Big Data

CAC170
Slides by Dr. Jeff Gray, University of Alabama
Edited by Dr. Amber Wagner and Dr. Anthony Winchester

Just for Fun – Check out this game

- http://en.akinator.com/

How did it work?

- How does the Akinator work?
- How is it able to predict your thoughts based on questions?
- Data…and a lot of it!

IBM's Watson

- Watson is a cognitive technology that can parse large amounts of natural language for Q&A
- In 2011, Watson competed against top Jeopardy champions for $1M prize. Watson had access to 200 million pages of text (4 terabytes)
- By having access to so much data, it was able to always find the answer. There are specific techniques for searching vast data quickly – something used in databases.

IBM's Watson

- Watson is now being used for cancer research (analyzing broad literature on medical research) among other areas
- Watson is an Artificial Intelligence that IBM makes available for businesses
- https://www.ibm.com/watson/ai-stories/index.html

What is Big Data?

- We create 2.5 quintillion bytes of data every day
  - From various kinds of sensors, media sites, transaction records, etc.
- Datasets are growing too large to manage with common software tools
  - This makes it difficult to derive meaning from them

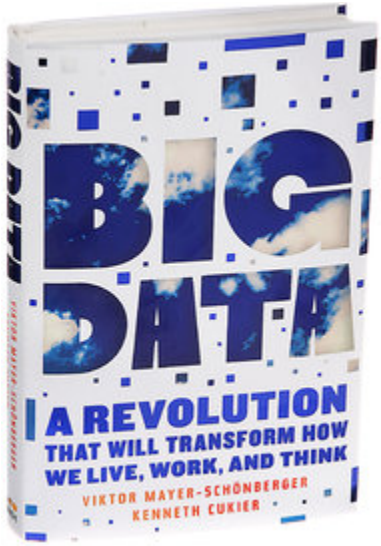
Why is it important?

Analyzing Big Data can provide us with important insights

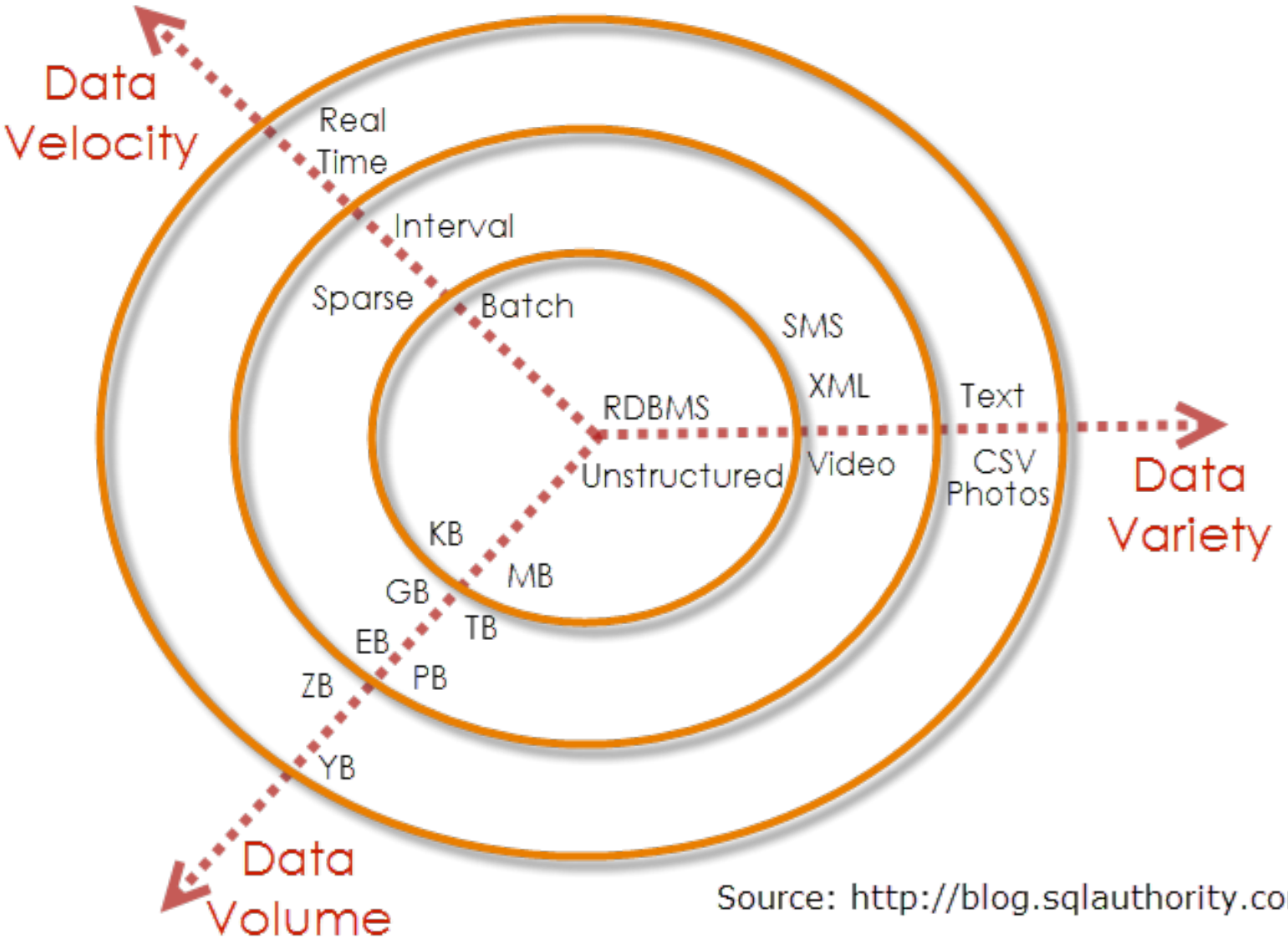
- When big data is properly captured and analyzed, it can provide important insights
- It has applications in:
  - Retail
  - Sports
  - Finance
  - Manufacturing
  - Disease control
  - And many more…

The 3 V's

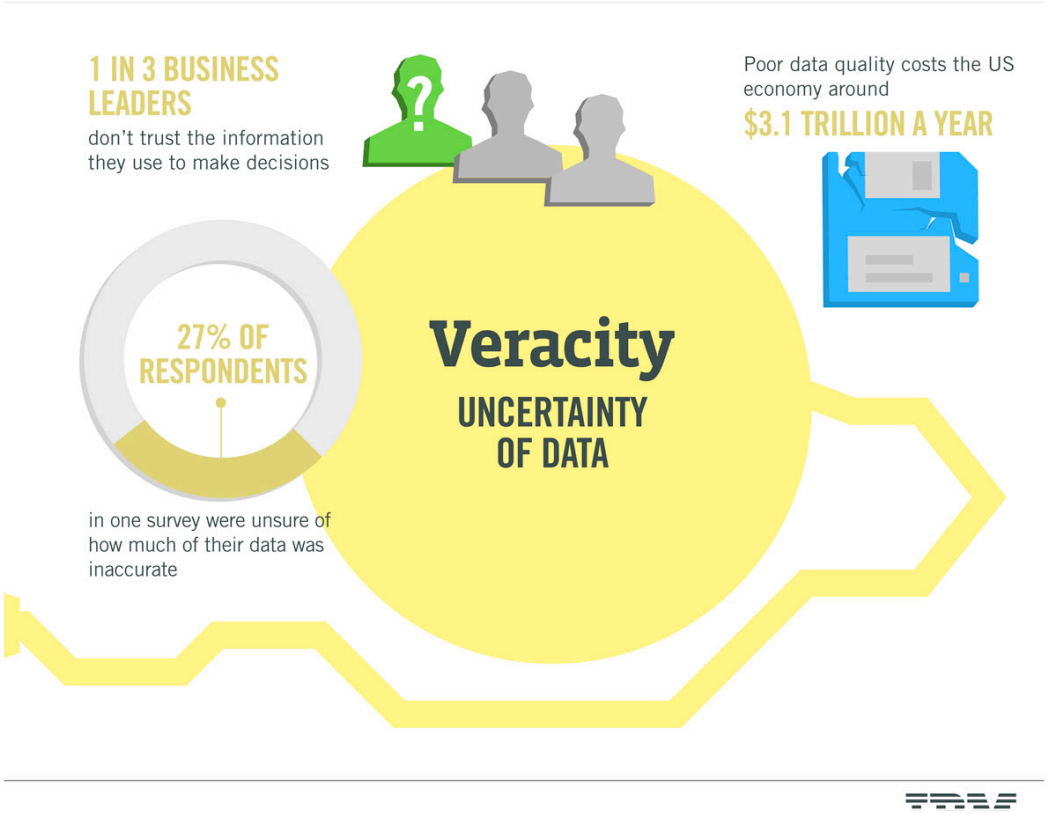
- **Volume** – the amount of data that is available for processing is overwhelming
- **Variety** – a large number of data types are now available, and data can be stored in a wide variety of formats, which makes processing it even more challenging
- **Velocity** – data is generated and captured at a much higher rate than it has been in the past



