

# DATA SCIENCE 1

VORLESUNG 1 - EINFÜHRUNG

PROF. DR. CHRISTIAN BOCKERMANN

HOCHSCHULE BOCHUM

SOMMERSEMESTER 2021

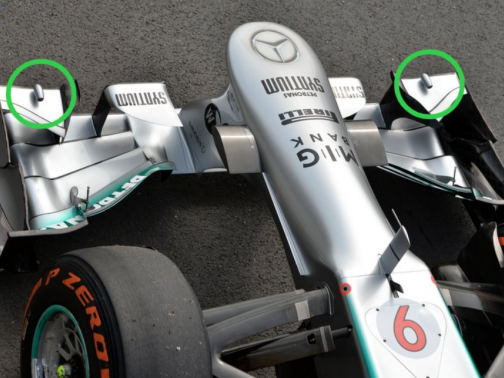
- 1 Was ist Data Science?
- 2 Maschinelles Lernen + CRSIP-DM
- 3 Software und Tools
- 4 Zusammenfassung





Data Science?





- 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>

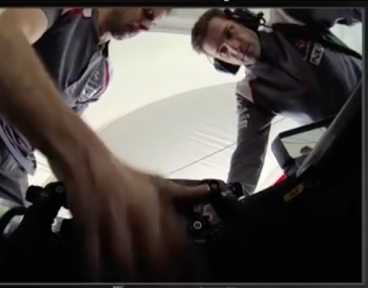
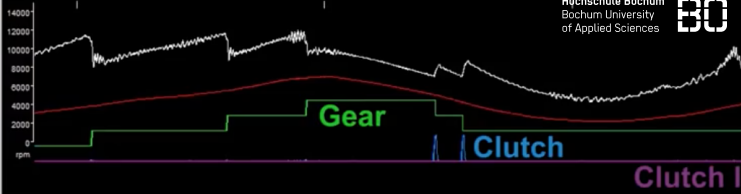
- 240 Sensoren, teilw. hohe Sampling Raten
- ATLAS Funksystem für Echtzeitdaten
  - Regelüberwachung (FIA)
  - Datenanalyse der Rennställe (Echtzeit)
- > 2 GB Daten pro Auto *pro Runde*

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

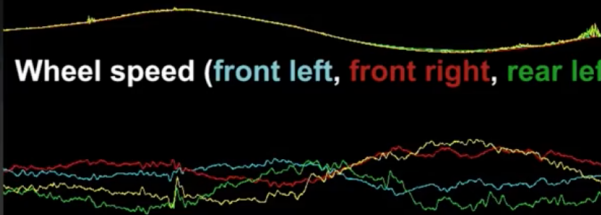


7430rpm  
102.3kgh  
3  
-0.000mm  
0.0%  
846rpm  
839rpm  
855rpm  
840rpm  
-4.23mm  
3.77mm  
-7.24°  
0.32°  
-14.83°  
22.90%  
-0.02bar  
NO

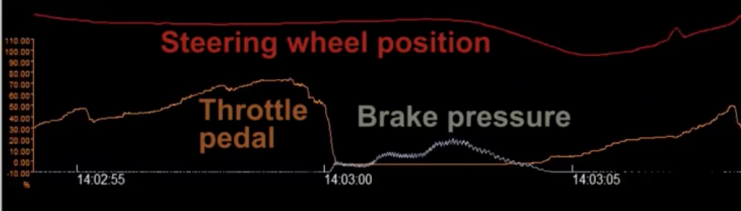
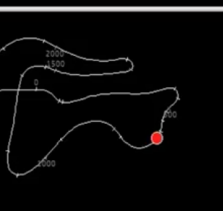
14.03.11./58



Wheel speed (front left, front right, rear left, rear right)



Dampers (front left, front right, rear left, rear right)



Steering wheel position

Throttle pedal

Brake pressure

7430rpm  
102.2kph  
3  
-0.000mm  
0.0%  
846rpm  
839rpm  
855rpm  
840rpm  
-4.23mm  
3.77mm  
-7.24°  
0.32°  
-14.83°  
22.90%  
-0.02bar  
NO

14.03.11./58



Video auf YouTube:

<https://youtu.be/0sR5oCI fXDI>

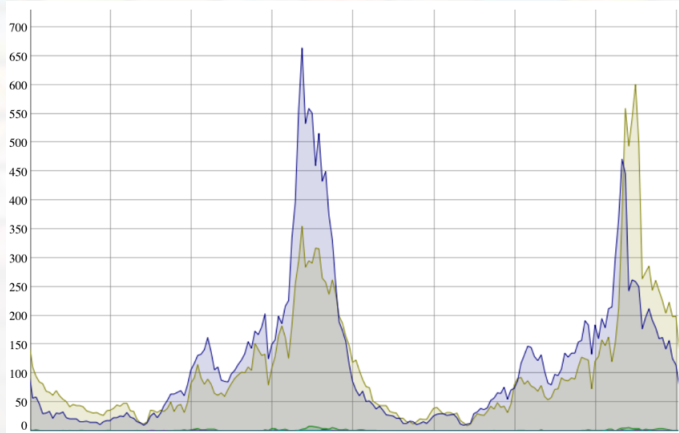




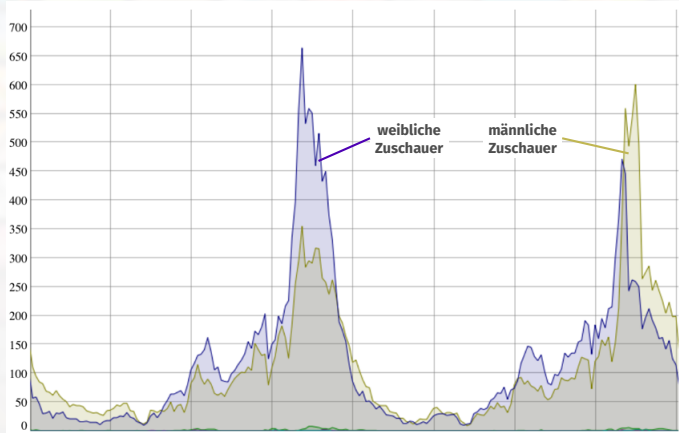


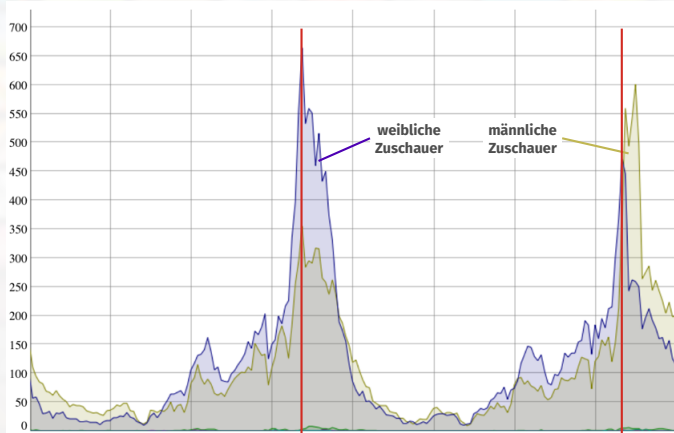
A silhouette of a person in a suit holding a smartphone up to take a picture. The background is a large wall composed of many small, square video screens, each displaying a different, colorful image. The overall scene is dimly lit, with the screens providing the primary light source.

Data Science?









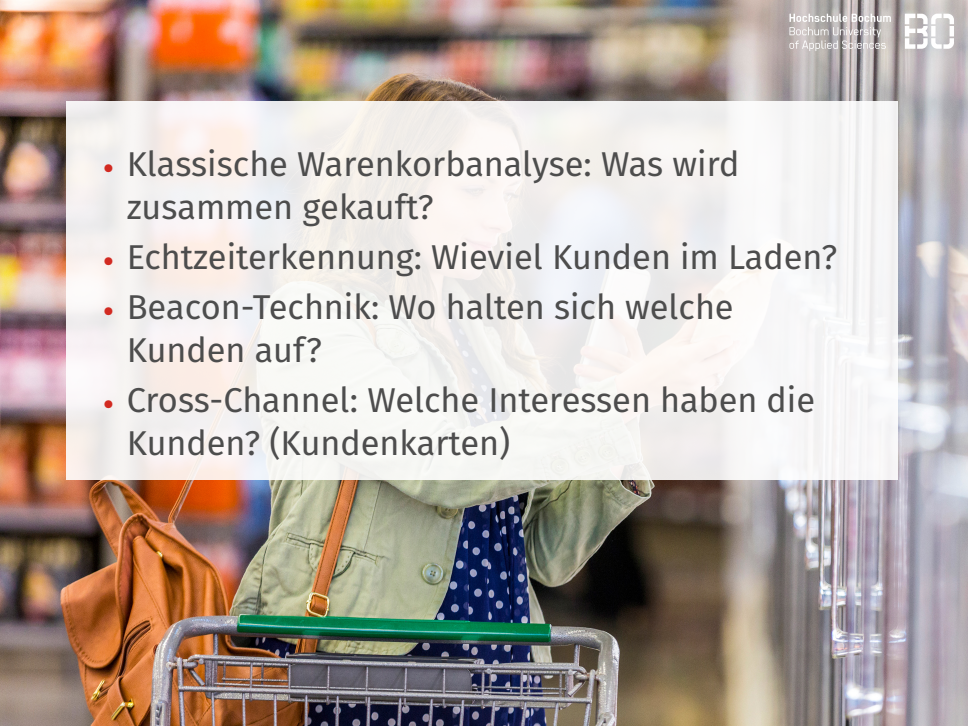
Mi, 14.3.  
20:15 Uhr

Do, 15.3.  
20:15 Uhr



A young woman with long brown hair, wearing a light green jacket over a blue polka-dot top, is standing in a supermarket aisle. She is holding a white smartphone in her right hand and a brown paper bag in her left hand, looking at the phone. She has a brown shoulder bag and is pushing a shopping cart. The background shows shelves of products, slightly out of focus.

Data Science?

- 
- A woman with brown hair, wearing a light green jacket over a blue polka-dot dress, is pushing a silver shopping cart with a green handle. She is looking down at a smartphone in her hands. The background is a blurred supermarket aisle with shelves of products.
- 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 x +

← → ↻ 🏠 🔒 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/](http://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](http://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/](http://www.easycosmetic.de/) ▼  
Markenkosmetik reduziert & schnell · Trusted-Shops Garantie · Retour gratis

**Boss Bottled für Herren bei Flaconi kaufen - flaconi.de**  
[Anzeige](#) [www.flaconi.de/](http://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 x +

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](http://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/](http://www.easycosmetic.de/) ▼  
Markenkosmetik reduziert & schnell - Trusted-Shops Garantie - Retour gratis

**Boss Bottled für Herren bei Flaconi kaufen - flaconi.de**  
[Anzeige](#) [www.flaconi.de/](http://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	Schoki
4	948.33	High	10 €	X
3	402.25	Mid	5 €	
5	1210.89	High	20 €	X
4	423.43	Mid	5 €	X
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	Katzen
X	X
X	X
X	X
	X
	X

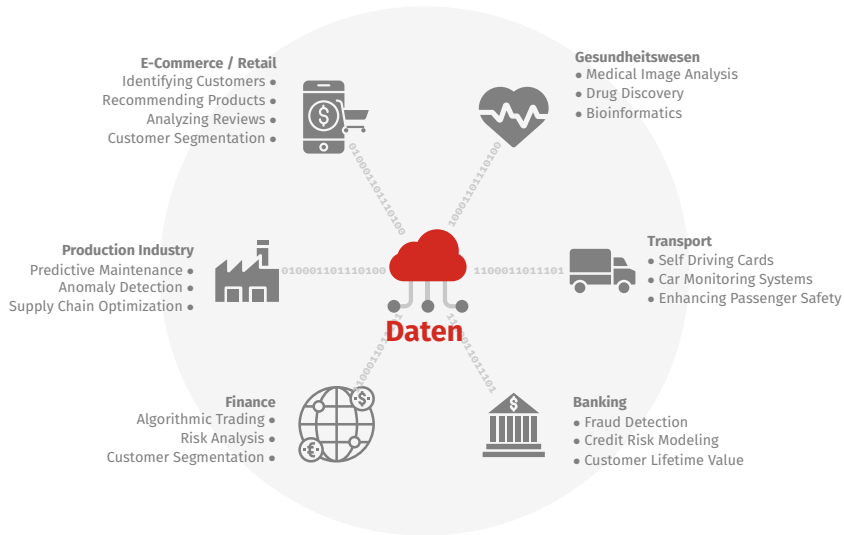


## Was wissen wir über die Besucher/Kunden?

# 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 €

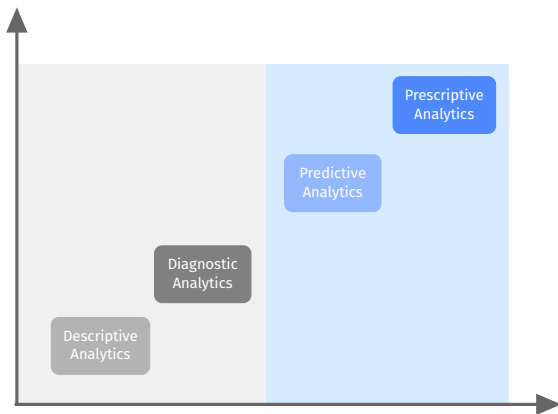
Schoki	Katzen	Hugo B
X	X	X
		X
X	X	X
X	X	X
	X	
	X	



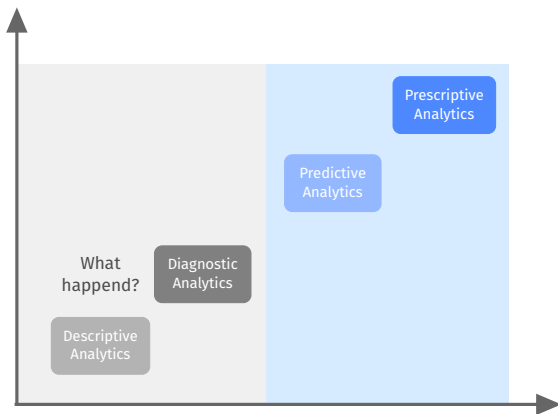




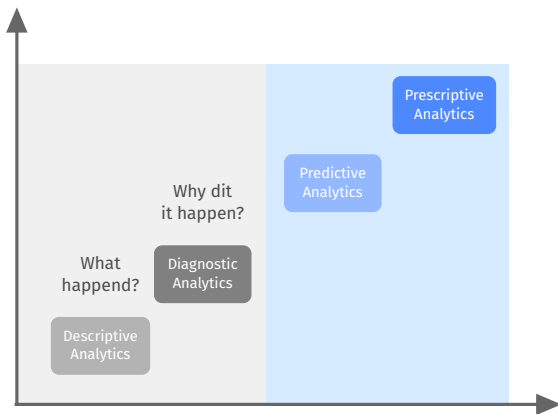
## Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



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

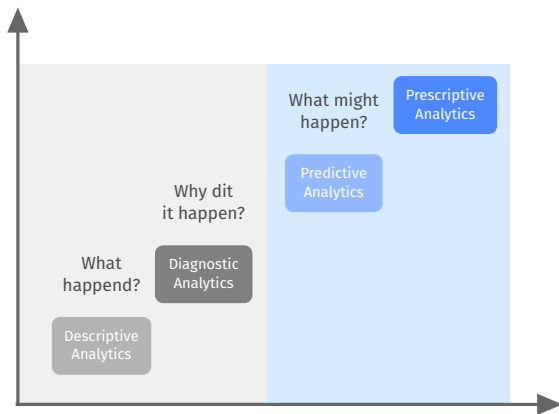


## Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*

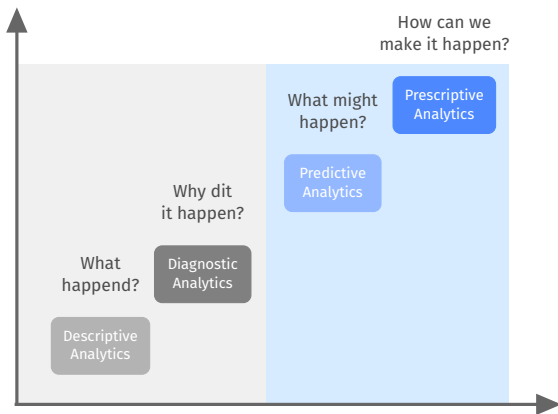




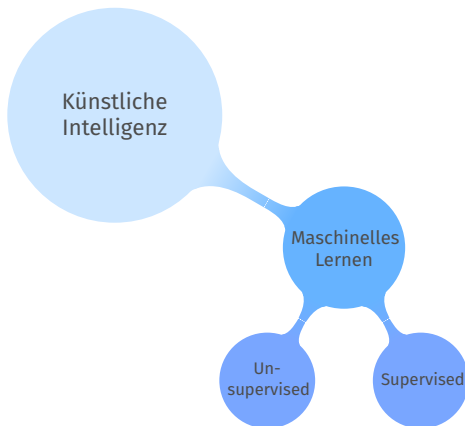
## Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



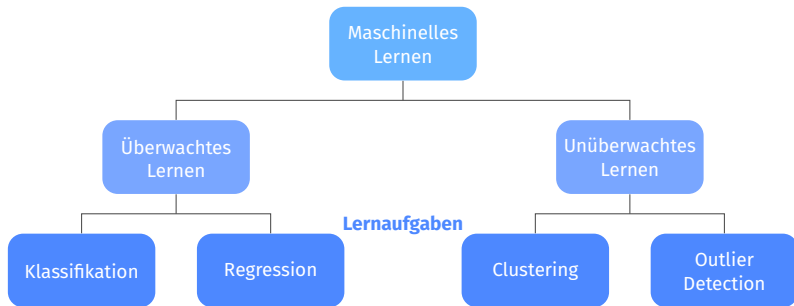
## Data Science bzw. *Advanced Analytics* als Weiterentwicklung von *Business Intelligence*



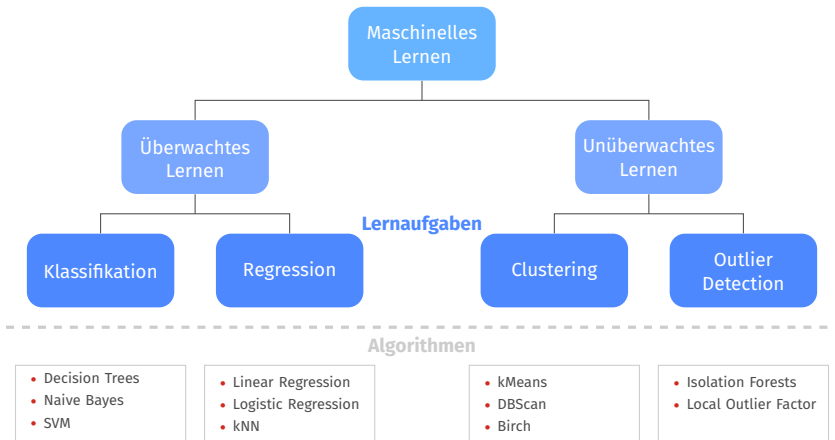
Maschinelles Lernen ist Teilgebiet der **künstlichen Intelligenz**



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  $\mathbf{x}$  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)  $f : \mathcal{X} \rightarrow \mathcal{Y}$  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)  $f : \mathcal{X} \rightarrow \mathcal{Y}$  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)  $f : \mathcal{X} \rightarrow \mathcal{Y}$  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	$a_1$	$a_2$	...	$a_d$	$y$
1	0	0	...	1	-1
2	0	1	...	1	+1
3	1	0	...	1	-1

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

d Merkmale					
ID	$a_1$	$a_2$	...	$a_d$	$y$
1	0	0	...	1	-1
2	0	1	...	1	+1
3	1	0	...	1	-1

$$\begin{aligned}\text{Beispiel } \mathbf{x}_2 &= (x_{a_1}, x_{a_2}, \dots, x_{a_d}, y) \\ &= (0, 1, \dots, 1, +1)\end{aligned}$$

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

d Merkmale					
ID	$a_1$	$a_2$	...	$a_d$	$y$
1	0	0	...	1	-1
2	0	1	...	1	+1
3	1	0	...	1	-1

$$\begin{aligned}\text{Beispiel } \mathbf{x}_2 &= (x_{a_1}, x_{a_2}, \dots, x_{a_d}, y) \\ &= (0, 1, \dots, 1, +1)\end{aligned}$$

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

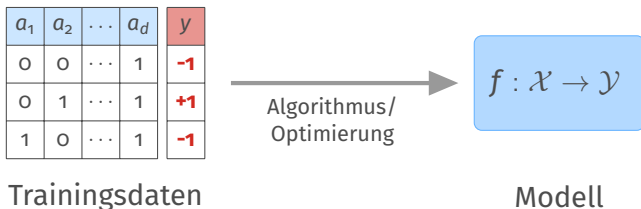
$a_1$	$a_2$	$\dots$	$a_d$	$y$
0	0	$\dots$	1	-1
0	1	$\dots$	1	+1
1	0	$\dots$	1	-1

Trainingsdaten

Algorithmus/  
Optimierung

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

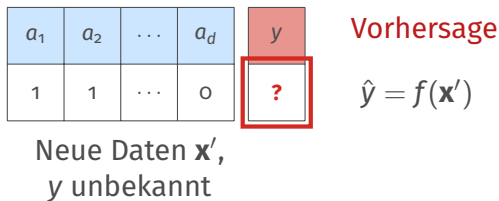
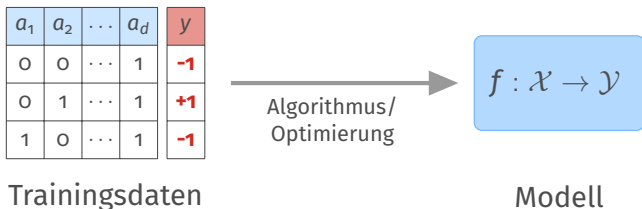
Modell



---

$a_1$	$a_2$	$\dots$	$a_d$	$y$
1	1	$\dots$	0	?

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





## 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

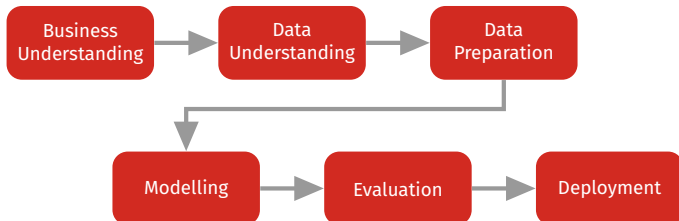
## Wie kommen wir zu einer Datenanalyse in einem Unternehmen?

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

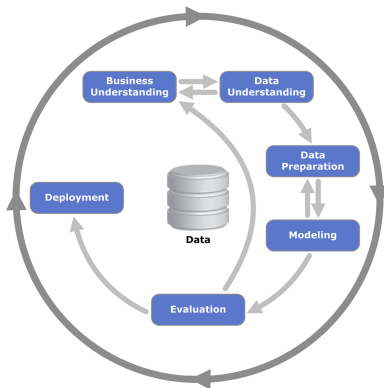
**Standardisierter Prozeß** für die Datenanalyse notwendig.

## CRISP-DM - **C**Ross **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

## **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>

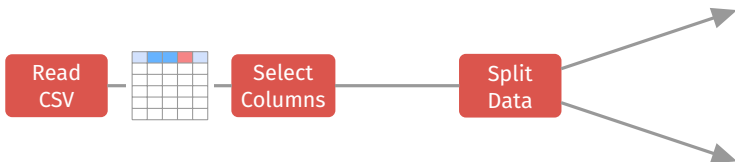
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

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



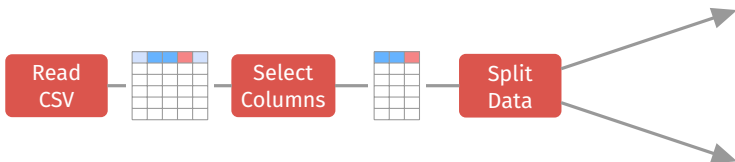
Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

- 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:

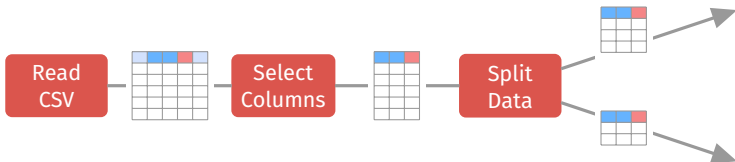
- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen





Werkzeuge um Prozesse mit graphischen Elementen zu entwerfen:

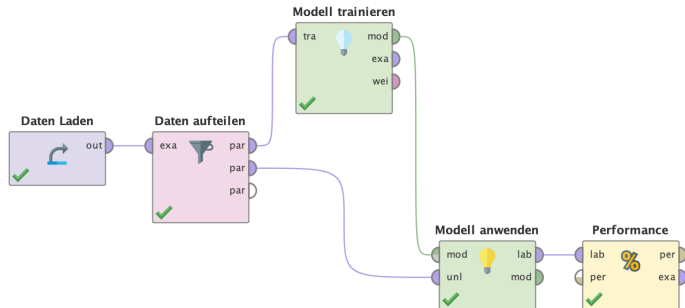
- Symbole für ausführbare Operationen
- Verbindungen zu Übergabe von Ergebnis-Objekten
- Einfache Start/Stopp Funktionen, Anhalten von Prozessen
- Möglichst ohne Programmierung auskommen



The screenshot displays the RapidMiner Studio Free 9.7.002 interface. The main workspace shows a workflow process with the following operators: 'Daten Laden' (Load Data), 'Daten aufteilen' (Split Data), 'Modell trainieren' (Train Model), 'Modell anwenden' (Apply Model), and 'Performance'. The 'Modell trainieren' operator is highlighted with a green box. The 'Parameters' panel on the right shows settings for the 'Modell trainieren (Decision Tree)' operator, including 'criterion' (gain\_ratio), 'maximal depth' (10), 'apply pruning' (checked), 'confidence' (0.1), 'apply prepruning' (checked), 'minimal gain' (0.01), and 'minimal leaf size' (2). The 'Help' panel at the bottom right provides information about the 'Decision Tree' operator, including its category (Supervised Classification, Regression, Model Trees) and a synopsis stating that it generates a decision tree.

**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

The screenshot displays the KNIME Analytics Platform interface with a workflow titled "Visual Analysis of Sales Data". The workflow consists of the following nodes:

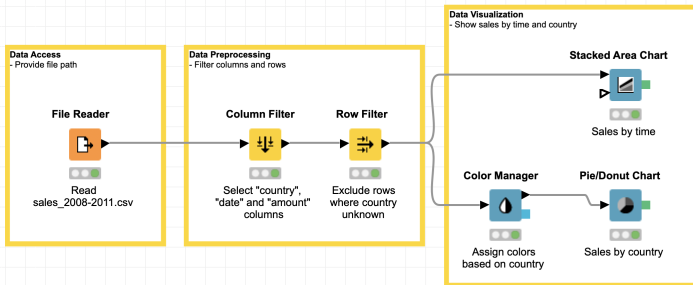
- Data Access:** "Provide file path" - File Reader (Read sales\_2008-2011.csv)
- Data Preprocessing:** "Filter columns and/or rows" - Column Filter (Select "country", "sales" and "amount" columns) and Row Filter (Exclude rows where country unknown)
- Data Visualization:** "Show sales by time and country" - Color Manager (Assign colors based on country), Stacked Area Chart (Sales by time), and Pie/Donut Chart (Sales by country)

The interface includes a KNIME Explorer on the left showing project structure, a Node Repository, an Outline view, and a Console window at the bottom displaying system messages and warnings.

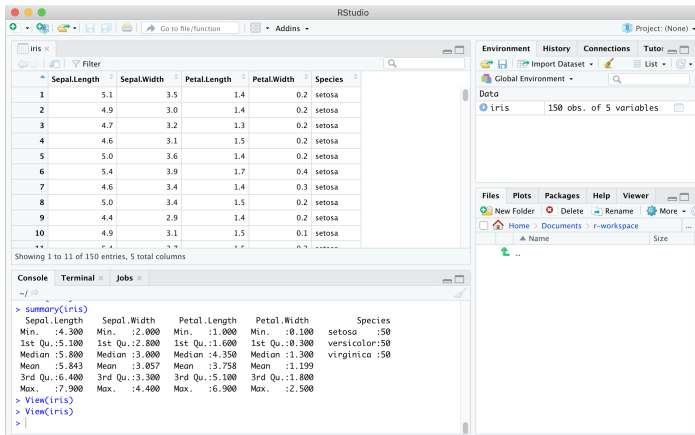
```

KNIME Console
=====
*** Welcome to KNIME Analytics Platform v4.2.2.v202009250800 ***
*** Copyright by KNIME AG, Zurich, Switzerland ***
=====
Log file is located at: /Users/chris/.knime-workspace/.metadata/knime/knime.log
WARN: Color Manager 3:2 Column "income" has no nominal values set
WARN: Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN: Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN: Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
WARN: Decision Tree Predictor 3:4 DataColumnSpec already contains a colo
  
```

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



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

Programmiersprache **R** für Statistik Aufgaben

The screenshot displays the RStudio interface with the following components:

- Environment:** Shows the loaded data frame 'iris' with 150 observations and 5 variables.
- Data View:** A table showing the first 10 rows of the 'iris' dataset.
- Console:** Contains the command `summary(iris)` and its output, which provides a statistical summary for each variable.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

```
> summary(iris)
  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa   :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500

> View(iris)
> View(iris)
>
```

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

```
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:]
```

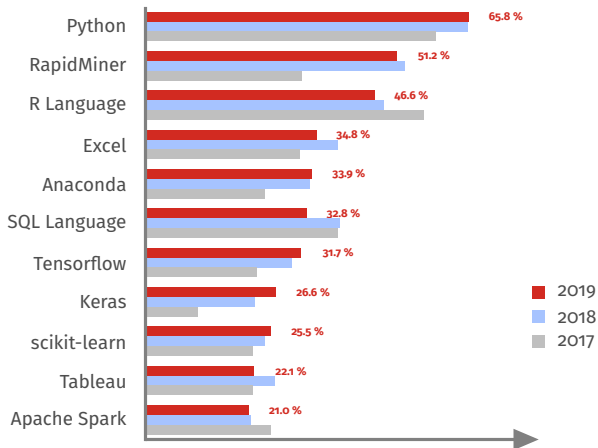


## Warum wird im Data Science Kurs **Python** benutzt?

- 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

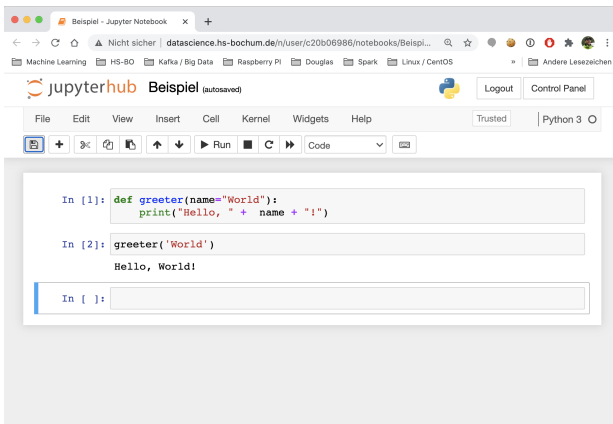
Datei HelloWorld.py:

```
# 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:



Beispiel - Jupyter Notebook

Nicht sicher | datascience.hs-bochum.de/n/user/c20b06986/notebooks/Beispi...

Machine Learning | HS-BO | Kafka / Big Data | Raspberry PI | Douglas | Spark | Linux / CentOS | \* | Andere Lesezeichen

jupyterhub Beispiel (autosaved) Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
In [1]: def greeter(name="World"):
        print("Hello, " + name + "!")

In [2]: greeter('World')
Hello, World!
```

In [ ]:

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

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

Demo: Python Jupyter Notebook

- 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