Private Machine Learning
The EPI Project

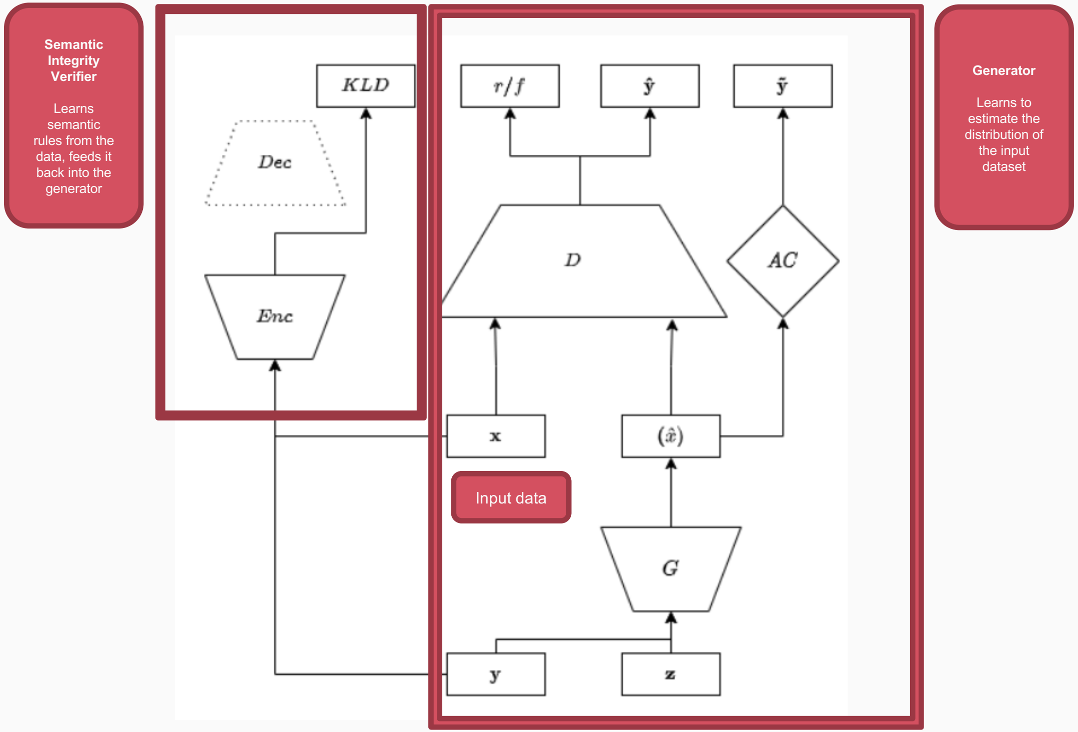
Saba Amiri
Adam Belloum, Eric Nalisnick, Sander Klous, Leon Gommans

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Problem
  - Alternative approach to making ML models private: generate private data, use without limitation
  - Task? estimating the true joint distribution of the input data
  - Output? a model that can generate unlimited synthetic records with the same statistical properties as real data
  - Evaluation? statistical tests, machine learning efficacy
  - Why focus on semantic integrity? generative models are probabilistic, even an effective model could possibly generate samples that are in distribution, but semantically incorrect, e.g. a patient over 200 years old, a female patient with prostate cancer
  - How?
    - Supervised: rule based, could be very expensive, we might not know all the rules out there
    - Unsupervised: learn the rules from the data itself

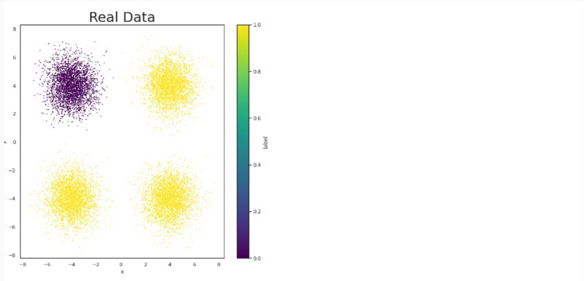
Tabular Synthetic Data Generation with Improved Semantic Integrity

