Introduction to Big Data

Jonathan Wayne Korn @ Columbia University

2021-04-30

What is Big Data?

- There are some things that are **so big** that they have implications for everyone, whether we want it or not.
- **Big Data** is one of those thongs, and is completely transforming the way we do business and is impacting most other parts of our lives.

Basic Idea

- Behind the phrase ***'Big Data*** is that everything we do is increasingly leaving a digital trace *(or data)*, which we *(and others)* can use/analyze.
- Big Data therefore refers to our ability to make use of the ever-increasing volumes of data.

The Scope of Big Data

- *"From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days . . . and the pace is accelerating."*
  - Eric Schmidt Executive Chairman, Google

Activity/Transaction Data

- Simple activities/transactions like listening to music or reading a book are now generating data.
  - Digital music players and eBoks collect data on our activities/transactions.
  - Your smart phone collects data on how you use it and your web browser collects information on what you are searching for.
  - Your credit card company collects data on where you shop and your shops collect data on what you buy.
  - It is hard to imagine any activity/transaction that does not generate data.

Conversation Data

- Our conversation are now digitally recorded *whether in text or speech*.
- It all started with emails, but nowadays most of our conversations leave a digital trail.
- Just think of all the conversations we have on social media sites.
- Even many of our phone conversations are now digitally recorded.

Photo and Image Data

- Just think about all the pictures we take on our smart phones or digital cameras.
- We upload and share 100s of thousands of them on social media sites every second.
- The increasing amounts of CCTV cameras take video images and we upload hundreds of hours of video images to YouTube and other sites every minute.

Sensor Data

- We are increasingly surrounded by sensors that collect and share data.
- Take your smart phone, it contains a global positioning sensor to track exactly where you are every second of the day.
  - It includes an accelometer to track the speed and direction at which you are traveling.
- We now have sensors in many devices and products.

Internet of Things Data

- We now have smart TVs that are able to collect and process data, we have smart watches, smart fridges, and smart alarms.
- The internet of Things, or Internet of Everything connects these devices so that
  - *i.e.* the traffic sensors on the road send data to your alarm clock whcih will wake you up earlier than planned because the blocked road means you have to leave earlier to make your 9am meeting...

Datafication

- With the datafication comes big data, which is often described using the four V_s:
  (1) Volume
  (2) Velocity
  (3) Variety
  (4) Veracity

Volume

- refers to the vast amounts of data generated every second.
- We are not talking Terabytes but Zettabytes or Bronobytes.
- If we take all the data generated in the world between the beginning of time and 2000, the same amount of data will soon be generated every minute.
- New bug data tools use distributed systems so that we can store and analyze the data across databases that are dotted around anywhere in the world.

Velocity

- refers to the speed at which new data is generated and the speed at which the data moves around.
- Just think of social media messages going viral in seconds.
- technology allows us now to analyze the data while it is being generated *(sometimes referred to as in memory analytics)*, without ever putting it into databases.

Variety

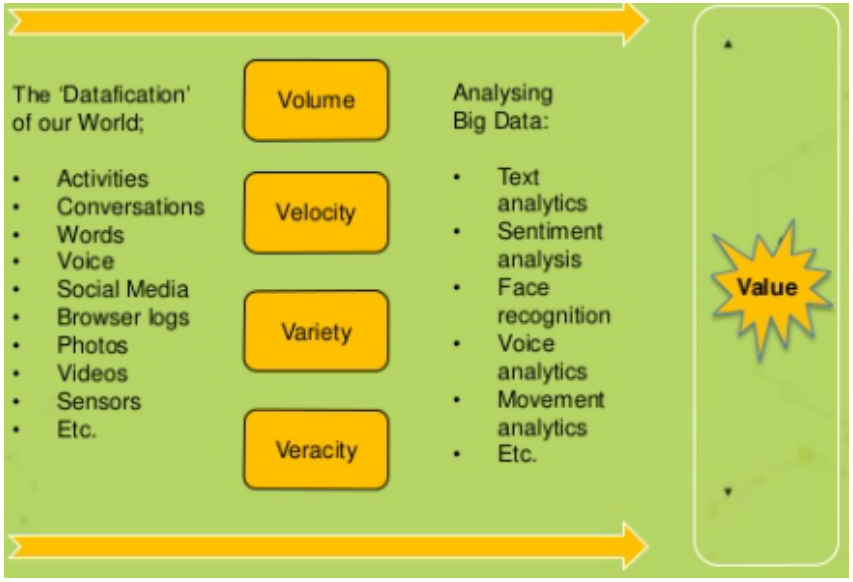
- refers to the different types of data we can now use.
- In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data.
- In fact, 80% of the world's data is unstructured *(text, images, video, voice, etc)*.
- With big data technology we can now analyze and bring together data of different types such as *messages, social media conversations, photos, sensor data, video, or voice recordings.*

