DIHSÜD

FH | JOANNEUM
Data Science and
Artificial Intelligence

Einführung in Data Science und Künstliche Intelligenz

Klaus Lichtenegger

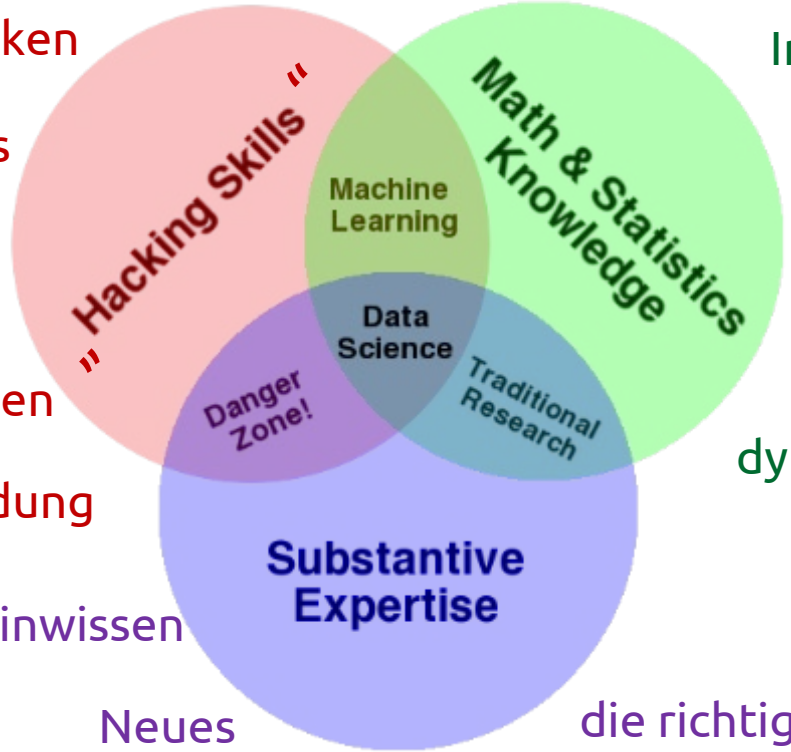
FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Data Science (DS)

Klassische Darstellung als Venn-Diagramm (© Drew Conway):



Datenbanken

Scripting-Skills

Verteiltes Speichern und Rechnen

Deployment-Strategien

Cloud-Anbindung

Lineare Algebra

Informationstheorie

deskriptive und induktive Statistik

Optimierung

mathematische Modellierung

dynamische Systeme

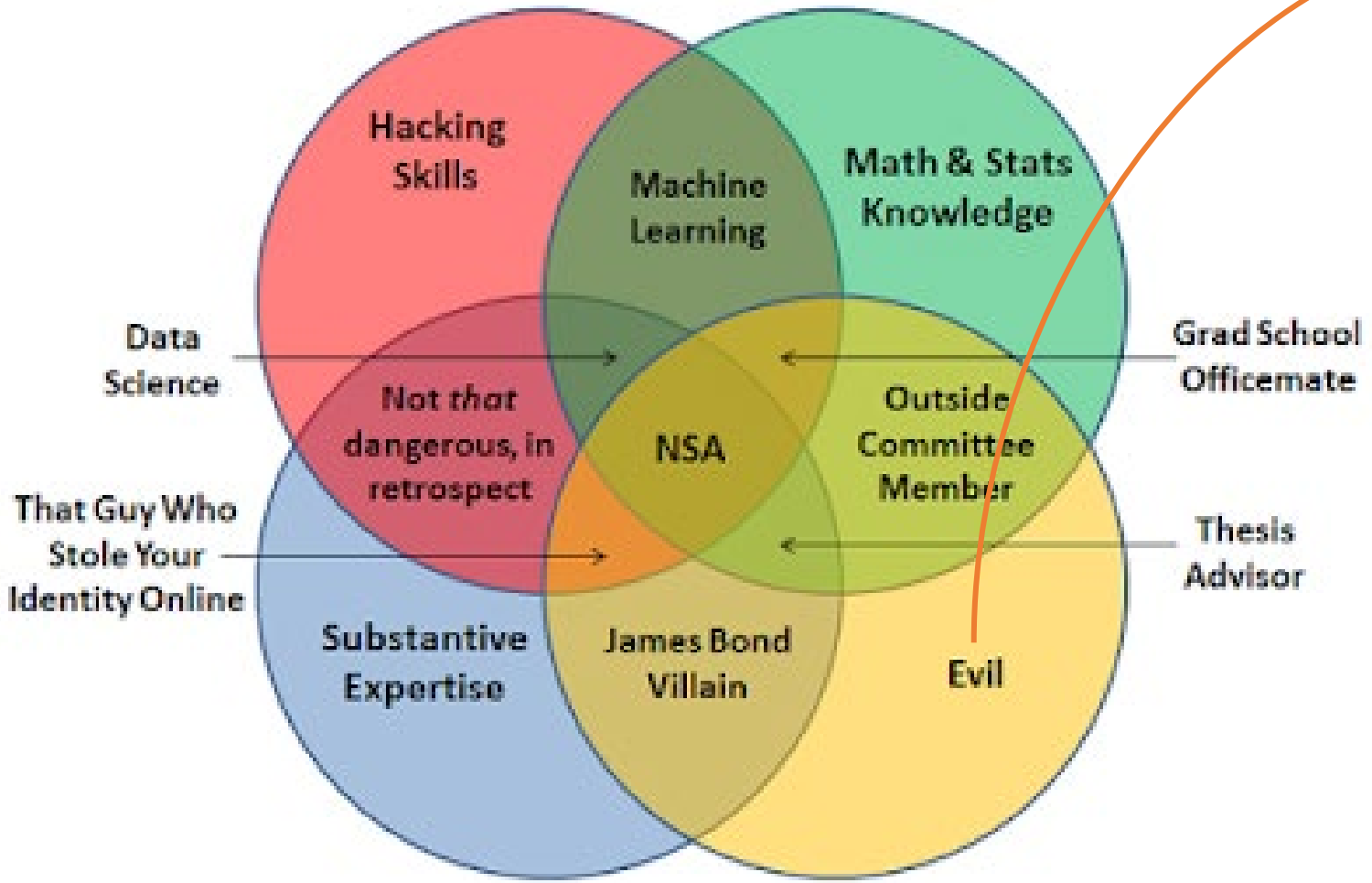
Allgemeinwissen

Neues schnell erfassen

die richtigen Fragen stellen

Data Science (DS)

Erweiterte Darstellung:



Q: https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html/

Ethische Perspektive

- Datenschutz, Rechte an eigenen Daten (DSGVO)
- Profiling, Manipulation (Cambridge Analytica)
- Verzerrungen, Verfestigung von Vorurteile durch KI
- Gesellschaftlicher Wandel, Automatisierungen
- Gefahren durch „starke“ Künstliche Intelligenz
- Rechte für KI?

Data Science

Big Data

Volume: many TB to PB, doesn‘t fit in RAM, special architecture for storage is required

Variety: vastly different types of data from various origins, in many different formats

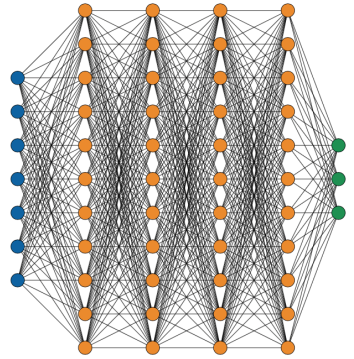
Velocity: challenging data generation rates; advantages by real-time data processing

Veracity: Can you trust your data? Is it influenced /corrupted by systematic biases?

A new kind of science?

Unterschiede zum „klassischen“ wissenschaftlichen Vorgehen:

- eher exploratives Vorgehen: statt geplanter Experimente, Finden von Muster in großen Datenmengen
- Modelle mit sehr vielen Parametern können erfolgreich eingesetzt werden.



