universität koblenz
weiter:denken

Big Data
Session 1: Introduction to Big Data

Frank Hopfgartner
Institute for Web Science and Technologies

Intended Learning Outcomes

universität koblenz
weiter:denken

At the end of this lecture, you will be able to:

- Define the concept of ‘Big Data’ and describe typical characteristics or properties
- Recognise benefits of using bigger data compared to smaller data
- Describe example applications and data products that can be developed using Big Data
- Describe challenges when dealing with Big Data
- Understand the implications of using Big Data

Outline

universität koblenz
weiter:denken

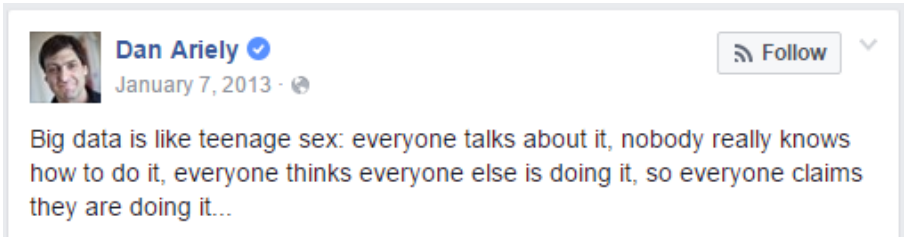
- Defining Big Data
- Big Data Analytics
- Big Data Use Cases
- Big Data Technologies

In 2013, Big Data seemed like teenage sex...

universität koblenz
weiter:denken

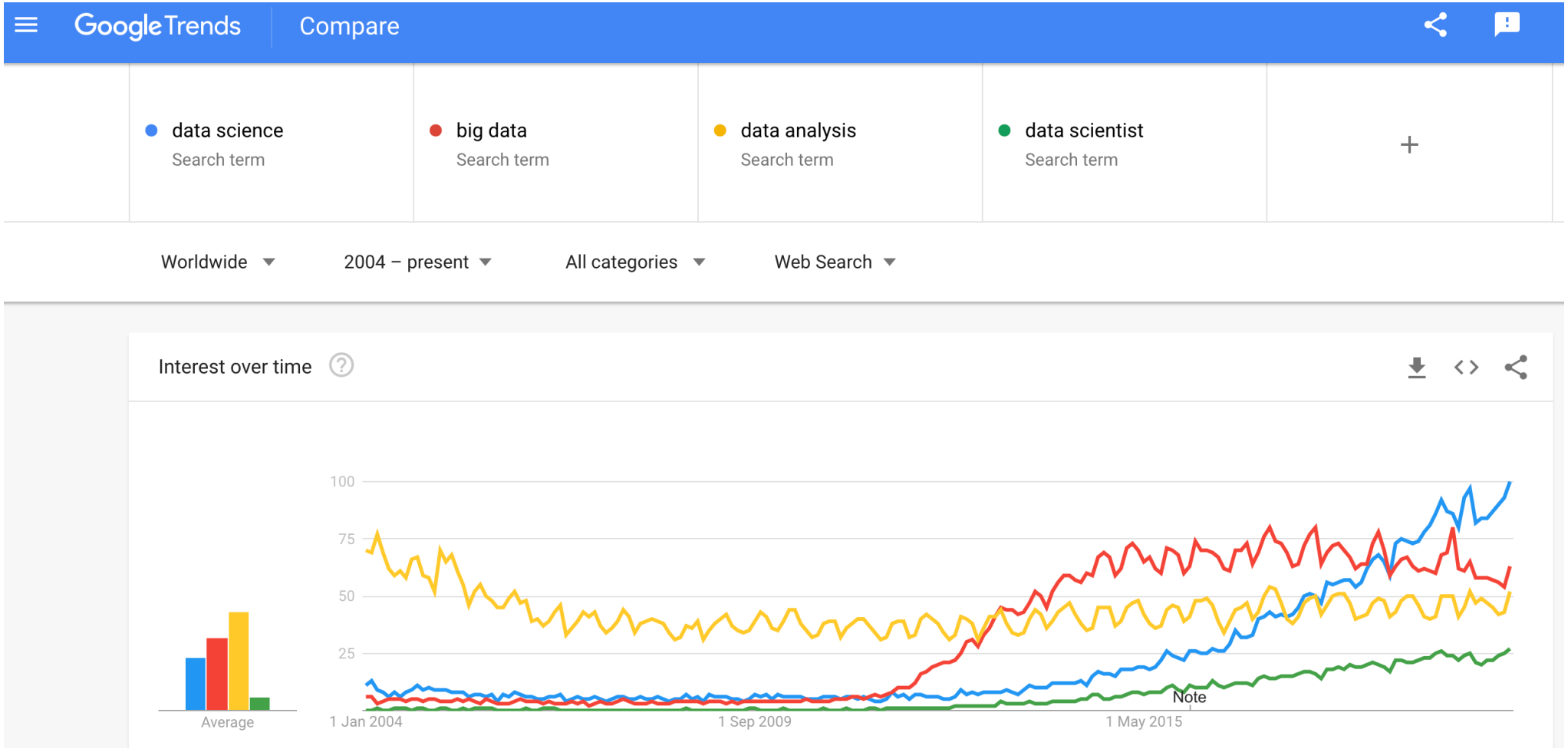
- everyone was talking about it
- nobody really knew how to do it
- everyone thought everyone else was doing it
- so everyone claimed they were doing it...

-Dan Ariely, Professor at Duke University



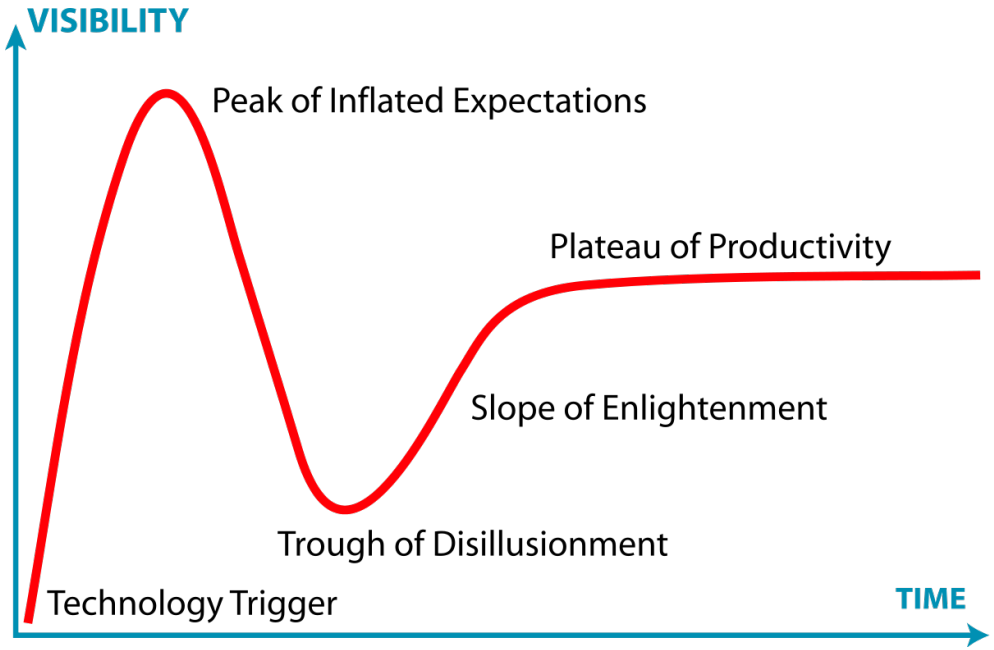
From around 2013 the term ‘Big Data’ was becoming more widely used

