Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics
Introduction

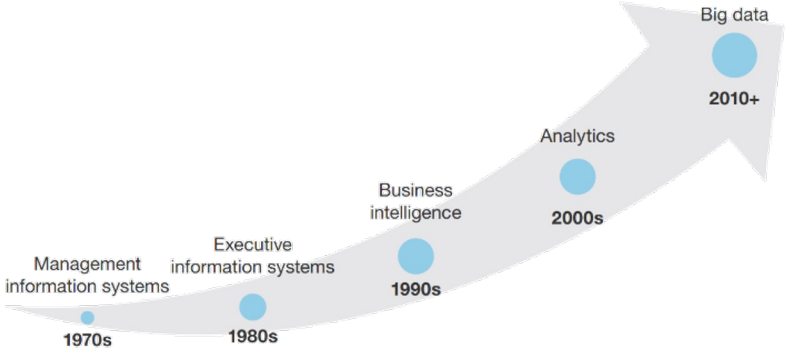
Dresden // 19.04.2023
Sommersemester 2023

Structuring the Field:
Understanding Data Science

A Brief History

From Management Support Systems to Big Data Analytics

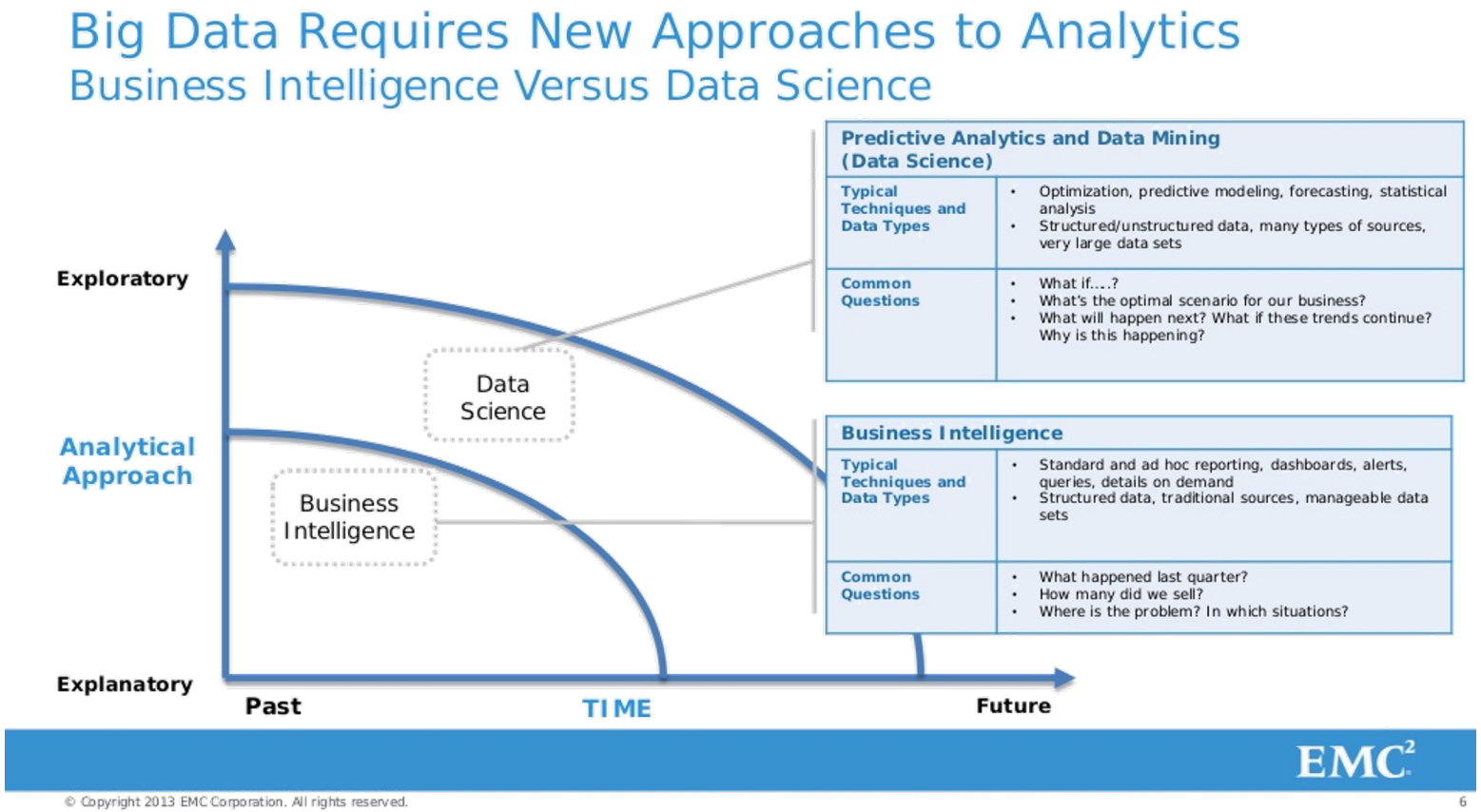
Evolution of Analytical Information Systems



	1970s	1980s	1990s	2000s	2010+
✓ Functions	Operational reporting	Executive dashboards	Operational, tactical, strategic, historical, and predictive reporting, analytics, and data visualization	What-if analysis, includes historical, operational (real time), and predictive	Data democratization, more data to more people with better and faster insights
! Limitations	• Information silos • No integration between data sources • Few strategic and historical reporting capabilities • Mostly printed reports and crosstabs	• Information silos • No integration between data sources • Few strategic and historical reporting capabilities • Mostly static data visualizations	• Limited predictive capabilities • Mostly standalone applications • Complex and inflexible • Based on RDBMS, SQL	• Mostly standalone applications • Mostly used by data scientists	• Mostly standalone applications • Mostly used by data scientists

(Source: Thomas, 2017)

Data Science vs. Business Intelligence



(Source: EMC Corporation, 2013)

A Definition of Data Science

by Provost & Fawcett (2013)

> *At a high level, **data science** is a **set of fundamental principles** that support and guide the principled extraction of **information** and **knowledge** from **data**. Possibly the most closely related concept to data science is **data mining** - the actual extraction of knowledge from data via **technologies** that incorporate these principles. There are hundreds of different **data mining algorithms**, and a great deal of detail to the methods of the field. We argue that underlying all these many details is a much smaller and more concise set of fundamental principles.*

> *Data science involves **principles, processes, and techniques** for understanding phenomena via the (automated) analysis of data. (…) the ultimate goal of data science is improving **decision making**, as this generally is of paramount interest to business. (…) Data-driven decision making refers to the practice of basing decisions on the analysis of data rather than purely on intuition.*

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

To understand Data Science, it requires to understand the 1) principles, 2) concepts, 3) techniques, 4) technologies and 5) processes behind it!