3Vs of Big Data



Source: http://blog.sqlauthority.com


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

Emerging Trends in Big Data (Cukier et al.)

- "Datafication" of many events/activities
  - Turning all aspects of life into data
  - Sometimes, unanticipated future use
- Change in how we approach data
  - Collecting large volume, rather than smaller samples; still need statistics, just not small samples
  - Moving away from need of pristine and curated samples, and accepting "messy" data (occasional inaccuracies are ok)
    - Benefits of using more data outweigh cost of using smaller data
  - Accepting the answer of "what" in place of "why"
    - Correlation rather than causation is often necessary when evaluating results; meant to inform rather than to explain
  - More room for human intuition will be needed

How is it used?

- Netflix and Amazon use big data to improve user recommendations

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

- The school system in Mobile County Alabama used big data to increase the graduation rate from 55% to 70%

- Domino's used big data to determine that customers order more pizza when it's raining. They now base their ad campaigns around local weather patterns

How is it used? (cont'd)

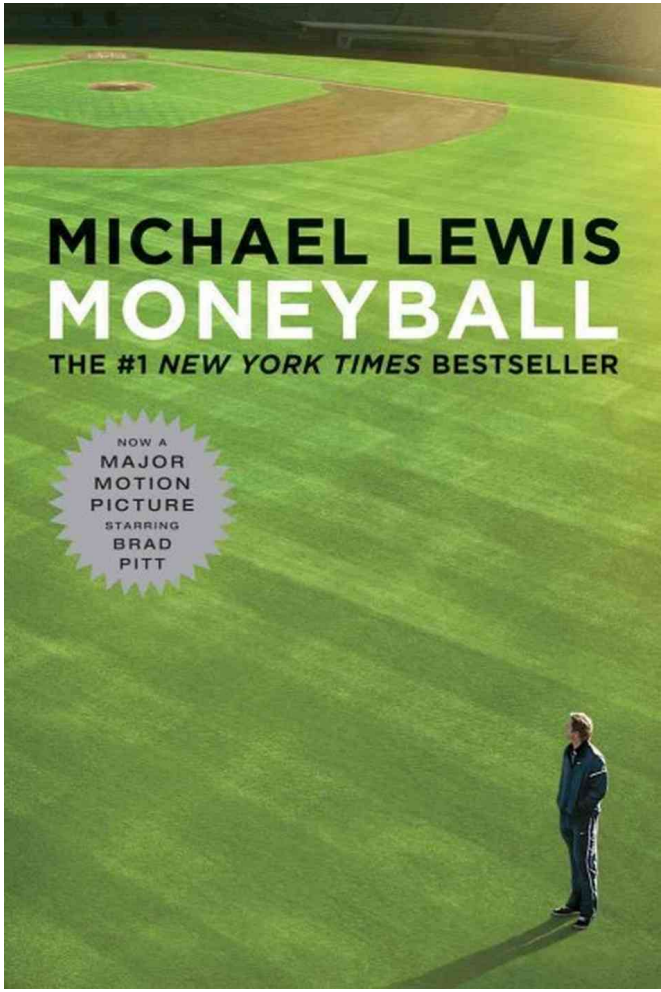
- Mitsui Knowledge Industry (Japan) uses big data to perform complex genome analysis that allows for customized cancer treatments

- A twitter algorithm can predict when a healthy person will get sick 8 days in advance, with 90% accuracy