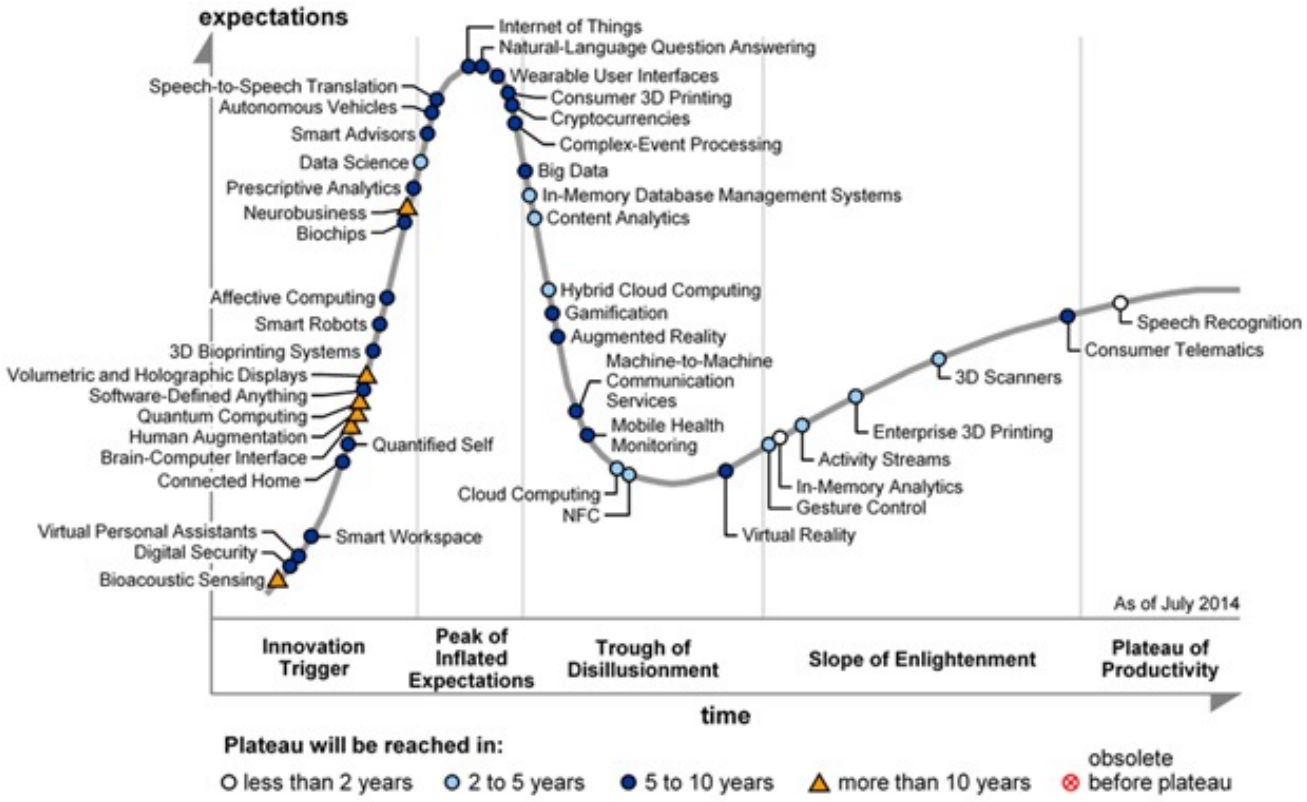
universität koblenz
weiter:denken



The Gartner Hype Cycle of Emerging Technologies, 2014

universität koblenz
weiter:denken



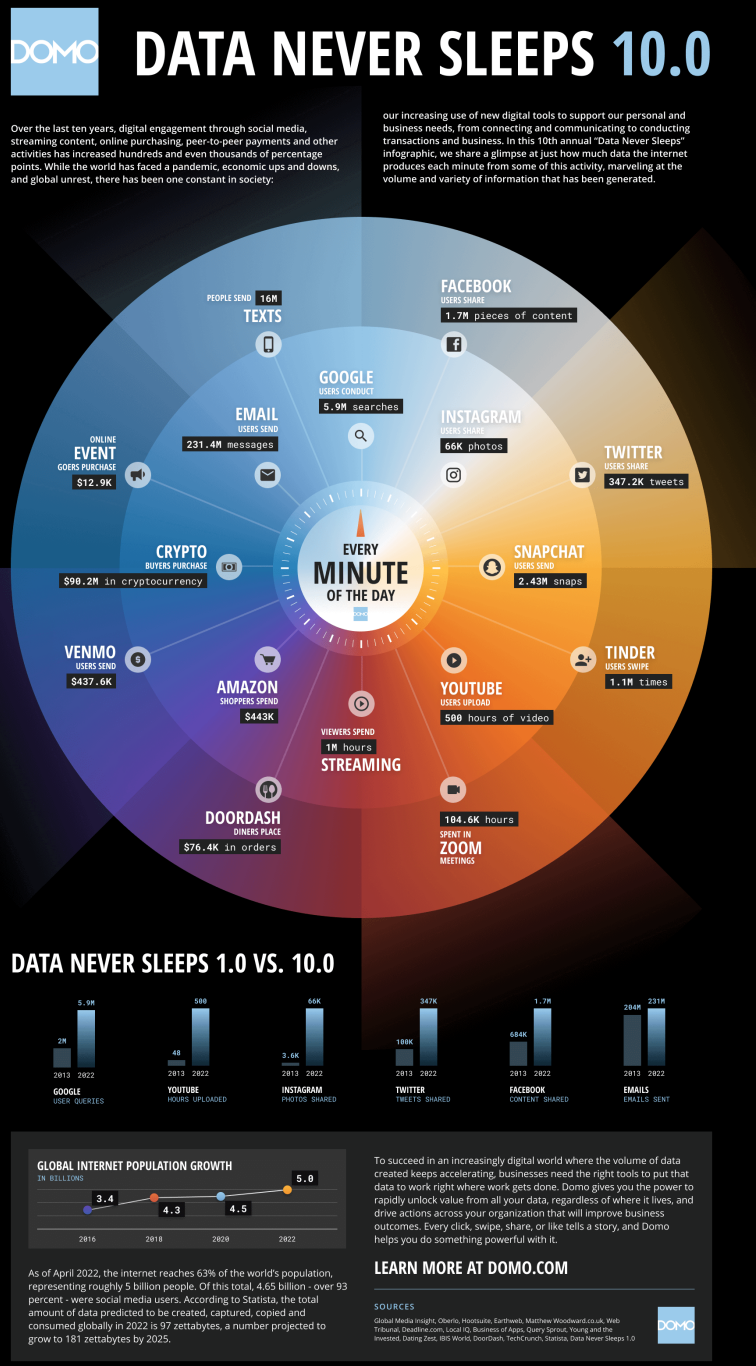


http://www.forbes.com/sites/gartnergroup/2014/09/17/digital-business-technologies-dominate-gartner-2014-emerging-technologies-hype-cycle/

Today, Big Data is reality

- Companies that are not investing in Big Data now might be out of business in 2021 (Gartner)
- Big data can (according to the European Commission):
  - increase the productivity of all sectors
  - address many challenges that face our society

https://www.domo.com/data-never-sleeps



According to the Oxford English Dictionary

universität koblenz
weiter:denken

Big data is "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges."

According to the World Economic Forum

universität koblenz
weiter:denken

Source: https://www.youtube.com/watch?v=eVSfJhssXUA

Your turn

universität koblenz
weiter:denken

In pairs or groups think about the following:

- How would you define ‘Big Data’?
- How does Big Data fit in with Data Science?
- What challenges might you face when collecting and working with Big Data?
- What can you do with bigger data that you can’t do with smaller data?

What is Big Data? An early definition from 2001

universität koblenz
weiter:denken

- ***Volume*** *(amount of data)*: the large amount of data being generated and stored (normally in the order of PBs or ZBs)
- ***Variety*** *(forms of data)*: the range of data types and sources being used, including unstructured data
- ***Velocity*** *(speed of data)*: the rate at which data is collected, shared and analysed - often real time streaming data (e.g., from social media)

Doug Laney, Vice President and Distinguished Analyst with the Gartner Chief Data Officer research and advisory team

Volume

universität koblenz
weiter:denken

Name	Equals to	Size in bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2
Byte	8 bits	1
Kilobyte	1024 bytes	1024
Megabyte	1024 kilobytes	1,048,576
Gigabyte	1024 megabytes	1,073,741,824
Terrabyte	1024 gigabytes	1,099,511,627,776
Petabyte	1024 terrabytes	1,125,899,906,842,624
Exabyte	1024 petabytes	1,152,921,504,606,846,976
Zettabyte	1024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1024 zettabytes	1,208,925,819,614,629,174,706,176

Volume Example Facebook in 2020

- 3.1 billion active users of their services (Facebook, Instagram, WhatsApp, etc.)
- 5 billion comments are left on Facebook pages monthly
- 100 million hours of videos watched every day
- 147,000 pictures uploaded every second

Inside Facebook's Oregon Data Center (CNET News)

universität koblenz
weiter:denken

Source: https://www.youtube.com/watch?v=4A_A-CmrqpQ

Velocity Examples

universität koblenz
weiter:denken

- **Alibaba in 2019**
  - 1.3 billion delivery orders on 11/11/2019
- **Google in 2019**
  - 3.5 billion searches on average per day

https://graphics.reuters.com/SINGLES-DAY-ALIBABA/0100B30E24T/index.html

Example: Most Tweeted Event in Sports History (35.6 million Tweets)

universität koblenz
weiter:denken

Variety Examples

Datafication: *"... taking all aspects of life and turning them into data. Google's augmented-reality glasses datafy the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks."*

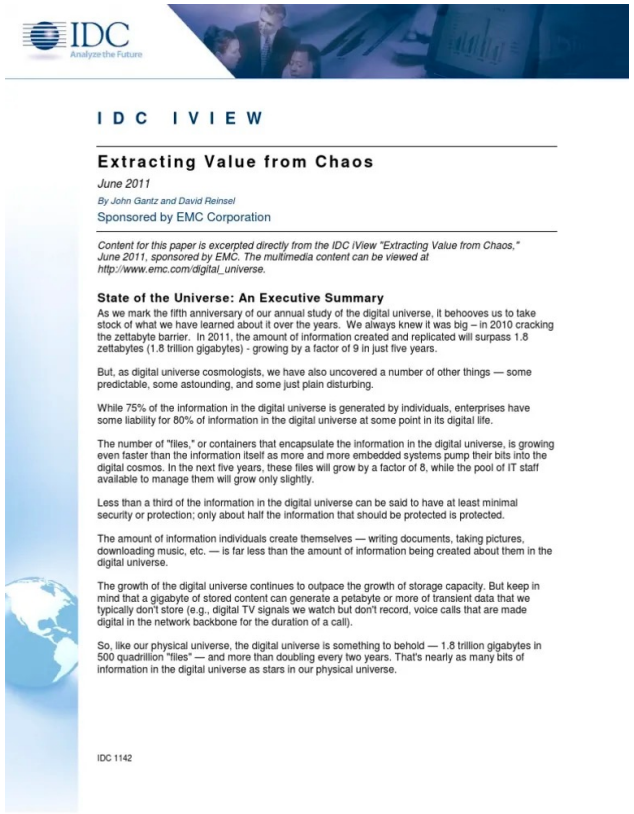
"A 2012 survey by NewVantage Partners of over fifty executives in large organisations suggests that for large companies, the lack of structure of data is more salient than addressing its size."
(Davenport, 2014)

Estimated that 95% of Big Data is unstructured

Numerical, categorical or binary
Text
Records
Chemical structures
Biomedical
Geo-spatial
Network
Sensor
Images
Sounds
Video
.....

Veracity

Veracity *(reliability of data)*: uncertainty in data quality (accuracy, provenance, relevance and consistently)



IDC

IDC IVIEW

Extracting Value from Chaos

June 2011

By John Gantz and David Reinsel

Sponsored by EMC Corporation

Content for this paper is excerpted directly from the IDC iView "Extracting Value from Chaos," June 2011, sponsored by EMC. The multimedia content can be viewed at http://www.emc.com/digital_universe.

State of the Universe: An Executive Summary

As we mark the fifth anniversary of our annual study of the digital universe, it behooves us to take stock of what we have learned about it over the years. We always knew it was big – in 2010 cracking the zettabyte barrier. In 2011, the amount of information created and replicated will surpass 1.8 zettabytes (1.8 trillion gigabytes) - growing by a factor of 9 in just five years.

But, as digital universe cosmologists, we have also uncovered a number of other things — some predictable, some astounding, and some just plain disturbing.

While 75% of the information in the digital universe is generated by individuals, enterprises have some liability for 80% of information in the digital universe at some point in its digital life.

The number of "files," or containers that encapsulate the information in the digital universe, is growing even faster than the information itself as more and more embedded systems pump their bits into the digital cosmos. In the next five years, these files will grow by a factor of 8, while the pool of IT staff available to manage them will grow only slightly.

Less than a third of the information in the digital universe can be said to have at least minimal security or protection; only about half the information that should be protected is protected.

