BAN 210: Essentials of Analytics

Shikhar Acharya, Ph.D.

Taming the Big Data Tidal Wave

- Bill Franks

https://www.youtube.com/watch?v=-VQu7-5Ucs0

Everybody Lies: Big Data, New Data, and What The Internet Can Tell Us About Who We Really Are

- Seth Stephens-Davidowitz

https://www.youtube.com/watch?v=s59SxTg0B98

Taming the Big Data Tidal Wave
- Bill Franks

https://www.youtube.com/watch?v=-VQu7-5Ucs0

https://www.youtube.com/watch?v=iXSbi2WsZoU

Chapter 1: What is Big Data and Why Does it Matter

- What is Big Data?
  - "Big Data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population" Merv Adrian

  - "Big Data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" McKinsey Global Institute

Big data is not only about volume. It is also about velocity and variety of data

http://www.youtube.com/watch?v=2D8oji5EKbM

https://www.youtube.com/watch?v=Hv397JnNWYc

According to the author of the blue book, from the perspective of organization on benefitting from big data, is the ***big*** part or the ***data*** part important?

How is big data different from traditional data?

- Big data is automatically generated without direct human involvement
- Big data is often new source of data than the ones that were collected by directed human intervention
- Big data is often disorganized and **not** in a user friendly format
- Large portion of big data may not provide us any useful information

How is big data same from traditional data?

- The “volume” of data analyzed in some cases tested the limit of resources at that time.

Risks of big data

- Organizations may not be able to “tame” big data
  - Bring the right person
- Too much resources wasted on capturing and storing data than on analyzing and extracting value from it
  - Start with small and manageable project

- Privacy

https://www.youtube.com/watch?v=M-kTFOBYlrQ

https://www.youtube.com/watch?v=nS_qKr-yGck

https://www.youtube.com/watch?v=GI1cPijv-l8

Opposing view:
https://www.youtube.com/watch?v=BJ8puU2VFqw

Should the government be allowed to collect personal data from all citizens for security reasons such as protecting from homeland terrorist attack?

Why to tame big data

- What was the need to tame horses, dogs, and oxen?

- What was the need to mine metals and petroleum products?

The structure of big data

- Structured data
  - Clear format and without any kind of ambiguity
  - Consistent on format

  https://www.census.gov/foreign-trade/statistics/historical/gands.pdf

- Unstructured data
  - Ambiguous data, extremely difficult to comprehend

    eg: picture/audio/video signal in digital format

- Semi-structured data
  - Between the two

  view-source:
  https://www.kaggle.com/c/digit-recognizer/data

Exploring Big Data

- Much time is spent collecting, formatting, and cleaning data (>80%) and the rest of the time in analyzing the data

http://www.datasciencecentral.com/profiles/blogs/what-are-the-greatest-inefficiencies-data-scientists-face-today

- Right question, right data, and incremental value creation
  - Example: second paragraph, page 17

Getting rid of useless data

- Data generated every second about someone's location
- Pixel data in handwritten digit recognition
- Do not need the location of each product with RFID tag
  - *but store data if feasible*

- ETL: Extract, Transform, and Load data
  - extract from data source
  - Transform to useful from
  - Load to proper platform for analysis
- EDW: Enterprise Data Warehouse
  - Centralized storage of all available data
  - Combination with traditional data

Need for data standards

- Decreases the resources required for data cleaning and preparation

- Enables transform of data between platforms

Big data is a relative concept. What is big data now may not be considered big data later

Examples of future ‘potential’ big data: page 25

Taming the Big Data Tidal Wave
- Bill Franks

Quiz on 09/06 (open book, no notes, no internet except eLearn)

- Foreword, Preface, Chapter 1 and Chapter 2
- **Please bring your charged laptop with internet connectivity**