

Reality-Centric AI

Mihaela van der Schaar

John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge //
Director, Cambridge Center for AI in Medicine // Turing Faculty Fellow, The Alan Turing Institute



van_der_Schaar
\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE



mv472@cam.ac.uk



@MihaelaVDS



linkedin.com/in/
mihaela-van-der-schaar/

A complex world

- Enormous leaps in AI (NLP/ChatGPT, computer vision, robotics, etc.)
- And yet...
 - Deliver personalized medicine?
 - Transform healthcare systems?
 - Traffic control?
 - Solve the energy crisis?
 - Transform education?



A complex world

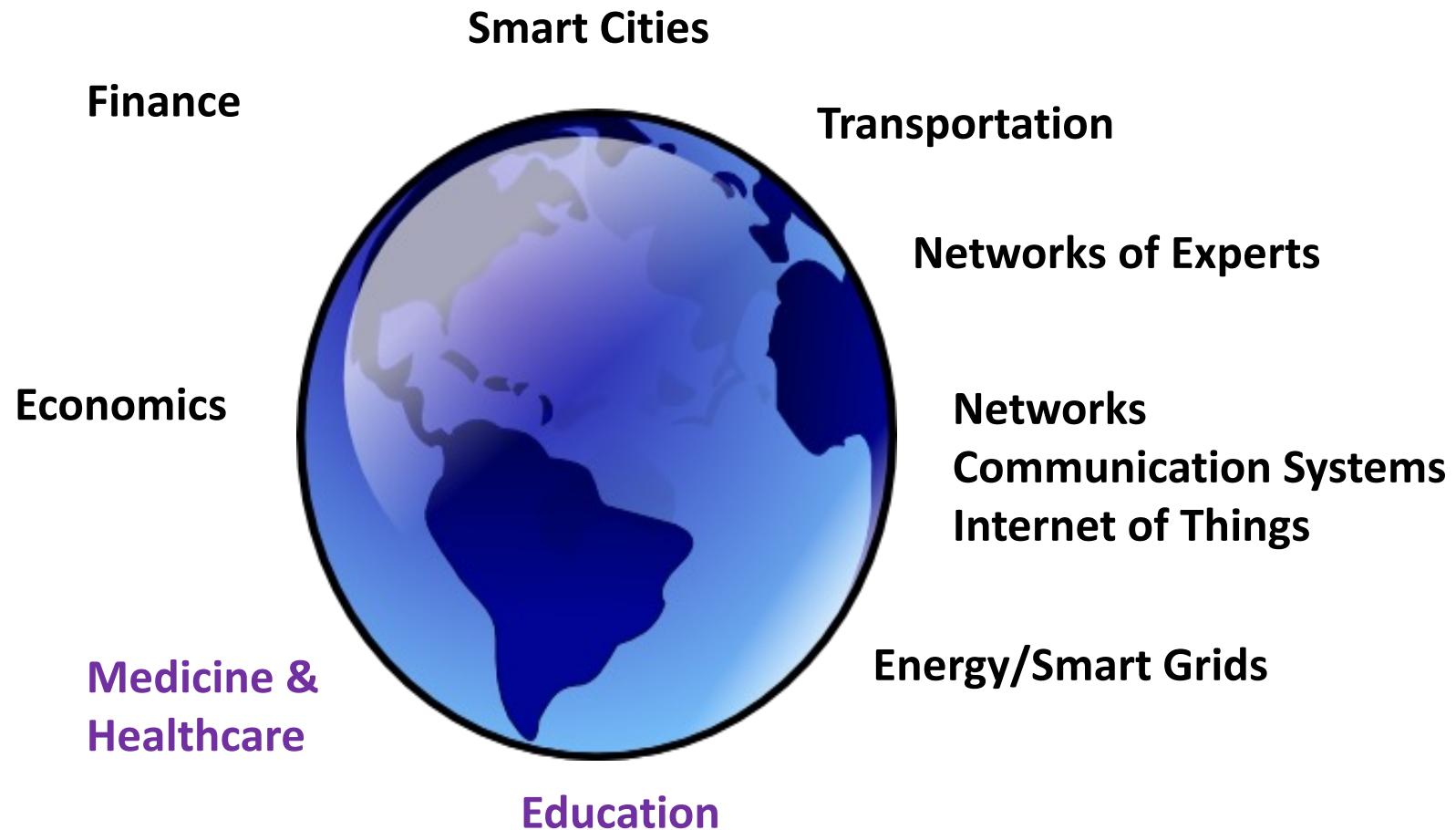
- Enormous leaps in AI (NLP/ChatGPT, computer vision, robotics, etc.)
- And yet...
 - Deliver personalized medicine?
 - Transform healthcare systems?
 - Traffic control?
 - Solve the energy crisis?
 - Transform education?

Not just a few breakthroughs away.....

Because of the **complexity of the real-world** – a different AI/ML paradigm



A complex world – our lab's work for the past 20 years



Solving these
complex
human-centric
problems is
our biggest task
as AI researchers!



A fundamentally new paradigm is needed for AI

- Reality-centric AI aims to reorientate AI towards the complexities of the real world



<https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/>

What is Reality-centric AI?

- AI which can operate effectively, reliably, and accountably
in the real world
- Focuses on inherent and unavoidable complexity of the real world at the heart of designing, training, testing, and deploying AI models.



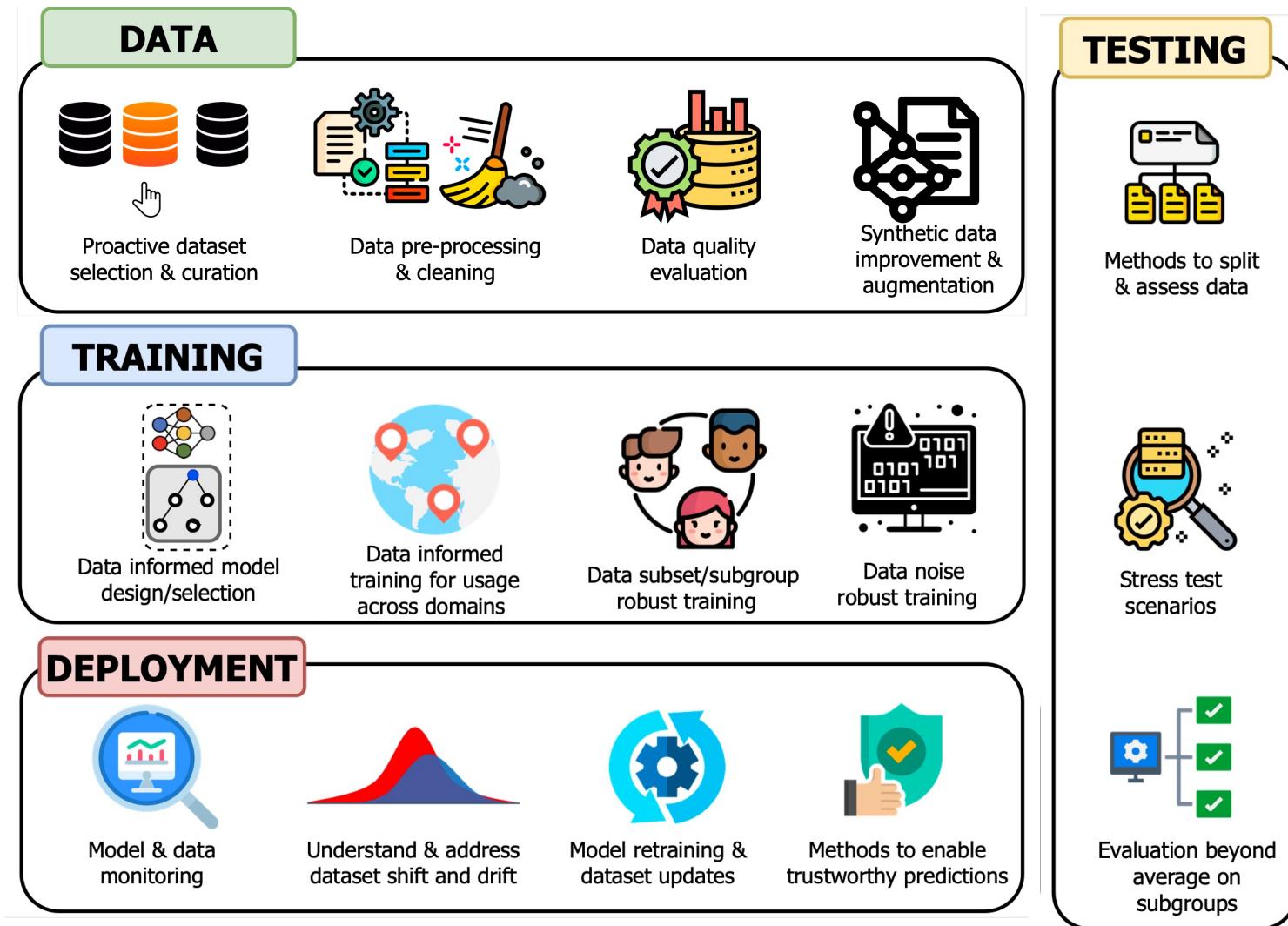
Reality-centric AI – Use of Real-world Data



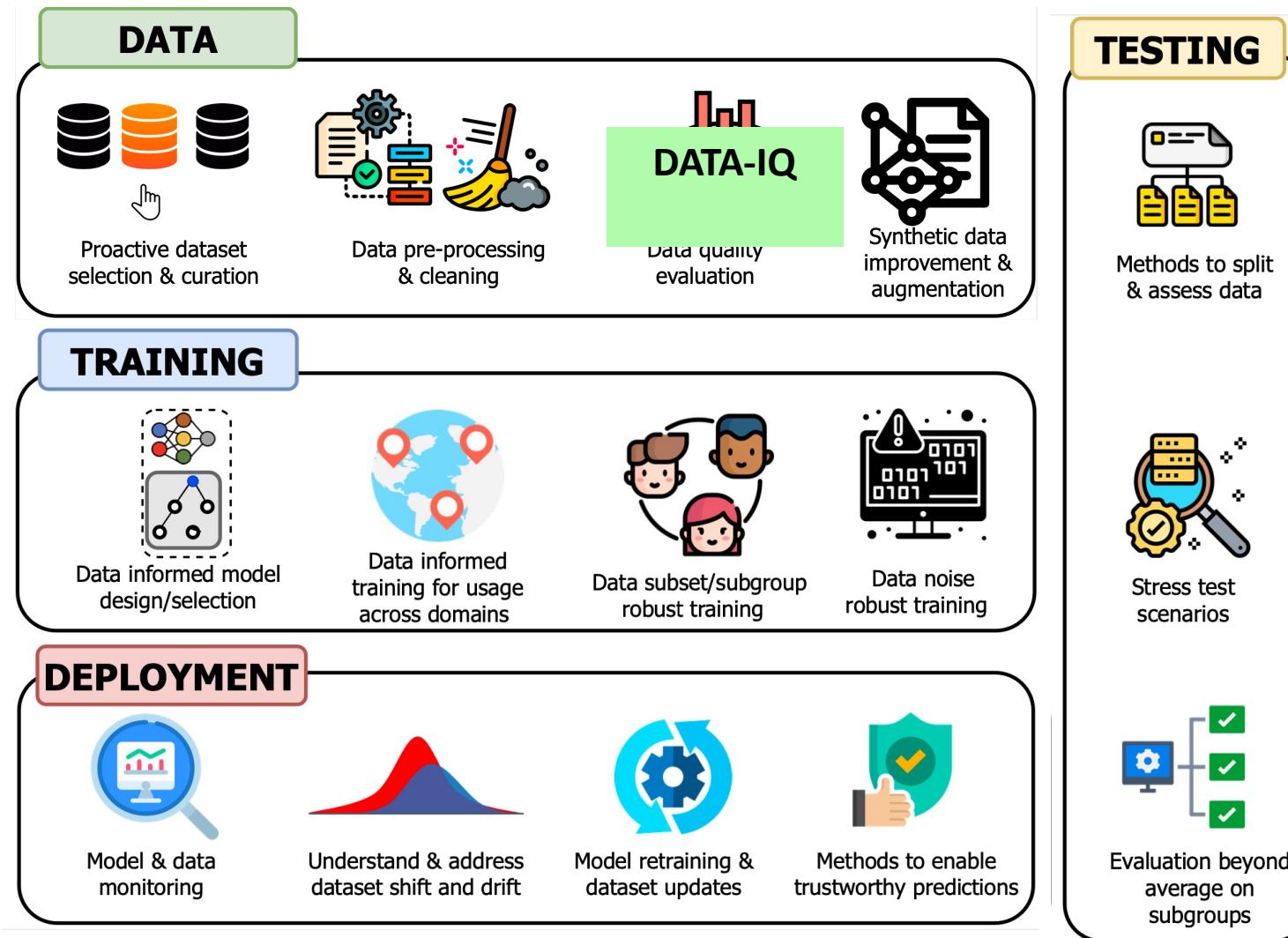
- messy, biased, noisy, erroneous
- costly to acquire
- limited
- incomplete
- changes over time

Construct Reality-Centric ML pipelines with Real-world data

Construct Reality-Centric ML pipelines with Real-world data

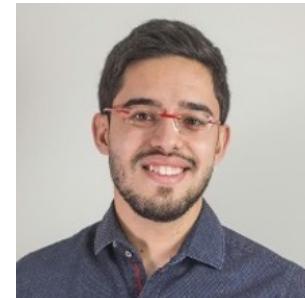


Construct Reality-Centric ML pipelines with Real-world data



vanderschaar-lab.com/
→ dc-check

DC-Check provides online & offline tools to actionably engage with data-centric considerations in ML development



Nabeel Seedat



Fergus Imrie



van_der_Schaar
LAB

vanderschaar-lab.com

UNIVERSITY OF
CAMBRIDGE

Data-IQ: Characterizing heterogeneous subgroups in tabular data

NeurIPS 2022

Nabeel Seedat, Jonathan Crabbe, Ioana Bica & Mihaela van der Schaar



Focus: Characterize the data to create higher quality training datasets!



van_der_Schaar
\ LAB

vanderschaar-lab.com



Auditing datasets



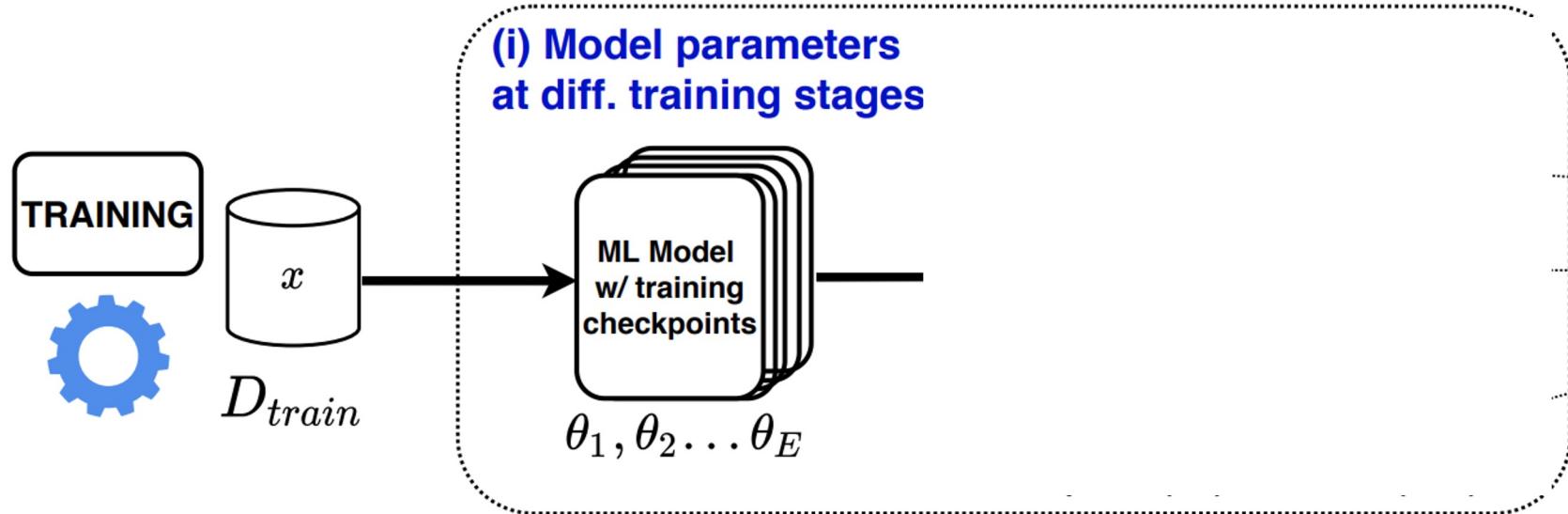
**Auditing datasets at scale
is hard & time-consuming!**

Contributions of Data-IQ

- **Data-IQ can characterize samples into subgroups based on the inherent uncertainty in data**
- **Useful tool for practitioners:** understand quality of data, guide feature acquisition and/or help judiciously sculpt datasets
- **Useful for many data modalities:** our focus is on tabular data, however, DataIQ also work for other data modalities – including images and text (see our paper)



Data-IQ pipeline



Two types of uncertainty

Data-IQ

$$v_{\text{al}} = \mathbb{E}_{\vartheta} [\mathbb{V}_{\tilde{Y}|X,\vartheta}[\tilde{Y}|X = x, \vartheta]]$$

Variability due to inherent inability to predict with certainty (linked to the **data**)

Aleatoric uncertainty

Data maps [23]

$$v_{\text{ep}} = \mathbb{V}_{\vartheta} [\mathbb{E}_{\tilde{Y}|X,\vartheta}[\tilde{Y}|X = x, \vartheta]]$$

Variability due to change in predictions via parameters (linked to the **model**)

Epistemic uncertainty



Data-IQ assesses uncertainty over training

Data-IQ

$$v_{\text{al}} = \mathbb{E}_{\vartheta} [\mathbb{V}_{\tilde{Y}|X,\vartheta}[\tilde{Y}|X = x, \vartheta]]$$

Variability due to inherent inability to predict with certainty (linked to the **data**)

EASY: PREDICTED CORRECTLY, WITH HIGH CONFIDENCE

HARD: PREDICTED INCORRECTLY, WITH HIGH CONFIDENCE

AMBIGUOUS: INHERENT AMBIGUITY,
CURRENT FEATURES INSUFFICIENT TO DISTINGUISH,
NO MATTER THE MODEL



Two types of uncertainty

Data-IQ

$$v_{\text{al}} = \mathbb{E}_{\vartheta} [\mathbb{V}_{\tilde{Y}|X,\vartheta}[\tilde{Y}|X = x, \vartheta]]$$

Variability due to inherent inability to predict with certainty (linked to the **data**)

Aleatoric uncertainty

Stratification based on data uncertainty

Data maps [23]

$$v_{\text{ep}} = \mathbb{V}_{\vartheta} [\mathbb{E}_{\tilde{Y}|X,\vartheta}[\tilde{Y}|X = x, \vartheta]]$$

Variability due to change in predictions via parameters (linked to the **model**)

Epistemic uncertainty



Differences between Data-IQ & Data Maps: Type of uncertainty matters!

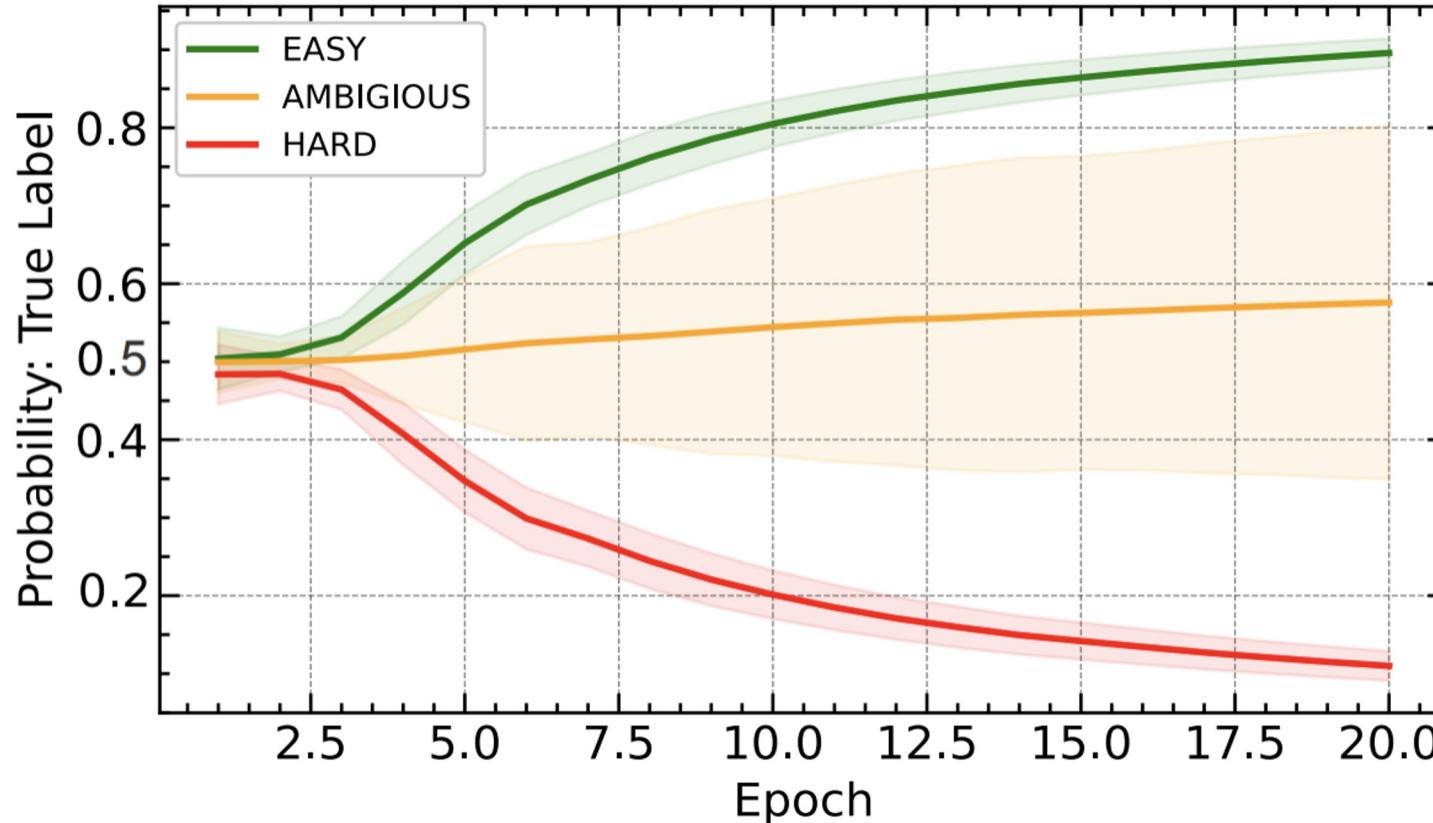


van_der_Schaar
\ LAB

vanderschaar-lab.com

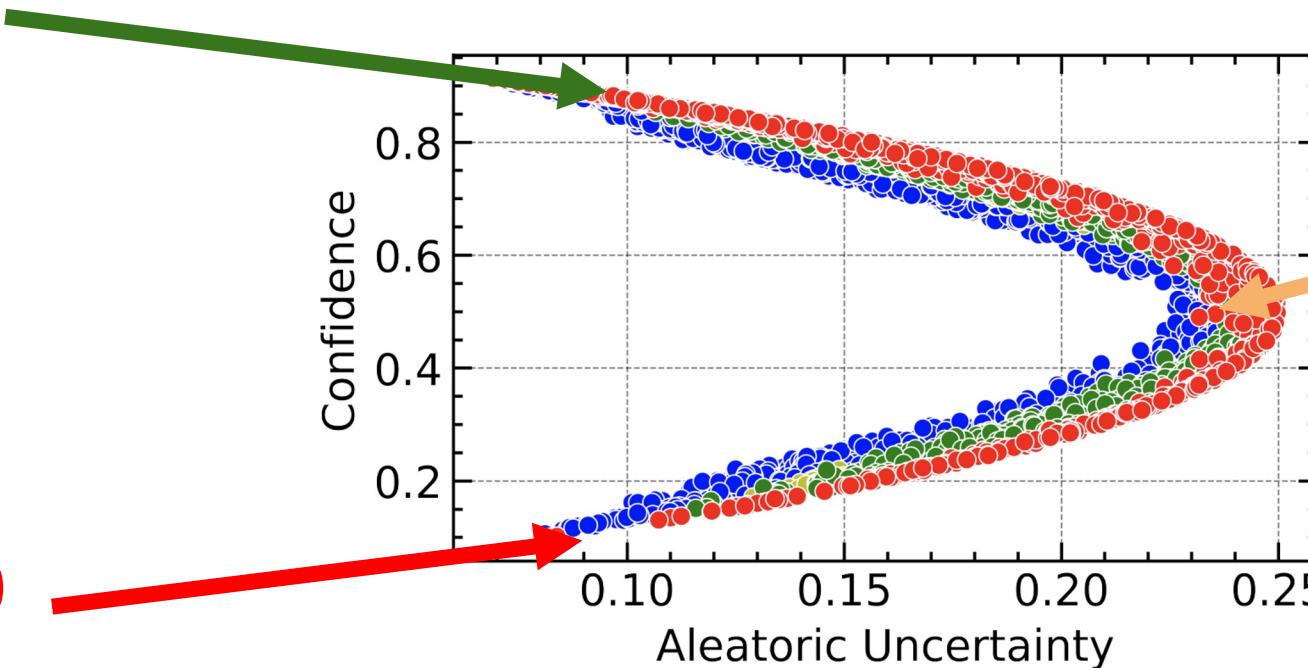
UNIVERSITY OF
CAMBRIDGE

Differences between Data-IQ & Data Maps: Type of uncertainty matters!



Differences between Data-IQ & Data Maps: Type of uncertainty matters!

EASY



AMBIGUOUS

HARD



van_der_Schaar
\LAB

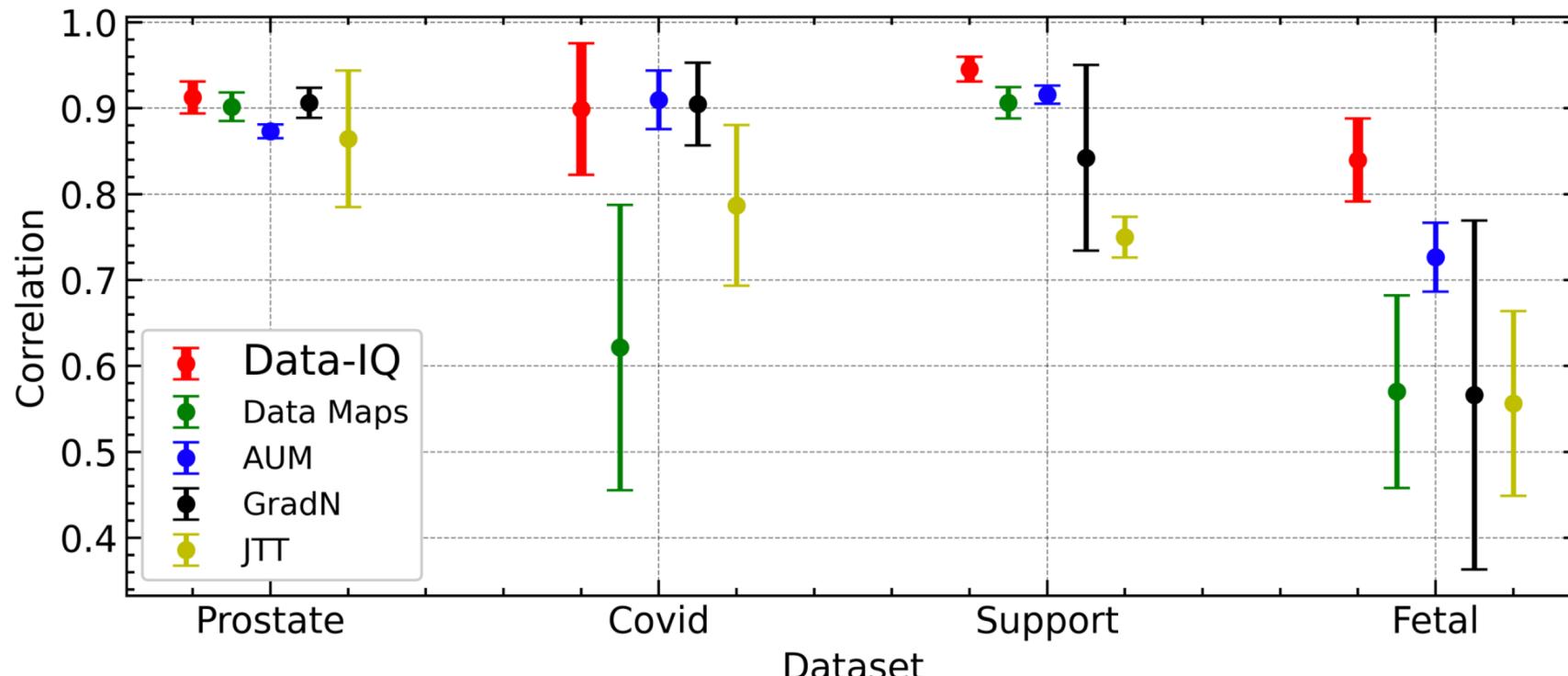
vanderschaar-lab.com

UNIVERSITY OF
CAMBRIDGE

Data-IQ: Impact

ROBUST DATA CHARACTERIZATION

Aleatoric (data) uncertainty: consistent sample characterization for similar performing models



Data-IQ: Impact

ROBUST DATA CHARACTERIZATION

Aleatoric (data) uncertainty: consistent sample characterization

Principled data collection

Utility for feature acquisition & dataset selection



van_der_Schaar
\ LAB

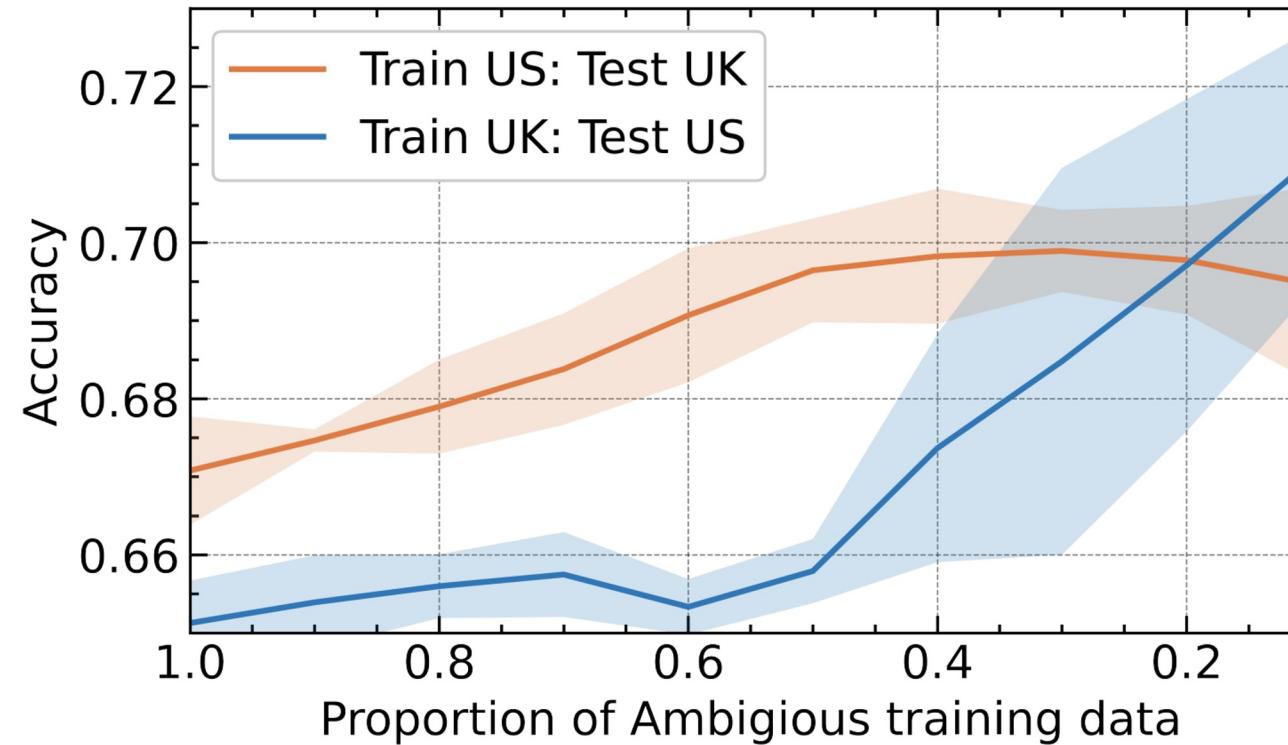
vanderschaar-lab.com

UNIVERSITY OF
CAMBRIDGE

Data-IQ: Impact

Reliable model deployment

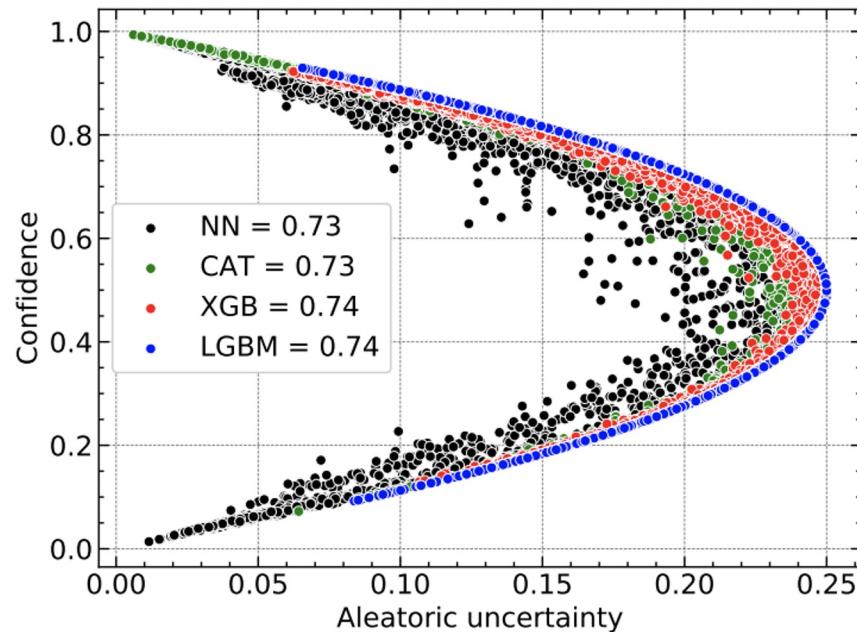
Guide sculpting of datasets for more reliable deployment of models (improved generalization)



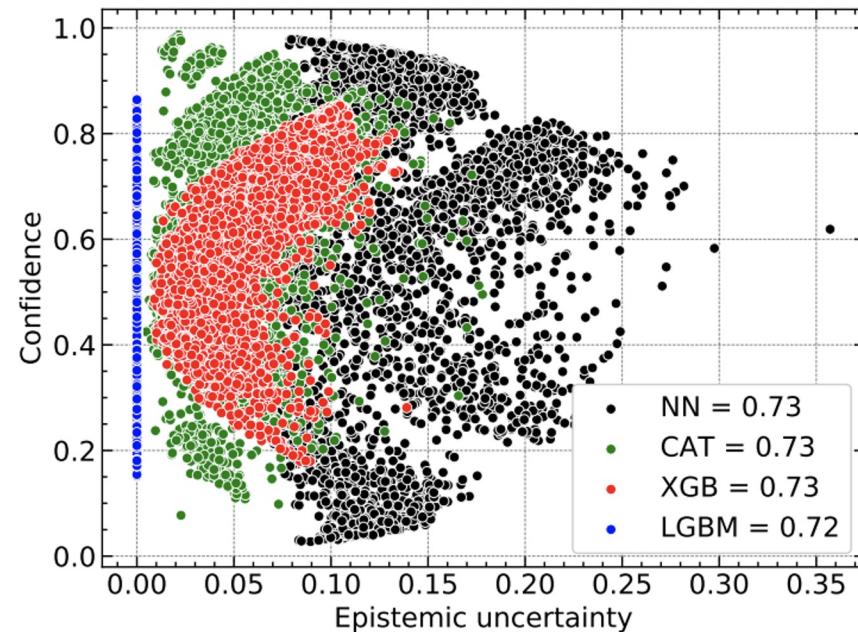
Data-IQ: Impact

Plug & Play

Data-IQ is usable with any model trained iteratively: neural nets, XGBoost etc



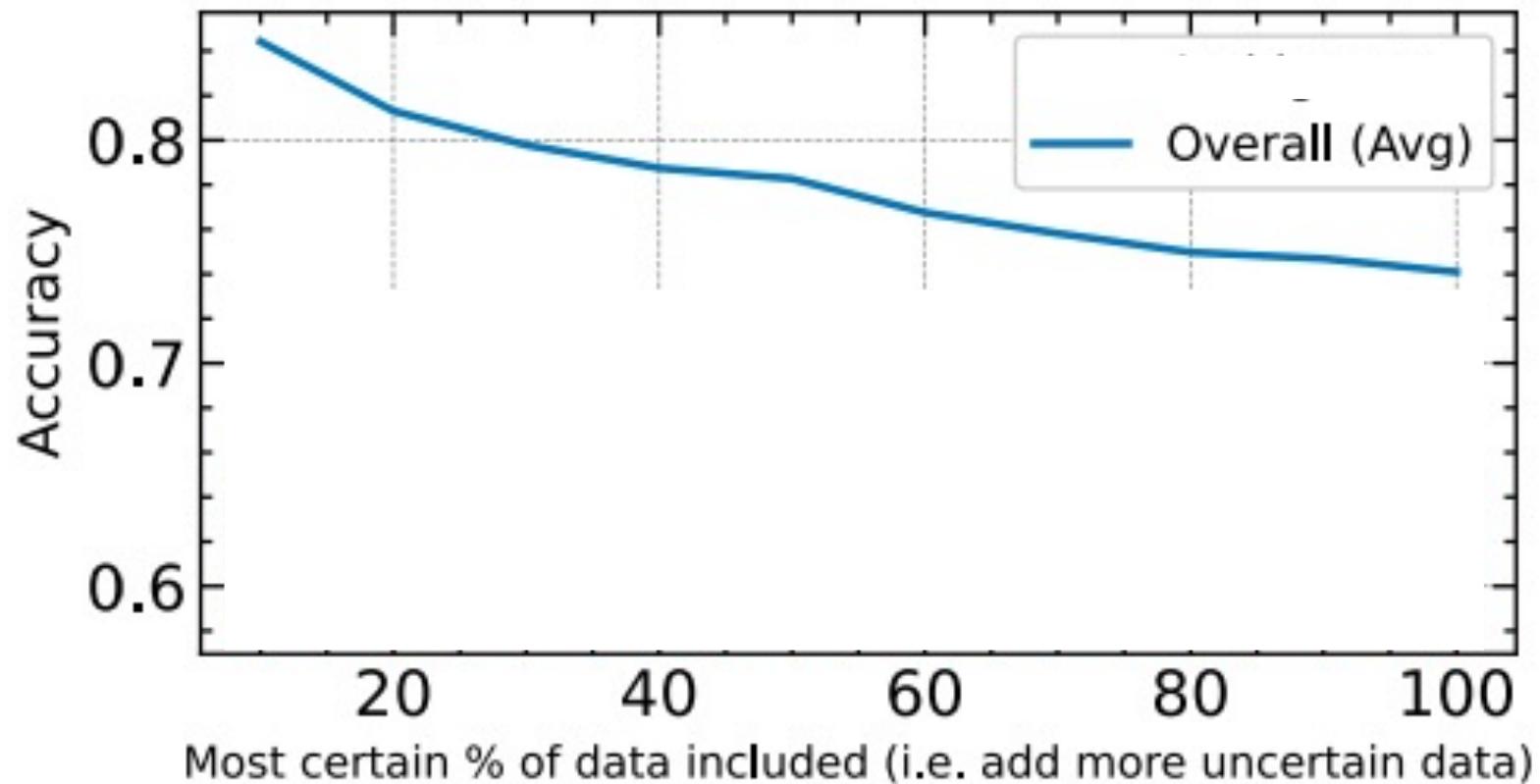
(a) Data-IQ



(b) Data Maps

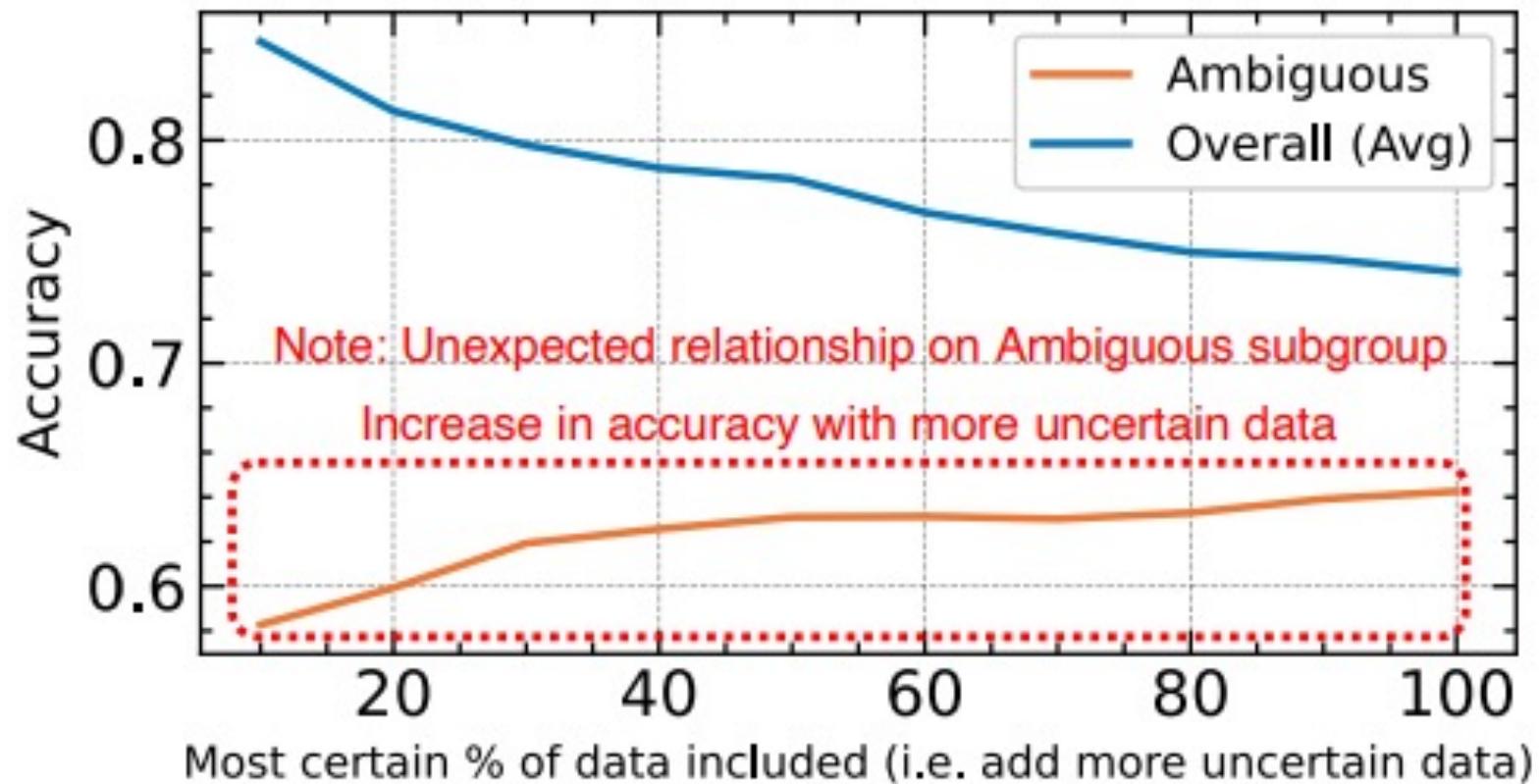
Data-IQ: Impact

Subgroup-informed usage of uncertainty estimation



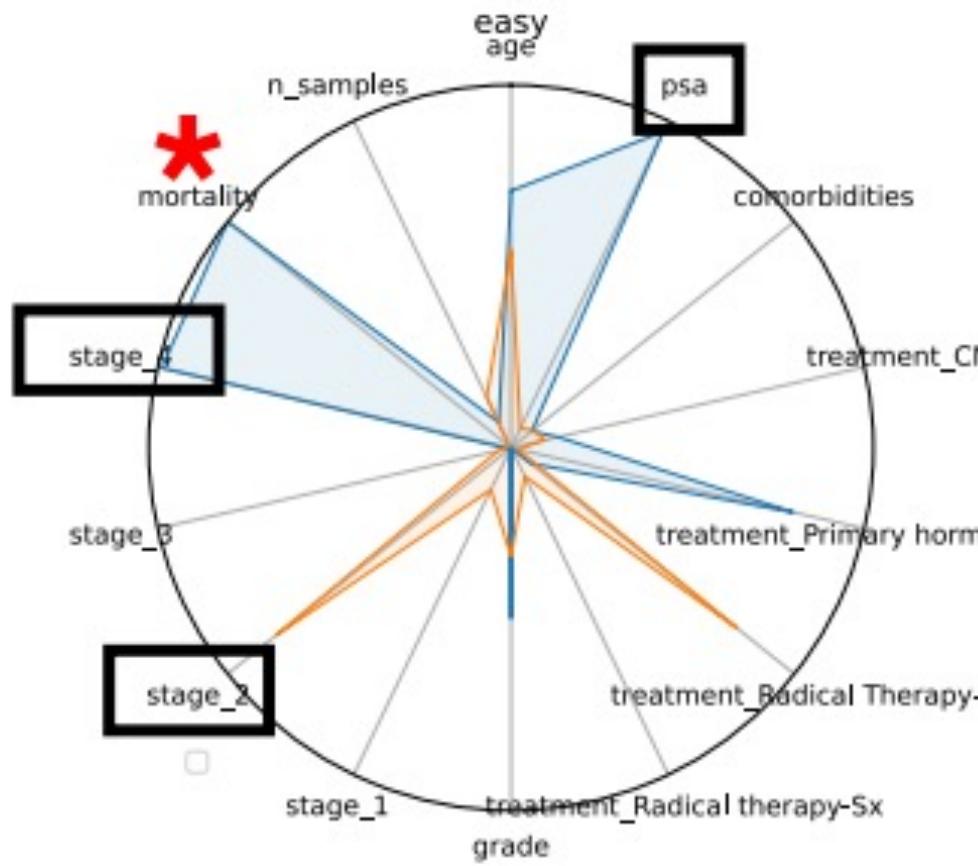
Data-IQ: Impact

Subgroup-informed usage of uncertainty estimation



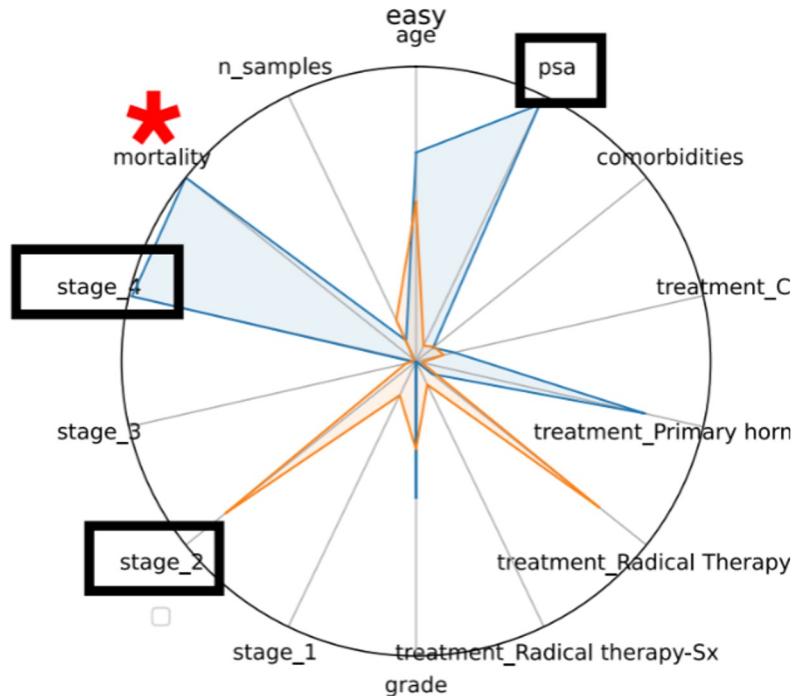
Data-IQ: Insights

Less Severe (low PSA, stage 2) = survive
Severe (high PSA, stage 4) = mortality



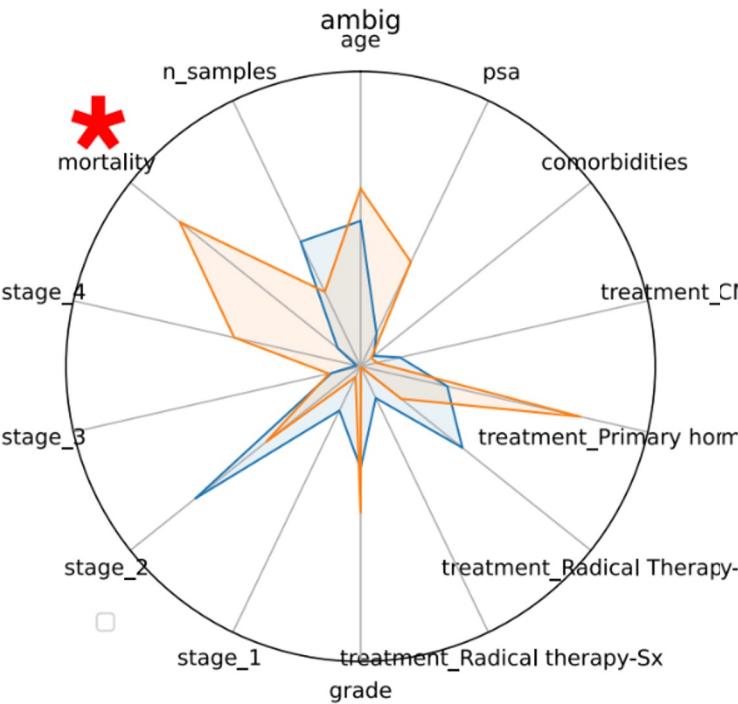
Data-IQ: Insights

Less Severe (low PSA, stage 2) = survival
Severe (high PSA, stage 4) = mortality



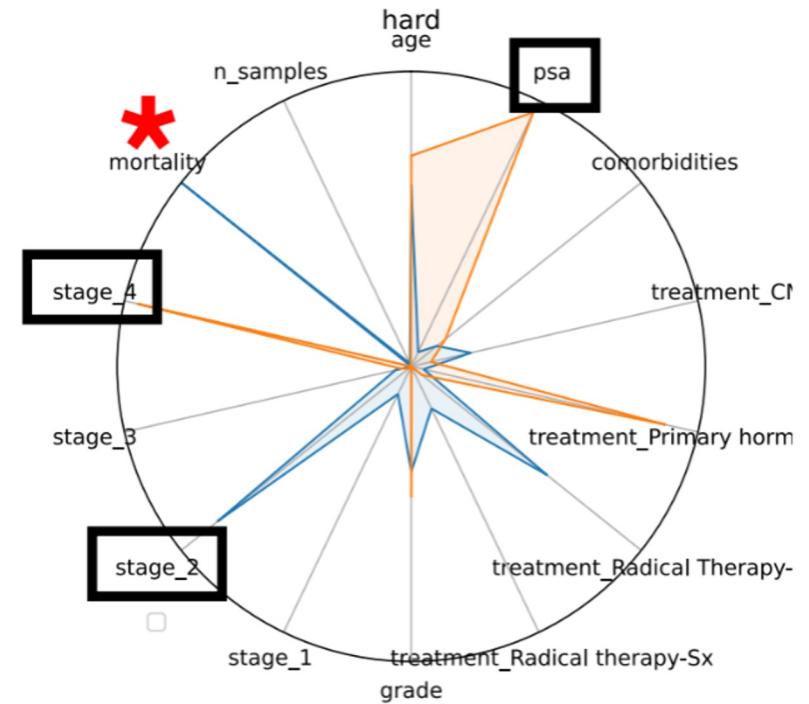
(a) EASY

Patients with similar features
BUT different outcomes



(b) AMBIGUOUS

Severe (high PSA, stage 2) = survive
Less Severe (low PSA, stage 2) = mortality



(c) HARD

Using Data-IQ on your own data

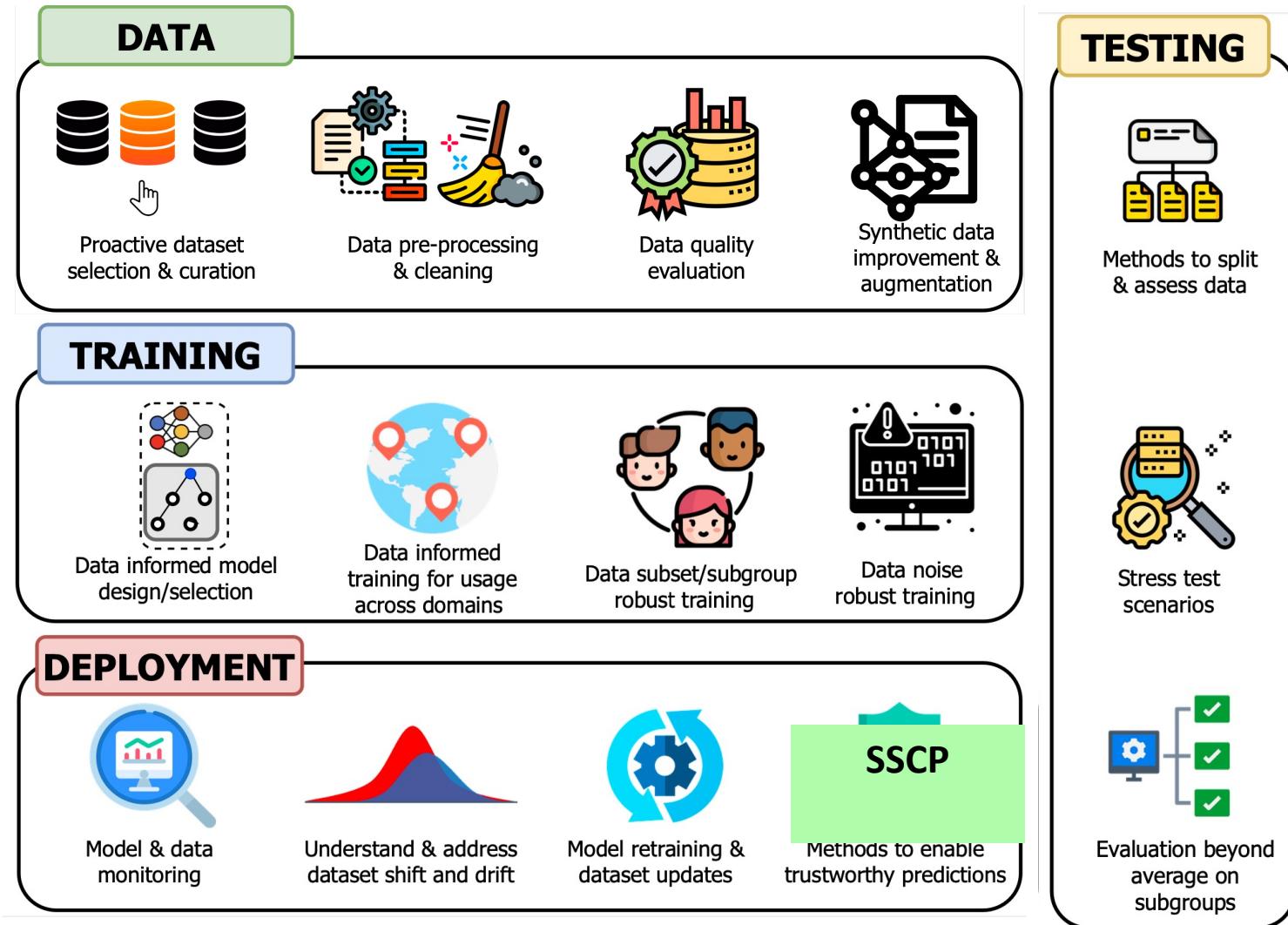
**Easily applied in your existing
model training loops**

Just 2 extra lines of code!