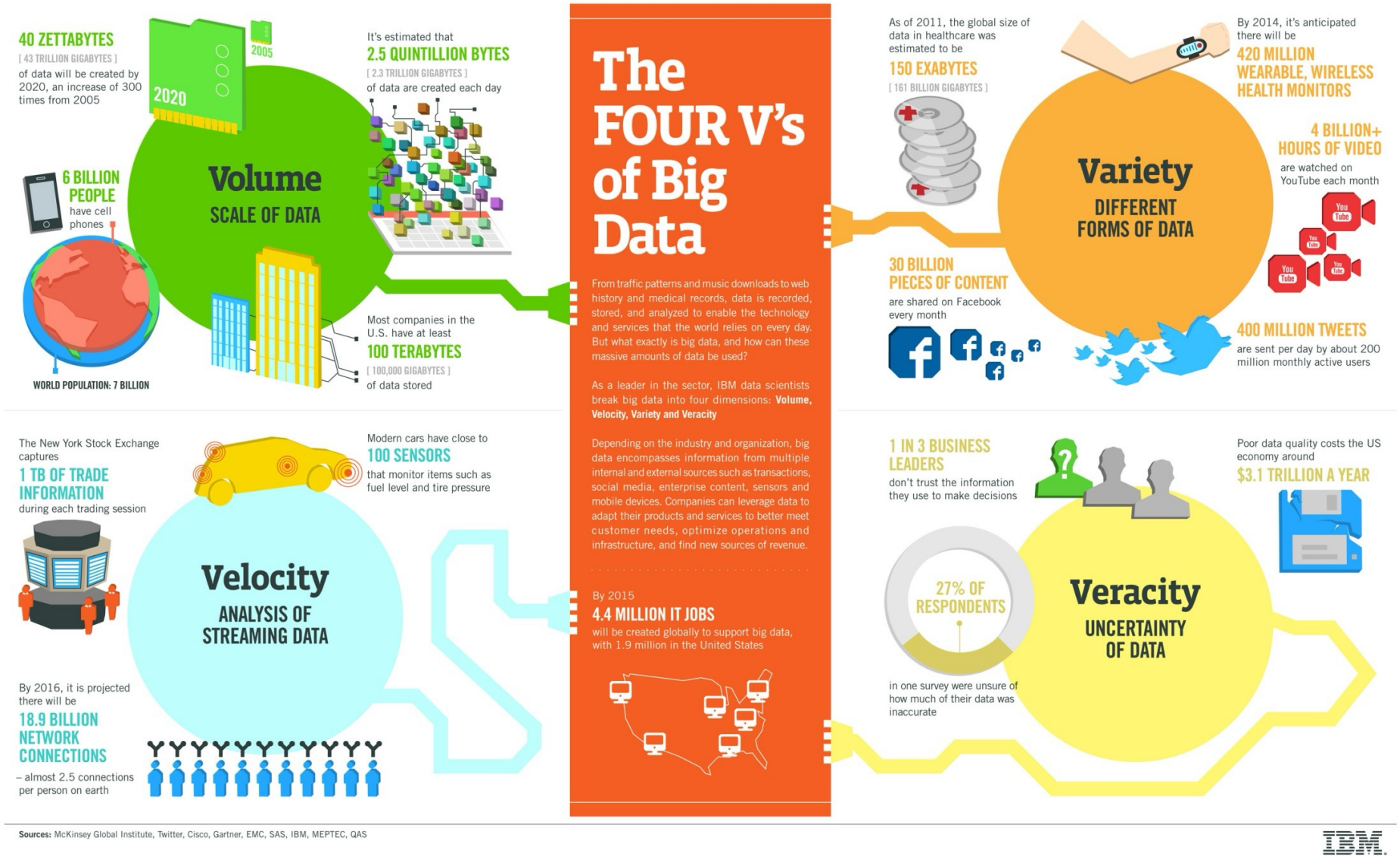
The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. — is far less than the amount of information being created about them in the digital universe.

The growth of the digital universe continues to outpace the growth of storage capacity. But keep in mind that a gigabyte of stored content can generate a petabyte or more of transient data that we typically don't store (e.g., digital TV signals we watch but don't record, voice calls that are made digital in the network backbone for the duration of a call).

So, like our physical universe, the digital universe is something to behold — 1.8 trillion gigabytes in 500 quadrillion "files" — and more than doubling every two years. That's nearly as many bits of information in the digital universe as stars in our physical universe.

IDC 1142

Gantz, J. and Reinsel, D. (2011). Extracting Value from Chaos

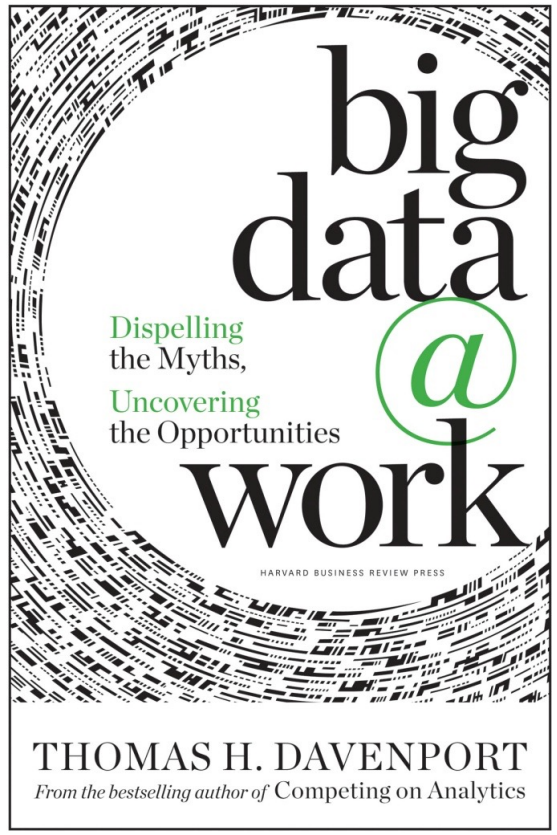


Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. P. (2012). *Analytics: The real-world use of big data*. NY, USA: IBM Institute for Business Value.

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

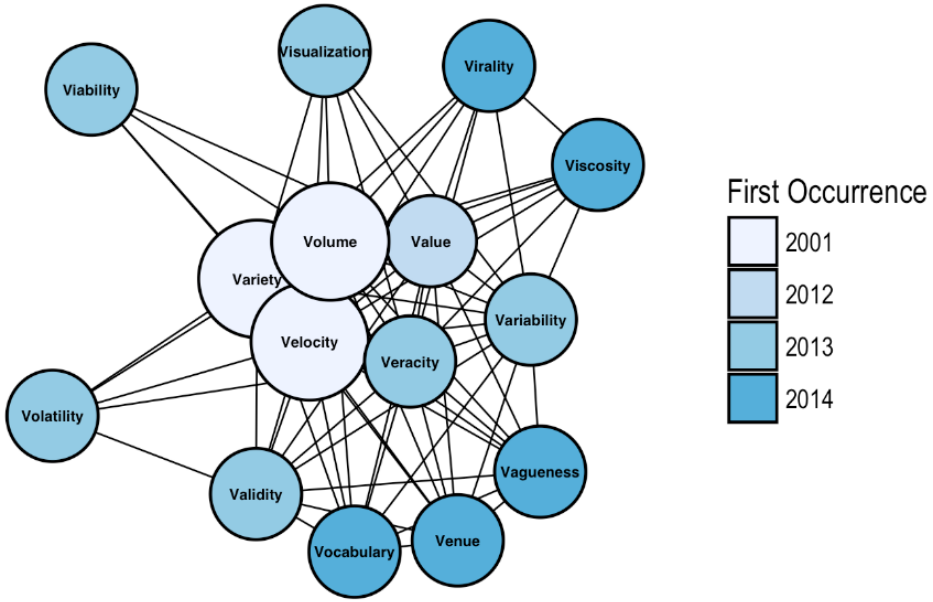
Value

- At least three classes of value (Davenport, 2014)
  - Cost reductions (e.g., use of Big Data technologies)
  - Improvements in decision-making
  - Improvements in products and services (e.g., People You May Know or PYMK feature in LinkedIn)



Additional V's

Validity: selection of appropriate data with respect to the intended use.
Volatility: the extent to which data can be relied upon to remain available
Variability: meaning of the data
Visualisation: transforming data into insights

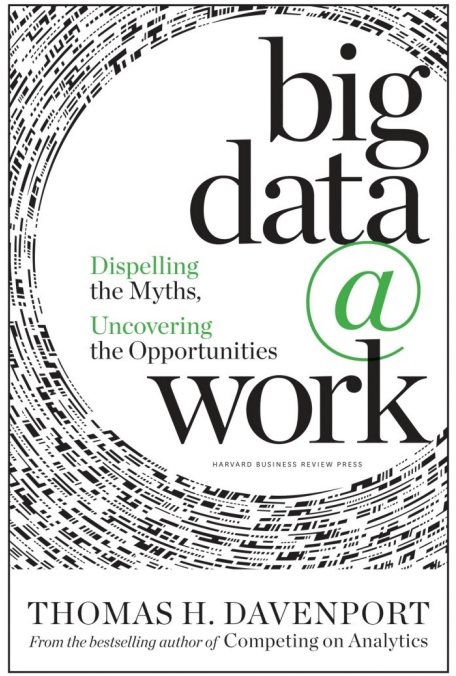


https://www.elderresearch.com/blog/42-v-of-big-data

What's in a name?

universität koblenz
weiter:denken

- Davenport (2014) highlights a number of problems with the name 'Big Data' (not the idea), including
  - 'Big' is only one aspect of what's distinctive about new forms of data (structure often the bigger problem)
  - 'Big' is relative and will change
  - If the data doesn't fit all the Vs is it still Big Data?
  - The term 'Big Data' is being misused by vendors and marking companies to refer to any analytics and reporting



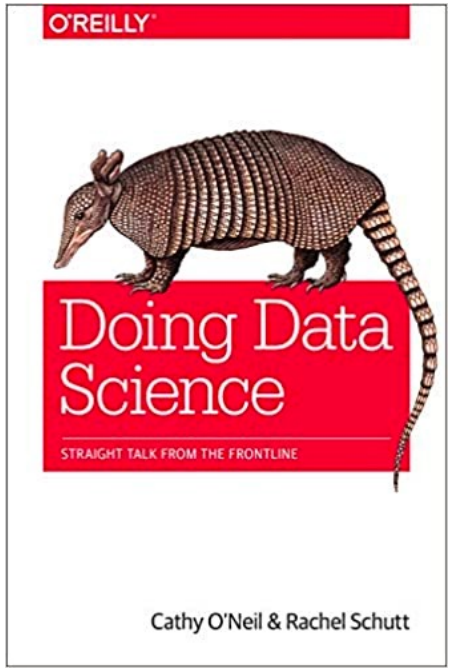
"The point is not to be dazzled by the volume of data, but rather to analyse it – to convert it into insights, innovations and business value." (Davenport, 2014)

Going beyond the V's…

universität koblenz
weiter:denken

- **"Big" is a moving target.**
  - Only when the size becomes a challenge is it worth referring to it as big
  - Big if size of the data outstrips current computational solutions
- **"Big" is when you can't fit it on one machine.**
  - Different individuals and companies have different resources
- **Big Data is a cultural phenomenon.**
  - It describes how much data is part of our lives, precipitated by accelerated advances in technology.

O'Neil, C. and Schutt. Doing Data Science, p. 24, O'Reilly, 2013.



