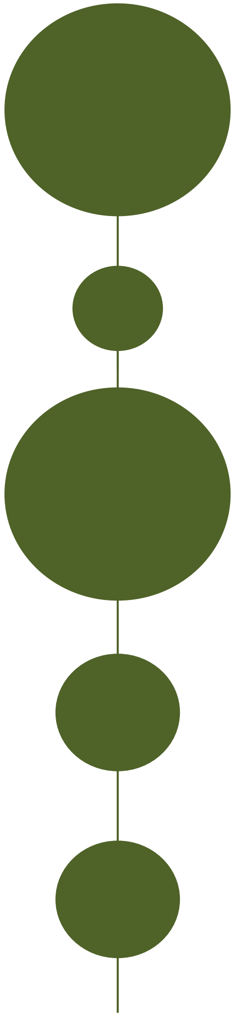
Introduction to Data Science

Lecture 1: Organizational Information and Introduction

Prof. Dr. Hendrik Meth
HdM, Stuttgart

About me



1995-2001

University of Mannheim
Diploma in Business Informatics

2002-2004

Icon, Karlsruhe
SAP APO & SAP BW-Consulting

2004-2010

Bosch, Stuttgart
Business Intelligence Consulting,
Product & Project Management

2010-2013

University of Mannheim
PHD in Business Informatics

2013-2016

BorgWarner ITSE
Manager Business Intelligence
Competence Center

since 03.2016

HdM Stuttgart
Professor
Big Data & Data Science

Teaching Portfolio

Master

BUSINESS INTELLIGENCE (WI3)

- Data Warehousing
- Reporting
- BI Trends
- Workshop

DATA SCIENCE (WI3)

- Foundations
- Processes
- Methods & Algorithms
- Workshop

INTRODUCTION TO DATA SCIENCE (DSM)

- Processes
- Methods
- Algorithms

Bachelor

BIG DATA SCENARIOS

- Methods and technologies for unstructured data analysis
- Natural Language Processing

PROJECTS, e.g. “DEEP LEARNING”, “SPORTS ANALYTICS”

- Neural Networks
- Pose Estimation
- …

ADVANCED DATA SCIENCE

- Data Preprocessing
- Advanced Methods
- Analytical Project

MATHEMATICS & STATISTICS

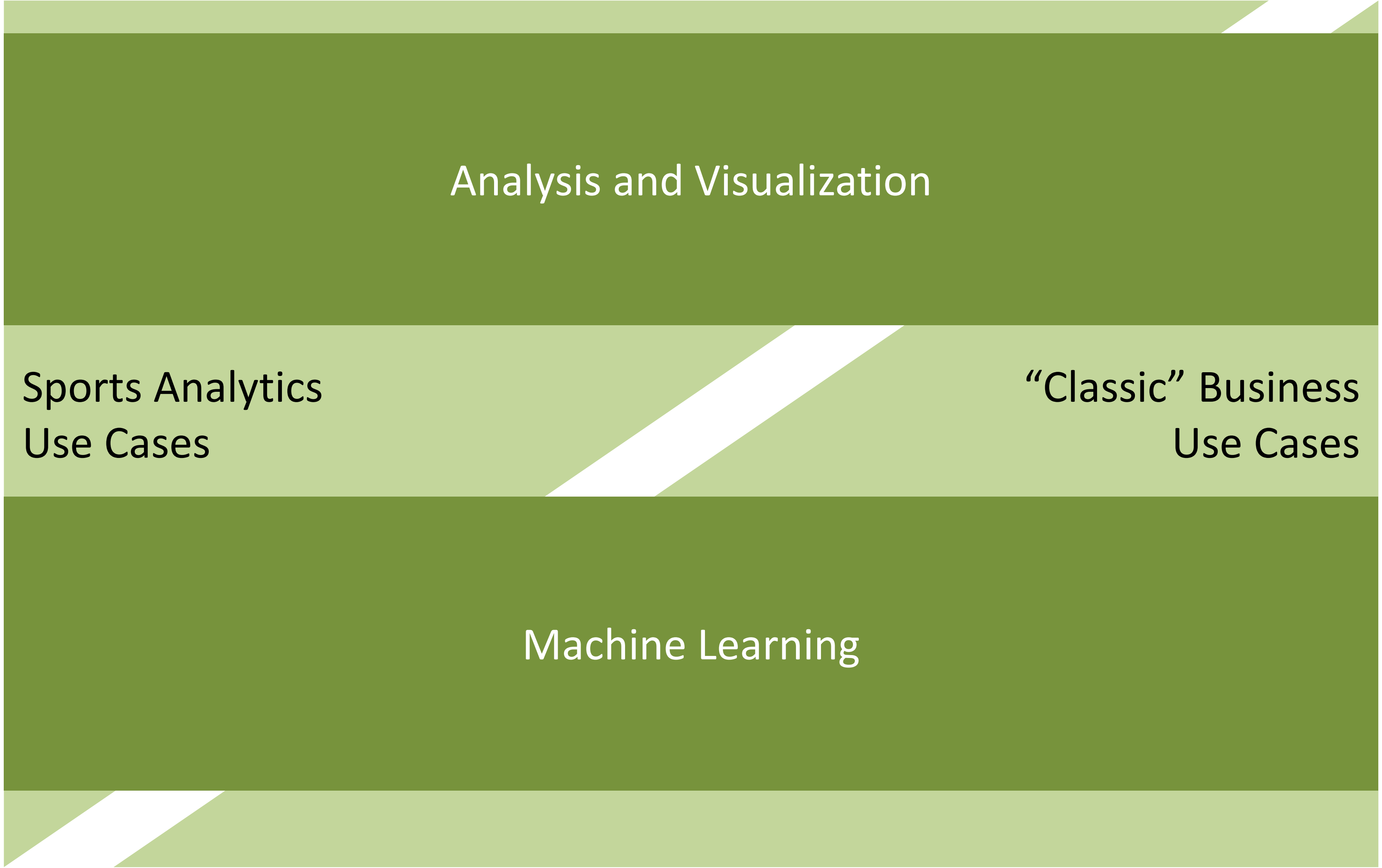
- Foundations of descriptive statistics
- Distributions
- Measures
- Analysis approaches and models

DATA SCIENCE

- Foundations
- Processes
- Methods & Algorithms for structured data

Research Direction

Mission: Conduct design-oriented Data Science and Business Intelligence studies in cooperation with local and global partners



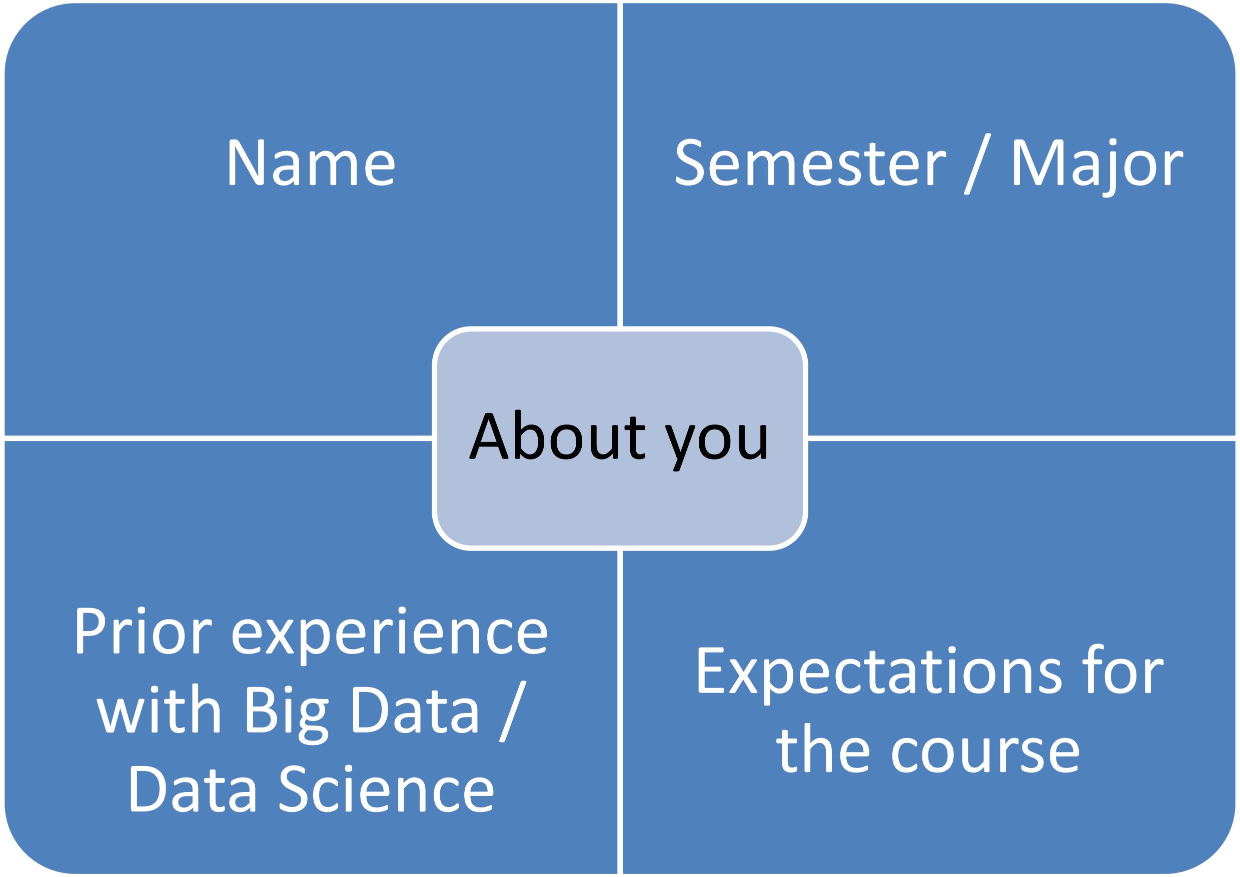
Analysis and Visualization

Sports Analytics
Use Cases

“Classic” Business
Use Cases

Machine Learning

About you



Name

Semester / Major

About you

Prior experience with Big Data / Data Science

Expectations for the course

HOCHSCHULE DER MEDIEN

© Prof. Dr. Hendrik Meth

Today’s session

Agenda

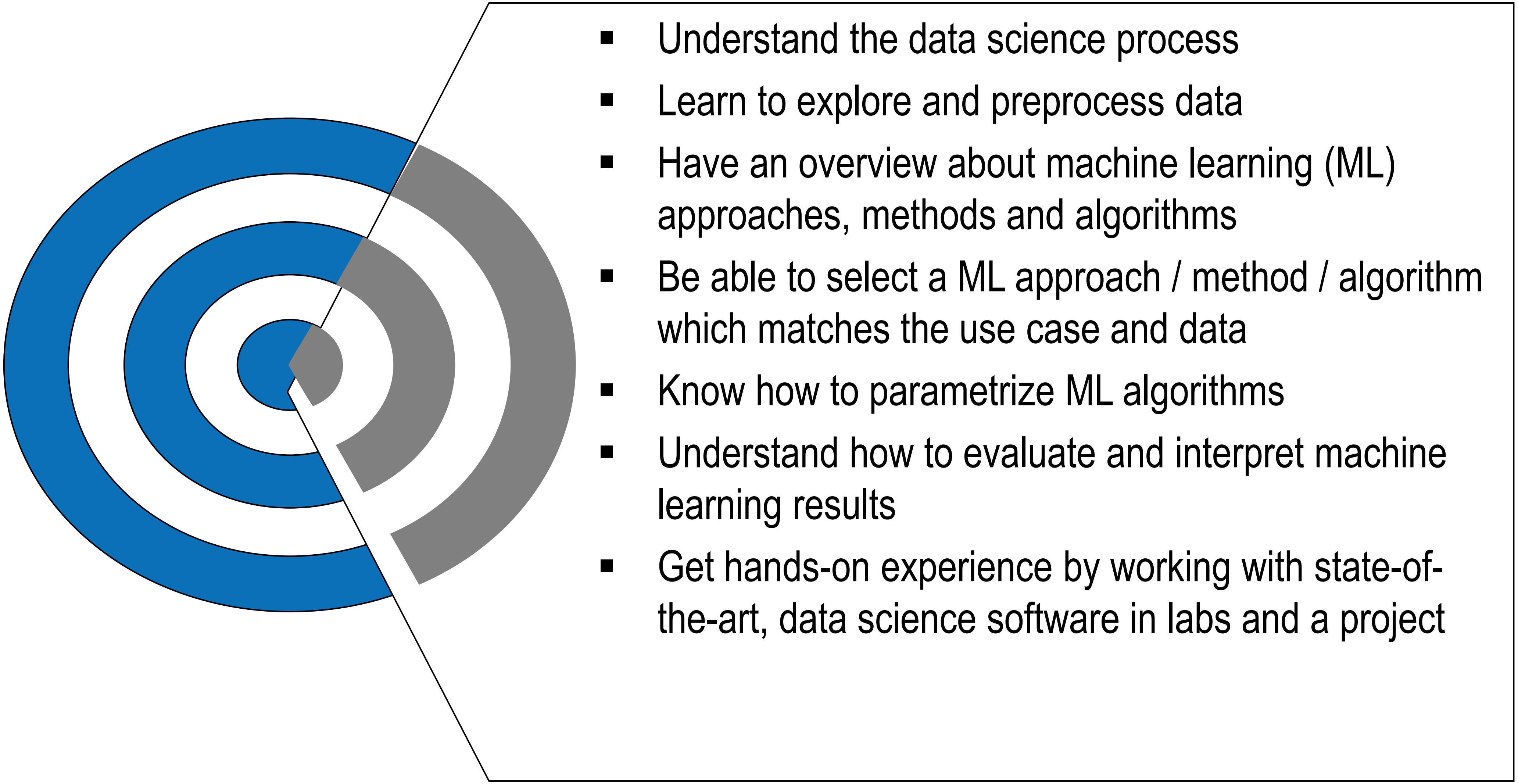
1 Organizational Information

2 Introduction

3 Data Science Process

4 Summary

Goals of this course



Know how to conduct a data science project!

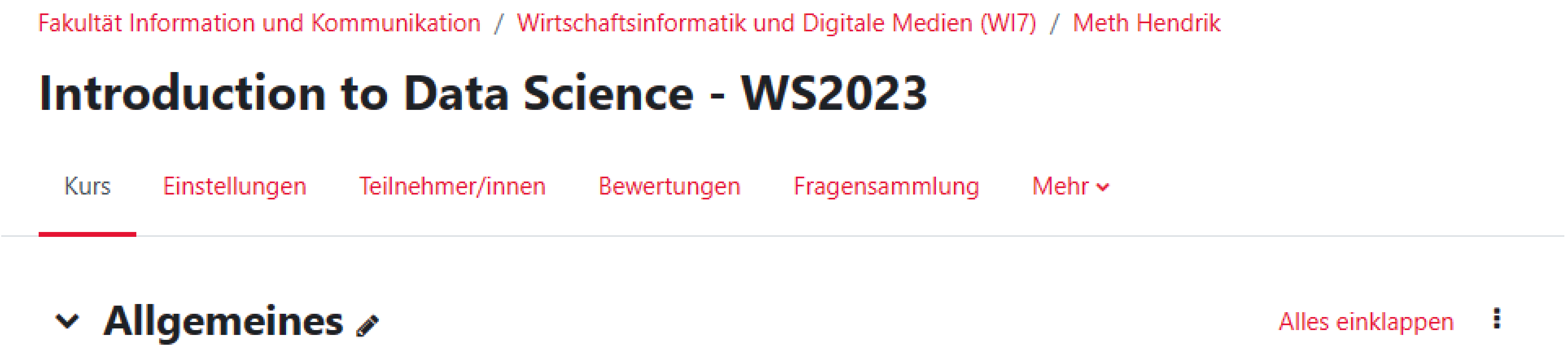
- Understand the data science process
- Learn to explore and preprocess data
- Have an overview about machine learning (ML) approaches, methods and algorithms
- Be able to select a ML approach / method / algorithm which matches the use case and data
- Know how to parametrize ML algorithms
- Understand how to evaluate and interpret machine learning results
- Get hands-on experience by working with state-of-the-art, data science software in labs and a project

Elements helping you achieve your goals

Lecture	▪ Introduces key concepts and provides an environment that enables and facilitates your learning
Lecture Materials	▪ Materials will be made available before each lecture ▪ Literature will provide background information on concepts discussed in class; should be considered as opportunity to extend and deepen your understanding
Labs	▪ Leverage real world Data Science software and apply the concepts introduced in the lecture
Project	▪ Learn to apply methods and technology in a more realistic, project-based context

Course Materials

- On Moodle all lecture and lab materials plus readings will be made available for download
- https://moodle.hdm-stuttgart.de/course/view.php?id=16307 Students registered for this lecture can enroll using registration key: ***

