Privacy-Preserving Collaborative Information Sharing through Federated Learning
– A Case of the Insurance Industry

Panyi Dong
ARC 2023,
Jul. 31, 2023

Acknowledgement

- Intel Labs
- Discovery Partners Institute, University of Illinois

intel

- Shih-han Wang
- Brandon Edwards
- Micah Sheller
- Wells Lee
- Patrick Foley
- Nageen Himayat
- Parviz Peiravi
- Jason Martin

DPI
DISCOVERY PARTNERS INSTITUTE
PART OF THE UNIVERSITY OF ILLINOIS SYSTEM

- Tianyang Wang
- Marc Goodman
- Renu Kulkarni

IRisk
Illinois Risk Lab

NCSA

- Panyi Dong
- Zhiyu Quan
- Runhuan Feng

Synopsis

- Challenges in the insurance industry
- How Federated Learning (FL) can help
- Key insights from real-life empirical experiments

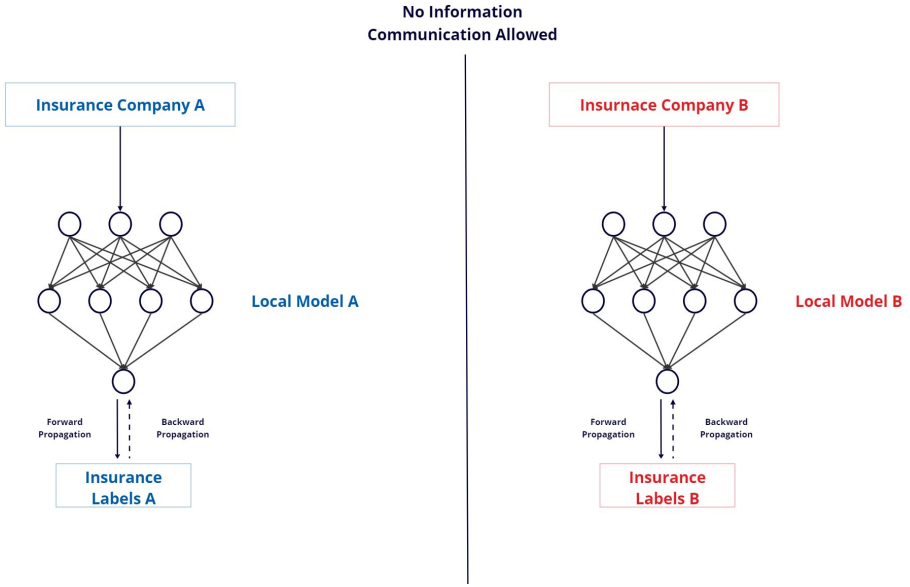
Motivation

Problems

- Academic literature shows that ML collaboration has proven beneficial
  - ML algorithms for better
    - Loss models, reserving, fraud detection, investment decisions, etc.
  - Industry-level insights
    - Regulators
    - Align common interests

- Data sharing with other companies is practically near impossible
  - Data security standards
    - Insurance Data Security Model Law (NAIC)
  - Growing attention in cyber-security
    - Infrastructures
  - Business concern
    - Proprietary information

