Institute
and Faculty
of Actuaries

IFoA GIRO Conference 2024

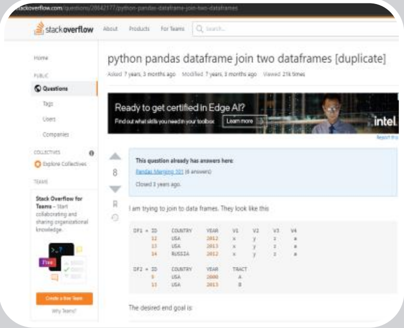
18–20 November, ICC, Birmingham

Hush hush: Keeping neural network claims modelling private, secret, and distributed using federated learning

Dr Małgorzata Śmietanka & Dylan Liew & Michelle Chen

IFoA GIRO Conference 2024

Who am I and motivation



What I think I do

What I really do

What My Boss Thinks I Do

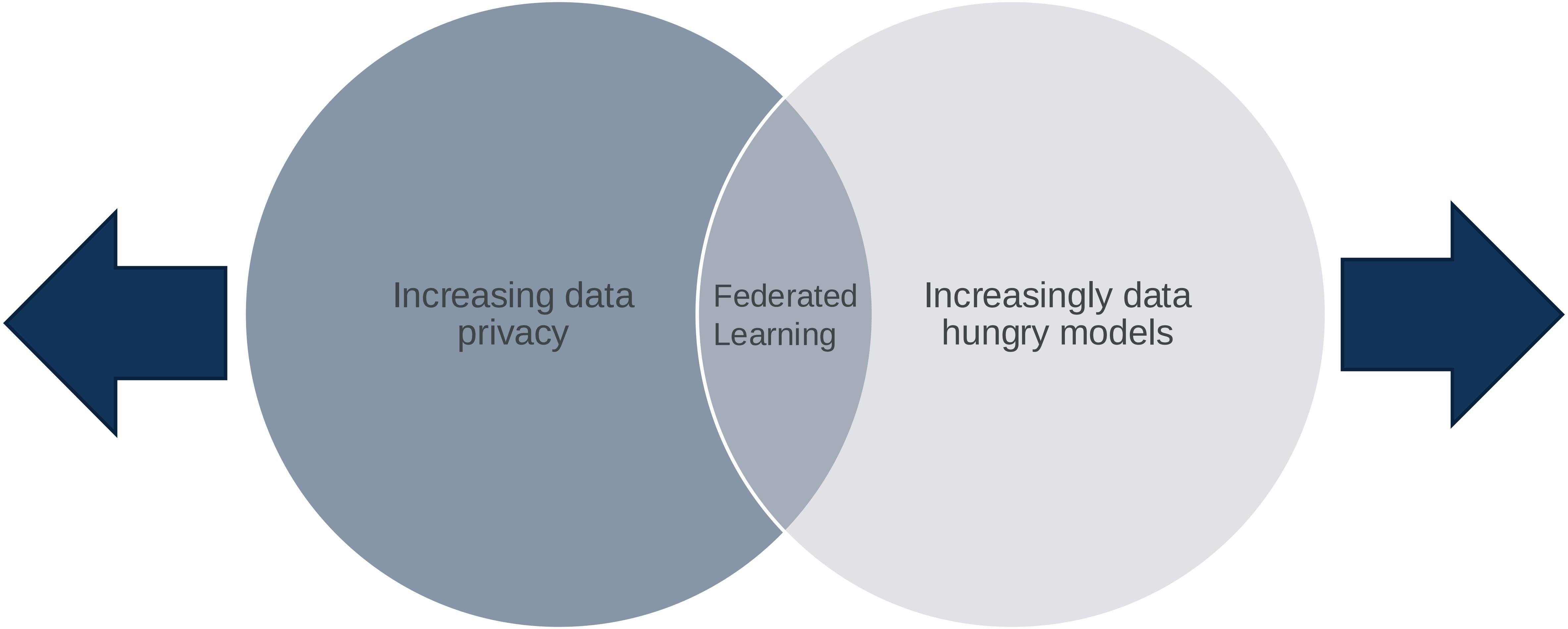
So what do these surveys actually say?

- Crowdflower, 2015: "*66.7% said cleaning and organizing data is one of their most time-consuming tasks*".
  - They didn't report estimates of time spent
- Crowdflower, 2016: "*What data scientists spend the most time doing? Cleaning and organizing data: 60%, Collecting data sets; 19% …*".
  - Only 80% of time spent if you also lump in collecting data as well
- Crowdflower, 2017: "*What activity takes up most of your time? 51% Collecting, labeling, cleaning and organizing data*"
  - Less than 80% and also now includes tasks like labelling of data
- Figure Eight, 2018: Doesn't cover this question.
- Figure Eight, 2019: "*Nearly three quarters of technical respondents 73.5% spend 25% or more of their time managing, cleaning, and/or labeling data*"
  - That's pretty far from 80%!
- Kaggle, 2017: Doesn't cover this question
- Kaggle, 2018: "*During a typical data science project, what percent of your time is spent engaged in the following tasks? ~11% Gathering data, 15% Cleaning data…*"
  - Again, much less than 80%

Do data scientists spend 80% of their time cleaning data? Turns out, no? – Lost Boy (ldodds.com)

GDPR

The Push and Pull



Netflix

NETFLIX

Netflix Prize

Home | Rules | Leaderboard | Register | Update | Submit | Download

Leaderboard

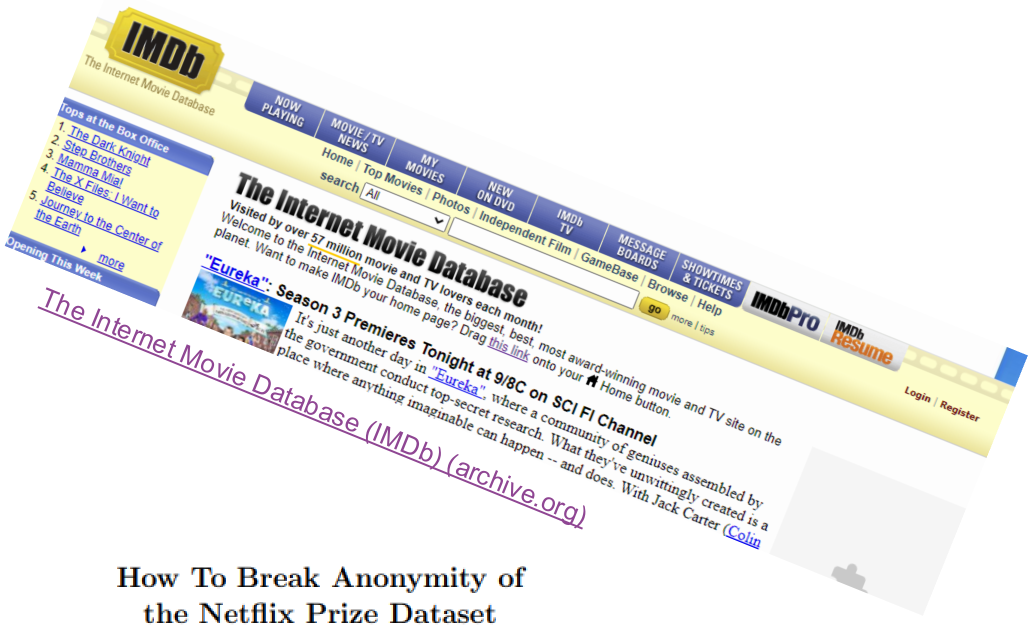
Display top 40 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
1	PragmaticTheory	0.8584	9.78	2009-06-16 01:04:47
2	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-06-15 18:03:55
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8620	9.40	2009-06-17 13:41:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8640	9.19	2009-06-09 22:24:53
9	xlvector	0.8640	9.19	2009-06-17 12:47:27
10	BruceDengDaoCiYiYou	0.8641	9.18	2009-06-17 17:08:31
11	Ces	0.8642	9.17	2009-06-12 23:04:25
12	majia2	0.8642	9.17	2009-06-15 03:35:00
13	xiangliang	0.8642	9.17	2009-06-13 16:35:35
14	Feeds2	0.8647	9.11	2009-06-16 22:21:19
15	Just a guy in a garage	0.8650	9.08	2009-05-24 10:02:54
16	Team ESP	0.8653	9.05	2009-06-16 05:25:11
17	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
18	NewNetflixTeam	0.8657	9.01	2009-05-31 07:30:22
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14

WIRED

```
For the Netflix Prize, your program must predict the all ratings the customers
gave the movies in the qualifying dataset based on the information in the
training dataset.

The format of your submitted prediction file follows the movie and customer id,
date order of the qualifying dataset.  However, your predicted rating takes the
place of the corresponding customer id (and date), one per line.

For example, if the qualifying dataset looked like:

111:
3245,2005-12-19
5666,2005-12-23
6789,2005-03-14
225:
1234,2005-05-26
3456,2005-11-07

then a prediction file should look something like:
111:
3.0
3.4
4.0
225:
1.0
2.0

which predicts that customer 3245 would have rated movie 111 3.0 stars on the
19th of Decemeber, 2005, that customer 5666 would have rated it slightly higher
```

Netflix Prize data | Kaggle



The Internet Movie Database (IMDb) (archive.org)

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

March 2, 2007

Netflix answers this question as follows:

> No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy, which you can review here. Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isn't a privacy problem is it?

