Do You Really Need Public Data?

Surrogate Public Data for Differential Privacy on Tabular Data

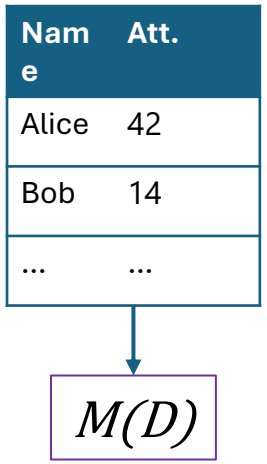
PPPC 2026

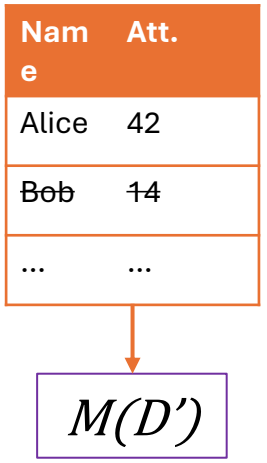
Shlomi Hod[‡] (BU) Lucas Rosenblatt[‡] (NYU) Julia Stoyanovich (NYU)

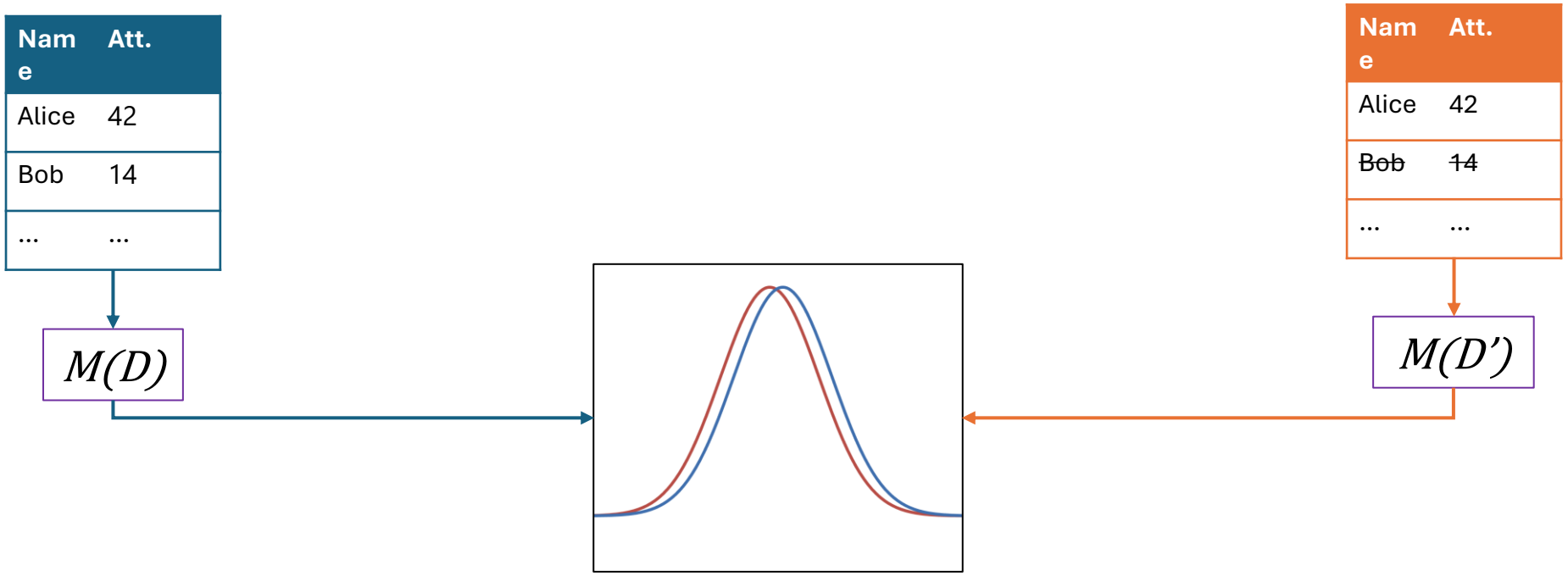
[‡]Equal contribution.

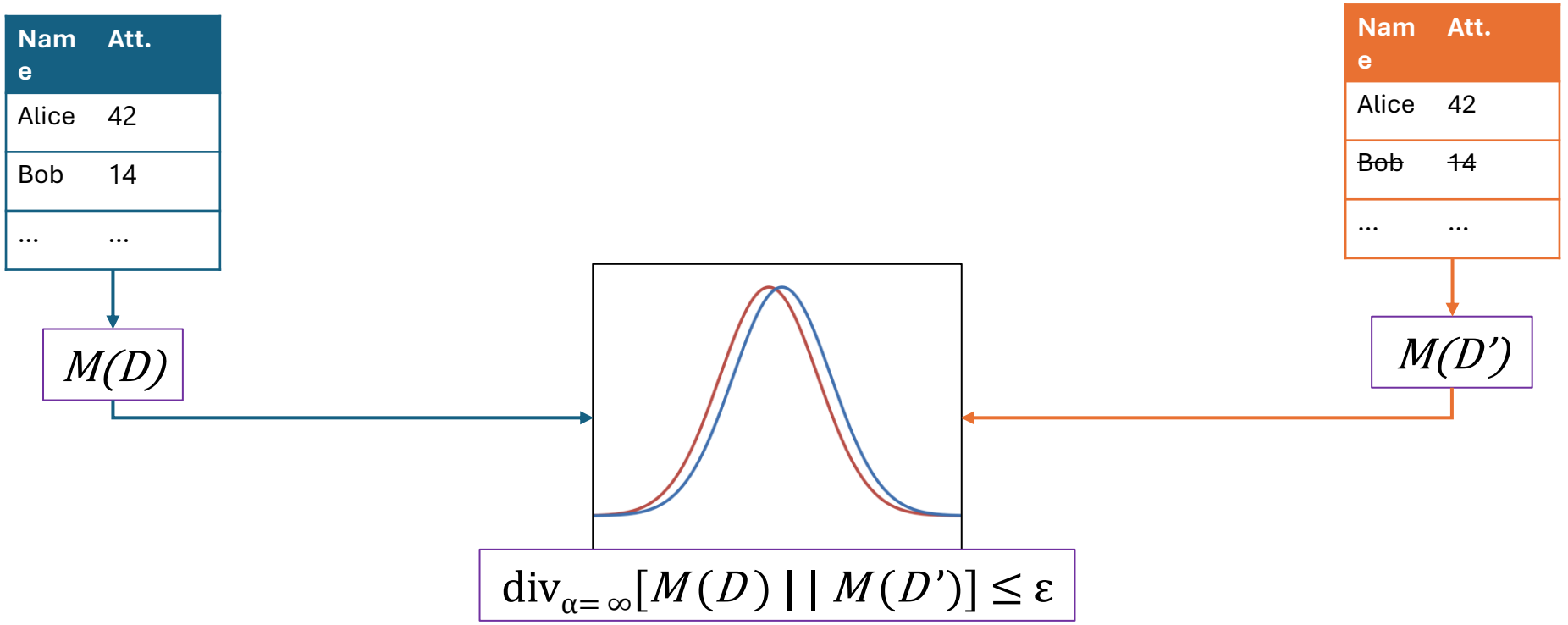
Name	Att.
Alice	42
Bob	14
...	...