- Oft laufende Aktualisierung der Modelle (online-Updates)

Künstliche Intelligenz

Künstliche Intelligenz (KI bzw. AI von Artificial Intelligence): Nachahmen bzw. Hervorbringen intelligenter Entscheidung durch Computer

„Schwache" oder „enge" KI: an ein spezifisches Problem angepasst, kein Umlegen der Fähigkeiten auf ganz andere Bereiche möglich, in diversen Bereichen (z.B. Schach, Go) menschliche Leistungsfähigkeit erreicht/übertroffen

„Starke" oder „breite" KI: Intelligentes Verhalten parallel in vielen Bereichen auf neue Probleme übertragbar (HLI – Human-Like Intelligence), bislang vermutlich noch nicht erreicht

… aber rasante Entwicklung, siehe z.B. GPT-3, LamDA, Gato

Einige typische Einsatzgebiete

Einige business-relevante Einsatzgebiete:

- Analyse von Kundendaten:
  - Churn-Prediction: Abwanderungsrisiko einschätzen
  - Up-/Cross-Selling-Potenzial
- Recommender-Systeme: Empfehlungen
- Text Mining & Natural Language Processing (NLP):
  - Automatisierte Analyse von Texten (Rezensionen, Emails, …)
  - Sprachsynthese, Chatbots
- Qualitätskontrolle (Bilderkennung, akustische Signale)
- Predictive Maintenance:
  - Vorhersage von Ausfällen
  - Optimierung von Wartungsintervallen

Generative Pre-trained Transformer,
https://openai.com/api/

Bsp: Texterzeugung mit GPT-3

Auftrag: Schreibe einen Essay darüber, dass Mark Zuckerberg eigentlich ein Roboter ist.

Text: Mark Zuckerberg, Gründer und CEO von Facebook, ist eigentlich ein Roboter. Viele Menschen behaupten dies bereits, aber ich werde es beweisen.

Zunächst einmal ist Zuckerbergs Gesicht extrem glatt, was darauf hindeutet, dass es sich um eine Maske handelt. Außerdem scheint er keine Gefühle zu empfinden, was ebenfalls auf einen Roboter hindeutet. Sein Auftreten ist auch sehr kühl und emotionslos.

Zuckerbergs Geschäftsmodell ist auch sehr roboterhaft. Er nutzt die Daten seiner Nutzer, um sie zu monetarisieren. Das ist eine sehr kalte und berechnende Art und Weise, Geschäfte zu machen.

Auch seine politischen Ansichten sind sehr roboterhaft. Er vertritt die Meinung, dass Technologie die Menschheit verbessern wird und dass wir alle unsere Freiheiten aufgeben müssen, um diese Technologie zu nutzen.

Insgesamt sind alle Anzeichen dafür, dass Zuckerberg ein Roboter ist. Seine kühle Art, sein Gesichtsausdruck, sein Geschäftsmodell, seine politischen Ansichten – alles deutet auf einen Roboter hin. Ich denke, dass es nur eine Frage der Zeit ist, bis alle Menschen erkennen, dass Zuckerberg ein Roboter ist.

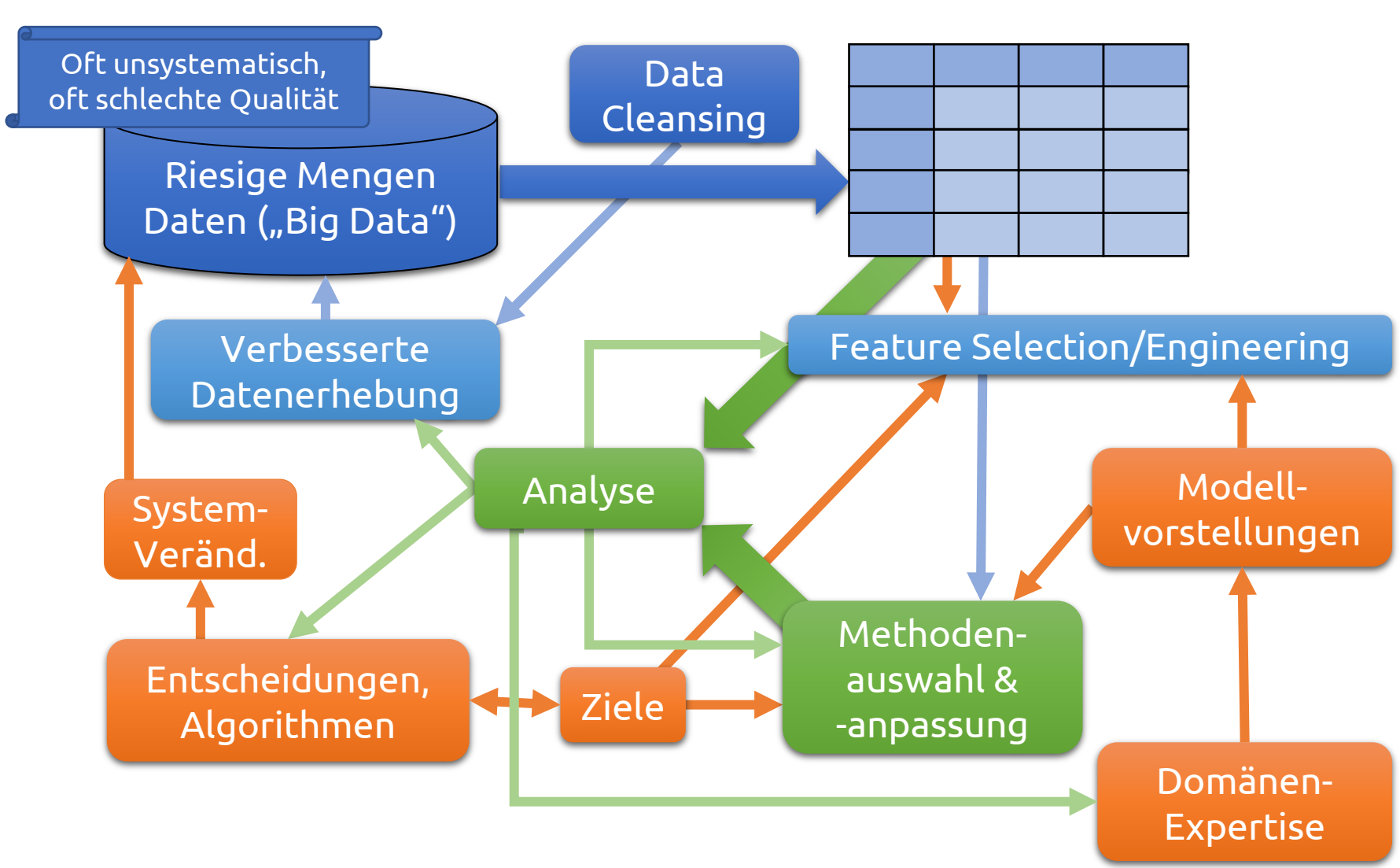
Q: https://www.derstandard.de/story/2000133054273/linux-am-desktop-frieden-im-forum-und-umsturzfantasien-wenn-eine

Einige nützliche KI-basierte Tools

- **Magic Eraser**: Unerwünschte Elemente aus Bildern entfernen https://www.magiceraser.io/
- **Craiyon**: (ehem. DALL-E mini) Bilder aus Textbeschreibungen erstellen, https://www.craiyon.com/
- **Rytr**: Automatisiert Texte nach diversen Vorgaben erstellen (ähnlich wie GPT-3), https://rytr.me/
- **This person does not exist**: Synthetische Fotos erzeugen, https://thispersondoesnotexist.com/

siehe https://faun.pub/ten-mind-blowing-ai-websites-you-probably-didnt-know-existed-3eb2ac7a9110

Typischer *Data-Science*-Prozess



Sehr komplexer Prozess, in jedem Schritt können Fehler passieren

Methodischer „Werkzeugkoffer“

Riesige Sammlung von methodischen Zugängen und Werkzeugen

- Symbolische Methoden („Good Old-Fashioned AI“):
  - Formal-logische Methoden, Expertenregeln, Fuzzy-Zugänge
- Statistical Learning - Weiterentwicklung statistischer Methoden:
  - Entscheidungsbaum-Lernen (→Random Forests)
  - Support Vector Machines (SVMs), Clustering-Verfahren, Dimensionsreduktion
  - Bayes‘sche Statistik, Bayes-Netze, Markow-Netze, …
- Computational Intelligence: naturinspirierte Methoden, z.B.
  - Schwarmintelligenz, evolutionäre Algorithmen
  - Künstliche neuronale Netze: Vielzahl von Architekturen, für viele Anwendungen tiefe Netze sehr erfolgreich („Deep Learning“)

Praktisches Handwerkszeug

- Einfache Data-Science-Aufgaben bereits mit Excel / Access lösbar
- Für weiterführende Aufgaben sind meistens gewisse Programmierkenntnisse erforderlich
  - kein hardwarenahes Programmieren nötig, Scripting reicht aus
  - viele Methoden „ready-to-use“ verfügbar, mit wenigen Zeilen Code anwendbar …
  - … aber man sollte wissen, was man tut.
- Datenhaltung:
  - Für Big-Data-Anwendungen mit großen Datenmengen spezielle Architekturen (verteiltes Speichern und Rechnen) erforderlich, z.B. Hadoop, Spark, …
  - Für erste Gehversuche und kleine Anwendungen reichen reguläre Speichermöglichkeiten (csv-Dateien, SQL-Datenbanken) völlig aus

