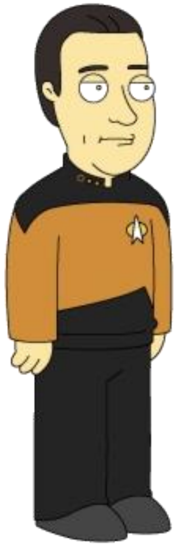
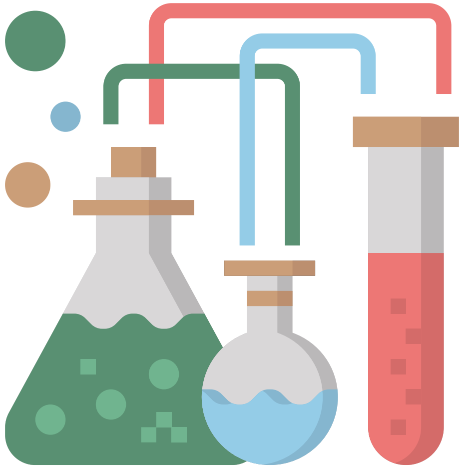
{Tabular} + {Data} + {Synthesis}





for Data Management

Frühjahrstreffen 2024 der Fachgruppe Datenbanken

Fabian Panse
fabian.panse@hpi.de

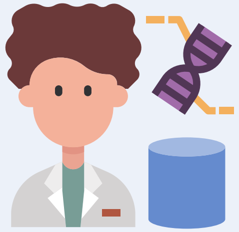
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Outline

Motivation & Challenges

Existing Solutions

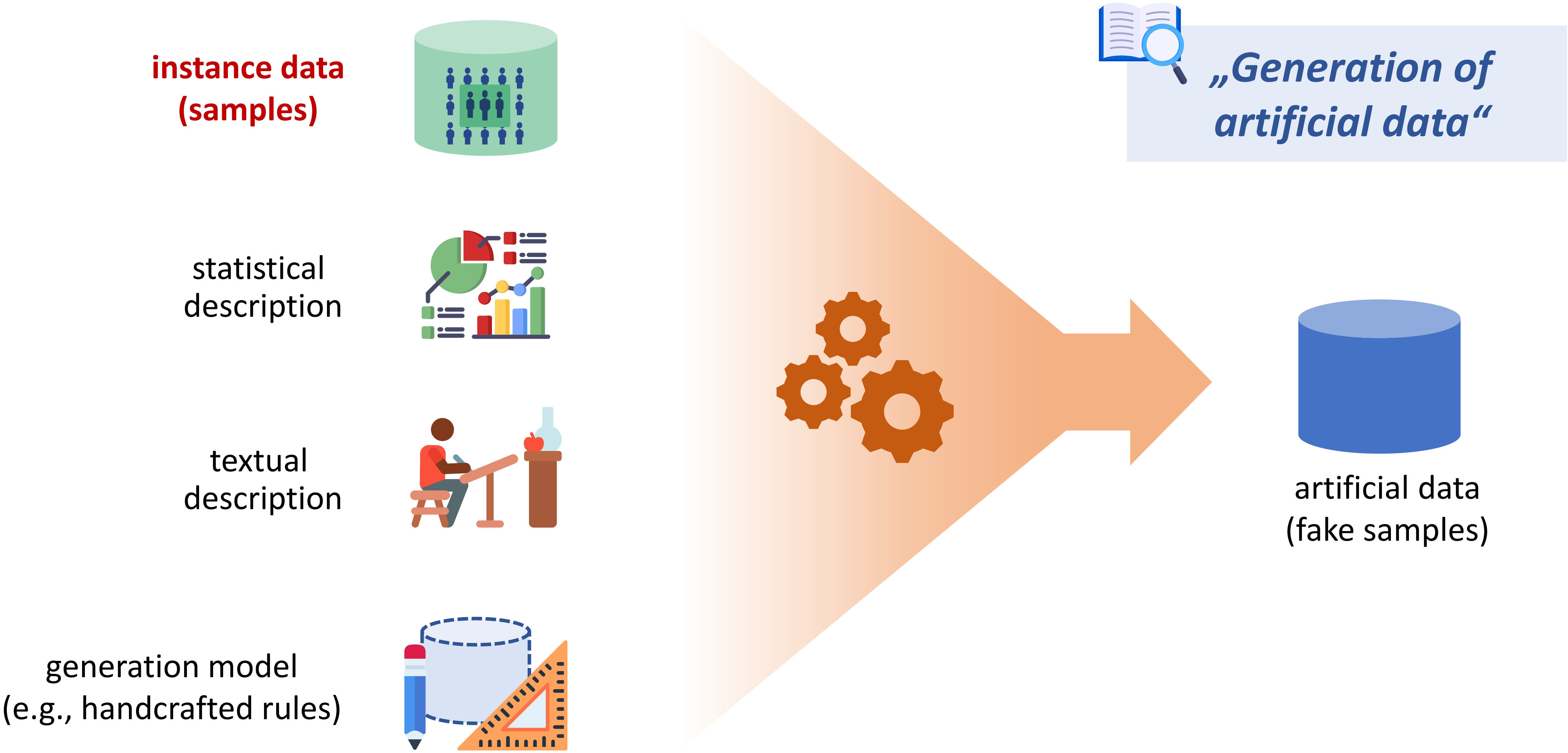
Challenges for Data Management



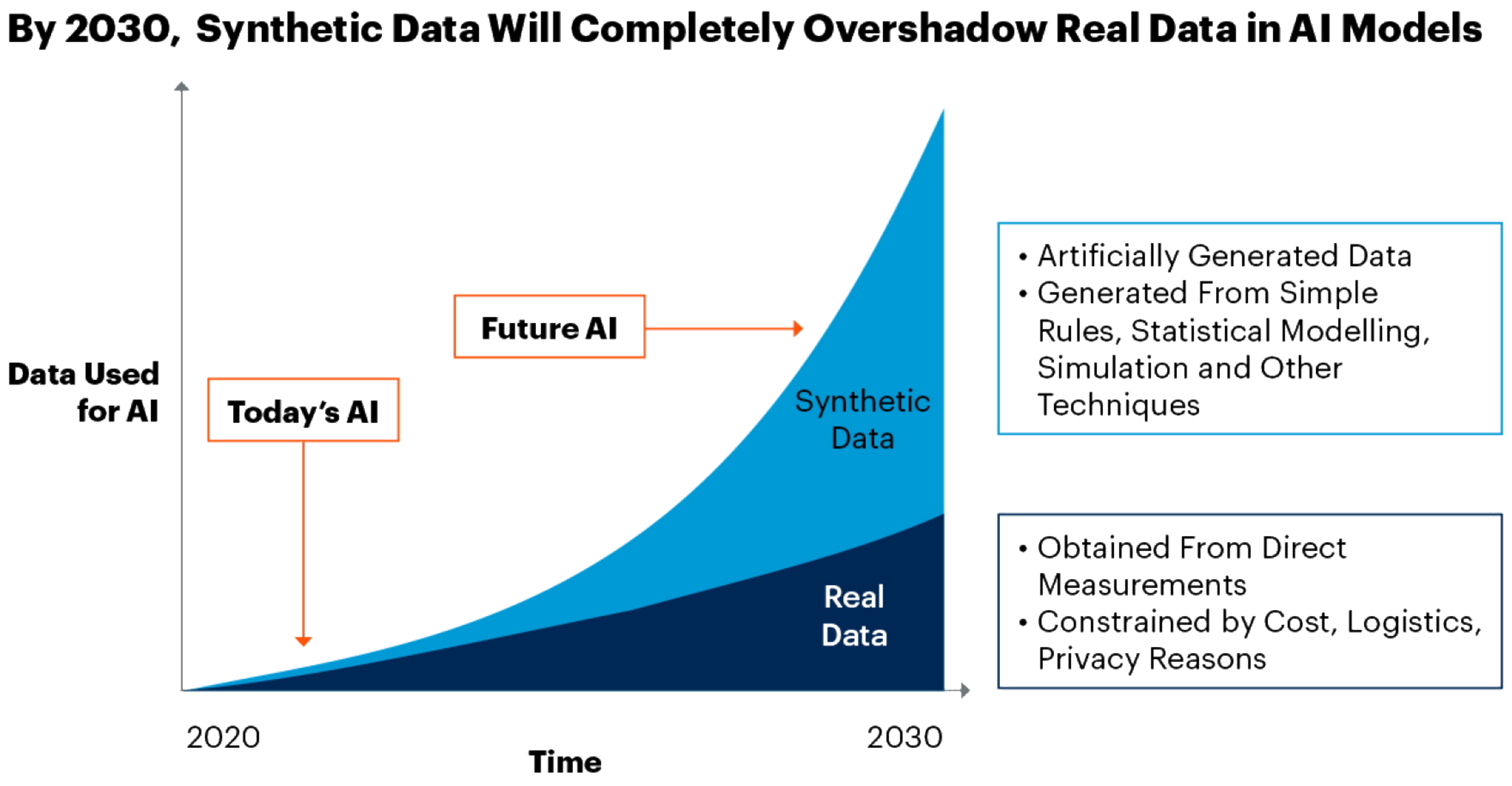
Data Synthesis for Research Data

MOTIVATION
& GENERAL CHALLENGES

What is Data Synthesis?



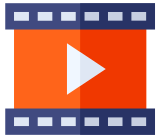
The Future Role of Synthetic Data



Gartner: "Maverick Research: Forget About Your Real Data – Synthetic Data Is the Future of AI," Leinar Ramos, Jitendra Subramanyam, 24 June 2021.

What Data can we synthesize?

KREA



"Imagine a scene where a federation space vessel is orbiting a Class M planet."

DALL·E

"Write a 4-Line poem that captures the essence of exploring the universe, inspired by "Star Trek.“."

ChatGPT 3.5

"In galaxies vast, we seek to roam,
Boldly venturing far from home.
With stars as guides, our spirits soar,
Infinite horizons, forever we explore."

Tabular Data Synthesis

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764

ChatGPT 3.5

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Samuel	White	Colonel	Human	Male	2290	2375	SW-522-619	1.65
Carlos	Rodriguez	Lieutenant	Human	Male	2256	Alive	CR-735-124	1.78
Mei-Ling	Chen	Major	Human	Female	2301	2378	MC-912-503	1.70
Mikhail	Ivanov	Captain	Human	Male	2240	2370	MI-623-818	1.82
Isabella	Santos	Lt. Commander	Human	Female	2285	2360	IS-409-278	1.68
Hiroshi	Tanaka	Lieutenant	Human	Male	2268	2385	HT-827-615	1.75

Tabular Data Synthesis

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
NFN	Spock	Captain	Android	Male	2263.0	Alive	-	1.78
Nyota	Uhura	Lieutenant	Human	Male	-	Alive	-	1.83
Jean-Luc	Picard	Captain	Human	Male	2305.0	2399	SP-937-215	1.6764
S'Chn T'Gai	Spock	Captain	Vulcan	Male	2399.0	2372	S 179-276SP	1.8
Nyota	Data	Lieutenant	Human	Male	-	Alive	-	1.8
James	Kirk	Captain	Human	Male	2245.0	2372	SP-937-215	1.83

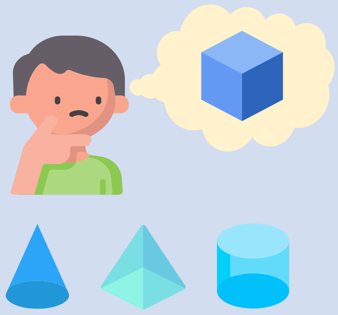
Why do we need to synthesize Tabular Data?

Many industrial and research datasets **cannot be shared** due to **privacy regulations**.

80% of industrial data is never used[1]

High-quality **training data** for machine learning is **hard to obtain**, especially **labeled** data.

Many training datasets contain **data biases**, leading to learned models **reinforcing** those biases.



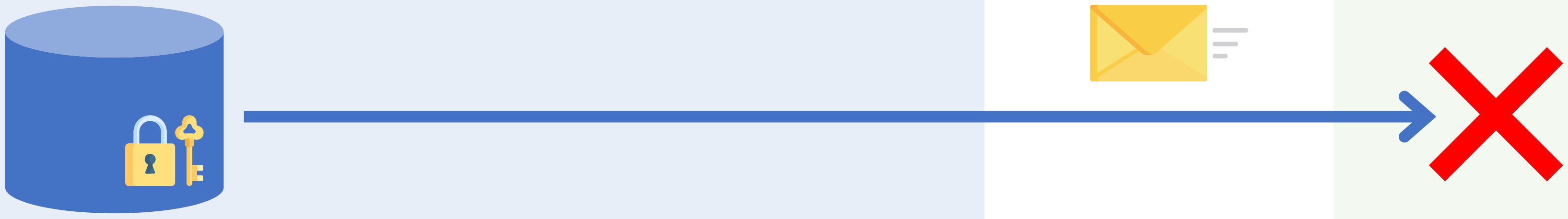
For many use cases, there is simply **no training data,** and data from **similar use cases** must be used instead.

1. Data Act: Commission proposes measures for a fair and innovative data economy (23.02.2022) https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113

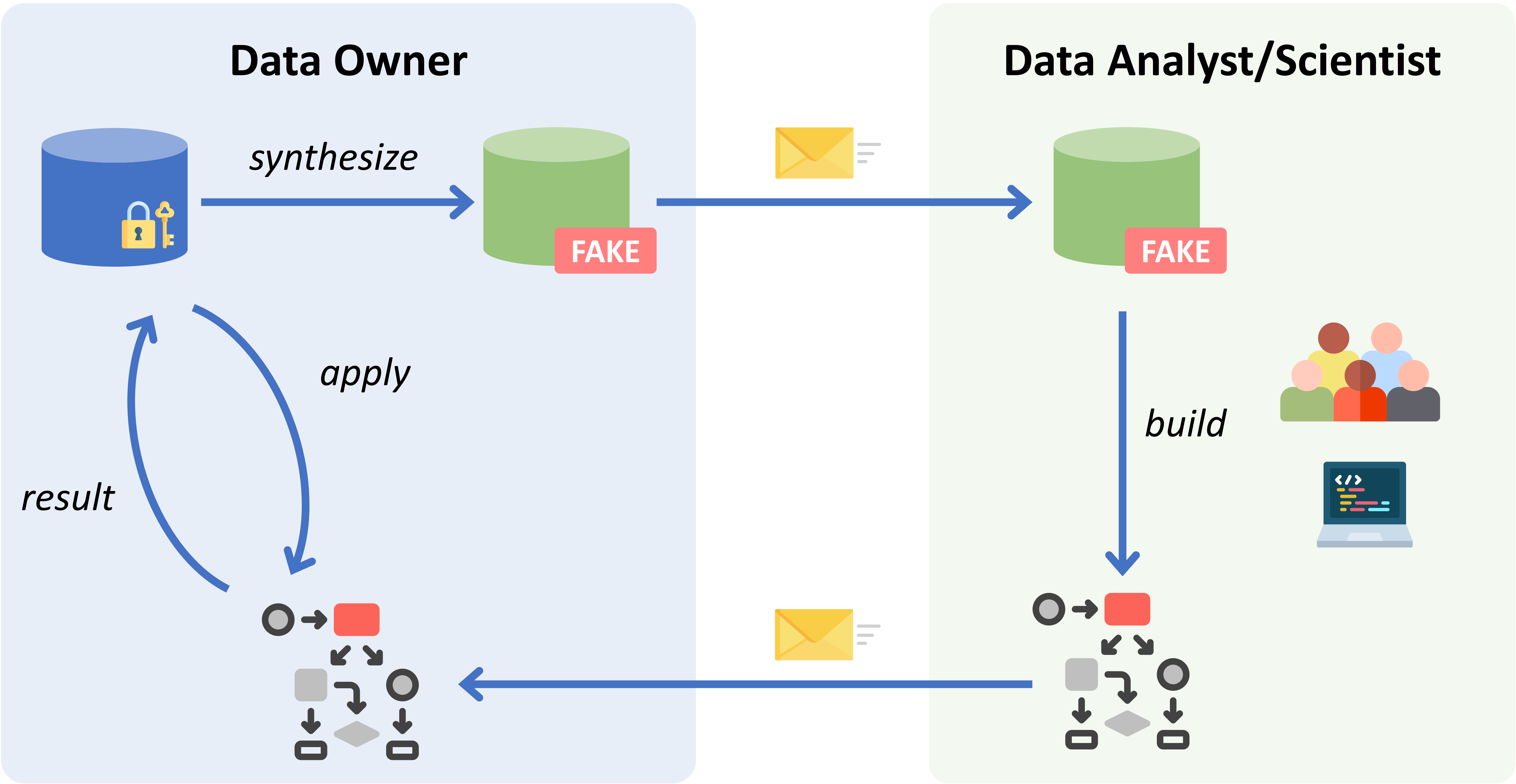
Privacy-Preserving Data Sharing

Data Owner

Data Analyst/Scientist

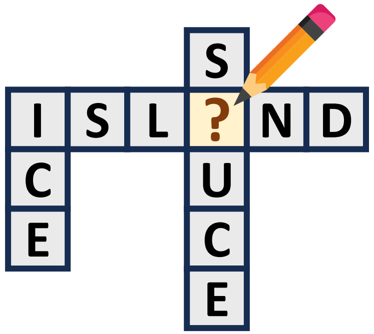


Privacy-Preserving Data Sharing



Further Purposes

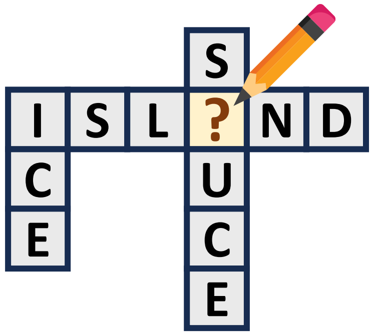
1 Missing Value Imputation



FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764

Further Purposes

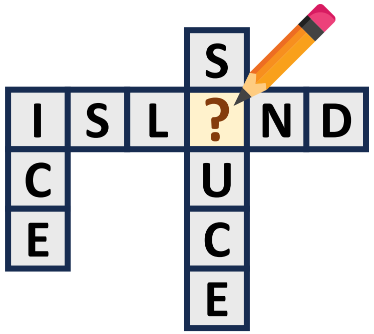
1 Missing Value Imputation



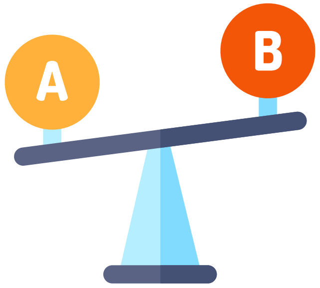
FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	1.78
James	Kirk	Captain	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	2233	Alive	NU-937-213	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	ND-937-218	1.8
Pavel	Chekov	Lieutenant	Human	Male	2245	Alive	656-5827B	1.6764

Further Purposes

1 Missing Value Imputation



2 Rebalancing

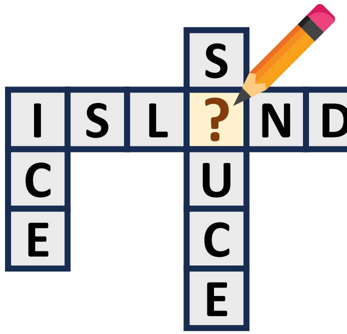


Data Imbalance

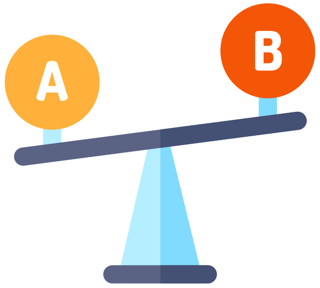
FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764
Isabella	Santos	Lt. Commander	Human	Female	2285	2360	IS-409-278	1.68
Mei-Ling	Chen	Major	Android	Female	2301	2378	MC-912-503	1.70
Aisha	Khan	Lieutenant	Android	Female	2297	Alive	AK-319-526	1.70

Further Purposes

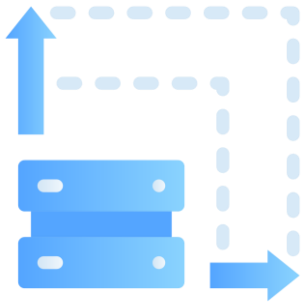
1 **Missing Value Imputation**



2 **Rebalancing**



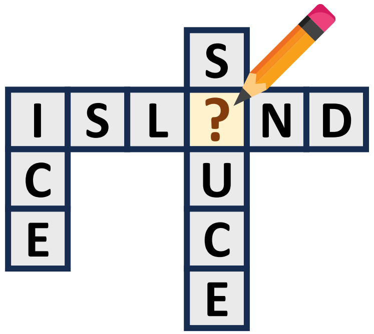
3 **Augmentation**



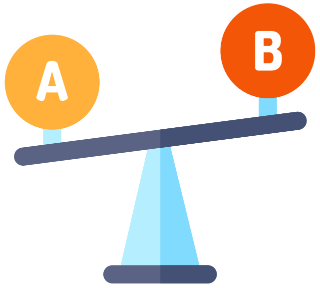
FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764
Samuel	White	Colonel	Human	Male	2290	2375	SW-522-619	1.65
Carlos	Rodriguez	Lieutenant	Human	Male	2256	Alive	CR-735-124	1.78
Mei-Ling	Chen	Major	Human	Female	2301	2378	MC-912-503	1.70

Further Purposes

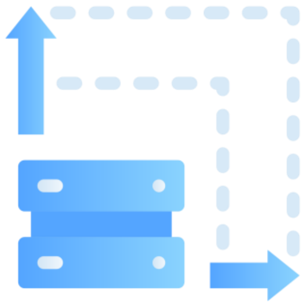
1 **Missing Value Imputation**



2 **Rebalancing**



3 **Augmentation**



4 **Customization**

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Samuel	White	Colonel	Human	Male	2320	Alive	SW-522-619	1.65
Carlos	Rodriguez	Lieutenant	Human	Male	2356	Alive	CR-735-124	1.78
Mei-Ling	Chen	Major	Human	Female	2301	Alive	MC-912-503	1.70
Mikhail	Ivanov	Captain	Human	Male	2340	Alive	MI-623-818	1.82
Isabella	Santos	Lt. Commander	Human	Female	2325	Alive	IS-409-278	1.68
Hiroshi	Tanaka	Lieutenant	Human	Male	2368	Alive	HT-827-615	1.75

„Only humans born after 2300 who are still alive"

Challenges

Col1	Col2	Col3	Col4	Col5	Col6
1	X			3	
1		Z	Q	4	X
2	X	Y			X

(1) Missing Values

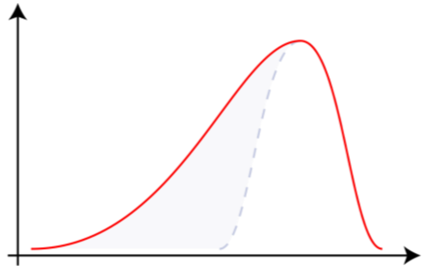
Type	Example
Int	{1,2,3}
Decimal	[-1.78,2.3]
Boolean	{t,f}
Categorical	{s,n,w,e}
Texts	[a-z]*
Mixed	{‚no‘,1,2,3}

(2) Diverse Column Types

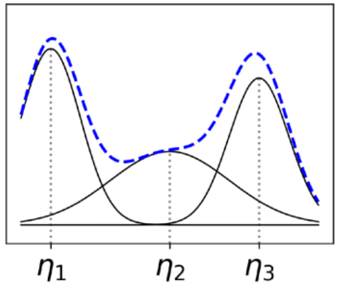
User Class	Frequency
Daily	38,769
Casual	101,398
Fraudulent	219

(3) Class Imbalance

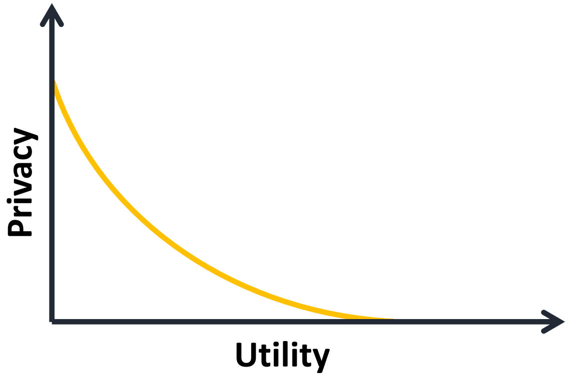
skewed [*]



multimodal [**]



(4) Diverse Column Distributions



(5) Privacy vs. Utility Trade-Off

[*] https://en.wikipedia.org/wiki/Skewness [**] L. Xu, M. Skoularidou, et al. Modeling Tabular Data using Conditional GAN. NeurIPS, 2019.

Tabular Data Synthesis

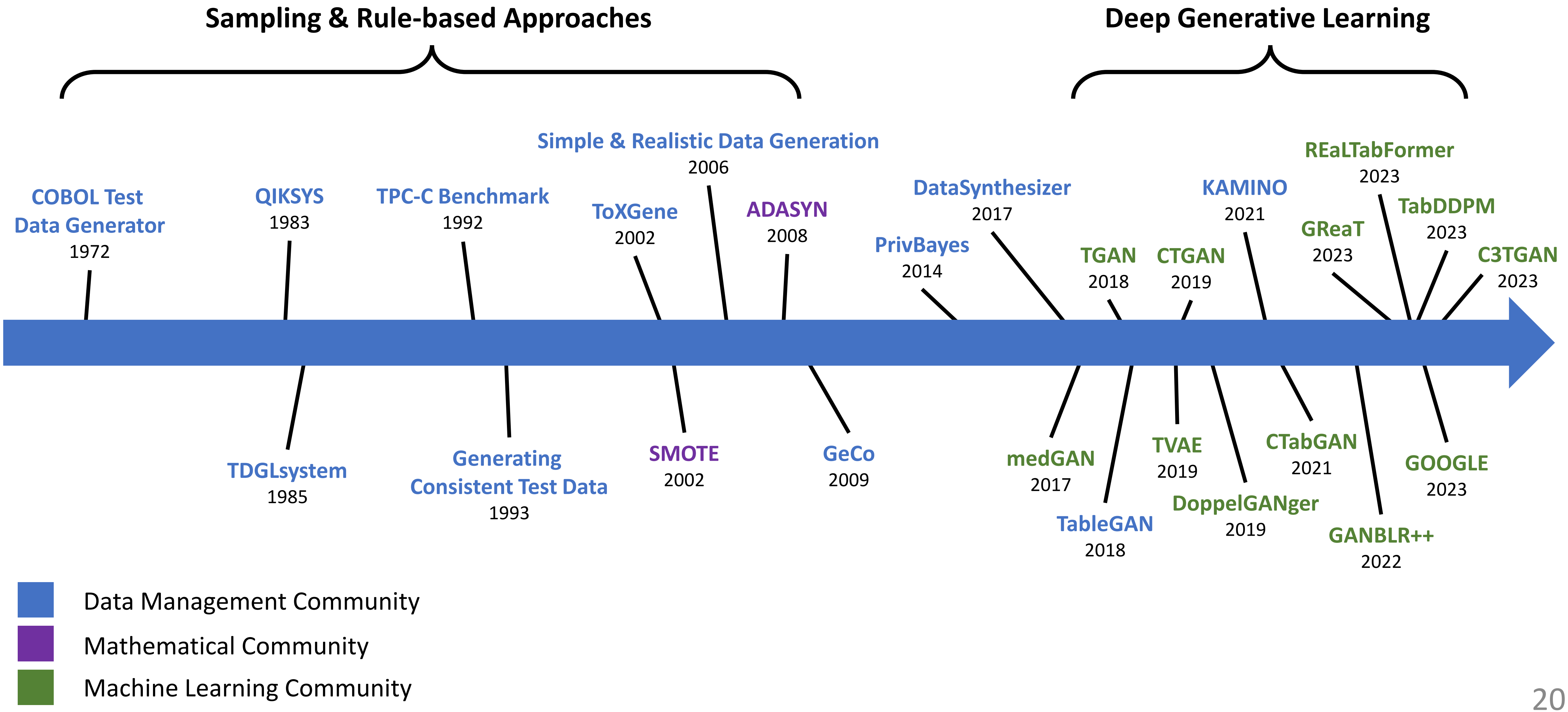
Mixed Data Type

FName	LName	Rang	Race	Gender	YoB	YoD	Service Number	Height
Jean-Luc	Picard	Captain	Human	Male	2305	2399	SP-937-215	-
James	Kirk	-	Human	Male	2233	2371	SC 937-0176 CEC	1.78
S'Chn T'Gai	Spock	Commander	{Vulcan, Human}	Male	2230	2263	S 179-276SP	1.83
Nyota	Uhura	Lieutenant	Human	Female	-	Alive	-	1.60
NFN	Data	Lt. Commander	Android	Male	2338	2399	-	1.8
Pavel	Chekov	Lieutenant	Human	-	2245	Alive	656-5827B	1.6764

Multi-modal distribution (one mode per race-gender combination)

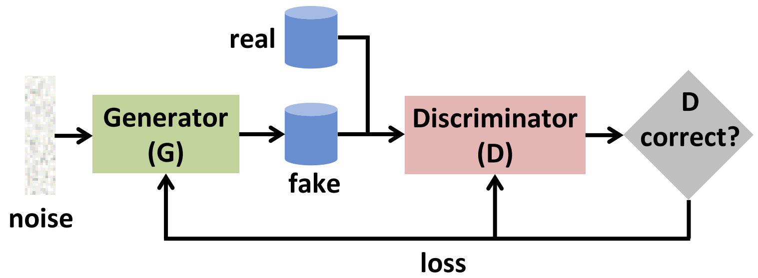
EXISTING SOLUTIONS

Timeline

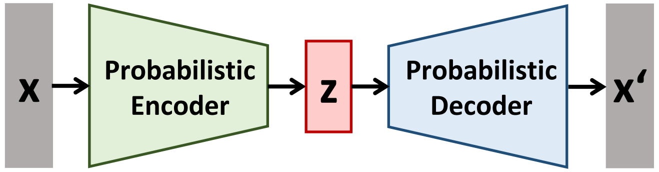


Sampling & Rule-based Approaches
Deep Generative Learning
COBOL Test Data Generator 1972
QIKSYS 1983
TDGLsystem 1985
TPC-C Benchmark 1992
Generating Consistent Test Data 1993
ToXGene 2002
SMOTE 2002
Simple & Realistic Data Generation 2006
ADASYN 2008
GeCo 2009
PrivBayes 2014
DataSynthesizer 2017
medGAN 2017
TGAN 2018
TableGAN 2018
CTGAN 2019
TVAE 2019
DoppelGANger 2019
KAMINO 2021
CTabGAN 2021
GANBLR++ 2022
GReaT 2023
REaLTabFormer 2023
TabDDPM 2023
C3TGAN 2023
GOOGLE 2023
Data Management Community
Mathematical Community
Machine Learning Community

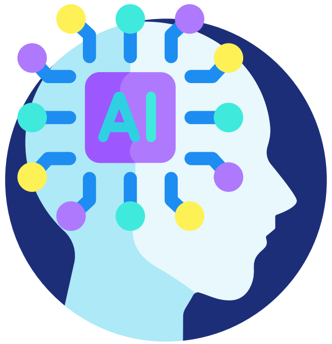
Generative Deep Learning Approaches

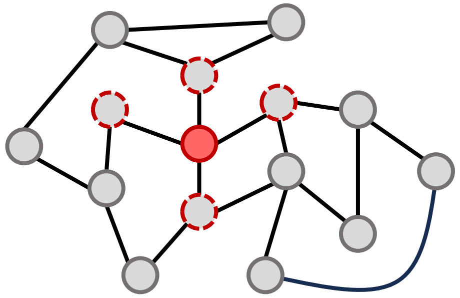


Generative Adversarial Networks

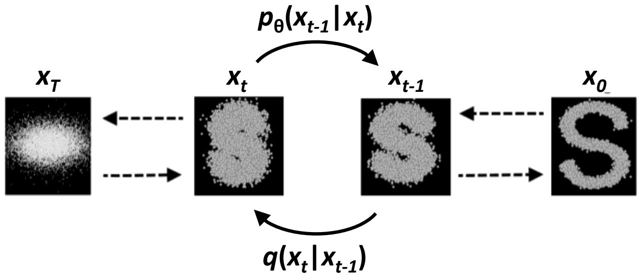


Variational Autoencoders

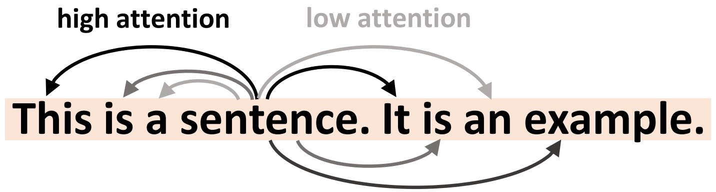




Graph Neural Networks

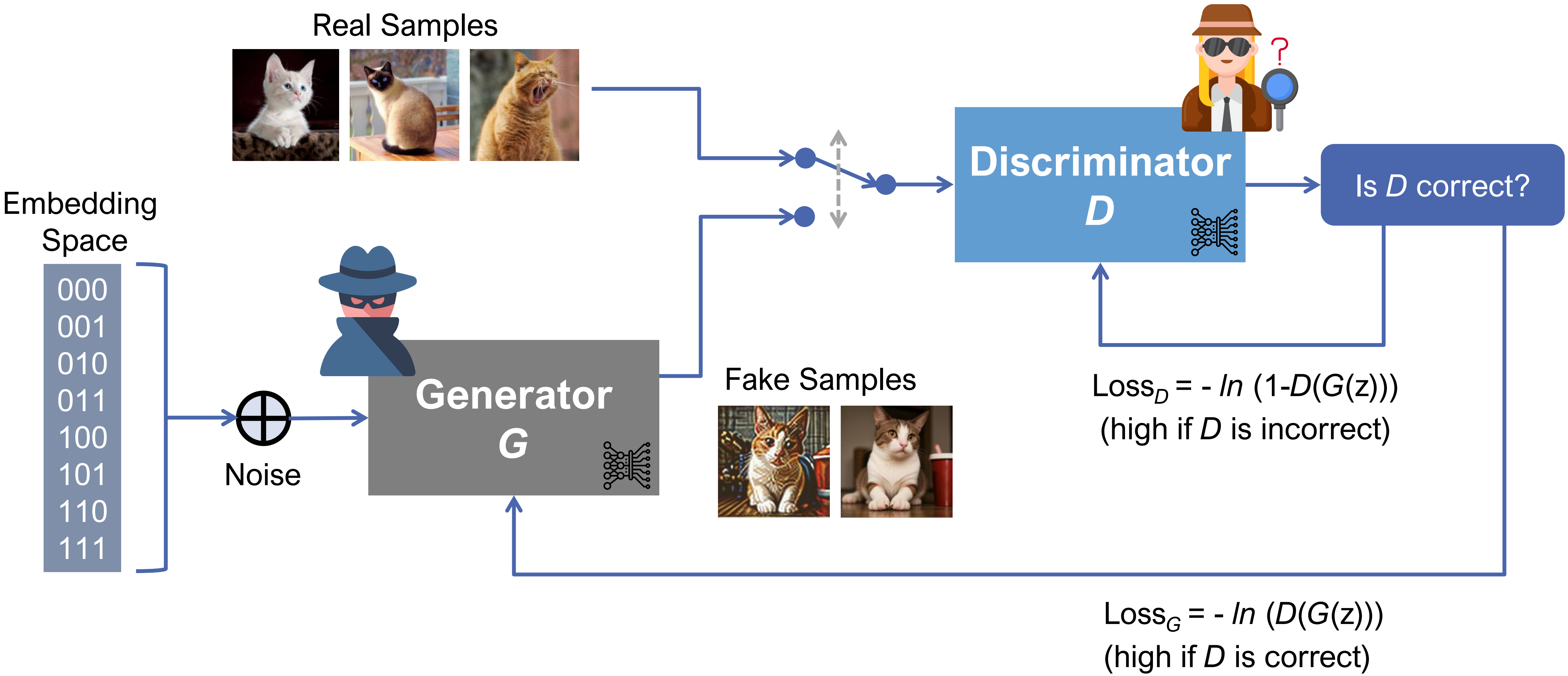


Denoising Diffusion Probabilistic Models

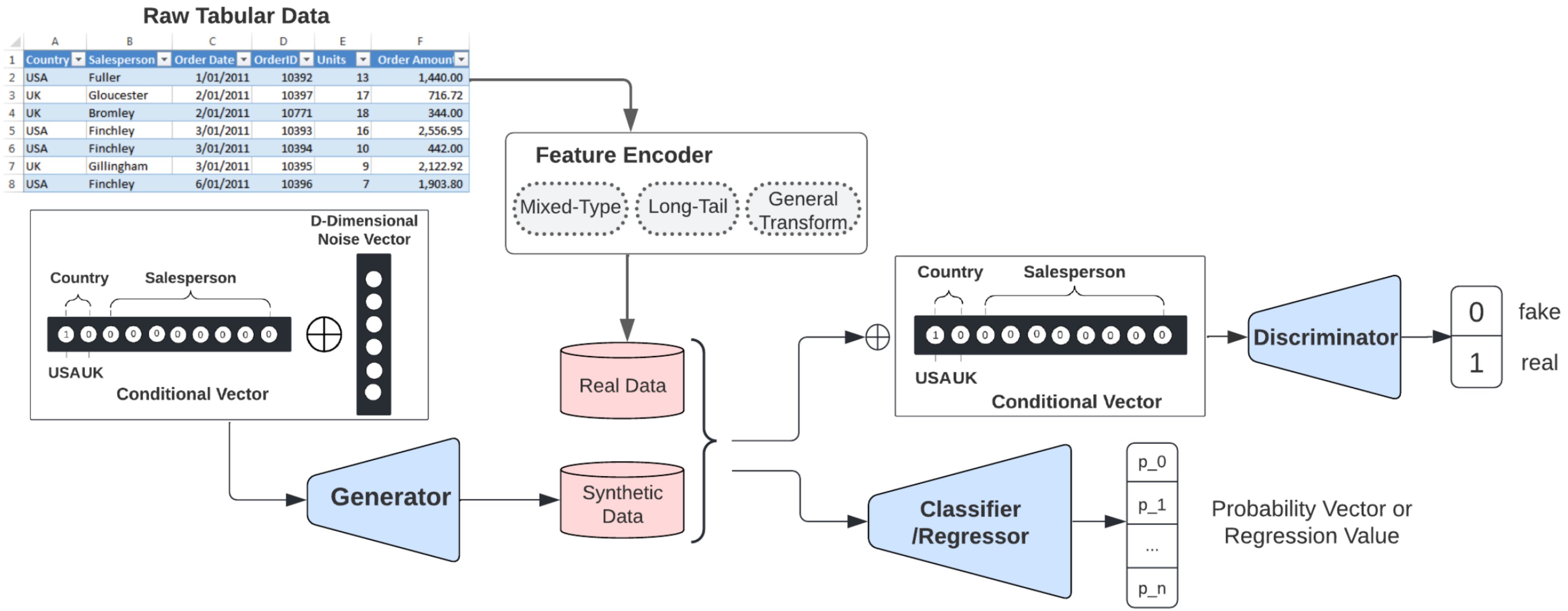


(Pretrained) Transformers / Large Language Models

Generative Adversarial Networks (GANs)



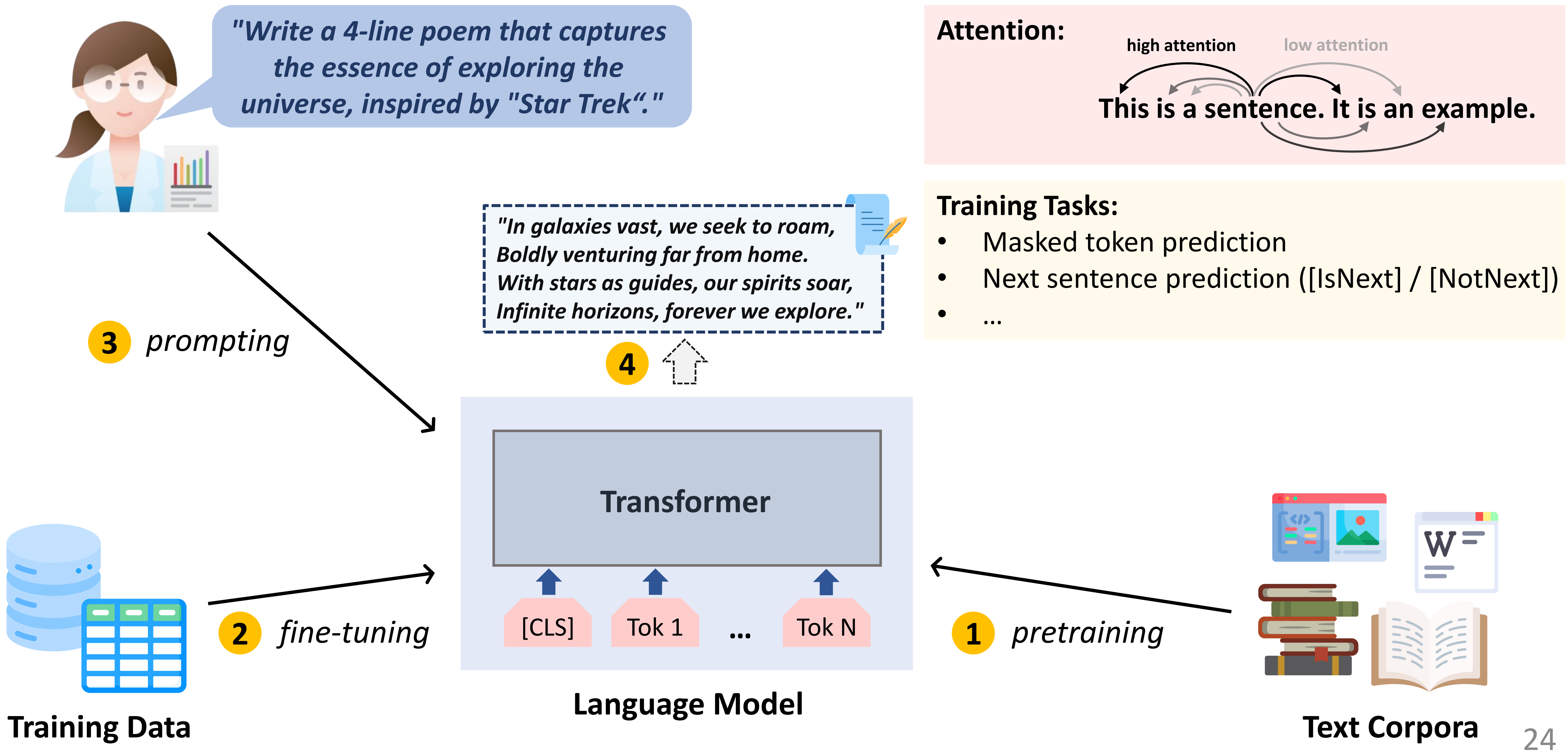
CTAB-GAN+



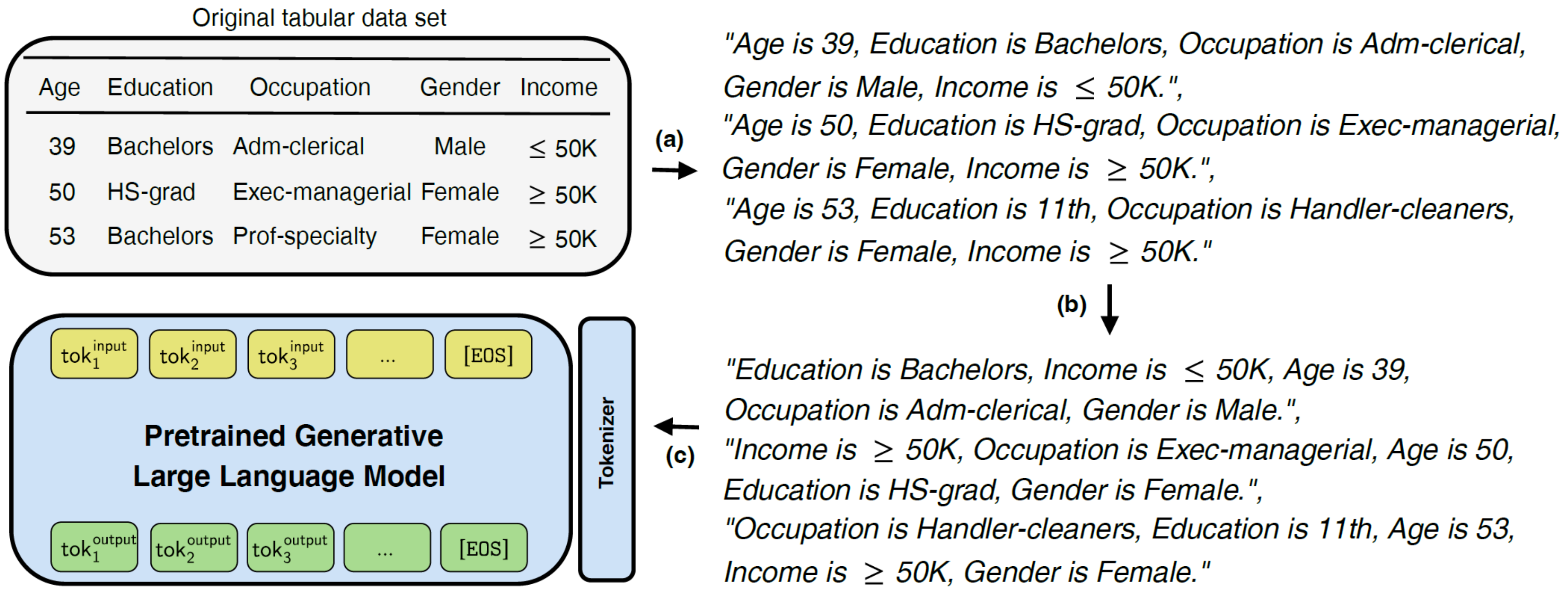
- Uses different encodings for different column types and distributions
- Uses a classifier/regressor for additional supervision
- Three types of losses: (1) information loss, (2) downstream loss & (3) generator loss

Z. Zhao, A. Kunar, et al. CTAB-GAN+: Enhancing Tabular Data Synthesis. Frontiers in Big Data. 2024.

Large Language Models



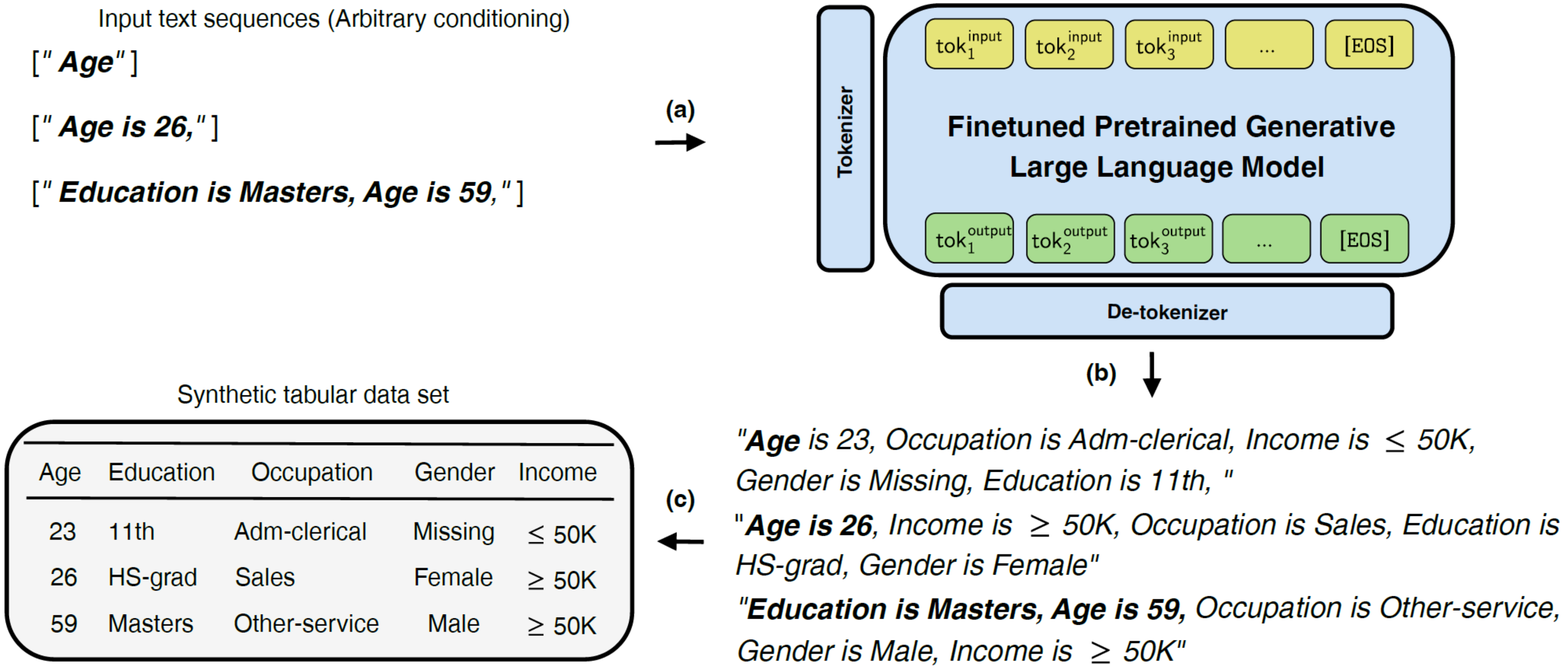
GReaT - Fine-tuning



(a) Transformation of tabular data into meaningful text
(b) Permutation of feature order
(c) Fine-tuning of the LLM

Y. Borisov, K. Seßler, et al. Language Models are Realistic Tabular Data Generators. ICLR. 2023

GReaT - Sampling

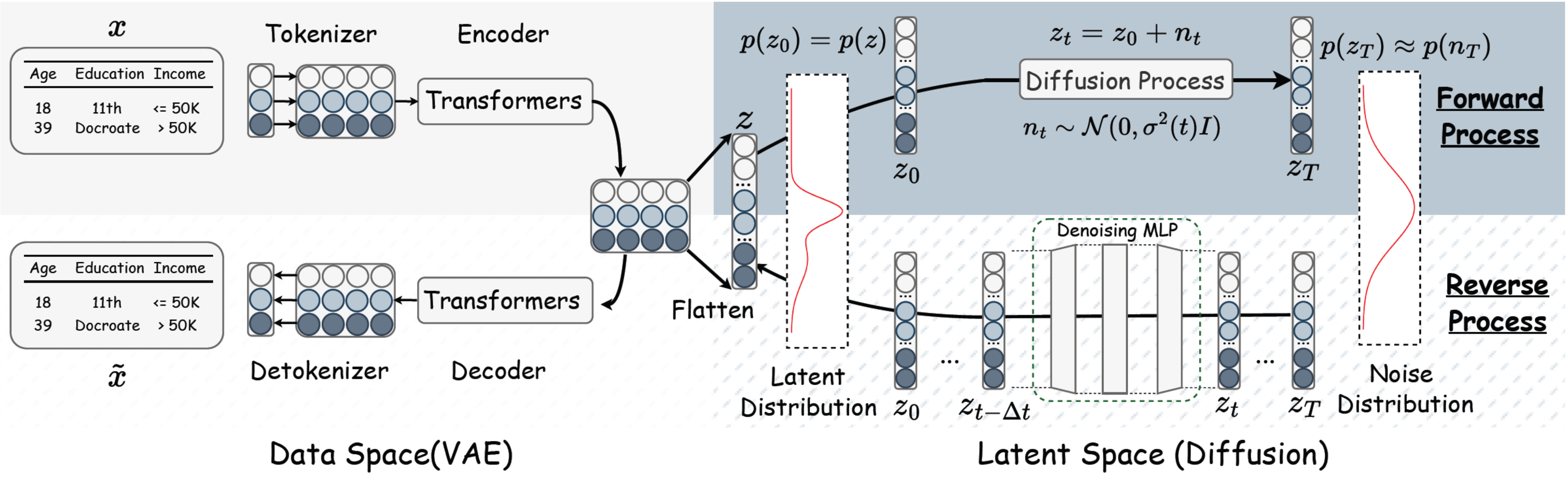


(a) Transformation of feature name or feature-value pairs into text (conditioning)
(b) LLM completes the input
(c) Transformation back to tabular representation

Out of the Box?

Y. Borisov, K. Seßler, et al. Language Models are Realistic Tabular Data Generators. ICLR. 2023

TABSYN



- Combines VAEs with Diffusion
- Uses Transformer for encoding & decoding

H. Zhang, J. Zhang, et al., Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. arXiv. 2023.

Validation of Synthetic Data

Statistical Properties
(similar value distributions, means, std. deviation, pairwise correlations)

ML Efficiency
(supervised model trained on fake data applied to real data)

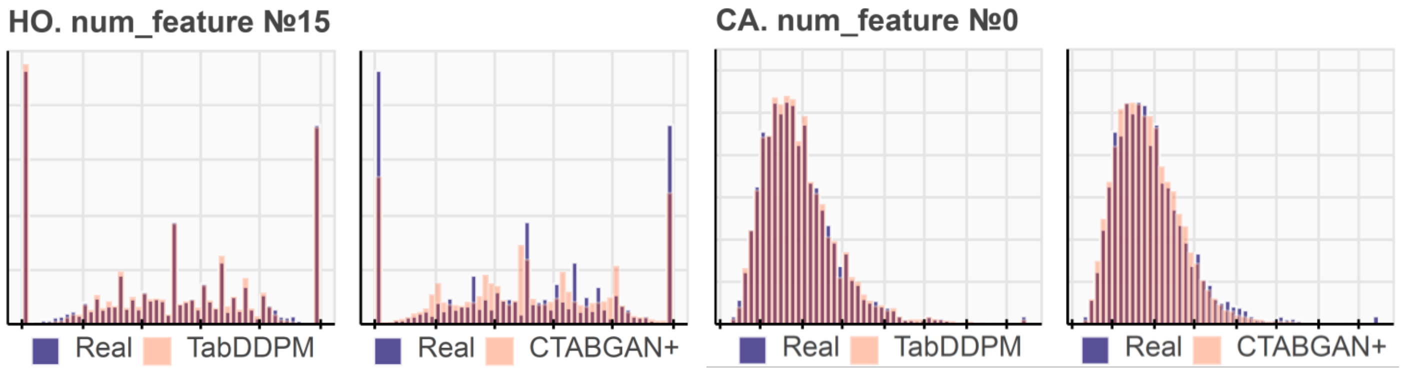
Discriminator
(classifier trained to distinguish fake from real records)

Clustering of Mixed Dataset
(entropy of resulting clusters w.r.t. real & fake records)

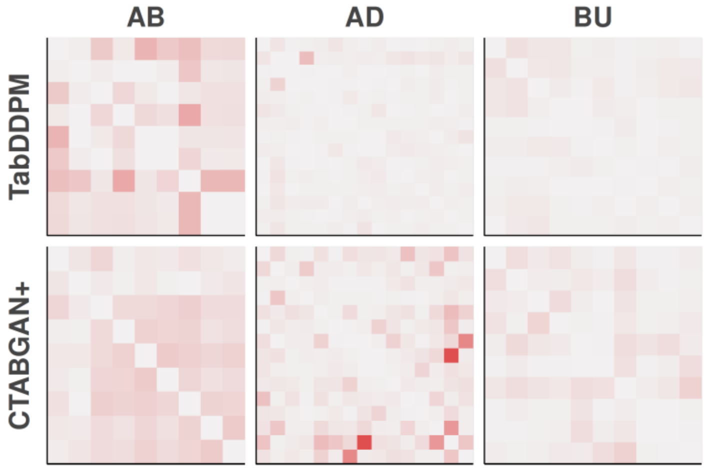
Distance to Closest Records
(similarity between fake and real records to measure potential privacy leakage)

Manual Inspection
(domain experts try to distinguish real from fake records)

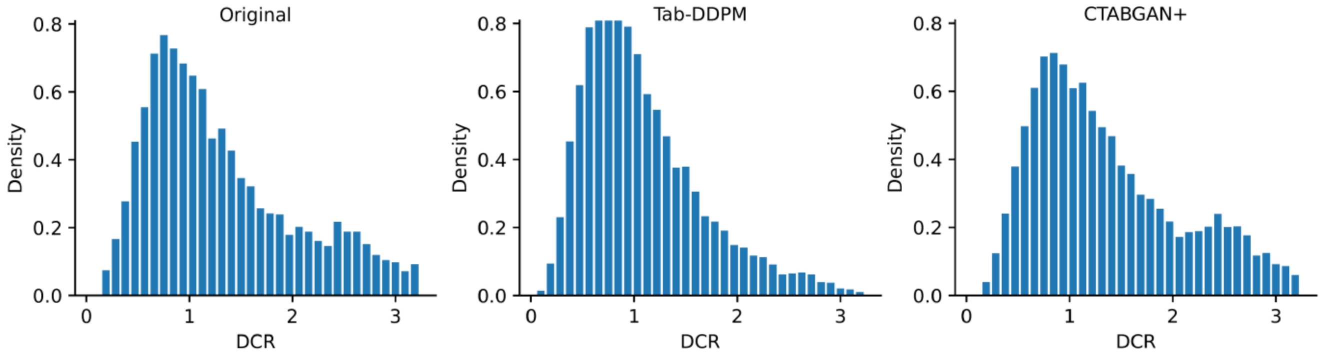
Generation Quality



Column Distributions



Correlation Coefficient



Distance to Closest Record

(Regression R2, Classification F1)

	AB (R2)	AD (F1)	BU (F1)
TabDDPM	0.392	0.758	0.851
CTABGAN+	0.316	0.730	0.837
Real	**0.423**	**0.750**	**0.845**

ML Efficiency

A. Kotelnikov, D. Baranchuk, et al., TabDDPM: Modelling Tabular Data with Diffusion Models. ICML. 2023

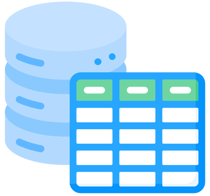
A. Solatoria, O. Dupriez: REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. arXiv. 2023.

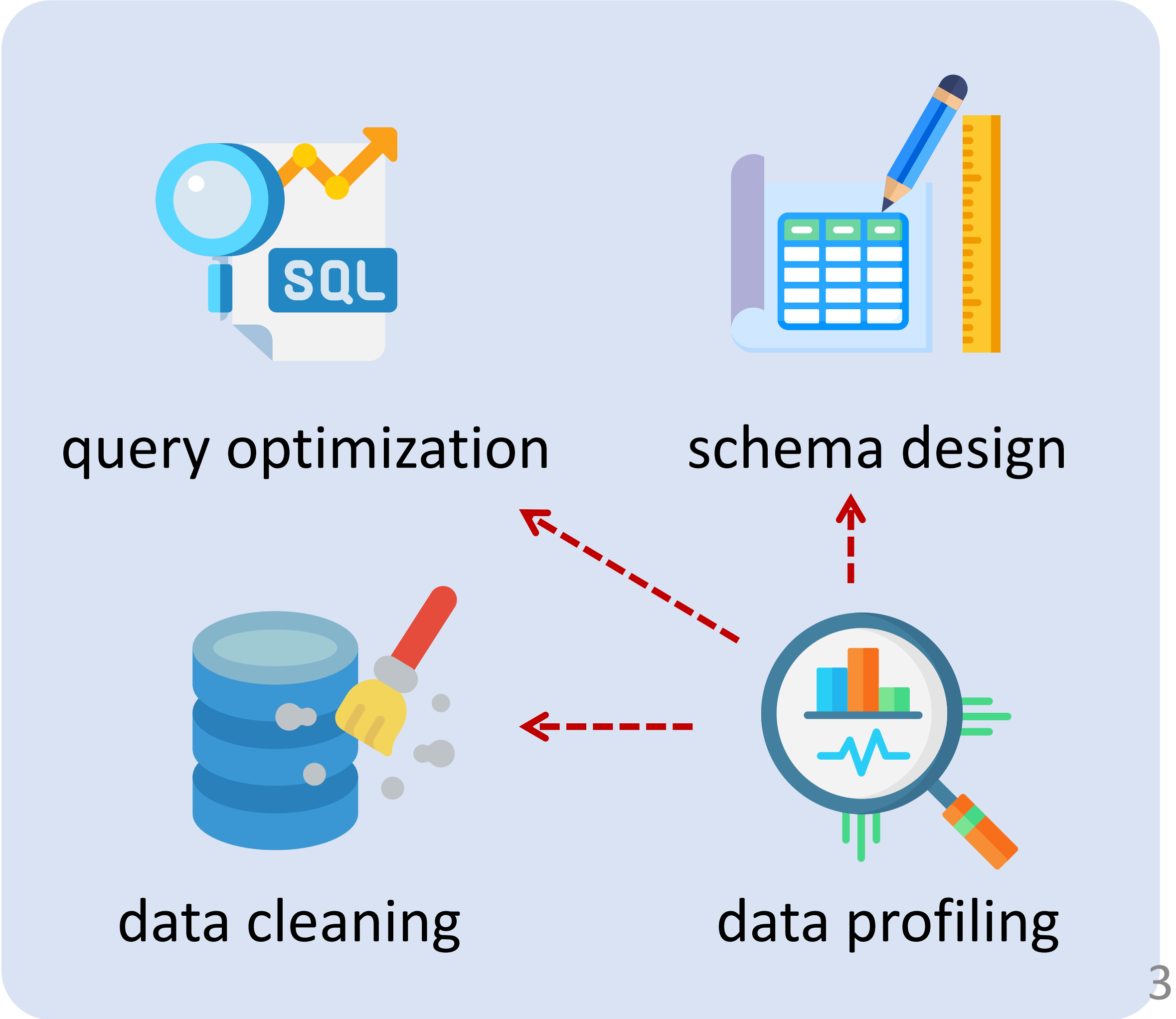
CHALLENGES FOR DATA MANAGEMENT

ML vs. DM

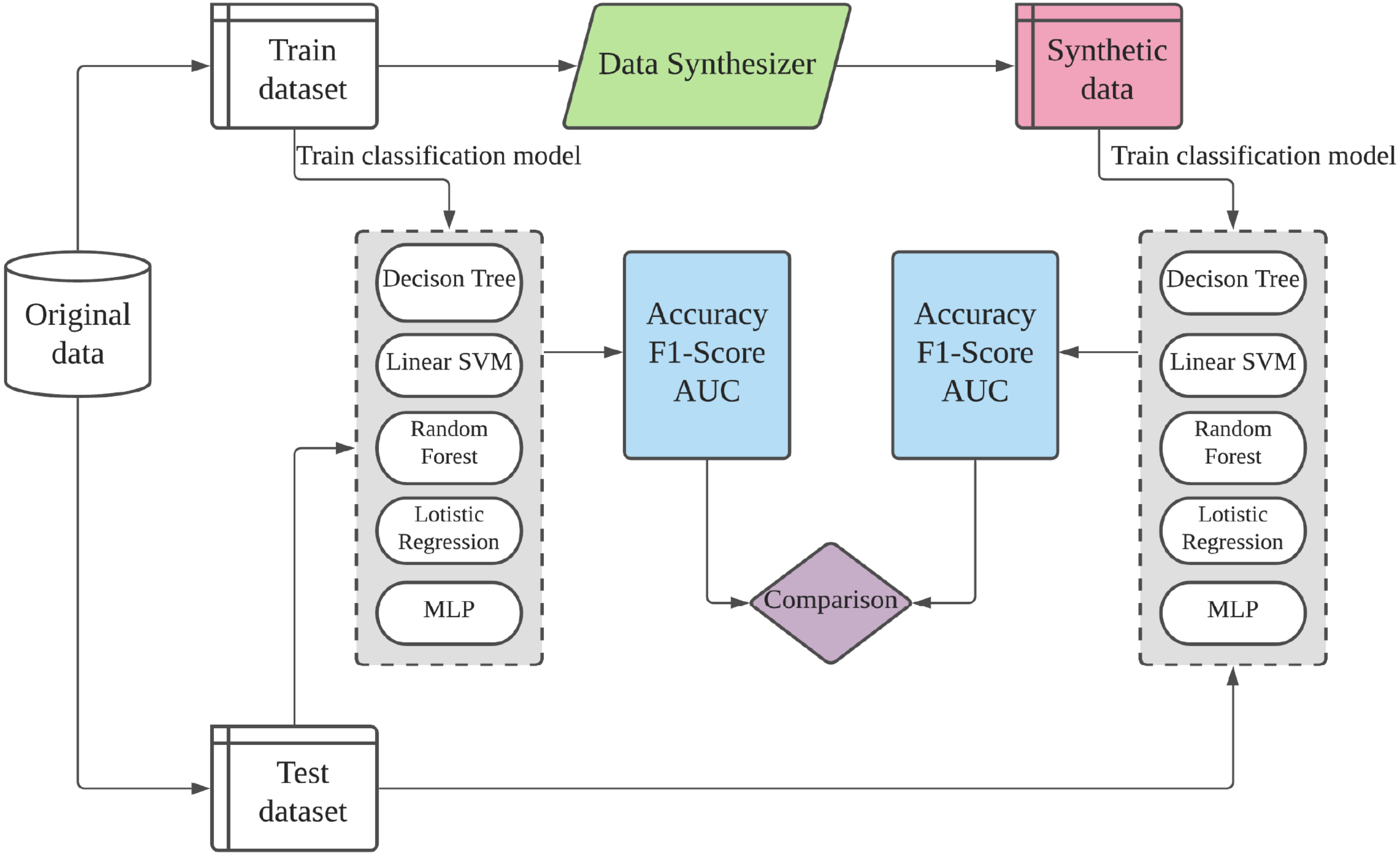
Machine Learning

Data Management





Evaluation: ML-Efficiency



Z. Zhao, A. Kunar, R. Birke, and L. Chen, CTAB-GAN: effective table data synthesizing, ACML, 2021.

Evaluation: Utility

Machine Learning

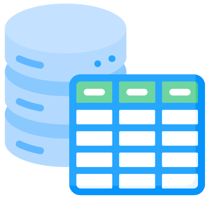
Classification

- Decision Tree Classifier
- Linear SVM
- Random Forest Classifier
- Multi-Layer Perceptron
- CatBoost

Regression

- Multinomial Logistic Regression
- Linear Regression

Data Management



Query Optimization

- Query Execution Plans
- Join Implementations

Schema Design

- Normalization
- Inheritance Modeling

Data Quality

- DQ Assessment
- Data Cleaning

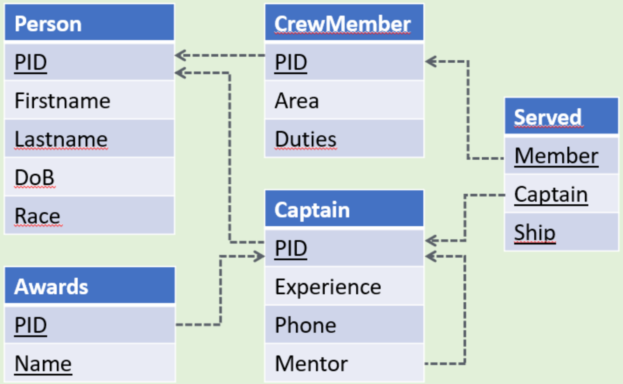
Data Profiling

- Statistics
- Integrity Constraints (FDs, UCCs, INDs)

Evaluation: Utility

M

Multi-Table Schemas



C

Integrity Constraints

- Functional Dependencies
- Unique Column Combinations
- Inclusion Dependencies
- Denial Constraints
- …

Data Management

Query Optimization M C
- Query Execution Plans
- Join Implementations

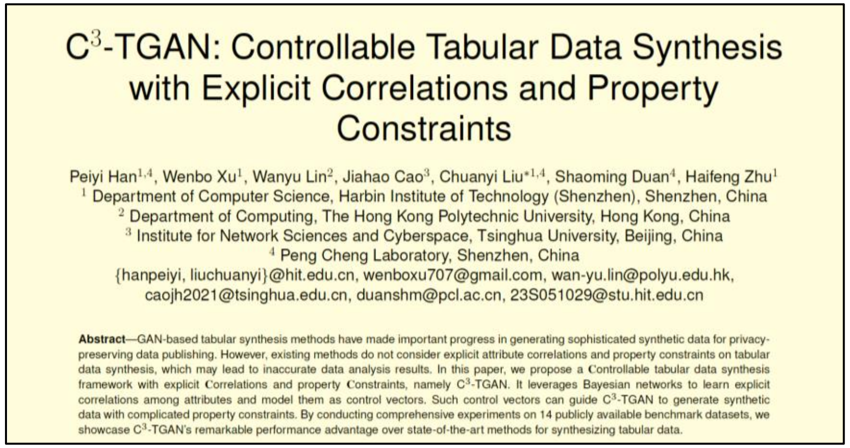
Schema Design M C
- Normalization
- Inheritance Modeling

Data Quality M C
- DQ Assessment
- Data Cleaning

Data Profiling M C
- Statistics
- Integrity Constraints (FDs, UCCs, INDs)

Integrity Constraints

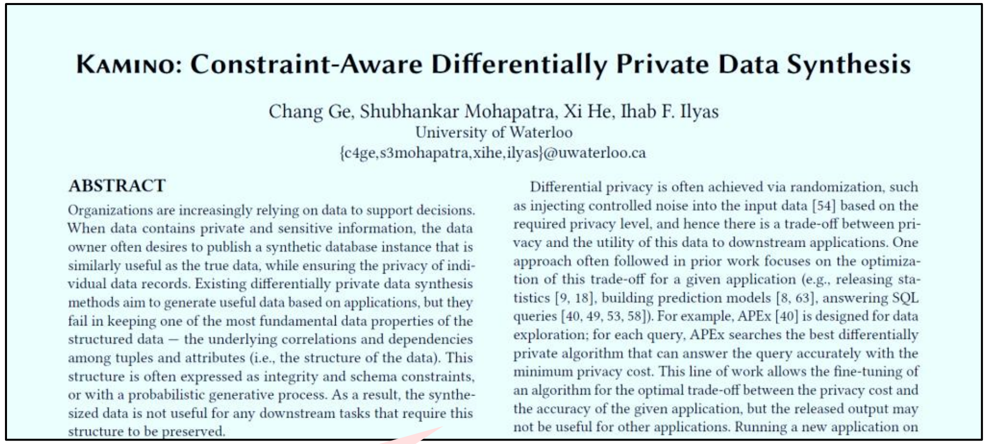
- Property constraints (e.g., Age > 0)
- Intra-record constraints
  - Comparisons (e.g., Age > Experience)
  - Arithmetic functions (e.g., Sal = nettoSal + bruttoSal)
- Inter-record constraints (single table)
  - Unique column combinations (UCCs)
  - Functional dependencies (FDs)
  - Denial constraints (DCs)
- Inter-record constraints (multiple tables)
  - Inclusion dependencies (INDs) / foreign keys (FKs)
  - Cardinality restrictions
    (e.g., a person can buy at most two books)



C^3-TGAN: Controllable Tabular Data Synthesis with Explicit Correlations and Property Constraints

Peiyi Han[1,4], Wenbo Xu[1], Wanyu Lin[2], Jiahao Cao[3], Chuanyi Liu*[1,4], Shaoming Duan[4], Haifeng Zhu[1]
[1] Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[3] Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China
[4] Peng Cheng Laboratory, Shenzhen, China
{hanpeiyi, liuchuanyi}@hit.edu.cn, wenboxu707@gmail.com, wan-yu.lin@polyu.edu.hk,
caojh2021@tsinghua.edu.cn, duanshm@pcl.ac.cn, 23S051029@stu.hit.edu.cn

Abstract—GAN-based tabular synthesis methods have made important progress in generating sophisticated synthetic data for privacy-preserving data publishing. However, existing methods do not consider explicit attribute correlations and property constraints on tabular data synthesis, which may lead to inaccurate data analysis results. In this paper, we propose a Controllable tabular data synthesis framework with explicit Correlations and property Constraints, namely C^3-TGAN. It leverages Bayesian networks to learn explicit correlations among attributes and model them as control vectors. Such control vectors can guide C^3-TGAN to generate synthetic data with complicated property constraints. By conducting comprehensive experiments on 14 publicly available benchmark datasets, we showcase C^3-TGAN's remarkable performance advantage over state-of-the-art methods for synthesizing tabular data.



KAMINO: Constraint-Aware Differentially Private Data Synthesis

Chang Ge, Shubhankar Mohapatra, Xi He, Ihab F. Ilyas
University of Waterloo
{c4ge,s3mohapatra,xihe,ilyas}@uwaterloo.ca

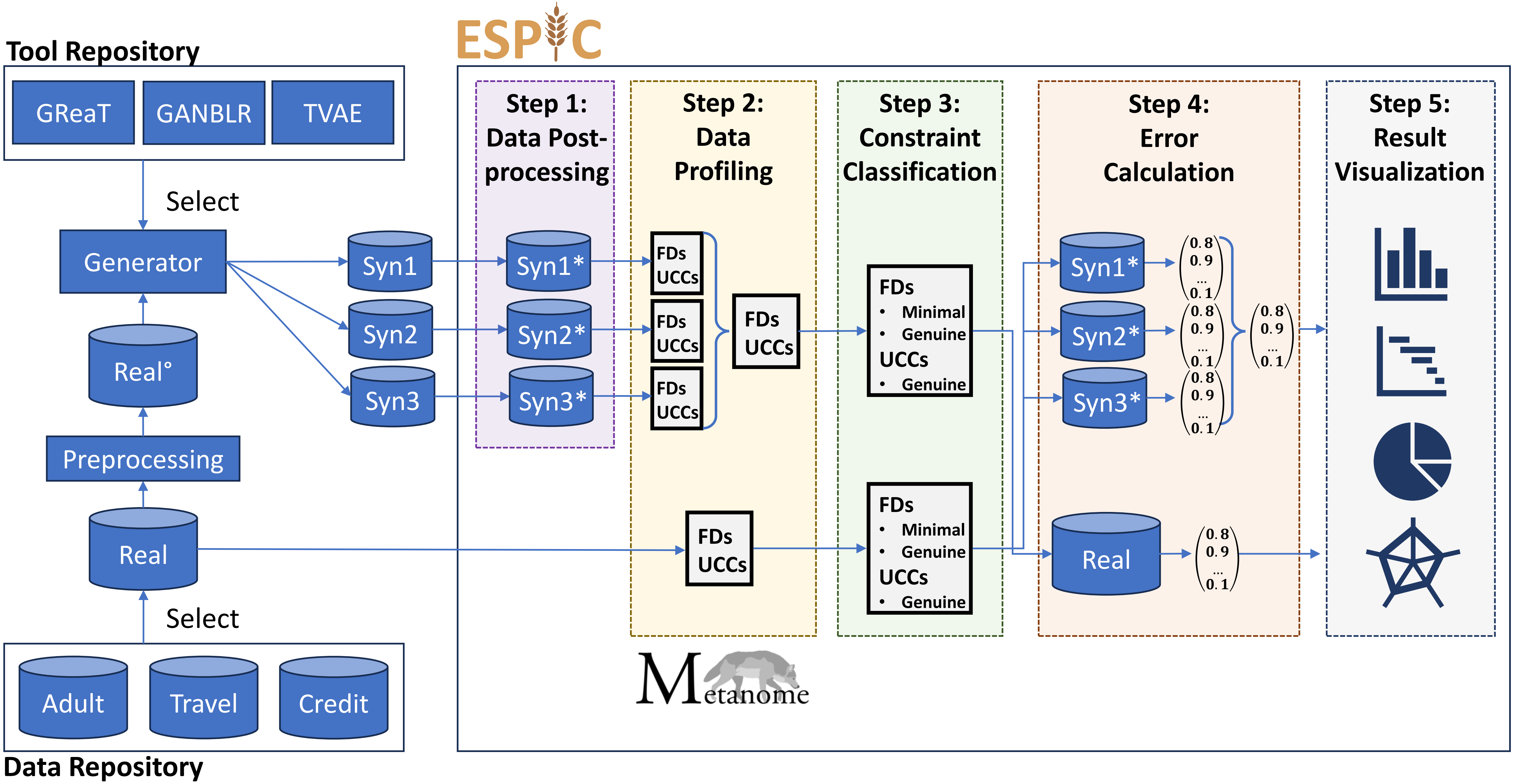
ABSTRACT

Organizations are increasingly relying on data to support decisions. When data contains private and sensitive information, the data owner often desires to publish a synthetic database instance that is similarly useful as the true data, while ensuring the privacy of individual data records. Existing differentially private data synthesis methods aim to generate useful data based on applications, but they fail in keeping one of the most fundamental data properties of the structured data – the underlying correlations and dependencies among tuples and attributes (i.e., the structure of the data). This structure is often expressed as integrity and schema constraints, or with a probabilistic generative process. As a result, the synthesized data is not useful for any downstream tasks that require this structure to be preserved.

Differential privacy is often achieved via randomization, such as injecting controlled noise into the input data [54] based on the required privacy level, and hence there is a trade-off between privacy and the utility of this data to downstream applications. One approach often followed in prior work focuses on the optimization of this trade-off for a given application (e.g., releasing statistics [9, 18], building prediction models [8, 63], answering SQL queries [40, 49, 53, 58]). For example, APEx [40] is designed for data exploration; for each query, APEx searches the best differentially private algorithm that can answer the query accurately with the minimum privacy cost. This line of work allows the fine-tuning of an algorithm for the optimal trade-off between the privacy cost and the accuracy of the given application, but the released output may not be useful for other applications. Running a new application on

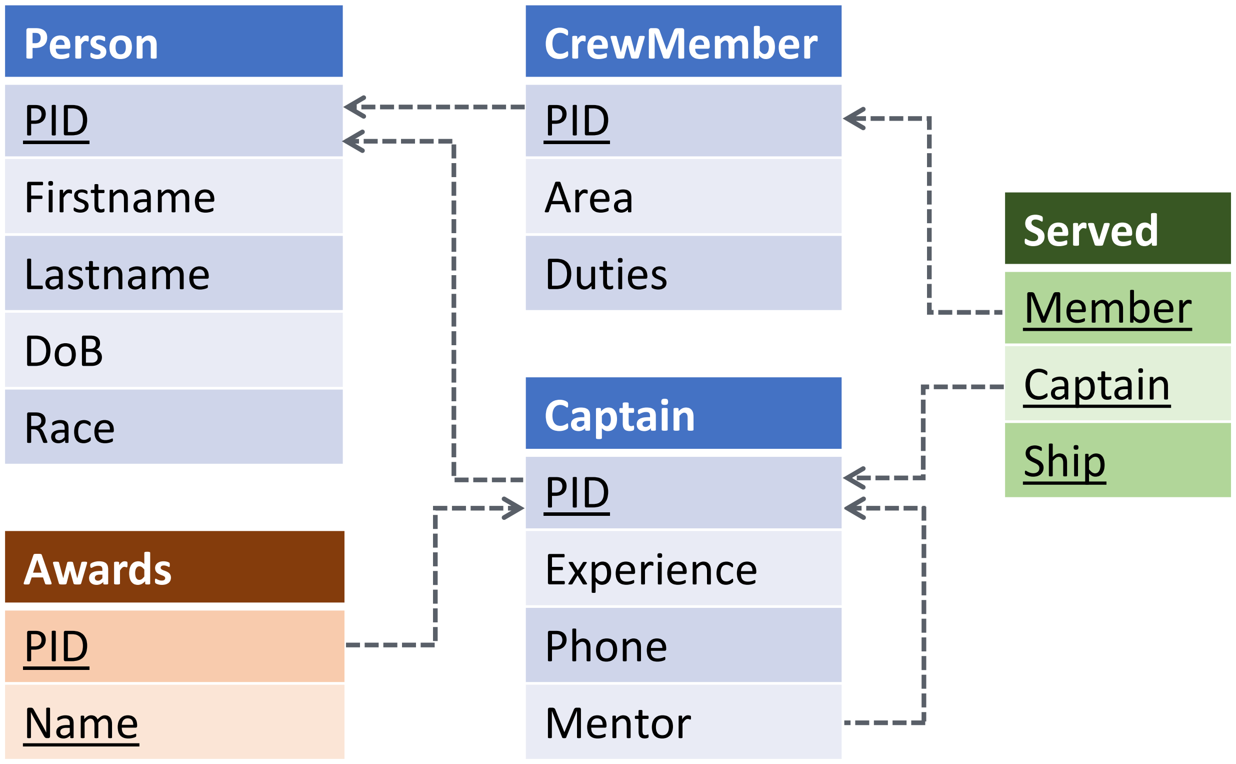
Specialized solution (not adoptable)

Evaluating Synthesizers for Preserving ICs



Multi-Table Schemas

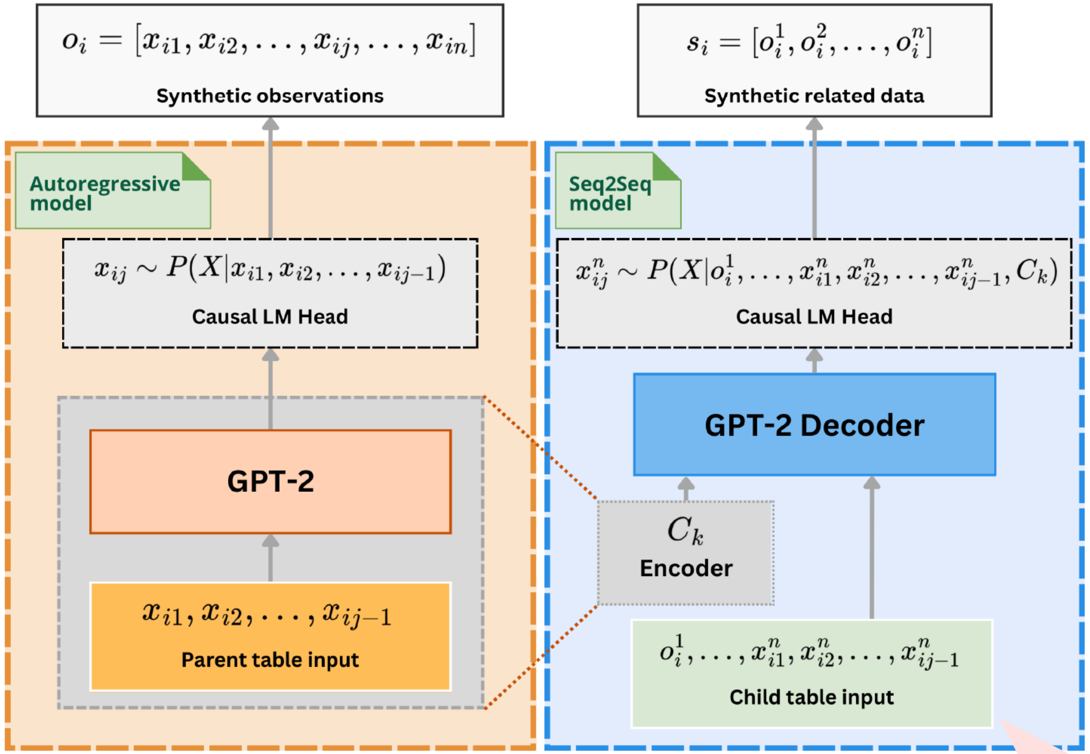
- Consists of tables modeling
  - Entity types (e.g., CrewMember)
  - n:m Relationship types (e.g., Served)
  - Multivalued attributes (e.g., Awards)
- Inheritance Hierarchies
  - ✗ Horizonal partitioning
  - ✓ Vertical partitioning
  - ✗ Full redundancy



Generation with Multi-Table Schemas

Two tables

- 2 Entity types (e.g., Spaceship & CrewMember)
- 1:n Relationship type (e.g., hasCaptain)



REaLTabFormer

What if we have more than two tables?

Potential Solutions:

1. **Iterative Synthesis:** Subsequent generation of tables conditioned by anchestor tables

   Problems: What if we have cycles? What if a table has multiple parent tables?

2. **Denormalization:** Requires preservation of Functional Dependencies (e.g., KAMINO)

   Problem: Table can become very large

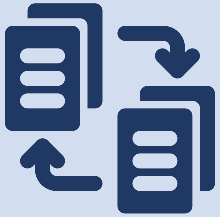
3. **Holistic Synthesis:** Embedding for whole database including multi-table structure

   Problem: Very complex embedding

A. Solatoria, O. Dupriez: REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers. arXiv. 2023.

DATA SYNTHESIS
FOR RESEARCH DATA

Rationales for Sharing Research Data

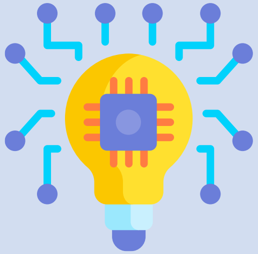


Reproduce or **verify** research.

Make results of **publicly funded** research **available**.

Enable others to ask **new questions** of extant data.

(reuse & secondary use)



Advance the state of research and **innovations** (e.g., training & benchmark datasets).

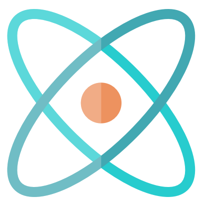
C. Borgman. *The conundrum of sharing research data,* Journal of the American Society for Information Science and Technology, 63(6), 2011.

What Characterizes Research Data?

Highly diverse data landscape!

Case Studies

- few data (expensive to acquire)
- often sensitive information
- high risk of bias



Particle Accelerator

- hugh volumes of data
- mostly less sensitive

Reuse: Different hypothesis than in original research, but data collection depends on studied hypothesis.

Bias often only recognizable if entire data collection pipeline is known (e.g., selection criteria of study participants).

What Characterizes Research Data?

id	RAdeg	DEdeg	Type	max	n_max	f_min	min	n_min	Epoch	Period	V
0	152.7375	-50.515	MIRA	8.65	V	(	3.64	V	2452630.1	241.0	0
1	271.0	-32.38833	MISC	12.33	V	(	1.12	V	2452966.7	210.881485	0
2	116.82833	-19.40111	LB	8.9	V		9.9	V			0
3	137.40958	44.77611	SN:	13.5	V		20.0	V			1
4	89.6125	-17.66333	EC	10.15	V	(	0.39	V	2451869.31	0.41459	0
5	70.56667	-25.825	EC	12.65	V	(	0.59	V	2451868.886	0.254893	0
6	59.36958	-1.15944	UV:	8.06	V	(	0.1	V			1
7	188.37917	-54.66	MIRA	8.63	V	(	4.36	V	2452263.2	221.463745	0
8	285.95542	-21.02778	RRAB	12.9	pg		13.9	pg	2426507.433	0.4789412	0
9	238.80833	-13.31861	RRAB	13.9	pg		15.6	pg	2435663.82	0.45295	0
10	338.49292	24.565	M	8.0	V		13.6	V	2445177.0	424.8	0

The International Variable Star Index (https://www.aavso.org/vsx/)

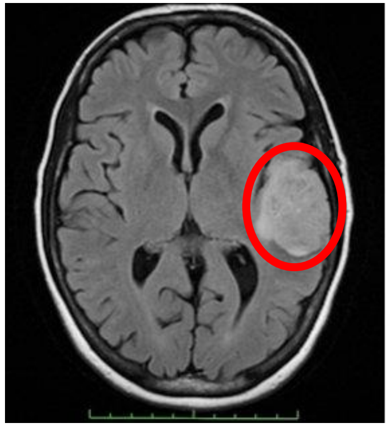
Research data are often:

- **numerical** (problem for LLMs?)
- **unstructured** (e.g., human text annotations) or **semi-structured** (e.g., JSON/XML-files)
- **multimedia** data (e.g., MRI images)

What Characterizes Research Data?

Depends on position in research pipeline!

MRI Image



extract

Requires image data

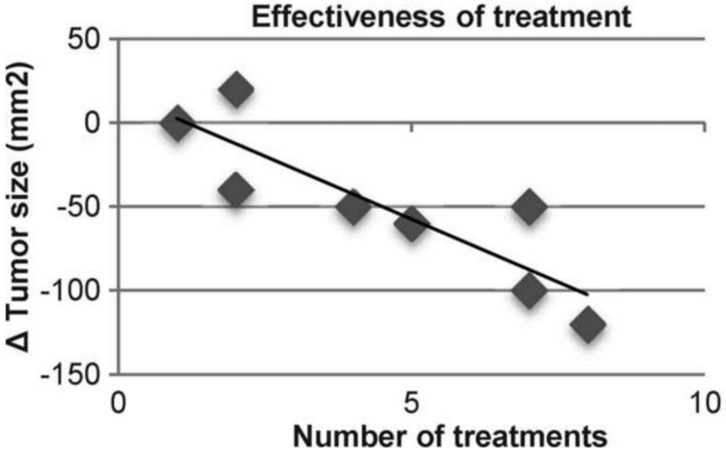
Tumor and Treatment Data

Patient ID	Tumor area pretreat (mm²)	Tumor area posttreat (mm²)	Number therapy sessions
1001	454	317	4
1002	234	82	7

analyze

Requires tabular data

Analysis Results

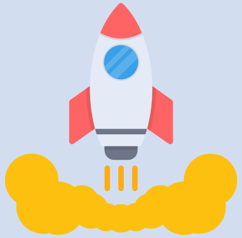


A. Surkis, K. Read. *Research data management*, Journal of the Medical Library Association 103(3), 2015.

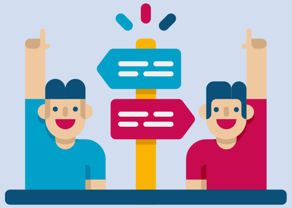
Summary

Many reasons to generate synthetic data

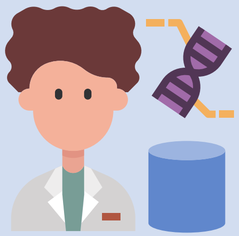
- privacy regulations
- missing values
- too few data
- imbalanced data



- Generative deep learning has great potential
- Approaches from image and language processing
- Is in a rapid process of continuous improvement

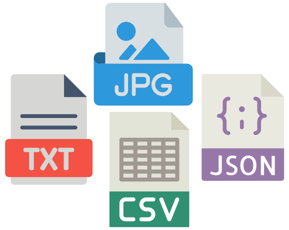


- Most approaches from ML community
- Different goals than DB community
  - no constraints
  - only single tables

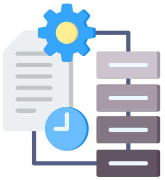


- Research data landscape is diverse
- Need for synthetic data of different types

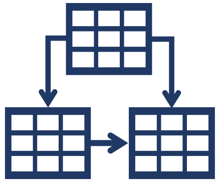
How is your Research Data?



What types of data do you work with? Tabular data?



What do you do with your research data? Just Machine Learning?



Do you have complex schemas?

Are integrity constraints important for your research?

Do you need synthetic data?