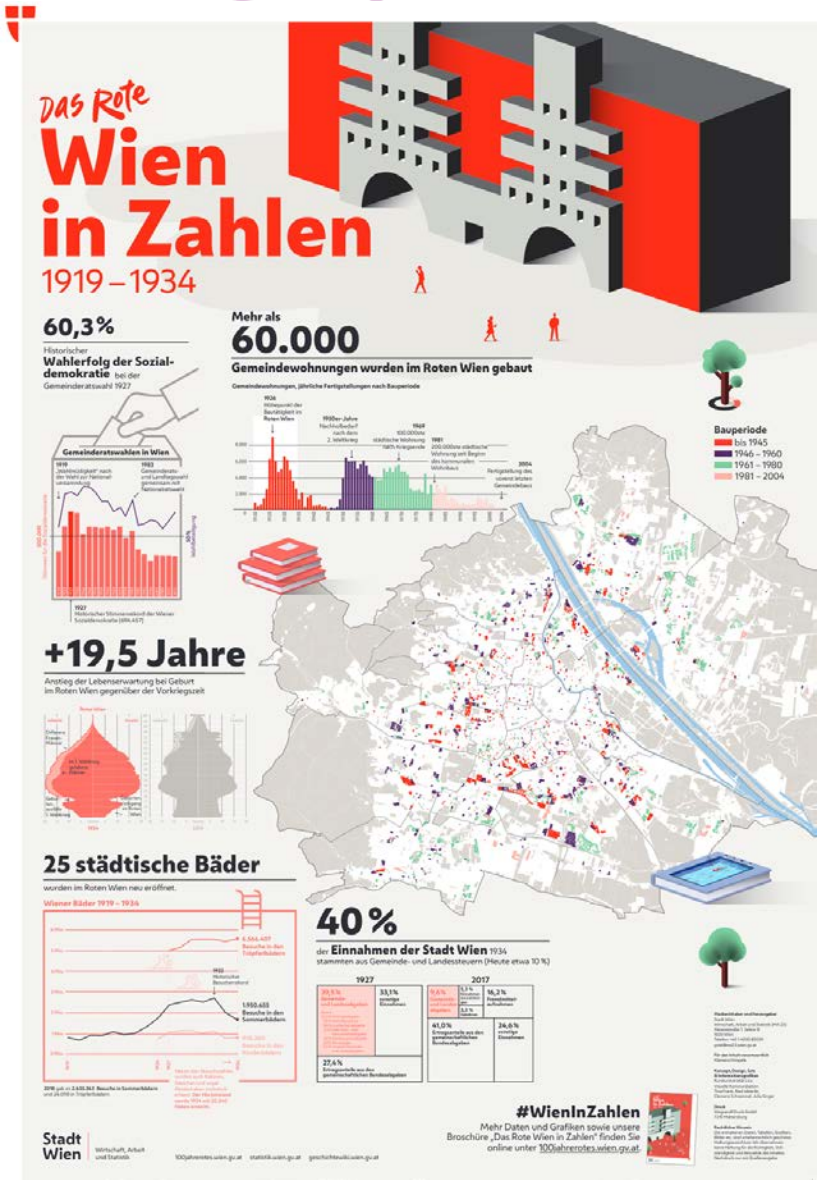
KI-generierte synthetische Daten als Data Governance Tool

