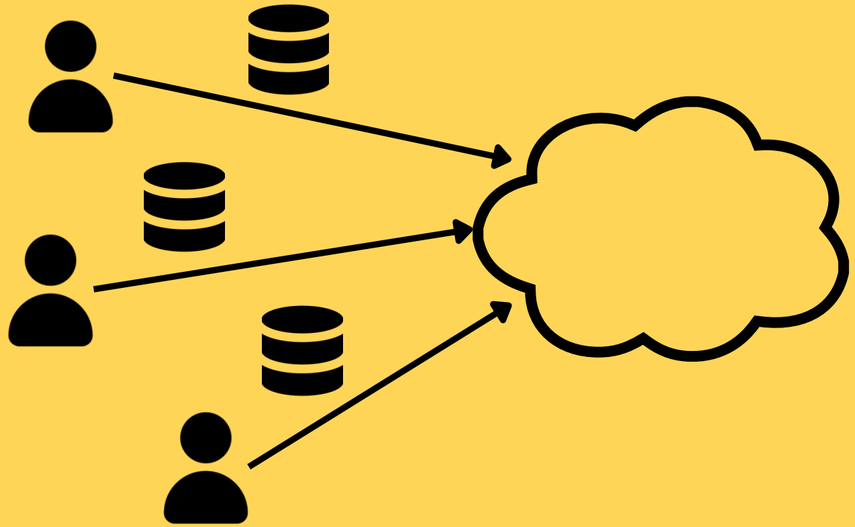
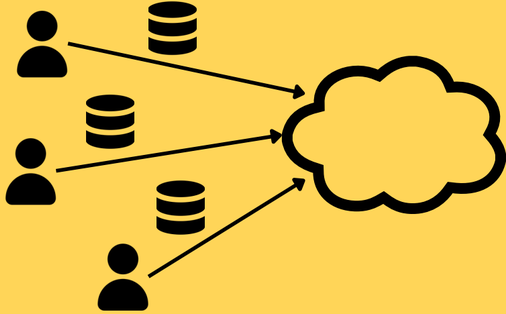
Building sustainable *data-ecosystems*

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.

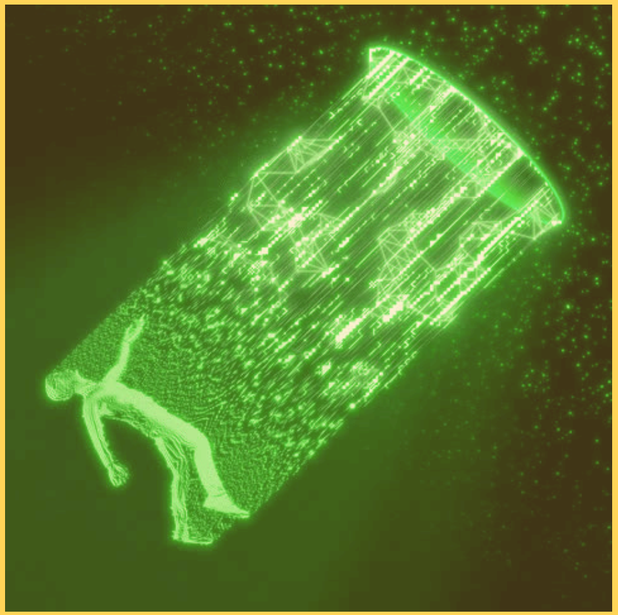


Data centralization is failing us

Failures to Individuals

Loss of Control

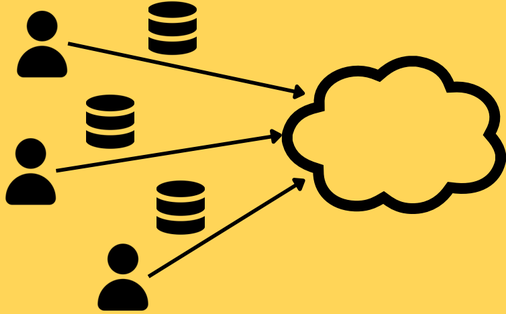
- No privacy
- Data stored indefinitely
- No control over usage which may be unaligned with my interests.



src: @perfectloop
used with permission

Data centralization is failing us

Failures to Society



Consent in Crisis:
The Rapid Decline of the AI Data Commons

Shayne Longpre[1], Robert Mahari[1], Ariel Lee[1], Campbell Lund[1], Hamidah Oderinwale[2], William Brannon[2], Nayan Saxena[2], Naana Obeng-Marnu[2], Tobin South[2], Cole Hunter[2], Kevin Klyman[2], Christopher Klamm[2], Hailey Schoelkopf[2], Nikhil Singh[2], Manuel Cherep[2], Ahmad Mustafa Anis[3], An Dinh[3], Caroline Chitongo[3], Da Yin[3], Damien Sileo[3], Deividas Mataciunas[3], Diganta Misra[3], Emad Alghamdi[3], Enrico Shippole[3], Jianguo Zhang[3], Joanna Materzynska[3], Kun Qian[3], Kush Tiwary[3], Lester Miranda[3], Manan Dey[3], Minnie Liang[3], Mohammed Hamdy[3], Niklas Muennighoff[3], Seonghyeon Ye[3], Seungone Kim[3], Shrestha Mohanty[3], Vipul Gupta[3], Vivek Sharma[3], Vu Minh Chien[3], Xuhui Zhou[3], Yizhi Li[3], Caiming Xiong[4], Luis Villa[4], Stella Biderman[4], Hanlin Li[4], Daphne Ippolito[4], Sara Hooker[4], Jad Kabbara[4], and Sandy Pentland[4]

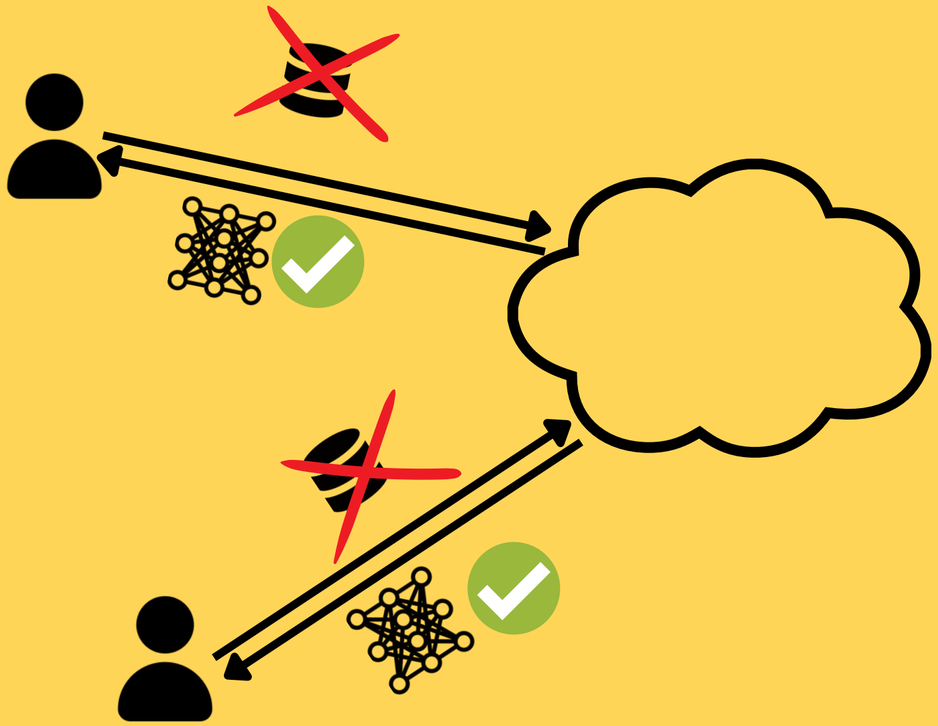
[1]Team Leads, [2]Top Contributors, [3]Contributors (alphabetized), [4]Advisors

Market Inefficiencies

- No sharing of gains => unsustainable data ecosystem

Collaborative Learning

User-centric ML

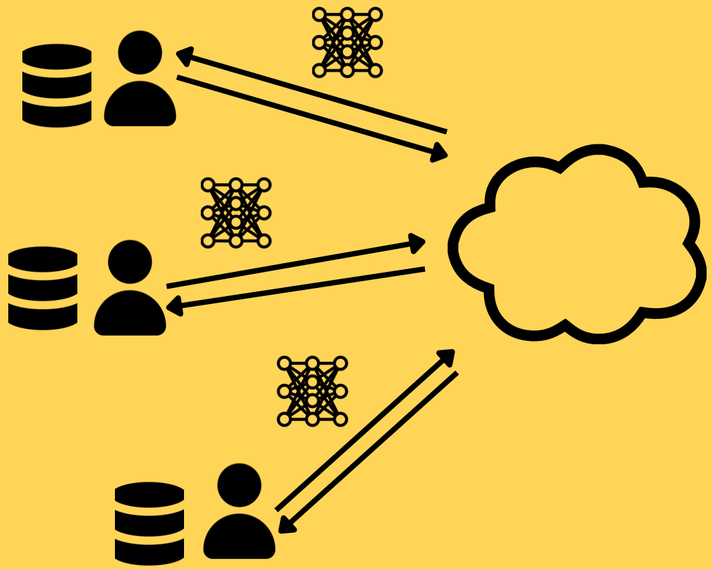


How to put users in control?

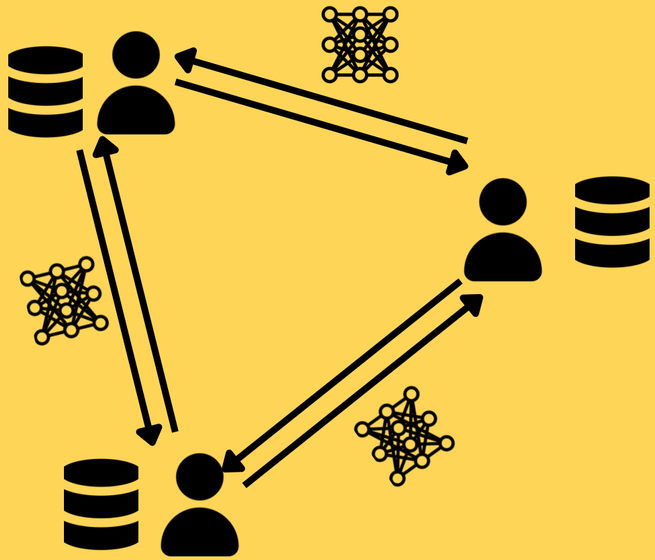
- Control over data => control over usage
  - don’t send your data
  - only aggregates (models, ...)
- Reconcile conflicting goals/preferences

Collaborative Learning

User-centric ML



Federated Learning
(central coordinator)



Decentralized Learning
(peer to peer)

Agenda

01

Distributed Training
Optimization

02

Aligning Goals
Economics

03

Data Markets
Opt+Econ

1

Scalable Distributed Training

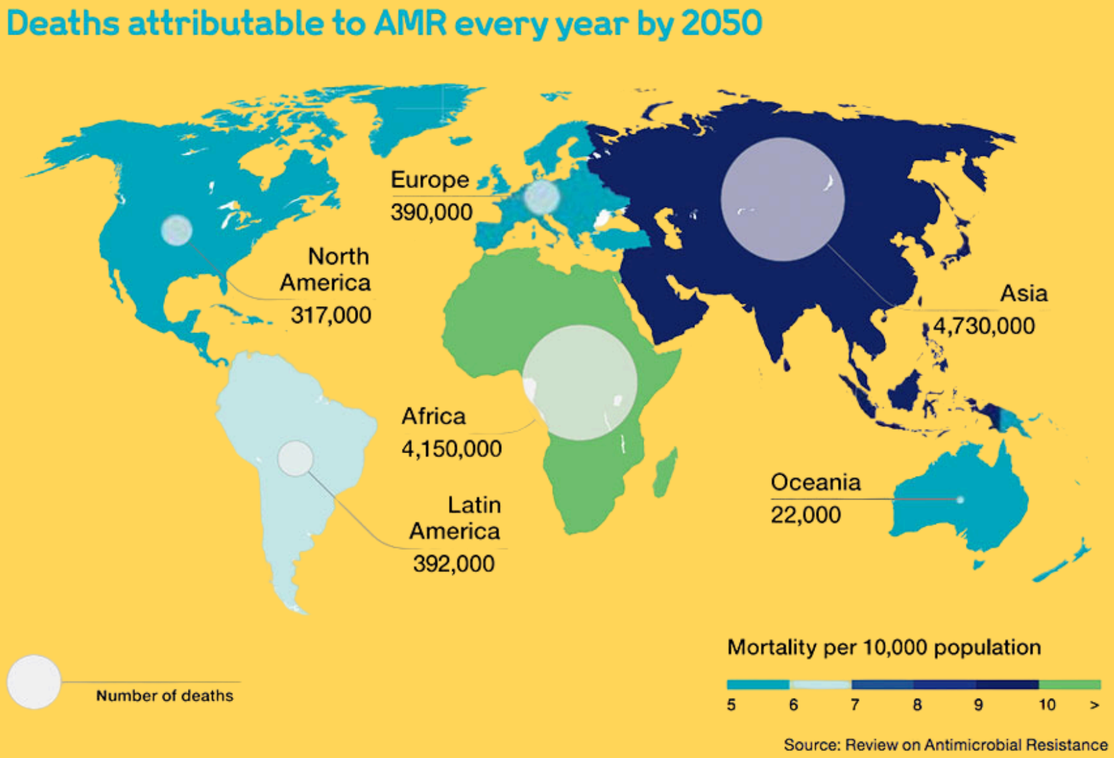
Fighting Superbugs



MEDECINS SANS FRONTIERES

Antimicrobial Resistance

A global crisis



"By 2050, 10 million deaths per year due to antimicrobial resistance."

- WHO 2014

- Outbreaks will start in low/middle-income regions and spread worldwide.

MEDECINS SANS FRONTIERES

Antimicrobial Resistance

Shortage of pathologists



MSF site in Yemen

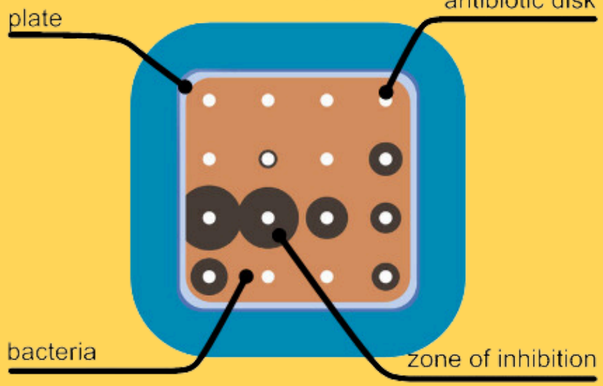
- WHO 2017 Recommendation: conduct an AST test before prescribing antibiotics.
- But, pathology experts are in **severe shortage**.

MEDECINS SANS FRONTIERES

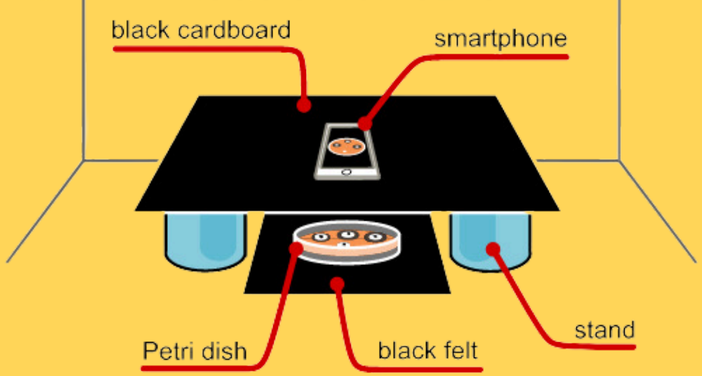
Antibiogo
Idea

Use **on-mobile ML** to automate the process.

a



b



Standardized data acquisition using a cheap kit

c



App analyzes image, gives results.

MEDECINS SANS FRONTIERES

Antibiogo

Goals

- Deploy in **15 countries** across Africa and Asia.
- Collect data in **real-time** for
  - epidemiological monitoring
  - improve ML model
- But, **privacy**. Data needs to remain distributed.



MEDECINS SANS FRONTIERES

Can we learn from distributed data?

Scalable, Resilient, Reliable Algorithms



Can we learn from distributed data?

Scalable, Resilient, Reliable Algorithms



Scale to massive networks

- [**K**KM+ ICML 20]
- [**K**JK+ NeurIPS 21]
- [YW**K**+ NeurIPS 22]
- ...

Resilient to data poisoning

- [**K**HJ ICML 2021]
- [**K**HJ ICLR 2022]
- [H**K**J SPICY-FL 2022]

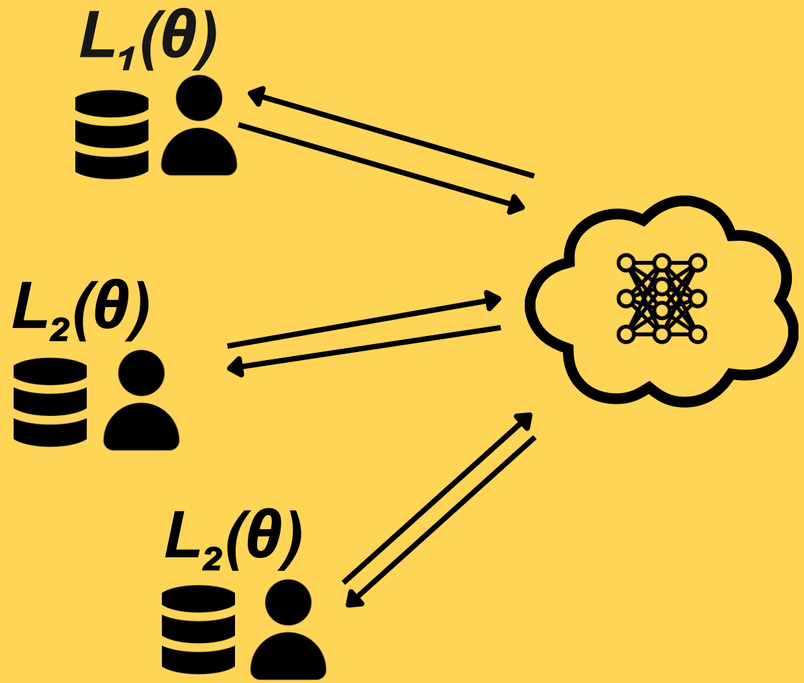
Collaborative conformal prediction

- [LY**K**+ ICML 23]

Scalable Collaborative Learning

How?

