Chapter 01

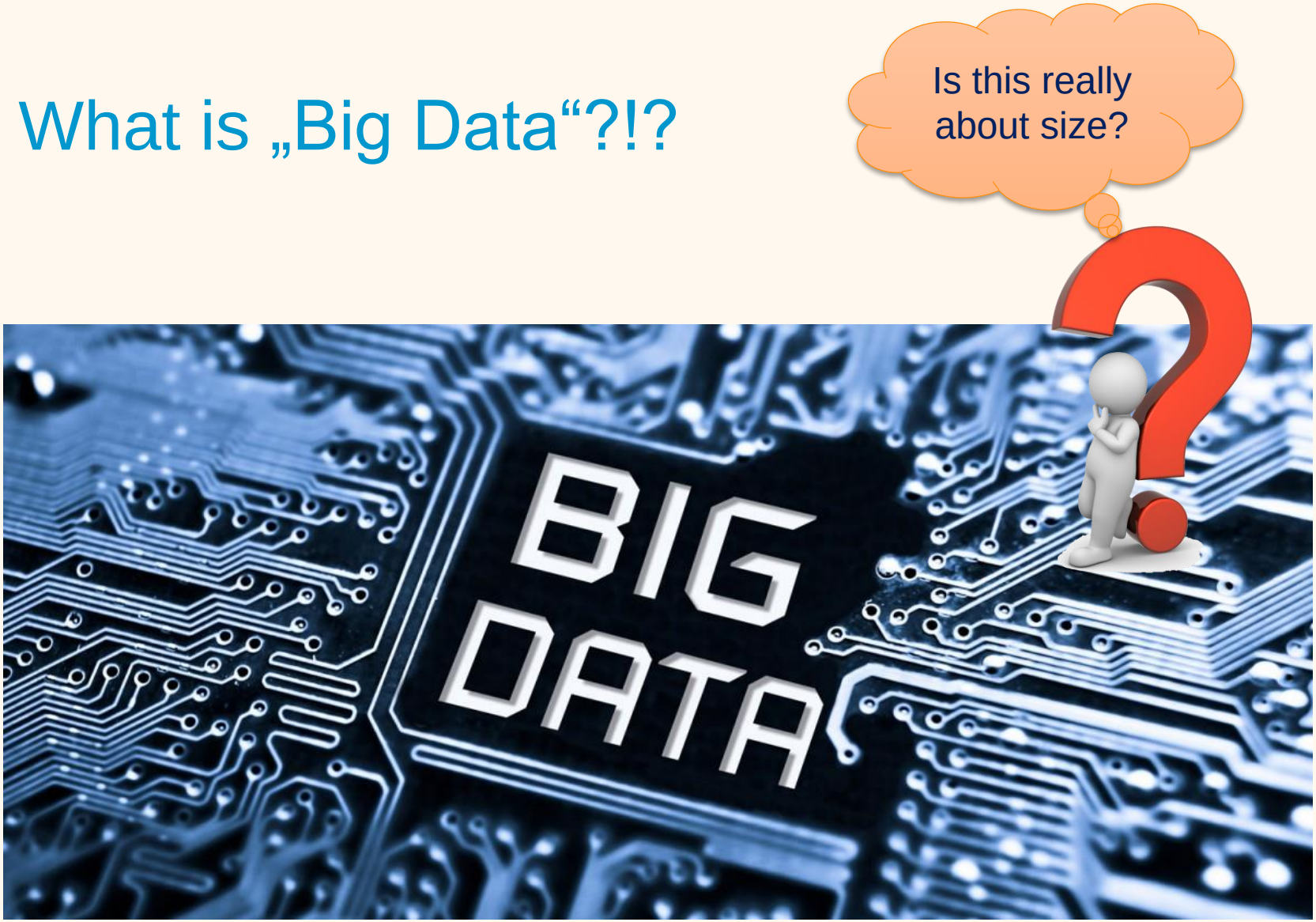
Introduction

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

Outline

- Introduction to Big Data
- Data Science and Business Intelligence
- The Skillset of Data Scientists
- Summary

What is „Big Data“?!?



Naive Definition

- Naive definition:
  - Big data only depends on the data size
  - 1 Gigabyte? 1 Terabyte? 1 Petabyte?

- Naive interpretation misses important aspects
  - Time:
    - Analyzing 1 Gigabyte of data per day is different from analyzing 1 Gigabyte of data per second
  - Diversity:
    - Analyzing spread sheets with numeric data is different from analyzing Web pages that contain a mixture of text and images
  - Distribution:
    - Analyzing data from a single source is different from analyzing data from multiple sources

Definition of Big Data

- Following Gartner‘s IT Glossary:
  - Big data is high-**volume**, high-**velocity** and/or high-**variety** information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