Even more sensitive

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

1. Abstract

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently. It was found that combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are publicly available in this form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

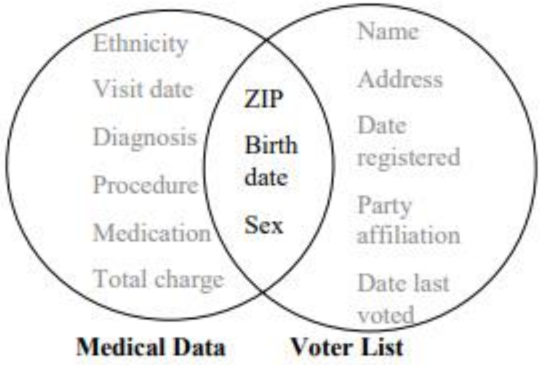


Figure 1 Linking to re-identify data

“

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.

“Anonymized” data really isn't—and here's why not | Ars Technica

Sweeney, Abu and Winn — Identifying Participants in the Personal Genome Project by Name

Identifying Participants in the Personal Genome Project by Name

Latanya Sweeney, Akua Abu, Julia Winn

Harvard College
Cambridge, Massachusetts
latanya@fas.harvard.edu, aabu@college.harvard.edu, jwinn@post.harvard.edu

We linked names and contact information to publicly available profiles in the Personal Genome Project. These profiles contain medical and genomic information, including details about medications, procedures and diseases, and demographic information, such as date of birth, gender, and postal code. By linking demographics to public records such as voter lists, and mining for names hidden in attached documents, we correctly identified 84 to 97 percent of the profiles for which we provided names. Our ability to learn their names is based on their demographics, not their DNA, thereby revisiting an old vulnerability that could be easily thwarted with minimal loss of research value. So, we propose technical remedies for people to learn about their demographics to make better decisions.

INTRODUCTION

The freedom to decide with whom to share one's own medical and genomic information seems critical to protecting personal privacy in today's data-rich networked society. Individuals are often in the best position to make decisions about sharing extensive amounts of personal information for many and thousands of people get subsequently harmed doing so, policy makers may respond and take away the freedom to make personal data sharing decisions, thereby depriving society of individual choice. To make smarter decisions, people need an understanding of actual risks and ways technology can help.

BACKGROUND

Launched in 2006, the Personal Genome Project (PGP) aims to sequence the genotypic and phenotypic information of 100,000 informed volunteers and display it publicly online in an extensive public database [1]. Information provided in the PGP includes DNA information, behavioral traits, medial conditions, physical characteristics, and environmental factors. A general argument for the disclosure of such information is its utility. The PGP founders believe this information will aid researchers in establishing correlations between certain traits and conducting research in personalized medicine. They also foresee its use as a tool for individuals to learn about their own genetic risk profiles for disease, uncover ancestral data, and examine biological

Smartphone Federated Learning Pipeline

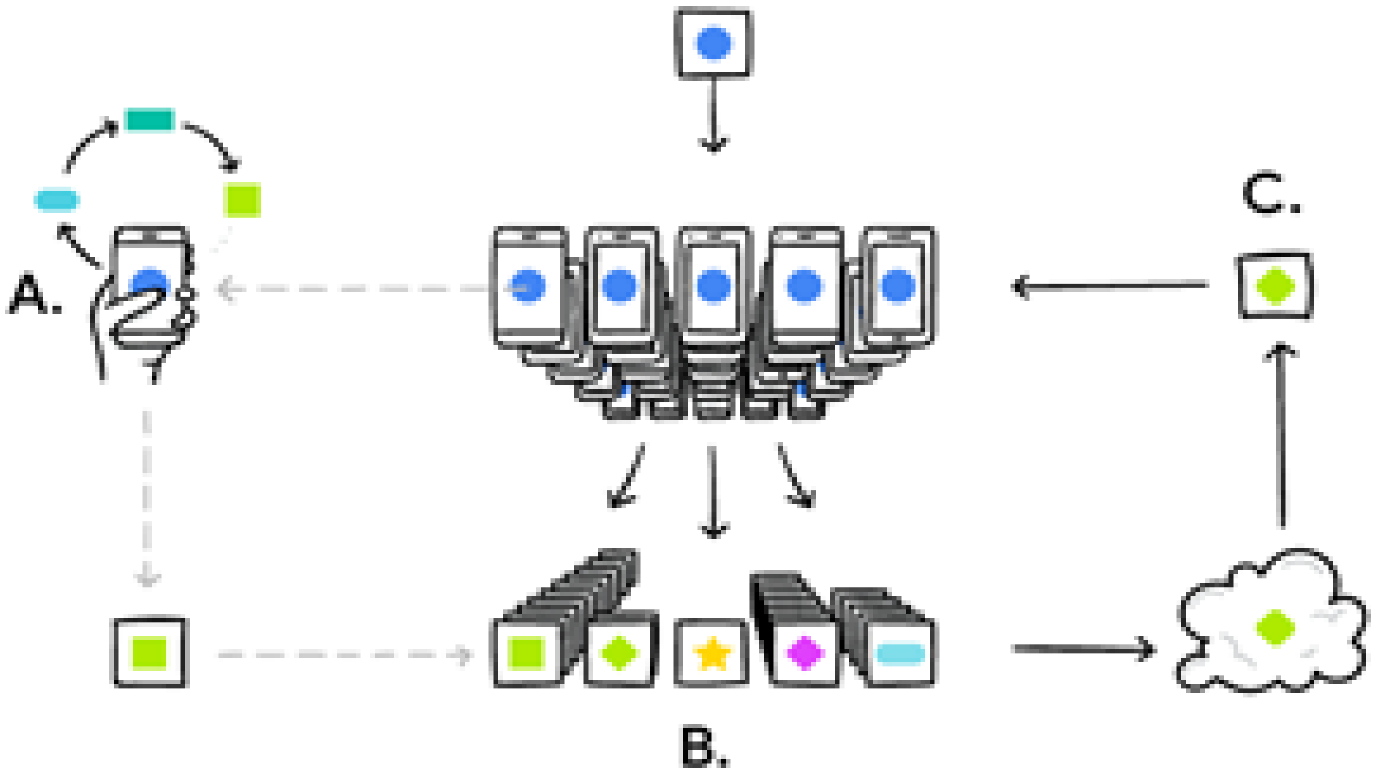


Figure 2: Google Federated Learning

A) your phone personalises the model locally depending on your usage;

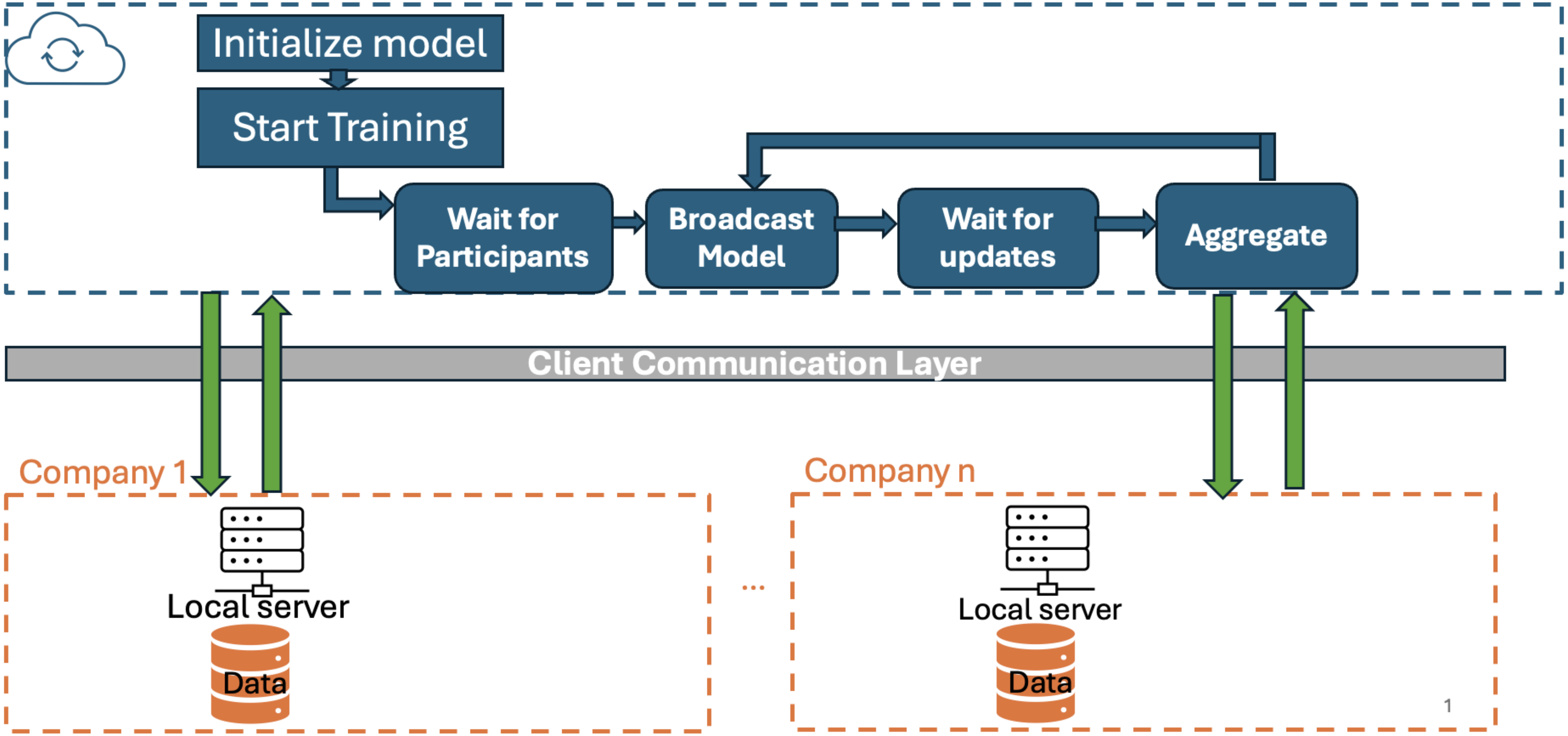
B) many users' updates are aggregated;

