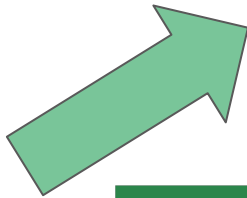


# Modul 2: Umgang mit Arbeitsumgebung, Software und Datenanalyse

Angewandte Datenanalyse für die öffentliche Verwaltung in Bayern (ADA Bayern)

[www.ada-oeffentliche-verwaltung.de](http://www.ada-oeffentliche-verwaltung.de)



**BERD**  
@NFDI



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN



Bayerisches Staatsministerium  
für Digitales



Einführung & Daten kennenlernen	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit	13:05 - 13:45
Pause	13:45 - 14:00
Stratifizierte Zufallsstichprobe	14:00 - 14:20
Teamarbeit	14:20 - 14:55
Ausblick & Tagesabschluss	14:55 - 15:30

# Abschlussstermin 30.04.

10:00 - 12:00: Nur Studis (LMU)

13:30 - 15:00: Abschlusspräsentationen mit allen (LMU / Zoom)

Die Studierenden stellen die Zwischenergebnisse der Gruppen vor + ihre Pläne für die Seminararbeit

(jeweils 10-12 Minuten pro Gruppe + 5 Minuten pro Seminararbeit)

Seminarvorträge: TBD (Was wünscht ihr euch?)

Abgabe der Seminararbeit: TBD (Was wünscht ihr euch?)

# Erste Schritte: Die Daten kennenlernen

Wie machen Sie das normalerweise?

# Erste Schritte: Die Daten kennenlernen

230817\_Abfrage\_Januar-Dezember.csv (read-only) - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

A1 fx Σ = Gericht

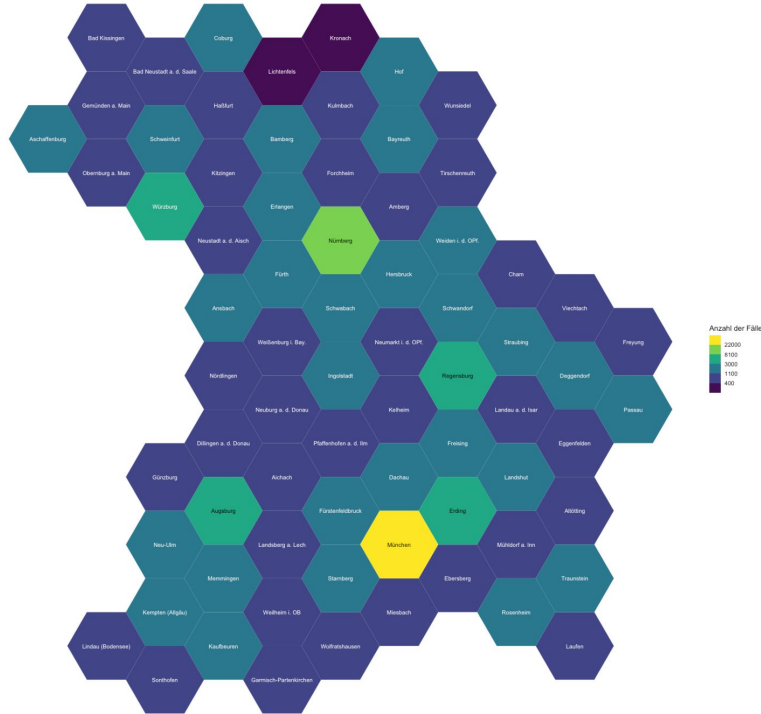
	A	B	C	D	E	F
1	Gericht	Aktenzeichen	Verfahrensstatus	Kurzrubrum	Streitwert in EURO	Gesamtstreitgegenstand
2	Amtsgericht Aichach	101 C 1/18 WEG	weggelegt	2bfeff69dafc780	2200	Forderung
3	Amtsgericht Aichach	101 C 1/18 WEG	weggelegt	2bfeff69dafc780	2200	Forderung
4	Amtsgericht Aichach	101 C 16/18	weggelegt	75e33dfe8630700	606.9	NA
5	Amtsgericht Aichach	101 C 16/18	weggelegt	75e33dfe8630700	606.9	NA
6	Amtsgericht Aichach	101 C 18/18	weggelegt	fa9d6c2f19cfe3a	390.61	NA
7	Amtsgericht Aichach	101 C 18/18	weggelegt	fa9d6c2f19cfe3a	390.61	NA
8	Amtsgericht Aichach	101 C 19/18	weggelegt	217770d2888f15c		500 NA
9	Amtsgericht Aichach	101 C 19/18	weggelegt	217770d2888f15c		500 NA
10	Amtsgericht Aichach	101 C 21/18	weggelegt	4658df2e2d8c17f	108.75	NA
11	Amtsgericht Aichach	101 C 21/18	weggelegt	4658df2e2d8c17f	108.75	NA
12	Amtsgericht Aichach	101 C 2/18	weggelegt	bf90e9aad079c70	608.59	NA
13	Amtsgericht Aichach	101 C 2/18	weggelegt	bf90e9aad079c70	608.59	NA
14	Amtsgericht Aichach	101 C 22/18	weggelegt	3d00543ae236f2c		500 NA
15	Amtsgericht Aichach	101 C 22/18	weggelegt	3d00543ae236f2c		500 NA
16	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
17	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
18	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
19	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
20	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
21	Amtsgericht Aichach	101 C 23/18	weggelegt	790f6ddc660b7e6		4400 Räumung
22	Amtsgericht Aichach	101 C 24/18	weggelegt	ea194871afc662	100.56	NA
23	Amtsgericht Aichach	101 C 24/18	weggelegt	ea194871afc662	100.56	NA
24	Amtsgericht Aichach	101 C 24/18	weggelegt	ea194871afc662	100.56	NA
25	Amtsgericht Aichach	101 C 26/18	weggelegt	f5c5126d79856ad		1167 NA
26	Amtsgericht Aichach	101 C 26/18	weggelegt	f5c5126d79856ad		1167 NA
27	Amtsgericht Aichach	101 C 28/18	weggelegt	6bd5bc00a13528	197.35	NA
28	Amtsgericht Aichach	101 C 28/18	weggelegt	6bd5bc00a13528	197.35	NA
29	Amtsgericht Aichach	101 C 31/18	weggelegt	8d7a25c4e1d69bf	163.14	NA
30	Amtsgericht Aichach	101 C 31/18	weggelegt	8d7a25c4e1d69bf	163.14	NA
31	Amtsgericht Aichach	101 C 3/18	weggelegt	5ad86a52dd1d5b7	709.35	NA
32	Amtsgericht Aichach	101 C 3/18	weggelegt	5ad86a52dd1d5b7	709.35	NA
33	Amtsgericht Aichach	101 C 33/18	weggelegt	94c0de40eb247f8	1926.58	NA
34	Amtsgericht Aichach	101 C 33/18	weggelegt	94c0de40eb247f8	1926.58	NA
35	Amtsgericht Aichach	101 C 34/18	weggelegt	f93f87ab1dcf5dd	333.2	NA
36	Amtsgericht Aichach	101 C 34/18	weggelegt	f93f87ab1dcf5dd	333.2	NA
37	Amtsgericht Aichach	101 C 37/18	weggelegt	9d36ec4eb6bd89d	470.98	NA
38	Amtsgericht Aichach	101 C 37/18	weggelegt	9d36ec4eb6bd89d	470.98	NA
39	Amtsgericht Aichach	101 C 40/18	weggelegt	85312ad4dd2ba8c	2272.05	NA

230817\_Abfrage\_Januar-Dezember

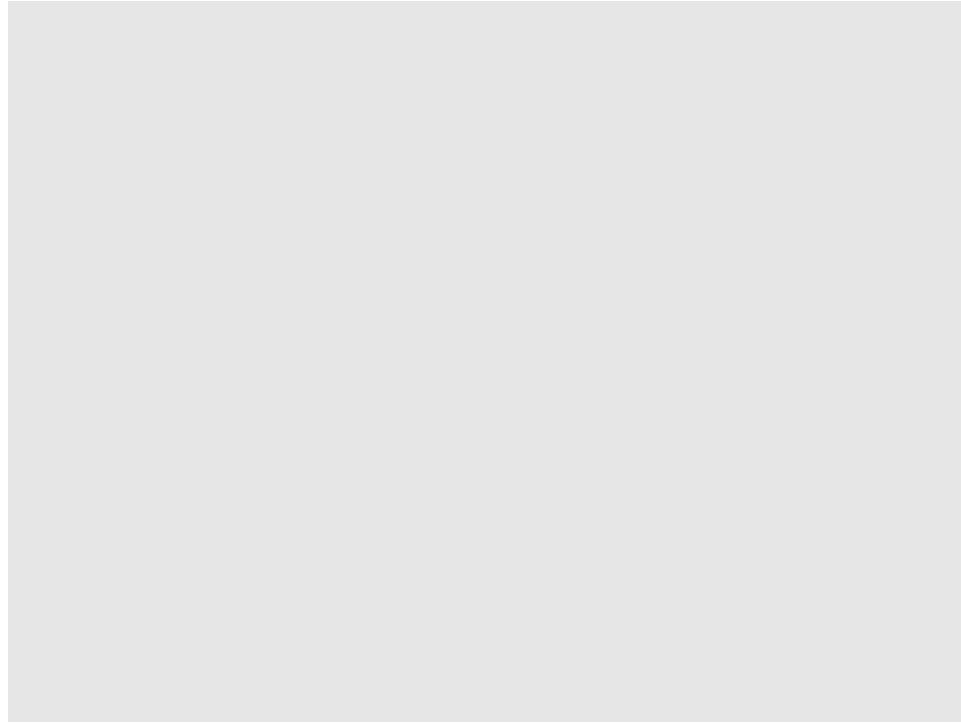
Sheet 1 of 1 | Default | Average: Sum: 0 | 100%

# Wie wir die Daten heute kennenlernen

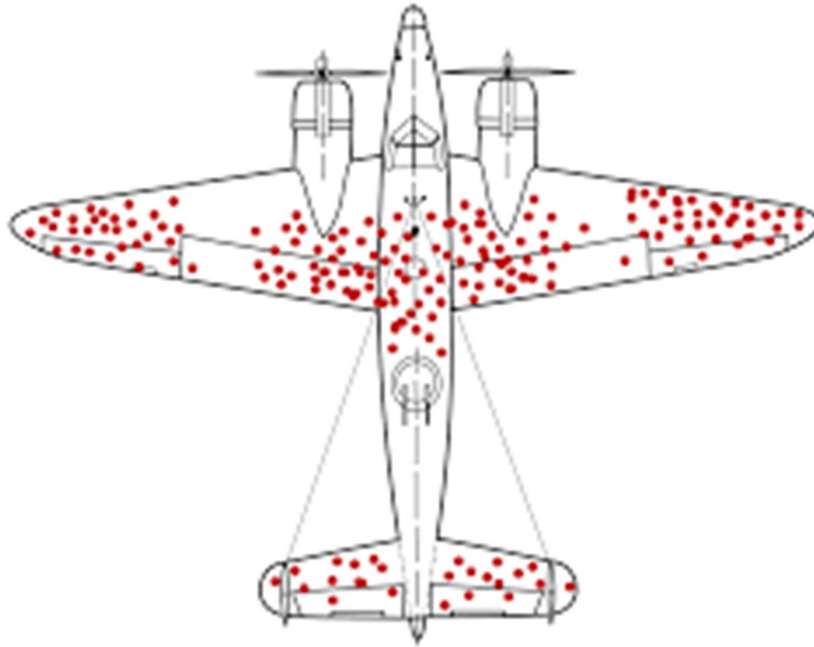
Anzahl der Fälle pro Amtsgerichtsbezirk  
Bayern 2018



Daten generierende Prozesse kennen



# Daten und Daten Generierung im Blick haben



- Wald, Abraham. (1943). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*.
  - Statistical Research Group, Columbia University.
- accessed 8.5.2022  
<https://apps.dtic.mil/docs/citations/ADA091073>
- Illustration of hypothetical damage pattern on a WW2 bomber, based on report above; picture concept by Cameron Moll (2005, claimed on [Twitter](#) and credited by [Mother Jones](#)), new version by [McGeddon](#) based on a Lockheed PV-1 Ventura drawing (2016) CC-BY-SA 4.0







# Wörterbuch für technische Begriffe

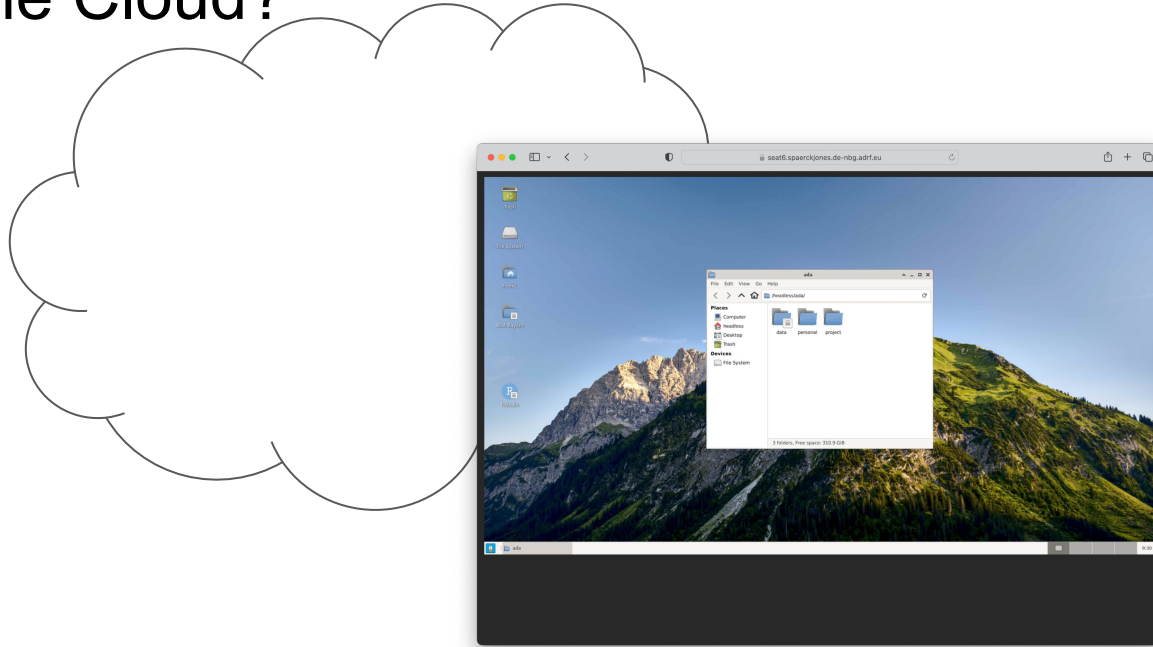
**Open Source (quelloffene) Software** ist Software, die von allen inspiziert, verändert und verbessert werden kann.

**R** ist eine quelloffene (Open Source) Programmiersprache und -umgebung insbesondere für statistische Berechnungen und Grafiken.

**RStudio** ist eine quelloffene integrierte Entwicklungsumgebung (IDE) für die Programmiersprache R (und andere Programmiersprachen).

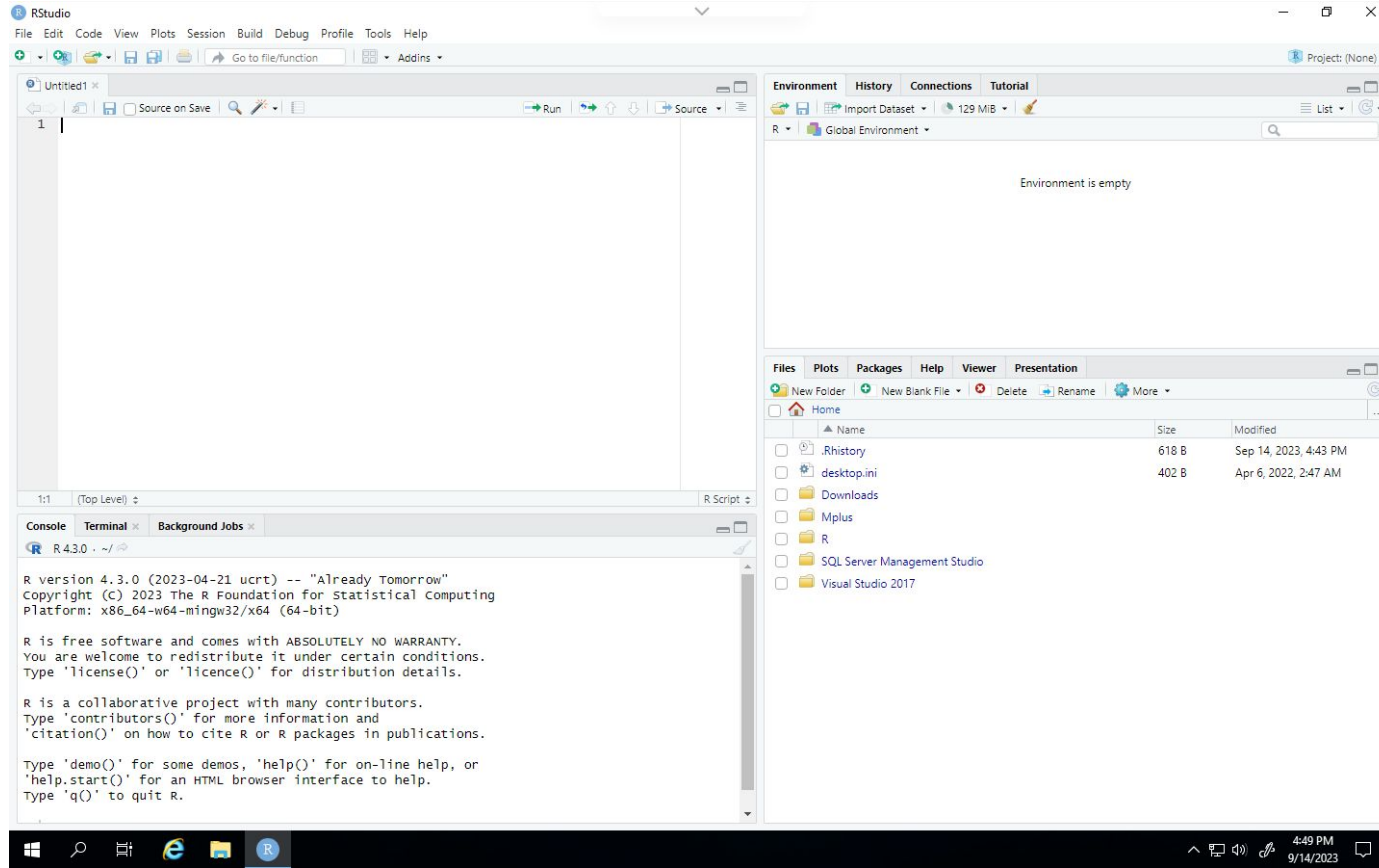
**Quarto** kann zur Erstellung dynamischer, interaktiver und reproduzierbarer wissenschaftlicher und technischer Dokumente verwendet werden.

# Was ist die Cloud?



Stark vereinfacht dargestellt, ist die Cloud nur ein Computer, der an einem anderen Ort steht.

# RStudio in der Cloud



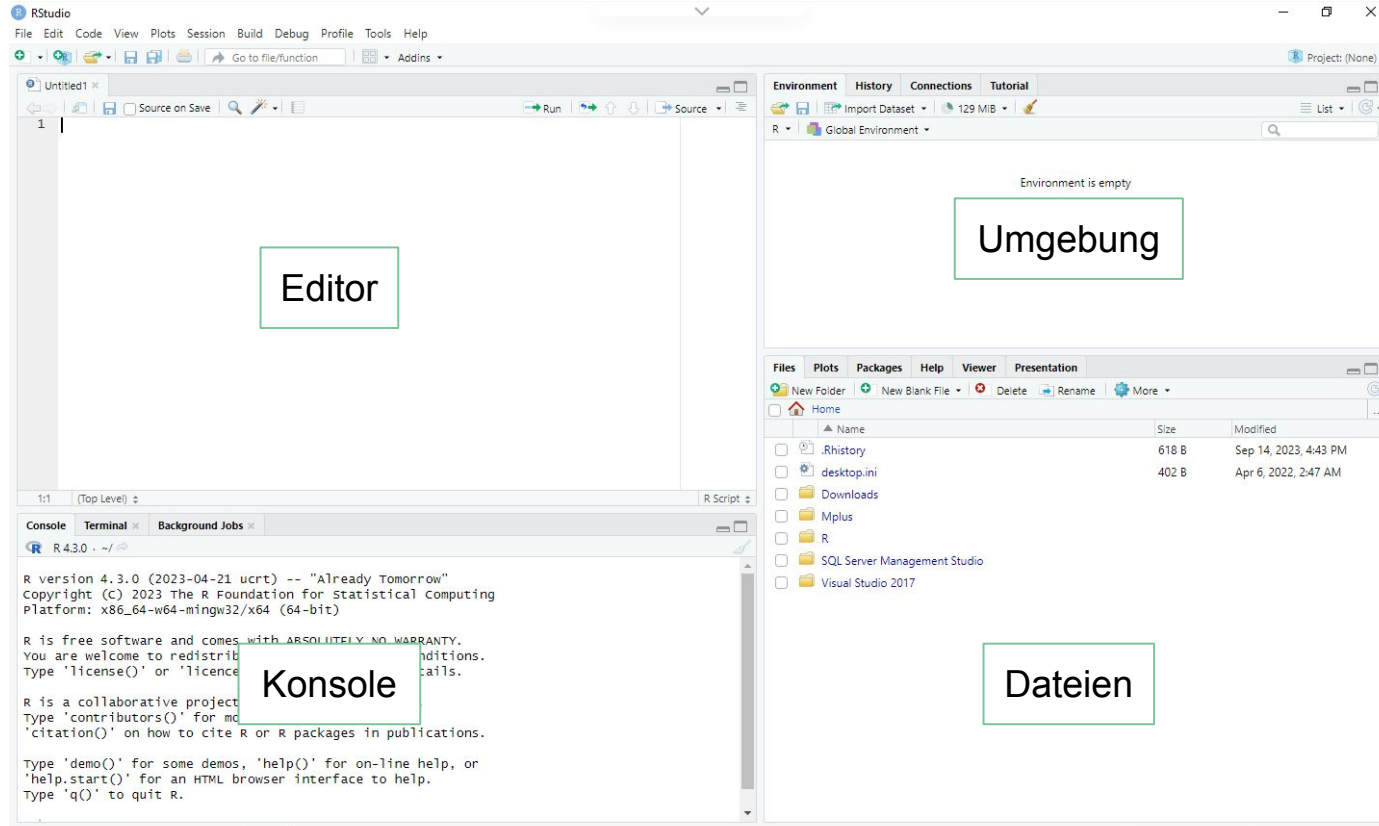
The screenshot displays the RStudio environment. The main editor window shows a blank script titled 'Untitled1'. The console window at the bottom displays the following output:

```
R 4.3.0 ~~/  
  
R version 4.3.0 (2023-04-21 ucrt) -- "Already Tomorrow"  
Copyright (c) 2023 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

The Environment pane on the right shows 'Global Environment' and 'Environment is empty'. The Files pane at the bottom right shows a file explorer view of the Home directory with the following table:

Name	Size	Modified
.Rhistory	618 B	Sep 14, 2023, 4:43 PM
desktop.ini	402 B	Apr 6, 2022, 2:47 AM
Downloads		
Mplus		
R		
SQL Server Management Studio		
Visual Studio 2017		

# RStudio in der Cloud



# Teamarbeit 3: Erste Analysen in R

In dieser Teamarbeit lernen wir zunächst R kennen.

Anschließend führen wir gemeinsam erste Analysen mit den forumSTAR Daten durch.

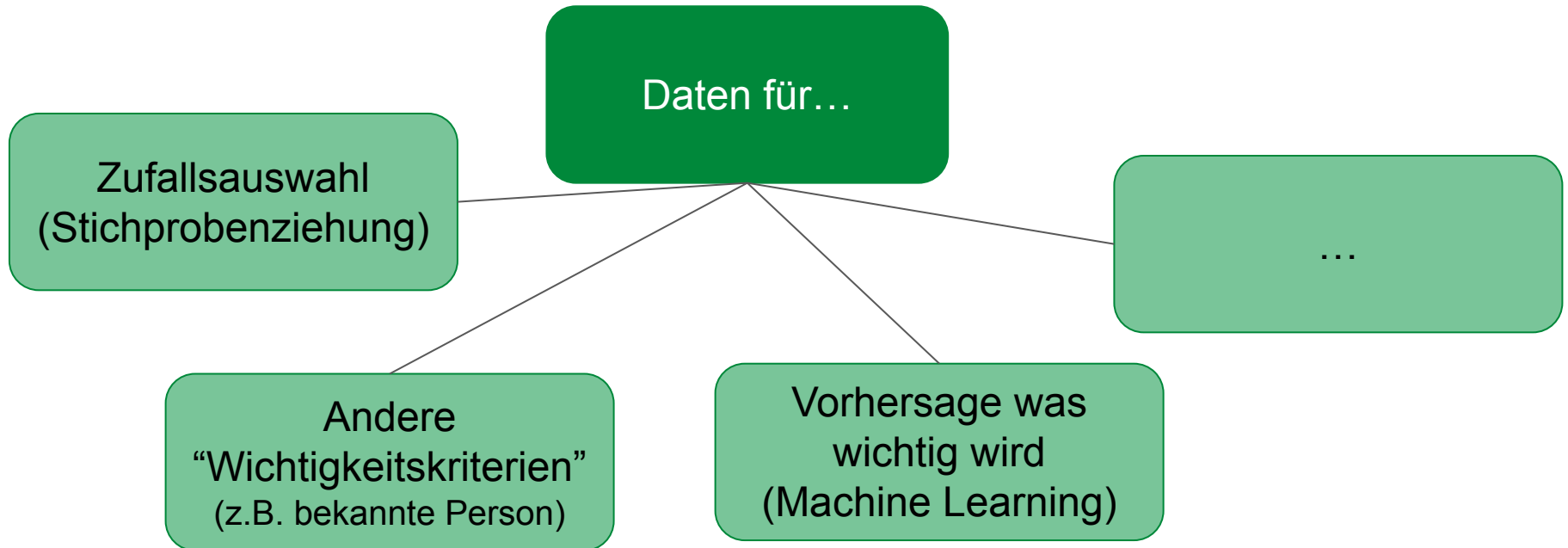
Das Ziel dabei ist es die Daten gut zu verstehen.



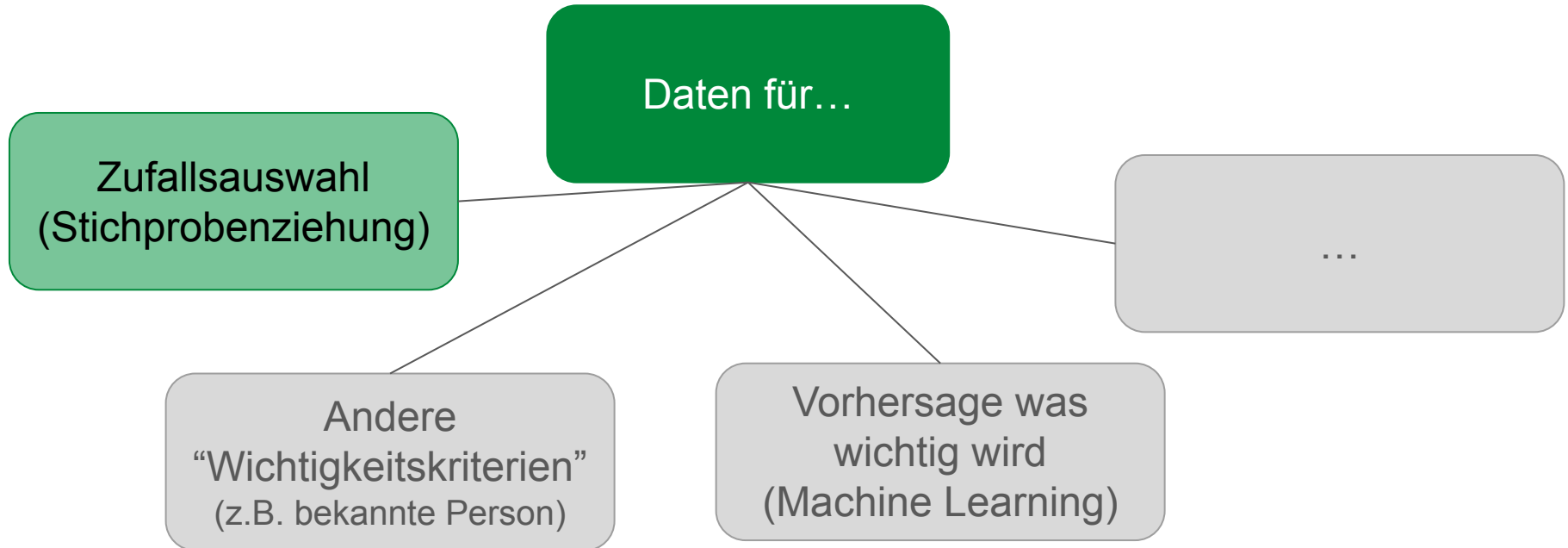
Einführung & Daten kennenlernen	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit	13:05 - 13:45
Pause	13:45 - 14:00
Stratifizierte Zufallsstichprobe	14:00 - 14:20
Teamarbeit	14:20 - 14:55
Ausblick & Tagesabschluss	14:55 - 15:30



# Wie können wir die vorliegenden Daten nutzen?



# Wie können wir die vorliegenden Daten nutzen?



# Im Zweifel Zufall

Terrorist Detektor: 99.9% korrekt

Terrorist klassifiziert als harmloser Passagier: 0.001

Harmloser Passagier klassifiziert als Terrorist: 0.001

1 Person in 1 Millionen ist ein Terrorist (0.000001)

Das heisst bei 999 von 1.000 Leuten wird der Detektor falschen Alarm schlagen.

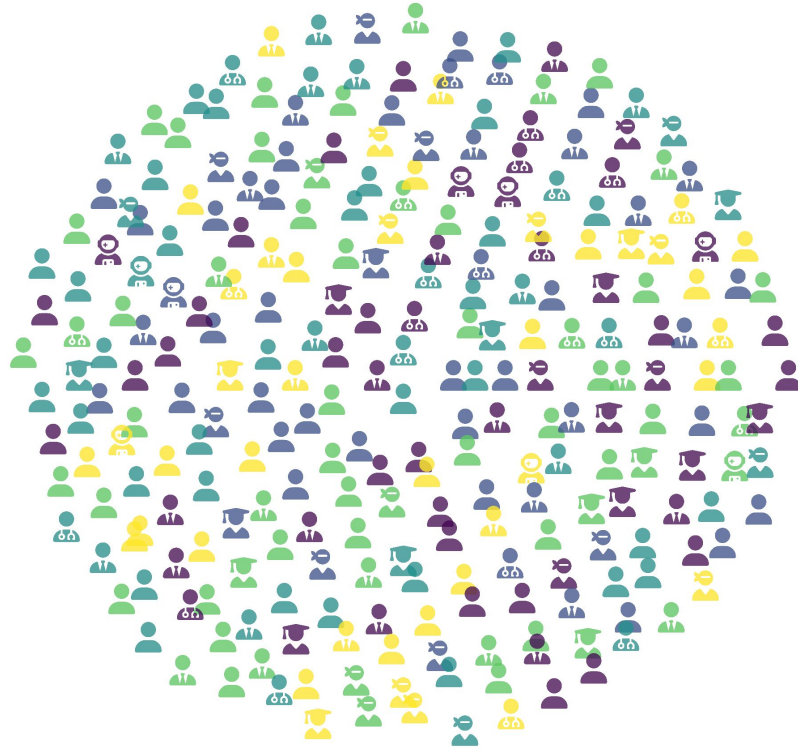


Wir kennen das Problem auch aus der Corona Zeit ...

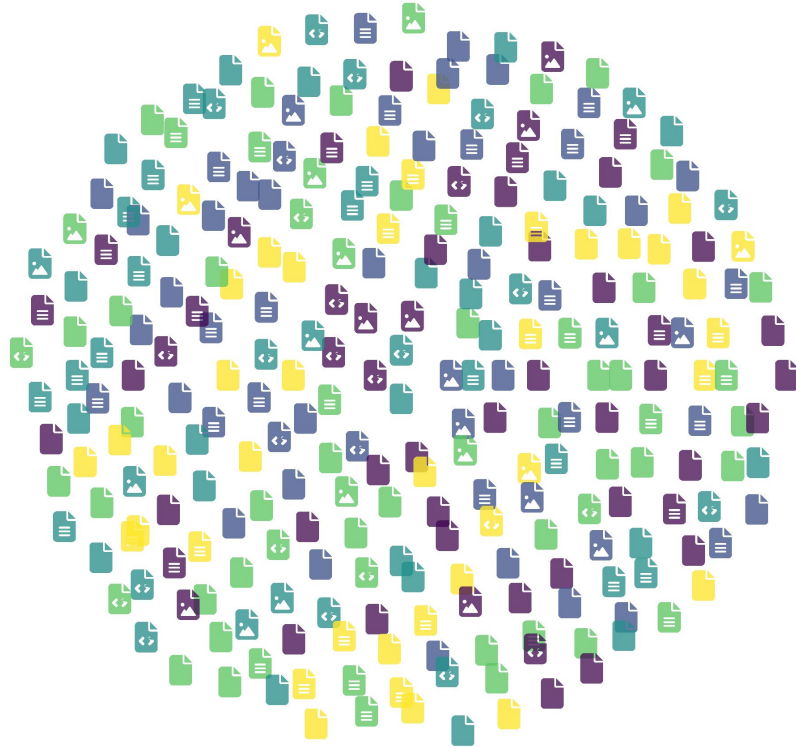


# Die Grundgesamtheit ist die Menge aller Personen...

Grundgesamtheit →

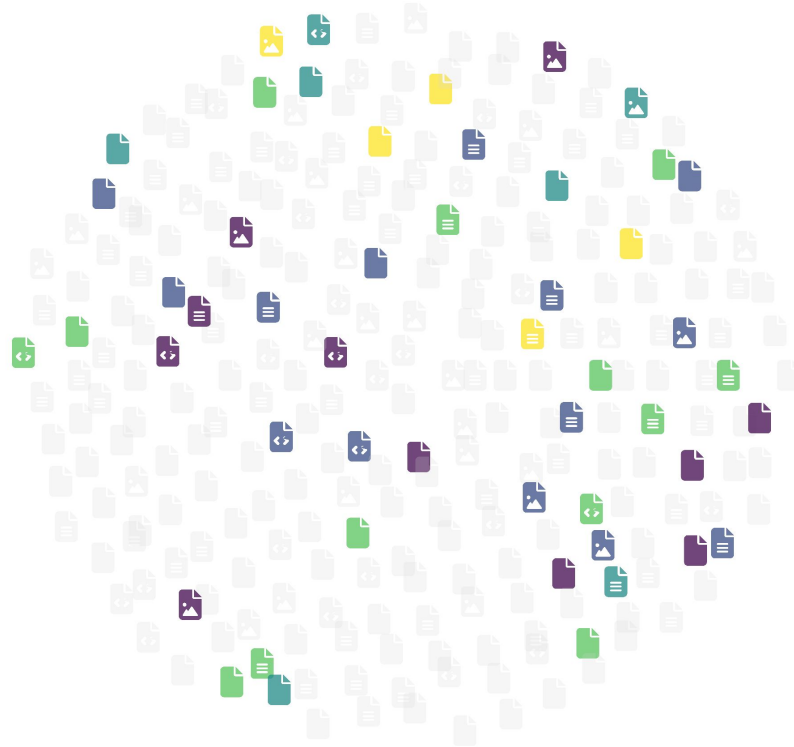


... oder Akten über die wir eine Aussage treffen wollen



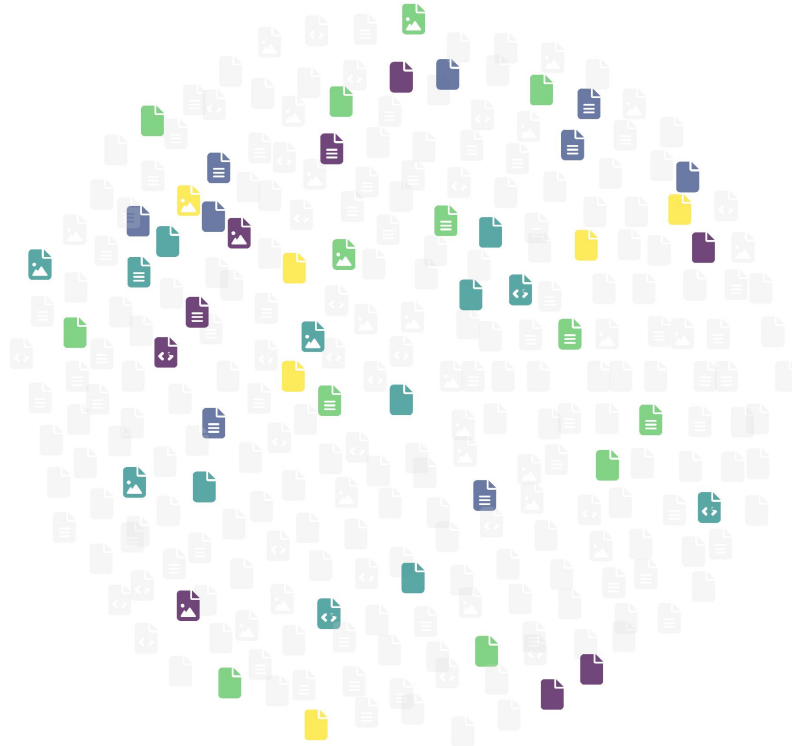
# Eine Stichprobe ist eine Teilmenge der Grundgesamtheit

Stichprobe



# Zufallsstichproben unterscheiden sich bei Wiederholung

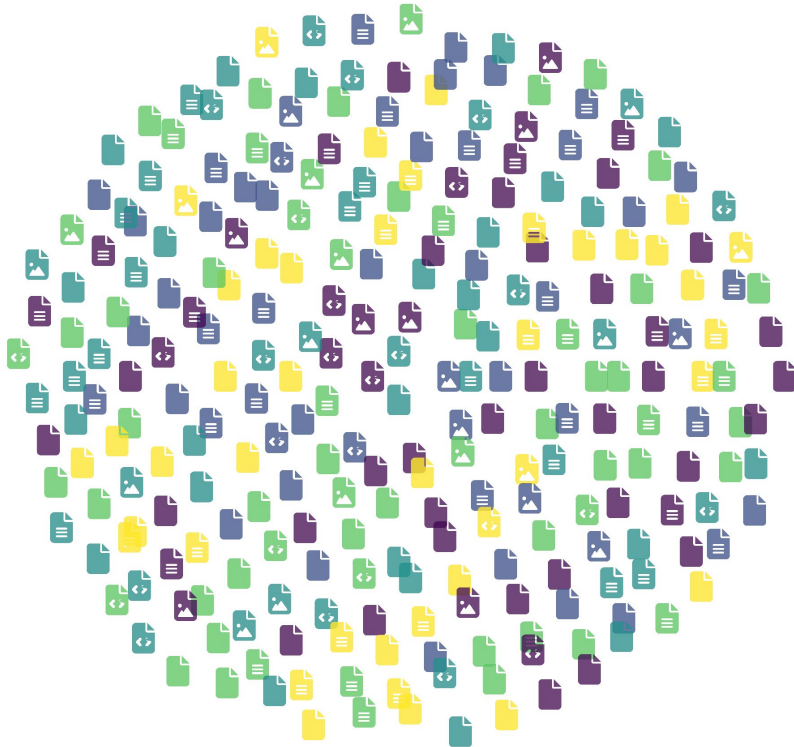
Stichprobe







# Wie können wir eine einfache Zufallsstichprobe ziehen?



Jede Akte soll die gleiche Chance haben!  
(simple random sample, SRS)

Soll jede Akte die gleiche Chance haben?  
(SRS with unequal probabilities)



Losverfahren  
Urne

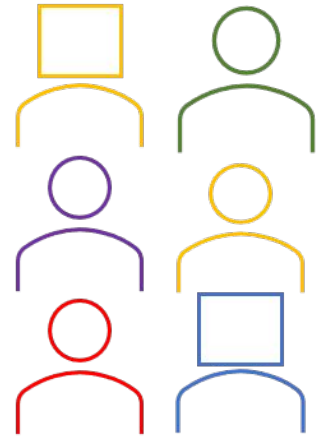
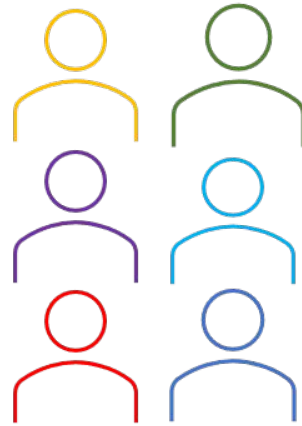
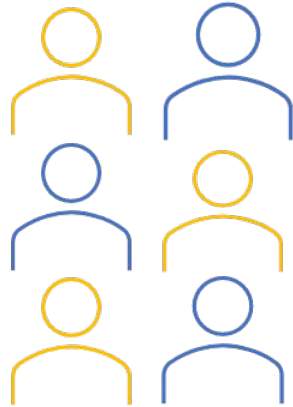
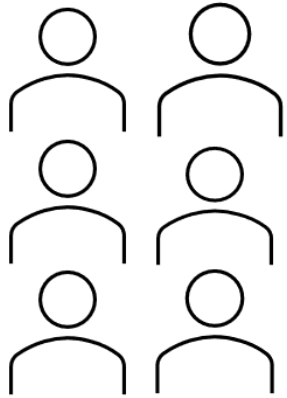


Systematische  
Zufallsauswahl



Zufallszahlentabelle/  
Generator

Wie groß soll unsere Stichprobe sein?



## Umfragen zur Landtagswahl in Bayern

Partei	Infratest dimap 12.09.	Forschungsgruppe Wahlen 08.09	GMS 06.09.
CSU	36%	36%	38%
SPD	9%	9%	8%
Grüne	15%	16%	13%
FDP	3%	4%	4%
Linke	-	-	1%
FW	17%	16%	16%
AfD	13%	12%	14%
Sonstige	7%	7%	6%

Tabelle: SZ • Quelle: [Wahlrecht.de](https://www.wahlrecht.de) • Erstellt mit [Datawrapper](https://www.datawrapper.de)

“Wahlumfragen sind keine Prognosen für das Wahlergebnis. Sie bilden lediglich die politische Stimmung ab. Dabei ist stets ein **statistischer Fehler von 1,5 bis 3 Prozentpunkten (Fehlertoleranz)** zu beachten, wobei sich die Höhe des statistischen Fehlers an der Höhe der Prozentpunkte einer Partei orientiert. Je mehr Prozent eine Partei erhält, desto größer ist auch die Fehlerwahrscheinlichkeit. Eine weitere Unsicherheit ist, dass ein Teil der Wahlberechtigten zum Zeitpunkt der Erhebung noch nicht entschieden haben, wen sie wählen oder ob sie überhaupt wählen.”

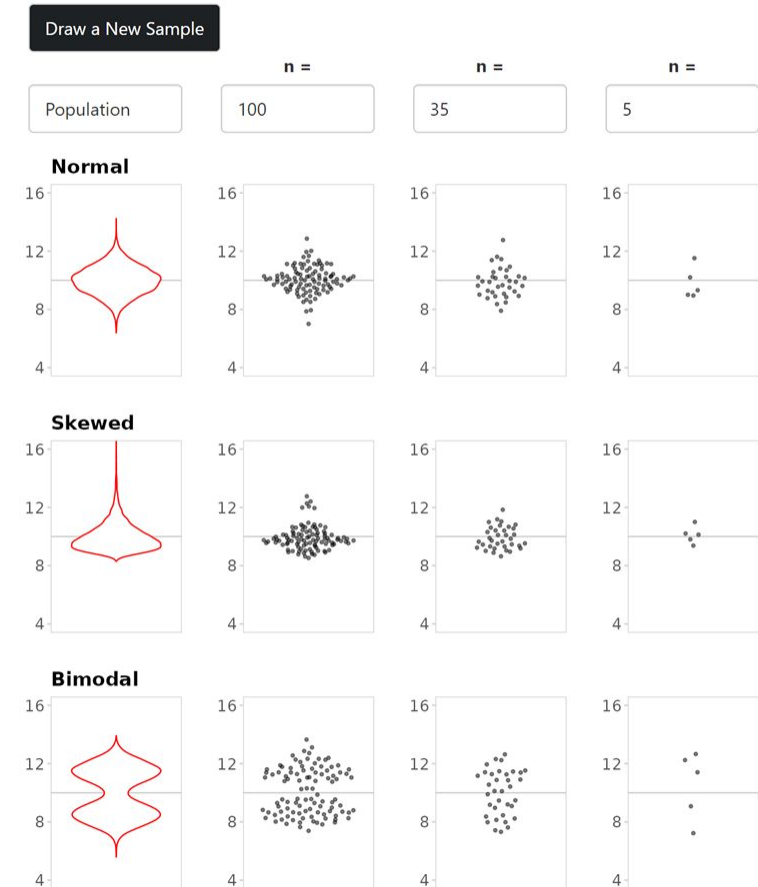
**"Mehr als ein Drittel der befragten Männer (34 Prozent) gibt an, dass sie gegenüber Frauen schon mal handgreiflich werden, um ihnen Respekt einzuflößen."**

*Ergebnis der Umfrage "Spannungsfeld Männlichkeit" der Hilfsorganisation Plan International*

# Verteilung der Streitwerte?

[https://rtools.mayo.edu/size\\_matters/](https://rtools.mayo.edu/size_matters/)

Show questions >>



# Teamarbeit 4: Einfache Stichprobenziehung

→ Gruppenarbeit



Einführung & Daten kennenlernen	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit	13:05 - 13:45
Pause	13:45 - 14:00
Stratifizierte Zufallsstichprobe	14:00 - 14:20
Teamarbeit	14:20 - 14:55
Ausblick & Tagesabschluss	14:55 - 15:30

# Vor- und Nachteile der einfachen Zufallsstichprobe

Alle Akten haben dieselbe  
Auswahlwahrscheinlichkeit.

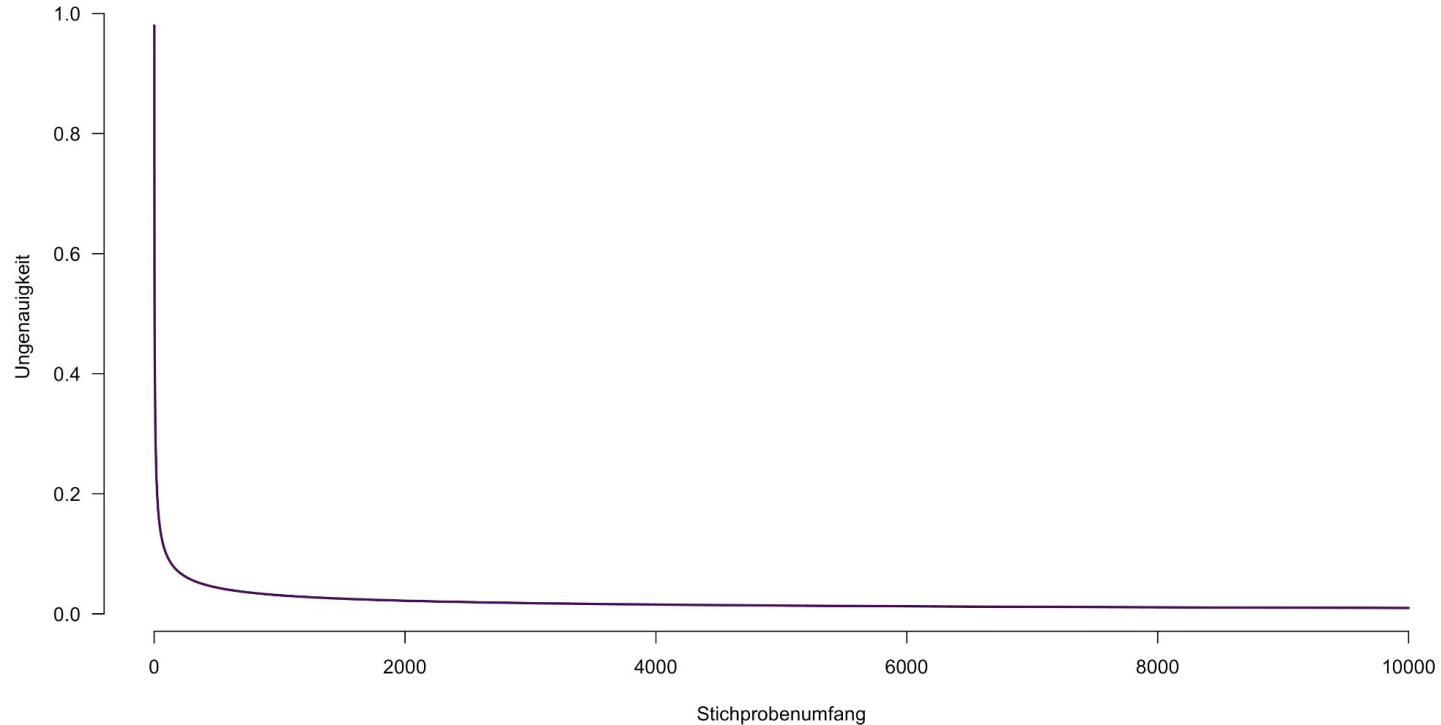
Alle Akten haben dieselbe  
Auswahlwahrscheinlichkeit.

Ist eine einfache Zufallsstichprobe mit den  
Kriterien der Aussonderungsbekanntmachung  
Justiz kompatibel?



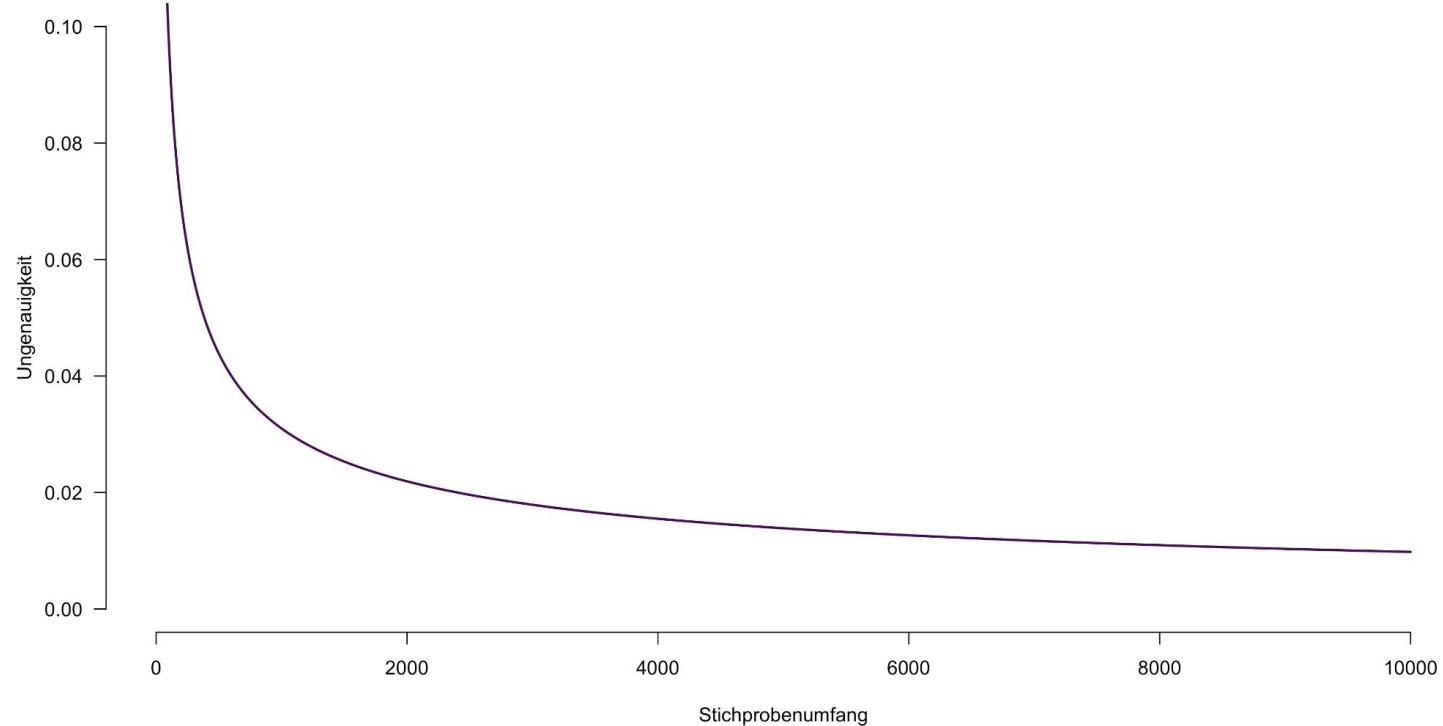
# Wie genau können wir Anteilswerte bestimmen?

Ungenauigkeit von Anteilswerten und Stichprobenumfang

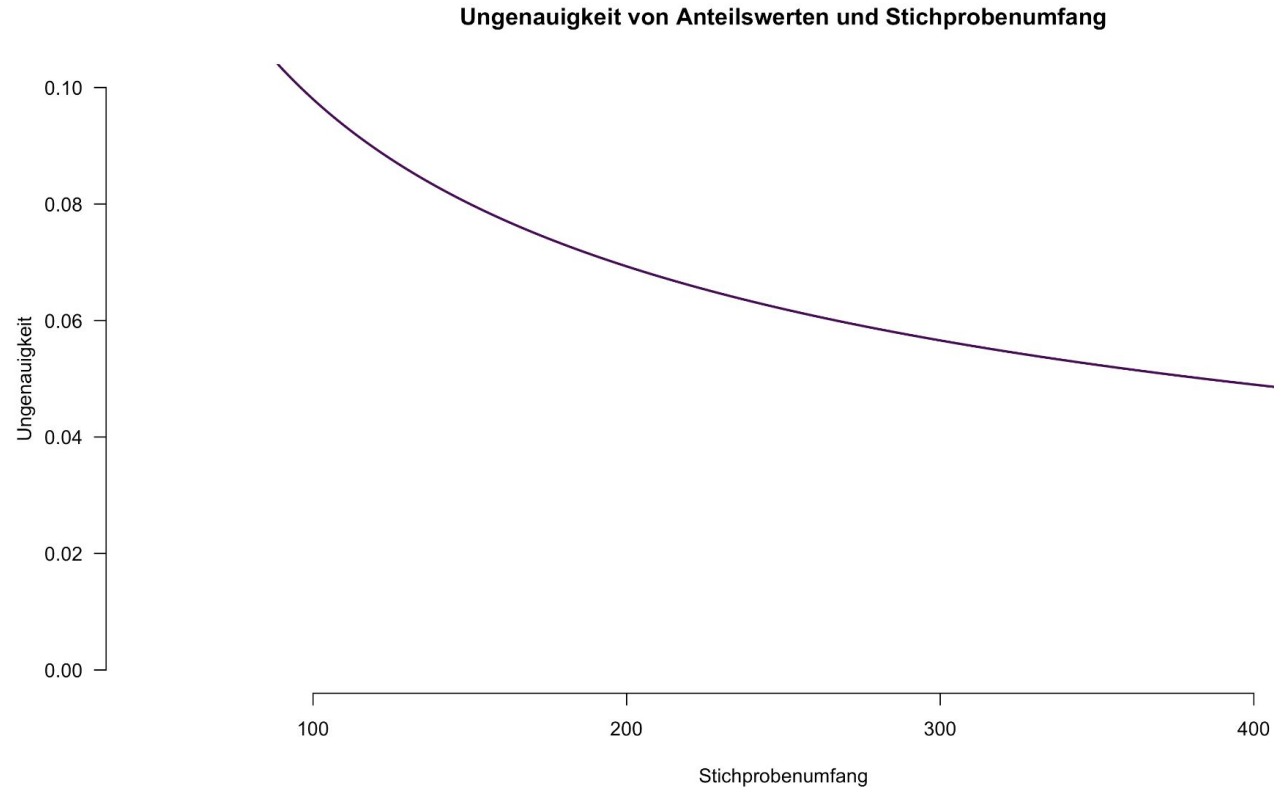


# Wie genau können wir Anteilswerte bestimmen?

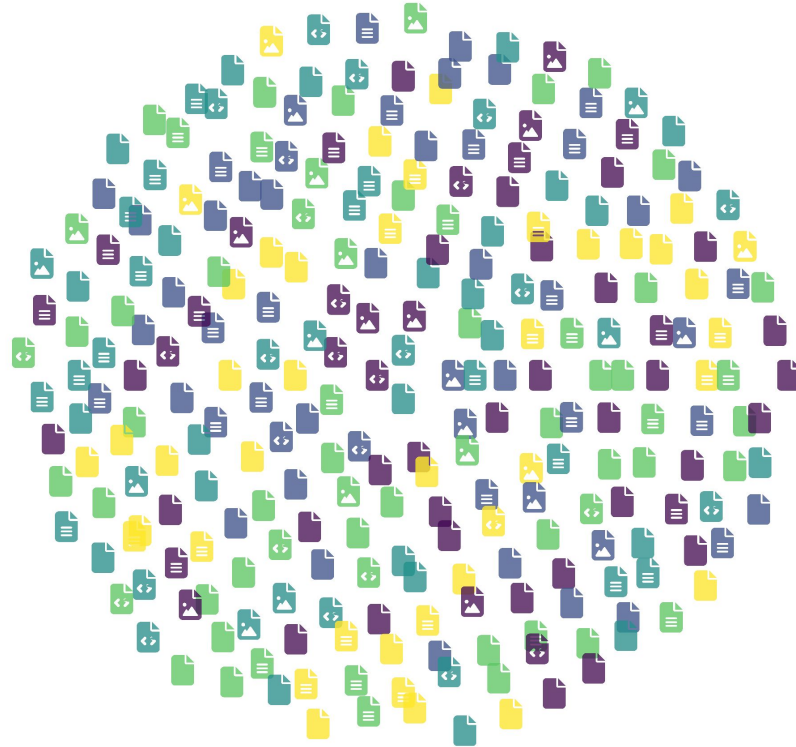
Ungenauigkeit von Anteilswerten und Stichprobenumfang



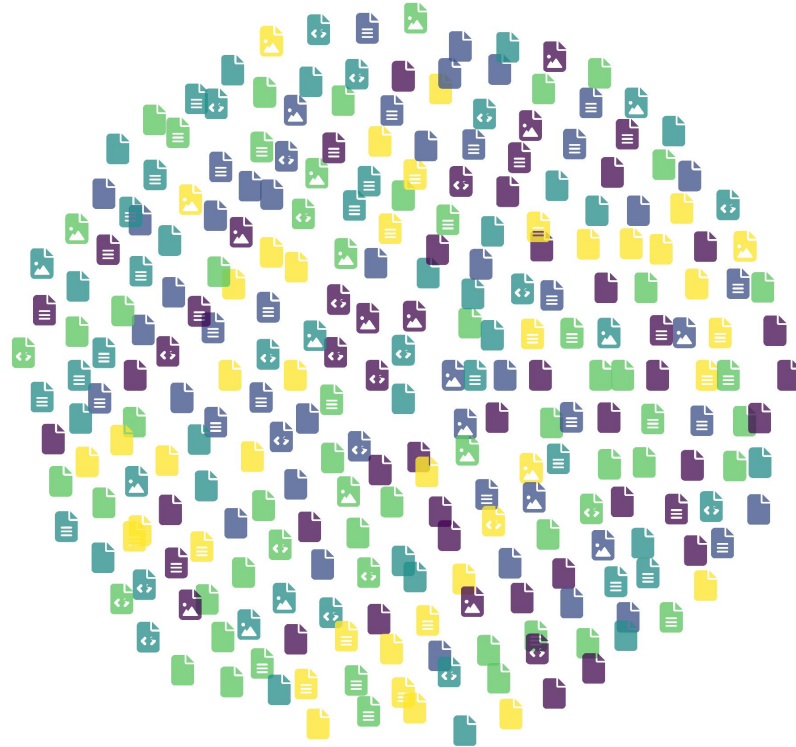
# Wie genau können wir Anteilswerte bestimmen?



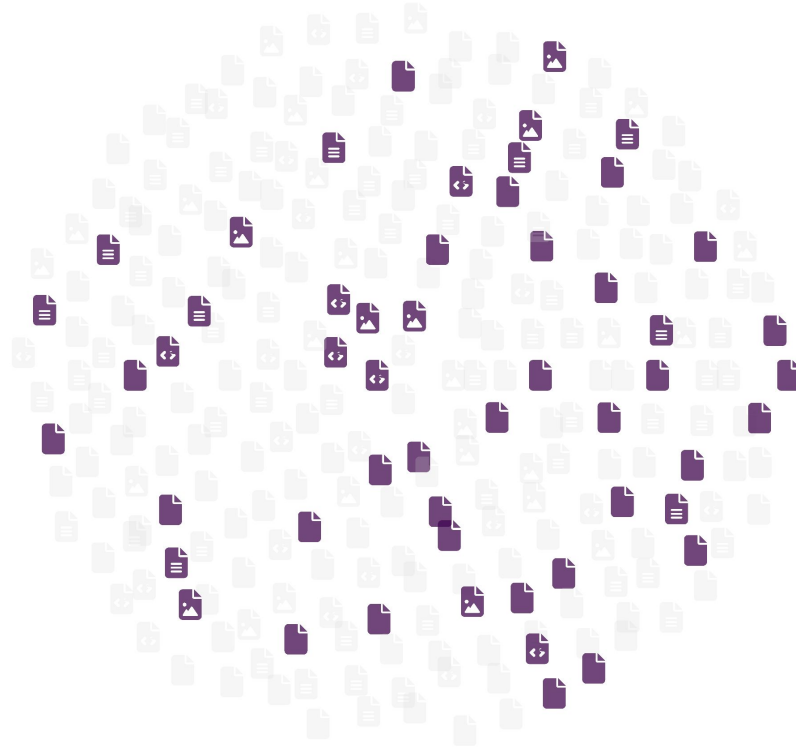
Mit geschichteten Zufallsstichproben können wir...



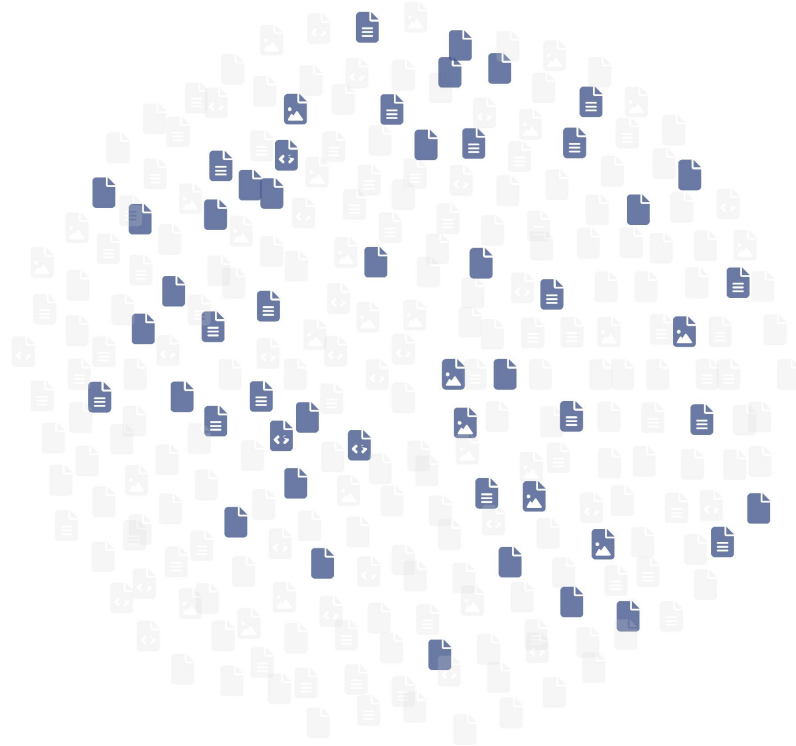
die Zufallsauswahl mit qualitativen Kriterien kombinieren



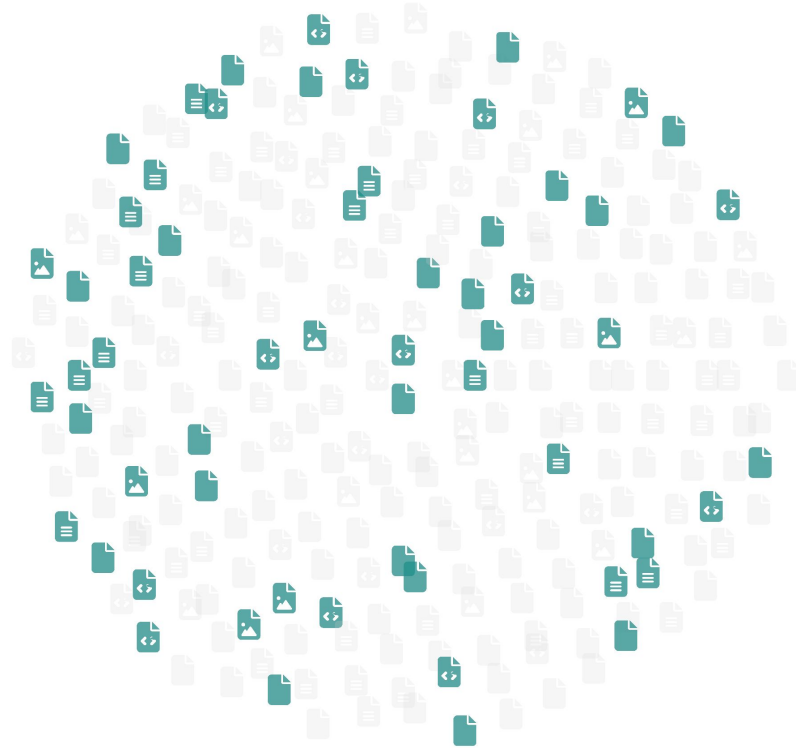
# Unsere Grundgesamtheit geschichtet nach Farben



# Unsere Grundgesamtheit geschichtet nach Farben

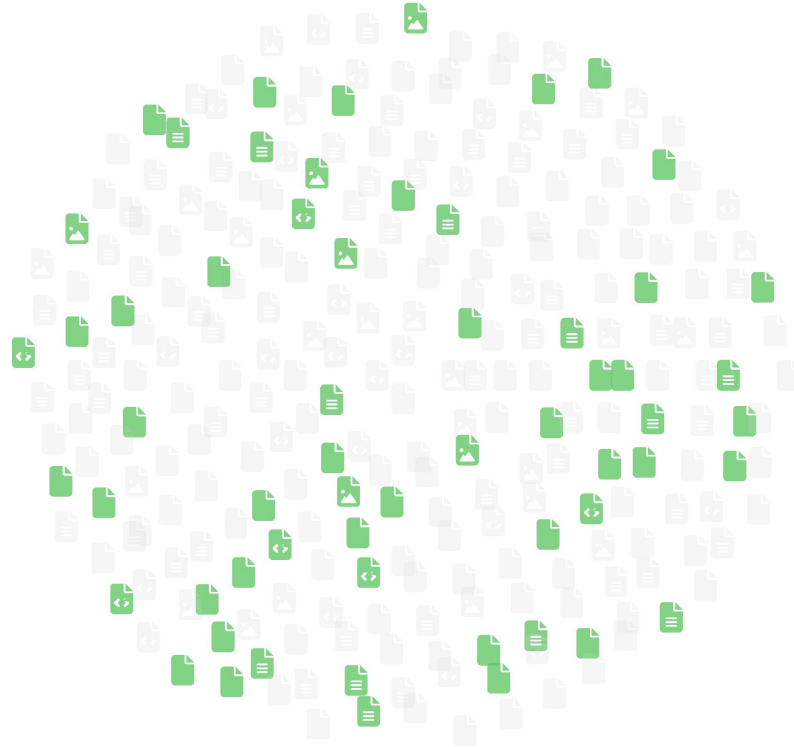


# Unsere Grundgesamtheit geschichtet nach Farben

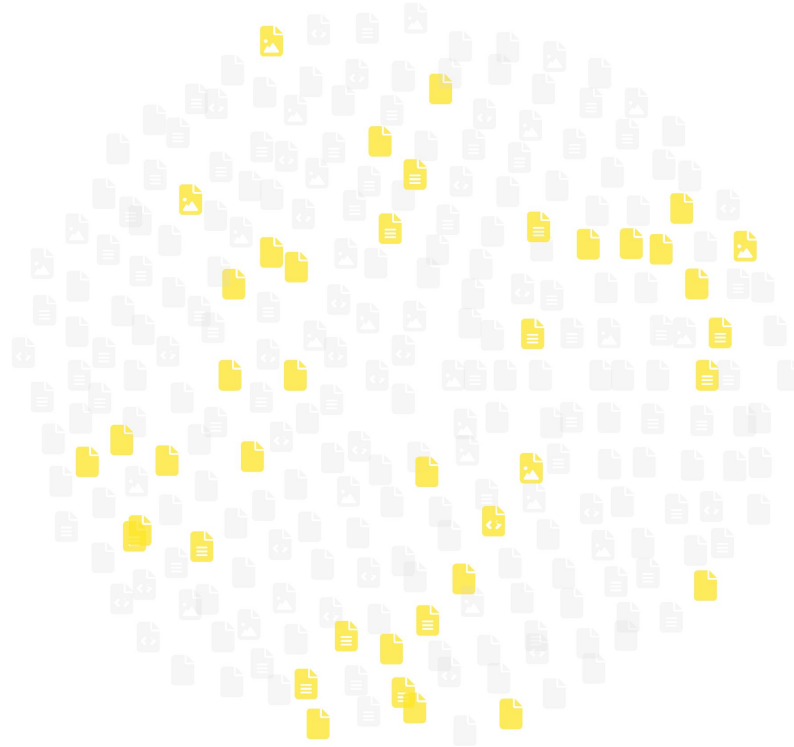




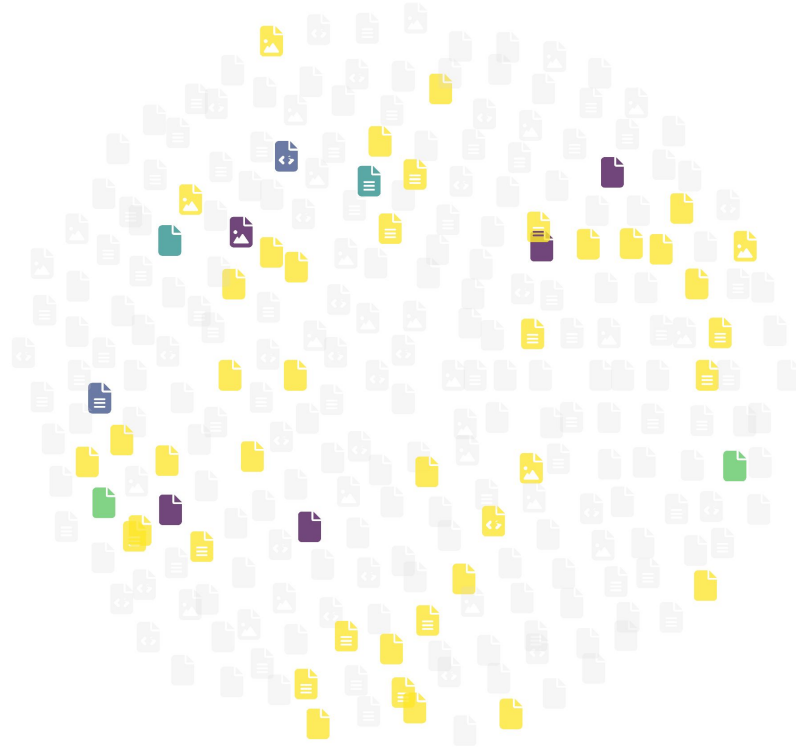
# Unsere Grundgesamtheit geschichtet nach Farben



# Unsere Grundgesamtheit geschichtet nach Farben

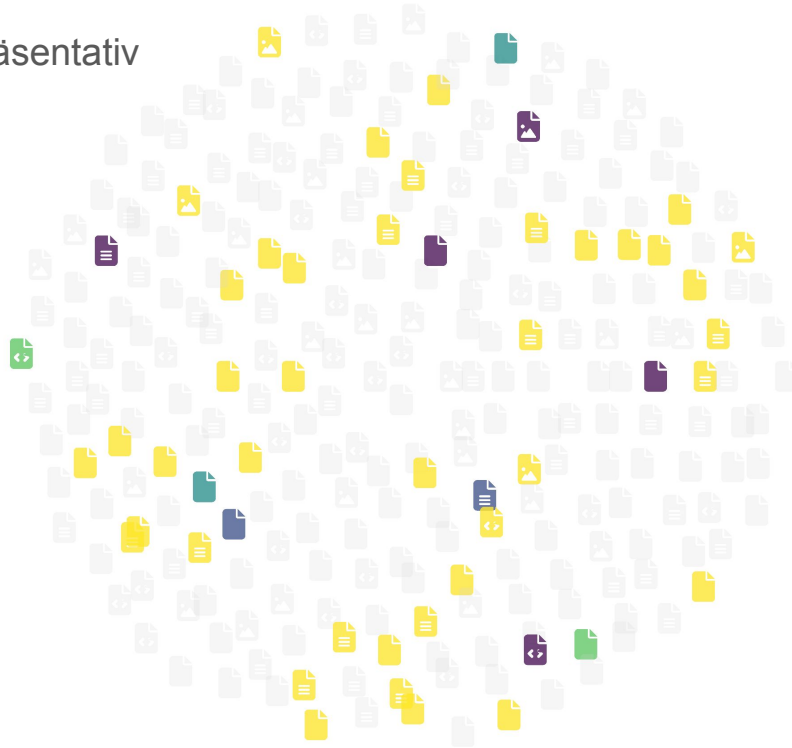


Wir wählen alle gelben Akten, fünf lila Akten und je zwei aus den anderen Schichten



# Eine weitere geschichtete Zufallsstichprobe

Sind diese Stichproben repräsentativ für die Grundgesamtheit?

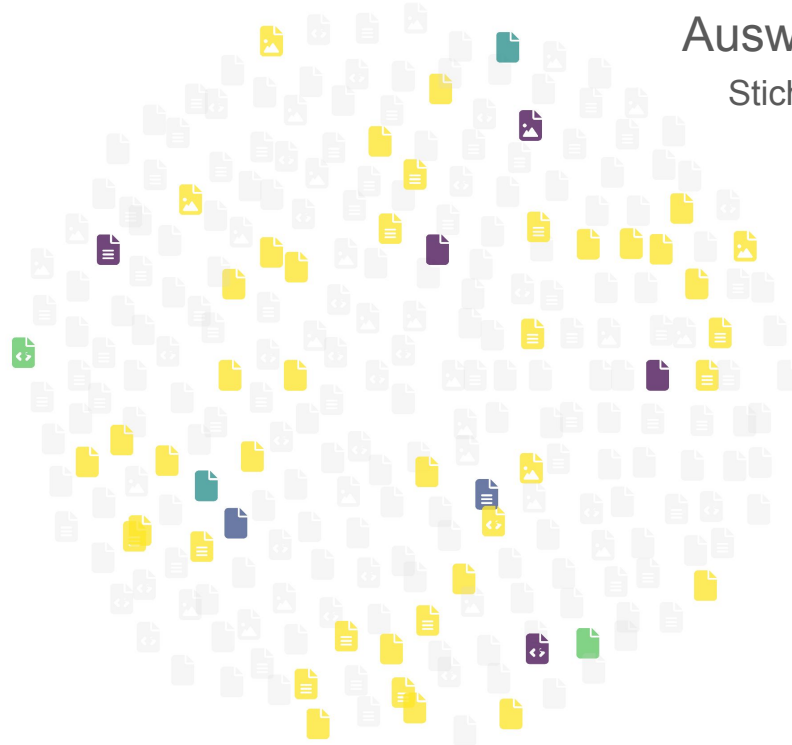


# Bei Transparenz zur Auswahl können wir hochrechnen

Transparenz über:

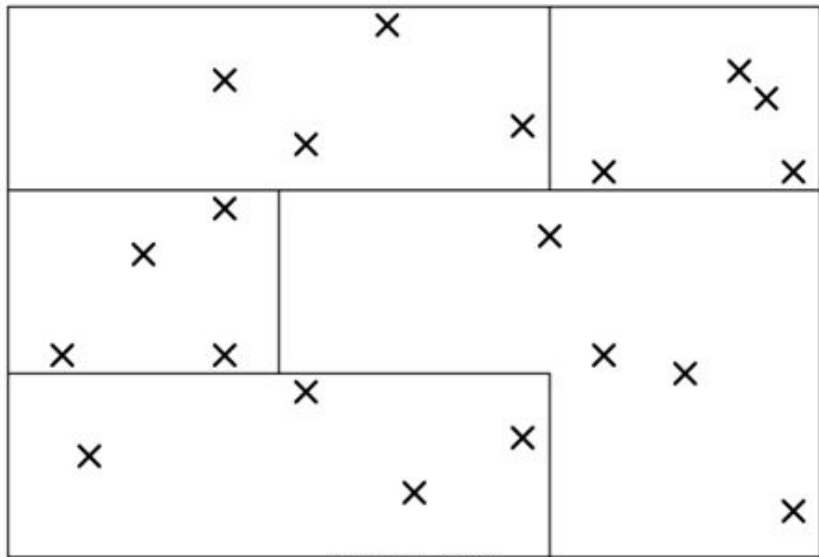
Größe der Schichten

Stichprobengröße pro  
Schicht

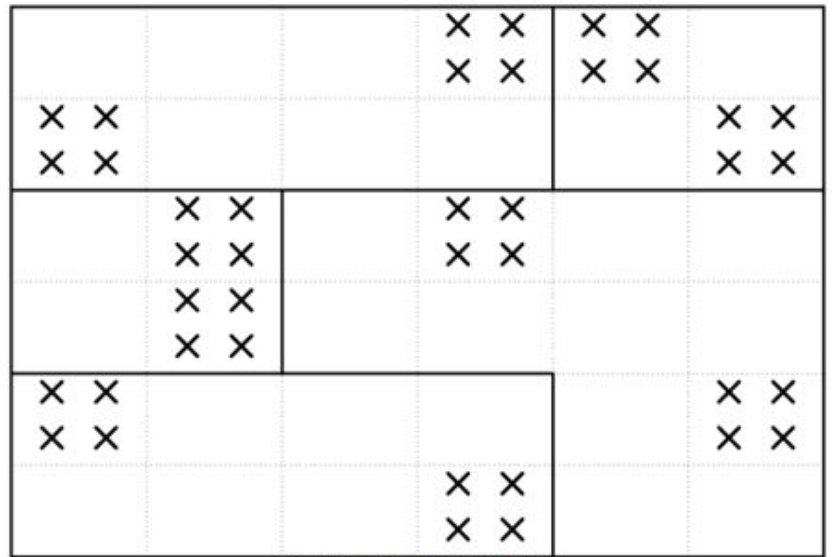


Auswahlwahrscheinlichkeit =  
Stichprobengröße/Größe der Schicht

Hochrechnungsfaktor =  
 $1/\text{Auswahlwahrscheinlichkeit}$

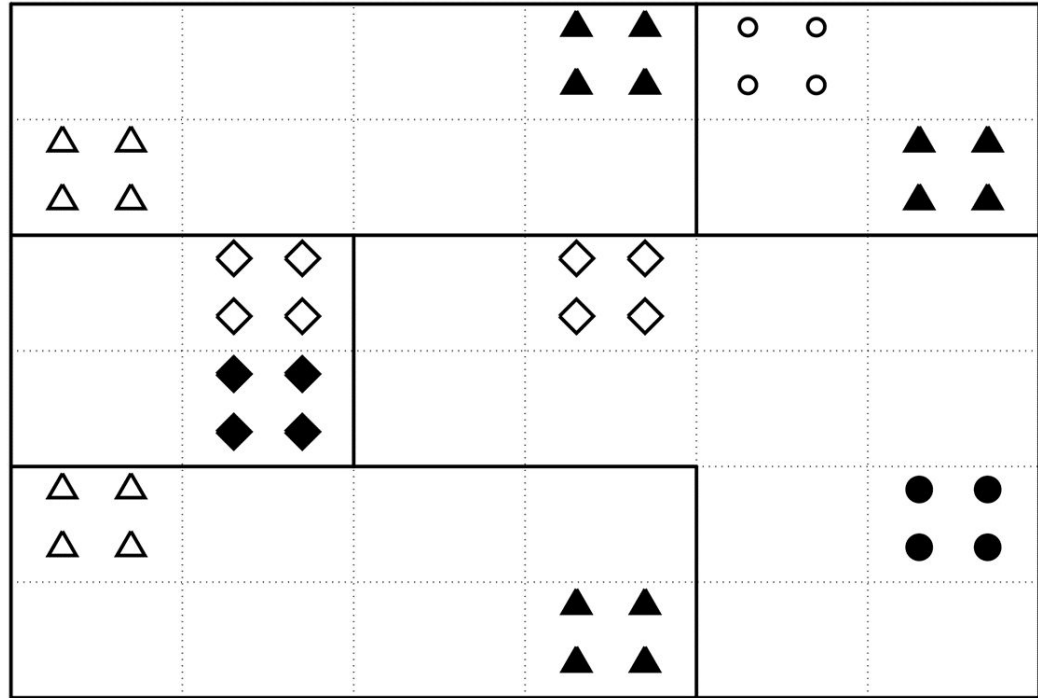


Stratified sample



Cluster sample within strata

Wie viele von wo?



Very homogeneous cluster

# Ergebnisse

## Archivierungsmodell für Zivilgerichtsverfahren der Amtsgerichte

- Erste Schicht: Verteilung der Gewichtung der Stichprobenziehung nach Anzahl der Termine (in längeren Verfahren erfahren wir mehr über den Lebenssachverhalt)
  - Auswahlwahrscheinlichkeit 1 bei 20 oder mehr Terminen – da übernehmen wir alle Verfahren (insgesamt wohl nicht mehr als 10% der Stichprobengröße)
  - Auswahlwahrscheinlichkeit bei 10 bis 20 Terminen – sollen überproportional vertreten sein (mind. 20% der Stichprobengröße)
  - Null bis 10 Terminen – Rest der Stichprobengröße



# Ergebnisse

## Archivierungsmodell für Zivilgerichtsverfahren der Amtsgerichte

- Zweite Schicht: Verteilung der Stichprobenziehung über Sachgebietsgruppen (gleiche Auswahlwahrscheinlichkeit für jede Gruppe)
  - **Verkehrsunfallsachen** (sehr viele Verfahren)
  - **Körper und Person** (Eingriff in die körperliche Unversehrtheit, besonders schwerwiegende Fälle)
    - Arzthaftungssachen
    - Körperverletzung
  - **Leben und Wohnen** (Zusammenleben)
    - Nachbarschaftssachen
    - Wohnungsmietsachen
    - Sonstige Mietsachen
    - Wohnungseigentumssachen
  - **Wirtschaftsrecht**
    - Bau-/Architektensachen (ohne Architektenhonorarsachen)
    - Kaufsachen
    - Reisevertragssachen
    - Kredit-/Leasingsachen
    - Ansprüche aus Versicherungsverträgen (ohne Verkehrsunfallsachen)
    - Gesellschaftsrechtliche Streitigkeiten
    - Honorarforderungen von Personen, für die eine besondere Honorarordnung gilt
  - **Sonstiges**
    - Schuldrechtsanpassungs- und Bodenrechtssachen der neuen Länder (ganz wenig)
    - Sonstiger Verfahrensgegenstand

# Teamarbeit 5:

→ Freie Gruppenarbeit

- Haben Sie ähnliche Probleme in Ihrer Arbeit?
- Womit könnten die Studierenden helfen?
- Alternativ: Stratifizierte Zufallsstichprobe in R



Ausblick:

Was kann man mit Data Science  
noch machen?

Beispiel: Record Linkage

# Ziel: Bekannte Persönlichkeiten zur Archivierung identifizieren

Anzubieten sind:

“Akten über Verfahren, an denen bekannte Persönlichkeiten des öffentlichen Lebens (Politiker, Wissenschaftler, Künstler usw.), bedeutende Familien, Stiftungen, Firmen oder sonstige Unternehmen beteiligt waren”

(Aus Artikel 10.2.2. der Aussonderungsbekanntmachung der Justiz)

# Idee: Zusätzliche Datenquellen nutzen

## Verwaltungsdaten

- Namen von Bürgermeistern
- Firmeninhaber aus Handelsregister
- Registerdaten aus Behörden

## Externe Daten (Internet)

- Zeitungen & Nachrichten
- Wikipedia
- Wikidata



```

1 SELECT ?person ?personLabel ?birthDate ?description
2 WHERE {
3   ?person wdt:P27 wd:Q183; # German citizenship
4     wdt:P569 ?birthDate. # Birth date property
5   FILTER (YEAR(?birthDate) > 1985). # Filter for birth after 1950
6
7   OPTIONAL {
8     ?person schema:description ?description. # Description property
9     FILTER (LANG(?description) = "de"). # Filter for English or German language description
10  }
11 SERVICE wikibase:label { bd:serviceParam wikibase:language "de". }
12 }

```

Table

26949 Ergebnisse in 23922 ms

&lt;/&gt; Code

Herunterladen

Link

Search

person	personLabel	birthDate	description
<a href="#">wd:Q888991</a>	Lena Malkus	6. August 1993	deutsche Weitspringerin
<a href="#">wd:Q89097</a>	Carmen Klaschka	8. Januar 1987	deutsche Tennisspielerin
<a href="#">wd:Q89128</a>	Laura Siegemund	4. März 1988	deutsche Tennisspielerin

# Was ist Wikidata?

Datenbank enthält 107 Millionen Datenobjekte

- 6 Millionen Objekte sind “instance of ‘human’”
- ~122 Tausend Personen mit Geburtsdatum nach 1950 und deutscher Staatsangehörigkeit (nachfolgend verwendet)

## Wozu dient Wikidata?

- Zentralisierter Datenspeicher für alle Wikipedia-Projekte
- Infoboxen und Listen in Wikipedia können aus Wikidata gefüllt & aktualisiert werden
- Daten dürfen für beliebige Zwecke genutzt werden

## Wer fügt Inhalte ein?

- Jede und Jeder

# Record Linkage: Verschiedene Datenquellen zusammenführen

Variablen aus der Justiz-Datenbank:

Anrede	Titel	Name	Vorname	Rufname	Geburtsname	weitere Namen, Künstlernamen, Ordensnamen, Hausnamen, frühere Namen etc.	Geburts-/Gründungsdatum	Geburtsland	Sterbe-/Löschdatum	Staatsangehörigkeit
Herr	NA	Müller	Thomas	Thomas	NA	NA	NA	NA	NA	NA

Variablen aus Wikidata:

person	personLabel	birthDate	description	birthName	birthPlace	alsoKnownAs
<a href="http://www.wikidata.org/entity/Q104178">http://www.wikidata.org/entity/Q104178</a>	Thomas Müller-Pering	1958-04-22	deutscher Gitarrist, Professor an der HfM „Franz Liszt“ Weimar	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	Müller-Pering
<a href="http://www.wikidata.org/entity/Q2426226">http://www.wikidata.org/entity/Q2426226</a>	Thomas Müller	1981-11-05	deutscher Schauspieler und Musiker	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	Tom Verhagen
<a href="http://www.wikidata.org/entity/Q688535">http://www.wikidata.org/entity/Q688535</a>	Thomas Müller	1961-03-05	deutscher Nordischer Kombinierer	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA
<a href="http://www.wikidata.org/entity/Q1440749">http://www.wikidata.org/entity/Q1440749</a>	Thomas Müller	1939-01-12	deutscher Komponist	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA
<a href="http://www.wikidata.org/entity/Q2426220">http://www.wikidata.org/entity/Q2426220</a>	Thomas Müller	1953-01-16	deutscher Physiker	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA
<a href="http://www.wikidata.org/entity/Q11897405">http://www.wikidata.org/entity/Q11897405</a>	Thomas Müller	1983-12-01	deutscher Sportschütze	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA
<a href="http://www.wikidata.org/entity/Q17326576">http://www.wikidata.org/entity/Q17326576</a>	Thomas Müller	1958-01-01	deutscher Militärhistoriker und Konservator	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA
<a href="http://www.wikidata.org/entity/Q20428259">http://www.wikidata.org/entity/Q20428259</a>	Thomas Müller	1966-04-07	deutscher Judoka	NA	<a href="http://www.wikidata.org/entity/Q8157228">http://www.wikidata.org/entity/Q8157228</a>	NA



# Exaktes Matching

1. Variablen standardisieren
2. Match = Perfekte Übereinstimmung auf allen Variablen

Datenbank 1:

<b>first_word</b>	<b>last_word</b>
Ben	Braun
Thomas	Müller

Datenbank 2:

<b>first_word</b>	<b>last_word</b>
Helmut	Fischer
Tomas	Müller
Thomas	Müller

# Exaktes Match: Ergebnisse mit unseren Daten

- People-Tabelle hat 349.492 Prozessbeteiligte
  - Ausschluss von 118.993 Unternehmen (identifiziert anhand des fehlenden Vornamens)
- 230.499 Personen bleiben zur weiteren Analyse
  - Wenn eine Person an mehreren Verfahren beteiligt ist, kann das mit den vorliegenden Daten nicht erkannt werden. Solche Namen werden mehrfach gezählt
- 14.713 Personen (6,4%) haben mindestens einen Namensvetter in Wikidata
  - 7064 unterschiedliche Namen
  - Beteiligung an 14.713 Akten (11,6%)

Diskussion: Wie bewerten Sie diese Ergebnisse?

# Datenprobleme in den Akten

## Duplikate

- Vor- und Nachname identisch bei 48912 Einträgen in people
  - Handelt es sich um unterschiedliche Personen oder war dieselbe Person in mehreren Verfahren beteiligt?
- Auch in Wikidata treten 5338 Namen mehrfach auf
- Verfügbare Variablen erlauben oft keine eindeutige Identifikation einzelner Personen

## Fehlende Werte

## Schreibweisen

# Maschinelles Lernen für Record Linkage

**TABLE 1. An Illustrative Example of Agreement Patterns.**

	Name				Address	
	First	Middle	Last	Date of birth	House	Street
<b>Data set <math>\mathcal{A}</math></b>						
1	James	V	Smith	12-12-1927	780	Devereux St.
2	Robert	NA	Martines	01-15-1942	60	16th St.
<b>Data set <math>\mathcal{B}</math></b>						
1	Michael	F	Martinez	02-03-1956	4	16th St.
2	James	D	Smithson	12-12-1927	780	Dvereux St.
<b>Agreement patterns</b>						
$\mathcal{A}.1 - \mathcal{B}.1$	Different	Different	Different	Different	Different	Different
$\mathcal{A}.1 - \mathcal{B}.2$	Identical	Different	Similar	Identical	Identical	Similar
$\mathcal{A}.2 - \mathcal{B}.1$	Different	NA	Similar	Different	Different	Identical
$\mathcal{A}.2 - \mathcal{B}.2$	Different	NA	Different	Different	Different	Different

Beispiel von Enamorado et al. (2019) zum Informationsgehalt von Tippfehlern

Lösung: Machine Learning Modell, das Typos “zulässt”. Hier: FastLink

# Was ist eigentlich maschinelles Lernen?

Checkers Beispiel (Samuel 1959)

Definition:

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

(Mitchell 1997)

Ca. 2015-2017: AlphaGo schlägt verschiedene Großmeister

# Ergebnis

- Erneut 14713 Personen mit exaktem Match (Jaro-Winkler-Distanz = 1)
- Zusätzlich unterscheiden sich 2708 Nachnamen minimal (Jaro-Winkler-Distanz > 0.97), z.B.
  - Schmidt <> Schmid <> Schmied
  - Müller <> Müllner
  - Schulz <> Schulze
  - Hoffmann <> Hofmann
- Zusätzlich unterscheiden sich 2346 Vornamen minimal (Jaro-Winkler-Distanz > 0.97)

# Diskussion

Was wird benötigt, damit Record Linkage zur Identifikation berühmter Persönlichkeiten verwendet werden kann?

Könnten zusätzliche Variablen zum Linkage verwendet werden?

Könnten besser geeignete externe Daten zum Linkage verwendet werden?

Was ist Prominenz? Ist es überhaupt möglich die Identifikation von berühmten Persönlichkeiten zu automatisieren?

Denkanstoß/Vorschlag: “Prominenzindikator” für Auswahl bereitstellen

Andere Möglichkeiten? Z.B. Medieninteresse am Verfahren messen?

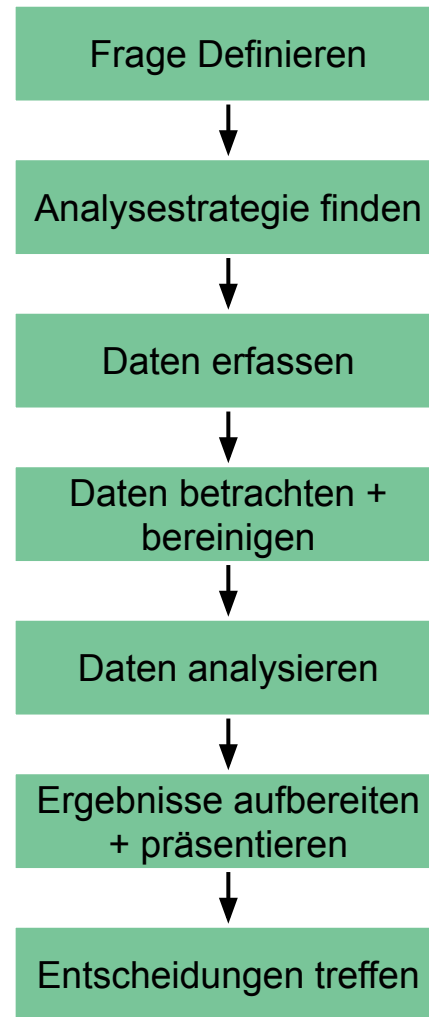
# Was kann man mit Data Science noch machen?

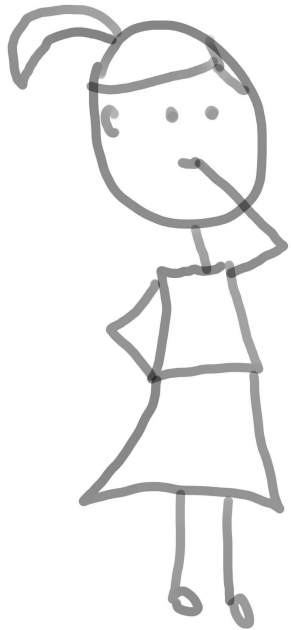
- Digitale Dokumente,
- Metadaten strukturieren mit LLMs,
- Reverse Image Search for Faces
- ...

Aber: Was dürfen Behörden und Archive? Was nicht?



# Vorgehen bei der Beantwortung von Fragen mit Daten





*Was war für Sie heute überraschend?*

# Morgen

Was bisher geschah + Visualisieren	10:00 - 10:30
Besuch des Digitalministers Dr. Fabian Mehring	10:30 - 11:15
Visualisieren und Vorbereitung Nachmittag	11:15 - 11:50
Mittagspause	11:50 - 12:50
Teamarbeit	12:50 - 15:10
Pause: Selbstbestimmt nach Bedarf der Teams	
Abschluss	15:10 - 15:30