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

Chuck Cartledge, PhD

23 February 2018

Table of contents (1 of 1)

What are we going to cover?

We're going to talk about:

- What is Big Data?
- What is Big Data, beyond the marketing hype?
- What sets Big Data apart?
- What is a practical definition of Big Data?

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 [8]

"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 [3]

Doug Laney, META Group

The origin of "Big Data" ideas and definitions.

- Started in the e-commerce Mergers and Acquisitions arena
- Used to explain why traditional Relational Database Management Systems (RDMS) wouldn't scale
- Intended audience was non-technical management

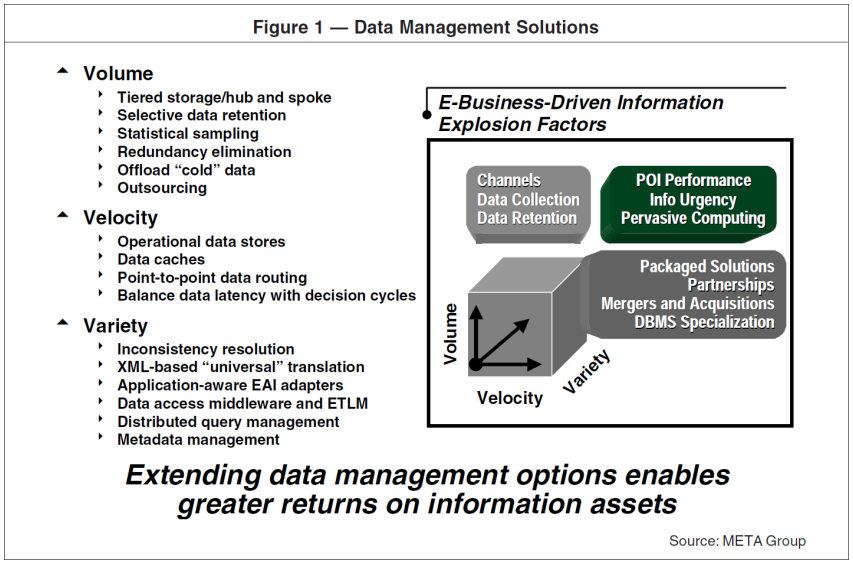


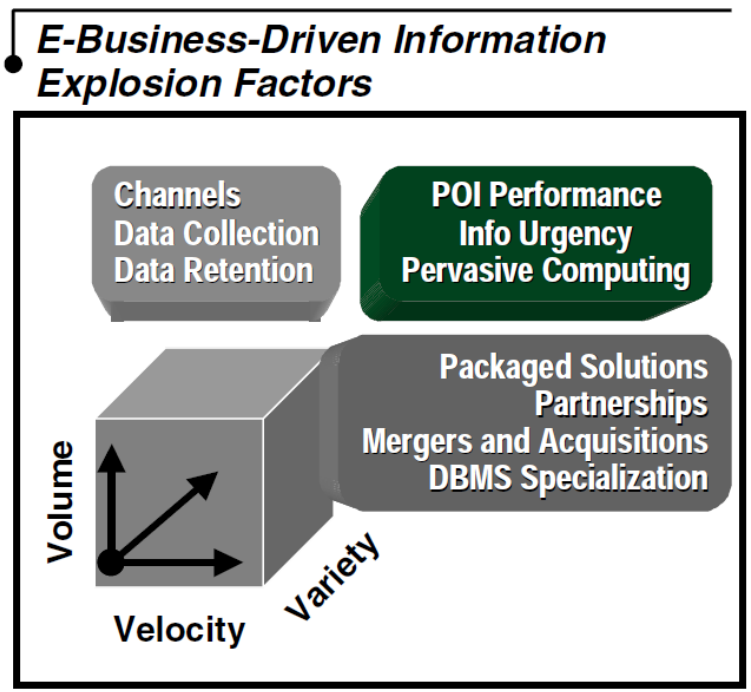
Image from [7].

Take away: traditional RDMS don't/won't scale and different approaches are needed.

Laney's original BD Vs

Figure 1 — Data Management Solutions

- **Volume**
  - Tiered storage/hub and spoke
  - Selective data retention
  - Statistical sampling
  - Redundancy elimination
  - Offload "cold" data
  - Outsourcing
- **Velocity**
  - Operational data stores
  - Data caches
  - Point-to-point data routing
  - Balance data latency with decision cycles
- **Variety**
  - Inconsistency resolution
  - XML-based "universal" translation
  - Application-aware EAI adapters
  - Data access middleware and ETLM
  - Distributed query management
  - Metadata management



Extending data management options enables greater returns on information assets

Source: META Group

Volume — what does it mean for Big Data?

