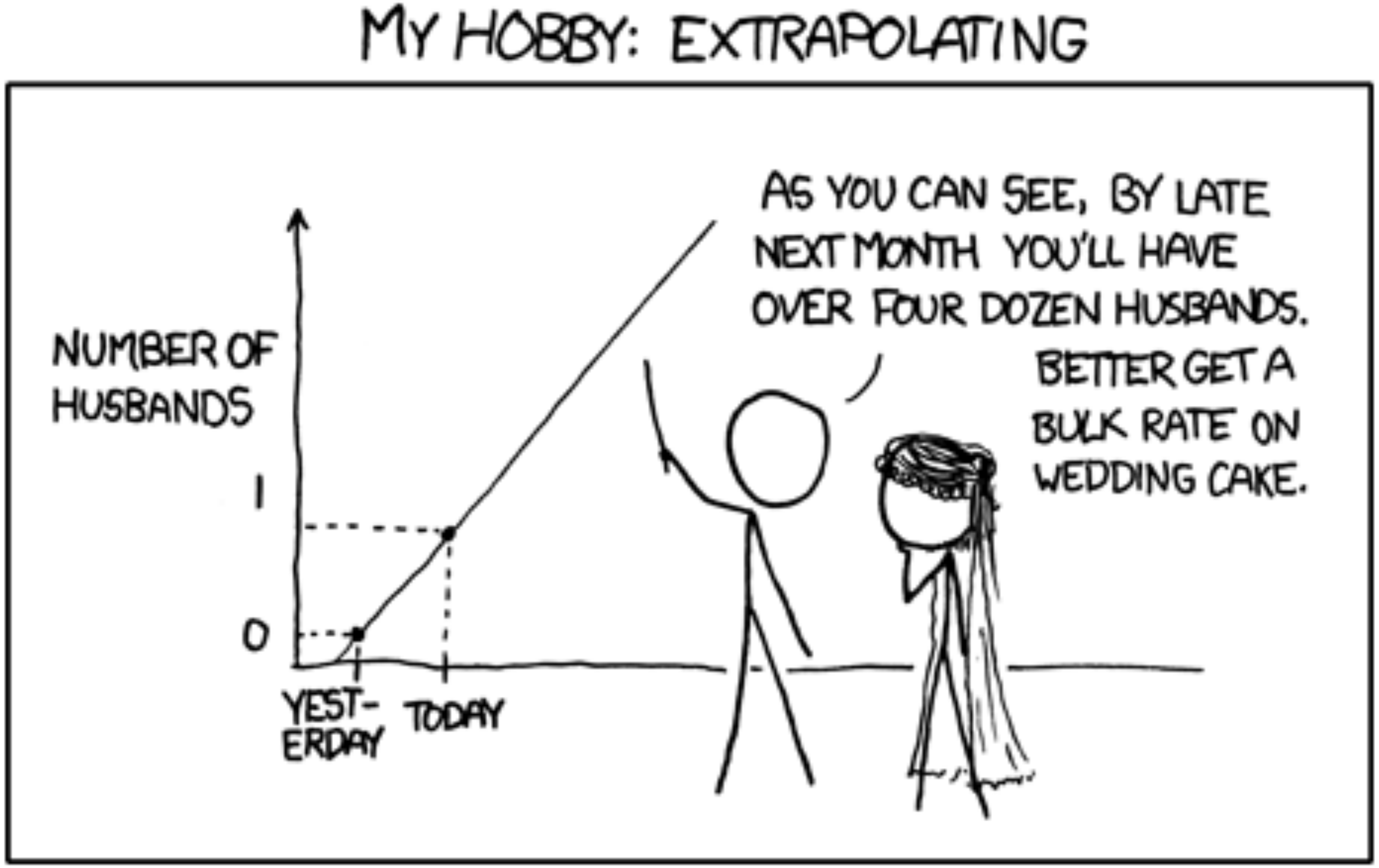
Will data science prevent deadly police shootings?
DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY
A Data Scientist's Real Job: Storytelling
Data Science Is Dead

