Security and Privacy of ML

Model & Data Confidentiality

Shang-Tse Chen

Department of Computer Science

& Information Engineering

National Taiwan University

Today's Topics

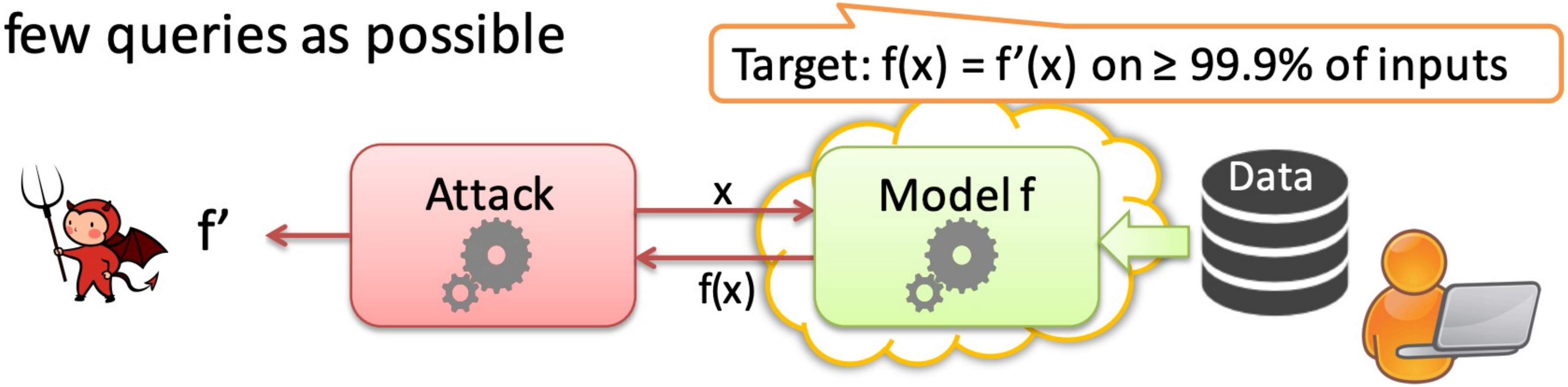
- Model Privacy
- Data Privacy

Machine Learning as a Service (MLaaS)

User uploads training data, and then gets access to a **black-box** prediction model. ($$ per query)

Model Extraction Attack [Tramèr et al. '16]

Goal: Adversarial client learns close approximation of f using as few queries as possible



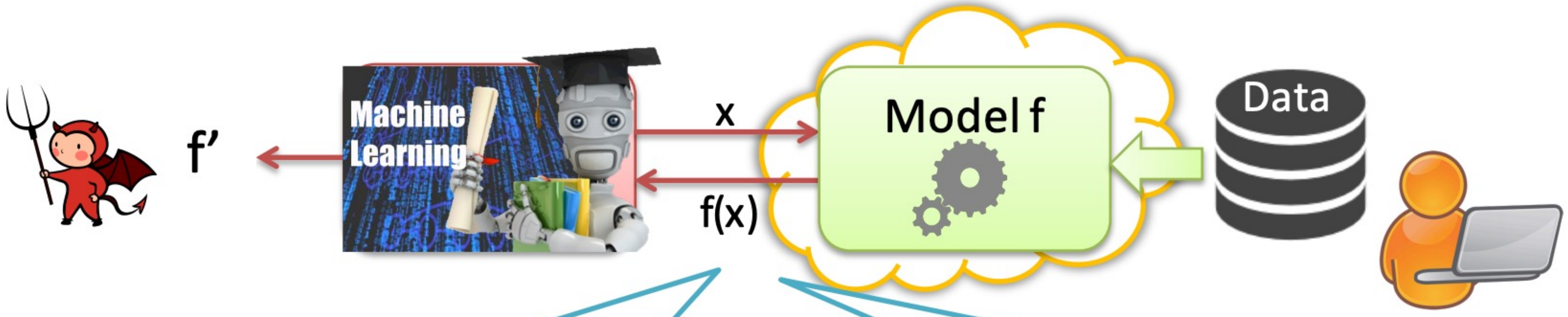
Applications:

1) Undermine pay-for-prediction pricing model
2) Facilitate privacy attacks
3) Stepping stone to model-evasion

[Stealing Machine Learning Models via Prediction APIs. Tramèr et al. Usenix Security Symposium 2016]

Model Extraction Attack

Goal: Adversarial client learns close approximation of f using as few queries as possible

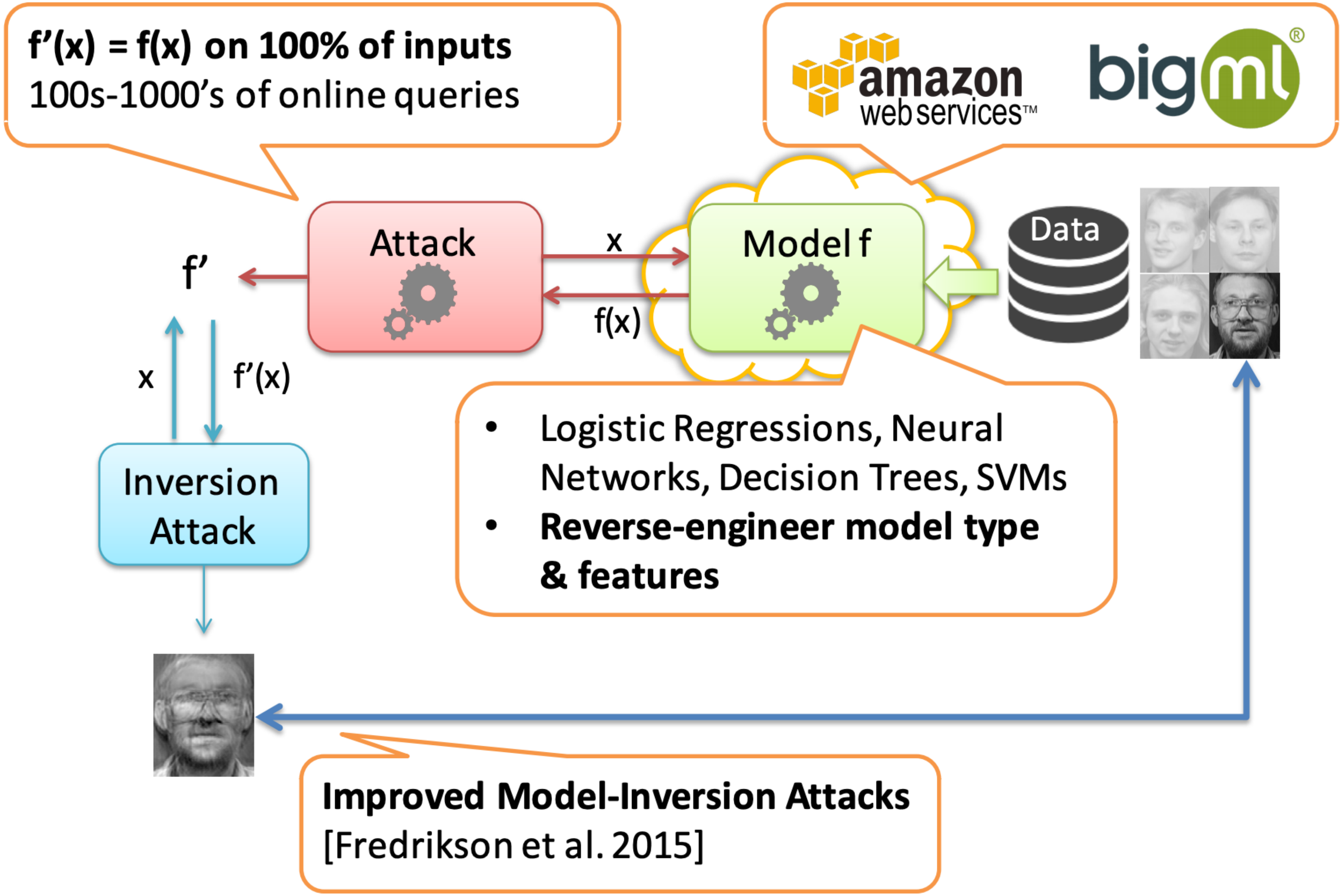


No! Prediction APIs return more information than assumed in prior work and “traditional” ML

Isn’t this “just Machine Learning”?

If f(x) is just a class label: learning with membership queries
- Boolean decision trees [Kushilevitz, Mansour – 1993]
- Linear models (e.g., binary regression) [Lowd, Meek – 2005]

Main Results



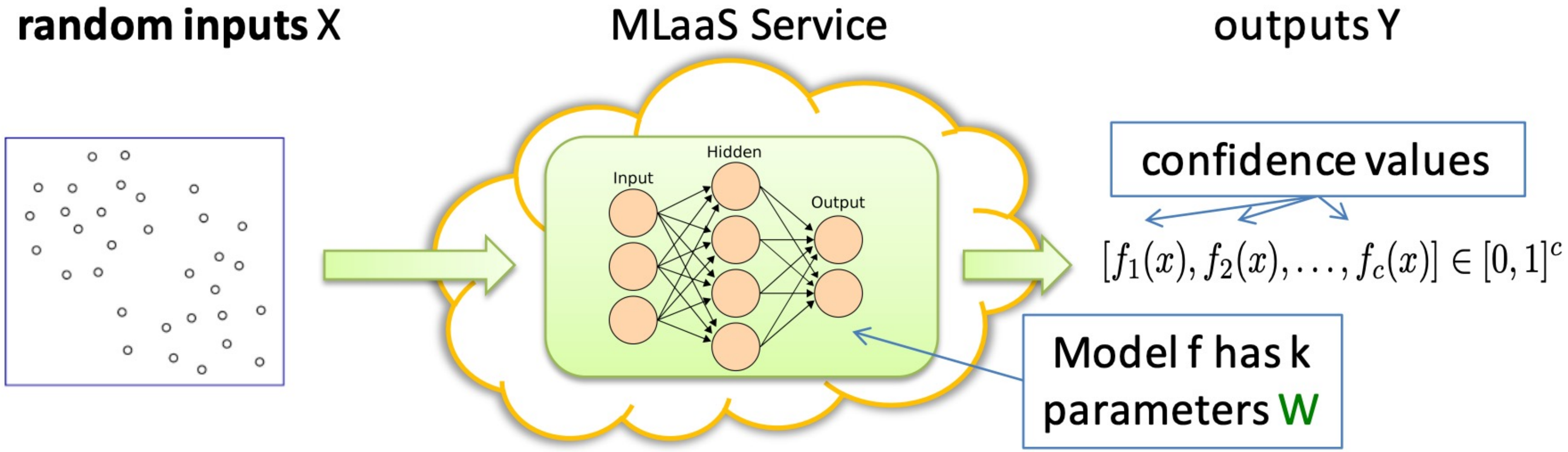
Example: Logistic Regression

$$f(x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \quad \Rightarrow \quad \ln\left(\frac{f(x)}{1 - f(x)}\right) = w \cdot x + b$$

linear equation with $d + 1$ unknown variables

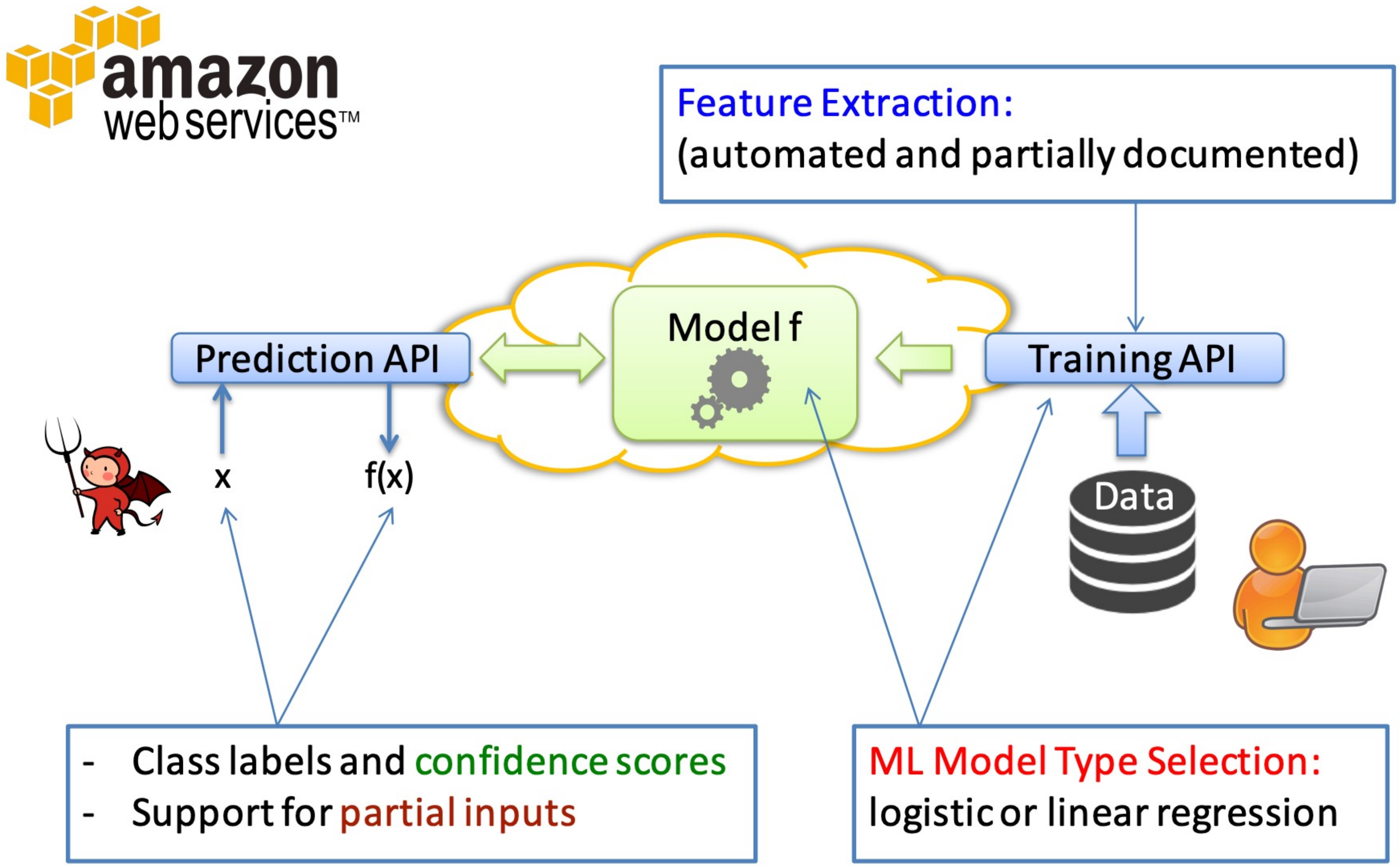
Model extraction algorithm: query $d + 1$ points and solve a linear system of $d + 1$ equations

Generic Equation-Solving Attack

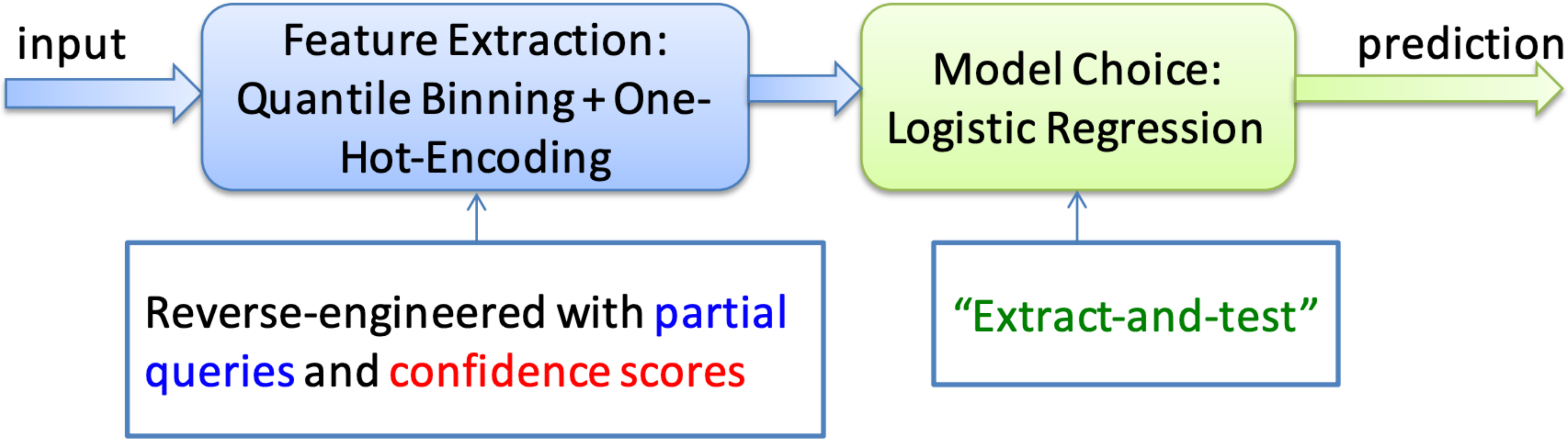


- Solve non-linear equations for weights W
  - Optimization + gradient descent
  - >99% agreement between f and f’
  - ~1 query per unknown weight

Case Study on AWS



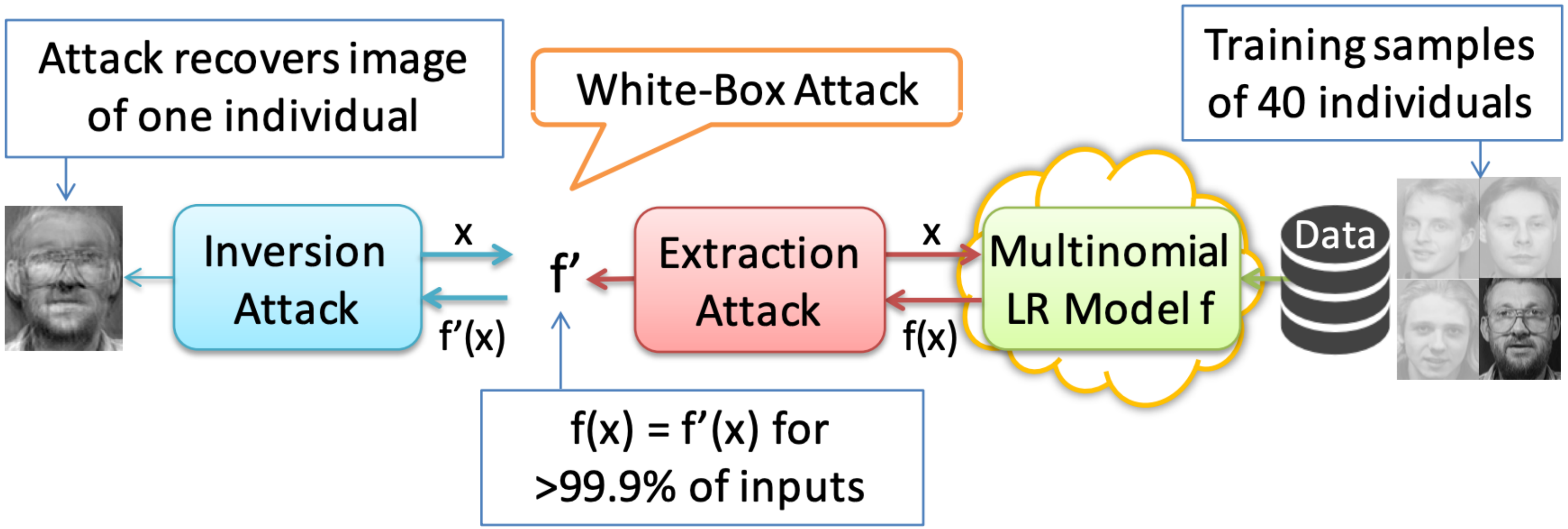
Case Study on AWS



Model	Online Queries	Time (s)	Price ($)
Handwritten Digits	650	70	0.07
Adult Census	1,485	149	0.15

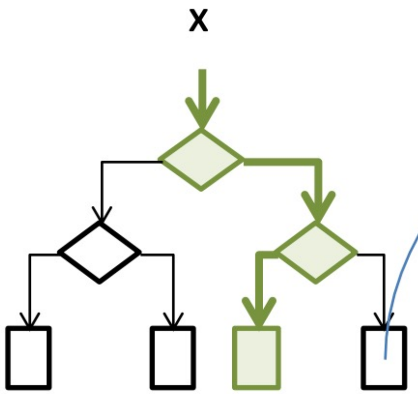
Application: Model Inversion Attack

Infer training data from trained models [Fredrikson et al. – 2015]



Strategy	Attack against 1 individual			Attack against all 40 individuals	
	Online Queries	Attack Time		Online Queries	Attack Time
Black-Box Inversion [Fredrikson et al.]	20,600	24 min	×40	800,000	16 hours
Extract-and-Invert (our work)	41,000	10 hours	×1	41,000	10 hours

Extracting a Decision Tree



Confidence value derived from class distribution in the training set

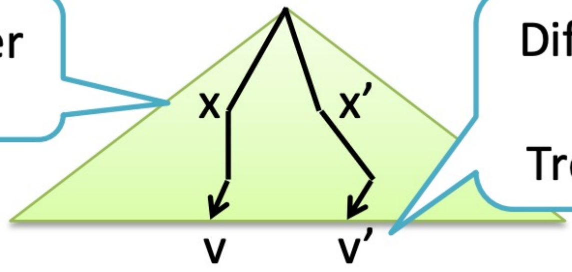
Kushilevitz-Mansour (1992)

- Poly-time algorithm with *membership queries* only
- Only for Boolean trees, impractical complexity

(Ab)using Confidence Values

- <u>Assumption:</u> all tree leaves have unique confidence values
- Reconstruct tree decisions with "differential testing"
- Online attacks on BigML

Inputs x and x' differ in a single feature



Different leaves are reached
⇔
Tree "splits" on this feature

Countermeasures

API Minimization

$f(x) = y$ ← Prediction

~~Confidence~~

- Prediction = class label only
- *Learning with Membership Queries*

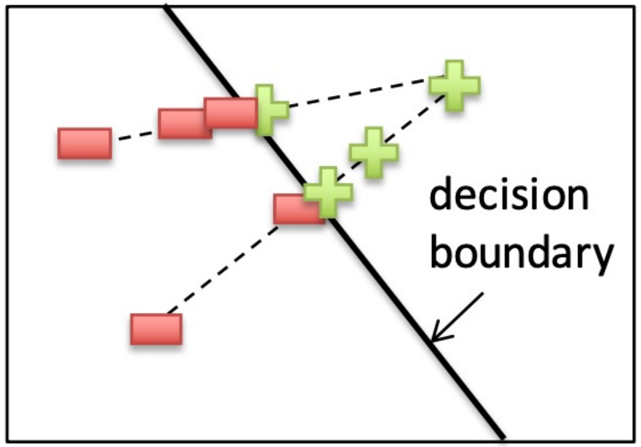
Attack on Linear Classifiers [Lowd,Meek – 2005]

classify as "+" if w*x + b > 0 and "-" otherwise

n+1 parameters w,b

f(x) = sign(w*x + b)

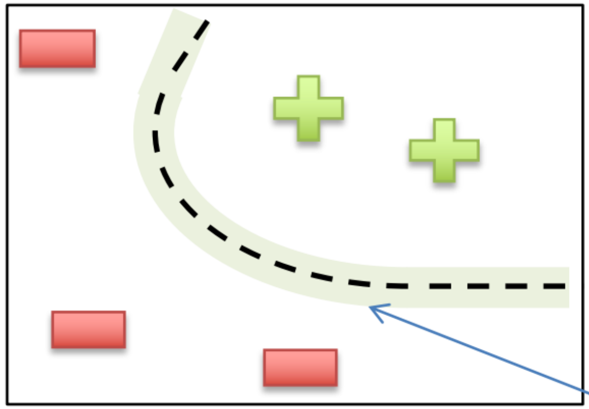
1. Find points on decision boundary (w*x + b = 0)
   - Find a "+" and a "-"
   - Line search between the two points
2. Reconstruct w and b (up to scaling factor)



Generic Model Retraining Attack

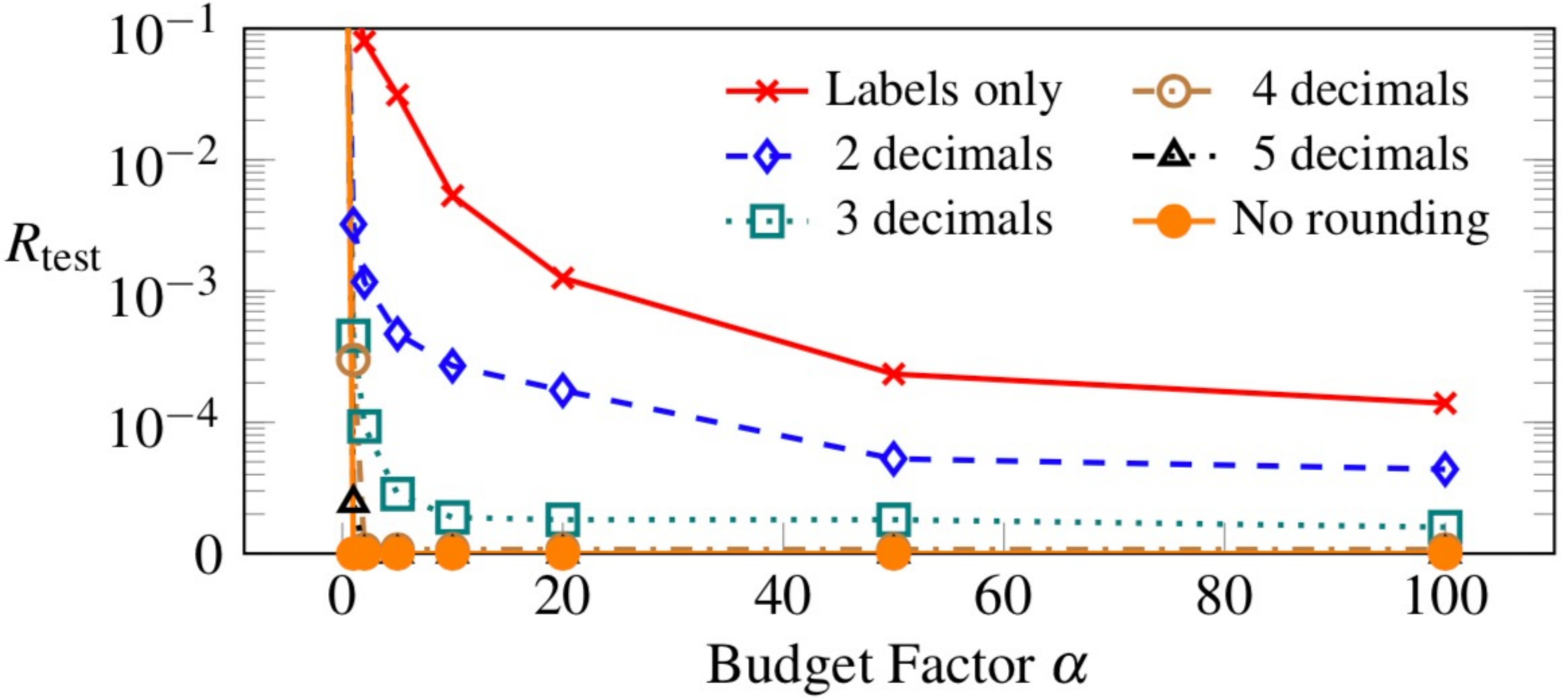
- Extend the Lowd-Meek approach to non-linear models
- Active Learning:
  - Query points close to "decision boundary"
  - Update f' to fit these points
- Multinomial Regressions, Neural Networks, SVMs:
  - >99% agreement between f and f'
  - ≈ 100 queries per model parameter of f

≈ 100× less efficient than equation-solving



query more points here

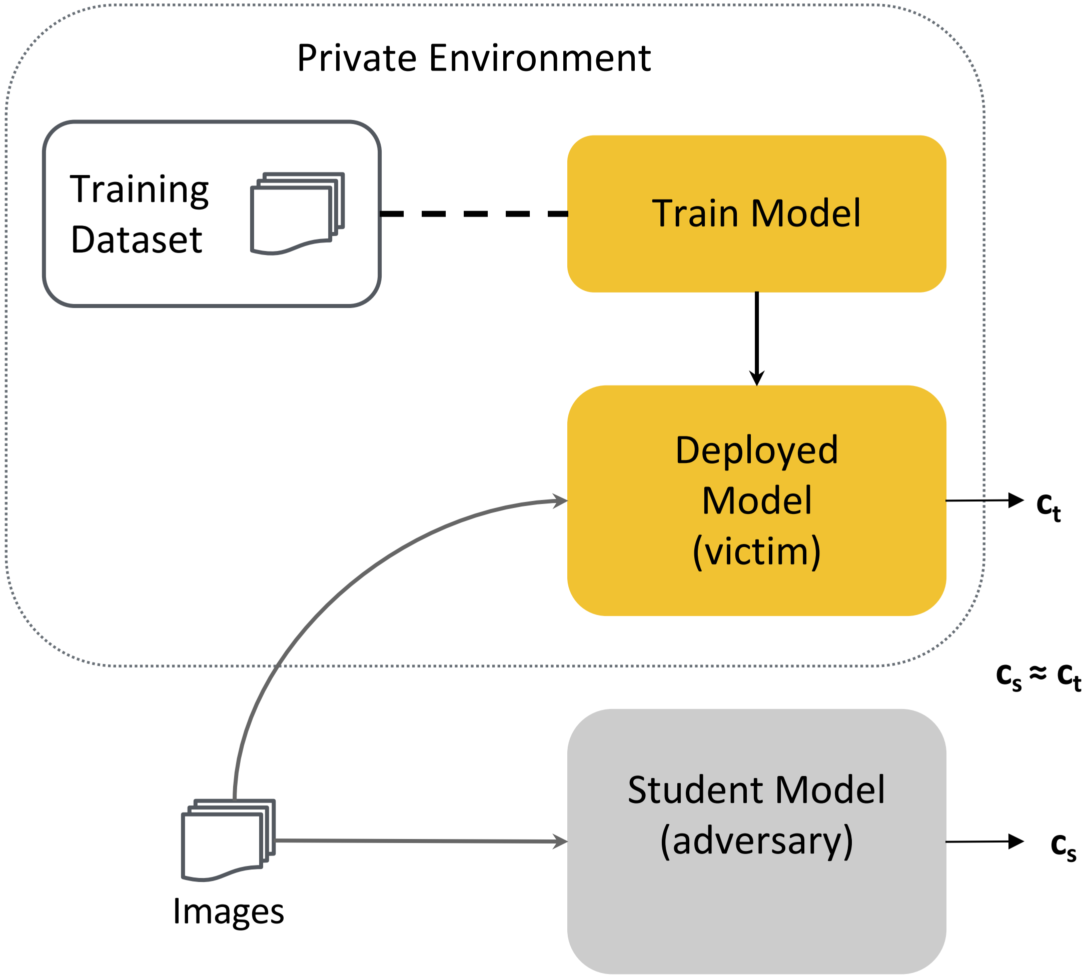
Attack performance with defenses



Data Free Model Extraction

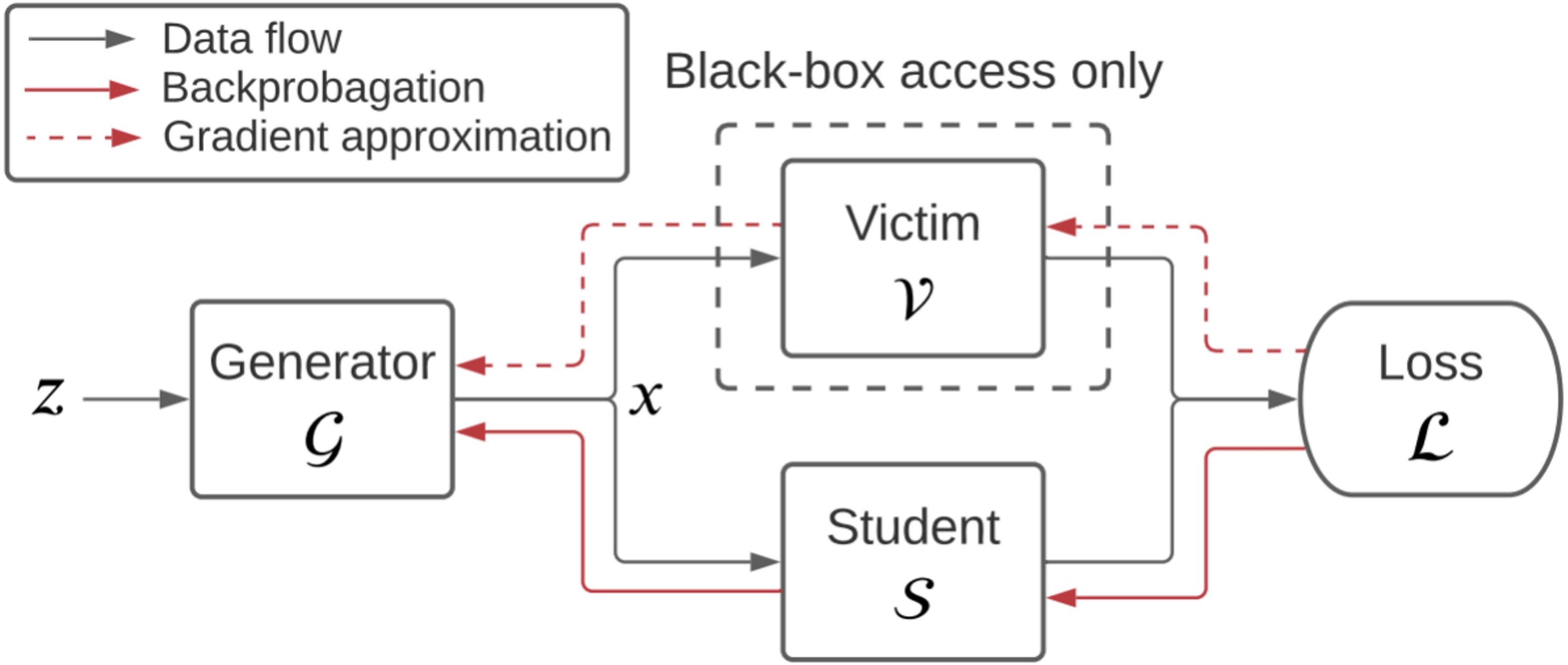
Attack performance depends on query image qualities

	Victim	CIFAR10	CIFAR100	SVHN	MNIST	SVHN_{skew}	Random
CIFAR10	95.5%	95.2%	93.5%	66.6%	37.2%	-	10.0%
SVHN	96.2%	96.0%	-	96.3%	89.5%	96.1%	84.1%



[Data-Free Model Extraction. Truong et at. CVPR 2021]

Data Free Model Extraction



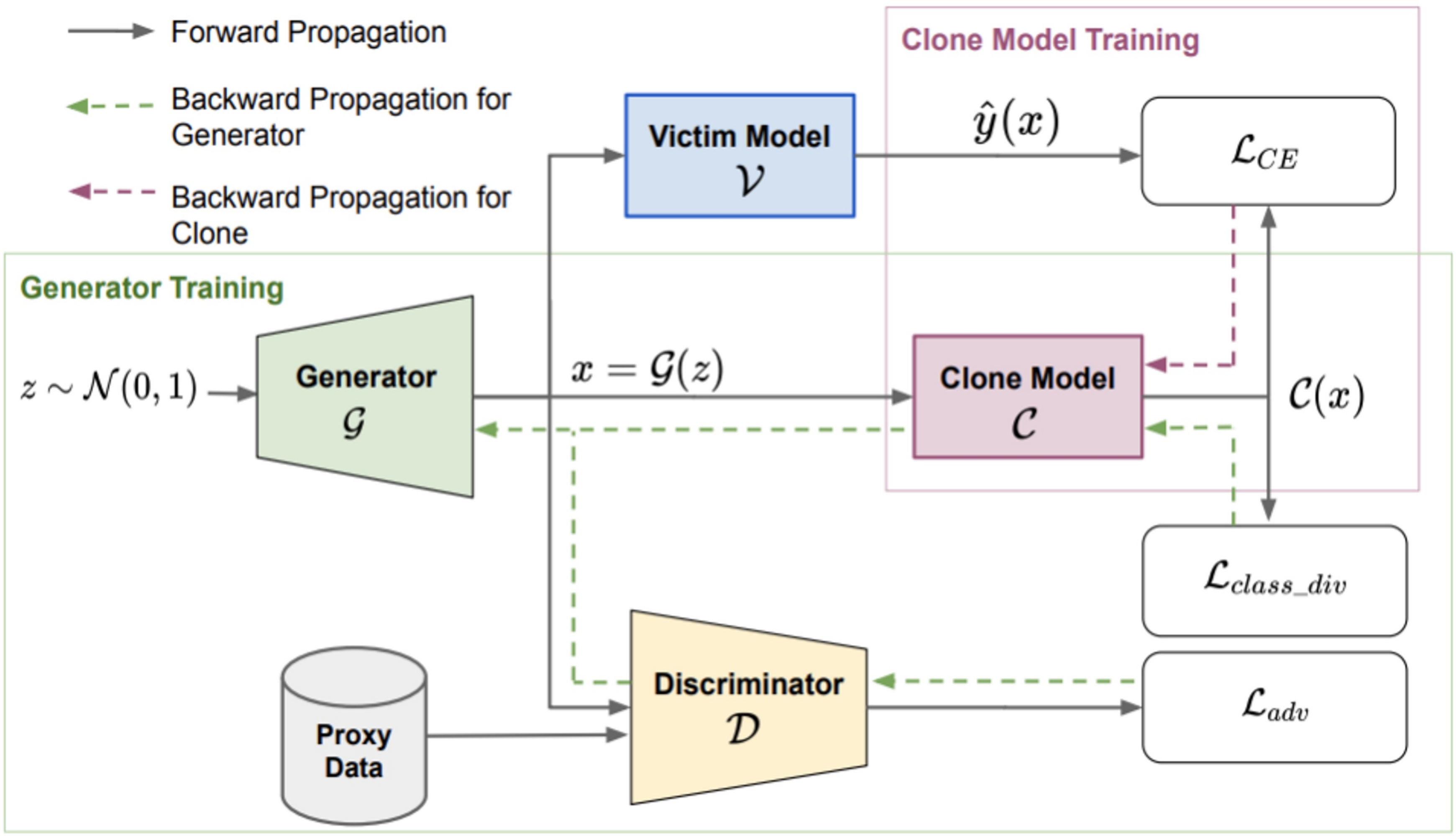
$$\min_{\mathcal{S}} \max_{\mathcal{G}} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\mathcal{L}(\mathcal{V}(\mathcal{G}(z)), \mathcal{S}(\mathcal{G}(z))) \right]$$

Data Free Model Extraction

Dataset (budget)	Victim accuracy	DFME	Random
CIFAR10 (20M)	95.5%	88.1% (0.92×)	10.0%
SVHN (2M)	96.2%	95.2% (0.99×)	84.1%

- Drawback:
  - Query budget is high (2M and 20M queries)
  - Not an issue when attacking on-device ML models

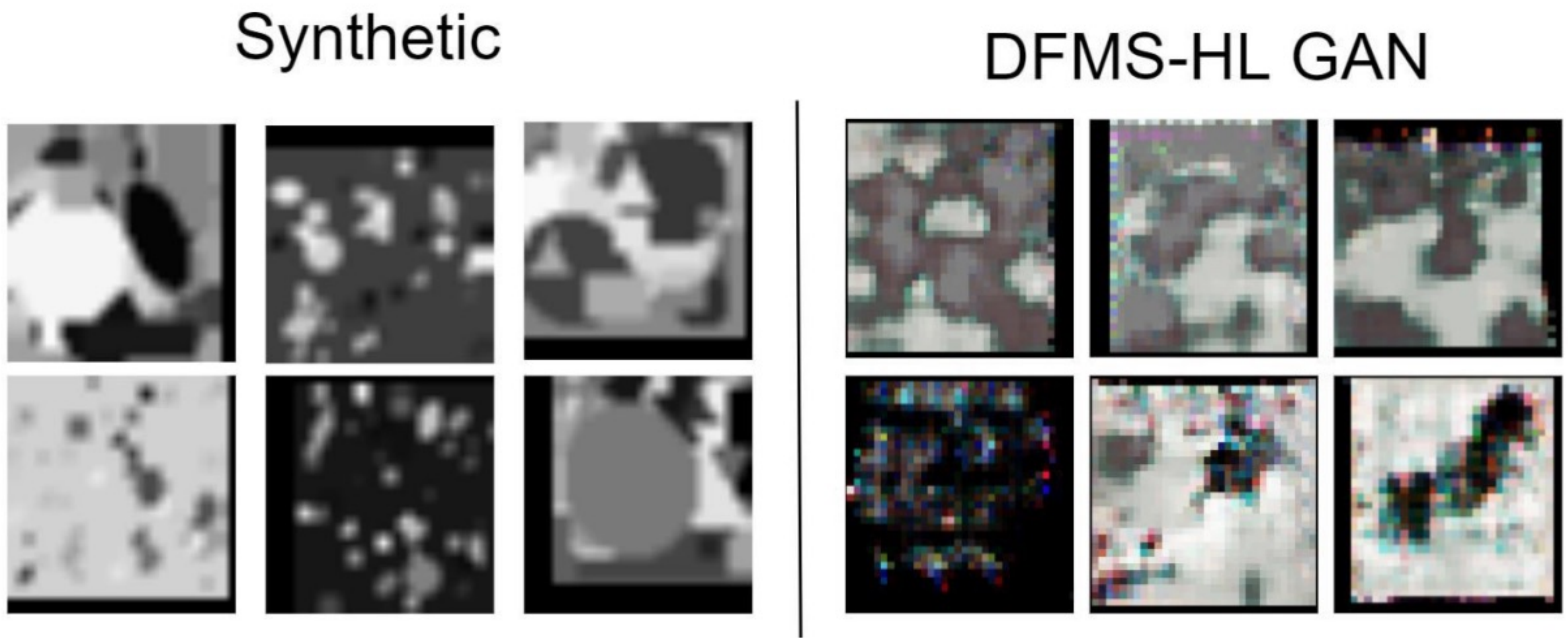
Data-Free Model Stealing with Hard Label



[Towards Data-Free Model Stealing in a Hard Label Setting. Sanyal et at. CVPR 2022]

Data-Free Model Stealing with Hard Label

- Proxy data



Data-Free Model Stealing with Hard Label

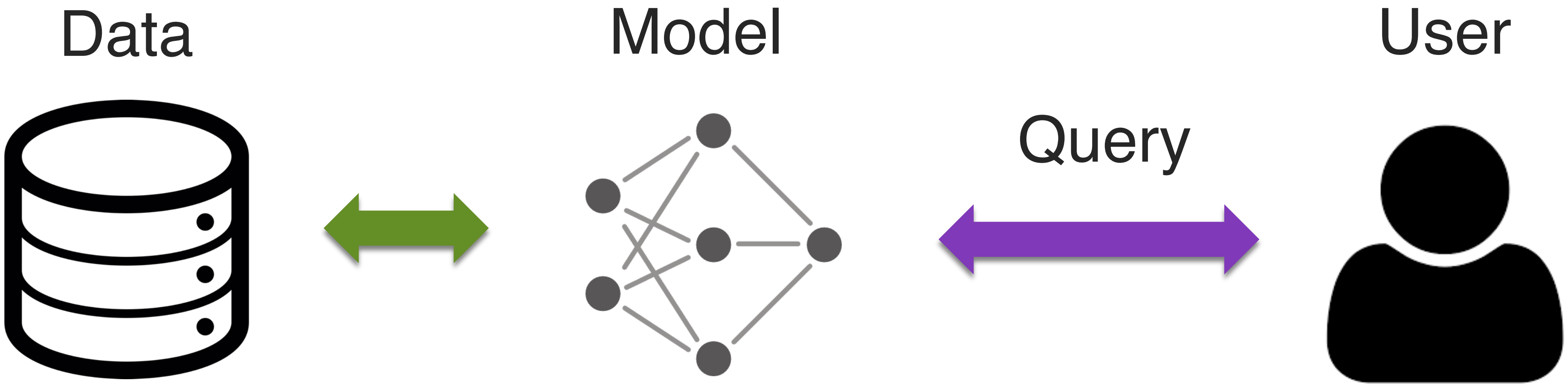
Method	Hard Label	Black-Box	Data-Free	Victim Accuracy	Synthetic/ Data-Free	CIFAR-100 (40C)	CIFAR-100 (10C)
			Victim Accuracy $\sim$ 95.5%, Victim Model: ResNet-34				
MAZE [17]	×	✓	✓	95.50	45.60	-	-
DFME [34]	×	✓	✓	95.50	88.10	-	-
DFMS-HL (Ours)	✓	✓	✓	95.59	84.51	**92.06**	**85.53**
DFMS-SL (Ours)	×	✓	✓	95.59	**91.24**	**93.96**	**90.88**

Let’s Move On to Data Privacy

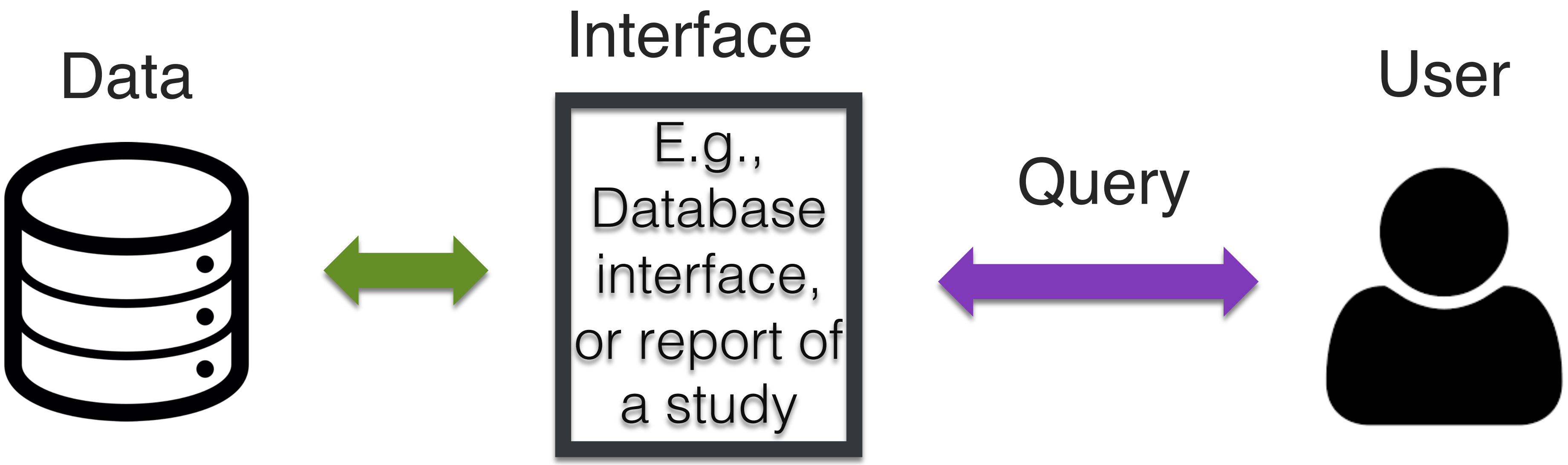
Data Privacy

- Common approach: anonymize sensitive data
- Many ways to de-anonymize
- Unprotected ML model may leak training data information

Generic Framework



Generic Framework



How do we provide useful information to user, while preserving privacy of individuals in the data?

Anonymization

國立臺灣大學
計算機及資訊網路中心
電子報
第0040期·2017.03.20 發行
ISSN 2077-8813

歷史回顧 訂閱/取消 校務服務 專題報導 技術論壇 推薦刊物

首頁 > 技術論壇

淺談個資遮罩

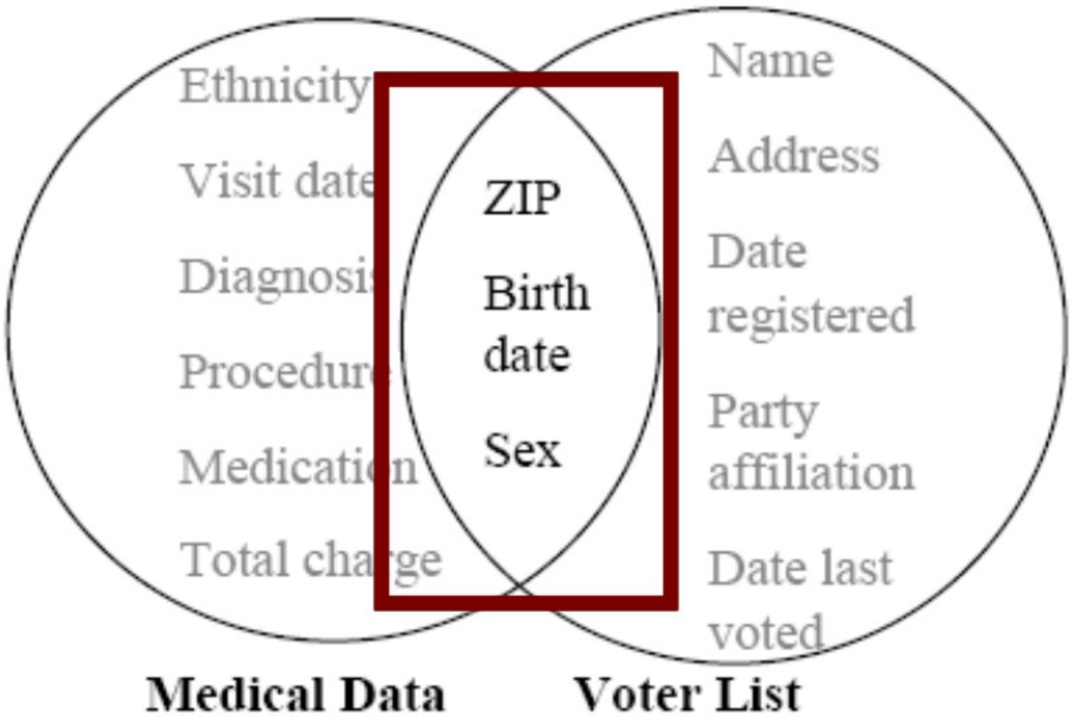
作者：許凱平 / 計算機及資訊網路中心作業管理組副組長

在個資法推行的這幾年，如何拿捏資訊揭露的尺度，一直是件不容易的事。根據條文，姓名是個資的一種，所以我們可以看到有些文件的姓名被遮罩起來，為的是避免觸法。這些改變，讓我們感到個資法的宣導有了成效，似乎民眾的隱私受到更好的保護，但另一方面，有些資訊的遮罩又會讓人覺得矯枉過正。例如在模範生的公告上學生姓名都成了陳○○或許○○，看到這樣一則公告，我想大部分的人都會懷疑這樣的公告，還有公告的必要嗎？雖然後來教育部做了解釋，獎懲公告屬於姓名的合理使用範圍，不需去識別化，但是已經成為新的積習，大家就繼續沿用，也不管是否真的需要。另外，由於大家對於相同的屬性資料（姓名、身分證號碼等），遮罩方式各有不同，有些遮頭，有些去尾；有心人士稍加拼圖就湊出原始資料的機率其實並不低。

https://www.cc.ntu.edu.tw/chinese/epaper/0040/20170320_4008.html

Linkage Attack

87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB



[Sweeney. '97]

Linkage Attack

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings (noisy)

	Titanic	Heidi
Bob	2	1

Used as auxiliary information

Weighted Scoring Algorithm

[Narayanan et al. '08]

K-anonymity

Ensure that each record is indistinguishable with other k-1 records

ID	Age	Zipcode	Diagnosis
1	28	13053	Heart Disease
2	29	13068	Heart Disease
3	21	13068	Viral Infection
4	23	13053	Viral Infection
5	50	14853	Cancer
6	55	14853	Heart Disease
7	47	14850	Viral Infection
8	49	14850	Viral Infection
9	31	13053	Cancer
10	37	13053	Cancer
11	36	13222	Cancer
12	35	13068	Cancer

k-anonymization

K=4

ID	Age	Zipcode	Diagnosis
1	[20-30]	130**	Heart Disease
2	[20-30]	130**	Heart Disease
3	[20-30]	130**	Viral Infection
4	[20-30]	130**	Viral Infection
5	[40-60]	148**	Cancer
6	[40-60]	148**	Heart Disease
7	[40-60]	148**	Viral Infection
8	[40-60]	148**	Viral Infection
9	[30-40]	13***	Cancer
10	[30-40]	13***	Cancer
11	[30-40]	13***	Cancer
12	[30-40]	13***	Cancer

K-Anonymity

- Optimal k-anonymity is an NP-hard problem
- May remove too much information

系所組代碼 **905**

系所組別:資訊工程學系碩士班

准考証號碼	姓　名	錄取別	身份別
905150072	林〇廷	正取	一般生
905150079	許〇瑋	正取	一般生
905150676	丁　〇	正取	一般生
905150659	韓〇駿	正取	一般生
905150671	余〇倫	正取	一般生
905150028	曾〇棠	正取	一般生
905170070	楊〇羽	正取	一般生
905150285	丁〇安	正取	一般生
905150480	潘〇辰	正取	一般生

k=2

系所組代碼 **905**

系所組別:資訊工程學系碩士班

准考証號碼	姓　名	錄取別	身份別
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生
9051[redacted]	[redacted]	正取	一般生

Attack to K-Anonymity

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l-Diversity

Extension of K-anonymity

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Attack to l-Diversity: Skewness Attack

- Suppose 10% of the population suffer from diabetes
- In this subset, the probability of diabetes is much higher

Race	DOB	Sex	ZIP	Disease
black	64	F	941**	diabetes
black	64	F	941**	short breath
black	64	F	941**	diabetes
black	64	F	941**	diabetes

Attack to l-Diversity: Similarity Attack

l-diversity does not consider the semantics of sensitive values!

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	⩾40	50K	Gastritis
4790*	⩾40	100K	Flu
4790*	⩾40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

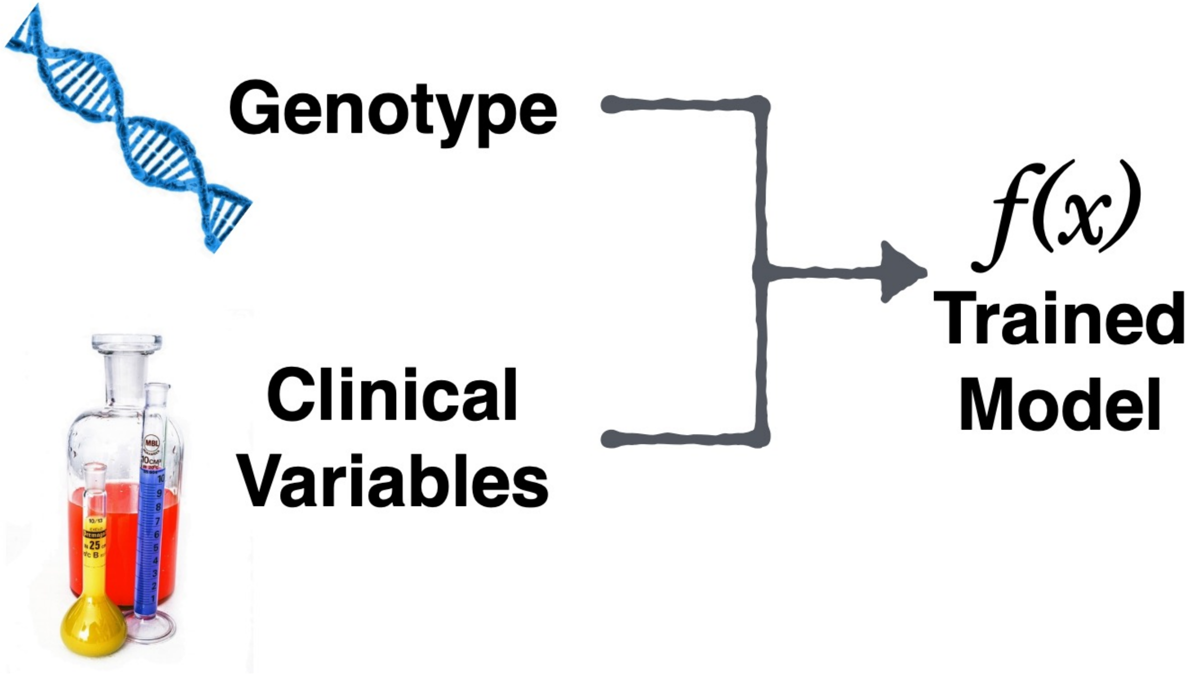
1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

Many subsequent work

- t-closeness, m-invariance, delta-presence, …
- Still an active research area

Model Inversion Attack [Fredrikson et al. ‘14]

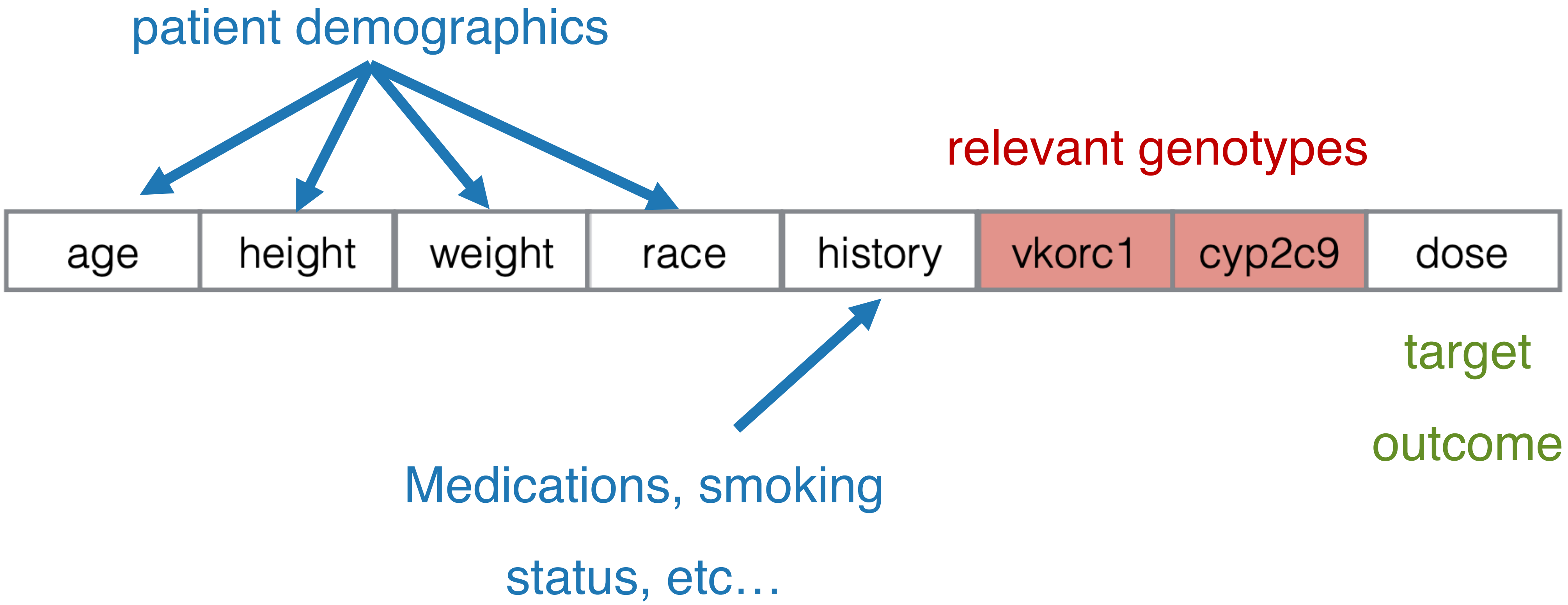
- Application in pharmacogenetics



[Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing.
Fredrikson et al. Usenix Security Symposium 2014]

Example Task: Warfarin Dosing

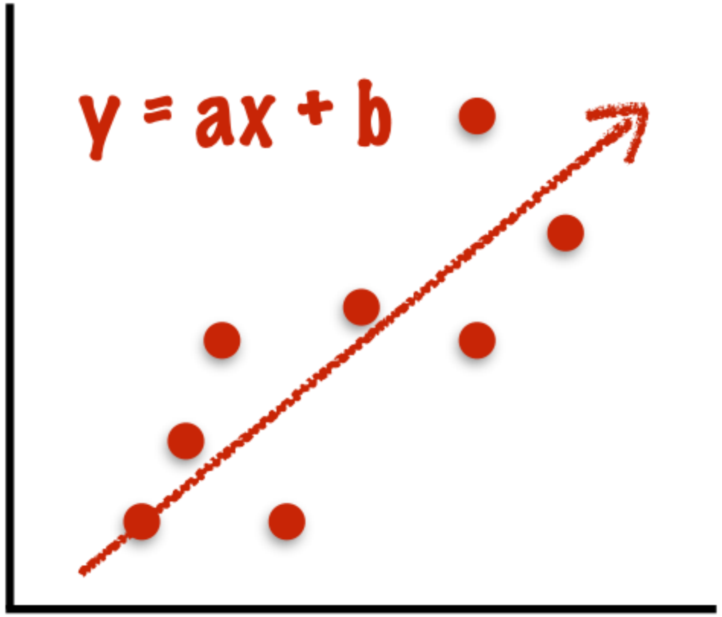
- Warfarin is the most popular anticoagulant in use today
- Warfarin is notoriously difficult to dose correctly



Example Task: Warfarin Dosing

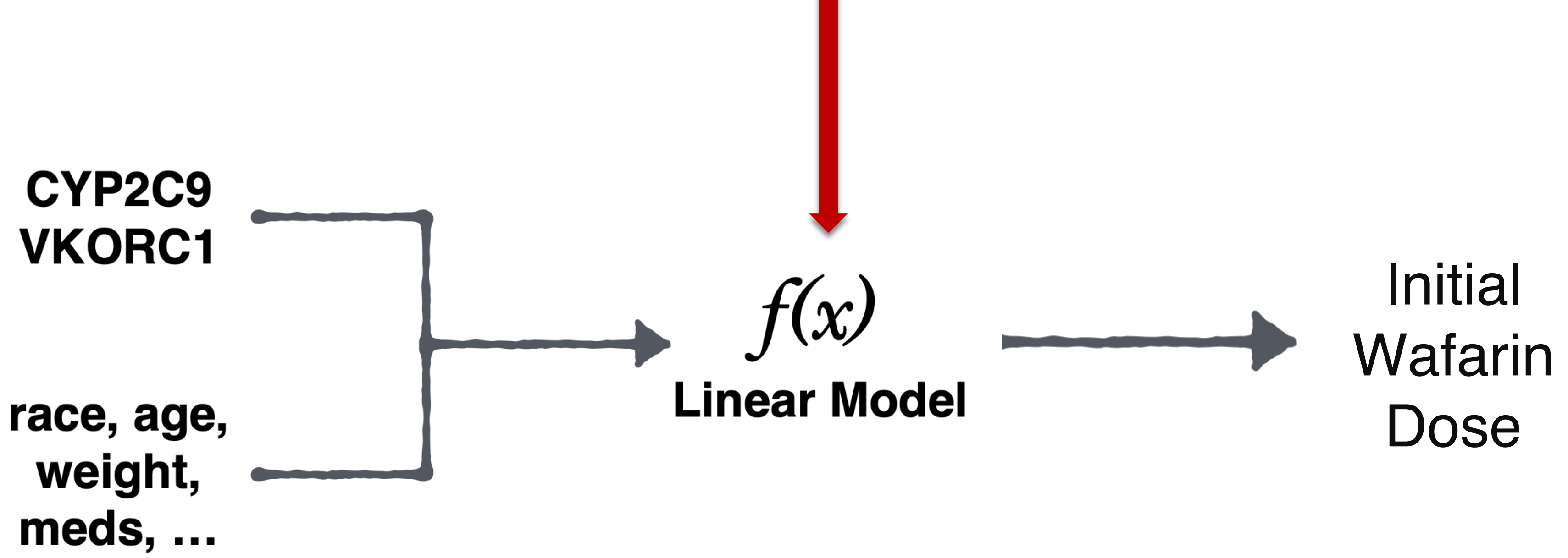
- Studies show linear regression performs best

age	height	weight	race	history	vkorc1	cyp2c9	dose



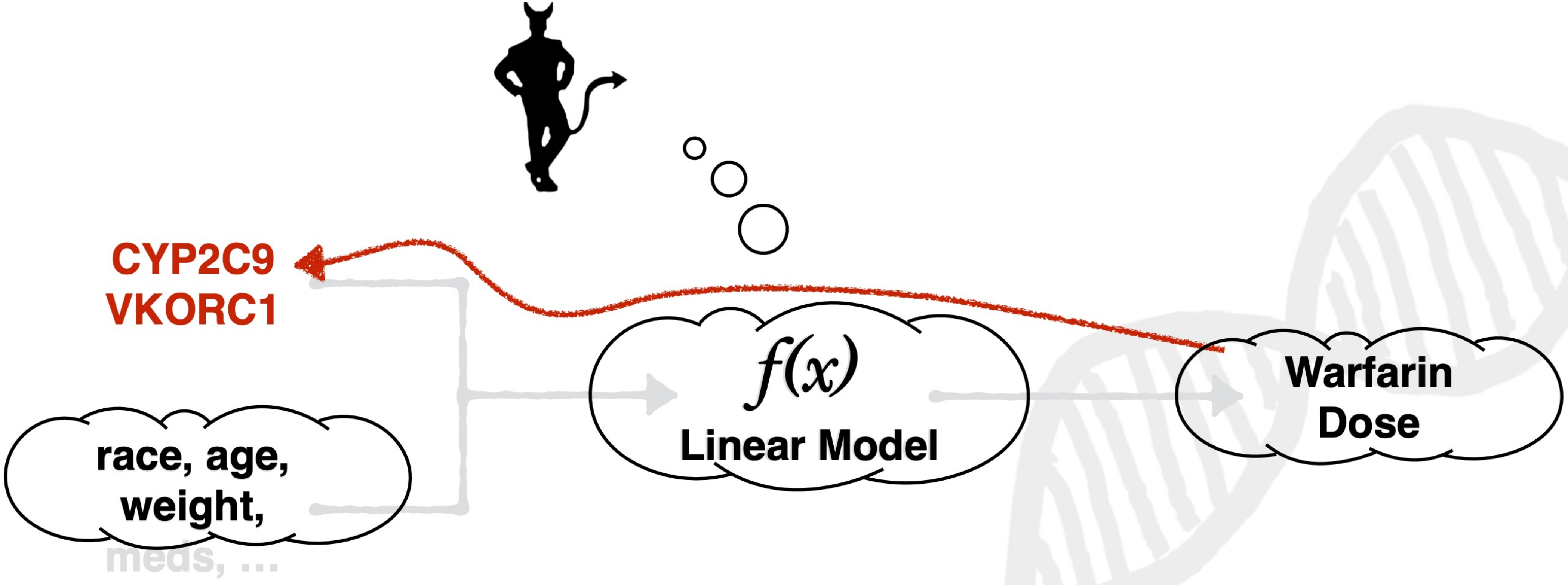
Pharmacogenetic Privacy

sqrt(dose) = **5.6044** + **0.2614** * age + **0.1092** * asian race - **0.2760** * black or african american -
0.8677 * vkorc1=A/G - **1.6974** * vkorc1=A/A - **1.9206** * cyp2c9=*2/*3 -
2.3312 * cyp2c9=*3/*3 + ···



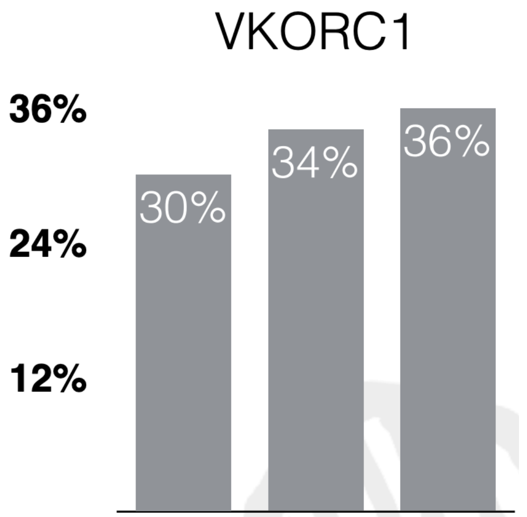
Pharmacogenetic Privacy

age	height	weight	race	history	vkorc1	cyp2c9	dose



Model Inversion

- Attacker knows:
  - basic demographics
  - black-box access to model
  - stable warfarin dose
  - marginal priors on patient distribution



- **Goal**: infer the patient’s genetic markers from this information

Model Inversion Algorithm

1. Compute all values that agree with given information

	age	height	weight	race	history	vkorc1	cyp2c9	dose		
$f(x)$	50-59	176.53	144.2	white				42.0	**49.7**	***p=0.23***
	50-59	176.53	144.2	white				42.0	**42.0**	***p=0.75***
	50-59	176.53	144.2	white				42.0	**39.2**	***p=0.01***

2. Find the most likely values among those that remain

Model Inversion Algorithm

When model is perfect

1. Input: $\mathbf{z}_K = (x_1, \ldots, x_k, y), f, p_{1,\ldots,d,y}$
2. Find the *feasible set* $\hat{\mathbf{X}} \subseteq \mathbf{X}$, i.e., such that $\forall \mathbf{x} \in \hat{\mathbf{X}}$
   (a) $\mathbf{x}$ matches $\mathbf{z}_K$ on known attributes: for $1 \leq i \leq k, \mathbf{x}_i = x_i$.
   (b) f evaluates to y as given in $\mathbf{z}_K$: $f(\mathbf{x}) = y$.
3. If $|\hat{\mathbf{X}}| = 0$, return $\perp$.
4. Return x_t that maximizes $\sum_{\mathbf{x} \in \hat{\mathbf{X}} : \mathbf{x}_t = x_t} \prod_{1 \leq i \leq d} p_i(\mathbf{x}_i)$

Model Inversion Algorithm

When model is imperfect

1. Input: $\mathbf{z}_K = (x_1, \ldots, x_k, y)$, f, $\boldsymbol{\pi}$, $p_{1,\ldots,d,y}$
2. Find the *feasible set* $\hat{\mathbf{X}} \subseteq \mathbf{X}$, i.e., such that $\forall \mathbf{x} \in \hat{\mathbf{X}}$
   (a) $\mathbf{x}$ matches $\mathbf{z}_K$ on known attributes: for $1 \leq i \leq k$, $\mathbf{x}_i = x_i$.
3. If $|\hat{\mathbf{X}}| = 0$, return $\perp$.
4. Return x_t that maximizes $\sum_{\mathbf{x} \in \hat{\mathbf{X}}: \mathbf{x}_t = x_t} \pi_{y,f(\mathbf{x})} \prod_{1 \leq i \leq d} p_i(\mathbf{x}_i)$

$\pi(y, y') = \mathbf{Pr}\left[\mathbf{z}_y = y | f(\mathbf{z}_\mathbf{x}) = y'\right]$ can be estimated by confusion matrices or standardized regression error

Limitation of This Method

- Inefficient if dimensions we want to recover are high
    - e.g., image domain

Model Inversion in Face Recognition

[Fredrikson et al. ‘15]

[Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures.

Fredrikson et al. CCS 2015]

How Do We Achieve This?

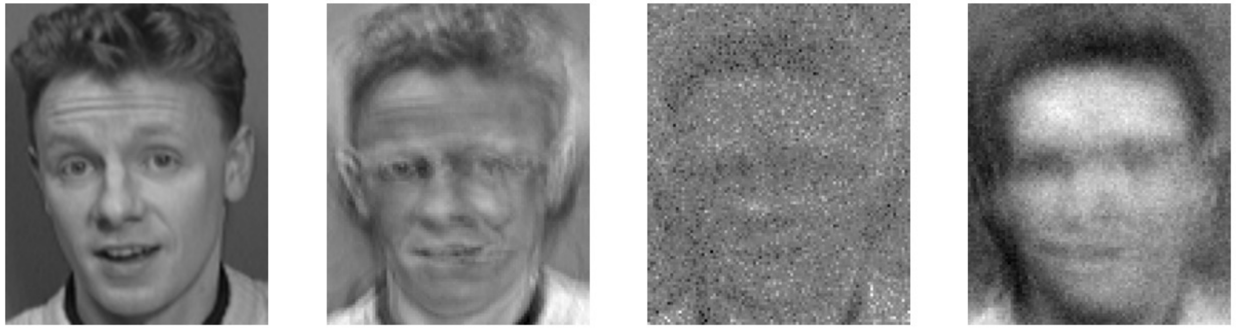
- Gradient Descent!
- Like adversarial attack, but needs some constraints in the direction that we move

$$\mathbf{x}_i \leftarrow \textsc{Process}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$$

- Follow the gradient until meets the confidence threshold

Experiments

- **Attack 3 models**: softmax regression, multi-layer perceptron, stacked denoising autoencoder network



Target **Softmax** **MLP** **DAE**

Algorithm 2 Processing function for stacked DAE.

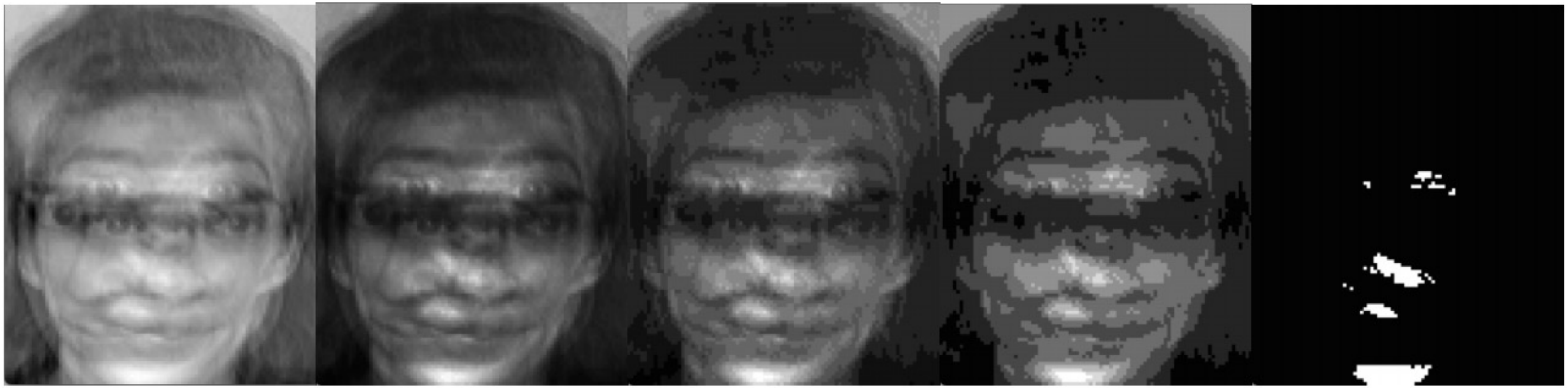
function PROCESS-DAE($\mathbf{x}$)
  encoder.DECODE($\mathbf{x}$)
  $\mathbf{x} \leftarrow$ NLMEANSDENOISE($\mathbf{x}$)
  $\mathbf{x} \leftarrow$ SHARPEN($\mathbf{x}$)
  return *encoder*.ENCODE($vecx$)

Black-box Attack

- Estimate each gradient with 2d black-box queries
- Works well for softmax regression (linear model)
- Takes too long for MLP and stacked DAE

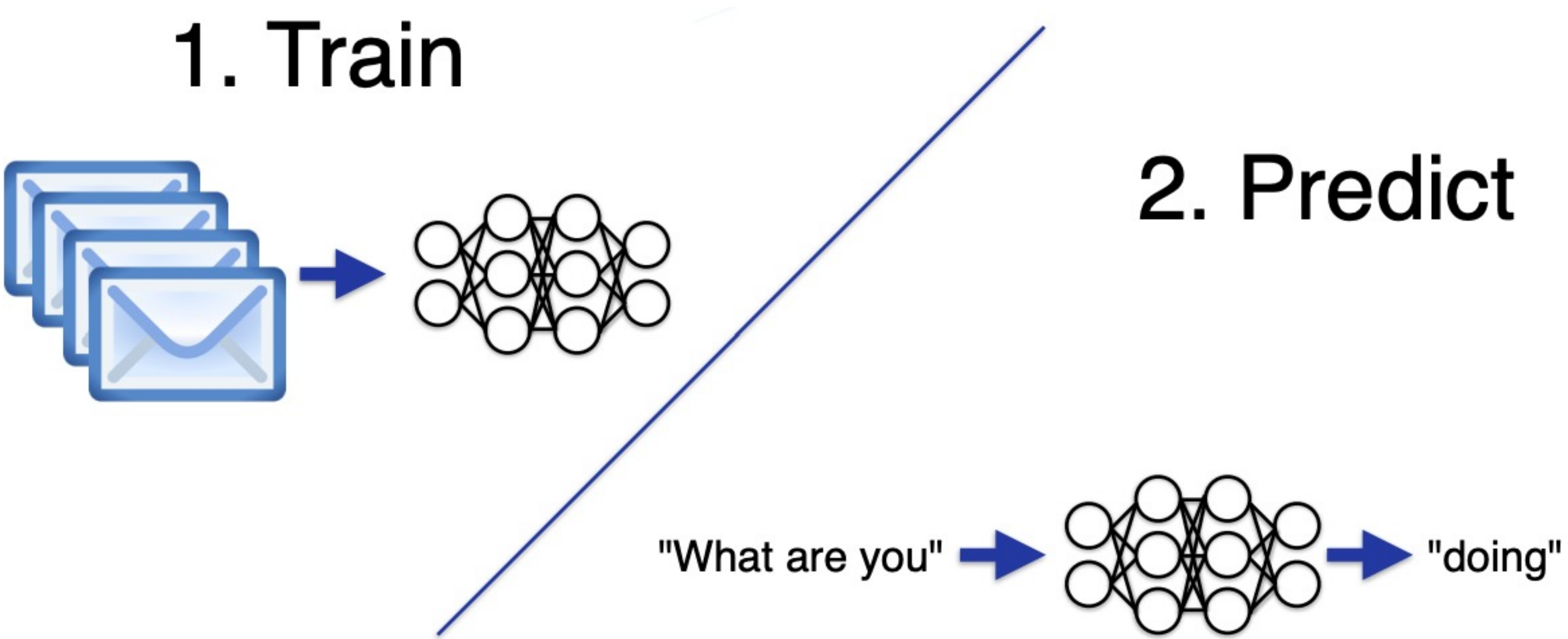
Possible Black-box Defense: Rounding

Output confidence values with less precision



no rounding $r = 0.001$ $r = 0.005$ $r = 0.01$ $r = 0.05$

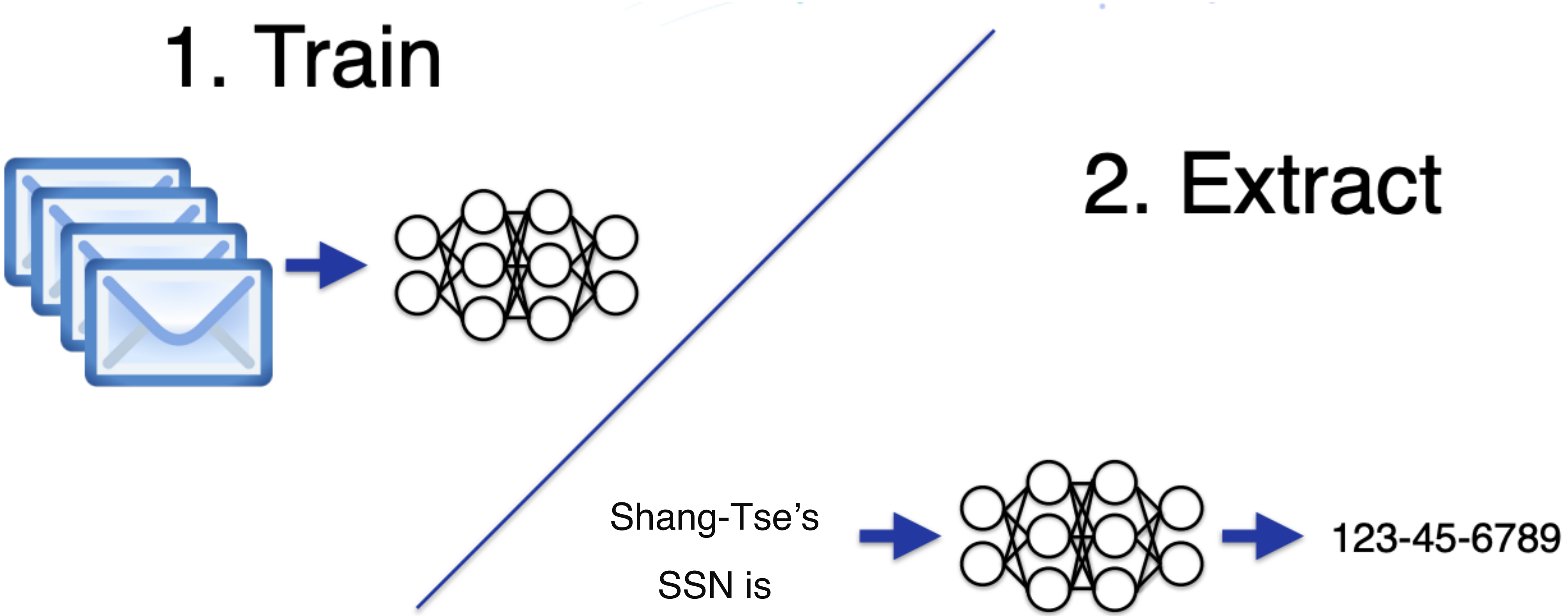
Other Applications



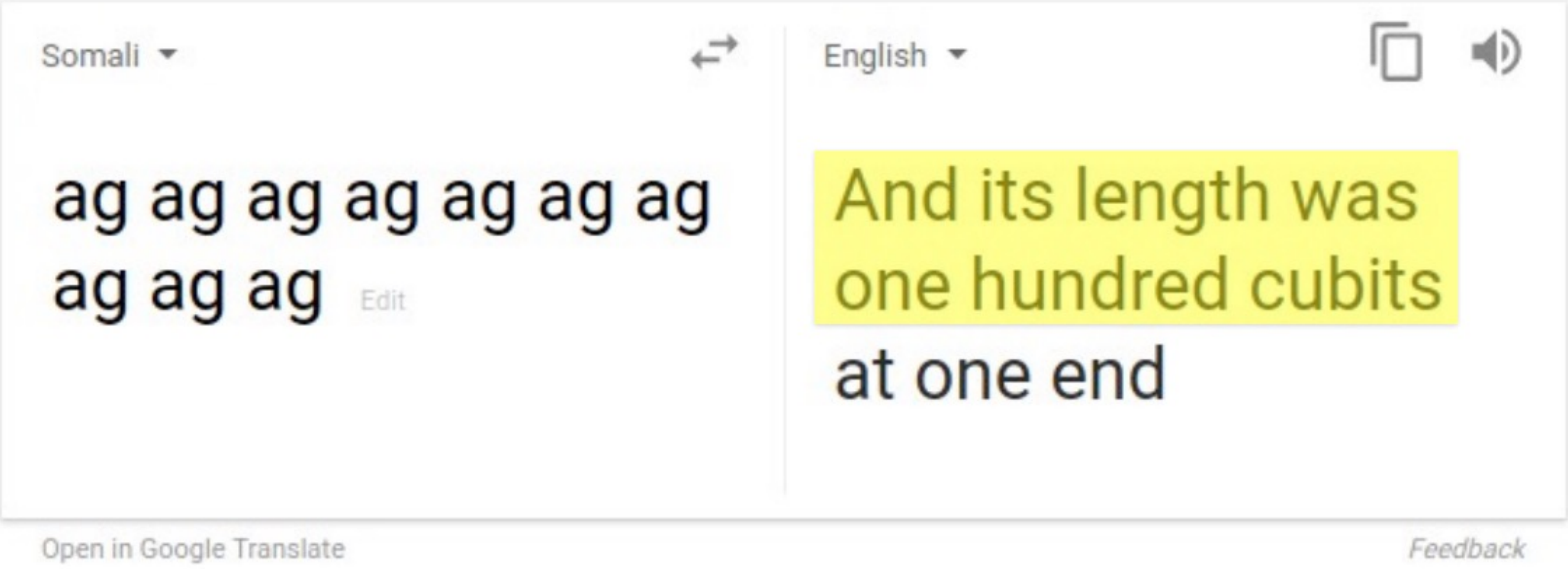
[The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Carlini et al. Usenix Security Symposium 2019]

Other Applications



[The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks
Carlini et al. Usenix Security Symposium 2019]


Somali
ag ag ag ag ag ag ag ag ag ag
Edit
English
And its length was one hundred cubits at one end
Open in Google Translate
Feedback

Google

"its length was one hundred cubits"

All Images News Shopping Videos More Settings Tools

About 2,850 results (0.17 seconds)

1 Kings 7:2 He built the House of the Forest of Lebanon a hundred ...
https://biblehub.com/1_kings/7-2.htm ▾
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

1 Kings 7:2 NLT: One of Solomon's buildings was called the Palace of ...
https://biblehub.com/nlt/1_kings/7-2.htm ▾
For he built the house of the forest of Lebanon; **its length was one hundred cubits**, and its breadth fifty cubits, and its height thirty cubits, on four rows of cedar ...

Extracting Training Data

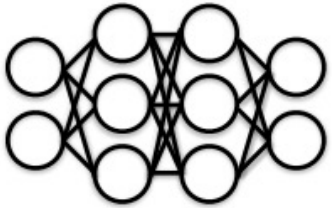
- P(My SSN is 000-00-0000) = 0.01
- P(My SSN is 000-00-0001) = 0.02
- P(My SSN is 000-00-0002) = 0.01
- ….
- P(My SSN is 123-45-6788) = 0.00
- P(My SSN is 123-45-6789) = **0.32**
- …
- P(My SSN is 999-99-9999) = 0.01

Does It Work in Practice?

- The brute-force search needs too many queries
- Better algorithm inspired by Dijkstra's shortest path search
  - Takes only 10^5 queries, four orders of magnitude fewer than the brute-force approach

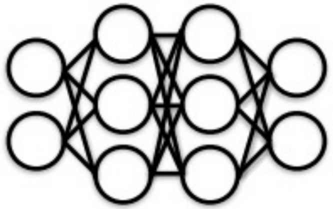
Choose Between

Model A



Accuracy: 96%
High Memorization

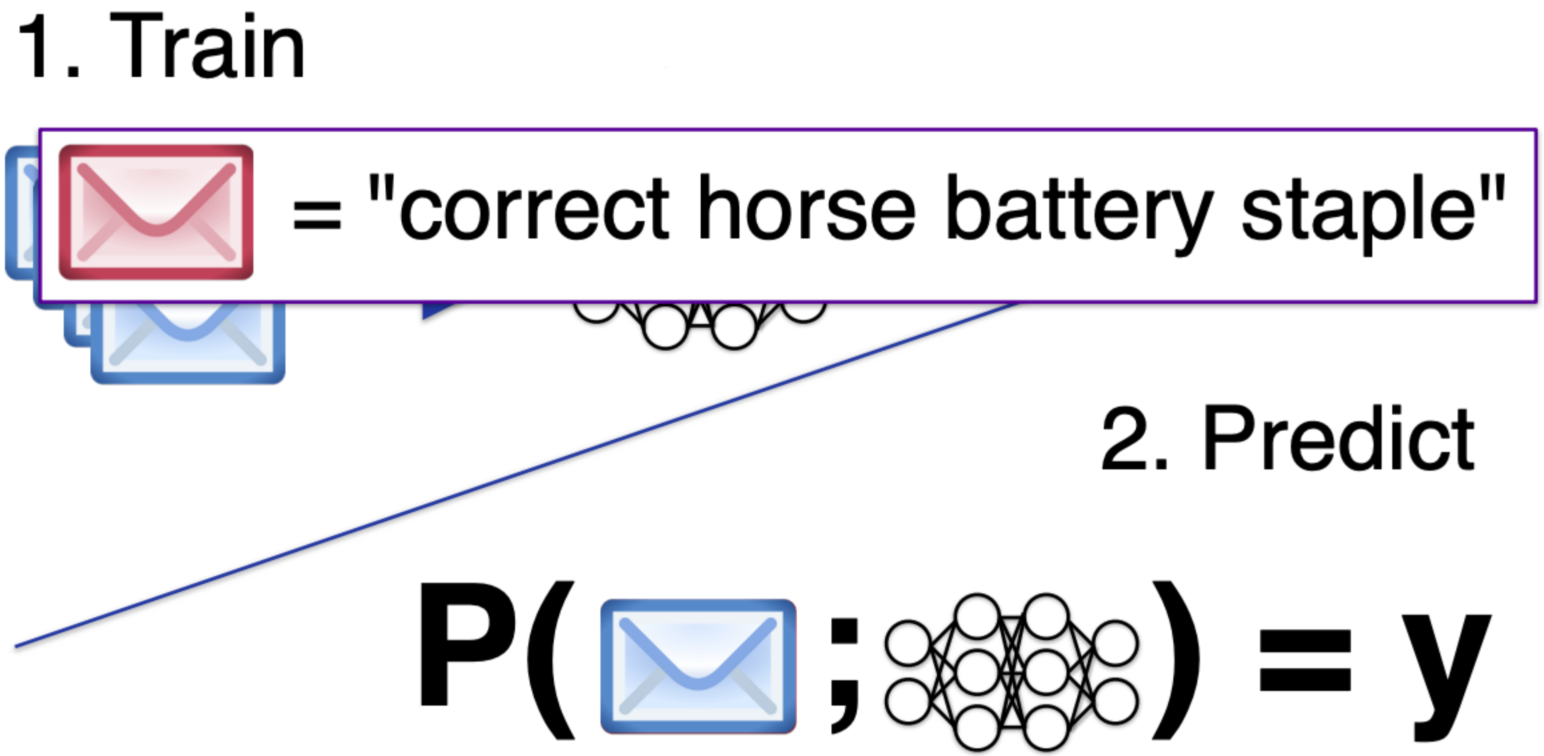
Model B



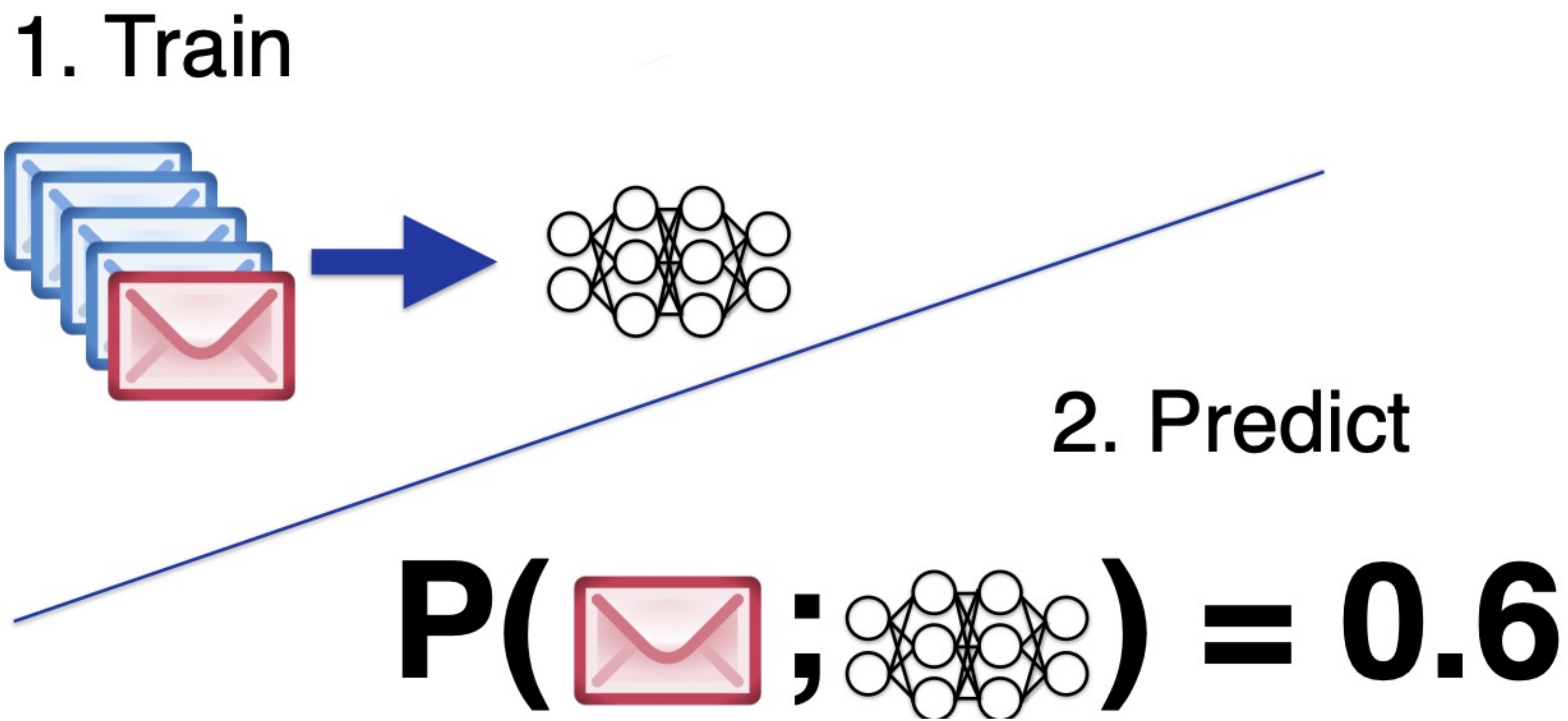
Accuracy: 92%
No Memorization

Exposure-based Testing Method

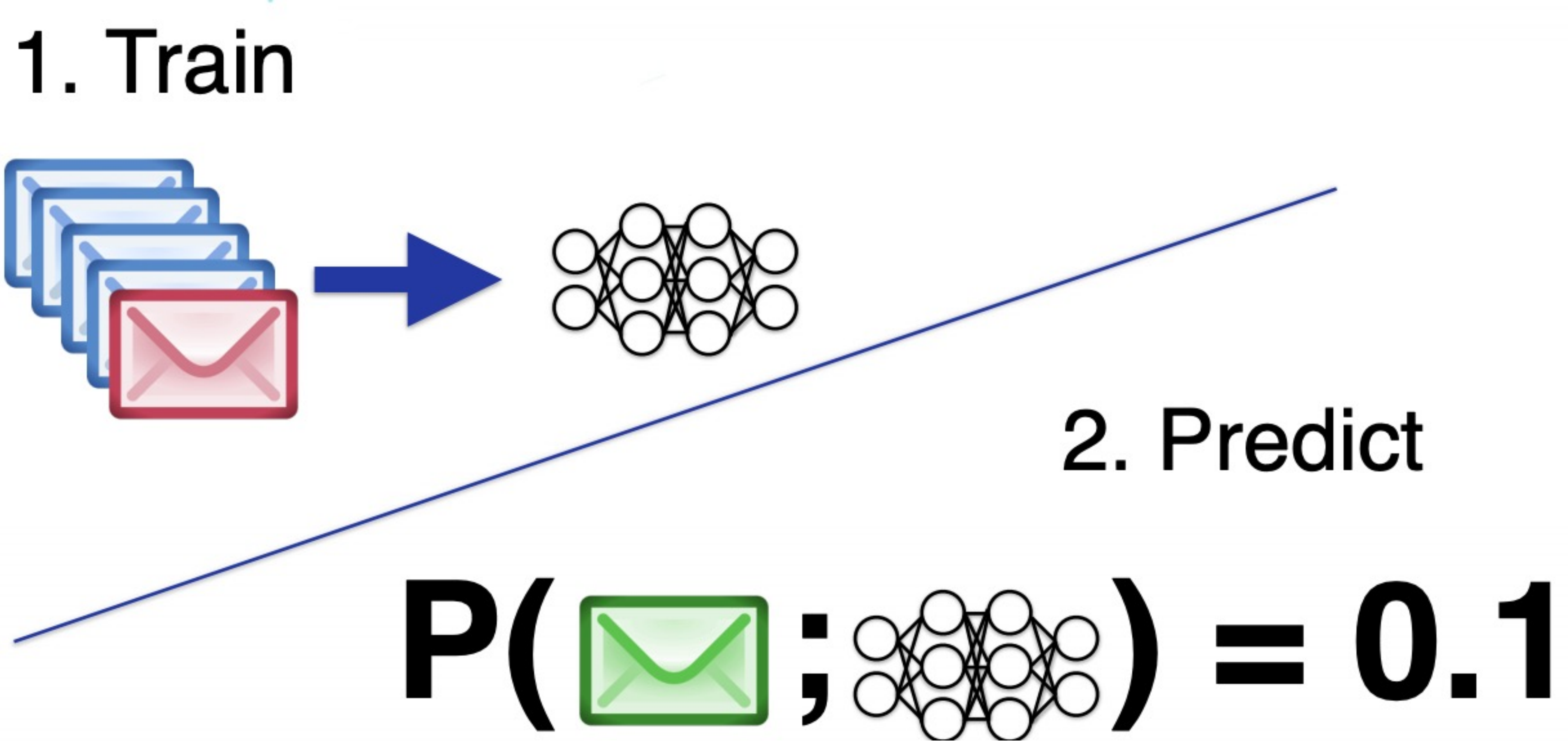
- If a model memorizes completely random canaries, it probably also is memorizing other training data



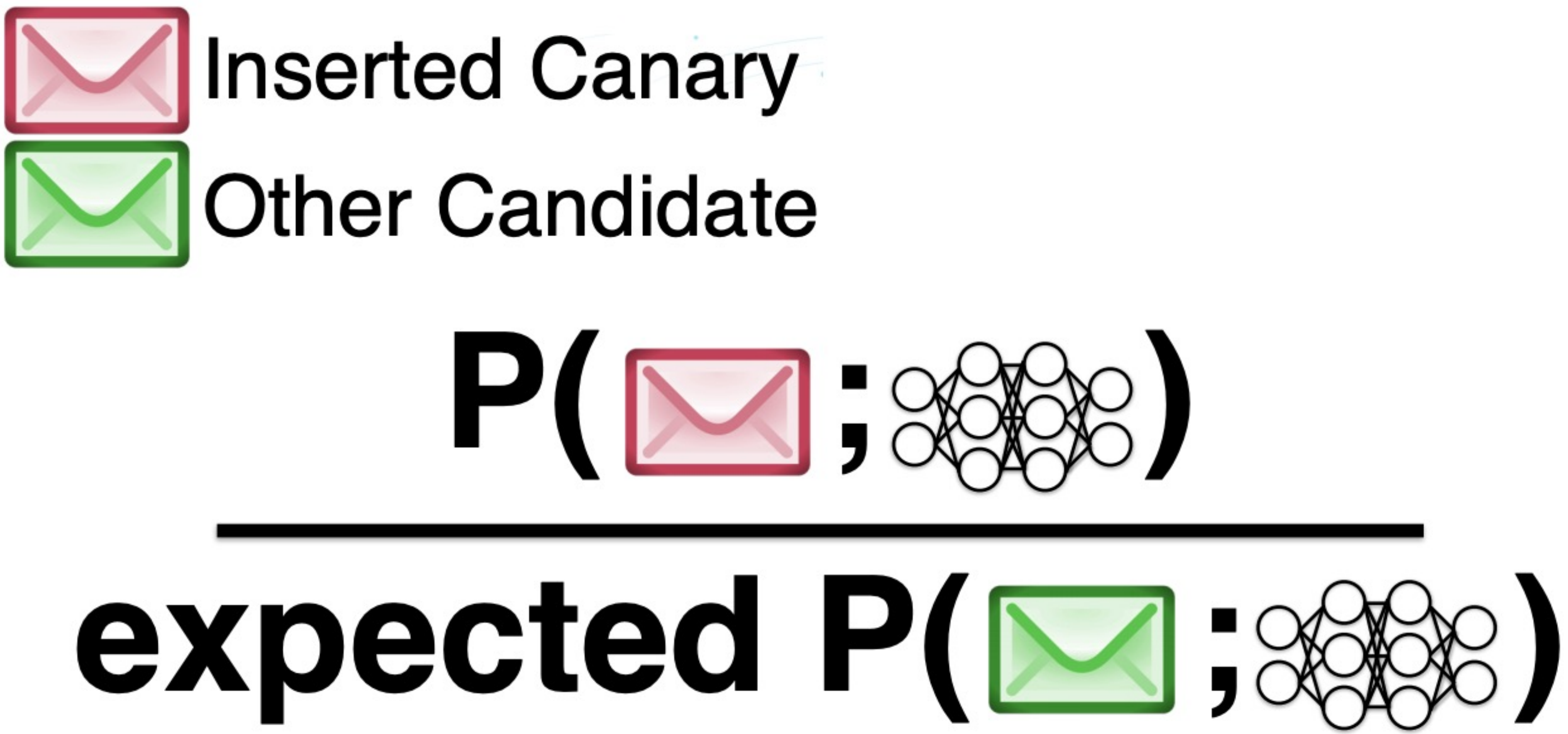
Exposure-based Testing Method



Exposure-based Testing Method



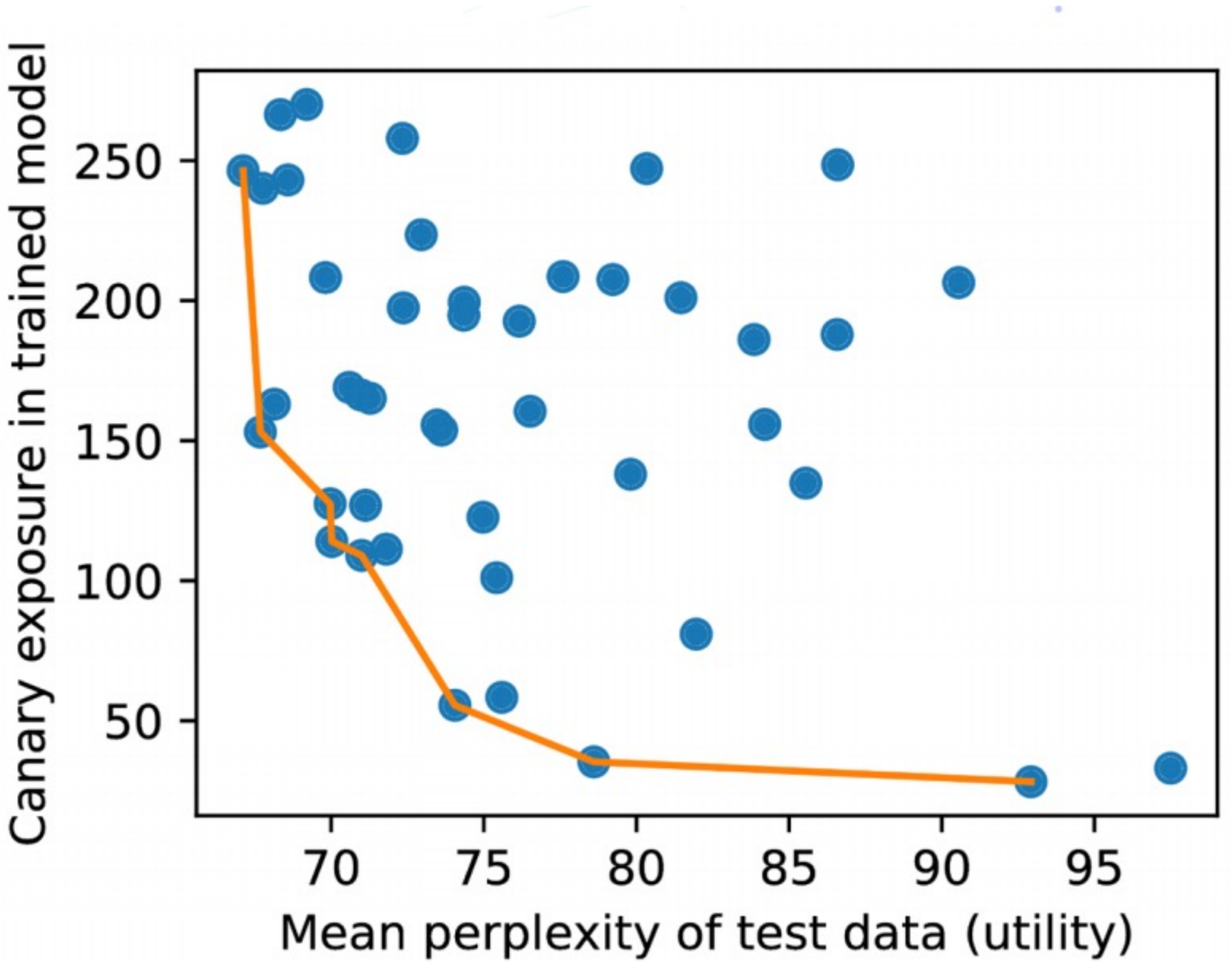
Exposure



Summary of the Testing Algorithm

1. Generate canary 
2. Insert  into training data
3. Train model
4. Compute exposure of 
   (compare likelihood to other candidates) 

How to Choose Models?



Provable Defense?

- Differential Privacy
  - We will introduce this framework later in this course