

# Chapter 01

# Introduction

Dr. Steffen Herbold  
[herbold@cs.uni-goettingen.de](mailto: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“?!?

Is this really  
about size?



# 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.

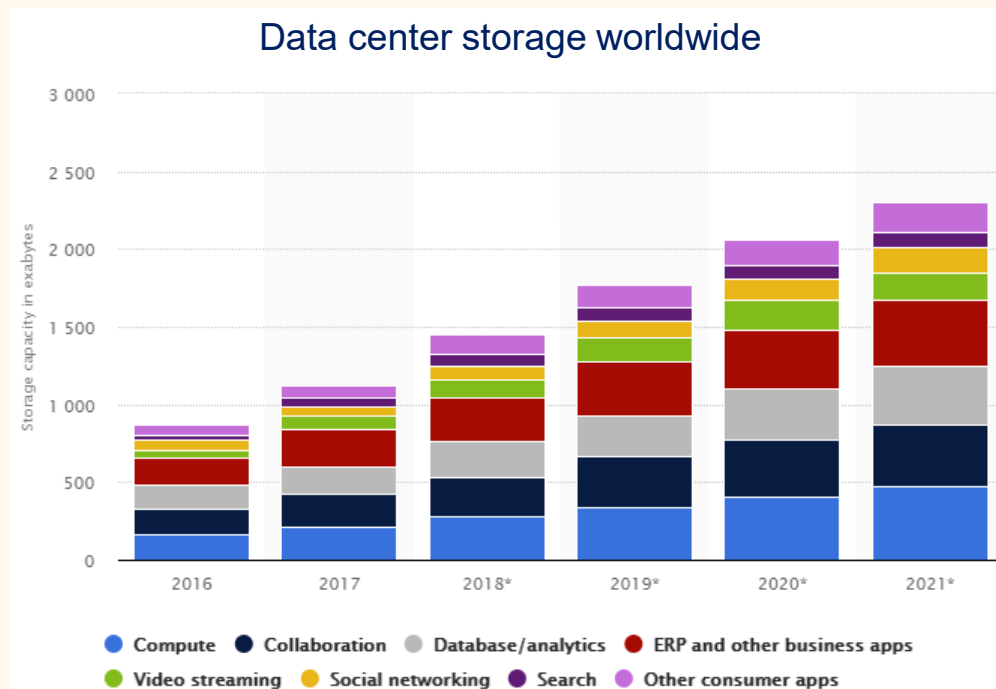
- The three Vs

- Volume
- Velocity
- Variety



# The 3 Vs: Volume

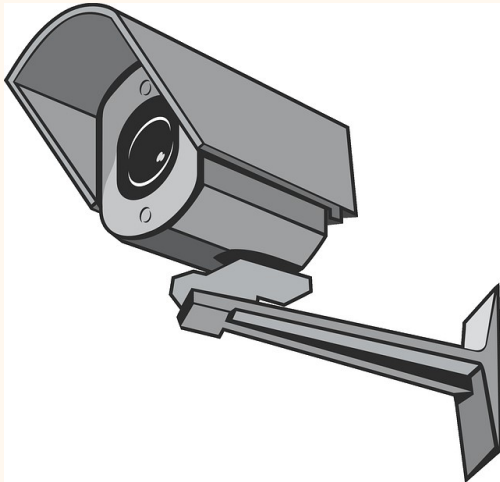
- Scale of the data must be „big“
  - No clear definition
  - „that demand [...] innovative forms of information processing“ (Gartner)



