Big Data Analytics

Mitschrift von Aaron Winziers

SS 2020 - Coronasemester

1 Introduction

1.1 3 Big Vs

- Volume
- Velocity - Data should be updated much more quickly - no longer work in batches
- Variety - Videos, text, from web etc

Veracity joins the other 3 Vs nowadays

1.2 Volume

- Average company has 100 TB of data
- 2.5 quintillion bytes created every day
- the amount of data created will be 300x greater in 2020 than 2005 (aggregate, estimate)

1.2.1 Challenges created by data volume

- Efficient storage
- Efficiant process queries
- Efficient learning with models
- What hardware and software architecture is needed for this?

1.3 Variety

- Data consists of different forms of data

1.3.1 Challenges created by data variety

- Syntactic heterogeneity - understadning different data types and formats
- Semantic heterogeneity - Differnt representations forthe same information
  Name abreviations - John Smith, J Smith, (Smith, John), Jon Smithe
- The prev 2 issues need to be understood because we need to combine:
  information from many different sources
  different types of information

1.4 Velocity

- The speed at which data is created and processed
- Data needs to be processed quickly or otherwise (sometimes) forgotten

1.4.1 Challenges created by data velocity

- Extremely fast flow of information
- Assessing the value of incoming information and drop "unimportant" information
- Quick integration of new information

1.5 Veracity

- Deals with the uncertainty of data
- Can you trust the data?

1.5.1 Challenges created by data velocity

- Differnet kinds of data defects:
  Data may be invalid (broken sensors, bad software)
  Data may be biased and not reflect the true population
  Data may be manipulated
- Methods are needed to identify and "repair" data defects

1.5.2 User-Generated Data

- Users may answer dishonestly or not take surveys seriously
- Users may try to purposely influence the results of surveys
- Must check the plausibility of the data before using