Federated Learning Systems: Towards Effective and Efficient Machine Learning Systems on Data Silos

Bingsheng He
School of Computing
hebs@comp.nus.edu.sg
Joint work with Qinbin Li, Zhaomin Wu, Sixu Hu, Naibo Wu, Yiqun Diao, Zeyi Wen (HKUST GZ), Quan Chen (SJTU), Dawn Song (UCB)

Outline

- **Introduction and Motivation**
- **Design principles of a federated learning system**
- **The vision and roadmap**
- **Our work**
  - A federated tree based system
  - Benchmarks, One-shot learning, etc…
- **Summary and on-going work**

Data Breach is NOT an Exception

TechRepublic. SEARCH IT Policy Downloads 5G Developer Top DaaS providers Security More Newsletters

Facebook data privacy scandal: A cheat sheet

by James Sanders | Dan Patterson in Security on July 24, 2019, 8:52 AM PST

Read about the saga of Facebook's failures in ensuring privacy for user data, including how it relates to Cambridge Analytica, the GDPR, the Brexit campaign, and the 2016 US presidential election.

Millions of Facebook user records exposed in data breach

share Save 2

T-Mobile just told some customers that there was a data breach of their personal information. Here's how to check if you're affected.

Aaron Holmes, Business Insider US November 22, 2019

Macy's tells customers their payment information may have been stolen by hackers

Shoshy Ciment, Business Insider US November 19, 2019

At Least 80% of Shopping Apps Leak Users' Data. Here's How to Protect Yourself

Singtel, Ninja Van among 5 firms fined $177,000 for leaking customers' passport details, signatures, birth certs and more

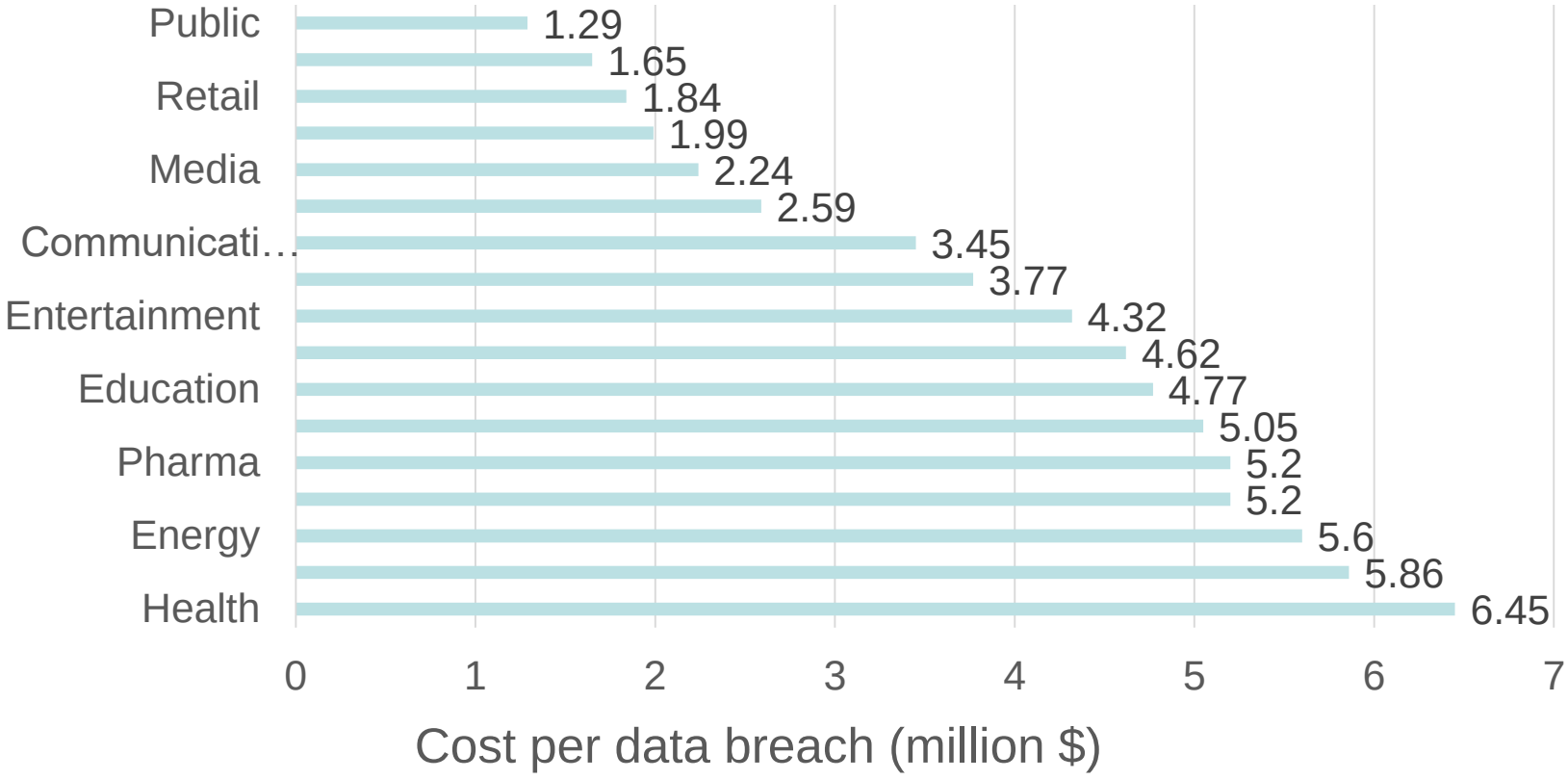
Rachel Genevieve Chia November 6, 2

Passport data of 30 million Malindo and Lion Air customers leaked: here's what we know

Rachel Genevieve Chia September 19, 2019

The Cost of Data Breach

➢ Average Cost of a data breach by industry



https://digitalguardian.com/blog/whats-cost-data-breach-2019

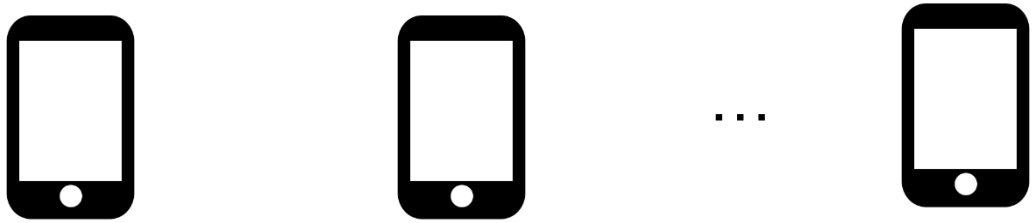
GDPR – General Data Protection Regulation

- **Right to be forgotten:** Be able to ask companies or platforms to delete your data.
- **Specific permission:** Unless or until you give permission to an app or website to use your details in a specific way, they can't use it for any other purpose or sell it to third parties.
- **Data portability:** Data subjects have the right to copy, transfer, or move personal data to a different company.
- **Privacy by design:** when you sign up for a service, you should not be asked for data that is not directly needed or relevant for the purposes of using that app or service.

Scenario of "Data Islands"

- Hospitals may have very different patients, but the patient records have similar measurements.

- The hospital, bank, and e-commerce company have different aspects of user information, but they can have the same users.



- User data are stored on their devices individually.

Big Challenge: Machine learning algorithms are data hungry.

Can We Learn a Model with Preserving the Data Privacy?

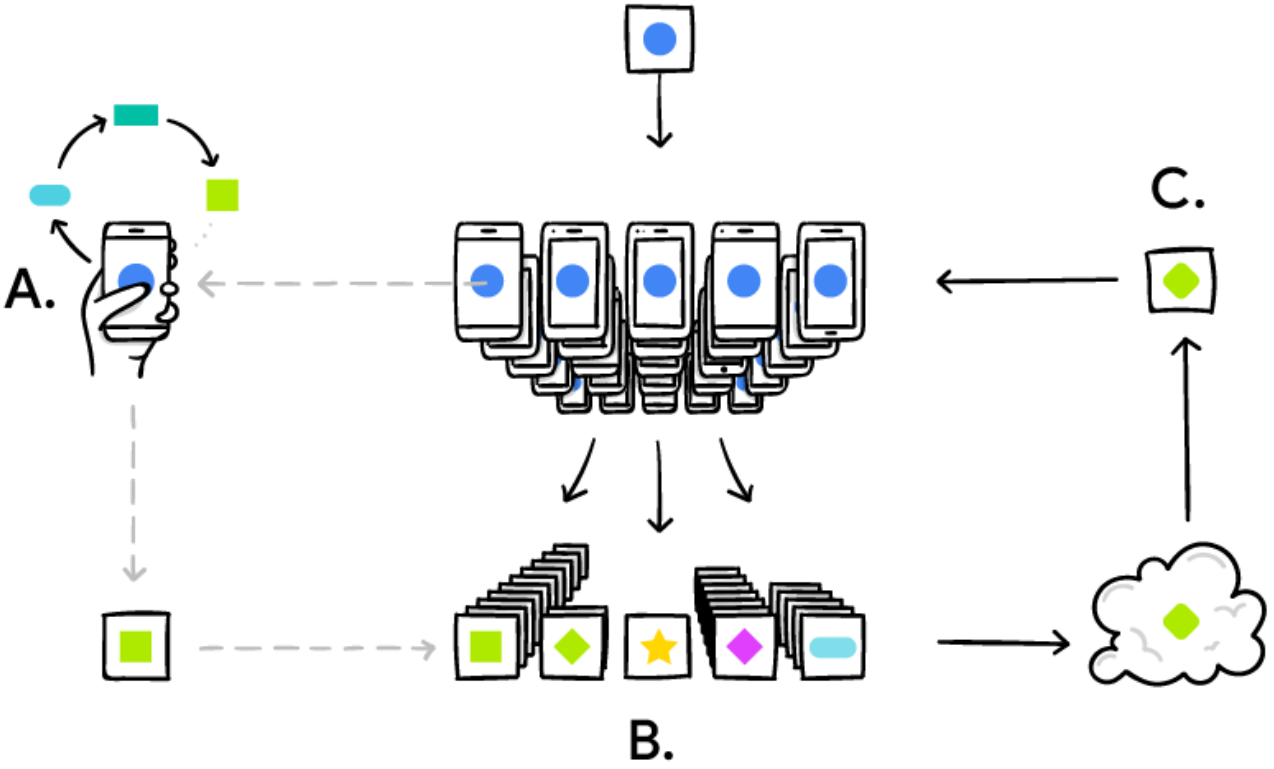
One Promising Paradigm: Federated Learning

- **According to Wiki:** Federated learning is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, **without exchanging their data samples**.

- **An alternative (more precise) definition:** enable the collaborative training of machine learning models among different organizations **under the privacy restrictions.**

Case Study

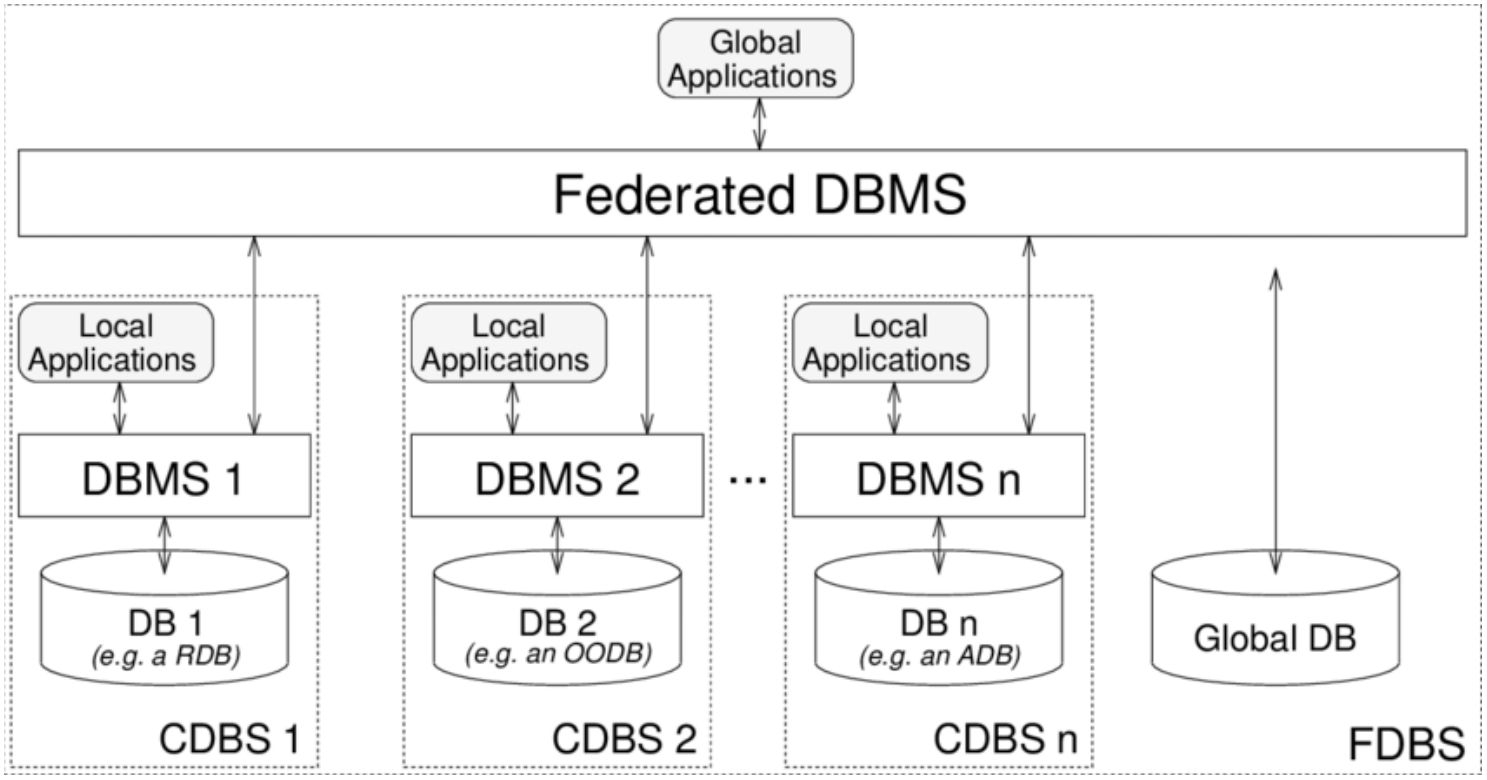
➢Keyboard Word Suggestion



Can We Build “TensorFlow/PyTorch” for Federated Learning?

Federated Database Systems (FDBSs) [2]

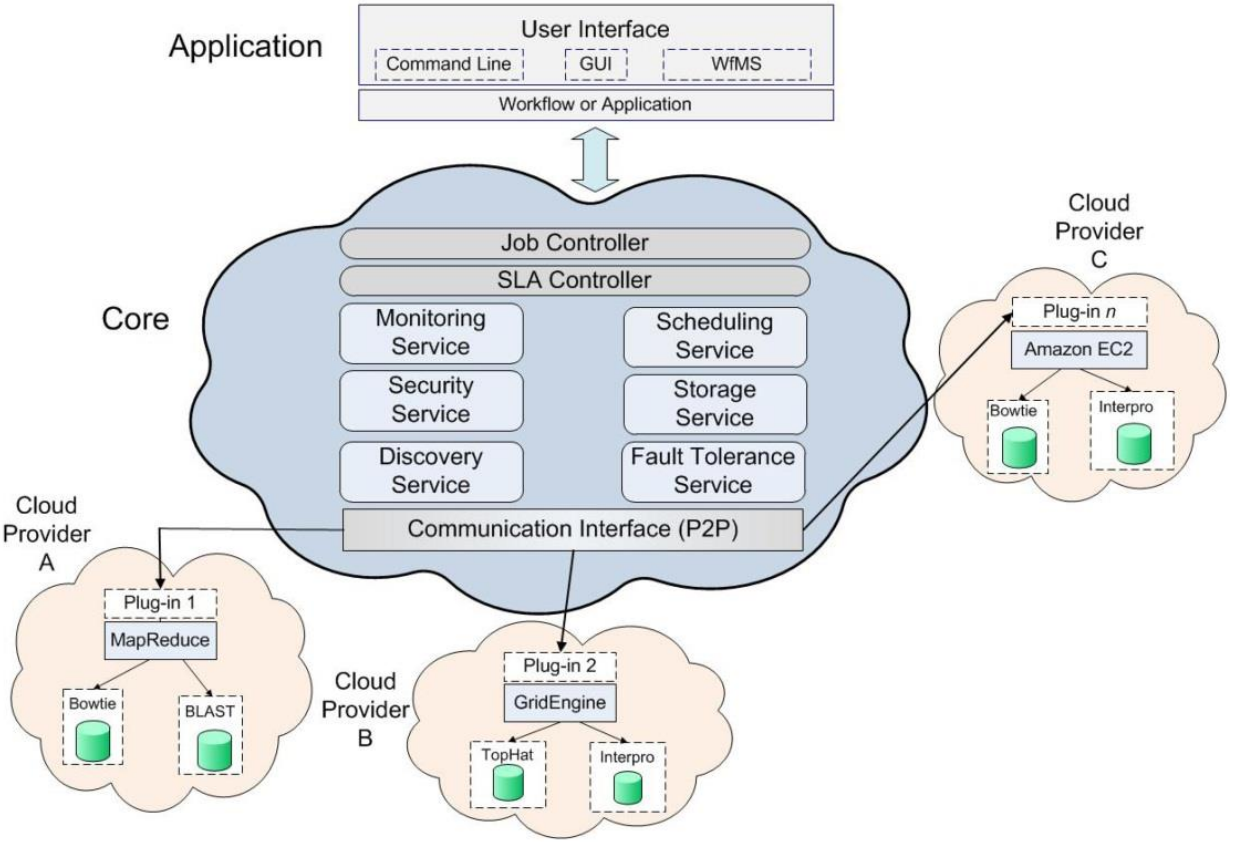
- **FDBS has been an effective way of connecting multiple databases.**



[2] Sheth, Amit P., and James A. Larson. "Federated database systems for managing distributed, heterogeneous, and autonomous databases." ACM Computing Surveys (CSUR) 22.3 (1990): 183-236.

Federated Cloud (FC)[3]

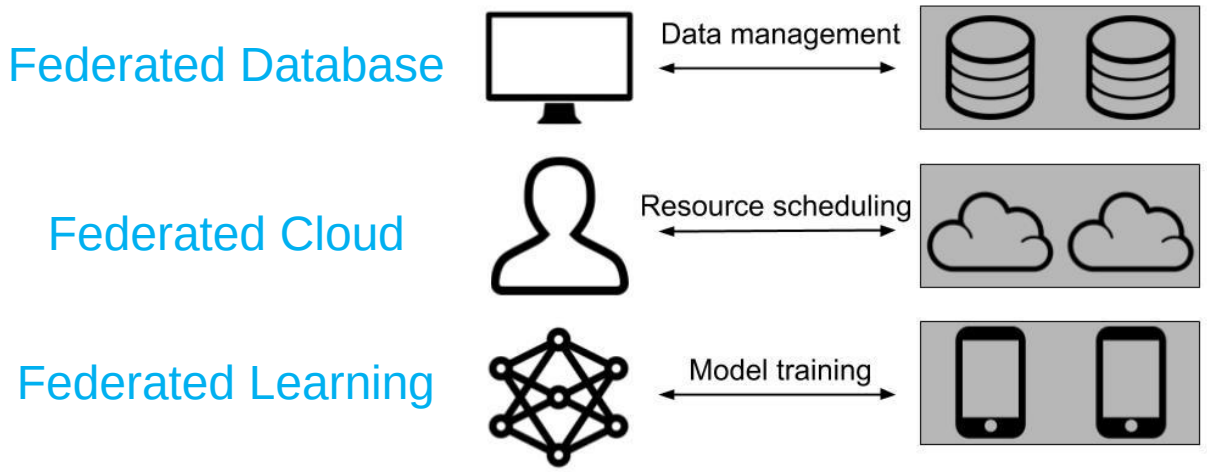
- **Resource migration and resource redundancy.**



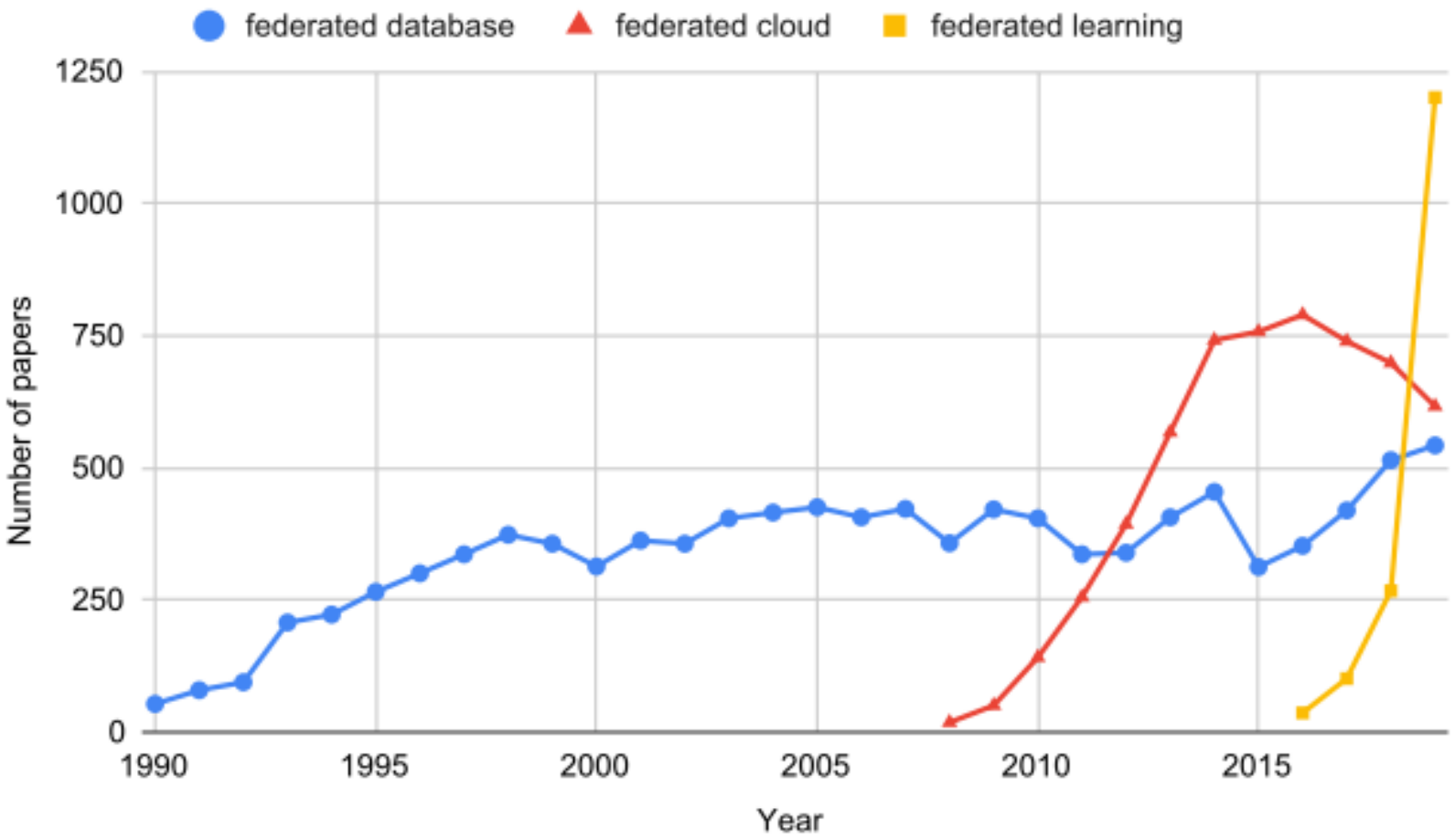
[3] Kurze, Tobias, et al. "Cloud federation." Cloud Computing 2011 (2011): 32-38.

Analogy among Federated Systems

- **Similarities:**
  - Heterogeneity
  - Autonomy
- **Differences:**
  - Federated learning systems focus on the secure model training under the privacy restrictions.

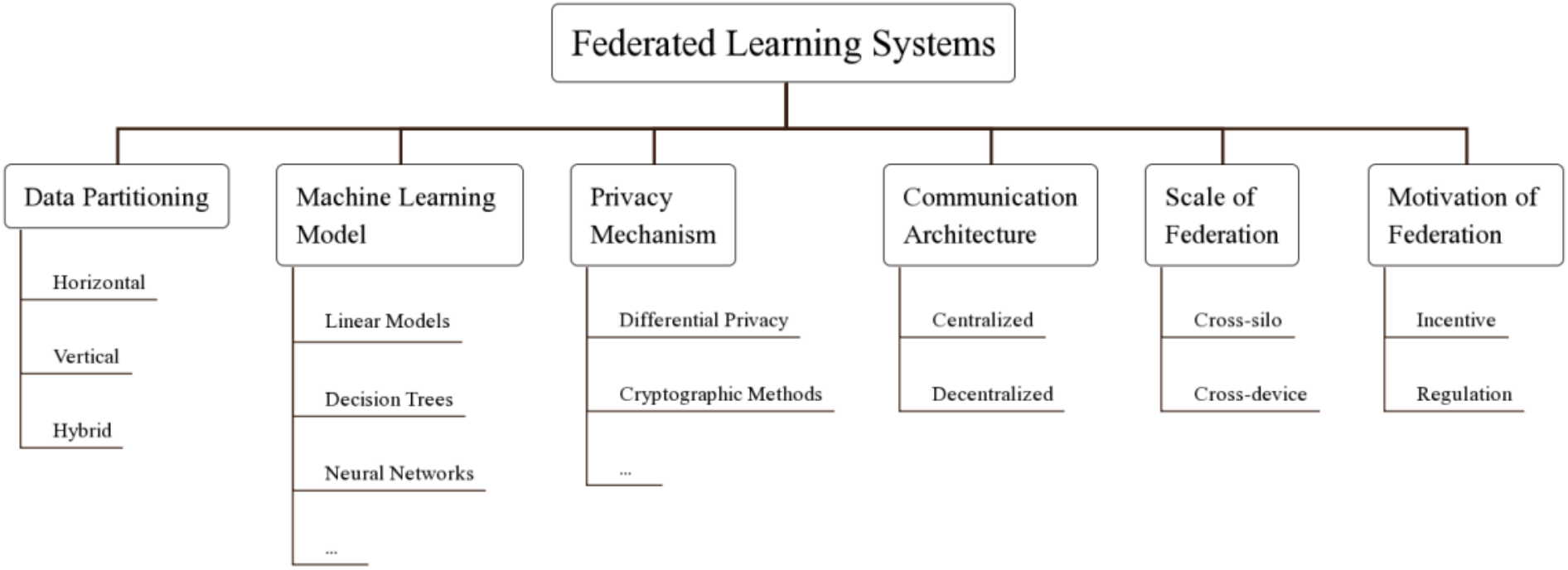


Trends of Research on Federated Systems



GoogleScholar: The number of related papers on “federated database”, “federated cloud”, and “federated learning”

Taxonomy



Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wu, Bingsheng He. "Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection." IEEE TKDE 2021.

Existing Systems

Federated Learning Systems	data partition	model implementation	privacy mechanism	communication architecture
Google TensorFlow Federated (TFF) [5], [157]	horizontal	LM, NN	CM, DP, MA	centralized
PySyft [6]				
PhotoLabeller[3]		NN	MA	
Federated AI Technology Enabler (FATE)[4]	hybrid	LM, DT, NN	CM	distributed

- **Most of them are in their early stage**
  - Functionality: limited support for data partitioning /model/privacy definition
  - Efficiency: Cryptographic methods can be costly.
  - Scalability and communication
  - Incentive?

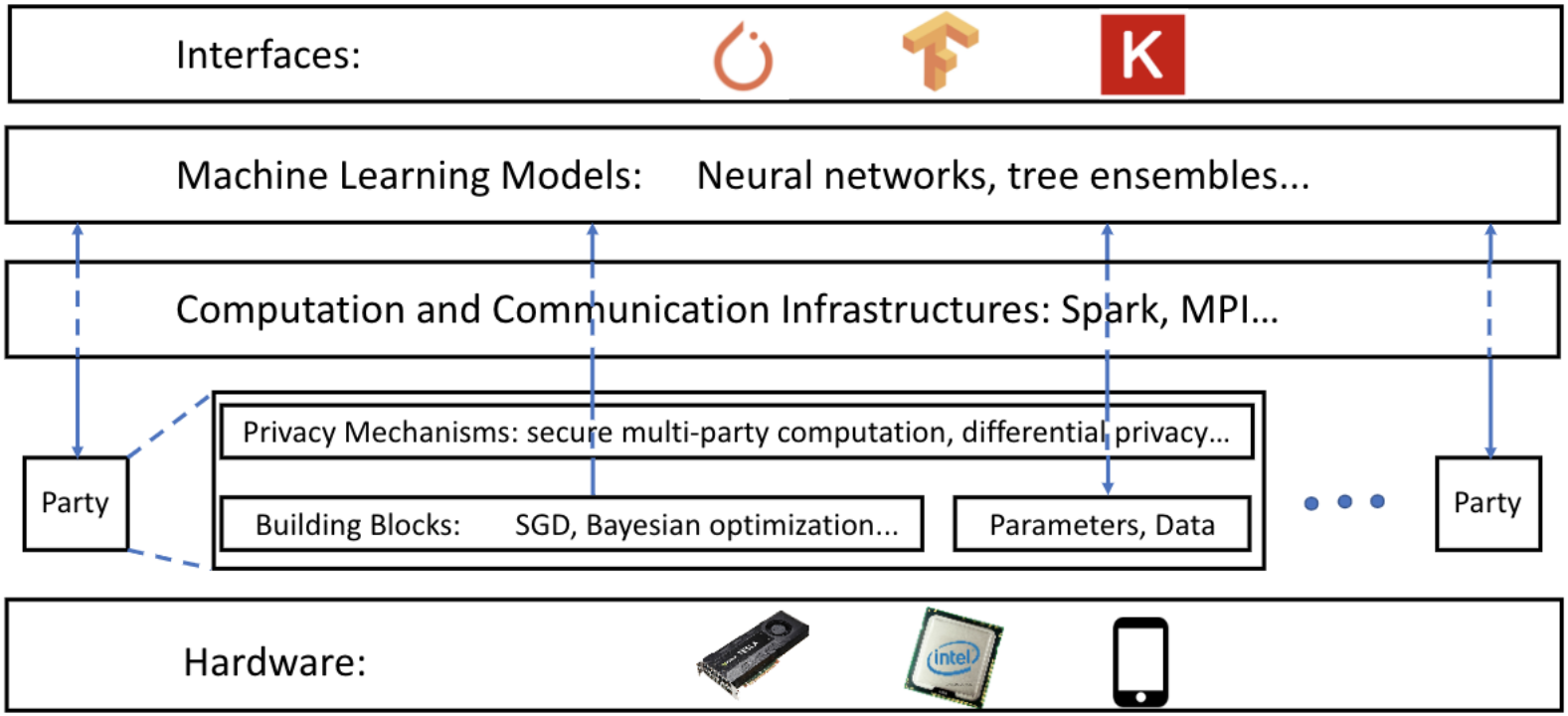
Future Research Directions

- Dynamic scheduling: Unfixed number of parties.
- Diverse privacy restrictions: Heterogeneous differential privacy[8].
- Intelligent benefits: Incentive mechanisms.
- Benchmark: LEAF[9].
- System architecture: Federated averaging[10].
- Programming abstractions and runtime
- Data life cycles: Data creation, storage, use, share, archive and destroy.
- Efficiency

[8] Alaggan, Mohammad, Sébastien Gambs, and Anne-Marie Kermarrec. "Heterogeneous differential privacy." arXiv preprint arXiv:1504.06998 (2015).
[9] Caldas, Sebastian, et al. "Leaf: A benchmark for federated settings." arXiv 2018.
[10] McMahan, H. Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." arXiv 2016.

ThunderFL: Practical and Efficient Federated Learning

- A general framework for differentially private machine learning algorithms.
- A federated learning system.

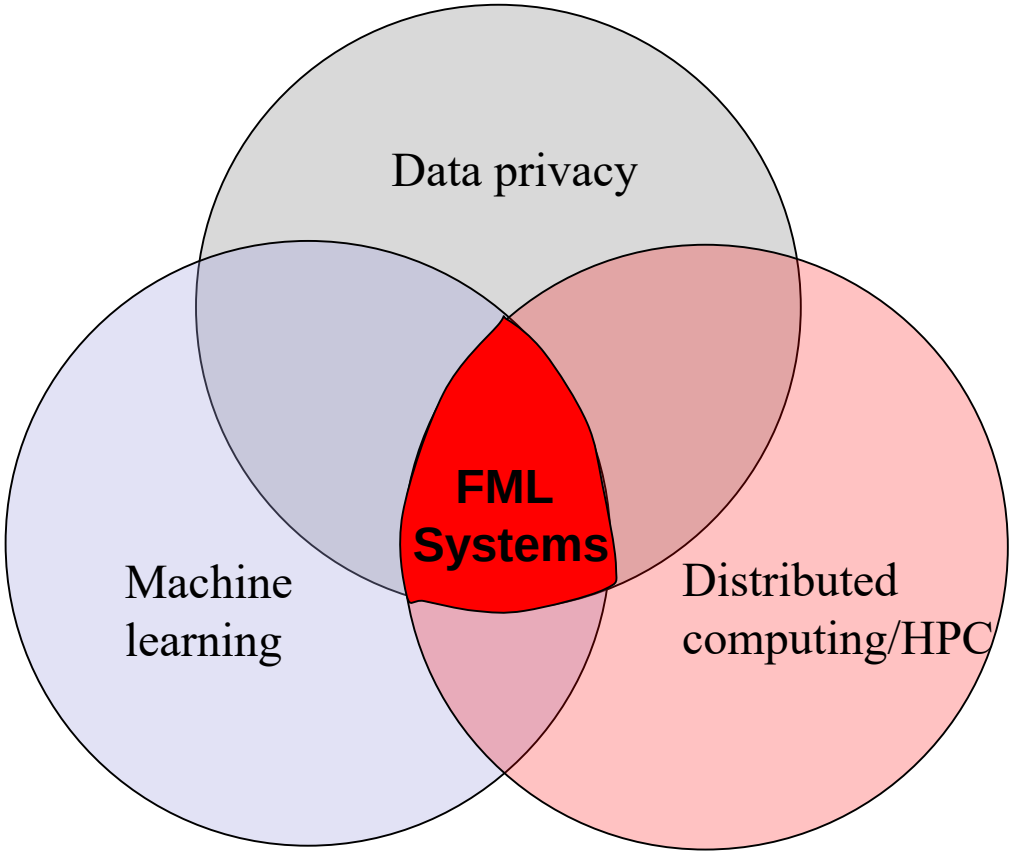


A Roadmap Towards "PyTorch" in FL

- **1.0: more on functionality**
  - A public benchmark and reference implementation
  - Programming abstractions and building blocks
  - FedTree: Secure and Efficient Tree FML systems
- **2.0: more on incentive designs and efficiency**
  - Incentive designs (~blockchain systems)
  - Efficiency (hardware accelerations)
- **Beyond: new architectures and new systems**
  - End-to-end system architectures from user privacy to system
  - New secure and efficient hardware…

Let me know if you are interested in this journey.

System Research for FML Systems



Our On-going Work

- **A tree based federated learning system**
  - Practical Federated Gradient Boosting Decision Trees [AAAI 2020]
  - Privacy-Preserving Gradient Boosting Decision Trees [AAAI 2020]
- **Benchmarks, Characterization and Implication for Federated Learning Systems:** OARF [ACM TIST2021], NIID-Bench [ICDE2022]
- **Single-shot FML systems** [AAAI2021, TBD2022]
- **Model-Contrastive Federated Learning** [CVPR2021]

Yet, Another Benchmark OARF?

Name	Federated Dataset	Partitioning Scheme	Various ML Models	Privacy Mechanism	Comm. Cost
LEAF (Caldas et al. 2018)	✗	✗	✓	✓	✓
FL Performance Evaluation (Nilsson et al. 2018)	✗	✗	✗	✗	✓
Street Image (Luo et al. 2019)	✗	✗	✓	✗	✓
Edge AIBench (Hao et al. 2019)	✗	✗	✓	✗	✗
Evaluation for Large-Scale FL (L. Liu et al. 2020)	✗	✗	✗	✗	✓
End-to-End Evaluation (Gao et al. 2020)	✗	✓	✗	✗	✓
FedML (He et al. 2020)	✓	✓	✓	✓	✗
Data Island Benchmark Suite (Liang et al. 2020)	✓	✗	✗	✗	✗
A Benchmark of Federated Forest (Y. Liu et al. 2019)	✓	✗	✓	✓	✗
FedEval (Chai et al. 2020)	✗	✓	✓	✓	✓
FL Person Re-Identification (Zhuang et al. 2020)	✓	✗	✓	✓	✗
Semi-Supervised FL (Zhang et al. 2020)	✗	✗	✗	✗	✓
OARF (our work) (Hu et al. 2020)	✓	✓	✓	✓	✓

Benchmark Design – Tasks and Data Sets

P[1]	D[2]	Task	Data Set	Size	Align
Horizontal	CV	Gender / Age Prediction	10K US Adult Face	~10k	—
			All-Age-Face	~13k	
			OUI Audience Face	~26k	
			IMDB-WIKI – 500k	~52k	
			Labeled Faces in the Wild	~13k	
		Face Recognition	BioID	~1.5k	
			FEI Face Database	~2.8k	
			FaceScrub	~100k	
		ASCII Character Recognition	MNIST	~60k	
			Chars 74K	~74k	
			Stanford OCR	~52k	
		Chinese Character Recognition	HIT-OR3C	~0.5M	
			CASIA-HWDB1.1	~1.1M	
	NLP	Sentiment Analysis	IMDB Movie Review	50k	
			Rotten Tomato Movie Review	~7k	
			Amazon Movie Review	~8M	
	GIS	Traffic Prediction	METR-LA	~6.5M	
			PEMS-BAY	~17M	
Vertical / Hybrid	General ML	Trend Prediction / Recommendation	Steam Game	~17k	Title
			IGN Rating	~18k	
			Video Game Sales 2019	~55k	
		Trend Prediction / Recommendation	MillionSong	~1M	Name
			Free Music Archive	~106k	
		Trend Prediction / Recommendation	MovieLens Rating	~20M	Title
			Movie Industry	~6.8k	
			IMDB Movie	~4M	
Any		Any	Synthetic Data Sets	—	

[1] Data partitioning [2] Domain

Experiment 2: Sentiment Analysis

Sentiment Analysis In this task, we train a LSTM model and predicts the sentiment of movie reviews. We split the Amazon and IMDB data sets for training and prediction.

Training Setup	Accuracy (%)	
	IMDB	Amazon
IMDB	85.1	-
Amazon	-	86.7
Combined[1]	86.8	87.2
FedAvg	85.4	87.2

[1] Combining two data sets horizontally and fed them into the model as one data set.

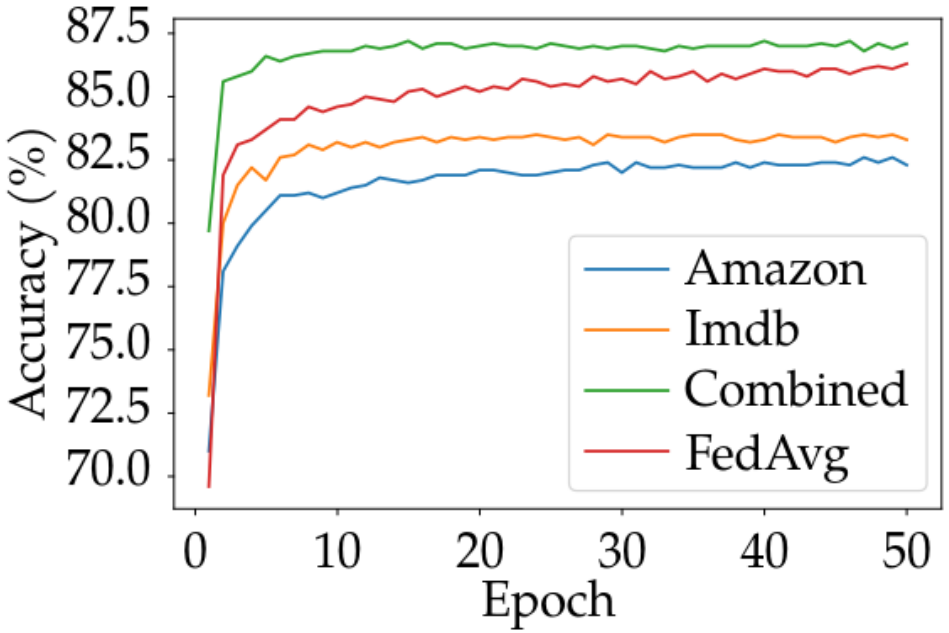
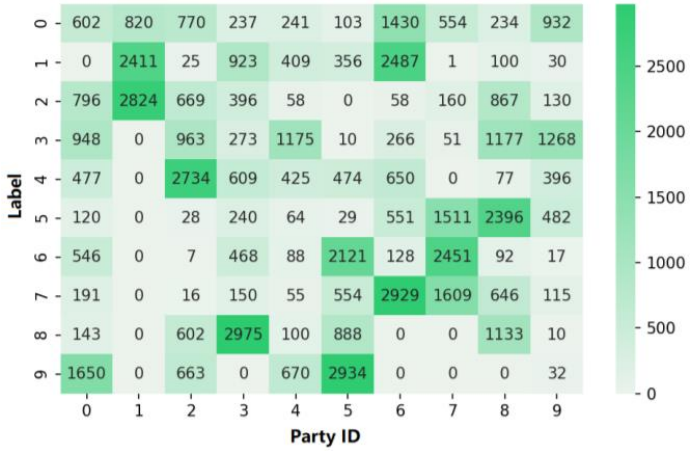


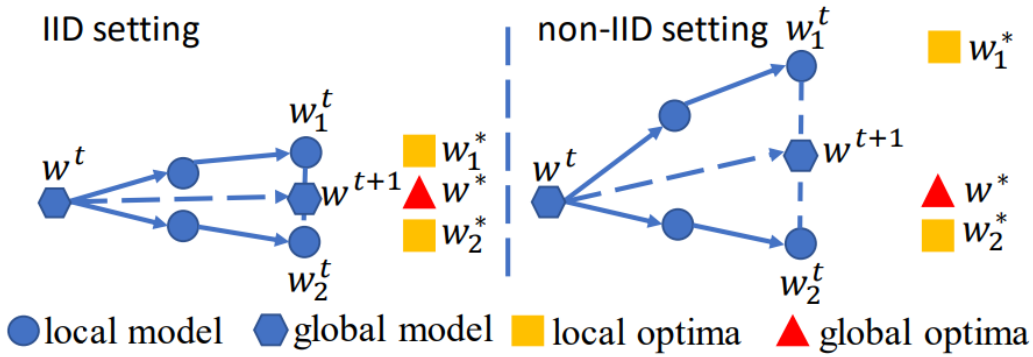
Figure 2: Accuracy on Combined Data

The Effect of Non-IID Data

- **Non-IID data significantly decreases the performance of FL.**

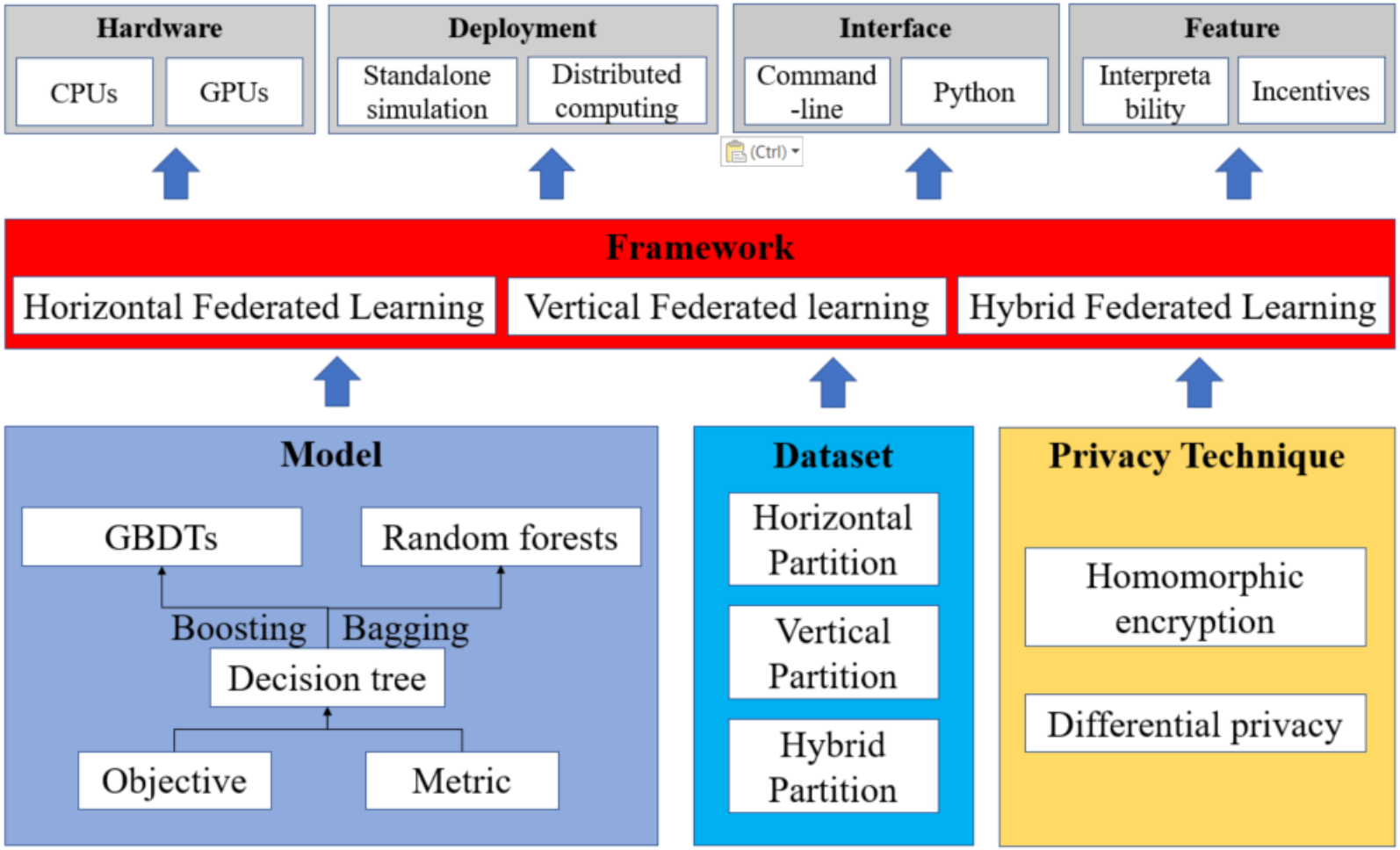


Datasets	IID	non-IID
CIFAR-10	70.4%	49.8%
SVHN	88.5%	80.2%



Qinbin Li*, Yiqun Diao*, Quan Chen, Bingsheng He. Federated Learning on Non-IID Data Silos: An Experimental Study. IEEE ICDE: IEEE International Conference on Data Engineering 2022

FedTree: Secure, Efficient and Flexible Tree FML



https://github.com/Xtra-Computing/FedTree

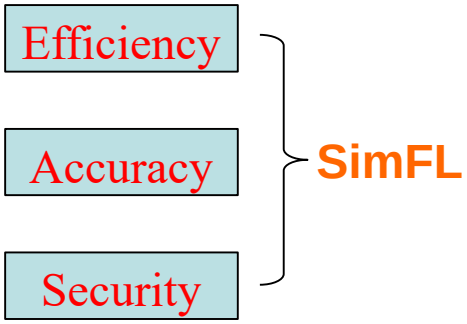
Gradient Boosting Decision Trees (GBDT)

- **Have won many awards in machine learning and data mining competitions.**
- **Achieve state-of-the-art performances in many machine learning tasks.**
  - Multi-class classification, click prediction, learning to rank
- **Some powerful libraries**
  - XGBoost, LightGBM, CatBoost, ThunderGBM,…

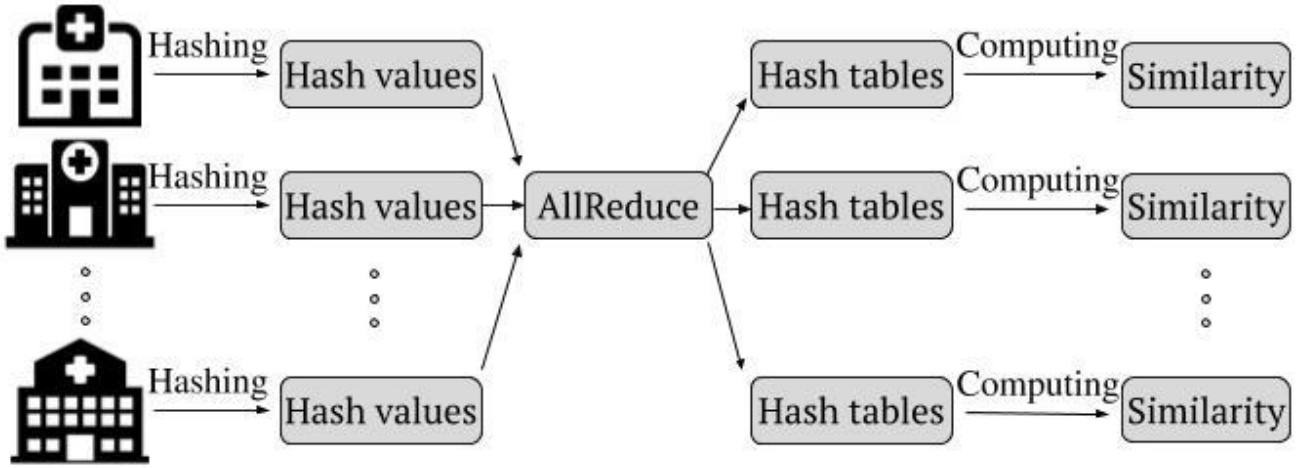
GBDT-based Federated Learning

- **Cryptographic approaches**
  - Homomorphic encryption, secret sharing…
  - High computation and communication cost

- **Differential privacy**
  - Adding random noise to the model parameters
  - Tend to produce less accurate models due to the noise

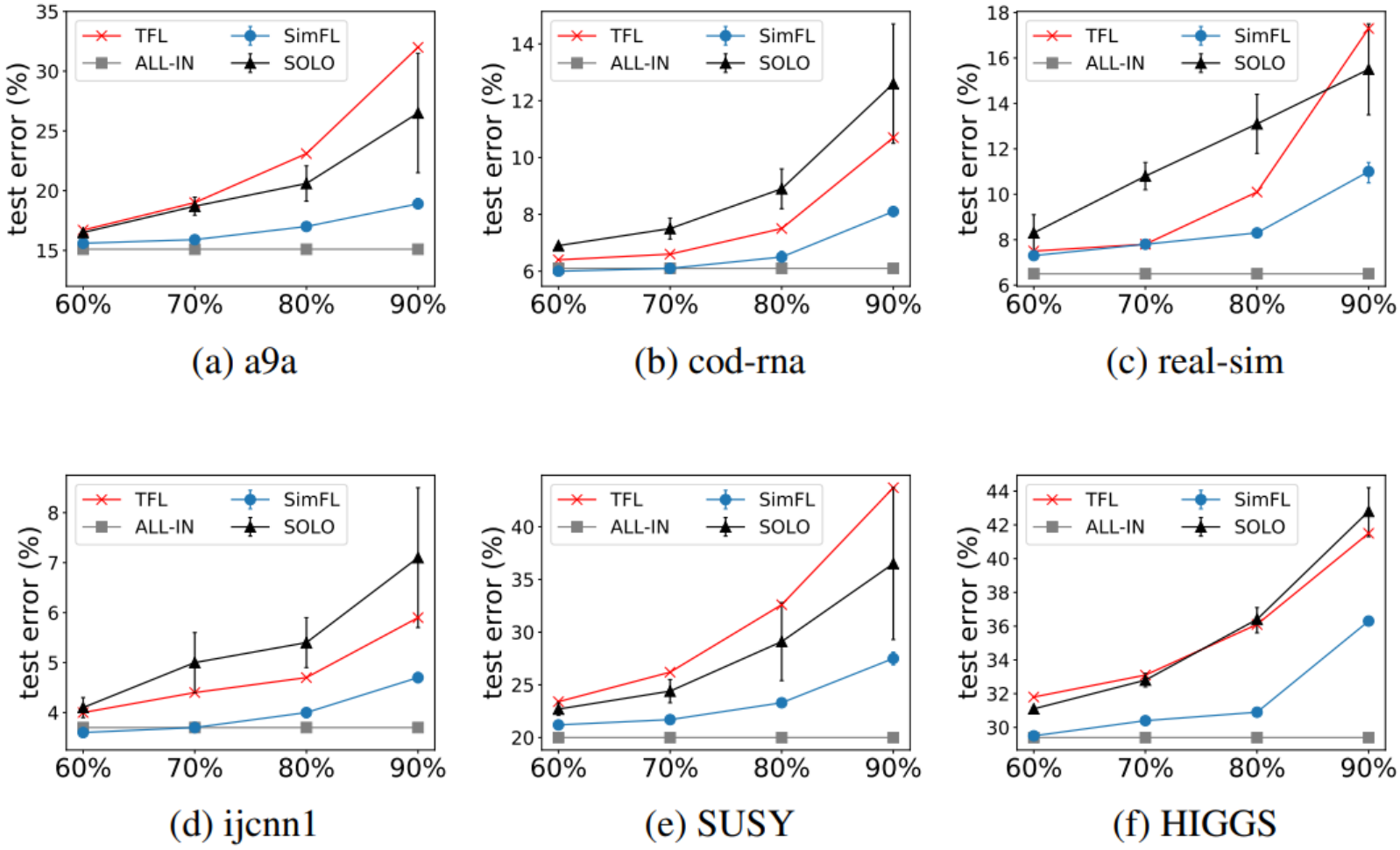


The Preprocessing Stage



- **Result:** Given an instance x_m^i in party P_i, and given a party P_j, we know an instance x_n^j from P_j that is similar with x_m^i.

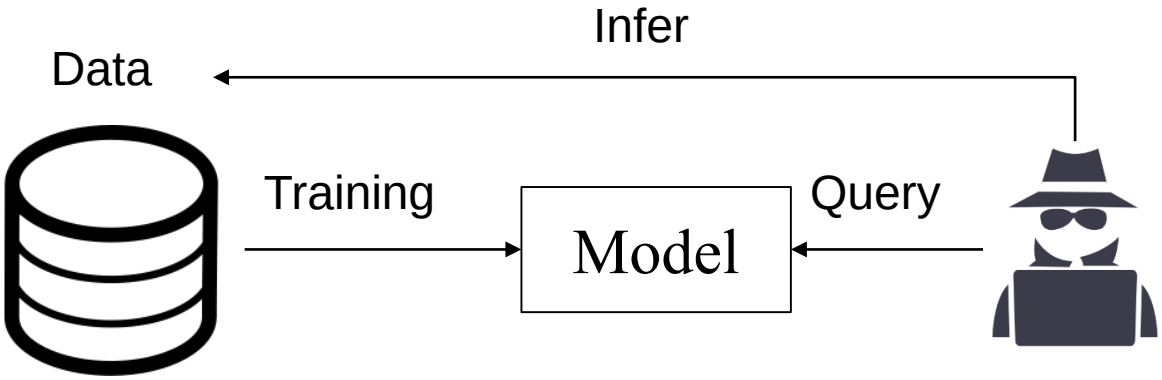
2-party Setting



The test errors with different ratio θ

Privacy Issues on Models

- **Model Inference attacks**
  - The attacker can infer sensitive information about the training data by only accessing to the model[1,2].
  - Model should also be protected.



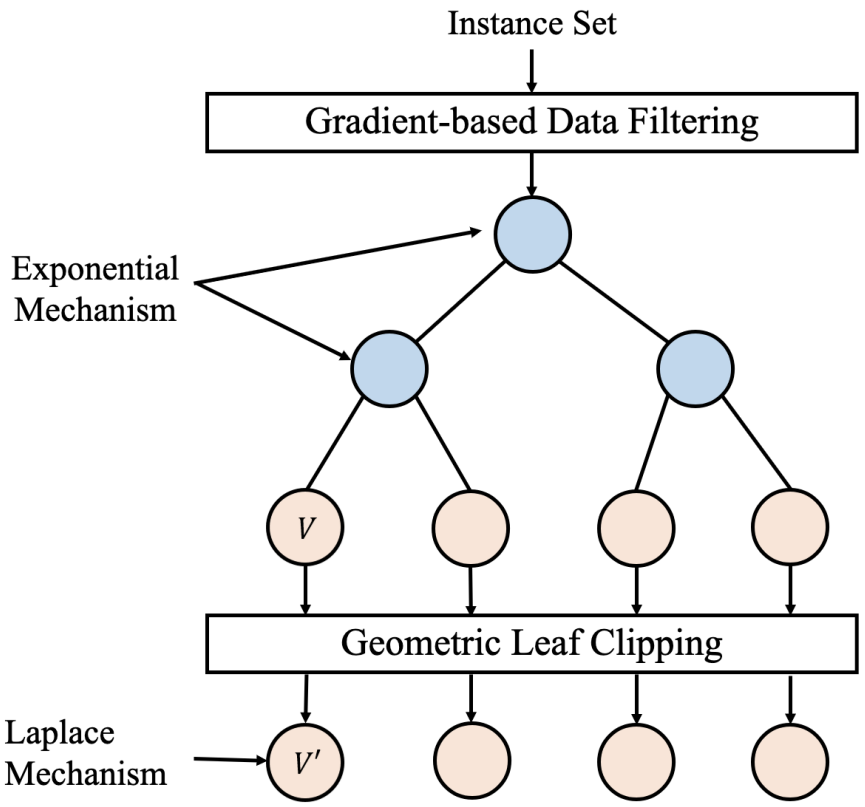
[1] Shokri, Reza, et al. "Membership inference attacks against machine learning models." S&P 2017.
[2] Nasr, Milad, Reza Shokri, and Amir Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." S&P 2019.

DPBoost: Privacy-Preserving GBDT

- **Idea: Apply differential privacy (DP) in the training of GBDT.**
  - DP: A single record does not influence the output of a function much.
  - Approaches: Apply Laplace mechanism and exponential mechanism.
- **Challenges**
  - Sensitivity bounding: How to bound the output range of the functions to compute gain and leaf values?
  - Budget allocation: With a fix total privacy budget, how to allocate the budget among different trees?

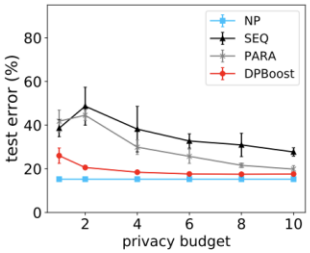
Apply DP in a Tree

- **Internal nodes**
  - Apply exponential mechanism to randomize the selection of split points.
- **Leaf nodes**
  - Apply Laplace mechanism to add noises to the leaf values.
- **Sensitivity**
  - Filtering and clipping techniques to bound the sensitivities.

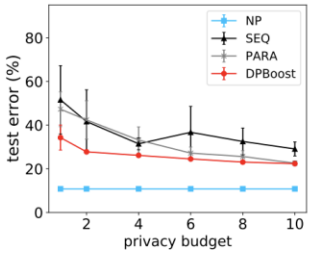


Evaluation

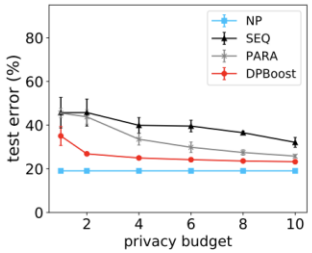
- **DPBoost significantly outperforms the other privacy-preserving approaches.**



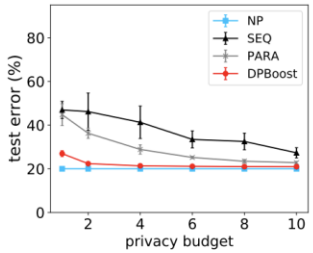
(a) adult



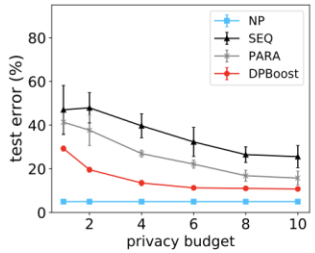
(b) real-sim



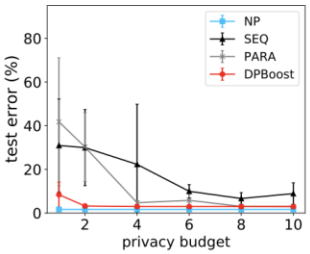
(c) covtype



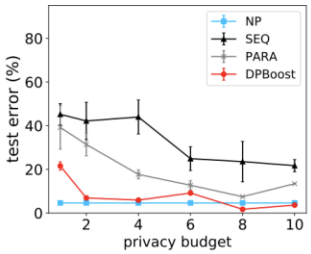
(d) susy



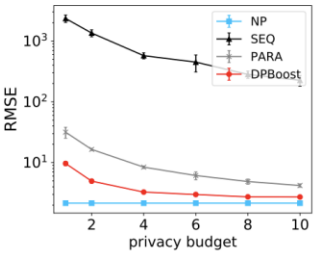
(e) cod-rna



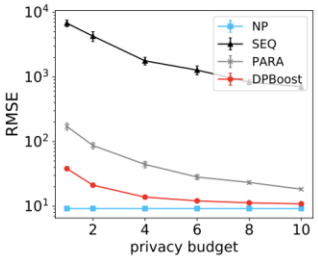
(f) webdata



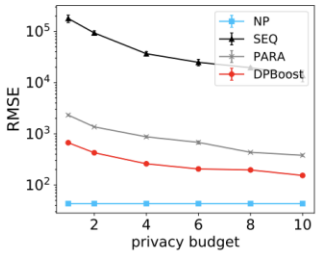
(g) synthetic_cls



(h) abalone



(i) YearPredictionMSD



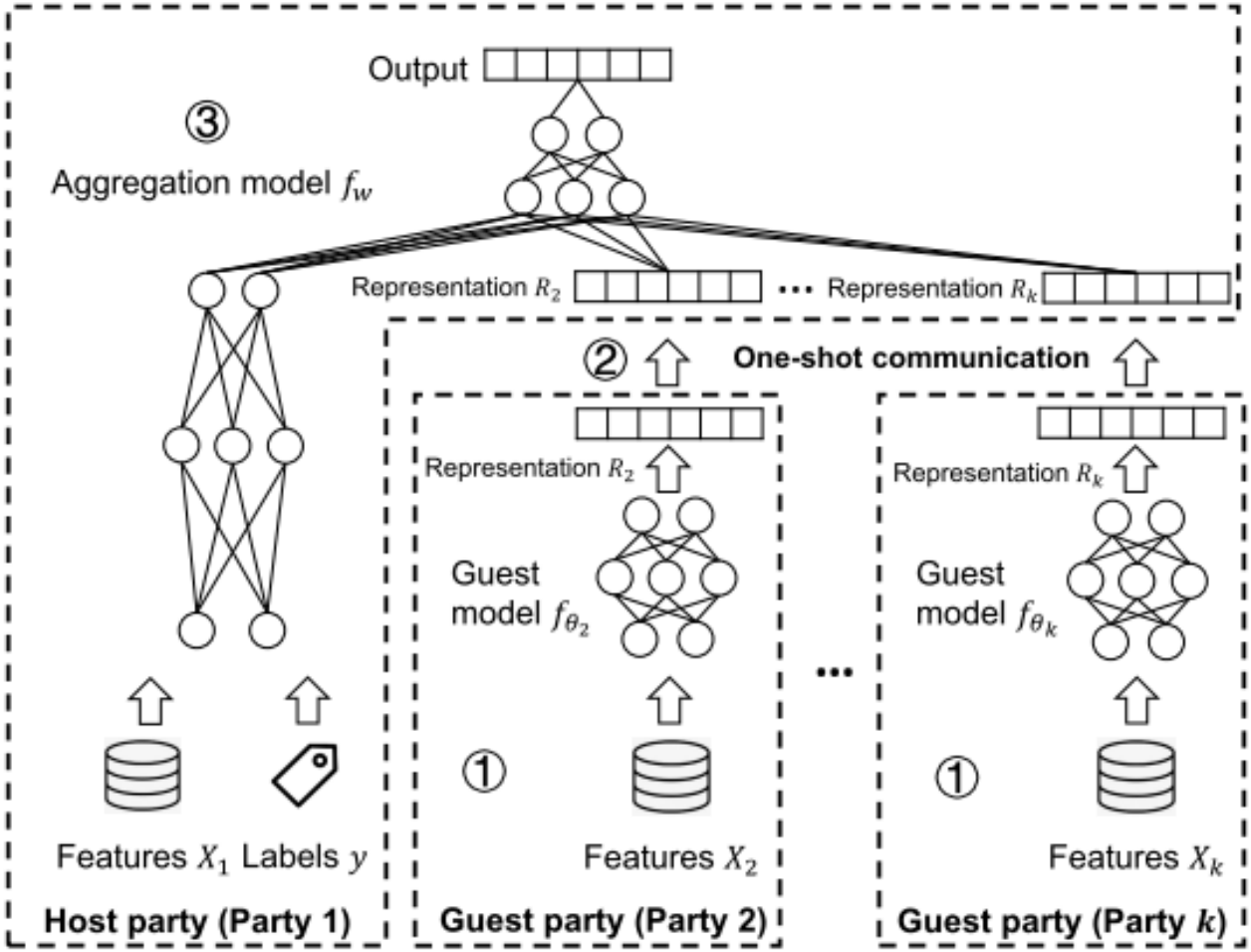
(j) synthetic_reg

One-Shot Vertical Federated Learning with Unsupervised Representation Learning

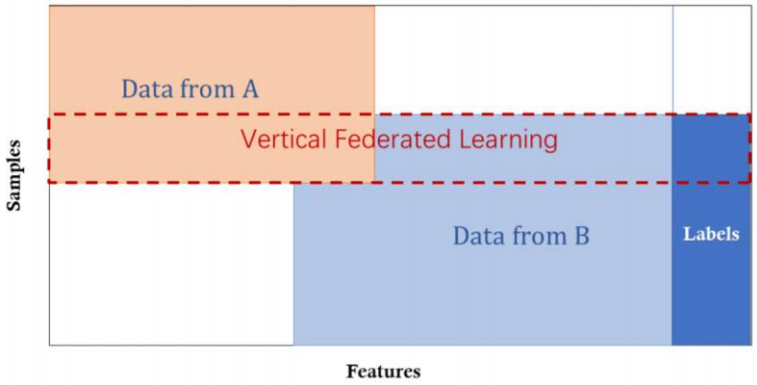
- Most previous studies are based on horizontal setting. However, relatively less attention has been drawn to vertical setting.
- The labels usually only exist in one single party.
- One shot: a single communication round, thus, optimal in communication.
- Core idea: Extract information from guest parties by using unsupervised learning, which can extract representative features from guest parties without exposing the raw data.

