Machine Learning and Privacy

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 · Commonsense

PERSONAL ASSISTANT
Brevia · DiligenceEngine

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

MEDICAL
Parzival · transcriptic · ZEPHYR · Genescient · grand round table · bina · TUTE · MICROSCAN · IVISYS · BOULDER IMAGING

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

FINANCE
Colorain · newsle · ARRIA · SAILTHRU · wavii · Owlin · NarrativeScience · Summly · Prismatic · ai · Affirm · inVenture · zestfinance · BILL GUARD · LendUp · LendingClub · Kabbage

PHILANTHROPIES
DataKind · thorn · DATA GUILD

AUTOMOTIVE
Google · Continental · TESLA · MOBILEYE · CRUISE

DIAGNOSTICS
enlitic · 3SCAN · lumiata · ENTOPSY

RETAIL
BAY SENSORS · PRISM SKYLABS · celect · euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence · MAGIC LEAP · 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 · evolution robotics

EMOTIONAL RECOGNITION
affectiva · BEYONDVERBAL · 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

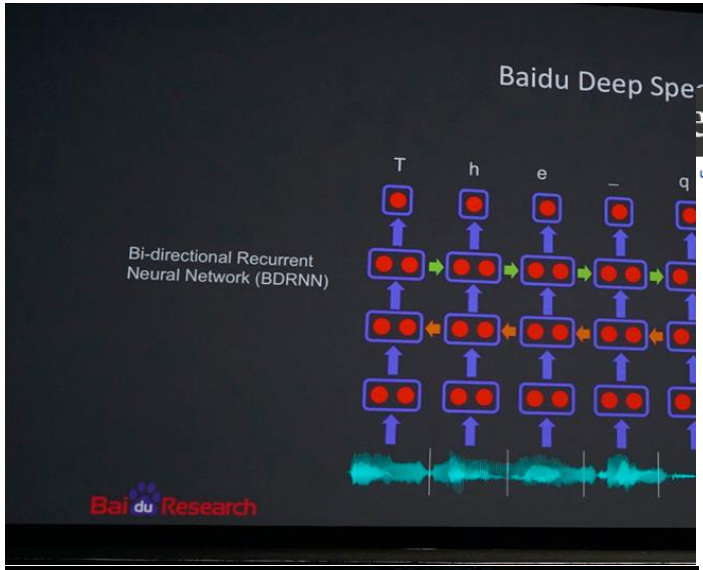
rbes | New Posts +1 | Most Popular | Lists | Video | Subscribe to Forbes

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

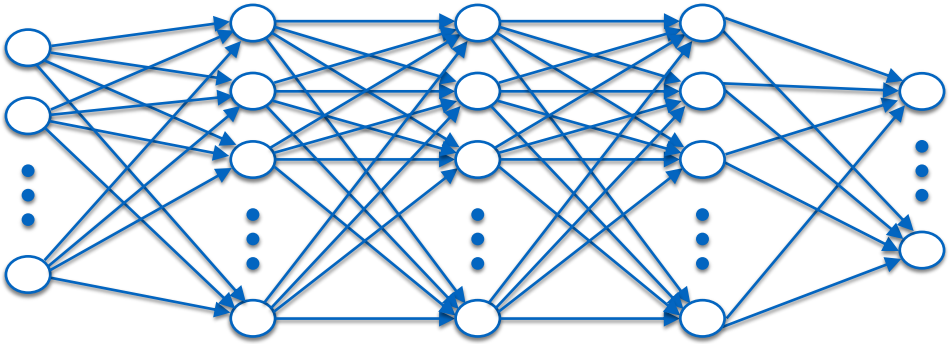
Michael Thomsen
Contributor

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

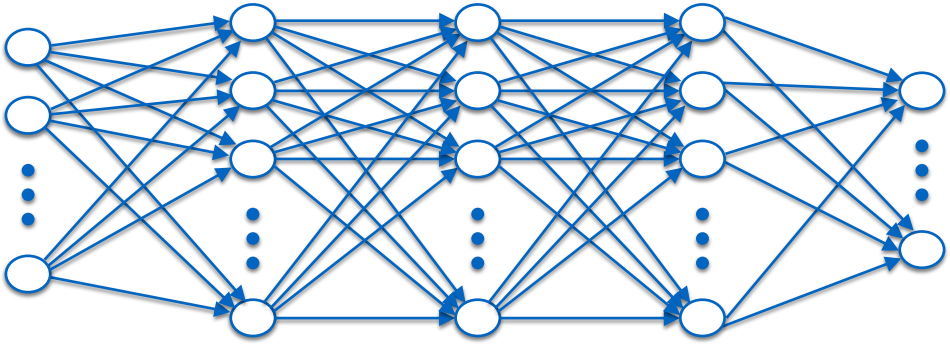
Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

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

Typical Task: Classification



Query

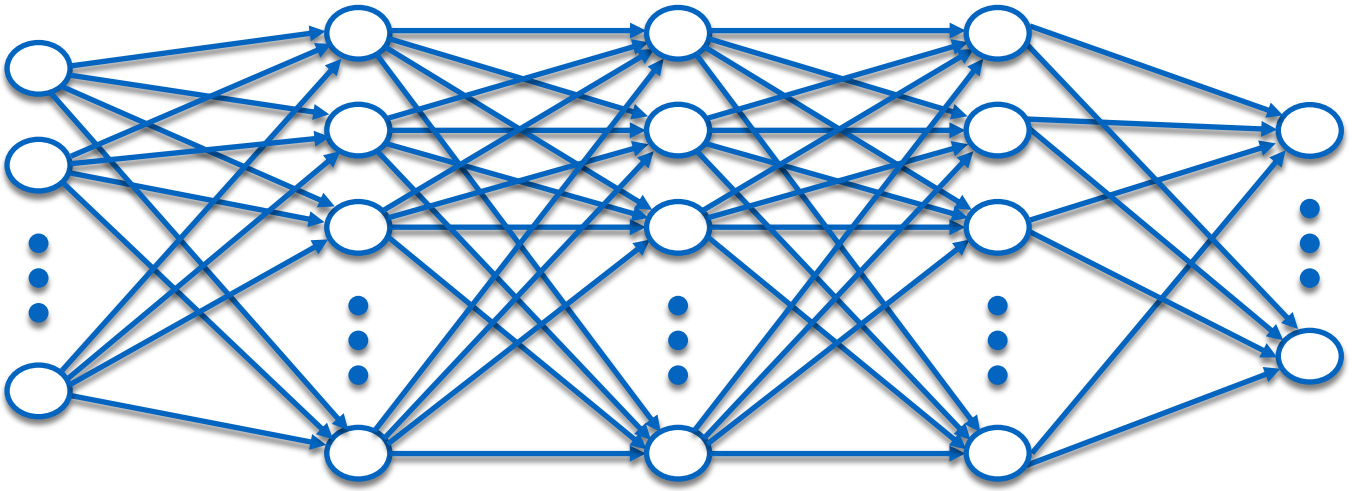


Classification result

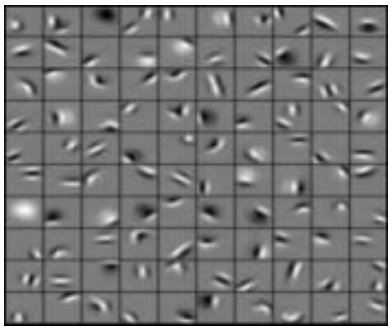
airplane
automobile
...
ship
truck

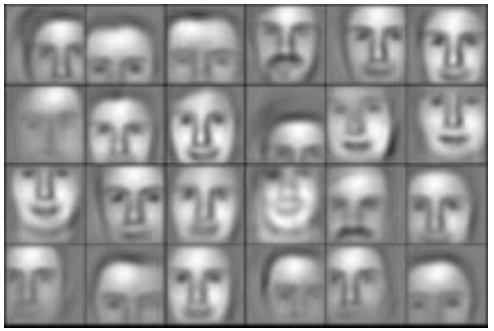
Deep Neural Networks

input

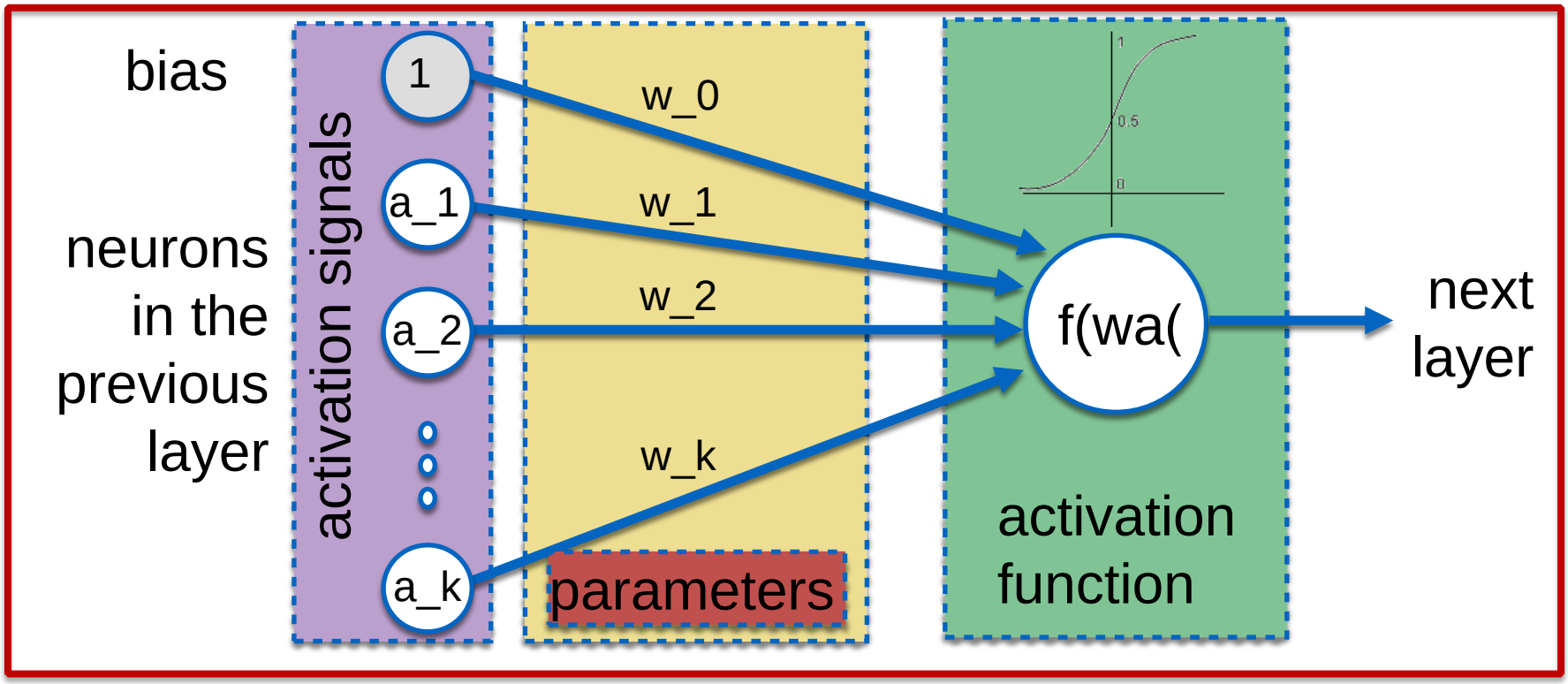


output



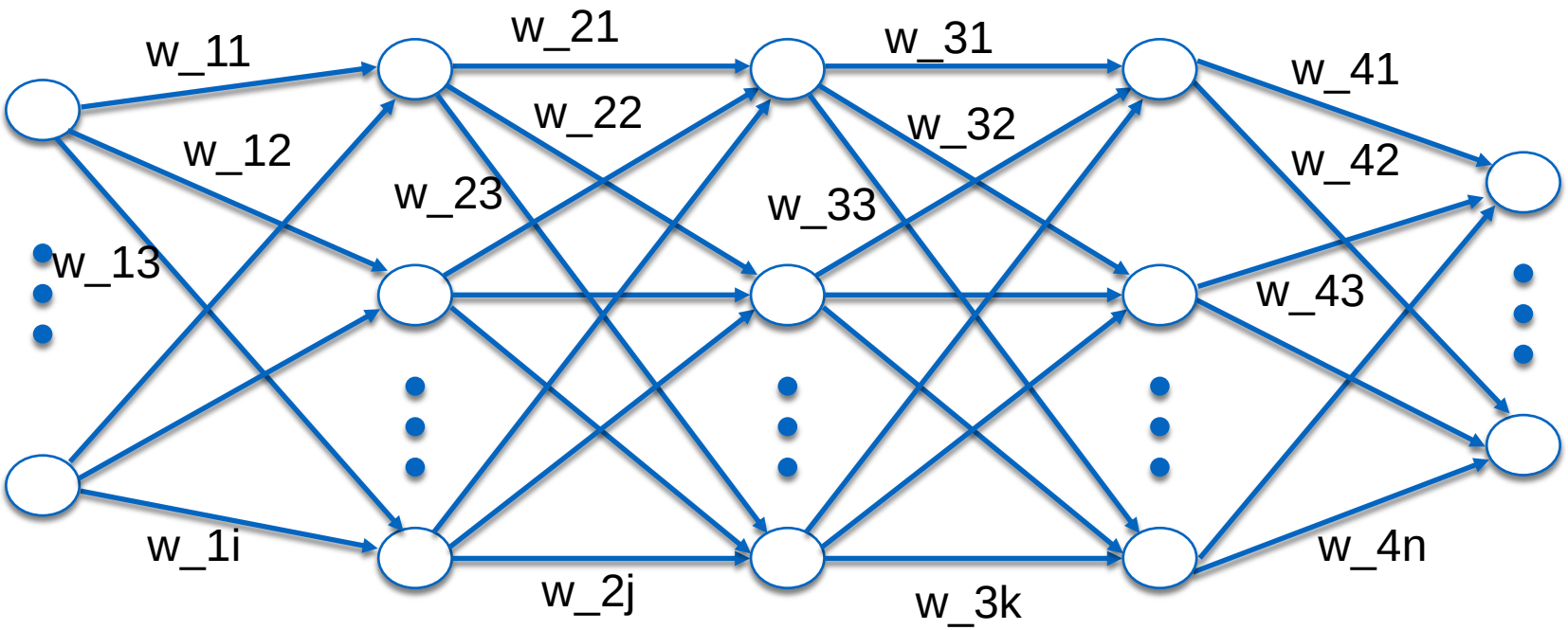


Deep Neural Networks



Learn parameters using
Stochastic Gradient Descent (SGD(

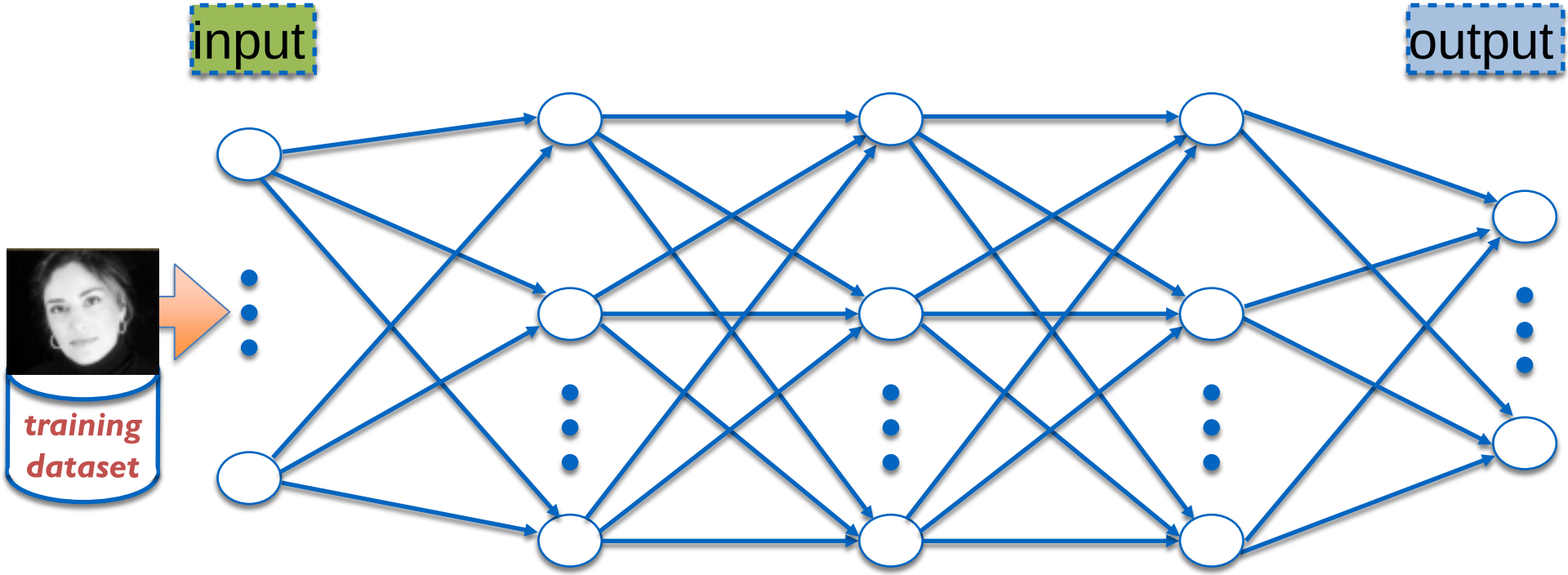
Parameter Training using SGD



Find parameters that minimize the classification error

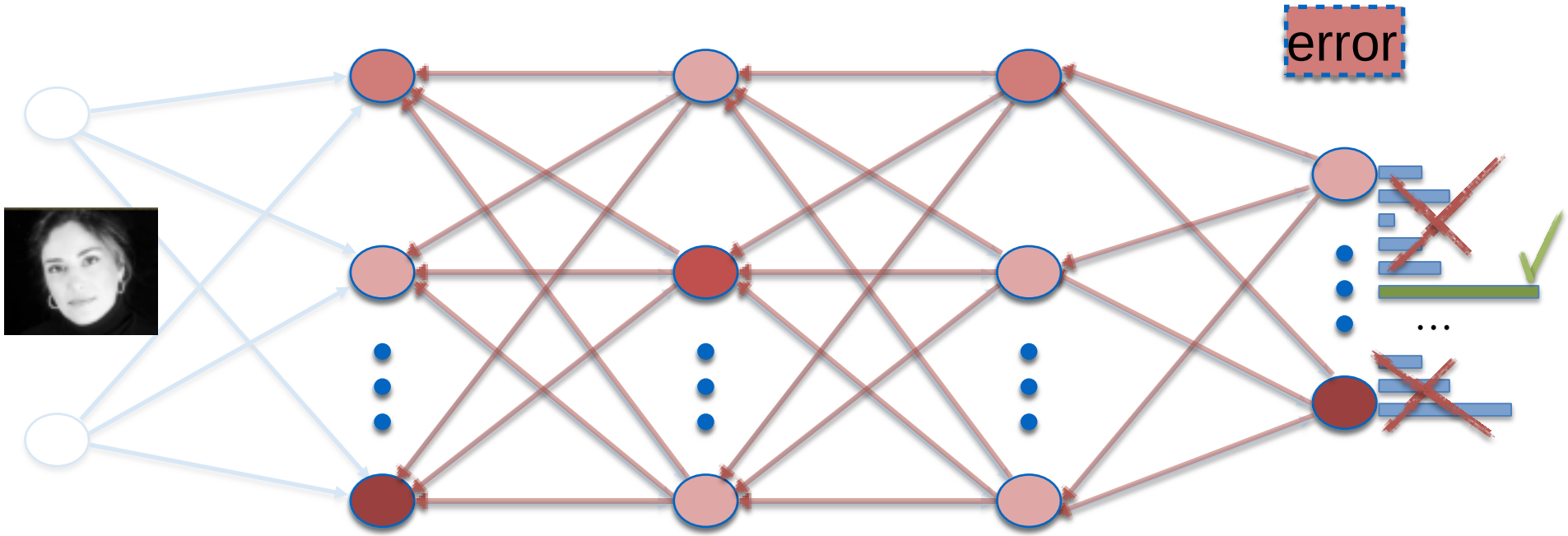
Parameter Training using SGD

(1Feed-Forward

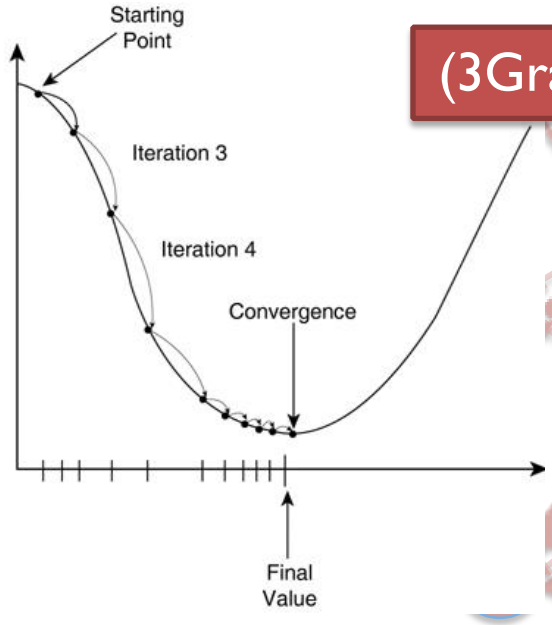


Parameter Training using SGD

(2Back-propagation

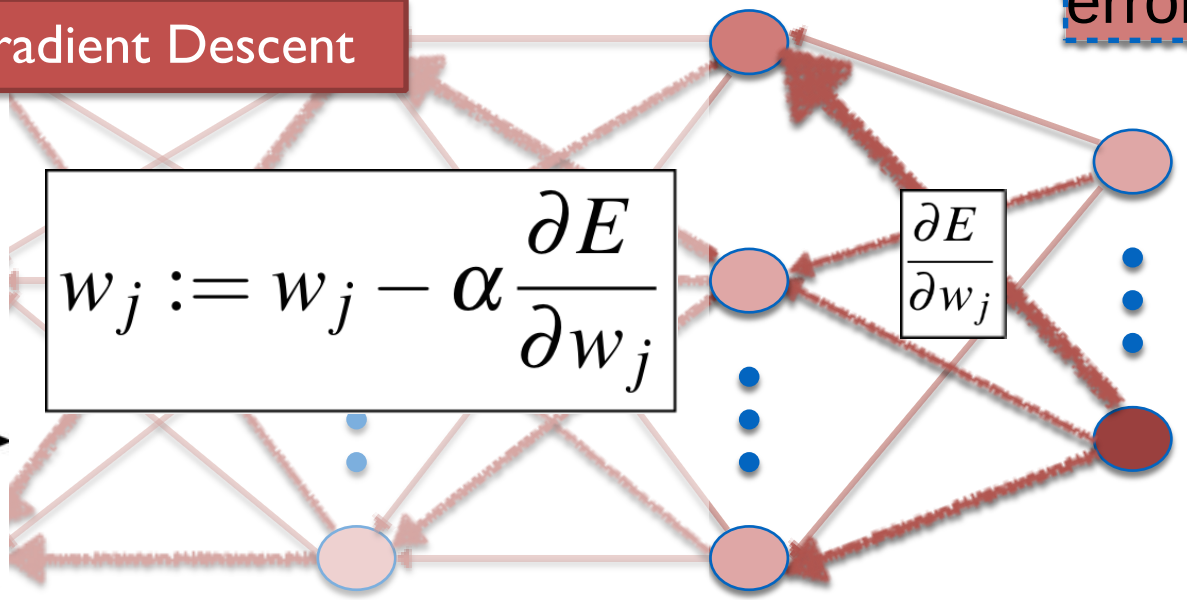


Parameter Training using SGD



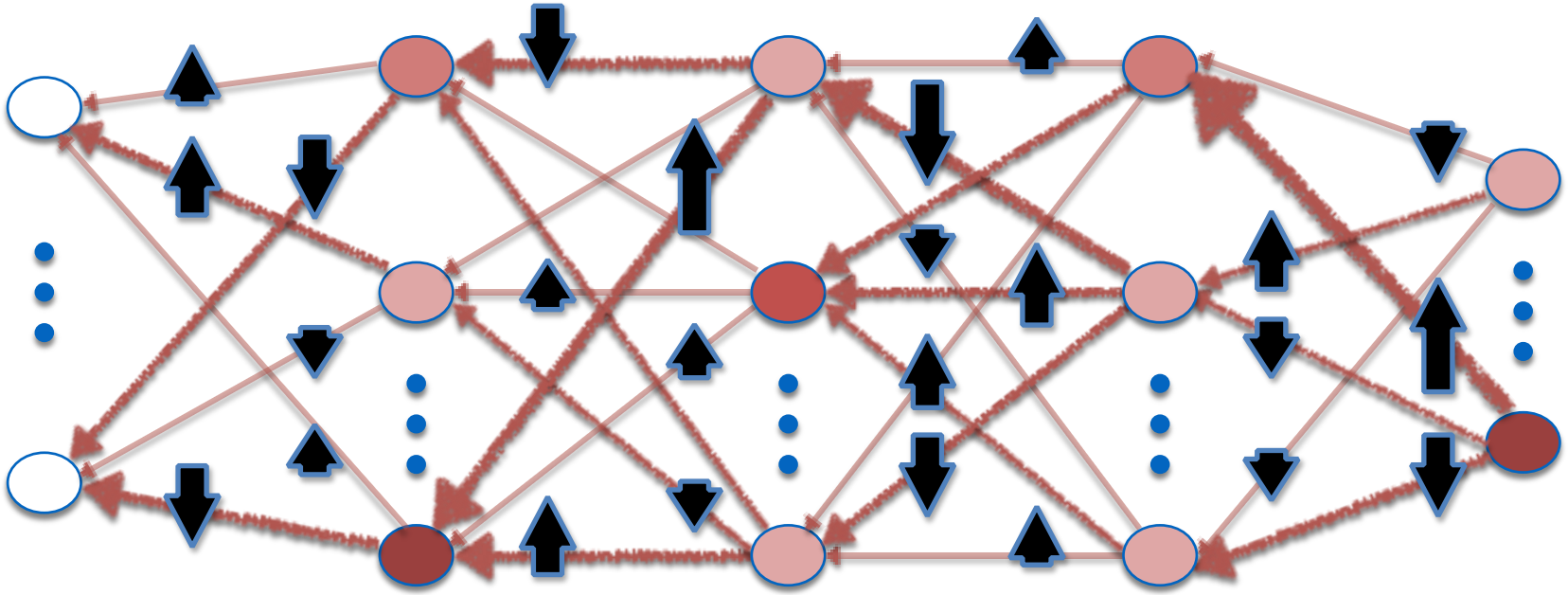
(3Gradient Descent

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



Parameter Training using SGD

Parameter Update



Repeat for new batches of training data

2014

Users' data

Services

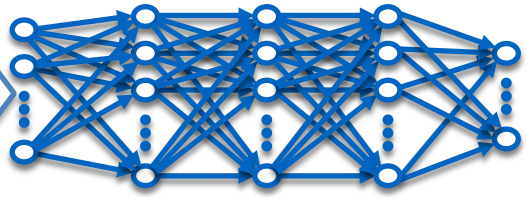
Threats

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

2018

Users' data

Machine learning



Services

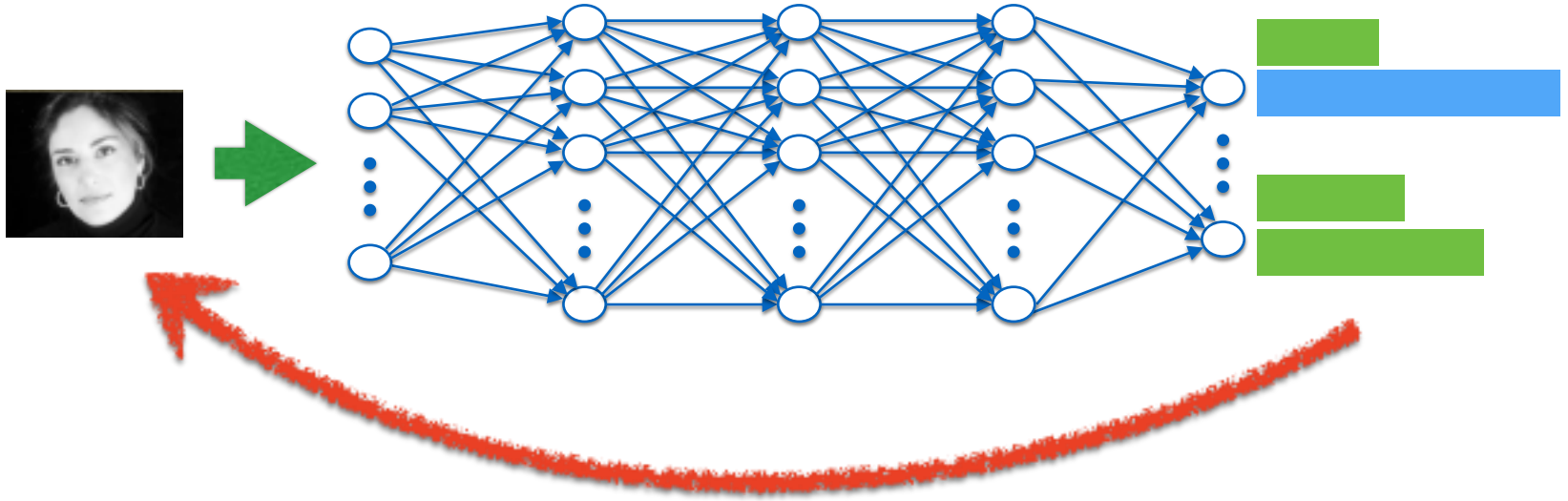
Do trained models leak sensitive data?

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

Is it possible to keep the model itself private ?

Model Inversion

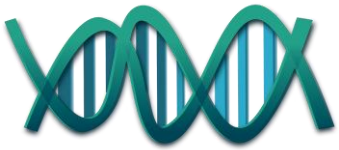
Fredrikson et al.

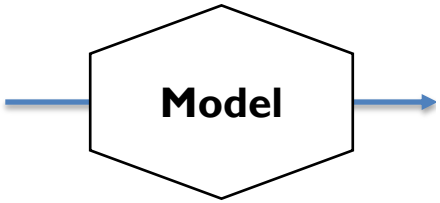


Given an output of a machine learning model, infer something about the input

"unexpected attributes"

Model Inversion in Action



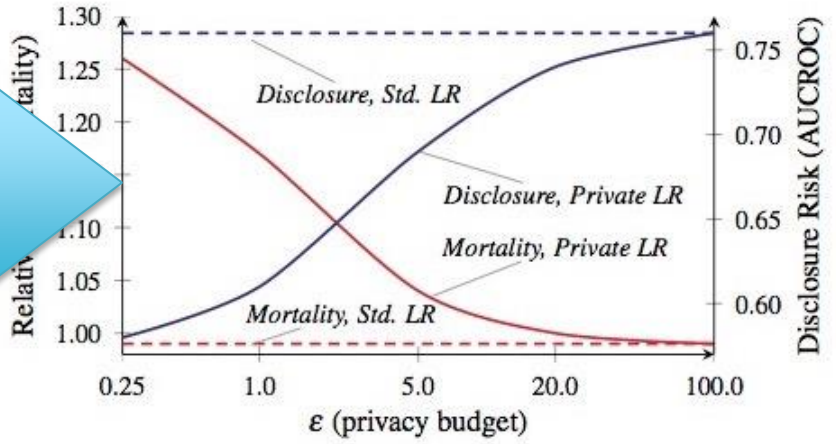


given patient's genome…

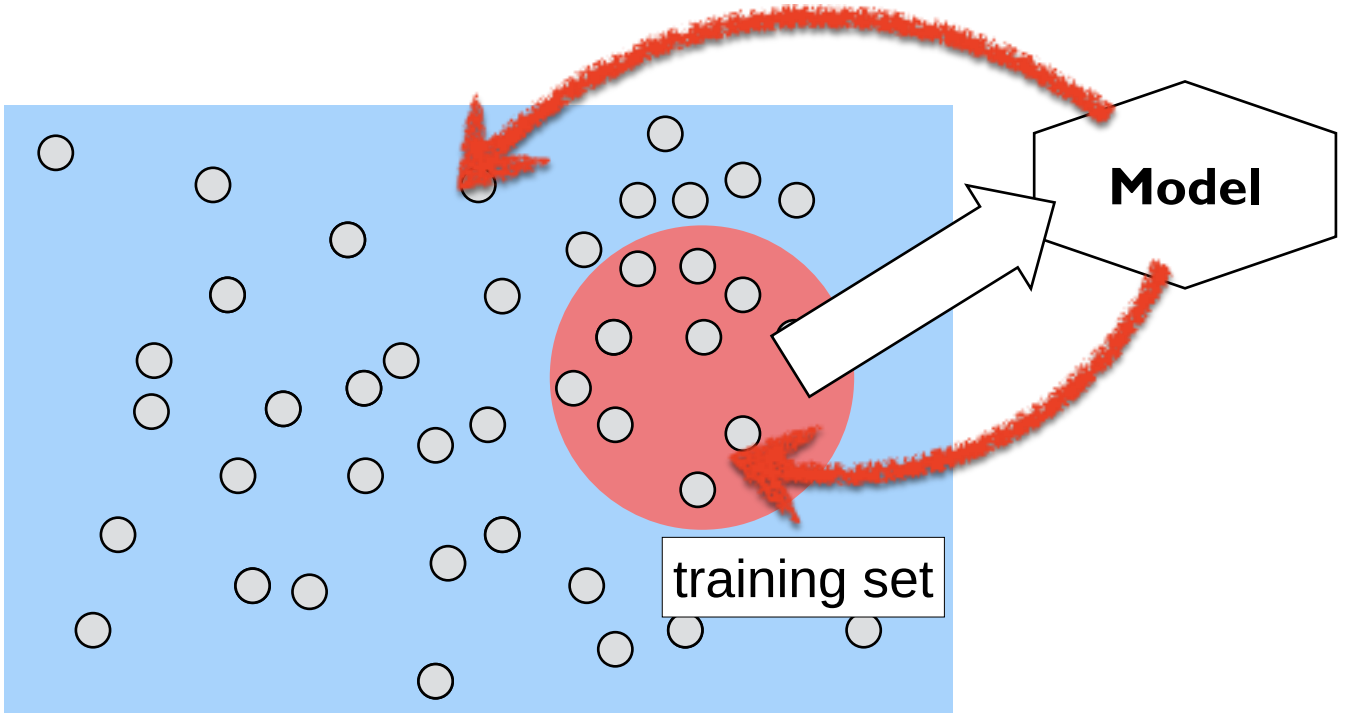
…determine correct warfarin dosage

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

What does this chart measure?



Does Inference Breach Privacy?



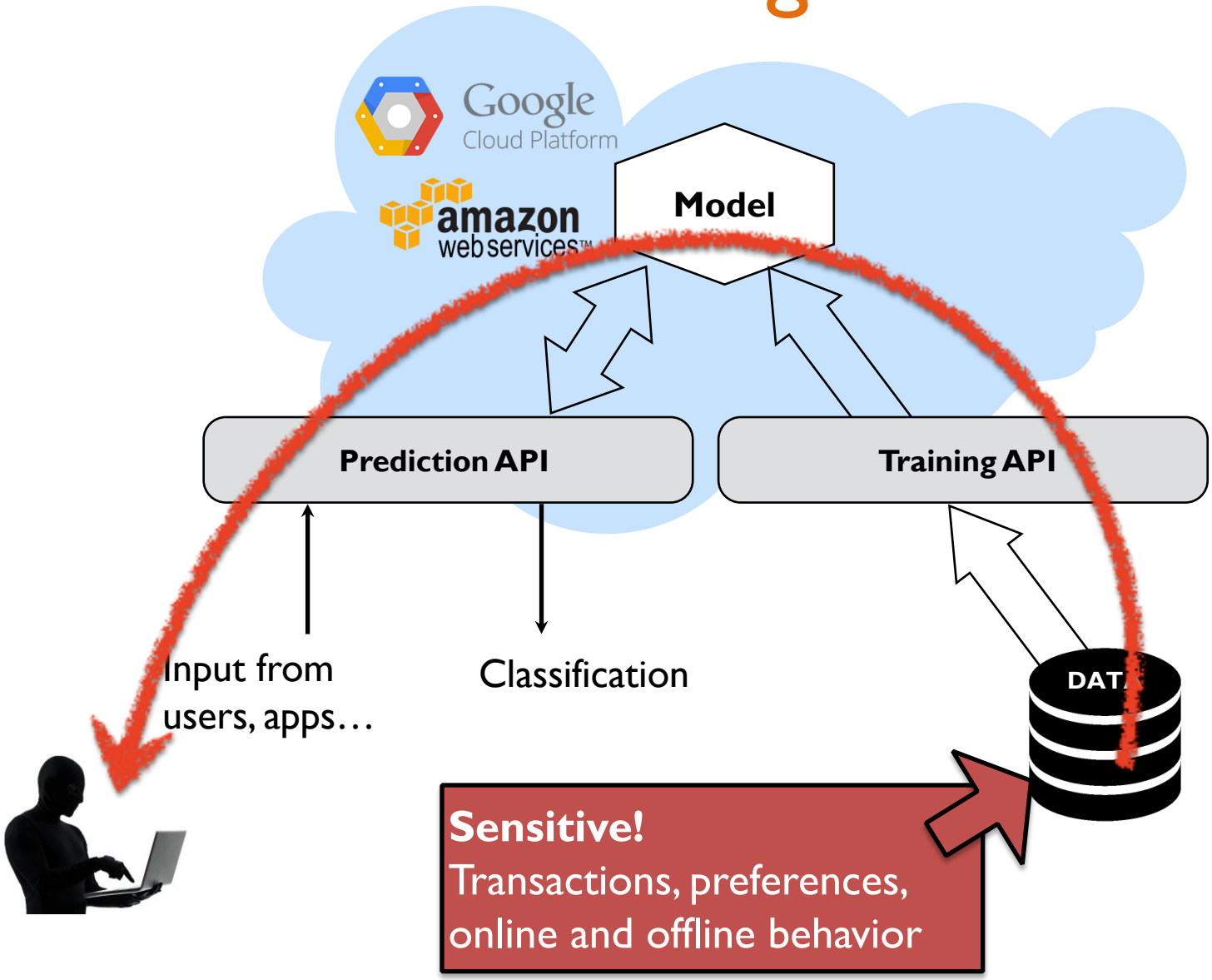
Recommended Reading

Frank McSherry .

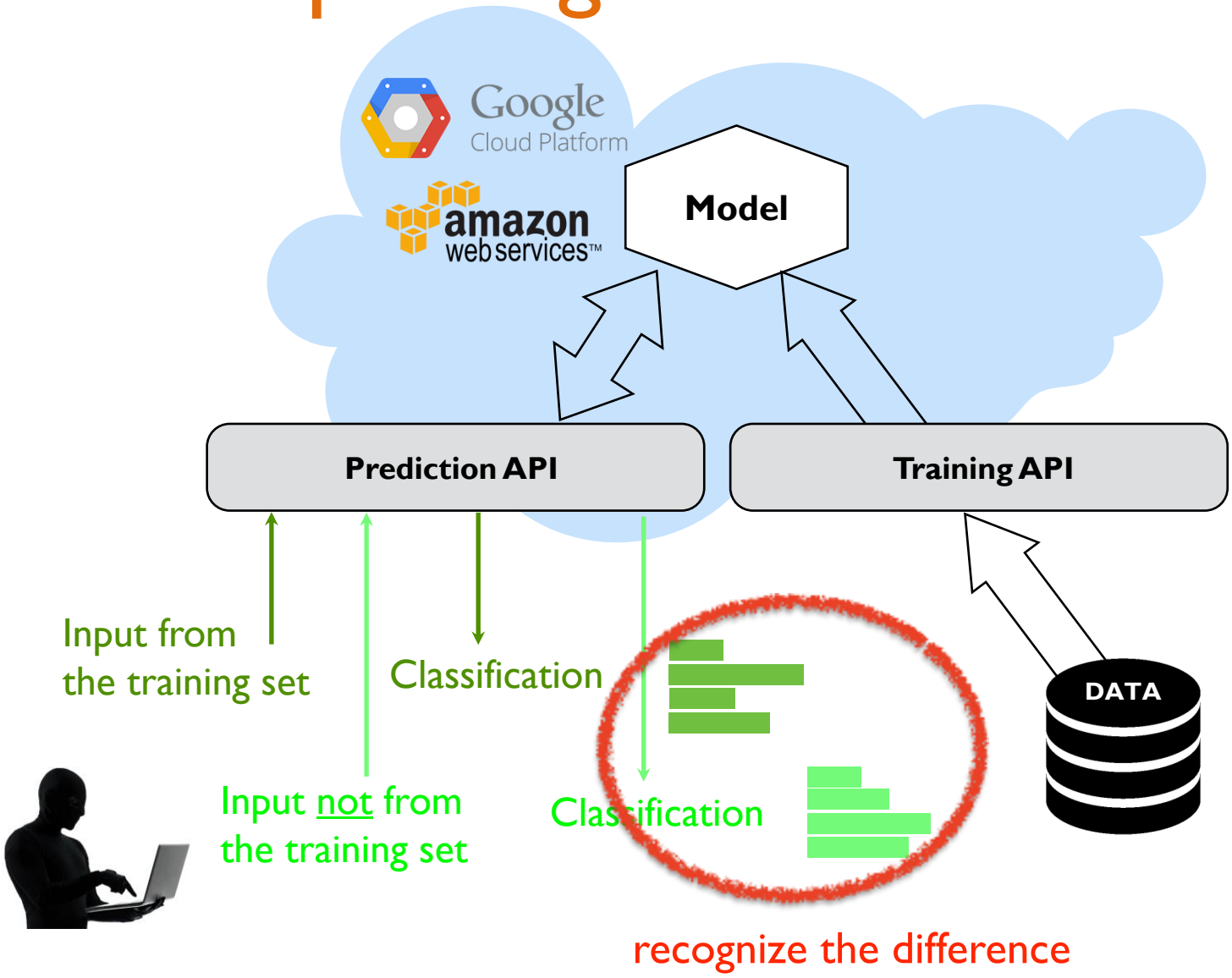
"Statistical inference considered harmful"

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

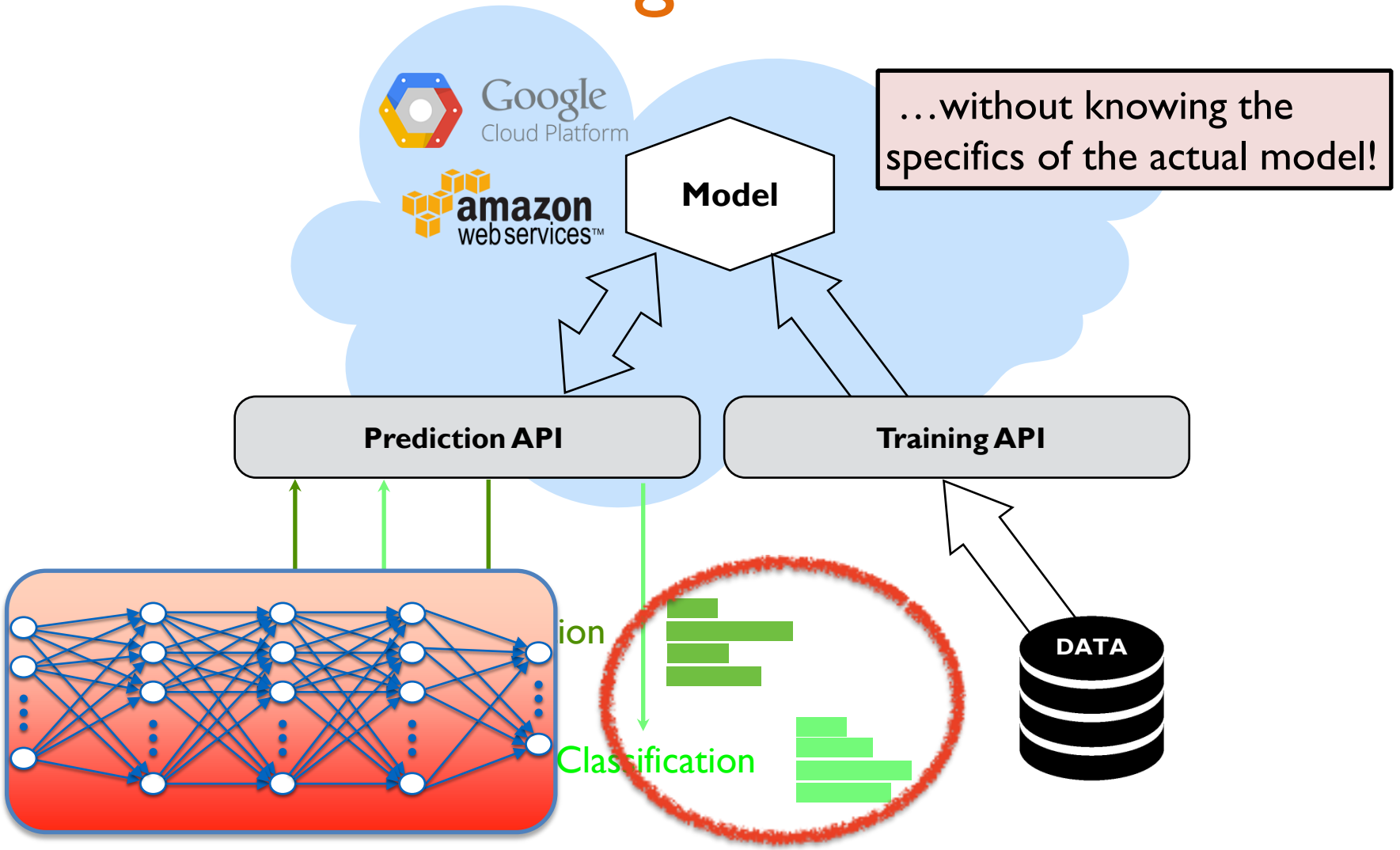
Machine Learning as a Service



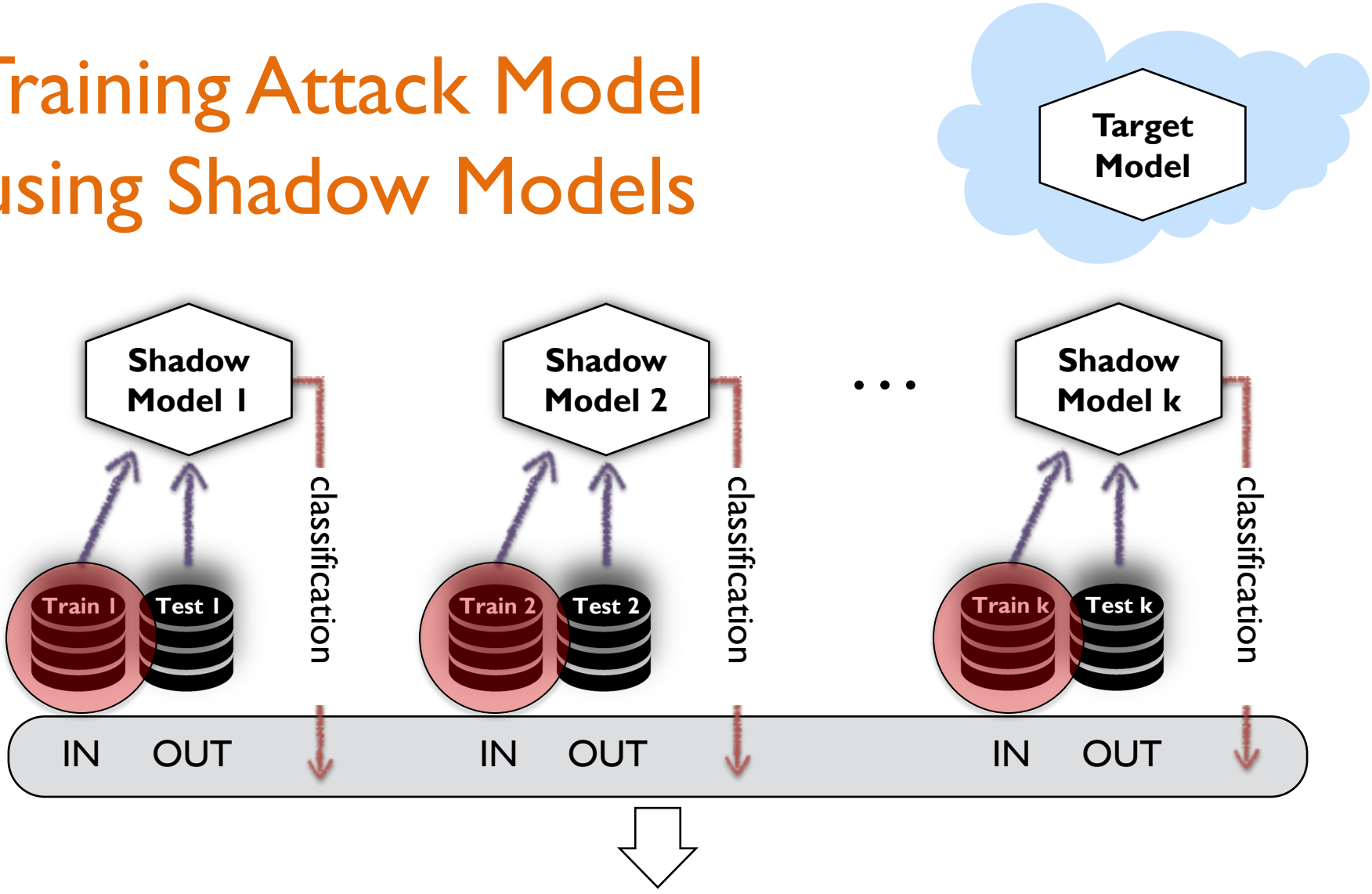
Exploiting Trained Models



ML Against ML



Training Attack Model using Shadow Models

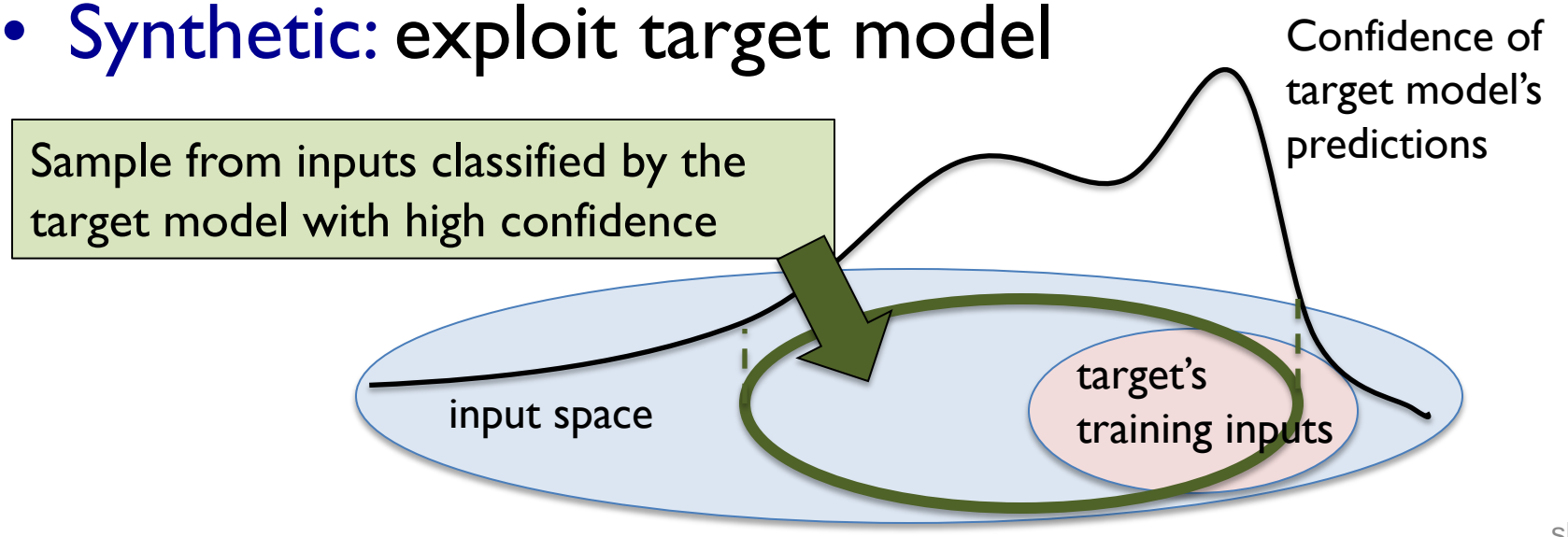


Train the attack model

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



Synthesizing Shadow Training Data

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

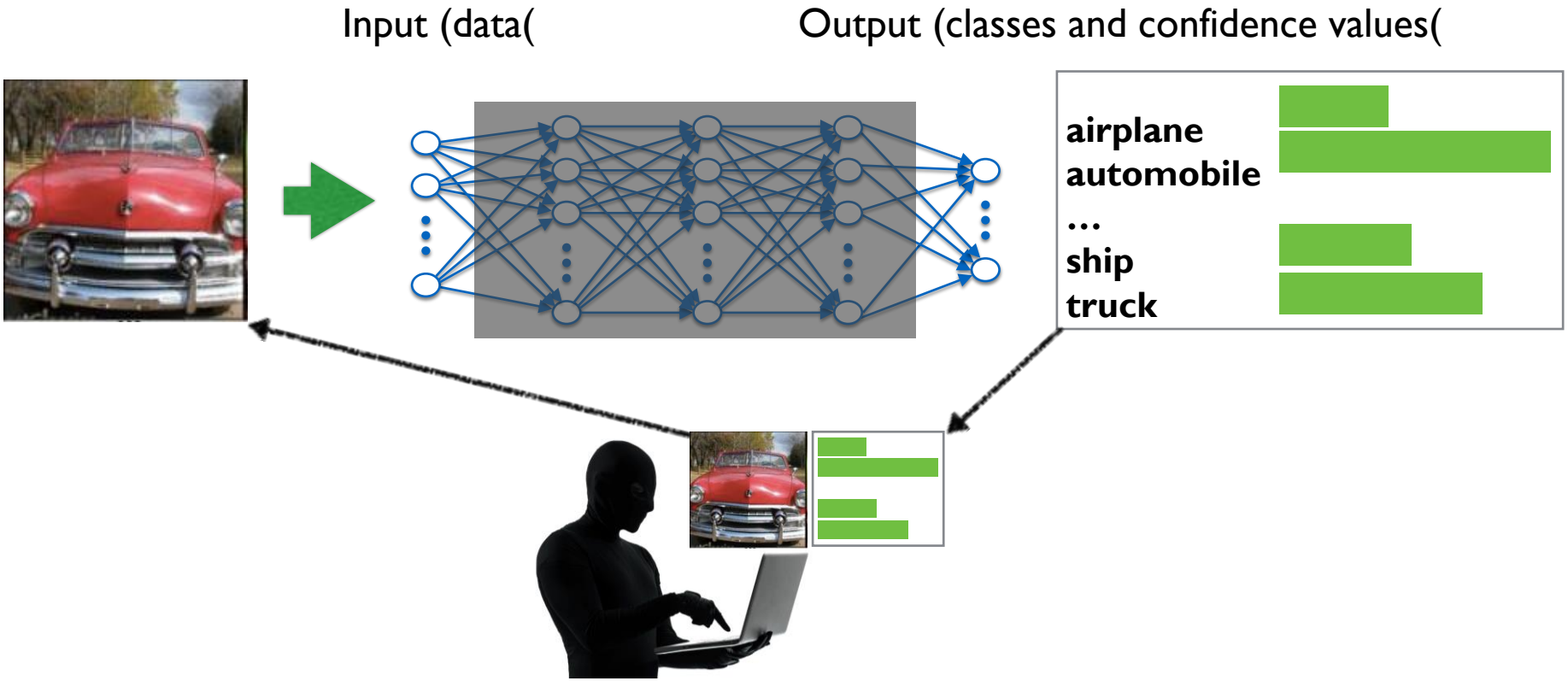
Sample from these inputs to synthesize the training dataset for shadow models

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

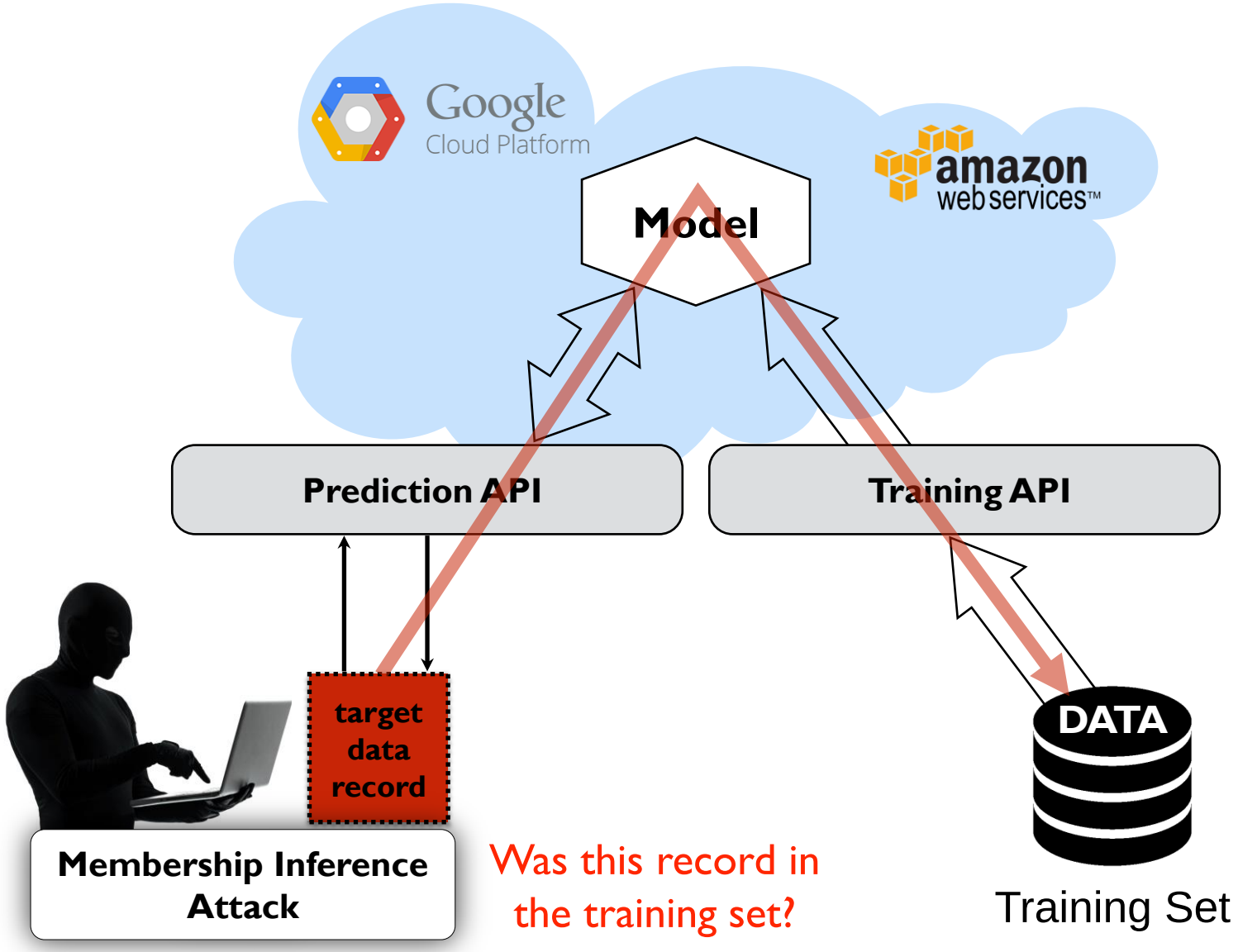
Algorithm 1 Data synthesis using the target model

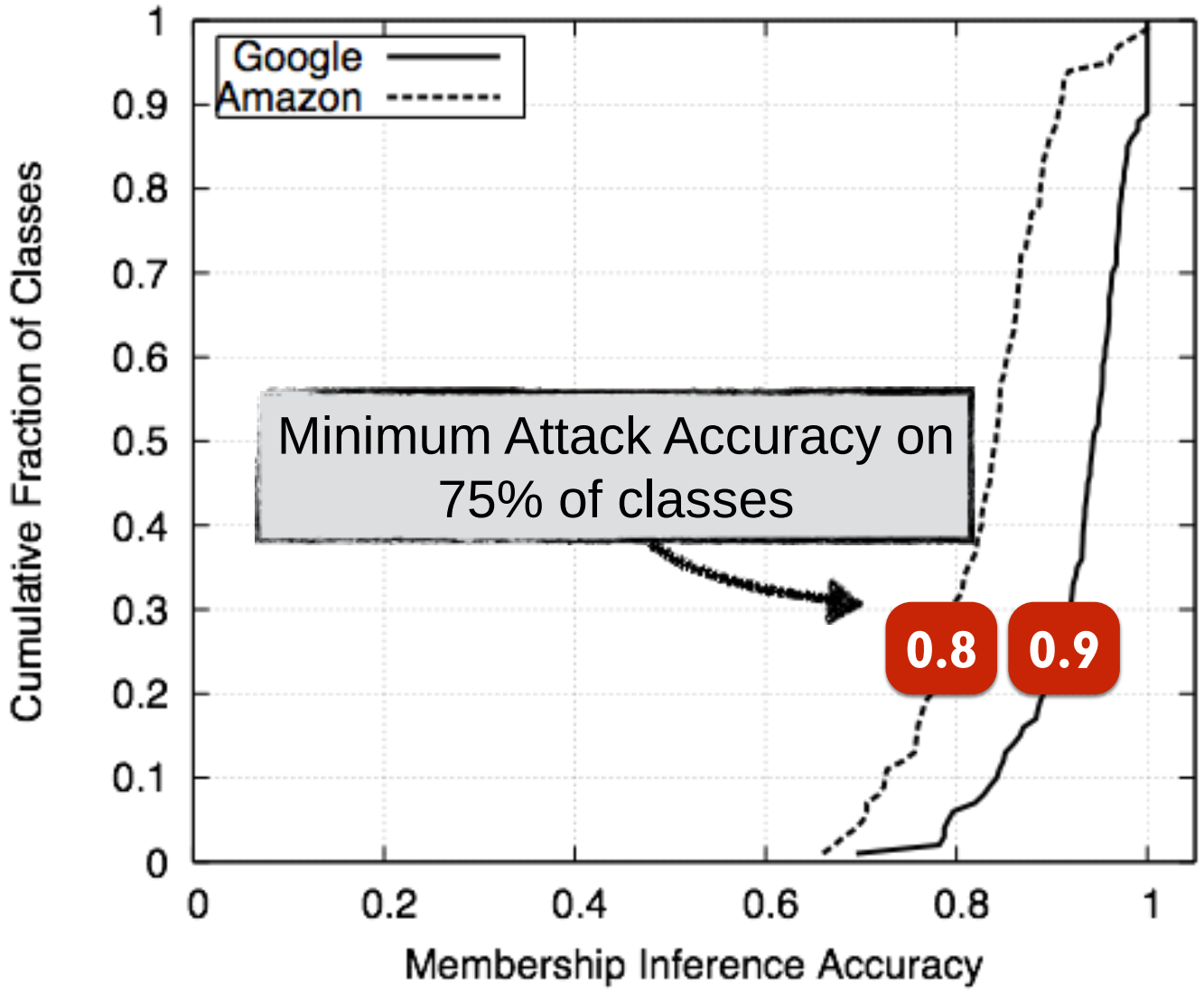
```
1:  procedure SYNTHESIZE(class : c)
2:      x ← RANDRECORD( )                    ▷ initilize a record randomly
3:      y*_c ← 0
4:      j ← 0
5:      k ← k_max
6:      for iteration = 1 ··· iter_max do
7:          y ← f_target(x)                  ▷ query the target model
8:          if y_c ≥ y*_c 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
                    k ← max(k_min, ⌈k/2⌉)
                    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



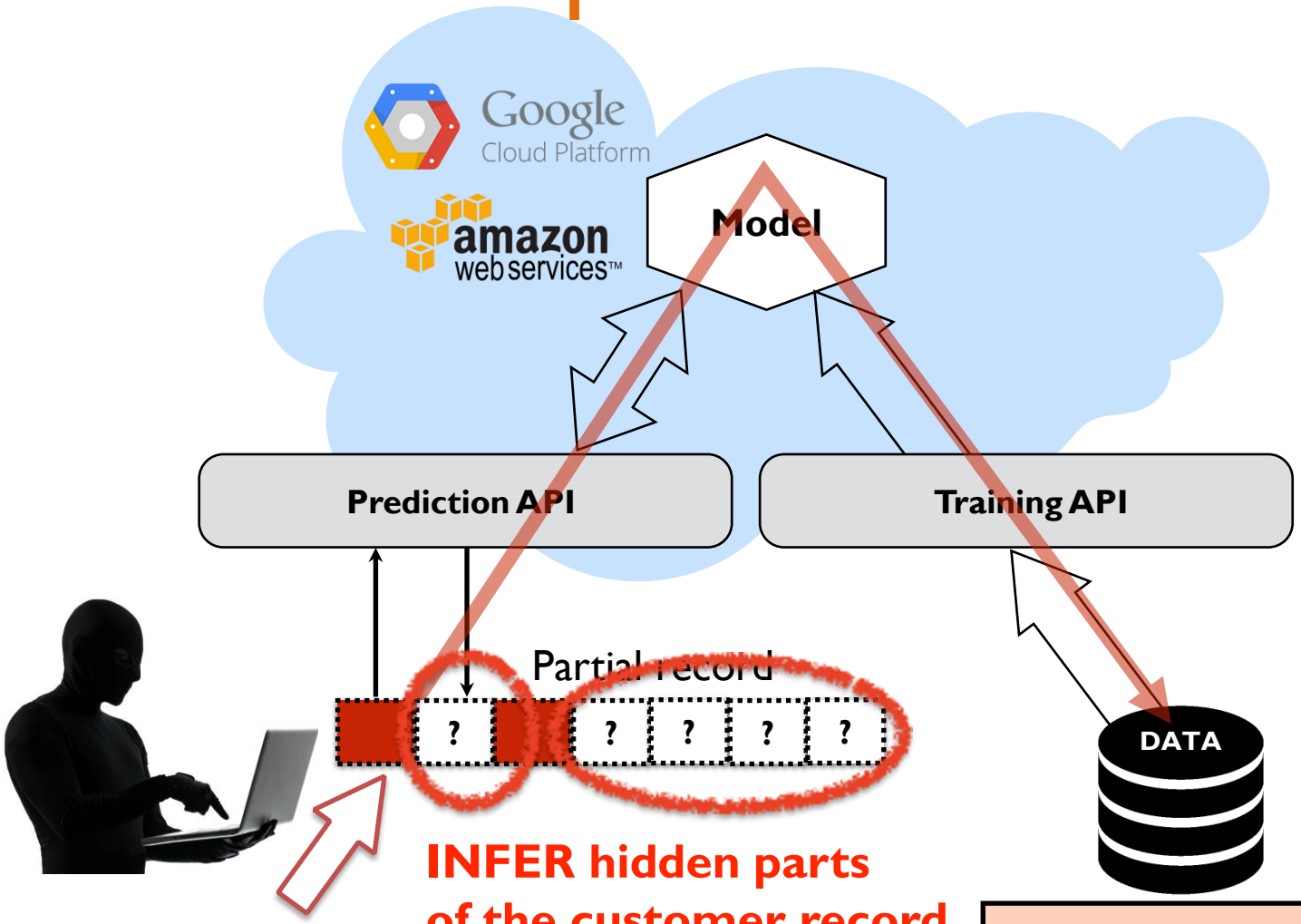
Was this image part of the training set?


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



Purchase Dataset — Classify Customers

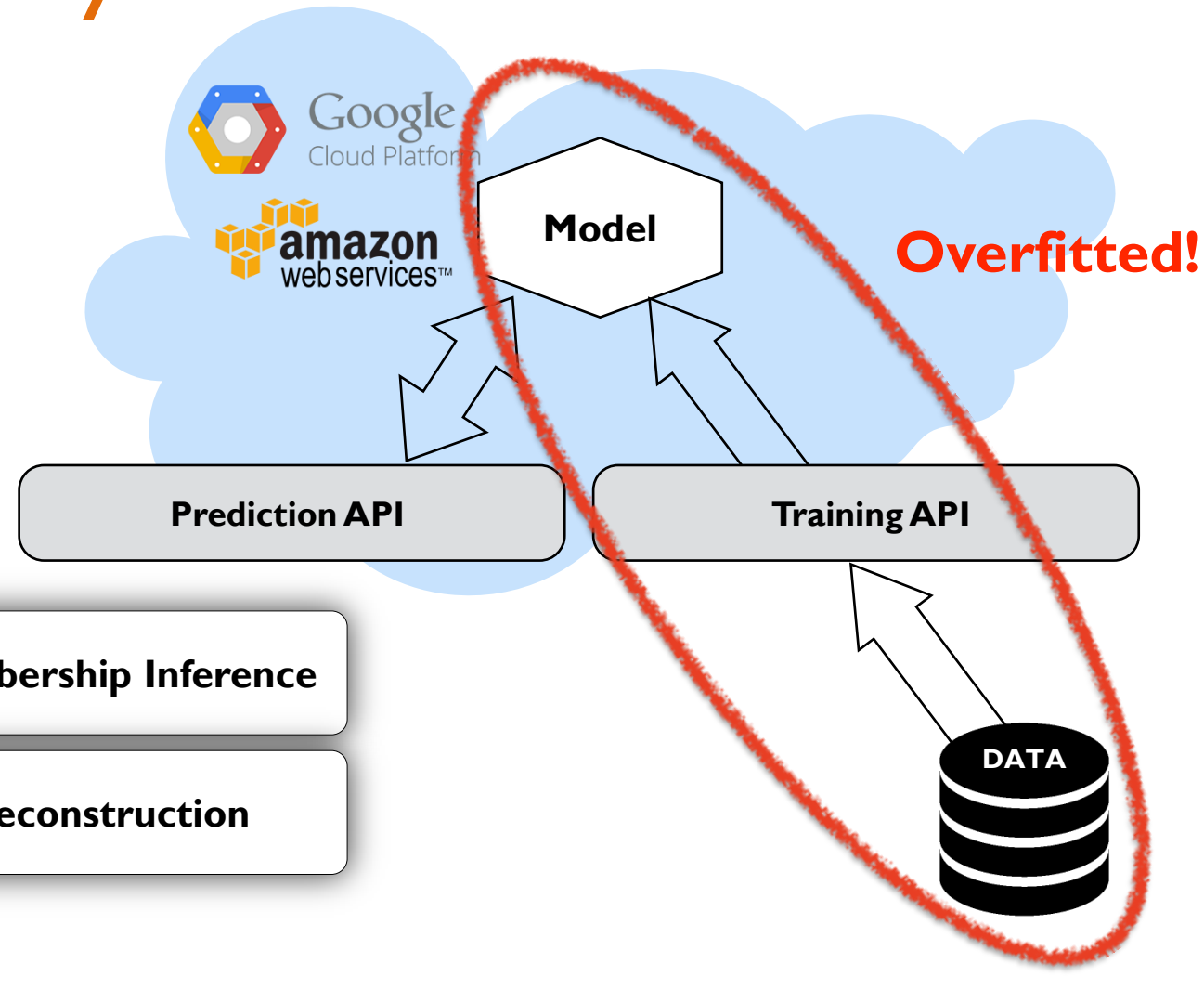
Next Step: Reconstruction



Auxiliary information, public databases , accidentally revealed data

Example: store purchases or mobile phone locations

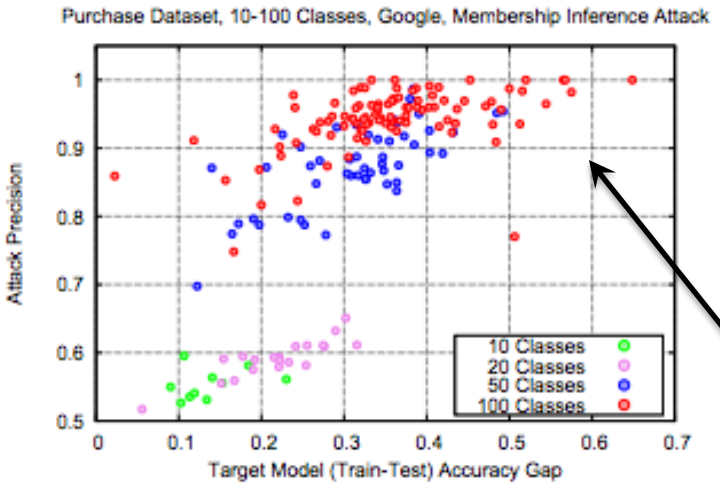
Why Do These Attacks Work?

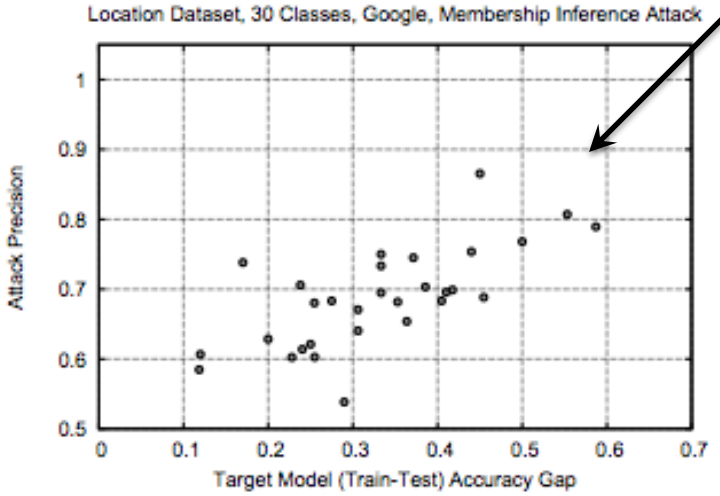


Membership Inference

Reconstruction

Attack Success vs. Test-Train Gap

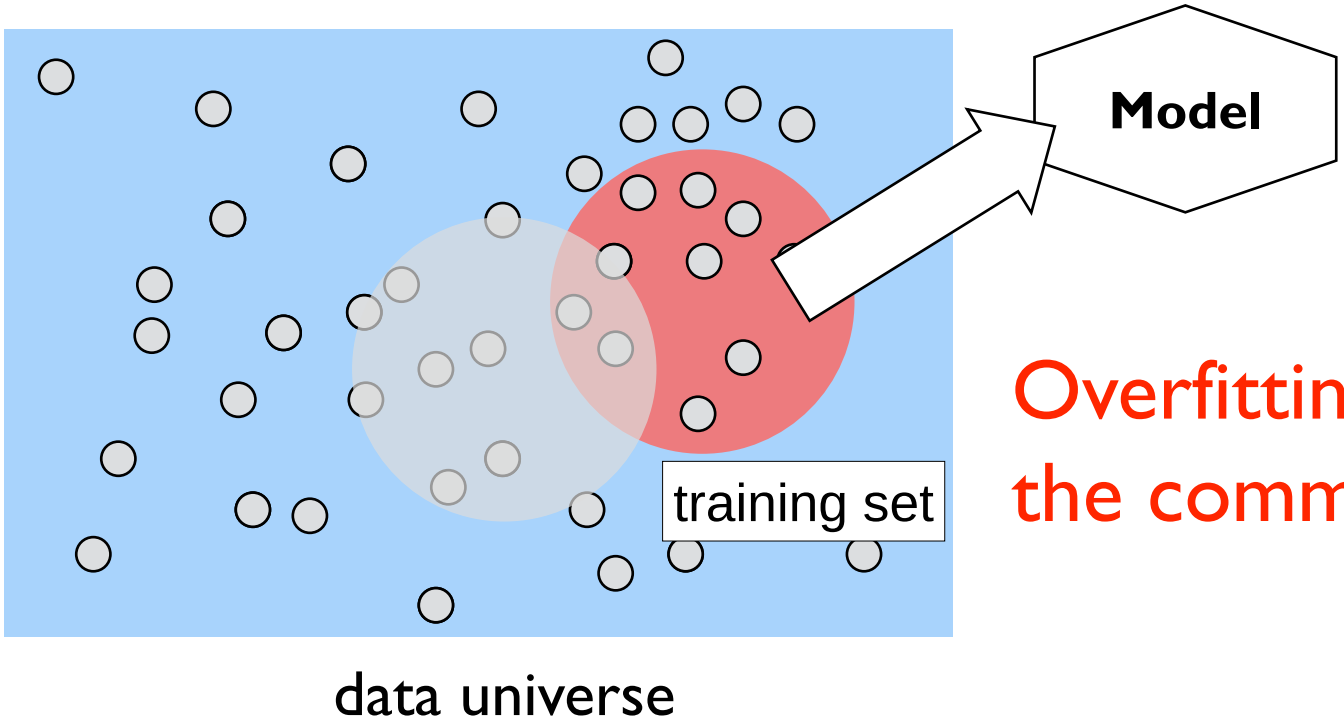




More overfitted

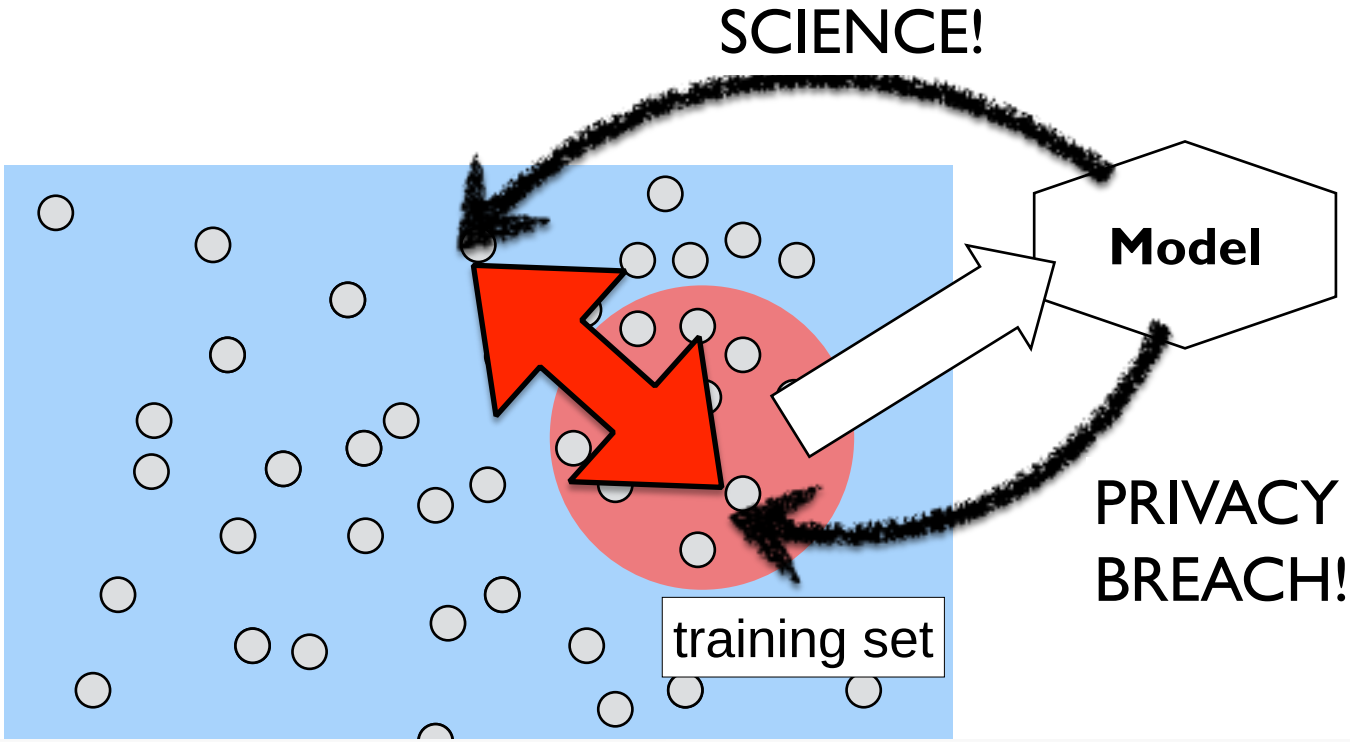
Privacy :
Does the model leak information about data in the training set?

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

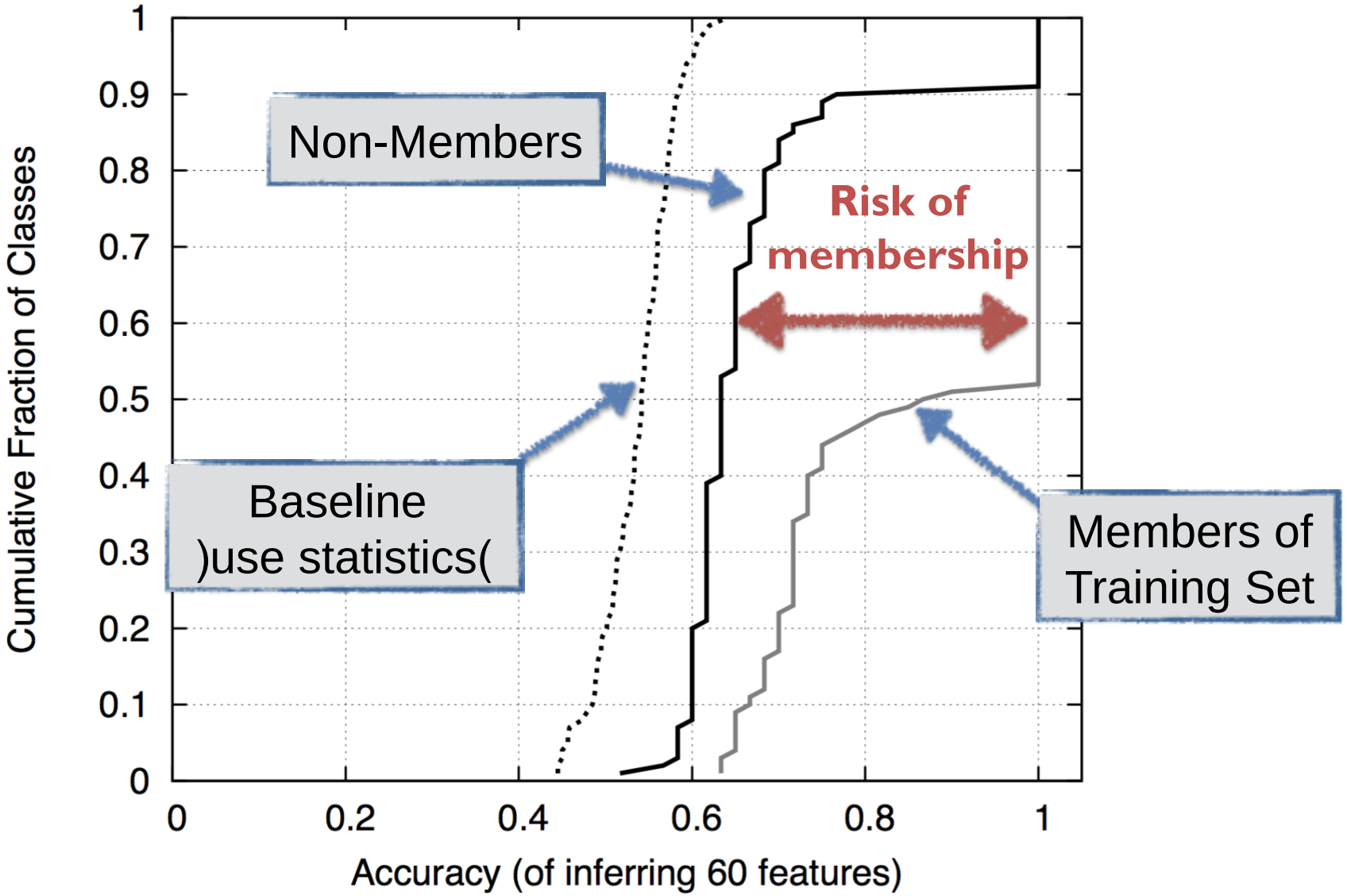


Overfitting is the common enemy

Does Inference Breach "Privacy?"



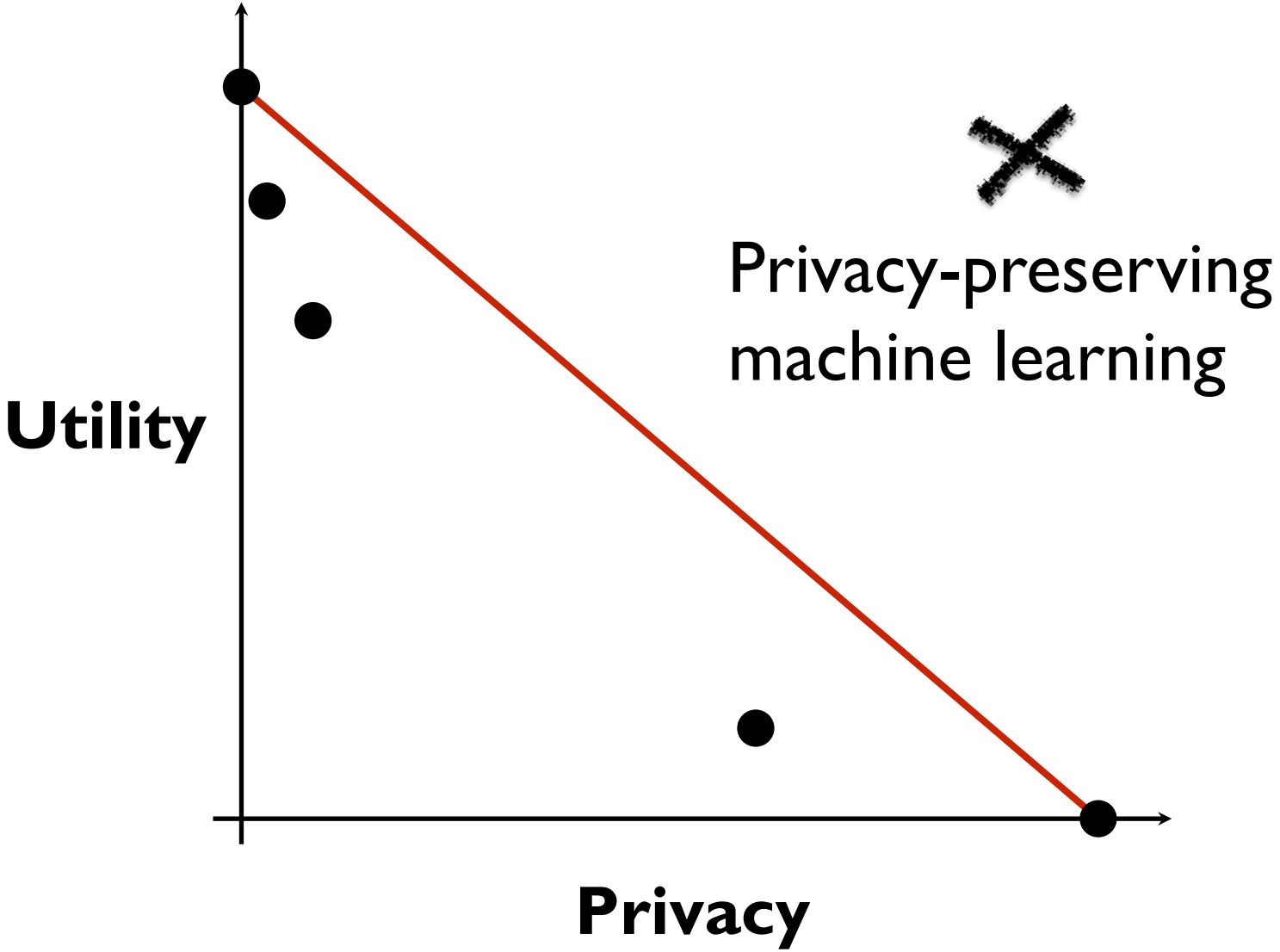
<u>Privacy breach = risk of membership:</u>
Gap between what can be inferred from the model about a member of the training set and an arbitrary input from the population



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