Privacy Techniques for Data Science
Or
Yes, Differential Privacy Is Useful
•••

DataEng Conference

Jim Klucar, Immuta
@jimmuta

10/30/17

Data Management for Data Science

Connect, Control, Comply, Accelerate

The Law is Coming For Your Data

“*We can math our way around this*”

William Shakespeare, *H*[redacted]

Jim Klucar, just now.

Three Data Privacy Scenarios

Release

Collect

Interact

K-Anonymization

“proc ... the

indiv

re-id

Sween

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

K-Anonymization Extras

L-Diversity

Ensure ample diversity in k-groups.

T-Closeness

Ensure Statistical Distribution of data in k-groups represents overall statistics of data set.

There are many attack vectors available to de-identify released data

Linkage, Temporal, Complementary Release

Releasing Data is Risky

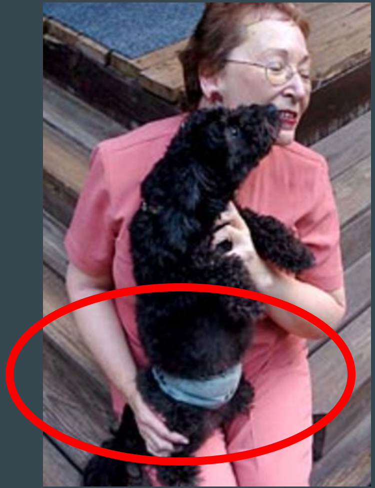
3 Months, 20 Million Terms, 650,000 users

AOL No. 4417749 Thelma Arnold

62 y.o widow in Georgia

“Numb fingers”

“dog that urinates on everything.”



CNNMoney Ranked this #57 in segment titled "101 Dumbest Moments in Business.”

$500 Million Class Action Suit

Randomized Response

Collect Sensitive Data Privately

Who is going to give me a red card?

Think of number between 1 and 6

Throw in Red if you picked a 3

Randomized Response

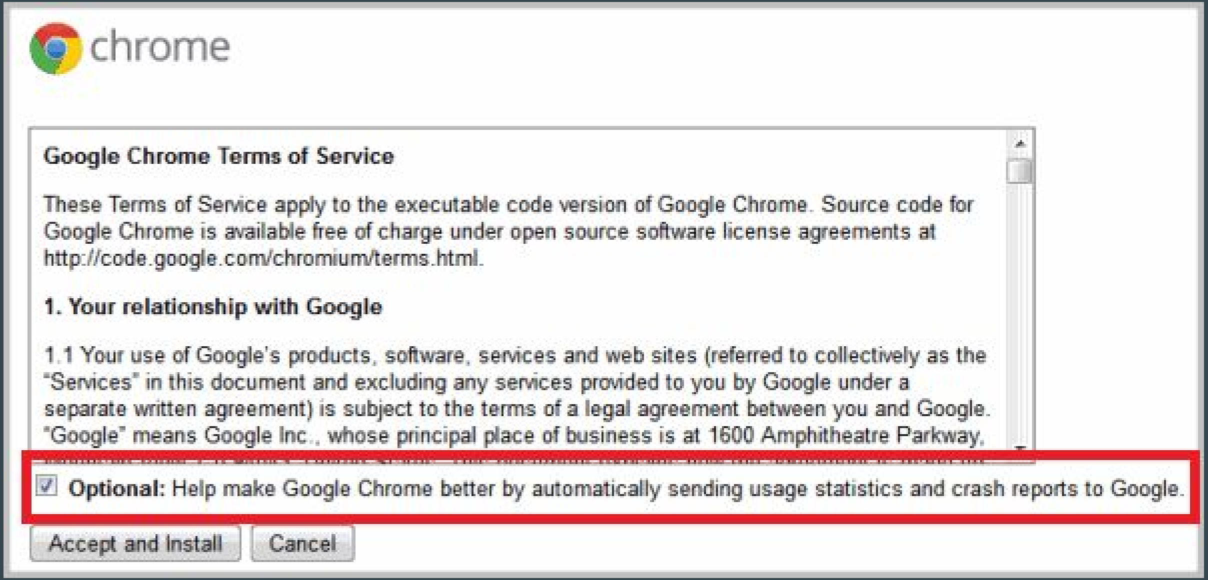
Plausible Deniability

$$\hat{p} = \frac{\mathbf{E}[y] - q}{1 - 2q}$$

p = True proportion

q = ⅙ E[y] = results

Randomized Response



[RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response]

Differential Privacy

‘Differential privacy formalizes the idea that a "private" computation should not reveal whether any one person participated in the input or not, much less what their data are.’ - [Frank McSherry]

(https://github.com/frankmcsherry/blog/blob/master/posts/2016-02-03.md)

Differential Privacy

$$Pr[\mathcal{M}(D_1) \in S] \leq \Pr[\mathcal{M}(D_2) \in S] \cdot e^{\epsilon}$$

Can We Have That In Words Please?

Trades privacy for usability. Aggregate queries only.

Protects against all past, current and future data releases.

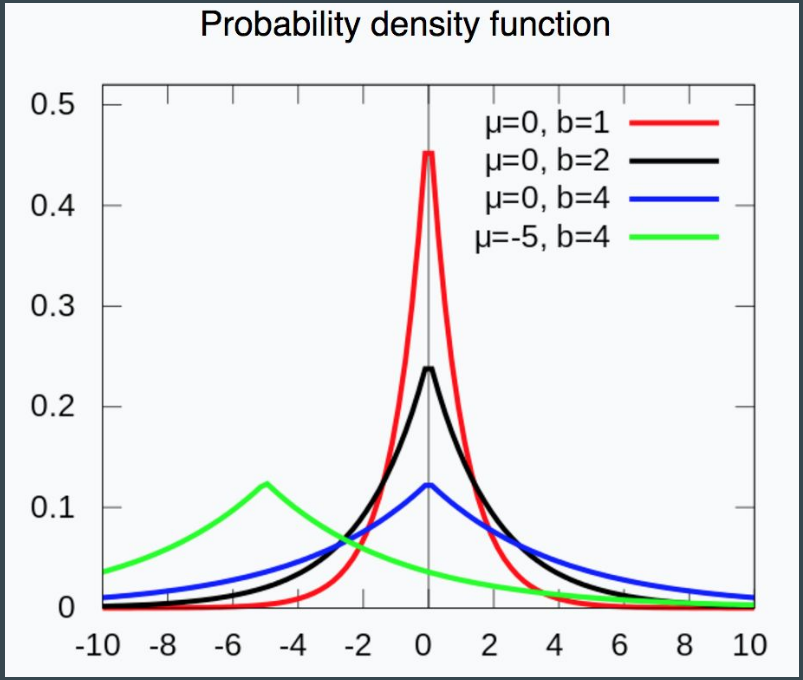
Statistically same result regardless of any single entry in database.

Your data is already a noisy sample of some infinite reality, so what’s more noise?

Laplacian Method

Add Laplacian Noise Proportional to Sensitivity of M

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



Let b = Δm/ε,

random draw from L(0, Δm/ε) as noise

Sensitivity and Robust Statistics

$$\Delta m = \max_{\substack{x,y \in N^{|\mathcal{X}|} \\ \|x-y\|_1 = 1}} \| \mathcal{M}(D_1) - \mathcal{M}(D_2) \|_1$$

\$320k $330k $340k

\$30M

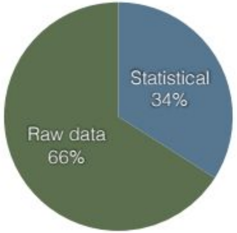
Δm (mean) ~ 30M

Δm (median) ~ 10k

Internal study of queries at Uber:

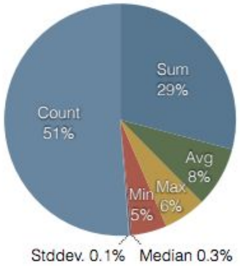
- SQL queries written by employees at Uber
- 8.1 million queries executed between March 2013 and August 2016
- broad range of sensitive data including rider and driver information, trip logs, and customer support data

Question 5: What fraction of queries use aggregations?



Results. Approximately one-third of queries are statistical, meaning they return only aggregations (count, average, etc.). The remaining queries return non-aggregated results (i.e., raw data) in at least one output column.

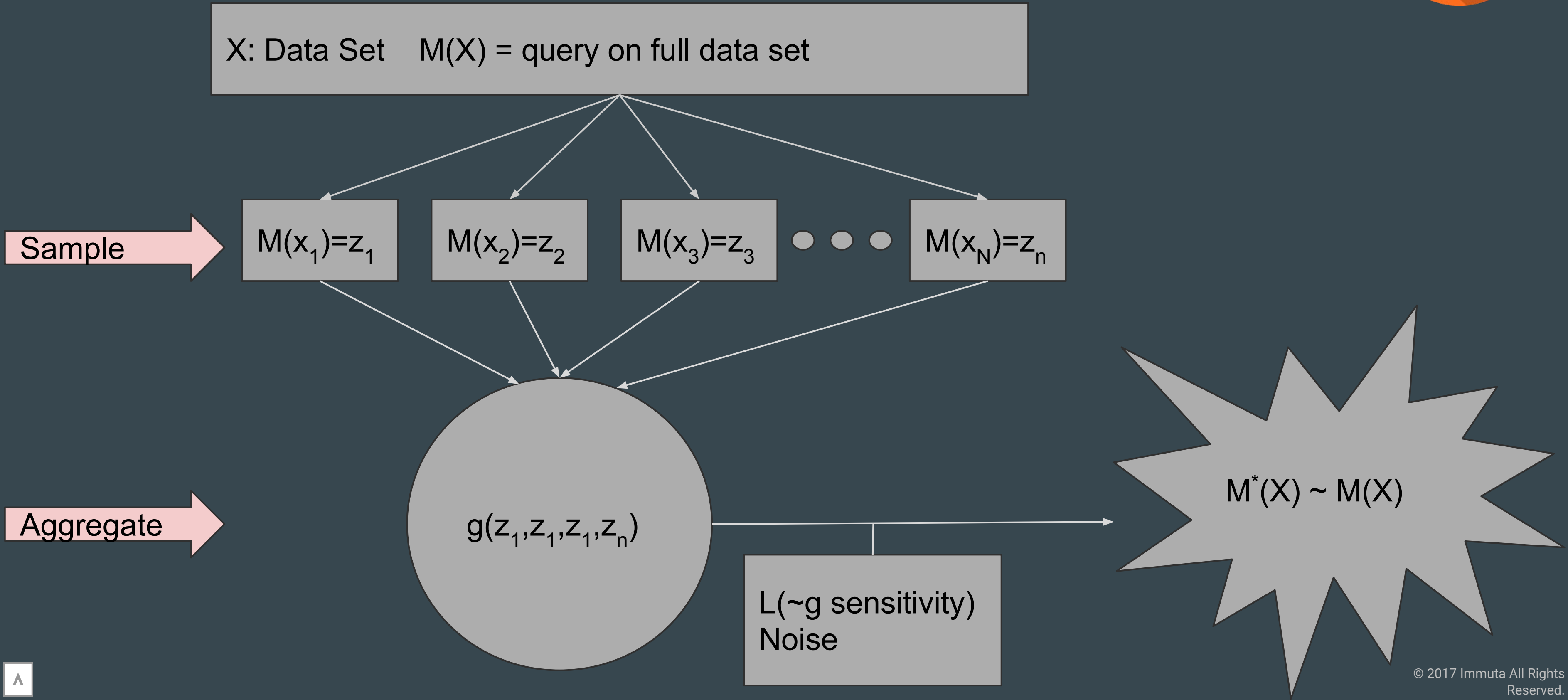
Question 6: Which aggregation functions are used most frequently?



Results. `Count` is the most common aggregation function (51%), followed by `Sum` (29%), `Avg` (8%), `Max` (6%) and `Min` (5%). The remaining functions account for fewer than 1% of all aggregation functions.

Towards Practical Differential Privacy for SQL Queries
Johnson, Near, Song, Aug 2017

Sample and Aggregate, Localize Sensitivity



Let’s Violate Someone’s Privacy *(kinda, not really)*

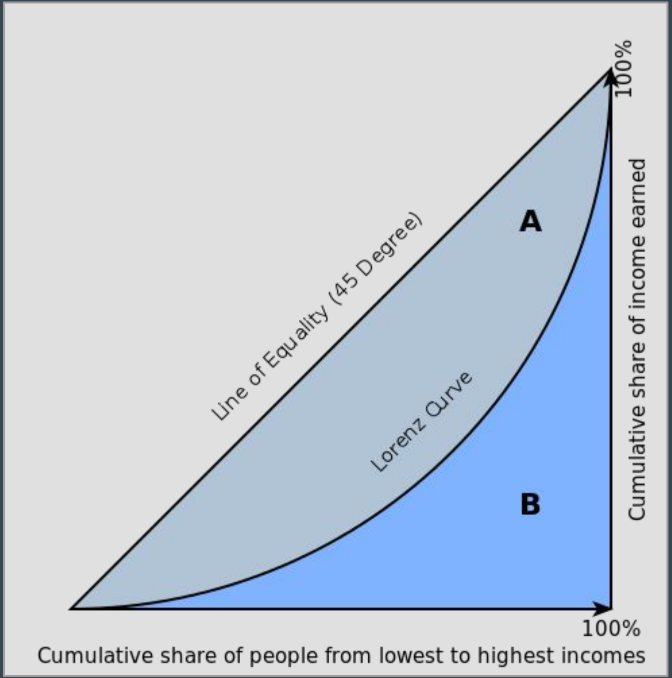
Demo: Simple Average

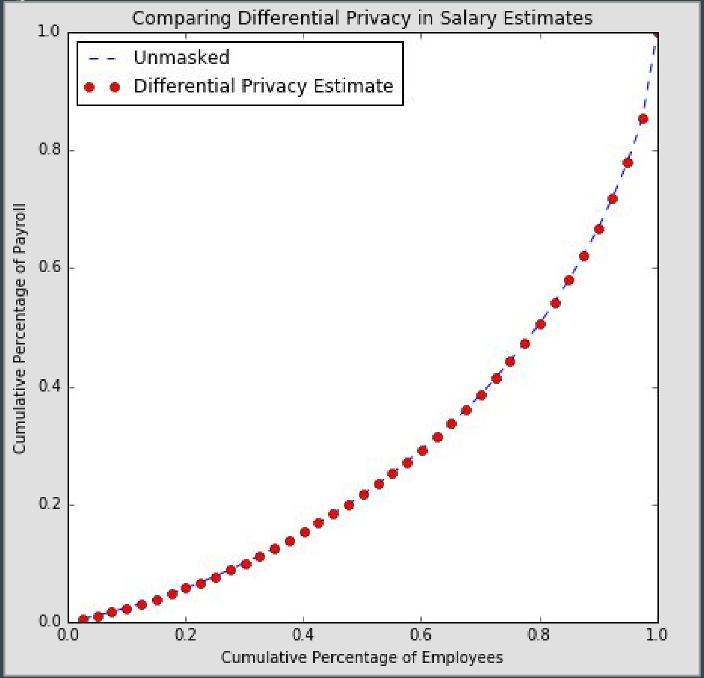
Demo: Too specific query.

Demo: State of Oklahoma Salaries

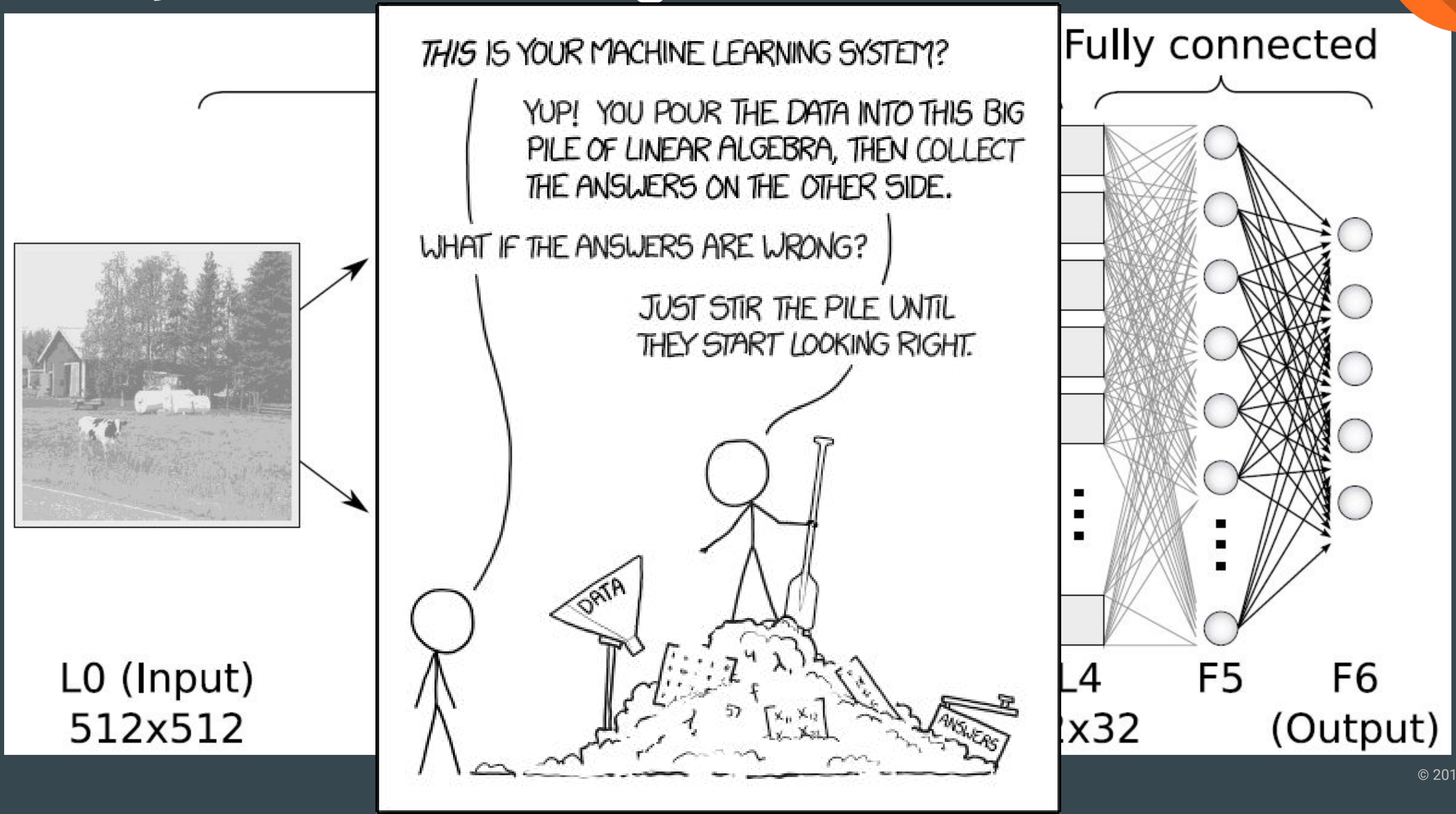
Gini Coefficient

Measures Imbalance of a Distribution, usually wealth



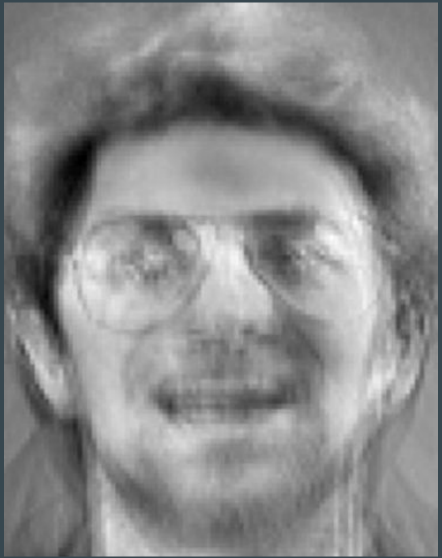


But Surely Machine Learning Is Private...



Machine Learning is Vulnerable Too

Model Inversion Attack - Exploiting Public APIs of SAAS Machine Learning Models



[Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Fredrikson, et al.]

Similar Method For DP Machine Learning

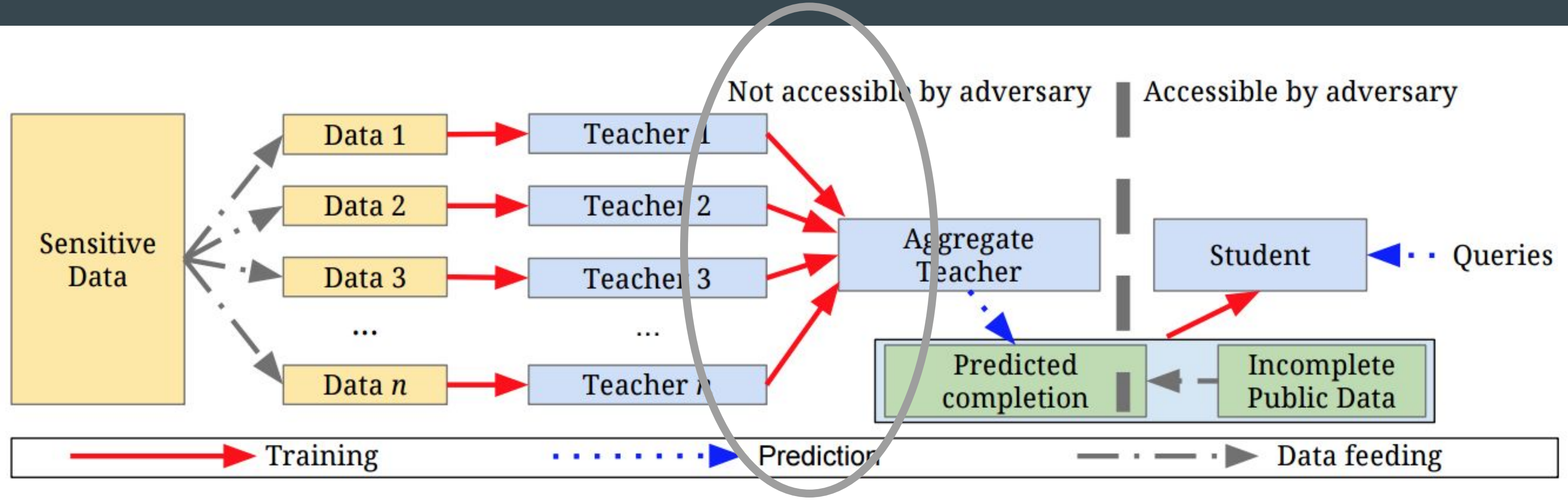


Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

'Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data', Papernot, et al https://arxiv.org/abs/1610.05755

Experimental Results

Dataset	ε	δ	Queries	Non-Private Baseline	Student Accuracy
MNIST	2.04	10^{-5}	100	99.18%	98.00%
MNIST	8.03	10^{-5}	1000	99.18%	98.10%
SVHN	5.04	10^{-6}	500	92.80%	82.72%
SVHN	8.19	10^{-6}	1000	92.80%	90.66%

MNIST = Standard Handwriting Database
SVHN = Street View House Numbers

More Experimental Results

UCI Diabetes dataset

Predict patient readmission

Model: random forest with 100 trees

Data:

- Training for teachers: train
- Training for student: test[:500]
- Testing: test[500:]

Non private model: 93.81%

Private model: 93.94% with $(1.44, 10^{-5})$ guarantee

Summary

Laws are catching up to your data science techniques.

Laws can be accommodated incorporating privacy techniques into Data Science

CONTACT

jim@immuta.com

@jimmuta

www.immuta.com