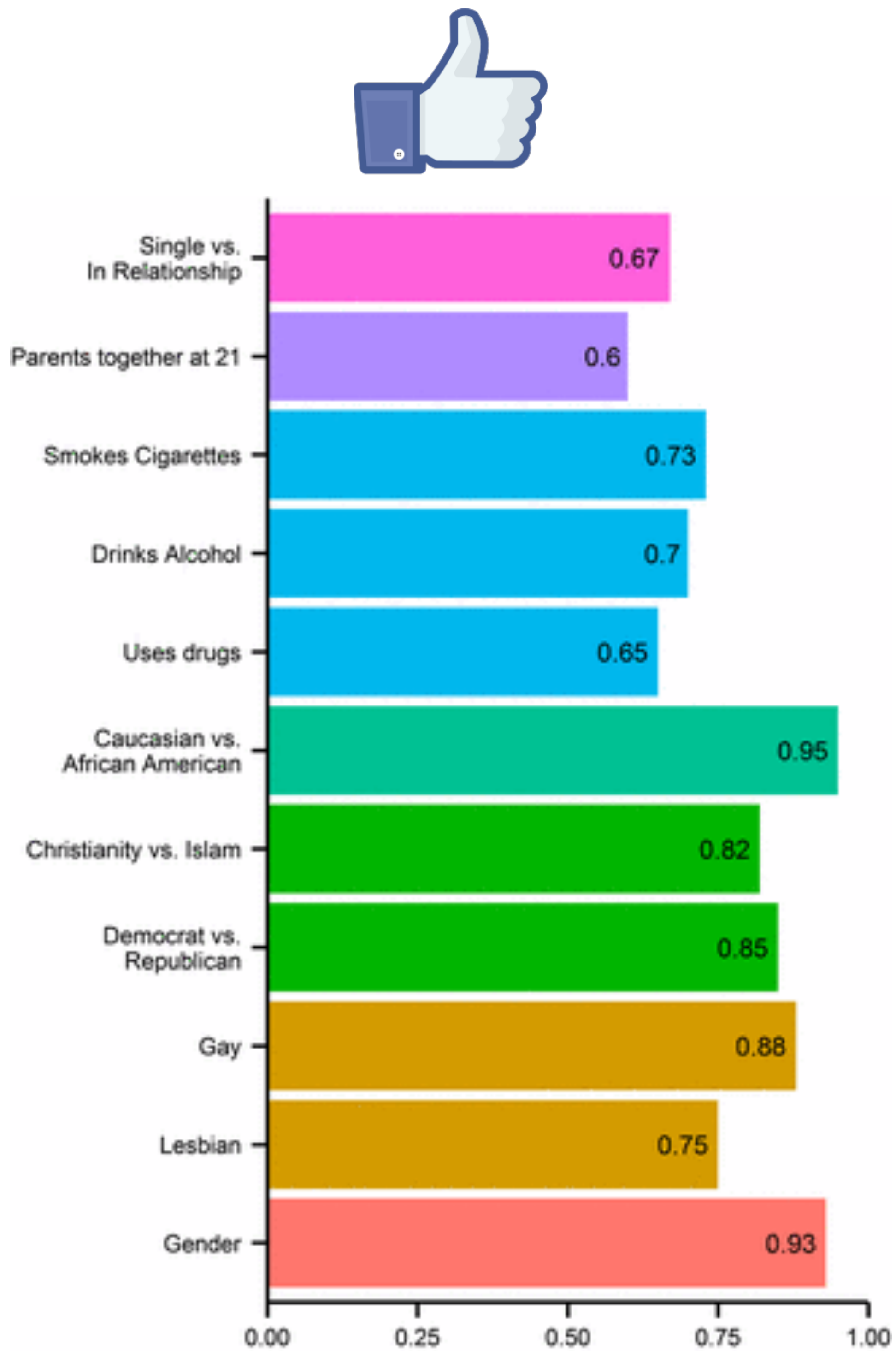
Data Privacy Risks of Machine Learning

Reza Shokri

reza@nus.edu.sg

Inference Attacks

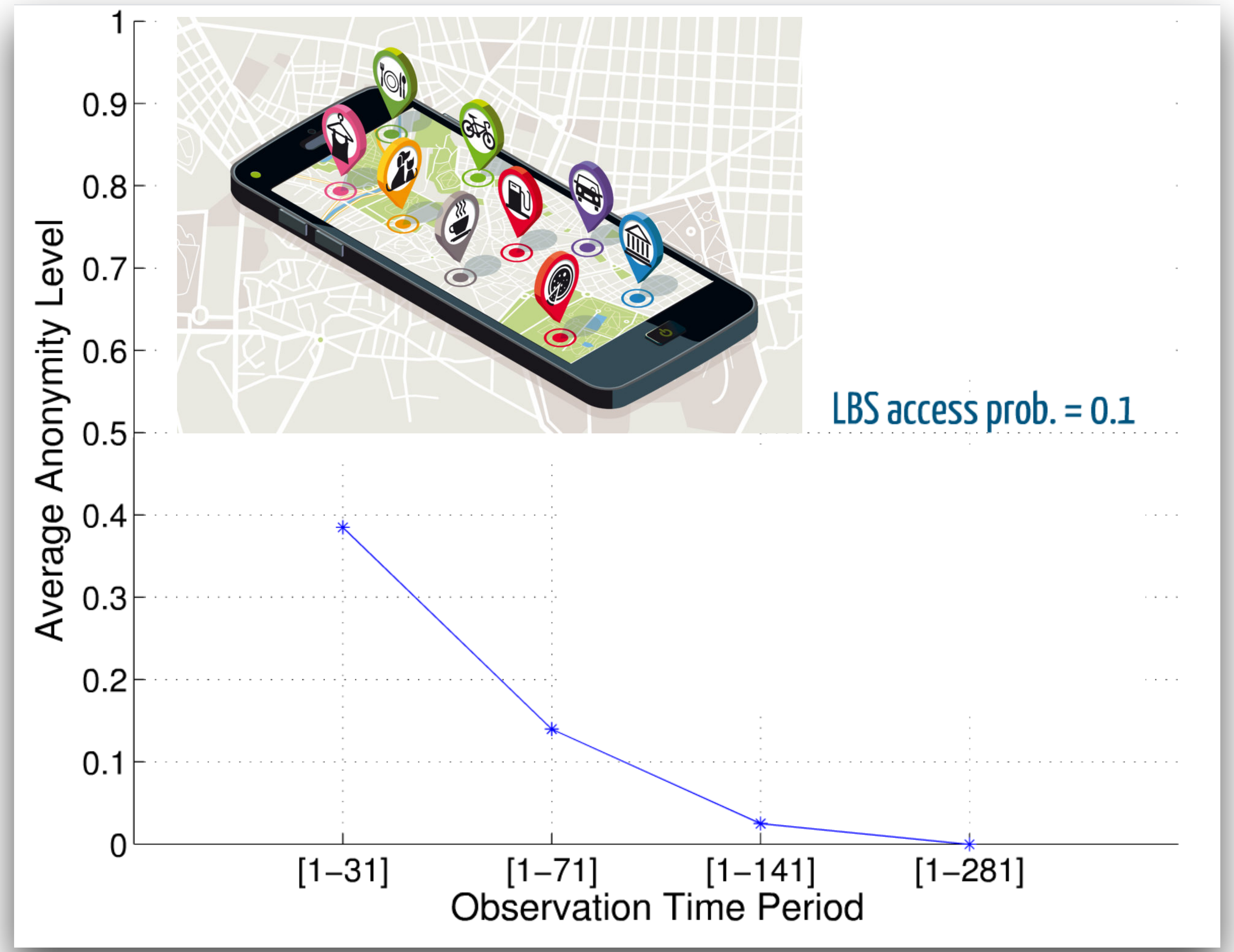


- The webpages that users 'like' on the Internet could be used to infer their personal traits
- Machine learning algorithms can be trained to find the connections

Kosinski, M., Stillwell, D. and Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior.

Inference Attacks

Identify individuals from their location traces



R. Shokri, et al., Quantifying Location Privacy, in IEEE S&P 2011

Data Sanitization

- Anonymize data, by removing personally identifying information. But, what does it mean?
- Randomize data, by perturbing the attributes in the dataset. But, what about data quality?

A huge business: Replace identities with random numbers, and remove sensitive information

Data Sanitization

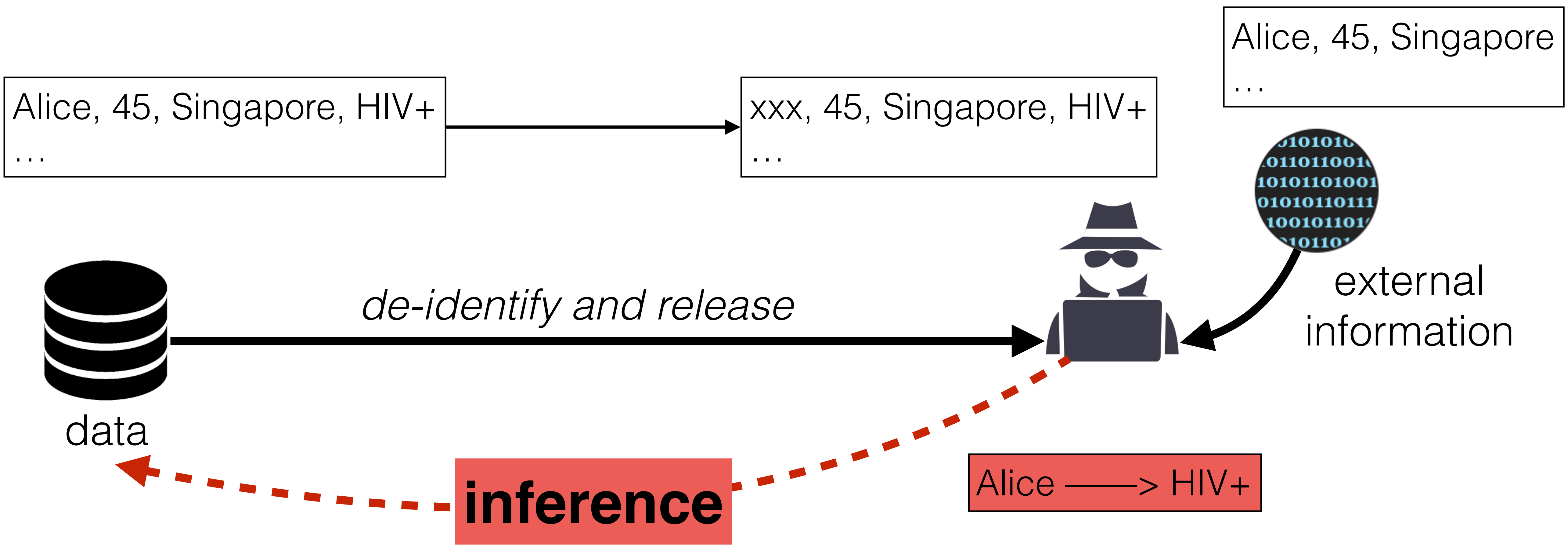
- Anonymize data, by removing personally identifying information. But, what does it mean?
- Randomize data, by perturbing the attributes in the dataset. But, what about data quality?

A huge business: Replace identities with random numbers, and remove sensitive information

What about privacy?

Bad news: Anonymized data isn't

- This is a proven **fact** in the computer science community



A. Narayanan, and V. Shmatikov, Robust de-anonymization of large datasets, in IEEE S&P 2008

WIRED AI Can Recognize Your Face Even If You're Pixelated

BUSINESS | CULTURE | GEAR | IDEAS | SCIENCE

SHARE

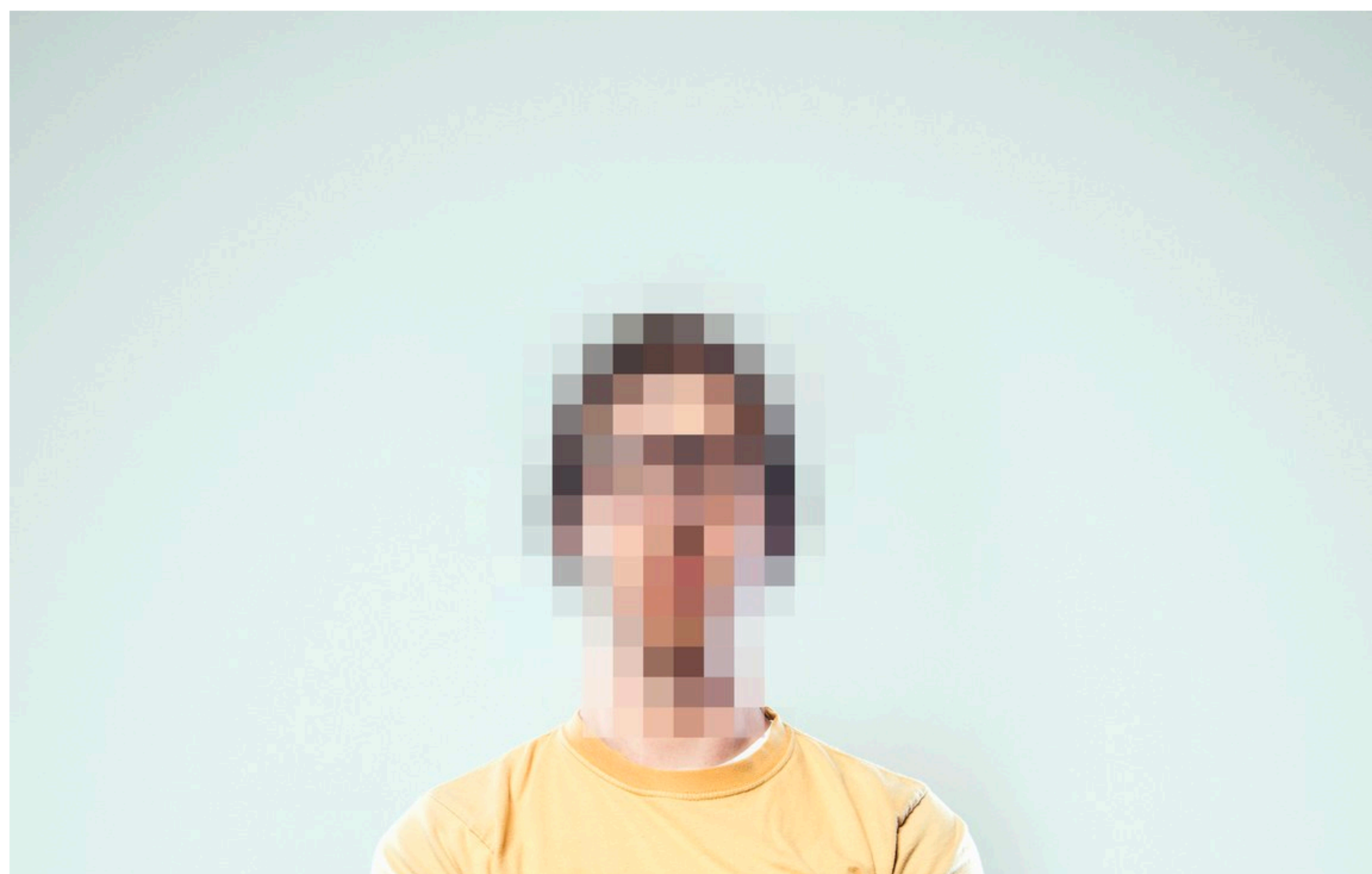
SHARE

TWEET

COMMENT 6

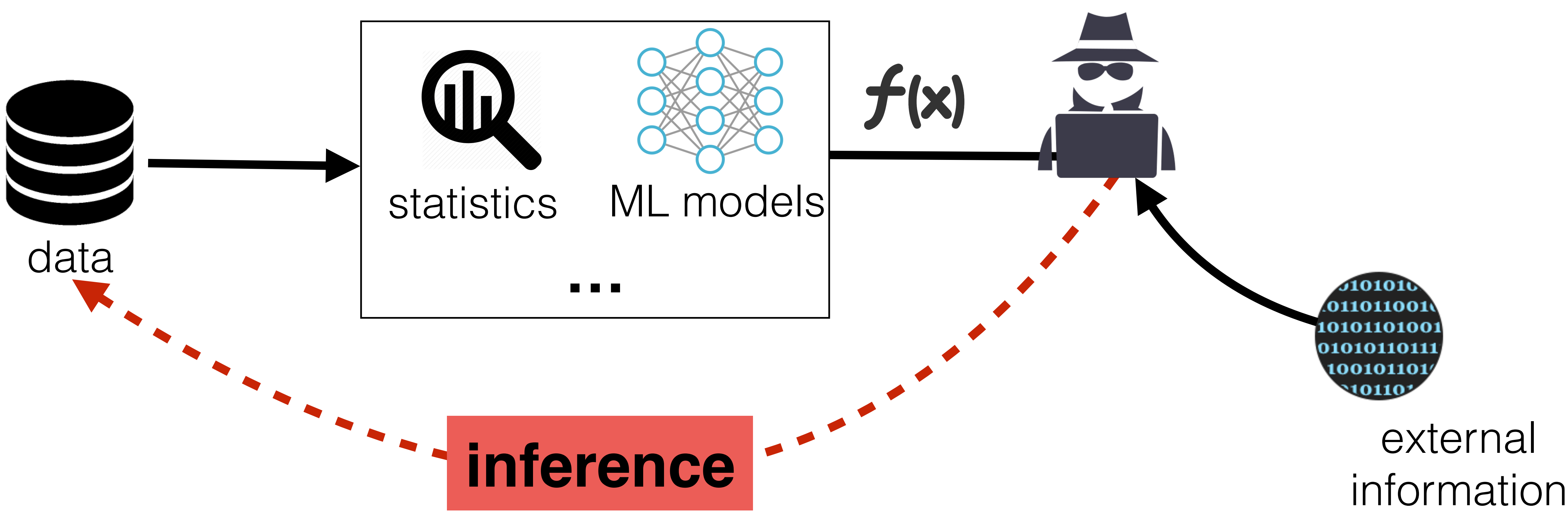
EMAIL

AI Can Recognize Your Face Even If You're Pixelated



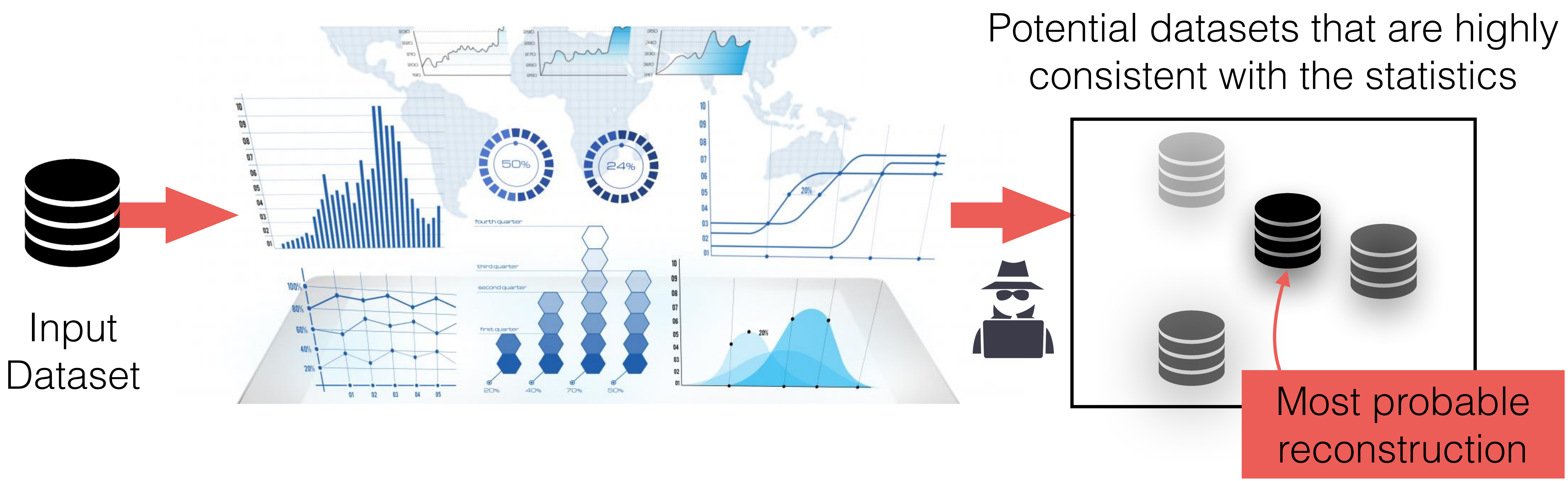
R. McPherson, R. Shokri, and V. Shmatikov, Defeating Image Obfuscation with Deep Learning, 2016

Inference Avalanche



Releasing (many) Statistics

- Can lead to identifying the records in the dataset, and eventually reconstructing the whole dataset



I Dinur, K Nissim, Revealing information while preserving privacy, PODS 2003
N. Homer, et al., Resolving individuals contributing trace amounts of DNA to highlycomplex mixtures using high-density SNP genotyping microarrays, in PLoS Genetics, vol. 4, no. 8, 2008.

Forbes

3/12/2010 @ 12:35PM | 3,837 views

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Back to the Future: NIH to Revisit Genomic Data-Sharing Policy

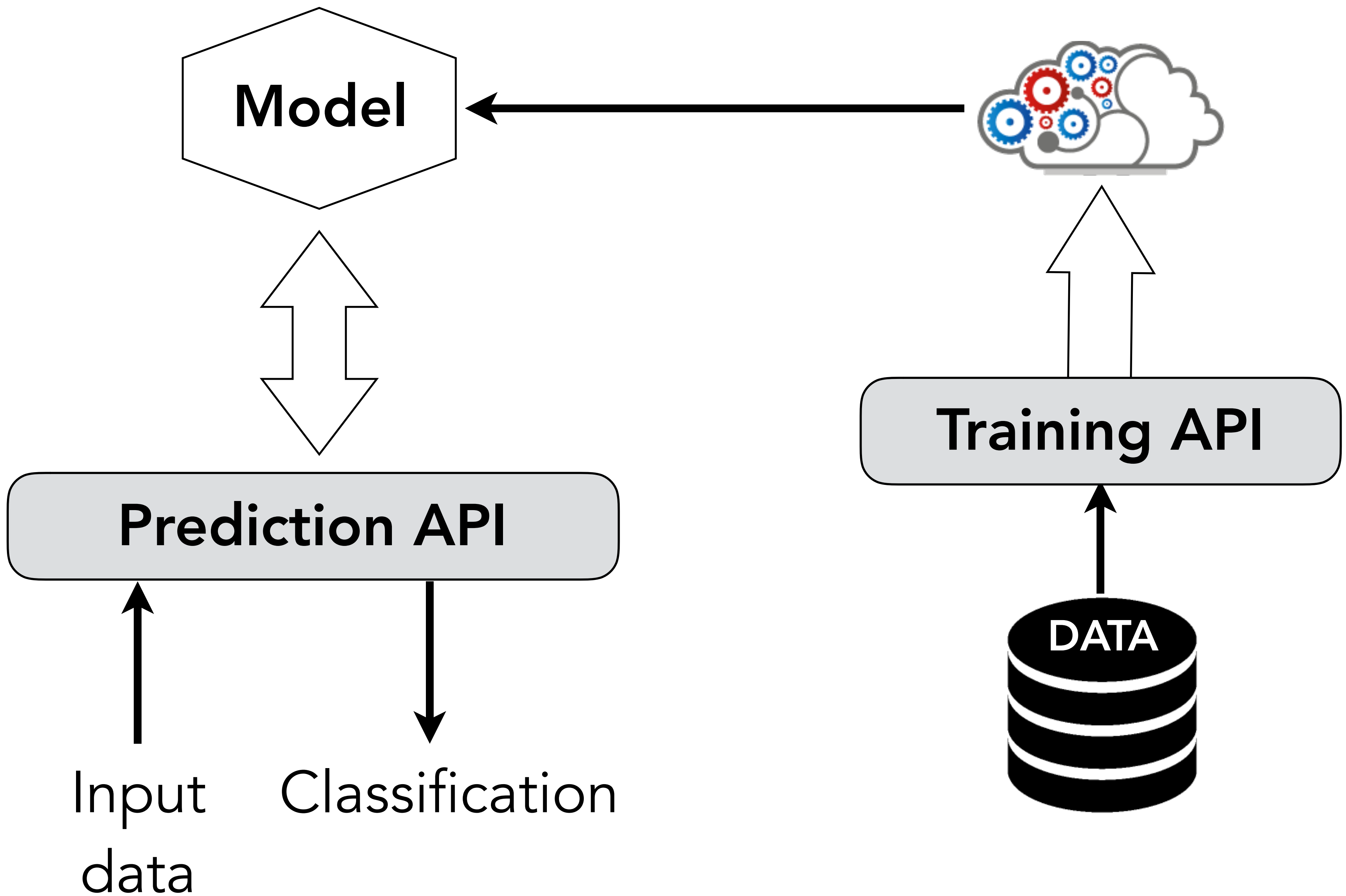
Posted by Dan Vorhaus on October 28, 2009

As first reported by GenomeWeb, last week the NIH issued a "Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data." Although the title doesn't exactly trip off of the tongue, the NIH's announcement provides an opportunity to review where we are and where we have already been when it comes to genomic data-

Breaches Lead to Push to Protect Medical Data

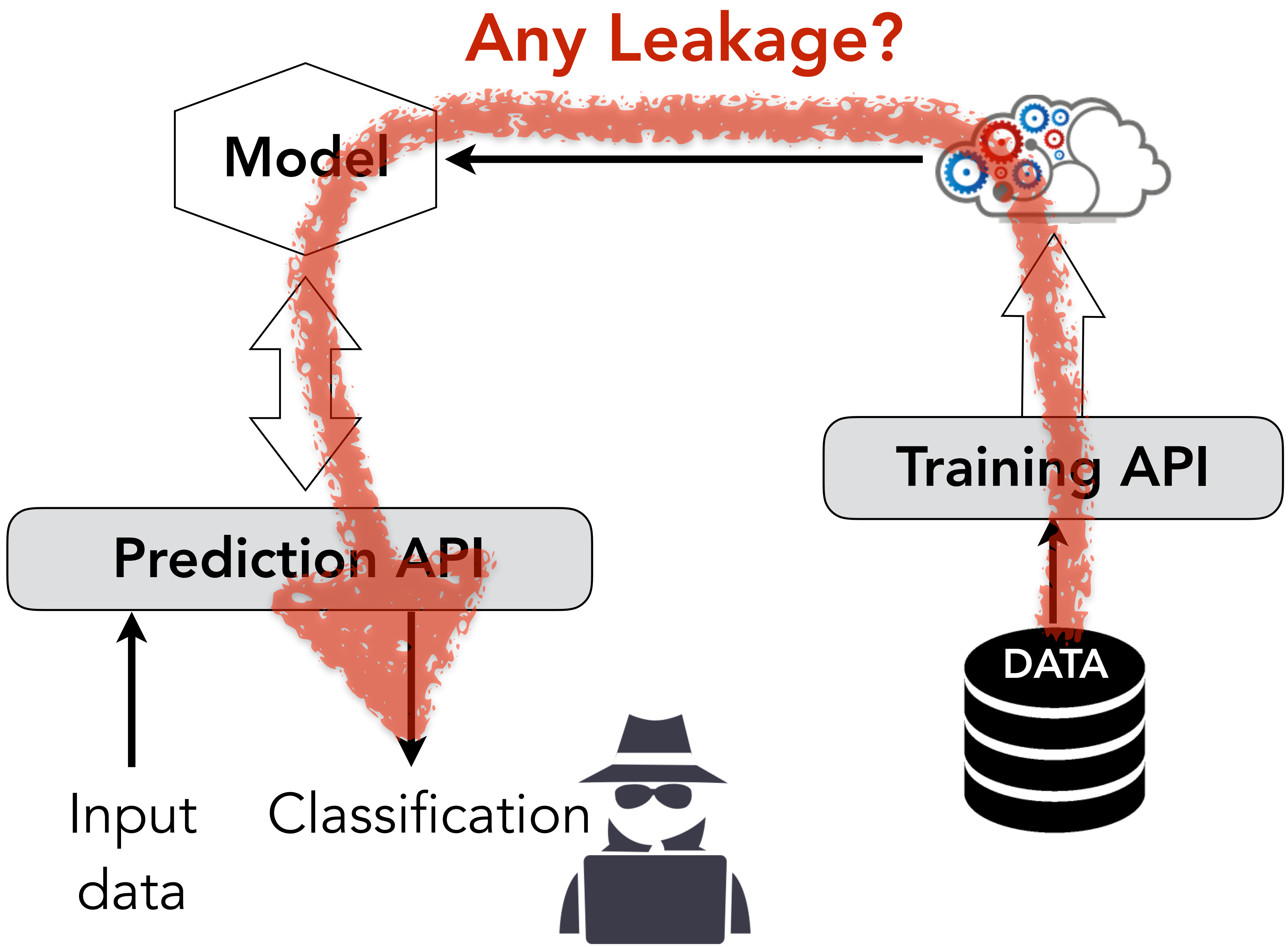
By MILT FREUDENHEIM MAY 30, 2011

Machine Learning



R. Shokri, Membership Inference Attacks against Machine Learning Models, in IEEE S&P 2017

Machine Learning



R. Shokri, Membership Inference Attacks against Machine Learning Models, in IEEE S&P 2017

Machine Learning

Any Leakage?

Membership Inference Attack

Google Cloud Platform

Customer classification using their purchase data

Attack accuracy: 93%

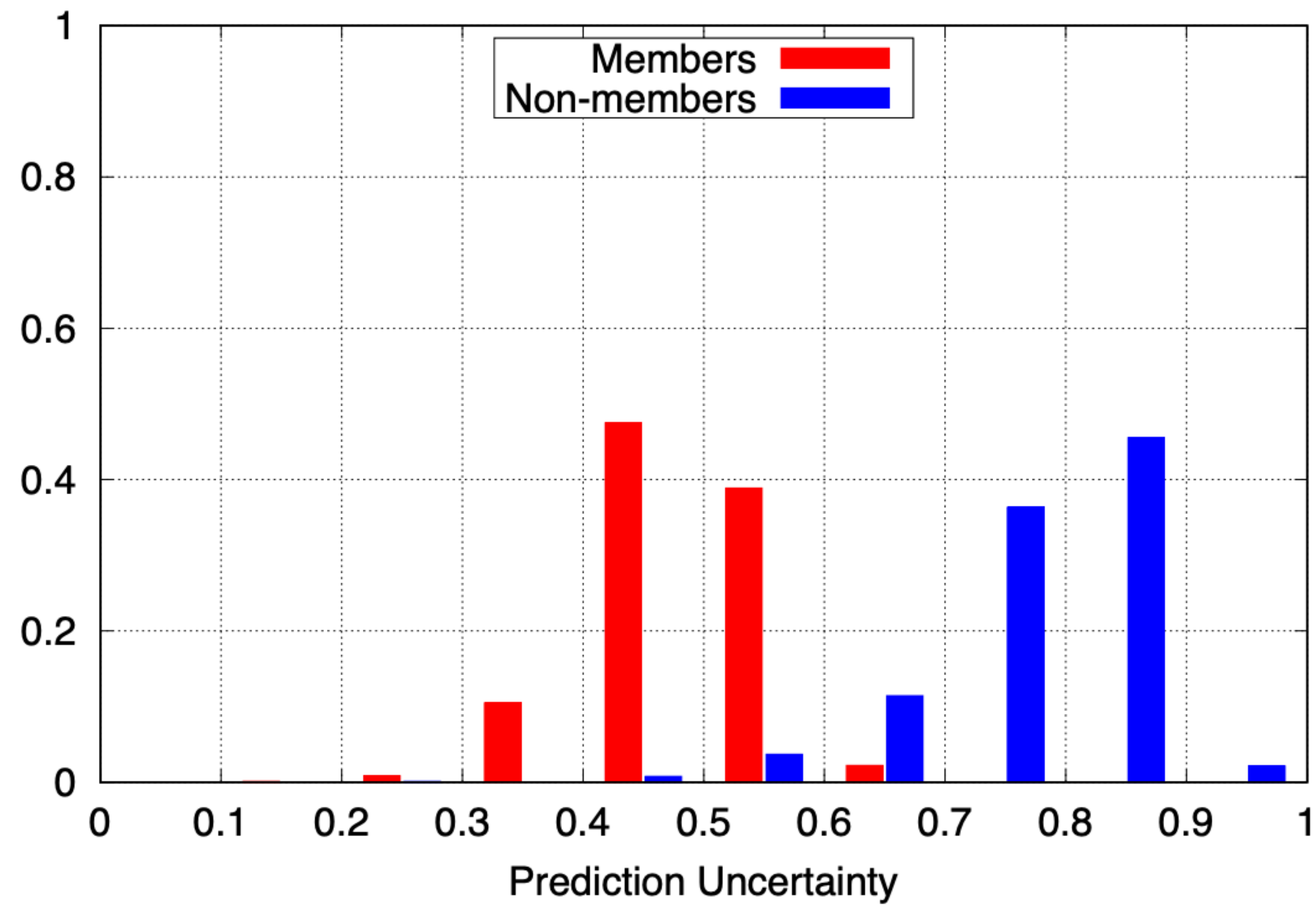
Determine if a data record was part of the training set

ML models leak significant amount of info

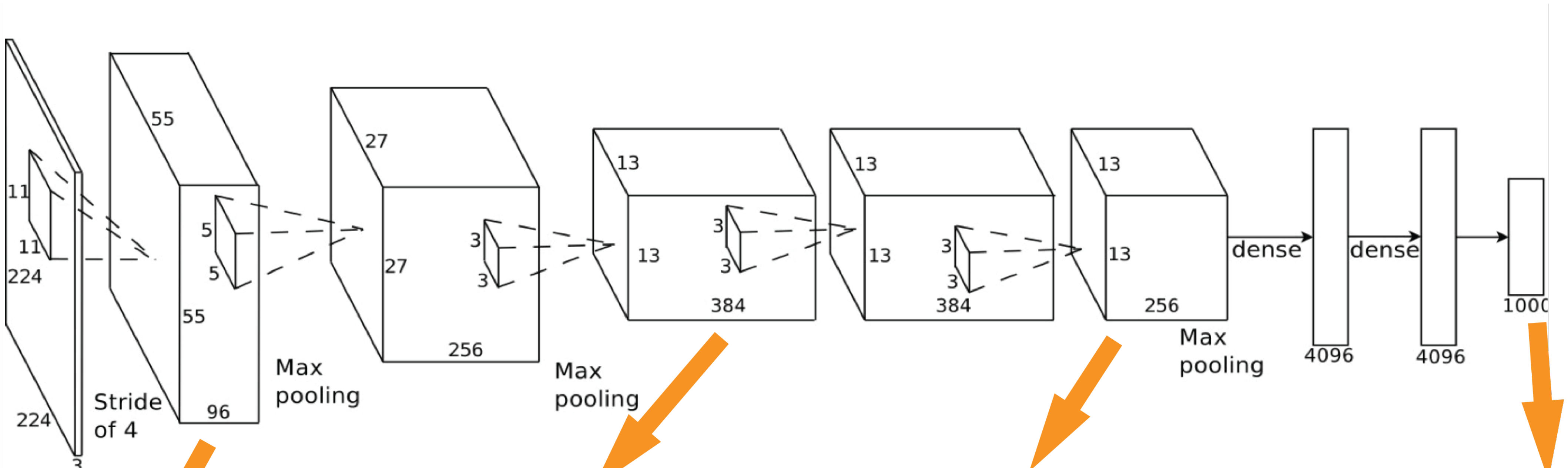
R. Shokri, Membership Inference Attacks against Machine Learning Models, in IEEE S&P 2017

Members are Distinguishable

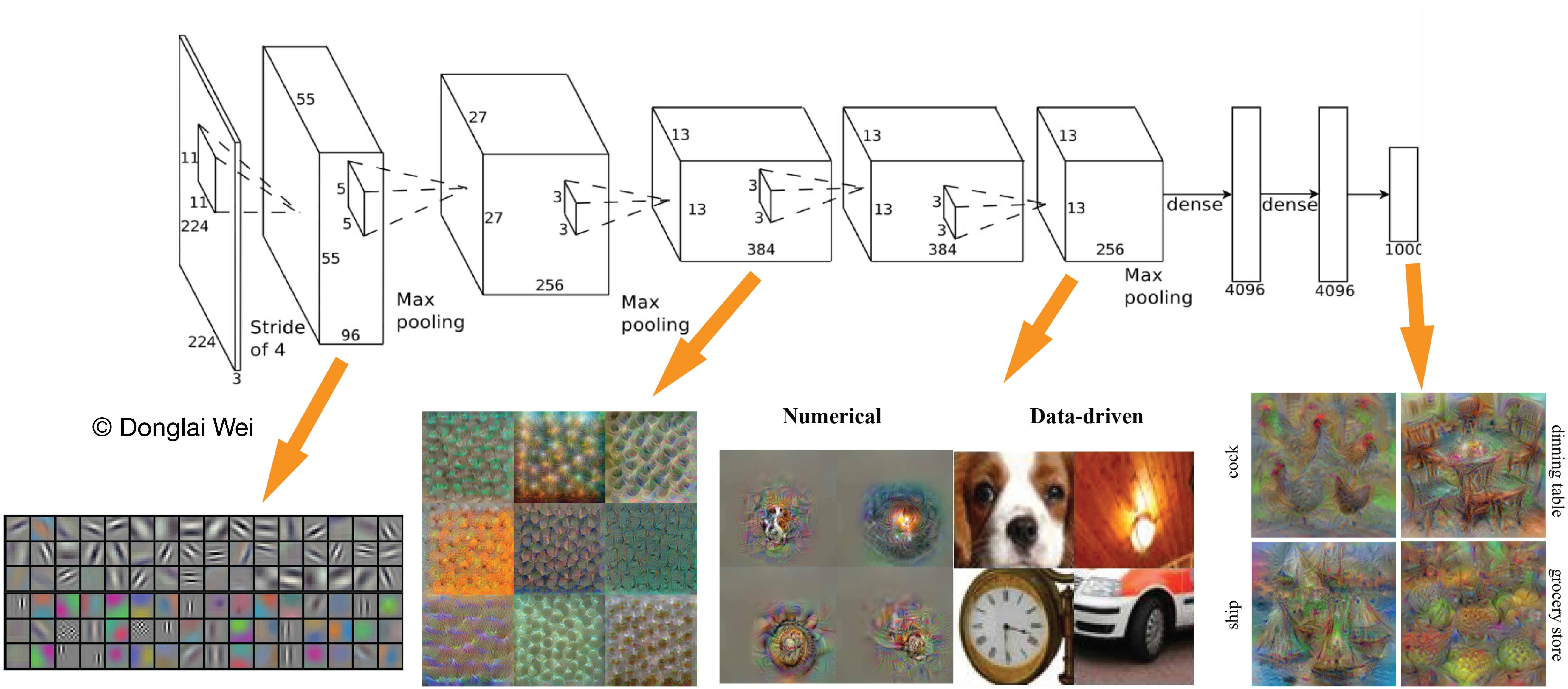
- Model's behavior is different for members of the training set vs non-members



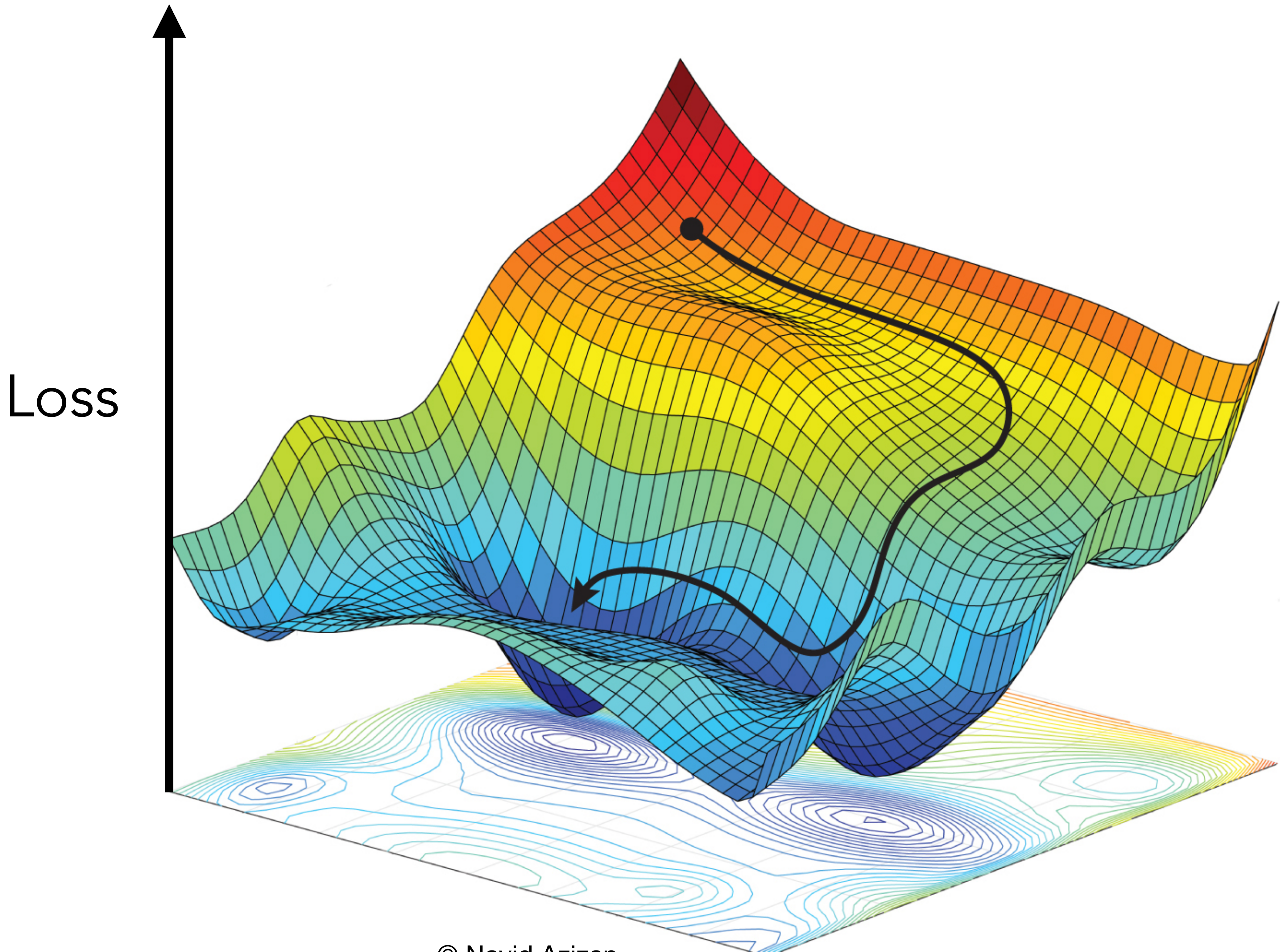
White-box Access to Parameters



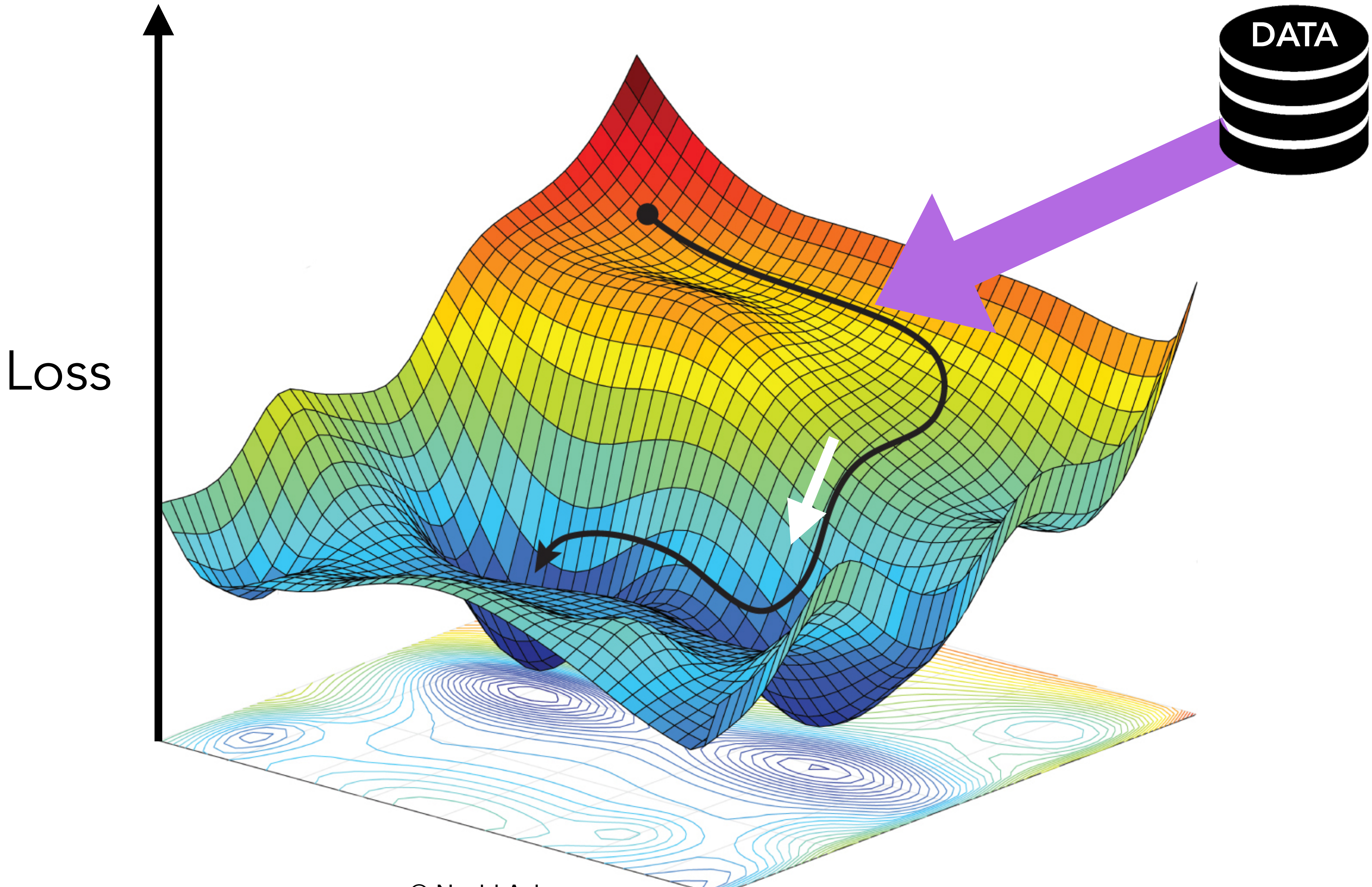
White-box Access to Parameters



Stochastic Gradient Descent

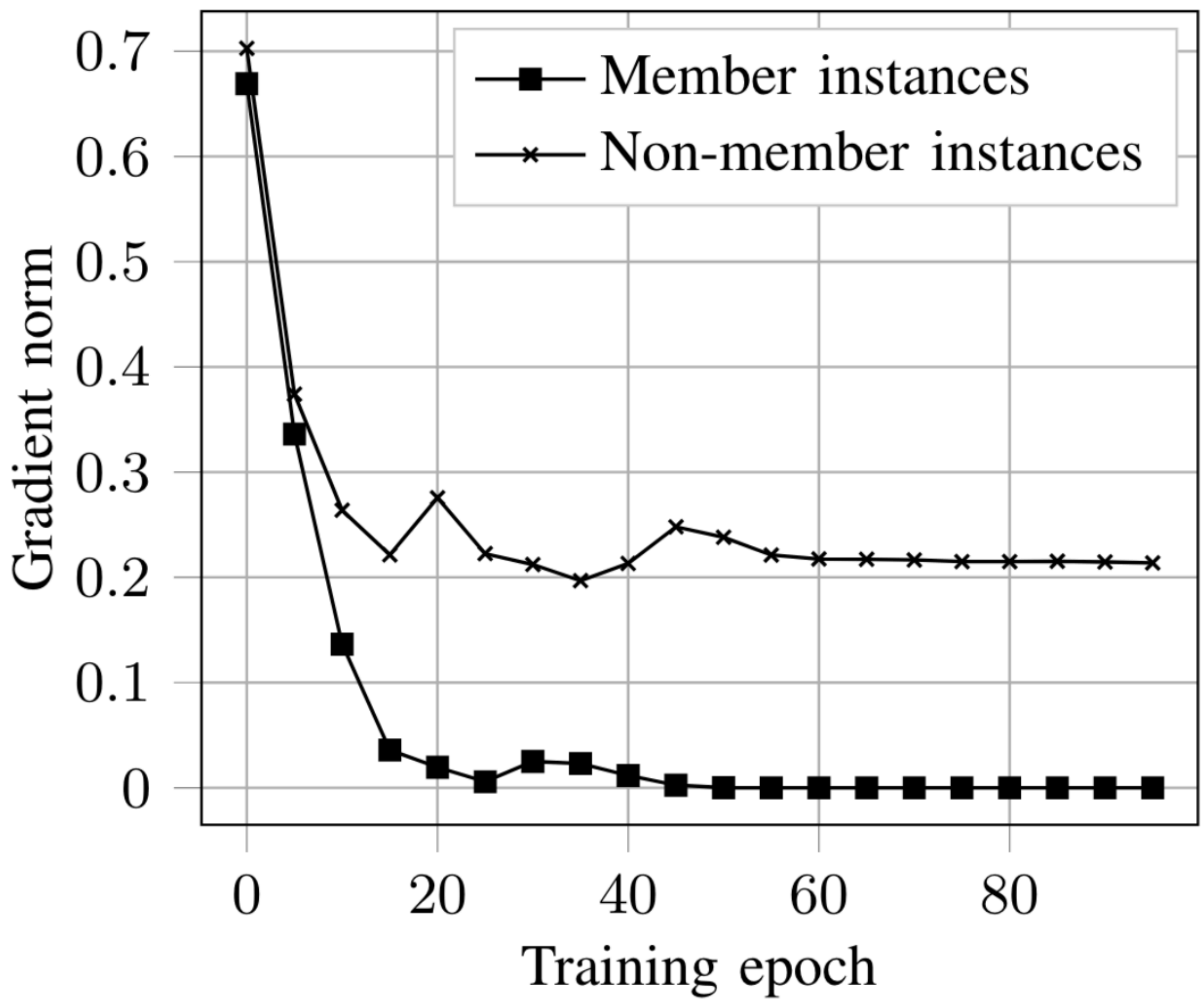


Stochastic Gradient Descent



© Navid Azizan

Gradient of Loss on Members vs. Non-members



Generalizability and Privacy

in the white-box setting

Pre-trained Target Model				Attack Accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	$\boxed{100\%}$	$\boxed{82\%}$	67.7%	67.7%	$\boxed{74.3\%}$
Texas100	Fully Connected	81.6%	52%	63.0%	63.3%	68.3%
Purchase100	Fully Connected	100%	80%	67.6%	67.6%	73.4%

High generalizability
(Best available models)

Low privacy
(Significant leakage
through parameters)

M. Nasr, R. Shokri, and A. Houmansadr, Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, in IEEE S&P 2019

Generalizability and Privacy

in the white-box setting

Pre-trained Target Model				Attack Accuracy		
Dataset	**Architecture**	**Train Accuracy**	**Test Accuracy**	**Black-box**	**White-box (Outputs)**	**White-box (Gradients)**
CIFAR100	Alexnet	99%	44%	74.2%	74.6%	75.1%
CIFAR100	ResNet	89%	73%	62.2%	62.2%	64.3%
CIFAR100	DenseNet	100%	82%	67.7%	67.7%	74.3%
Texas100	Fully Connected	81.6%	52%	63.0%	63.3%	68.3%
Purchase100	Fully Connected	100%	80%	67.6%	67.6%	73.4%

Large capacity

High generalizability
(Best available models)

Low privacy
(Significant leakage through parameters)

M. Nasr, R. Shokri, and A. Houmansadr, Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, in IEEE S&P 2019

Models as Personal Data?

- GDPR: Personal data are any information which are (directly or indirectly) related to an identified or identifiable natural person.

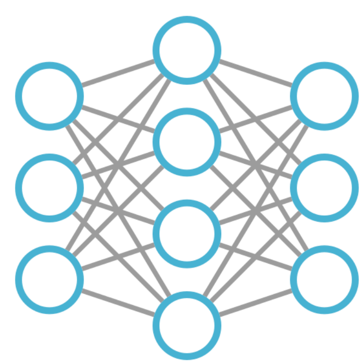
- The models enable identifying whose data has been part of the training data
- They can also be used to partially reconstruct training data

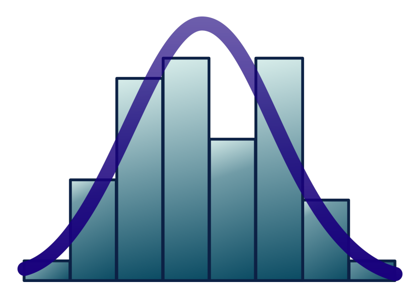
Adapt Mechanisms to Use-Cases

Privacy-Preserving Computation

Outsourcing

Collaborative Computation





Data Analytics and Machine Learning

Data Exploration and Visualization

Privacy-Preserving Computation

- The result of the computation does not leak significant amount of information about the input data
- If no information is leaked, the outcome is useless
- What information should we hide and what should we share?
- **Differential Privacy**
    - Hide individuals' information
    - Allow information sharing about global patterns in the data

C Dwork, F McSherry, K Nissim, A Smith, Calibrating noise to sensitivity in private data analysis, in Theory of cryptography 2006

Privacy-Preserving Data Synthesis

Sharing without Sharing

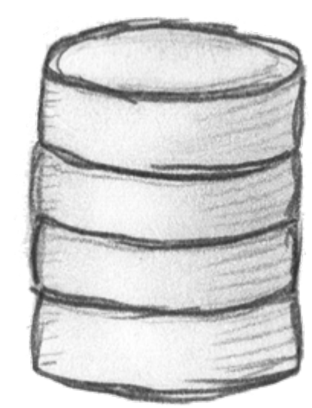
- Given real (sensitive) data, generate synthetic data that satisfy differential privacy, and are also useful (preserve utility)

- same format
- similar - but not same! - features

Real Data

generative model

Differential Privacy

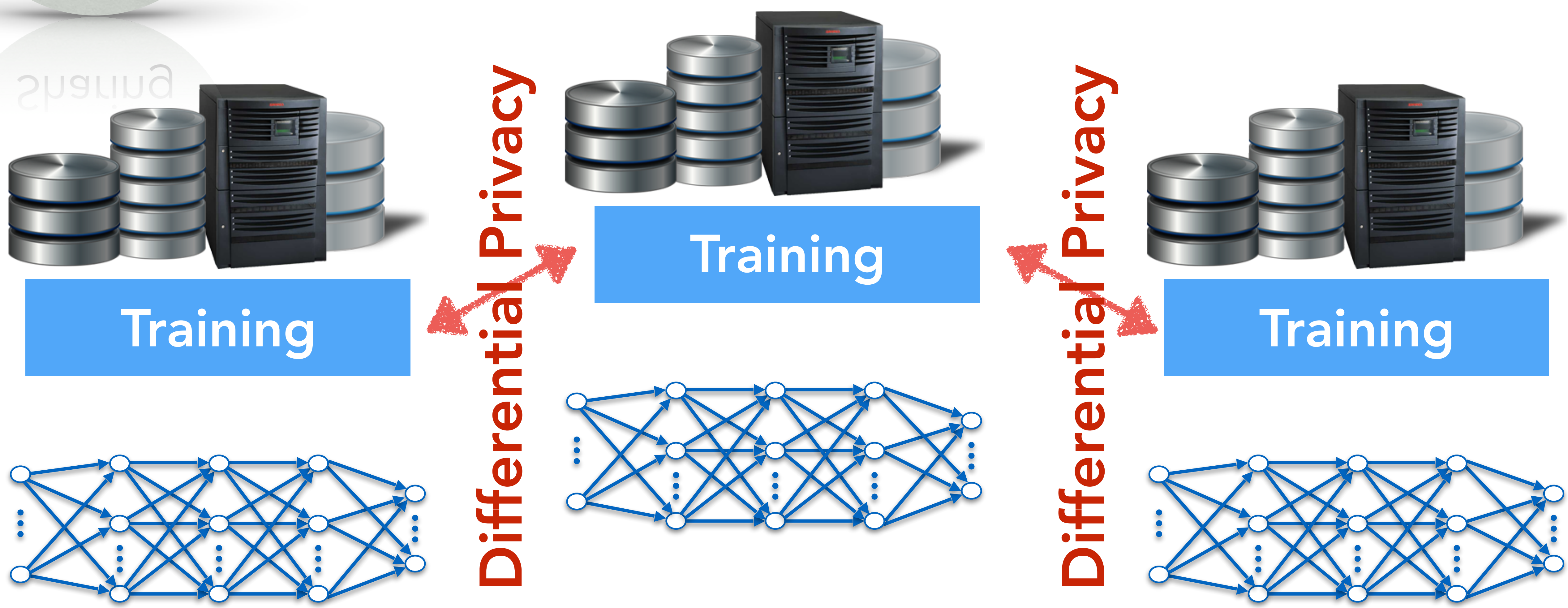


Synthetic Data

V. Bindschaedler, R. Shokri, and C. Gunter, Plausible Deniability for Privacy-Preserving Data Synthesis, in VLDB 2017

Collaborative Learning (a.k.a. Federated Learning)

Sharing without Sharing



exchange DP **hints** about **models** during training

R. Shokri and V. Shmatikov, Privacy-Preserving Deep Learning, in CCS 2015
HB McMahan, Communication-efficient learning of deep networks from decentralized data, 2016