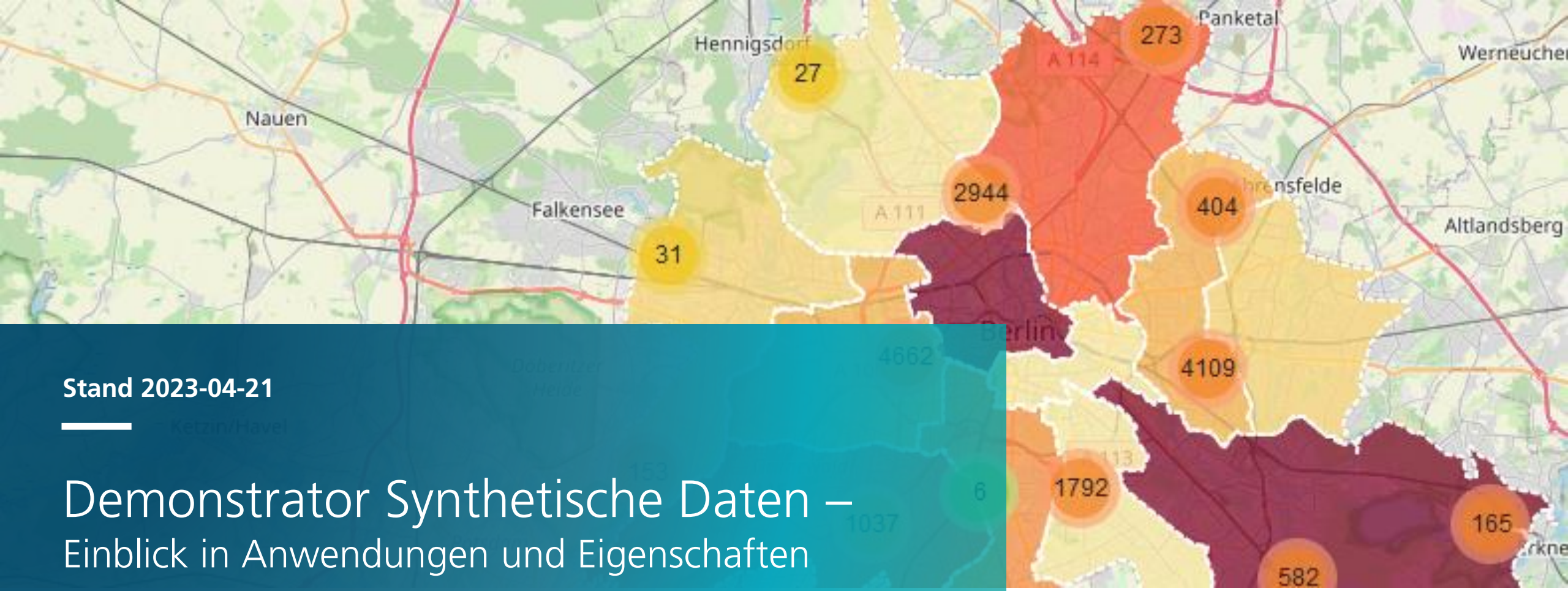


Stand 2023-04-21

Demonstrator Synthetische Daten –
Einblick in Anwendungen und Eigenschaften

Dorian Wachsmann, Jens Tiemann, Jan Dennis Gumz, Fabian Manzke

Gefördert durch:
Bundesministerium des Innern und für Heimat
Fraunhofer FOKUS

Übersicht

Was sind synthetische Daten?

- Herausforderung & Lösungsansatz
- Algorithmen im Demonstrator

3 Zugänge zum Demonstrator

- schützenswerte Daten anonymisieren
- synthetische mit realen Daten vergleichen
- Illustration von Re-Identifizierbarkeit

Detail zum Demonstrator

Siehe auch Blogbeitrag:

Kompetenzzentrum Öffentliche IT

Abschnitt 1

Theorie: Was sind synthetische Daten?

Was sind synthetische Daten?

Herausforderungen im Umgang mit Daten

- enge Grenzen bei Nutzung personenbezogener Daten
- sensible / schützenwerte Daten in allen Bereichen

Daten als (begrenzte) Ressource

- ggf. eingeschränkte Verfügbarkeit
- ggf. Unvollständigkeit, Stichwort Fairness
- multiple Quellen nötig?

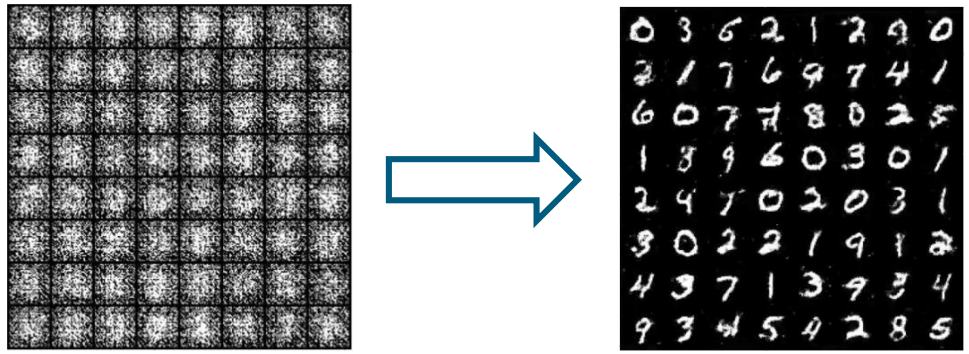
Herausforderung des „Daten-Labeling“

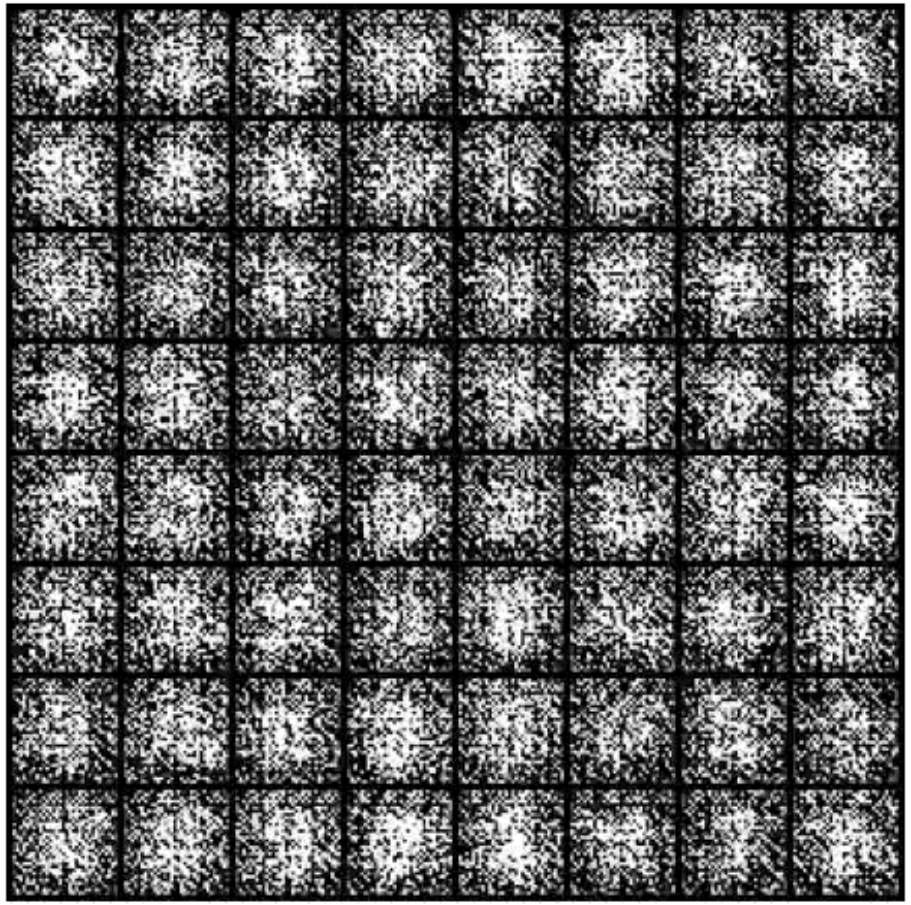
- Überwachtes Lernen von KI Modellen benötigt „gelabelte“ Daten: Woher kommen die Labels, »wer tut die Arbeit«?

Lösungsansatz

Synthetische Daten

- sind „künstlich“ erzeugt, also algorithmisch
- sollten idealerweise Struktur & Eigenschaften realer Daten abbilden
  - abhängig von Ziel der Nutzung und verwendeter Methode
- Realdaten bleiben geschützt
  - Schutz von Individuen bzw. Datenpunkten



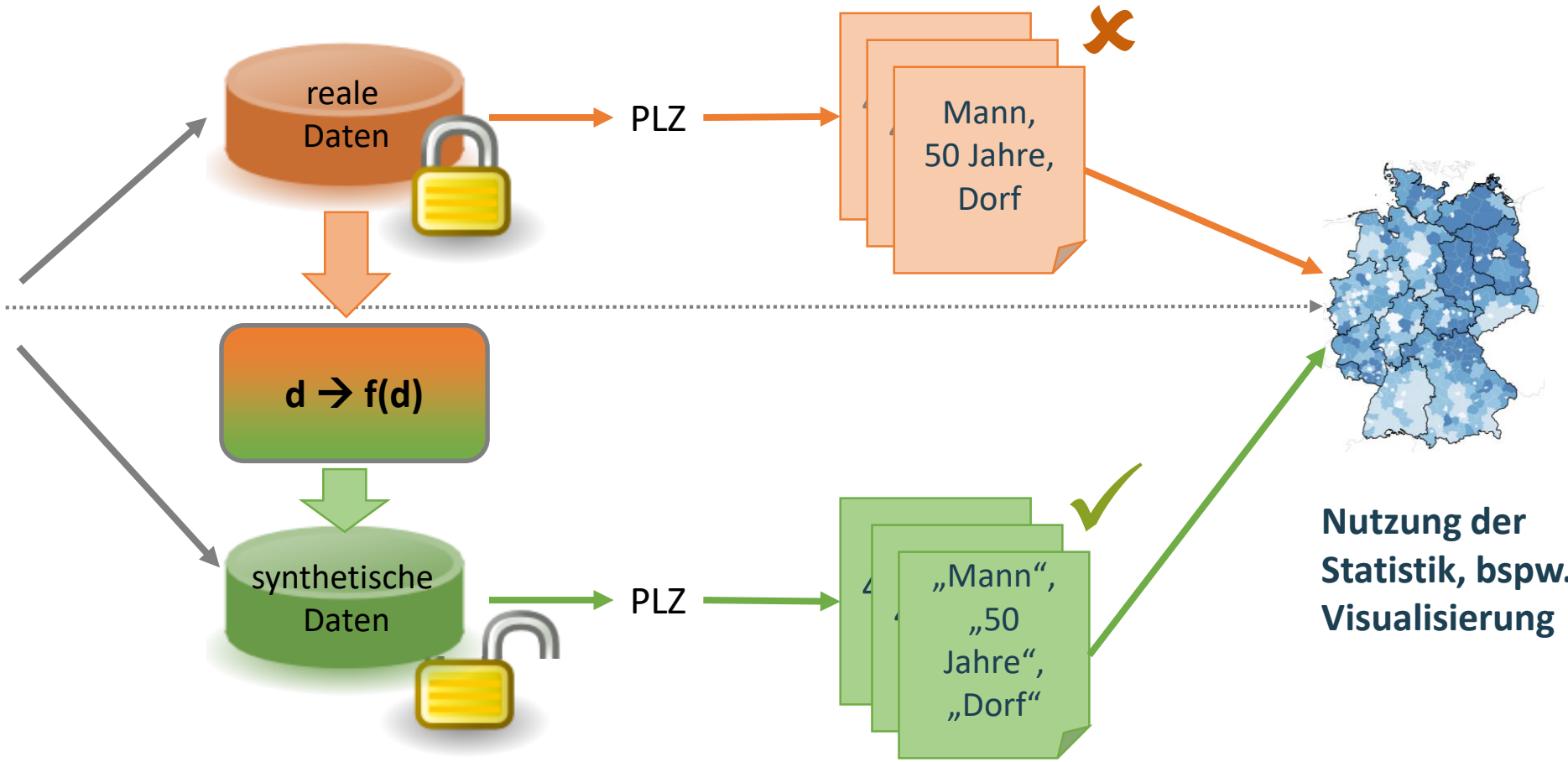


(Video)

Altersverteilung nutzen, ohne mit realen Datensätzen arbeiten zu müssen

Beispiel zur Verwendung synthetischer Daten

Fragestellung: Wohnorte(„männlich, 45-64 Jahre“)



Nutzung der Statistik, bspw. Visualisierung

Abschnitt 2

Der Demonstrator: Methoden und Zugänge

Demonstrator gibt ersten Einblick

Einleitung

reale Daten als Basis → Erzeugung synthetischer Daten

- statistische Eigenschaften der Originaldaten bestmöglich abbilden

Genutzte Datenbasis: Baumbestand Berlin

- Annahme: Eigenschaften von Bäumen enthalten sensible Informationen
- vergleichbar mit personenbezogenen Daten
  z. B. Pflanzjahr := Geburtsjahr

Anmerkung:
Die hier vorgestellten Verfahren dienen der Illustration und müssen für konkrete Anwendungen angepasst werden

Kompetenzzentrum Öffentliche IT

Demonstrator gibt ersten Einblick

Was ist unser Ansatz?

reale Daten als Basis → Erzeugung synthetischer Daten

- „Anonymisierung“ durch Vergröberung
- 3 generative Methoden:
  - Gaussian Copula
  - Generative Adversariale Netze
  - Variationeller Autoencoder

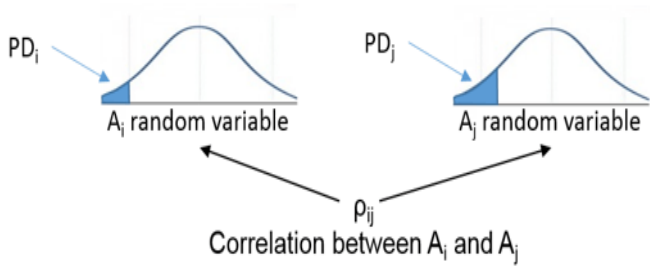
Jede Methode hat Vorteile und Nachteile, der Demonstrator ermöglicht Vergleich anhand einer praktischen Anwendung

Methoden im Detail

Generierung synthetischer Daten

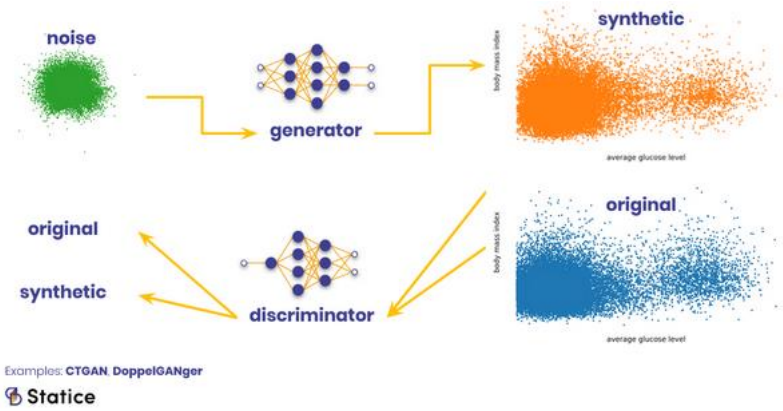
Allen Methoden gemeinsam: Rekonstruktion der unbekannten Wahrscheinlichkeitsverteilung, die den Originaldaten zugrunde liegt

Gaussian Copula



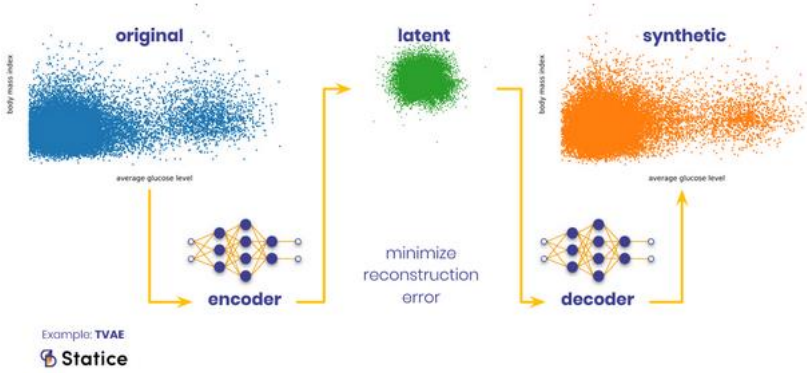
- Funktionale Zusammenhänge zwischen Randverteilungen
- Modellierung stochastischer Abhängigkeiten

Generative Adversariale Netze



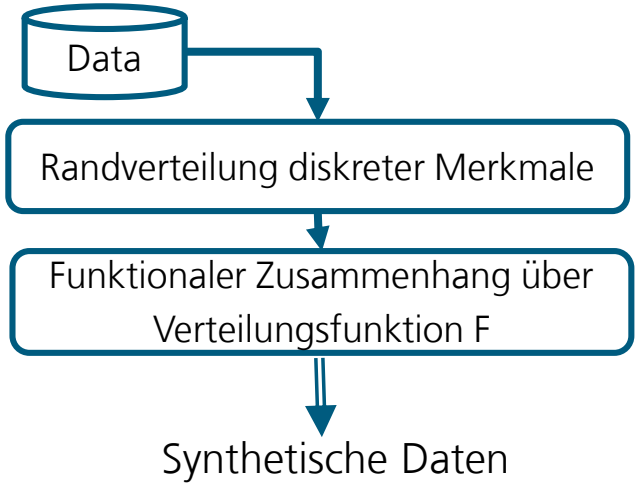
- Deep Learning Methode
- Spieltheoretischer Ansatz: Generator vs. Diskriminator
- Nachteil: Relative kompliziertes Training

Variationeller Autoencoder



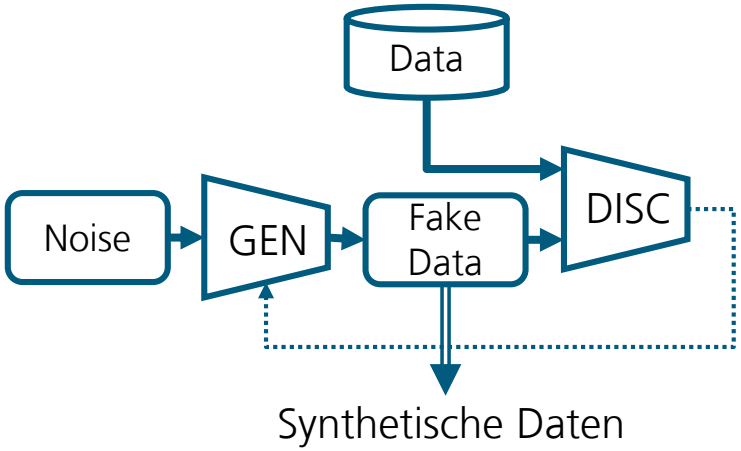
- Deep Learning Methode
- Encoder kodiert Basisdaten, Decoder decodiert Daten → lernt über Ähnlichkeit
- Daten Neu-Generierung durch Variieren des latent-space (Input des Decoders)

Generierung synthetischer Daten



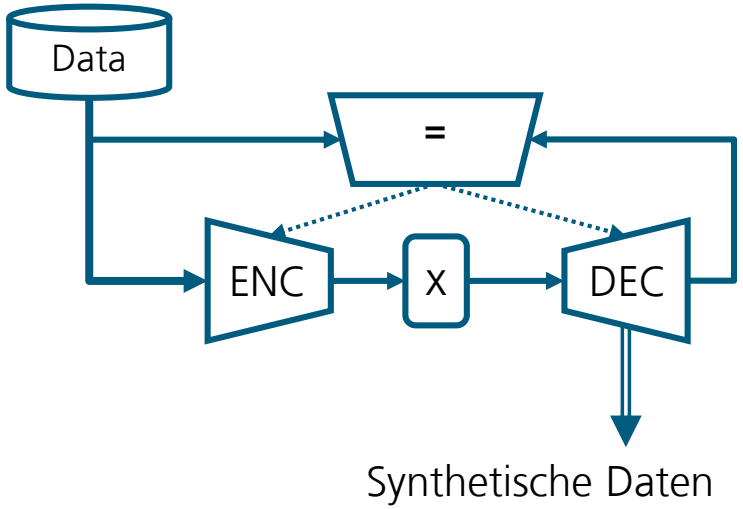
Gaussian Copula

- Sampling von der (multivariaten) Verteilungsfunktion F zur Datengenerierung



Generative Adversariale Netze

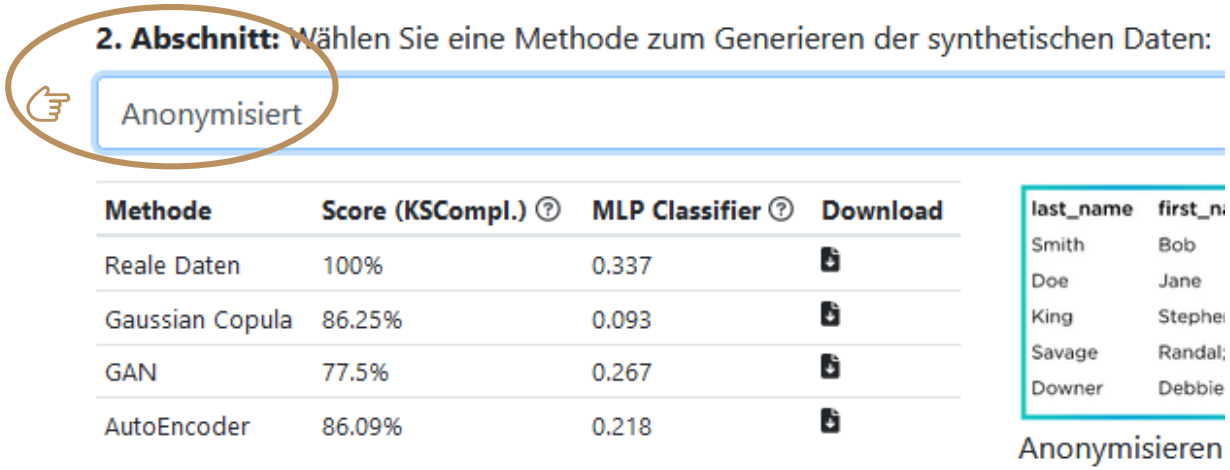
- Deep Learning Methode
- Spieltheoretischer Ansatz: Generator (GEN) vs. Diskriminator (DISC)

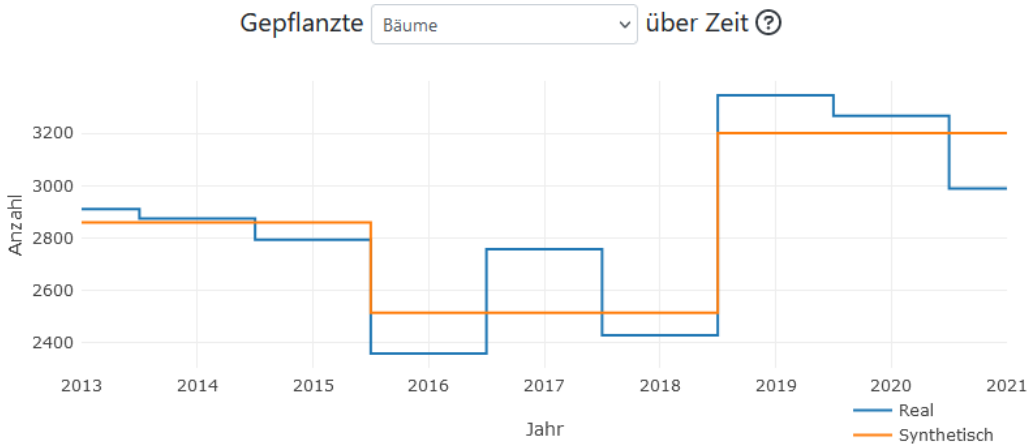


Variationeller Autoencoder

- Deep Learning Methode
- Idee Training: ENC kodiert Data d zu x, DEC decodiert x zu ~d. Abstandsmaß d zu ~d definiert den Loss
- Idee Generierung: wenn x ausreichend strukturiert, sample random x und nutze DEC für Generierung

Erster Zugang: Klassische Anonymisierung





klassische Methode zum Schutz sensibler Informationen: Vergröberung

- Effekt: Individuum „verstecken“ in einer Gruppe

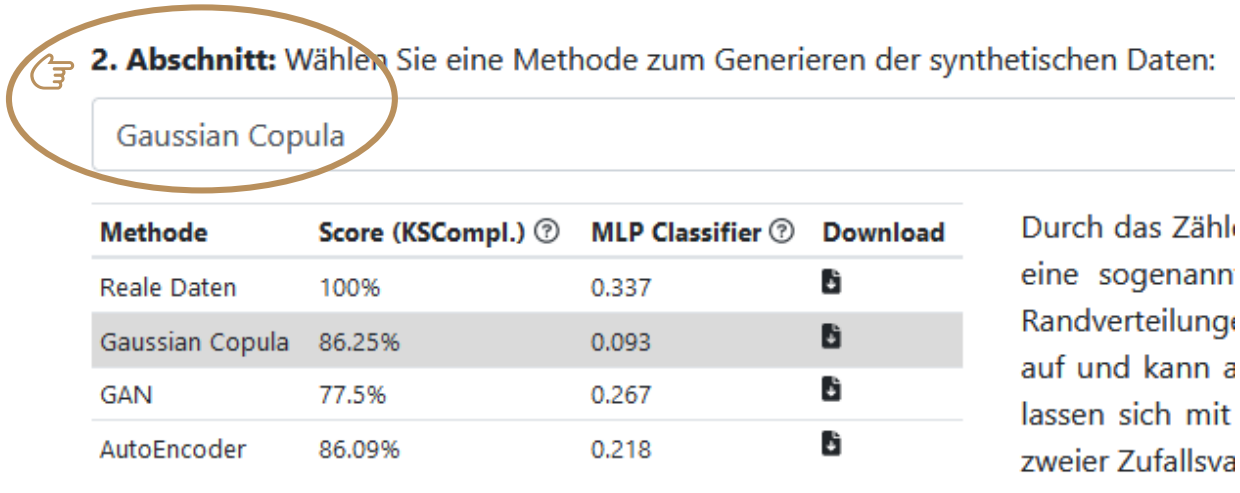
☞ Auswahl der Methode »Anonymisiert« im 2. Abschnitt

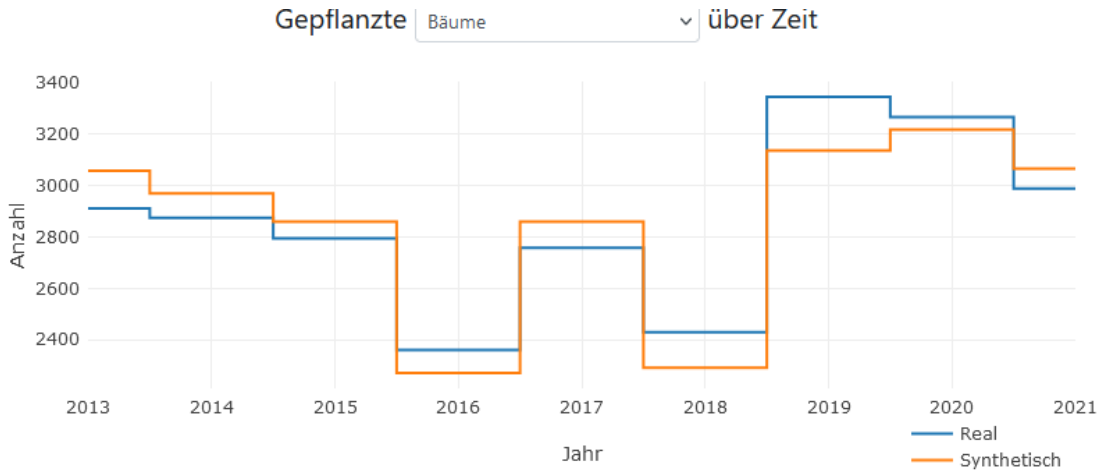
☞☞ Auswahl „Gepflanzte Bäume über Zeit“ im 3. Abschnitt

Demonstrator zeigt im 3. Abschnitt den Vergleich realer und – in diesem Fall – einfach anonymisierten Daten

- Wirkung: jahresgenaue Auflösung ist nicht mehr möglich
- Nachteil: Auswertung entsprechend ungenau unmöglich

Erster Zugang: Anonymisierung mit synthetischen Daten





Ansatz zum Schutz sensibler Daten:
Generierung künstlicher Daten mit ähnlichen statistischen Eigenschaften

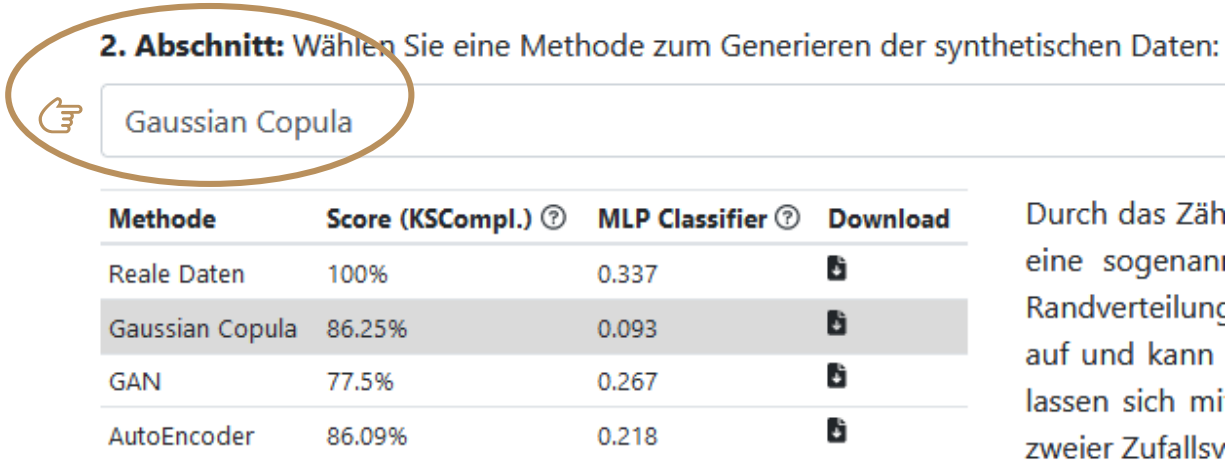
- Effekt: idealerweise sind die Datenobjekte komplett künstlich und lassen keine Rückschlüsse auf reale Echtdaten zu

☞ Wechsel der Methode im 2. Abschnitt zu »Gaussian Copula«

Hier: genaue Darstellung der Eigenschaft „Pflanzjahr“

Verschiedene Methoden zeigen eine mehr oder weniger gute Annäherung an die Eigenschaft „Jahr“

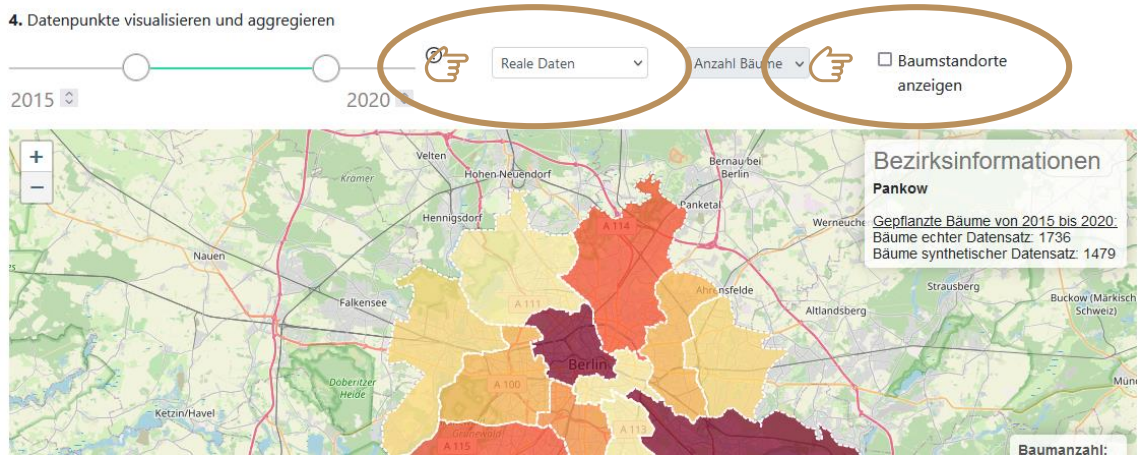
Zweiter Zugang: Vergleichbarkeit Synthetische Daten & Echtdaten



Nutzung einer Methode zur Generierung synthetischer Daten:

- Gaus
- GAN
- Auto

☞ im 2. Abschnitt »Gaussian Copula« wählen

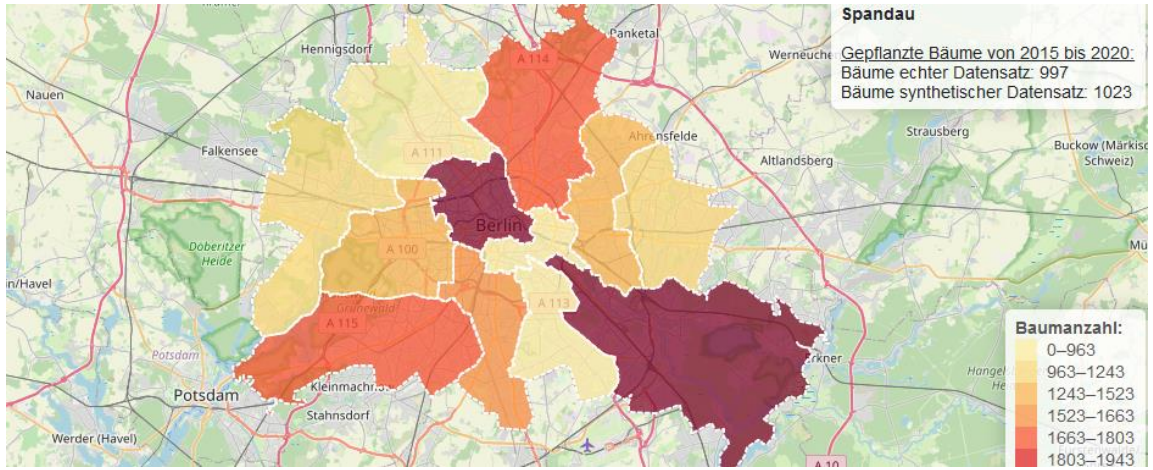


Karte zeigt im 4. Abschnitt reale Daten: Anzahl der Bäume pro Bezirk der Pflanzjahre 2015 bis 2020

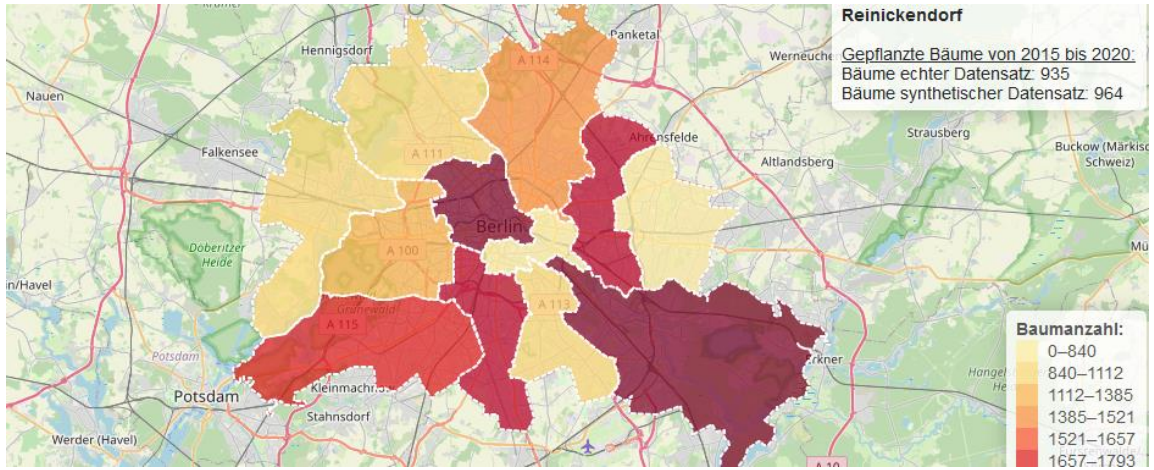
Bedienelemente:

☞ »reale / synthetische Daten« ☞ »Baumstandorte anzeigen«

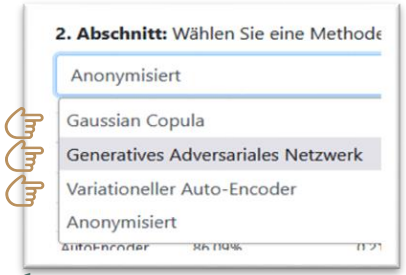
Zweiter Zugang: Vergleichbarkeit mit realen Daten – aggregierte Datensätze



☞ Einstellung: reale Daten



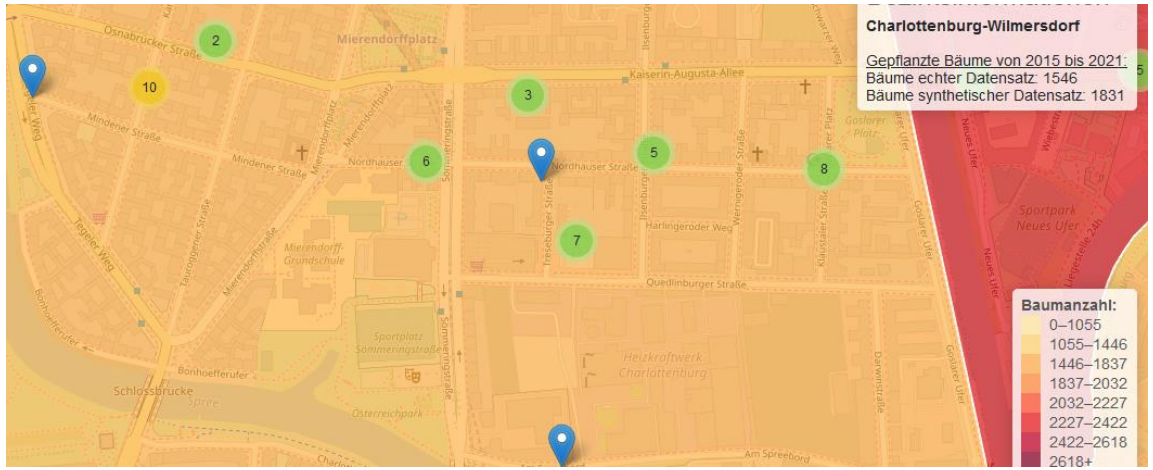
☞ Einstellung: synthetische Daten



Wechsel der Methode im 2. Abschnitt zeigt unterschiedlich ähnliche Verteilungen in Bezug auf reale Daten

Zugänge zum Demonstrator

Zweiter Zugang: Vergleichbarkeit mit realen Daten – einzelne Bäume

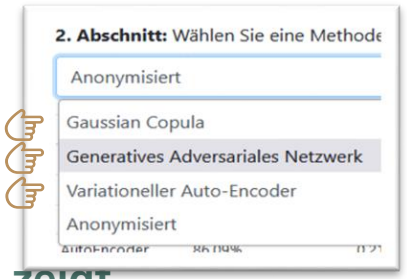


☞ Einstellung: Baumstandorte anzeigen
☞ Einstellung: reale Daten

Anmerkung: Die Datenbasis enthält nur Bäume, die in den letzten 10 Jahren gepflanzt wurden, geringe Anzahl von Standorten.

☞ Einstellung: synthetische Daten

Clusterbildung aufgrund von Reverse Geocoding



Wechsel der Methode im 2. Abschnitt zeigt unterschiedlich ähnliche Verteilungen in Bezug auf reale Daten

Dritter Zugang: Illustration der Re-Identifizierbarkeit

1. Abschnitt: Wählen Sie eine Datenmenge: (vorerst nur Baumdaten 2022)

Berliner Baumdaten 2022

Daten	25958 Berliner Bäume
Erhebung	2022
Infos	Sachdaten zum Baumbestand - Straßenbäume - mit Angaben zur Baumart, Adresse, Pflanzjahr, Höhe etc. für alle in den letzten 10 Jahren gepflanzten Bäume: Mehr Infos hier

Bezirk	Str.Name	Baumart	Pflanzjahr	Umfang	Höhe
Pankow	Blankenburger Straße	Kaiser-Linde	2014	32cm	3m
Neukölln	Gerlachsheimer Weg	Birne	2017	23cm	Keine Angabe
Treptow-Köpenick	Puschkinallee	Erle	2016	26cm	5m
Treptow-Köpenick	Am Falkenberg	Schwarz-Ahorn 'Faassen's Black'	2021	20cm	6m
Mitte	Neue Hochstraße	Esche	2016	20cm	6m

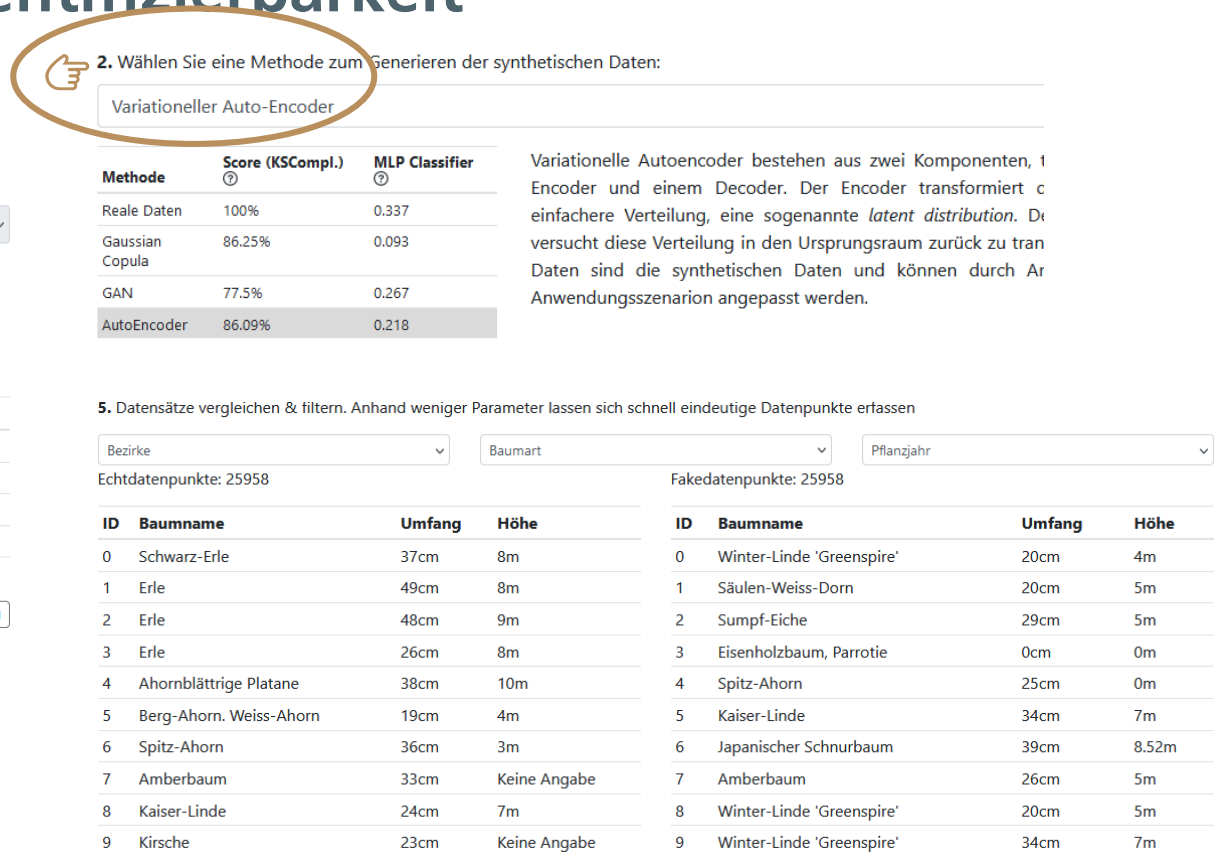
5 Beispieldaten

Zufällige Auswahl

Zur Datenbasis im 1. Abschnitt werden zur Illustration der Daten zufällig ausgewählte Datensätze angezeigt

- Datenbasis enthält Bezirk, Straße, Baumart, Pflanzjahr usw.

☞ Auswahl der Methode »Auto Encode« im 2. Abschnitt



2. Wählen Sie eine Methode zum Generieren der synthetischen Daten:

Variationeller Auto-Encoder

Methode	Score (KSCompl.)	MLP Classifier
Reale Daten	100%	0.337
Gaussian Copula	86.25%	0.093
GAN	77.5%	0.267
AutoEncoder	86.09%	0.218

Variationelle Autoencoder bestehen aus zwei Komponenten, t
Encoder und einem Decoder. Der Encoder transformiert d
einfachere Verteilung, eine sogenannte *latent distribution*. De
versucht diese Verteilung in den Ursprungsraum zurück zu tran
Daten sind die synthetischen Daten und können durch Ar
Anwendungsszenarion angepasst werden.

5. Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

Bezirke | Baumart | Pflanzjahr

Echtdatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Schwarz-Erle	37cm	8m
1	Erle	49cm	8m
2	Erle	48cm	9m
3	Erle	26cm	8m
4	Ahornblättrige Platane	38cm	10m
5	Berg-Ahorn. Weiss-Ahorn	19cm	4m
6	Spitz-Ahorn	36cm	3m
7	Amberbaum	33cm	Keine Angabe
8	Kaiser-Linde	24cm	7m
9	Kirsche	23cm	Keine Angabe

Fakedatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Winter-Linde 'Greenspire'	20cm	4m
1	Säulen-Weiss-Dorn	20cm	5m
2	Sumpf-Eiche	29cm	5m
3	Eisenholzbaum, Parrotie	0cm	0m
4	Spitz-Ahorn	25cm	0m
5	Kaiser-Linde	34cm	7m
6	Japanischer Schnurbaum	39cm	8.52m
7	Amberbaum	26cm	5m
8	Winter-Linde 'Greenspire'	20cm	5m
9	Winter-Linde 'Greenspire'	34cm	7m

Im 5. Abschnitt werden Einträge der realen und der synthetischen Datenbasen miteinander verglichen und können gefiltert werden

Dritter Zugang: Illustration der Re-Identifizierbarkeit

5. Datensätze vergleichen & filtern. Anhand weniger Parameter lassen sich schnell eindeutige Datenpunkte erfassen

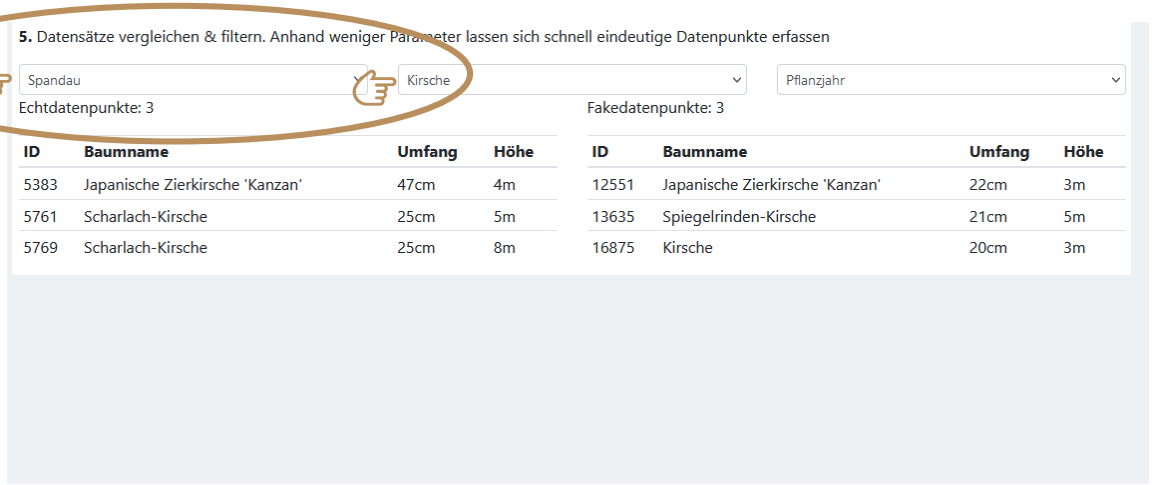
Bezirke | Baumart | Pflanzjahr

Echtdatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Schwarz-Erle	37cm	8m
1	Erle	49cm	8m
2	Erle	48cm	9m
3	Erle	26cm	8m
4	Ahornblättrige Platane	38cm	10m
5	Berg-Ahorn. Weiss-Ahorn	19cm	4m
6	Spitz-Ahorn	36cm	3m
7	Amberbaum	33cm	Keine Angabe
8	Kaiser-Linde	24cm	7m
9	Kirsche	23cm	Keine Angabe

Fakedatenpunkte: 25958

ID	Baumname	Umfang	Höhe
0	Winter-Linde 'Greenspire'	20cm	4m
1	Säulen-Weiss-Dorn	20cm	5m
2	Sumpf-Eiche	29cm	5m
3	Eisenholzbaum, Parrotie	0cm	0m
4	Spitz-Ahorn	25cm	0m
5	Kaiser-Linde	34cm	7m
6	Japanischer Schnurbaum	39cm	8.52m
7	Amberbaum	26cm	5m
8	Winter-Linde 'Greenspire'	20cm	5m
9	Winter-Linde 'Greenspire'	34cm	7m



Anhand von bekannten Merkmalen lässt sich ein Objekt „einkreisen“ und schließlich identifizieren

Beispiel: Seltener Nachname in einer kleinen Stadt – Person kann anhand dieser Informationen identifiziert werden

Synthetische Daten: keine reale Personen, sondern realistische statistische Eigenschaften nach festgelegten Kriterien

legt man einen Bezirk und eine (seltene) Baumart fest, reduziert sich die Anzahl infrage kommender Bäume schnell

☞ Auswahl Bezirk »Spandau« und Baumart »Kirsche«

Listen zeigen ähnliche Bäume, aber die synthetischen Bäume existieren nicht

Abschnitt 3

Weiterführendes

Baumdaten Berlin

Datenbasis

- Bestand Straßenbäume: über 900.000
- Reduzierung nötig: Pflanzjahr <= 10 Jahre: ~ 25.000
- Informationsgehalt: Baumart, Pflanzjahr, Umfang, Höhe, Bezirk, Str. Name, Hausnr.
- Keine Geokoordinate (lat/lng): manuelle Berechnung anhand von Str. Name (+Hausnr.)

  → Kartendarstellung führt zu »Clustering«

- Quelle:

Seite 1 von 87

	Baumart deutsch	Baumart botanisch	Gattung botanisch	Straße	Hausnr.	Pflanzjahr	Standalte
	Berg-Ahorn, Weiss-Ahorn	Acer pseudoplatanus	ACER	Kögelstraße	1	2018	5
eil II)	Spitz-Ahorn	Acer platanoides	ACER	Hansastraße	-	2014	9
ße	Kaiser-Linde	Tilia intermedia 'Pallida'	TILIA	Rheinsteinstraße	-	2015	8
	Kirsche	Prunus spec.	PRUNUS	Badstraße	37	2014	9
FSTR.	Winter-Linde 'Greenspire'	Tilia cordata 'Greenspire'	TILIA	Benekendorffstraße	169	2014	9
	Amerikanische Linde	Tilia americana	TILIA	Lübarser Straße	4	2014	9
	Eiche 'Fastigiata Koster'	Quercus robur 'Fastigiata Koster'	QUERCUS	Bayernallee	33/A	2018	5
	Eiche 'Fastigiata Koster'	Quercus robur 'Fastigiata Koster'	QUERCUS	Bayernallee	33A	2018	5
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
Straße	Berg-Ahorn, Weiss-Ahorn	Acer pseudoplatanus	ACER	Rudolf-Reusch-Straße	54	2015	8
Straße	Berg-Ahorn, Weiss-Ahorn	Acer pseudoplatanus	ACER	Rudolf-Reusch-Straße	34	2015	8
Straße	Berg-Ahorn, Weiss-Ahorn	Acer pseudoplatanus	ACER	Rudolf-Reusch-Straße	52	2015	8
g	Feld-Ahorn	Acer campestre	ACER	Bitterfelder Weg	59	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Kleinkronige Winter-Linde	Tilia cordata 'Rancho'	TILIA	Treskowallee	-	2015	8
	Krim-Linde	Tilia euchlora	TILIA	Siegfriedstraße	60	2015	8
	Krim-Linde	Tilia euchlora	TILIA	Siegfriedstraße	61-63	2015	8
	Amerikanische Rot						

Nutzung entlang der Abschnitte

Übersicht Demonstrator

1. Abschnitt: Datenmenge

- vorerst nur Baumdaten 2022

2. Abschnitt: Methode zum Generieren der synthetischen Daten

- Gaussian Coppula
- Generative Adversariale Netze (GAN)
- Variationelle Autoencoder
- Anonymisierung – zum Vergleich

3. Abschnitt: Szenarien zum qualitativen Vergleich der Daten

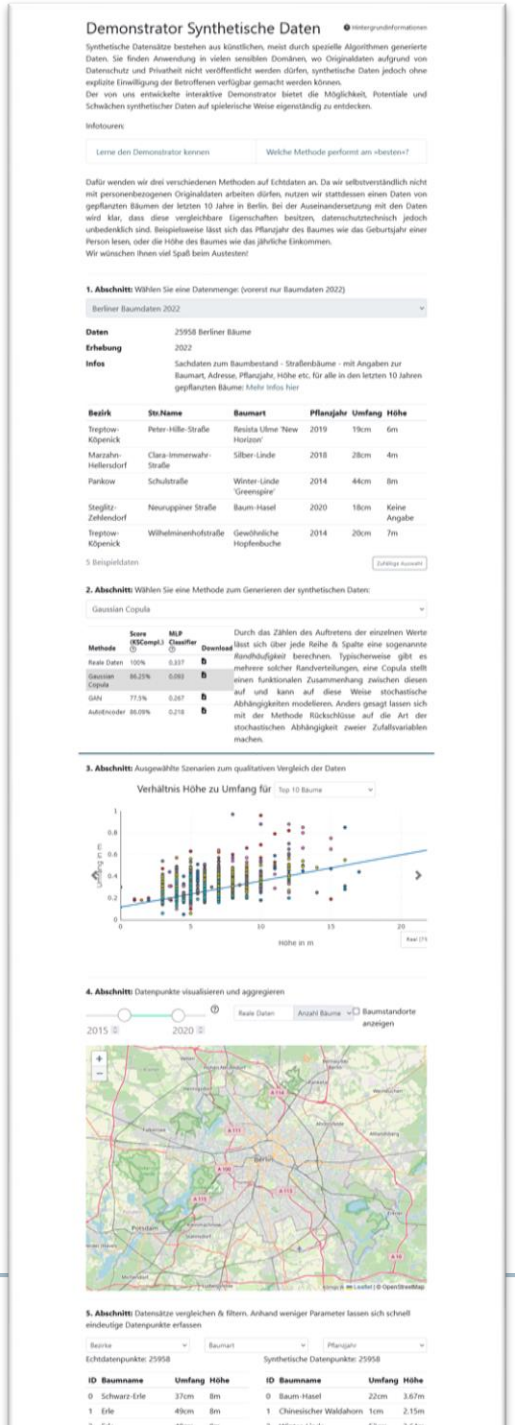
- Verschiedene statistische Auswertungen, Vergleich mit Realdaten

4. Abschnitt: Datenpunkte visualisieren und aggregieren

- Reale oder synthetische Datenpunkte auf Karten

5. Abschnitt: Datensätze vergleichen und filtern

- Filter nach Eigenschaften kann Datenpunkte effektiv eingrenzen



Kompetenzzentrum Öffentliche IT

https://www.oeffentliche-it.de/

Kompetenzzentrum Öffentliche IT

Öffentliche IT

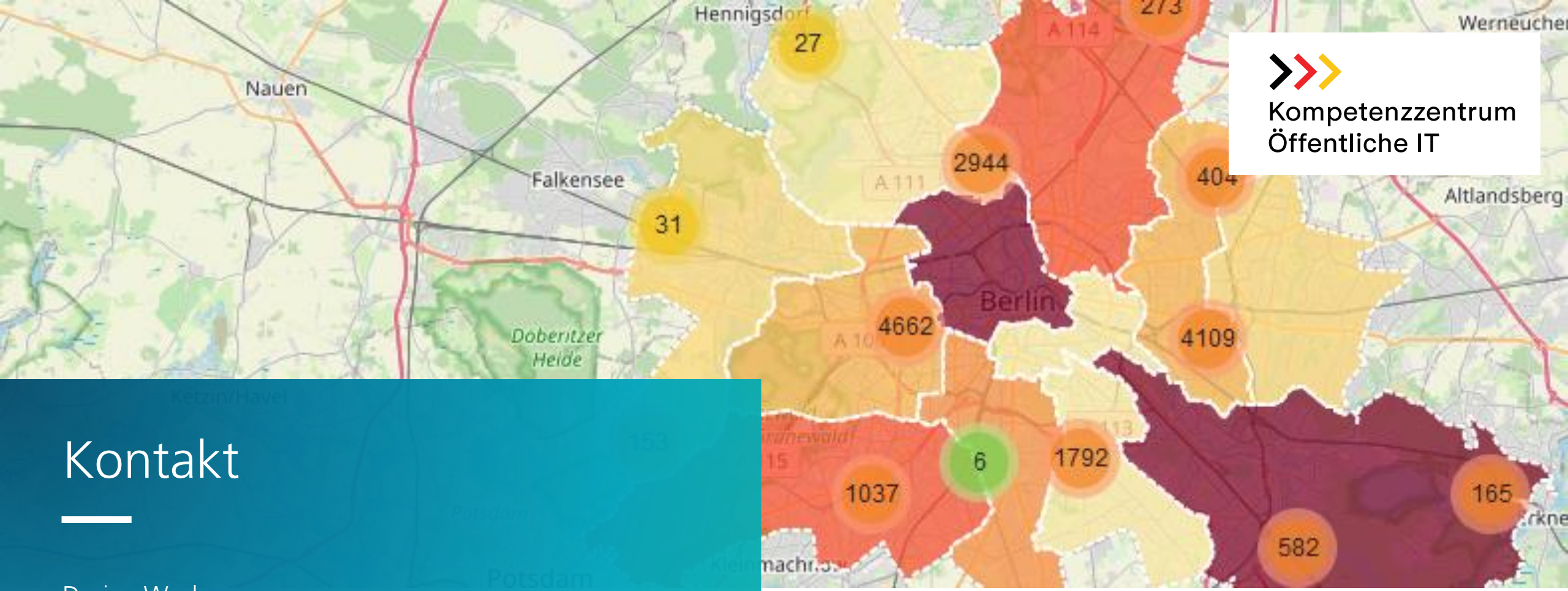
- IT im öffentlichen Raum
- IT der öffentlichen Verwaltung

Kompetenzzentrum Öffentliche IT (ÖFIT)

- angesiedelt am Fraunhofer-Institut für offene Kommunikationssysteme (FOKUS)
- gefördert vom Bundesministerium des Innern und für Heimat (BMI)

Demonstrator Synthetische Daten

https://www.oeffentliche-it.de/werkstatt/synthetische-daten-demonstrator



Kompetenzzentrum
Öffentliche IT

Kontakt

Dorian Wachsmann
Kompetenzzentrum Öffentliche IT (ÖFIT)
Tel. +49 30 3463-7338
dorian.wachsmann@fokus.fraunhofer.de

Gefördert durch:
Bundesministerium des Innern und für Heimat
Fraunhofer FOKUS