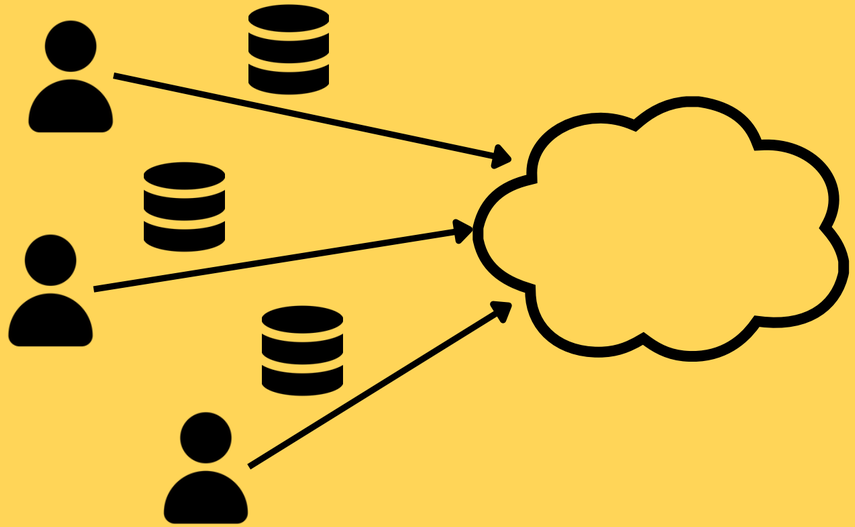


Building collaborative *data-ecosystems*

Sai Praneeth Karimireddy

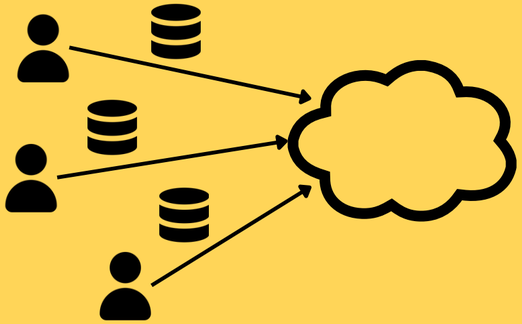




- Firms gather vast pools of user data
  - Tesla and Uber are tech companies due to their exabytes of driving data.
  - *"We are a data business"* - Dominos '20.
- This data embodies human labor, but little to no value is given back.

Data centralization is failing us

Incentives



The Data That Powers A.I. Is Disappearing Fast

July 19, 2024

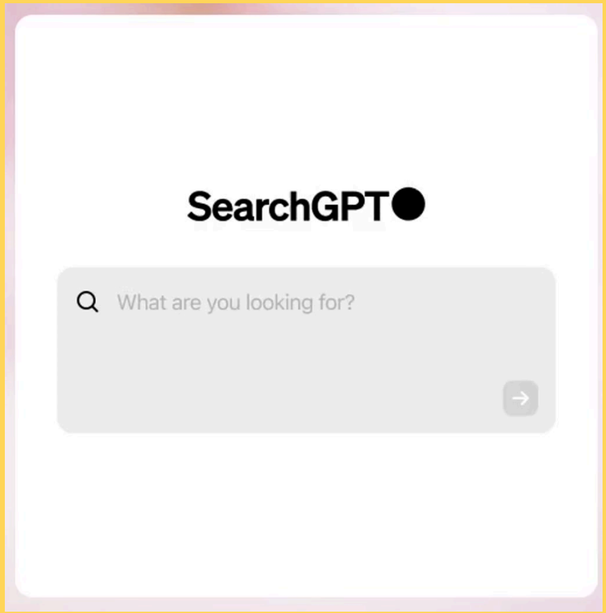
Source: New York Times
Author: Kevin Roose
Coverage Type: Reporting

AI search could break the web

Developers should act before governments fall back on blunt tools.

By Benjamin Brooks — October 31, 2024

Reddit has a new AI training deal to sell user content / Over a decade of valuable user content is now for sale as Reddit preps to go public.

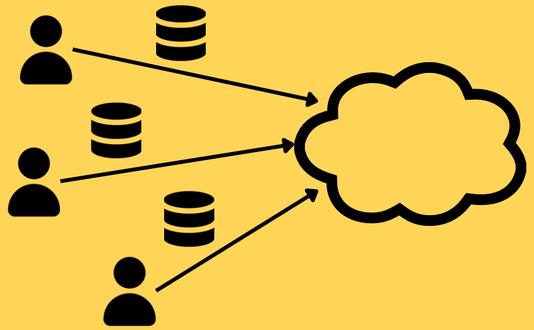


Twitter's $42,000-per-Month API Prices Out Nearly Everyone

Tiers will start at $500,000 a year for access to 0.3 percent of the company's tweets. Researchers say that's too much for too little data.

Data centralization is failing us

Privacy & Security



A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too

TECHNOLOGY

07/04/2024

Here's How to Freeze Your Credit After the Change Healthcare Data Breach

Freezing your credit won't protect you from all scams. But it does keep cybercriminals from opening accounts in your name.

Danni Santana

Oct. 30, 2024 7:00 a.m. PT

6 min read

Hacker behind Snowflake customer data breaches remains active

The hacker has extorted $2.7 million as part of the attacks on Snowflake customers, according to a researcher tracking the case.

BY AJ VICENS • SEPTEMBER 21, 2024

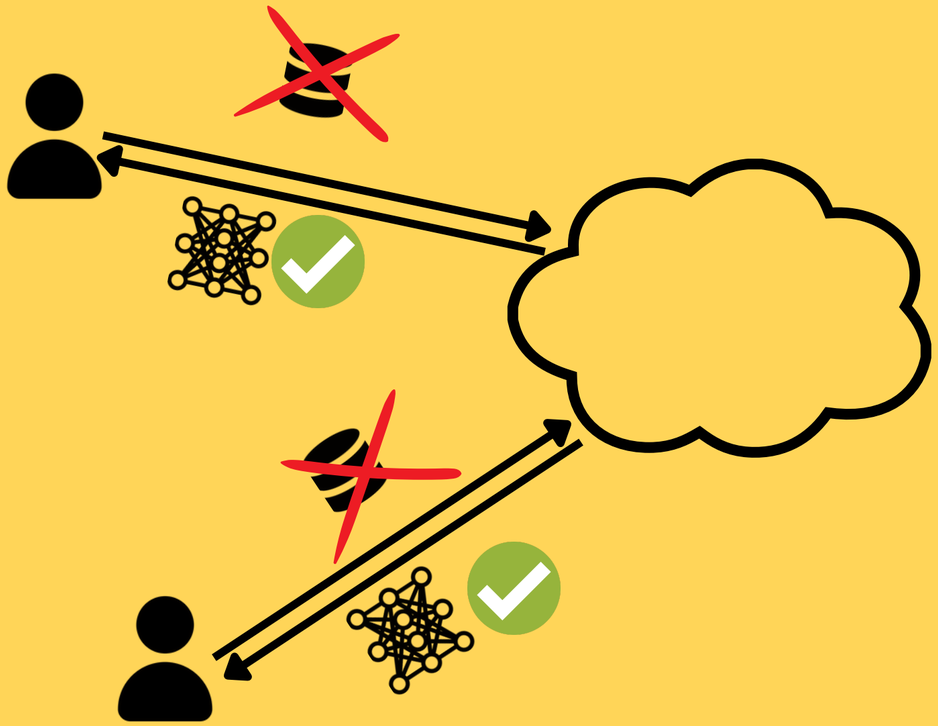
Amazon confirms employee data breach after vendor hack

