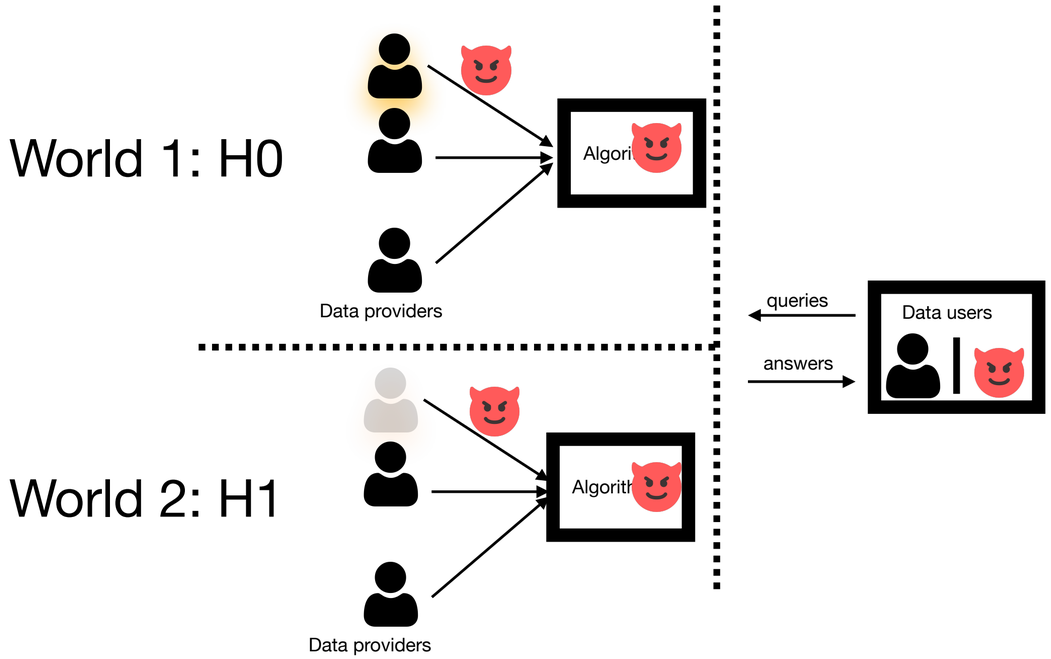
CSCI 699: Privacy Preserving Machine Learning - Week 10

Federated Learning

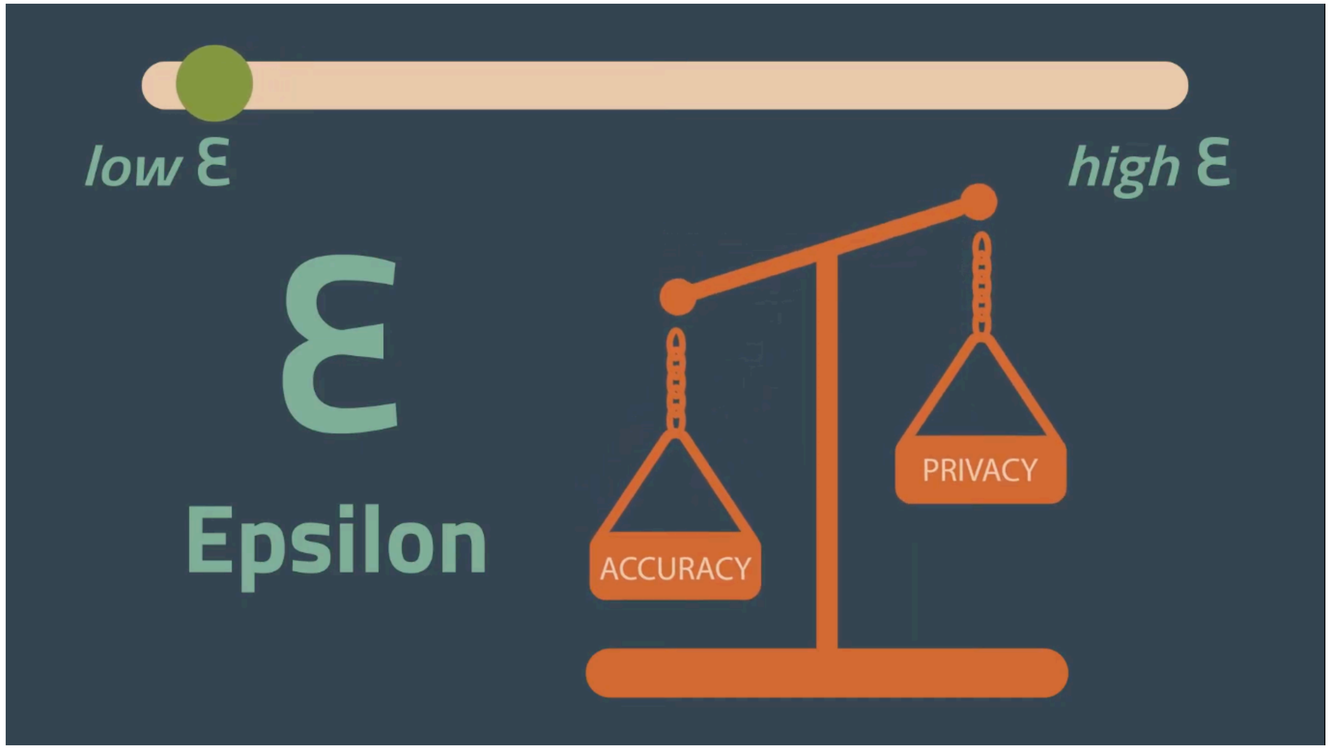
Sai Praneeth Karimireddy, Nov 1 2024. Slides partly from FL tutorial NeurIPS 20.

Recap: Local Differential Privacy

- New: how well can the adversary guess which world I am in by looking at my communication
- Ok if central server gets hacked etc. Only trust yourself.
- But needs a lot of noise. Can we get guarantees similar to local DP, but not add as much noise?