© Statista 2018

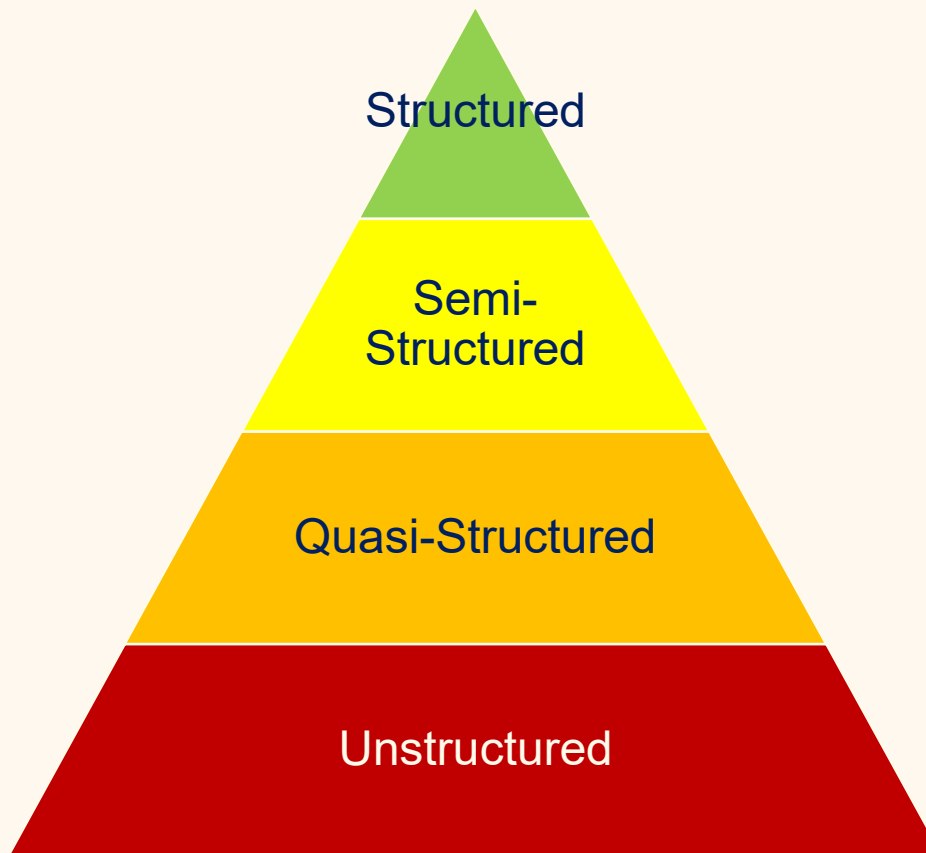
# The 3 Vs: Velocity

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



# Examples for data types

# Structured

	A	B	C	D	E	F	G	H	I	J	K	L	M	N							
1	FLWS	"1.800 FLWS Corp."	"NasdaqNM"	3.955	3.55	3.67	0.94879	0.00%	N/A	6.313	3.2656	3.5407	N/A	"12/31/2012"	"4:00pm"	FLWS	"FLWS" 0.00 - 0.00%	FLWS	0		
2	FCTY	"First Century Banc"	"NasdaqNM"	46.64	46.64	0.012698	"0.00%	N/A	6.0114	6.012698	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"1:31pm"	FCTY	"FCTY" 0.00 - 0.00%	FCTY	0
3	FISV	"First Solar Inc"	"NasdaqNM"	5.51	5.51	0.000000	"0.00%	N/A	6.0000	6.0000	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	FISV	"FISV" 0.00 - 0.00%	FISV	0
4	SRCE	"1st Source Bancorp"	"NasdaqNM"	27.25	27.25	0.000000	"0.00%	N/A	4.2071	4.2071	0.00%	N/A	22.12	22.12	0.00%	"12/31/2012"	"4:00pm"	SRCE	"SRCE" 0.00 - 0.00%	SRCE	0
5	FUBC	"1st United Banc"	"NasdaqNM"	6.97	6.97	0.000000	"0.00%	N/A	6.25	6.25	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	FUBC	"FUBC" 0.00 - 0.00%	FUBC	0
6	VNET	"21st Century Group, P"	"NasdaqNM"	11.00	11.00	0.000000	"0.00%	N/A	6.19	6.19	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	VNET	"VNET" 0.00 - 0.00%	VNET	44830
7	SSBK	"360bio Inc"	"NasdaqNM"	13.96	13.96	0.000000	"0.00%	N/A	13.94	13.97	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	SSBK	"SSBK" 0.00 - 0.00%	SSBK	0
8	ESBK	"East Shore Banc"	"NasdaqNM"	5.51	5.51	0.000000	"0.00%	N/A	6.47	6.47	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ESBK	"ESBK" 0.00 - 0.00%	ESBK	0
9	EGHT	"848 Inc"	"NasdaqNM"	7.07	7.07	0.000000	"0.00%	N/A	3.76	3.76	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	EGHT	"EGHT" 0.00 - 0.00%	EGHT	100
10	AVHI	"A V Homes, Inc."	"NasdaqNM"	16.00	16.00	0.000000	"0.00%	N/A	12.22	13.54	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	AVHI	"AVHI" 0.00 - 0.00%	AVHI	0
11	SHM	"A Scholtes, Inc"	"NasdaqNM"	29.67	29.67	0.000000	"0.00%	N/A	29.96	26.02	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	SHM	"SHM" 0.00 - 0.00%	SHM	0
12	ASTM	"Astron BioSciences"	"NasdaqNM"	41.41	41.41	0.000000	"0.00%	N/A	3.96	3.96	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ASTM	"ASTM" 0.00 - 0.00%	ASTM	0
13	ASTM	"Astron BioSciences"	"NasdaqNM"	41.41	41.41	0.000000	"0.00%	N/A	3.96	3.96	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ASTM	"ASTM" 0.00 - 0.00%	ASTM	0
14	ABAX	"ABAXIS, Inc."	"NasdaqNM"	40.81	40.81	0.000000	"0.00%	N/A	37.13	37.13	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ABAX	"ABAX" 0.00 - 0.00%	ABAX	0
15	ABMD	"ABIONMED, Inc."	"NasdaqNM"	14.00	14.00	0.000000	"0.00%	N/A	14.00	14.00	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ABMD	"ABMD" 0.00 - 0.00%	ABMD	500
16	AXAS	"Axasis Petroleum"	"NasdaqNM"	2.35	2.35	0.000000	"0.00%	N/A	2.35	2.35	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	AXAS	"AXAS" 0.00 - 0.00%	AXAS	6000