Veracity

- refers to the messiness or trustworthiness of the data.
- With many forms of big data quality and accuracy are less controllable *(just think of Twitter posts with hash tags, abbreviations, typos, and colloquial speech as well as reliability and accuracy content)* but technology now allows us to work with this type of data.

Turning Big Data Into Value

- The datafication of our world gives us unprecedented amounts of data terms of Volume, Velocity, Variety, and Veracity.
- The latest technology such as cloud computing and distributed systems together with the latest software and analysis approaches allow us to leverage all types of data to gain insight and add value.



Use Case Example of Big Data ~ Better Understand and Target Customers:

- To better understand and target customers, companies expand their traditional datasets with *social media data, browser, text analytics or sensor data* to get a more complete picture of their customer.
- The big objective, in many cases, is to create predictive models.
- Using big data, Telecom companies can now better predict customer churn; retailers can predict what products will sell, and car insurance companies understand how well their customers actually drive.

Use Case Example of Big Data ~ Understand and Optimize Business Processes:

- Big data is also increasingly used to optimize business processes.
- Retailers are able to optimize their stock based on predictive models generated from *social media data*, *webs search trends and weather forecasts.*
- Another example is supply chain or delivery route optimization using data from geographic positioning and radio frequency identification sensors.

Use Case Example of Big Data ~ Improving Health:

- The computing power of big data analytics enables us to find new cures and better understand and predict disease patterns.
- We can use all the data from smart watches and wearable devices to better understand links between lifestyles and diseases.
- Big data analytics also allow us to monitor and predict epidemics and disease outbreaks, simply by listening to what people are saying, *i.e. "Feeling rubbish today - in bed with a cold"* or searching for on the internet, *i.e. "cures for flu"*.

Some Examples of Big Data

A place i think you will like is Kaggle. It provides both a source of data and code examples to work from for various problems the data fits.

- Lets take sometime to sign up a haggle account *(. . . we have to in order to download data.)* and browse what it has to offer.

Another place that is ripe with data for working with machine learning is the Machine Learning Repository $\sim$ UCI.

- Lets see what this site has to offer and use the tool to navigate the types of data it hosts.

What is Data?

`Data` is a set of *(values/measurements)* of *(quantitative or qualitative)* variables.

- It is information in a raw or unstructured form.
- It can consist of fact, figure, characters, symbols, etc.
- Take a look at the example below. It reflects measurements of BJ sales performance.
  - Think of all the questions that could be addressed using this dataset:

```
library(datasets)
head(BJsales)
```

```
## [1] 200.1 199.5 199.4 198.9 199.0 200.2
```

Types of Measurements

In data, we can distinguish two types of variables:

- (1) Categorical and,
- (2) Continuous

(1) - Categorical Variables

(def) - entities that are divided into distinct categories.

- Includes the following:
  - binary variables
  - nominal variables
  - ordinal variables
- A categorical variable can be *countries*, *year*, *gender*, *occupation*.
- For instance in R, it stores categorical variables into a factor format.
  - Factors are the variables in R which take on a limited number of different values.

Categorical Levels of Measurement - Binary

(def): a binary variable is only two categories.

- i.e. dead or alive.
- i.e. gender *(female or male)*.