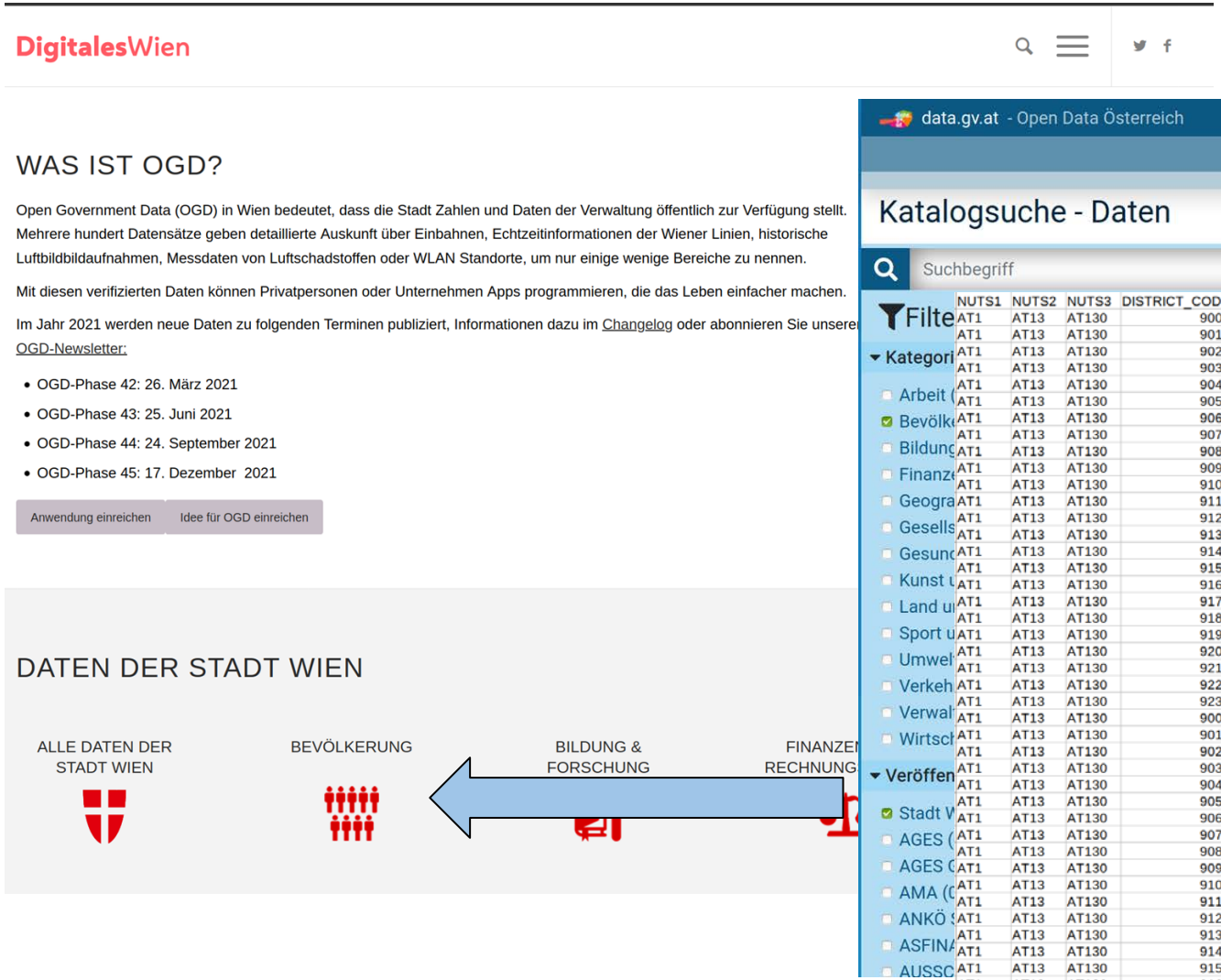
Roman Seidl
Klaudius Kalcher

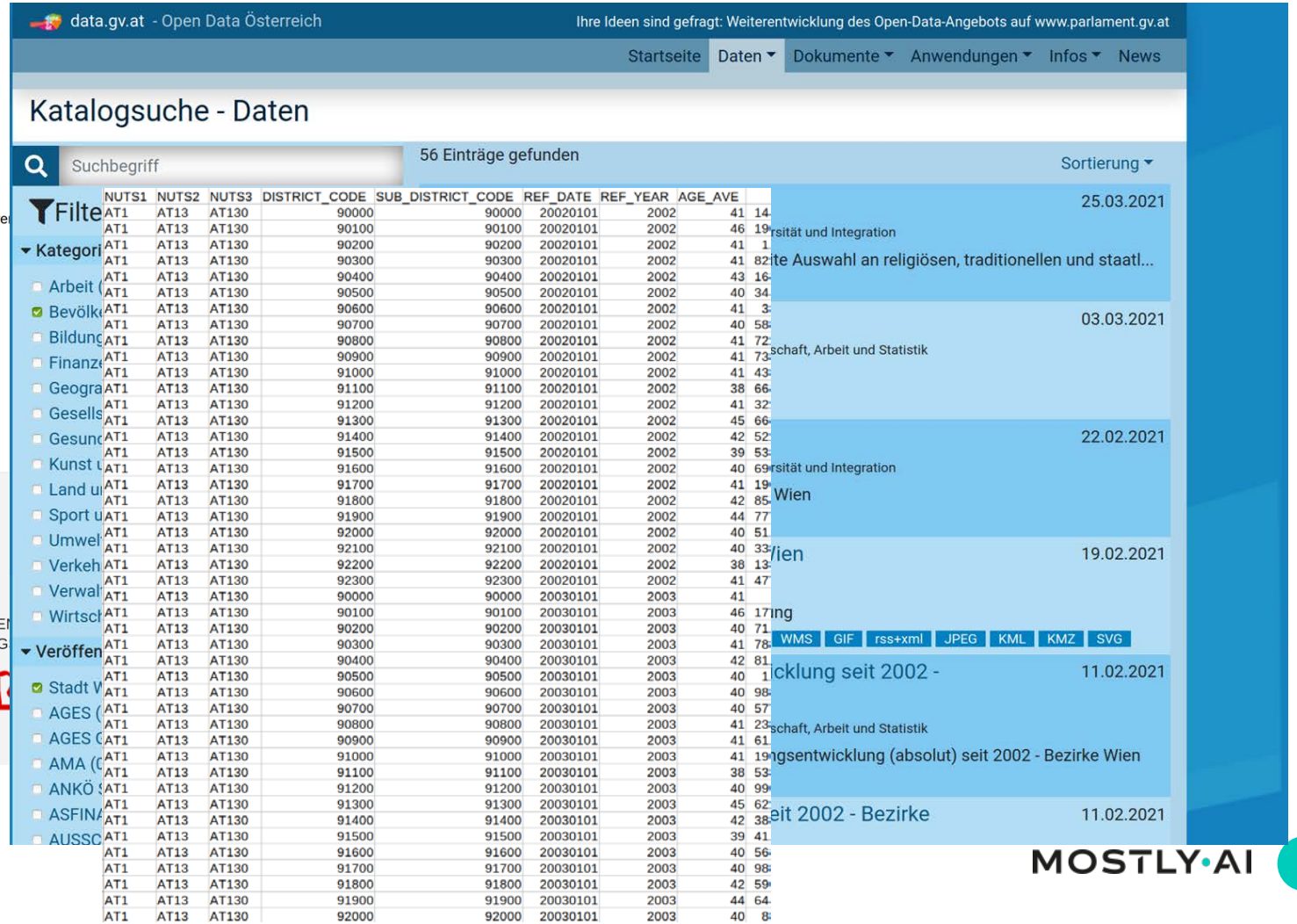
MOSTLY·AI

Wirtschaft, Arbeit und Statistik - Themen



Ziel: Open Data vs Privacy





Privatsphäre in Big Data - Herausforderungen

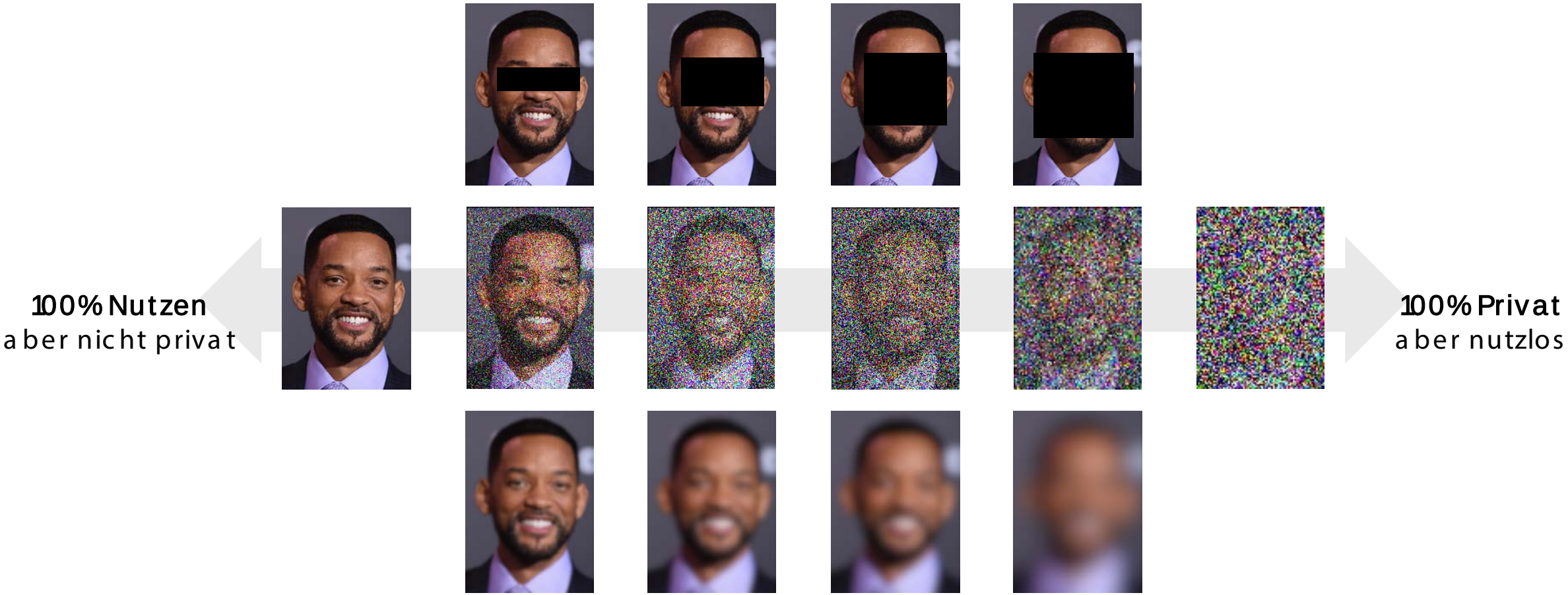
Klassische Anonymisierung schlägt fehl

- Die meiste Information wird zerstört
- Weniger als 5% der Information kann erhalten werden
- Dennoch oft hohes Deanonymisierungsrisiko

Feingranulare Daten nicht zugänglich

Klassische Anonymisierungsverfahren Schlagen Fehl

Der Trade- off zwischen Privatsphäre und Nutzen ist riesig, schränkt Verwendung der Daten stark ein



MOSTLY GENERATE Synthetic Data Platform

1 **Original behavioral customer data**

- highly valuable, but
- privacy-sensitive
- locked away (GDPR / CCPA)

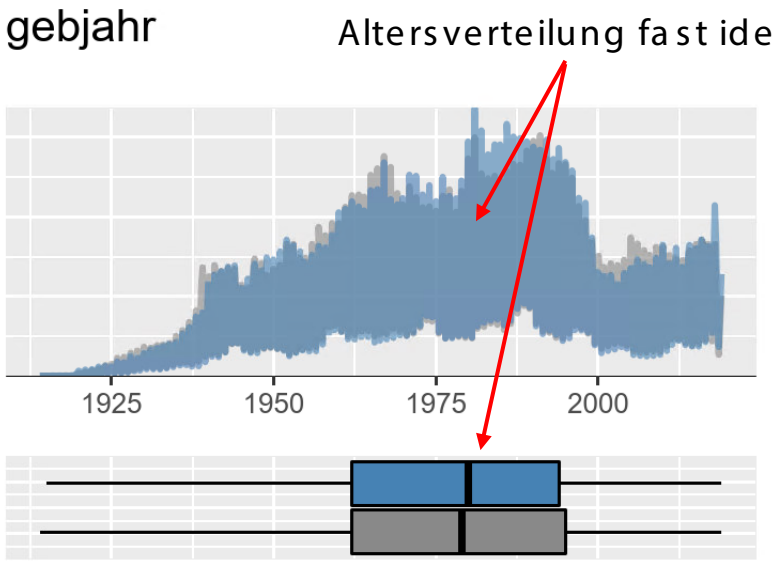
2 **Synthetization**

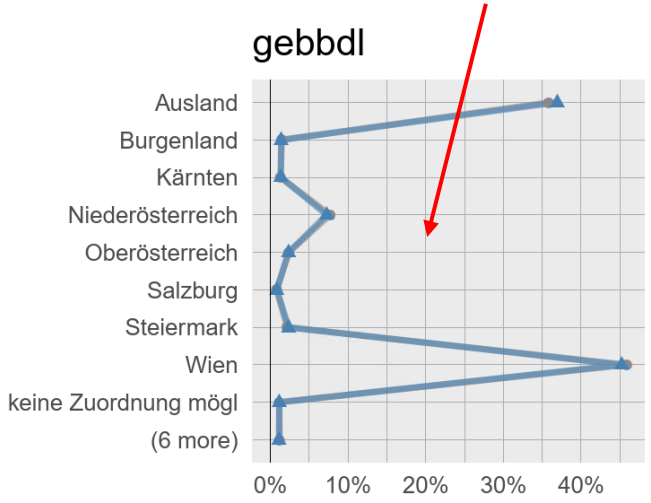
- API integration
- Cloud bucket support
- Deployment possibilities
  - on-prem
  - your cloud environment
  - SaaS

3 **Synthetic behavioral customer data**

- as-good-as-real
- fully anonymous
- free to use (GDPR / CCPA compliant)

Eine synthetische Wiener Bevölkerung





Legende

Originaldaten

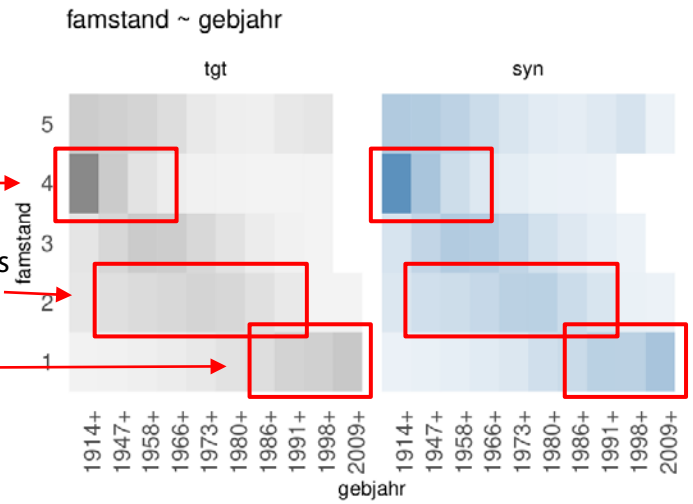
Synthetische Daten

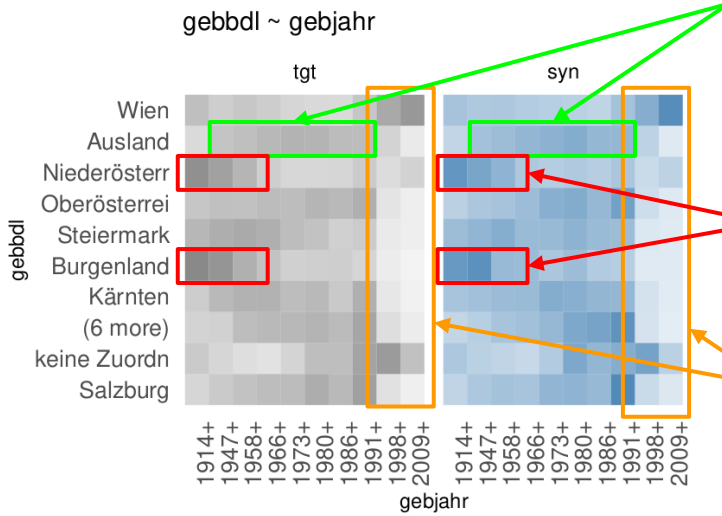
Zusammenhänge zwischen Attributen

Verwitwete älter

Verheiratete mittleren alters

Junge meist ledig



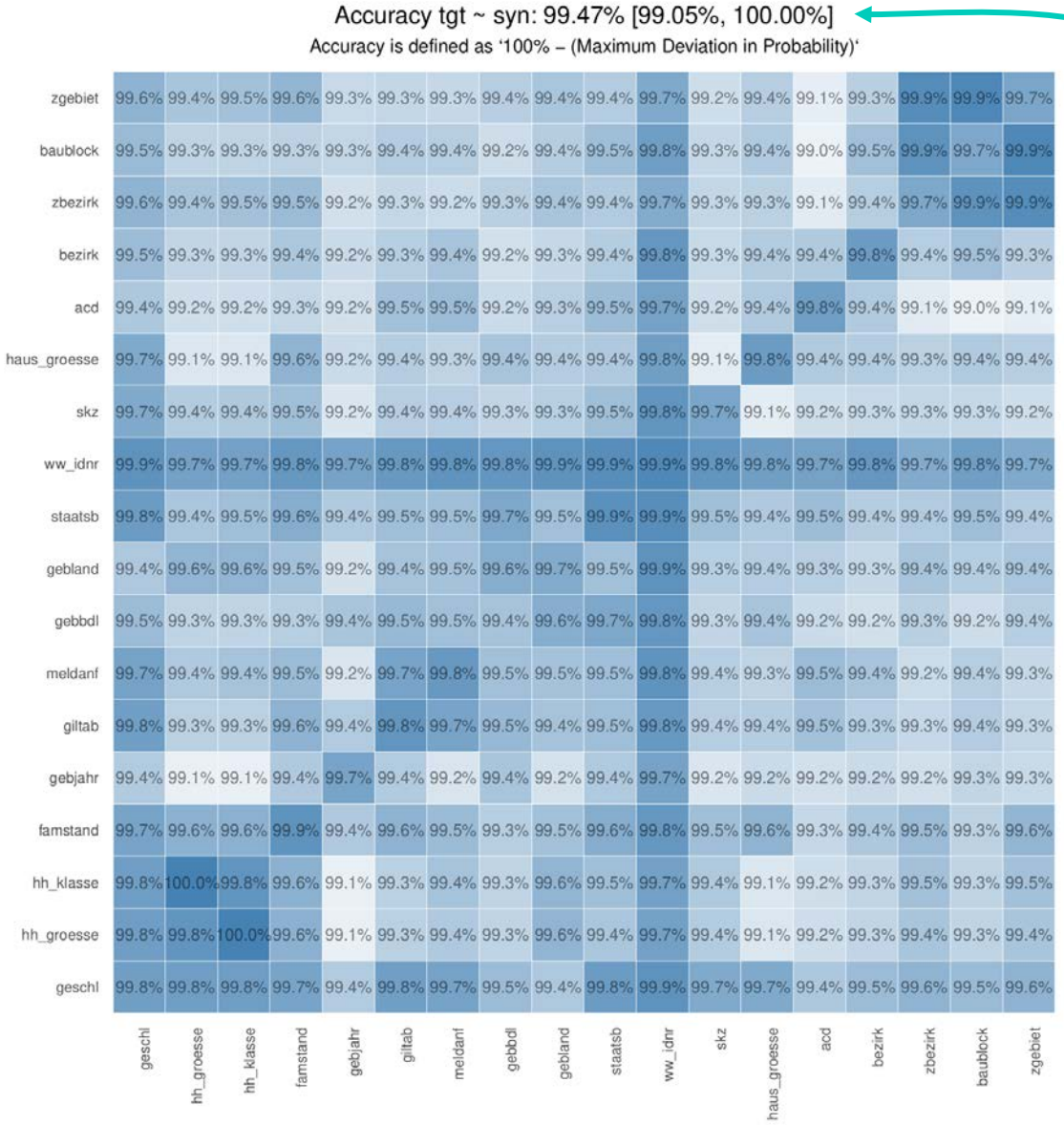


Im Ausland geborene Wiener überwiegend 20-60

Mehr Ältere in NÖ oder Bgld als in anderen Bundesländern geboren

Wiener Kinder überwiegend in Wien geboren

Eine synthetische Wiener Bevölkerung (II)



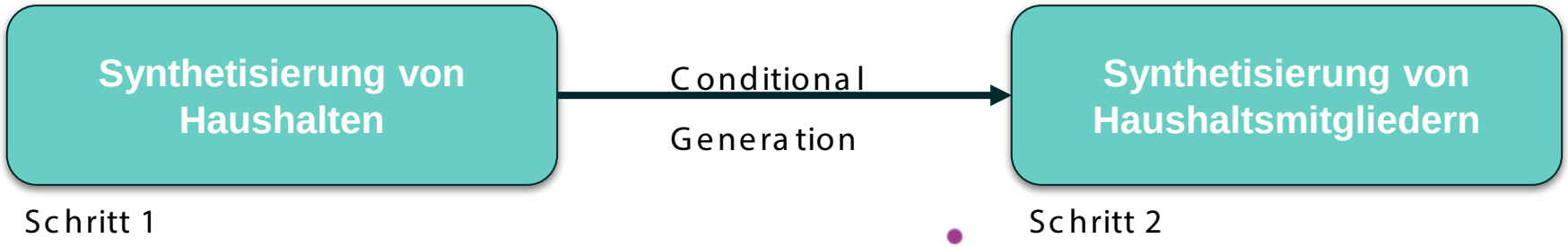
Alle univariaten Verteilungen und bivariaten Zusammenhänge zu mind. 99% erhalten

