Big Data: Data Wrangling Boot Camp
What is Big Data?

Chuck Cartledge, PhD

16 September 2016

Table of contents I

And, why is it interesting?

Big data has emerged as a technology term and trend that is complementary to and considered to be equally as transformational as the cloud computing model.
. . . represented as an "old" or "new" capability depending on the perspective of those defining it, . . .

Lee Badger [5]

Big Data can be characterized by the three V's: volume (large amounts of data), variety (includes different types of data), and velocity (constantly accumulating new data).

Jules. J. Berman [2]

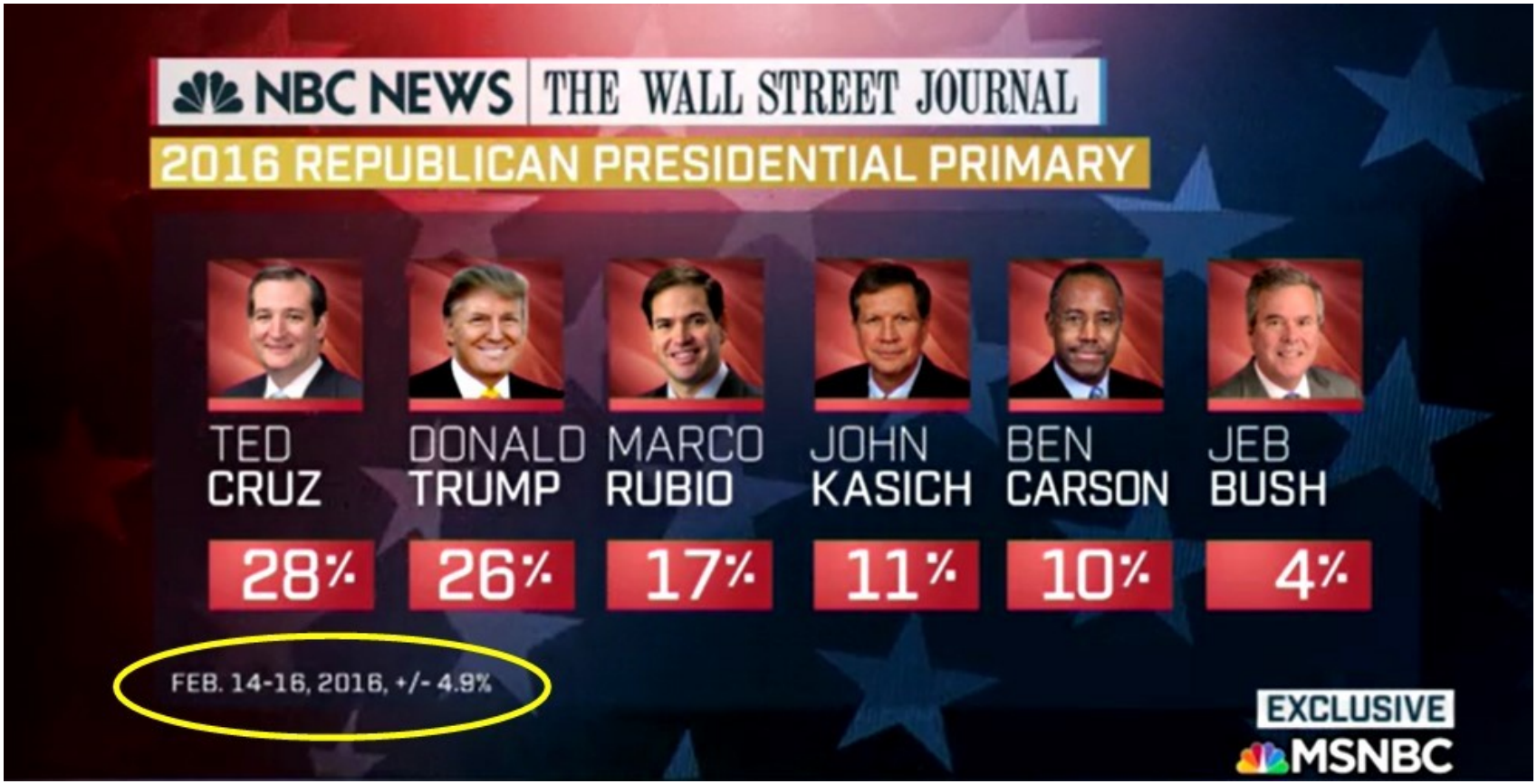
Important ideas from statistics

How "good" an answer do you want?
Questions that need to be answered:

- How accurately do you need the answer?
- What level of confidence do you intend to use?
- What is your current estimate of the answer you're after?

Image from [4].

The greater the tolerance for error, the few samples needed.



If you have some pre-knowledge of the "population" then you only need to sample a very small number of "individuals" to get a good enough answer.[7]

How sampling differs from "Big Data"

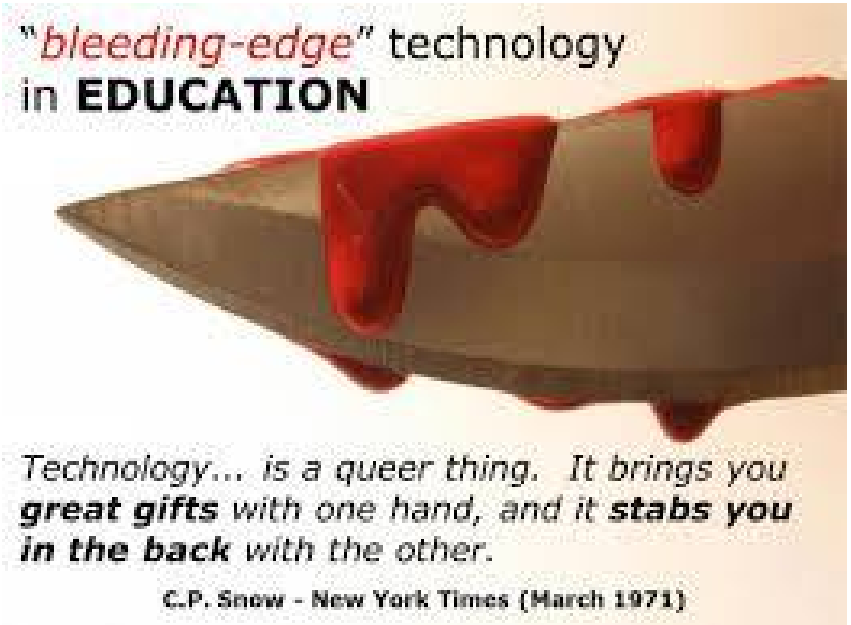
- Sampling – start with a preconceived idea of the outcome
- Sampling – few data points extremely valuable ($n = 1000$)
- Big data – you don't know what the data holds
- Big data – many data points extremely cheap ($n = all$)

Leadership role changes from investigator to data [6].

Large data sets are messy, incomplete, inconsistent, and error prone. Require lots of data munging and **data wrangling**.

We'll be covering virtually "bleeding edge" stuff.

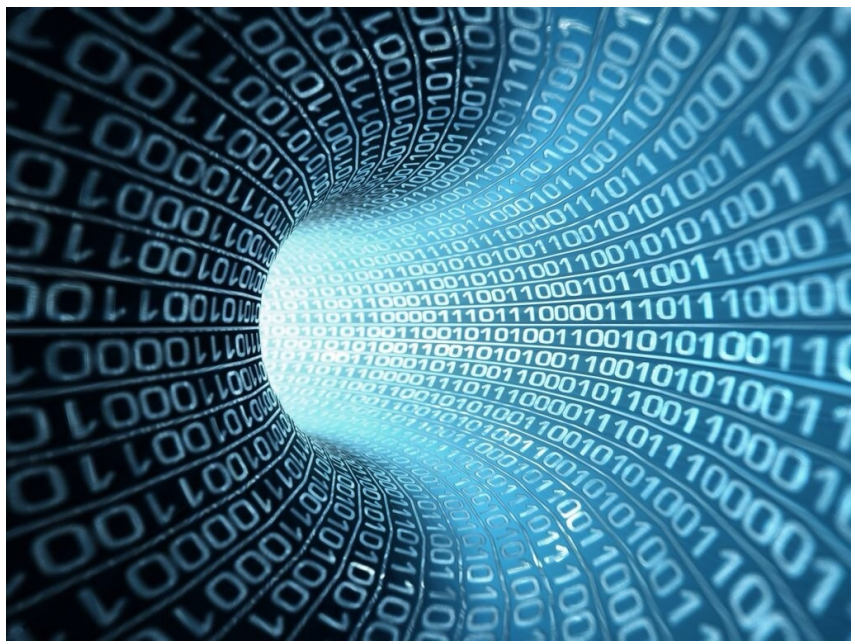
- Data too big for a single machine.
- Processing too long for a single machine.
- Question/analysis is paralizable.



Lots of places, lots of it, and fast.

We are "drowning" in Big Data.

- 230,000,000 tweets per day [3]
- 2,700,000,000 Facebook likes per day [1]
- 100 hours of YouTube video every minute [8]
- Clickstream left on servers



Our wearable devices are contributing to this avalanche of data.

With all this data, what kinds of questions can we ask?

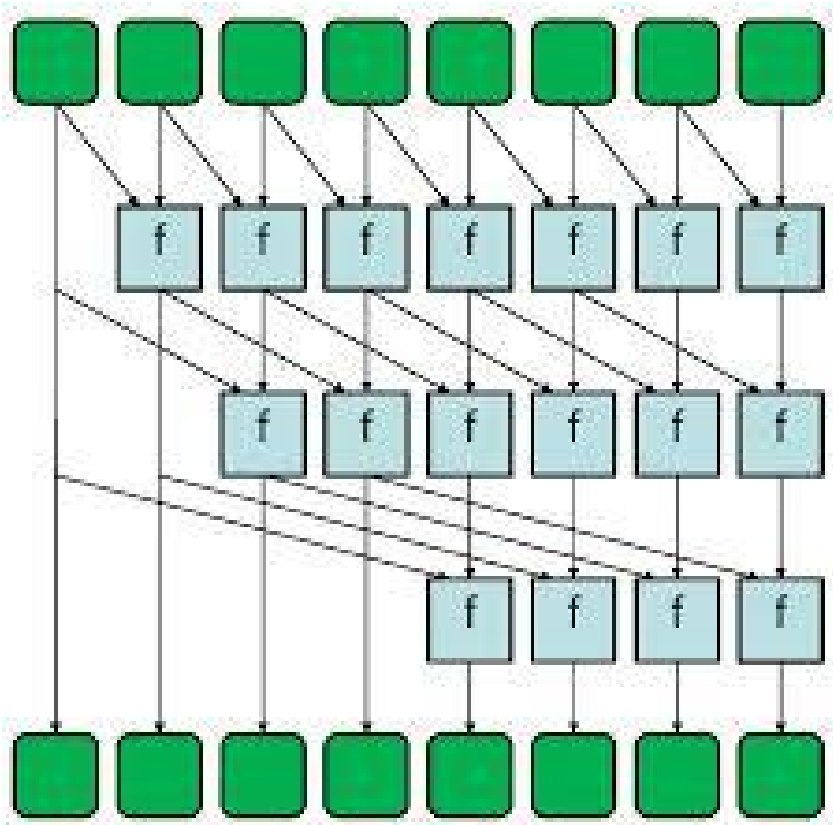
- How is data from one data set related to data in another?
- Are the relationships one-to-one or, one-to-many, or many-to-many?
- Is the data "clean" or not?
- What are we trying to find from the data?

The details of the questions depend on the data and what we are interested in finding.

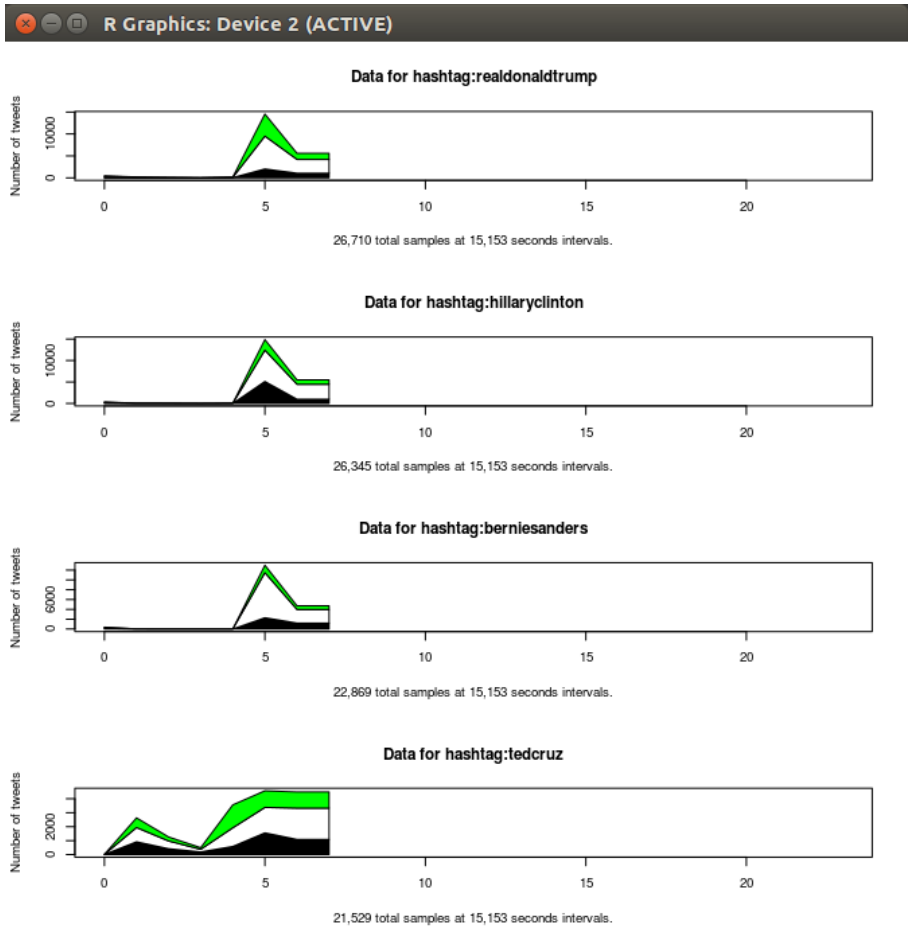
Some questions are easily stated, . . .

Which of these questions are amenable to Big Data processing (and why)?

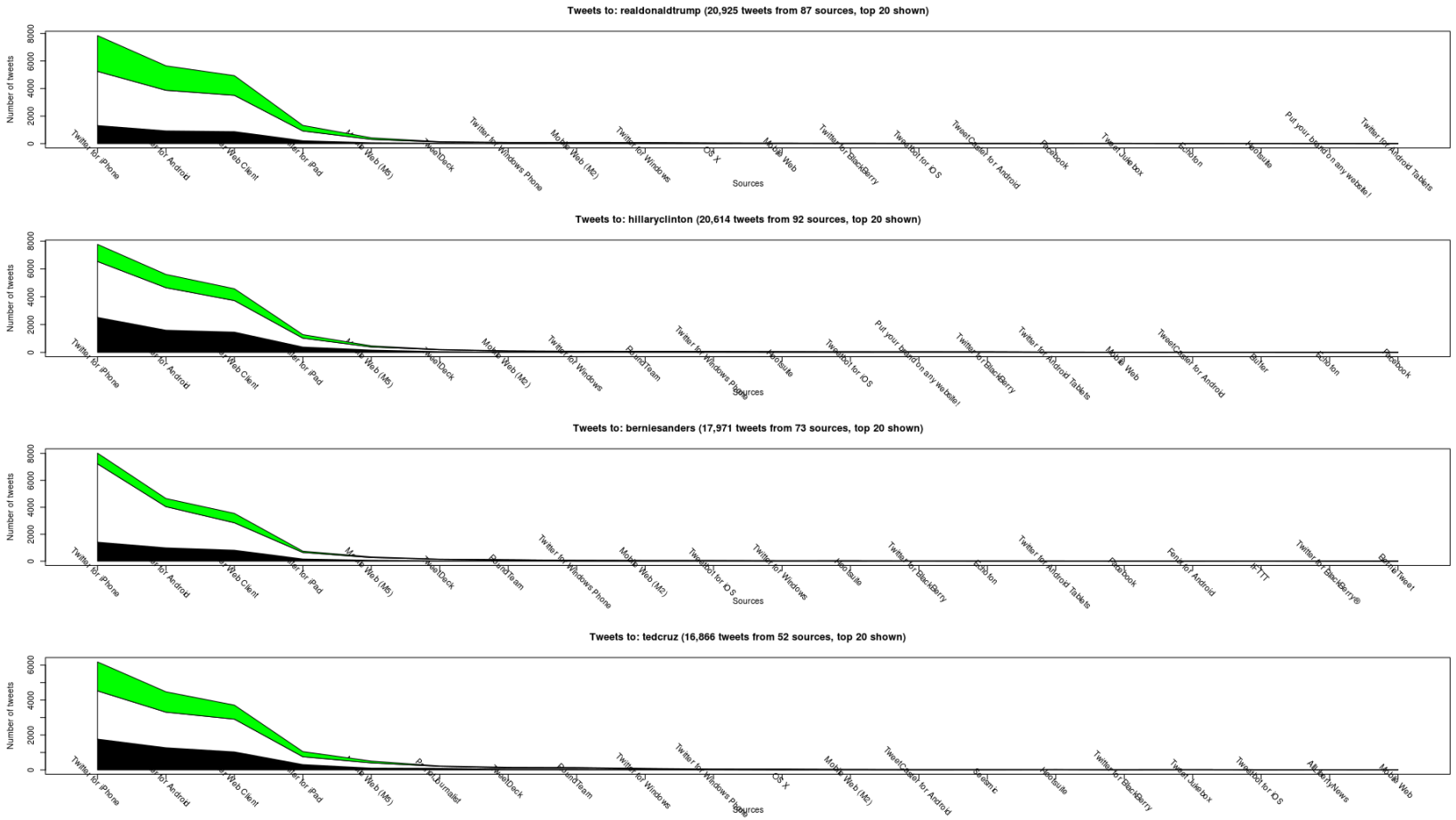
1. $a[i] = b[i] + c[i]$
2. $a[i] = f(b)$
3. $a[i] = a[i-1] + b[i-1]$
4. $a = b + c$



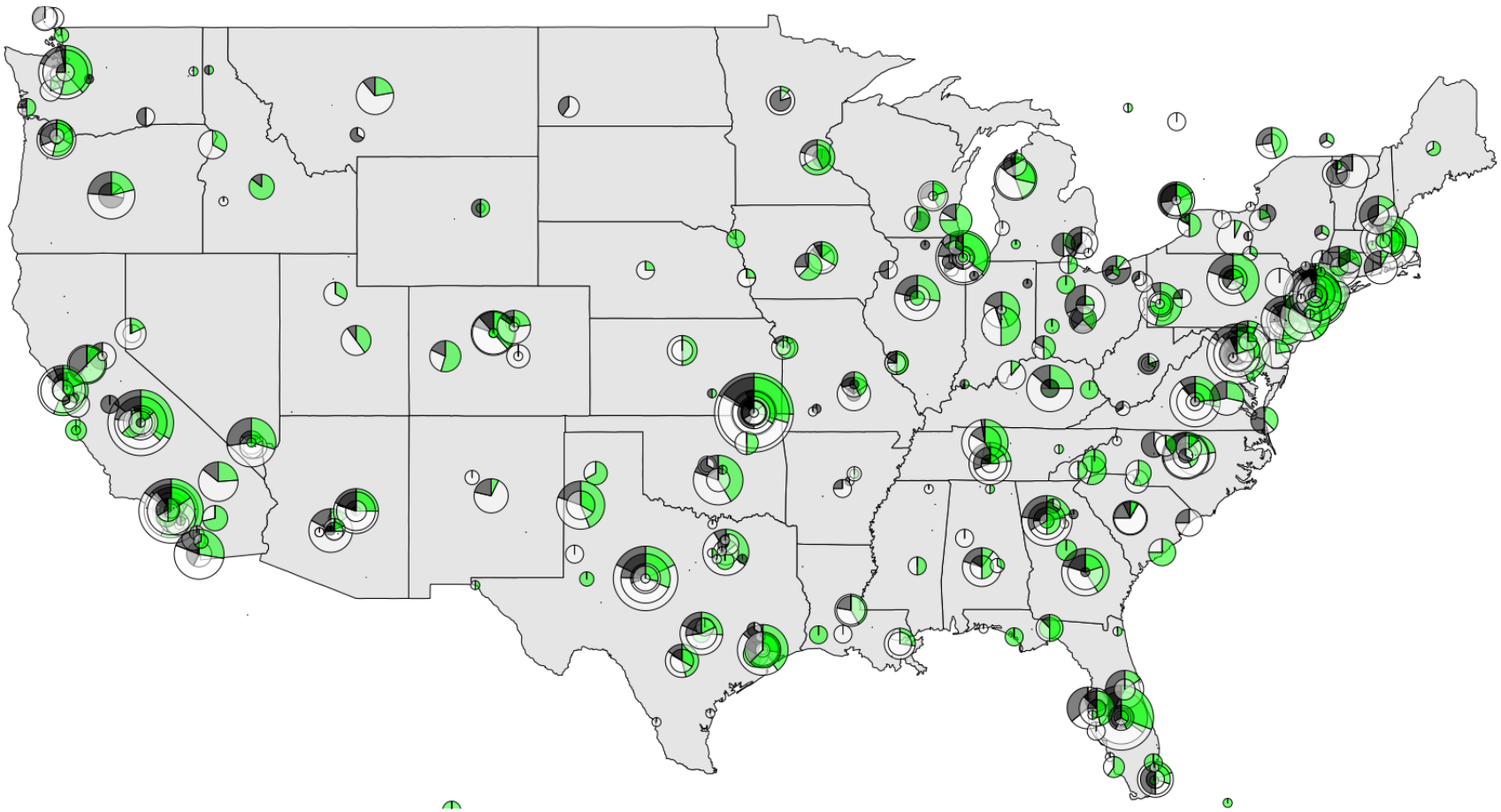
Does the tweet sentiment change over time?



What sends what type of tweet?



Where do tweets come from?



A pragmatic definition

"...big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more."

Mayer-Schönberger and Cukier [6]

A practical definition based on people time.

If:

- your data won't fit into one machine or application, or
- you are waiting too long for an answer

then:

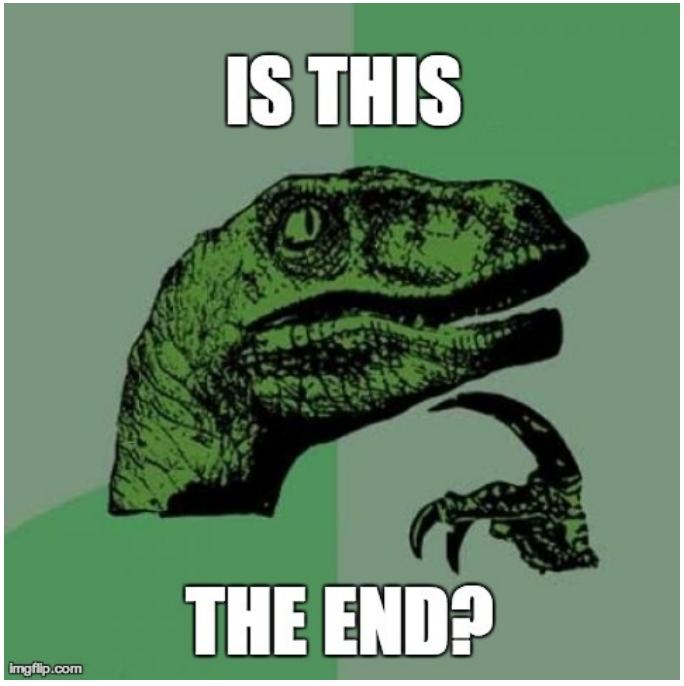
you have a Big Data problem that requires Big Data tools and techniques.

Q & A time.

"'The Answer to the Great Question . . . Of Life, the Universe and Everything . . . is . . . forty-two,' said Deep Thought, with infinite majesty and calm."
Douglas Adams, The Hitchhiker's Guide to the Galaxy

What have we covered?

- Big Data is all around us.
- Big Data is about volume, variety, velocity, and getting answers quickly.
- Some Big Data questions are easy to state, but impossible to answer.



Next: Digging into Big Data overview and concepts.

References I

[1] Anson Alexander, Facebook user statistics 2012 [infographic], ansonAlex.com (2012).

[2] Jules J Berman, Principles of big data: Preparing, sharing, and analyzing complex information, Newnes, 2013.

[3] Joab Jackson, The big promise of big data, Business Software (2012).

[4] James Klurfeld, Making sense of the campaign: The truth about polling, http://drc.centerfornewsliteracy.org/resource/making-sense-campaign-truth-about-polling, 2016.

References II

[5] Robert Bohn Lee Badger, David Bernstein, Us government cloud computing technology roadmap volume i, Tech. report, National Institute of Standards and Technology, 2014.

[6] Viktor Mayer-Schönberger and Kenneth Cukier, Big data: A revolution that will transform how we live, work, and think, Houghton Mifflin Harcourt, 2013.

[7] Mario F Triola, Essentials of statistics, Pearson Addison Wesley Boston, MA, USA:, 2008.

[8] YouTube, Statistics, http://www.youtube.com/yt/press/statistics.html.