How much is there? And, how do we store it?

- Store relational records?
- Store transactional records?
- How long to keep data available?
- How to access data?
- How to migrate data?

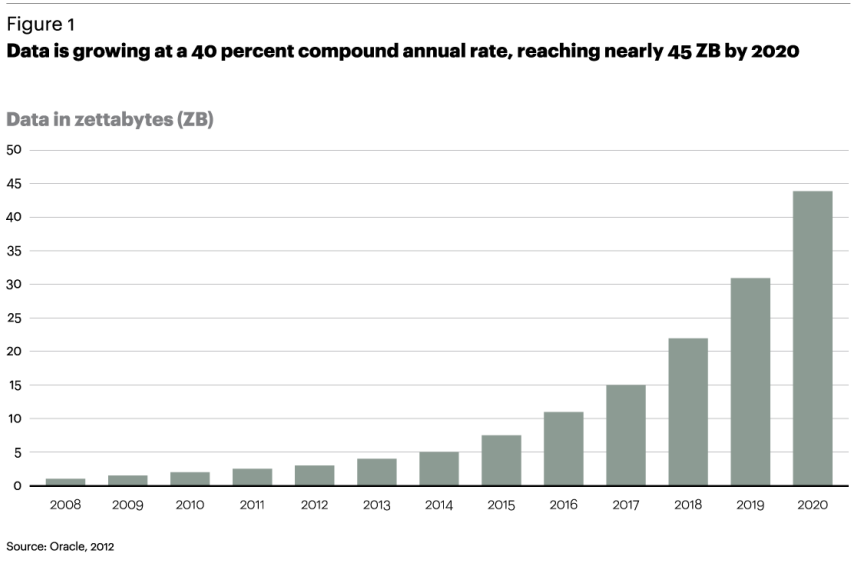
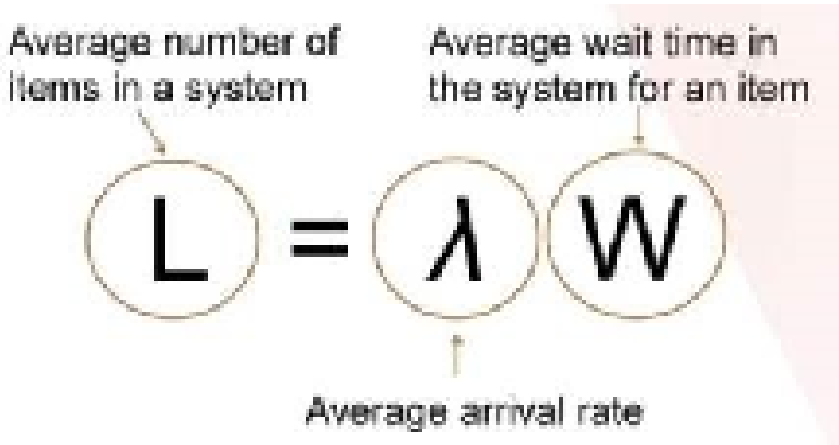


Image from [4].

See http://en.wikipedia.org/wiki/Metric_prefix for list of prefixes.

Velocity — what does it mean for Big Data?

- Frequency of data generation/delivery
- Think of data from a device, or sensor, robots, clicklogs
- Real-time analysis is small (9%) [12].
- Most Big Data analytics is batch



Known as "Little's Law" [9]

Take away: data is generated at a high speed, it must be analyzed before the next set of data is delivered.

Variety — what does it mean for Big Data?

Not all data is the same.

- Data from a multitude of different sources.
- Not all data is useful.
- Data is lost during "normalization"
- Hopefully not important data, when in doubt: keep it somehow
- Gets away from relational databases

The original Vs have been expanded

Lots more Vs.

1. Vagueness
2. Validity
3. Value
4. Variability
5. Variety
6. Velocity
7. Venue
8. Veracity
9. Viability
10. Vincularity
11. Virility
12. Viscosity
13. Visibility
14. Visible
15. Visualization
16. Vitality
17. Vocabulary
18. Volatility
19. Volume

We'll talk about these later.

The Big Data challenges.

- Heterogeneity
- Scale
- Timeliness
- Complexity
- Privacy

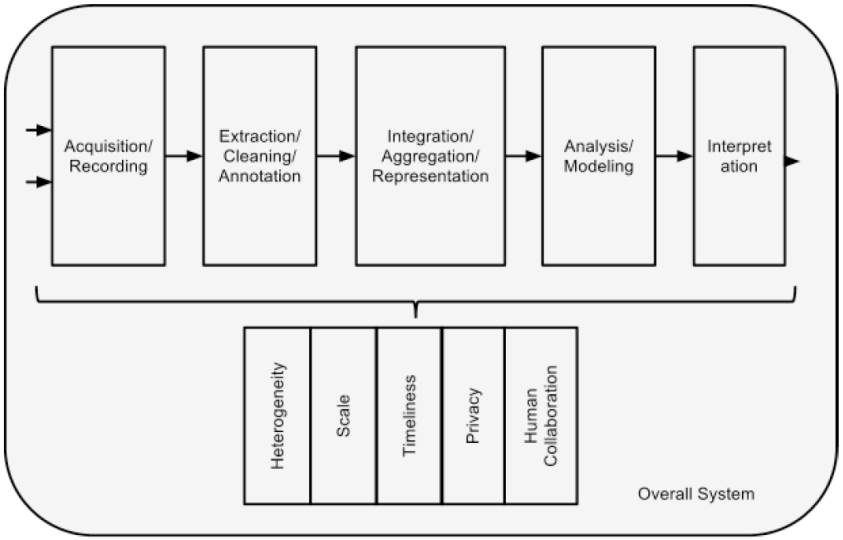


Image from [15].

The Big Data user changes the question[1].

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?

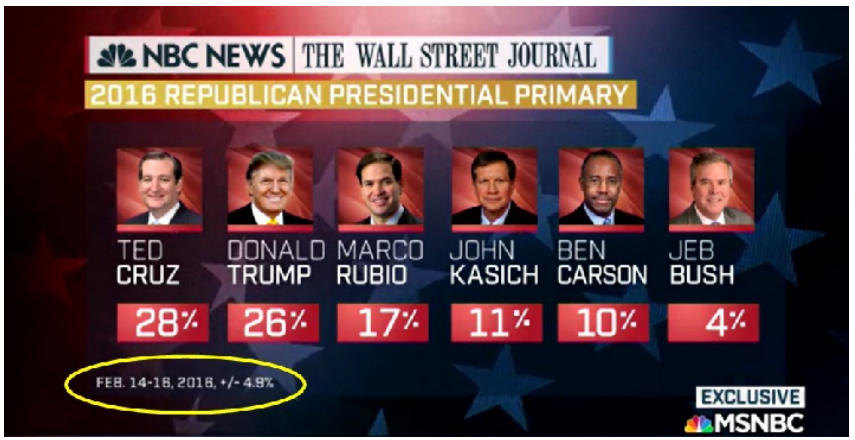
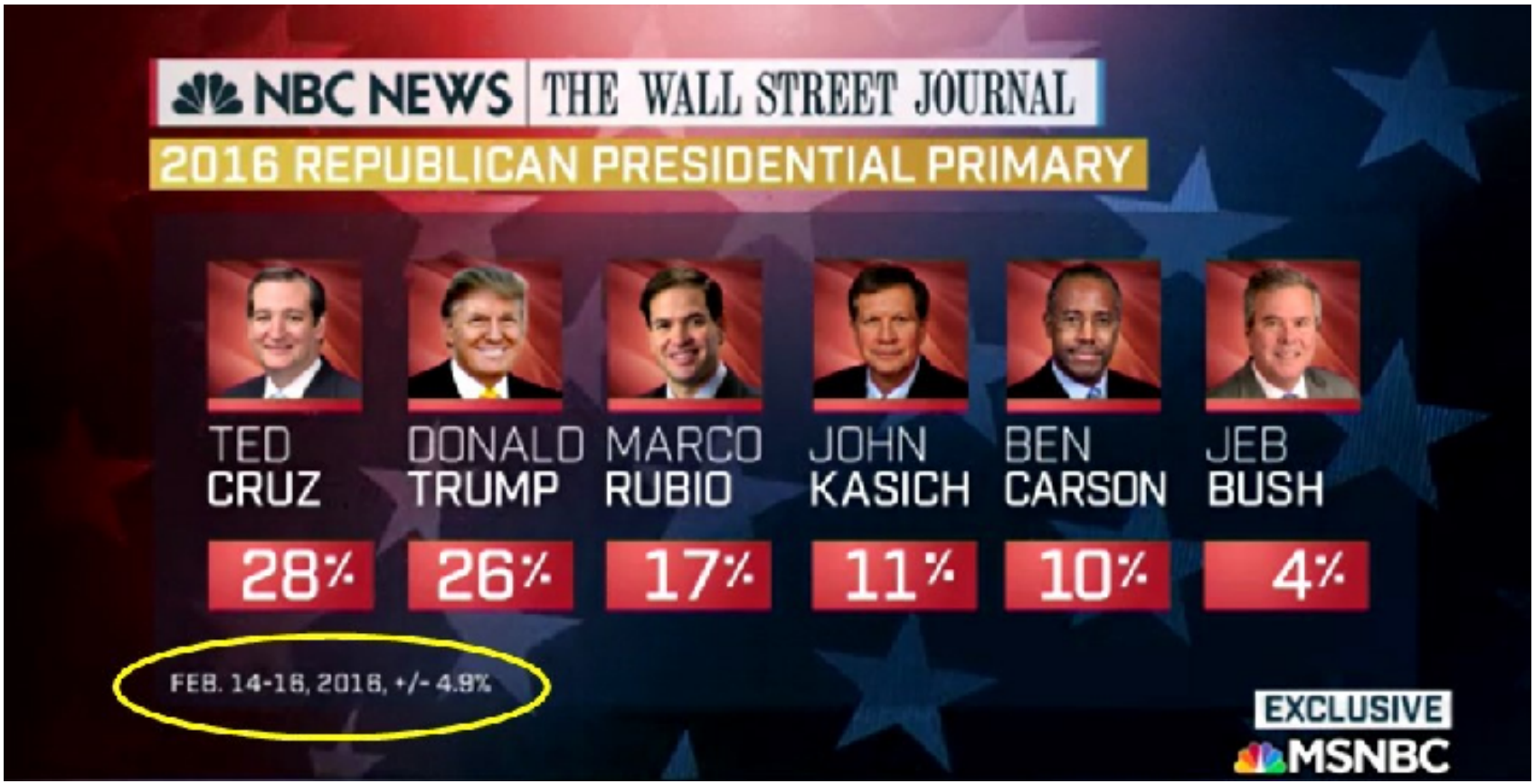