17	AXAS	"Axasis Petroleum"	"NasdaqNM"	2.35	2.35	0.000000	"0.00%	N/A	2.35	2.35	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	AXAS	"AXAS" 0.00 - 0.00%	AXAS	6000
18	ACAD	"Acadia Healthcare"	"NasdaqNM"	24.25	24.25	0.000000	"0.00%	N/A	24.25	24.25	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACAD	"ACAD" 0.00 - 0.00%	ACAD	0
19	ACAD	"Acadia Healthcare"	"NasdaqNM"	24.25	24.25	0.000000	"0.00%	N/A	24.25	24.25	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACAD	"ACAD" 0.00 - 0.00%	ACAD	0
20	AXDX	"Accelerate Diagnostics"	"NasdaqNM"	10.00	10.00	0.000000	"0.00%	N/A	10.00	10.00	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"3:54pm"	AXDX	"AXDX" 0.00 - 0.00%	AXDX	0
21	ACCL	"Accell Systems"	"NasdaqNM"	9.63	9.63	0.000000	"0.00%	N/A	9.63	9.63	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACCL	"ACCL" 0.00 - 0.00%	ACCL	0
22	ACCL	"Accell Systems"	"NasdaqNM"	9.63	9.63	0.000000	"0.00%	N/A	9.63	9.63	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACCL	"ACCL" 0.00 - 0.00%	ACCL	0
23	ARAY	"Accura Corporation"	"NasdaqNM"	17.72	17.72	0.000000	"0.00%	N/A	17.72	17.72	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ARAY	"ARAY" 0.00 - 0.00%	ARAY	0
24	ARAY	"Accura Corporation"	"NasdaqNM"	17.72	17.72	0.000000	"0.00%	N/A	17.72	17.72	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ARAY	"ARAY" 0.00 - 0.00%	ARAY	0
25	ACRQ	"Accellix Pharmaceuticals"	"NGM"	44.44	44.44	0.000000	"0.00%	N/A	42.63	9.23	3.043	N/A	"12/31/2012"	"3:53pm"	ACRQ	"ACRQ" 0.00 - 0.00%	ACRQ	0			
26	ACET	"Acetate Corporation"	"NasdaqNM"	10.10	10.10	0.000000	"0.00%	N/A	10.09	9.79	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACET	"ACET" 0.00 - 0.00%	ACET	300
27	ACW	"Academy Sports & Outdoors"	"NasdaqNM"	16.00	16.00	0.000000	"0.00%	N/A	16.00	16.00	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACW	"ACW" 0.00 - 0.00%	ACW	0
28	ACW	"Academy Sports & Outdoors"	"NasdaqNM"	16.00	16.00	0.000000	"0.00%	N/A	16.00	16.00	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACW	"ACW" 0.00 - 0.00%	ACW	0
29	APKT	"Acme Packet, Inc."	"NasdaqNM"	27.25	27.25	0.000000	"0.00%	N/A	22.12	22.12	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	APKT	"APKT" 0.00 - 0.00%	APKT	0
30	ACNB	"ACNB Corporation"	"NasdaqNM"	10.00	10.00	0.000000	"0.00%	N/A	16.18	16.18	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"3:04pm"	ACNB	"ACNB" 0.00 - 0.00%	ACNB	0
31	ACOR	"Acorda Therapeutics"	"NasdaqNM"	26.81	26.81	0.000000	"0.00%	N/A	24.86	24.86	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACOR	"ACOR" 0.00 - 0.00%	ACOR	0
32	ACOR	"Acorda Therapeutics"	"NasdaqNM"	26.81	26.81	0.000000	"0.00%	N/A	24.86	24.86	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"4:00pm"	ACOR	"ACOR" 0.00 - 0.00%	ACOR	0
33	ACTS	"Actinium Solutions"	"NasdaqNM"	2.25	2.25	0.000000	"0.00%	N/A	1.61	1.60	0.00%	N/A	12.31	12.31	0.00%	"12/31/2012"	"8:00pm"	ACTS	"ACTS" 0.00 - 0.00%	ACTS	0