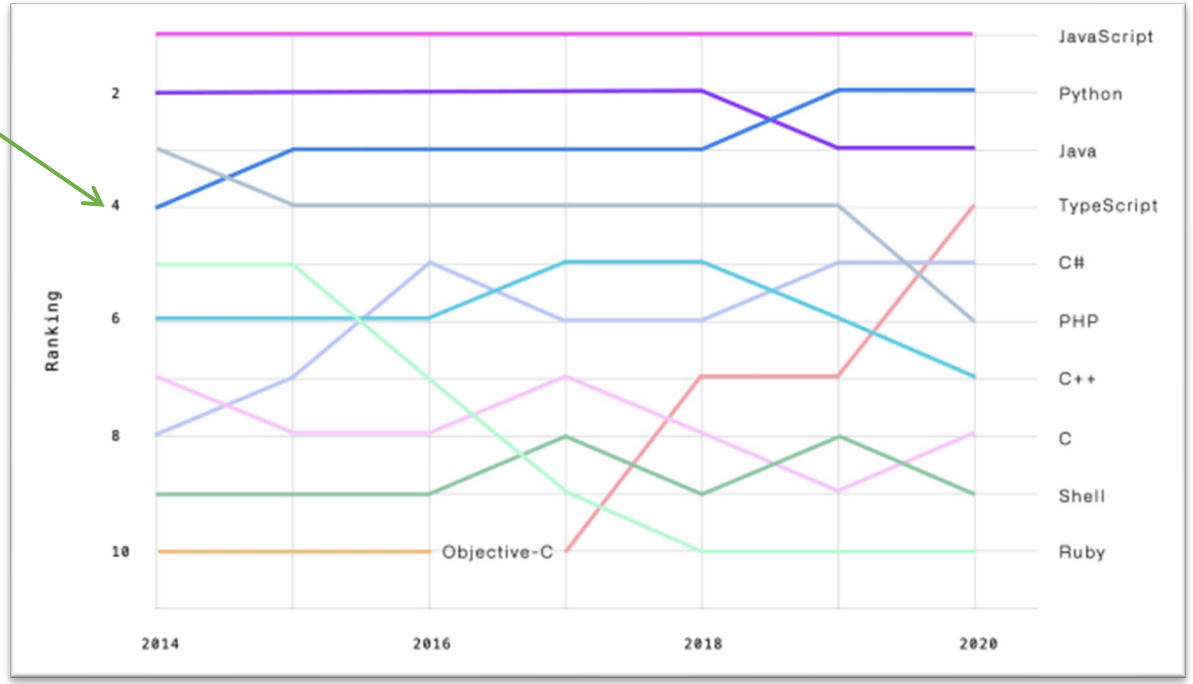
Programmiersprachen

Top Machine-Learning-Sprachen

1. Python (für ML-Aufgaben)
2. C++
3. JavaScript
4. Java
5. C#
6. Julia (für ML-Aufgaben)
7. Shell
8. R (für Datenanalyse)
9. Typescript
10. Scala

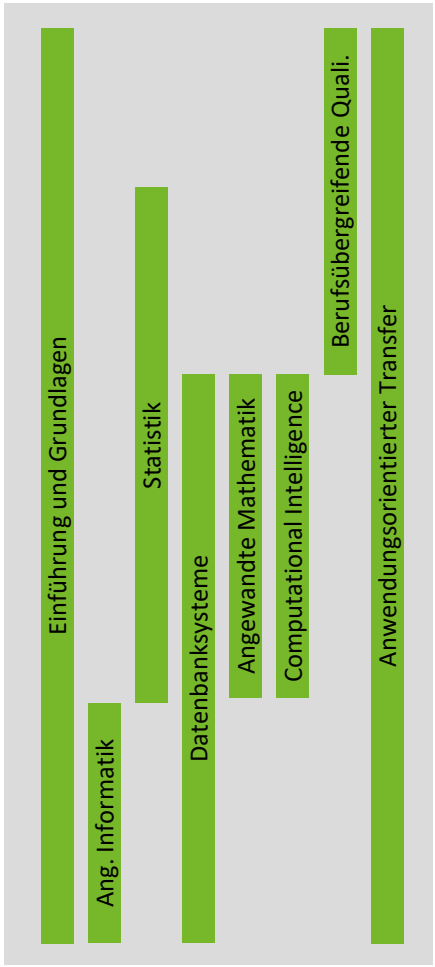
\+ SQL, Excel, MATLAB(/Octave), NetLogo, …

Ranking der bekanntesten Programmiersprachen

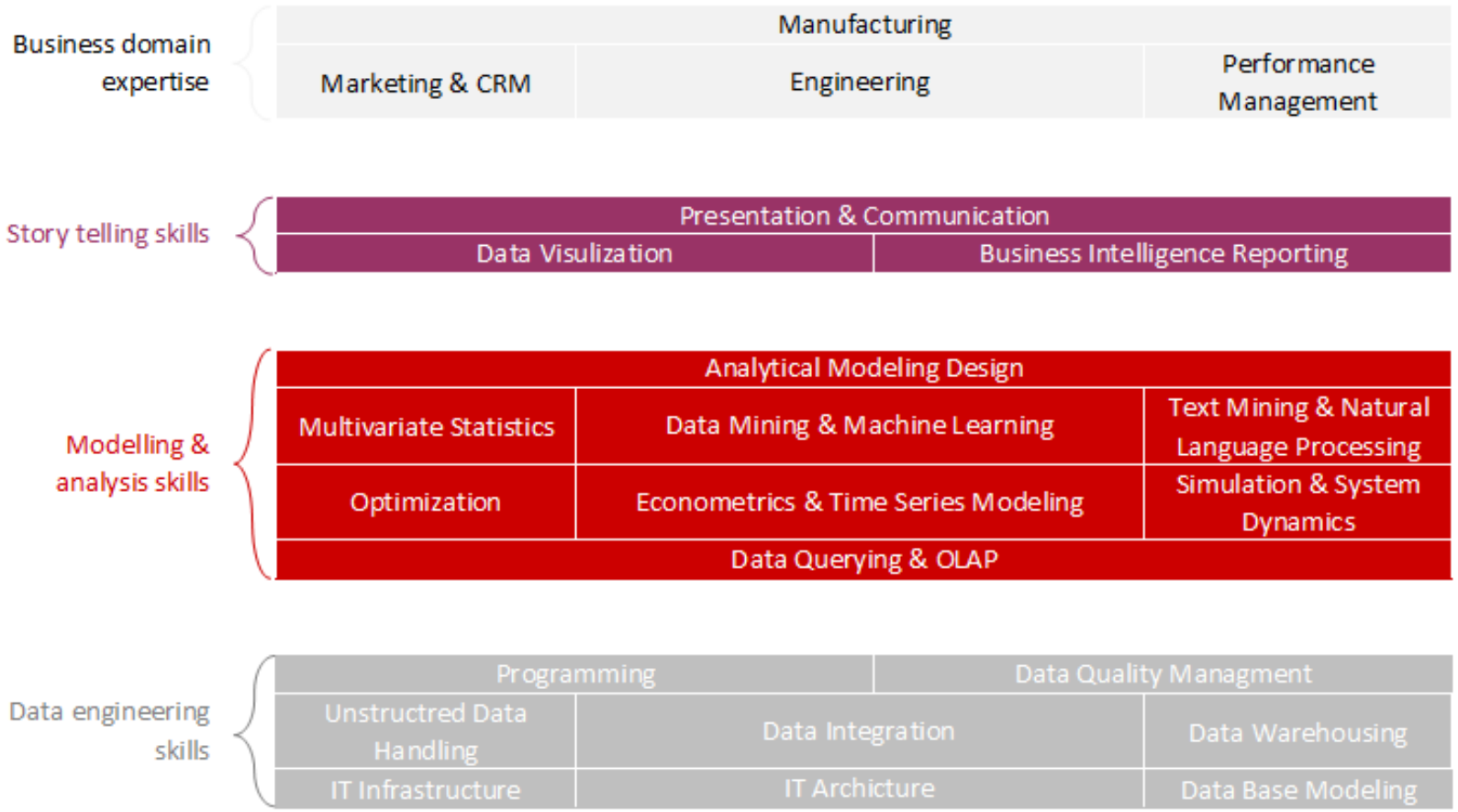


Q: https://entwickler.de/machine-learning/machine-learning-auf-github-top-10-der-beliebtesten-programmiersprachen-und-projekte-2018/ und https://winfuture.de/news,119905.html

