Cryptography for Privacy-Preserving Machine Learning

PhD Defense · Théo Ryffel

Directors

David Pointcheval
Francis Bach

Jury

Aurélien Bellet
Yuval Ishai
Renaud Sirdey
Mariya Georgieva
Laurent Massoulié
Jonathan Passerat-Palmbach

23/06/2023
Paris

Your everyday life... is fueled with ML

- Morning music
- Route planning
- Email filtering
- Automatic translation

Machine Learning

Definition of Machine Learning

The process of computers changing the way they carry out tasks by **learning from new data**, **without** a human being needing to give **instructions in the form of a program** - *Cambridge Dictionary*

Example

Classify a skin tumor as benign or cancerous => no simple rules

ML in healthcare?

ML in Marketing
(2018)

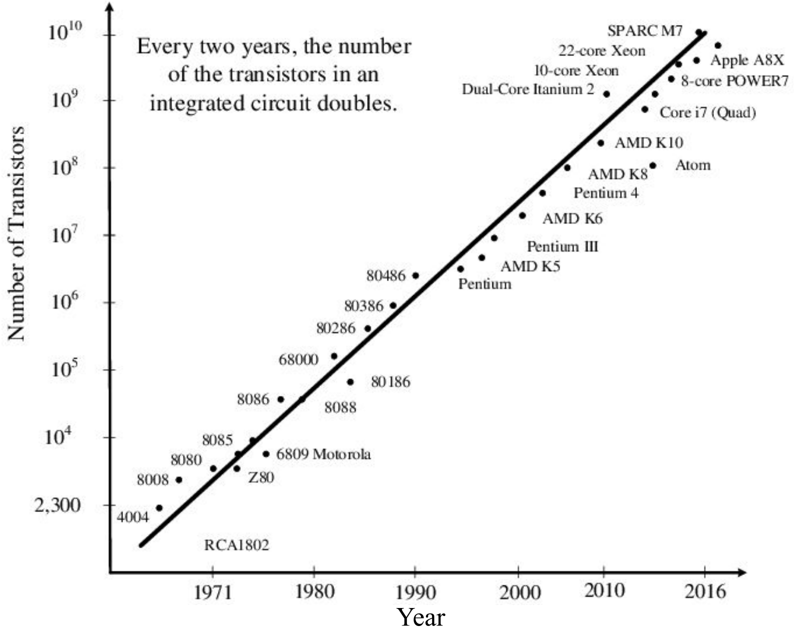
ML in Healthcare
(2018)

Refs. https://www.marketresearchfuture.com/ https://www.marketsandmarkets.com/

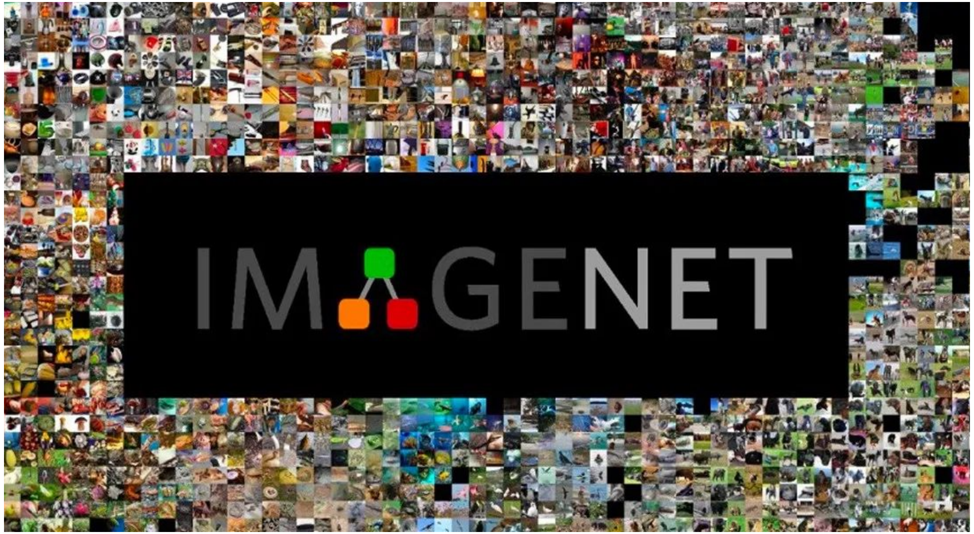
Machine Learning

Machine Learning needs powerful processors and data in large quantities

Moore's Law



Imagenet (2009)



Contextual Integrity [Nis04]

Contextual Integrity (CI)

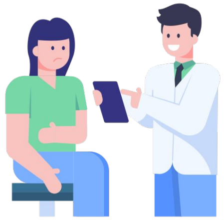
Privacy is respected when an **information flow** from one individual to another via a dedicated channel is **appropriate**, with respect to the sender, the recipient, the person concerned, the type of information and the transmission principle.

Remarks

- Not only about one's own information => not secrecy
- Positive definition: information flows => collaboration
- Ethical dimension

Contextual Integrity [Nis04]

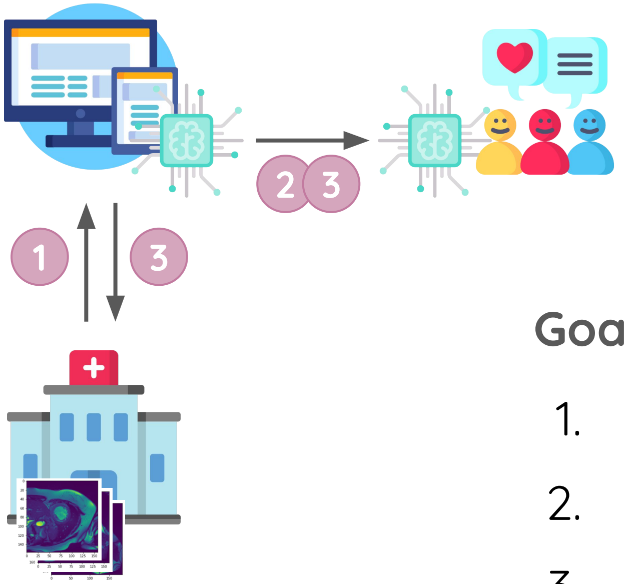
Example which satisfies Contextual Integrity:



A patient sends their own medical report to their doctor
via a secure messaging system

What if you replace:
- doctor *by* employer ?
- secure messaging app *by* public communication channel ?
- their own medical report *by* the one of a relative ?

Privacy-preserving ML through the prism of CI



Motivation:

Explore how privacy enhancing technologies can provide contextual integrity to machine learning workflows.

Goals:

1. Data used for training should not directly be exposed
2. ML models trained should not disclose private data
3. ML models should not be disseminated or exposed

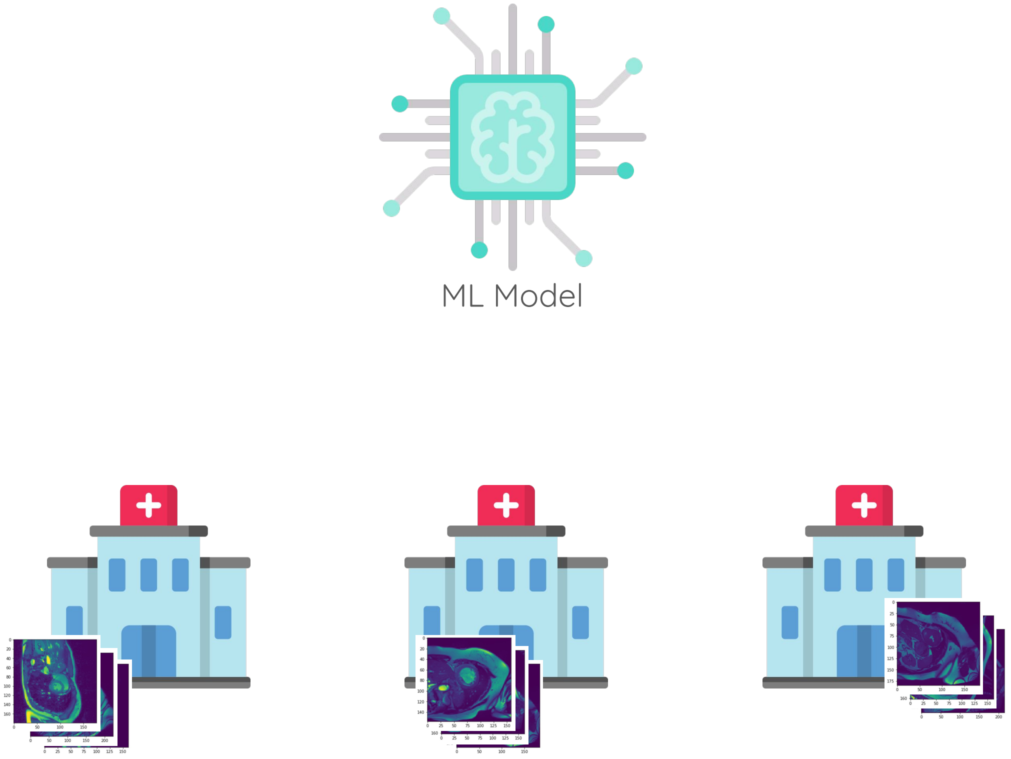
Privacy-preserving ML through the prism of CI

1. Federated Learning (and attacks on models)
2. Differential Privacy
3. Encrypted Computation

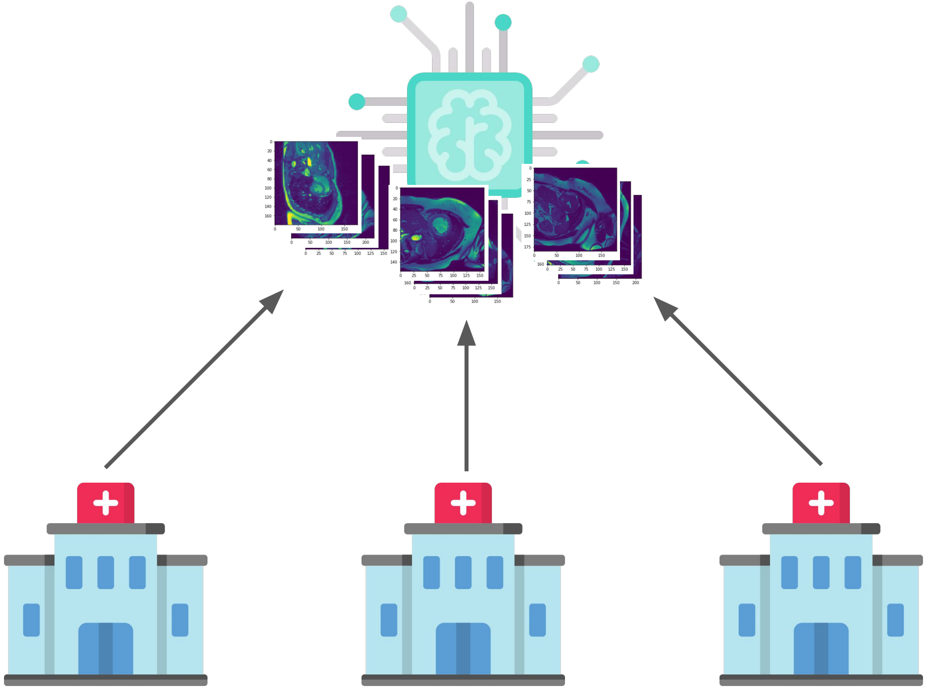
1

Federated Learning (and attacks on models)

Federated Learning [MMR+17]



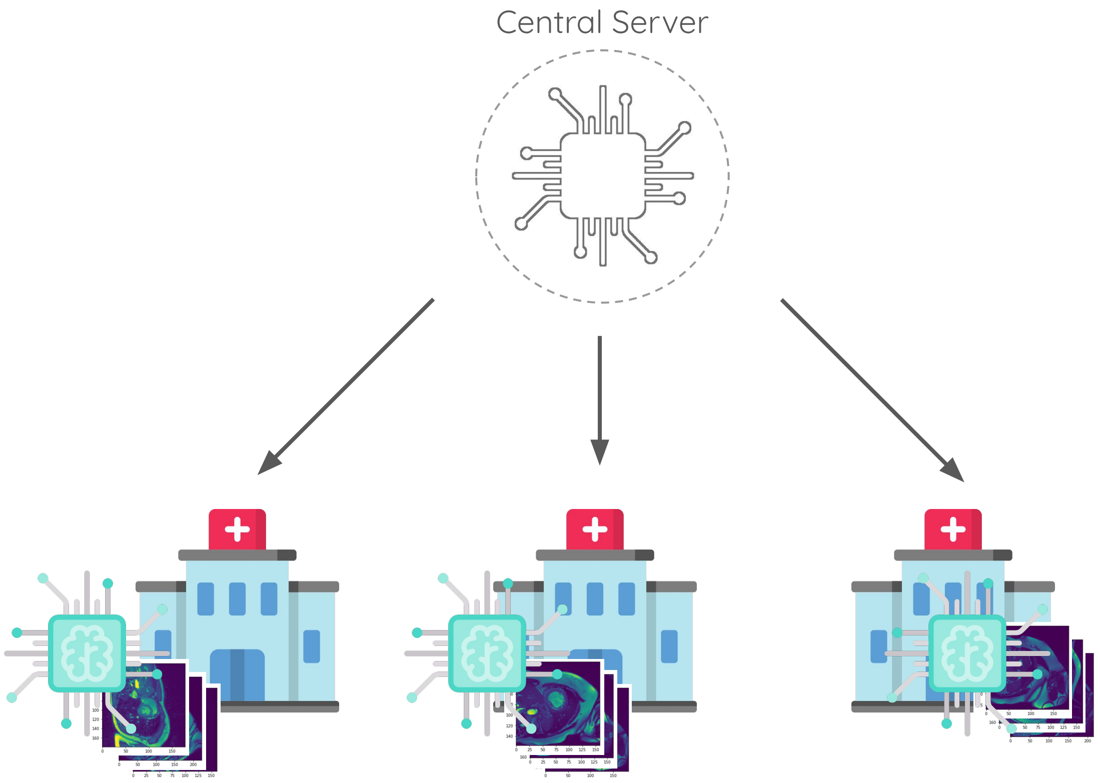
Federated Learning



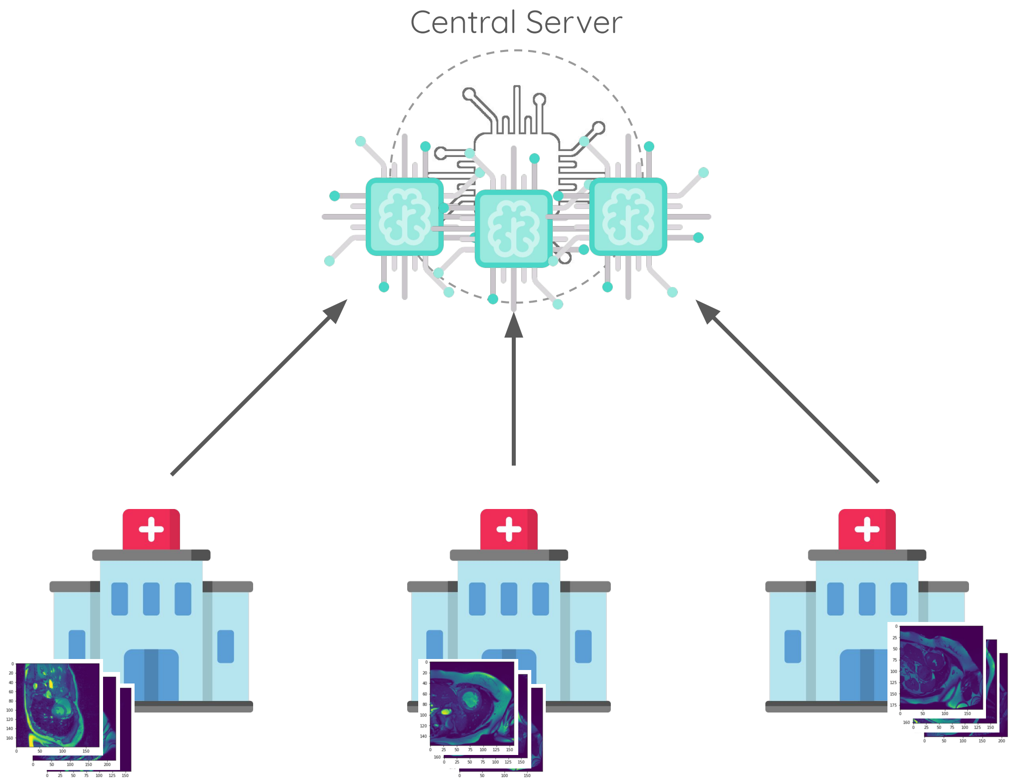
Current situation

- No sovereignty nor control on how data is used
- Important quantity of data to store

Federated Learning



Federated Learning



Benefits

- Sovereignty over data
- Transparency on computations

Our contribution:

PySyft [RTD+18]

Federated Learning <> Contextual Integrity

- Personal data used for training should not be directly exposed
- ML models should not disclose private training data
- ML models should not be disseminated or exposed

Threats against ML models

Example #1: Model inversion

1. On a fully trained network [FJR15]
   Black box setting

Original Reconstructed

2. During a federated training [GBDM20]
   Attack by the central server, white box setting

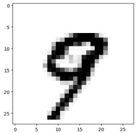
Original Reconstructed

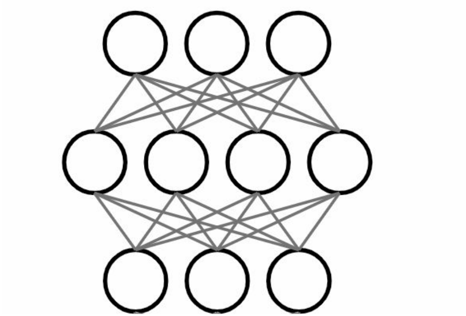
Threats against ML models

Example #2: Membership Inference [SSSS17]

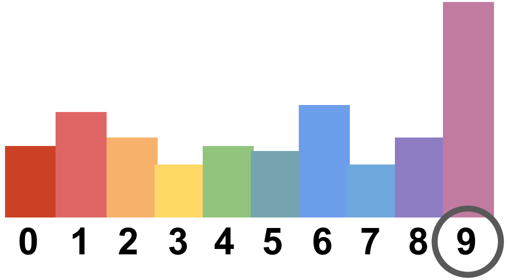
Threats against ML models

Our contribution: “Collateral Learning” [RPB+19]



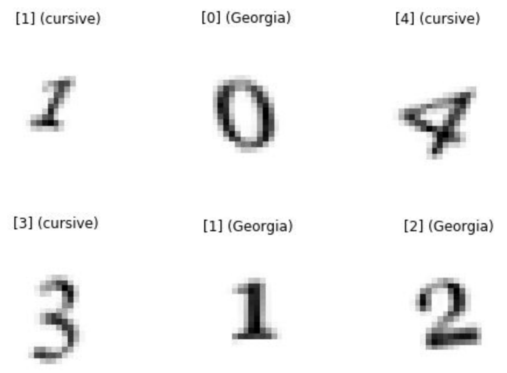


10 output classes $c_0, c_1 \dots c_9$



Collateral Learning [RPB+19]

A dataset with two classification tasks



Main Task

1 0 4

3 1 2

Collateral Task

cursive georgia cursive

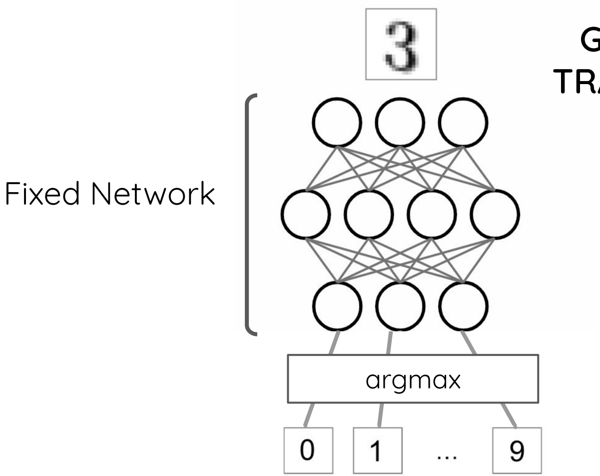
cursive georgia georgia

Collateral Learning [RPB+19]

Main Task

Collateral Task

GLOBAL TRAINING

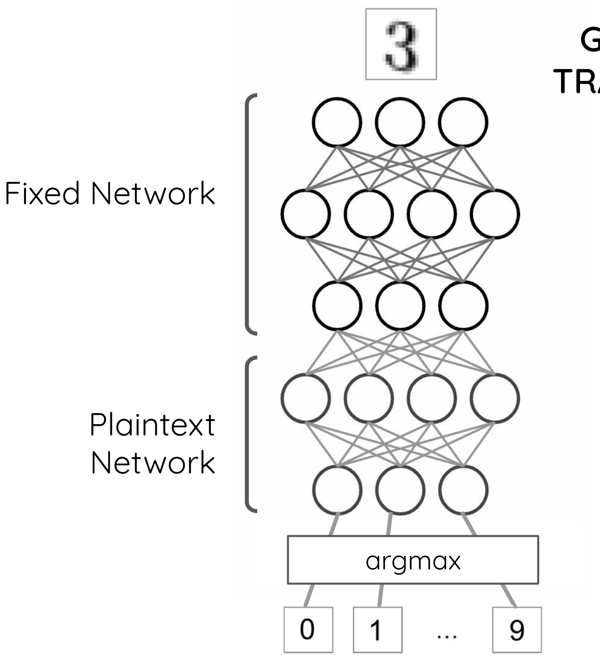


Collateral Learning [RPB+19]

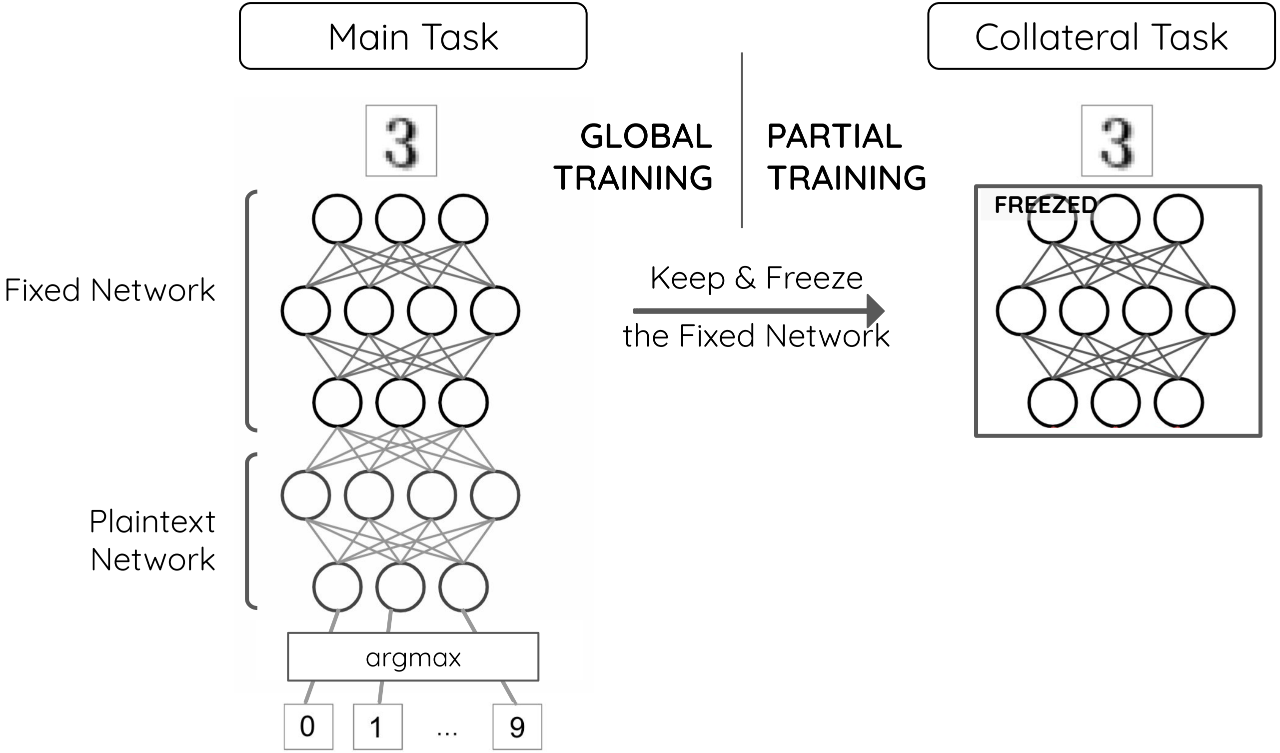
Main Task

Collateral Task

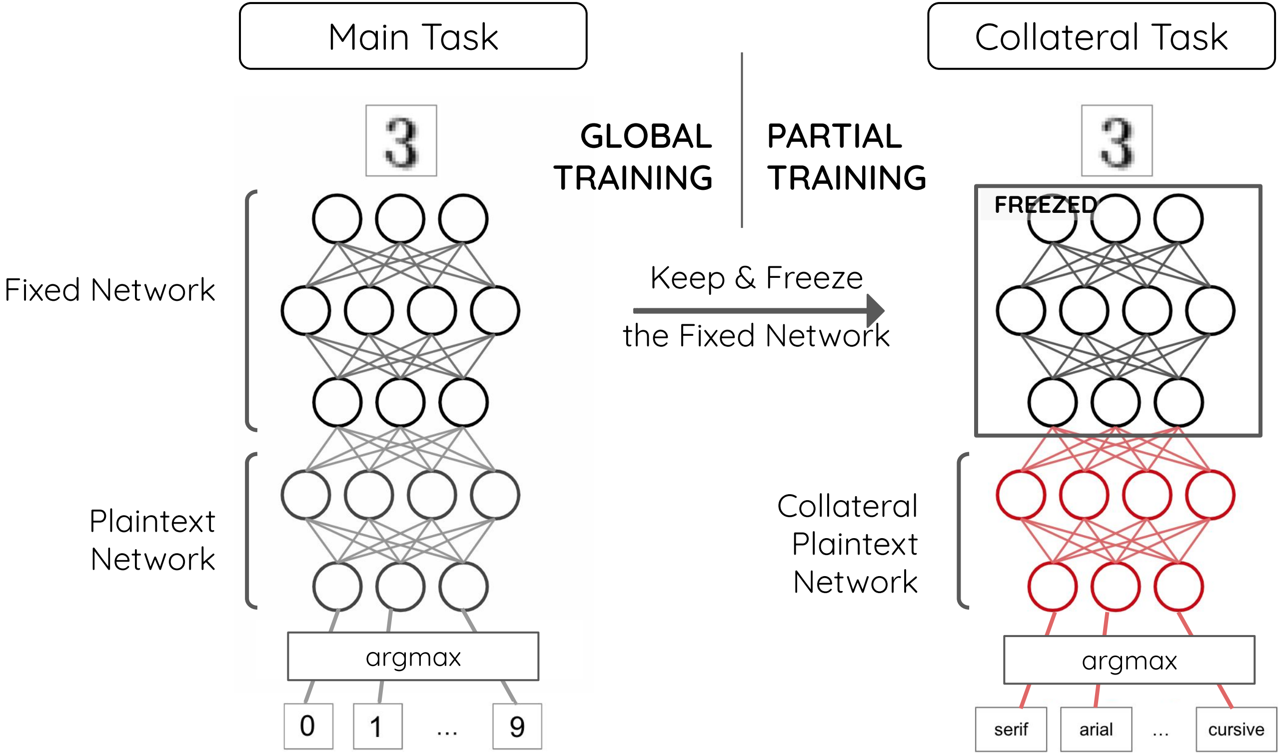
GLOBAL TRAINING



Collateral Learning [RPB+19]



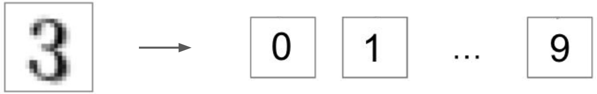
Collateral Learning [RPB+19]



Collateral Learning [RPB+19]

The collateral task achieves a high accuracy

Main Task



Random baseline	10%
Accuracy with a CNN	**99,1%**

Collateral Task

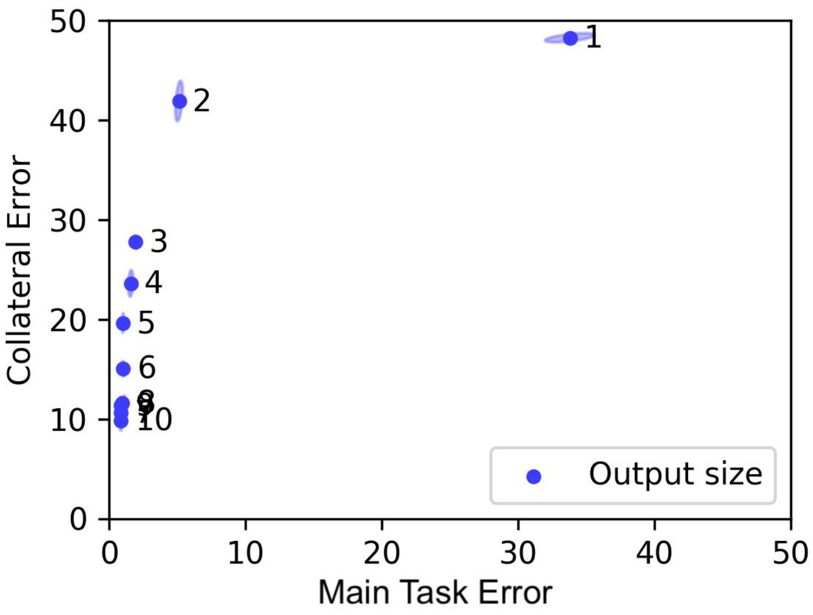
Random baseline	20%
Accuracy with a CNN	**93.5%**

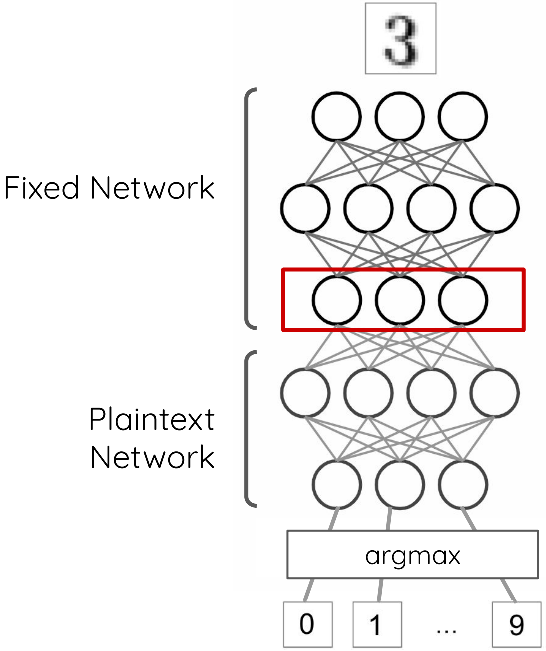
Implication of Collateral Learning

What solution can we propose?

Collateral learning [RPB+19]

Mitigation #1: reducing the fixed network output



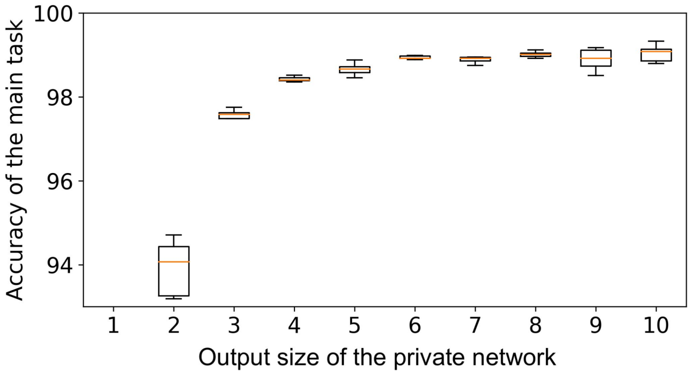


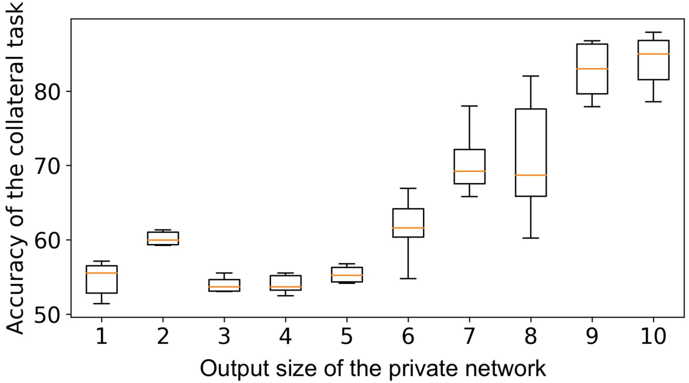
Collateral learning [RPB+19]

Mitigation #2: Adversarial learning [DDSV04] against a simulated adversary

Perform a joint optimisation using the loss that we imagine an adversary would try to optimize:

$$\boxed{L_{joint} = L_{main} - \alpha \cdot L_{collateral}}$$





Collateral learning [RPB+19]

Mitigation #2: Adversarial learning against a simulated adversary

Accuracy of an attacker to distinguish between 2 fonts, using different classifiers.

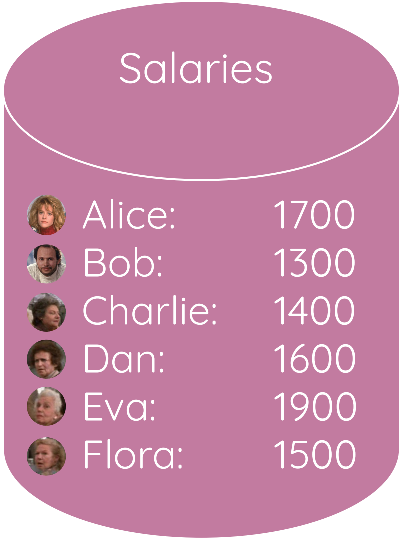
Linear Ridge Regression	$53.5 \pm 0.5\%$
Logistic Regression	$52.5 \pm 0.6\%$
Quad. Discriminant Analysis	$54.9 \pm 0.3\%$
SVM (RBF kernel)	$57.9 \pm 0.4\%$
Gaussian Process Classifier	$53.8 \pm 0.3\%$
Gaussian Naive Bayes	$53.2 \pm 0.5\%$
K-Neighbors Classifier	$58.1 \pm 0.7\%$
Decision Tree Classifier	$56.8 \pm 0.4\%$
Random Forest Classifier	$58.9 \pm 0.2\%$
Gradient Boosting Classifier	$58.9 \pm 0.2\%$

(*Baseline:* 50%)

2

Differential Privacy

Intuition



Mean with bob: 1567
Mean without bob: 1620

Deduction of bob's salary: 1567 * 6 - 1620 * 5 = 1300

Objective: to ensure that statistical analysis does not compromise the privacy of individuals

Analyse: function(data) *example : mean of salaries*

Perfect confidentiality: the result of the query is indistinguishable if you add or remove a single individual in the dataset

If you add noise to the calculation, it becomes difficult to determine Bob's salary or even if Bob is part of the dataset

(ε,δ)-Differential Privacy [DMNS06]

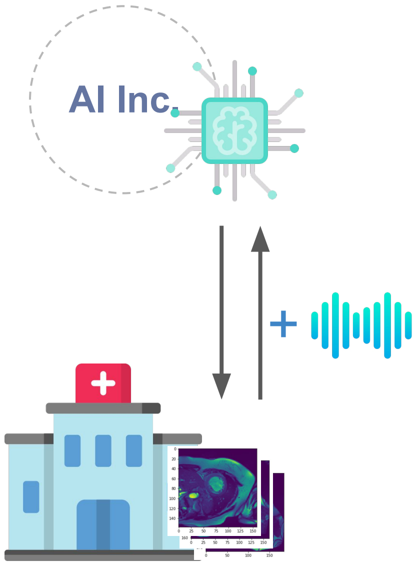
A randomized algorithm $\mathcal{A}$ satisfies (ε,δ)-differential privacy if for any datasets $\mathcal{D}$ and $\mathcal{D}'$ only differing in one item, we have:

$$\mathrm{P}[\mathcal{A}(\mathcal{D}) \in S] \leq e^{\varepsilon}\mathrm{P}[\mathcal{A}(\mathcal{D}') \in S] + \delta$$

Privacy budget (ε,δ) “small” => increased privacy

Differential Privacy

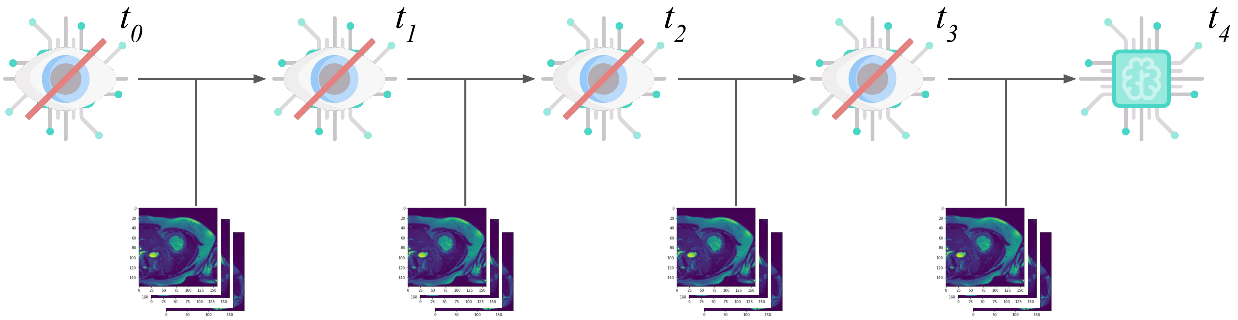
Differentially Private Stochastic Gradient Descent (DP-SGD) [BST14]



- Works by adding Gaussian noise on the model updates
- Limited access to data for a given privacy budget
- More noise: better privacy budget but worse model utility => **trade-off**

Stochastic Gradient Langevin Dynamics [RBP22]

Our working hypothesis: DP-SGD assumes that the model is public at each iteration. If we can hide the model during training and only disclose it at the end, less information should leak.



Stochastic Gradient Langevin Dynamics [RBP22]

Extension of [CYS21] which leverages Langevin diffusion to achieve:

- Exponentially fast convergence of the privacy (instead of $\sqrt{K}$)
- Under smooth and strongly convex objectives
- For full gradient descent

Our contribution: a stochastic version more practical for ML users [RBP22]

Differential Privacy <> Contextual Integrity

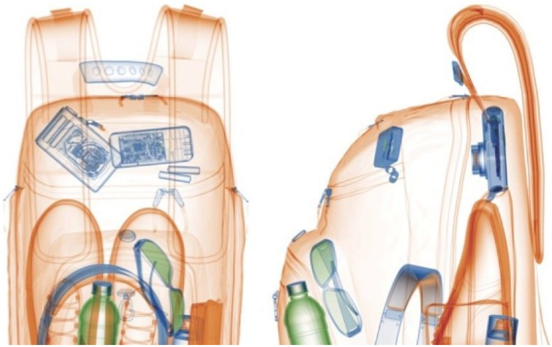
- Personal data used for training should not be directly exposed
- ML models should not disclose private training data

  The model is sanitized to prevent attacks like model inversion or membership attack
- ML models should not be disseminated or exposed

  The model is still shared directly to the data owners => IP issue

Differential Privacy <> Contextual Integrity

Example of risk on the model: Airport X-ray security scan



3

Encrypted Computation

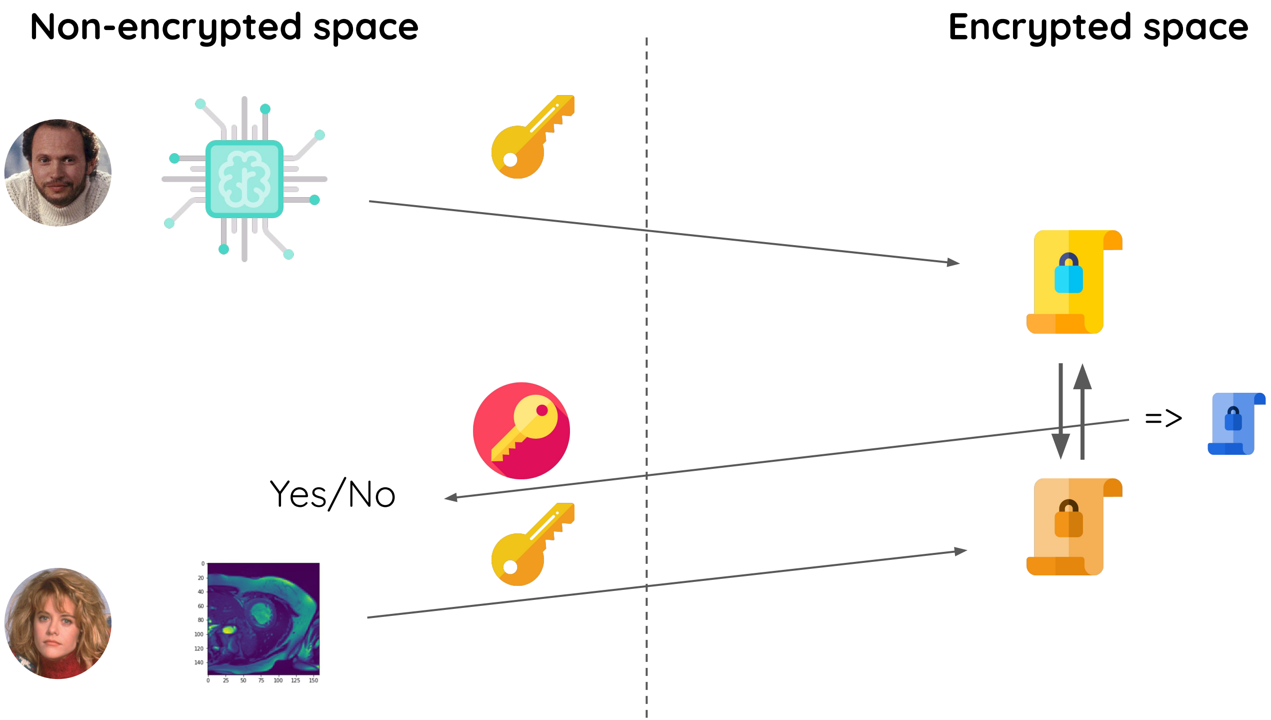
Overview of available methods

The ML model should be “encrypted”, meaning **usable but not visible**

Several methods:

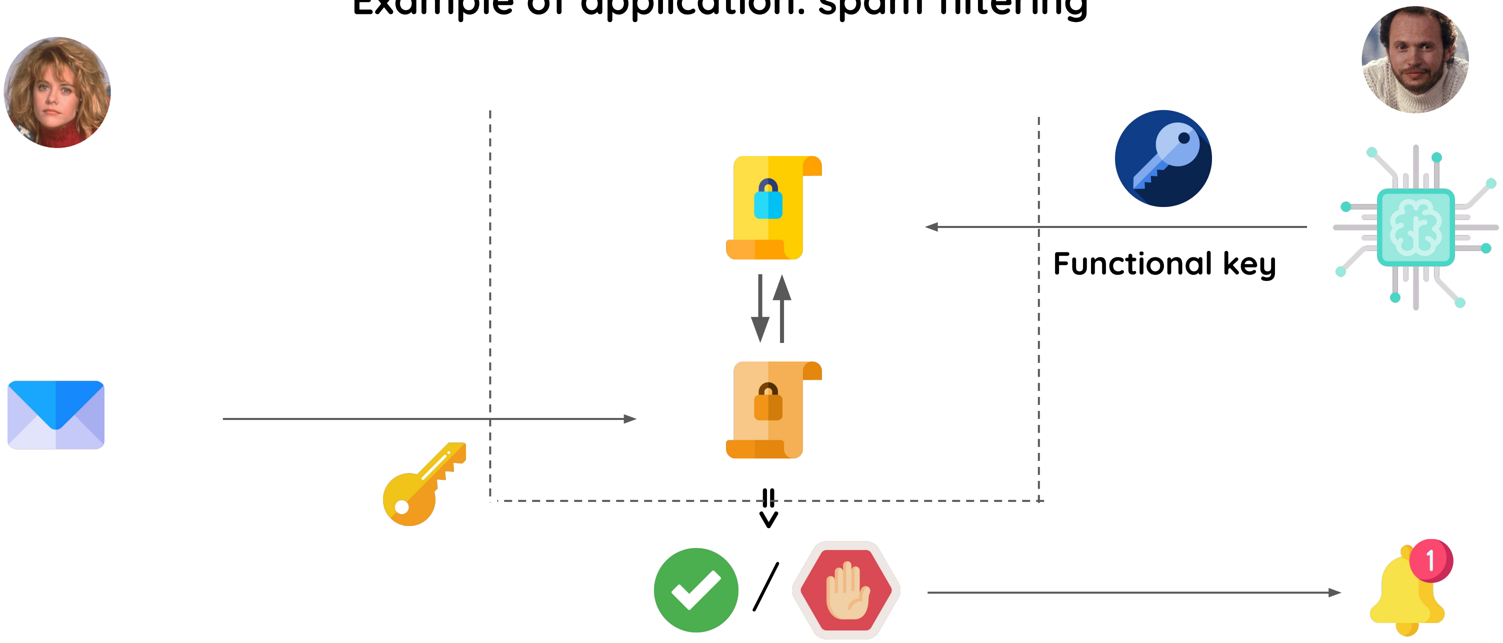
- Homomorphic Encryption
- Functional Encryption
- Secure Multi-Party Computation

Homomorphic Encryption



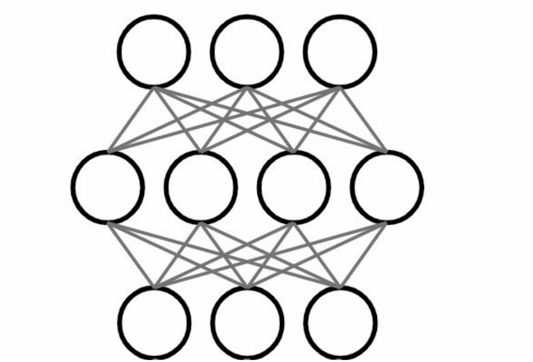
Functional Encryption

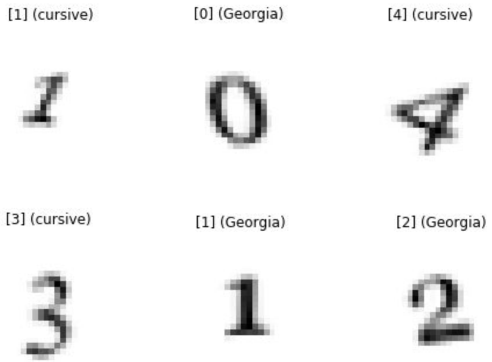
Example of application: spam filtering



A Quadratic Functional Encryption Scheme

Our contribution: Extension of [DGP18], the functional scheme can be viewed as a neural network with one hidden layer and a square activation. [RPB+19]



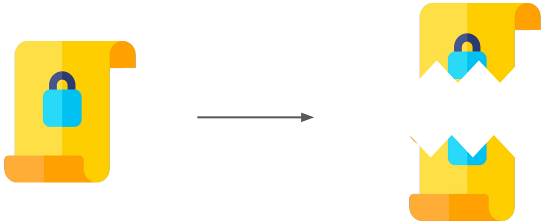


Secure Multi-Party Computation

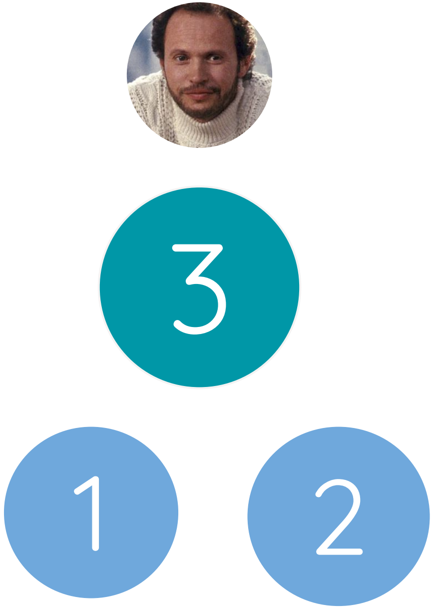
Definition

The set of methods for parties to jointly compute a function over their inputs while keeping those inputs private.

Focus on additive secret-sharing:

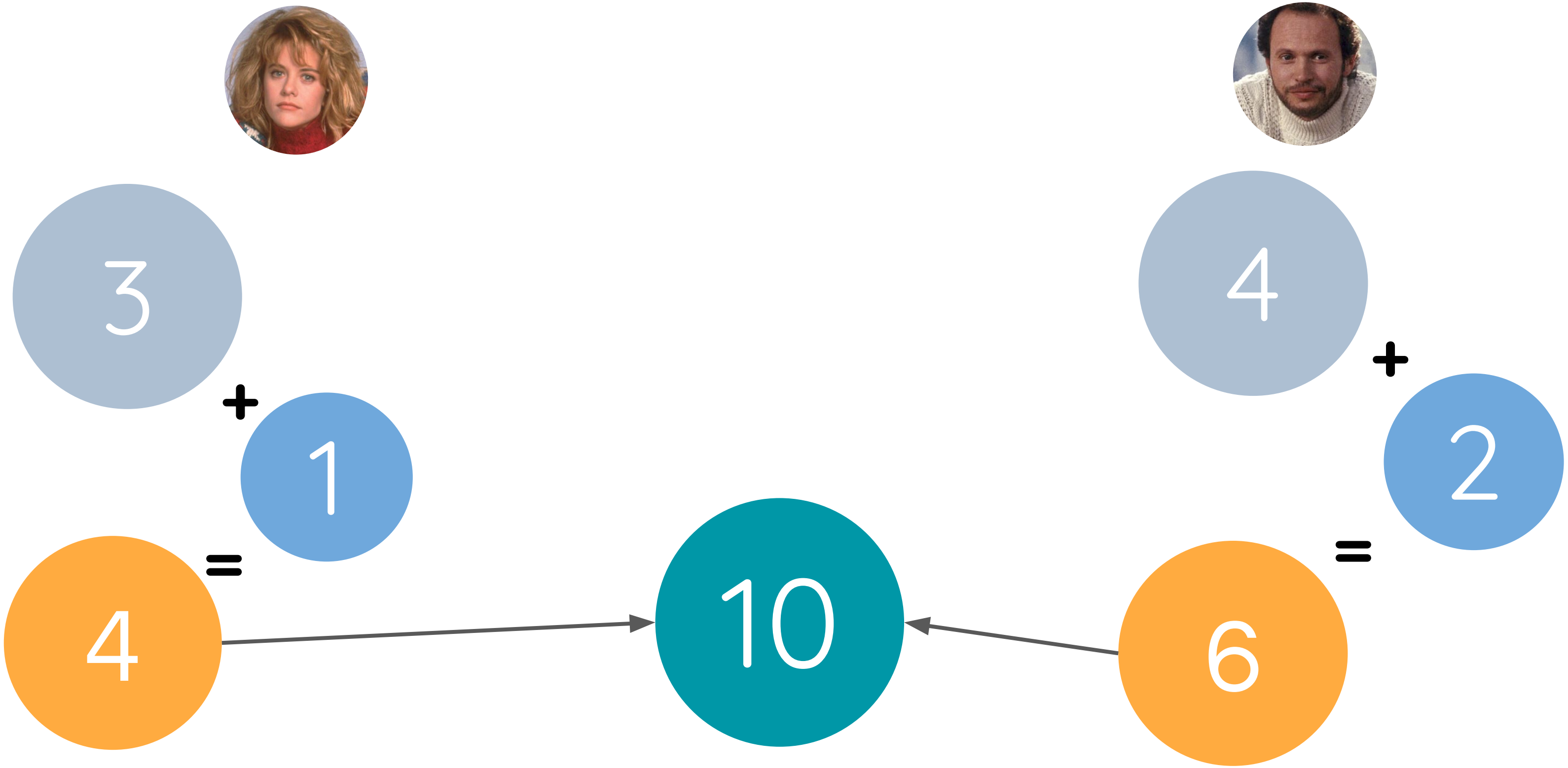


Additive Secret Sharing



Additive Secret Sharing

Additive Secret Sharing



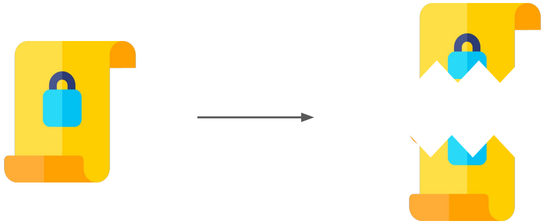
Additive Secret Sharing

Secret sharing: no single party can reconstruct sensitive data alone

Shared governance: data can only be used or decrypted if everyone agrees

Additive Secret Sharing

Data and models can be secret shared in the same way



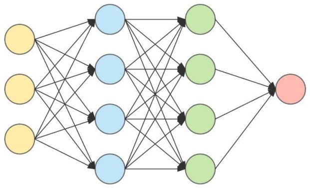
Operations needed for ML models

- Addition (already seen)
- Multiplication & matrix multiplication (not very difficult)
- **Comparisons**

Additive Secret Sharing

$$x \qquad\qquad f \quad \longrightarrow \quad f(x)$$

Example $f: x \mapsto x \geq 0$



Function Secret Sharing [BGI15]

A different perspective

$$[\![x]\!]_0$$

$$\Pi_f$$

$$[\![x]\!]_1$$

Additive Secret Sharing

$$[\![x]\!]_0 \qquad \Pi_f$$

$$[\![x]\!]_1 \qquad \Pi_f$$

Additive Secret Sharing

$$[\![x]\!]_0 \quad \Pi_f \quad \longrightarrow \quad [\![f(x)]\!]_0$$

$$\Sigma \quad f(x)$$

$$[\![x]\!]_1 \quad \Pi_f \quad \longrightarrow \quad [\![f(x)]\!]_1$$

Function Secret Sharing

A different perspective

$$[\![f]\!]_0$$

$$x$$

$$[\![f]\!]_1$$

Function Secret Sharing

A different perspective

$$x \nearrow x \quad [\![f]\!]_0 \longrightarrow [\![f]\!]_0(x)$$

$$x \searrow x \quad [\![f]\!]_1 \longrightarrow [\![f]\!]_1(x)$$

$$\Sigma \quad f(x)$$

Example $f: x \mapsto x \geq \alpha$

Function Secret Sharing

A different perspective

$$\llbracket y \rrbracket_0 + \llbracket \alpha \rrbracket_0 \qquad\qquad x \qquad\qquad \llbracket f \rrbracket_0 \longrightarrow \llbracket f \rrbracket_0(x)$$

$$\Sigma \quad x \qquad\qquad\qquad\qquad \Sigma \quad f(x)$$

$$\llbracket y \rrbracket_1 + \llbracket \alpha \rrbracket_1 \qquad\qquad x \qquad\qquad \llbracket f \rrbracket_1 \longrightarrow \llbracket f \rrbracket_1(x)$$

$x \geq \alpha \Rightarrow y \geq 0$ with $x = y + \alpha$

Secure Comparison with Function Secret Sharing

[RTPB22]

Reminder on the binary notation:

$$x = x_1 x_2 \dots x_n = \sum 2^{n-k} \cdot x^k$$

Example:

$$n = 3, \qquad x = 010_2 = 4 \cdot 0 + 2 \cdot 1 + 1 \cdot 0 = 2$$

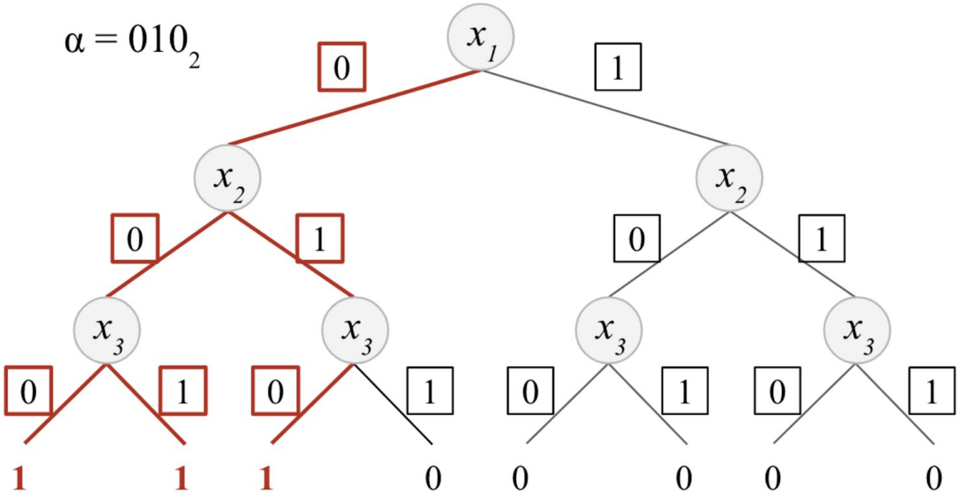
Using the bit notation, x and α write:

$$x = x_1 x_2 \dots x_n, \quad \alpha = \alpha_1 \alpha_2 \dots \alpha_n$$

Secure Comparison with Function Secret Sharing

[RTPB22]

Example with $n = 3$ of $x \leq \alpha$



Bit per bit private comparison

For each k from 1 to n

- if $x_k >$, then $x \leq \alpha$ is false
- if $x_k <$, then $x \leq \alpha$ is true
- if $x_k =$, then we need to compare the bit $k+1$ to decide

Secure Comparison with Function Secret Sharing

Correctness & Security of our protocol [RTPB22]

- Honest but curious, 2 party computation with trusted dealer
- Small error rate (that can be avoided with extra computation) => [BCG+21]

Secure Comparison with Function Secret Sharing [RTPB22]

Why FSS is promising for private ML

Pros

- Enjoys the efficiency of MPC protocols (only light cryptographic primitives)
- Can be run on GPUs
- Considerably reduces the number of communication rounds compared to other MPC protocols: 1 for private comparison

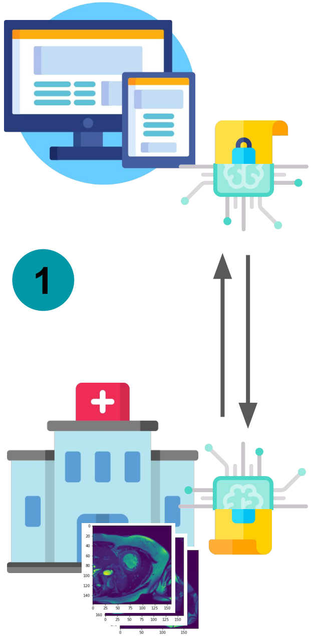
Cons

- Requires big preprocessing keys (correlated random strings)
  size of key of a 32 bits integer comparison ~ 32 λ bits

Example: 224x224 image through ResNet-18
3311620 comparisons => ~1.7 Go per key

Machine Learning with Function Secret Sharing

AriaNN - Private training using Function Secret Sharing [RTPB22]

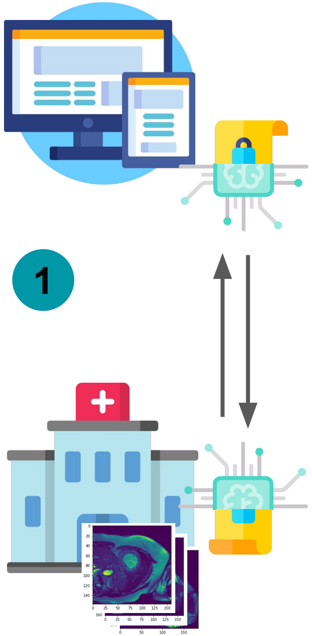


Dataset	Model	Accuracy (%)	Time / epoch (h)
28x28 MNIST	Linear	98.0	0.8
28x28 MNIST	LeNet	99.2	4.2

github.com/OpenMined/sycret

Machine Learning with Function Secret Sharing

AriaNN - Private evaluation using Function Secret Sharing [RTPB22]



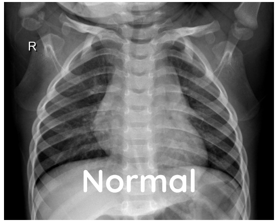
Dataset	Model	Time over LAN using CPU (s)
32x32 CIFAR10	AlexNet	0.15
32x32 CIFAR10	VGG16	1.75
224x224 Imagenet	ResNet18	19.9

Time over LAN using GPU (s)
0.078
1.55
13.9

github.com/OpenMined/sycret

Machine Learning with Function Secret Sharing

Study using AriaNN for private evaluation [KZPR+21]



End-to-end privacy preserving deep learning on multi-institutional medical imaging

Function Secret Sharing used for **Secure inference-as-a-service**, a scenario where latency matters

To evaluate privately a neural network at expert level accuracy

Function Secret Sharing <> Contextual Integrity

- Personal data used for training should not be directly exposed
- ML models should not disclose private training data
  If differential privacy is used
- ML models should not be disseminated or exposed

Additional remarks

- Only honest but curious security
- Model not visible during training: condition for the DP methods exposed
  => allows powerful combinations

4

Conclusion

Conclusions & Perspectives

- Impact on the run-time or accuracy: the use of PETs depends on the context, hence the concept of contextual integrity
- Cross domain research: a challenge and an opportunity
- Need for user-friendly open-source implementations to accelerate awareness
- Real life data needs intensive cleaning and structuration to be useable, this can't be done once data is encrypted
- Also a social, political, legal and economic challenge

Acknowledgments

Directors
David Pointcheval
Francis Bach

Jury
Aurélien Bellet
Yuval Ishai
Renaud Sirdey
Mariya Georgieva
Laurent Massoulié
Jonathan Passerat-Palmbach

DI ENS
Michel Abdalla
Léonard Assouline
Hugo Beguinet
Céline Chevalier
Baptiste Cottier
Guillaume Gette
Lénaïck Gouriou
Paul Hermouet
Jianwei Li
Brice Minaud
Phong Nguyen
Thanh-Huyen Nguyen
Ky Nguyen
Paola de Perthuis
Duong-Hieu Phan
Michael Reichle
Robert Schaedlich
Hugo Senet
Quentin Sopheap
Quoc Huy Vu
Hoeteck Wee
Xiaxi Ye

DI ENS (Former students, post-docs)
Balthazar Bauer
Jérémy Chotard
Geoffroy Couteau
Aurélien Dupin
Pooya Farshim
Chloé Hébant
Louiza Khati
Romain Gay
Anca Nitulescu
Michele Orrù
Mélissa Rossi
Azam Soleimanian

DI ENS / INRIA (Support)
Lise-Marie Bivard
Linda Boulevart
Martine Girardot
Sophie Jaudon
Valerie Mongiat
Nathalie Gaudechoux
Jacques Beigbeder
Ludovic Ricardou

OpenMined
Andrew Trask
Robert Wagner
Jason Mancuso
Morten Dahl
Patrick Cason
Pierre Tholoniat
George Muraru
Tudor Cebere
Rasswanth S
Hrishikesh Kamath
Arturo Marquez
Yugandhar Tripathi
S P Sharan
Nicolas Remerscheid
Jason Paumier
Muhammed Abogazia
Alan Aboudib
Ayoub Benaissa
Sukhad Joshi
and many other

Arkhn
Corneliu Malciu
Emeric Lemaire
Alexis Thual
Margaux Françoise
Elsie Hoffet
Jason Paumier
Simon Vadée
Nicolas Riss
Céline Thiriez
Ilan Zana
Jean-Baptiste Laval
Lucile Saulnier
Marion Bladier
Melina Chamayou
Pierrick Bellut
Clément Jumel
Teva Riou
Valentin Matton
Sophie Peltier
Shamira Ruesche
Horace Blanc
Michel Giboreau
Maud Bisserier
Louise Naudin
Elena Mylona
Louise Garnier
Yasmine Marfoq
Manon Chatal
Thierry Chanet
Abdellah Lidghi
Charlotte Delhomme
Ronan Sy
Monon Ongena
Simon Meoni
Florian Sebal

My friends and family ❤️

Bibliography (our contributions)

[**R**TD+18] Théo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. In *NeurIPS 2018 Workshop Privacy-Preserving Machine Learning*, 2018.

[**R**PB+19] Théo Ryffel, David Pointcheval, Francis Bach, Edouard Dufour-Sans, and Romain Gay. Partially encrypted deep learning using functional encryption. *Advances in Neural Information Processing Systems*, 32, 2019

[KZP**R**+21] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jung- mann, Marc-Matthias Steinborn, et al. End-to-end privacy preserving deep learn- ing on multi-institutional medical imaging. *Nature Machine Intelligence*, pages 1–12, 2021.

[**R**BP22] Théo Ryffel, Francis Bach, and David Pointcheval. Differential privacy guarantees for stochastic gradient langevin dynamics, ArXiv 2022.

[**R**TPB22] Théo Ryffel, Pierre Tholoniat, David Pointcheval, and Francis Bach. Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. *Pro- ceedings on Privacy Enhancing Technologies*, 2022.

Bibliography

[BCG+21] Elette Boyle, Nishanth Chandran, Niv Gilboa, Divya Gupta, Yuval Ishai, Nishant Kumar, and Mayank Rathee. Function secret sharing for mixed-mode and fixed- point secure computation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 871–900. Springer, 2021.

[BGI15] Elette Boyle, Niv Gilboa, and Yuval Ishai. Function secret sharing. In *Annual International Conference on the Theory and Applications of Cryptographic Tech- niques*, pages 337–367. Springer, 2015.

[CYS21] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 2021.

[DDSV04] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99-108. 2004.

[DGP18] Edouard Dufour-Sans, Romain Gay, and David Pointcheval. Reading in the Dark: Classifying Encrypted Digits with Functional Encryption. ArXiv 2018

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[GBDM20] onas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. In- verting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Nis04] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Member- ship inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.