Challenges and Opportunities in Data-Centric AI

Guang Cheng
MLU Talk
Search-M5 & UCLA
Dec, 2022

Part 1: Opportunities in data-centric AI

Part 2: Synthetic data as an example

Part 3: Challenges ahead of us

Part 4: Followup activities

What is data-centric AI?

Deep learning have matured to the point that high-performance model architectures are widely available and increasingly convergent, while approaches to engineering datasets have lagged. One modern solution is data-centric AI.

"Data-centric AI encompasses methods and tools to systematically characterise, evaluate, and monitor the underlying data used to train and evaluate models. At the ML pipeline level, this means that the considerations at each stage should be informed in a data-driven manner.... We should leverage the data-driven insights as feedback to systematically improve the model." – Seedat et al (2022)

What is data-centric AI?

Deep learning have matured to the point that high-performance model architectures are widely available and increasingly convergent, while approaches to engineering datasets have lagged. One modern solution is data-centric AI.

"Data-centric AI encompasses methods and tools to systematically characterise, evaluate, and monitor the underlying data used to train and evaluate models. At the ML pipeline level, this means that the considerations at each stage should be informed in a data-driven manner.... We should leverage the data-driven insights as feedback to systematically improve the model." – Seedat et al (2022)

- In NeurIPS, 99% model-centric paper (e.g., network architecture & algorithm design & feature engineering) v.s. 1% data-centric paper (e.g., data sourcing & augmentation & generation):

- In NeurIPS, 99% model-centric paper (e.g., network architecture & algorithm design & feature engineering) v.s. 1% data-centric paper (e.g., data sourcing & augmentation & generation):

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

- In NeurIPS, 99% model-centric paper (e.g., network architecture & algorithm design & feature engineering) v.s. 1% data-centric paper (e.g., data sourcing & augmentation & generation):

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

- The current status of data-centric AI
  - fairly ambiguous since annotators are the source of data and ground truth;
  - rapidly developing, but still no active research community yet;

- In NeurIPS, 99% model-centric paper (e.g., network architecture & algorithm design & feature engineering) v.s. 1% data-centric paper (e.g., data sourcing & augmentation & generation):

	Steel defect detection	Solar panel	Surface inspection
Baseline	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

- The current status of data-centric AI
  - fairly ambiguous since annotators are the source of data and ground truth;
  - rapidly developing, but still no active research community yet;
- What is lacking in this field includes absence of tooling, best practices, and infrastructure for managing data in ML systems.

Data size matters[1]: an empirical language model (LM) scaling law

Let N denotes model size and D denotes training data size.

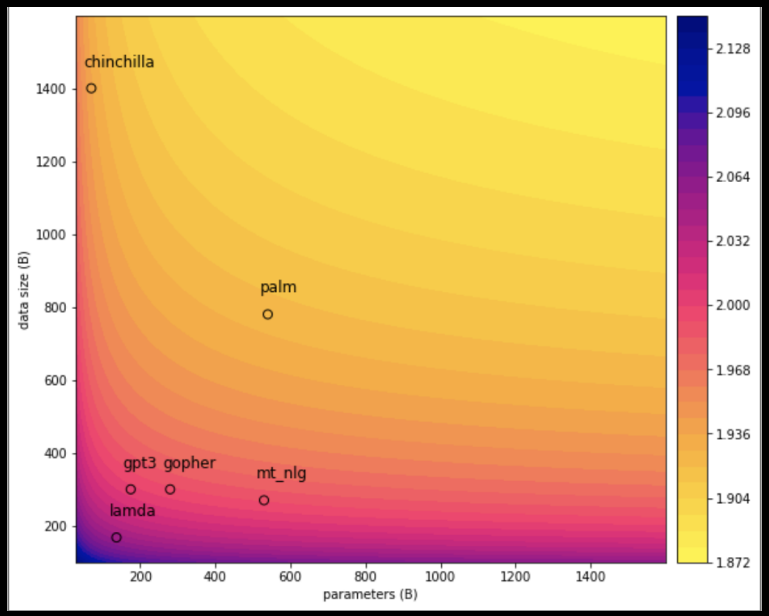
$$\text{LM loss } L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + C.$$

[1] Training compute-optimal large language models by DeepMind

Data size matters[1]: an empirical language model (LM) scaling law

Let N denotes model size and D denotes training data size.

$$\text{LM loss } L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + C.$$

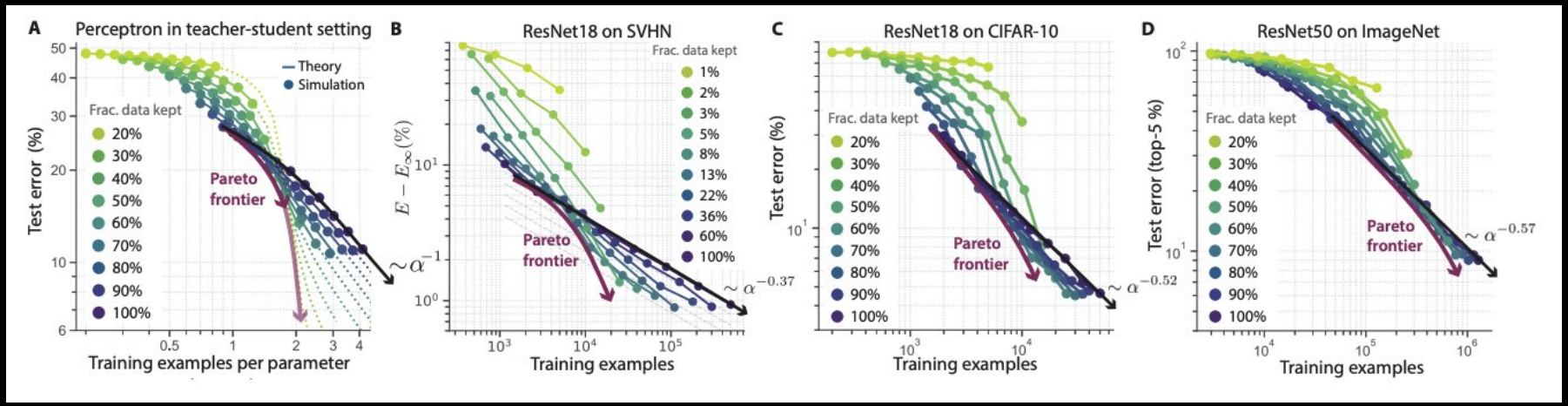


If we can leverage enough data, there is no reason to train 500B+ param models.

[1] Training compute-optimal large language models by DeepMind

Data quality matters more[2]: beyond the scaling law

Removing the "easy training samples" (defined in a self-supervised manner) improves scaling law from power level to exponential level. It means that we could go from 3% to 2% error by only adding a few carefully chosen examples, rather than collecting 10× more random ones.



[2]Beyond neural scaling laws: beating power law scaling via data pruning by Meta AI

Other examples of data-centric AI include

- a. data sourcing & labeling & pre-processing & cleaning
  – 1st generation of data industry;
- b. data generation & augmentation & pruning
  – algorithm-based, 2nd generation of data industry;
- c. data governance & compliance
  – regulation-driven, e.g., fairness and privacy.

Other examples of data-centric AI include

- a. data sourcing & labeling & pre-processing & cleaning
  – 1st generation of data industry;
- b. data generation & augmentation & pruning
  – algorithm-based, 2nd generation of data industry;
- c. data governance & compliance
  – regulation-driven, e.g., fairness and privacy.

Other examples of data-centric AI include

- a. data sourcing & labeling & pre-processing & cleaning
  – 1st generation of data industry;
- b. data generation & augmentation & pruning
  – algorithm-based, 2nd generation of data industry;
- c. data governance & compliance
  – regulation-driven, e.g., fairness and privacy.

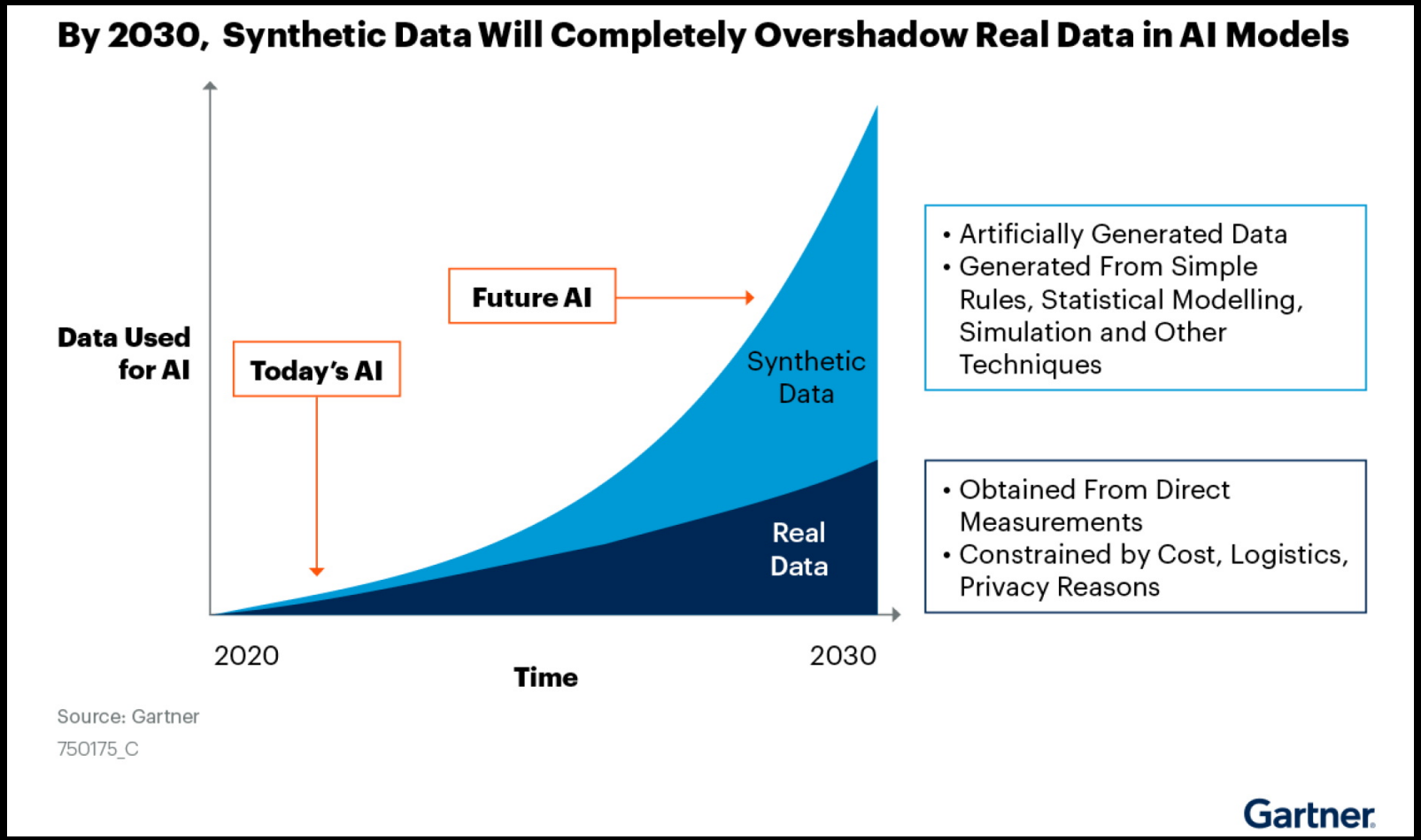
Part 1: Opportunities in data-centric AI

Part 2: Synthetic data as an example

Part 3: Challenges ahead of us

Part 4: Followup activities

Synthetic data is future



Synthetic data is everywhere: image, text, graph and tabular data.

Use cases of synthetic data

- Synthetic data for training: augment the original training data, especially when certain class instances are rare
  I. Fraud instances when training a fraud detection model;
  II. Cybersecurity attack instances when training an anomaly detection model;
- Synthetic data for testing: test operational applications development
  I. Test auto-driving in the corner cases;
  II. Test an existing healthcare policy in the new situation of Covid-19;
- Synthetic data for auditing: identify operational anomalies and risks
  I. Synthesize images or videos for products in and out of stock in Amazon Go (retail auditing);
  II. Synthesize loan applicants with diverse backgrounds, say race, to ensure fairness of AI loan decisions (compliance auditing).

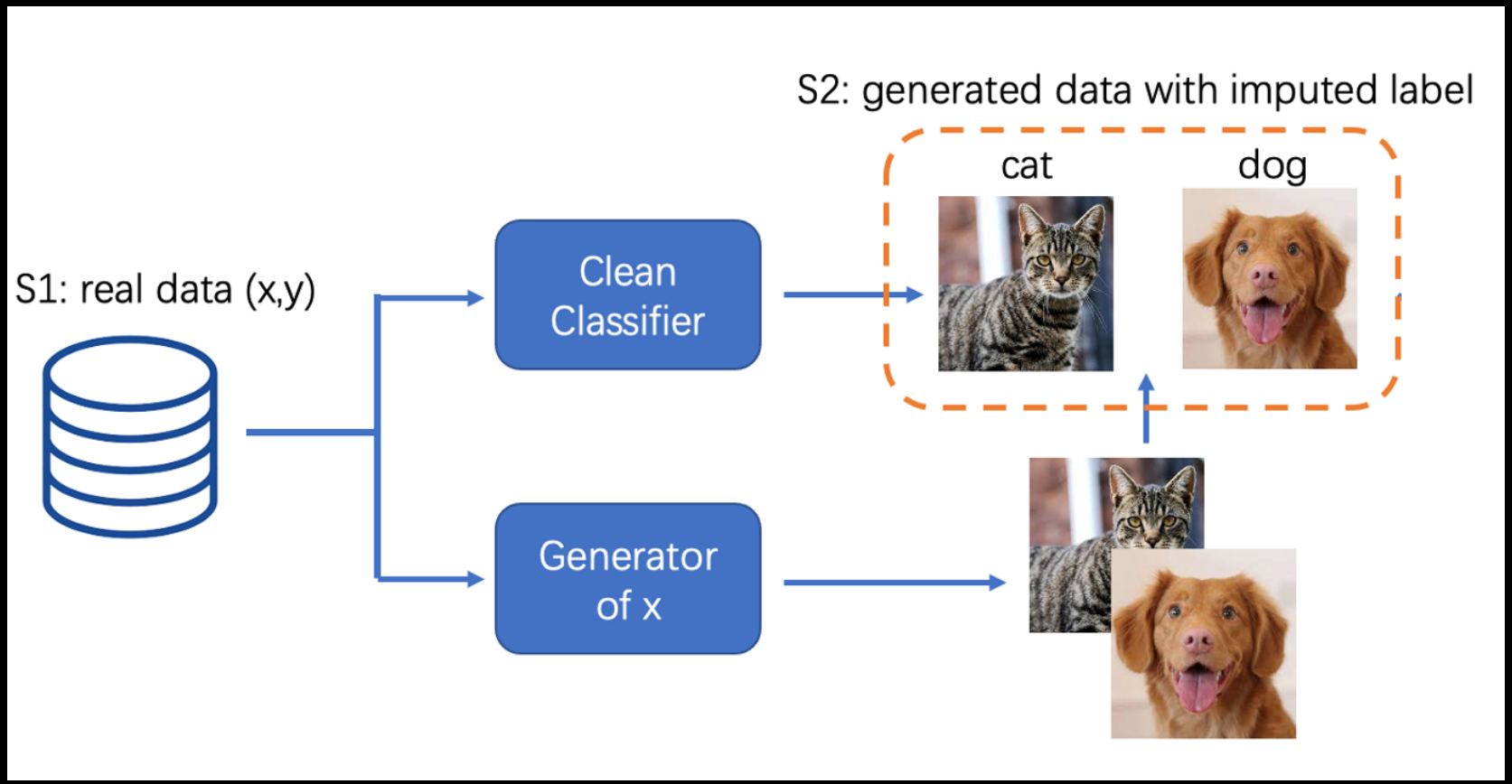
Use cases of synthetic data

- Synthetic data for training: augment the original training data, especially when certain class instances are rare
  I. Fraud instances when training a fraud detection model;
  II. Cybersecurity attack instances when training an anomaly detection model;
- Synthetic data for testing: test operational applications development
  I. Test auto-driving in the corner cases;
  II. Test an existing healthcare policy in the new situation of Covid-19;
- Synthetic data for auditing: identify operational anomalies and risks
  I. Synthesize images or videos for products in and out of stock in Amazon Go (retail auditing);
  II. Synthesize loan applicants with diverse backgrounds, say race, to ensure fairness of AI loan decisions (compliance auditing).

Use cases of synthetic data

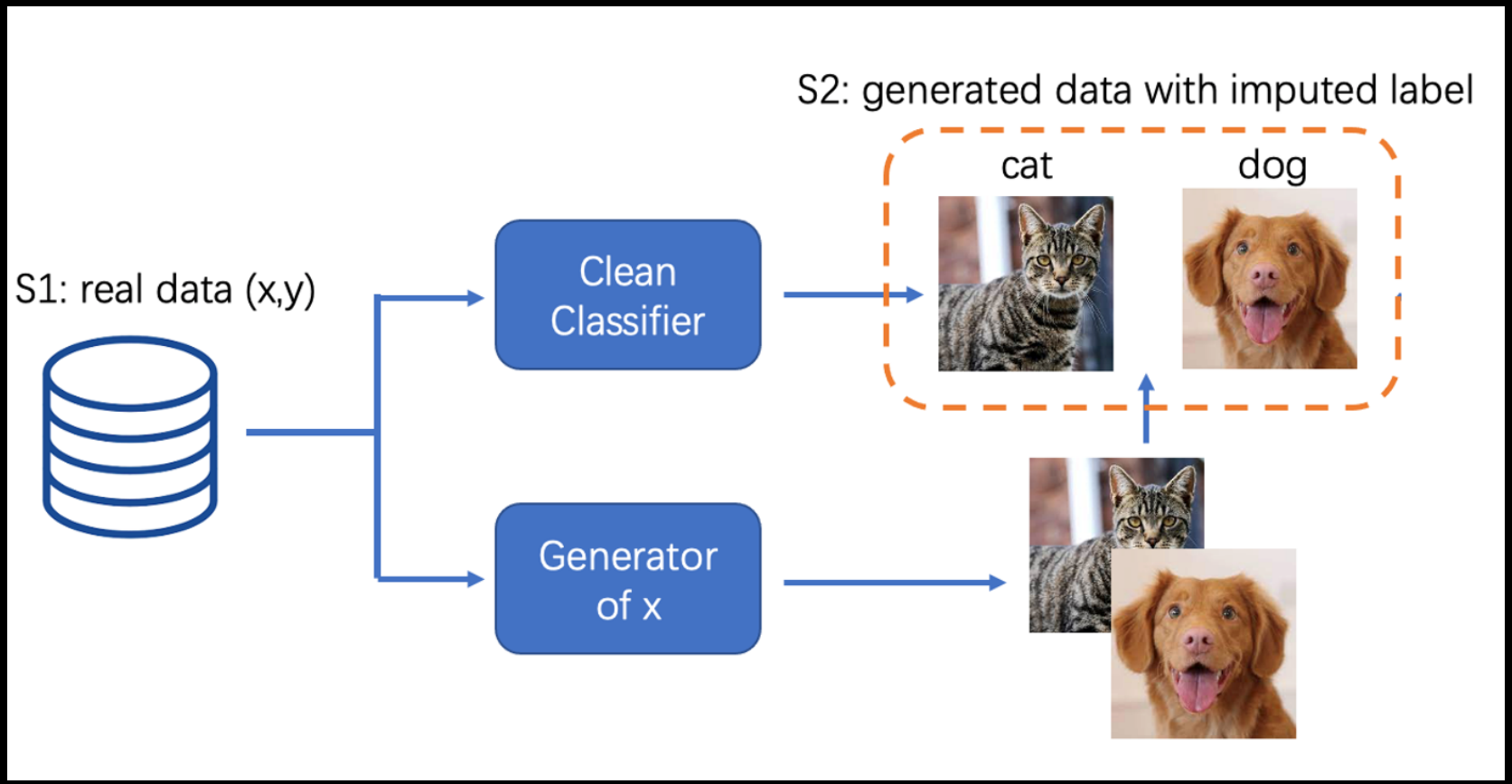
- Synthetic data for training: augment the original training data, especially when certain class instances are rare
  I. Fraud instances when training a fraud detection model;
  II. Cybersecurity attack instances when training an anomaly detection model;
- Synthetic data for testing: test operational applications development
  I. Test auto-driving in the corner cases;
  II. Test an existing healthcare policy in the new situation of Covid-19;
- Synthetic data for auditing: identify operational anomalies and risks
  I. Synthesize images or videos for products in and out of stock in Amazon Go (retail auditing);
  II. Synthesize loan applicants with diverse backgrounds, say race, to ensure fairness of AI loan decisions (compliance auditing).

An illustration on how synthetic data is generated



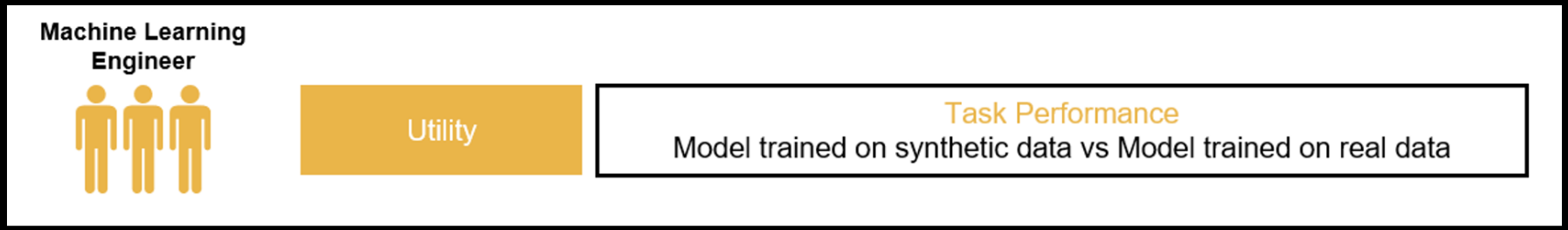
Gowal et al. (2021) observed that the mixture of real data and synthetic data, say $S_1 \cup S_2$, can be used to enhance adversarial robustness.

An illustration on how synthetic data is generated

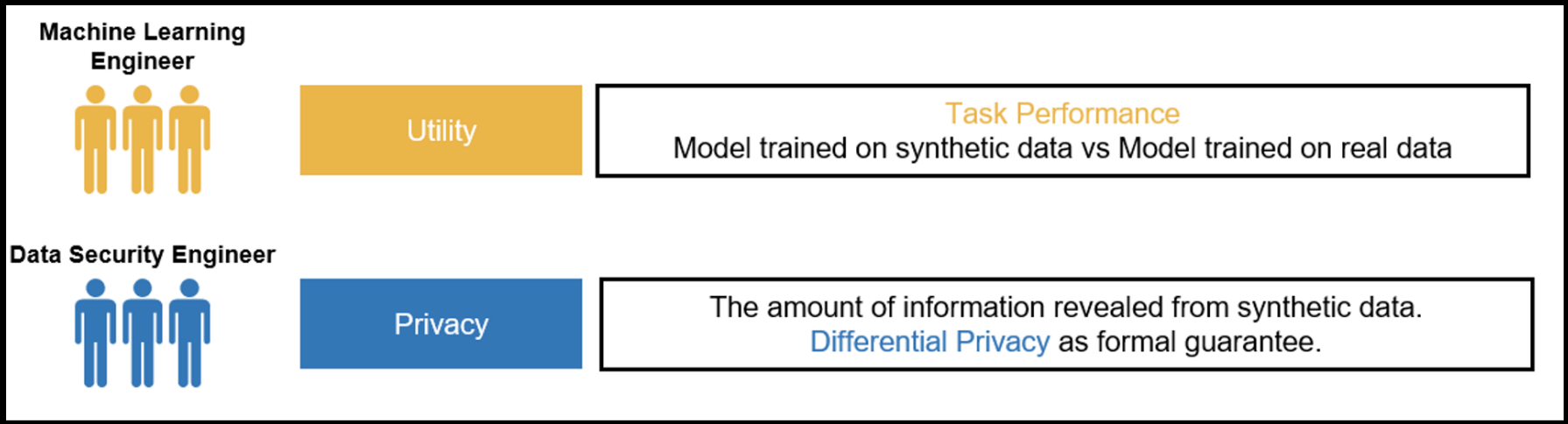


Gowal et al. (2021) observed that the mixture of real data and synthetic data, say $S_1 \cup S_2$, can be used to enhance adversarial robustness.

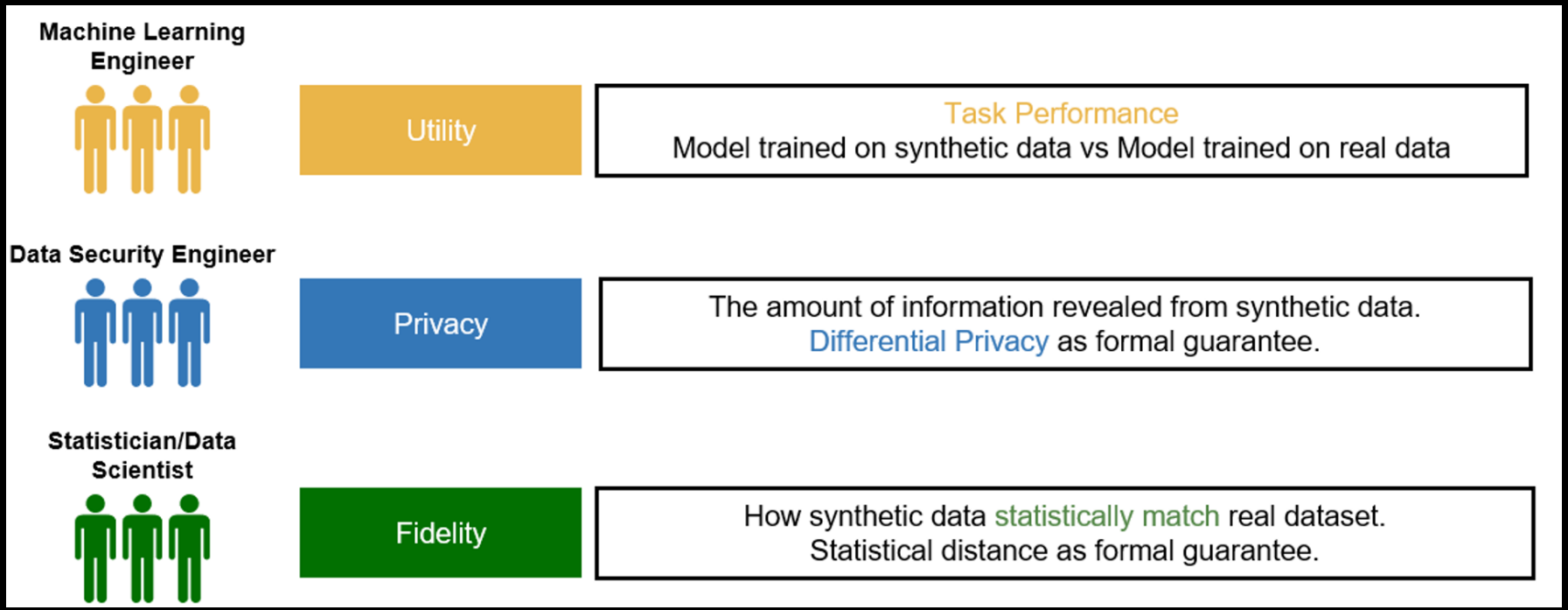
Evaluation of synthetic data: utility, privacy, fidelity



Evaluation of synthetic data: utility, privacy, fidelity



Evaluation of synthetic data: utility, privacy, fidelity



Result 1: the choice of utility metric is subtle[3]

Does synthetic dataset need to be a perfect twin of real dataset?

[3] Cheng, Wang, Potluru, Balch and C. (2022)

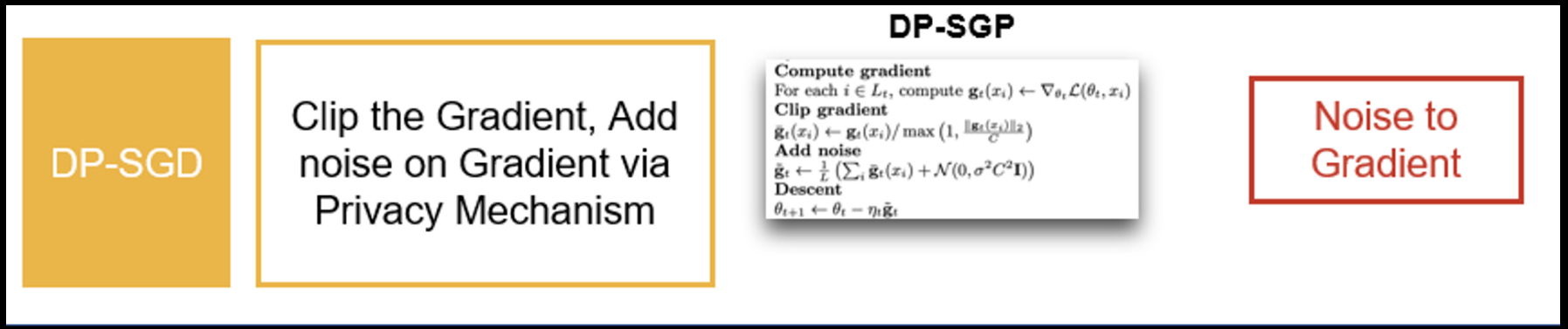
Result 1: the choice of utility metric is subtle[3]

Does synthetic dataset need to be a perfect twin of real dataset?

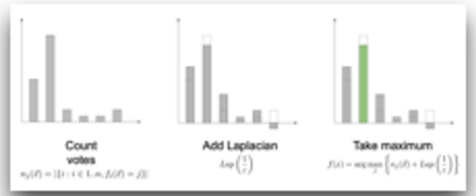
Training fraud detection model via real (blue) and synthetic data (red, green).

[3] Cheng, Wang, Potluru, Balch and C. (2022)

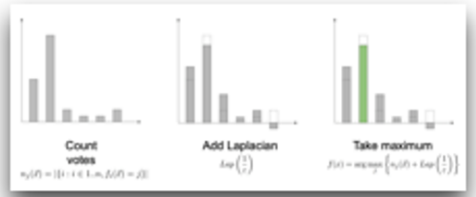
Result 2: differentially private tabular data synthesizers



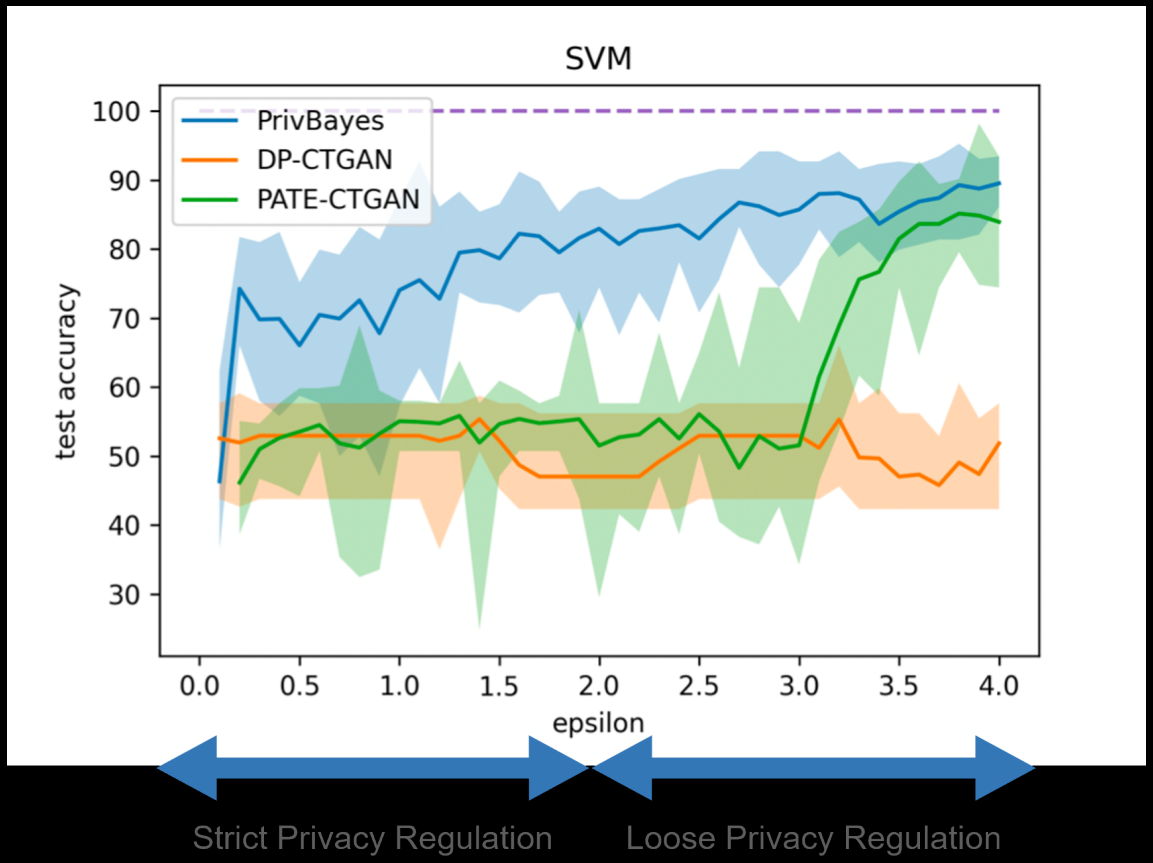
Result 2: differentially private tabular data synthesizers

DP-SGD	Clip the Gradient, Add noise on Gradient via Privacy Mechanism	**DP-SGP** **Compute gradient** For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$ **Clip gradient** $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i)/\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$ **Add noise** $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$ **Descent** $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$	Noise to Gradient
PATE	Ensemble Teacher's model, Add noise on Teacher's vote (Prediction) via Privacy Mechanism	**PATE** 	Noise to Vote

Result 2: differentially private tabular data synthesizers

DP-SGD	Clip the Gradient, Add noise on Gradient via Privacy Mechanism	**DP-SGP** **Compute gradient** For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$ **Clip gradient** $\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i)/\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$ **Add noise** $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$ **Descent** $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$	Noise to Gradient
PATE	Ensemble Teacher's model, Add noise on Teacher's vote (Prediction) via Privacy Mechanism	**PATE**  	Noise to Vote
PrivBayes	Add noise on Data's marginal distribution.	**PrivBayes** Materialize the joint distribution $\Pr[X_i, \Pi_i]$ Generate differentially private $\Pr^*[X_i, \Pi_i]$ by adding Laplace noise $\mathrm{Lap}\left(\frac{4(d-k)}{n\cdot\varepsilon}\right)$ Set negative values in $\Pr^*[X_i, \Pi_i]$ to 0 and normalize; Derive $\Pr^*[X_i \mid \Pi_i]$ from $\Pr^*[X_i, \Pi_i]$; add it to $\mathcal{P}^*$	Noise to Marginal Distribution

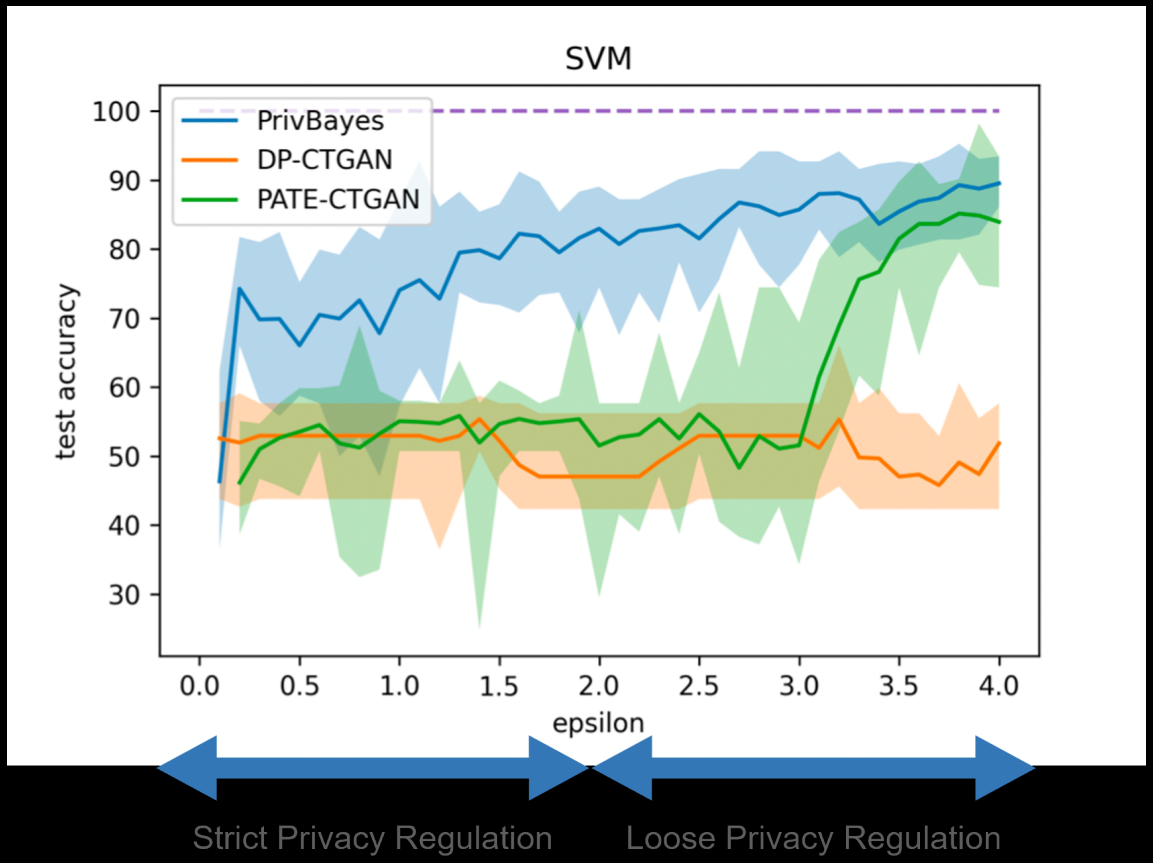
Result 2: the subtle trade-off between utility and privacy[4]



- DP-SGD makes permanent utility damage;
- PATE is good at loose privacy requirement;
- PrivBayes is great on both loose and strict privacy requirement;
- This finding persists across different classifiers, utility metrics & datasets.

[4] Liu, Wang and C. (2022)

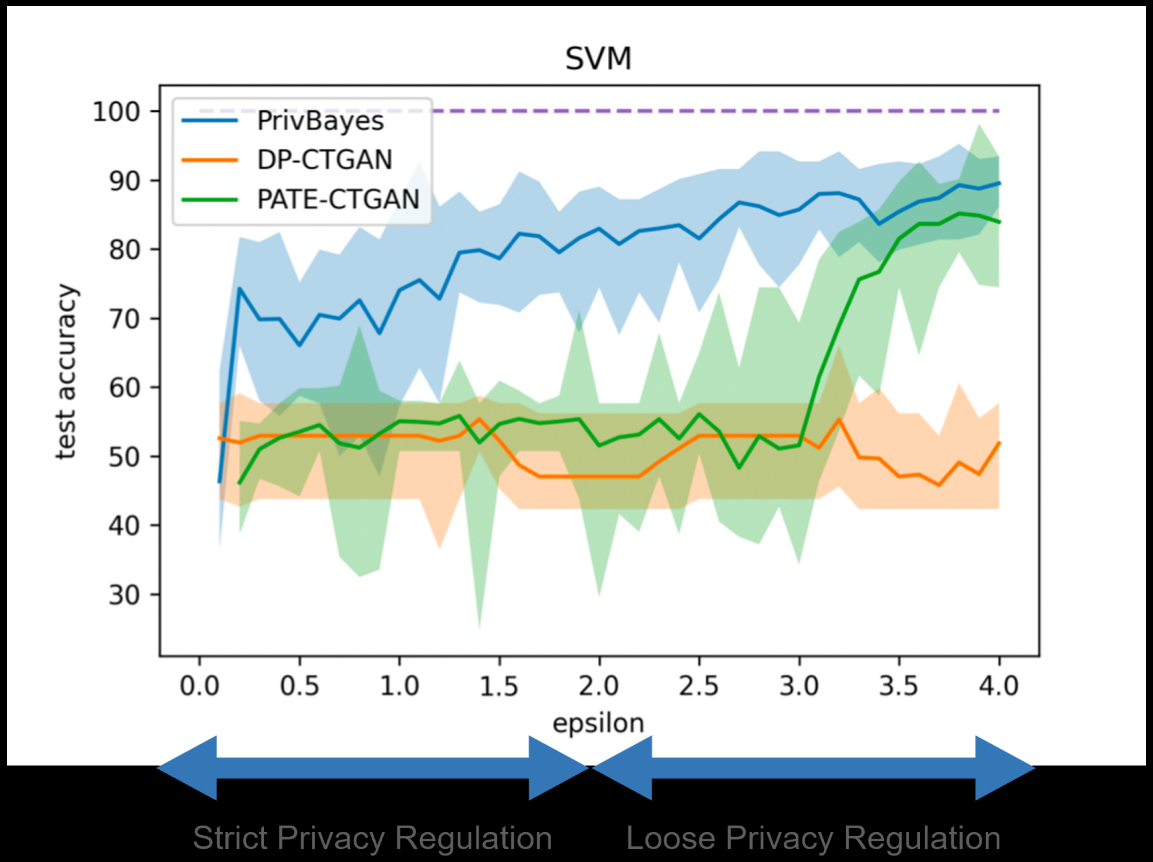
Result 2: the subtle trade-off between utility and privacy[4]



- DP-SGD makes permanent utility damage;
- PATE is good at loose privacy requirement;
- PrivBayes is great on both loose and strict privacy requirement;
- This finding persists across different classifiers, utility metrics & datasets.

[4] Liu, Wang and C. (2022)

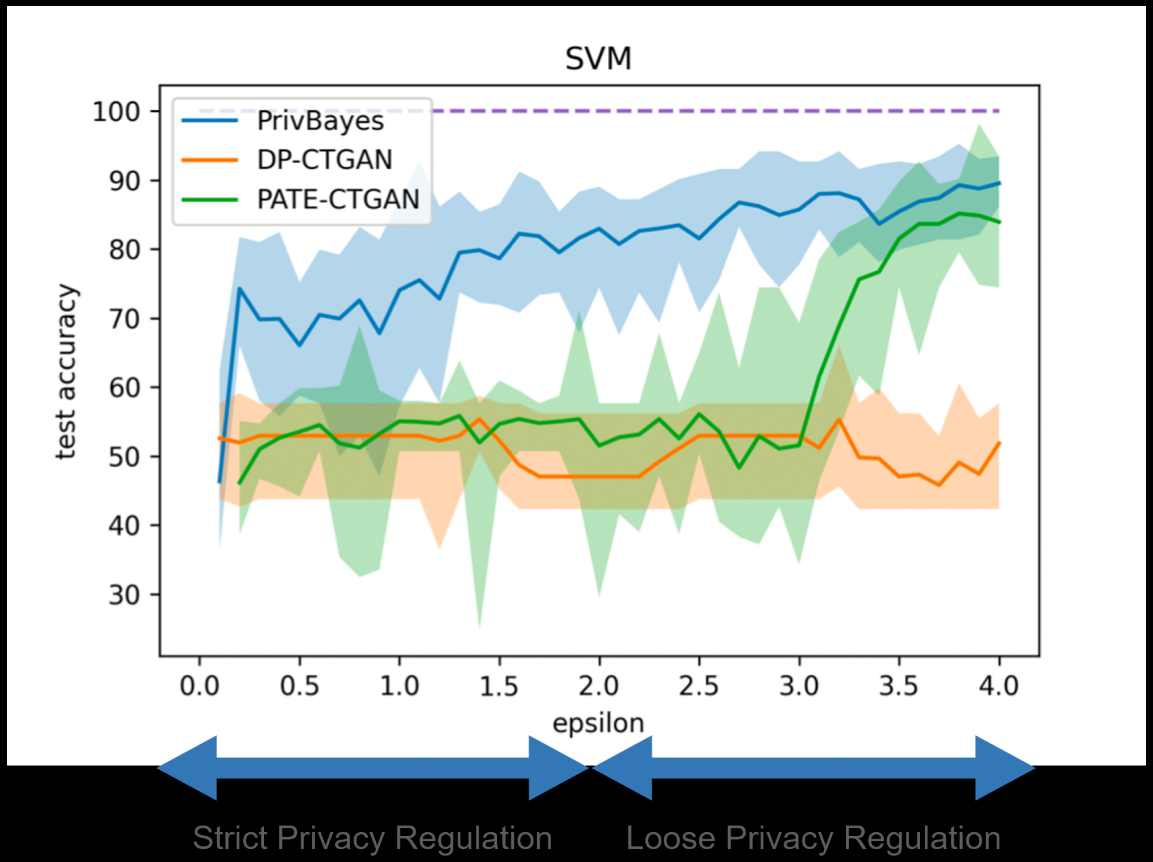
Result 2: the subtle trade-off between utility and privacy[4]



- DP-SGD makes permanent utility damage;
- PATE is good at loose privacy requirement;
- PrivBayes is great on both loose and strict privacy requirement;
- This finding persists across different classifiers, utility metrics & datasets.

[4] Liu, Wang and C. (2022)

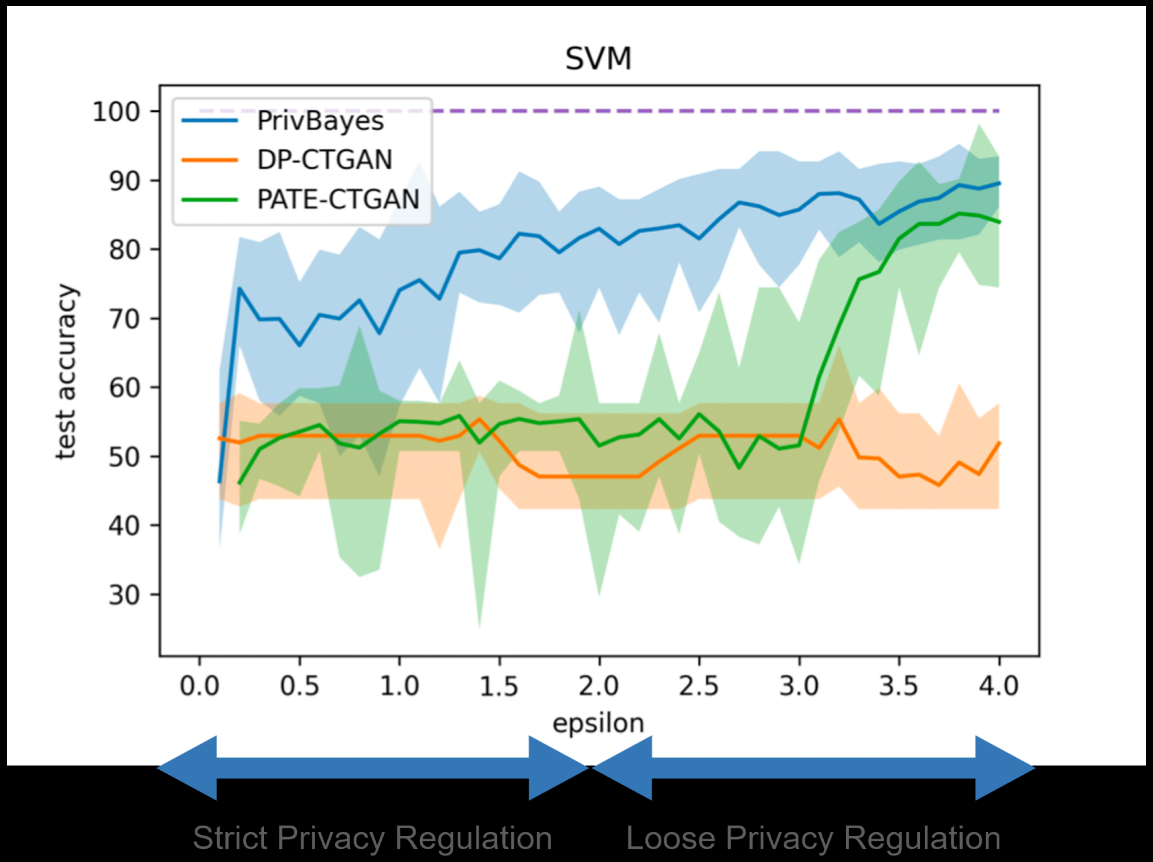
Result 2: the subtle trade-off between utility and privacy[4]



- DP-SGD makes permanent utility damage;
- PATE is good at loose privacy requirement;
- PrivBayes is great on both loose and strict privacy requirement;
- This finding persists across different classifiers, utility metrics & datasets.

[4] Liu, Wang and C. (2022)

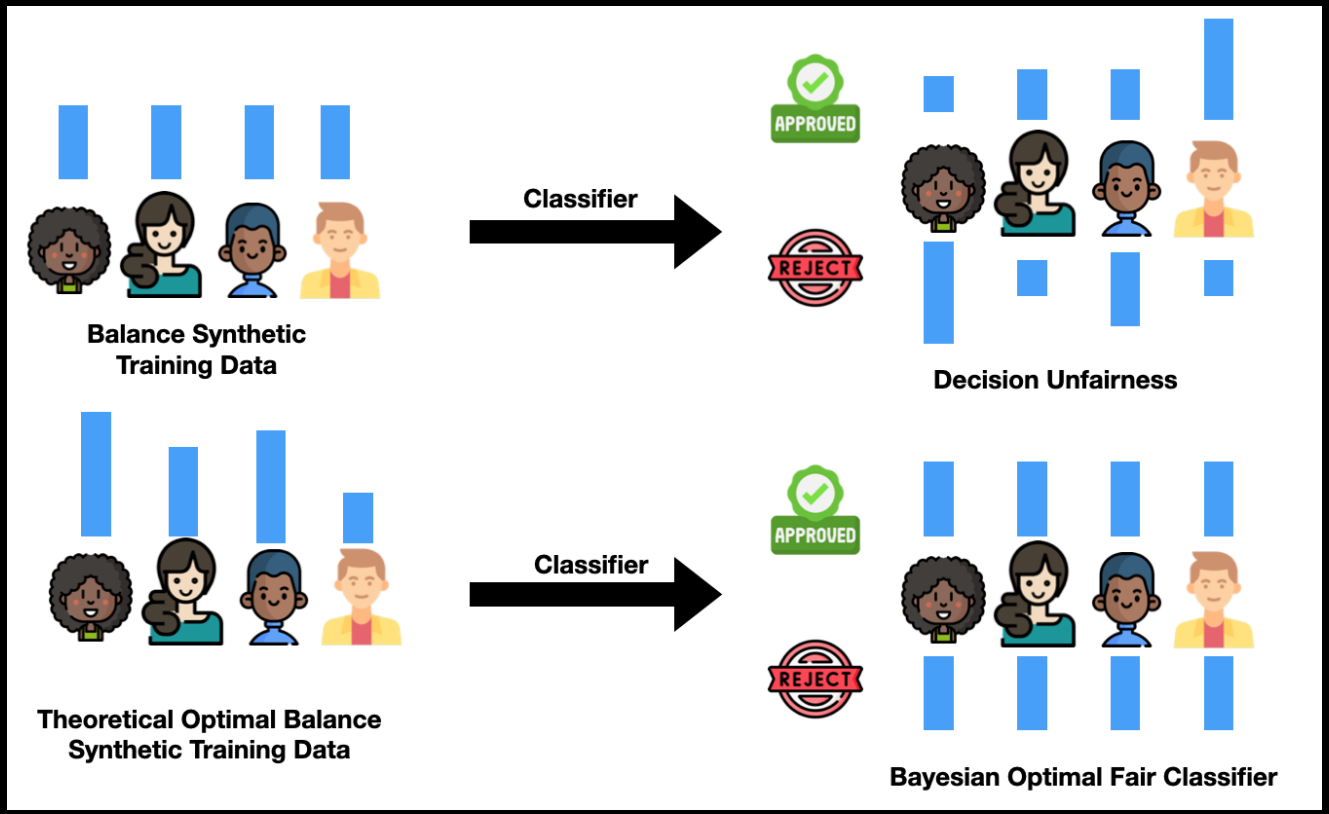
Result 2: the subtle trade-off between utility and privacy[4]



- DP-SGD makes permanent utility damage;
- PATE is good at loose privacy requirement;
- PrivBayes is great on both loose and strict privacy requirement;
- This finding persists across different classifiers, utility metrics & datasets.

[4] Liu, Wang and C. (2022)

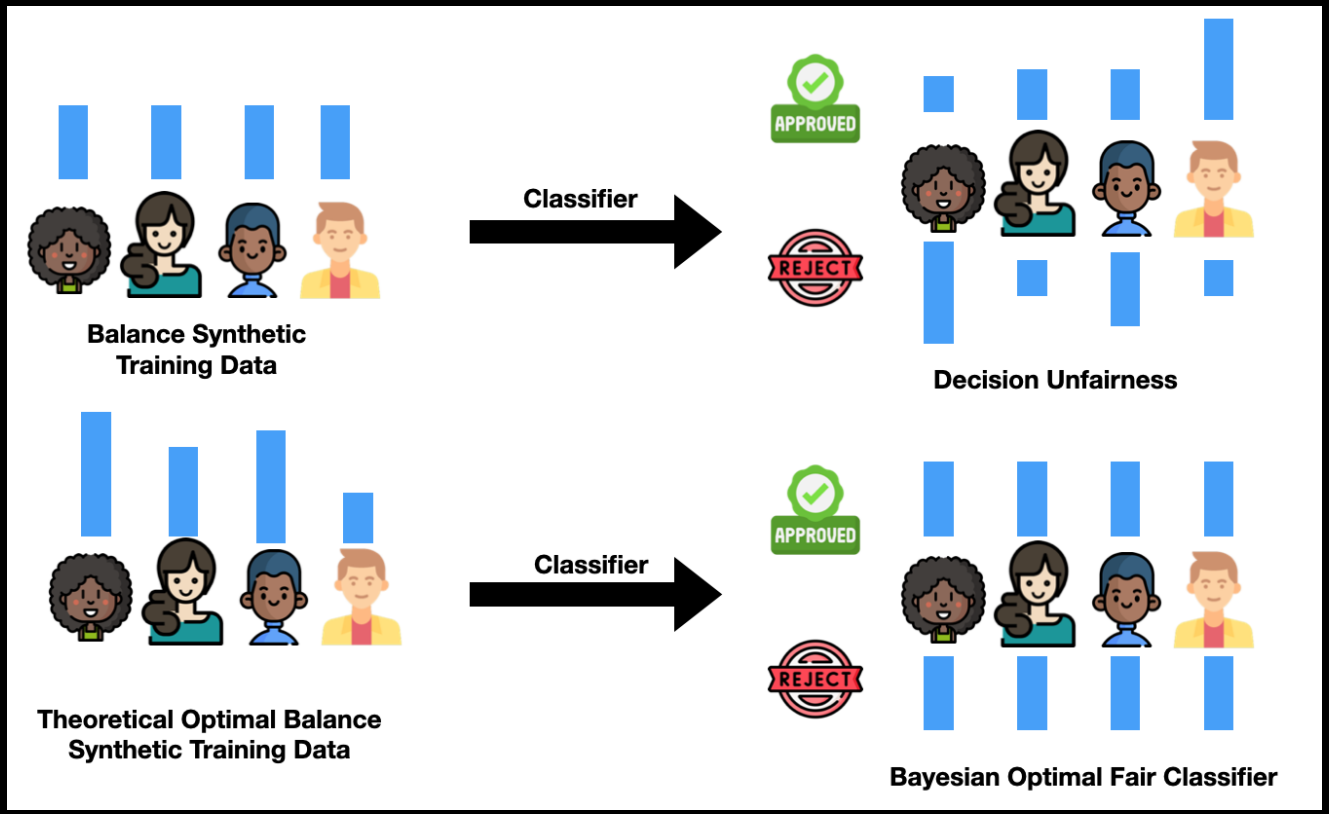
Result 3: balanced (synthetic) data does not necessarily imply a fair decision[5]



Upper figure: synthetic training dataset with demographic fairness leads to an unfair decision! Lower figure: synthetic training dataset with imbalanced demographic distribution leads to a fair decision.

[5] Zeng, Wang, Zhang and C. (2022)

Result 3: balanced (synthetic) data does not necessarily imply a fair decision[5]



Upper figure: synthetic training dataset with demographic fairness leads to an unfair decision! Lower figure: synthetic training dataset with imbalanced demographic distribution leads to a fair decision.

[5] Zeng, Wang, Zhang and C. (2022)

Part 1: Opportunities in data-centric AI

Part 2: Synthetic data as an example

Part 3: Challenges ahead of us

Part 4: Followup activities

Challenges in data-centric AI

- Foundational model vs Foundational (high quality & diverse) dataset; – how to improve quality of pre-training dataset in the GPT-3?
- Incremental re-training after data quality got improved;
- Trace AI success or failure through data;
- Interpretable synthetic data;
- New regulation, e.g., ownership, on synthetic data.

Challenges in data-centric AI

- Foundational model vs Foundational (high quality & diverse) dataset; – how to improve quality of pre-training dataset in the GPT-3?
- Incremental re-training after data quality got improved;
- Trace AI success or failure through data;
- Interpretable synthetic data;
- New regulation, e.g., ownership, on synthetic data.

Challenges in data-centric AI

- Foundational model vs Foundational (high quality & diverse) dataset; – how to improve quality of pre-training dataset in the GPT-3?
- Incremental re-training after data quality got improved;
- Trace AI success or failure through data;
- Interpretable synthetic data;
- New regulation, e.g., ownership, on synthetic data.

Challenges in data-centric AI

- Foundational model vs Foundational (high quality & diverse) dataset; – how to improve quality of pre-training dataset in the GPT-3?
- Incremental re-training after data quality got improved;
- Trace AI success or failure through data;
- Interpretable synthetic data;
- New regulation, e.g., ownership, on synthetic data.

Challenges in data-centric AI

- Foundational model vs Foundational (high quality & diverse) dataset; – how to improve quality of pre-training dataset in the GPT-3?
- Incremental re-training after data quality got improved;
- Trace AI success or failure through data;
- Interpretable synthetic data;
- New regulation, e.g., ownership, on synthetic data.

Followup activities

- Data-centric AI seminar (organized by Guang Cheng and Lihong Li): please signup datacentricai@amazon.com for future talks;
- Synthetic data workshop in the April of 2023 at UCLA, sponsored by Amazon-UCLA science hub;
- UCLA Stat 213 on Synthetic Data offered in the Spring of 2023.

Followup activities

- Data-centric AI seminar (organized by Guang Cheng and Lihong Li): please signup datacentricai@amazon.com for future talks;
- Synthetic data workshop in the April of 2023 at UCLA, sponsored by Amazon-UCLA science hub;
- UCLA Stat 213 on Synthetic Data offered in the Spring of 2023.

Followup activities

- Data-centric AI seminar (organized by Guang Cheng and Lihong Li): please signup datacentricai@amazon.com for future talks;
- Synthetic data workshop in the April of 2023 at UCLA, sponsored by Amazon-UCLA science hub;
- UCLA Stat 213 on Synthetic Data offered in the Spring of 2023.


GRAZAS
DANKIE
KIITOS
TODA
MERCI
TAKK
QUJAN
THANK YOU
QUJANARSUAQ
TODA
ASANTE
TACK
MERCI
XIE XIE
MITELA
TÄNAN
THANKS
TAKK
ASANTE
GRAZAS
多謝
MAHALO
THANK YOU
GRACIAS
OBRIGADO
APACHE
TODA
SPASIBO
MERCI
どうも
OBRIGADA
THANK YOU
GRACIAS
OBRIGADO
MUCHAS GRACIAS
YEWO
DANKIE
THANK YOU
SUKRIA
TODA
ASANTE
GRACIAS
TACK
TAKK
XIE XIE
MITELA
GRAZAS
TÄNAN
TACK
TAKK
KIITOS
ДЯКУЮ
SPASIBO
DANKE
ДЯКУЙ
HVALA
PAXMET!
多謝
GRAZIE
DZIEKUJE
DEKUJI
TAK
TESEKKÜR EDERIM
GRACIAS
THANKS
TODA
KIITOS
SPAS
唔該
DO JEH
FOOLE TUNK
THANK YOU
DANKIE
CHOKRAN
SHUKRAN
SPAS
WILIWNI
BAIE DANKIE
NGIYABONGA
FAKAAUE
NGIYABONGA
HVALA
SUKRIA
MERCI
धनयवाद
XIE XIE
SUKRIA
ASANTE
KOP KUN
SPASIB
AITEH
PAXMAT
СПАСИБО
TESEKKÜR EDERIM
ДЗЯКУЙ
הבר הדות
GRAZAS
THANK YOU
धनयवाद
SPAS
XIE XIE
TODA
THANK YOU
多謝
TA
SALAMAT PO
TAK