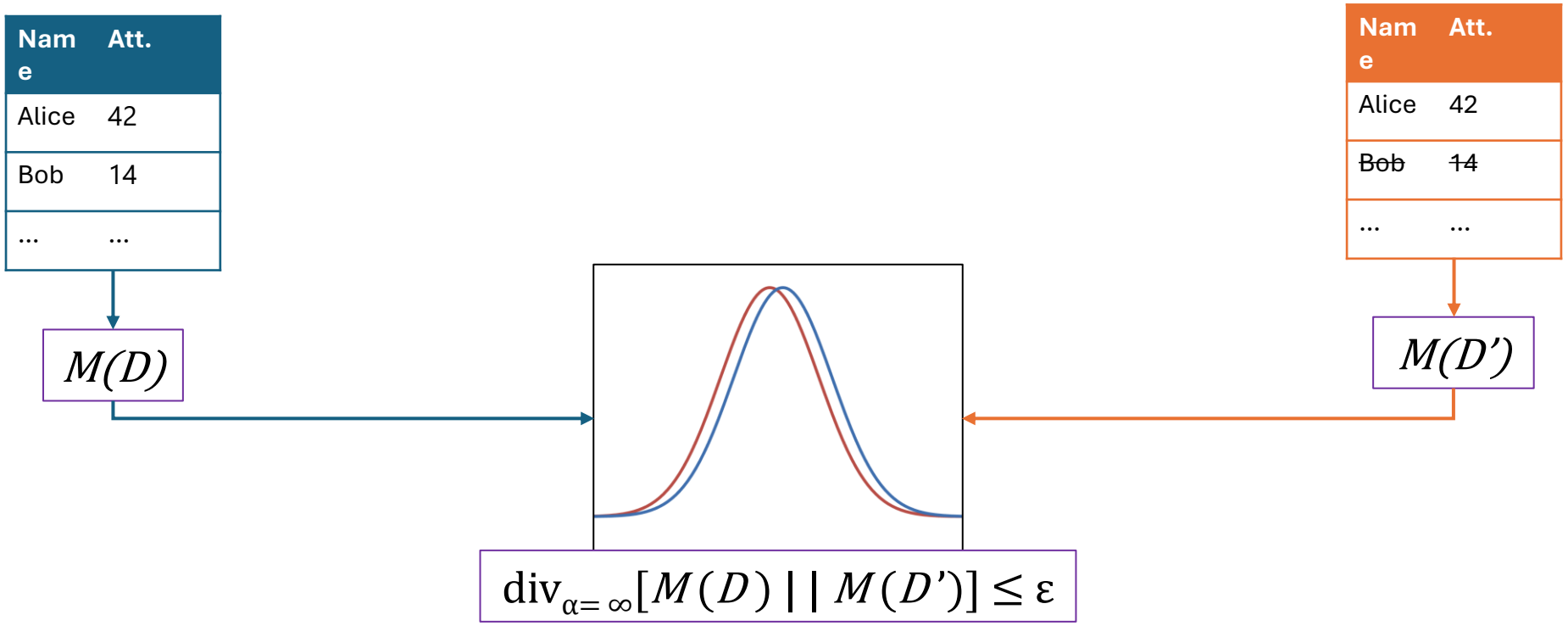
Name	Att.
Alice	42
~~Bob~~	~~14~~
...	...


Name
Att.
Alice
42
Bob
14
...
...
M(D)


Name
Att.
Alice
42
Bob
14
...
...
M(D')


Name
Att.
Alice
42
Bob
14
...
...
M(D)
Name
Att.
Alice
42
Bob
14
...
...
M(D')


Name
Att.
Alice
42
Bob
14
...
...
M(D)
Name
Att.
Alice
42
Bob
14
...
...
M(D')
$\mathrm{div}_{\alpha=\infty}[M(D) \,||\, M(D')] \leq \varepsilon$

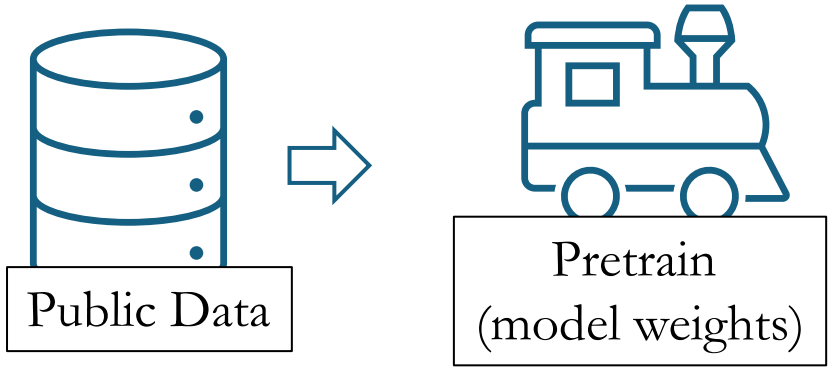


M satisfies **differential privacy (DP)** if its output remains **nearly unchanged** when the data of a single individual is modified/removed.

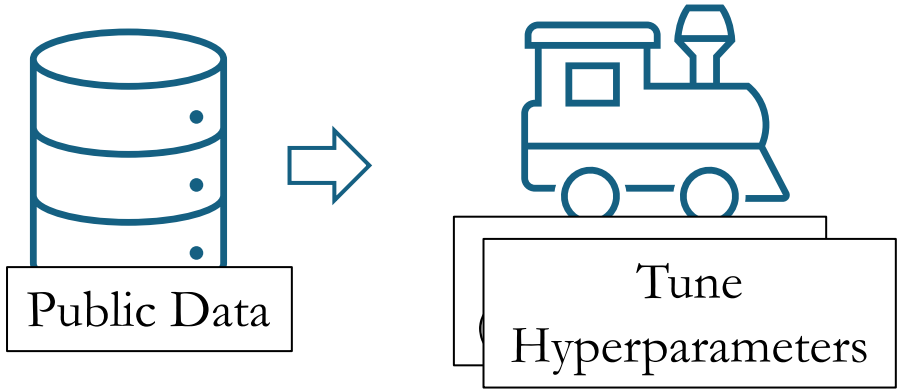
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

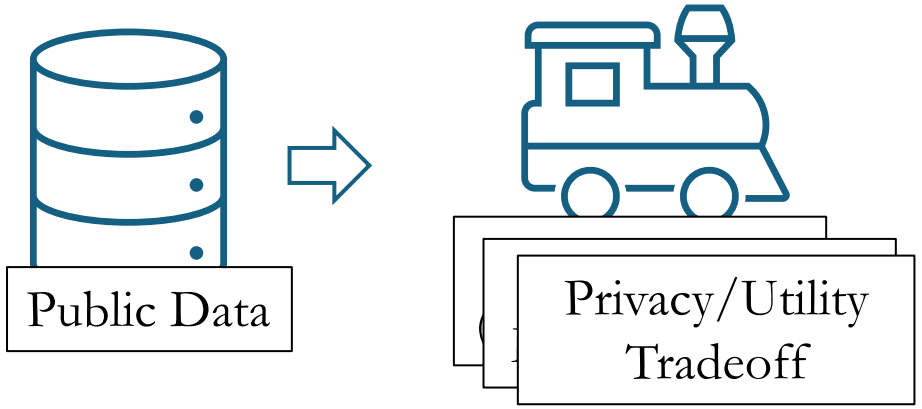
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



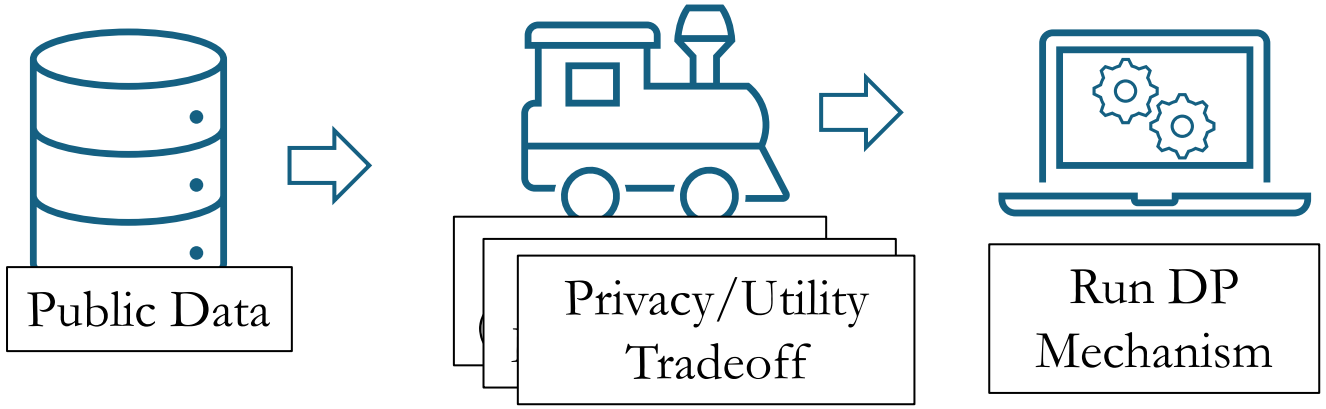
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



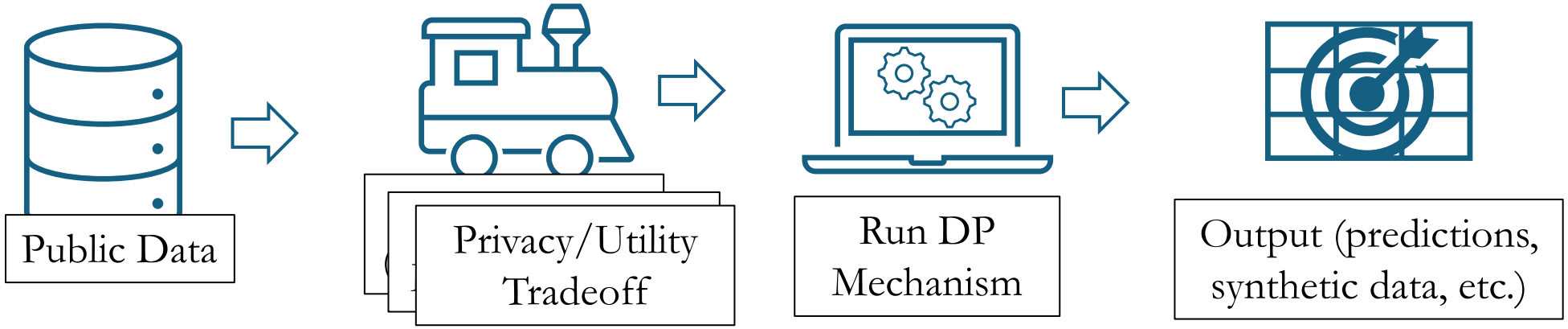
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



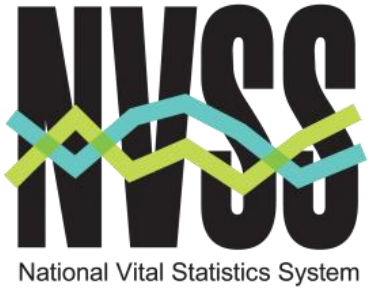
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

Private data: **2014 Israel’s live birth registry**

In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

Private data: **2014 Israel’s live birth registry**

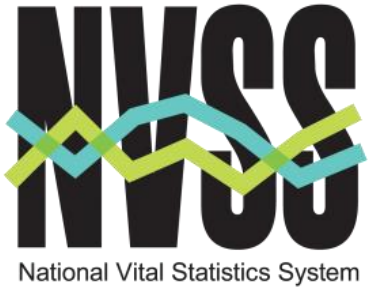
Public data: **2019 US’s NVSS data**



In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

Private data: **2014 Israel’s live birth registry**

Public data: **2019 US’s NVSS data**

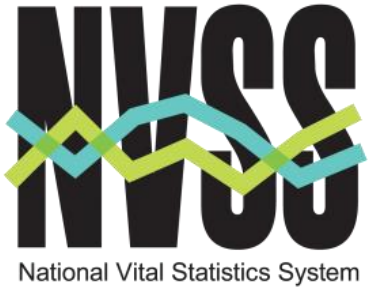


Private data: **Census 2020**

In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

Private data: **2014 Israel’s live birth registry**

Public data: **2019 US’s NVSS data**



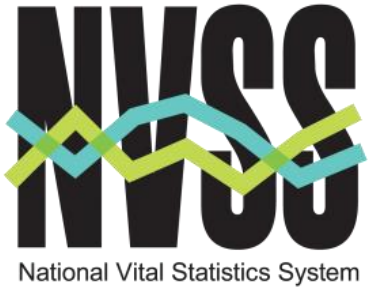
Private data: **Census 2020**

Public data: **Census 2010**

In "differential privacy world," we are allowed to use **public** data to improve our model outputs.

Private data: **2014 Israel's live birth registry**

Public data: **2019 US's NVSS data**

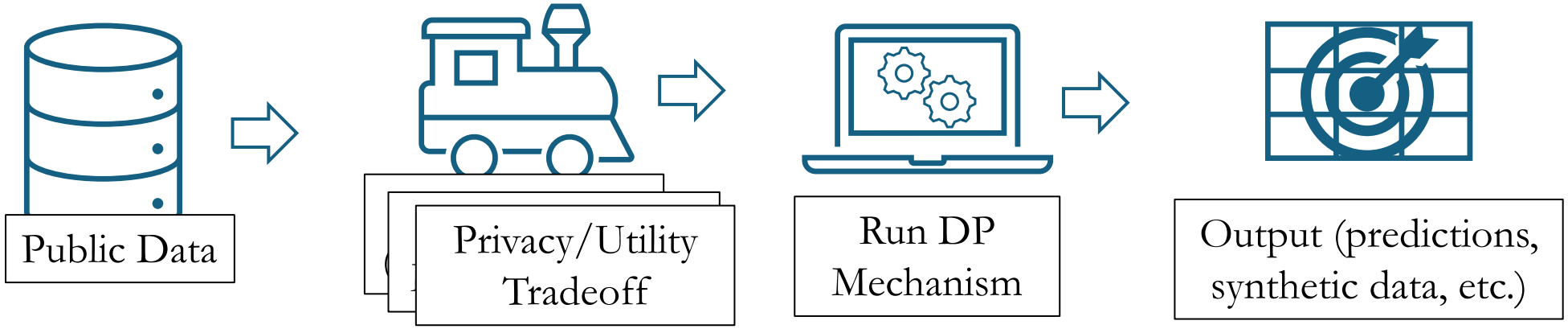


Private data: **Census 2020**

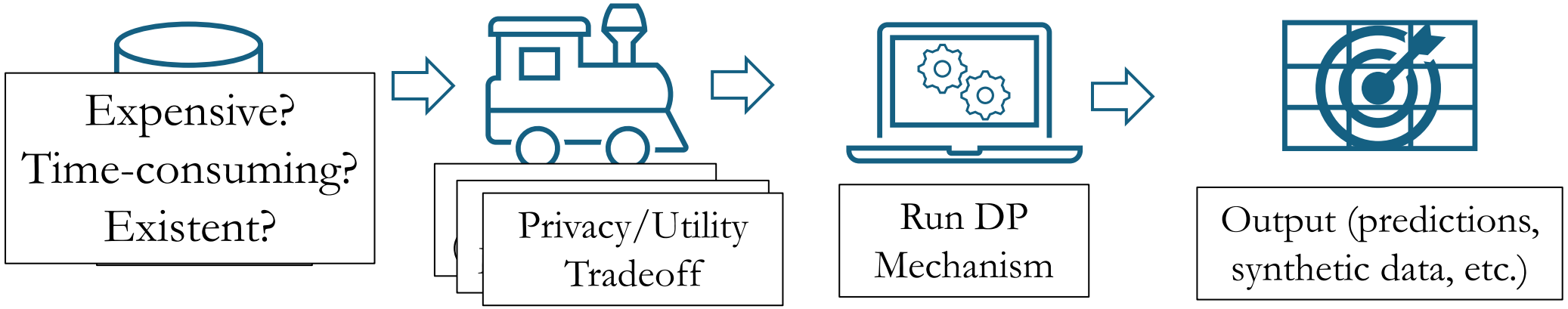
Public data: **Census 2010**

"… the simplest approach, when possible, is to do all model architecture search and hyperparameter tuning on a proxy public dataset (with a distribution similar to the private data), and only use the private training dataset to train the final DP model."

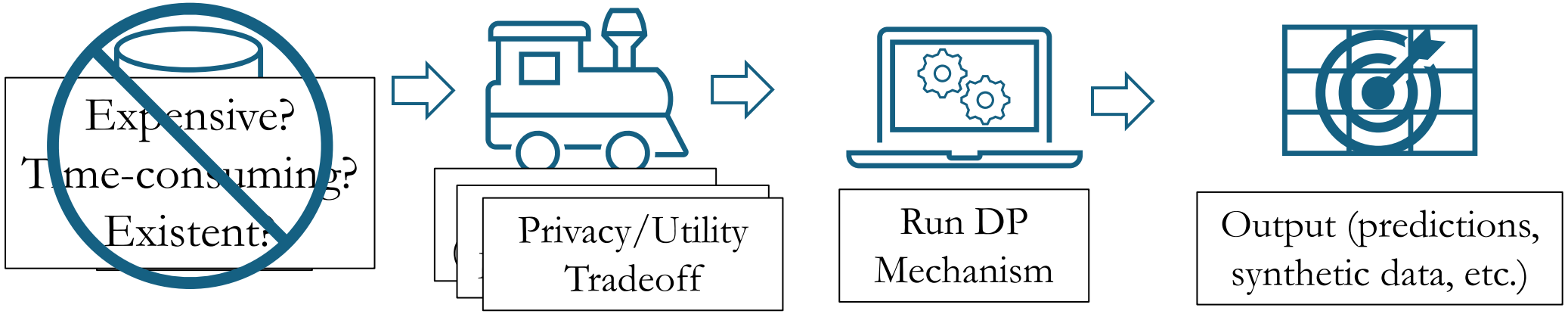
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



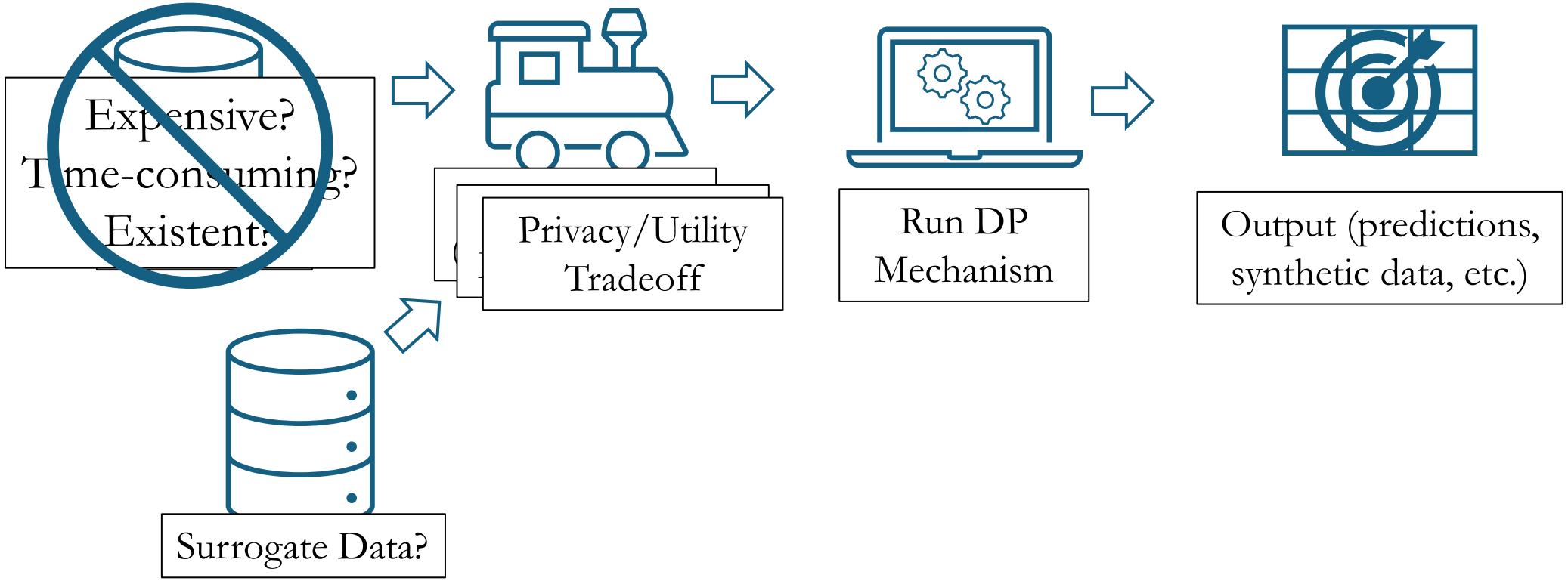
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



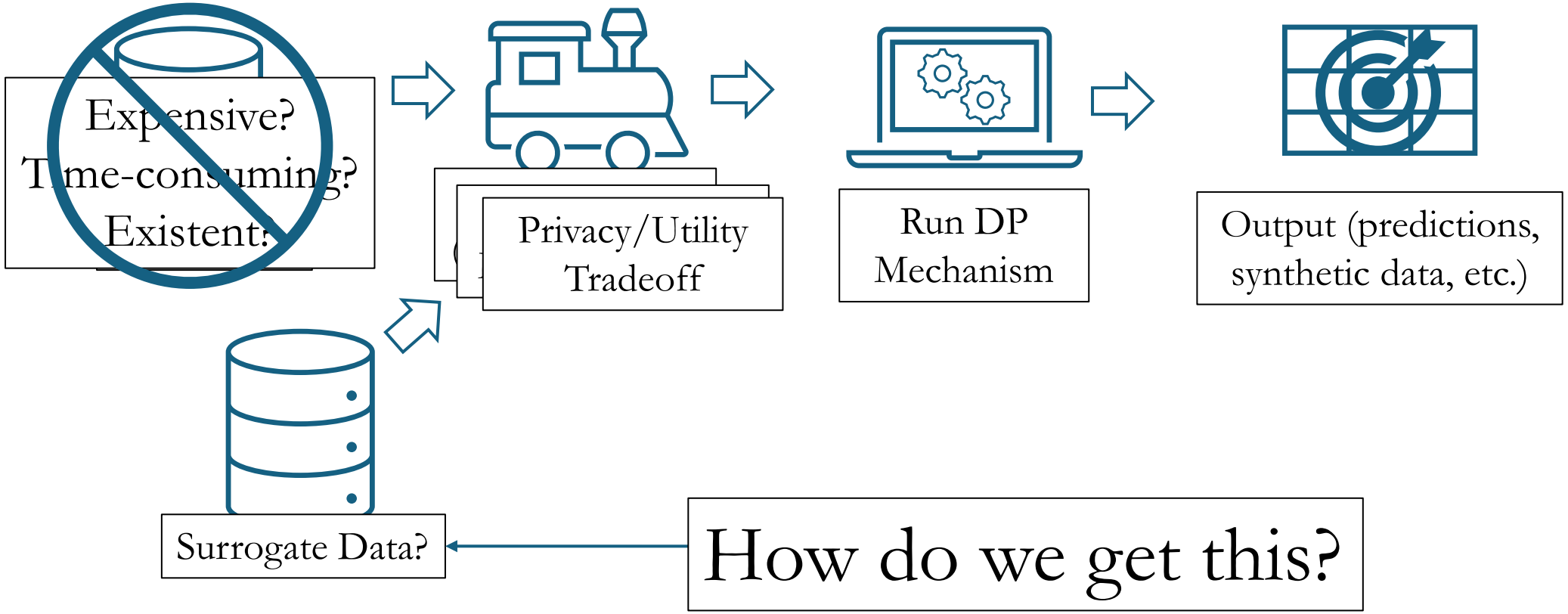
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.



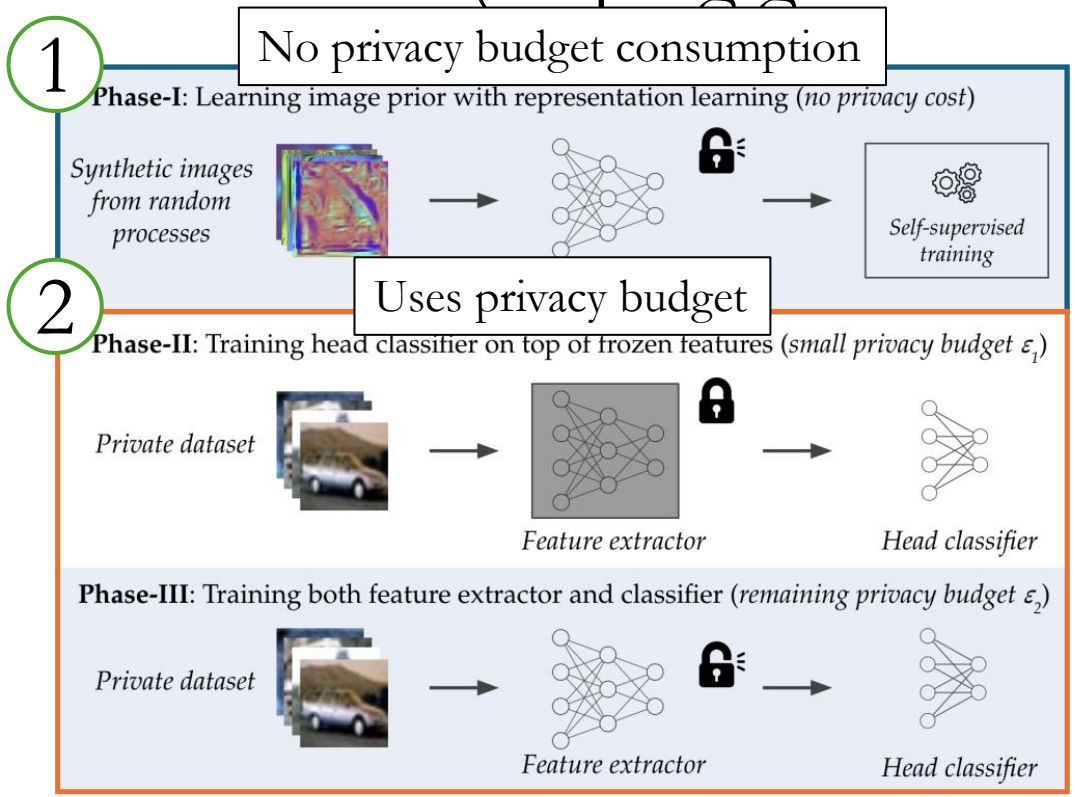
In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

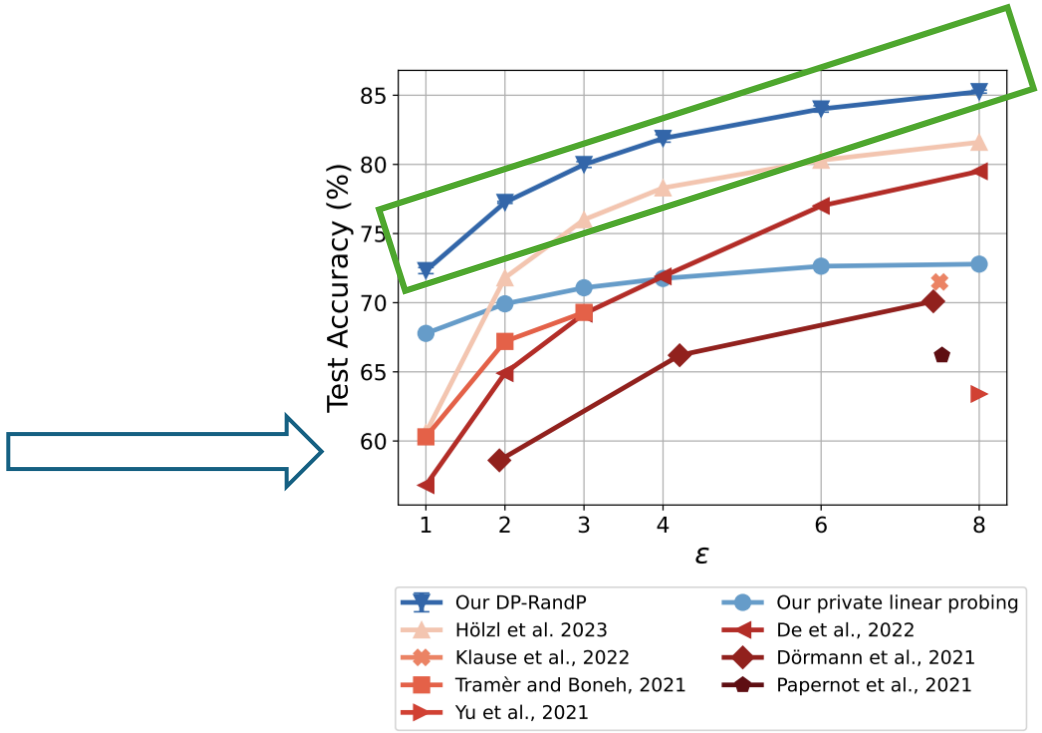


In “differential privacy world,” we are allowed to use **public** data to improve our model outputs.

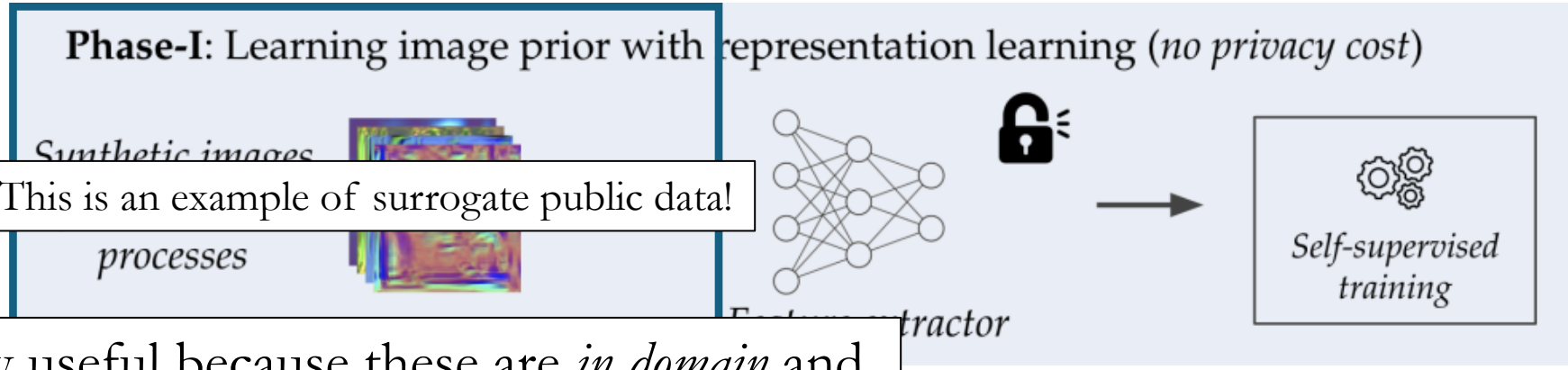


Generally, this has been **hard** for tabular data, due to domain heterogeneity…but! It's known to be useful (helping get SOTA results) for image/text data.





How do you do this for **tabular data**?



This is an example of surrogate public data!

Only useful because these are *in domain* and have interesting artifacts (like real images)

Do we have a strong prior over tabular data?

Will let us make **surrogate public data.**

Surrogate Public Data

Public Data (informal, [BKS23])
A computation taking a dataset as input does **not consume privacy loss budget** with respect to any other private, sensitive dataset

Surrogate Public Data (working def)
A dataset **generated** independently of a sensitive dataset, consuming **no privacy loss budget**, and based only **on publicly available schema or metadata**.

How useful is **surrogate public data** relative to **"traditional" public data** for **DP auxiliary tasks**?

{}


```
{
  ...
  "RELACT": {
    "description": "Main labour market activity status",
    "dtype": "int64",
    "values": {
      "1": "Employed",
      "2": "Unemployed",
      "3": "Retired",
      "4": "Student",
      "5": "Unable to work",
      "6": "Doing unpaid social work or charitable activities",
      "7": "Other inactive person"
    }
  },
  "CERTIG": {
    "description": "Degree of disability",
    "dtype": "int64",
    "values": {
      "1": "0-32%",
      "2": "33-44%",
      "3": "45-64%",
      "4": "65-74%",
      "5": "75% or more",
      "6": "Not known"
    }
  },
  "AUDI_7_1": {
    "description": "Has significant difficulty hearing a
       ↪ conversation with several people without a hearing aid",
    "dtype": "int64",
    "values": {
      "1": "Yes",
      "2": "No"
    }
  },
  ...
}
```


r/ai

NYU

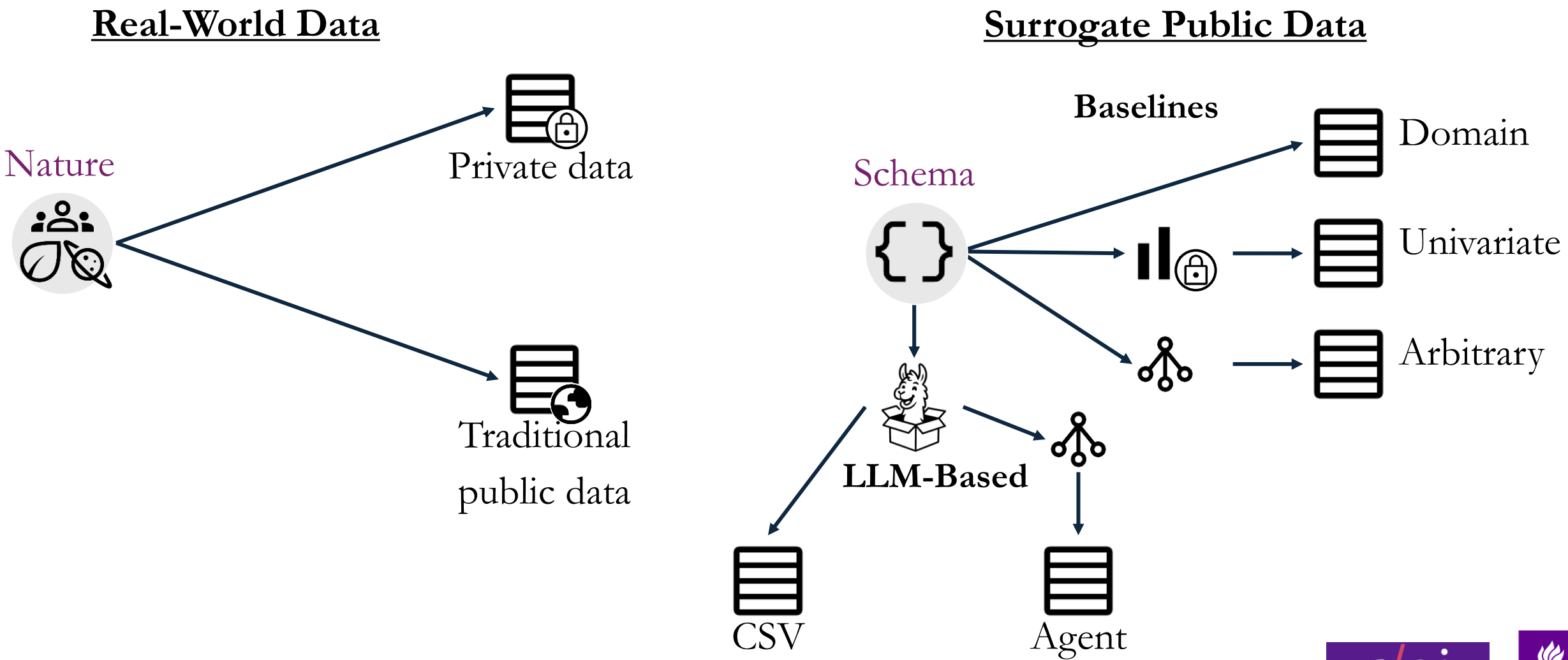
Our Contributions

Design an **empirical evaluation framework** of surrogate public data for DP auxiliary tasks.

Build an **agent-based generation method (LLMs!)** for surrogate public data.

Characterize the **tasks and settings where surrogate public data can be useful** for DP auxiliary tasks.

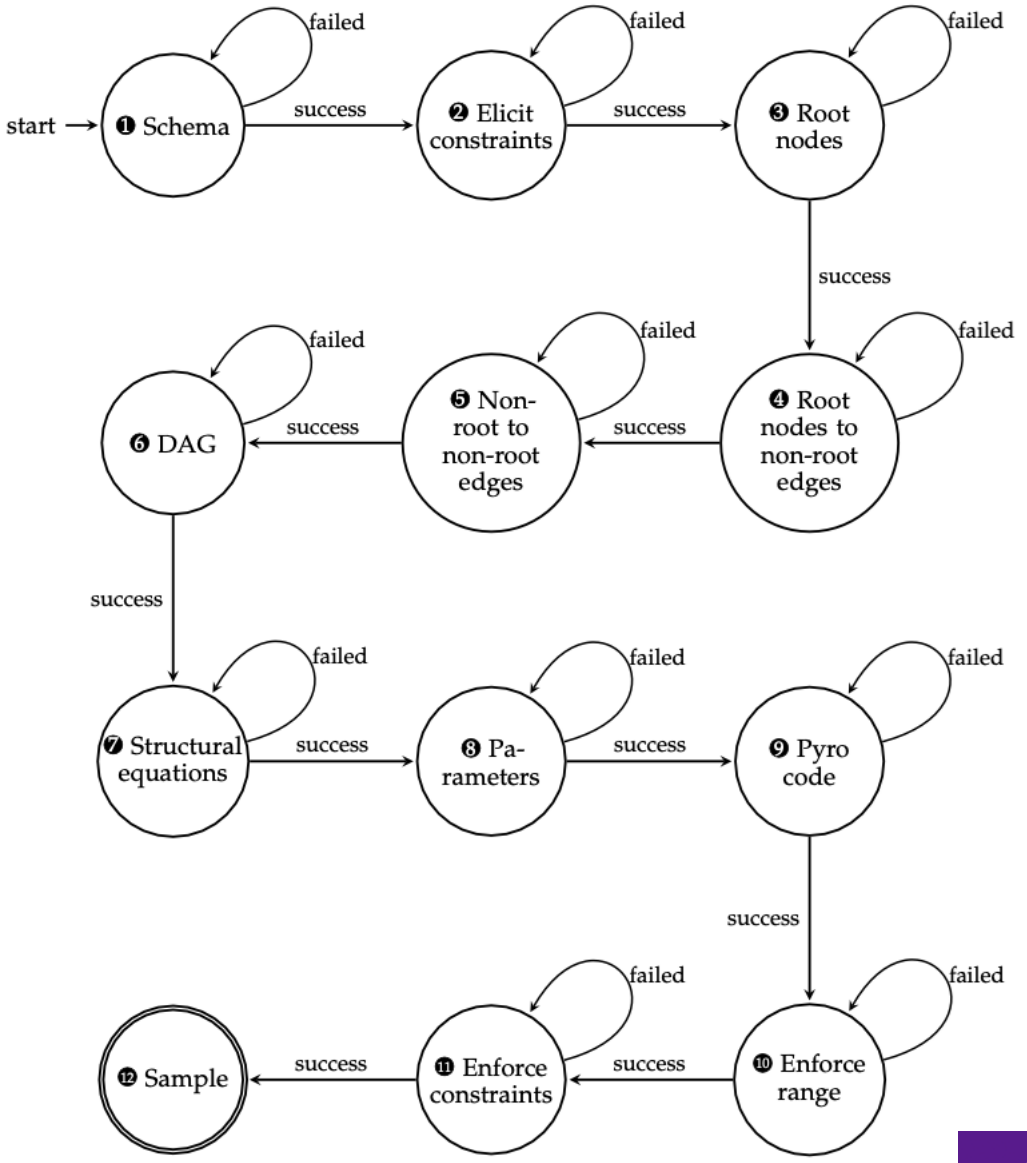
Types of Surrogate Public Data



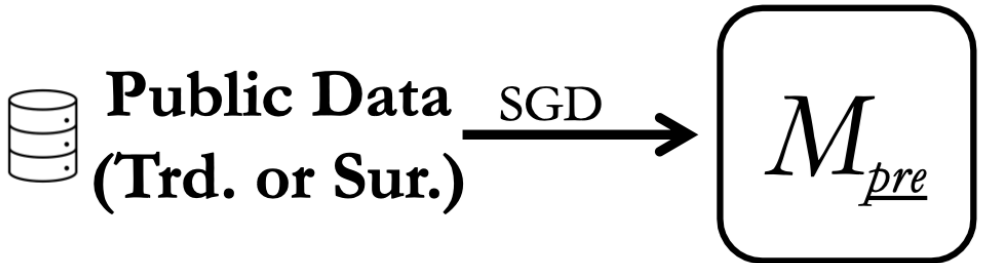
Surrogate Data: CSV

```
System: You are an expert in {domain} who generates synthetic data
    ↪ that closely mirrors real-world {domain} data. Your goal is
    ↪ to create data that would be indistinguishable from real {
    ↪ domain} records.

Follow exactly these rules:
1. Only output the CSV data with no additional text or
    ↪ explanations
2. Always include a header row matching the schema exactly
3. Strictly adhere to the provided schema's data types and
    ↪ possible values for all fields
4. Use comma as the separator
5. Ensure all values and relationships between fields are
    ↪ realistic and statistically plausible
6. Generate diverse data while maintaining real-world patterns and
    ↪ constraints
7. Include occasional edge cases at realistic frequencies

User: Generate {num_rows} rows of data with these fields:

{schema}
```

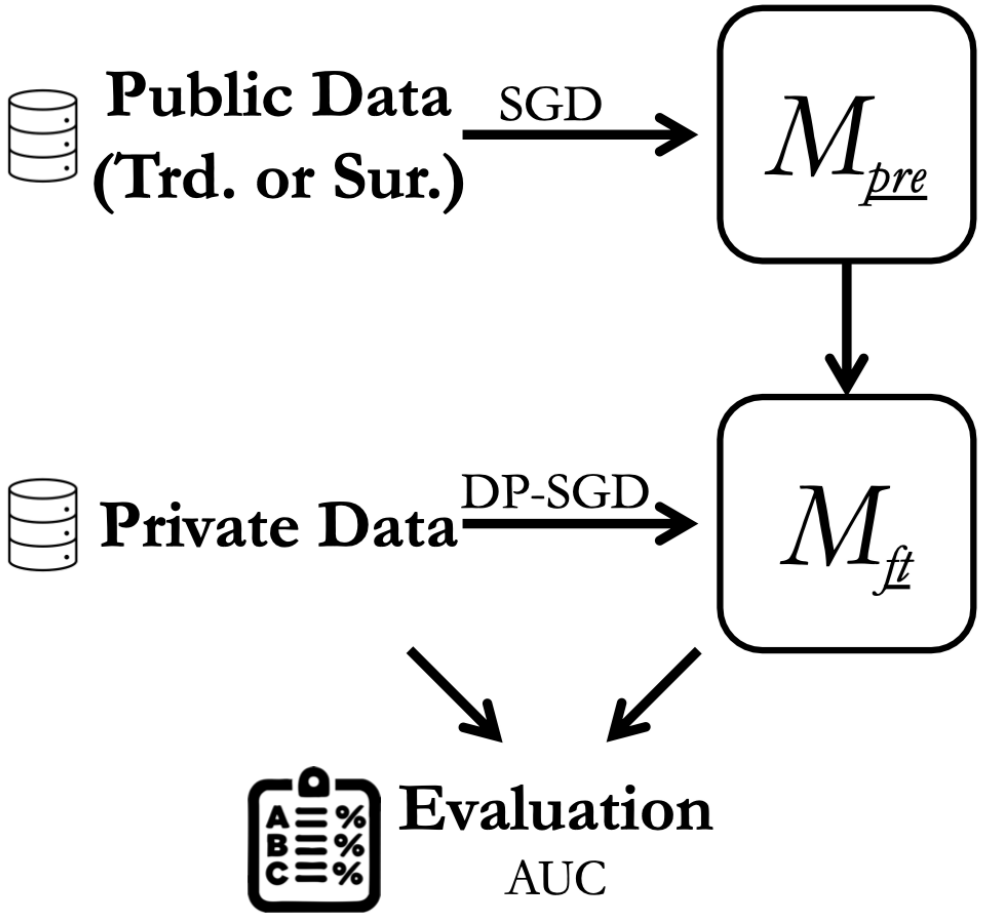
Surrogate Data: Agent



Example DP Auxiliary Task: Pretraining Classifier

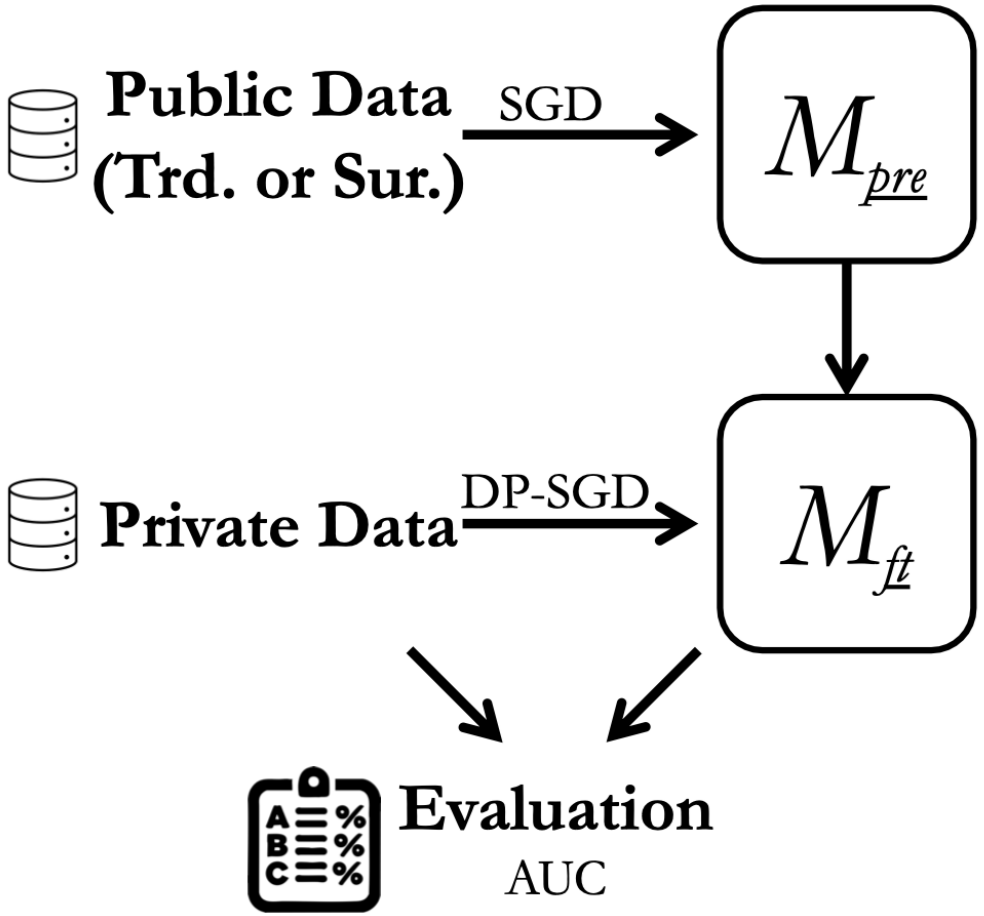


Example DP Auxiliary Task: Pretraining Classifier



Revisiting Deep Learning Models for Tabular Data. Gorishniy et al. 2022

Example DP Auxiliary Task: Pretraining Classifier



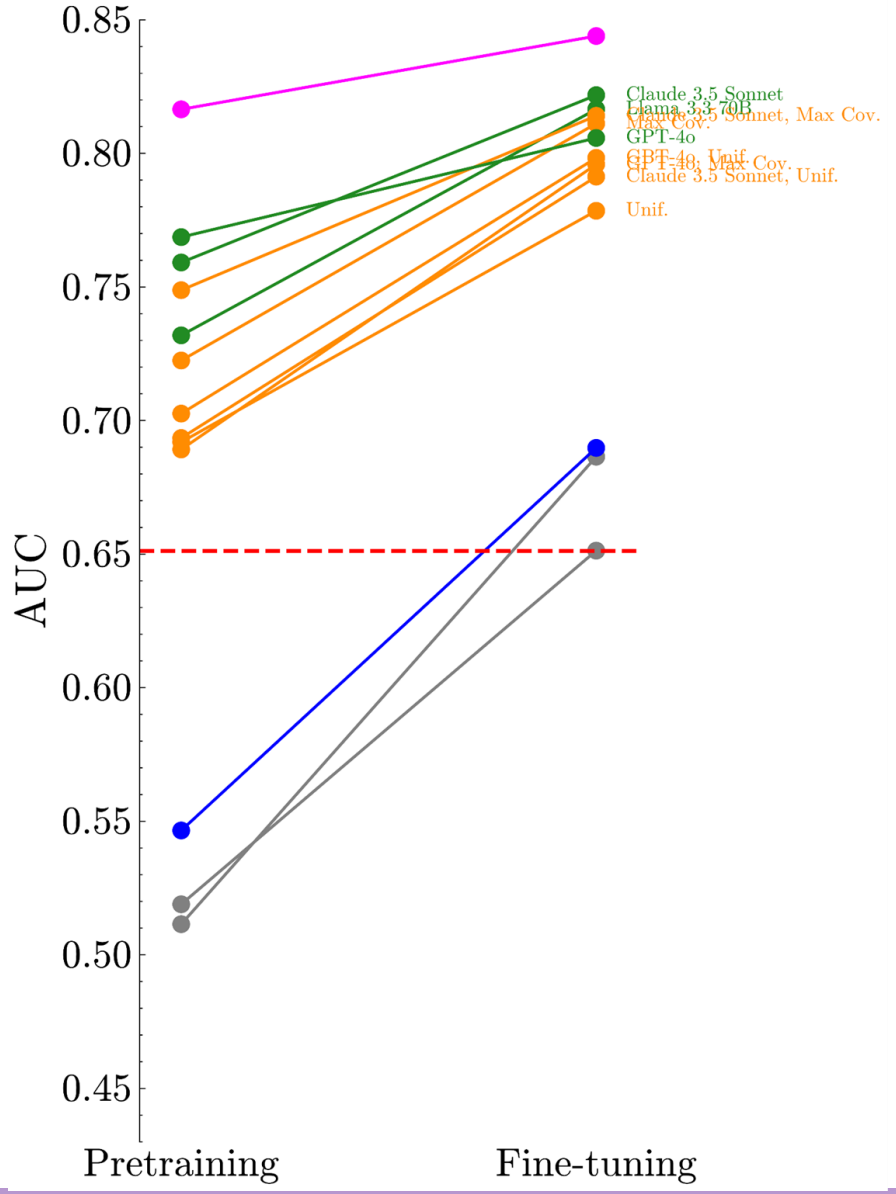
EDAD
Disability survey
11 categorical columns
2.5K records

Classifier
(DP-SGD) FTTransformer

Revisiting Deep Learning Models for Tabular Data. Gorishniy et al. 2022

Results: EDAD
$\varepsilon = 1$

- Without pretraining
- Public
- Baseline
- Arbitrary
- CSV
- Agent



Notes

Pretraining performance varies widely.

DP-SGD finetuning improves by *about* the same (consistent slope).

Starting point matters a lot!

Caveats…

Caveats…

1. Is this LLM generated / queried data truly “Public”? What does that mean?

Caveats…

1

Is this LLM generated / queried data truly "Public"? What does that mean?

See Tramer et al. 2023 + we discuss a paradigm shift in what privacy means with prevalent+powerful LLMs

Could be foundation models with DP going forward? Vault Gemma as the first and prominent example!

Caveats…

1. Is this LLM generated / queried data truly "Public"? What does that mean?

See Tramer et al. 2023 + we discuss a paradigm shift in what privacy means with prevalent+powerful LLMs

2. Other places, LLM generated data less useful, and the results less impressive.

Caveats…

1. Is this LLM generated / queried data truly "Public"? What does that mean?

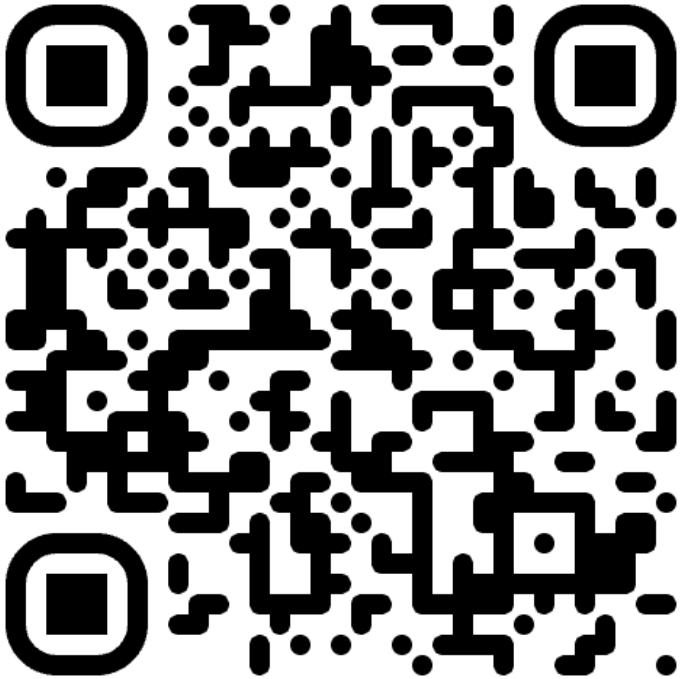
See Tramer et al. 2023 + we discuss a paradigm shift in what privacy means with prevalent+powerful LLMs

2. Other places, LLM generated data less useful, and the results less impressive.

Probably somewhat useful for hyperparameter tuning for DP synthetic data, less useful in assessing Priv/Util tradeoff

…

Paper!



Thanks!