- Our method



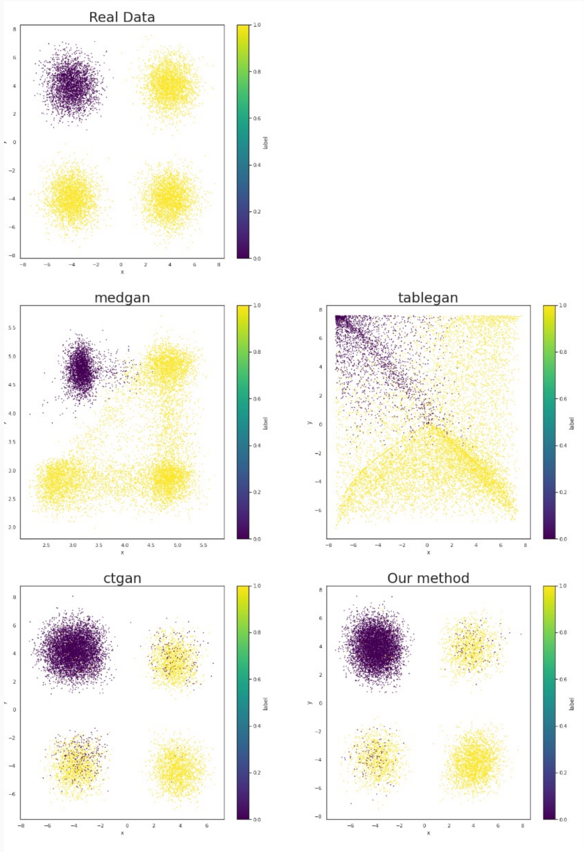
Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
    - Toy dataset
    - Two classes, four modes
    - Aim: estimate the distribution while labelling the samples accurately



Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
  - Toy dataset
  - Two classes, four modes
  - Aim: estimate the distribution while labelling the samples accurately

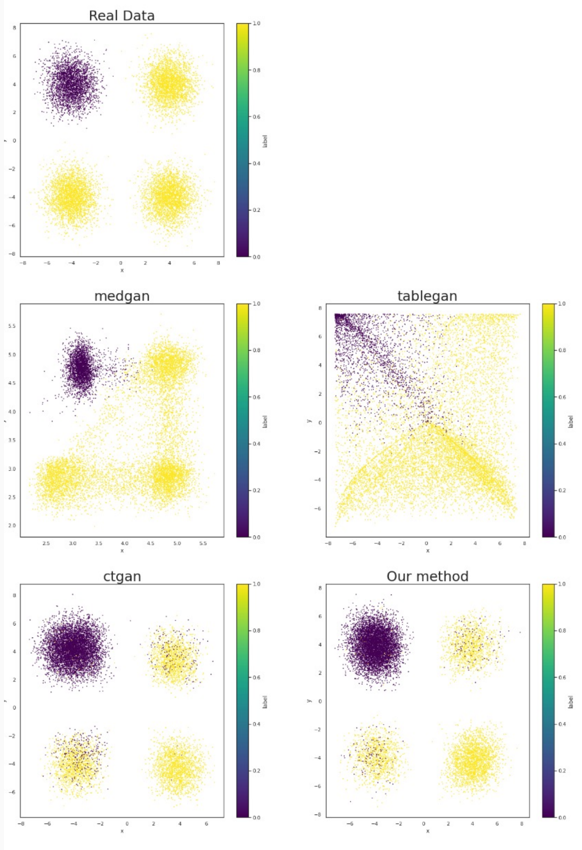


Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
  - Toy dataset
  - Two classes, four modes
  - Aim: estimate the distribution while labelling the samples accurately

Table 1: Label accuracy for toy dataset

Model	Correct labels
medgan	74.8%
tablegan	92.9%
ctgan	92.1%
Our method	**97.4%**



Tabular Synthetic Data Generation with Improved Semantic Integrity

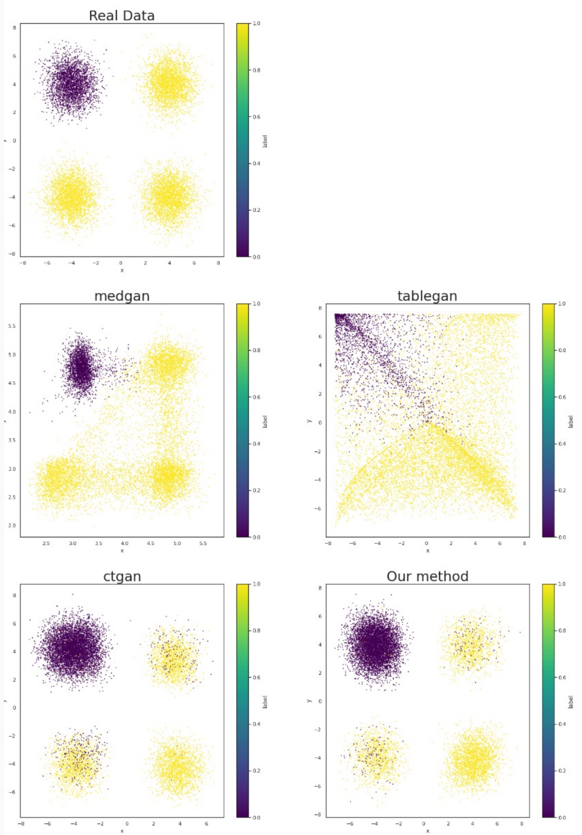
- Results compared with four state of the art models
  - Toy dataset
  - Two classes, four modes
  - Aim: estimate the distribution while labelling the samples accurately

Table 1: Label accuracy for toy dataset

Model	Correct labels
medgan	74.8%
tablegan	92.9%
ctgan	92.1%
Our method	**97.4%**

Table 2: Record distance

Model	Record distance
medgan	802
tablegan	6322
ctgan	2148
Our method	3941



Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Table 3: Memorization results

Model	Detection score
medgan	0.001
tablegan	0.49
ctgan	0.48
Our method	**0.50**

- How similar are the two datasets?
- Can we distinguish between the real and fake data samples successfully?

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results compared with four state of the art models
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Table 3: Memorization results

Model	Detection score
medgan	0.001
tablegan	0.49
ctgan	0.48
Our method	**0.50**

Table 4: Machine learning efficacy scores delta

Model	Accuracy	F1-score	F2-score
medgan	-0.32	-0.29	-0.22
tablegan	-0.35	-0.27	-0.21
ctgan	-0.24	-0.25	-0.18
Our method	**+0.01**	**+0.04**	**+0.03**

- **If we train a ML model on two datasets, one real and one synthetic, will there be a noticable difference?**
- **Will the performance drop if the model is trained on synthetic data?**

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Table 3: Memorization results

Model	Detection score
medgan	0.001
tablegan	0.49
ctgan	0.48
Our method	**0.50**

Table 4: Machine learning efficacy scores delta

Model	Accuracy	F1-score	F2-score
medgan	-0.32	-0.29	-0.22
tablegan	-0.35	-0.27	-0.21
ctgan	-0.24	-0.25	-0.18
Our method	**+0.01**	**+0.04**	**+0.03**

Table 5: Semantic integrity - C1 and C2 features; Record distance

Model	Females - C1	Females - C2	Males - C1	Males - C2
Real data	*70%*	*5%*	*0%*	*0%*
medgan	59.2%	0%	0%	0%
tablegan	72.8%	1.3%	0%	0.4%
ctgan	71.8%	14.9%	3.2%	0.9%
Our method	72%	14.7%	**1.7%**	**0.3%**

How much semantically incorrect samples are we generating?

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Table 3: Memorization results

Model	Detection score
medgan	0.001
tablegan	0.49
ctgan	0.48
Our method	**0.50**

Table 4: Machine learning efficacy scores delta

Model	Accuracy	F1-score	F2-score
medgan	-0.32	-0.29	-0.22
tablegan	-0.35	-0.27	-0.21
ctgan	-0.24	-0.25	-0.18
Our method	**+0.01**	**+0.04**	**+0.03**

Table 5: Semantic integrity - C1 and C2 features; Record distance

Model	Females - C1	Females - C2	Males - C1	Males - C2
Real data	*70%*	*5%*	*0%*	*0%*
medgan	59.2%	0%	0%	0%
tablegan	72.8%	1.3%	0%	0.4%
ctgan	71.8%	14.9%	3.2%	0.9%
Our method	72%	14.7%	**1.7%**	**0.3%**

How much semantically incorrect samples are we generating?

Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Adult dataset: US census data
  - Long tailed features, minority classes
  - Two binary control features
    - C1: 5% of females positive, all men negative
    - C2: 70% of females positive, all men negative
  - Aim: estimate the distribution accurately
    - Without generating samples of males with C1/C2 positive
    - Without suppressing the female C1/C2 positives – erasing the problem

Table 3: Memorization results

Model	Detection score
medgan	0.001
tablegan	0.49
ctgan	0.48
Our method	**0.50**

Table 4: Machine learning efficacy scores delta

Model	Accuracy	F1-score	F2-score
medgan	-0.32	-0.29	-0.22
tablegan	-0.35	-0.27	-0.21
ctgan	-0.24	-0.25	-0.18
Our method	**+0.01**	**+0.04**	**+0.03**

Table 5: Semantic integrity - C1 and C2 features; Record distance

Model	Females - C1	Females - C2	Males - C1	Males - C2
Real data	*70%*	*5%*	*0%*	*0%*
medgan	59.2%	0%	0%	0%
tablegan	72.8%	1.3%	0%	0.4%
ctgan	71.8%	14.9%	3.2%	0.9%
Our method	72%	14.7%	**1.7%**	**0.3%**

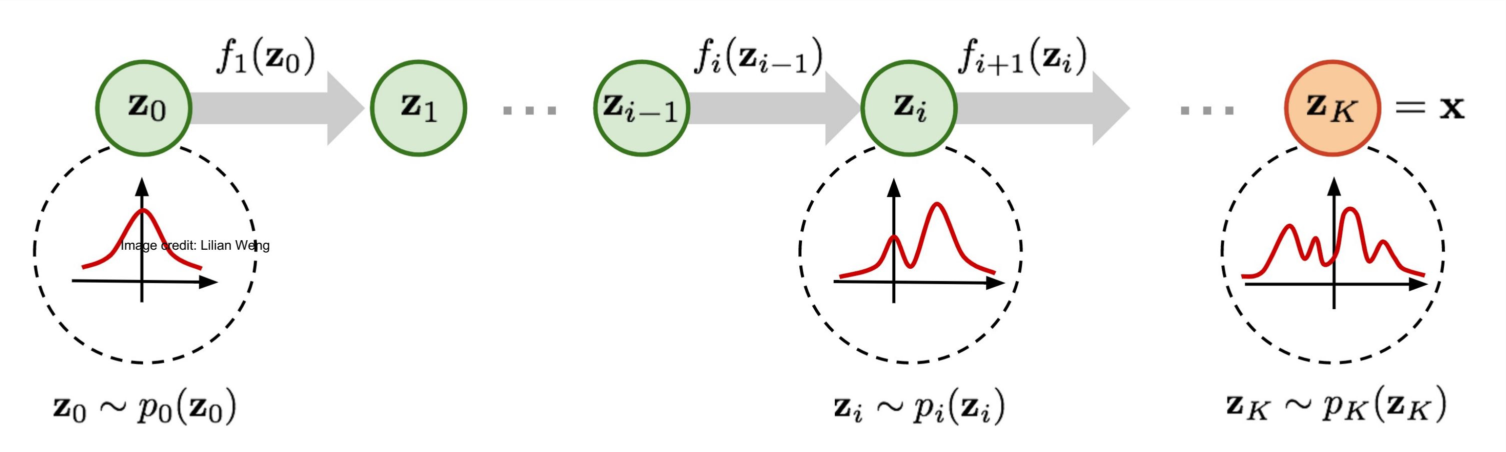
How much semantically incorrect samples are we generating?

