Data Privacy in the Age of AI

Srini Devadas

Data Privacy

- Given a dataset with sensitive information, such as:
  - Census data
  - Health records
  - Social network activity
  - Telecommunications data
  - Financial records
- How can we:
  - enable desirable uses of the data
  - while protecting the privacy of the data subjects?

- Health diagnostics
- Intrusion detection
- Computing financial risk
- Combating misinformation

Privacy Risk: Querying ChatBots Can Leak Personal Information

6 Things You Should Never Share with ChatGPT

ChatGPT is a great tool, but you shouldn't trust it with your sensitive and private data.

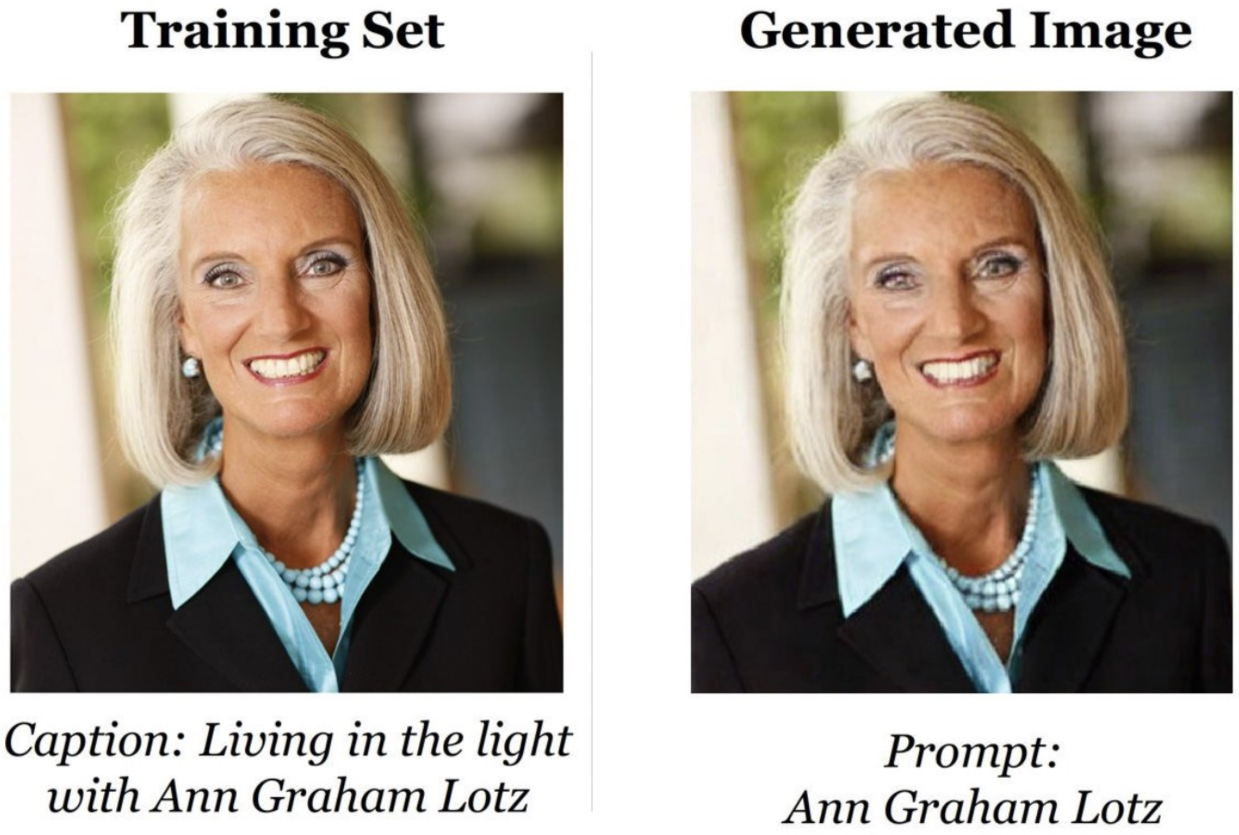
Artificial Intelligence | Posted 03/15/2023 | Articles

Bing's Issues Aren't Funny, They're Dangerous—And May Spawn Bad Regulation

ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned

Published: February 7, 2023 8:06pm EST Updated: February 9, 2023 11:30pm EST

Privacy Risk: Training Data Leakage from Stable Diffusion (Vision) Model



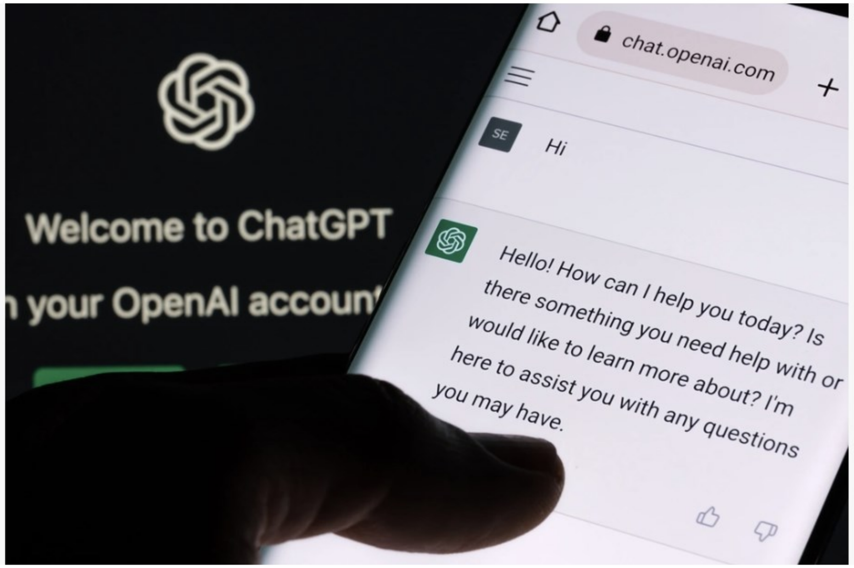
https://usa.kaspersky.com/blog/neural-networks-data-leaks/28191/

Privacy Risk: Training Data Leakage from Large Language Model (LLM)

Regurgitate large amounts of training data

Jai Vijayan, Contributing Writer
December 1, 2023

5 Min Read

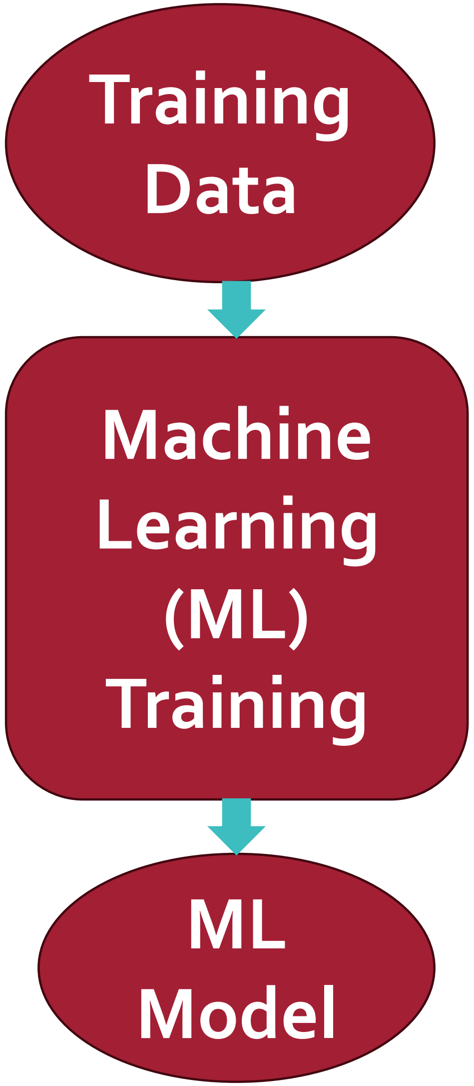


SOURCE: ASCANNIO VIA SHUTTERSTOCK

Can getting ChatGPT to repeat the same word over and over again cause it to regurgitate large amounts of its training data, including personally identifiable information and other data scraped from the Web?

https://www.darkreading.com/cyber-risk/researchers-simple-technique-extract-chatgpt-training-data

Aspects to Privacy



Privacy Concern

Needs to happen privately

Confidential computing

Exposure can hurt privacy!

Aspects to Privacy

Don't put in personal information!

Private Inference

Query from Client

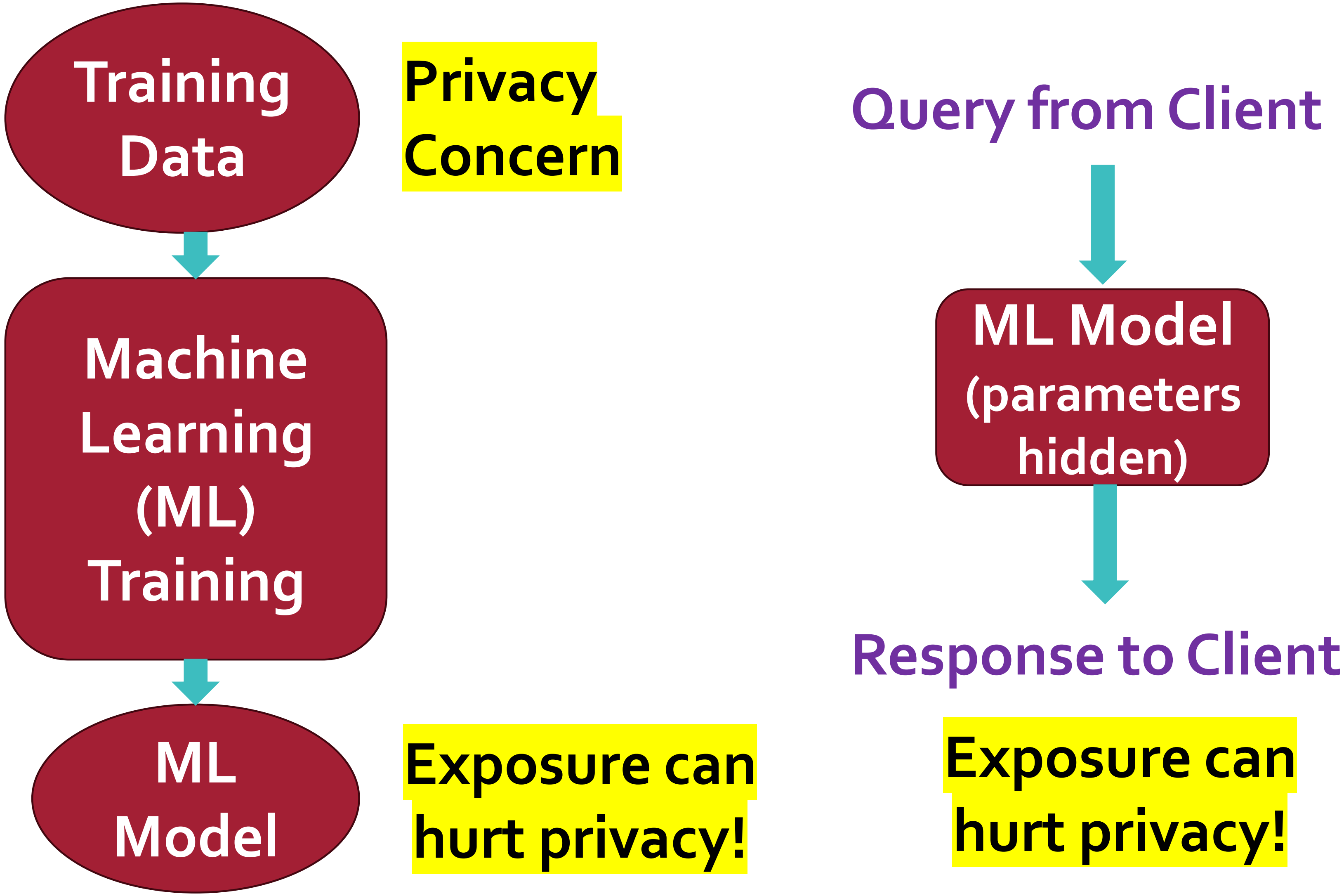
ML Model (parameters hidden)

Response to Client

Private Inference

- Use Fully Homomorphic Encryption (FHE) to train model on encrypted training data to produce an encrypted model
- Main challenge: Multiple orders of magnitude slowdown
- **Approach: Hardware acceleration of FHE**

Aspects to Privacy



Information leakage r

Information leakage from arbi
can be described as th

- X:

nation disclosed

  - stics (mean/median) of a sensitive dataset X
  - Neural networks (and their model parameters) learned from samples X
  - Responses to inference queries from neural network learned from samples X

Automatically privatize any algorithm without looking inside of it.

Defining Privacy

An adversary cannot recover our secret given the leakage

Defining Privacy

An adversary cannot recover our secret given the leakage

Worst-case guarantee against **<u>arbitrary</u>** adversary strategy

Defining Privacy

An adversary cannot recover our secret given the leakage

Upper bounds of **<u>success rate or probability</u>** (security parameter)

Defining Privacy

An adversary cannot recover our secret given the leakage

Successful statistical inference by returning a **satisfied estimation**

Defining Privacy

An adversary cannot recover our secret given the leakage

Output of a **black-box** processing mechanism

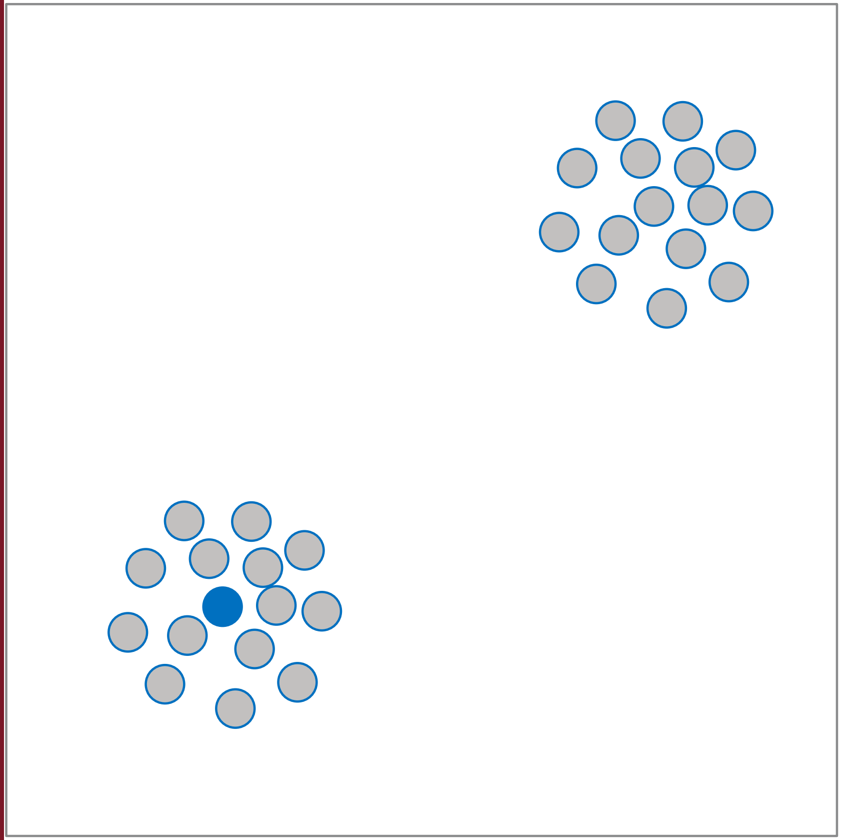
Differential Privacy

A. Independent of prior knowledge (data entropy)
B. The privacy/security proof is hard to generate (could be NP-hard)
C. Utility-privacy tradeoff can be loose, requiring large noise

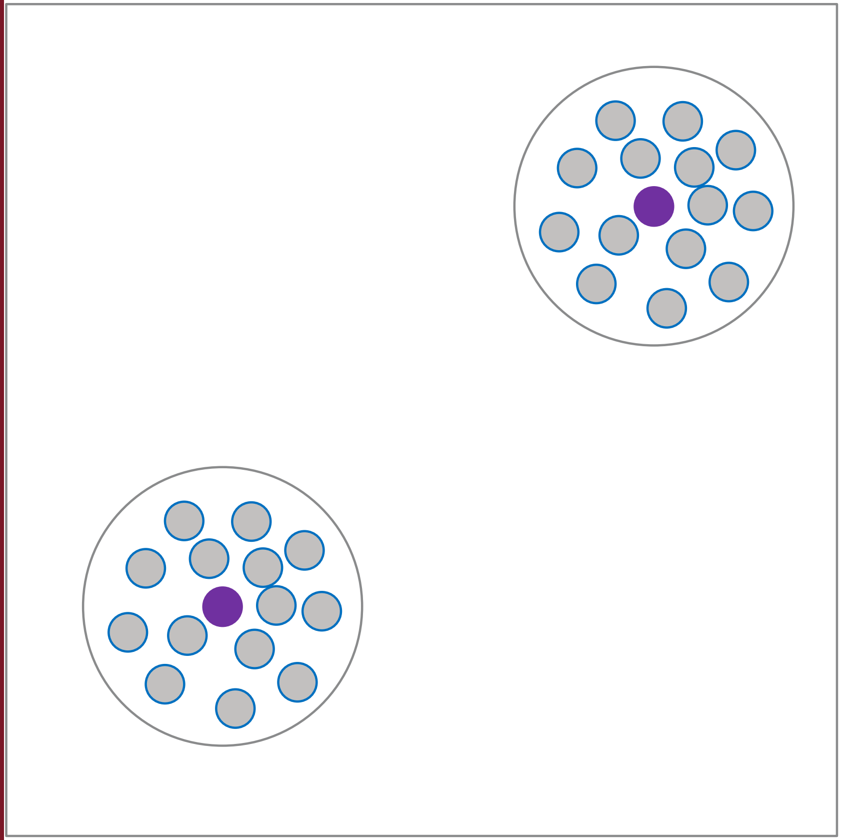
PAC Privacy

A. Still worst-case (against arbitrary adversarial strategy) but exploits and depends on data entropy obtained from assumed distribution
B. Automatic PAC Privacy proof and privatization scheme
C. Tight utility-privacy tradeoff: only adding necessary noise

Consider K-Means

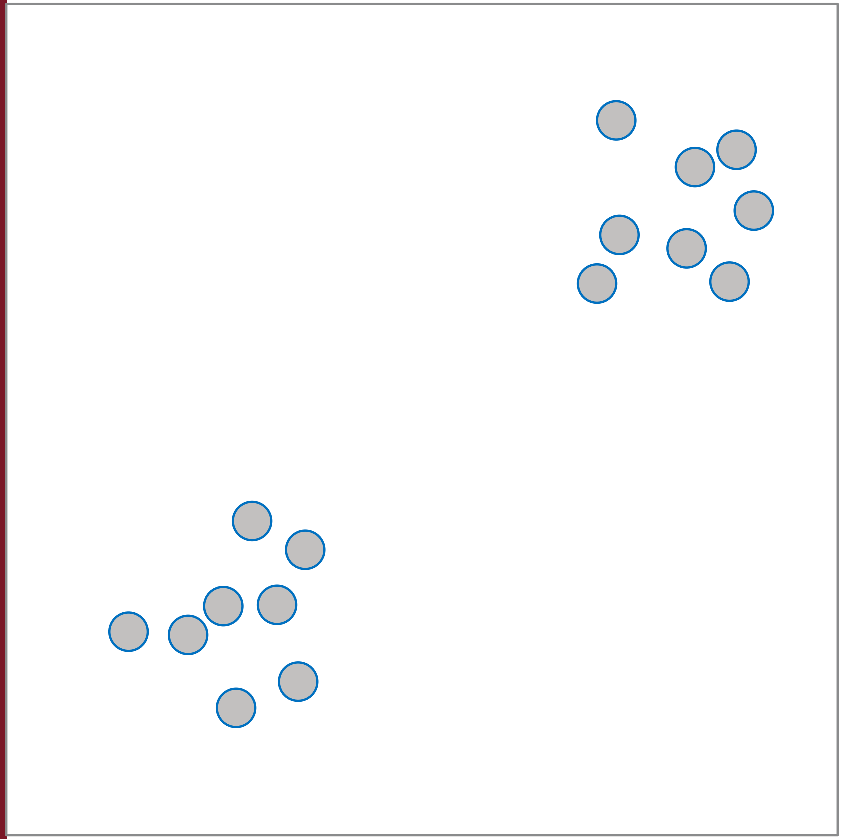


Consider K-Means



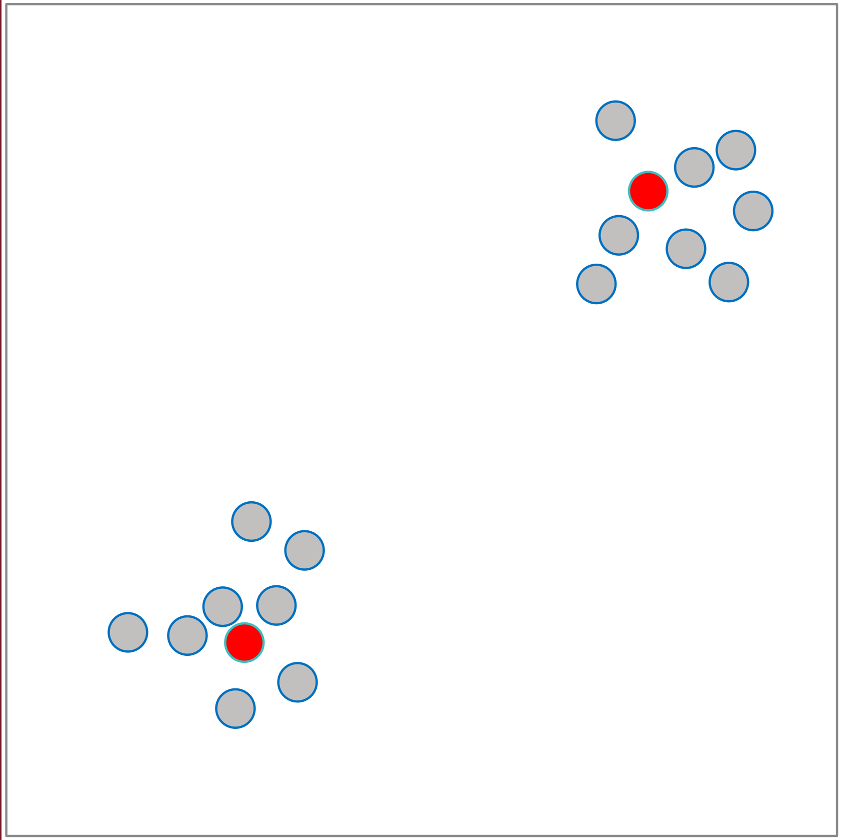
- Non-private 2-means clustering produces purple centroids that define 2 clusters
- Want to release centroids **with guarantees that data points cannot be reconstructed**

Consider K-Means



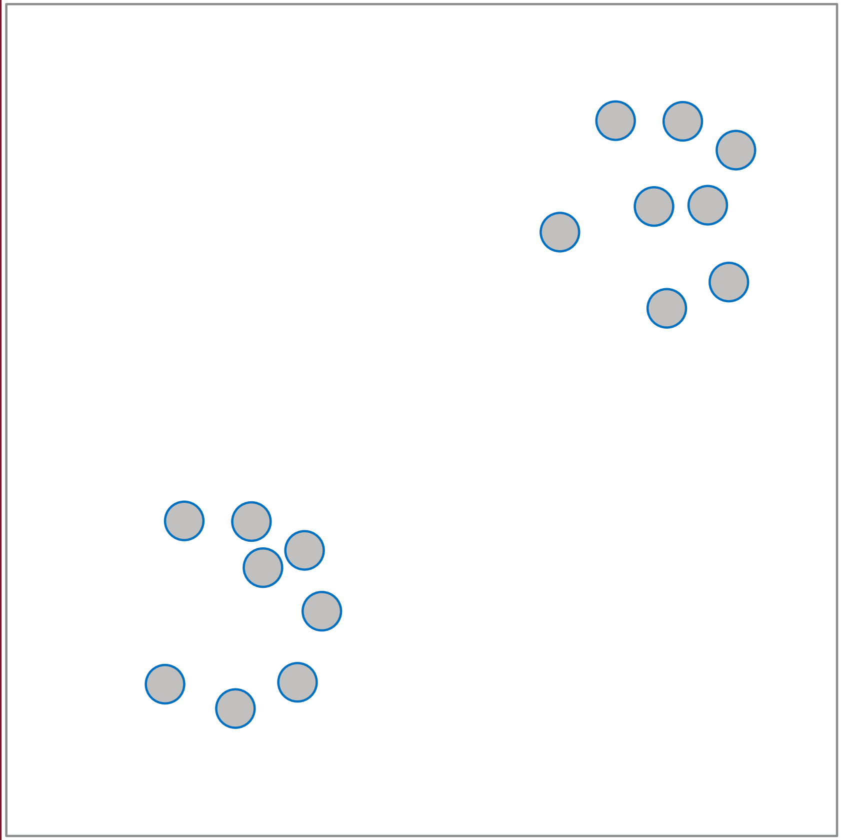
- Select half the points randomly in a trial

Consider K-Means



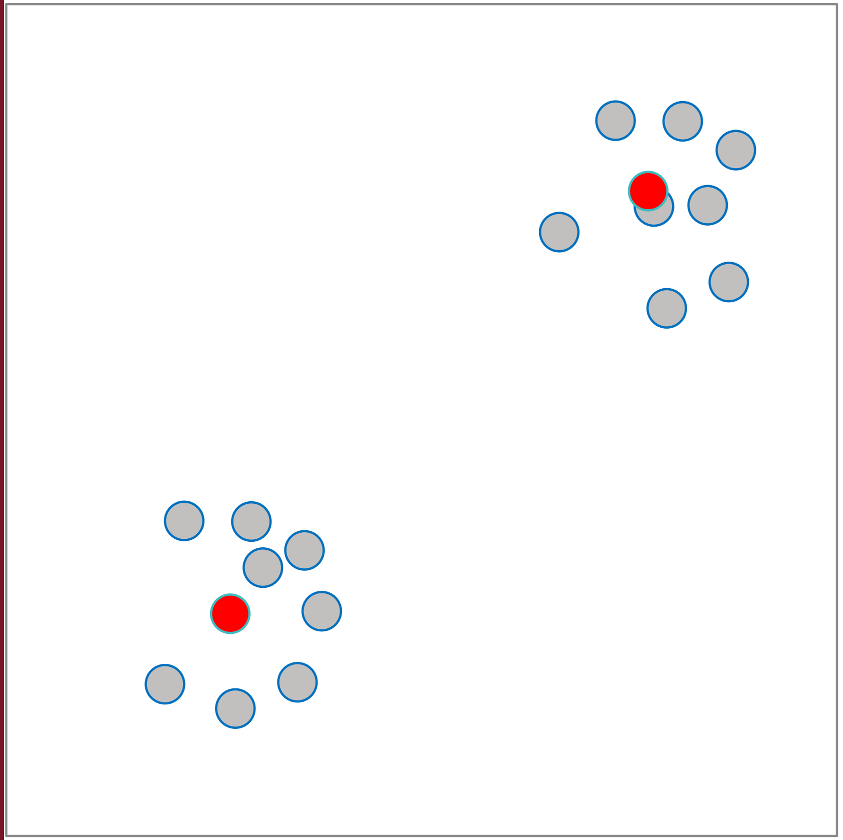
- Select half the points randomly in a trial
- Run K-means algorithm on only the selected points to produce centroids

Consider K-Means



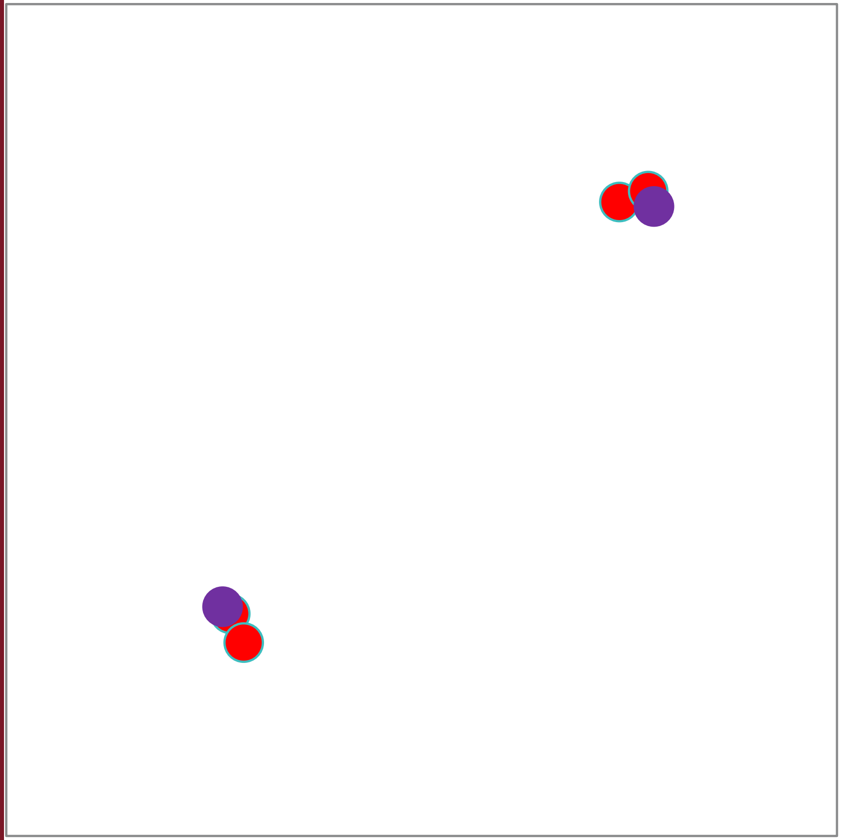
- Select a different half the points randomly in a different trial

Consider K-Means



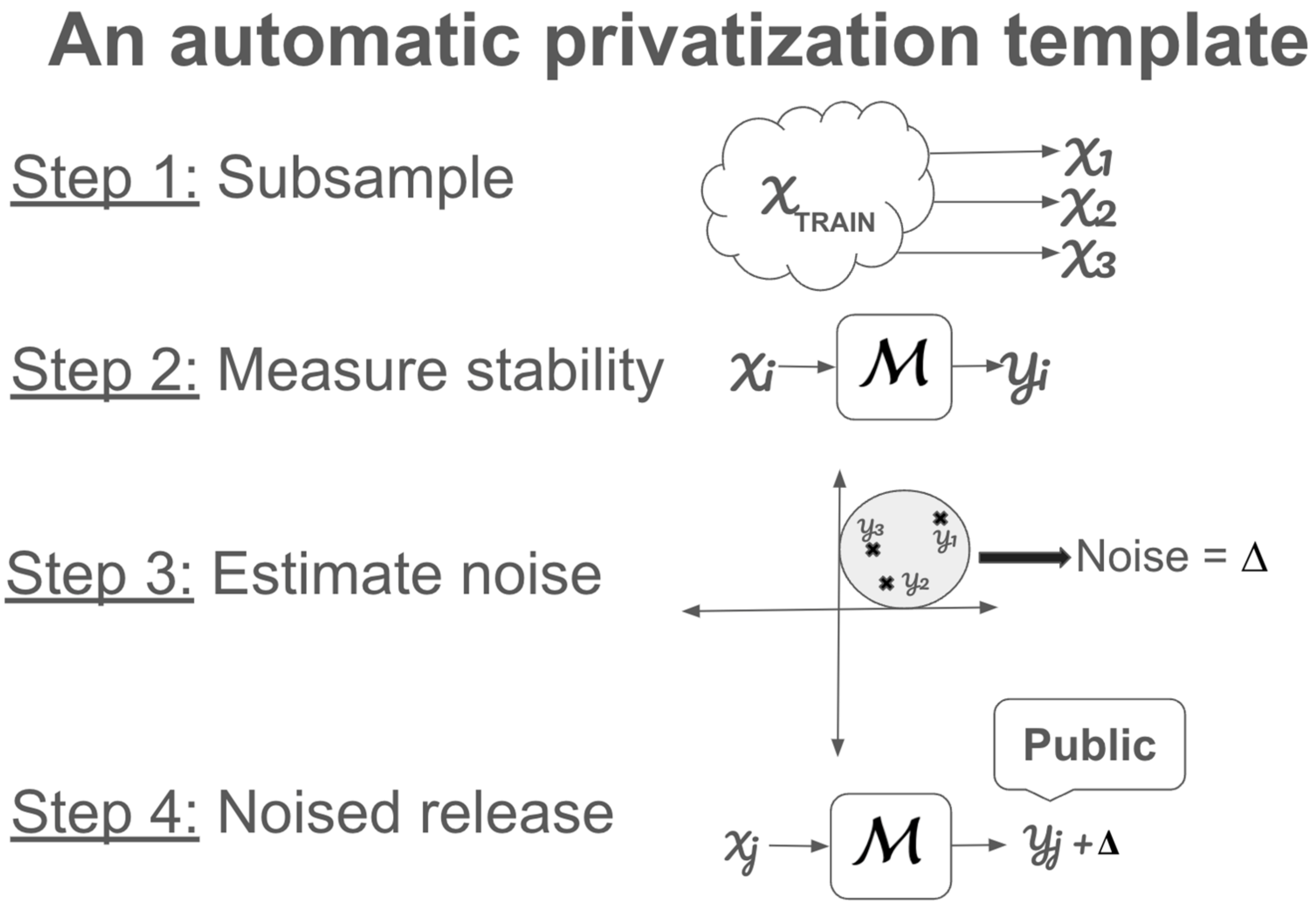
- Select a different half the points randomly in a different trial
- Run K-means algorithm on only the selected points to produce centroids

Consider K-Means



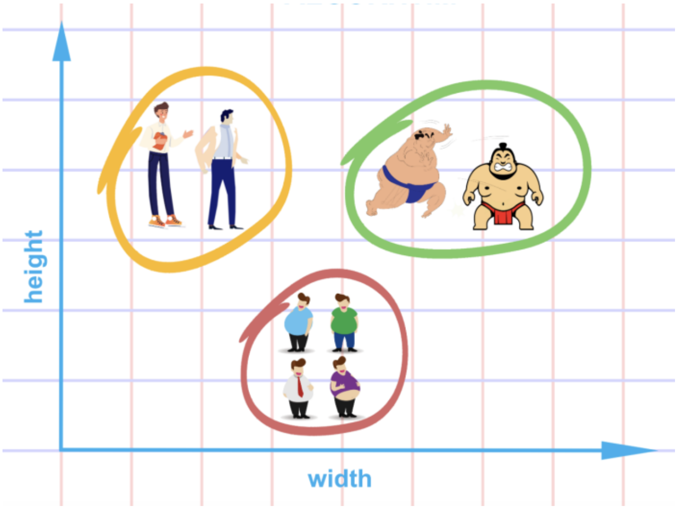
- Since the two clusters are far apart, we will get close to the same centroids for each trial, which are also close to the non-private "ground truth" centroids
- Release either red centroid pair after adding small noise; this will ensure high PAC Privacy for all data points

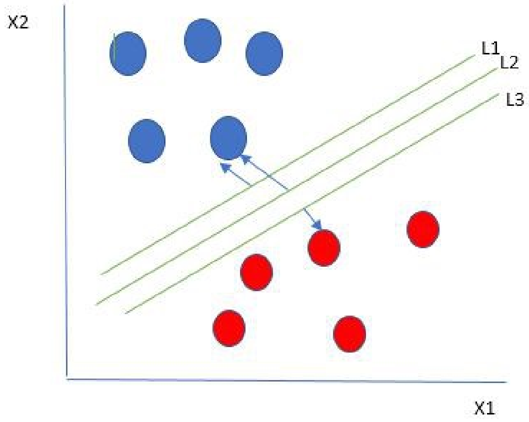
Privatizing Black-box Algorithms



Lots of Ongoing Work

- Privatized many classic algorithms
  - Support Vector Machines (SVM)
  - K-Means
  - Principal Component Analysis
  - Random Forest
- The more stable the algorithm, the easier it is to privatize
- Current: Privatize Deep Learning, Private Database analytics





Long-term Goals

- Allow deep learning model that detects disease to be deployed guaranteeing privacy of patients whose data was used to train the model
- Large Language Model (LLM) responses to queries guarantee privacy of training data
- And more …

Thank you!

Srini Devadas

devadas@mit.edu