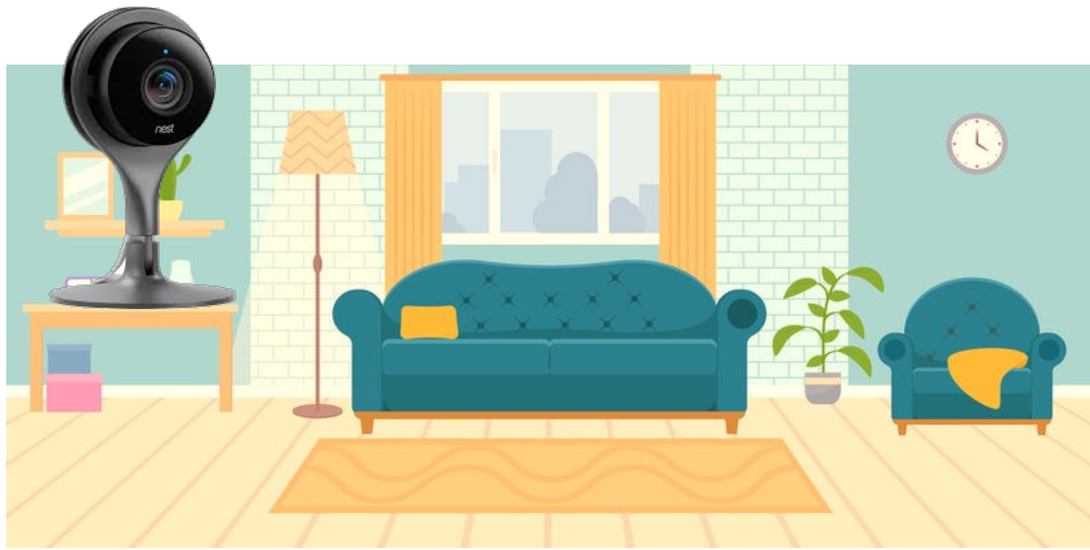


Federated Learning



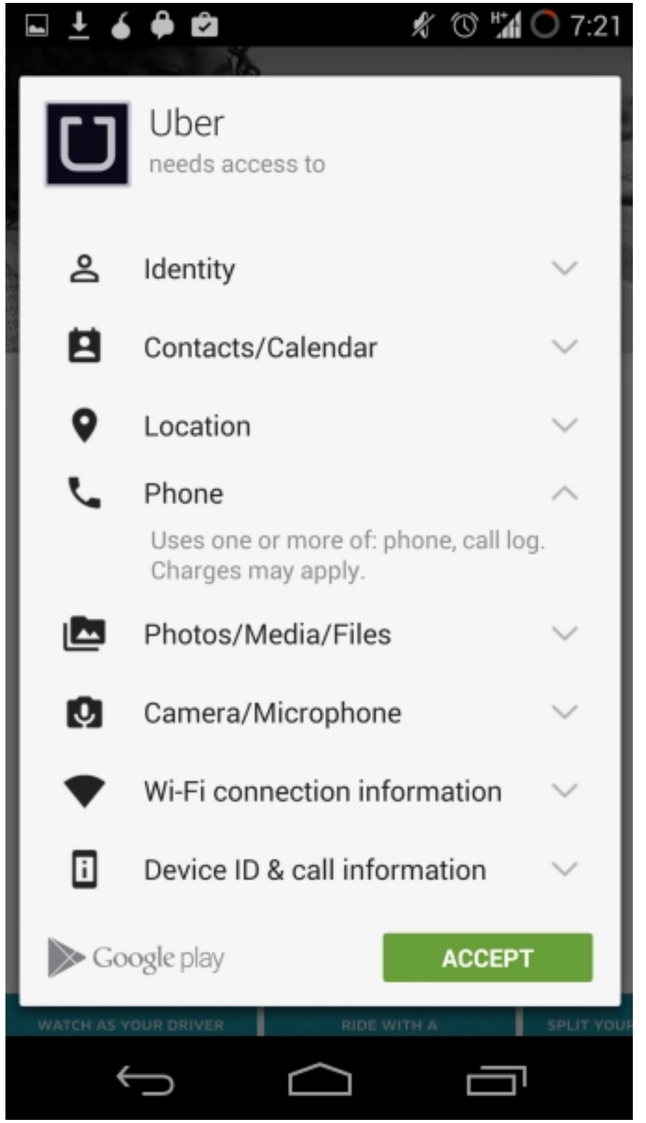
Data minimization

- Roommate and I agree to get smart light.
- Roommate installs a web camera + cloud ML model.
- I am uncomfortable.



Data minimization

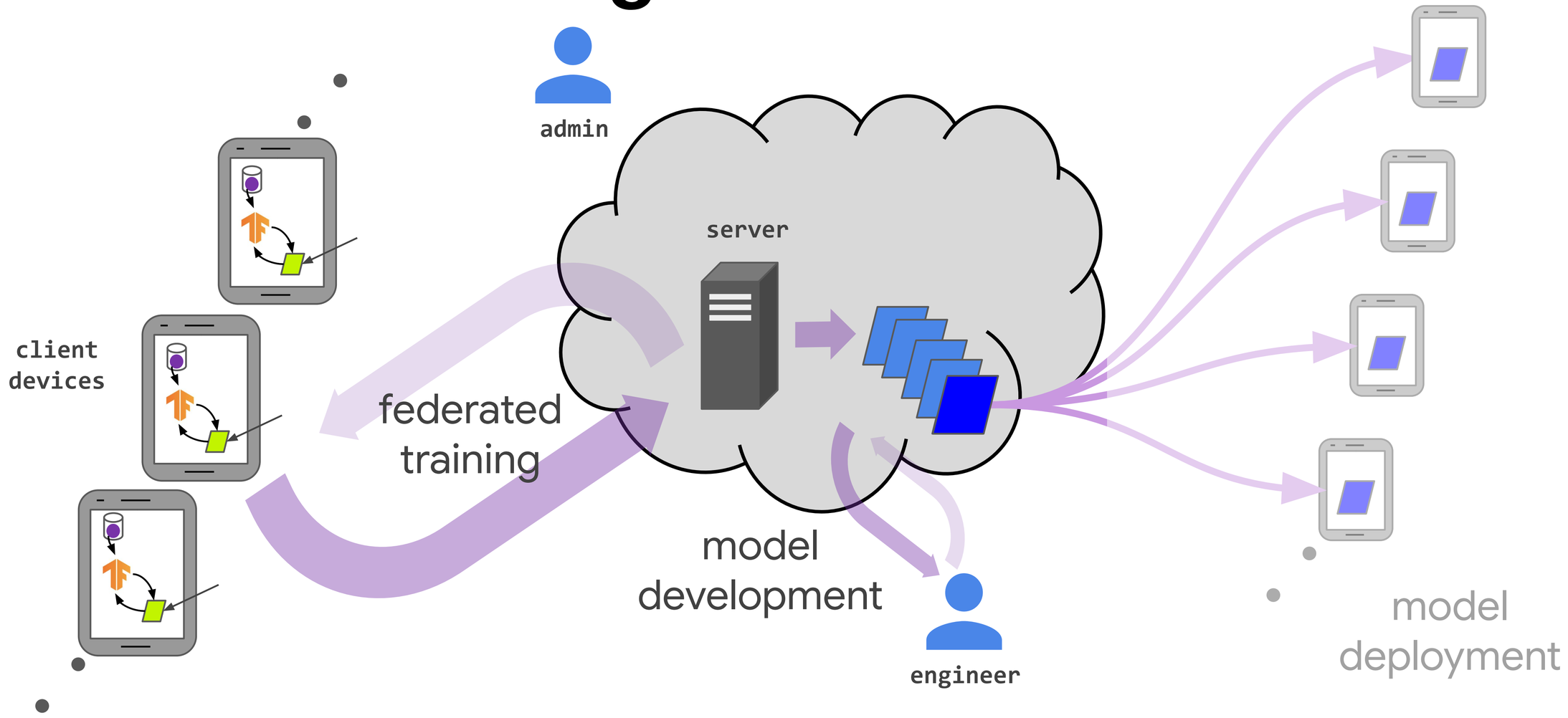
- Currently, ML follows the webcam model.
- This is bad
    - Servers can and do get hacked.
    - No control over privacy
    - No ownership over data
- Alternative?



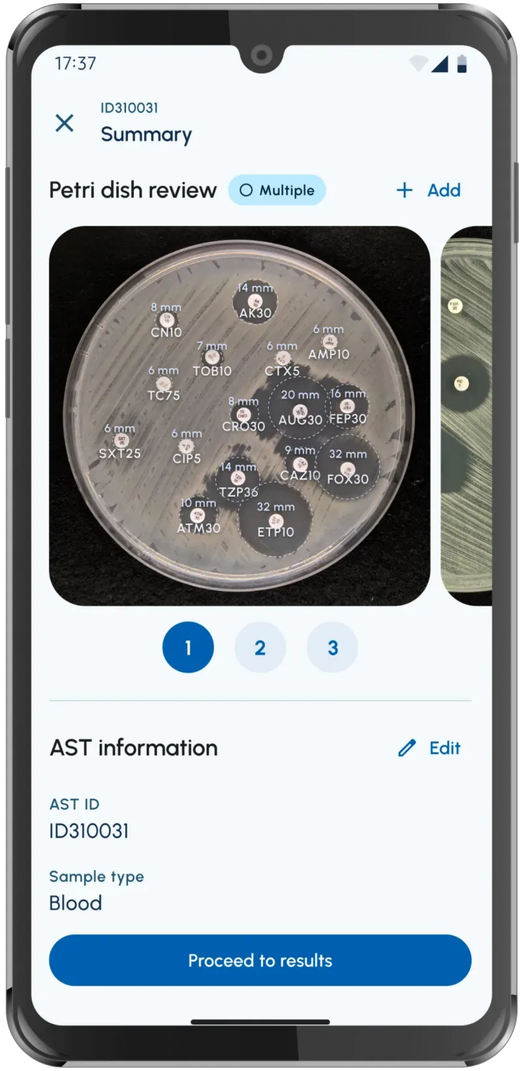
Data minimization

- Use a motion detector instead.
- ***Data minimization:***
  - Tie consent to a specific use case.
  - Collect only the data absolutely necessary.
  - Store locally as much as necessary.

Federated Learning



Case study 1:
Fighting Superbugs

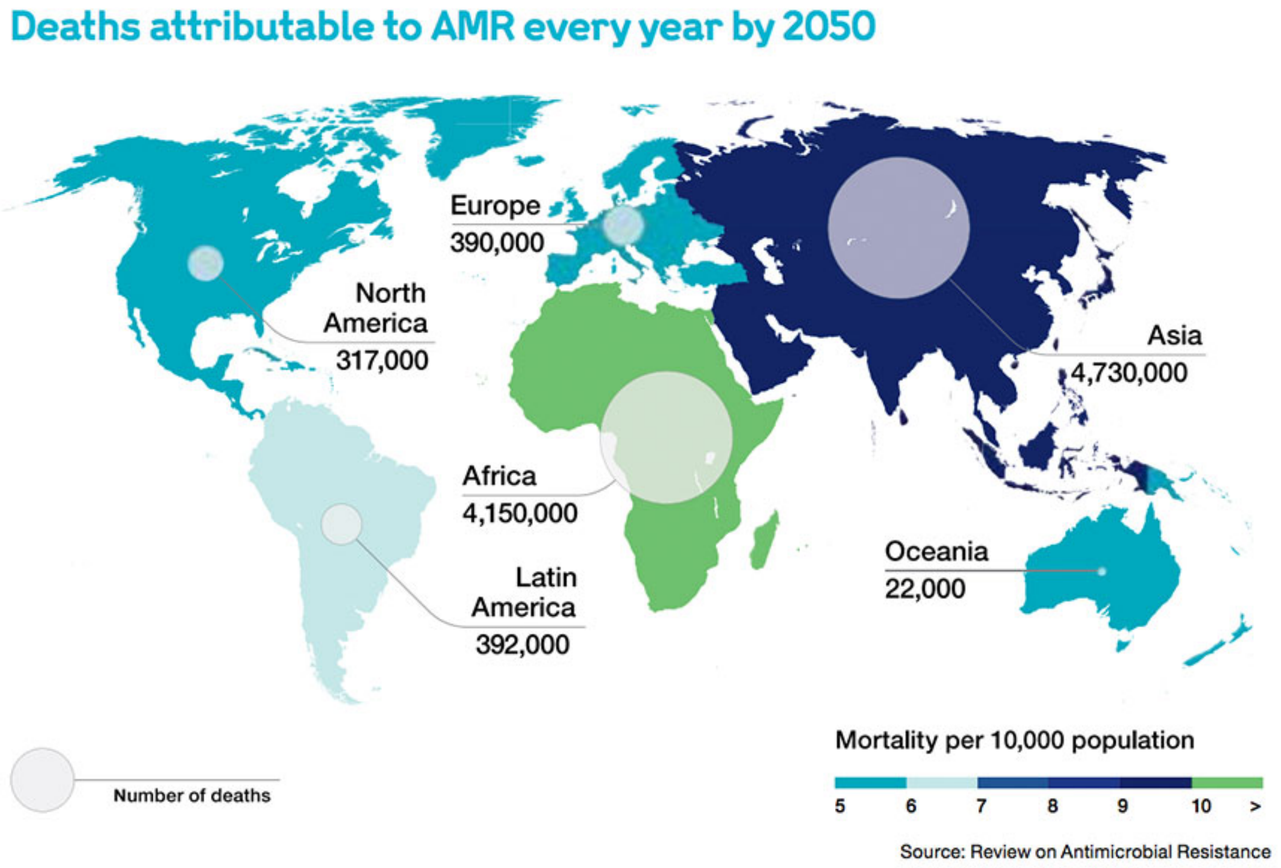


MEDECINS SANS FRONTIERES

Antimicrobial Resistance

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

- [WHO 2014]



MEDECINS SANS FRONTIERES

Antimicrobial Resistance

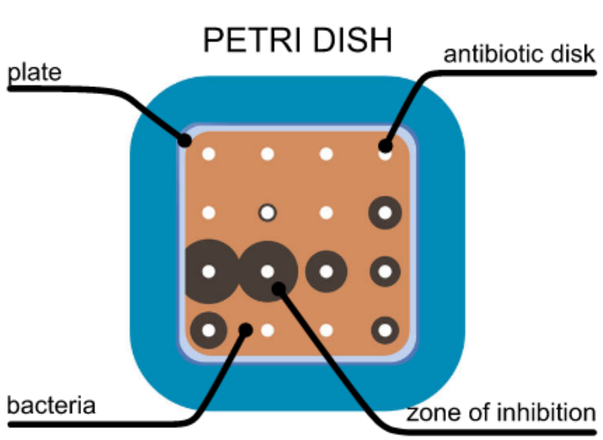
Requires pathology experts :(

MEDECINS SANS FRONTIERES

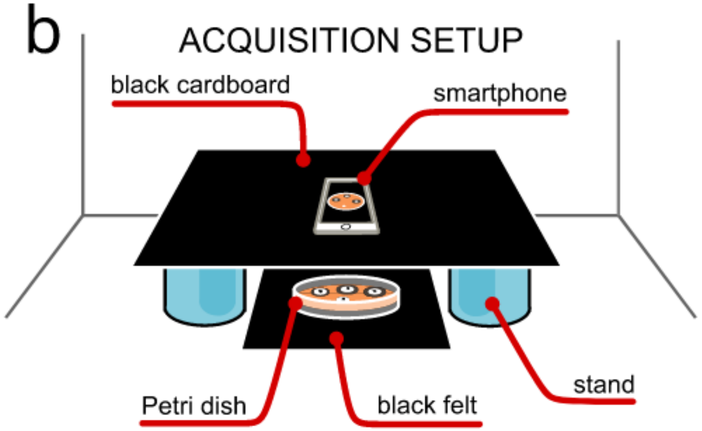
Antibiogo

Use on-mobile ML to automate the process.

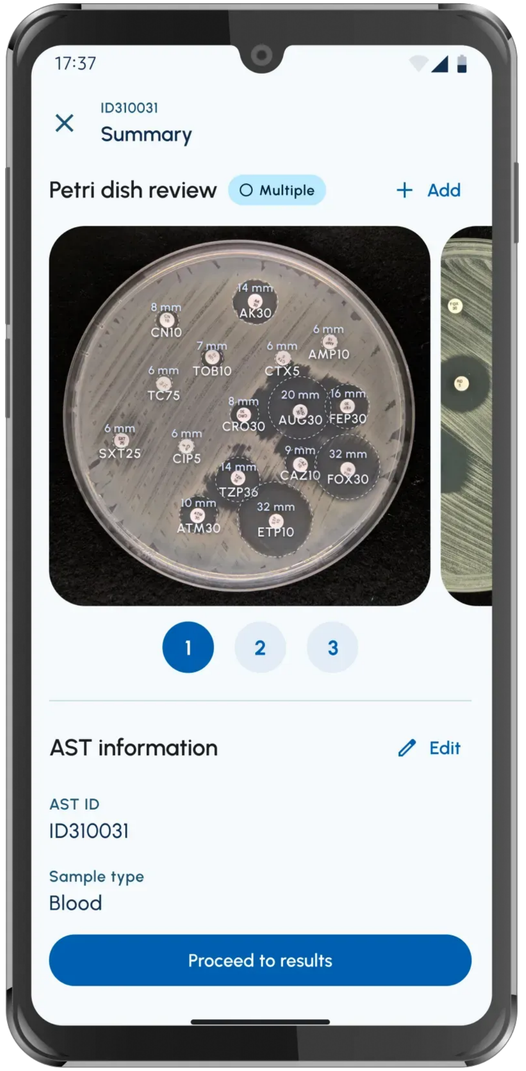
a



b



c

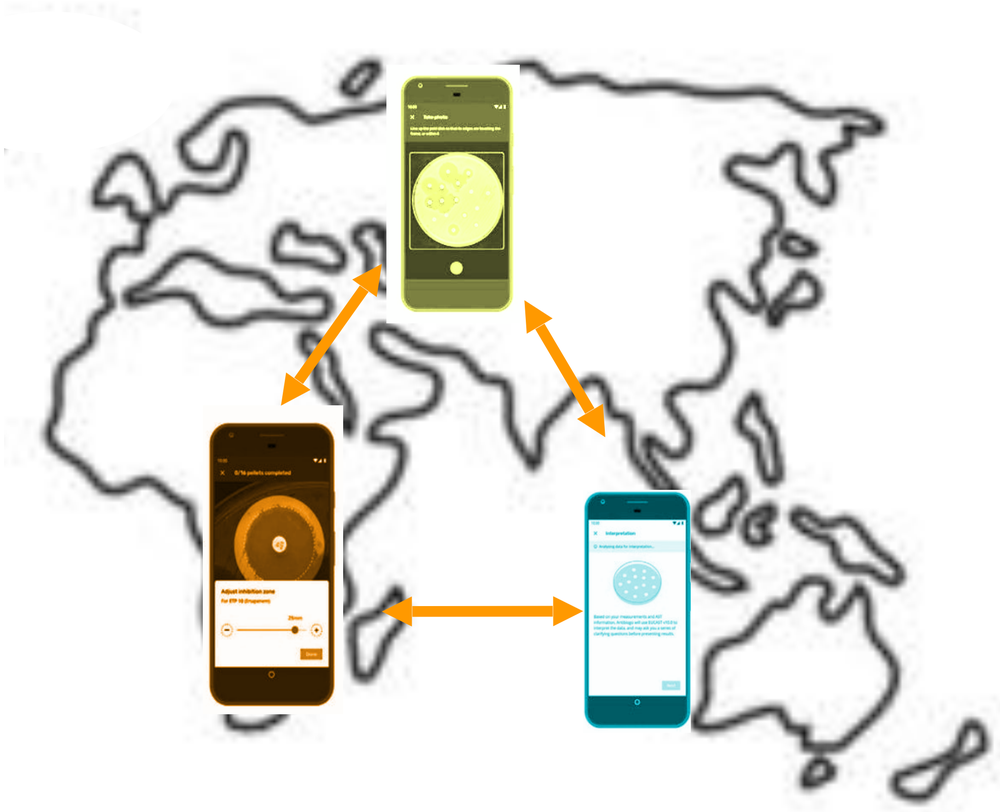


MEDECINS SANS FRONTIERES

Antibiogo Goal

- Multi-continent collaboration
  15 countries in Africa and Asia

- Continuously in real-time
  - improve ML models,
  - epidemiological monitoring



MEDECINS SANS FRONTIERES

Antibiogo Challenges

Continual learning across 15 countries, but

1. **Privacy** data can't leave the device.

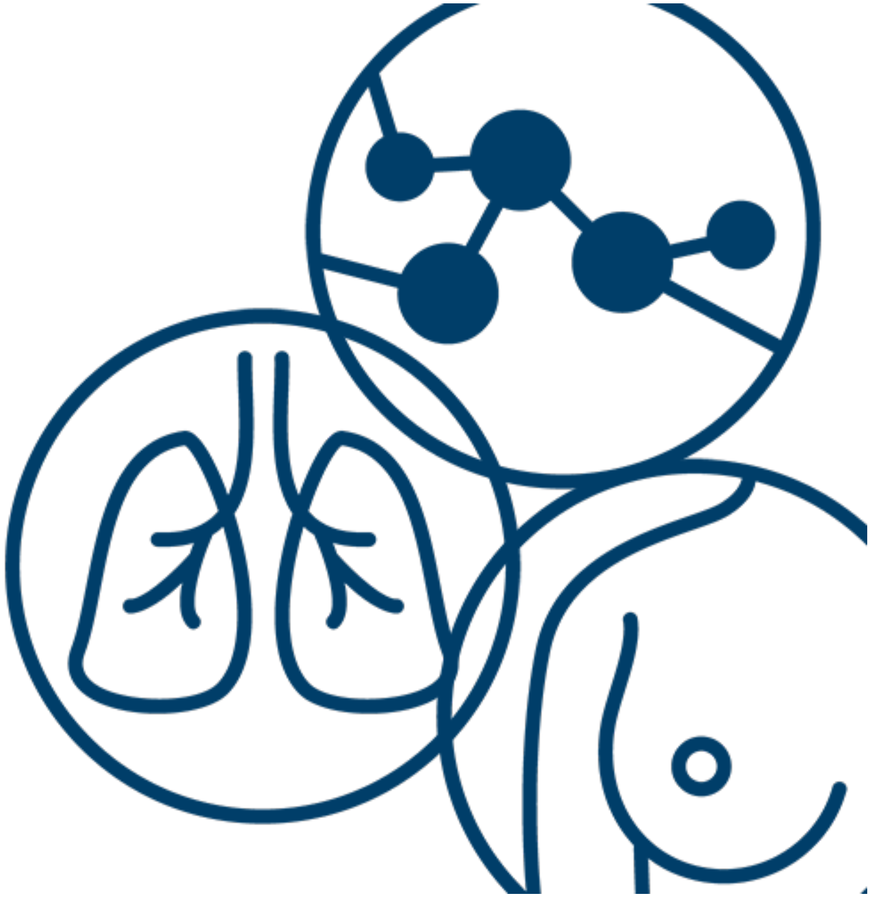
Resilience

2. Bad network, low end phones.

3. Very noisy data

MEDECINS SANS FRONTIERES

Case study 2:
Fighting Cancer



Cancer Registry of Norway

Shortages in Cancer Expertise

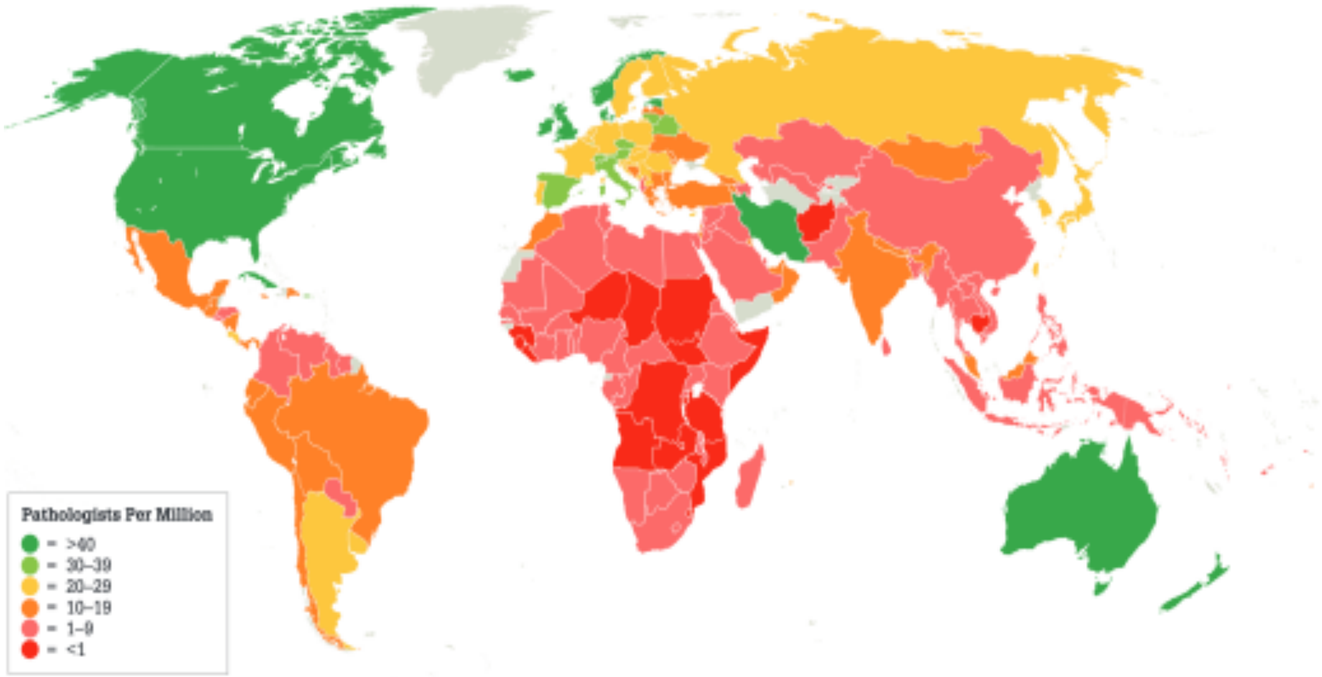
Global expertise per capita is falling, leading to deadly shortages.

"Cancer patients face life-threatening delays due to lack of staff, say UK radiologists."

— The Guardian, 8 Jun 2023

Cancer patients face worsening treatment delays due to lack of staff, report finds

— Sky News, 8 June 2023



Cancer Registry of Norway

Hanna, Timothy P., et al. "Mortality due to cancer treatment delay: systematic review and meta-analysis." bmj 371 (2020).

AI for Cancer

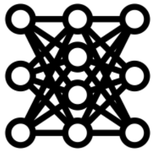
Sensitive patient data

scans

tests

genome

ML models



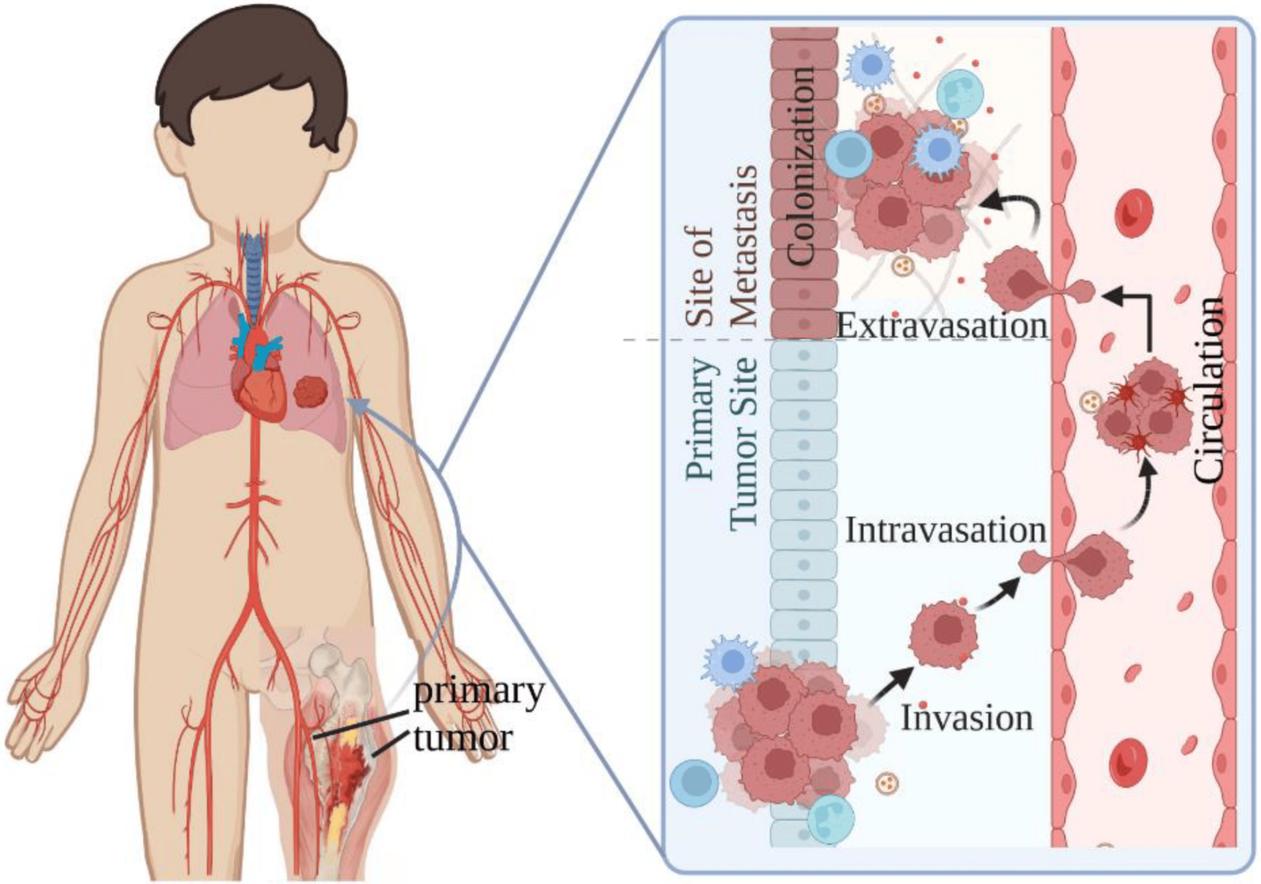
Tasks
Classify the cancer grade/stages
Segment the cancer region in radiographs
Predict the risk of developing a cancer type

Cancer Registry of Norway

Data Scarcity

Sarcoma is 1% of cancer diagnosis.
In 2022, 562 cases in Norway.

Same problem in low-middle income countries.



Cancer Registry of Norway

Data Scarcity

Ethics and Justice, Healthcare, Machine Learning

The Geographic Bias in Medical AI Tools

Patient data from just three states trains most AI diagnostic tools.

Sep 21, 2020 | Shana Lynch

Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review

Lisa N. Guo, MD • Michelle S. Lee, BA • Bina Kassamali, BA • Carol Mita, MSLIS • Vinod E. Nambudiri, MD, MBA

FDA NEWS RELEASE

FDA Takes Important Steps to Increase Racial and Ethnic Diversity in Clinical Trials

Agency's Focus on Inclusion in Trials for All Medical Products Aligns with Biden Administration's Cancer Moonshot Goal of Addressing Inequities and Beyond

Share | Post | Linkedin | Email | Print

For Immediate Release: April 13, 2022

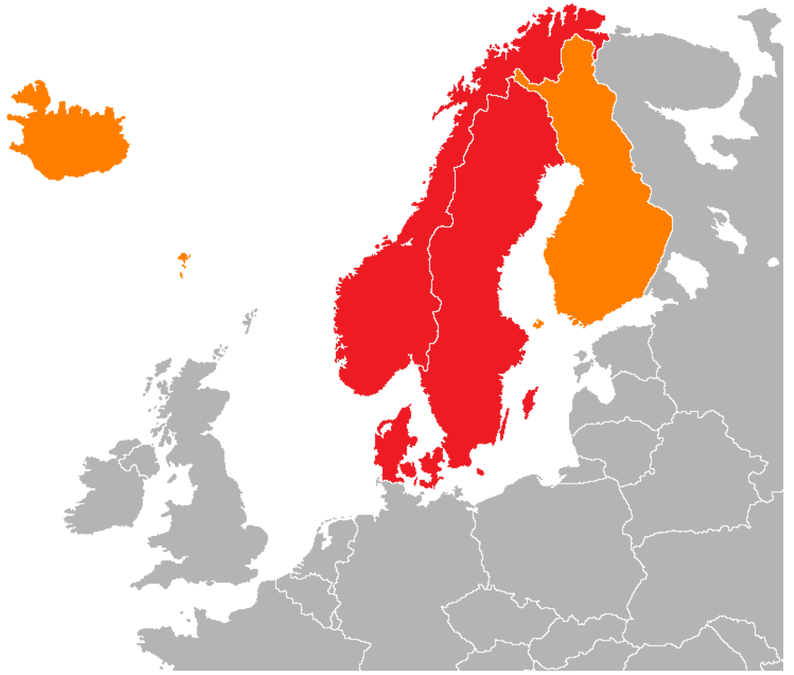
Cancer Registry of Norway

NORDCAN
Association of the Nordic Cancer Registries

NORDIC CANCER UNION

Position paper [Karimireddy et al. FMEC 2023]

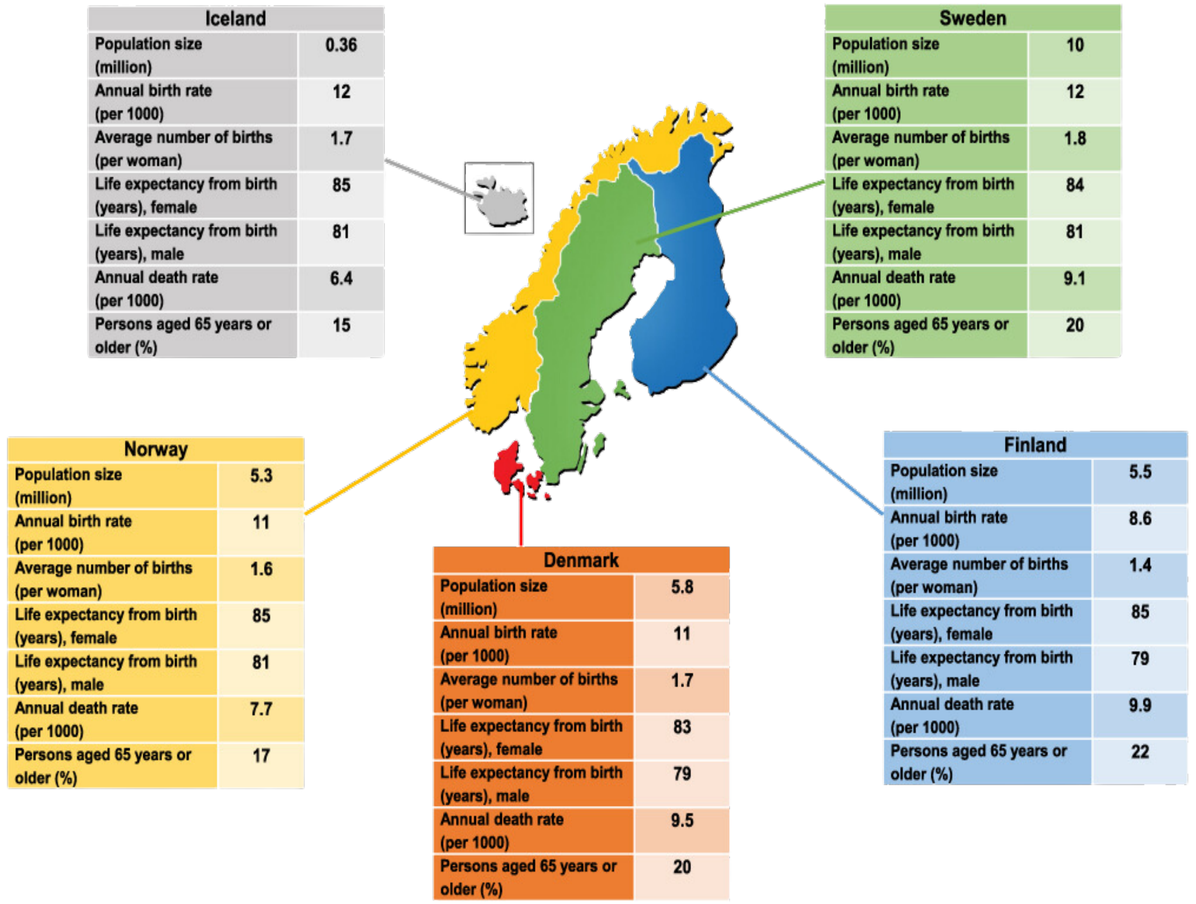
- 7+ countries, 21 registries collaboration to train ML models, monitoring.
- Strict **privacy** regulations to share data, especially genomic.
- But data collected is extremely **heterogeneous**.
- Also, **strategic** concerns like fairness and accountability.



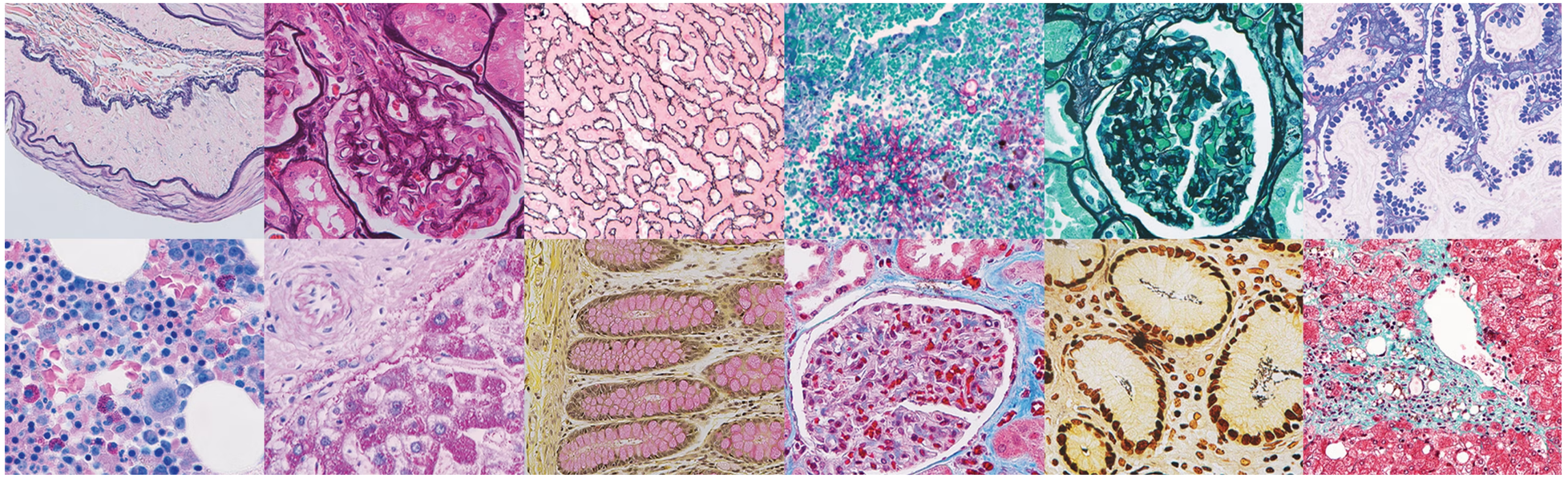
The Danish Cancer Registry
Finnish Cancer Registry
Icelandic Cancer Registry
The Cancer Registry of Norway
The Swedish Cancer Registry
The Faroe Islands Cancer Registry
The National Board of Health of Greenland

Heterogenous data

- Each site will have different demographics and populations



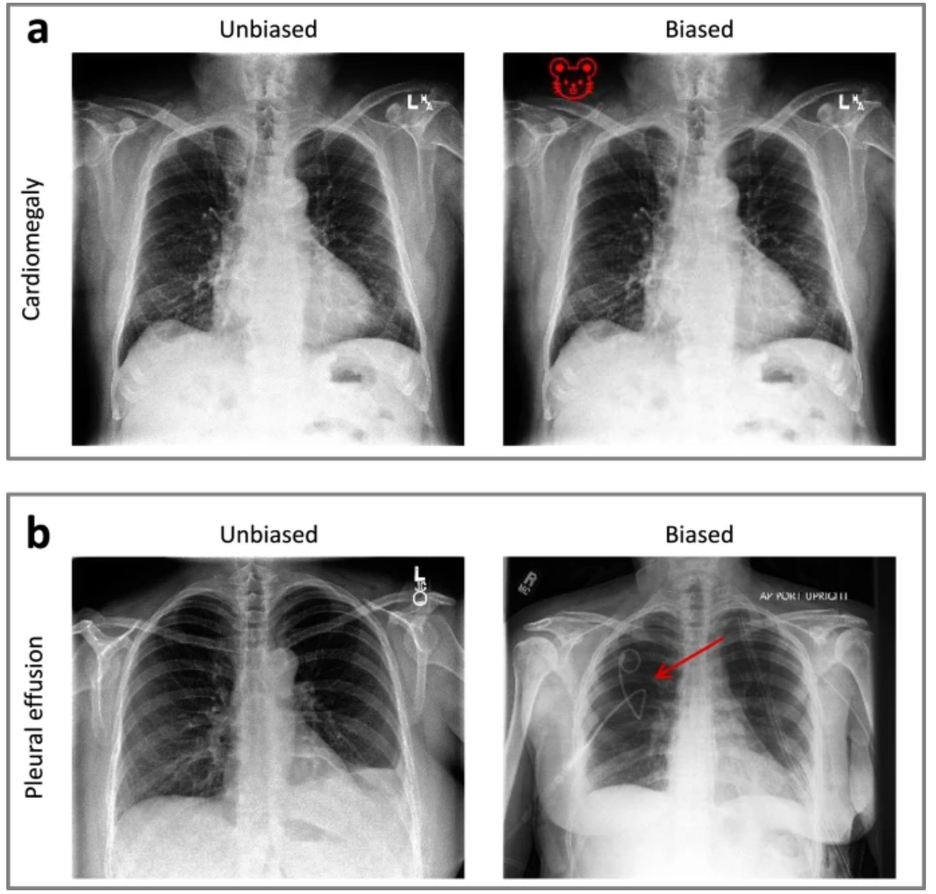
Heterogenous data



- Each Hospital collects data differently:
  - Different technologies (MRI machines, CT scanners, staining .)
  - Different procedures (is covid +ve: self-reported, PCR test, antigen, CT scan, doctor’s diagnosis?)

Heterogenous data

- Spurious/shortcuts learning. Turns out models were predicting based on:
    - The TAG of the hospital.
    - whether the patient was lying down or sitting up.



[Klaudia et al. 2024]

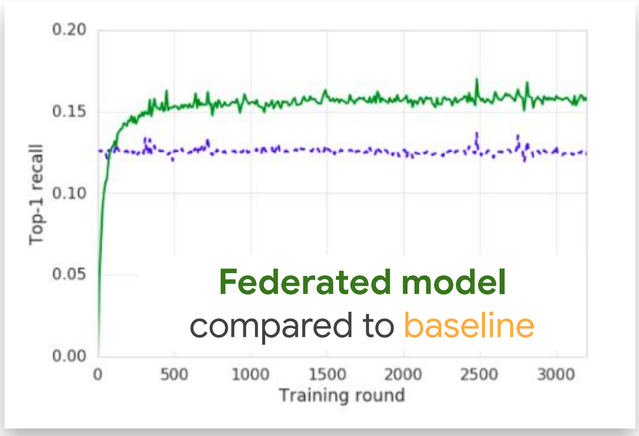
Cross-device and Cross-silo FL

	Cross-device FL	**Cross-silo FL**
Challenges	Data Privacy	Data privacy
	Scale + Resilience: large and unreliable networks	
	Noisy unreliable data	Heterogenous data and population
		Strategic concerns like fairness and accountability

Gboard: next-word prediction

Federated RNN (compared to prior n-gram model):

- Better next-word prediction accuracy: +24%
- More useful prediction strip: +10% more clicks



Cross-device federated learning at Apple


MIT Technology Review

Sign in Subscribe

Artificial intelligence / Machine learning

How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

by **Karen Hao** December 11, 2019

*"Instead, **it relies primarily on a technique called federated learning**, Apple's head of privacy, Julien Freudiger, told an audience at the Neural Processing Information Systems conference on December 8. Federated learning is a privacy-preserving machine-learning method that was first introduced by Google in 2017. It allows Apple to train different copies of a speaker recognition model across all its users' devices, using only the audio data available locally. It then sends just the updated models back to a central server to be combined into a master model. In this way, raw audio of users' Siri requests never leaves their iPhones and iPads, but the assistant continuously gets better at identifying the right speaker."*

https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/

Cross-silo federated learning from NVIDIA


NVIDIA

PLATFORMS DEVELOPERS INDUSTRIES SHOP DRIVERS SUPPORT ABOUT NVIDIA EMAIL SIGN-UP

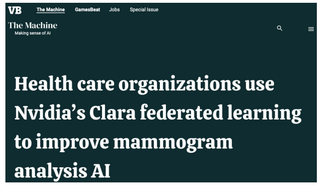
HOME DEEP LEARNING NETWORKING DRIVING GAMING PRO GRAPHICS AUTONOMOUS MACHINES HEALTHCARE AI PODCAST

Medical Institutions Collaborate to Improve Mammogram Assessment AI with NVIDIA Clara Federated Learning

In a federated learning collaboration, the American College of Radiology, Diagnosticos da America, Partners HealthCare, Ohio State University and Stanford Medicine developed better predictive models to assess breast tissue density.

April 15, 2020 by MONA FLORES

"Federated learning addresses this challenge, enabling different institutions to collaborate on AI model development without sharing sensitive clinical data with each other. The goal is to end up with more generalizable models that perform well on any dataset, instead of an AI biased by the patient demographics or imaging equipment of one specific radiology department."


VB The Machine GamesBeat Jobs Special Issue

The Machine
Making sense of AI

Health care organizations use Nvidia's Clara federated learning to improve mammogram analysis AI


MENU SEARCH MedCityNews

Theranos founder Elizabeth Holmes may seek 'mental disease' defense in trial, document shows

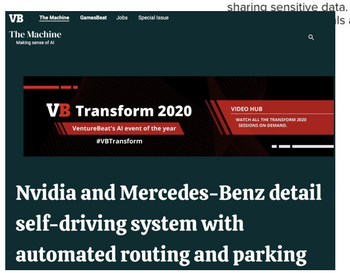
Covid-19 shows primary care's value to diverse communities and how to pay for it

3 ways t

HOSPITALS, ARTIFICIAL INTELLIGENCE, HEALTH TECH

Nvidia says it has a solution for healthcare's data problems

The chipmaker touted a new framework that would allow hospitals and pharmaceutical companies to collaborate on AI projects without sharing sensitive data. Nvidia said the framework is already gaining
ls and drug developers.


VB The Machine GamesBeat Jobs Special Issue

The Machine
Making sense of AI

VB Transform 2020
VentureBeat's AI event of the year
#VBTransform

VIDEO HUB
WATCH ALL THE TRANSFORM 2020 SESSIONS ON-DEMAND.

Nvidia and Mercedes-Benz detail self-driving system with automated routing and parking

[1] https://blogs.nvidia.com/blog/2020/04/15/federated-learning-mammogram-assessment/
[2] https://venturebeat.com/2020/04/15/healthcare-organizations-use-nvidias-clara-federated-learning-to-improve-mamm
[3] https://medcitynews.com/2020/01/nvidia-says-it-has-a-solution-for-healthcares-data-problems/
[4] https://venturebeat.com/2020/06/23/nvidia-and-mercedes-benz-detail-self-driving-system-with-automated-routing-and

Cross-silo federated learning from Intel


ARTIFICIAL INTELLIGENCE, DIAGNOSTICS

UPenn, Intel partner to use federated learning AI for early brain tumor detection

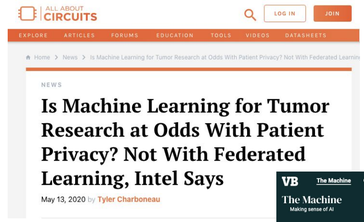
The project will bring in 29 institutions from North America, Europe and India and will use privacy-preserved data to train AI models. Federated learning has been described as being born at the intersection of AI, blockchain, edge computing and the Internet of Things.

By ALARIC DEARMENT

Post a comment / May 11, 2020 at 10:03 AM

"The University of Pennsylvania and chipmaker Intel are forming a partnership to enable 29 healthcare and medical research institutions around the world to train artificial intelligence models to detect brain tumors early."

"The program will rely on a technique known as federated learning, which enables institutions to collaborate on deep learning projects without sharing patient data. The partnership will bring in institutions in the U.S., Canada, U.K., Germany, Switzerland and India. The centers – which include Washington University of St. Louis; Queen's University in Kingston, Ontario; University of Munich; Tata Memorial Hospital in Mumbai and others – will use Intel's federated learning hardware and software."


ALL ABOUT CIRCUITS

EXPLORE ARTICLES FORUMS EDUCATION TOOLS VIDEOS DATASHEETS

Home > News > Is Machine Learning for Tumor Research at Odds With Patient Privacy? Not With Federated Learning

NEWS

Is Machine Learning for Tumor Research at Odds With Patient Privacy? Not With Federated Learning, Intel Says

May 13, 2020 by Tyler Charboneau


Bio-IT World

Next-Gen Technology • Big Data • Personalized Medicine

Subscribe News Advertise Free Downloads Events About Bio-IT

RELATED STORIES

Intel, Penn Medicine Launch Federated L

By Allison Proffitt

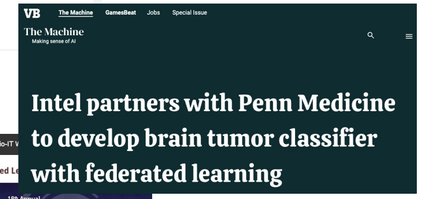
May 28, 2020 | The University of Pennsylvania and Intel have built a federation of 30 institutions to use federated learning to train artificial intelligence (AI) models to identify boundaries of brain tumors.

Led by Spyridon Bakas at the Center for Biomedical Image Computing and Analytics (CBICA) at the Perelman School of Medicine at the University of Pennsylvania, the federation is the next step forward in a years-long effort to gather data that would empower AI in brain image analysis.

18th Annual Discovery on TARGET VIRTUAL

The Industry's Preeminent Event on Novel Drug Targets and Technologies

September 16-18, 2020


VB The Machine GamesBeat Jobs Special Issue

The Machine
Making sense of AI

Intel partners with Penn Medicine to develop brain tumor classifier with federated learning

[1] https://medcitynews.com/2020/05/upenn-intel-partner-to-use-federated-learning-ai-for-early-brain-tumor-detection/
[2] https://www.allaboutcircuits.com/news/can-machine-learning-keep-patient-privacy-for-tumor-research-intel-says-yes-with-federated-learning/
[3] https://venturebeat.com/2020/05/11/intel-partners-with-penn-medicine-to-develop-brain-tumor-classifier-with-federated-learning/
[4] http://www.bio-itworld.com/2020/05/28/intel-penn-medicine-launch-federated-learning-model-for-brain-tumors.aspx

The Dream vs. Current Reality of FL

- The dream:
  - A private distributed global protocol
  - That unites the world's data and compute

- The current reality
  - Data is extremely messy and even actively harmful - need **data quality and valuation.**
  - Cannot train LLMs over commodity hardware - need better **sysML.**

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven

July 30, 2021

Wynants, Laure, et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal." bmj 369 (2020).