Privacy in the Year 2027

Vitaly Shmatikov

Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

ARTIFICIAL INTELLIGENCE
IBM WATSON, MetaMind, Numenta, ai-one, Cycorp, Microsoft Research, nara, Reactor, SCALED INFERENCE

DEEP LEARNING
vicarious, Vision Factory, facebook, Baidu IDL, Google, ersatz, SKYMIND, SignalSense

MACHINE LEARNING
rapidminer, context relevant, 0xdata H2O, DATARPM, LIftIgniter, SPARKBEYOND, Azure ML, yhat, wise.io, Sense, GraphLab, Alpine, Ayasdi

NLP PLATFORMS
cortical.io, idibon, LUMINOSO, wit.ai, MaluubA

PREDICTIVE APIS
AlchemyAPI, MINDOPS, Google, bigml, indico, ALGORITHMIA, Expect Labs, PredictionIO

IMAGE RECOGNITION
clarifai, MADBITS, DNNresearch, DEXTRO, VISENZE, lookflow

SPEECH RECOGNITION
GRIDSPACE, popUP archive, NUANCE

RETHINKING ENTERPRISE

SALES
Preact, AVISO, RelateIQ, NGDATA, CLARABRIDGE, FRAMED, infer, ATTENTIVITY, causata

SECURITY / AUTHENTICATION
CROSSMATCH, conjur, EYEVERIFY, BITSIGHT, CYLANCE, bionym

FRAUD DETECTION
sift science, SOCURE, ThreatMetrix, feedzai, Brighterion, VERAFIN

HR / RECRUITING
TalentBin, entelo, predikt, Connectifier, gild, hiQ

MARKETING
brightfunnel, bloomreach, CommandIQ

PERSONAL ASSISTANT

INTELLIGENCE TOOLS
ADATAO, Palantir, Quid, Digital Reasoning, FirstRain

Tech 2015: Deep Learning And Machine Intelligence Will Eat The World

ADTECH
METAMARKETS, dstillery, rocketfuel, YieldMo, ADBRAIN

OIL AND GAS
kaggle, AYASDI, TACHYUS, biota, Flutura

FINANCE
Cocorain, newsle, ARRIA, SAILTHRU, wavii, Owlin, NarrativeScience, Quid, Summly, Prismatic, ai, Affirm, inVenture, zestfinance, BILLGUARD, LendUp, LendingClub, Kabbage

PHILANTHROPIES
DataKind, thorn, The DATA GUILD

AUTOMOTIVE
Google, Continental, TESLA, MOBILEYE, CRUISE

DIAGNOSTICS
enlitic, 3SCAN, lumiata, ENLITIC

RETAIL
BAY SENSORS, PRISM SKYLABS, celect, euclid

MEDICAL
Parzival, transcriptic, ZEPHYR HEALTH, Genescient, grand round table, bina, TUTE GENOMICS

MICROSCAN, IVISYS, BOULDER IMAGING, Brevia, DiligenceEngine

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence, APX, blippAR, META, layar

GESTURAL COMPUTING
THALMICLABS, omek, FLUTTER, LEAP MOTION, eyeSight, 3Gear, GestureTek, nod

ROBOTICS
intel, LIQUID ROBOTICS, SoftBank, iRobot, Boston Dynamics, jibo, anki

EMOTIONAL RECOGNITION
affectiva, BEYOND VERBAL, EMOTIENT, cogito

SUPPORTING TECHNOLOGIES

HARDWARE
NVIDIA, XILINX, QUALCOMM, NERVANA SYSTEMS, TERADEEP, Artificial Learning, rigetti

DATA PREP
TRIFACTA, Paxata, tamr, Alation

DATA COLLECTION
diffbot, kimono, CrowdFlower, Connotate, WorkFusion, import.io

www.shivonzilis.com/machineintelligence

Bloomberg BETA

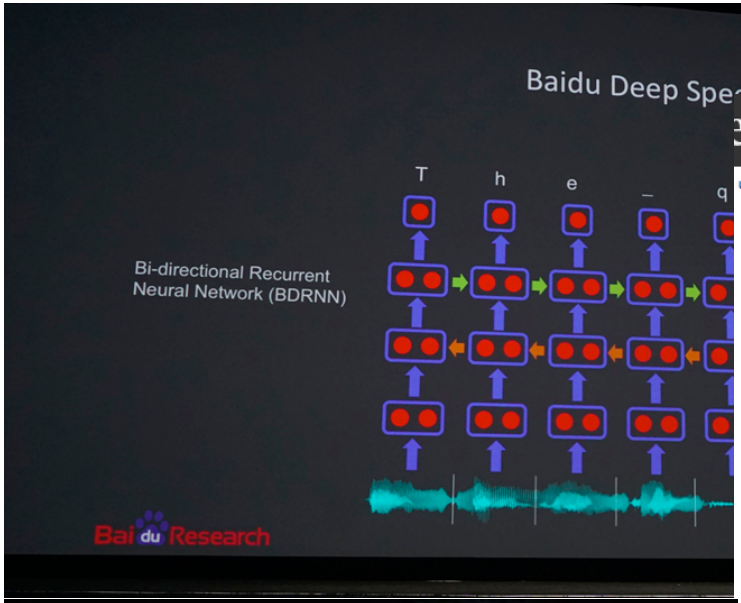
rbes ▾ | New Posts +1 | Most Popular | Lists | Video | Subscribe to Forbes

Sign up

Robert Hof
Contributor

TECH 12/18/2014 @ 9:00AM | 48,377 views

Baidu Announces Breakthrough In Speech Recognition, Claiming To Top Google And Apple



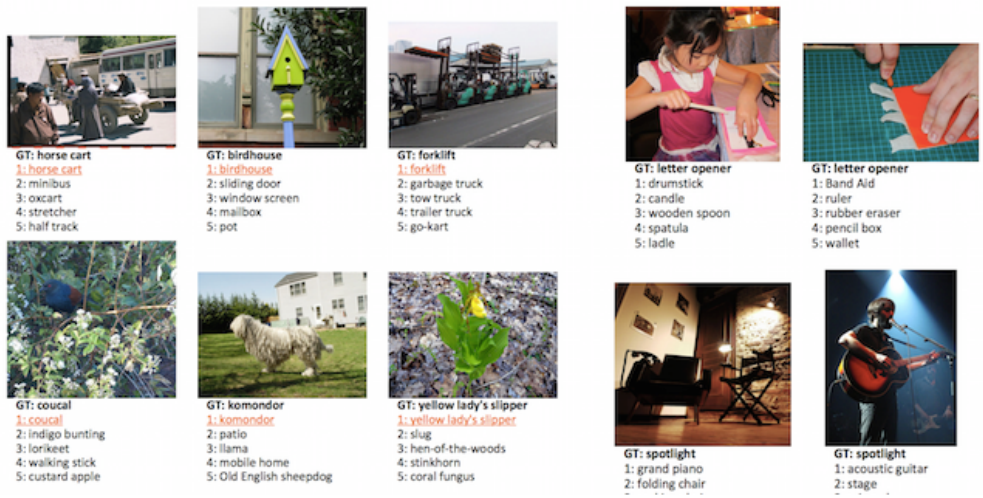
rbes ▾ | New Posts +1 | Mos

up

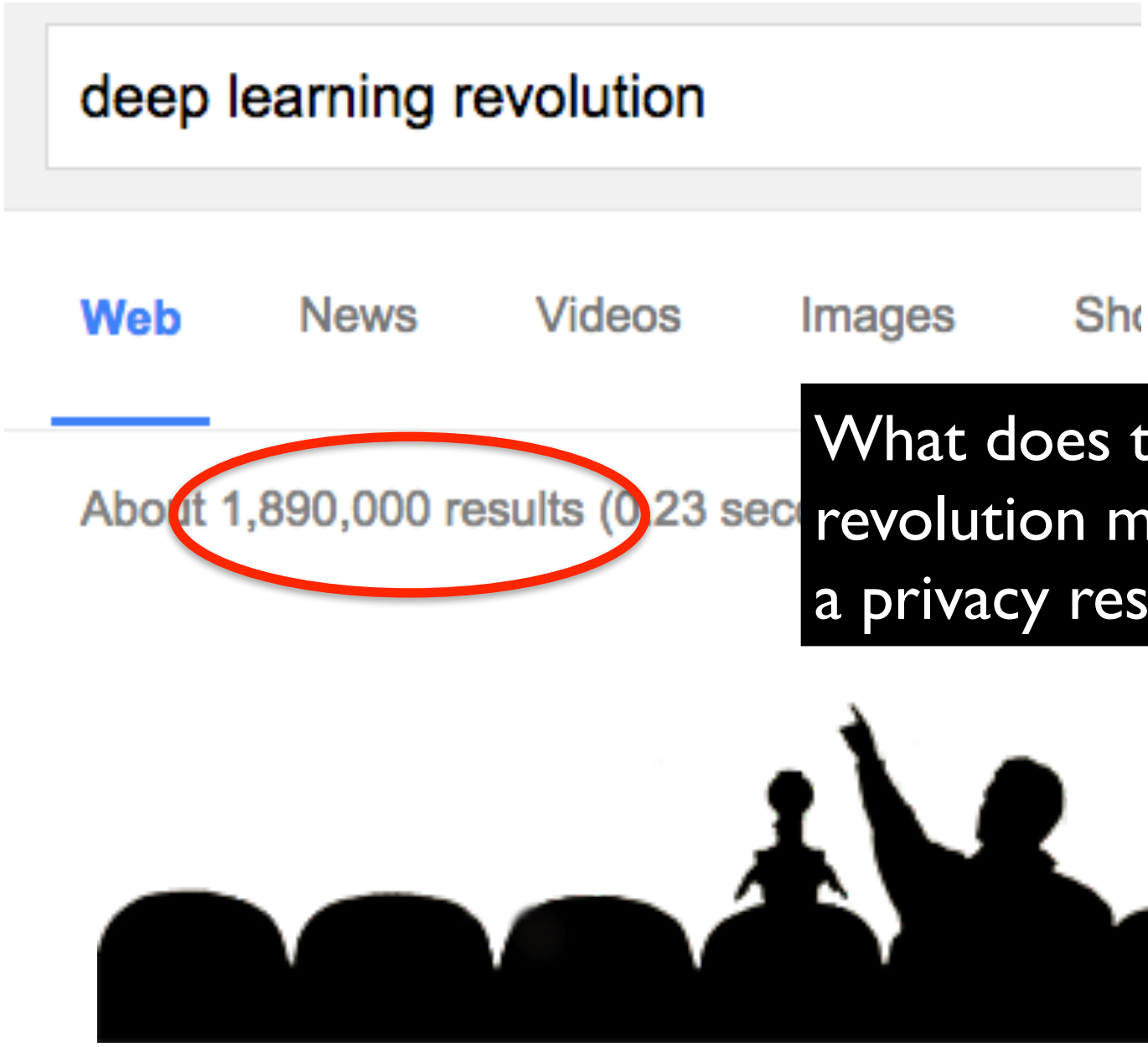
Michael Thomsen
Contributor

TECH 2/19/2015 @ 1:06PM | 4,996 views

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

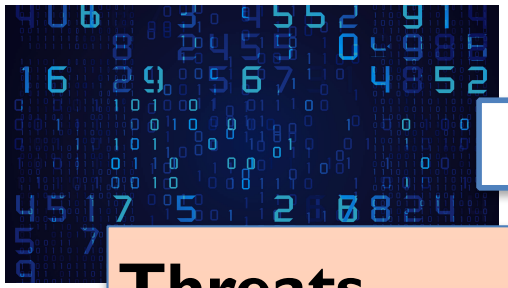


Images used to test Microsoft's deep learning AI. Image via Microsoft.


deep learning revolution
Web
News
Videos
Images
About 1,890,000 results (0.23 sec
What does this revolution mean for a privacy researcher?

2014

Users' data



Services

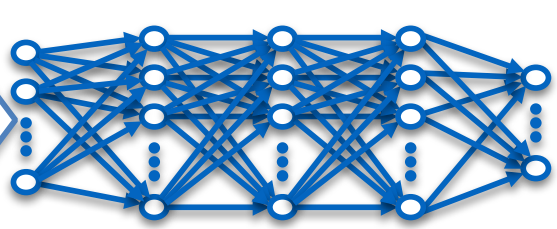
Threats

- Collection of sensitive personal data
- Anonymization and re-identification
- Inference attacks
- Side channels

2020

Users' data

Machine learning



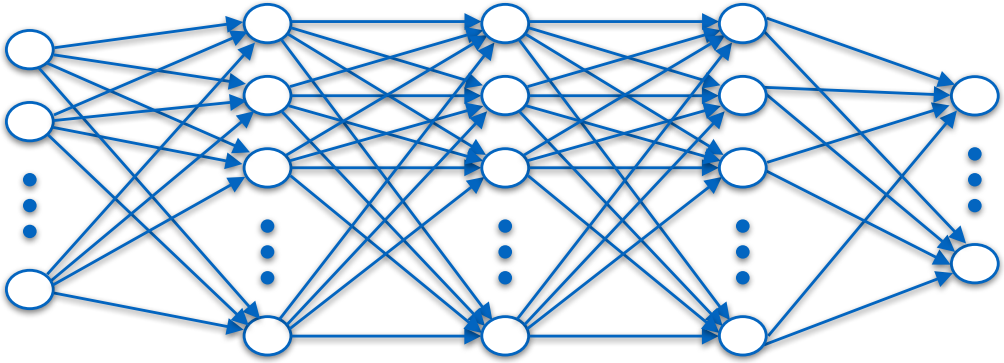
Services

Do trained models leak sensitive data?

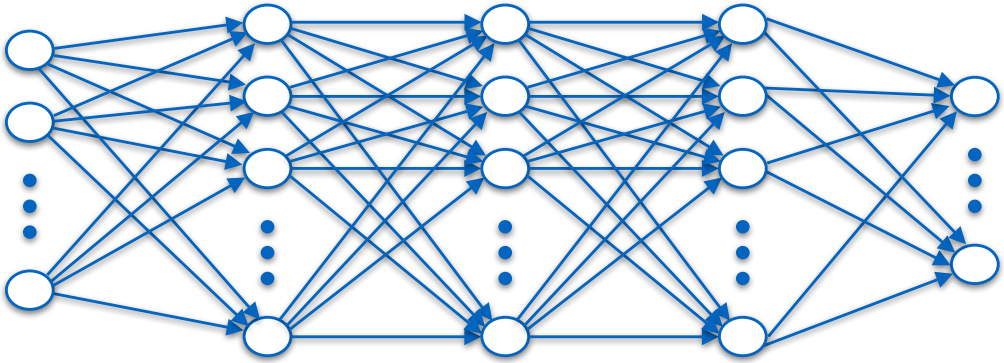
Is it possible to train a "good" model while respecting privacy of training data?

Is it possible to keep the model itself private?

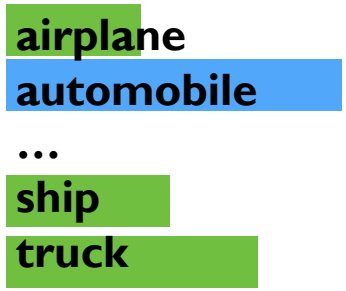
Typical Task: Classification



Query

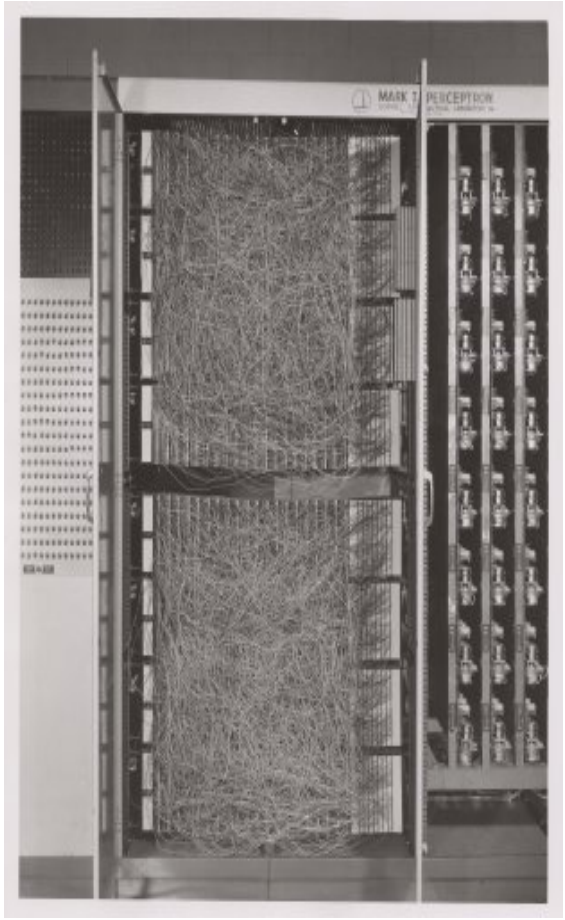


Classification result

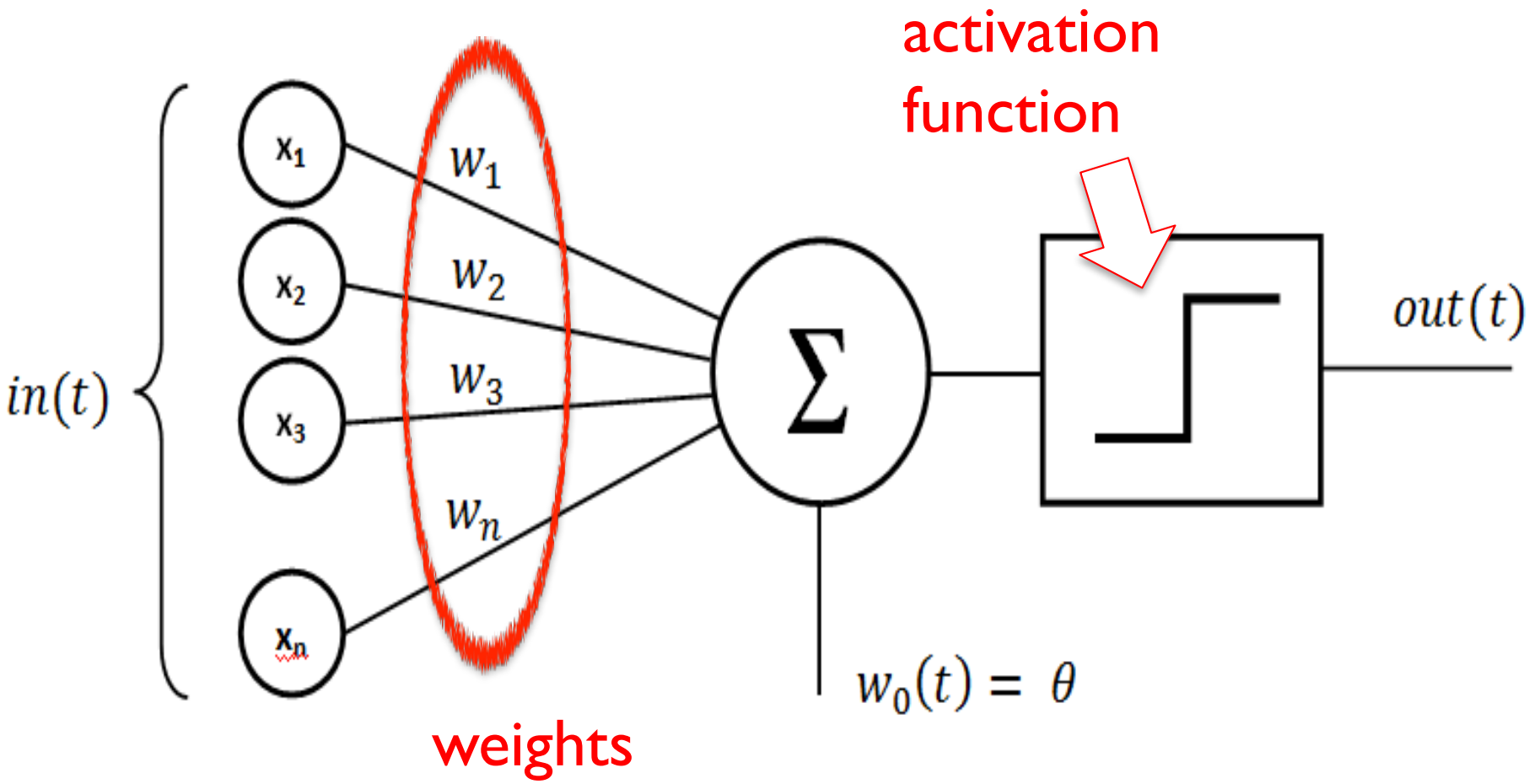


Perceptron (1957, Cornell)

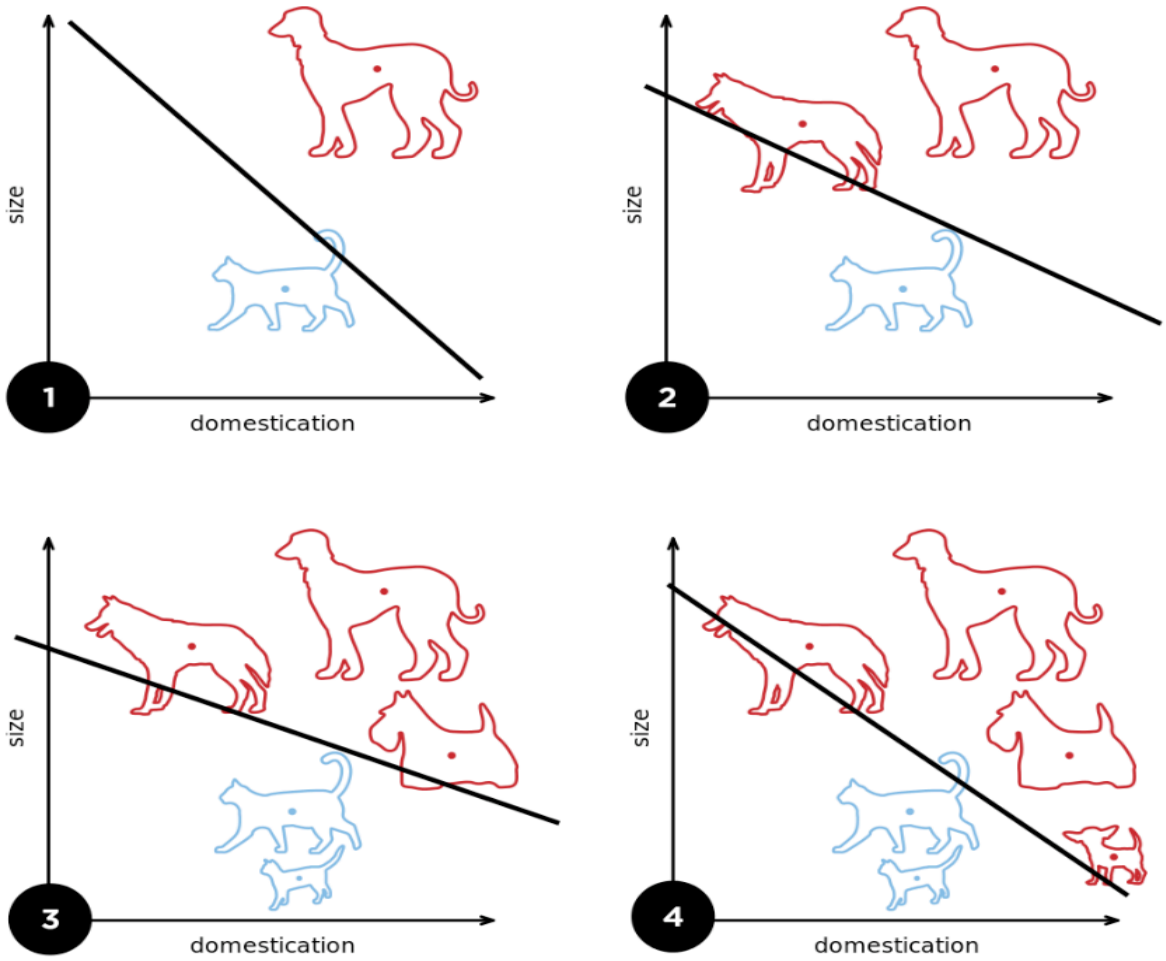
- 400 pixel camera
- Designed for image recognition
- "Knowledge" encoded as weights in potentiometers (variable resistors)
- Weights updated during learning performed by electric motors



Perceptron

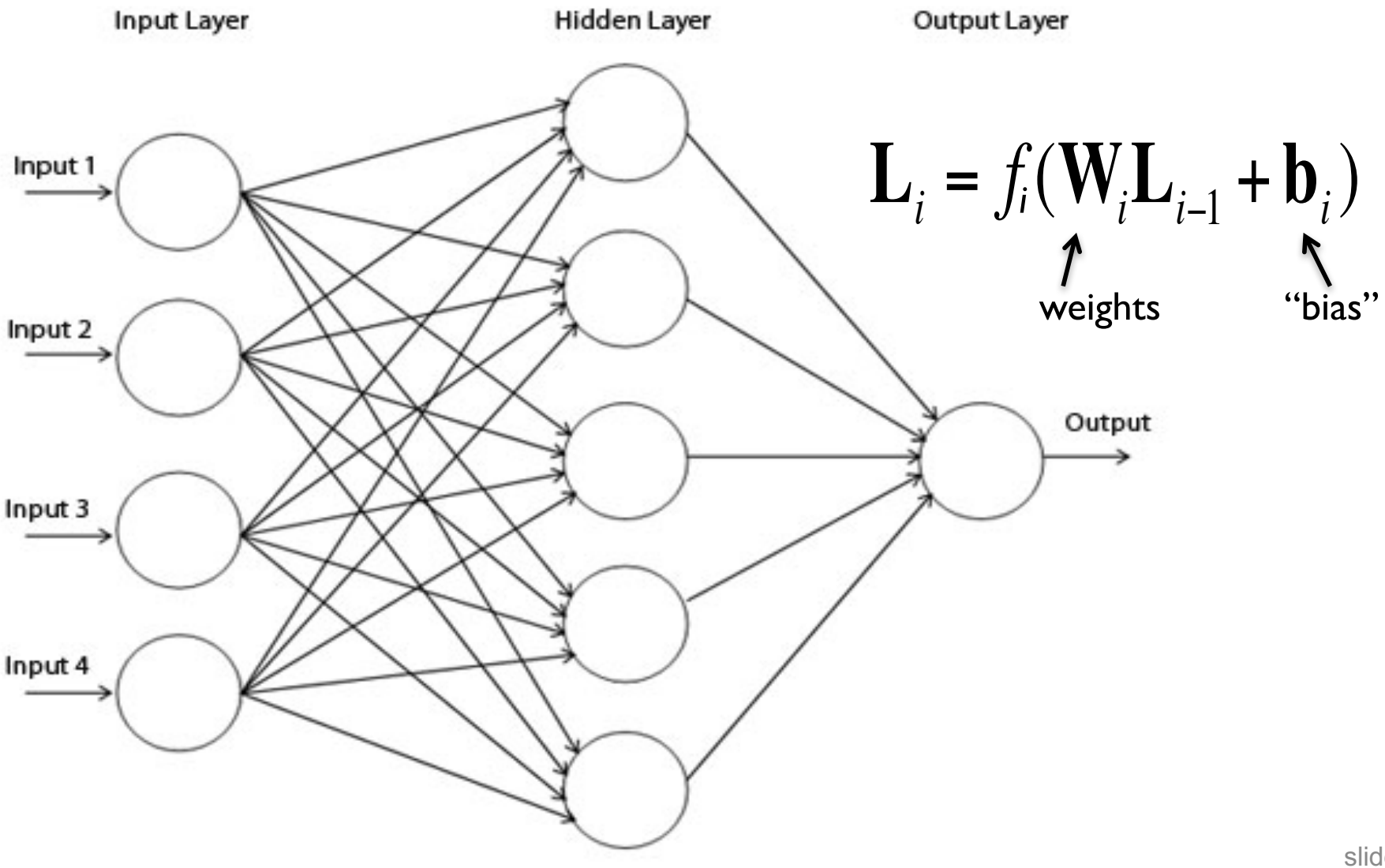

activation
function
$in(t)$
x_1
x_2
x_3
x_n
w_1
w_2
w_3
w_n
Σ
$out(t)$
$w_0(t) = \theta$
weights

What Perceptron Can Learn

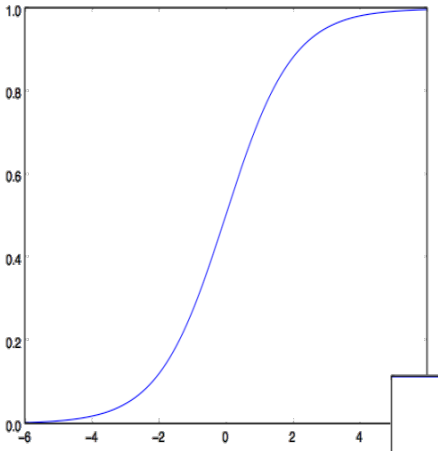


Binary classifier, linearly separable patterns

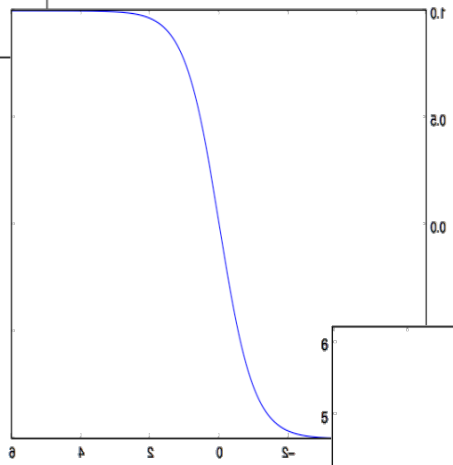
Feedforward Neural Networks



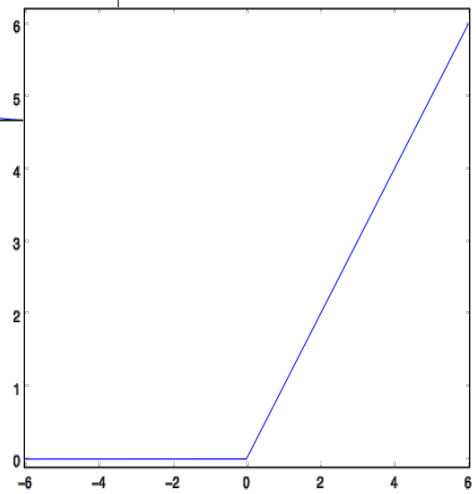
Activation Functions



Sigmoid / logistic: $f(x) = 1 / (1-e^{-x})$



Tanh: $f(x) = \tanh(x)$



ReLU: $f(x)=\max(x,0)$

Universal Approximation Theorem

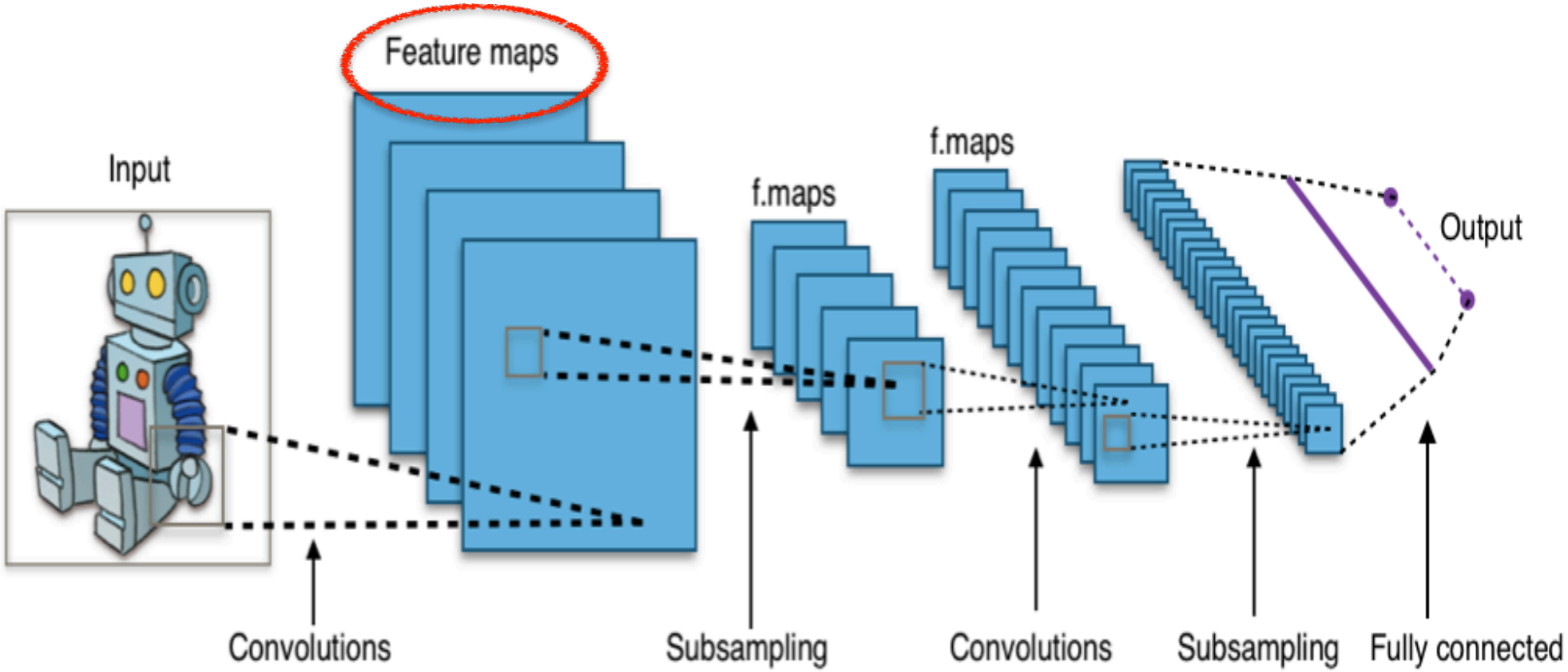
- Multilayer perceptron with a single hidden layer and linear output layer can approximate any continuous function on a compact subset of R^n to within any desired degree of accuracy

  – Under some assumptions about activation functions

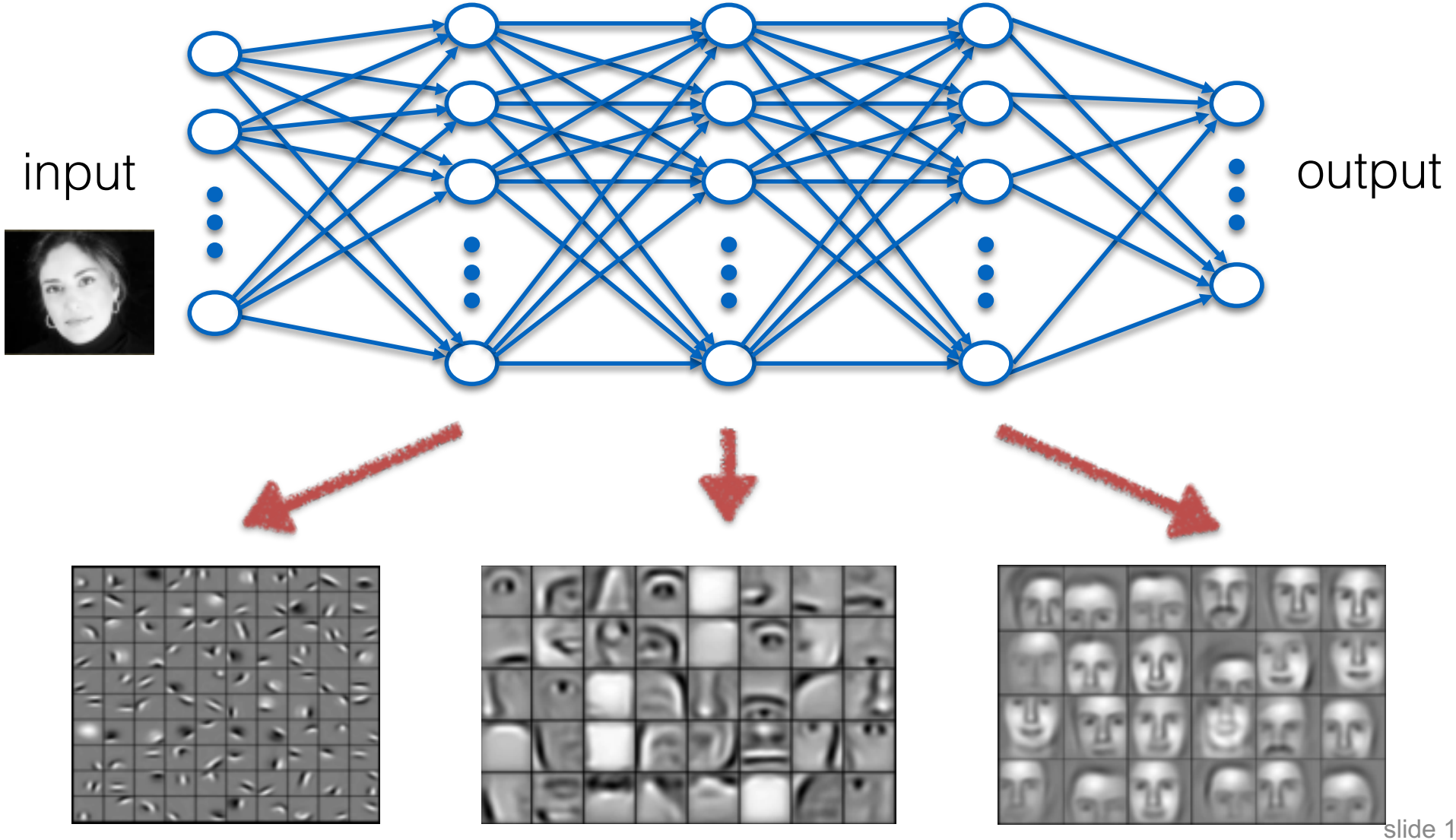
Why Deep Neural Networks

- Functions representable with a deep network can require exponential number of hidden units with a shallow (single hidden layer) network
- Piecewise linear networks (e.g., using ReLU) can represent functions that have a number of regions exponential in depth of network
  - Capture repeated, mirroring, symmetric patterns in the data
  - Often better generalization

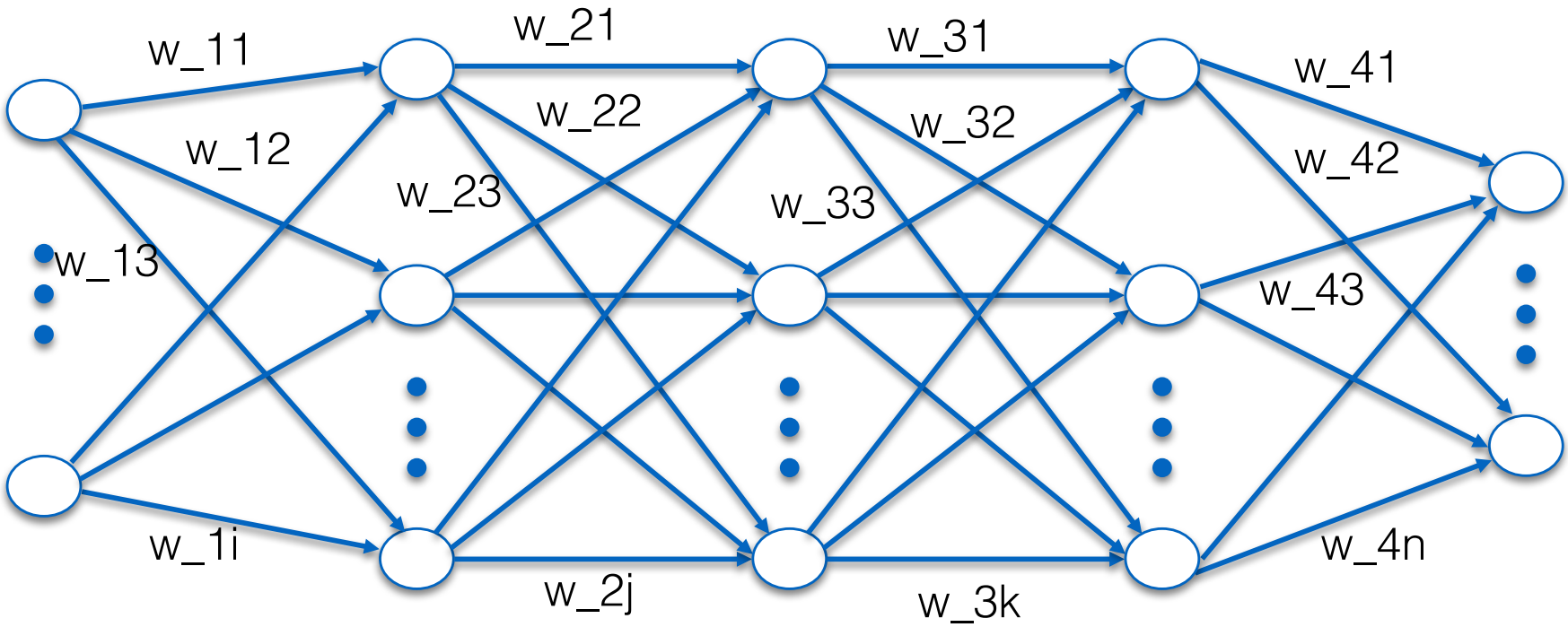
Convolutional Neural Networks



Example: Face Recognition

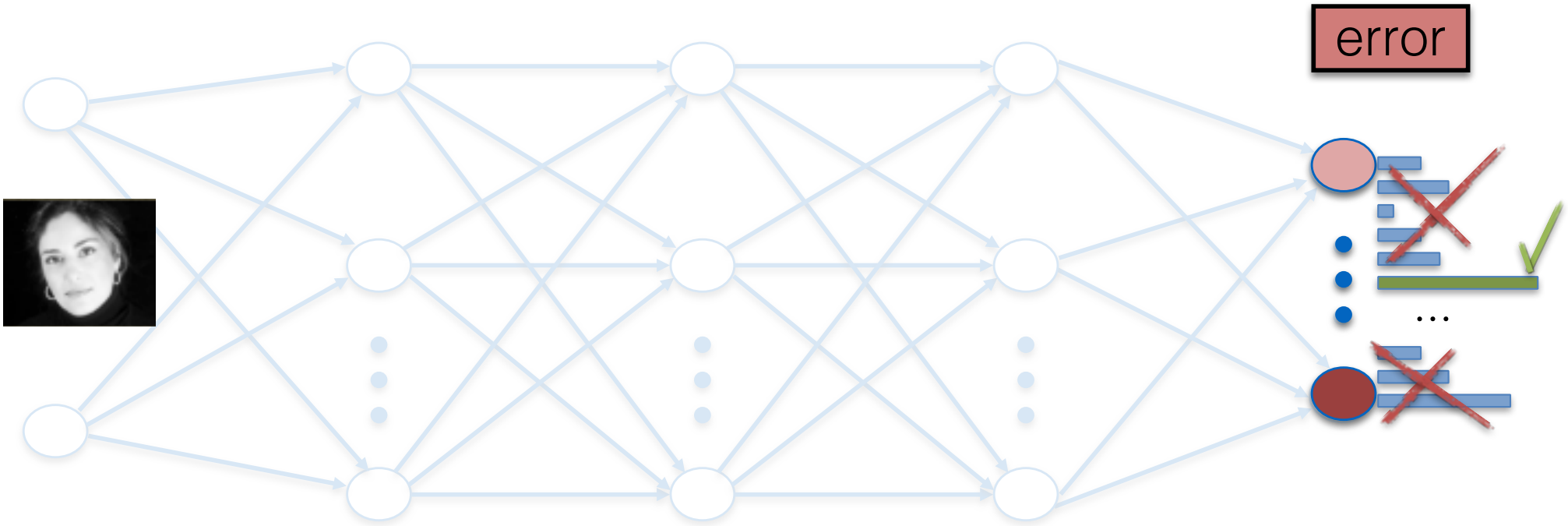


Parameter Training



Find parameters that minimize the classification error

Loss Function
(Cost Function, Objective Function)



Loss Function

- Measures the "cost" of fitting a model to data
- Examples:
  - L^2 squared difference between model output and known correct output ("ground truth")
  - L^1 absolute difference between model output and known correct output
  - Cross-entropy between model output (interpreted as probability) and correct output

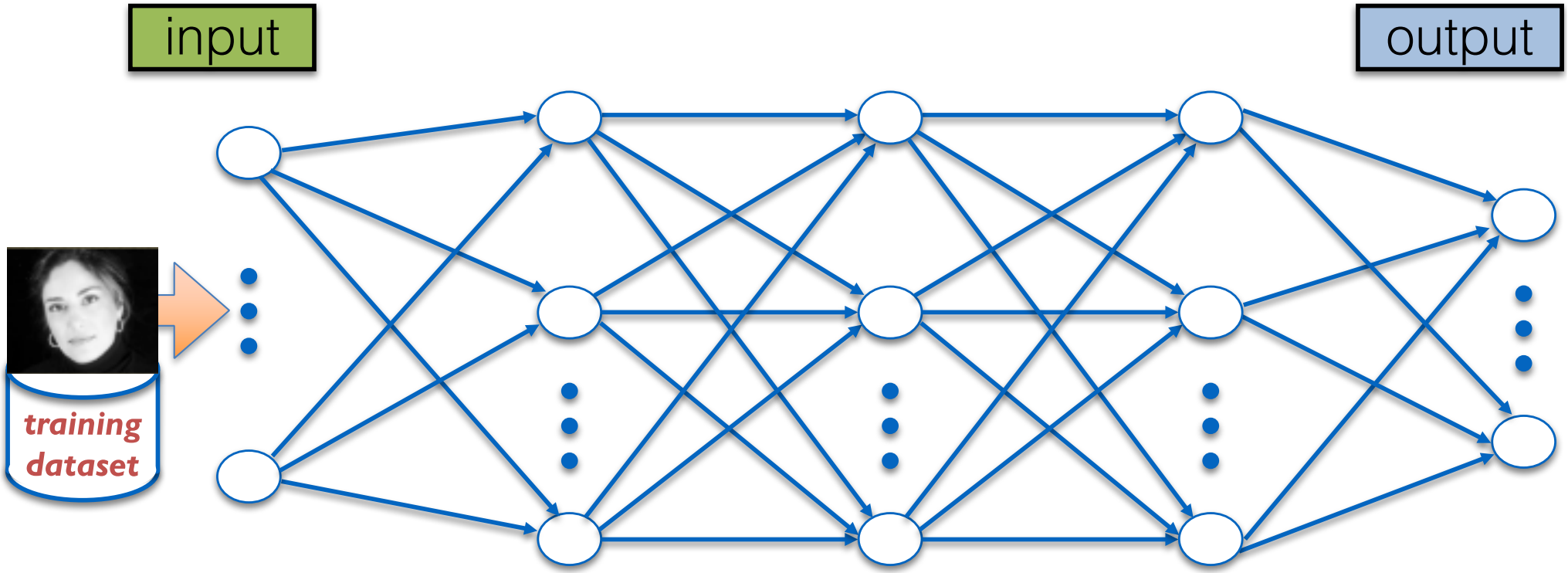
Measuring Accuracy

- Training dataset: model is "trained" to fit this
- Validation dataset: model is repeatedly tested on this data during training
- Testing dataset: measure accuracy of a trained model

Testing accuracy vs. training accuracy

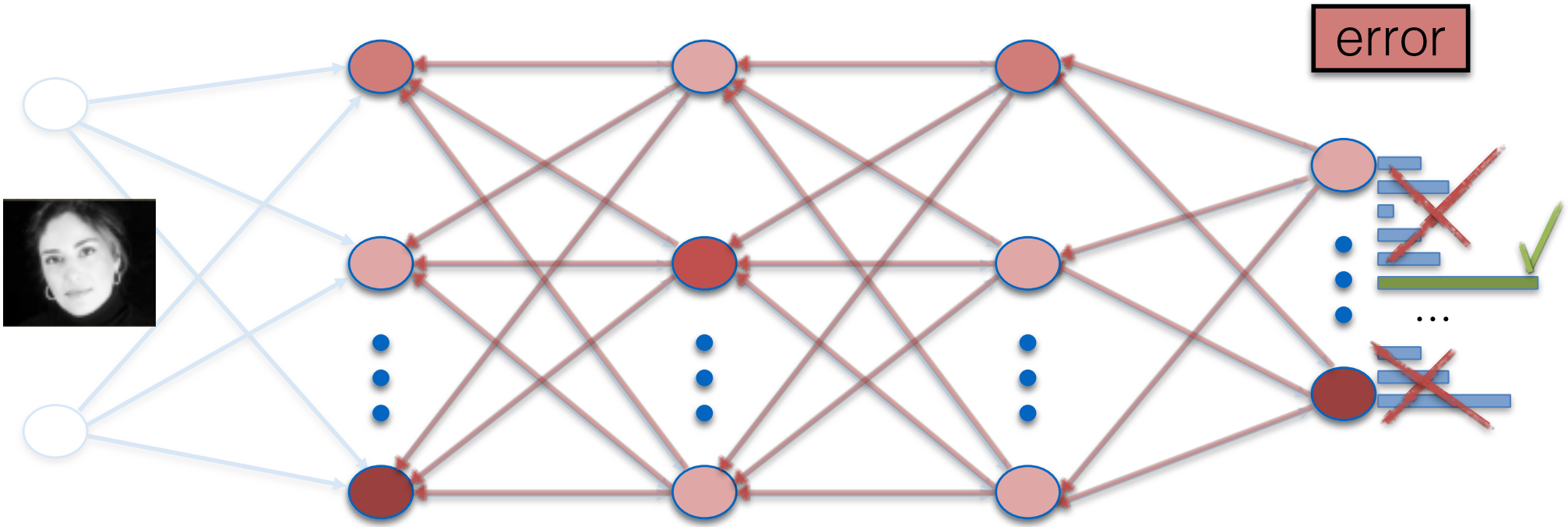
Parameter Training

1) Feed-forward

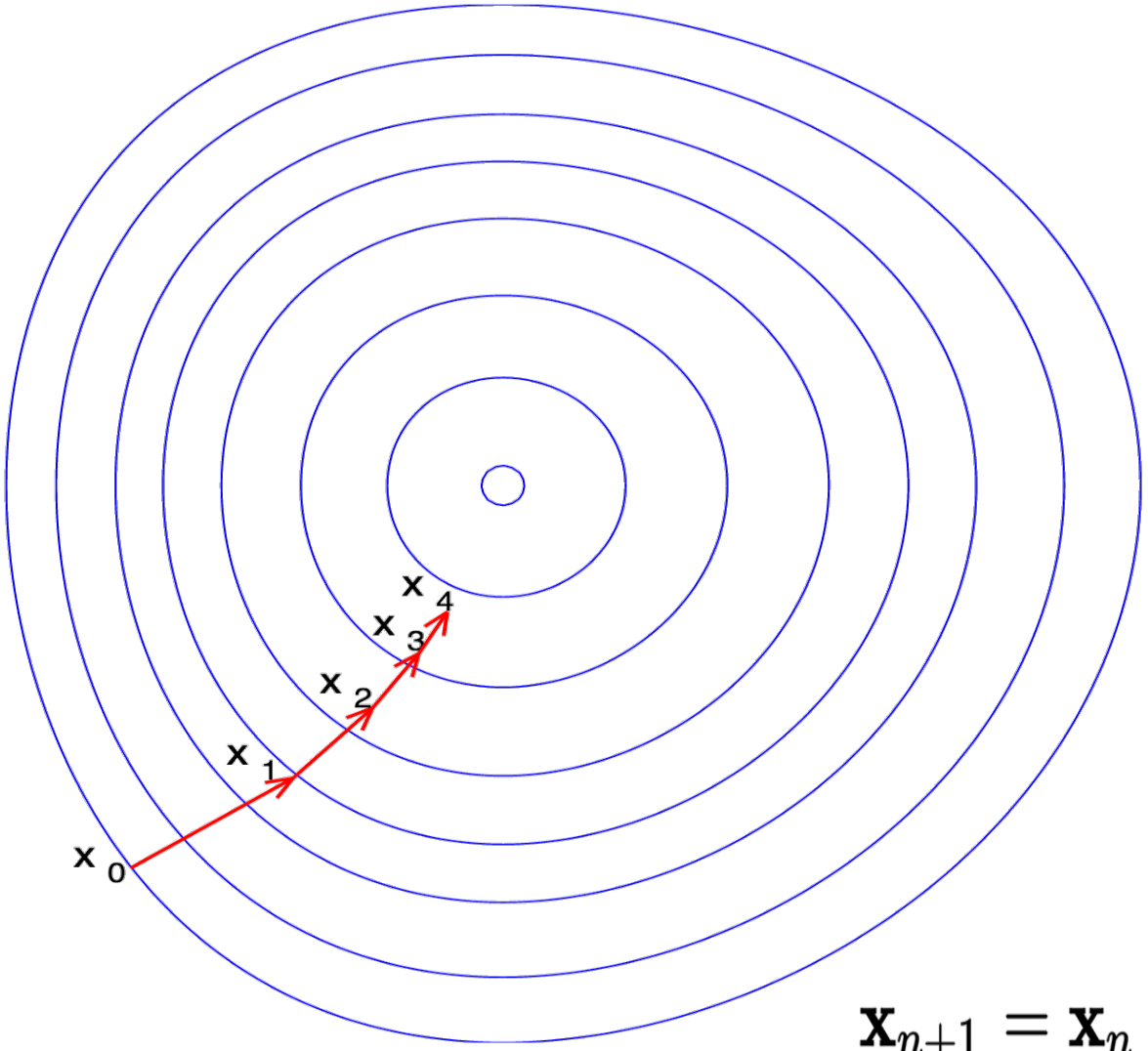


Parameter Training

2) Back-propagation



“Batch” Gradient Descent



Step size
(“learning rate”)

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n)$$

Stochastic Gradient Descent

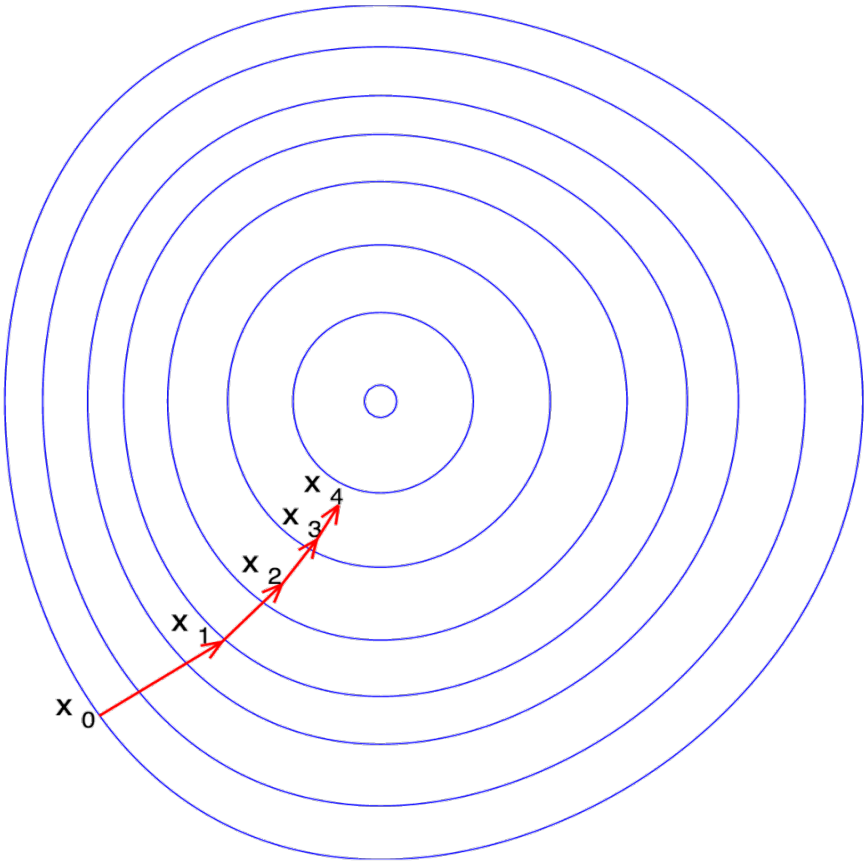
- Need to compute sum of n terms, n is large
- Sample instead of computing the full sum
  - Example: train on 100 photos to compute an estimate, then repeat and update the estimates

Repeat until an approximate minimum is reached:

- Pick a random sample of training examples
- For i in 1, 2 … n

$$w := w - \eta \nabla Q_i(w)$$

Gradient Descent with Backpropagation

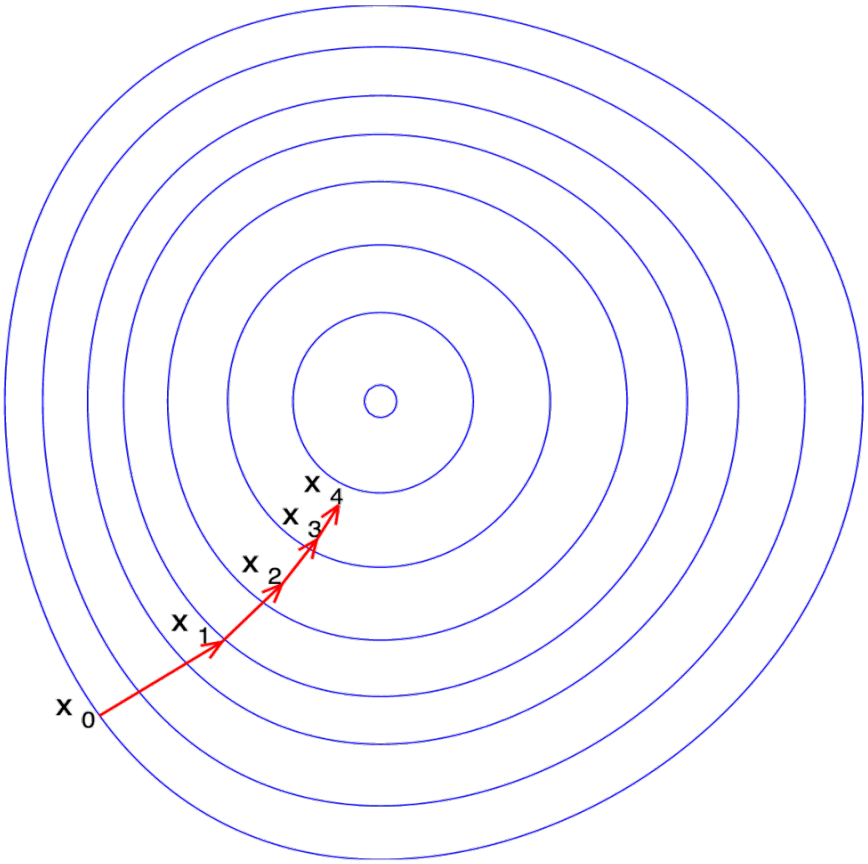


- Initialize weights w_0
- Repeatedly apply gradient descent:

$$w_{n+1} = w_n - \gamma_n \nabla_w E(w_n)$$

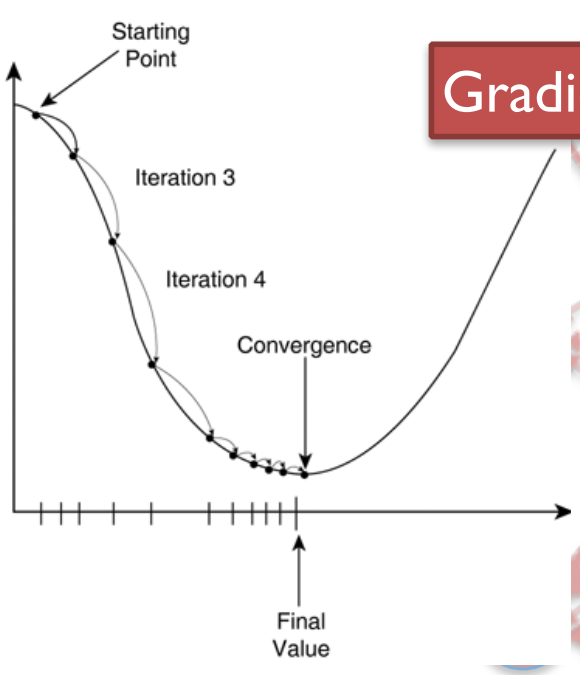
- Stop when validation error is "small"

Stochastic Gradient Descent with Backpropagation



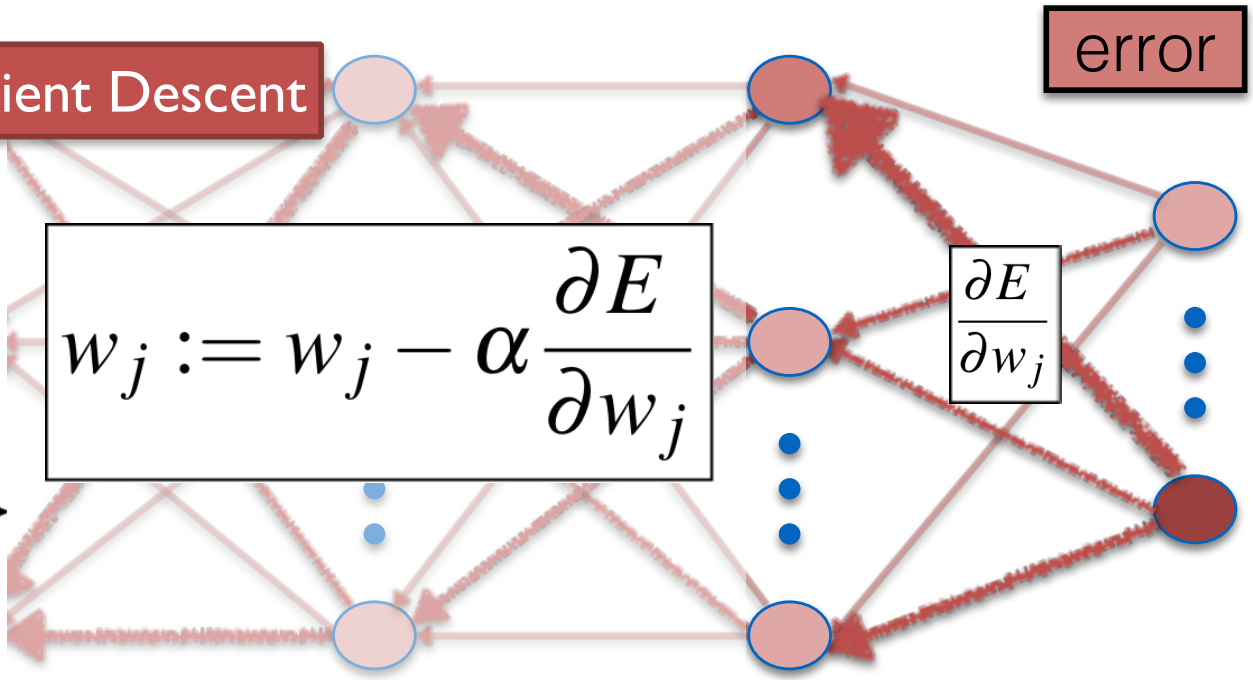
- Initialize weights w_0
- Randomly shuffle dataset
- For each batch i calculate gradient descent using backpropagation and apply $w_{n+1} = w_n - \gamma_n \nabla E_i(w_n)$
- Stop when validation error is "small"

Parameter Training using SGD

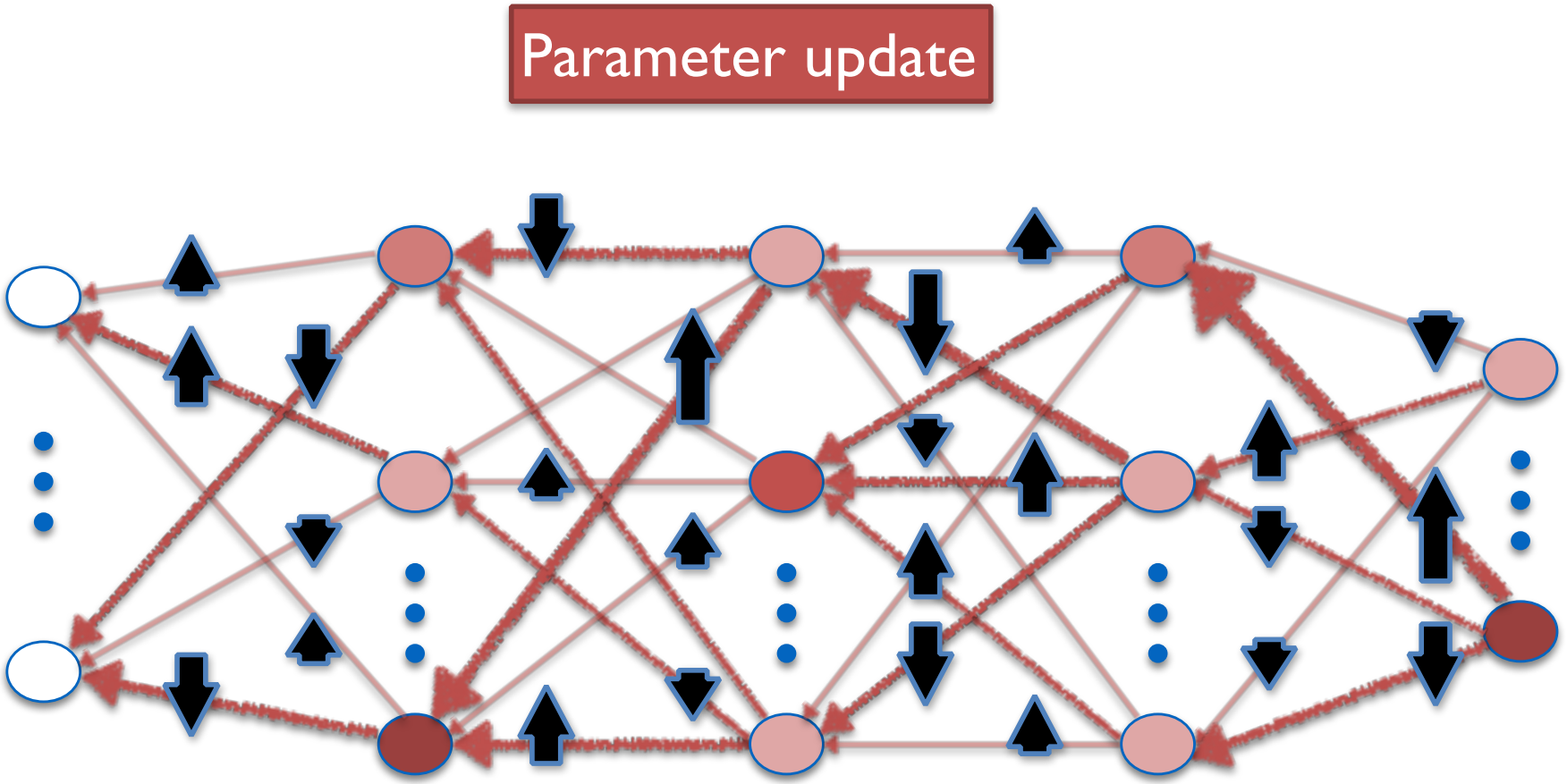


Gradient Descent

$$w_j := w_j - \alpha \frac{\partial E}{\partial w_j}$$



Parameter Training using SGD

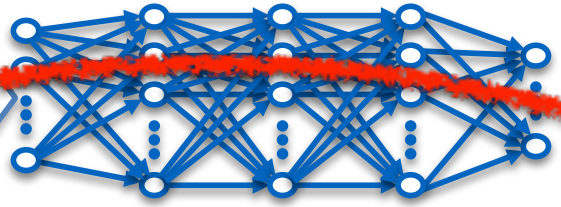


Repeat for new batches of training data

Privacy?

Sensitive data

Machine learning

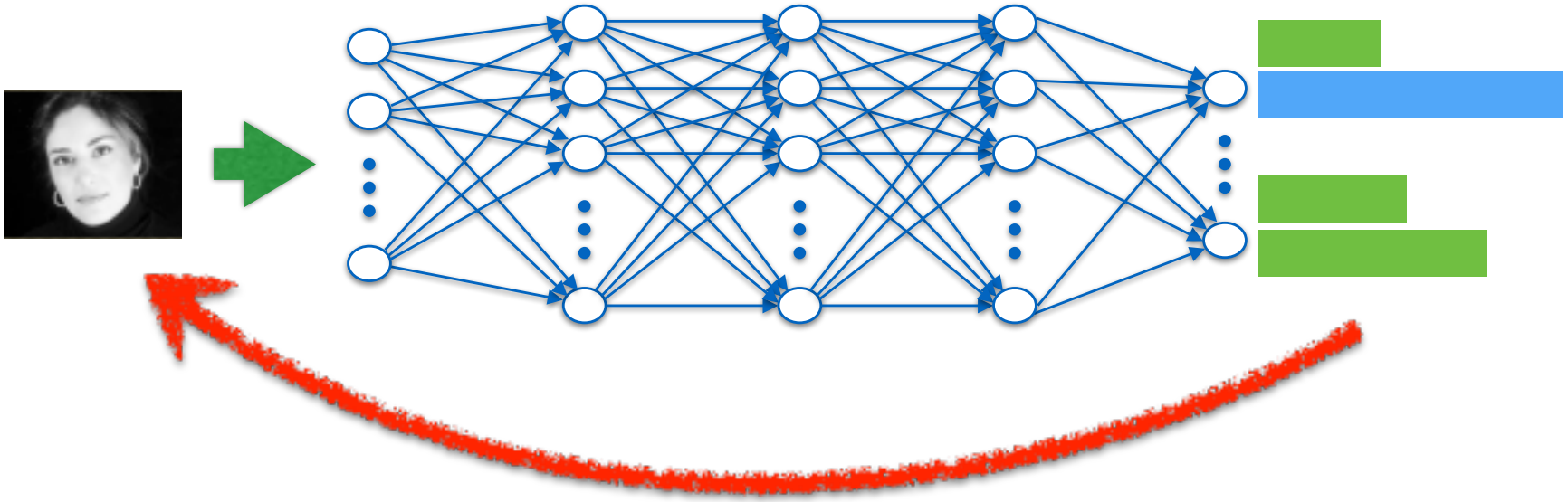


?

Medical images
Clinical records
Text documents
Personal photos
Retail purchases

Fredrikson et al.

Model Inversion

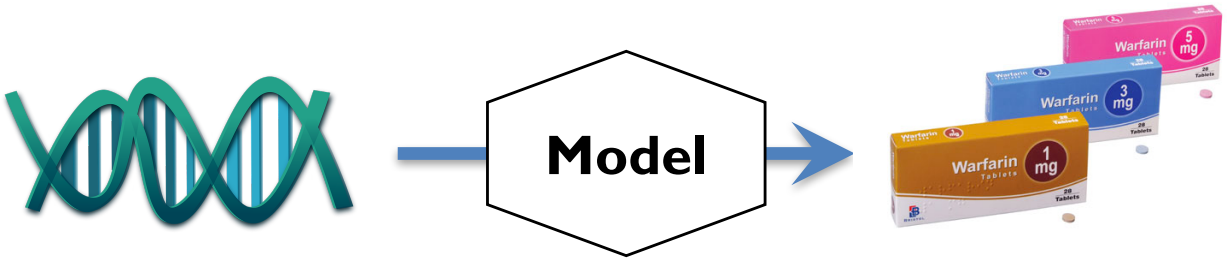


Given an output of a machine learning model, infer something about the input

"unexpected attributes"

Fredrikson et al.

Model Inversion in Action

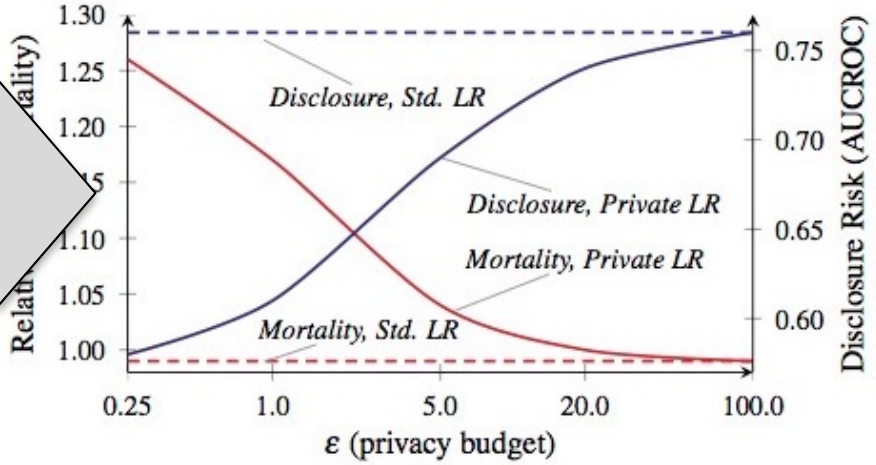


given patient's genome …

… determine correct warfarin dosage

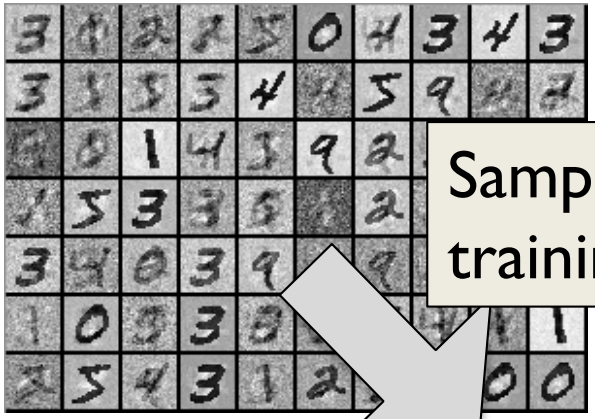
Privacy breach:
given patient's warfarin dosage,
infer information about patient's genome

What does this chart measure?

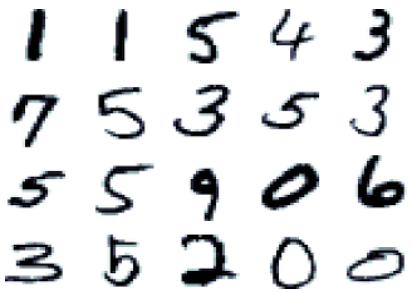


Hitaj et al.

Deep Models under the GAN

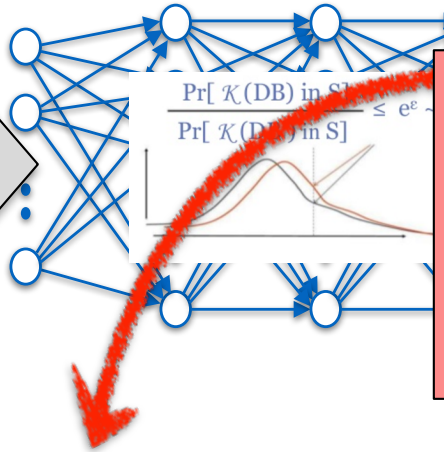


Sample a training dataset



Train a (differentially private) model

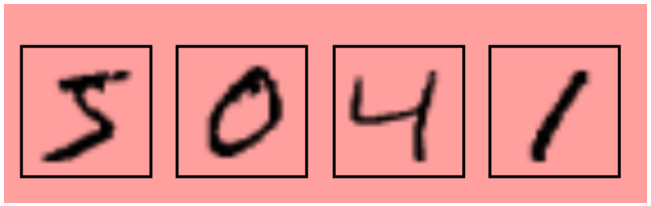
Abadi et al. (CCS 2016)



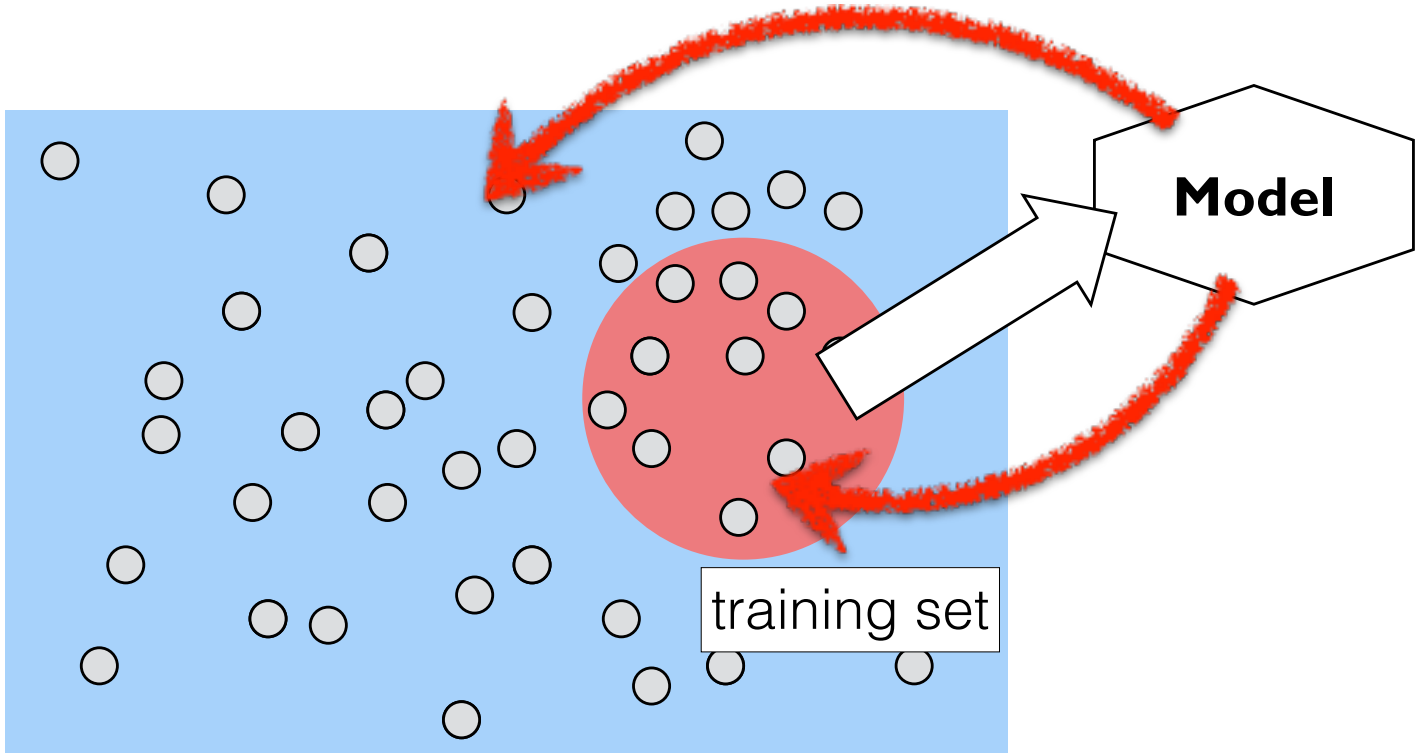
Use adversarial learning to generate representatives of every class

They look very similar to training data…

Privacy is broken?



Does Inference Breach "Privacy"?



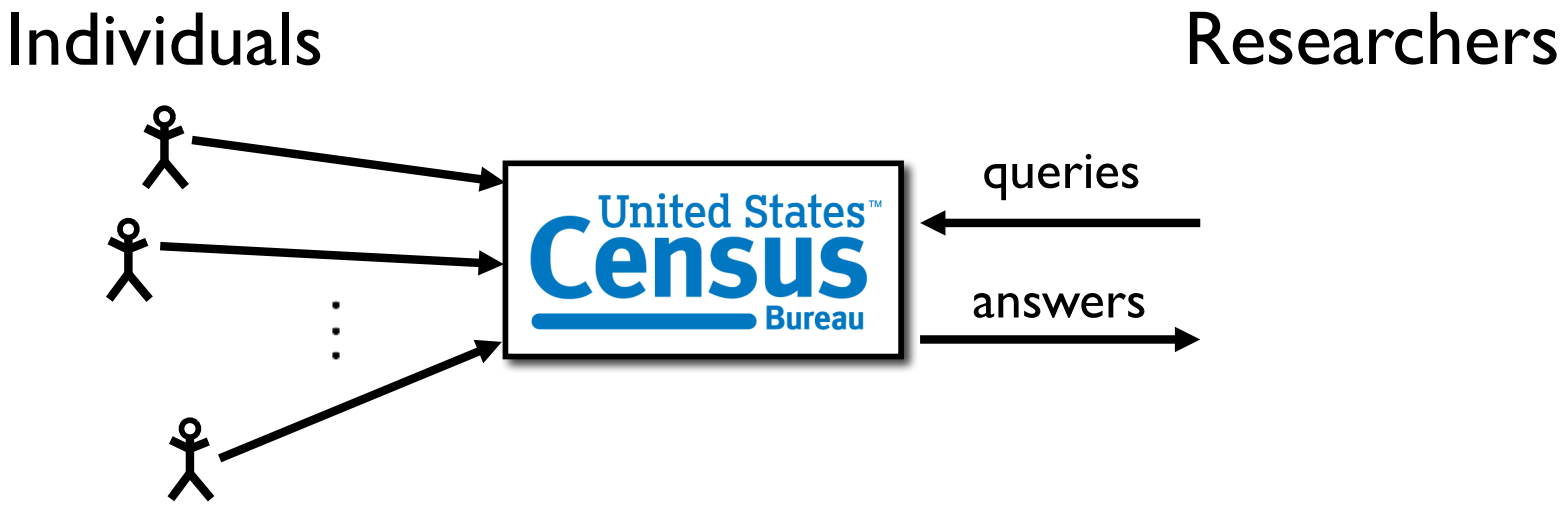
Recommended Reading

Frank McSherry.

"Statistical inference considered harmful"

https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md

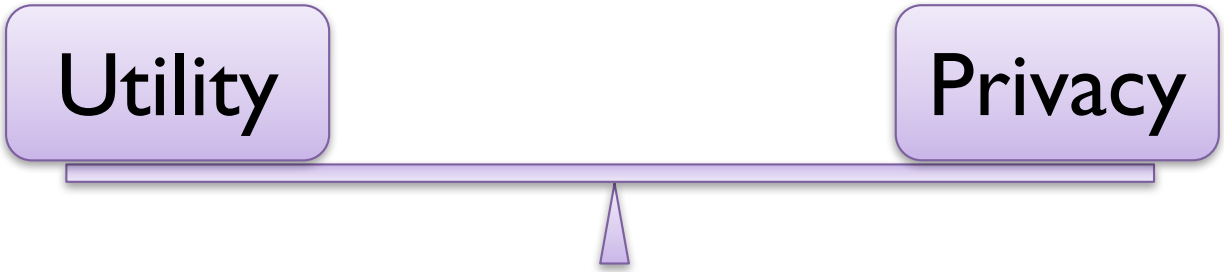
Privacy in Statistical Databases



Large collections of personal information

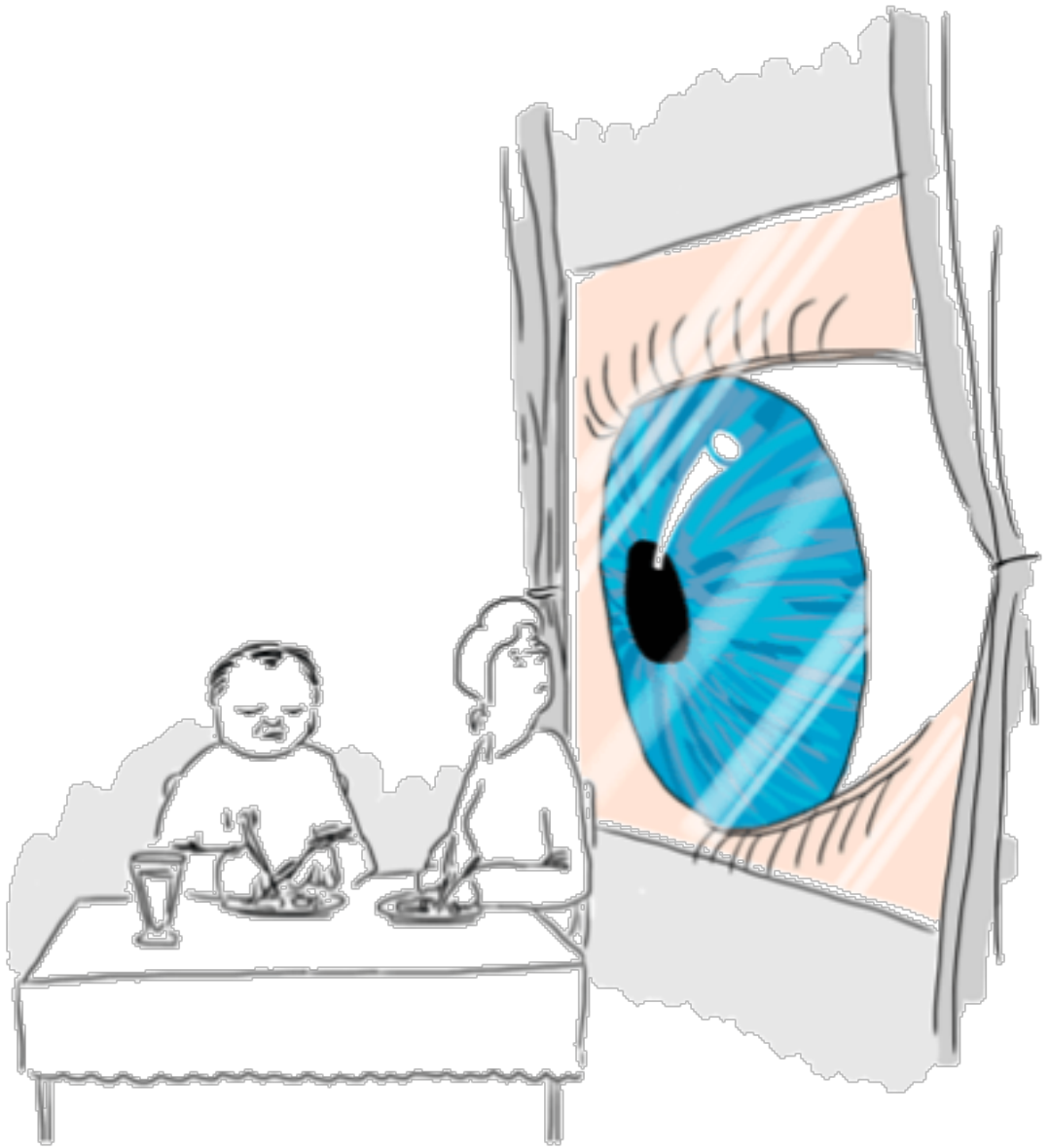
- census data
- national security
- medical/public health
- social networks
- recommendation systems
- education

Conflicting Goals



Utility: release aggregate statistics

Privacy: ??? (intuition: individual information stays "hidden")

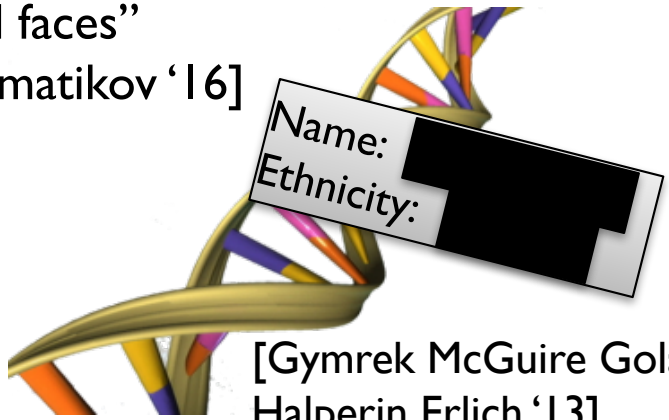


"Relax – it can only
see metadata."

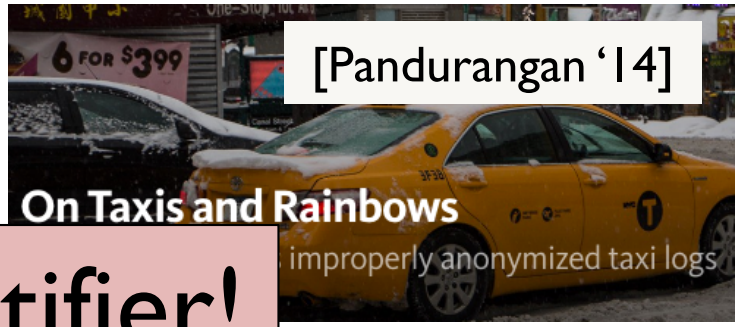
Slide credit: Adam Smith

Remove Obvious Identifiers?

"AI recognizes blurred faces"
[McPherson Shokri Shmatikov '16]

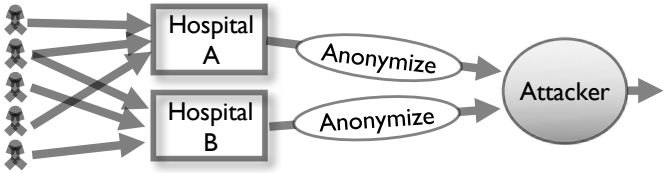


[Gymrek McGuire Golan Halperin Erlich '13]



[Pandurangan '14]

Everything is an identifier!



[Ganta Kasiviswanathan Smith '08]

Membership Inference Attacks

- Exact high-dimensional summaries allow an attacker to test membership in a data set

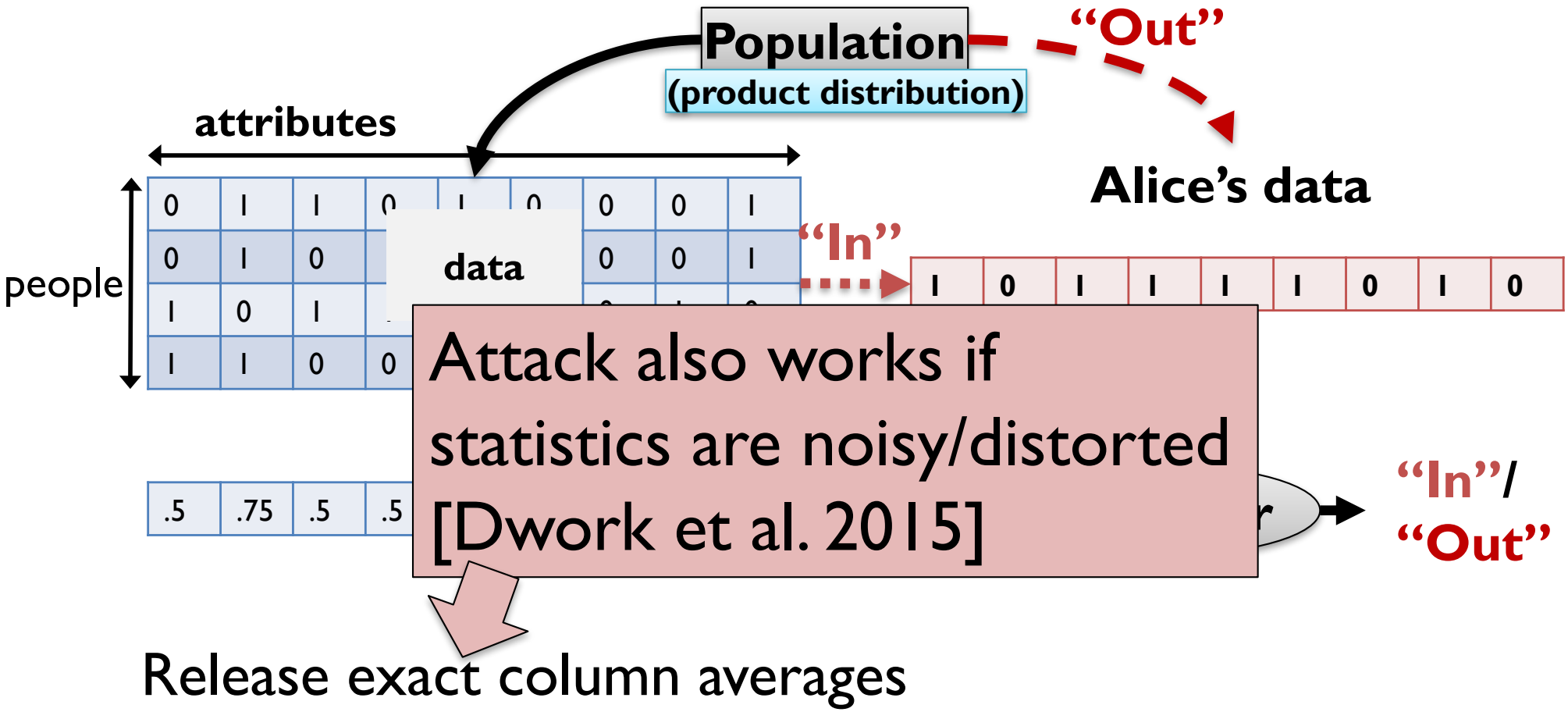
  [Homer et al. 2008]

  – Caused US NIH to change data sharing practices for genomic data

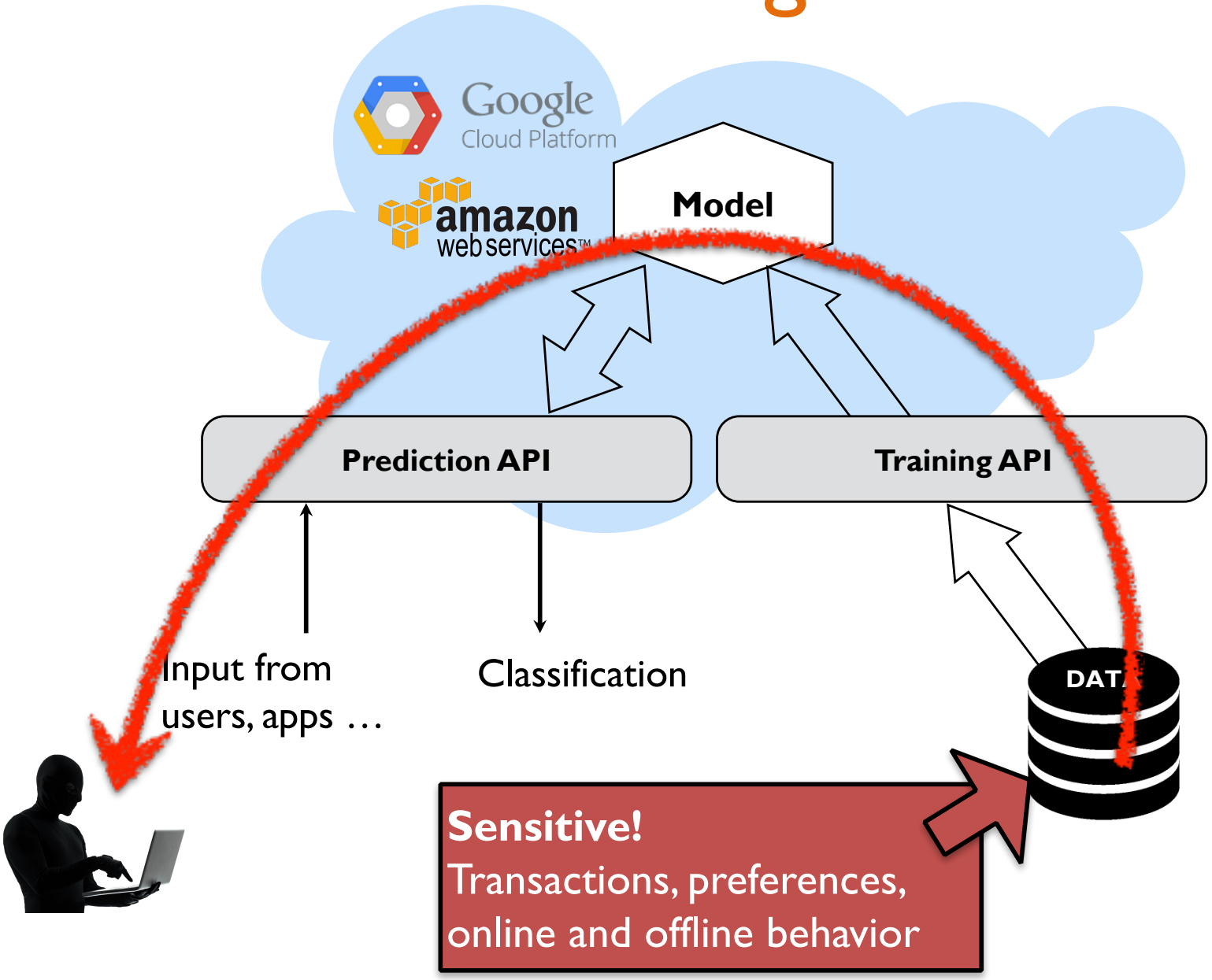
- Distorted high-dimensional summaries allow an attacker to test membership in a data set

  [Dwork et al. FOCS 2015]

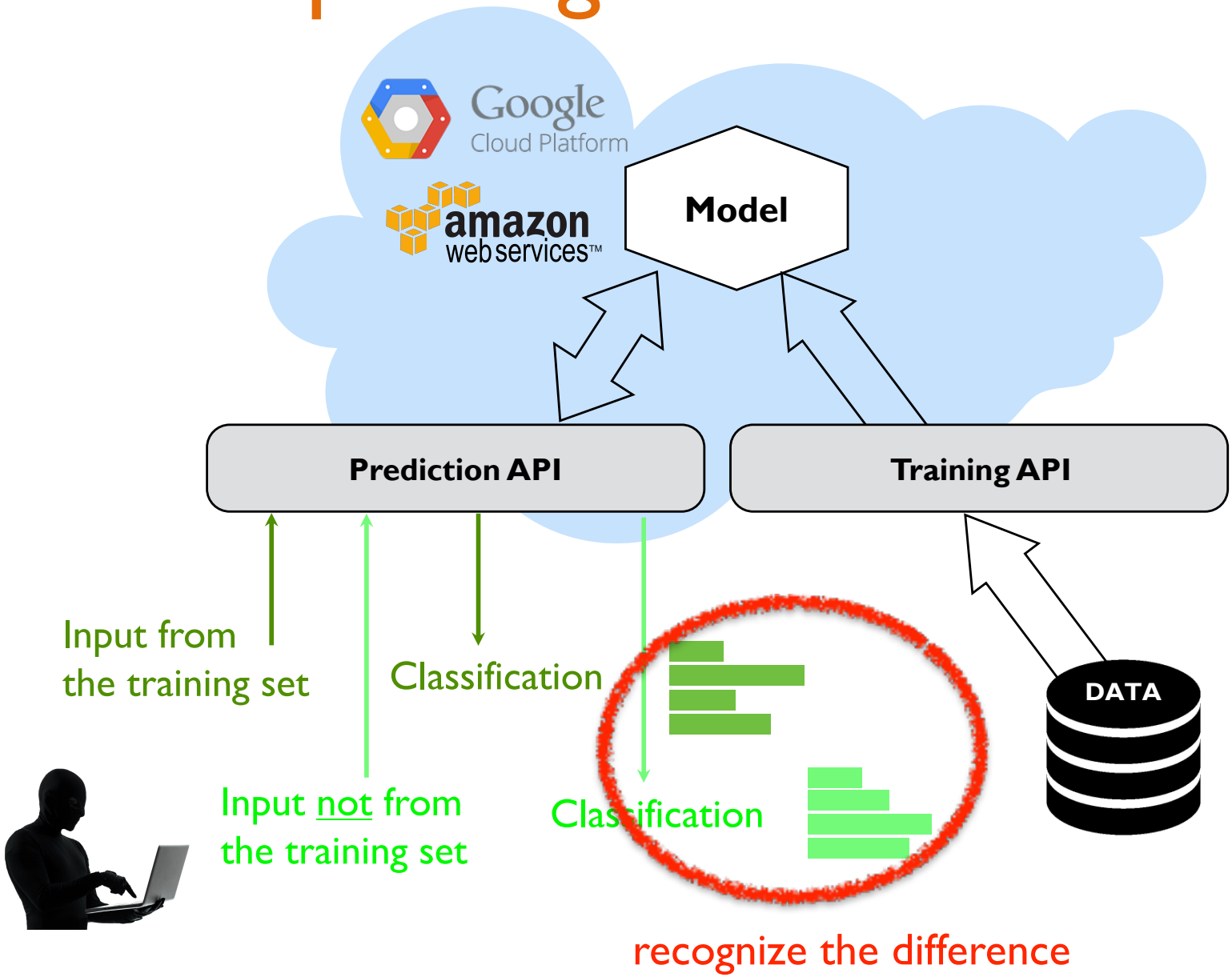
Homer at al. Attack



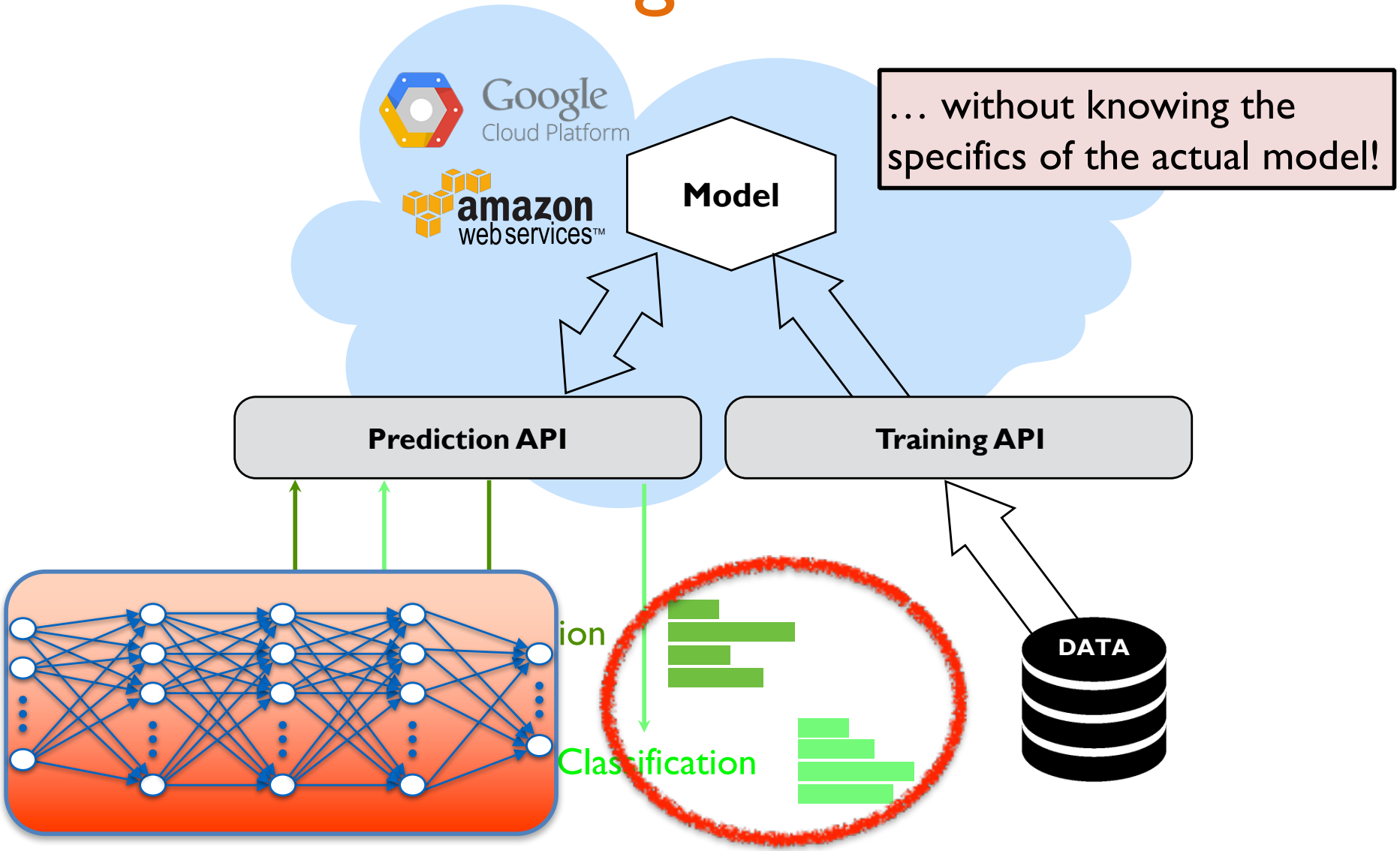
Machine Learning as a Service



Exploiting Trained Models



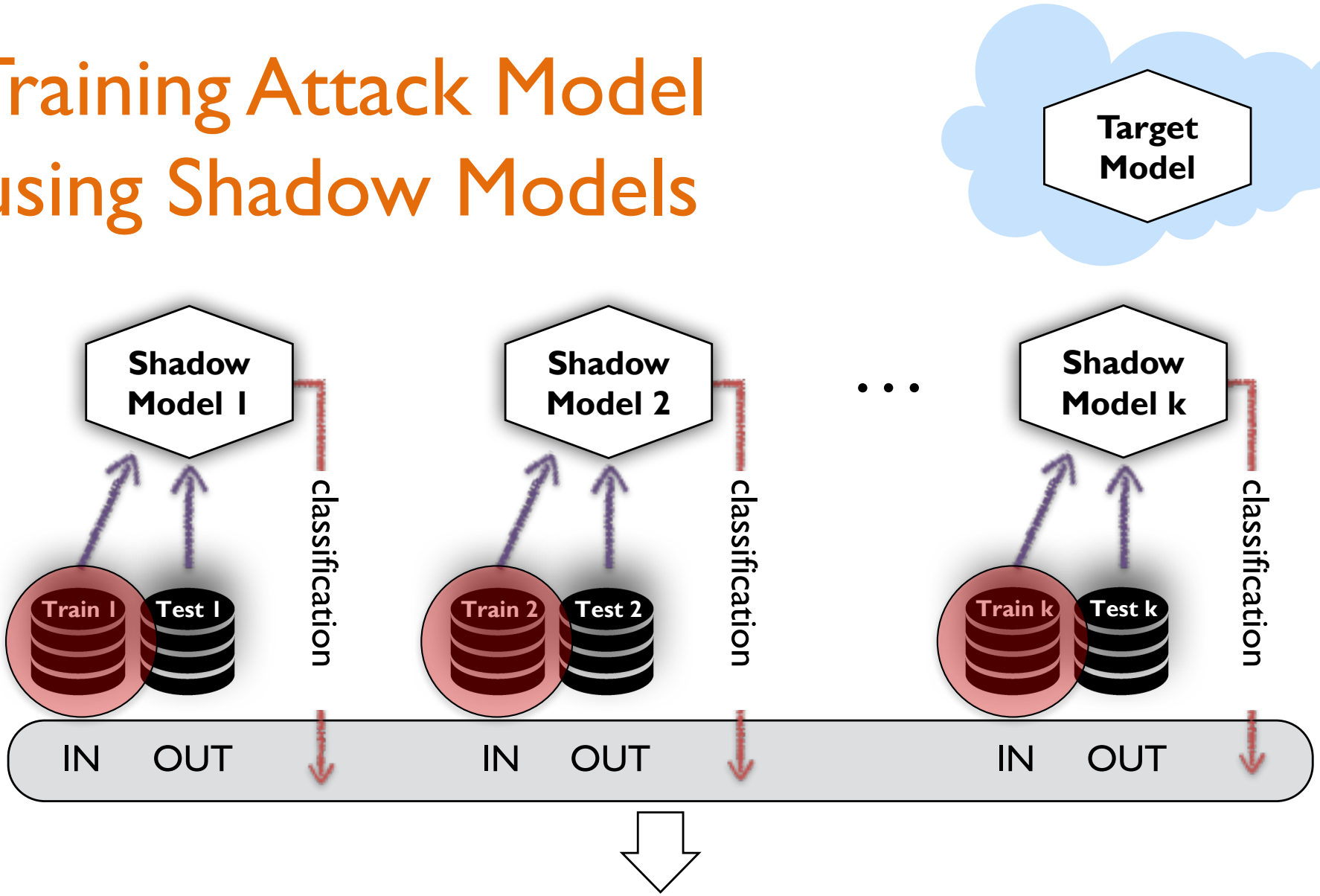
ML Against ML



Train a model to…

recognize the difference

Training Attack Model using Shadow Models

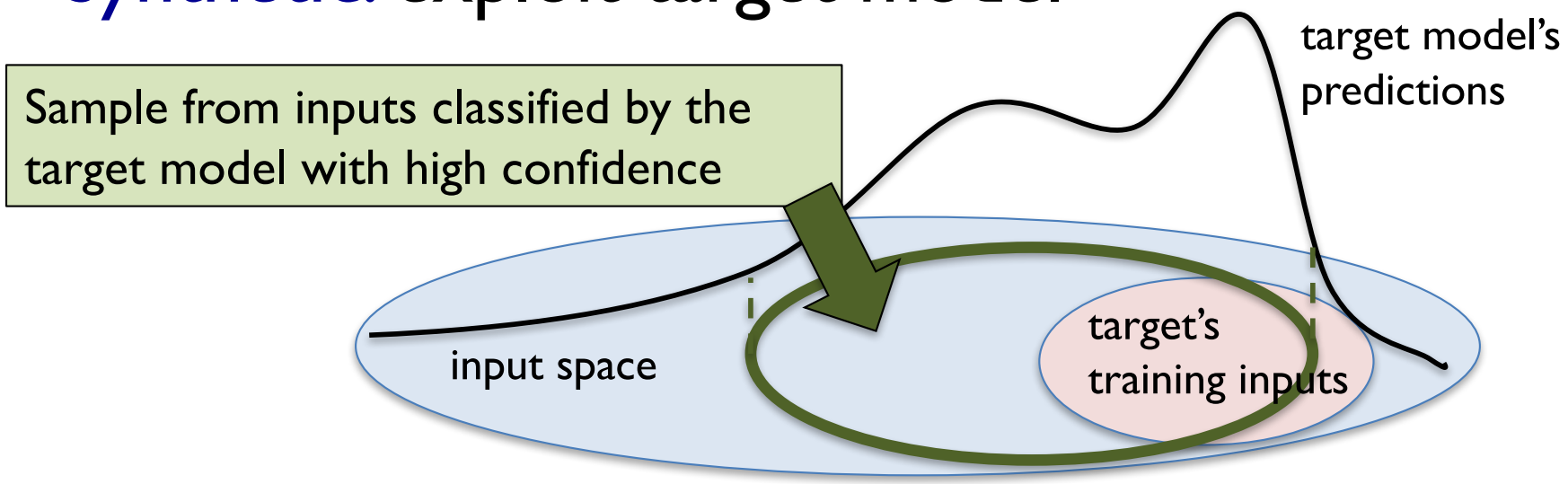


Train the attack model

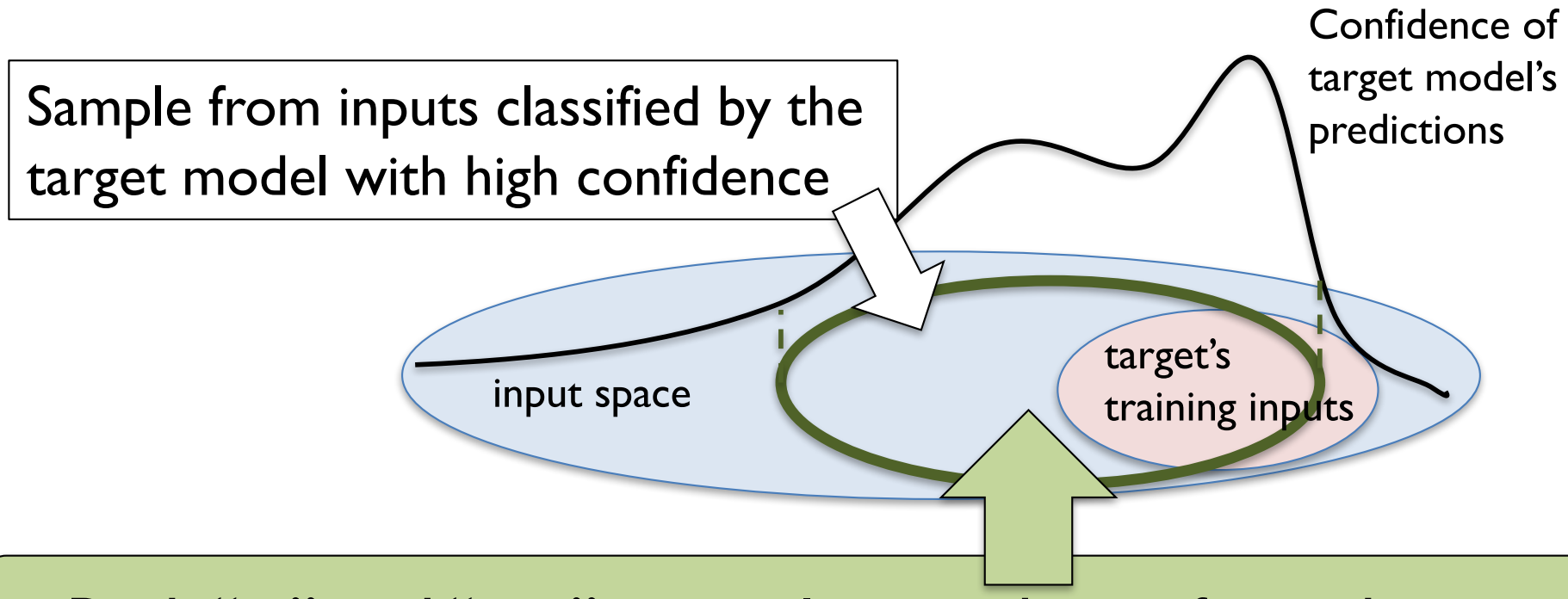
to predict if an input was a member of the training set (in) or a non-member (out)

Training Data for Shadow Models

- Real: must be similar to training data of the target model (drawn from same distribution)
- Synthetic: sample feature values from (known) marginal distributions
- Synthetic: exploit target model



Important Point



- Both "in" and "out" examples are drawn from this space

- It is <u>not</u> the case that attack model simply learns to say that all inputs classified by the target model with high confidence belong to its training dataset

Synthesizing Shadow Training Data

Hill-climb the space of possible inputs to find those classified by the target model with high confidence

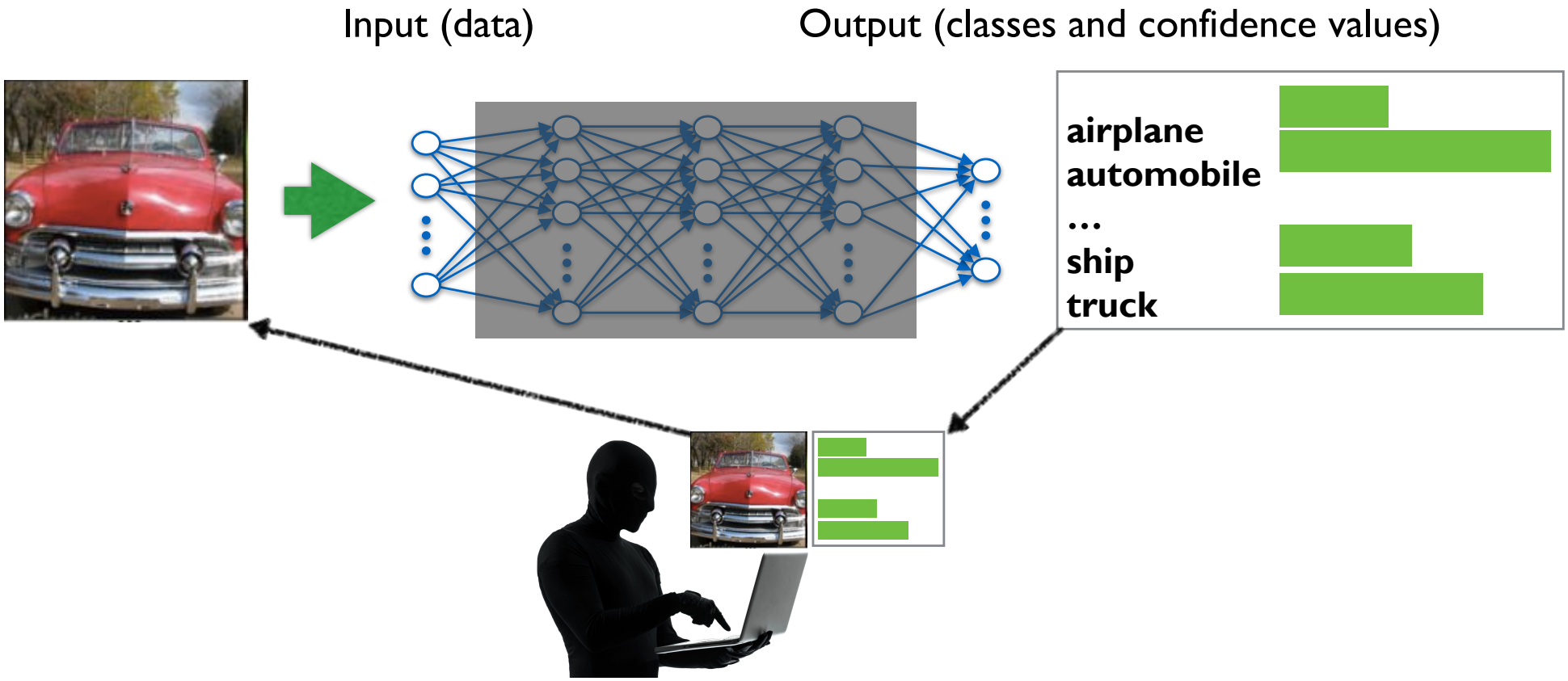
Sample from these inputs to synthesize the training dataset for shadow models

If many candidate inputs rejected by the target model, re-randomize some features and try again

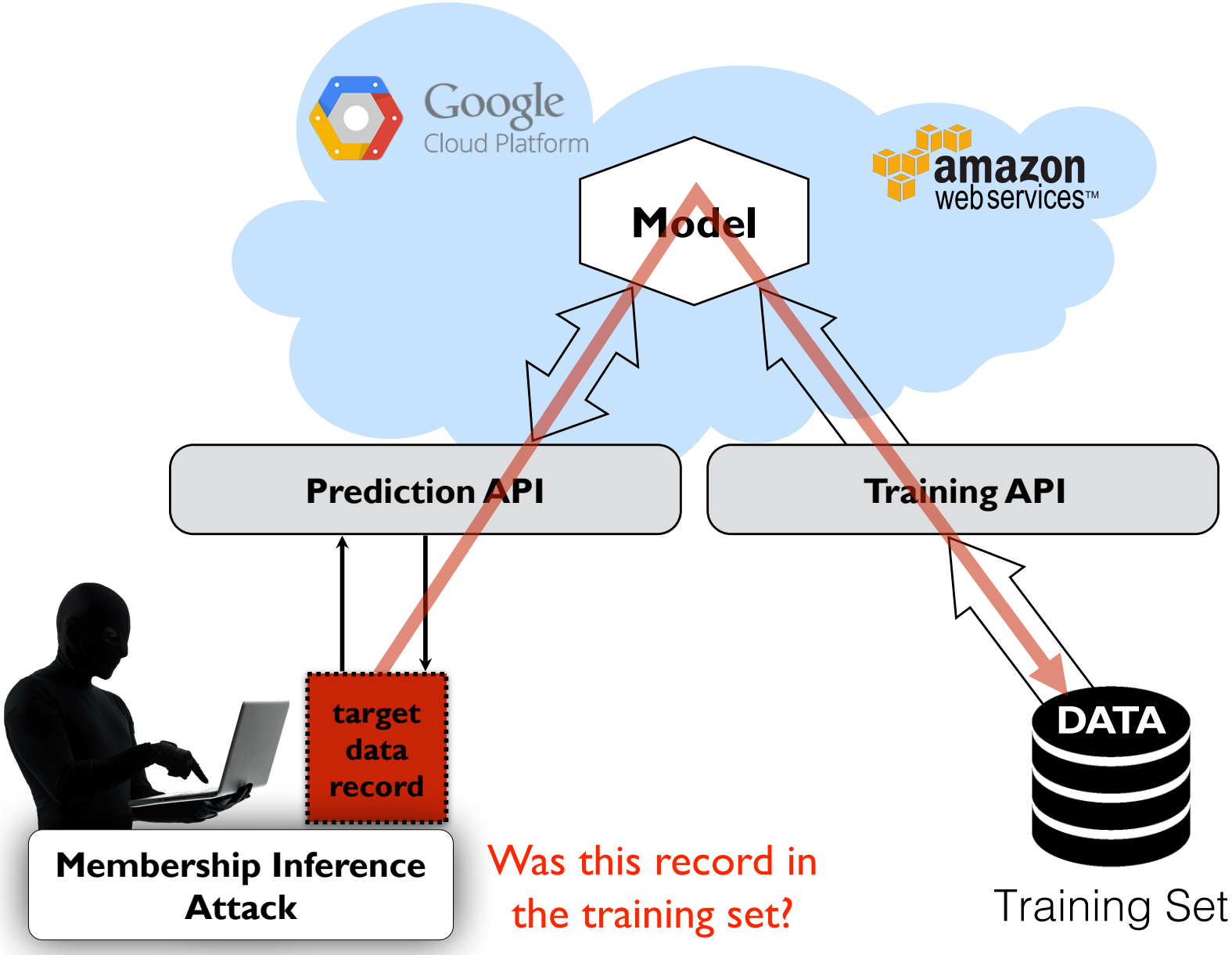
Algorithm 1 Data synthesis using the target model

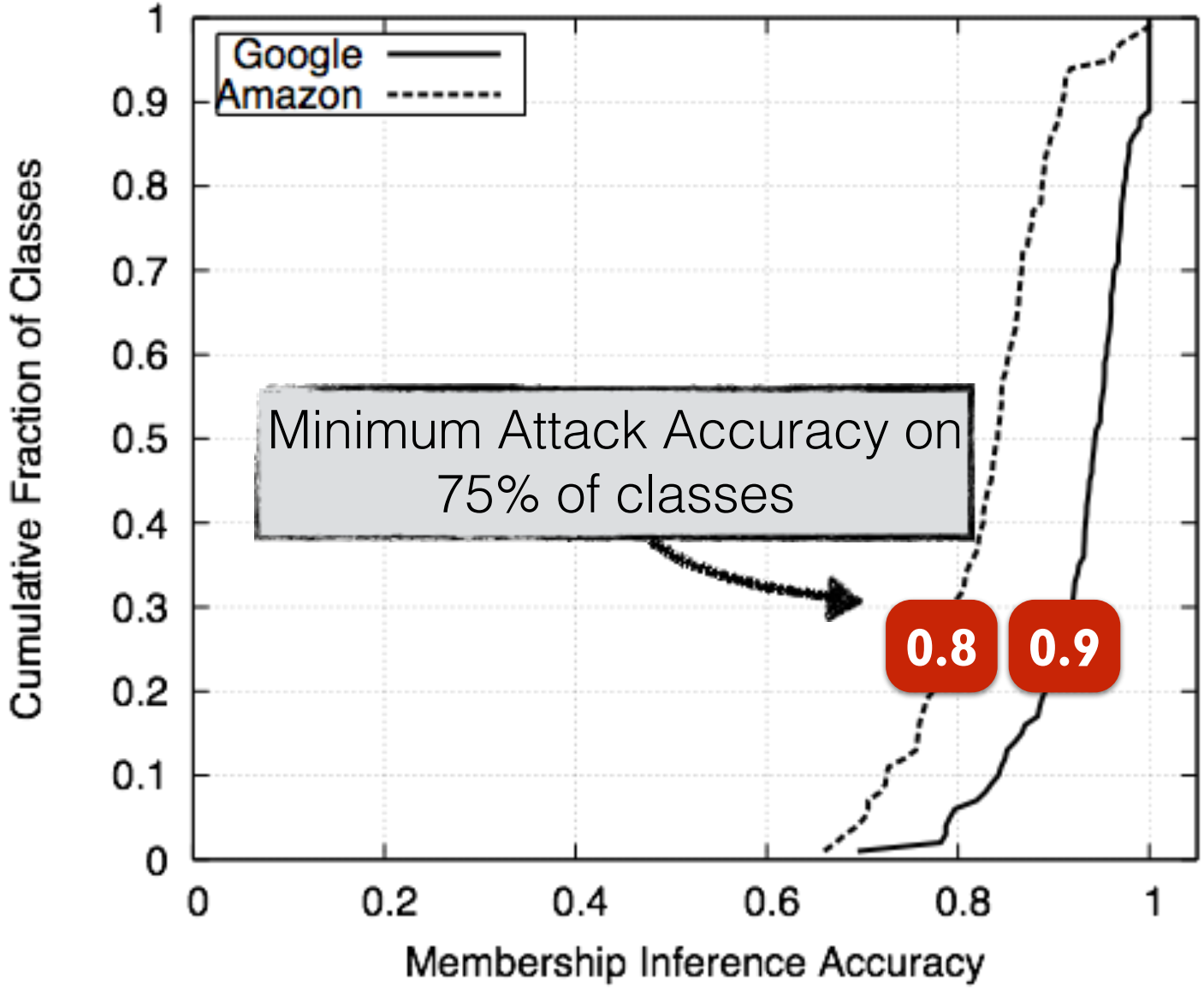
```
1:  procedure SYNTHESIZE(class : c)
2:      x ← RANDRECORD( )                          ▷ initilize a record randomly
3:      y*_c ← 0
4:      j ← 0
5:      k ← k_max
6:      for iteration = 1 ··· iter_max do
7:          y ← f_target(x)                        ▷ query the target model
8:          if y_c ≥ y*_c then                     ▷ accept the record
9:              if y_c > conf_min and c = arg max(y) then
10:                 if rand() < y_c then           ▷ sample
:                       return x                   ▷ synthetic data
:                   end if
:               end if
:               x* ← x
:               y*_c ← y_c
:               j ← 0
17:         else
18:             j ← j + 1
19:             if j > rej_max then                ▷ many consecutive rejects
20:                 k ← max(k_min, ⌈k/2⌉)
21:                 j ← 0
22:             end if
23:         end if
24:         x ← RANDRECORD(x*, k)                  ▷ randomize k features
25:     end for
26:     return ⊥                                   ▷ failed to synthesize
27: end procedure
```

Membership Inference Attack



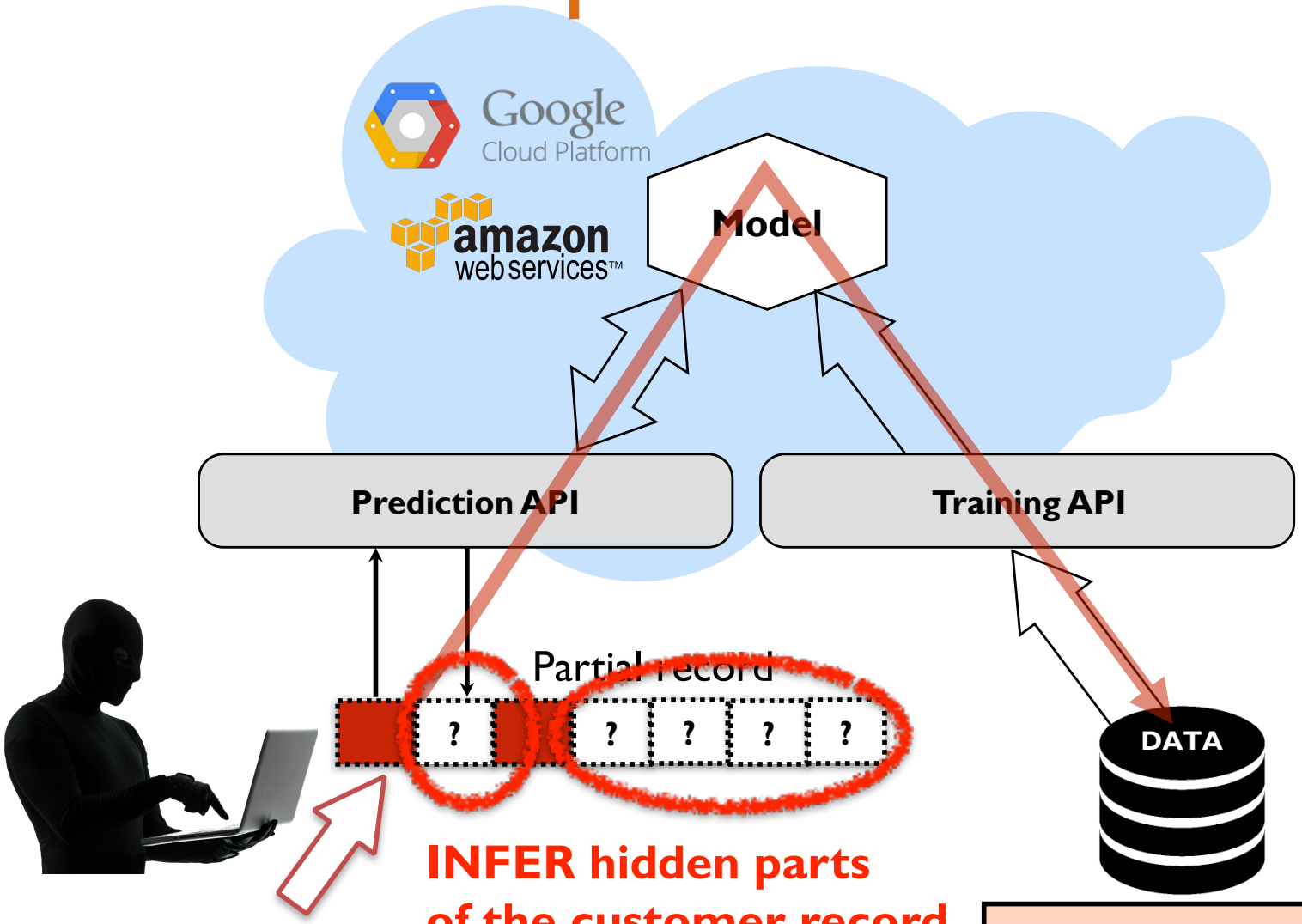
Was this image part of the training set?


Google
Cloud Platform
amazon
web services™
Model
Prediction API
Training API
target
data
record
DATA
Membership Inference
Attack
Was this record in
the training set?
Training Set



Purchase Dataset — Classify Customers

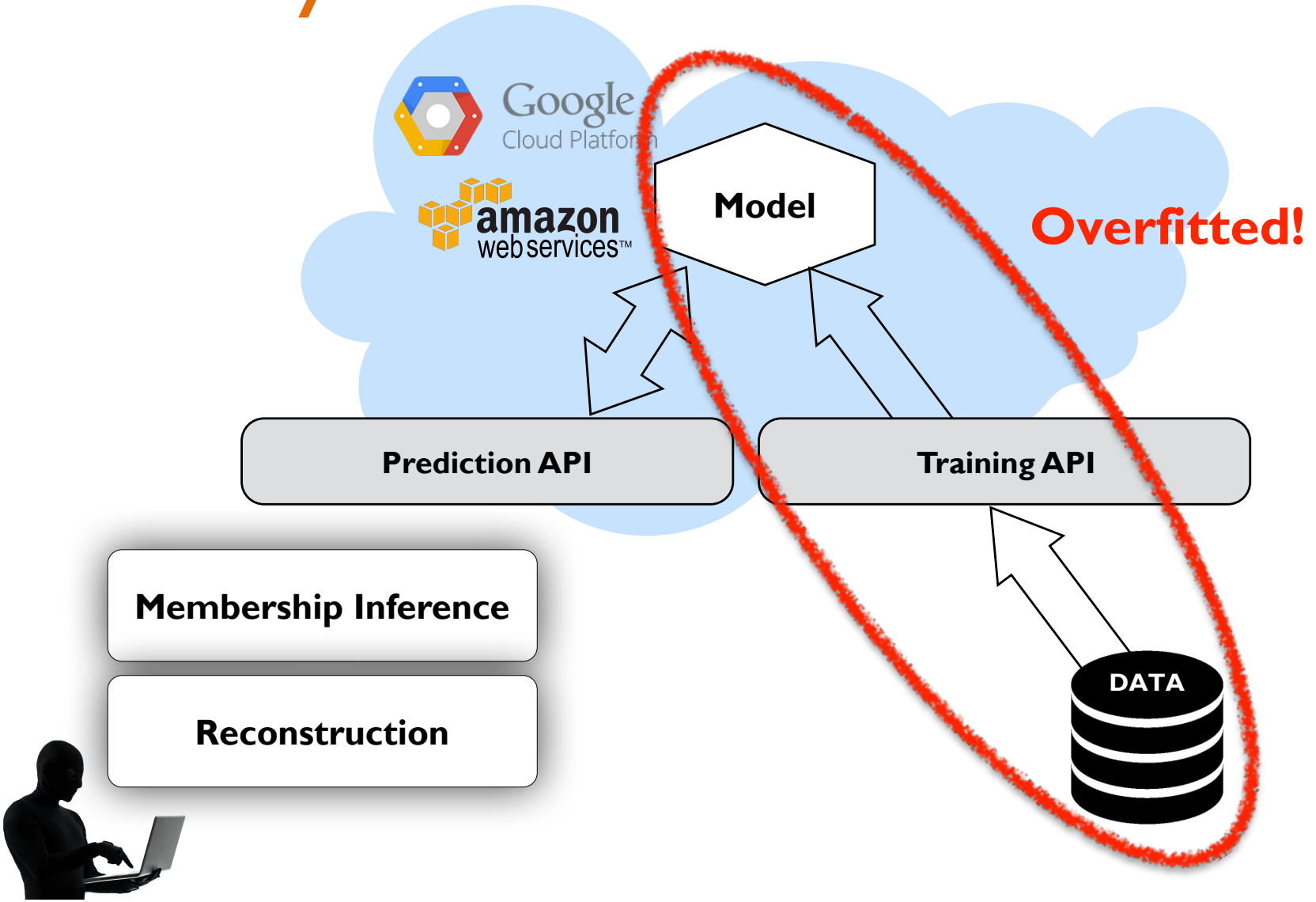
Next Step: Reconstruction



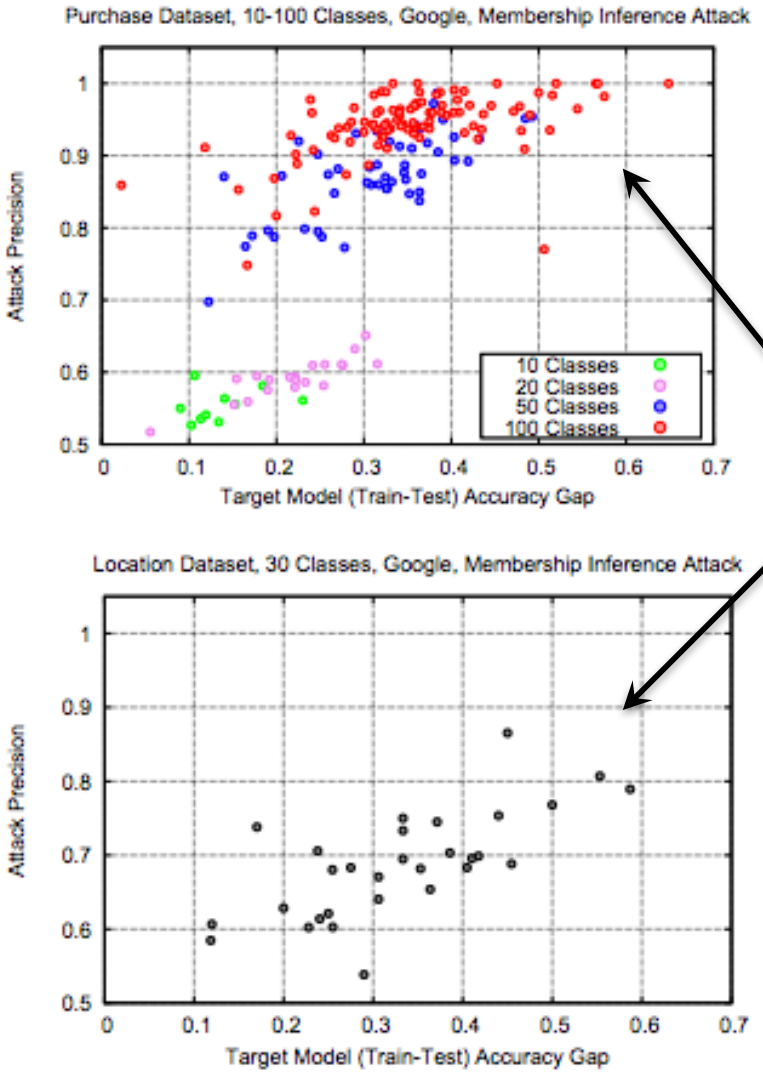
Auxiliary information, public databases, accidentally revealed data

Example: retail purchases or mobile phone locations

Why Do These Attacks Work?



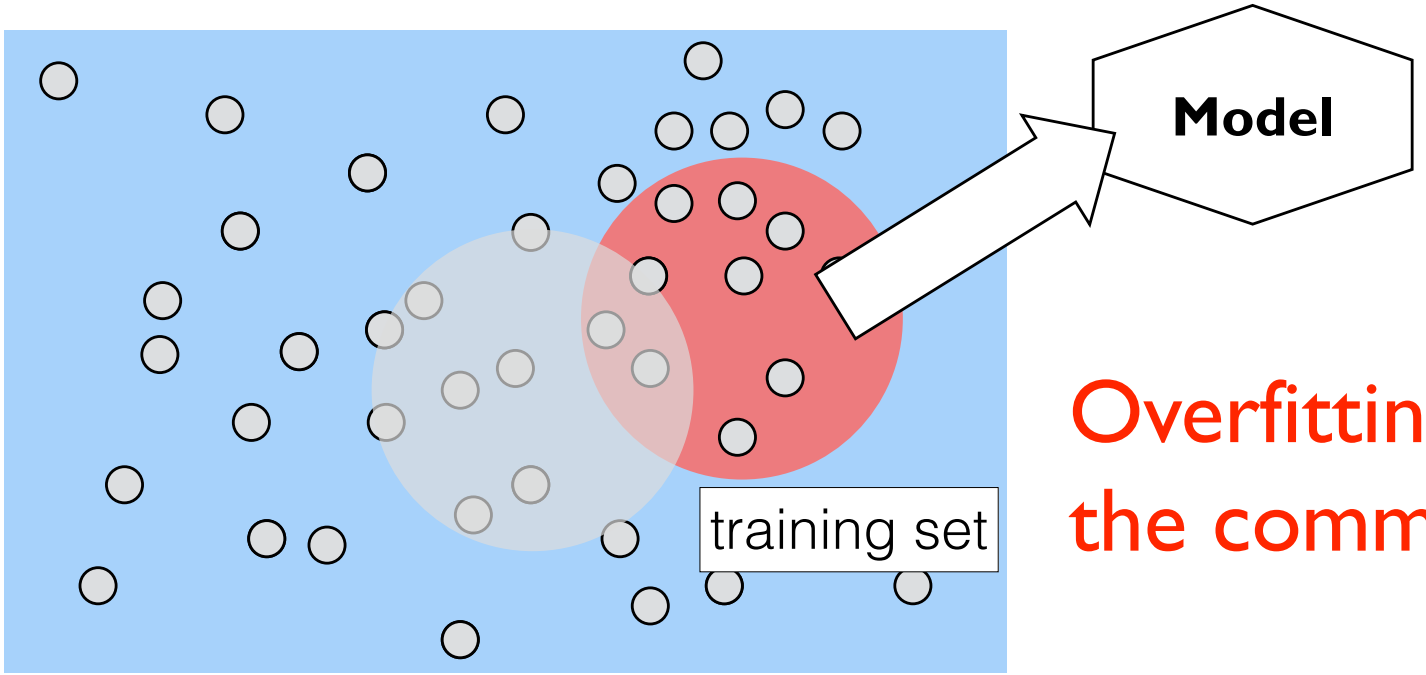
Attack Success vs. Test-Train Gap



More overfitted

Privacy:
Does the model leak information about data in the training set?

Learning:
Does the model generalize to data outside the training set?

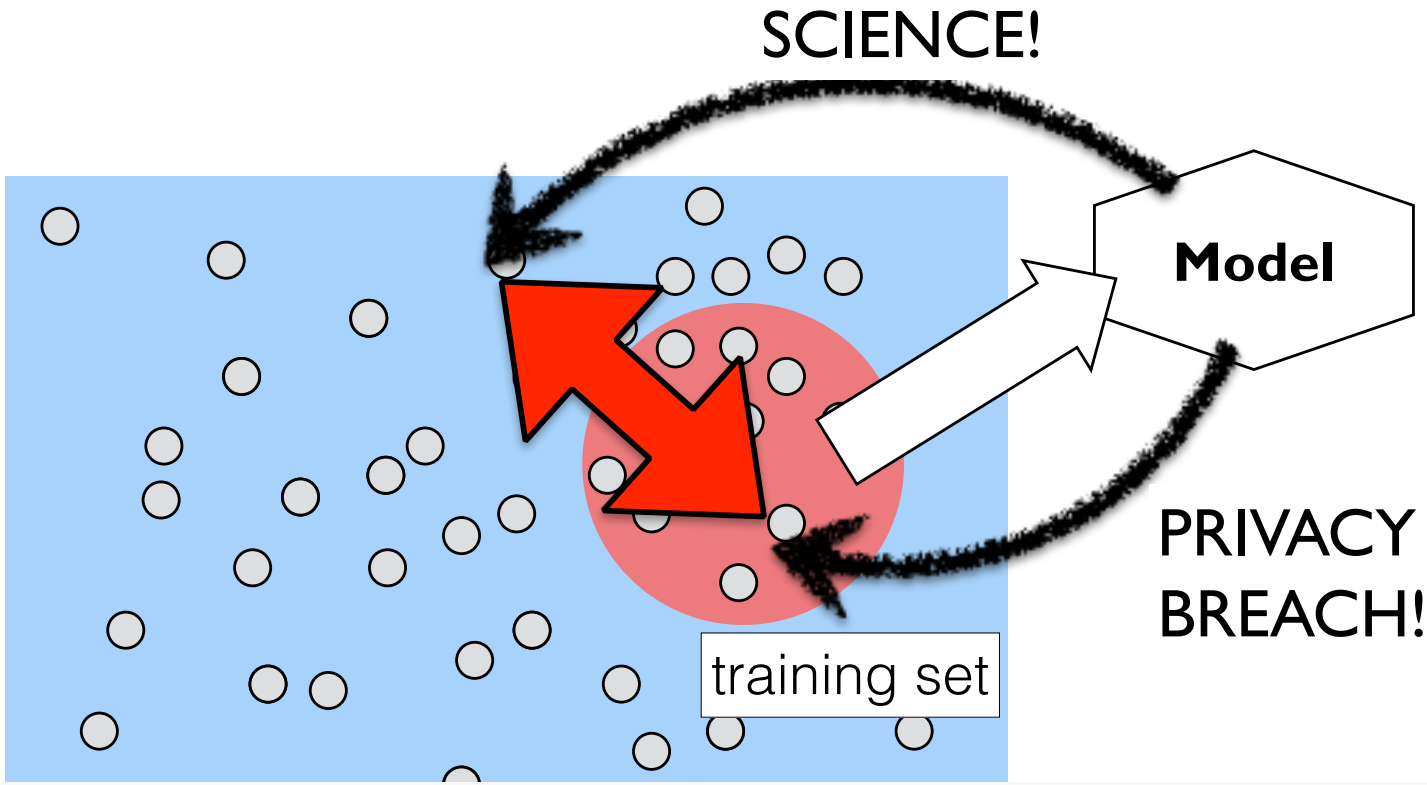


Overfitting is the common enemy

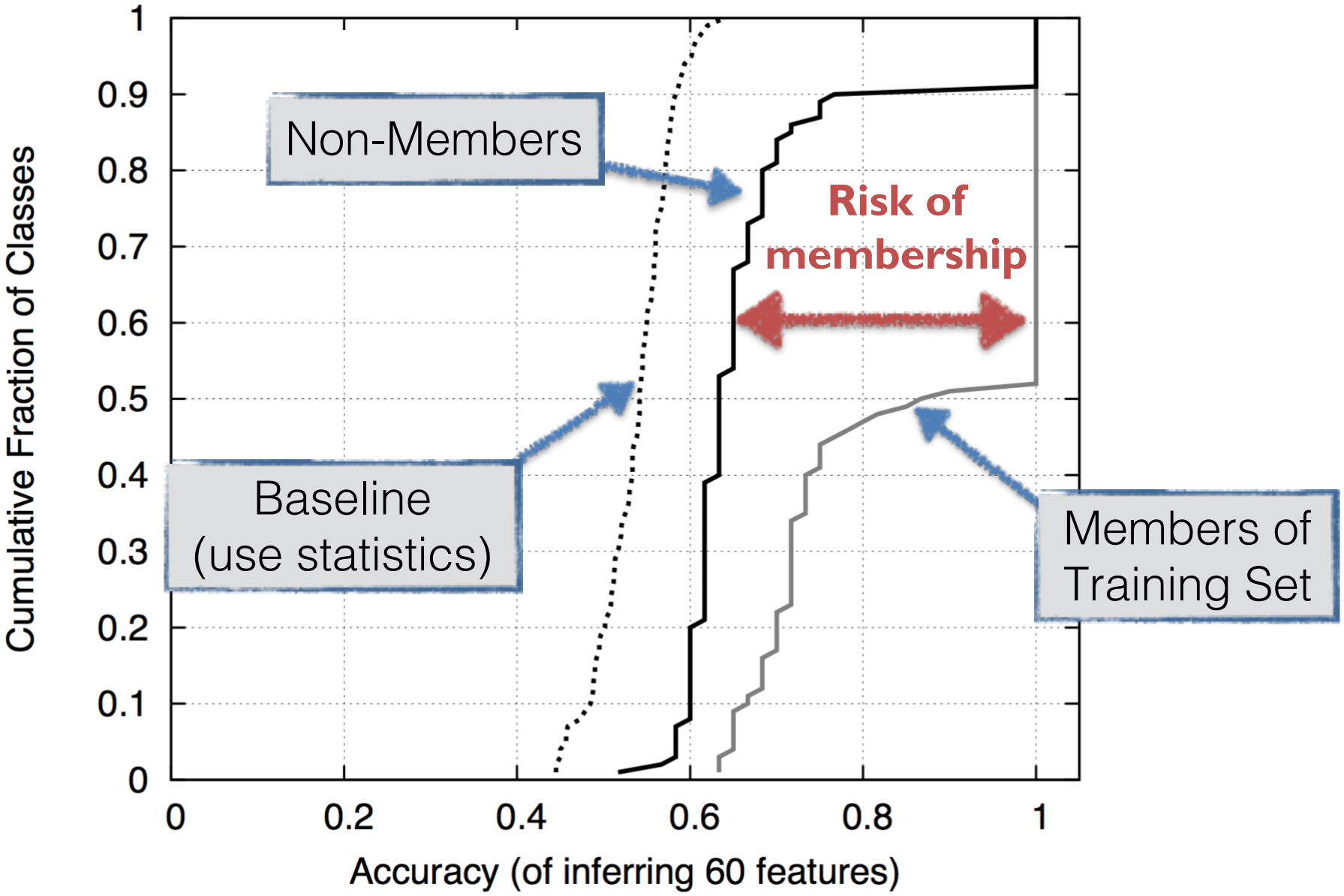
Generalizability Is Not Privacy

- Deep neural networks have huge memorization capacity
- A well-generalized model can still leak information about its training dataset
  - Good test performance on the primary task does not preclude good performance on another task (e.g., membership inference or reconstruction)

Does Inference Breach "Privacy"?



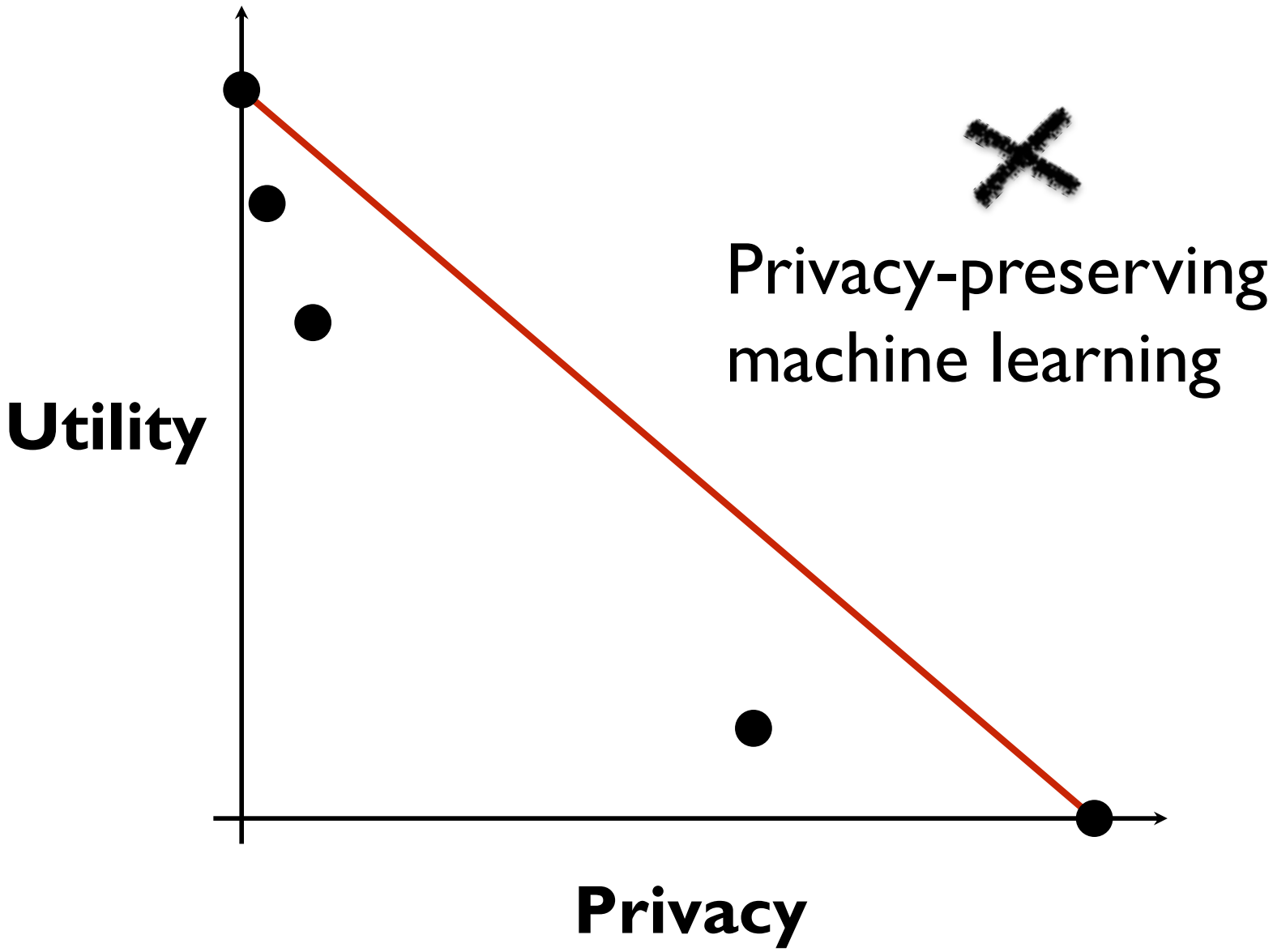
Privacy breach = risk of membership:
Gap between what can be inferred from the model about a member of the training set and an arbitrary input from the population



Purchase Dataset — Classify Customers — Google API

Future

- Modern machine learning is both a threat and an opportunity for data privacy
- For once, privacy and utility are not in conflict: overfitting is the common enemy
  - Overfitted models leak training data
  - Overfitted models lack predictive power
- Need generalizability and accuracy


Utility
Privacy
Privacy-preserving
machine learning

"Classical" Intuition for Privacy

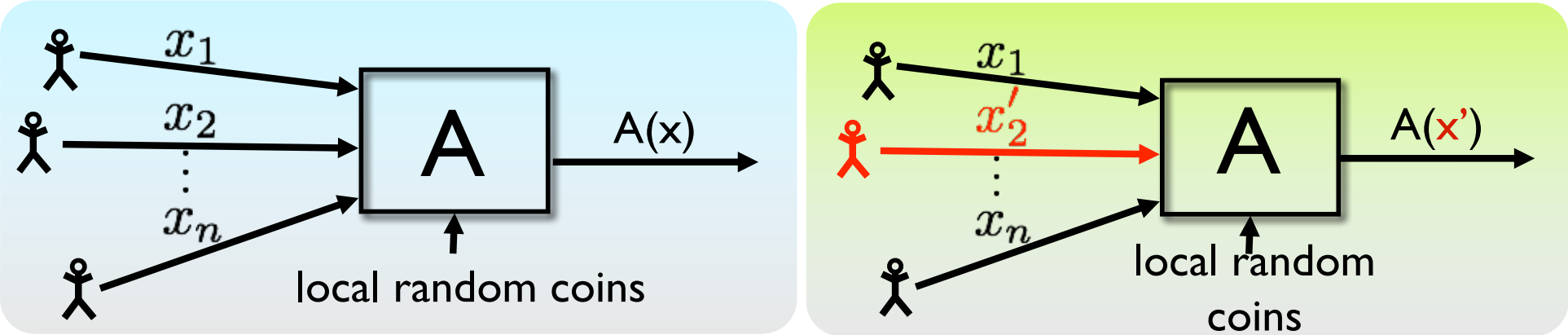
- Dalenius (1977): "If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S, a disclosure has taken place"
  - Privacy means that anything that can be learned about a respondent from the statistical database can be learned without access to the database

Problems with Classical Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
  - What if my (incorrect) prior is that every student in this room has three arms?
- How much is "too much?"
  - Can't achieve cryptographically small levels of disclosure and keep the data useful
  - Users are supposed to learn unpredictable things about the data

Dwork et al.

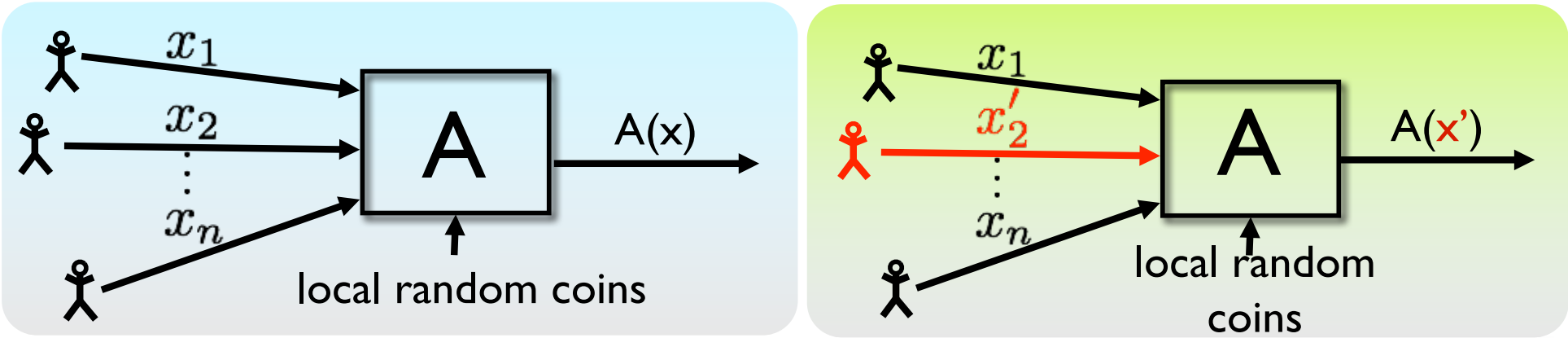
Differential Privacy: Intuition



If you change or remove one person's data, distribution of outputs should not change much

Dwork et al.

Differential Privacy



Neighboring databases induce **close** distributions on outputs

Differ in one data point

Definition: A is ε-differentially private if,

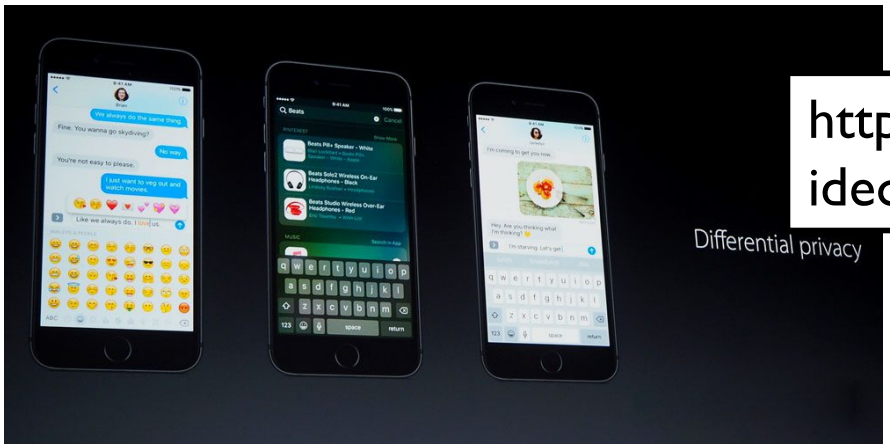
for all neighbors x, x',

for all subsets S of outputs

Measure of information leakage

$$\Pr(A(x) \in S) \leq e^{\epsilon} \cdot \Pr(A(x') \in S)$$

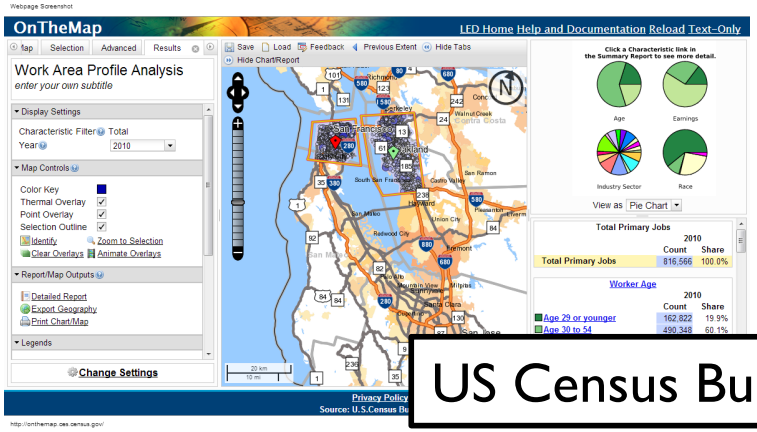
Deployed Differential Privacy



https://developer.apple.com/videos/play/wwdc2016/709/

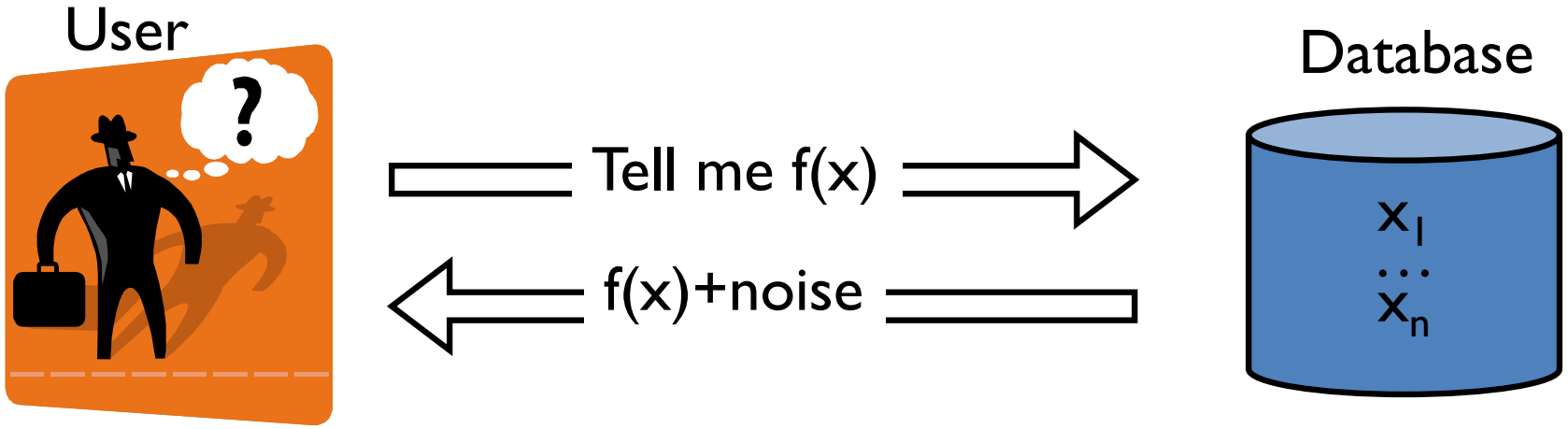
Chrome

https://github.com/google/rappor



US Census Bureau (goal: full census in 2020)

Laplace Mechanism



- Intuition: f(x) can be released accurately when f is insensitive to individual entries $x_1, \ldots x_n$
- Global sensitivity $GS_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$
  - Example: $GS_{average} = 1/n$ for sets of bits
- Theorem: $f(x) + Lap(GS_f/\varepsilon)$ is ε-indistinguishable
  - Noise generated from Laplace distribution

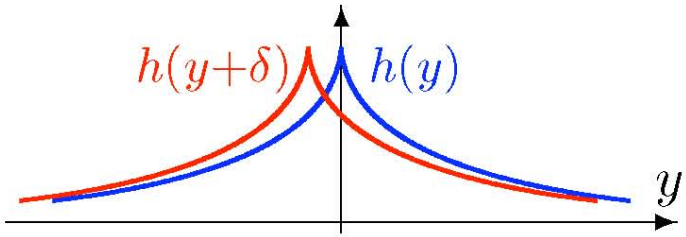
Lipschitz constant of f

Achieving Privacy

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$



Sliding property of $\mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

Proof idea:

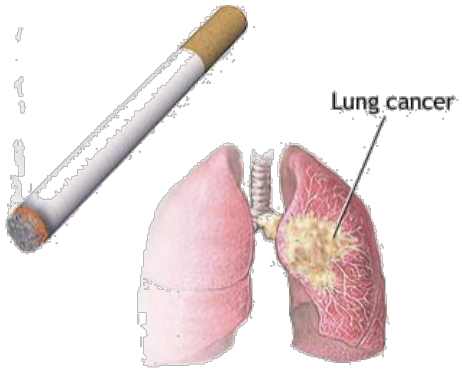
$A(x)$: blue curve

$A(x')$: red curve

$\delta = f(x) - f(x') \leq \mathsf{GS}_f$

What Differential Privacy Is Not

Your beliefs about me are the same after you see the output as they were before



Impossible

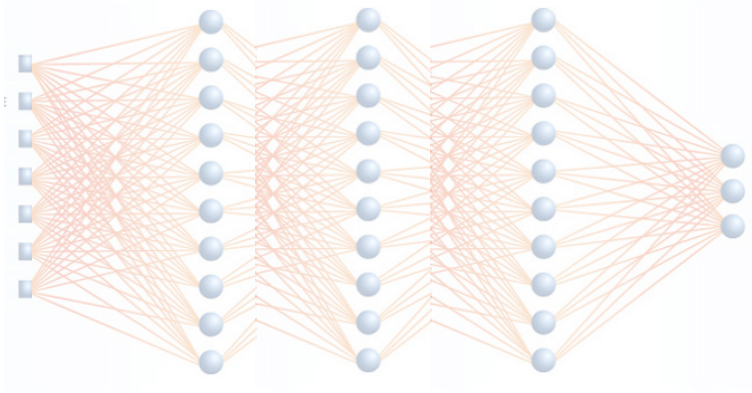
- Suppose you know that I smoke
- Clinical study: "smoking and cancer correlated"
- You learn something about me
  … whether or not my data were used

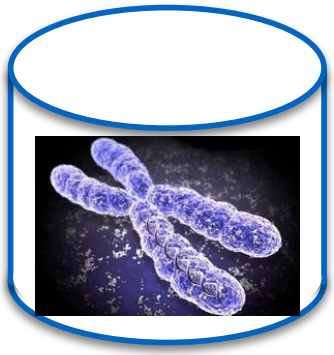
What Differential Privacy Means

You learn (almost) the same things about me whether or not my data are used

No matter what you know ahead of time

Learning today



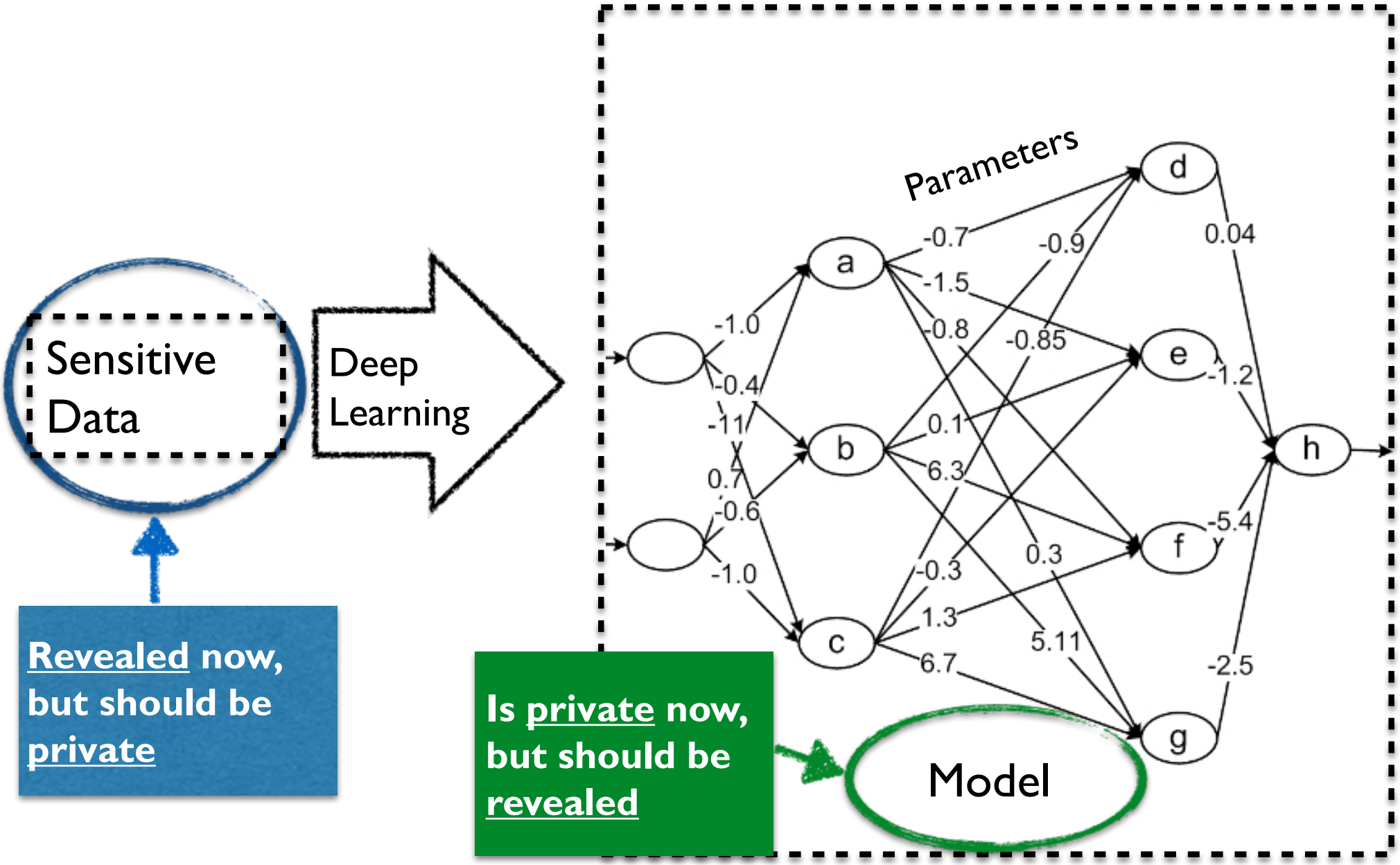


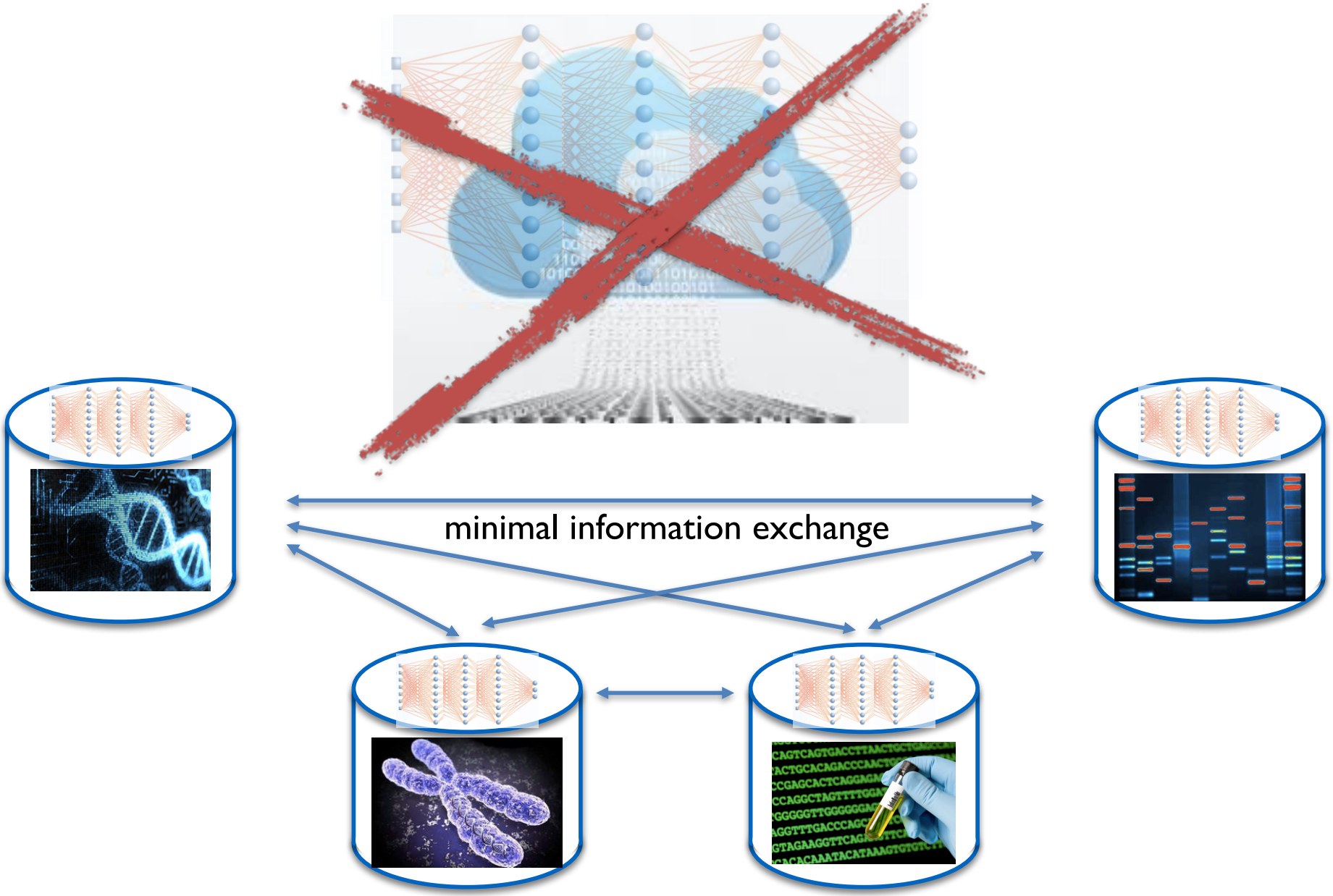
Privacy Concerns

- Training data is sensitive
  - speech, photo images, written documents
- Users have no control over the learning objective
- Using trained networks requires users to share their private data with service providers

Possible Consequences

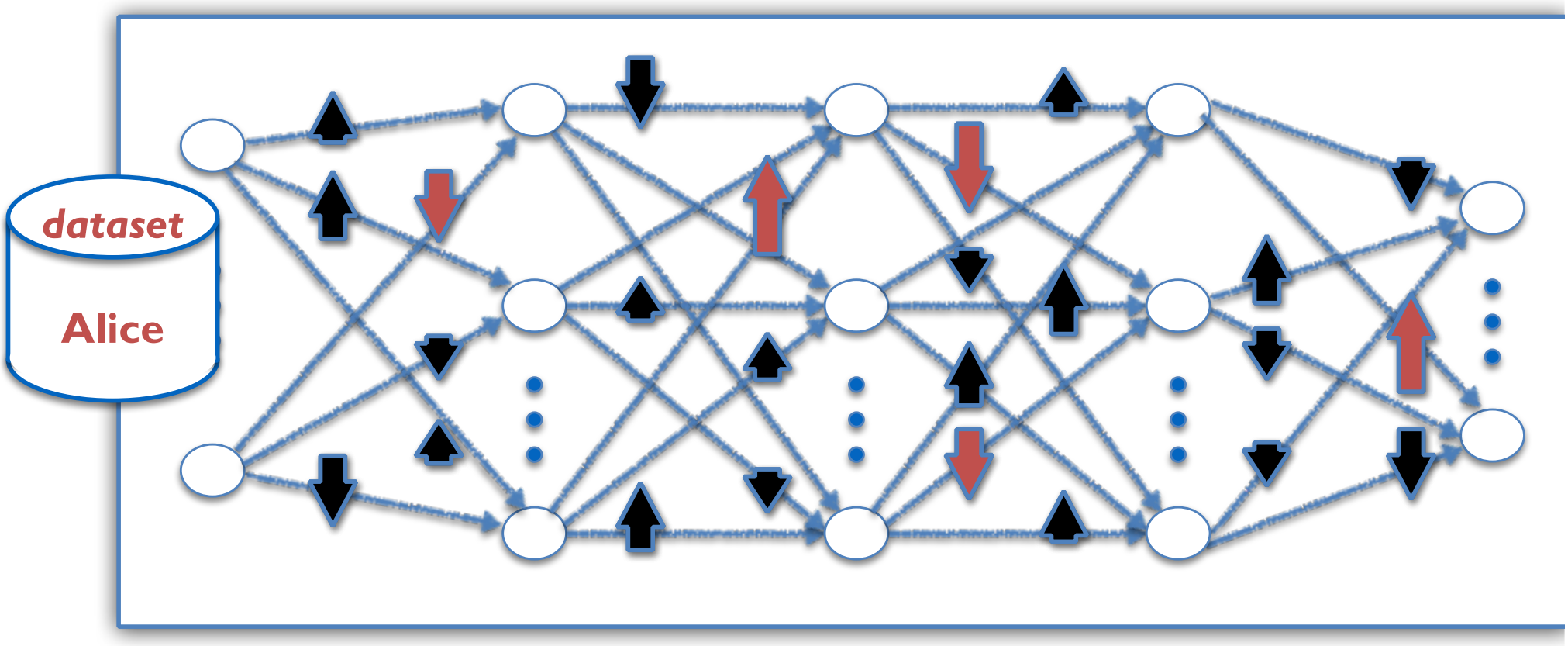
- Users' data might be used in wrong context
  - Compromises and data breaches
  - Inference of sensitive information
  - Training of intrusive models
- Holders of sensitive data cannot benefit from large-scale deep learning because they may not share or pool their datasets for training
  - Biomedical researchers?
  - Social scientists?


Sensitive Data
Deep Learning
Parameters
a
b
c
d
e
f
g
h
-1.0
-0.4
-11
0.7
-0.6
-1.0
-0.7
-1.5
-0.8
-0.9
-0.85
0.1
6.3
0.3
-0.3
1.3
6.7
5.11
0.04
-1.2
-5.4
-2.5
Revealed now, but should be private
Is private now, but should be revealed
Model



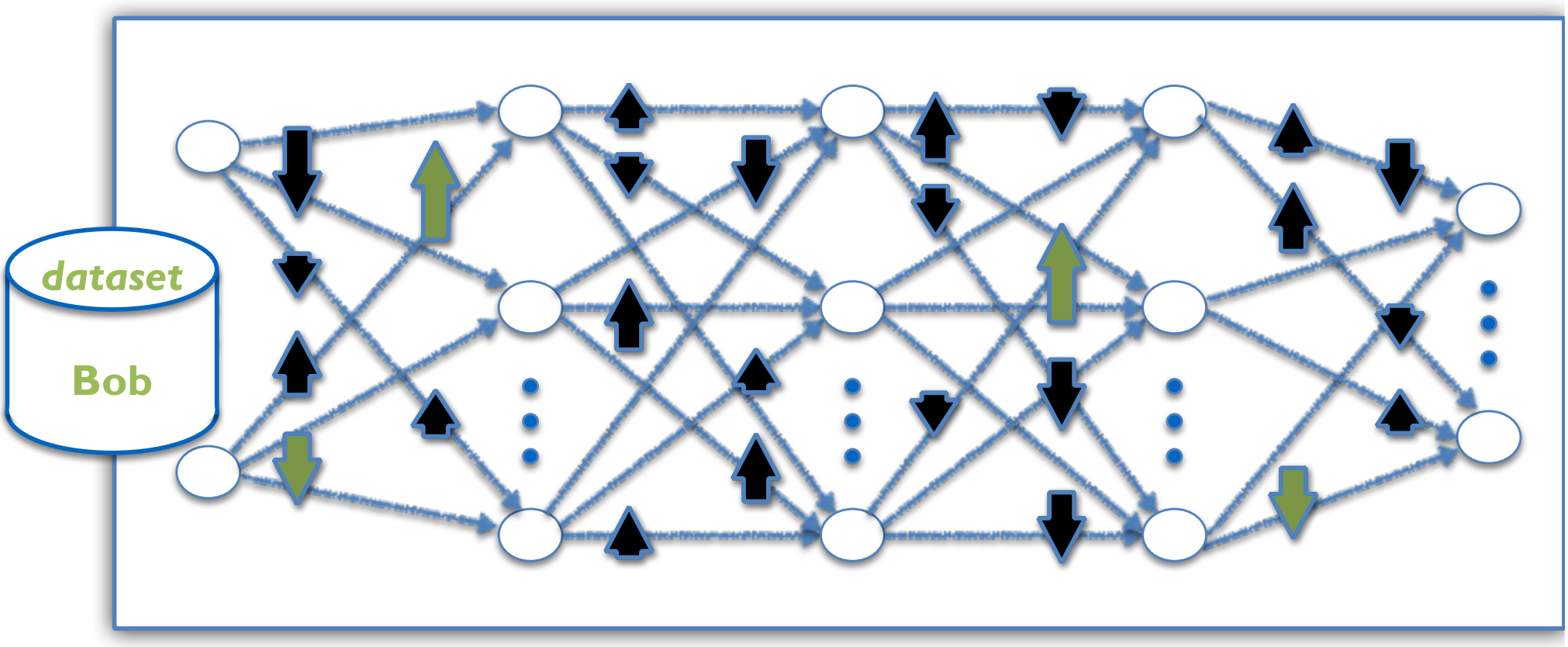
Distributed Selective SGD (DSSGD)

Selective SGD

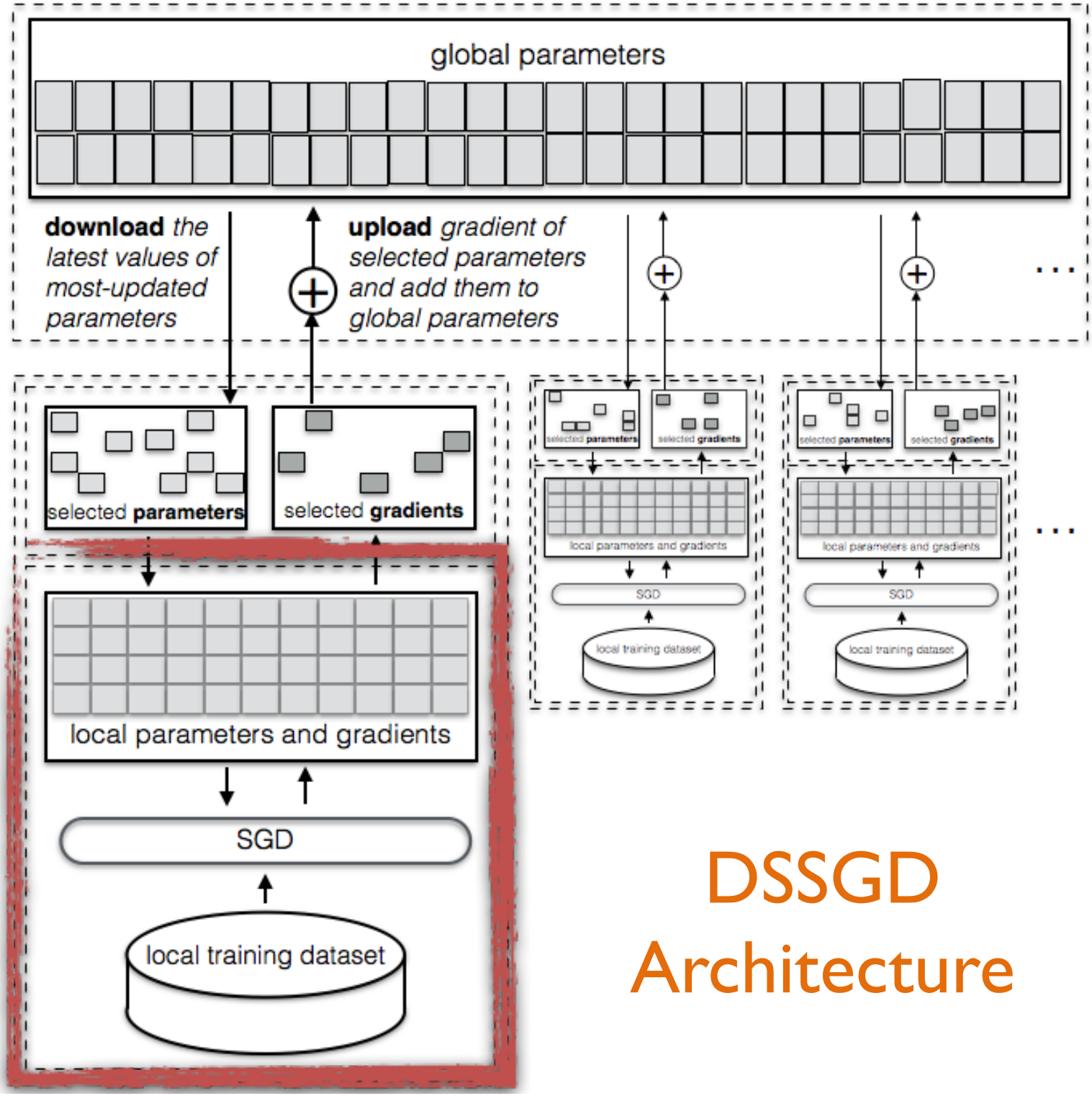


Share with others

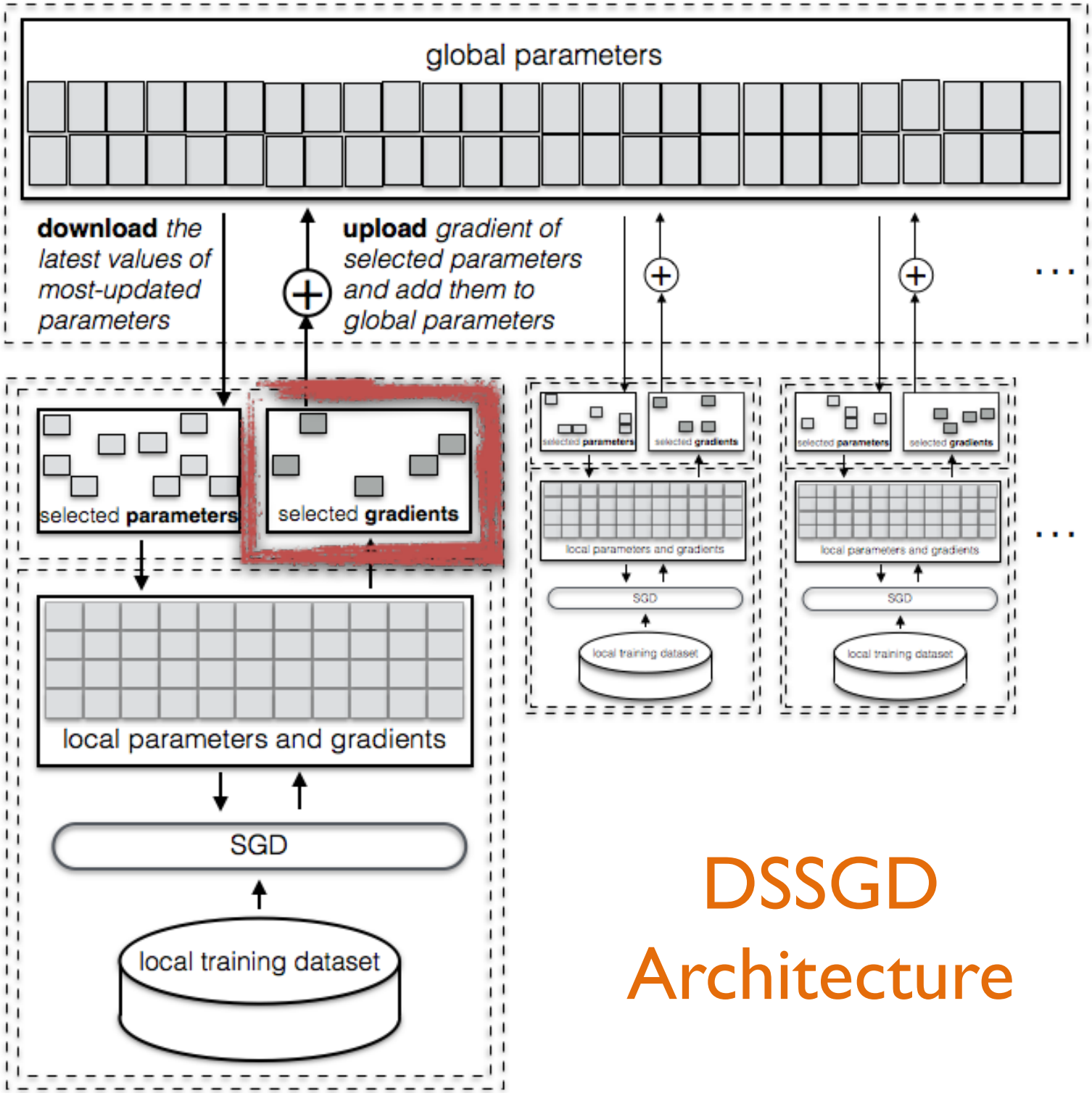
Selective SGD



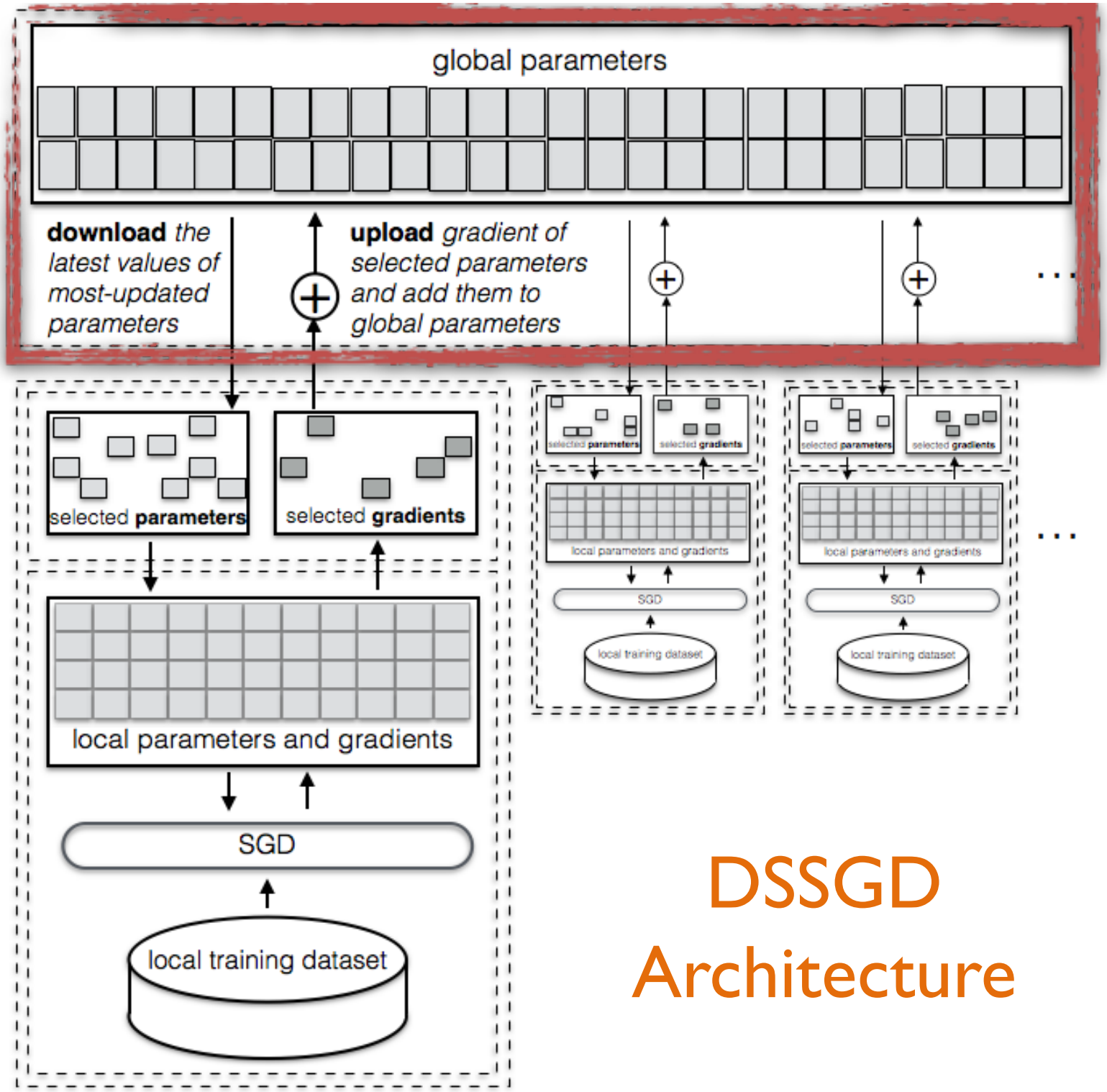
Share with others



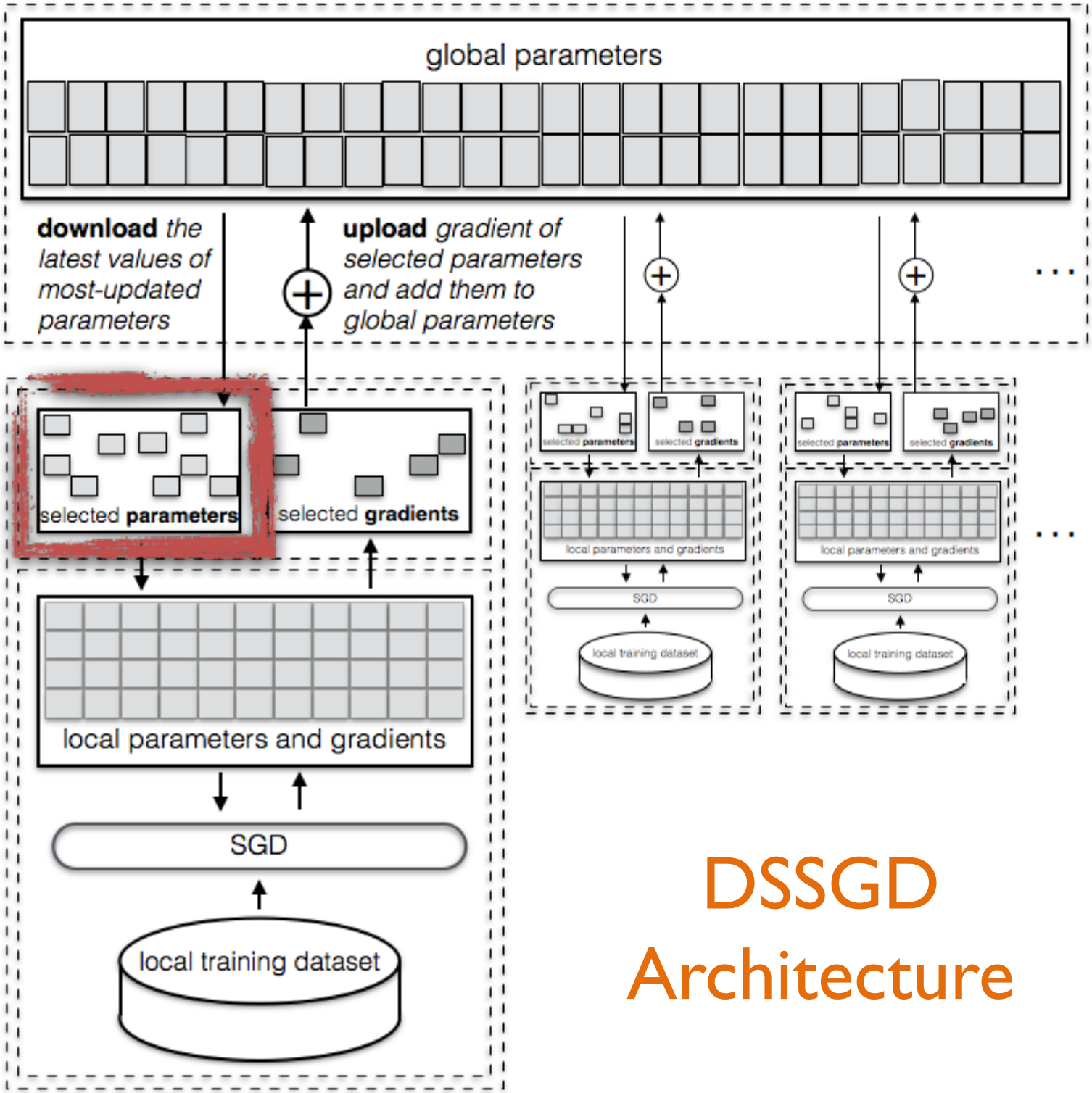
DSSGD Architecture



DSSGD Architecture



DSSGD Architecture



DSSGD Architecture

Distributed Selective SGD

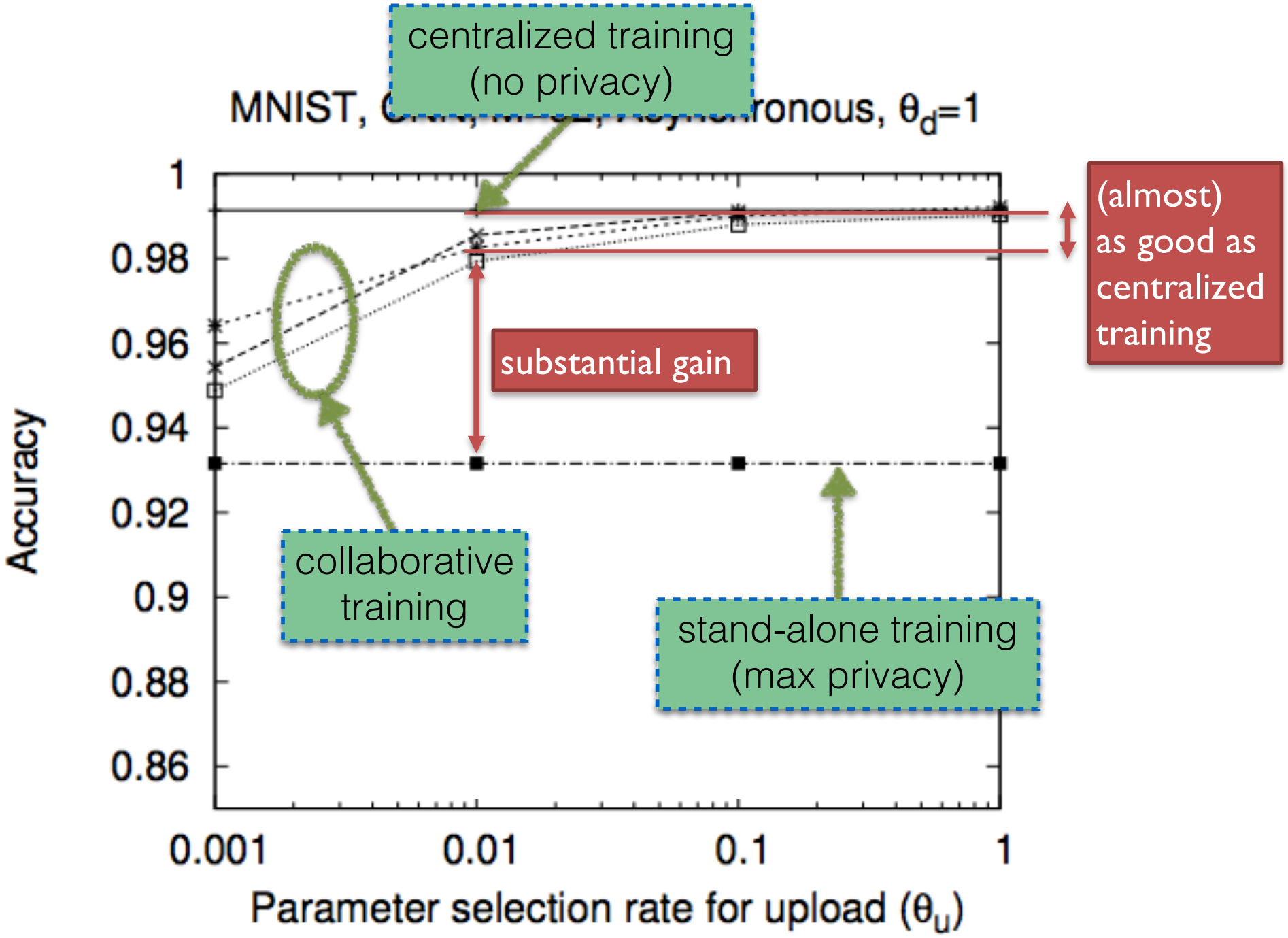
- Local training, global convergence
- High training stochasticity
- Less overfitting

Evaluation Datasets

MNIST

SVHN

Task: Find the digit in the image — 10 class classification

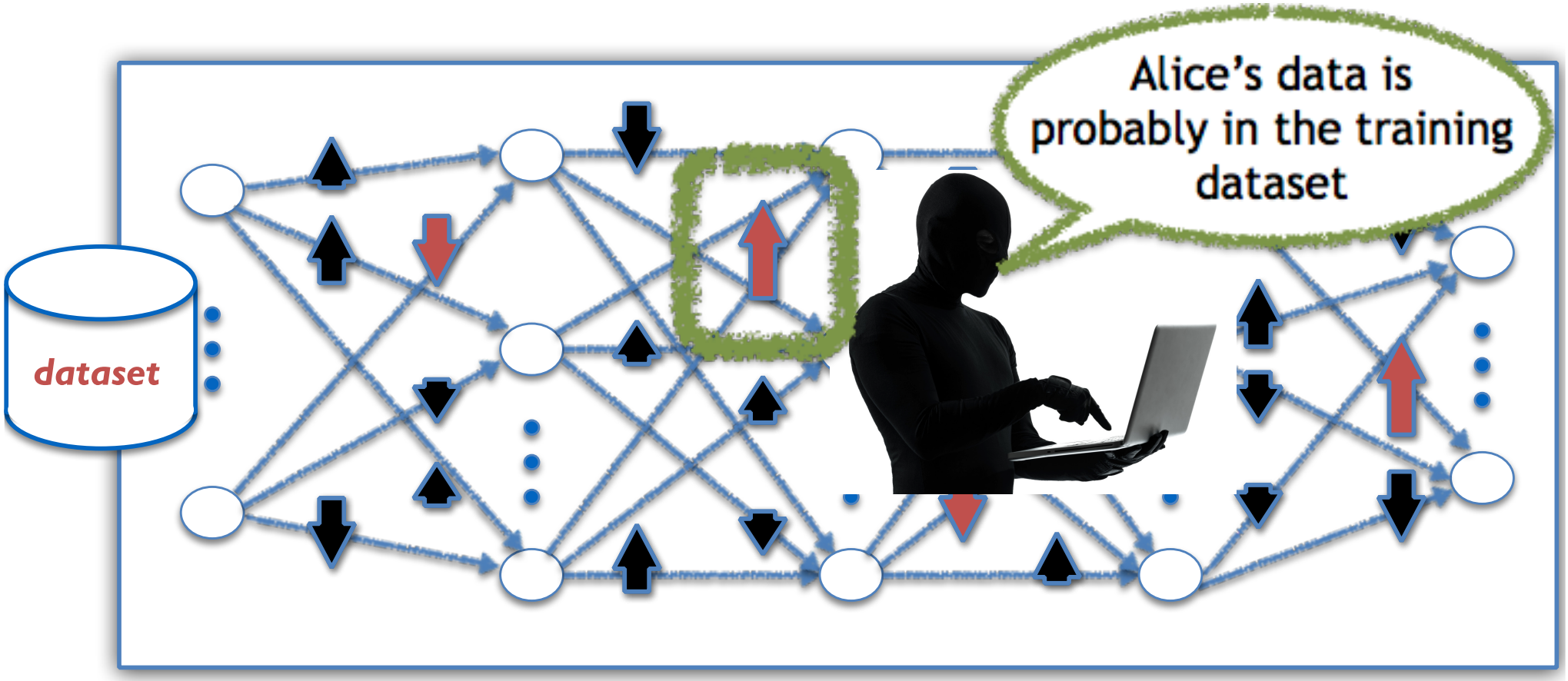

centralized training
(no privacy)
MNIST, ronous, θd=1
(almost)
as good as
centralized
training
substantial gain
collaborative
training
stand-alone training
(max privacy)
Accuracy
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.001
0.01
0.1
1
Parameter selection rate for upload (θu)

Privacy Properties

- Participants' datasets remain private
- Full control over parameter selection
- Known learning objective
- Resulting model available to all parties

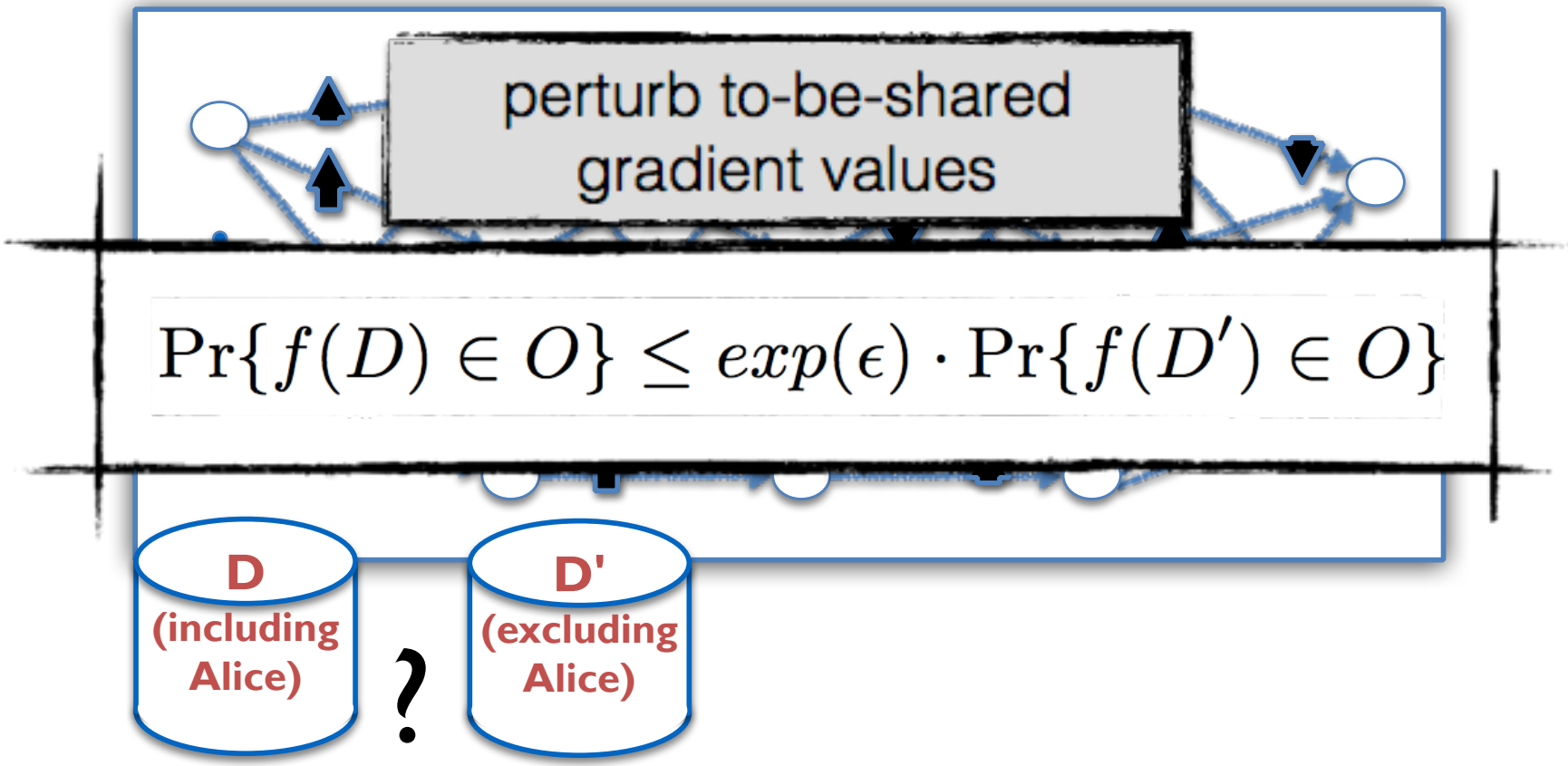
Indirect Information Leakage

through gradient sharing



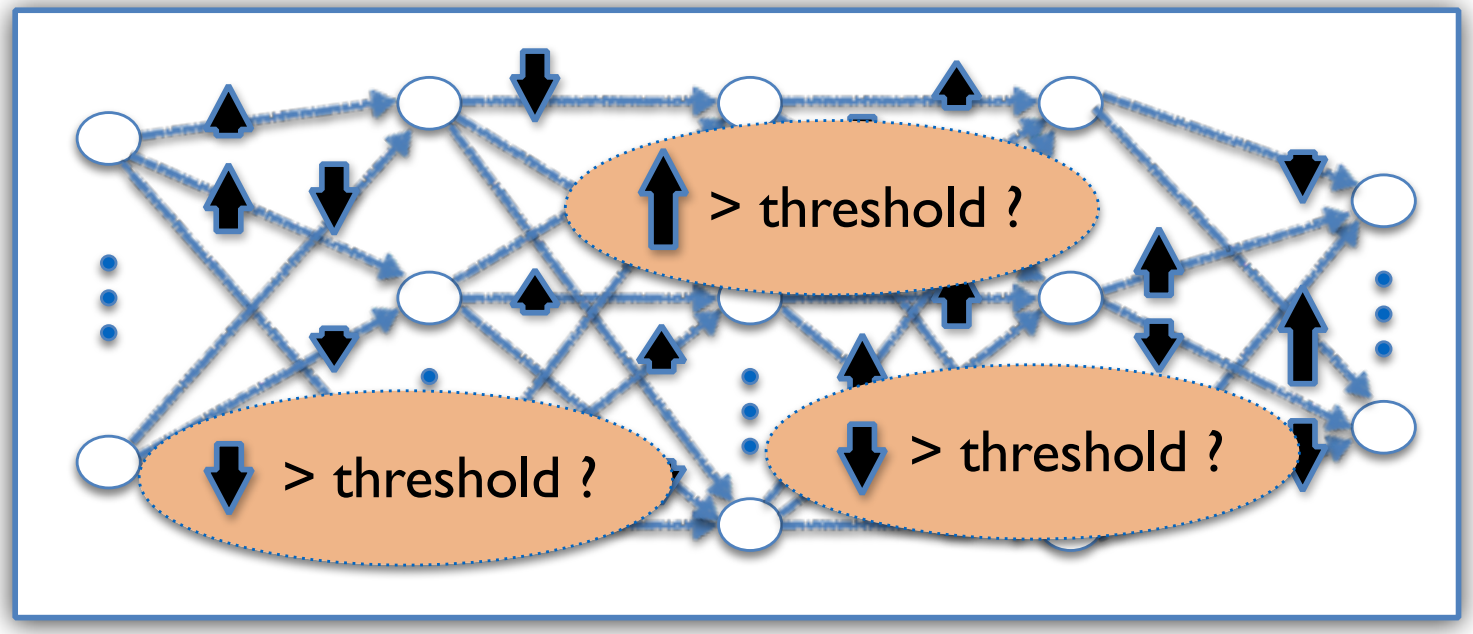
Prevent Indirect Leakage

- Differentially private parameter selection and gradient sharing



Sparse Vector Technique

- Select a small fraction of (perturbed) gradients that are above a given (perturbed) threshold



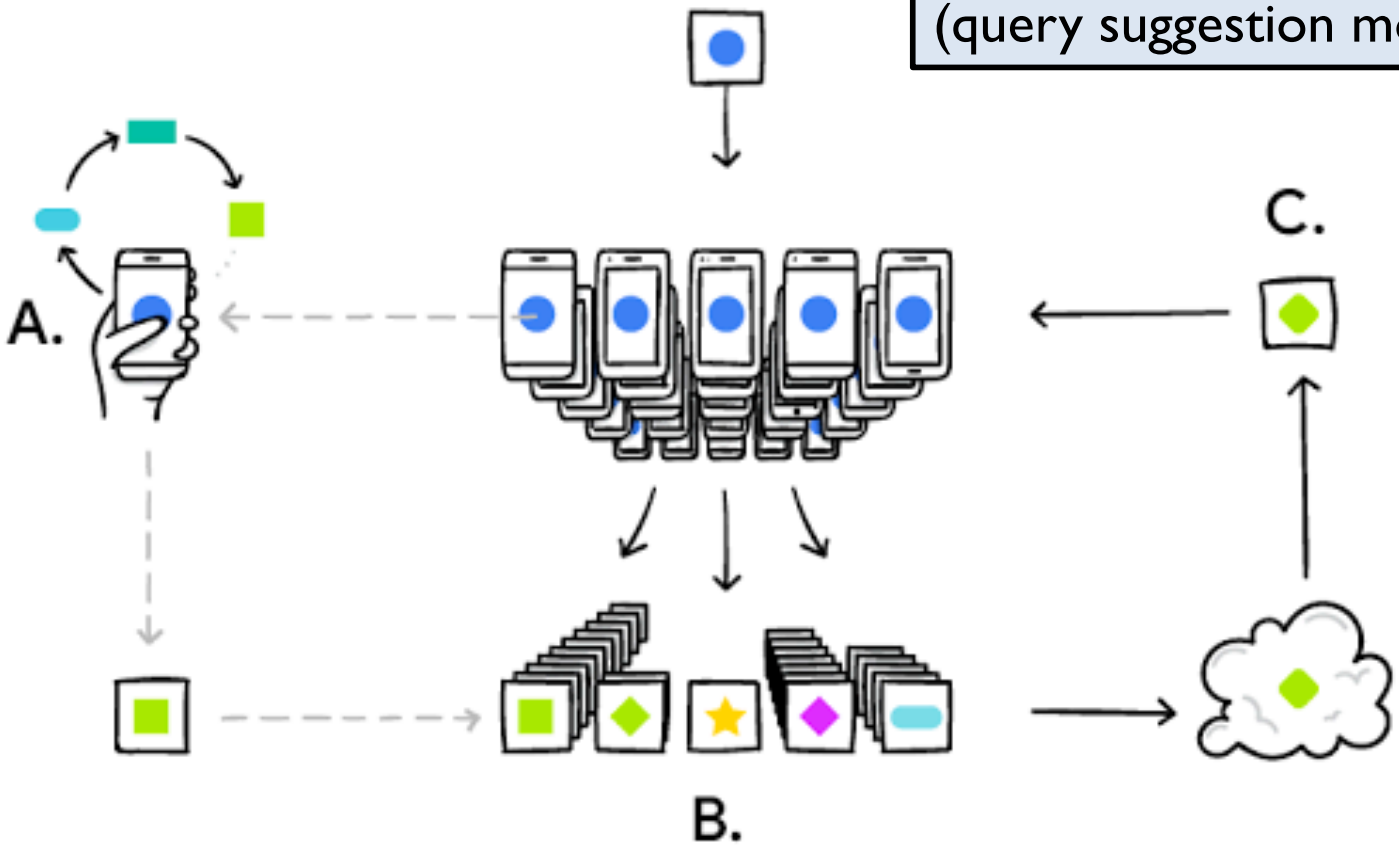
C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

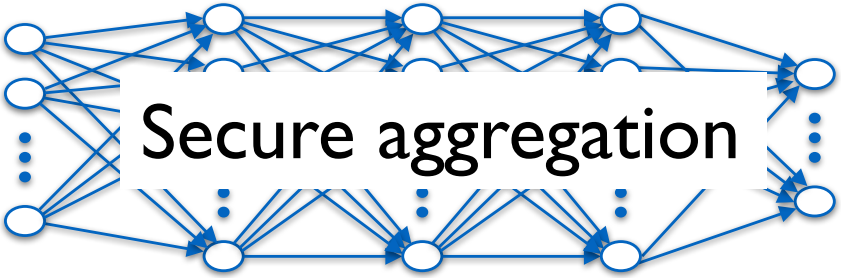
differentially private comparison

Federated Learning: Collaborative Machine Learning without Centralized Training Data

Thursday, April 06, 2017

Currently being tested by Google in Gboard on Android (query suggestion model)





Compressed parameter updates
(not just raw gradient steps)

Each phone trains
a local model