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

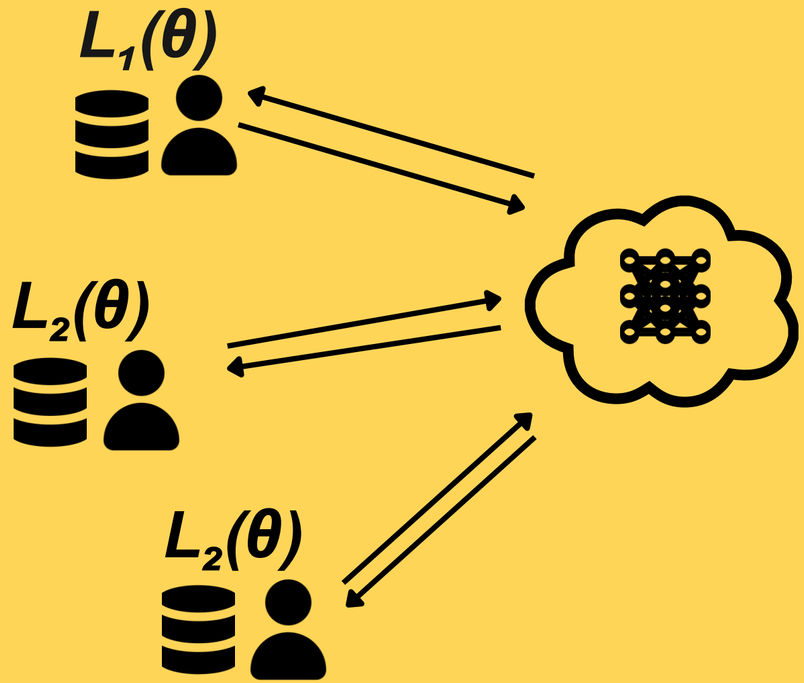


Federated Learning
[McMahan et al. 2016]

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

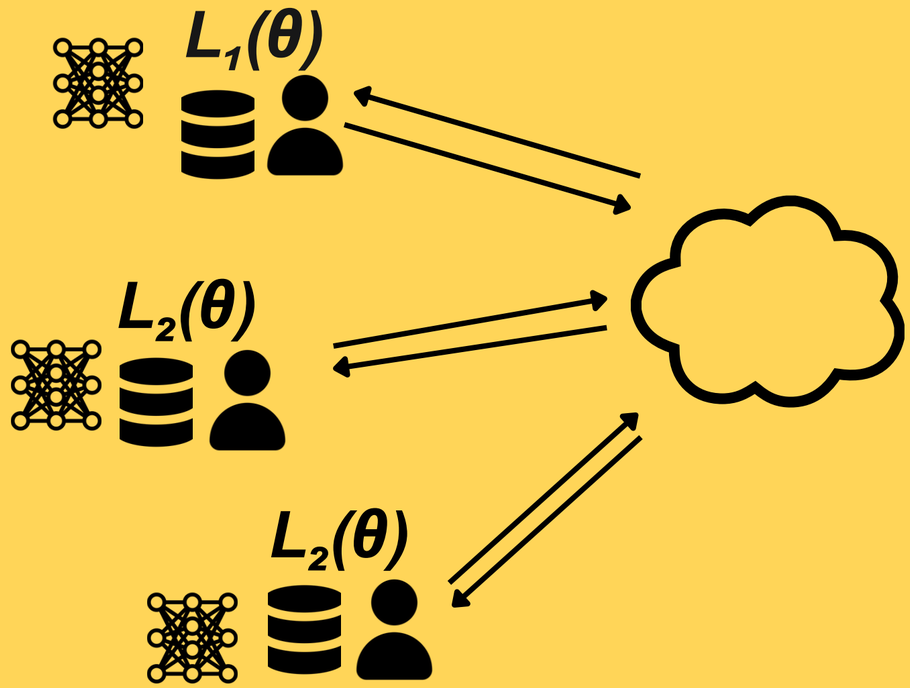


Federated Learning
[McMahan et al. 2016]

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 $\mathcal{L}_i(\theta_i)$



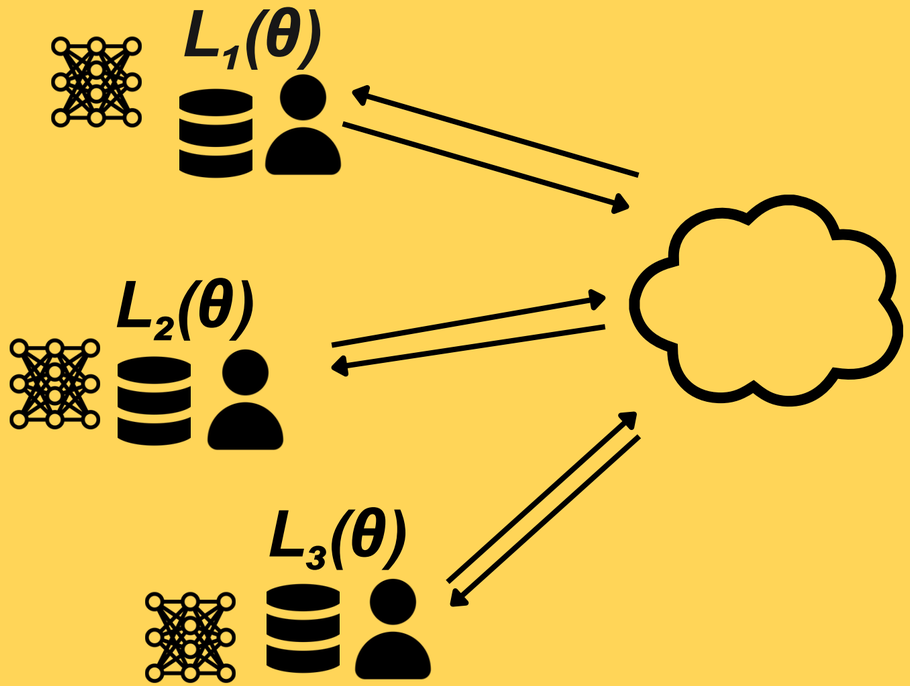
Federated Learning
[McMahan et al. 2016]

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

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



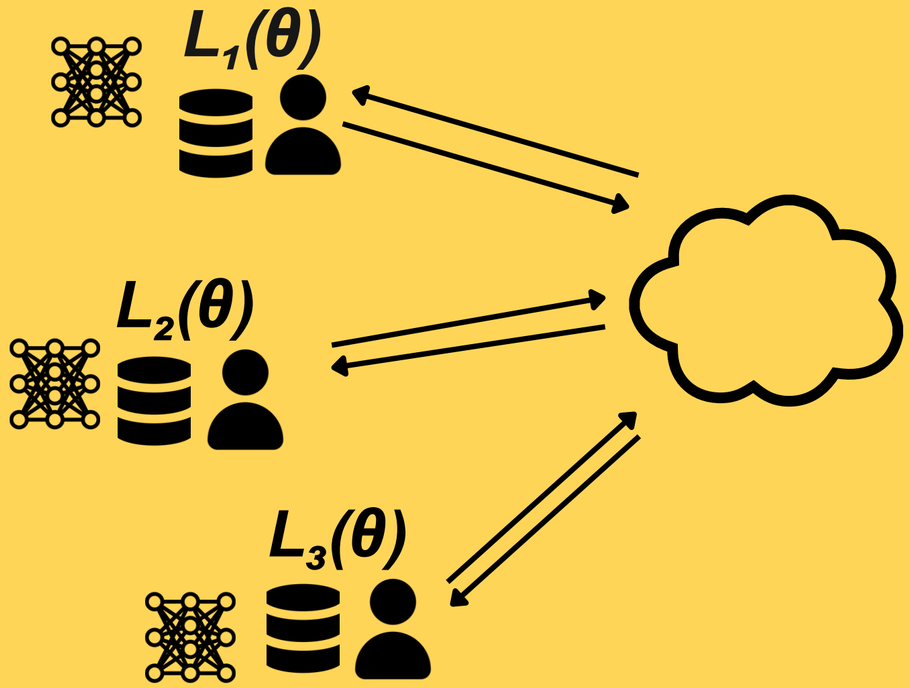
Formal security and privacy **important**, ask me later.

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

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



Scalable Collaborative Learning

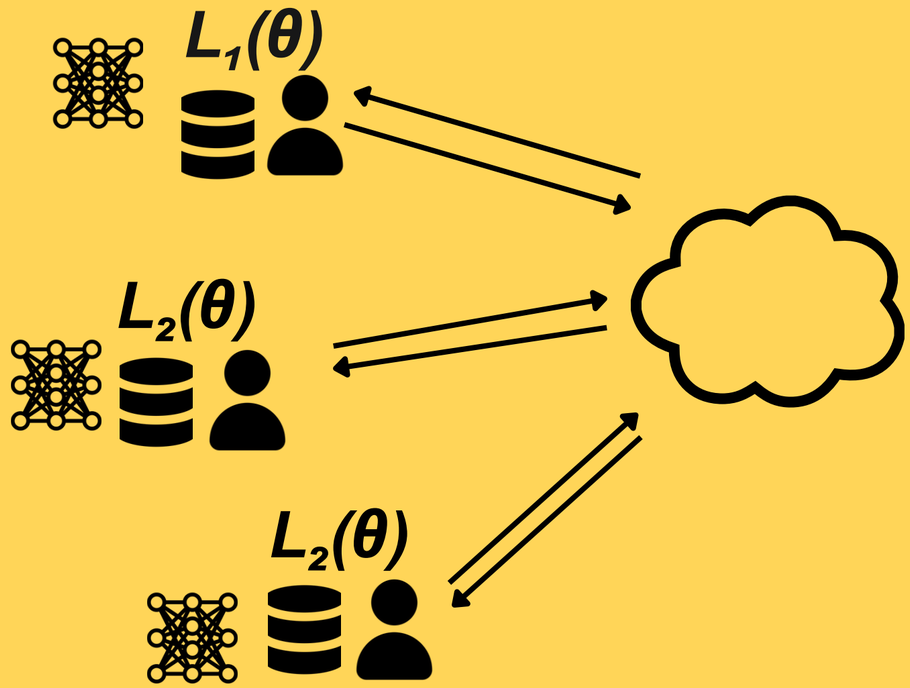
Distributed SGD?

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 using SGD**

$$\theta_i = \theta - \nabla \mathcal{L}_i(\theta)$$

   - Synchronize parameters

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



Scalable Collaborative Learning

Communication bottleneck

	Single GPU	Collaboration between USA and EU	
computation (per round)	1ms	1ms	keep same
communication per round	1ms	**1s**	reduce by 1000x?
training time	1 hr	**5 days!**	

Scalable Collaborative Learning

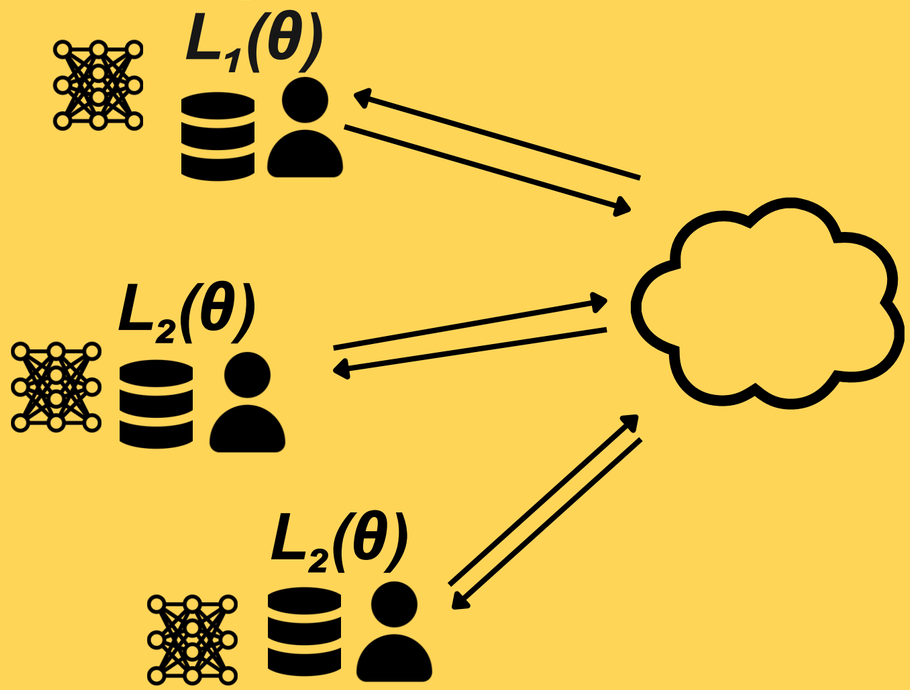
FedAvg: communcation-efficient learning

1. Iteratively,
   - Copy global θ
   - **Update θ_i using 1k SGD steps**

$$\overset{1k}{\circlearrowleft}\ \theta_i = \theta_i - \nabla \mathcal{L}_i(\theta_i)$$

   - Synchronize parameters

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



FedAvg

[McMahan et al. 2016]

Scalable Collaborative Learning

FedAvg - the problem

1. Iteratively,
   - Copy global θ
   - **Update θ_i using 1k SGD steps**

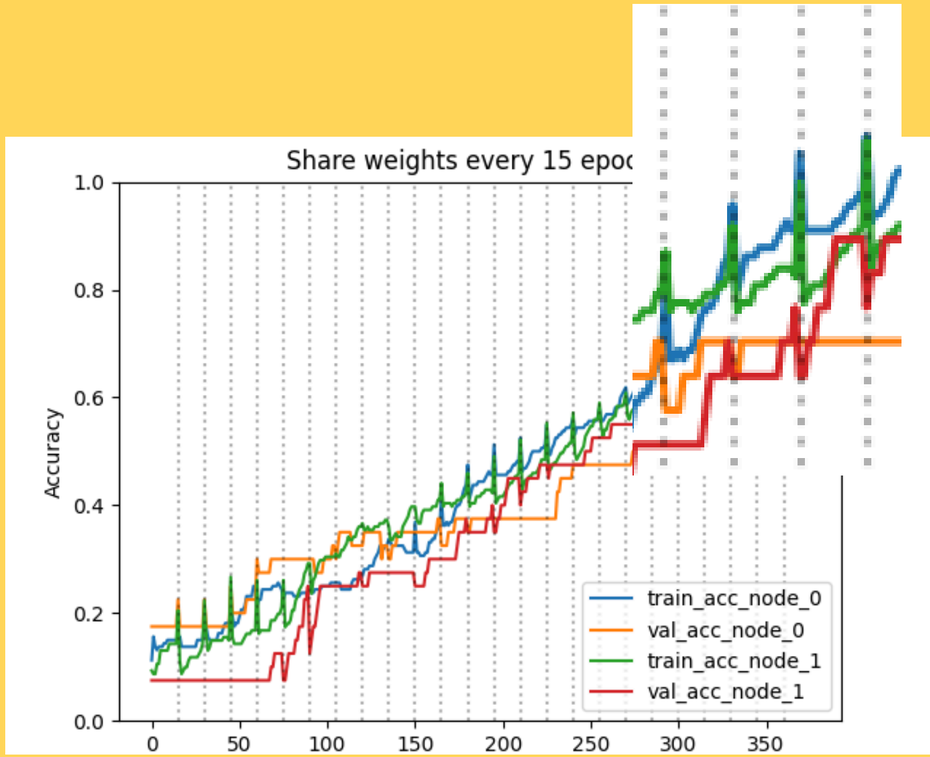


$$\theta_i = \theta_i - \nabla \mathcal{L}_i(\theta_i)$$

   - Synchronize parameters

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

Overfits to local data :(



Scalable Collaborative Learning

Prior work

- Empirical work: [Zhao et al. 2018; Hsieh . 2019, ...]
- Theoretical analysis:
  - i.i.d data [Zinkevich et al., 2010; Stich 2018,; Patel & Dieuleveut 2019, ...]
  - general [Wang et al., 2019; Yu et al., 2019; Li et al., 2018; Haddadpour & Mahdavi, 2019; Khaled et al., 2020, ...]
  - Lower bounds [Arjevani & Shamir 2015] - complexity depends on Hessian similarity **δ**, not smoothness (**L**) where **δ << L.**

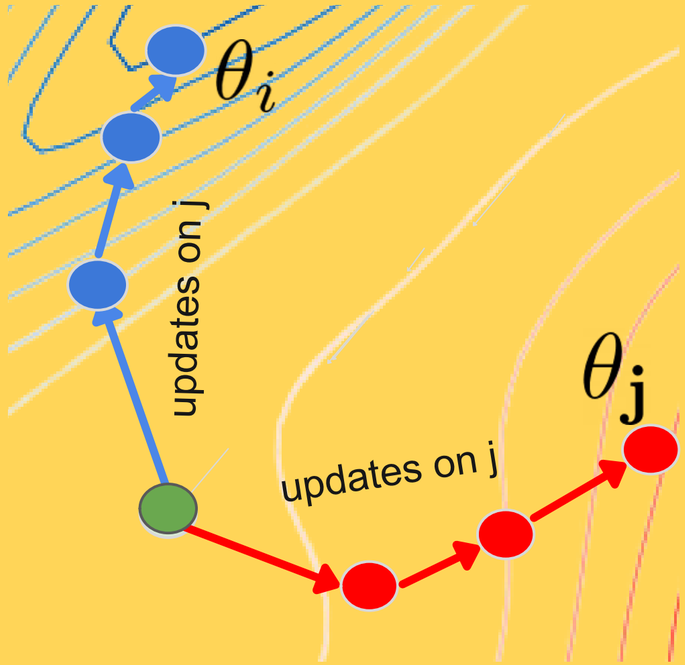
- None explain why FedAvg overfits or how to fix it.
- None match lower bounds, taking advantage of Hessian similarity δ.