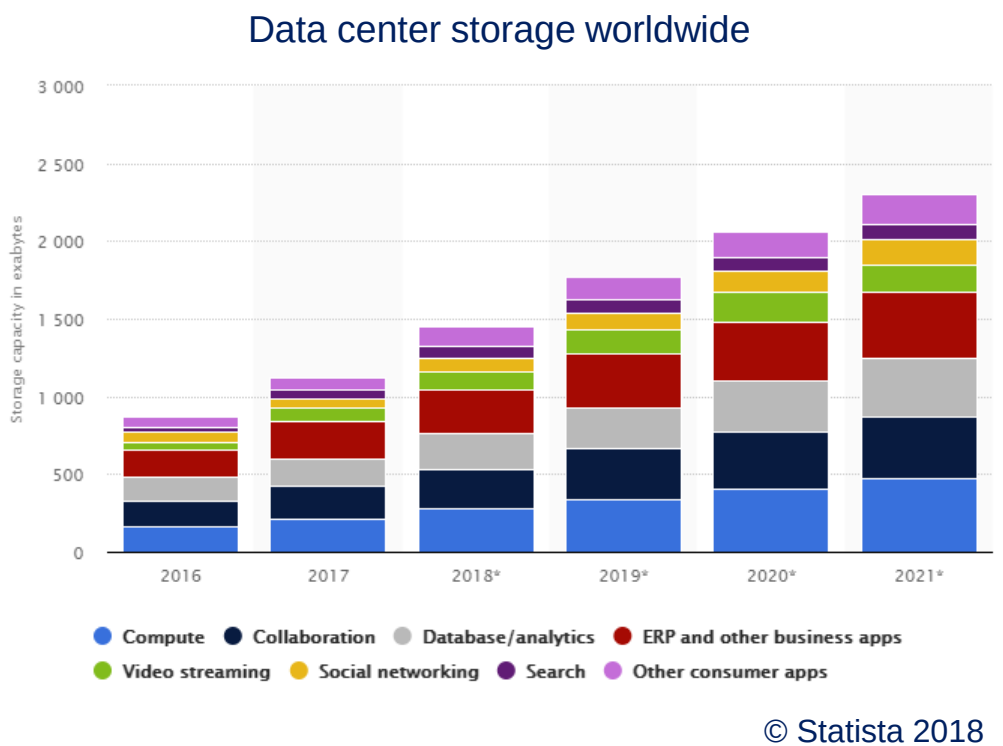
- The three Vs
  - Volume
  - Velocity
  - Variety

Some people actually use 10 Vs to define big data!

- Variability
- Veracity
- Validity
- Vulnerability
- Volatility
- Visualization
- Value

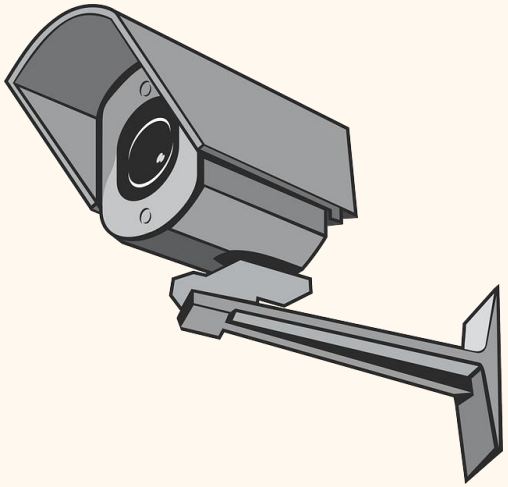
The 3 Vs: Volume

- Scale of the data must be „big“
  - No clear definition
  - „that demand […] innovative forms of information processing“ (Gartner)



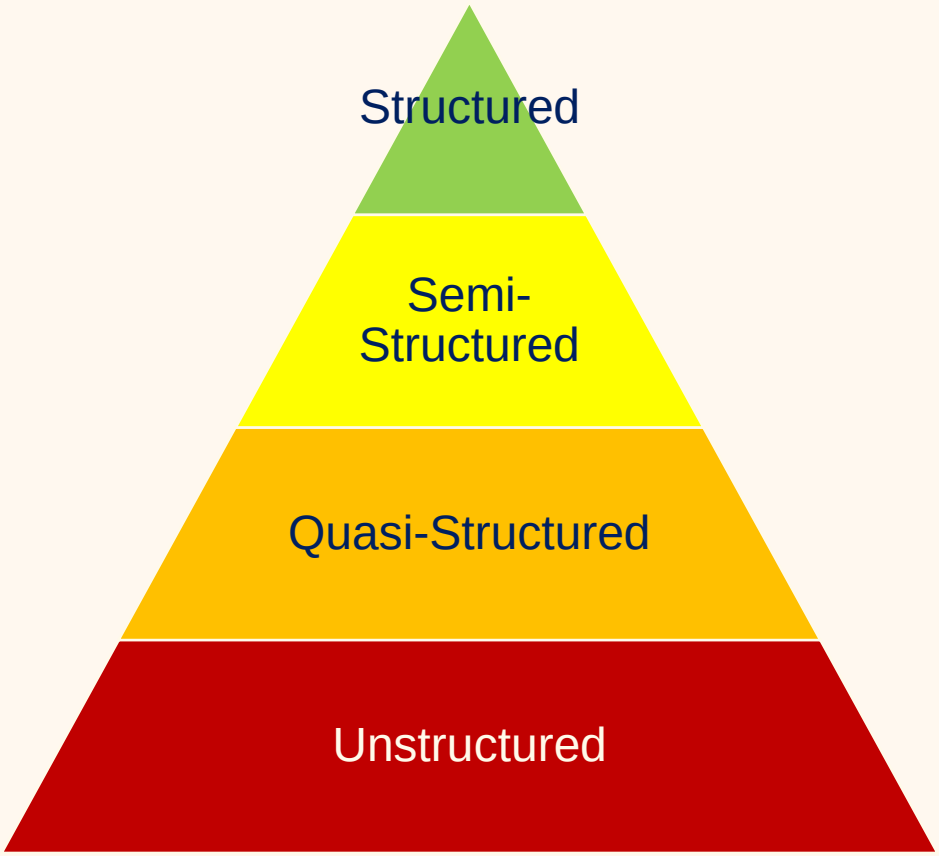
The 3 Vs: Velocity

- Speed at which new data is created
- Speed at which data must be processed and analyzed
  - Often close to real-time