Plug-ins for Pytorch, Tensorflow & Sklearn



Construct Reality-Centric ML pipelines



vanderschaar-lab.com/
→ dc-check

DC-Check provides online & offline tools to actionably engage with data-centric considerations in ML development



Improving Adaptive Conformal Prediction Using Self-Supervised Learning

AISTATS 2023

Nabeel Seedat, Alan Jeffares*, Fergus Imrie & Mihaela van der Schaar*



van_der_Schaar
LAB

vanderschaar-lab.com

UNIVERSITY OF
CAMBRIDGE

Self- and Semi-Supervised Learning for tabular data

VIME (NeurIPS 2020)

First self- and semi-supervised method for tabular data

Self-supervised pretext tasks:

- Feature vector estimation
- Mask vector estimation

Masking process can be used for semi-supervised learning

SEFS (ICLR 2022)

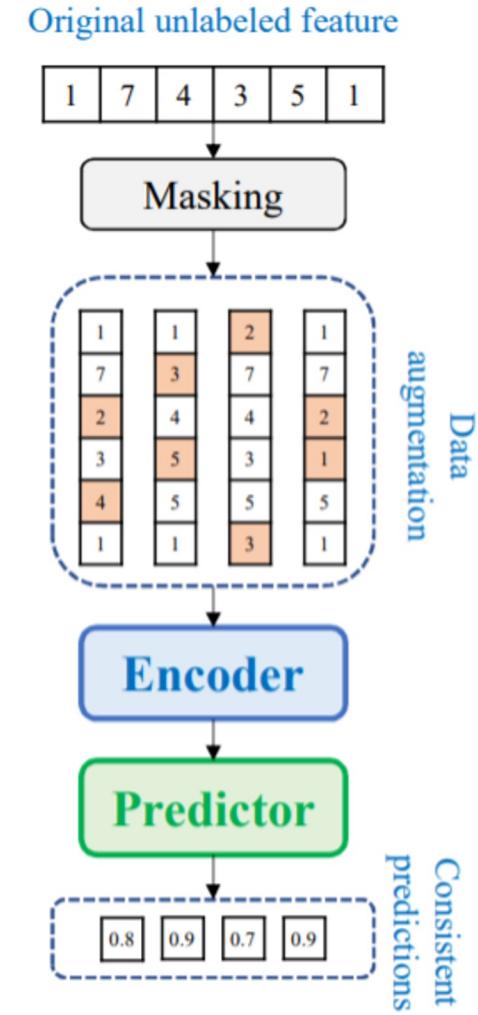
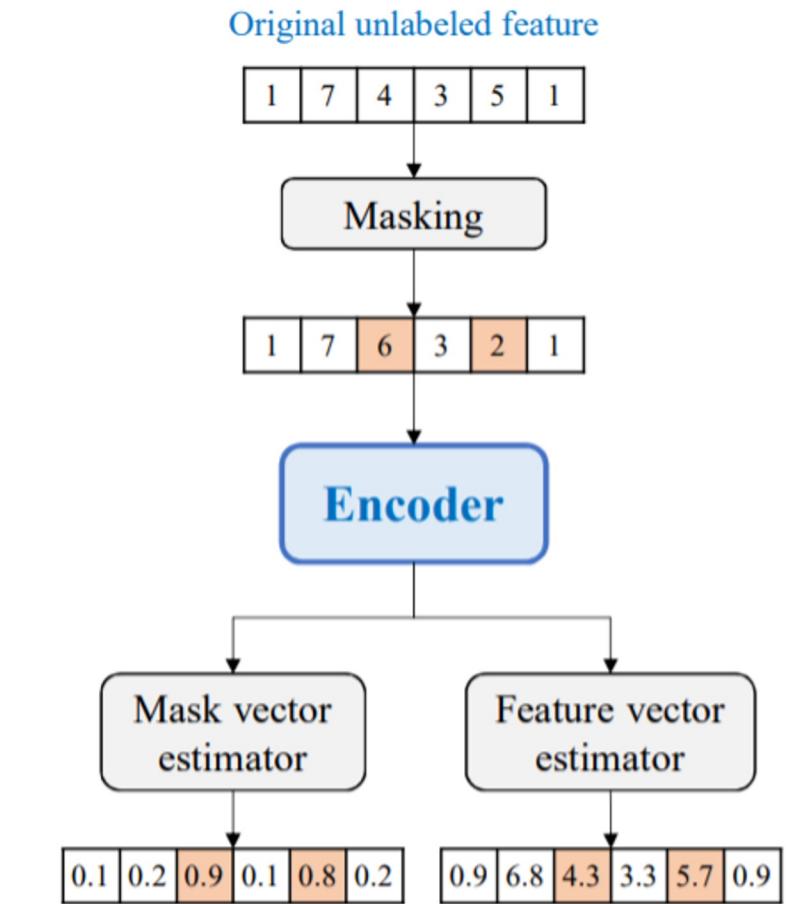
Improvements to masking process

→ Independent vs. Correlated masking

Extension to feature selection



van_der_Schaar
\\ LAB



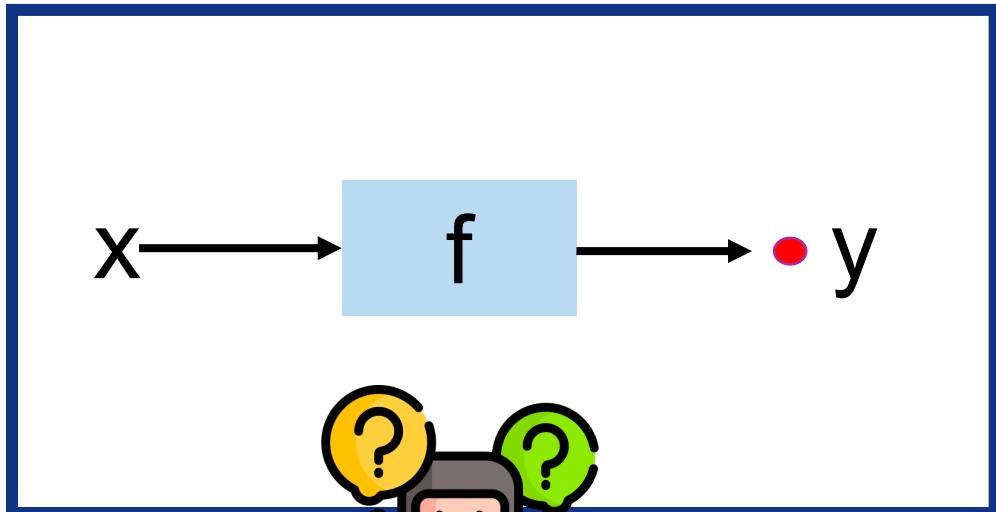
vanderschaar-lab.com



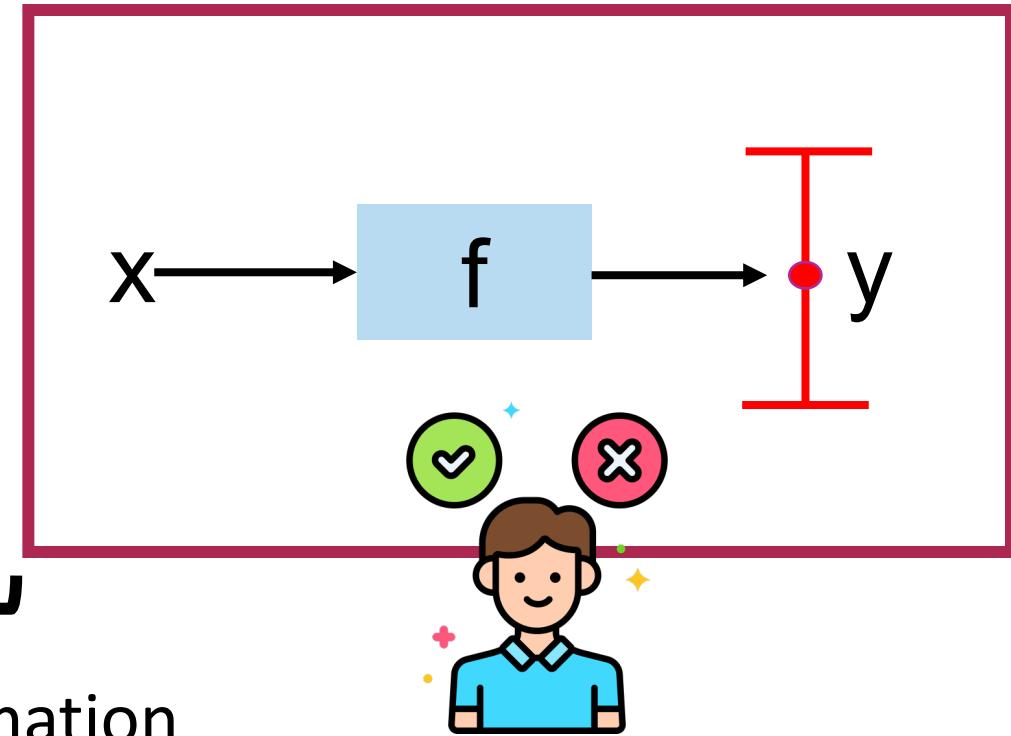
UNIVERSITY OF
CAMBRIDGE

Trustworthy ML

Typical ML: Point Estimates



Trustworthy ML: Point Estimates + Uncertainty



Uncertainty estimation

This work: Conformal Prediction

vanderschaar-lab.com



van_der_Schaar
LAB



UNIVERSITY OF
CAMBRIDGE

Conformal Prediction (CP) – Vovk et al (2005)



Prediction Intervals, with guarantees on coverage

i.e. true value will be contained within the interval with a specified probability

$$\mathbb{P}\{Y \in \hat{C}(X)\} \geq 1 - \alpha.$$

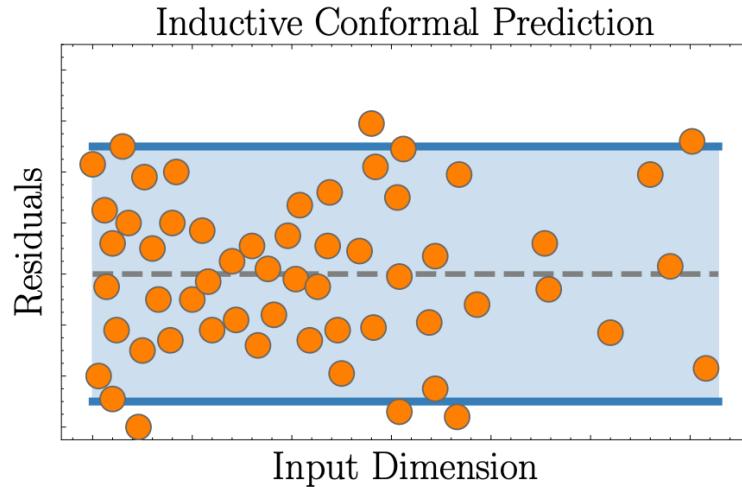


Powerful: Model-agnostic & distribution-free assumptions

Paradigms of Conformal Prediction

Inductive CP:

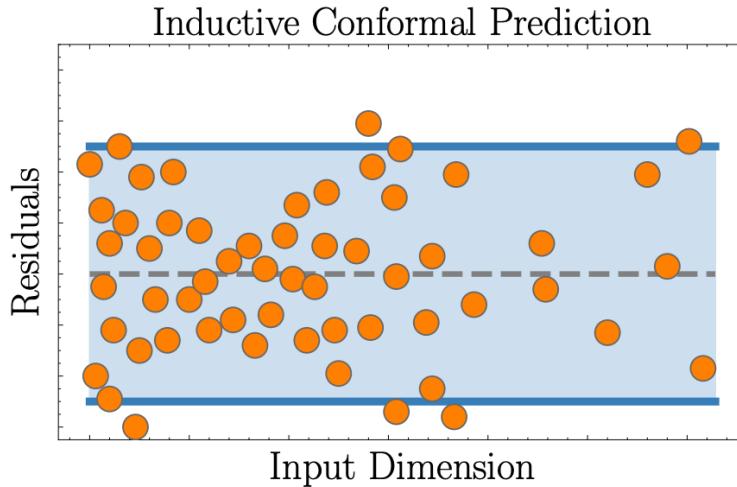
Constant width intervals



Paradigms of Conformal Prediction

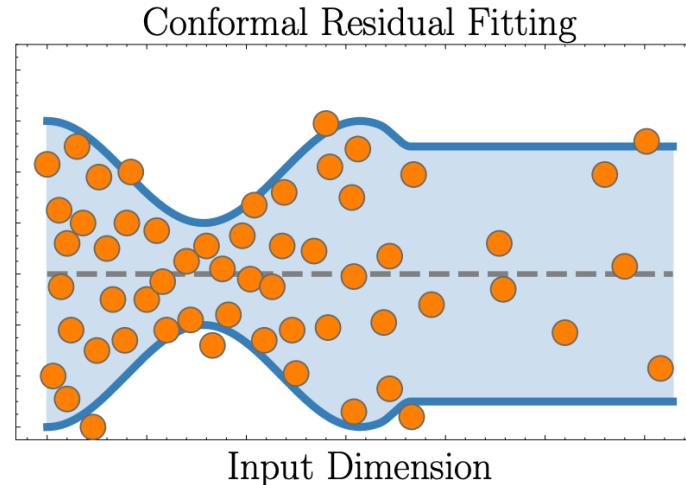
Inductive CP:

Constant width intervals



Conformal Residual Fitting:

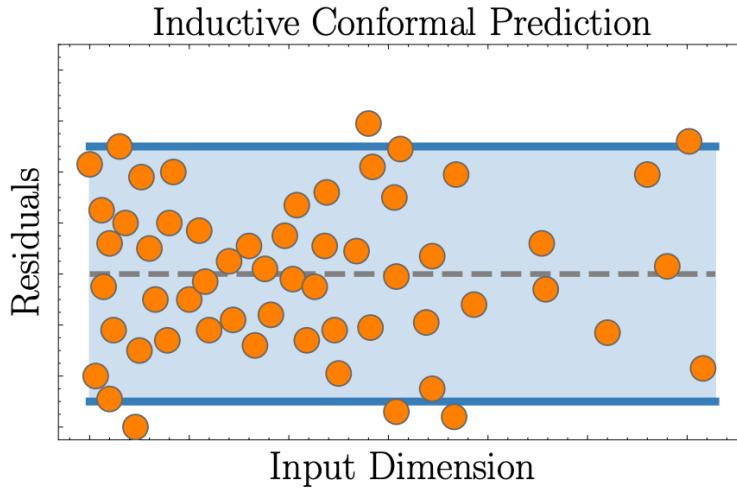
Adaptive but inefficient
in some regions



Paradigms of Conformal Prediction

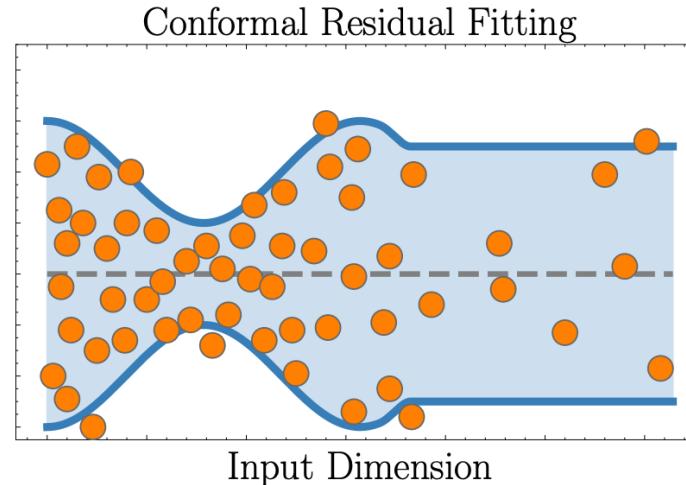
Inductive CP:

Constant width intervals



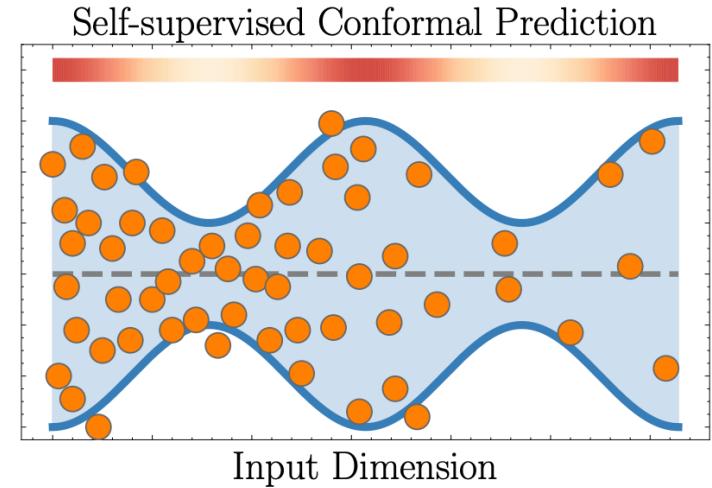
Conformal Residual Fitting:

Adaptive but inefficient
in some regions



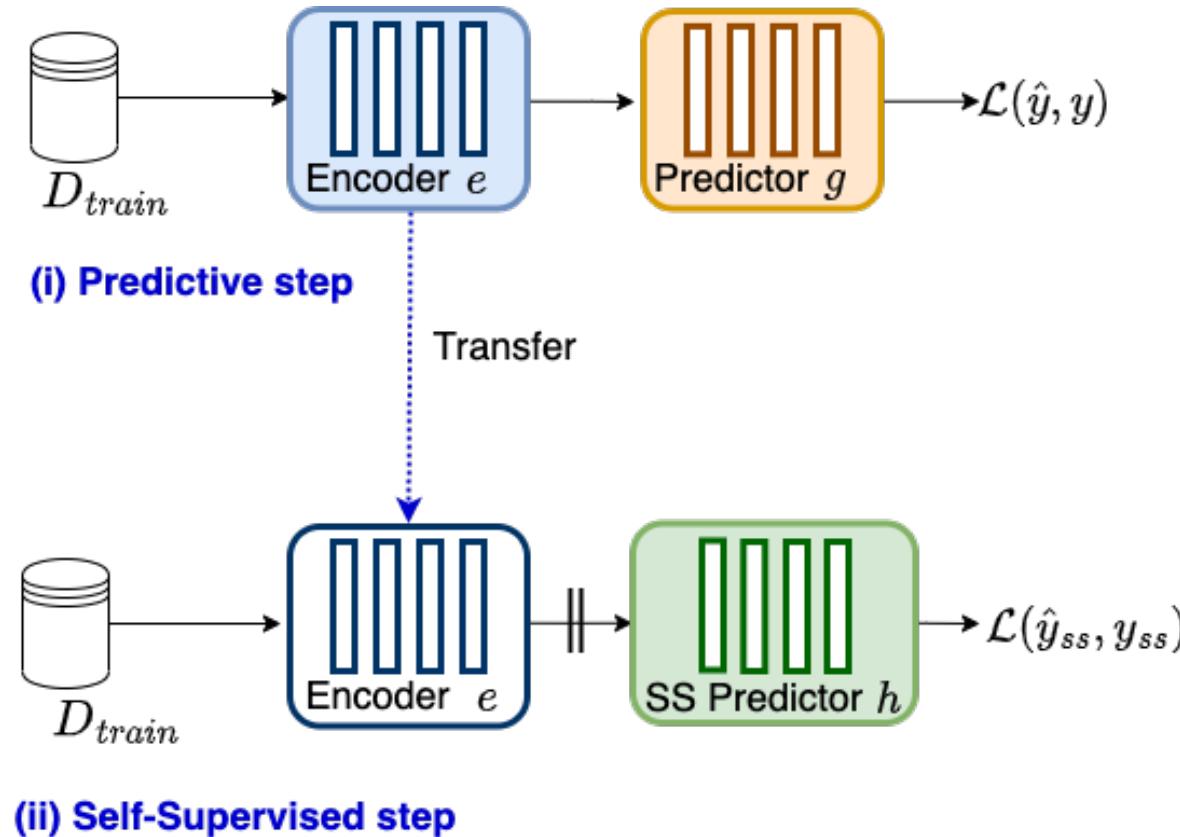
Self-supervised CP (SSCP):

Improved efficiency,
particularly for sparser regions



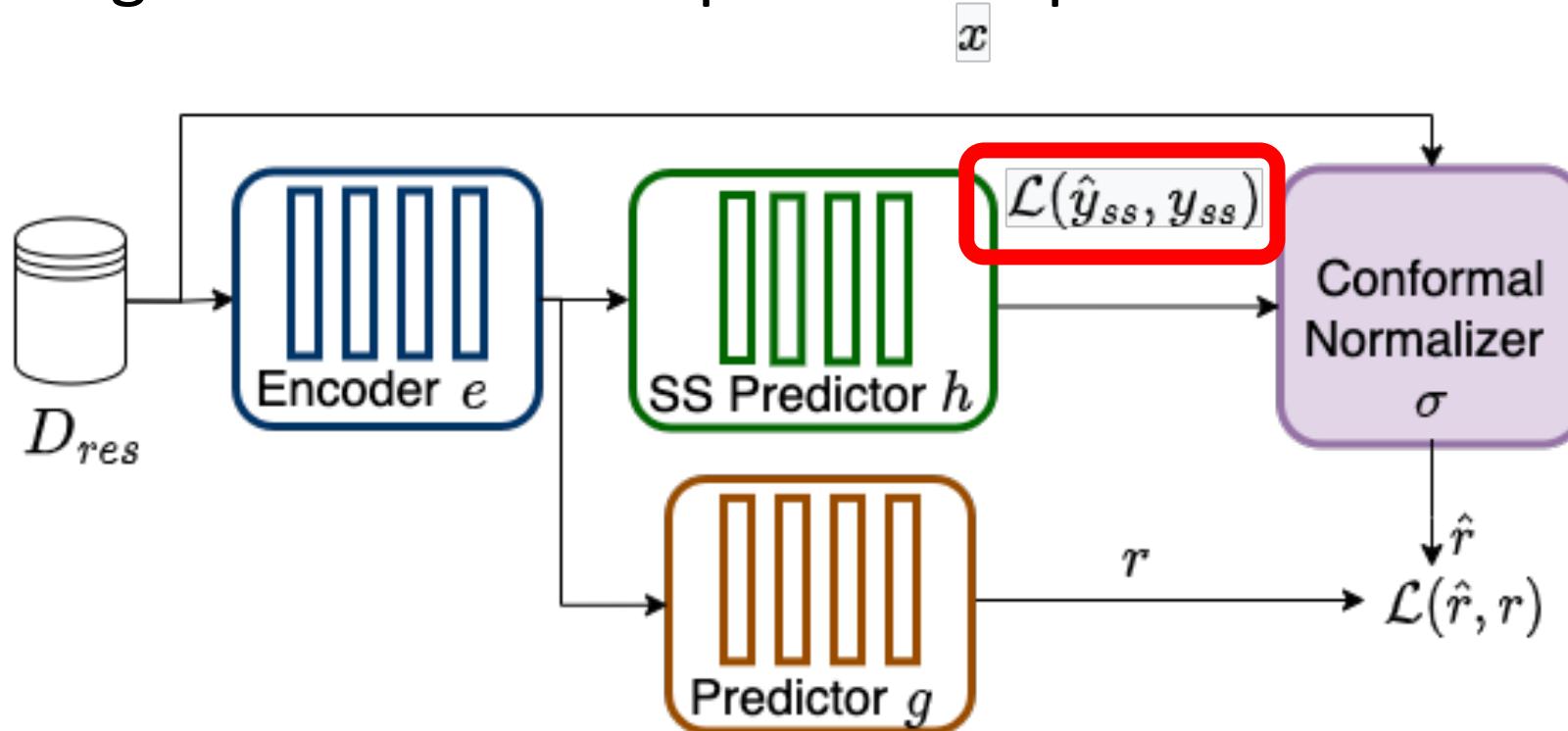
SSCP: Predictive & Self-Supervised phase

- Train predictor $f = e \circ g$
- Transfer encoder e for self-supervised task
- Train self-supervised predictor h



