Synthetic Data Generation with PostgreSQL

OSS DB Talks 2023

23. March 2023 17:30, SIX ConventionPoint Zürich

Prof. Stefan Keller OST

Presentation Overview

Synthetic Data Generation with PostgreSQL

1. Introduction
2. Generation
3. Evaluation
4. Generation for PostgreSQL
5. pgsynthdata Tool

Synthetic Data - Introduction

What is synthetic data?

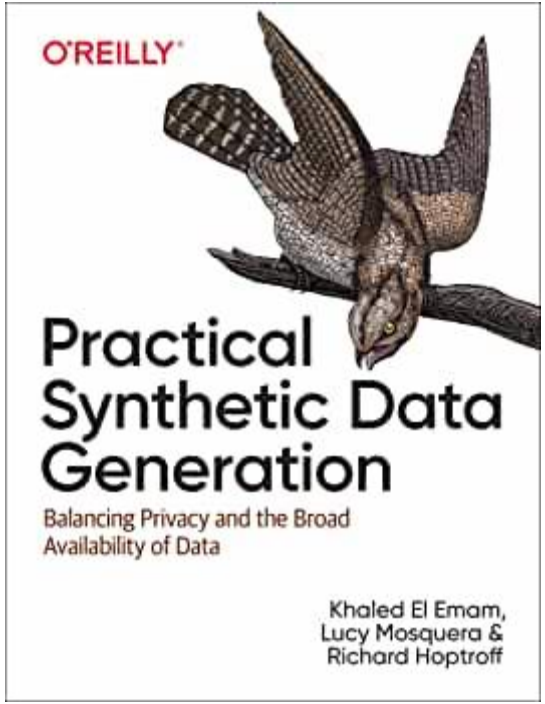
- Definition: "Data applicable to a given situation that are not obtained by direct measurement"
  (Source: https://en.wikipedia.org/wiki/Synthetic_data)

- Synthetic data …
  - is often explained through it's creation process (generation)
  - is a broader term for narrower terms like anonymized, artificial and fully/pure synthetic data (see later) –
  - sometimes also defined a subset of anonymized data

What is synthetic data?

- Synthetic Data tries to
  - preserve the overall properties and characteristics of the original data
  - without revealing information about actual individual data samples

- Other statement:
  - "Balancing privacy and the demand for data availability" (Practical Synthetic Data Generation (2020), Emam, Mosquera & Hoptroff, O'Reilly)



Why synthetic data?

- Real data might have gaps and structural mismatches (compared to data that will be processed later)
- Access to real data might be restricted
- Real data might be subject to privacy
- Real data might be non-existent

Applications of synthetic data

- Model training
  (machine learning, AI)
- Data anonymization
  (open data publishing)
- Software testing for reliability or scalability
  (SW development, SW engineering)
- Database performance optimization
  (data engineering)

Synthetic Data: Companies and industries

- Some industries (no claim to be complete):
  - Medicine
  - Government
  - Computer Science and Data Science, etc. …

- Some companies (no claim to be complete):
  - MOSTLY AI, Wien AT, https://mostly.ai
  - Synthesized, London UK, www.synthesized.io
  - Gretel.ai, San Diego USA, https://gretel.ai
  - Datacebo, Boston USA, https://datacebo.com/
  - Syntheticus.ai, Zürich CH, https://syntheticus.ai
  - itopia AG, Zürich CH, www.itopia.ch/en/key-issues/synthetic-data/

Factsheet

iSynth

itopia
corporate information technology

Efficient generation of customized, complex and consistent synthetic data

It is often not possible or permissible to use production data, anonymized or otherwise, for test purposes and doing so can lead to unwanted side effects.

Synthetic data is the alternative. Until now, however, creating synthetic data has been too complex, of insufficient quality and, often, simply not practicable.

iSynth enables you to create synthetic data efficiently and cost-effectively.

Key Features

Support of all test levels: iSynth can supply synthetic data for all test levels. From unit tests to fully integrated system tests on large system landscapes. iSynth's synthetic data can be used as primary data or to complement existing data and is suitable for generating both concise and large test data sets.

Flexible and extensible: iSynth is easy to adapt and extend according to

Challenges

Regulatory requirements: Requirements for handling private, health and bank data are becoming increasingly restrictive. In the age of big data analytics, solutions based on data anonymization will most likely fail to meet the requirements in terms of data and confidentiality protection.

Agile projects and DevOps testing: In addition to appropriate development methods and tools, frameworks, service simulators and test utilities, it is essential to have the right amount of consistent, high-quality test data in order to perform meaningful tests.

External sourcing: Software components are often developed by external partners. Whether these partners are active onshore, nearshore or offshore is

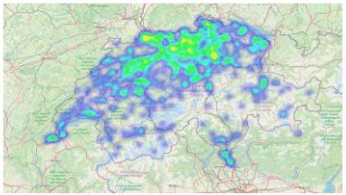
Use Case of a point GAN by Syntheticus.ai

Use Case

	FIRST 5 ENTRIES	HEATMAP BY CITY	HEATMAP BY POPULATION

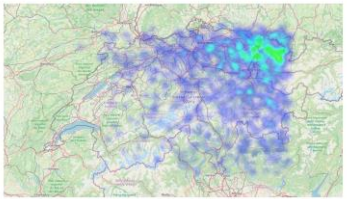
ORIGINAL

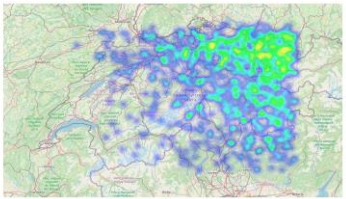
	city	lat	lng	population
0	Zürich	47.3786	8.5400	434008
1	Geneva	46.2000	6.1500	201818
2	Basel	47.5606	7.5906	177595
3	Lausanne	46.5333	6.6333	138905
4	Bern	46.9480	7.4474	133798



SYNTHETIC 50 EPOCH

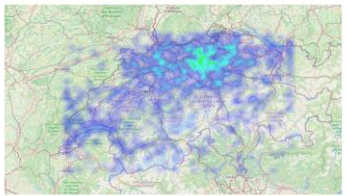
	city	lat	lng	population
0	Oberhelfenschwil	47.4818	8.6384	3243
1	Lungern	47.0919	9.2770	722
2	La Roche	47.3226	6.9592	878
3	Oberdorf	47.6232	9.6130	1257
4	Riva San Vitale	46.5774	10.1743	1224

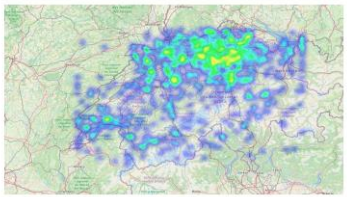




SYNTHETIC 200 EPOCH

	city	lat	lng	population
0	Oberhelfenschwil	47.6242	7.4491	4094
1	Lungern	46.3720	8.1040	1854
2	La Roche	47.5201	6.7359	1151
3	Oberdorf	47.7135	7.7865	10004
4	Riva San Vitale	47.0594	9.4420	1376





Methods for generating synthetic data

- Statistical methods:
  - Mathematical models to generate data that have the same statistical properties as real data.
  - Frequency distribution classification, Monte Carlo simulation, Gaussian mixture modeling, Markov chain.
- Rule-based methods:
  - Define rules or constraints that generate synthetic data.
  - A rule might be that a zip code must match the corresponding city.
- Data augmentation methods:
  - This involves adding noise, perturbations, or transformations to real data to create new, synthetic data.
  - Fuzzification, Differential Privacy.
- Database methods:
  - A database management system is used to generate synthetic data. Databases in turn use statistical and rule-based methods (functional dependency). Random number generators.
- Machine learning methods:
  - Training ML models on real data and then using these to generate synthetic data w/ similar characteristics.
  - Generative adversarial network (GAN) to generate new images that are similar

Types of synthetic data (1 of 3)

- Fully synthetic data:
  - completely artificially generated
  - doesn't contain original data
- Partially synthetic data:
  - only values of the selected sensitive attribute are replaced with synthetic data
- Hybrid synthetic data:
  - generated using both original and synthetic data

(Source: Surendra & Mohan, 2017)

Types of synthetic data (2 of 3)

- Synthesized Ltd. (2018) defined following computer-supported data generation types:
- Anonymized data, produced by a 1-to-1 transformation from original data. Examples include noise obfuscation, masking, or encryption
- Artificial data, produced by an explicit probabilistic model via data sampling
- Synthetic data, produced by a model (configuration, rules) which in turn can be learned by statistics from original data

Types of synthetic data (3 of 3)

Original Data

account_id	trans_id	Name	type	amount	balance
5132412	2451	Max Rogue	34	545.65	2456.56

1 Anonymised Data

Hashing, noise obfuscation, masking, encryption

513XXX	24XX	X	[30-40]	480.34	2300.56

2 Artificial Data

Manual manufacturing

41554523	4324	Rachel McDonald	100	1043	514

3 Synthetic Data

Generative model of original data

3253215	456	Jack Ma	33	558	2152
5453472	623	Mike Brown	34	603	27600

3.1 Anonymized data: Intro.

- Anonymization
  - Modification of personal such that the individual details of personal circumstances can only be attributed to a natural person with a disproportionate expense of resources
  - Goal 1: Maximize accuracy of responses to queries to databases
  - Goal 2: Minimize probability of identifying the records used to respond
- Quantification of anonymity:
  - "Differential Privacy"
- Nice property: Applicable in realtime as part of queries as view on tables with productive data

3.1 Anonymized data cont.

- Approaches
  - Add noise or dummy data (similar pseudo-random GPS noise)
  - Aggregate data: classic approach, expose only sum of at least 3
  - Suppress/delete or swap data: see Differential Privacy

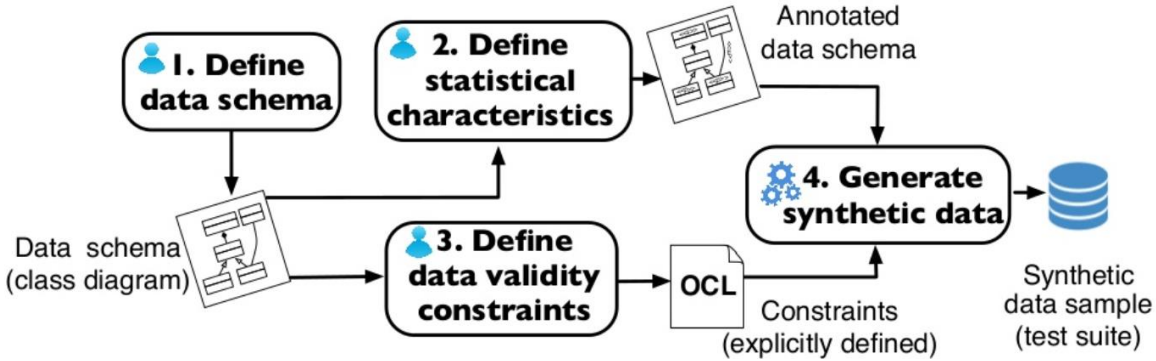
- Differential privacy:
  - Technique for ensuring that individual data points are protected when aggregate information is shared by adding or deleting data (noise) up to a certain value of parameter epsilon (ε)
  - Extension of K- and ε-anonymization

- Tool:
  - Differential Privacy by Google Repo; implements e.g. sum(), avg() etc.

3.2 Artificial data generation

- A model is created by hand
- which describes an observed behavior
- or by configuring statistic values
- (Agent-based modeling)

Source: "Synthetic Data Generation for Statistical Testing" by Soltana, Sabetzadeh, and Briand, University of Luxembourg



3.3 Pure synthetic data generation

- Drawing model by example:
  - Observing real statistical distributions of original data
  - Manual configuration is optional

- Tools:
  - pgsynthdata

3.3 Pure synthetic data: Challenges

- Outliers may be missing: Synthetic data can only mimic the real-world data, it is not an exact replica of it. Outliers can be more important than regular data points.
- Quality of the model depends on the data source: Quality of synthetic data is highly correlated with the quality of the input data and the data generation model. Synthetic data may reflect the biases in source data.
- User acceptance: It's an emerging concept and may be new to users
- Synthetic data generation (still) requires time and effort.

(Adapted from: https://research.aimultiple.com/synthetic-data/)

Synthetic data: Statistical Evaluation

- How to generate test data that meet both requirements, validity and representativeness, at the same time in a scalable manner?

Validity

Exhaustive search	Metaheuristic-search
- Constraint programming [Cabot et al., JSS 2014] - Alloy [Sen et al., ICMT 2019]	- Alternating Variable Method (AVM) [Ali et al., TSE 2013] [Ali et al., ESE 2016]

- **Representativeness**

Heuristics	Sampling
- Rule-based [Hartmann et al., SmartGridComm 2014] - Model-based [Soltana et al., SoSyM 2016]	- Boltzmann's random sampling [Mougenot et al., ECMDA-FA 2009]

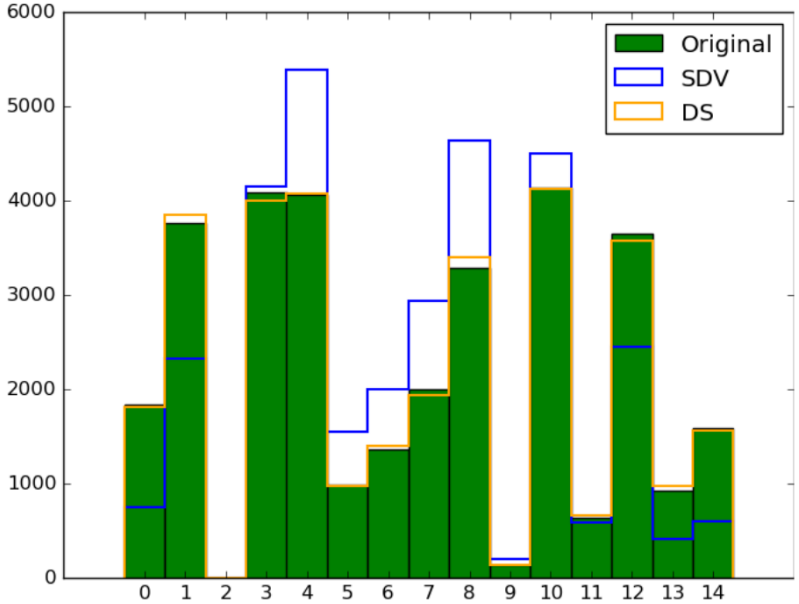
(Source https://www.slideshare.net/briand_lionel/synthetic-data-generation-for-statistical-testing

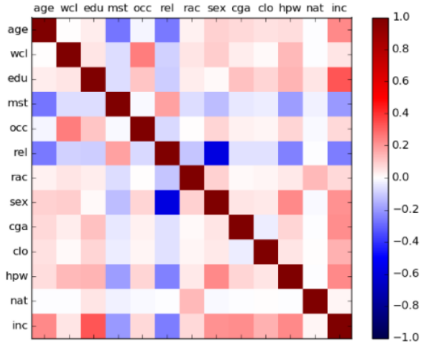
Synthetic data: Statistical Evaluation cont.

- For each of the datasets, perform following steps:
  - For each attribute, generate a histogram visualizing both the distribution of the real and the synthetic data
  - Compute the correlation coefficients and generate a heat map to visualize dependencies between attributes
  - Measure the distance between the real and the synthetic data via row-by-row computations of nearest neighbors

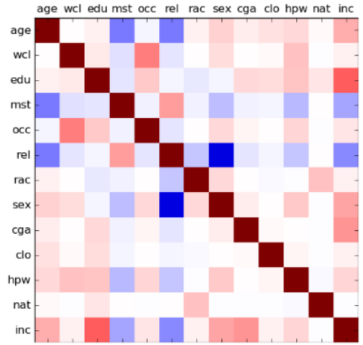
(Hittmeir et al. 2019)

Synthetic data: Statistical Evaluation cont.





(a) Original



(b) DS

(Source: Hittmeir et al. 2019)

Synthetic data: Experimental Evaluation

- Train various machine learning models
  - with the original data w\ test set
  - with synthesized data of same size
- Test them on the test set (taken from original data)
- Compare

(Hittmeir et al. 2019)

Experimental Evaluation: Outlook

- Understand why synthetization works better on some dataset than others
- Influence
  - Generation method
  - Differential privacy
  - Level of differential privacy (ε)
- Defining and quantifying the privacy levels and guarantees achieved by synthetic data

(Hittmeir et al. 2019)

Synthetic data generation tools

- Faker
  - Package in Python that generates fake data, like "first names"
  - MIT License
  - https://faker.readthedocs.io/
- Synthetic Data Vault (SDV)
  - Tools in Python to generate tabular synthetic data
  - Business source lic. ("anti service" not OSS), maintanied by Datacebo https://docs.sdv.dev/sdv/
- others…?

… generation tools for PostgreSQL

- Tool PostgreSQL Anonymizer
  - PostgreSQL license, Ruby, by Damien Clochard, Dalibo, Paris
  - https://labs.dalibo.com/postgresql_anonymizer
- Tool PGFaker
  - MIT license, TypeScript, by Imanpal Singh, India
  - Weiterentwicklung von pg-anonymizer
  - https://github.com/imanpalsingh/pg-faker
- Tool Google Differential Privacy
  - Apache 2.0 license C++, by Google (not officially supported)
  - https://github.com/google/differential-privacy
- Tool pgsynthdata
  - MIT license, Python, by Institute for Software, OST Rapperswil
  - https://gitlab.com/geometalab/pgsynthdata

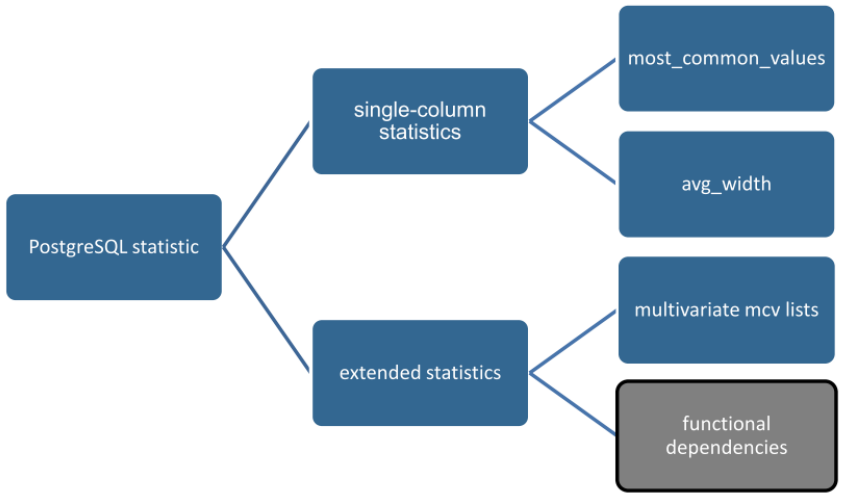
The pgsynthdata Tool 

PostgreSQL internal statistics

```
SELECT reltuples, relpages, relallvisible
FROM pg_class WHERE relname = 'flights';
 reltuples | relpages | relallvisible
-----------+----------+---------------
    214867 |     2624 |          2624
(1 row)
```

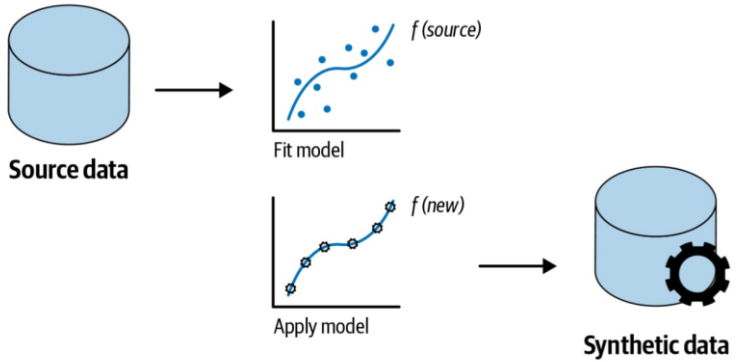
```
SELECT most_common_vals AS mcv,
  left(most_common_freqs::text,60) || '...' AS mcf
FROM pg_stats
WHERE tablename = 'flights' AND attname = 'aircraft_code' \ gx
-[ RECORD 1 ]-----------------------------------------------------------
mcv | {CN1,CR2,SU9,321,763,733,319,773}
mcf | {0.2783,0.27473333,0.25816667,0.059233334,0.038533334,0.0370...
```

```
SELECT left(histogram_bounds::text,60) || '...' AS histogram_bounds
FROM pg_stats s
WHERE s.tablename = 'boarding_passes' AND s.attname = 'seat_no';
                      histogram_bounds
-------------------------------------------------------------------
 {10B,10D,10D,10F,11B,11C,11H,12H,13B,14B,14H,15H,16D,16D,16H...
(1 row)
```



What is pgsynthdata?

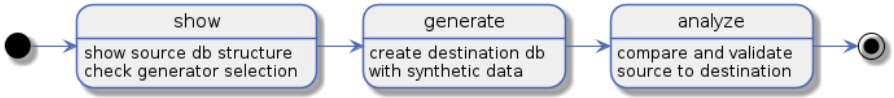
- CLI tool for PostgreSQL, which creates synthetic data
- Mathematic Model taken from PostgreSQL internal statistics
- Little configuration needed (configuration by comments, e.g. NAME_GENERATOR)
- Written in Python
- MIT Open Source License
- Maintainer: Institute for Software OST



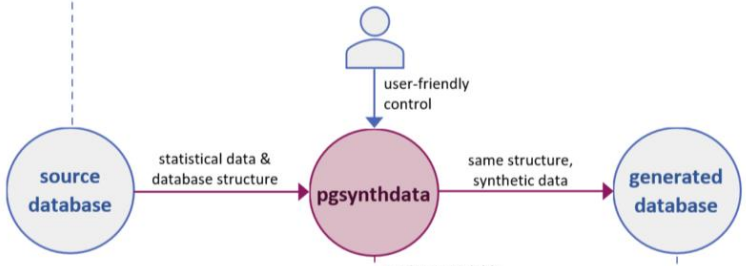
Source: Practical Synthetic Data Generation (2020), Emam, Mosquera & Hoptroff, O'Reilly

How does pgsynthdata work?

Workflow:



id [PK] integer	name text	salary real	kids ages smallint[]	location point	dob date
1001	Peter Pan	5443	{2,5,7}	(27.9455...	1983-12-24
1002	Nora Niemand	6543	[null]	(40.9242...	1992-05-12
1003	Max Muster	3799	{16}	(65.4916...	1957-07-05



id [PK] integer	name text	salary real	kids ages smallint[]	location point	dob date
1001	Paulina Michel	4223	{6,8,4}	(361,797)	1964-05-26
1002	Thomas Frick	5324	{4}	(696,524)	1976-12-30
1003	Nikolaus Forst...	4537	[null]	(324,48)	1976-08-18

pgsynthdata: Characteristics

- Able to generate synthetic data from various PostgreSQL databases and with generators for a wide range of data types
- Maintainable and extensible with a plugin system and own generators (to be programmed in Python)
- Suitable at least for benchmarking
- In short:
  - Easy-to-use, low config
  - Easy-to-extend

pgsynthdata: Outlook

- Possible extensions:
  1. Generic faker for all base data types
  2. Support for composite primary keys
  3. Support for self-referencing foreign keys
  4. Spatial data types Point, LineString, Polygon
  5. Make it suitable for ML functionality (new training phase)
  6. A bit under maintained

- Current developments spring semester 2023:
  - Student project of André Von Aarburg about (4) Point and (5) ML

pgsynthdata extended by GeoPointGAN

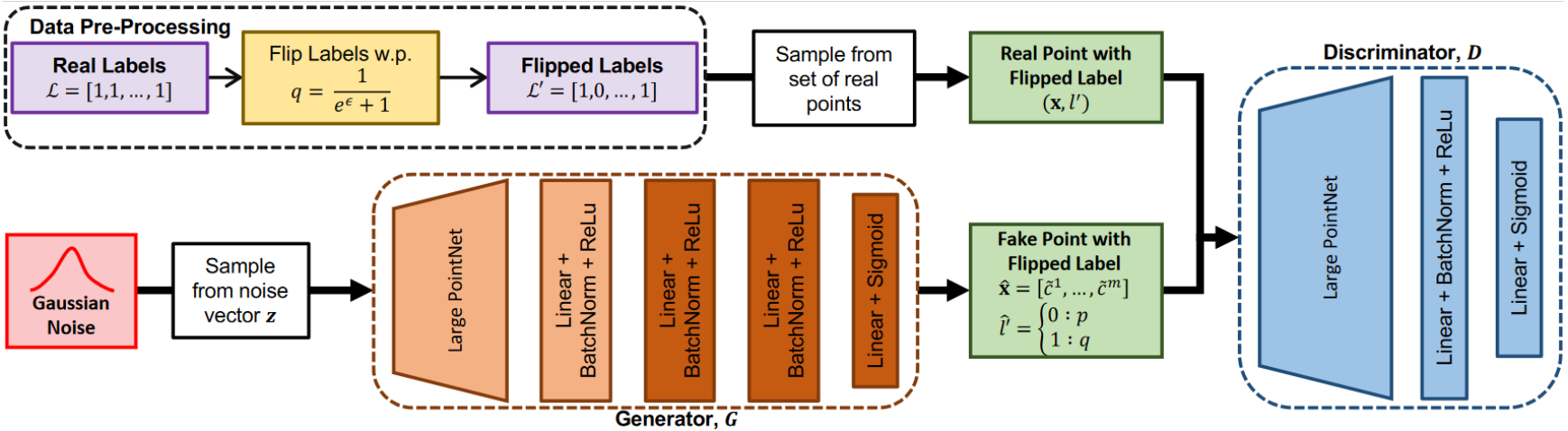
- Student project spring semester 2023 (ongoing)
- Goal: To implement the paper by
  - Cunningham, Klemmer, Wen & Ferhatosmanoglu (2022). GeoPointGAN: Synthetic Spatial Data with Local Label Differential Privacy
  - a generative model for geographic point coordinates with a privacy mechanism
- …and to integrate it in pgynthdata

GeoPointGAN: Visualization

GeoPointGAN generated and privatized data: 311 caller locations in New York.
(Source: Cunningham et al. (2022). GeoPointGAN…)

GeoPointGAN: Processing workflow

GeoPointGAN pipeline including privacy mechanism.
(Source: Cunningham et al. (2022). GeoPointGAN…)



Discussion

Prof. Stefan Keller
Institute for Software
OST Campus Rapperswil
Oberseestrasse 10
CH-8640 Rapperswil
stefan.keller@ost.ch
www.ost.ch/ifs
@sfkeller @SwissPUGOrg