Talk Overview

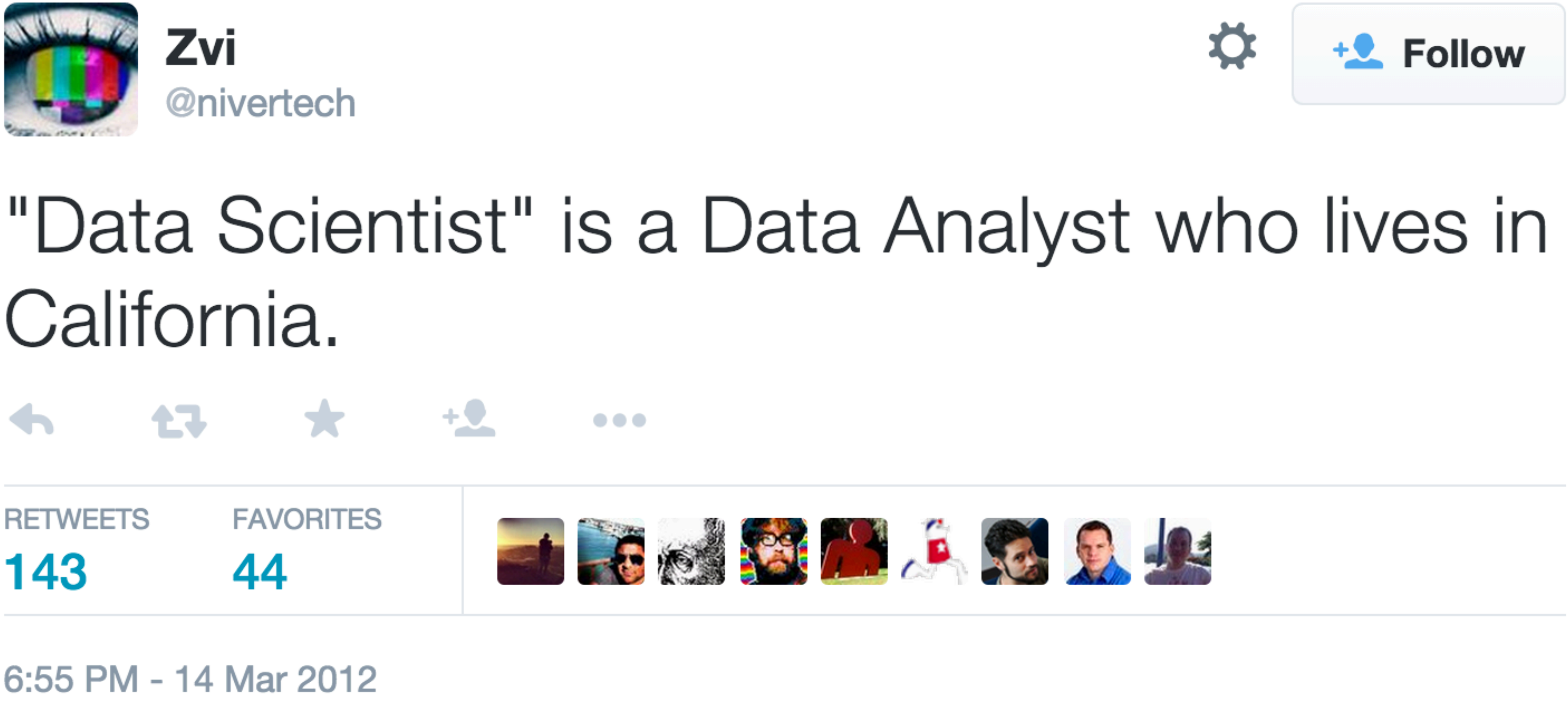
- What is a Data Scientist?
- Introduction to Data Products
- Understanding Data Processing Engines: SQL, MapReduce, and Spark
- Data Modeling for Data Science
- Algorithms Every Data Scientist Should Know
- Executing a Data Product Strategy

What is a Data Scientist?

One Definition…


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

…and Another…



Zvi
@nivertech

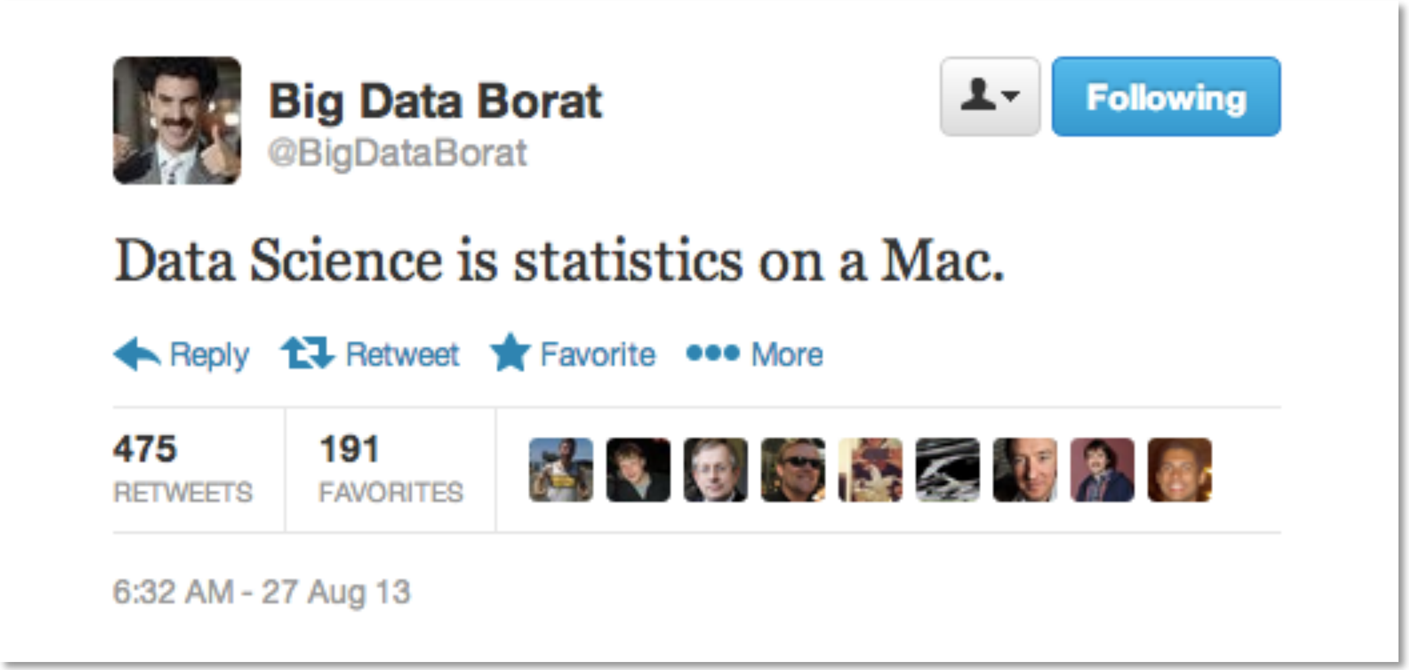
"Data Scientist" is a Data Analyst who lives in California.