SSCP: Conformal Normalized Phase

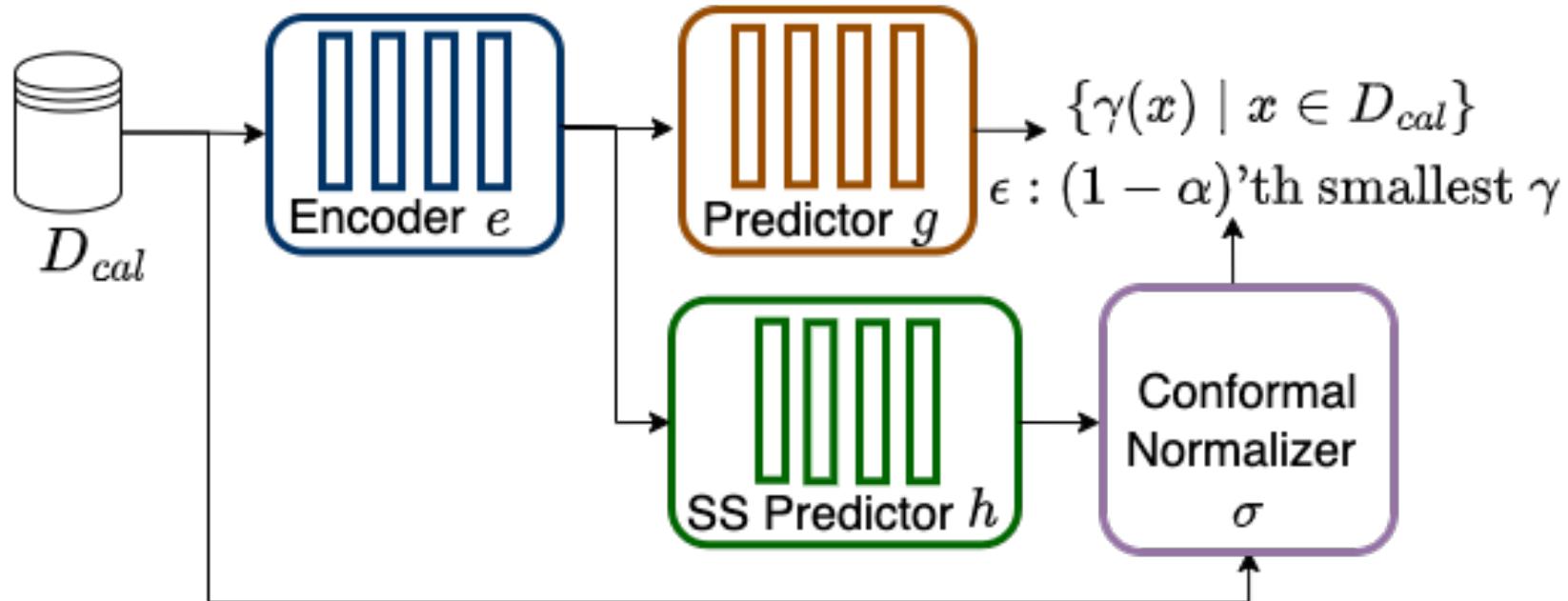
- Augment with self-supervised input



(S2) Conformal Normalizer phase

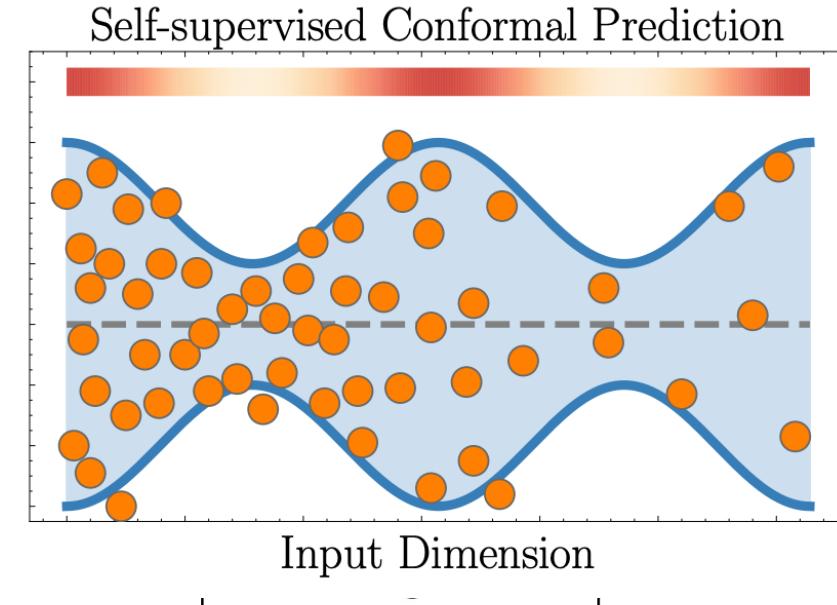
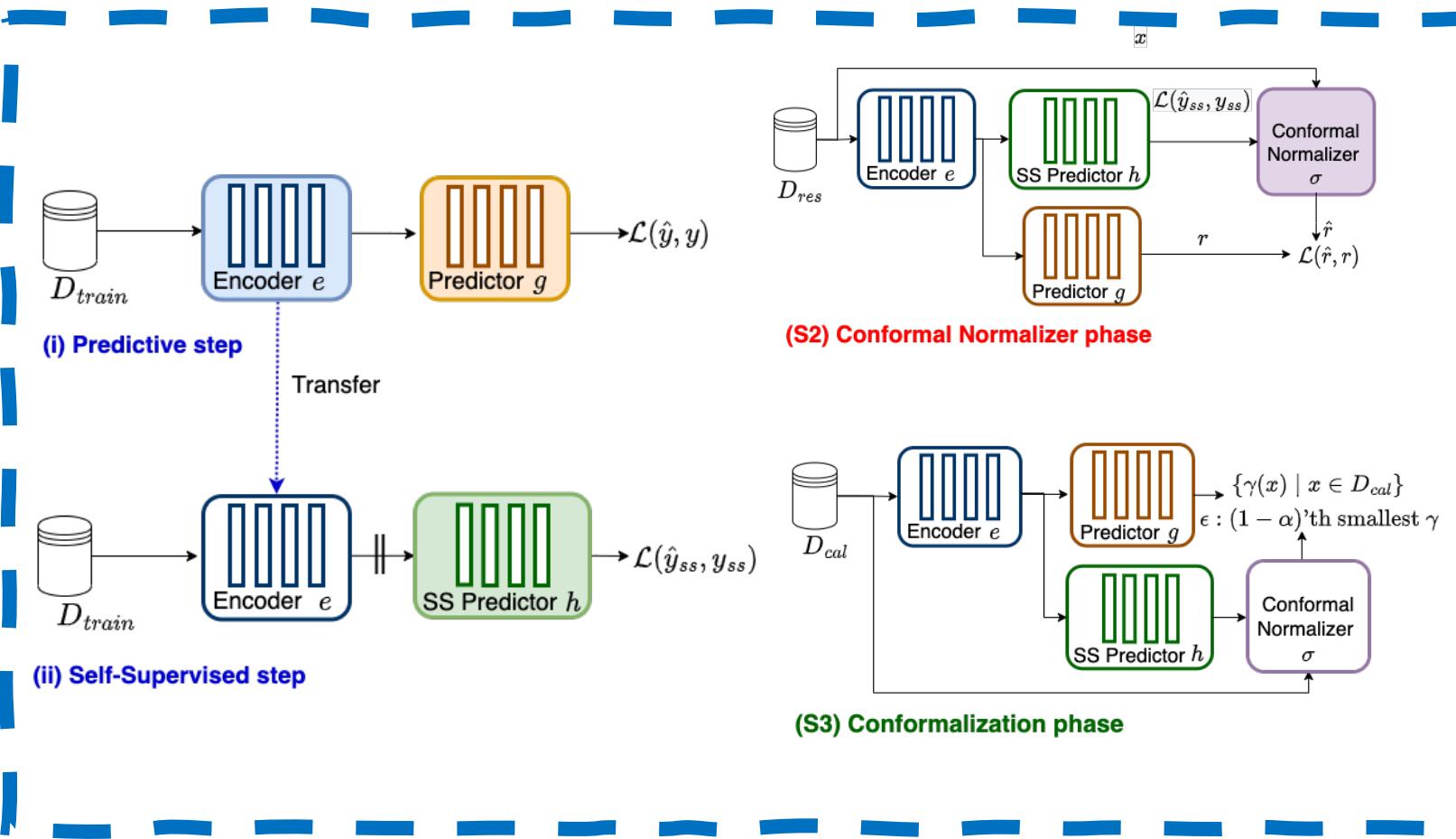
SSCP: Conformalization phase

- Conformalize using the calibration set



(S3) Conformalization phase

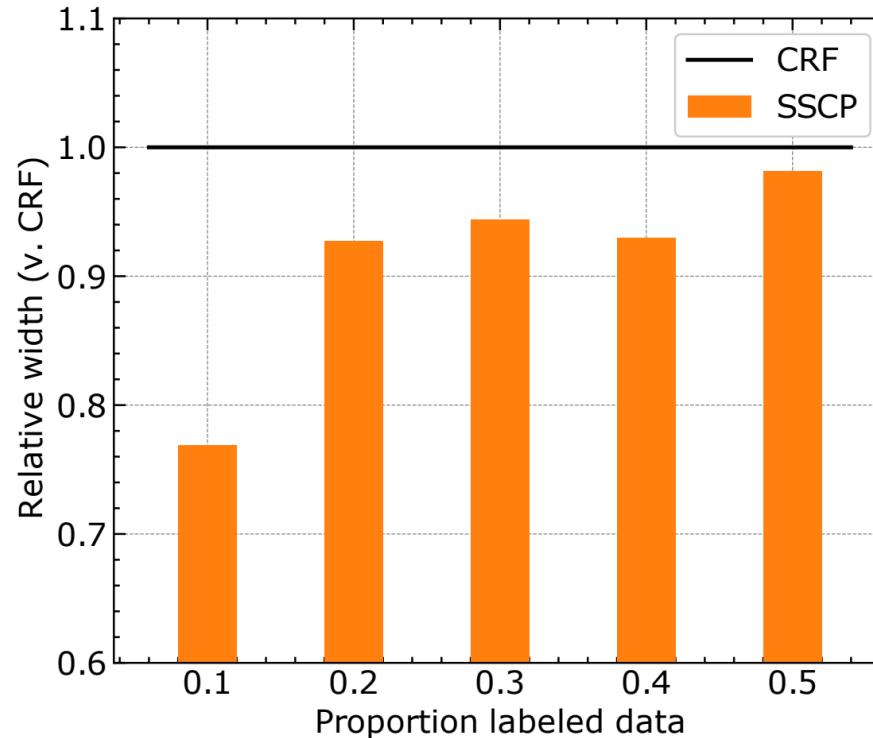
SSCP produces Adaptive Intervals



Adaptive conformal prediction intervals

SSCP produces more Adaptive Intervals

UNLABELED DATA*

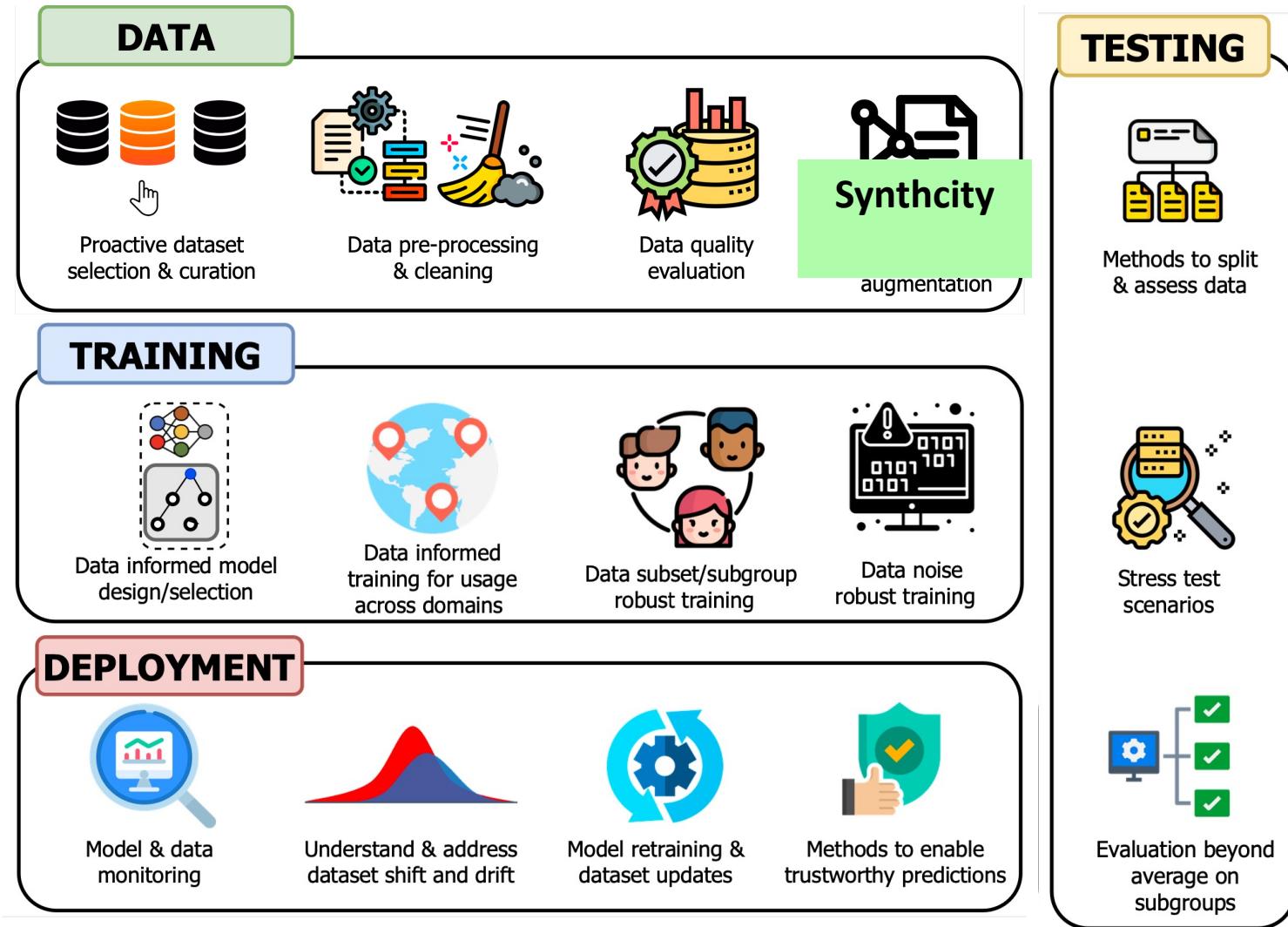


LABLED DATA*

Dataset	Metric	CRF	SSCP
concrete (n=1030)	Avg.Width	0.768	0.743
	Avg.Deficit	0.099	0.098
	Avg.Excess	0.263	0.253
community (n=1994)	Avg.Width	2.602	2.462
	Avg.Deficit	0.355	0.361
	Avg.Excess	1.030	0.967
star (n=2161)	Avg.Width	0.293	0.263
	Avg.Deficit	0.036	0.031
	Avg.Excess	0.120	0.100
bike (n=10886)	Avg.Width	0.720	0.690
	Avg.Deficit	0.164	0.161
	Avg.Excess	0.244	0.232
Blog data (n=52397)	Avg.Width	3.474	3.360
	Avg.Deficit	3.084	3.155
	Avg.Excess	1.292	1.227
Facebook (n=81311)	Avg.Width	1.917	1.860
	Avg.Deficit	1.956	1.998
	Avg.Excess	0.584	0.554

* SSCP improves both Conformal Residual Fitting (CRF) & Conformalized Quantile Regression (CQR)

Construct Reality-Centric ML pipelines



vanderschaar-lab.com/
→ dc-check

DC-Check provides online & offline tools to actionably engage with data-centric considerations in ML development



Why is my data not good enough?



Relevancy – is my data fit for purpose?

- Dataset does not contain all **relevant variables**
- Dataset does not faithfully represent the **population of interest**
- **Historical** data may not be **future-proof**



Quality – how accurate is my data?

- Measurement, recording, and linking **errors**
- **Missing** values (not at random)
- **Bias** (unfair data)



Privacy – can I easily access or share the data?

- Personal **identifiable** and **sensitive** information



van_der_Schaar
\ LAB

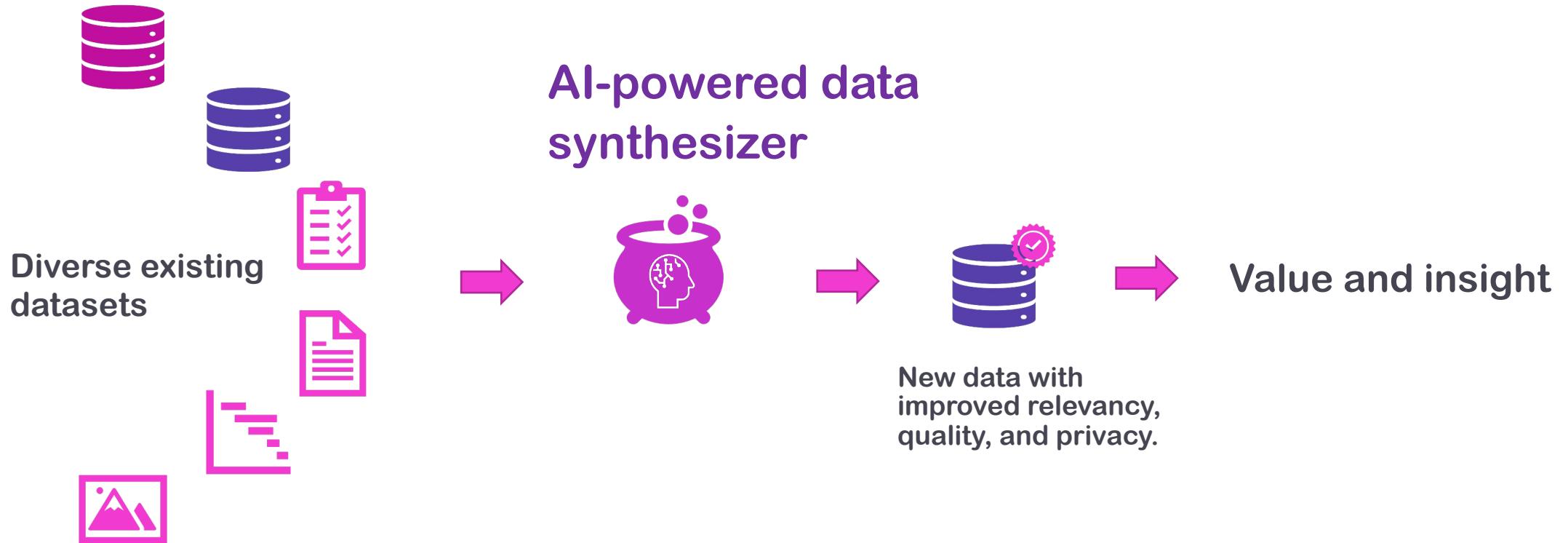
vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE

Data synthesis promises to bridge the data gap

Turns **the data you have** into **the data you need!**



van_der_Schaar
LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE

Synthetic Data: Introduced the idea in 2017



Synthetic Data: What is it and why do we need it?

Mihaela van der Schaar

John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine, University of Cambridge
Director, Cambridge Center for AI in Medicine
Turing Faculty Fellow, The Alan Turing Institute

A screenshot of a YouTube video player. The video title is "Play (k) van_der_Schaar \ LAB". Below the title, the University of Cambridge logo is displayed. The video progress bar shows 4:06 / 1:12:10. The video content starts with "Introducto...". To the right of the video player are social media links: an envelope icon for mv472@cam.ac.uk, a Twitter icon for @MihaelaVDS, and a LinkedIn icon for linkedin.com/in/mihaela-van-der-schaar/. There are also standard YouTube controls for play/pause, volume, and full screen.

