



Digital Applications & Data Management

WS25/26

Dr. Jens Kohl



Roadmap Vorlesung

1. Einführung und Übersicht
2. Grundlagen Data Science
3. Vorgehen Data Science Use Case
4. Case Study Data Science
5. Grundlagen unüberwachtes Lernen
6. Grundlagen überwachtes Lernen
(tabellarische Daten)
7. Case Study überwachtes Lernen
(tabellarische Daten)
8. Grundlagen überwachtes Lernen (Bilddaten)
9. Case Study überwachtes Lernen und Transfer Learning (Bilddaten)
10. Grundlagen Generative AI
11. Generative AI mit Texten und Prompt Engineering
12. Agentic AI
13. Ausblick: Machine Learning in der Cloud und Reinforcement Learning



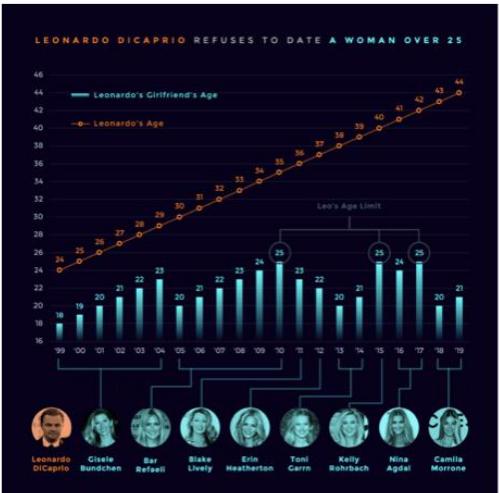
Vorlesung 2: Grundlagen Data Science



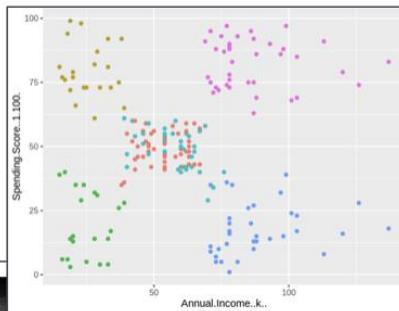
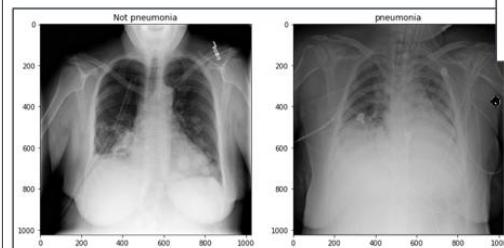
Was machen wir heute?

Motivation

Data Science: Daten analysieren, um daraus Erkenntnisse zu gewinnen und diese zu visualisieren.



Machine Learning: aus Daten ein Modell lernen, das eigenständig Erkenntnisse gewinnt.

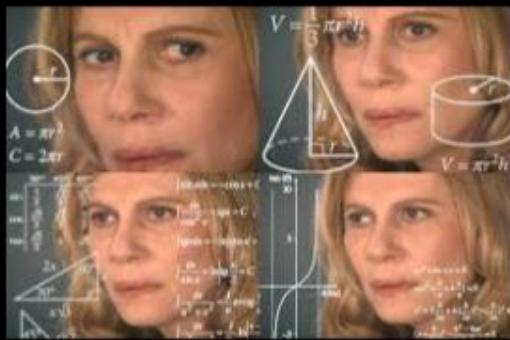




Data Science

Was ist das?

what my friends think I do



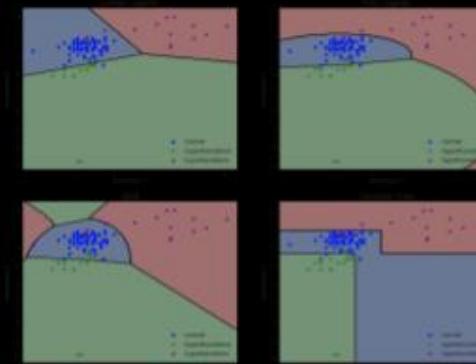
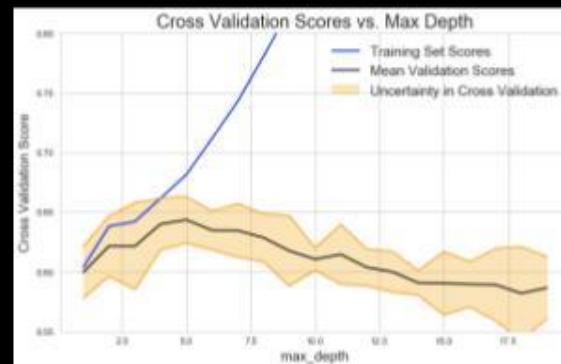
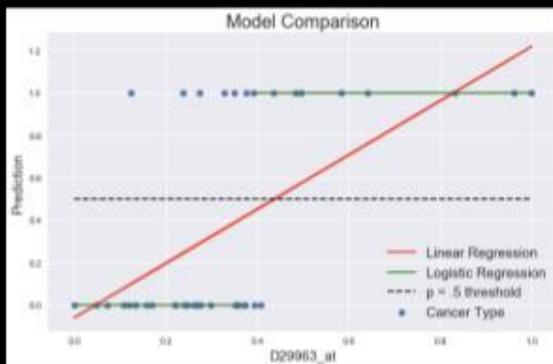
what my family thinks I do



what society thinks I do



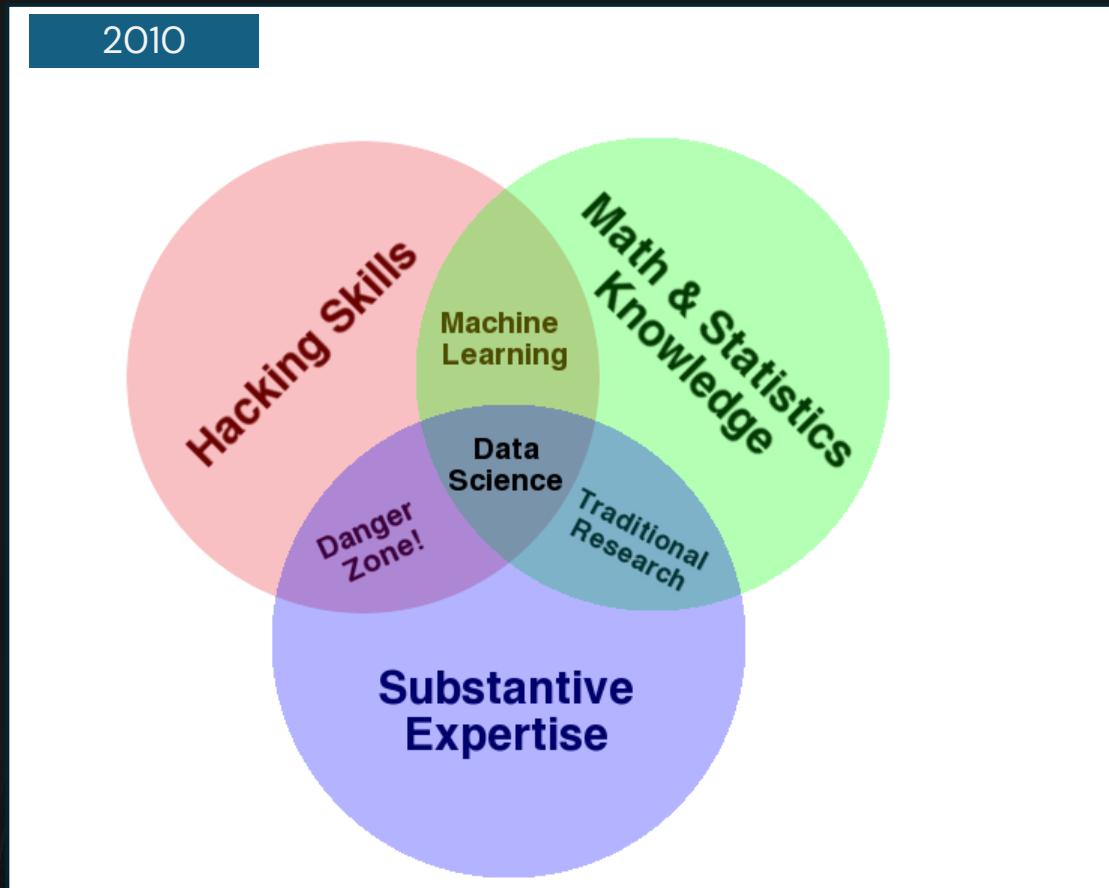
what I actually (will) do in Data Science 1



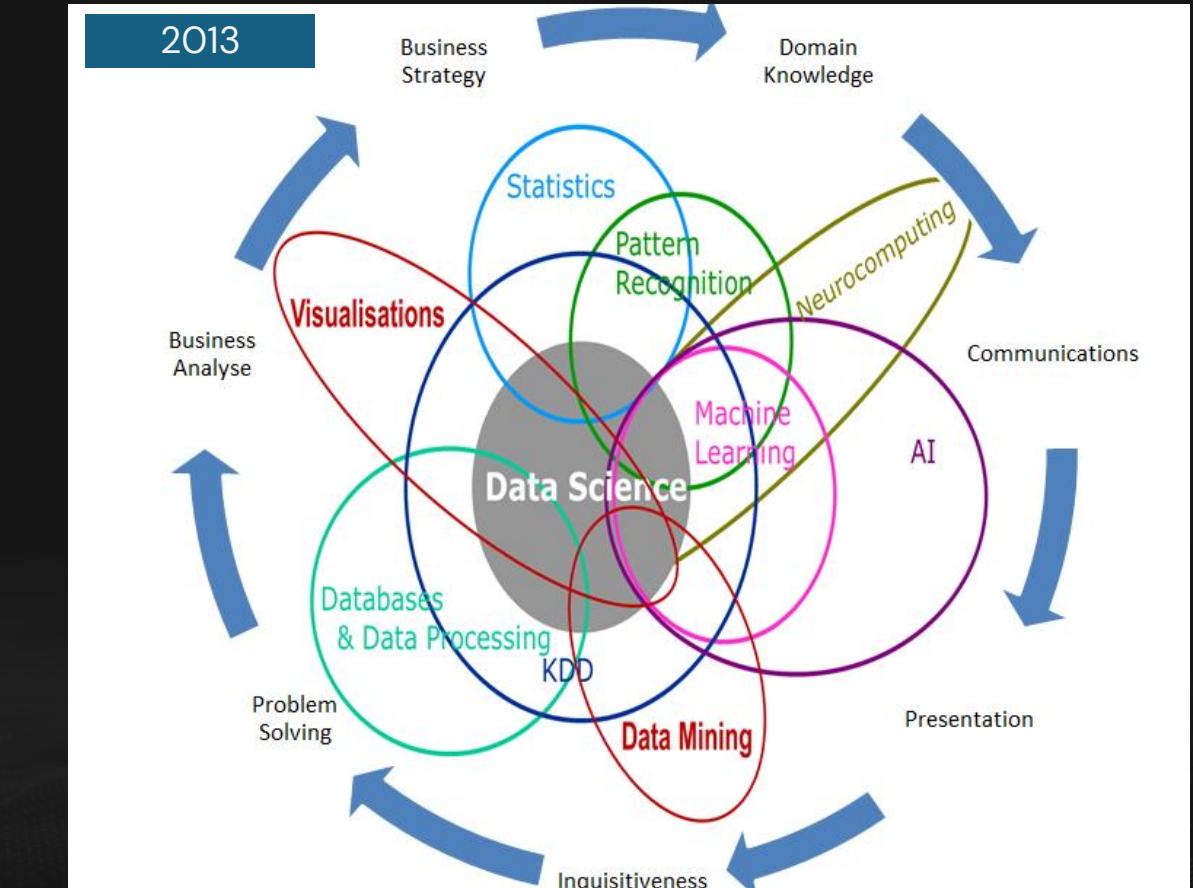


Data Science

Starke Änderung für Verständnis des Begriffs sowie Umfangs Data Science



Quelle: Drew Conway 2010, verfügbar unter: [Link](#)

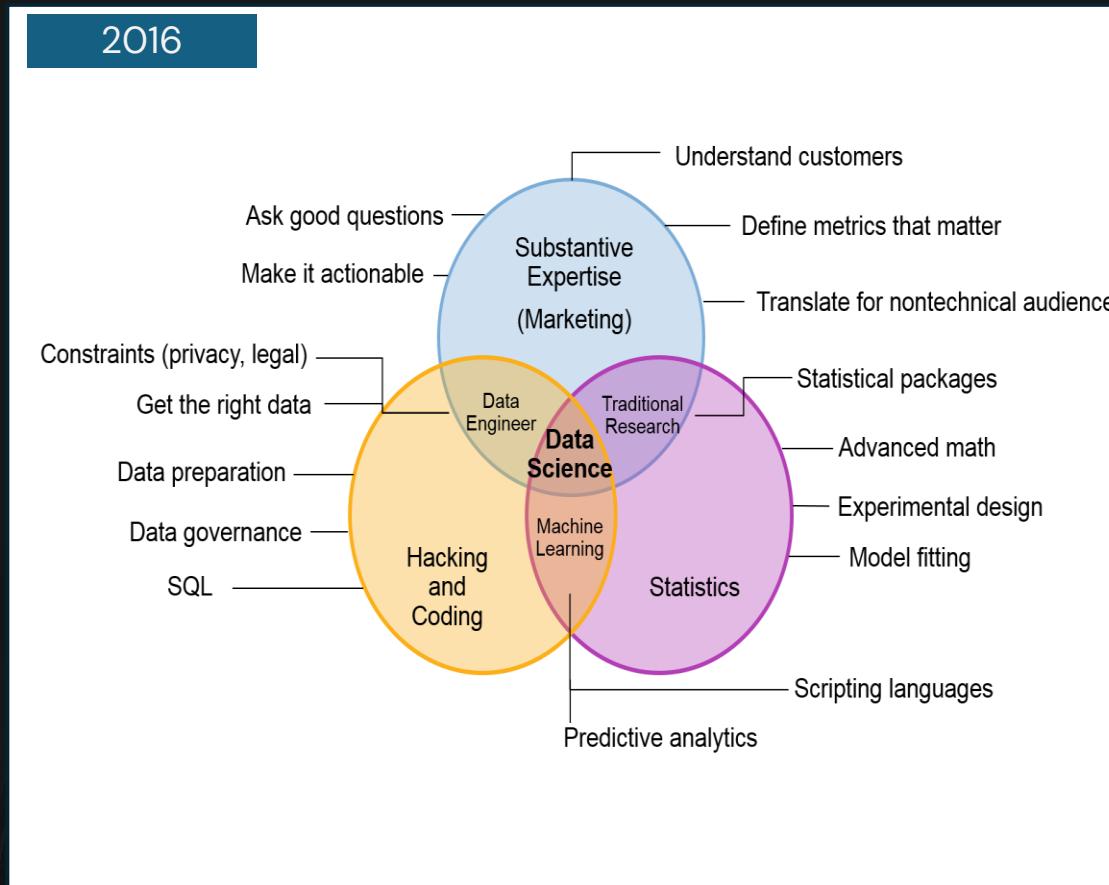


Quelle: B. Tierney, 2013, verfügbar unter: [Link](#)



Data Science

Starke Änderung für Verständnis des Begriffs sowie Umfangs Data Science



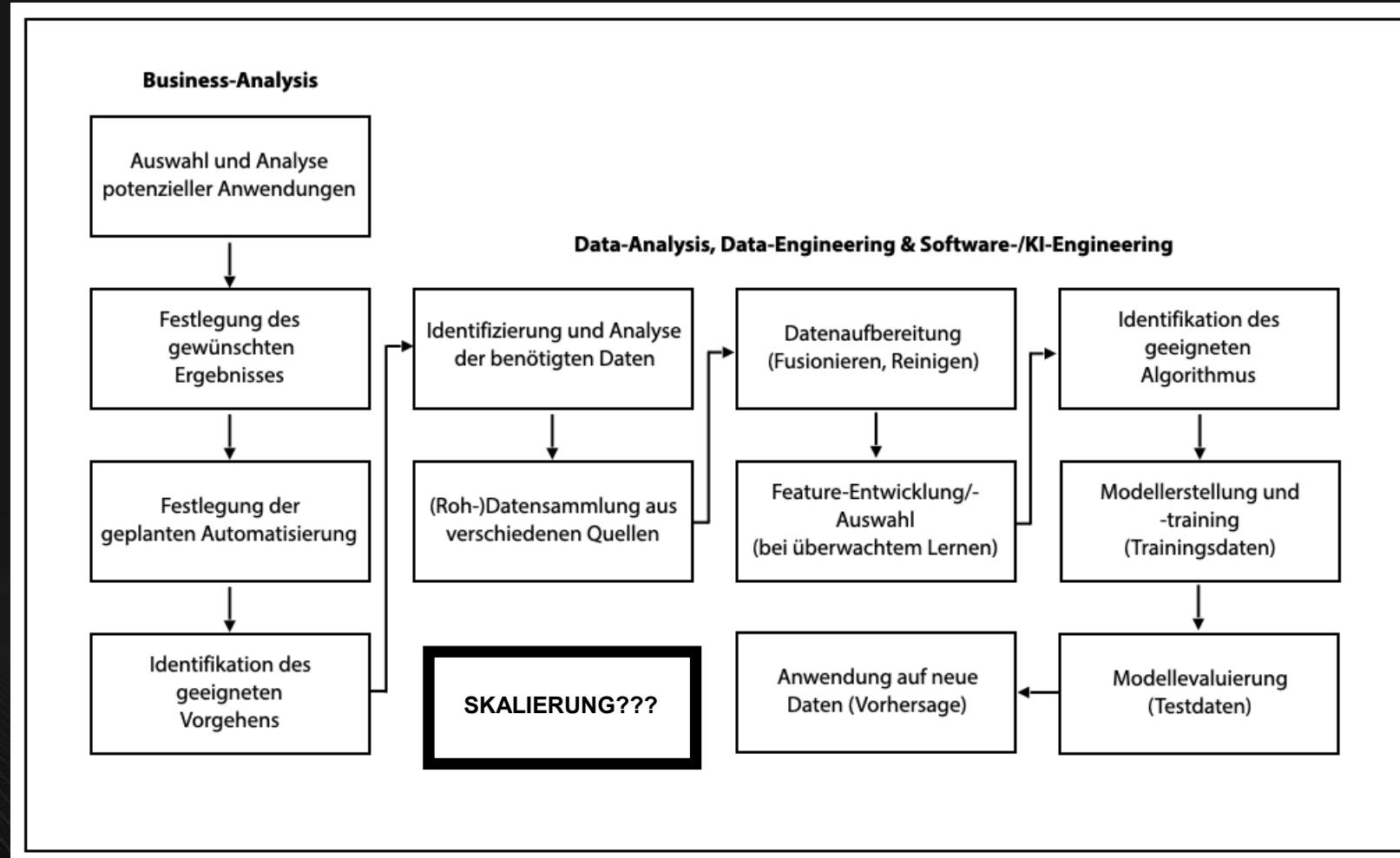
Quelle: Gartner 2016, verfügbar unter: [Link](#)

Quelle: NIST big data workgroup, 2019, verfügbar unter: [Link](#)



Data Science

Ablauf eines Data Science Use Case

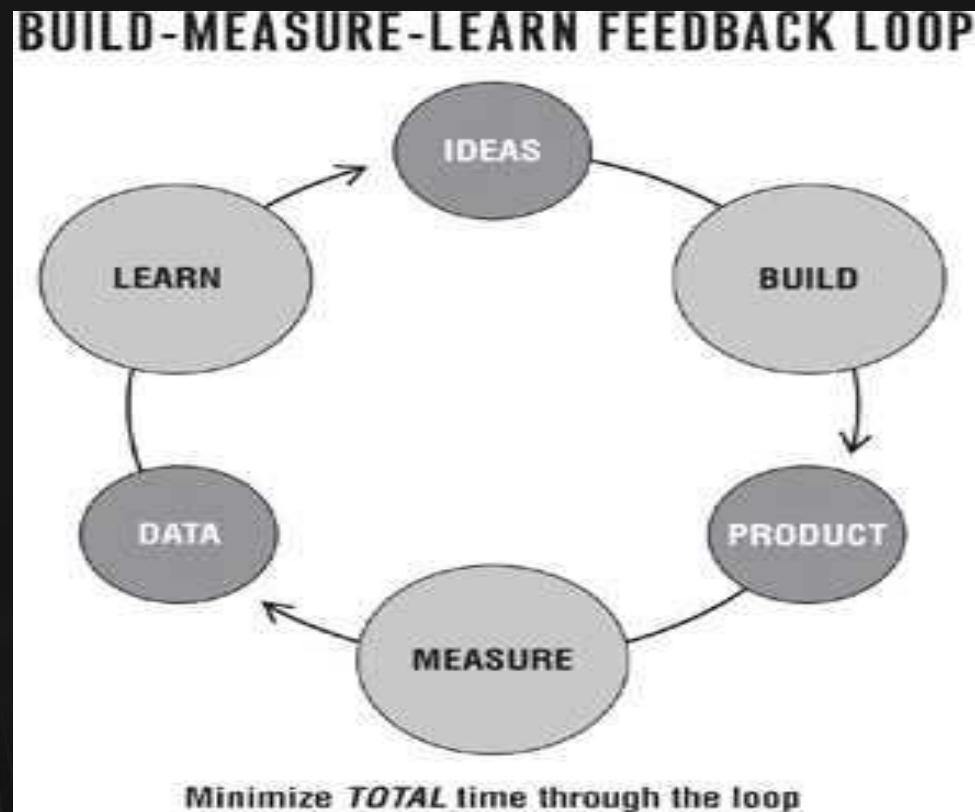




Data Science

Vorgehensweise Use Case

Vorgehensweise „data-driven company“



Quelle: E. Ries, „The Lean Start-up“, 2011

Generische Vorgehensweise Data Science

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

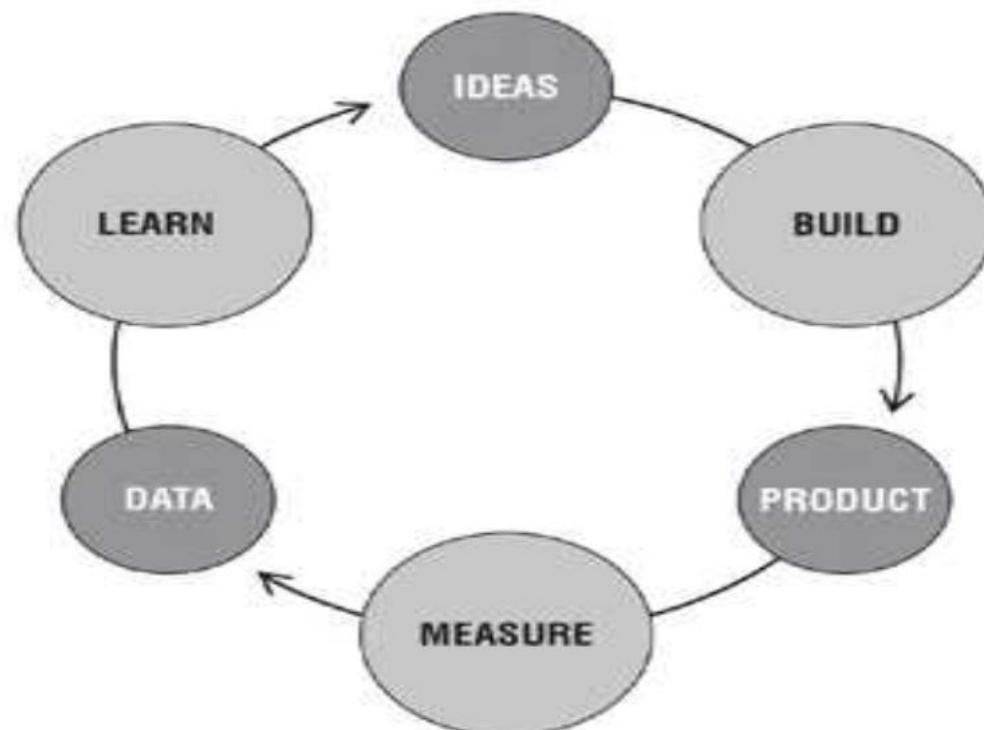
Quelle: Protopapas, Rader, Tanner, CS109 Data Science, 2020, [Link](#)



Build-measure-learn feedback loop

Übersicht

BUILD-MEASURE-LEARN FEEDBACK LOOP



"The fundamental activity of a startup is to turn ideas into products, measure how customers respond, and then learn whether to pivot or persevere. All successful startup processes should be geared to accelerate that feedback loop".

"Startups exist not just to make stuff, make money, or even serve customers. They exist to **learn how to build a sustainable business**. This **learning** can be **validated** scientifically by running **frequent experiments** that allow entrepreneurs to test each element of their vision."



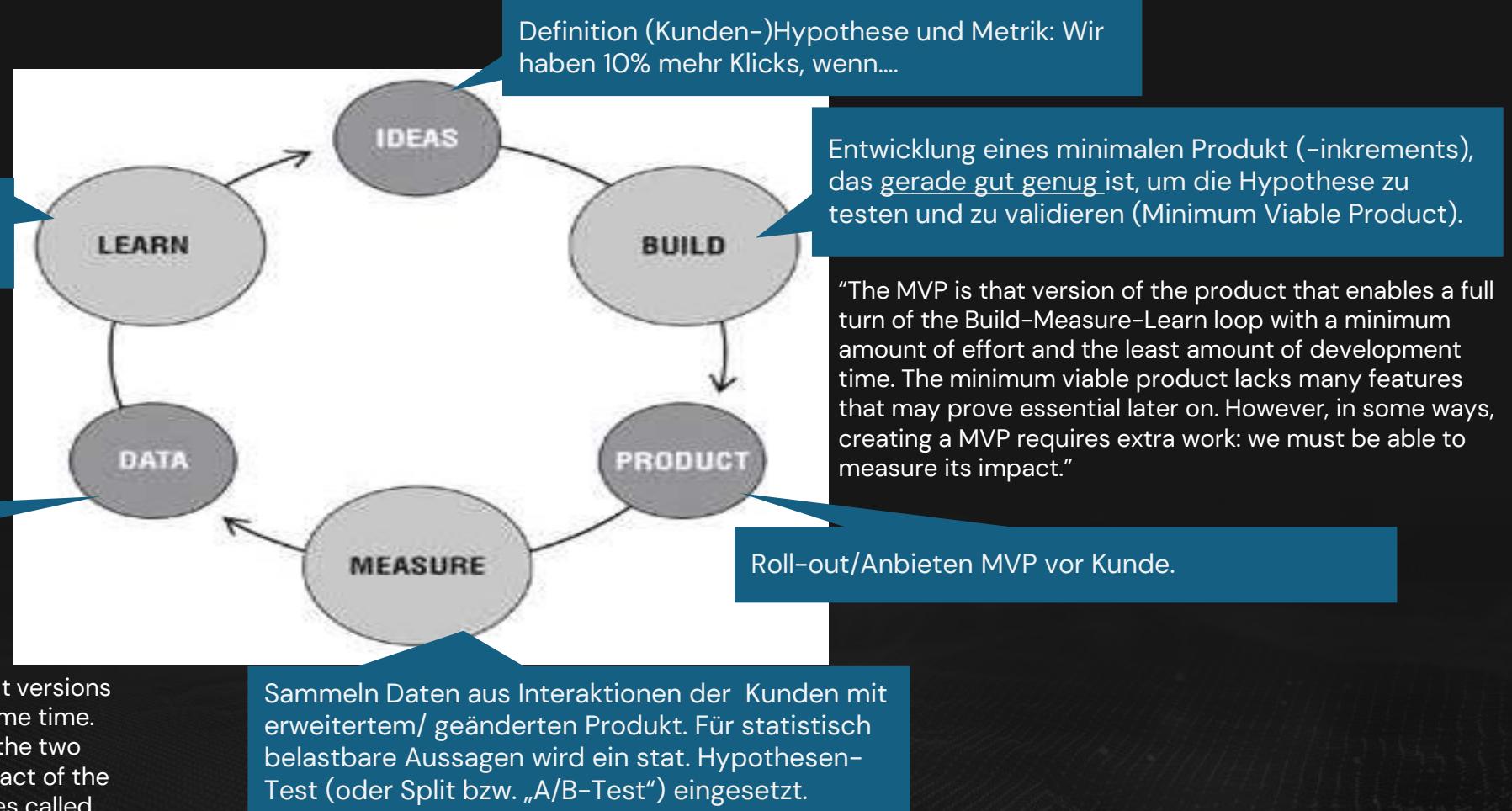
Iterativer Prozess mit Ziel des kontinuierlichen Lernens durch Daten



Build-measure-learn feedback loop

Durchführen **Datenanalysen** mit Ziel Lernen: Hypothese weiterverfolgen oder ändern (pivot)?

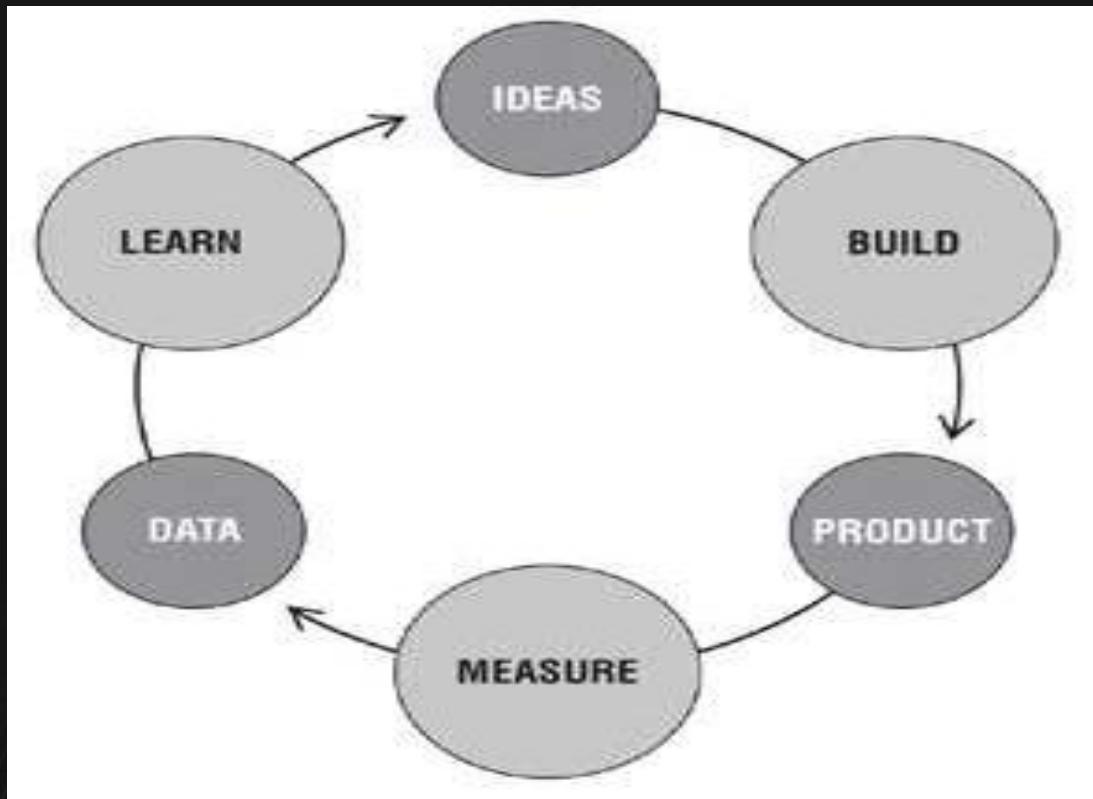
Data Engineering: Aufbereiten und Säubern Daten (Datenqualität!)



"A split-test experiment is one in which different versions of a product are offered to customers at the same time. By observing the changes in behavior between the two groups, one can make inferences about the impact of the different variations. (This technique is sometimes called A/B testing after the practice of assigning letter names to each variation.)"



Build-measure-learn feedback loop



Sie sind verantwortlicher Manager eines Online-Shops/ ...

- Wofür wären Kunden bereit (mehr) zu zahlen?
Welche Kundenhypothese haben Sie?
- Was wäre Ihr minimales Produkt (MVP), um diese Hypothese zu testen?
- Was wären (beispielhafte) Metriken für Messen dieser Hypothese?

Am Beispiel WhatsApp:

- Hypothese: Versenden beliebiger Handy-Nachrichten per Internet statt SMS/ MMS liefert Mehrwert für Kunden (für den Kunden auch zahlen¹ würden).
- MVP: eine App, die nur Text versenden kann (Roll-out erst für iPhone um Aufwand zu sparen und mehr Nutzer).
- Metriken: Anzahl Downloads für App, Anzahl versendeter Nachrichten, Anzahl zahlender Kunden, Anzahl Power-User (Kunden mit mehr als x Nachrichten),

Definieren Sie einen Durchlauf des Loop für einen Online-Shop oder eine andere digitale Firma



Data Science

Vorgehen bei einem Use Case

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Generische Vorgehensweise Data Science

Was ist die Fragestellung?

Was würde ich tun, wenn ich alle verfügbare Daten hätte? Was möchte ich abschätzen/ vorhersagen?



Data Science

Vorgehen bei einem Use Case

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Generische Vorgehensweise Data Science

Wie wurden die Daten generiert?

Welche Daten sind relevant?

Datenschutz??



Data Science

Vorgehen bei einem Use Case

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Generische Vorgehensweise Data Science

Daten darstellen (visuelles Verständnis)

Gibt es Anomalien? Unplausible Werte?

Sehen Sie Muster in den Daten?



Data Science

Vorgehen bei einem Use Case

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Generische Vorgehensweise Data Science

Modell erstellen

Modell trainieren auf Daten ("fitten")

Modell validieren



Data Science

Vorgehen bei einem Use Case

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Generische Vorgehensweise Data Science

Was habe ich gelernt?

Machen die Ergebnisse Sinn?

Ergebnisse verständlich kommunizierbar? Story??



Deep dive Daten



Datenformate

Audio

Datentypen:

- Sprache
- Musik (iTunes, MP3, OGG, WAV, ...)
- Geräusche

Datenstruktur:

- Binäre (nicht direkt lesbar) Datei,
- Größe abhängig von Sample rate (Frequenzen pro Sekunde) und Bitrate (Abtastung in Bits pro Sekunde)

Typische Anwendungsgebiete:

- Spracherkennung (Siri, GoogleNow, Cortana)
- Computersprache (Amazon Polly)
- Automatisches Übersetzen/ Untertitel



Datenformate

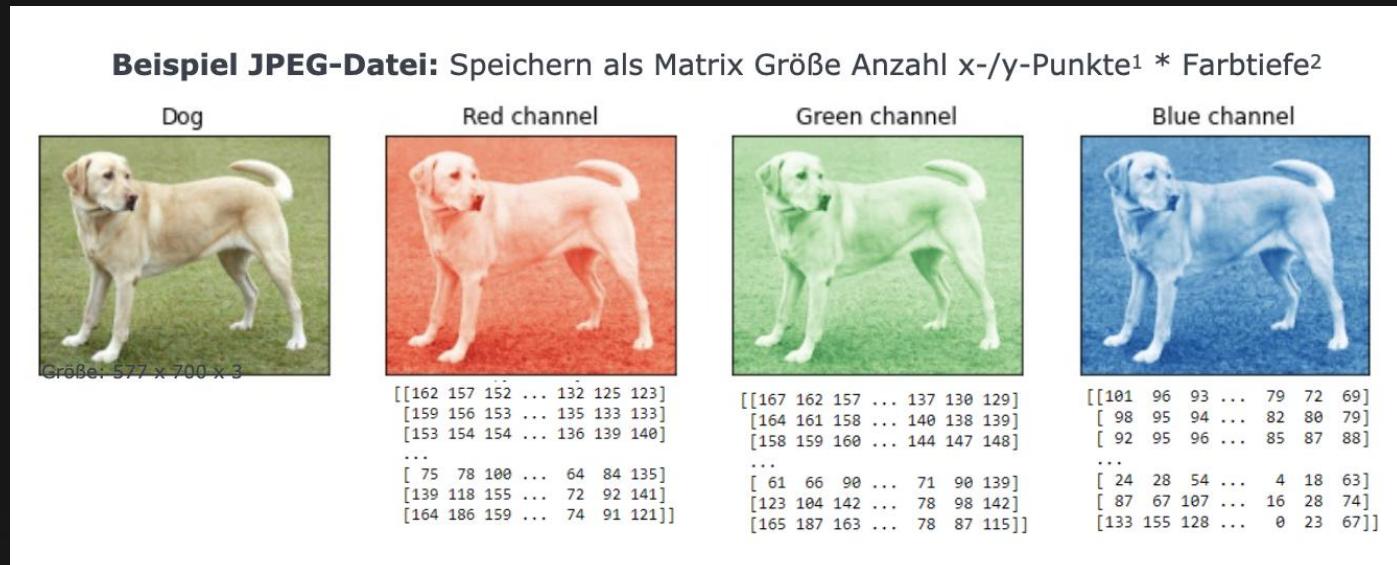
Bilder

Datentypen:

- Einzelne Bilder (RAW, JPEG)
- Video (MPEG)

Typische Anwendungsgebiete:

- Erkennen Inhalte eines Bildes
- Industrieroboter
- Autonomes Fahren



1 Anzahl Punkte gleichbedingt mit Begriff Pixel. Kamera mit 9 Mio. Pixel also 3000 x 3000 Punkte.
2 Häufig eingesetzt wird RGB mit 8 Bit/ 256 Farben je Farbkanal = 16,7 Mio. Farben
Quelle: Bild von Wikipedia-Artikel Hund ([Link](#)), Zerlegung in Farbkanäle per Python-Skript.



Datenformate

Texte

Datentypen:

- Strukturierte Texte (MS Office,
- Webseiten in XML/HTML,
- Social Media)
- Messdateien
- Unstrukturierte Texte
- ...

Beispiel Unicode UTF-8¹ (häufigste Codierung im Web²)

Unicode-Zeichen	Binäre Codierung	Was?	Beispiel
U-00000000 – U-0000007F:	0xxxxxxx	Lateinisches Alphabet mit Satzzeichen ohne Umlaute	U+0021 → 100001 → ! U+0041 → 1000001 → A
U-00000080 – U-000007FF:	110xxxxx 10xxxxxx	Erweiterung um Sprachen mit Akzenten, Umlaute,	U+00A9 → 1100001010101001 → ©
U-00000800 – U-0000FFFF:	1110xxxx 10xxxxxx 10xxxxxx	Weitere Sprachen z.B. Chinesisch oder Japanisch	U+3231 → 11100011 10001000 10110001 → 株 U+4E76 → 111001001011100110110110 → 豐

Typische Anwendungsgebiete:

- Spamfilter
- Übersetzungen
- Suchanfragen



Datenqualität

Wann sind Daten von guter Qualität?

- Fehlerfreiheit: ... wenn sie mit der Realität übereinstimmen
- Eindeutige Auslegbarkeit: ...wenn sie in gleicher, korrekter Art und Weise begriffen werden
- Einheitliche Darstellung: ...wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden
- Übersichtlichkeit: ...wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.
- Vollständigkeit:wenn sie nicht fehlen und zu festgelegten Zeitpunkten zur Verfügung stehen
- Verständlichkeit: ...wenn sie unmittelbar von den Anwendern verstanden werden können
- Relevanz: ...wenn sie für den Anwender notwendige Informationen liefern
- Glaubwürdigkeit: wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und –verbreitung vertrauensvoll ist
- Aktualität: wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden
- Wertschöpfung: wenn ihre Nutzung zu einer quantifizierbaren monetären Steigerung führen kann



Datenqualität

Fallbeispiel

Kunden-Nr.	Name	Geburts-datum	Alter	Geschlecht	Email	PLZ	Stadt	Letzter Kontakt	T645fet	Umsatz 2015
20456	Tina Huber	10.01.2010	21	W		8000	München	01.08.2021	Ja	100€
20456	Teddy Test	6.8.1490	20	M	test@test.de	80797	Freising	05.03.2008	Nein	
23578	B. Trüger	08.07.1979	41	D	trueger@gmx.de	D-80793	Muenchen	01.07.2020	bald	10000
28903	Amy Doe	03/12/2003		F	amyd@yahoo.com		Düsseldoof	15.07.2020	ja	4000\$

Was sind mögliche Probleme?



Datenqualität

Beispielhafte Regeln in einem Online-Store

	Kunden-ID	Name	Geboren	Alter	Adresse	Kreditkartennummer	Einkäufe 2020	Umsätze 2020
Regel für Sicherstellen Datenqualität	ID definiert und eindeutig (d.h. darf max. 1 mal vorkommen)	Liegt vor	Geburtsdatum in europäischem Format: TT.MM.YY., sonst umwandeln	Alter < 120	muß vorliegen	1. 12 ≤ Anzahl Ziffern ≤ 16 2. Korrekte Prüfsumme (bspw. Luhn-Algorithmus ¹)		Währung in EUR, sonst umwandeln
Relevant für Wertschöpfung per Service/Empfehlung	-	-	Altersgruppen	Ja, für Empfehlungen Aber bspw. auch für Ansprache Kunde	Ja, bspw. Wohnort		Ja, für Empfehlungen	Ja, für Empfehlungen

Kriterien für Datenqualität und Relevanz (Wertschöpfung)



Datenbasierte Geschäftsmodelle

Übersicht Geschäftsmodelle

Data-informed¹ Geschäftsmodelle: Optimierung bestehender Wertschöpfungsprozesse durch Daten.

- Prozessoptimierung durch Automatisierung (gesamte Industrie).
- Reduktion Entwicklungszeit/ -kosten durch Simulation (Luft- & Raumfahrttechnik, Automobilbereich).
- Online-Vertrieb für physische Produkte (Otto, Lieferando, Zalando, Amazon).
- Mobility Dienste (Uber, Lyft).

Data-infused¹ Geschäftsmodelle: Wertschöpfungsprozesse hängen wesentlich von Daten ab.

- Personalisierte Werbung (Facebook und Google).
- Personalisierte Produktempfehlungen (Amazon).
- Quantitative Analysis/ Algorithmic Trading.

Data driven¹ Geschäftsmodelle: Wertschöpfung vollständig digital.

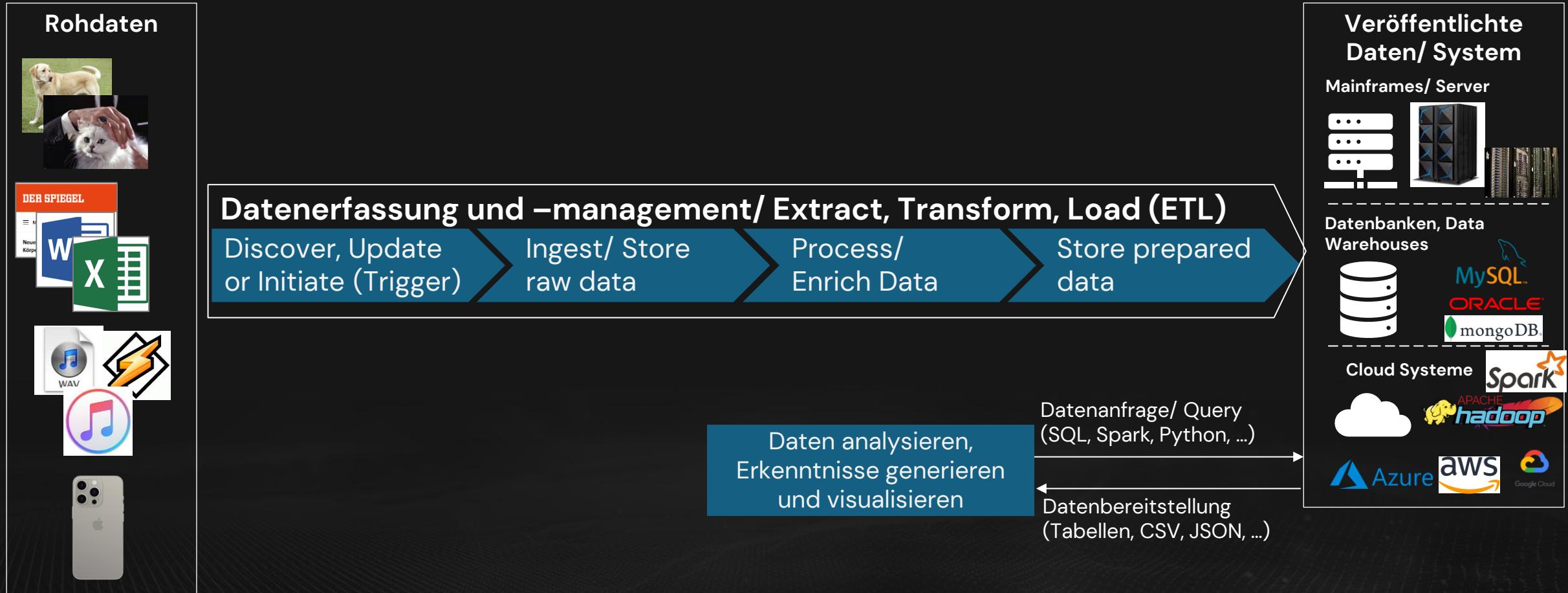
- Online-Vertrieb digitaler Produkte (Netflix, Spotify, Steam,).
- Software-Geschäftsmodelle (Werbebasiert, Freeware, Freemium, Shareware, Mieten, Kauf).

 Datenbasierte Geschäftsmodelle ermöglichen per Skalierung mehr Effizienz/ Profit bei gleichbleibender Kostenstruktur
„Data is the world's most valuable resource“²



Daten

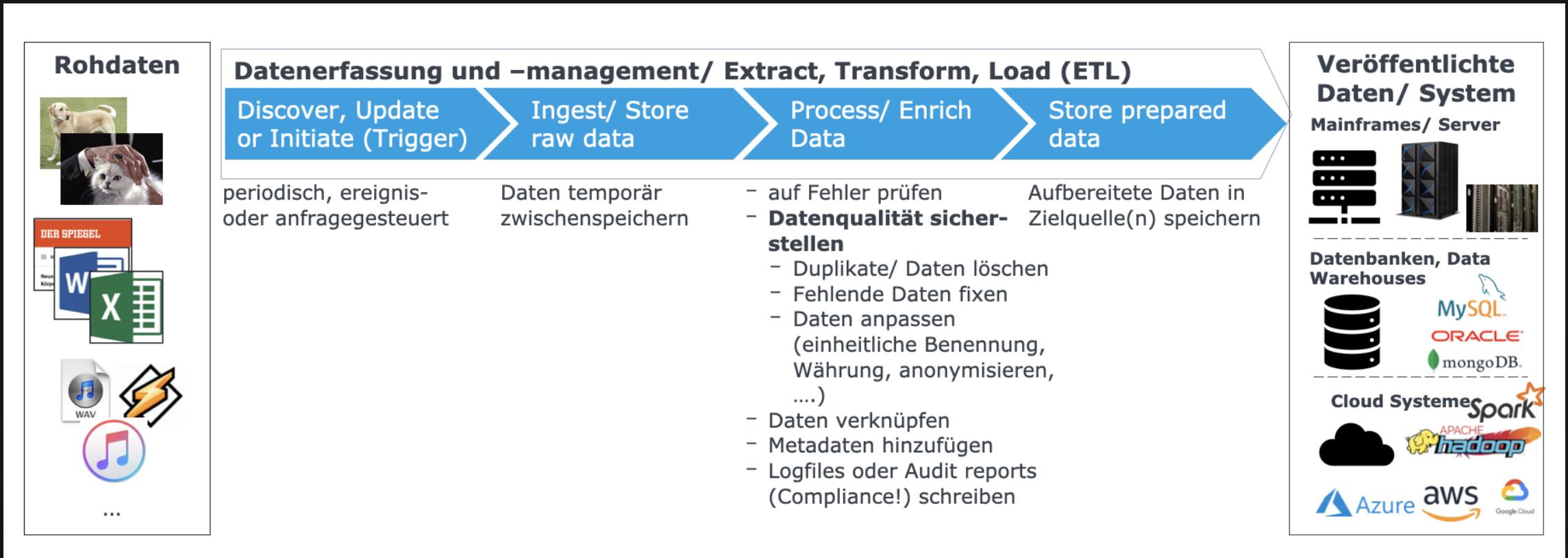
Workflow Datenmanagement at scale





Daten

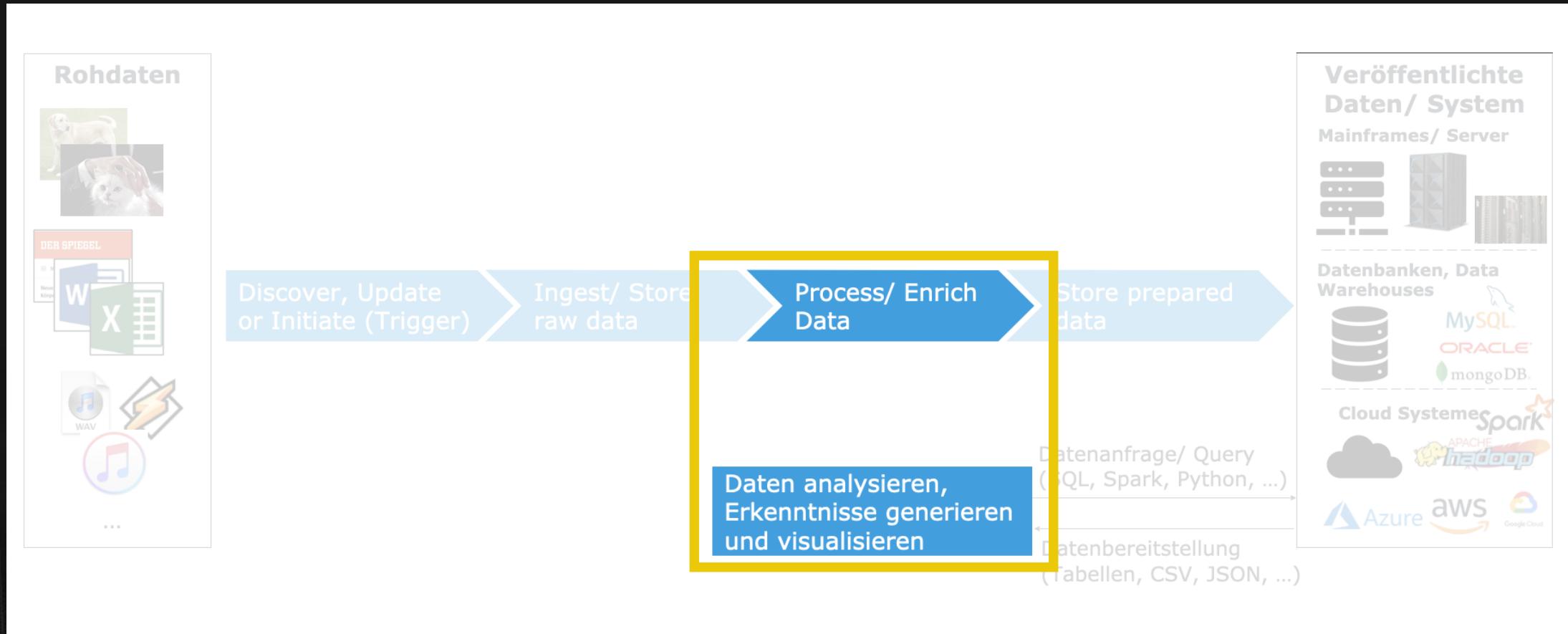
Workflow Datenmanagement at scale





Daten

Fokus des Kurses



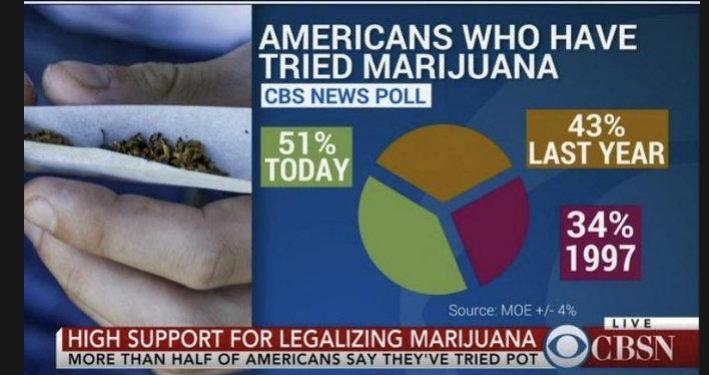
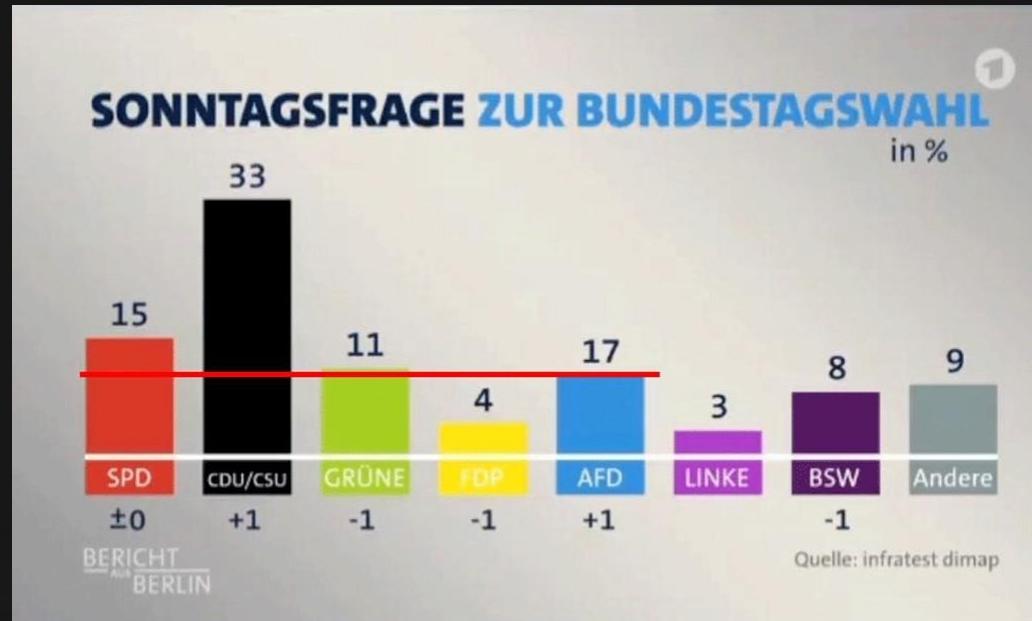
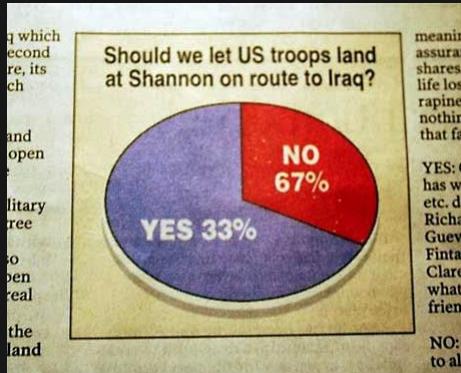


Deep dive Visualisierung



Data Science

Fallbeispiele schlechte Visualisierungen.

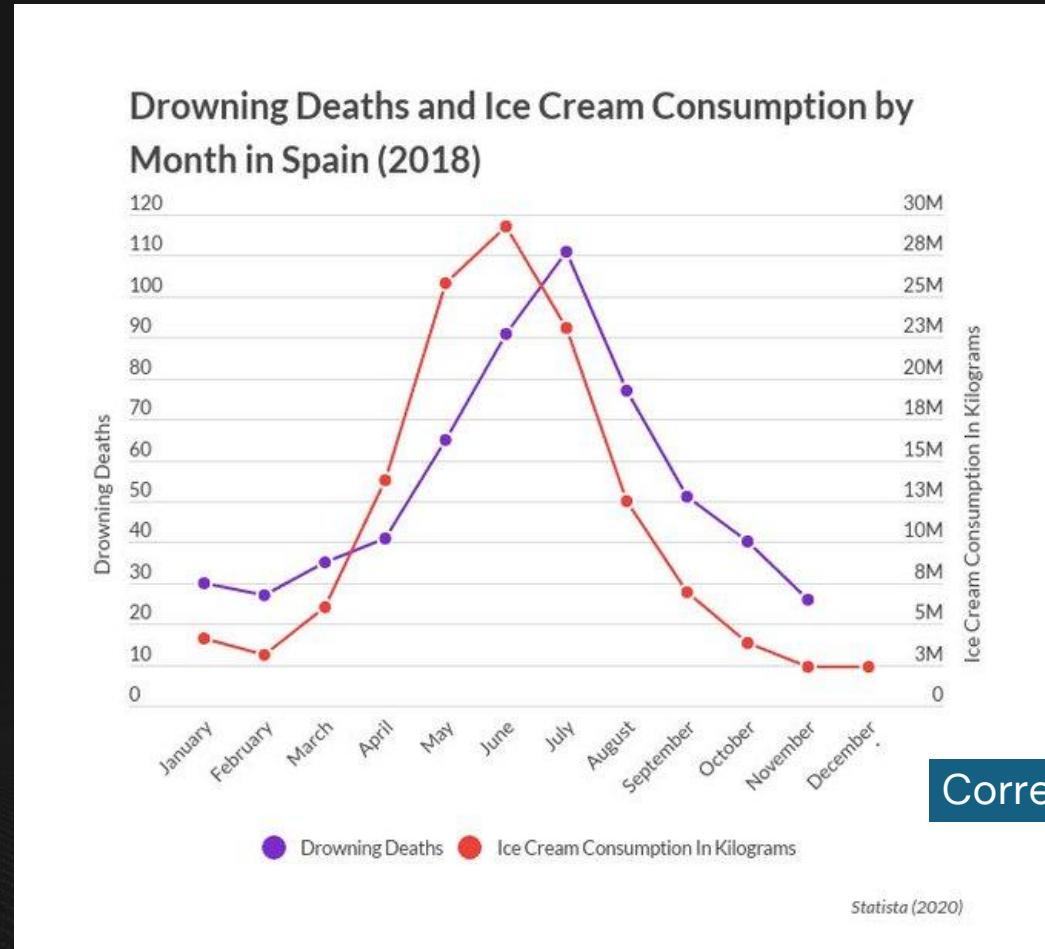


Quellen: <https://en.rattibha.com/thread/1530501932439244801>,
<https://www.welt.de/kultur/medien/article253433298/Bericht-aus-Berlin-Zu-hohe-Saeulendiagramme-fuer-SPD-und-Gruene-ARD-entschuldigt-sich.html>,
https://www.reddit.com/r/CrappyDesign/comments/bemumv/this_chart_on_marijuana_usage/



Data Science

Fallbeispiele schlechte Visualisierungen

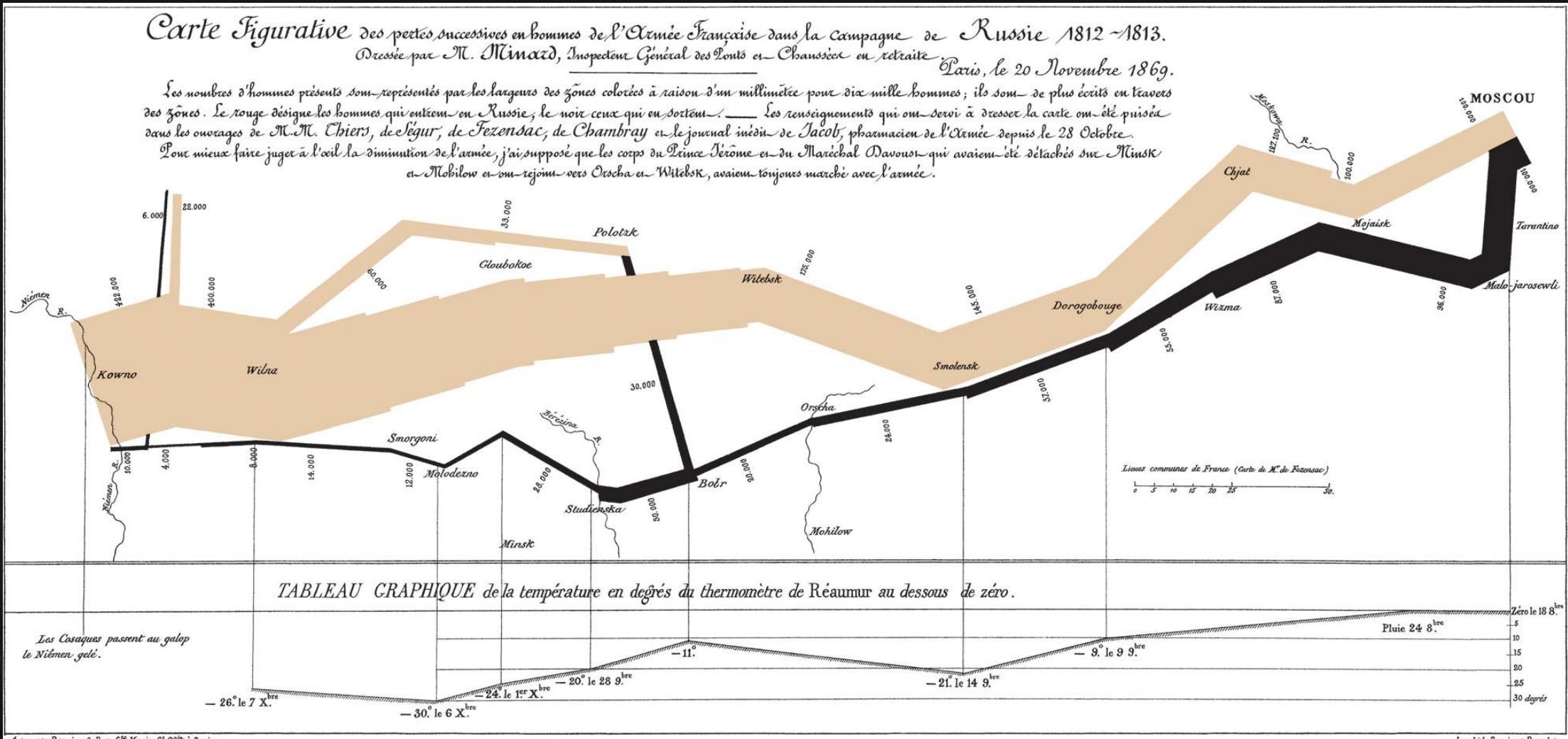


Correlation does not imply causation!!



Data Science

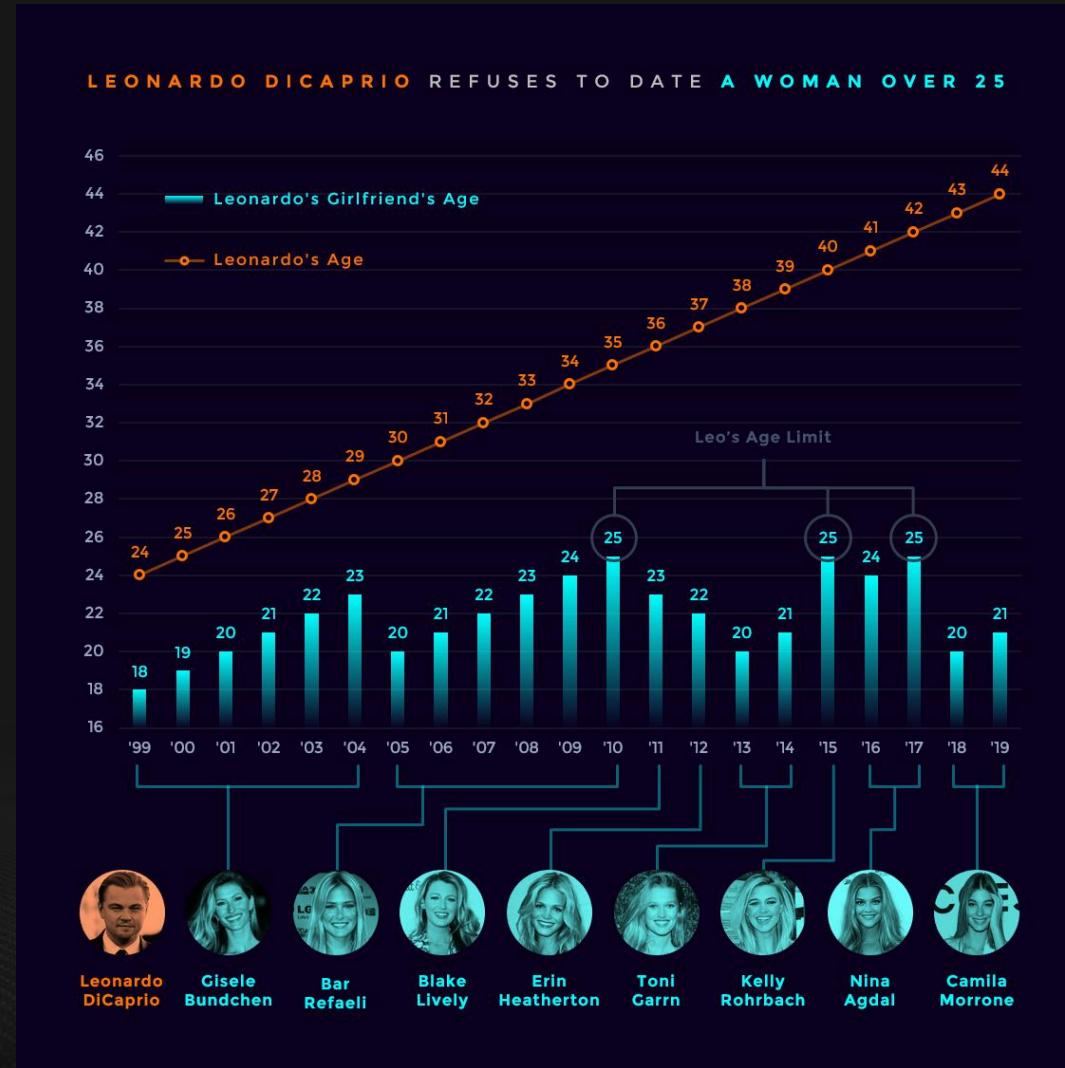
Fallbeispiele gute Visualisierungen.





Data Science

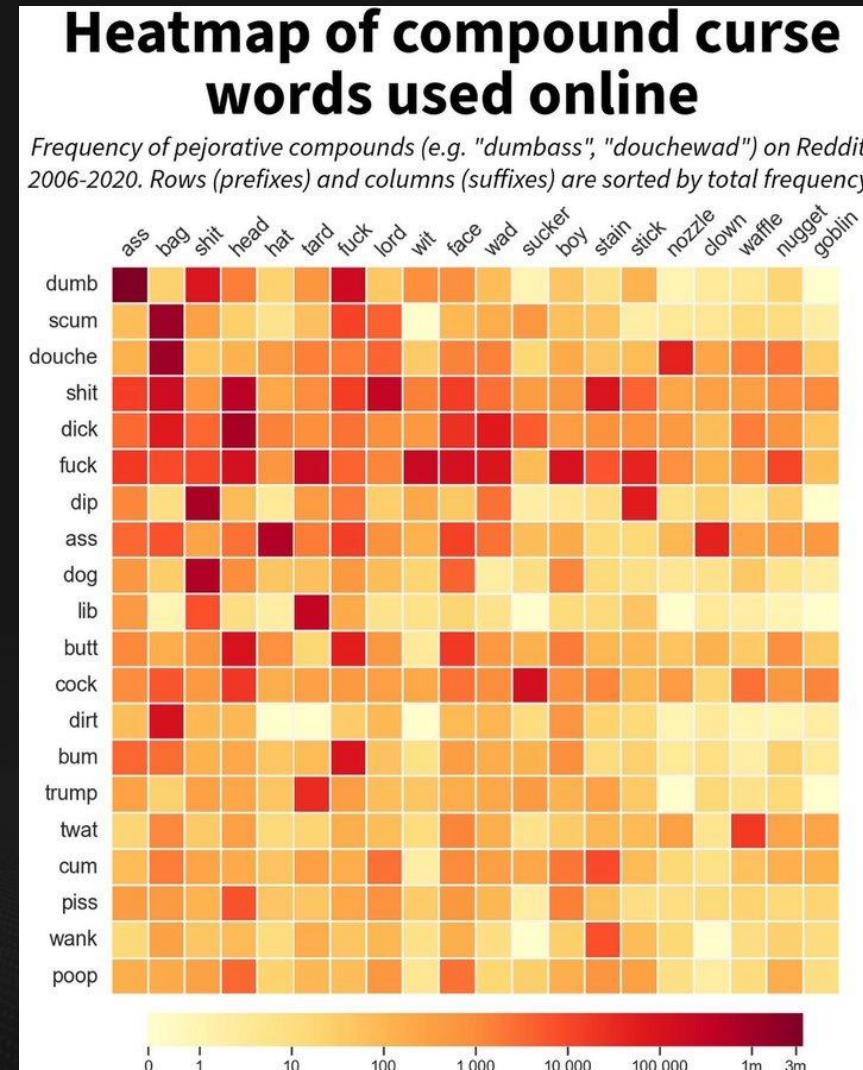
Fallbeispiele gute Visualisierungen.





Data Science

Fallbeispiele gute Visualisierungen





ToDo bis zur nächsten Vorlesung

- Google Account eröffnen
- Google Chrome als Browser einrichten
- <https://colab.research.google.com/> (schon mal reinschauen)



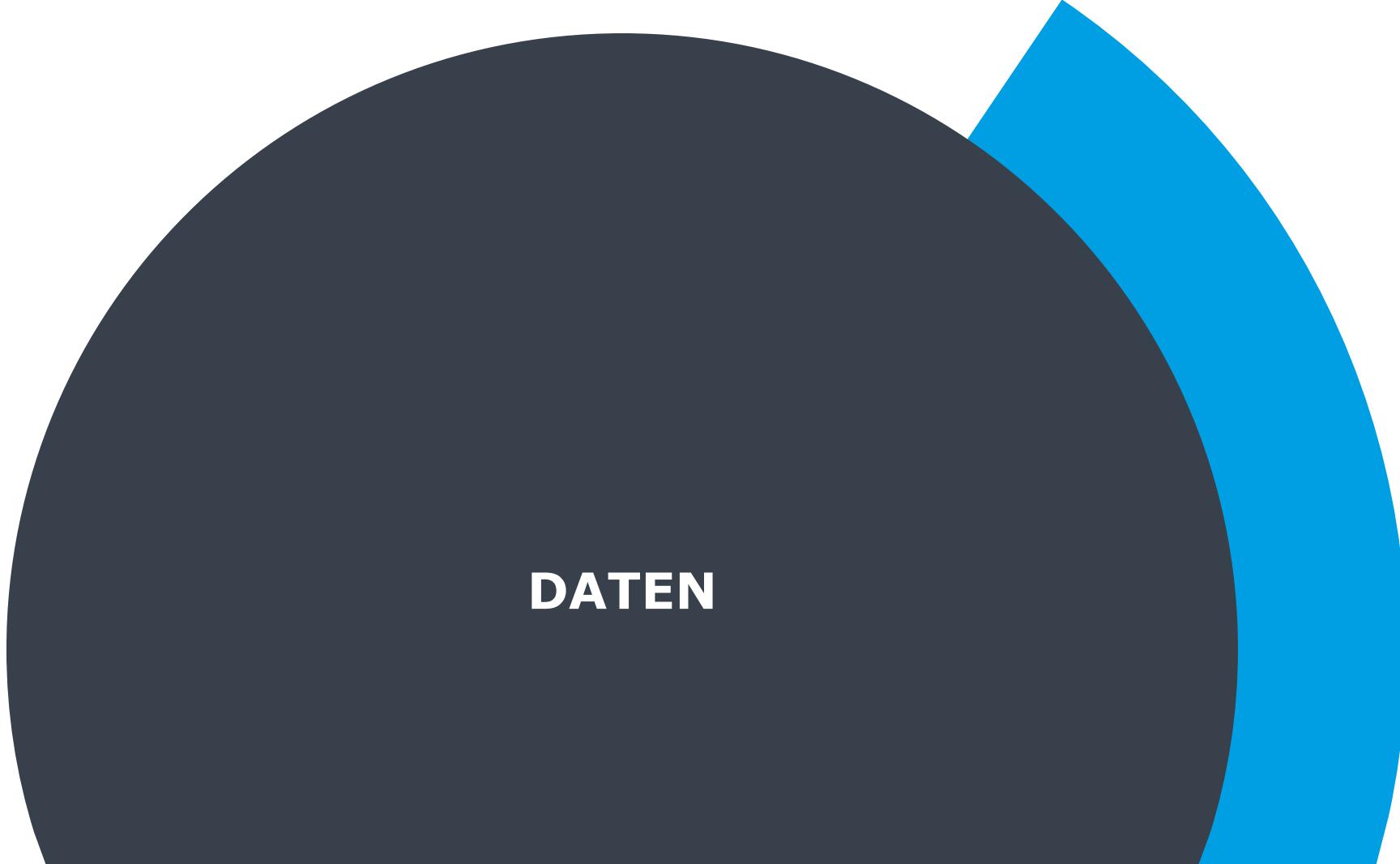
Literatur

Statistik:

- James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning (freies Ebuch unter: [Link](#))
- Spiegelhalter: The Art of Statistics: Learning from data
- Witte: Statistics (10th Edition)
- Silver: The Signal and the noise
- Taleb: Black Swan
- Huff: How to Lie with Stastistics
- Wheelan: Naked statistics

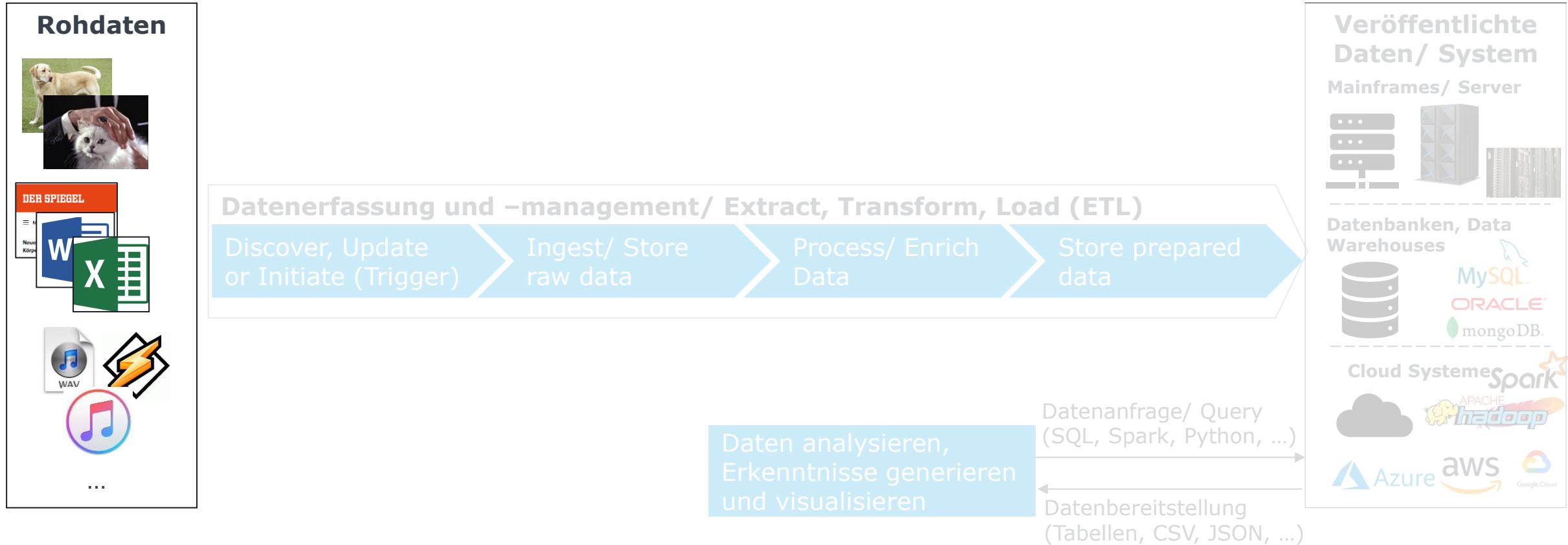


Backup



DATEN

DATENFORMATE



AUDIODATEIEN.

Datentypen:

- Sprache
- Musik (iTunes, MP3, OGG, WAV, ...)
- Geräusche

Datenstruktur:

- Binäre (nicht direkt lesbar) Datei,
- Größe abhängig von Sample rate (Frequenzen pro Sekunde) und bitrate (Abtastung in Bits pro Sekunde)

Typische Anwendungsgebiete:

- Spracherkennung (Siri, GoogleNow, Cortana)
- Computersprache (Amazon Polly)
- Automatisches Übersetzen/ Untertitel

Nicht im Fokus Vorlesung

BILDER/ VIDEO

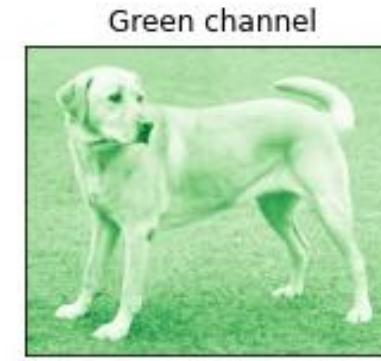
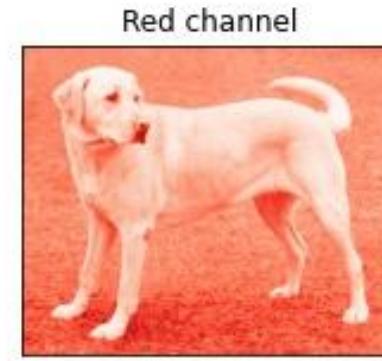
Datentypen:

- Einzelne Bilder (RAW, JPEG)
- Video (MPEG)

Typische Anwendungsgebiete:

- Erkennen Inhalte eines Bildes
- Industrieroboter
- Autonomes Fahren

Beispiel JPEG-Datei: Speichern als Matrix Größe Anzahl x-/y-Punkte¹ * Farbtiefe²



```
[[162 157 152 ... 132 125 123]
 [159 156 153 ... 135 133 133]
 [153 154 154 ... 136 139 140]
 ...
 [ 75  78 100 ...  64  84 135]
 [139 118 155 ...  72  92 141]
 [164 186 159 ...  74  91 121]]
```

```
[[167 162 157 ... 137 130 129]
 [164 161 158 ... 140 138 139]
 [158 159 160 ... 144 147 148]
 ...
 [ 61  66  90 ...  71  90 139]
 [123 104 142 ...  78  98 142]
 [165 187 163 ...  78  87 115]]
```

```
[[101  96  93 ...  79  72  69]
 [ 98  95  94 ...  82  80  79]
 [ 92  95  96 ...  85  87  88]
 ...
 [ 24  28  54 ...   4  18  63]
 [ 87  67 107 ...  16  28  74]
 [133 155 128 ...   0  23  67]]
```

TEXTE

Datentypen:

- Strukturierte Texte (MS Office, Webseiten in XML/HTML, Social Media)
- Messdateien
- Unstrukturierte Texte
- ...

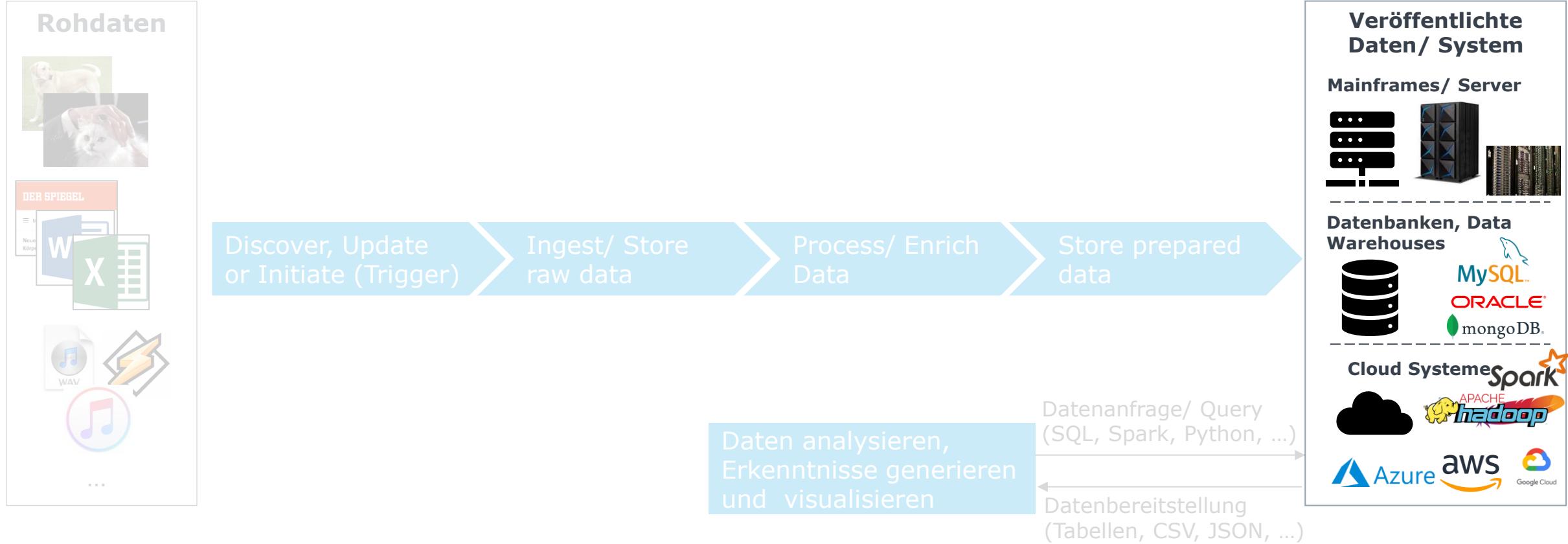
Beispiel Unicode UTF-8¹ (häufigste Codierung im Web²)

Typische Anwendungsgebiete:

- Spamfilter
- Übersetzungen
- Suchanfragen

Unicode-Zeichen	Binäre Codierung	Was?	Beispiel
U-00000000 – U-0000007F:	0xxxxxx	Lateinisches Alphabet mit Satzzeichen ohne Umlaute	U+0021 → 100001 → ! U+0041 → 100001 → A
U-00000080 – U-000007FF:	110xxxxx 10xxxxxx	Erweiterung um Sprachen mit Akzenten, Umlaute,	U+00A9 → 1100001010101001 → ©
U-00000800 – U-0000FFFF:	1110xxxx 10xxxxxx 10xxxxxx	Weitere Sprachen z.b. Chinesisch oder Japanisch	U+3231 → 11100011 10001000 10110001 → 株 U+4E76 → 111001001011100110110110 → 喬

VERÖFFENTLICHTE DATEN/ SYSTEM



SPEICHERN DER DATEN.

- [PC oder lokale Speichermedien]
- Mainframes
- Server im privaten oder Firmennetzwerk
- **relationale und nicht-relationale Datenbanken**
- **Data Warehouses** (Amazon RedShift, Snowflake, Google BigQuery, SAP, ...)
- Cloud Systeme:
 - Buckets (Amazon S3, Azure Storage, ...): enthält Objekte bis zu 5 TB Größe, Zugriff Web-Interface
 - Distributed Datasets (Spark, Hadoop, ...): Daten (meist Tabellen) verteilt auf mehrere Systeme

**Hauptsächliches Unterscheidungskriterium:
on-prem (vor Ort bei Person/Firma) oder Cloud.**



Cloud Systeme ermöglichen (beliebige) Anpassung bereitgestellte Ressourcen an Nachfrage (elastic/ Skalierbarkeit), stellen aber höhere Anforderungen an Datensicherheit (DSGVO) und Datenhaltung (bspw. China)

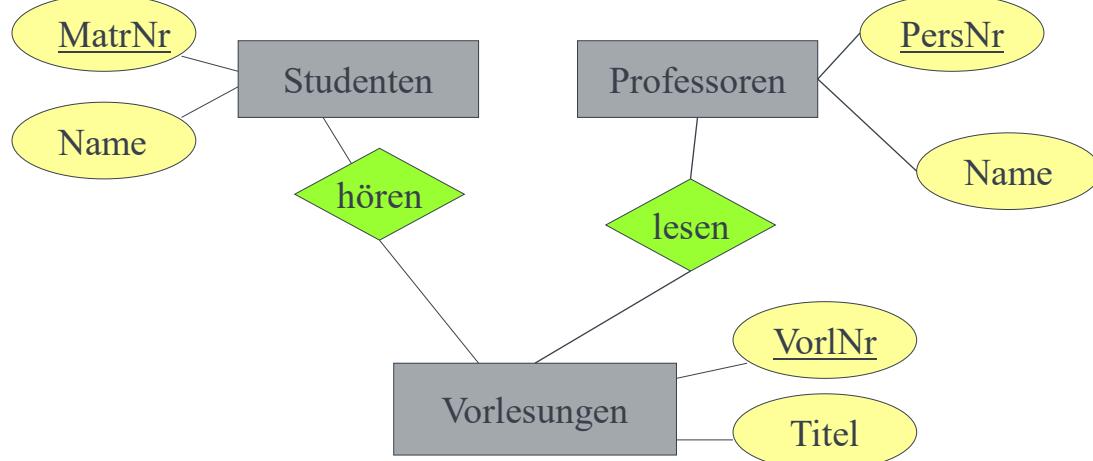
DATENBANKEN: RELATIONALE DATENBANKEN.



MySQL™

ORACLE

Konzeptuelle Modellierung in Form von Beziehungen (Relationen)



Darstellung im sogenannten „Entity-Relationship“-Diagramm¹

Speicherung der Daten in Form von Tabellen

Studenten		hören		Vorlesungen	
MatrNr	Name	MatrNr	VorlNr	VorlNr	Titel
26120	Fichte	25403	5022	5001	Grundzüge
25403	Jonas	26120	5001	5022	Glaube & Wissen
...	

Bearbeiten der Daten mit SQL (Structured Query Language)

```

Select Name
From Studenten, hören, Vorlesungen
Where Studenten.MatrNr = hören.MatrNr
and hören.VorlNr = Vorlesungen.VorlNr
and Vorlesungen.Titel = `Grundzüge`;
    
```

```

Update Vorlesungen
Set Titel = `Logik`
Where VorlNr = 5001;
    
```

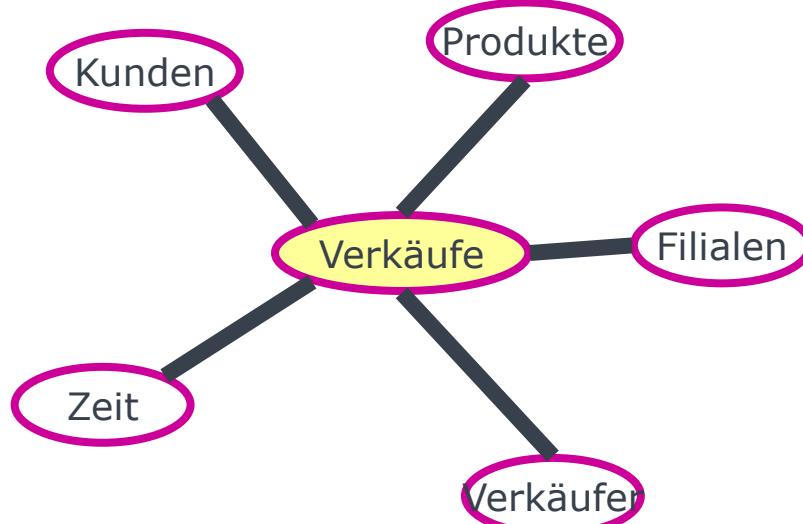


Vorteile: weit verbreitet, robust, garantiert konsistente Daten und benötigt wenig Speicherplatz.
Nachteile: hoher Aufwand beim Speichern, Ändern und Abfragen von Daten → schlecht skalierbar.

DATENBANKEN: DATA WAREHOUSES.



Logische Struktur Data Warehouse (Star schema)



Speicherung der Daten im Data Warehouse

Produkte					
ProduktNr	Produkttyp	Produktgruppe	Produkthauptgruppe	Hersteller	...
1347	Handy	Mobiltelekom	Telekom	Siemens	...
...

Kunden			
KundenNr	Name	wiealt	...
4711	Kemper	43	...
...

Filialen			
Filialenkennung	Land	Bezirk	...
Passau	D	Bayern	...
...

Verkäufe					
VerkDatum	Filiale	Produkt	Anzahl	Kunde	Verkäufer
25-Jul-00	Passau	1347	1	4711	825
...

Zeit								
Datum	Tag	Monat	Jahr	Quartal	KW	Wochentag	Saison	...
...
25-Jul-00	25	Juli	2000	3	30	Dienstag	Hochsommer	...
...
18-Dec-01	18	Dezember	2001	4	52	Dienstag	Weihnachten	...
...

Sehr große **Faktentabelle** enthält Daten des Geschäftsprozesses, verlinkt auf einzelne, kleine **Dimensionstabellen** mit den Daten.



Vorteile: schnelles Auswerten verschiedenster, zusammenhängender Daten.

Nachteile: Aufbereiten und Aktualisieren von Daten aufwendig. Hohe Wartungskosten. Struktur nur schwer änderbar.

DATENBANKEN: NICHT-RELATIONALE DATENBANKEN.

Key-Value



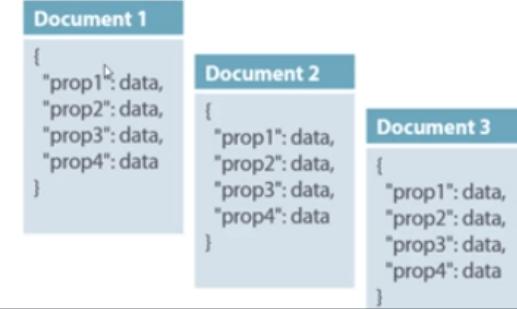
Key	Value
Name	Joe Bloggs
Age	42
Occupation	Stunt Double
Height	175cm
Weight	77kg

Wide-Column Store



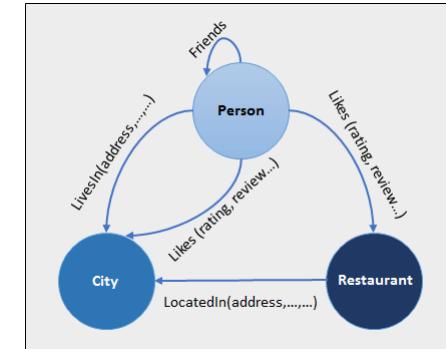
ColumnFamily			
Row Key	Column Name		
	Key	Key	Key
Value			
Column Name	Key	Key	Key
	Value	Value	Value
Value			

Document-oriented

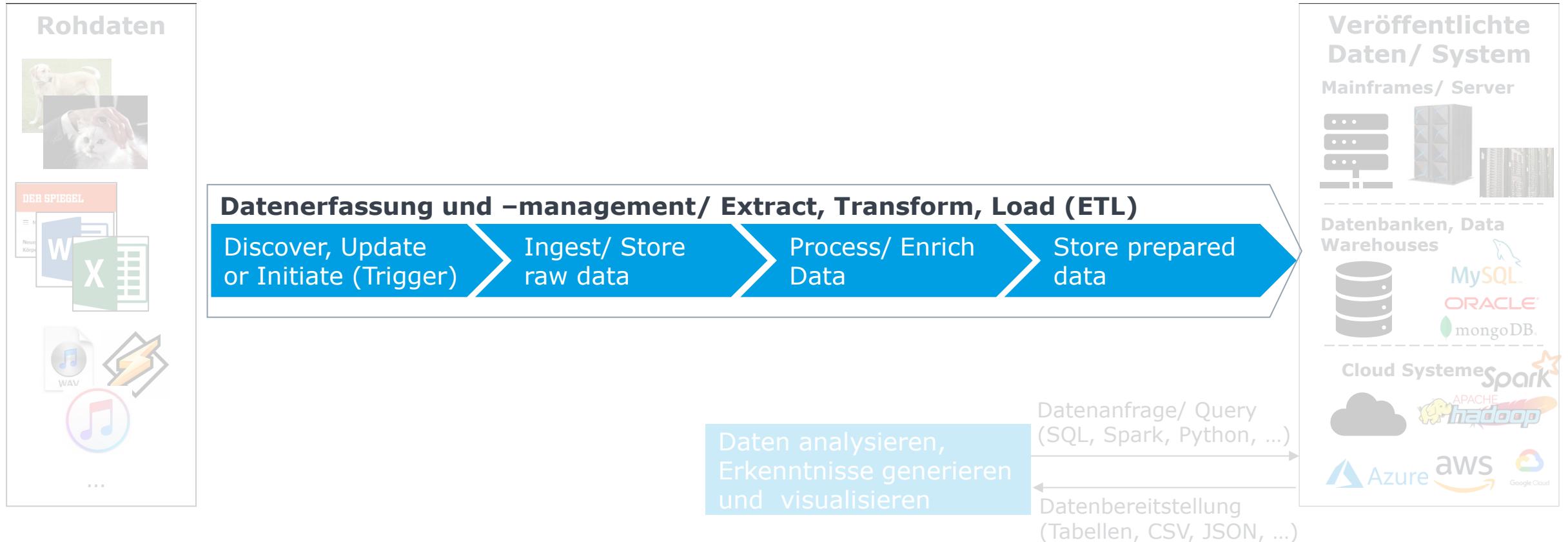


Vorteile: sehr schneller Lese- und Schreibzugriff auch bei sehr großen Datenmengen. Ausfallsicher durch Replikationen.
Nachteile: Daten können inkonsistent oder veraltet sein. Fixes Datenschema, keine Verknüpfungen Daten möglich!

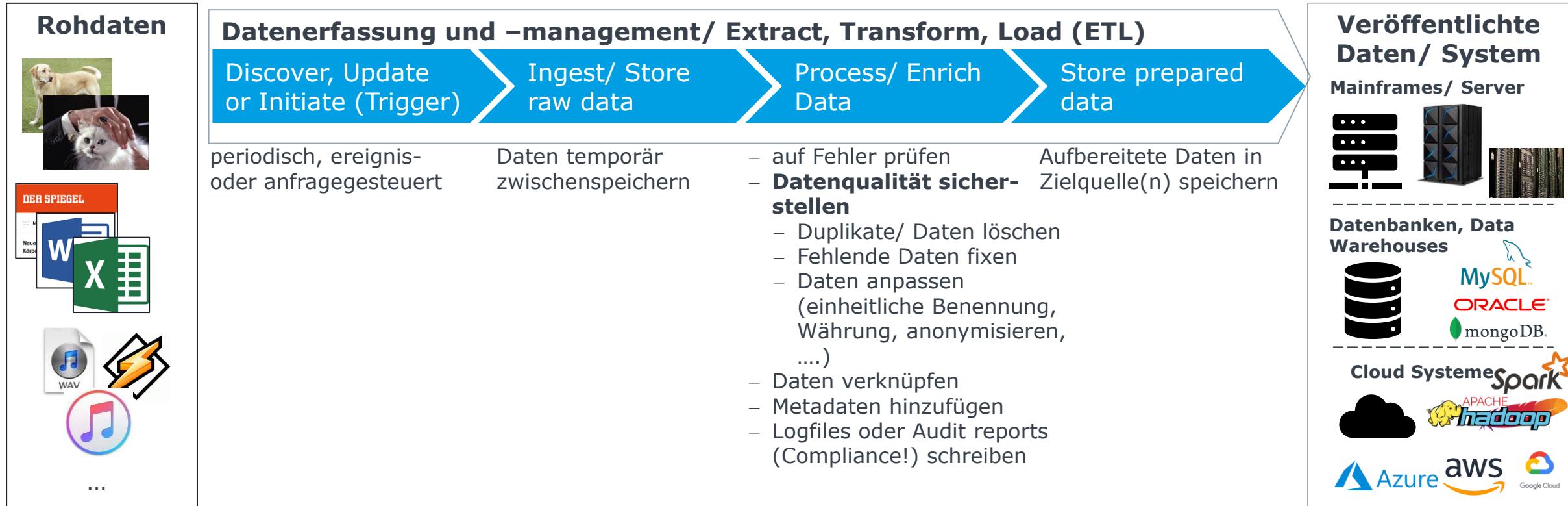
Graph-Based



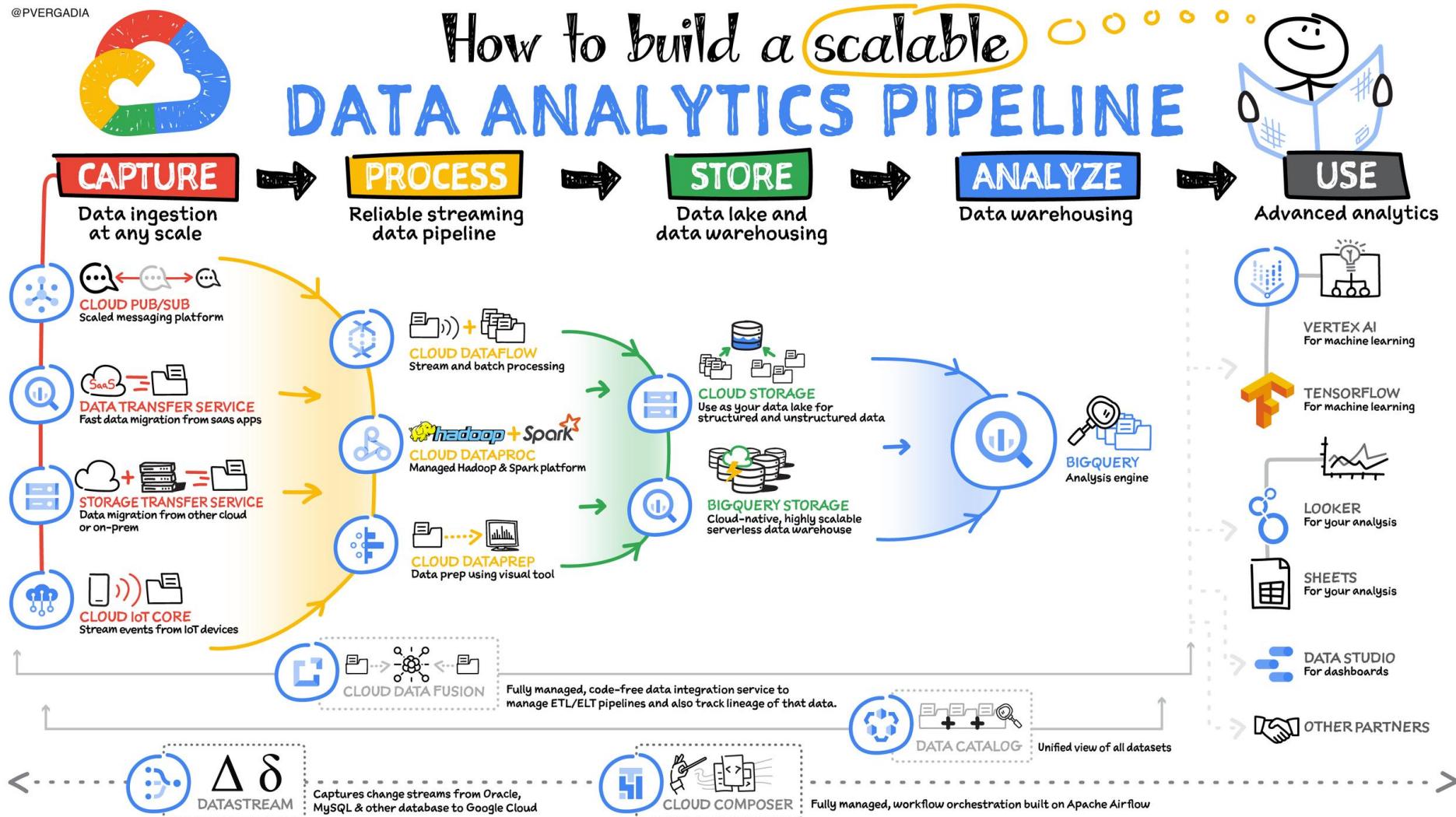
WORKFLOW DATENERFASSUNG- UND MANAGEMENT/ ETL (EXTRACT, TRANSFORM, LOAD).



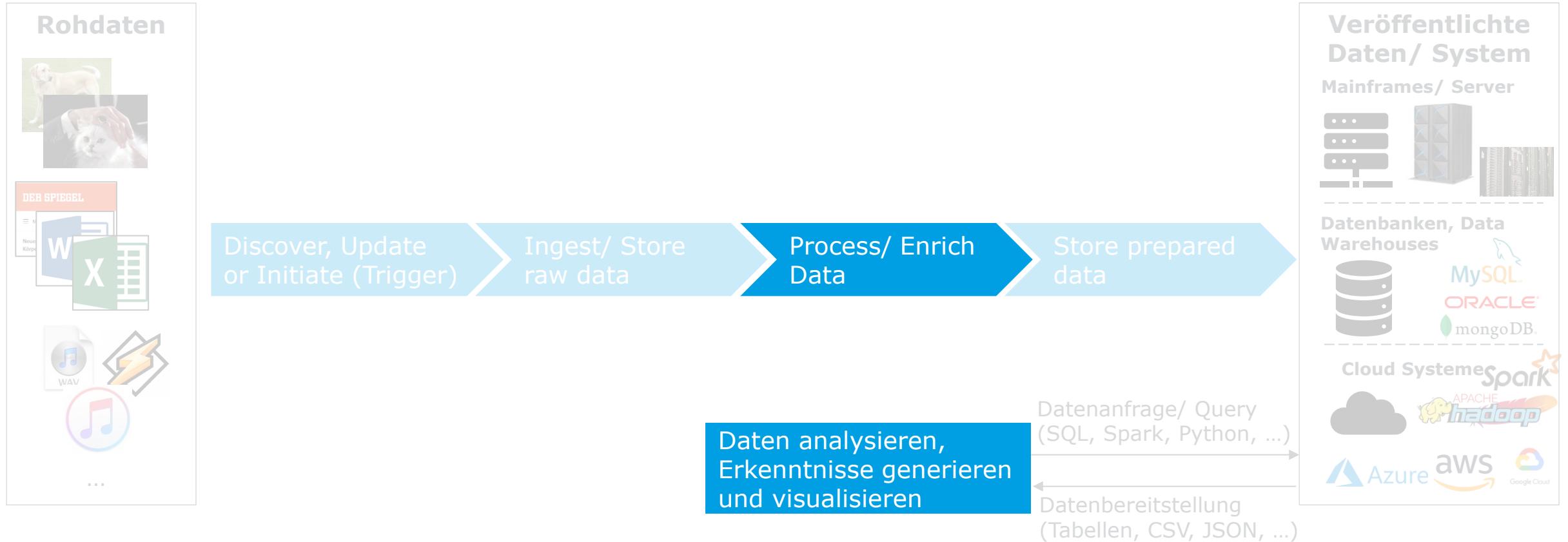
ETL IM DETAIL.



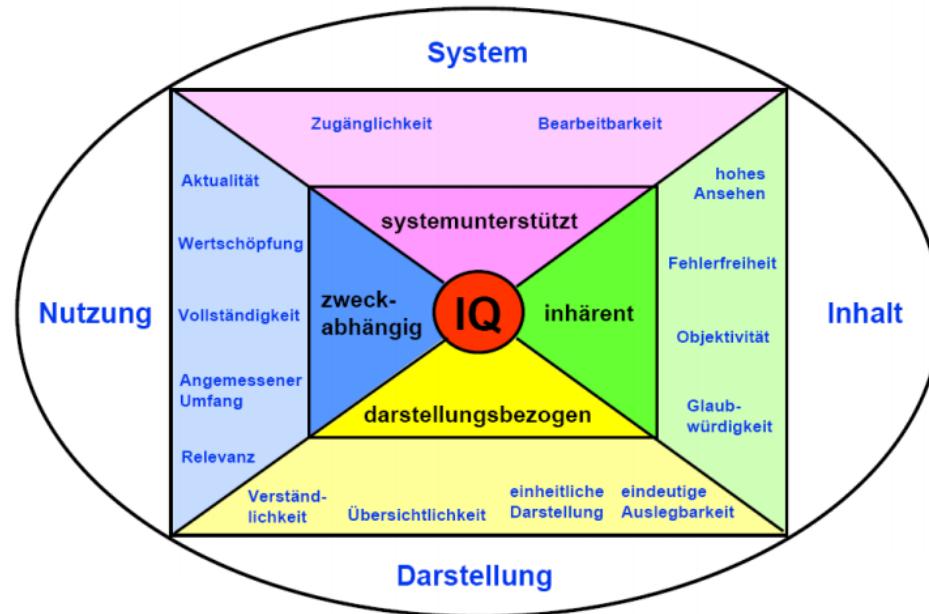
Daten aus mehreren Datenquellen extrahieren, an (Geschäfts-)Bedürfnisse anpassen und in neuer Quelle ablegen



DATENQUALITÄT



6. ÜBERSICHT DATENQUALITÄT



Detaillierung Kriterien
im Backup



Es gibt viele verschiedene Kriterien für Datenqualität, die o.a. Kriterien sind bekannte Beispiele.
Es werden auch nicht immer alle verwendet.

BEISPIELHAFTE KRITERIEN FÜR DATENQUALITÄT.

Fehlerfreiheit: ... wenn sie mit der Realität übereinstimmen

Eindeutig. Auslegbarkeit: ...wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden

Einheitliche Darstellung: ...wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden

Übersichtlichkeit: ...wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.

Vollständigkeit:wenn sie nicht fehlen & zu festgelegten Zeitpunkten in den jeweiligen Prozessschritten zur Verfügung stehen

Verständlichkeit: ...wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können

Relevanz: ...wenn sie für den Anwender notwendige Informationen liefern.

Glaubwürdigkeit: wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und –verbreitung mit hohem Aufwand betrieben werden.

Aktualität: wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden.

Wertschöpfung: wenn ihre Nutzung zu quantifizierbaren Steigerung einer monetären Zielfunktion führen kann.



Datenaufbereitung und –bearbeitung beträgt ca. 70-80% der Zeit eines Use Case Data Science oder AI!

PERSONALISIERTE KAUFEMPFEHLUNGEN ONLINE-SHOP - PRÄMISSEN.

Ziel: Generieren Einnahmen für einen Online-Shop durch personalisierte Kaufempfehlungen (Was kauften ähnliche Kunden?).

Dazu benötigen wir (Auszug...):

- Für jeden Kunden eine Liste seiner Einkäufe, aus der wir per Abgleich mit ähnlichen Kunden Empfehlungen generieren.
- (viele) soziographische Daten je Kunde. Durch aggregieren dieser Kundendaten, lernen wir ein Modell für Bestimmen:
 - Wie solvent ein individueller Kunde ist (bspw. anhand Wohnviertel, Umsatz in den letzten Jahren,)
 - Ähnlicher Kunden zu einem individuellen Kunden („Was für Kunde A relevant ist, ist es vielleicht auch für Kunde B...“)
- Unser Geschäftsmodell funktioniert nur mit qualitativ guten Daten, da sonst die Kaufempfehlungen nicht überzeugen.
- Da wir viele Kunden haben, brauchen wir automatisiert auswertbare Regeln für das Prüfen der Daten (übernächste Folie).



Wie solche Regeln sowie Empfehlungsmodell programmiert wird, schauen wir uns in den weiteren Vorlesungen noch an..

PERSONALISIERTE KAUFEMPFEHLUNGEN ONLINE-SHOP – ANWENDEN DER AUSGEWÄHLTE KRITIERIEN FÜR DATENQUALITÄT.

- Fehlerfreiheit:** für jeden Eintrag/ Zeile ergeben die definierten Prüfkriterien keinen Fehler.
- Einheitl. Darstellung:** Geldsummen immer in Euro, Telefonnummern immer mit internationaler Vorwahl, ...
- Übersichtlichkeit:** genau die für Betreuung relev. Eigenschaften in leicht fassbarem Format (z.B.: Adresse liegt vor, nicht zu viele Infos)
- Verständlichkeit:** die Attribute und Werte des Kunden sind für jeweilige Bearbeiter der Firma verständlich (Support, Werbeabteilung, ...)
- Vollständigkeit:** für jeden Kunden sind alle Attribute befüllt.
- Relevanz:** die für die Anwendungsfälle (bspw. Betreuung, Kaufempfehlung, ...) notwendigen Eigenschaften des Kunden sind vorhanden. Das ist das Zweckbindungsprinzip aus der Datenschutzgrundverordnung rein (Art. 5-1b¹).
- Angemessener Umfang:** nur die für die Anwendungsfälle notwendigen Daten werden erfaßt (Minimalprinzip aus der DSGVO, Art. 5-1c¹)
- Glaubwürdigkeit:** die Daten sind vertrauenswürdig. Dieses Kriterium ist oft schwammig. In der Praxis geht man oft davon aus, daß falls die Postadresse existiert, Kreditkarte gültig ist (bspw. per Minibuchung 0,01€), die Daten des Kunden glaubwürdig sind.
- Aktualität:** Kundendaten sind auf dem letzten Stand (bspw. seiner letzten Transaktionen/ Interaktionen mit der Firma)
- Wertschöpfung:** siehe vorige Seite

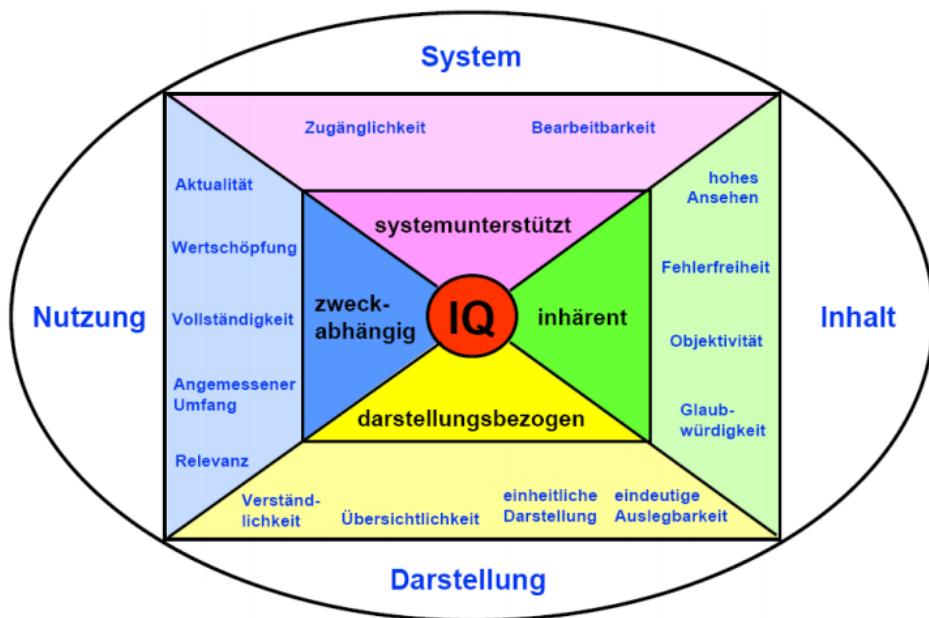
PERSONALISIERTE KAUF-EMPFEHLUNGEN ONLINE-SHOP – DATENARCHITEKTUR UND REGELN ZUR SICHERSTELLUNG DATENQUALITÄT.

	Kunden-ID	Name	Geboren	Alter	Adresse	Kreditkartennummer	Einkäufe 2020	Umsätze 2020
Regel für Sicherstellen Datenqualität	ID definiert und eindeutig (d.h. darf max. 1 mal vorkommen)	Liegt vor	Geburtsdatum in europäischem Format: TT.MM.YY., sonst umwandeln	Alter < 120	muß vorliegen	1. 12 ≤ Anzahl Ziffern ≤ 16 2. Korrekte Prüfsumme (bspw. Luhn-Algorithmus ¹)		Währung in EUR, sonst umwandeln
Relevant für Wertschöpfung per Service/Empfehlung	-	-	Altersgruppen	Ja, für Empfehlungen Aber bspw. auch für Ansprache Kunde	Ja, bspw. Wohnort		Ja, für Empfehlungen	Ja, für Empfehlungen



Es gibt für Anzahl, Art und Umfang der Features kein richtig oder falsch.
Art und Umfang entwickelt sich über die Jahre, bspw. aufgrund gesetzlicher Anforderungen, Business Logic, ...

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE SYSTEM.

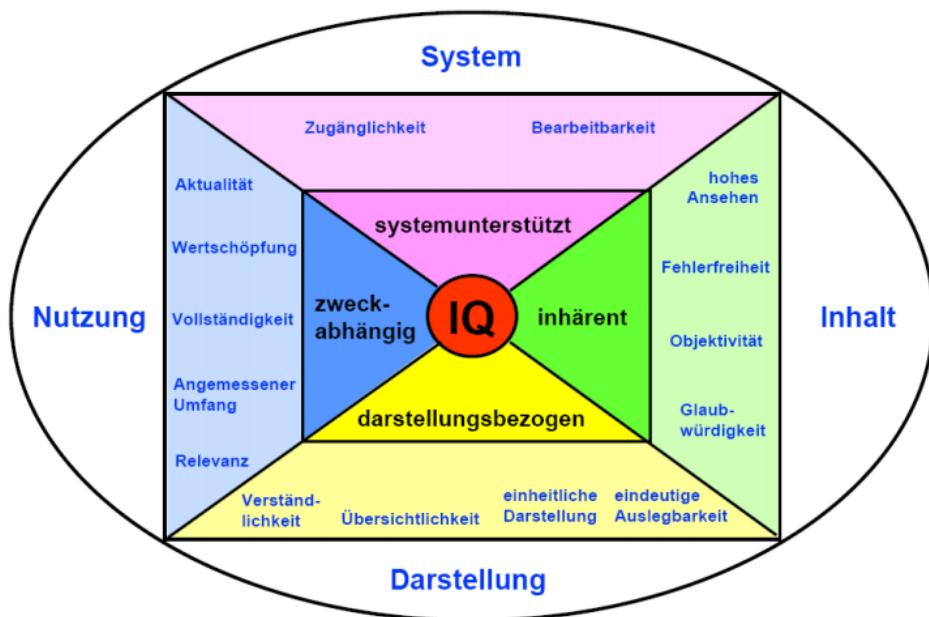


Informationen haben...

Zugänglichkeit (accessibility): wenn sie anhand einfacher Verfahren auf direktem Weg für den Anwender abrufbar sind.

(leicht) Bearbeitbarkeit (ease of manipulation): wenn sie leicht zu ändern/ für unterschiedliche Zwecke zu verwenden sind.

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE INHALT.



Informationen haben...

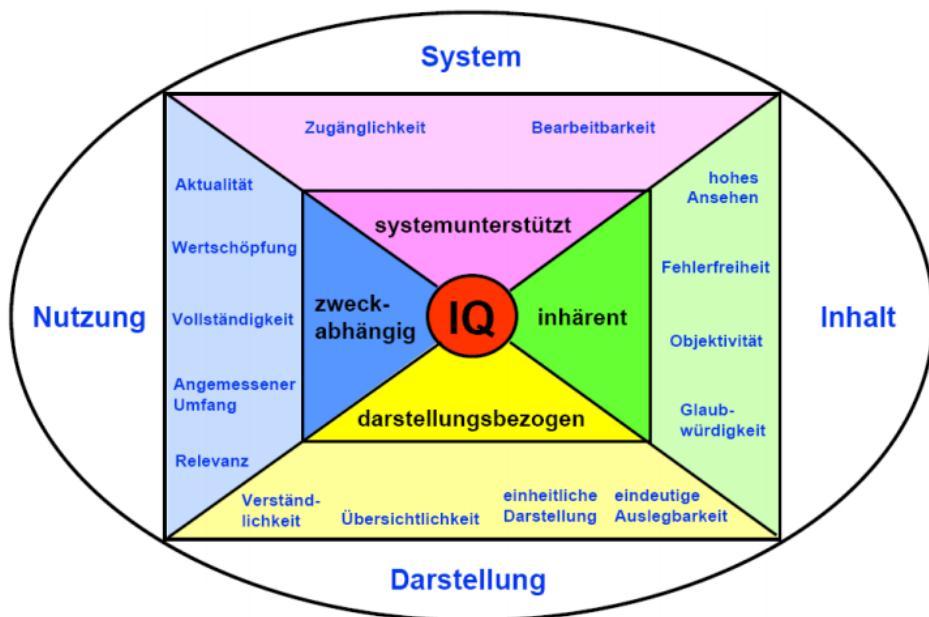
Hohes Ansehen: (reputation): wenn die Informationsquelle, das Transportmedium und das verarbeitende System im Ruf einer hohen Vertrauenswürdigkeit und Kompetenz stehen.

Fehlerfreiheit (free of error): wenn sie mit der Realität übereinstimmen.

Objektivität (objectivity): wenn sie streng sachlich und wertfrei sind

Glaubwürdigkeit (believability): wenn Zertifikate einen hohen Qualitätsstandard ausweisen oder die Informationsgewinnung und –verbreitung mit hohem Aufwand betrieben werden.

6. ÜBERSICHT DATENQUALITÄT. KATEGORIE DARSTELLUNG.



Informationen haben...

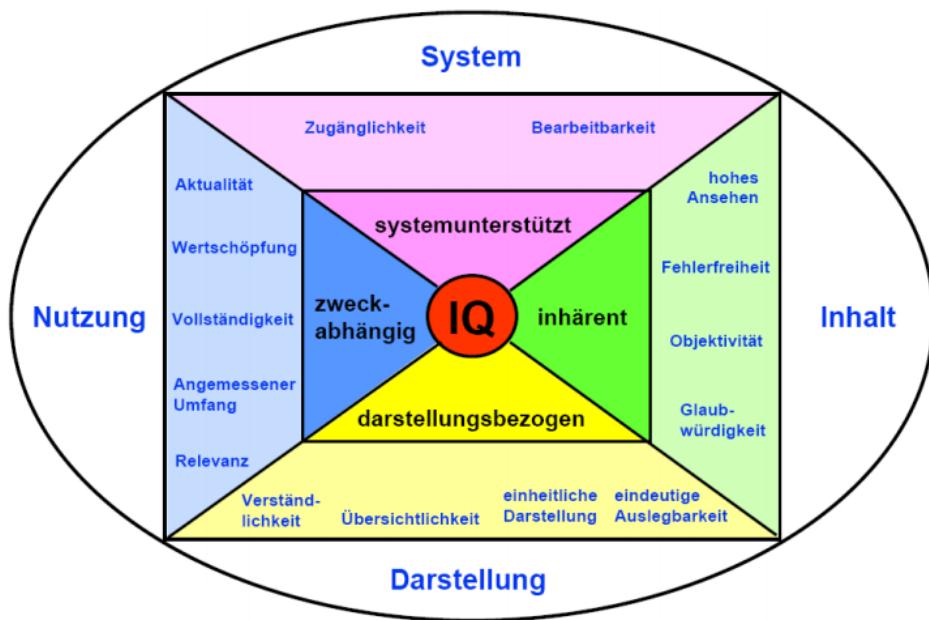
Eindeutig. Auslegbarkeit (interpretability): wenn sie in gleicher, fachlich korrekter Art und Weise begriffen werden

Einheitl. Darstellung (consistent representation): wenn die Informationen fortlaufend auf dieselbe Art und Weise abgebildet werden.

Übersichtlichkeit (concise representation): wenn genau die benötigten Informationen in einem passenden und leicht fassbaren Format dargestellt sind.

Verständlichkeit (understandability): wenn sie unmittelbar von den Anwendern verstanden und für deren Zwecke eingesetzt werden können.

6. ÜBERSICHT DATENQUALITÄT KATEGORIE NUTZUNG.



Informationen haben...

Aktualität (timeliness): wenn sie die tatsächliche Eigenschaft des beschriebenen Objektes zeitnah abbilden.

Wertschöpfung (value-added): wenn ihre Nutzung zu quantifizierbaren Steigerung einer monetären Zielfunktion führen kann.

Vollständigkeit (completeness): wenn sie nicht fehlen und zu den festgelegten Zeitpunkten in den jeweiligen Prozessschritten zur Verfügung stehen.

Angemessener Umfang (appropriate amount of data): wenn die Menge der verfügbaren Information den gestellten Anforderungen genügt.

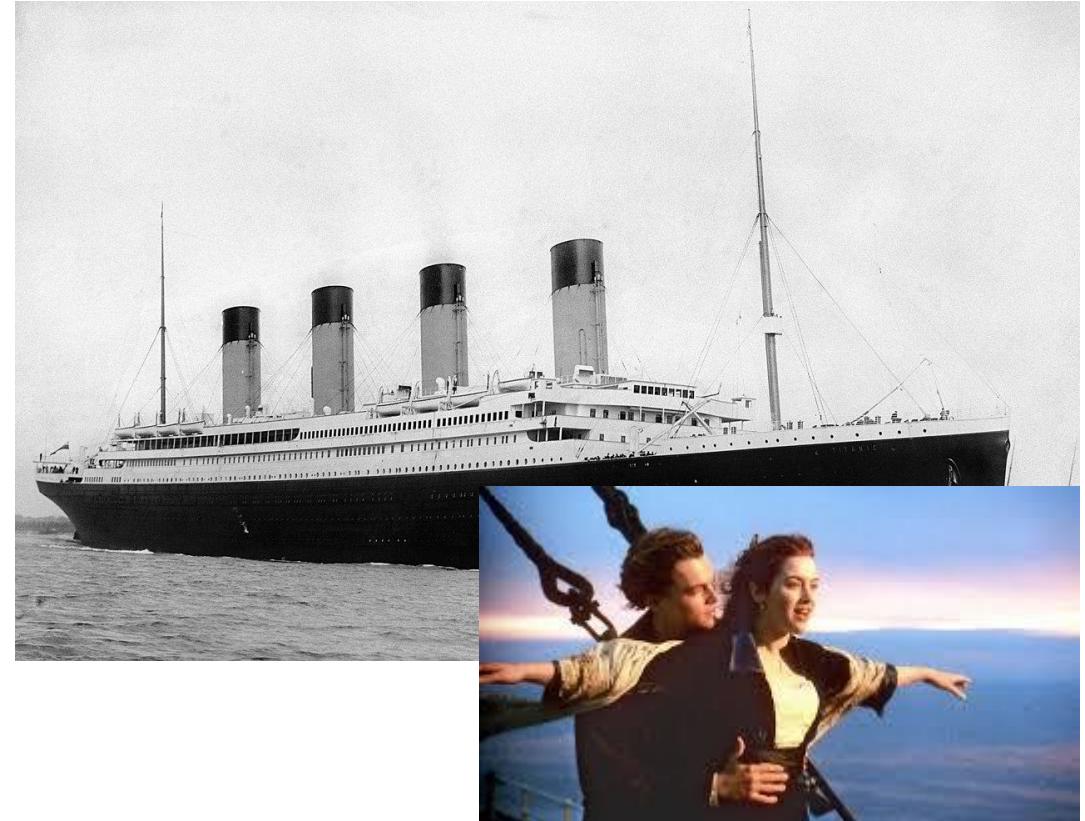
Relevanz (relevancy): wenn sie für den Anwender notwendige Informationen liefern.



DATA SCIENCE CASE STUDY PASSAGIERE TITANIC

ÜBERSICHT DATA SCIENCE AM FALLBEISPIEL TITANIC.

- Passagierliste Titanic ist beliebter Datensatz für Data Science:
 - Kleiner Datensatz (1310 Zeilen à 14 Spalten)
 - Deckt ganzen Workflow inkl. üblicher Probleme ab
 - Fragestellung einfach verständlich und interessant
- Im Notebook zur Vorlesung wird folgendes gemacht:
 - Import/ Laden der Daten
 - Data Engineering:
 - Daten säubern und aufbereiten
 - neue Features erstellen
 - Univariate Datenanalysen (Analyse eines Features)
 - Multivariate Datenanalysen (Analyse mehrerer Features)
 - Annäherung Zielvariable per manueller Optimierung

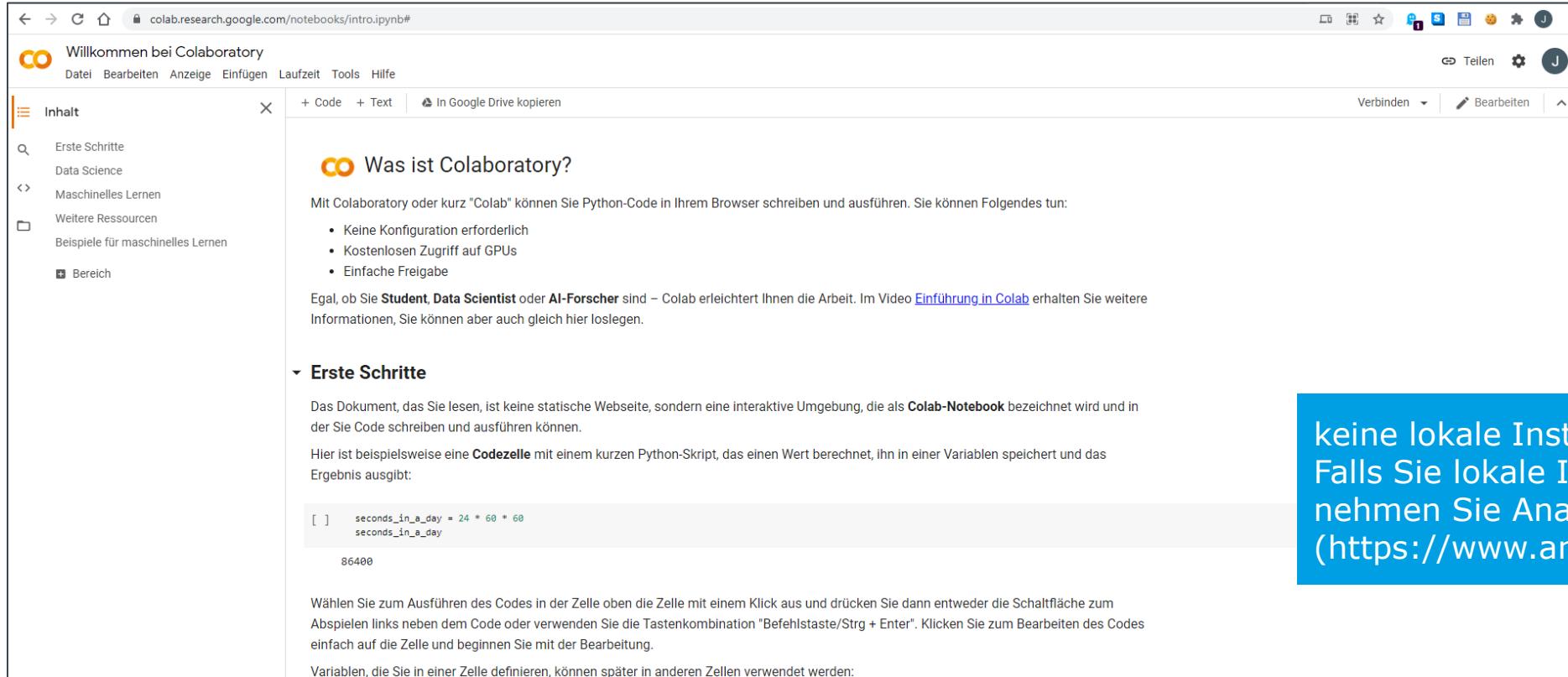


Das Titanic-Notebook ist sehr detailliert, wir werden uns in der heutigen Vorlesung nur die wichtigsten Sachen ansehen

ALS PROGRAMMIERSPRACHE WERDEN WIR PYTHON EINSETZEN.

- Einfach zu erlernen und zu benutzen.
- Kostenfrei verfügbar.
- Sehr viele kostenfreie, leistungsfähige Bibliotheken, die viel Programmierarbeit abnehmen.
- Flexibel und weit einsetzbar.
- Sehr häufig für Data Science und Künstliche Intelligenz eingesetzt.
- Sehr viele frei verfügbare Beispiele und Tutorials.

IM RAHMEN DER VORLESUNG WERDEN SIE PROGRAMMIEREN, EMPFEHLUNG PROGRAMMIERUMGEBUNG IST GOOGLE COLAB.



The screenshot shows the Google Colab interface. On the left, there's a sidebar titled 'Inhalt' with sections like 'Erste Schritte', 'Data Science', 'Maschinelles Lernen', 'Weitere Ressourcen', and 'Beispiele für maschinelles Lernen'. The main content area has a title 'Was ist Colaboratory?' and text explaining what Colab is and how it works. It includes a bulleted list: 'Keine Konfiguration erforderlich', 'Kostenlosen Zugriff auf GPUs', and 'Einfache Freigabe'. Below this, there's a note about being a student, data scientist, or AI-researcher, and a link to a video introduction. A code cell is shown with the Python code:

```
[ ] seconds_in_a_day = 24 * 60 * 60
```

 and the output:

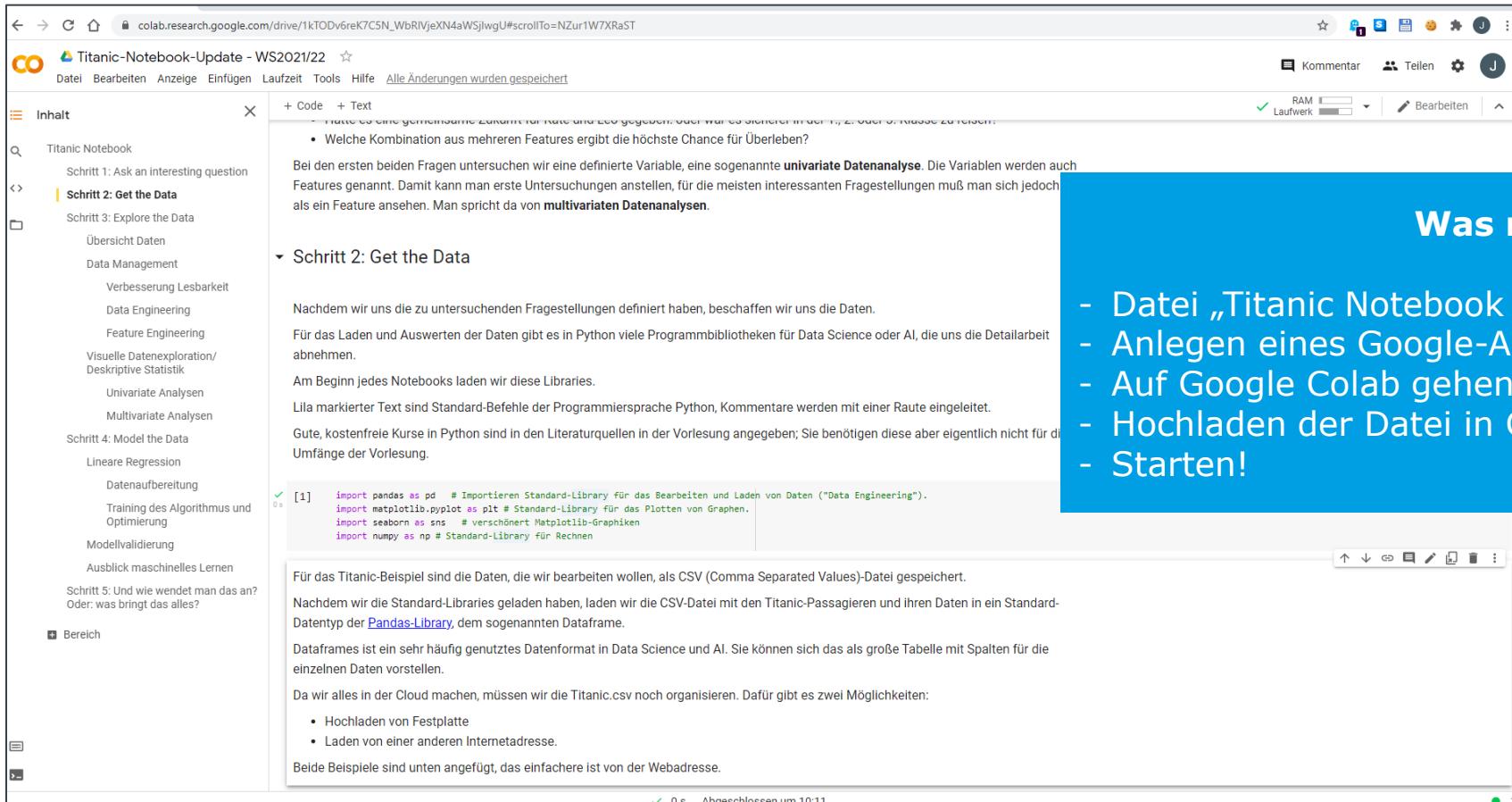
```
86400
```

. At the bottom, there's a note about executing code in cells and defining variables.

**keine lokale Installation notwendig.
Falls Sie lokale Installation bevorzugen,
nehmen Sie Anaconda
(<https://www.anaconda.com/>)**

<https://colab.research.google.com/notebooks/intro.ipynb#>

WIR SCHAUEN UNS DIE EINZELNEN SCHRITTE ANHAND EINES NOTEBOOKS AUF COLAB AN.



The screenshot shows a Google Colab notebook titled "Titanic-Notebook-Update - WS2021/22". The notebook structure is as follows:

- Inhalt** (Content) sidebar:
 - Schritt 1: Ask an interesting question
 - Schritt 2: Get the Data** (highlighted)
 - Schritt 3: Explore the Data
 - Schritt 4: Model the Data
 - Ausblick maschinelles Lernen
 - Schritt 5: Und wie wendet man das an? Oder: was bringt das alles?
 - Bereich
- Schritt 2: Get the Data** section:
 - Nachdem wir uns die zu untersuchenden Fragestellungen definiert haben, beschaffen wir uns die Daten.
 - Für das Laden und Auswerten der Daten gibt es in Python viele Programmobilitheken für Data Science oder AI, die uns die Detailarbeit abnehmen.
 - Am Beginn jedes Notebooks laden wir diese Libraries.
 - Lila markierter Text sind Standard-Befehle der Programmiersprache Python, Kommentare werden mit einer Raute eingeleitet.
 - Gute, kostenfreie Kurse in Python sind in den Literaturquellen in der Vorlesung angegeben; Sie benötigen diese aber eigentlich nicht für die Umfänge der Vorlesung.
- Code cell [1]:

```
import pandas as pd # Importieren Standard-Library für das Bearbeiten und Laden von Daten ("Data Engineering").
import matplotlib.pyplot as plt # Standard-Library für das Plotten von Graphen.
import seaborn as sns # verschönert Matplotlib-Grafiken
import numpy as np # Standard-Library für Rechnen
```
- Schritt 5: Und wie wendet man das an? Oder: was bringt das alles?** section:
 - Für das Titanic-Beispiel sind die Daten, die wir bearbeiten wollen, als CSV (Comma Separated Values)-Datei gespeichert.
 - Nachdem wir die Standard-Libraries geladen haben, laden wir die CSV-Datei mit den Titanic-Passagieren und ihren Daten in ein Standard-Datentyp der [Pandas-Library](#), dem sogenannten Dataframe.
 - Dataframes ist ein sehr häufig genutztes Datenformat in Data Science und AI. Sie können sich das als große Tabelle mit Spalten für die einzelnen Daten vorstellen.
 - Da wir alles in der Cloud machen, müssen wir die `Titanic.csv` noch organisieren. Dafür gibt es zwei Möglichkeiten:
 - Hochladen von Festplatte
 - Laden von einer anderen Internetadresse.
 - Beide Beispiele sind unten angefügt, das einfachere ist von der Webadresse.

At the bottom of the notebook, it says "0 s Abgeschlossen um 10:11".

Was müssen Sie tun:

- Datei „Titanic Notebook“ runterladen aus Github.
- Anlegen eines Google-Accounts (falls Sie nicht schon haben).
- Auf Google Colab gehen: [Link](#)
- Hochladen der Datei in Google Colab.
- Starten!

HANDS ON DATA SCIENCE-FALLBEISPIEL

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Was ist die Fragestellung?

Was würde ich tun, wenn ich alle Daten hätte?

Was möchte ich abschätzen/ vorhersagen?

- Wie hoch war die Überlebens-Chance eines Passagiers der Titanic?
- Hätte es eine gemeinsame Zukunft für Kate und Leonardo gegeben: oder war es sicherer, in der 1., 2. oder 3. Klasse zu reisen?
- Was ist der sicherste Indikator für das Überleben eines Passagiers?

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Wie wurden die Daten generiert?

Welche Daten sind relevant?

Privacy??

- Daten sind aus Passagierliste abgetippt
- Relevante Daten: schauen wir es uns an
- Privacy: eher nicht relevant

ÜBERSICHT WICHTIGER DATA ENGINEERING TECHNIKEN

- **Verbessern Lesbarkeit/ Erklärbarkeit:** sprechende Namen für Eigenschaften (**Features**) und Werte.
- **Imputation:** fehlende Werte löschen oder ersetzen (bspw. Mittelwert, definierter Wert,)
- **Typumwandlungen:** beim Einlesen der Daten werden numerische Features oft als Text erkannt.
- **Diskretisation:** Einteilen von Werten mit großem Wertebereich in Gruppen, bspw. Alter (Kind, Teenager, Erwachsene, ...)
- **Categorization:** Werte mit beschränktem Wertebereich zusammenfassen (bspw. Farben, Wochentage, Geschlecht).
- **Outliers:** Werte, die sehr unterschiedlich zu restlichen Werten sind löschen oder per Standardwert ersetzen (bspw. Größe)
- **Normalisation/Scaling:** Werte innerhalb gewissen Wertebereichs bringen für Vermeiden Verzerrungen (bspw. Größe)
- **Feature Splitting:** Aufteilen Features für Infogewinn (bspw. Name in Vor-/Nachname, Adresse in Stadt und Straße)
- **One-hot-encoding:** Generieren neuer Features aus einem Feature (bspw. einzelne Tage statt Wochentage).
- Neue Features: Bauen neuer Features aus bestehenden oder Berechnungen (bspw. BMI aus Gewicht und Größe).
- **Feature removal:** Unwichtige Features löschen.



Data Engineering ist ein iterativer Prozess entlang des gesamten Use Cases und umfaßt oft 60 - 80% der Arbeit!

VERANSCHAULICHUNG DATA ENGINEERING ANHAND PASSAGIERLISTE TITANIC

index	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.55	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0	1	2	113781	151.55	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
5	1	1	Anderson, Mr. Harry	male	48.0	0	0	19952	26.55	E12	S	3	NaN	New York, NY
6	1	1	Andrews, Miss. Kornelia Theodosia	female	63.0	1	0	13502	77.9583	D7	S	10	NaN	Hudson, NY
7	1	0	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S	NaN	NaN	Belfast, NI
8	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0	2	0	11769	51.4792	C101	S	D	NaN	Bayside, Queens, NY
9	1	0	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C	NaN	22.0	Montevideo, Uruguay

- **Verbessern Lesbarkeit/ Erklärbarkeit:** sprechende Namen für sibsp, parch, pclass, fare.
- **Categorization:** Einsetzen von beschränkten Werten für Embarked wie Southampton, Cherbourg, ...
- **One-hot-encoding:** Generieren neuer Features aus einem Feature (bspw. einzelne Tage statt Wochentage).
- **Diskretisation:** Einteilen von Alter in Altergruppen wie Kind, Teenager, Erwachsene,
- **Neue Features/ Feature splitting:** Aufbau neues Feature HomeCountry aus Homedest.
- Mehrdeutigkeiten auflösen: cabin oder Name.
- **Imputation:** fehlende Werte löschen oder ersetzen für boat oder body. → Aufwendig, sehr oft erfahrungsgesetzten.
- **Normalisation/Scaling:** bspw. für Alter und Ticketpreis.
- **Feature removal:** Löschen bspw. von cabin

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Daten aufzeichnen (visuelles Verständnis)

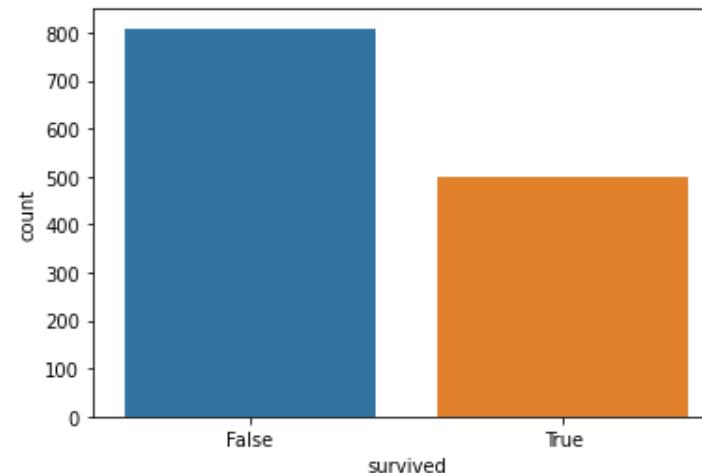
Gibt es Anomalien? Unplausible Werte?

Sehen Sie Muster in den Daten?

DATENEXPLORATION: UNTERSUCHEN EINZELNER MERKMALE (UNIVARIATE ANALYSEN).

Wie viele Passagiere haben insgesamt überlebt?

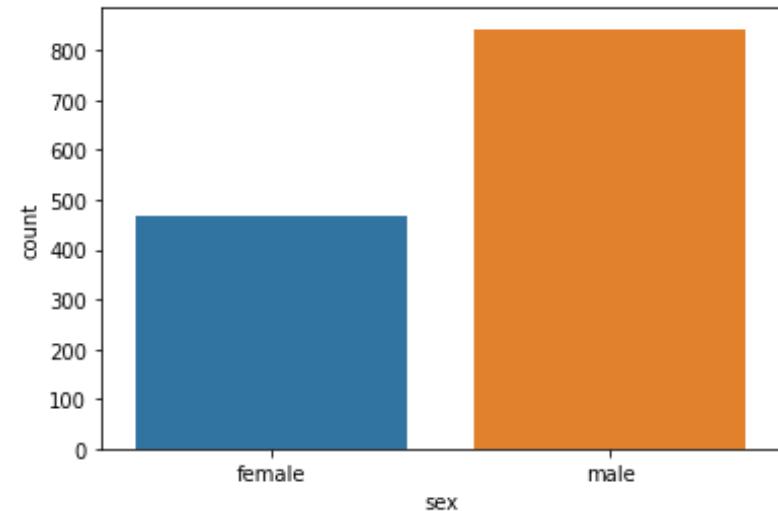
500 von 1309



Wie viele Frauen/ Männer waren an Bord?

Frauen: 466

Männer: 843



► Empfehlung: Einsatz univariater Analyse am Anfang jeder Datenanalyse, um Muster oder Anomalien in Daten zu erkennen.

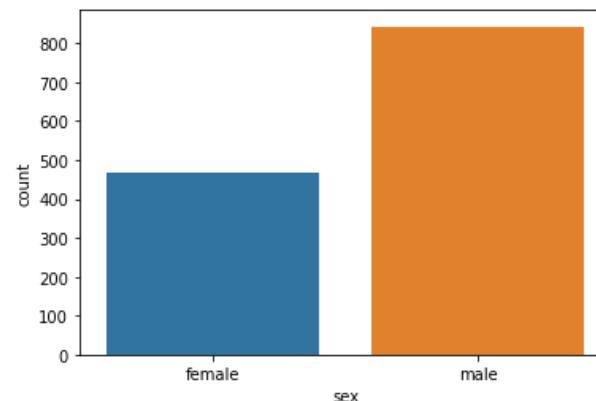
DATENEXPLORATION: UNTERSUCHEN MEHRERER MERKMALE (MULTIVARIATE ANALYSEN).

Zwei Attribute: wie viele Frauen/Männer überlebten?

Geschlecht an Bord: Frauen 466, Männer 843

Überlebt:

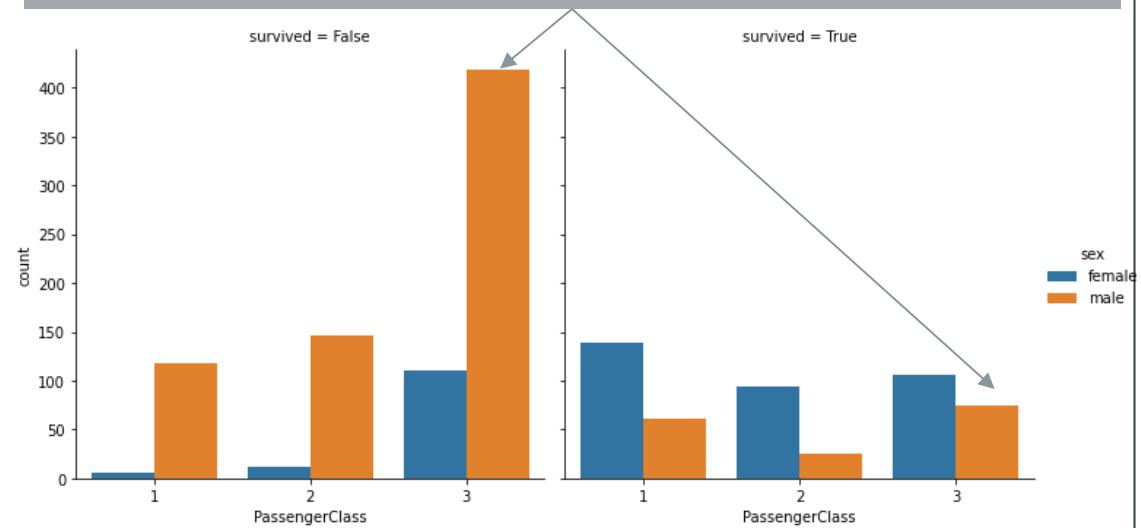
- Frauen überlebt: $339/466 = 72\%$
- Männer überlebt: $161/843 = 19\%$



Bedingte Wahrscheinlichkeit daß, gegeben ein Mann, er in Passagierklasse 3 war

Drei Attribute: Wie viele Männer/Frauen überlebten in den verschiedenen Passagierklassen?

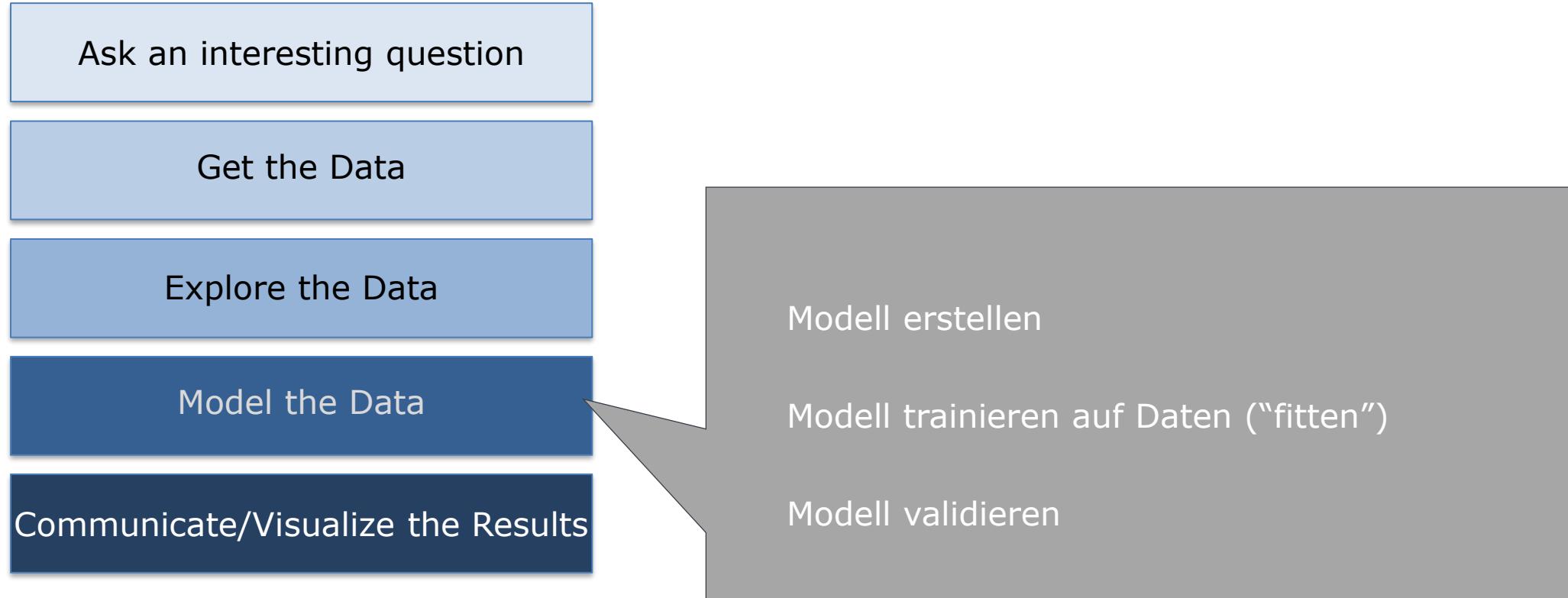
Wahrscheinlichkeit, daß ein Mann in Passagierklasse 3 überlebt:
 $\Pr(\text{Mann} \mid \text{PC}=3, \text{überlebt})$



Verteilung Geschlechter aus Passagierklassen:

- Anzahl Männer & Passagierklasse: PC1 = 179, PC2 = 171, PC3 = 493
- Anzahl Frauen & Passagierklasse: PC1 = 144, PC2 = 106, PC3 = 216
- $\Pr(\text{Mann} \mid \text{PC}=3) = \Pr(\text{Mann und PC3}) / \Pr(\text{Mann}) = 493/843 = 58\%$
- $\Pr(\text{Frau} \mid \text{PC}=1) = \Pr(\text{Frau und PC1}) / \Pr(\text{Frau}) = 144/466 = 30\%$

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



MODEL THE DATA: AUSBLICK AUF DIE SPÄTEREN VORLESUNGEN.

Trainieren Machine Learning Model (vereinfacht)

```
from sklearn.ensemble import RandomForestClassifier
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
random_forest = RandomForestClassifier(n_estimators=200)
random_forest.fit(X_train, y_train)
Y_prediction = random_forest.predict(X_test)
```

Messen Modellgüte per Metriken

Einsatz Confusion Matrix als Metrik für Klassifikation:

		Predicted	
		Ja	Nein
Tatsächlich	Ja	True Positive	False Negative
	Nein	False Positiv	True Negative

$$\text{Accuracy} = \frac{\text{korrekt vorhergesagt}}{\text{Gesamtzahl}} = 97.49\%$$

Anwenden Modell



```
[ ] # notwendigen Features sind: PassengerClass, Sex (0 Frau, 1 Mann), Age, # SiblingSpousesPresent, ParentsChildrenPresent, fare, AgeGroup
Kate = [[1, 0, 17., 0, 1, 150., 2]]
Leo = [[3, 1, 20., 0, 0, 15., 3]]
Billy = [[1, 1, 30., 0, 0, 150., 3]]

# make a prediction
print("Prädiktion Überlebenschance Kate:", logmodel.predict(Kate))
print("Prädiktion Überlebenschance Leo:", logmodel.predict(Leo))
print("Prädiktion Überlebenschance Billy:", logmodel.predict(Billy))

Prädiktion Überlebenschance Kate: [1]
Prädiktion Überlebenschance Leo: [0]
Prädiktion Überlebenschance Billy: [1]
```

Wir werden uns die einzelnen Schritte im Detail in den nächsten Vorlesungen ansehen