Continued... Categorical Levels of Measurement - Binary

- Example of a binary variable:

```
gender_vector <- c("Male", "Female", "Female")
str(gender_vector)
```

```
##  chr [1:3] "Male" "Female" "Female"
```

```
factor_gender_vector <-factor(gender_vector)
str(factor_gender_vector)
```

```
##  Factor w/ 2 levels "Female","Male": 2 1 1
```

Categorical Levels of Measurement - Nominal

(def): A nominal variable is more than two categories.

- i.e. whether someone is an omnivore, vegetarian, vegan, or fruitarian.
- Example of a nominal variable:

```
data = iris
str(data$Species)
```

```
##  Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1
```

Categorical Levels of Measurement - Ordinal

(def): A ordinal variable is the same as a nominal, but the categories have a logical order.

- i.e. whether people got a fail, a pass, a merit or a distinction in their exam.
- i.e suppose you have a variable, economic status, with three categories *(low, medium and high)*.

In addition to being able to classify values into categories, you can order the categories, for instance:

- *(low, medium and high.)*

Continued... Categorical Levels of Measurement - Ordinal

Example of a ordinal variable:

```
credit_scores = c(650, 700, 750, 641, 800)
credit_status = ifelse(credit_scores<699, "low", ifelse(cre
credit_pt = ifelse(credit_status=="low", 1,ifelse(credit_st
credit_data = data.frame(credit_scores,credit_status, credi
credit_data[order(credit_pt),]
```

```
##   credit_scores credit_status credit_pt
## 1           650           low         1
## 4           641           low         1
## 2           700           med         2
## 3           750           med         2
## 5           800          high         3
```

(2) Continous Variables

(def): entities get a distinct score.

- Includes the following:
  - interval variables
  - ratio variables

Continous Levels of Measurement - Interval

(def): A interval variable is equal intervals on the variable. It represent equal differences in the property being measured.

- i.e. the difference between 6 and 8 is equivalent to the difference between 13 and 15.
- i.e. What is the average precip. reading in the South West of America?

Continued... Continous Levels of Measurement - Interval

- Example of a interval variable:

```
library(datasets)
precip = precip
head(precip)
```

```
##      Mobile      Juneau     Phoenix Little Rock Los Ange
##        67.0        54.7         7.0        48.5        1
```

Continous Levels of Measurement - Ratio

(def): A ratio variable is the same as an interval variable, but the ratios of scores on the scale must also make sense.

- i.e. a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8.

Continued... Continous Levels of Measurement - Ratio

- Example of a ratio variable:

```
library(datasets)
trees = trees
trees$Height
```

```
##  [1] 70 65 63 72 81 83 66 75 80 75 79 76 76 69 75 74 85
## [26] 81 82 80 80 80 87
```

Consider Measurement Error:

The accuracy of the measurements are key to your solutions.

- Measurement Error: - aka observational error
  - ***(def):*** The discrepancy between the actual value we're trying to measure, and the number we use to represent that value.
    - i.e. You (in reality) weigh 80 kg.
    - i.e. You stand on your bathroom scales and they say 83 kg.
    - i.e. The measurement error is 3 kg.

How Valid Are My Measures?

Validity:

- ***(def):*** - Whether an instrument measures what it set out to measure.
  - Including the following:
    - Content validity - Evidence that the content of a test corresponds to the content of the construct it was designed to cover.
    - Ecological validity - Evidence that the study, experiment or test can be applied, and allow inferences, to real-world conditions.

Are My Measures Reliable?

Reliability:

- ***(def):*** The ability of the measure to produce the same results under the same conditions.

Test-Retest Reliability:

- ***(def):*** The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

Population vs Sample

First, populations and samples should be understood so that your analysis is not mis-leading when interpreting results.

- ***Population***:
  - ***(def):*** includes all of the elements from a set of data.
- ***Sample***:
  - ***(def):*** consists one or more observations drawn from the population.

Structured Data

When you think of sturtcured data think of organized formatted data in tables, lists, etc.

Unstructured Data

When you think of unstructured data you should think of any data sources that is not in a table, list, etc. Typically the data is either text or images as will discuss in the course.