The 3 Vs: Variety

- Diversity in data types and data sources



- Data with defined types and structure
- Example: comma separated values

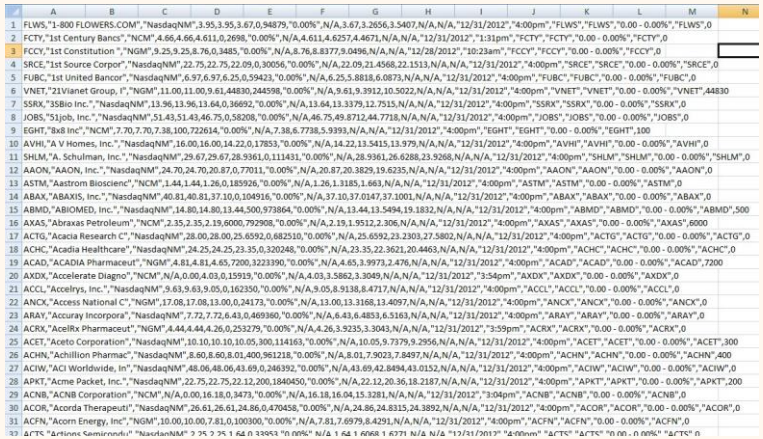
- Textual data with parseable pattern
- Example: XML files with schema

- Textual data with erratic formats that can be formated with effort
- Example: Clickstream data

- Data that has no inherent structure, often with multiple formats
- Example: Web site, videos

