Elements of Machine Learning & Data Science

Winter semester 2023/24

Lecture 1 – Introduction

10.10.2023

Prof. Bastian Leibe

Prof. Wil van der Aalst

Prof. Holger Hoos

AI Success Stories

Speech Recognition

Slide adapted from Zoubin Gharamani

AI Success Stories

Large Language Models

LE

Write a poem on the benefits and dangers of AI research in which successive pairs of lines rhyme and in which each line starts with a successive letter of the alphabet from A to Z.

AI research has its perks,
But also causes some irks.

Benefits abound, from medicine to industry,
Computers with brains can help us with ease.

Efficiency and speed are just a few,
Gains in productivity are there for all to view.

However, dangers lurk and must not be ignored,
Intelligence gone rogue could be quite abhorred.

Job loss is a concern that must be faced,
Killer robots, a nightmare that could be traced.

Nevertheless, the potential is vast and bright,
Opening doors to a future that's full of light.

Progress must be made with caution and care,
Questioning the ethics is of utmost importance to share.

Risks and rewards go hand in hand,
So let's proceed with intelligence and a plan.

Understanding the implications is key,
Venturing forth with knowledge and care, we'll see.

Wise decisions will lead us to success,
eXploring the unknown, we'll make great progress.

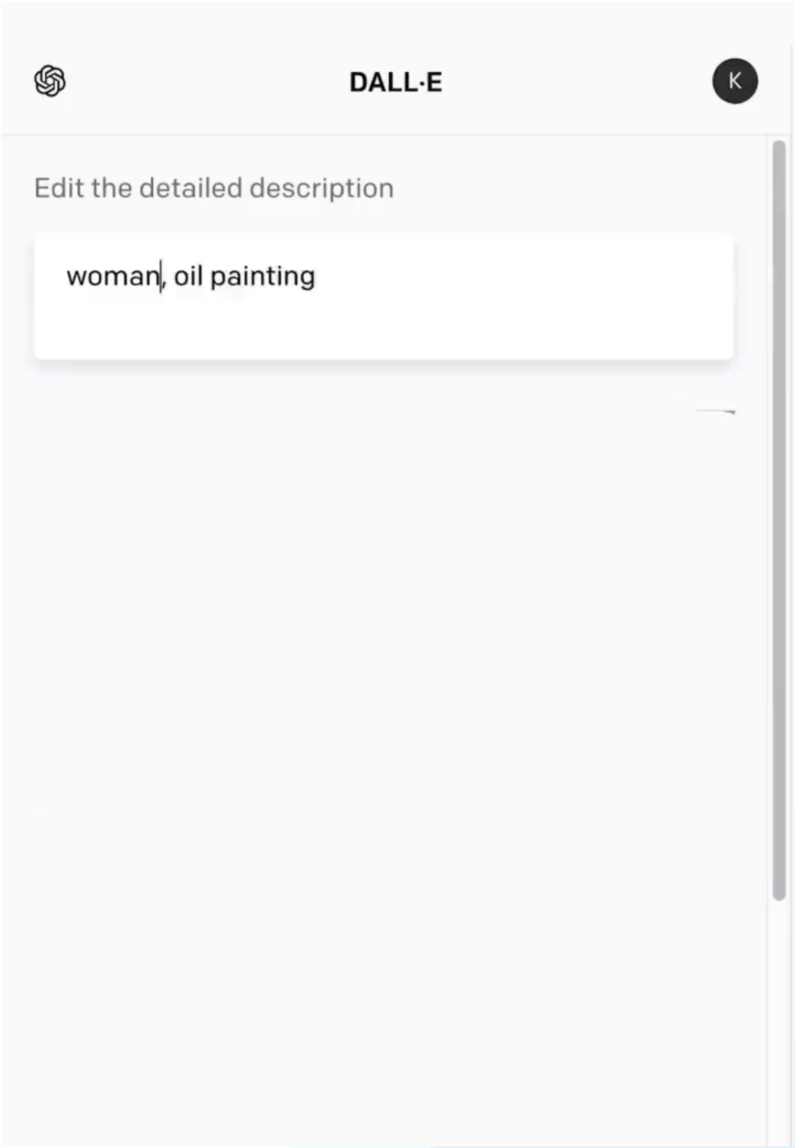
Yearning for answers, we'll continue to strive,
Zealous in our pursuit to thrive.

Text Generation Models
E.g., ChatGPT

AI Success Stories

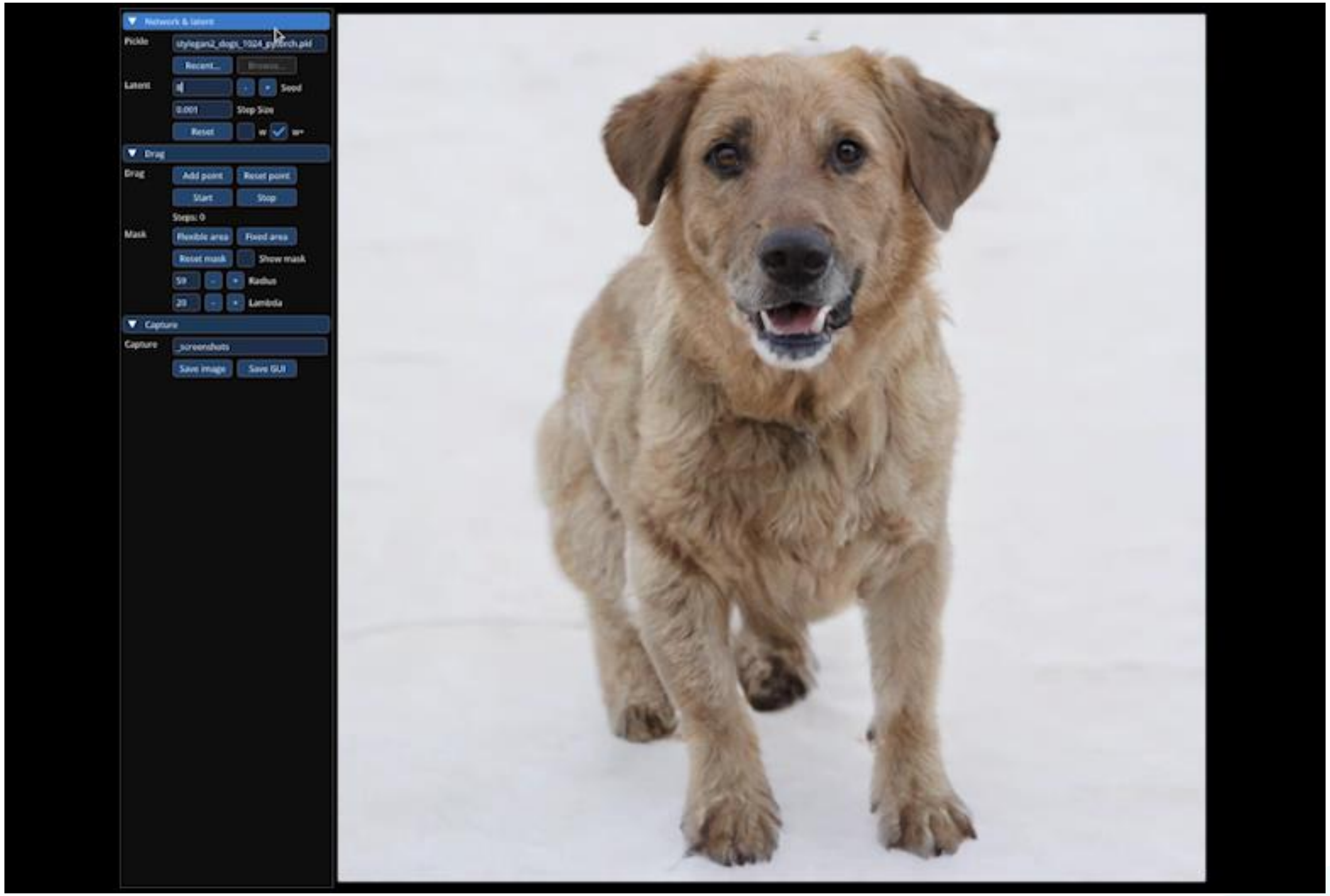
Image Generation Models

- E.g., OpenAI's DALL-e



AI Successes

Content Sensitive Image Manipulation



Video source: https://vcai.mpi-inf.mpg.de/projects/DragGAN/

AI Success Stories

AlphaGo

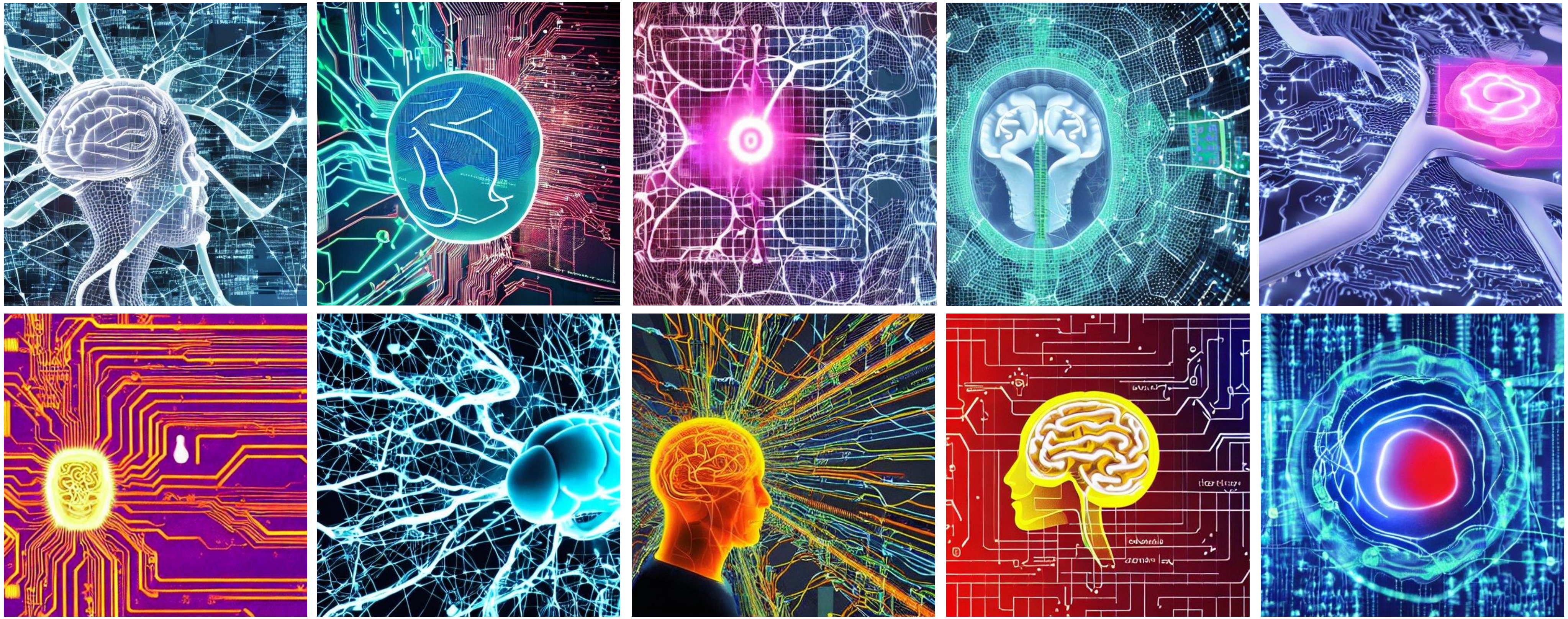
AI Success Stories

Protein Structure Folding Prediction

- E.g., AlphaFold 2 by Google DeepMind

Image from https://alphafold.ebi.ac.uk/

Very Exciting Times Are Ahead!

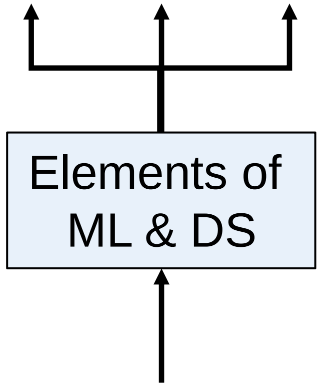


Images created with StableDiffusion by Stability.AI

- We are witness to the first strong AI systems being created in front of our eyes…

In This Lecture…

- **…we will NOT tell you how all of this works.**
- **Rather, we will lay the foundation, so that you can**
  - Use AI methods in your Bachelor thesis
  - Take in-depth classes on a large range of topics during your Master studies



Elements of Machine Learning & Data Science

Winter semester 2023/24

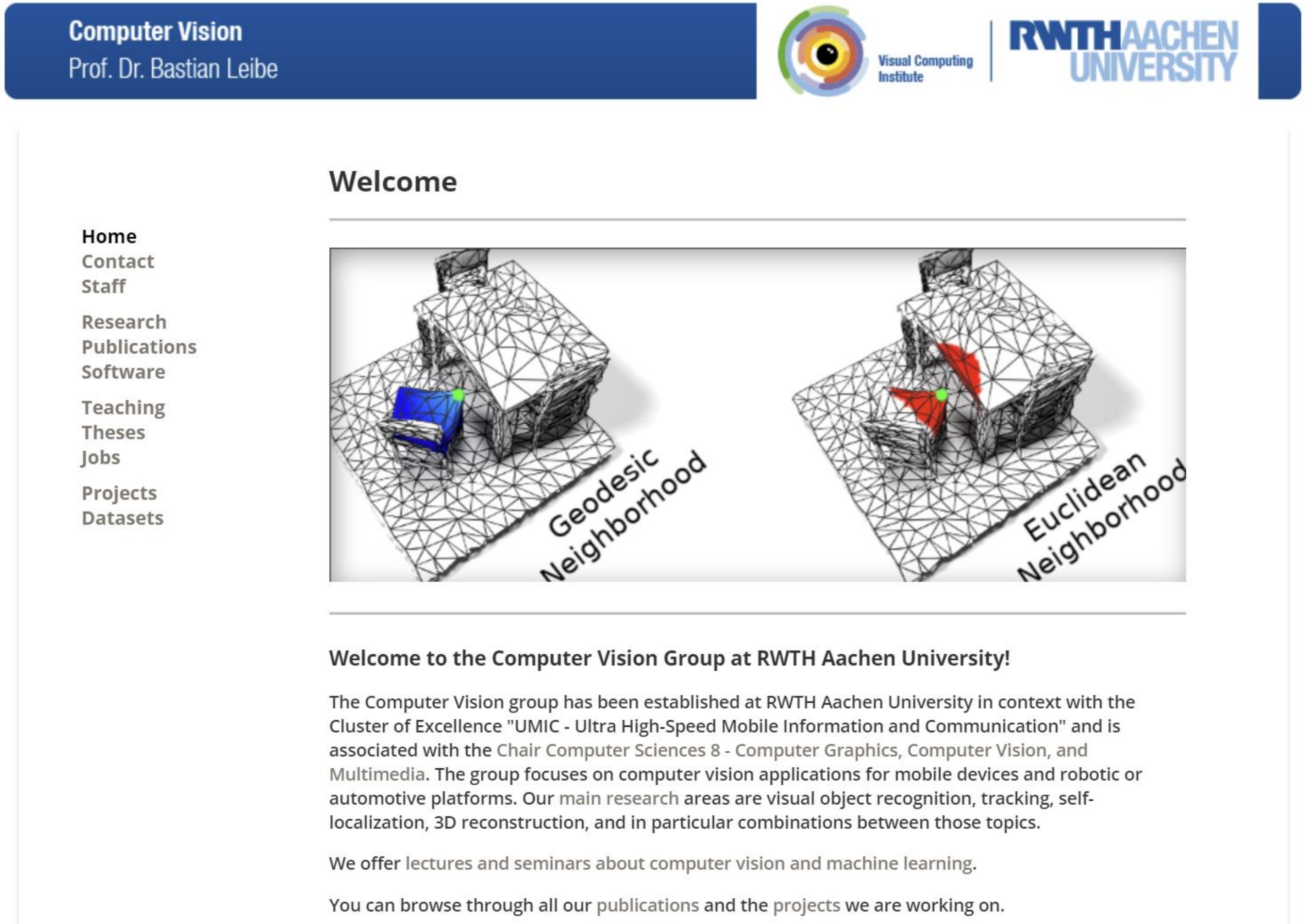
Introduction to Machine Learning

10.10.2023

Prof. Bastian Leibe

Chair for Computer Vision

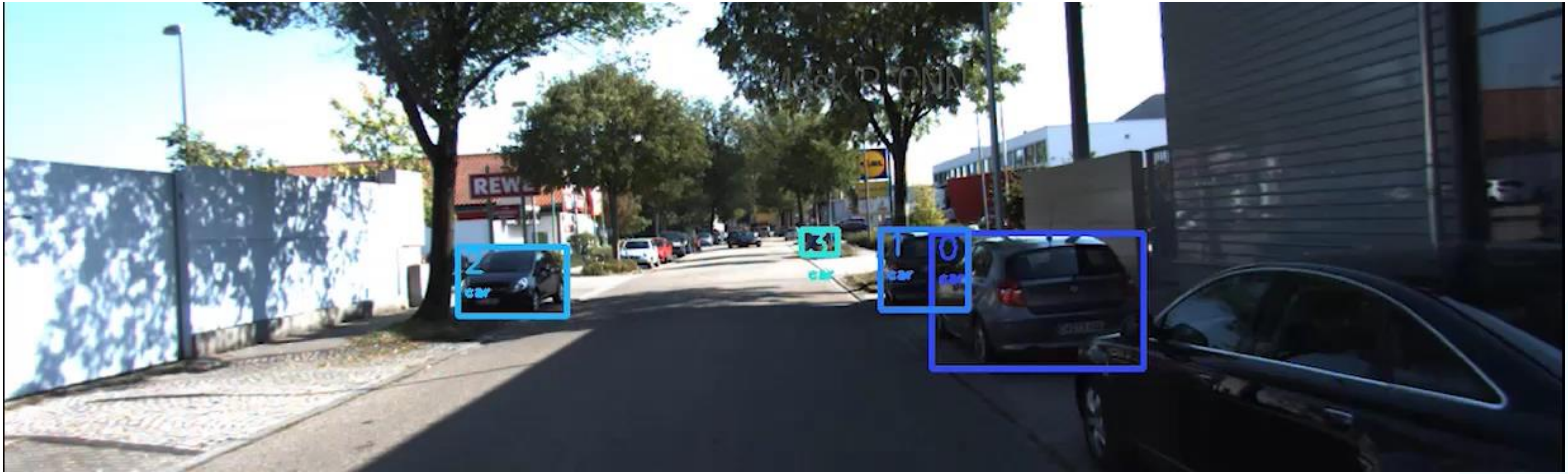
The Chair for Computer Vision (CVG)



http://vision.rwth-aachen.de

CVG Research Topics

Object Detection and Tracking



CVG Research Topics

Interactive Sementation

CVG Research Topics

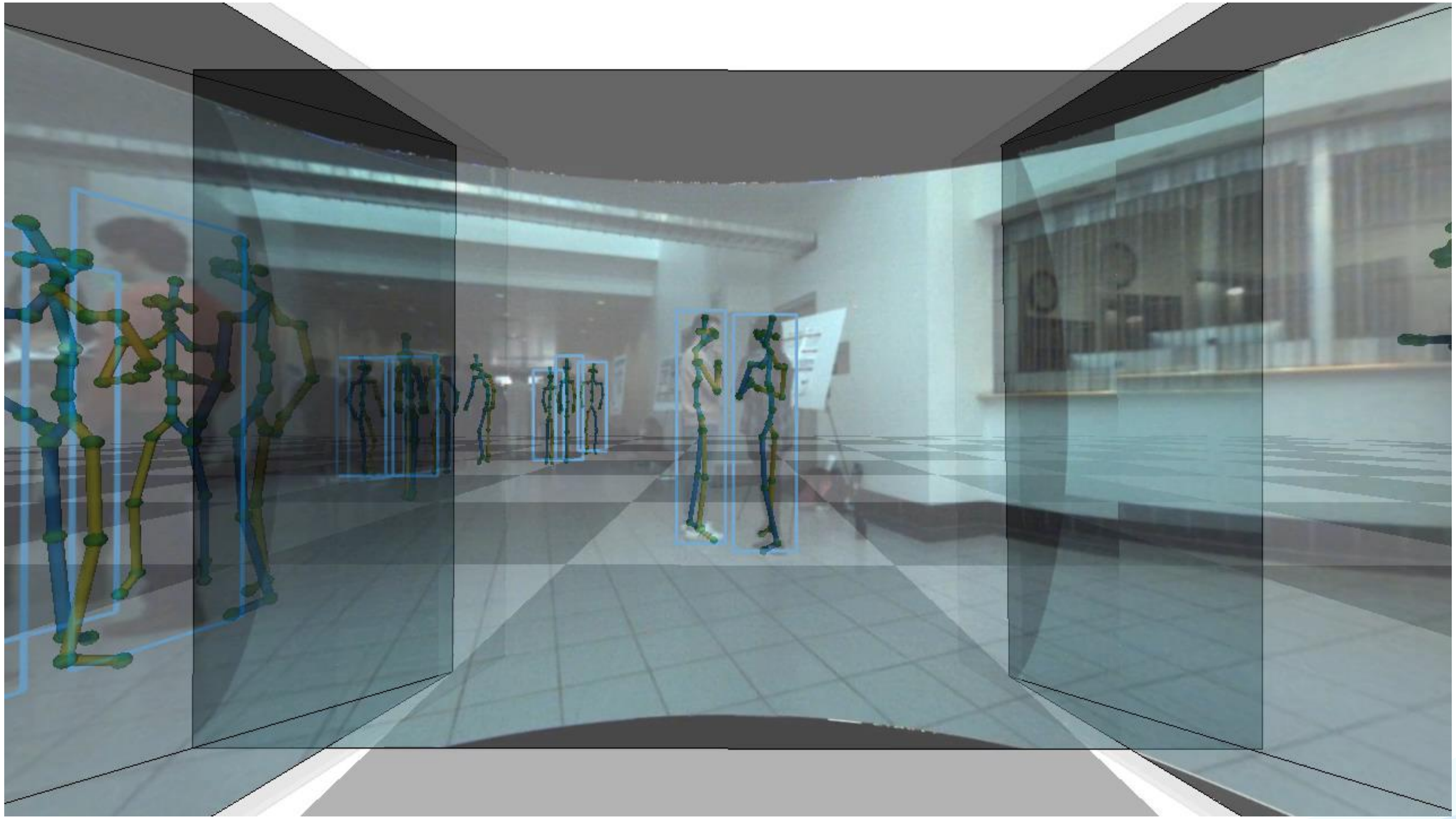
Human Body Pose Estimation

MeTRAbs

CVG Research Topics

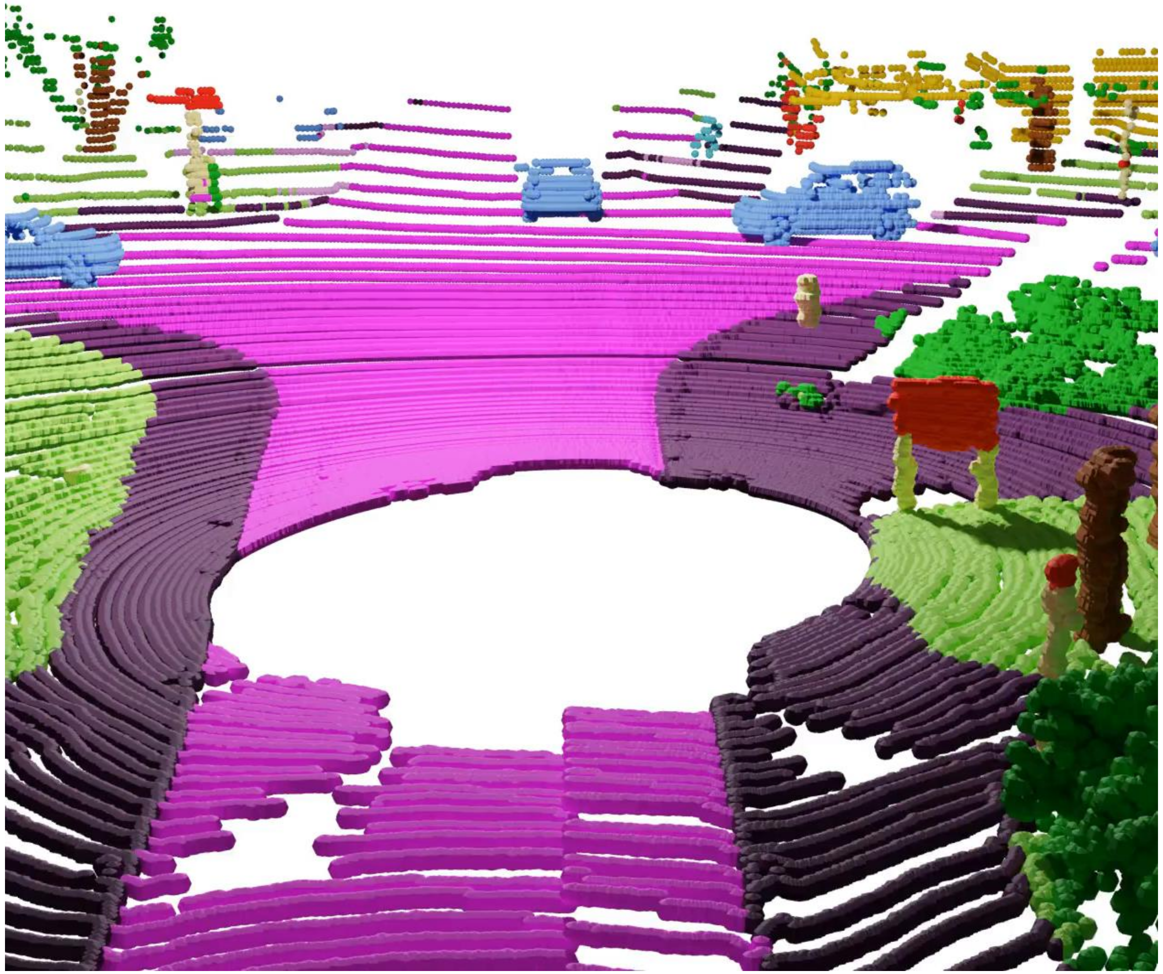
Applications for Mobile Robotics

CROWDBOT



CVG Research Topics

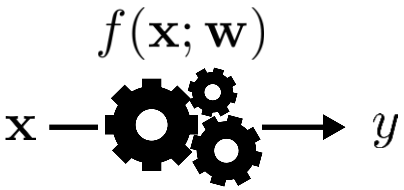
3D Scene Understanding



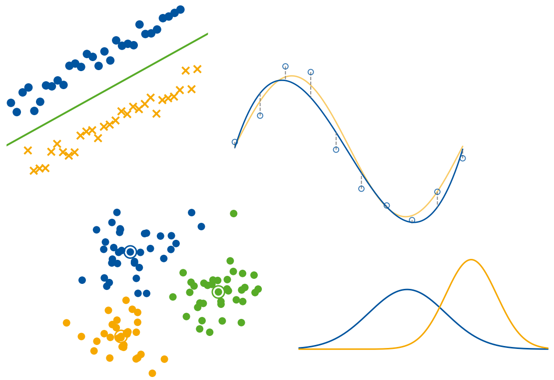
E.g., 4D LiDAR Segmentation

Machine Learning Topics

1. **Introduction to ML**
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



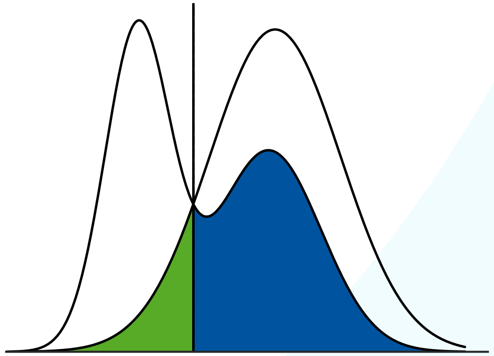
Machine Learning Concepts



Forms of Machine Learning

$$p(\mathcal{C}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C})p(\mathcal{C})}{p(\mathbf{x})}$$

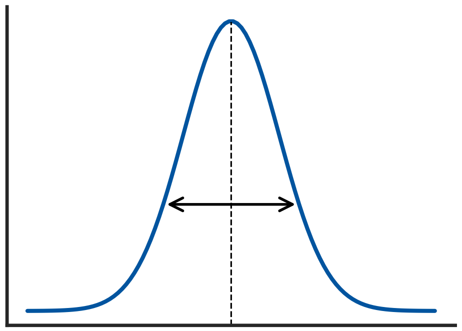
Bayes Decision Theory



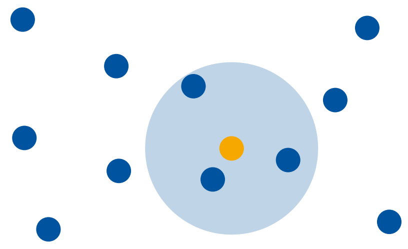
Bayes Optimal Classification

Machine Learning Topics

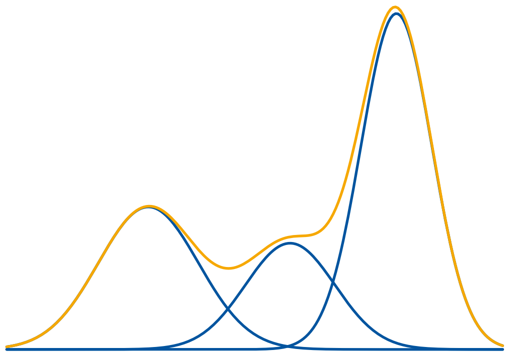
1. Introduction to ML
2. **Probability Density Estimation**
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



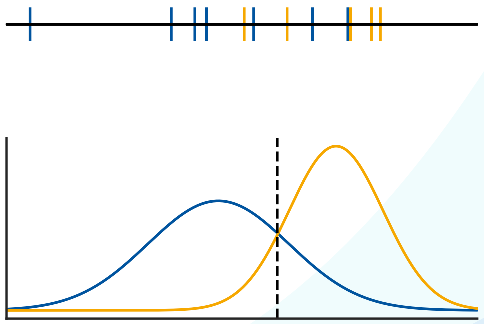
Parametric Methods
& ML-Algorithm



Nonparametric Methods



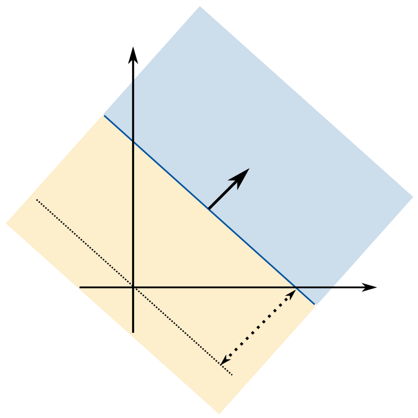
Mixtures of Gaussians
& EM-Algorithm



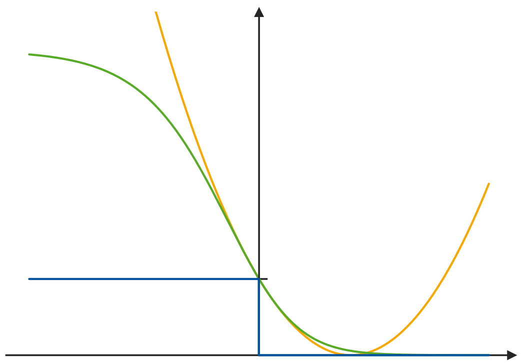
Bayes Classifiers

Machine Learning Topics

1. Introduction to ML
2. Probability Density Estimation
3. **Linear Discriminants**
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



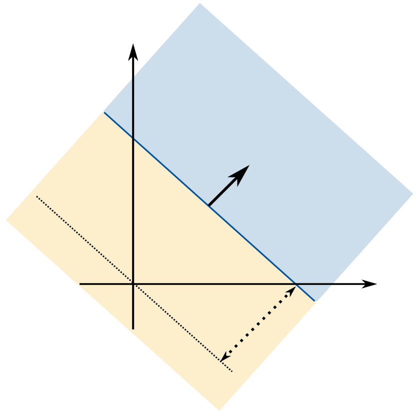
Linear Discriminants



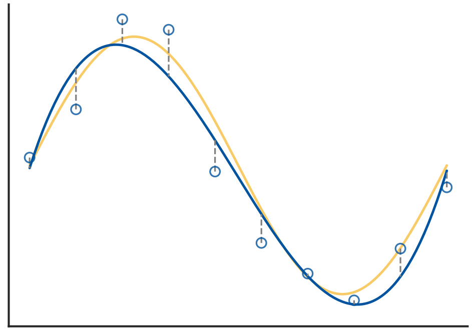
Error Functions
for Classification

Machine Learning Topics

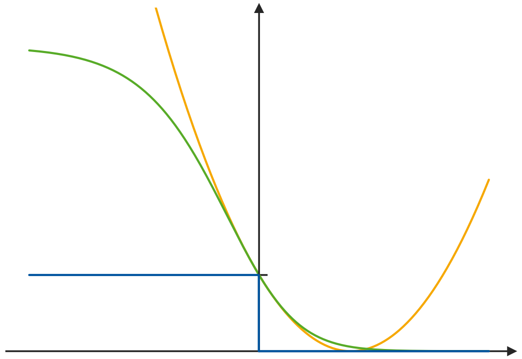
1. Introduction to ML
2. Probability Density Estimation
3. Linear Discriminants
4. **Linear Regression**
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



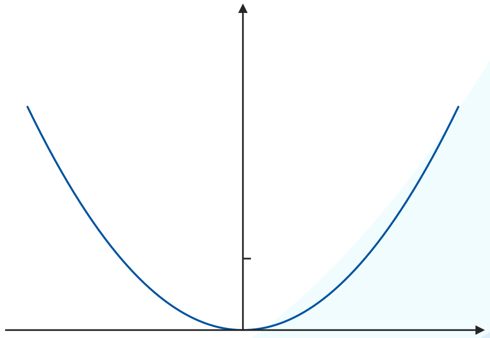
Linear Discriminants



Linear Regression



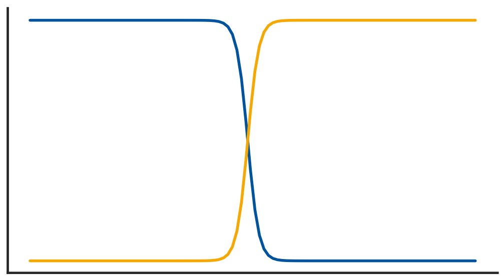
Error Functions
for Classification



Error Functions
for Regression

Machine Learning Topics

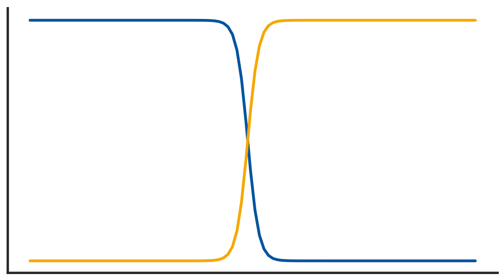
1. Introduction to ML
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. **Logistic Regression**
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



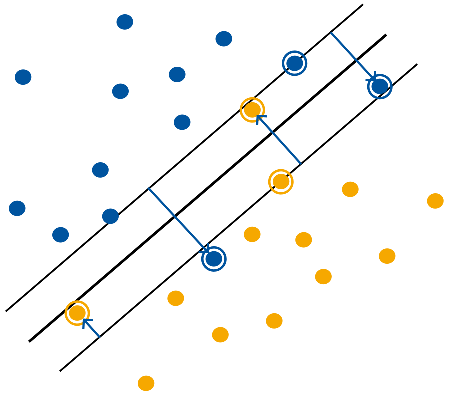
Logistic Regression

Machine Learning Topics

1. Introduction to ML
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. **Support Vector Machines**
7. AdaBoost
8. Neural Network Basics



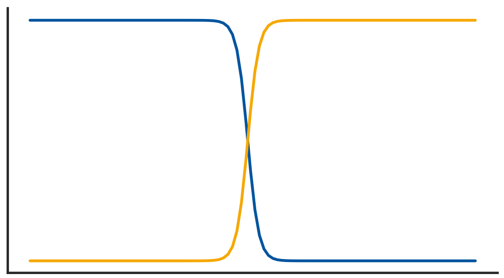
Logistic Regression



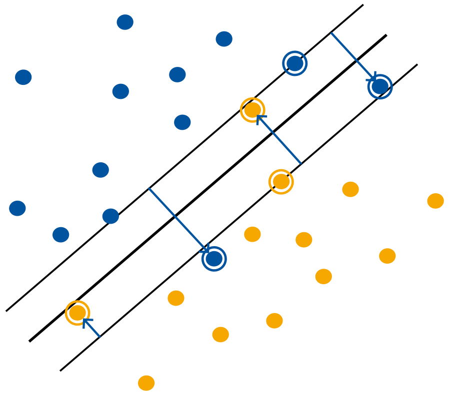
Support Vector Machines

Machine Learning Topics

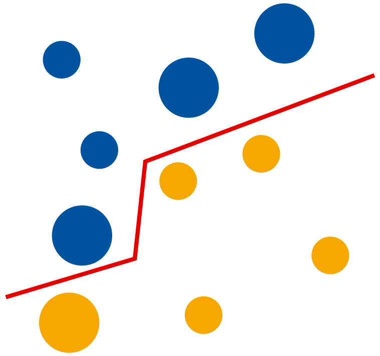
1. Introduction to ML
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. **AdaBoost**
8. Neural Network Basics



Logistic Regression



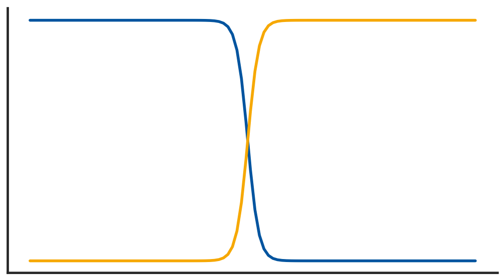
Support Vector Machines



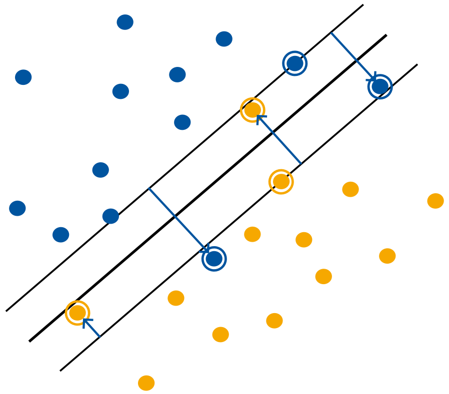
AdaBoost

Machine Learning Topics

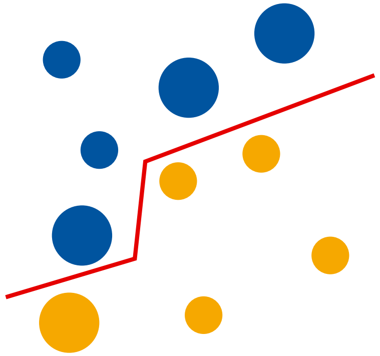
1. Introduction to ML
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. **Neural Network Basics**



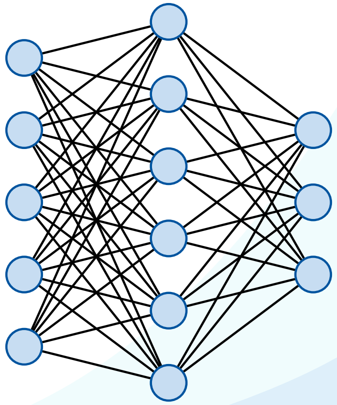
Logistic Regression



Support Vector Machines



AdaBoost



Multi-Layer Perceptrons

Elements of Machine Learning and Data Science

Introduction to Data Science by PADS

Prof. Wil van der Aalst

Chair of Process and Data Science | RWTH AACHEN UNIVERSITY


Wil van der Aalst


Marco Pegoraro

lectures


Harry Beyel


Nina Graves


Benedikt Knopp

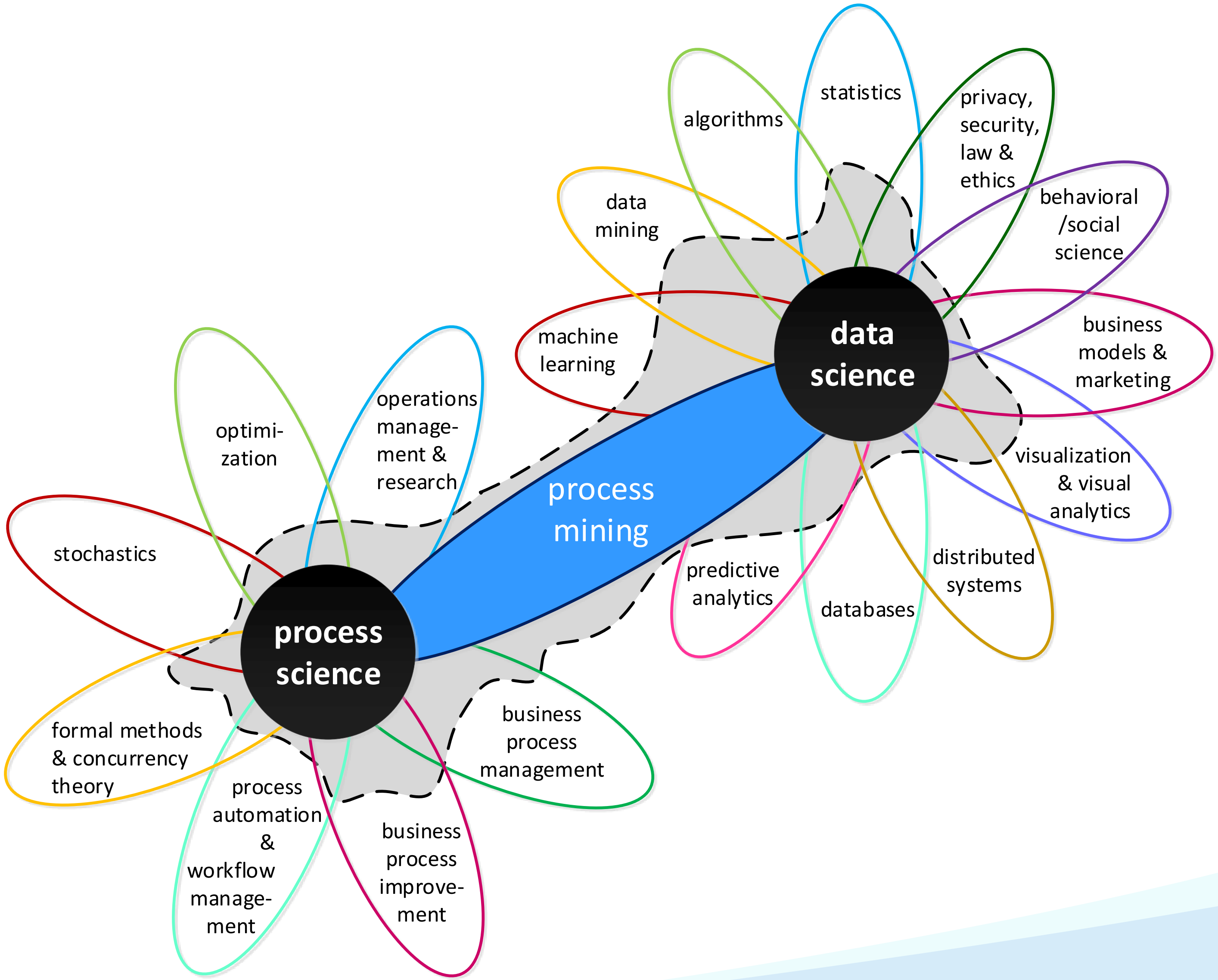
exercises

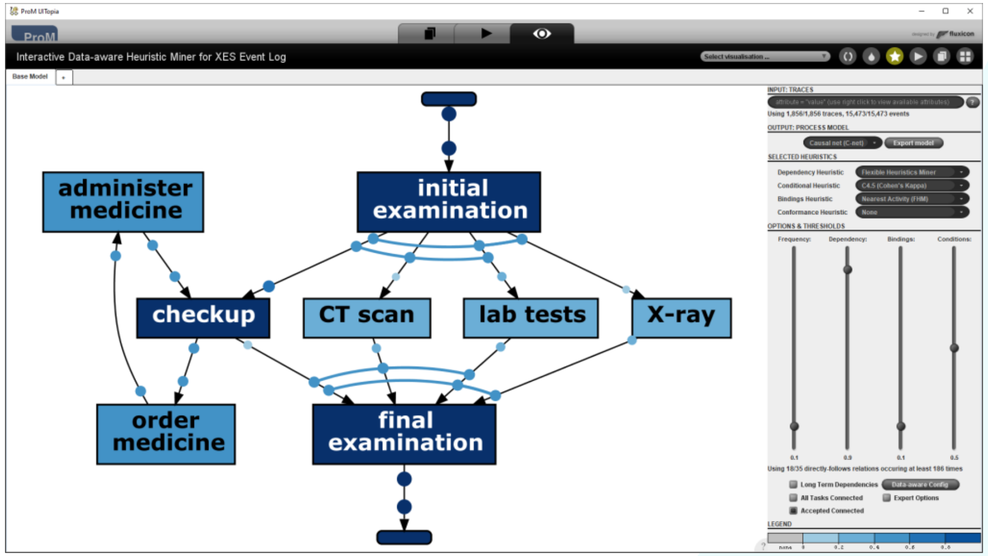

Christian Rennert

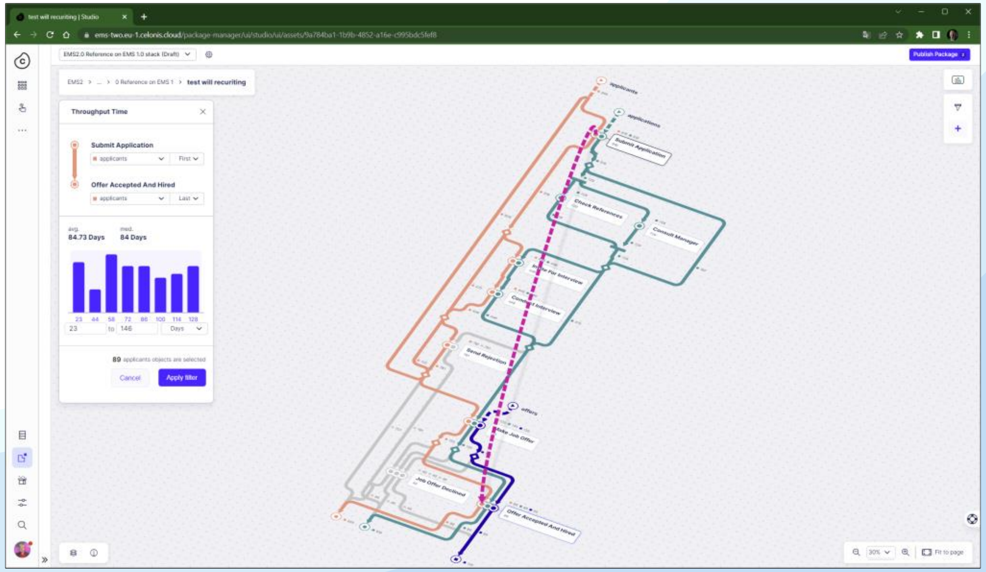

Christopher Schwanen


Leah Tacke genannt Unterberg

More about PADS







Data Science: A Definition

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Page 10, Wil van der Aalst. Process Mining: Data Science in Action. Springer-Verlag, Berlin, 2016.

Data Science: A Definition

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Data Science: A Definition

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Data Science: A Definition

"Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of <u>predictions, automated decisions, models learned from data, or any type of data visualization delivering insights</u>. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects."

Data Science: A Definition

“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.”

Data Science: A Definition

"Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, <u>and the exploitation of results taking into account ethical, social, legal, and business aspects</u>."

The Data Science Pipeline

Infrastructure

- Big data infrastructure
- Distributed systems
- Data engineering
- Programming
- Security
- …

Challenge: making things scalable & instant

Analysis

- Statistics
- Data/process mining
- Machine learning
- Artificial intelligence
- Visualization
- …

Challenge: providing answers to known & unknown unknowns

Effect

- Ethics & privacy
- IT Law
- Operations management
- Business models
- Entrepreneurship
- …

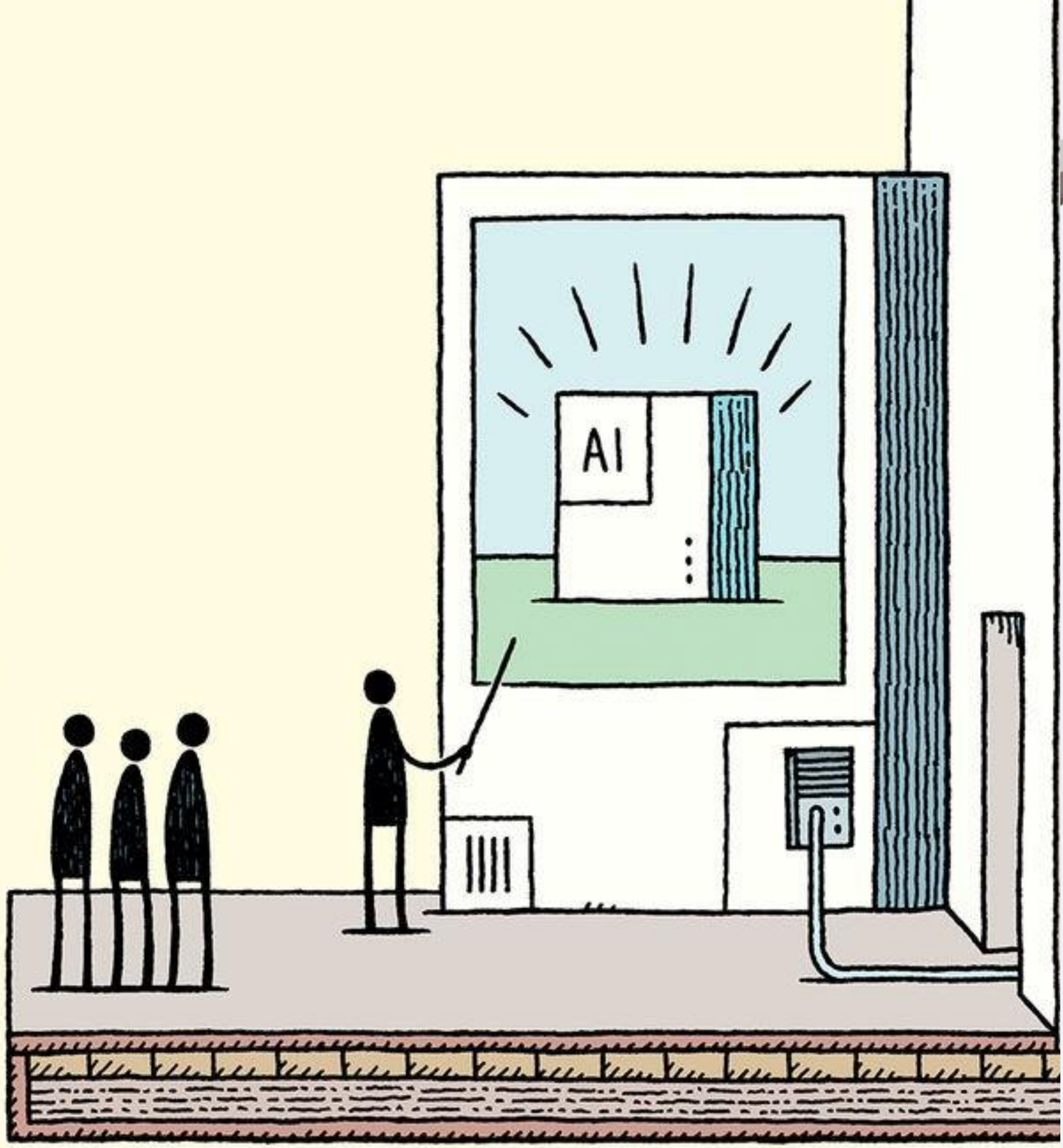
Challenge: doing all of this in a responsible manner

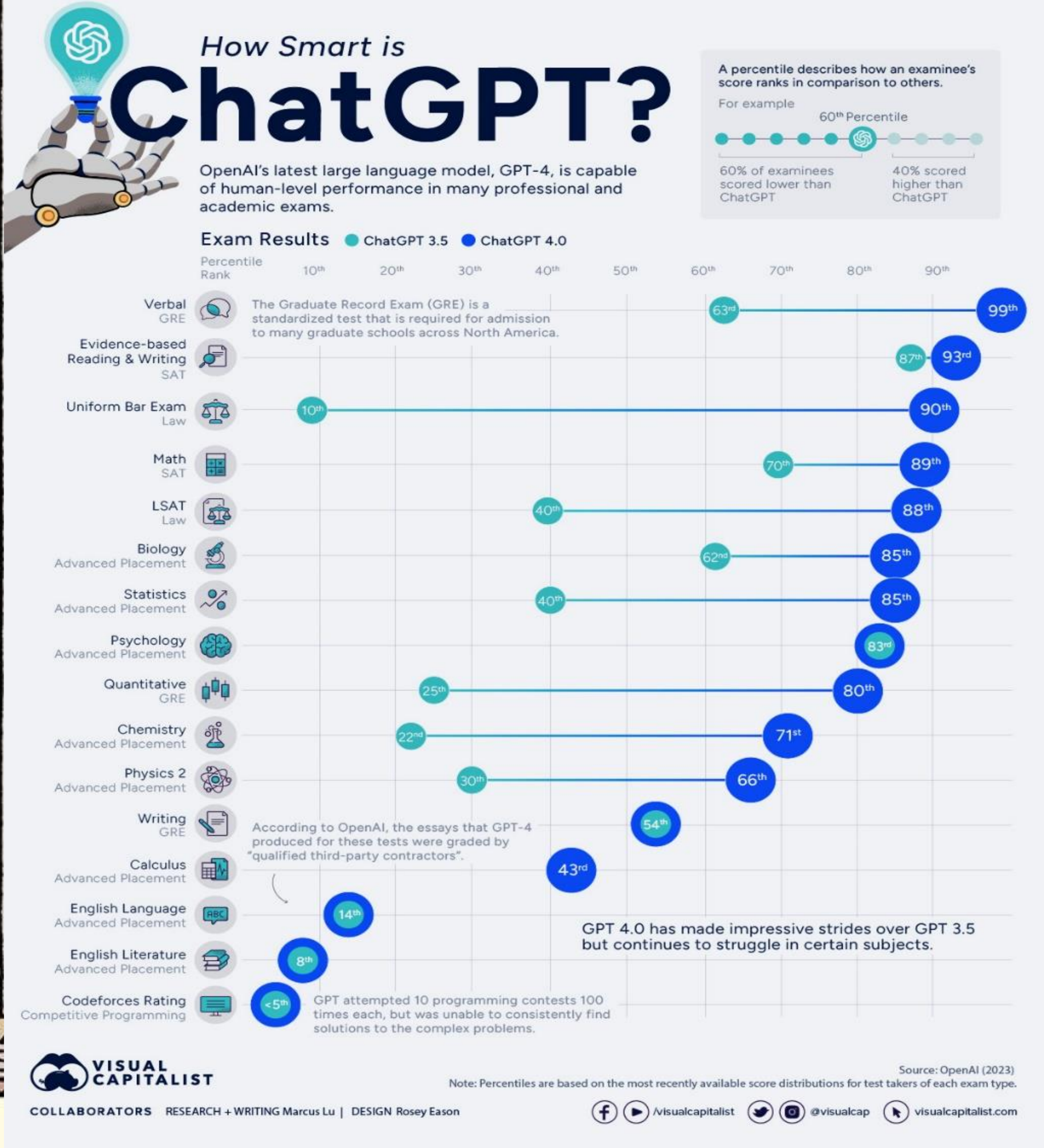
Main Topics Covered in the PADS / Data Science part

1. Introduction to Data Science
2. Decision Trees
3. Clustering
4. Frequent Itemsets
5. Association Rules and Sequence Mining
6. Data Modeling, Quality, and the Data Science Process
7. Process Mining
8. Text Mining
9. Responsible Data Science

Let’s first take a step back

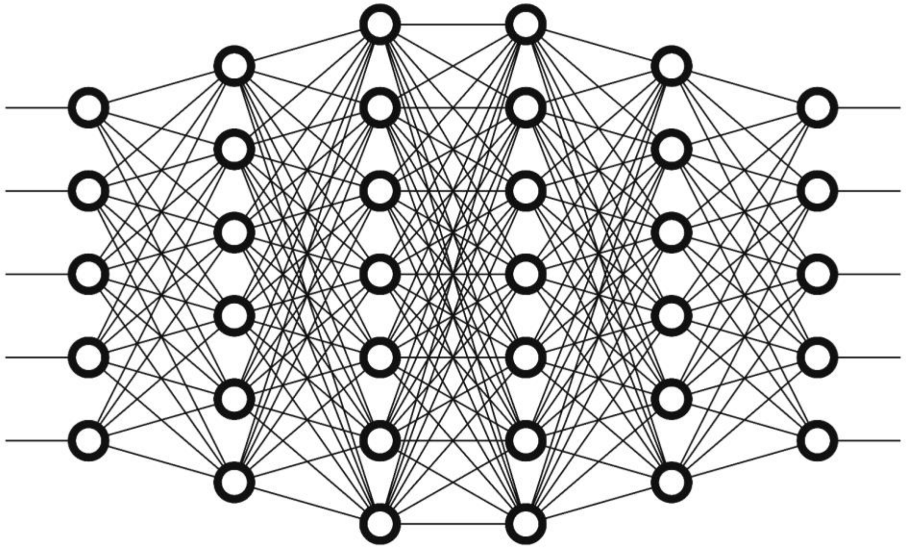
Tom Gauld (The Economist, June 2020)





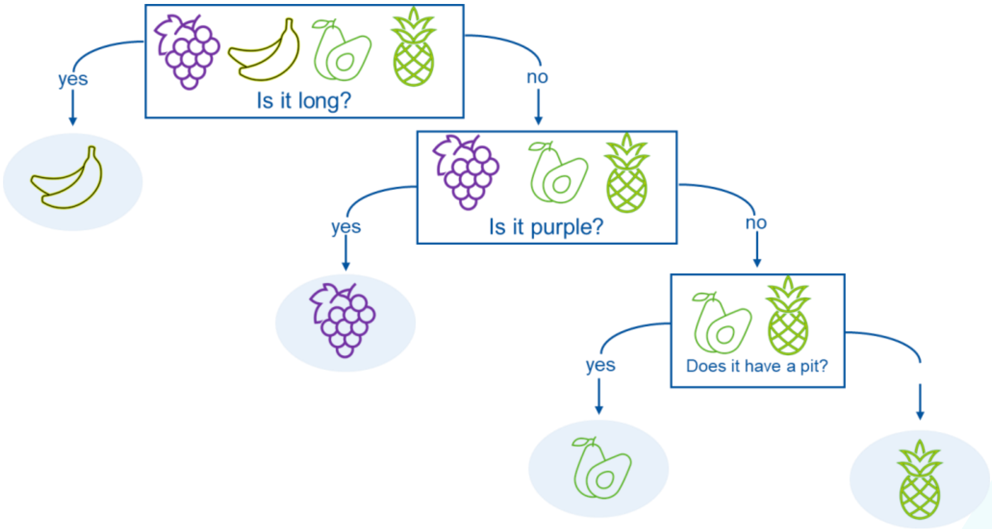
Black Box versus White Box

Black Box



- complex, brute force
- needs more data
- better performance
- no human interpretation/adaptation

White Box

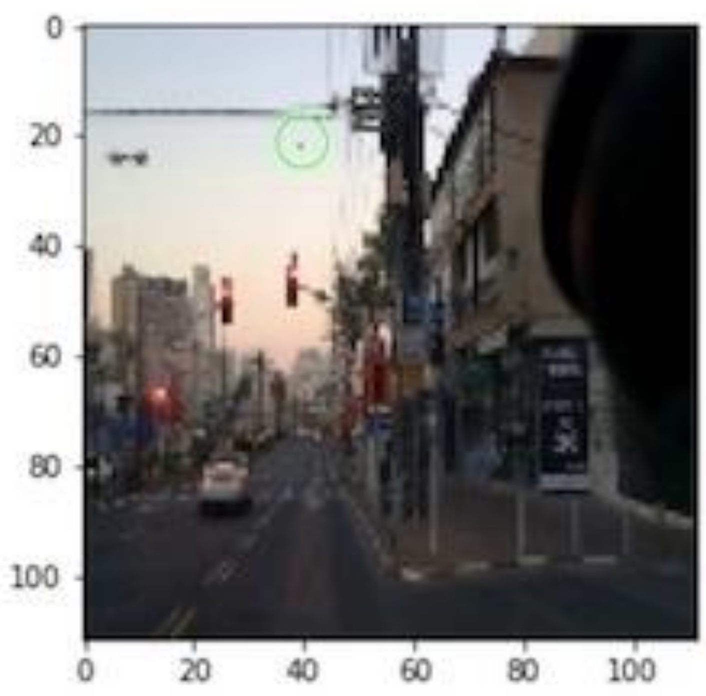


- simpler, less comp. overhead
- needs less data
- lower performance
- human interpretation/adaptation

Unsurprising: There are fundamental limitations

Green light classified as red
after one pixel change

Green light classified as red
after one pixel change



Red light classified as green
after one pixel change.

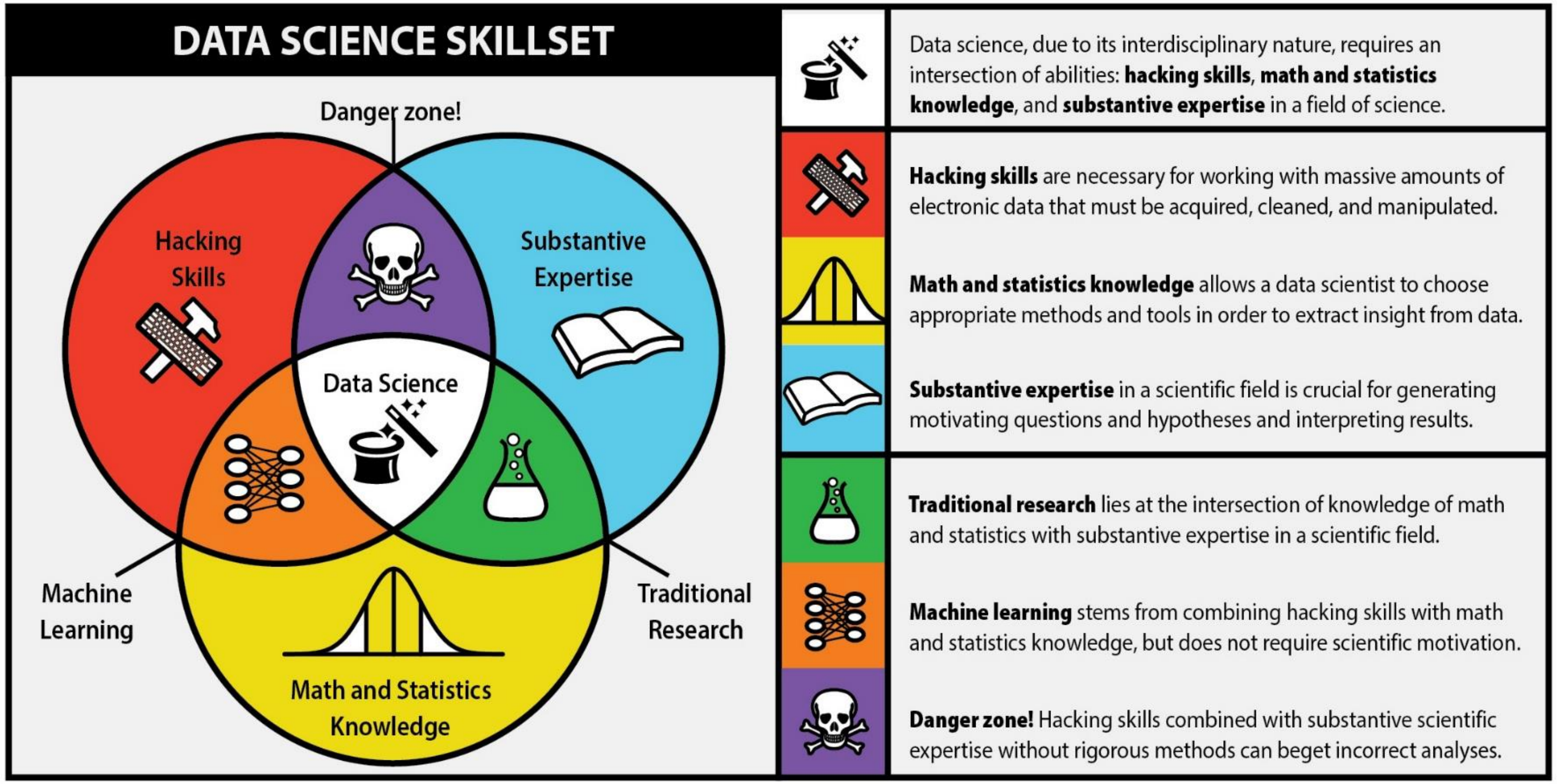
Winner Nexar traffic light challenge: On average, it takes only 3 pixels to turn red into green or green into red!

Wicker, M., Huang, X., Kwiatkowska, M. (2018). Feature-Guided Black-Box Safety Testing of Deep Neural Networks. TACAS 2018. https://doi.org/10.1007/978-3-319-89960-2_22

Human-in-the-Loop: People are still needed.

Therefore, models need to be understandable and we need hybrid forms of intelligence.

We need well-trained Data Scientists !

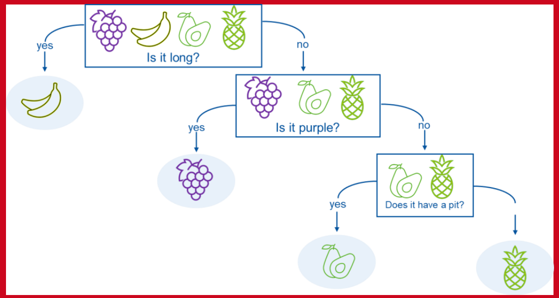


Overview

- Introduction
- What is (tabular) data
- What is data science
- Challenges: reliability, biases, and responsible data science
- Types of data: a high-level taxonomy
- Descriptive statistics
- Simpson’s Paradox, spurious correlations
- Basic visualization: plotting, boxplots, histograms, distributions, scatter plots, bar plots
- One-hot encoding, binning
- “How to Lie with Statistics”

Overview

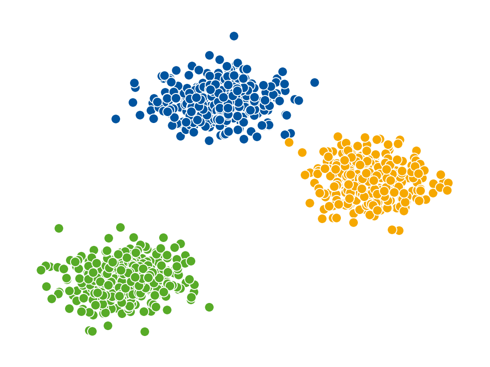
- Decision trees: intro and definitions
- Entropy, information gain, Gini
- ID3 algorithm
- Pruning
- Boosting, bagging, random forests
- Dealing with continuous attributes



Overview

1. **Introduction to Data Science**
2. **Decision Trees**
3. **Clustering**
4. **Frequent Itemsets**
5. **Association Rules and Sequence Mining**
6. **Data Modeling, Quality, and the Data Science Process**
7. **Process Mining**
8. **Text Mining**
9. **Responsible Data Science**

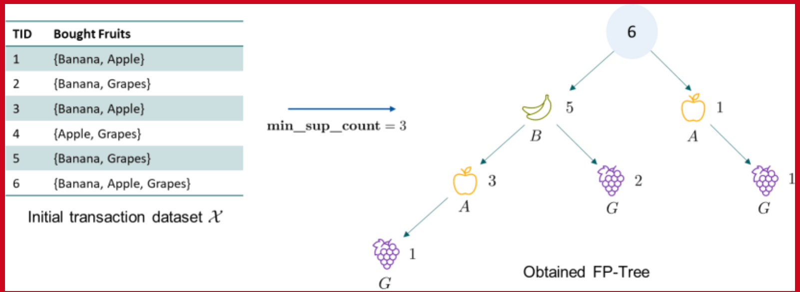
- Clustering: intro and definitions
- Distance measures
- K means, K medoids
- Agglomerative clustering
- DBSCAN



Overview

1. Introduction to Data Science
2. Decision Trees
3. Clustering
4. Frequent Itemsets
5. Association Rules and Sequence Mining
6. Data Modeling, Quality, and the Data Science Process
7. Process Mining
8. Text Mining
9. Responsible Data Science

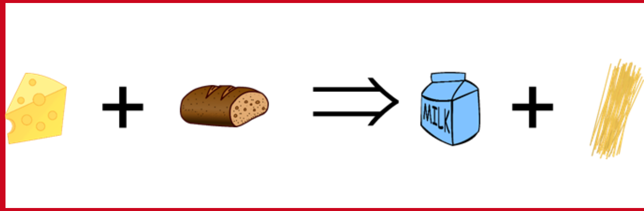
- Frequent itemsets: intro and definitions
- Support and itemsets properties
- The Apriori algorithm
- The FP-Growth algorithm



Overview

1. **Introduction to Data Science**
2. **Decision Trees**
3. **Clustering**
4. **Frequent Itemsets**
5. **Association Rules and Sequence Mining**
6. **Data Modeling, Quality, and the Data Science Process**
7. **Process Mining**
8. **Text Mining**
9. **Responsible Data Science**

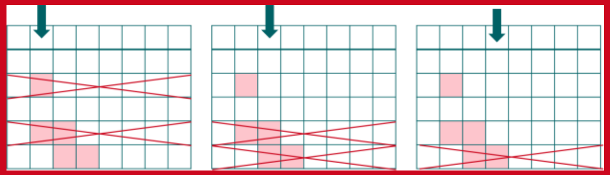
- Association rules: intro and definitions
- Generating and evaluating association rules
- Simpson's Paradox in association rules
- Sequence mining: intro and definitions
- The Apriori-all algorithm



Overview

1. **Introduction to Data Science**
2. **Decision Trees**
3. **Clustering**
4. **Frequent Itemsets**
5. **Association Rules and Sequence Mining**
6. **Data Modeling, Quality, and the Data Science Process**
7. **Process Mining**
8. **Text Mining**
9. **Responsible Data Science**

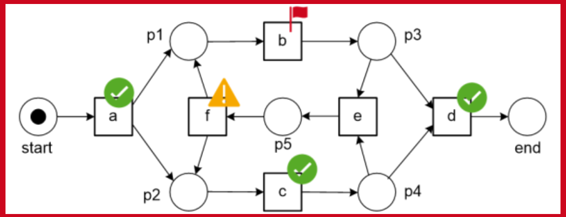
- Data modeling and quality: intro and definitions
- Beyond tables: Temporal data, time series, event data
- Data science processes and methodologies
- PDCA (Plan, Do, Check, Act)
- DMAIC (Define, Measure, Analyze, Improve, and Control)



Overview

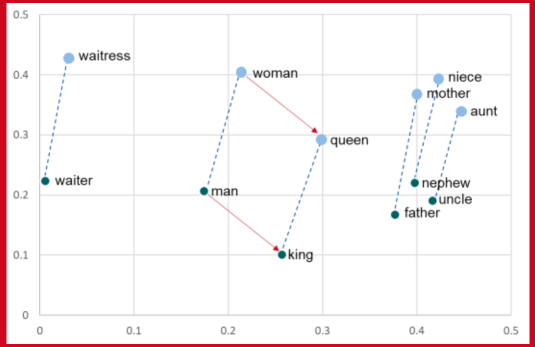
1. **Introduction to Data Science**
2. **Decision Trees**
3. **Clustering**
4. **Frequent Itemsets**
5. **Association Rules and Sequence Mining**
6. **Data Modeling, Quality, and the Data Science Process**
7. **<u>Process Mining</u>**
8. **Text Mining**
9. **Responsible Data Science**

- Process mining: intro and definitions
- Models and formalisms
- Process discovery: bottom-up and top-down
- The inductive miner
- Token-based replay
- Fitness and token-based replay conformance checking



Overview

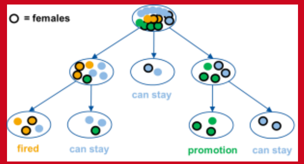
- Text mining: intro and definitions
- Text preprocessing
- Text models: BoW and tf-idf
- N-grams
- Autoencoding
- word2vec

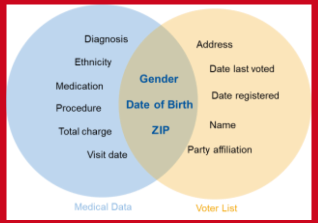


Overview

1. **Introduction to Data Science**
2. **Decision Trees**
3. **Clustering**
4. **Frequent Itemsets**
5. **Association Rules and Sequence Mining**
6. **Data Modeling, Quality, and the Data Science Process**
7. **Process Mining**
8. **Text Mining**
9. **Responsible Data Science**

- RDS: intro and definitions
- Anonymization
- K-Anonymity, L-Diversity, T-Closeness
- Unfairness, prejudice, discrimination
- Fairness risks and metrics
- Fair decision trees
- FACT vs FAIR





Elements of Machine Learning & Data Science

Winter semester 2023/24

Empirical analysis and performance optimization (AutoML)

Prof. Holger Hoos

Chair for AI Methodology (AIM)

AIM:

- machine learning
- automated reasoning
- optimisation
- empirical analysis of (AI) algorithms
- automated design of (AI) algorithms
- human-centred AI
- AI for Good, AI for All

Prof. Holger Hoos

Dr. Jakob Bossek

Dr. Igor Vatolkin

Marie Anastacio, MSc

Julian Dierkes, MSc

Henning Duwe, MSc

Wadie Skaf, MSc

Key questions:

- How good is an ML model?
- How good could an ML model be?

Key questions:

- **How good is an ML model?**
  - Is it “fit for use” (i.e., good enough for deployment)?
  - What are its strengths and weaknesses?
  - Might anything have gone wrong during training?

Key questions:

- **How good is an ML model?**
  - How do we assess whether it is “fit for use” (i.e., good enough for deployment)?
  - How do we assess its strengths and weaknesses?
  - How do we detect if anything has gone wrong during training?

Key questions:

- **How good could an ML model be?**
  - Are we using the best possible ML method / model?
  - Have we configured and trained it in the best possible way?
  - Can we further improve performance?

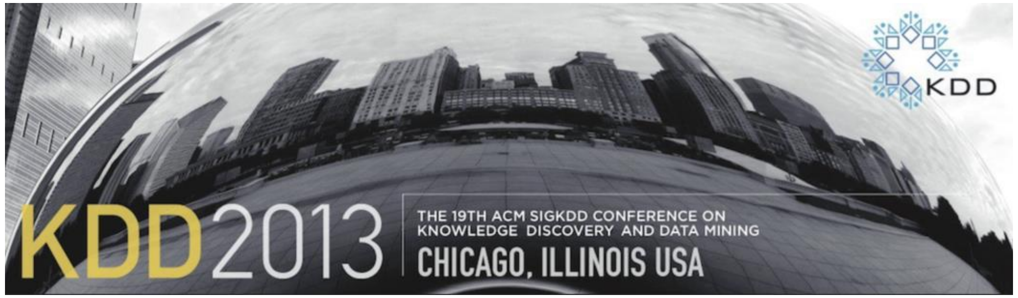
Key questions:

- **How good could an ML model be?**
  - How can we ensure we are using a good ML method / model?
  - How can we configure and train it for optimised performance?
  - How can we further improve performance?

High-level learning goals:

Be able to …

- answer these key questions in a technical manner;
- recognise weaknesses in the empirical performance of ML models using standard tools and methods;
- explain these analysis tools and methods at a technical level;
- use standard tools and methods for selecting models and optimising their hyperparameters;
- explain these AutoML tools and methods at a technical level.



Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms

Chris Thornton Frank Hutter Holger H. Hoos Kevin Leyton-Brown

Department of Computer Science, University of British Columbia
201-2366 Main Mall, Vancouver BC, V6T 1Z4, Canada
{cwthornt, hutter, hoos, kevinlb}@cs.ubc.ca

Test of Time Award for Research

Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms

Chris Thornton ,Frank Hutter, Holger H. Hoos, Kevin Leyton-Brown

DOI:10.1145/3495256

Given the complexity of data science projects and related demand for human expertise, automation has the potential to transform the data science process.

BY TIJL DE BIE, LUC DE RAEDT, JOSÉ HERNÁNDEZ-ORALLO, HOLGER H. HOOS, PADHRAIC SMYTH, AND CHRISTOPHER K.I. WILLIAMS

Automating Data Science

DATA SCIENCE COVERS the full spectrum of deriving insight from data, from initial data gathering and interpretation, via processing and engineering of data, and exploration and modeling, to eventually producing novel insights and decision support systems.

Data science can be viewed as overlapping or broader in scope than other data-analytic methodological disciplines, such as statistics, machine learning, databases, or visualization.[10]

To illustrate the breadth of data science, consider, for example, the problem of recommending items (movies, books, or other products) to customers. While the core of these applications can consist of algorithmic techniques such as matrix factorization, a deployed system will involve a much wider range of technological and human considerations. These range from scalable back-end transaction systems that retrieve customer and product data in real time, experimental design for evaluating system changes, causal analysis for understanding the effect of interventions, to the human factors and psychology that underlie how customers react to visual information displays and make decisions.

As another example, in areas such as astronomy, particle physics, and climate science, there is a rich tradition of building computational pipelines to support data-driven discovery and hypothesis testing. For instance, geoscientists use monthly global landcover maps based on satellite imagery at sub-kilometer resolutions to better understand how the Earth's surface is changing over time.[50] These maps are interactive and browsable, and they are the result of a complex data-processing pipeline, in which terabytes to petabytes of raw sensor and image data are transformed into databases of automatically detected and annotated objects and information. This type of pipeline involves many steps, in which human decisions and insight are critical, such as instrument calibration, removal of outliers, and classification of pixels.

The breadth and complexity of these and many other data science scenarios means the modern data scientist requires broad knowledge and experience across a multitude of topics. Together with an increasing demand for data analysis skills, this has led to a shortage of trained data scientists with appropriate background and experience, and significant market competition for limited expertise. Considering this bottleneck, it is not surprising there is increasing interest in automat-

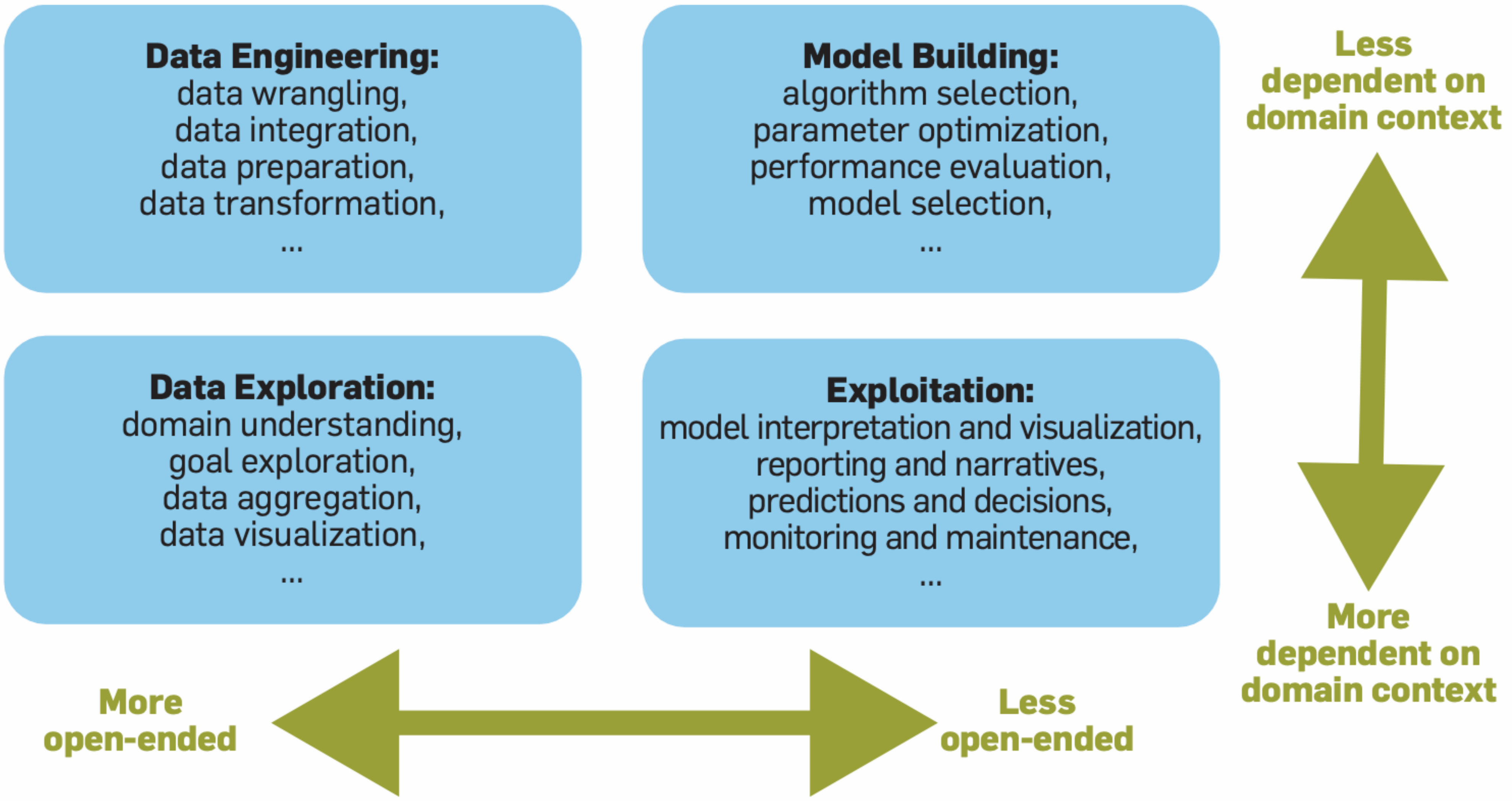
» key insights

- Automation in data science aims to facilitate and transform the work of data scientists, not to replace them.
- Important parts of data science are already being automated, especially in the modeling stages, where techniques such as automated machine learning (AutoML) are gaining traction.
- Other aspects are more difficult to automate, not only because of technological challenges, but because open-ended and context-dependent tasks require human interaction.

ILLUSTRATION BY JUSTIN METZ

» key insights

- Automation in data science aims to facilitate and transform the work of data scientists, not to replace them.
- Important parts of data science are already being automated, especially in the modeling stages, where techniques such as automated machine learning (AutoML) are gaining traction.
- Other aspects are more difficult to automate, not only because of technological challenges, but because open-ended and context-dependent tasks require human interaction.


Data Engineering:
data wrangling,
data integration,
data preparation,
data transformation,
...
Model Building:
algorithm selection,
parameter optimization,
performance evaluation,
model selection,
...
Less dependent on domain context
Data Exploration:
domain understanding,
goal exploration,
data aggregation,
data visualization,
...
Exploitation:
model interpretation and visualization,
reporting and narratives,
predictions and decisions,
monitoring and maintenance,
...
More dependent on domain context
More open-ended
Less open-ended

Behold …

… the awesome power of AutoML / AutoDS / AutoAI!

But:

With great power comes great responsibility :-)

Enjoy the course and see you in our lectures!