# Data Science 1

## Vorlesung 1 - Einführung

Prof. Dr. Christian Bockermann

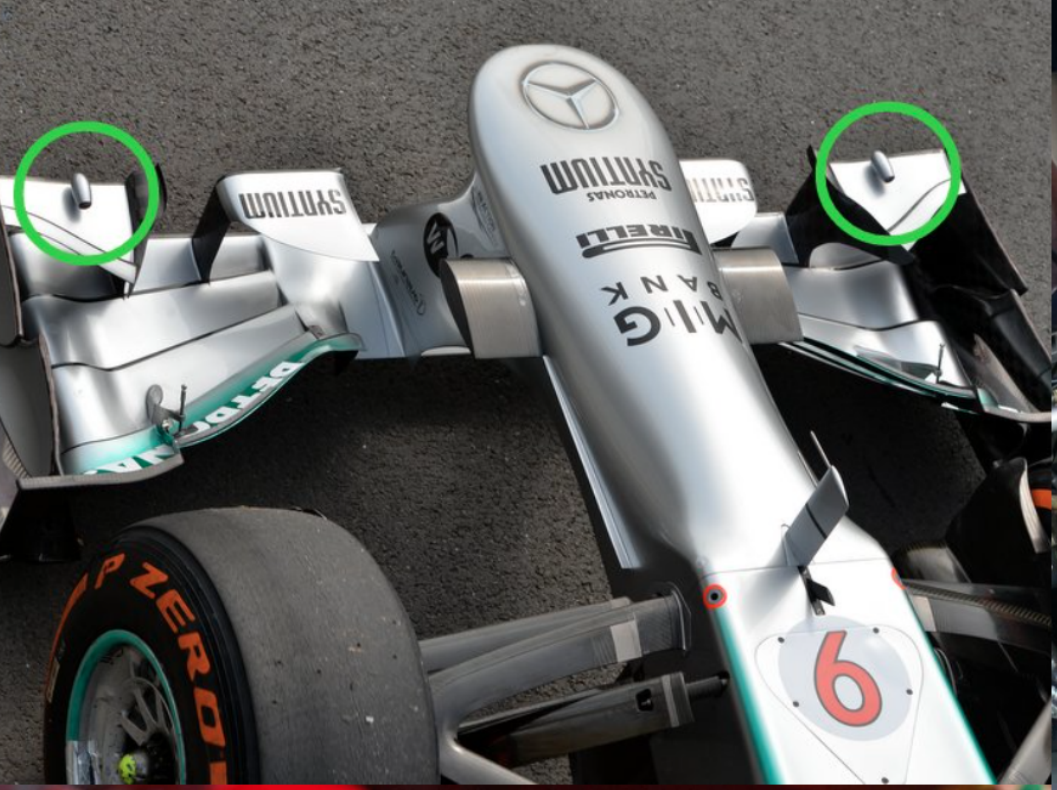
Hochschule Bochum

Sommersemester 2021

# Inhalt


Hochschule Bochum
Bochum University
of Applied Sciences
RENAULT
Red Bull
Pepe Jeans


# Data Science?




**Hochschule Bochum**
Bochum University
of Applied Sciences

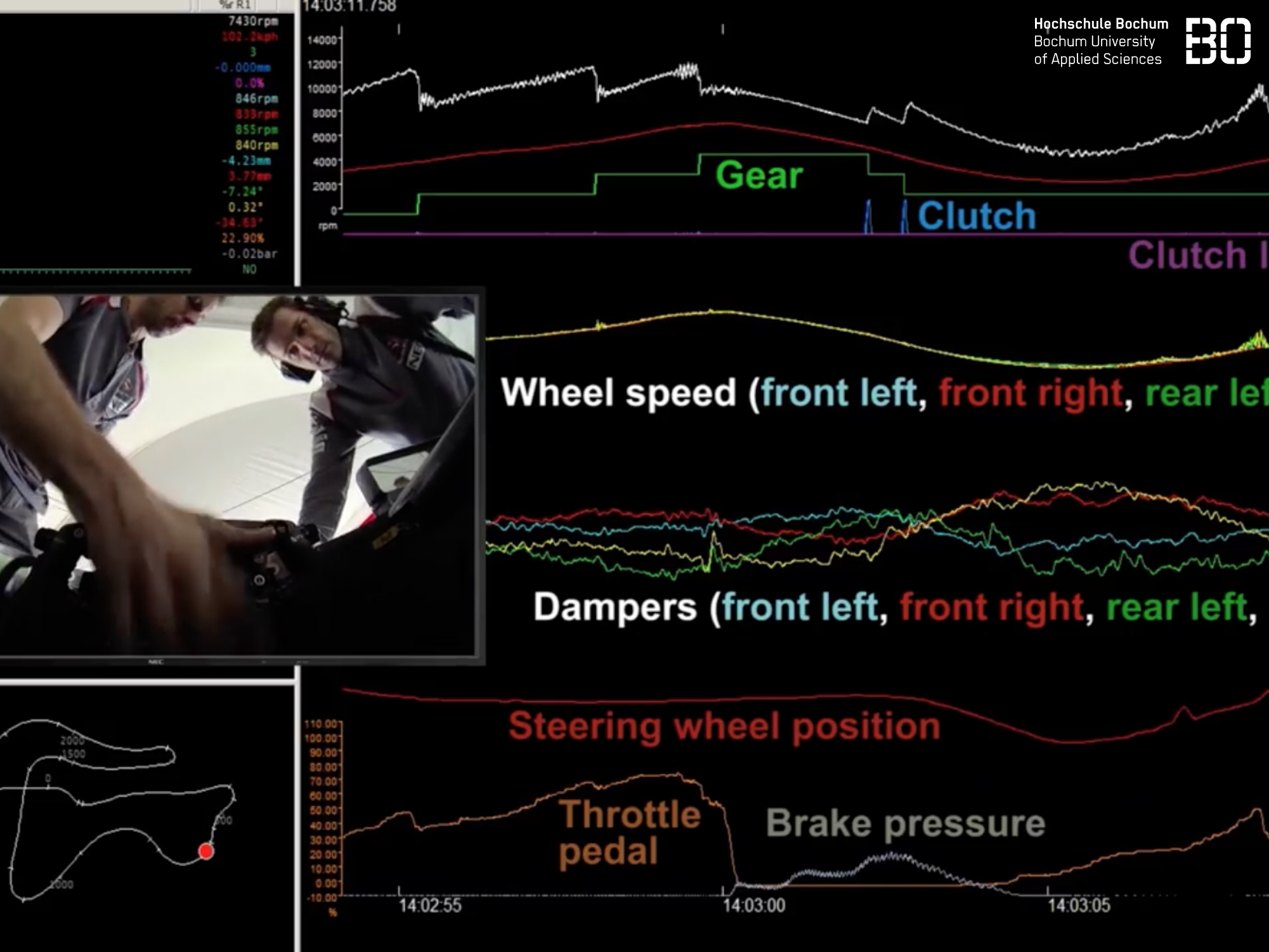
- 240 Sensoren, teilw. hohe Sampling Raten
- ATLAS Funksystem für Echtzeitdaten
  - Regelüberwachung (FIA)
  - Datenanalyse der Rennställe (Echtzeit)
- > 2 GB Daten pro Auto *pro Runde*

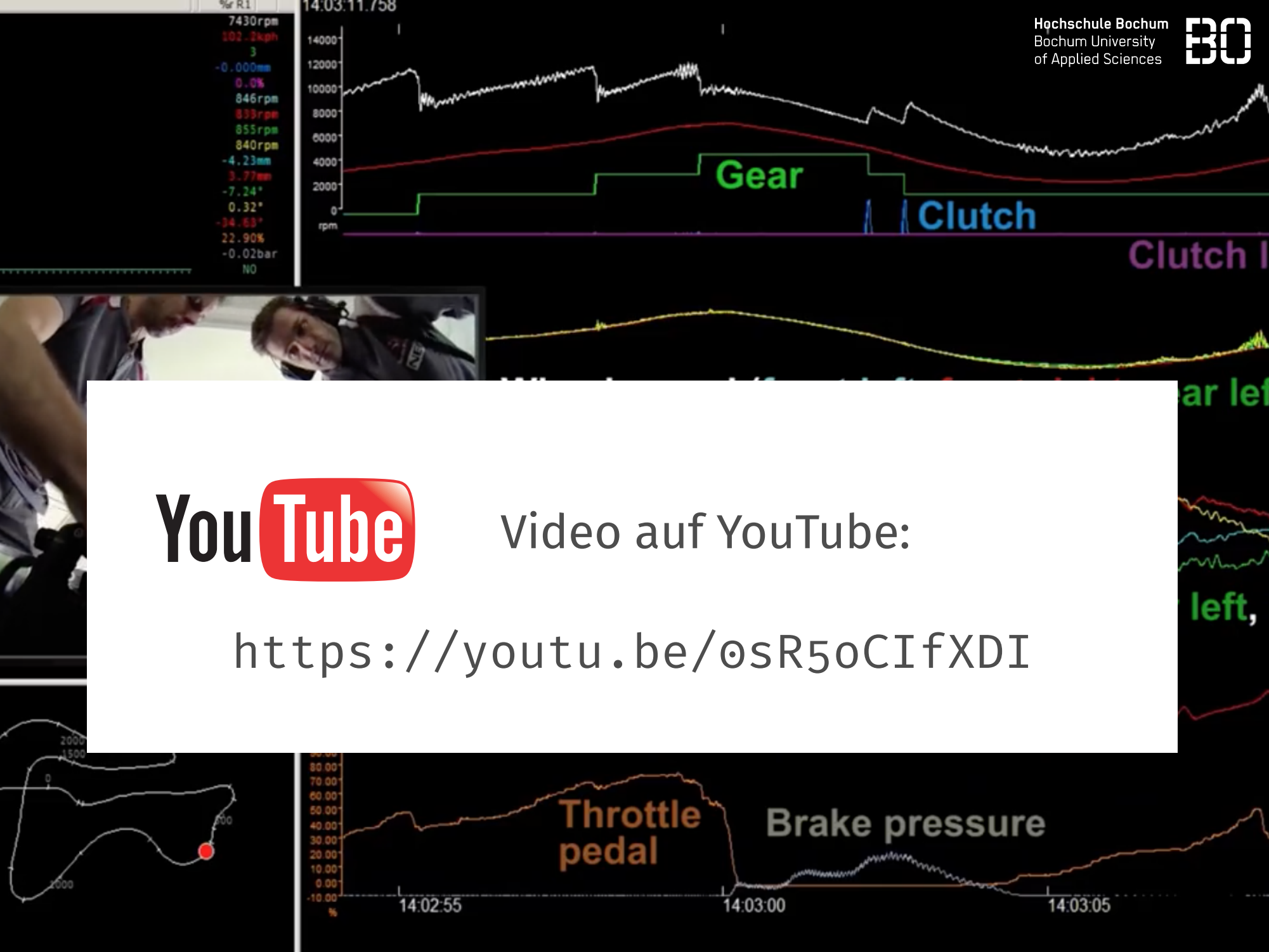
https://www.computerwoche.de/a/it-in-der-formel-1,3213160
https://gigaom.com/2013/05/29/formula-1-racing-changes-pose-big-data-challenge/
https://www.intel.co.uk/content/www/uk/en/it-management/cloud-analytic-hub/big-data-powers-f1.html

- 240 Sensoren, teilw. hohe Sampling Raten
- ATLAS Funksystem für Echtzeitdaten
  - Regelüberwachung (FIA)
  - Datenanalyse der Rennställe (Echtzeit)
- > 2 GB Daten pro Auto *pro Runde*

- Vorhersage: Reifen/Material-Ermüdung
- Erkennung von Leistungsverlust
- Optimierung durch Echtzeitanalyse!

https://www.computerwoche.de/a/it-in-der-formel-1,3213160
https://gigaom.com/2013/05/29/formula-1-racing-changes-pose-big-data-challenge/
https://www.intel.co.uk/content/www/uk/en/it-management/cloud-analytic-hub/big-data-powers-f1.html


Hochschule Bochum
Bochum University
of Applied Sciences
14:03:11.758
7430rpm
102.2kph
3
-0.000mm
0.0%
846rpm
839rpm
855rpm
840rpm
-4.23mm
3.77mm
-7.24°
0.32°
-34.63°
22.90%
-0.02bar
NO
14000
12000
10000
8000
6000
4000
2000
0
rpm
Gear
Clutch
Clutch I
Wheel speed (front left, front right, rear lef
Dampers (front left, front right, rear left,
Steering wheel position
Throttle pedal
Brake pressure
110.00
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
-10.00
%
14:02:55
14:03:00
14:03:05



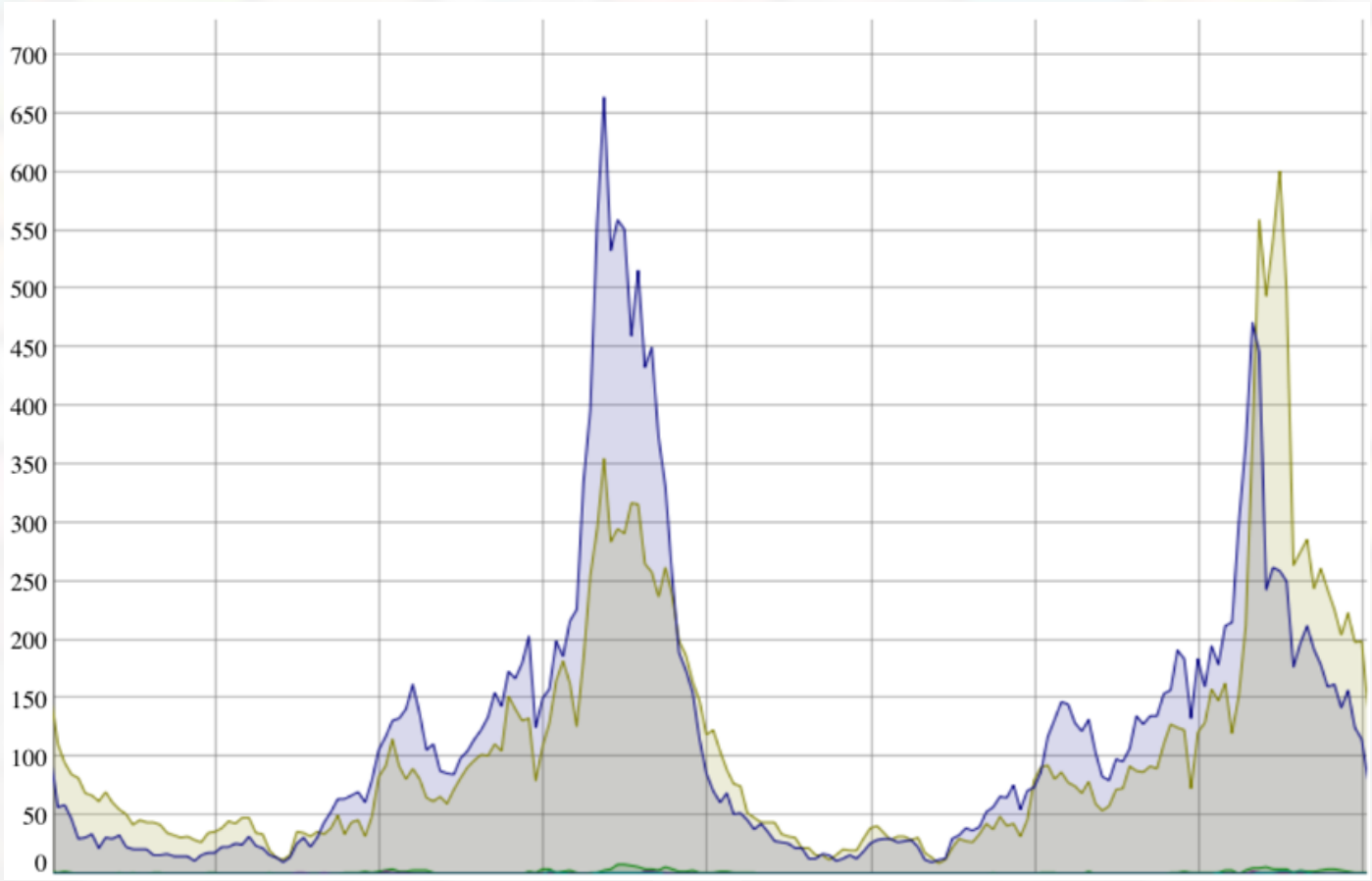

YouTube Video auf YouTube:

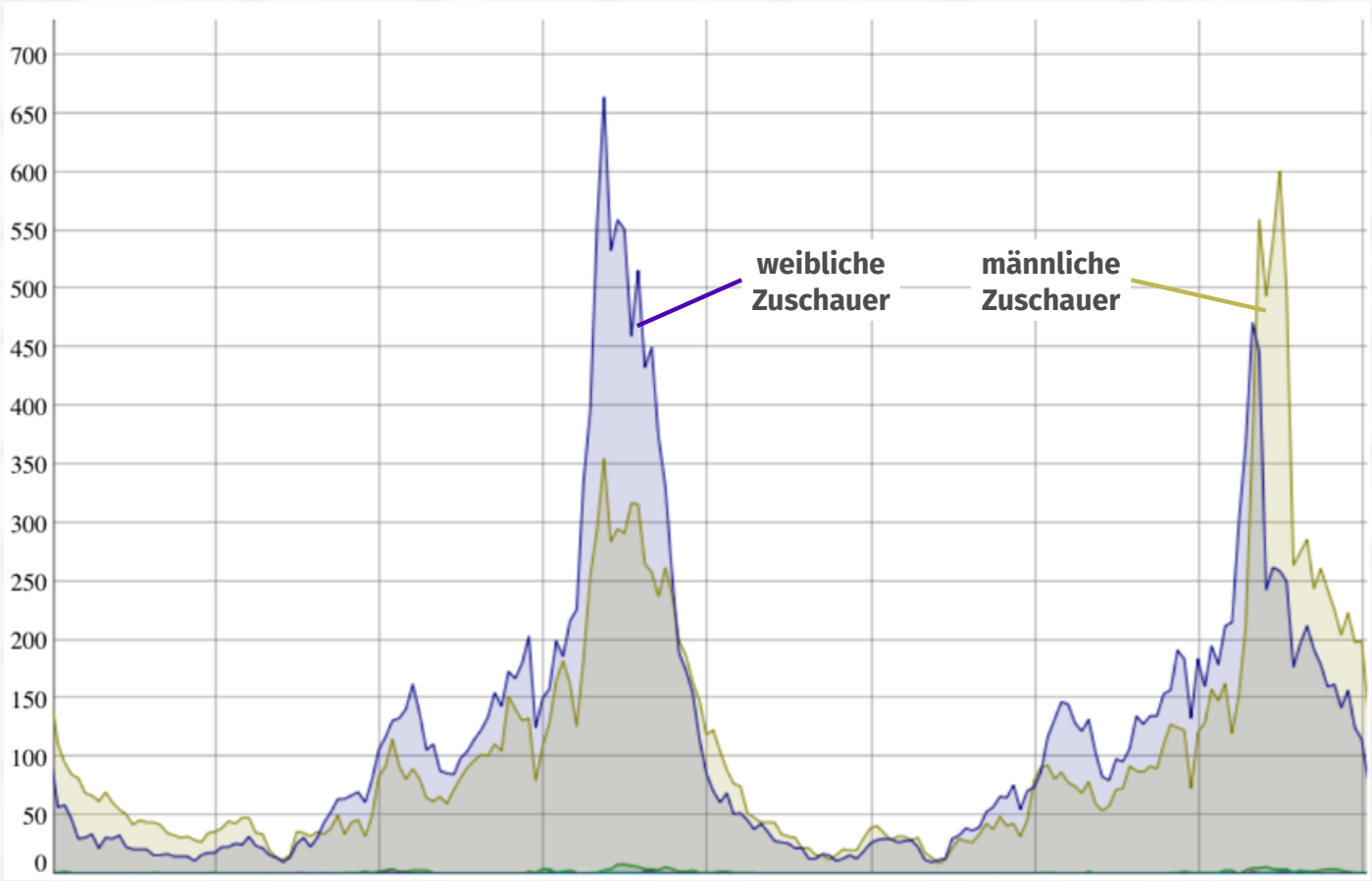
https://youtu.be/osR5oCIfXDI

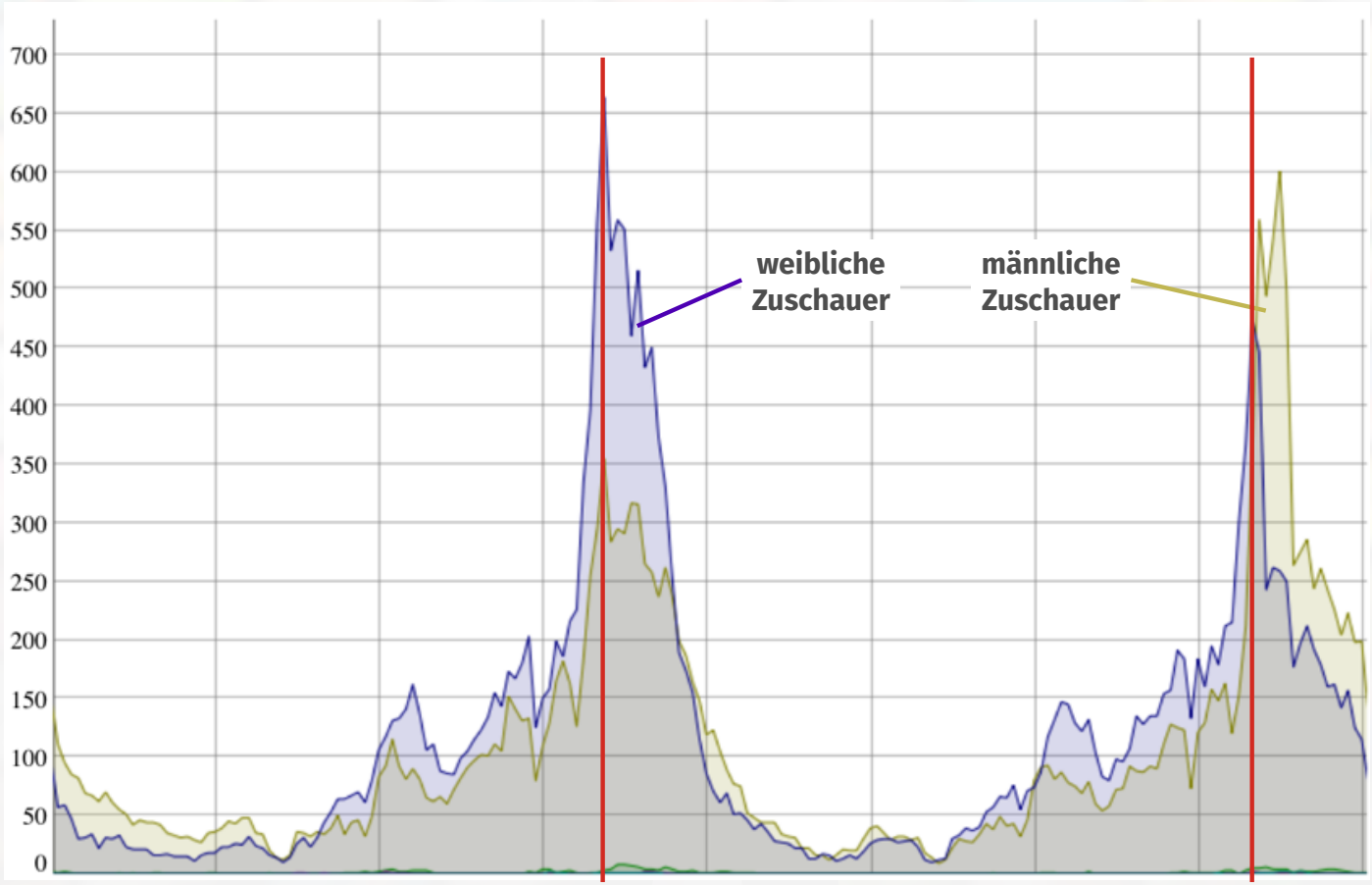

Hochschule Bochum
Bochum University
of Applied Sciences

Hochschule Bochum
Bochum University
of Applied Sciences

# Data Science?

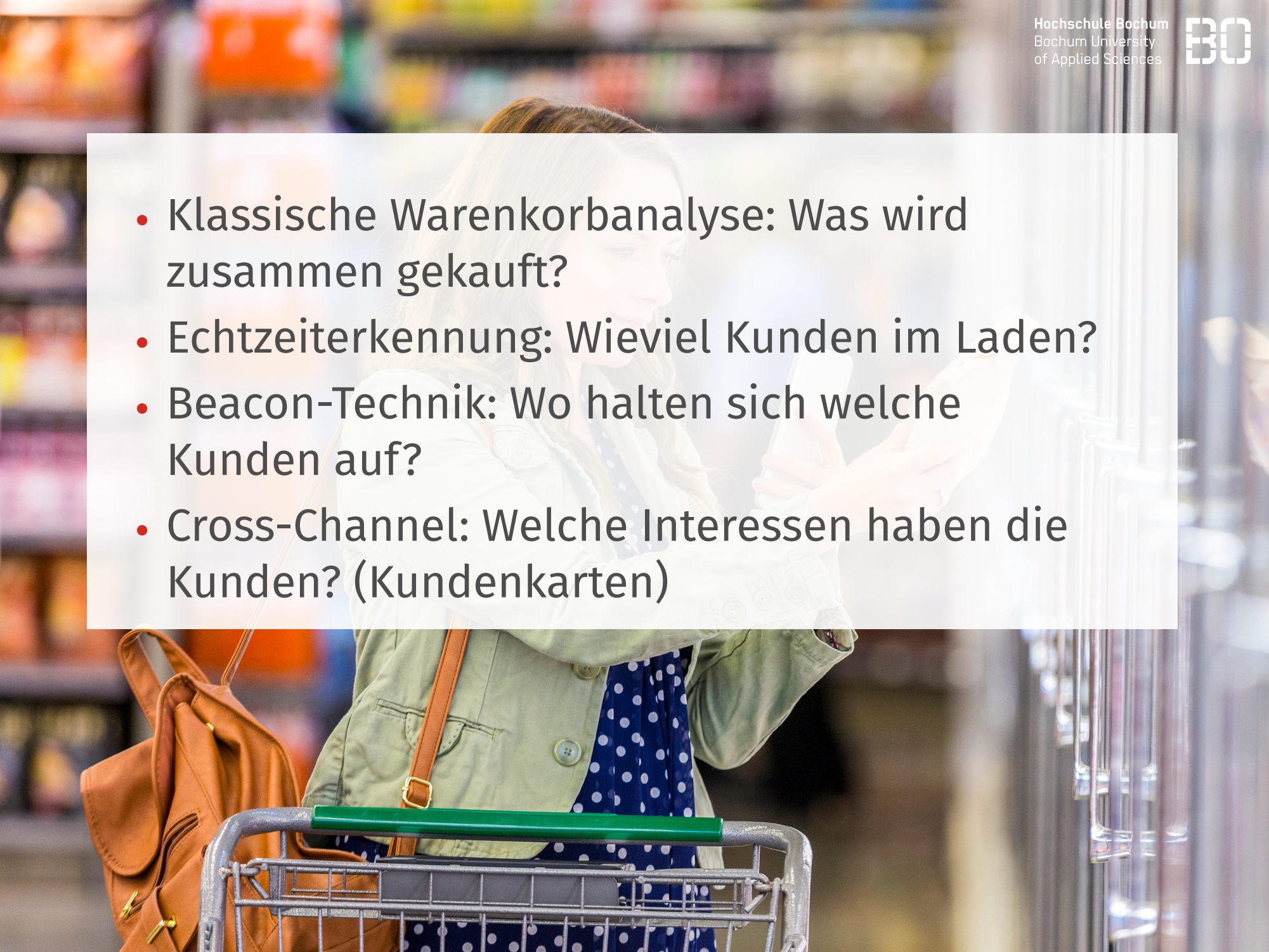

700
650
600
550
500
450
400
350
300
250
200
150
100
50
0


700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
weibliche Zuschauer
männliche Zuschauer

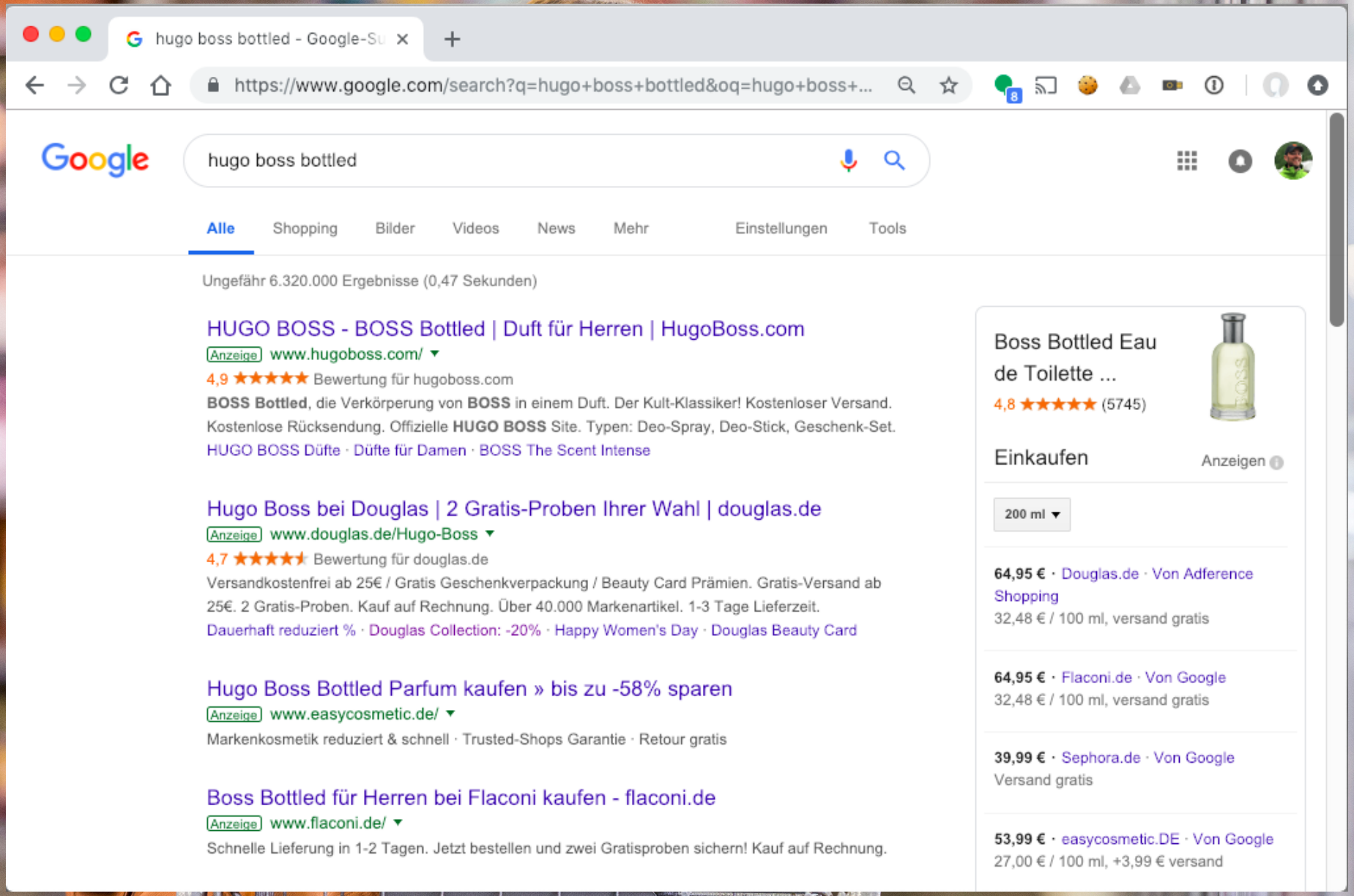

700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
weibliche Zuschauer
männliche Zuschauer
Mi, 14.3.
20:15 Uhr
Do, 15.3.
20:15 Uhr

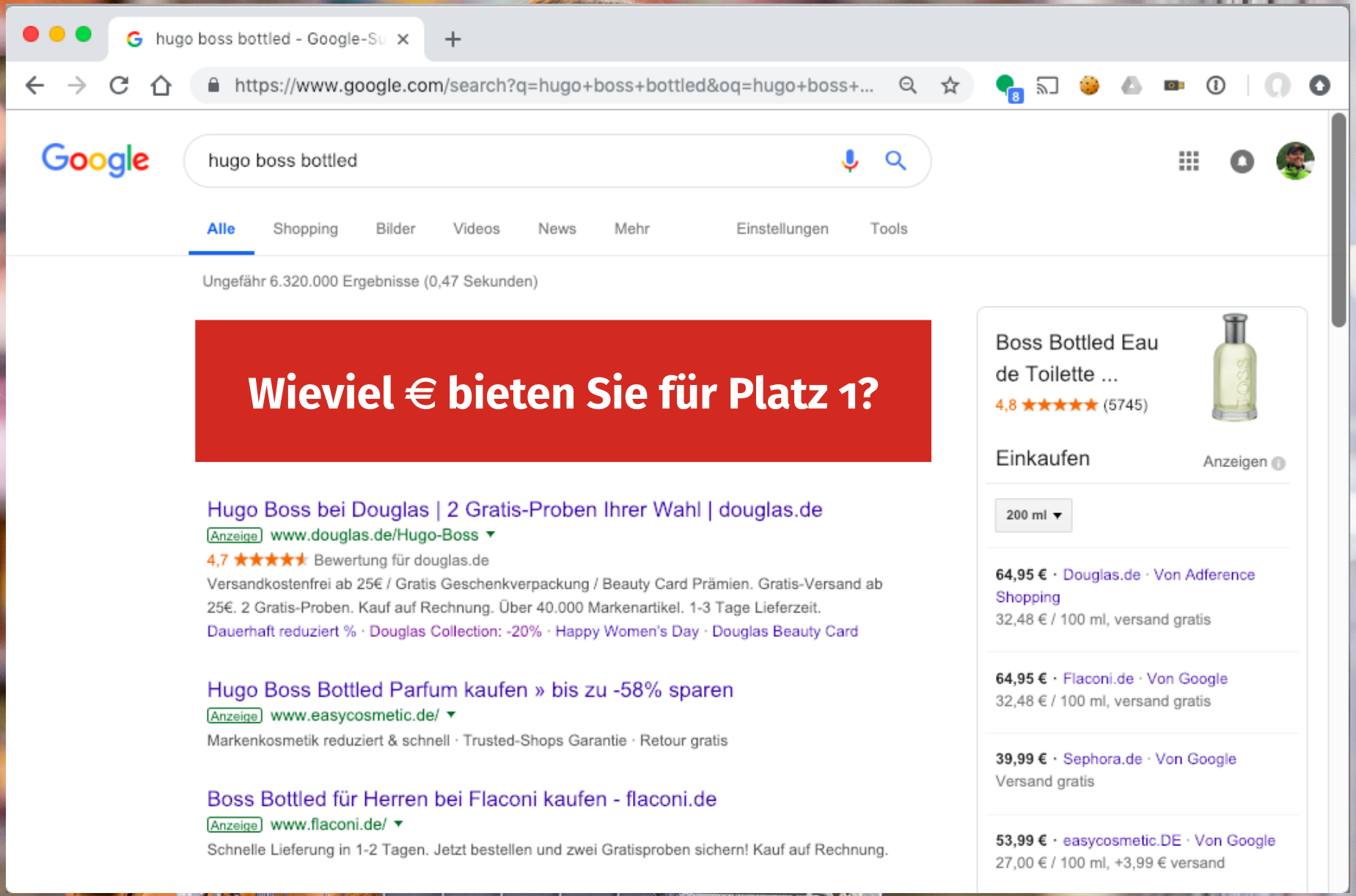

Hochschule Bochum
Bochum University
of Applied Sciences


Hochschule Bochum
Bochum University
of Applied Sciences
Data Science?



- Klassische Warenkorbanalyse: Was wird zusammen gekauft?
- Echtzeiterkennung: Wieviel Kunden im Laden?
- Beacon-Technik: Wo halten sich welche Kunden auf?
- Cross-Channel: Welche Interessen haben die Kunden? (Kundenkarten)


hugo boss bottled - Google-Su
https://www.google.com/search?q=hugo+boss+bottled&oq=hugo+boss+...
Google
hugo boss bottled
Alle
Shopping
Bilder
Videos
News
Mehr
Einstellungen
Tools
Ungefähr 6.320.000 Ergebnisse (0,47 Sekunden)
HUGO BOSS - BOSS Bottled | Duft für Herren | HugoBoss.com
Anzeige www.hugoboss.com/
4,9 Bewertung für hugoboss.com
BOSS Bottled, die Verkörperung von BOSS in einem Duft. Der Kult-Klassiker! Kostenloser Versand. Kostenlose Rücksendung. Offizielle HUGO BOSS Site. Typen: Deo-Spray, Deo-Stick, Geschenk-Set.
HUGO BOSS Düfte · Düfte für Damen · BOSS The Scent Intense
Hugo Boss bei Douglas | 2 Gratis-Proben Ihrer Wahl | douglas.de
Anzeige www.douglas.de/Hugo-Boss
4,7 Bewertung für douglas.de
Versandkostenfrei ab 25€ / Gratis Geschenkverpackung / Beauty Card Prämien. Gratis-Versand ab 25€. 2 Gratis-Proben. Kauf auf Rechnung. Über 40.000 Markenartikel. 1-3 Tage Lieferzeit.
Dauerhaft reduziert % · Douglas Collection: -20% · Happy Women's Day · Douglas Beauty Card
Hugo Boss Bottled Parfum kaufen » bis zu -58% sparen
Anzeige www.easycosmetic.de/
Markenkosmetik reduziert & schnell · Trusted-Shops Garantie · Retour gratis
Boss Bottled für Herren bei Flaconi kaufen - flaconi.de
Anzeige www.flaconi.de/
Schnelle Lieferung in 1-2 Tagen. Jetzt bestellen und zwei Gratisproben sichern! Kauf auf Rechnung.
Boss Bottled Eau de Toilette ...
4,8 (5745)
Einkaufen
Anzeigen
200 ml
64,95 € · Douglas.de · Von Adference Shopping
32,48 € / 100 ml, versand gratis
64,95 € · Flaconi.de · Von Google
32,48 € / 100 ml, versand gratis
39,99 € · Sephora.de · Von Google
Versand gratis
53,99 € · easycosmetic.DE · Von Google
27,00 € / 100 ml, +3,99 € versand


hugo boss bottled - Google-Su
https://www.google.com/search?q=hugo+boss+bottled&oq=hugo+boss+...
Google
hugo boss bottled
Alle
Shopping
Bilder
Videos
News
Mehr
Einstellungen
Tools
Ungefähr 6.320.000 Ergebnisse (0,47 Sekunden)
Wieviel € bieten Sie für Platz 1?
Hugo Boss bei Douglas | 2 Gratis-Proben Ihrer Wahl | douglas.de
Anzeige www.douglas.de/Hugo-Boss
4,7 Bewertung für douglas.de
Versandkostenfrei ab 25€ / Gratis Geschenkverpackung / Beauty Card Prämien. Gratis-Versand ab 25€. 2 Gratis-Proben. Kauf auf Rechnung. Über 40.000 Markenartikel. 1-3 Tage Lieferzeit.
Dauerhaft reduziert % · Douglas Collection: -20% · Happy Women's Day · Douglas Beauty Card
Hugo Boss Bottled Parfum kaufen » bis zu -58% sparen
Anzeige www.easycosmetic.de/
Markenkosmetik reduziert & schnell · Trusted-Shops Garantie · Retour gratis
Boss Bottled für Herren bei Flaconi kaufen - flaconi.de
Anzeige www.flaconi.de/
Schnelle Lieferung in 1-2 Tagen. Jetzt bestellen und zwei Gratisproben sichern! Kauf auf Rechnung.
Boss Bottled Eau de Toilette ...
4,8 (5745)
Einkaufen
Anzeigen
200 ml
64,95 € · Douglas.de · Von Adference Shopping
32,48 € / 100 ml, versand gratis
64,95 € · Flaconi.de · Von Google
32,48 € / 100 ml, versand gratis
39,99 € · Sephora.de · Von Google
Versand gratis
53,99 € · easycosmetic.DE · Von Google
27,00 € / 100 ml, +3,99 € versand

# Was wissen wir über die Besucher/Kunden?

| # 12 M | € 12 M | Class | Budget |
|---|---|---|---|
| 4 | 948.33 | High | 10 € |
| 3 | 402.25 | Mid | 5 € |
| 5 | 1210.89 | High | 20 € |
| 4 | 423.43 | Mid | 5 € |
| 1 | 89.99 | Low | 0 € |
| 1 | 125.37 | Low | 0 € |

# Was wissen wir über die Besucher/Kunden?

| # 12 M | € 12 M | Class | Budget |
|---|---|---|---|
| 4 | 948.33 | High | 10 € |
| 3 | 402.25 | Mid | 5 € |
| 5 | 1210.89 | High | 20 € |
| 4 | 423.43 | Mid | 5 € |
| 1 | 89.99 | Low | 0 € |
| 1 | 125.37 | Low | 0 € |

| Schoki |
|---|
| X |
| |
| X |
| X |
| |
| |

PAYBACK

# Was wissen wir über die Besucher/Kunden?

| # 12 M | € 12 M | Class | Budget | Schoki | Katzen |
|---|---|---|---|---|---|
| 4 | 948.33 | High | 10 € | X | X |
| 3 | 402.25 | Mid | 5 € | | |
| 5 | 1210.89 | High | 20 € | X | X |
| 4 | 423.43 | Mid | 5 € | X | X |
| 1 | 89.99 | Low | 0 € | | X |
| 1 | 125.37 | Low | 0 € | | X |

PAYBACK

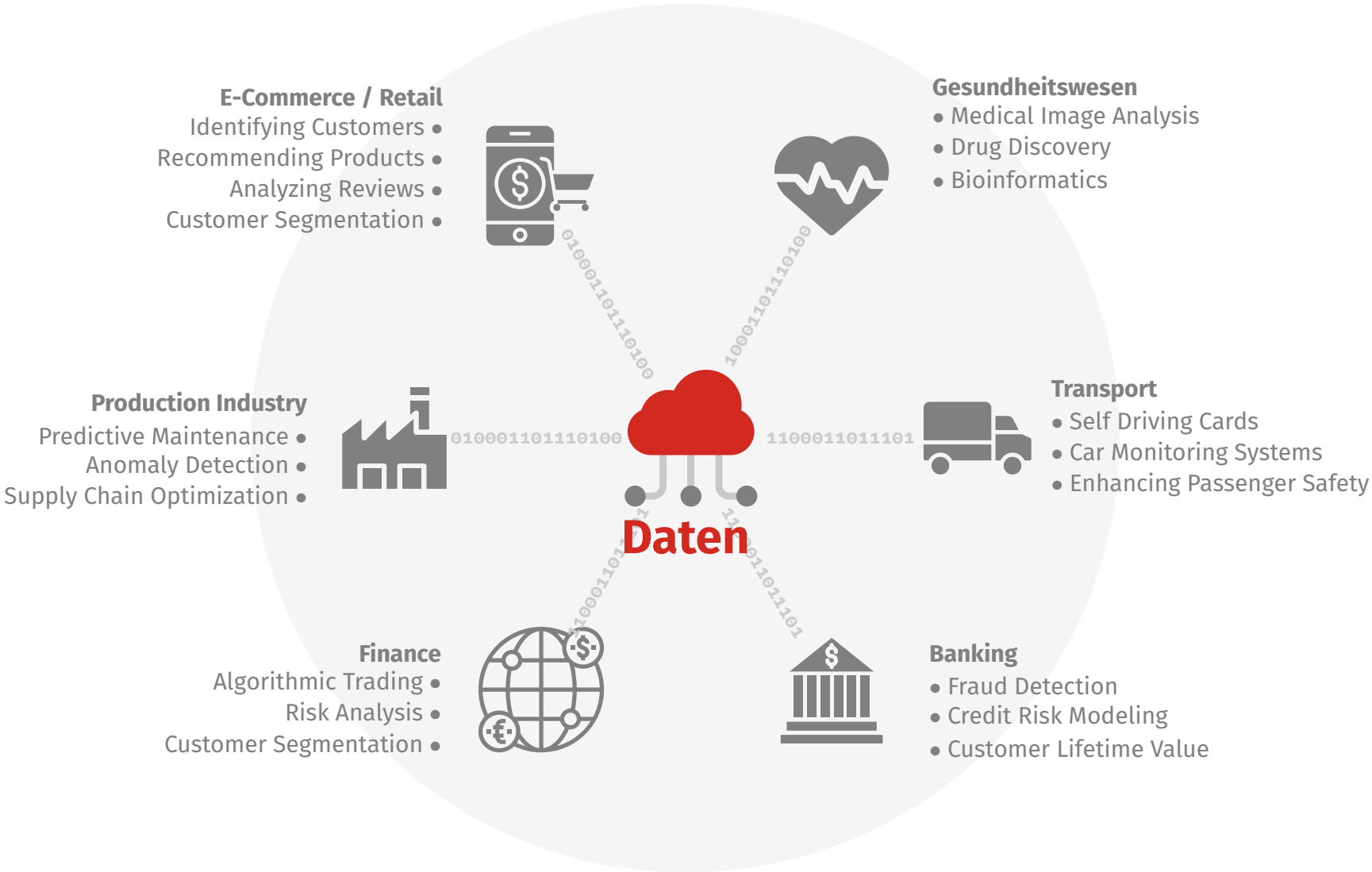
# Was wissen wir über die Besucher/Kunden?