Structuring the Field:
1) Data Science Principles

Some Examples of Data Science Principles

by Provost & Fawcett (2013)

> *"Extracting useful knowledge from data to solve **business problems** can be treated **systematically** by following **a process** with reasonably **well-defined stages**."*

1. **Structured approach towards analytical problem**
2. **Careful analysis of the problem**

Examples of common process/ procedure models:

Knowledge Discovery in Databases (KDD)

Sample, Explore, Modify, Model, Asess (SEMMA)

More details later…

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

Some Examples of Data Science Principles

by Provost & Fawcett (2013)

> *"**Evaluating** data-science results requires careful **consideration of the context** in which they will be used."*

Critical consideration on the application in question

Example questions:

Does the pattern lead to better decisions than **some reasonable alternative**?

How well would one have done **by chance**?

How well would one do with a **smart "default" alternative**?

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

Some Examples of Data Science Principles

by Provost & Fawcett (2013)

> *"The relationship between the **business problem** and the **analytics solution** often can be decomposed into **tractable sub-problems** via the framework of analyzing **expected value**."*

Breaking down the business problem into components

Reduce complexity/ increase transparency

> *"**Information technology** can be used to find **informative data items** from within a large body of data."*

Usage of information technology to compute standard measures to quickly identify objects of interest

For example: Computation of correlation measures between hundreds of variables

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

Some Examples of Data Science Principles

by Provost & Fawcett (2013)

> *"Entities that are **similar with respect to known features** or attributes often are **similar with respect to unknown features** or attributes."*

Similarity is one of the most important measures in Data Science

Many ways and specifications to quantify similarity

> *"If you look too hard at a set of data, you will find something - but it might **not generalize beyond the data** you're observing."*

Problem of representation, also called „overfitting"

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

Some Examples of Data Science Principles

by Provost & Fawcett (2013)

> *"To draw **causal conclusions**, one must pay very close attention to the presence of **confounding factors**, possibly unseen ones."*

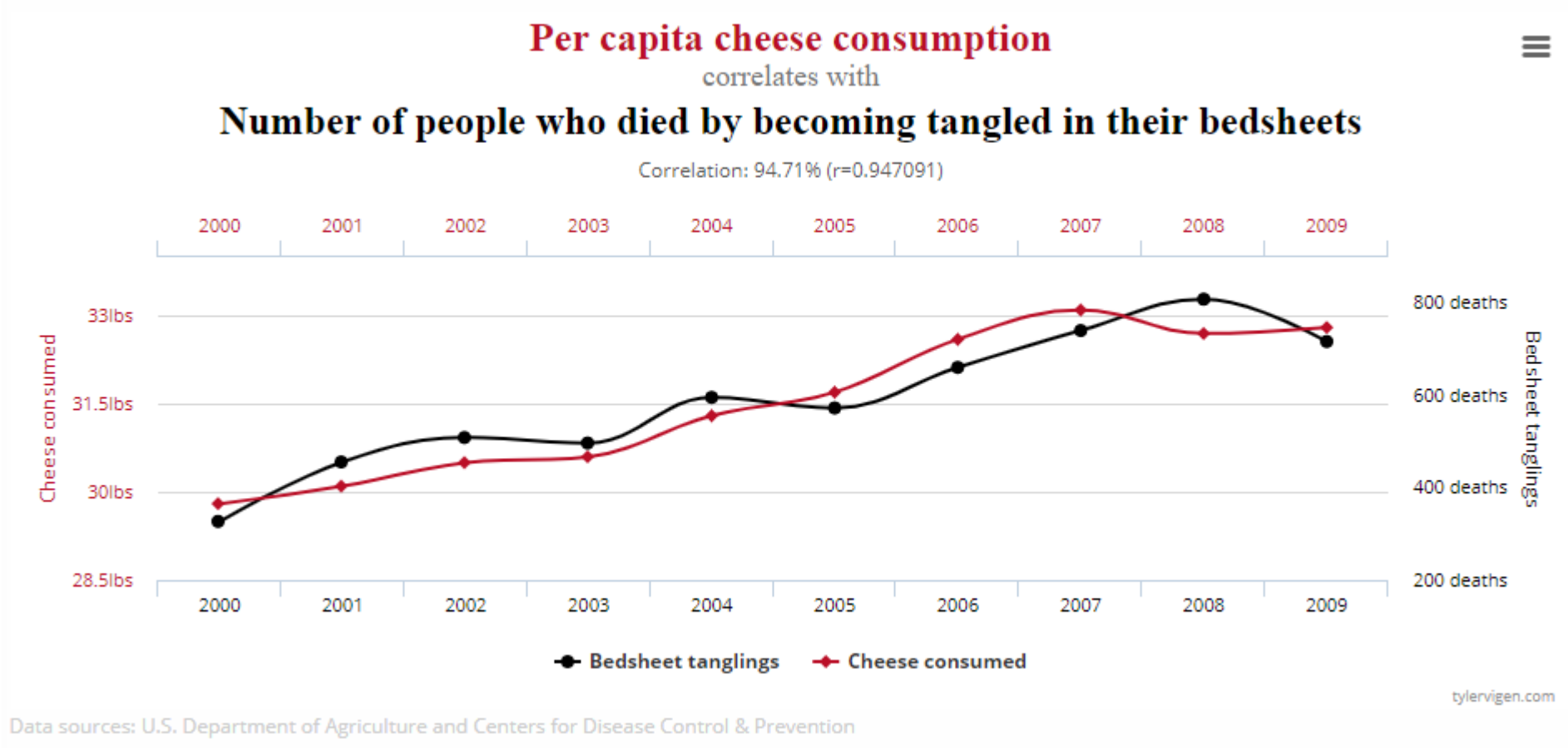
Correlation does not imply causation!

It requires clear understanding of underlying data analysis techniques and their assumptions

What are the conclusions of the analysis and how can they be used for decision making?

(Source: Provost & Fawcett, 2013: Data Science and its Relationship to Big Data and Data-Driven Decision Making)

Example: Spurious Correlations



Structuring the Field:
2) Data Science Concepts

Buzzword Bingo or Solid Concepts?

Many buzzwords in this area!

