

Opportunities and Challenges of Synthetic Data for Enterprise Workflows

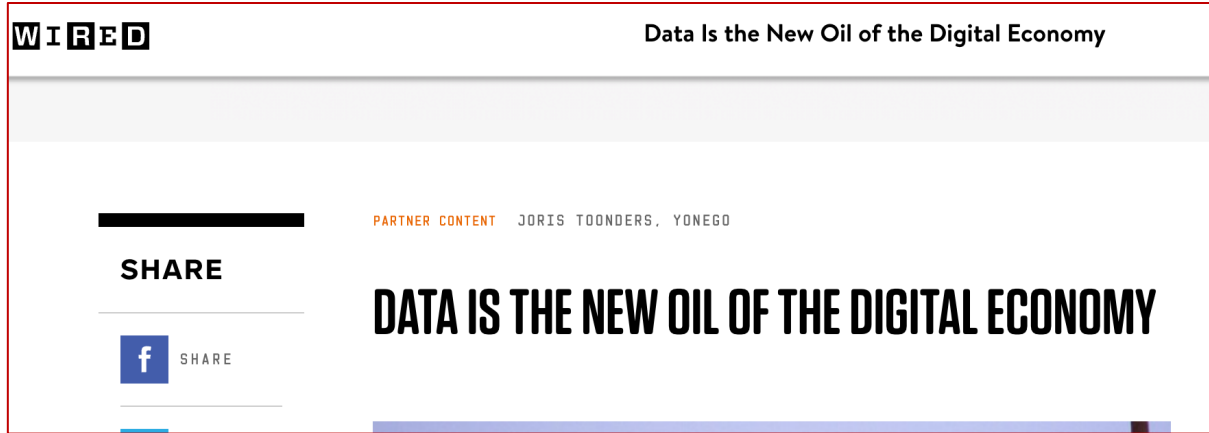
Vyas Sekar

Credits: Giulia Fanti, Zinan Lin, Yucheng Yin

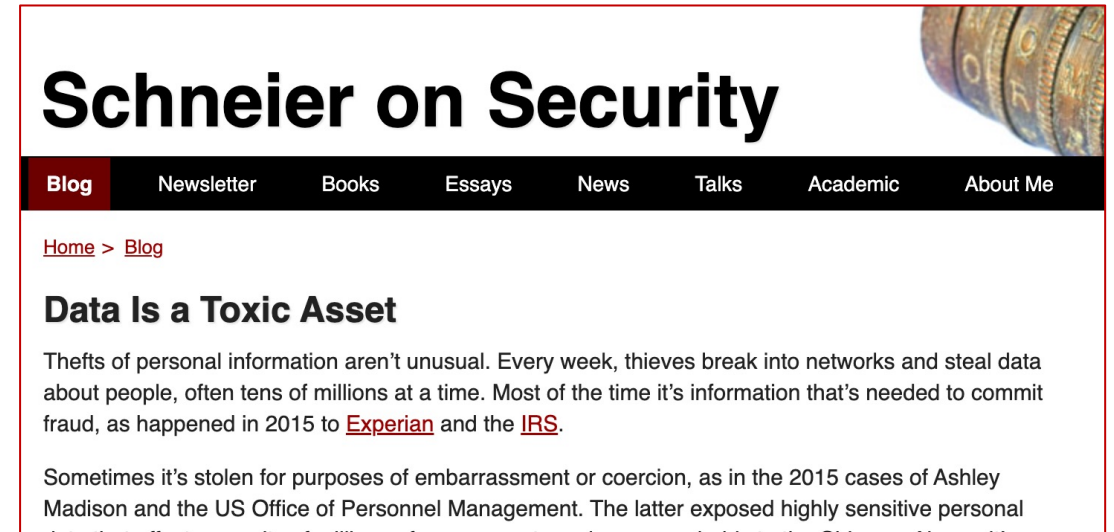
Carnegie Mellon University



Two Conflicting Views on Data!



VS.



Data Governance Requirements

REGULATIONS

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Version 1.0

NIST PRIVACY FRAMEWORK: A TOOL FOR
IMPROVING PRIVACY THROUGH ENTERPRISE
RISK MANAGEMENT, VERSION 1.0

January 16, 2020

CALIFORNIA CONSUMER PRIVACY ACT OF 2018

Contents

1798.100. General Duties of Businesses that Collect Personal Information.....	3
1798.105. Consumers' Right to Delete Personal Information	4
1798.106. Consumers' Right to Correct Inaccurate Personal Information.....	6
1798.110. Consumers' Right to Know What Personal Information is Being Collected. Right to Access Personal Information.....	7
1798.115. Consumers' Right to Know What Personal Information is Sold or Shared and to	



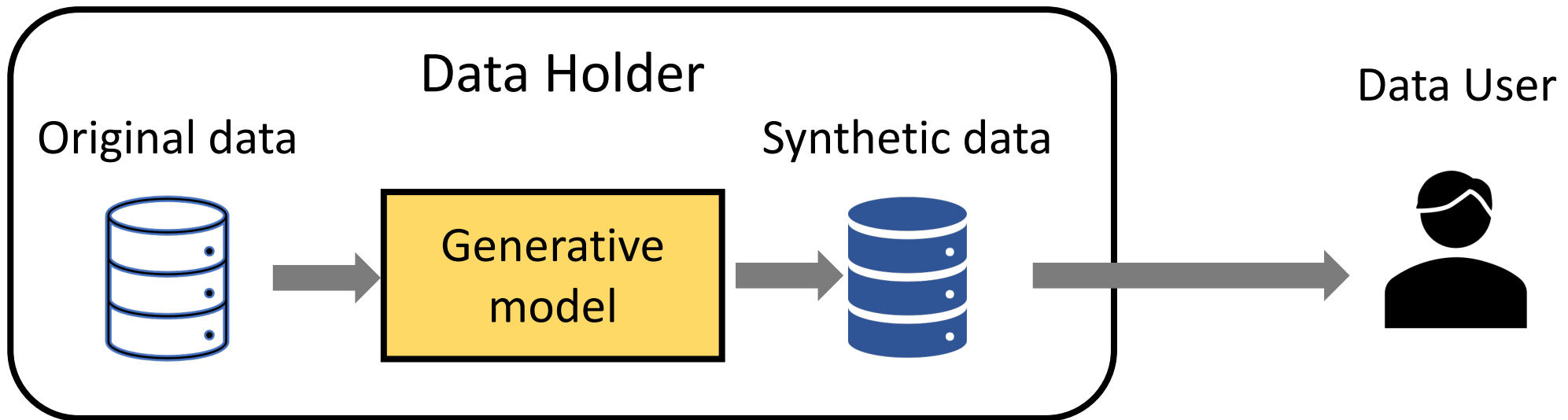
OCR PRIVACY BRIEF

SUMMARY OF THE HIPAA PRIVACY RULE

Synthetic data: A promising alternative

“Synthetic data: production data...
that are not obtained by direct measurement”

- Merriam-Webster



There is a lot of hype around synthetic data!

TECHNOLOGY > AI AND AUTOMATION | August 2, 2023

The majority of AI training data will be synthetic by next year, says Gartner

FORBES > INNOVATION > AI

Synthetic Data Is About To Transform Artificial Intelligence

Agencies Explore Policy Barriers, Synthetic Data to Maintain Health Care Privacy

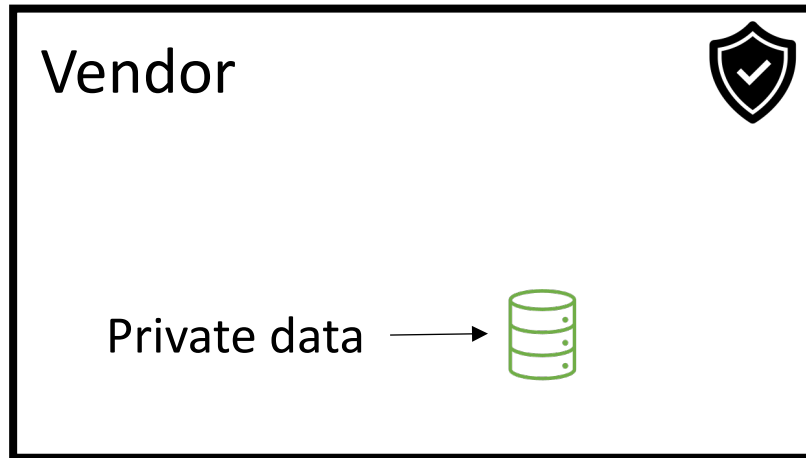
Policy changes and synthetic data could play a key role in eliminating risks while improving data access and security.

Why should you care: Goals of this talk

- Opportunities to use Synthetic Data to Unlock Data-driven Workflows
- Background: AI/ML Science Behind Synthetic Data
- What are the practical challenges with Synthetic Data?
 - Fidelity
 - Privacy, Compliance, and Governance implications

Opportunities for Synthetic Data

Use Case 1: Data sharing

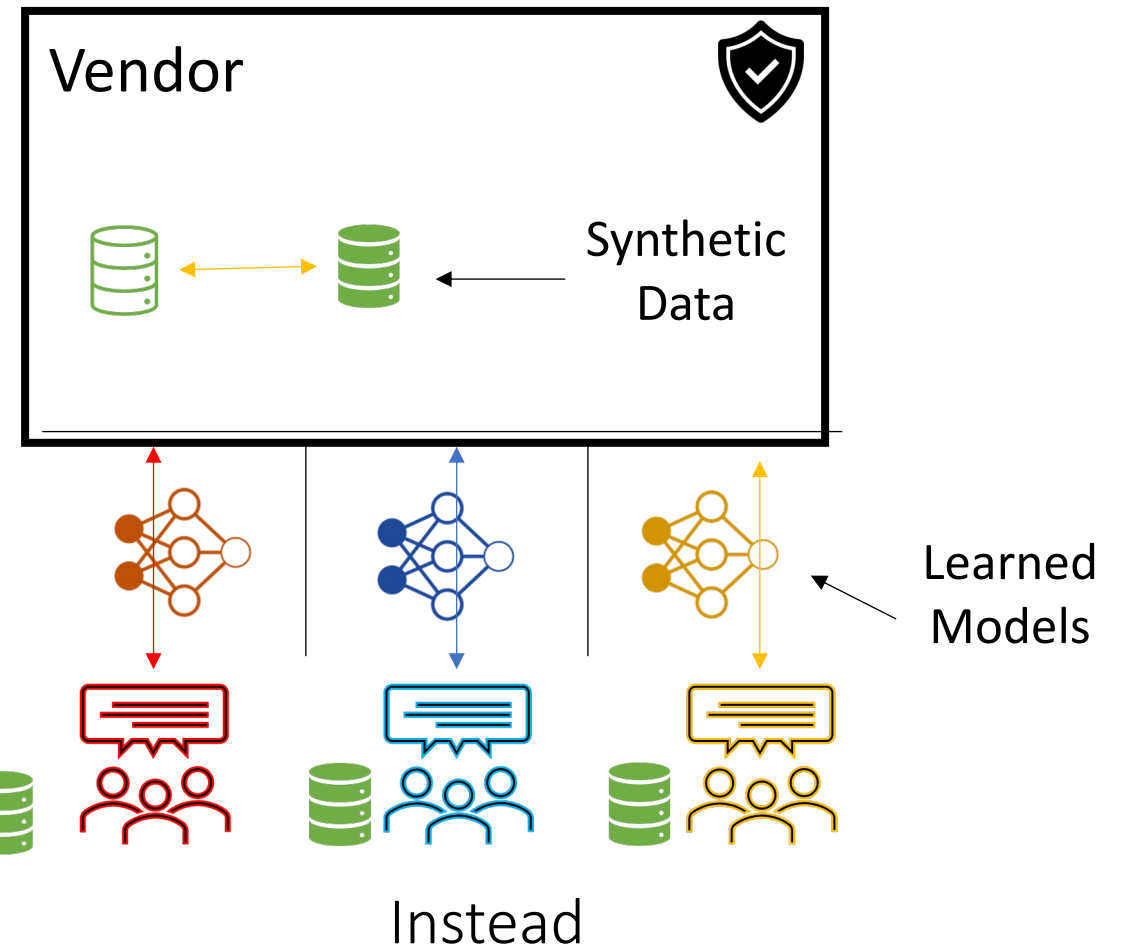


3rd party
researchers



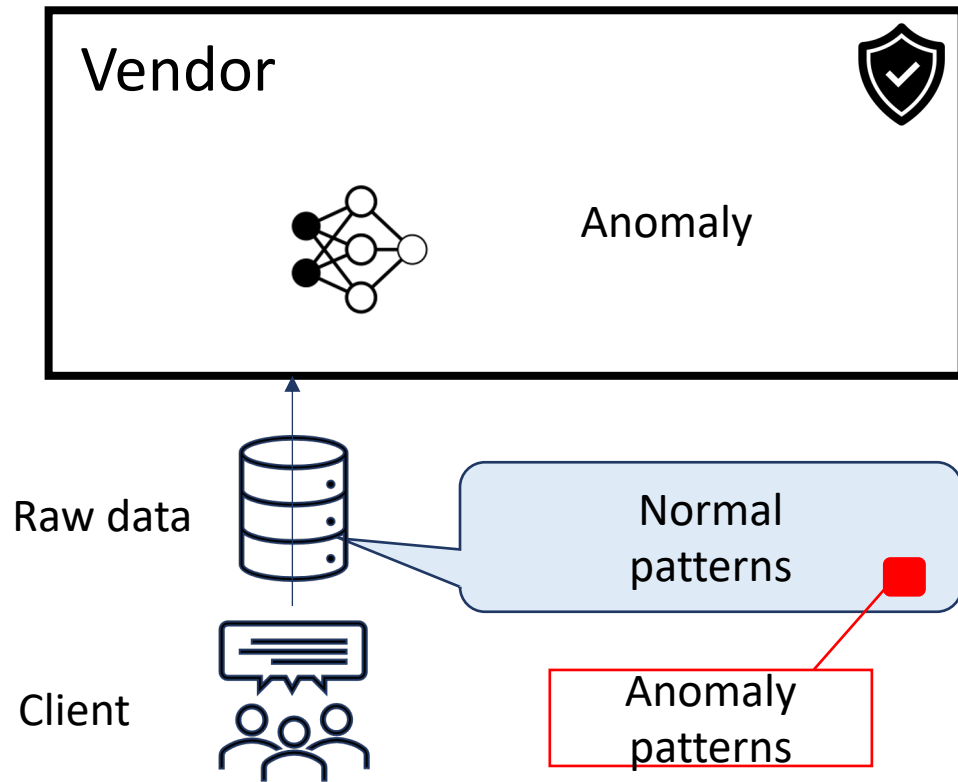
Today

Researchers don't have good enough data to
develop good algorithms



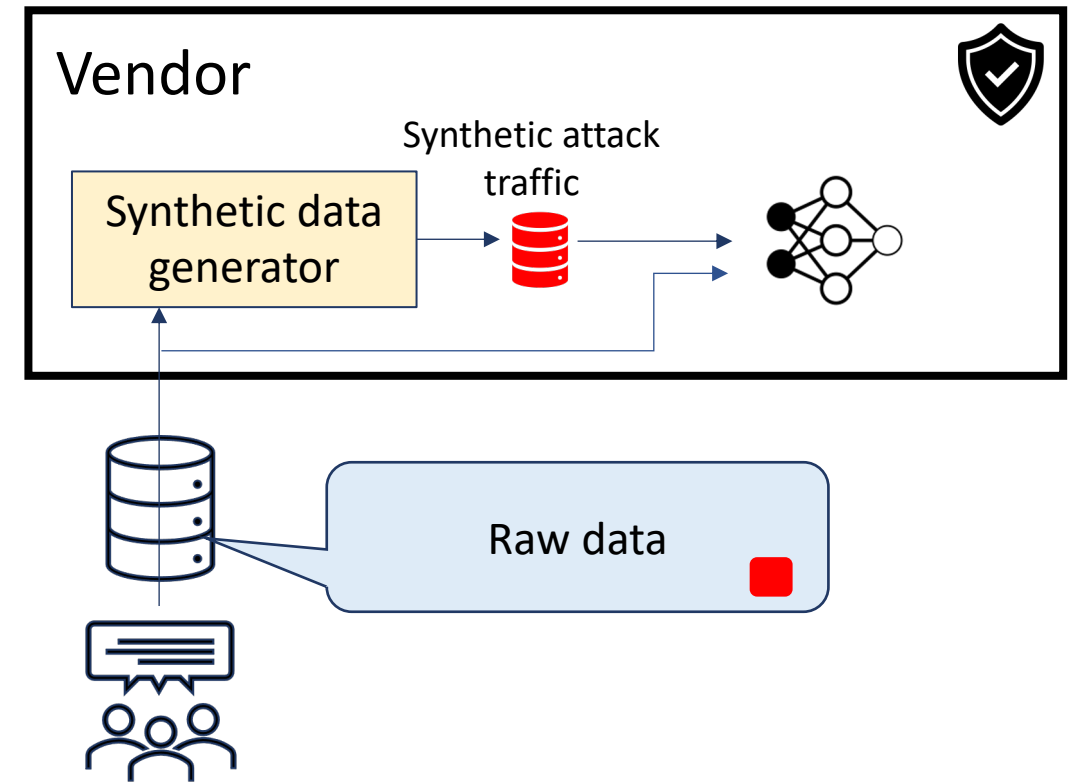
Researchers use **synthetic data** to
learn algorithms

Use case 2: Modeling rare data (e.g., robust training, unbiasing)



Today

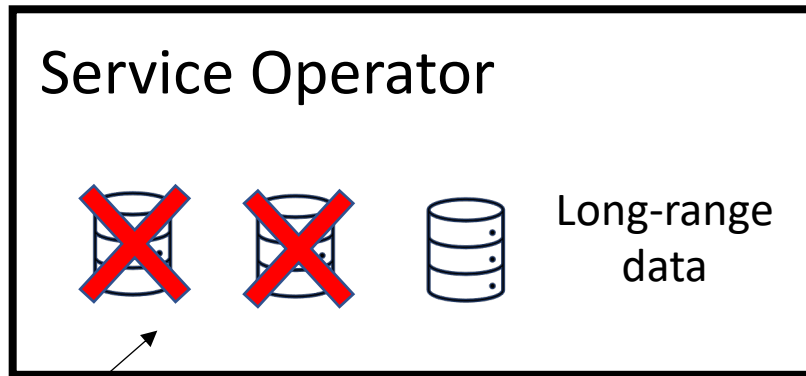
Vendors train data from highly imbalanced training datasets



Instead

Vendors use **synthetic data** to balance training data

Use case 3: Tackle Data Retention Challenge

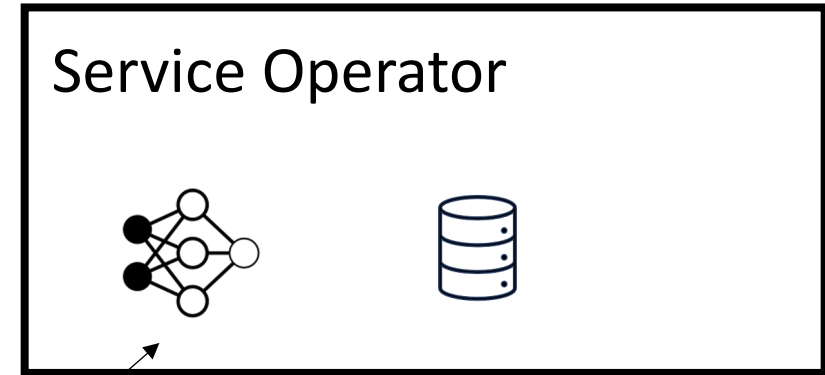


Retention
limits

- Expensive to store
- Data use restrictions

Today

Operators can't do retroactive analysis or
incident response due to compliance
Or retention restrictions



Synthetic model
parameters

Instead

Vendors store **generative model parameters**
to model older data

Use case 4: Unlock In-Enterprise Data Silos

Can't build robust models

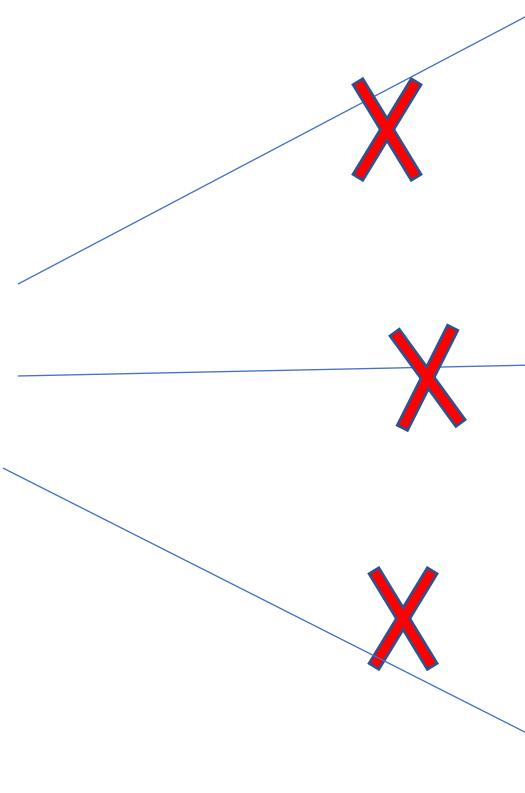
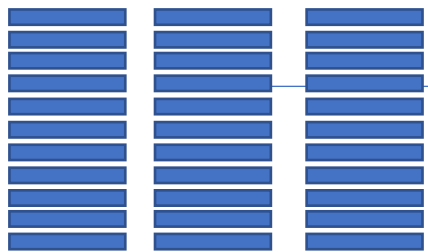
Data science Team

*Spends a lot of time trying
to reproduce incidents*

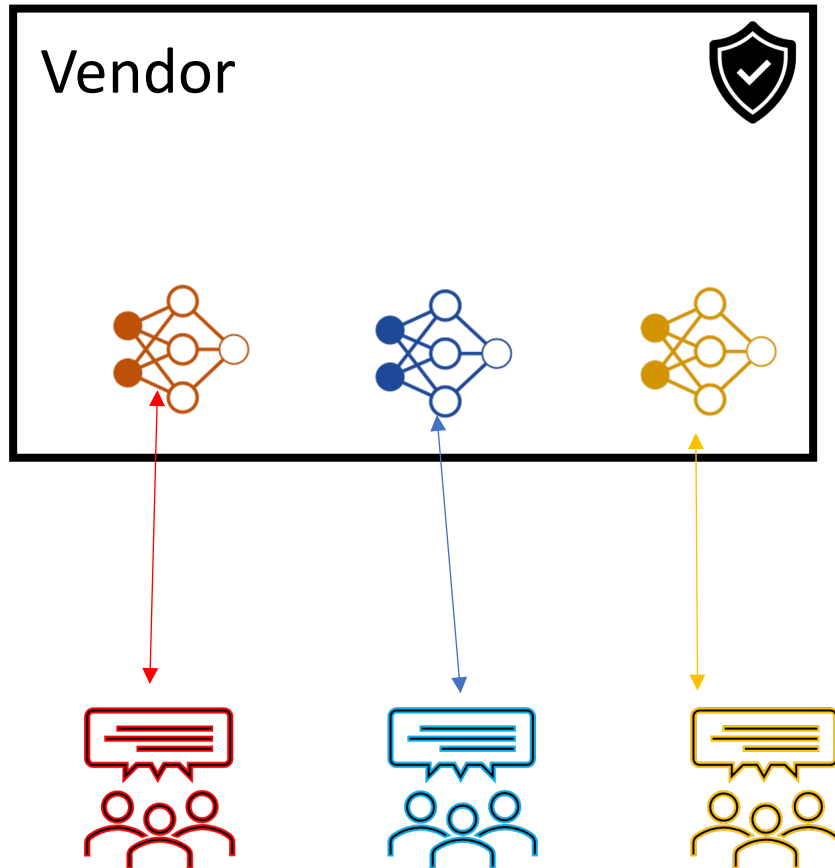
Test & Response
Team

Cant access data to demo
for new customers

Sales Team

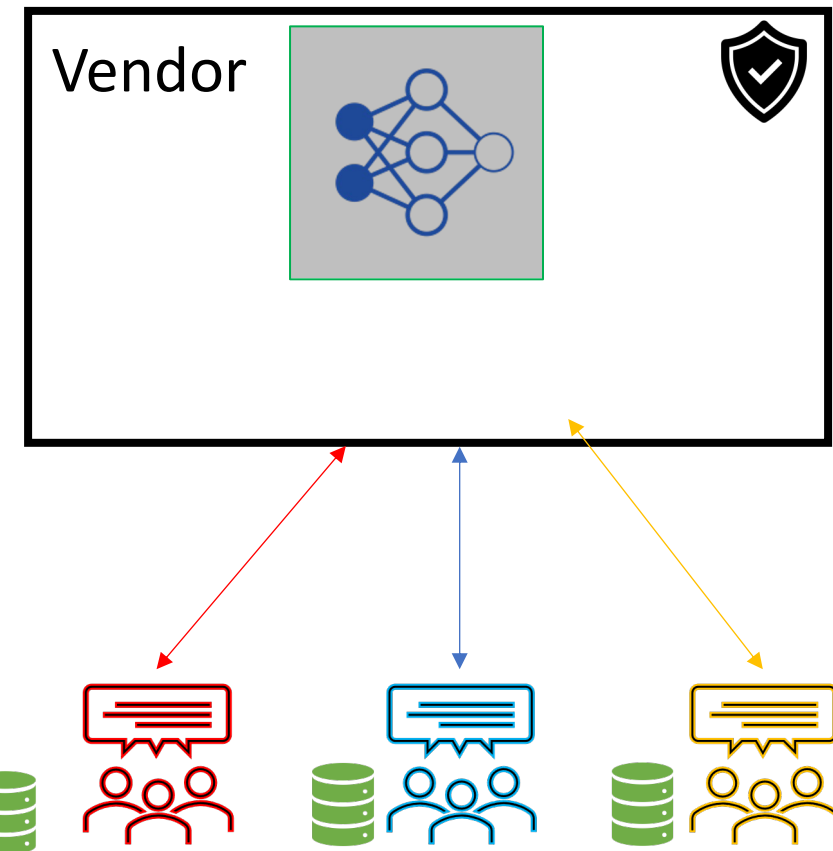


Use Case 5: Cross-customer insights



Today

Siloed Models Per Customer



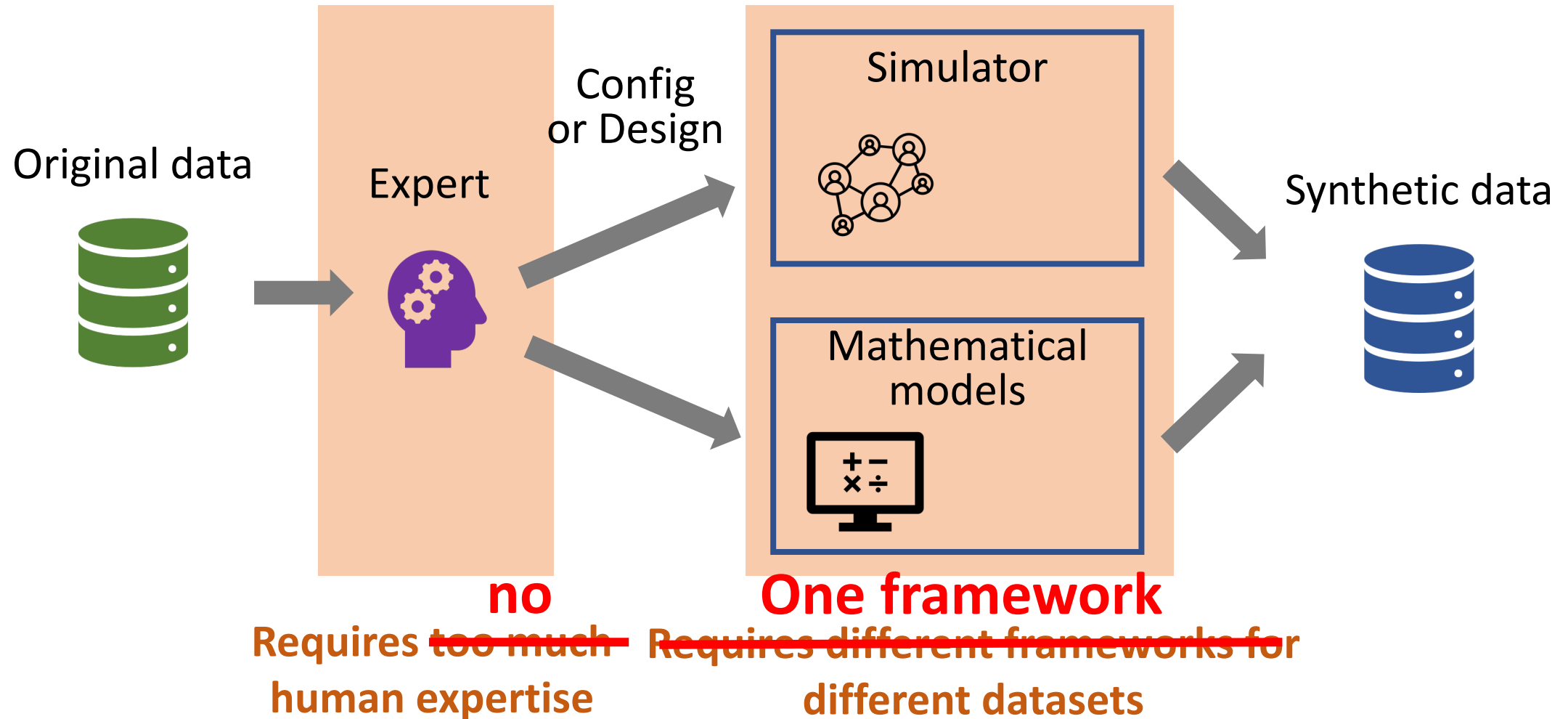
Instead

Potential to learn across customers

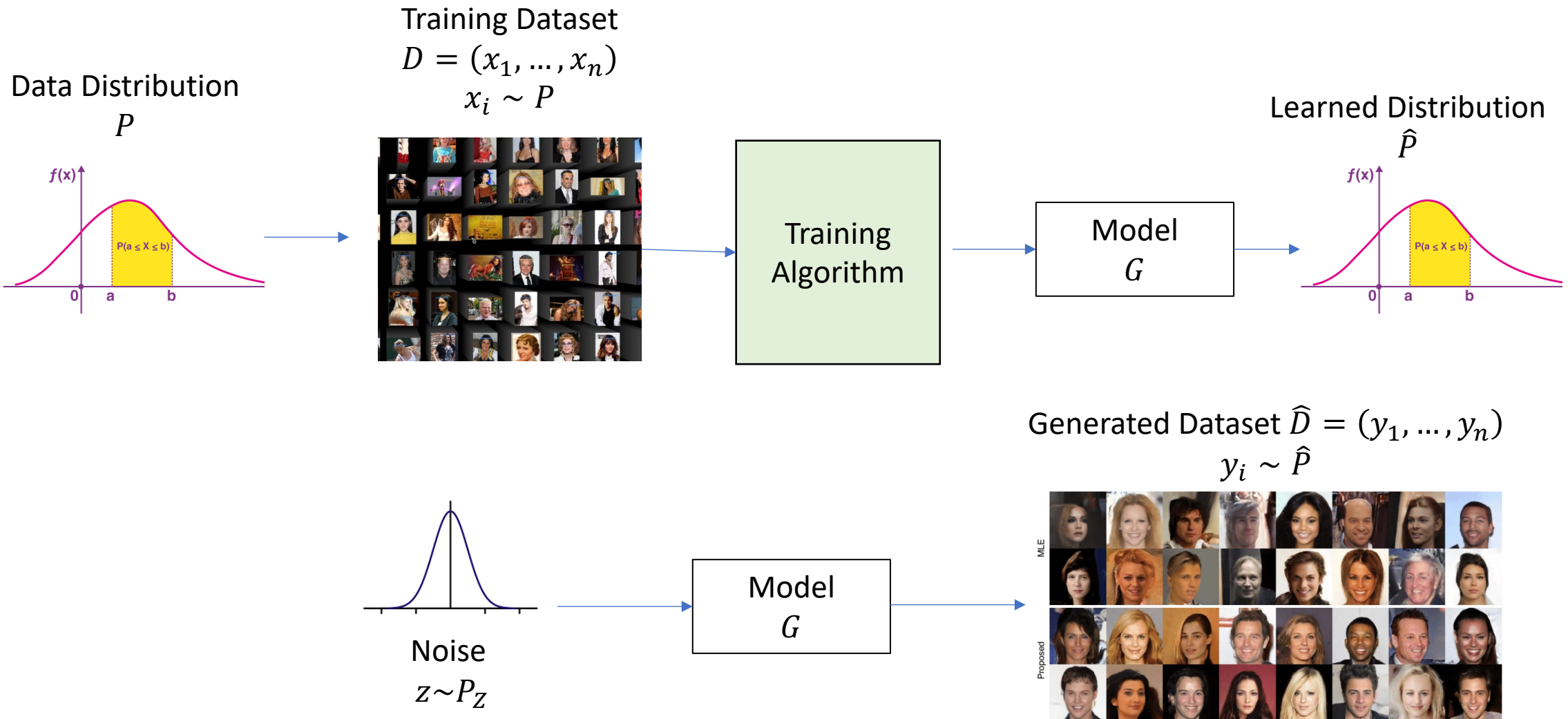
Background:

The Science Behind Synthetic Data

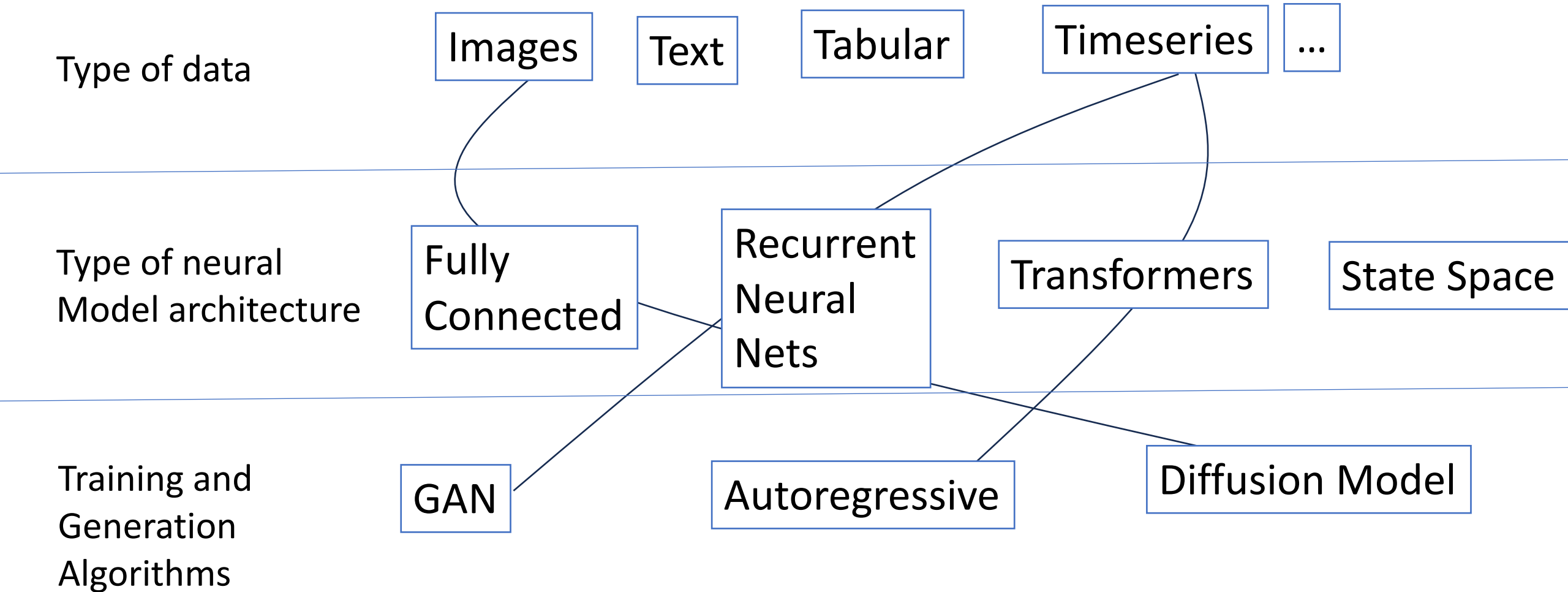
Traditional Approaches for Synthetic Data



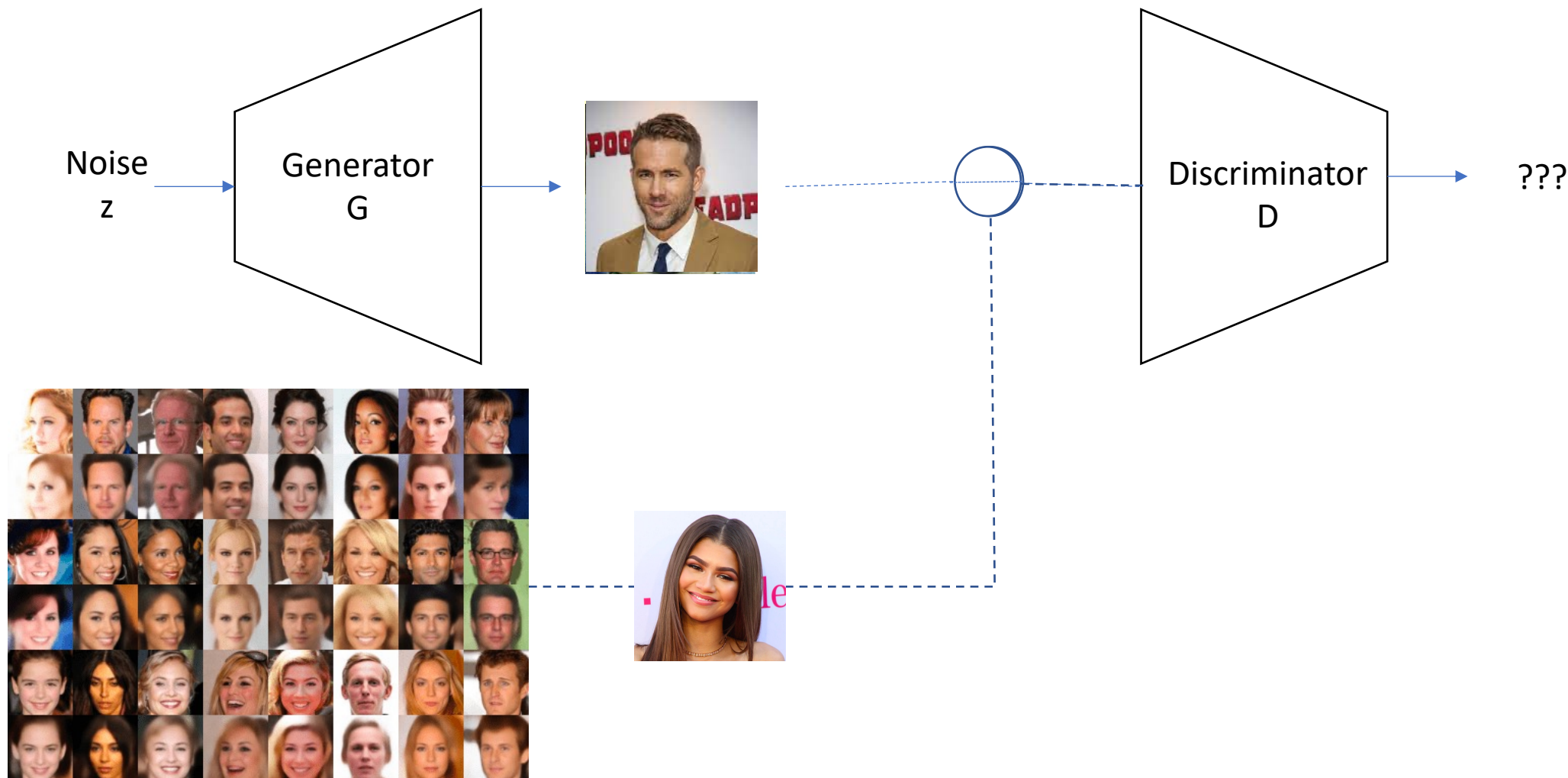
New Opportunity: Deep Generative models



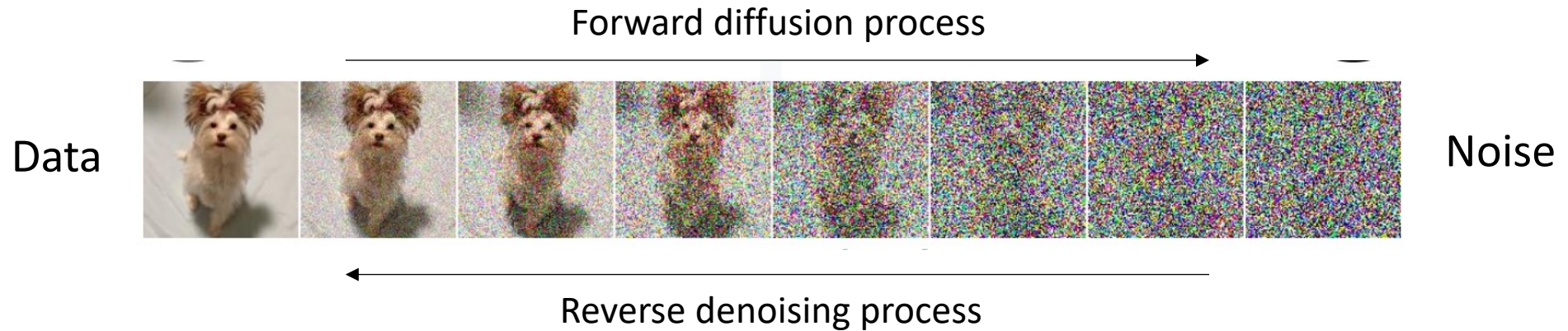
High Level Taxonomy of Deep Generative Models



Generative Adversarial Networks



Denoising Diffusion Probabilistic Models

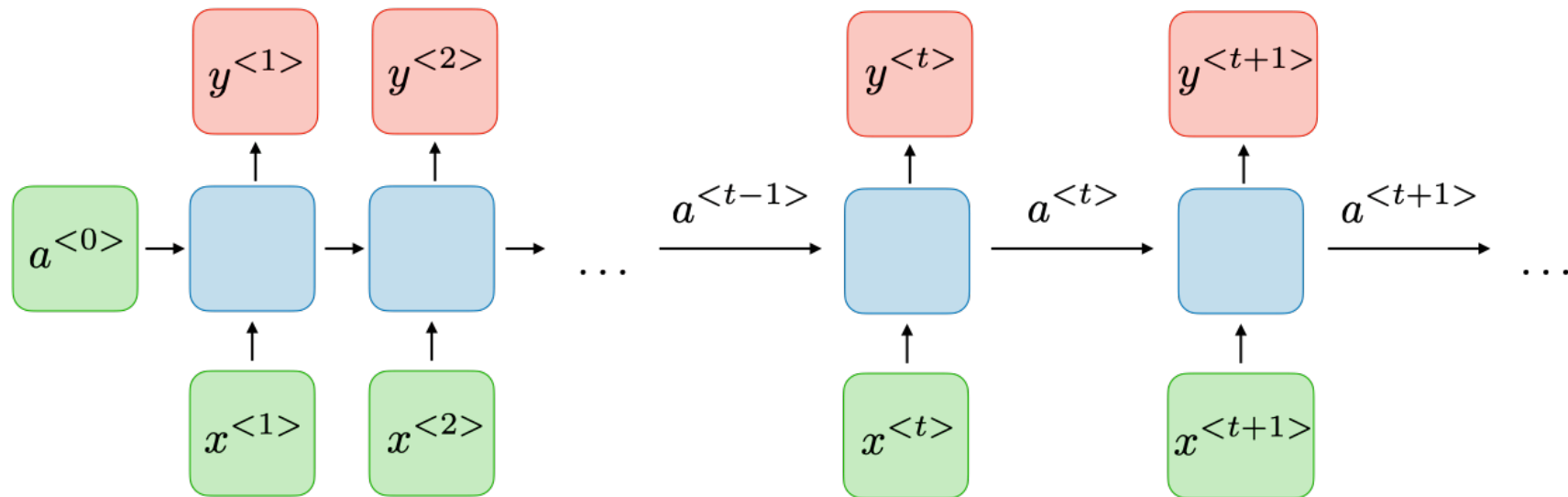


Construct two processes:

- 1) Forward diffusion process:
Progressively map data to a known noise distribution (e.g., Gaussian)
- 2) Reverse denoising process
Progressively reverse the noise process

Recurrent models

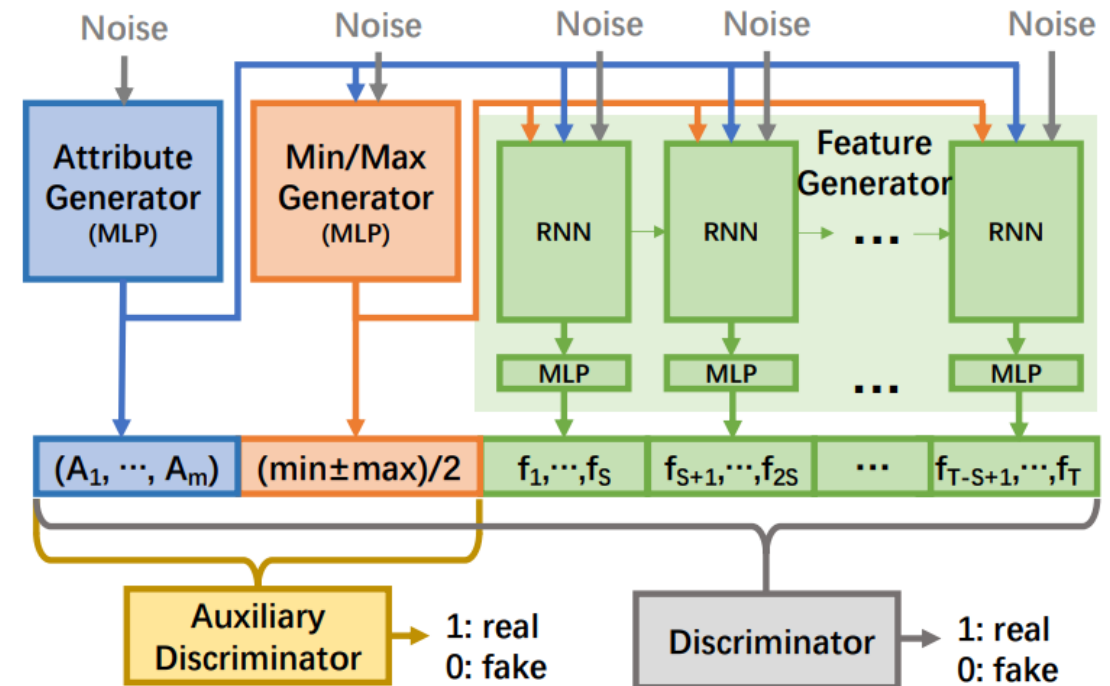
- Basic RNN structure



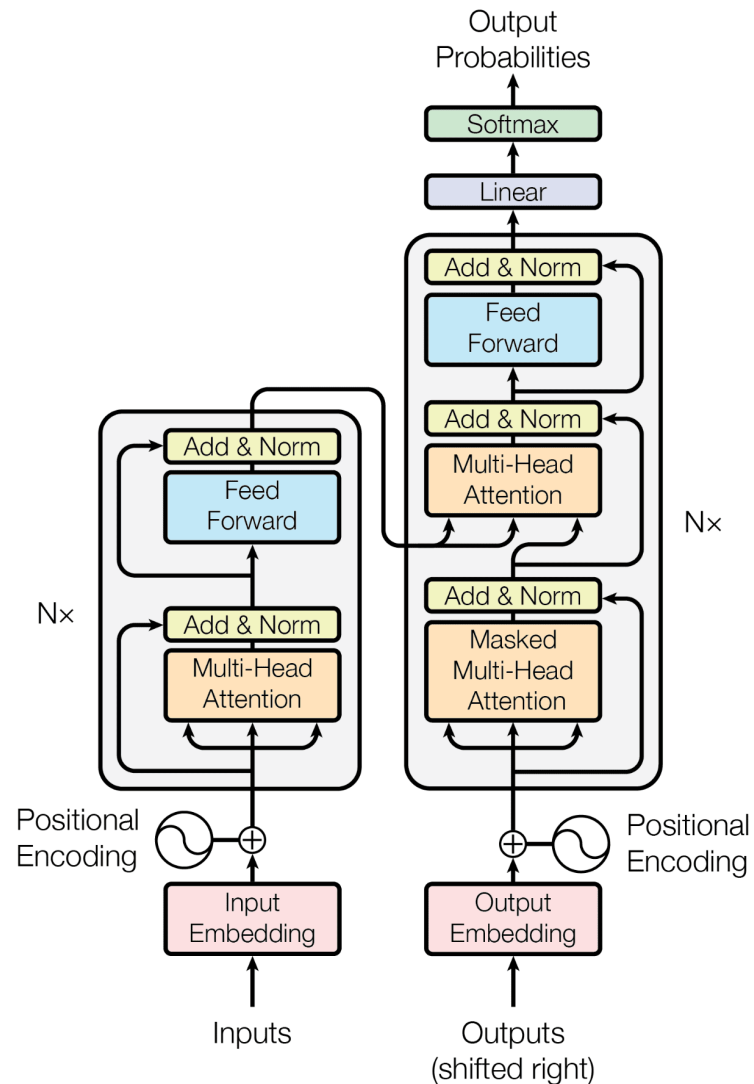
x – input sequence
 y – output sequence
 a – hidden state

You can combine RNNs with different training methods

- E.g., DoppelGANger (2020)
- RNN architecture + adversarial training (like GANs)
- Currently widely used in synthetic data industry to model time series data



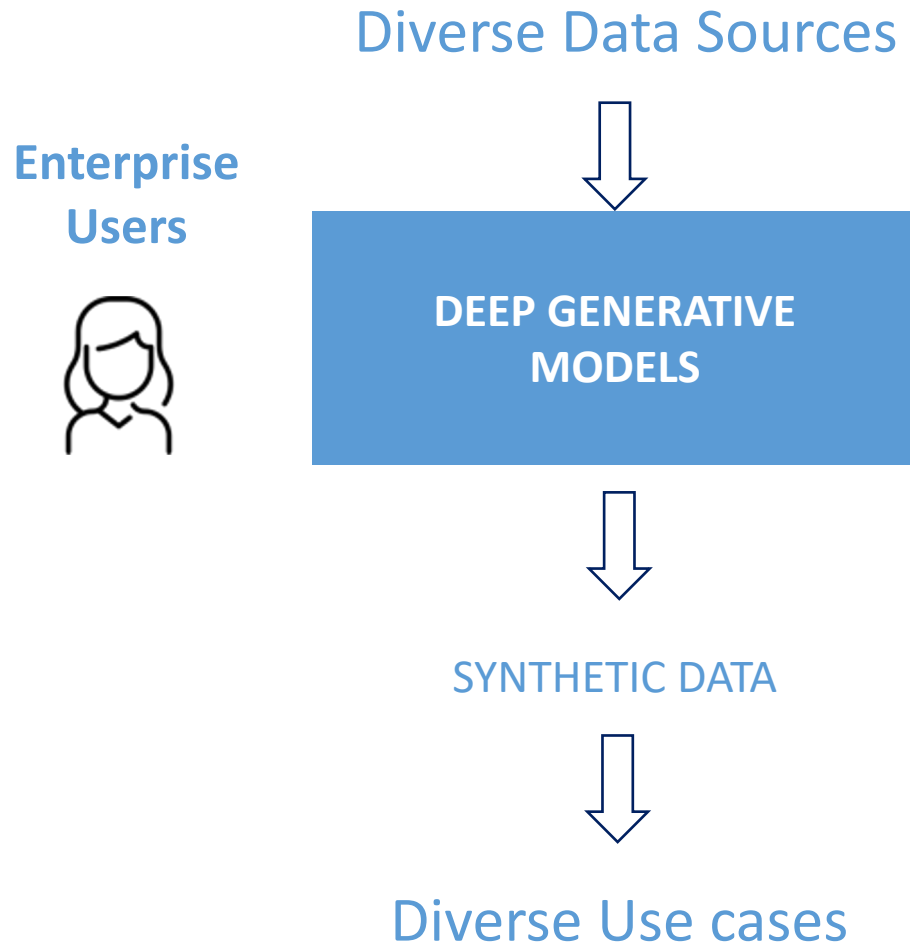
Autoregressive: The Transformer (2017)



Main idea: use an inner product score function for the attention mechanism

Challenges in synthetic data for Enterprises

Mind The Gap



Modeling fidelity and privacy needs

Use case and dataset adaption

Enterprise pipelines and compliance requirements

Our Academic Work

- Modeling Privacy and Fidelity
 - Fundamental privacy bounds: [AISTATS paper](#)
 - Statistical privacy: [In submission](#)
 - Model fidelity: [NetFID](#)
- Use case and dataset adaptation
 - State of art Timeseries Models: [DoppelGANger](#)
 - GANs for network traffic data: [NetShare](#)
 - Models for embedded systems traces: [CANGen](#)
 - Models for deep generative storage and retention: [DeepStore](#)
 - Rare Sample Generation: [RareGAN](#)
 - Improving stability and learning fidelity: [SpectralNormalization](#)

Privacy considerations with synthetic data

Attribute Privacy (Protecting Values)

- Hide Metadata fields or values

Sample-Level Privacy (User Privacy)

- Differential privacy
- Membership inference

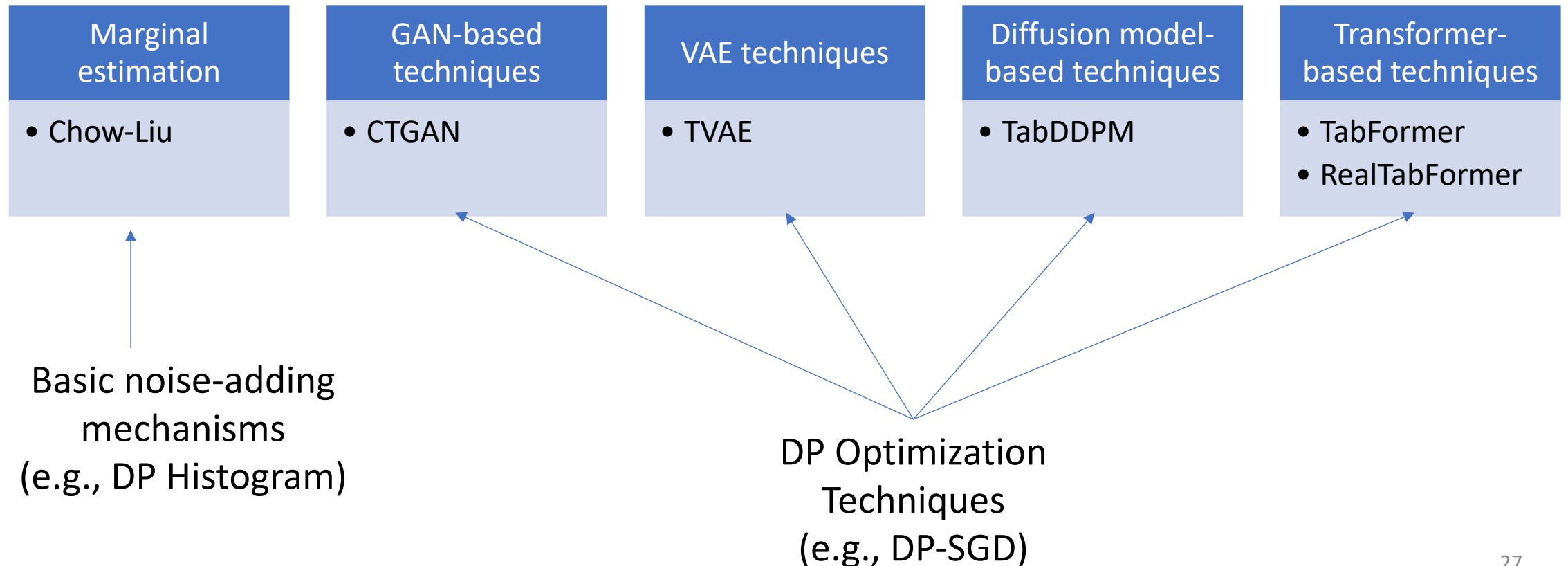
Distributional Privacy (Business Secrets)

- Aggregate functions of the data

Academic Work in Privacy

- Metadata hiding/transformations is easy with our architecture
- GAN-generated data satisfy intrinsic privacy due to generalizability
- Basic GAN satisfies some weak DP property
- Diff Private training is possible but may hurt fidelity
- Robust to membership inference attacks
- Improving privacy with pretraining with public/consortium data
- Statistical distribution privacy

Private Synthetic Tabular Data



Benchmarking Diff. Private Synthetic Data

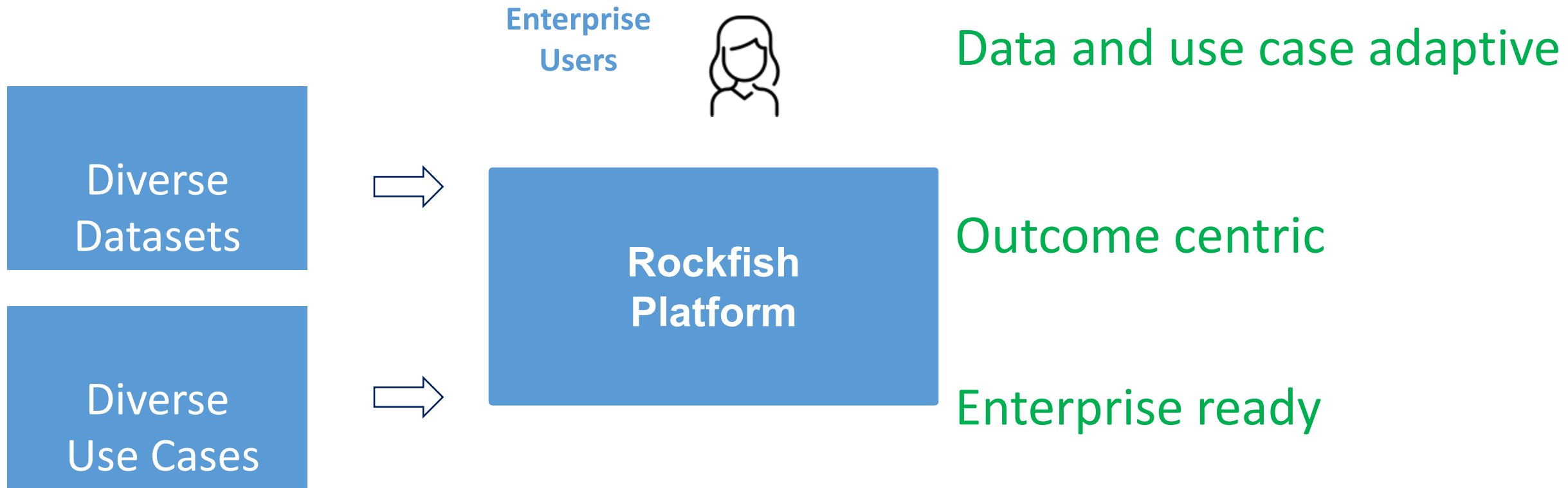
Clear winner

Mechanism	GT	Ind	Pair	Corr	F1	GT	Ind	Pair	Corr	F1
MST	69%	95%	81%	52%	44%	1.56	1.05	1.24	2.00	2.00
MWEM-PGM	19%	0%	14%	29%	33%	2.88	2.76	2.62	3.86	2.17
PrivBayes	9%	0%	0%	19%	17%	4.54	5.43	5.67	3.29	3.67
Kamino	1%	0%	0%	0%	6%	5.26	4.27	4.93	7.87	3.67
FEM	0%	0%	0%	0%	0%	4.91	4.30	4.35	5.95	5.06
RAP	1%	0%	0%	5%	0%	5.94	5.83	5.39	7.17	5.27
PATECTGAN	4%	5%	5%	5%	0%	6.17	6.45	5.90	4.90	7.65
DPCTGAN	1%	0%	0%	5%	0%	6.56	6.84	6.68	5.16	7.75
RonGauss	0%	0%	0%	0%	0%	7.35	7.06	7.11	7.61	7.61
DPGAN	0%	0%	0%	0%	0%	8.46	9.06	9.44	6.78	8.60
PATEGAN	0%	0%	0%	0%	0%	8.99	9.85	9.70	7.05	9.41

(a) **Optimal Rate**

(b) **Average Ranking**

Our work at Rockfish Data: Democratize Synthetic Data for Enterprises



Wrap-Up: Takeaway Messages

- Synthetic data using generative models have a wide range of uses
- We are entering an exciting period for synthetic data
- Early academic work has shown promise, guides intuition and design
- At Rockfish, we are democratizing this impactful tech for enterprises!