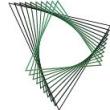


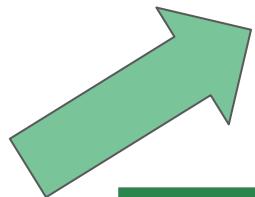
# Modul 2: Umgang mit Arbeitsumgebung, Software und Datenanalyse

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

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



**BERD**  
@NFDI



Bayerisches Staatsministerium  
für Digitales



Vortrag: Einführung	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Vortrag: Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit: Einfache Zufallsstichprobe	13:05 - 13:30
Pause	13:30 - 13:45
Vortrag: Stratifizierte Zufallsstichprobe	13:45 - 14:15
Teamarbeit: Stratifizierte Zufallsstichprobe	14:15 - 14:45
Vortrag: Abschluss	14:45 - 15:00

# Erste Schritte: Die Daten kennenlernen

Wie machen Sie das normalerweise?

# Erste Schritte: Die Daten kennenlernen

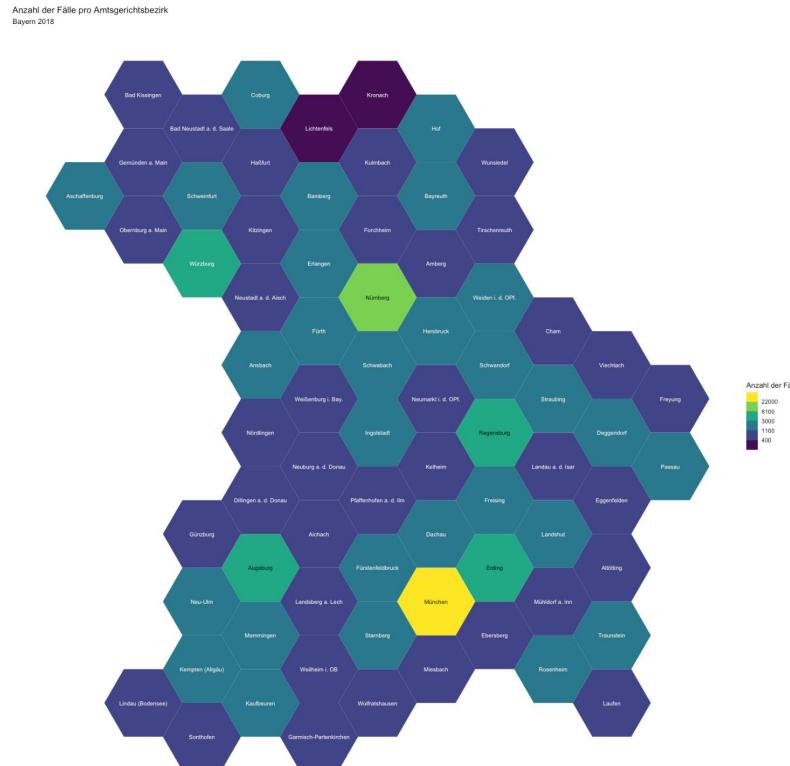
230817\_Abfrage\_Januar-Dezember.csv (read-only) - LibreOffice Calc

File Edit View Insert Format Styles Sheet Data Tools Window Help

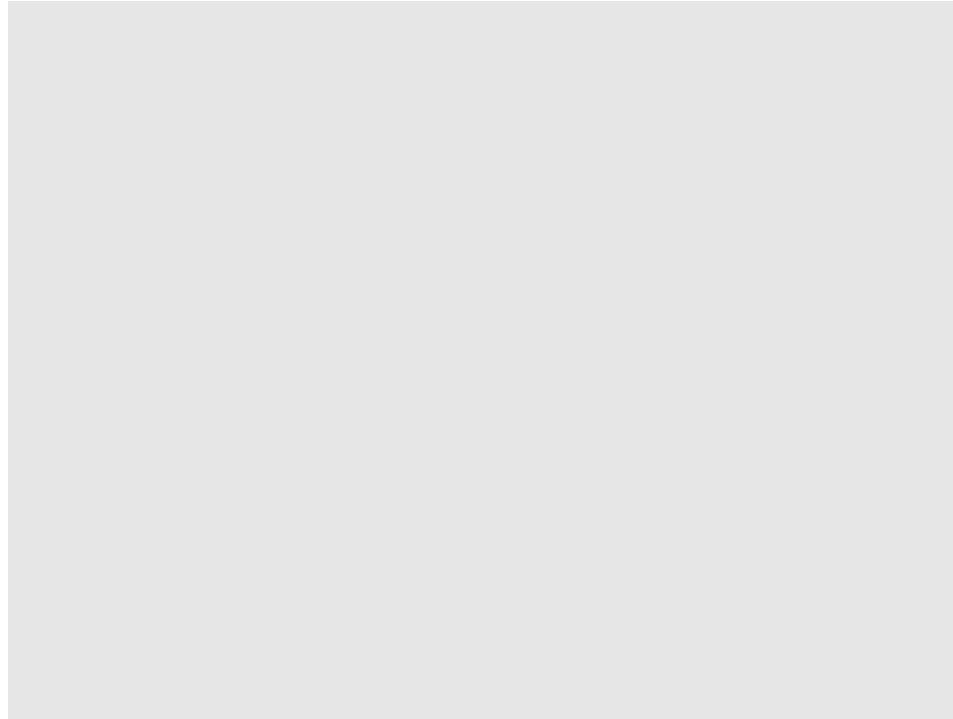
A1

	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	75e33dfe6630700	606.9	NA
5	Amtsgericht Aichach	101 C 16/18	weggelegt	75e33dfe6630700	606.9	NA
6	Amtsgericht Aichach	101 C 18/18	weggelegt	fa9d6c219cfe3a	390.61	NA
7	Amtsgericht Aichach	101 C 18/18	weggelegt	fa9d6c219cfe3a	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	bf90e98ad079c70	608.59	NA
13	Amtsgericht Aichach	101 C 2/18	weggelegt	bf90e98ad079c70	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	ea194871afcd662	100.56	NA
23	Amtsgericht Aichach	101 C 24/18	weggelegt	ea194871afcd662	100.56	NA
24	Amtsgericht Aichach	101 C 24/18	weggelegt	ea194871afcd662	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	6bdd5bc00a13528	197.35	NA
28	Amtsgericht Aichach	101 C 28/18	weggelegt	6bdd5bc00a13528	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	f93f7ab1dcf5dd	333.2	NA
36	Amtsgericht Aichach	101 C 34/18	weggelegt	f93f7ab1dcf5dd	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

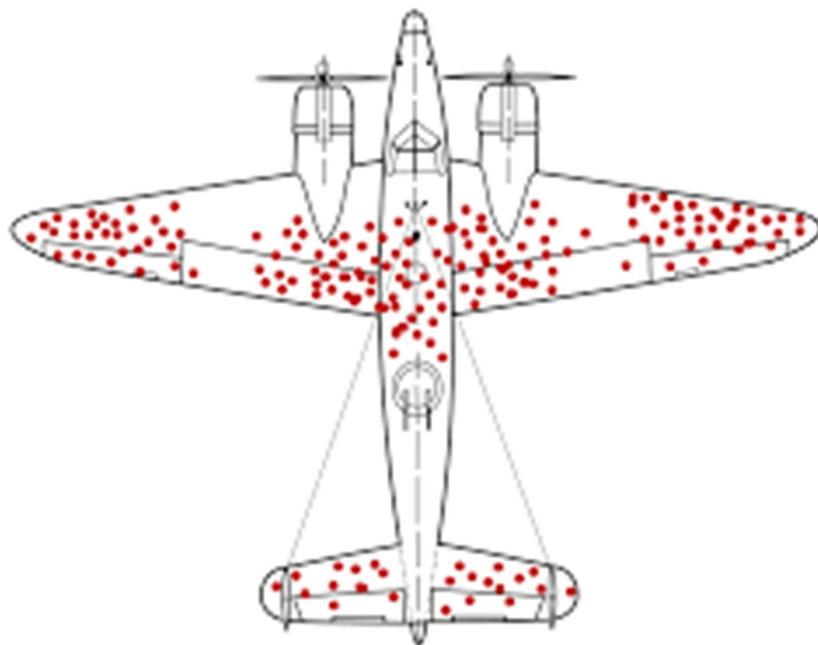
# Wie wir die Daten heute kennenlernen



# 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

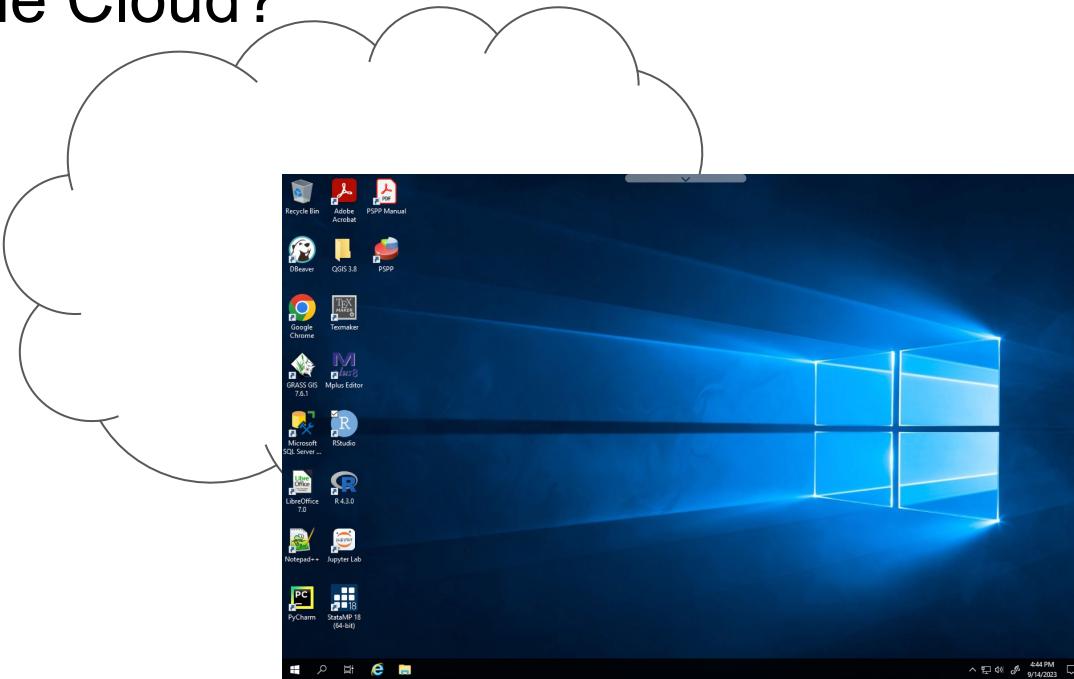
**Open Source (quelloffene) Software** ist Software, die von allen inspiziert, verändert und verbessert werden kann.

**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.

# Sichere Coleridge Cloud

**Administrative Data Research Facility**

**Democratizing in Data**

**5 Safes Framework**

- Projects
- People
- Settings
- Data
- Exports

**Data Stewardship**  
Manage Projects, People, Datasets, & Agreements

**Governance**  
Enterprise Data Catalog - Traditional Metadata Management with Rich Context Disclosure Review for Exports

**Transparency**  
Usage Based Pricing Model User Accessible Usage Statistics Flexible Performance Options

**Technology**  
Secure Remote Access, FedRAMP Authorized GovCloud

The diagram illustrates the architecture of the Coleridge Cloud. It features a central circular icon divided into four quadrants, each containing a blue icon: a bar chart (top), a briefcase (top-right), a microchip (bottom-left), and a database (bottom-right). A dashed line connects the center to five surrounding boxes: "Data Stewardship" (red), "Governance" (blue), "Transparency" (orange), "Technology" (green), and "5 Safes Framework" (dark blue). Below the Technology box is a detailed network diagram showing the flow from user access through Okta and HTTPS to the "Administrative Data Research Facility" application, which then connects via HTTPS to an "AppStream Stack" (containing EC2 Auto-Scaling) and a "Secure Gateway". The stack also includes "Coleridge Initiative Applications" like "Data Stewardship" and "Informatica Extract Load Transform". Finally, the stack connects to various cloud services at the bottom: Databases, S3, File Server, and Network Firewall.



# RStudio in der Cloud

The screenshot shows the RStudio interface running in a cloud environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Project dropdown set to (None). The main workspace contains an 'Untitled1' R script file with one line of code ('1'). The Environment tab in the top right shows an empty Global Environment. The Files tab in the bottom right displays a file tree under 'Home' with the following contents:

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		

The Console tab at the bottom shows the R startup message:

```
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)
```

The 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 taskbar at the bottom shows the Windows Start button, a search icon, a file icon, the Internet Explorer icon, a folder icon, and the RStudio icon. The system tray shows the date and time as 4:49 PM on 9/14/2023.

# RStudio in der Cloud

The screenshot shows the RStudio interface running in a cloud environment. The interface is divided into several panes:

- Editor**: The top-left pane contains a code editor window titled "Untitled1".
- Umgebung**: The top-right pane displays the "Environment" tab, which shows an empty global environment.
- Konsole**: The bottom-left pane is the "Console" tab, showing the R command-line interface and its output. The output includes the R version information, copyright notice, and a welcome message about the software being free and coming with no warranty.
- Dateien**: The bottom-right pane is the "Files" tab, showing a file tree with the contents of the current directory, including files like "Rhistory", "desktop.ini", "Downloads", "Mplus", "R", "SQL Server Management Studio", and "Visual Studio 2017".

The status bar at the bottom right indicates the time as 4:49 PM and the date as 9/14/2023.

```
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 more details.

R is a collaborative project.
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.
```

# Teamarbeit 3: Erste Analysen in R

In dieser Teamarbeit lernen wir zunächst Quarto und R in der Coleridge Cloud kennen.

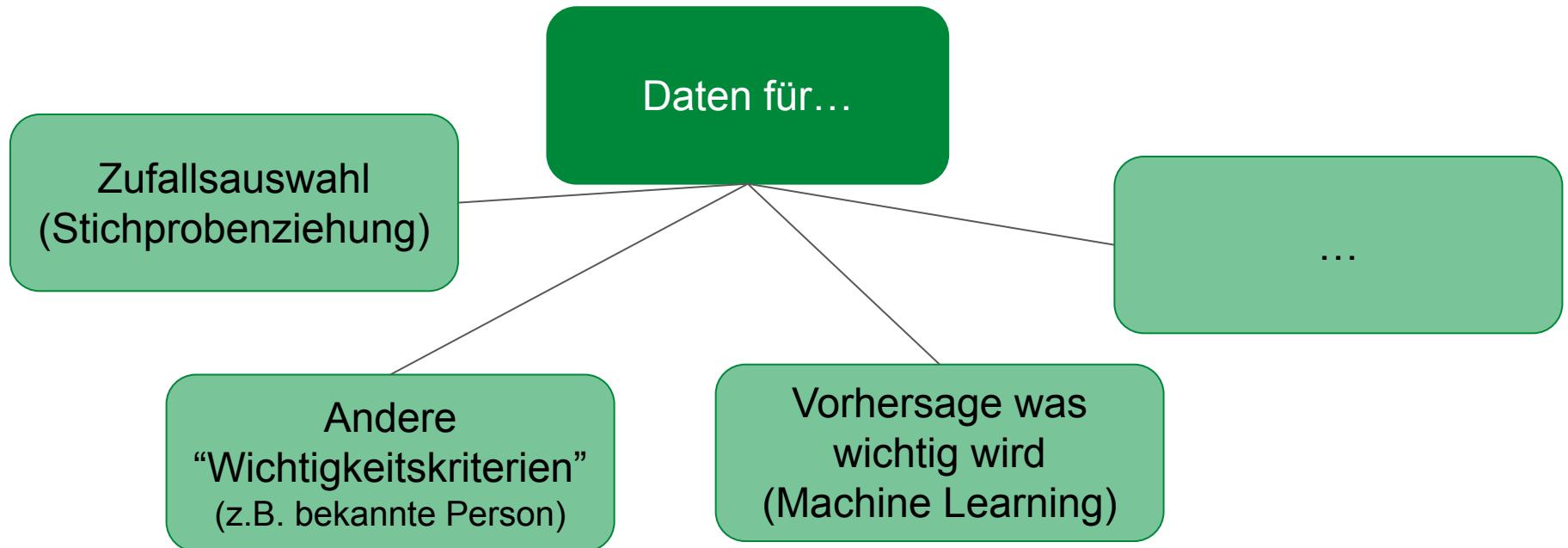
Anschließend führen wir gemeinsam erste Analysen mit den forumSTAR Daten durch.

Das Ziel dabei ist es die Daten gut zu verstehen.

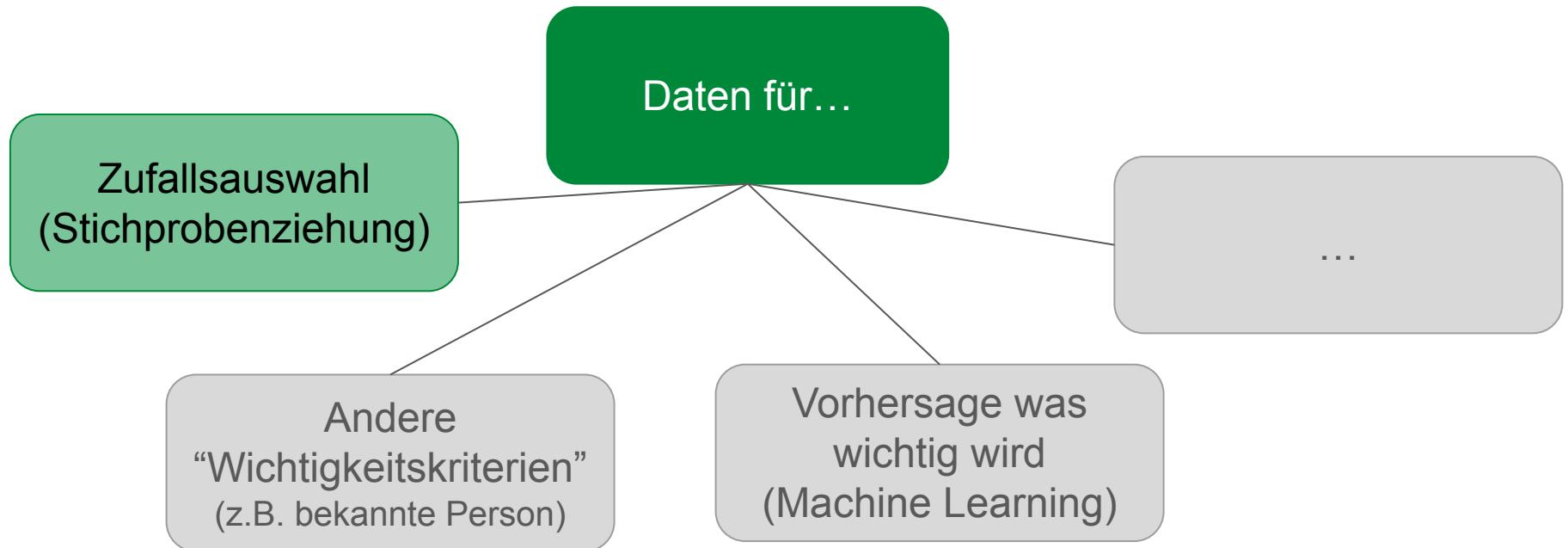


Vortrag: Einführung	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Vortrag: Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit: Einfache Zufallsstichprobe	13:05 - 13:30
Pause	13:30 - 13:45
Vortrag: Stratifizierte Zufallsstichprobe	13:45 - 14:15
Teamarbeit: Stratifizierte Zufallsstichprobe	14:15 - 14:45
Vortrag: Abschluss	14:45 - 15:00

# 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.

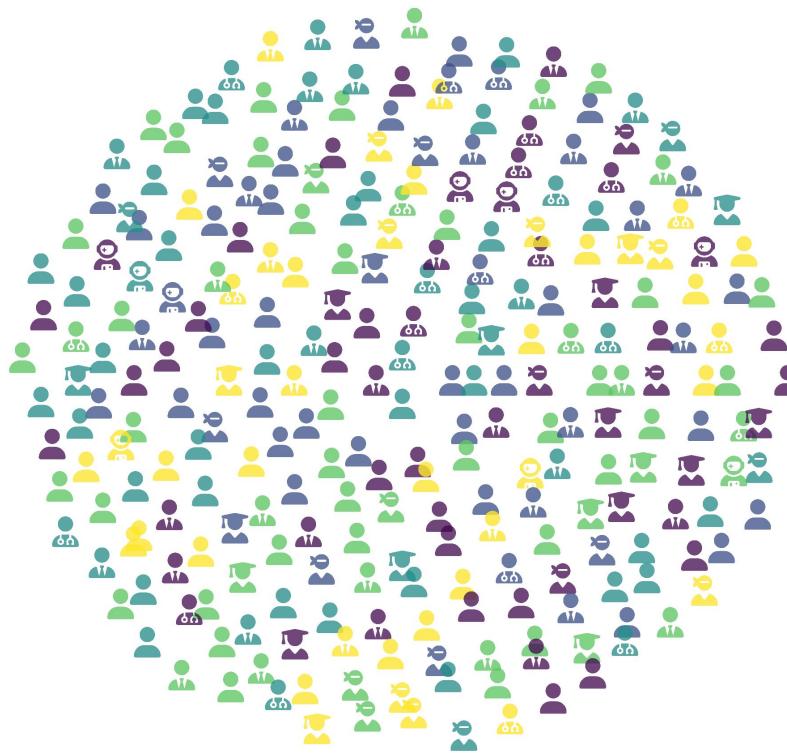


Wir kennen das Problem auch aus der Corona Zeit ...

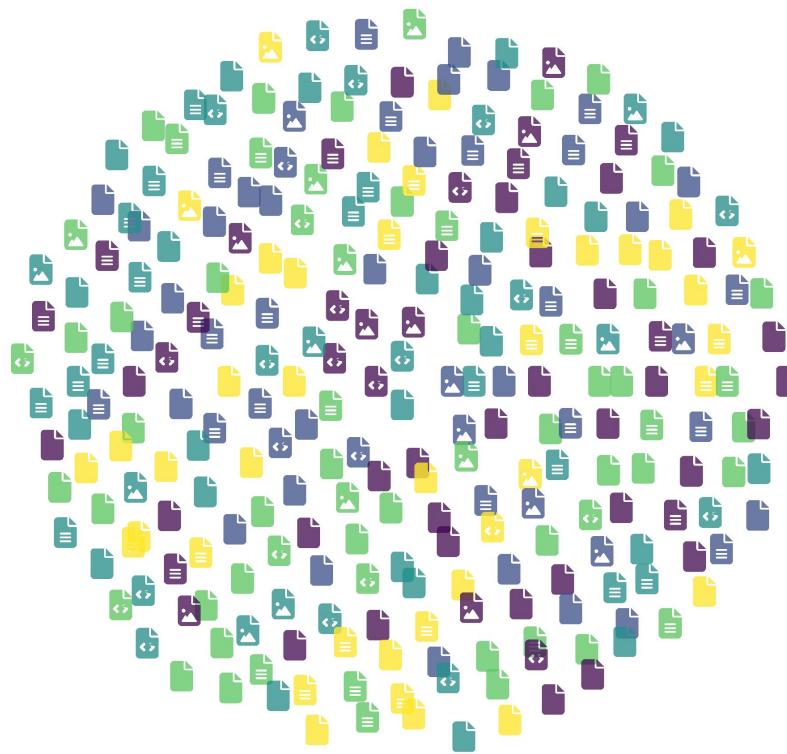


# Die Grundgesamtheit ist die Menge aller Personen...

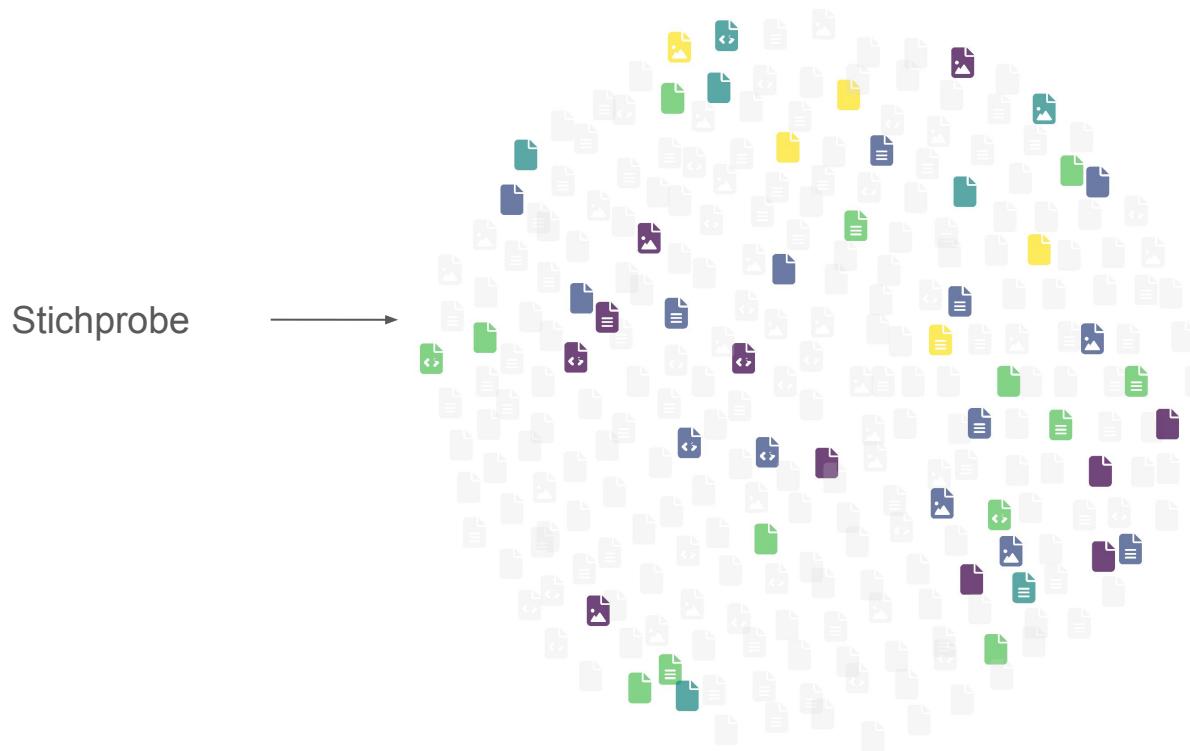
Grundgesamtheit →



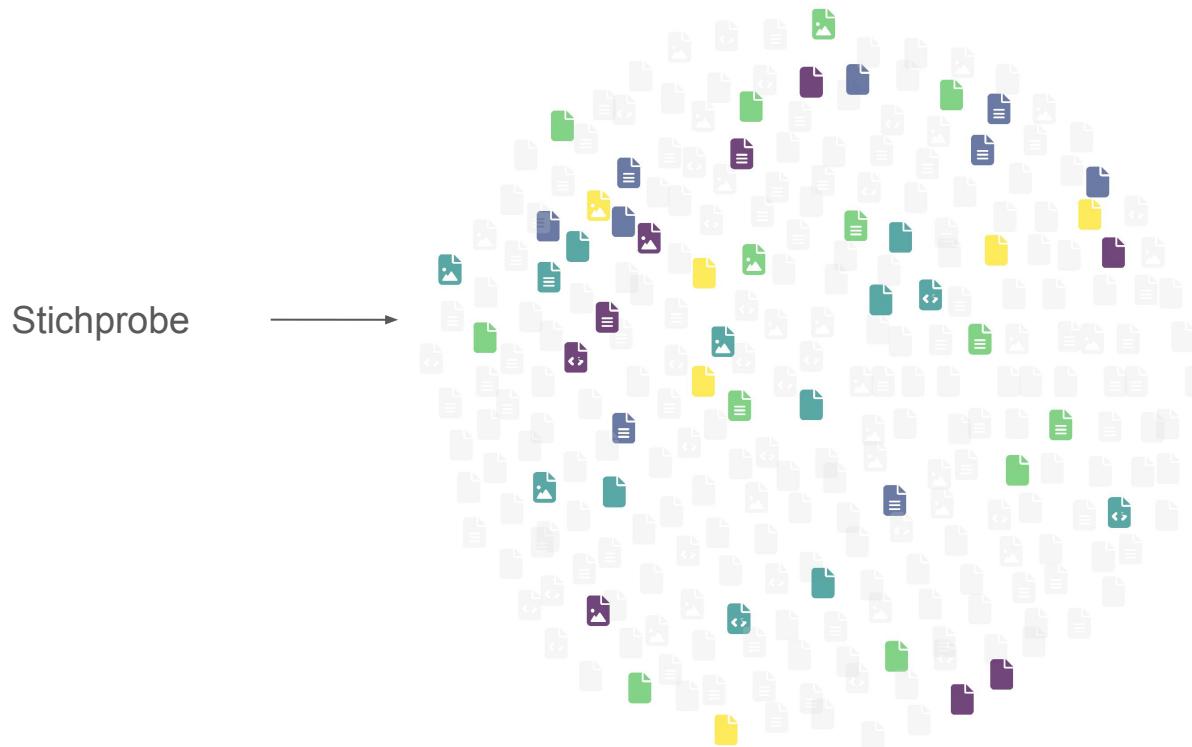
... oder Akten über die wir eine Aussage treffen wollen



# Eine Stichprobe ist eine Teilmenge der Grundgesamtheit



# Zufallsstichproben unterscheiden sich bei Wiederholung



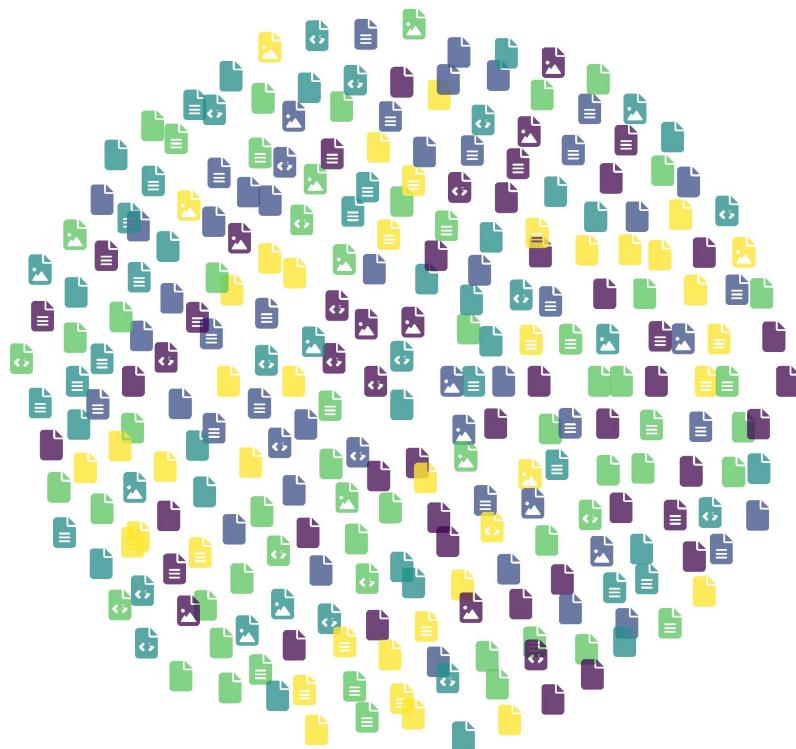
Wir nutzen die Stichprobe um Merkmale zu untersuchen

## Erste Stichprobe



## Zweite 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

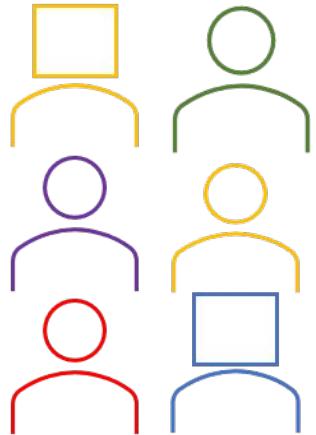
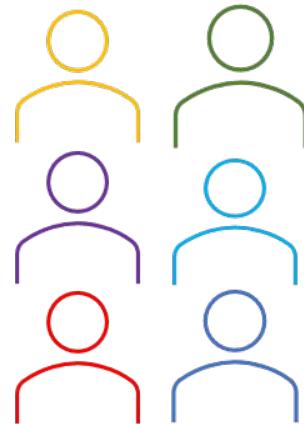
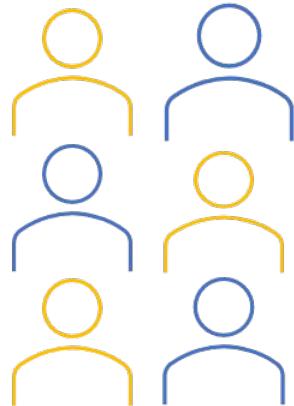
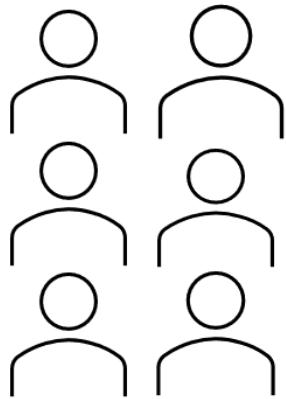


Systematische  
Zufallsauswahl



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](#) • Erstellt mit [Datawrapper](#)

“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."**

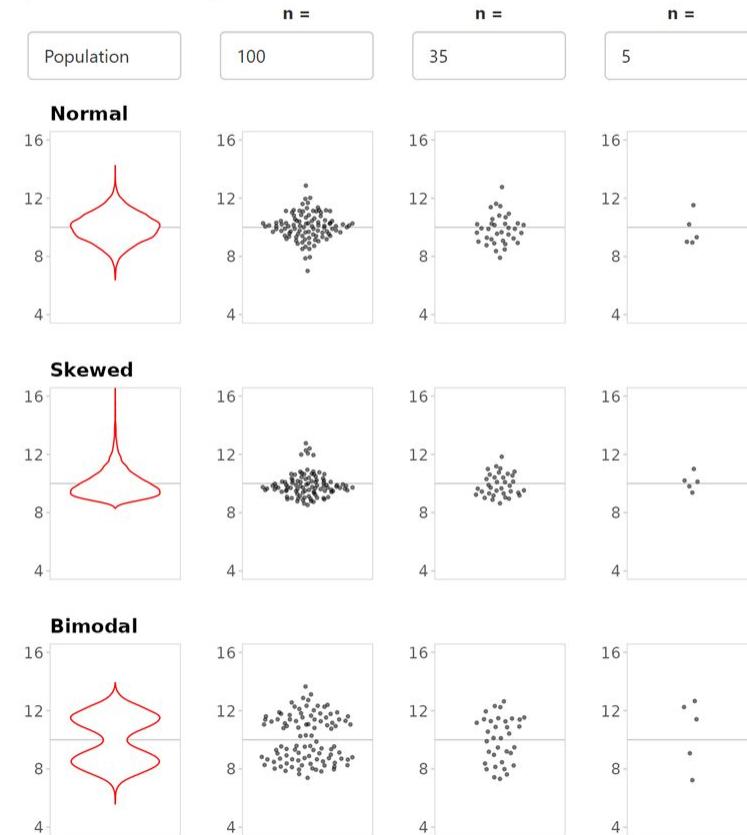
*Eraebnis der Umfrage "Spannungsfeld Männlichkeit" der Hilfsorganisation Plan International*

Show questions >>

# Verteilung der Streitwerte?

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

Draw a New Sample



# Teamarbeit 4: Einfache Stichprobenziehung

→ Gruppenarbeit

A large, green, five-pointed starburst shape with a white outline, positioned in the bottom-left corner of the slide.

Projekt

Vortrag: Einführung	10:00 - 10:30
Pause	10:30 - 10:40
Teamarbeit: Erste Analysen in R	10:40 - 12:00
Mittagspause	12:00 - 12:45
Vortrag: Einfache Zufallsstichprobe	12:45 - 13:05
Teamarbeit: Einfache Zufallsstichprobe	13:05 - 13:30
Pause	13:30 - 13:45
Vortrag: Stratifizierte Zufallsstichprobe	13:45 - 14:15
Teamarbeit: Stratifizierte Zufallsstichprobe	14:15 - 14:45
Vortrag: Abschluss	14:45 - 15:00

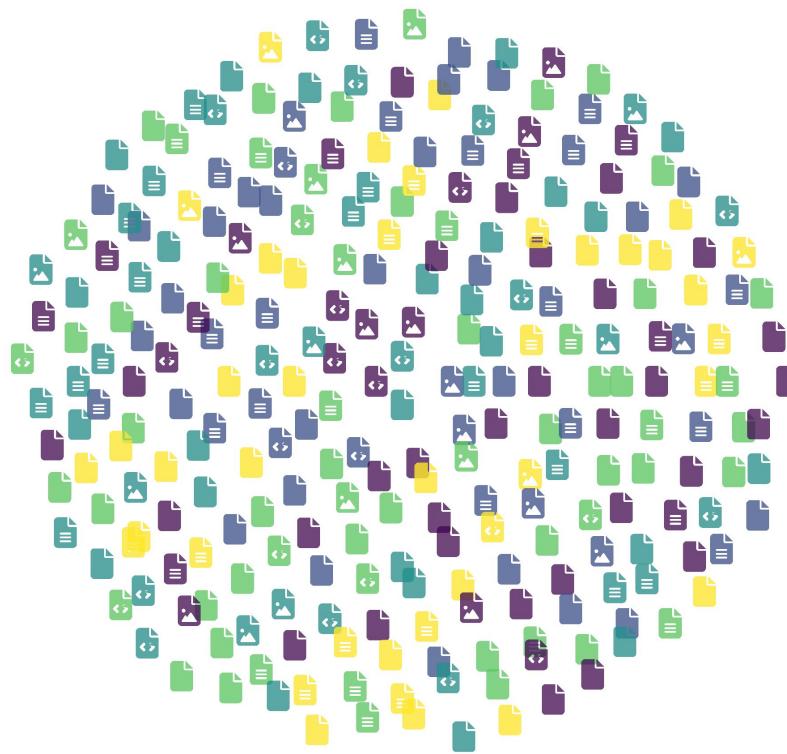
# Vor- und Nachteile der einfachen Zufallsstichprobe

Alle Akten haben dieselbe  
Auswahlwahrscheinlichkeit.

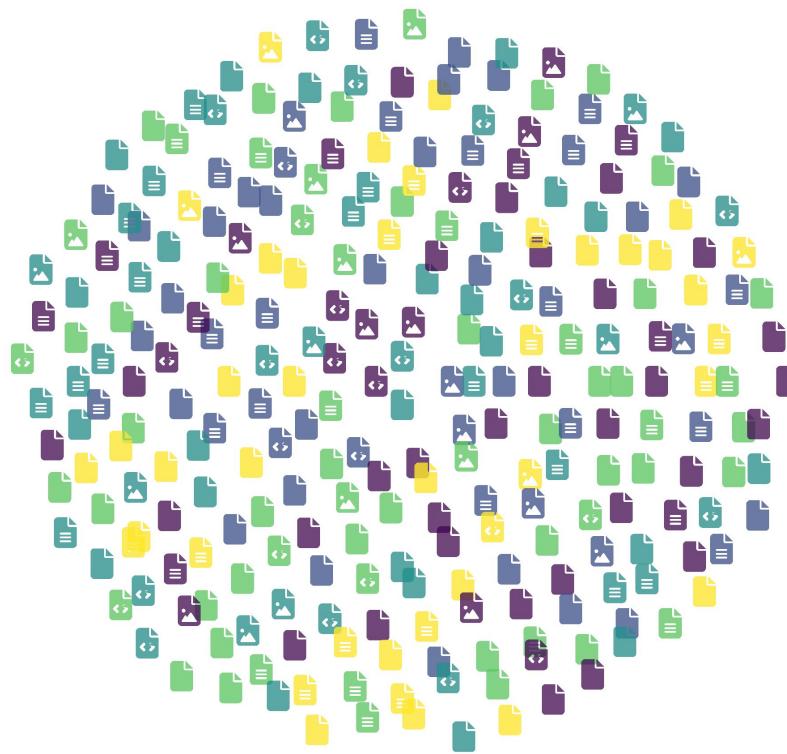
Alle Akten haben dieselbe  
Auswahlwahrscheinlichkeit.

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

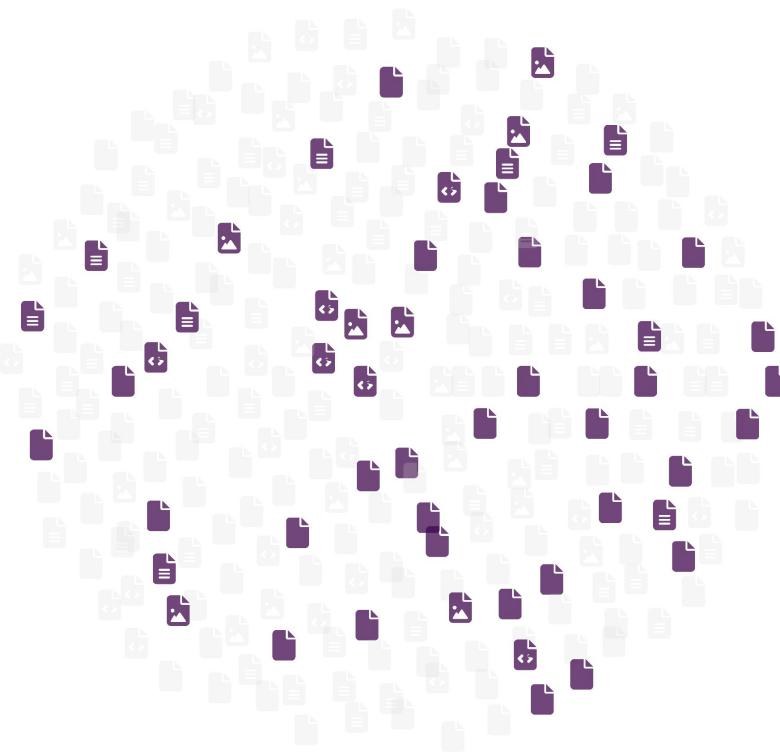
# Mit geschichteten Zufallsstichproben können wir...



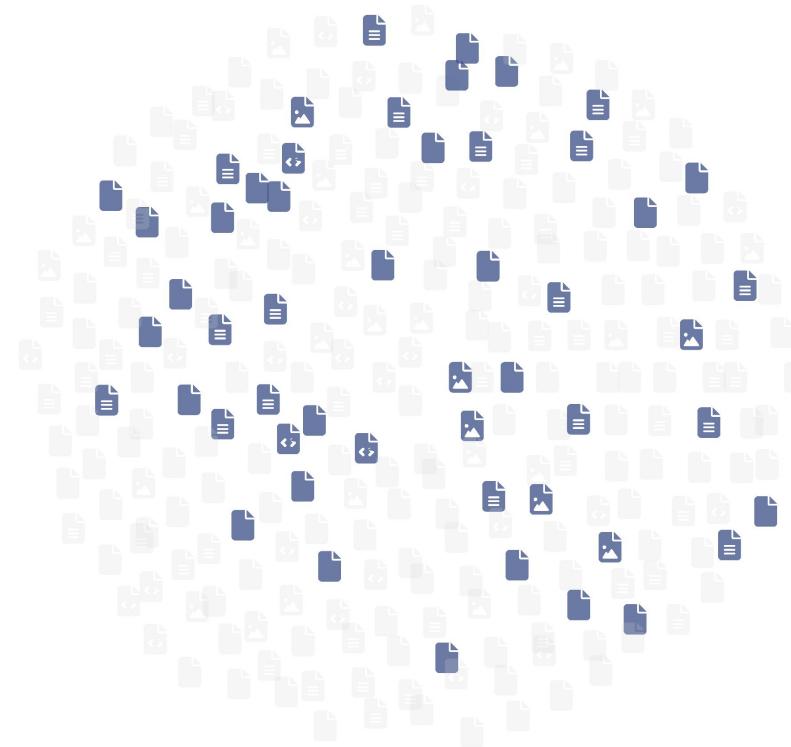
die Zufallsauswahl mit qualitativen Kriterien kombinieren



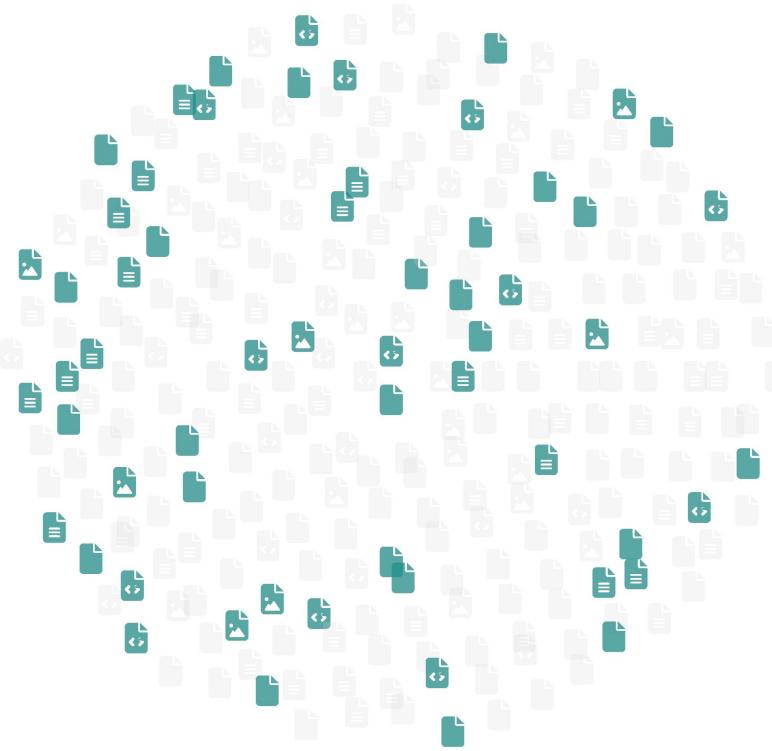
# Unsere Grundgesamtheit geschichtet nach Farben



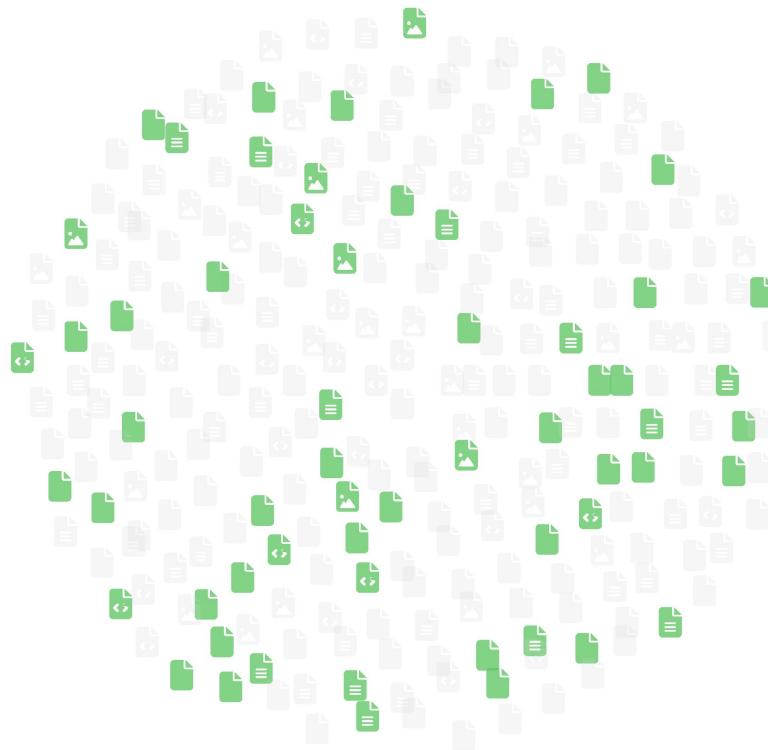
# Unsere Grundgesamtheit geschichtet nach Farben



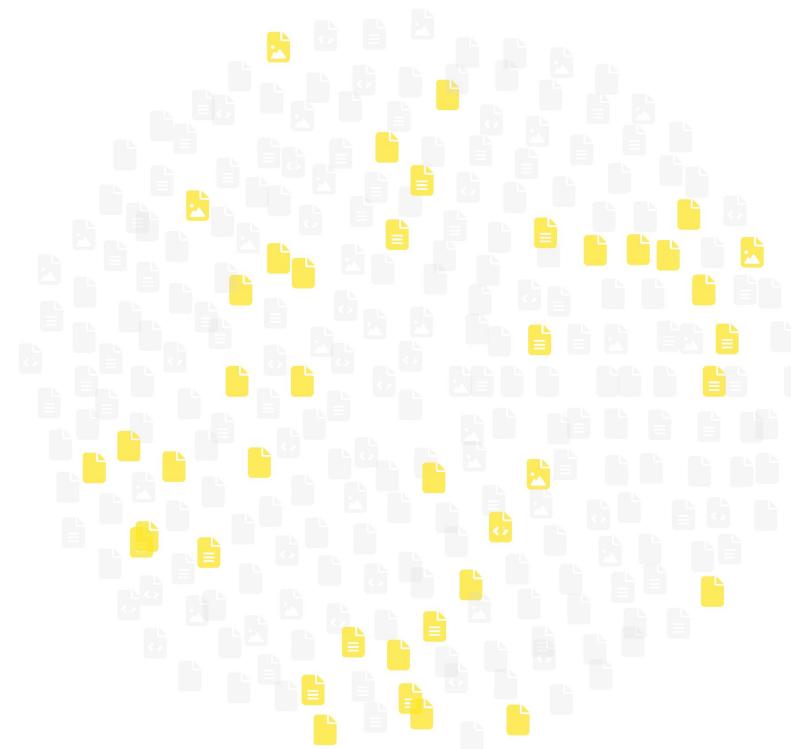
# Unsere Grundgesamtheit geschichtet nach Farben



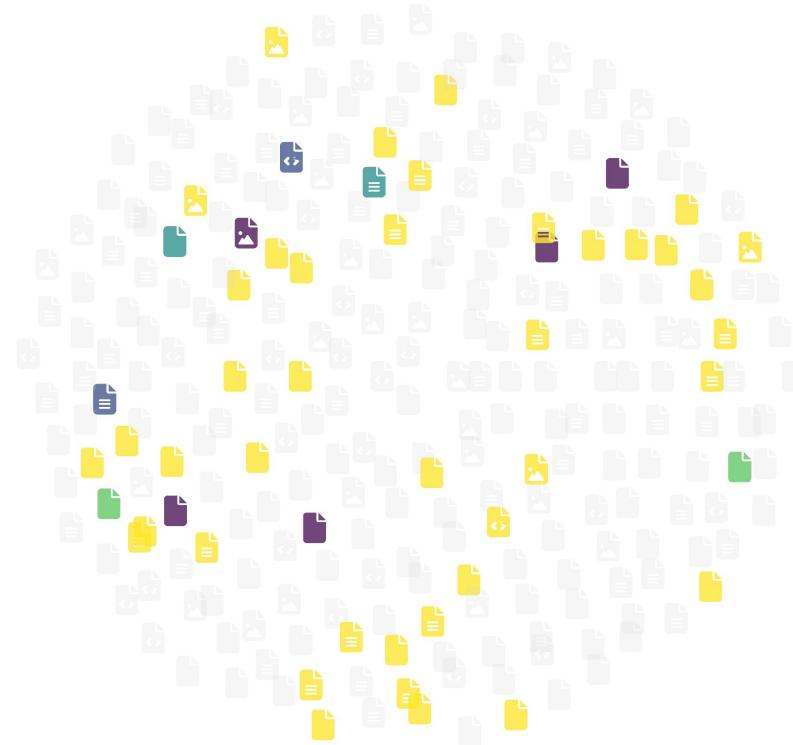
# Unsere Grundgesamtheit geschichtet nach Farben



# Unsere Grundgesamtheit geschichtet nach Farben

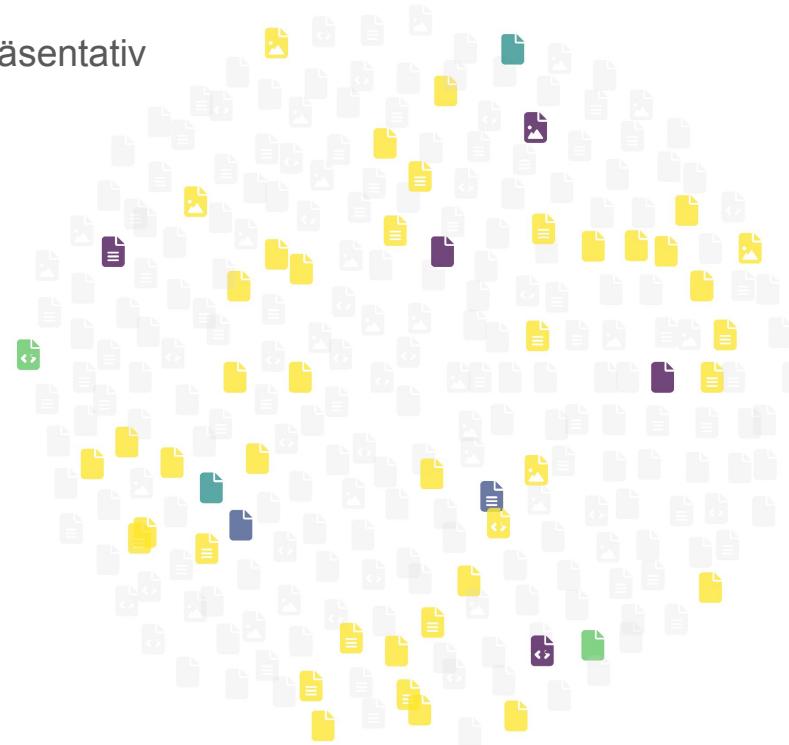


Wir wählen alle gelben Akten, fünf lila Akten und je zwei aus den anderen Schichten



# Eine weitere geschichtete Zufallsstichprobe

Sind diese Stichproben repräsentativ  
für die Grundgesamtheit?

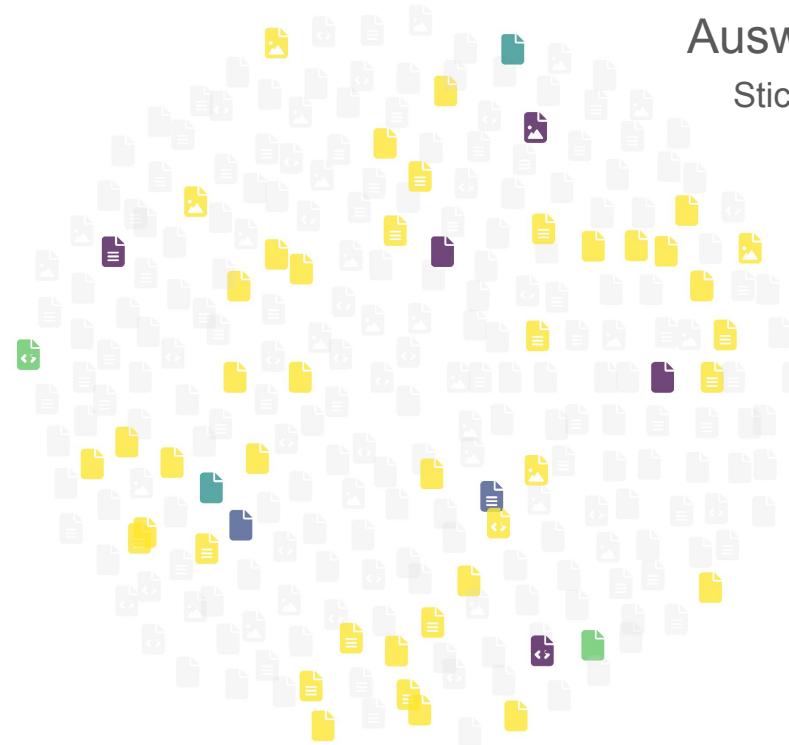


# Bei Transparenz zur Auswahl können wir hochrechnen

Transparenz über:

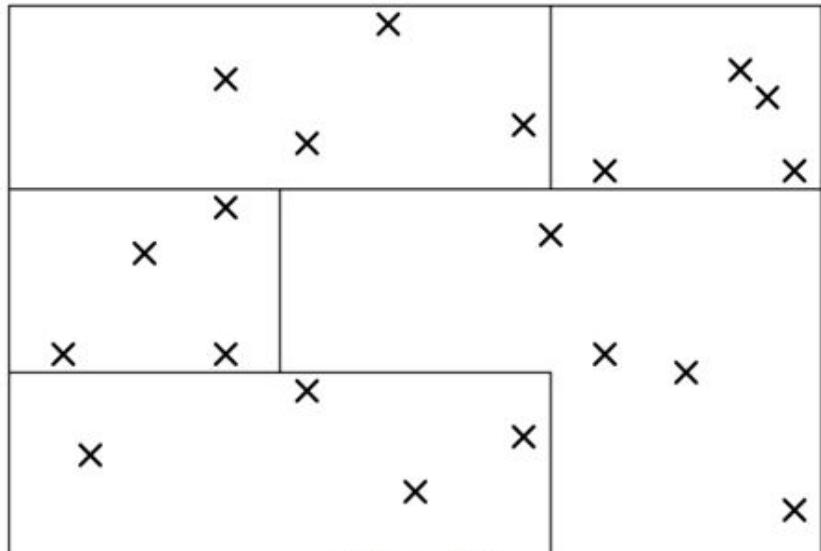
Größe der Schichten

Stichprobengröße pro  
Schicht

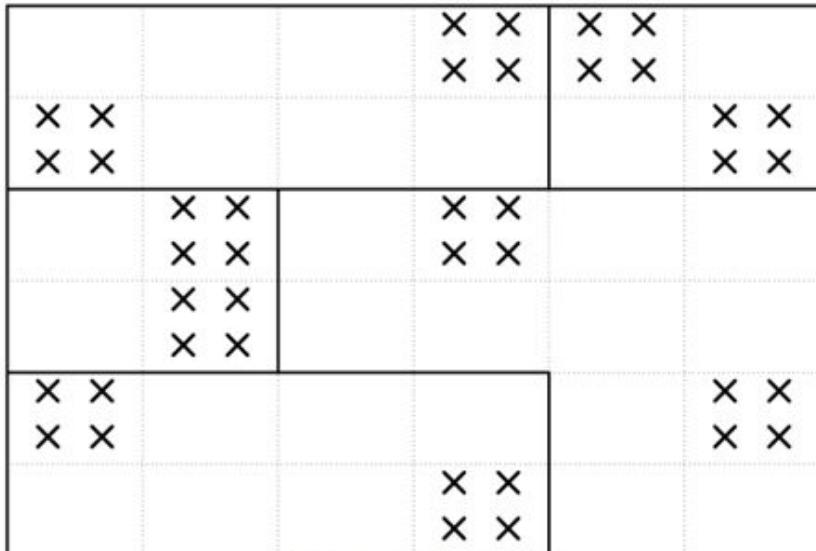


Auswahlwahrscheinlichkeit =  
Stichprobengröße/Größe der Schicht

Hochrechnungsfaktor =  
1/Auswahlwahrscheinlichkeit

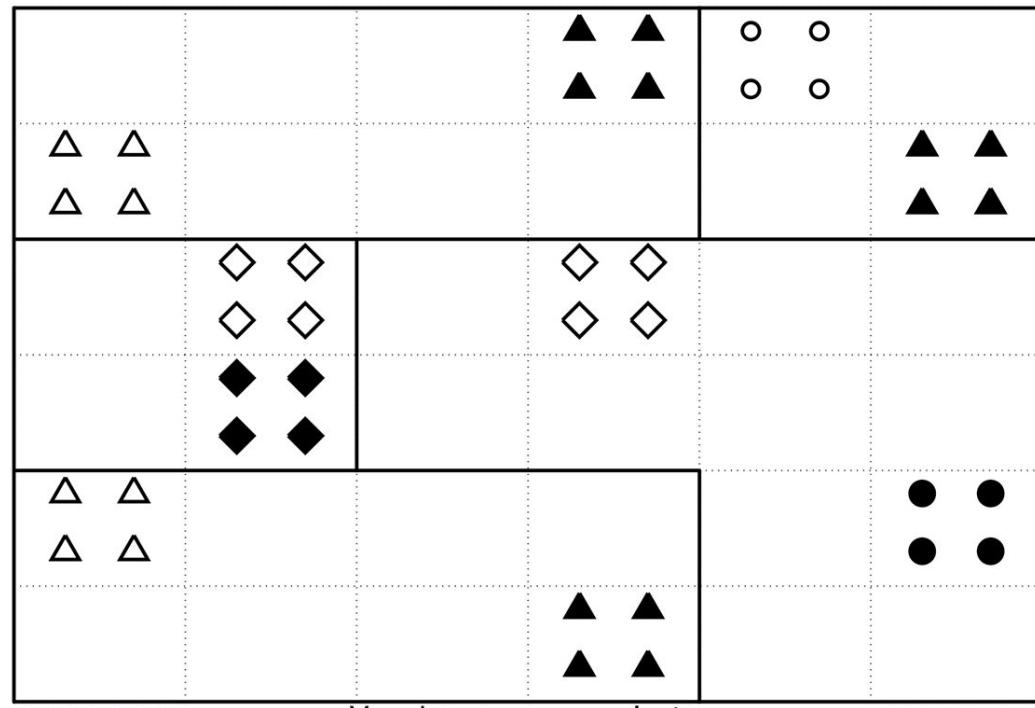


Stratified sample



Cluster sample within strata

# Wie viele von wo?



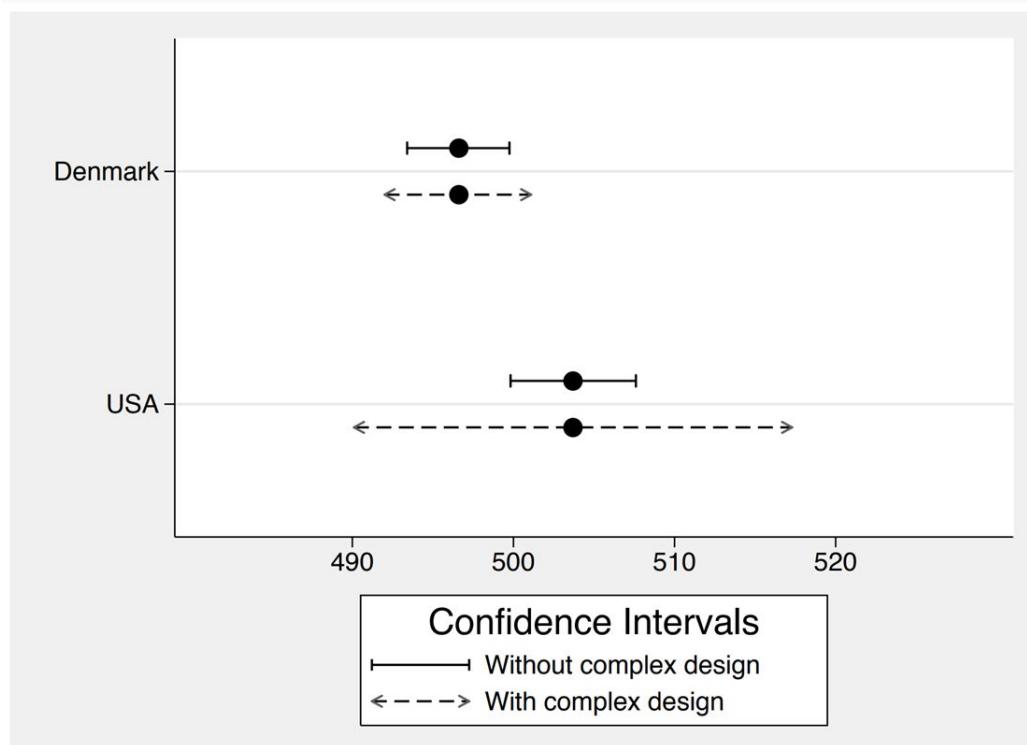
Mehrstufig (1) Gerichte (2) Akten

# Länder und Parteien Vergleiche - Wovon hängt das ab?

Werte (bei Prozenten)

Fallzahl

Stichprobenverfahren

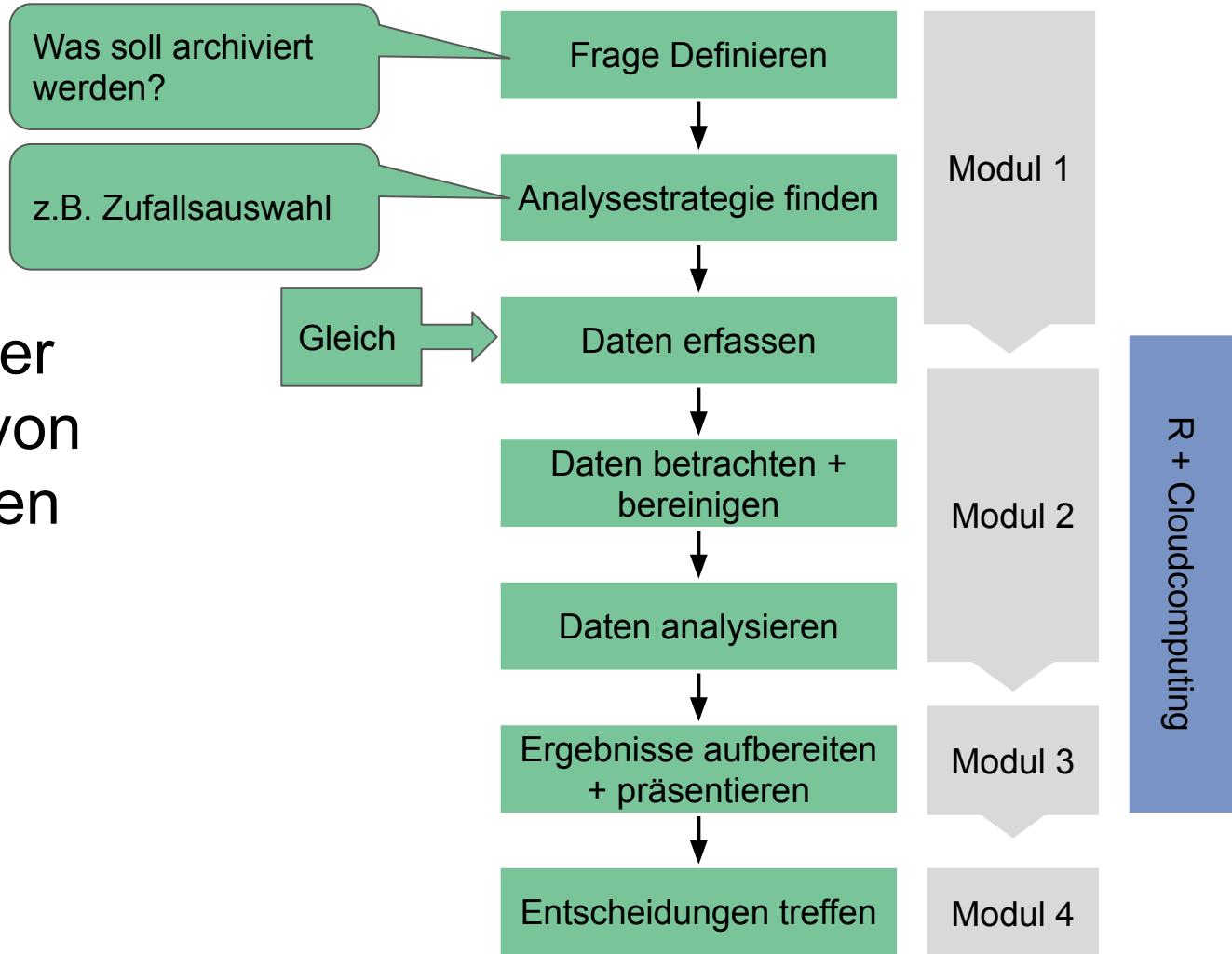


Werte aus der Pisa-Studie 2000

# Vorhersagemodell auf Basis von Daten im Archiv

Ist das sinnvoll?

# Vorgehen bei der Beantwortung von Fragen mit Daten



# Teamarbeit 5: Stratifizierte Stichprobenziehung

→ Gruppenarbeit