- Big Data can help police predict where and when crimes will occur
  - Ethical Issues: More like "Big Brother"
- Big Data has helped astronomers identify the first Earth-like planets outside of our solar system

Moneyball Example

- Collected wisdom of baseball insiders (coaches, owners) is flawed and incorrect at times
- Used analytics over large data sets to determine more important factors in talent appraisal
- Oakland A's 2002 team were more competitive than teams with 3 times their salary budget

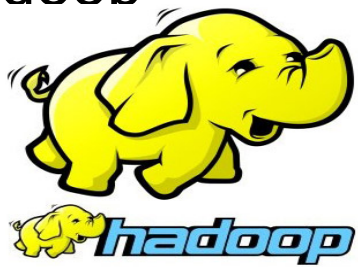


Big Data Processing Steps

1. Identify appropriate data source and form questions
2. Extract data source into format supported by underlying tools
3. Normalize data (remove redundancies, irrelevant details)
4. Import data into tool
5. **Perform analysis**
6. Visualize results

Tools for Processing Big Data

- **Microsoft Excel** - the well-known and trusted excel can be used for processing large data sets
- **Hadoop** – the most well-known Big Data tool, provided by Apache, requires extensive programming knowledge to set up and use
- **SAS** – provides a more intuitive interface and better graphical representations of data
- **Google's Cloud Machine Learning Engine**– takes advantage of machine learning to extract meaning from data
- **BitDeli** – lightweight, easier to use version of Hadoop
- **And many more…**

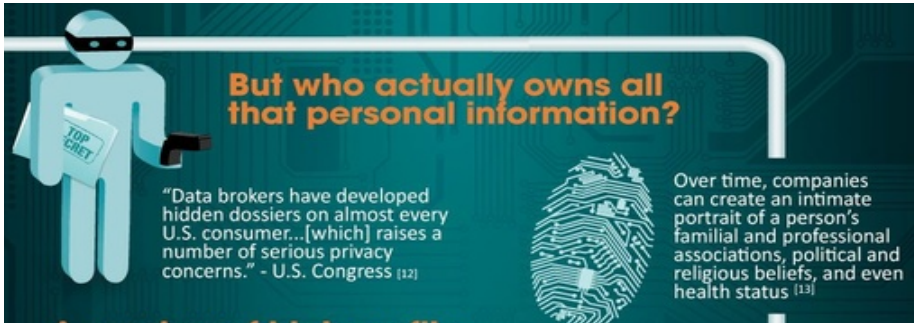


Data Sources

- Sites that offer users the option to download data in a variety of formats:
  - https://data.govloop.com/
  - https://data.austintexas.gov/
  - https://data.cityofchicago.org/
  - https://data.seattle.gov/
  - https://data.medicare.gov/
  - https://data.sfgov.org/
  - https://data.sunlightlabs.com/
- For a list of links to other sites with free data:
  - http://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public

Ethical Concerns
Is big data "creepy"?

- The collection of big data allows corporations to collect large amounts of data on their customers



- As of yet, there are few restrictions on how companies may use that data
- It can be:
  - Sold to other companies
  - Sold to government agencies
  - Used to de-anonymize individuals

Documentary Recommendation
http://tacma.net/

Big Data Gone Wrong

- There are some very notable cases of big data being used in a way that violates users' privacy
  - Target outing a teenager's pregnancy
    - http://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2
  - In 2012 Google spent $22.5 million on a settlement over allegations that they secretly tracked users' web surfing via Apple's safari browser
  - In 2012, Facebook paid $20 million to settle a lawsuit that alleged that Facebook users' pictures were used without their knowledge to endorse products they had 'liked'
  - Verizon received a "non-final" rejection for a patent for a technology that would serve targeted ads to users based on what they do or say in front of their television
  - In 2013, the revelation of the NSA using Big Data for national security concerns led to an international outcry (See Time articles)

Activity

- Find a partner
- Find a data source that interests you from slide 17
- Complete the following:
  - Write a brief summary about what the data represents
  - What are the “variables” in the data?
  - List five questions you want to answer about the data. What information do you think you need to answer each question?
  - Without performing an analysis, what do you think the answers to the questions will be?