C) the aggregated updates form a consensus change to the shared model; and

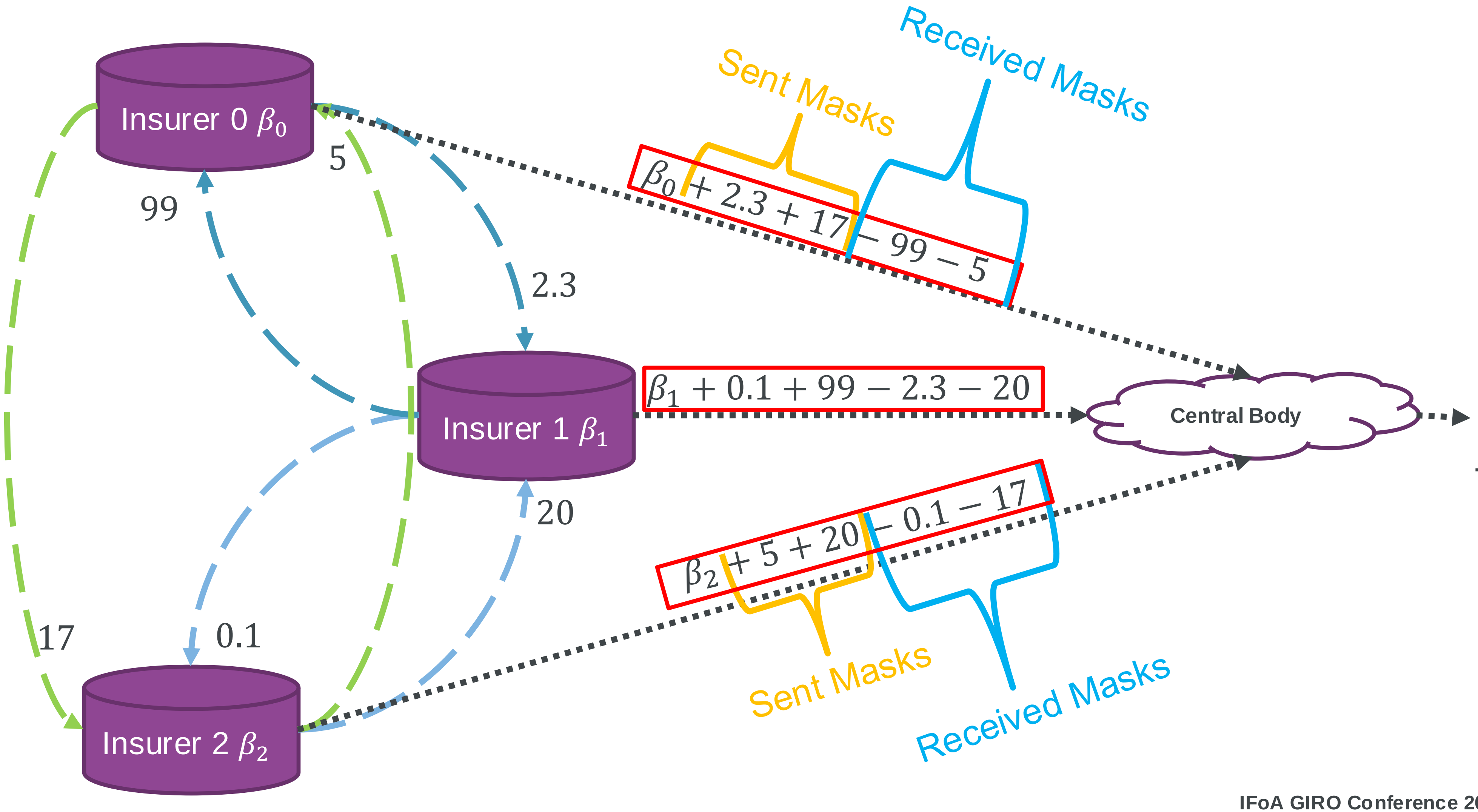
D) the shared models are updated.

[Google AI Blog: Federated Learning: Collaborative Machine Learning without Centralized Training Data (googleblog.com)]

Insurance Federated Learning Pipeline



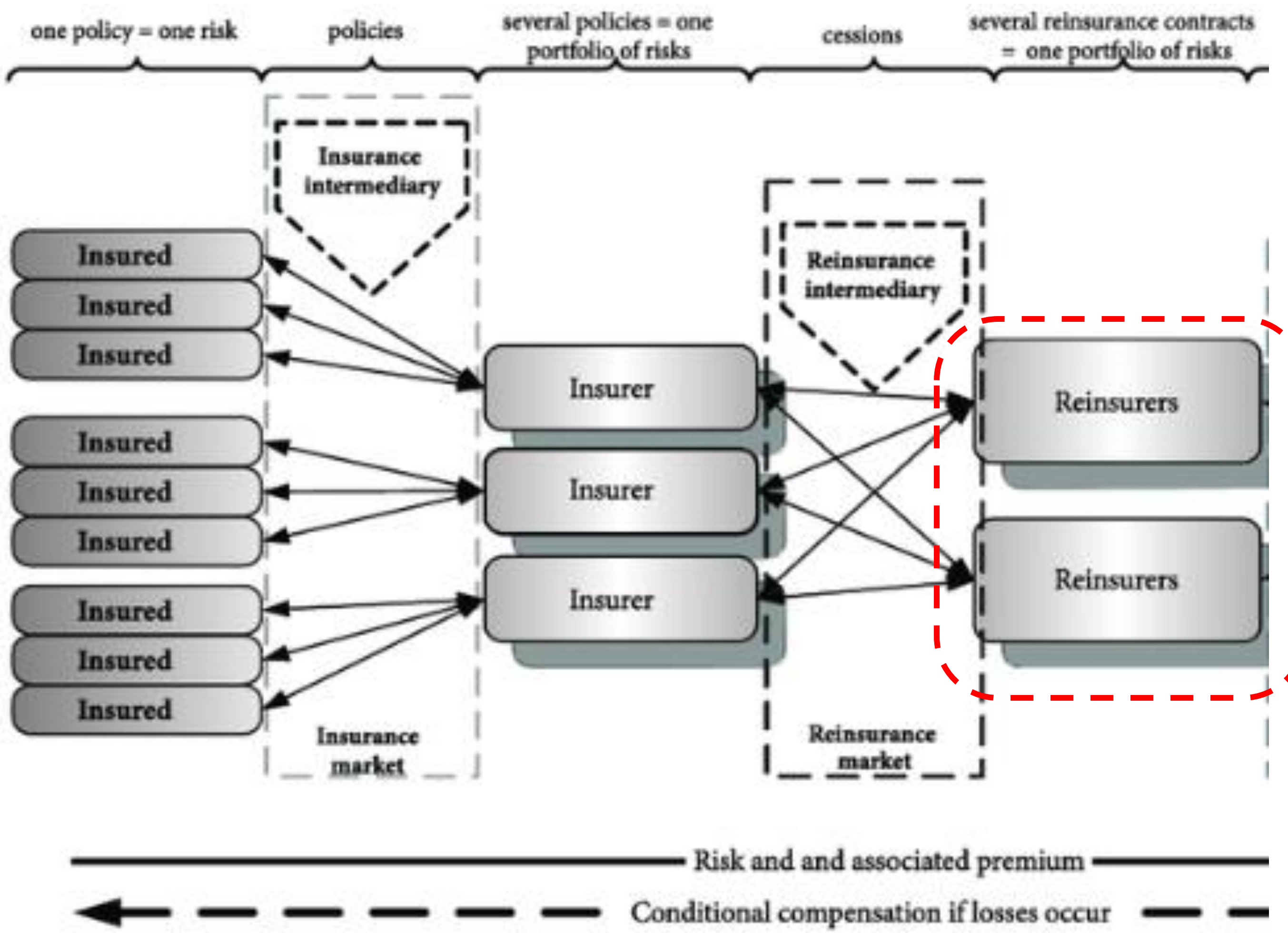
Need to encrypt parameters but maintain the average