Often driven by marketing purposes

We will look at a selection of buzzwords:
A) Big Data, B) Data Mining, C) Analytics, D) Machine Learning, E) Statistic, F) Visualization

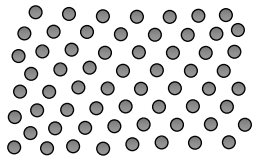
A) Big Data

Giving the phenomenon a name

Big Data

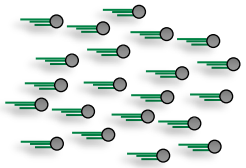
The 7 Vs

Volume



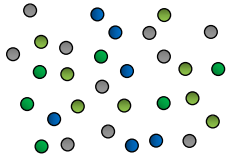
Scale of Data

Velocity



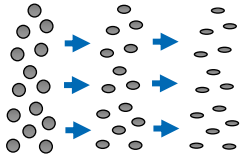
Streaming Data

Variety



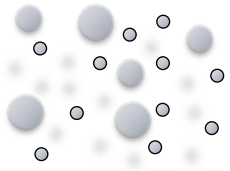
Different Forms of Data

Variability



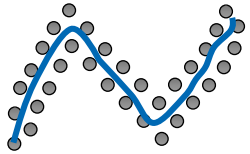
Change of Data

Veracity



Uncertainty of Data

Visualization



Representation of Data

Value

Value of Data

Traditional Data vs. Big Data

A Comparison by Ramannavar & Sidnal (2016)

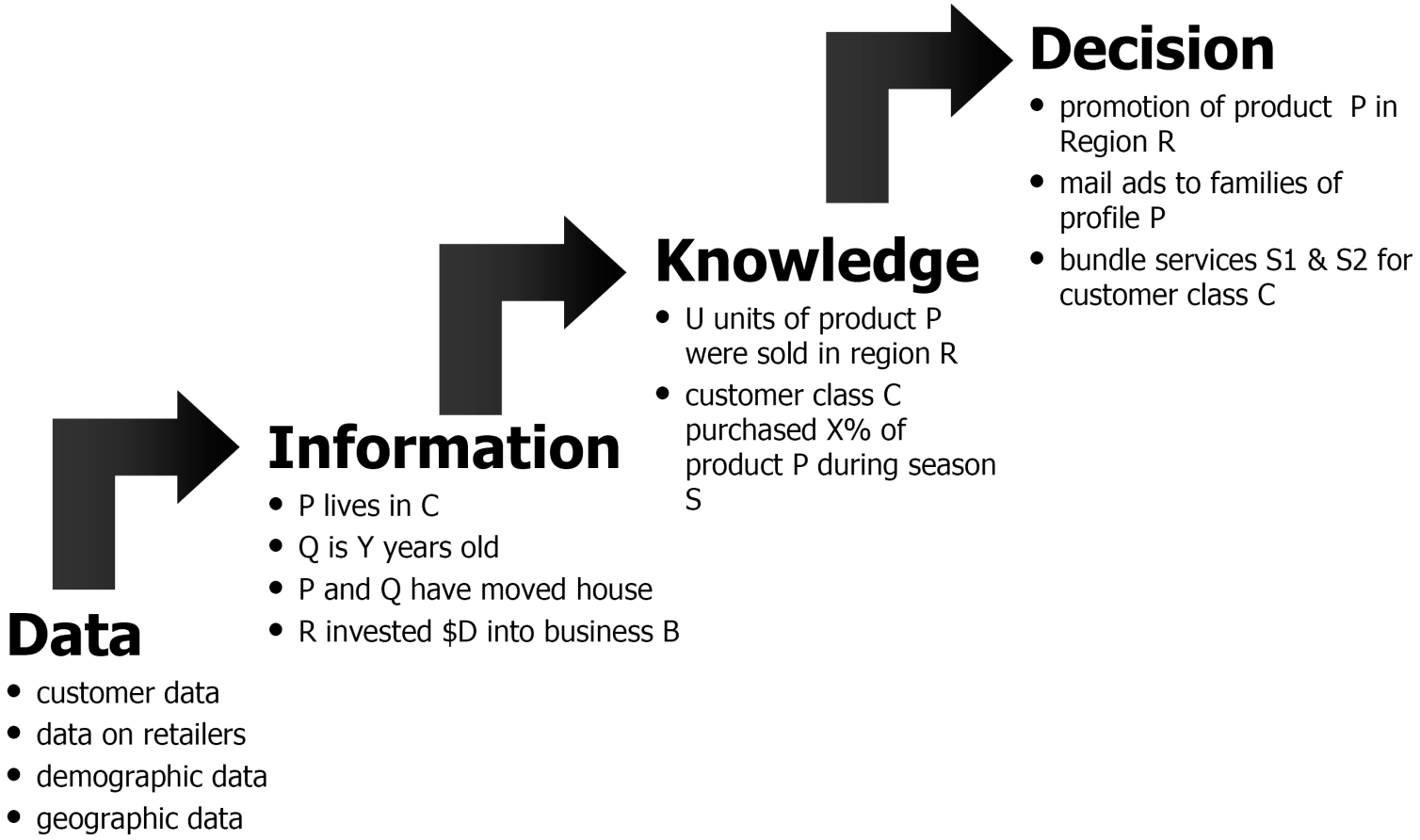
Characteristic	Traditional data	Big data
Volume	Gigabytes to a few terabytes	Hundreds of terabytes, petabytes, exabytes, and beyond
Nature and location	Typically limited to an organization or an enterprise and intrinsically centralized	Distributed across multiple entities; spread throughout the electronic space such as Internet servers and cloud servers
Structure and content	Highly structured and has a single discipline or may be a sub discipline. This is often in uniform records, such as an ordered spreadsheet	Semi-structured or highly unstructured (free form text, images, motion pictures, sound recordings, etc.). The subject matter may cross multiple disciplines that are not related
Data models	Stable data models	Flat schemas
Goal	Designed to answer very specific questions or serve a particular goal	Designed with a flexible goal to answer protean questions. The goal is vague but is obvious that there really is no way to completely specify what big data resource will contain, how data types are organized, gets connected to other resources, etc.
Data preparation	Data user prepares her own data for her purposes in case of traditional data	Big data comes from many diverse sources and prepared by many people
Longevity	Traditional data projects have a well established time frame and end in finite time	Big data projects typically contain data that must be stored perpetuity. Ideally, data stored in big source will be absorbed into another resource when original resource terminates. Hence, big data projects extend into future and past, accruing data prospectively and retrospectively

Traditional Data vs. Big Data

A Comparison by Ramannavar & Sidnal (2016)