Problems

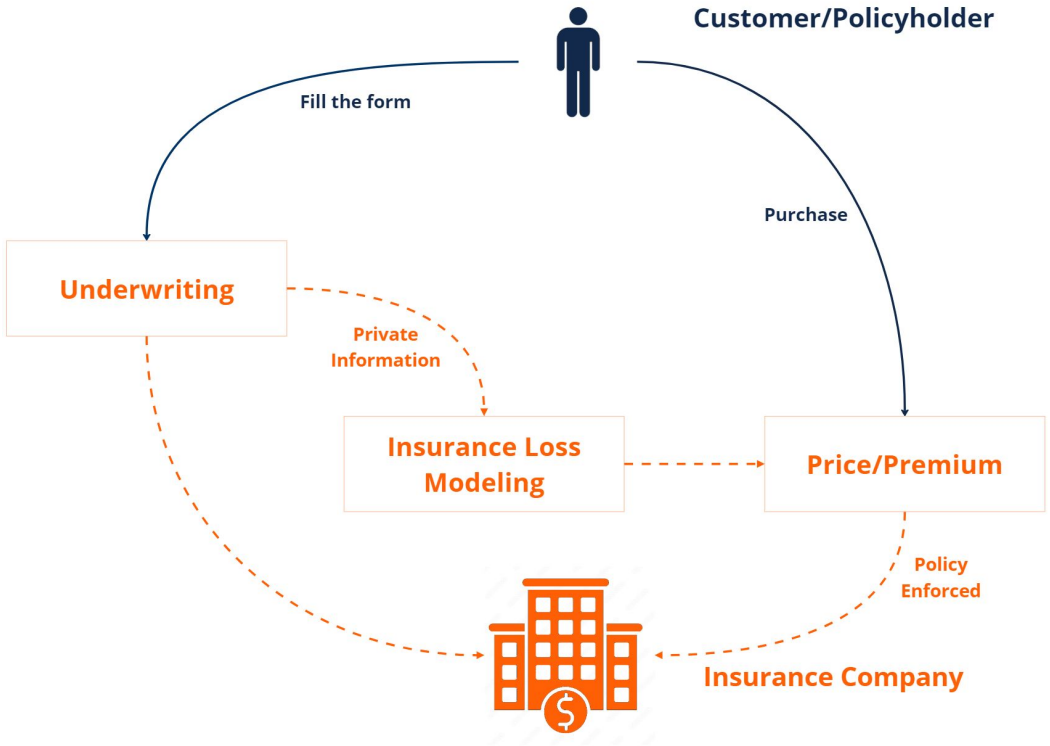


- Insurance companies are in the local mode
  - Rarely observe industry collaborations
- ML collaboration + privacy-preserving solution is in demand
  - Federated Learning (FL) can fill in the gap
  - Real-life applications can be lucrative with high ROI if feasible

Insurance use cases for FL

- Claim Loss Modeling
  - Address *Data Shortage*
  - *The focus of our paper*
- Fraud Detection
  - Cross-industry collaboration to solve industry pain points
  - Extend to problems like money laundering
    - Bank
- Catastrophe Modeling
  - Reinsurance company
  - Climate risk modeling

FL for Claim Loss Modeling



- Claim Loss Modeling
  - Learn from historical claim events
  - An imbalanced supervised regression task

Problems Shortage in Data

- Insurance companies collect conventional information from policyholders
  - Private personal information
    - Names
    - Addresses
    - Credits
  - Private business information
    - Sales records
    - Annual revenue
    - Properties
- But still in shortage of (the focus of our work)
  - Data volume
  - Data variety

Problems Data Volume

- Claim loss events have rare occurrences (imbalance nature)
  - Less than 10% in practice (extreme as 0.1%)
    - Car accidents
    - Rare diseases
  - Individual insurance companies lack sufficient loss events
    - Depends on the market share
    - New to the line of business/Step into new regions
- *More prominent the data volume, the more claim loss events to learn from*
  - Better insurance loss model

Solutions HFL

- To address the *Shortage of Data Volume*
  - Horizontal Federated Learning (HFL) as a solution
- HFL,
  - Learn industry-level insights
  - Simulate the centralized training utilizing all datasets from collaborators
  - By iterative local training and model aggregation
- In business,
  - Multiple insurance companies collaborate on the same line of business
  - Single company presents heterogeneous groups of policyholders
  - The centralized model can learn from all claim loss events

Problems Data Variety

- Every insurer is chasing the perfect feature set
  - 100% accurately map risk factors (features) to risks
  - A challenge to the entire insurance industry
- However,
  - The risks may come from multiple external sources
    - Social media
    - Telematics
    - US census data
  - Insurance companies by themselves can’t get their hands on everything
- A cross-industry partnership may be a solution
  - Insurance-InsurTech
  - Insurance-Banking
  - Insurance-Government

Solutions VFL

- To address the *Shortage of Data Variety*
  - Vertical Federated Learning (VFL) as a solution
- VFL,
  - Learn cross-industry insights
- In business,
  - Multiple companies from different industries
  - Same group of policyholders
  - Expand the feature space

Data & Experiments

Experiments

- Real-life Datasets
  - Two Insurance companies
    - Insurance features + Insurance labels
  - One InsurTech company
    - InsurTech features
- Learning task
  - Insurance claim loss modeling
    - Regression
- Experiment framework: OpenFL
  - Open-source FL framework by Intel Labs
  - FNN as model architecture

Datasets Insurance

- Two Insurance companies (A/B)

Company	A	B
Product Coverage	Liability in Business Owners' Policy (BOP) products	General Liability products
Data Size	392,726 policies	210,857 policies
	26 features	39 features

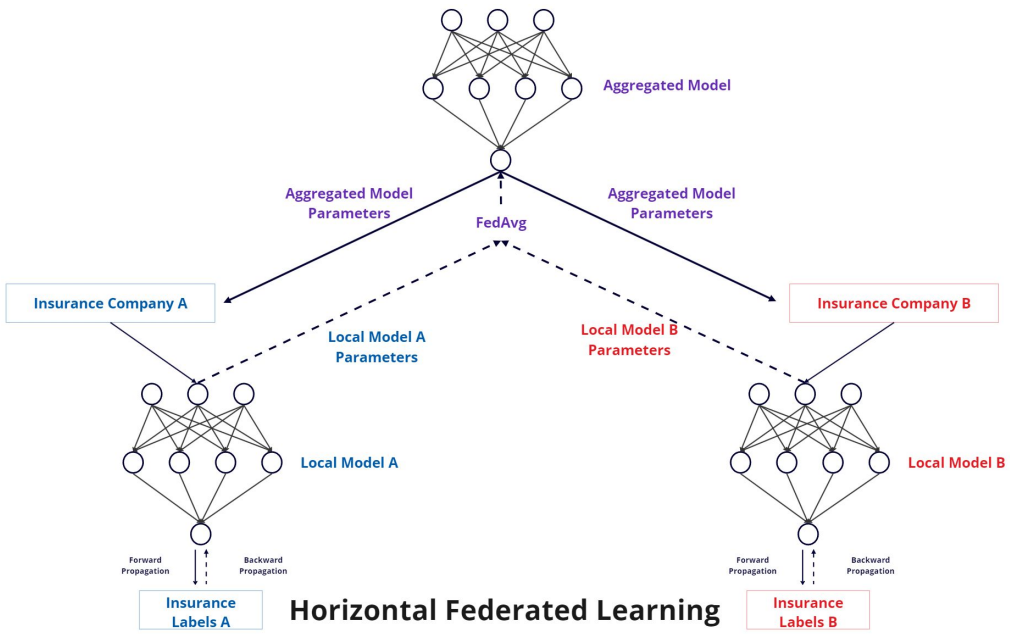
- Features (limited policy information)
  - Coverage Limit
  - Exposure
  - Category of business

Datasets InsurTech

- InsurTech datasets
  - Consists of hundreds of features from multiple data sources
  - Covering policyholders of both insurance company A/B
  - Aligned for each policy

- Data Size
  - For Company A
    - 555 features
    - 392,726 policies
  - For Company B
    - 555 features
    - 210,857 policies
  - 555 features are the same

Experiments HFL



- Two-collaborator HFL
  - Common line of the business - Liability products
  - *FedAvg*
- Enables
  - Protection against raw data leakage
  - Collaboration among insurance companies
  - Improvement in insurance models

Experiments HFL

- Algorithm: FedAvg
  - Earliest FL algorithm
  - Iterative local training + central model aggregation
    - same local training
    - central takes the average of models

$$f_{central} = \frac{f_1 + f_2}{2}$$

  - The consensus model inhere the global insights

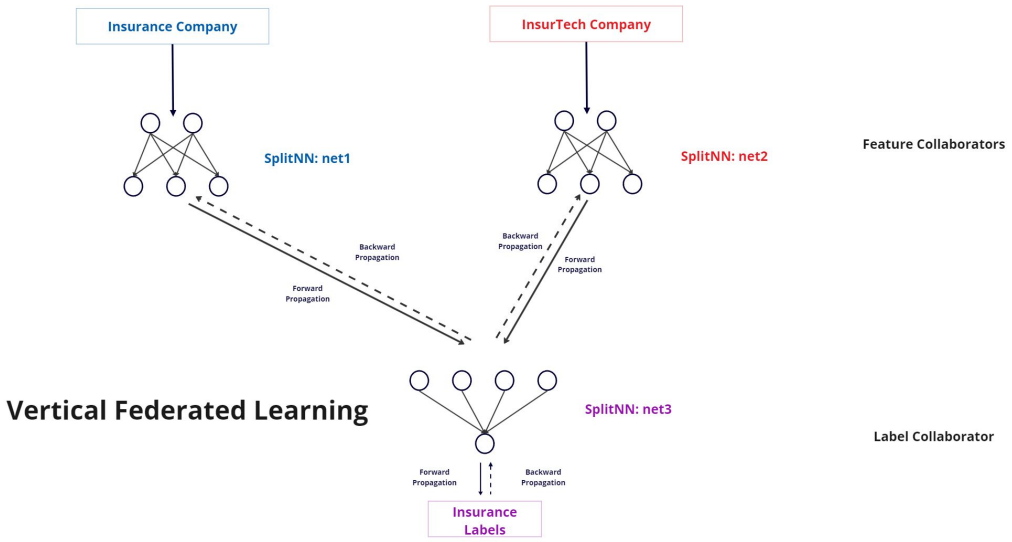
Performance HFL

Collaborator	Split	Mode	PE
A	Train*	Local	-0.16
		HFL	**-0.07**
	Test	Local	-0.18
		HFL	**-0.09**
B	Train	Local	0.22
		HFL	**0.13**
	Test	Local	0.23
		HFL	**0.16**

Performance metrics of HFL by Percentage Error $PE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i y_i}$

* To ensure robustness and reliable model performance, the training process was conducted using 10-fold cross-validation with random sampling.

Experiments VFL



- Two-collaborator VFL
  - Insurance-InsurTech
  - Liability products
  - *SplitNN*
- Enables
  - Protection against raw data leakage
  - Collaboration for cross-industry
  - Comprehensive insights

Experiments VFL

- Algorithm: Split Neural Network (SplitNN)
  - Split one NN into multiple segments
    - Easy incorporation into NNs
  - Differentiate collaborators by roles
    - Feature collaborator has only features
    - Label collaborator provides labels
  - Forward/Backward propagation
    - Feature collaborator forwards raw data to unidentifiable embeddings
    - Label collaborator forwards all embeddings to predictions
    - Backward propagation inverses sequentially

Performance VFL

Collaborator	Split	Mode	PE
A	Train	Local	-0.16
		VFL	**0.07**
	Test	Local	-0.18
		VFL	**0.04**
B	Train	Local	0.22
		VFL	**0.12**
	Test	Local	0.23
		VFL	**0.16**

Performance metrics of VFL by Percentage Error ($PE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_i (y_i - \hat{y}_i)}{\sum_i y_i}$

Conclusion

Conclusion

- Identify potential real-world use cases
  - Solving data shortages in loss modeling
- Propose Federated Learning as a solution
  - HFL for the increase in data volume
  - VFL for the increase in data variety
- Experiments have demonstrated improved insurance claim loss models
  - Better portfolio predictions
  - More efficient risk management

Thank you!
Q&A

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Appendix

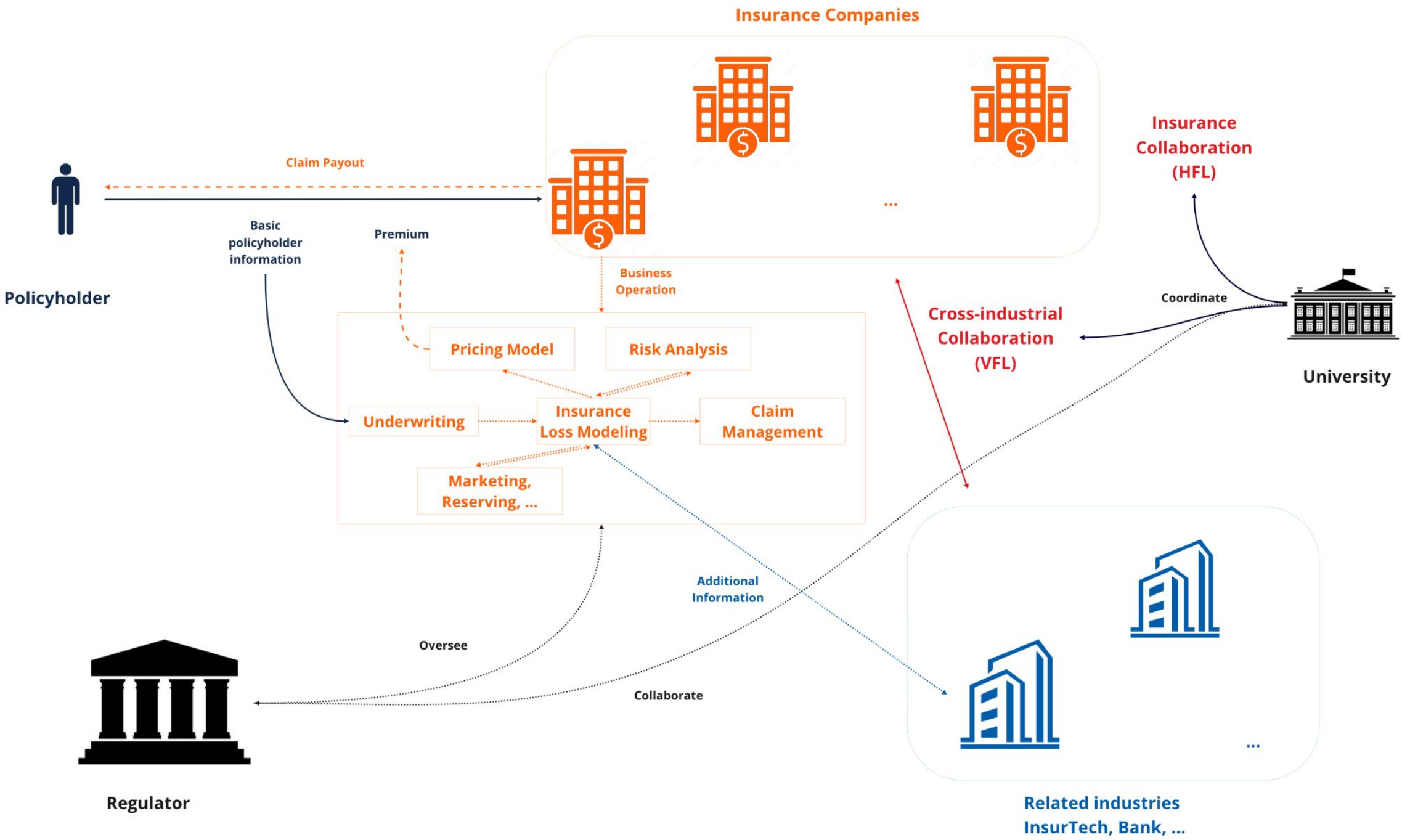
Insurance in short

- Insurance is
  - Data-driven industry
    - Natural for Machine Learning (ML) innovations
    - Many informed business decisions rely on data
  - Offers protection against various risks
    - It requires massive data to assess and manage the underlying risk
    - Abundant partnership opportunities, e.g., external data vendors
  - Overlapping business lines and target segments
    - Collaboration becomes rational and necessary
    - Operational efficiency and cost optimization

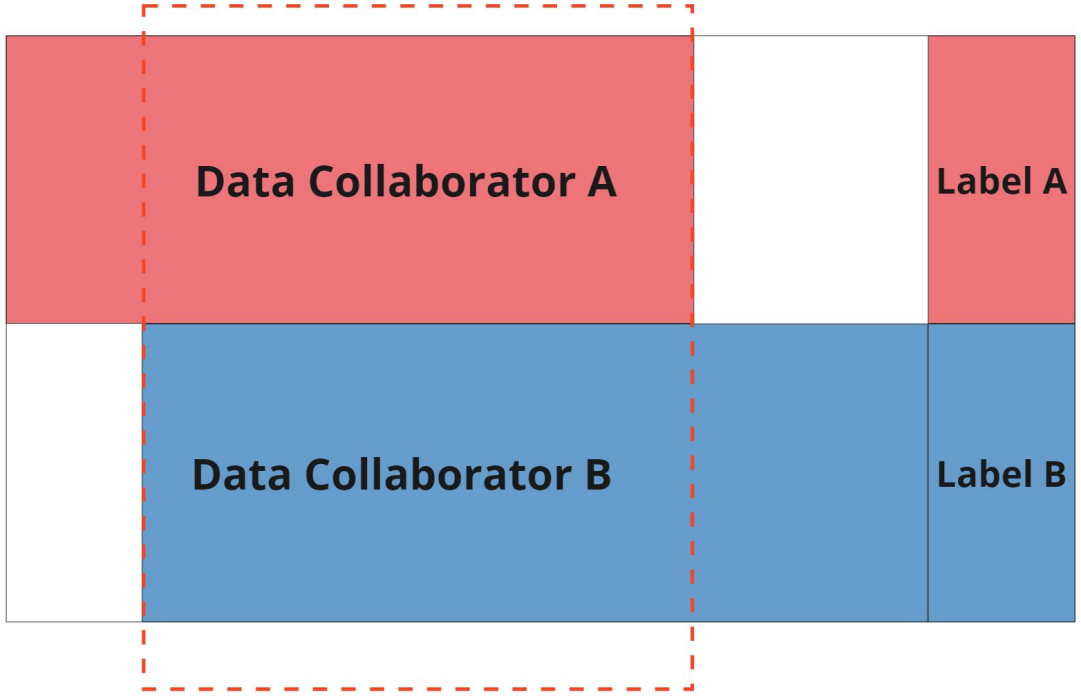
Insurance in short

- Insurance industry is highly-regulated
  - Regulators oversee insurance industry: solvency and fairness
  - Increasing data privacy concerns and regulatory requirements

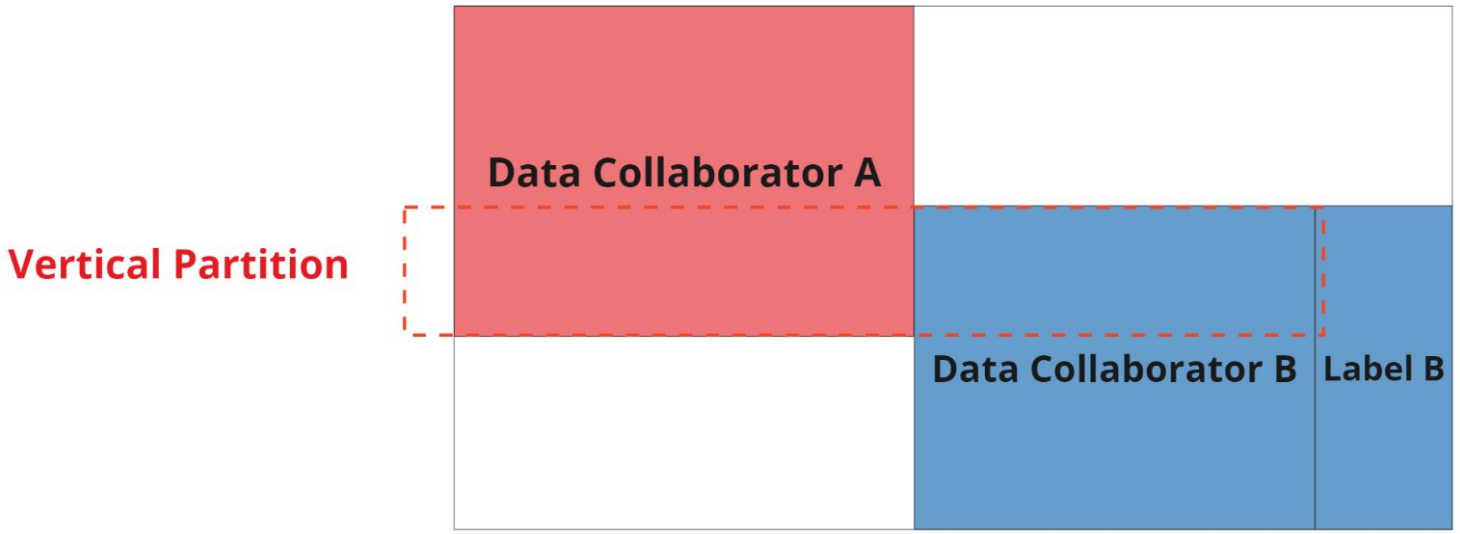
- Insurance companies are risk-averse
  - Prioritizing stability and minimizing uncertainties
  - Technical advancements bring risk-mitigation and competitive advantage
  - Business insights and value propositions
  - Success stories!


