LNMIIT

Introduction to Data Science

In God we trust, all others bring data

- W.E. Deming

Data Science

- Take care of 3V's of data - Velocity, Volume and Variety
- Integrating data from various sources
- Store it in an uniform format to make it easier to analyse
- Analyse data to find patterns or trend
- Machine learning techniques to predict and forecast
- Make decisions based on the findings

Every 2 days we create as much as we did from the beginning of time until 2003!

If you burned all of the data created in just one day onto DVDs, you could stack them on top of each other and reach the moon - twice!

The total amount of data being captured and stored in industry doubles every 1.2 years!

Source: http://www.slideshare.net/BernardMarr/big-data-25-facts

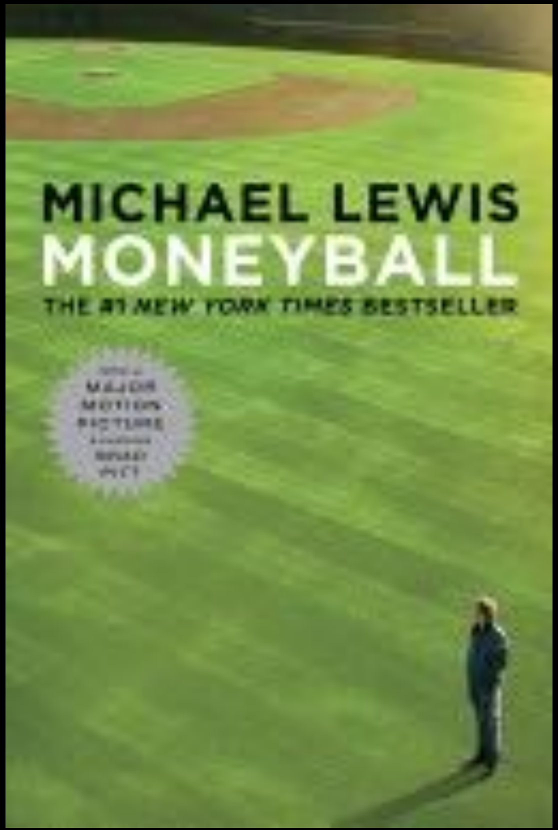
Data Flood!

- 500 Million tweets sent every day
- 241 million monthly users active
- 76% access on mobile
- Over 35 languages

NETFLIX

- On-demand internet streaming media (from 2007)
- It was mostly available in North America and Europe only but now in India too
- DVD services
- Subscribers in 40 countries
- **Movie recommendation engine**
- **2006 - Netflix prize for recommendation also**
- 2010 - deanonimization problem
- Throtling issue

Analytics in Baseball



- Oakland A’s baseball team - poorest team - but managed to win so many games
- Billy Beane was the manager
- **Did baseball statistics - defined new metrics for player selection - Sabermetrics**
- Now Boston Red Sox team uses analytics everywhere in ALCS

CAESARS ENTERTAINMENT

Harrah's

- Casinos - Loveman (COO) - Harvard Business School Professor
- **Repeat slot players, not high rollers, were most profitable**
- Loyalty reward card program - diamond, platinum and gold - customer relationship management
- Customer's preference are captured and analysed in real time
- **82% of revenue comes from 26% of customers - most of them old ladies!**
- Incentivising staff on customer satisfaction not on financial performance
- Pain threshold → Anticipate dissatisfaction → free meal or movie time

DS in the Business Context

- Change is constant in business environment
- It gets more and more complex
- Companies have to be agile and take quick strategic, tactical and operational decisions
- It needs data, data and data
- But data is only the first step, it needs information and knowledge to make decisions and then actions

DS in Business Scenario

- Business pressures due to competitive market
- Take responses to counter the pressure
- Computerised support to monitor the environment, and
- To enhance the actions taken by organization
- Complexity is two way - it provides opportunity and problems. That’s the challenge!
    - Example - globalisation has brought in more competition together with more complexities

Scenario and Examples

- Rule-based systems that provides solutions in some specific area - finance, manufacturing and so on
- Example: To determine the price of an item
- Example: If 90% of the seats on a flight are full 10 hours prior to the departure then offer a discount of x
- Example: Recommender systems

Scenario and Examples

- Warning or action that is initiated when a predetermined unusual event occurs
- Example: Credit card payments when it is unusual additional authorisation may be required
- Example: When a longtime customer deposits large amount of money he may get higher interest rate for the deposit
- Alerts may be sent to managers also who are responsible to manage the performance of the transactions

Scenario and Examples

1. What was the sales figures of the shop A in this period?
2. Which shop is the most profitable in this locality?
3. Which products are selling the most in this locality?
4. Why was the sales figure low for the shop A in this period?
5. Why was that shop more profitable in this locality?
6. Why were those products selling like anything in this locality

Can you order the above in terms of increasing order of difficulty?

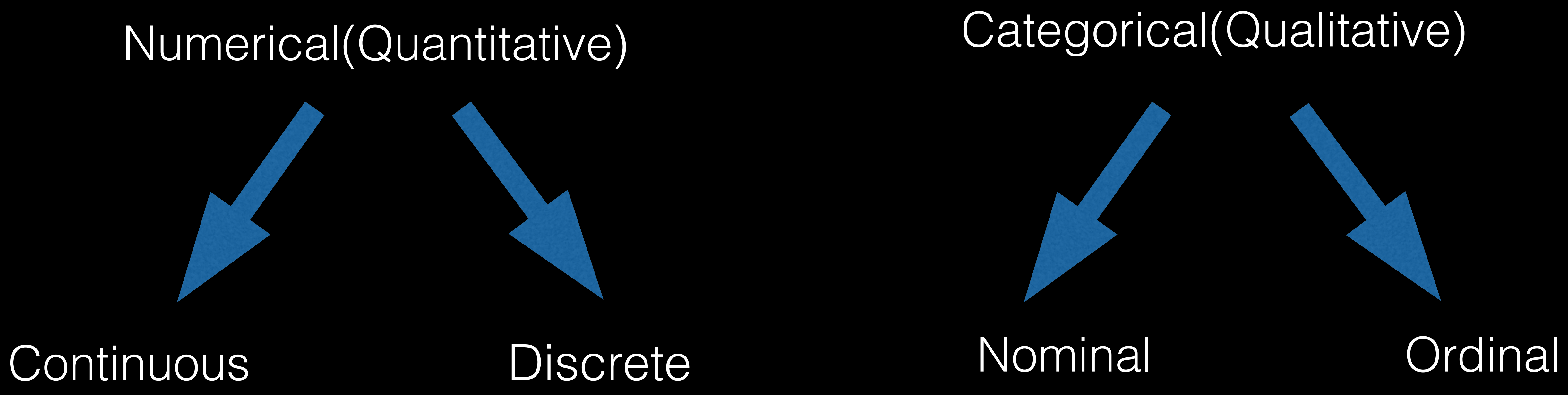
Scenario and Examples

1. Sentiment analysis of newspaper handles with respect to politics in Twitter
2. Sentiment analysis of newspaper handles in Twitter
3. Sentiment analysis of newspaper handles with respect to various political parties in Twitter - this is called Bias!

Can you order the above in terms of increasing order of difficulty?

Data

Types of Variables



Numerical(Quantitative)

Continuous

Discrete

Categorical(Qualitative)

Nominal

Ordinal

Descriptive Statistics

- Data described in a format that is concise and crisp
  - Graphical representation
  - Tabular representation
  - Summary Statistics

Descriptive versus Inferential Statistics

Descriptive Statistics

- Single variable
- Multiple variable
    - Distribution
    - Relationship