RETWEETS 143 FAVORITES 44

6:55 PM - 14 Mar 2012

cloudera

…and Another…



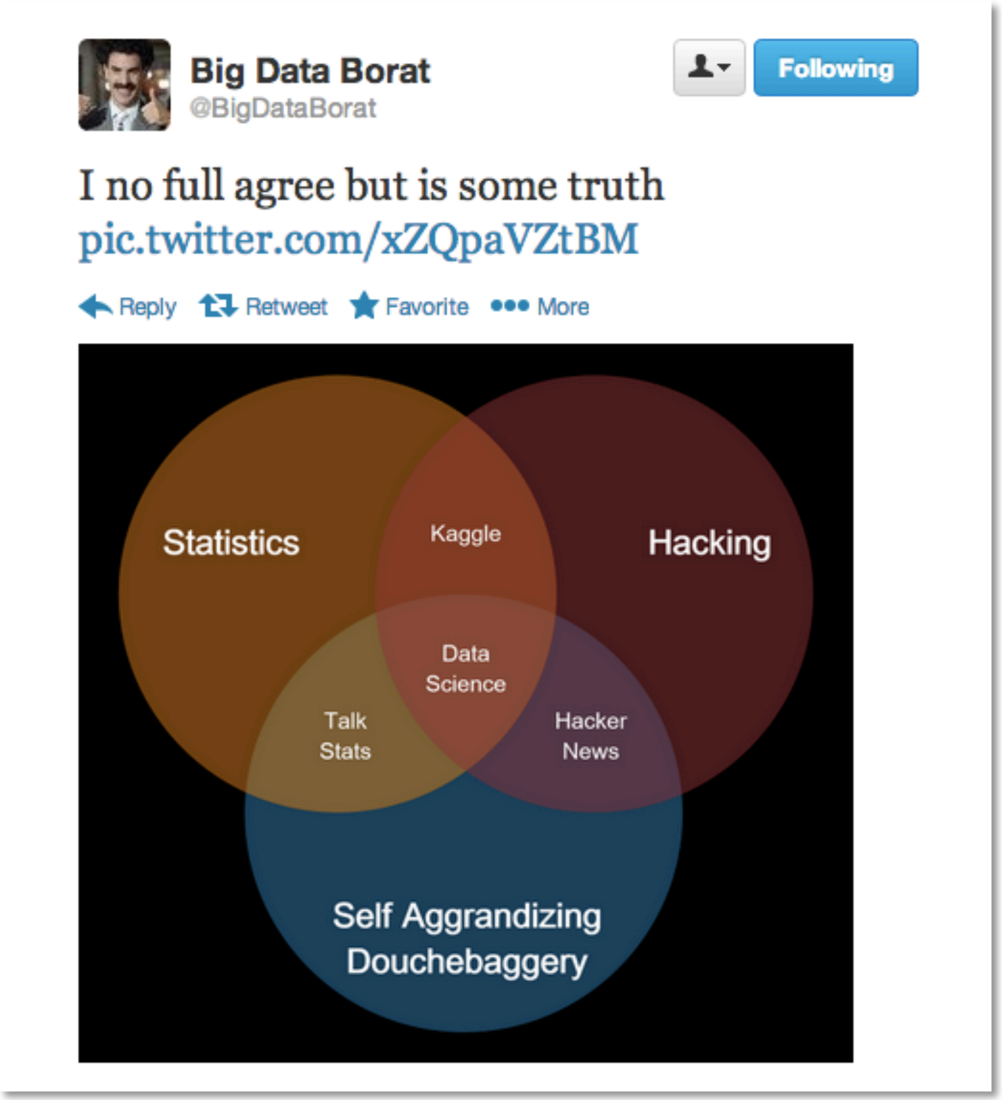
Big Data Borat
@BigDataBorat

Data Science is statistics on a Mac.

