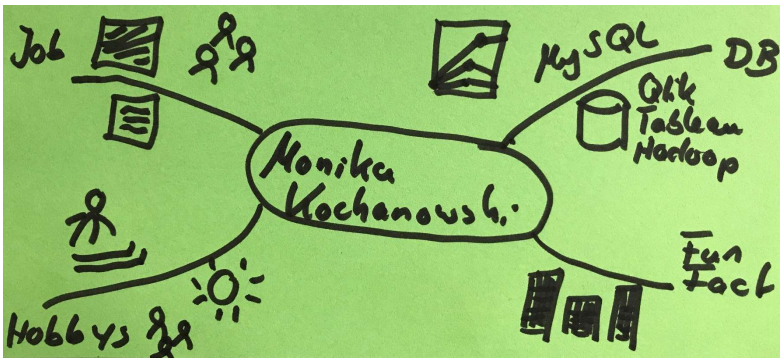
# Data Science
# 1. Teil – Einführung

Vorlesung an der DHBW Stuttgart, Prof. Dr. Monika Kochanowski



1

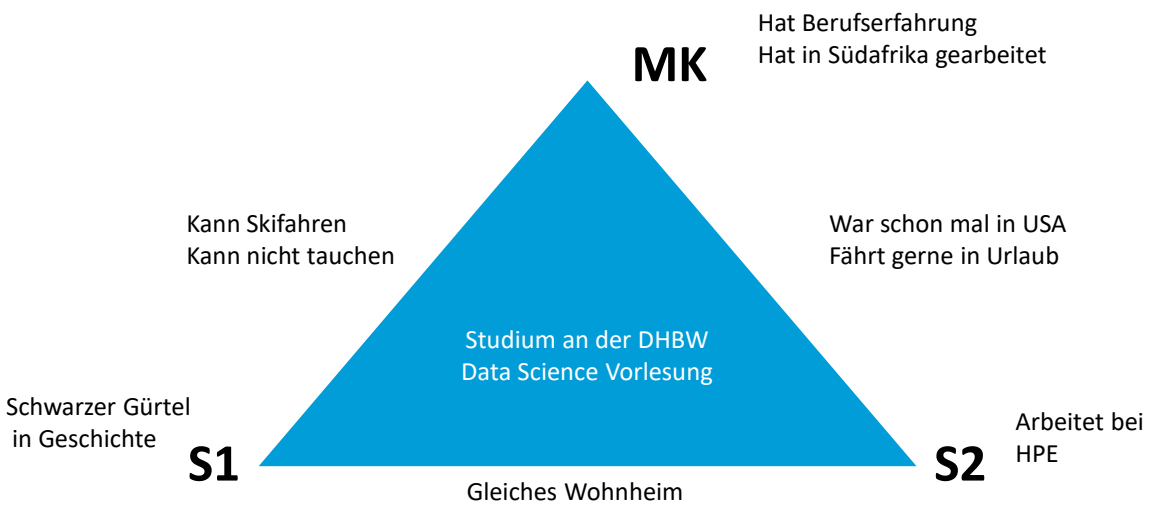
## Selbstvorstellung






2

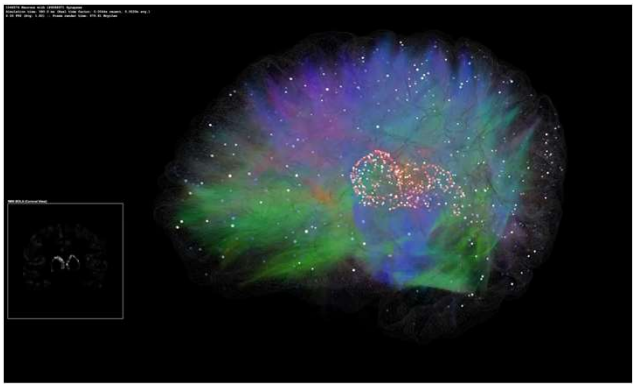
## Selbstvorstellung






3

## Inspiration



Simulation des Thalamokortikalen Systems mit
1 Mio. Neuronen und 189 Mio. Synapsen (www.digicortex.net)
Gehirn: ~ 86 Milliarden Neuronen und ~100 Billionen Synapsen

Lt. Com. Data
(Star Trek TNG)
1987

Frankenstein (1818)

Ava (Ex Machina) 2015

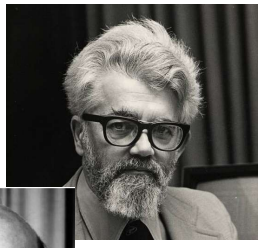


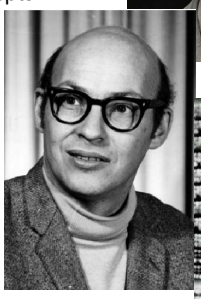
5

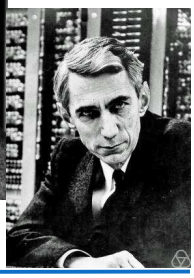
## Künstliche Intelligenz – 1956
## „A Proposal for the Dartmouth Summer Research Projekt on Artificial Intelligence”

1. Simulating higher functions of the human brain
2. Programming a computer to use general language ✓
3. Arranging hypothetical neurons in a manner so that they can form concepts
4. A way to determine and measure problem complexity ✓
5. Self-improvement ✓
6. Abstraction: Defined as the quality of dealing with ideas rather than events
7. Randomness and creativity

- ✓ – “mehr oder weniger” zum heutigen Stand erreicht


J. McCarthy


M. L. Minsky

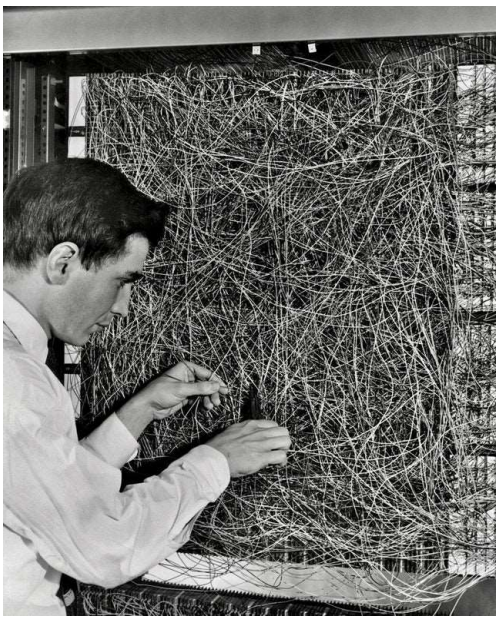

C. E. Shannon



## Künstliche Intelligenz im zeitlichen Kontext

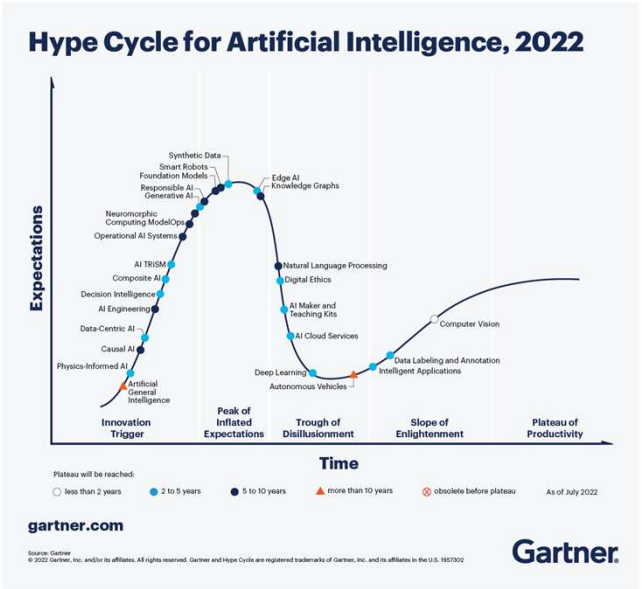
- Frank Rosenblatt, Mark-I Perceptron (1957, 1962)

- Eine der ersten Beschreibungen und Inbetriebnahme eines Neuronalen Netztes mit einer und mehreren Schichten.



## Hintergrund: Data Science.. und KI

- Aktuelle Forschungsthemen mit Schnittstellen zu Data Science im Wandel der Zeit
  - Internet of Things, Cloud, Big Data, Artificial Intelligence, …
  - Daten spielen in vielen Projekten eine wichtige Rolle
    - Energieprognosen
    - Schadenhöhe vorhersagen in Kfz-Schadenfällen
    - Manuelle Arbeit optimal unterstützen
    - Bildklassifikation
- Motivation für Data Science Vorlesungen
  - Viele Daten, Hype steigt, Jobangebote
  - »*sexiest job of the 21st century*« *(Harvard Business Review 2012)*






## Inhalte der heutigen Vorlesung

- Organisatorisches
- Einführung und Inhalte
- Erwartungen und Ziele
- Crisp-DM
- Grundlagen maschinelles Lernen
- Anwendungsbeispiele mit Übung
- Wrap-up

## Organisatorisches (1 von 2)

- **Rahmen**
  - Termine werden in Rapla gepflegt
- **Prüfungsleistung**
  - Wird auf Moodle zur Verfügung gestellt
  - **Python, v 3.9, Anaconda3 2022.05**
  - Kompatibilität und Lauffähigkeit ist !Pflicht! zum Bestehen
  - sklearn sollte für fast alles reichen (keine Zusatzpakete)
- **Kontaktdaten:**
  - per Moodle oder Mail, monika.kochanowski@dhbw-stuttgart.de
    Gerne: Direkt nach der Vorlesung, sonst: B3.10



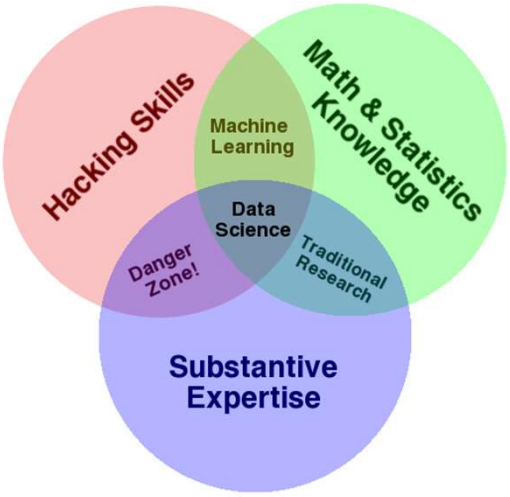
## Organisatorisches (2 von 2)

- **Moodle**
  - Folien
  - Übungen – wenn notwendig Material
  - **Prüfungsleistungsabgabe**
- **Skripte**
  - Folien werden als pdf zur Verfügung gestellt
  - Fotoprotokoll o. ä. in eigener Verantwortung
- **Programmierumgebung für Übungen**
  - Anaconda mit Python 3
  - Installation: *selbstverantwortlich*
  - Es kann gerne alles andere verwendet werden während der Entwicklung, so lange es unter der vorgegebenen VM zum Ende hin kompatibel gestaltet wird..