Characteristic	Traditional data	Big data
Longevity	Traditional data projects have a well established time frame and end in finite time	Big data projects typically contain data that must be stored perpetuity. Ideally, data stored in big source will be absorbed into another resource when original resource terminates. Hence, big data projects extend into future and past, accruing data prospectively and retrospectively
Stakes	Being limited to a lab or an institution, can usually recover from occasional small data failures	Being varied in scope, can be irrevocably expensive to recover. Failure may lead to huge losses. Example: the failure of NCI cancer biomedical informatics grid caused $350 million loss for fiscal years 2004–2010
Measurements	Measured using one experimental protocol and the data can be represented using one set of standard units	Delivered in many different electronic formats. Measurements when present may be obtained by many protocols. Verifying quality of big data is a major challenge
Reproducibility	Traditional data projects can be typically repeated	Replication of big data is seldom feasible
Introspection	Individual data points (row and column) can be easily identified	Unless exceptionally organized, it is impossible to locate a data entity
Analysis	Traditional project itself needs to be analyzed along with data	Big data is analyzed orderly in incremental steps. Data can be extracted, reviewed, reduced, normalized, transformed, visualized, interpreted, and reanalyzed with different methods
Tools and technologies	RDBMS, row-oriented databases, ETL tools, SQL	HDFS, Column-oriented databases, NoSQL, Hadoop, Storm, Spark

Information Value Chain



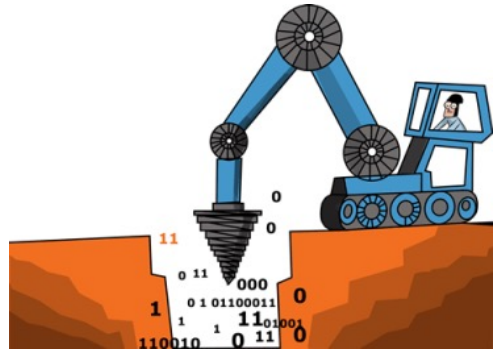
B) Data Mining

The search for the digital treasures

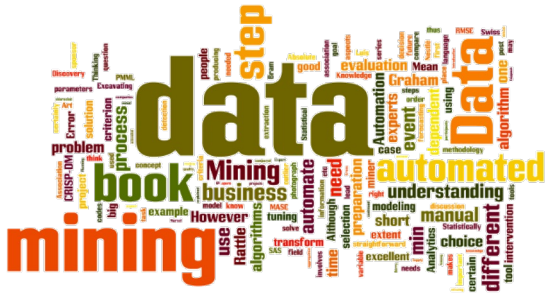
Data Mining – A metaphor from the mining industry

Mining: The industry or activity of removing substances, such as coal, metal, gold or diamonds, from the ground by digging.

https://dictionary.cambridge.org/de/worterbuch/englisch/mining



http://djitz.com



http://pinterest.com

Data Mining: the process of using special software to look at large amounts of computer data in order to find out useful information, for example what types of product a company's customers buy.

https://dictionary.cambridge.org/de/worterbuch/englisch/data-mining

Data Mining

Definition and context

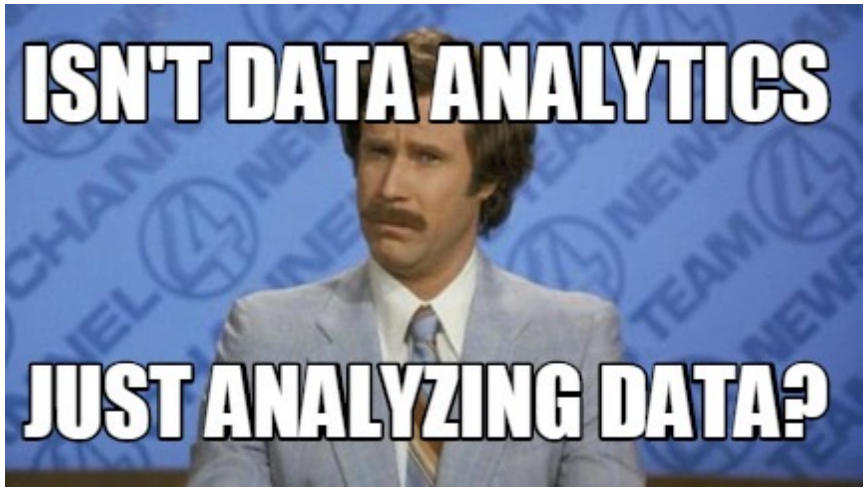
„Data Mining is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions.“
(Cabena et al. 1997)

„Data Mining, as we use the term, is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.“

(Berry et al. 1997)

C) Analytics

Is it just about analyzing data?

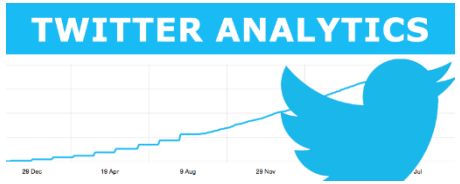


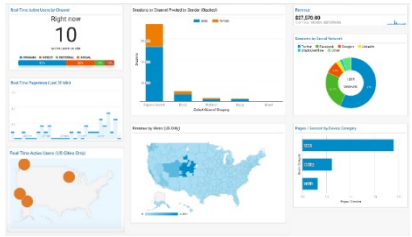
http://www.digitalvidya.com

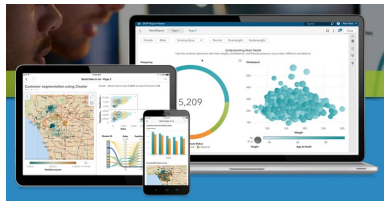
Analytics in Different Business Domains

Examples of application domains:

Marketing Analytics
Finance Analytics
Sales Force<< Analytics
HR/ People Analytics
Supply Chain Analytics
Fraud Analytics
Web Analytics
Social Media Analytics
Energy Analytics
Maintenance Analytics
Learning Analytics
...







Different Types of Analytics

Taking responsibilities for dedicated analytical tasks

Descriptive Analytics

Question: "*What is happening right now/ what happened in the past?*"

Summarizing collected data from various sources and providing **aggregated measures** and **visualizations**

Including traditional **business intelligence** tools: e.g. reports, dashboards, querying, online analytical processing (OLAP)

Diagnostic Analytics

Question: "*Why did/ does it happen?*"

Explanatory empirical modeling based on statistical inference and causal hypotheses testing

Predictive Analytics

Question: "*What is likely to happen?*"

Developing **prediction models** with high accuracy by using a large amount of historical data, detecting **complex** and **non-trivial relationships** and providing predictions and equivalent probability scores on new unclassified observations

Prescriptive Analytics

Questions: "*What should be done / why should it be done?*"

Mathematical techniques (e.g. optimization, simulation) that computationally determine a set of **high-value alternative actions** or decisions given a complex set of objectives, requirements and constraints

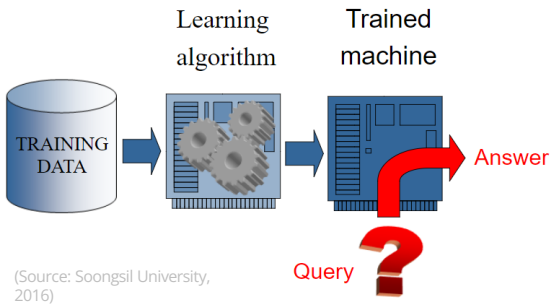
D) Machine Learning

Machine Learning

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." (Samuel, 1959)

"A subspecialty of computer science (within a field historically called "artificial intelligence") concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data." (Manyika et al., 2011)

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions through building a model from sample inputs. (Bishop, 2006)



(Source: Soongsil University, 2016)

Machine Learning

A formal problem definition proposed by Mitchell (1997)

„A computer program is said to learn from **experience *E*** with respect to some class of **tasks *T*** and **performance measure *P*** if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

Goal: Generalize the experience in a way that allows to improve your performance on the task.

Given:

a task T,

a performance measure P,

some experience E with the task

Examples:

A game learning problem:

- **Task T**: playing Backgammon
- **Performance measure P**: percent of games won against opponent
- **Training experience E**: playing practice games against itself

A handwriting recognition learning problem:

- **Task T**: recognizing and classifying handwritten words within images
- **Performance measure P**: percent of words correctly classified
- **Training experience E**: a database of handwritten words with given classifications

A robot driving learning problem:

- **Task T**: driving on public four-lane highways using vision sensors
- **Performance measure P**: average distance traveled before an error (as judged by human overseer)
- **Training experience E**: a sequence of images and steering commands recorded while observing a human driver

Machine Learning Categorization

Different types of learning problems

Machine Learning tasks can be classified into three broad categories of learning problems:

Supervised Learning:

The algorithm is given training data which contains examples of the **input vectors** along with their corresponding **target vectors** as the "correct answer" and the goal is to learn a general rule/ model that maps inputs to outputs.

Examples: classification, regression

Unsupervised Learning:

Principle: No labels (i.e. target values) are given to the algorithm, leaving it on its own to **find structure in its input**. Unsupervised learning can be a goal by itself (discovering hidden patterns) or a means towards a supervised approach (feature learning).

Examples: cluster analysis, dimensionality reduction, density estimation

Reinforcement Learning:

"(...) problem of finding suitable actions to take in a given situation in order to **maximize a reward**. Here the learning algorithm is not given examples of optimal outputs, in contrast to supervised learning, but must instead discover them by a **process of trial and error**. Typically there is a sequence of states and actions in which the learning algorithm is interacting with its environment." (Bishop, 2006)

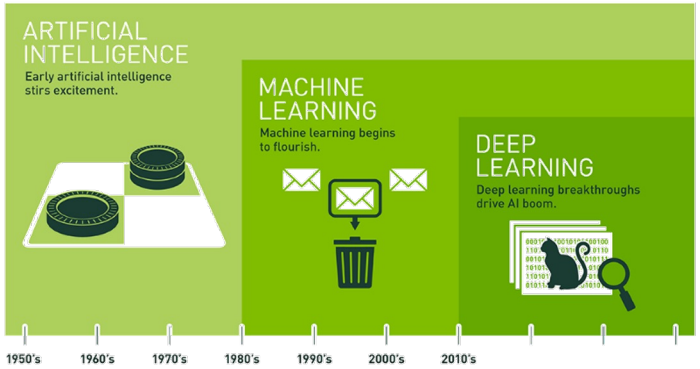
Machine Learning & Deep Learning

Deep Learning as a new disruptive technology

A special class of Machine Learning algorithms is Deep Learning.

“The main concept in deep leaning algorithms is automating the extraction of representations (abstractions) from the data. Deep learning algorithms use a huge amount of unsupervised data to automatically extract complex representation.” (Najafabadi et al., 2016)

“These algorithms are largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain’s ability to observe, analyze, learn, and make decisions, especially for extremely complex problems.” (Najafabadi et al., 2016)



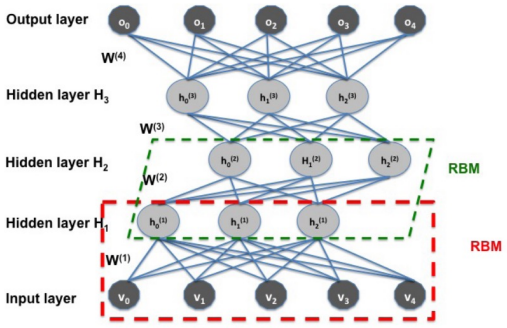
https://blogs.nvidia.com

Deep Learning

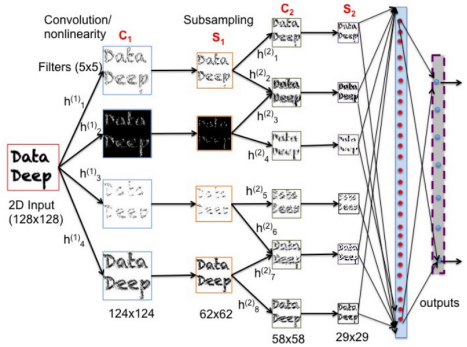
Deep neural network architectures

"Deep learning is a relatively new term, although it has existed prior to the recent dramatic uptick in online searches. Enjoying a surge in research and industry, due mainly to its incredible successes in a number of different areas, deep learning is the process of applying deep neural network technologies - that is, neural network architectures with multiple hidden layers of neurons - to solve problems."

(KDnuggets, 2016)



(Source: Chen & Lin, 2014)



(Source: Chen & Lin, 2014)

E) Statistics

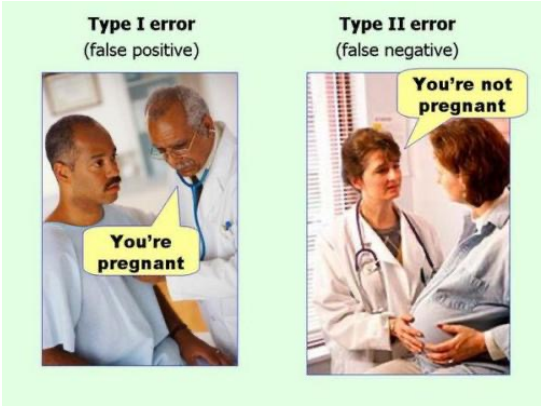
http://www.truthfacts.com

Statistics

Definition

"Statistics:

- **The science of the collection, organization, and interpretation of data, including the design of surveys and experiments.**
- **Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the "null hypothesis"), and what relationships between variables likely result from some kind of underlying causal relationship (i.e., that are "statistically significant").**
- **Statistical techniques are also used to reduce the likelihood of Type I errors ("false positives") and Type II errors ("false negatives")."**
  (Manyika et al., 2011)



Statistics

Descriptive or Inductive

Statistics = teaching in dealing with quantitative information (= data)

- **Descriptive or Inductive**

Descriptive Statistics

Task: Data should be prepared and graphically displayed.

- **Concentration of the underlying extensive database on a few meaningful measures.**
- **In extreme cases, it is possible that a single number characterizes the entire data set**

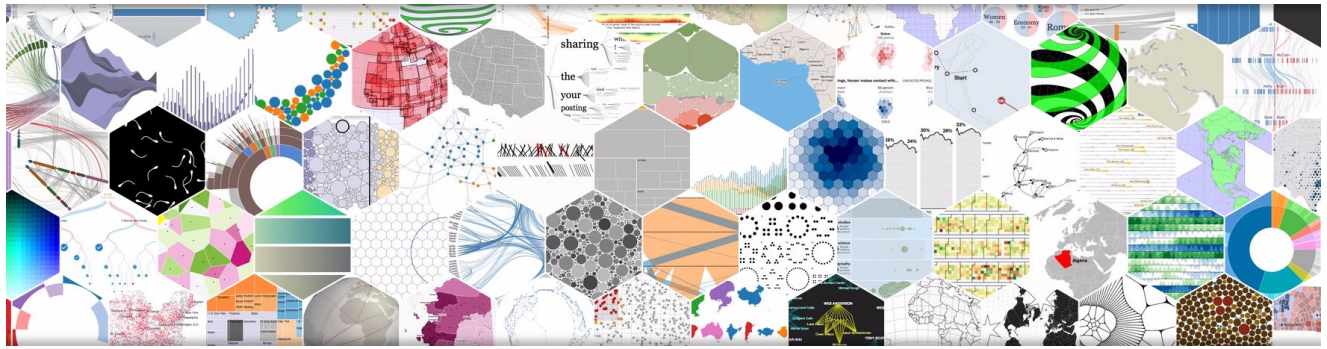
Inductive statistics

Summary of all methods needed for rational decision making under uncertainty or risk.

Task: The properties of the population are derived from the data of the sample.

- **Basis of inductive statistics: probability theory**
- **Objective: The statistical methods of inductive statistics serve to make predictions (e.g., about risks).**

F) Visualization



https://d3js.org/

Importance of Visualization

Why is data visualization so important?

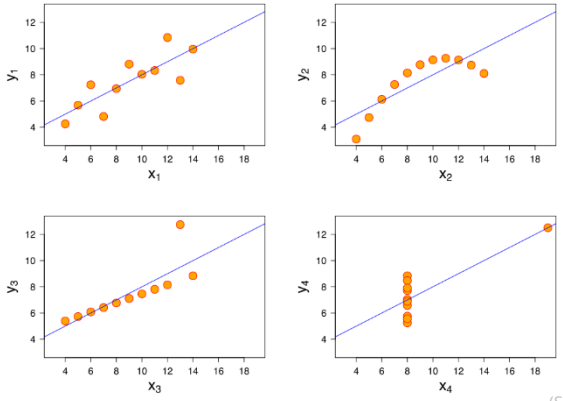
“Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.” (SAS, 2017)

Another Example: Anscombe’s Quartet

4 data sets of 11 elements

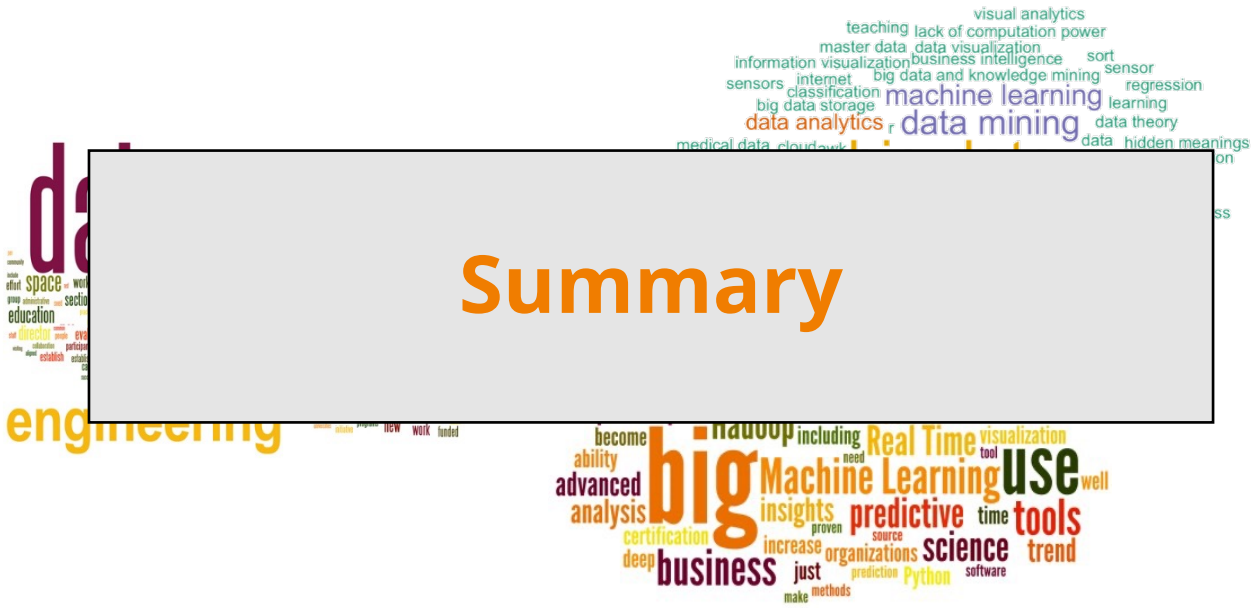
same descriptive statistics for all 4 data sets

- mean of x = 9
- variance of x = 11
- mean of y = 7.50
- variance of y = 4.125
- correlation = 0.816
- linear regression line: y = 3.00 + 0.500x



(Source: Anscombe, 1973)

