

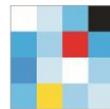
# Modul 4: Ergebnispräsentation und Umsetzung

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

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



**BERD**  
@NFDI

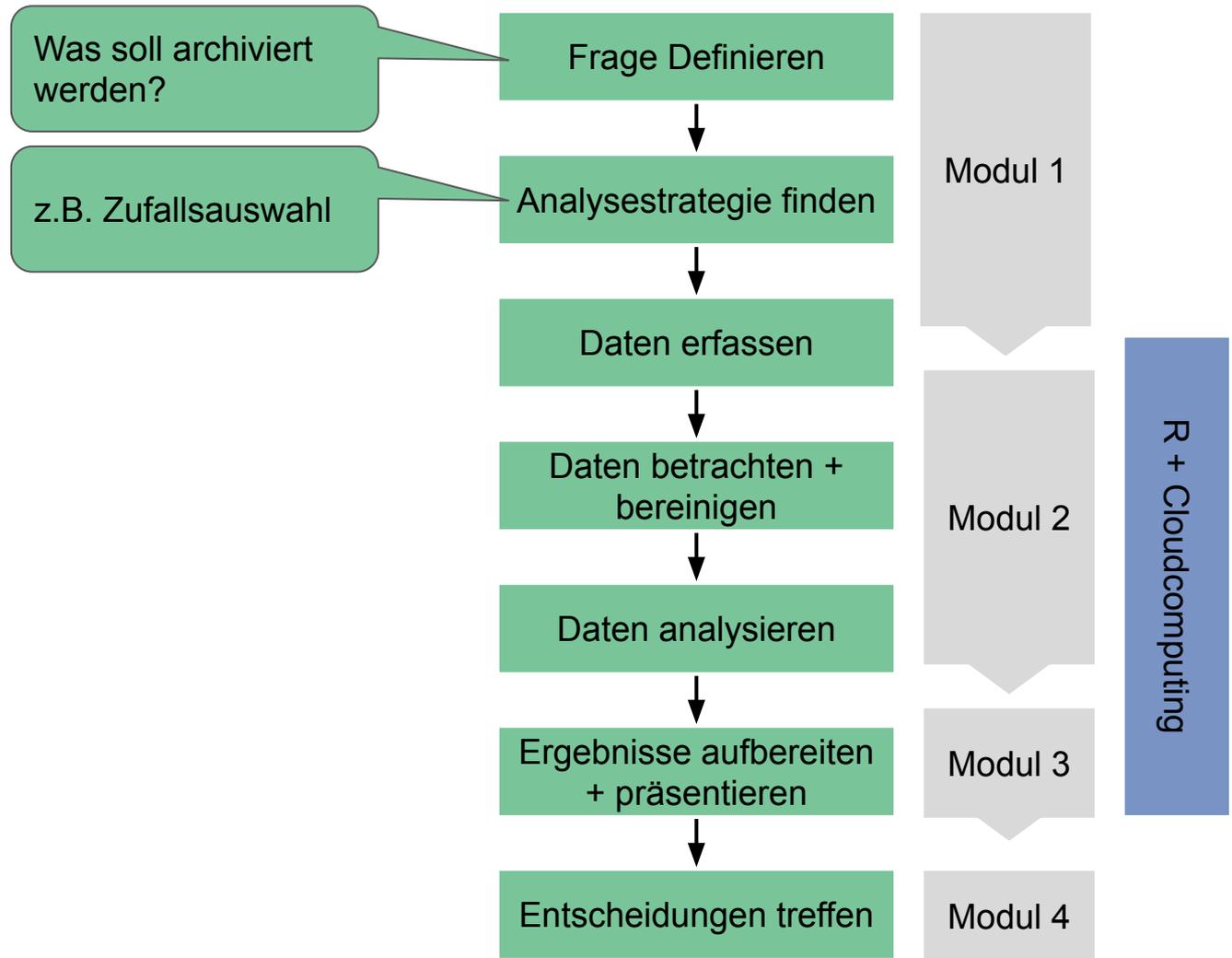


Bayerisches Staatsministerium  
für Digitales

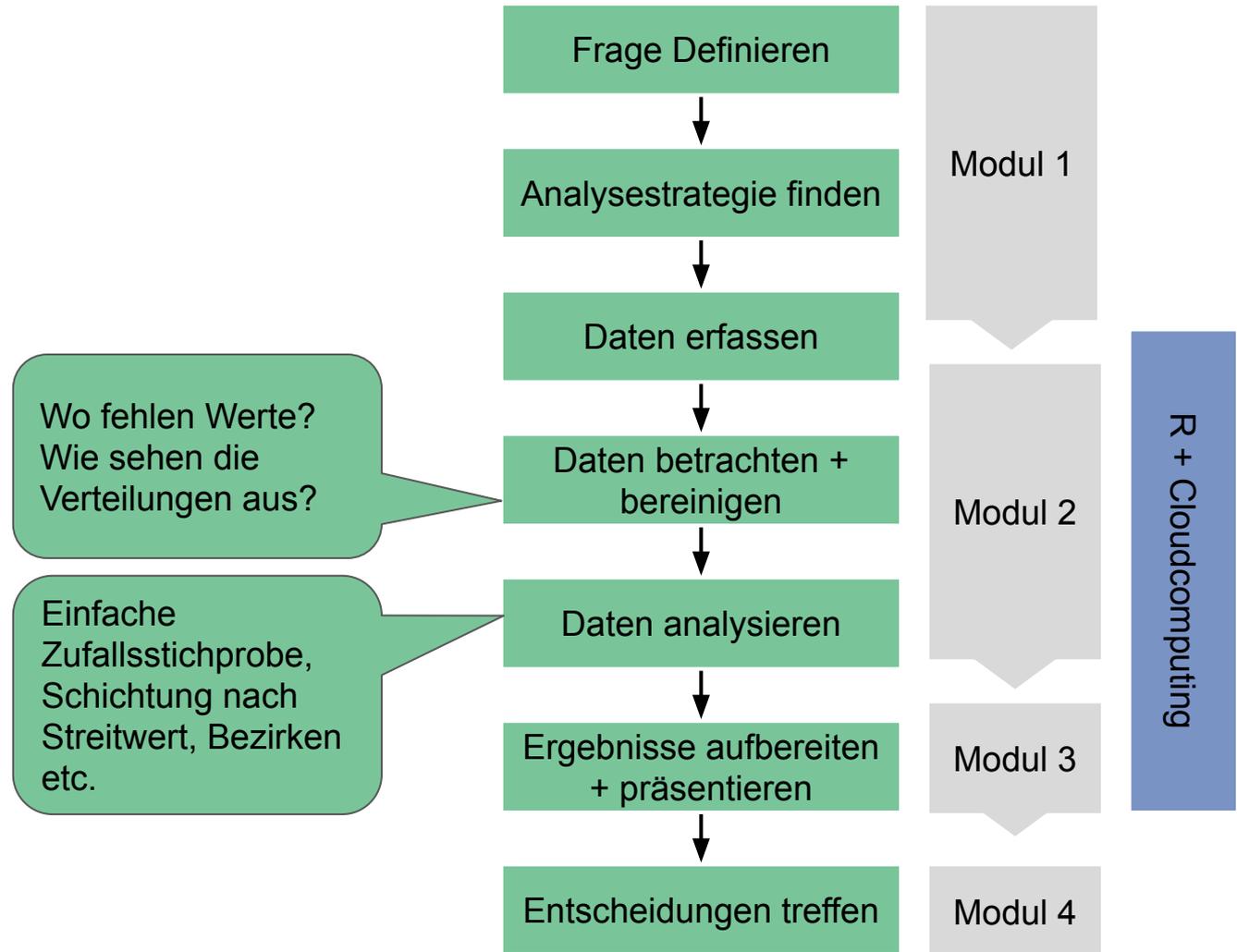


Vortrag: Was bisher geschah + Record Linkage	13:00 - 13:45
Pause	13:45 - 13:55
Teamarbeit: Ergebnisse + Präsentation fertigstellen	13:55 - 15:15
Pause	15:15 - 15:30
Teampräsentationen: 6-7 Minuten pro Team	15:30 - 16:10
Pause	16:10 - 16:15
Diskussion: Wie geht es weiter?	16:15 - 16:40
Abschlussrunde	16:40 - 17:00

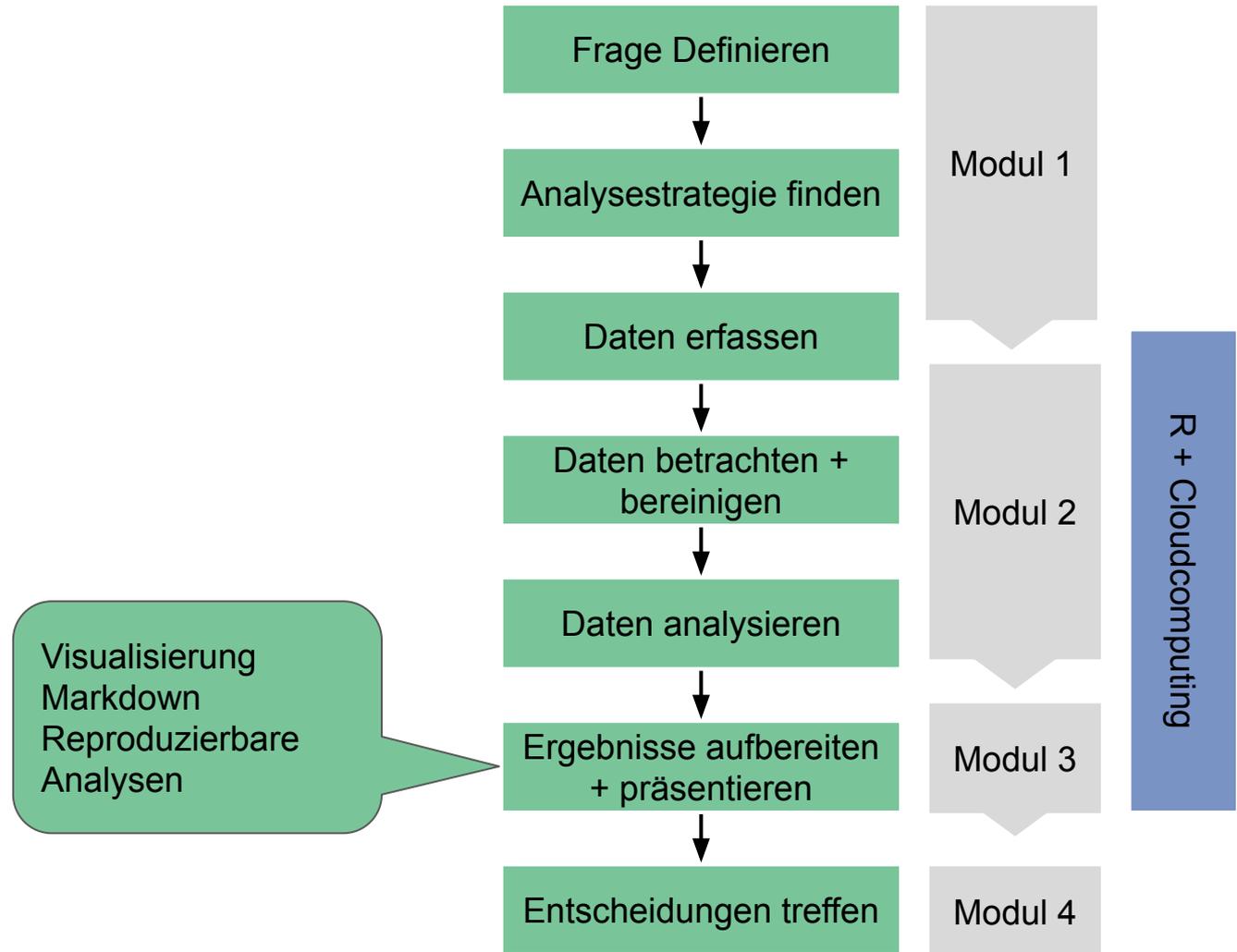
# Rückblick



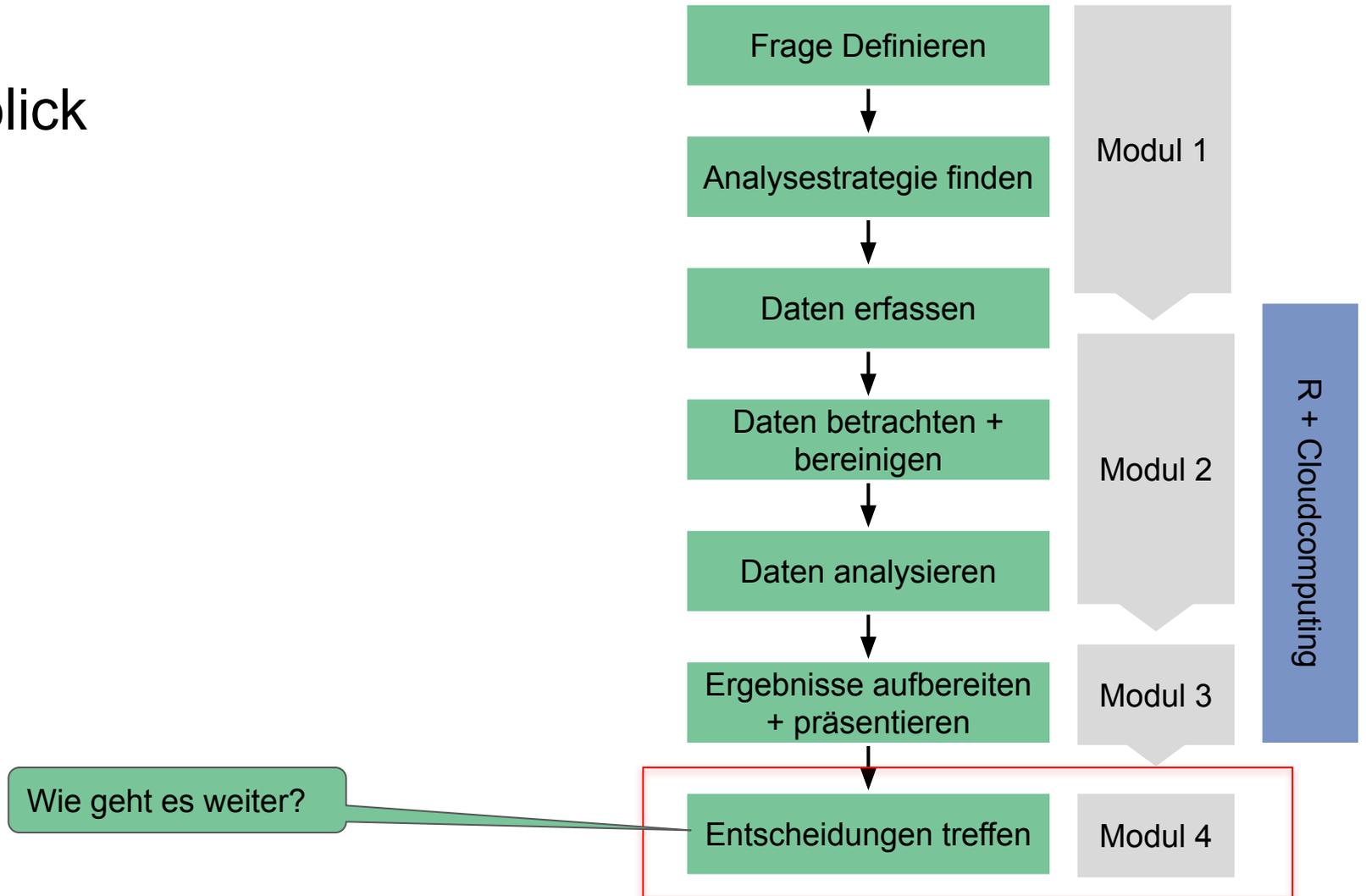
# Rückblick



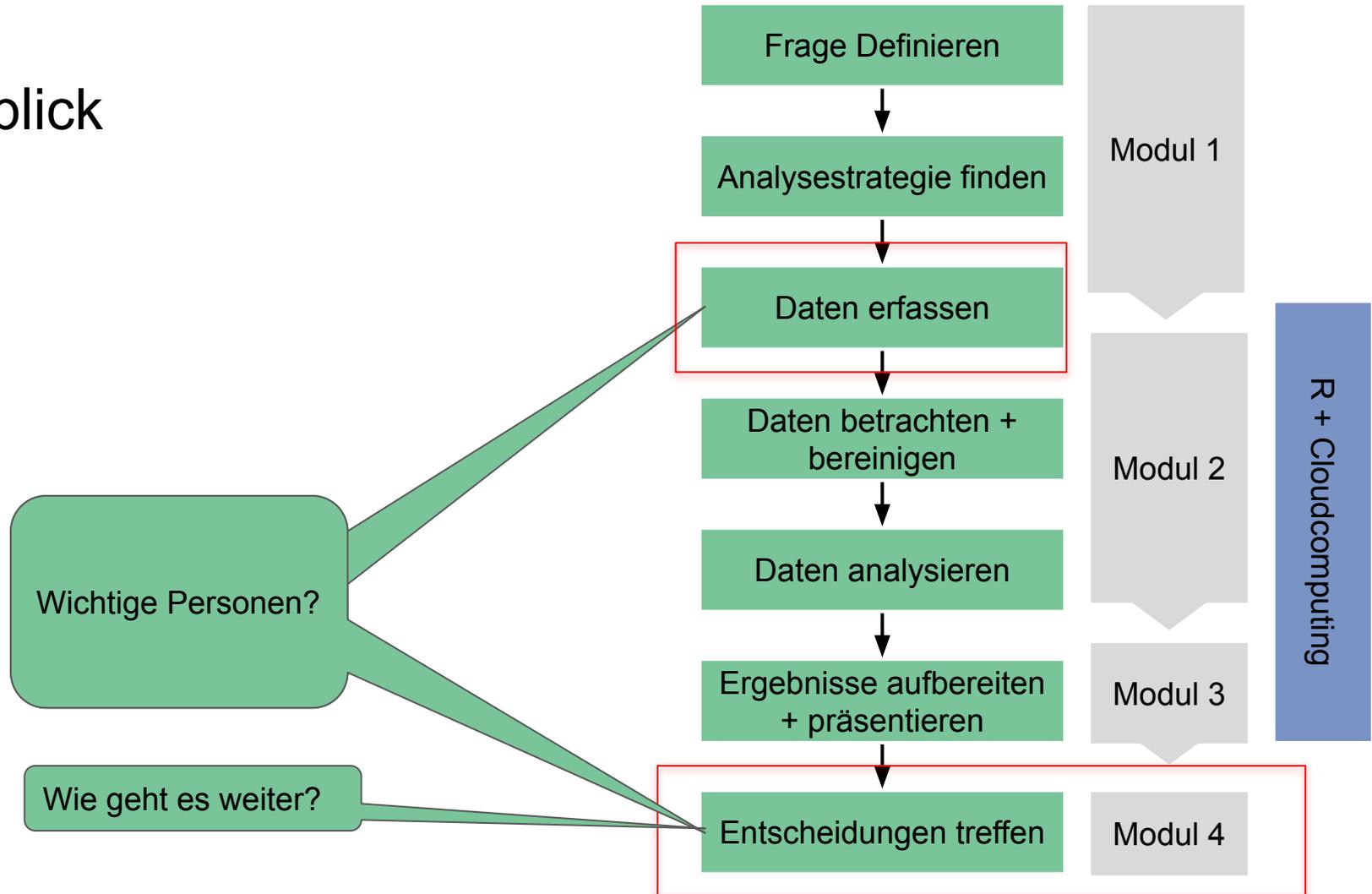
# Rückblick



# Rückblick



# Rückblick



# 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?

# Bedeutung für die Archive. Was folgt daraus?

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

Was ist Prominenz? (lokale Prominenz identifizieren?)

Denkanstoß/Vorschlag: “Prominenzindikator” für Auswahl bereitstellen.

Andere Möglichkeiten? Medieninteresse an Verfahren messen?

# Zukunft: Digitale Akte

Was ändert sich? Datenqualität verbessert? ...

# Teamarbeit

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



3 Slides pro Team:

- Ergebnisse
- Plan für Umsetzung
- Zukunftsvision: Was bräuchte man noch für das "perfekte" Archivierungs-System?

Vortrag: Was bisher geschah + Record Linkage	13:00 - 13:45
Pause	13:45 - 13:55
Teamarbeit: Ergebnisse + Präsentation fertigstellen	13:55 - 15:15
Pause	15:15 - 15:30
Teampräsentationen: 6-7 Minuten pro Team	15:30 - 16:10
Pause	16:10 - 16:15
Diskussion: Wie geht es weiter?	16:15 - 16:40
Abschlussrunde	16:40 - 17:00

# Team-Präsentationen

Team 1:

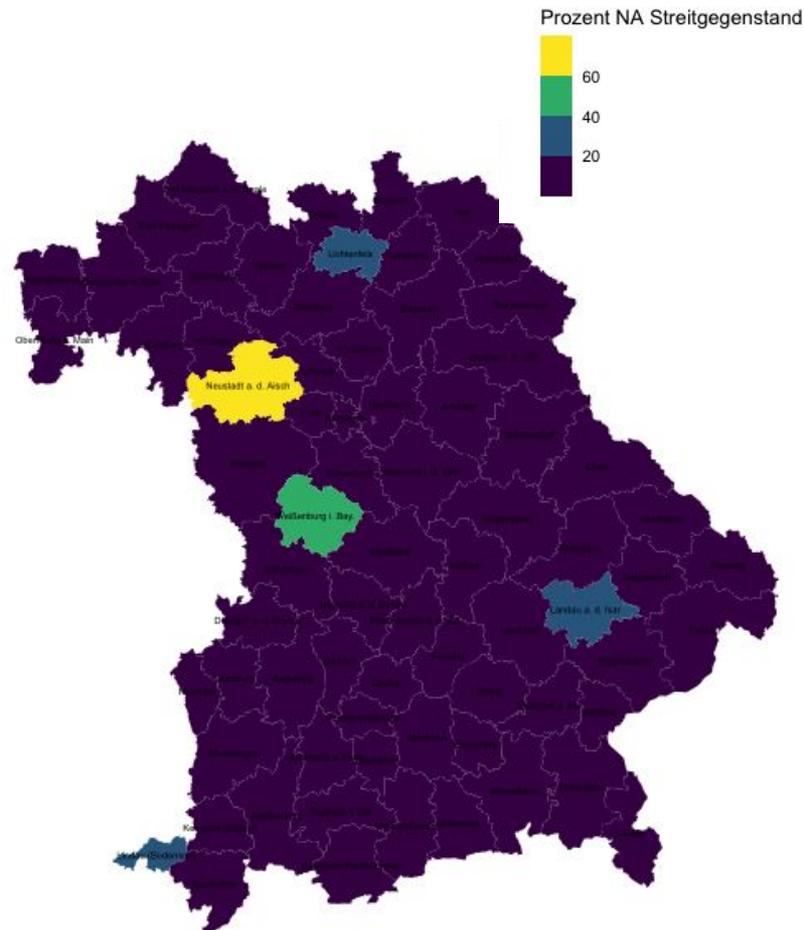
Wanhua Her, **Markus Schmalzl**, Hannah  
Hien, Dominique Mergen, Matthias Bogner,  
Heidi Seibold