# 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>
```

## Quasi-Structured

Home / user / sandbox / Omniture\_0.tsv.gz

Registered User SWID (if logged in)

View As Binary

Stop

preview


Download

View File Location

Refresh

1331799426	2012-03-15 01:17:06	2860005755985467733	4611687631106657821	FAS-2.8-A53
N 0	99.122.210.248	0	10	http://www.acme.com/SH55126545/VD5517036
4	{7AAB8415-E803-3C5D-7100-E362D7F67CA7}			
		U	en-us,en;q=0.5	516 575 1366 Y
N Y	2 0 304	sbcbglobal.net	15/2/2012 4:16:0 4 240 45 41	10002,00
011,10020,00007	Mozilla/5.0 (Windows; U; windows NT 6.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6			
48 0	2 3 0	homestead usa 528 f1	0 0 0	0
				WPLG

# Unstructured



# Software Engineering for Distributed Systems


Prof. Dr. phil.-nat. Jens Grabowski

Institute of Computer Science, University of Göttingen

---

Home ▾
Staff ▾
Research ▾
Publications ▾
Awards ▾
Teaching ▾

## Our Research



### News

- Paper accepted at SAM 2018
- Article accepted in the Springer Software Quality Journal
- Two Presentations and a tutorial accepted at the ICAAT 2018
- DFG grant for DEFECTS project
- Paper accepted at the European Conference of Software Engineering Education (ECSEE 2018)
- Papers accepted for the Post-Proceedings of SimScience 2017
- Another paper accepted at CLOSER 2018
- DFG grant for GAUIS project
- Journal First Presentation at ICSE 2018
- Paper accepted at CLOSER 2018

[More news...](#)

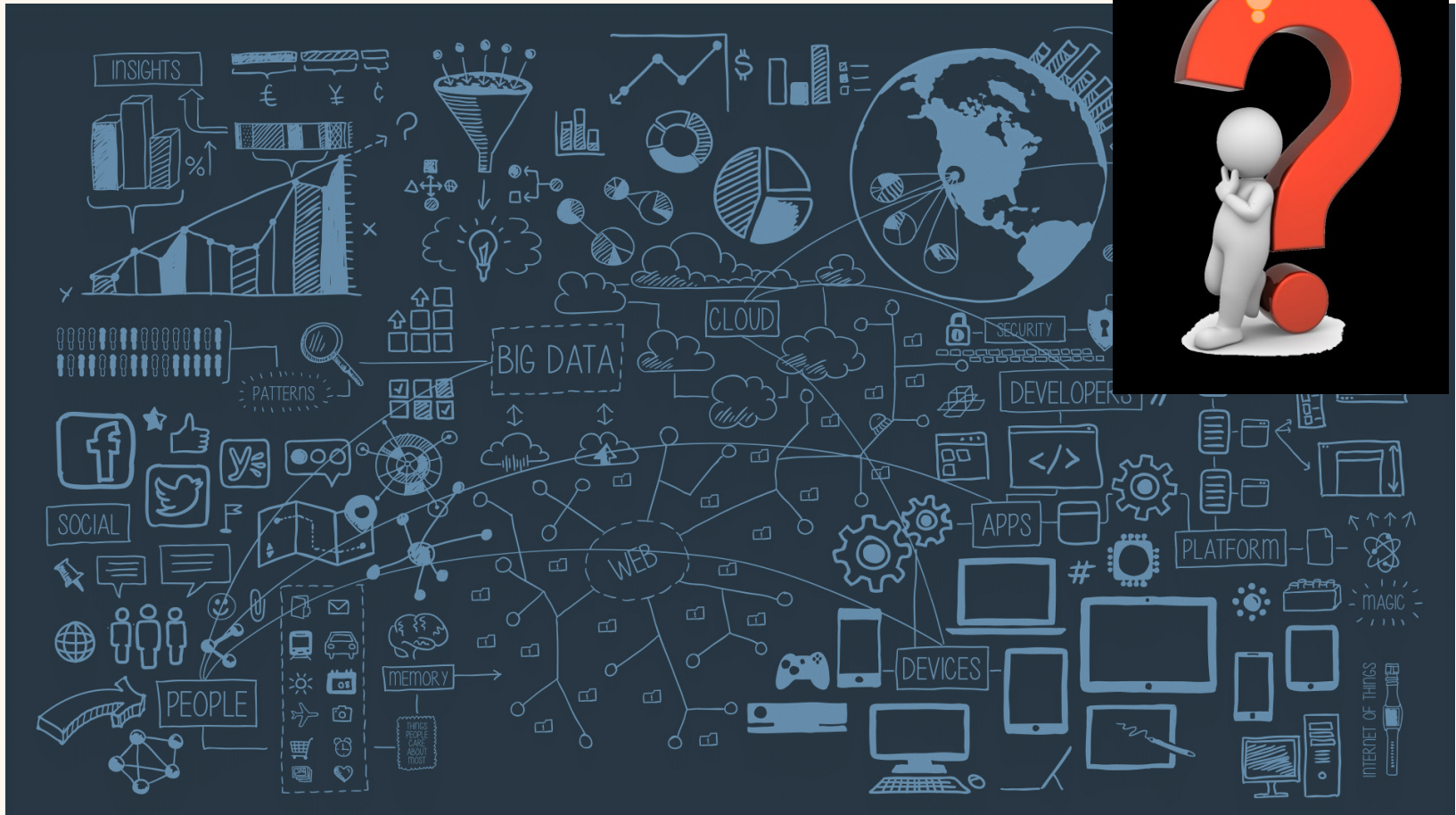
# Outline

- 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

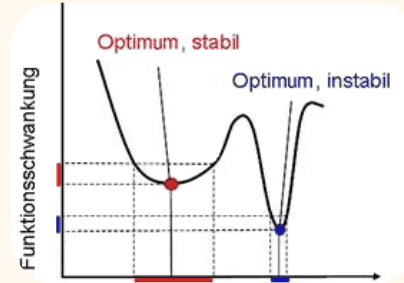
# Tools? Big Data? Machine Learning?



# Mathematical Aspects



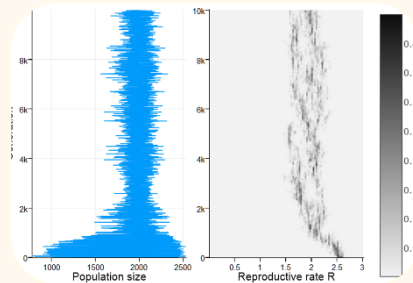
Computational  
Geometry



Optimization



Stochastics



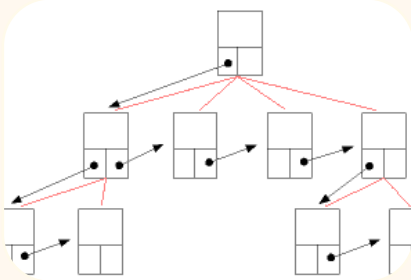
Scientific  
Computing



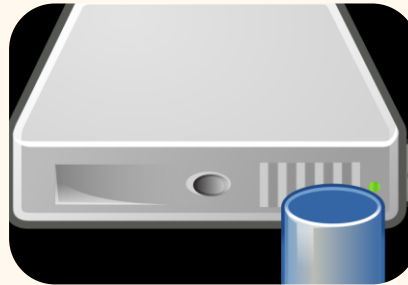