Similar definitions

universität koblenz
weiter:denken

> "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.
>
> *Dumbill, E. Making Sense of Big Data, Big Data, 1(1):1-2, 2012, https://doi.org/10.1089/big.2012.1503*

> "[We] characterize big data using the idea of a "big data trajectory," which refers to the activities of a group of researchers who (1) set out to change the collective set of data available to address one or more shared problems of interest in such a way that (2) the researchers believe existing methods or resources available to the group are not adequate for the project and (3) they believe acquiring these methods or resources poses specific research problems separate from the original problems of shared interest.
>
> *Sterner & Franz. Taxonomy for Humans or Computers: Cognitive Pragmatics for Big Data. Biological Theory, 12:99-111, 2017, https://doi.org/10.1007/s13752-017-0259-5*

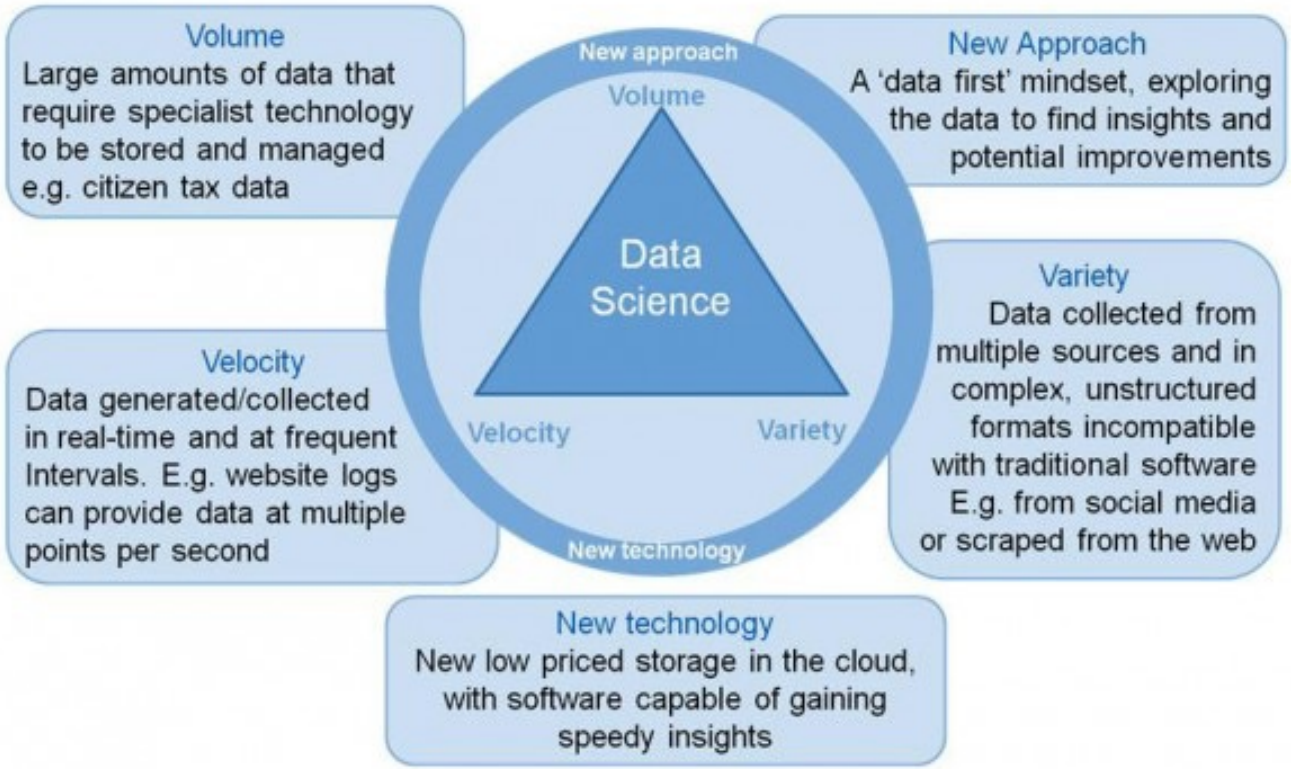
Outline

universität koblenz
weiter:denken

- Defining Big Data
- **Big Data Analytics**
- Big Data Use Cases
- Big Data Technologies

Big Data has changed how organisations operate

universität koblenz
weiter:denken



https://openpolicy.blog.gov.uk/2014/05/30/data-science-1/

GOV.UK

Blog

Open Policy Making

Organisations: Civil Service

The Data Science in Government programme: using data in new ways to improve what government does

Evaluation of data analytics

universität koblenz
weiter:denken

Term	Time frame	Specific meaning
Decision support	1970-1985	Use of data analysis to support decision making
Executive support	1980-1990	Focus on data analysis for decisions by senior executives
Online Analytical Processing (OLAP)	1990-2000	Software for analysing multidimensional data tables
Business Intelligence	1989-2005	Tools to support data-driven decision, with emphasis on reporting
Analytics	2005-2010	Focus on statistical and mathematical analysis for decisions
Big Data	2010-present	Focus on very large, unstructured, fast-moving data

The idea of analysing data to **make sense** *of what's happening in our businesses has been with us for a long time (in corporations since at least 1954 when UPS started an analytics group), so why do we have to keep coming up with new names to describe it?* (Davenport, 2014:10)

Does Big Data = traditional data analytics?

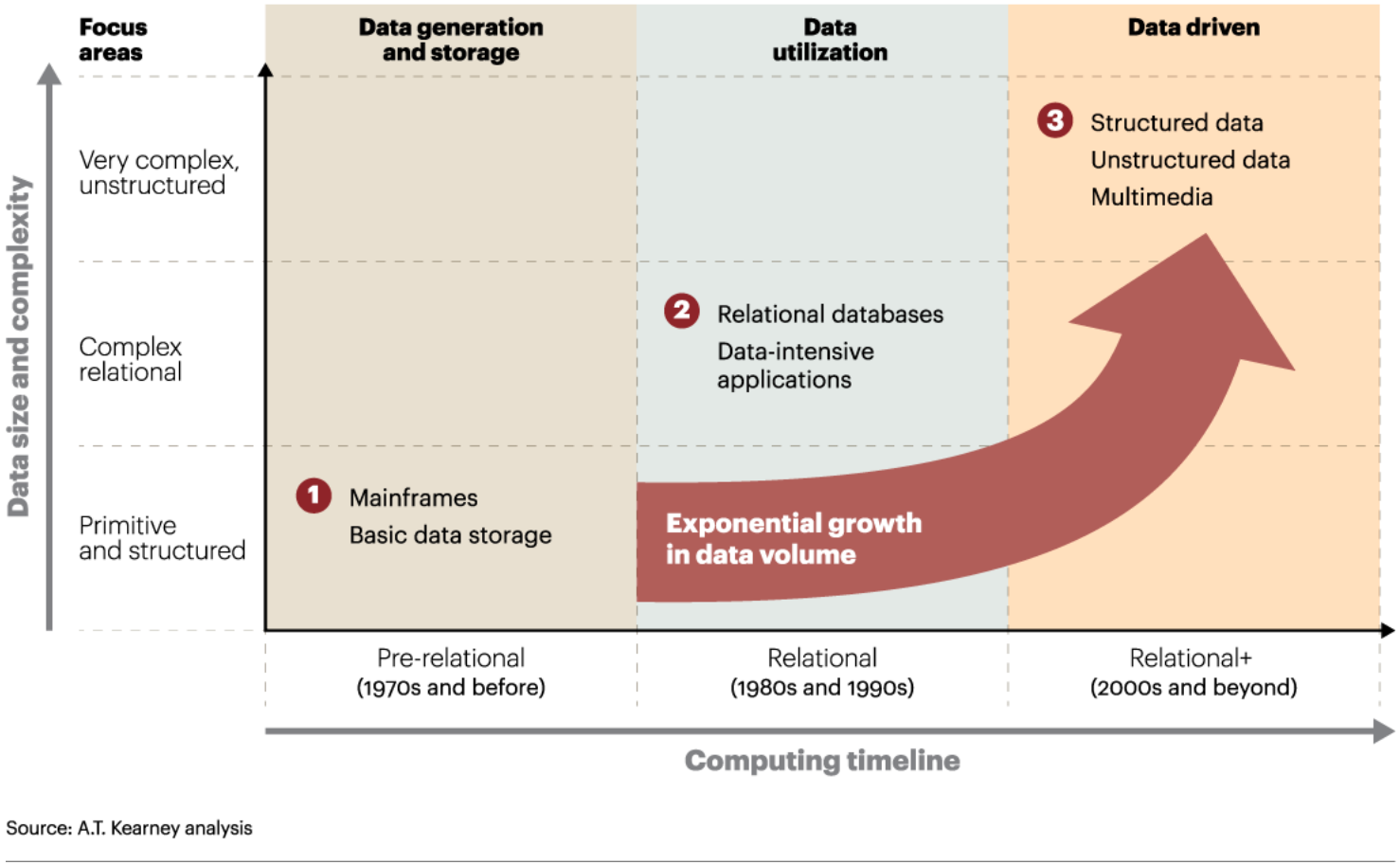
universität koblenz
weiter:denken

	Big Data	Traditional analytics
Type of data	Unstructured formats	Formatted in rows and columns
Volume of data	100 Terabytes to Petabytes	Tens of Terabytes or less
Flow of data	Constant flow of data	Static pool of data
Analysis methods	Machine learning	Hypothesis-based
Primary purpose	Data-based products	Internal decision support and services

Source: (Davenport, 2014:4)

The evolution of Big Data

universität koblenz
weiter:denken



Source: https://www.co.kearney.com/analytics/article?/a/big-data-and-the-creative-destruction-of-today-s-business-models

Big Data revolution

Mayer-Schönberger & Cukier (2013) argue that the Big Data revolution consists of

- Collecting large amounts of data rather than smaller samples (from some to all, i.e. N=ALL)
- Tolerating inaccuracies in larger amounts of data compared to higher quality smaller amounts (from clean to messy)
- Giving up on knowing the causes and accept only associations
- *"Using big data will sometimes mean forgoing the quest for why in return for knowing what"*

V. Mayer-Schönberger and K. Cukier (2013). Big Data. A revolution that will transform how we live, work, and think. John Murray.

Outline

universität koblenz
weiter:denken

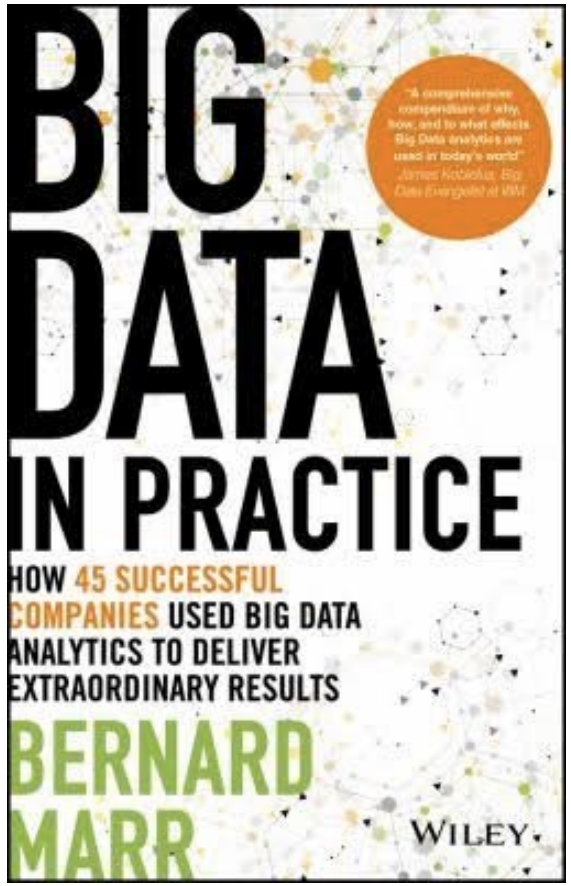
- Defining Big Data
- Big Data Analytics
- **Big Data Use Cases**
- Big Data Technologies

universität koblenz
weiter:denken

Who uses Big Data?

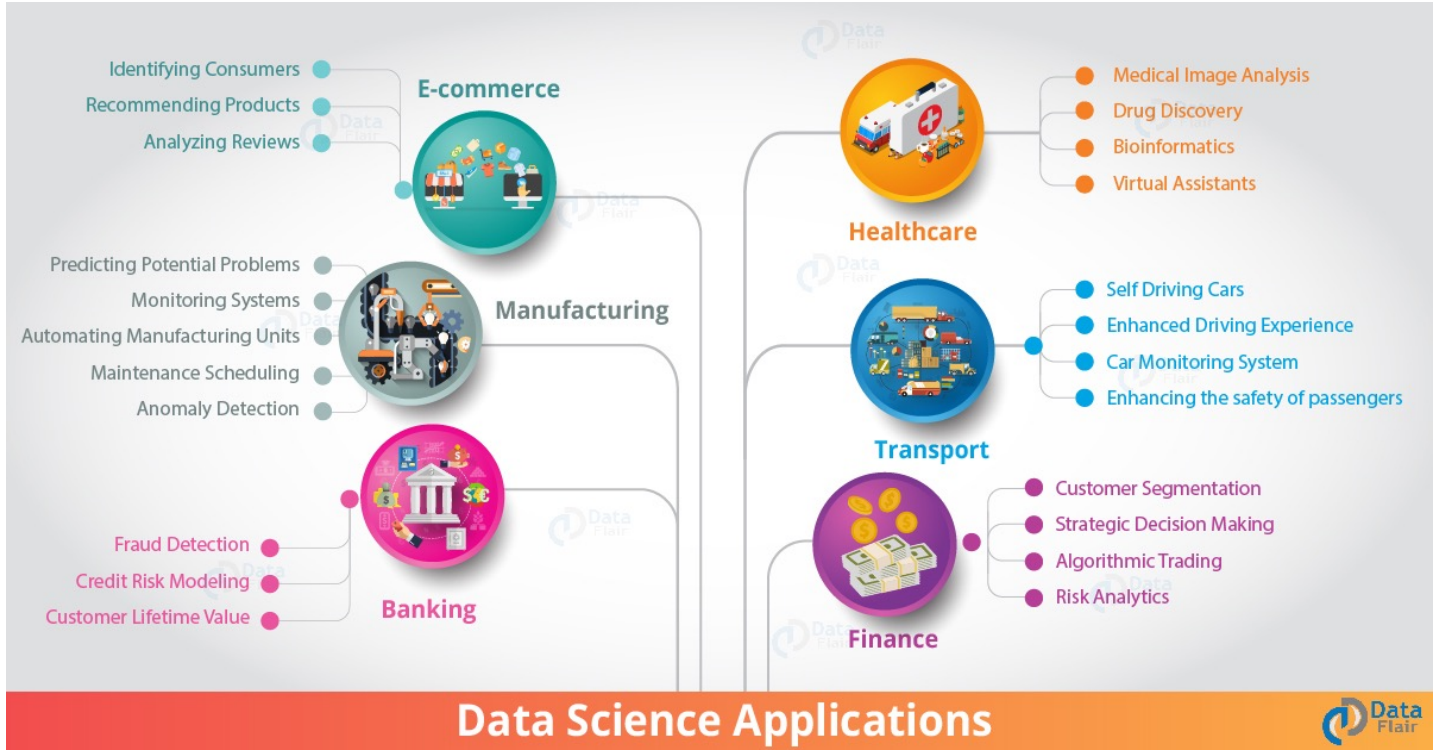
universität koblenz
weiter:denken

- Internet Companies
- Banking
- Government
- Health Care
- Manufacturing
- Retail



B. Marr. Big Data in Practice. Wiley

Use of Big Data in Business



Big data will spell the death of customer segmentation and force the marketer to understand each customer as an individual within 18 months or risk being left in the dust.

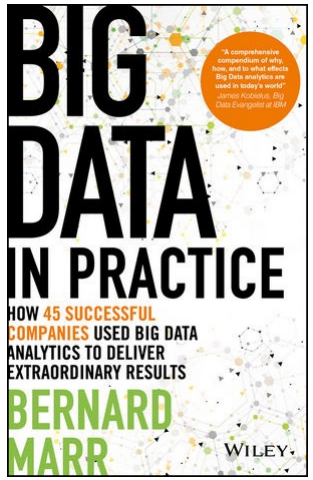
Ginni Rometty, CEO
IBM

Measurable benefits

universität koblenz
weiter:denken

- Businesses that use big data saw a **profit increase of 8-10%** (Entrepreneur, 2019)
- Businesses that use big data saw a **10 percent reduction in overall costs** (Entrepreneur, 2019)
- Nearly 50 percent of businesses say big data and analytics have fundamentally **changed business practices** in their sales and marketing departments. (Forbes, 2018)
- 62 percent of retail businesses see **competitive advantages** from information and analytics. (Towards Data Science, 2018)
- …

€ $
¥ £



Source: https://learn.g2.com/big-data-statistics

Target: Predictive analytics (in 2012)

universität koblenz
weiter:denken

FEB 16, 2012 @ 11:02 AM 3,308,494 The Little Black Book of Billionaire Secrets

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF
Welcome to The Not-So Private Parts where technology & privacy collide **FULL BIO**

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT +2.92%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial

Source: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#6f4aee0a6668

Walmart

universität koblenz
weiter:denken

- Data: Supermarket product sales
- Having the right products in stock (logistics)
- 40PB of transactional data (database of 200B sales)
- Need for real-time analysis:
  - “If you can’t get insights until you’ve analysed your sales for a week or a month, then you’ve lost sales within that time.”
  - Spotting problems within 20 minutes

Walmart: Big Data, Big Opportunities

universität koblenz
weiter:denken

Source: https://www.youtube.com/watch?v=m5Iij3zQp_w

IBM Watson (in 2011)

universität koblenz
weiter:denken

- Natural Language Processing
- Machine Learning
- 90 Servers / Hadoop
- Data: Wikipedia, news, stats

Details: Rob High. The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. Redbooks, IBM, 2012.

IBM Watson (in 2011)

universität koblenz
weiter:denken

Source: https://www.youtube.com/watch?v=WFR3lOm_xhE

Netflix

universität koblenz
weiter:denken

- World's most valued media company ($164 million)
- 151 million subscribers, **93% customer retention rate**
- 75% of viewer activity based off personalization recommendations
- **Big data analytics used to identify user preferences**
  - content recommendation
  - custom marketing
  - investing in original content

Details: Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (December 2015), 19 pages. DOI: http://dx.doi.org/10.1145/2843948

Amazon

universität koblenz
weiter:denken

- World's largest online marketplace, AI assistant provider, …
- Big Data analytics at its core, e.g.,
  - **Personalised recommendation system** (responsible for 35% of annual sales)
  - **Anticipatory Shipping model** predicts which products you are likely to purchase (and when) and distributes them to local distribution centres
  - **Supply Chain optimization** to identify warehouse closest to their customers, reducing shipping costs by 10-40%

Uber

universität koblenz
weiter:denken

- App+Data about journeys
  - No cars owned
  - No drivers employed
- Predict demand, allocate resources, set fares
  - Dynamic pricing
  - Rating data
  - External data: traffic, public transport
- Legal/ethical aspects

Details: Judd Cramer and Alan B. Krueger. Disruptive Change in the Taxi Business: The Case of Uber. Working Paper 22083, 2016. DOI: 10.3386/w22083

SMEs

universität koblenz
weiter:denken

- Local butcher in London
- Competition from supermarkets
- Installed sensor at store window
  - How many people pass by / stop and when
  - SaaS to avoid initial HW/SW setup cost
- Google trends
  - What food items are popular
- Open at night and offer sandwiches and meals

Details: Chapter 8, Big Data in Practice

DHL - Logistics

universität koblenz
weiter:denken

- Inventory & Asset Optimisation
- Transport & Delivery Optimisation
- Supplier Risk & Due Diligence Assessment
- Customer Management

Big Data Analytics at DHL

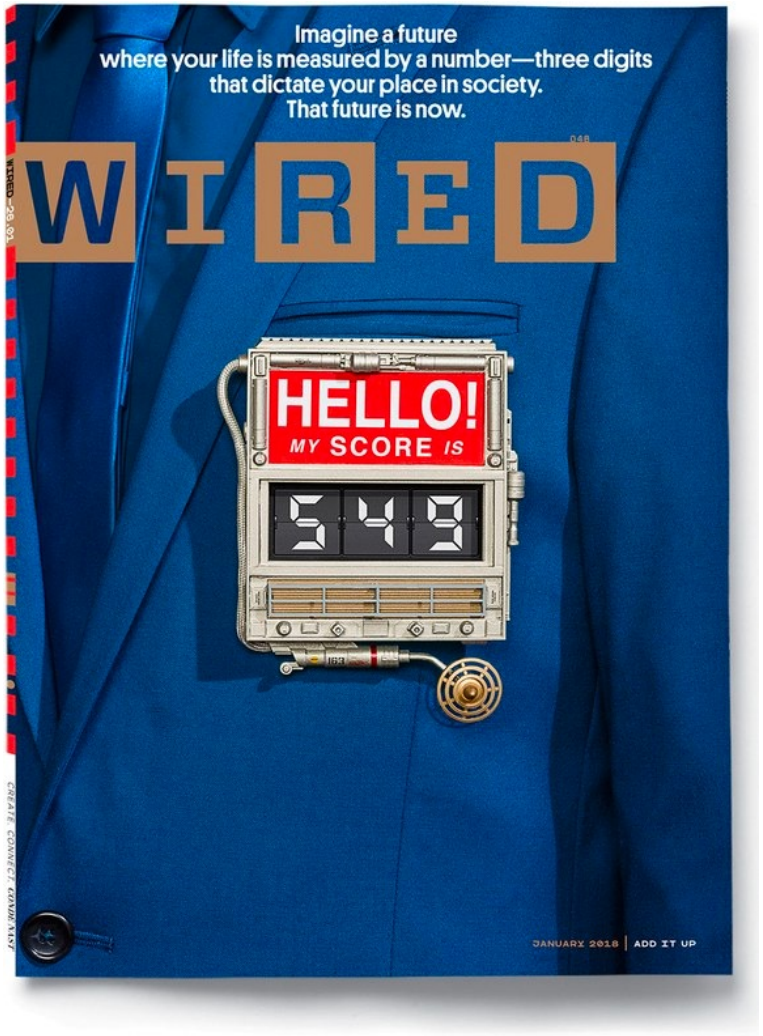
Sports Analytics

universität koblenz
weiter:denken

N. J. Kim and J. K. Park, "Sports analytics & risk monitoring based on hana platform: Sports related big data & IoT trends by using HANA In-memory platform," *2015 International SoC Design Conference (ISOCC)*, Gyeongju, Korea (South), 2015, pp. 221-222

Goverments: Zhisama (Sesame) Credit Score

‘people will be penalized for the crime of spreading online rumours, among other offenses, and that those deemed “seriously untrustworthy” can expect to receive substandard services’
‘ensure that the bad people in society don’t have a place to go, while good people can move freely and without obstruction’



https://www.wired.com/story/age-of-social-credit/

Outline

universität koblenz
weiter:denken

- Defining Big Data
- Big Data Analytics
- Big Data Use Cases
- **Big Data Technologies**

universität koblenz
weiter:denken

How to get started?

Options: Scale-up vs. scale-out

universität koblenz
weiter:denken

- Scale-up
  - Increasing the power of your computer (i.e, disk, memory, processor) only possible up to a certain extent
- Scale-out
  - Use many standard computers and distribute data and computation over them

Facebook's Data Centre in Sweden

Amazon Web Services (AWS)

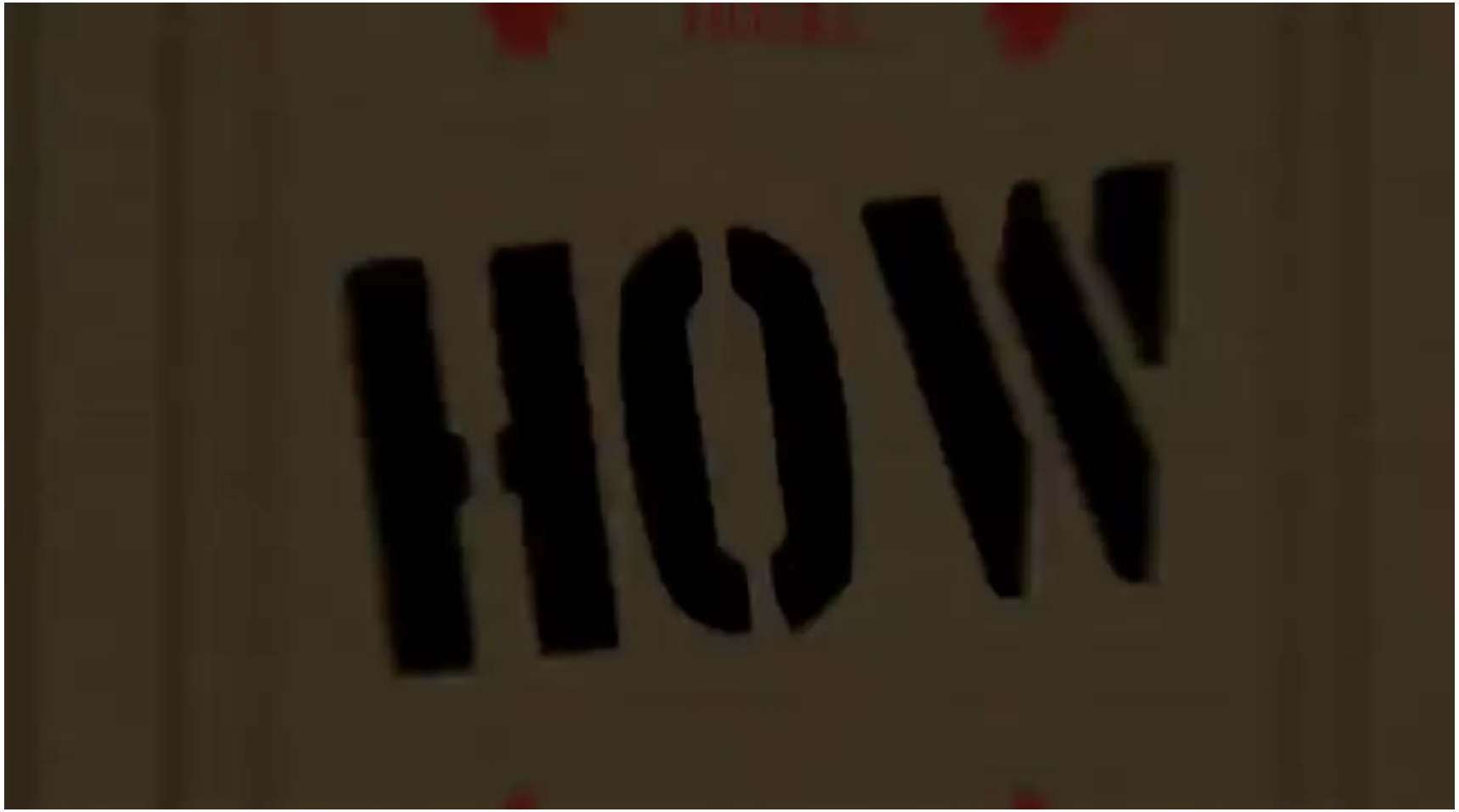
universität koblenz
weiter:denken

AWS 2022 revenues: $80 billion

https://aws.amazon.com/about-aws/global-infrastructure/

Amazon: How big is the Web's biggest e-Retailer? (Bloomberg)

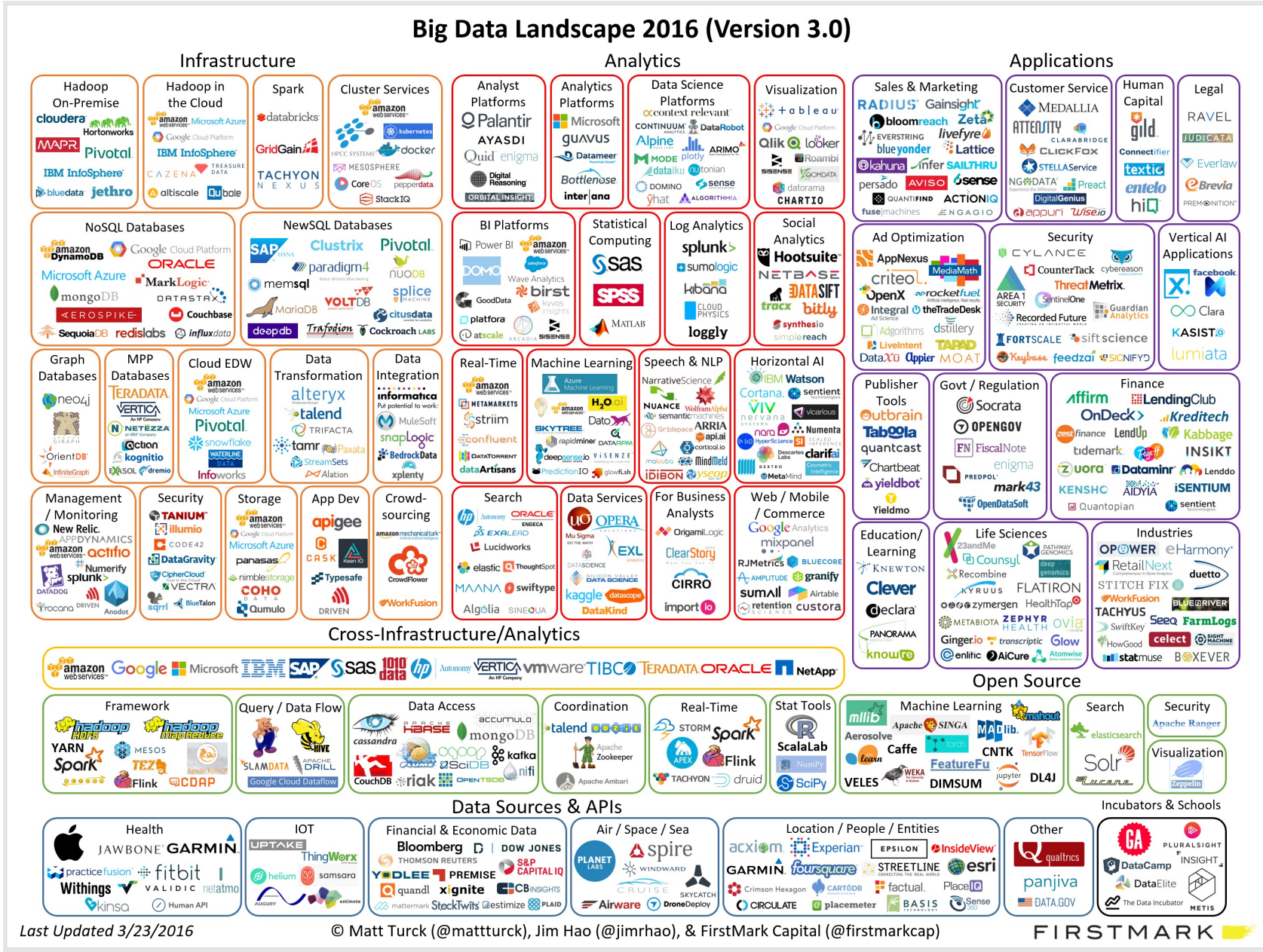
universität koblenz
weiter:denken



Source: https://www.youtube.com/watch?v=58KUVIn7KoE

Big Data Landscape in 2016

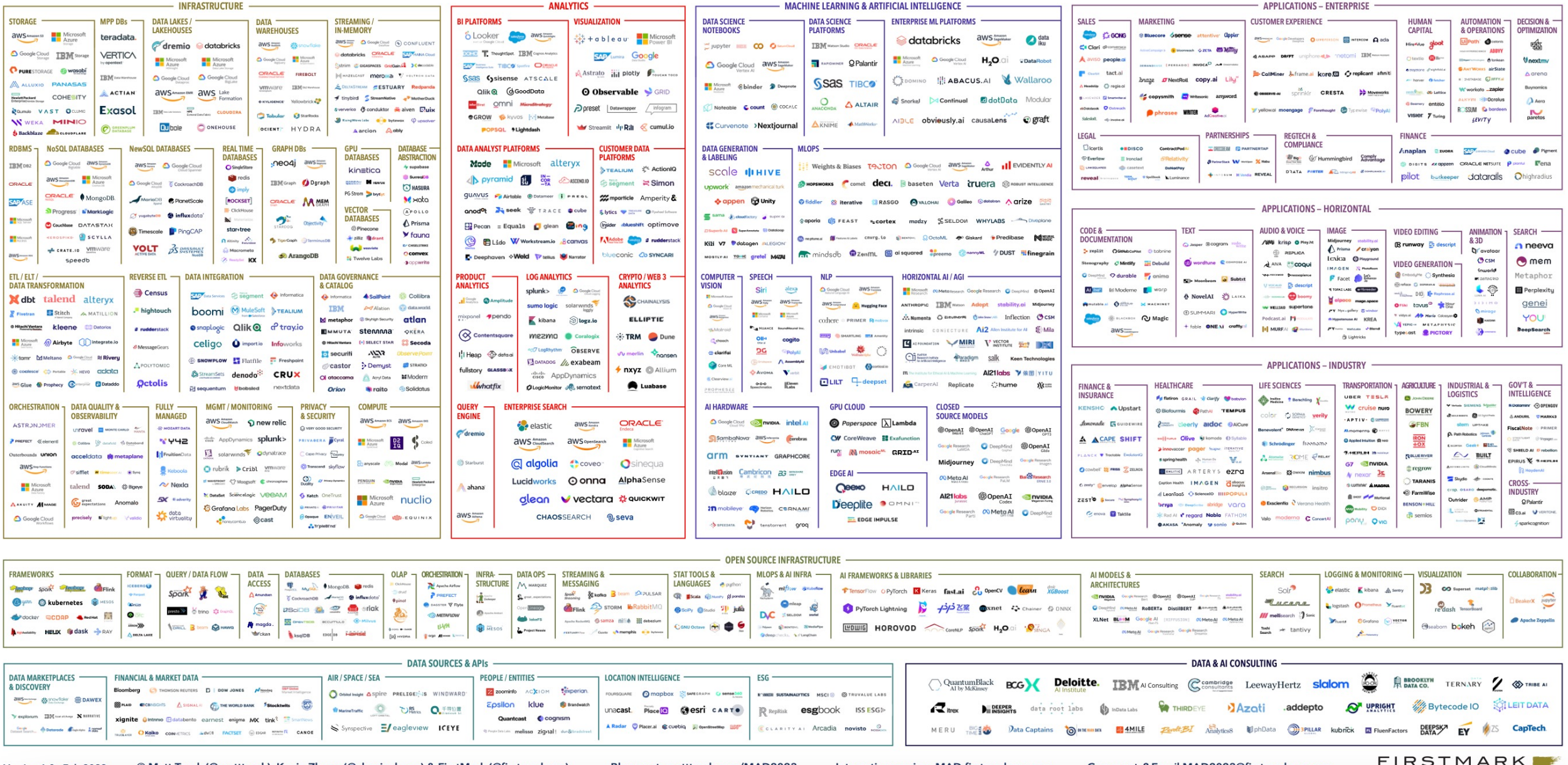
universität koblenz



Big Data Landscape in 2023

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE

universität



Version 1.0 - Feb 2023

Blog post: mattturck.com/MAD2023

Interactive version: MAD.firstmarkcap.com

Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

At its core: Hadoop

universität koblenz
weiter:denken

- Open source framework for reliable, scalable, distributed computing
- http://hadoop.apache.org/

- Developed by Apache Software Foundation
  - Decentralised open source community of developers
  - Supports collaborative, consensus-based development process

Hadoop core modules

universität koblenz
weiter:denken

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects

universität koblenz
weiter:denken

- **HBase:** A scalable, distributed database that supports structured data storage for large tables.
- **Hive:** A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Pig:** A high-level data-flow language and execution framework for parallel computation.
- **Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **ZooKeeper:** A high-performance coordination service for distributed applications.
- **And others**

Operating Hadoop

universität koblenz
weiter:denken

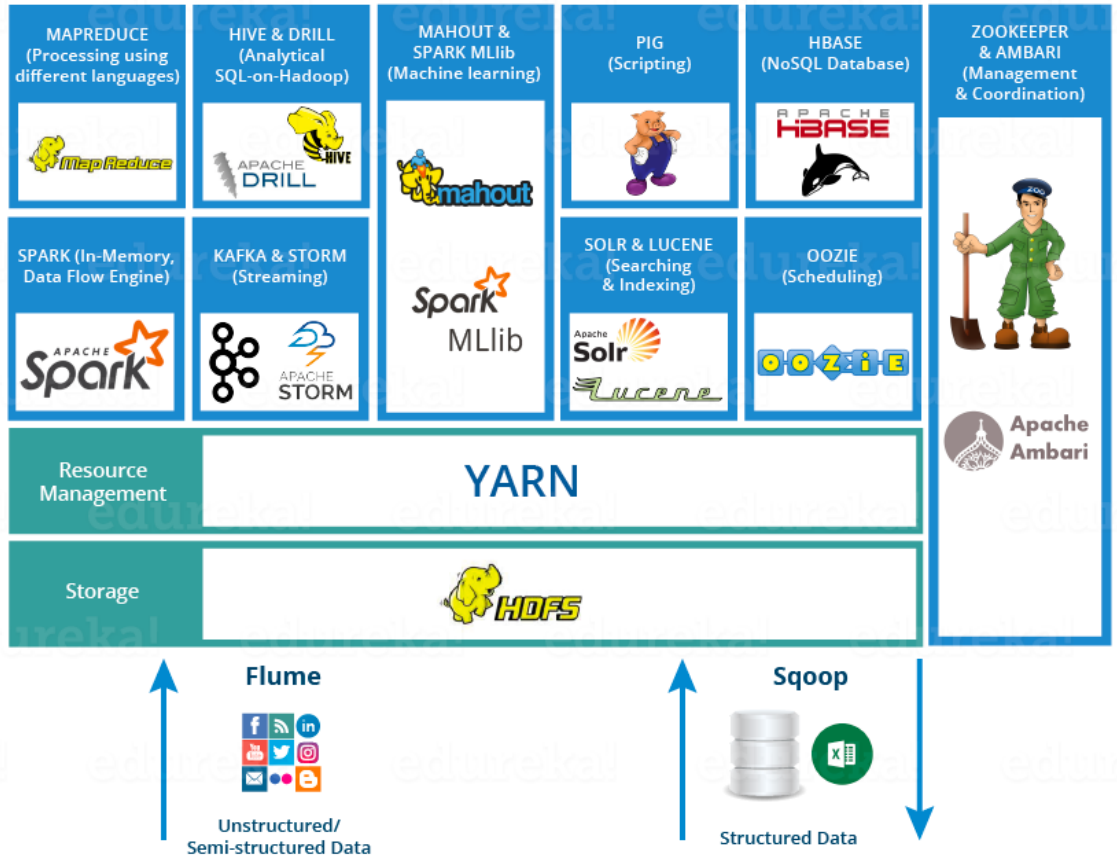
- In on-site data centres (e.g., FB4's own cluster)
- On the cloud
  - Microsoft Azure
  - Amazon EC2/S3 services
  - Amazon Elastic Map Reduce
  - Google Cloud Platform
  - Oracle Cloud Platform
  - …

Who uses Hadoop?

universität koblenz
weiter:denken

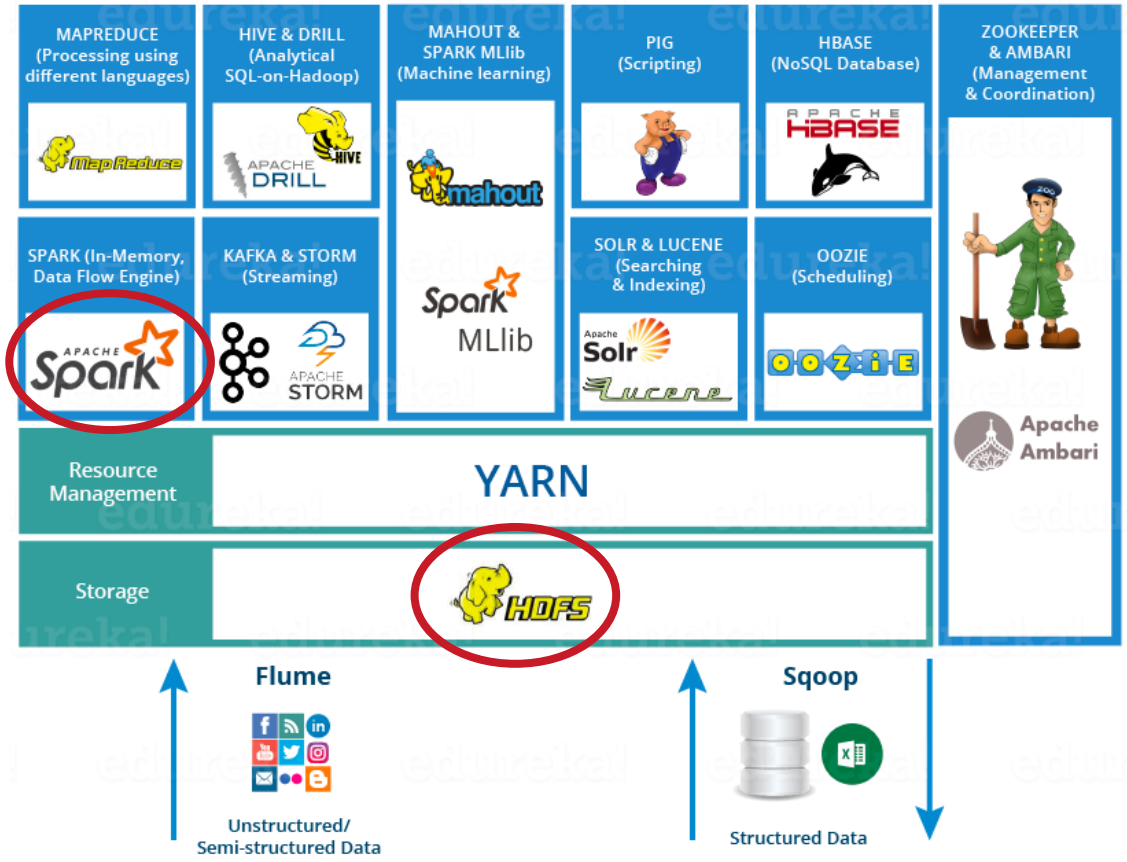
- Amazon/A9
- Facebook
- Google
- New York Times
- Veoh
- Yahoo!
- …. many more

The Hadoop ecosystem



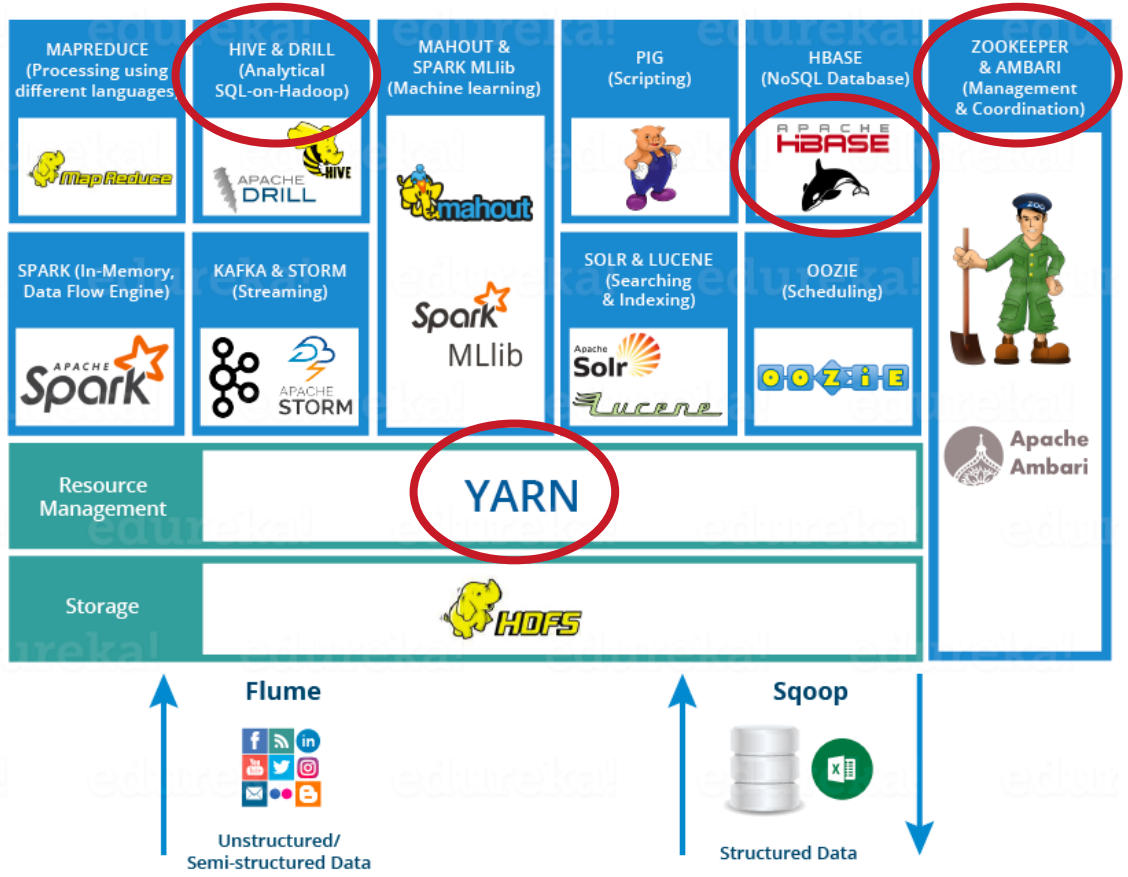
Session 2: Storage Infrastructures

universität koblenz
weiter:denken



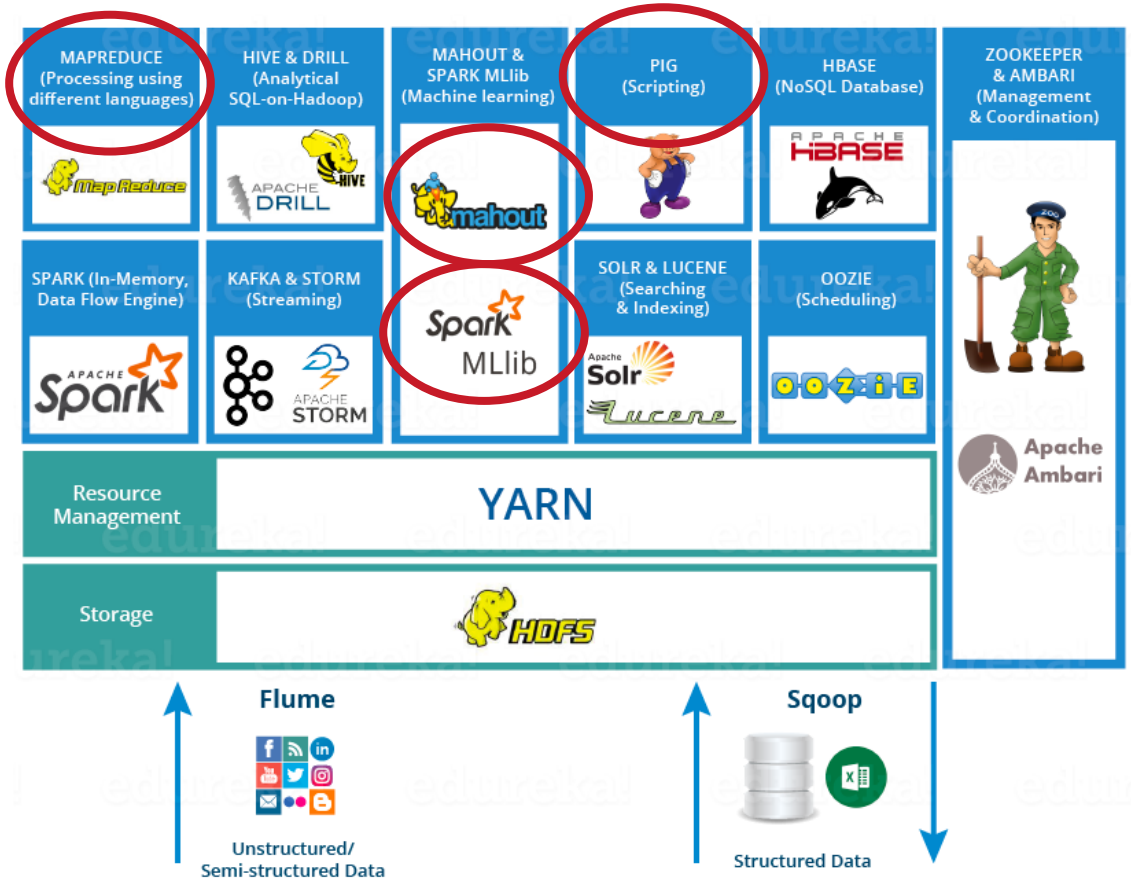
Session 3: Column Stores and Coordination

universität koblenz
weiter:denken



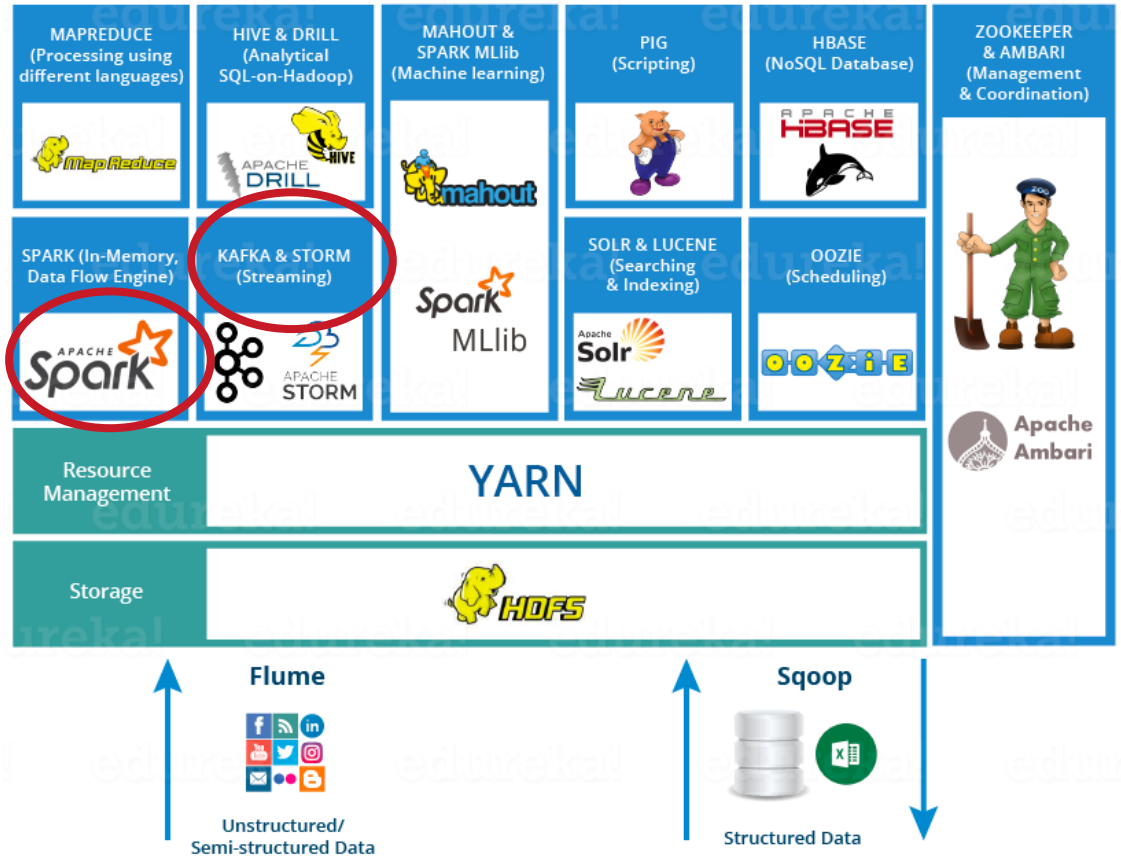
Session 4: Distributed Data Processing

universität koblenz
weiter:denken



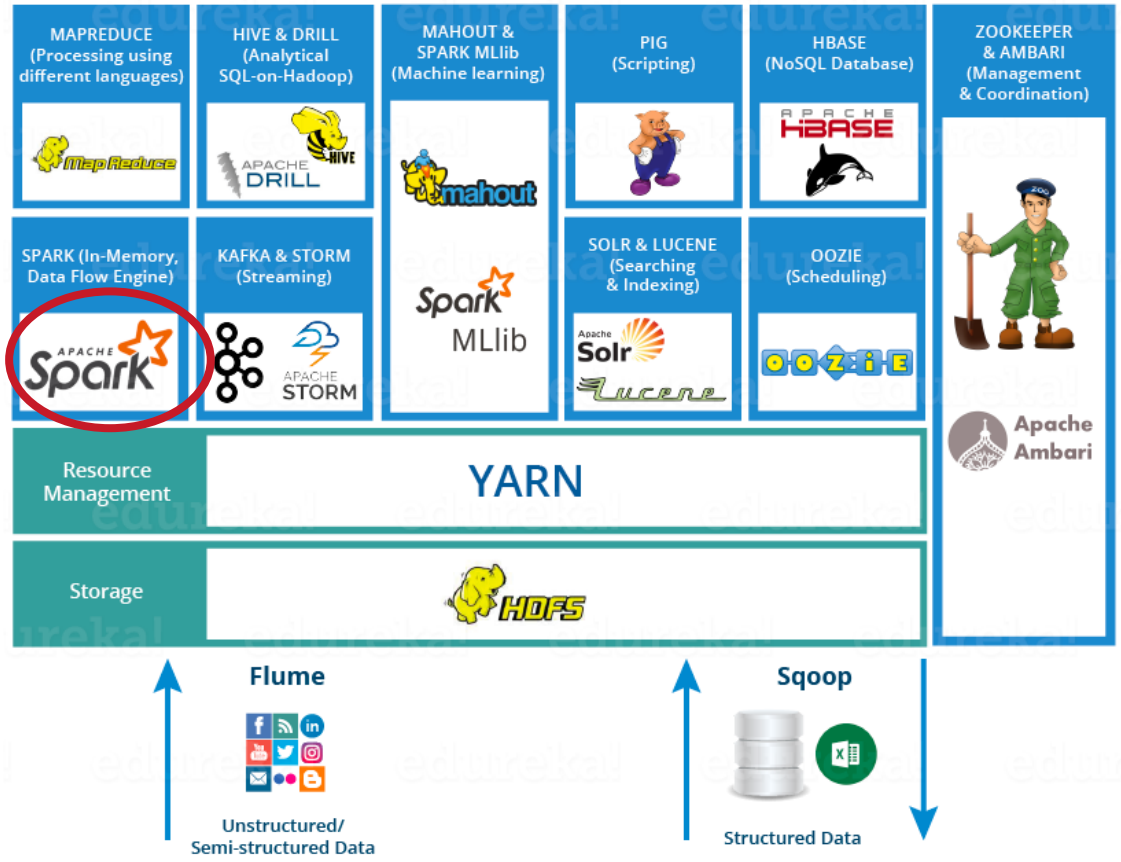
Session 5: Processing Large Data Streams

universität koblenz
weiter:denken



Session 6: Processing Large Graph Data

universität koblenz
weiter:denken



Session 7: Link Analysis at Scale
Session 8: Recommender Systems

universität koblenz
weiter:denken

- Graph analysis applications
  - Link Analysis algorithms
  - Single-Source Shortest Path
  - PageRank
  - Community Detection

- Recommender systems
  - Collaborative filtering
  - Content-based recommendation

Session 9: Challenges & Opportunities
Session 10: Q&A

universität koblenz
weiter:denken

Challenges & Opportunities

- Keeping track of research challenges
- Case studies