Zhaomin Wu*, Qinbin Li*, Bingsheng He. Practical Vertical Federated Learning with Unsupervised Representation Learning. IEE TBD: IEEE Transactions on Big Data (Special Issue on Trustable, Verifiable, and Auditable Federated Learning) 2022.

One-Shot Vertical Federated Learning with Unsupervised Representation Learning



Vertical Federated Learning (VFL)



(Yang et al. TIST 2019)

Share the same sample space
Own a subset of features
Only one party has labels

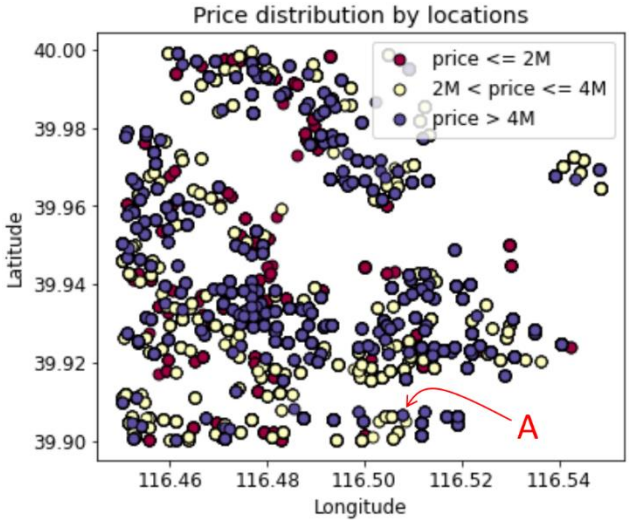
How to determine which instances should be involved in training?
Privacy-Preserving Record Linkage (PPRL) [1]

How existing studies use PPRL in VFL?
Train exactly/top1 matched records

According to the study in German record linkage center [2], 72.7% of the applications suffer information loss by exact/top1 linkage

[1] Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. Handbook of big data technologies, 851-895.
[2] Manfred Antoni and Rainer Schnell. The past, present and future of the german record linkage center (grlc). Jahrbücher für Nationalökonomie und Statistik, 239(2):319–331, 2019.

Record Linkage



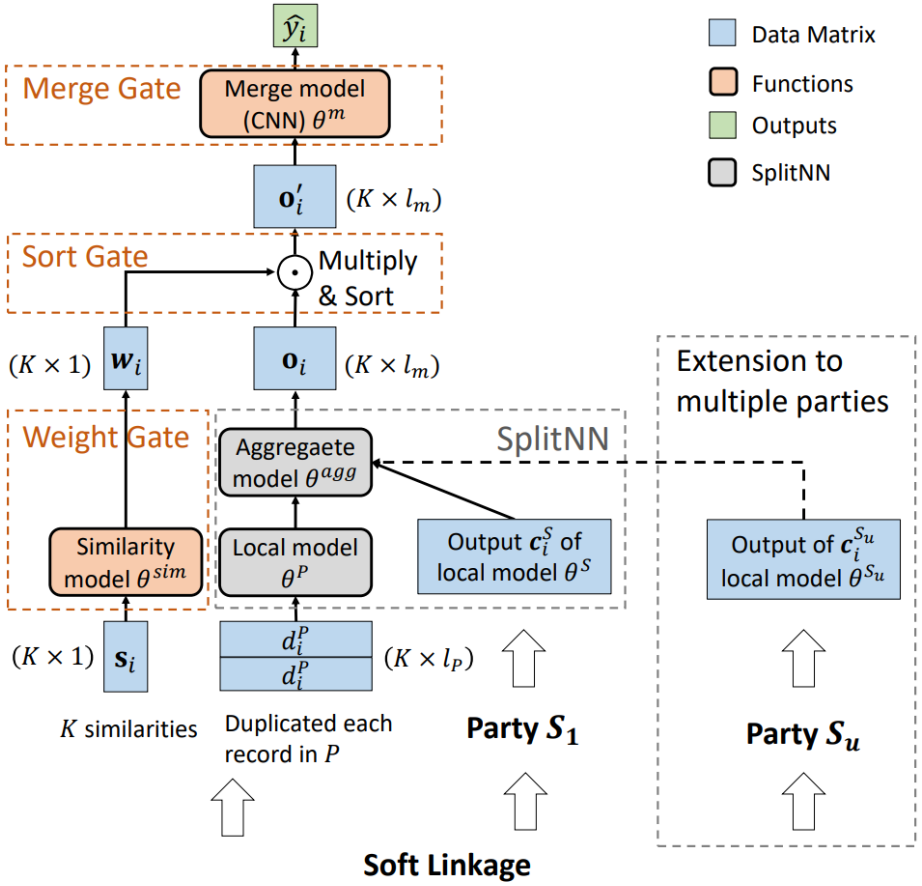
Housing price by geolocations in Beijing

Real estate company & Airbnb
Linked on housing address

Task: Predict housing price

Only linking records with top1 similarity may not capture key features

Our Design: FedSim



Weight Gate: grant each record pair a weight according to its similarity

Sort Gate: sort the record pairs according to weights

Merge Gate: a CNN with $n \times 1$ kernel to merge the record pairs with similar weights

Zhaomin Wu*, Qinbin Li*, Bingsheng He. A Coupled Design of Exploiting Record Similarity for Practical Vertical Federated Learning. NeurIPS: Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022).

Our Design: FedSim

Table 1: Performance on real-world datasets

Algorithms	house (numeric) $\Delta = 34.05$	bike (numeric) $\Delta = 14.26$	hdb (numeric) $\Delta = 20.69$	game (string) $\Delta = 4.14$	company (string) $\Delta = 10.50$
Solo	58.31±0.28	272.83±1.50	29.75±0.15	85.27±0.29%	42.67±0.66
Exact	-	-	-	89.25±0.12%	44.44±1.95
Top1Sim	58.54±0.35	256.19±1.39	31.56±0.21	92.71±0.08%	42.84±0.77
FeatureSim	66.39±0.15	273.29±0.37	37.39±0.29	91.13±0.23%	39.24±1.80
AvgSim	51.92±0.65	239.85±0.40	34.12±0.19	90.84±0.14%	38.19±0.91
FedSim (w/o Weight)	42.82±0.20	236.79±0.29	27.18±0.08	92.79±0.13%	41.00±1.19
FedSim (w/o Sort)	52.14±0.58	238.30±0.81	36.35±0.42	92.79±0.10%	38.28±1.56
FedSim (w/o CNN)	42.62±0.20	235.97±0.42	27.76±0.13	92.50±0.12%	39.63±1.31
FedSim	**42.12±0.23**	**235.67±0.27**	**27.13±0.06**	**92.88±0.11%**	**37.08±0.61**

FedSim outperforms all the baselines in five real world datasets

Conclusion

- A comprehensive survey on **categorization** and **comparison** for federated learning systems.
- Our initial study on GBDT and FedTree has demonstrated that secure and practical machine learning systems are possible.
- Federated learning systems will be an **exciting research journey**, which call for the system research from machine learning, distributed computing, and data privacy communities.

Other On-going Works

- **Improve the efficiency of vertical GBDT**
- **One-shot communication-efficient FML systems**
- **Hardware accelerated FML systems**
- **Open-source systems and benchmarks:**
  - https://github.com/Xtra-Computing/FedTree
  - https://github.com/Xtra-Computing/NIID-Bench
  - https://github.com/Xtra-Computing/OARF
- **AI Singapore Research Grant 2020 Award: "Toward Trustable Model-centric Sharing for Collaborative Machine Learning“ [We are hiring…]**

References

- Sixu Hu*, Yuan Li*, Xu Liu*, Qinbin Li*, Zhaomin Wu*, Bingsheng He. The OARF Benchmark Suite: Characterization and Implications for Federated Learning Systems. ACM TIST: ACM Transactions on Intelligent Systems and Technology. https://arxiv.org/abs/2006.07856
- Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wu, Bingsheng He. "Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection." in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3124599. https://arxiv.org/abs/1907.09693
- Qinbin Li*, Zeyi Wen^, Bingsheng He. Practical Federated Gradient Boosting Decision Trees. AAAI: Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) (1,591/7,737=20.6%).
- Qinbin Li*, Zhaomin Wu*, Zeyi Wen^, Bingsheng He. Privacy-Preserving Gradient Boosting Decision Trees. AAAI: Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) (1,591/7,737=20.6%).
- Qinbin Li*, Zeyi Wen^, Bingsheng He. Adaptive Kernel Value Caching for SVM Training. IEEE TNNLS: IEEE Transactions on Neural Networks and Learning Systems.
- Zhaomin Wu*, Qinbin Li*, Bingsheng He. Practical Vertical Federated Learning with Unsupervised Representation Learning. IEE TBD: IEEE Transactions on Big Data (Special Issue on Trustable, Verifiable, and Auditable Federated Learning) 2022.
- Qinbin Li*, Bingsheng He, Dawn Song. Practical One-Shot Federated Learning for Cross-Silo Setting. IJCAI: 30th International Joint Conference on Artificial Intelligence, 2021. https://arxiv.org/abs/2010.01017
- Qinbin Li*, Yiqun Diao*, Quan Chen, Bingsheng He. Federated Learning on Non-IID Data Silos: An Experimental Study. IEEE ICDE: IEEE International Conference on Data Engineering 2022
- Qinbin Li*, Bingsheng He, Dawn Song. Model-Contrastive Federated Learning. CVPR: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Zhaomin Wu*, Qinbin Li*, **Bingsheng He**. A Coupled Design of Exploiting Record Similarity for Practical Vertical Federated Learning. NeurIPS: Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022).

Backup Slides