Generating Heavy-Tailed Synthetic Data with Normalizing Flows

- Problem
  - Alternative approach to making ML models private: generate private data, use without limitation
  - Task? estimating the true joint distribution of the input data
  - Output? a model that can generate unlimited synthetic records with the same statistical properties as real data
  - Evaluation? statistical tests, machine learning efficacy

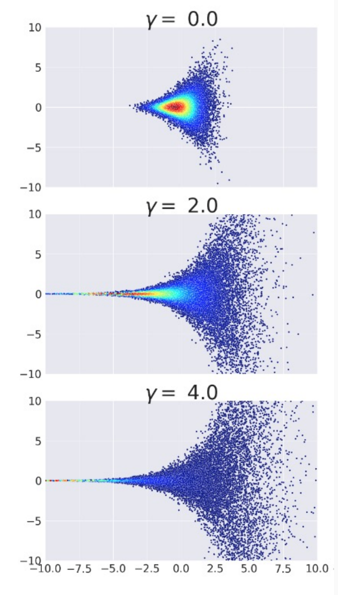
Tabular Synthetic Data Generation with Improved Semantic Integrity

- Our method
  - We use normalizing flows
  - We propose changes in the architecture to help the model better capture the tail of the input data



Tabular Synthetic Data Generation with Improved Semantic Integrity

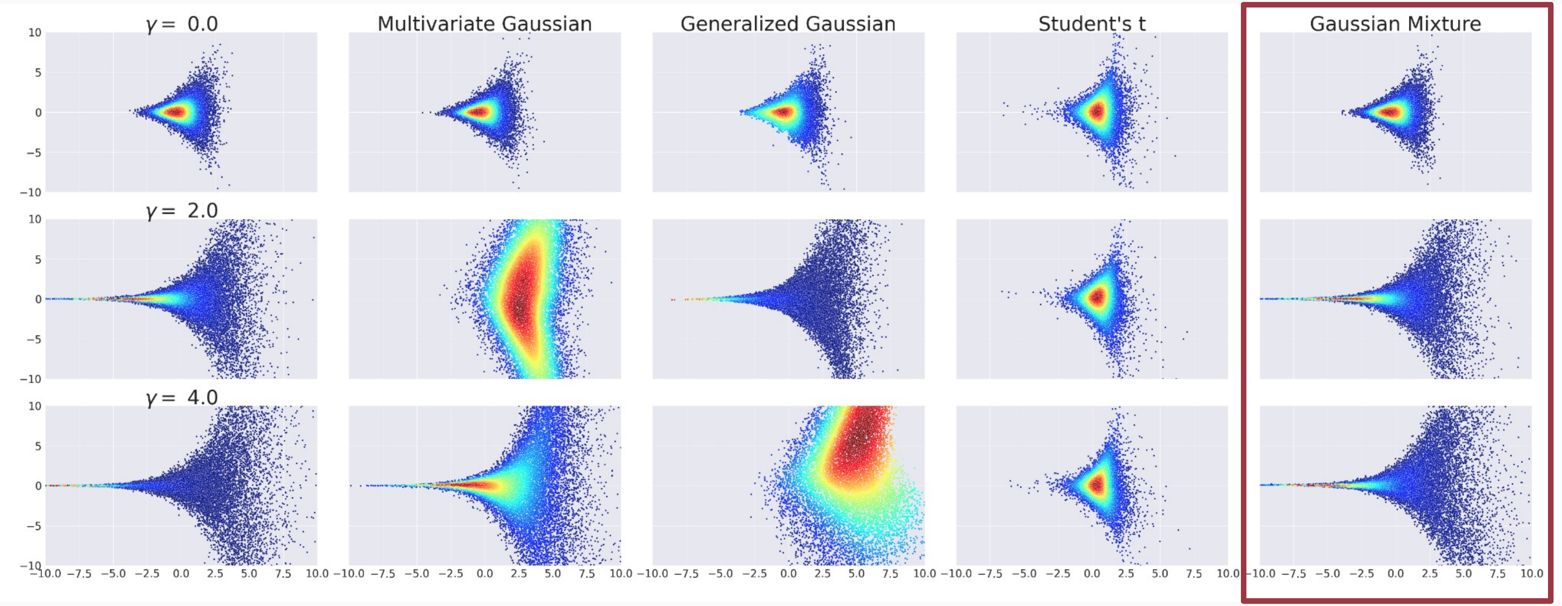
- Results
  - Toy dataset: samples from Neal's funnel
  - Aim is to accurately estimate the input data distribution and its tail properties



Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Toy dataset: samples from Neal's funnel
  - Aim is to accurately estimate the input data distribution and its tail properties

Our proposed method replacing the simple base distribution with a mixture

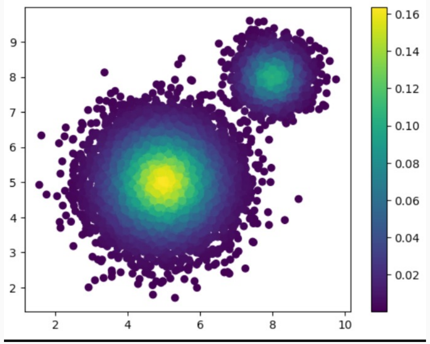


Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Toy dataset: samples from Neal's funnel
  - Aim is to accurately estimate the input data distribution and its tail properties
  - Experiments on real datasets also show the same sort of improvement in both general performance and in capturing the tail behaviour
  - Second contribution: targeted sampling

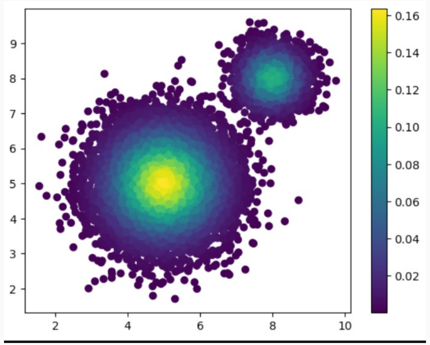
Tabular Synthetic Data Generation with Improved Semantic Integrity

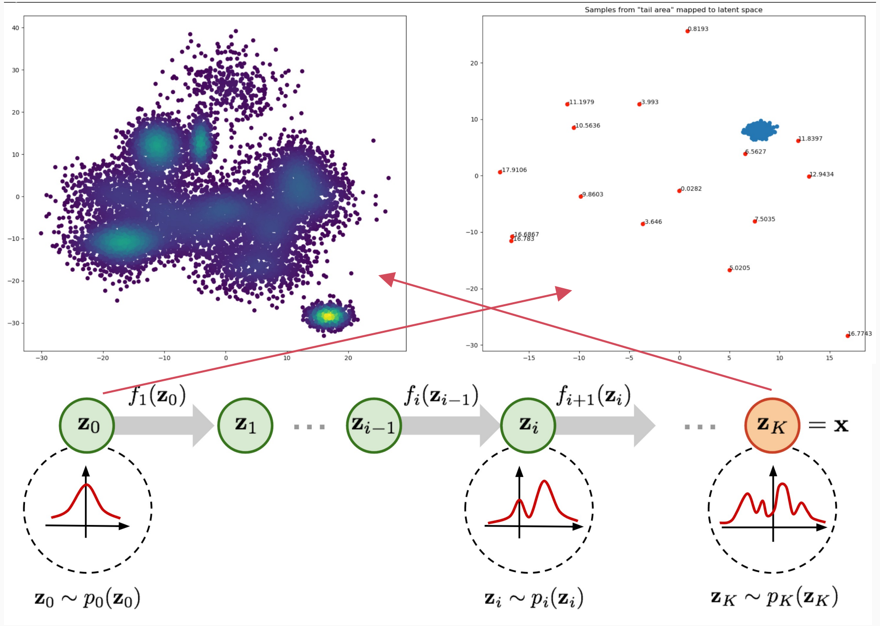
- Results
  - Toy dataset: mixture of Gaussians
  - Second contribution: targeted sampling



Tabular Synthetic Data Generation with Improved Semantic Integrity

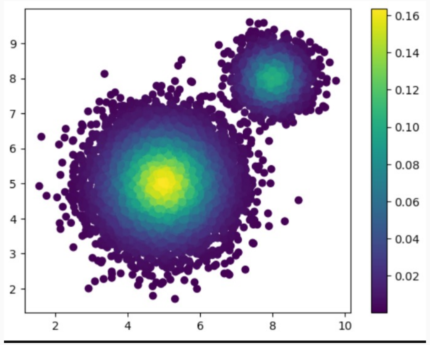
- Results
  - Toy dataset: mixture of Gaussians
  - Second contribution: targeted sampling

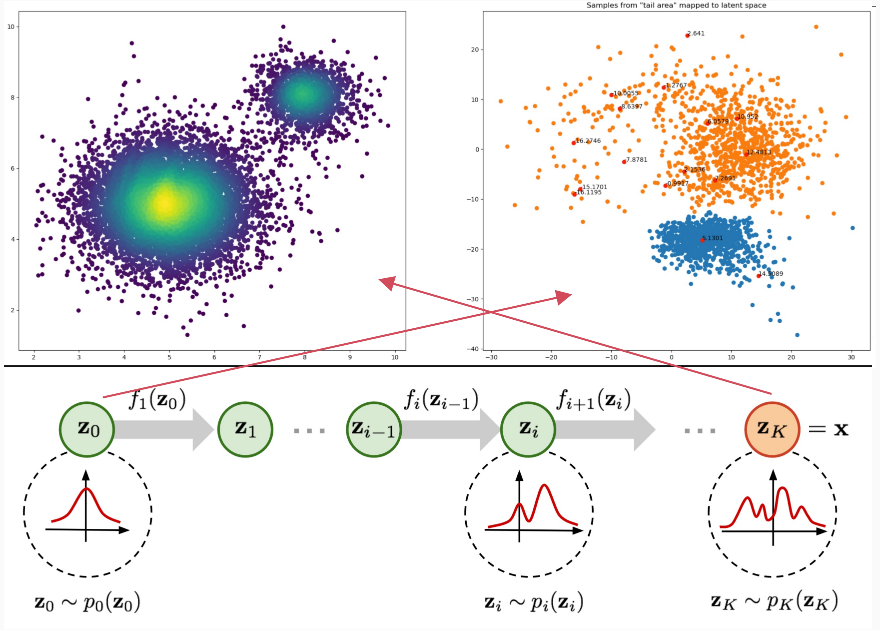




Tabular Synthetic Data Generation with Improved Semantic Integrity

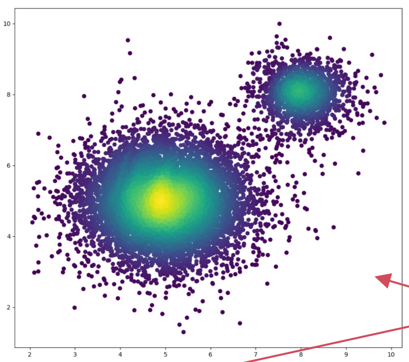
- Results
  - Toy dataset: mixture of Gaussians
  - Second contribution: targeted sampling

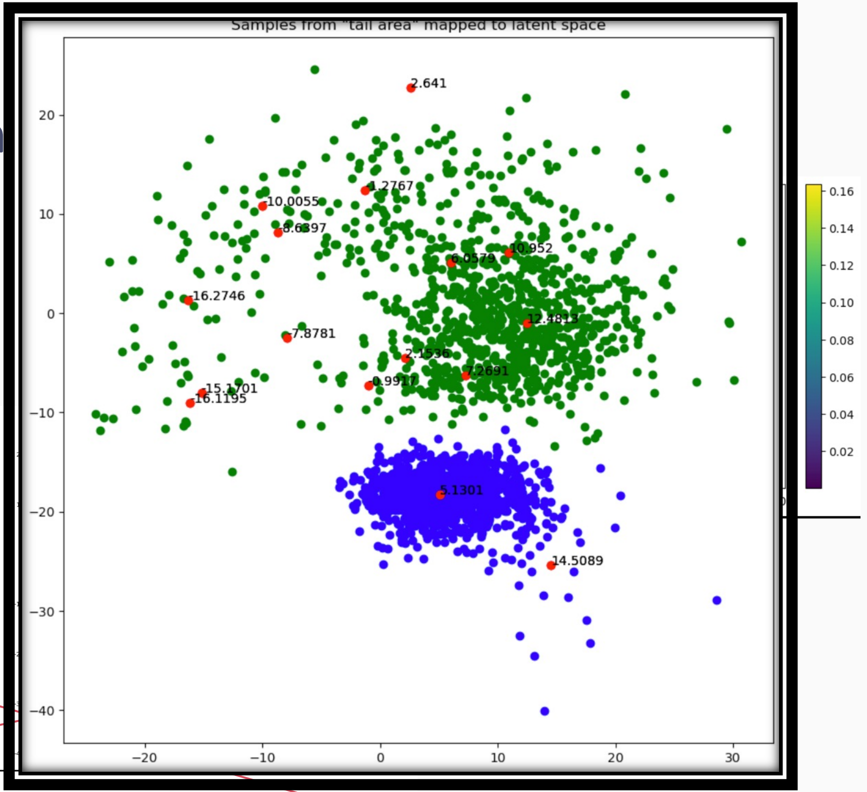


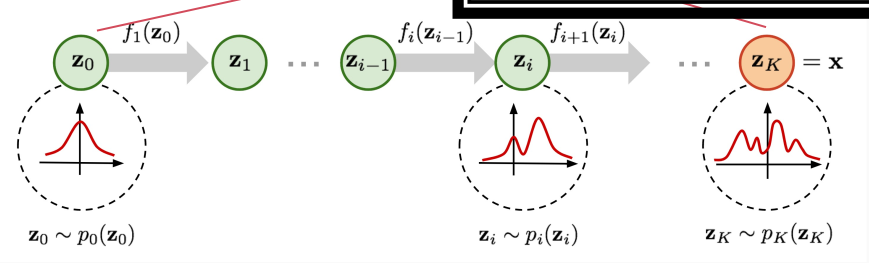


Tabular Synthetic Data Generation with

- Results
  - Toy dataset: mixture of Gaussians
  - Second contribution: targeted sampling

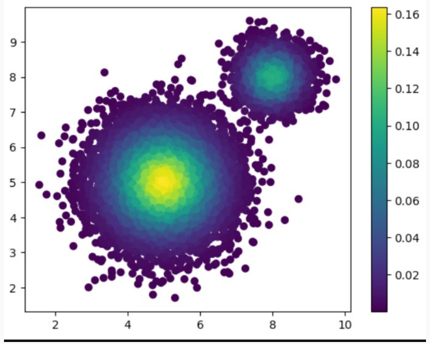


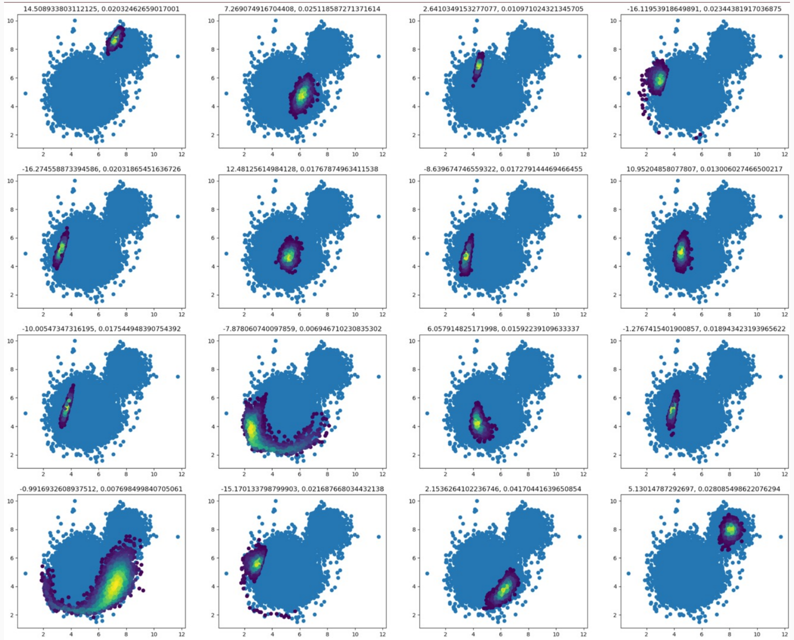




Tabular Synthetic Data Generation with Improved Semantic Integrity

- Results
  - Toy dataset: mixture of Gaussians
  - Second contribution: targeted sampling





Thank You!