| # 12 M | € 12 M | Class | Budget | Schoki | Katzen | Hugo B |
|---|---|---|---|---|---|---|
| 4 | 948.33 | High | 10 € | X | X | X |
| 3 | 402.25 | Mid | 5 € | | | X |
| 5 | 1210.89 | High | 20 € | X | X | X |
| 4 | 423.43 | Mid | 5 € | X | X | X |
| 1 | 89.99 | Low | 0 € | | X | |
| 1 | 125.37 | Low | 0 € | | X | |

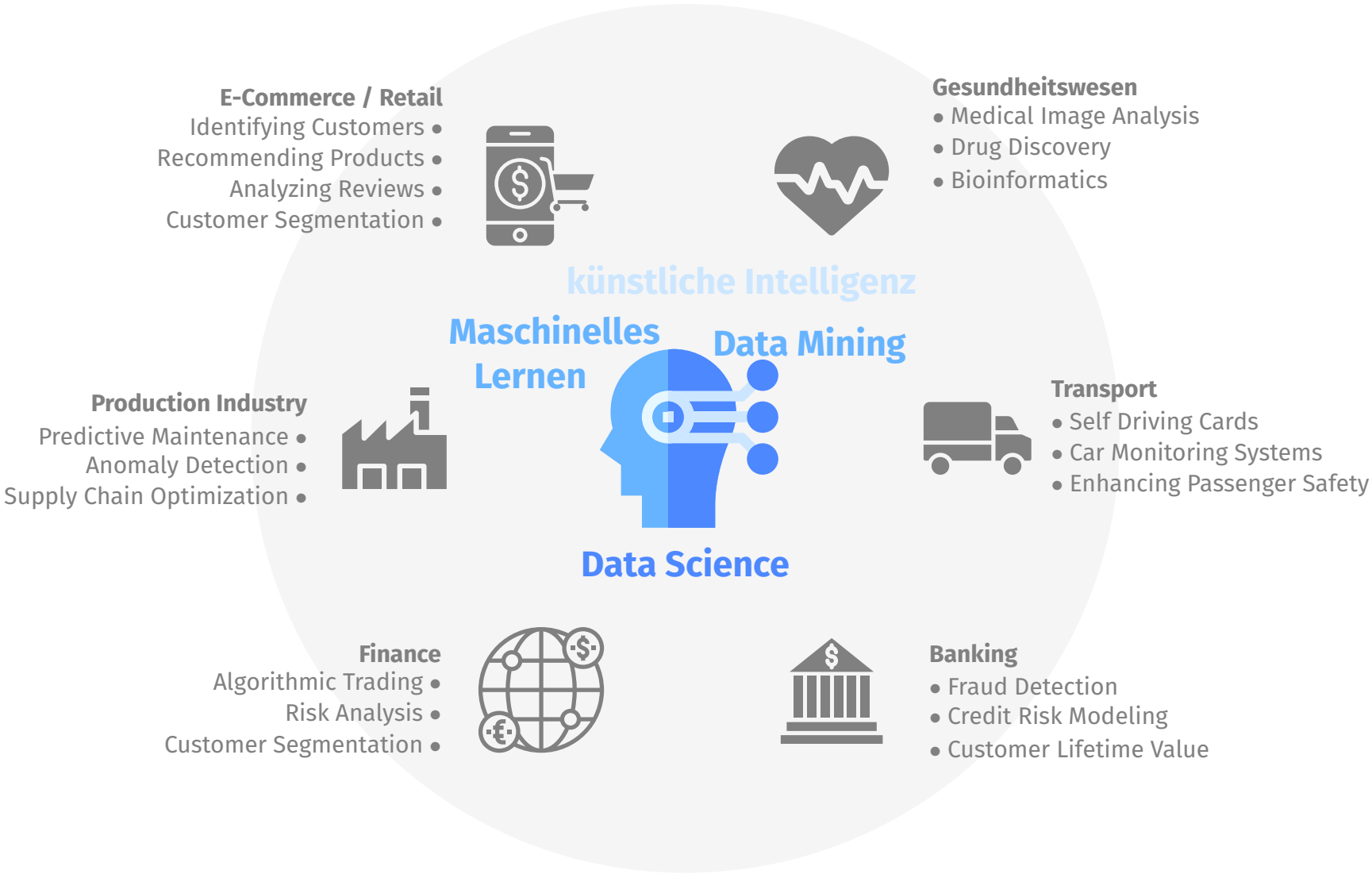
PAYBACK

Google

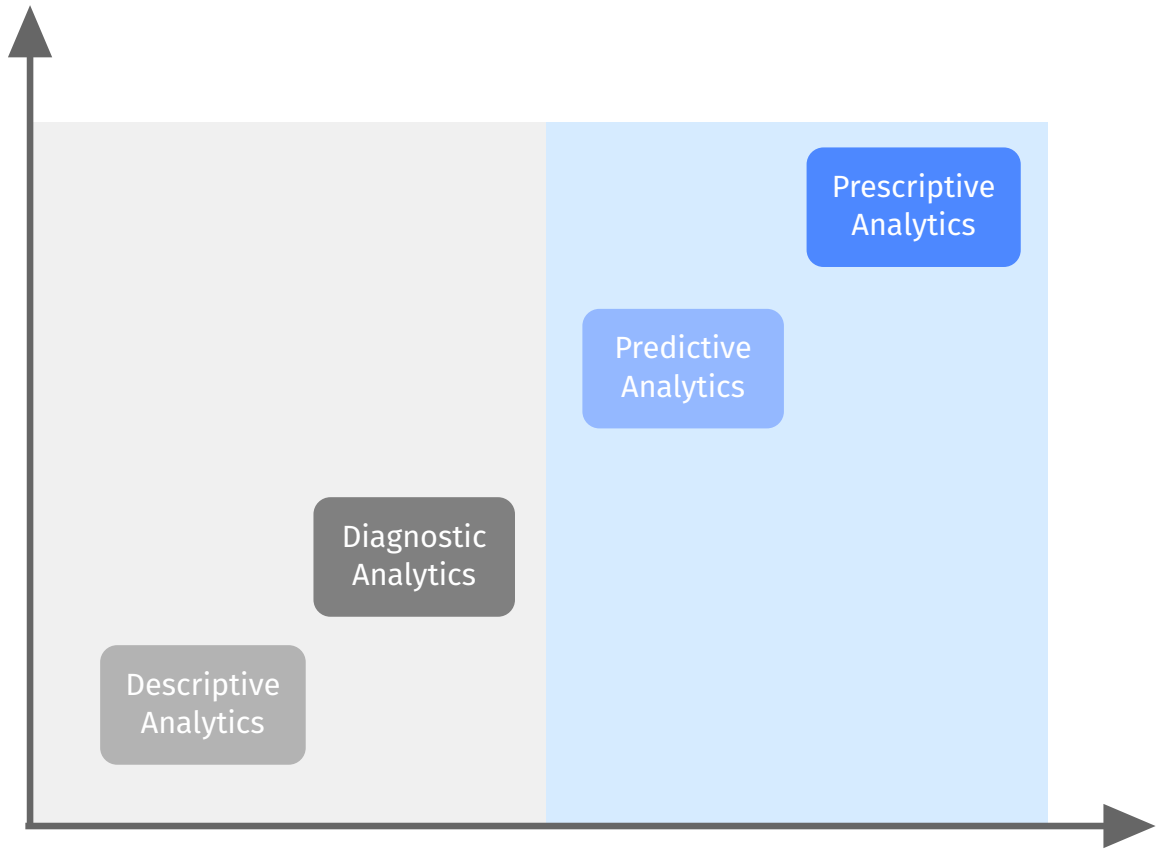
# Data Science, Data Mining und KDD



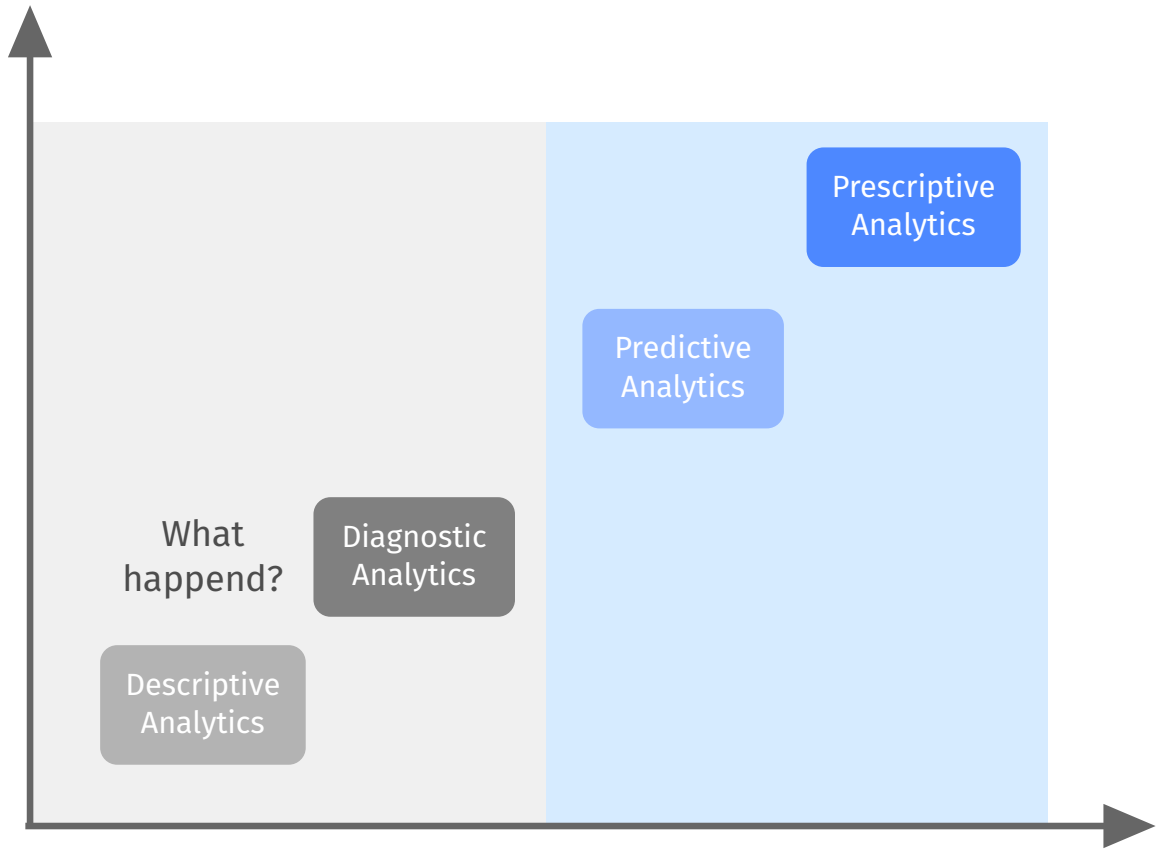
# Data Science, Data Mining und KDD



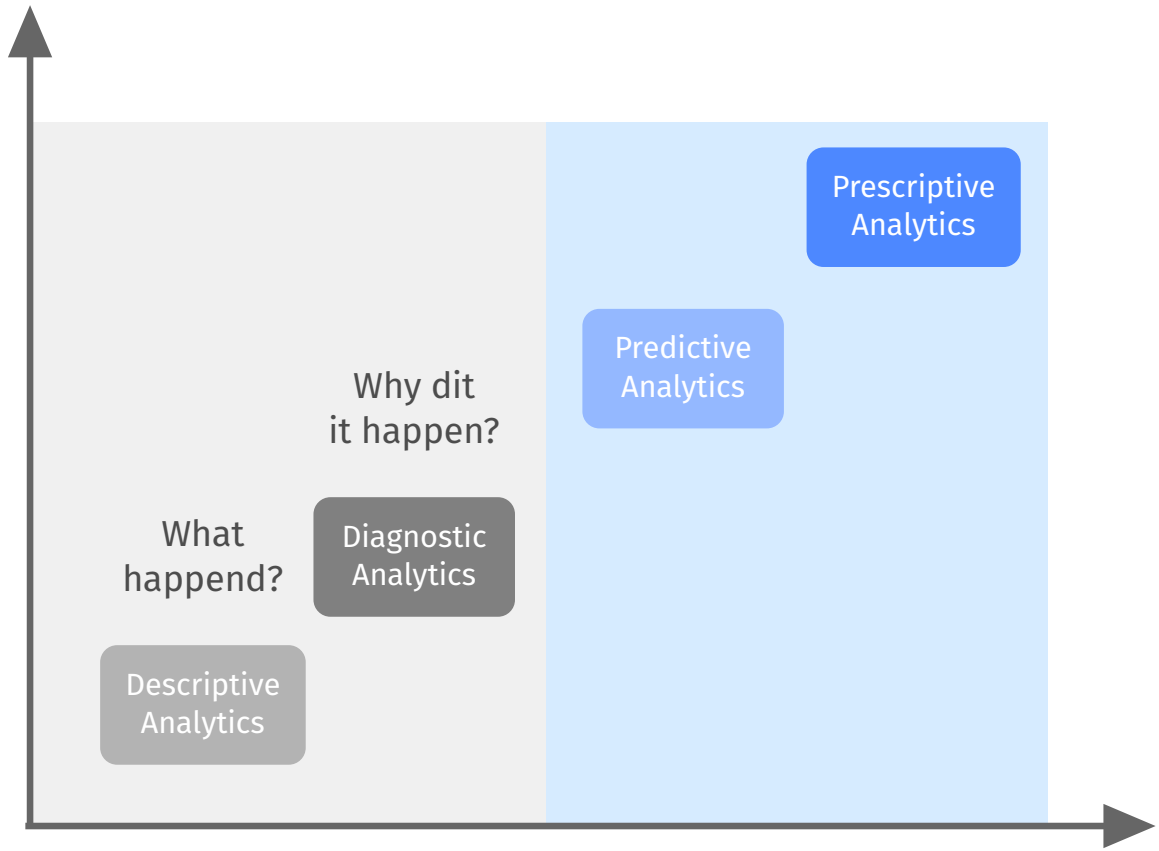
Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



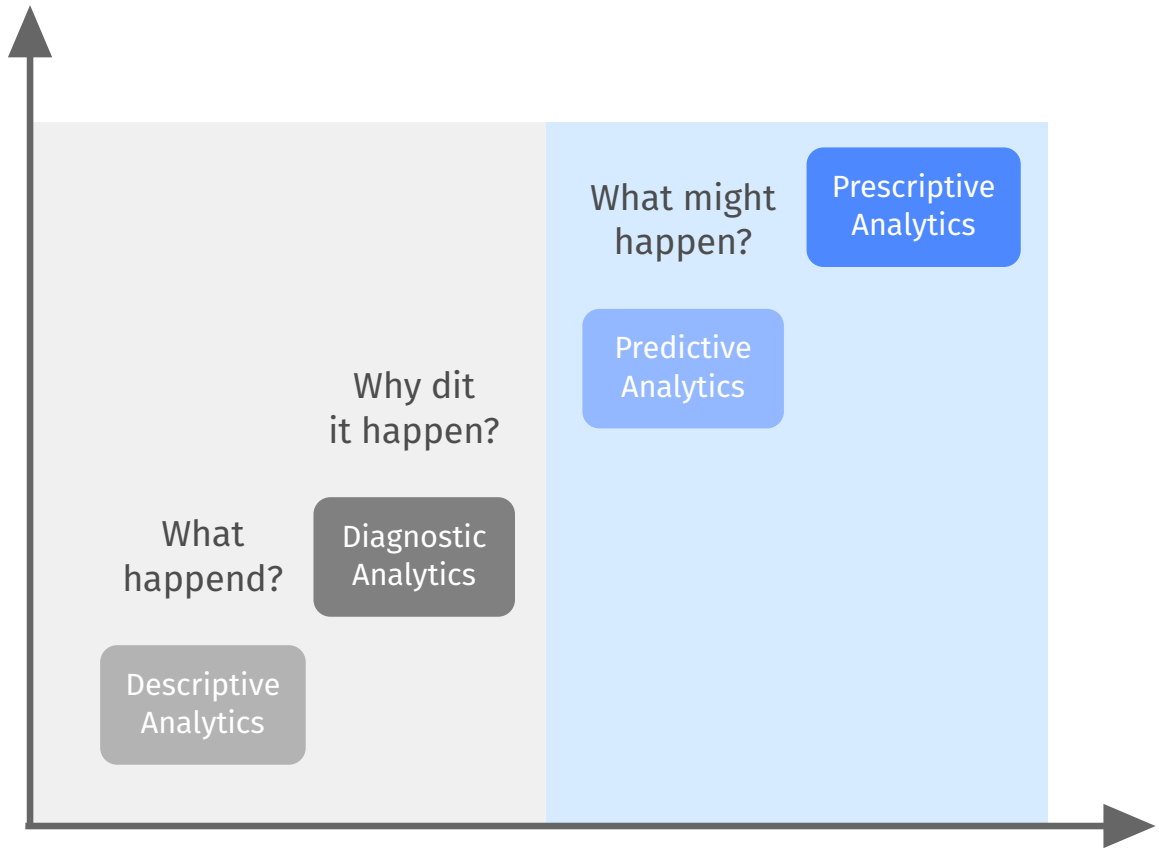
Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



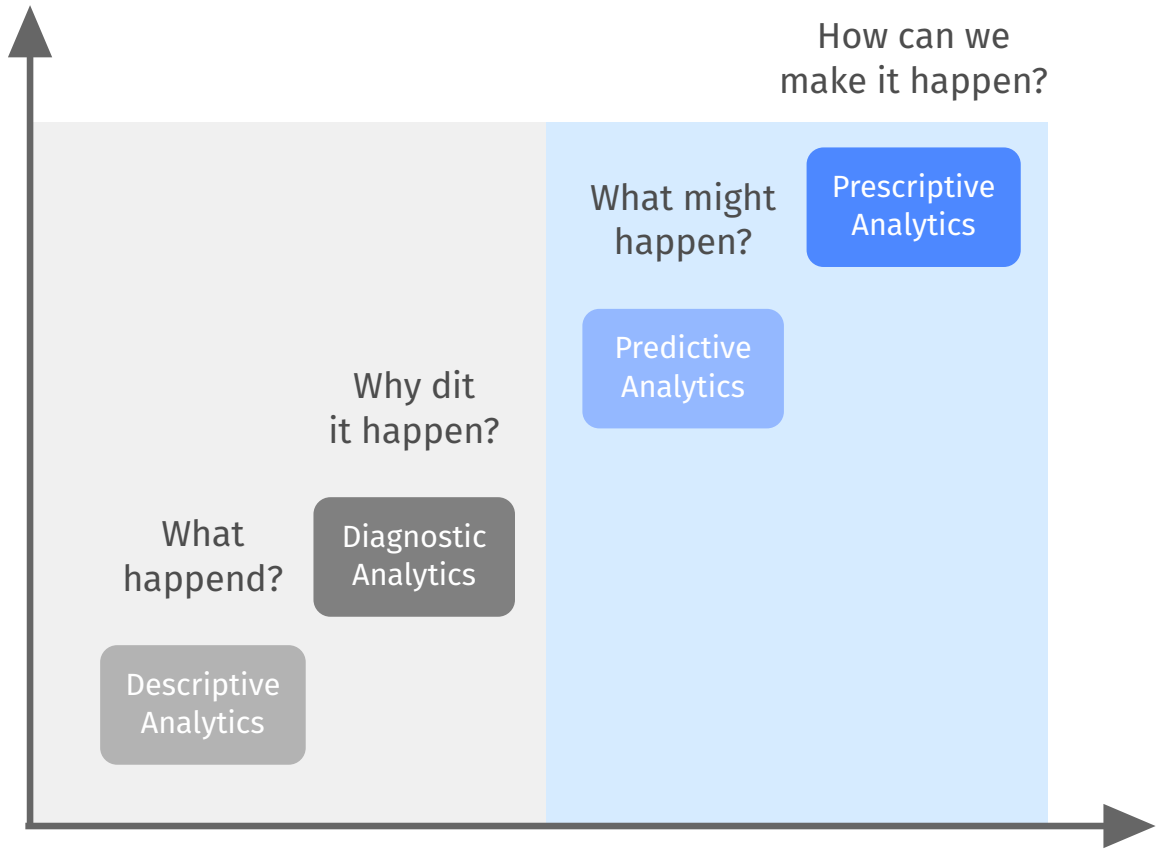
Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



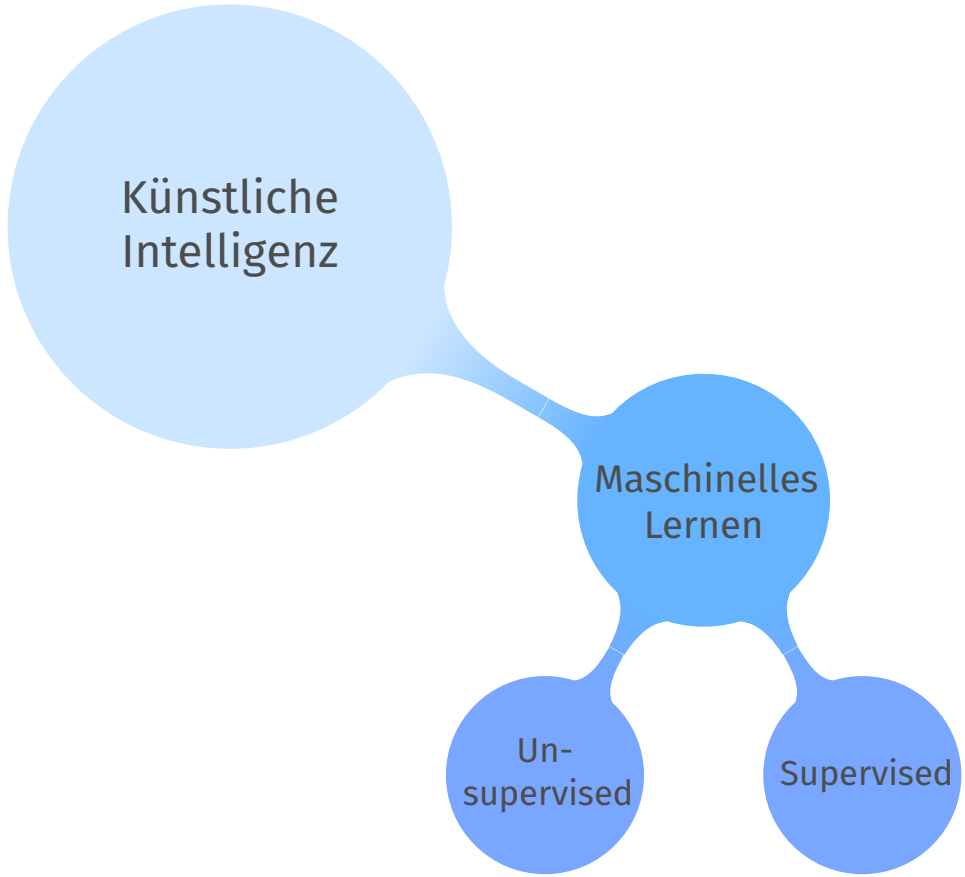
Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



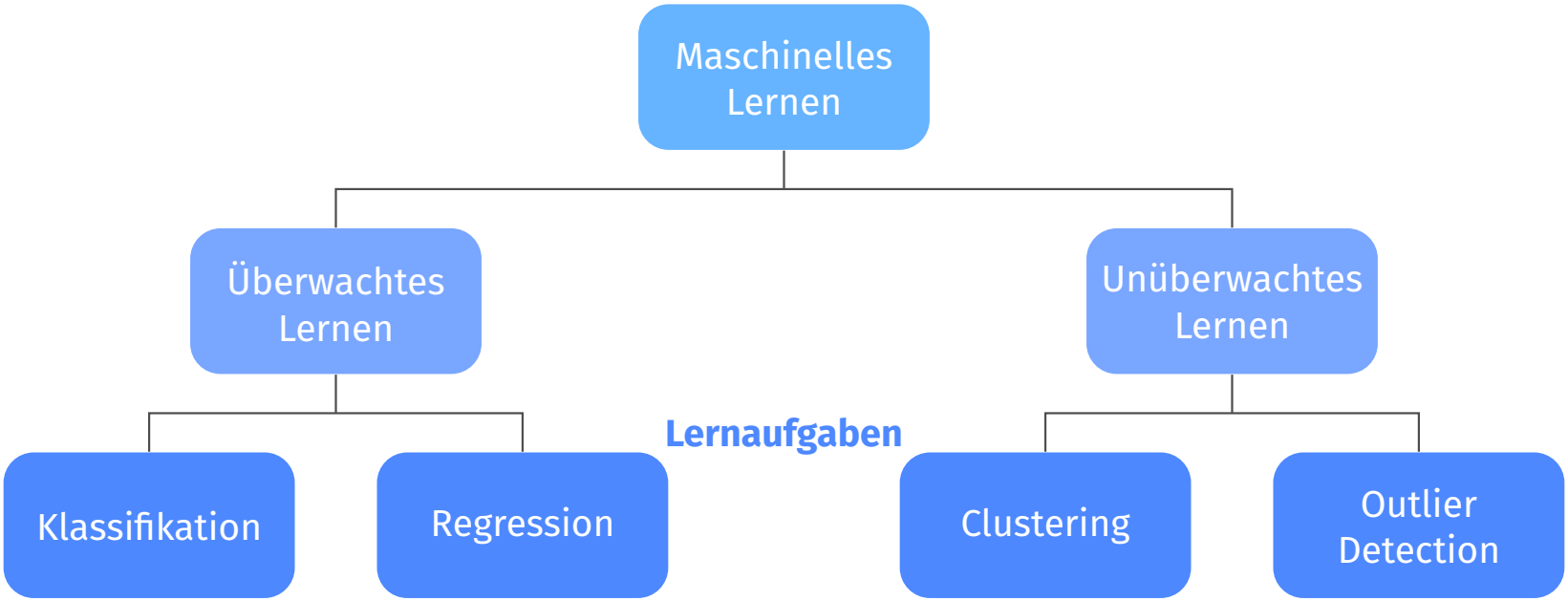
Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



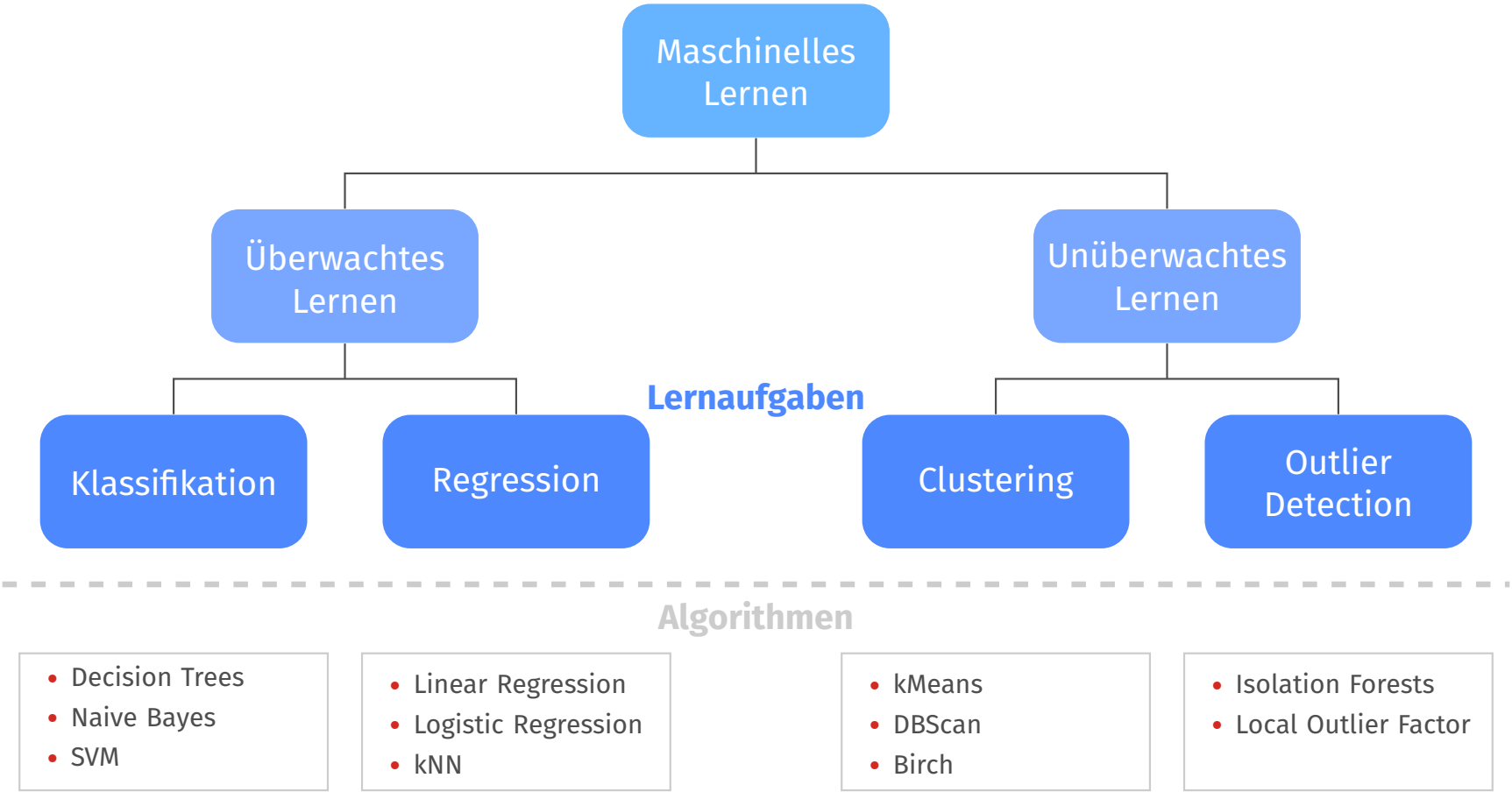
Maschinelles Lernen ist Teilgebiet der künstlichen Intelligenz



Maschinelles Lernen ist Teilgebiet der künstlichen Intelligenz



## Maschinelles Lernen ist Teilgebiet der künstlichen Intelligenz



**Lernaufgaben** definieren Ein- und Ausgabe, sowie das Ziel der Modellierung, z.B.

> "Entscheide für einen Text **x** ob er zur Klasse *Spam* oder zur Klasse *KeinSpam* gehört."