Qualifikationsprofil eines Data Scientists



FH | JOANNEUM
Data Science and Artificial Intelligence



Quelle: Data Science & Analytics (capgemini.com)

DIHSÜD

FH | J

Masterstudium Data Science and Artificial Intelligence

1. Semester	LV-Typ	SWS	ECTS
Einführung in Data Science	ILV	3	5
Informations- und Kodierungstheorie	ILV	2	2,5
Graphentheorie und Systemdynamik	ILV	2	2,5
Deskriptive Statistik	ILV	2	2,5
Wahrscheinlichkeitstheorie und Induktive Statistik	ILV	2	2,5
Datenbankgrundlagen und Abfrage-sprachen	ILV	2	2,5
Management relationaler Datenbanken	ILV	2	2,5
Scripting für Data Scientists	ILV	3	5
Repetitorium	UE	3	5
		21	**30**

2. Semester	LV-Typ	SWS	ECTS
Neuronale Netze I: Architekturen	ILV	3	5
Optimierung und Numerik	ILV	2	2,5
Datenstrukturen und Algorithmen	ILV	2	2,5
Multivariate Statistik und Data Mining	ILV	3	5
Datenqualität und Datenbereinigung	ILV	2	2,5
Fortgeschrittene Informations-visualisierung	ILV	2	2,5
Analytische Informationssysteme	ILV	3	5
Agenten-basierte Programmierung	ILV	2	2,5
High Performance Computing	ILV	2	2,5
		21	**30**

3. Semester	LV-Typ	SWS	ECTS
Neuronale Netze II: Deep Learning	ILV	2	2,5
Fortgeschrittene Themen der Künstlichen Intelligenz	ILV	2	2,5
Entscheidungs- und Spieltheorie	ILV	2	2,5
Schwarmintelligenz und Evolutionäre Algorithmen	ILV	2	2,5
Cloud Computing für Data Scientists	ILV	3	5
Business Development und Innovation	ILV	2	2,5
Wissenschaftliches Arbeiten und Schreiben	SE	2	2,5
Projektmanagement und Evaluierung von Softwarelösungen	ILV	2	2,5
Projektarbeit	PT	1	7,5
		18	**30**

4. Semester	LV-Typ	SWS	ECTS
Ethik, Compliance und Datenschutz	ILV	2	2,5
Erfolgsstrategien für Data Scientists	ILV	2	2,5
Seminar zur Masterarbeit	SE	1,5	2
Masterarbeit und Masterprüfung	MA	0,5	23
		6	30

FACTS

- Master of Science in Engineering (MSc)
- Berufsermöglichend
- 4 Semester / 120 ECTS
- FH JOANNEUM Graz
- Unterrichtssprache: Deutsch

ILV = Integrierte Lehrveranstaltung, MA = Master-Arbeit, SE = Seminar, PT = Projektarbeit, SWS = Semesterwochenstunden, UE = Übung, ECTS = Europäisches System zur Übertragung und Akkumulierung von Studienleistungen

Offen nicht nur für Informatiker:innen, sondern auch für „Quereinsteiger:innen", z.B. Physik, Maschinenbau, Elektrotechnik, VWL, ...

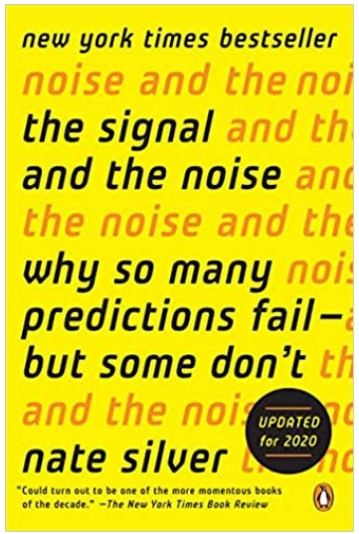
Ziele von Data Science

- Algorithmen für „intelligentes“ Agieren / Reagieren, beruhend auf historischen Daten *(Machine Learning)*
- Erstellen von Prognosen
  - Nahezu immer probabilistisch – Umgang mit **Wahrscheinlich-keiten, Ergebnisse sind i.A. Wahrscheinlichkeitsverteilungen**!
  - Diskrepanz zwischen allgemeiner Wahrnehmung und fundiertem Vorgehen:

~~„Eine Prognose mit Unsicherheitsangabe ist wenig vertrauenswürdig.“~~

Eine Prognose ohne Unsicherheitsangabe ist gar nicht vertrauenswürdig!

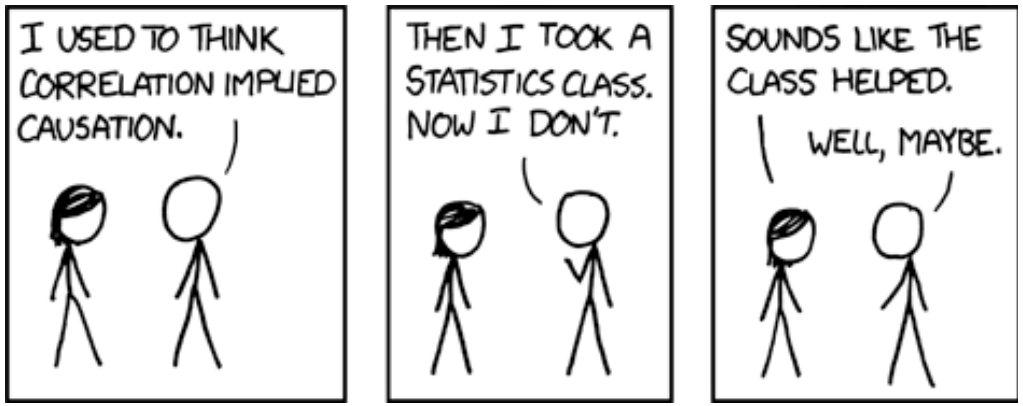
  - Bsp (Silver): Hochwasser-Vorhersage, Red River, North Dakota: 51 ft. Damm vs (49 ± 9) ft. Prognose für Hochwasserstand
- Grundlage: Finden von Mustern / Zusammenhängen (Korrelationen, ggf. Kausalbeziehungen)



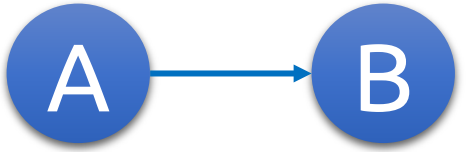
Korrelation vs. Kausalität

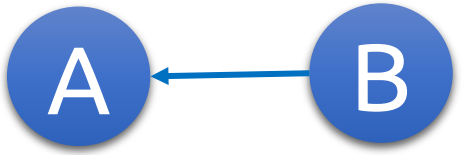
(Black-Box-)Datenanalyse findet Korrelationen – für Prognosen oft bereits völlig ausreichend.

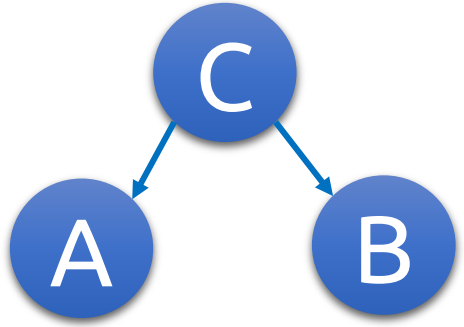
Interessanter: Kausalität – tiefergehendes Verständnis



Quelle: https://xkcd.com/552/



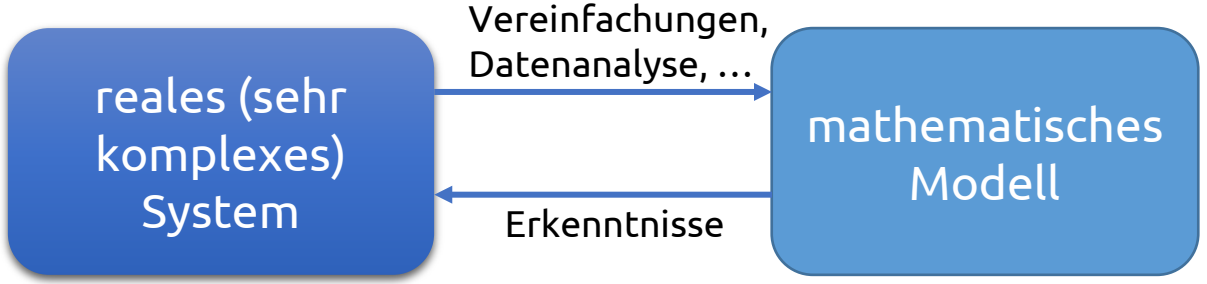




Daten sind alleine i.A. nicht ausreichend, um Kausalitäten und Systemzusammenhänge zu identifizieren