# Ergebnisse

Streitgegenstand	Anzahl
Vollstreckungsabwehrklage	305
Beseitigung	362
Abgabe einer Willenserklärung	388
Entschädigung	444
Vornahme einer Handlung	589
Urheberrecht	737
Herausgabe	784
Schmerzensgeld	801
Feststellung	826
Zustimmung zur Mieterhöhung	1015
Unterlassung	1069
Beschlussanfechtung	1269
einstweiliger Verfügung	1331
Räumung und Forderung	1998
Räumung	3126
Räumung und Herausgabe	3938
Duldung	4876
NA	10065
Schadensersatz	28051
Forderung	63277

## Relevante Spalten:

- Gericht / Bezirk
- Streitwert
- Gesamtstreitgegenstand
- Streitgegenstand
- Sachgebiet
- Anzahl Termine
- Daten (Datum)
- Anzahl an Beteiligten
- Es ging voraus... (bisherige Verfahren)
- Informationen über die Person (Minderjährige, berühmte Personen, Beteiligungsart, etc.)
- ggf. weitere



# Umsetzung

## Datenqualität verbessern. Ideen:

- Klarstellen, was wichtig ist → Pflichtfelder
- Vorbelegung löschen
- Kommunizieren, in welchen Bezirken die Qualität niedrig ist
- Kategorisierung vereinfachen (weniger einfachere Kategorien)
- Kategorien sinnvoller gestalten (z.B. Forderung als Kategorie löschen)
- ~~Weiterbildungen für die Personen, die die Daten eintragen?~~
- ~~Relevanz von Archivierung deutlich machen, damit Motivation steigt~~
- Automatisierung, z.B. Vorausfüllung durch KI (wie bei ebay Kleinanzeigen)

## Außerdem:

- Zufallsauswahl implementieren
- Fortbildung: wie kann man das jetzt gelernte in die Fläche bringen?
- Austausch über Verwaltungsgrenzen hinweg weiterführen (diverse Teams benötigt, um gute Entscheidungen zu treffen → alle Kompetenzen an einem Tisch)

Anzeigendetails

Gebot / Gesuch  Ich biete  Ich suche

Titel

Kategorie Haus & Garten > Wohnzimmer [Kategorie ändern](#)

Art  ▾

Farbe (optional)  ▾

# Zukunftsvision

- **Weiteren Wissensaustausch:** Archivar:innen mit Skills in Data Science ausbilden.
- **Cloud-Plattform**, auf der man **einfach** und **sicher** zusammenarbeiten kann.
- **Aufträge:** Kollaborationen mit Expert:innen in Digitalisierung.
- **Qualitative hochwertige Daten** von den Daten-Produzenten.
- **Digitale Akte:** mit KI durchsuchen, um auch das “besondere” und nicht nur das “typische” zu finden.

Team 2:

Christian Pfrang, **Andreas Nestl**,  
Andreas Hutterer, Dominique Mergen,  
Jan Simson, Frauke Kreuter

# Ergebnisse:

**Ziel:** Interessensneutral die Tätigkeit des Staat abbilden ... womit sich Staat beschäftigt hat

**Grundgesamtheit** kennen  
und Populationswerte und **Ziehungswahrscheinlichkeiten** aufbewahren (Justizstatistik)

**Gesamtzahl der Akten:** 126.945  
"Certainty Strata" (Angeboten): 372

Reduktion der Population um Erledigungsgrund "Rücknahme" 16.520  
Quote etwa: 1-5%

Streitwert (>4000 JN) x Länge (>10, JN)  
**100% aus den besonderen**

Sachgebiet (17) x Erledigungsgrund (29) x  
6% alles was nicht besonders ist

	<10 Tage	>10
<4000 Euro	71.745	199
> 4000 Euro	12.250	<b>103</b>

# Umsetzung

## **Kurzfristig**

In der Cloudumgebung arbeiten - Strata definieren

Testlauf mit den Archivaren

Erweiterung auf andere Datenbestände

**Auszug aus dem Forumstar regelmäßig aufsetzen**

## **Umgebung klären**

Alternative 1 - Software installieren im Haus

Alternative 2 - Cloudumgebung Bayern (bite)

Alternative 3 - Datenraum Kultur oder Datenraum Archiv

# Zukunftsvision

Auf Richter verzichten und

E-Akten inhaltlich auslesen

- **Alle mit Pressemitteilungen**
- **Stichworte definieren**
  - Dieselskandal
  - Wirecard
  - Namen



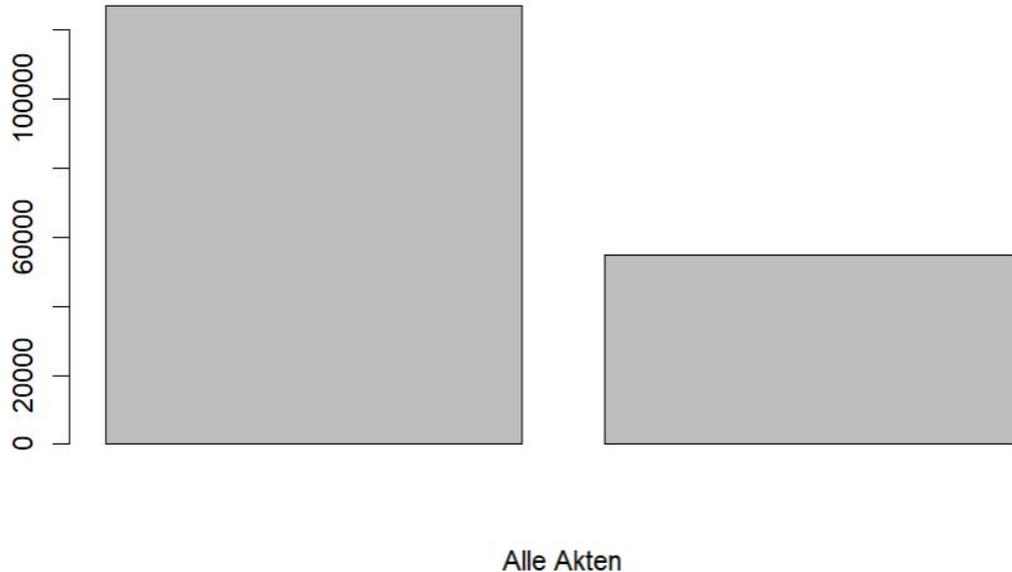
Team 3:

**Daniel Gietl, Michael Unger,**  
Johannes Stoiber, Tilman Janzarik,  
Malte Schierholz

## Schritt 1:

Berücksichtige nur Verfahren, bei denen ein Prozess tatsächlich stattgefunden hat

(Erledigungsgrund == "Endurteil", "Vergleich", "Beschluss", da dies am häufigsten ist)



Archivierung erfolgt immer, bei besonderen Fällen:

1. Besonders hervorstechende Verfahren (**Berühmte Persönlichkeiten, Rechtlich bedeutsame Fälle (Präzedenzfall)**)
  - a. Manuelle Erfassung erforderlich
  - b. Technisch in eigener Schicht, wird mit Wskt. 1 gezogen (Archivsachenvermerk)
2. Weitere bedeutsame Verfahren,
  - a. **`Dauer des Verfahrens in Tagen`  $\geq 1500$**  Tage (menschliche Entscheidung, da Umfang von ca. 150 Fällen brauchbar ist)
  - b. dann muss es sich um ein wichtiges Verfahren handeln, welches in jedem Fall archiviert wird.
  - c. Alternative (nicht zu bevorzugende) Ansätze: Lange **Aufbewahrungsfrist über 50 Jahre** oder ein besonders **hoher Streitwert** (in Abhängigkeit von Verfahrensdauer & Streitgegenstand) oder Prozesse bei denen ein **Gutachten eingeholt** wurde (Info liegt in einer anderen Datenbank und würde eine Verknüpfung erfordern) deuten ebenfalls auf ein bedeutsames Verfahren hin.

## Zufallsziehung (für typische Fälle)

- Einfache Zufallsstichprobe getrennt aus jedem Bezirk stellt flächendeckende Abdeckung sicher
- Stichprobenumfang: Sollen *gleich viele* Akten von jedem Gericht gezogen werden oder soll die Anzahl *proportional* zur Anzahl der Akten sein?
  - Kompromisslösung: Stichprobengröße = Wurzel(N) für jedes Gericht, da der Grenznutzen einer zusätzlichen Akte mit steigender Stichprobengröße abnimmt.

A tibble: 74 × 3

<b>strata</b> <chr>	<b>akten_pro_stratum</b> <int>	<b>anzahl_akten_pro_schicht</b> <dbl>
Aichach	457	22
Amberg	489	23
Ansbach	846	30
Aschaffenburg	690	27
Augsburg	2449	50
Bad Kissingen	297	18
Bad Neustadt a. d. Saale	244	16
Bamberg	796	29
Bayreuth	811	29
besonderes Verfahren	49	49



# Umsetzung

- Zusätzliches Feature in Forumstar (wenn vom Ministerium Geld zur Programmierung gegeben wird)
- Zeitige Kennzeichnung (am 2.1. des Folgejahres) zur Generierung/Kennzeichnung der zu archivierenden Akten

# Zukunftsvision

## Bessere Meta-Datenerfassung

- Beteiligung Pressestelle
- Sammlung entscheidender Fälle aus Juris-Datenbank / Juristische Fachzeitschriften?
- Weitere externe Ressourcen?
- Tenor des Urteils (sehr strukturierter Text, rechtliche Relevanz ableitbar?) - KI?

Was passiert bei E-Akten? Haben wir dann Zugriff auf detaillierte Akteninhalte?

## Team 4:

Christian Brück, Claudia Kalesse, Franziska  
Armbruster, Nicola Humphreys, Gunther  
Friedrich, Ulrike Claudia Hofmann, Dominik  
Mayer, **Marcel Neunhoeffer**

# Ergebnisse

In den vorliegenden Daten haben wir **zu wenig Information** für die sinnvolle Auswahl von Akten zur Archivierung.

Eine **einfache Zufallsstichprobe ist nicht ausreichend** für die Auswahl im Einklang mit der Aussonderungsbekanntmachung.

## **ABER**

Wir können/wollen eine einfache Zufallsstichprobe nutzen um **besser zu verstehen was besondere Akten ausmacht**.

Mit den gelernten Kriterien können Akten **in Zukunft** mit Hilfe einer **geschichteten Zufallsstichprobe** ausgewählt werden.

- Akten die auf die Beschreibung in der Aussonderungsbekanntmachung passen werden immer (mit höhere Wahrscheinlichkeit) ausgewählt.
- Darüber hinaus werden weitere Akten mit geringerer Auswahlwahrscheinlichkeit zufällig gezogen.  
→ So können zeittypische Verfahren abgedeckt werden

# Umsetzung

1. Um zu lernen inwieweit wir mit den ForumStar Daten archivwürdige Akten identifizieren können, betrachten wir eine kleine Zufallsstichprobe (n = 100) an Akten in einem **Pilotprojekt**. Insbesondere interessieren wir uns für:
  - a. Gibt es einen **Zusammenhang zwischen Bedeutung des Prozesses** und der **Anzahl der Prozesstage** oder **anderen inhaltlichen Kriterien** (Umfang des Prozesses [Blattzahl], Anzahl Gutachten, Anzahl der Beteiligten...)
  - b. Gibt es einen **Zusammenhang zwischen diesen Kriterien** und **Medienberichterstattung** über den Prozess. (Wunsch in Zukunft Berichterstattung zu digitalen Akten anfügen. → Externe Datenquellen)
2. Mit den **Ergebnissen** aus dem **Pilotprojekt** ist eine **Anpassung der übermittelten Daten** von den Gerichten möglich, sodass die **Auswahl zur Archivierung in Zukunft vereinfacht wird**.
3. Zur geschichteten Zufallsstichprobe: Neben den von Gerichten mit Hilfe der oben entwickelten Kriterien mit Archivsachenvermerk gekennzeichneten Akten, werden – zum Beispiel am Ende jedes Geschäftsjahres – mit Hilfe einer Zufallsstichprobe einige Akten (zeittypische Verfahren) vor der anstehenden Ausdünnung zusätzlich mit dem Vermerk gekennzeichnet. → Für diese Akten könnte eine spätere Abgabe an die Archive in Frage kommen.



Wie geht es weiter?

# Diskussion

Wie können die Ergebnisse in bestehende Prozesse eingebaut werden?

Wie könnte man die Stichprobenziehung in die Praxis einbauen? Wer braucht was, wann, wo?

Wie kann man es Technologie-unabhängig machen (falls ForumSTAR abgestellt wird)?

Was können wir mal machen, wenn die digitale Akte kommt?

Abschlussrunde

# Unser Fazit

- Tolles Pilotprojekt → wir haben viel gelernt
- Es gibt noch viel zu tun
- Wir können noch einiges verbessern: v.a. Cloud-Zugang
- Gemischte Expertise in Team: super!
- Erste "easy wins" durch Zugang zu den Daten in neuer Form und mit neuem Blickwinkel

# Was nehmen Sie aus dieser Workshopreihe mit?

→ **Feedback-Formular**

(auf Papier oder unter  
<https://soda.limesurvey.net/981929>)



# Was kommt noch von uns

- Website wird ergänzt mit Präsentationen
- Unsere Erkenntnisse Aufschreiben: z.B. für Archiv-Zeitschrift, Harvard Data Science Review, ...
- Nach dem Kurs ist vor dem Kurs. Termine: TBD