## Data Science – Begriffsbestimmung und Historie

- Bei der Erstellung eines Curriculums für Data Science wurden die **Fähigkeiten** wie in dem Diagramm dargestellt aufgeteilt
- Datenvisualisierung
- Maschinelles Lernen
- Mathematik
- Statistik
- Informatik
- Kommunikation
- Domänenwissen

- .. is a blend of Red-Bull-fueled hacking and espresso-inspired statistics. […] Data science is the civil engineering of data. […]
  Meteamarket CEO Mike Driscoll in [O'Neil and Schutt 2013]



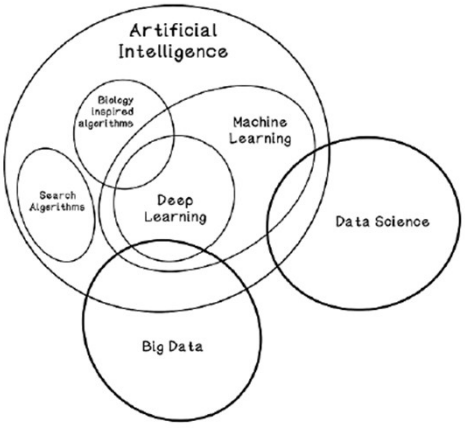

Drew Conway 2010, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, Creative Commons



16

## KI – Data Science – Big Data – ML … und was machen wir in der Vorlesung?

- **Data Science**
- „ … is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is related to data mining and big data."
- „viertes Paradigma" der Wissenschaft – empirical, theoretical, computational and data-driven (Jim Gray)
- Quelle: Wikipedia




Grokking Artificial Intelligence Algorithms, Manning 2019



17

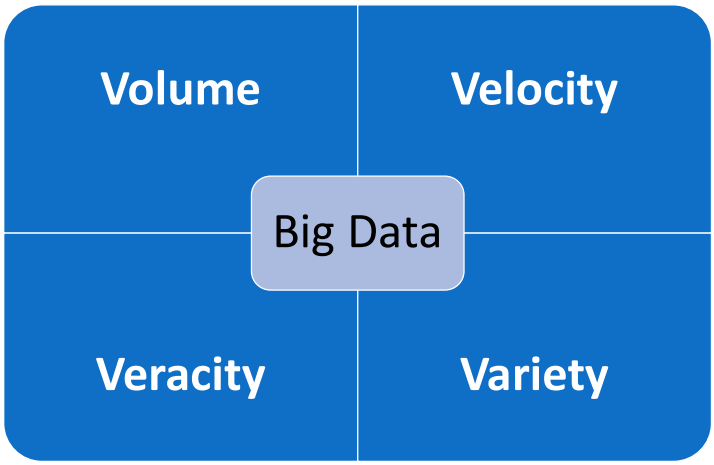
## Big Data
### Abgrenzung zu Data Science

- Analyse von »sehr großen« Datenmengen
  - Kein Konsens über »sehr groß«, auch abhängig von der Entwicklung
  - Häufig auch: **Die 4 V's**
  - Beispiel Hadoop bei Facebook[1] (»das zweitgrößte Hadoop Cluster der Welt«):
    - 9 TB im Hauptspeicher
    - 2 PB Daten
    - 10 TB neue Daten pro Tag
    - 2500 Prozessoren
- Fraunhofer IAIS
  - Big Data Systeme sind verteilt
  - Big Data Algorithmen und Applikationen müssen parallelisierbar sein
  - **Was folgt daraus für Big Data und Data Science?**

[1] Tom White, »Hadoop the definitive guide«, O'Reilly, 2012



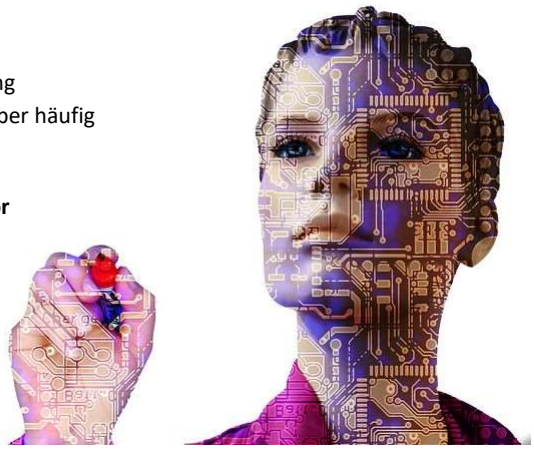
## Big Data und die vier Vs
### Eigenschaften von Big Data




z. B. in **Klein, Tran-Gia und Hartmann, 2013** KLEIN, DOMINIK ; TRAN-GIA, PHUOC ; HARTMANN, MATTHIAS: Big Data. In: *Informatik-Spektrum* Bd. 36 (2013), Nr. 3, S. 319–323

## Künstliche Intelligenz
### Abgrenzung zu Data Science

- Künstliche Intelligenz
  - Keine einheitliche Definition aktuell in der Forschung
  - Ist nicht ausschließlich maschinelles Lernen (wird aber häufig synonym verwendet)
  - Häufig:
    - **Aufgaben mit Computern anzugehen die zuvor menschliche Intelligenz erfordert haben**
- Data Science..
  - Ist nicht nur maschinelles Lernen
  - Ist auch maschinelles Lernen
  - Ist die Wissenschaft der Analyse von Daten

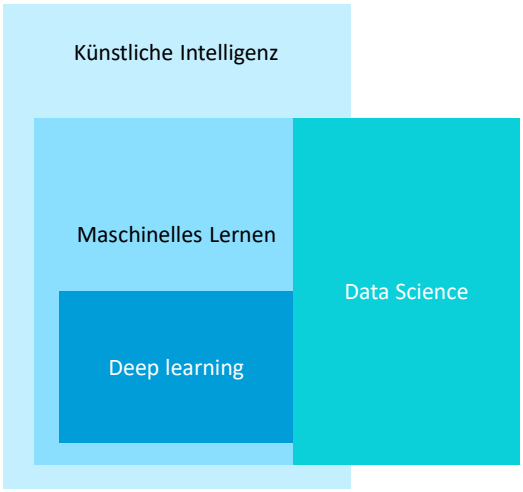




21

## Data Science
### Künstliche Intelligenz und Data Science – eine Abgrenzung

- Eine **künstliche Intelligenz** ist ein Stück Software, das ein Problem mit Verfahren löst, die sich an menschlichen Kognitionsprozessen orientieren. (Synonym / Marketing-Begriff: kognitive Systeme)
- **Maschinelles Lernen** ist die Untermenge von künstlicher Intelligenz, die auf einer Menge Trainingsdaten arbeitet.
- **Deep Learning** ist die Untermenge von maschinellem Lernen, die mit künstlichen neuronalen Netzen mit vielen Hidden Layers arbeitet.
- **Data Science** ist ein interdisziplinäres Wissenschaftsfeld, das durch Methoden, Prozesse, Algorithmen die Extraktion von Erkenntnissen, Mustern aus strukturierten und unstrukturierten Daten ermöglicht[1)]
- **In der VL: Fokus auf überwachte grundlegende Methoden**