By Sergiu Gatlan

November 11, 2024 02:10 PM

Collaborative Learning

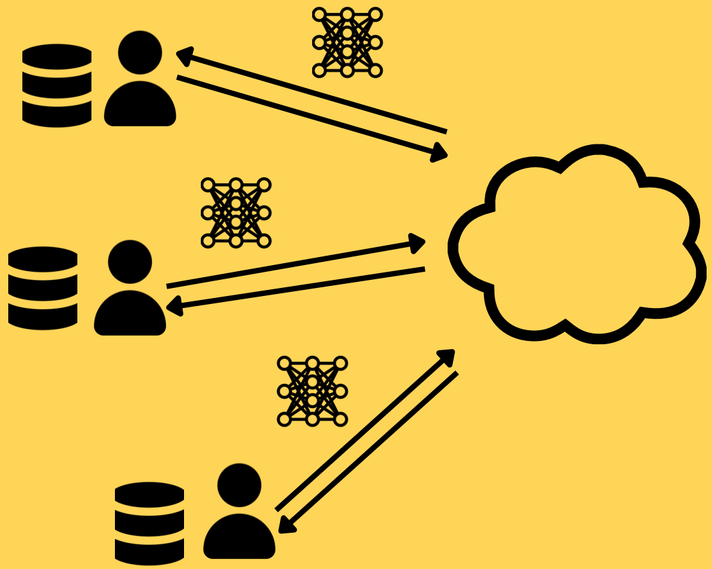
A user-centric ML paradigm



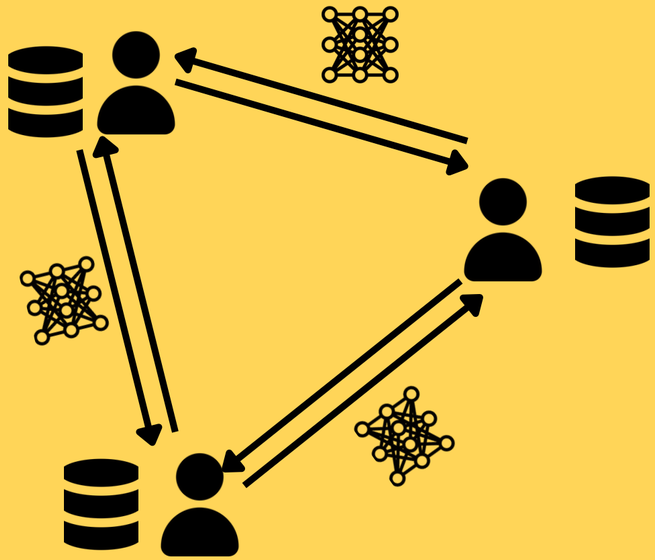
- Keep data distributed
- C*ollaboratively* train ML models
- Dont’ send data, only aggregate information (models, ...)

Collaborative Learning

A user-centric ML paradigm



Federated Learning
(central coordinator)

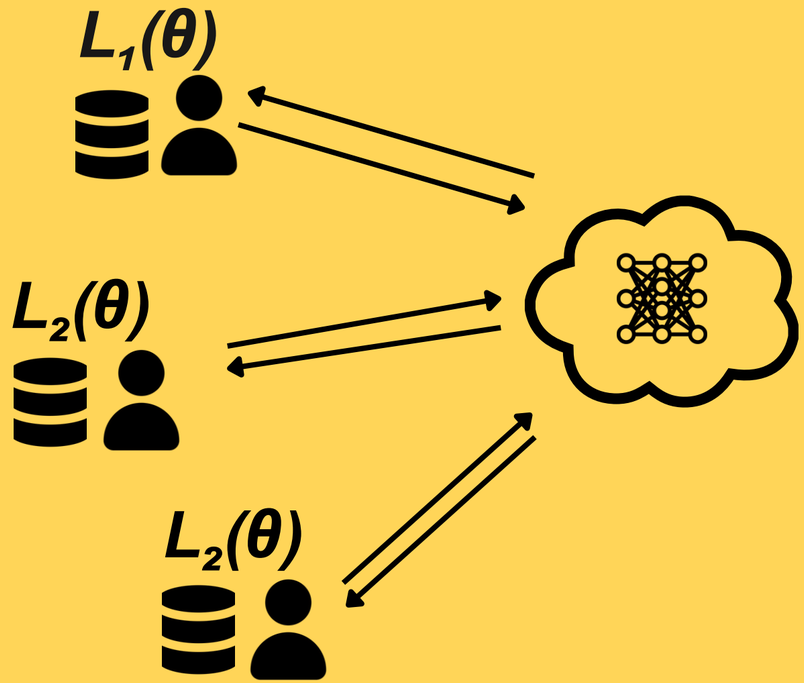


Decentralized Learning
(peer to peer)

Scalable Collaborative Learning

How?

1. Define user's loss using local data distribution $\mathcal{L}_i(\theta_i)$

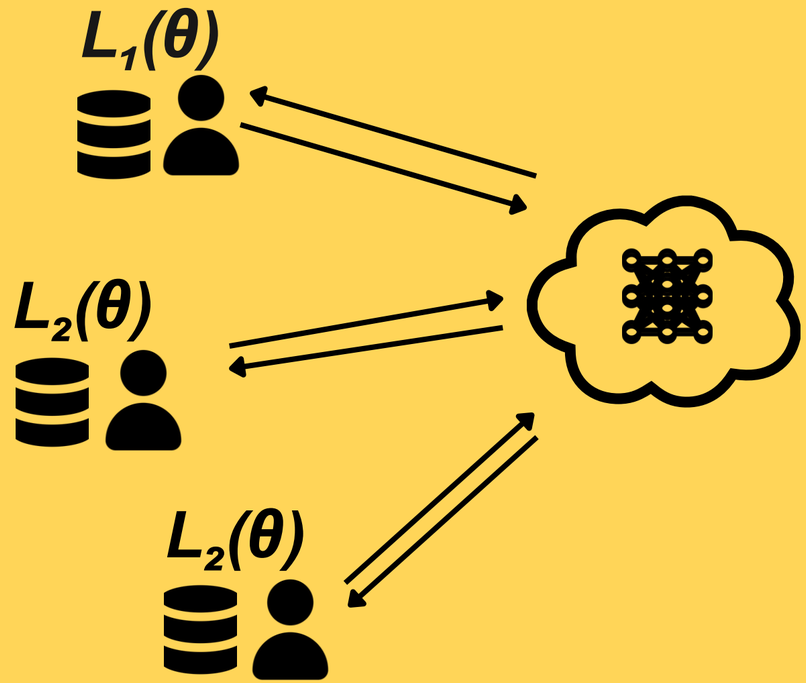


Federated Learning
[McMahan et al. 2016]

Scalable Collaborative Learning

How?

1. Define user's loss using local data distribution $\mathcal{L}_i(\theta_i)$
2. Iteratively on every user,
   - Copy global θ

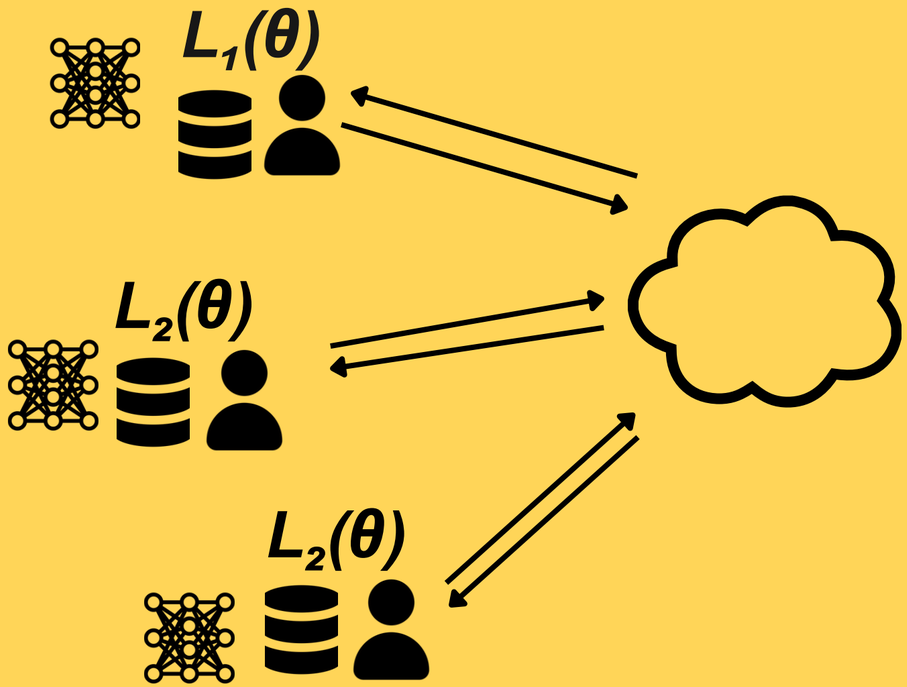


Federated Learning
[McMahan et al. 2016]

Scalable Collaborative Learning

How?

1. Define user's loss using local data distribution $\mathcal{L}_i(\theta_i)$
2. Iteratively on every user,
   - Copy global θ
   - Update θ_i to minimize local loss $\mathcal{L}_i(\theta_i)$



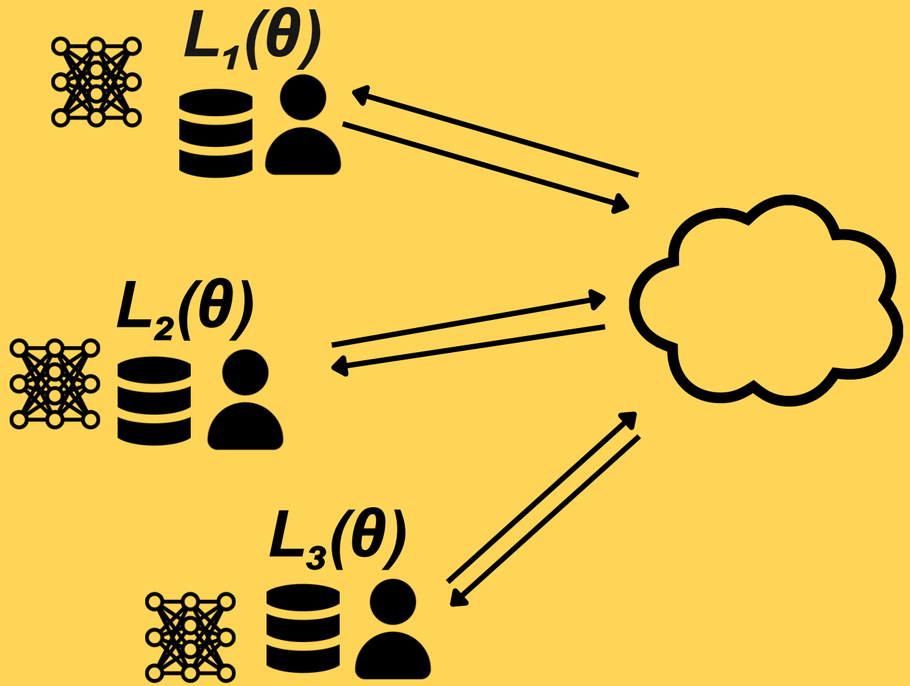
Federated Learning
[McMahan et al. 2016]

Scalable Collaborative Learning

How?

1. Define user's loss using local data distribution $\mathcal{L}_i(\theta_i)$
2. Iteratively on every user,
   - Copy global θ
   - Update θ_i to minimize local loss $\mathcal{L}_i(\theta_i)$
   - Synchronize parameters θ_i

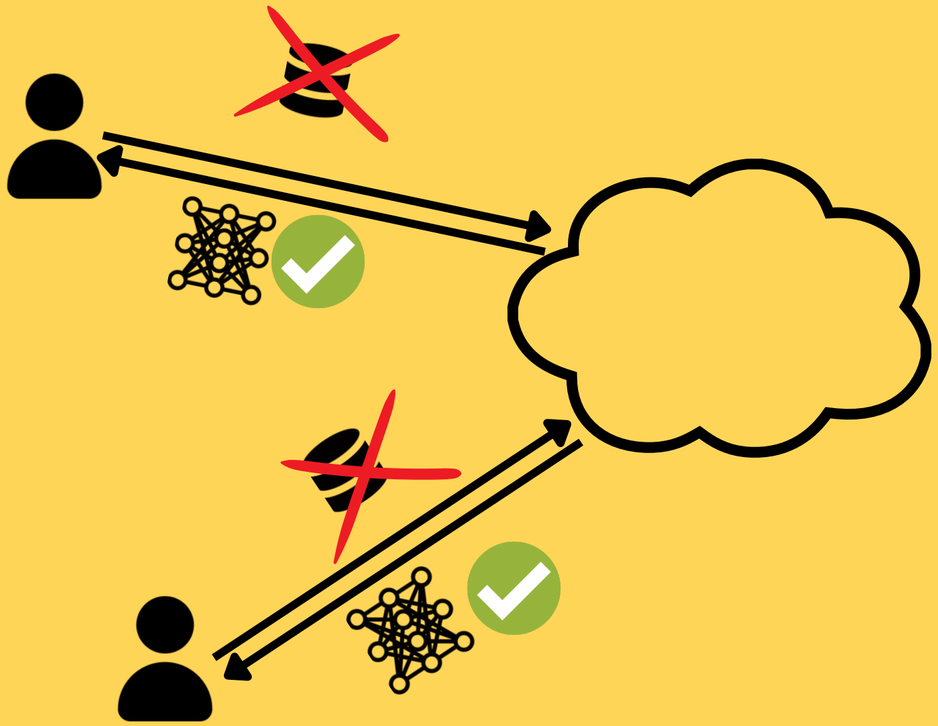
$$\theta = \frac{1}{m} \sum_i \theta_i$$



Formal security and privacy **important**, ask me later.

Collaborative Learning

Incentives



- What if we both have conflicting interests and goals?
- What is my data worth?
- Many important, challenging, and opn questions!

Agenda

01

Aligning Goals
Economics

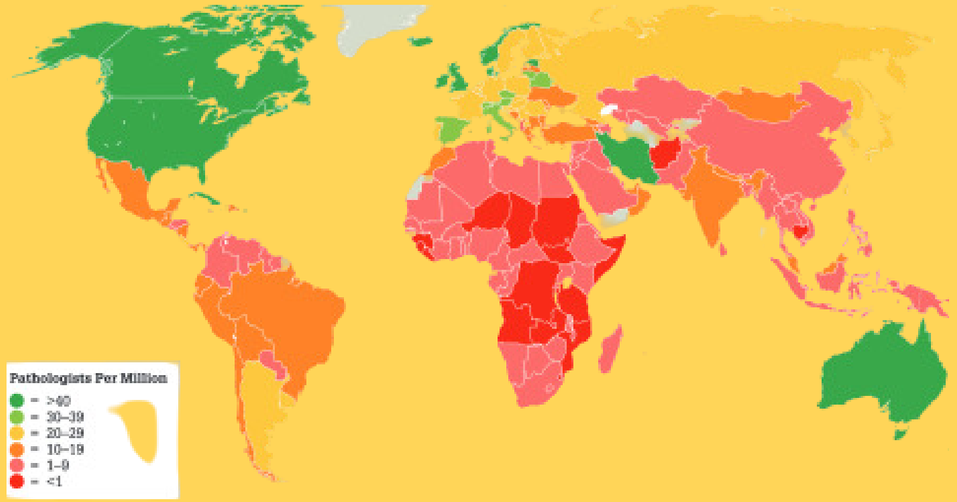
02

Data Markets
Opt+Econ

1

Aligning Diverse Goals

Shortage of Cancer expertise



> **"Cancer patients face life-threatening delays due to lack of staff, say UK radiologists."**
>
> — The Guardian, 8 Jun 2023

- Global shortage of expertise to diagnose and manage cancer even in high-income regions.
- Life-threatening consequences.
- Can AI help? Where to get enough diverse data?