Machine  
Learning



# Computer Science Aspects



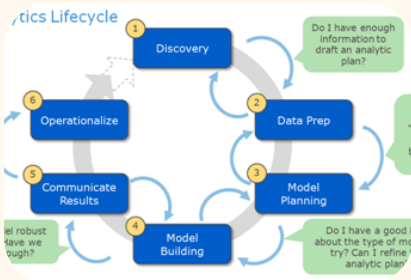
Data Structures and Algorithms



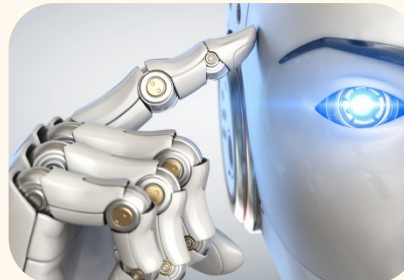
Data bases



Distributed Computing



Software Engineering

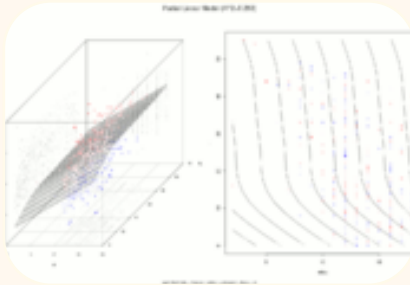


Artificial Intelligence

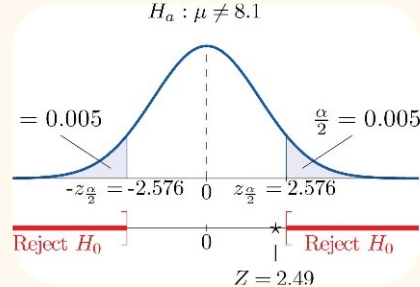


Machine Learning

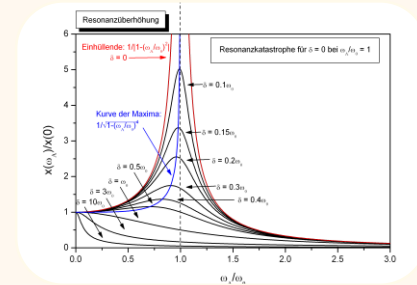
# Statistical Aspects



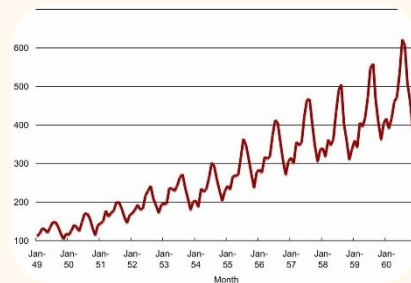
Linear Models



Statistical Tests



Inference



Time Series Analysis



Machine Learning

# Applications



Intelligent Systems