$$\frac{[\beta_0 \cancel{+2.3+17-5-99}] + [\beta_1 \cancel{-2.3-20+0.1+99}] + [\beta_2 \cancel{-17+20+5-0.1}]}{3}$$

$$= \frac{\beta_0 + \beta_1 + \beta_2}{3}$$

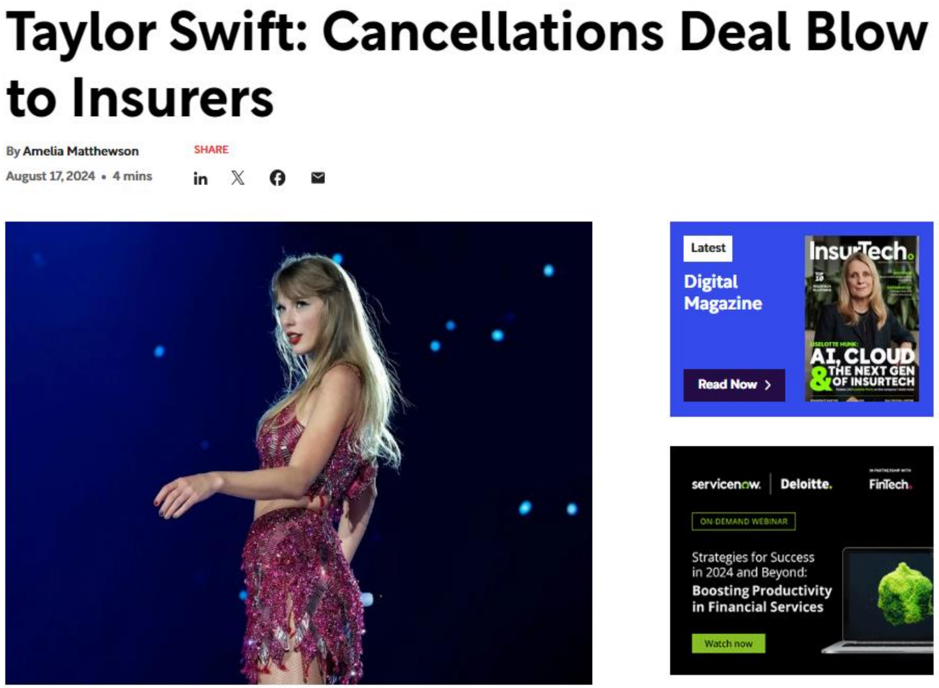
Application: Reinsurance



- Reinsurer provides protection to insures
- The pricing is determined by collecting data from different insures on the loss experience
- With Federated Learning, reinsurance could **better** comply with data privacy.

Pooling data to determine price on the reinsurance contracts for different products

Application: Lloyd's of London



How does Lloyd’s benefit from Pooling Data?

- Different companies that write the same insurance products uses their own internal datasets to predict risk and sharing data through FL can help aggregate data to enhance risk pricing
- Some insurance products (e.g. space shuttle insurance) can have very little insurance claims data for building pricing models due to the nature of the product
- Lloyd’s of London operates globally, they may be able to share diverse datasets via FL without centralising data

Data – French motor claims

- The **freMTPL2freq** car insurance claims dataset – Publicly available
- 677,991 motor third-party liability policies (observed on a year)

Table 1. Description of data, fields, and preprocessing transformations used in experiment

Field	Description	Transformation
IDpol	Unique policy number	Dropped
ClaimNb	Number of claims on the given policy	Capped at 4
Exposure	Total exposure in yearly units	Capped at 1
Area	France area code (categorical, ordinal)	Ordinally encoded e.g. $A:1, B:2, C:3$ etc.
VehPower	Horse power of the car (categorical, ordinal)	*MinMaxScaler*
VehAge	Age of the car in years	*MinMaxScaler*
DrivAge	Age of the driver in years	*MinMaxScaler*
BonusMalus	Bonus-malus (i.e. No Claims Discount) level between 50 - 230	*MinMaxScaler* after capping at 150
VehBrand	Car brand (categorical, nominal)	One-hot-encoded
VehGas	Diesel or petrol car (binary)	Ordinally encoded i.e. $Regular:1, Diesel:2$
Density	Density of inhabitants per km2 in the city of the residential address of the driver	*MinMaxScaler* after log transforming
Region	Regions in France prior to 2016 (categorical)	One-hot-encoded

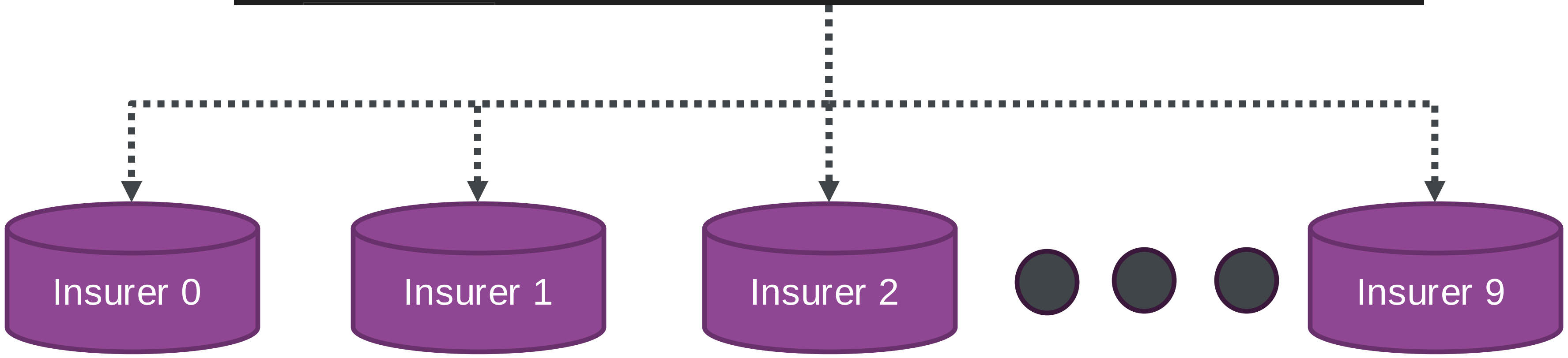
Insurance Federated Learning Use Case

```
df = pd.read_csv('../data/freMTPL2freq.CSV')
df
```