Image from [6].

The greater the tolerance for error, the fewer 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.[16]

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 [10].

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.

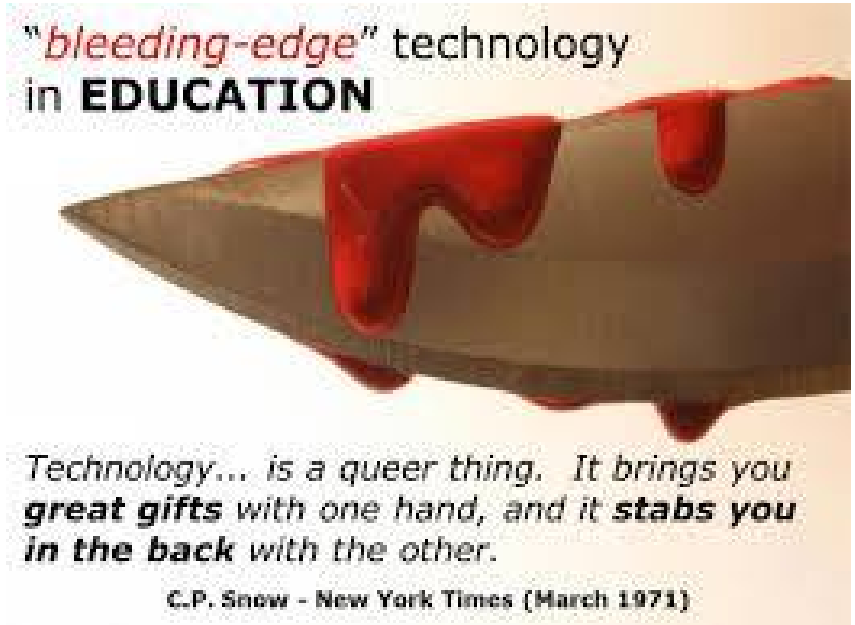
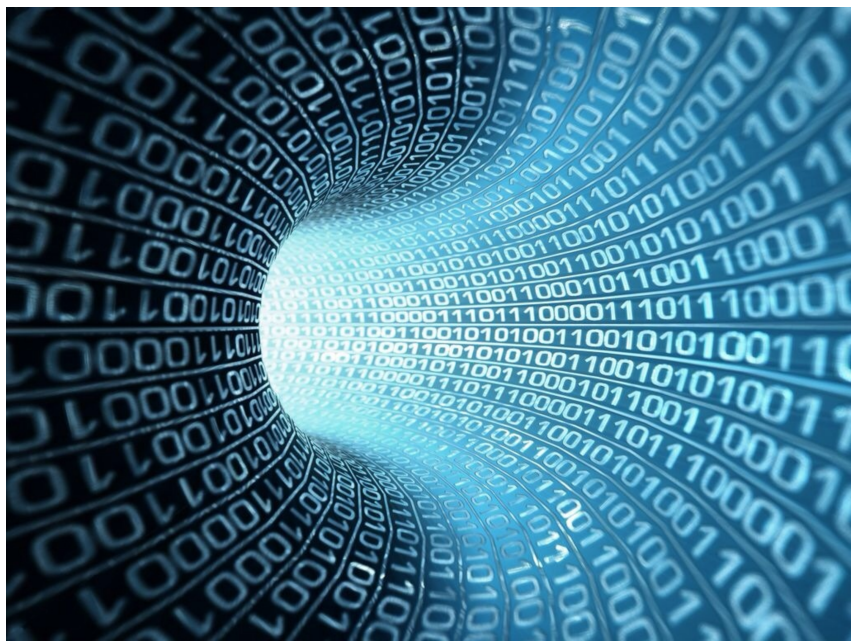


Image from [13].

Lots of places, lots of it, and fast.

We are "drowning" in Big Data.

- 230,000,000 tweets per day [5]
- 2,700,000,000 Facebook likes per day [2]
- 100 hours of YouTube video every minute [17]
- 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?

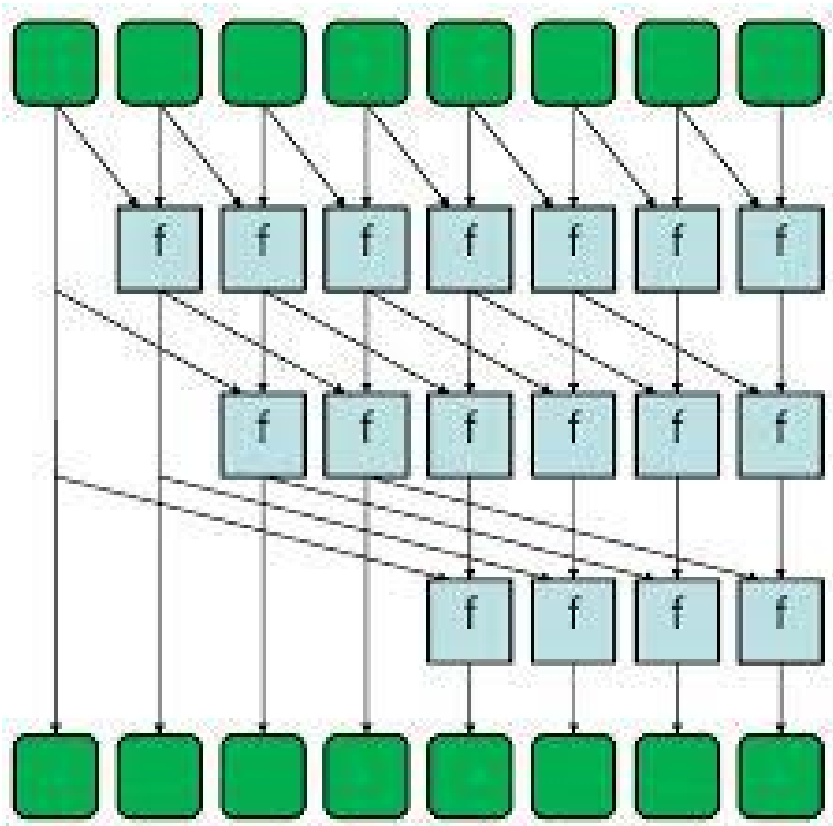
- How is data from one data set related to data in another?
- Are the relationships one-to-one or, one-to-many, or many-to-many?
- Is the data "clean" or not?
- What are we trying to find from the data?

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

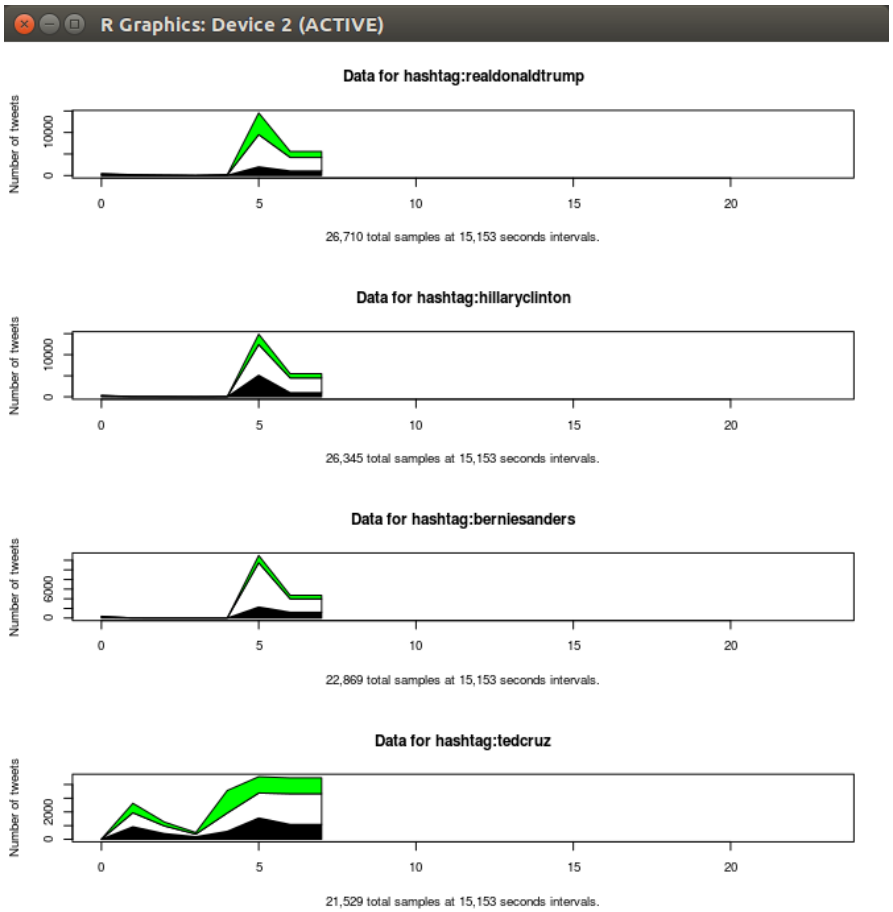
Some questions are easily stated, . . .

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

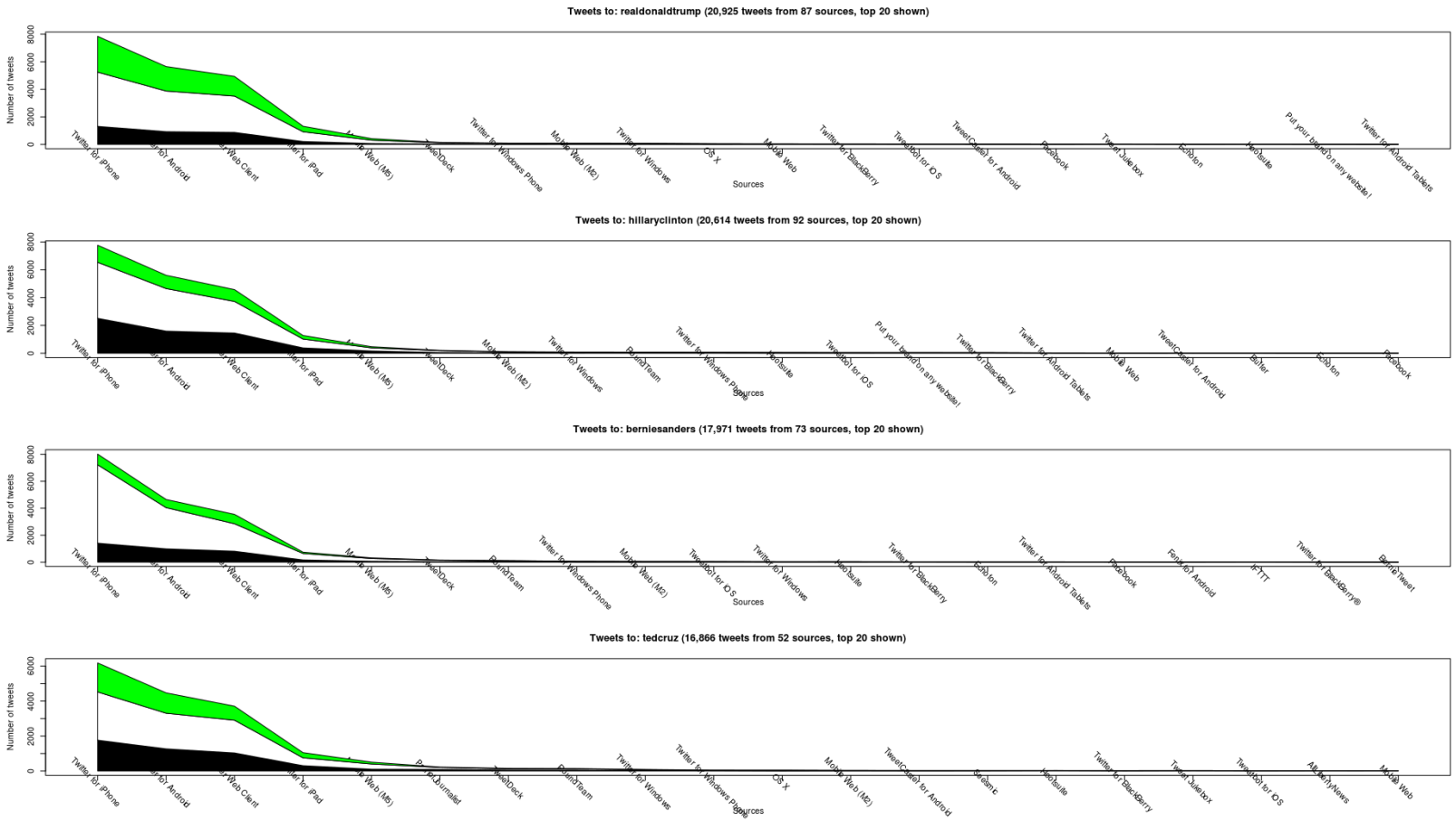
1. $a[i] = b[i] + c[i]$
2. $a[i] = f(b)$
3. $a[i] = a[i-1] + b[i-1]$
4. $a = b + c$



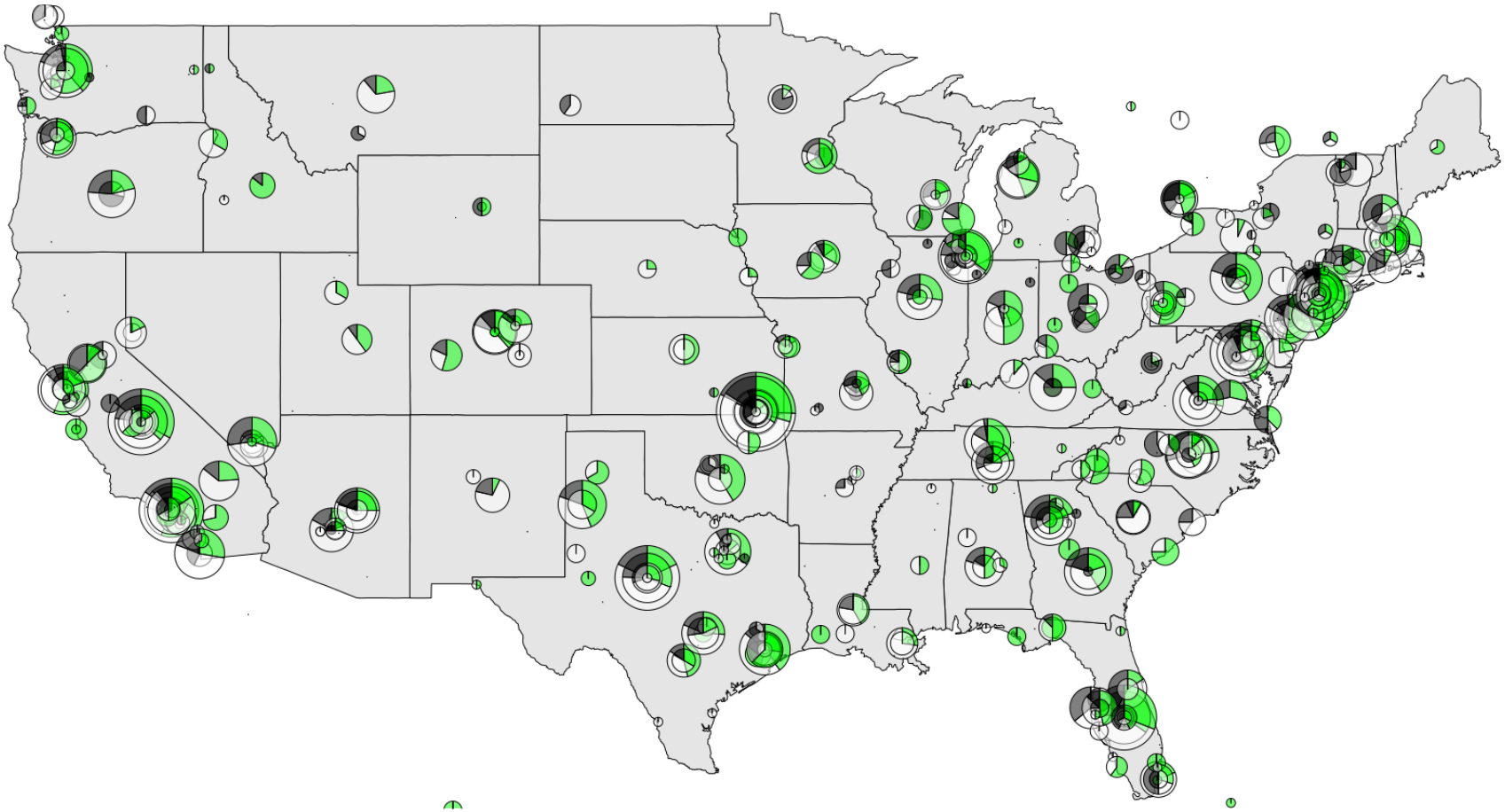
Does the tweet sentiment change over time?



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 [10]

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.

A simple idea in pictures

"In addition to the foundational and translational skills training that students receive, they would also benefit from a better understanding of ethics and social context of data . . ."

NAS [11]

Image from [14].

Ethics in Big Data[11]

- Data confidence – avoiding overconfidence and the inclination to draw stronger-than-appropriate conclusions
- Data context – understand the context of data sets before they are processed
- Fairness – treat all [data] equitably and avoid bias
- Privacy – privacy with respect to how data are collected and analyzed
- Stewardship – supervision of a data set at all stages of existence
- Validity – ensure that the data set contains valid information

This area could be a full credit college course.

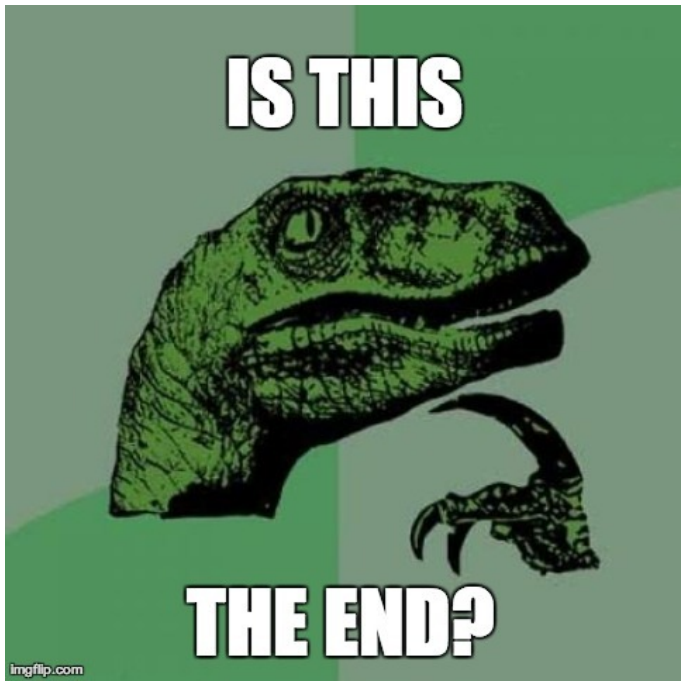
Q & A time.

Q: What is the square root of $4b^2$?

A: To be or not to be.

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.
- Dealing with Big Data can raise real ethical questions



Next: Digging into Big Data overview and concepts.

References (1 of 5)

[1] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, and Michael Franklin, Challenges and Opportunities with Big Data, Purde e-Pubs (2011).

[2] Anson Alexander, Facebook User Statistics 2012 [Infographic], ansonAlex.com (2012).

[3] Jules J Berman, Principles of Big Data: Preparing, Sharing, and Analyzing Complex I Newnes, 2013.

[4] Applied Innovations, Track website visitors, `http://www.appliedi.net/blog/track-website-visitors/`, 2010.

References (2 of 5)

[5] Joab Jackson, The Big Promise of Big Data, Business Software (2012).

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

[7] Doug Laney, 3D Data Management: Controlling Data Volume, Velocity and Varie META Group Research Note **6** (2001).

[8] Robert Bohn Lee Badger, David Bernstein, US Government Cloud Computing Technology Roadmap Volume I, Tech. report, National Institute of Standards and Technology, 2014.

References (3 of 5)

[9] John DC Little, A Proof for the Queuing Formula: L= λ W, Operations Research **9** (1961), no. 3, 383–387.

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

[11] National Academies of Sciences Engineering and Medicine, Envisioning the data science discipline: The undergraduate perspective: Interim report, National Academies of Science, 2017.

[12] Philip Russom, Big Data Analytics, TDWI Best Practices Report, Fourth Quarter (2011).

References (4 of 5)

[13] Andy Shaw, Leading edge or bleeding edge? (reflecting on innovation), http://poopengineer.blogspot.com/2015/04/leading-edge-or-bleeding-edge.html, 2015.

[14] European Union Staff, Ethics, https://ec.europa.eu/programmes/horizon2020/en/h2020-section/ethics, 2017.

[15] Valsoft Services Staff, Challenges and Opportunities with Big Data, http://valsoftservices.com/big-data-implementation/, 2016.

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

References (5 of 5)

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