

# KI-generierte synthetische Daten als Data Governance Tool

*Roman Seidl*

*Klaudius Kalcher*

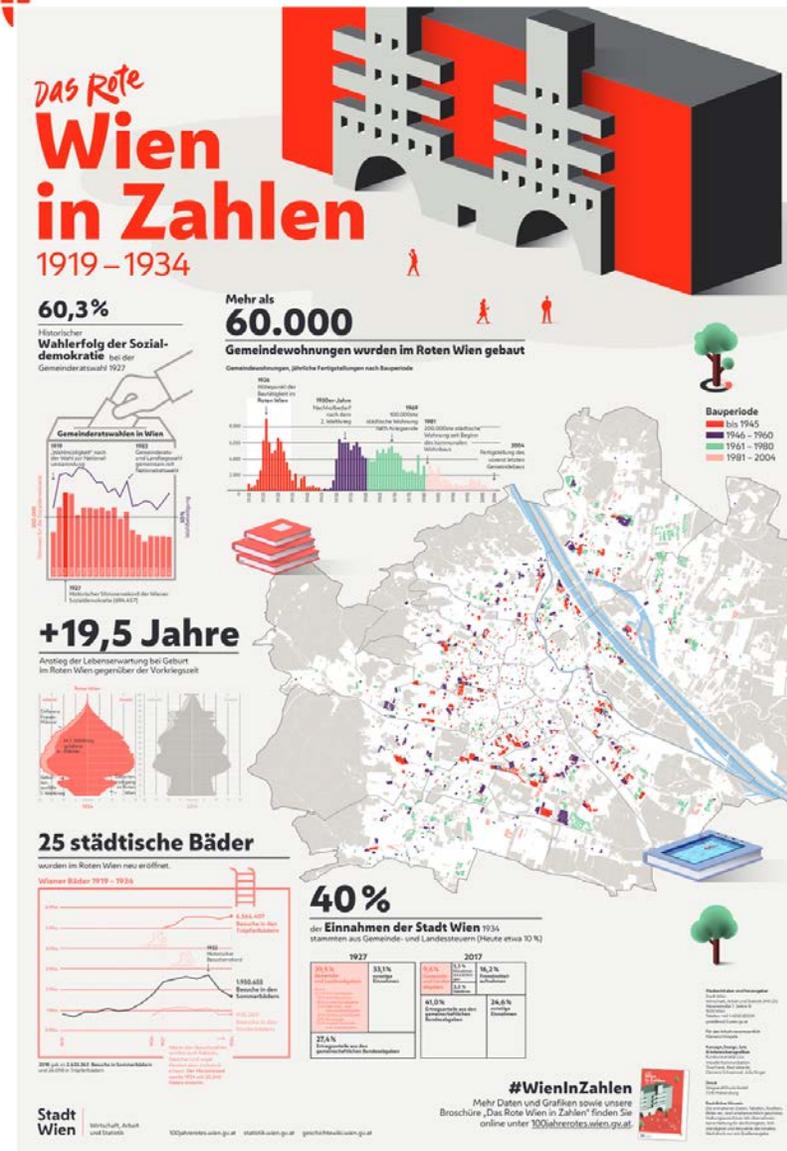


**Stadt  
Wien**

Wirtschaft, Arbeit  
und Statistik

**MOSTLY·AI**

# Wirtschaft, Arbeit und Statistik - Themen



# Ziel: Open Data vs Privacy

## WAS IST OGD?

Open Government Data (OGD) in Wien bedeutet, dass die Stadt Zahlen und Daten der Verwaltung öffentlich zur Verfügung stellt. Mehrere hundert Datensätze geben detaillierte Auskunft über Einbahnen, Echtzeitinformationen der Wiener Linien, historische Luftbildaufnahmen, Messdaten von Luftschadstoffen oder WLAN Standorte, um nur einige wenige Bereiche zu nennen.

Mit diesen verifizierten Daten können Privatpersonen oder Unternehmen Apps programmieren, die das Leben einfacher machen.

Im Jahr 2021 werden neue Daten zu folgenden Terminen publiziert, Informationen dazu im [Changelog](#) oder abonnieren Sie unsere [OGD-Newsletter](#).

- OGD-Phase 42: 26. März 2021
- OGD-Phase 43: 25. Juni 2021
- OGD-Phase 44: 24. September 2021
- OGD-Phase 45: 17. Dezember 2021

Anwendung einreichen

## DATEN DER STADT WIEN

ALLE DATEN DER STADT WIEN



BEVÖLKERUNG



BILDUNG & FORSCHUNG



FINANZEN & RECHNUNG



data.gv.at - Open Data Österreich Ihre Ideen sind gefragt: Weiterentwicklung des Open-Data-Angebots auf [www.parlament.gv.at](http://www.parlament.gv.at)

Startseite Daten Dokumente Anwendungen Infos News

### Katalogsuche - Daten

Suchbegriff  56 Einträge gefunden Sortierung ▾

Filter	NUTS1	NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	REF_DATE	REF_YEAR	AGE_AVE		
	AT1	AT13	AT130	90000	90000	20020101	2002	41 14		25.03.2021
	AT1	AT13	AT130	90100	90100	20020101	2002	46 19	rsität und Integration	
	AT1	AT13	AT130	90200	90200	20020101	2002	41 1	te Auswahl an religiösen, traditionellen und staatl...	
	AT1	AT13	AT130	90300	90300	20020101	2002	41 82		
	AT1	AT13	AT130	90400	90400	20020101	2002	43 16		
	AT1	AT13	AT130	90500	90500	20020101	2002	40 34		
	AT1	AT13	AT130	90600	90600	20020101	2002	41 3		
	AT1	AT13	AT130	90700	90700	20020101	2002	40 58		03.03.2021
	AT1	AT13	AT130	90800	90800	20020101	2002	41 72	schaft, Arbeit und Statistik	
	AT1	AT13	AT130	90900	90900	20020101	2002	41 73		
	AT1	AT13	AT130	91000	91000	20020101	2002	41 43		
	AT1	AT13	AT130	91100	91100	20020101	2002	38 66		
	AT1	AT13	AT130	91200	91200	20020101	2002	41 32		
	AT1	AT13	AT130	91300	91300	20020101	2002	45 66		
	AT1	AT13	AT130	91400	91400	20020101	2002	42 52		22.02.2021
	AT1	AT13	AT130	91500	91500	20020101	2002	39 53		
	AT1	AT13	AT130	91600	91600	20020101	2002	40 69	rsität und Integration	
	AT1	AT13	AT130	91700	91700	20020101	2002	41 19	Wien	
	AT1	AT13	AT130	91800	91800	20020101	2002	42 85	Wien	
	AT1	AT13	AT130	91900	91900	20020101	2002	44 77		
	AT1	AT13	AT130	92000	92000	20020101	2002	40 51		
	AT1	AT13	AT130	92100	92100	20020101	2002	40 33	Wien	19.02.2021
	AT1	AT13	AT130	92200	92200	20020101	2002	38 13	Wien	
	AT1	AT13	AT130	92300	92300	20020101	2002	41 47		
	AT1	AT13	AT130	90000	90000	20030101	2003	41		
	AT1	AT13	AT130	90100	90100	20030101	2003	46 17	Wien	
	AT1	AT13	AT130	90200	90200	20030101	2003	40 71	WMS GIF rss+xml JPEG KML KMZ SVG	
	AT1	AT13	AT130	90300	90300	20030101	2003	41 78	icklung seit 2002 -	11.02.2021
	AT1	AT13	AT130	90400	90400	20030101	2003	42 81		
	AT1	AT13	AT130	90500	90500	20030101	2003	40 1		
	AT1	AT13	AT130	90600	90600	20030101	2003	40 98		
	AT1	AT13	AT130	90700	90700	20030101	2003	40 57		
	AT1	AT13	AT130	90800	90800	20030101	2003	41 23	schaft, Arbeit und Statistik	
	AT1	AT13	AT130	90900	90900	20030101	2003	41 61	ngsentwicklung (absolut) seit 2002 - Bezirke Wien	
	AT1	AT13	AT130	91000	91000	20030101	2003	41 19		
	AT1	AT13	AT130	91100	91100	20030101	2003	38 53		
	AT1	AT13	AT130	91200	91200	20030101	2003	40 99		
	AT1	AT13	AT130	91300	91300	20030101	2003	45 62		
	AT1	AT13	AT130	91400	91400	20030101	2003	42 38	eit 2002 - Bezirke	11.02.2021
	AT1	AT13	AT130	91500	91500	20030101	2003	39 41		
	AT1	AT13	AT130	91600	91600	20030101	2003	40 56		
	AT1	AT13	AT130	91700	91700	20030101	2003	40 98		
	AT1	AT13	AT130	91800	91800	20030101	2003	42 59		
	AT1	AT13	AT130	91900	91900	20030101	2003	44 64		
	AT1	AT13	AT130	92000	92000	20030101	2003	40 8		

# Privatsphäre in Big Data - Herausforderungen

## Klassische Anonymisierung schlägt fehl

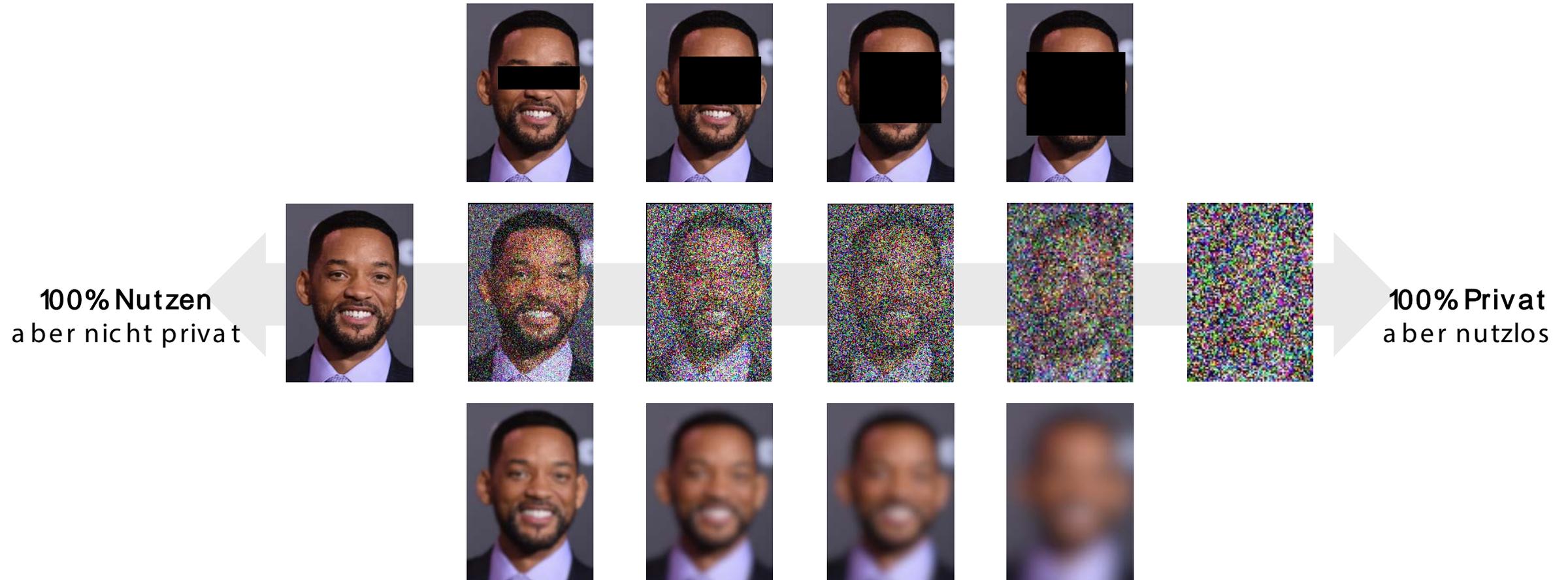
- Die meiste Information wird zerstört
- Weniger als 5% der Information kann erhalten werden
- Dennoch oft hohes Deanonymisierungsrisiko

## 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

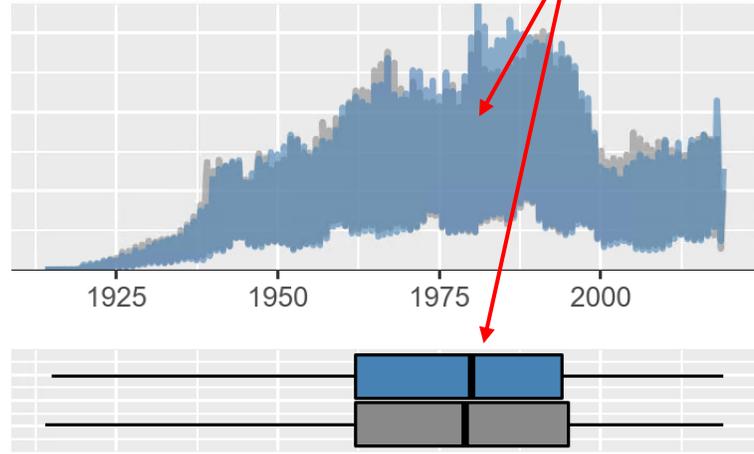
## Synthetic behavioral customer data

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

# Eine synthetische Wiener Bevölkerung

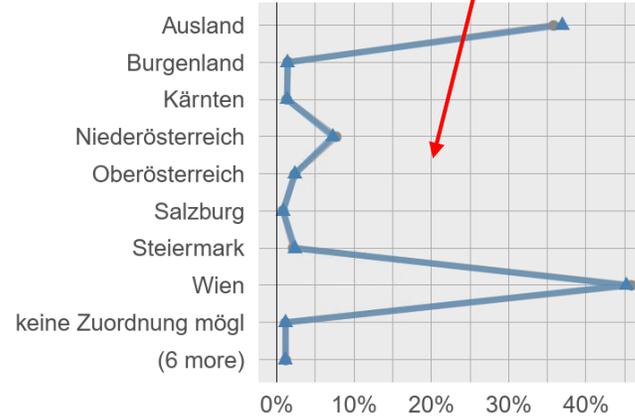
gebjahr

Altersverteilung fast identisch

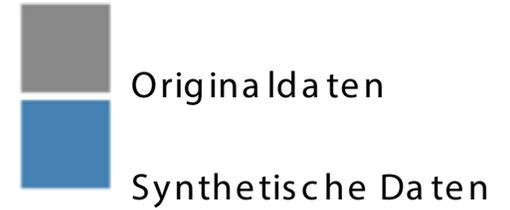


Herkunftsverteilung fast identisch

gebddl



Legende



## Zusammenhänge zwischen Attributen

Verwitwete älter

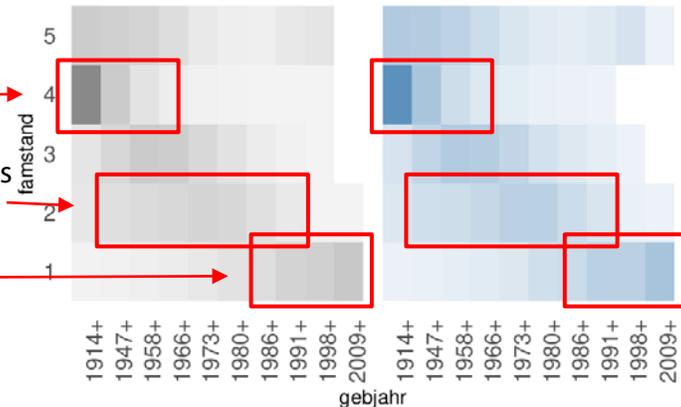
Verheiratete mittleren alters

Junge meist ledig

famstand ~ gebjahr

tgt

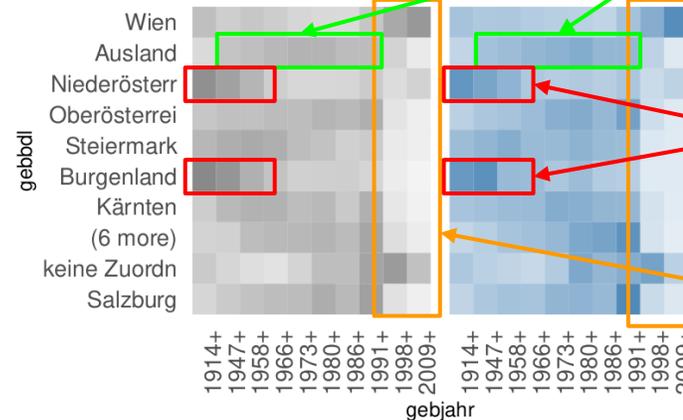
syn



gebddl ~ gebjahr

tgt

syn



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)

Accuracy tgt ~ syn: 99.47% [99.05%, 100.00%]  
 Accuracy is defined as '100% - (Maximum Deviation in Probability)'

zgebiet	99.6%	99.4%	99.5%	99.6%	99.3%	99.3%	99.3%	99.4%	99.4%	99.4%	99.7%	99.2%	99.4%	99.1%	99.3%	99.9%	99.9%	99.7%
baublock	99.5%	99.3%	99.3%	99.3%	99.3%	99.4%	99.4%	99.2%	99.4%	99.5%	99.8%	99.3%	99.4%	99.0%	99.5%	99.9%	99.7%	99.9%
zbezirk	99.6%	99.4%	99.5%	99.5%	99.2%	99.3%	99.2%	99.3%	99.4%	99.4%	99.7%	99.3%	99.3%	99.1%	99.4%	99.7%	99.9%	99.9%
bezirk	99.5%	99.3%	99.3%	99.4%	99.2%	99.3%	99.4%	99.2%	99.3%	99.4%	99.8%	99.3%	99.4%	99.4%	99.8%	99.4%	99.5%	99.3%
acd	99.4%	99.2%	99.2%	99.3%	99.2%	99.5%	99.5%	99.2%	99.3%	99.5%	99.7%	99.2%	99.4%	99.8%	99.4%	99.1%	99.0%	99.1%
haus_groesse	99.7%	99.1%	99.1%	99.6%	99.2%	99.4%	99.3%	99.4%	99.4%	99.4%	99.8%	99.1%	99.8%	99.4%	99.4%	99.3%	99.4%	99.4%
skz	99.7%	99.4%	99.4%	99.5%	99.2%	99.4%	99.4%	99.3%	99.3%	99.5%	99.8%	99.7%	99.1%	99.2%	99.3%	99.3%	99.3%	99.2%
ww_idnr	99.9%	99.7%	99.7%	99.8%	99.7%	99.8%	99.8%	99.8%	99.9%	99.9%	99.9%	99.8%	99.8%	99.7%	99.8%	99.7%	99.8%	99.7%
staatsb	99.8%	99.4%	99.5%	99.6%	99.4%	99.5%	99.5%	99.7%	99.5%	99.9%	99.9%	99.5%	99.4%	99.5%	99.4%	99.4%	99.5%	99.4%
gebland	99.4%	99.6%	99.6%	99.5%	99.2%	99.4%	99.5%	99.6%	99.7%	99.5%	99.9%	99.3%	99.4%	99.3%	99.3%	99.4%	99.4%	99.4%
gebld	99.5%	99.3%	99.3%	99.3%	99.4%	99.5%	99.5%	99.4%	99.6%	99.7%	99.8%	99.3%	99.4%	99.2%	99.2%	99.3%	99.2%	99.4%
meldanf	99.7%	99.4%	99.4%	99.5%	99.2%	99.7%	99.8%	99.5%	99.5%	99.5%	99.8%	99.4%	99.3%	99.5%	99.4%	99.2%	99.4%	99.3%
giltab	99.8%	99.3%	99.3%	99.6%	99.4%	99.8%	99.7%	99.5%	99.4%	99.5%	99.8%	99.4%	99.4%	99.5%	99.3%	99.3%	99.4%	99.3%
gebjahr	99.4%	99.1%	99.1%	99.4%	99.7%	99.4%	99.2%	99.4%	99.2%	99.4%	99.7%	99.2%	99.2%	99.2%	99.2%	99.2%	99.3%	99.3%
famstand	99.7%	99.6%	99.6%	99.9%	99.4%	99.6%	99.5%	99.3%	99.5%	99.6%	99.8%	99.5%	99.6%	99.3%	99.4%	99.5%	99.3%	99.6%
hh_klasse	99.8%	100.0%	99.8%	99.6%	99.1%	99.3%	99.4%	99.3%	99.6%	99.5%	99.7%	99.4%	99.1%	99.2%	99.3%	99.5%	99.3%	99.5%
hh_groesse	99.8%	99.8%	100.0%	99.6%	99.1%	99.3%	99.4%	99.3%	99.6%	99.4%	99.7%	99.4%	99.1%	99.2%	99.3%	99.4%	99.3%	99.4%
geschl	99.8%	99.8%	99.8%	99.7%	99.4%	99.8%	99.7%	99.5%	99.4%	99.8%	99.9%	99.7%	99.7%	99.4%	99.5%	99.6%	99.5%	99.6%
geschl	geschl	hh_groesse	hh_klasse	famstand	gebjahr	giltab	meldanf	gebld	gebland	staatsb	ww_idnr	skz	haus_groesse	acd	bezirk	zbezirk	baublock	zgebiet

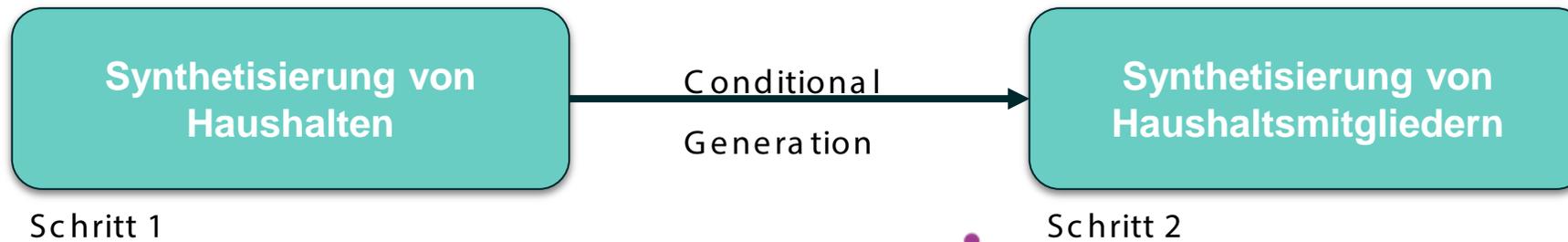
Alle univariaten Verteilungen  
 und bivariaten Zusammenhänge  
 zu mind. 99% erhalten

(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	MfOS6T713cL4Yc0d68Abdl29
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

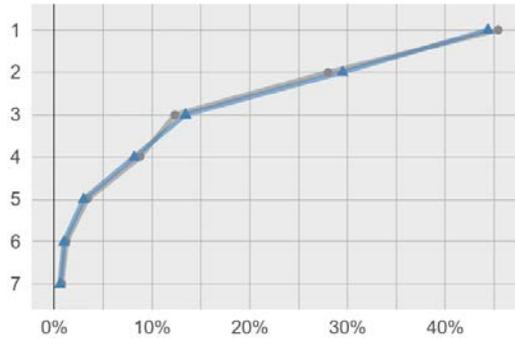
Key		
HAUSHALTE_ID	MfOS6T713cL4Yc0d68Abdl29	MfOS6T713cL4Yc0d68Abdl29
GESCHL	2	1
GEBJAHR	1966	1962
GEBBDL	Wien	Niederösterreich
GEBLAND	40	40
STAATSB	40	40
FAMSTAND	2	2
MELDANF	1965-11-13	1997-08-03

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

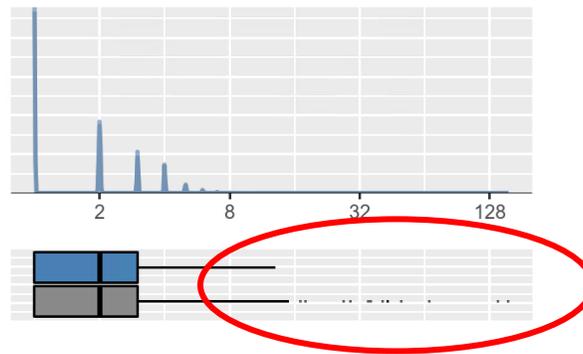
**Schritt 2:** Zu jedem Haushalt wird eine zu den Attributen des Haushalts passende Gruppe von synthetischen Personen generiert

# Statistiken von Haushalten

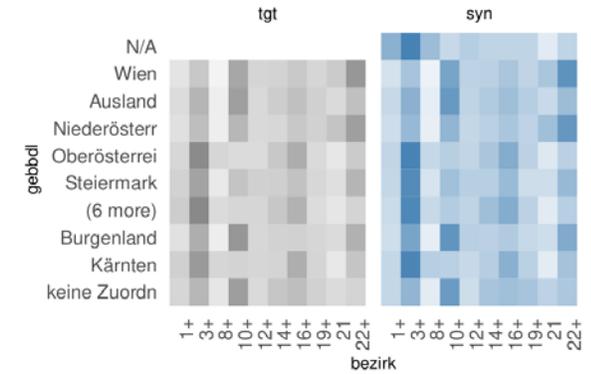
hh\_klasse



hh\_groesse

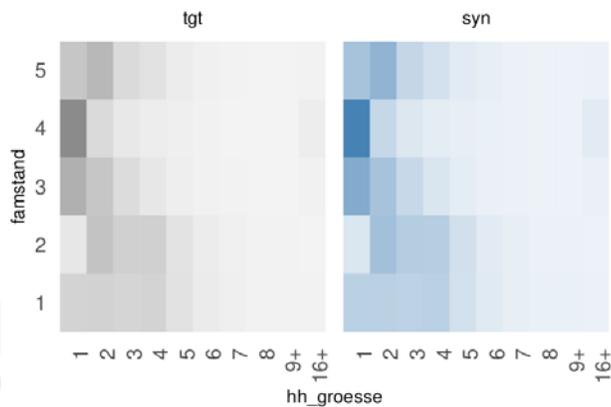


gebdbl ~ bezirk

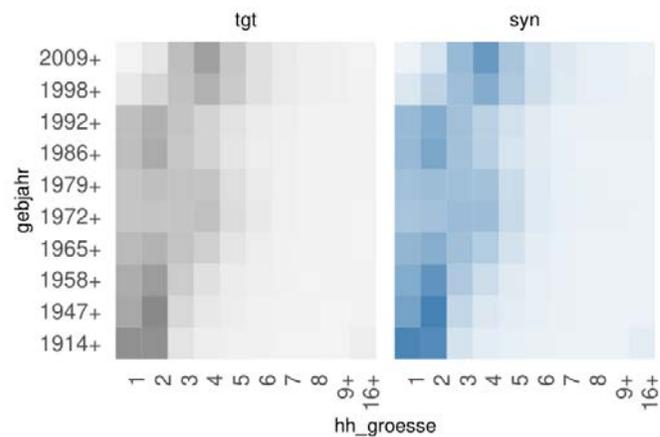


Auch Statistiken auf Haushaltsebene sehr gut getroffen

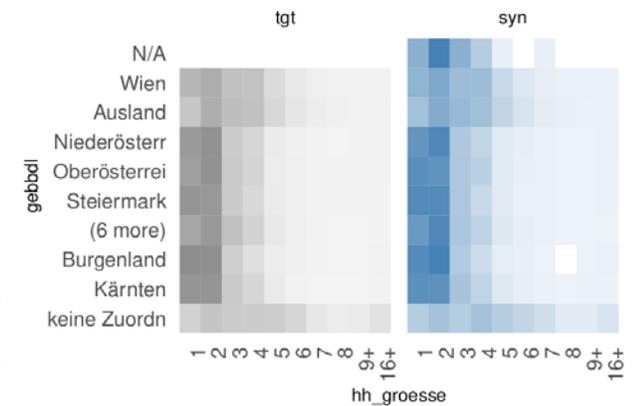
famstand ~ hh\_groesse



gebjahr ~ hh\_groesse

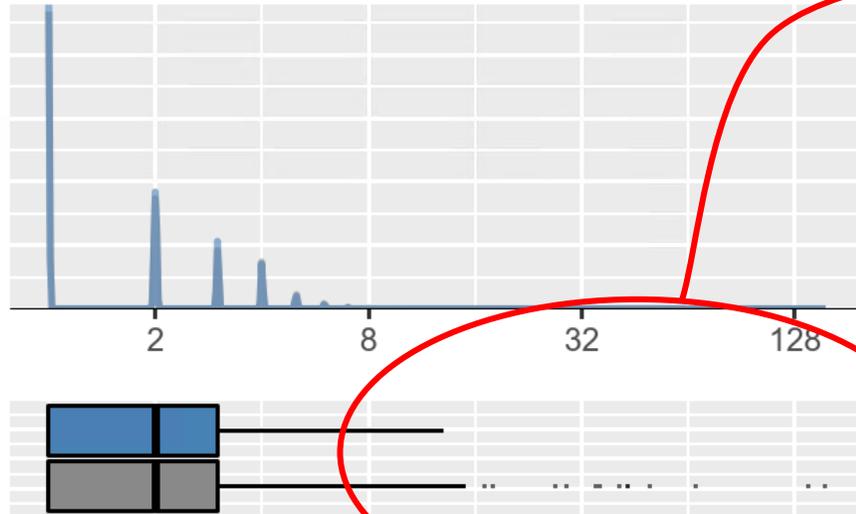


gebdbl ~ hh\_groesse



# Statistiken von Haushalten

hh\_groesse



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

		<i>stimmt nicht überein mit</i>			
		<u>bezirk</u>	<u>zbezirk</u>	<u>zgebiet</u>	<u>baublock</u>
Wert aus	<u>bezirk</u>	0,000%	0,001%	0,005%	2,645%
	<u>zbezirk</u>	0,001%	0,000%	0,013%	3,307%
	<u>zgebiet</u>	0,005%	0,013%	0,000%	3,472%
	<u>baublock</u>	2,645%	3,307%	3,472%	0,000%



# Synthetisierungsablauf im Überblick



Anzahl Haushalte } pro Bezirk  
Anzahl Personen }  
Zuordnung Baublocks zu Bezirken

**Haushalte** mit  
exakter Geografie,  
Hausgröße, etc.

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



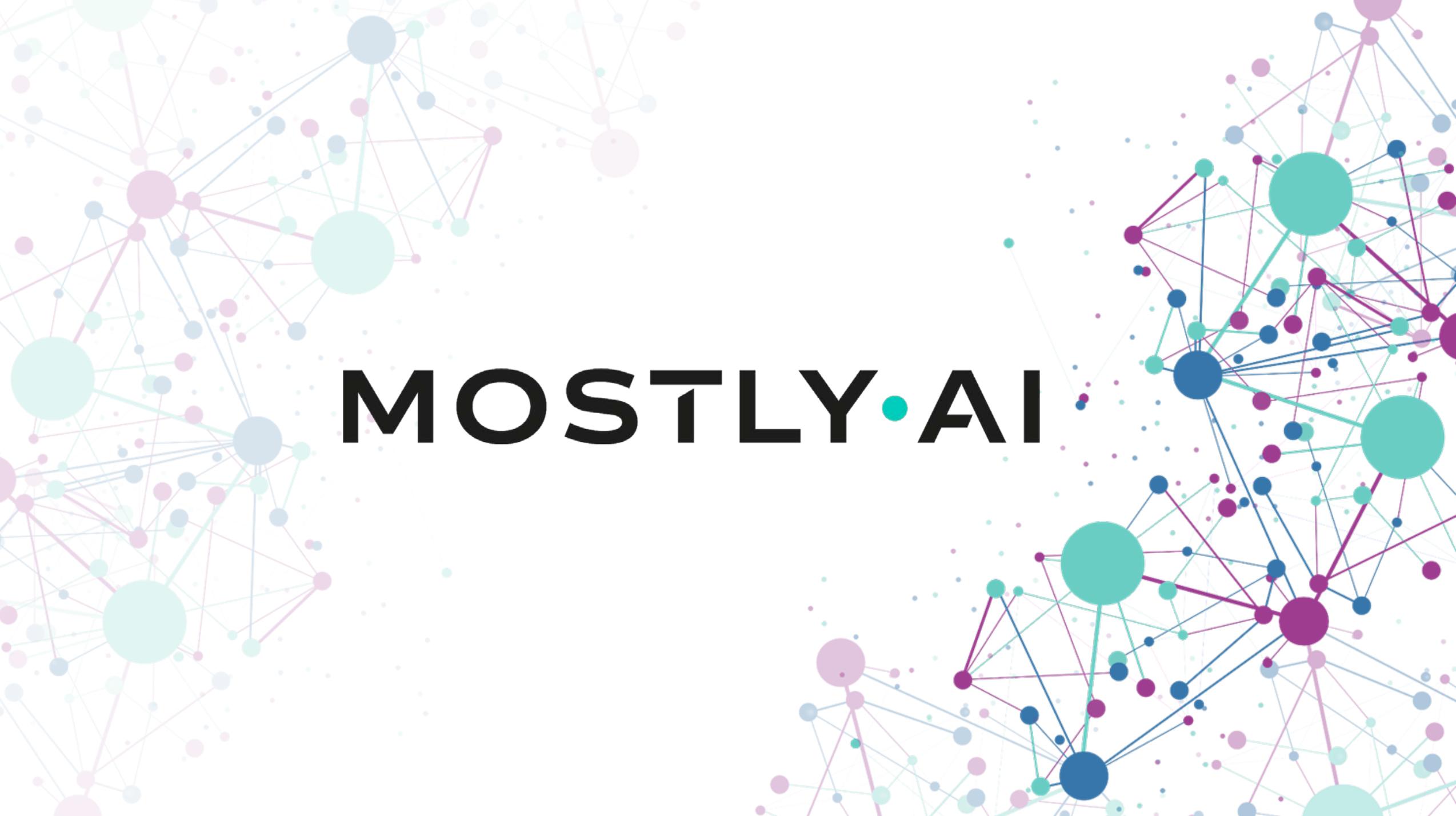
# Nächste Schritte & Ziele

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**