(im Durchschnitt zu 99.47%)

Challenges der Bevölkerungsstatistik

- Personen gehören zu Haushalten
- **Privatsphäre von Haushalten** als Ganzes ist ebenso zu schützen wie die von Individuen

Herangehensweise:



Generierung von konsistenten Haushalten

Schritt 1: Synthetische Haushalte generieren

Tabelle A: HAUSHALTE

Key	
ID	MfOS6T713cL4Yc0d68AbdI29
BEZIRK	16
ZGEBIET	1603
ZBEZIRK	16033
BAUBLOCK	16033001
ACD	48089
WW	0
SKZ	903736
HH_KLASSE	2
HH_GROESSE	2
HAUS_GROESSE	18

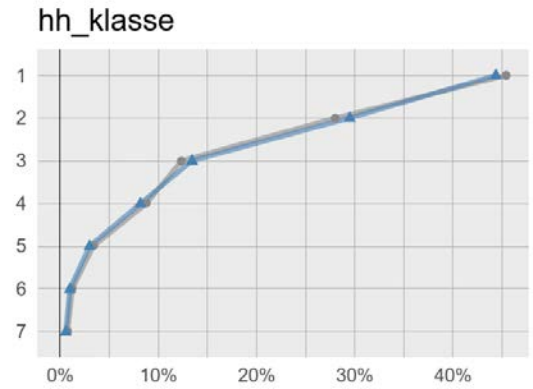
Tabelle B: PERSONEN

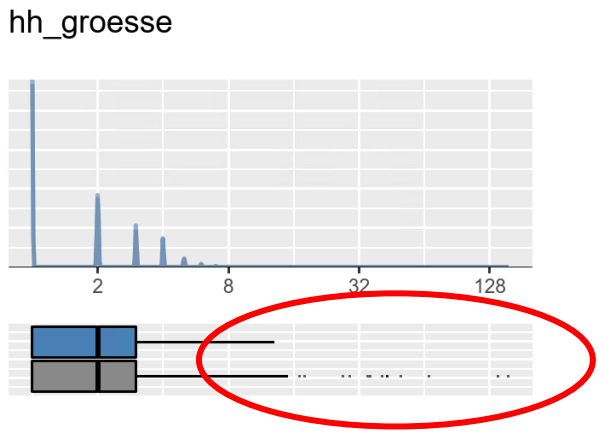
Key		
HAUSHALTE_ID	MfOS6T713cL4Yc0d68AbdI29	MfOS6T713cL4Yc0d68AbdI29
GESCHL	2	1
GEBJAHR	1966	1962
GEBBDL	Wien	Niederösterreich
GEBLAND	40	40
STAATSB	40	40
FAMSTAND	2	2
MELDANF	1965-11-13	1997-08-03

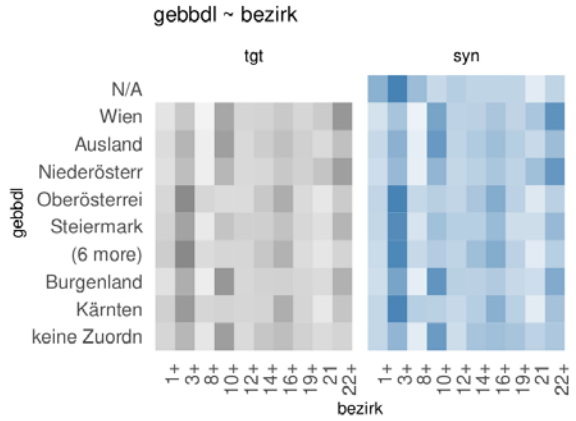
Hier: Haushaltsgröße = 2, Bezirk = 16
→ Zwei Personen wurden generiert, die plausibel zusammen ein 2-Personen-Haushalt in Ottakring sein könnten

Schritt 2: Zu jedem Haushalt wird eine zu den Attributen des Haushalts passende Gruppe von synthetischen Personen generiert

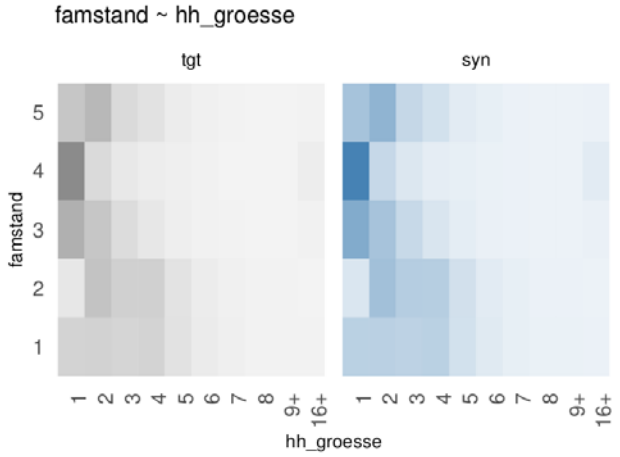
Statistiken von Haushalten

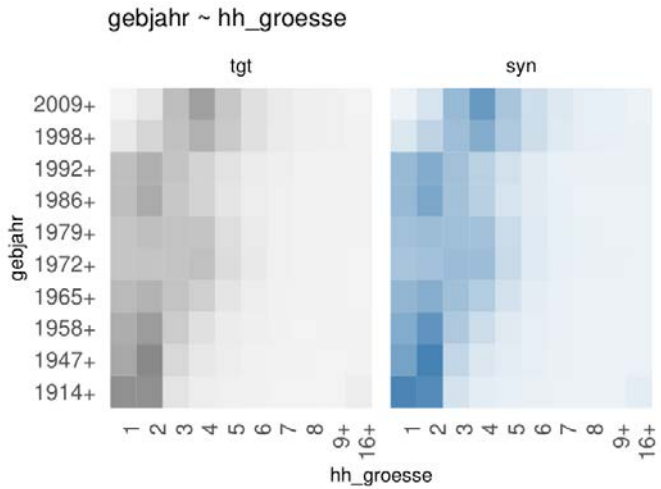


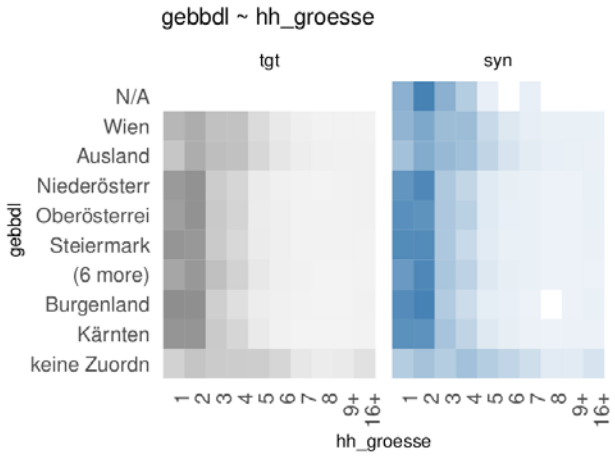




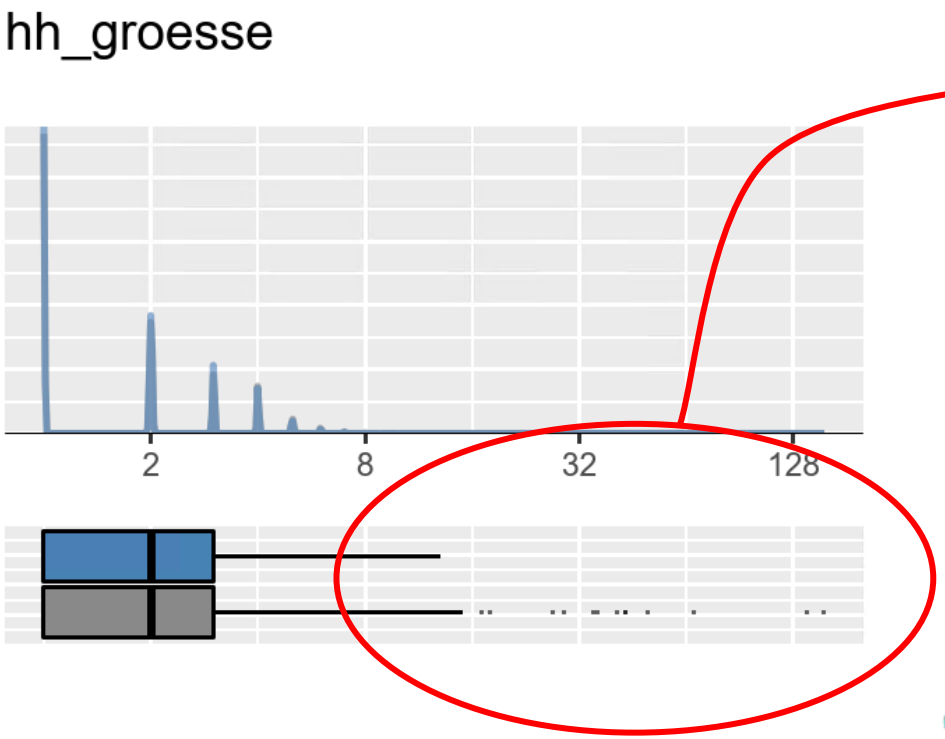
Auch Statistiken auf Haushaltsebene sehr gut getroffen







Statistiken von Haushalten



Ausreißer werden bei Synthetisierung besonders geschützt

Hier aber ein Spezialfall - "**Anstaltshaushalte**"
z.B. Justizvollzugsanstalten, Pflegeheime

Zählen in der Statistik zwar als ein Haushalt,
ihre Existenz muss aber nicht geheim gehalten werden

Mögliches **Vorgehen**, um diese Ausreißer doch zu erhalten:

Haushalte beibehalten mit tatsächlichem Ort & Größe,
Personen darin synthetisch generieren

Erhalten bestimmter exakter Zusammenhänge

Exakt erhalten werden sollten

- **Anzahl der Haushalte** pro Bezirk
- **Gesamtbevölkerung** pro Bezirk
- Zuordnung geografischer Einheiten

} nicht sensibel

Angepasste Lösung:
Erstellung **teilsynthetischer** Daten

- Nicht-sensible Information wird aus den Originaldaten übernommen
- Haushalte werden synthetisiert
- Personen werden synthetisiert

Zuordnung von bezirk, zbezirk, zgebiet, baublock

		stimmt nicht überein mit			
		bezirk	**zbezirk**	**zgebiet**	**baublock**
Wert aus	**bezirk**	0,000%	0,001%	0,005%	2,645%
	zbezirk	0,001%	0,000%	0,013%	3,307%
	zgebiet	0,005%	0,013%	0,000%	3,472%
	baublock	2,645%	3,307%	3,472%	0,000%

Synthetisierungsablauf im Überblick

Schritt 0

Anzahl Haushalte
Anzahl Personen
} pro Bezirk
Zuordnung Baublocks zu Bezirken

Schritt 1

Haushalte mit
exakter Geografie,
Hausgröße, etc.

Schritt 2

Personen mit
Geburtsjahr, Geschlecht,
Familienstand, Geburtsort,
Staatsbürgerschaft, etc.

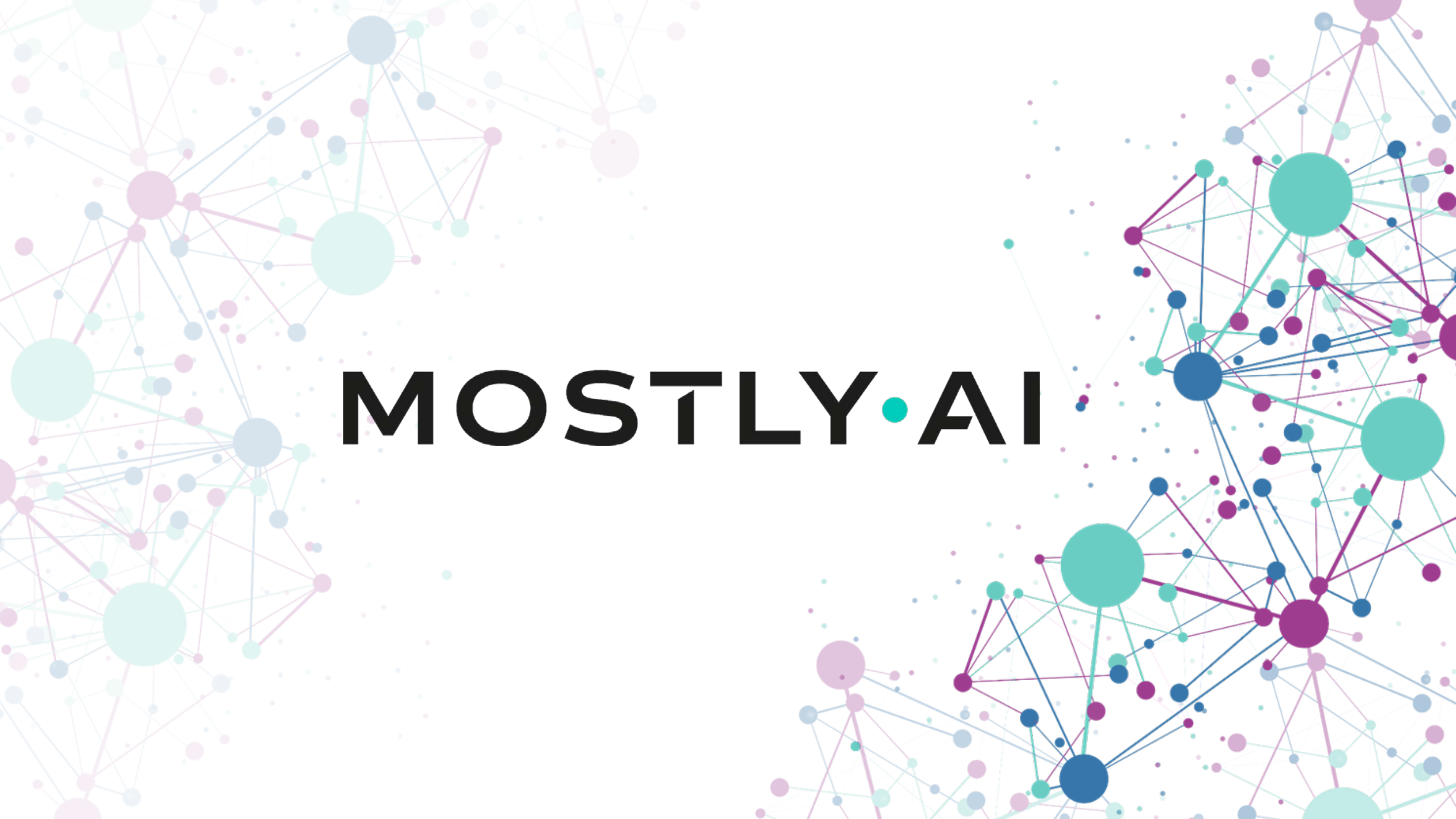
Nächste Schritte & Ziele

Bis Ende 2021

- Publizierbaren **Testdatensatz** zu Daten von 2020 erstellen
- Für diesen Datensatz und die Methodik ein **Review** durchführen

Längerfristiges Ziel

- **Publikation**
- Weitere Anwendungen identifizieren


MOSTLY·AI