Is Big Data Bigger than a Bread Box?

Bradley Strauss

Chitika, Inc.

January 14, 2014

The Basic Problem

The basic problem we face is simple to state: the “big” in “big data” is not well-defined, and perhaps “data” is not very well-defined either.

Additionally, there are two common but extreme reactions to the phenomenon of “big data”:

- Big data is new and different and will solve all the world’s problems.
- Big data is nothing but marketing hype from people who don’t like traditional database solutions.

Is "Big Data" about Large Data sets ?

Is the talk about "big data" really concerned with the size of the data? That is, is the most important aspect the large quantity of data available?

- Big data as:
  1. An approach to data analysis
  2. A set of technologies
  3. An approach to decision making [1]
- But all of this begin with the question of how big is "big data," and why is the bigness important?

→ We will find it impossible to separate discussions of data size from the implications of data size on data analysis and technology.

[1] bits.blogs.nytimes.com/2013/06/19/sizing-up-big-data-broadening-beyond-the-internet/

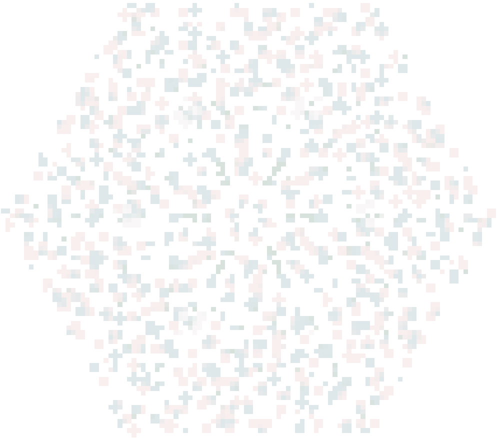
What we’re going to do

Most of this presentation is about the first of the three big data interpretations: big data as an approach to and phenomenon within data analysis.

→ Although not everything about big data analysis is new, there are aspects of big data analysis that are different from traditional data analysis.

- Topics:
    1. Practical and theoretical definitions of large data sets
    2. Examples from advertising technology
    3. Challenges of analyzing large data sets
    4. Toolkit for analysts and data scientists working with big data

First a Warning: Noise as Signal



[2]

[2] Ben Klemens, apophenia.info

The Skeptical View

Argument about the value or interpretation of "big data" often revolves around the second of the two areas above, namely, technologies typical of big data, such as Hadoop and NoSQL databases, and decision-making.

- Small data sets are still very important.
- Practical challenges of drawing useful insights from large data sets prevent organizations from using their data.
- "I was doing databases for 40 years, and now I discover it is called Big Data." —Michael Stonebraker[3]

[3] www.kdnuggets.com/2013/04/big-data-techcon-hadoop-is-not-dead-yet.html

The Optimistic View

- "Big Data" influencing all areas of decision-making.
- Having more data makes us more objective. We can "follow the data."
- Example: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Chris Anderson, *Wired*[4]

→ From the point of view of the data analyst, both of these extreme views are probably not accurate. (To be fair, the optimistic view is probably much more incorrect and dangerous.)

[4] wired.com/science/discoveries/magazine/16-07/pb_theory; see also Peter Norvig's comments.

A Common Lament

- An increasingly common refrain: it is time for “big data” to move beyond Internet advertising.
- This, alas, is not that talk.
- Small warning: there is going to be a little bit of math. But really only a very little.

Practical Definitions

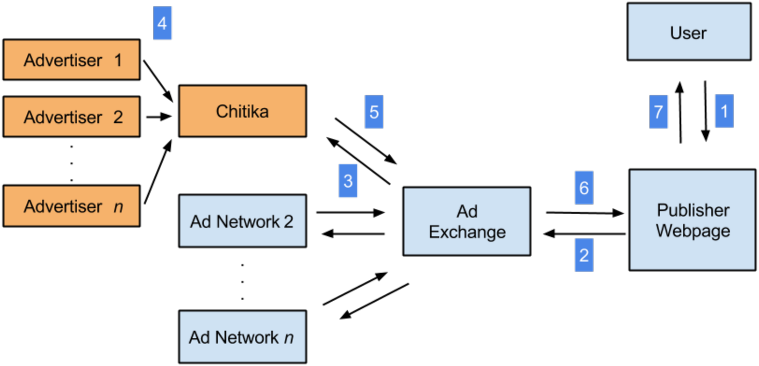
- A lower bound: if you can fit your data in an Excel spreadsheet, it is not big.
  - This data might still be extremely important!
- Data sets that don't fit into memory on your workstation
  - Loose but useful definition
  - Practical lower bound (currently) of maybe 10 gigabytes
- Implication: data management is extremely important
- An upper bound: total amount of data produced globally
  1. How big is this?
  2. An estimate I've read (is it accurate?) is about 10^{18} bytes in two days.
  3. Number of atoms in the observable universe: maybe about 10^{80} atoms.