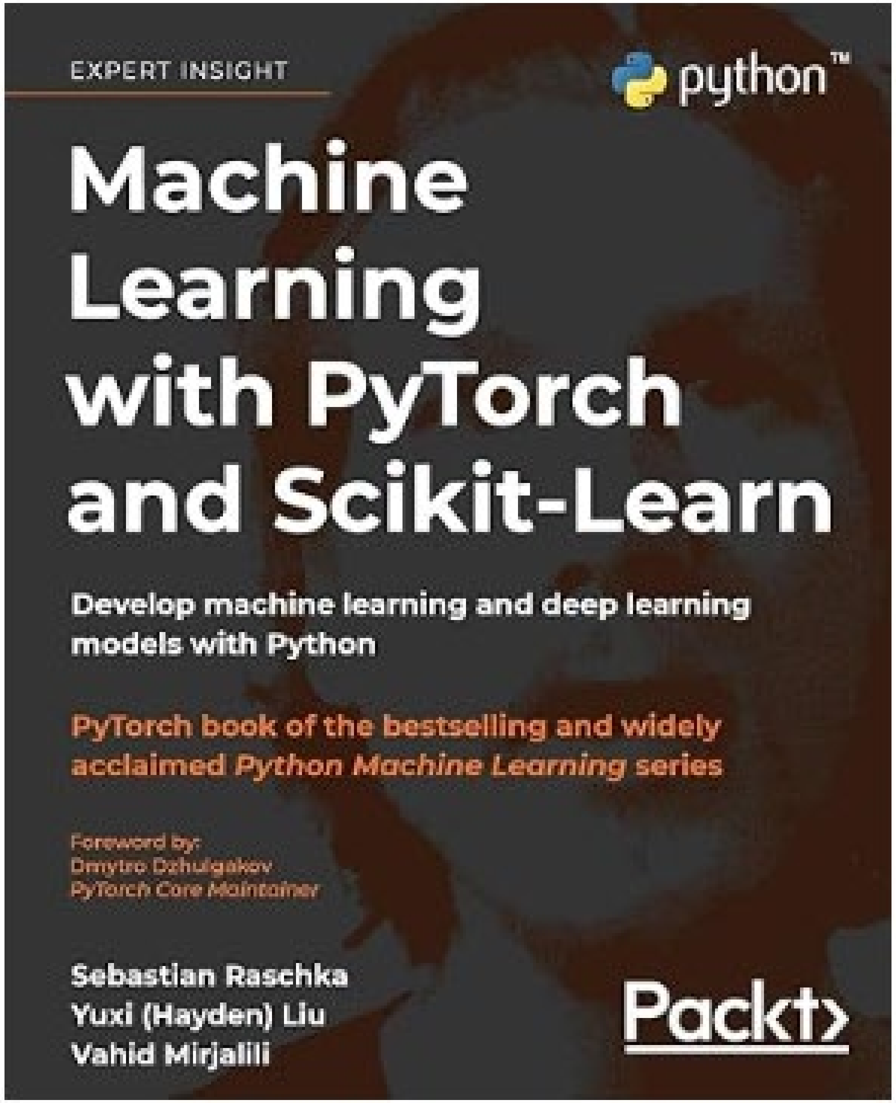


Fakultät Information und Kommunikation / Wirtschaftsinformatik und Digitale Medien (WI7) / Meth Hendrik

Introduction to Data Science - WS2023

Key: IDS23

Text book – Data Science & Python



EXPERT INSIGHT

python™

Machine Learning with PyTorch and Scikit-Learn

Develop machine learning and deep learning models with Python

PyTorch book of the bestselling and widely acclaimed *Python Machine Learning* series

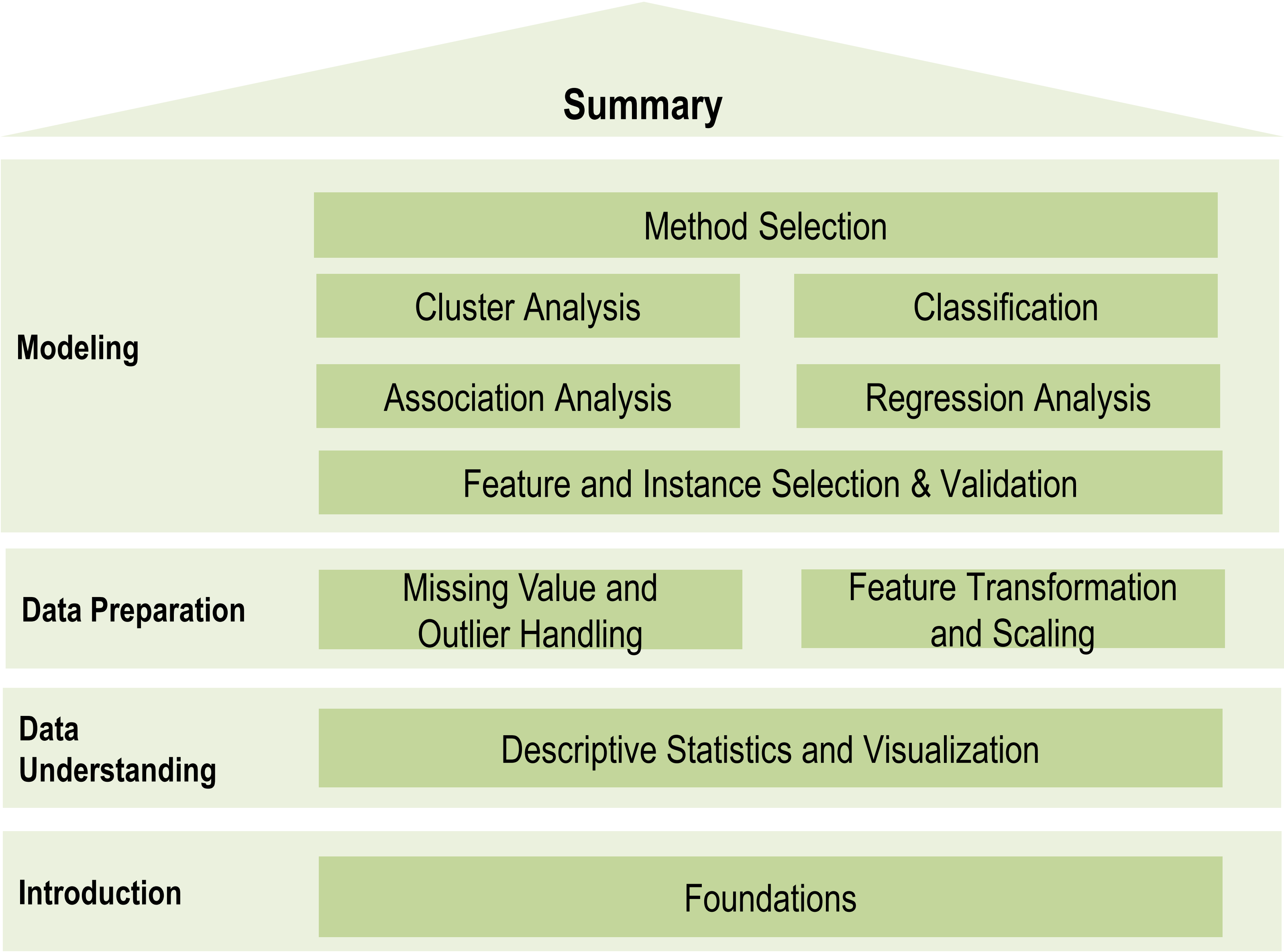
Foreword by:
Dmytro Dzhulgakov
PyTorch Core Maintainer

Sebastian Raschka
Yuxi (Hayden) Liu
Vahid Mirjalili

Packt>

Raschka, S., Liu, Y. H., Mirjalili, V., & Dzhulgakov, D. (2022). *Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing Ltd.

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

Data Understanding

Descriptive Statistics and Visualization

Introduction

Foundations

Schedule

	s202	In class	At home
1	13.10.2023	L: Course Organization & Introduction	T: PANDAS1
2	20.10.2023	L: Data Science Process & Case Introduction	T: PANDAS2 & PANDAS 3
3	27.10.2023	L: Data Exploration & PANDAS Exploration	
4	03.11.2023	L: Data Preparation - Transformation & Scaling	T: Data Preparation - Transformation & Scaling
5	10.11.2023	L: Data Preparation - Missing Values & Outliers	T: Data Preparation - Missing Values & Outliers
6	17.11.2023	**P: Data Exploration**	
7	24.11.2023	L: Data Preparation- Feature & Instance Selection	T: Data Preparation- Feature & Instance Selection
8	01.12.2023	L: Regression	T: Regression
9	08.12.2023	L: Clustering	T: Build and Evaluate a Clustering Model
10	15.12.2023	L: Association Rules	T: Derive and Evaluate Association Rules
11	22.12.2023	**P: Data Preparation**	
	29.12.2023	Christmas Break	
	05.01.2024	Christmas Break	
12	12.01.2024	L: Classification – Decision Trees	T: Build and Evaluate a DT (Ensemble) Model
13	19.01.2024	L: Classification – Log. Regression, Naive Bayes, KNN	T: Build and Evaluate a LogRegression and NaiveBayes Model
14	26.01.2024	**P: Model Optimization**	

L = Lecture / T = Technology Lab / P = Presentation

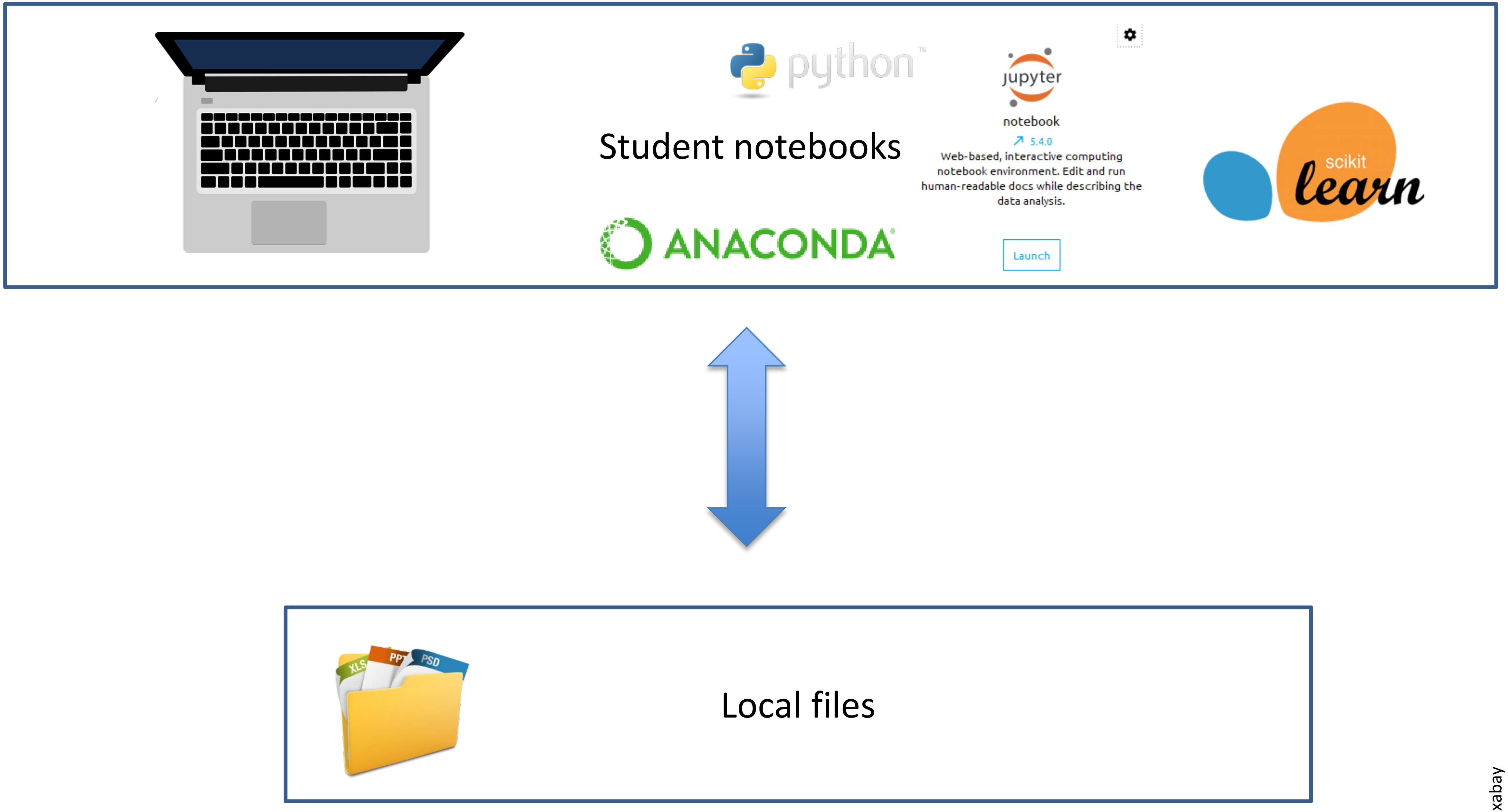
Course Grading

Element	Description	Exam / Due Date
Project (4 teams of 4)	• Analysis case study to be solved with Python in teams • Teams can be chosen by yourself	Due Dates: • WP1 Data Exploration • Moodle upload / GitHub Freeze: 16.11.2023 • Presentation: 17.11.2023 • WP2 Data Preparation • Moodle upload / GitHub Freeze: 21.12.2023 • Presentation: 22.12.2023 • WP3 Modeling & Validation • Moodle upload / GitHub Freeze: 25.1.2024 • Presentation: 26.1.2024

16-17 participants

HOCHSCHULE DER MEDIEN

Python Labs - Infrastructure



Student notebooks

notebook

↗ 5.4.0

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Local files

Picture Source: Pixabay

Contact

Prof. Dr. Hendrik Meth
Consultation hour: per request
Phone: +49 711 8923-3287
E-Mail: meth@hdm-stuttgart.de

HOCHSCHULE DER MEDIEN

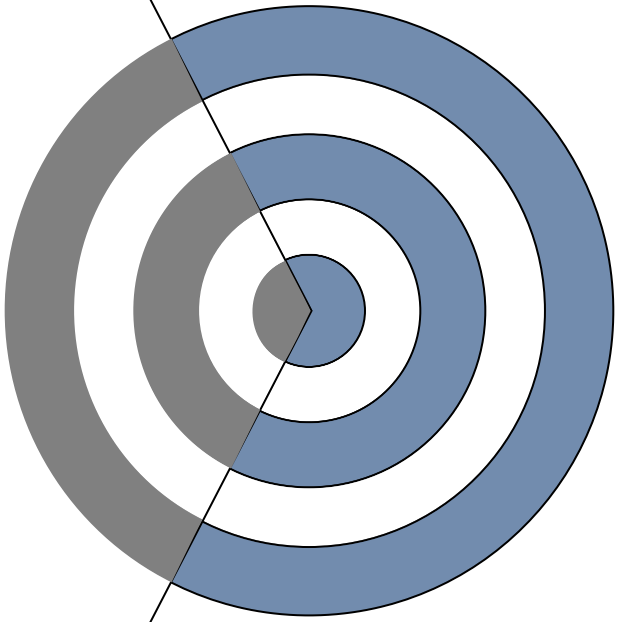
Today’s session

	Agenda
1	Organizational Information
2	Introduction
3	Data Science Process
4	Summary

Goals of today‘s session

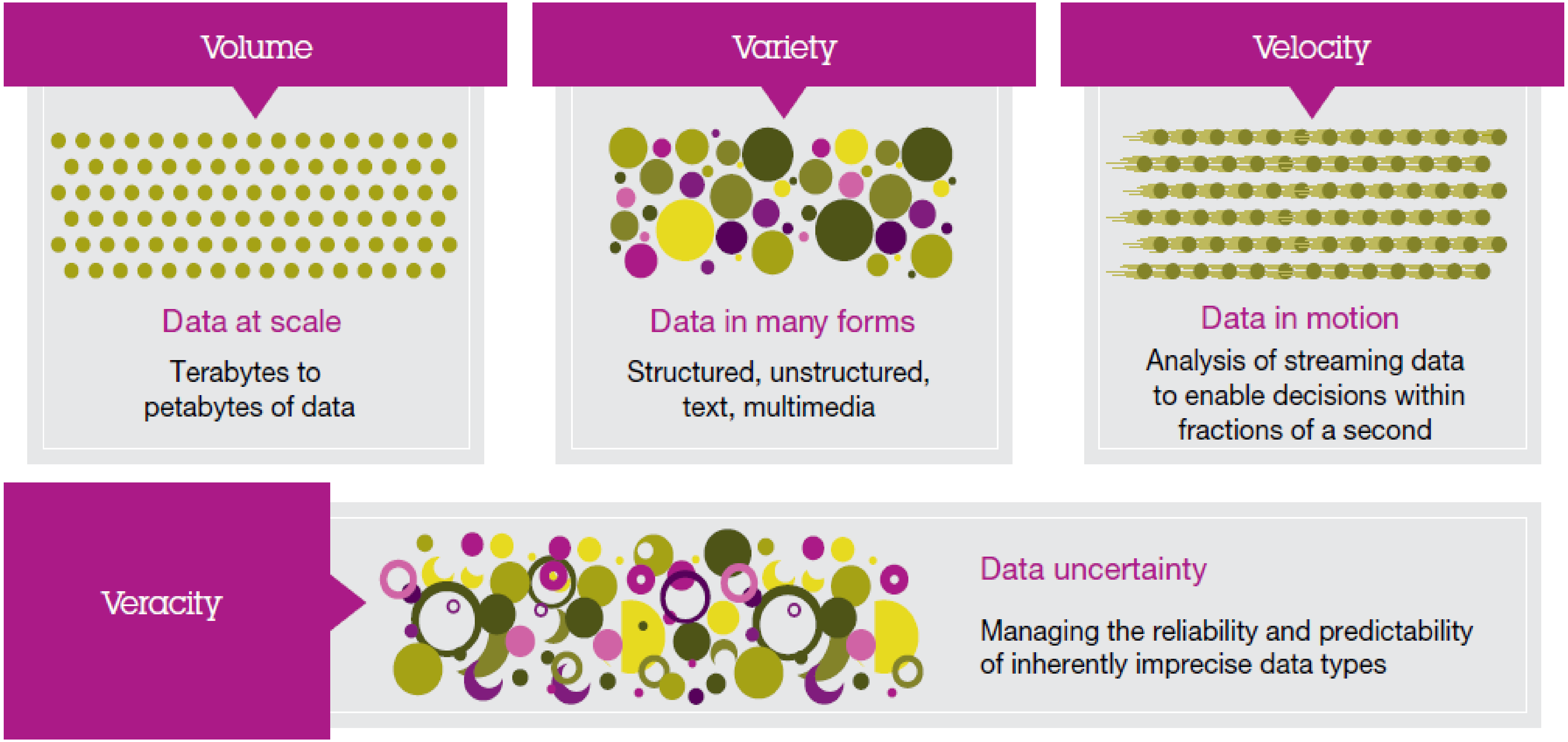
After completing this session, you should be able to

- Define Data Science
- Describe different types of Data Science
- Characterize the Data Science process
- Give examples of Data Science applications



What is Big Data ?