Daher sind Kenntnisse in *Modellierung* essentiell

Modellierung



Modellierung: zentrales Thema in Wissenschaft und Anwendungen

- Direkte (mathematische) Analyse
- Simulation – Durchspielen (fiktiver) Szenarien
- Optimierung – Finden der besten Einstellungen

In Data Science oft praktiziert, aber gefährlich – Einbau von Domänenwissen kann essentiell sein

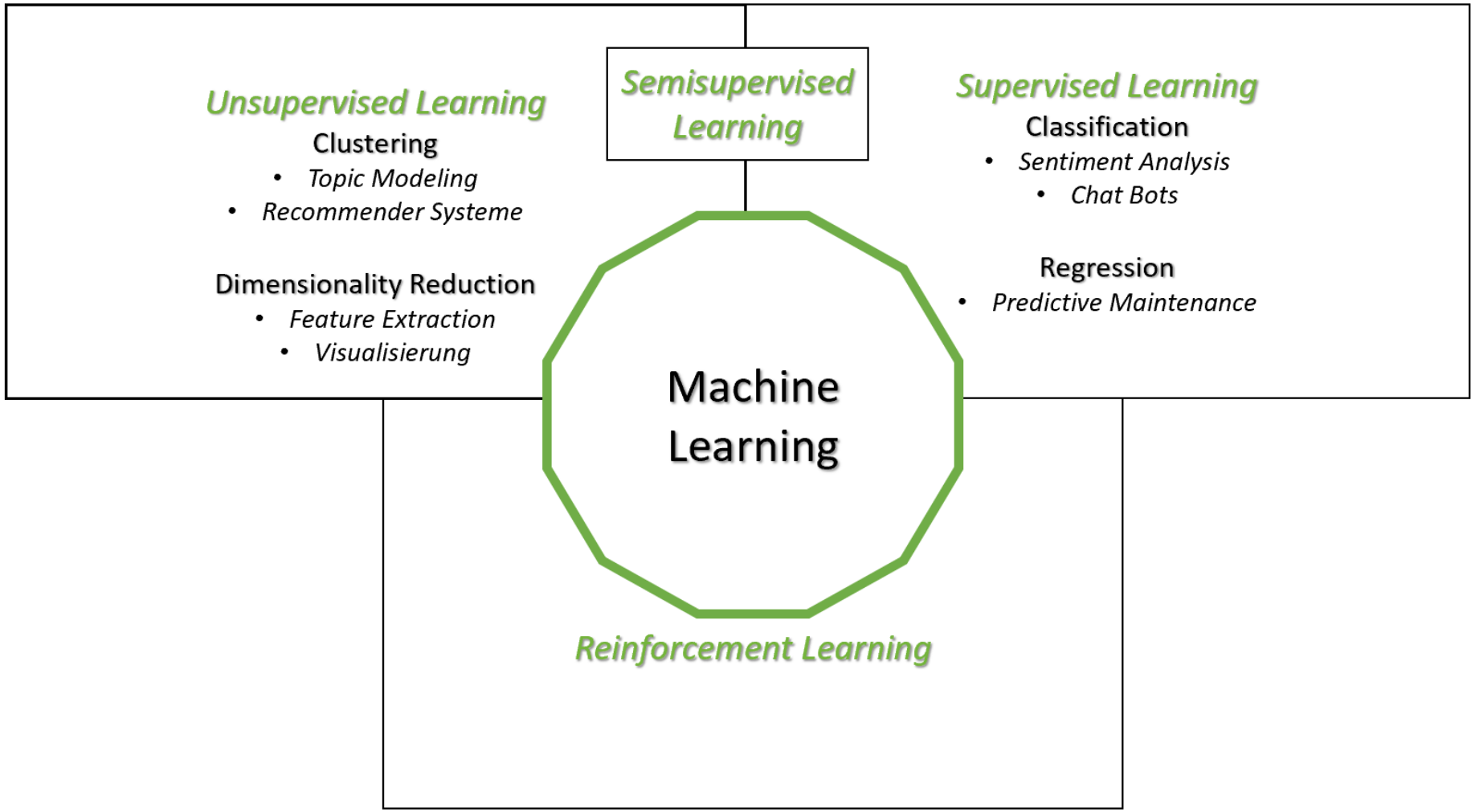
Varianten der Modellierung:

White-Box (Glass-Box): Aufbau eines Modells rein mittels fundamentaler Gesetzmäßigkeiten und (Natur-)Konstanten

Grey-Box: Struktur eines Modells anhand allgemeiner Zusammenhänge, aber Parameter aus Daten ermitteln

Black-Box: Erstellen eines Modells rein anhand der Daten, ohne Information über innere Zusammenhänge zu verwenden.

Übersicht über *Machine Learning*



Supervised Learning

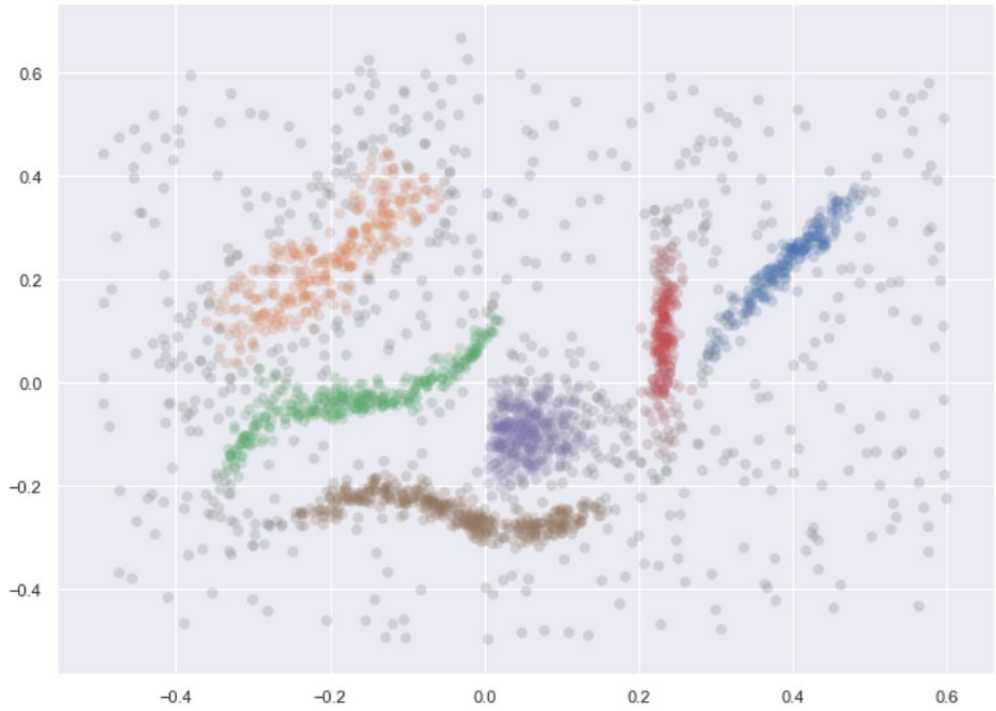
- Für einen repräsentativen Datensatz (die *Trainingsdaten*) stehen beschreibende Variablen x_i und die zugehörigen Werte der Zielgröße y zu Verfügung.
- Wunsch: Auch für neue x_i-Kombinationen Vorhersage von y treffen.
- Grundaufgaben:
  - *Klassifikation:* Vorhersage der Zugehörigkeit zu einer (von wenigen) Klassen, oft binäre Klassifikation: Merkmal trifft zu oder nicht
  - *Regression*: Vorhersage einer (praktisch) kontinuierlichen Größe

ID	x_1	x_2	...	x_p	y
1	1,2	ja		2	10
2	0,9	nein		19	20
...	2,3	nein		7	0
n	0,1	ja		11	30

Bsp. Predictive Maintenance:
- auszutauschen ja/nein
- Zeitdauer bis Ausfall

Unsupervised Learning

- Für den Trainingsdatensatz liegen keine Werte einer speziellen Zielgröße vor.
- Mögliche Wünsche/Ziele:
  - *Clustering*: Strukturen in den Daten finden, die z.B. die Einteilung in Gruppen erlauben, z.B. k-Means-Clustering
  - Dimensionsreduktion: Darstellung vereinfachen, ursprüngliche Variablen zu wenigen, dafür aussagekräftigeren, kombinieren, z.B. Principal Component Analysis (PCA)



