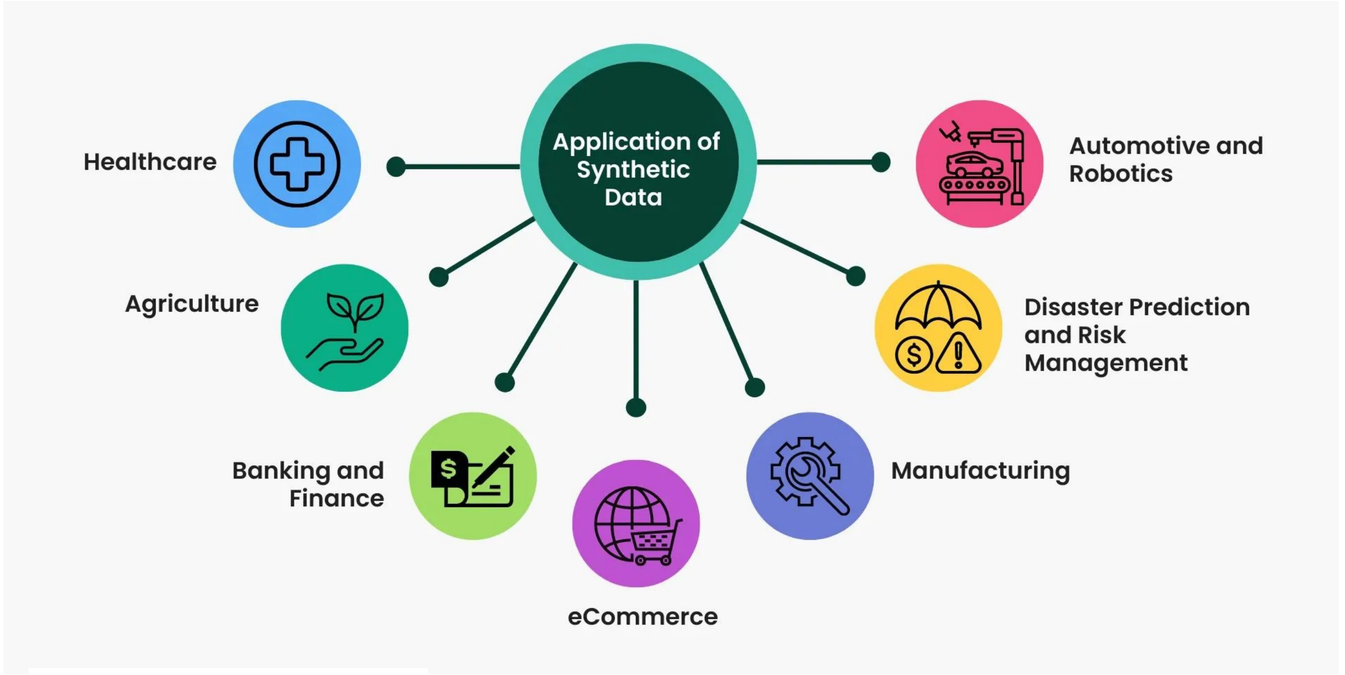
SiloFuse

Cross-silo Synthetic Data Generation with Latent Tabular Diffusion Models

Aditya Shankar, Hans Brouwer, Rihan Hai, Lydia Chen

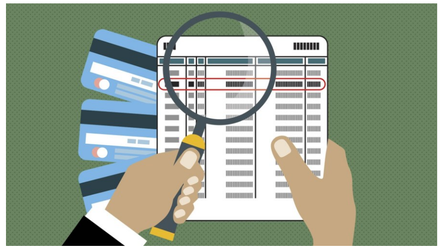
Synthetic data is everywhere

Predictive maintenance



Simulating crop patterns

Patient risk classification

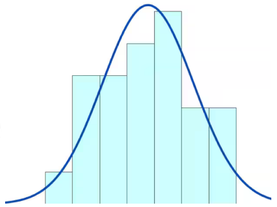


Fraud detection

WHY synthetic data

- **Privacy**

- Mitigating **class imbalance**

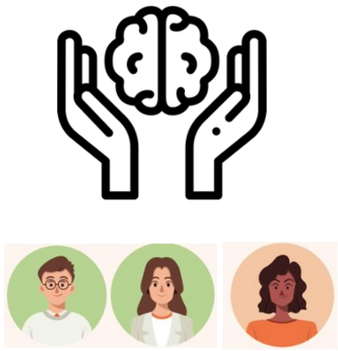


- **Faster** collection

- **Cheaper** generation

Our problem scenario

- Dataset often **distributed across silos**
- Our case**: Vertically partitioned data**
  - ***Feature*** partitioned data
- Joint learning beneficial
  - High rate **AND** high stress = higher risk
- ***How* to synthesize?**



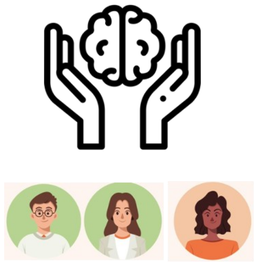
Patient ID	Stress level
#78349	High
#21567	Low
#32171	**High**