Cancer Registry of Norway

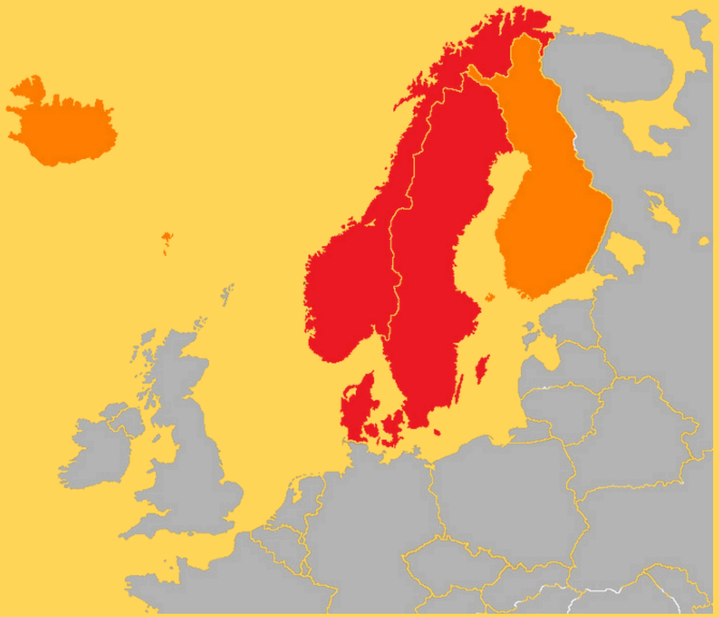
NORDCAN Collaboration

Passive users to active agents

7+ countries and 21 registries want to collaborate.

But,

- Strict data regulations. Use collaborative learning.
- **Beneficence:** risk-benefit analysis for each sub-population. So what loss should we minimize?



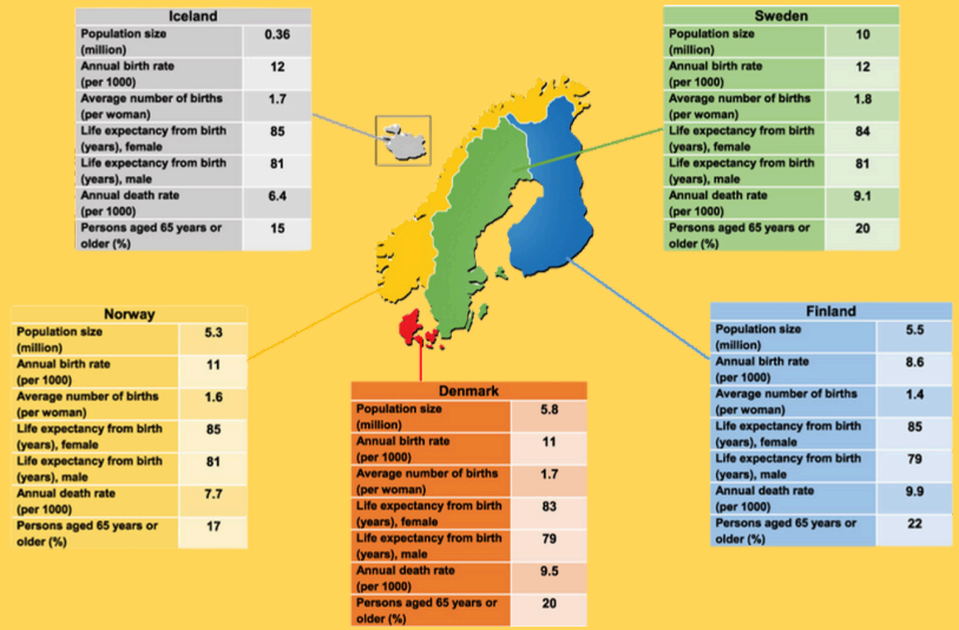
Cancer Registry of Norway

Position paper with Norway cancer registry

[*Karimireddy* et al. FMEC 2023]

Multi-Agent Formalism

- ***n*** countries (agents), each can ***freely choose***
  - ***how much*** data
  - ***which*** samples
- Data incurs real costs (privacy, side effects)
- Efficiency critical, but over which subpopulation?
  Fo each objective I pick, different countires might be angry.
- This is a multi-agent game.

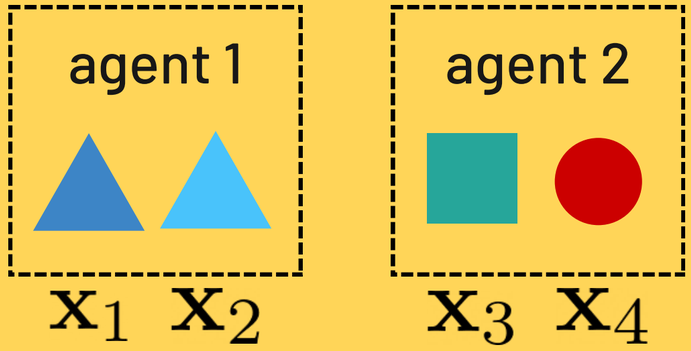


[*Karimireddy*, Guo, Jordan. **best paper** FL-NeurIPS 2022]

[Huang et al. 24. Evaluating and Incentivizing Diverse Data Contributions in Collaborative Learning]

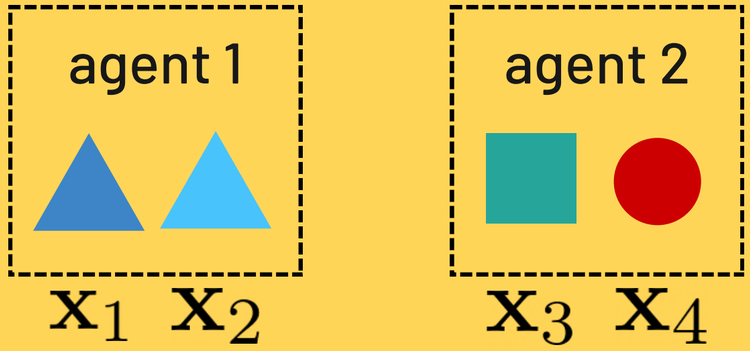
Experiment Design
Heterogeneous populations

- Data types
  - **type** encoded by $\mathbf{x}$ = (30 yrs, Male, Asian, ...)
  - Total population has many types:
    $$\mathcal{X} = (\mathbf{x}_j)_{j=1,\dots}$$
  - Partitioned between ***n*** agents. Agent i only has subtypes $\mathcal{X}_i \subseteq \mathcal{X}$.



Experiment Design
Sampling strategy

- Data types
  - Total population has many types:
    $$\mathcal{X} = (\mathbf{x}_j)_{j=1,\dots}$$
  - Partitioned between ***n*** agents. Agent i only has subtypes $\mathcal{X}_i \subseteq \mathcal{X}$.
- Find the “best” sampling strategy:
  - m_j is **number of samples** of type $\mathbf{x}_j$.
  - Agent i shares $\mathbf{m}_i = (m_j, \text{ for } \mathbf{x}_j \in \mathcal{X}_i)$



$$\mathbf{m} = (2.3,\ 0.7,\ 1.8,\ 0.1)$$
$$= (\mathbf{m}_1,\ \mathbf{m}_2)$$

Classic Experiment Design

Estimating error

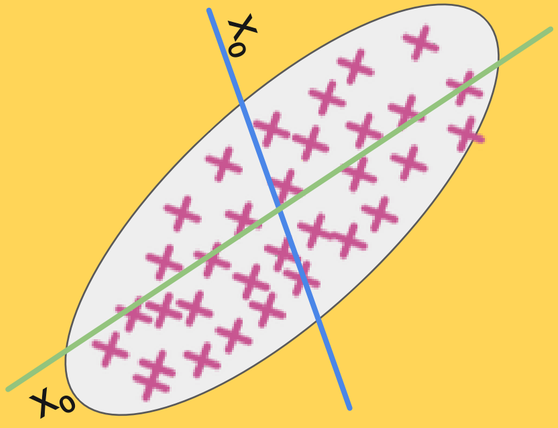
- Data space $\mathcal{X}$
- Total data sampled $\mathbf{m}$
- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

- Error on any test $\mathbf{x_0}$ determined by $\mathcal{I}(\mathbf{m})$

$$\mathcal{E}(\mathbf{x_0}, \mathbf{m}) = \mathbf{x_0}^\top (\sum_{j=1}^{k} m_j \mathbf{x_j} \mathbf{x_j}^\top)^\dagger \mathbf{x_0}$$
$$= \mathbf{x_0}^\top \mathcal{I}(\mathbf{m})^\dagger \mathbf{x_0}$$

- Information matrix $\mathcal{I}(\mathbf{m})$ captures error on all possible test points.

Error determined by information matrix



[Wald 1943; Pukelsheim 2006]

Classic Experiment Design

Choosing a metric

- Data space $\mathcal{X}$
- Total data sampled $\mathbf{m}$
- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

- Error on any test $\mathbf{x_0}$ determined by $\mathcal{I}(\mathbf{m})$

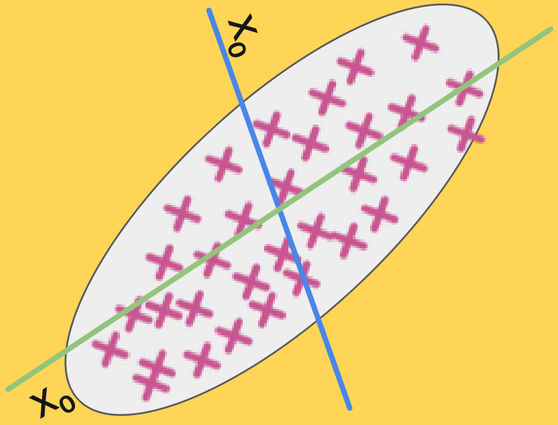
$$\mathcal{E}(\mathbf{x_0}, \mathbf{m}) = \mathbf{x_0}^\top (\sum_{j=1}^{k} m_j \mathbf{x_j} \mathbf{x_j}^\top)^\dagger \mathbf{x_0}$$
$$= \mathbf{x_0}^\top \mathcal{I}(\mathbf{m})^\dagger \mathbf{x_0}$$

- Metric compresses into a scalar error

$$\mathcal{E}_f(\mathcal{X}, \mathbf{m}) = \mathbf{f}\left(\mathcal{I}(\mathbf{m})^\dagger\right)$$

- $\mathbf{f}$ can be
  - Tr (A criterion)
  - log det (D criterion)
  - λmin (E criterion)
  - ...

Error determined by information matrix



[Wald 1943; Pukelsheim 2006]

Multi-Agent Experiment Design

Estimating agent's error

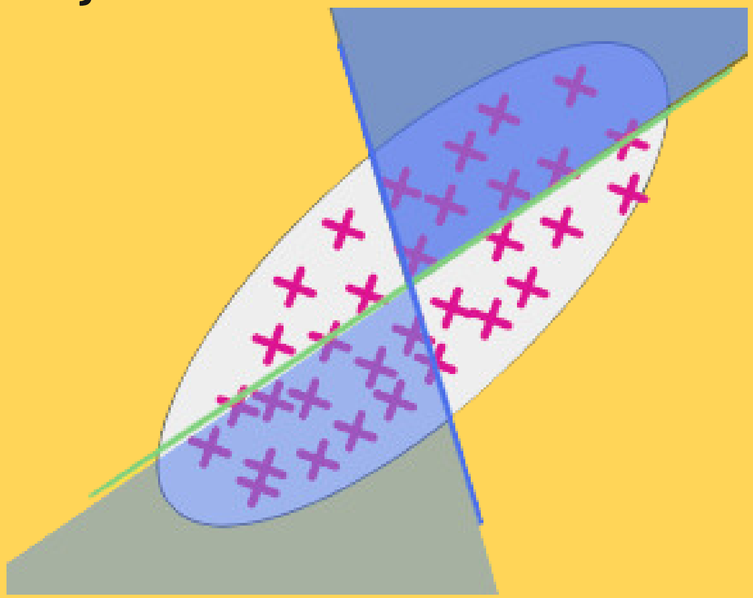
- Data space $\mathcal{X}$
- Total data sampled $\mathbf{m}$

- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

- Agent i only cares about errors on $\mathcal{X}_\mathbf{i}$

$$\mathcal{E}_f(\mathcal{X}_i, (\mathbf{m}_i, \mathbf{m}_{\neg i})) = \mathbf{f}\left(\Pi_i \mathcal{I}(\mathbf{m})^\dagger \Pi_i\right)$$

Projected information matrix



Multi-Agent Experiment Design

Efficiency games

- Data space $\mathcal{X}$:
- Total data sampled $\mathbf{m}$
- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

Global: $$\min_{\mathbf{m}} \underbrace{\mathcal{E}_f(\mathcal{X}, \mathbf{m})}_{\text{net error}} + \underbrace{c\|\mathbf{m}\|_1}_{\text{net cost}}$$

Agents: $$\min_{\mathbf{m}_i} \underbrace{\mathcal{E}_f(\mathcal{X}_i, (\mathbf{m}_i, \mathbf{m}_{\neg i}))}_{\text{agent error}} + \underbrace{c\|\mathbf{m}_i\|_1}_{\text{agent cost}}$$

Global and local objectives misaligned => inefficiency => ***Price of Anarchy***



Incentive-compatible efficiency

Which objective function f is incentive-compatible?

- *Not minimizing avg error:* Global efficiency prioritizes majority agents 1, 2 who have same type vs. minority agent 3 who has diff type.



Incentive-compatible efficiency

Which objective function f is incentive-compatible?

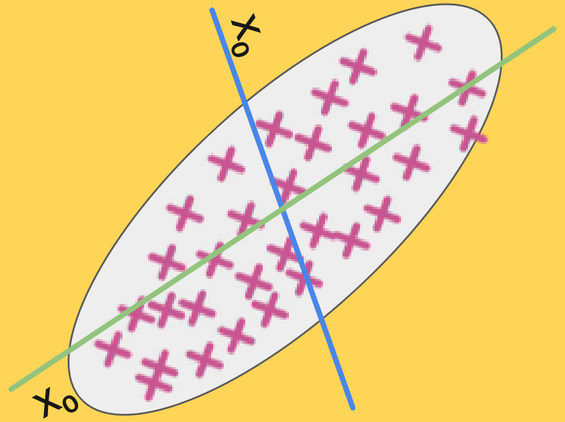
- *Not minimizing avg error*
- *Not minimizing worst-case error:*
  - Agent 2 has a harder task.
  - Global efficiency means 95% acc on all.
  - But agent 2 has 2x data costs.
  - Likely prefers lower acc for lower data costs.



Main Result on Efficiency

- For **f = log det** (D-criterion),
  - unique equilibrium
  - always Price of Anarchy = 1!
  - invariant to population splits.
- Not true for others: avg (A, V), nor min-max (G, E).

$$\log\det(\Pi\,\mathcal{I}(m)) = \log\det(\mathcal{I}(m)) + c$$



Upshot: Incentive-compatible only if everyone uses D-criterion. So, use it.

[Huang et al. 24. Evaluating and Incentivizing Diverse Data Contributions in Collaborative Learning]

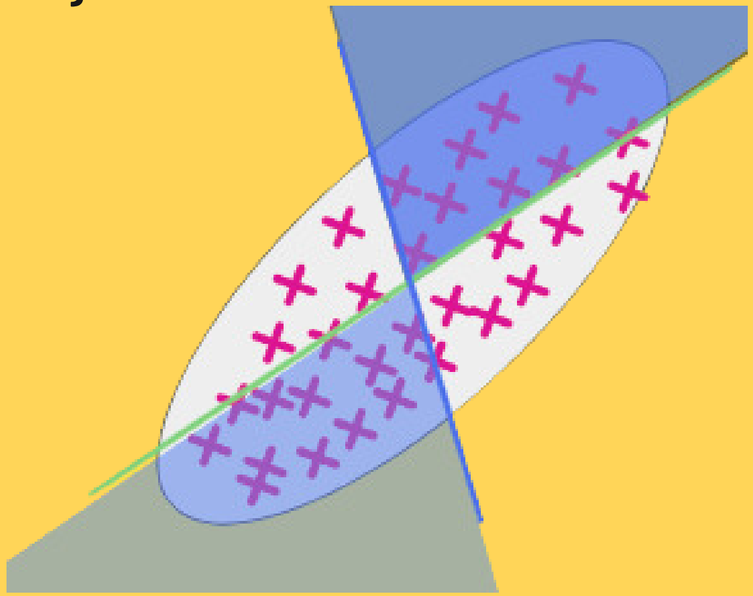
Fairness and Free-Riding

- Data space $\mathcal{X} = (\mathbf{x_1}, \ldots, \mathbf{x_k})$
- Total data $\mathbf{m} = (m_1, \ldots, m_k)$
- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

Agents: $\min_{\mathbf{m}_i} \mathcal{E}_f(\mathcal{X}_i, (\mathbf{m}_i, \mathbf{m}_{\neg i})) + c\|\mathbf{m}_i\|_1$

- Agents may end up with very different utilities
- Others' data is free => unequal net costs => **free-riding**.

Projected information matrix



Efficiency-Fairness Tradeoff

- Efficient equilibrium may result in unequal benefits.
- Agent 3 can simply use data from 1 & 2 => efficient but unfair.
- Our Results
  - Mechanisms to ensure fairness i.e. **equal benefit for all.**
  - Tradeoff fairness with efficiency.



Results on Fairness-Efficiency

1. Fairness is incompatible with efficiency.

2. A simple penalization scheme guarantees
   a) optimal data contributions, and
   b) equal net utility for all.

3. Can trade-off efficiency/fairness.

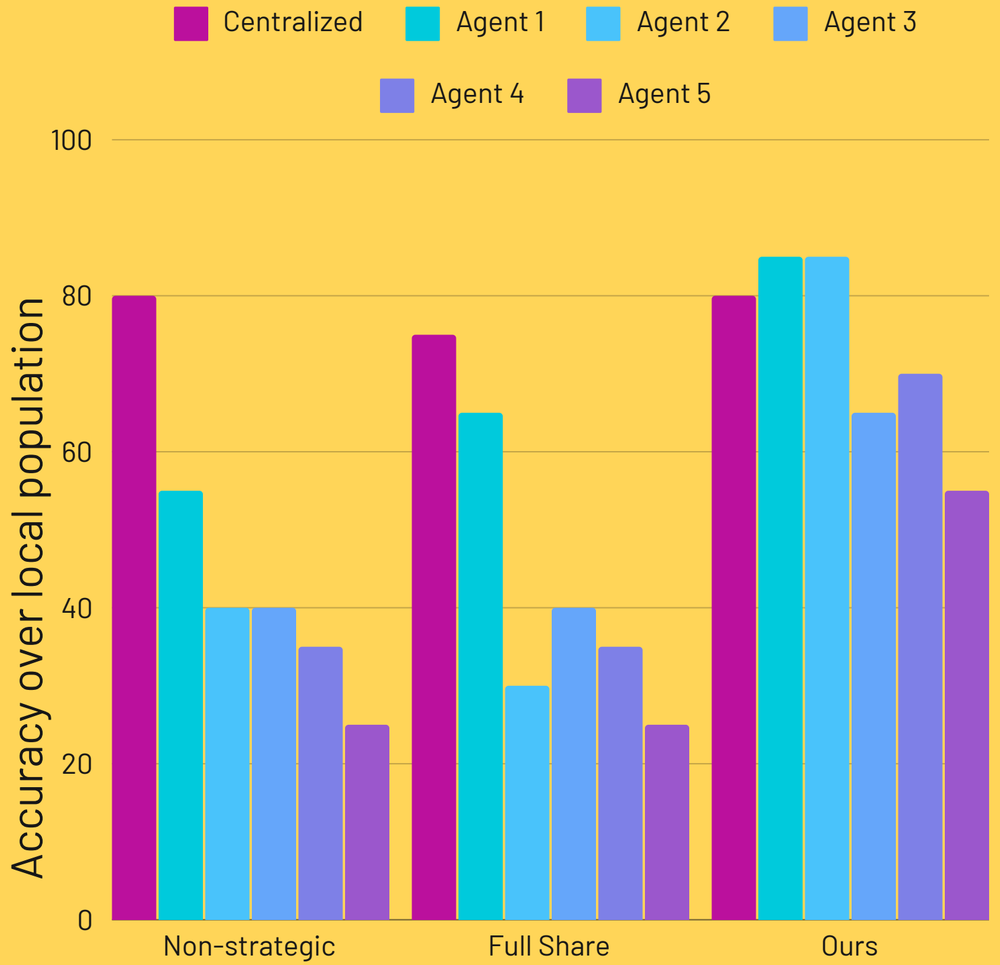


[Huang et al. 24. Evaluating and Incentivizing Diverse Data Contributions in Collaborative Learning]

Principal-Agent Experiment Design

Our Results

- Finetuning ResNet-18 on MedMNIST
- Types = classes. Split heterogeneously using Dirichlet sampling.
- All work if centralized, o.w.
  - Non-strategic: inefficient
  - Full sharing: unequal
  - Ours: efficient and “fair”



Takeaways

- Taking human agency seriously requires combining **econ+ML**.
- Gives concrete answers to thorny questions:
  - What definition of fair performance? Use ***log det***, not min-max or average.
  - How to design systems which align diverging interests?

2

Open data marketplace

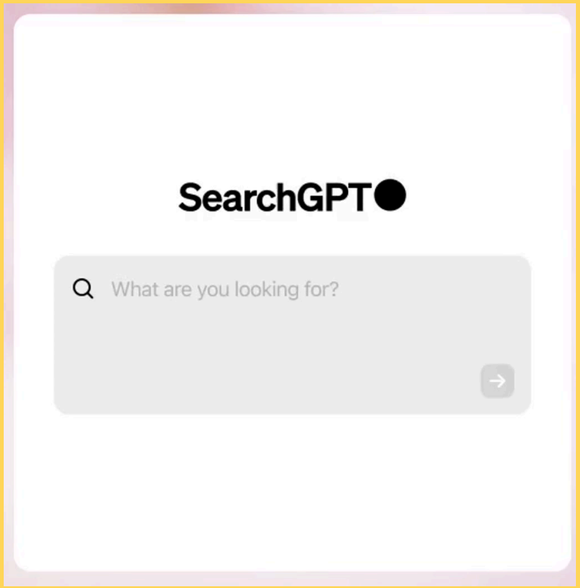


How to compensate for user data?

AI search could break the web

Developers should act before governments fall back on blunt tools.

By Benjamin Brooks October 31, 2024

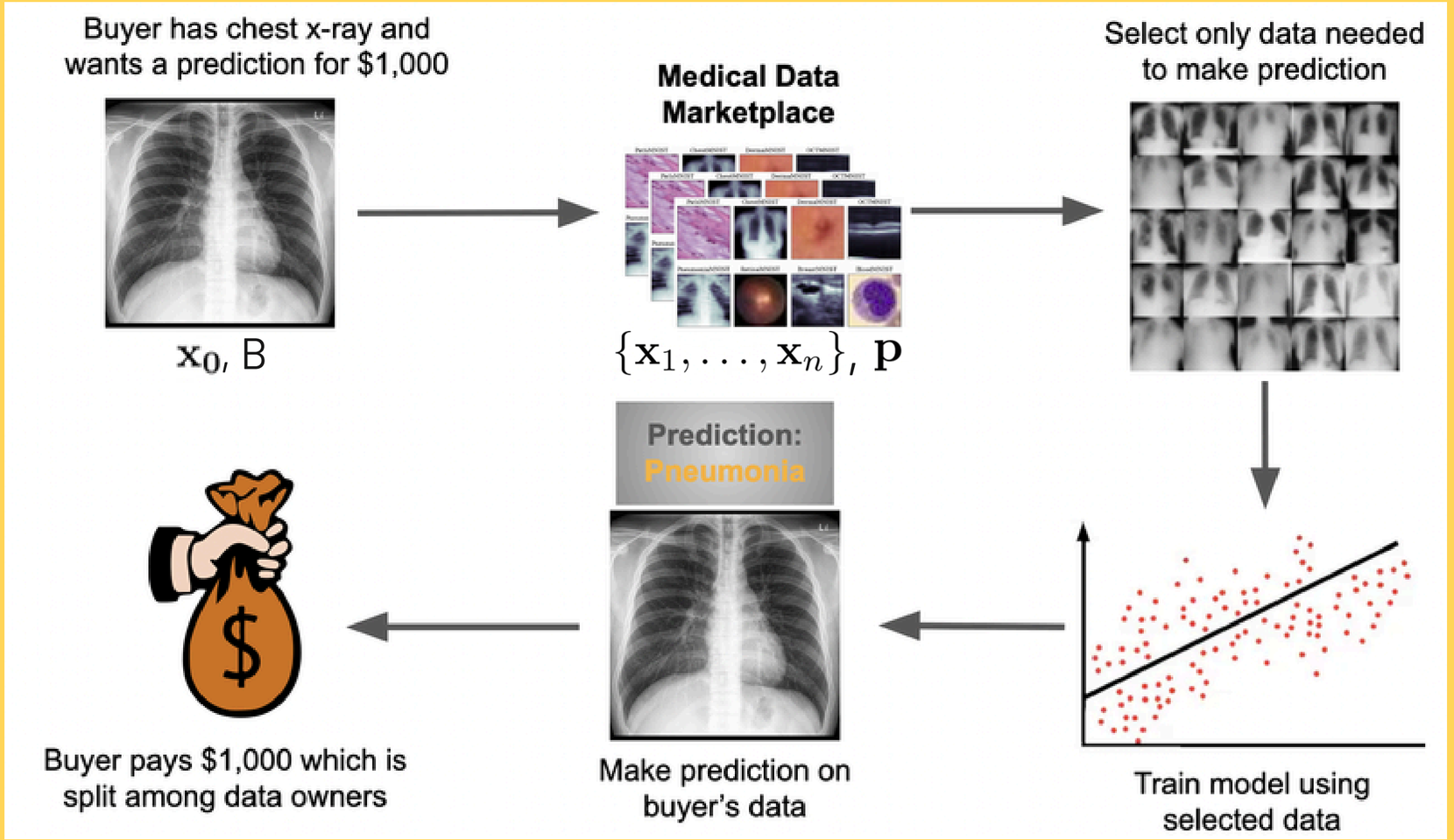


Staff Reporters | Aug 22, 2023

YouTube unveils strategy to compensate artists for AI-generated music

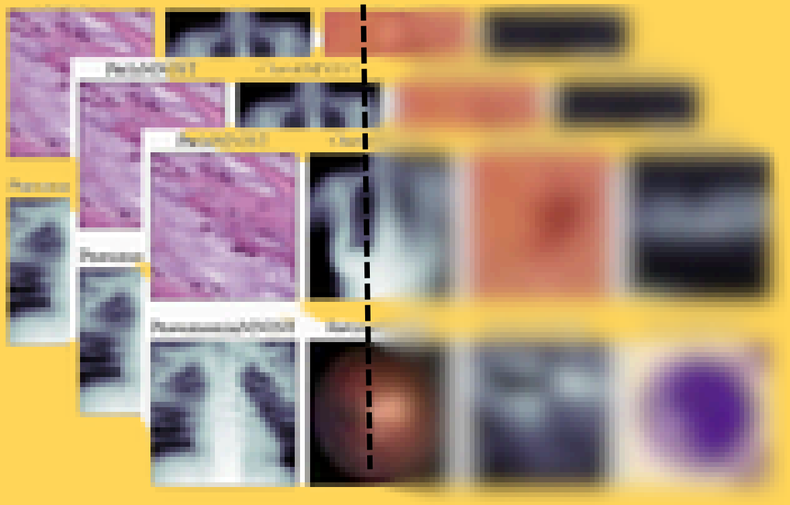


Open data marketplace



- Buyer test data $\mathbf{x}_0$ and budget B
- n sellers' data points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with prices $\mathbf{p}$
- Select best data within budget

Open or private?



- But, seller can’t reveal data
  - privacy concerns
  - IP concerns: data can be easily copied
- How can a buyer tell which data is useful for them?
- **Collaborative** data markets.

Collaborative data marketplaces

1

Model value of each seller datapoint

?

2

Collaborative data selection within budget

?

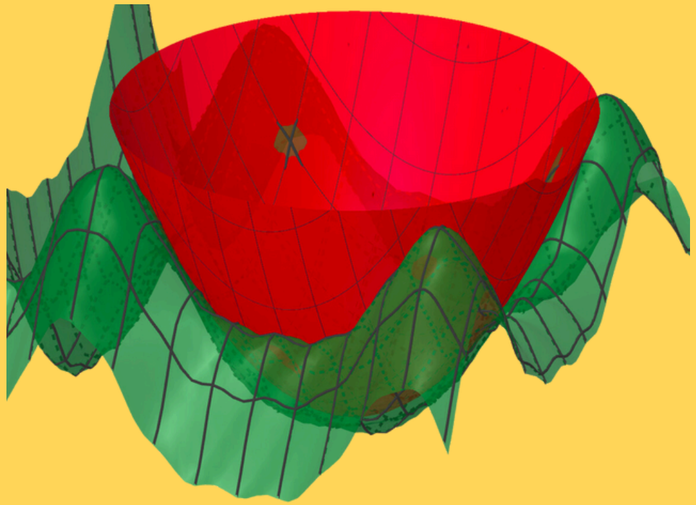
3

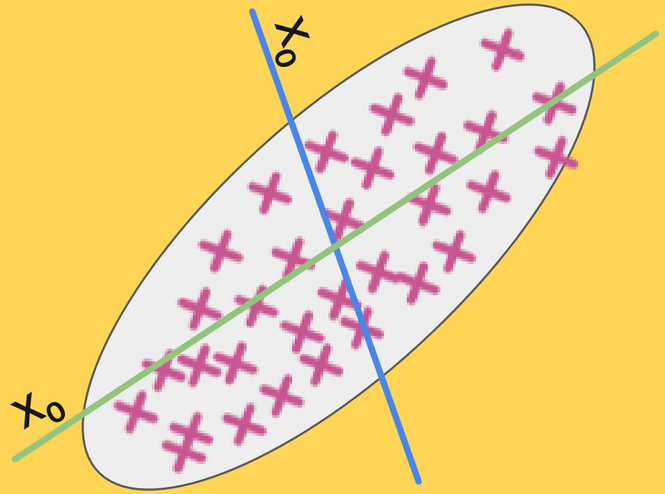
Train a model using CL on selected data

?

Collaborative data discovery

Modeling data value





a. Construct a linear proxy-model using

- Neural Tangent Kernel (NTK)
- Embeddings

b. Use information matrix to estimate error to buyer on $\mathbf{x_0}$

$$\min_{\substack{\mathbf{p}^\top \mathbf{w} \leq B \\ w_j \in \{0,1\}}} \mathbf{x_0}^\top \left(\sum_j w_j \mathbf{x_j} \mathbf{x_j}^\top\right)^\dagger \mathbf{x_0}$$

Collaborative data discovery

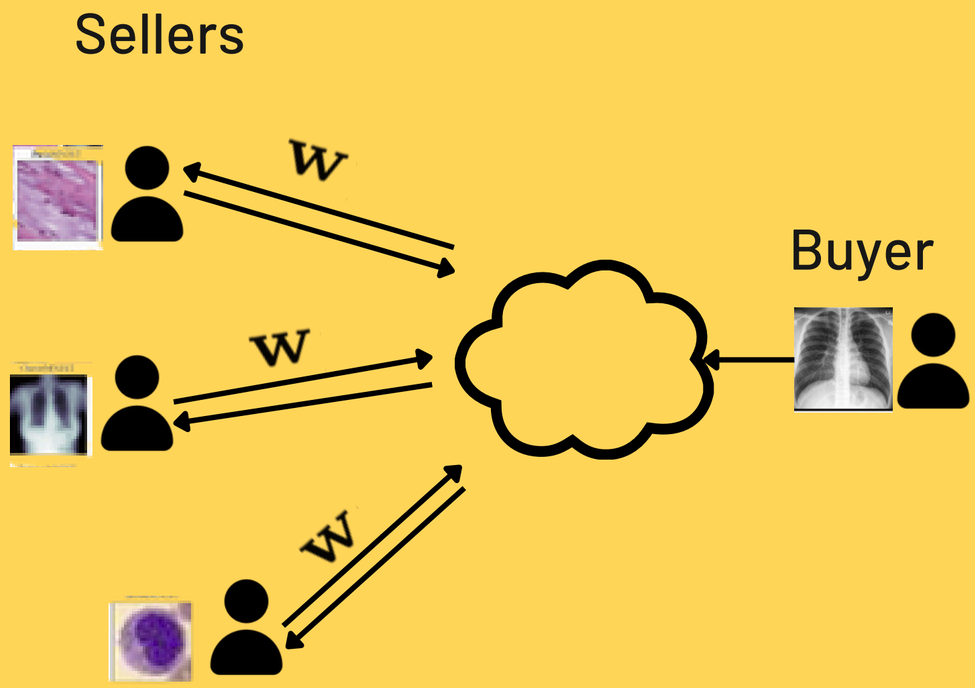
Iterative, collaborative algorithms

- Minimize error on buyer $\mathbf{x_0}$ within budget

$$\min_{\mathbf{p}^\top \mathbf{w} \leq B} \quad \mathbf{x_0}^\top \left(\sum_j w_j \mathbf{x_j} \mathbf{x_j}^\top\right)^\dagger \mathbf{x_0}$$

$w_j \in \{0, 1\}$

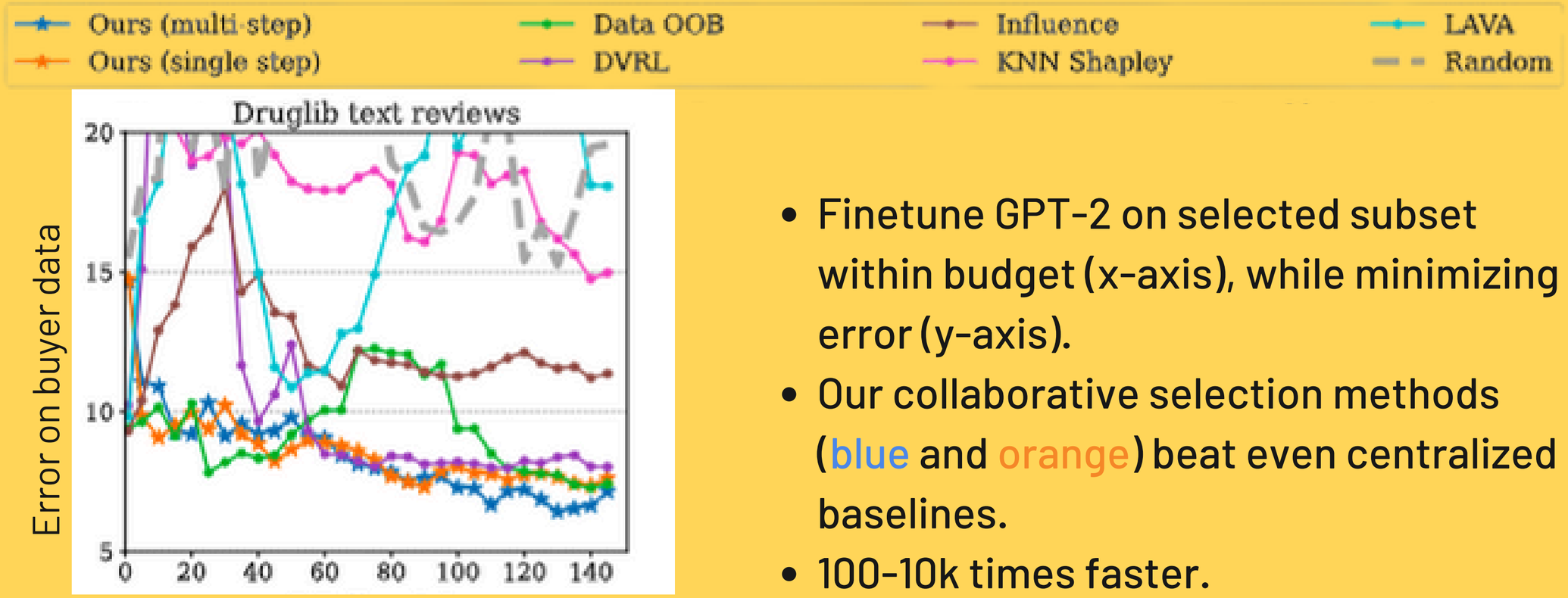
- NP Hard => O(1/B) convex approximation
- Using ***collaborative conditional gradient*** [Frank and Wolfe 1956]



[Lu, Huang, *Karimireddy*, Jordan, Raskar NeurIPS 2024]

Collaborative data discovery

Results



- Finetune GPT-2 on selected subset within budget (x-axis), while minimizing error (y-axis).
- Our collaborative selection methods (blue and orange) beat even centralized baselines.
- 100-10k times faster.

[Lu, Huang, *Karimireddy*, Jordan, Raskar NeurIPS 2024]

3

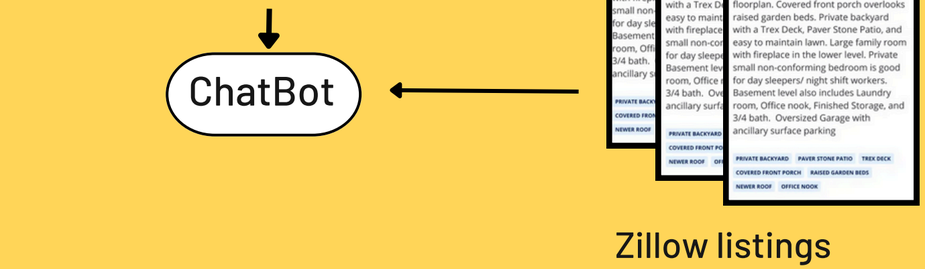
Conclusion

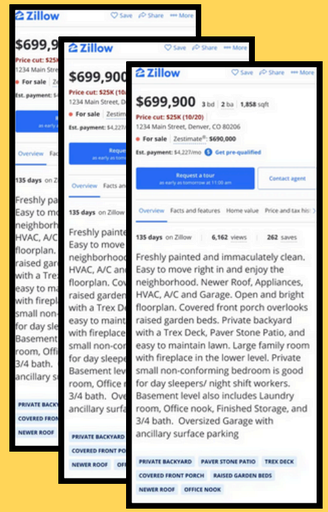


Better data understanding

Future work

Prompt
Which house should I buy?





Zillow listings

1. A theory of data utility
   - Beyond linear models
   - Statistically sound
   - Incentive compatible

2. What if data is manipulated or fake? (peer prediction/Bayesian persuasion + ML)

3. Establish authencity & provenance (watermarks, ZKP)

Collaborative Learning

Opportunity to

- rethink existing digital infrastructure
- make ML truly participatory
- center interests of people & society



Thank you.