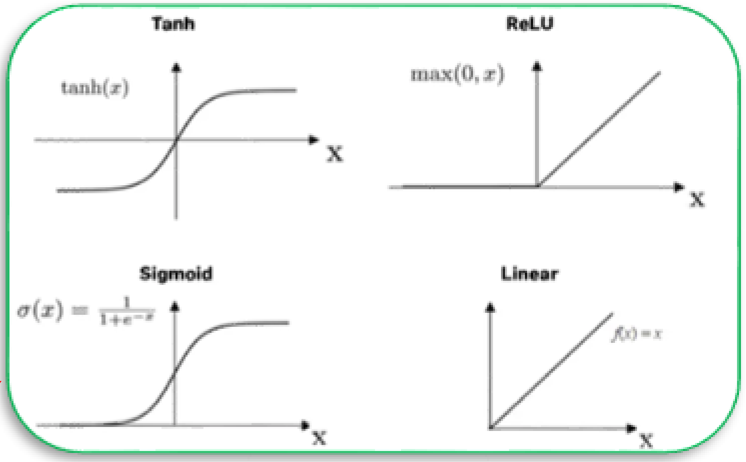
[108] ✓ 0.5s

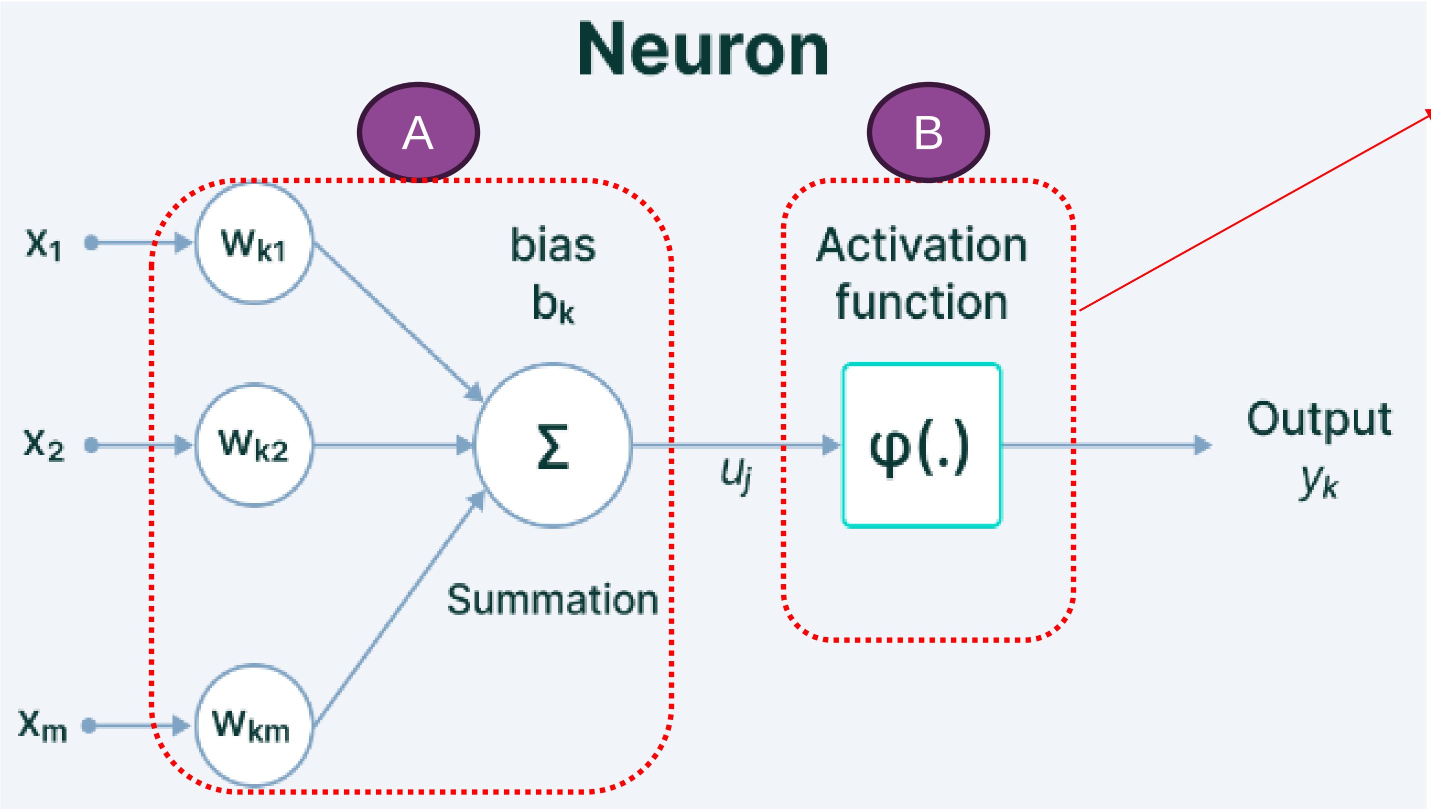
	IDpol	ClaimNb	Exposure	Area	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Density	Region
0	1.0	1	0.10000	D	5	0	55	50	B12	Regular	1217	R82
1	3.0	1	0.77000	D	5	0	55	50	B12	Regular	1217	R82
2	5.0	1	0.75000	B	6	2	52	50	B12	Diesel	54	R22
3	10.0	1	0.09000	B	7	0	46	50	B12	Diesel	76	R72
4	11.0	1	0.84000	B	7	0	46	50	B12	Diesel	76	R72
...	...	...	...	...	...	...	...	...	...	...	...	...
678008	6114326.0	0	0.00274	E	4	0	54	50	B12	Regular	3317	R93
678009	6114327.0	0	0.00274	E	4	0	41	95	B12	Regular	9850	R11
678010	6114328.0	0	0.00274	D	6	2	45	50	B12	Diesel	1323	R82
678011	6114329.0	0	0.00274	B	4	0	60	50	B12	Regular	95	R26
678012	6114330.0	0	0.00274	B	7	6	29	54	B12	Diesel	65	R72

678013 rows × 12 columns



Neural Networks





Neural Networks		GLMs
Weights & Biases (point A)	↔	Coefficients
Activation function (point B)	↔	Link function
Loss objective	↔	Response distribution

[Neural Network from Scratch. Previously in the last article, I had… | by SARVESH DUBEY | Becoming Human: Artificial Intelligence Magazine]

https://www.v7labs.com/blog/neural-network-architectures-guide

https://www.researchgate.net/figure/Fig-3-The-basic-activation-functions-of-the-neural-networksNeural-Networks_fig3_350567223

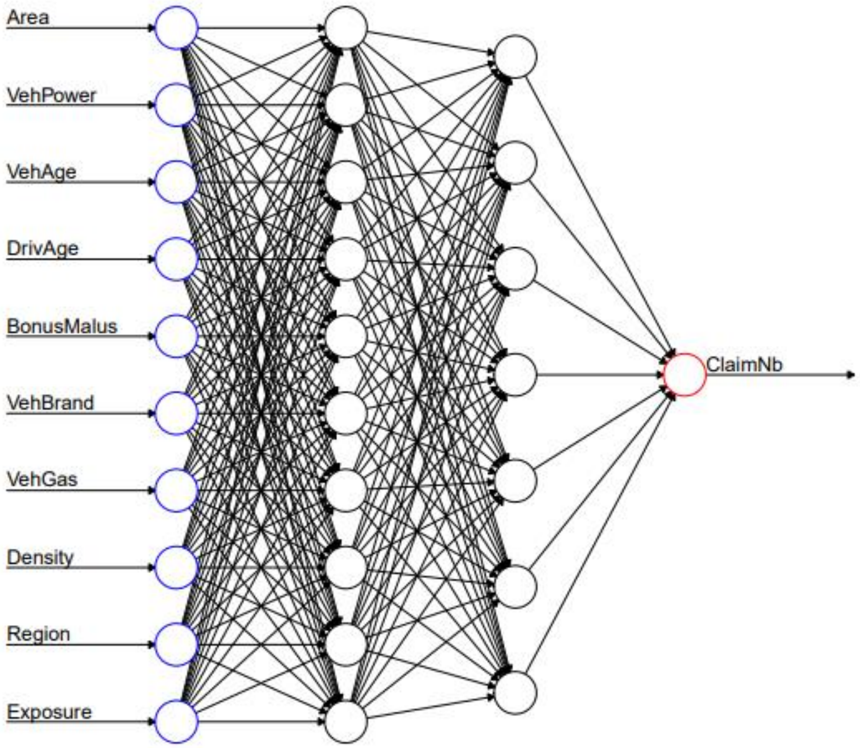
Neural Network Model Setup

Table 2. Neural Network Architecture used in all 3 Scenarios

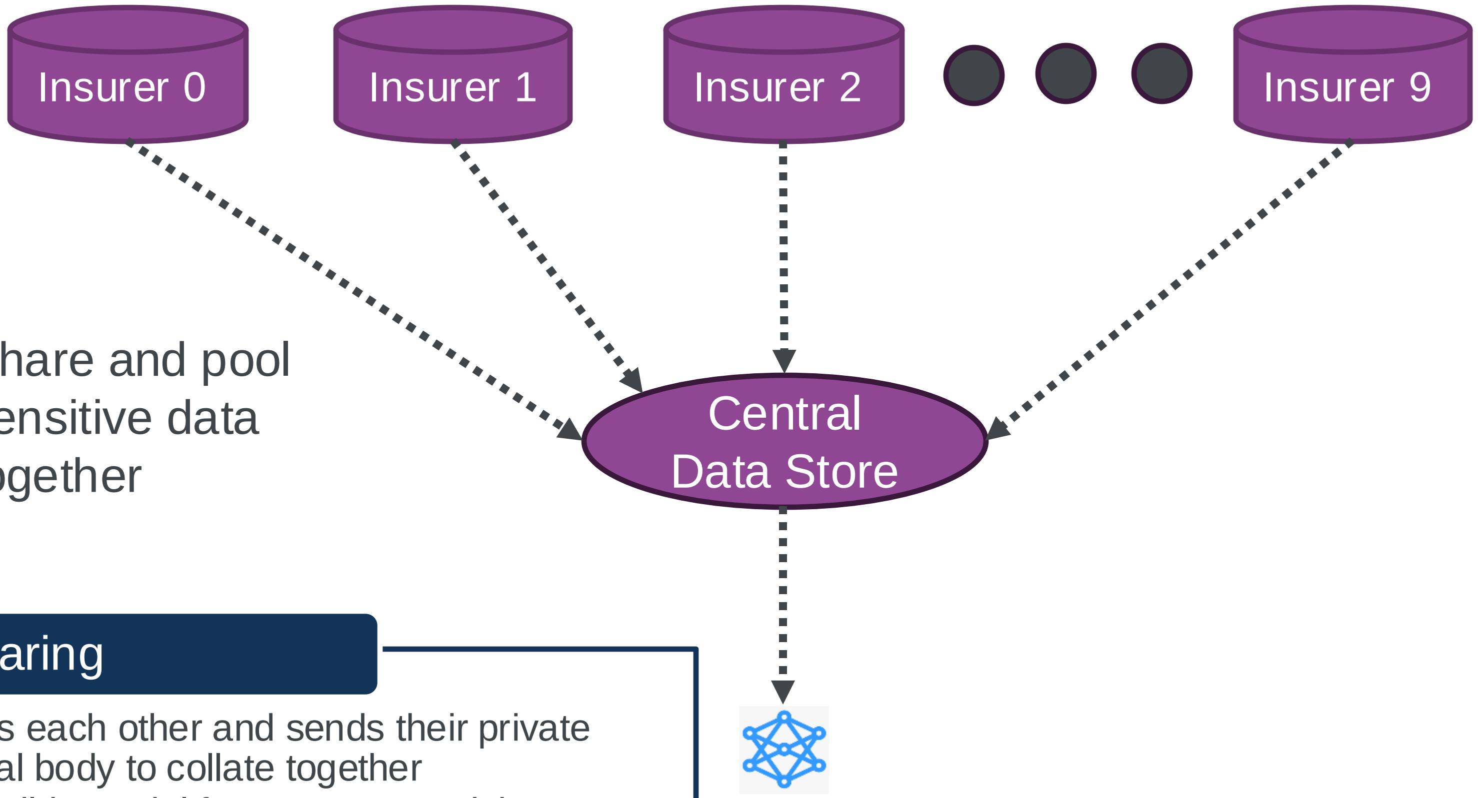
Hyperparameter	Selection
Input neurons	39 based on the preprocessing done in Section 5.2.2
Hidden Layers	2
Output Layer	1 output neuron with exponential link function (to ensure only positive frequencies are predicted)
Optimiser	*NAdam*
Activation Function	*tanh*
Loss Function	Negative Poisson Log Likelihood
Initialisation	Xavier
Epochs	300

Table 3. Hyperparameter Search Space Considered in all 3 Scenarios

Hyperparameter	Search Space
Learning Rate	[0.001, 0.002, 0.01]
Number neurons in Hidden Layer 1	[5, 10, 15, 20]
Number neurons in Hidden Layer 2	[5, 10, 15, 20]
Batch Size	[500, 1,000, 5,000, 10,000]



Global Model Scenario – 10 insurers, 1 models

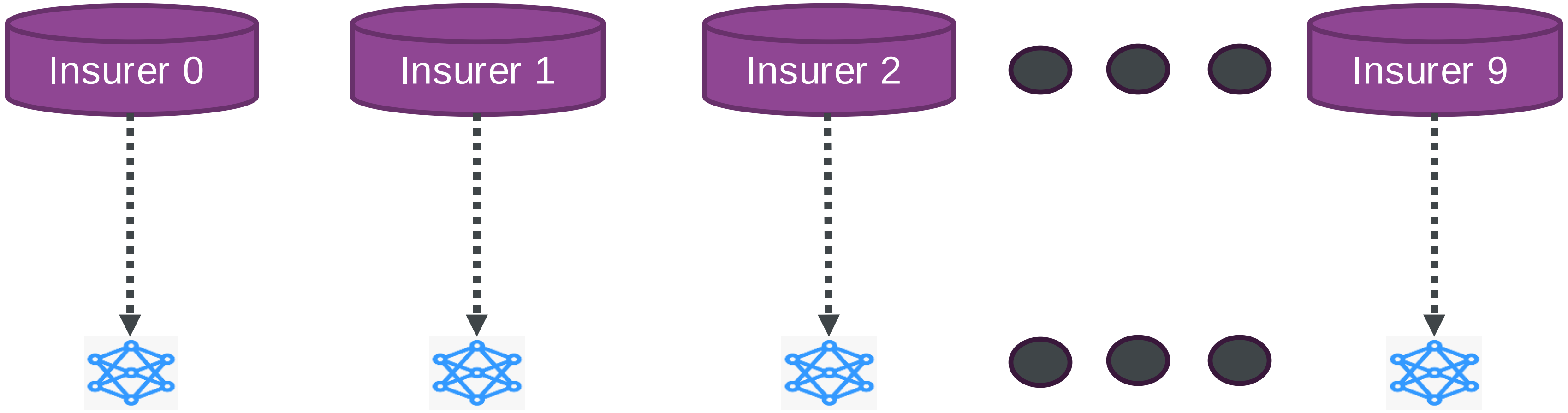


Share and pool sensitive data together

Sharing is Caring

- Everyone trusts each other and sends their private data to a central body to collate together
- Central body builds model for everyone and then sends back to companies
- **A.k.a. 1 "Global" model** as it uses all the data and applies to everyone

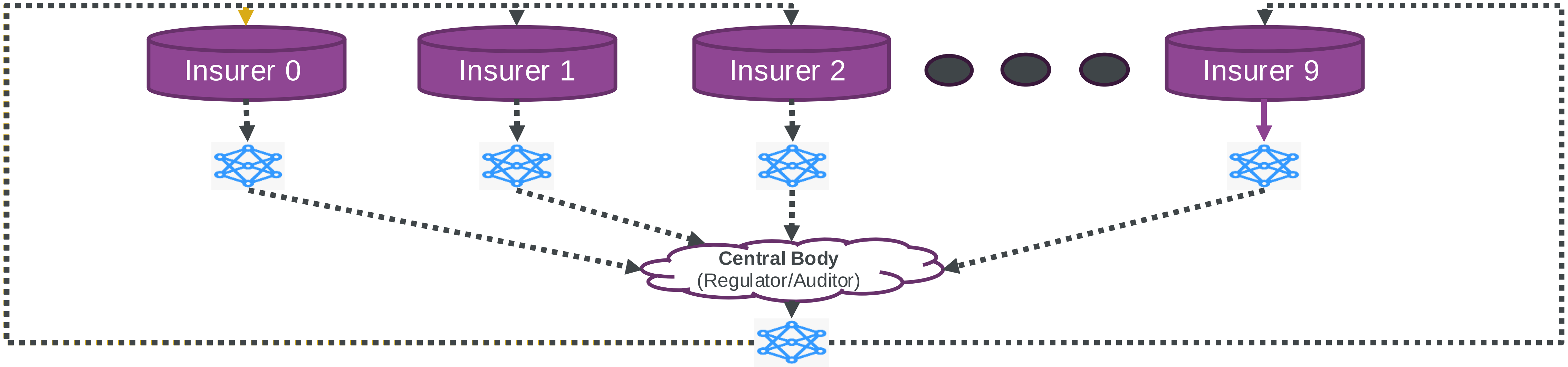
Partial Model Scenario – 10 insurers, 10 models



Each insurer builds their own model just using their data

- No one trusts anyone
- Low volume of data used to build models which could be more relevant to company although may not be credible
- **A.k.a. 10 "Partial" models** as each company's model only has partial access to the whole market data

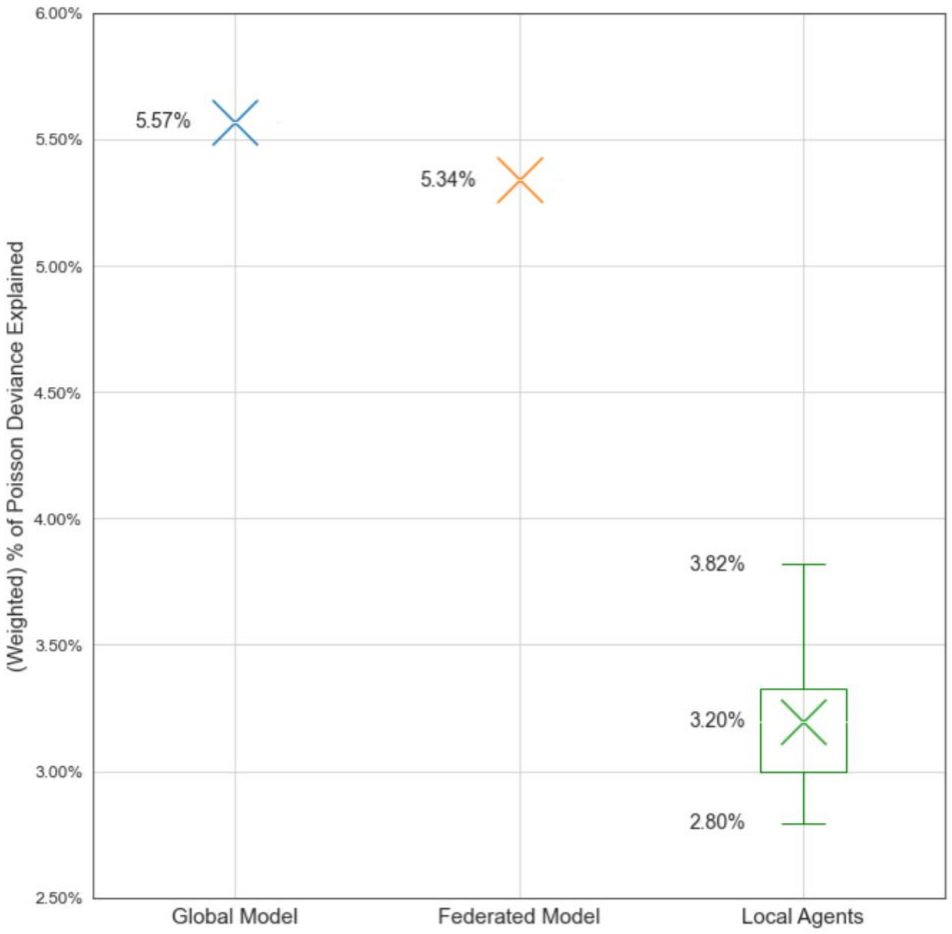
Federated Model Scenario – 10 insurers, 1 model



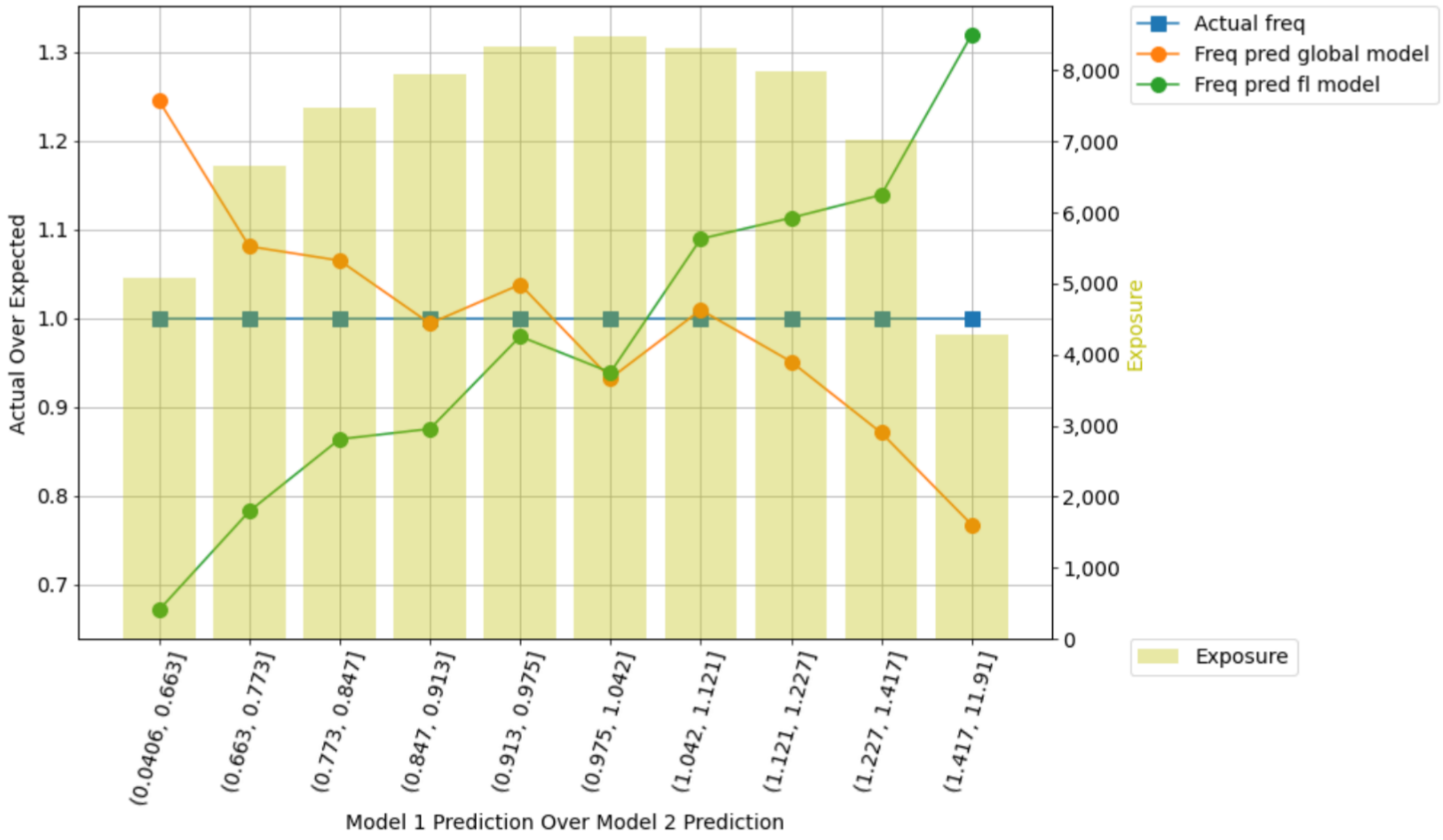
"United Federation"

- Everyone keeps their 10th of their data to themselves
- However they securely share their parameters with central body
- Central body securely averages all the insurer's parameters and shares back
- Bringing the model to the data rather than bringing the data to the model
- **A.k.a. 1 "Federated" model**

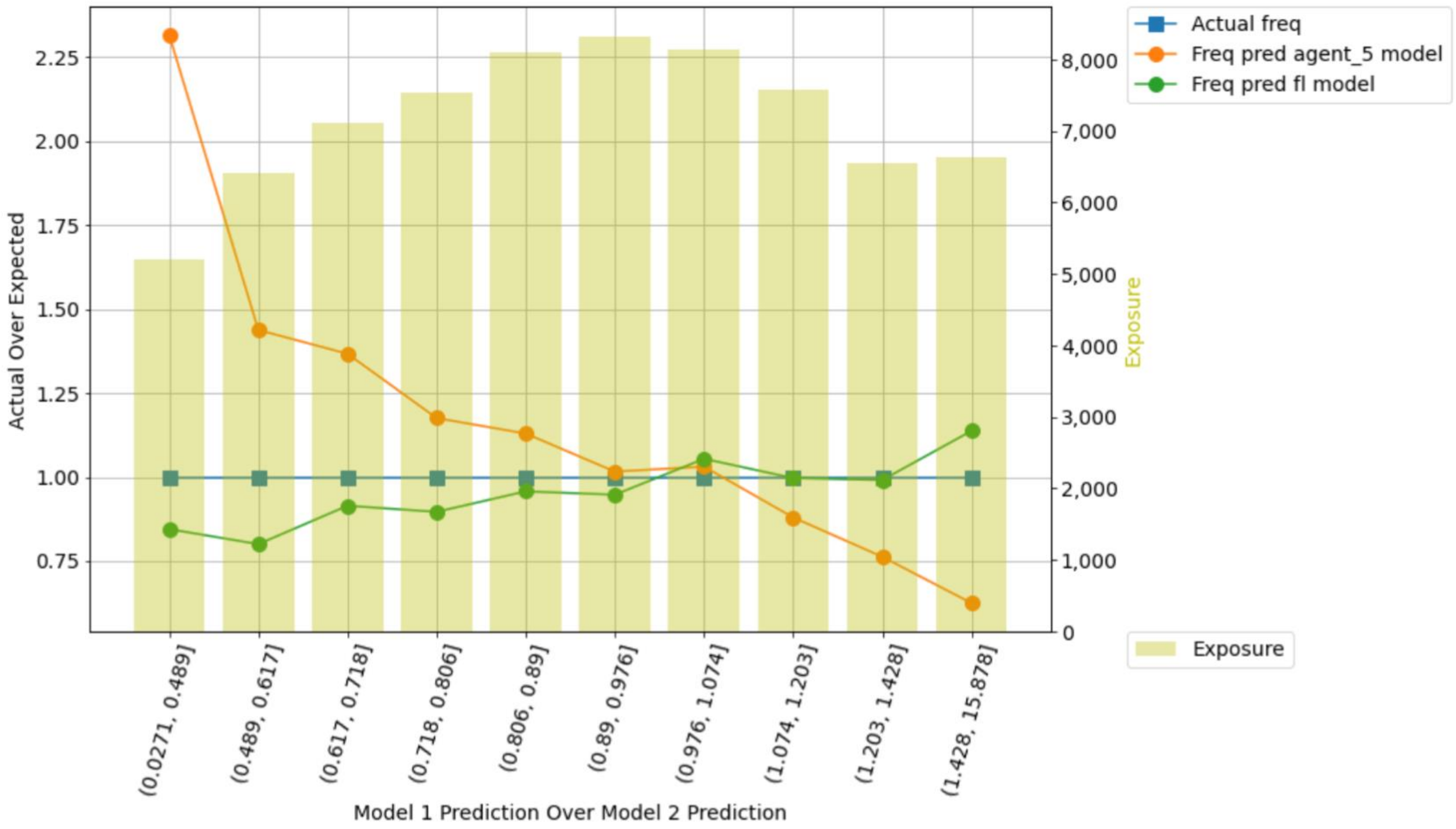
Comparison of results



Comparison of results



Comparison of results



Limitations

Redistribution weights

1.Data

- Imbalance – quality, size, etc
- Heterogeneity – non IID

Information on Age & gender? ….*Vertical FL?*

2.Feature

- Uniformity of feature space across insurers
- i.e. Same number of column

Identical transformation

3.Scale

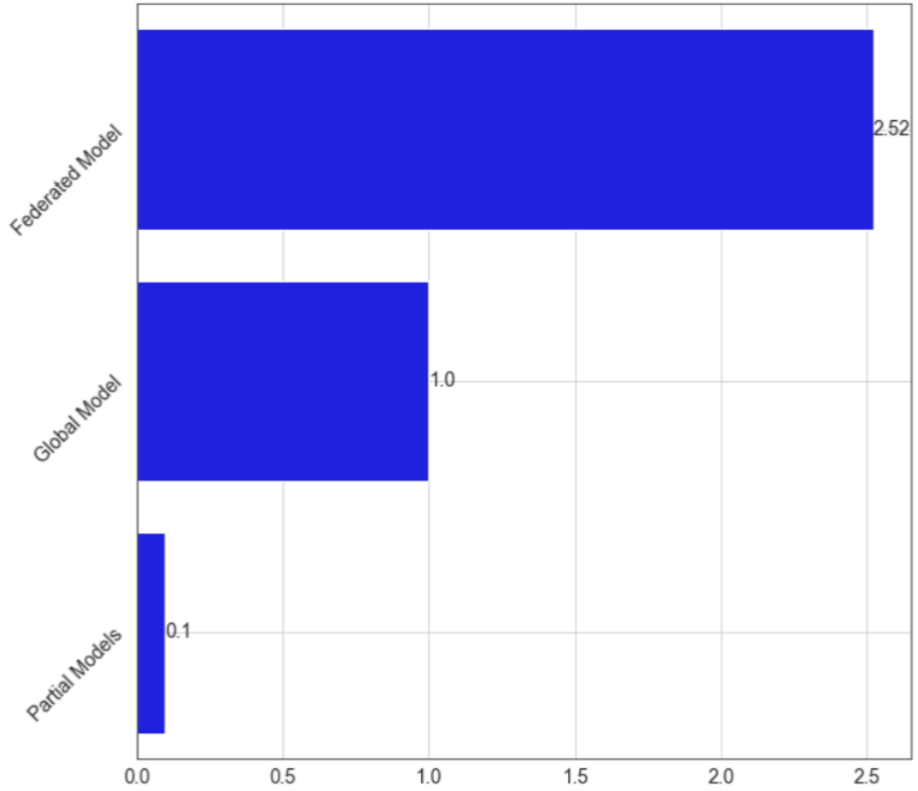
- Individual custom features doesn't work
- e.g. MinMaxScaler – same range applied to all insurers

Uniform naming

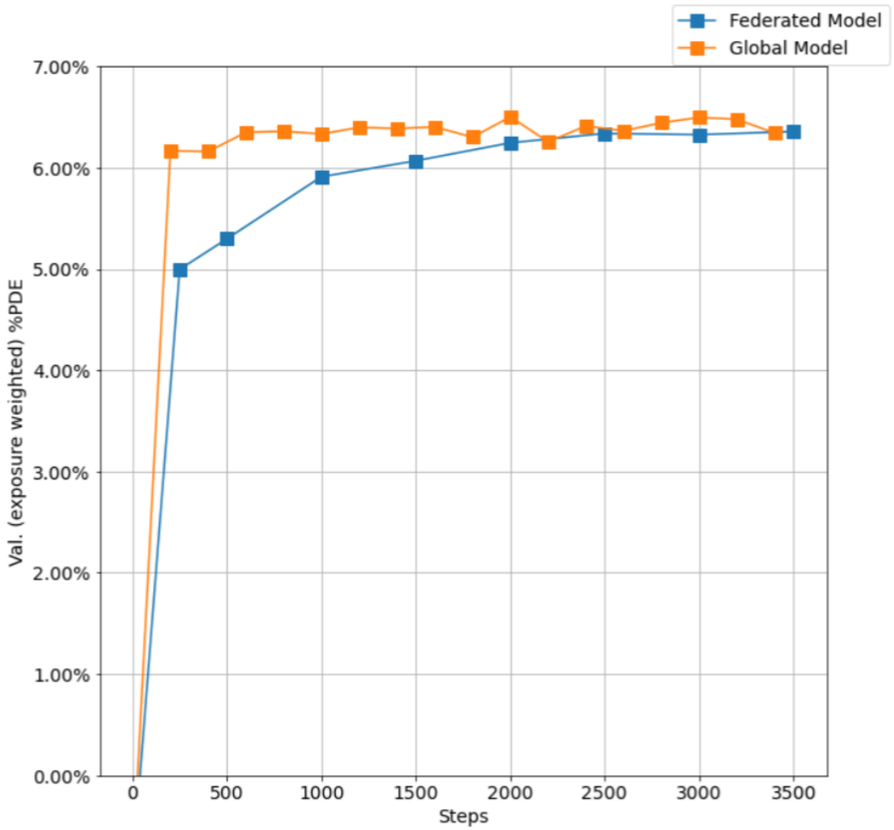
4.Encode

- Aligned definition of division and granularity
- e.g. Car brand as "B2" and "B3"?

What’s the catch?

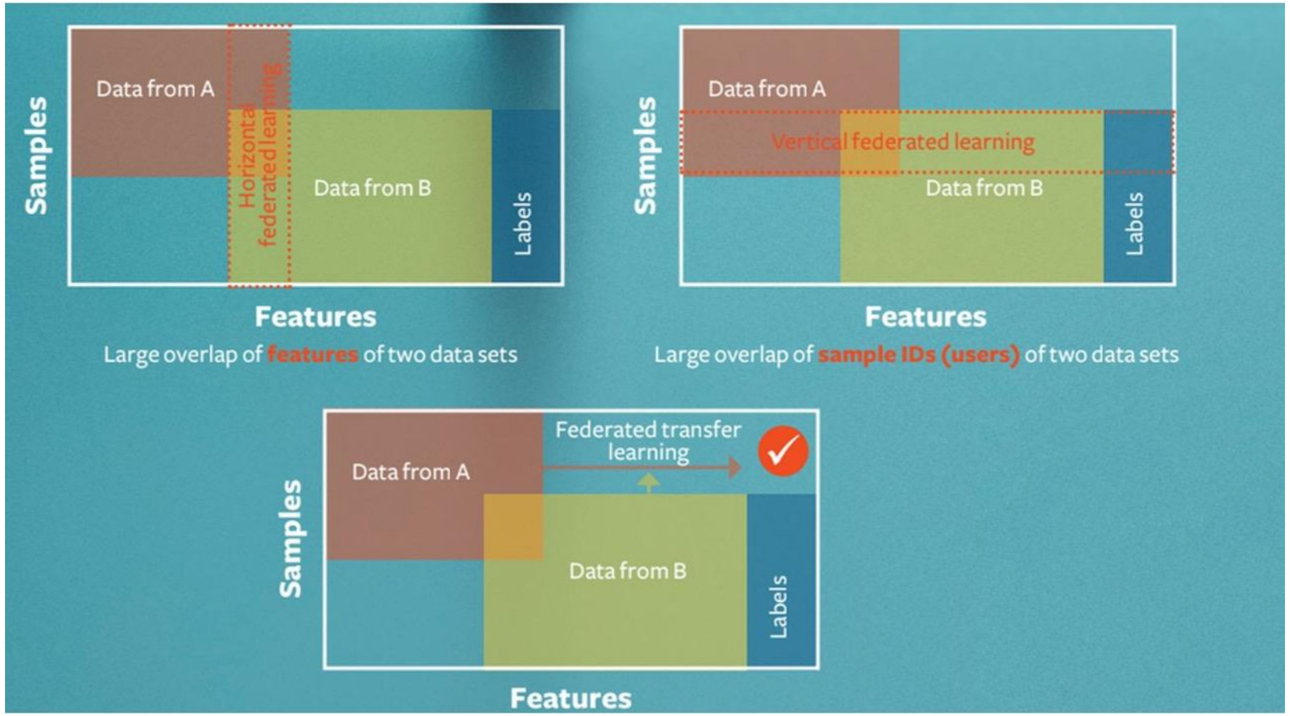


(a) Relative increase in observed wall time to train the models compared to training the Global Model.



(b) Exposure Weighted Validation % PDE of the Global and Federated Models over different number of parameter update steps.

Federated Learning Types



Federated Learning Challenges

Beyond data...

Data

Covered in the previous slide

System and Operational

- Model Convergence
- Fault Tolerance
- Client Dropout

Adoption Barriers

- Integration complexity
- Cultural Resistance
- Skill Gap

Questions

Comments

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter.