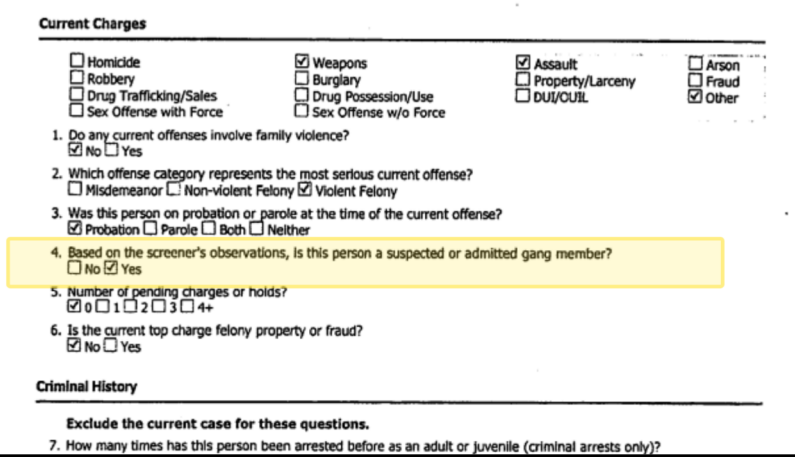
Fairness and Privacy in Machine Learning

Mark Craven and David Page
Computer Sciences 760
Spring 2019

The COMPAS system

- used by many governments (including state of Wisconsin) to predict risk that those convicted of crimes will commit future crimes
- scores derived from 137 questions that are either answered by defendants or pulled from criminal records.



Current Charges

☐ Homicide ☐ Robbery ☐ Drug Trafficking/Sales ☐ Sex Offense with Force
☑ Weapons ☐ Burglary ☐ Drug Possession/Use ☐ Sex Offense w/o Force
☑ Assault ☐ Property/Larceny ☐ DUI/OUIL
☐ Arson ☐ Fraud ☑ Other

1. Do any current offenses involve family violence?
☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
☐ No ☑ Yes
5. Number of pending charges or holds?
☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
☑ No ☐ Yes

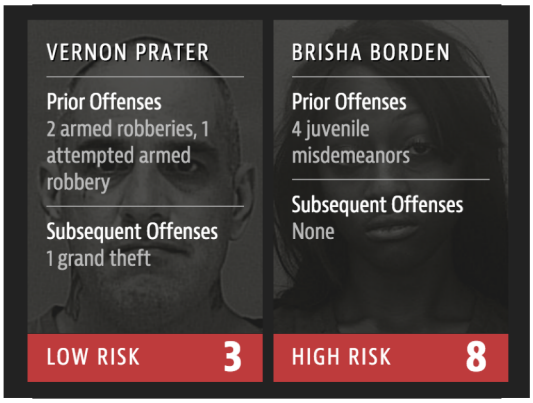
Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?

The COMPAS system

- ProPublica obtained the risk scores assigned to > 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over next 2 years



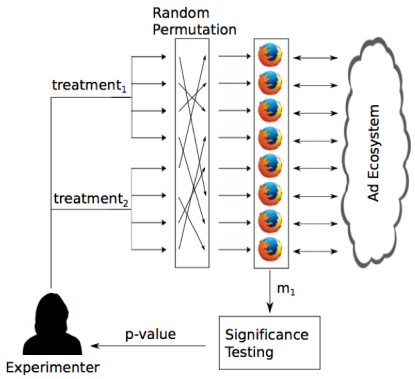
PROPUBLICA

The COMPAS system

- ProPublica obtained the risk scores assigned to > 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over next 2 years
- The system was particularly likely to falsely flag black defendants as future criminals
  - wrongly labeling them this way at almost twice the rate as white defendants
  - white defendants were mislabeled as low risk more often than black defendants

Google Ads Settings

- Datta et al. [PPET 2015] studied how user behaviors, Google's ads, and Ad Settings interact
- Setting gender to female in Google Ad Settings made it less likely that user would be shown ads for high paying jobs



Isn't discrimination the point of machine learning?

Yes, but we should be aware of

- unjustified bases for discrimination
- legal reasons to avoid unjust discrimination
- moral reasons to avoid unjust discrimination

Certain domains are legally regulated

- credit, education, employment, housing, public accommodation

Certain classes are legally protected in specific contexts

- race, color, sex, religion, national origin, citizenship, age, pregnancy, familial status, disability status, veteran status, genetic information

See http://mrtz.org/nips17/ for more detail

How does unfair bias arise in machine learning systems?

- selection, sampling, reporting bias in the data set
- bias in the objective function

Biases in data sets example #1

Garg et al. [*PNAS* 2017] "Word embeddings quantify 100 years of gender and ethnic stereotypes"

- tested relationships among concepts in Google word2vec vectors
- e.g. relatedness of occupations and words representing gender

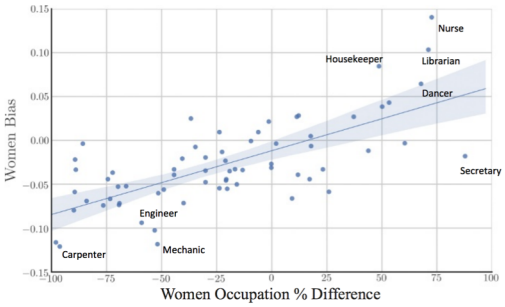


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

Biases in data sets example #1

Garg et al. [*PNAS* 2017] "Word embeddings quantify 100 years of gender and ethnic stereotypes"

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

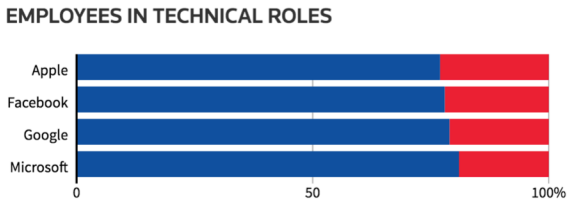
1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Biases in data sets example #2

BUSINESS NEWS OCTOBER 9, 2018 / 10:12 PM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ



How to achieve fairness in ML

1. Blindness approach: don't use features that enable unfair classifications/predictions
   - this approach is generally not effective; the data usually contains many surrogates for such protected features
   - e.g. the COMPAS system does not explicitly use race
   - e.g. word embeddings case illustrates a lot of dependence between gender words and other words

How to achieve fairness in ML

2. Group fairness approach
   - given two groups, G_1 and G_2
   - enforce that $P(Outcome = o \mid G_1) \approx P(Outcome = o \mid G_2)$ by incorporating such a term into objective function

How to achieve fairness in ML

3. Individual fairness approach
   - treat similar individuals similarly
   - $\mathrm{f}(\boldsymbol{x}^{(i)}) \approx \mathrm{f}(\boldsymbol{x}^{(j)}) \mid d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \approx 0$
   - where $d: X \times X \to \mathbb{R}$ is a distance metric for individuals

An individual fairness approach

[Dwork et al. ITCS 2012]

- model outputs a probability distribution over set of outcomes $P(y \mid \boldsymbol{x})$
- the notion of individual fairness can be captured by a (D, d)-Lipschitz property

  $$D\big(P(y \mid \boldsymbol{x}^{(i)}),\ P(y \mid \boldsymbol{x}^{(j)})\big) \leq d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$$

  where D is a distance measure for distributions
- learning is then a constrained optimization problem

An individual fairness approach

[Dwork et al. ITCS 2012]

- model outputs a probability distribution over set of outcomes $P(y \mid \boldsymbol{x})$
- the notion of individual fairness can be captured by a (D, d)-Lipschitz property

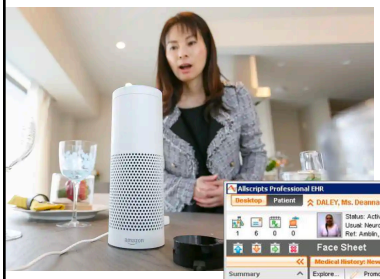
$$D\big(P(y \mid \boldsymbol{x}^{(i)}),\ P(y \mid \boldsymbol{x}^{(j)})\big) \leq d\big(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\big)$$

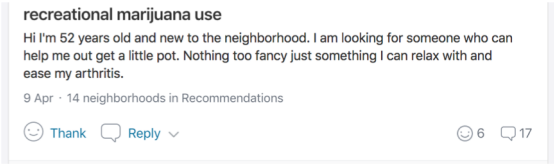
  where D is a distance measure for distributions

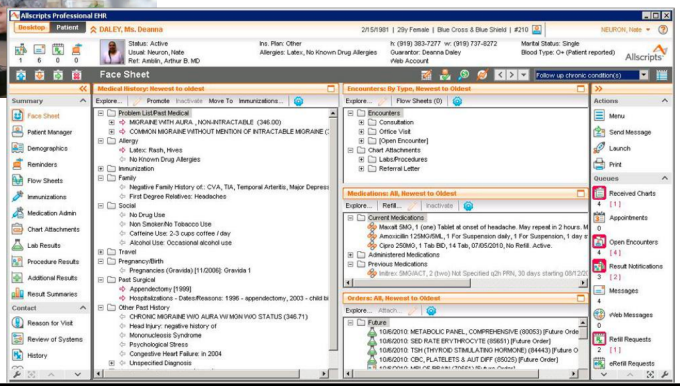
- learning is then a constrained optimization problem

Privacy in machine learning

- in many applications training data contains sensitive information about individuals

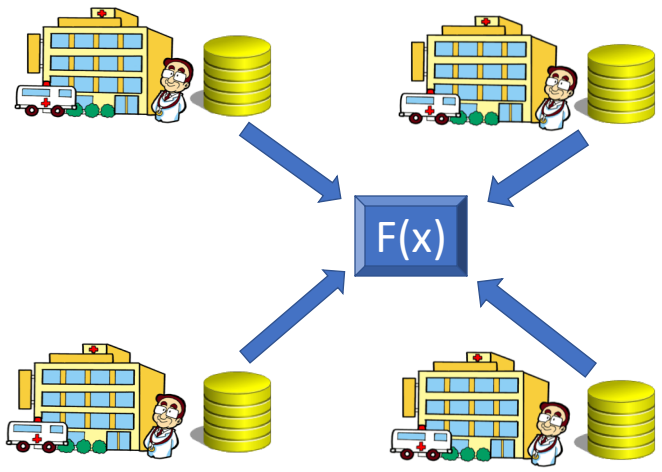




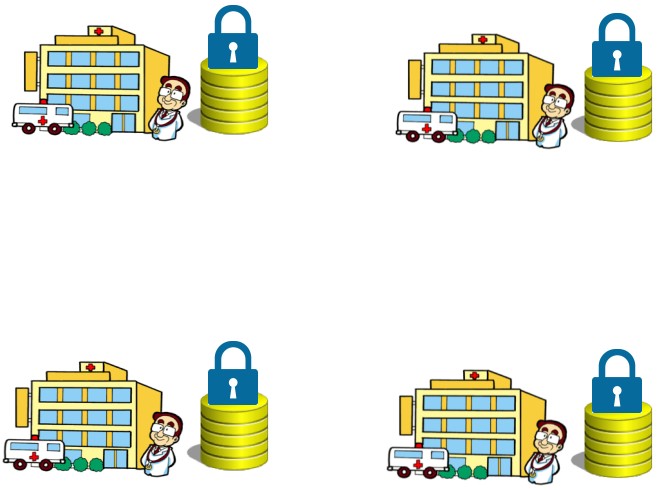


Need for privacy with health data

- large databases of patient information
  - regulations and expectations of privacy
  - large potential gains from data mining
  - How to balance utility and privacy?
- privacy approaches
  - *k*-anonymity (Sweeney, 2002), *l*-diversity (Machanavajjhala, 2007), *t*-closeness (Li, 2007)
  - homomorphic encryption
  - differential privacy (Dwork, 2006)



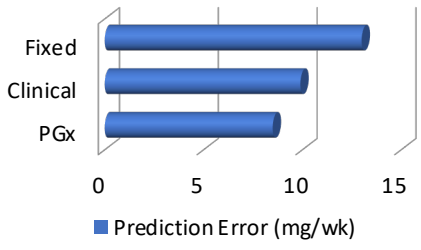
18



19

Recall: IWPC Warfarin dosing algorithm

- Over a dozen real-value prediction techniques were used
- Linear regression and support vector regression were the best performers



5.6044
-0.2614 Age in decades
+0.0087 Height in cm
+0.0128 Weight in kg
-0.8677 *VKORC1* A/G
-1.6974 *VKORC1* A/A
-0.4854 *VKORC1* genotype unknown
-0.5211 *CYP2C9* *1/*2
-0.9357 *CYP2C9* *1/*3
-1.0616 *CYP2C9* *2/*2
-1.9206 *CYP2C9* *2/*3
-2.3312 *CYP2C9* *3/*3
-0.2188 *CYP2C9* genotype unknown
-0.1092 Asian race
-0.2760 Black or African American
-0.1032 Missing or Mixed race
+1.1816 Enzyme inducer status
-0.5503 Amiodarone status
= square root of final dose

20

Recall: Ridge Regression

Data point: $(\mathbf{x}, y)$, $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$

Model: $\mathbf{w} \in \mathbb{R}^d$ vector of weights

$$y \approx f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \sum_{j=1}^{d} \mathbf{w}(j)\mathbf{x}(j)$$

Training: find argmin of $F(\mathbf{w}) = \underbrace{\sum_{i=1}^{n} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2}_{\text{mean squared error}} + \lambda \underbrace{||\mathbf{w}||_2^2}_{\text{regularization}}$

Public-Key Encryption

$sk \rightarrow$ secret key
$pk \rightarrow$ public key

Encryption:

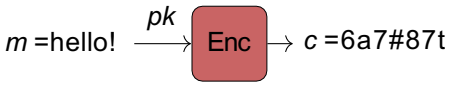
Decryption:

Public-Key Encryption

sk → secret key
pk → public key

Encryption: $\mathbf{c} = \text{Enc}_{pk}(\mathbf{m})$
c → hides **m** to everyone that does NOT have *sk*

Decryption:



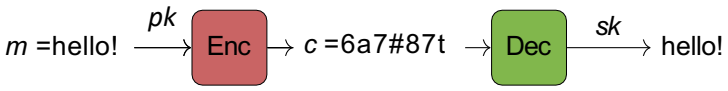
Public-Key Encryption

sk → secret key
pk → public key

Encryption: $\mathbf{c} = \text{Enc}_{pk}(\mathbf{m})$
c → hides **m** to everyone that does NOT have *sk*

Decryption:

c → reveals **m** to everyone that has *sk*



Linearly-Homomorphic Encryption

Addition of ciphertexts

$Enc_{pk}(\mathbf{m}_1) \boxplus Enc_{pk}(\mathbf{m}_2) = Enc_{pk}(\mathbf{m}_1 + \mathbf{m}_2)$

Multiplication of a ciphertext by a plaintext

$\mathbf{m}_1 \boxplus Enc_{pk}(\mathbf{m}_2) = Enc_{pk}(\mathbf{m}_1 \times \mathbf{m}_2)$

Linearly-Homomorphic Encryption

Addition of ciphertexts

$Enc_{pk}(\mathbf{m}_1) \boxplus Enc_{pk}(\mathbf{m}_2) = Enc_{pk}(\mathbf{m}_1 + \mathbf{m}_2)$

Multiplication of a ciphertext by a plaintext ($\mathbf{m}_1$ is public!)

$\mathbf{M}_1 \boxplus Enc_{pk}(\mathbf{m}_2) = Enc_{pk}(\mathbf{m}_1 \times \mathbf{m}_2)$

Linearly-Homomorphic Encryption

Addition of ciphertexts

$\mathrm{Enc}_{pk}(\mathbf{m}_1) \boxplus \mathrm{Enc}_{pk}(\mathbf{m}_2) = \mathrm{Enc}_{pk}(\mathbf{m}_1 + \mathbf{m}_2)$

Multiplication of a ciphertext by a plaintext ($\mathbf{m}_1$ is public)

$\mathbf{M}_1 \boxplus \mathrm{Enc}_{pk}(\mathbf{m}_2) = \mathrm{Enc}_{pk}(\mathbf{m}_1 \times \mathbf{m}_2)$

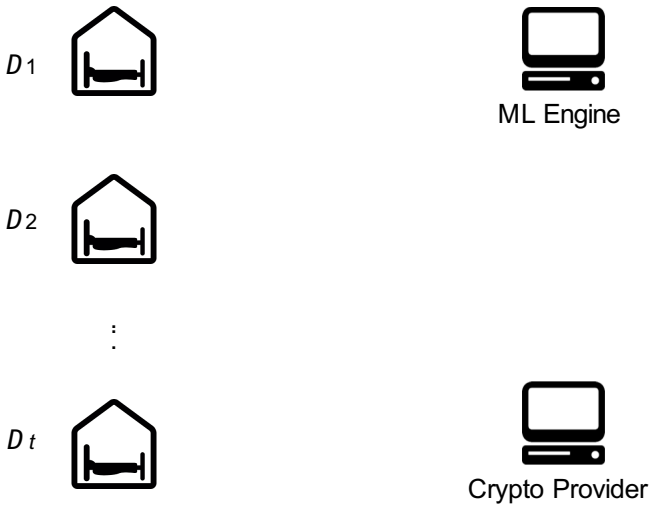
Fully homomorphic requires *multiplication* analog of $\boxplus$

and currently is ***much*** slower.

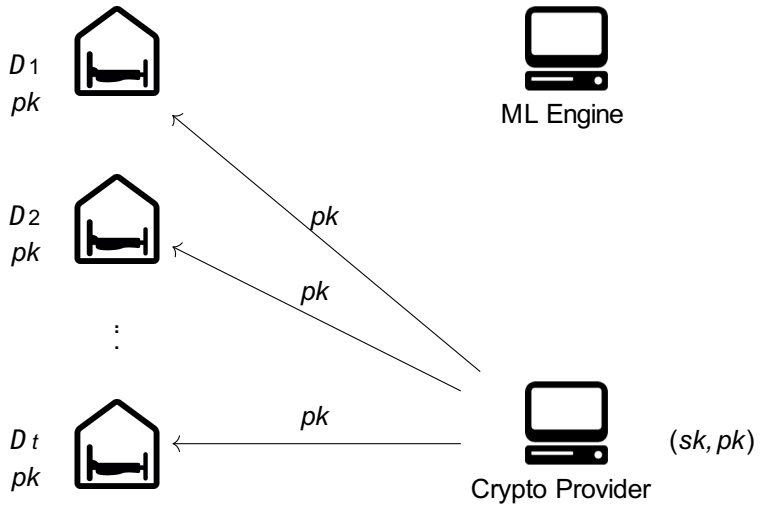
Database (DB): $10^5 \times 10^2$ real numbers in [−2000, 2000] with 3 digits in the fractional part. Times using linearly-homomorphic encryption:

- encrypt the DB: 40 minutes
- sum of two DBs: 3 seconds
- mult. by a constant: 25 mins

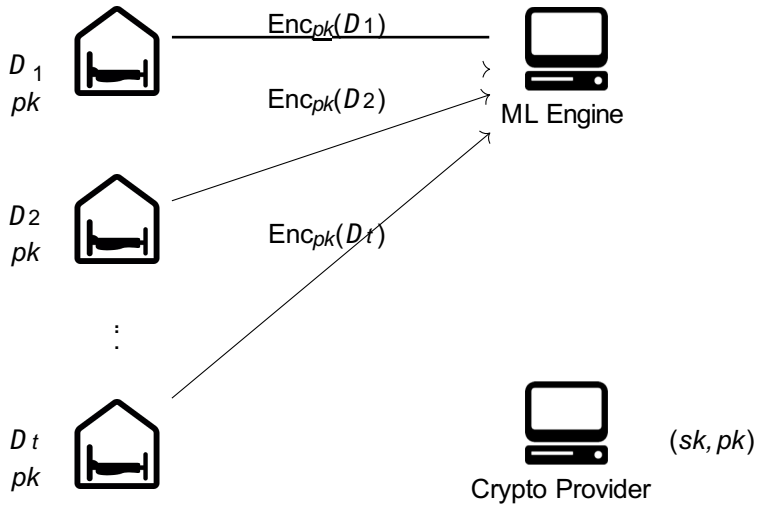
Illustration



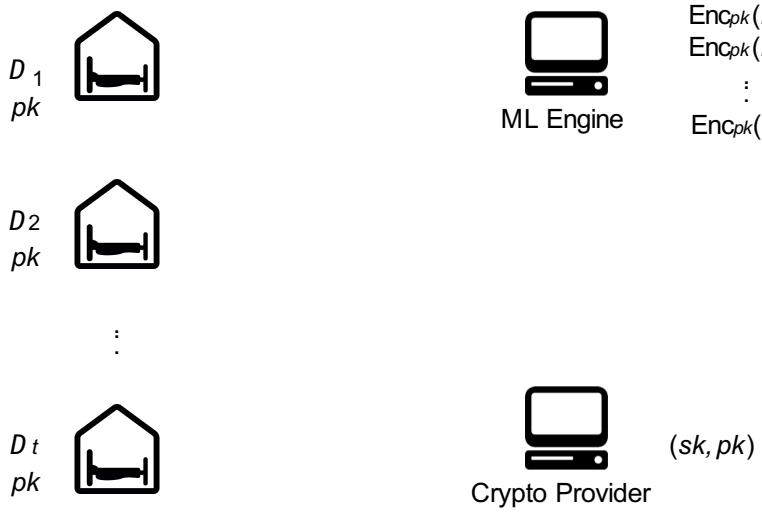
Illustration



Illustration

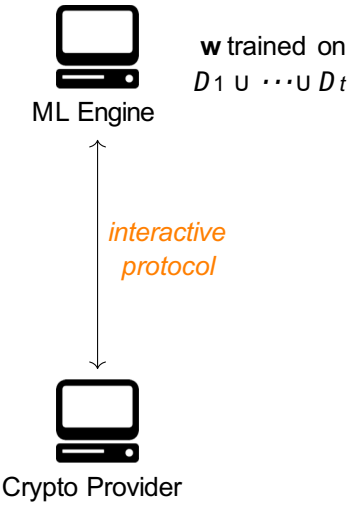


Illustration



11/ 29

Illustration



11/ 29

Illustration

Interactive protocol:

1. the ML engine "masks inside the encryption"
   $\mathrm{Enc}_{pk}(D) \rightarrow \mathrm{Enc}_{pk}(\tilde{D})$
2. the crypto provider decrypts, gets $\tilde{D}$ and computes a "masked model", $\tilde{\mathbf{w}}$
3. the ML engine computes the real model **w** from the masked one

Results [Giacomelli et al., *ACNS* 2018]

Results for seven UCI datasets (time in seconds):
(phase 1 = encryption, phase 2 = interactive protocol)

Dataset	n	d	ℓ	$\log_2(N)$	R_{MSE}	Phase 1		Phase 2	
						Time	kB	Time	kB
air	6252	13	1	2048	4.15E-09	1.99	53.24	3.65	96.51
beijing	37582	14	2	2048	5.29E-07	2.37	60.93	4.26	110.10
boston	456	13	4	2048	2.34E-06	2.00	53.24	3.76	96.51
energy	17762	25	3	2724	5.63E-07	12.99	238.26	37.73	451
forest	466	12	3	2048	3.57E-09	1.66	46.08	2.81	82.94
student	356	30	1	2048	4.63E-07	9.36	253.44	30.40	483.84
wine	4409	11	4	2048	2.62E-05	1.71	39.42	2.38	70.40

n = training data (number of data points)
d = number of features

Comments on Homomorphic Encryption

- Benefits
  - High utility – because No Noise!!!
  - No one sees data "in the clear"
- Disadvantages
  - Models (or even just predictions) may still give away more information about training examples (e.g., patients) than about other examples (patients)
  - Very high (as of now, completely impractical) runtimes for some methods (fully homomorphic encryption)
  - Feasible approaches (e.g., linearly homomorphic encryption) require re-developing each learning algorithm (e.g., ridge regression) from scratch with limited operations
  - Protections may be lost if/when Quantum Computers become available

35

Differential Privacy (Dwork, 2006)

- Goal
  - Small added risk of adversary learning (private) information about an individual if his/her data in the private database versus not in the database
- Informally
  - Query output does not change much between neighboring databases
  - E.g.: what is fraction of people in clinic with diabetes?

Name	Has Diabetes (X)
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1

36

Differential Privacy Definition

- Given
  - Input database *D*
  - Randomized algorithm *f* : *D* -> *Range*(*f*)
  - *f* is (e , δ)-differentially private iff

$$\Pr(f(D) \in S) \leq e^{\epsilon} \Pr(f(D') \in \mathcal{S}) + \delta$$

- For any *S* ∈ *Range*(*f*) and *D*' where d(*D*,*D*')=1
  - ε and δ are privacy budget
    - Smaller means more private

37

Obtaining Differential Privacy

- Note: Definition requires stochastic output… how to achieve?
- Perturbation {Laplace Mechanism} (Dwork, 2006)
  - Calculate correct answer f(D)
  - Add noise f(D) + η
- Soft-max {Exponential Mechanism} (McSherry and Talwar, 2007)
  - Quality function q(D,s)
  - Exponential weighting exp(e q(D,s))
- In both cases, noise is proportional to the *sensitivity* of the function

38

Global Sensitivity

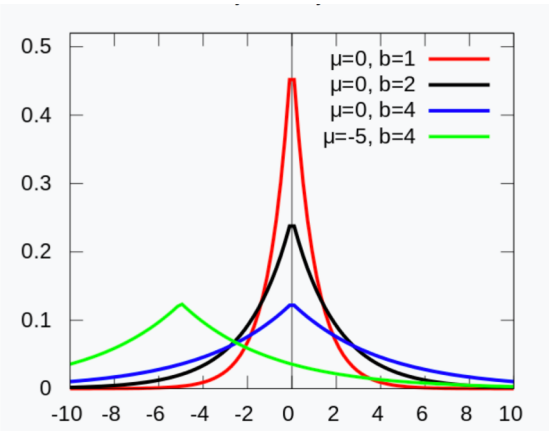
- Given $f : D \rightarrow R$, global sensitivity of f is

$$GS_f = \max_{d(D,D')=1} |f(D) - f(D')|$$

- Worst case
- Once f and the domain of D are chosen, global sensitivity is fixed

39

Add Laplace Noise, μ=0, *b* a function of sensitivity and ε



$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

40

Comments on Differential Privacy

- Provable guarantees, regardless of side information adversary has
- Elegant formulation that leads to many attractive algorithms
- Has insights for other areas such as fairness
- Poor intuition for how to select ε
- Can kill utility (e.g., accuracy, AUC) unless we have very many examples… so good fit for age of Big Data but not for medium data
- How to set privacy budget? If release DP dataset, can update with new release without adding to previous ε, so must plan far ahead