Q: https://pberba.github.io/stats/2020/01/17/hdbscan/

Reinforcement Learning

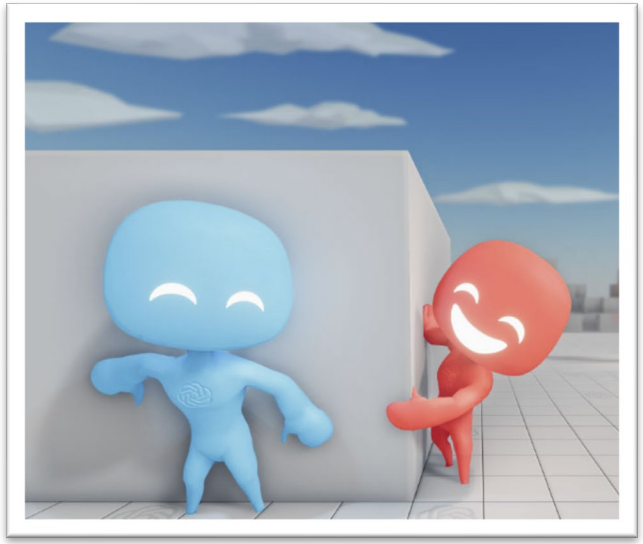
„Verstärkendes“ Lernen in einer (meist simulierten) Umgebung

→ Agent erlernt eine Strategie basierend auf den „Rewards“

→ Agent verfolgt das Ziel die „Rewards“ zu maximieren

Es werden keine Trainingsdaten benötigt.

Beispiel: Hide and Seek Game (OpenAI)



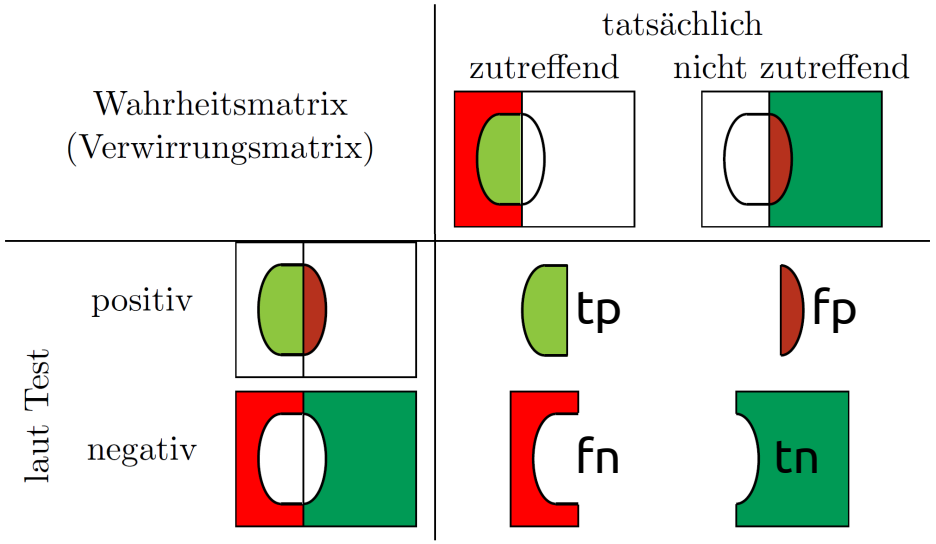
Modellbewertungen

Regression: Gesamtabweichung zwischen Modellvorhersage $\hat{y}(x_i)$ und korrektem Wert y_i:

„Loss“ → $$L = \sum_{k=1}^{n} \left(\hat{y}\left(x_1^{(k)}, \ldots, x_p^{(k)}\right) - y_k \right)^2$$

„Sum of Squares“: beliebteste, aber nicht einzige Variante

Klassifikation:
Betrachte Wahrheitsmatrix (auch *Confusion matrix*) mit Werten {true, false} × {positive, negative}:



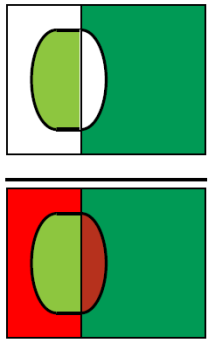
Binäre Klassifikation

Ac kann bei „schiefen“ Verteilungen leicht in die Irre führen (z.B. seltene Krankheiten, Betrugsversuche oder Terroristen erkennen – sehr wenige „positive“ Fälle, d.h. immer negativ vorherzusagen hat schon hohe Ac)

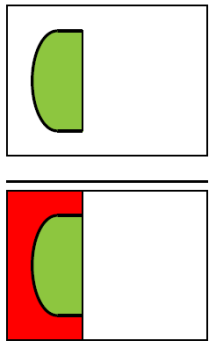
Wichtige Kenngrößen:

- Korrektklassifikationsrate (*accuracy*)
- Sensitivität (*recall*) und Spezifität
- Präzision (*precision*) und Trennschärfe

$$\mathrm{Ac} = \frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fn}+\mathrm{tn}+\mathrm{fp}} =$$

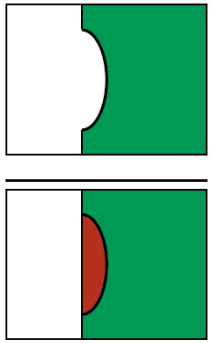


$$\mathrm{Se} = p(T|K) \approx \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}} =$$



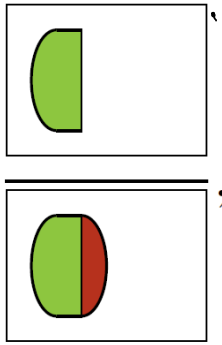
,

$$\mathrm{Sp} = p(\neg T|\neg K) \approx \frac{\mathrm{tn}}{\mathrm{tn}+\mathrm{fp}} =$$



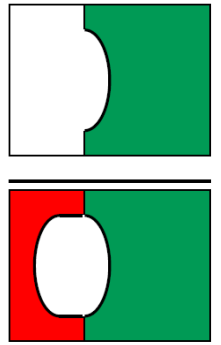
F-Score → harm. Mittelwert

$$\mathrm{Pr} = p(K\,|\,T) \approx \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}} =$$



,

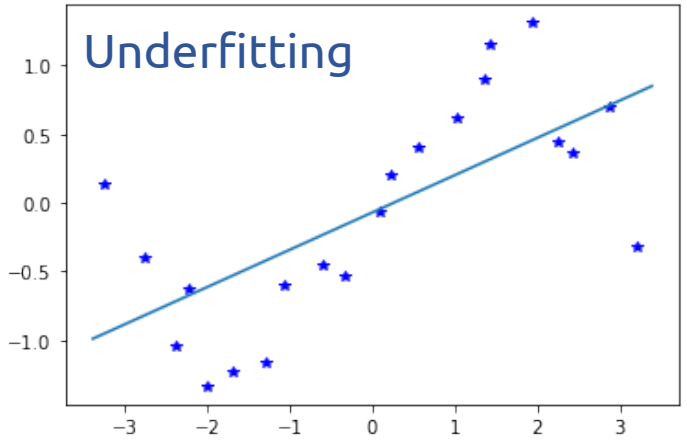
$$\mathrm{Tr} = p(\neg K\,|\,\neg T) \approx \frac{\mathrm{tn}}{\mathrm{tn}+\mathrm{fn}} =$$

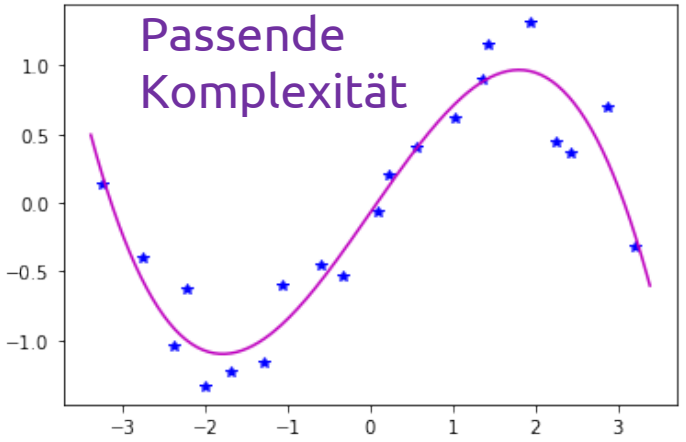


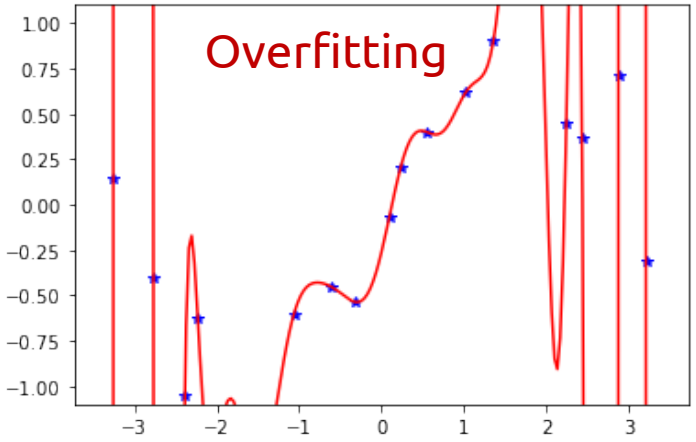
Under- und Overfitting (Unter- und Überanpassung)

- Wesentliche Aufgabe: Modellkomplexität (z.B. Zahl der Parameter) „richtig“ wählen.
  - Zu einfaches („starres“) Modell: Wichtige Eigenschaften werden nicht wiedergegeben (zu starker „*bias*“)
  - Zu komplexes („überflexibles“) Modell: Allzu detaillierte Anpassung an die Trainingsdaten („*variance*“), Generalisierung geht verloren
- Klassisches Beispiel: Polynominterpolation

Bias-Variance-Tradeoff



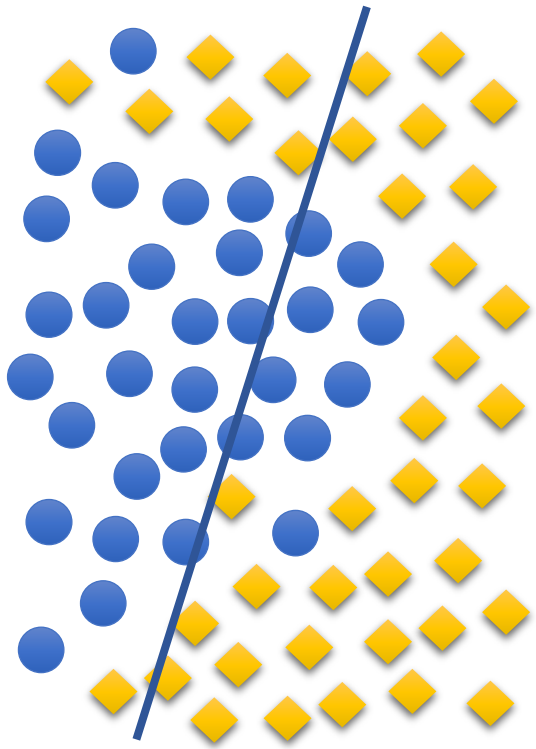




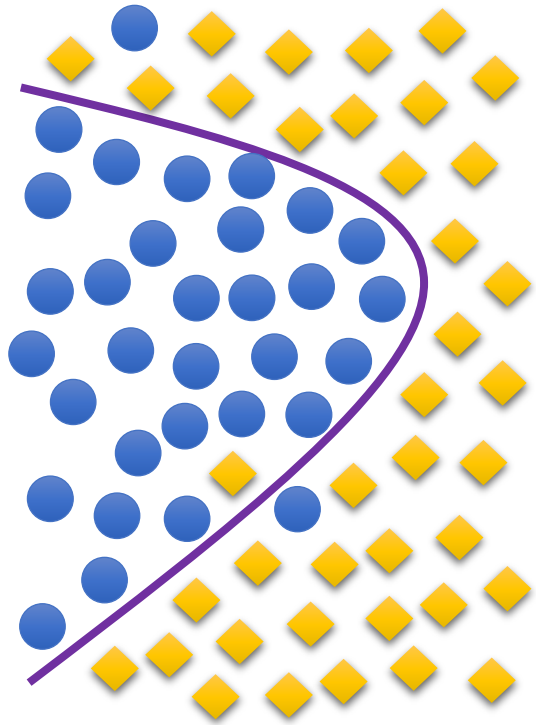
Under- vs. Overfitting

... ist auch für Klassifikationsaufgaben relevant.

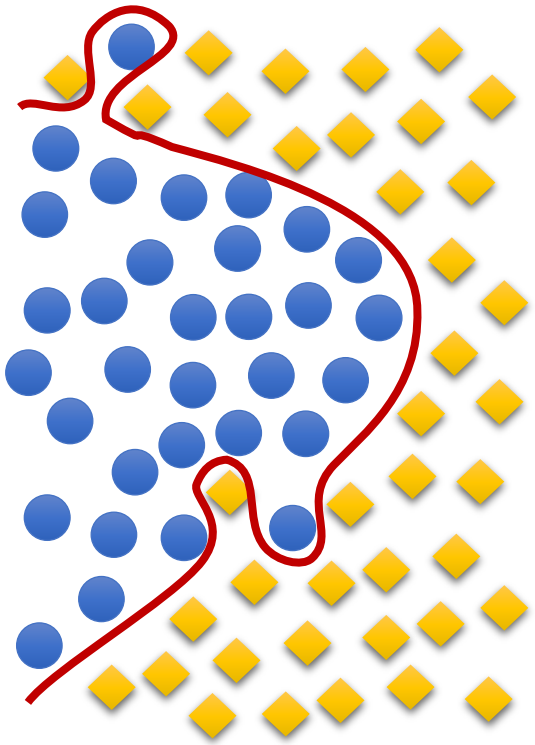
Aufgabe: Trenne mit einer einzelnen Kurve von



Underfitting



passende Komplexität

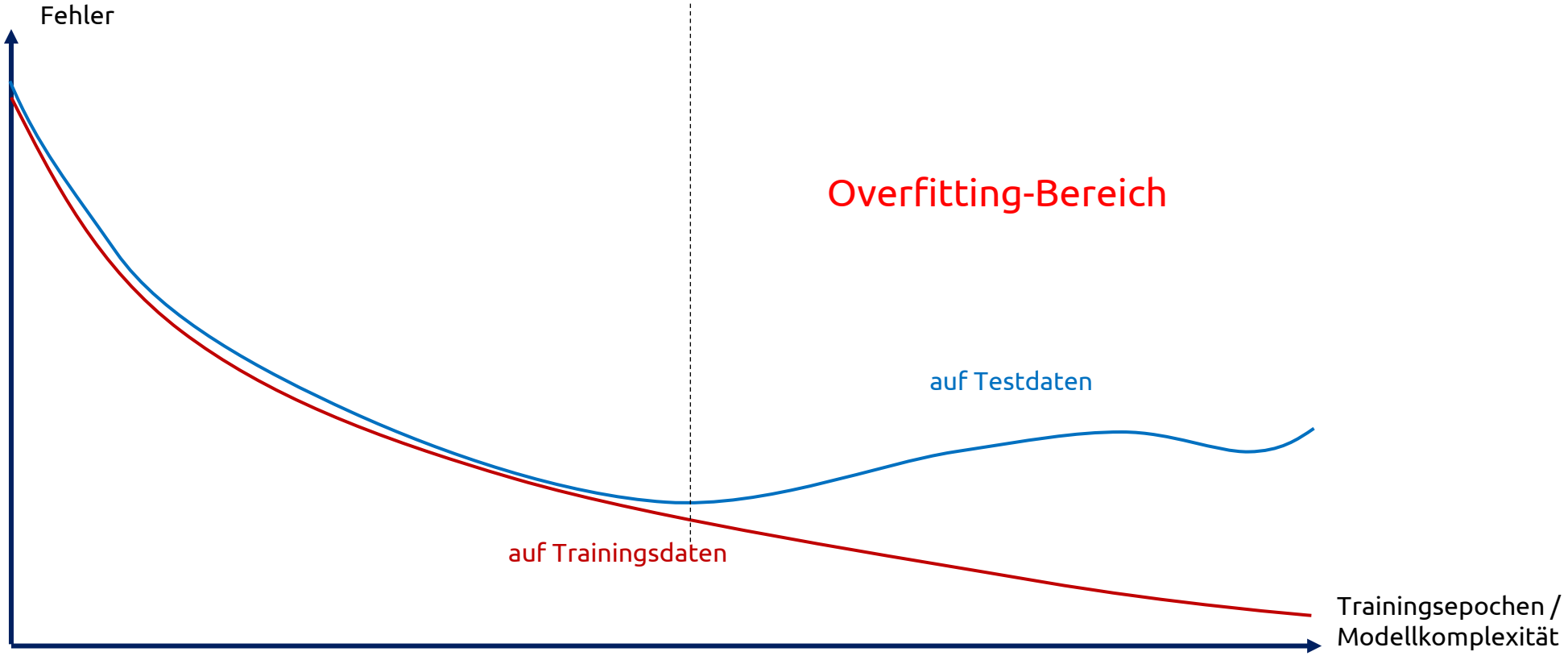


Overfitting

Strategien gegen Overfitting

Training eines Modells ist Optimierungsaufgabe,
Suche nach Minimum des Fehlers

Betrachte Fehler auf Trainingsdaten und auf Testdaten



Strategien gegen Overfitting

Alternative: **Kreuzvalidierung** (*k*-fold *cross validation*), vor allem bei geringen Datenmengen – Aufteilen in *k* Blöcke, jeweils „Rollentausch“

z.B. *k* = 3

Train-Test-Split:
Aufspaltung der ursprünglichen Trainingsdaten in einen Datensatz, mit dem tatsächlich trainiert wird, und einem, mit dem die Ergebnisse während des Trainings validiert werden.

