Three Pillars of Health Data Science
Transfer Learning, Federated Learning, and Imbalanced Learning

Ju Sun, PhD
Computer Science & Engineering
Jan 12, 2023

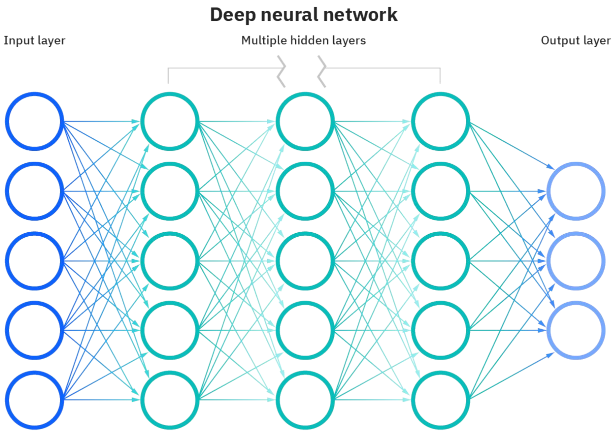
GLOVEX https://glovex.umn.edu/

University of Minnesota
Driven to Discover℠

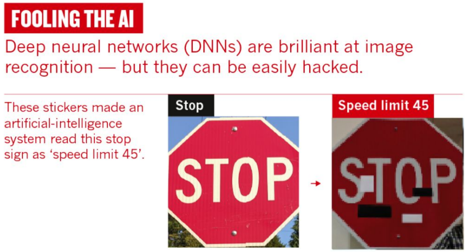
Research in the group

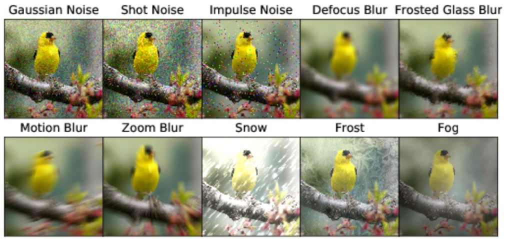
GLOVEX

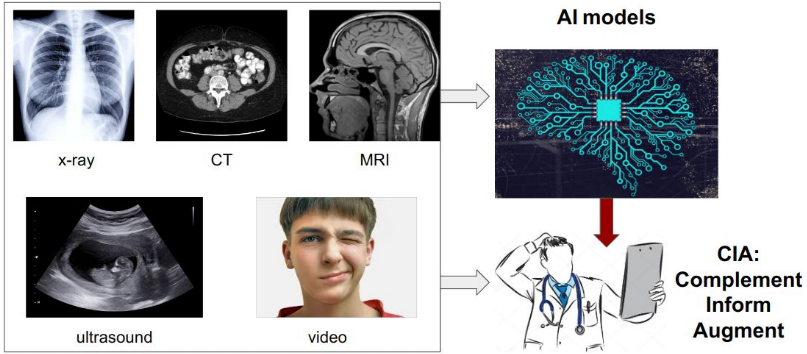
(Machine) **Learning**, (Numerical) **Optimization**, (Computer) **Vision**, healthcar**E**, + **X**



Our research themes

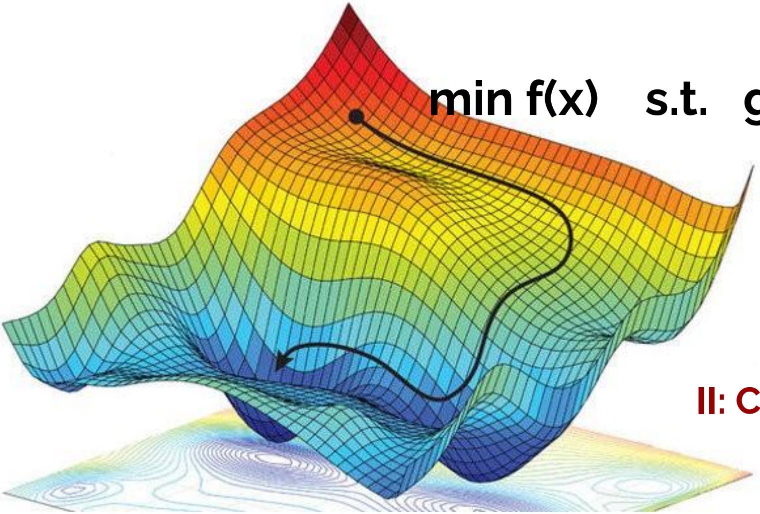




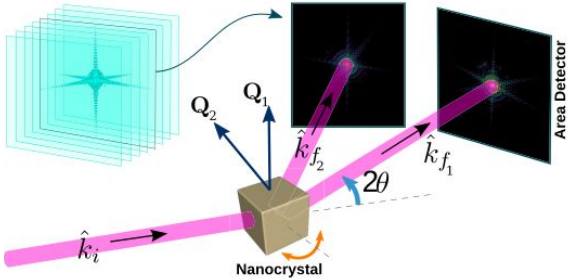


I: Trustworthy AI

III: AI for Healthcare



$$\min f(x) \quad \text{s.t.} \quad g(x) <= 0$$

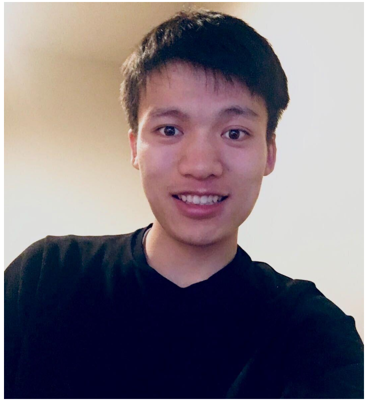


II: Computation for AI

IV: AI for Science and Engineering

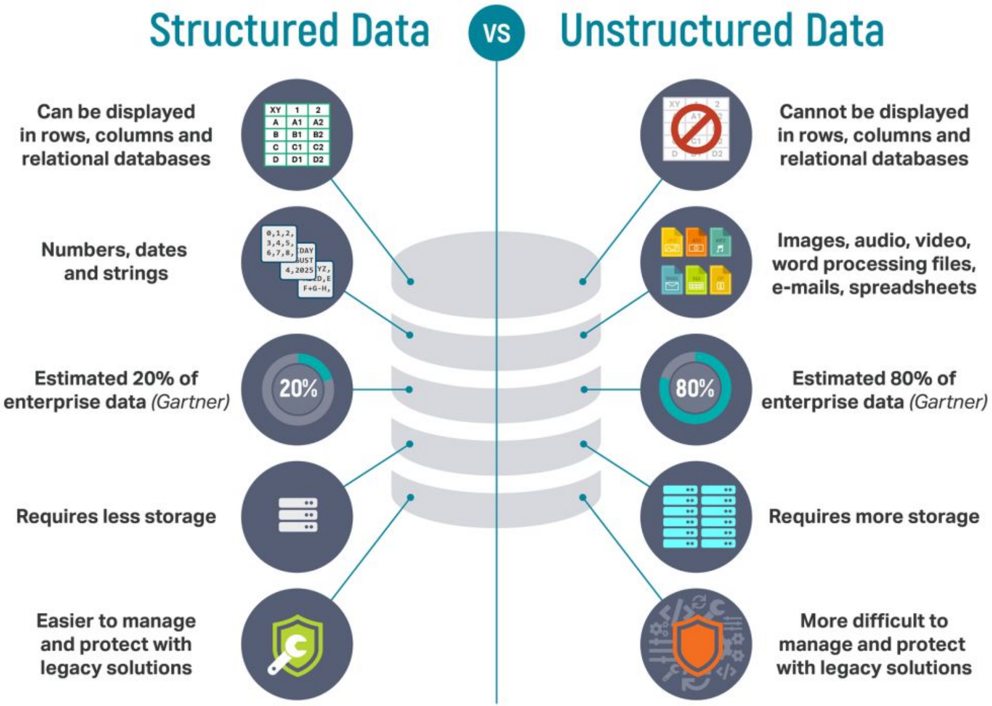
Thanks to

Thanks to



Le Peng (CS&E, PhD)

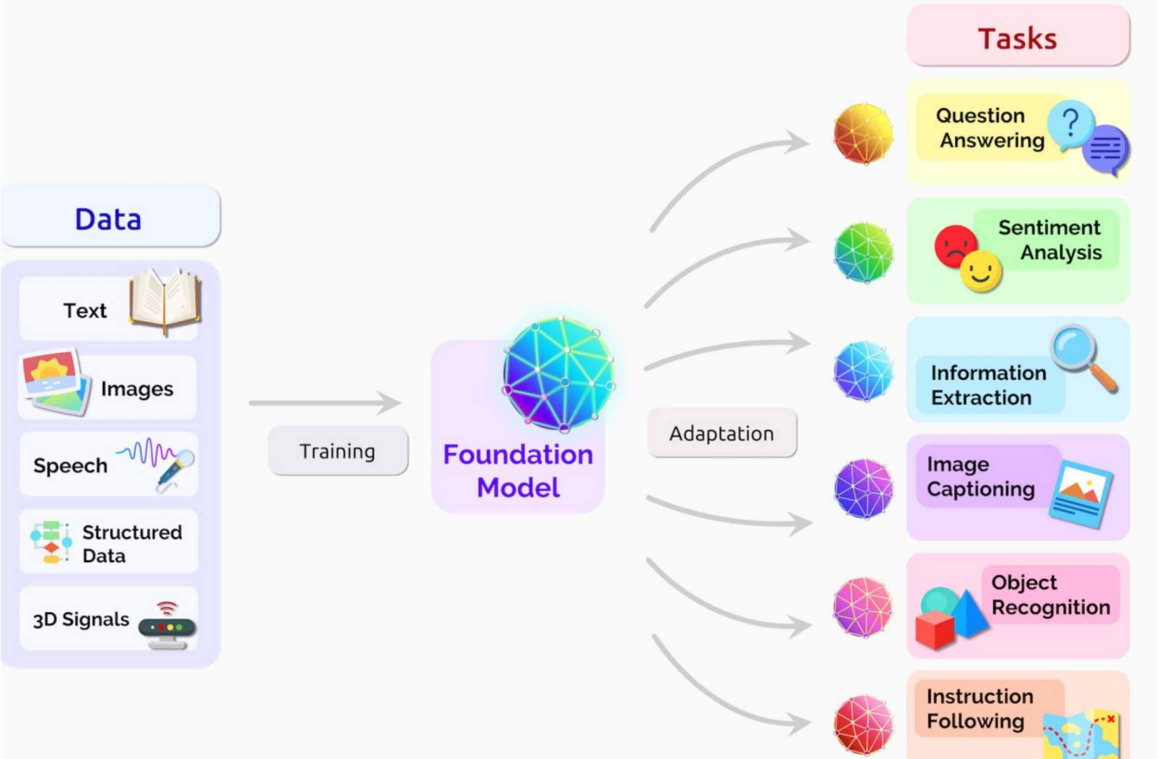
Deep learning is mostly for unstructured data



- Structured data directly go to classical MLDS tools
- Success of modern DL lies in **representation learning**

Credit: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

Deep learning is data-hungry



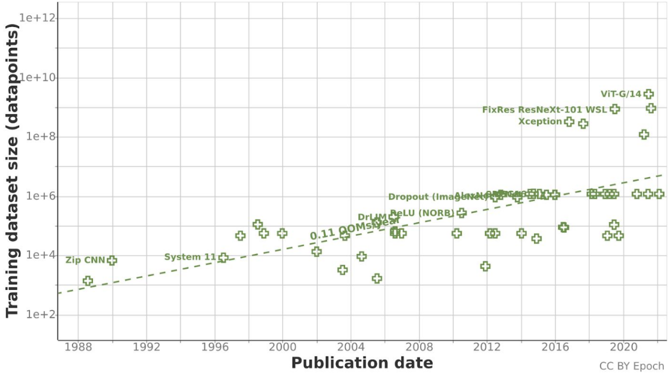
Credit: https://arxiv.org/abs/2108.07258

NLP models

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-Gen (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Credit: https://dl.acm.org/doi/10.1145/3442188.3445922

CV models



Credit: https://epochai.org/blog/trends-in-training-dataset-sizes

Deep learning is data-picky

Need **well-curated** datasets for training and evaluation

GLUE

NYU ML² UWNLP DeepMind

The General Language Understandi
resources for training, evaluating, and a
consists of:

SQuAD2.0

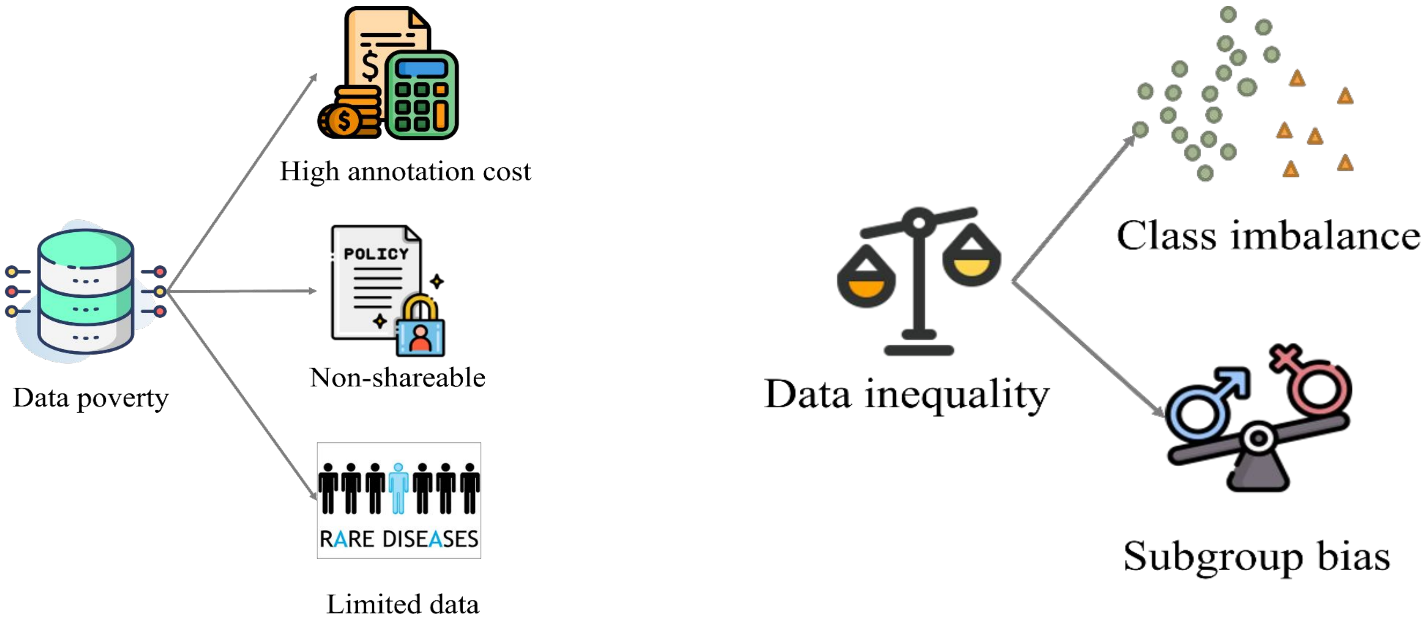
The Stanford Question Answering Dataset

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

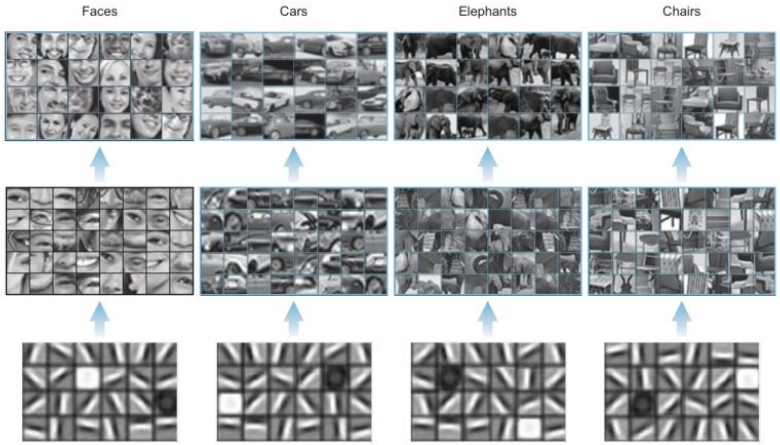
Data poverty and inequality (DPI) in healthcare



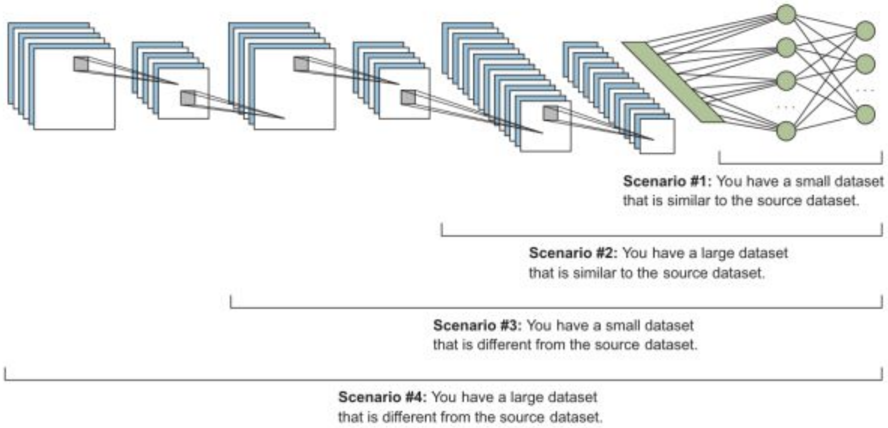
Outline

- **Addressing data poverty—transfer learning**
- Addressing data poverty—federated learning
- Addressing data inequality—imbalanced learning
- Perspective: toward human-in-the-loop health data science

Addressing data poverty—transfer learning



(Credit: [Elgendy, 2020])



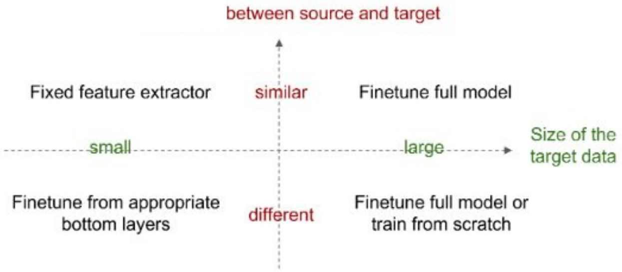


Fig. 2. Illustration of different DCNN-based TL scenarios and strategies

Truncated transfer learning (TTL)

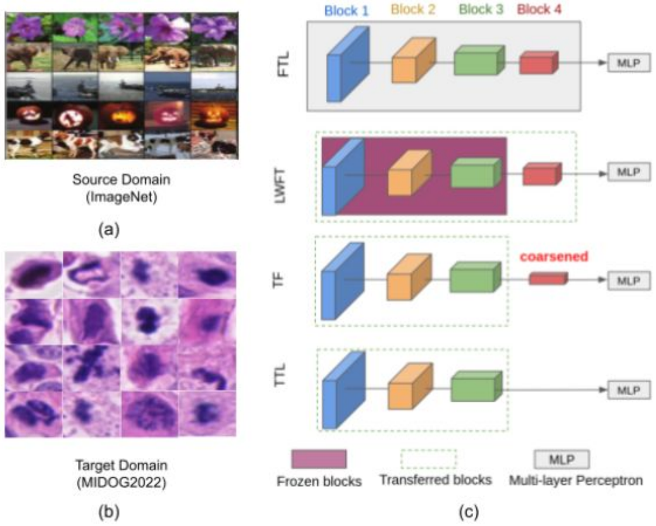


Fig. 3. Overview of typical TL setup, and the four TL methods that we focus on in this paper. (a) TL source domain: e.g., ImageNet object recognition; (b) TL target domain: e.g., mitotic cells classification; (c) Four TL methods: FTL, LWFT, TF, our TTL applied to ResNet50 pretrained on ImageNet.

Rethinking Transfer Learning for Medical Image Classification

Le Peng, Hengyue Liang, Gaoxiang Luo, Taihui Li, Ju Sun

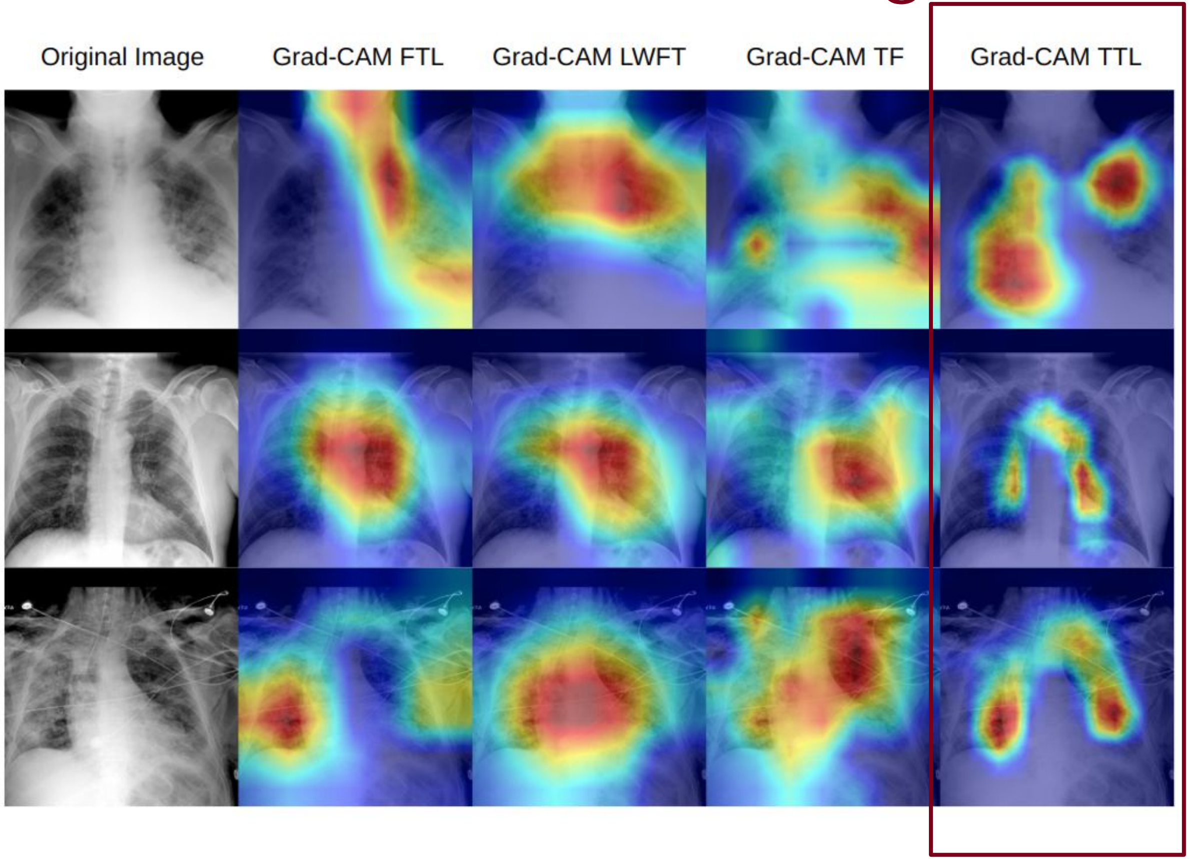
https://arxiv.org/abs/2106.05152

3D PULMONARY EMBOLISM CLASSIFICATION WITH DIFFERENT TL STRATEGIES. THE BEST RESULT OF EACH COLUMN IS COLORED IN RED. ↑ INDICATES LARGER VALUE IS BETTER AND ↓ INDICATES LOWER VALUE IS BETTER. "-1" MEANS WITH THE BLOCK-WISE SEARCH ONLY, AND "-2" MEANS WITH THE TWO-STAGE BLOCK-LAYER HIERARCHICAL SEARCH. NOTE THAT THE RUN TIME FOR THIS TABLE IS IN SECONDS, NOT MILLISECONDS.

Method	AUROC↑	AUPRC↑	Params(M)↓	MACs(G)↓	CPU(s)↓	GPU(s)↓
PENet	0.822 ± 0.010	0.855 ± 0.007	28.4	**51.7**	1.50	**1.59e-2**
FTL	0.821 ± 0.010	0.867 ± 0.006	47.5	66.3	1.44	1.96e-2
TF-1	0.849 ± 0.020	0.886 ± 0.017	36.1	64.9	1.41	1.93e-2
LWFT-1	0.817 ± 0.005	0.855 ± 0.003	47.5	66.3	1.44	1.96e-2
TTL-1	**0.854 ± 0.013**	**0.889 ± 0.015**	**26.11**	60.17	**1.32**	1.68e-2
TF-2	0.849 ± 0.020	0.886 ± 0.017	36.1	64.9	1.41	1.93e-2
LWFT-2	0.835 ± 0.038	0.870 ± 0.028	47.5	66.3	1.44	1.96e-2
TTL-2(ours)	**0.854 ± 0.013**	**0.889 ± 0.015**	**26.11**	60.17	**1.32**	1.68e-2

Smaller DNN model, boosted performance!

Truncated transfer learning (TTL)



For COVID-19 prediction:

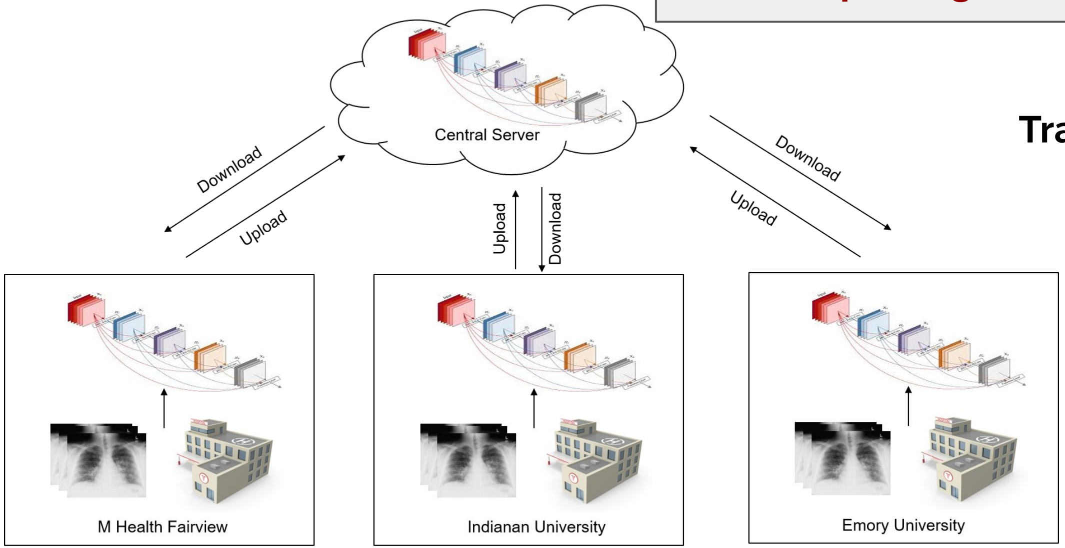
TTL correctly focuses more on texture (lesion) in the lung area!

Outline

- Addressing data poverty—transfer learning
- **Addressing data poverty—federated learning**
- Addressing data inequality—imbalanced learning
- Perspective: toward human-in-the-loop health data science

Addressing data poverty—federated learning

While respecting data privacy

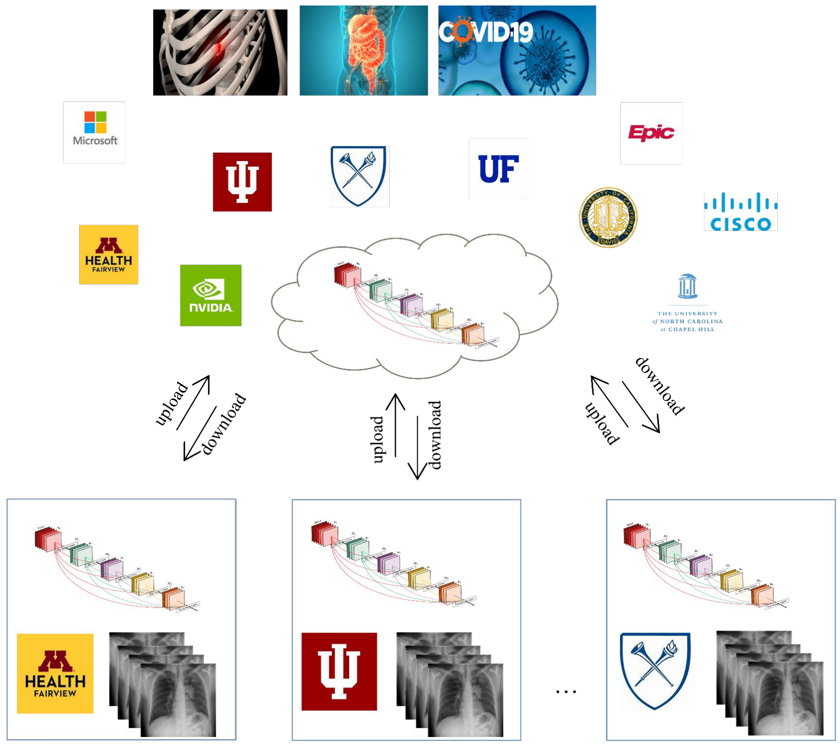


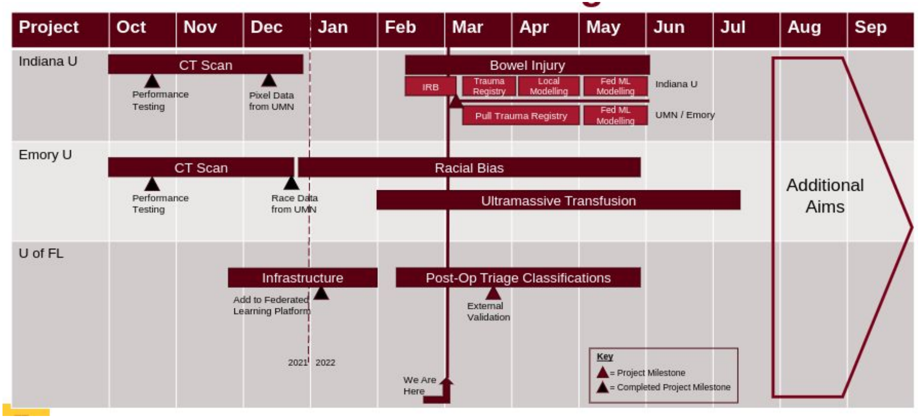
Traditional distributed learning:
Distribute the computing loads

Federated learning:
- Respect data privacy
- Share the intermediate MLDS models, **not the raw data**

Starting 2020 for medical imaging data

Our medical CV federation





Status of our CV federation

- ✓ (UMN) COVID-19 detection (UF, Emory, IU and MHealth Fairview)
- ✓ (Emory) Racial Bias study (Emory, IU and Mhealth Fairview)
- ❏ (UMN) RibFrac detection (Emory, IU and Mhealth Fairview)

FL COVID-19 detection

JOURNAL ARTICLE

Evaluation of federated learning variations for COVID-19 diagnosis using chest radiographs from 42 US and European hospitals

Le Peng, Gaoxiang Luo, Andrew Walker, Zachary Zaiman, Emma K Jones, Hemant Gupta, Kristopher Kersten, John L Burns, Christopher A Harle, Tanja Magoc ... Show more

Journal of the American Medical Informatics Association, ocac188,
https://doi.org/10.1093/jamia/ocac188
Published: 20 October 2022 **Article history** ▾

Federated learning (Journal of American Medical Informatics Association; 2022)

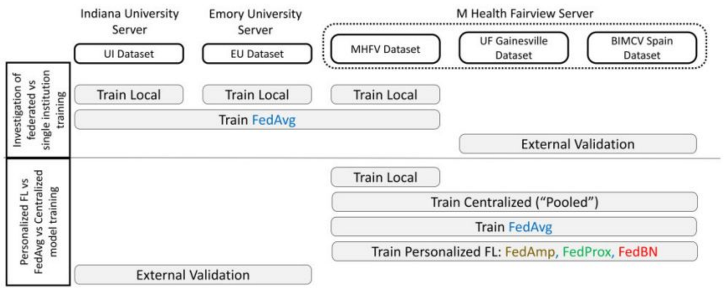


Figure 1. Schematic representation of the available datasets and the analysis conducted for this study. IU: Indiana University; EU: Emory University; MHFV: M Health Fairview; UF: University of Florida; BIMCV: Valencian Region Medical ImageBank.

Table 2. Internal and external validation of federated model

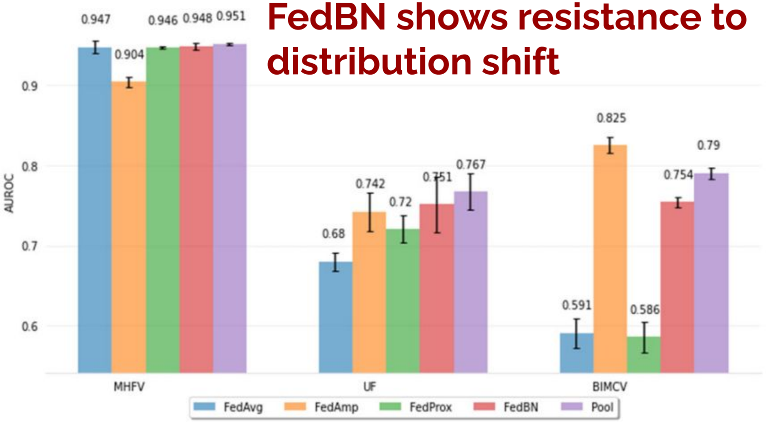
		N	AUROC	AUPRC	95% CI	Precision	Recall	F1 score
Internal	MHFV	9102	0.951	0.838	0.940–0.963	0.616	0.840	0.711
	IU	3179	0.871	0.886	0.857–0.885	0.828	0.748	0.786
	EU	4051	0.832	0.801	0.813–0.851	0.681	0.784	0.729
External	BIMCV	3822	0.601	0.611	0.585–0.617	0.616	0.471	0.533
	UF	2469	0.713	0.651	0.692–0.734	0.629	0.592	0.610

FL shows good generalization on external validation

Table 3. Performance comparison between single institution model (SIM) and federated learning model (FLM)

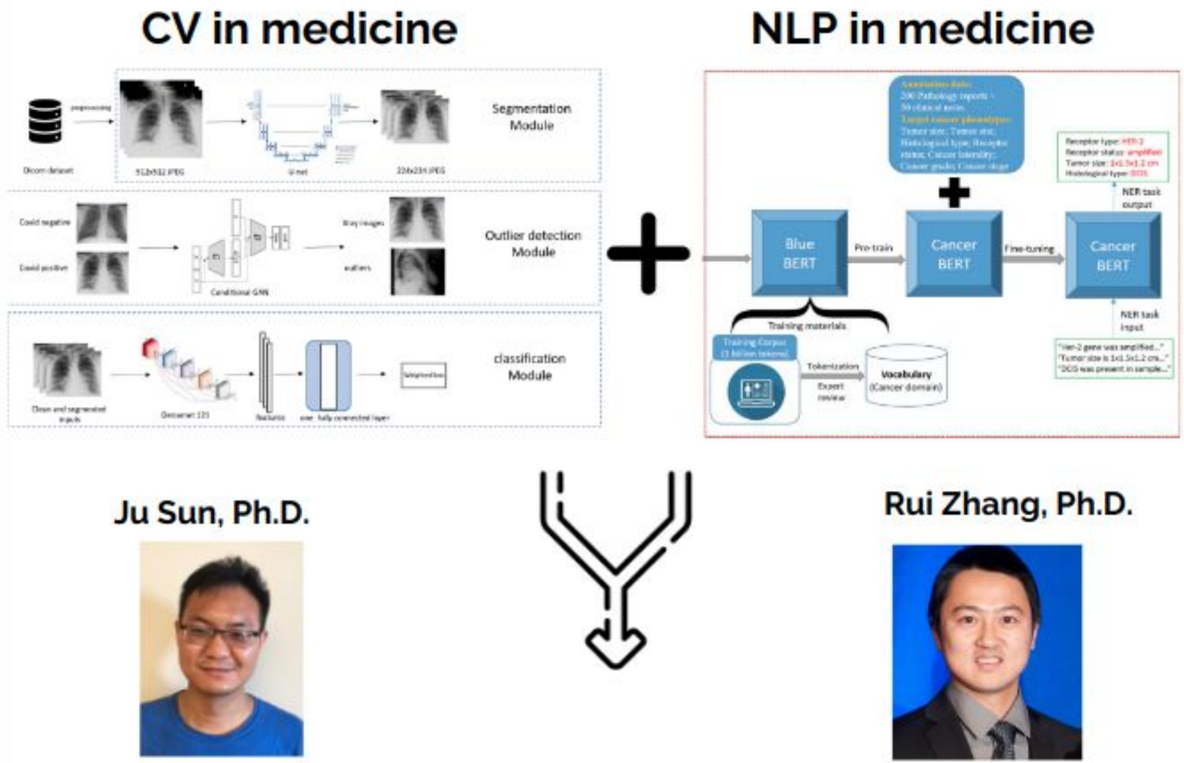
	AUROC			Sensitivity			Specificity		
	SIM	FLM	*P* value	SIM	FLM	*P* value	SIM	FLM	*P* value
MHFV	0.944	0.951	.492	0.870	0.840	.020	0.939	0.950	<.05
BIMCV	0.557	0.601	<.05	0.301	0.471	<.05	0.833	0.730	<.05
UF	0.667	0.713	<.05	0.548	0.592	<.05	0.721	0.759	<.05

Note: We use Delong's test to compare the difference of AUROC and McNemar's test to compare specificity and sensitivity.



FedBN shows resistance to distribution shift

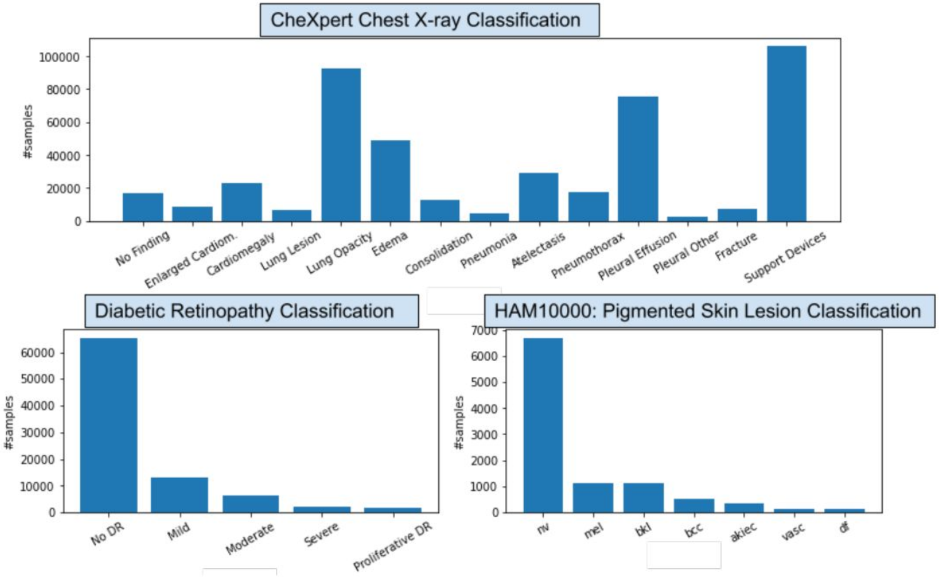
Next: FL for CV + NLP



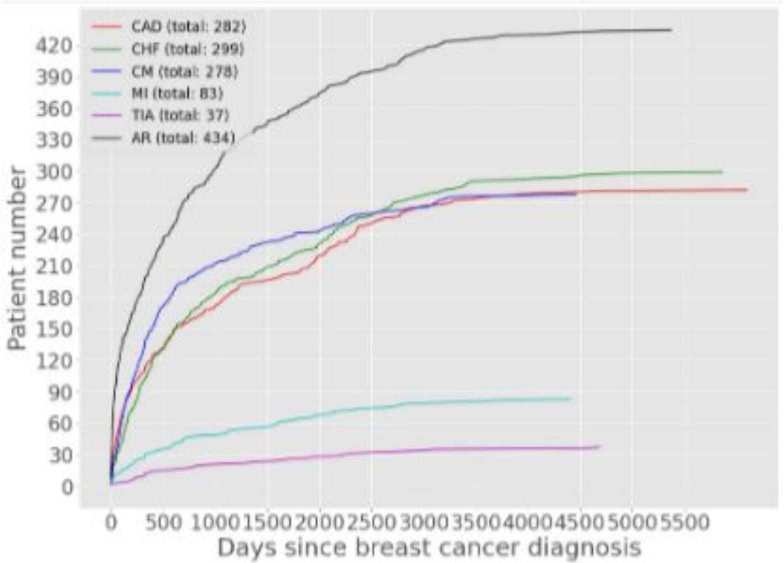
Outline

- Addressing data poverty—transfer learning
- Addressing data poverty—federated learning
- **Addressing data inequality—imbalanced learning**
- Perspective: toward human-in-the-loop health data science

Addressing data inequality—imbalanced learning



Imbalanced classification (IC)



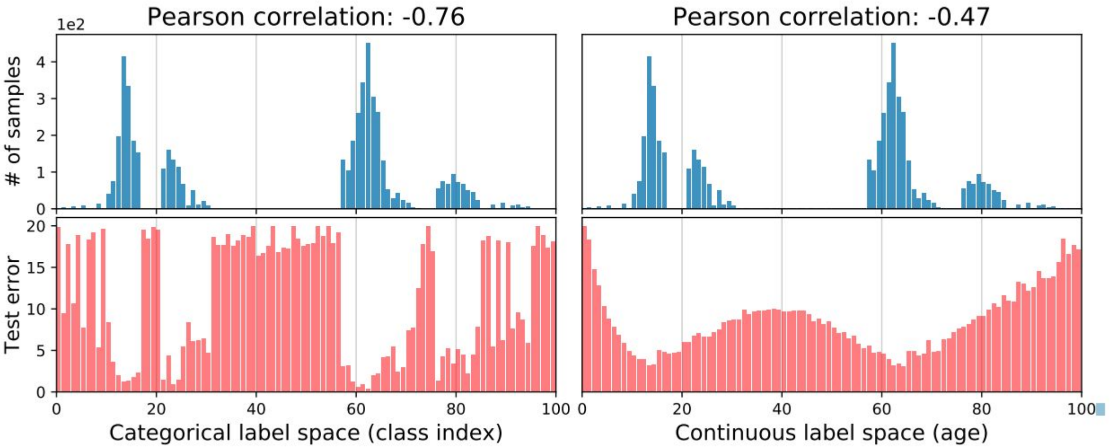
Imbalanced regression (IR)

While imbalance learning is challenging?

	Predicted POS	Predicted NEG
POS	70	30
NEG	1000	9000

Accuracy: 9070/10100 = 0.898
True Positive Rate (Sensitivity, Recall): 0.7
True Negative Rate (Specificity): 0.9
Balanced Accuracy: (0.7 + 0.9)/2 = 0.80
Precision (POS): 70/1070 = 0.065
F1 Score: 2*0.065*0.7/(0.065 + 0.7) = 0.119

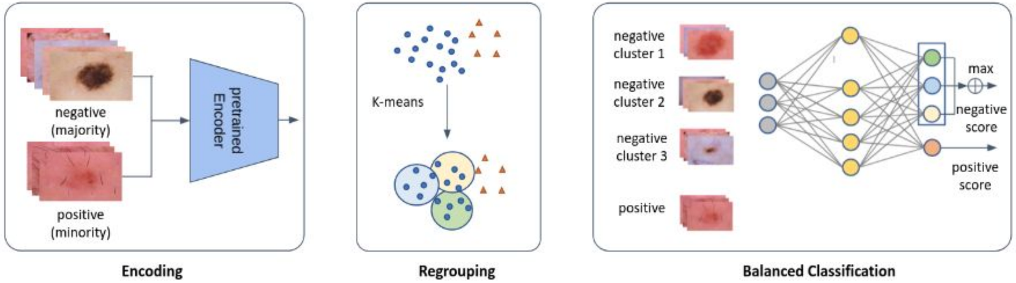
Figure 2: An example confusion table for binary classification, and the various associated performance metrics. POS: positive; NEG: negative.



(a) CIFAR-100 (subsampled) (b) IMDB-WIKI (subsampled)

Evaluation metrics ⇒ Learning goals matter!

SOTA methods for IC is (substantially?) suboptimal



Imbalanced Classification in Medical Imaging via Regrouping

Le Peng[1], Yash Travadi[2], Rui Zhang[3], Ying Cui[4], Ju Sun[1]
[1]Computer Science & Engineering, University of Minnesota, Twin Cities
[2]School of Statistics, University of Minnesota, Twin Cities
[3]Department of Surgery, University of Minnesota, Twin Cities
[4]Industrial and Systems Engineering, University of Minnesota, Twin Cities
{peng0347,trava029,zhan1386,yingcui,jusun}@umn.edu

Imbalanced learning (NeurIPS'22 Workshop: When Medical Imaging Meets NeurIPS) https://arxiv.org/abs/2210.12234

Binary Classification

Method	binary CIFAR-100			binary HAM10000		
	BA (%) ↑	AP (%) ↑ Neg (45,000)	AP (%) ↑ Pos (500)	BA (%) ↑	AP (%) ↑ Neg (9,688)	AP (%) ↑ Pos (327)
CE	81.9	99.9	68.1	76.6	99.6	67.3
WCE	84.5	99.9	58.2	84.9	99.7	56.5
Focal	80.4	99.7	70.5	51.9	90.8	37.0
LDAM	77.4	**100**	62.8	50.0	98.9	20.8
LA	81.9	**100**	[illegible]	[illegible]	99.5	[illegible]
AP	73.8	99.9	54.6	50.0	99.5	34.1
RUSC	84.4	99.7	16.8	**89.7**	99.6	35.6
DSMT	58.0	99.7	48.7	76.0	99.5	66.2
ROS	83.4	99.4	68.8	81.1	99.4	74.7
RG+CE$_m$	**87.9** +6.0	99.8 -0.1	**77.2** +9.1	83.7 +7.1	99.2 -0.4	79.9 +12.5
RG+CE$_s$	86.9 +5.0	99.9 +0.0	76.2 +8.1	80.6 +4.0	**99.9** +0.3	79.9 +12.5
RG+WCE$_m$	84.9 +3.0	99.8 -0.1	74.6 +6.5	85.0 +8.4	99.1 -0.5	**83.9** +16.5
RG+WCE$_s$	83.4 +1.5	99.8 -0.1	74.6 +6.5	80.8 +8.4	**99.9** +0.3	**83.9** +16.5

Our simple method outperforms SOTA!

Multi-class Classification

Method	BA (%) ↑	AP (%) ↑ nv 6705	mel 1113	bkl 1099	bcc 514	bakiec 327	vasc 142	df 115
CE	62.5	96.7	66.4	73.5	79.1	59.2	86.0	53.8
WCE	66.3	96.3	46.5	58.5	67.6	54.9	88.2	57.8
Focal	60.3	96.9	62.5	69.2	74.9	48.7	84.3	50.0
LDAM	56.5	96.0	62.9	66.2	71.0	51.6	83.6	10.0
LA	[illegible]	96.0	67.9	72.3	71.1	65.5	84.2	19.3
RUSC	59.4	92.4	30.9	29.0	39.8	24.9	74.9	39.7
DSMT	60.5	97.2	65.9	70.5	76.8	58.3	81.4	51.0
ROS	71.5	97.5	**73.3**	**82.8**	**88.2**	71.2	94.2	61.8
RG+CE$_m$	66.6	95.6	72.8	82.2	78.1	70.0	92.7	62.4
RG+CE$_s$	67.5	95.6	72.8	82.2	78.1	70.0	92.7	62.4
RG+WCE$_m$	**72.8**	94.3	72.6	76.0	82.0	68.9	**95.2**	**72.5**
RG+WCE$_s$	67.9	**98.0**	72.7	78.0	82.8	**71.4**	91.1	69.8

Ongoing: principled learning goals

$$\textbf{fix precision, optimize recall (FPOR):} \quad \max_{\boldsymbol{\theta},t} \ \mathrm{recall}(f_{\boldsymbol{\theta}}, t) \quad \mathrm{s.t.}\ \ \mathrm{precision}(f_{\boldsymbol{\theta}}, t) \geq \alpha,$$

$$\textbf{fix recall, optimize precision (FROP):} \quad \max_{\boldsymbol{\theta},t} \ \mathrm{precision}_t \quad \mathrm{s.t.}\ \mathrm{recall}(f_{\boldsymbol{\theta}}, t) \geq \alpha,$$

$$\textbf{optimize } F_\beta \textbf{ score (OFBS):} \quad \max_{\boldsymbol{\theta},t} \ F_\beta(f_{\boldsymbol{\theta}}, t),$$

$$\textbf{optimize AP (OAP):} \quad \max_{\boldsymbol{\theta}} \ \mathrm{AP}(f_{\boldsymbol{\theta}}).$$

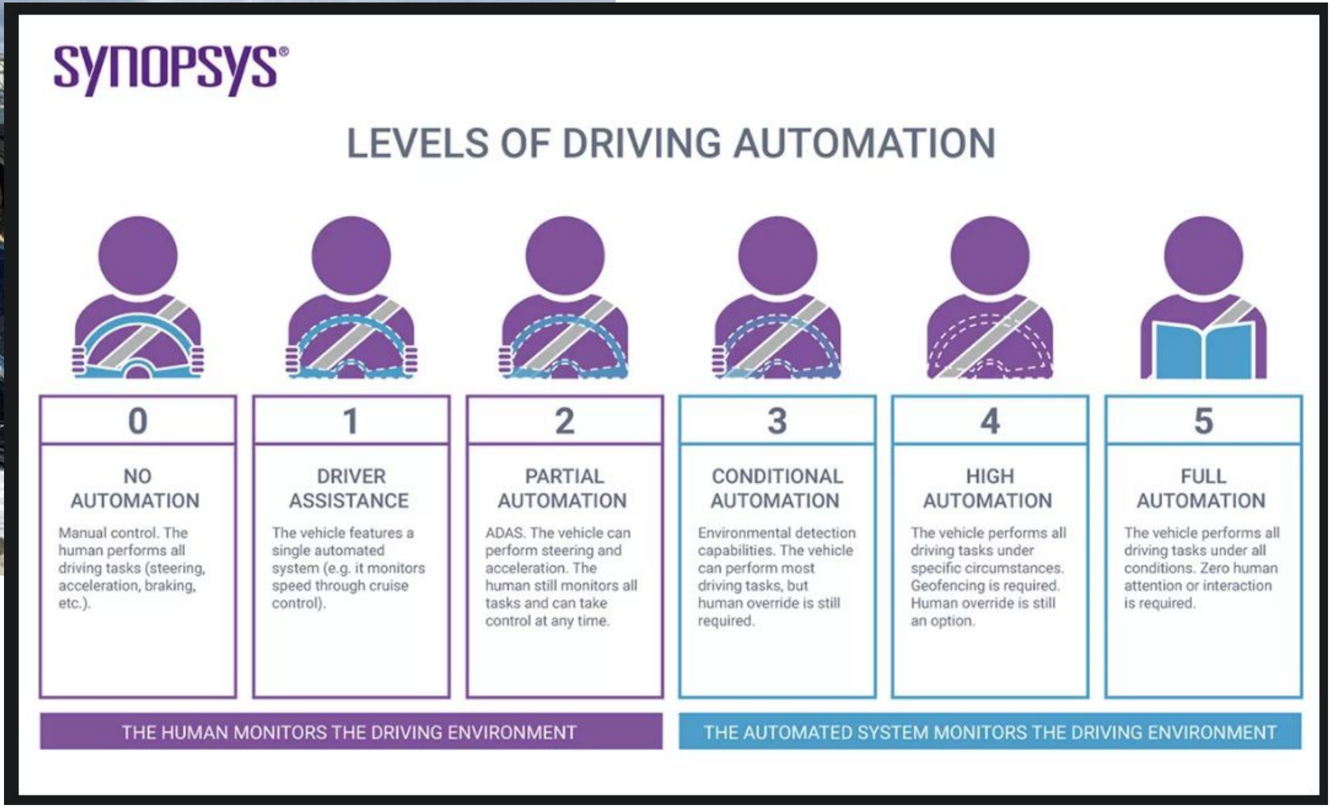
$$\textbf{optimize multiclass performance (OMCP):} \quad \max_{\boldsymbol{\theta},\boldsymbol{t}} \ \text{multiclass-metric}(f_{\boldsymbol{\theta}}, \boldsymbol{t}).$$

$$\textbf{optimize regression performance (OREGP):} \quad \max_{\boldsymbol{\theta}} \ \text{regression-metric}(f_{\boldsymbol{\theta}});$$

Outline

- Addressing data poverty—transfer learning
- Addressing data poverty—federated learning
- Addressing data inequality—imbalanced learning
- **Perspective: toward human-in-the-loop health data science**

Different levels of self-driving cars



Toward different levels of AI-assisted healthcare

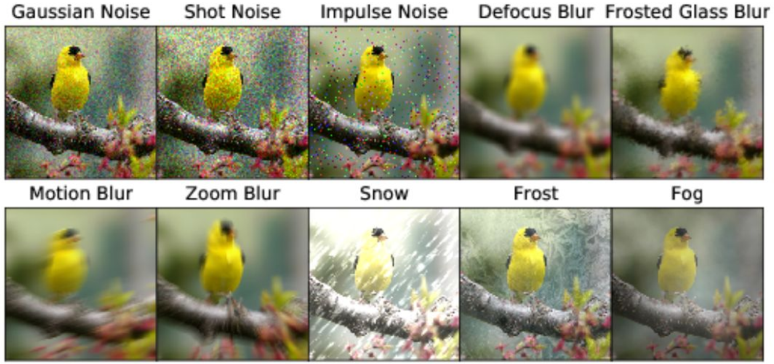
$$\max_{\boldsymbol{\theta},t}\ \mathrm{recall}(f_{\boldsymbol{\theta}},t) \quad \mathrm{s.t.}\ \ \mathrm{precision}(f_{\boldsymbol{\theta}},t) \geq \alpha,$$

$$\max_{\boldsymbol{\theta},t}\ \mathrm{precision}_t \quad \mathrm{s.t.}\ \mathrm{recall}(f_{\boldsymbol{\theta}},t) \geq \alpha,$$

$$\max_{\boldsymbol{\theta},t}\ F_{\beta}(f_{\boldsymbol{\theta}},t),$$

$$\max_{\boldsymbol{\theta}}\ \mathrm{AP}(f_{\boldsymbol{\theta}}).$$

Setting realistic goals: to be aligned with practical clinical demand



Addressing robustness: identifying most common nuisance factors in medical AI

Machine Learning with a Reject Option: A survey

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, Jesse Davis

Machine learning models always make a prediction, even when it is likely to be inaccurate. This behavior should be avoided in many decision support applications, where mistakes can have severe consequences. Albeit already studied in 1970, machine learning with a reject option recently gained interest. This machine learning subfield enables machine learning models to abstain from making a prediction when likely to make a mistake.

Allowing abstention: refraining from making prediction when sensing uncertainty/robustness issues

computer
science & Engineering

GLOVEX

glovex.umn.edu

GROUP OF LEARNING, OPTIMIZATION, VISION, HEALTHCARE, AND X

UNIVERSITY OF MINNESOTA
Driven to Discover℠