1) Vasant Dhar: Communications of the ACM, December 2013, Vol. 56 No. 12, Pages 64-73



22

## Künstliche Intelligenz
**Zeitstrahl**

**50er – 60er**
- Frühe Phase (50er – 60er) Jahre „subsymbolisches Lernen“:
  - *Grundidee*: Lernen mit so wenig Vorwissen wie möglich
  - *Methode*: Künstliche Neuronale Netze

**60er – 80er**
- Symbolische Lernansätze (60er – 80er Jahre):
  - *Grundidee*: Wissen und Strukturen werden *a priori* benötigt
  - *Methode*: Lernen von logischen Konzeptbeschreibungen

**Heute**
- Moderne Ansätze:
  - Probabilistisches Lernen – Statistisches Lernen
  - Rückkehr der Neuronalen Netze – „Deep Learning“



## Turing Award 2018

**Sieger:**

Yann LeCun

Geoffry Hinton

Yoshua Bengio

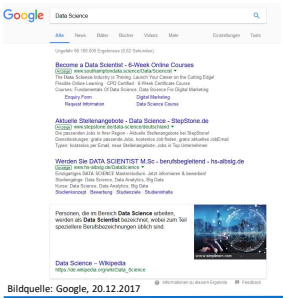
Vergeben für:

„The conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.“

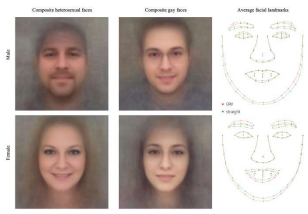
… und für‘s Durch- und Festhalten an NN‘s in schwierigen Zeiten

# Data Science – Anwendungsbeispiele


Bildquelle: Amazon, 15.12.2017



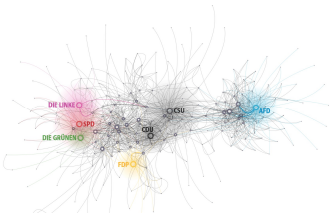

Bildquelle: Google, 20.12.2017



Bildquelle: Wang, Y., & Kosinski, M. (2017, October 16). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.. Retrieved from psyarxiv.com/hv28a

Bildquelle: Amazon, 20.12.2017

Bildquelle: https://www.retaildoctor.com.au/custome-persuasion-on-the-shelf/, Ursprünglich Think Neuro, Neuro Retail Revolution Conference Amsterdam, 2013

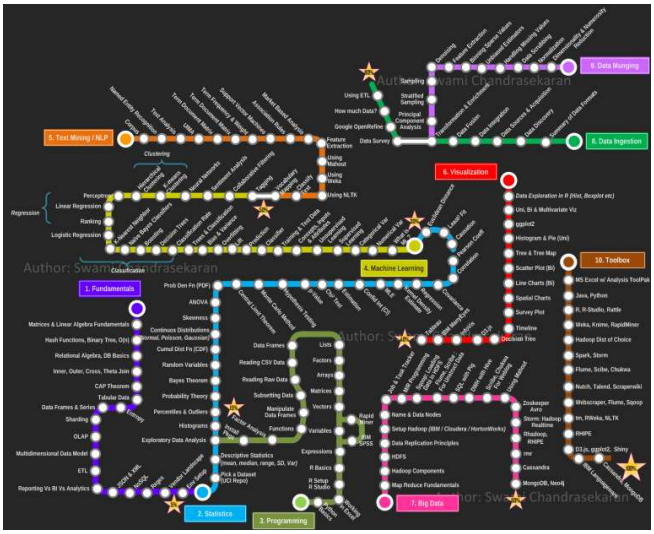



http://www.sueddeutsche.de/politik/politik-im-netz-in-der-rechten-echokammer-1.3485685



25

# Data Science – ein weites Feld




Bildquelle: Swami Chandrasekaran

Semester 1

- M1: New Business Models and Strategies (Okt/Nov)
- M2: Introduction to Business Analytics (Dez/Jan)
- M3: Introduction to Data Science (Feb/März)

Semester 2

- M4: Ethics and Law (April/Mai)
- M5: Data-Warehouse-Workshop (Juni/Juli)
- M6: BI- and Big-Data-Design-Workshop (Aug/Sept)

Semester 3

- M7: Programming for Data Scientists (Okt/Nov)
- M8: Business- und CRM-Analytics (Dez/Jan)
- M9: BI- and Big-Data-Architectures (Feb/März)

Semester 4

- M10: Applied Statistics (April/Mai)
- M11: Web- and Social-Media-Analytics (Juni/Juli)
- M12: Data-Mining-Process: Algorithmns and Implementation (Aug/Sept)

Semester 5

- M13: Thesis Coaching (Okt/Feb)
- M14: Master Thesis (Okt/Feb)

Masterstudiengang Data Science, HDM Stuttgart, 90 ETCS-Punkte, abgerufen 17.01.2018



26


Bildquelle: Swami Chandrasekaran



27

## Data Science Vorlesung
## Und was möchten Sie?

Bitte mit Smartphone ins Internet gehen unter

**menti.com** mit Code **wie angegeben**

Das Ziel der Vorlesung ist es, einen **Überblick** über das Feld Data Science zu geben und zu **motivieren**, die Themen weiter zu vertiefen.



28

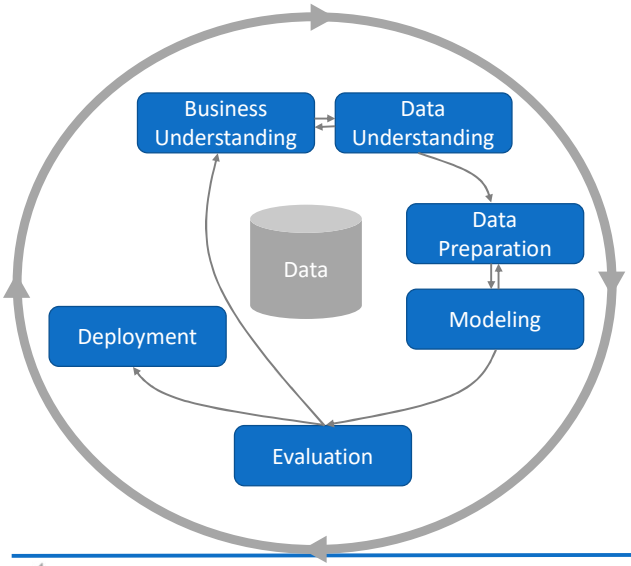
## Übungen

- Sie sind ein kleiner Online-Spezialversandhandel für [WÄHLEN SIE EIN THEMA AUS]. Aktuell haben Sie 100 Kunden, die regelmäßig bei Ihnen bestellen, und 1000, die einmal bestellt haben. Das sind die vorliegenden Daten. Sie haben von Data Science gehört, und möchten Ihre Daten **»gewinnbringend«** einsetzen.
- **Erster Teil (5 Minuten, Einzelübung)**
  - Überlegen Sie sich ein Szenario.
  - Überlegen Sie in groben Schritten (5 – 10), wie Sie auf dieser Basis vorgehen würden.
- Zweiter Teil (10 Minuten, in Gruppen von 4 Personen)
  - Definieren Sie die **Ziele** Ihres Data Science Vorhabens.
    - Worum geht es? Was wollen wir lernen? Wer will was lernen?
- **Skizzieren** Sie Ihre Ergebnisse auf einem DIN A4 Blatt Papier.



29

## Crisp-DM: Cross-Industry Standard Process for Data Mining

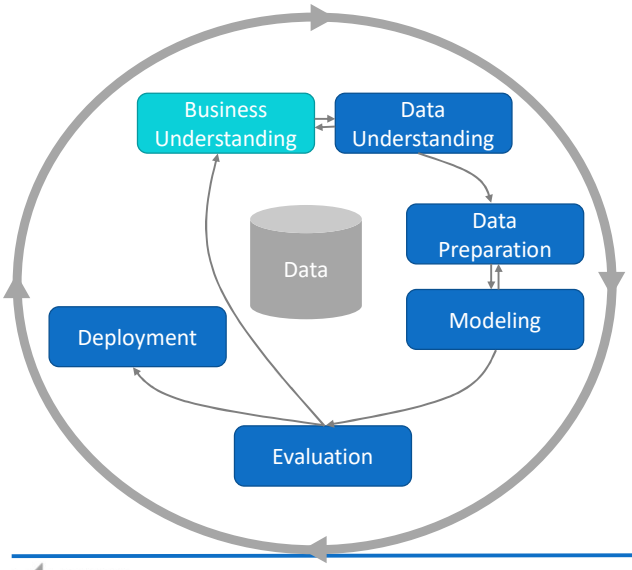



- 1996 in einem Forschungsprojekt verfeinertes Vorgehensmodell u. a. von Daimler, Teradata, SPSS
- Sechs Phasen in Data-Mining-Projekten werden beschrieben



30

## Crisp-DM: Die einzelnen Phasen

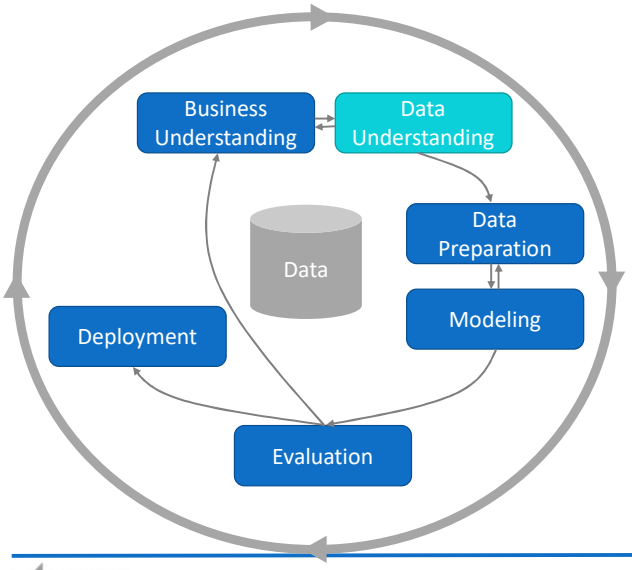



- Business Understanding = **Geschäftsverständnis**
- Was sind die Ziele auf Geschäftsebene?
- Welche Anforderungen an das Ergebnis gibt es?
- Welche offenen Fragen sollen beantwortet werden?
- Wie könnten beispielhafte Antworten oder Ergebnisse aussehen?



31

## Crisp-DM: Die einzelnen Phasen

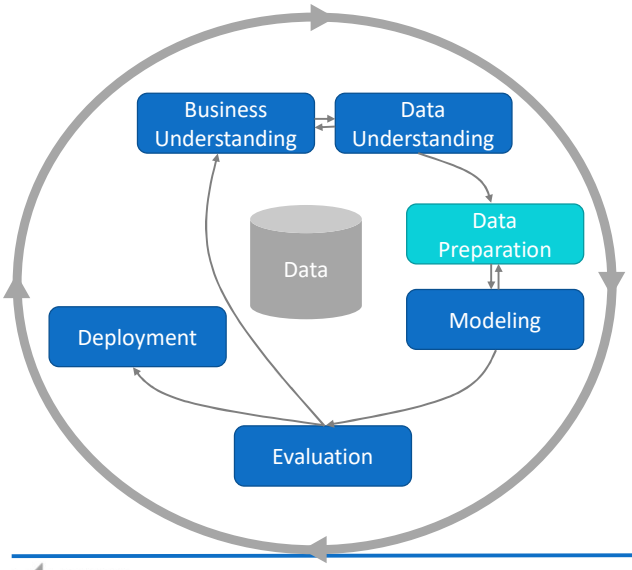



- Data Understanding = **Datenverständnis**
- Welche Daten liegen vor?
- Wie sehen diese aus? Könnte es Probleme mit den Daten geben?
- Kann man »auf den ersten Blick« bereits Zusammenhänge erkennen?
- Wie könnten beispielhafte Antworten oder Ergebnisse aussehen?



32

## Crisp-DM: Die einzelnen Phasen

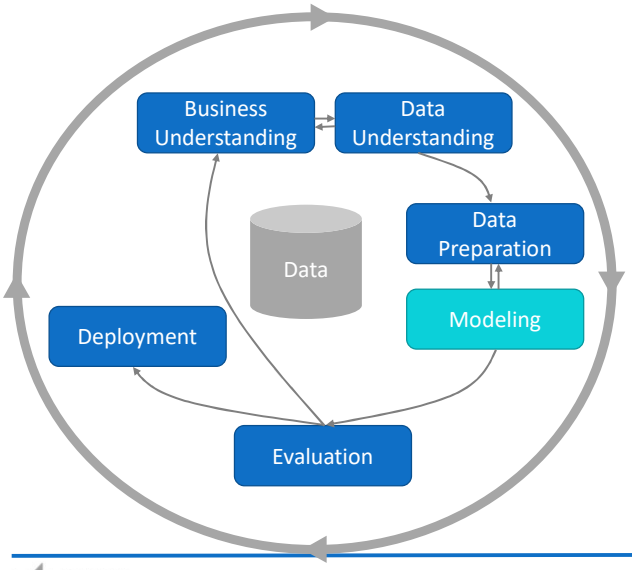



- Data Preparation = **Datenvorbereitung**
- Können die Daten in der vorliegenden Form verwendet werden? (meistens: nein)
- Wie können diese vorverarbeitet werden, um sie zu verwenden?



33

## Crisp-DM: Die einzelnen Phasen

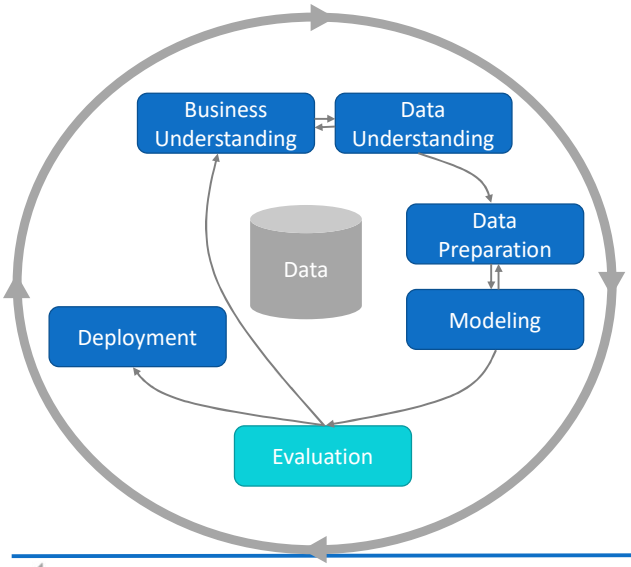



- Modeling = **Modellierung**
- Welche Verfahren lösen mein Problem?
- Wie kann man die Verfahren verbessern?
- Welche Alternativen gibt es?
- Achtung: Es gibt unter Umständen einen (sehr starken) Zusammenhang mit Data Preparation -> Iterationszyklen



34

## Crisp-DM: Die einzelnen Phasen

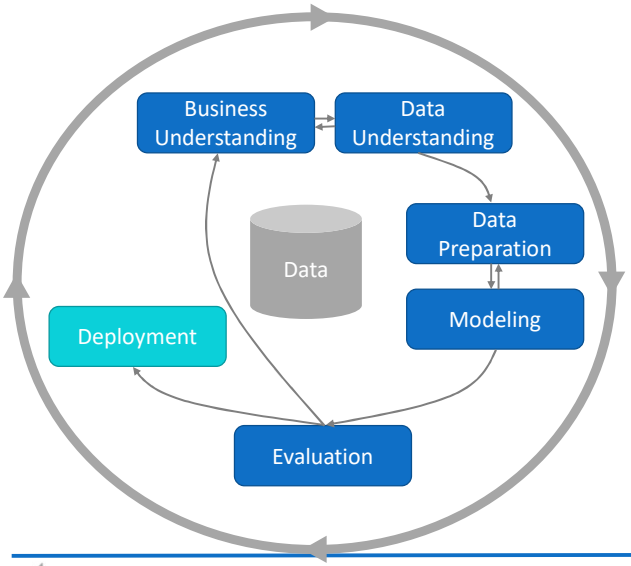



- Evaluation = **Evaluierung**
- Welches Verfahren ist am Besten?
- Beantwortet es die Fragen aus dem Geschäftsverstehen?



35

## Crisp-DM: Die einzelnen Phasen




- Deployment = **Bereitstellung**
- Wie können die Ergebnisse präsentiert werden?
- Wie können die Ergebnisse integriert werden?



36

## Pause – 10 Minuten

- Organisatorisches
- Einführung und Inhalte
- Erwartungen und Ziele
- Crisp-DM
- Grundlagen maschinelles Lernen
- Anwendungsbeispiele mit Übung
- Wrap-up




## Bedeutung und Nutzen des CRISP-DM Modells

- **Diskussion**
- Vorgehensmodelle sind bekannt (z. B. aus der Softwareentwicklung, Projektmanagement)
  - **Was ist ein Vorgehensmodell?**
- Vorteile
  - Erfahrungswissen ist in dem Modell wiedergegeben
  - Es wird nichts vergessen, z. B. das Business Understanding nicht aus den Augen zu verlieren
  - Kommunikation von Status, Zwischenergebnissen, etc. innerhalb eines interdisziplinären Teams und über die Hierarchieebenen hinweg leichter
  - Verbessert die Qualität von Projekten
  - Hier: Strukturiert auch die Vorlesung ganz gut

## Data Science: Anwendungsbeispiele
**Kreditausfall**

- Ziel
  - Gewinn (Zinsen) maximieren
- Mögliche Daten
  - Krediteigenschaften (Höhe, Laufzeit, ..)
  - Historische Daten (Häufigkeit der Beauftragung, ..)
  - Sozioökonomische Merkmale (Bildung, Lohn, Milieu, ..)
  - Demografie (Alter, ..)
  - Geografie (Stadt, Land, Staat, ..)
- Andere Beispiele
  - Marketing (Payback)
  - Medizin
  - Spamfilter




39

## Übungen

- Gruppenübung: Sie sind Dienstleister für einen kleinen Online-Spezialversandhandel für [**WÄHLEN SIE EIN THEMA AUS**]. Aktuell hat dieser 100 Kunden, die regelmäßig bestellen, und 1000, die einmal bestellt haben. Das sind die vorliegenden Daten. Sie haben von Data Science gehört, und möchten die Daten **»gewinnbringend«** einsetzen.
- **Erster Teil (max. 5 Minuten)**
  - Überlegen Sie sich ein Szenario.
- **Zweiter Teil (20 Minuten, in Gruppen von 4 Personen)**
  - Definieren Sie die **Ziele** Ihres Data Science Vorhabens.
    - Worum geht es? Was wollen wir lernen? Wer will was lernen?
  - Skizzieren Sie Ihr Vorgehen. Ordnen Sie die Schritte den CRISP-DM Phasen zu. Beschreiben Sie wichtige Meilensteine Ihres Projektes.
  - Gerne können Sie bereits Toolvorschläge, Algorithmen, Bewertungsmethoden etc. vorschlagen.
  - Spielen Sie die Fragestellungen zukünftig zu erfassender Daten durch.
- **Skizzieren** Sie Ihre Ergebnisse auf einem Flipchart. Stellen Sie diese vor (max. 2 – 3 Minuten).



40

## Maschinelles Lernen
### Begriffe

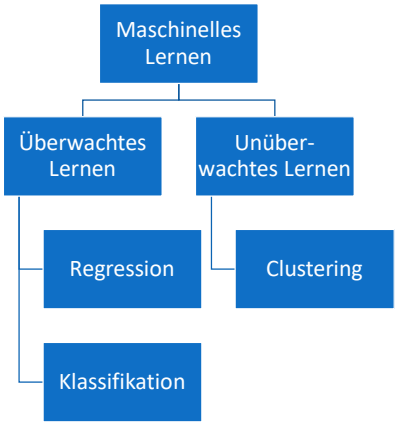
- Maschinelles Lernen (»Statistical Learning«) ist..
  - »Das Erstellen und Verwenden von Modellen, die mit Daten trainiert werden« [Joel Grues 2016]
  - Ein **Datensatz** hat N **Datenpunkte**
    - Diese haben p **Merkmale**, Variablen, Features, Attribute
    - Jeder dieser Merkmale hat eine Ausprägung / einen **Wert**
    - Wir sagen dann: der Datensatz hat p **Dimensionen**
- Ziele
  - **Vorhersage**: Vorhersage von Werten (wieviel ist das Fahrzeug wert?)
  - **Inferenz**: Verständnis von Zusammenhängen (wenn ein Fahrzeug ein Jahr altert, verliert es soviel Wert)

| Automarke | Neuwert | Baujahr | Erwerber | Datum | Wohnort | Preis |
|---|---|---|---|---|---|---|
| DAIMLER | 50.000 EUR | 2014 | Meier | 12.03.2018 | Stuttgart | 15.000 EUR |
| Mercedes | 60.000 EUR | 2015 | Müller | 07.02.2018 | Hamburg | 10.000 EUR |
| .. | .. | .. | .. | .. | .. | .. |



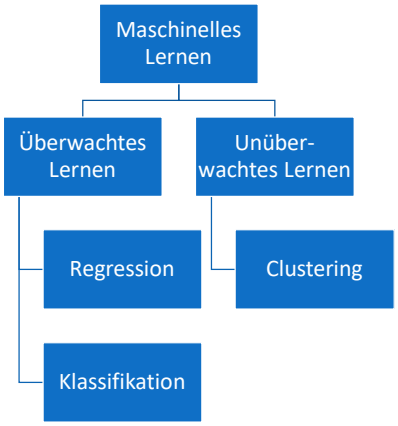
## Maschinelles Lernen
### Motivation

- **Maschinelles Lernen** (»Statistical Learning«) ist..
  - »das Erstellen und Verwenden von Modellen, die mit Daten trainiert werden.« [Joel Grues 2016]
  - Ein **Modell** ist hierbei ein »mathematischer (oder statistischer) Zusammenhang, der zwischen Variablen besteht.«
- Motivation
  - Modell ist nicht bekannt, aber ein Zusammenhang wird vermutet
    - Dieser wird als hilfreich eingestuft
  - Es ist unklar, ob es einen Zusammenhang gibt – aber wenn ja würde man ihn gerne finden



## Maschinelles Lernen
### Einordnung

- **Überwachtes Lernen (supervised)**
  - Für Datenpunkte liegen für die Merkmale (oder Variablen) sowohl Eingaben aus auch Ausgaben vor
    - **Eingabe**: Eingangsvariablen, Input, Prädiktor, unabhängige Variablen
    - **Ausgabe**: Reaktionsvariable, Output, abhängige Variablen
  - »For each **observation** of the **predictor** measurement(s) there is an associated **response** measurement.« [G. James et. al 2013]
- **Unüberwachtes** Lernen (unsupervised)
  - Clustering
- **Halbüberwachtes** Lernen (semi-supervised)
  - **Motivation?**






44

## Maschinelles Lernen
### Überwachtes Lernen

- Wiederholung: Beim überwachten Lernen liegen für die Datenpunkte für die Merkmale (oder Variablen) sowohl **Eingaben (predictor measurement) X** aus auch **Ausgaben (response measurement) Y** vor
- **Annahme**: es gibt eine Abbildung (oder eine) Funktion der Form $Y = f(X) + \varepsilon$
  - $\varepsilon$: stochastischer Fehlerterm $\varepsilon$
  - 15.000 EUR = f(DAIMLER, Rot, 2014, Meier, 12.03.2018, Stuttgart) + $\varepsilon$
- Um $f$ zu finden ohne Vorwissen, können zwei Arten von Methoden eingesetzt werden
  - **Parametrische** Methoden (Annahme über die Form von $f$ wird getroffen)
  - **Nicht-parametrische** Methoden (keine Annahme über die Form von $f$)

| Automarke | Neuwert | Baujahr | Erwerber | Datum | Wohnort | Preis |
|---|---|---|---|---|---|---|
| DAIMLER | 50.000 EUR | 2014 | Meier | 12.03.2018 | Stuttgart | 15.000 EUR |
| Mercedes | 60.000 EUR | 2015 | Müller | 07.02.2018 | Hamburg | 10.000 EUR |
| .. | .. | .. | .. | .. | .. | .. |

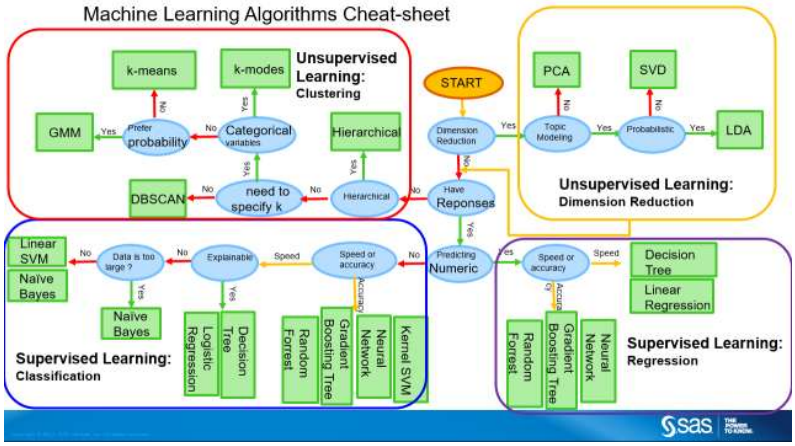


45

## Algorithmen für Maschinelles Lernen
**Als Beispiel eines von (sehr) vielen Cheat Sheets**

- Regression
- KNN
- Naive Bayes
- SVM
- ..

- Ziel: Grundlagen anhand bekannter Algorithmen legen und die »wichtigsten« kennenlernen




Bildquelle: https://whatsthebigdata.com/2017/05/02/types-of-machine-learning-algorithms-and-when-to-use-them/, am 10.02.2018



46

## Gruppenübung

- **10 Minuten Zeit – 2er Gruppen – Diskussion in der großen Runde**
- Entscheiden Sie für die folgenden drei Szenarien (1) ob es sich um eine **Klassifikations-** oder **Regressionsaufgabe** handelt und geben Sie an, ob wir mehr in (2) **Inferenz** oder **Vorhersage** interessiert sind. Weiterhin geben Sie die (3) Anzahl der **Dimensionen** p und die Anzahl der **Datenpunkte** N an.
- **(a)** Wir haben eine Liste der Top-500 Unternehmen in Deutschland. Für jede Firma liegt der Gewinn, Umsatz, die Anzahl der Mitarbeiter, die Branche und das Gehalt des CEOs vor. Wir interessieren uns für die Faktoren welche das CEO-Gehalt beeinflussen.
- **(b)** Wir möchten in neues Produkt im Markt einführen und wollen wissen, ob es ein Erfolg oder ein Ladenhüter wird. Wir haben Daten von 20 ähnlichen Produkten welche eingeführt wurden. Wir wissen ob es ein Erfolg oder ein Ladenhüter war, den Preis, das Budget für Marketing, die Preise der Konkurrenz und zehn weitere Variablen.
- **(c)** Wir möchten gerne die prozentuale Veränderung des Euros in Zusammenhang mit den Börsen weltweit vorhersagen. Dafür liegen Daten aus ganz 2018 vor. Für jede Woche ist bekannt wie sich der Euro prozentual verändert hat, sowie die Börsenwerte er USA, GB und Deutschland.

Quelle: Frei nach [James et al. 2013]



47

## Und nächste Stunde sehen Sie..

- Statistische Grundlagen
- Übung
- Mehr über Regression
- Das Bias- und Varianz Dilemma
- Erste Installation von Python



## Literaturliste


- [James et al. 2013] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: An introduction to statistical learning
  - Favorit: Sehr gut gemachte Einführung, jedoch Beispiele in R, verständlich mit Mathematik, als pdf frei erhältlich
- [Hastie et al. 2008] Trevor Hastie, Robert Tibshirani, Jerome Friedman: The elements of statistical learning
  - DIE Referenz, für Mathematiker geschrieben, als pdf frei erhältlich
- [O'Neil and Schutt 2013] Cathy O'Neil and Rachel Schutt: Doing Data Science
  - Spannend zu lesen, teilweise Erfahrungsberichte (durch Drittautoren)
- [Mueller and Guido 2017] Andreas C. Müller & Sasha Guido: An Introduction to Machine Learning with Python
  - Interessant da Python 3 tatsächlich genutzt wird für die Einführung inklusive der üblichen Bibliotheken
- [Grues 2016] Joel Grues (übersetzt von Kristian Rother): Einführung in Data Science
  - Auf deutsch gut übersetzt, nutzt Python für grundlegendes Verständnis ohne die üblichen Bibliotheken, extrem leicht lesbar
- [Alpaydin 2008]: Ethem Alpaydin (übersetzt von Simone linke): Maschinelles Lernen
  - Auf deutsch gut übersetzt, relativ viel Mathematik, in Deutschland scheint das weit verbreitet zu sein