Revolutionizing Healthcare - Synthetic Data in Healthcare



van der Schaar Lab
1.32K subscribers

Subscribe



Share

...



van_der_Schaar
\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE

Synthetic data tutorial – ICML 2021

“Synthetic Healthcare Data Generation and Assessment: Challenges, Methods, and Impact on Machine Learning”

vanderschaar-lab.com/
→ Tutorials

→ ICML 2021 tutorial: Synthetic healthcare
data generation and assessment

The screenshot shows a webpage for an ICML 2021 tutorial. At the top, there are 'EVENTS' and 'VIDEO' tabs. The title 'ICML 2021 tutorial: Synthetic healthcare data generation and assessment' is displayed. Below the title, it says 'This ICML tutorial, entitled "Synthetic Healthcare Data Generation and Assessment: Challenges, Methods, and Impact on Machine Learning," was given by Mihaela van der Schaar and Ahmed Alaa on July 19, 2021.' There is a video player with the text 'ICML 2021 - tutorial on synthetic data' and 'ICML 2021 Tutorial on synthetic data'. Below the video player, the names 'Mihaela van der Schaar and Ahmed Alaa' are listed. Logos for 'van_der_Schaar \ LAB', 'ICML International Conference On-Machine Learning', and 'UNIVERSITY OF CAMBRIDGE' are present. A button at the bottom right says 'Full slide deck (7 MB)'.



van_der_Schaar
\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE

Synthetic data tutorial – AAAI 2023

“Innovative Uses of Synthetic Data Tutorial”

Synthetic data fixes the issues with real data

vanderschaar-lab.com/
→ Tutorials

→ AAAI-23: Synthetic Data Tutorial



van_der_Schaar
\\ LAB

vanderschaar-lab.com

AAAI EVENTS NEWS

AAAI-23: Synthetic Data Tutorial

 Mihaela van der Schaar  Zhaozhi Qian  January 5, 2023  4 min read

This [AAAI tutorial](#) will be presented by [Mihaela van der Schaar](#) and [Zhaozhi Qian](#) on [Wednesday, 8 February 2023](#) 2 – 6 pm EST. This is a hybrid event (in person/online) you can register for [here](#).

Title

Innovative Uses of Synthetic Data Tutorial

About

One of the biggest barriers to AI adoption is the difficulty to access high quality training data. Synthetic data has been widely recognised as a viable solution to this problem. It allows sharing, augmenting and de-biasing data for building performant and socially responsible AI algorithms. However, despite the significant progress in the theory and algorithm, the community still lacks a unified software that enables practical data sharing and access with synthetic data.

This lab aims to bridge this gap by introducing *synthcity*, an open source Python library that implements an array of cutting edge synthetic data generators to address the problems of data generation due to its commonality in various applications.



UNIVERSITY OF
CAMBRIDGE

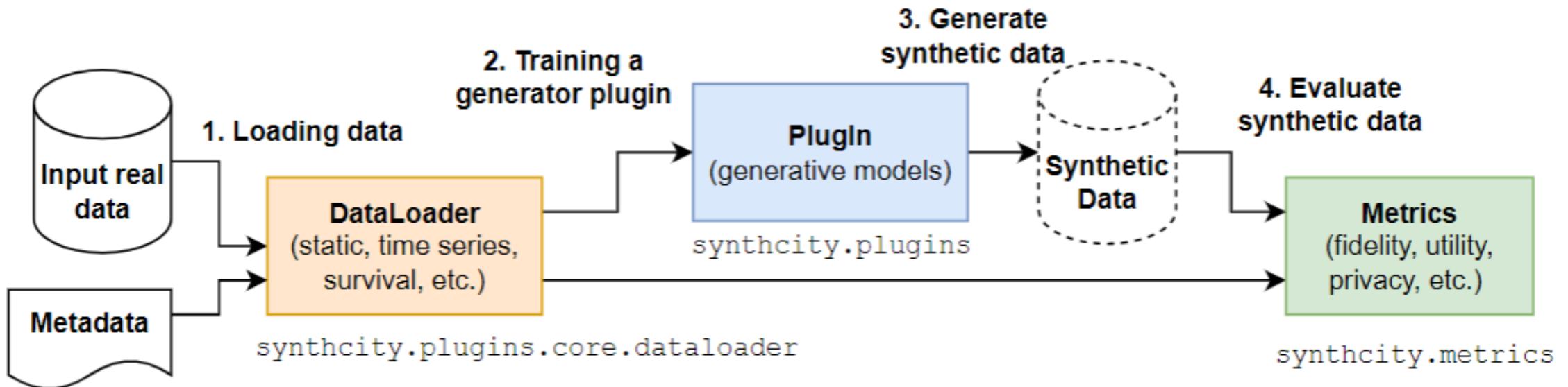
Synthetic data fixes the issues with real data

Synthetic data is a unified way to address various issues with real data

- Contains preexisting bias → Creating debiased synthetic data
- Small sample size → Augmenting with synthetic data
- Collected from different domains → Transfer and domain adaptation with synthetic data
- Future-proof → Forward looking data and future scenarios



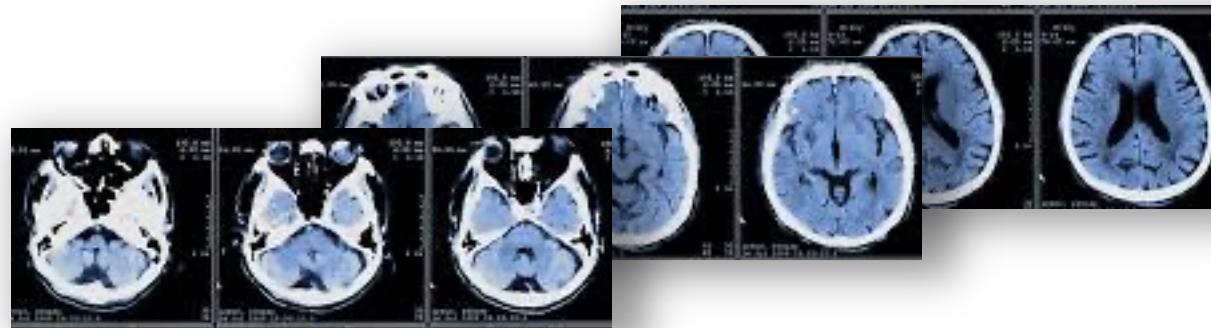
Synthcity: the most comprehensive open-source synthetic data generation and evaluation library – focused on tabular data



<https://www.vanderschaar-lab.com/synthcity-and-using-synthetic-data/>

Synthcity: Contains metrics for measuring the quality of generative models/synthetic data?

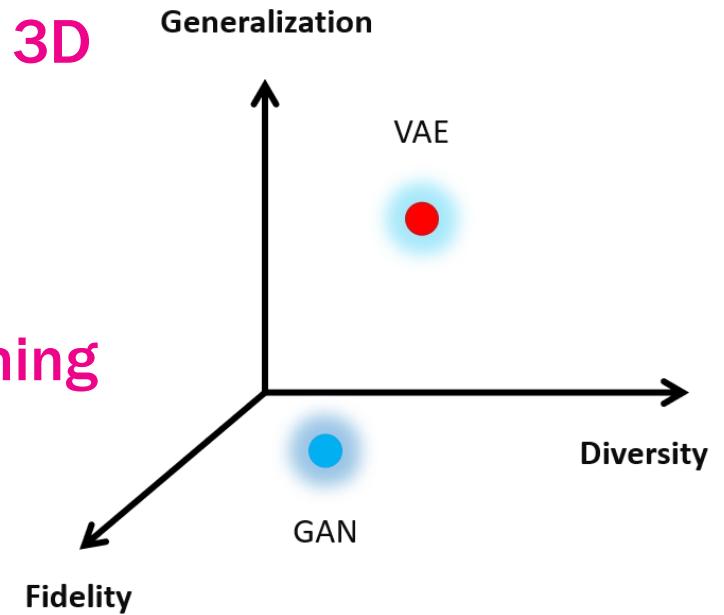
- We have followed the recipe and generated the desired synthetic data...



- How do we know if the synthetic data is of a high quality? What does “quality” mean?
 - Inspect individual samples? We may have generated millions!
 - Train on synthetic, test on real? Does not tell us the full picture...
- What makes a good synthetic dataset? How do we define “quality”?

Different ways in which a generative model may fail

- A single-dimensional metric is not enough...
- Every model's performance can be viewed as a point in a 3D space
 - **Fidelity:** How “good” the synthetic samples are?
 - **Diversity:** How much of the real data is covered?
 - **Generalization:** How often does the model copy training data?
- Need probabilistic, interpretable, multi-dimensional quantities



A. Alaa, B. van Breugel, E. Saveliev, vdS,
**How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and
Auditing Generative Models, ICML 2022, (in archive since 2021)**

Reality-centric AI – Use of Real-world Data



- messy, biased, noisy, erroneous
- costly to acquire
- limited
- incomplete
- changes over time

Construct Reality-Centric ML pipelines with Real-world data

Data-Centric AI: Resources

vanderschaar-lab.com/
→ data-centric-ai

The screenshot shows the DC-CHECK web application. At the top, there's a navigation bar with links: DC-CHECK, WHAT IS DATA-CENTRIC AI, ABOUT THE TOOL, RESOURCES, USE DC-CHECK →, and PAPER. Below the navigation is a large diagram titled "DC-CHECK" showing a checklist across four stages: DATA, TRAINING, TESTING, and DEPLOYMENT. The DATA stage includes icons for proactive dataset selection & curation, data pre-processing & cleaning, data quality evaluation, and synthetic data improvement. The TRAINING stage includes icons for data-informed model design/selection, data-informed training for usage across domains, data subset/subgroup robust training, and data noise robust training. The TESTING stage includes icons for methods to split & assess data. The DEPLOYMENT stage includes icons for model & data monitoring, understanding and addressing dataset shift and drift, and methods to enable trustworthy predictions. Below the diagram, a section titled "DATA" contains a question and answer form. The question is "1. Q1: How did you select, collect or curate your dataset?". The answer area lists three bullet points: "Have you conducted forensics on the dataset (i.e. provenance)?", "Did you assess the pertinence of the dataset for the task?", and "Is your dataset curation a once-off?". There is also a note about exporting the tool as a PDF.



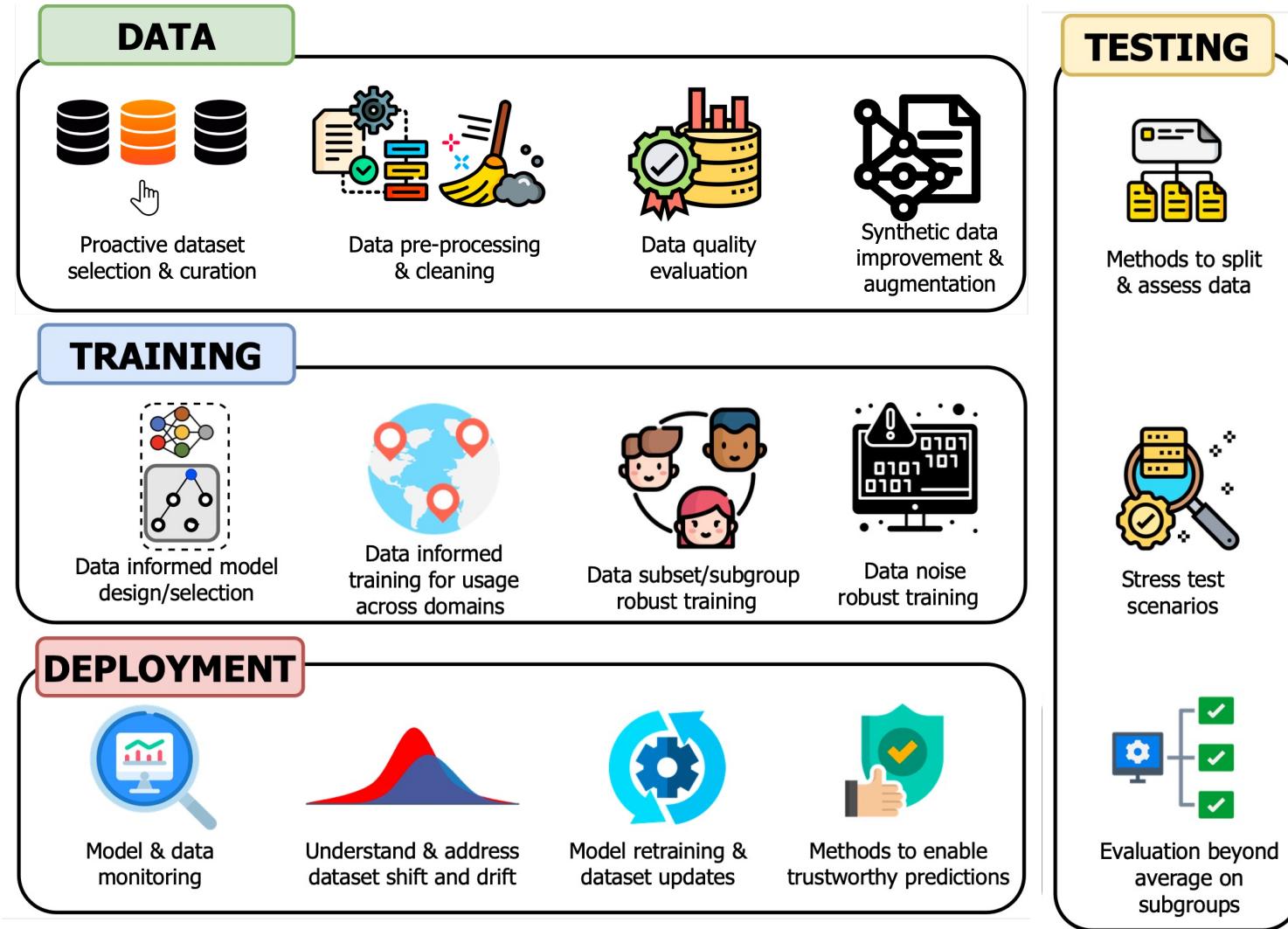
van_der_Schaar
\\ LAB

vanderschaar-lab.com



UNIVERSITY OF
CAMBRIDGE

Construct Reality-Centric ML pipelines



DC-Check: “Making ML work” to “Making ML real”



Find out more & engage with DC-Check resources @
<https://www.vanderschaar-lab.com/dc-check>

An actionable & standardized process to design data-centric AI

For BOTH researchers & practitioners

A recommendation tool for enabling transparency & accountability



van_der_Schaar
LAB

UNIVERSITY OF
CAMBRIDGE

Join us in doing Reality-Centric AI!

- Reality-centric AI aims to reorientate AI towards the complexities of the real world



<https://www.vanderschaar-lab.com/the-case-for-reality-centric-ai/>