Robotics



Marketing



Medicine



Autonomous Driving

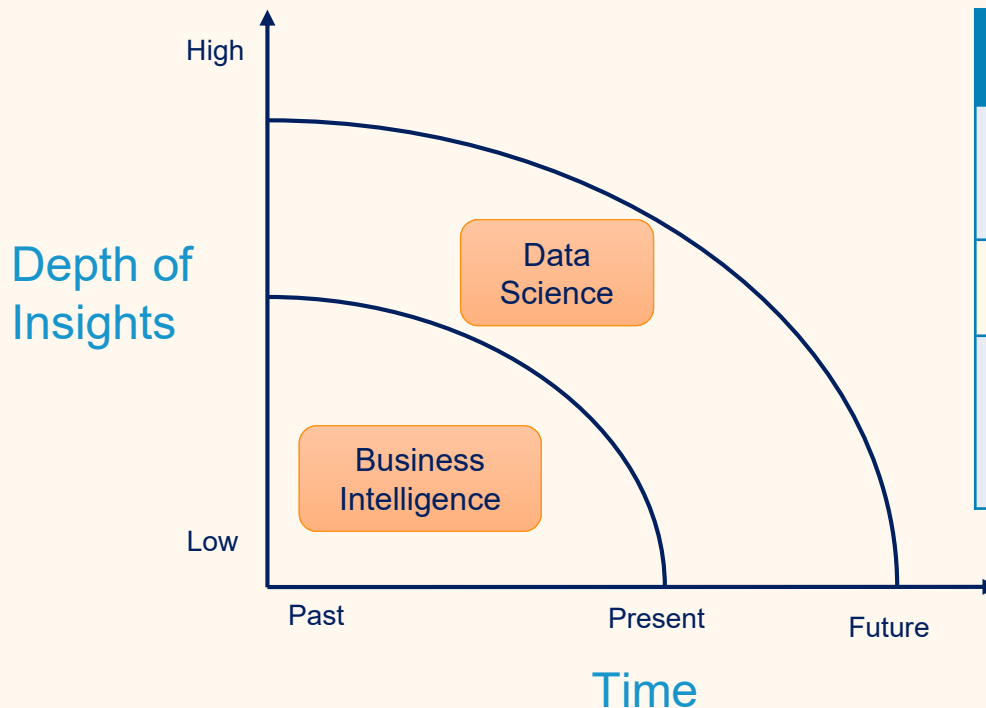


Social Networks

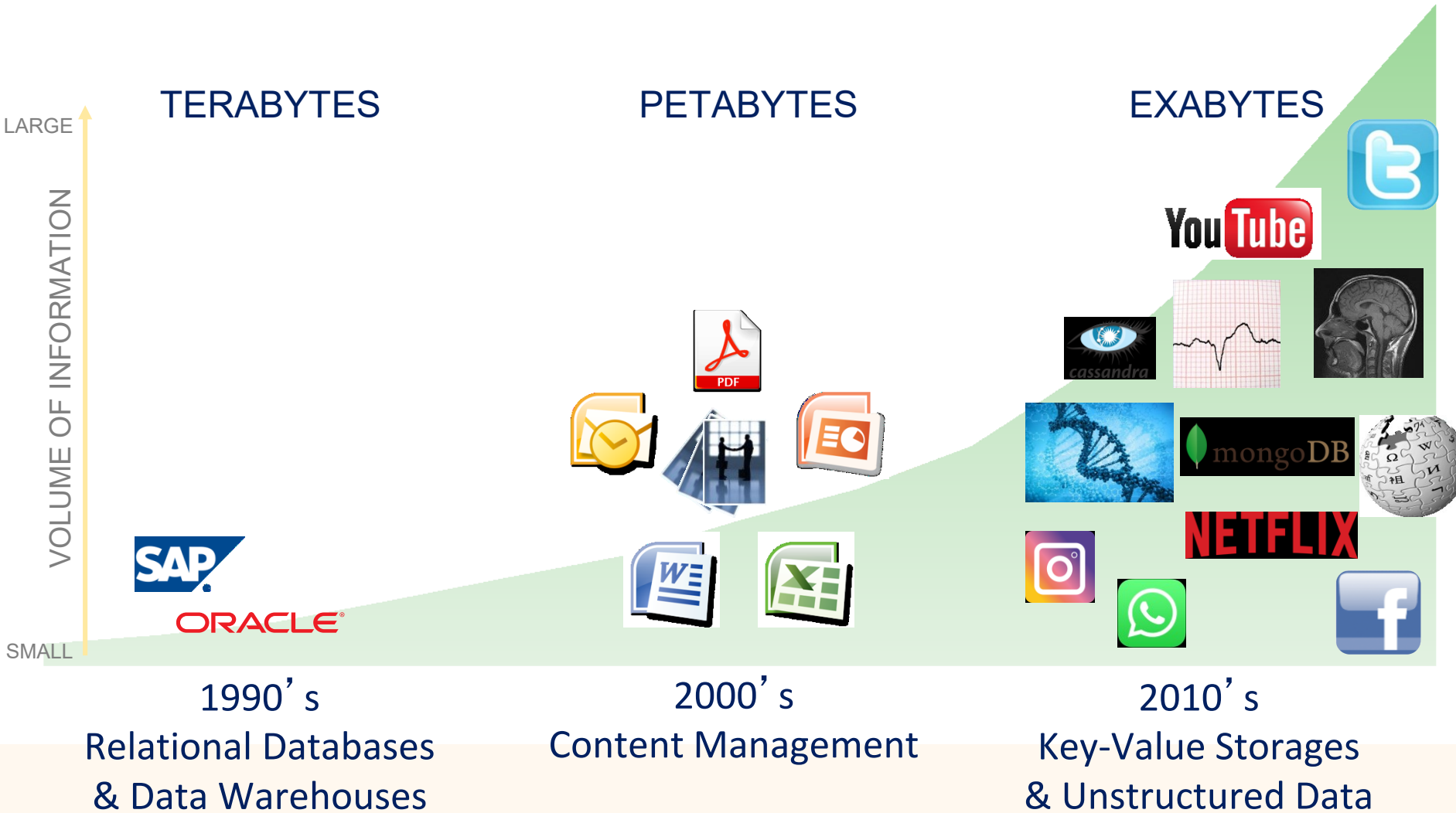


# Data Science vs. Business Intelligence

- 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	Optmization, 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...?



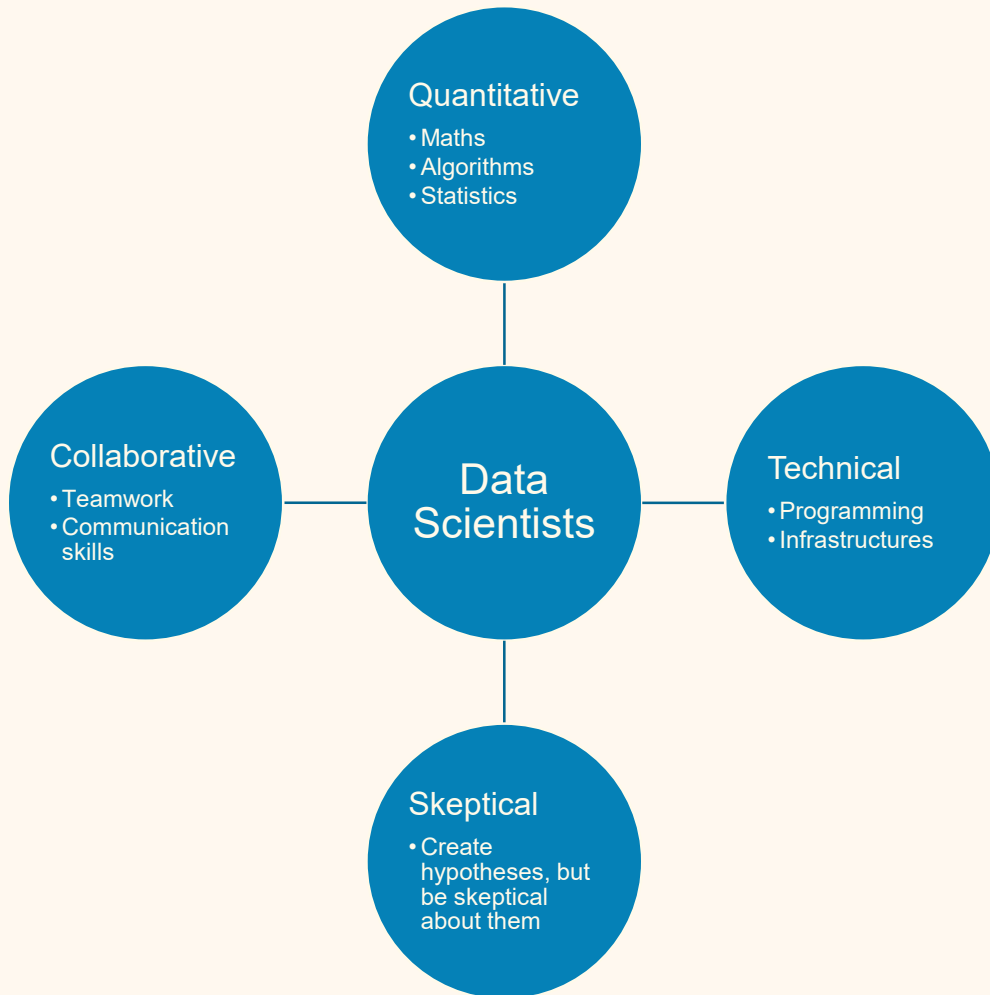
# Outline

- 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
  - „Spare time“ data scientists

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)

Data Science and Big Data Analytics

# 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