Statistics-Powered Safe ML

Yaniv Romano

Computer Science Department
Electrical and Computer Engineering Department

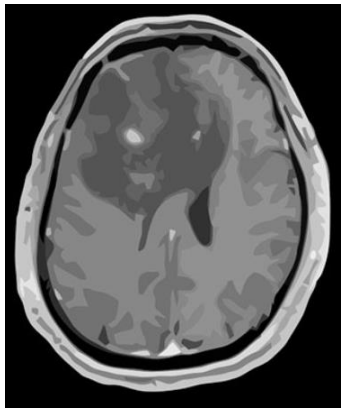
Technion – Israel Institute of Technology

May 2025

SIPL Conference, Technion

The ones who truly make this research matter

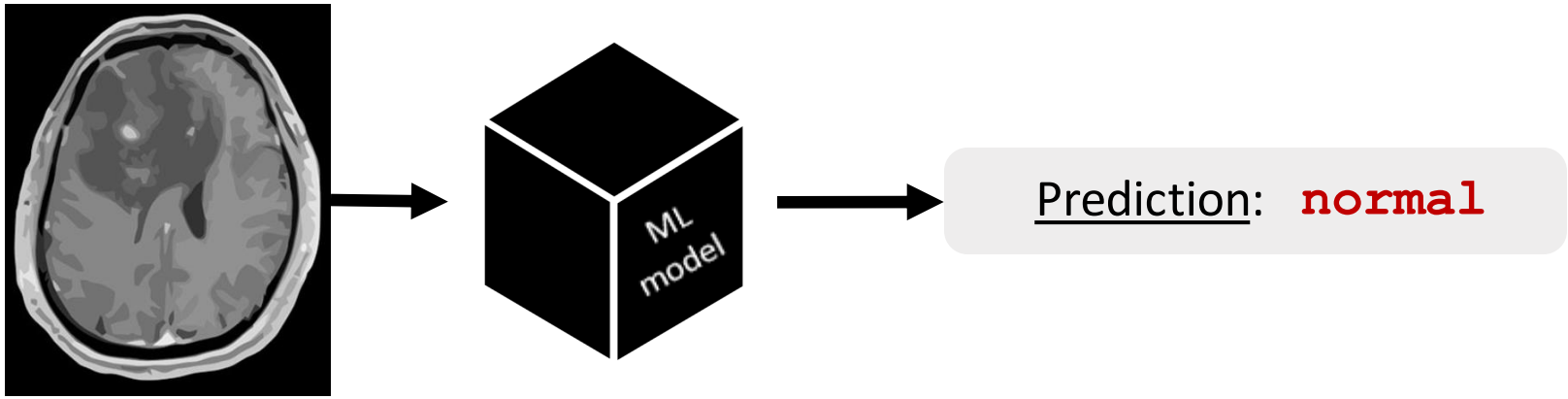
The ML revolution: a tension between potential & risks



True diagnosis = ?

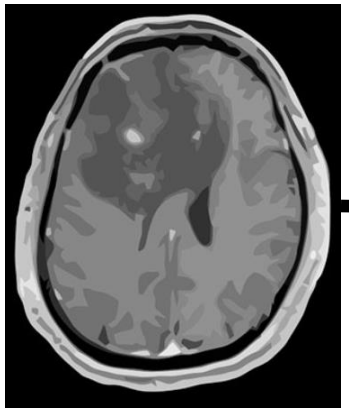
`normal` / `concussion` / `cancer` / …

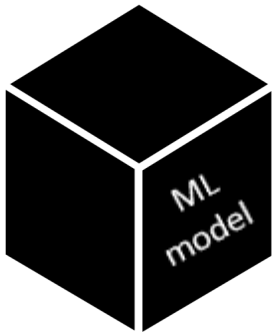
The ML revolution: a tension between potential & risks



Predictive model

The ML revolution: a tension between potential & risks

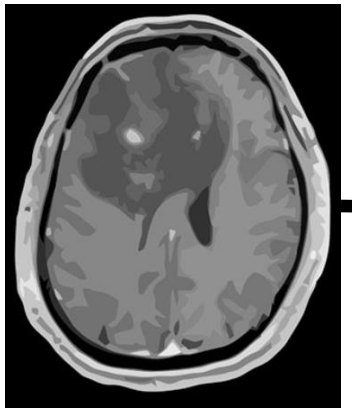


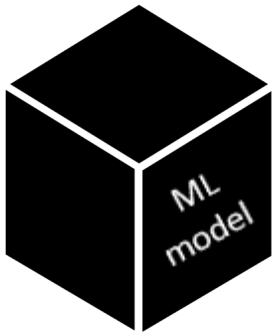


Predictive model

Prediction: **normal**

The ML revolution: a tension between potential & risks





Predictive model

Prediction: **normal**

- **Black-box effect**: hard to interpret
- **Deployment issues**: unpredictable behavior, test-time failures
- **Low-quality training data**: inaccurate labels, missing labels, synthetic, …

Unprecedented need to build confidence in ML systems

Overarching goal

put precise error bounds on ML predictions,
honestly reporting what can be inferred from data

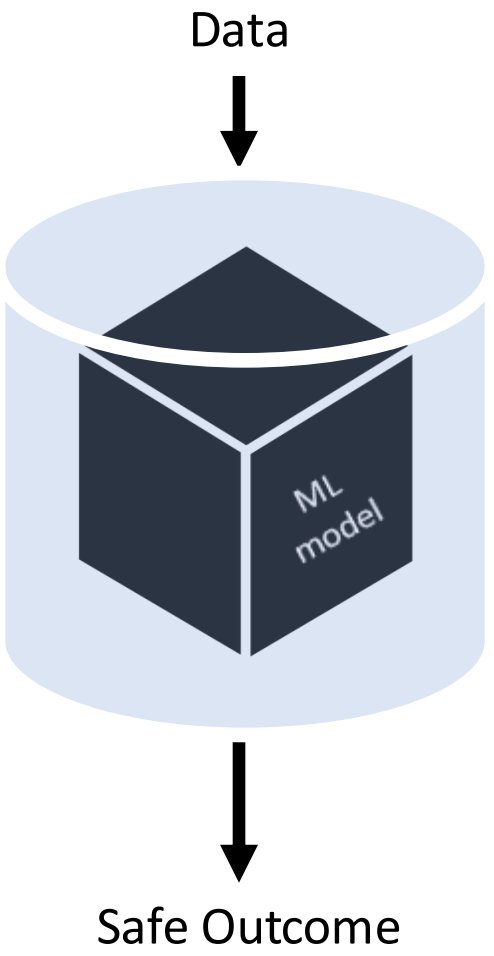
Unprecedented need to build confidence in ML systems

Overarching goal

put precise error bounds on ML predictions,
honestly reporting what can be inferred from data

How?

Novel protection tools that leverage black-box algorithms
and guarantee their reliability



Unprecedented need to build confidence in ML systems

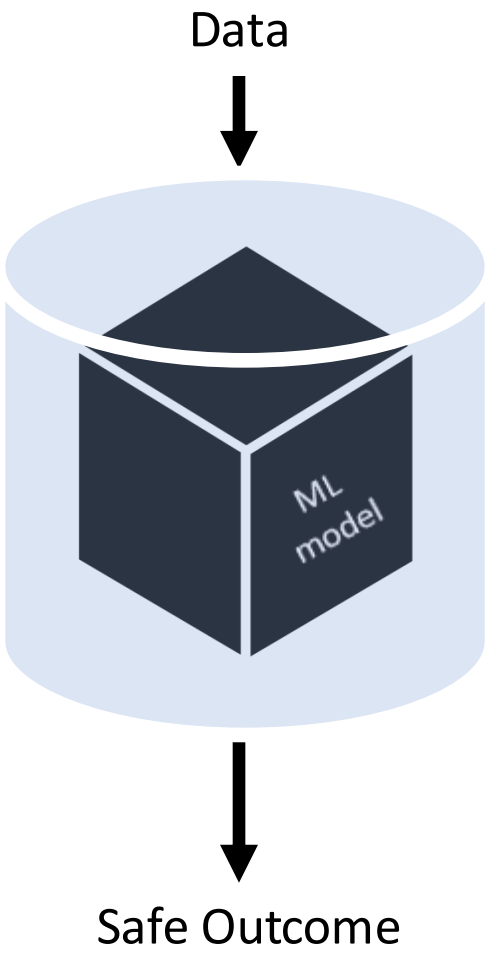
Overarching goal

put precise error bounds on ML predictions,
honestly reporting what can be inferred from data

How?

Novel protection tools that leverage black-box algorithms
and guarantee their reliability

- ✓ Under finite samples
- ✓ Any data: distribution-free
- ✓ Any black-box



Unprecedented need to build confidence in ML predictions

Overarching goal

put precise error bounds on ML predictions,
honestly reporting what can be inferred from data

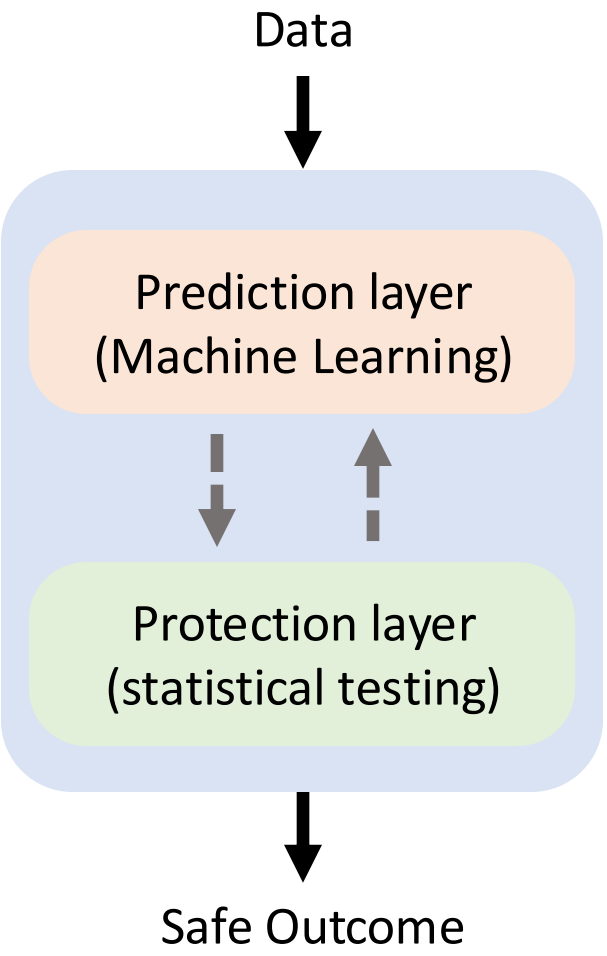
How?

Novel protection tools that leverage black-box algorithms
and guarantee their reliability

- ✓ Under finite samples
- ✓ Any data: distribution-free
- ✓ Any black-box

Novel approach

Revealing a unique interplay between statistics and ML



Real-world application of our statistical wrapper [CQR, Romano et al. ('19)]

- The *Washington Post* used **our method** to reliably project the 2020 US election results

Pennsylvania

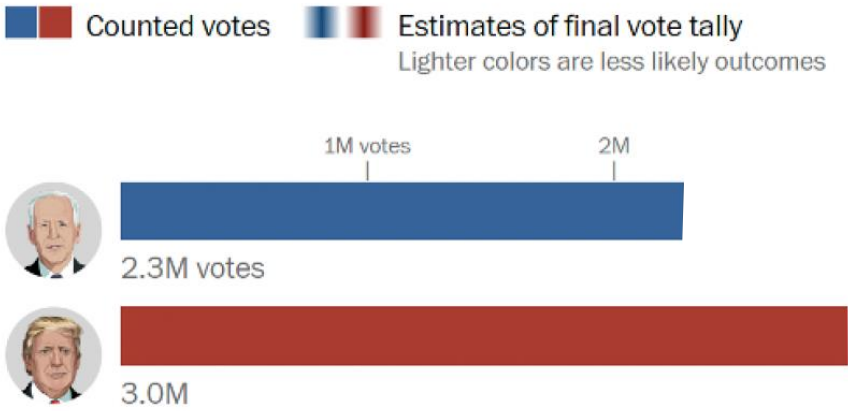
20 ELECTORAL VOTES

LIVE: Donald Trump (R) is leading. An estimated 78 percent of votes have been counted.

Biden
43.0%
2,283,656

Trump
55.7%
2,956,791

Where the votes could end up



The Washington Post
Democracy Dies in Darkness

Election night model results
(4 November 2020, 11:50 PM CA Time)

Real-world application of our statistical wrapper [CQR, Romano et al. ('19)]

- The *Washington Post* used **our method** to reliably project the 2020 US election results

Pennsylvania

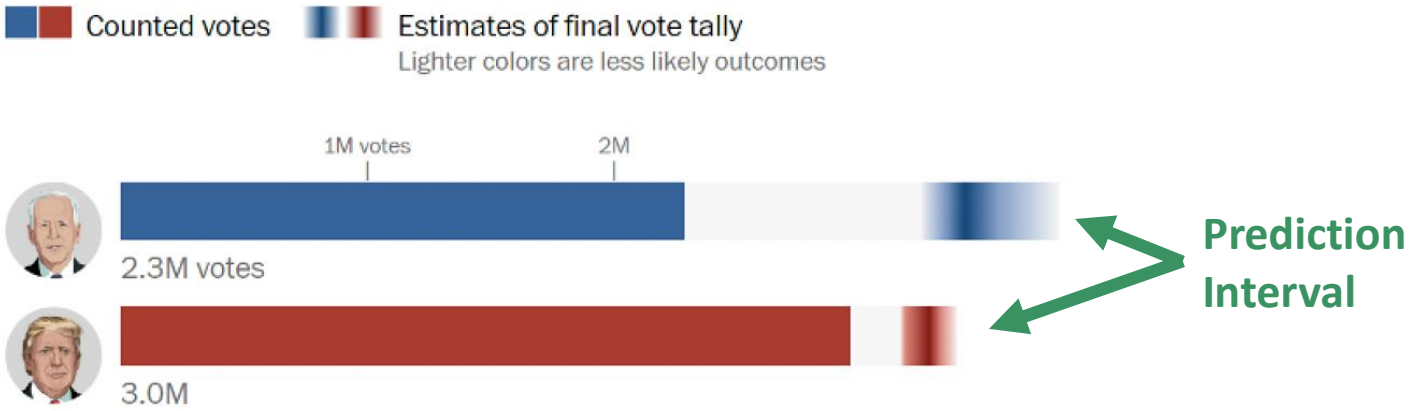
20 ELECTORAL VOTES

LIVE: Donald Trump (R) is leading. An estimated 78 percent of votes have been counted.

Biden
43.0%
2,283,656

Trump
55.7%
2,956,791

Where the votes could end up



The Washington Post
Democracy Dies in Darkness

Election night model results
(4 November 2020, 11:50 PM CA Time)

Real-world application of our statistical wrapper [CQR, Romano et al. ('19)]

- The *Washington Post* used **our method** to reliably project the 2020 US election results
- Same for the **2024 US election night**, with enhanced technology [Cherian, Bronner, and Candes ('24)]

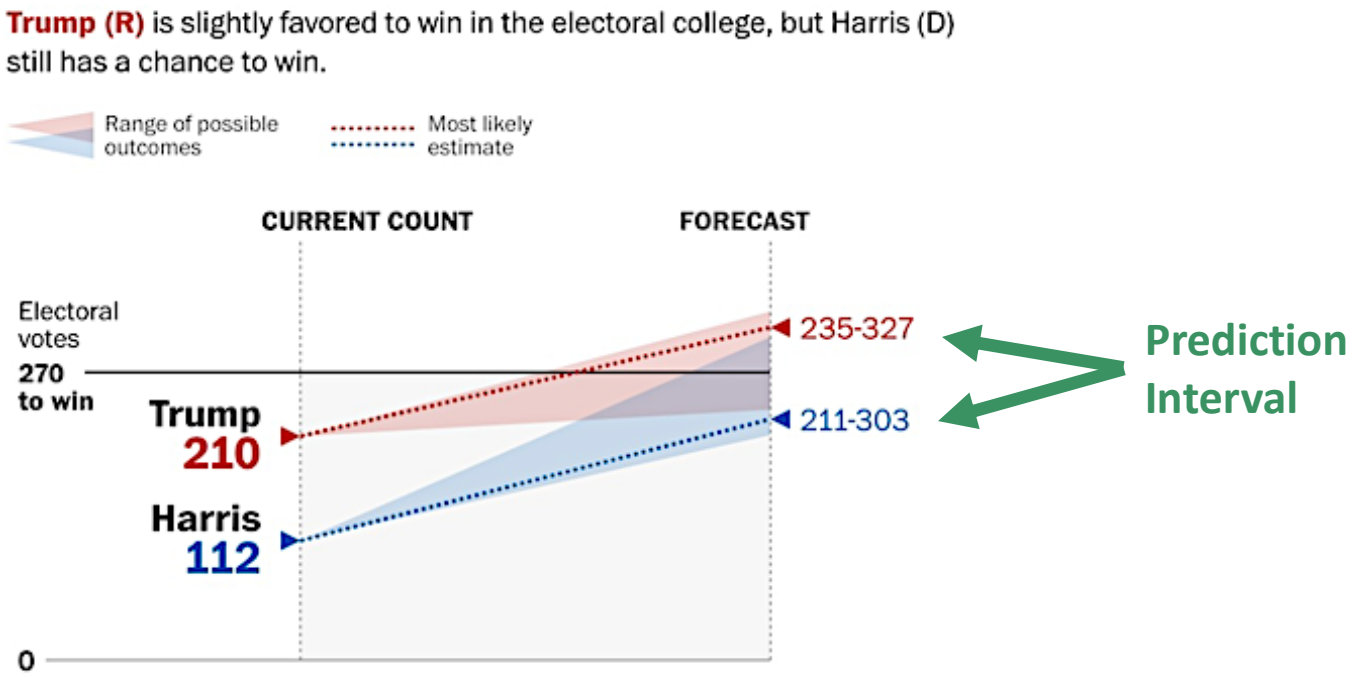
Post Pulse

Our forecast

Likely outcomes based on counted votes

Our forecast analyzes votes counted so far, along with historical results and demographic data, to estimate how many votes are outstanding and which candidate or party those votes will ultimately benefit. Read more.

Election night model results
(6 November 2024, 6:00 AM ISR Time)



Standard conformal prediction: key idea

Standard conformal prediction: key idea

Holdout **calibration data**

i.i.d. labeled samples

$$\{(X_i, Y_i)\}_{i=1}^{n}$$

features label

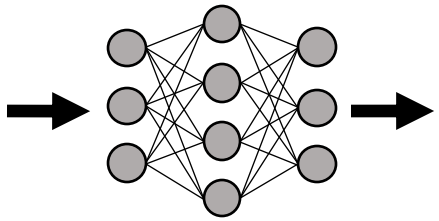
Standard conformal prediction: key idea

Holdout **calibration data**

i.i.d. labeled samples

$\{(X_i, Y_i)\}_{i=1}^{n}$

features label



Predictive model (black-box)

Calibrate a heuristic notion of prediction error

Standard conformal prediction: key idea

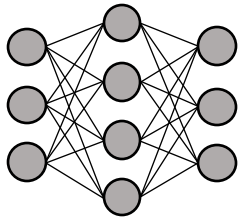
Test time inference

Holdout **calibration data**

i.i.d. labeled samples

$\{(X_i, Y_i)\}_{i=1}^{n}$

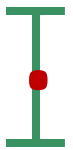
features label



Predictive model (black-box)

Calibrate a heuristic notion of prediction error

$C(X_{\text{test}})$



Prediction interval for a test point

$(X_{\text{test}}, Y_{\text{test}} = ?)$

Standard conformal prediction: key idea

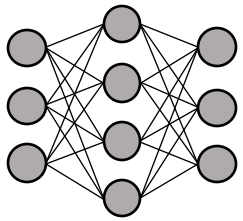
Test time inference

Holdout **calibration data**

i.i.d. labeled samples

$\{(X_i, Y_i)\}_{i=1}^{n}$

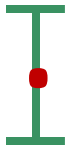
features label



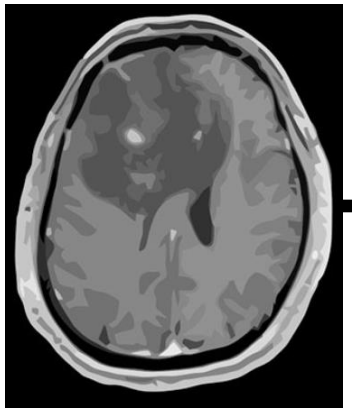
Predictive model (black-box)

Calibrate a heuristic notion of prediction error

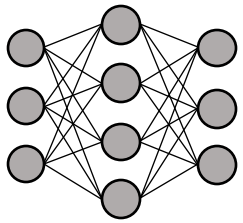
$C(X_{\text{test}})$



Prediction interval for a test point $(X_{\text{test}}, Y_{\text{test}} = ?)$



X_{test}



Naïve prediction

$\hat{Y}_{\text{test}} = \texttt{normal}$

Options: `normal` / `cancer` / `concussion` / …

Standard conformal prediction: key idea

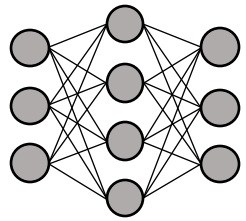
Test time inference

Holdout **calibration data**

i.i.d. labeled samples

$\{(X_i, Y_i)\}_{i=1}^{n}$

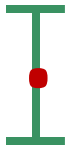
features label



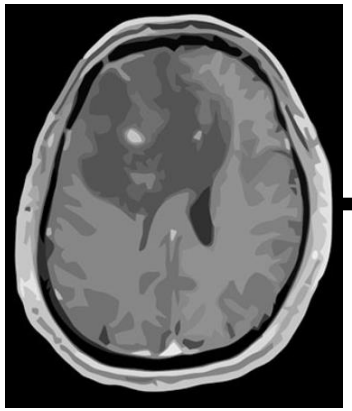
Predictive model (black-box)

Calibrate a heuristic notion of prediction error

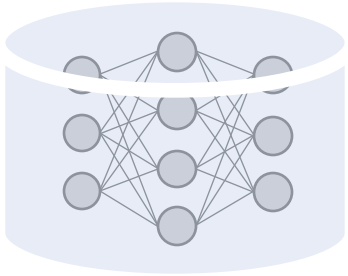
$C(X_{\text{test}})$



Prediction interval for a test point $(X_{\text{test}}, Y_{\text{test}} = ?)$



X_{test}



Prediction set

$C(X_{\text{test}}) = \{\texttt{normal}, \texttt{concussion}\}$

Options: `normal` / `cancer` / `concussion` / …

Standard conformal prediction: key idea

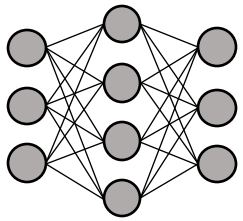
Test time inference

Holdout **calibration data**

i.i.d. labeled samples

$\{(X_i, Y_i)\}_{i=1}^{n}$

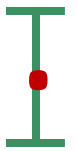
features label



Predictive model (black-box)

Calibrate a heuristic notion of prediction error

$C(X_{\text{test}})$



Prediction interval for a test point $(X_{\text{test}}, Y_{\text{test}} = ?)$

✓ The **prediction set** is guaranteed to cover the unseen test label w.h.p:

$$\mathbb{P}[Y_{\text{test}} \in C(X_{\text{test}})] \geq 1 - \alpha \quad (\text{e.g. } 95\%)$$

Limitations

✗ Holds under the **i.i.d. assumption** (holdout/test)

✗ Holds **marginally** over the population represented by the holdout data

Research landscape: reliable predictive inference

Q1: how to achieve more personalized safety guarantees?

→ **Challenge 1:** limited availability of labeled data

Research landscape: reliable predictive inference

Q1: how to achieve more personalized safety guarantees? → **Challenge 1:** limited availability of labeled data

Q2: how to ensure reliability when handed low-quality holdout data? → **Challenge 2:** violation of the i.i.d. assumption

Research landscape: reliable predictive inference

Q1: how to achieve more personalized safety guarantees? → **Challenge 1:** limited availability of labeled data

Q2: how to ensure reliability when handed low-quality holdout data? → **Challenge 2:** violation of the i.i.d. assumption

Q3: how to enhance robustness to test-time drifting data in an online manner? → **Challenge 3:** lack of up-to-date labels

This (overview) talk

Q1: how to achieve more personalized safety guarantees? → **Challenge 1:** limited availability of labeled data

Q2: how to ensure reliability when handed low-quality holdout data? → **Challenge 2:** violation of the i.i.d. assumption

Q3: how to enhance robustness to test-time drifting data in an online manner? → **Challenge 3:** lack of up-to-date labels

More personalized safety guarantees

Want valid UQ regardless of age, race, ethnicity, …

nature medicine

Explore content | About the journal | Publish with us | Subscribe

nature > nature medicine > comment > article

Comment | Published: 11 October 2023

Clinical AI tools must convey predictive uncertainty for each individual patient

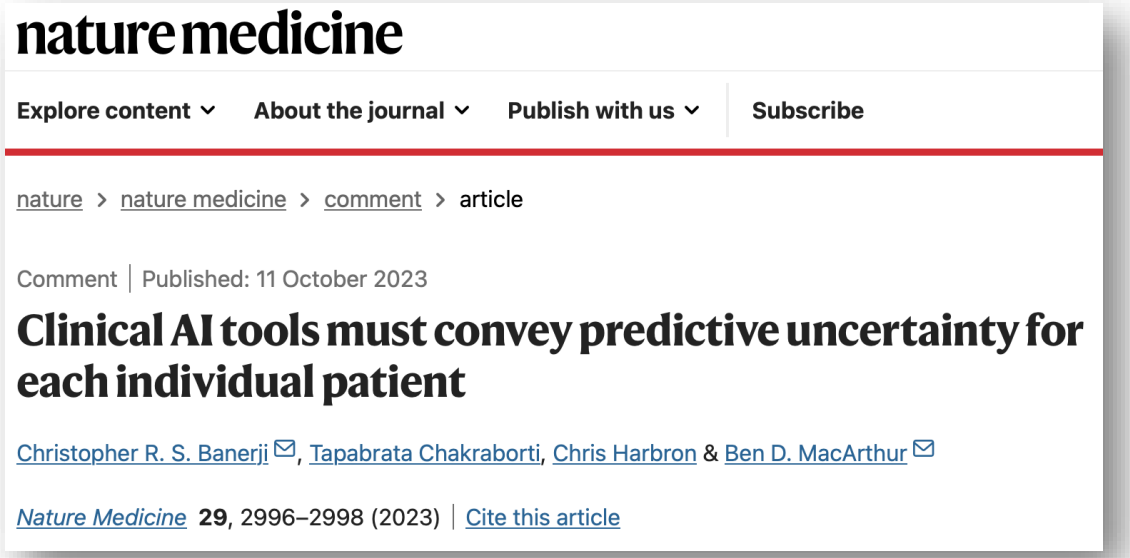
Christopher R. S. Banerji, Tapabrata Chakraborti, Chris Harbron & Ben D. MacArthur

Nature Medicine **29**, 2996–2998 (2023) | Cite this article

More personalized safety guarantees

Want valid UQ regardless of age, race, ethnicity, …

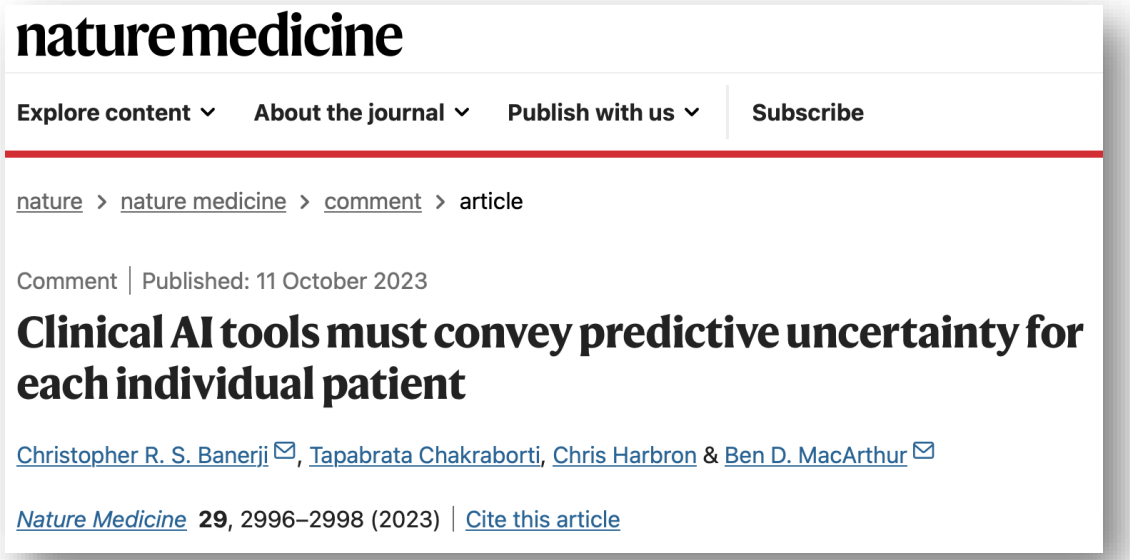
Challenge: localized ML calibration → sample size barriers → uninformative UQ



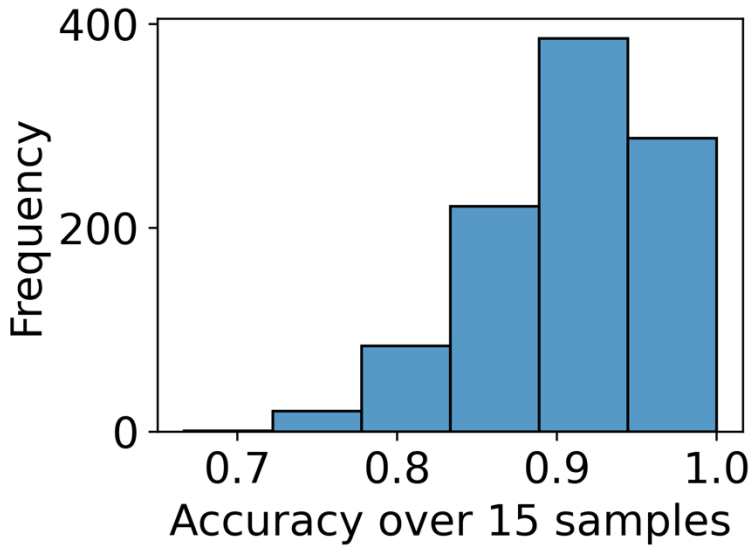
More personalized safety guarantees

Want valid UQ regardless of age, race, ethnicity, …

Challenge: localized ML calibration → sample size barriers → uninformative UQ



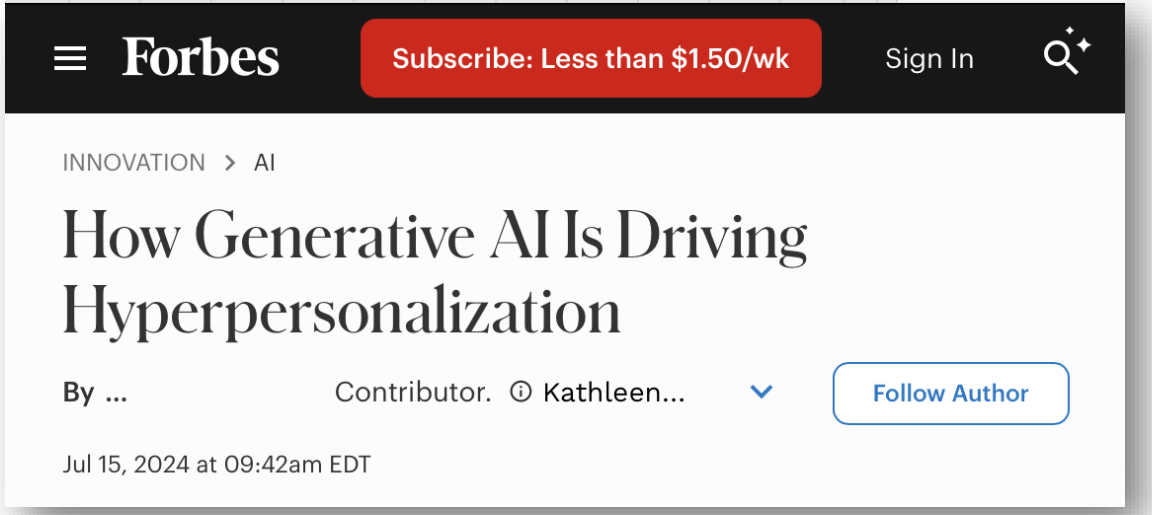
Fundamental sample size limitation



(**Data**: ImageNet; **Model**: VLM)

Recent breakthroughs in generative AI

- GenAI unlocks the ability to generate realistic images, text, …
- Unlocks the ability to fit more accurate, personalized models



GenAI is allowing for hyperpersonalization, one of the seven patterns of AI.
GETTY

Recent breakthroughs in generative AI

- GenAI unlocks the ability to generate realistic images, text, …
- Unlocks the ability to fit more accurate, personalized models
- **Problem:** we can't blindly trust synthetic data: biased, introducing unknown & undesired dist. shifts

AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Nature **631**, 755–759 (2024) | Cite this article

433k Accesses | **3161** Altmetric | Metrics

NEWS AND VIEWS | 24 July 2024

AI produces gibberish when trained on too much AI-generated data

Generative AI models are now widely accessible, enabling everyone to create their own machine-made something. But these models can collapse if their training data sets contain too much AI-generated content.

Recent breakthroughs in generative AI

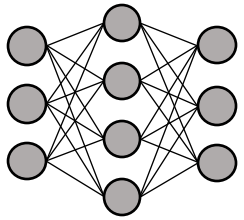
- GenAI unlocks the ability to generate realistic images, text, …
- Unlocks the ability to fit more accurate, personalized models
- **Problem:** we can't blindly trust synthetic data: biased, introducing unknown & undesired dist. shifts

How can we **safely** use synthetic data while achieving personalized reliability guarantees?

Our method: synthetic-powered predictive inference [BLLDR ('25)]

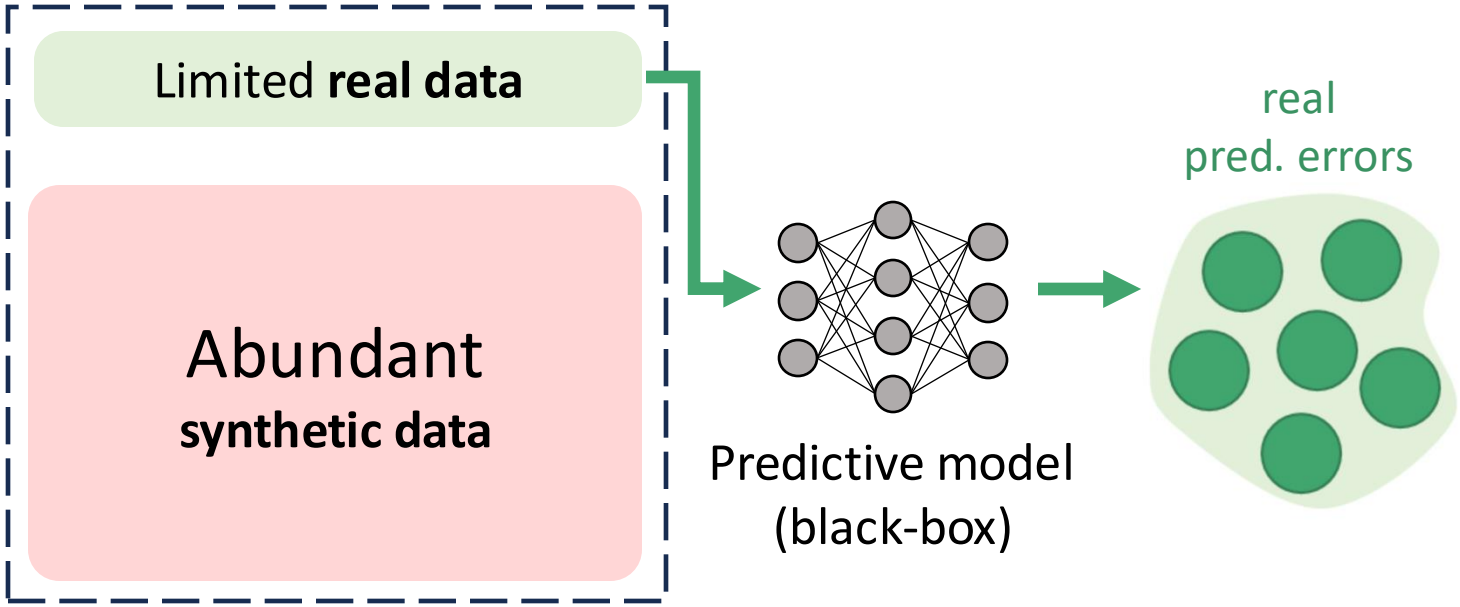
Limited **real data**

Abundant
synthetic data

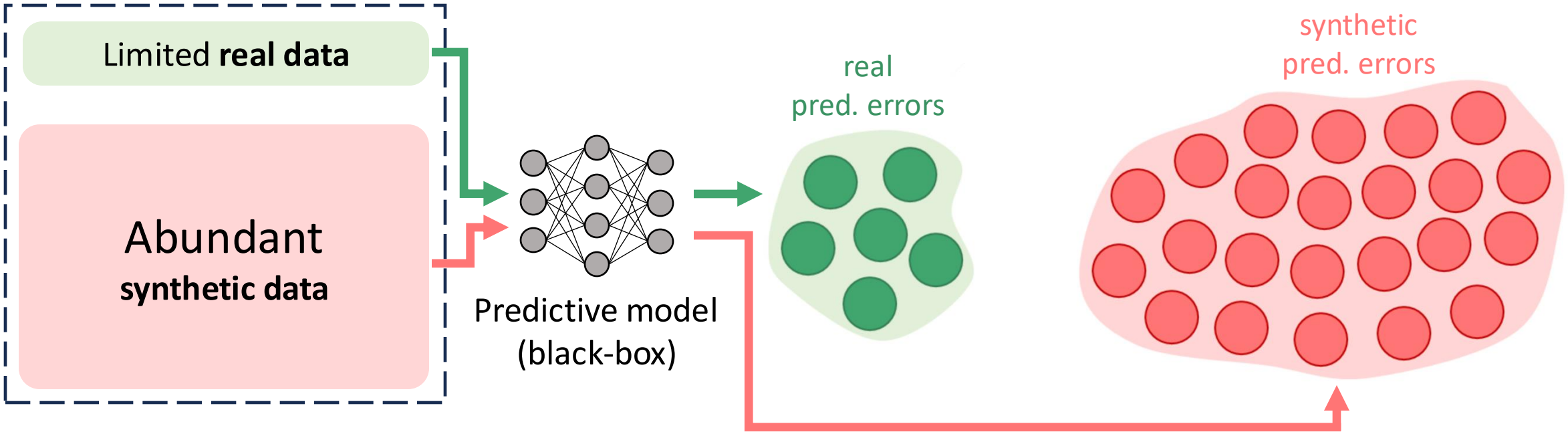


Predictive model
(black-box)

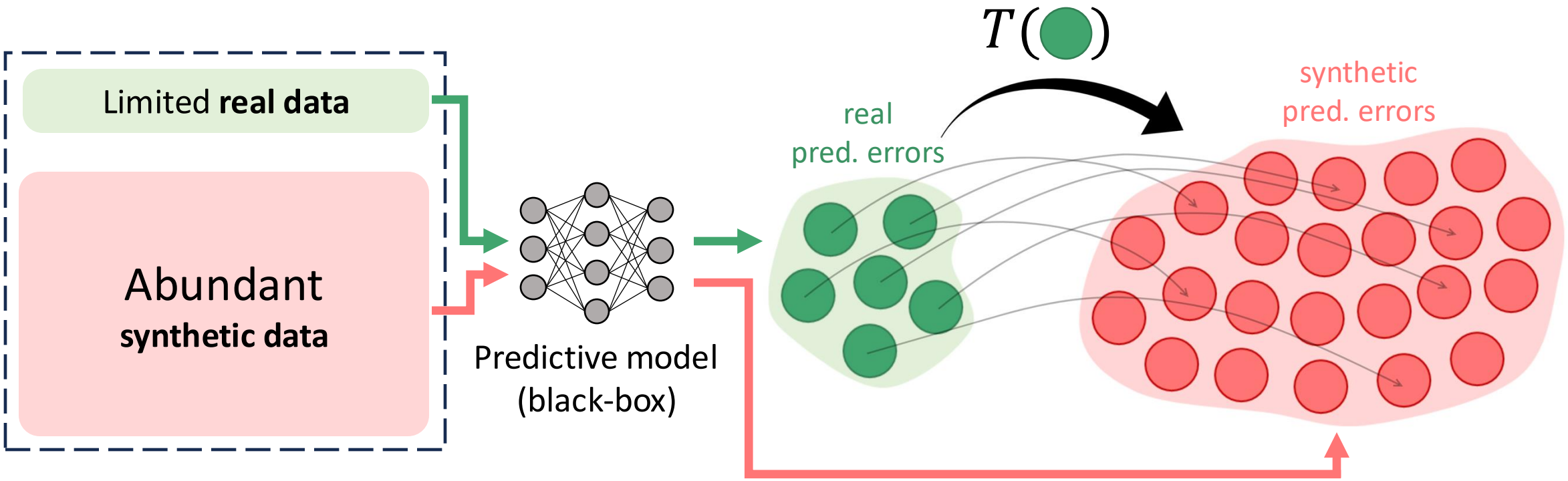
Our method: synthetic-powered predictive inference [BLLDR ('25)]



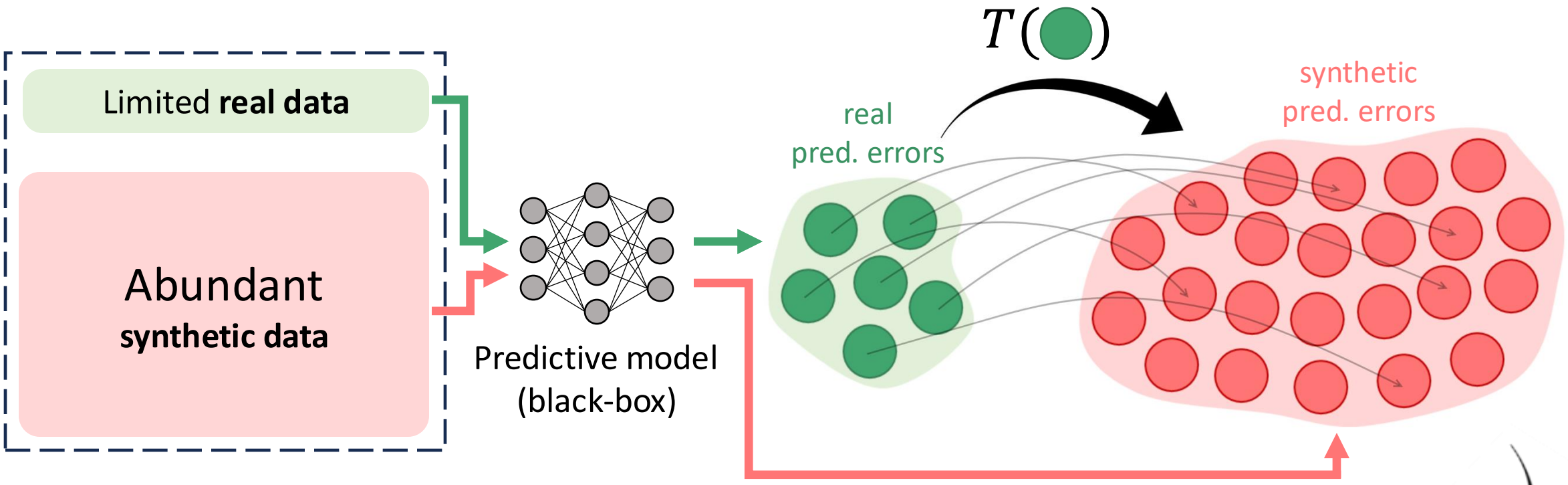
Our method: synthetic-powered predictive inference [BLLDR ('25)]



Our method: synthetic-powered predictive inference [BLLDR ('25)]



Our method: synthetic-powered predictive inference [BLLDR ('25)]

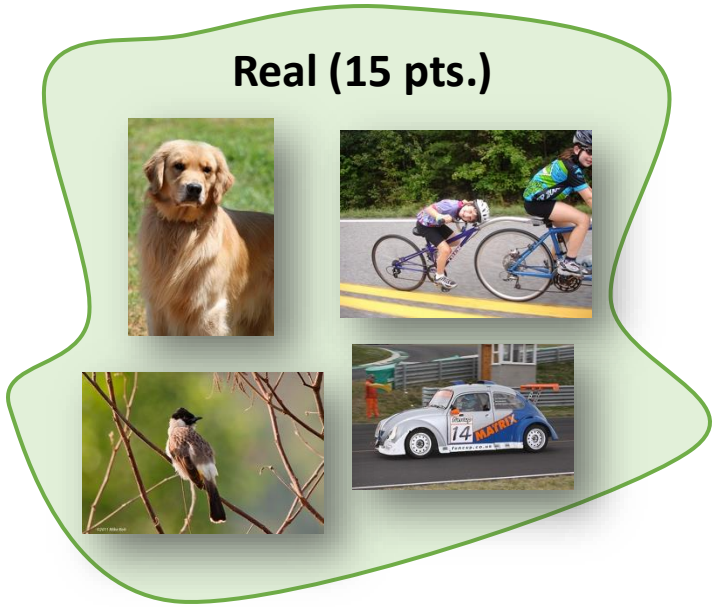


Theorem

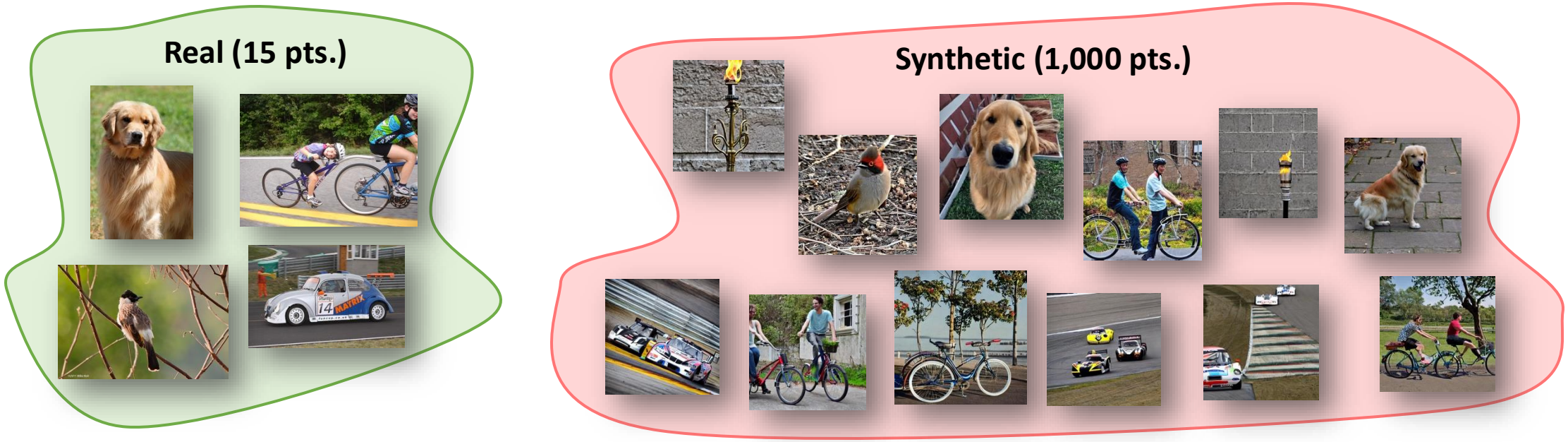
- ✓ Valid coverage (no matter how good GenAI is)
- ✓ High-quality synthetic data? accurate calibration, breaking sample-size barriers

model calibration
on the real-to-synthetic
transported pred. errors

Our method in action: ImageNet (VLM + Stable Diffusion)

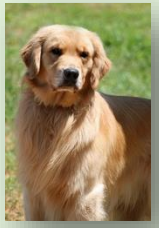


Our method in action: ImageNet (VLM + Stable Diffusion)



Our method in action: ImageNet (VLM + Stable Diffusion)

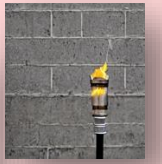
Real (15 pts.)




Synthetic (1,000 pts.)











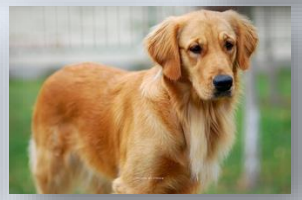

Test set (unlabeled)





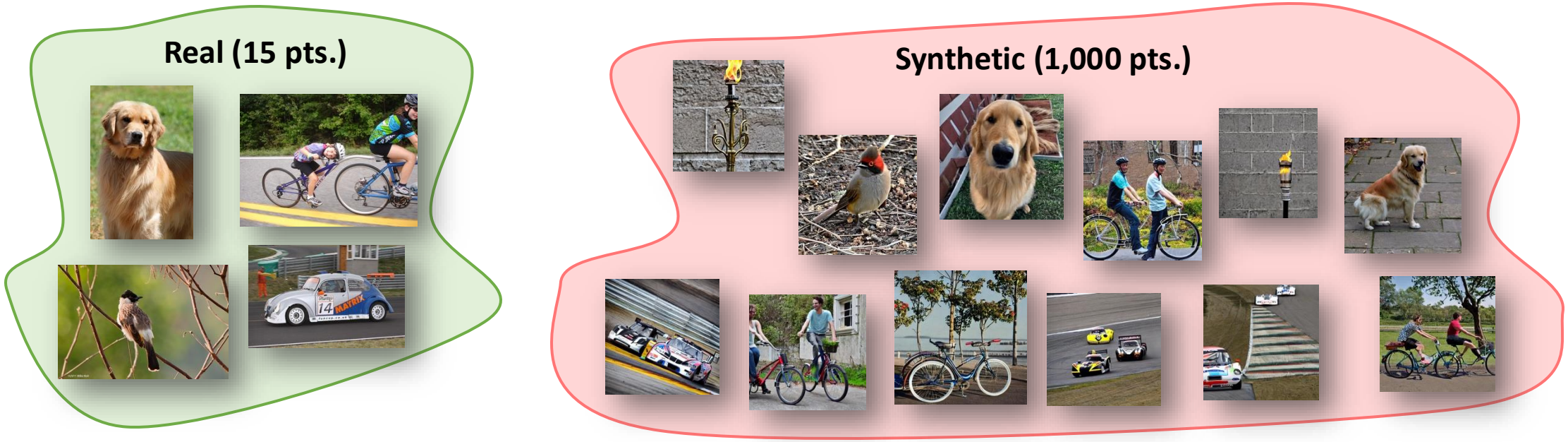


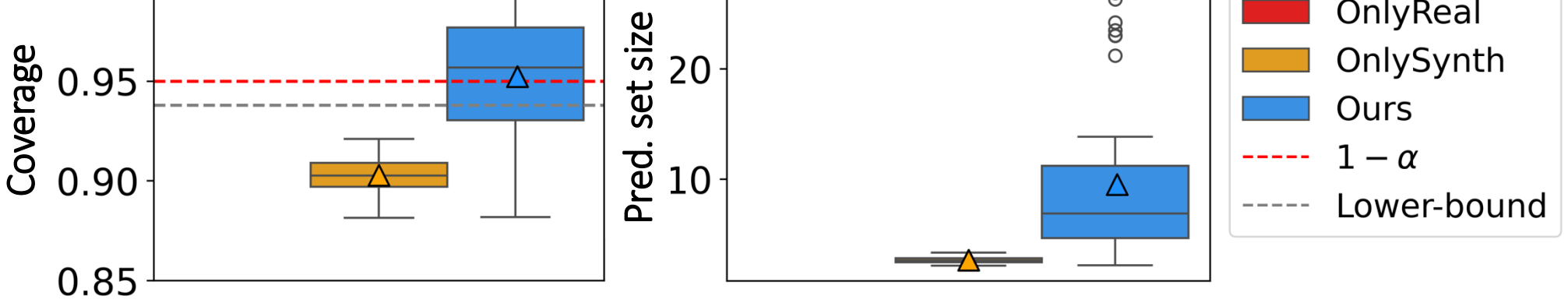


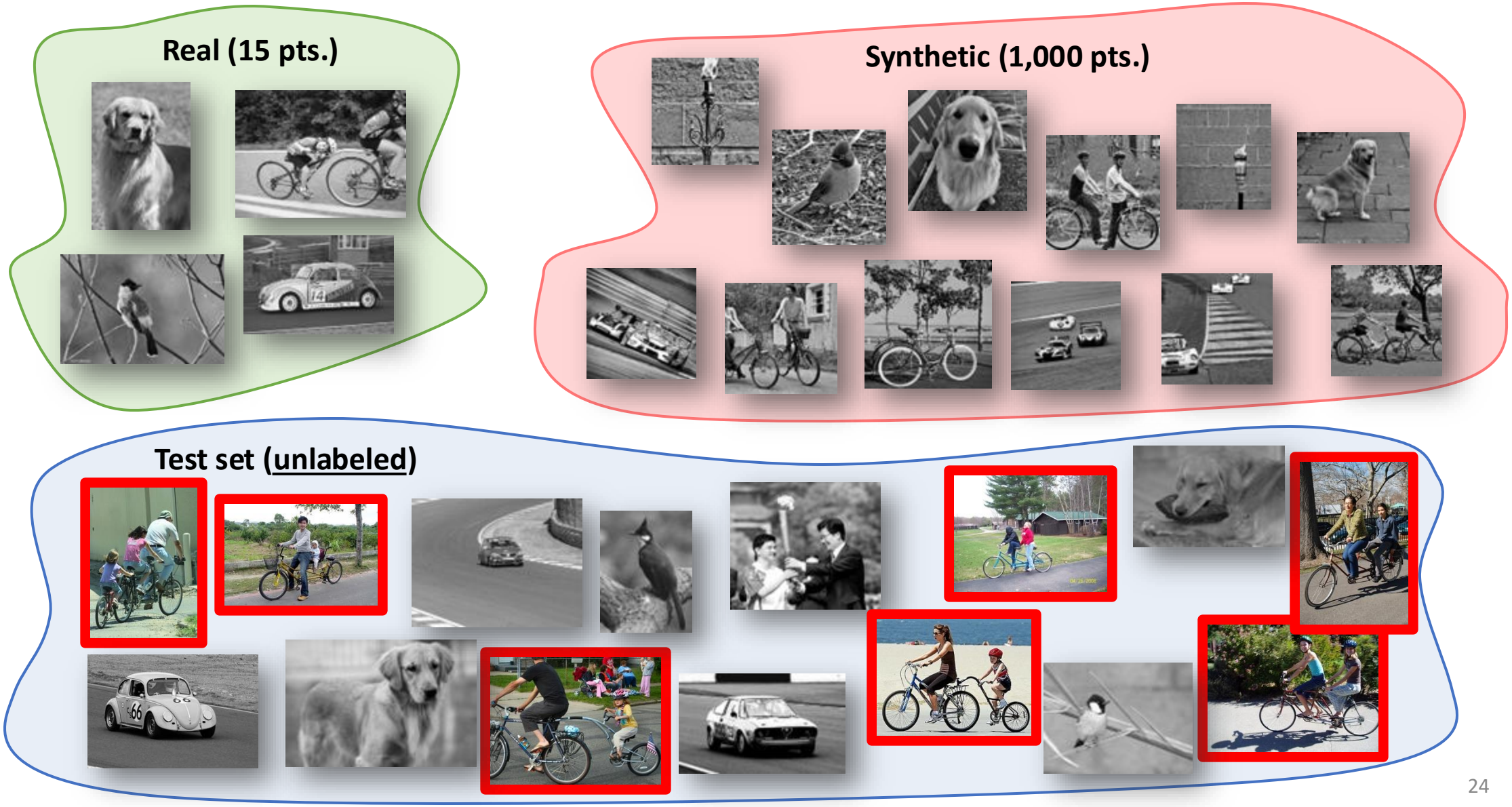



Our method in action: marginal coverage

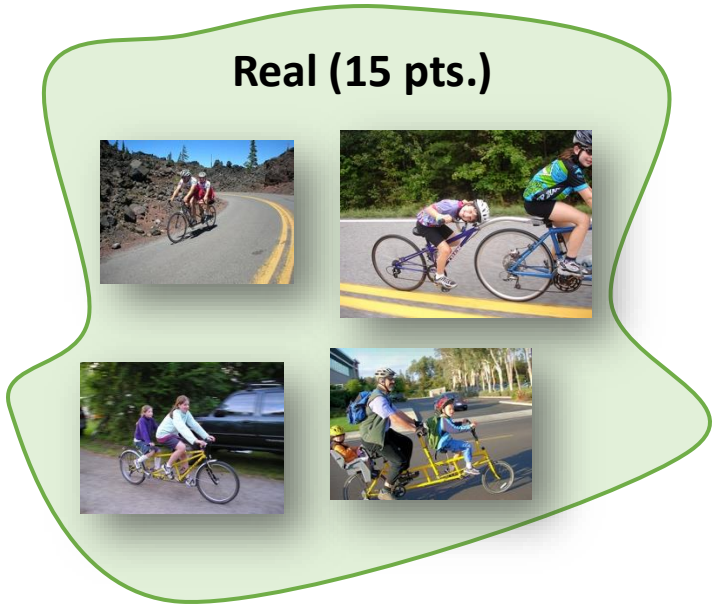


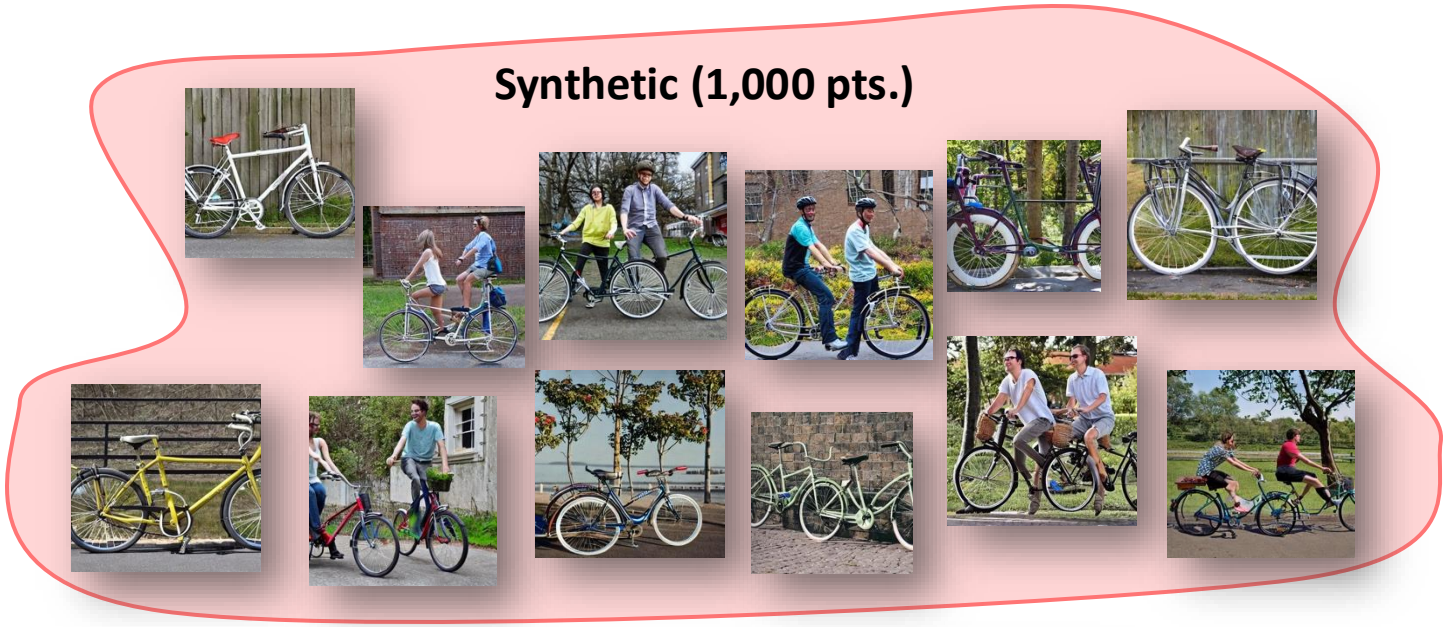


Our method in action: class (=bike) conditional coverage

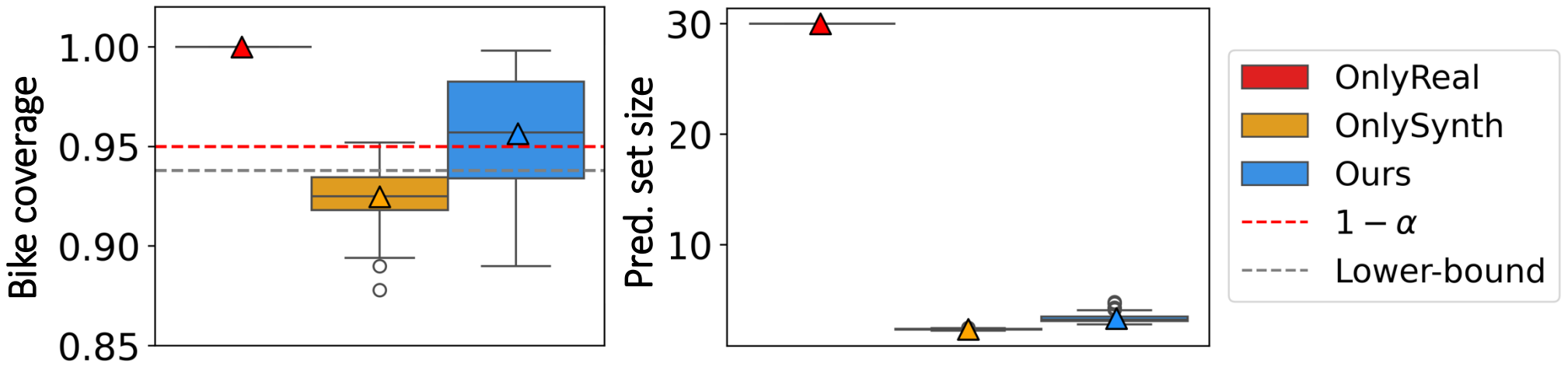


Our method in action: class (=bike) conditional coverage





Our method in action: class (=bike) conditional coverage



A historical perspective

DETERMINATION OF SAMPLE SIZES FOR SETTING TOLERANCE LIMITS

By S. S. Wilks

Princeton University, Princeton, N. J.

The Annals of Mathematical Statistics, **March 1941**

Q. How many samples do we need to obtain a **stable** prediction interval for a quality characteristic of a product?

A. About 1,000 (real) samples

Samuel S. Wilks (1906-1964)

A historical perspective

DETERMINATION OF SAMPLE SIZES FOR SETTING TOLERANCE LIMITS

By S. S. Wilks

Princeton University, Princeton, N. J.

The Annals of Mathematical Statistics, **March 1941**

Q. How many samples do we need to obtain a **stable** prediction interval for a quality characteristic of a product?

A. About 1,000 (real) samples

Samuel S. Wilks (1906-1964)

Fast forward to 2025… we have a new result!

Can break this sample size limit and obtain stable pred. intervals via synthetic data

Time to conclude

Our focus

Supporting black-box ML systems with formal safety guarantees

- Personalization
- Robustness
- Online adaptation

Time to conclude

Our focus

Supporting black-box ML systems with formal safety guarantees

- Personalization
- Robustness
- Online adaptation

Research horizons

- Stat empowers ML and ML empowers Stat
- Integrate protection layers within ML training

Time to conclude

Our focus

Supporting black-box ML systems with formal safety guarantees

- Personalization
- Robustness
- Online adaptation

Research horizons

- Stat empowers ML and ML empowers Stat
- Integrate protection layers within ML training

Social impact

Trustworthy data-driven insights, extracted from the most advanced ML systems

Time to conclude

Our focus

Supporting black-box ML systems with formal safety guarantees

- Personalization
- Robustness
- Online adaptation

Research horizons

- Stat empowers ML and ML empowers Stat
- Integrate protection layers within ML training

Social impact

Trustworthy data-driven insights, extracted from the most advanced ML systems

Thank you!