Scalable Collaborative Learning

Our lower bound analysis

- Main culprit is **data heterogeneity**.
- The updates to $\theta_{\mathbf{i}}$ on each user pull them far apart.
- Gradients computed at vastly different points become incompatible.

$$\theta_{\mathbf{i}} = \theta_{\mathbf{i}} - \gamma \nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}})$$
$$\theta_{\mathbf{j}} = \theta_{\mathbf{j}} - \gamma \nabla \mathcal{L}_{\mathbf{j}}(\theta_{\mathbf{j}})$$



[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning

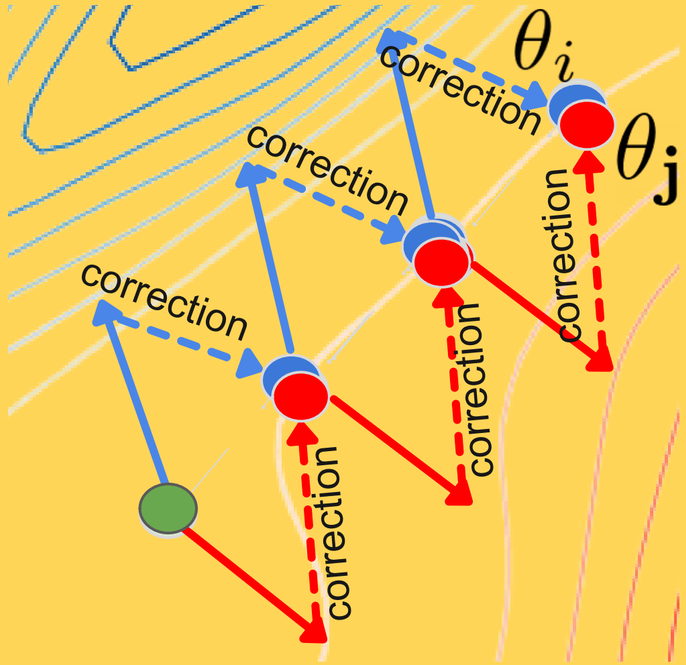
SCAFFOLD correction

1. Compute correction (using history) to reduce separation.
2. Iteratively,
   - Update θ_i using 1k SGD**+correction**

1k $\theta_i = \theta_i - \gamma(\nabla \mathcal{L}_i(\theta_i) + c - c_i)$

   - Synchronize parameters

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



[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning
SCAFFOLD performance

- The convergence rate improves from $O(1/t^2)$ to $O(\exp(-t))$.
- Nearly optimal communication-complexity, taking advantage of Hessian similarity $\delta \ll L$.

[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning

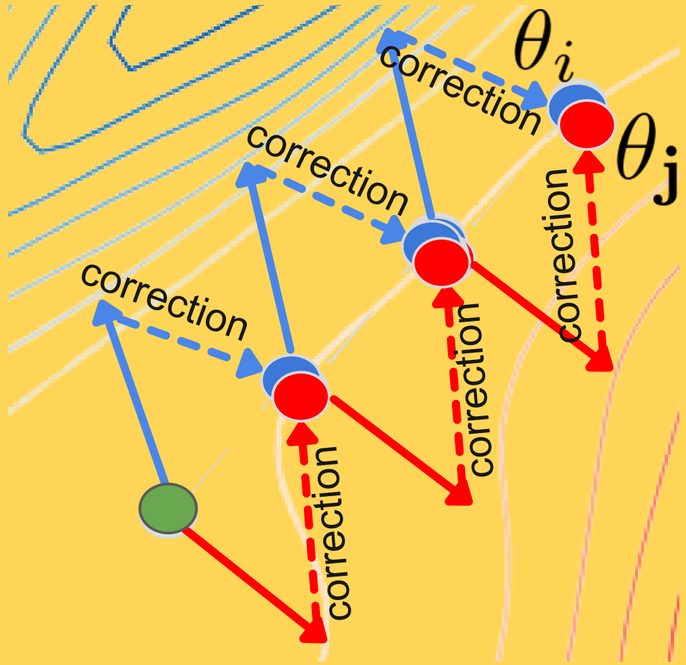
Bias correction proof sketch

Update

$$\nabla \mathcal{L}_i(\theta_i) + c - c_i$$

Target

$$- \nabla \mathcal{L}(\theta_i)$$



[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning

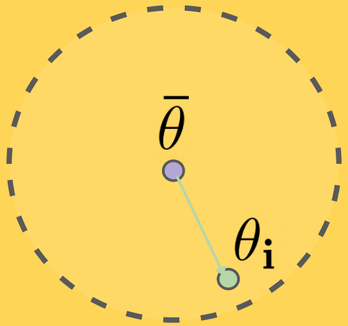
Bias correction proof sketch

$$\nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}}) + \mathbf{c} - \mathbf{c_i} - \nabla \mathcal{L}(\theta_{\mathbf{i}})$$

$$\approx \nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}}) + \nabla \mathcal{L}(\bar{\theta}) - \nabla \mathcal{L}_{\mathbf{i}}(\bar{\theta}) - \nabla \mathcal{L}(\theta_{\mathbf{i}})$$

$$= \nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\theta_{\mathbf{i}}) - \nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\bar{\theta})$$

$$\approx \delta \|\theta_{\mathbf{i}} - \bar{\theta}\|_2$$

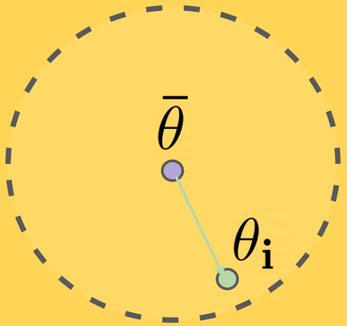


[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning

Bias correction proof sketch

$$
\begin{aligned}
&\nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}}) + \mathbf{c} - \mathbf{c_i} - \nabla \mathcal{L}(\theta_{\mathbf{i}}) \\
&\approx \boxed{\nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}})} + \nabla \mathcal{L}(\bar{\theta}) - \nabla \mathcal{L}_{\mathbf{i}}(\bar{\theta}) - \boxed{\nabla \mathcal{L}(\theta_{\mathbf{i}})} \\
&= \boxed{\nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\theta_{\mathbf{i}})} - \nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\bar{\theta}) \\
&\approx \delta \|\theta_{\mathbf{i}} - \bar{\theta}\|_2
\end{aligned}
$$



[*Karimireddy* et al. ICML 2020]

Scalable Collaborative Learning

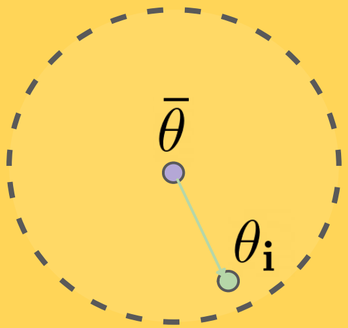
Bias correction proof sketch

$$\nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}}) + \mathbf{c} - \mathbf{c_i} - \nabla \mathcal{L}(\theta_{\mathbf{i}})$$

$$\approx \nabla \mathcal{L}_{\mathbf{i}}(\theta_{\mathbf{i}}) + \nabla \mathcal{L}(\bar{\theta}) - \nabla \mathcal{L}_{\mathbf{i}}(\bar{\theta}) - \nabla \mathcal{L}(\theta_{\mathbf{i}})$$

$$= \nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\theta_{\mathbf{i}}) - \nabla \left(\mathcal{L}_i - \mathcal{L}\right)(\bar{\theta})$$

$$\approx \delta \|\theta_{\mathbf{i}} - \bar{\theta}\|_2$$



[*Karimireddy* et al. ICML 2020]

Bias Correction
Extensions

source correction target

$$\nabla \mathcal{L}_i(\theta_i) + \mathbf{c} - \mathbf{c}_i - \nabla \mathcal{L}(\theta_i)$$

Adapt source gradients to target

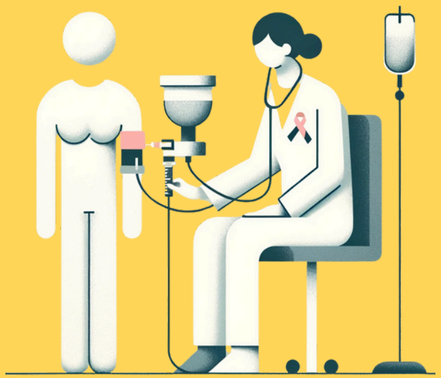
- multi-task learning
- domain adaptation
- debiasing synthetic data?

[Chayti & *Karimireddy* TMLR 24]

Scalable Collaborative Learning

SCAFFOLD impact

Neoadjuvant chemotherapy (NACT) for triple-negative breast cancer (TNBC)



OWKIN



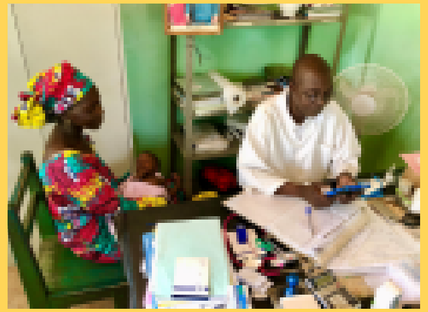
*"Scaffold is the **most promising** collaborative strategies* (sic)"
[Terrail et al. Nat Medicine 2023]

Scalable Collaborative Learning

Antibiogo deployment



- Scalable and robust collaborative learning.
- **Clinical trials** underway in Africa!
- Next,
  - EU trials and global deployment.
  - Integrate CL everywhere?



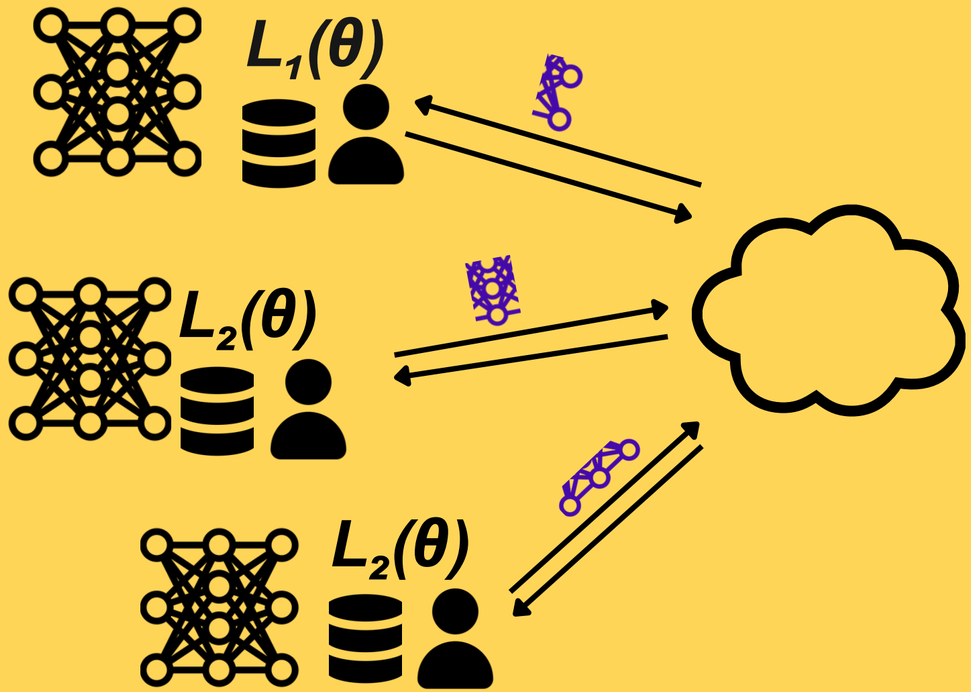
[Under preparation]

MEDECINS SANS FRONTIERES

Scalable Collaborative Learning

Compressed communication

- Iteratively need to
  - Synchronize parameters.
- But GPT-2 is 5GB!
- Idea: Send only the most important bits using **compression**.
- Correct for bias using **error-feedback.**



[*Karimireddy* et al. ICML 2019,
Stich and *Karimireddy*. JMLR 2020]
PowerSGD [Vogels, *Karimireddy*, Jaggi. NeurIPS 2019]

Also used outside CL

Scalable CL => scalable general learning

Accelerating PyTorch DDP by 10X With PowerSGD

PyTorch Follow

Nov 4, 2021 · 8 min read

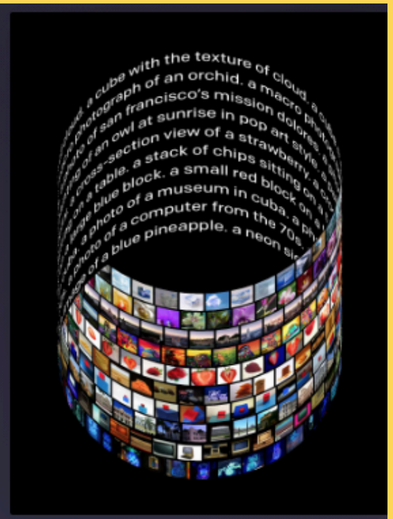
Authors: *Yi Wang (Facebook AI), Alex Iankoulski (Amazon AWS), Pritam Damania (Facebook AI), Sundar Ranganathan (Amazon AWS)*

Meta

DALL·E: Creating Images from Text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

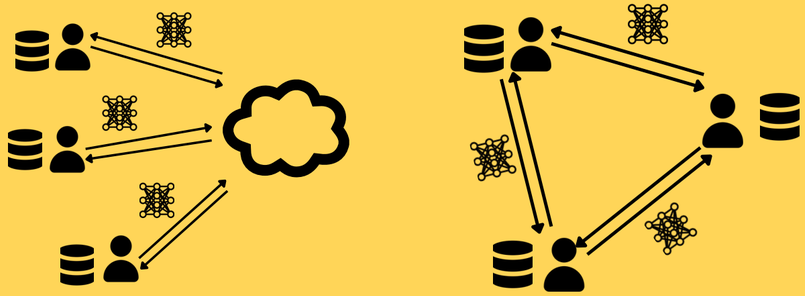
January 5, 2021
27 minute read



OpenAI

PowerSGD [Vogels, *Karimireddy*, Jaggi. NeurIPS 2019]

Data Decentralization
Are we there yet?



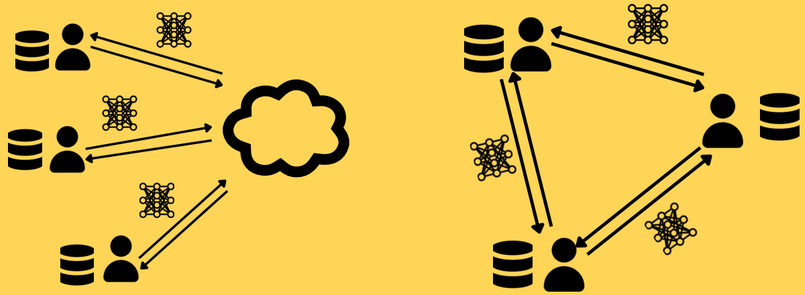
Loss of Control

- No privacy
- Data stored indefinitely
- No control over usage which may be unaligned with my interests.

Market Inefficiencies

- Generates monopolies, prevents competition.
- Data is not reused or repurposed.

Data Decentralization
Are we there yet?



Loss of Control

- No privacy
- Data stored indefinitely
- **No control over usage** which may be unaligned with my interests.

Market Inefficiencies

- No sharing of profit => unsustainable data ecosystem

Agenda

01

Distributed Training
Optimization

02

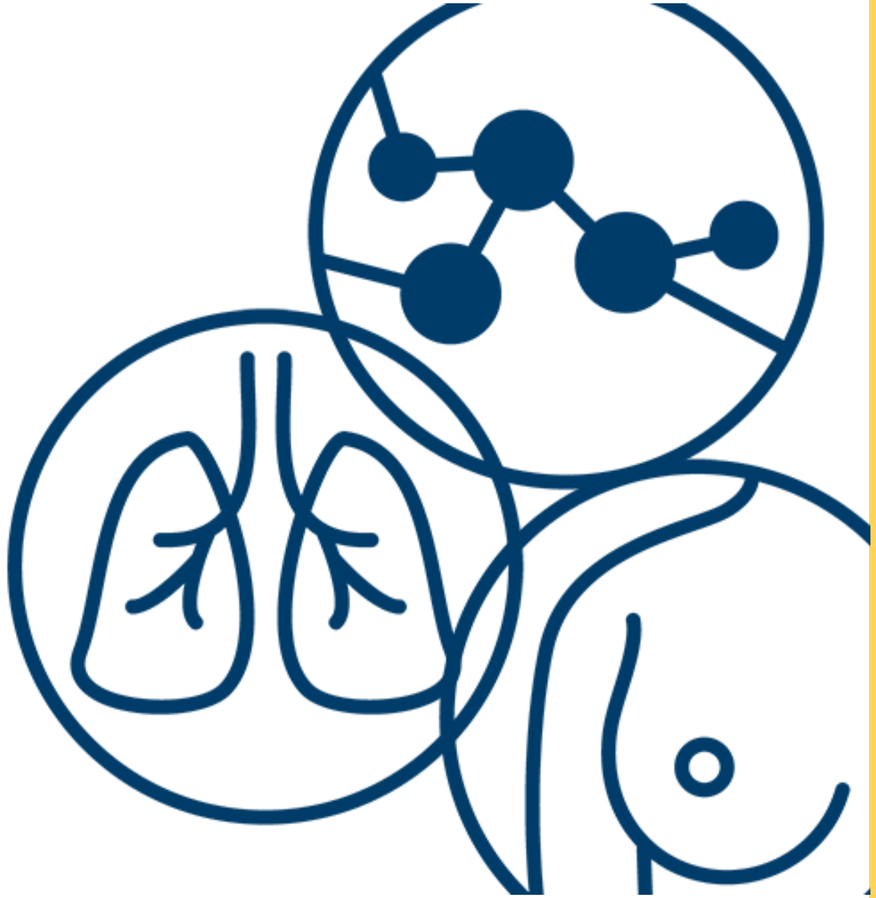
Aligning Goals
Economics

03

Data Markets
Opt+Econ

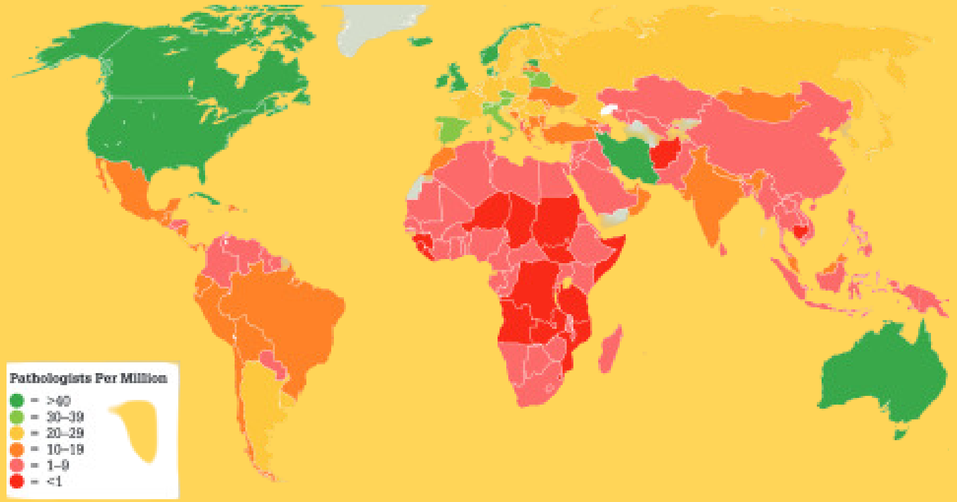
2

Aligning Diverse Goals
Fighting Cancer



Cancer Registry of Norway

Shortage of Cancer expertise



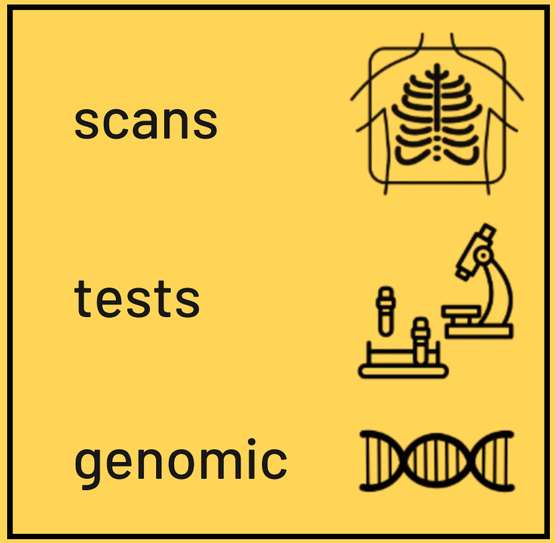
> **"Cancer patients face life-threatening delays due to lack of staff, say UK radiologists."**
>
> — The Guardian, 8 Jun 2023

- 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

Use AI to meet shortages

AI tools to aid and stretch expertise show great promise, but they need a lot of highly private data.



Sensitive patient data

ML model

- **Classify** cancer grades/stages
- **Segment** regions of interest in scans
- **Predict** future risk

Cancer Registry of Norway

Data scarcity & diversity

- Very little data. 562 cases of Sarcoma in 2022 across Norway.
- Worse for marginal groups. No dark-skinned skin cancer data.
- True in the US as well. Gets dire in low-income regions.





"There are two big areas... One is data systems..The second is ...reaching populations that have been left out."
- Bertagnolli 2024 (new NIH Chief)

Cancer Registry of Norway

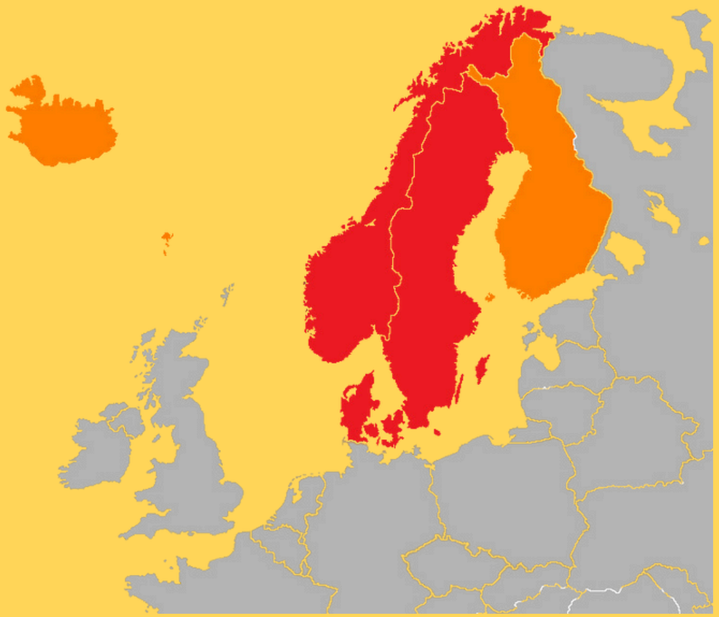
NORDCAN Collaboration

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. Need econ/game theory.



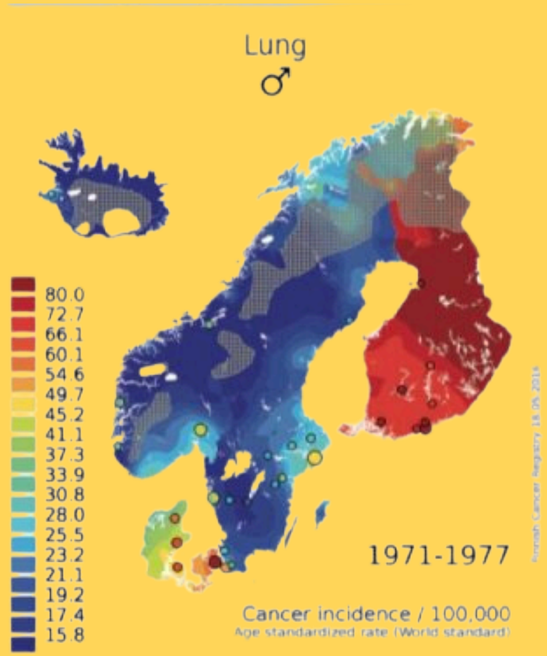
Cancer Registry of Norway

Position paper with Norway cancer registry

[*Karimireddy* et al. FMEC 2023]

Principal-Agents 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?
- Is any strategy both locally and globally optimal?



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

[Huang, *Karimireddy*, Jordan. Under review JMLR]

Principal-Agents Formalism

Efficient data sampling strategies

Is the sampling strategy 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.



Principal-Agents Formalism

Efficient data sampling strategies

Is the sampling strategy 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.



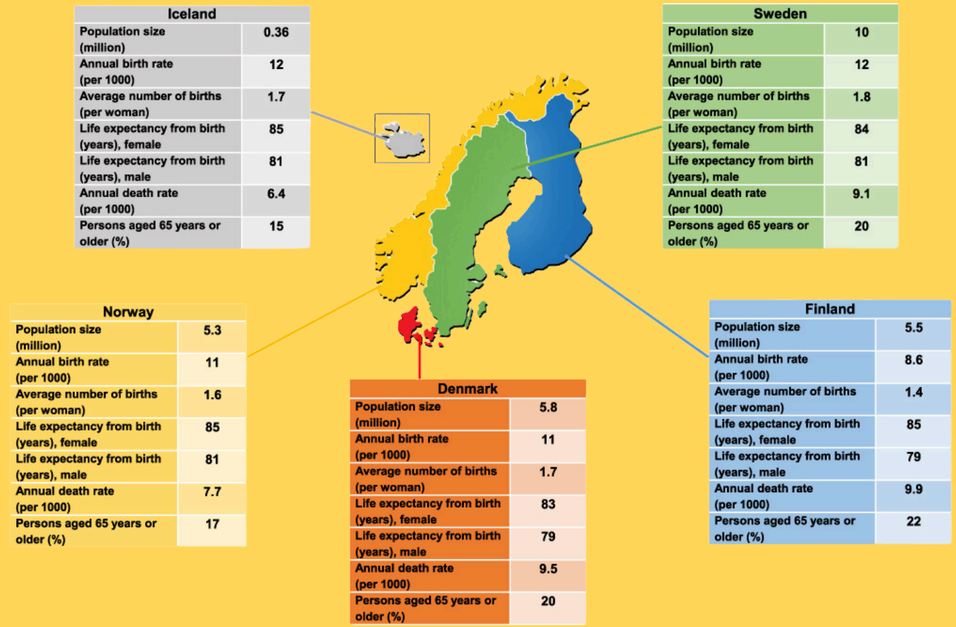
Principal-Agent Experiment Design

Heterogeneous populations

- Data types
  - **type** encoded by $\mathbf{x}$ = (30 yrs, Male, Asian, ...)
  - Total population has many types:

$$\mathcal{X} = (\mathbf{x}_j)_{j=1,\dots}$$

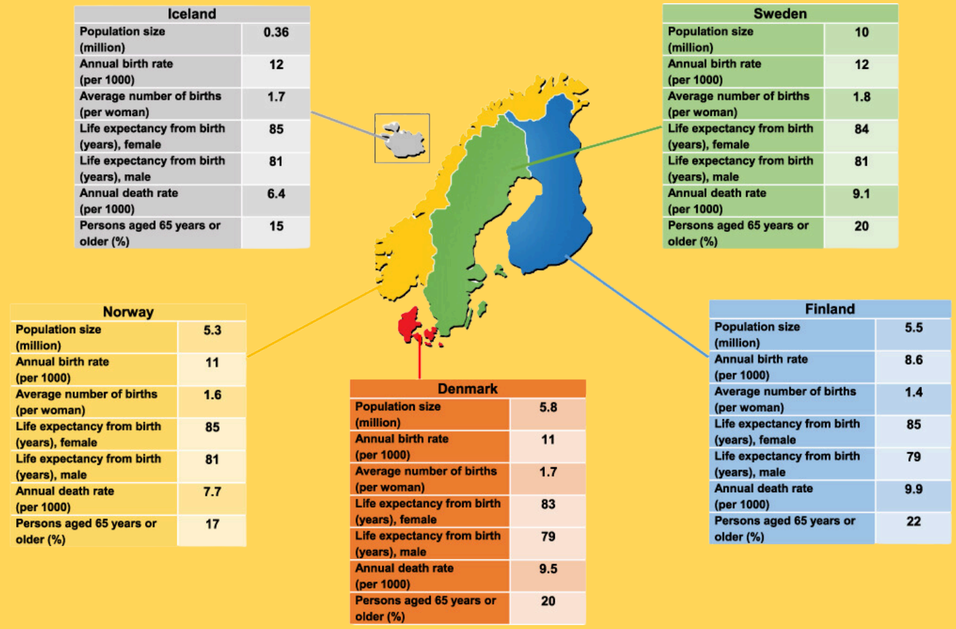
  - Partitioned between ***n*** agents. Agent i only has subtypes $\mathcal{X}_i \subseteq \mathcal{X}$.



Principal-Agent Experiment Design

Heterogeneous populations & contributions

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



Iceland	
Population size (million)	0.36
Annual birth rate (per 1000)	12
Average number of births (per woman)	1.7
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	6.4
Persons aged 65 years or older (%)	15

Sweden	
Population size (million)	10
Annual birth rate (per 1000)	12
Average number of births (per woman)	1.8
Life expectancy from birth (years), female	84
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	9.1
Persons aged 65 years or older (%)	20

Norway	
Population size (million)	5.3
Annual birth rate (per 1000)	11
Average number of births (per woman)	1.6
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	81
Annual death rate (per 1000)	7.7
Persons aged 65 years or older (%)	17

Denmark	
Population size (million)	5.8
Annual birth rate (per 1000)	11
Average number of births (per woman)	1.7
Life expectancy from birth (years), female	83
Life expectancy from birth (years), male	79
Annual death rate (per 1000)	9.5
Persons aged 65 years or older (%)	20

Finland	
Population size (million)	5.5
Annual birth rate (per 1000)	8.6
Average number of births (per woman)	1.4
Life expectancy from birth (years), female	85
Life expectancy from birth (years), male	79
Annual death rate (per 1000)	9.9
Persons aged 65 years or older (%)	22

Principal-Agent Experiment Design

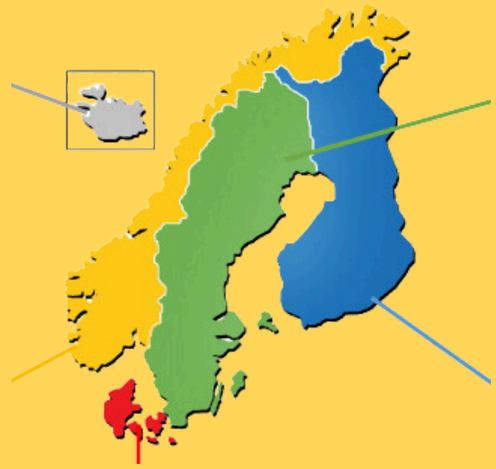
Efficiency games

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

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

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

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



Classic Experiment Design

Estimating error

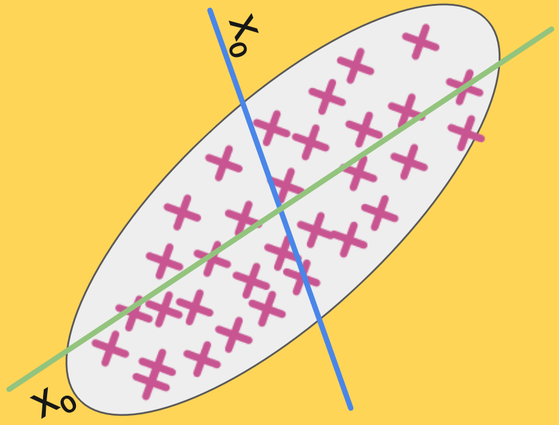
- Data space $\mathcal{X}$
- Total data sampled $\mathbf{m}$
- Agent i has types $\mathcal{X}_\mathbf{i}$
- Contributes $\mathbf{m}_i = (m_j \,|\mathbf{x}_j \in \mathcal{X}_i)$

- 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 $\mathcal{I}(\mathbf{m})$

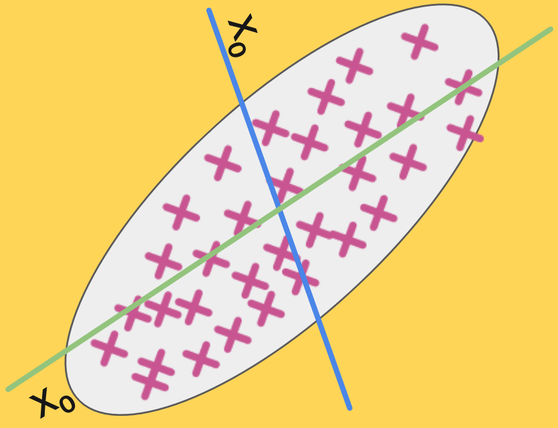
$$\mathcal{E}(\mathbf{x_0}, \mathbf{m}) = \mathbf{x_0}^\top (\sum_{j=1}^{k} m_j \mathbf{x_j} \mathbf{x_j}^\top)^\dagger \mathbf{x_0}$$
$$= \mathbf{x_0}^\top \mathcal{I}(\mathbf{m})^\dagger \mathbf{x_0}$$

- 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]

Principal-Agent Experiment Design

Estimating agent's error

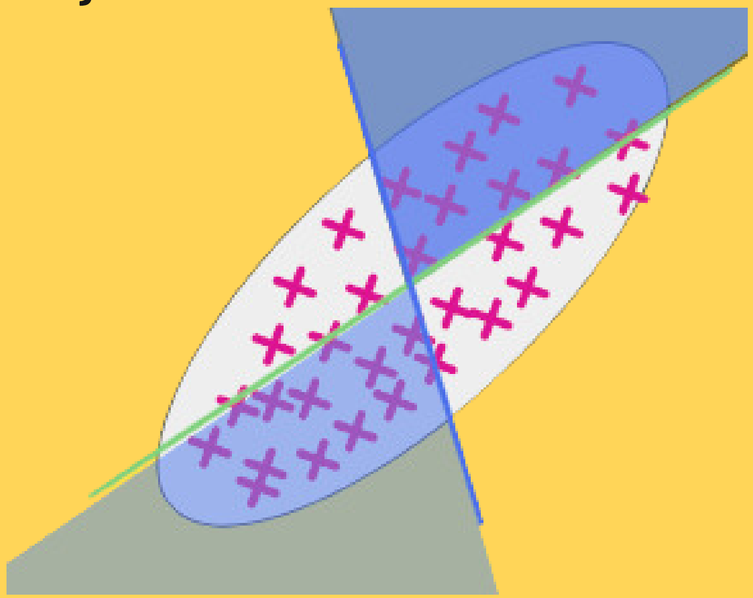
- Data space $\mathcal{X}$
- Total data sampled $\mathbf{m}$

- 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

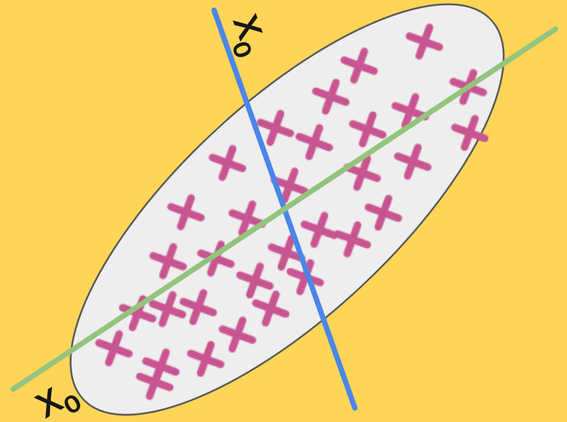


Principal-Agent Experiment Design

Our Theorem

- 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, *Karimireddy*, Jordan. Under review JMLR]

Principal-Agent Experiment Design

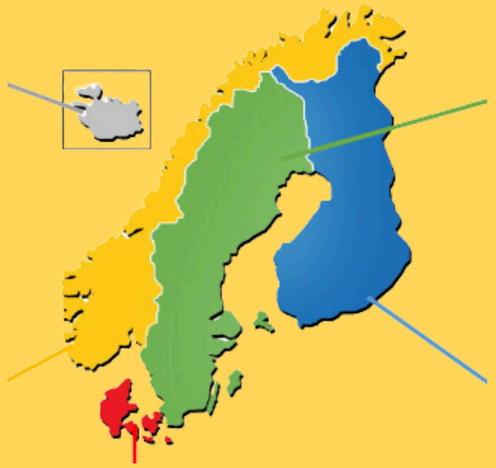
Proof Sketch

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

$$\frac{d}{d\,m_j}\left(\log\det(\Pi_i \mathcal{I}(\mathbf{m})^\dagger \Pi_i) + c\|\mathbf{m}\|_1\right)$$

$$= -\underbrace{\mathbf{x_j}^\top \mathcal{I}(\mathbf{m})^\dagger \mathbf{x_j}}_{\mathcal{E}(\mathbf{x_j},\mathbf{m})} + c$$

$$\Rightarrow \mathcal{E}(\mathbf{x_j}, \mathbf{m}) = c \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_i.$$



[Huang, *Karimireddy*, Jordan. Under review JMLR]

Principal-Agent Experiment Design

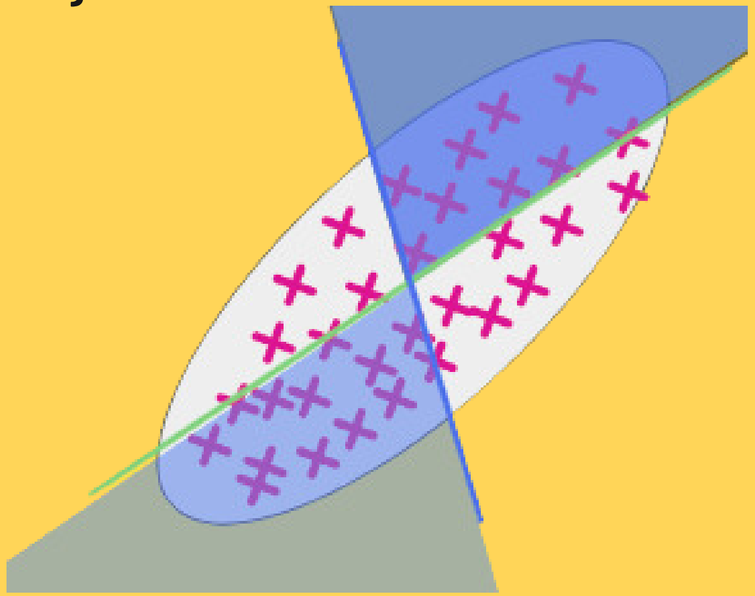
Free-riding games

- 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



[Huang, *Karimireddy*, Jordan. Under review JMLR]

Principal-Agents Formalism

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.



[Huang, *Karimireddy*, Jordan. ArXiv 2023]

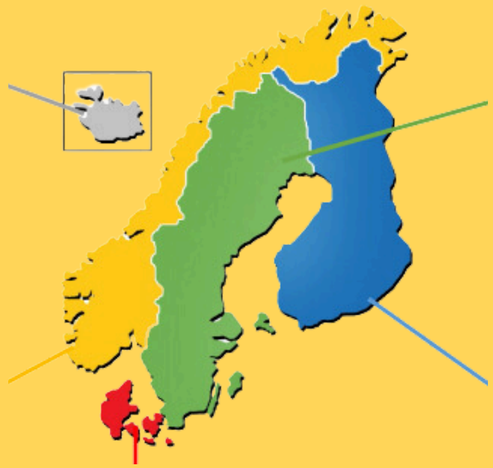
Principal-Agent Experiment Design

Our Theorems

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.

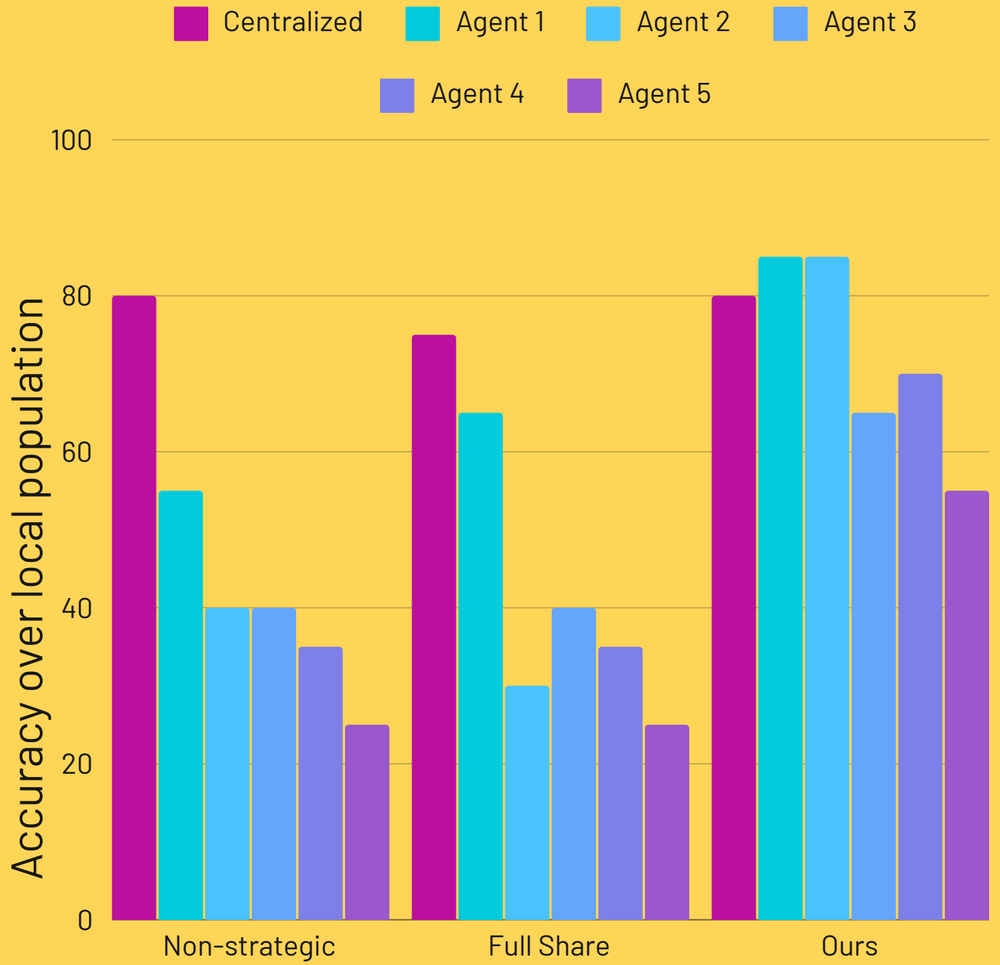


[Huang, *Karimireddy*, Jordan. ArXiv 2023]

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”

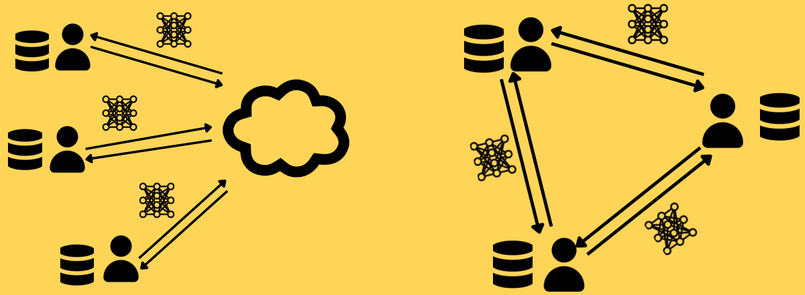


[Huang, *Karimireddy*, Jordan. ArXiv 2023]

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.
  - What definition and level of privacy? [FORC 24]

Data Decentralization
Are we there yet?



Loss of Control

- No privacy
- Data stored indefinitely
- **No control over usage** which may be unaligned with my interests.

Market Inefficiencies

- No sharing of profit => unsustainable data ecosystem

Agenda

01

Distributed Training
Optimization

02

Aligning Goals
Economics

03

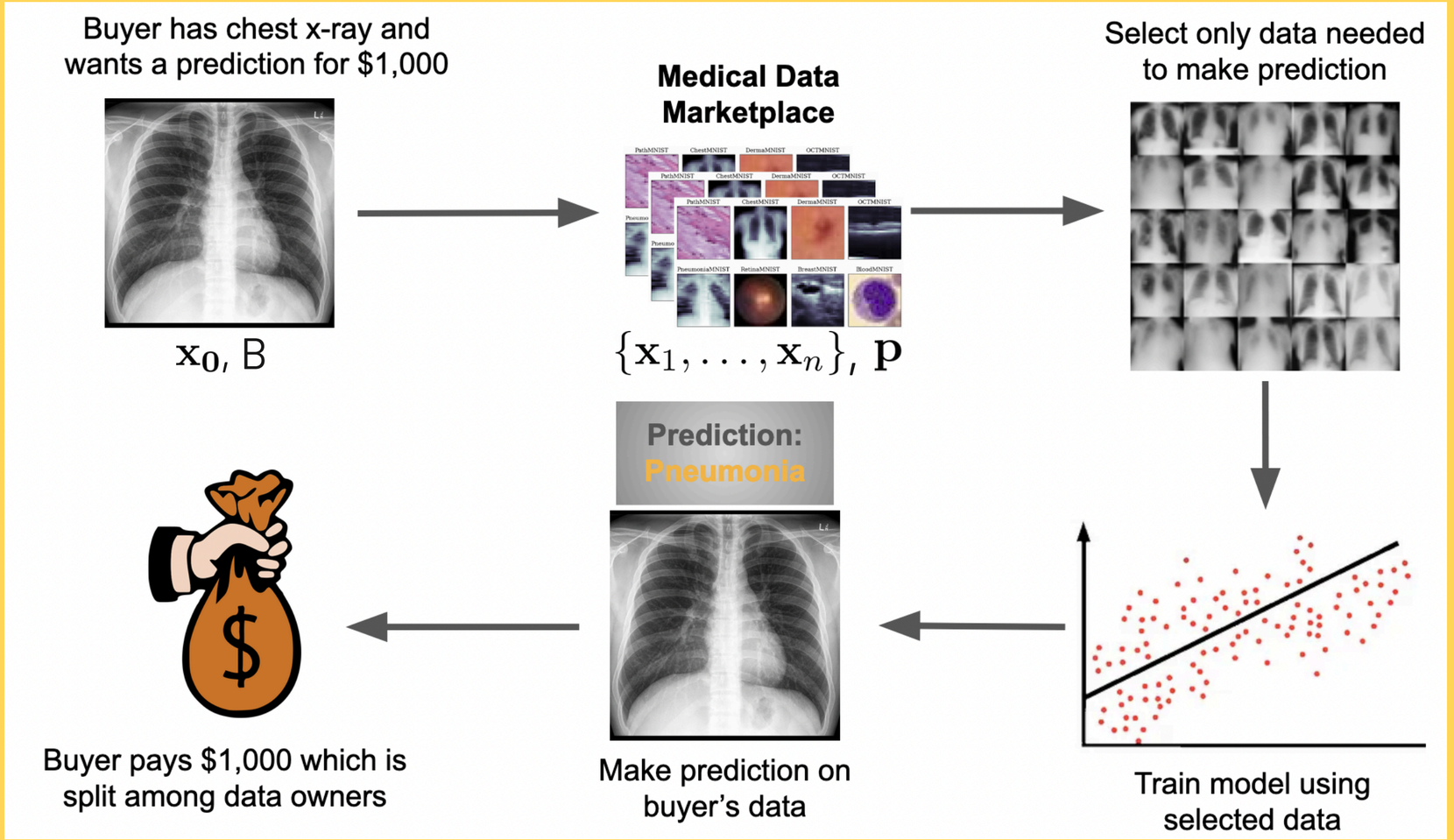
Data Markets
Opt+Econ

3

Open data marketplace

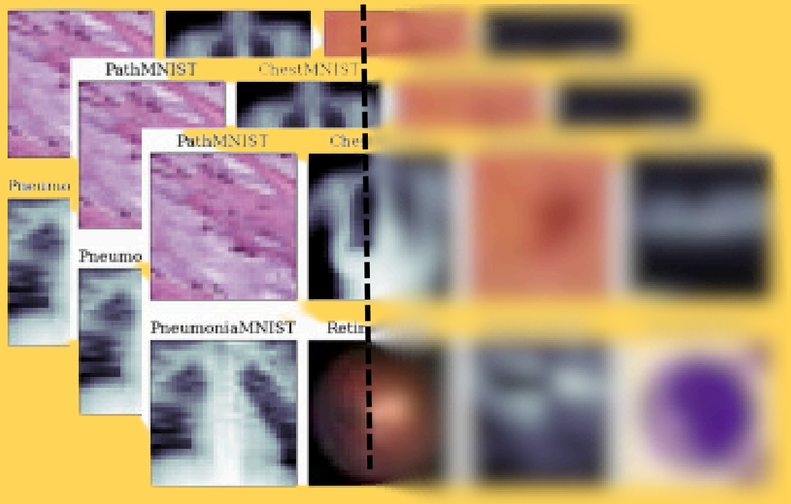


Open data marketplace



- 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 selection.

Collaborative data marketplaces

1

Model value of each seller datapoint



2

Collaborative data selection within budget



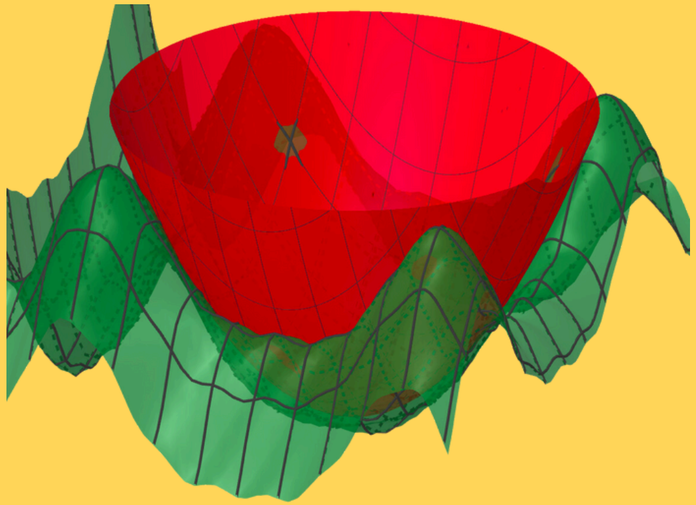
3

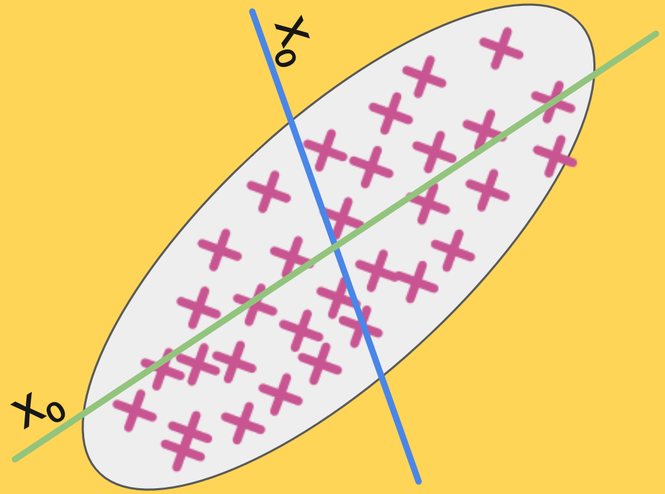
Train a model using CL on selected data



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

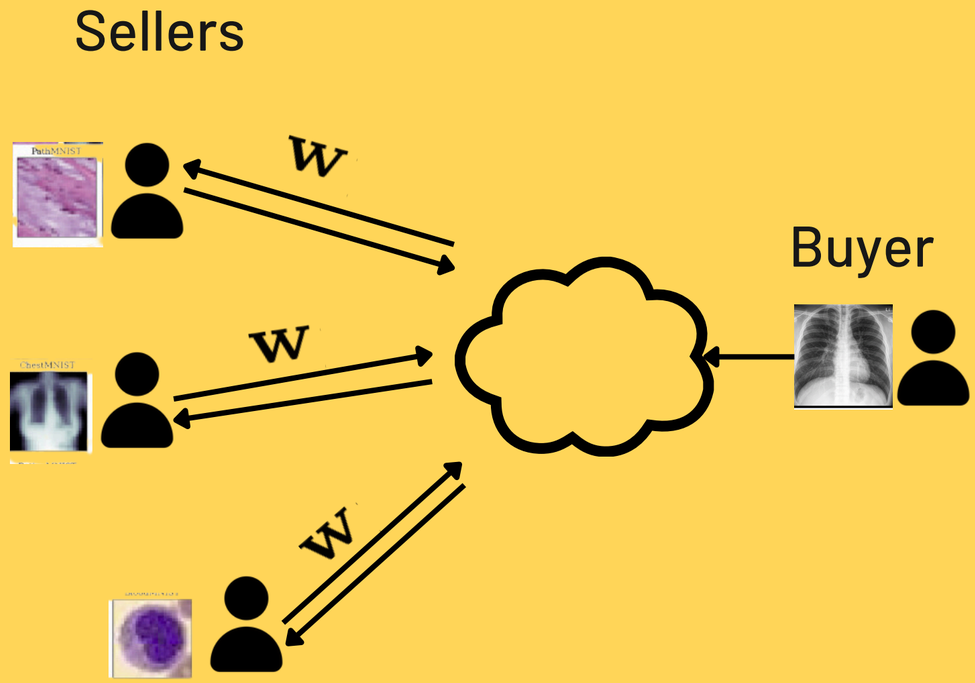
Iterative, collaborative algorithms

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

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

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

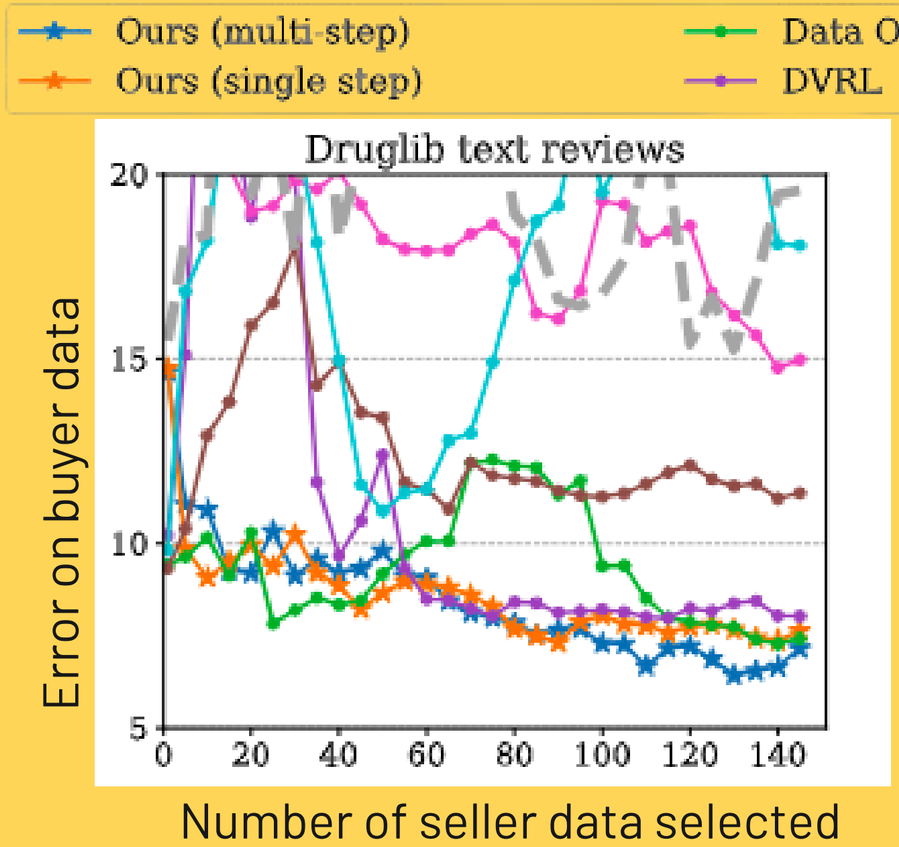
- NP Hard => O(1/B) convex approximation
- Using ***collaborative conditional gradient*** [Frank and Wolfe 1956]



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

Collaborative data discovery

Results

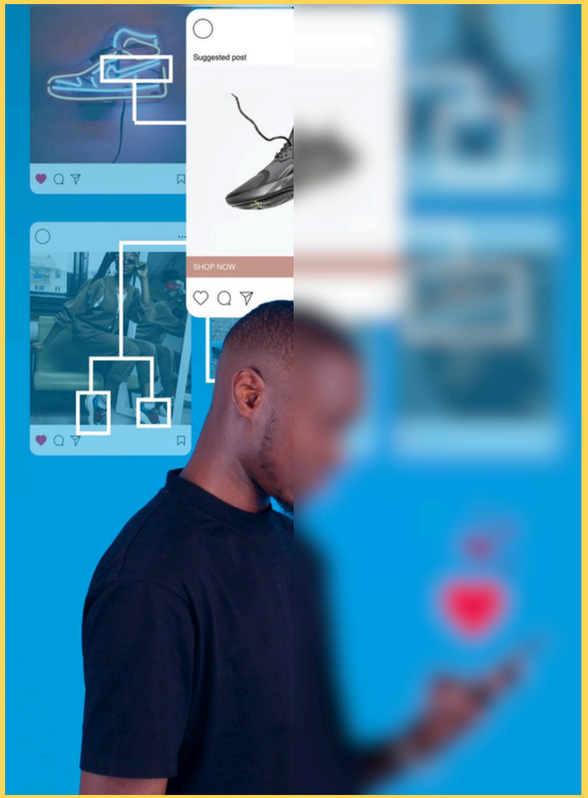


- Finetune GPT-2 on selected subset within budget (x-axis), while minimizing error (y-axis).
- Our collaborative selection methods (blue and orange) beat even centralized baselines.
- 100-10k times faster.

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

Aside: What is privacy?
An "econ" view of privacy

- A repeated game:
  - Agent is "vulnerable" (**V**) or "normal" (**N**).
  - Platform can exploit **V** and profit.
  - **V** can mask their data at some cost.
- Our results
  - *Revenue/welfare maximizing* privacy mechanisms.



Edited, orginial by Comuzi / © BBC / CC-BY 4.0

[ADWKJ FORC 24. *"Privacy Can Arise Endogenously in an Economic System with Learning Agents"*]

Takeaways

1. A universal open protocol for data needs collaborative data discovery.
2. Experiment-design lens provides effective data utility estimates.

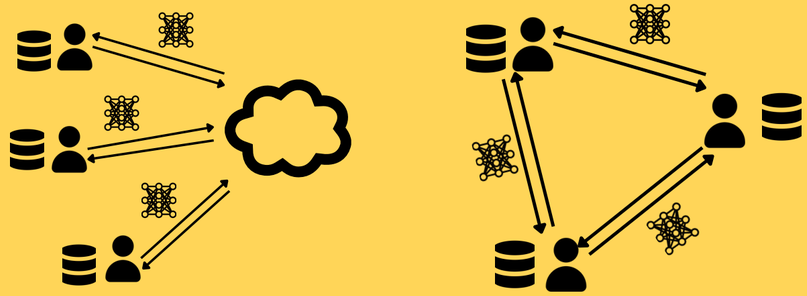
4

Conclusion



Data Decentralization

Just getting started



Loss of Control

- No privacy
- Data stored indefinitely
- **No control over usage** which may be unaligned with my interests.

Market Inefficiencies

- No sharing of profit => unsustainable data ecosystem

Better CL algorithms

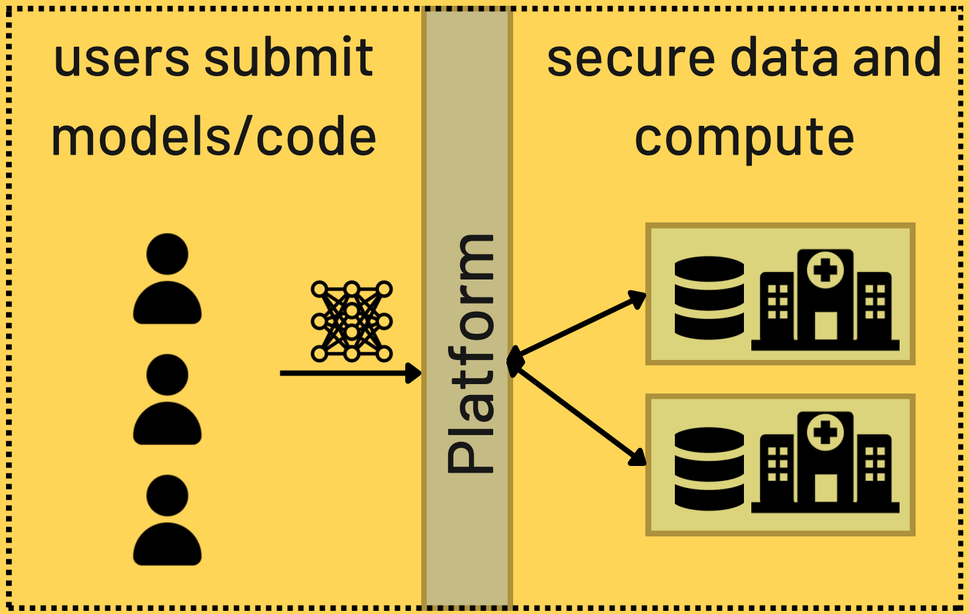
Future work

- Improved ML theory => improved CL algorithms
  - PAC learning
  - Stochastic optimization
  - NTK
  - Feature learning



Collaborative inference

Future work



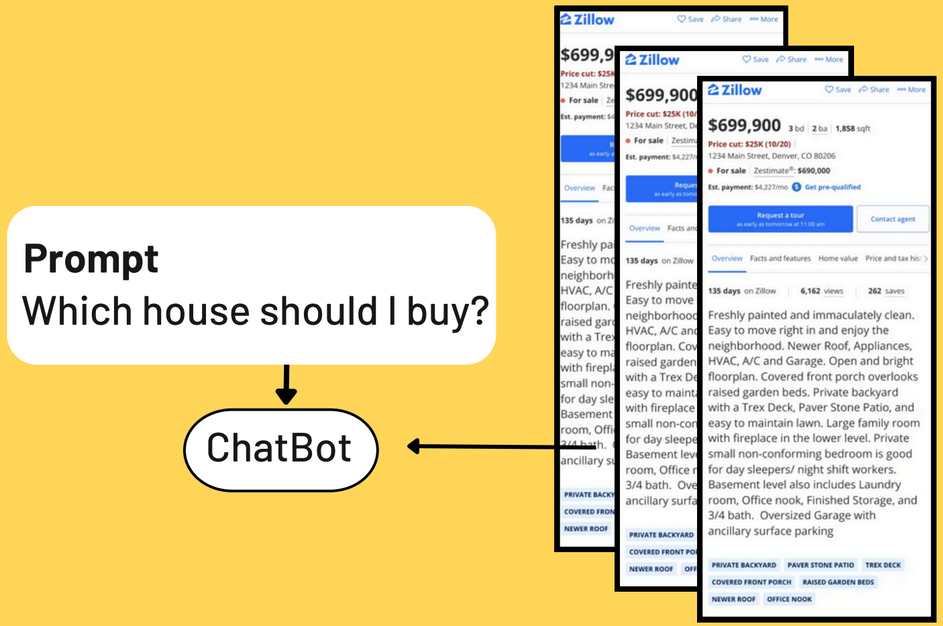
Collaborative testing and analytics
e.g. causal inference [GKJ 24]

[GKJ Arxiv 24. *"Collaborative Heterogenous Causal Inference beyond Meta Learning"*]

Better data understanding

Future work

1. A theory of data utility
   - beyond linear models, incentive compatible.
   - Value of particular data = reduction in uncertainity it provides. How to quanitfy uncertainity reduction? conformal prediction [ICML 23]



Zillow listings

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

[LY**K**VJR ICML 23. *"Federated conformal predictors for distributed uncertainty quantification"*]

Better collaboration mechanisms

Future work

- **Learn** and reconcile diverse user preferences
  - privacy safeguards?
  - risk apetites (e.g. cars in traffic)?
  - value systems and alignment?



Better collaboration mechanisms

Future work

Collaborate/negotiate while learning each other's objectives.

- What is the intention of the big triangle?
- What is the circle doing?



[Heider & Simmel '44]

Collaborative ~~Learning~~ Data Governance

Who decides how data impacts our lives?

- what is collected?
- what tradeoffs are made?
- what metrics are optimized?
- ...



Data “may” do a lot of good

“The insights we can get from this unprecedented access to data can be a great thing: we can get new understanding about how our society works, and improve public health, municipal services, and consumer products.”

- The Ethical Algorithm, Kearns & Roth

Building Collaborative Intelligence

User centric ML

1. Better Algorithms

Scale to massive networks

[**K**KM+ ICML 20], [**K**JK+ NeurIPS 21] , [YW**K**+ NeurIPS 22], ...

Resilient to data poisoning

[**K**HJ ICML 2021], [**K**HJ ICLR 2022], [H**K**J SPICY-FL 2022]

Uncertainty quantification

[LY**K**+ ICML 23]

2. Better Mechanisms

Theory of "good data"

[GH**K**J ***best paper*** FL-ICML 21], [WHJJ**K** TMLR 23], [C**K** TMLR 24],

Data markets

[LH**K**VJR Arxiv 24]

Collaboration

[C**K** TMLR 24],

Collaborative Learning

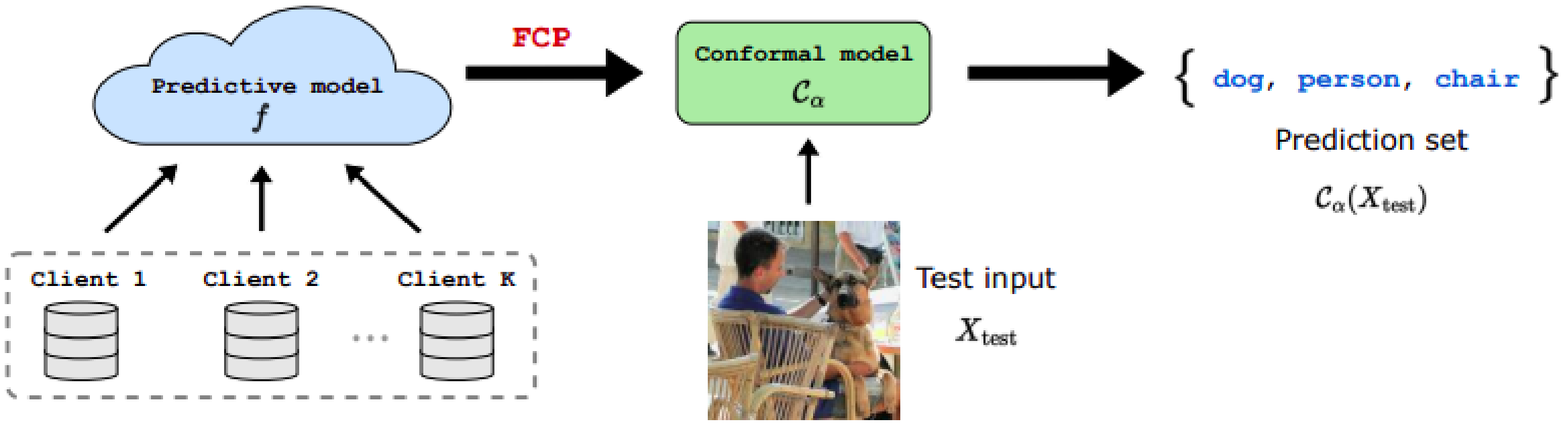
Opportunity to

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



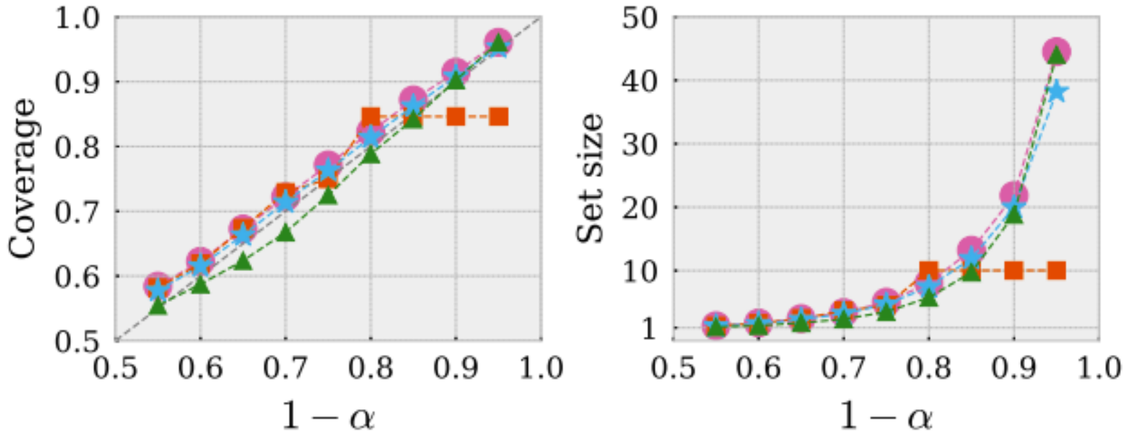
Thank you.

Uncertainty Estimation for Test-time



Even if training goes well, need uncertainty estimation.

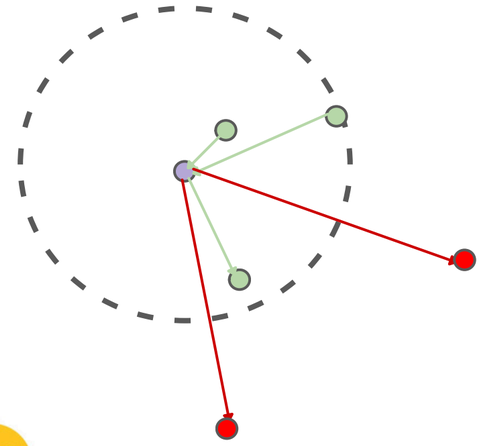
Use collaborative conformal prediction.

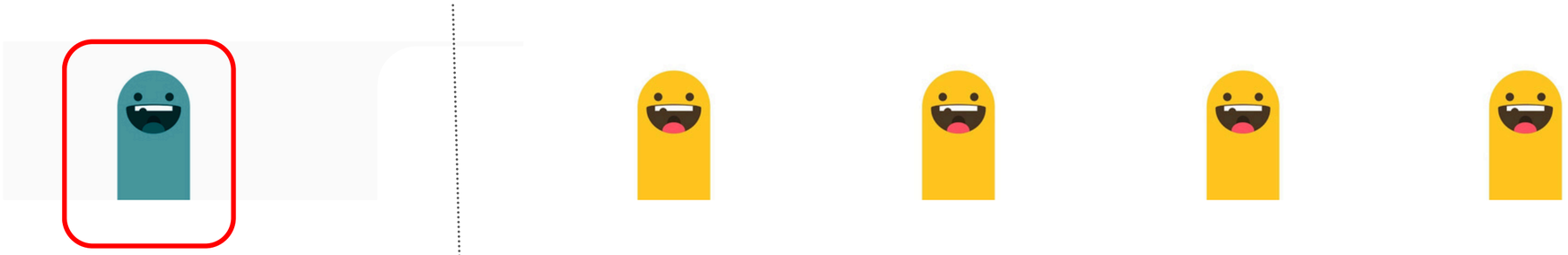


Lu, Yu, Karimireddy et al. [ICML 2023]

An optimization theory of dataset relevance

many users with different tasks.
which datasets are useful to me?





me: *min* $\boldsymbol{L_1(x)}$

other users: **Li(x)**

Karimireddy et al. [ICML 2021], [ICLR 2022]
[Werner, …, Karimireddy TMLR 2023]

[Werner, …, Karimireddy TMLR 2023]

An optimization theory of dataset relevance

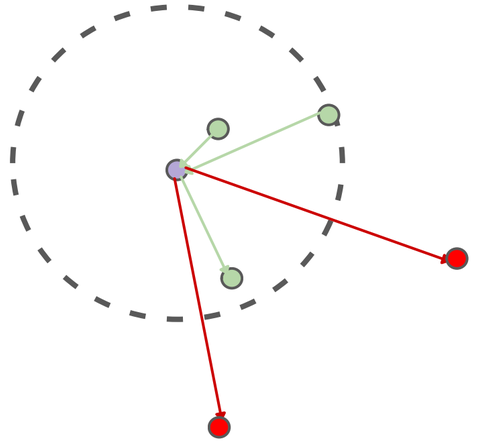
“Consistent” datasets belong together.

We develop a gradient based criterion

$$\nabla L_1(x_i^*) = 0$$ Data L*i* is inlier

$$\|\nabla L_1(x_j^*)\|^2 > 0$$ Data Lj is outlier

$$\Delta^2 = \min_j \|\nabla L_1(x_j^*)\|^2$$



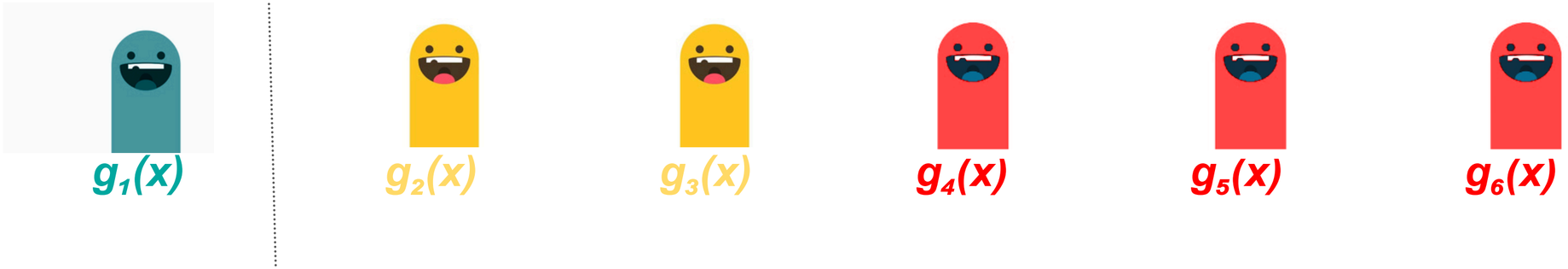
First optimization theoretic characterizing of data relevance!

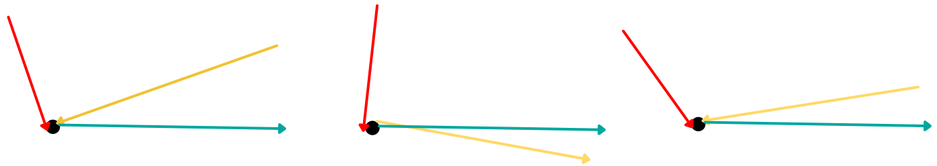
[Werner, …, Karimireddy TMLR 2023]

An optimization theory of dataset relevance

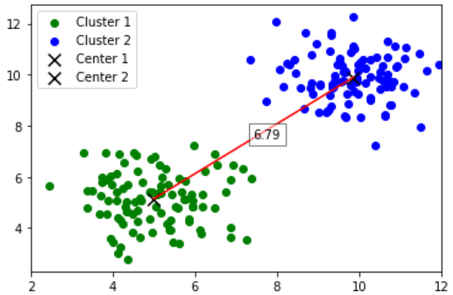
Cluster into datasets which are consistent with each other

using stochastic gradients.





compare gradients across time



cluster datasets

Distributed method
simultaneously clusters
+ trains models.

Federated clustered minimization

Many users, each with ***n*** stochastic gradients.

$\nabla L_1(x_i^*) = 0$ **Data L*i* is inlier**

$\|\nabla L_1(x_j^*)\|^2 > 0$ Data Lj is outlier

$\Delta^2 = \min_j \|\nabla L_1(x_j^*)\|^2$

Theorem. For a given user **i,** let **m** others share optima.

$$\frac{\sigma^2}{\mathbf{n}}\left(\frac{\mathbf{1}}{\mathbf{m}} + \frac{\sigma^2}{\mathbf{n\Delta^2}}\right) \lesssim \textbf{Error on } \mathbf{L_i} \lesssim \frac{\sigma^2}{\mathbf{n}}\left(\frac{\mathbf{1}}{\mathbf{m}} + \frac{\sigma}{\mathbf{\Delta}}\right)$$

[Werner, …, Karimireddy TMLR 2023]

First theoretical justification for gradient comparison methods in multi-task learning.

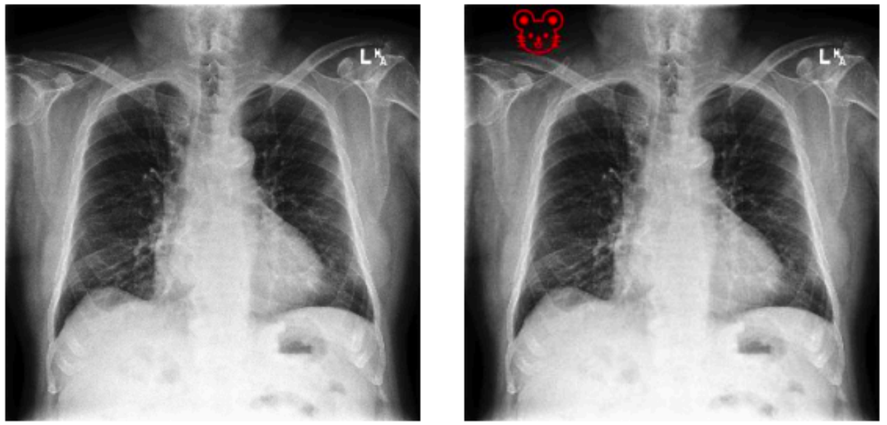
Application: Distributed dataset understanding

Understand how is my data
different from rest
(debug biased/bad datasets)

without looking at others data.

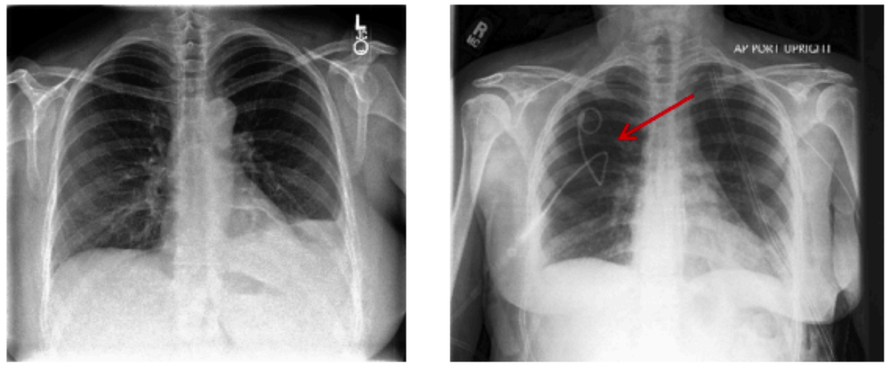
Cardiomegaly classification

marker



(a)

pleural effusion



(b)

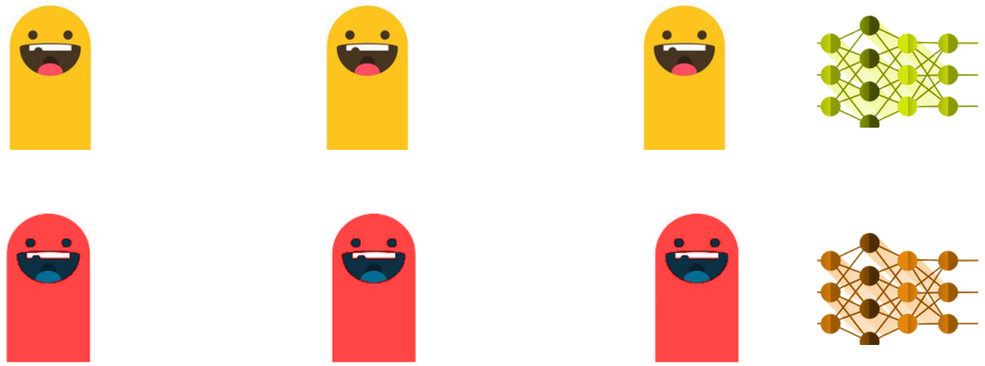
normal

biased

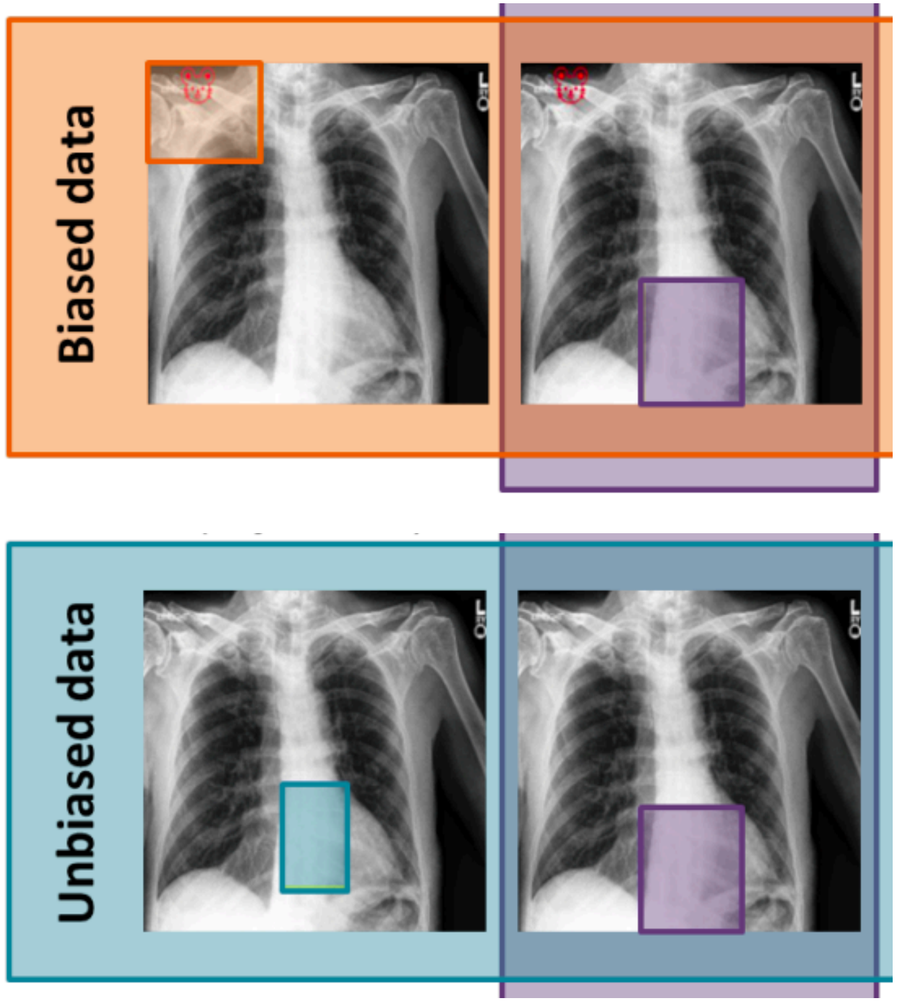
[Hartley, .., Karimireddy et al. 2024]

Application: Distributed dataset understanding

Idea: Run clustered training of interpretable models (ProtoPNet).

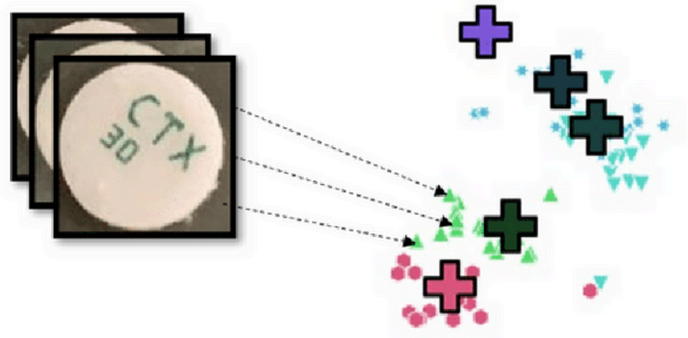


explanations differ = datasets different



[Hartley, .., Karimireddy et al. 2024]

Applications
antimicrobial resistance

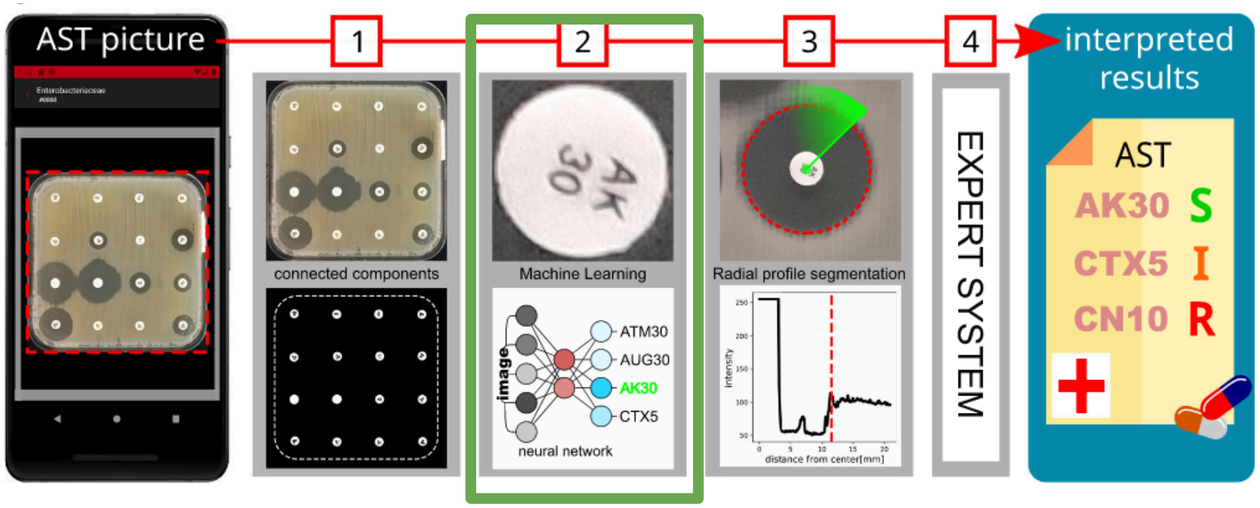


MEDECINS SANS FRONTIERES

Antibiogo

Use on-mobile ML to automate antimicrobial resistance testing.

1,2. identify antibiotics
3. measure “region of inhibition”
4. plug into an expert-system



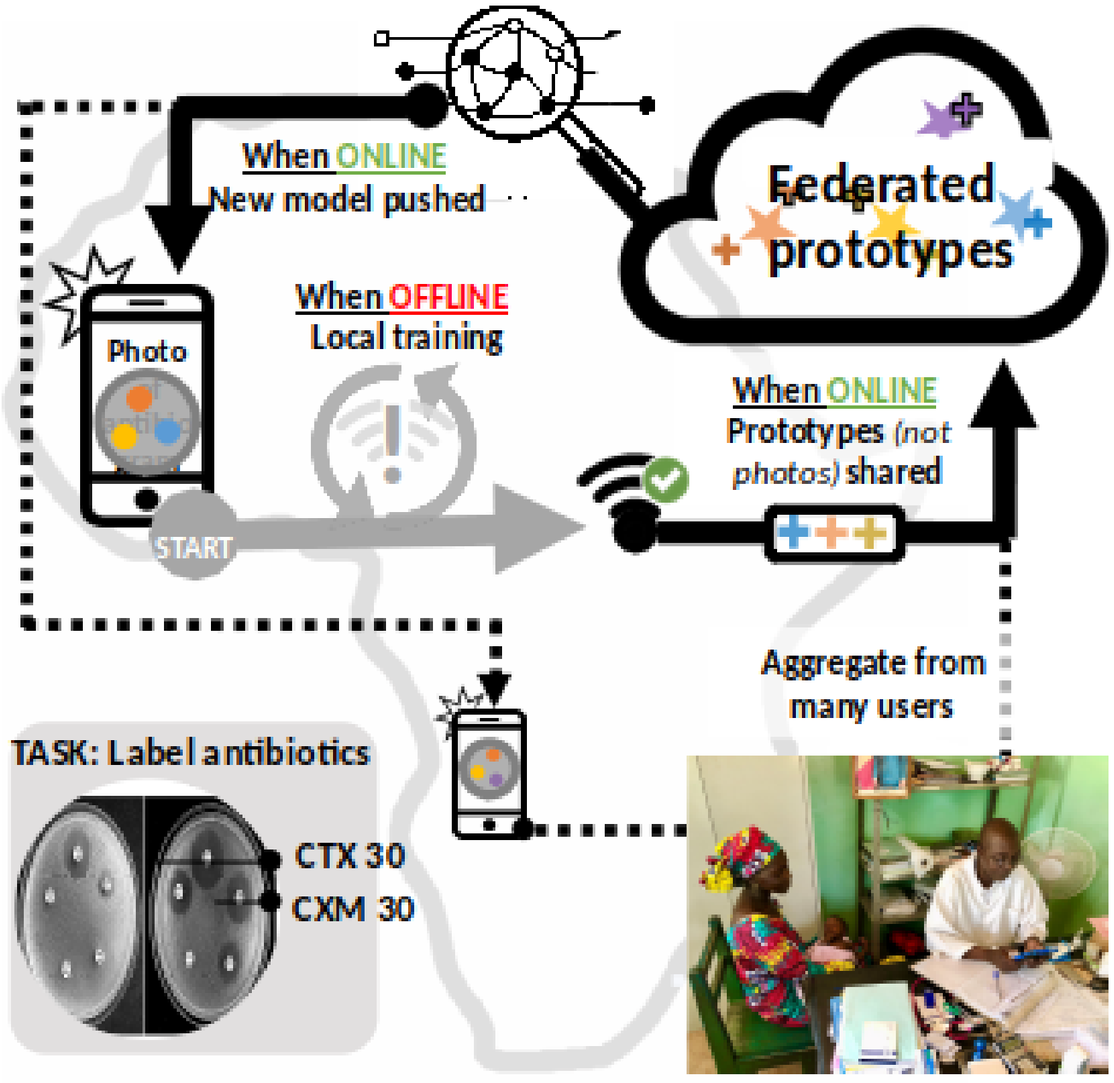


MEDECINS SANS FRONTIERES

Antibiogo

MEDECINS SANS FRONTIERES

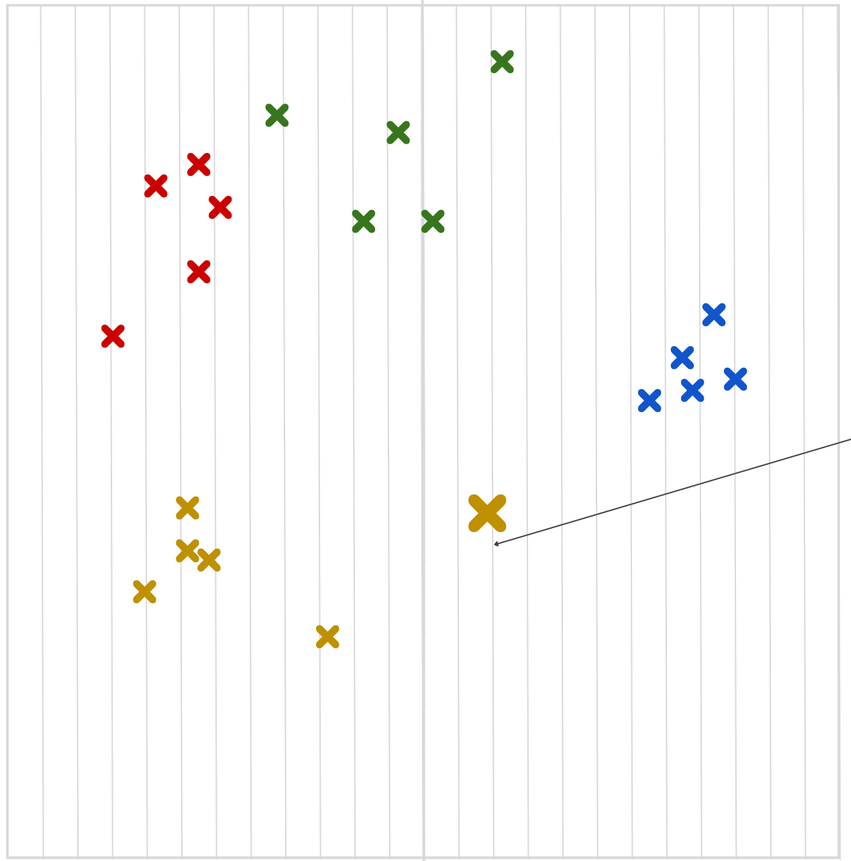
Mobile app to diagnose antimicrobial resistance,
continually improves using
distributed collaborative learning.

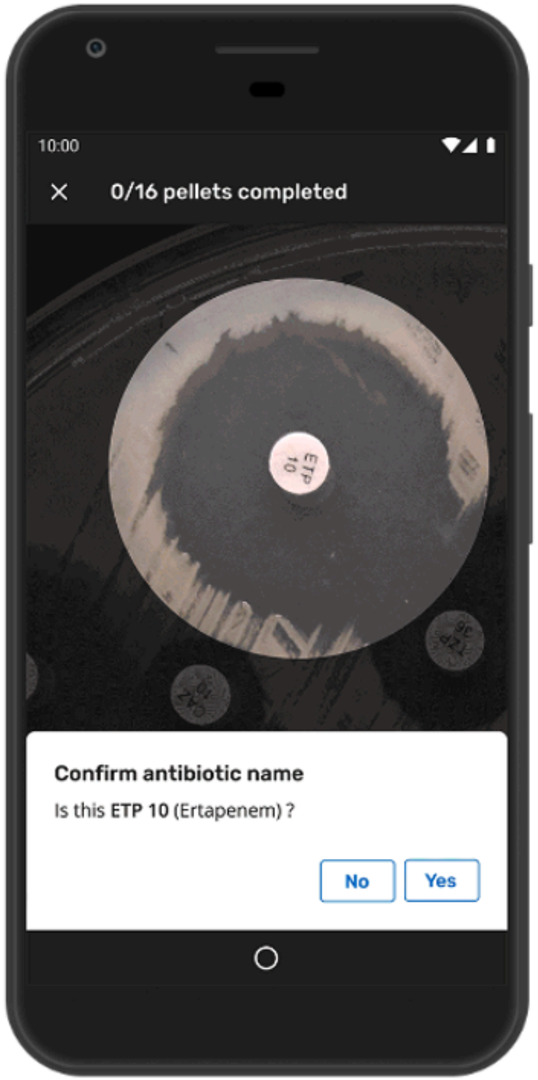


[submitting to Nature Comm.]

MEDECINS SANS FRONTIERES

Component 1: Updating Prototypes

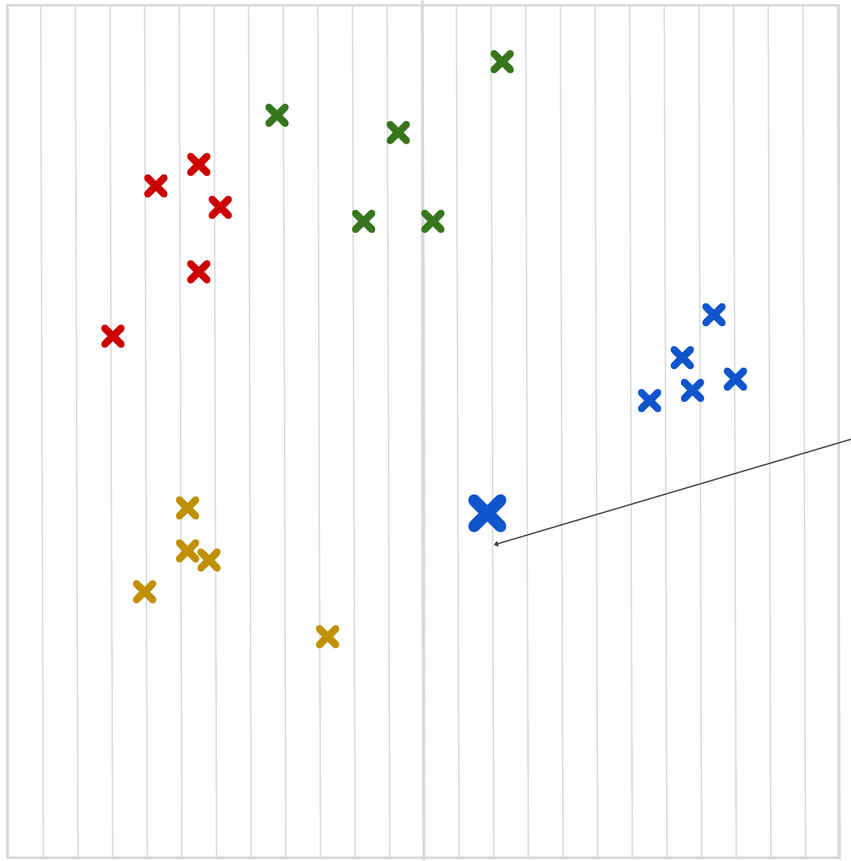




Yes.

New data point

Component 1: Updating Prototypes

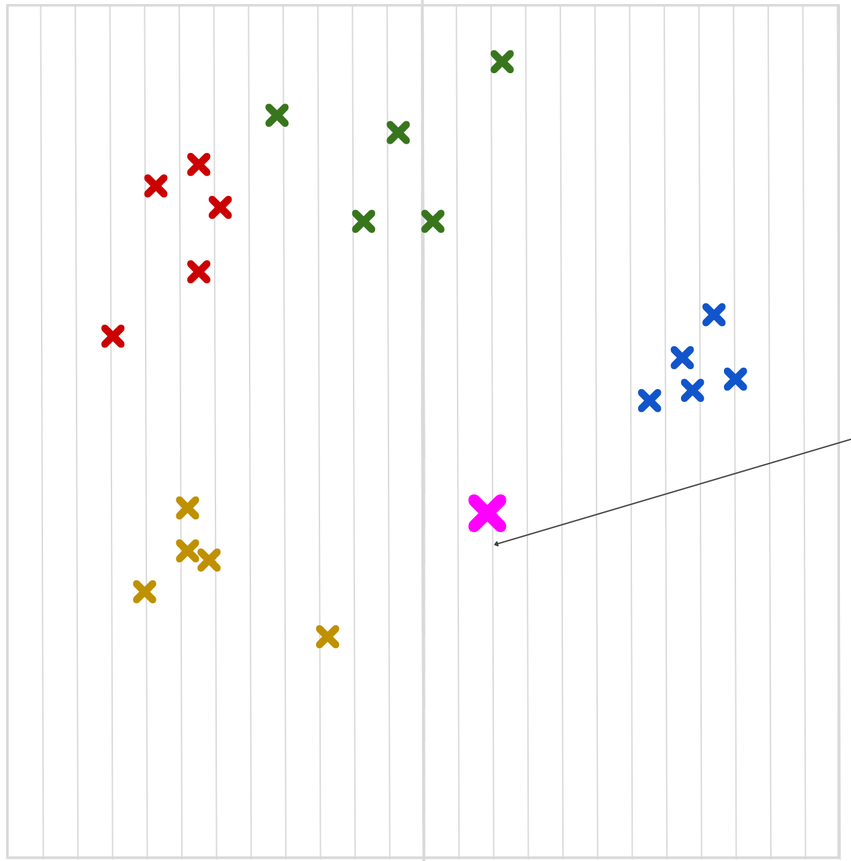




No. Existing class.

Also, new data point.

Component 1: Updating Prototypes

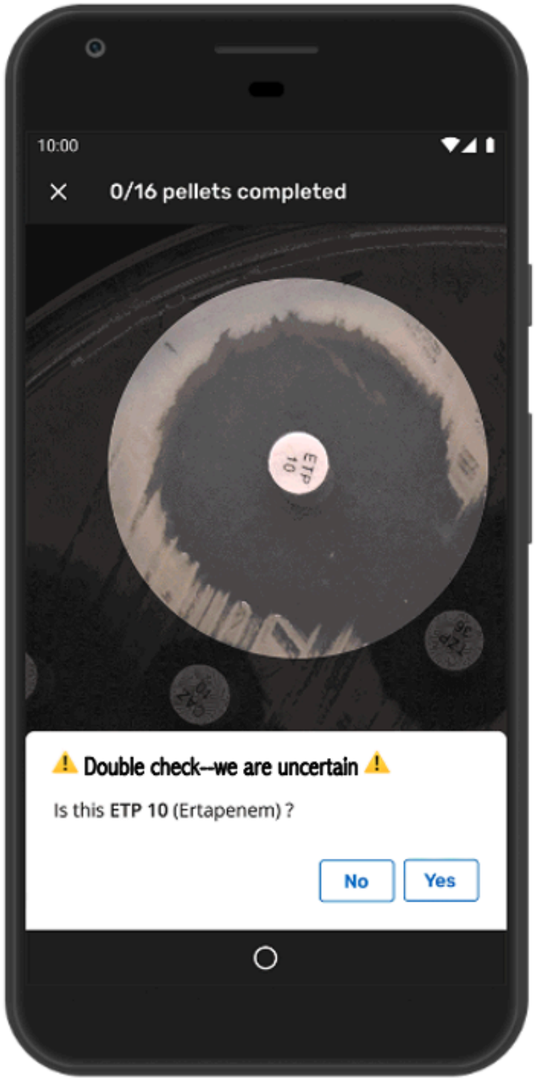




No. New Class.

New prototype!

Component 3: Uncertainty Estimation



- Ability to detect faulty images

Component 3: Uncertainty Estimation



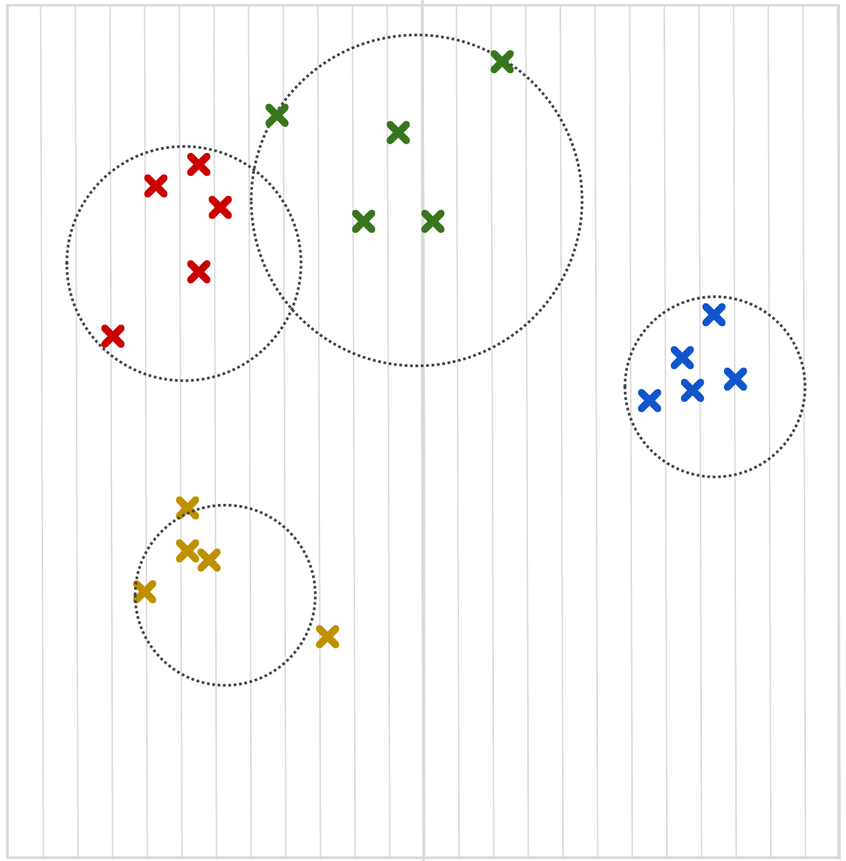
- Ability to detect faulty/bad images
- Give a list of “most likely” candidates

Component 3: Uncertainty Estimation



- Ability to detect faulty/bad images
- Give a list of “most likely” candidates
- Detect unknown classes

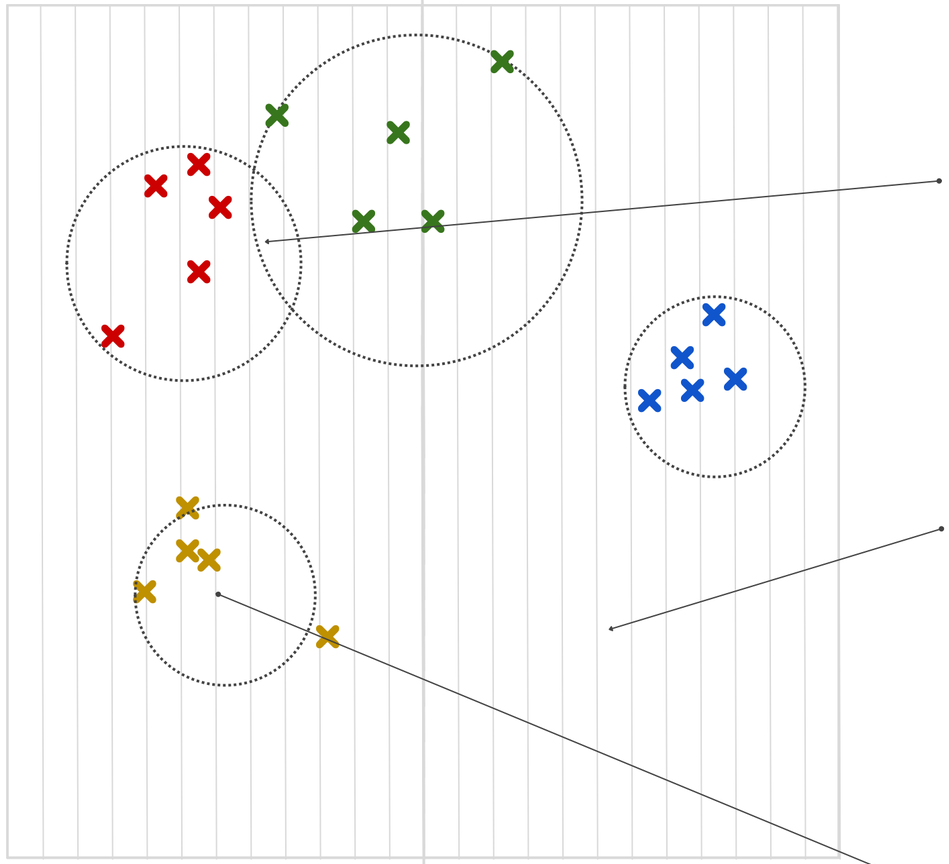
Component 3: Uncertainty Estimation



Idea: Use conformal prediction

Compute 98% cut-off distances to prototype.

Component 3: Uncertainty Estimation

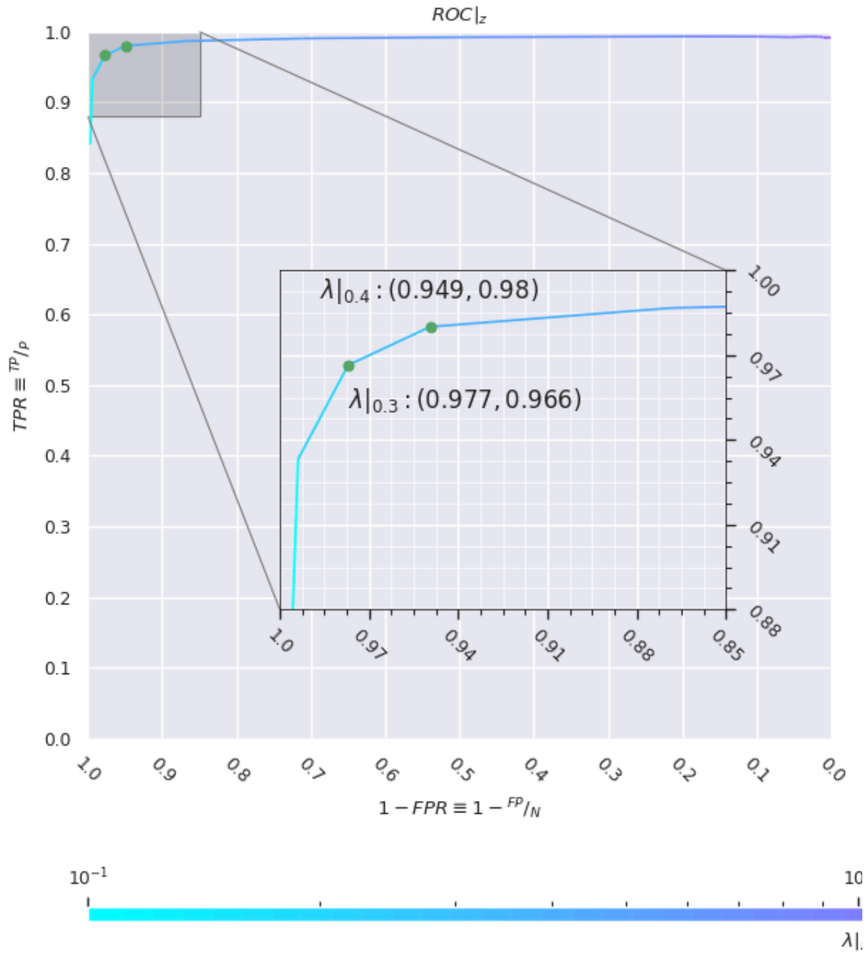


On user request:

- multiple possible prototypes, ask user to choose
- New prototype is created after asking the user
- Assign to ‘yellow’ prototype and adjust centroid

Antibiogo

Results



Sensitivity .98
Specificity .95

Filters out .77 of OOD test inputs.

MEDECINS SANS FRONTIERES