475 RETWEETS
191 FAVORITES

6:32 AM - 27 Aug 13

…and Another…



I no full agree but is some truth pic.twitter.com/xZQpaVZtBM

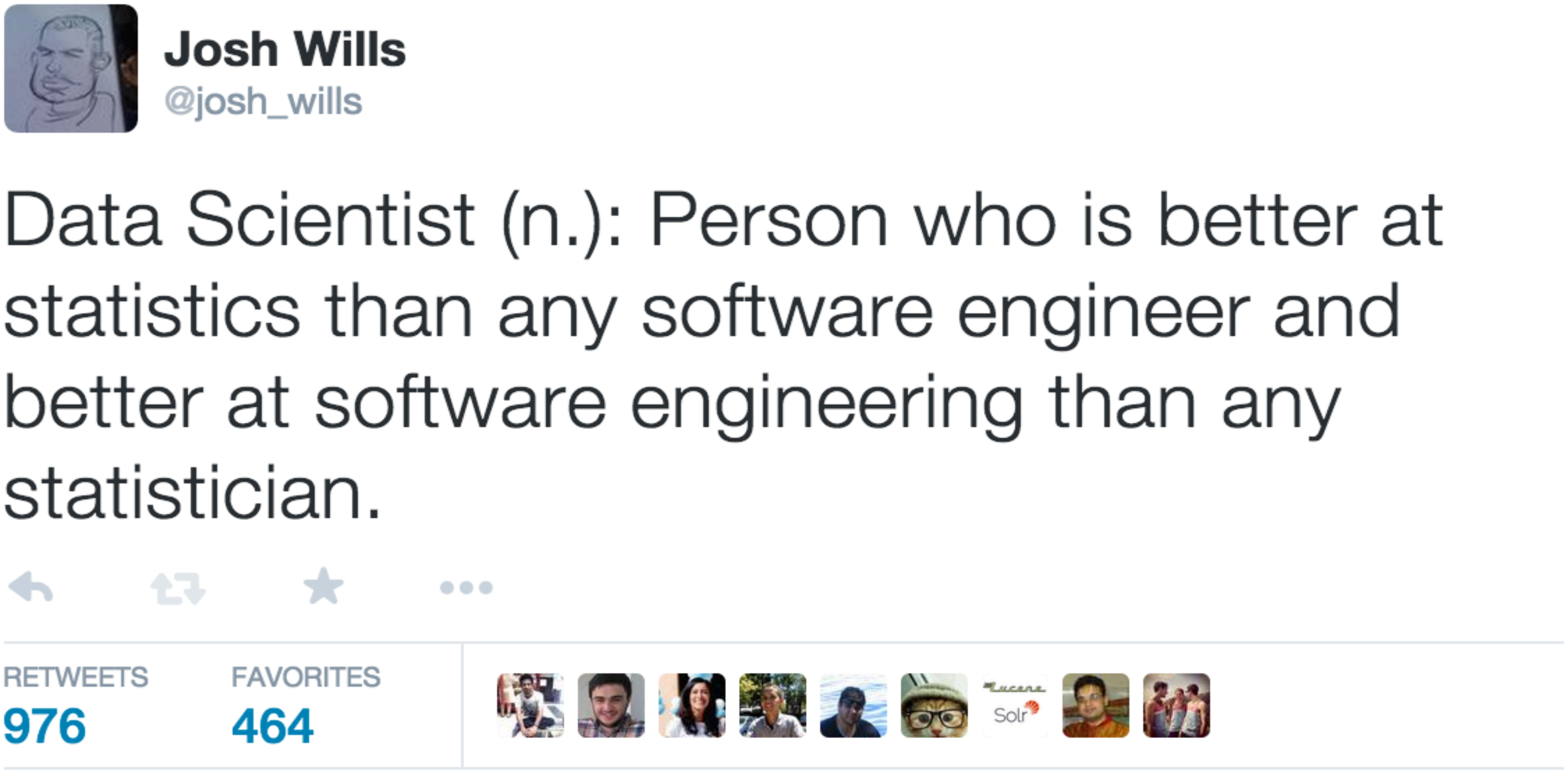
…and Another…

Phase 1. Collect Data

Phase 2. Data Science?

Phase 3. Profit!