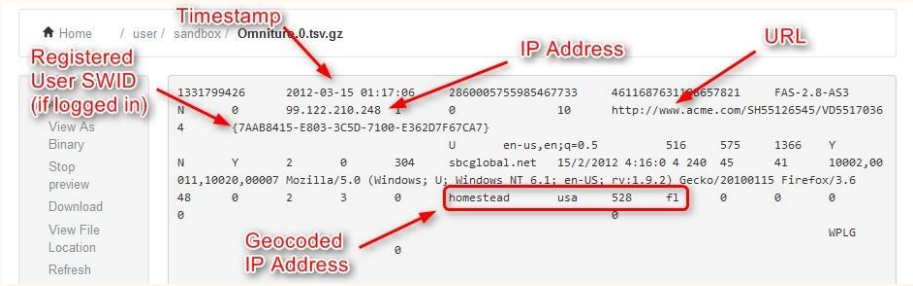
Examples for data types

Structured



```
FLWS,"1-800 FLOWERS.COM","NasdaqNM",3.95,3.95,3.67,0,94879,"0.00%",N/A,3.67,3.2656,3.5407,N/A,N/A,"12/31/2012","4:00pm","FLWS","FLWS","0.00 - 0.00%","FLWS",0
FCTY,"1st Century Bancs","NCM",4.66,4.66,4.611,0,2698,"0.00%",N/A,4.611,4.6257,4.4671,N/A,N/A,"12/31/2012","1:31pm","FCTY","FCTY","0.00 - 0.00%","FCTY",0
FCCY,"1st Constitution ","NGM",9.25,9.25,8.76,0,3485,"0.00%",N/A,8.76,8.8377,9.0496,N/A,N/A,"12/28/2012","10:23am","FCCY","FCCY","0.00 - 0.00%","FCCY",0
SRCE,"1st Source Corpor","NasdaqNM",22.75,22.75,22.09,0,30056,"0.00%",N/A,22.09,21.4568,22.1513,N/A,N/A,"12/31/2012","4:00pm","SRCE","SRCE","0.00 - 0.00%","SRCE",0
FUBC,"1st United Bancor","NasdaqNM",6.97,6.97,6.25,0,59423,"0.00%",N/A,6.25,5.8818,6.0873,N/A,N/A,"12/31/2012","4:00pm","FUBC","FUBC","0.00 - 0.00%","FUBC",0
VNET,"21Vianet Group, I","NGM",11.00,11.00,9.61,44830,244598,"0.00%",N/A,9.61,9.3912,10.5022,N/A,N/A,"12/31/2012","4:00pm","VNET","VNET","0.00 - 0.00%","VNET",44830
SSRX,"3SBio Inc.","NasdaqNM",13.96,13.96,13.64,0,36692,"0.00%",N/A,13.64,13.3379,12.7515,N/A,N/A,"12/31/2012","4:00pm","SSRX","SSRX","0.00 - 0.00%","SSRX",0
JOBS,"51job, Inc.","NasdaqNM",51.43,51.43,46.75,0,58208,"0.00%",N/A,46.75,49.8712,44.7718,N/A,N/A,"12/31/2012","4:00pm","JOBS","JOBS","0.00 - 0.00%","JOBS",0
EGHT,"8x8 Inc","NCM",7.70,7.70,7.38,100,722614,"0.00%",N/A,7.38,6.7738,5.9393,N/A,N/A,"12/31/2012","4:00pm","EGHT","EGHT","0.00 - 0.00%","EGHT",100
AVHI,"A V Homes, Inc.","NasdaqNM",16.00,16.00,14.22,0,17853,"0.00%",N/A,14.22,13.5415,13.979,N/A,N/A,"12/31/2012","4:00pm","AVHI","AVHI","0.00 - 0.00%","AVHI",0
SHLM,"A. Schulman, Inc.","NasdaqNM",29.67,29.67,28.9361,0,111431,"0.00%",N/A,28.9361,26.6288,23.9268,N/A,N/A,"12/31/2012","4:00pm","SHLM","SHLM","0.00 - 0.00%","SHLM",0
AAON,"AAON, Inc.","NasdaqNM",24.70,24.70,20.87,0,77011,"0.00%",N/A,20.87,20.3829,19.6235,N/A,N/A,"12/31/2012","4:00pm","AAON","AAON","0.00 - 0.00%","AAON",0
ASTM,"Aastrom Bioscienc","NCM",1.44,1.44,1.26,0,185926,"0.00%",N/A,1.26,1.3185,1.663,N/A,N/A,"12/31/2012","4:00pm","ASTM","ASTM","0.00 - 0.00%","ASTM",0
ABAX,"ABAXIS, Inc.","NasdaqNM",40.81,40.81,37.10,0,104916,"0.00%",N/A,37.10,37.0147,37.1001,N/A,N/A,"12/31/2012","4:00pm","ABAX","ABAX","0.00 - 0.00%","ABAX",0
ABMD,"ABIOMED, Inc.","NasdaqNM",14.80,14.80,13.44,500,973864,"0.00%",N/A,13.44,13.5494,19.1832,N/A,N/A,"12/31/2012","4:00pm","ABMD","ABMD","0.00 - 0.00%","ABMD",500
AXAS,"Abraxas Petroleum","NCM",2.35,2.35,2.19,6000,792908,"0.00%",N/A,2.19,1.9512,2.306,N/A,N/A,"12/31/2012","4:00pm","AXAS","AXAS","0.00 - 0.00%","AXAS",6000