Patient ID	Heart rate
#78349	80
#21567	67
#32171	**110**

Naive solution

Centralise then synthesise

Patient ID	Heart rate
#78349	80
#21567	67
#32171	**110**

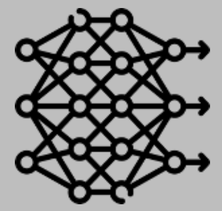


Patient ID	Stress level
#78349	High
#21567	Low
#32171	**High**

Server

Patient t ID	Rate	Stress
#78349	80	High
#21567	67	Low
#32171	**110**	**High**

Generator



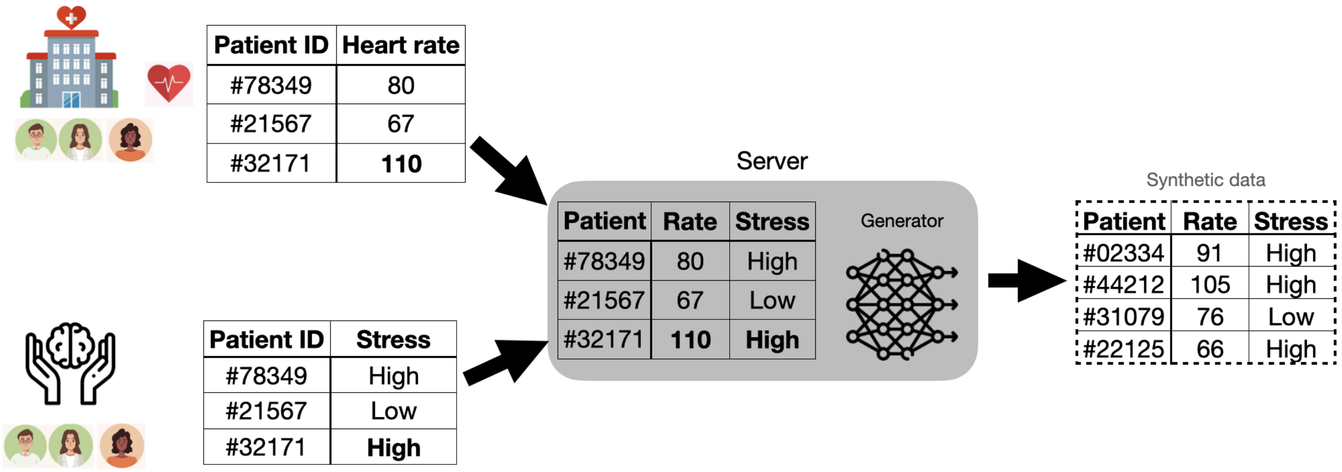
Synthetic data

Patient t ID	Rate	Stress
#02334	91	High
#44212	105	High
#31079	76	Low
#22125	66	High

Challenges

1 Data Privacy

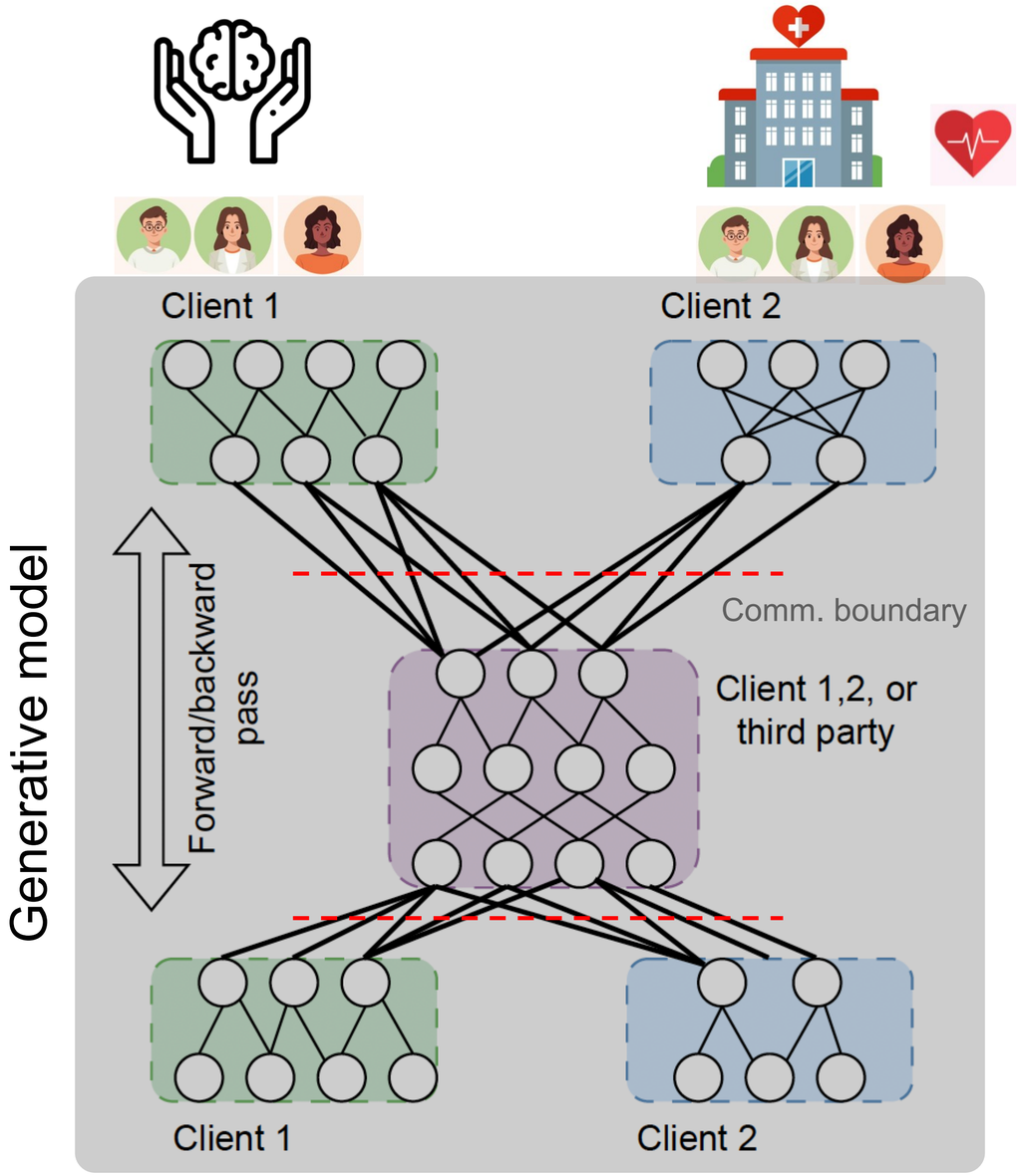
- Centralizing violates privacy
- **Who** plays the server role?
- Train **without moving** data?
  - ***Federated Learning? Encryption?***



#2 Model training

Without moving data

- Keep data local: **Federated learning**
- More **iterations** —> More **comm.**
- ***Improve training efficiency***



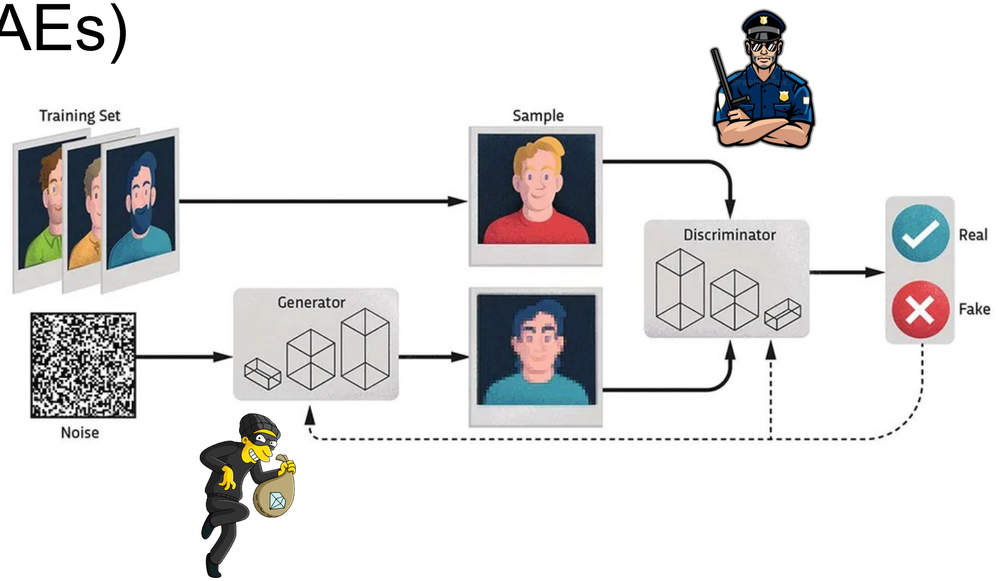
#3 What generative models?

- Classic: Bayesian, Variational autoencoders (VAEs)
- **Generative Adversarial Networks** (GANs)
  - Unstable training
- Capture feature links **across silos?**

Party 1

[Red, Yellow]

Party 2



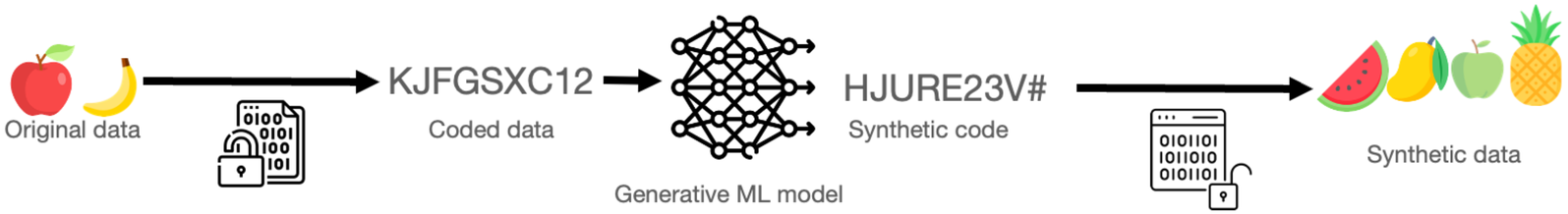
- ***Need stable generator while preserving cross-silo associations***

Solution

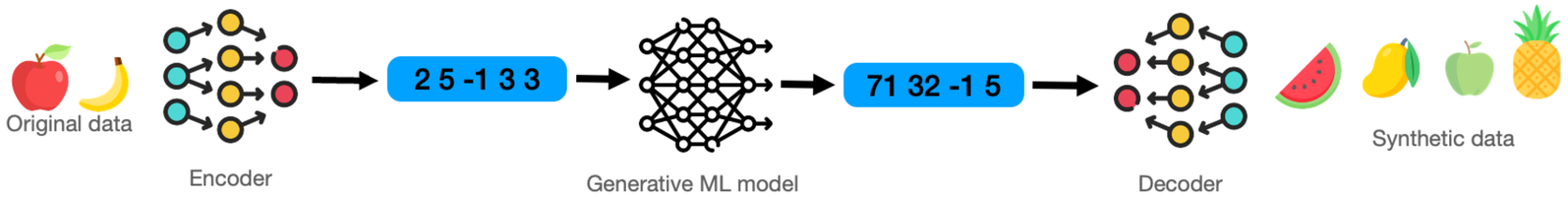
1 Data privacy

Latent models

- Inspiration from cryptography



- ***Latent*** generative models

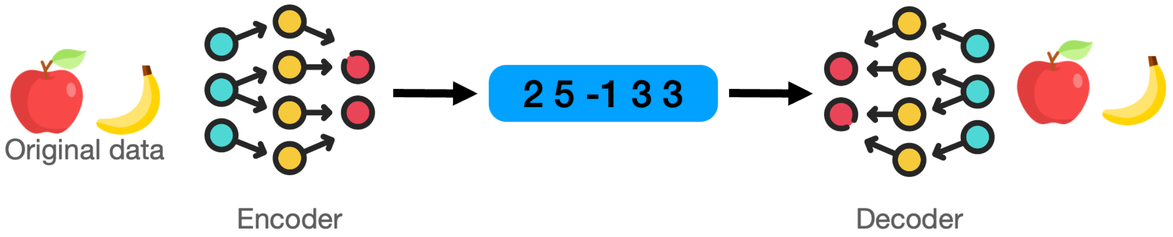


- ***Model is the "key" to your data***

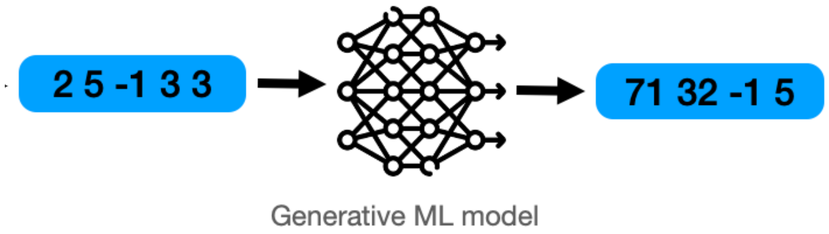
#2 Model training

Stacked training

- Step 1: **Local training (client)**



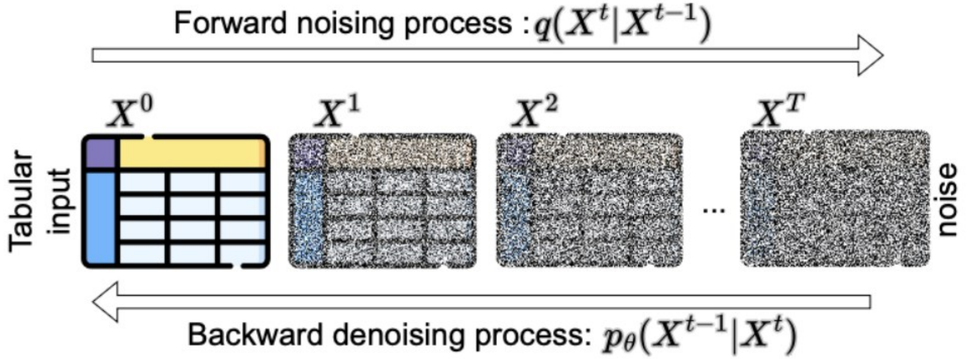
- Step 2: **Latent generator training (server)**

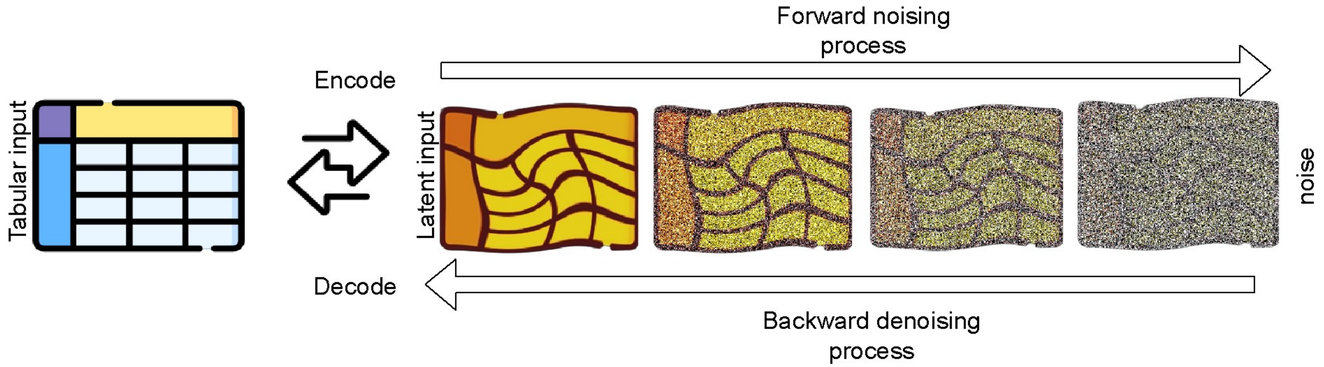


#3 Generative model

Tabular diffusion

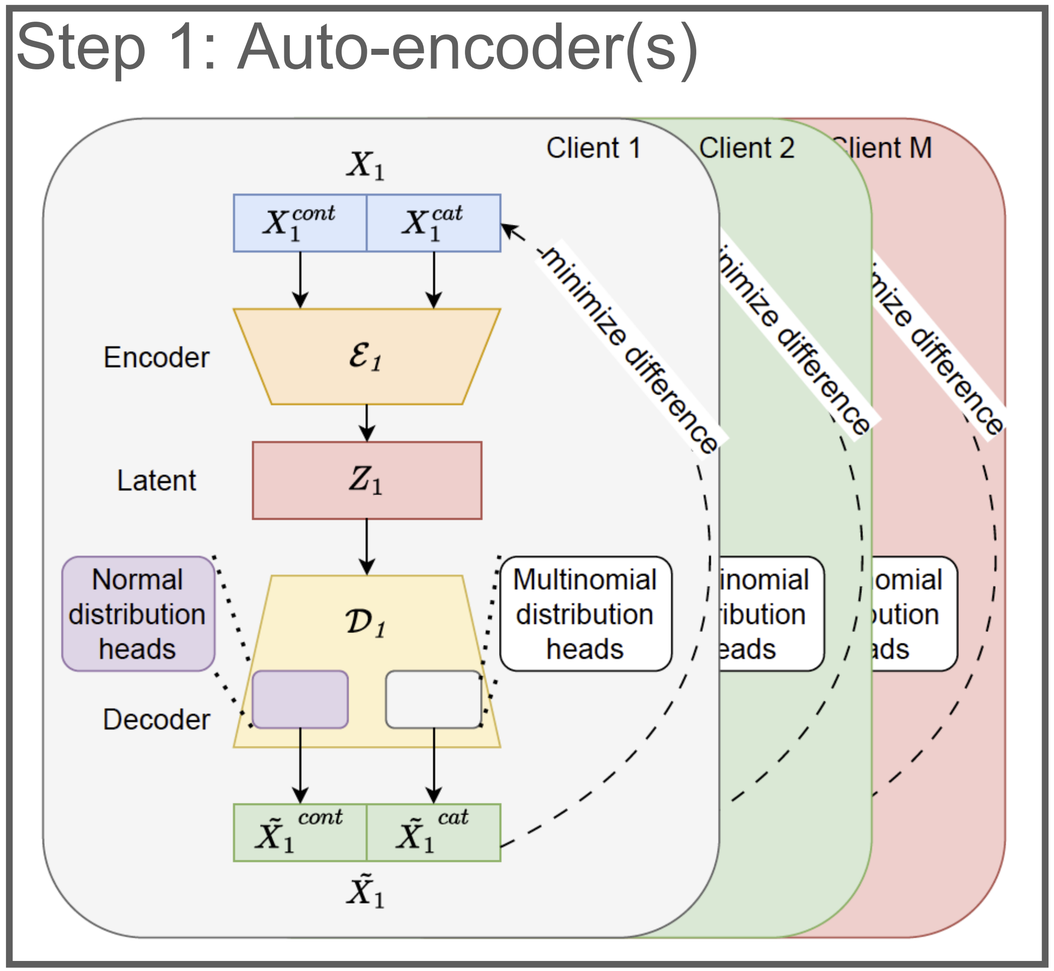
- **Stable** training process
- **Privacy?** ***Latent tabular diffusion***

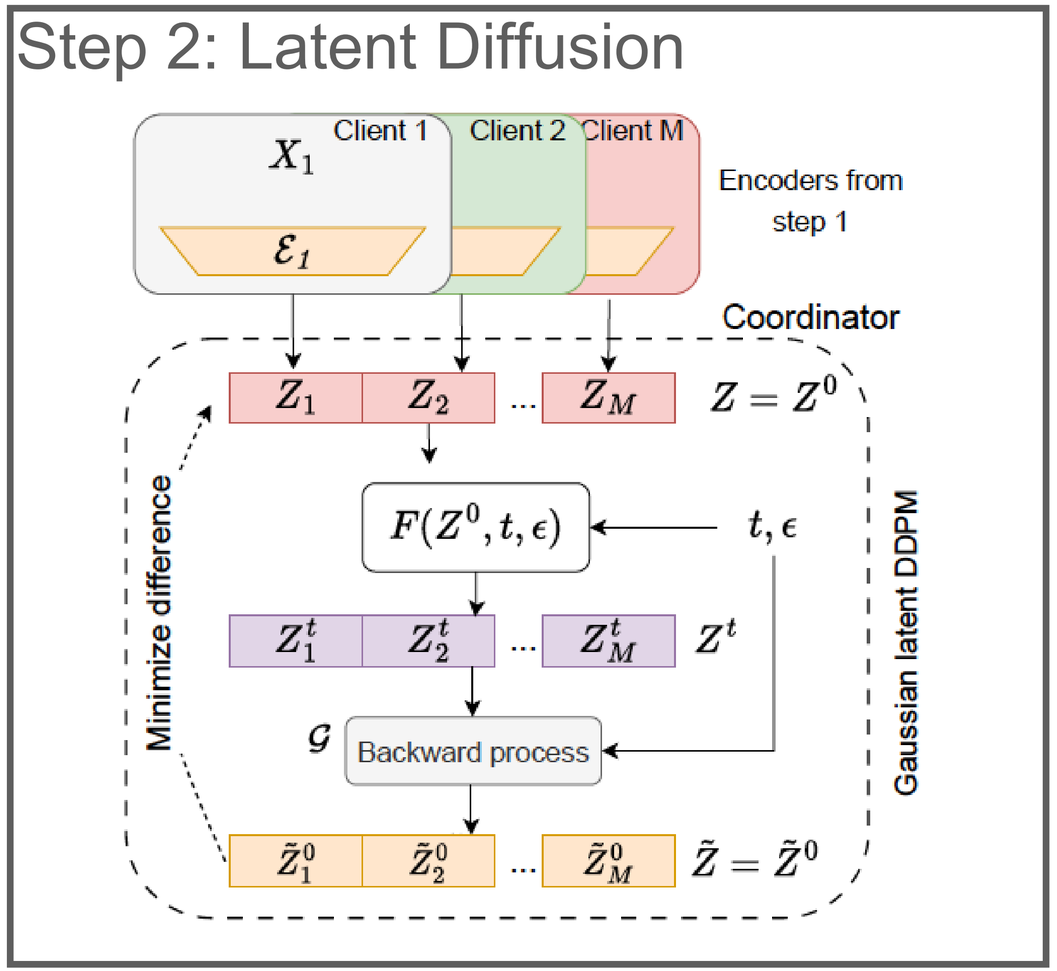




SiloFuse

Training

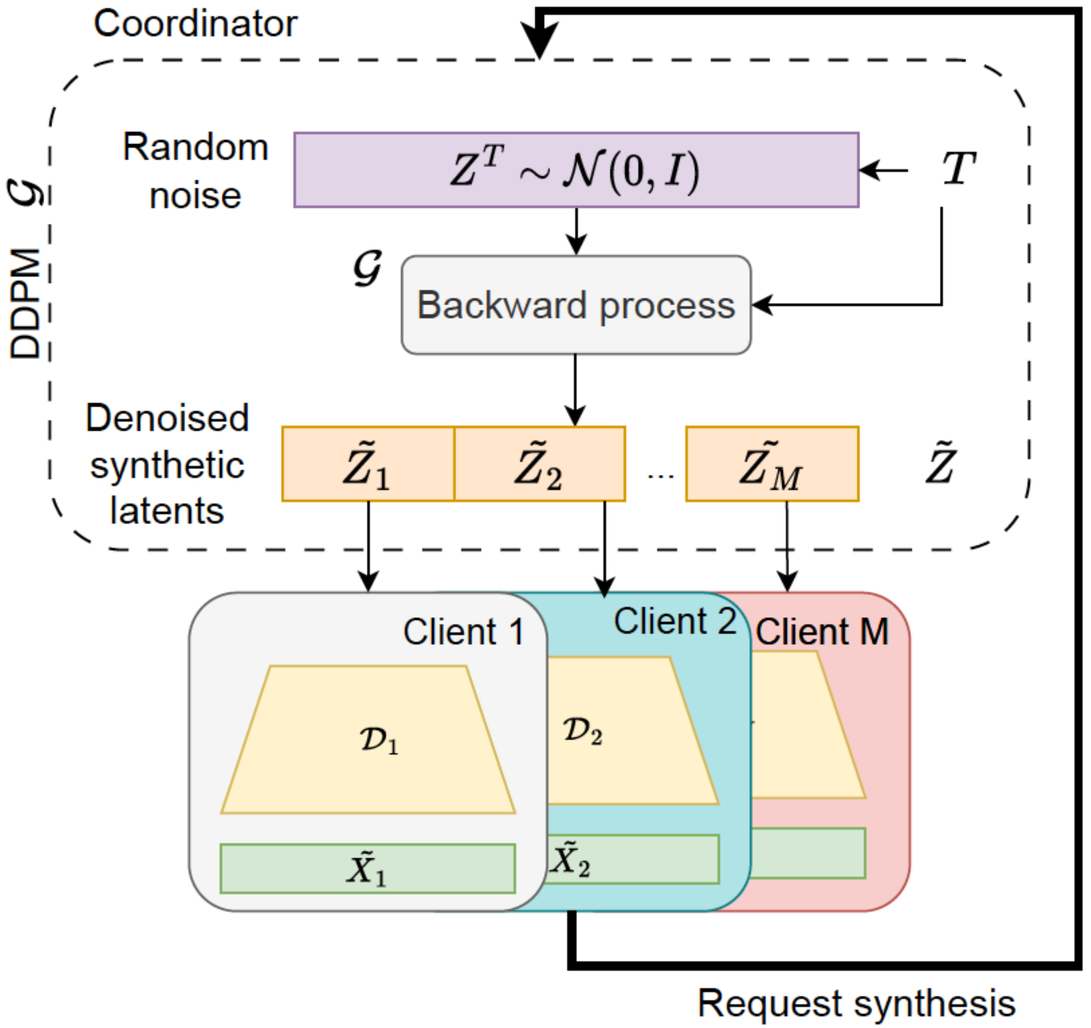




SiloFuse

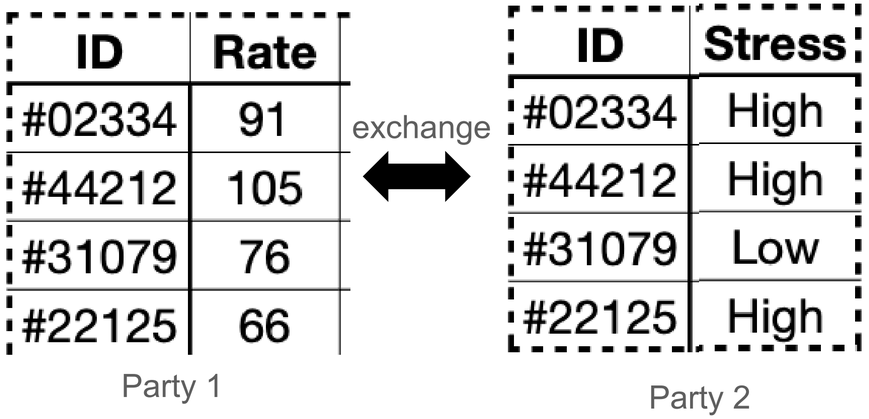
Synthesis

- **Two** scenarios
  - **Share** post synthesis (**weaker** privacy)
  - **Retain** vertical partitioning (**stronger** privacy)



Privacy

- **Semi-honest**
- Vertical partitioning post synthesis: **Theoretical irreversibility**
- Quantify **risks of sharing** post synthesis (3 attacks)
  - **Single out** rows
  - **Links** between rows
  - Infer **other columns** from local ones



Experiments

Resemblance

- Similarity between real and synthetic **distributions**
- Avg. of **5** metrics
  - Column **similarity** (correlation)
  - Probability distribution **distance**
  - **External discriminator** model
- **6** baselines. **9** datasets
- Centralized GANs, Stacked vs End-to-End, Centralized diffusion

Resemblance

- **Diffusion** outperform **GANs**
- **Latent** models **competitive** against **centralised diffusion**

Model	Abalone	Adult	Cardio	Churn	Cover	Diabetes	Heloc	Intrusion	Loan
GAN(conv)	64.0 ± 0.00	38.0 ± 0.00	59.8 ± 0.40	43.4 ± 0.80	45.2 ± 0.40	75.8 ± 0.40	54.0 ± 0.00	47.2 ± 0.40	76.4 ± 0.49
GAN(linear)	54.2 ± 0.40	28.6 ± 0.49	29.0 ± 0.00	30.8 ± 0.39	36.0 ± 0.00	51.0 ± 0.00	48.0 ± 0.00	39.0 ± 0.00	40.0 ± 0.00
E2E	85.2±0.40	60.0±0.00	60.2±0.40	88.2±0.40	51.0±0.40	72.4±0.80	68.4±0.49	48.0 ±0.00	81.2±0.40
E2EDistr	56.4±0.80	46.0±1.09	44.0±0.89	78.0±0.00	40.8±0.40	61.8± 3.12	61.0±0.00	37.0±0.00	49.8±1.16
TabDDPM	91.2±0.75	**97.0±0.00**	**98.0±0.00**	63.6±0.49	78.0±0.00	**94.6±0.49**	**88.0±0.00**	44.0±0.00	**98.0±0.00**
LatentDiff	**92.0±0.00**	78.0±0.00	72.2±0.40	**89.0±0.00**	**92.0±0.00**	90.0±0.63	83.4±0.49	**68.0±0.00**	83.4±0.49
SiloFuse	91.0±0.00	73.0±0.00	71.0±0.00	87.0±0.00	89.0 ± 0.00	84.0±0.63	79.0±0.00	67.0±0.00	81.2±0.40
PPD (vs GAN)	27.0	35.0	11.2	43.6	43.8	8.2	25.0	19.8	4.8

Downstream Utility

- Ratio of **classifier accuracy** - synthetic : real (capped at 100.0)
- **Diffusion** outperform GANs
- Comparable to centralized auto-encoder and diffusion

Model	Abalone	Adult	Cardio	Churn	Cover	Diabetes	Heloc	Intrusion	Loan
GAN(conv)	71.0 ± 0.63	82.6 ± 11.30	94.0 ± 3.63	85.0 ± 1.09	82.6 ± 1.01	96.2 ± 2.56	45.0 ± 0.63	36.2 ± 0.98	82.6 ± 0.49
GAN(linear)	65.2 ± 0.40	30.6 ± 1.62	47.6 ± 0.49	78.4 ± 1.02	36.2 ± 0.40	84.6 ± 1.62	38.4 ± 0.49	25.4 ± 0.49	82.0 ± 0.63
E2E	70.0±1.41	45.0±1.67	75.2±1.47	87.8±0.74	89.8±0.40	84.2±5.11	39.8±0.40	31.0±0.63	77.0±0.89
E2EDistr	70.8±1.72	33.6±1.02	55.3±0.33	87.0±0.89	56.8±1.46	89.0±2.28	39.2±0.40	24.2±1.16	71.2±2.92
TabDDPM	98.4±0.49	89.2±1.60	**99.6±0.49**	44.4±6.28	80.4±7.94	**100.0±0.00**	**96.4±0.49**	23.0±4.98	**98.8±1.17**
LatentDiff	**100.0±0.00**	**100.0±0.00**	86.4±2.33	**100.0±0.00**	95.8±0.40	99.6±0.49	76.4±0.49	61.2±1.16	94.8±1.93
SiloFuse	97.2±1.16	96.6±5.04	93.2±4.95	90.4±0.49	**96.4±0.49**	95.2±3.92	74.8±0.74	**64.2±0.74**	90.0±0.89
PPD (vs GAN)	26.2	14.0	-0.8	5.4	13.8	-1.0	29.8	28.0	7.4

Privacy

- Tradeoff

Model	Abalone	Adult	Cardio	Churn	Cover	Diabetes	Heloc	Intrusion	Loan
TabDDPM	48.2±0.48	70.1±2.61	76.1±2.19	86.7±0.24	59.1±3.85	46.2±0.69	**56.7±1.60**	**92.1±2.62**	57.9±1.87
LatentDiff	50.3 ± 0.58	73.7 ± 2.29	88.7 ± 2.82	78.1 ± 4.18	55.7 ± 1.81	62.4 ± 1.32	51.9 ± 1.46	70.5 ± 2.89	64.8 ± 2.36
SiloFuse	**55.9±0.46**	**92.1±0.60**	**93.2±4.97**	**92.3±1.84**	**65.1±2.25**	**78.1±2.40**	56.4±1.58	70.5±2.46	**79.3±5.35**

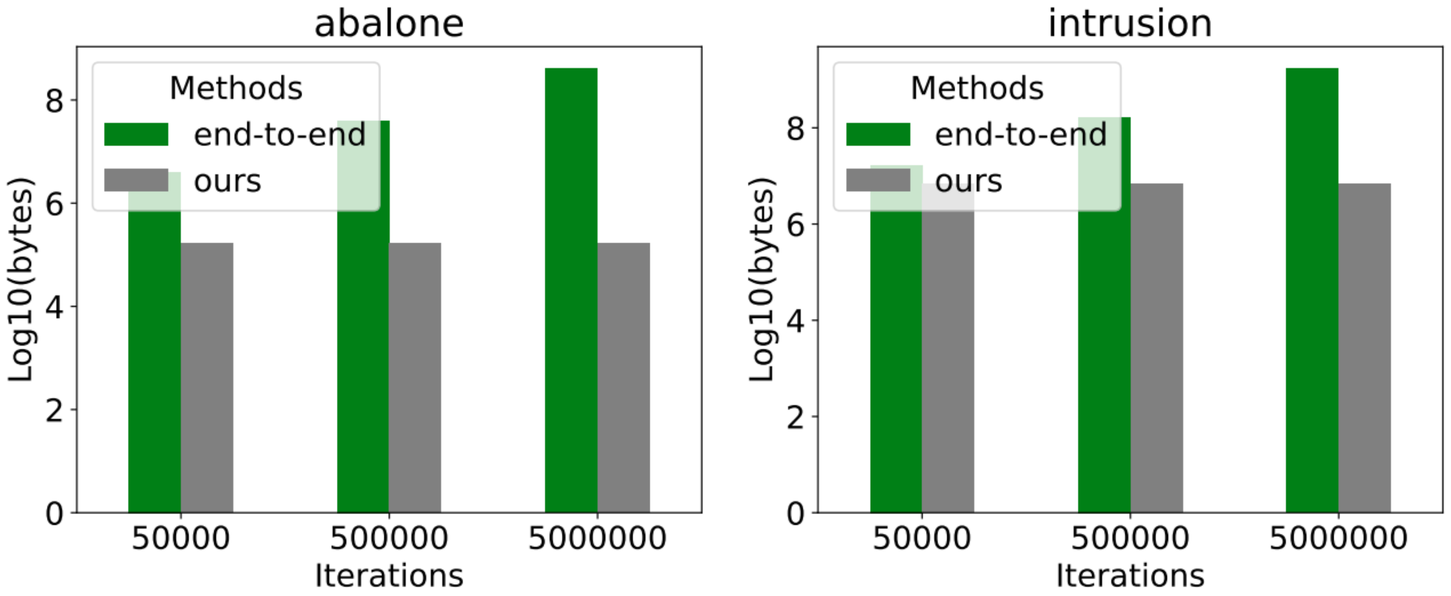
- Recommendations:
  - **Strong privacy** —> Retain partitioning. **FL** on downstream
  - **Weaker** guarantees —> Share post synthesis. **Local** model downstream

Communication

Stacked vs End-to-end training

- Bytes transferred during training
- End-to-end training: Increasing iterations —> Increasing comm.
- Stacked training: **Single** comm. round

Fixed cost should be the takeaway point



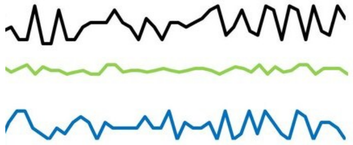
Future work and Conclusions

Conclusions

- Synthesize, yet maintain privacy
- Latent architecture
  - Links in real space **preserved** in latent space
  - **Efficient** training paradigm
- Privacy
  - **Spectrum**
  - Tradeoffs involved

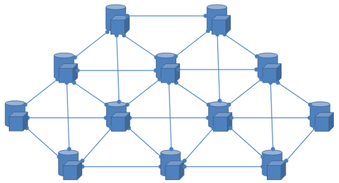
Future work

- **Beyond tabular** data

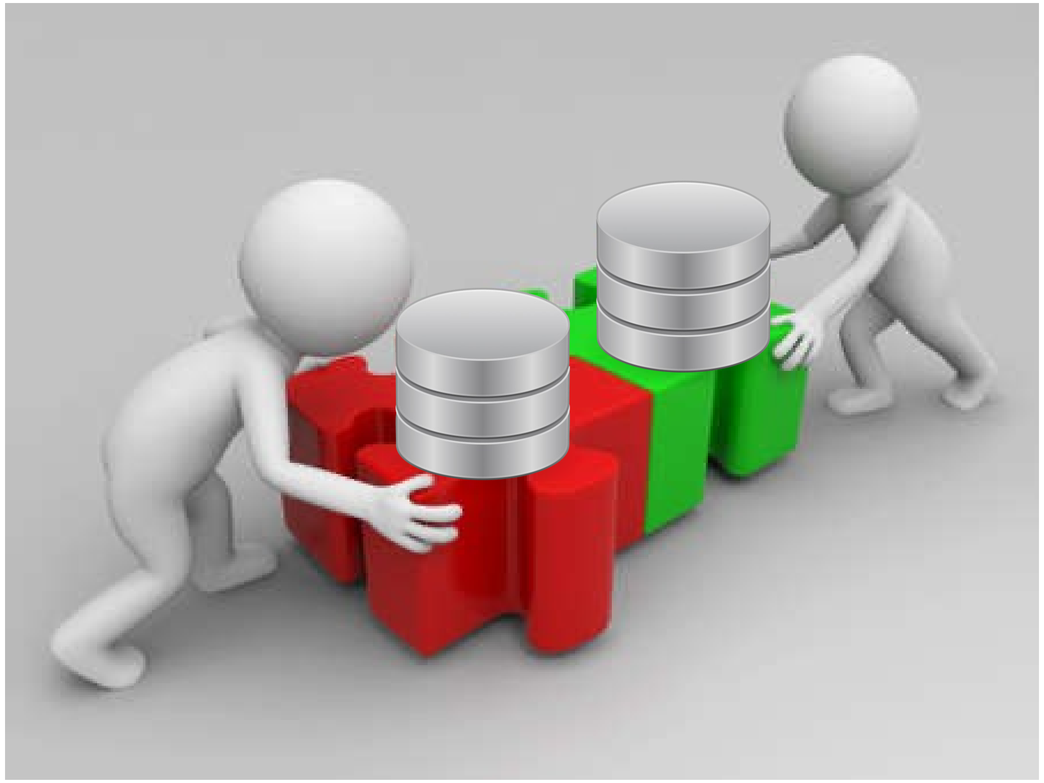


- **Malicious** adversaries

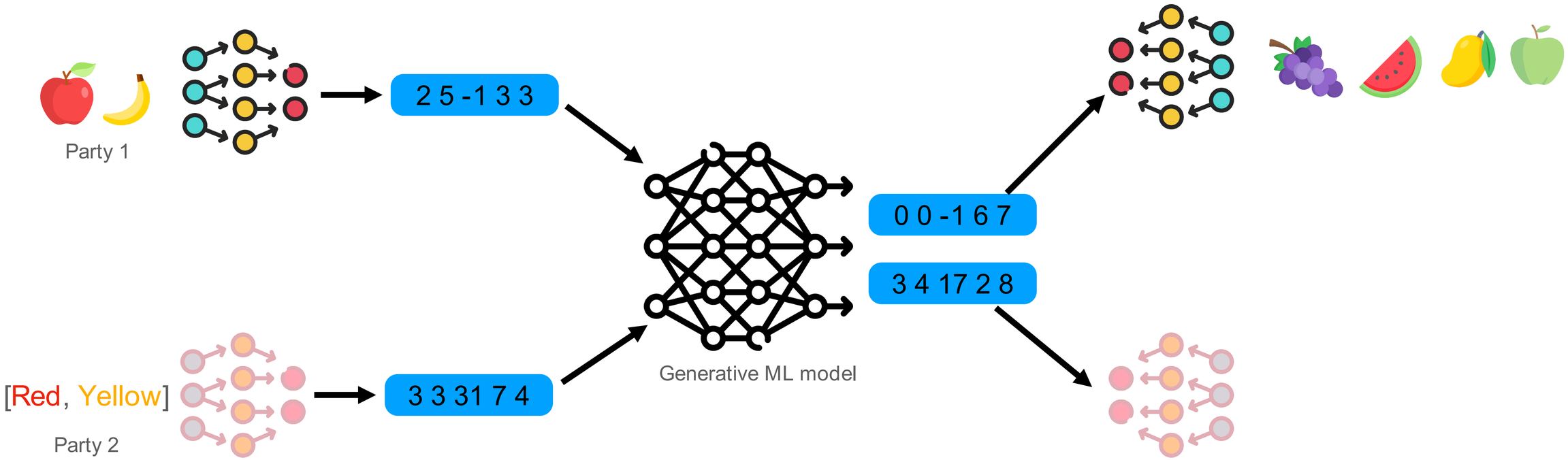
- **Distributed servers**



Thank you



Learning from multiple sources



- **New** data **without** knowing private info