- Four dimensions to be differentiated



Volume

Data at scale

Terabytes to petabytes of data

Variety

Data in many forms

Structured, unstructured, text, multimedia

Velocity

Data in motion

Analysis of streaming data to enable decisions within fractions of a second

Veracity

Data uncertainty

Managing the reliability and predictability of inherently imprecise data types

Source: Schroeck et al. 2012 – IBM Institute for Business Value

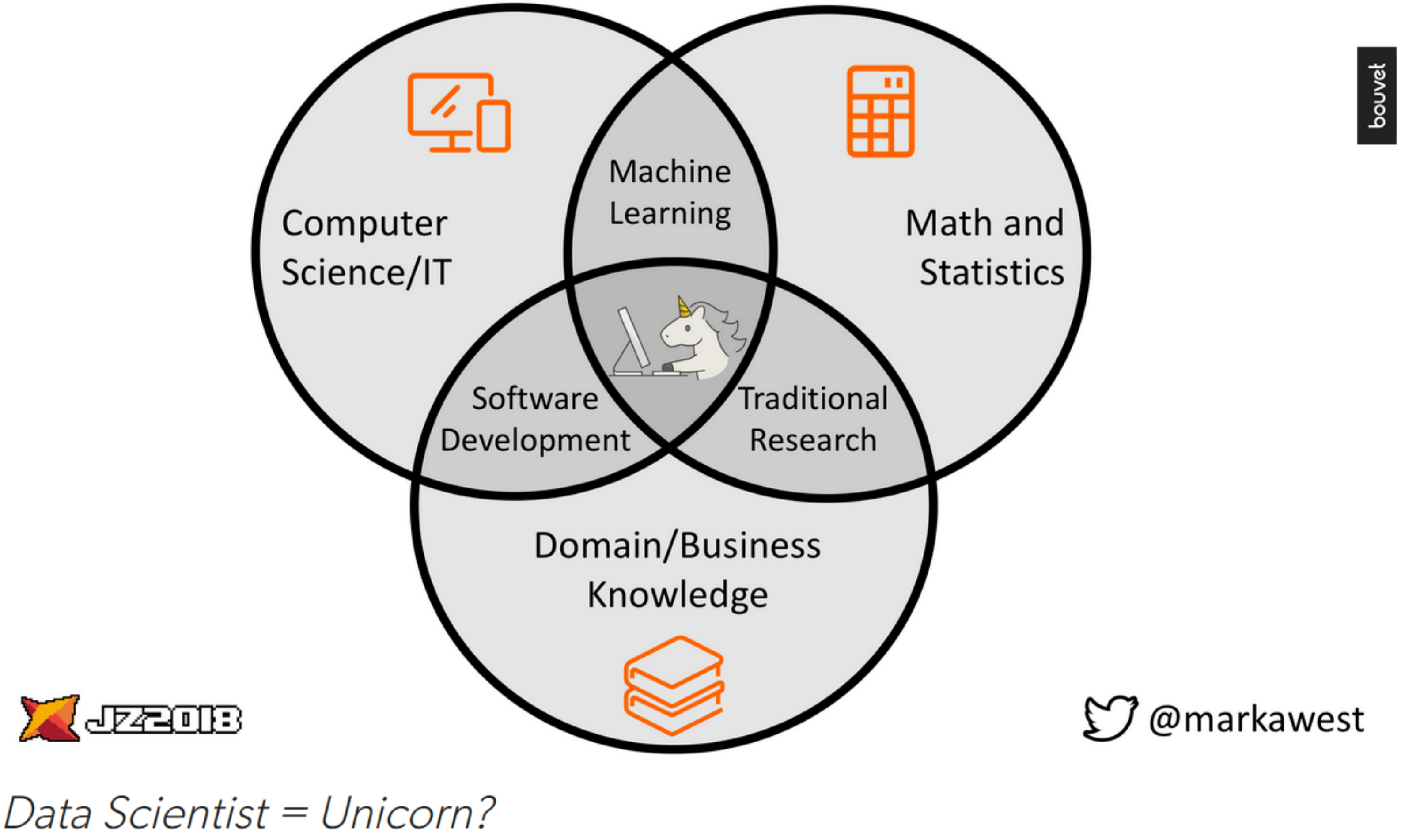
What is Data Science?

- Data science involves using automated methods to analyze massive amounts of data in different forms and to extract knowledge from them
- It is a continuation of some of the data analysis fields such as analytics, statistics and data mining
- It includes predictive methods (e.g. regression analysis) and exploratory ones (e.g. cluster analysis)

The Data Scientist

adventures in big data world

What is a Data Scientist ?



Computer Science/IT

Machine Learning

Math and Statistics

Software Development

Traditional Research

Domain/Business Knowledge

bouvet

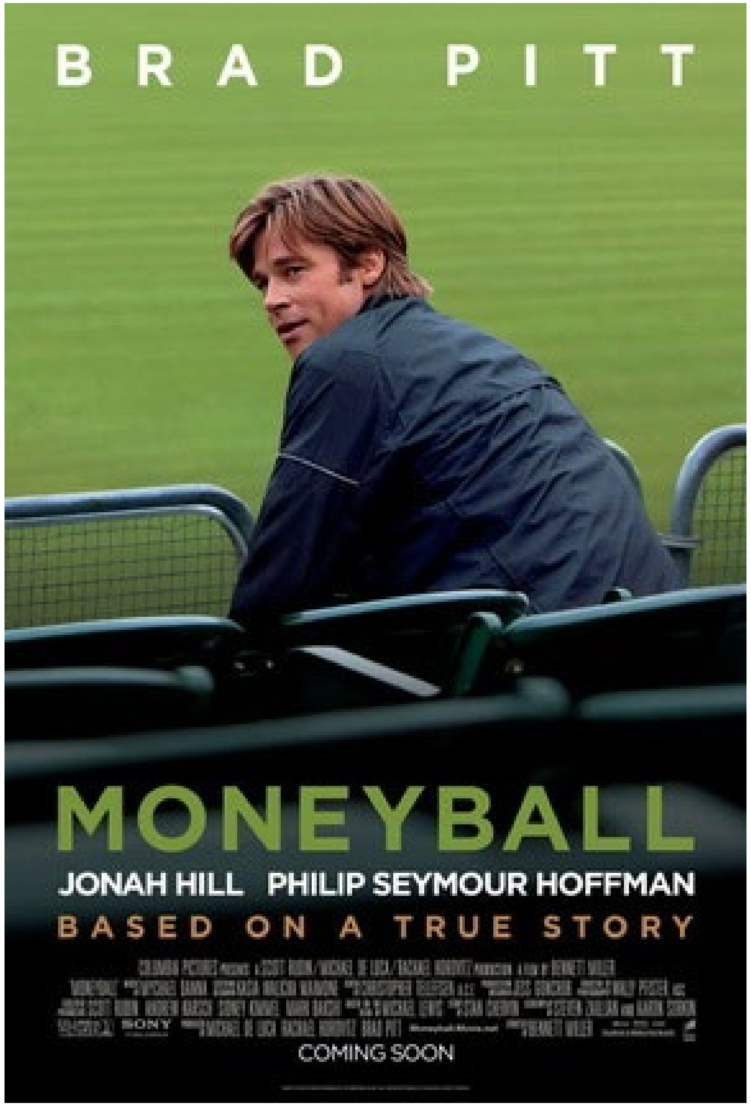
JZ2018

@markawest

Data Scientist = Unicorn?

Source: https://www.bouvet.no/bouvet-deler/roles-in-a-data-science-project

Moneyball

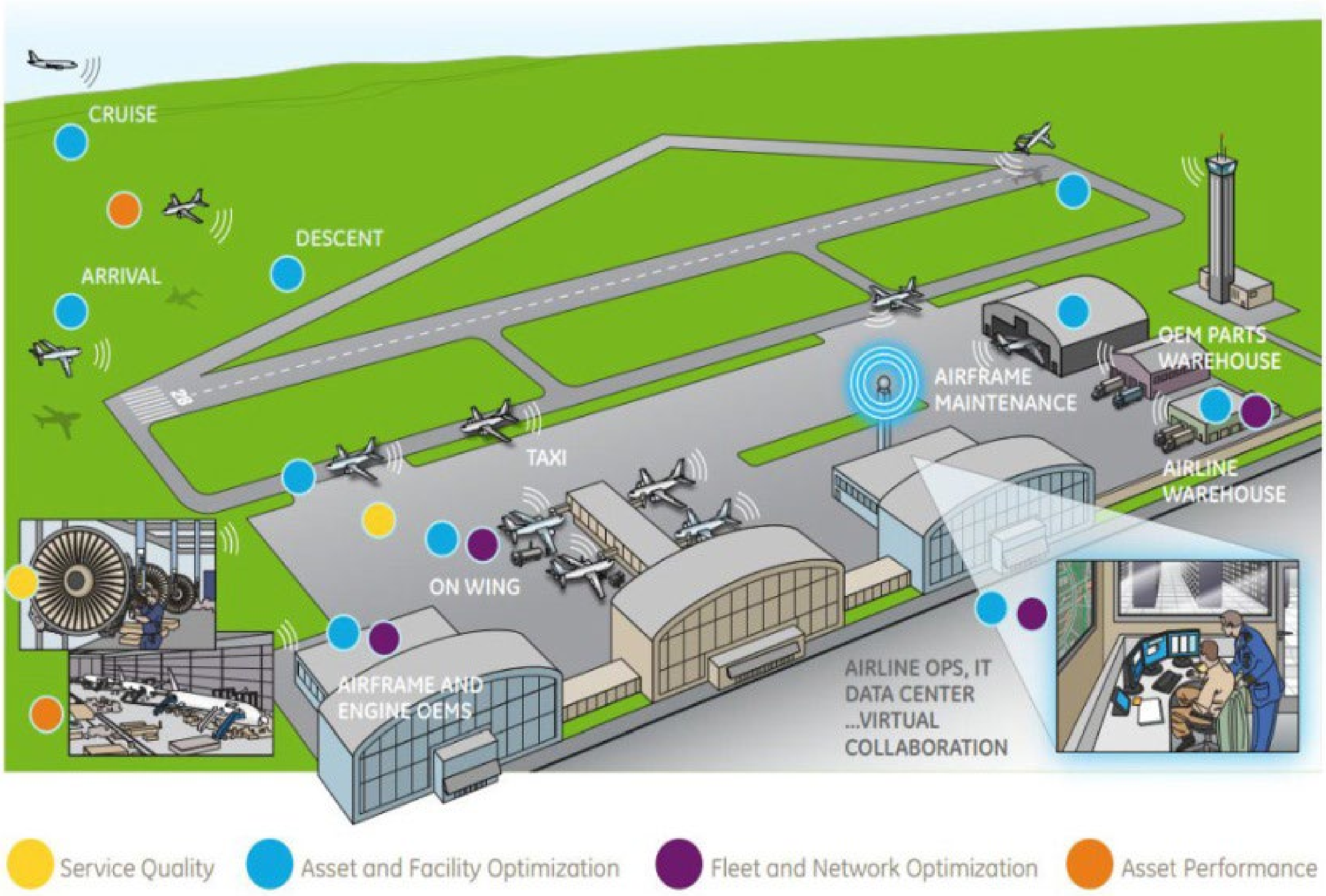


Picture source: wikipedia.com

- Application of Data Science methods to Major League Baseball
- In 2002 Oakland Athletics built a team of undervalued talent by taking a new data science-driven approach towards scouting and analyzing players

Etihad

Etihad Airways uses data science to analyze vast amounts of data that are generated in **real-time** by the **sensors on every plane**.



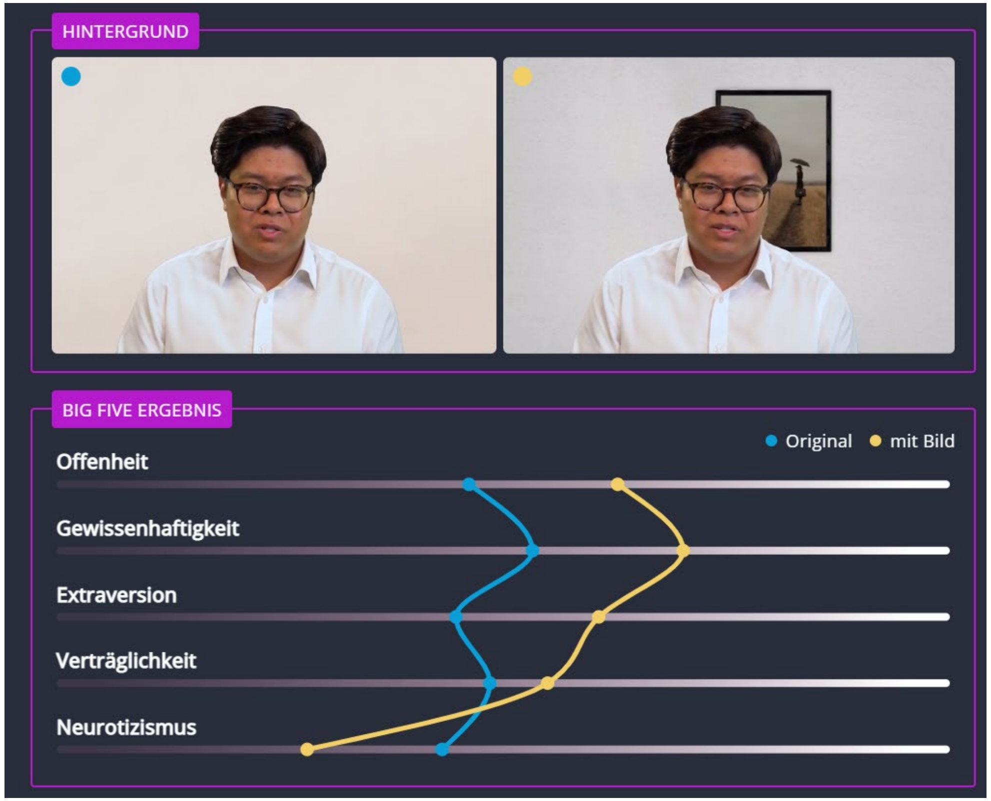
الإتحاد
ETIHAD
AIRWAYS
ABU DHABI

HOCHSCHULE DER MEDIEN

© Prof. Dr. Hendrik Meth

Source: http://www.bigdata-startups.com/BigData-startup/etihad-airways-big-data-reach-destination/

Data Science / AI in Human Resources context

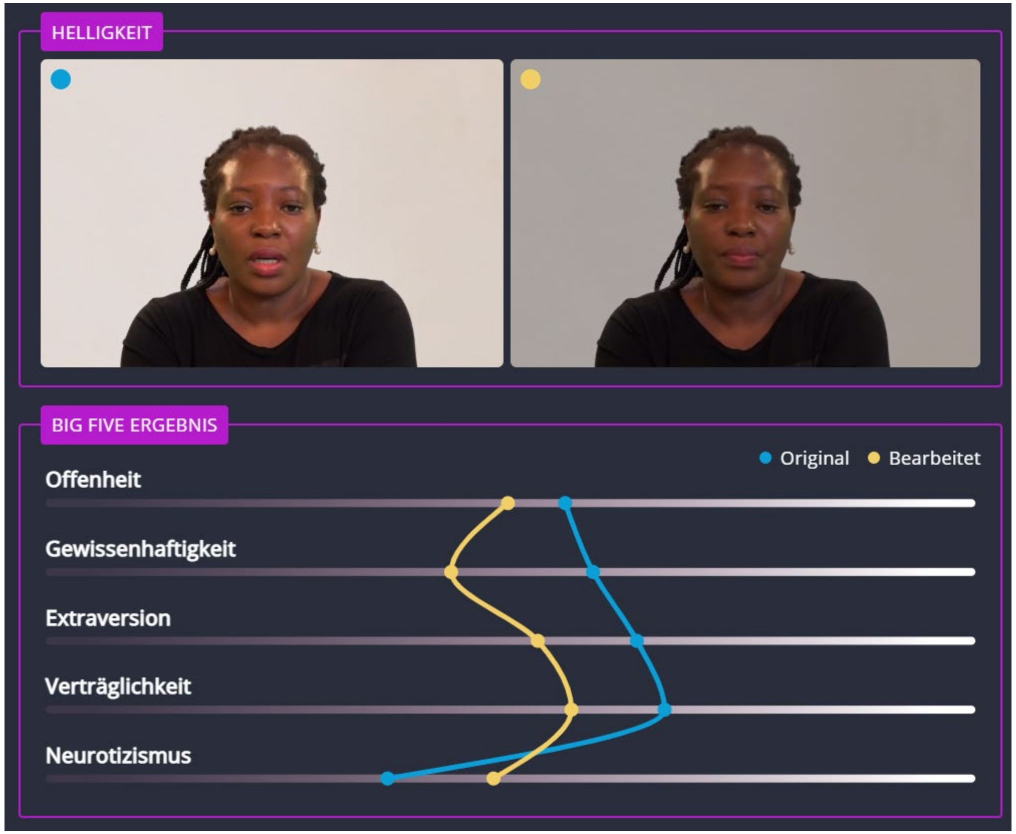


HINTERGRUND

BIG FIVE ERGEBNIS

● Original ● mit Bild

- Offenheit
- Gewissenhaftigkeit
- Extraversion
- Verträglichkeit
- Neurotizismus



HELLIGKEIT

BIG FIVE ERGEBNIS

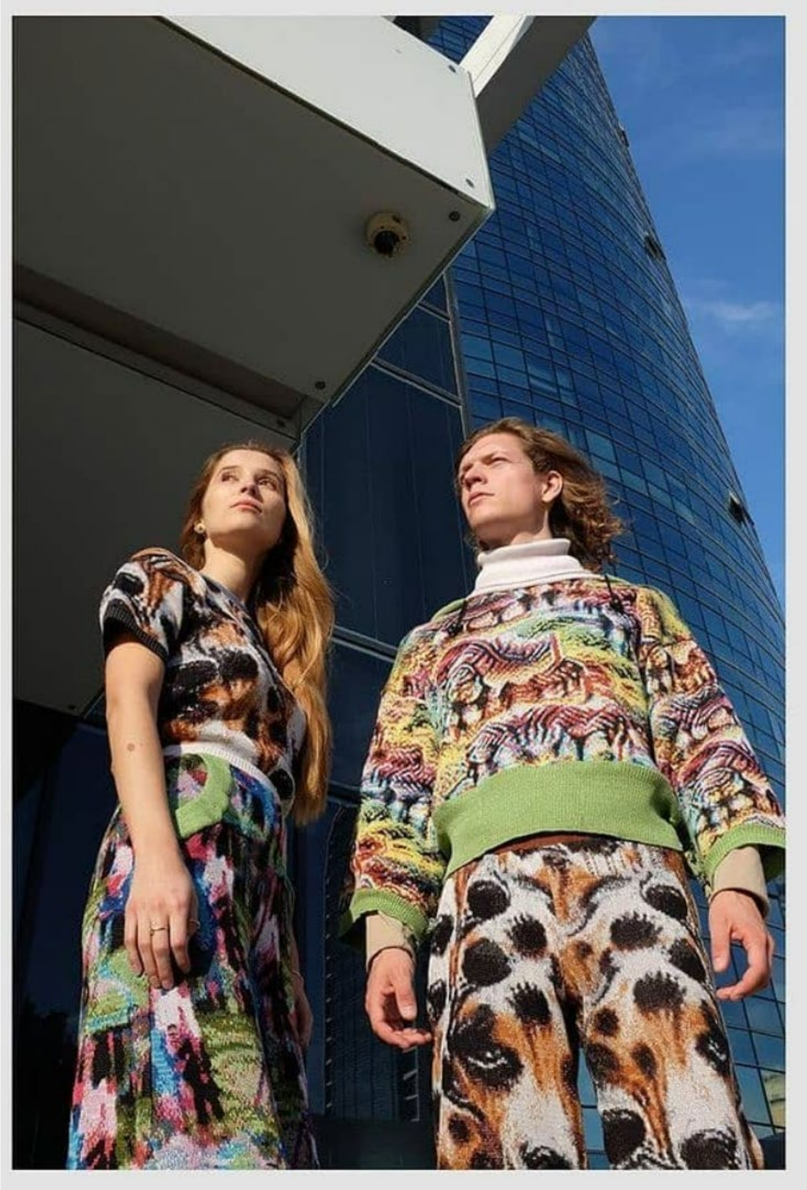
● Original ● Bearbeitet

- Offenheit
- Gewissenhaftigkeit
- Extraversion
- Verträglichkeit
- Neurotizismus

https://web.br.de/interaktiv/ki-bewerbung/

HOCHSCHULE
DER MEDIEN

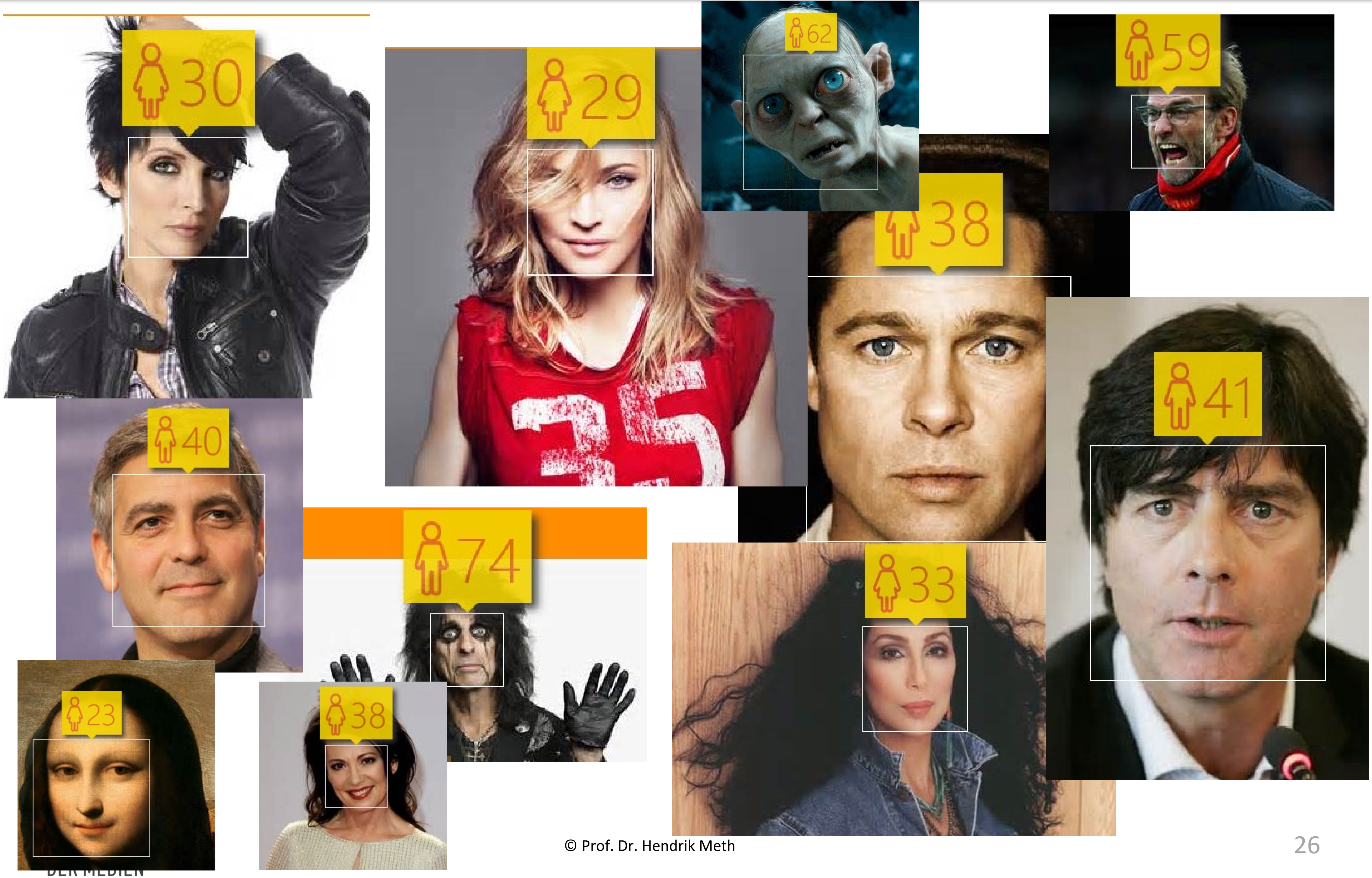
Fashion vs. Face Recognition



Source: https://www.instagram.com/capable.design/?utm_source=ig_embed&ig_rid=1ce6a251-98bd-4d16-b76b-177748c7f1dc

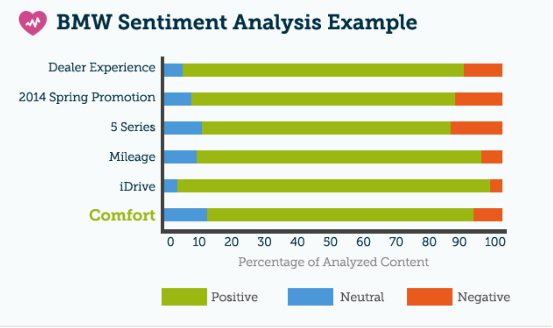
How-Old.net

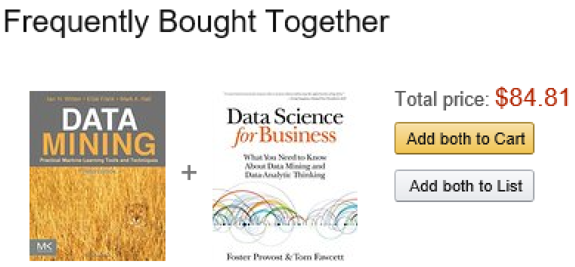
Some surprising insights..



Other Applications of Data Science

- Business
  - Customer relationship management, e-commerce,
  - Fraud detection, targeted marketing, sentiment analysis
- Web and Social Media
  - Advertising, search engine optimization, spam detection, web site optimization, personalization,
- Government
  - Crime detection, profiling tax cheaters, …

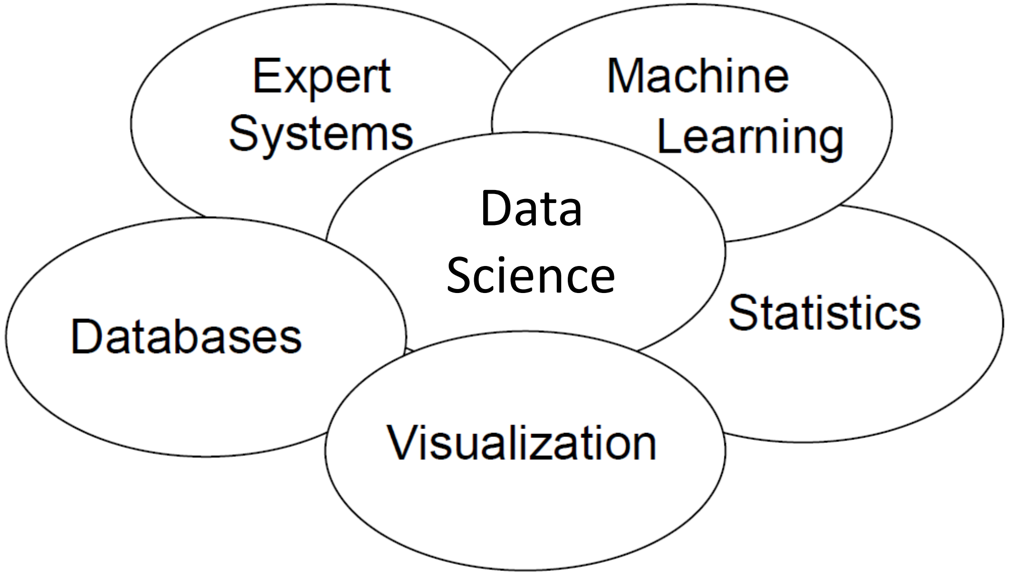



Frequently Bought Together

Total price: $84.81

DATA MINING
+
Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett
Add both to Cart
Add both to List

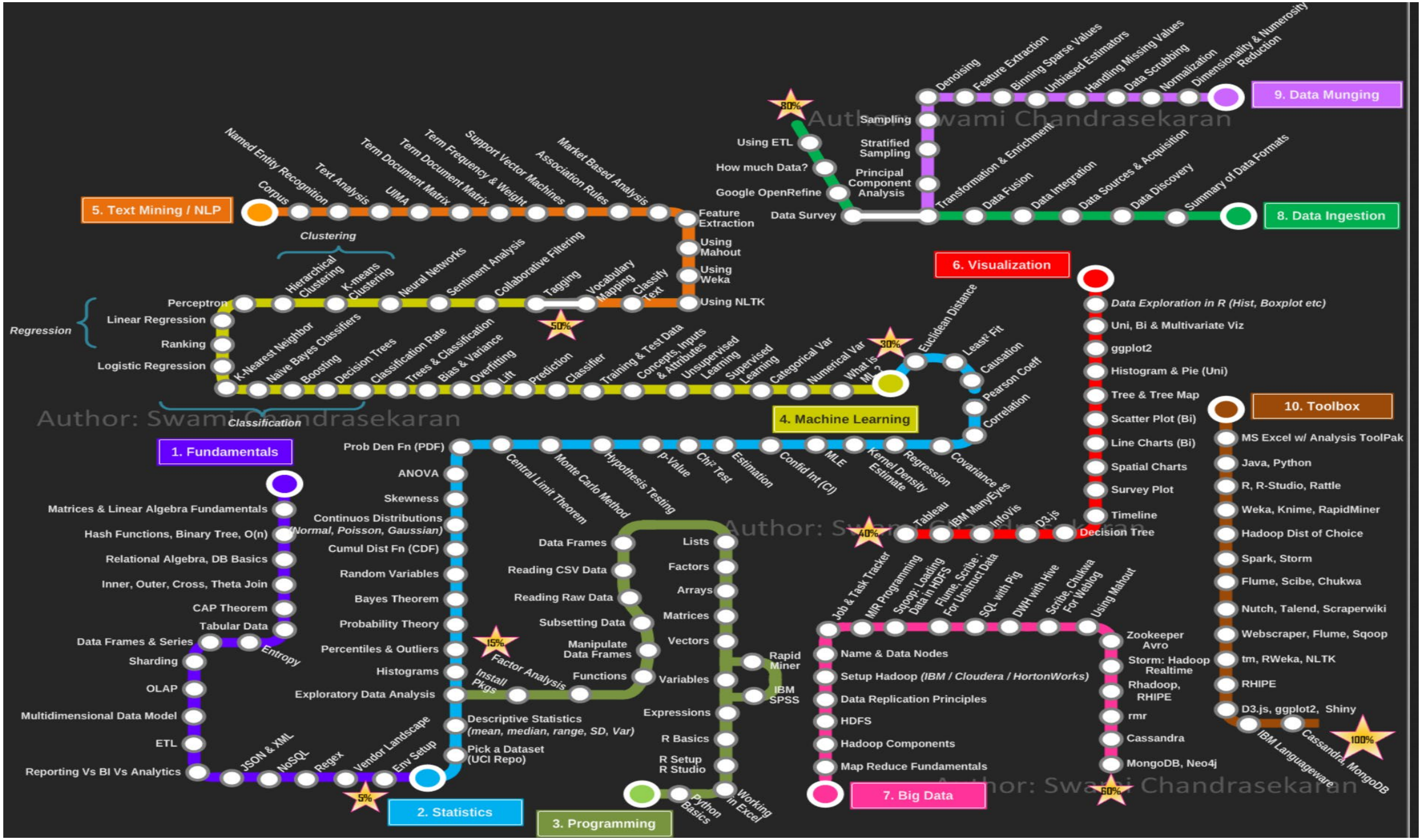
Data Science is Inter-Disciplinary



Traditional techniques may not work due to

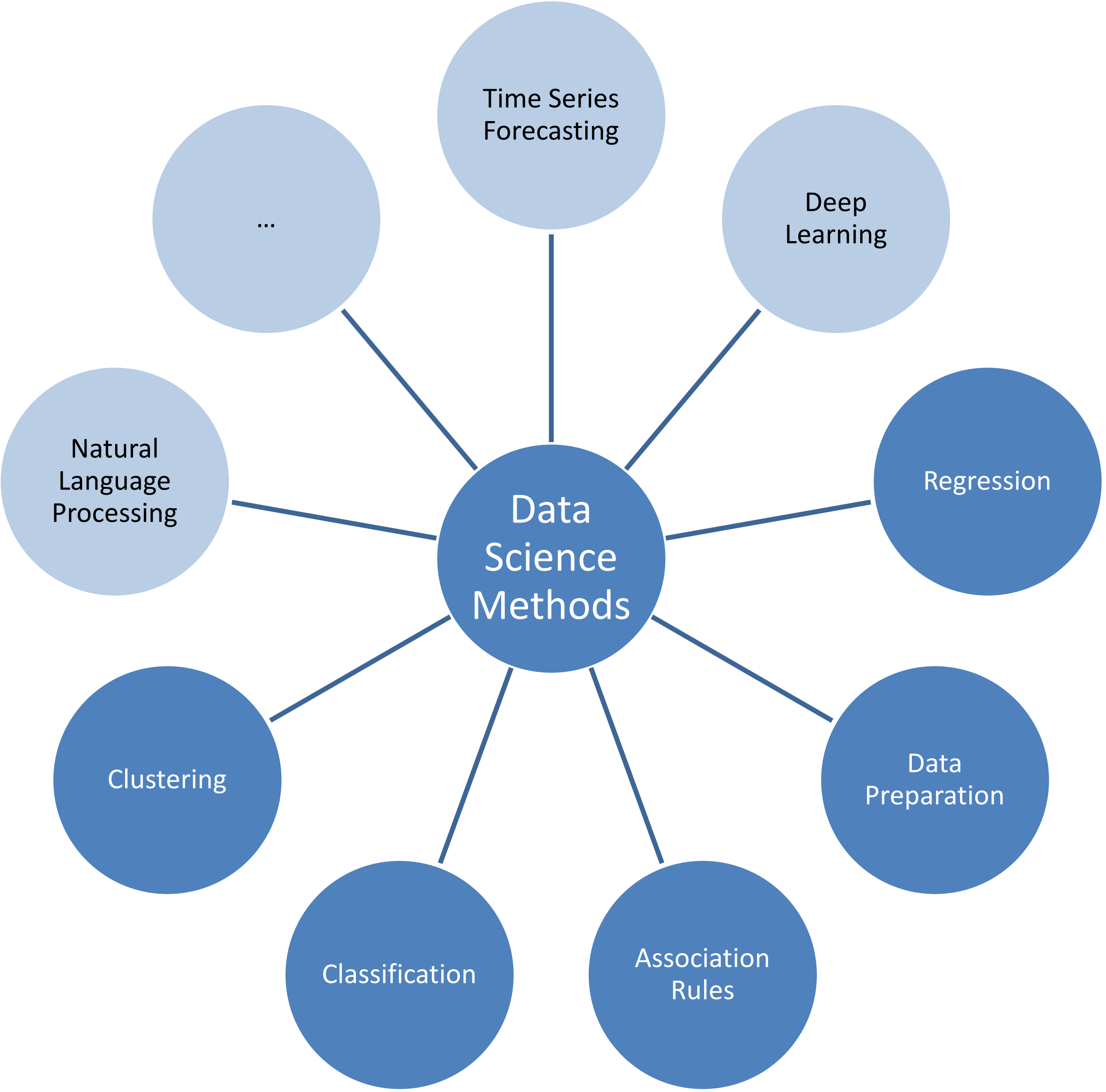
- large amount of data
- high dimensionality of data
- heterogeneous,distributed nature of data

Data Science is Inter-Disciplinary



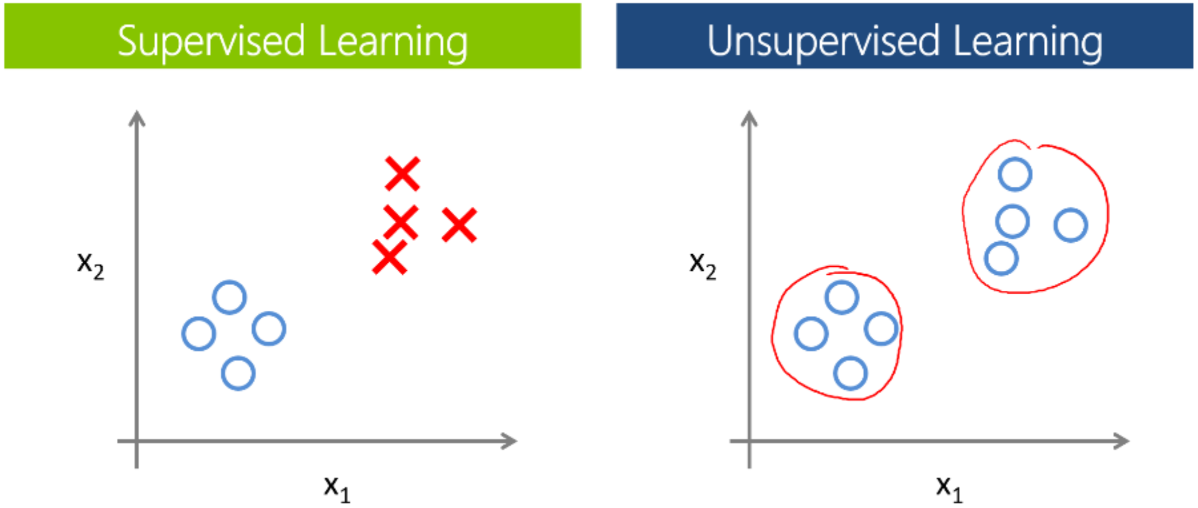
Source: https://cdn.datafloq.com/cms/2015/10/26/long-road-to-data-scientist.png?utm_source=datafloq&utm_medium=ref&utm_campaign=datafloq

Data Science Methods - Overview



Supervised vs. Unsupervised

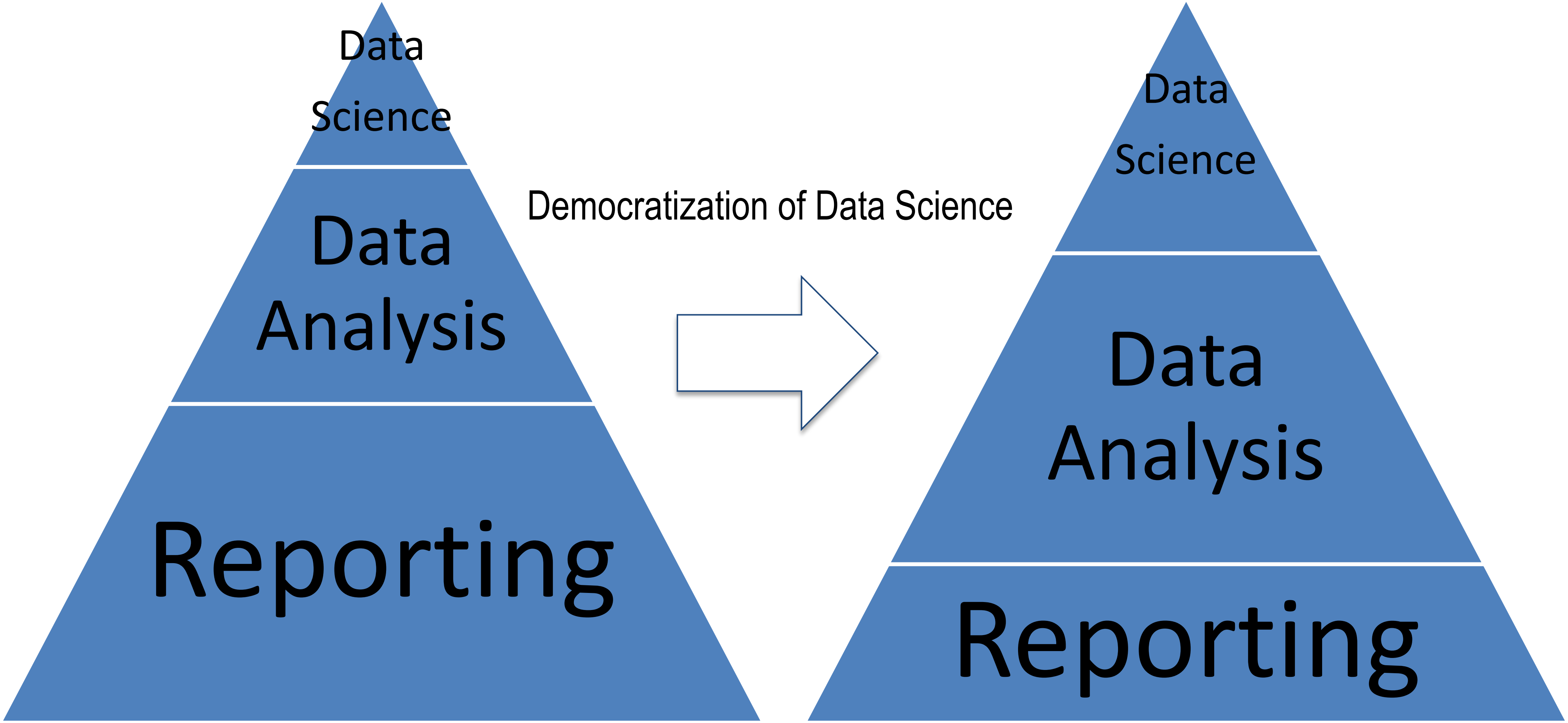
- **Supervised learning:**
  - Goal: predict data with unknown target attribute value with minimal error
  - Search for dependencies of a target attribute on the input data.
- **Unsupervised learning:**
  - Goal: Create a pattern of a more compact description of the data
  - No reference to target attribute, error not measureable.



© Prof. Dr. Hendrik Meth

Graph: Lecture 1, Andrew Ng's Machine Learning course on Coursera

Data Science Usage and Users



Drivers for Democratization of Data Science:

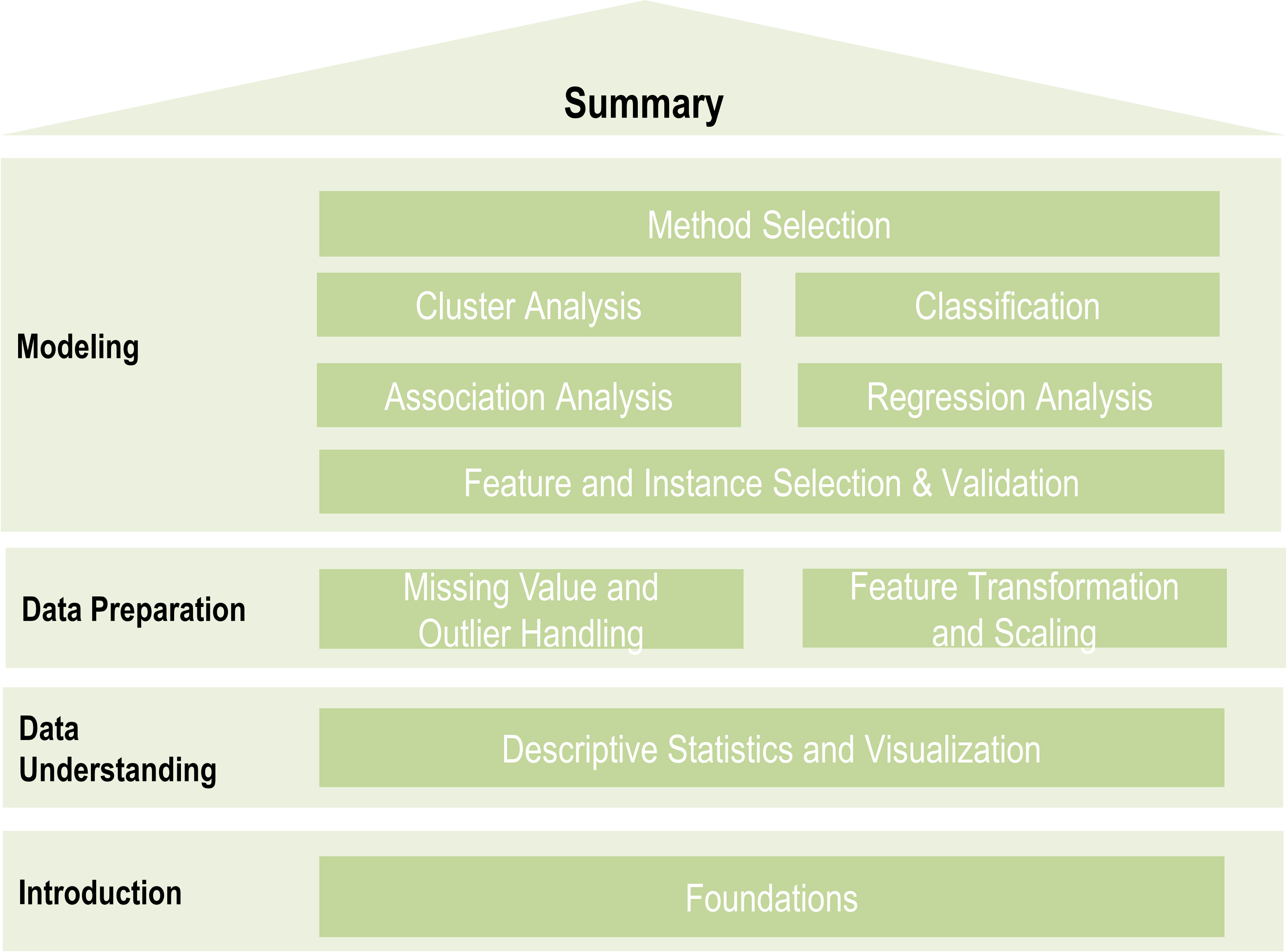
- Increasing Data Volumes
- Data Science capabilities as a competitive advantage
- Partial automation of analyses
- Improvement of tool support

HOCHSCHULE DER MEDIEN

Today's session

	Agenda
1	Organizational Information
2	Introduction
3	Course Overview
4	Summary

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

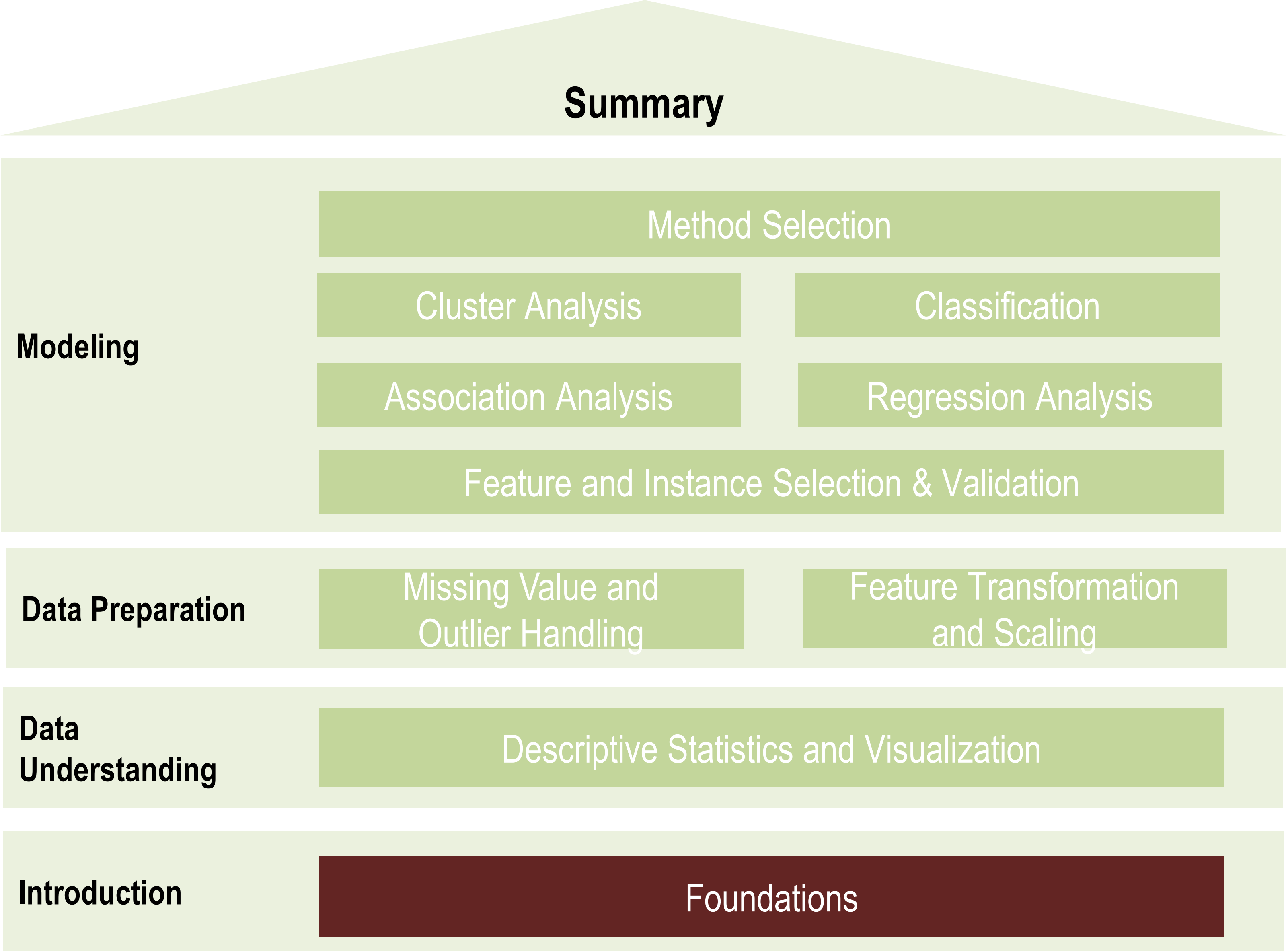
Data Understanding

Descriptive Statistics and Visualization

Introduction

Foundations

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

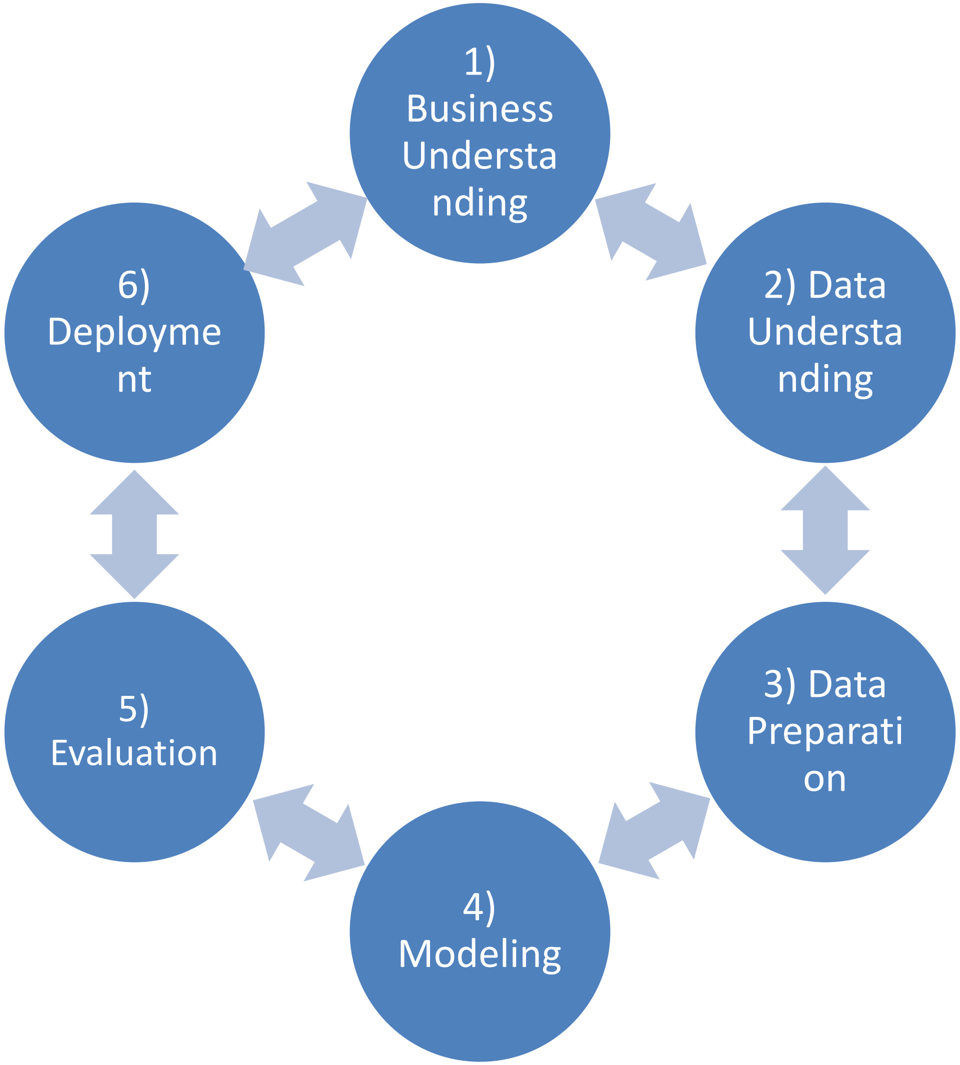
Data Understanding

Descriptive Statistics and Visualization

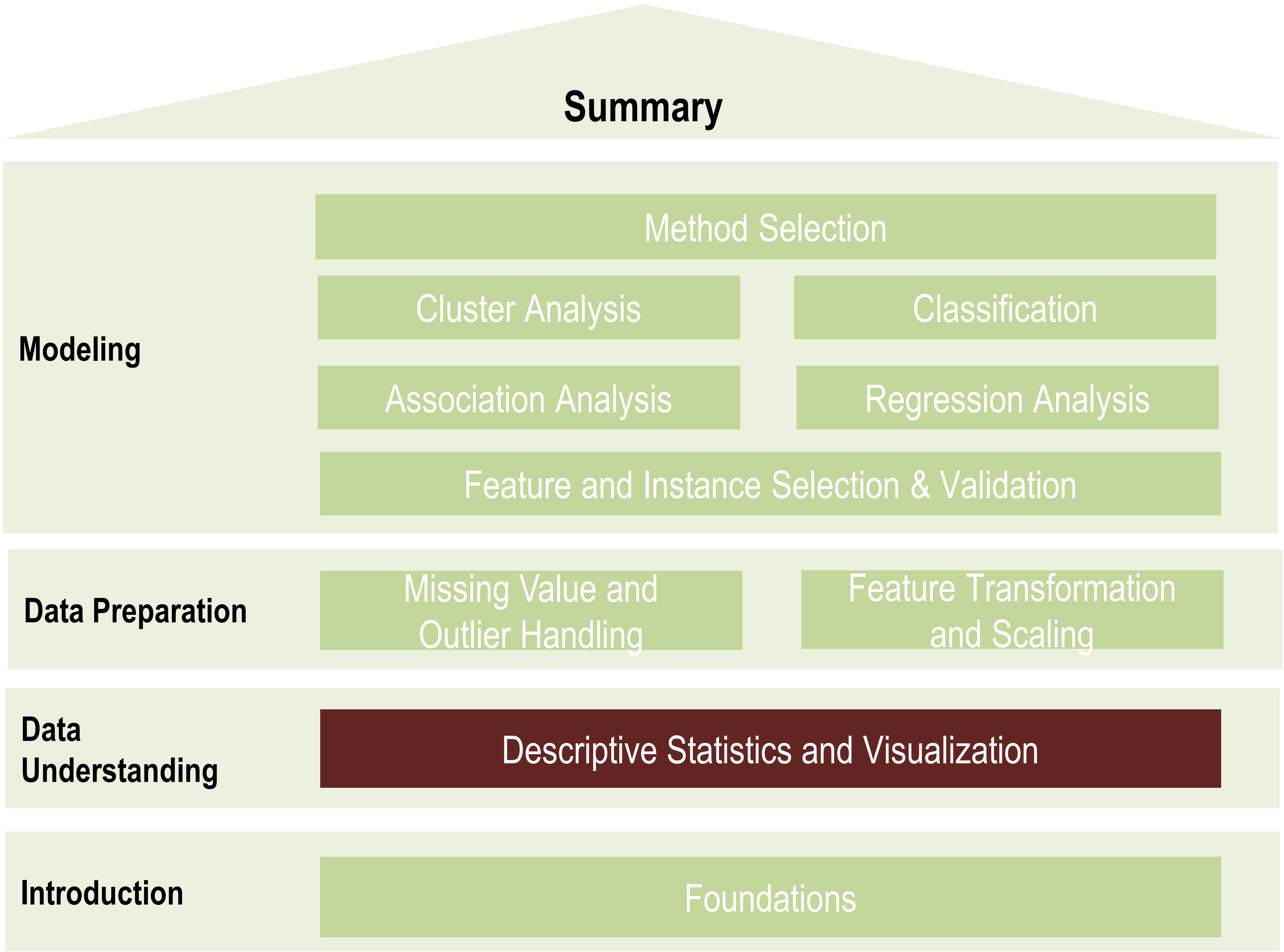
Introduction

Foundations

CRISP-DM


1) Business Understa nding
2) Data Understa nding
3) Data Preparati on
4) Modeling
5) Evaluation
6) Deployme nt

Lectures: Overview



Summary

Modeling
- Method Selection
- Cluster Analysis
- Classification
- Association Analysis
- Regression Analysis
- Feature and Instance Selection & Validation

Data Preparation
- Missing Value and Outlier Handling
- Feature Transformation and Scaling

Data Understanding
- Descriptive Statistics and Visualization

Introduction
- Foundations

Data Exploration

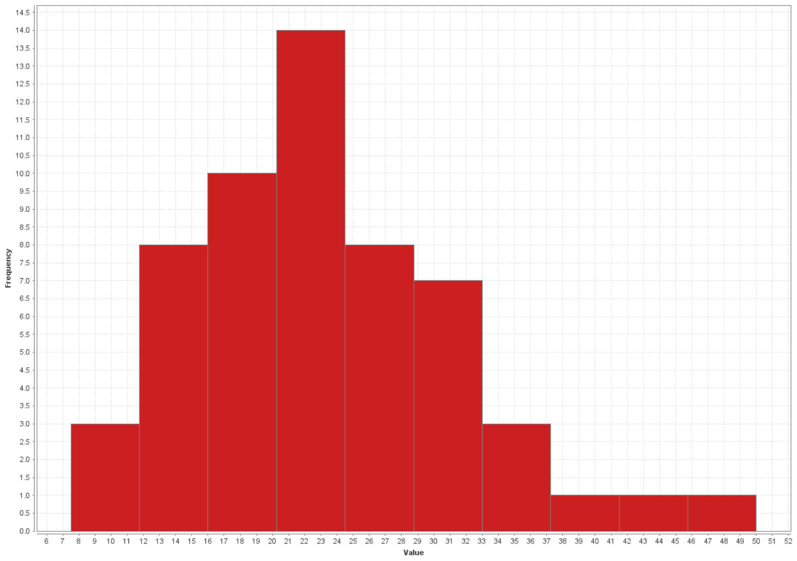
- Before applying advanced data science methods (such as clustering or classification) it is essential to perform basic **data exploration** to study the main characteristics of the data.
- Data exploration helps to
  - Understand the data better
  - Prepare the data for advanced analysis
  - Get insights sometime faster than using advanced methods
  - Interpret results of advanced methods

HOCHSCHULE DER MEDIEN

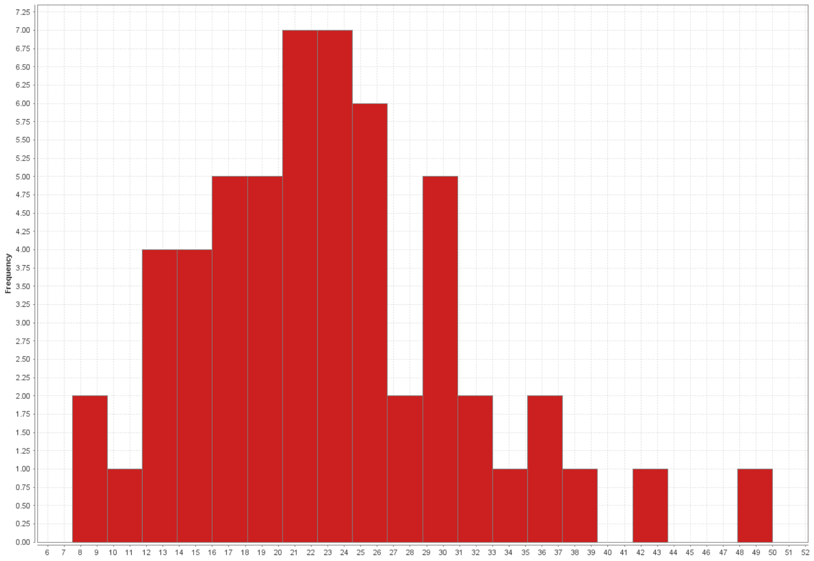
Univariate Exploration

Data Exploration - Histogram

- One of the most basic visual ways to understand the frequency of occurrence of a range of values for one variable
- Numeric variable to be analyzed takes the horizontal and its frequency of occurrence the vertical axis
- Histograms are used to find the central location, range, and shape of distribution.

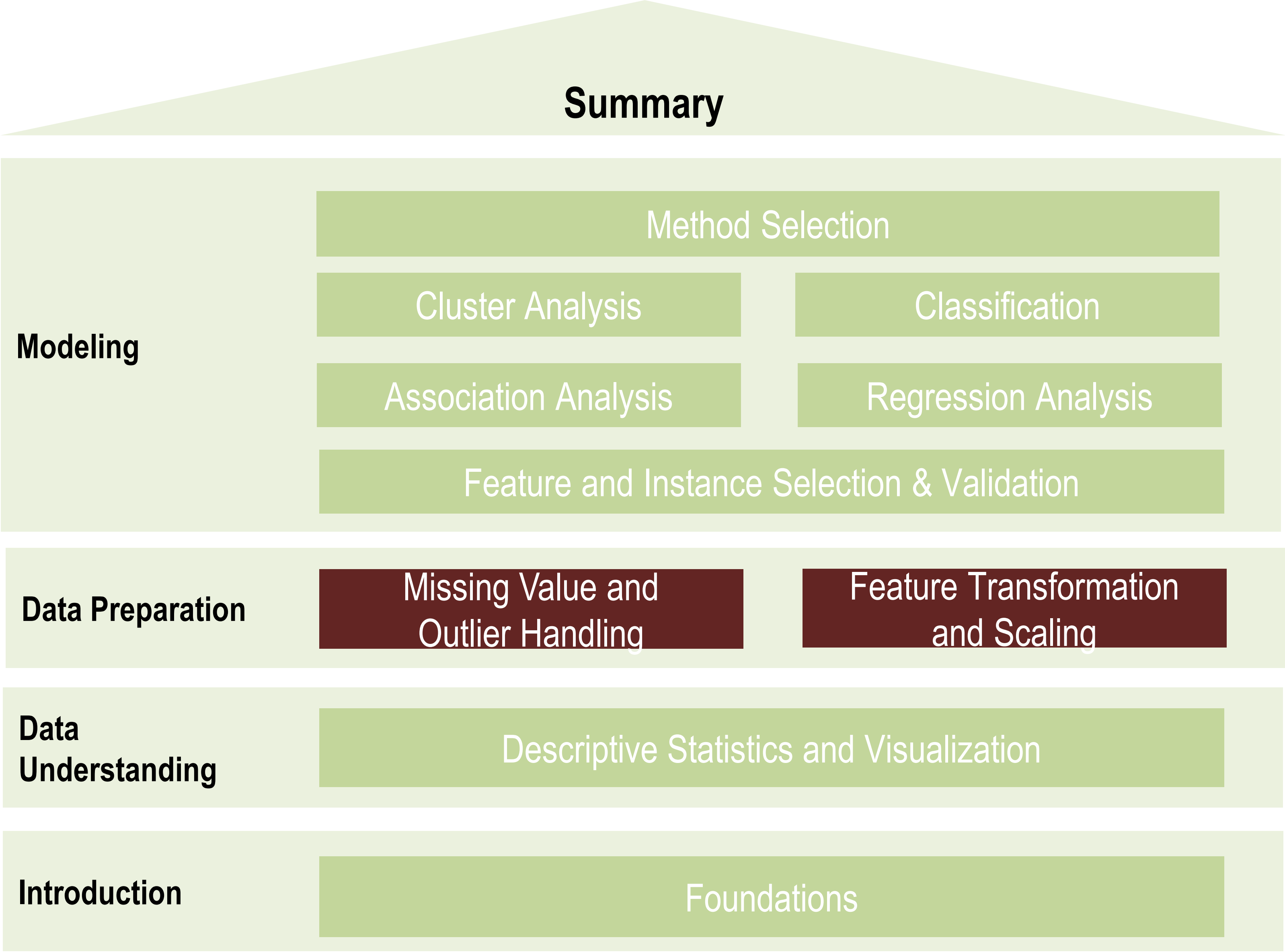


Histogram with 10 bins



Histogram with 20 bins

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

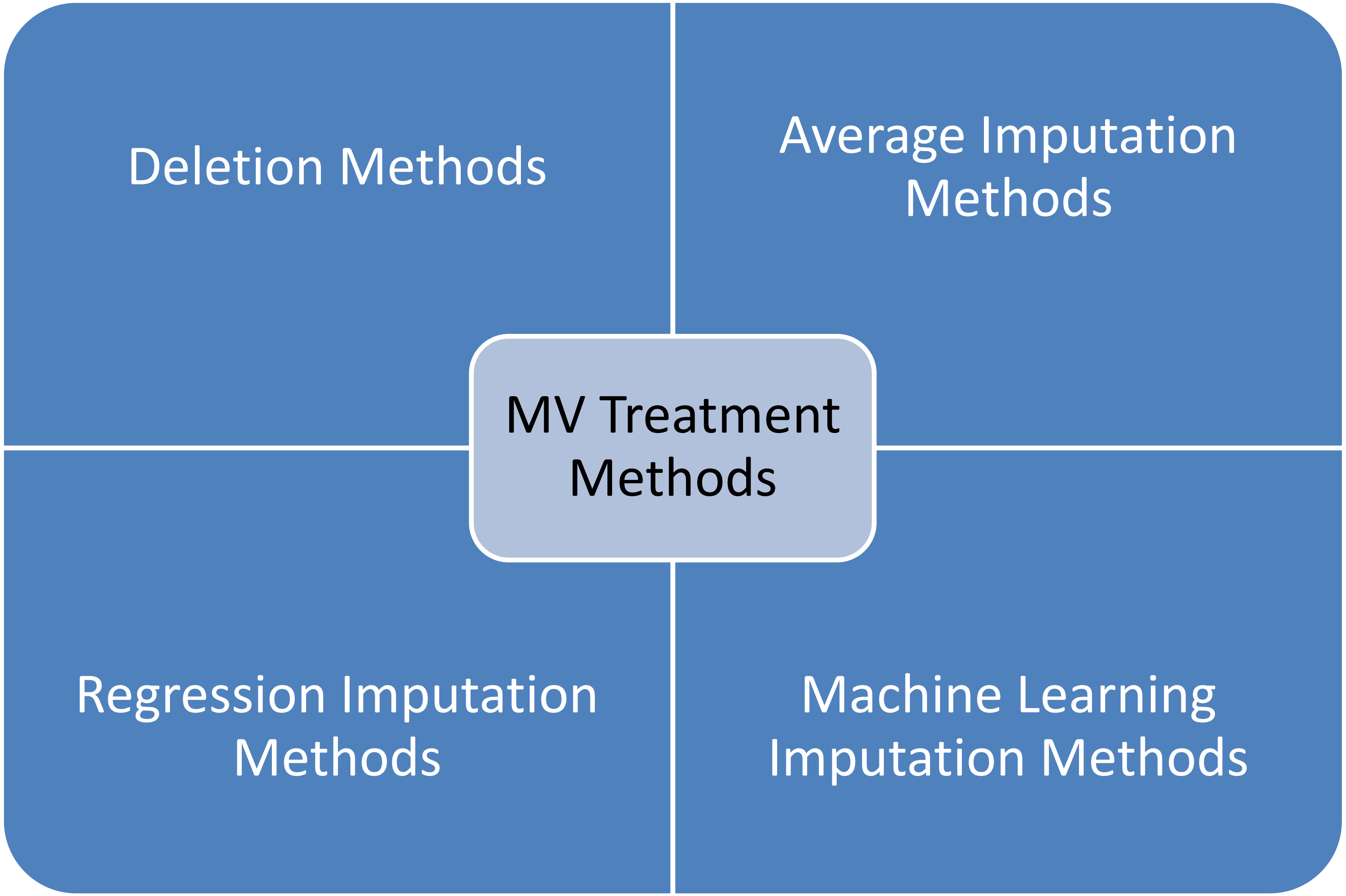
Data Understanding

Descriptive Statistics and Visualization

Introduction

Foundations

Alternative Treatments of MV



Deletion Methods

Average Imputation Methods

MV Treatment Methods

Regression Imputation Methods

Machine Learning Imputation Methods

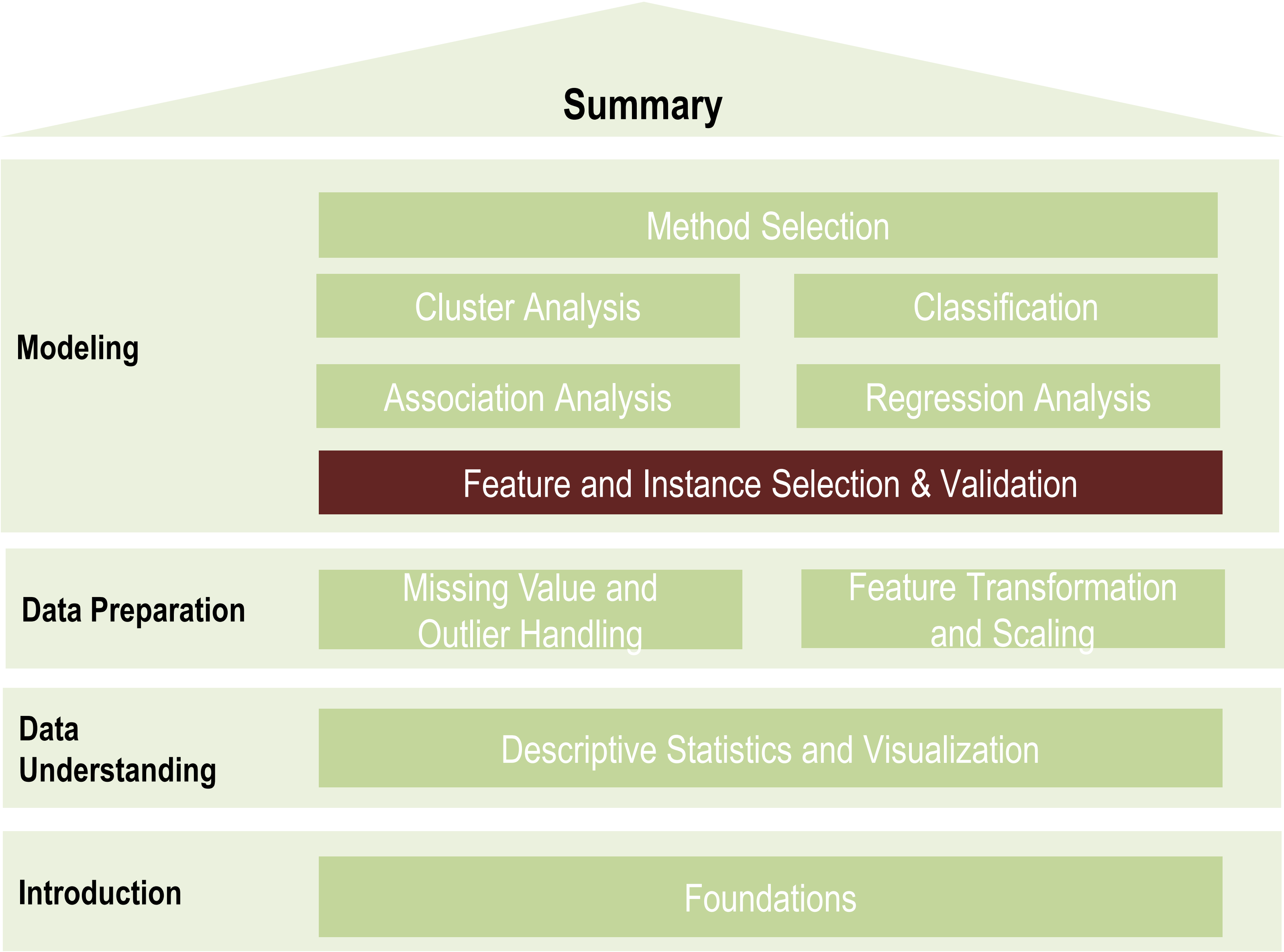
© Prof. Dr. Hendrik Meth

Outliers

- Outlier: Data object which is significantly different from other objects in data set
- Important: Definition is based on the **context** of other objects in the data set
- Example: A car with > 800 hp (horse power) will be an outlier in a database for used consumer cars, but a standard value in a database with formula 1 cars

Picture source: Pixabay

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

Data Understanding

Descriptive Statistics and Visualization

Introduction

Foundations

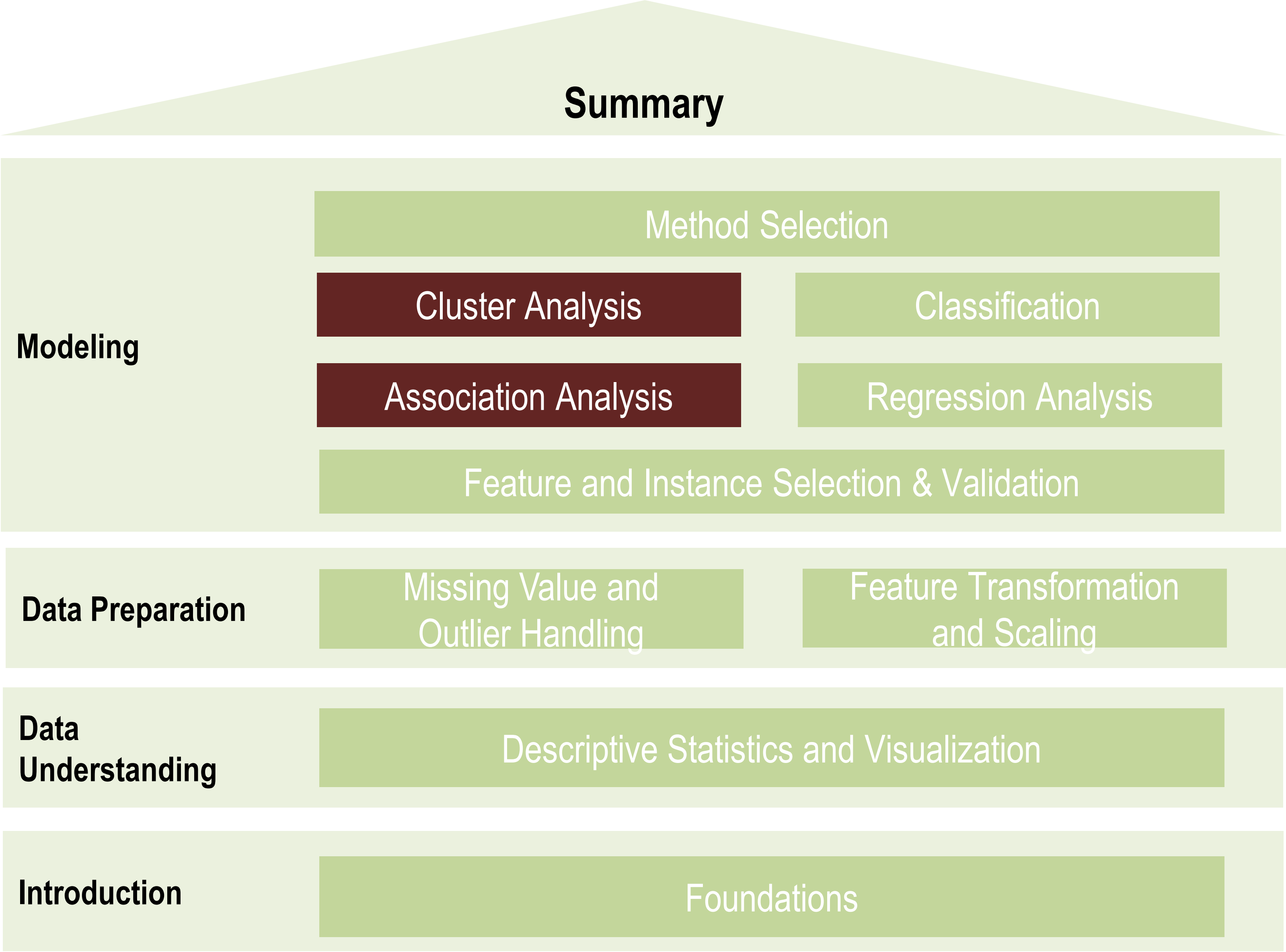
Motivation for Feature Selection

- Optimize performance of the data science algorithm, especially
  - select variables that are strongly correlated to dependent variable / label to be predicted
  - remove independent variables that are strongly correlated to each other (a requirement of many data science methods to work properly)
- Makes it easier for the analyst to interpret the outcome of the modeling

Data Reduction

- Data reduction consists of removing or grouping data
- Different types
  - **dimensionality reduction** focuses at reduction of attributes / features
  - **data sampling** focuses at reduction of examples / instances
- Data reduction aims to produce the same (or almost the same) outcome with reduced data and therefore more efficient processing

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

Data Understanding

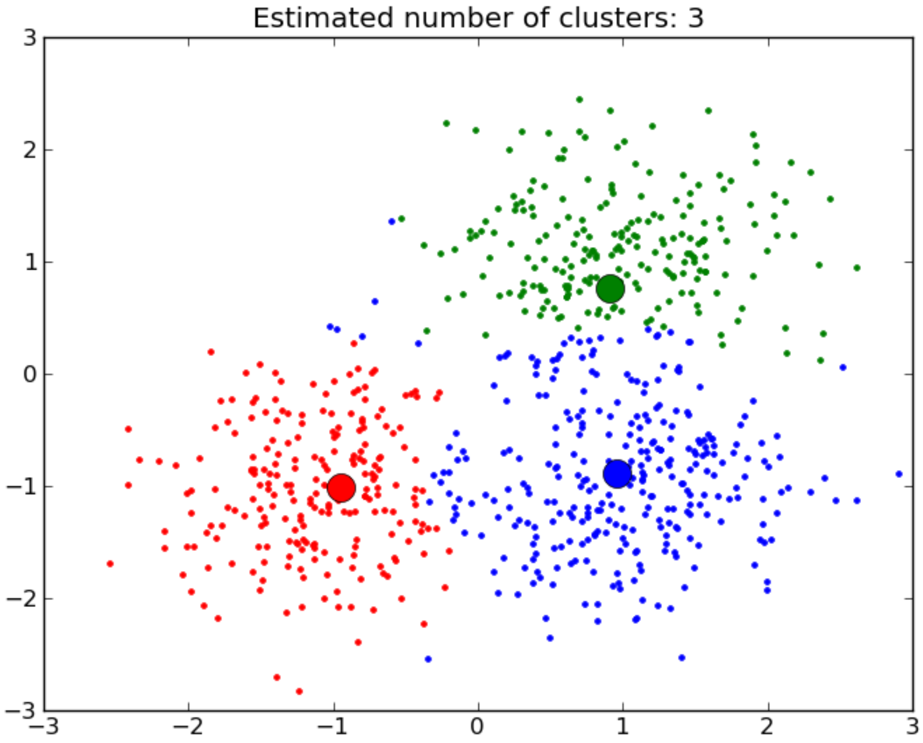
Descriptive Statistics and Visualization

Introduction

Foundations

Clustering

- Goal: Find groups of objects (=clusters)
- Pre-Requisites:
  - Set of objects / data points
  - Similarity measure to compare objects
- Conditions:
  - Objects within one cluster are similar to each other
  - Objects in different clusters are different from each other



Graph: http://scikit-learn.sourceforge.net/0.5/_images/plot_mean_shift.png

Clustering Application: Market Segmentation

- Cluster Analysis is a very useful method in market segmentation
- Market segmentation is based on the notion that
  - Customers in one market segment are similar to each other based on a given set of characteristics
  - Customers in different market segment aren’t
- Based on this segmentation the marketing mix (product, price, place, promotion) can be individually tailored to each segment

Association Rules

- Association rules describe relationships between attributes appearing together in transactions.
- Typical application areas:
  - Retailers
  - Tourism
  - eCommerce platforms (e.g. Amazon, ebay)

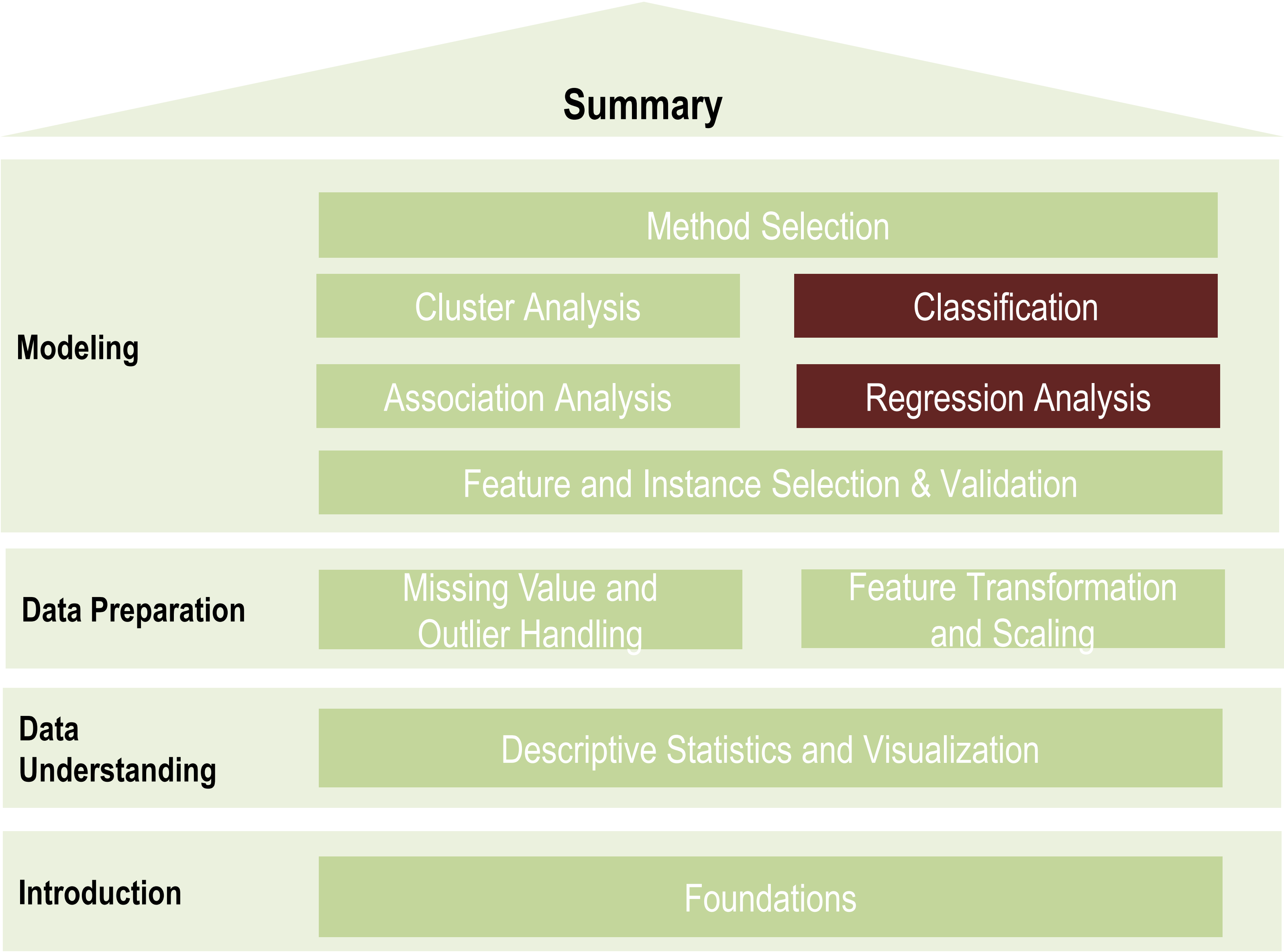
Transaction	Items
t1	Juice, coke, beer
t2	Juice, coke, wine
t3	Juice, water
t4	coke, beer, Juice
t5	Juice, coke, beer, wine
t6	water

Frequent Itemsets
Coke, Beer
Beer, Juice
Coke, Juice
Coke, Beer, Juice

Association Rules Application – Market Basket Analysis

- Typical questions to be answered by association rules:
  - Which products are often bought together?
  - What do customers buy who are similar to a certain customer?

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

Data Understanding

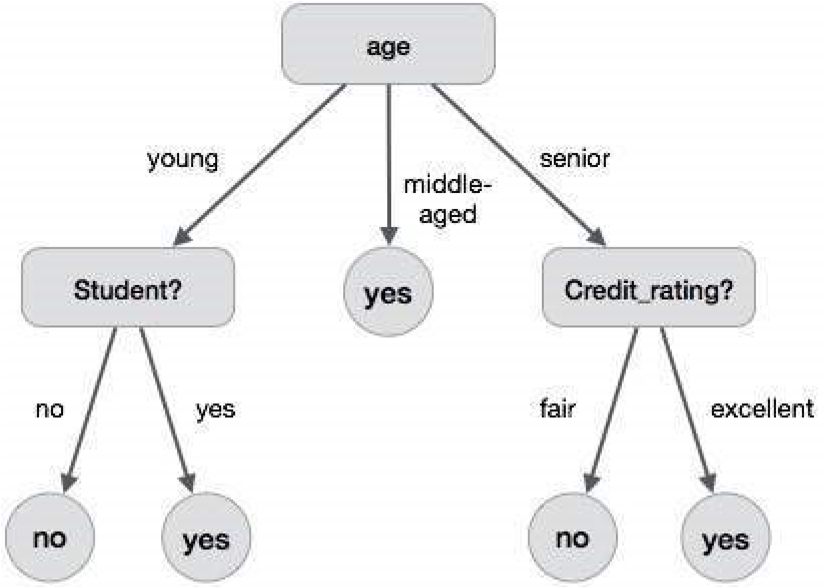
Descriptive Statistics and Visualization

Introduction

Foundations

Classification - Introduction

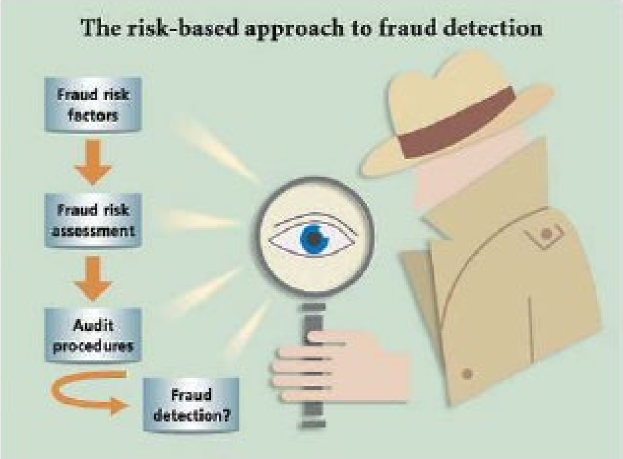
- Goal: Sort data records into two or more distinct *classes*
  - *Spam Mail / No Spam Mail*
  - *Potential Buyer / No Potential Buyer*
  - *Rainy Day / Sunny Day / Cloudy Day*
- Classification uses a training data set of already labeled records to „learn“ which records belong to a specific class
- Training data consists of independent variables, e.g. age of a customer and the class variable e.g. *potential buyer (yes/no)*
- Example: Decision tree classifier for a purchasing decision

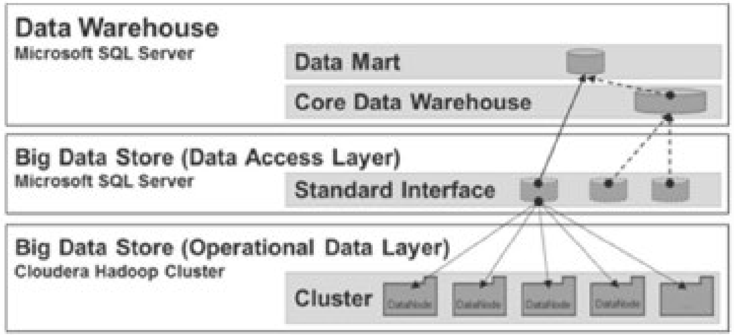


Graph: www.tutorialspoint.com

Classification Application: Fraud Detection

- Fraud is a major contributor to loss in this insurance industry
- In former times insurance companies often relied on random sample to detect fraud
- Decision Trees can be used to predict if a transaction is regular or fraud
- Technological challenge: Integrate external and internal data sources and provide results in an acceptable time frame



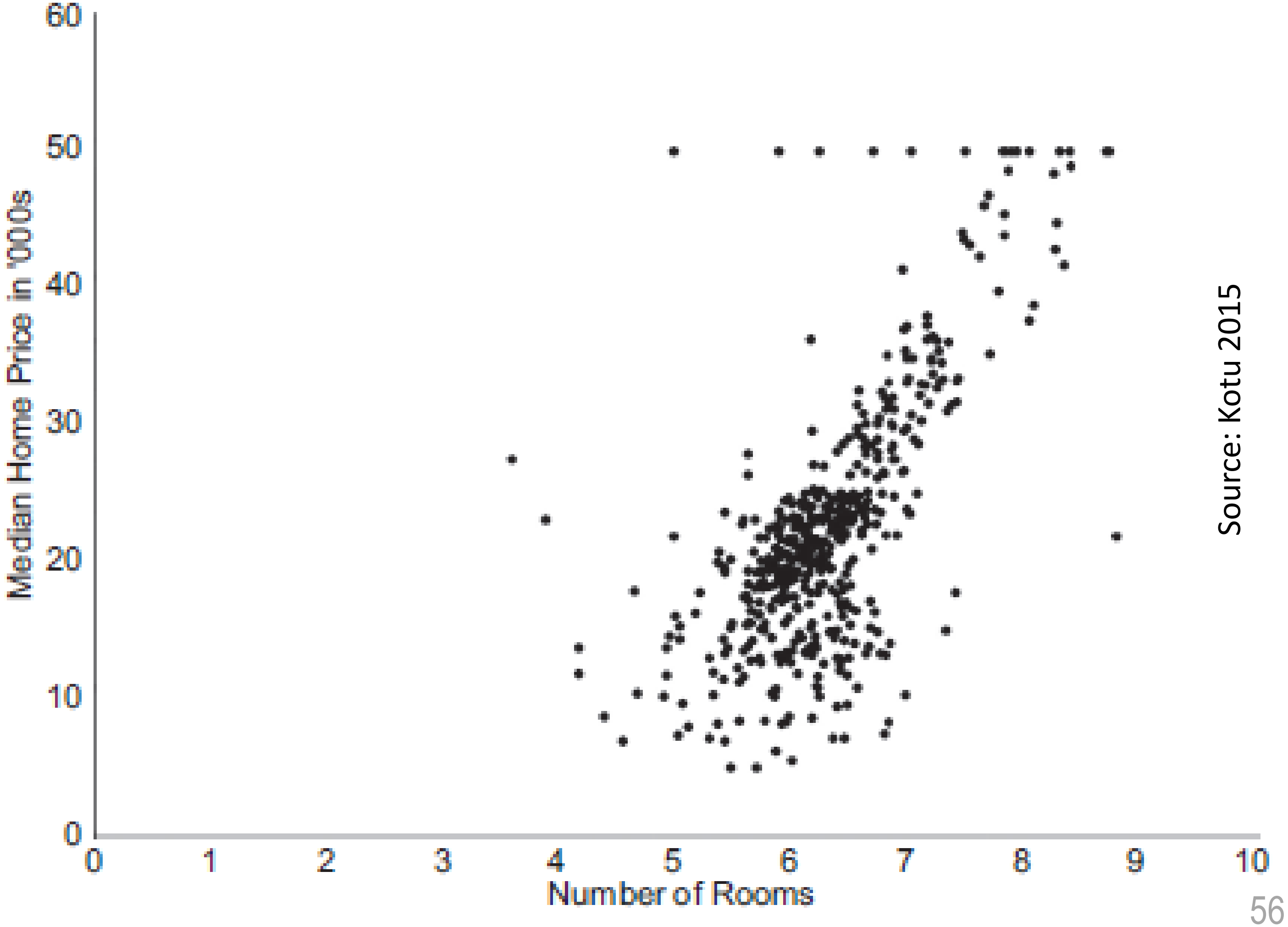


Regression Analysis

- Purpose: Identify a function that explains and predicts the value of the output variable when given the values of the input variables
- Types of regression
  - Linear regression: Numeric prediction
  - Logistic regression: Class prediction

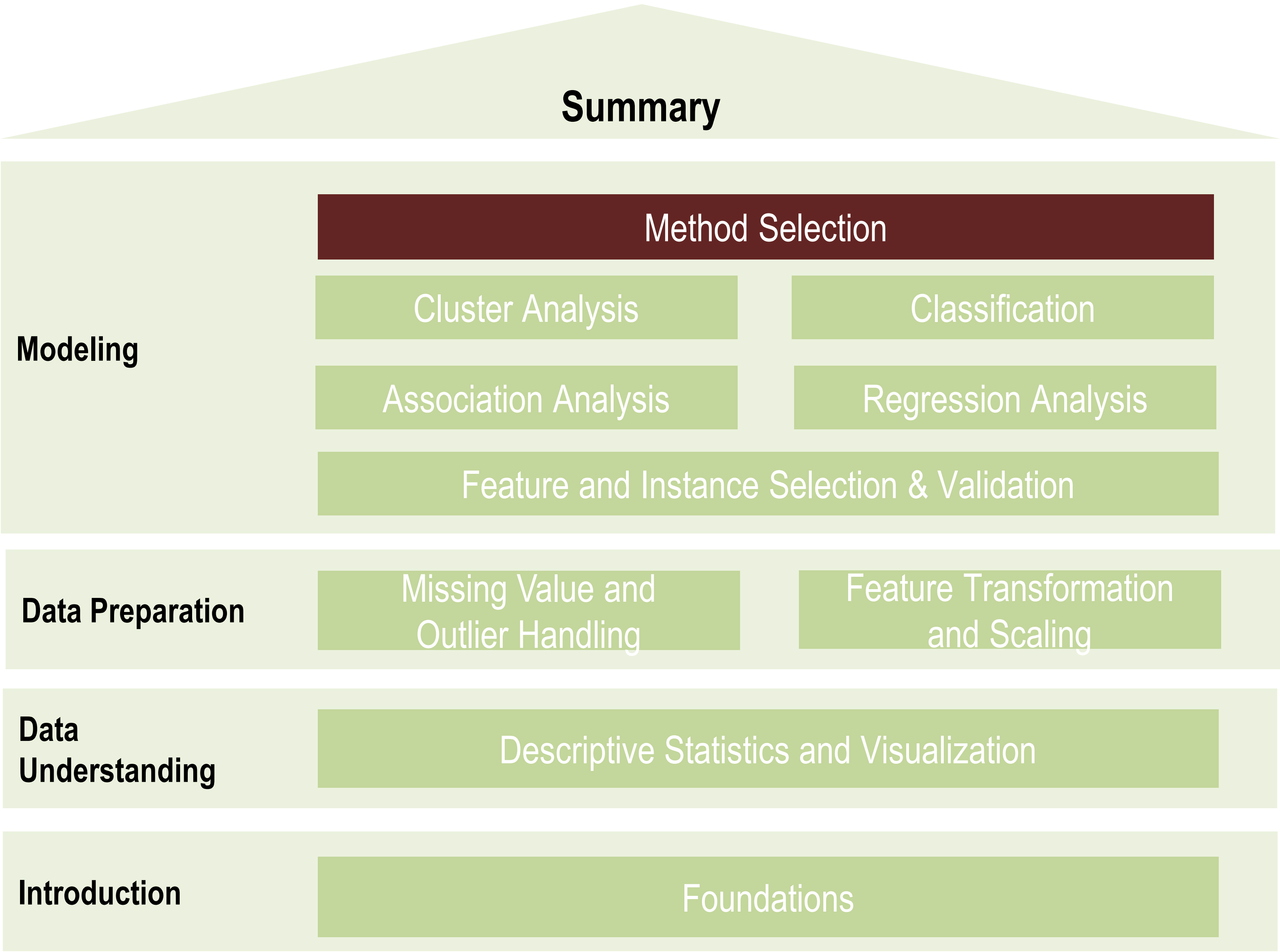
Regression Application: Demand forecasting

- Real estate: Predict real estate price based on living area, number of bedrooms, district ranking etc.



Source: Kotu 2015

Lectures: Overview



Summary

Modeling

- Method Selection
- Cluster Analysis
- Classification
- Association Analysis
- Regression Analysis
- Feature and Instance Selection & Validation

Data Preparation

- Missing Value and Outlier Handling
- Feature Transformation and Scaling

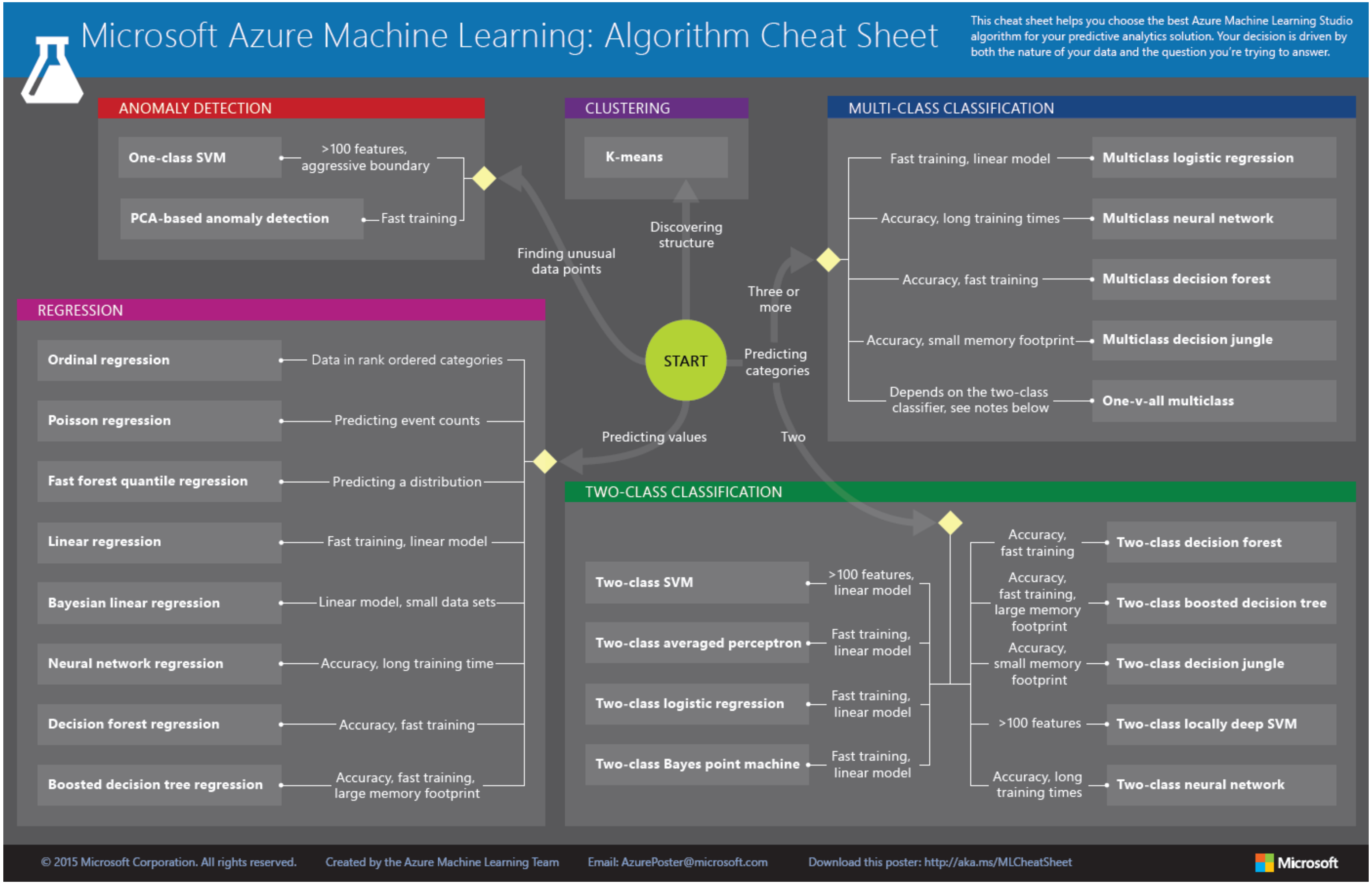
Data Understanding

- Descriptive Statistics and Visualization

Introduction

- Foundations

Method Selection



Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

- Finding unusual data points → ANOMALY DETECTION
- Discovering structure → CLUSTERING
- Predicting categories → Three or more → MULTI-CLASS CLASSIFICATION
- Predicting categories → Two → TWO-CLASS CLASSIFICATION
- Predicting values → REGRESSION

ANOMALY DETECTION

- >100 features, aggressive boundary → One-class SVM
- Fast training → PCA-based anomaly detection

CLUSTERING

- K-means

MULTI-CLASS CLASSIFICATION

- Fast training, linear model → Multiclass logistic regression
- Accuracy, long training times → Multiclass neural network
- Accuracy, fast training → Multiclass decision forest
- Accuracy, small memory footprint → Multiclass decision jungle
- Depends on the two-class classifier, see notes below → One-v-all multiclass

REGRESSION

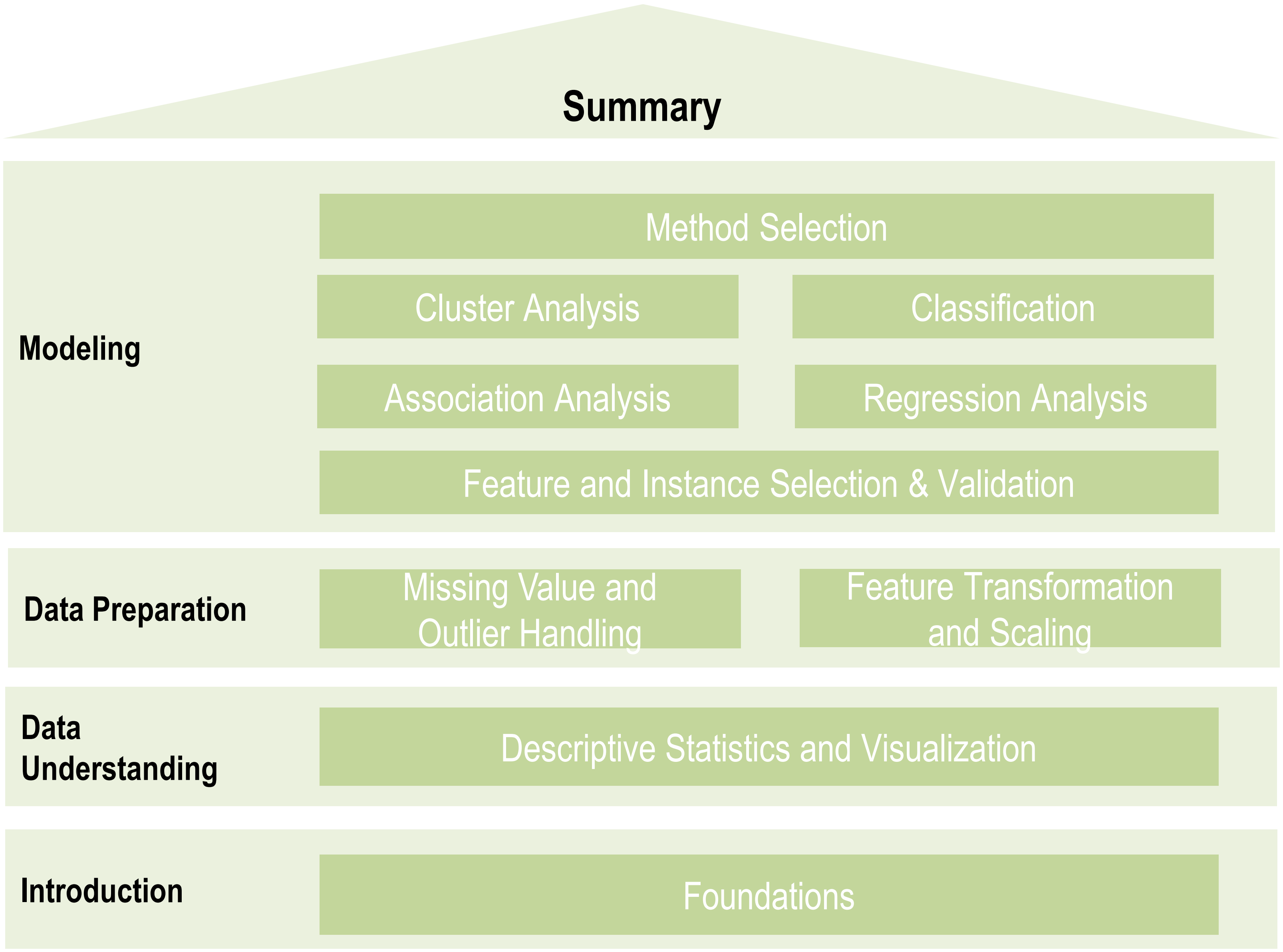
- Data in rank ordered categories → Ordinal regression
- Predicting event counts → Poisson regression
- Predicting a distribution → Fast forest quantile regression
- Fast training, linear model → Linear regression
- Linear model, small data sets → Bayesian linear regression
- Accuracy, long training time → Neural network regression
- Accuracy, fast training → Decision forest regression
- Accuracy, fast training, large memory footprint → Boosted decision tree regression

TWO-CLASS CLASSIFICATION

- >100 features, linear model → Two-class SVM
- Fast training, linear model → Two-class averaged perceptron
- Fast training, linear model → Two-class logistic regression
- Fast training, linear model → Two-class Bayes point machine
- Accuracy, fast training → Two-class decision forest
- Accuracy, fast training, large memory footprint → Two-class boosted decision tree
- Accuracy, small memory footprint → Two-class decision jungle
- >100 features → Two-class locally deep SVM
- Accuracy, long training times → Two-class neural network

Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet Microsoft

Lectures: Overview



Summary

Modeling

Method Selection

Cluster Analysis

Classification

Association Analysis

Regression Analysis

Feature and Instance Selection & Validation

Data Preparation

Missing Value and Outlier Handling

Feature Transformation and Scaling

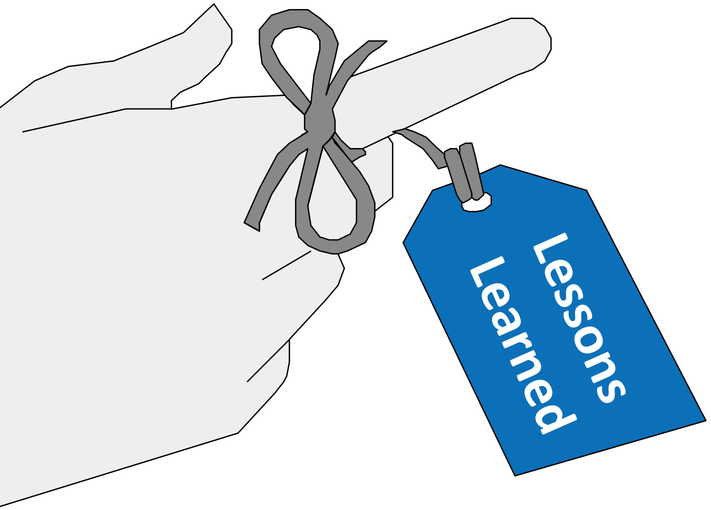
Data Understanding

Descriptive Statistics and Visualization

Introduction

Foundations

Today‘s session in review



Today you learned…

- …how to define Data Science
- …about different types of Data Science
- …how to characterize the Data Science process
- …about some examples of Data Science applications