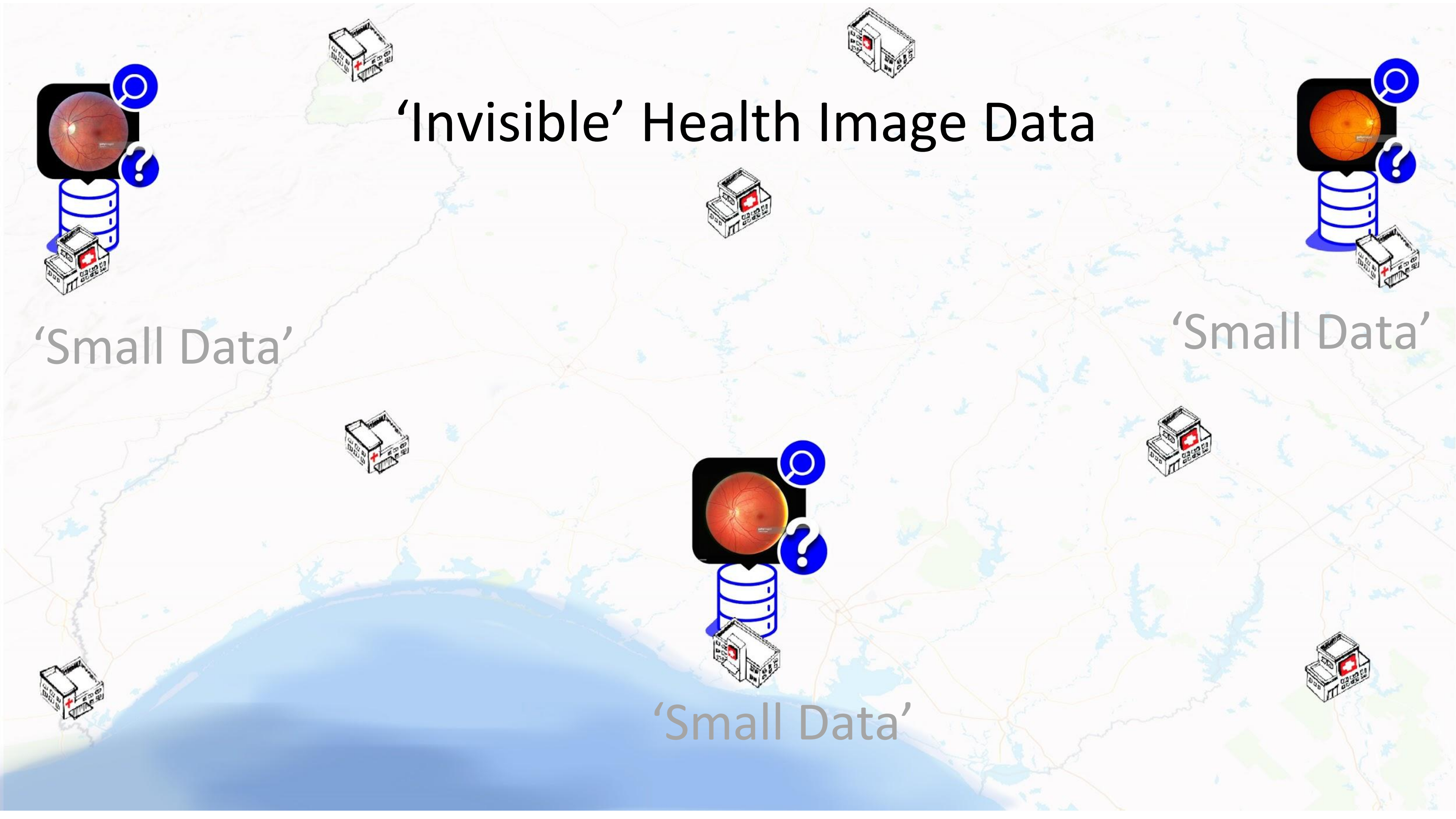
Massachusetts Institute of Technology

Split Learning

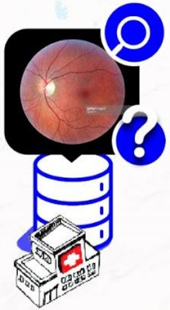
A resource efficient distributed deep learning method without sensitive data sharing

Praneeth Vepakomma
vepakom@mit.edu

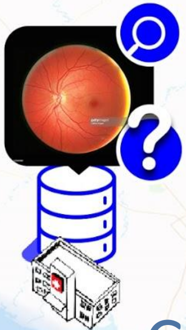
'Invisible' Health Image Data



ML for Health Images



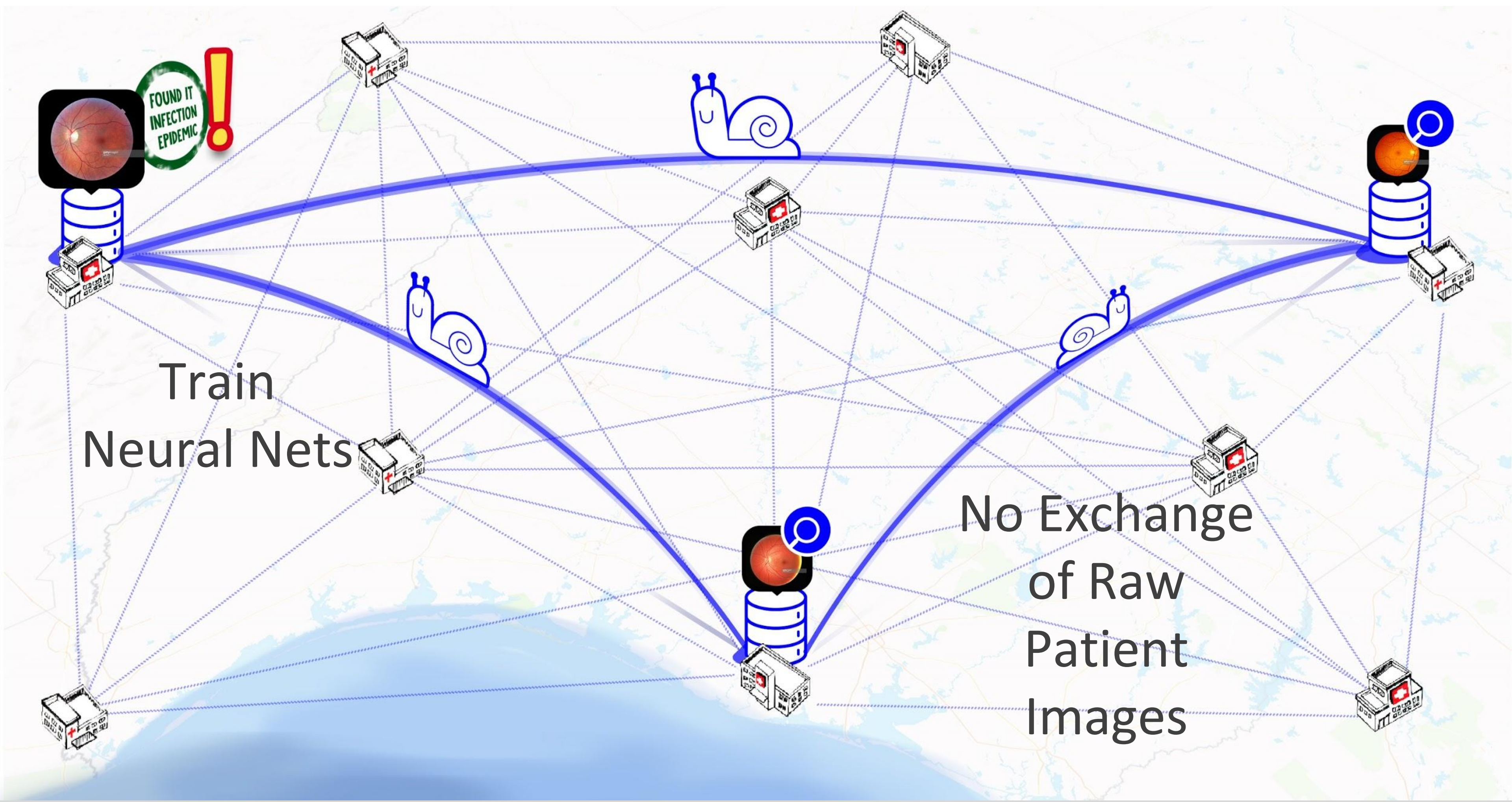
a. Distributed Data
b. Patient privacy
c. Incentives
d. ML Expertise
e. Efficiency



Low Bandwidth

'Small' Data

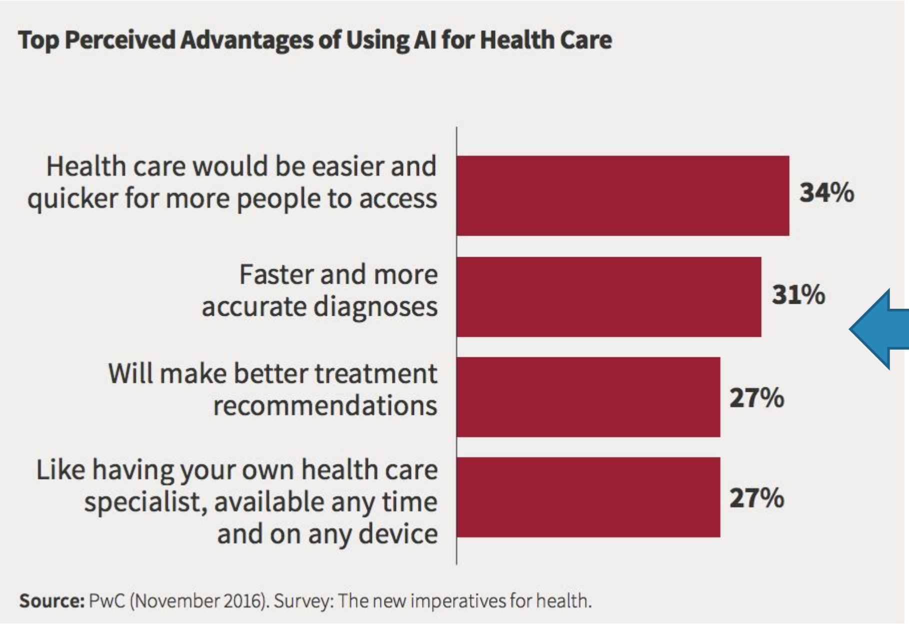
Low Compute



Gupta, Raskar ‘Distributed training of deep neural network over several agents’, 2017

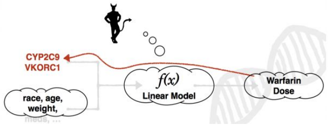
Intelligent Computing

Security, Privacy & Safety



… and predictive models can breach privacy too

Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing



Regulations

GDPR: General Data Protection Regulation

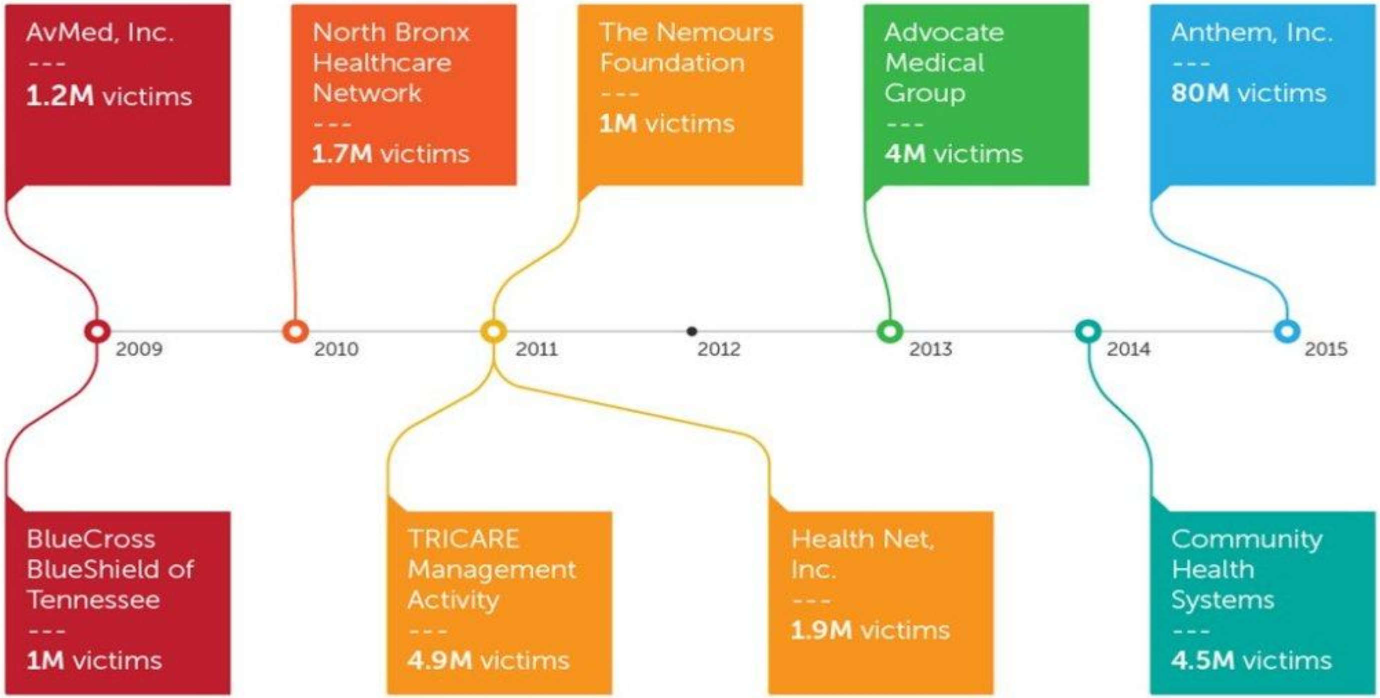
HIPAA: Health Insurance Portability and Accountability Act, 1996

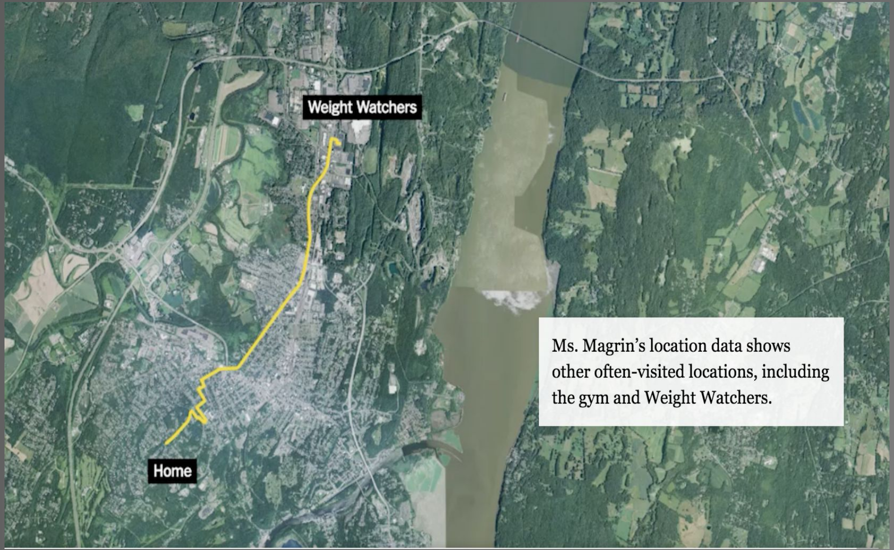
SOX: Sarbanes-Oxley Act, 2002

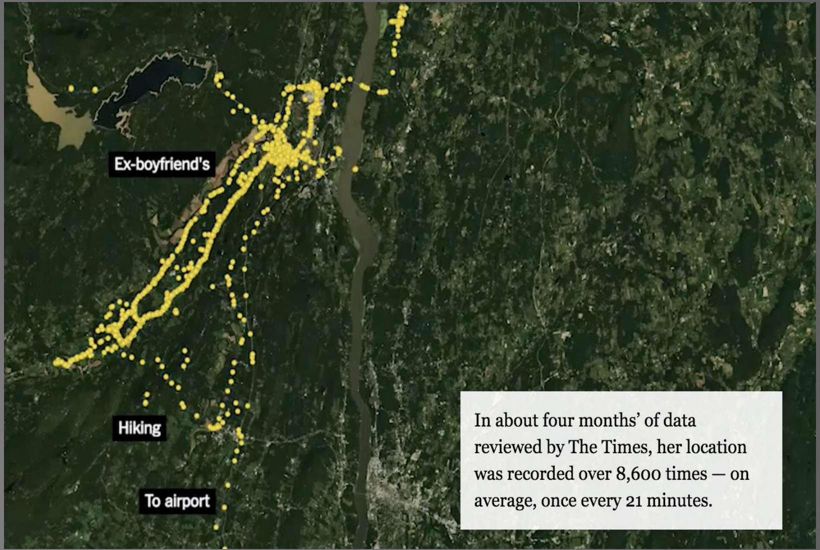
PCI: Payment Card Industry Data Security Standard, 2004

SHIELD: Stop Hacks and Improve Electronic Data Security Act, Jan 1 2019

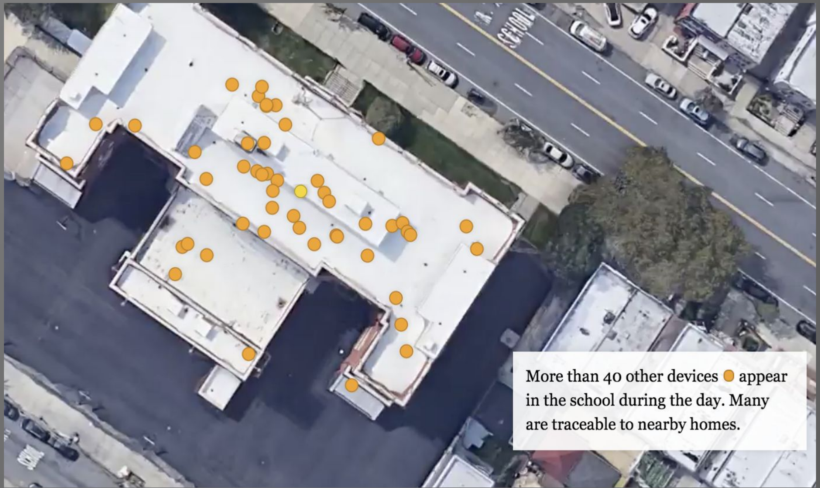
NOTABLE HEALTHCARE BREACHES




Weight Watchers
Home
Ms. Magrin's location data shows other often-visited locations, including the gym and Weight Watchers.


Ex-boyfriend's
Hiking
To airport
In about four months' of data reviewed by The Times, her location was recorded over 8,600 times — on average, once every 21 minutes.


Middle school, New York
A car pulls up at 7:30 a.m.
The device enters the school yard
Data reviewed by The Times includes dozens of schools. Here a device, most likely a child's, is tracked from a home to school.


More than 40 other devices appear in the school during the day. Many are traceable to nearby homes.

Challenges for Distributed Data + AI + Health

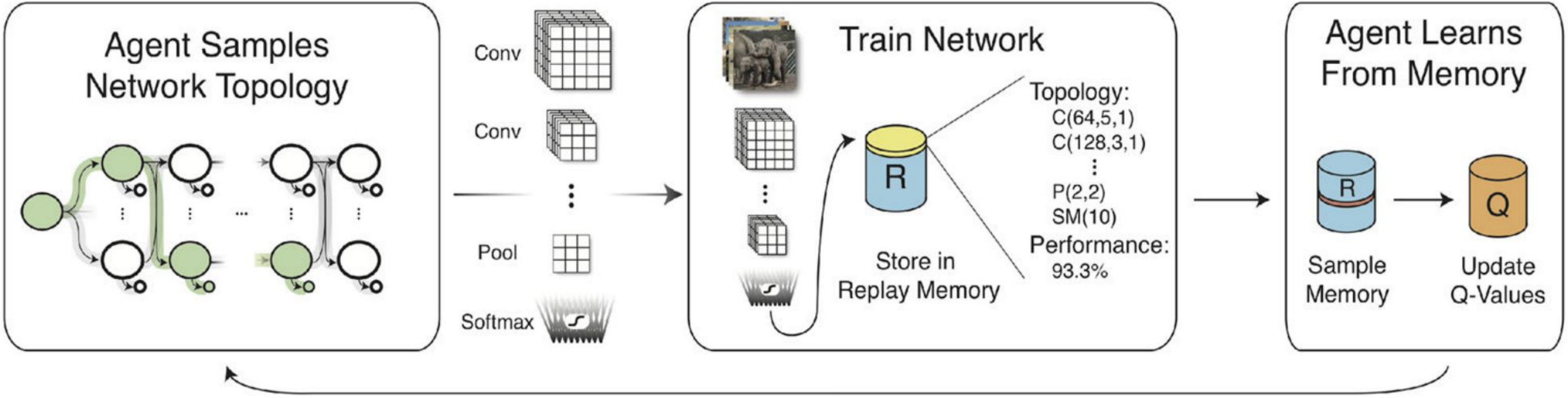
Distributed Data
Multi-Modal
Incomplete Data

Regulations
Incentives
Cooperation
Ease

Ledgering
Smart contracts
Maintenance

Resource-constraints
Memory, Compute, Bandwidth,
Convergence, Synchronization, Leakage

Automating ML : AI building AI



Teacher

Student

Otkrist Gupta, Baker, Naik, Raskar, ICLR 2017

AI: Bringing it all together

Training
Deep
Networks

No sharing of
Raw Images

Server

Client

Invisible Data / Data Friction

Overcoming Data Friction

Ease | Incentive | Trust | Regulation

Blockchain

AI/ SplitNN

Anonymize | Obfuscate | Encrypt

Protect Data

Anonymity is not enough …

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS

NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

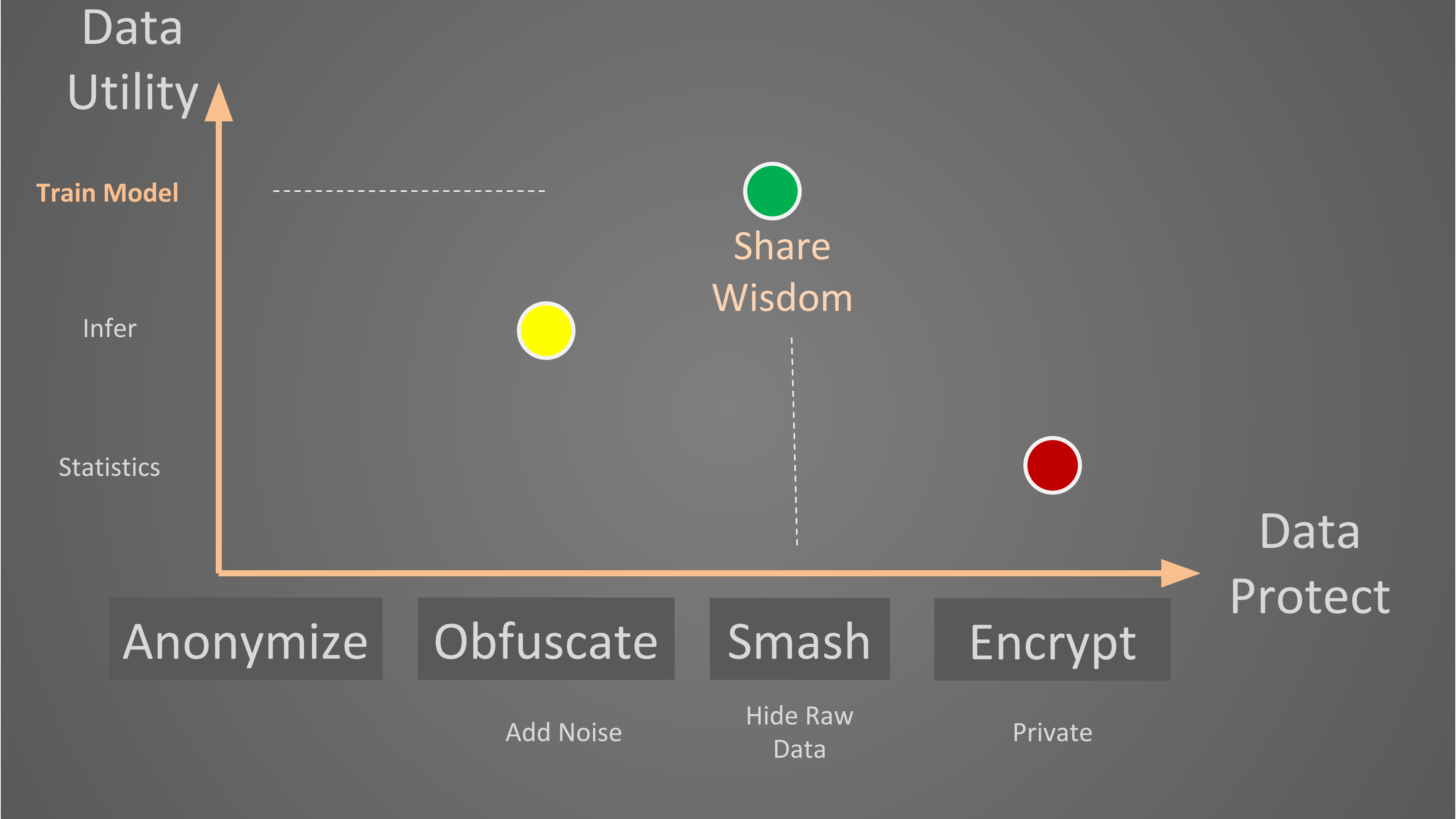
Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The Scientist » The Nutshell

"Anonymous" Genomes Identified

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013


Data
Utility
Train Model
Infer
Statistics
Share
Wisdom
Data
Protect
Anonymize
Obfuscate
Smash
Encrypt
Add Noise
Hide Raw
Data
Private

Federated Learning

Nets trained at Clients

Merged at Server

Differential Privacy

Obfuscate with noise

Hide unique samples

Split Learning (MIT)

Nets split over network

Trained at both

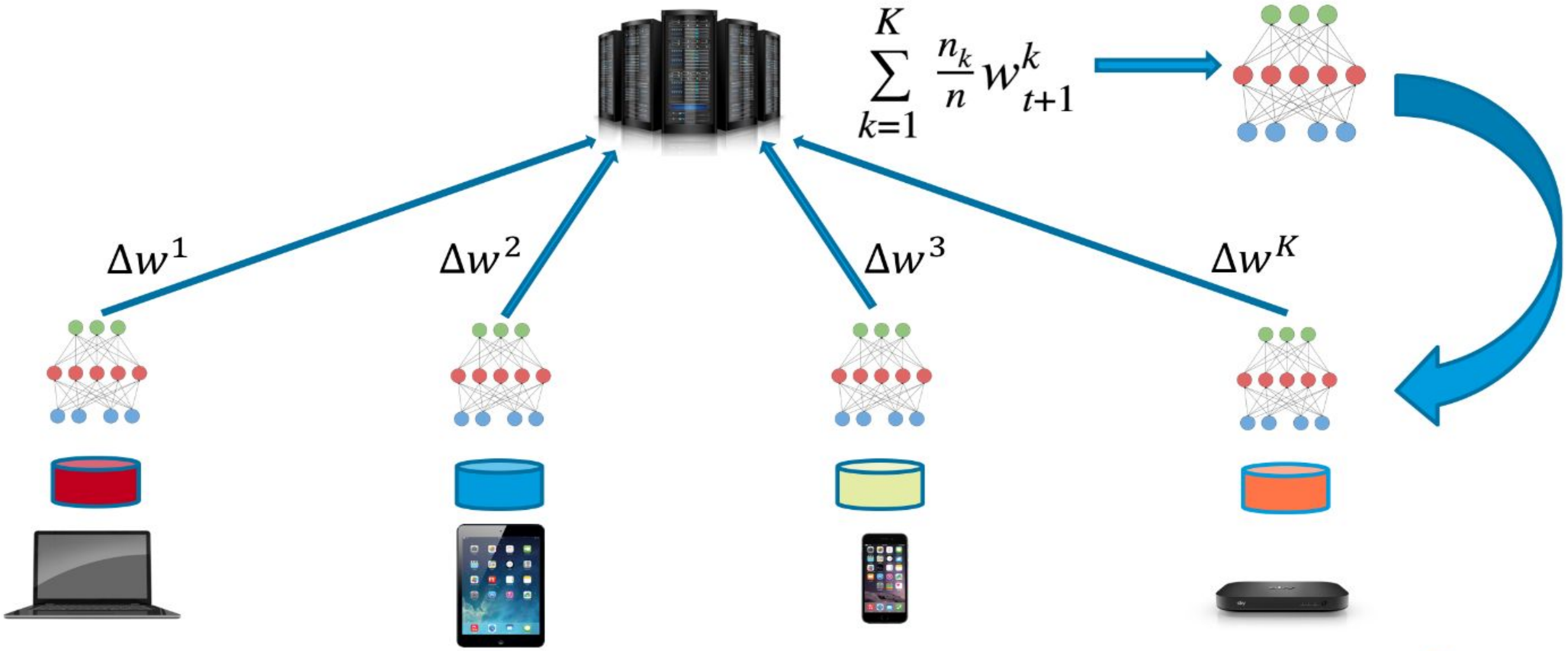
Homomorphic Encryption

Basic Math over Encrypted Data (+, x)

Federated Learning

Server

How does it work?

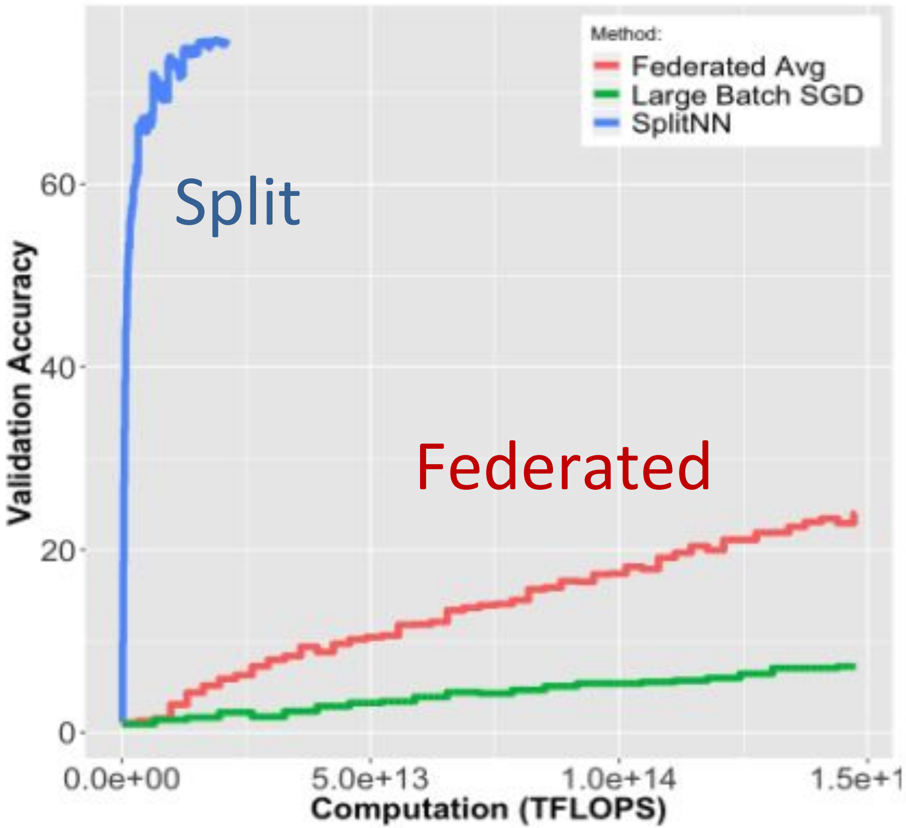


Distributed Training \ Protect data	Partial Leakage	Differential Privacy	Homomorphic Encryption	Oblivious Transfer, Garbled Circuits
Federated Learning				
Split Learning				

Inference but no training

Praneeth Vepakomma, Tristan Swedish, Otkrist Gupta, Abhi Dubey, Raskar 2018

When to use split learning?



Large number of clients:
Split learning shows positive results

Memory
Compute
Bandwidth
Convergence

Project Page and Papers:
https://splitlearning.github.io/

QUANTITATIVE RESULTS

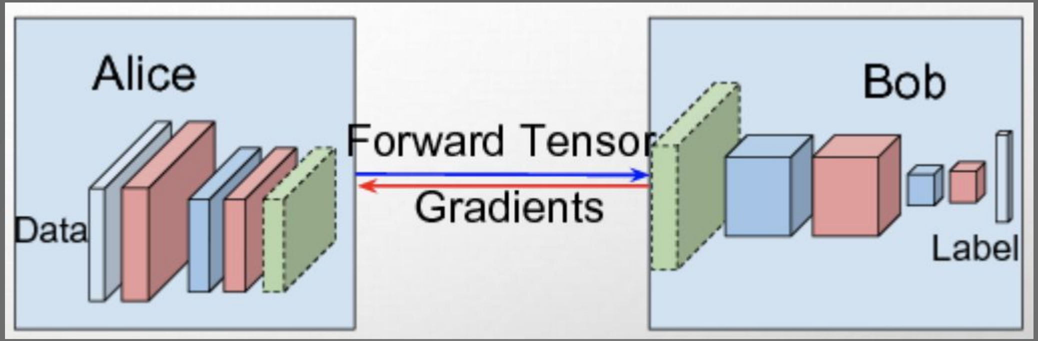
Method	100 Clients	500 Clients
Large Scale SGD	29.4 TFlops	5.89 TFlops
Federated Learning	29.4 TFlops	5.89 TFlops
Our Method (SplitNN)	0.1548 TFlops	0.03 TFlops

Table 1. Computation resources consumed per client when training CIFAR 10 over VGG (in teraflops)

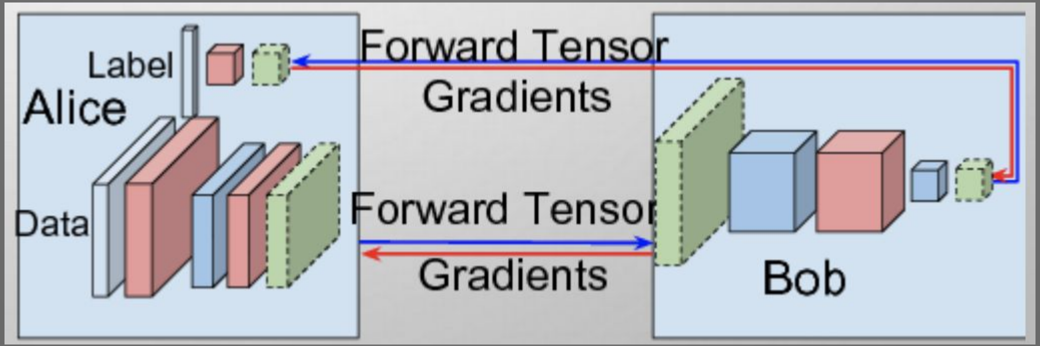
Method	100 Clients	500 Clients
Large Scale SGD	13 GB	14 GB
Federated Learning	3 GB	2.4 GB
Our Method (SplitNN)	6 GB	1.2 GB

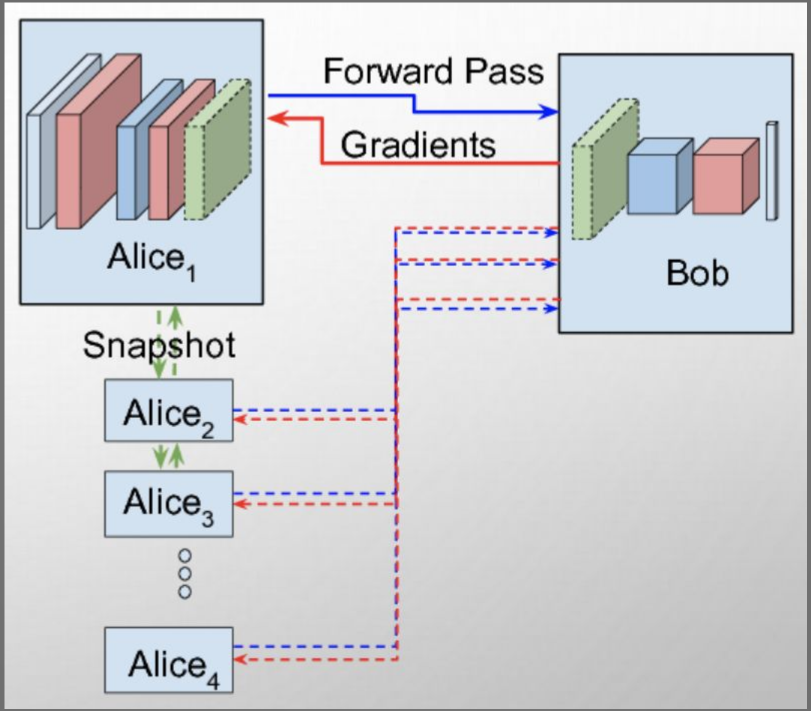
Table 2. Communication Bandwidth consumed per client when training CIFAR 100 and Resnet 50 (in gigabytes)

Label Sharing



No Label Sharing



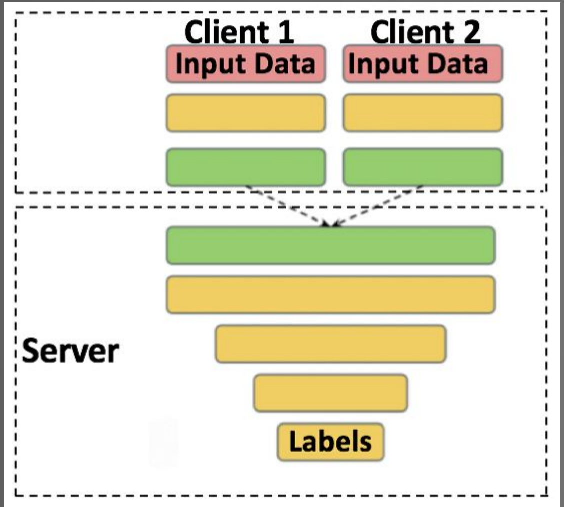


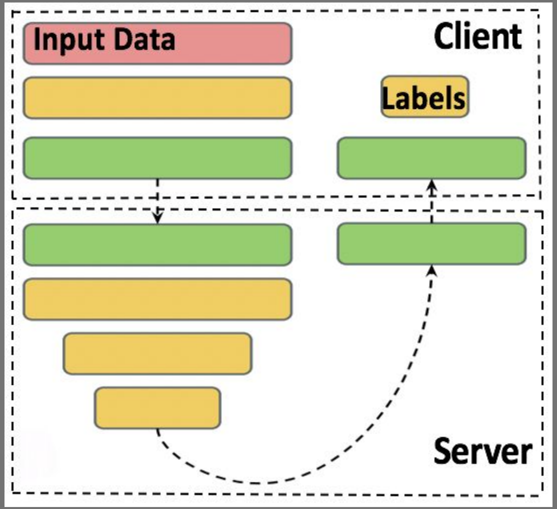
Gupta, Otkrist, and Raskar, Ramesh. "Secure Training of Multi-Party Deep Neural Network." U.S. Patent Application No. 15/630,944.

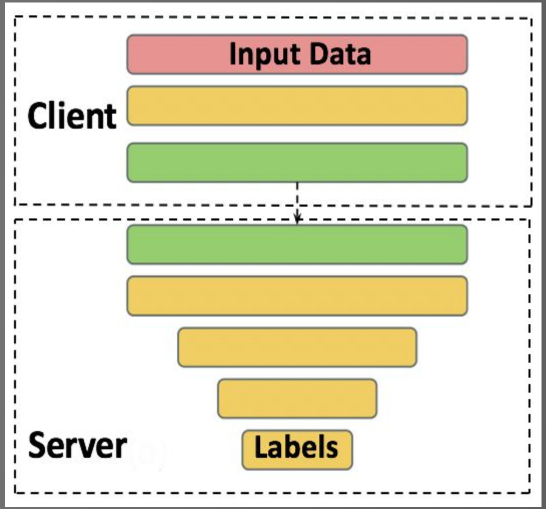
Distribution of parameters in AlexNet

Layer Name	Tensor Size	Weights	Biases	Parameters
Input Image	227x227x3	0	0	0
Conv-1	55x55x96	34,848	96	34,944
MaxPool-1	27x27x96	0	0	0
Conv-2	27x27x256	614,400	256	614,656
MaxPool-2	13x13x256	0	0	0
Conv-3	13x13x384	884,736	384	885,120
Conv-4	13x13x384	1,327,104	384	1,327,488
Conv-5	13x13x256	884,736	256	884,992
MaxPool-3	6x6x256	0	0	0
FC-1	4096×1	37,748,736	4,096	37,752,832
FC-2	4096×1	16,777,216	4,096	16,781,312
FC-3	1000×1	4,096,000	1,000	4,097,000
Output	1000×1	0	0	0
Total				**62,378,344**

Versatile Configurations of Split Learning

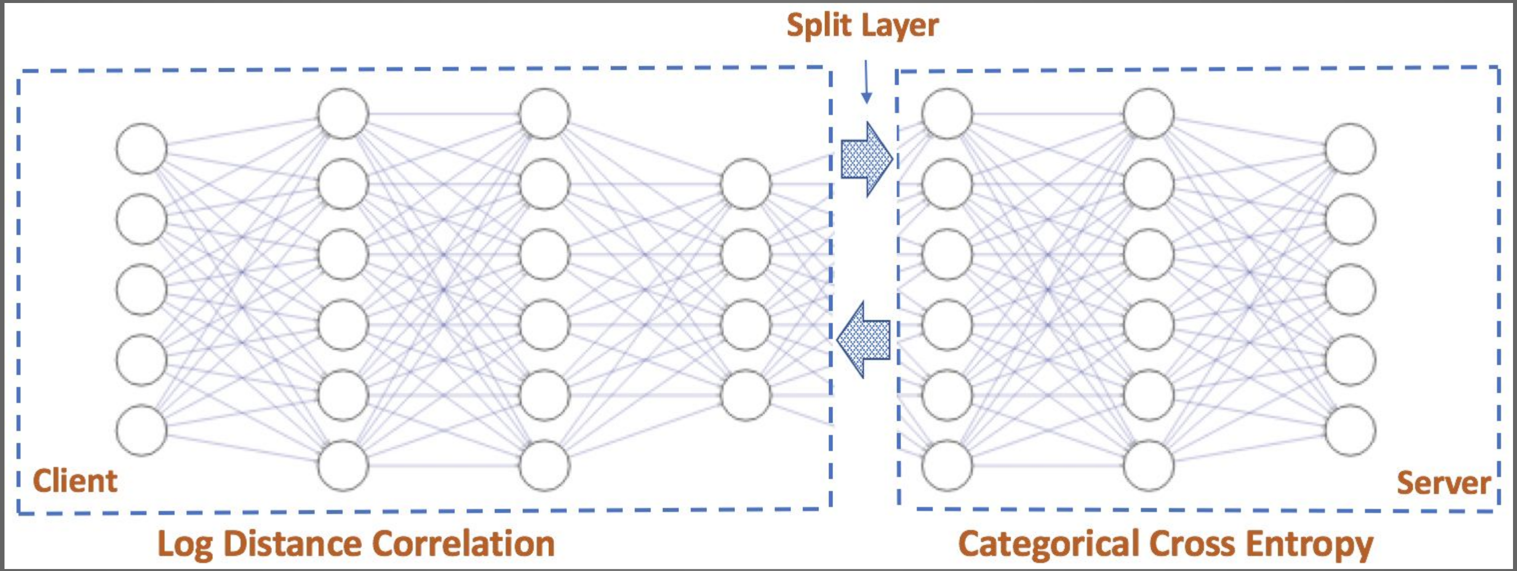






Split learning for health: Distributed deep learning without sharing raw patient data, Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, Ramesh Raskar, (2019)

NoPeek SplitNN: Reducing Leakage in Distributed Deep Learning



$$\alpha_1 DCOR(\mathbf{X_n}, \hat{\mathbf{Z}}) + \alpha_2 CCE(\hat{\mathbf{Y}}, \mathbf{Y_n})$$

Reducing leakage in distributed deep learning for sensitive health data, Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, Ramesh Raskar (2019)

No peak deep learning with conditioning variable

Setup:

- Supervised: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\} \subset X \times Y$
- Output: $y \in \mathbb{R}$
- *Goal* : To find a projection $\mathcal{S}_{Y|X}$ such that, $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}$.

Ideal Goal: To find such a conditioning variable Z within the framework of deep learning such that the following directions are approximately satisfied:

1. Y ⫫ X | Z (Utility property as X can be thrown away given Z to obtain prediction E(Y|Z))
2. X ⫫ Z (One-way property preventing proper reconstruction of raw data X from Z)

Note: ⫫ denotes statistical independence

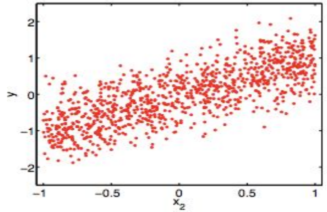
Possible measures of non-linear dependence

- COCO: Constrained Covariance
- HSIC: Hilbert-Schmidt Independence Criterion
- DCOR: Distance Correlation
- MMD: Maximum Mean Discrepancy
- KTA: Kernel Target Alignment
- MIC: Maximal Information Coefficient
- TIC: Total Information Coefficient

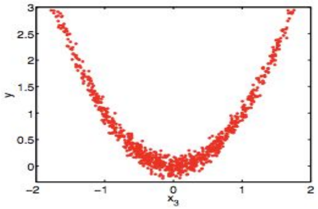
Why is it called distance correlation?

Definition 3.1. Sample Distance Covariance [3]: Given i.i.d samples $\mathcal{X}\times\mathcal{Y} = \{(\mathbf{x}_k, \mathbf{y}_k)|k = 1, 2, 3, \ldots, n\}$ and corresponding double centered Euclidean distance matrices $\widehat{\mathbf{E}}_{\mathbf{X}}$ and $\widehat{\mathbf{E}}_{\mathbf{Y}}$, then the squared sample distance correlation is defined as,

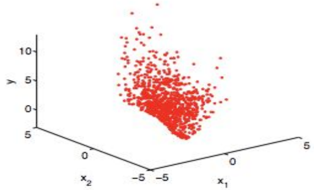
$$\hat{\nu}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n}[\widehat{\mathbf{E}}_{\mathbf{X}}]_{k,l}[\widehat{\mathbf{E}}_{\mathbf{Y}}]_{k,l},$$



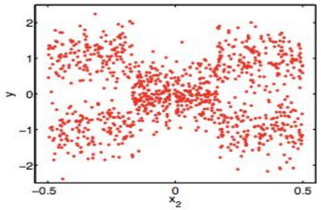
(a) linear



(b) non-linear



(c) non-linear



(d) non-linear

Distance Covariance (Székely, G. (2007))

$$\nu^2(\mathbf{X}, \mathbf{Y}; w) = \int_{\mathbb{R}^{h+m}} |f_{\mathbf{X},\mathbf{Y}}(t, s) - f_{\mathbf{X}}(t) f_{\mathbf{Y}}(s)|^2 w(t, s) dtds$$

where $f_{\mathbf{X}}, f_{\mathbf{Y}}, f_{\mathbf{X},\mathbf{Y}}$ are the characteristic functions of $\mathbf{X}, \mathbf{Y}, \mathbf{X} \times \mathbf{Y}$ and $w(t, s)$ is a suitably chosen weight function.

Sample Distance Covariance (2nd order)

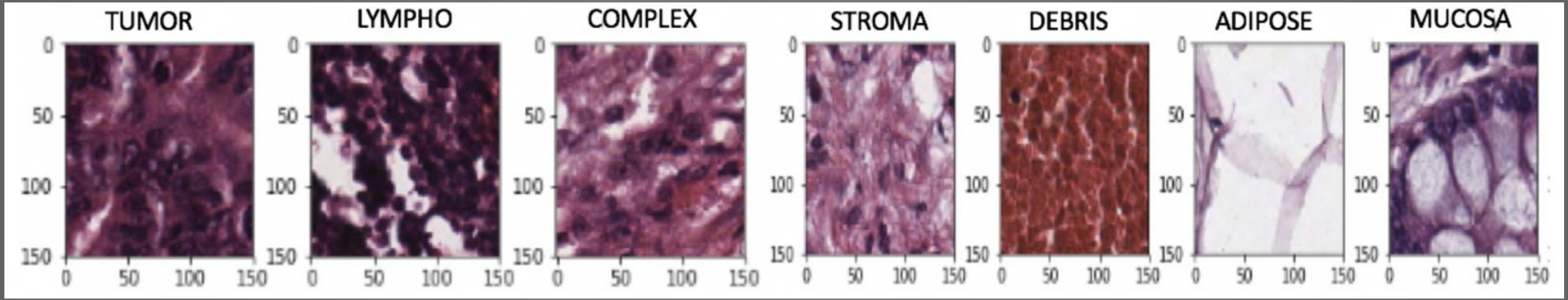
$$\hat{\nu}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \mathbf{Tr} \left(\mathbf{L_x}^T \mathbf{L_y} \right)$$

where $\mathbf{L}_x = \mathbf{D}_x - \mathbf{H}\mathbf{E}_\mathbf{x}\mathbf{H}$ and $\mathbf{L}_y = \mathbf{D}_y - \mathbf{H}\mathbf{E}_\mathbf{y}\mathbf{H}$.

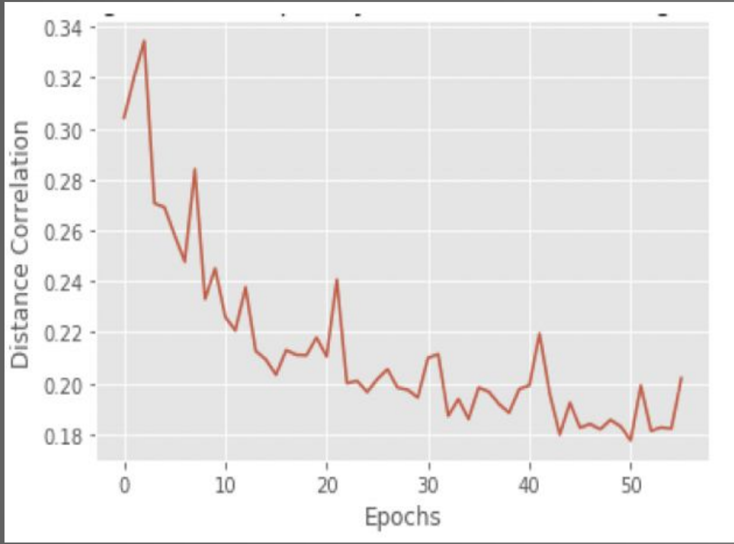
Lemma 3.1. *Given matrices of squared Euclidean distances* $\mathbf{E}_\mathbf{X}$ *and* $\mathbf{E}_\mathbf{Y}$ *and Laplacians* $\mathbf{L}_\mathbf{X}$ *and* $\mathbf{L}_\mathbf{Y}$ *formed over adjacency matrics* $\widehat{\mathbf{E}}_\mathbf{X}$ *and* $\widehat{\mathbf{E}}_\mathbf{Y}$*, the square of sample distance correlation* $\hat{\rho}^2(\mathbf{X}, \mathbf{Y})$ *is given by*

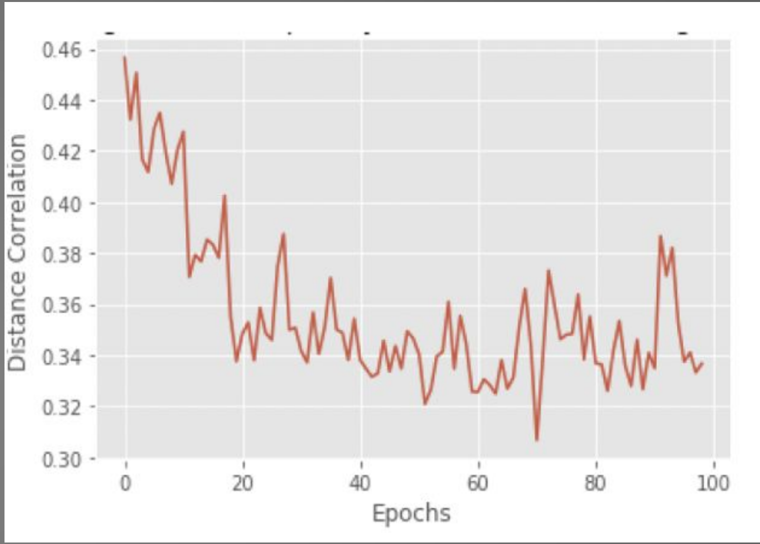
$$\hat{\rho}^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{Tr}\left(\mathbf{X}^T \mathbf{L}_\mathbf{Y} \mathbf{X}\right)}{\sqrt{\mathbf{Tr}\left(\mathbf{Y}^T \mathbf{L}_\mathbf{Y} \mathbf{Y}\right) \mathbf{Tr}\left(\mathbf{X}^T \mathbf{L}_\mathbf{X} \mathbf{X}\right)}}.$$

Colorectal histology image dataset (Public data)



Leakage Reduction in Action



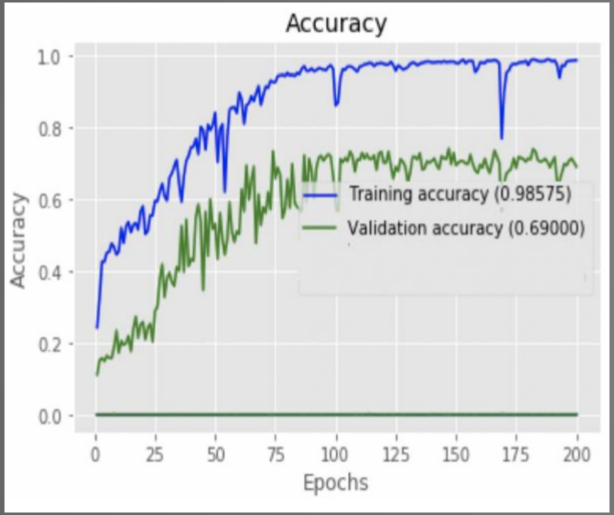


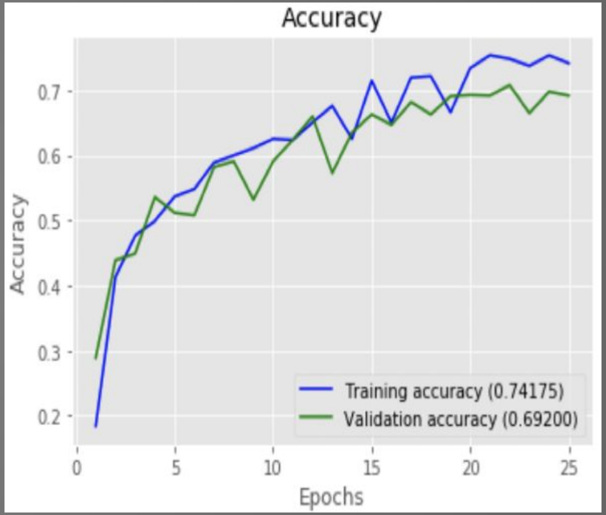
Reduced leakage during training over colorectal histology image data from 0.96 in traditional CNN to 0.19 in NoPeek SplitNN

Reduced leakage during training over colorectal histology image data from 0.92 in traditional CNN to 0.33 in NoPeek SplitNN

Reducing leakage in distributed deep learning for sensitive health data, Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, Ramesh Raskar (2019)

Similar validation performance





Effect of leakage reduction on convergence

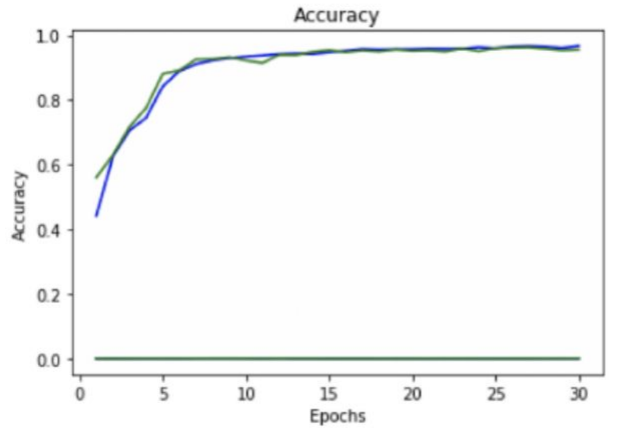


Figure 9: $\alpha_1 = 0.05$

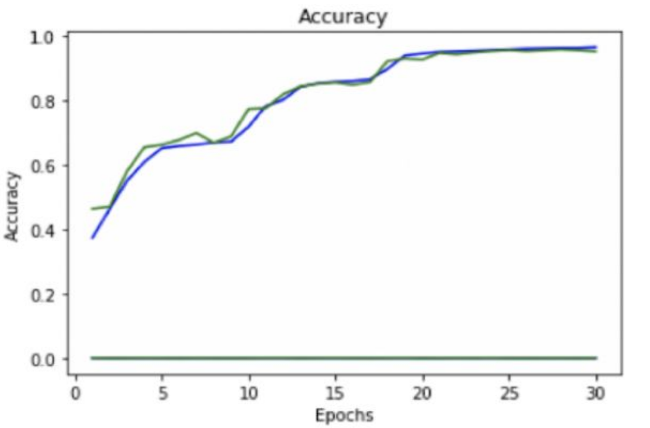


Figure 10: $\alpha_1 = 0.15$

Robustness to reconstruction

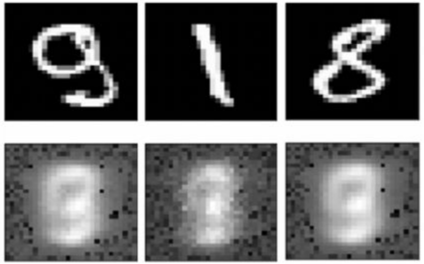


Figure 7: $\alpha_1 = 0.1$

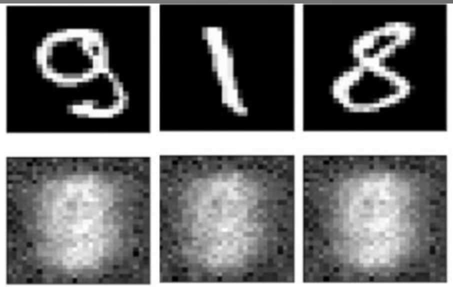


Figure 8: $\alpha_1 = 0.9$

Proof of one-Way Property:

$$DCOV(\mathbf{X}, \mathbf{Z}) = Tr(\mathbf{X}\mathbf{X}^{\mathbf{T}}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}) + \|\mathbf{X} - \mathbf{Z}\| + \|\mathbf{Z}\|$$

$$D_{KL}(\mathbf{Z}||\mathbf{X}) - D_{KL}(\mathbf{X}||\mathbf{Z}) = H(\mathbf{Z}, \mathbf{X}) - H(\mathbf{Z}) - H(\mathbf{X}, \mathbf{Z}) + H(\mathbf{X})$$

We show: Minimizing regularized distance covariance minimizes the difference of Kullback-Leibler divergences

$$= \det(\mathbf{Z^T X}) - \det(\mathbf{Z^T Z}) - \det(\mathbf{X^T Z}) + \det(\mathbf{X^T X})$$

This can be bounded using Hadamard's inequality as

$$\det(\mathbf{Z^T X}) - \det(\mathbf{Z^T Z}) + \det(\mathbf{X^T X}) - \det(\mathbf{X^T Z}) \leq \left\|\mathbf{Z^T X} - \mathbf{Z^T Z}\right\|_2 \frac{\left\|\mathbf{Z^T X}\right\|_2^n - \left\|\mathbf{Z^T Z}\right\|_2^n}{\left\|\mathbf{Z^T X}\right\|_2 - \left\|\mathbf{Z^T Z}\right\|_2} + \left\|\mathbf{X^T Z} - \mathbf{X^T X}\right\|_2 \frac{\left\|\mathbf{X^T Z}\right\|_2^n - \left\|\mathbf{X^T X}\right\|_2^n}{\left\|\mathbf{X^T Z}\right\|_2 - \left\|\mathbf{X^T X}\right\|_2}$$

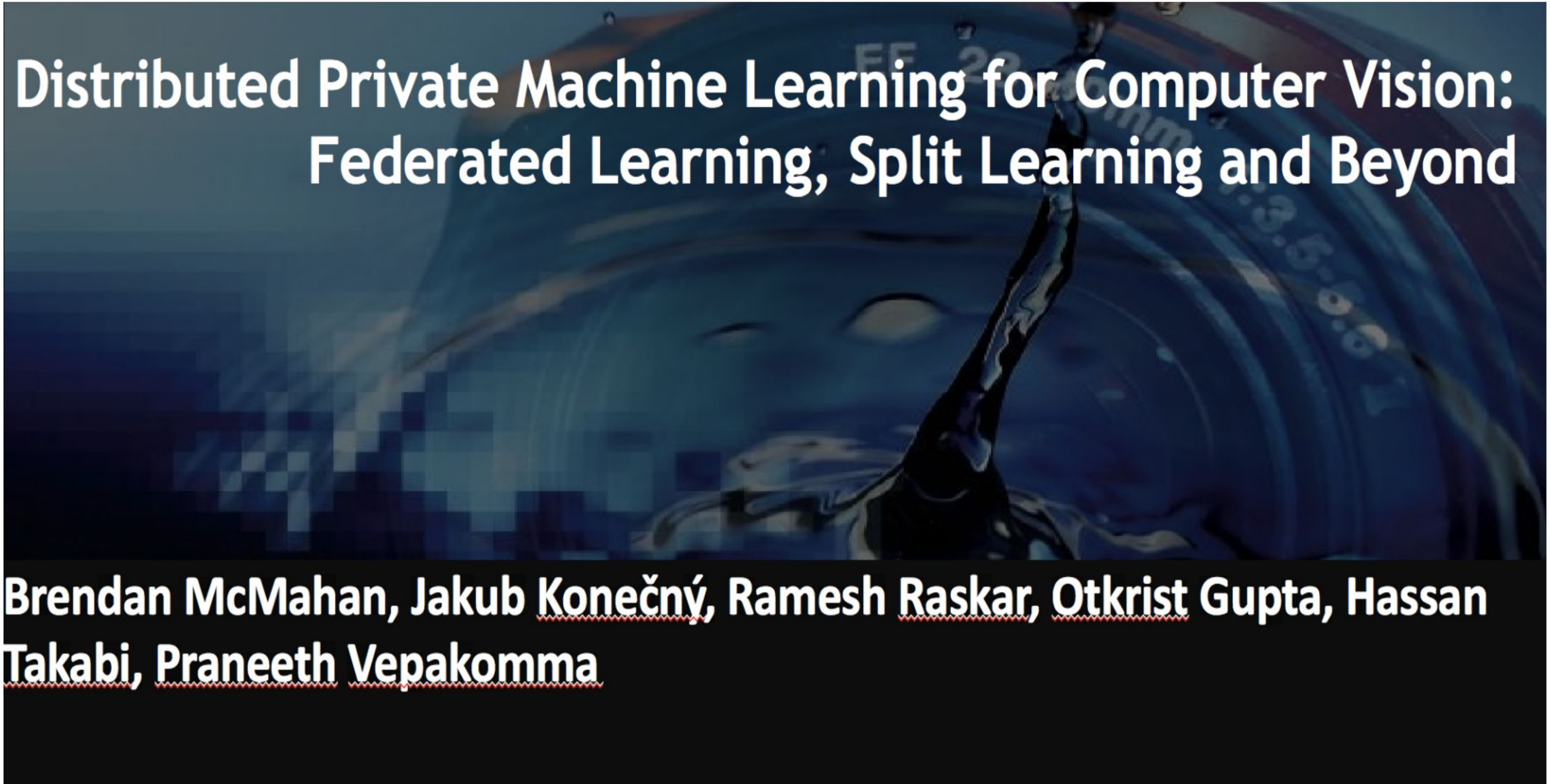
The fractional terms $\frac{\left\|\mathbf{Z^T X}\right\|_2^n - \left\|\mathbf{Z^T X}\right\|_2^n}{\left\|\mathbf{Z^T X}\right\|_2 - \left\|\mathbf{Z^T Z}\right\|_2}$, $\frac{\left\|\mathbf{X^T Z}\right\|_2^n - \left\|\mathbf{X^T X}\right\|_2^n}{\left\|\mathbf{X^T Z}\right\|_2 - \left\|\mathbf{X^T X}\right\|_2}$ can be written as a sum of geometric-series, with factors of change of $\frac{\left\|\mathbf{Z^T X}\right\|}{\left\|\mathbf{Z^T Z}\right\|}$, $\frac{\left\|\mathbf{X^T Z}\right\|}{\left\|\mathbf{X^T X}\right\|}$ respectively because

$$\frac{\left\|\mathbf{Z^T X}\right\|_2^n - \left\|\mathbf{Z^T Z}\right\|_2^n}{\left\|\mathbf{Z^T X}\right\|_2 - \left\|\mathbf{Z^T Z}\right\|_2} = \frac{1 - \left(\frac{\left\|\mathbf{Z^T X}\right\|_2}{\left\|\mathbf{Z^T Z}\right\|_2}\right)^n}{1 - \frac{\left\|\mathbf{Z^T X}\right\|_2}{\left\|\mathbf{Z^T Z}\right\|_2}} = \sum_{p=0}^{n-1} \left\|\mathbf{Z^T X}\right\|_2^p \left\|\mathbf{Z^T Z}\right\|_2^{p-1}$$

Therefore these fractional terms can be minimized by minimizing $\left\|\mathbf{Z^T X}\right\|_2$ and $\left\|\mathbf{Z^T Z}\right\|_2$ as the sums of products of decreasing functions of norms are also decreasing. By Cauchy-Schwarz inequality $\left\|\mathbf{Z^T}(\mathbf{X} - \mathbf{Z})\right\| \leq \left\|\mathbf{Z}\right\| \left\|\mathbf{X} - \mathbf{Z}\right\|$.

Therefore the upper-bound on difference of KL-divergence can be minimized by minimizing $\left\|\mathbf{Z}\right\|$ and $\left\|\mathbf{X} - \mathbf{Z}\right\|$ to minimize terms $\left\|\mathbf{Z^T X} - \mathbf{Z^T Z}\right\|$,$\left\|\mathbf{X^T Z} - \mathbf{X^T X}\right\|$ in addition to minimizing $\left\|\mathbf{Z^T Z}\right\|$, $\left\|\mathbf{Z^T X}\right\|_2 = Tr(\mathbf{Z^T X X^T Z}) = DCOV(\mathbf{X}, \mathbf{Z})$ to minimize terms $\frac{\left\|\mathbf{Z^T X}\right\|_2^n - \left\|\mathbf{Z^T X}\right\|_2^n}{\left\|\mathbf{Z^T X}\right\|_2 - \left\|\mathbf{Z^T Z}\right\|_2}$, $\frac{\left\|\mathbf{X^T Z}\right\|_2^n - \left\|\mathbf{X^T X}\right\|_2^n}{\left\|\mathbf{X^T Z}\right\|_2 - \left\|\mathbf{X^T X}\right\|_2}$.

CVPR Tutorial On Distributed Private Machine Learning for Computer Vision: Federated Learning, Split Learning and Beyond



Project Page and Papers: https://splitlearning.github.io/

Thanks and acknowledgements to: Otkrist Gupta (MIT/LendBuzz), Ramesh Raskar (MIT), Jayashree Kalpathy-Cramer (Martinos/Harvard), Rajiv Gupta (MGH), Brendan McMahan (Google), Jakub Konečný (Google), Abhimanyu Dubey (MIT), Tristan Swedish (MIT), Sai Sri Sathya (S20.ai), Vitor Pamplona (MIT/EyeNetra), Rodmy Paredes Alfaro (MIT), Kevin Pho (MIT), Elsa Itambo (MIT)

Massachusetts Institute of Technology

THANK

YOU