Theoretical Definitions

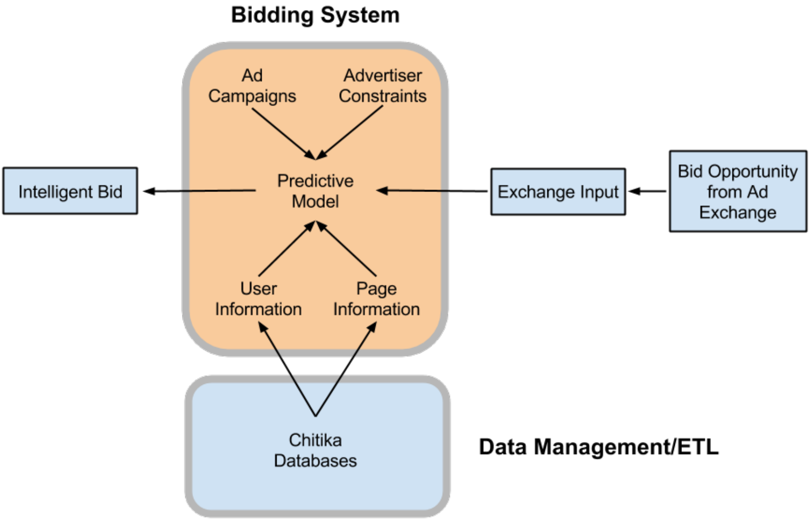
- Big data versus traditional business intelligence: descriptive versus inferential statistics
  - This influences the way you think about your data management needs
- Hadley Wickham:
  - Small data: cognition time $\gg$ computation time
  - Big data: computation time $\gg$ cognition time[5]
- In working with large data sets, you become very concerned, even sometimes obsessed, with things like dimensionality and sparsity.

[5] simplystatistics.org/unconference/

An Example: Real-time Bidding



Real-time Bidding (2)



Challenges of Large Data Sets

- How do statistical problems scale to very large sample sizes?
  - The example of statistical hypothesis testing.
- How do statistical problems scale to high-dimensional spaces? Why is this problematic?
- High-dimensional feature spaces become very, very sparse.

Penguins

Parallelism

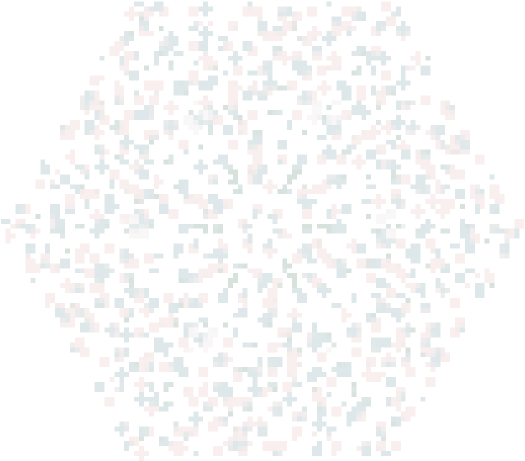
The penguin counting problem: how many penguins of a particular kind live in a given region?

- Dividing up the map and counting the penguins.
- This scales very well, since researchers can "divide and conquer."
- Why does this work?
  - $(1+2)+3=(1+3)+2$
- Many important problems scale in this fashion, for example, search results.
- MapReduce and Hadoop

Dimensionality Problems

- What if I want to know, given data I've collected, whether a penguin is of a particular kind?
  - Number of features: feather features, beaks, eyes, feet, behaviors.
  - How does this scale?
    - How large is 1000!?
- Or, what if I don't know where the penguins are?
- This is actually similar to the RTB problem above.

Noise as Signal



Skills for Careers in Data Science

- Key questions:
  1. Can you sit down in front of a new data set and get started?
  2. Do you understand the model you're using, why you're using it, and the potential pitfalls?
  3. Can you code your solution?
- Very practical suggestions:
  1. Take a statistics class (or several)
  2. Take a computer science class (or lots)
  3. Learn to write computer programs
     - Lots of options, but here's one suggestion: htdp.org