Oft Verhältnis von 80% zu 20% gewählt; Aufspaltung der Daten sollte möglichst „zufällig“ erfolgen.

Finaler Check sollte nochmals mit „frischen“ (bislang nicht verwendeten) Daten durchgeführt werden, weil auch die Suche nach den den besten Hyperparametern (z.B. Modellkomplexität, Lernraten) ein Teil des Trainingsprozesses ist.

Manche Zugänge brauchen keinen spezielle Maßnahmen gegen Overfitting:
- Sehr einfache Modell mit wenigen Parametern
- Bayes‘sches Statistical Learning hat einen Overfitting-Schutz schon „eingebaut“

Weitere hilfreiche Strategien

Baseline-Modelle konstruieren:

- Am Anfang möglichst einfache (= robuste) Modelle definieren, mit denen weitere Vorhersagen verglichen werden können.
- *Beispiel*: Wettervorhersage (sehr komplex). Welche Baseline-Modell finden Sie für die Vorhersage des Wetters von morgen?
  - *Mittelwerte über vielen Jahre,*
  - *Bauernkalender („Expertenregeln"),*
  - *Persistenzmodell: „Wetter gleich wie am Vortag"*

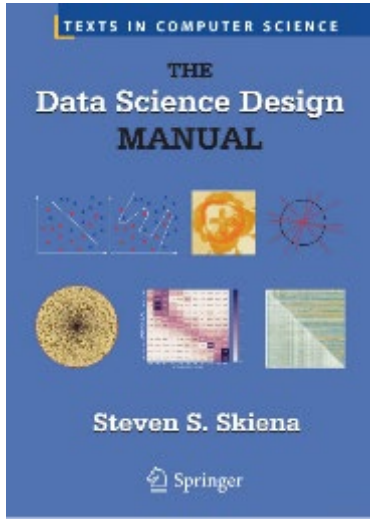
Plausibilitätschecks: Modell an besonders einfachen Problemen bzw. mit selbst konstruierten synthetischen Daten testen.

Herausforderungen

- Datenschutz:
  - Profiling („gläserner Mensch“): Überwachung / Manipulation;
    Gefahr vor allem durch Verknüpfung von Daten aus vielen verschiedenen Quellen.
  - Training mit Daten von verschiedenen Eigentümern (Federated Learning)
- Gefahren durch „starke KI“ („human-like AI“, „general AI“)
- Gesellschaftliche Transformation: womöglich Verschwinden von ganzen Berufsgruppen und Geschäftsbereichen
- *Biases*: Verzerrungen von KIs durch Training mit verzerrten Daten (z.B. fehlende Balance zwischen Geschlechtern, Ethnien, …)
- Vertrauenswürdige KI: Wie kann man KI-Entscheidungen begründen? Balance zwischen Erklärbarkeit und Performance finden.
- Nachhaltigkeit: Training von Deep-Learning-Systemen erfordert u.U. sehr viel Energie – Hardware-Verbesserungen, fundierte Methodenauswahl
- Manipulationen (Deep Fakes, Adversarial Attacks, …)

siehe aktuell z.B. Bilder von DALL-E

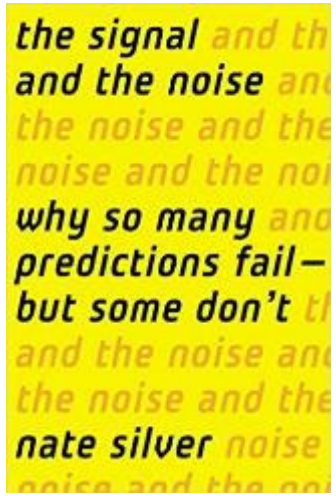
Literatur: Einstiege



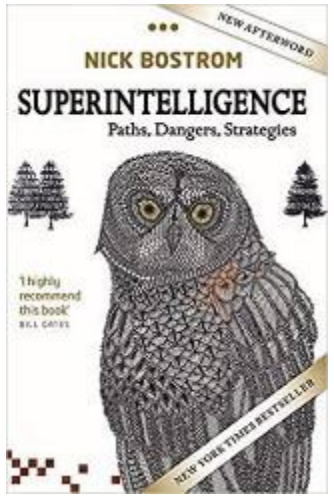
Skiena: *The Data Science Design Manual:* Sehr gutes Buch für den Einstieg in den Bereich *Data Science*, mit grundlegender Diskussion vieler zentraler Konzepte, inkl. „Metathemen“. Geht klarerweise bei den Methoden nicht sehr in die Tiefe, liefert aber ein sehr prägnantes „*big picture*“ und diskutiert viele Anwendungsfälle samt der Probleme und Schwierigkeiten.

Ertel: *Grundkurs künstliche Intelligenz*:
Eher „anfänger:innenfreundliche“ Einführung in diverse Methoden; relative umfangreiche Behandlung von klassisch-symbolischen KI-Ansätzen (formale Logik, Theorembeweiser), schöne Darstellung des Entscheidungsbaum-Lernens, Grundzügen der neuronalen Netze und Bayes-Netze

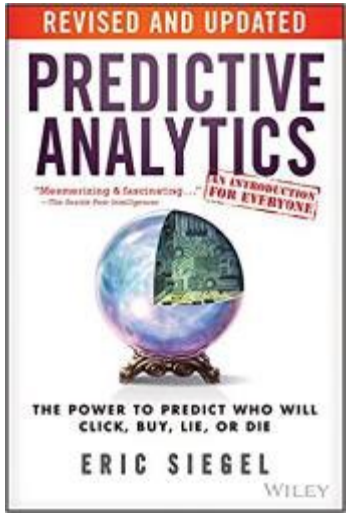
Popular Science; Gefahren



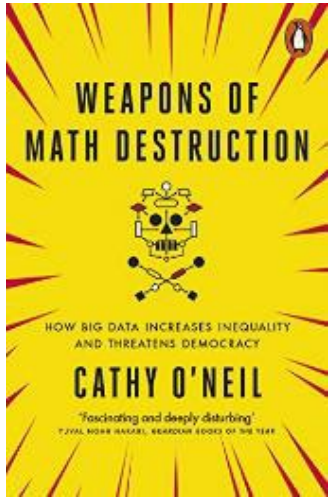
Silver: *The Signal and the Noise: Why so many predictions fail – but some don't*: Gut geschriebener (populärer, aber trotzdem sehr lehrreicher) Einblick in angewandte Prognosemethoden.



Bostrom: *Superintelligence*: Darstellung der Gefahren durch dem menschlichen Verstand überlegene Künstliche Intelligenz und Diskussion verschiedener Szenarien.



Siegel: *Predictive Analytics – The Power to Predict who will Click, Buy, Lie or Die:* Gut lesbare Einführung in das Gebiet der *Predictive Analytics*, der Anwendung von ML-Methoden insbesondere im Business-Bereich.



O'Neil: *Weapons of Math Destruction*: Kritischer Blick auf *Big Data* und *Machine-Learning*-Methoden hinsichtlich Datenschutz, Diskriminierung, Manipulationsmöglichkeiten sowie der Wirkung kognitiver und statistischer Verzerrungen bei der Erzeugung selbsterfüllender Prophezeiungen.

Bitte um Review/Feedback

Link: https://fragebogen.joanneum.at/dihsued_feedback/?q=base&r=BB361928

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN