Privacy-Preserving Systems (a.k.a., Private Systems)

CU Graduate Seminar

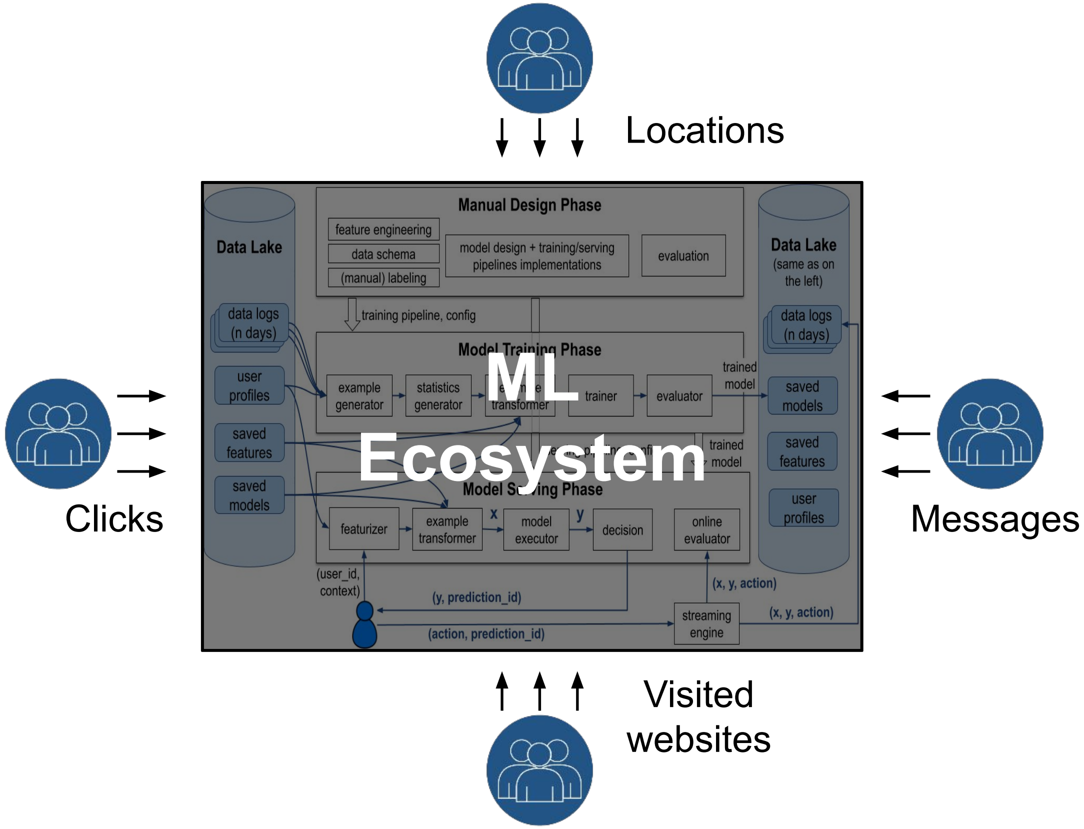
Instructor: Roxana Geambasu

Privacy Attacks

Running Example:
Anatomy of Modern ML Ecosystems

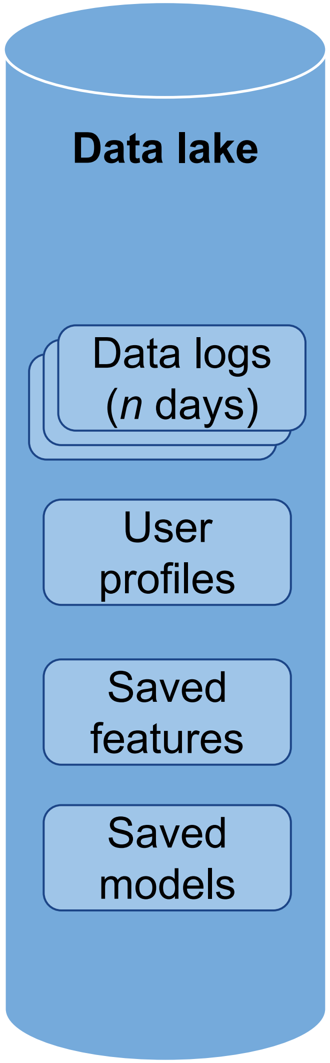
Complex and Data Hungry

- Companies collect immense user data and process it with sophisticated pipelines.
- These pipelines form complex ML ecosystems that raise serious privacy risks for the users.
- To understand the risks, it's useful to inspect the structure of ML ecosystems.



Starting Points

- Public papers of ML production platforms at two large companies:
  1. Google’s TensorFlow Extended
  2. Meta’s FBLearner



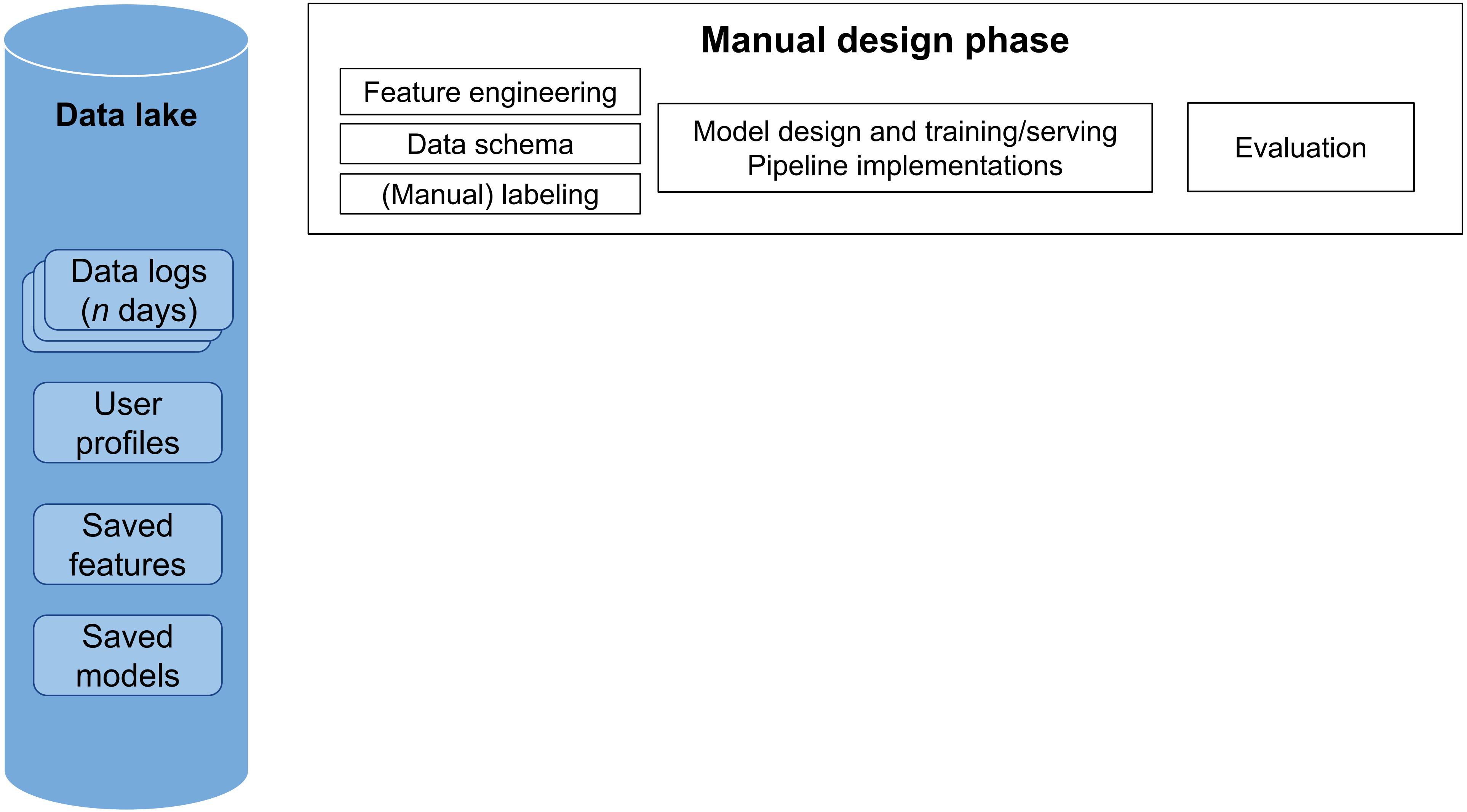
Data lake

Data logs (*n* days)

User profiles

Saved features

Saved models



Data lake

Data logs (*n* days)

User profiles

Saved features

Saved models

Manual design phase

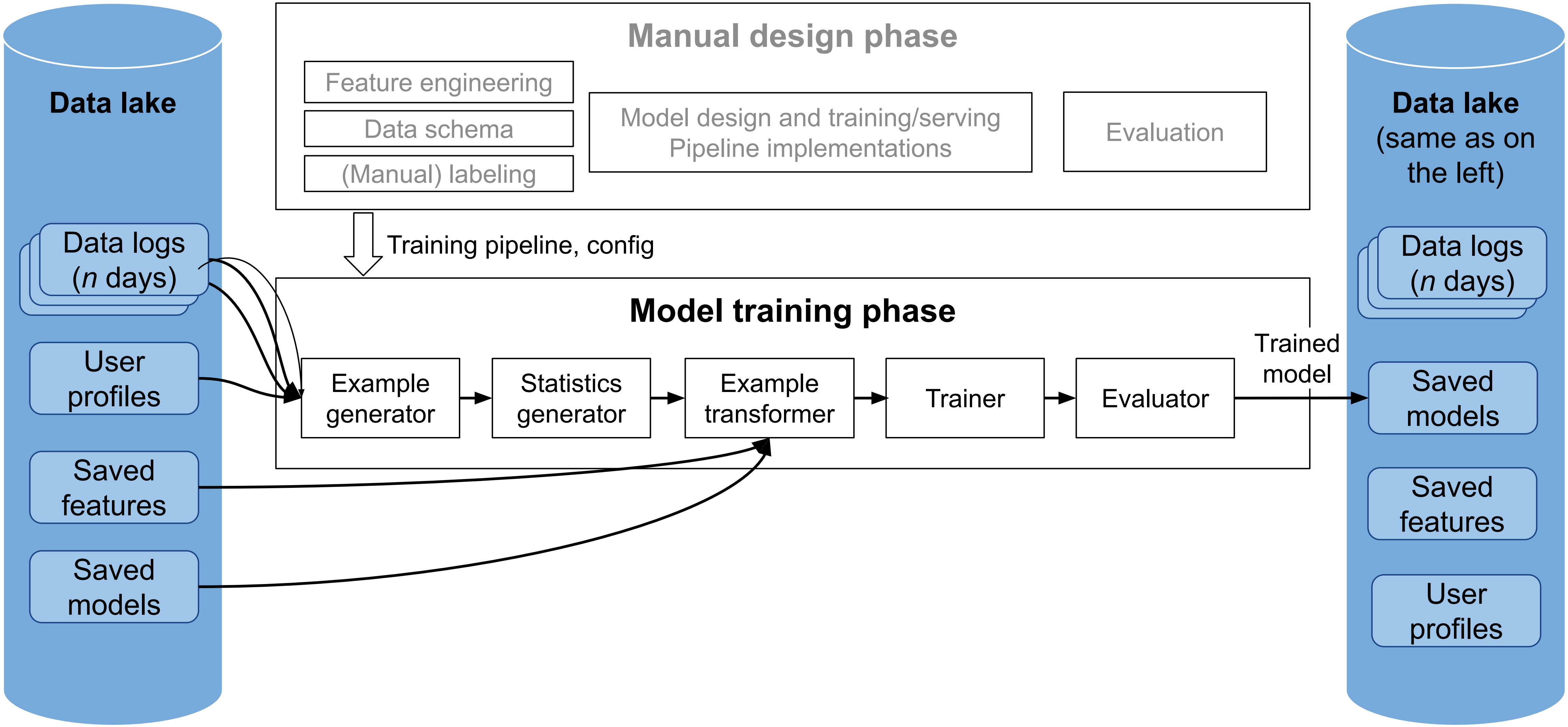
Feature engineering

Data schema

(Manual) labeling

Model design and training/serving Pipeline implementations

Evaluation



Data lake

Data logs (*n* days)

User profiles

Saved features

Saved models

Manual design phase

Feature engineering

Data schema

(Manual) labeling

Model design and training/serving Pipeline implementations

Evaluation

Training pipeline, config

Model training phase

Example generator

Statistics generator

Example transformer

Trainer

Evaluator

Trained model

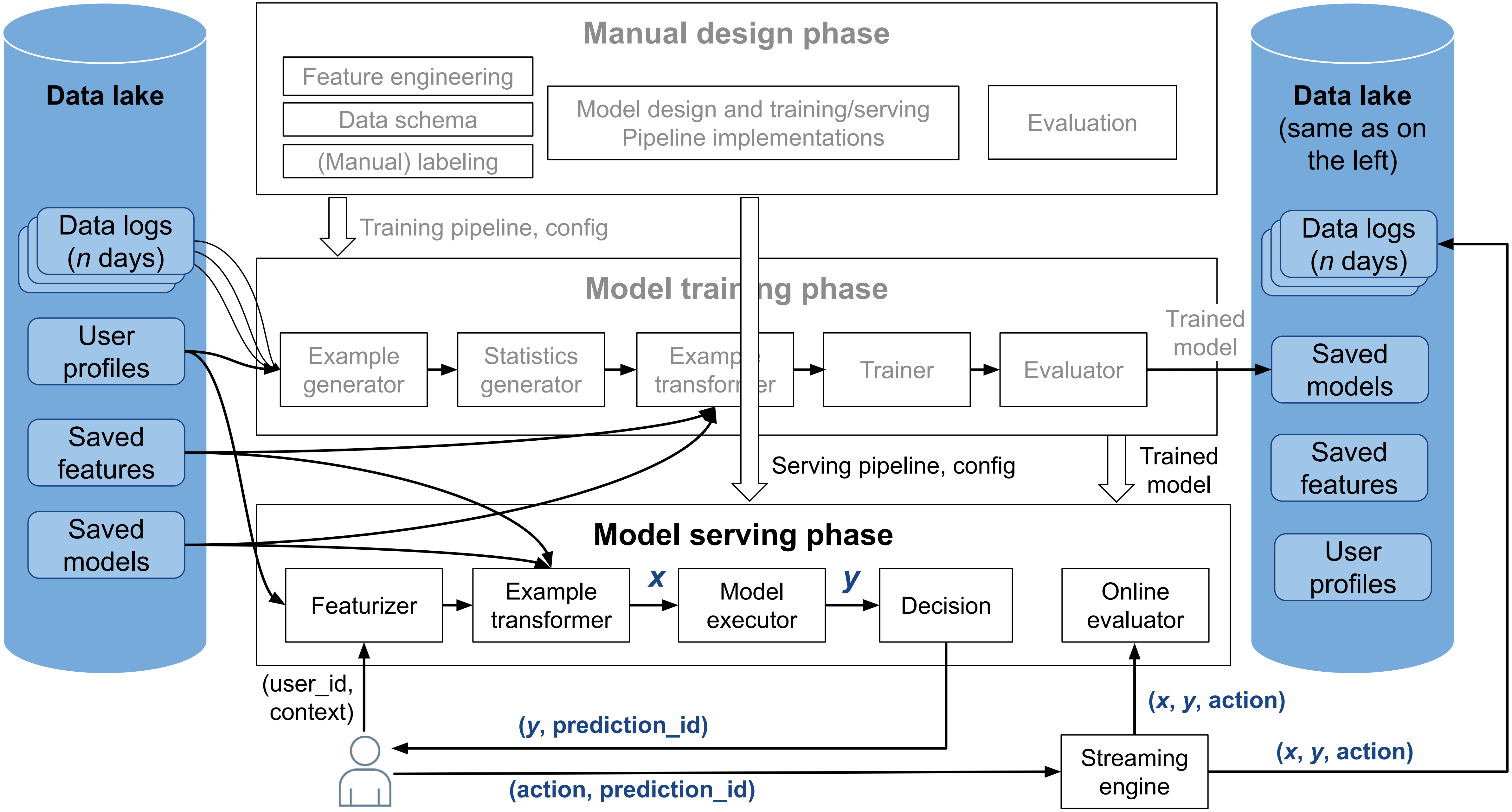
Data lake (same as on the left)

Data logs (*n* days)

Saved models

Saved features

User profiles



Data lake

Data logs
(*n* days)

User profiles

Saved features

Saved models

Manual design phase

Feature engineering

Data schema

(Manual) labeling

Model design and training/serving
Pipeline implementations

Evaluation

Training pipeline, config

Model training phase

Example generator

Statistics generator

Example transformer

Trainer

Evaluator

Trained model

Serving pipeline, config

Trained model

Model serving phase

Featurizer

Example transformer

x

Model executor

y

Decision

Online evaluator

(user_id, context)

(*y*, prediction_id)

(action, prediction_id)

(*x*, *y*, action)

Streaming engine

(*x*, *y*, action)

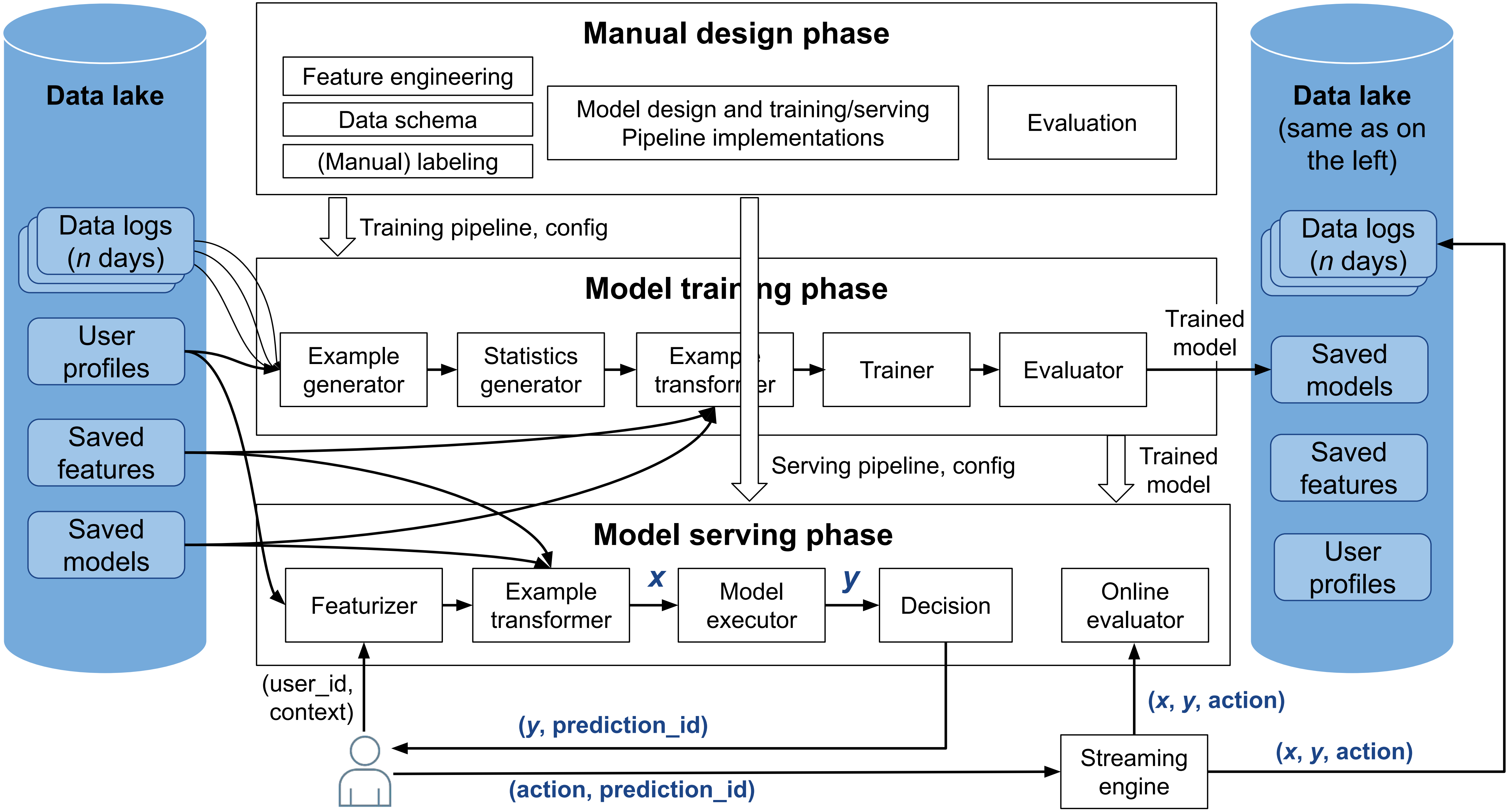
Data lake
(same as on the left)

Data logs
(*n* days)

Saved models

Saved features

User profiles



Data lake

Data logs
(*n* days)

User profiles

Saved features

Saved models

Manual design phase

Feature engineering

Data schema

(Manual) labeling

Model design and training/serving
Pipeline implementations

Evaluation

Training pipeline, config

Model trair ng phase

Example generator

Statistics generator

Examp e transforr er

Trainer

Evaluator

Trained model

Serving pipeline, config

Trained model

Model serving phase

Featurizer

Example transformer

x

Model executor

y

Decision

Online evaluator

(user_id, context)

(*y*, prediction_id)

(action, prediction_id)

Streaming engine

(*x*, *y*, action)

(*x*, *y*, action)

Data lake
(same as on the left)

Data logs
(*n* days)

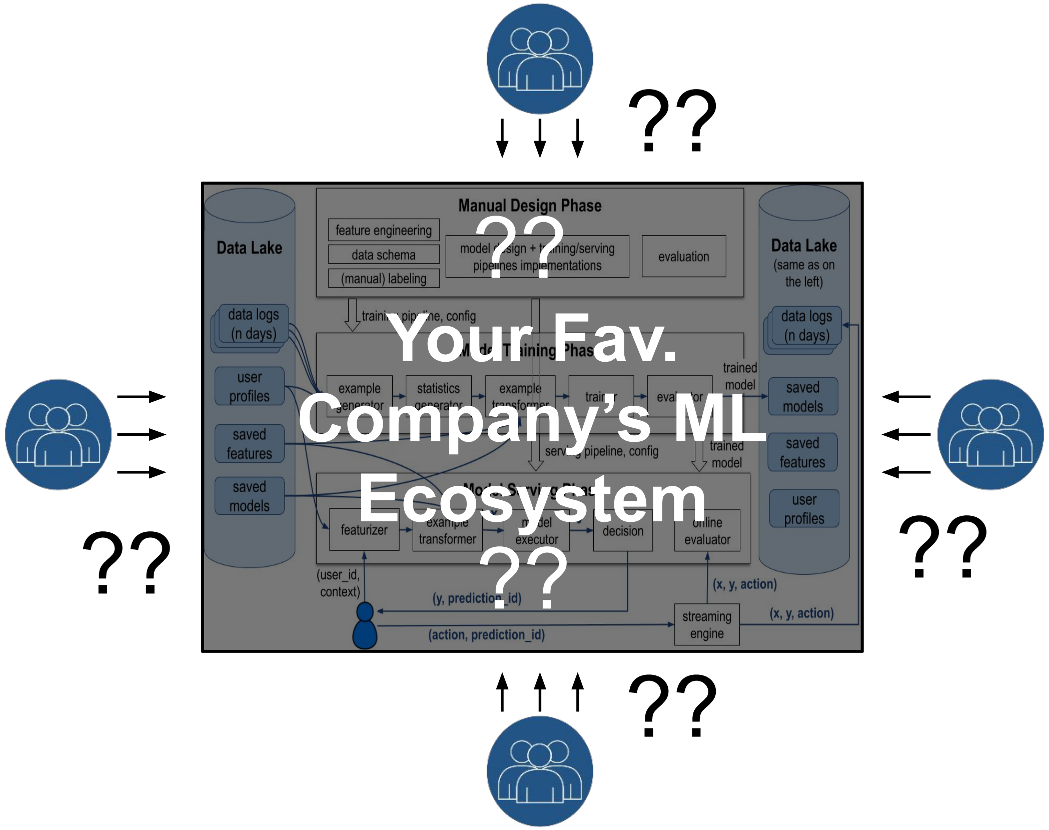
Saved models

Saved features

User profiles

Questions

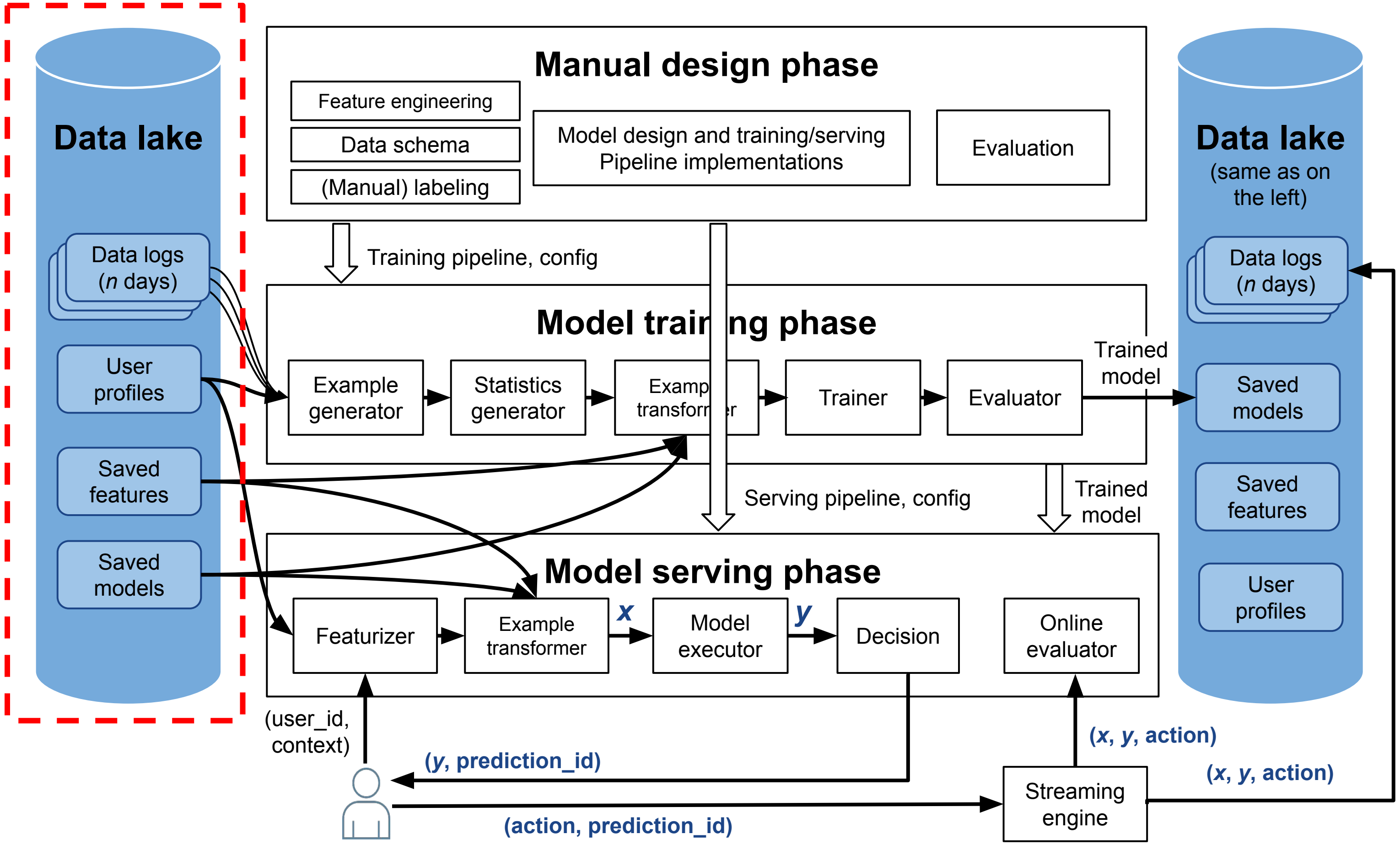
- Describe the structure of an ML ecosystem you know of.
- How does it compare to the structure described here?



Anatomy of Modern ML Ecosystems

The End

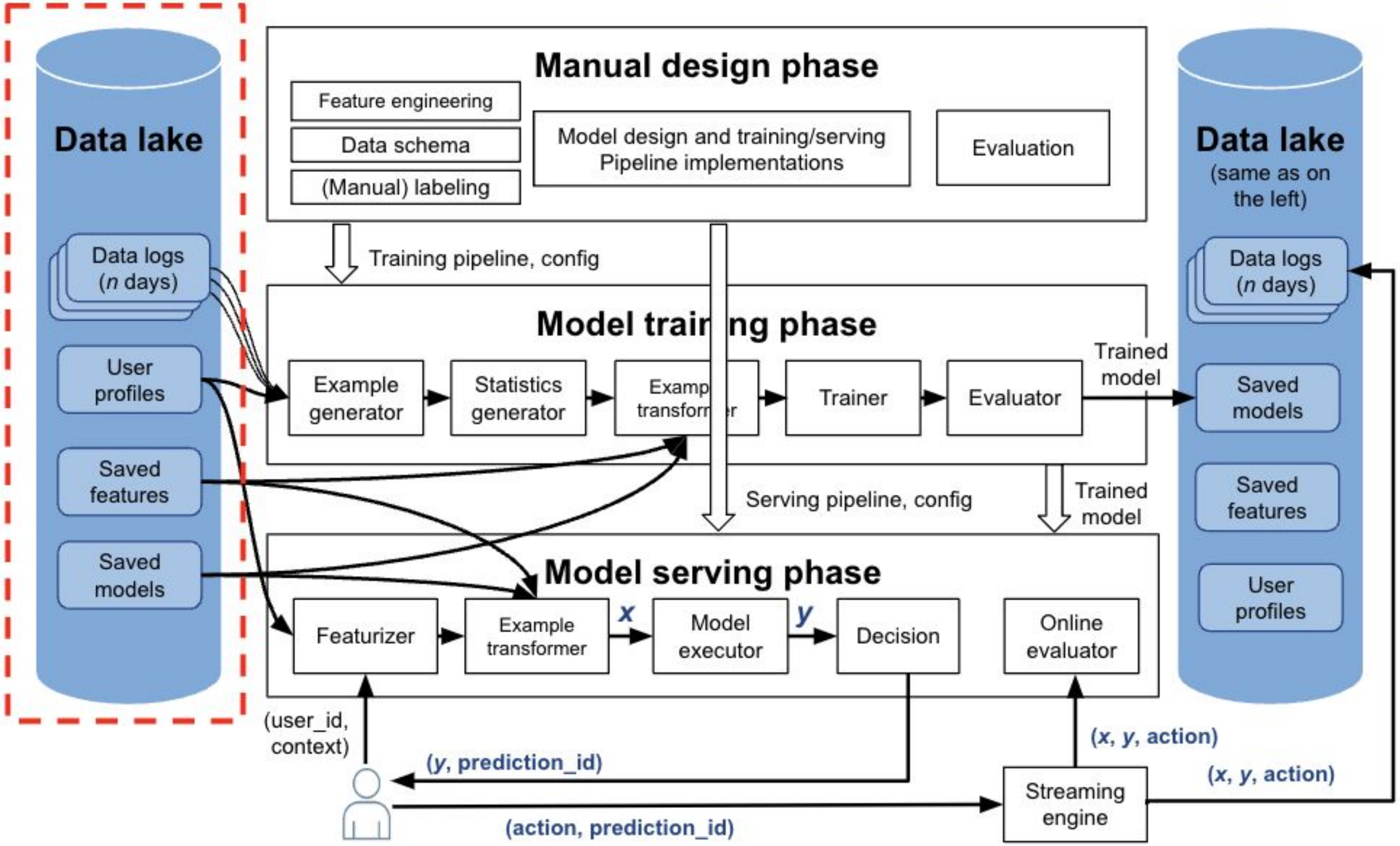
Data Exposure Risks in ML Ecosystems



Data lake
Data logs
(*n* days)
User profiles
Saved features
Saved models
Manual design phase
Feature engineering
Data schema
(Manual) labeling
Model design and training/serving Pipeline implementations
Evaluation
Training pipeline, config
Model training phase
Example generator
Statistics generator
Example transformer
Trainer
Evaluator
Trained model
Serving pipeline, config
Trained model
Model serving phase
Featurizer
Example transformer
x
Model executor
y
Decision
Online evaluator
(user_id, context)
(*y*, prediction_id)
(action, prediction_id)
(*x*, *y*, action)
Streaming engine
(*x*, *y*, action)
Data lake
(same as on the left)
Data logs
(*n* days)
Saved models
Saved features
User profiles

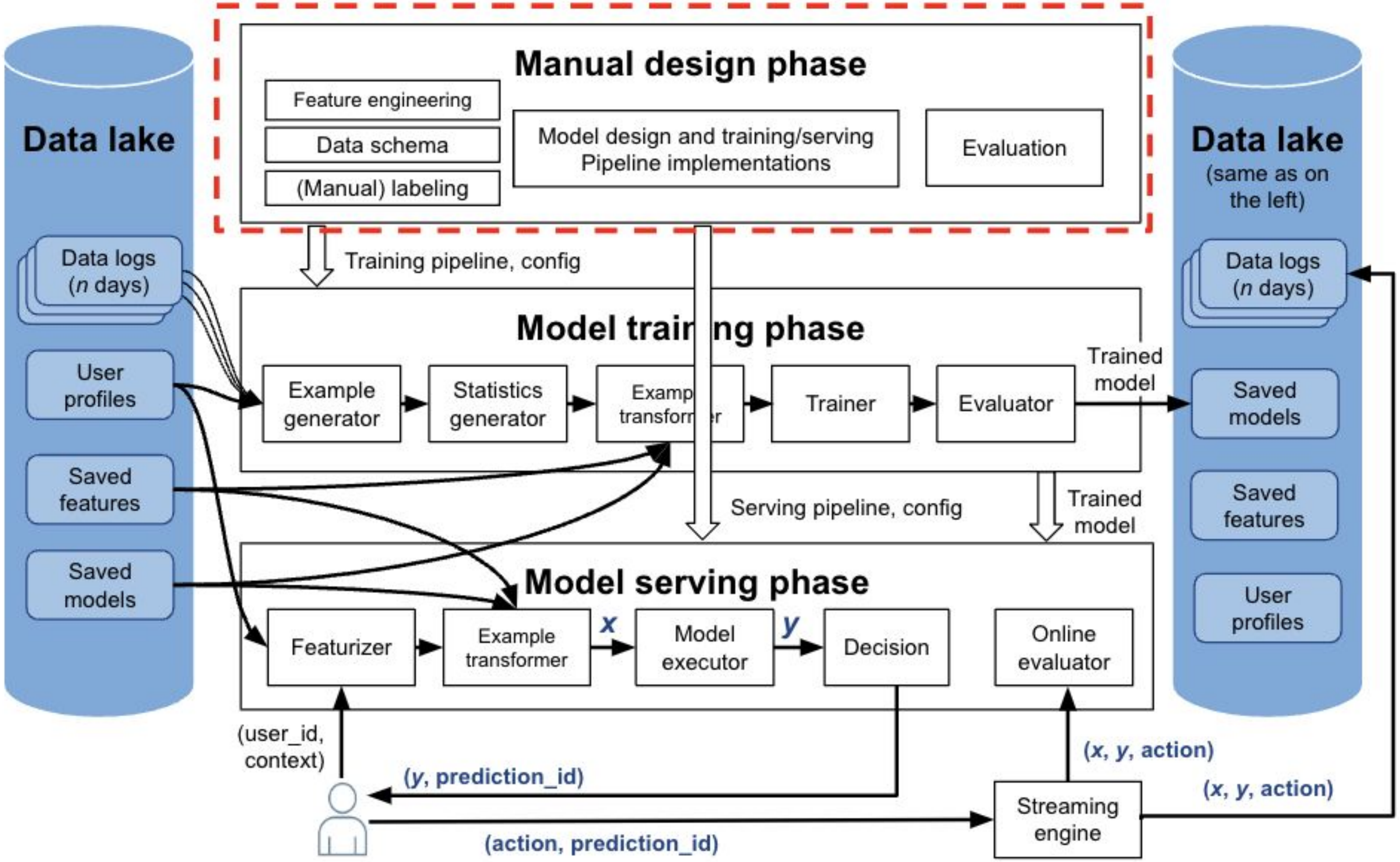
1. Exposure Through the Data Lake

- Data lake accumulates logs and data products (models, features).
- Encryption at rest and access control are used, but there is pressure to make logs/models/ features widely accessible.
- Wide access leads to **wide attack surface**.



2. Exposure Through the Manual Design

- Access to data logs is given to select engineers.
- Access comes with strict instructions not to abuse it, but privacy transgressions happen.
- Solutions: **Audit data accesses** to deter transgressors. Open privacy-preserving interface to data (e.g., **differential privacy**).

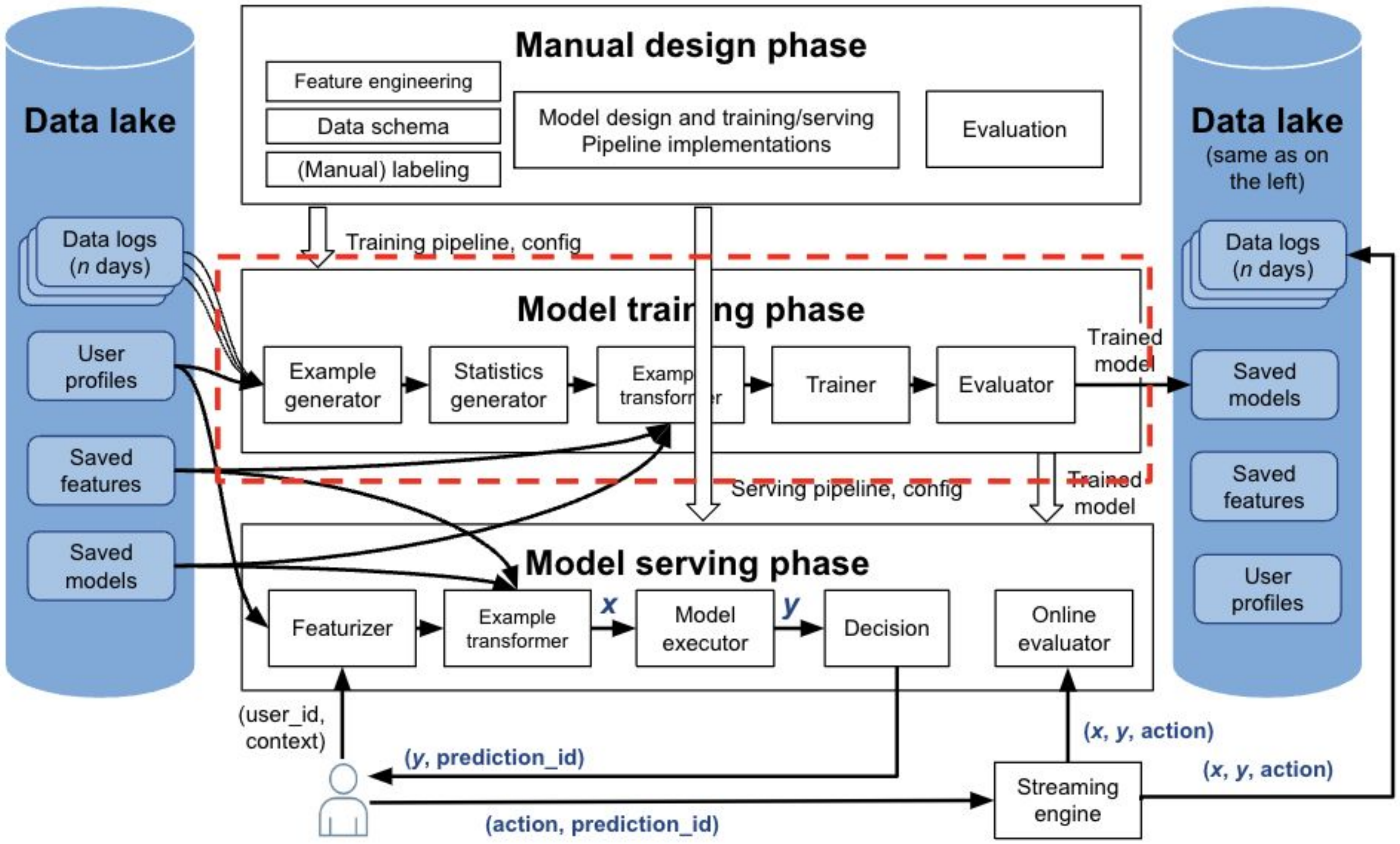


3. Exposure Through the Model Training

- Model training pipelines need access to logs, features in the data lake.

- If these processes are hacked, the hacker gains access to data.

- Auditing pipelines' data accesses is ineffective, but **advanced cryptography** (such as homomorphic encryption) can help.

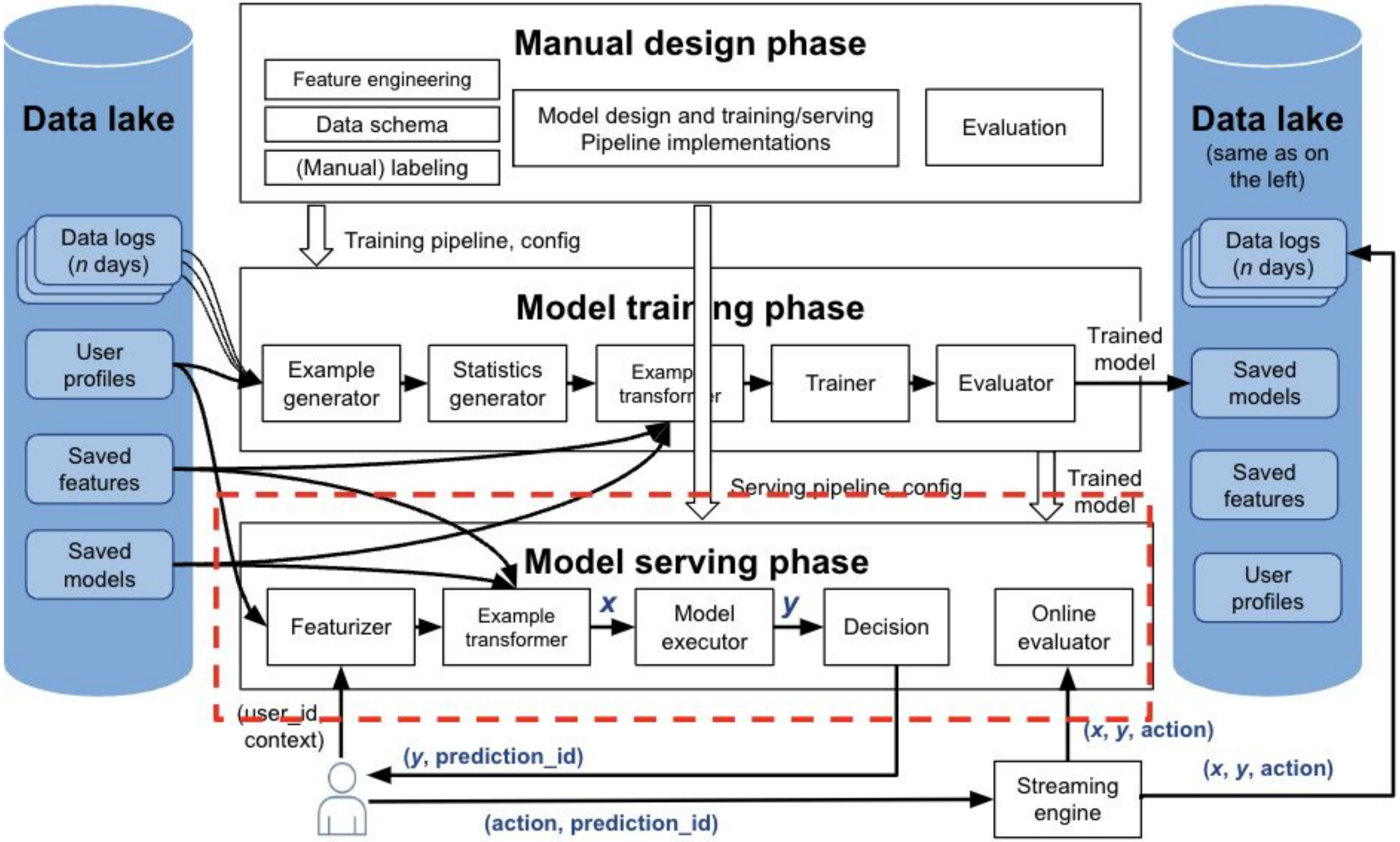


4. Exposure Through the Model Serving

- Serving pipelines distribute models to servers and mobiles all over the world.

- Model parameters and predictions can leak information from the training sets.

- Cryptography may not suffice, but **differential privacy** helps.

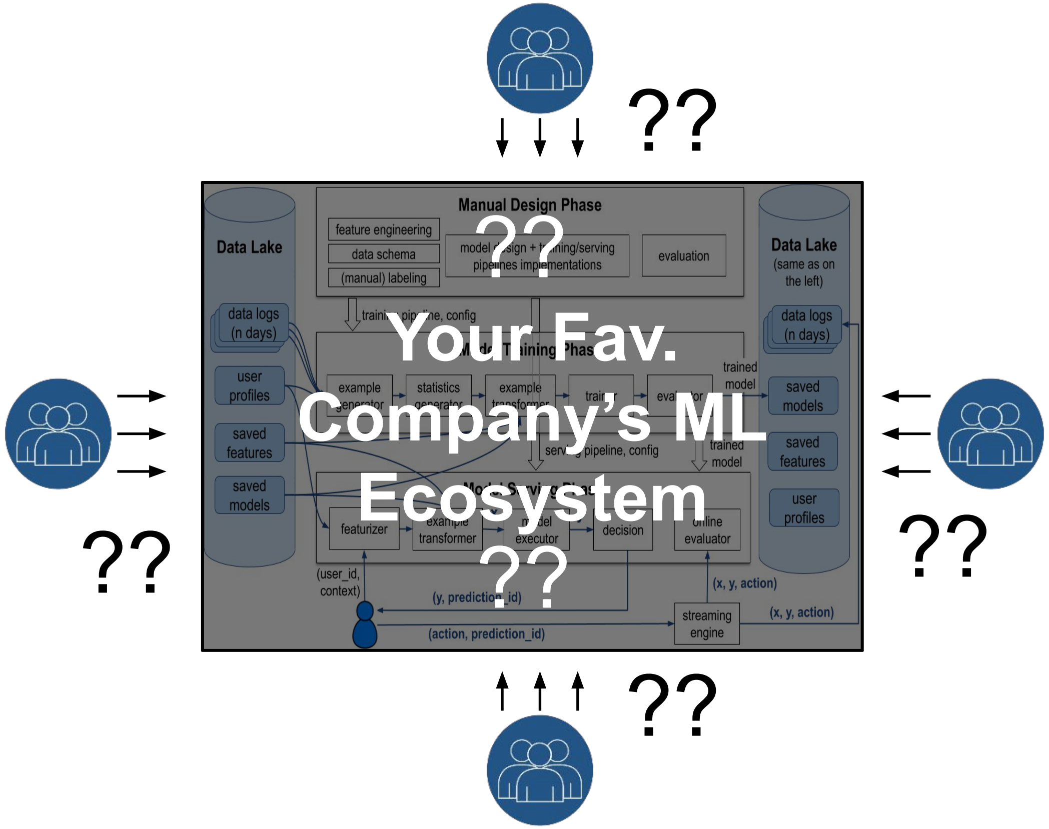


This Course

Advanced privacy technology	**Purpose**
Differential privacy	Stop training data exposure through manual design phase, model serving, and model/feature sharing within and across companies. Can also address data/model retention problem in big data.
Homomorphic encryption	Stop data exposure through the data lake, model training, and model inference.
Hardware enclaves	
Secure multi-party computation	Avoid aggregating immense data in the data lake and therefore avoid exposing data through wide-access data lakes and manual design phase.
Federated learning	
Combinations	Prevent broader sets of risks through combinations of privacy technologies

Questions

- What data exposure risks can you identify within your favorite ML ecosystem?
- How do they map onto the threats we discussed here?
- Any concerns left unaddressed?

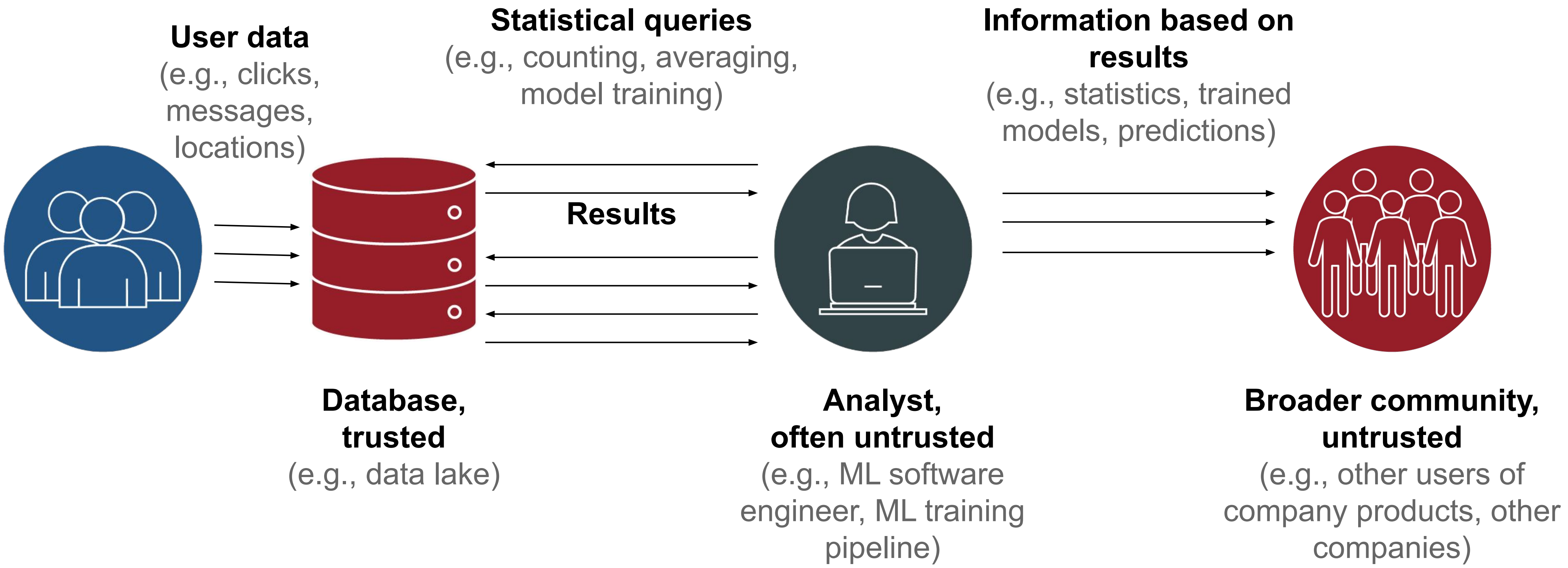


Data Exposure Risks in ML Ecosystems

The End

Differential Privacy: Threat Model and Alternative Common Practices

General Threat Model



User data
(e.g., clicks, messages, locations)

Statistical queries
(e.g., counting, averaging, model training)

Results

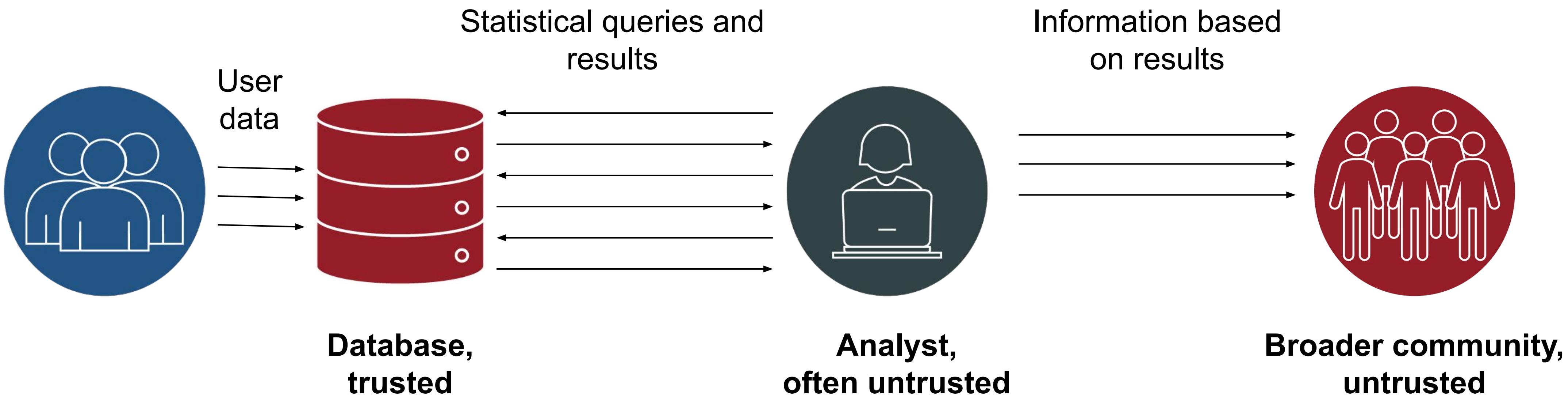
Information based on results
(e.g., statistics, trained models, predictions)

Database, trusted
(e.g., data lake)

Analyst, often untrusted
(e.g., ML software engineer, ML training pipeline)

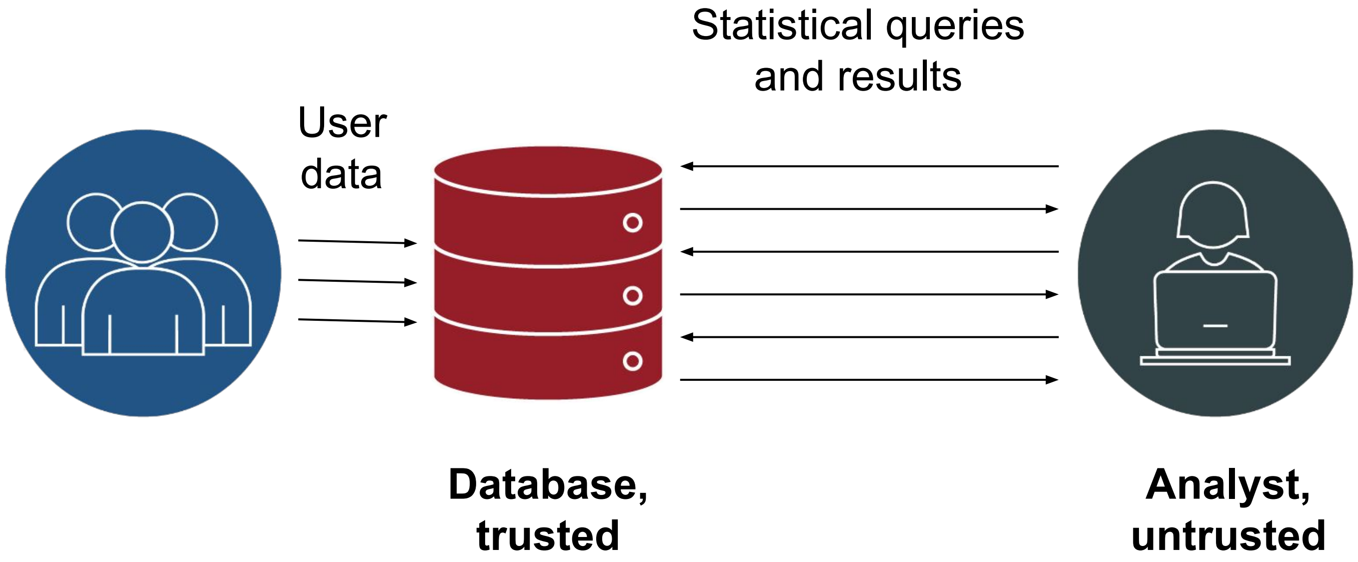
Broader community, untrusted
(e.g., other users of company products, other companies)

Generic Threat Model (cont.)

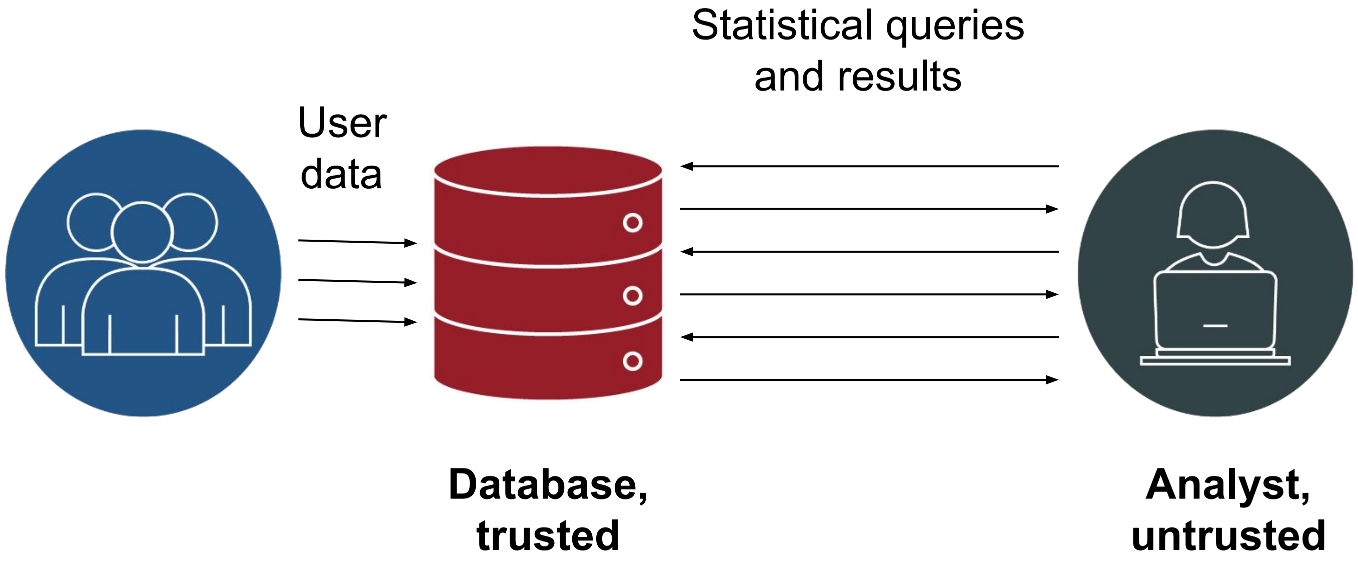

User data
Statistical queries and results
Information based on results
Database, trusted
Analyst, often untrusted
Broader community, untrusted

Goal: *allow statistical queries without increasing the privacy exposure of individuals in the database to the analyst or to the broader community*

This Lecture: Simplified Threat Model

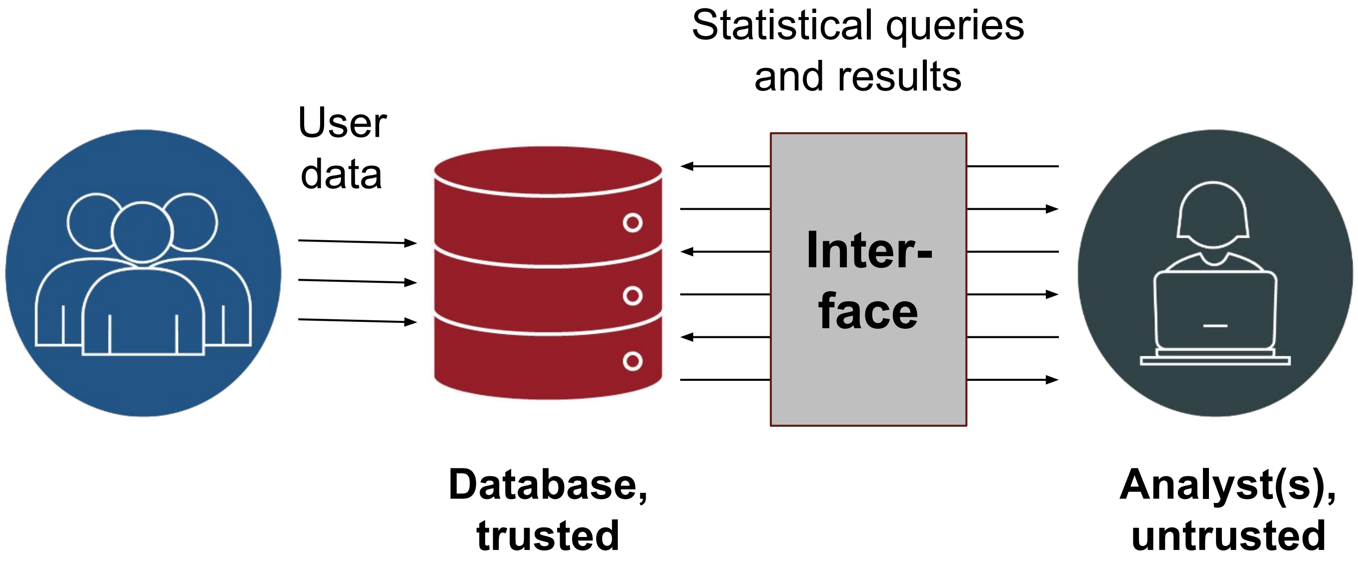


This Lecture: Simplified Threat Model



Goal: allow statistical aggregate queries without increasing the privacy exposure of individuals in the database to the analyst(s)

Approach



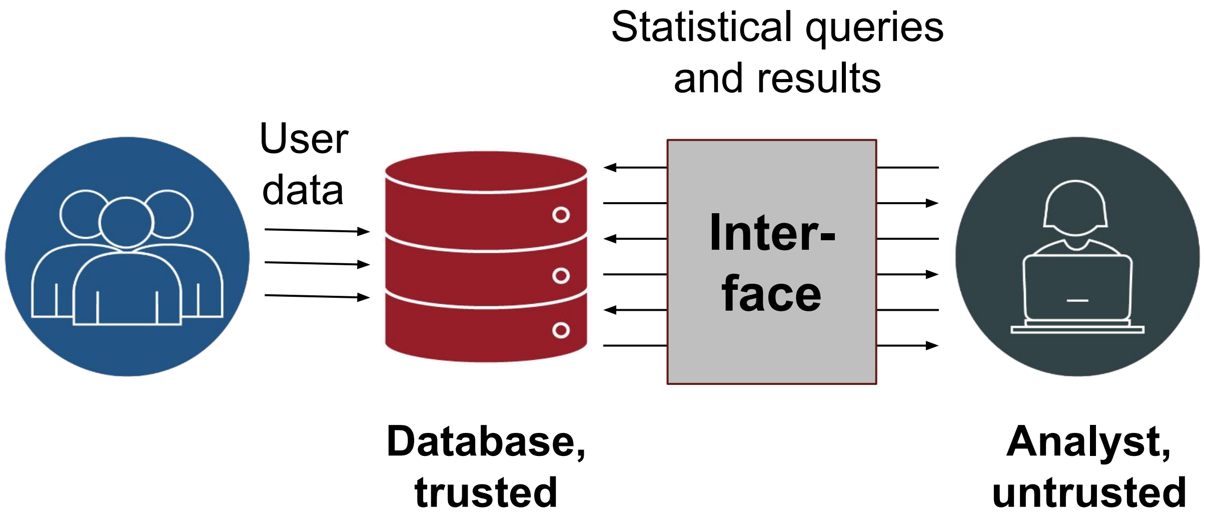
Approach: restrict analyst’s access through an interface that determines what’s “okay to release”

Common Solutions

- Two main categories:
  1. **Anonymization**
  2. **Aggregates-only**

- Both fail spectacularly with **side information** and **multiple queries**

- Whether a vulnerability is problematic depends on **context** (privacy is contextual)



What We’ll Discuss

1. Anonymization and attacks against it
2. Aggregates and attacks against them
3. Fundamental attack

Plan

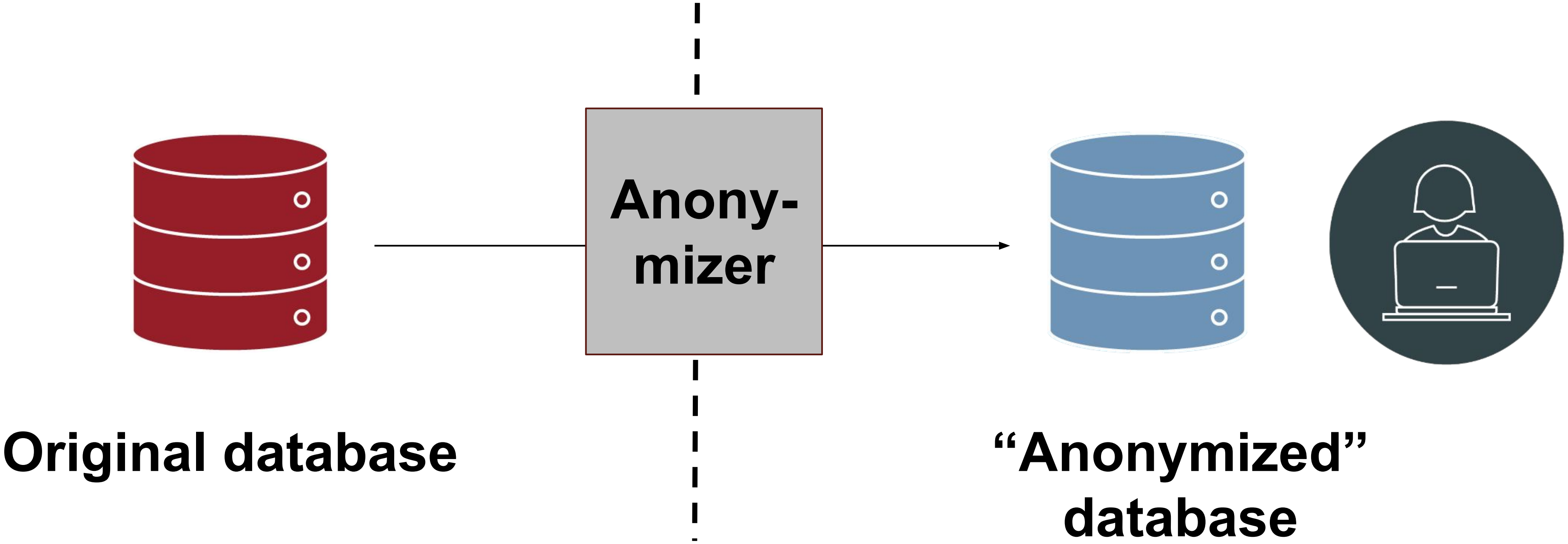
- We’ll discuss both by example, highlighting the “key ingredients” of the attacks: data distribution, side information, multiple queries
- Next time, we will discuss differential privacy, the only known privacy technology that rigorously addresses the private statistical data release problem. Works by addressing the key ingredients head-on

Differential Privacy: Threat Model and Alternative Common Practices

The End

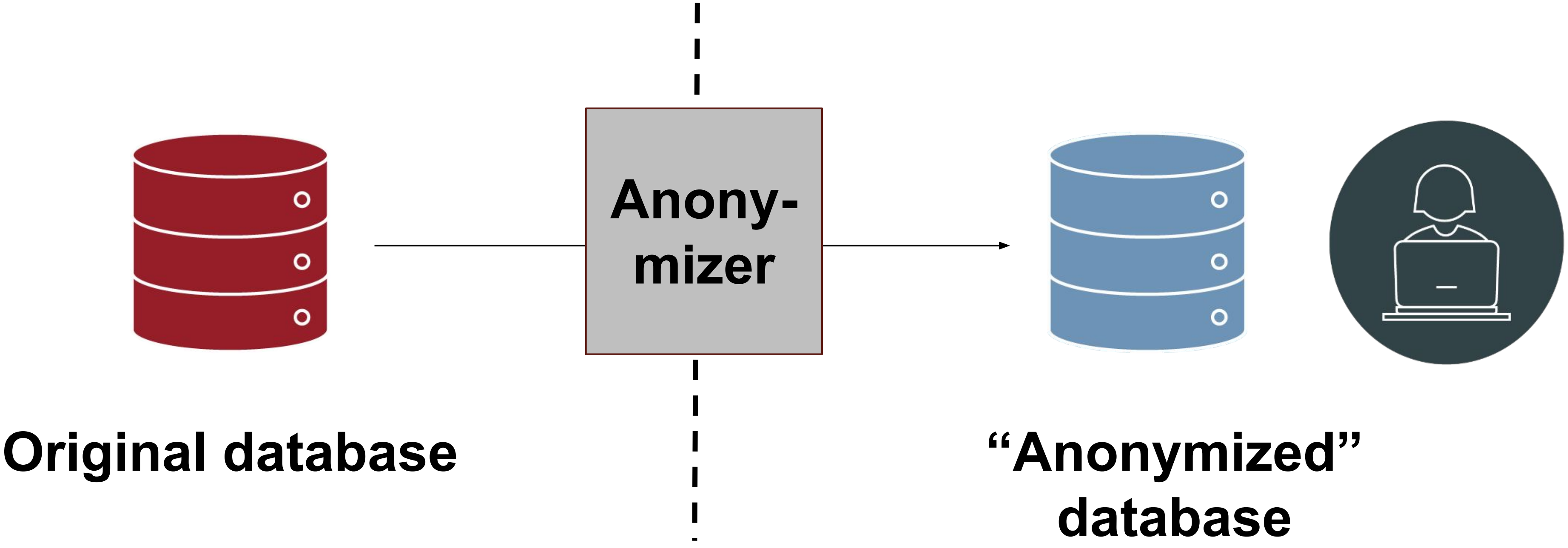
Anonymization and Attacks against It

Anonymization



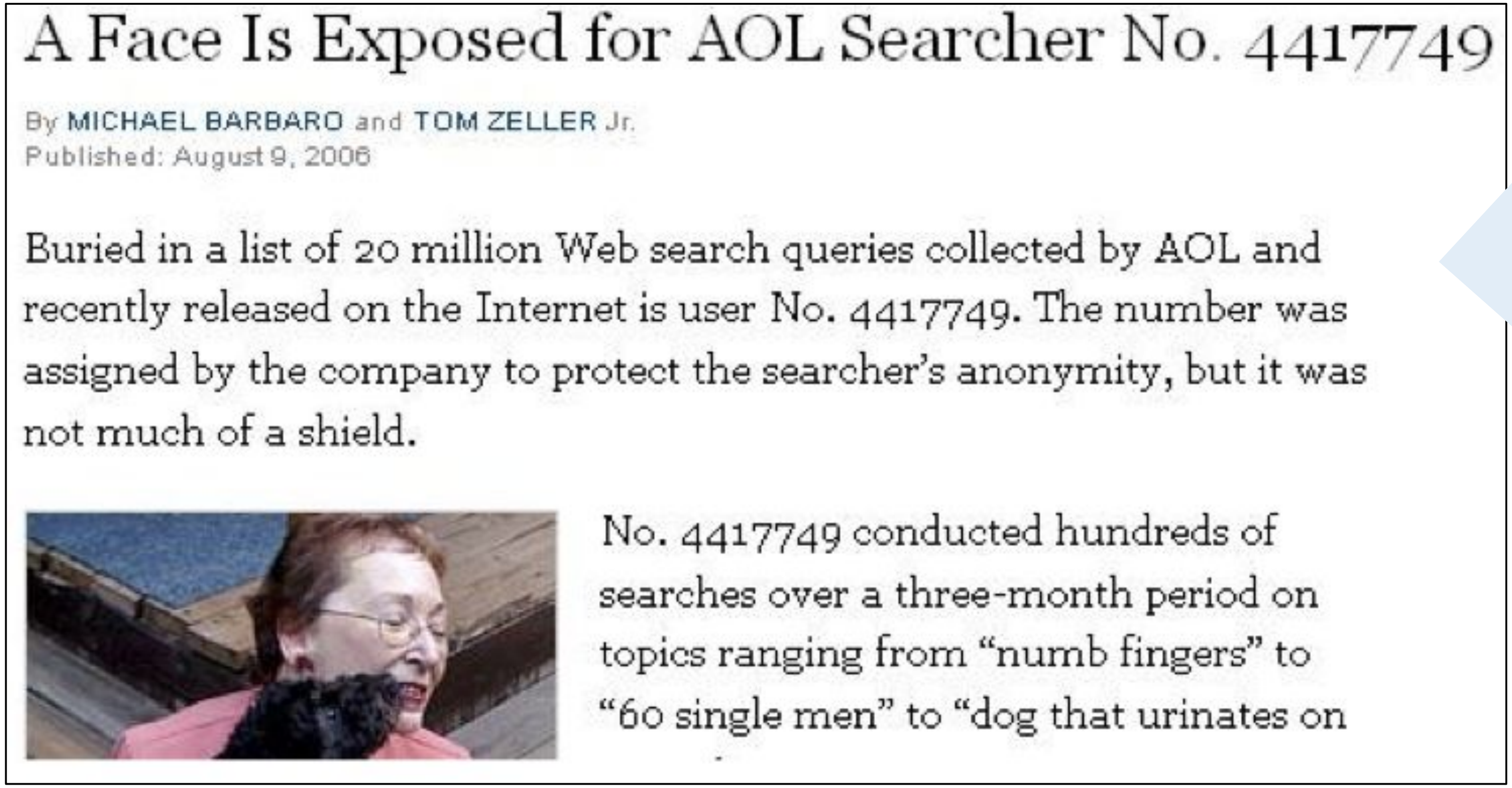
- For example: Anonymizer removes names, phone numbers, home addresses, and other “obviously” personally identifiable information (PII)
- Better-defined approaches exist, such as k-anonymity, l-diversity, …

Anonymization



Problem: “Anonymized data isn’t.” [Cynthia Dwork]

Re-Identification Attack: AOL Example



A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

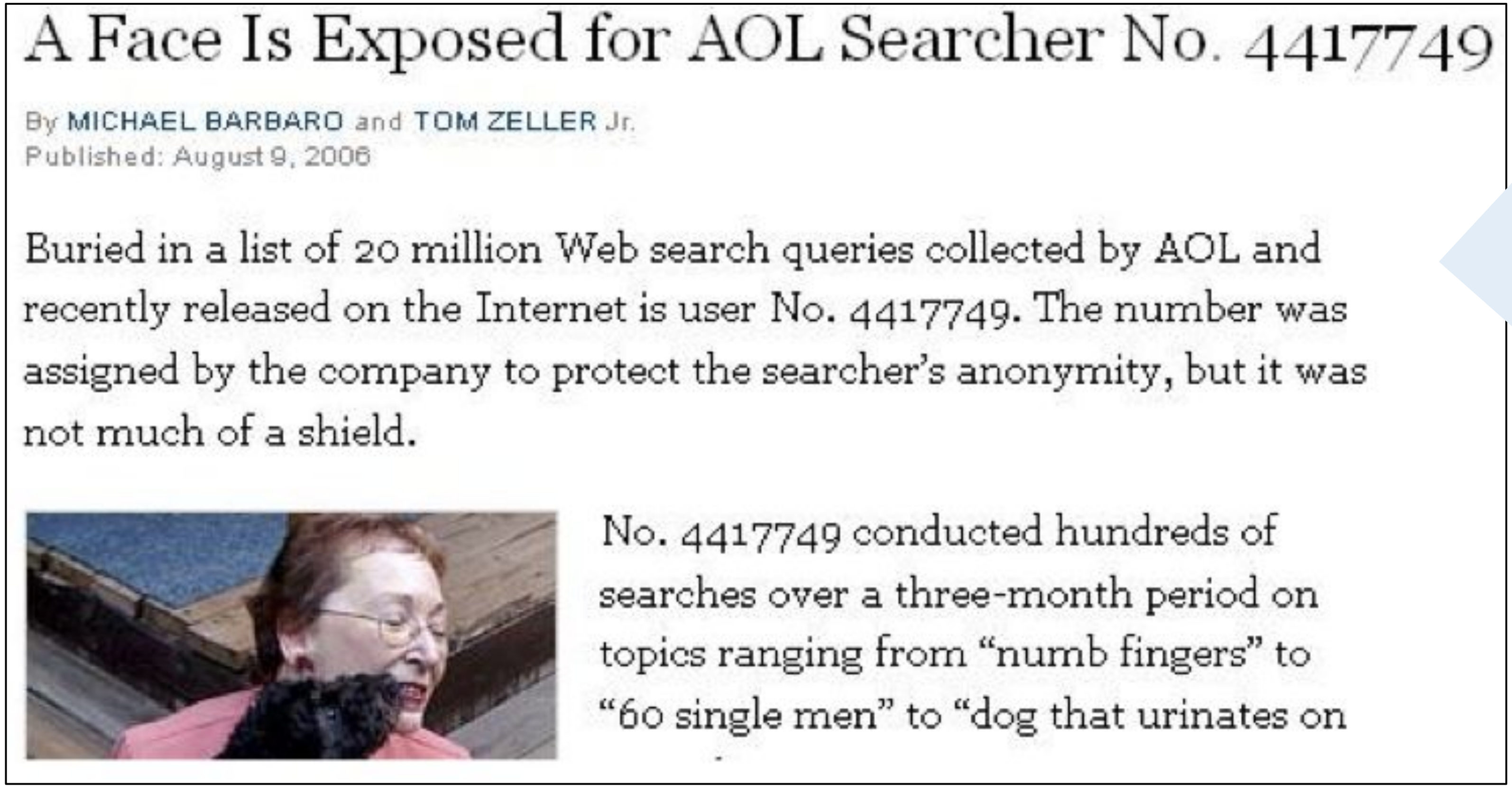
No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on

Thelma Arnold from Lilburn, Georgia, was reidentified from ***a few searches***.

- "Landscapers in Lilburn, Georgia"
- People with the last name Arnold
- Homes sold in Shadow Lake

[NYT'06]

Re-Identification Attack: AOL Example (cont.)



A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on

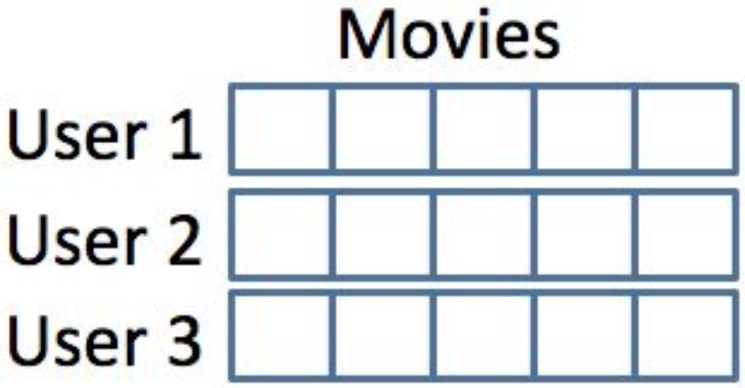
The reidentification of the user's ID exposed other of her searches.

- "Numb fingers"
- "60 single men"
- "Dog that urinates on everything"

[NYT'06]

Re-Identification Attack: Netflix Example

NETFLIX



Movies

User 1

User 2

User 3

- Users were reidentified through ratings they posted ***publicly on IMDb*** under their own names.
  - 2–8 public ratings and dates sufficient to re-identify 99% of the dataset
  - Once you re-identify X, you can learn ratings X didn’t post publicly
- In response, Netflix canceled its second competition.

[Narayanan+08]

Attack “Ingredients”

- AOL attack:
  - **Data distribution:** a few searches uniquely identify a person (pseudo-PII)
  - **Side information:** access to public directory of residents

- Netflix attack:
  - **Data distribution:** a few movie ratings uniquely identify a person (pseudo-PII)
  - **Side information:** access to public IMDb records

- Same “ingredients” enable attacks against many “anonymized” datasets
  - Human mobility traces are particularly de-anonymizable: e.g., knowing **4 spatio-temporal locations** or **3 most-visited locations** (side information) is sufficient to uniquely identify an individual in large metro traces (because that’s how user data distribution is…) [Wang+18]

Examples from CitiBike Data

(results obtained by previous Private Systems students!)

Examples from CitiBike Data

(results obtained by previous Private Systems students!)

- CitiBike releases monthly data about all rides to enable public analyses of how this publicly funded project is used by New Yorkers

- For each ride, they release:

Released Data
`start date/time`
`start-station`
`end date/time`
`end-station`
~~`age,`~~
~~`gender`~~

Question: Is this “okay to share”?

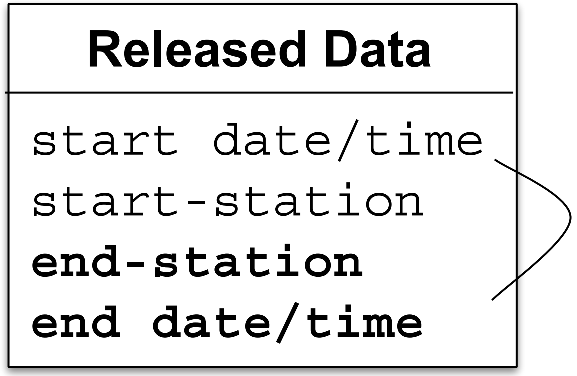
- no personally identifiable information (PII) is released, so data is “anonymous”

Examples from CitiBike Data

(results obtained by previous Private Systems students!)

- Suppose someone knows the start date/time and start-station of a ride

- Then, for 81% of NYC rides, they can learn with 100% confidence the **end-station** and **end date/time** of that ride



Released Data

start date/time
start-station
end-station
end date/time

Examples from CitiBike Data

(results obtained by previous Private Systems students!)

- Suppose someone knows the start date/time and start-station of a ride

Released Data

```
start date/time
start-station
```
`end-station`

`end date/time`

- Then, for 81% of NYC rides, they can learn with 100% confidence the **end-station** and **end date/time** of that ride

Question: Is *that* okay?

- consider spying neighbor
- consider an abusive partner

Examples from CitiBike Data

(results obtained by previous Private Systems students!)

- Suppose someone knows the start date/time and start-station of a ride **(side information)**
- Then, for 81% of NYC rides, they can learn with 100% confidence the **end-station** and **end date/time** of that ride **(data distribution)**

Released Data

```
start date/time
start-station
end-station
end date/time
```

Question: Is *that* okay?

- consider spying neighbor
- consider an abusive partner

attacker context

Coarser Time Doesn’t Eliminate Problem

Attacker side information	New info learned	Efficacy

start-station + start-date/time (to the minute)	end-station
	end-date/time
		81% of the rides are vulnerable

start-station + start-date/time (within 10 min)		35% of the rides are vulnerable
start-station + start-date/time (within 30 min)		15% of the rides aree vulnerable
start-station + start-date/time (within 60 min)		7% of the rides are vulnerable

Other vulnerabilities include:

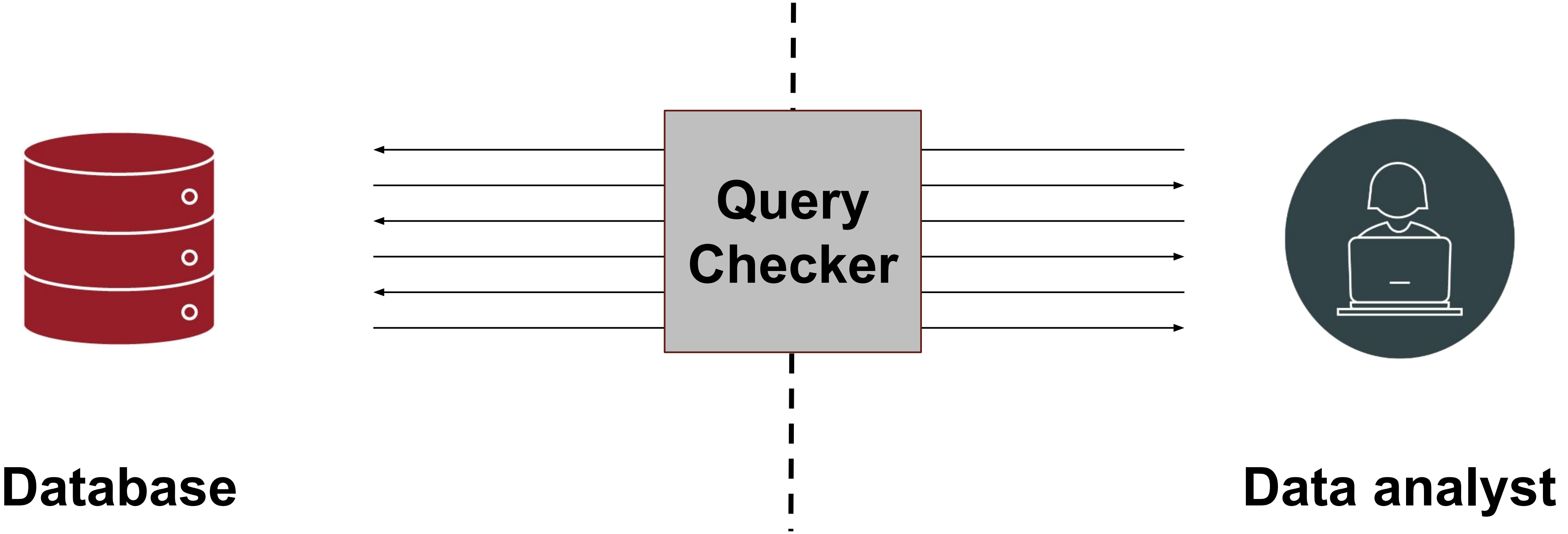
- attacker knows end-station + end-date/time; learns start-station + start-date/time
- attacker knows you rode with a friend; learns when and where

Anonymization and Attacks against It

The End

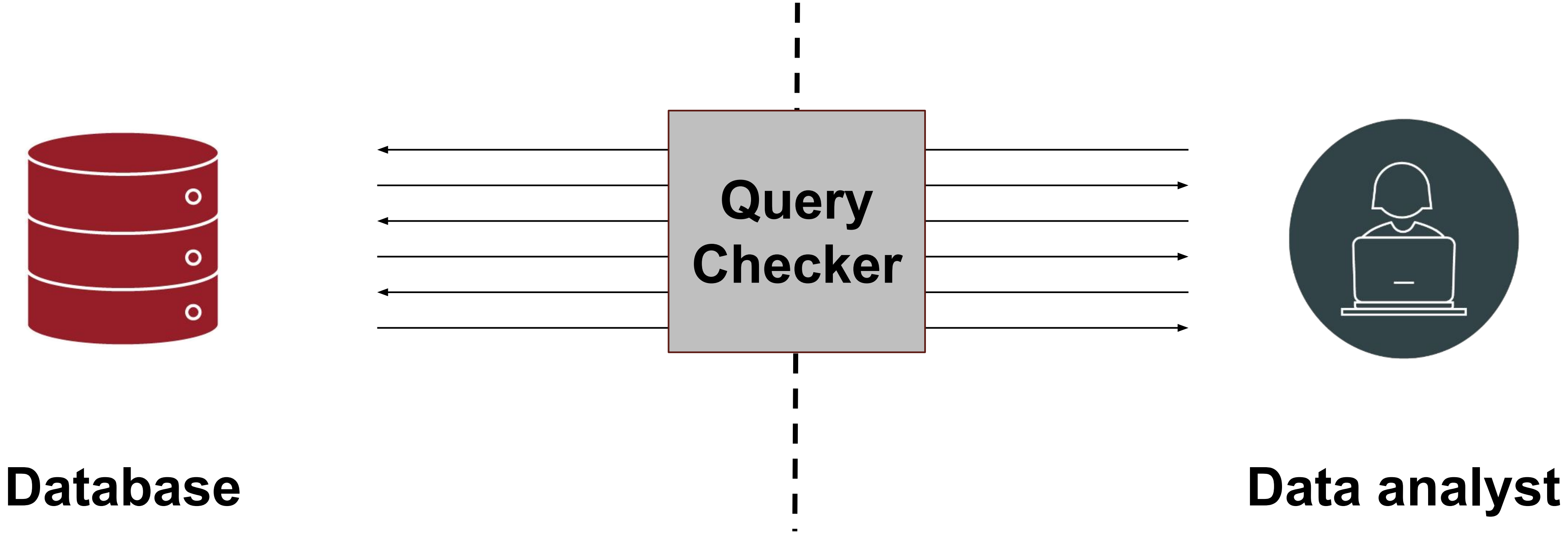
Aggregates and Attacks against Them

Aggregates “Only”



For example: Query Checker only permits statistical queries that aggregate information from large groups

Aggregates “Only” (cont.)



Problem: “The aggregates masquerade.” [Cynthia Dwork]

Intuition

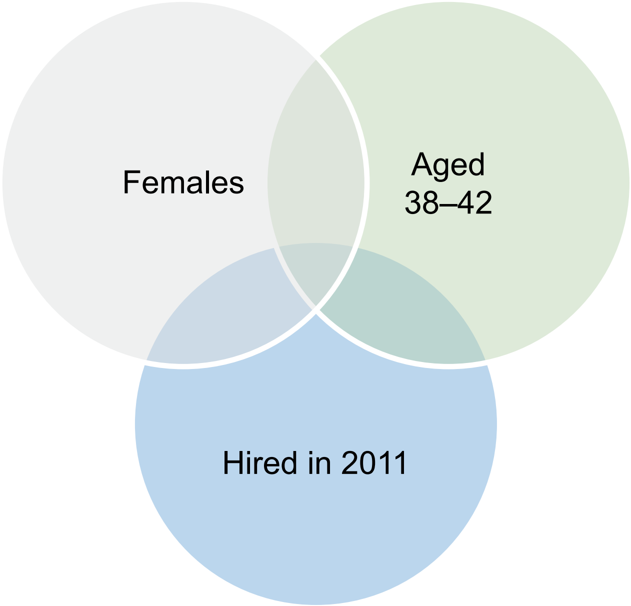
- Query 1: What is the average salary of faculty at CU?

Intuition

- Query 1: What is the average salary of faculty at CU?
- Query 2: What is the average salary of faculty, *excluding Roxana Geambasu*?

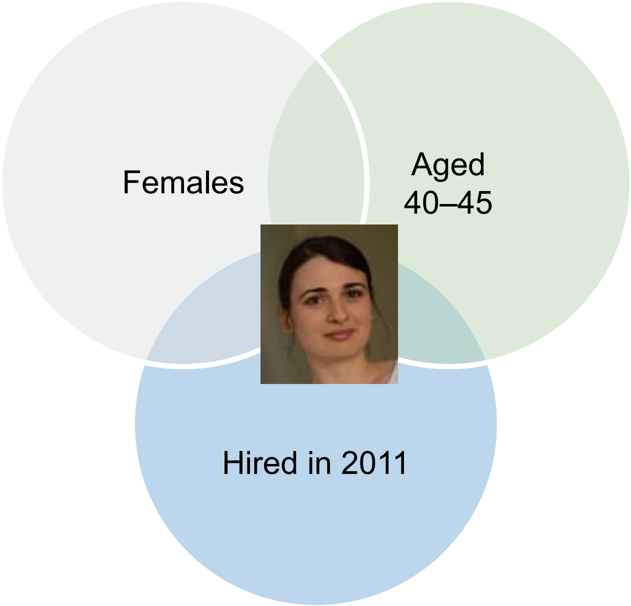
Intuition

- Query 1: What is the average salary of faculty at CU?
- Query 2: What is the average salary of faculty, *excluding Roxana Geambasu*?
- Alternatively, you may ask these three queries:



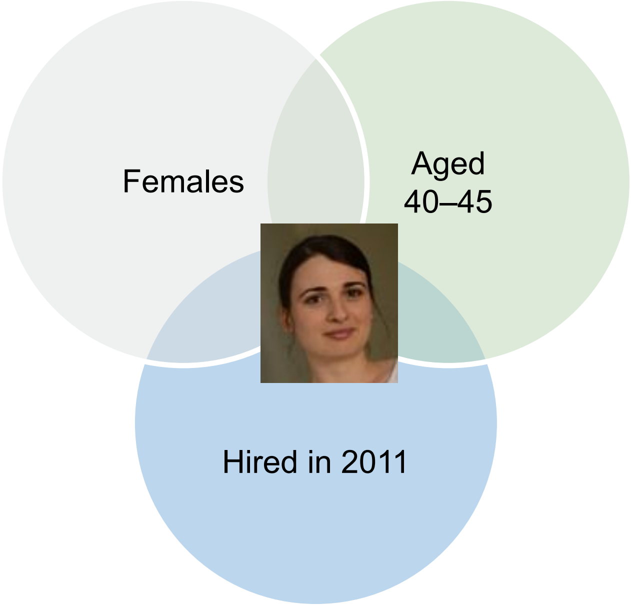
Intuition

- Query 1: What is the average salary of faculty at CU?
- Query 2: What is the average salary of faculty, *excluding Roxana Geambasu*?
- Alternatively, you may ask these three queries:



Intuition

- Query 1: What is the average salary of faculty at CU?
- Query 2: What is the average salary of faculty, *excluding Roxana Geambasu*?
- Alternatively, you may ask these three queries:



With **multiple queries**, or some **auxiliary information**, one can usually find (adaptive) statistical queries that **in combination** reveal private information.

Privacy Attacks

- Two types of attacks can be mounted against aggregates:
  - ***Membership inference***
  - ***Database reconstruction***

- We describe both by example next

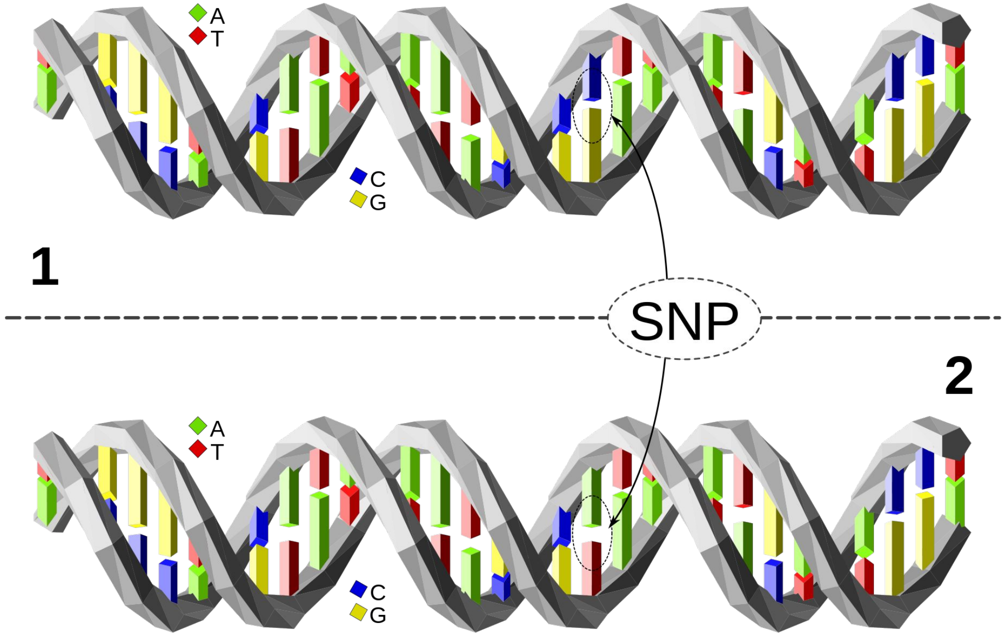
- CA will then demo some

Membership Inference

- **Example:** Genome-Wide Association Studies (GWAS), which find associations between single nucleotide polymorphisms (SNPs) and a certain disease

- Studies compare allele frequencies in each SNP for the diagnosed group vs. a reference group

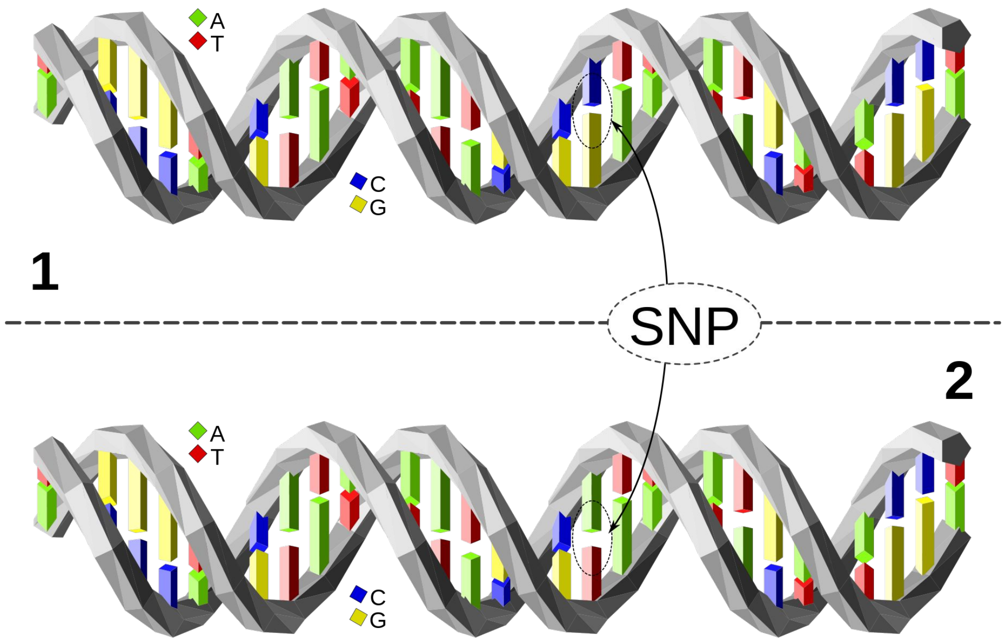
- They release allele frequencies for hundreds of thousands of SNPs



[Homer+08]

Membership Inference (cont.)

- [Homer+08] showed:
  - By having access to DNA of X, one can reverse whether X was in the ***diagnosed group*** from published GWAS statistics
    - $D(Yij) = |Yij - Refj| - |Yij - Diagj|$
  - Response: National Institutes of Health (NIH) changed rules about statistical data release from funded studies



[Homer+08]

Database Reconstruction

- **Example:** US Census Bureau's reconstruction of the 2010 Census database from the released statistics
- Idea: find a database consistent with the released statistics
- Mixed linear program formulation: find an assignment for all database cells under the constraints given by the released statistics

[Garfinkel+18]

Database Reconstruction (cont.)

1. Reconstructed <*block, sex, age*, **race, ethnicity**> for 46% of the population (142 million of 308 million people)
2. Linked <*block, sex, age*> to commercial data, resulting in putative reidentification of 45% of the population (138 million people)
3. Validated reidentification for 17% of the population (52 million people)

For 52 million Americans, self-declared **<race, ethnicity>** can now be recovered by anyone in the world!

[Garfinkel+18]

Attack “Ingredients”

- GWAS attack:
  - **Data distribution:** expression at a few SNPs are unique to a person
  - **Side information:** access to (partial) DNA of target
  - **Multiple queries:** allele frequencies at many SNPs were released

- Census attack:
  - **Data distribution:** <block, sex, age> can uniquely identify a person in certain areas
  - **Side information:** access to commercial data
  - **Multiple queries:** contingency tables over many groupings of the data were released

Attack “Ingredients”

- GWAS attack:
  - **Data distribution:** expression at a few SNPs are unique to a person
  - **Side information:** access to (partial) DNA of target
  - **Multiple queries:** allele frequencies at many SNPs were released

- Census attack:
  - **Data distribution:** <block, sex, age> can uniquely identify a person in certain areas
  - **Side information:** access to commercial data
  - **Multiple queries:** contingency tables over many groupings of the data were released

attacker context

Aggregates and Attacks against Them

The End

Fundamental Attack

Dinur-Nissim Attack

- A landmark theoretical result by Dinur and Nissim in 2003 shows that the data reconstruction attack is **fundamental**, not just a possibility in obscure cases

- Informally, their result can be stated as follows: "Releasing **overly accurate** estimates of **too many** linear statistics from a dataset fundamentally enables reconstruction of the dataset"
  - Proof gave a general albeit inefficient algo for reconstruction

- Many privacy attacks are instantiations of this fundamental attack, more efficient but less general

[Dinur&Nissim'03]

ML Also Leaks

- ML is more complex than linear statistics, but don’t let that fool us.
- Large networks, particularly deep neural networks (DNNs), have been shown to memorize specific examples in the training set [Nasr+23], [Carlini+19], [Nasr+19], [Shokri+17].
- You will read one paper as an example.
- Practical demonstrations of attacks are a highly active area of research, so expect further progress!

"Lay" Take-Aways

1. Even without "PII," it's possible to learn sensitive info about individuals from data releases, especially with **side information** or **multiple queries** (a.k.a. **attacker context**).
2. It's difficult to determine what's "okay to release," because vulnerability to attack depends on **data distribution** and **attacker context.**
3. Ad-hoc solutions (incl. anonymization, k-anonymity, aggregates-only) are unreliable, because they too depend on **data distribution** and **attacker context.**
4. Next time: **differential privacy**, a rigorous privacy technology to establish "what's okay to release" that does NOT depend on these!

Fundamental Attack

The End

References Cited So Far

[NYT’06] Barbaro, Zeller. A face is exposed for AOL searcher*. The New York Times*, 2006.

[Narayanan+08] Narayanan & Schmatikov. *Robust deanonymization of sparse datasets (Netflix Prize Data).* IEEE S&P, 2008.

[Garfinkel+18] Garfinkel, Abowd, & Martindale. *Understanding database reconstruction attacks on public data*. Communications of the ACM, 2018.

[Homer+08] Homer, Szelinger, Redman, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In *PLOS Genetics*, 2008.

References Cited So Far

[Carlini+19] Carlini, Liu, Erlingsson, Kos, & Song. *The secret sharer: Evaluating and testing unintended memorization in neural networks*. USENIX Security, 2019.

[Dinur&Nissim'03] Dinur & Nissim. *Revealing information while preserving privacy.* PODS, 2003.

[Kasiviswnathan+13] Kasiviswanathan, Rudelson, & Smith. The power of linear reconstruction attacks. *SODA*, 2013.

[Nasr+19] Nasr, Shokri, & Houmansadr. *Comprehensive privacy analysis of deep learning*. IEEE S&P, 2019.

[Shokri+17] Shokri, Stronati, Song, & Shmatikov. *Membership inference attacks against machine learning models*. IEEE S&P, 2017.

[Nasr+23]]Nasr, Carlini, Hayase, et al., ArXiv 2023. Scalable Extraction of Training Data from (Production) Language Models.

ML Privacy Attacks: Mechanics and Demos

Privacy Attack Types

Two main types:

- Membership inference
- Data reconstruction

Attacks are relevant for any type of statistical analyses, but we focus here on ML.

Membership Inference

- Determine whether a given record was part of the training dataset.
- This standard attack is a privacy yardstick
  - Revealing membership can directly be harmful in some settings
  - Also useful as a building block for other attacks
  - Simple and universal definition

MI security game [Carlini+22]

Challenger C, adversary A, data distribution D

1. C samples a training dataset d from D and trains a model f = T(d)
2. C flips a bit b:
   - If b = 0: sample (x,y) from D
   - If b =1: sample (x,y) from d
3. C sends (x,y) to A
4. A outputs a guess g – informed by queries on f and D

Evaluating an attack:

- Balanced accuracy: Pr[g = b]. E.g. ½ if the adversary is random
- True Positive and False Positive rates:
  - TPR=1 if A always correctly identifies a training sample
  - FPR=0 if A never mistakes a non-member for a training sample

MI example: the LOSS attack [Yeom+18]

- Idea:
  - When you train an ML model, the train loss is usually lower than the test loss.
  - If I give you a point, you can compute its loss
  - Low loss = likely to be a train point (member)
  - High loss = likely to be a test point (non-member)
- Simple and lightweight attack: see demo notebook on Courseworks.
- Other attacks exist:
  - [Carlini+22] optimizes for high TPR at low FPR
  - [Shokri+17] doesn’t need access to the model’s loss

Data Reconstruction

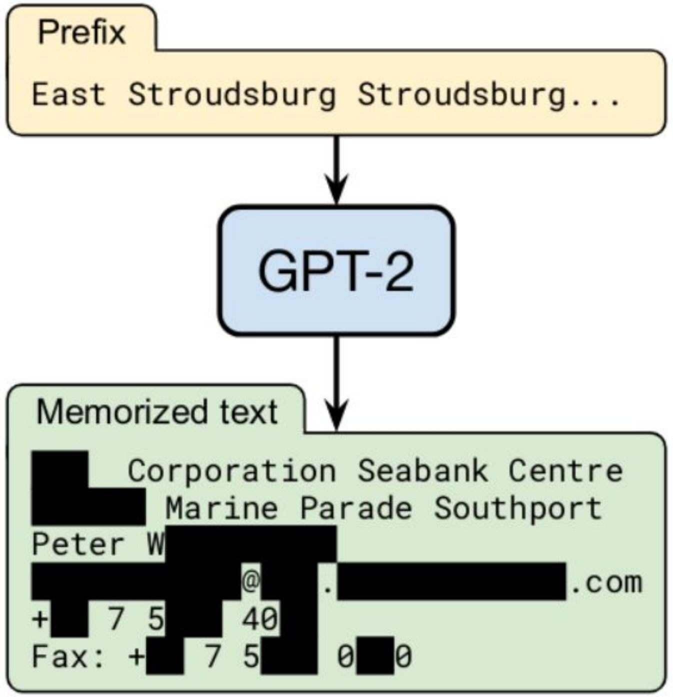
- Recover a training sample from a model
  - Also known as training data extraction
  - Related attacks: model inversion (approximate information about the training dataset), attribute inference (recover sensitive features given public features)
- Harder than membership inference
  - Reduction: an adversary who knows how to reconstruct data can use reconstruction to do membership inference [Yeom+18]
- See paper about ChatGPT attacks discussed in class, or [Carlini+21]

Extracting data from GPT2 [Carlini+21]

- GPT2, a large language model:
  - Transformer with 1.5B parameters
  - Trained on 40GB of Internet text
  - Publicly accessible data can still cause privacy harms
- Generating text:
  - Given a prompt, predict the next word
  - $\Pr[x_1, \ldots, x_n] = \Pi_i \Pr[x_i \mid x_1, \ldots, x_{i-1}]$
- Attack idea:
  - Generate many possible outputs
  - Run a membership inference attack on each output
  - Keep the top outputs, they are likely memorized training samples

Extracting data from GPT2 [Carlini+21]

- Results:
  - Names, addresses, phone numbers
  - Even a string of 87 characters that appears only 10 times (in a single page)
  - 67% true positive rate for the best variant of the attack (confirmed by Googling the top 100 candidates)
- Some techniques to get better samples:
  - Condition on promising prompts
  - Compare to another language model
  - Sample with temperature instead of taking the top-1 or top-n most likely next tokens



Prefix

East Stroudsburg Stroudsburg...

GPT-2

Memorized text

[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Other Attacks

There are other attacks against ML models that can be used to strengthen privacy attacks. For example:

- Model extraction attack:
  - Recover the model weights from an inference API
  - Poses a security threat (intellectual property theft)
  - Can also lead to data extraction for some classes of models (e.g. kernel logistic regression [Tramer+16])
- Poisoning attack:
  - Give special samples to trigger a particular model behavior
  - Can strengthen data extraction attacks (e.g. the Truth Serum attack [Tramer+22])

References Cited

[Yeom+18] Yeom, Giacomelli, Fredrikson, Jha. *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting.* IEEE CSF, 2018.

[Tramer+16] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, *Stealing Machine Learning Models via Prediction APIs*, USENIX Security 16

[Tramer+22] F. Tramèr et al., *Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets*, CCS 2022.

[Shokri+17] Shokri, Stronati, Song, & Shmatikov. *Membership inference attacks against machine learning models*. IEEE S&P, 2017.

[Carlini+22] Carlini, Chien, Nasr, Song, Terzis, Tramèr. *Membership Inference Attacks From First Principles*. IEEE S&P, 2022.

[Carlini+21] Carlini, Tramèr, Wallace, Jagielski, Herbert-Voss, Lee, Roberts, Brown, Song, Erlingsson, Oprea, Raffel. *Extracting Training Data from Large Language Models*. USENIX Security Symposium 2021.

ML Privacy Attacks: Mechanics and Demos

The End

Homework 1 Overview

(CA walks through HW1 notebook posted on courseworks)