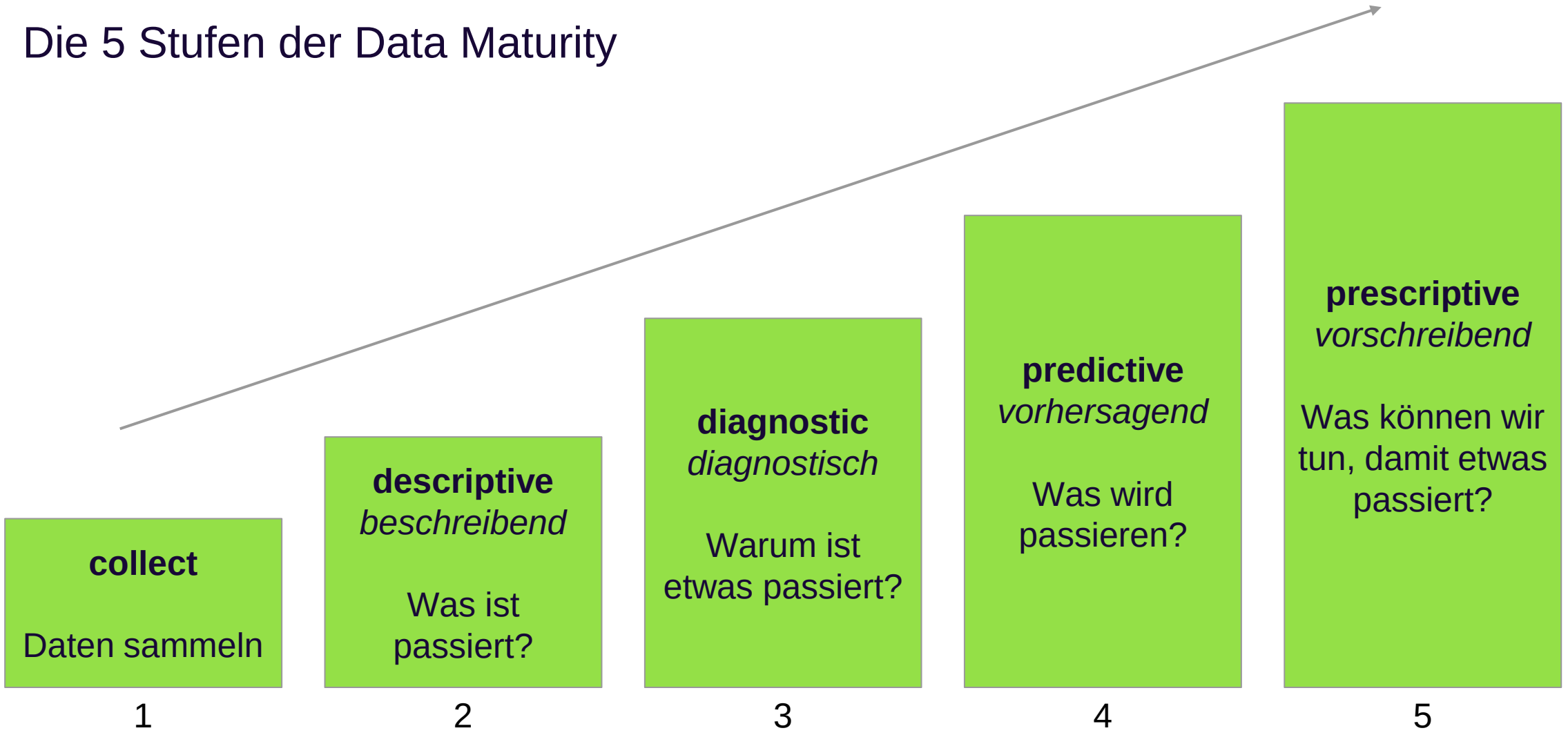
DataCraft

Data Maturity

DataCraft

Die 5 Stufen der Data Maturity



DataCraft

Stufe 1: **Daten sammeln**

Stammdaten

Daten aus operativen Prozessen

Marktforschung

Social Media

Konkurrenz (Webscraping)

Servicedienste(z.B. Wetter)

Bewertungs-plattformen

DataCraft

Stufe 2: **descriptive** – Was ist passiert?

- Wie viele Kunden waren letzte Woche in meiner Filiale?
- Wie viele Waren hatten wir gestern im Lager ?
- Welchen Umsatz habe ich erzielt?
- Welche Ware war besonders begehrt?
- Welche Temperatur/Luftfeuchtigkeit herrschte im Museum?
- Welche Größe hatten die gefangenen Fische?
- Wie viele Mitarbeiter haben wir zur Verfügung?

DataCraft

Stufe 3: **diagnostic** – Warum ist etwas passiert?

- Gestern waren zu wenig Mitarbeiter in der Fabrik, weil es einen Unfall gab
- Wir mussten viel Ware weg schmeißen, weil eine Baustelle vor dem Laden war
- Die Anzahl vermieteter Boote war letzte Woche gering, weil das Wetter regnerisch war
- Öl und Reis sind immer mittags schon ausverkauft, weil die Nachfrage wegen des Kriegs groß ist
- Diebstahlrate ist höher als in einem anderen Bezirk, weil die Armuts-Rate hoch ist

DataCraft

Stufe 4: **predictive** – Was wird passieren?

- Wie hoch sind die Eisverkäufe morgen?
- Wie wird das Wetter in den nächsten Tagen?
- Mit welcher Kundenanzahl kann ich rechnen?
- Wie viel Lachs wird in der nächsten Saison zur Verfügung stehen?
- Wie hoch wird die Lachsnachfrage in der kommenden Saison?
- Wie viele Patienten kommen mit einem Beinbruch in den nächsten Tagen in die Notaufnahme?

DataCraft

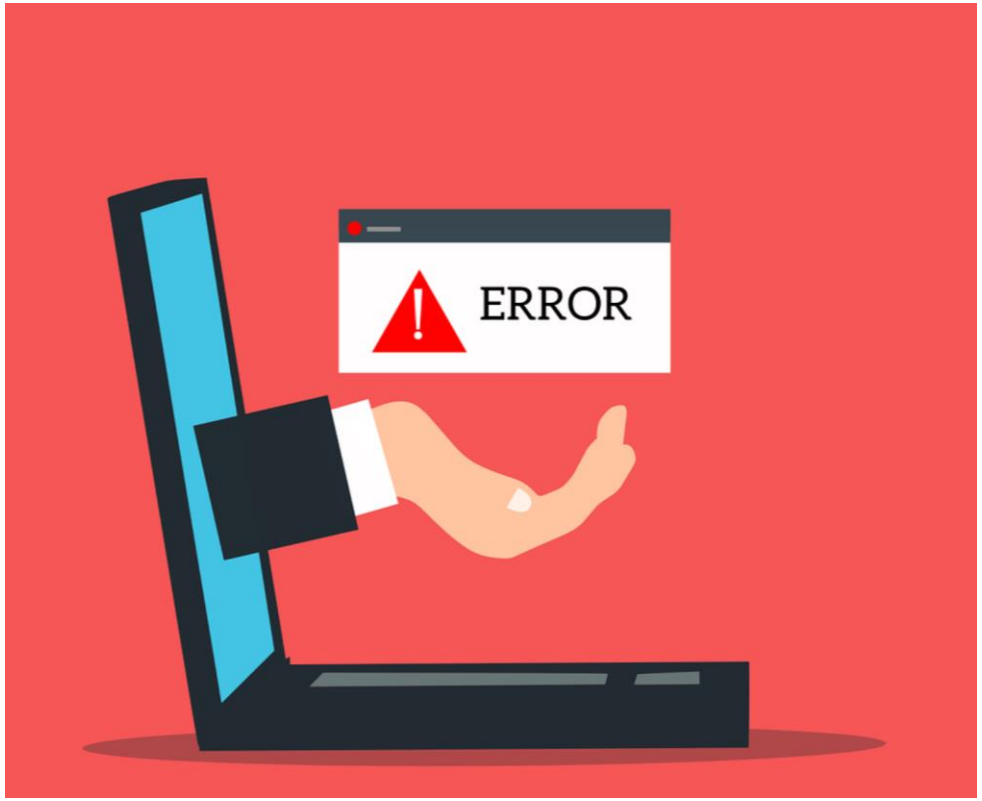
Stufe 5: **prescriptive** – Was können wir tun, damit etwas passiert?

- Um den Umsatz zu erhöhen, können wir einen Artikel als Aktion anbieten
- Um die Kundenfluktuation aufzuhalten können wir mehr Mitarbeiter in die Schicht schicken
- Durch Social Media Werbung wird die Kundenzahl im Bootsverleih erhöht
- Durch Recommendation Engine ("Kunden kauften auch") kann der Warenkorbwert erhöht werden
- Durch ein Bewertungssystem kann mehr Vertrauen geschaffen werden, was zu mehr Käufen führt
- Wir können aktuelle Trends nutzen, um den Absatz zu erhöhen
- Werbemittelallokation (Verteilung des Budgets), um mehr Umsatz zu machen als letztes Jahr

Typische Fehler im Umgang mit Daten

DataCraft

- Garbage In - Garbage Out
- zu starke Komplexitätsreduktion / eindimensionale KPIs
- Stichprobengröße / Datenmenge
- Overfitting
- Bias / Verzerrung
- Korrelation vs. Kausalität
- p-Wert Hacking / exzessive Suche nach Mustern



Garbage In – Garbage Out

DataCraft

- **Was können Gründe für fehlerhafte Daten sein?**

- verwechselte Felder
- falscher Datentyp (1 oder 1.0)
- Datumsformate (18.05.22, 18.05.2022, 2022-05-18, 2022/18/5)
- Zeitangabe: Zeitzone!
- äöü und Sonderzeichen (UTF-8)
- Dubletten
- Ausreißer/Ausnahmen
- nur Sonderzeichen bzw. nicht weiter verarbeitbar

DataCraft

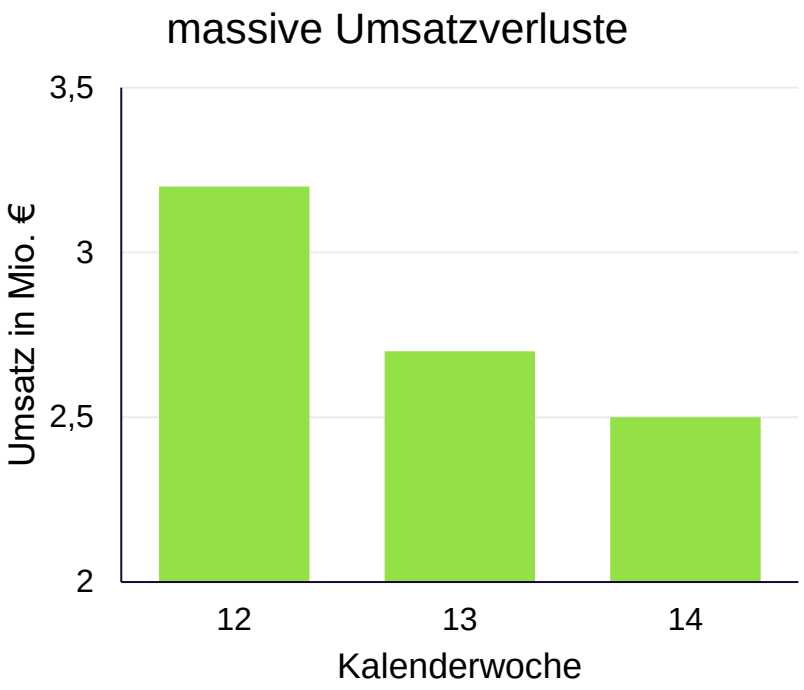
Beispiele

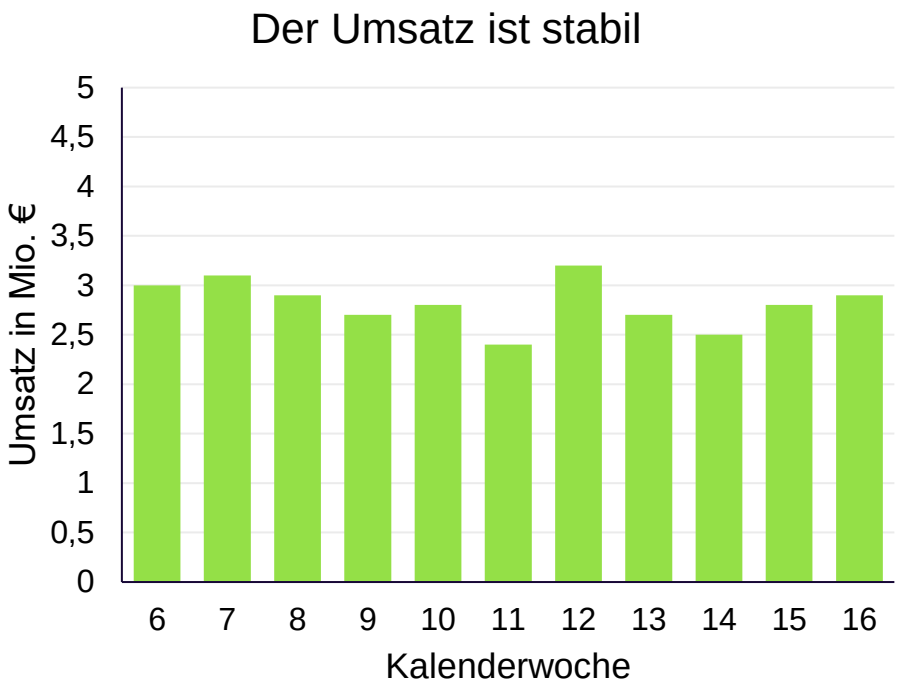
- Anzahl Telefonate für Callcentermitarbeiter: Problem Qualität wird nicht berücksichtigt, Mitarbeiter bekommt den Anreiz, schnell wieder aufzulegen, Zufriedenheit des Kunden
- Einzelhandel: Verkaufszahlen, aber Warenkorbwert viel höher
- Fussball: Bewertung der einzelnen Spieler nach Ergebnis. Problem bei Ersatzspielern: kaum Einsätze
- Absturzsicherung: Dicke der Nähte

Durch KPIs in Reportings werden Anreize geschaffen

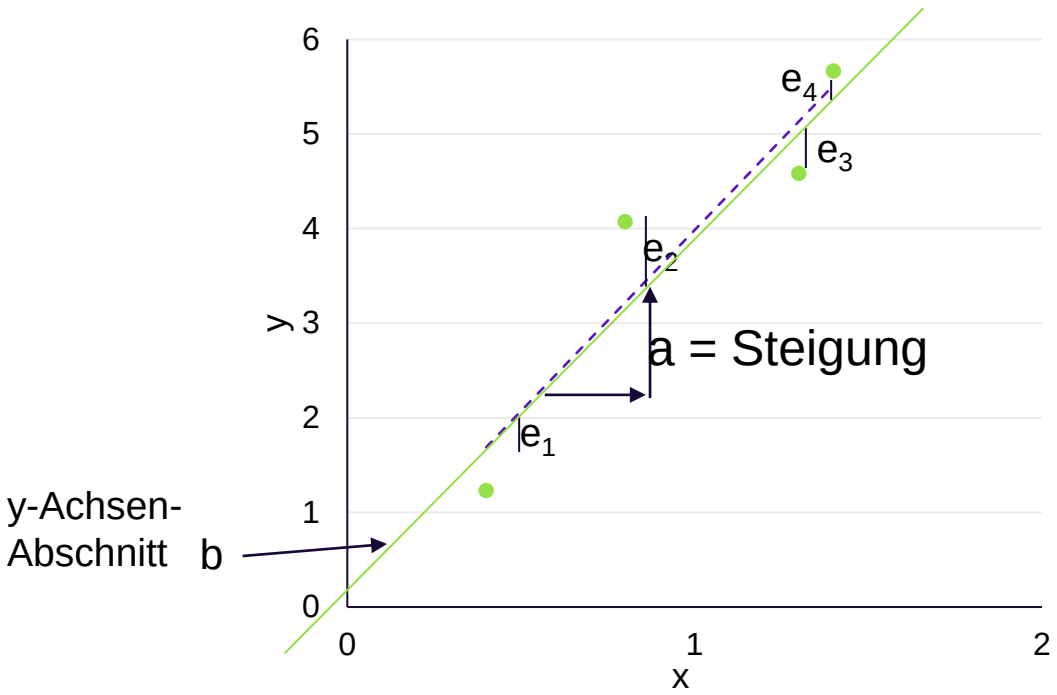
Stichprobengröße

- zu wenige Daten
- nur passende Beispiele raussuchen





Lineare Regression = Methode der kleinsten Quadrate
Lege eine Gerade durch die Punkte, so dass die Summe der quadrierten Fehler möglichst klein ist



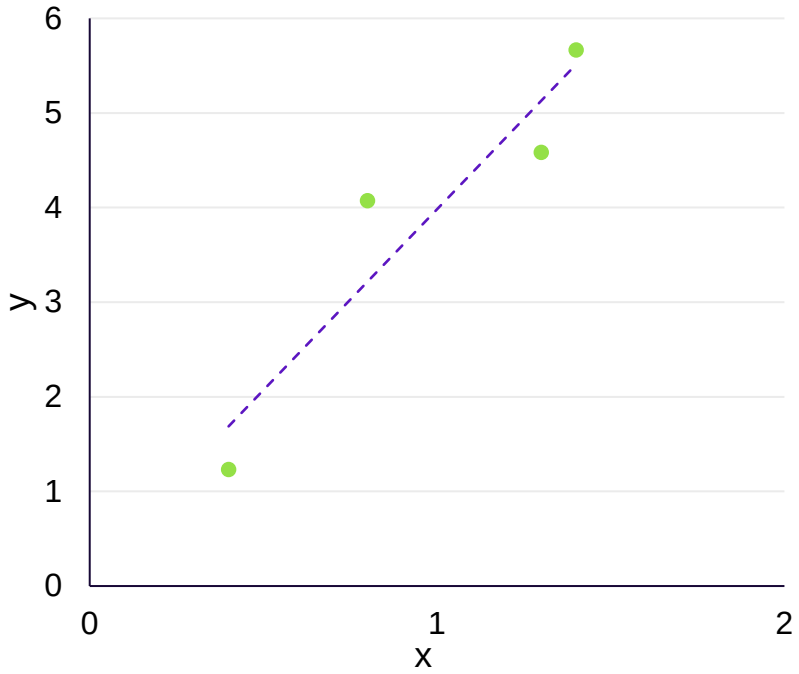
Geradengleichung: $y = a*x + b$

Finde a und b, so dass
$e_1^2 + e_2^2 + e_3^2 + e_4^2$
minimal wird

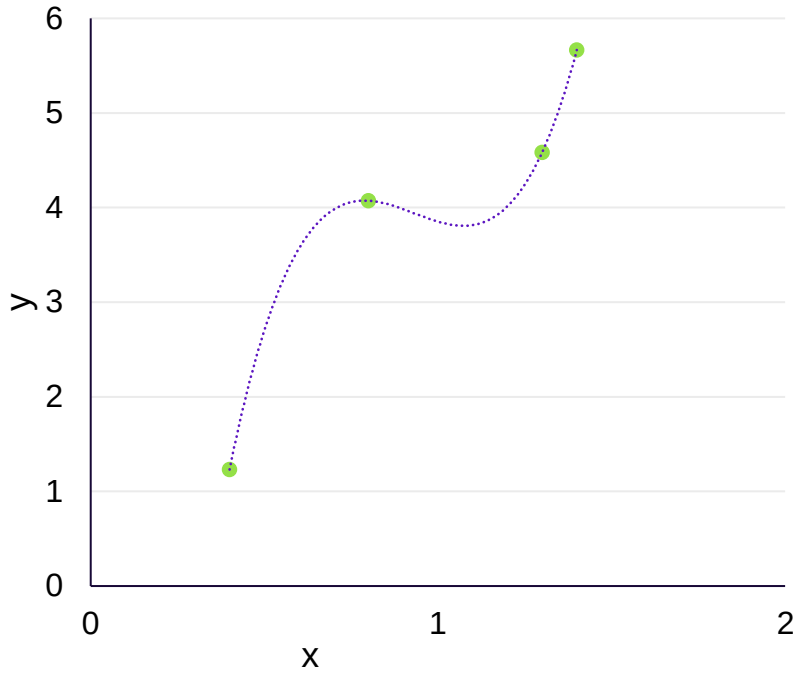
DataCraft

Overfitting = Überanpassung an die vorhandenen Daten

Gerade

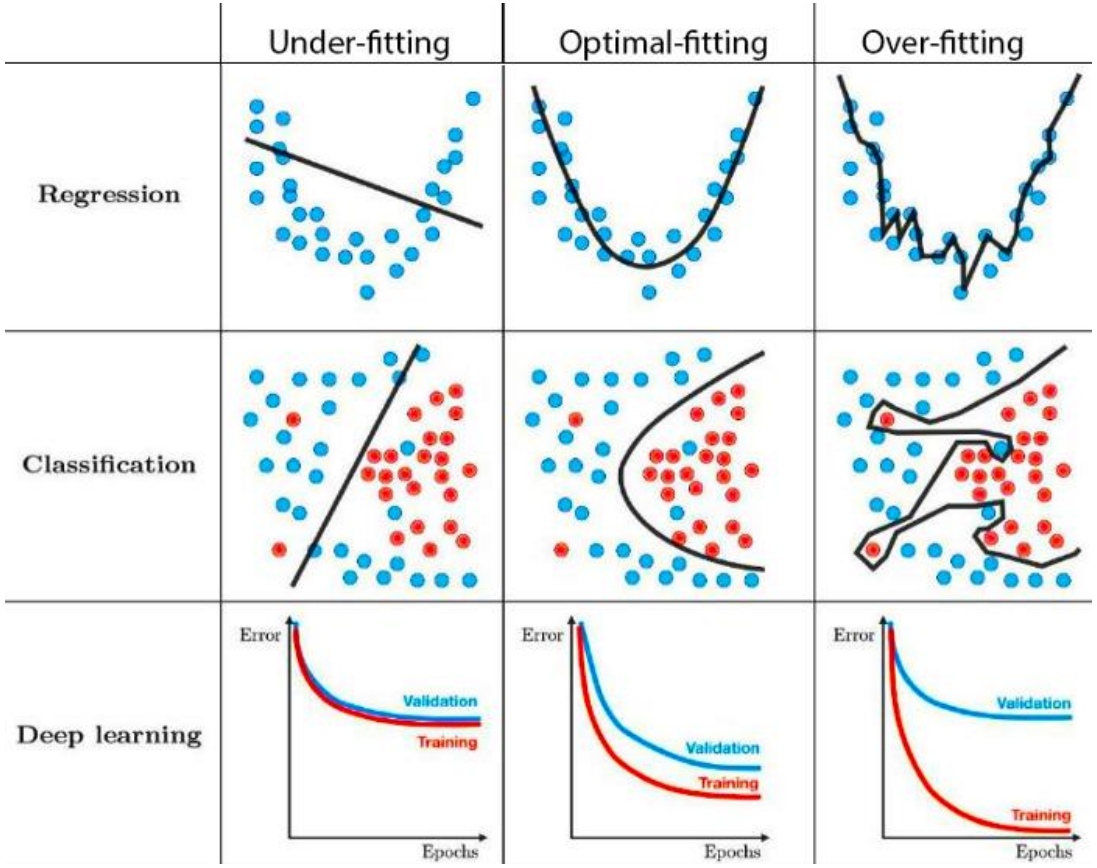


Polynom 3. Grades



DataCraft

Overfitting = Überanpassung an die vorhandenen Daten



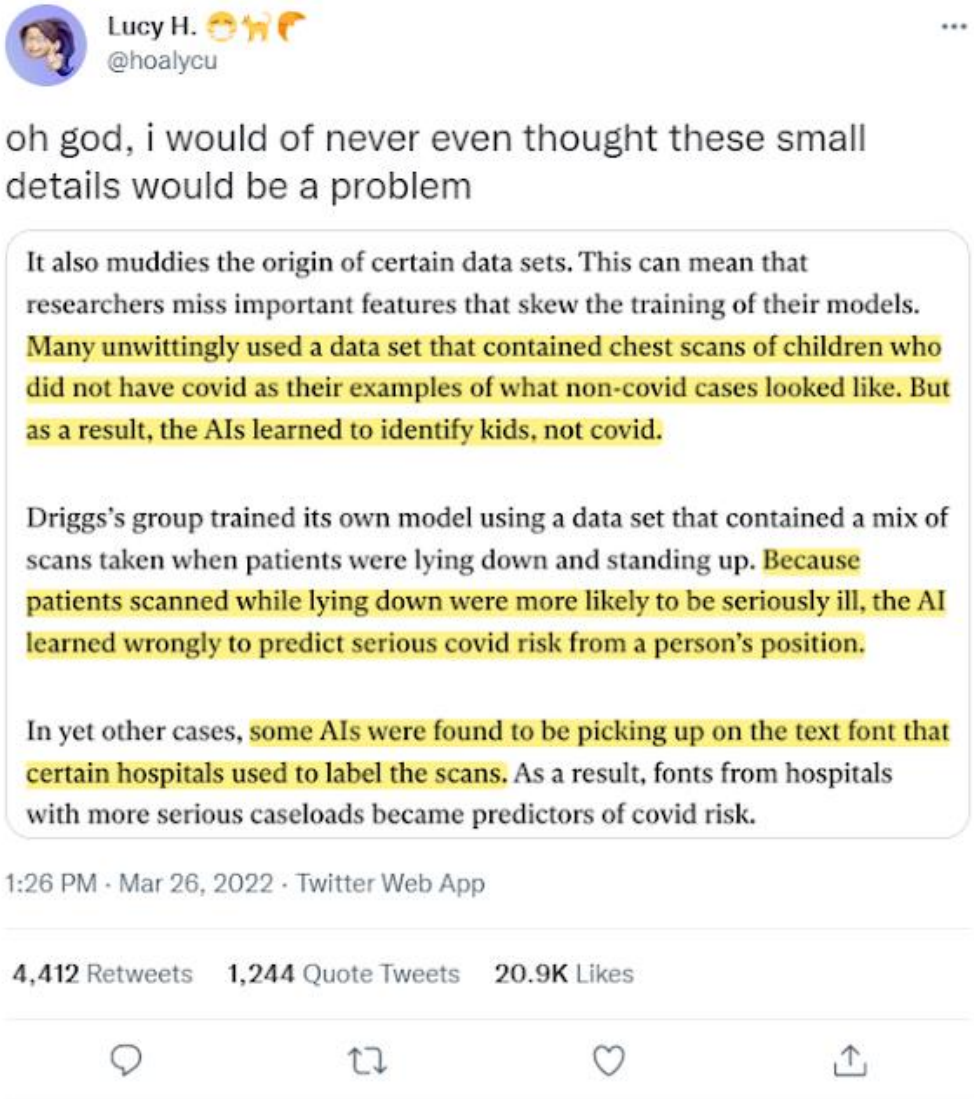
DataCraft

Bias = Verzerrungen / Vorurteile

Die Gefahr ist, dass das nicht bewusst passiert

- **Confirmation Bias**: Suche nach Beispielen/Fakten, die unsere Meinung bestätigen
- **Availability Bias:** Präsenz im Gedächtnis
- **Selection Bias:** Auswahl der Testgruppe, z.B. Gehaltsangaben in Umfragen
- **Historical Bias:** Erkenntnisse anhand vergangener Daten
- **Survivorship Bias:** nur Überlebende, z.B. erfolgreiche Unternehmer
- **Outlier Bias:** Im Mittel sieht alles gut aus

DataCraft



Lucy H.
@hoalycu

oh god, i would of never even thought these small details would be a problem

It also muddies the origin of certain data sets. This can mean that researchers miss important features that skew the training of their models. Many unwittingly used a data set that contained chest scans of children who did not have covid as their examples of what non-covid cases looked like. But as a result, the AIs learned to identify kids, not covid.

Driggs's group trained its own model using a data set that contained a mix of scans taken when patients were lying down and standing up. Because patients scanned while lying down were more likely to be seriously ill, the AI learned wrongly to predict serious covid risk from a person's position.

In yet other cases, some AIs were found to be picking up on the text font that certain hospitals used to label the scans. As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.

1:26 PM · Mar 26, 2022 · Twitter Web App

4,412 Retweets 1,244 Quote Tweets 20.9K Likes

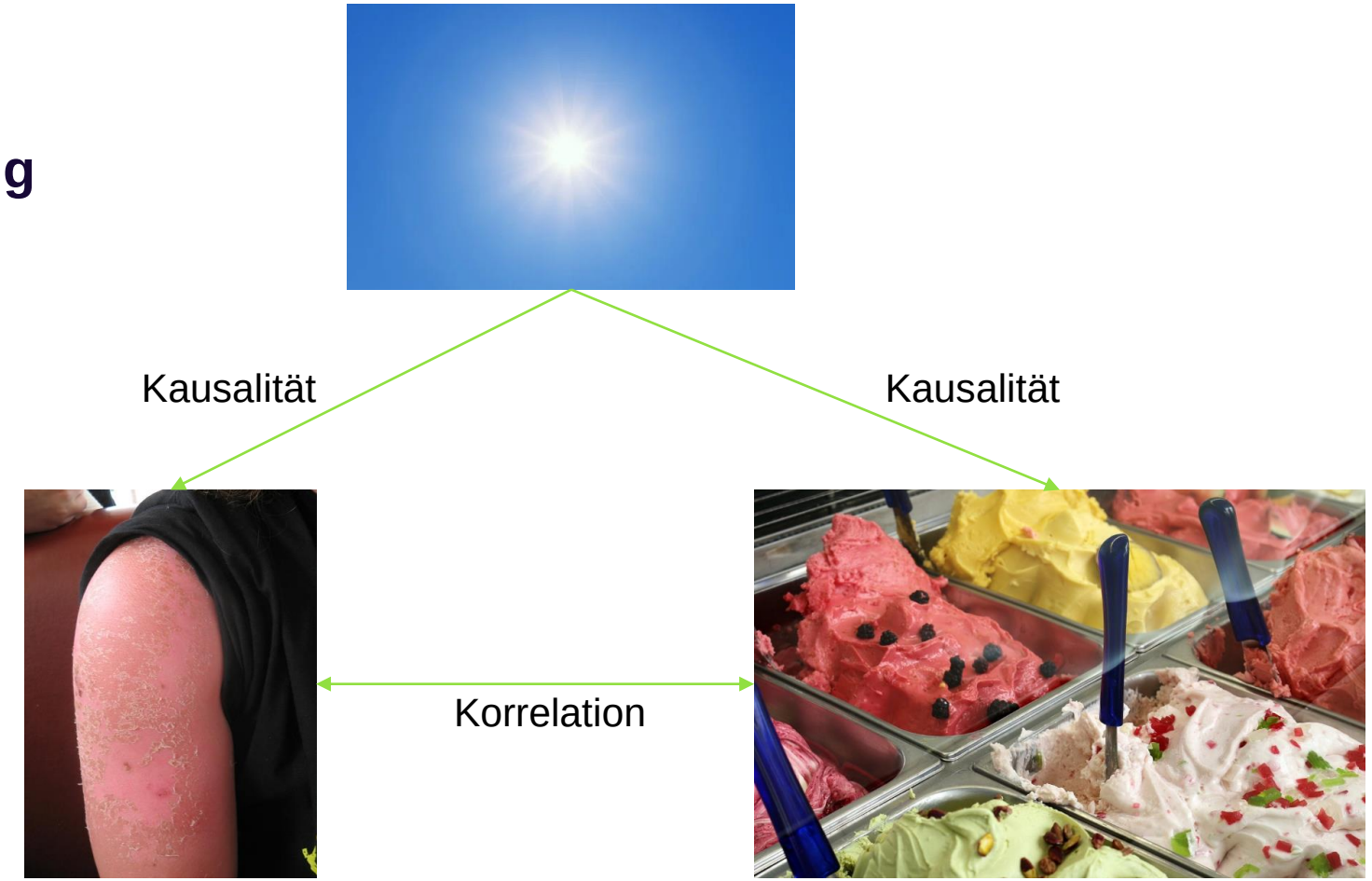
Korrelation vs. Kausalität

DataCraft

Korrelation ≠ Kausalität

Korrelation = Zusammenhang zwischen Variablen

Kausalität = Ursache



p-Wert Hacking

DataCraft

- p-Wert ist eine statistische Maßzahl zwischen 0 und 1, die bei Hypothesentests verwendet wird
- Der p-Wert gibt die Wahrscheinlichkeit an, dass die Daten unter der sogenannten Nullhypothese beobachtet werden. Ist diese Wahrscheinlichkeit klein, wird die Nullhypothese abgelehnt.
- Werden nun viele Tests durchgeführt, dann ist die Wahrscheinlichkeit, dass einer der Tests falsch ist, groß.

Test	Wahrscheinlichkeit, dass das Test-Ergebnis falsch ist	W., dass eines der Ergebnisse falsch ist
1	5%	5%
2	5%	9,75%
3	5%	14,26%
4	5%	18,55%
5	5%	22,62%
6	5%	26,49%

Allgemeiner: Suche so lange in den Daten, bis Du ein Muster bzw. eine Auffälligkeit findest!