…versus Another



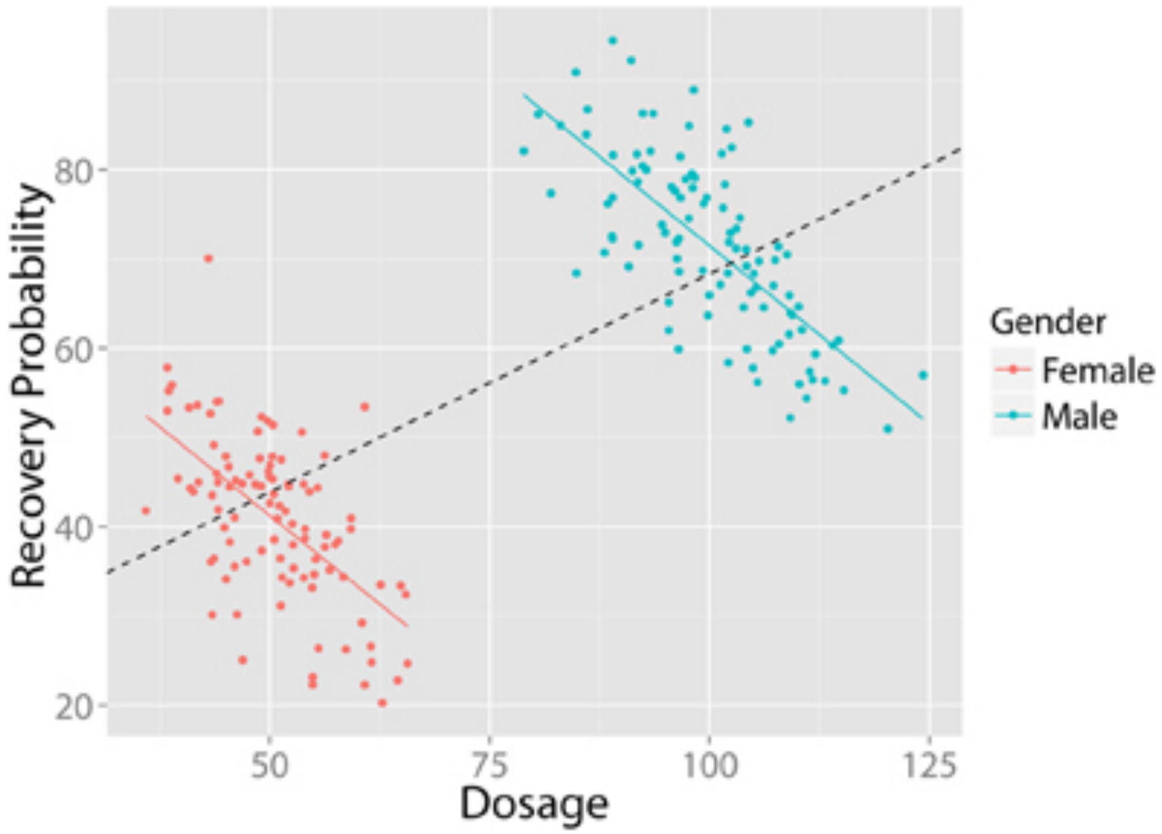
Josh Wills
@josh_wills

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

RETWEETS	FAVORITES
976	464

Data Scientist vs. Software Engineer

- Most software engineers do not understand the assumptions behind statistical models
  - Independence
  - Normally distributed errors
- Common errors/pitfalls when conducting data analysis
  - Correlation != causation
  - Simpson’s Paradox



cloudera

Introduction to Data Products

What is a Data Product?

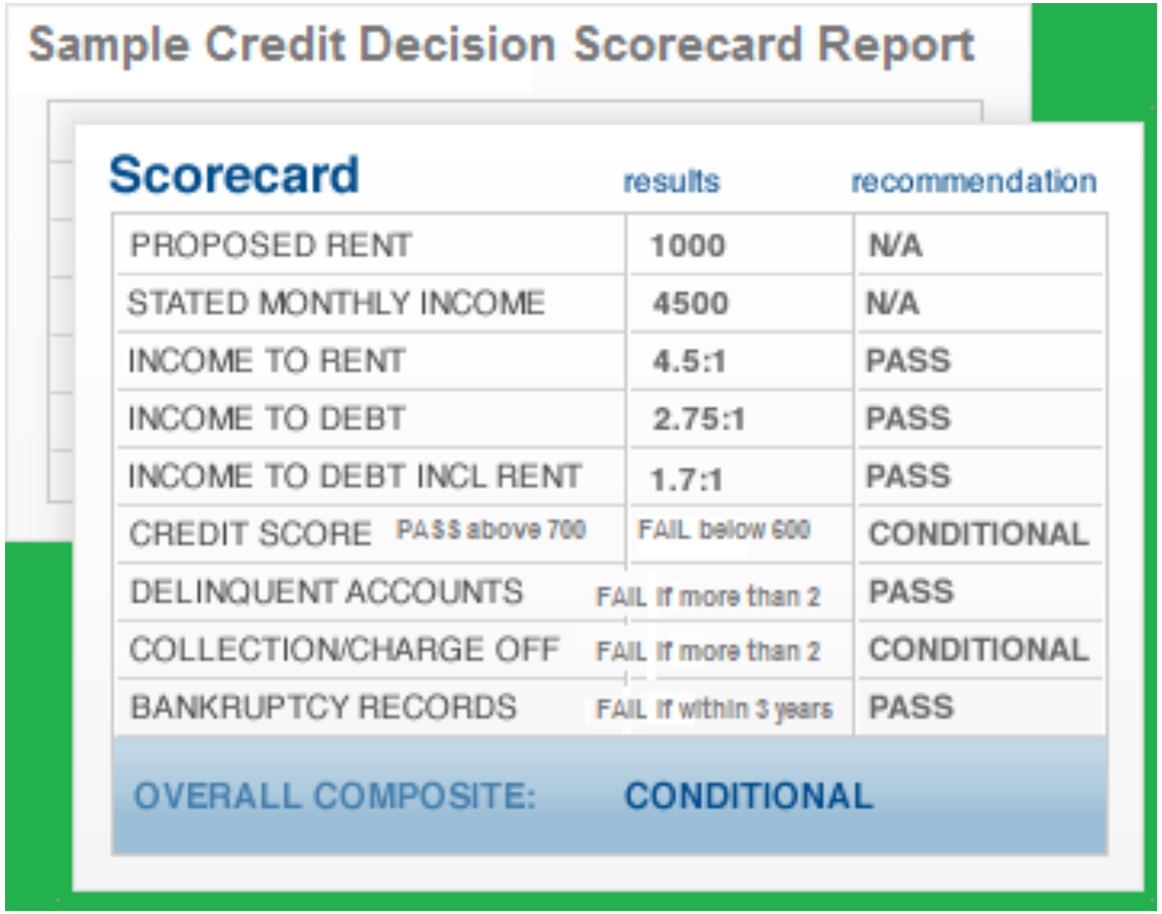
- Created from data
- Using the product generates more data
- This data can be used to improve the functionality of the product
  - Ideally in an automated fashion