**Lernaufgaben** definieren Ein- und Ausgabe, sowie das Ziel der Modellierung, z.B.

> "Entscheide für einen Text **x** ob er zur Klasse *Spam* oder zur Klasse *KeinSpam* gehört."

Eingabedaten werden typischerweise in einen Merkmalsraum $\mathcal{X}$ der Dimension $d$ abgebildet

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$$

Die Ausgabemenge $\mathcal{Y}$ kann eine Menge von Klassen oder eine reelle Zahl sein, z.B.

$$\mathcal{Y} = \{\text{Spam}, \text{KeinSpam}\}$$

Das Ziel besteht darin, eine Funktion (Modell) $f : \mathcal{X} \to \mathcal{Y}$ zu lernen, mit

$$f(\mathbf{x}) = \begin{cases} +1, & \text{falls } \mathbf{x} \text{ Spam Nachricht} \\ -1, & \text{sonst} \end{cases}$$

Das Ziel besteht darin, eine Funktion (Modell) $f : \mathcal{X} \to \mathcal{Y}$ zu lernen, mit

$$f(\mathbf{x}) = \begin{cases} +1, & \text{falls } \mathbf{x} \text{ Spam Nachricht} \\ -1, & \text{sonst} \end{cases}$$

Bei der binären Klassifikation wird häufig $\mathcal{Y} = \{-1, +1\}$ gewählt.

Das Ziel besteht darin, eine Funktion (Modell) $f : \mathcal{X} \to \mathcal{Y}$ zu lernen, mit

$$f(\mathbf{x}) = \begin{cases} +1, & \text{falls } \mathbf{x} \text{ Spam Nachricht} \\ -1, & \text{sonst} \end{cases}$$

Bei der binären Klassifikation wird häufig $\mathcal{Y} = \{-1, +1\}$ gewählt.

Für die Regression gilt $\mathcal{Y} = \mathbb{R}$.

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

| | $d$ Merkmale | | | | |
|---|---|---|---|---|---|
| ID | $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
| 1 | 0 | 0 | $\cdots$ | 1 | **-1** |
| 2 | 0 | 1 | $\cdots$ | 1 | **+1** |
| 3 | 1 | 0 | $\cdots$ | 1 | **-1** |

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

| | $d$ Merkmale | | | | |
|---|---|---|---|---|---|
| ID | $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
| 1 | 0 | 0 | $\cdots$ | 1 | **-1** |
| 2 | 0 | 1 | $\cdots$ | 1 | **+1** |
| 3 | 1 | 0 | $\cdots$ | 1 | **-1** |

Beispiel $\mathbf{x}_2 = (x_{a_1}, x_{a_2}, \ldots, x_{a_4}, y)$
$= (0, 1, \ldots, 1, \mathbf{+1})$

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

| | $d$ Merkmale | | | | |
|---|---|---|---|---|---|
| ID | $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
| 1 | 0 | 0 | $\cdots$ | 1 | **-1** |
| 2 | 0 | 1 | $\cdots$ | 1 | **+1** |
| 3 | 1 | 0 | $\cdots$ | 1 | **-1** |

Beispiel $\mathbf{x}_2 = (x_{a_1}, x_{a_2}, \ldots, x_{a_4}, y)$
$= (0, 1, \ldots, 1, \mathbf{+1})$

- Beispiele werden auch *examples* oder *instances* genannt
- Merkmale (engl. *features*) werden auch *attributes* oder *Variablen* (Statistik) bezeichnet

# ML - Modell Training

| $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
|---|---|---|---|---|
| 0 | 0 | $\cdots$ | 1 | -1 |
| 0 | 1 | $\cdots$ | 1 | +1 |
| 1 | 0 | $\cdots$ | 1 | -1 |

Trainingsdaten

Algorithmus/
Optimierung

$$f : \mathcal{X} \to \mathcal{Y}$$

Modell

# ML - Modell Training

| $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
|---|---|---|---|---|
| 0 | 0 | $\cdots$ | 1 | -1 |
| 0 | 1 | $\cdots$ | 1 | +1 |
| 1 | 0 | $\cdots$ | 1 | -1 |

Trainingsdaten

Algorithmus/
Optimierung

$$f : \mathcal{X} \to \mathcal{Y}$$

Modell

| $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
|---|---|---|---|---|
| 1 | 1 | $\cdots$ | 0 | ? |

Neue Daten $\mathbf{x}'$,
$y$ unbekannt

| $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
|---|---|---|---|---|
| 0 | 0 | $\cdots$ | 1 | -1 |
| 0 | 1 | $\cdots$ | 1 | +1 |
| 1 | 0 | $\cdots$ | 1 | -1 |

Trainingsdaten

Algorithmus/
Optimierung

$f : \mathcal{X} \to \mathcal{Y}$

Modell

| $a_1$ | $a_2$ | $\cdots$ | $a_d$ | $y$ |
|---|---|---|---|---|
| 1 | 1 | $\cdots$ | 0 | ? |

Neue Daten $\mathbf{x}'$,
$y$ unbekannt

Vorhersage

$\hat{y} = f(\mathbf{x}')$

**Wie kommen wir zu einer Datenanalyse in einem Unternehmen?**

- Daten meist in verschiedenen Fachabteilungen verteilt

**Wie kommen wir zu einer Datenanalyse in einem Unternehmen?**

- Daten meist in verschiedenen Fachabteilungen verteilt
- Unterschiedliche Datenformate / Datenbanken

**Wie kommen wir zu einer Datenanalyse in einem Unternehmen?**

- Daten meist in verschiedenen Fachabteilungen verteilt
- Unterschiedliche Datenformate / Datenbanken
- verschiedene Granularitäten / unterschiedliche Qualität

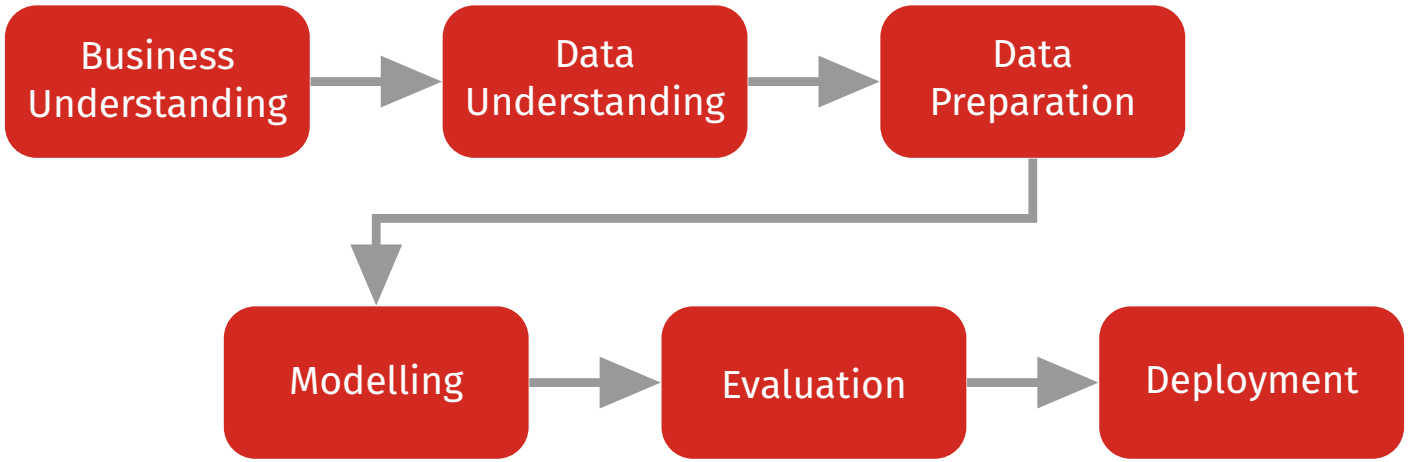
**Wie kommen wir zu einer Datenanalyse in einem Unternehmen?**

- Daten meist in verschiedenen Fachabteilungen verteilt
- Unterschiedliche Datenformate / Datenbanken
- verschiedene Granularitäten / unterschiedliche Qualität

**Standardisierter Prozeß** für die Datenanalyse notwendig.

**CRISP-DM** - **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

- Standardisierter Prozess zur Datenanalyse
- Initiative von Data Mining Anbietern (IBM, SPSS), Beratern (Cap Gemini,..) und Anwendern (Daimler AG,...)
- Daten Mining Prozess in Phasen zerlegt



Prozeß erlaubt Iterationen/Rücksprünge

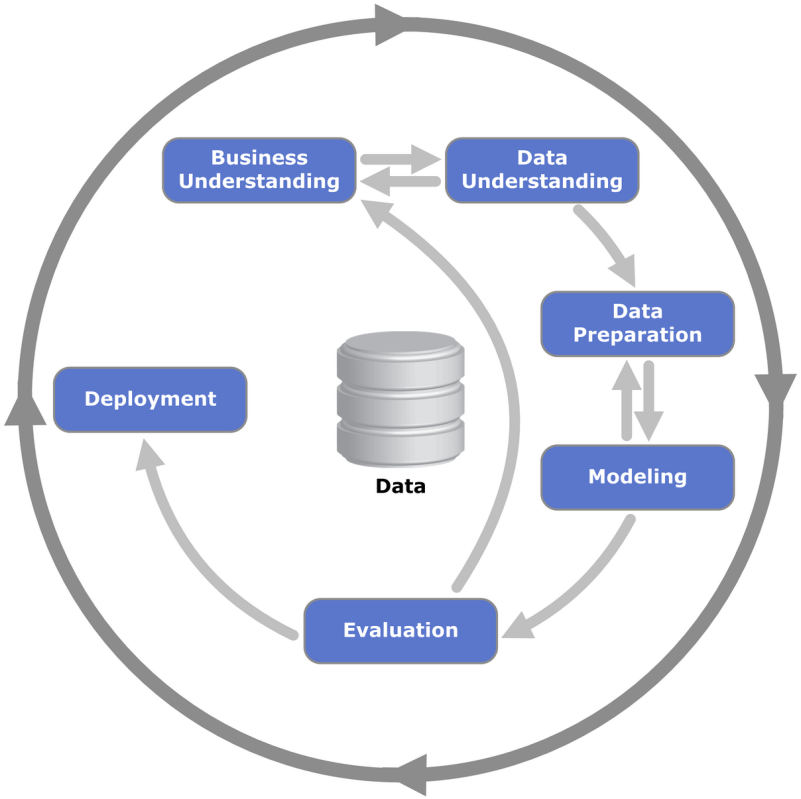



**Abbildung:** Das CRISP-DM Phasen-Modell

# CRISP-DM - Die Phasen

**Business Understanding**:

- Definition des Geschäftsziels
- Definition des Analyse-Ziels
- Festlegen von Erfolgskriterien

**Data Understanding**

- Datenerhebung verstehen, Merkmale verstehen
- Datenqualität untersuchen

**Data Preparation**

- Daten sammeln und zusammenführen (Daten-Silos)
- Normalisieren, *Data Cleaning*
- ggf. neue Merkmale definieren *Feature Engineering*

**Data Preparation**

- Daten sammeln und zusammenführen (Daten-Silos)
- Normalisieren, *Data Cleaning*
- ggf. neue Merkmale definieren *Feature Engineering*

*Data Preparation* nimmt bis zu 90% des Aufwandes ein

**Modelling**

- Modell-Wahl / *Algorithm Selection*
- Modell Training
- Iterativer Prozess, Kreuzvalidierung(!)

**Evaluation**

- Evaluation des Modells auf Test-Daten
- Fehlermaß ggf. anwendungsspezifisch
- Interpretation des Modells

**Deployment**

- Modell in Geschäftsprozesse integrieren
- Offline-Vorhersage
- Integration in online-Prozesse

## Weiterentwicklung von CRISP-DM

- Keine weitere Entwicklung
- Immer noch am meisten verbreitetes Prozess-Modell für Data Mining
- IBM veröffentlichte 2015 *Analytics Solutions Unified Method for Data Mining* - ASUM-DM
- ASUM-DM erweitert/verfeinert CRISP-DM in einigen Bereichen

## Graphische Tools

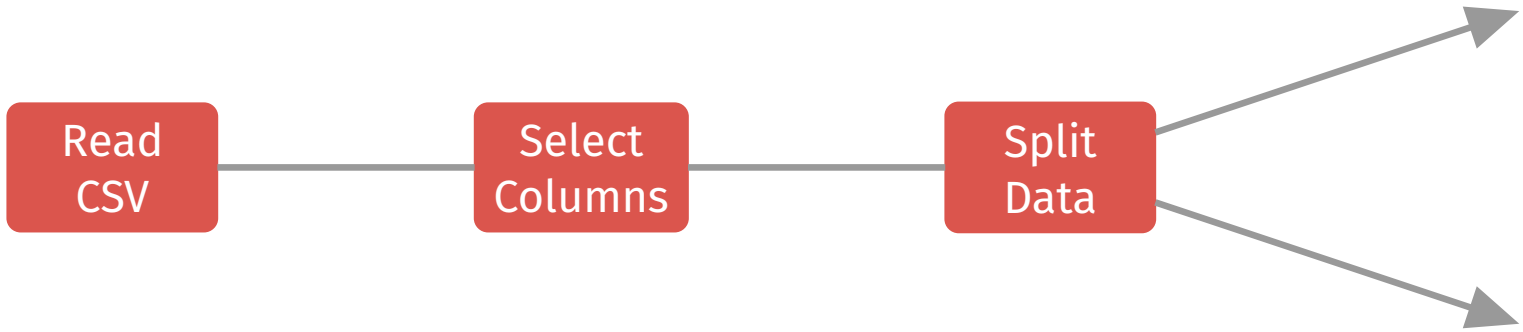
- RapidMiner, `http://rapidminer.com`
- Knime, `http://www.knime.com`
- R-Studio, `http://rstudio.com`
- WEKA, MOA, `http://www.cs.waikato.ac.nz/ml/weka`

## Programmiersprachen

- Julia, `http://julialang.org`
- Python mit Pandas, SciKit Learn
  `http://scikit-learn.org`
- R, `http://www.r-project.org`

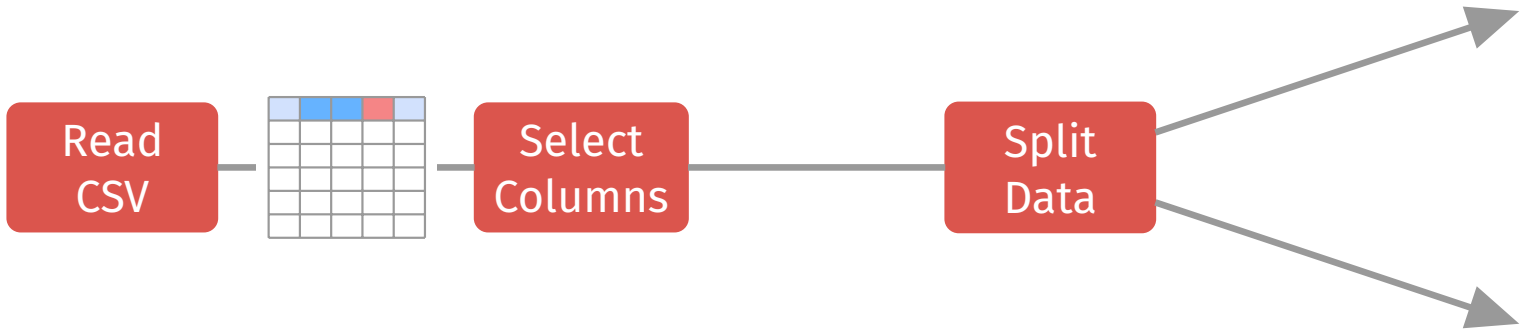
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



