Exploiting Leakage
in Searchable Encryption and Machine Learning

Tom Ristenpart

CORNELL TECH

Covering joint work with:
David Cash, Paul Grubbs, Jason Perry (Searchable encryption)
Matthew Fredrikson, Eric Lantz, Simon Lin, David Page, Somesh Jha (ML)

Plaintext keyword search

Email client

Keyword stemming

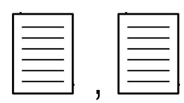
The attached contract is ready for signature. Please print 2 documents and have Atmos …

Keyword	Documents
contract	1, 7
signatur	8, 9, 1, 15, 200

Email storage provider

Upload documents

Search: “contract”



Appended-PRF Searchable Encryption

Email client

Encrypt plaintext & keyed hash of keywords

89123fdbf32a665befg8819890fbacda
4320182321a1343187fabaedf3140fba

H_K(attach) H_K(contract) H_K(ready) …

Upload encrypted documents

Keyword	Documents
H_K(contract)	1, 7
H_K(signatur)	8, 9, 1, 15, 200

Email storage provider

Appended-PRF Searchable Encryption

Email client

Encrypt plaintext & keyed hash of keywords

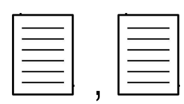
$7813fed = H_K(contract)$

89123fdbf32a665befg8819890fbacda
4320182321a1343187fabaedf3140fba

ab34df 7813fed 873f63 …

Upload encrypted documents

Search: "7813fed"



Keyword	Documents
7813fed	1, 7
456abc3	8, 9, 1, 15, 200

Email storage provider

Legacy compatible:
Works with existing plaintext storage interfaces

Two more schemes to consider

(2) Unordered appended-PRFs

Randomize order of PRF values

The attached contract is ready for signature. Please print 2 documents and have Atmos …

H_K(contract) H_K(ready) H_K(attach) …

(3) Encrypted index

Keyword	Documents
H_K(contract)	
H_K(signatur)	

Encrypt each document list under keyword-specific key

Qualitative comparison of schemes

Appended-PRF scheme used in industry

bitglass CipherCloud Trust in the Cloud™ skyhigh CRYPTERON SECURE YOUR CLOUD DATA

Unordered appended-PRF used in research literature

Mimesis Aegis [Lau et al. 2014]
ShadowCrypt [He et al. 2014]

Encrypted index in literature & starting to appear in industry

[Cash et al. 2014]

skyhigh

Qualitative comparison of schemes

Appended-PRF scheme used in industry

Unordered appended-PRF used in research literature

Encrypted index in literature & starting to appear in industry

Ease of deployment

Provable security claims

Leakage-abuse attacks

[Islam, Kuzu, Kantarcioglu – 2013]
[Cash, Grubbs, Perry, R. – 2015]

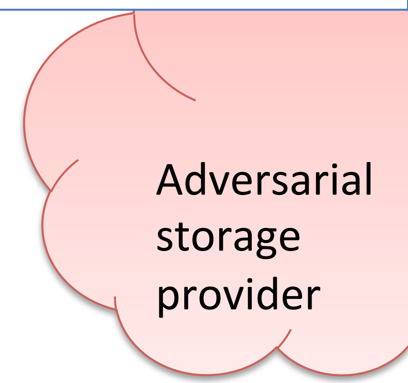
All searchable encryption leaks information about plaintexts and queries. Appended-PRF case:

Keyword	Documents
H_K(contract)	1, 7
H_K(signatur)	8, 9, 1, 15, 200

H_K(attach) H_K(contract) H_K(ready) …

Upload encrypted documents

Search: "H_K(contract)"



Leakage-abuse attacks

[Islam, Kuzu, Kantarcioglu – 2013]
[Cash, Grubbs, Perry, R. – 2015]

All searchable encryption leaks information about plaintexts and queries. Appended-PRF case:

"Keyword 7813fed came second in Document 1"
(Keyword location)

ab34df 7813fed 873f63 …

Upload encrypted documents

Search: "7813fed"

"Keyword 7813fed searched often"
(Search frequency)

Keyword	Documents
7813fed	1, 7
456abc3	8, 9, 1, 15, 200

Adversarial storage provider

"Document 1 and 7 both contain 7813fed" (Co-occurrence relationships)

Unordered appended-PRF: order of keywords not leaked

Encrypted index: order of keywords not leaked & leakage only after queries made

We don’t know answers to basic security questions:

- Does leakage damage confidentiality?
- How much more security does one achieve via more complex schemes?
- What adversarial capabilities are likely to arise in practice?

Leakage-abuse attack taxonomy

Attacker goal	Query recovery	
	Plaintext recovery	
Attacker capabilities	Passive	Observe queries and stored ciphertexts
	Active	Force insertion of documents and/or queries
Document knowledge	Full	Know all plaintexts exactly
	Partial	Know some plaintexts
	Distributional	Know similar plaintexts

IKK 2013 against encrypted index: Query recovery | Passive | Full

Simulations with Enron email corpus: 80% of queries recoverable
We'll come back to this

Partial plaintext recovery against appended-PRF

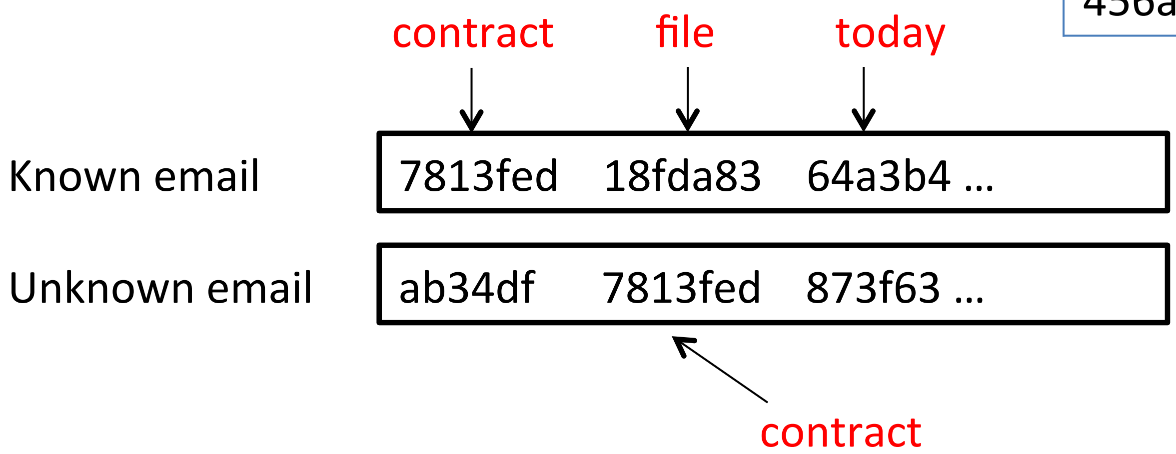
[Cash, Grubbs, Perry, R. – 2015]

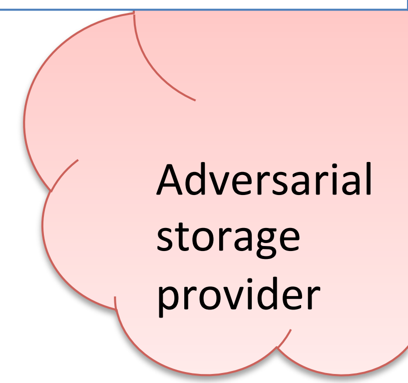
Plaintext recovery

Passive

Partial

Keyword	Documents
7813fed	1, 7
456abc3	8, 9, 1, 15, 200





Partial plaintext recovery against appended-PRF

[Cash, Grubbs, Perry, R. – 2015]

Plaintext recovery

Passive | Partial

Simulations with Enron email corpus
- 30,109 emails from employee sent_mail folders
- Adversary knows 20 random emails (0.06%)
- Simply match keywords in known emails to unknown

Unknown email plaintext	The attached contract is ready for signature. Please print 2 documents and have Atmos execute both and return same to my attention. I will re- turn an original for their records after ENA has signed. Or if you prefer, please provide me with the name / phone # / address of your customer and I will Fed X the Agreement.
Recovered information	attach contract signatur pleas print 2 document have execut both same will origin ena sign prefer provid name agreement

Randomizing hash order

Plaintext recovery

Passive Partial

Leaving hashes in document order makes attack easy

Simple change: randomize order of hashes to leak less information (sort by hash value)

contract file today

Known email: 7813fed 18fda83 64a3b4 …

Unknown email: ab34df 7813fed 873f63 …

contract

Randomizing hash order

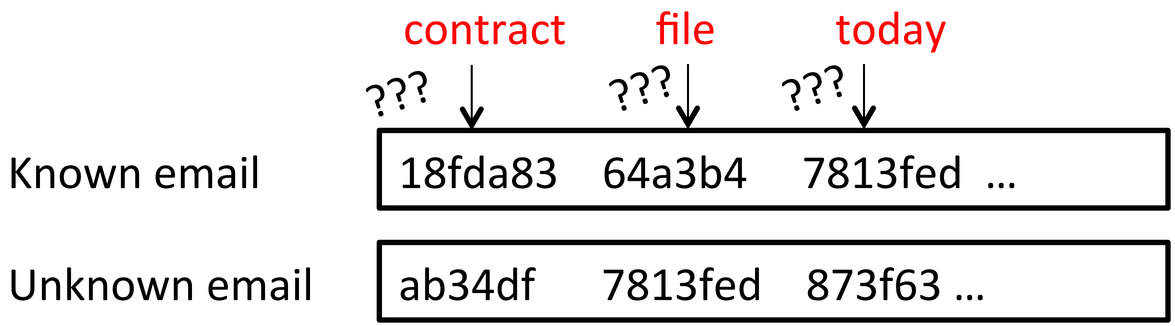
Plaintext recovery

Passive

Partial

Leaving hashes in document order makes attack easy

Simple change: randomize order of hashes to leak less information
(sort by hash value)



Order issue left implicit in prior work

Mimesis Aegis: randomizes order due to Bloom filter

ShadowCrypt: implementation randomizes order, paper does not discuss

Chosen-email attacks

Plaintext recovery	
Active	Distributional

Email client

89123fdbf32a665befg8819890fbacda
4320182321a1343187fabaedf3140fba

H_K(signatur) H_K(contract)

Insert new email

Keyword	Documents
H_K(contract)	1, 7
H_K(signatur)	8, 9, 1, 15, 200

Adversarial storage provider

Send victim an email

To: victim@victim.com
From: sally@sally.net

Contract signature

Chosen-email attacks

Plaintext recovery

Active | Distributional

Email client

89123fdbf32a665befg8819890fbacda
4320182321a1343187fabaedf3140fba

456abc3 7813fed

Insert new email

Keyword	Documents
7813fed	1, 7
456abc3	8, 9, 1, 15, 200

Adversarial storage provider

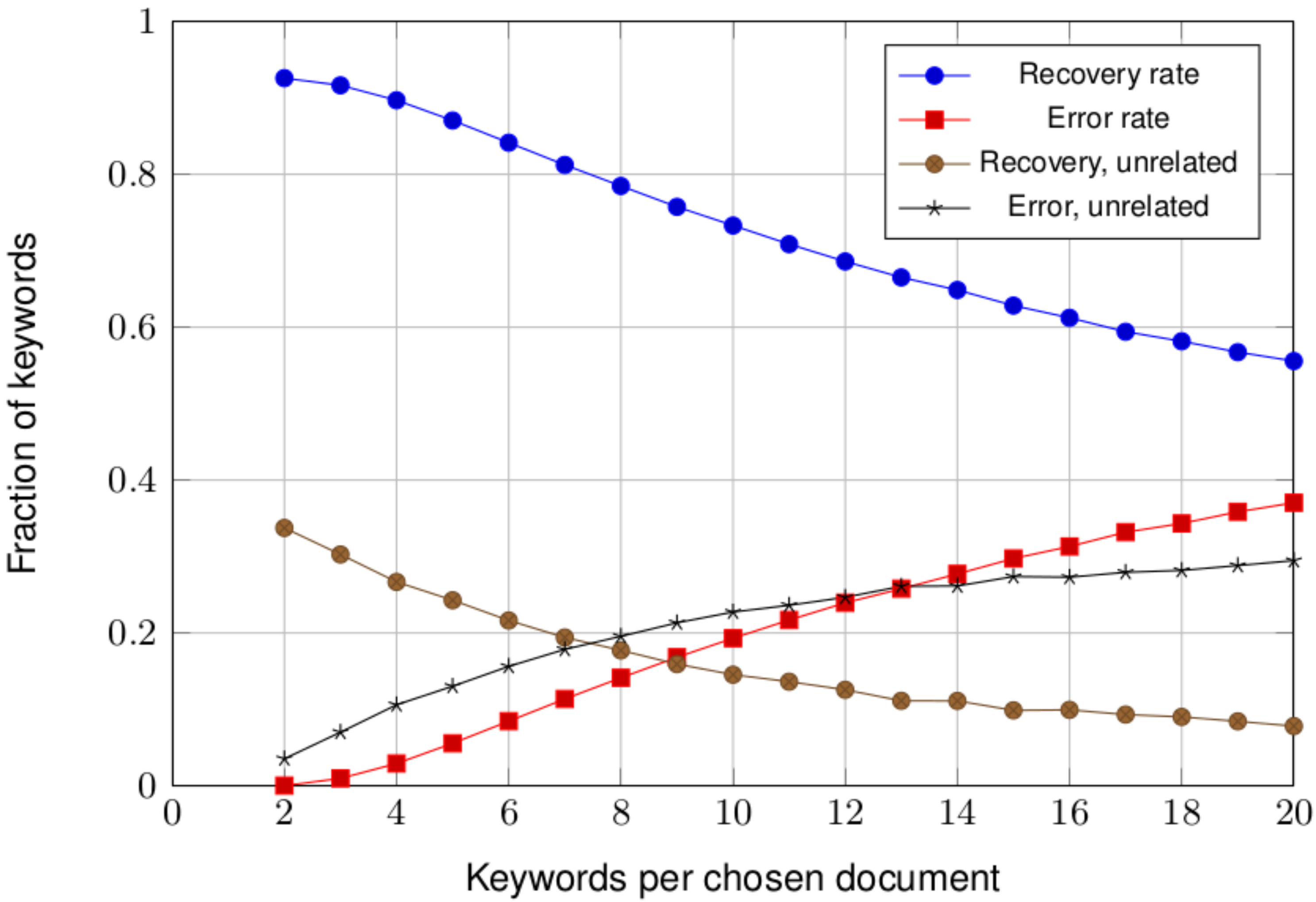
Disambiguate 2 keywords
by their expected frequency

Send victim an email

To: victim@victim.com
From: sally@sally.net

Contract signature

Disambiguation performance



Related: split Enron into training and testing sets, train frequency on training
Unrelated: train on distinct email corpus (Apache corpus)

Case studies of three attacks

1. Simple attack against *appended-PRF*

Plaintext recovery	
Passive	Partial

2. Chosen-email attack against *unordered appended-PRF*

Plaintext recovery	
Active	Distributional

3. Query recovery against *encrypted index schemes*

Query recovery	
Passive	Full

IKK query recovery attack

Query recovery

Passive | Full

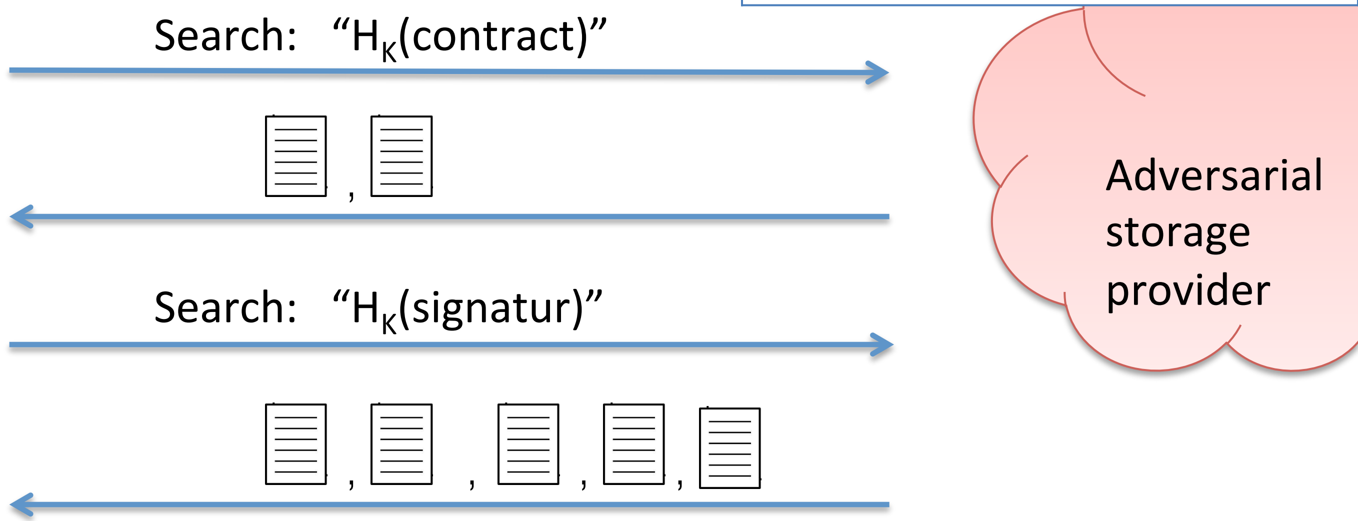
Adversary knows ***full plaintext corpus***

Goal is to uncover search query keywords used by client

Keyword	Documents
H_K(contract)	1, 7
H_K(signatur)	8, 9, 1, 15, 200

Email client

Uniformly selects keywords to search



IKK detail expensive attack using simulated annealing to solve NP-complete problem sufficient to reveal queries

We give way simpler attack

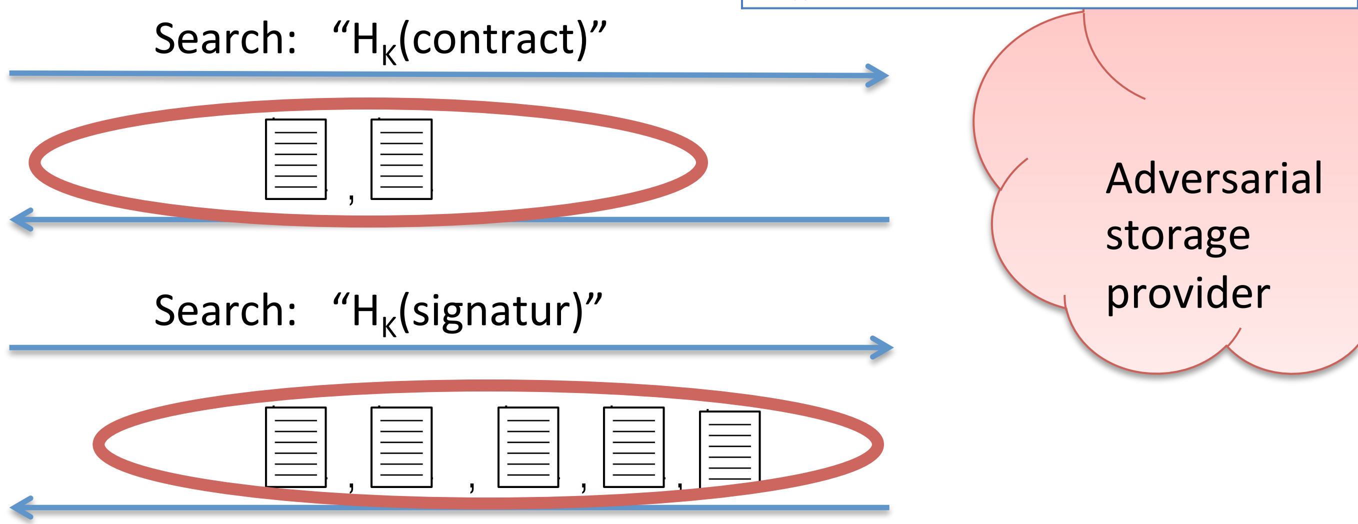
Query recovery | Passive | Full

Adversary knows ***full plaintext corpus***
Goal is to uncover search query keywords used by client

Keyword	Documents
H_K(contract)	1, 7
H_K(signatur)	8, 9, 1, 15, 200

Email client

Uniformly selects keywords to search



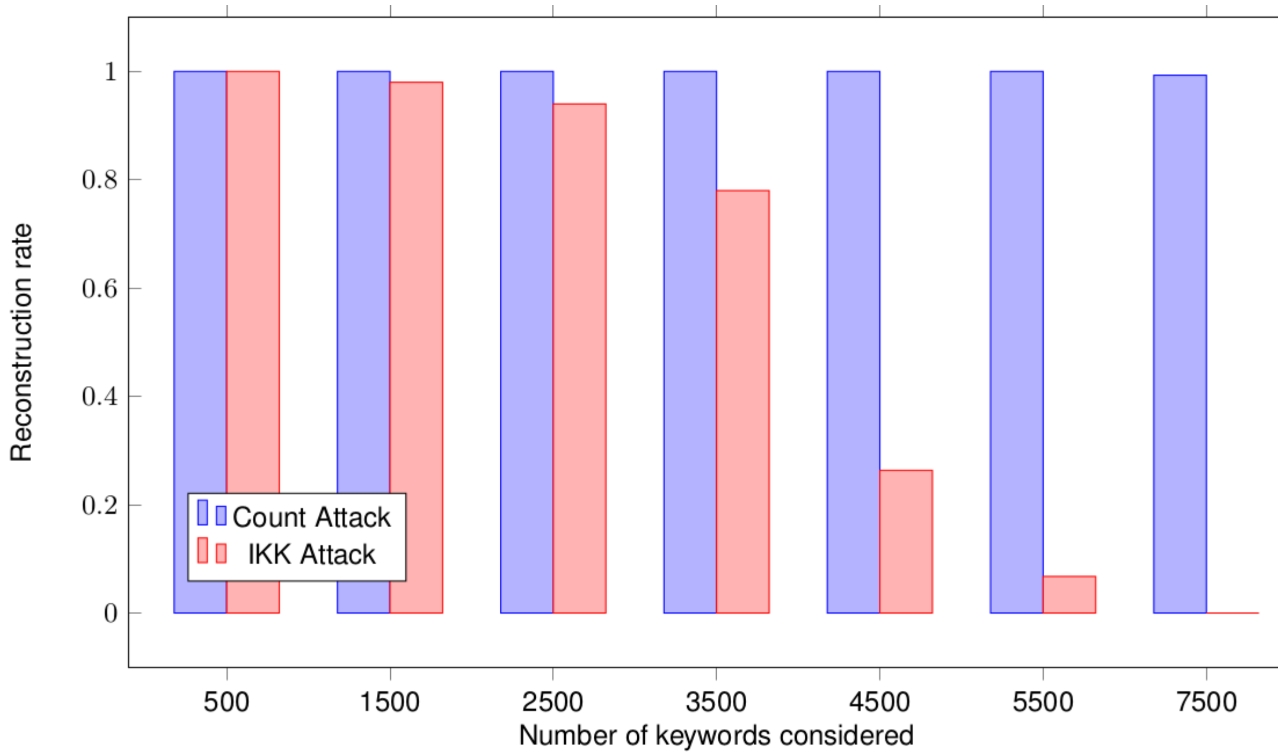
Attacker sees number of documents returned
Many keywords appear in a unique number of documents
Disambiguate with co-occurrence relationships

IKK vs "count" attack

Query recovery

Passive

Full



Subset of Enron emails (known to attacker)
Most popular x keywords considered
10% of keywords uniformly sampled and queried

Summary of leakage-abuse attacks

Provable security must be (at least) paired with empirical security analyses

Lots of open questions:

- Leakage of richer queries
- Role of updates
- Effect of re-encryption
- Viability of active attacks in practice

And challenges:

- Better data sets for simulations
- Query traces
- Countermeasures

Part 2:
Machine learning model inversion

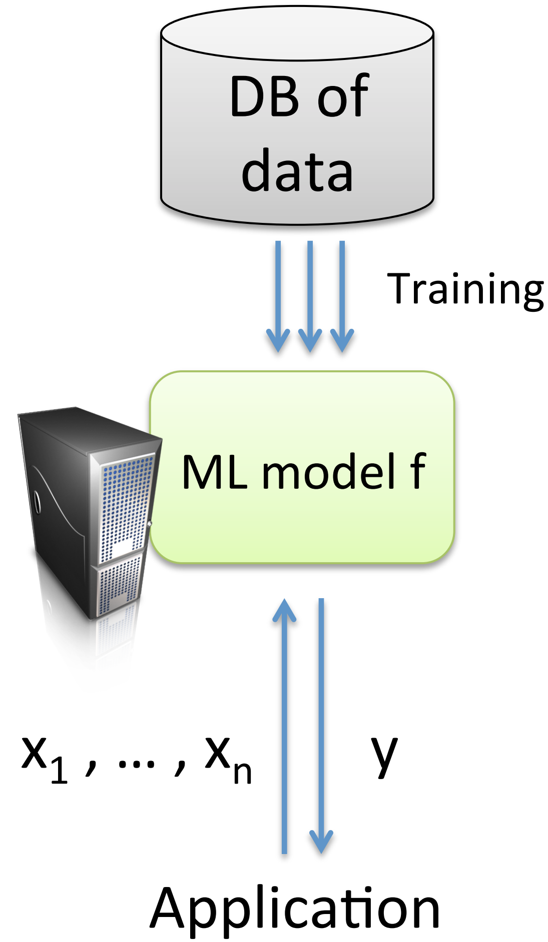
Machine learning (ML) systems

(1) Gather some labeled data

(2) Train ML model f from data

$$f(x_1, \ldots, x_n) = y$$

(3) Use f in some application or publish it for others to use



Increasing use of ML

Medical applications

WARFARINDOSING www.WarfarinDosing.org

Facial recognition

facebook

SkyBiometry

Cloud-based Face Detection and Recognition API

Cloud computing

Prediction API

bigml®

Microsoft Azure Machine Learning

Privacy concerns in machine learning?

Release of sensitive data?

Even de-identified data dangerous

[Sweeney ‘00]

[Naranayan & Shmatikov ‘08] …

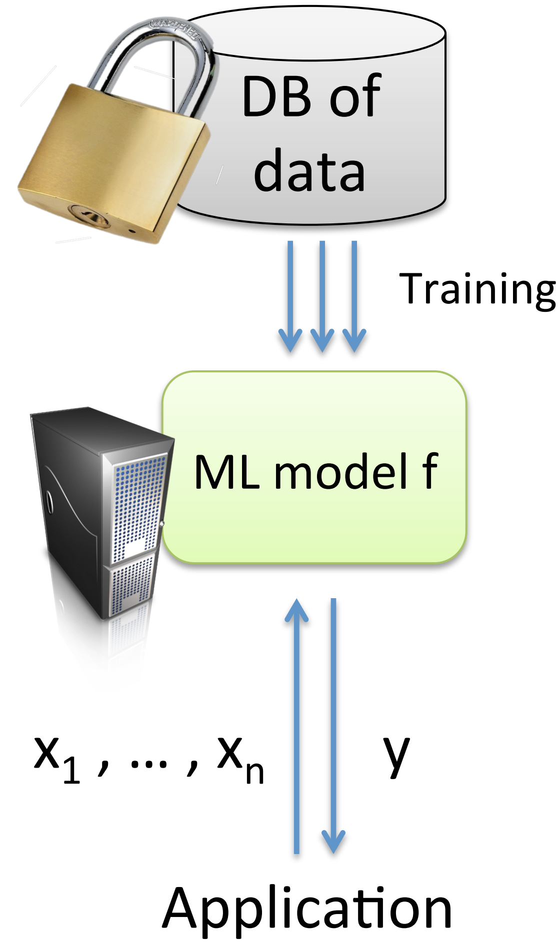
k-anonymity [Sweeney ‘02]

Differential privacy

[Dwork, McSherry, Nissim, Smith ‘06]

…

Overarching lesson:
Don’t release sensitive data sets without due care



Privacy concerns in machine learning?

Release of sensitive data?

Even de-identified data dangerous

[Sweeney '00]

[Naranayan & Shmatikov '08] …

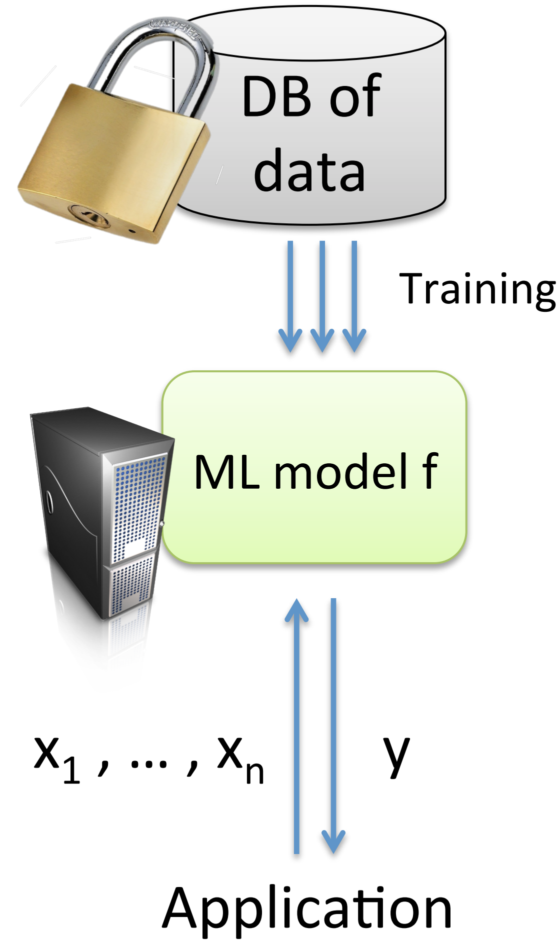
k-anonymity [Sweeney '02]

Differential privacy

[Dwork, McSherry, Nissim, Smith '06]

…

What about risks related to adversarial access to (just) model f?



[Ateniese et al. 2013]: Determine one bit of info about DB given ability to download f

New privacy concerns in ML

[Fredrikson, Lantz, Lin, Jha, Page, R. – Security `14]
[Fredrikson, Jha, R. – CCS `15]

Model inversion attacks:

(1) Linear regression for personalized medicine

Predict genotypes of patients

(2) Decision trees trained from lifestyle surveys

Predict marital infidelity of training set members

(3) Neural networks for facial recognition

Recover recognizable images of training set members

Preliminary investigation of countermeasures

Differential privacy

Sensitive-feature-aware CART decision trees

Rounded confidence values

Privacy in pharmacogenetics

[Fredrikson, Lantz, Lin, Jha, Page, R. – Security `14]

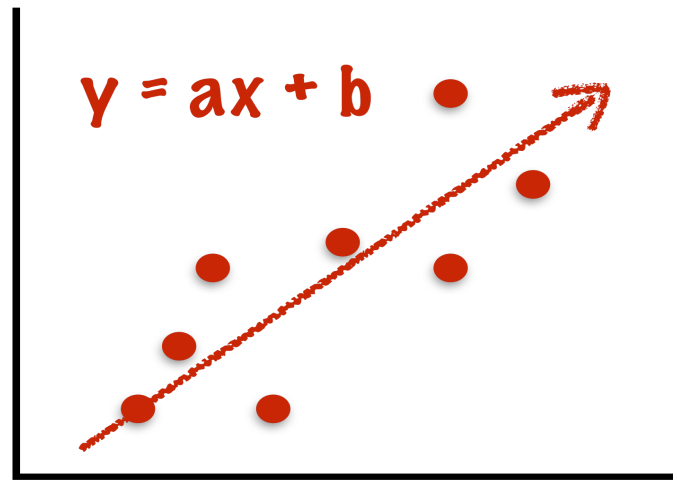
Case study in context of ***personalized medicine***

WARFARINDOSING

www.WarfarinDosing.org

<u>IWPC study:</u>

- Linear regression based classifier
- Trained on demographics, health history, and genetic markers
- Predicts initial dose of warfarin
- [IWPC] researchers showed evidence that this outperformed clinical practice



Data set is publicly available (in de-identified form), but similar data sets must be private

WARFARINDOSING

www.WarfarinDosing.org

> Warfarin Dosing

> Clinical Trial

> Outcomes

> Hemorrhage Risk

> Patient Education

> Contact Us

> References

> Glossary

> About Us

User:
Patient:
Version 2.42
Build : Feb 05, 2014

Required Patient Information

Age: [ ] **Sex:** -Select- **Ethnicity:** -Select-

Race: -Select-

Weight: [ ] lbs **or** [ ] kgs

Height: ([ ] feet and [ ] inches) **or** ([ ] cms)

Smokes: -Select- **Liver Disease:** -Select-

Indication: -Select-

Baseline INR: [ ] **Target INR:** [ ] Randomize & Blind

Amiodarone/Cordarone® Dose: [ ] mg/day

Statin/HMG CoA Reductase Inhibitor:
-Select-

Any azole (eg. Fluconazole): -Select-

Sulfamethoxazole/Septra/Bactrim/Cotrim/Sulfatrim: -Select-

Genetic Information

VKORC1-1639/3673: Not available/pending

CYP4F2 V433M: Not available/pending

GGCX rs11676382: Not available/pending

CYP2C9*2: Not available/pending

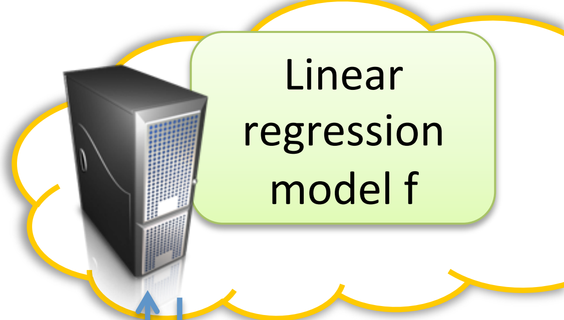
CYP2C9*3: Not available/pending

CYP2C9*5: Not available/pending

CYP2C9*6: Not available/pending

Warfarin model inversion attack

[Fredrikson, Lantz, Lin, Jha, Page, R. – Security `14]



$$f\,(\,x_1\,,\,\ldots\,,\,\boxed{x_n}\,)\ =\ y$$

Demographic information
Health history
Genotype

Suggested initial dose of warfarin

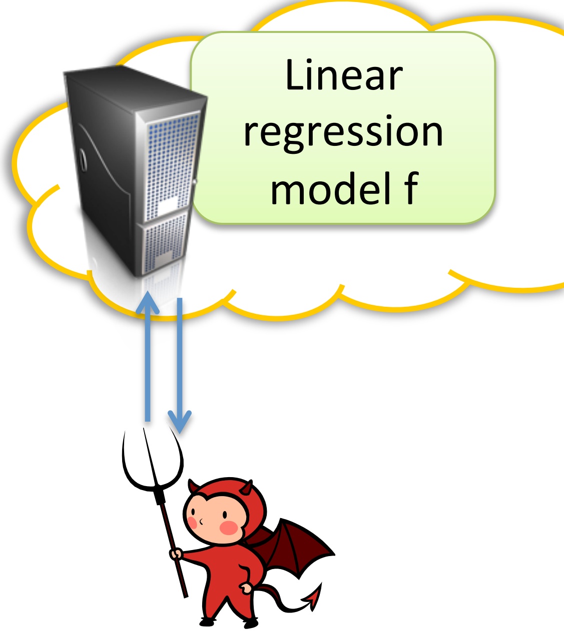
Info on x_1 , … , x_{n-1}
Stable dose y’ (y’ ≠ y)
Model f

Model inversion algorithm

Target person’s genotype

Warfarin model inversion attack

[Fredrikson, Lantz, Lin, Jha, Page, R. – Security `14]



x_n takes on values in set $\{v_1,\ldots,v_s\}$

(1) Compute feasible set of input vectors:

$$z_1 = (x_1,\ldots,x_{n-1},v_1)$$
$$z_2 = (x_1,\ldots,x_{n-1},v_2)$$
$$\ldots$$
$$z_s = (x_1,\ldots,x_{n-1},v_s)$$

(2) Compute $y_j = f(z_j)$ for each j

(3) Output v_j that maximizes

Realizes MAP estimator
(optimal subject to info available)

$$\sum_{j=1}^{s}\left(\pi(y, y_j)\cdot\prod_{i=1}^{n}p(z_j[i])\right)$$

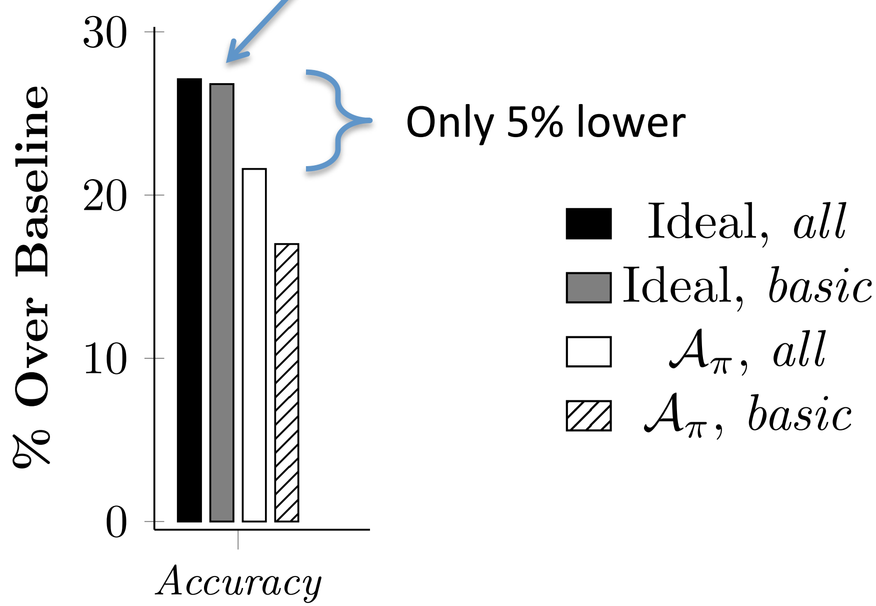
Weight by error

Independent priors

Model inversion results for IWPC model

Linear regression model directly trained from dataset

Baseline is guessing without access to model (36% accuracy)



Only 5% lower

Everything but genotype

Basic demographics about person

VKORC1

Model aids attacker in prediction almost as much as training directly on data set

New privacy concerns in ML

Model inversion attacks:

(1) Linear regression for personalized medicine

Predict genotypes of patients

(2) Decision trees trained from lifestyle surveys

Predict marital infidelity of training set members

(3) Neural networks for facial recognition

Recover recognizable images of training set members

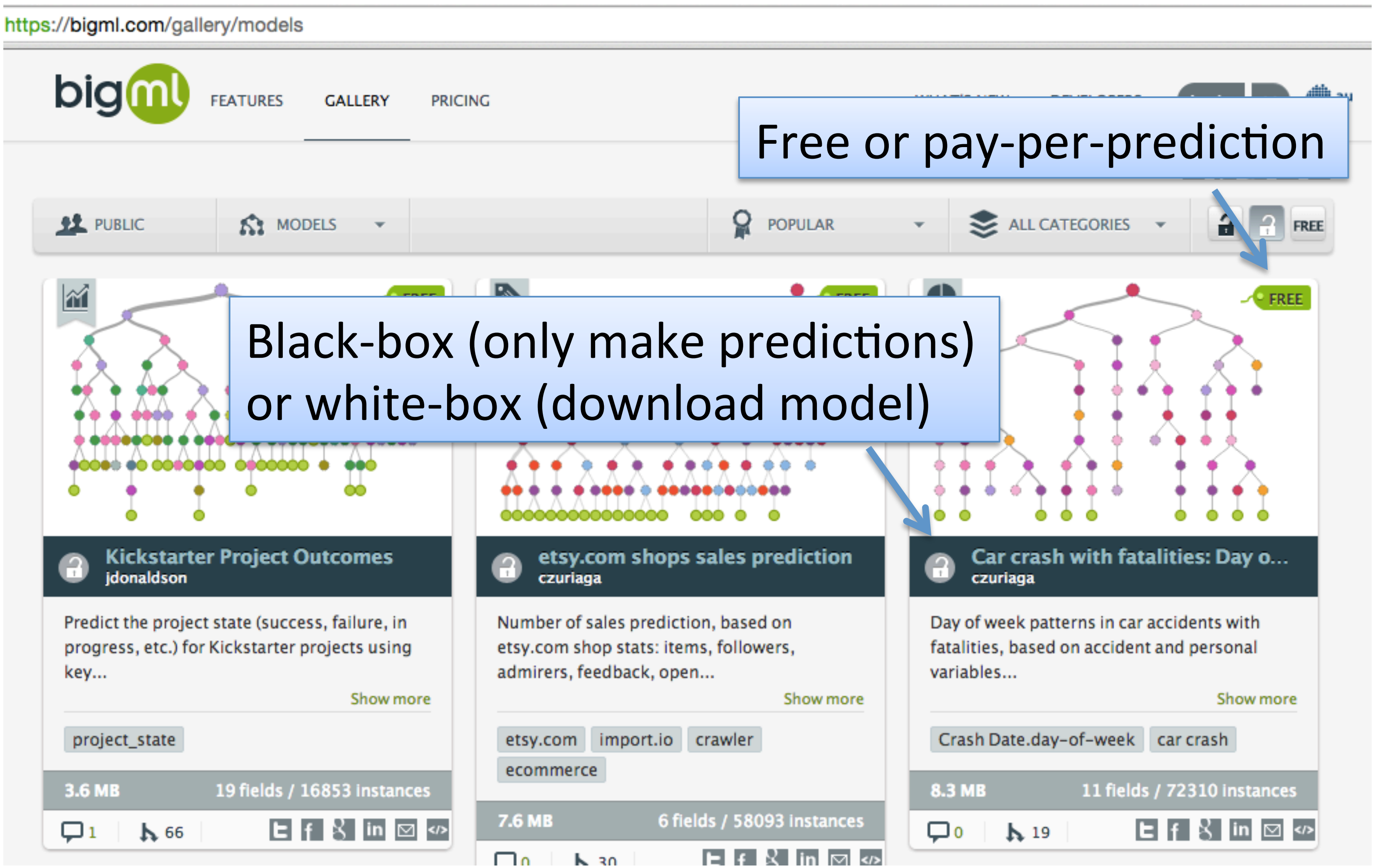
Preliminary investigation of countermeasures

Differential privacy

Sensitive-feature-aware CART

Rounded confidence values

ML-as-a-service APIs

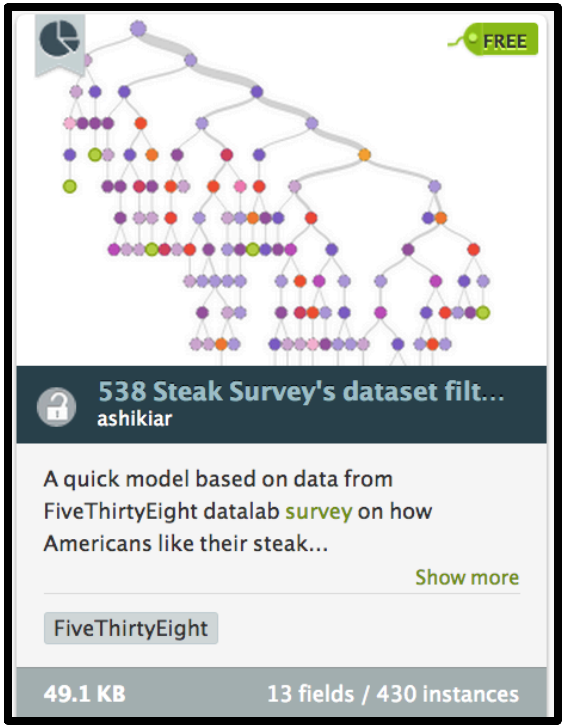


Free or pay-per-prediction

Black-box (only make predictions) or white-box (download model)

Sensitive decision tree models

538 steak survey
GSS marital happiness study (see paper)



Survey of 332 people to determine if "risky" lifestyle choices correlates with steak preferences

$$f\,(\,x_1\,,\,\ldots\boxed{,\,x_n\,)}\;=\;y$$

Household income
Whether person gambles
Whether cheated on significant other
…

Prediction of how person likes steak prepared:
- rare
- medium-rare
- medium
- medium-well
- well-done

De-identified training dataset available, we use to simulate attacks

Black-box warfarin-like attack for 538 survey

Given:

$x_1 , \ldots , x_{n-1}$
Actual steak preference *y'*
Marginal priors, queries to *f*
Confusion matrix **C** for *f*

Model inversion algorithm

Predict:

Infidelity status x_n

$\mathbf{C}_{y',y}$ = # training instances w/ steak type y' predicted as y

Simple black-box MAP estimator (like the warfarin one):

$$\arg\max_{x_n} \frac{\mathbf{C}_{y',f(x_1,\ldots,x_n)}}{\sum_{l\in Y}\mathbf{C}_{y',l}} \cdot \Pr[\,x_n\,]$$

Black-box warfarin-like attack for 538 survey

<u>Given:</u>

x_1 , … , x_{n-1}
Actual steak preference *y'*
Marginal priors, queries to *f*
Confusion matrix **C** for *f*

Model inversion algorithm

<u>Predict:</u>

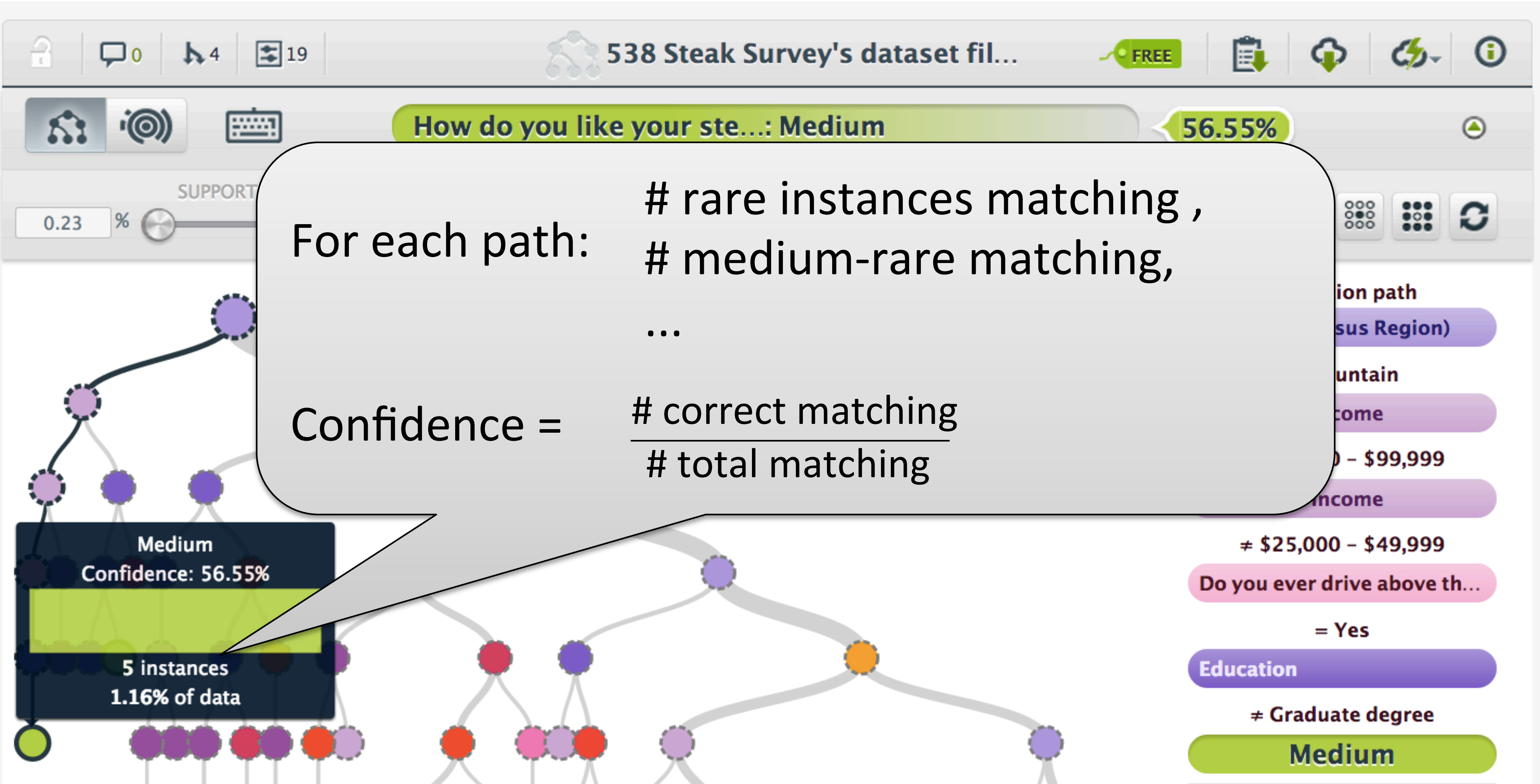
Infidelity status x_n

$\mathbf{C}_{y',y}$ = # training instances w/ steak type y' predicted as y

Performance:

	Accuracy	Precision	Recall
Baseline guessing	82.9%	0.0%	0.0%
MI attack	85.8%	85.7%	21.1%

BigML reveals confidence values



For each path: # rare instances matching ,
medium-rare matching,
...

Confidence = $\dfrac{\text{\# correct matching}}{\text{\# total matching}}$

New MI attack using granular confidence data

<u>Given:</u>

x_1 , … , x_{n-1}
Actual steak preference *y'*
Marginal priors, queries to *f*
Confusion matrix **C** for *f*
Path counts

New model
inversion algorithm

<u>Predict:</u>

Infidelity status x_n

$\mathbf{C}_{y',y}$ = # training instances w/ steak type y' predicted as y

	Accuracy	Precision	Recall
Baseline guessing	82.9%	0.0%	0.0%
MI attack	85.8%	85.7%	21.1%
MI attack w/ confidences	86.4%	**100%**	21.1%

New privacy concerns in ML

Model inversion attacks:

(1) Linear regression for personalized medicine

Predict genotypes of patients

(2) Decision trees trained from lifestyle surveys

Predict marital infidelity of training set members

(3) Neural networks for facial recognition

Recover recognizable images of training set members

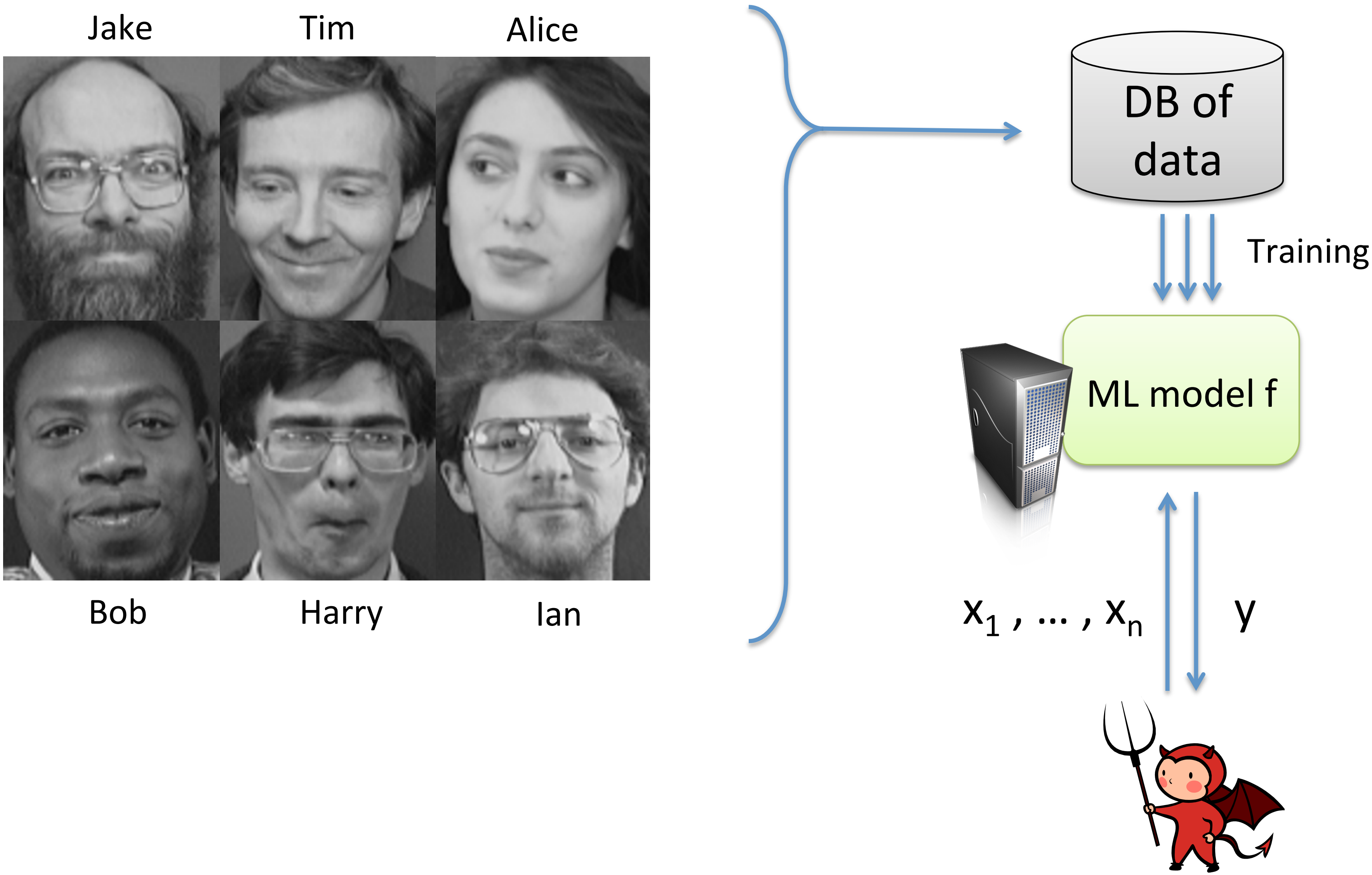
Preliminary investigation of countermeasures

Differential privacy

Sensitive-feature-aware CART

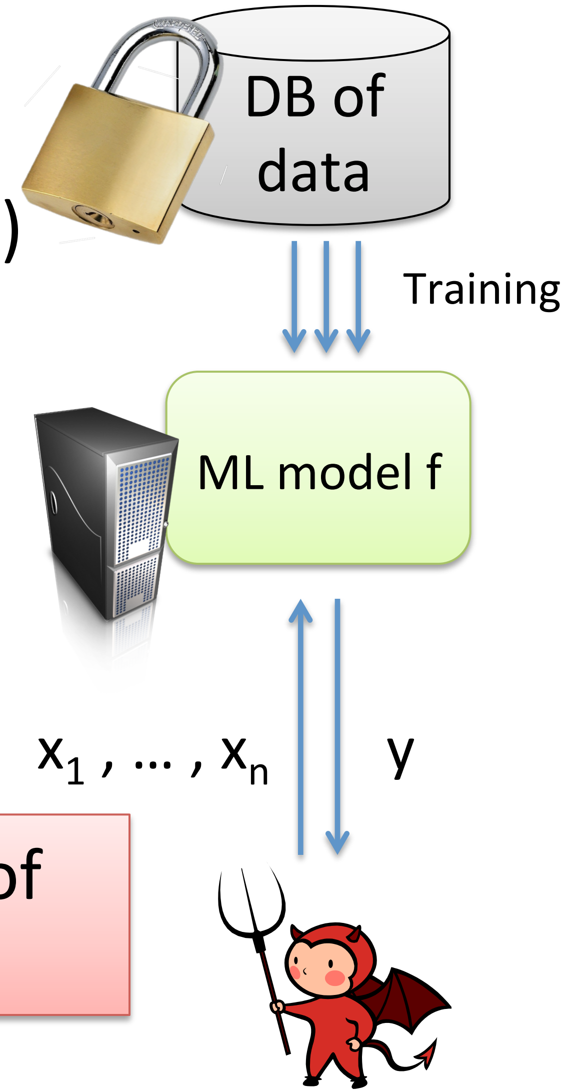
Rounded confidence values

Model inversion for facial recognition

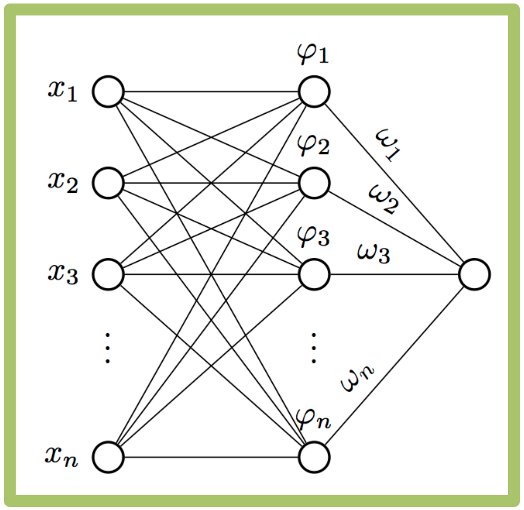


Model inversion for facial recognition

Softmax
Multi-layer perceptron (MLP)
Stacked de-noising auto-encoder (DAE)



Pixel data



Prediction

Can attacker use f to recover images of training member's faces?

Taking advantage of confidence values

$$f(\boxed{x_1, \ldots, x_n}) = [y_{Bob}, \ldots, y_{Jake}]$$

Unknown pixel data

Vector of class confidences each in [0,1]
Output label of highest confidence class

AT&T faces dataset:

$n = 92 * 112 = 10{,}304$

$|x_i| = 8$ bits (grayscale intensity value)

$8^{10,304}$ possible images

Naïve brute-force search won't work

Taking advantage of confidence values

$$f\left(\boxed{x_1 , \ldots , x_n}\right) = [y_{Bob} , \ldots , y_{Jake}]$$

Unknown pixel data

Vector of class confidences each in [0,1]
Output label of highest confidence class

<u>Insight:</u>

confidences allows efficient gradient descent-based search

Find $x_1,\ldots,x_n$ with highest confidence for 'Bob'

Gradient descent:

- White-box we calculate symbolically
- Black-box need to do numerical estimation

Model (trained on AT&T faces)	**Local white-box time (seconds)**
Softmax	1
Multi-layer perceptron	1,298
Denoising autoencoder	692

Example outputs of MI attack for different models

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

Inversion for three neural-network classifiers :

Softmax, Multi-layer perceptron, De-noising auto-encoder

Trained on AT&T faces dataset (40 individuals, 400 images)

Recognizability?

Amazon Mechanical Turk to evaluate image reconstruction recognizability

The image on the left is a face that was altered by computer processing. It may or may not correspond to one of the faces displayed to the right of it.

If you believe that it does correspond to one of the other faces, please select the corresponding image. If you do not believe that it corresponds to one of the other faces, select "Not Present".

Altered Image

Not Present

Re-identification accuracy up to 95% for skilled workers

New privacy concerns in ML

Model inversion attacks:

(1) Linear regression for personalized medicine

Predict genotypes of patients

(2) Decision trees trained from lifestyle surveys

Predict marital infidelity of training set members

(3) Neural networks for facial recognition

Recover recognizable images of training set members

Preliminary investigation of countermeasures

Differential privacy

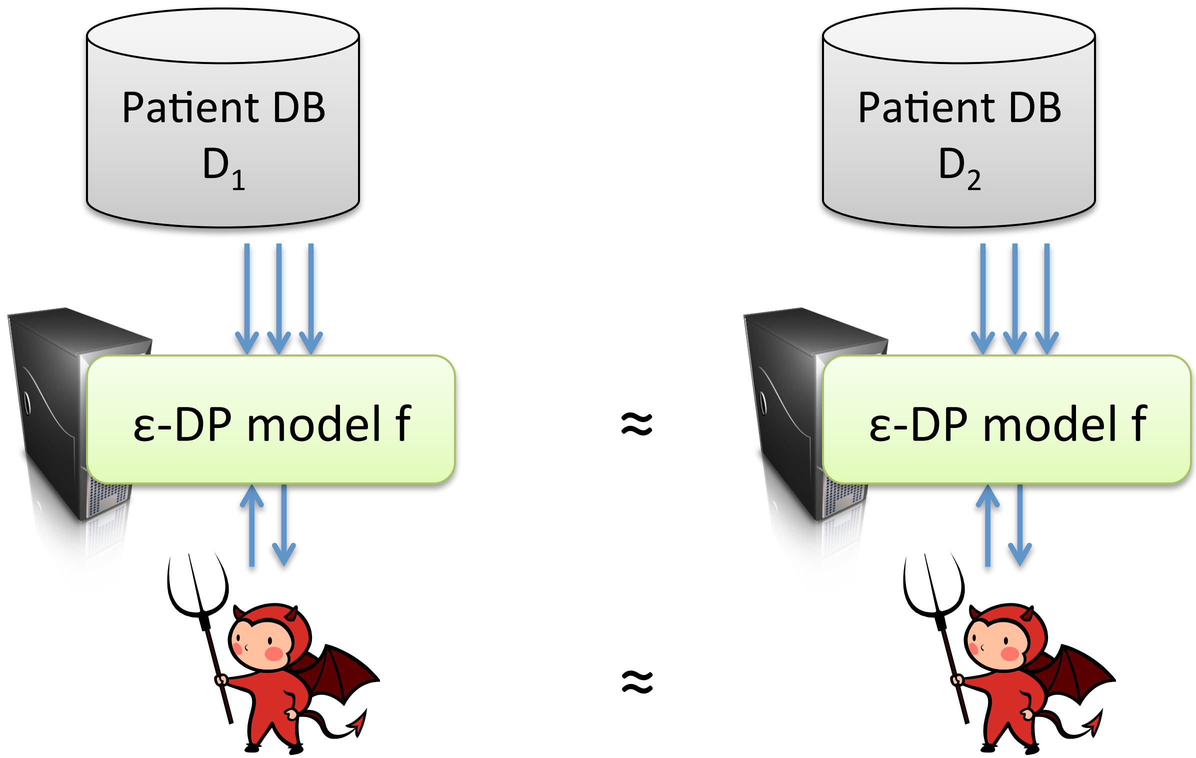
Sensitive-feature-aware CART

Rounded confidence values

Differential privacy

[Dwork, McSherry, Nissim, Smith '06]

Given model f adversary can't learn whether any single individual contributed to training data set



Inversion success: Can't vary by > e^{ε} for dataset with or w/o individual

Guarantees nothing about absolute success

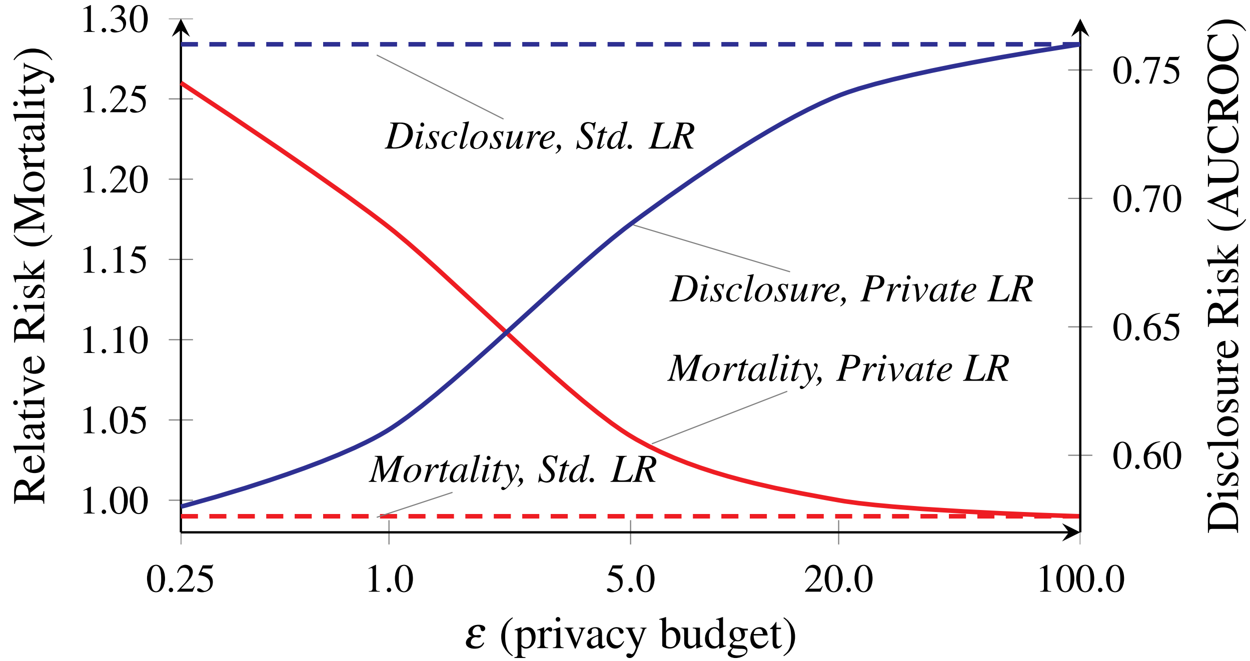
End-to-end analysis of DP in warfarin case

Differentially private version of model hides whether individual contributed to training data set with efficacy a function of privacy budget ε

[Zhang et al.] functional mechanism for private linear regression

We performed end-to-end case study:

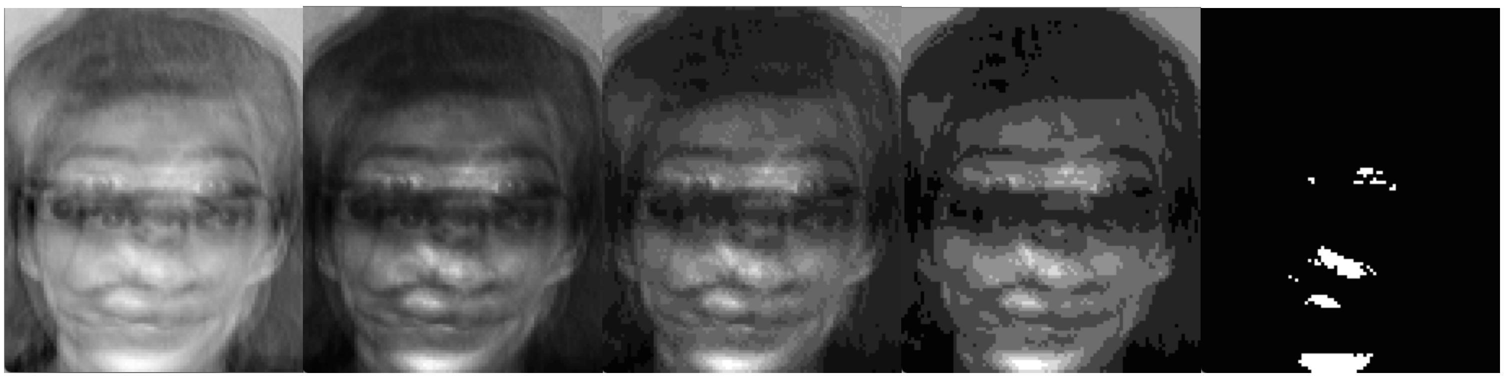
- Evaluate model inversion disclosure risk for DP models
- Use simulated clinical trials to evaluate utility of DP models



Other simple countermeasures?

Attacks that rely on confidence data: degrade it

Our MI attack against softmax with rounded confidences:



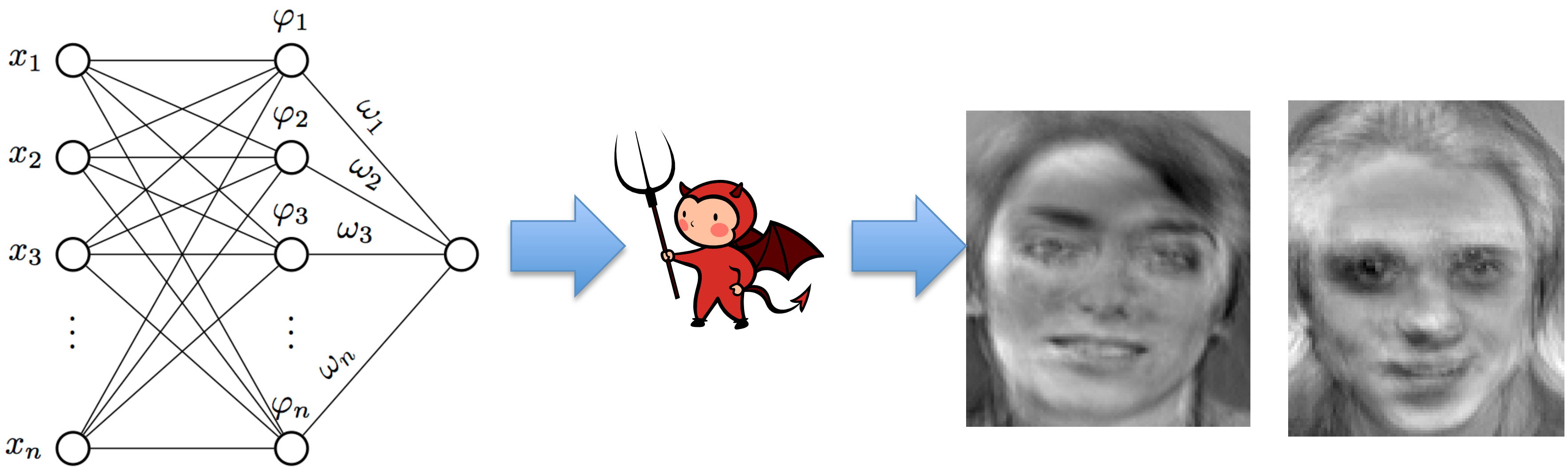
no rounding | $r = 0.001$ | $r = 0.005$ | $r = 0.01$ | $r = 0.05$

Rounding confidence values to nearest r

Sensitive-feature-aware CART decision tree training (see paper)

Model inversion and ML privacy

Adversarial access to models has subtle implications



Open questions: better attacks, handling more sophisticated ML models, principled countermeasures

Exploiting Leakage
in Searchable Encryption and Machine Learning

Tom Ristenpart

CORNELL TECH

Covering joint work with:
David Cash, Paul Grubbs, Jason Perry (Searchable encryption)
Matthew Fredrikson, Eric Lantz, Simon Lin, David Page, Somesh Jha (ML)