Insurance Companies
...
Claim Payout
Basic policyholder information
Premium
Policyholder
Business Operation
Pricing Model
Risk Analysis
Underwriting
Insurance Loss Modeling
Claim Management
Marketing, Reserving, ...
Insurance Collaboration (HFL)
Coordinate
University
Cross-industrial Collaboration (VFL)
Additional Information
Oversee
Collaborate
Regulator
...
Related industries
InsurTech, Bank, ...

Solutions HFL



Solutions VFL



Datasets InsurTech

- InsurTech
  - Tailored technology-enabled innovations for the insurance industry
  - Integrated with the entire value chain of every business lines
    - Marketing, Underwriting, Claim management
    - Property & Casualty, Life & Health

- Industry examples
  - Mobile devices with apps
    - Reporting claims, customer service
  - Wearable technology
    - Telematics, health tracking
  - Internet-enhanced features
    - Real-time, dynamic information from emerging public data sources
    - Characterize operations, products, services, etc.

Projection

Allstate, Commercial lines, Year 2022

Earned Premium	$919 M	
Loss ratio	120.7%	
Claims (expenses)	$1,109 M	
PE	20.7%	
	Through HFL	Through VFL
Relative improvement in PE (in experiments)	44.4%	52.5%
Average improvement in PE	9.2%	10.9%
Improvement in Dollars	**$84 M**	**$100 M**

Projection

Allstate, Commercial lines, Same Calculation, earlier years

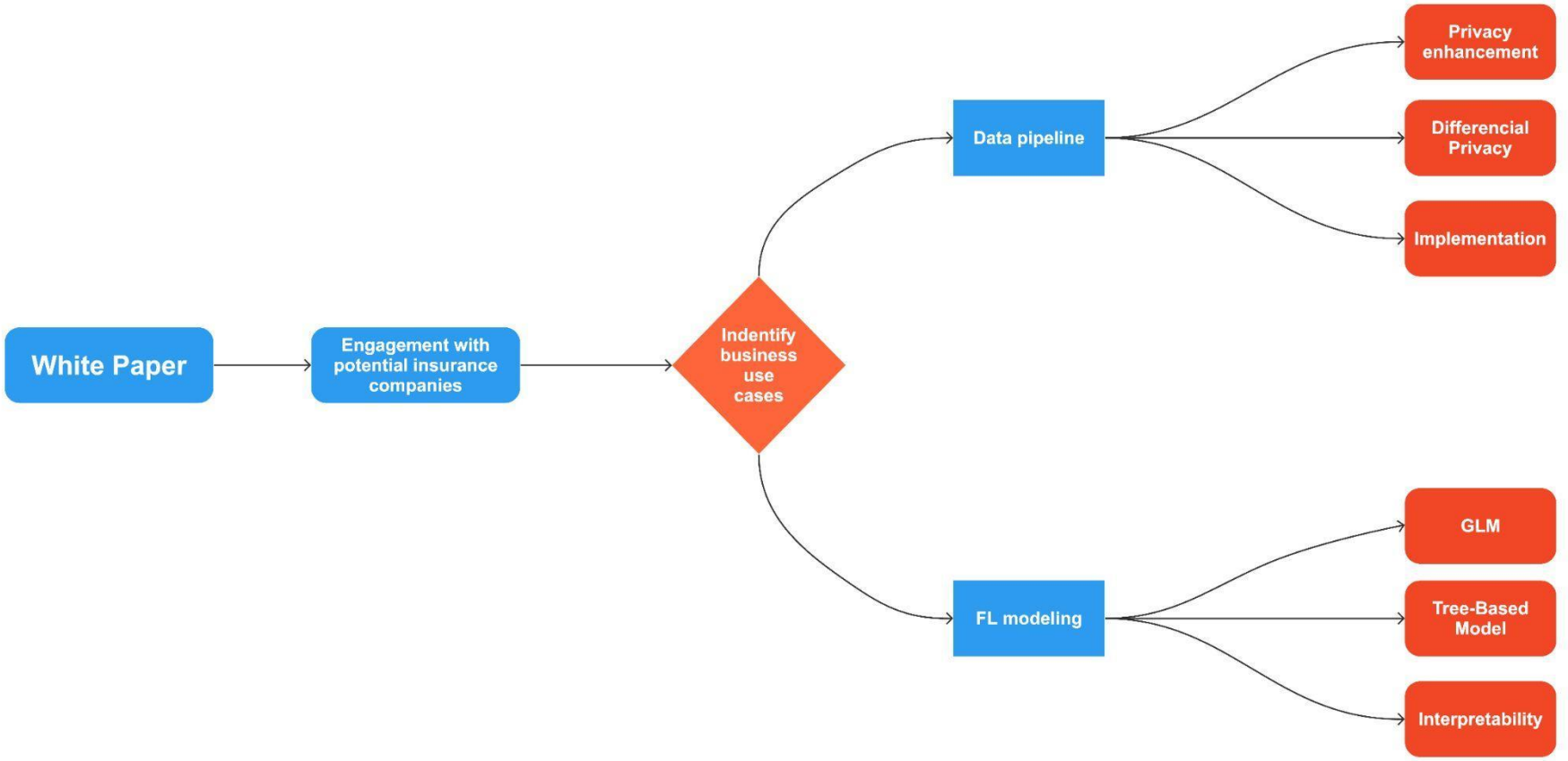
Year	Earned Premium	Loss Ratio	Improvement in Dollar	
			Through HFL	Through VFL
2021	$827 M	97.5%	$9 M	$11 M
2020	$767 M	82.4%	$60 M	$71 M
2019	$882 M	81.3%	$73 M	$87 M
2018	$655 M	91.3%	$25 M	$30 M

Projection

- **More Potentials**
  - Allstate, Earned Premiums in 2022
    - Auto: $25,286 M
    - Homeowners: $ 9,249 M
    - Personal: $ 2,016 M
    - Commercial: $ 919 M

- **Solve problems from source**
  - Better underwriting, pricing, claim management
  - Further improvement

Future


White Paper
Engagement with potential insurance companies
Indentify business use cases
Data pipeline
FL modeling
Privacy enhancement
Differencial Privacy
Implementation
GLM
Tree-Based Model
Interpretability

Future

- Adoption of insurance friendly algorithms
  - GLM
  - Tree-based models
- Explainability/Interpretability in the context of FL
  - Business insights for insurance companies
  - Future trends for regulators
- Explore more use cases
  - Fraud Detection
  - Catastrophe Modeling
- Enhancement of privacy protection tailored for insurance
  - Protection of features
  - Differential Privacy (DP)
  - Trusted Execution Environment (TEE)

IRisk Lab projects in parallel

- Customers may switch coverages among different insurance companies
  - May interfere with models
  - Deeper analysis is needed
- “Weird” privacy concerns may come up
  - Feature names > All raw values
    - Sharing feature engineering means expose business strategy
  - Seeking for more secure solutions
    - Not willing to share even parameters
  - …
- Those problems has been studied as a project at IRisk Lab at UIUC
  - Try to integrate insurance-specific solutions with FL