Classic Data Products

- Data products aren’t a new idea, just a new phrase

- Credit scorecards and rules-based fraud detection engines have been around for a long time

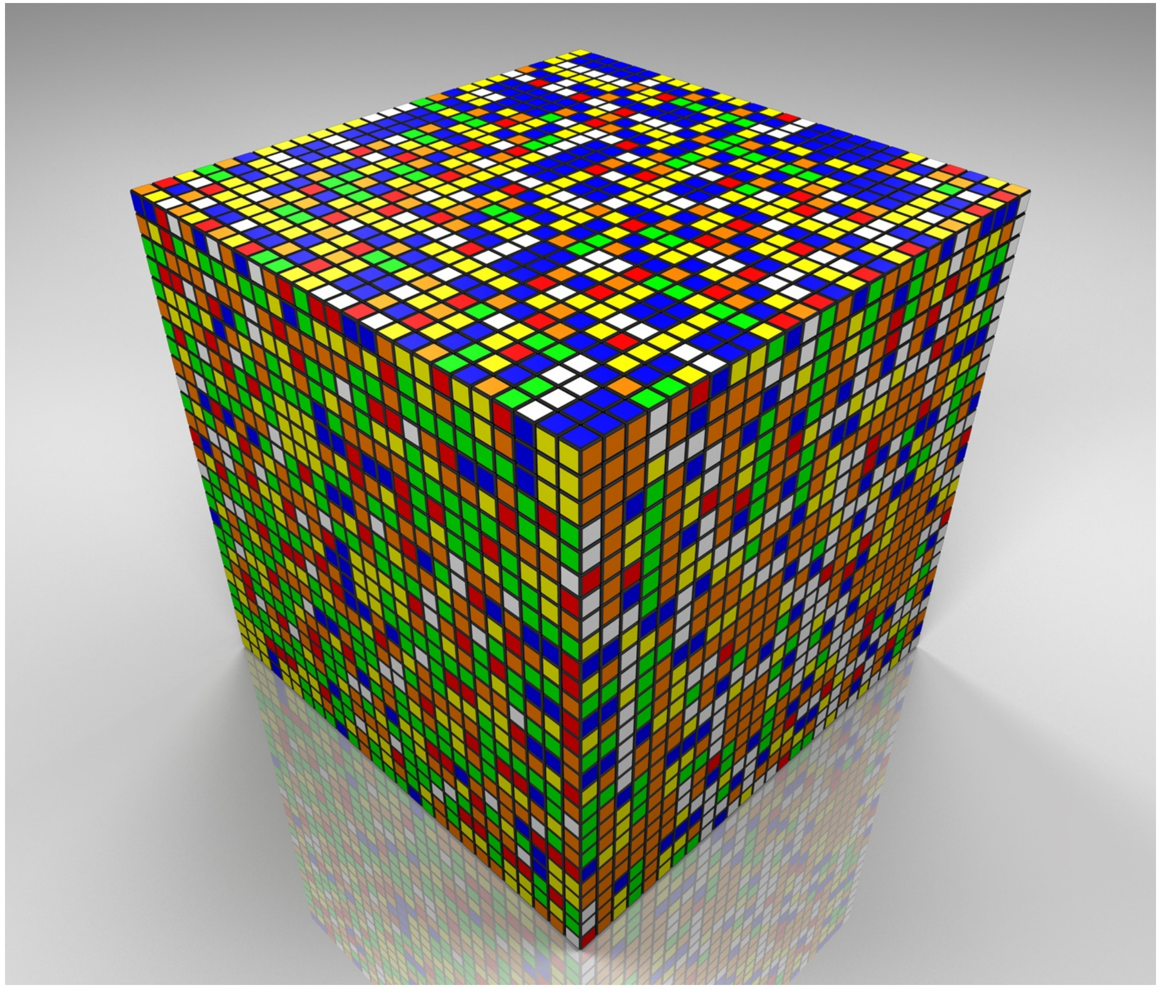
- What has changed?



Sample Credit Decision Scorecard Report

Scorecard	results	recommendation
PROPOSED RENT	1000	N/A
STATED MONTHLY INCOME	4500	N/A
INCOME TO RENT	4.5:1	PASS
INCOME TO DEBT	2.75:1	PASS
INCOME TO DEBT INCL RENT	1.7:1	PASS
CREDIT SCORE PASS above 700	FAIL below 600	CONDITIONAL
DELINQUENT ACCOUNTS	FAIL if more than 2	PASS
COLLECTION/CHARGE OFF	FAIL if more than 2	CONDITIONAL
BANKRUPTCY RECORDS	FAIL if within 3 years	PASS
OVERALL COMPOSITE:	CONDITIONAL	

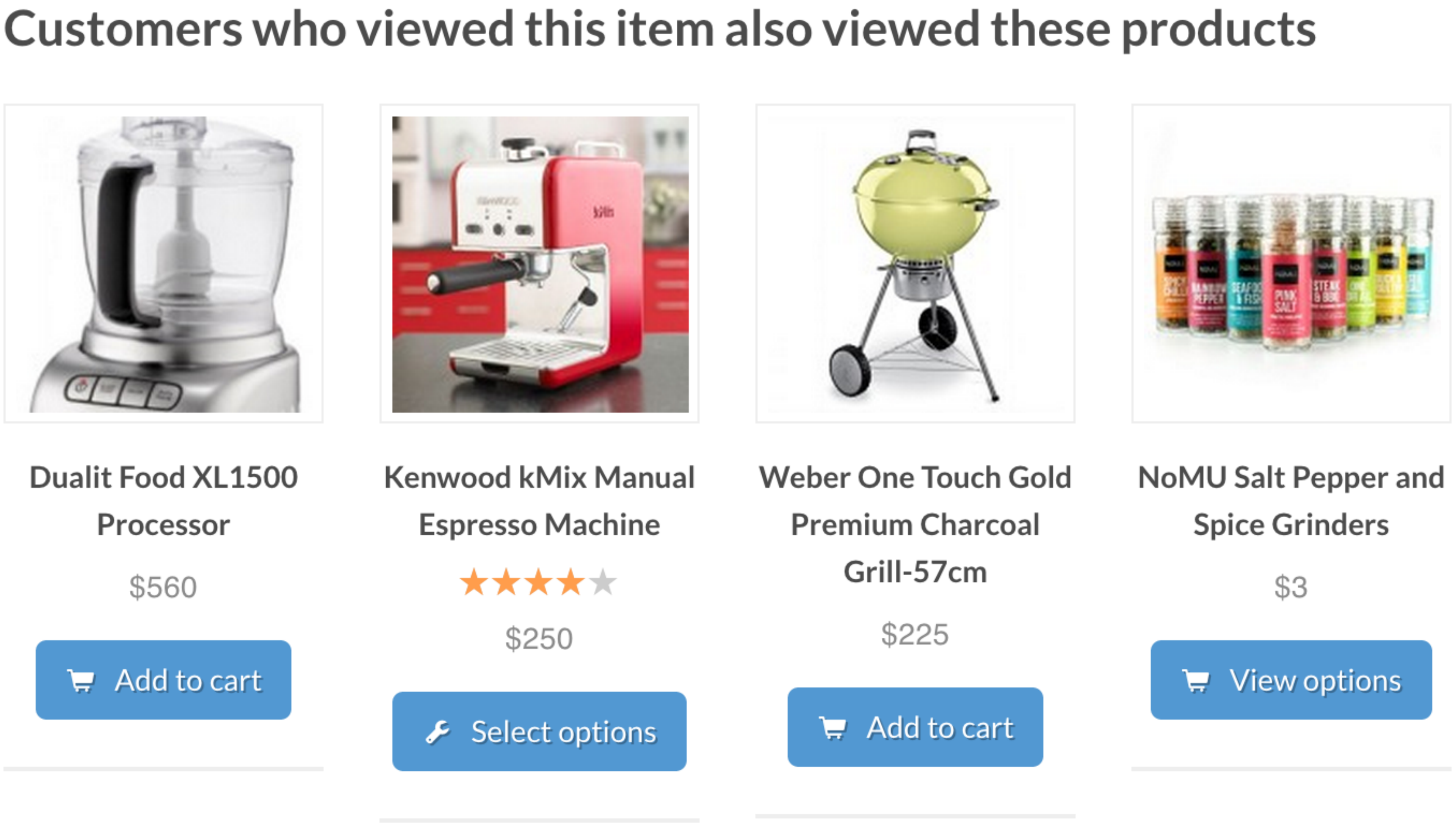
Big Data and Data Products



- What makes “big data” big?
  - Volume?
  - Variety?
  - Velocity?

- Data becomes big when we take one or more large data sets and start to analyze *relationships* between observations

Recommendation Engines



Customers who viewed this item also viewed these products

Dualit Food XL1500 Processor

$560

Kenwood kMix Manual Espresso Machine

★★★★☆

$250

Weber One Touch Gold Premium Charcoal Grill-57cm

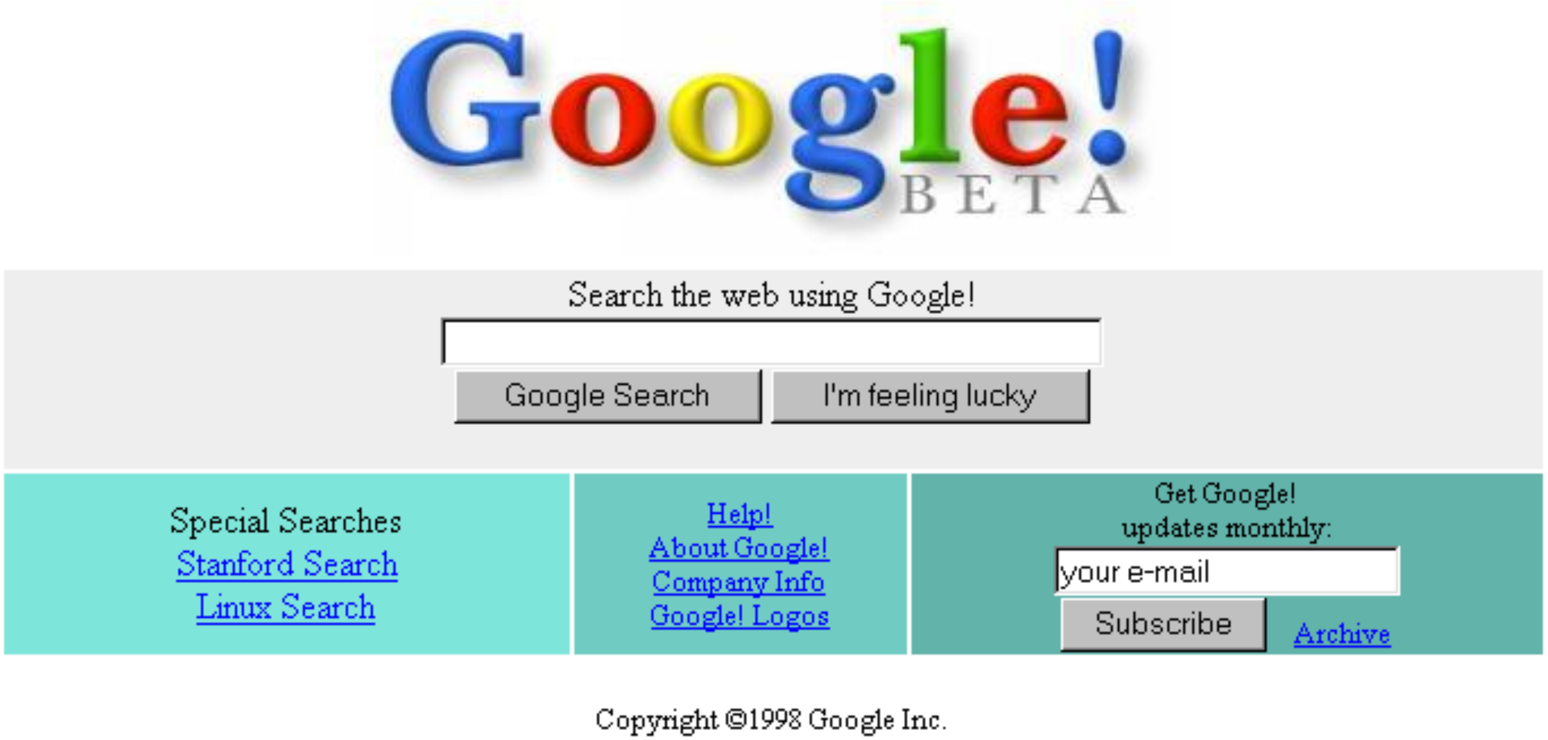
$225

NoMU Salt Pepper and Spice Grinders

$3

cloudera

Search Engines



Copyright ©1998 Google Inc.

Data Products Are The New Competitive Advantage

Designing for Iteration

cloudera

cloudera

Thanks!