ACTG,"Acacia Research C","NasdaqNM",28.00,28.00,25.6592,0,682510,"0.00%",N/A,25.6592,23.2303,27.5802,N/A,N/A,"12/31/2012","4:00pm","ACTG","ACTG","0.00 - 0.00%","ACTG",0
ACHC,"Acadia Healthcare","NasdaqNM",24.25,24.25,23.35,0,320248,"0.00%",N/A,23.35,22.3621,20.4463,N/A,N/A,"12/31/2012","4:00pm","ACHC","ACHC","0.00 - 0.00%","ACHC",0
ACAD,"ACADIA Pharmaceut","NGM",4.81,4.81,4.65,7200,3223390,"0.00%",N/A,4.65,3.9973,2.476,N/A,N/A,"12/31/2012","4:00pm","ACAD","ACAD","0.00 - 0.00%","ACAD",7200
AXDX,"Accelerate Diagno","NCM",N/A,0.00,4.03,0,15919,"0.00%",N/A,4.03,3.5862,3.3049,N/A,N/A,"12/31/2012","3:54pm","AXDX","AXDX","0.00 - 0.00%","AXDX",0
ACCL,"Accelrys, Inc.","NasdaqNM",9.63,9.63,9.05,0,162350,"0.00%",N/A,9.05,8.9138,8.4717,N/A,N/A,"12/31/2012","4:00pm","ACCL","ACCL","0.00 - 0.00%","ACCL",0
ANCX,"Access National C","NGM",17.08,17.08,13.00,0,24173,"0.00%",N/A,13.00,13.3168,13.4097,N/A,N/A,"12/31/2012","4:00pm","ANCX","ANCX","0.00 - 0.00%","ANCX",0
ARAY,"Accuray Incorpora","NasdaqNM",7.72,7.72,6.43,0,469360,"0.00%",N/A,6.43,6.4853,6.5163,N/A,N/A,"12/31/2012","4:00pm","ARAY","ARAY","0.00 - 0.00%","ARAY",0
ACRX,"AcelRx Pharmaceut","NGM",4.44,4.44,4.26,0,253279,"0.00%",N/A,4.26,3.9235,3.3043,N/A,N/A,"12/31/2012","3:59pm","ACRX","ACRX","0.00 - 0.00%","ACRX",0
ACET,"Aceto Corporation","NasdaqNM",10.10,10.10,10.05,300,114163,"0.00%",N/A,10.05,9.7379,9.2956,N/A,N/A,"12/31/2012","4:00pm","ACET","ACET","0.00 - 0.00%","ACET",300
ACHN,"Achillion Pharmac","NasdaqNM",8.60,8.60,8.01,400,961218,"0.00%",N/A,8.01,7.9023,7.8497,N/A,N/A,"12/31/2012","4:00pm","ACHN","ACHN","0.00 - 0.00%","ACHN",400
ACIW,"ACI Worldwide, In","NasdaqNM",48.06,48.06,43.69,0,246392,"0.00%",N/A,43.69,42.8494,43.0152,N/A,N/A,"12/31/2012","4:00pm","ACIW","ACIW","0.00 - 0.00%","ACIW",0
APKT,"Acme Packet, Inc.","NasdaqNM",22.75,22.75,22.12,200,1840450,"0.00%",N/A,22.12,20.36,18.2187,N/A,N/A,"12/31/2012","4:00pm","APKT","APKT","0.00 - 0.00%","APKT",200
ACNB,"ACNB Corporation","NCM",N/A,0.00,16.18,0,3473,"0.00%",N/A,16.18,16.04,15.3281,N/A,N/A,"12/31/2012","3:04pm","ACNB","ACNB","0.00 - 0.00%","ACNB",0
ACOR,"Acorda Therapeuti","NasdaqNM",26.61,26.61,24.86,0,470458,"0.00%",N/A,24.86,24.8315,24.3892,N/A,N/A,"12/31/2012","4:00pm","ACOR","ACOR","0.00 - 0.00%","ACOR",0
ACFN,"Acorn Energy, Inc","NGM",10.00,10.00,7.81,0,100300,"0.00%",N/A,7.81,7.6979,8.4291,N/A,N/A,"12/31/2012","4:00pm","ACFN","ACFN","0.00 - 0.00%","ACFN",0
ACTS,"Actions Semicondu","NasdaqNM",2.25,2.25,1.64,0,33953,"0.00%",N/A,1.64,1.6068,1.6721,N/A,N/A,"12/31/2012","4:00pm","ACTS","ACTS","0.00 - 0.00%","ACTS",0
```

Quasi-Structured



Semi-Structured

```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

Unstructured

Outline

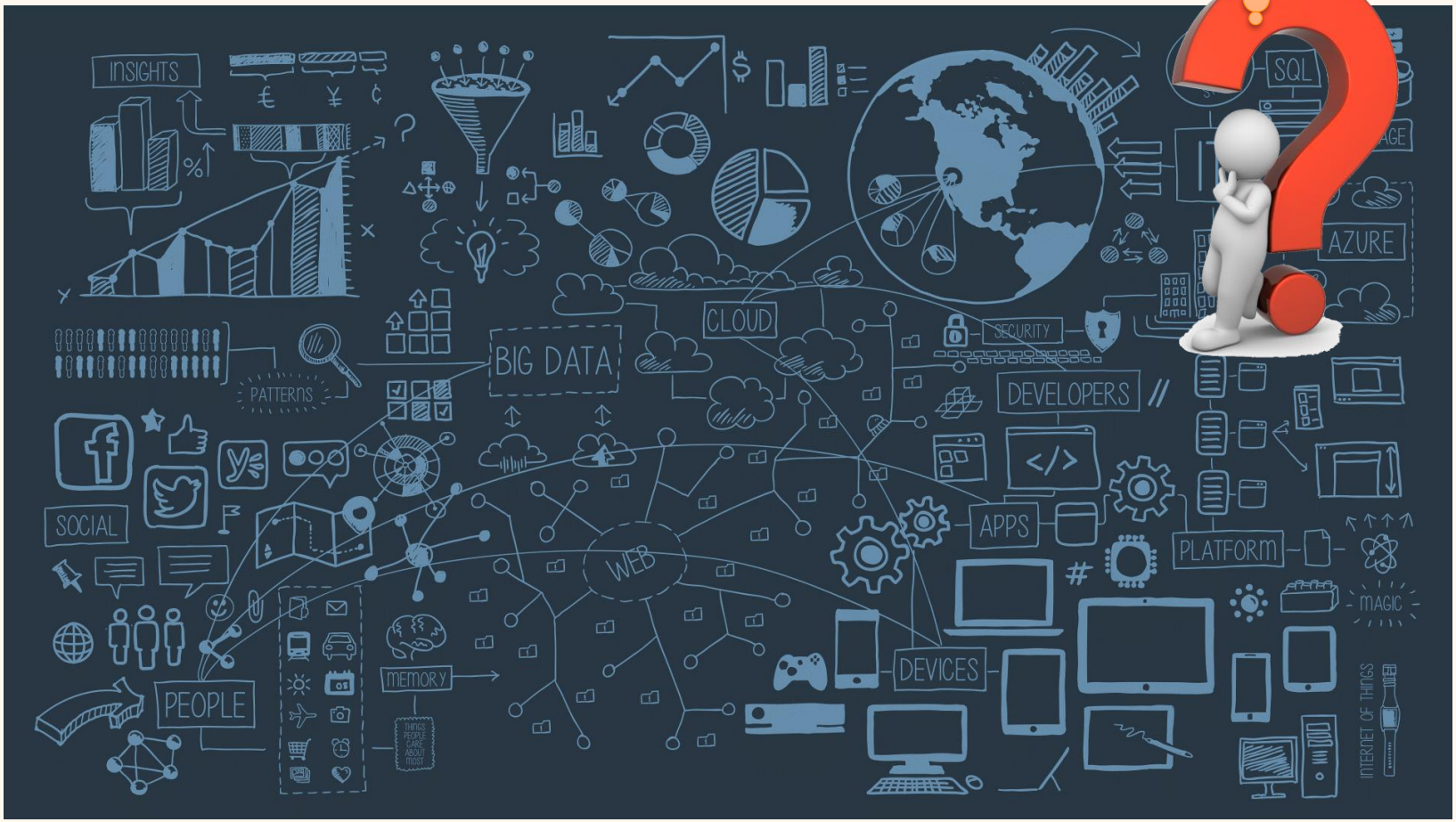
- Introduction to Big Data
- **Data Science and Business Intelligence**
- The Skillset of Data Scientists
- Summary

Defining Data Science

- Unfortunately, there is no clear definition (yet?)
- Goal is the extraction of knowledge from data
- Combination of techniques from different disciplines
- Scientific principles guide the data analysis

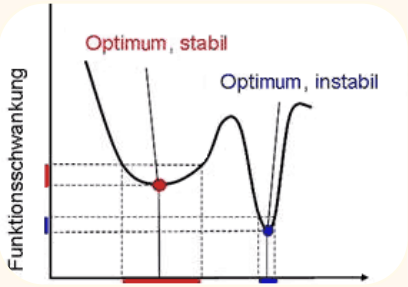
What is „Data Science“?!?

Tools? Big Data?
Machine Learning?



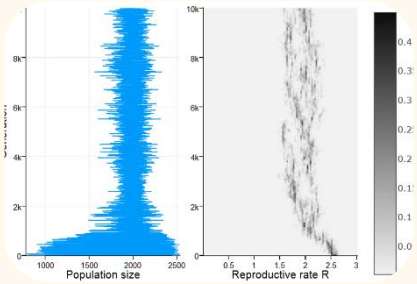
Mathematical Aspects

Computational Geometry



Optimization

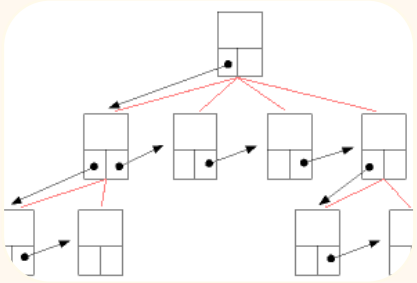
Stochastics



Scientific Computing

Machine Learning

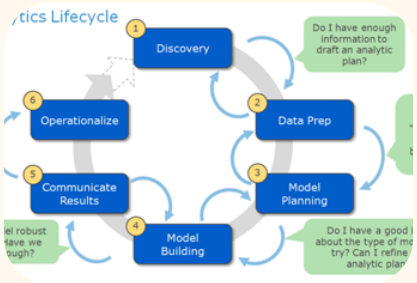
Computer Science Aspects



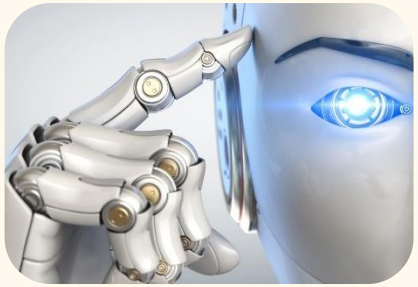
Data Structures and Algorithms

Databases

Distributed Computing



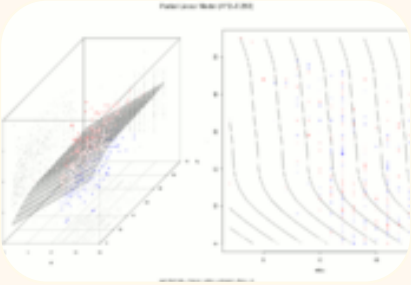
Software Engineering



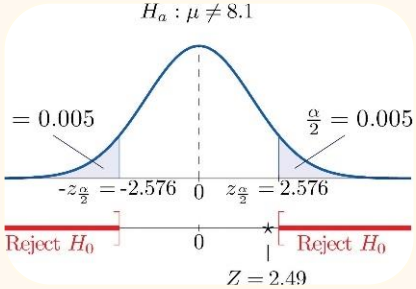
Artificial Intelligence

Machine Learning

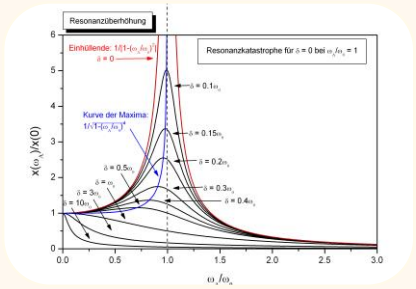
Statistical Aspects



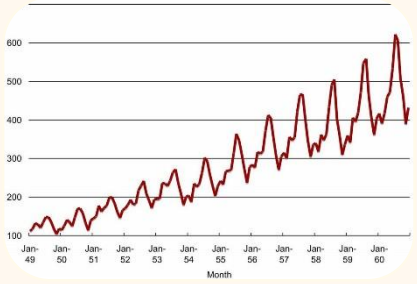