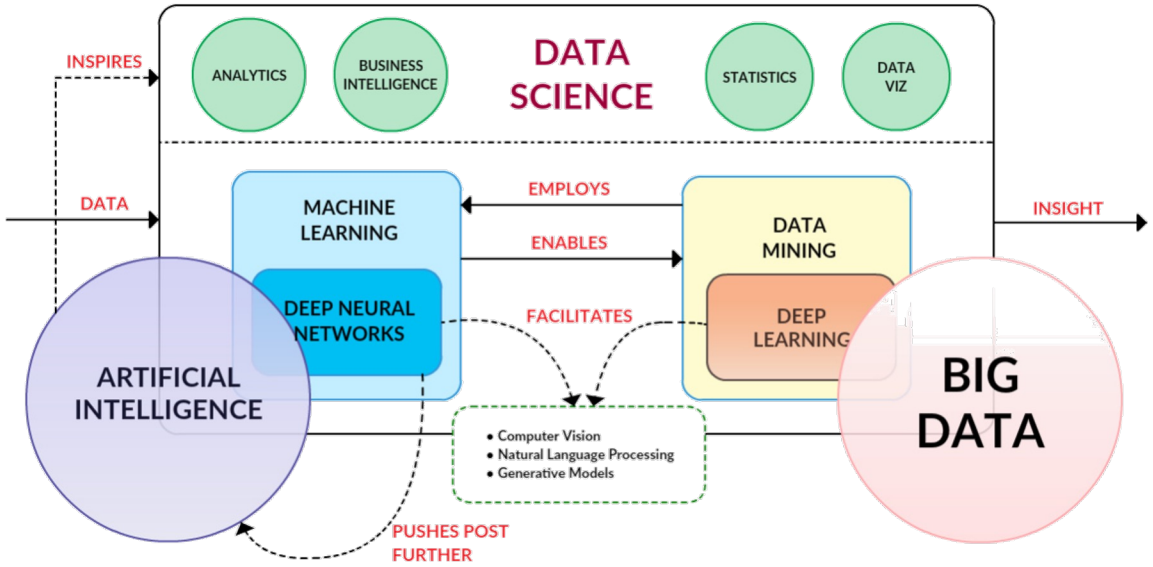
Buzzword Bingo or Solid Concepts?



Summary

Bringing It Together

Alignment of introduced concepts



(Source: KDnuggets, 2017)
Further reading: http://www.kdnuggets.com/2017/01/data-science-puzzle-revisited.html?utm_content=bufferf7b23&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer

Structuring the Field:
3) Data Science Techniques

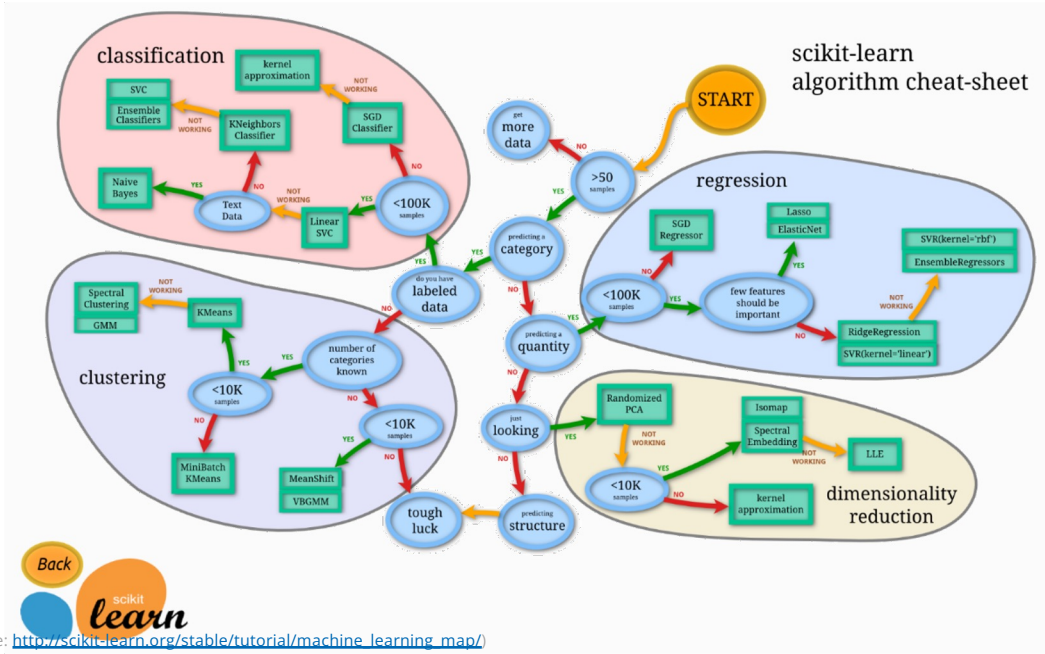
Variety of Data Analysis Techniques

Which one to choose?

How do we know which technique/algorithm to choose?

Algorithm Cheat Sheet

proposed by Python scikit-learn



(Source: http://scikit-learn.org/stable/tutorial/machine_learning_map/)

Variety of Data Analysis Techniques

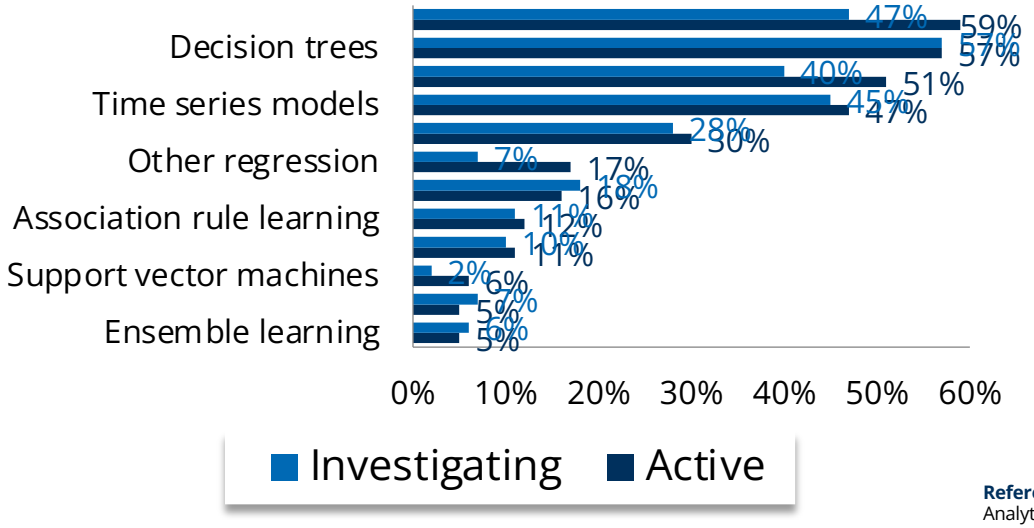
What is best practice?

What are the most commonly used techniques/algorithms?

Top Techniques

identified by TDWI Research

"What are the most popular techniques for predictive analytics in your organization?"



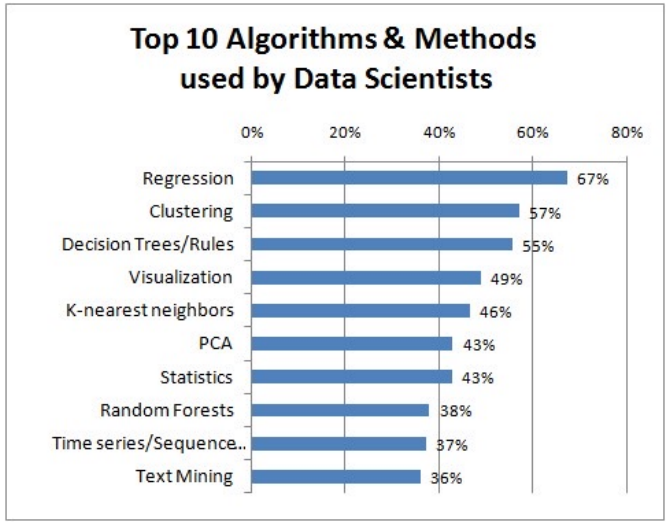
Reference: Fern Halper: Predictive Analytics for Business Advantage. TDWI Best Practice Report. TDWI Research, First Quarter, 2014.

Top Techniques

identified by KDnuggets (2011 vs. 2016)

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Source: http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html



Source: http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html

Structuring the Field:
4) Data Science Technologies

Data Science Technologies

Our focus on selected technologies

Advanced Enterprise Solutions

SAS

Microsoft

Code-based Open Source Solutions

python

R

SAS Enterprise Guide (for Business Intelligence)

User interfaces and wizards

Reporting, graphics and analysis tasks

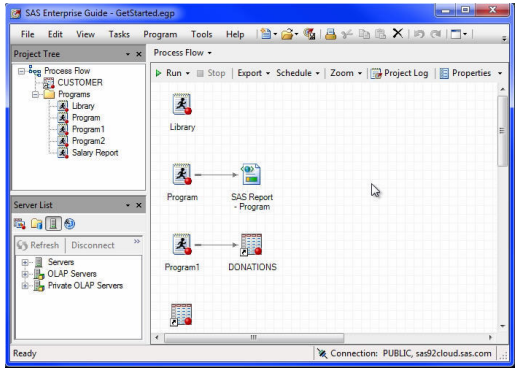
Data management

OLAP access, visualization and manipulation

Distribution of results

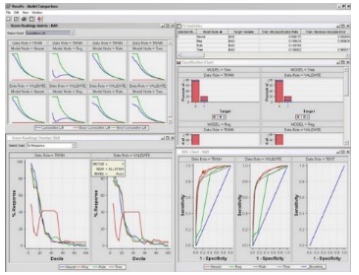
Administration and security

Support for high-performance computing in the grid



SAS Enterprise Miner (for Analytics)

- SEMMA-based modeling: sample, explore, manipulate, model, assess
- Advanced predictive modeling algorithms including decision trees, neural networks, logistic regression, etc.
- Descriptive functionalities including statistics, plotting, scorecards, etc.



Microsoft

Microsoft Business Intelligence

SQL Server as relational database management system

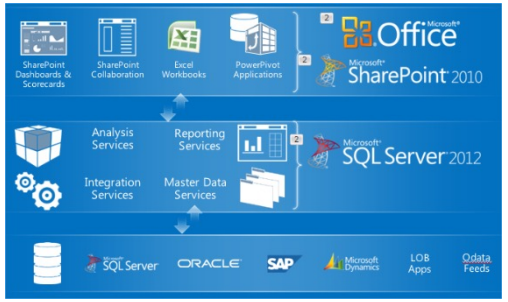
Integration Services for ETL processes

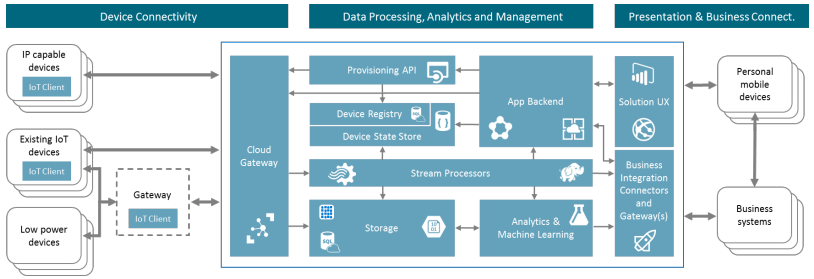
Analysis Services for multidimensional OLAP analysis

Reporting Services for frontend development

SharePoint Server as content management system

PowerBI/ Excel/ PowerPivot for Self-Service BI





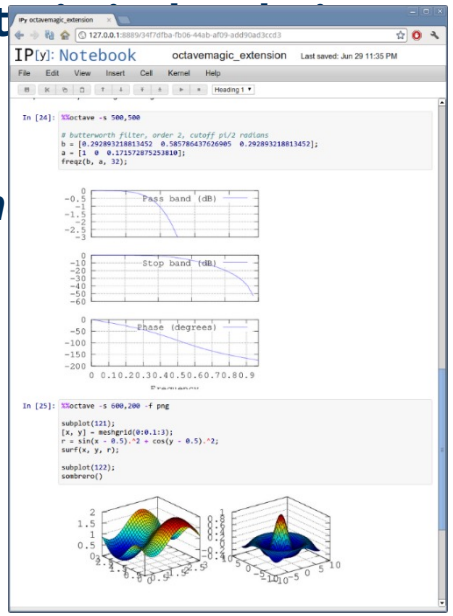
Microsoft Azure IoT Analytics

- **IoT Hub** as cloud conncection gateway
- **Stream Analytics** for stream processing
- **Machine Learning** for analytical tasks
- **Cosmos DB** for storing big data
- **... and more** Azure Cloud Services

Python is an open source programming language with focus on productivity and code readability

Python was inspired by C, Modula 3 and particularly by ABC and created by Guido Van Rossum, released in 1991

Python is used by programmers that want to delve into data analysis or apply st and by developers that turn to data science



"The closer you are to working in an engineering environm
the more you might prefer Python."

Many IDEs for Python (most popular: Spyder, iPythonNotebook)

Main advantage: Python is a full-fledged programming language

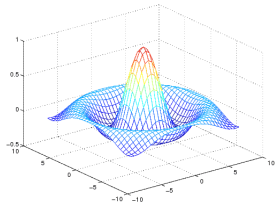
(Source: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis)

R is an open source programming language with focus on better user friendly data analysis, statistics and graphical models

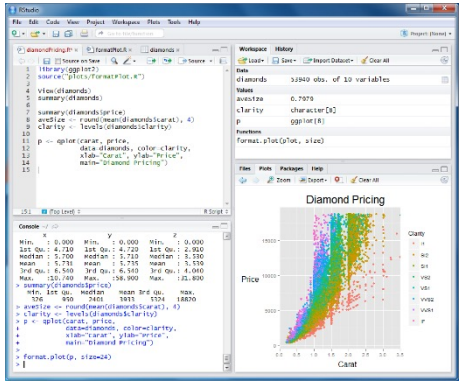
R is an implementation of S programming language

Created by Ross Ihaka and Robert Gentleman Released in 1995

R has been used primarily in academics and research, but it is rapidly expanding into the enterprise market



"The closer you are to statistics, research and data science, the more you might prefer R."



Standard IDE is R Studio

Main advantage: graphical capabilities for data visualization

(Source: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis)

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention