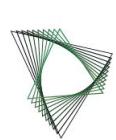


Modul 1: Einführung, Problemerfassung und Daten verstehen

Angewandte Datenanalyse für die öffentliche Verwaltung in Bayern (ADA Bayern)
www.ada-oeffentliche-verwaltung.de



BERD
@NFDI



Bayerisches Staatsministerium
für Digitales



Über uns



Prof. Dr. Frauke Kreuter



Dr. Malte Schierholz



Dr. Marcel Neunhoeffer



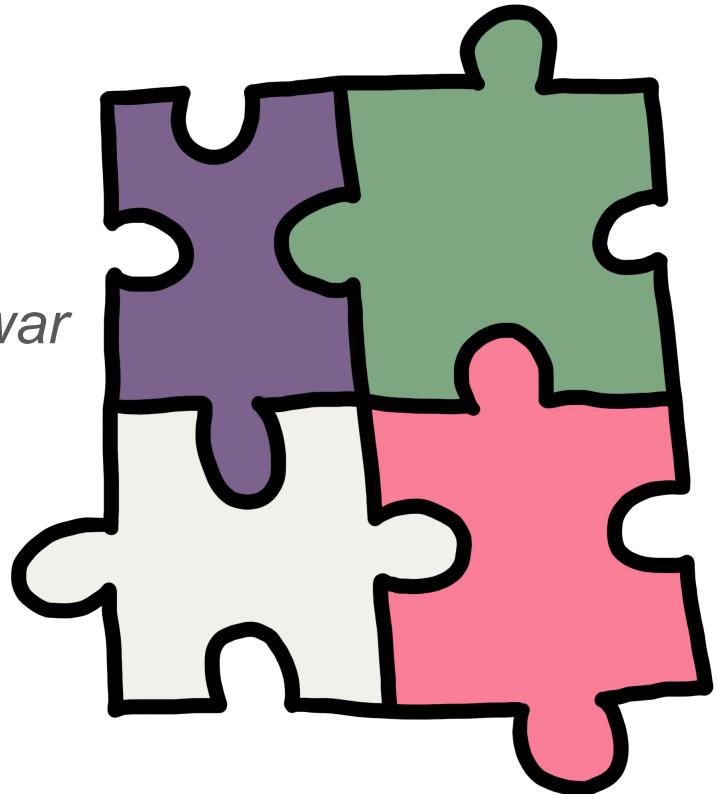
Dr. Heidi Seibold



Felix Henninger

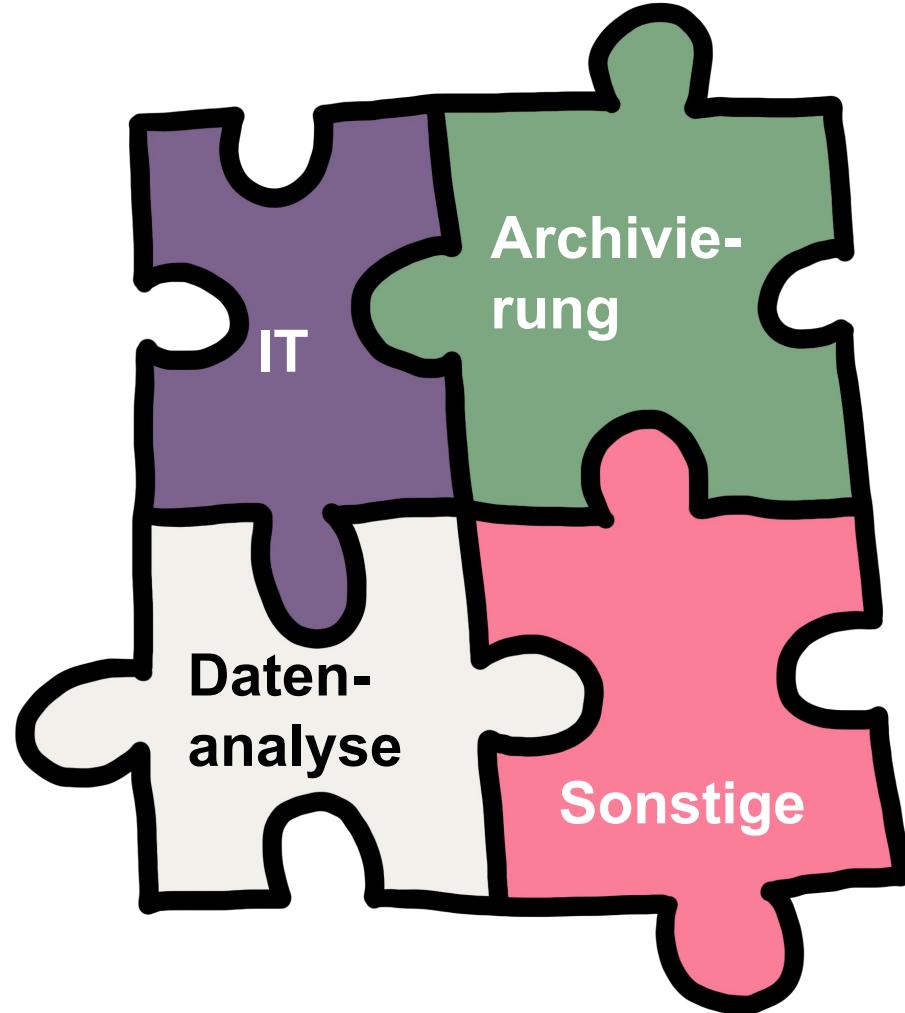
Über euch/Sie

- Name
- Fachlicher Hintergrund (kurz!)
- Grüner Zettel: *Wenn das passiert, war es ein guter Workshop*
- Roter Zettel: *Das soll hier auf gar keinen Fall passieren*



Einführung, Kennenlernen, Problemstellung	10:00 - 11:00
Pause	11:00 - 11:15
Teamarbeit	11:15 - 12:00
Mittagspause	12:00 - 12:50
Vorstellung des Projekt von Seiten der GDA	12:50 - 13:10
Daten Nutzen	13:10 - 14:00
Cloud Check	14:00 - 14:15
Pause	14:15 - 14:35
Teamarbeit	14:35 - 15:15
Wrap-up und Ausblick	15:15 - 15:30

Data Science ist ein Team Sport!



Teamarbeit

Die Teams arbeiten an einem Projekt, das am Ende in einer Präsentation vorgestellt werden soll:



- Wählen sie 1-2 Fragen, die man mit Daten bearbeiten kann
- Erörtern Sie mit dem gemeinsamen Wissen, wie die Frage(n) angegangen werden könnte(n)

Die Teams

Gruppe 1	Malte	Marilen Zilian	Constantin Breß	Aida Kreider	Xiangyu Wang	Nina Mayer
Gruppe 2	Marcel	Ines Müller	Markus Wingerath	Fabian Fuchs	Sven Morlock	Shanshan Zhou
Gruppe 3	Heidi	Christine Friederich	Julia Siebel		Benjamin Dornow	Anna Fuchs

Evidenzbasierte Verwaltung



THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

June 4, 2019

M-19-18

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Russell T. Vought
Acting Director

SUBJECT: Federal Data Strategy - A Framework for Consistency

AUTHENTICATED

- *Building a Culture that Values Data and Promotes Public Use (practices 1-10)*
- *Governing, Managing, and Protecting Data (practices 11-26)*
- *Promoting Efficient and Appropriate Data Use (practices 27-39)*

Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

- (a) **SHORT TITLE.**—This Act may be cited as the “Foundations for Evidence-Based Policymaking Act of 2018”.
(b) **TABLE OF CONTENTS.**—The table of contents for this Act

Foundations for
Evidence-Based
Policymaking Act
of 2018.
5 USC 101 note.

[ENGLISH](#)[FRANÇAIS](#)[KONTAKT](#)[DATENSCHUTZHINWEIS](#)

...

[☰ Menü](#) | [Bundesregierung](#) | [Startseite](#) [Suche](#)

Datenstrategie der Bundesregierung

Eine Innovationsstrategie für gesellschaftlichen Fortschritt und nachhaltiges Wachstum

- Kabinettfassung, 27. Januar 2021



...



Broschüre, 120 Seiten

Stand: 27. Januar 2021

Sprachen: Deutsch

Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz 2022

Datenlabor

Das BMUV initiiert und betreut unter anderem zwei zentrale Bausteine der Datenstrategie der Bundesregierung, die zugleich zentrale Modernisierungsvorhaben des BMUVs sind: das „BMUV Datenlabor“ und das „Anwendungslabor für KI und Big Data (KI-Lab)“.

Ziel beider Vorhaben ist, die Datengrundlage für die politische Entscheidungsfindung des Ministeriums zu verbessern. Beide Maßnahmen sind thematisch eng miteinander verbunden und sollen in Zukunft noch besser voneinander profitieren.

Ziel des Projekts war daher die bedarfs- und nutzerzentrierte Entwicklung von gemeinsamen Prozessen und Dienstleistungen beider Labore mit Fokus auf den Aufbau des Datenlabors.



24.05.2022

Hintergrundinformation



Was ist eigentlich die Frage? Problemstellung definieren.



Ziel: Definiere das Projektziel



Aktion: Welche Maßnahmen/Handlungen wird dieses Projekt informieren?



Daten: Welche Daten stehen dazu intern zur Verfügung? Welche müssen extern ergänzt werden? Was brauchst du um an die Daten zu kommen?



Analyse: Welche Analyse ist notwendig? Handelt es sich dabei um eine Beschreibung, Intervention, Vorhersage oder Verhaltensänderung

Ethische Aspekte: Privacy, Transparenz, Gleichstellung, Verantwortlichkeiten

Es beginnt mit einer Frage...

Wir haben 100 KiTa Plätze. Wer soll die bekommen?

Welche Flächen müssen wir in Zukunft wegen Hochwasser anders bewirtschaften

Wir haben 100 Professuren in der High-Tech Agenda finanziert. War das sinnvoll?

Wie verteilen wir die Impfdosen?

Es gibt Berichte von Sozialarbeitern. Wo müssen wir intervenieren?

Sollen wir diese Maßnahme nochmal machen?

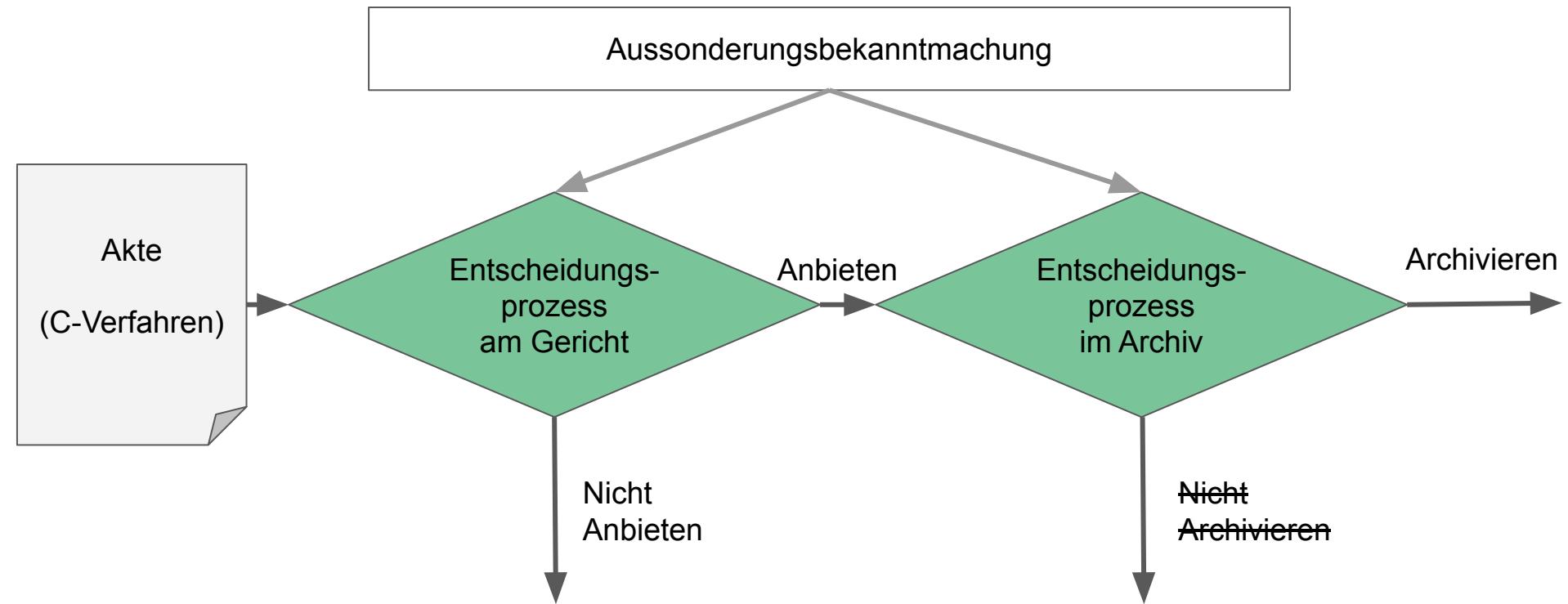
Was soll archiviert werden?

Was ist eigentlich die Frage? Problemstellung definieren.

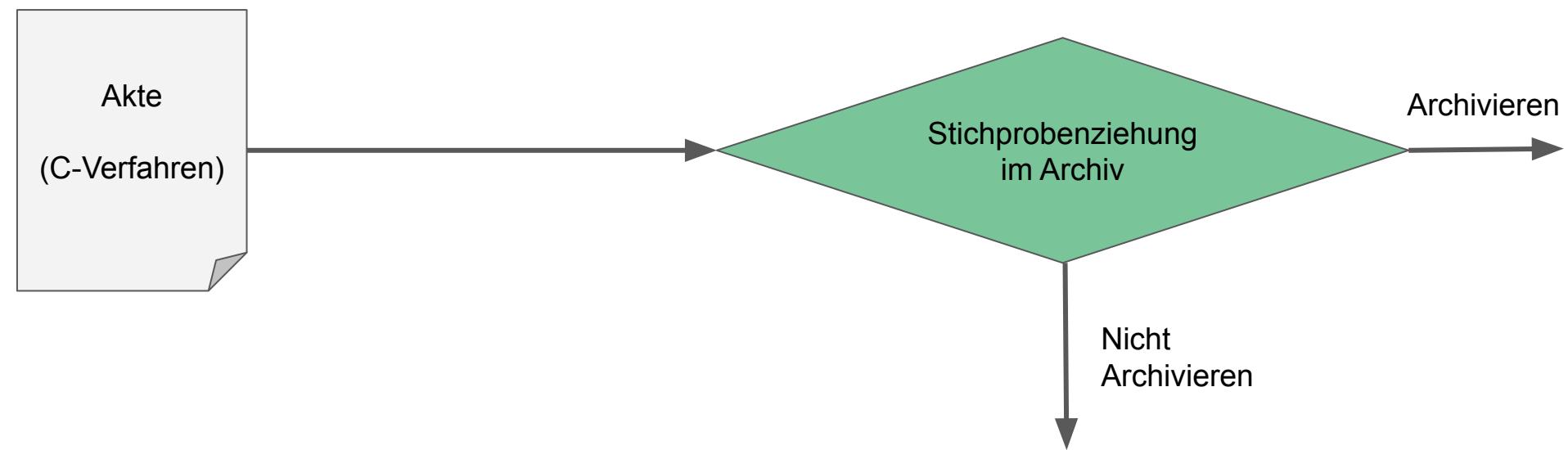
Hier:



Entscheidungsprozess heute



Entscheidungsprozess mit Hilfe von Stichprobenziehung: Typische Fälle



Fragen beantworten mit Daten



- In den US verbleiben weniger als die Hälfte der HIV-Positiven in Versorgungsprogrammen.
- KI hilft ambulanten vorherzusagen, wer aus Vorsorge herausfällt.
- Ressourcen für Interventionen werden entsprechend des Ausfallrisikowertes geplant

Ramachandran, A., Kumar, A., Koenig, H. et al. Predictive Analytics for Retention in Care in an Urban HIV Clinic. *Sci Rep* 10, 6421 (2020).
<https://doi.org/10.1038/s41598-020-62729-x>

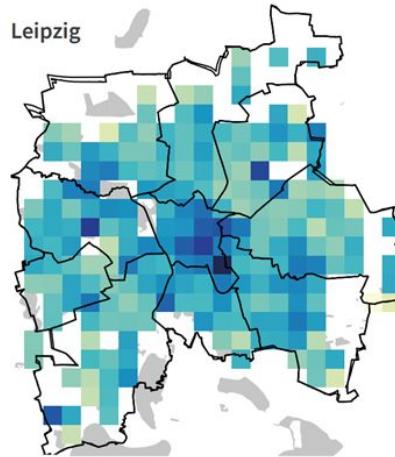
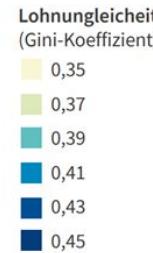
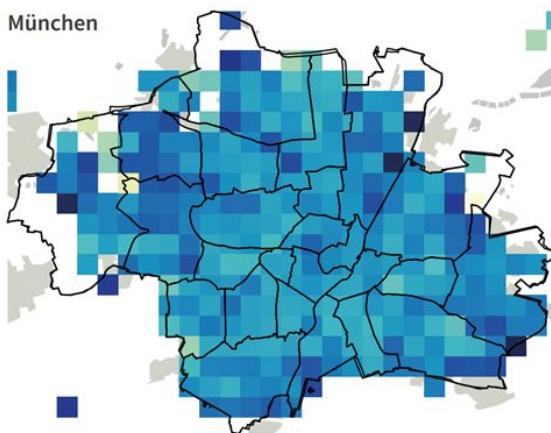
- Kooperation: Johnson County, Kansas Carnegie Mellon University
- Ziel: Spirale unbehandelter psychiatrischer Erkrankungen und Verhaftungen zu durchbrechen.
- KI-gestützte Vorhersage der Wahrscheinlichkeit ins Gefängnis zu kommen
- Personen mit hohem Risikowert werden je nach vorhandenen Ressourcen priorisiert behandelt



Rodolfa, K.T., Lamba, H. & Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nat Mach Intell* 3, 896–904 (2021).
<https://doi.org/10.1038/s42256-021-00396-x>

Fragen beantworten mit administrativen Daten

Kleinräumige Verteilung der Lohnungleichheit in München und in Leipzig 2017



■ geografische Flächen, die schwach besiedelt, bewaldet oder Gewässer sind

Die Datenbasis beruht auf den administrativen Daten der Bundesagentur für Arbeit und zeigt 1x1-Kilometer-Gitterzellen. Insgesamt werden für München 294 und für Leipzig 218 unzensierte Gitterzellen (mit mehr als 20 Personen) abgebildet. Dargestelltes Maß für die kleinräumige Lohnungleichheit ist der zellenspezifische Gini-Koeffizient.

Interpretationshilfe: Hellere Gitterzellen weisen auf einen niedrigeren Gini-Koeffizienten – also niedrigere Lohnungleichheit – hin, dunklere auf einen höheren Gini-Koeffizienten. Flächen ohne Gitterzellen sind nicht oder zu schwach besiedelt und somit censiert. Die zugrunde liegenden Karten zeigen schwach besiedelte und bewaldete Flächen sowie Gewässer (grau). Schwarze Linien kennzeichnen die einzelnen Bezirke.

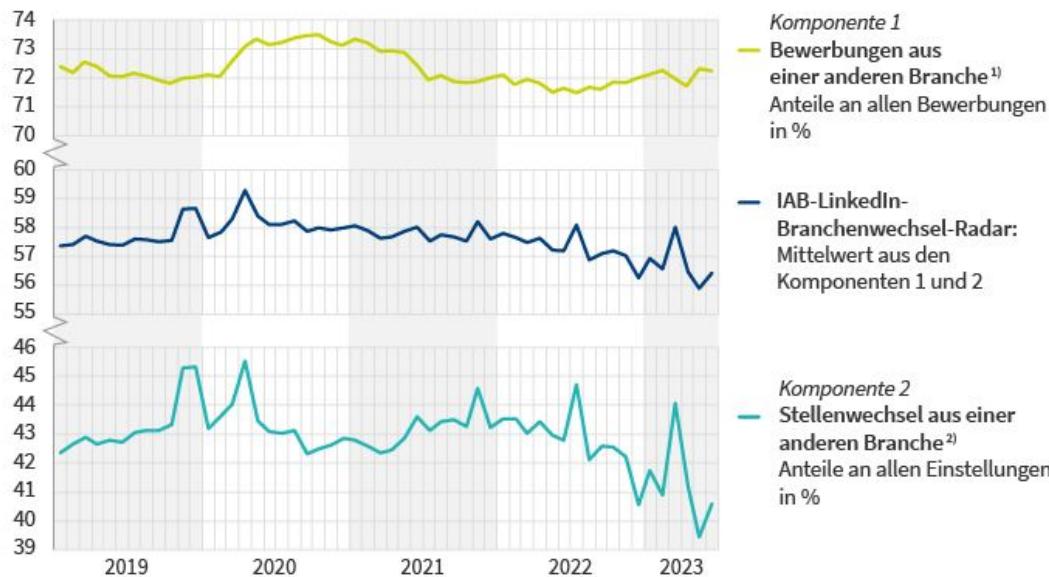
Quellenangabe: GridAB v2.1, eigene Darstellung. © IAB

“Die
Lohnungleichheit
ist in
ostdeutschen
Nachbarschaften
kleiner als in
westdeutschen”

<https://doku.iab.de/kurzber/2023/kb2023-09.pdf>

Fragen beantworten mit (den richtigen) Daten

Abb: Das IAB-LinkedIn-Branchenwechsel-Radar



¹⁾ LinkedIn-Mitglieder, die sich mit dem Online-Bewerbungsformular von LinkedIn auf eine Stelle in einer anderen Branche bewerben, als sie laut Mitgliederprofil zum Zeitpunkt der Bewerbung beschäftigt sind.

²⁾ LinkedIn-Mitglieder, in deren Mitgliederprofil vor Antritt der neuen Stelle eine Stelle in einer anderen Branche eingetragen war.

“Das IAB-LinkedIn-Branchenwechsel-Radar: „Great Resignation“ ist kein Trend”

Fragen beantworten mit kombinierten Daten

Average adult income rank
for a poor child

60th percentile

50th

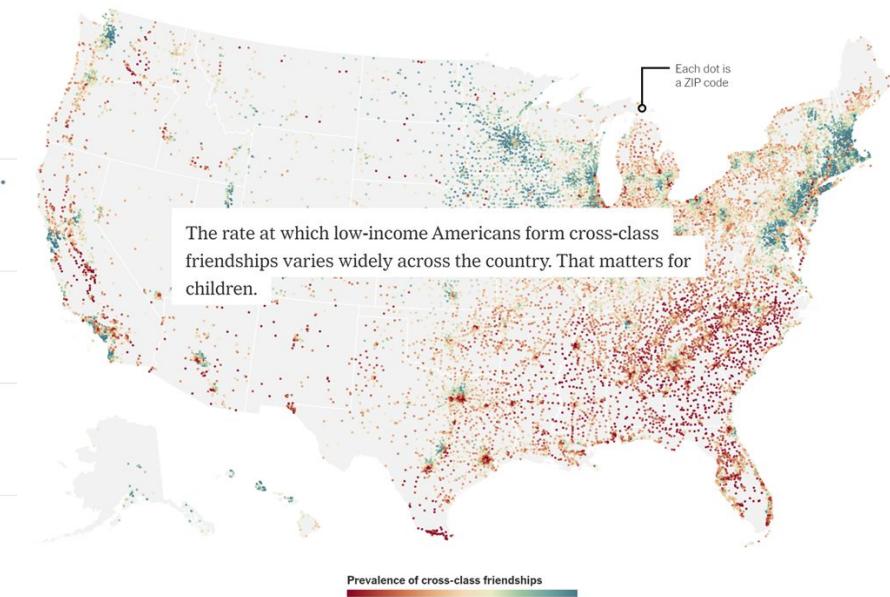
40th

30th

Winsted, Minn. (55395)
High mobility

The new research has an answer: economic connectedness. In places where poor people have more rich friends, outcomes are better.

Conley, Ga. (30288)
Low mobility



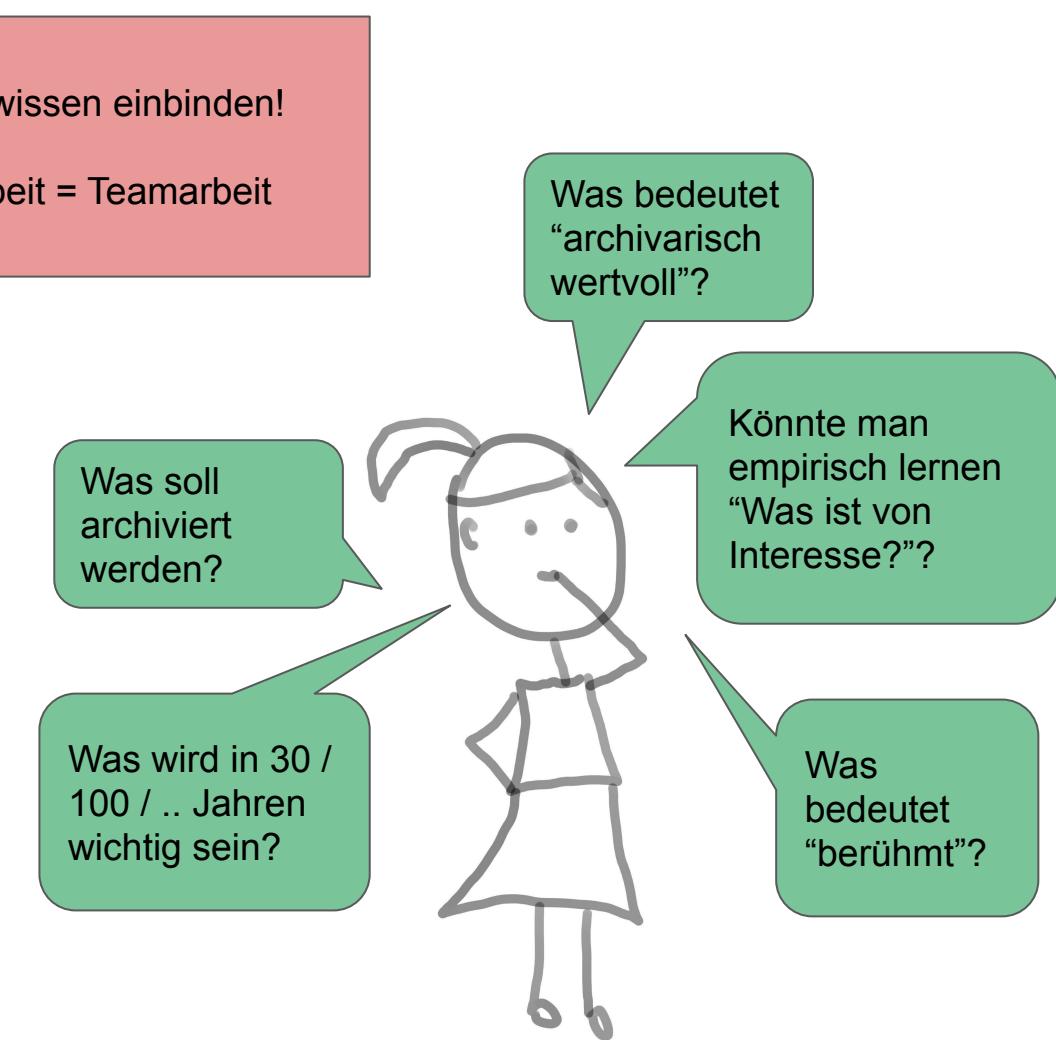
US-Steuerdaten kombiniert mit Facebook-Daten
<https://www.nature.com/articles/s41586-022-04996-4>

Teamarbeit 1

Expertenwissen einbinden!

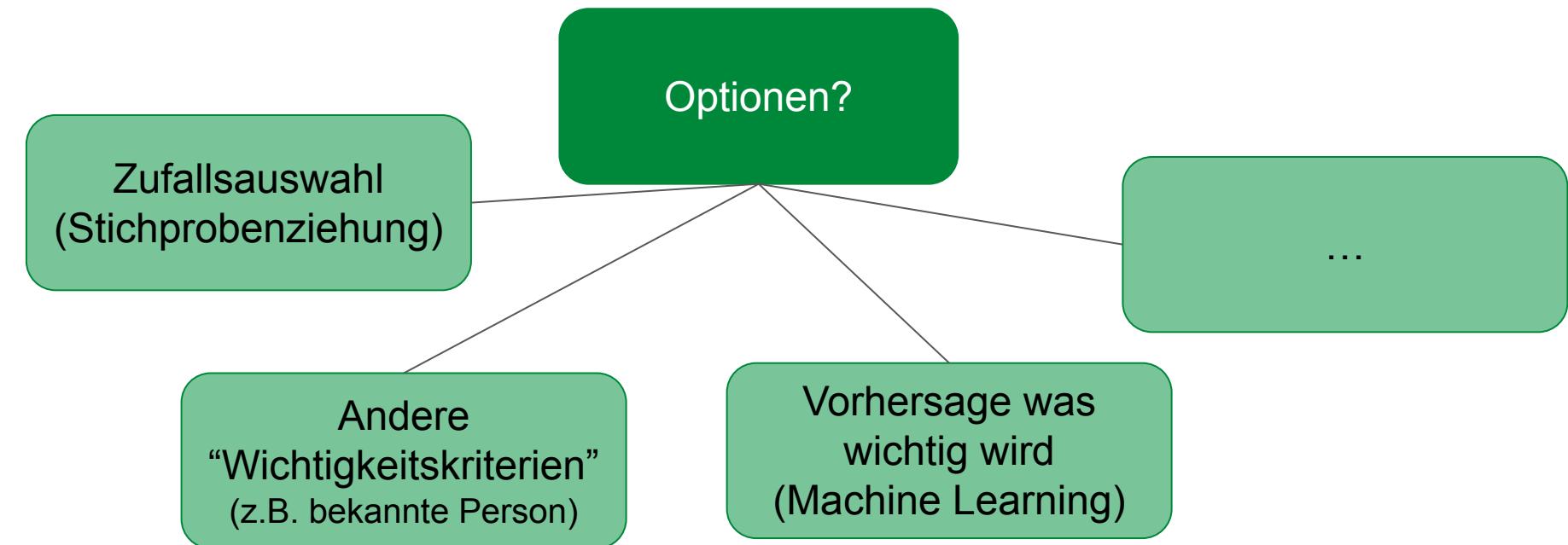
Datenarbeit = Teamarbeit

1. Besprechen Sie die Expertisen der Team-Mitglieder
2. Besprechen Sie: Wie kann man entscheiden, welche Akten langfristig archiviert werden sollen?
3. Was sind die Kriterien für "archivwürdig"?

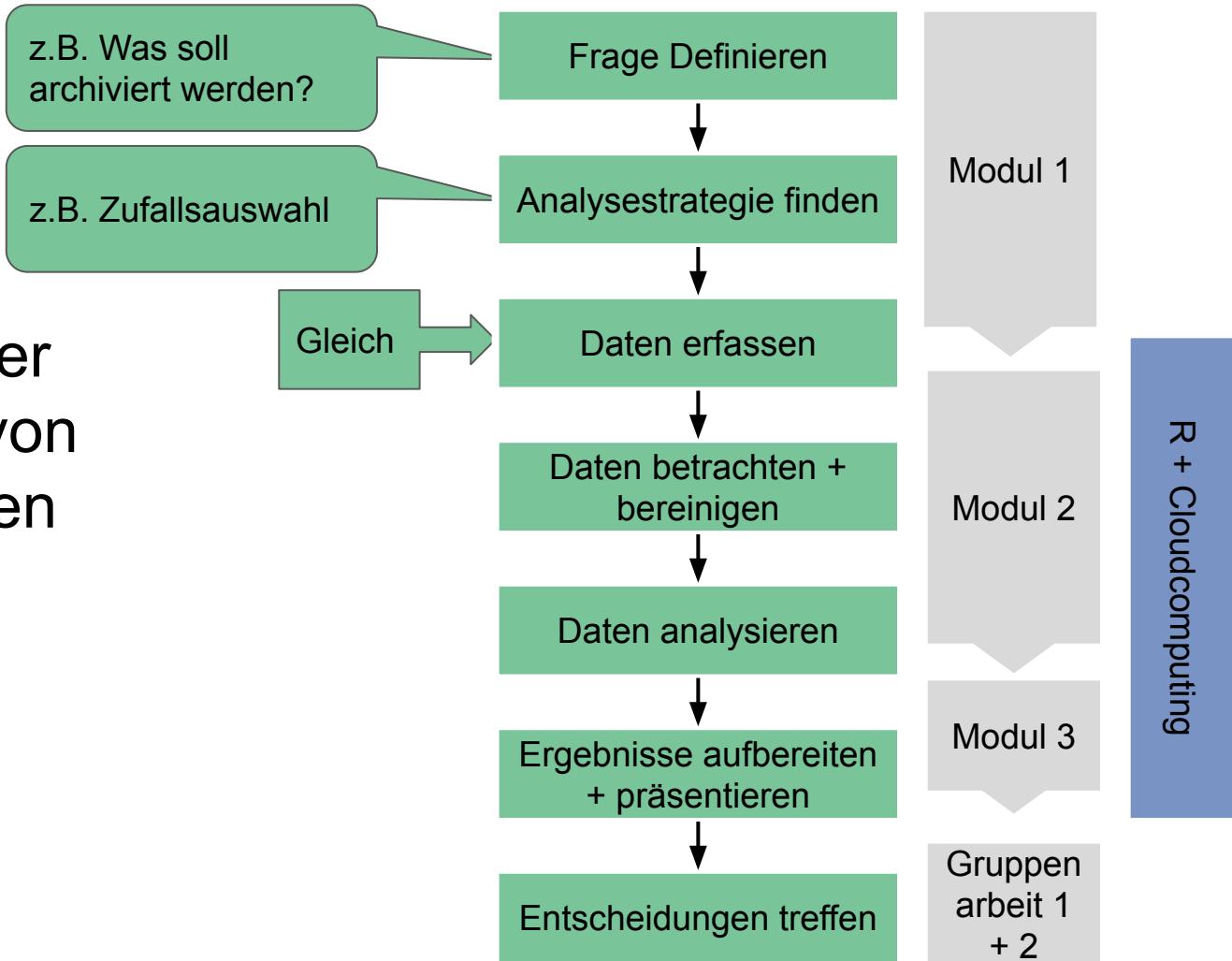


Daten nutzen, um eigene Probleme zu lösen.

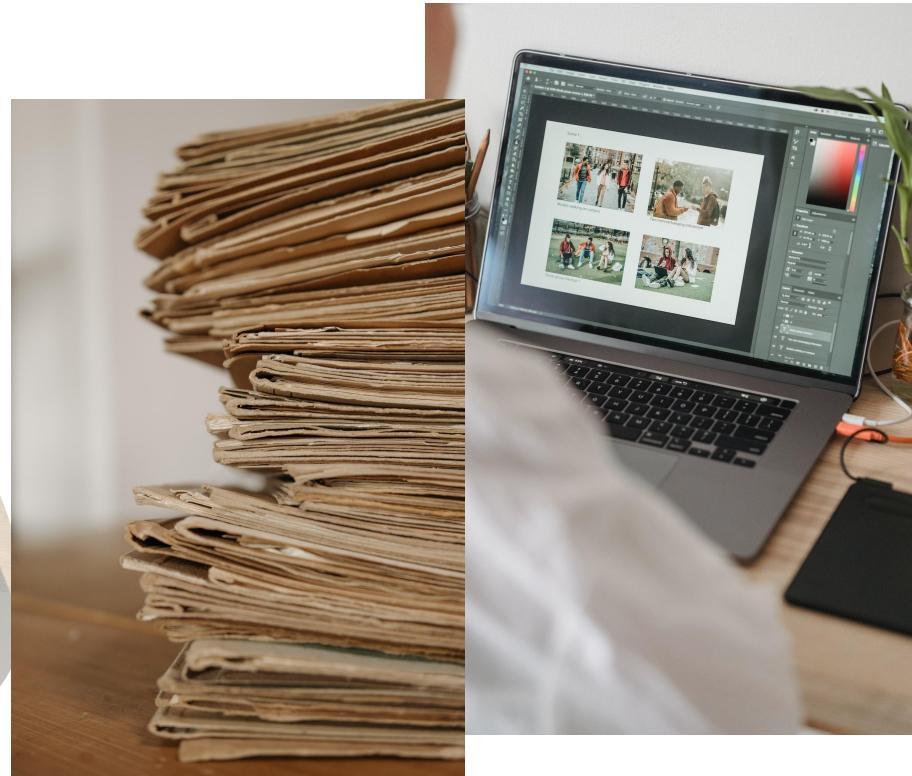
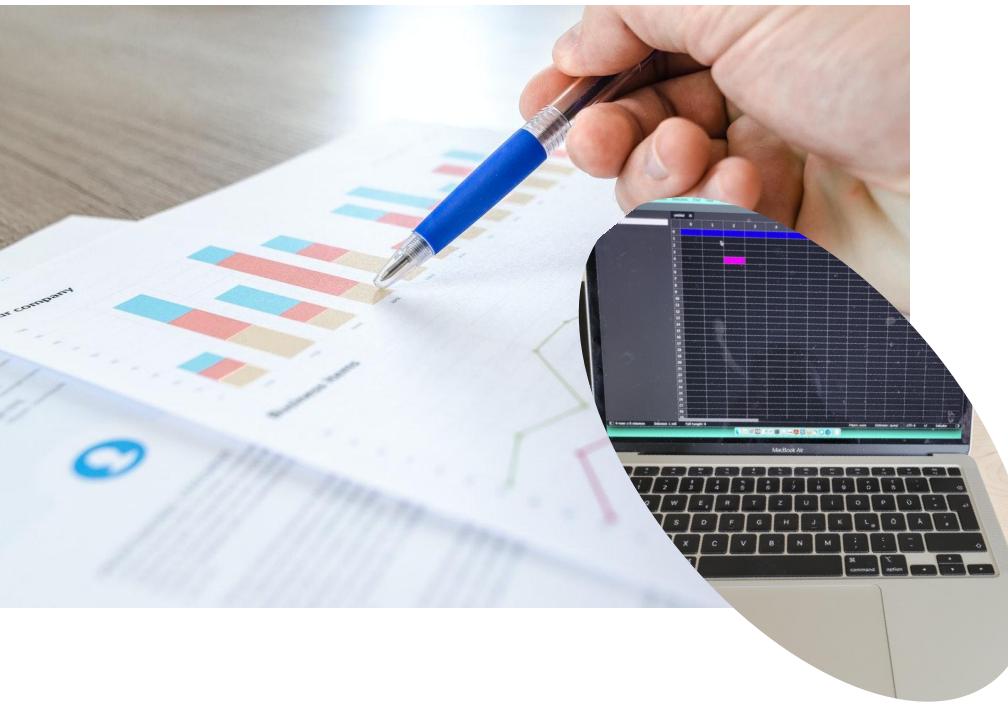
Datenbasiert entscheiden welche Akten langfristig archiviert werden sollen.



Vorgehen bei der Beantwortung von Fragen mit Daten

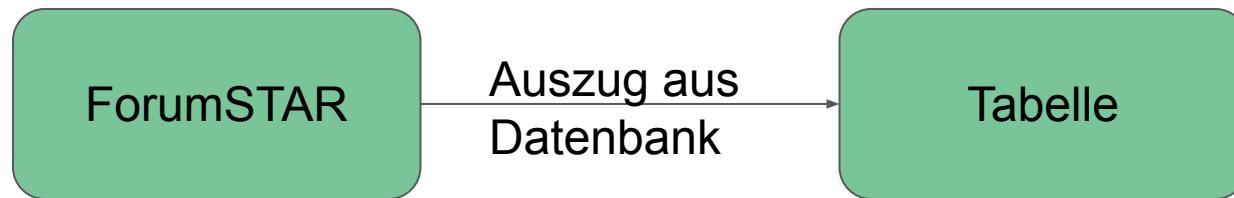


Was sind Daten?



Mit welchen Arten von Daten beschäftigen Sie sich?

Welche Daten gibt es in unserem Projekt?



AZ: 1 F 1/15

Daten betr. Scheidung

Wurde dieses Verfahren nach dem 20.06.2012 eingeleitet? ja nein

Sind beide Ehegatten Iraner? ja nein

örtliche Zuständigkeit [?](#)

F_Sch_ROM III

Heirat: 15.08.1990 **Ehezeitbeginn:** 01.08.1990 [?](#) [§](#) **ZU SchAntrag:** 18.01.2015 **Ehezeitende:** 31.12.2014 [?](#)

vor dem Standesbeamten? Ja Nein

Standesamt: Rosenheim **Staat:** Deutschland [?](#)

HeiratRegNr: 15/1990 **(wenn Heiratsort im Ausland)**

Hatten die Ehegatten zwischen Rechtshängigkeit und letzter mtl. Verhandlung zu irgendeinem Zeitpunkt - gemeinsam ihren gew. Aufenthaltsort in Deutschland? ja nein [?](#)

Haben die Ehegatten eine Rechtswahl getroffen? ja nein

Hat es früher eine Trennungsentscheidung gegeben (Art. 9 ROM III-VO)? ja nein [?](#)

Hatten beide Ehegatten bei Anrufung des Gerichts ihren gew. Aufenthaltsort in Deutschland? ja nein

Staatsangehörigkeit bei ZU des Scheidungsantrags (wegen VA gem. Art. 17 Abs. 3 EGBGB):

<input checked="" type="checkbox"/> Ehemann: Türkei	<input checked="" type="checkbox"/> Ehefrau: Türkei
<input type="checkbox"/> außerdem - nicht "effektiv" -	<input type="checkbox"/> außerdem - nicht "effektiv" -
<input type="checkbox"/> Ehemann ist staatenlos	<input type="checkbox"/> Ehefrau ist staatenlos

War einer der Ehegatten bei ZU des Scheidungsantrags Flüchtling oder Asylberechtigter?

Antragsteller/in <input checked="" type="checkbox"/> Ehefrau <input type="checkbox"/> Ehemann	<input type="checkbox"/> ja <input checked="" type="checkbox"/> nein
VB Antragsgegner/in <input checked="" type="checkbox"/> kein VB oder VB stellt keinen Antrag <input type="checkbox"/> stellt Scheidungsantrag <input type="checkbox"/> beantragt Abweisung Scheidungsantrag	<input type="checkbox"/> widerspricht <input type="checkbox"/> widerspricht nicht <input type="checkbox"/> äußert sich nicht
Antragsgegner/in persönlich <input checked="" type="checkbox"/> stimmt zu <input type="checkbox"/> widerspricht	<input type="checkbox"/> ja <input type="checkbox"/> nein

Soll der Verfahrensteil Scheidung begründet werden? ja nein

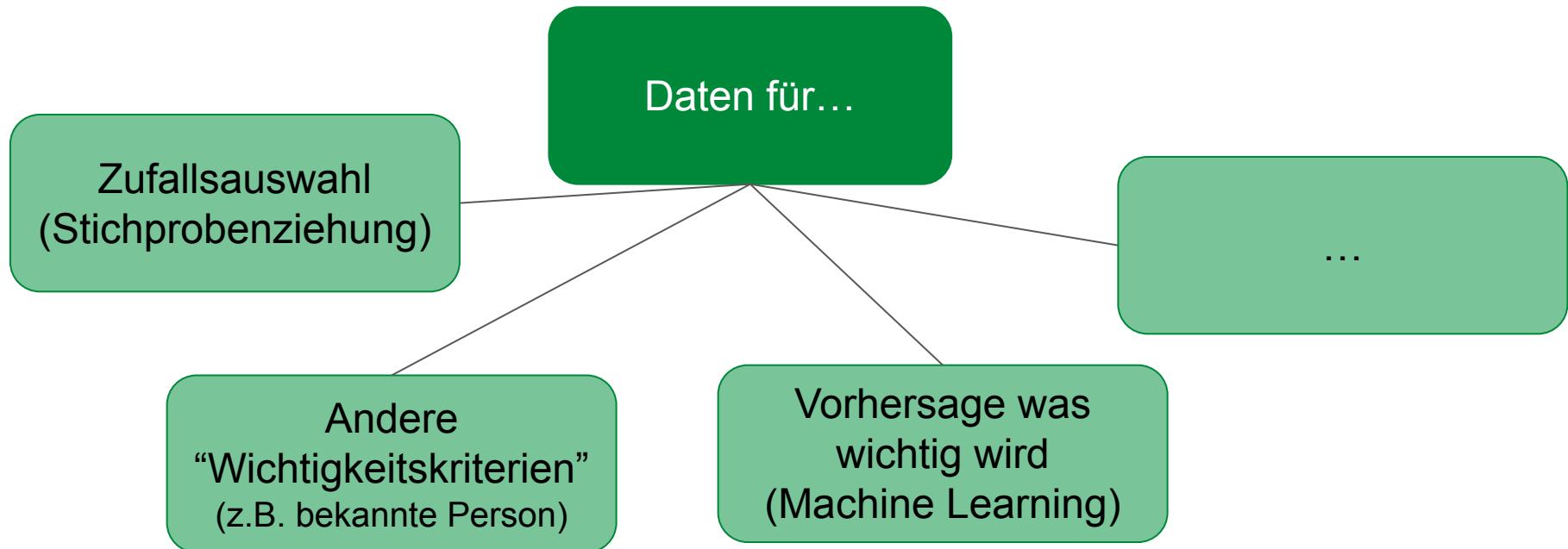
Feststellungen des Gerichts zur Dauer des Getrenntlebens weniger als 1 Jahr (unzumutbare Härte) mindestens 1 Jahr

[Generelle Vorgaben ändern](#)

Weiter [E](#)

DE 15:46 18.01.2015

Welche Daten könnte man noch nutzen?



Daten: das neue Gold

Was macht Daten so wertvoll?
(auch in unserem Projekt)

Zeitersparnis
(Automatisierung)

Verbesserung der
Archivierung

Bürger:innen glücklich
machen

Forschung
verbessern

...

Fragen?

Daten sammeln und verstehen

Selbst erhoben

Von anderen
Organisationen

Aus dem Internet

Woher können Daten kommen?

Von Kollaborations-
Partnern

...

Open Data

<https://bund.dev/apis>

bundDEV
VERWALTUNG DIGITAL



Egal ob Sie als Privatperson eine Open-Source-Anwendung bauen oder als Unternehmen Open-Data in Ihre Produkte integrieren wollen. Unsere Schnittstellen stehen der gesamten Gesellschaft offen.

<https://www.opendata.bayern>

Open Data
Bayern

Offene Daten aus Bayern

Hier finden Datenbegeisterte freie Datensätze und die Verwaltung Unterstützung, um mehr Daten zu teilen. Damit schaffen wir gemeinsam - Verwaltung, Unternehmen, aber auch Wissenschaft und Zivilgesellschaft - Mehrwert für uns alle.

Daten entdecken

[Daten liefern](#)

F A I R

FINDABLE



ACCESSIBLE



INTEROPERABLE



REUSABLE



Where is
the data?

How can
I access it

How can I
read and work
with it?

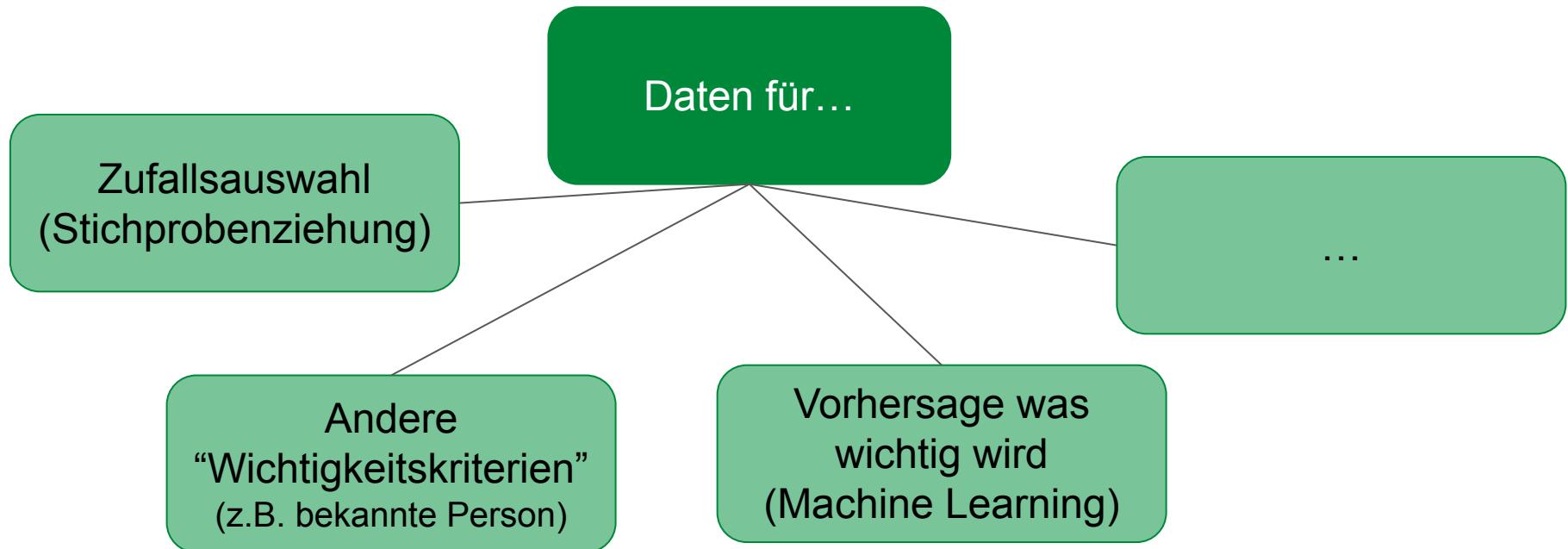
How can I
reuse it?

F A I R

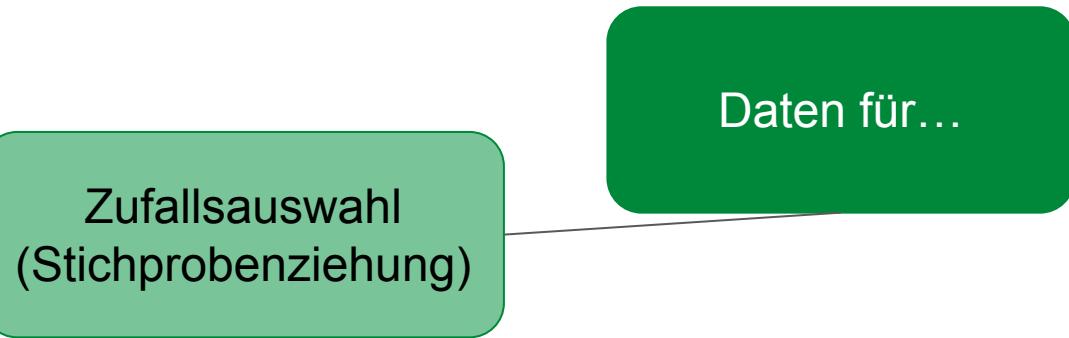
≠

OPEN

Woher könnten Daten kommen?



Woher könnten Daten kommen?



Wie entstehen die verfügbaren Daten?



Was ist eine Datenbank?

- Sammlung von Daten über Entitäten
- Datenmodell von “Relationalen Datenbankmanagementsystemen” (RDBMS):
 - Daten sind in **Tabellen** gespeichert
 - Beziehungen zwischen Tabellen und deren Spalten werden in einem **Schema** definiert
 - RDBMS verwenden die **Structured Query Language** (SQL) um Nutzer mit den Daten interagieren zu lassen
- RDBMS sind weit verbreitet, da
 - viele Nutzer **gleichzeitig** Daten abrufen können
 - Daten bleiben jederzeit **konsistent** und fehlerfrei (*ACID*), selbst bei gleichzeitigen Updates und bei Netzwerkfehlern
- Alternativ: Vielfältige NoSQL-Datenbanken

Ein simples Datenbank-Schema für Gerichtsverfahren

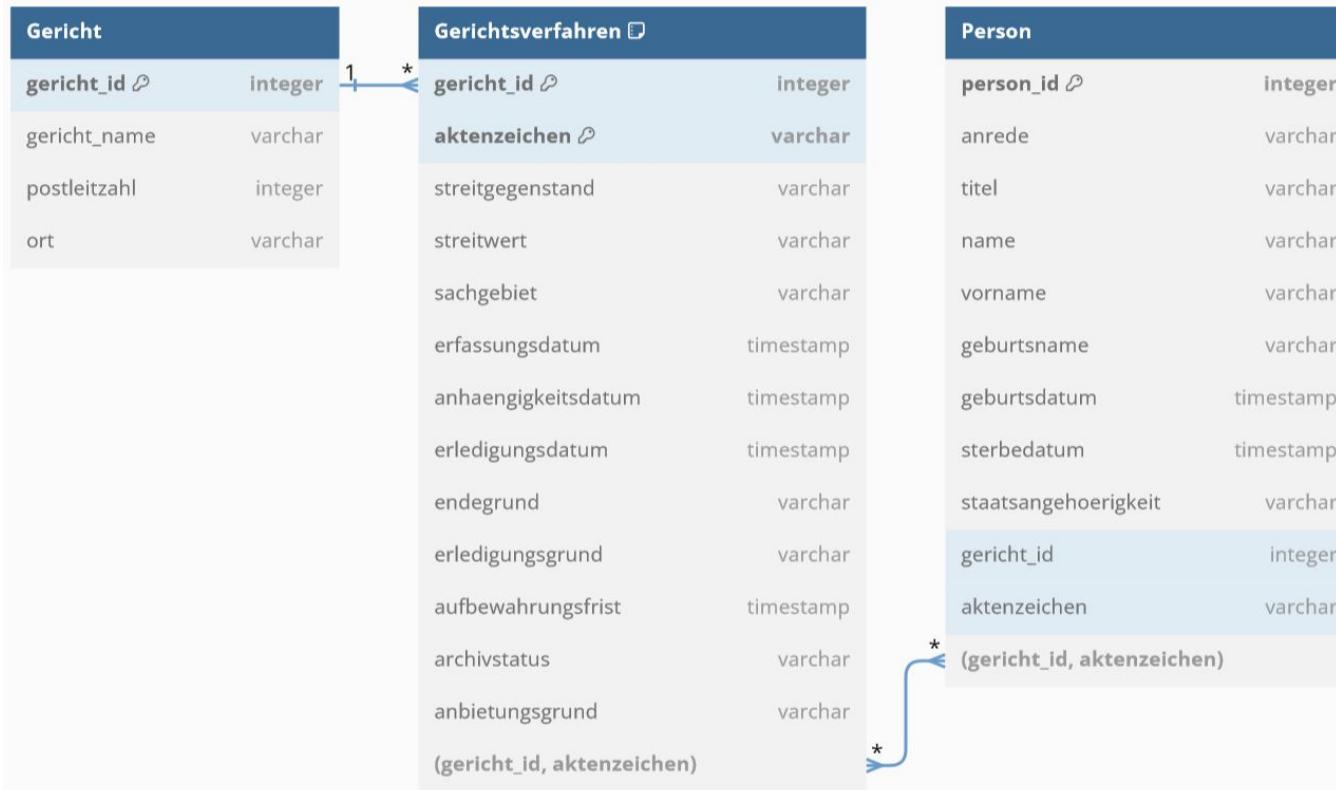


Tabelle: Akten

```
1 SELECT Gericht, Aktenzeichen, Verfahrensstatus, Kurzrubrum, "Streitwert in EURO", Gesamtstreitgegenstand, Streitgege  
2 FROM cases;
```

	Gericht	Aktenzeichen	Verfahrensstatus	Kurzrubrum	Streitwe	Gesamtst
1	Amtsgericht Erlangen	1 C 100/18	weggelegt	ANONYMISIERT	7.920 [NULL]	
2	Amtsgericht Erlangen	1 C 105/18	weggelegt	ANONYMISIERT	1.777,15 [NULL]	
3	Amtsgericht Erlangen	1 C 11/18	weggelegt			
4	Amtsgericht Erlangen	1 C 16/18	weggelegt			
5	Amtsgericht Erlangen	1 C 21/18	weggelegt			
6	Amtsgericht Erlangen	1 C 27/18	weggelegt			
7	Amtsgericht Erlangen	1 C 3/18	weggelegt			
8	Amtsgericht Erlangen	1 C 7/18	weggelegt			

Tabelle: Schema

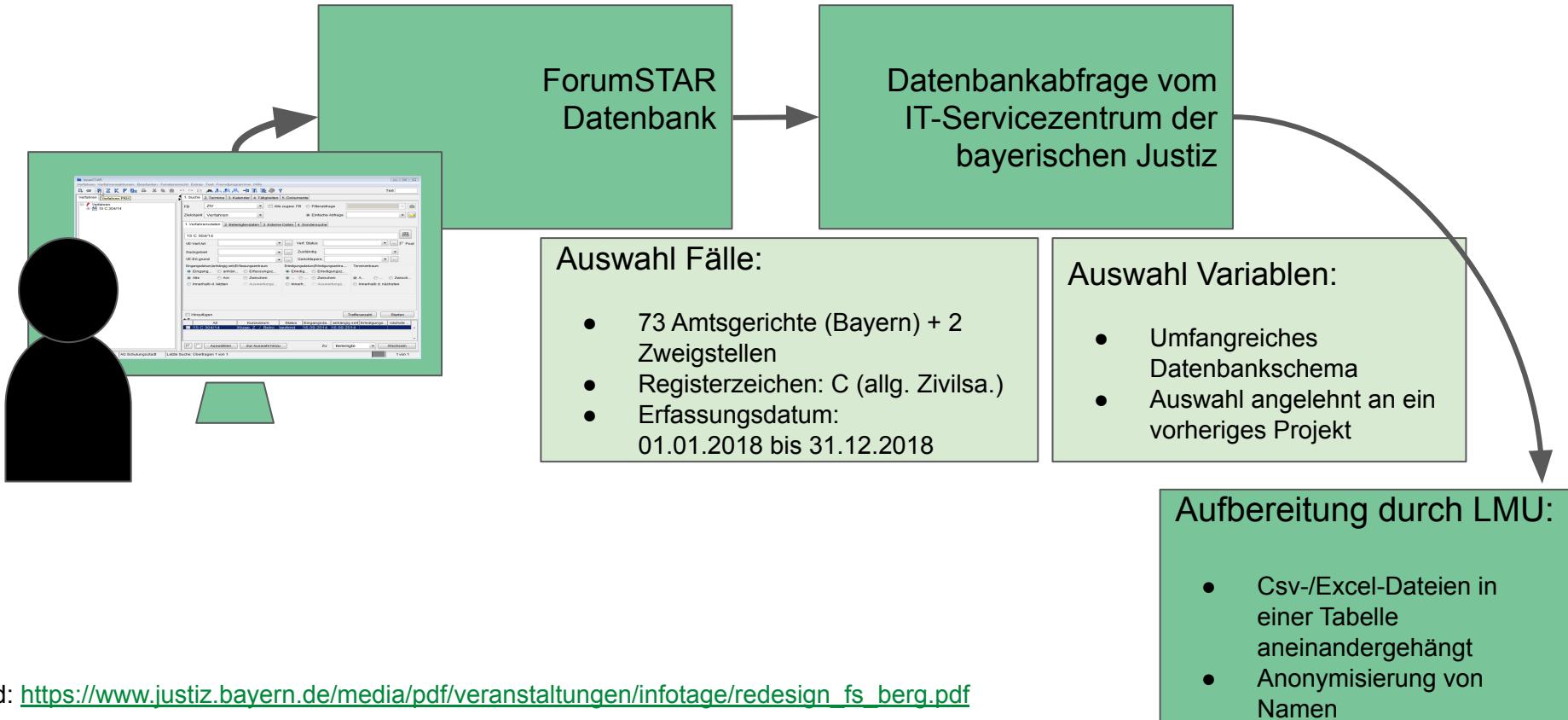
```
5 SELECT * FROM sqlite_schema;
```

	type	name	tbl_name	rootpage	sql
1	table	sqlite_stat1	sqlite_stat1	121	CREATE TABLE sqlite_stat1(tbl,idx,stat)
2	table	sqlite_stat4	sqlite_stat4	122	CREATE TABLE sqlite_stat4(tbl,idx,neq,nlt,ndlt,sample)
3	table	cases	cases	2	CREATE TABLE `cases`(`Gericht` TEXT,`Aktenzeichen` TEXT,`Verfahrensstatus` TEX
4	table	people	people	123	CREATE TABLE `people`(`Gericht` TEXT,`Aktenzeichen` TEXT,`Verfahrensbeteiligu

Tabelle: Personen

	Gericht	Aktenzeichen	Verfahrensbeteiligungsart	Anrede
1	Amtsgericht Erlangen	1 C 100/18	Kläger	Herr
2	Amtsgericht Erlangen	1 C 100/18	Beklagter	Herr
3	Amtsgericht Erlangen	1 C 100/18	Dritter	Frau
4	Amtsgericht Erlangen	1 C 100/18	Dritter	Herr
5	Amtsgericht Erlangen	1 C 100/18	Dritter	Herr
6	Amtsgericht Erlangen	1 C 100/18	Dritter	Frau

Wie entstehen die verfügbaren Daten?



Warum ist es wichtig
zu wissen, wie die
Daten entstanden
sind?

Welche externen Datenquellen können genutzt werden?

Daten für...

Andere
“Wichtigkeitskriterien”
(z.B. bekannte Person)



```
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 }
```

26949 Ergebnisse in 23922 ms

[Code](#)[Herunterladen](#)[Link](#)

Table ▾

Search

person	personLabel	birthDate	description
wd:Q88991	Lena Malkus	6. August 1993	deutsche Weitspringerin
wd:Q89097	Carmen Klaschka	8. Januar 1987	deutsche Tennisspielerin
wd:Q89128	Laura Siegemund	4. März 1988	deutsche Tennisspielerin

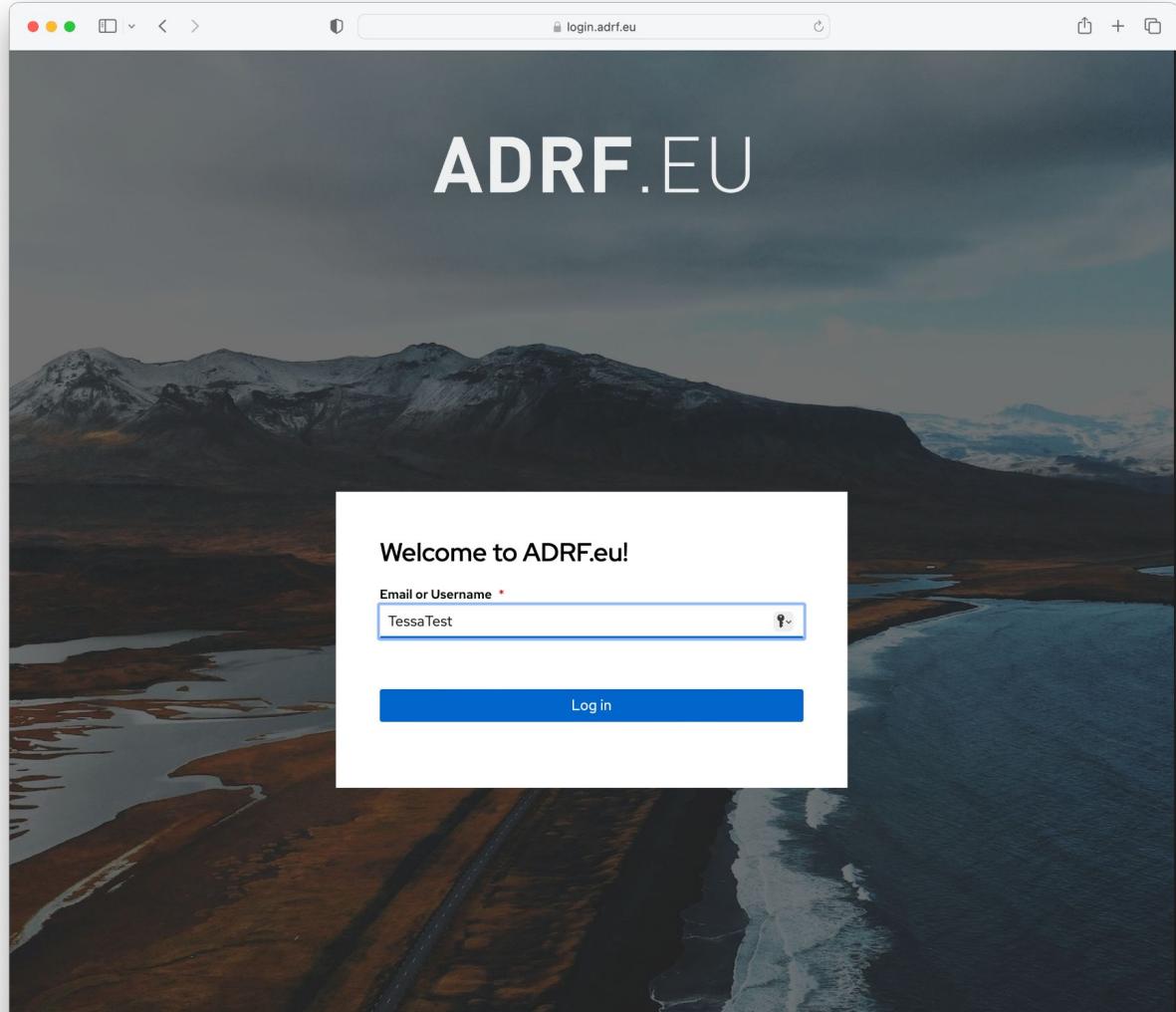
Cloud-Plattform Check



Login

Zugang erhalten Sie
immer über die
Adresse
<https://login.adrf.eu>

Bitte loggen Sie sich
mit dem
Benutzernamen
VornameNachname
ein (TessaTestuserin)



Zugang

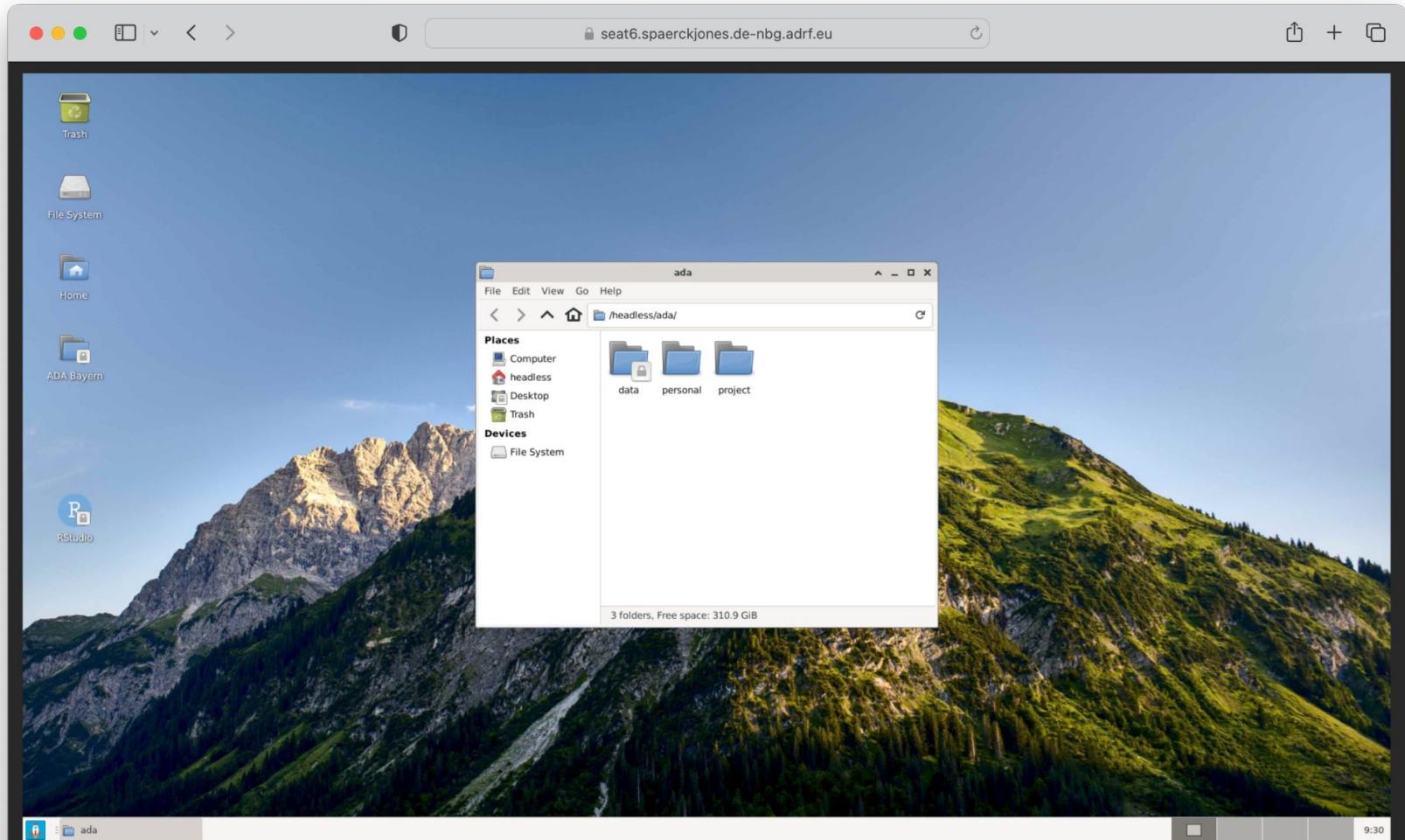
Nach dem Login
sehen Sie den
Zugang für unseren
Workshop.

The screenshot shows a web browser window with the URL `login.adrf.eu` in the address bar. The page title is "My applications". There is a single application entry: "ADA Bayern – Workshop 2", which features a large letter "A" icon and the text "ADA Cloudzugang". To the right of the application list is a search bar with the placeholder "Search...". The top right corner of the browser window shows user information: "TessaTest" and a profile picture. The overall interface is clean and modern, typical of a corporate intranet or service provider's login portal.



Virtuelle Analyseumgebung

Die **Cloudplattform** können Sie nutzen wie Ihren eigenen PC



My applications

ADA Bayern – Workshop 2

ADA Clouzugang

Search...

login.adrf.eu

TessaTest



login.adrf.eu



0



TessaTest



User details

Sessions

Consent

MFA Devices

Connected services

Tokens and App passwords

Update details

Username *

TessaTest

Name *

Tessa Testuserin

Email *

Email

Locale *

Auto-detect (based on your browser)

Save

Change your password

Change password



0



- User details
- Sessions
- Consent
- MFA Devices
- Connected services
- Tokens and App passwords

Enroll RefreshDelete

Type

- WebAuthn Authenticator Setup Stage
- Static Authenticator Stage
- TOTP Authenticator Setup Stage

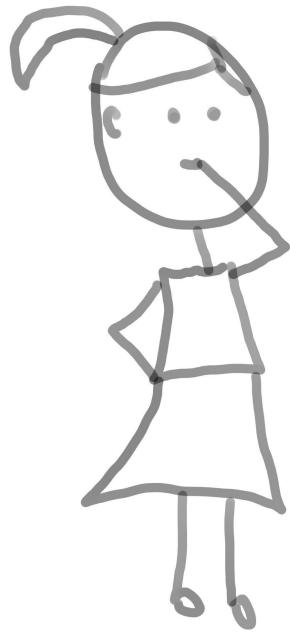
No objects found.

1 - 0 of 0

Teamarbeit 2

- Was wären optimale Möglichkeiten für die Auswahl (basierend auf Daten)?
 - Was müssten wir dafür digitalisieren?
 - Wie würde eine optimale Welt hier aussehen?
- Warum archivieren wir nicht einfach alle Akten?
- Falls noch Zeit ist: Gibt es bei Ihnen ähnliche Fragen/Projekte, die man mit Hilfe von Daten angeht/angehen könnte?





Was nehmen Sie von heute mit?

Welche Fragen haben Sie?

Ausblick: Morgen

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

Ausblick: Übermorgen

Was bisher geschah + Visualisieren	10:00 - 10:30
Besuch des Digitalministers Dr. Fabian Mehring	10:30 - 11:15
Visualisieren und Vorbereitung Nachmittag	
Mittagspause	12:00 - 12:45
Teamarbeit	12:45 - 15:10
Pause: Selbstbestimmt nach Bedarf der Teams	
Abschluss	15:10 - 15:30