Linear Models



Statistical Tests



Inference



Time Series Analysis

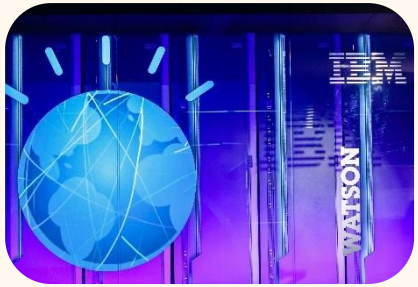
Machine Learning

Applications

Intelligent Systems

Robotics

Marketing



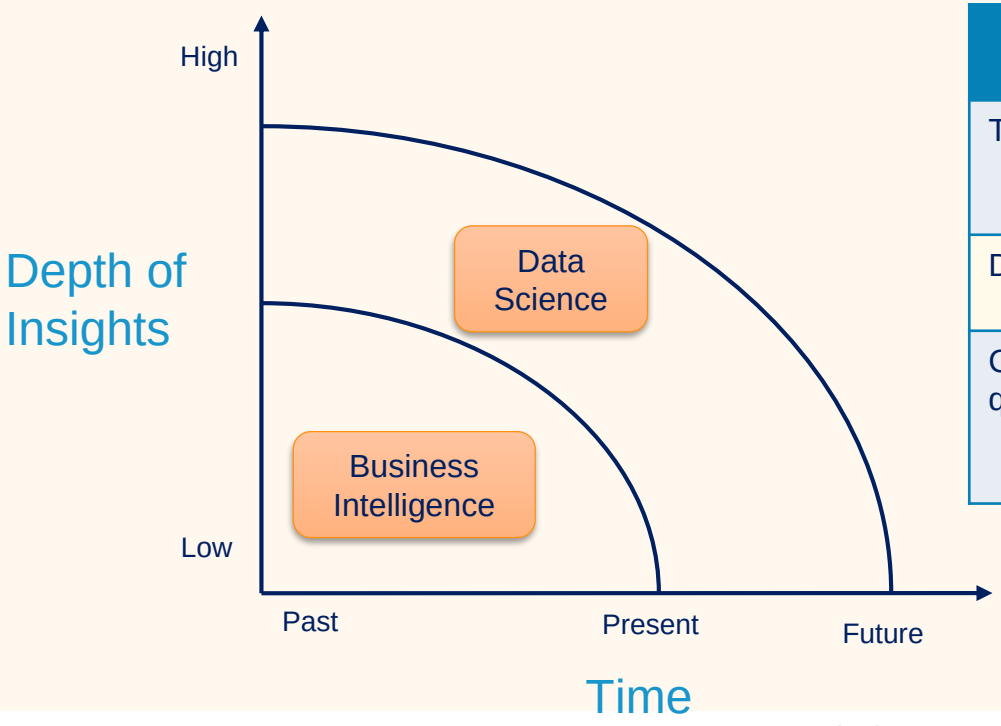
Medicine

Autonomous Driving

Social Networks

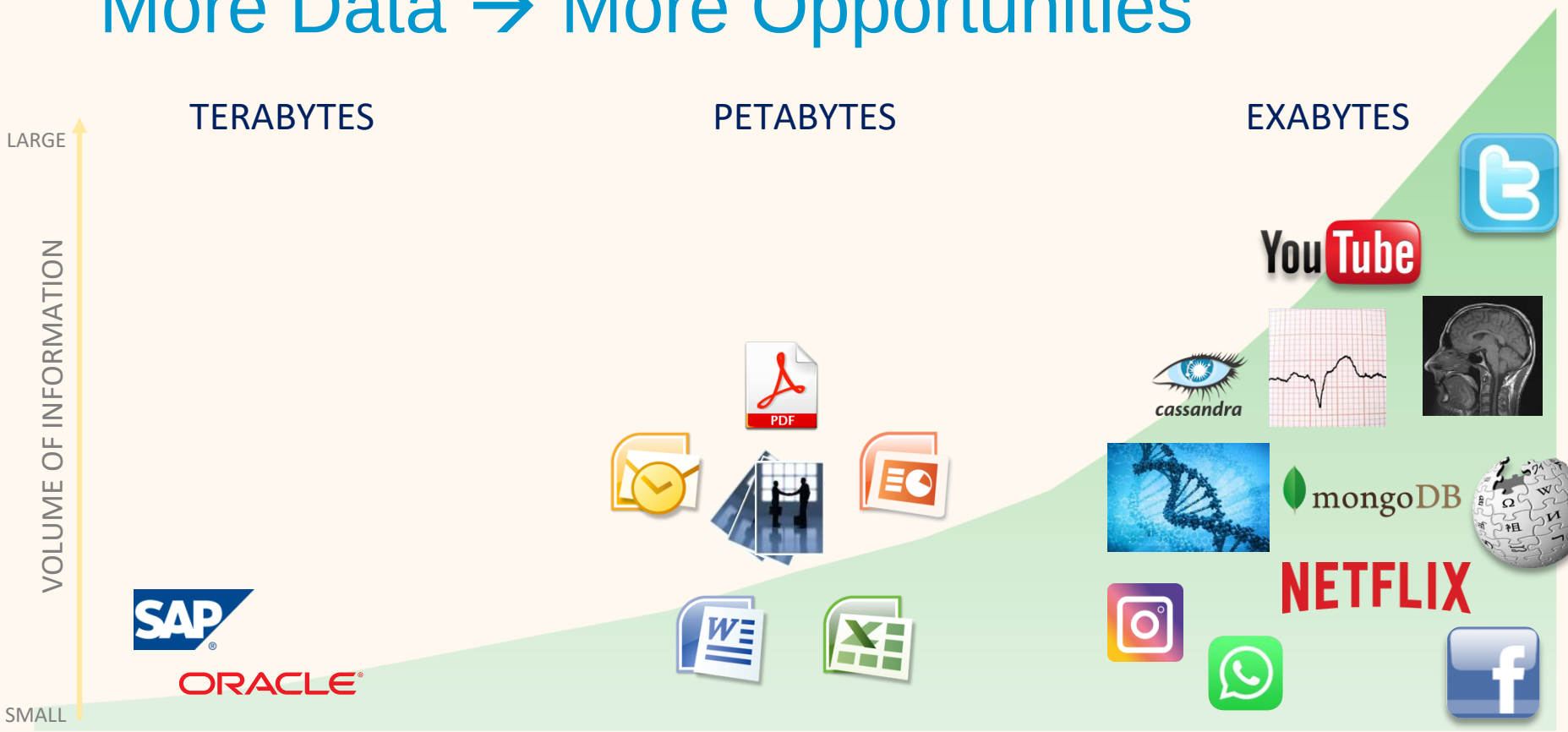
Data Science vs. Business Intelligence

- Business Intelligence (Gartner IT Glossary)
  - […] best practices that enable access to and analysis of information to improve and optimize decisions and performance.



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?

More Data → More Opportunities



1990's
Relational Databases
& Data Warehouses

2000's
Content Management

2010's
Key-Value Storages
& Unstructured Data

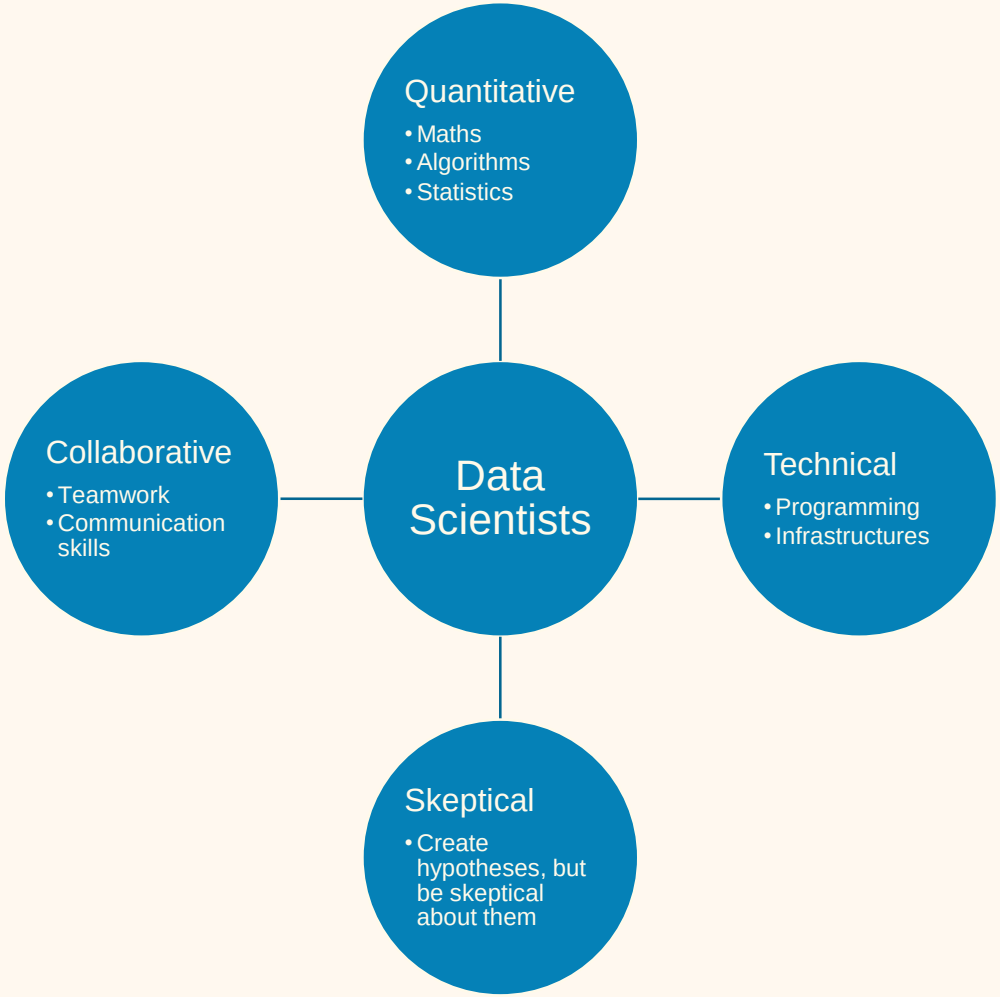
Outline

- Introduction to Big Data
- Data Science and Business Intelligence
- **The Skillset of Data Scientists**
- Summary

What are Data Scientists?

- Not computer scientists
  - But should know about databases, data structures, algorithms, etc.
- Not mathematicians
  - But should know about optimization, stochastics, etc.
- Not statisticians
  - But should know about regression, statistical tests, etc.
- Not domain experts
  - But must work together with them

Skills of Data Scientists



A bit of everything

… but actually as much as possible of everything

Different types of Data Scientists

- According to Microsoft Research:
  - Polymath
    - „Do it all“
  - Data Evangelist
    - Data analysis, disseminating and acting on insights
  - Data Preparer
    - Querying existing data, preparing data for analysis
  - Data Shapers
    - Analyzing and preparing data
  - Data Analyzer
    - Analyzing data
  - Platform Builder
    - Collect data and create infrastructures
  - Moonlighters (50%/20%)
    - „Spare time“ data scientists
  - Insight Actors
    - Use the outcome and act on insights.

Miyung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel: Data Scientists in Software Teams: State of the Art and Challenges, IEEE Transactions on Software Engineering (Online First)

Outline

- Introduction to Big Data
- Data Science and Business Intelligence
- The Skillset of Data Scientists
- **Summary**

Summary

- Big data has a high volume, velocity, and variety
- Different data structures
  - Structured, semi-structured, quasi-structured, unstructured
- Data science is a very diverse discipline
  - Maths, computer science, statistics, applications

→ Data scientists require a diverse skillset