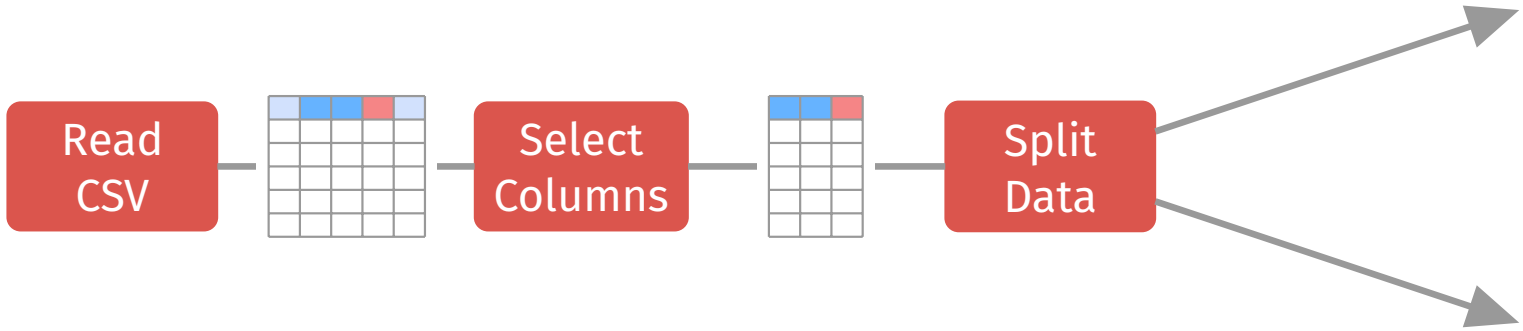
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



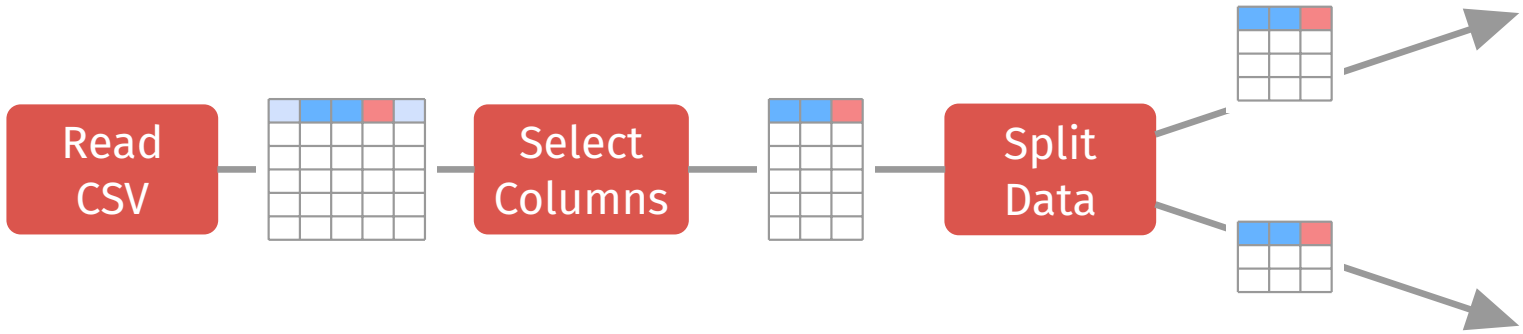
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



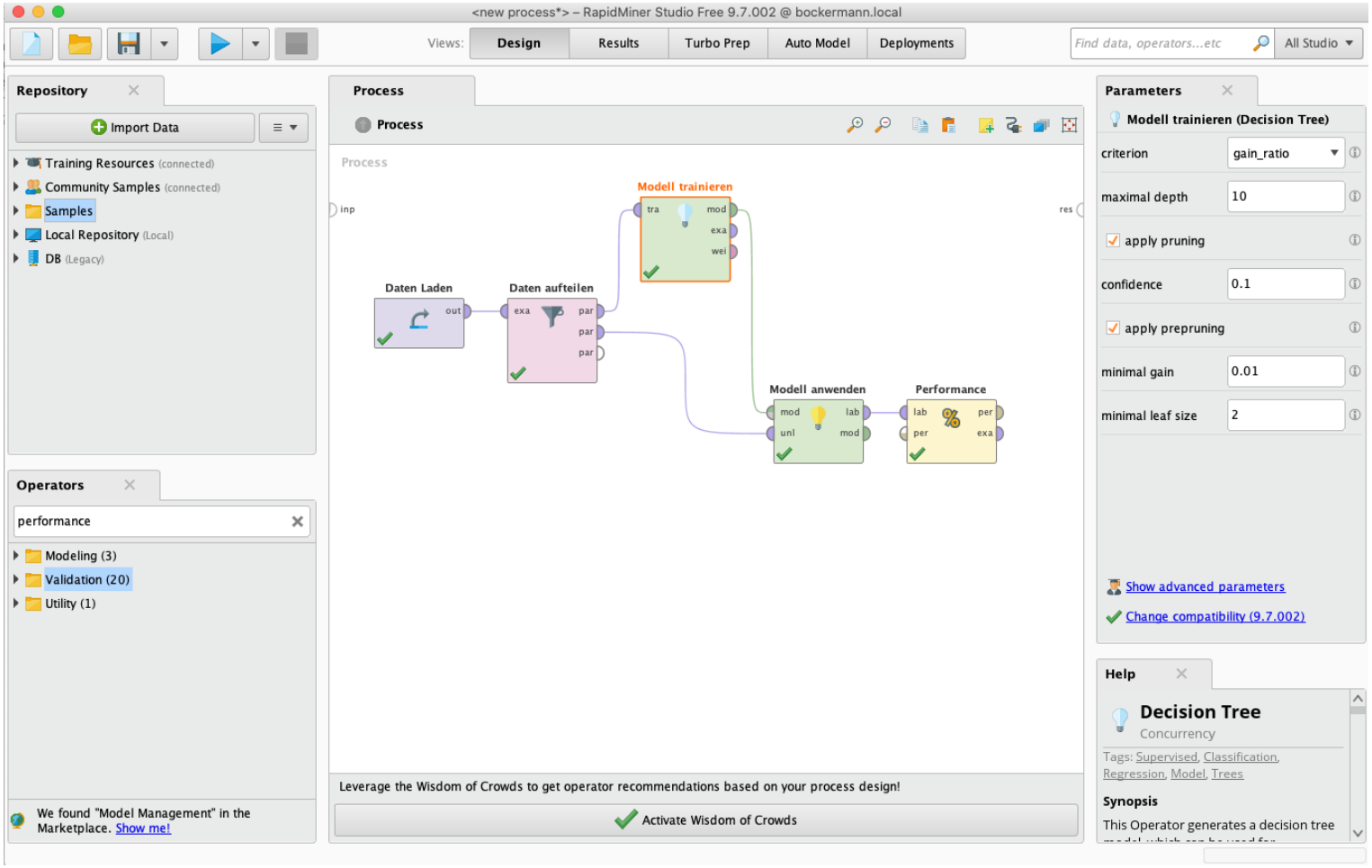



**Abbildung:** Die graphische Schnittstelle von RapidMiner.

# Prozesse werden als Graph mit vordefinierten Operator-Bausteinen gebaut

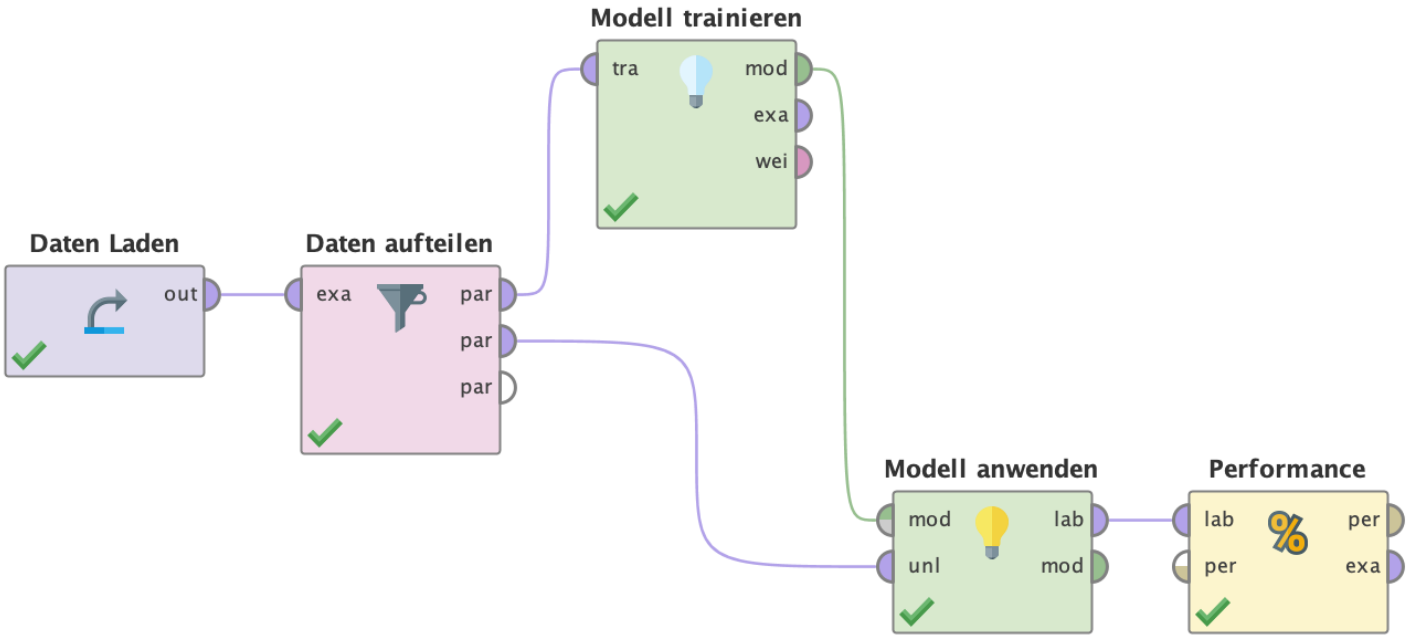



**Abbildung:** Ein Prozeß als Graph in RapidMiner.

RapidMiner wurde als OpenSource Tool am Lehrstuhl für künstliche Intelligenz der TU Dortmund entwickelt

- Prozess-Definition für ETL, Modellierung und Auswertung
- Einfaches Inspizieren / Exploration von Daten
- Enterprise Version für Unternehmen verfügbar
- Marktplatz mit Vielzahl von Erweiterungen
- *Wisdom of the crowds* Ansatz für schnellen Start

# KNIME ist ebenfalls ein graphisches Tool für Prozess-Design

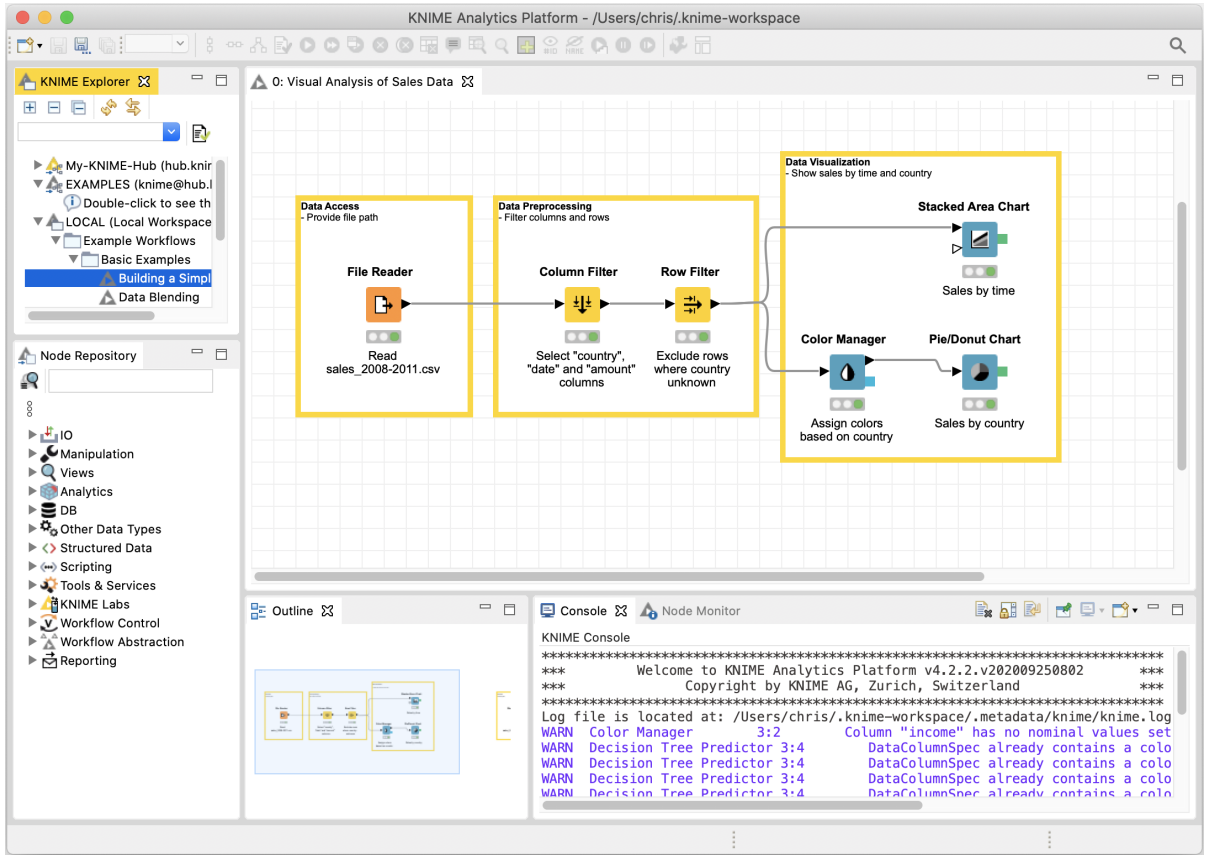



**Abbildung:** Die graphische Schnittstelle von KNIME.

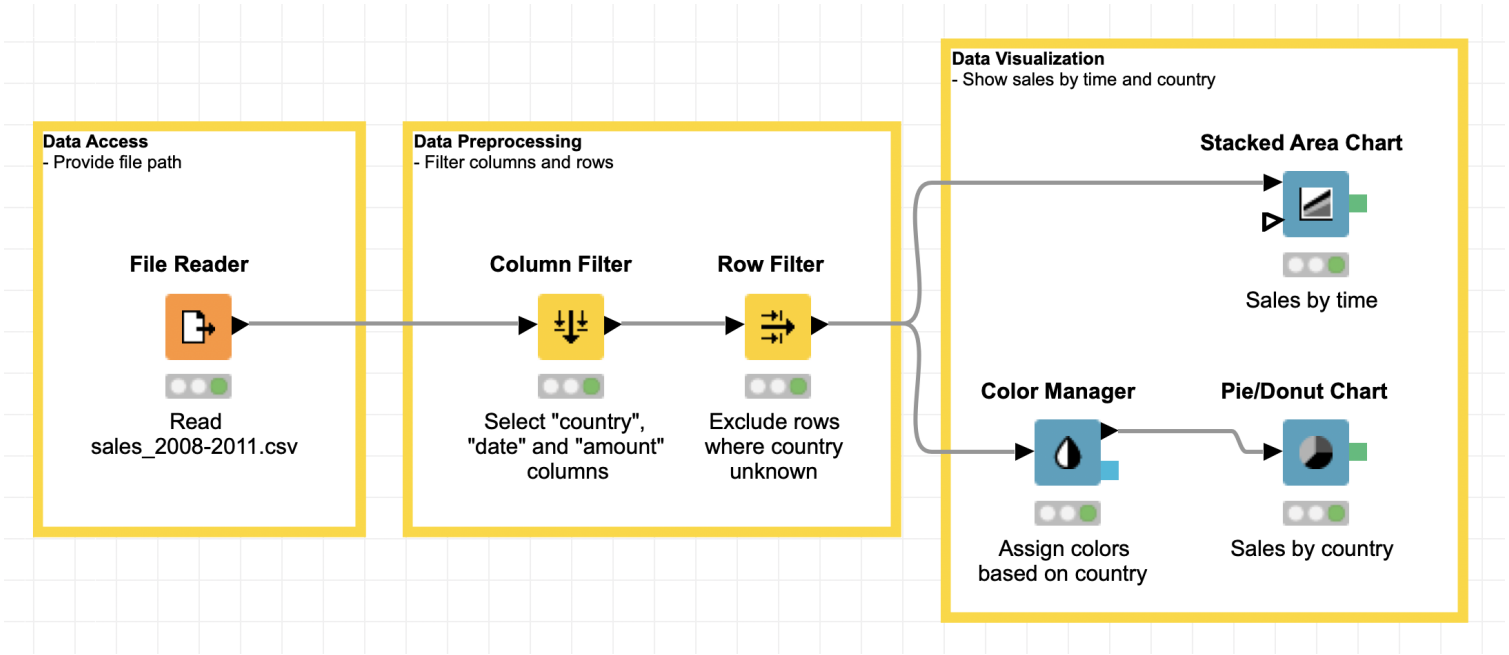



**Abbildung:** Ein Prozess zur Visualisierung mit KNIME.

# Programmiersprache **R** für Statistik Aufgaben

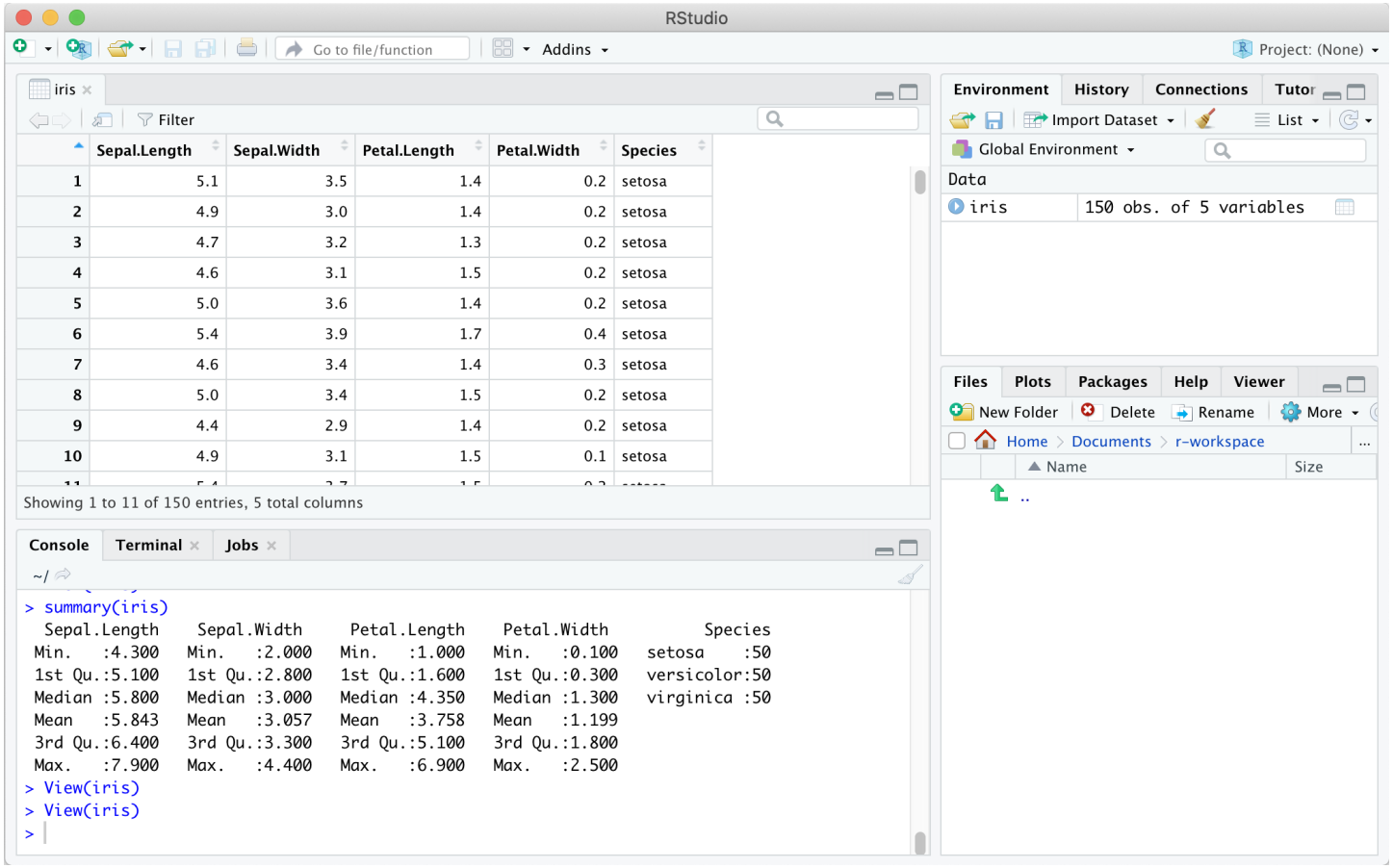



**Abbildung:** RStudio Umgebung für die Sprache R.

```python
import pandas as pd

# read data from csv
table = pd.read_csv('daten.csv')

# select columns
table = table[['col1', 'col2', 'colY']]

# split data into two sub-tables
# (first 50 row and remaining rows)
tab1 = table[:50]
tab2 = table[50:]
```

**Warum wird im Data Science Kurs Python benutzt?**

- Leicht erlernbare Sprache
- Universell einsetzbar
- Hersteller unabhängig
- Weit verbreitete Sprache für Rapid Prototyping

Viele etablierte Data Science Module:

- NumPy
- Pandas
- SciKit-Learn

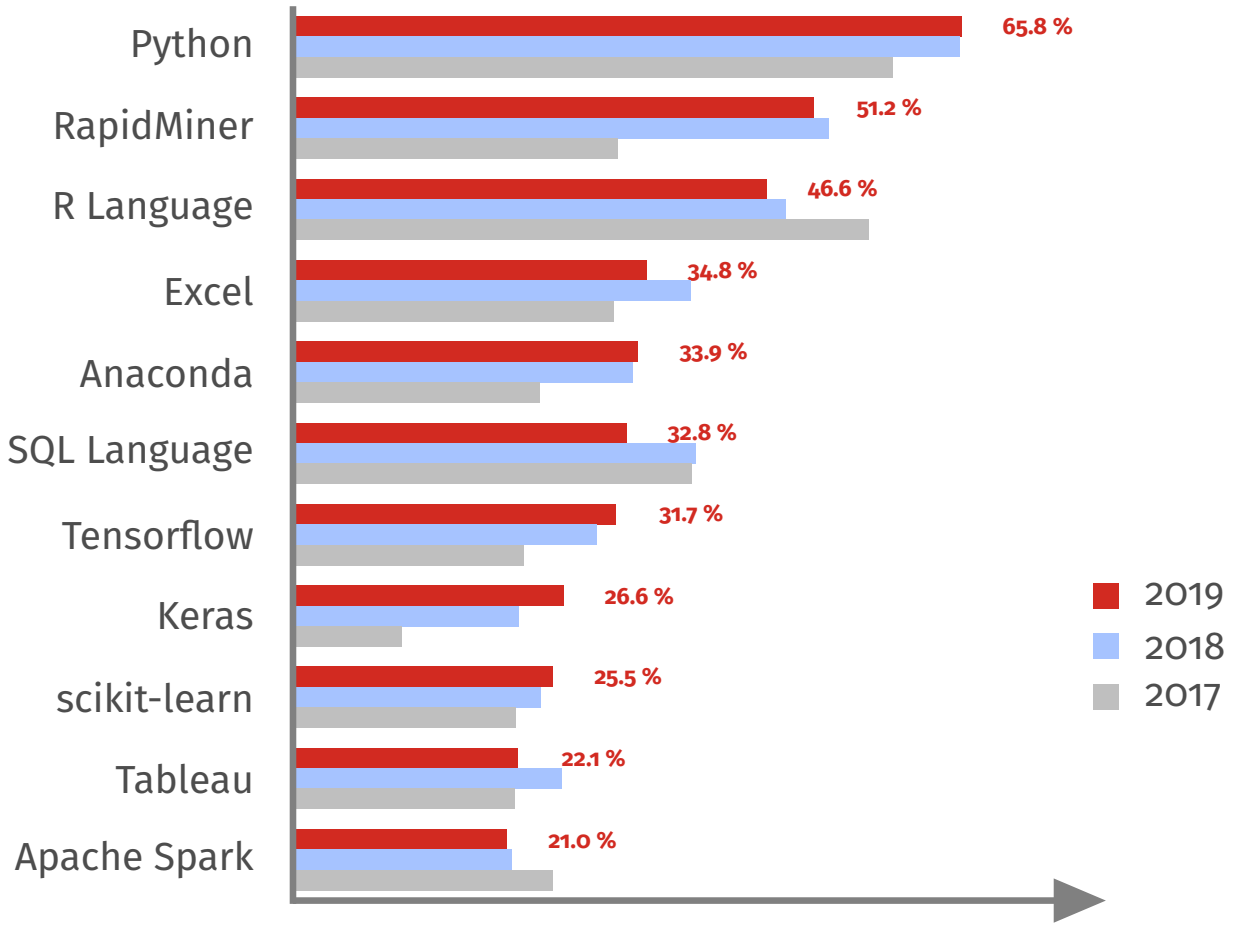



**Abbildung:** KDNuggets Umfrage der beliebtesten DataScience Tools

## Python ist eine Skript-Sprache

Datei `HelloWorld.py`:

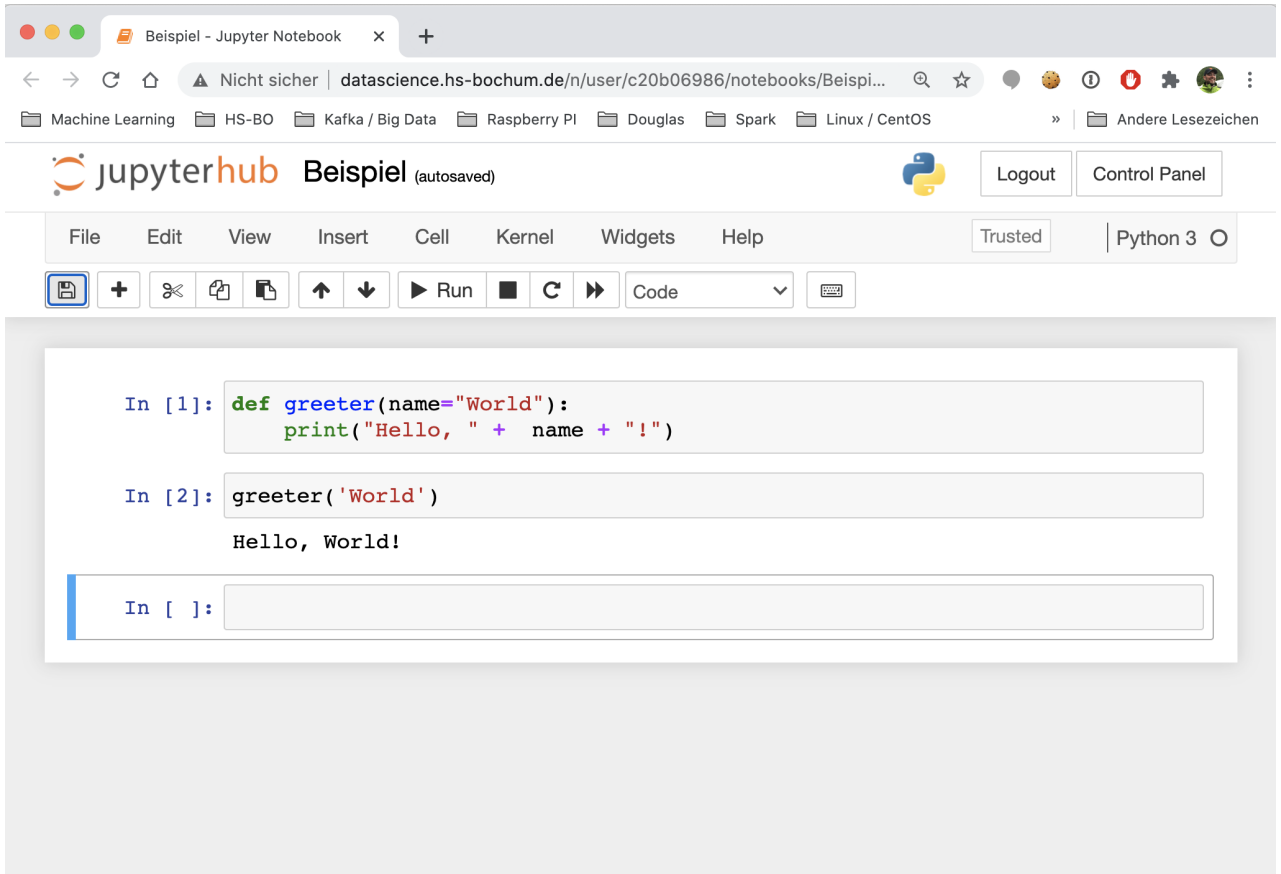
```
# Ein Beispiel fuer eine einfache Funktion
#
def greeter(name="World"):
  print("Hello, " + name + "!")

greeter('World')
```

Starten eines Skripts, z.B. mit

```
python3 HelloWorld.py
```

# Jupyter Notebooks bieten Python-Umgebung im Browser:



**Jupyter Notebooks bieten Python-Umgebung im Browser**

`https://datascience.hs-bochum.de/`

Demo: Python Jupyter Notebook

# Zusammenfassung

- Beispiele für Data Science / ML
- Überblick ML: Lernaufgaben, Datenrepräsentation, Modell-Training
- Prozess-Modell für Datenanalyse (CRISP-DM)
- Überblick über Software / Tools für Data Science und ML