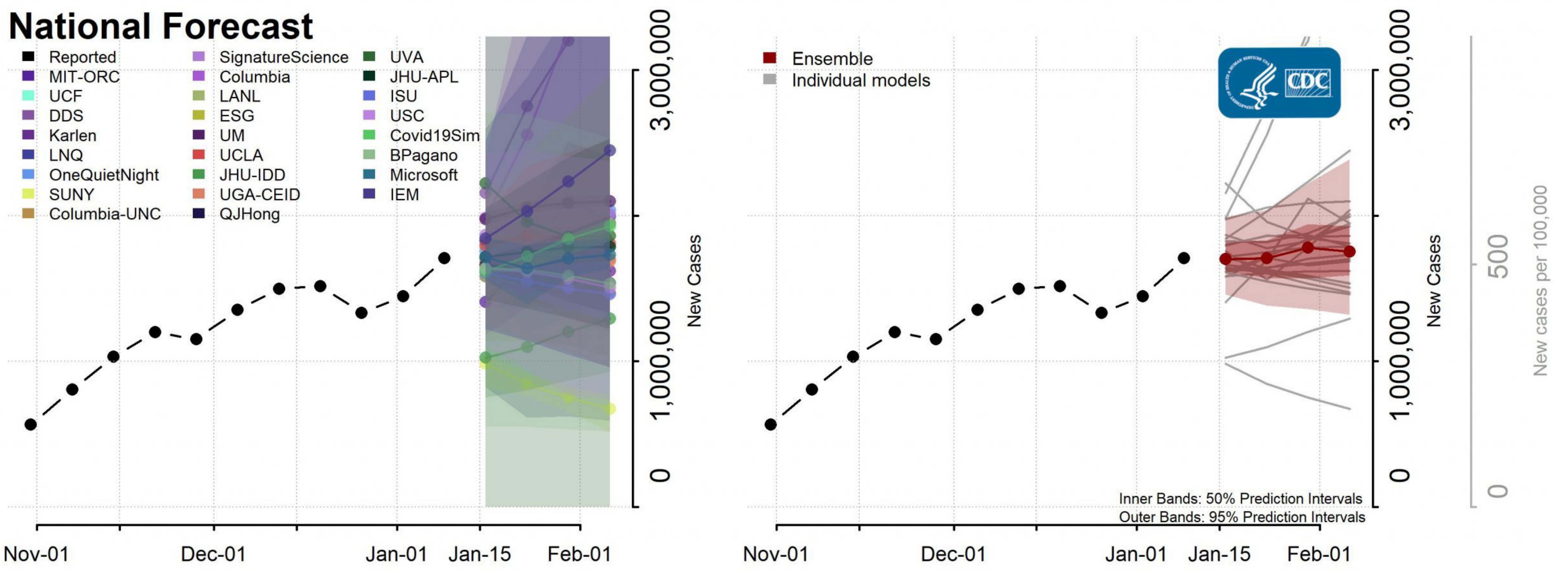
Probabilistic Deep Learning for Uncertainty Quantification and Decision Making

Final Defense for Sophia Sun
Advised by Professor Rose Yu
Nov. 21, 2025

UCSDCSE
Computer Science and Engineering
spatiotemporal Machine Learning

Motivation: ML for Critical Applications



Motivation: ML for Critical Applications

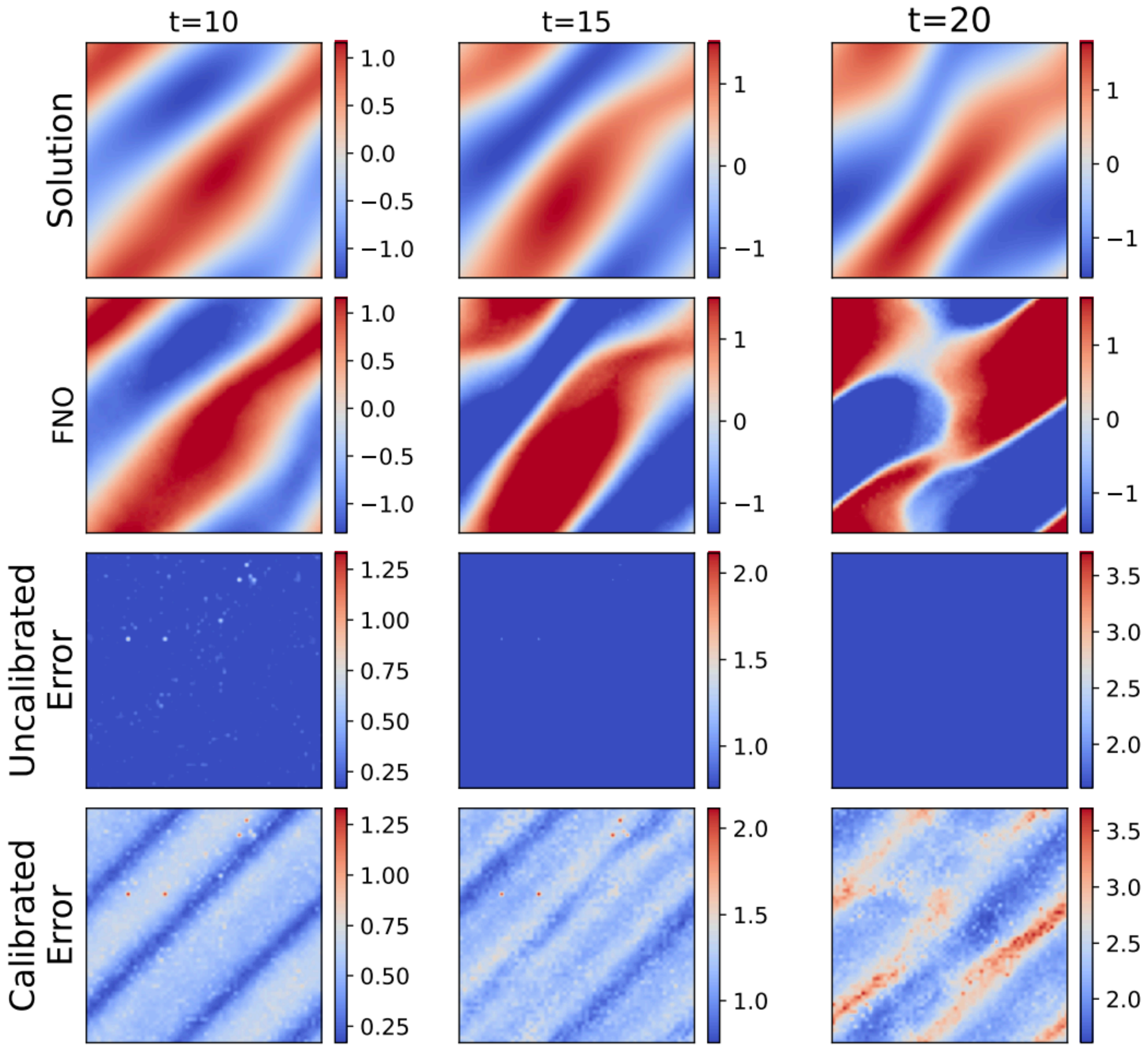


Figure 9: Calibrating the uncertainty captured by using MC dropout (STD) within the FNO in modelling out-of-distribution data for the Navier–Stokes case. The top row shows the ground truth, the second row the output of the FNO, the third row the error (taken as the standard deviation here) captured by the probabilistic FNO, and the final row shows the calibrated error obtained using the CP framework over the probabilistic outputs showing 67 % coverage.

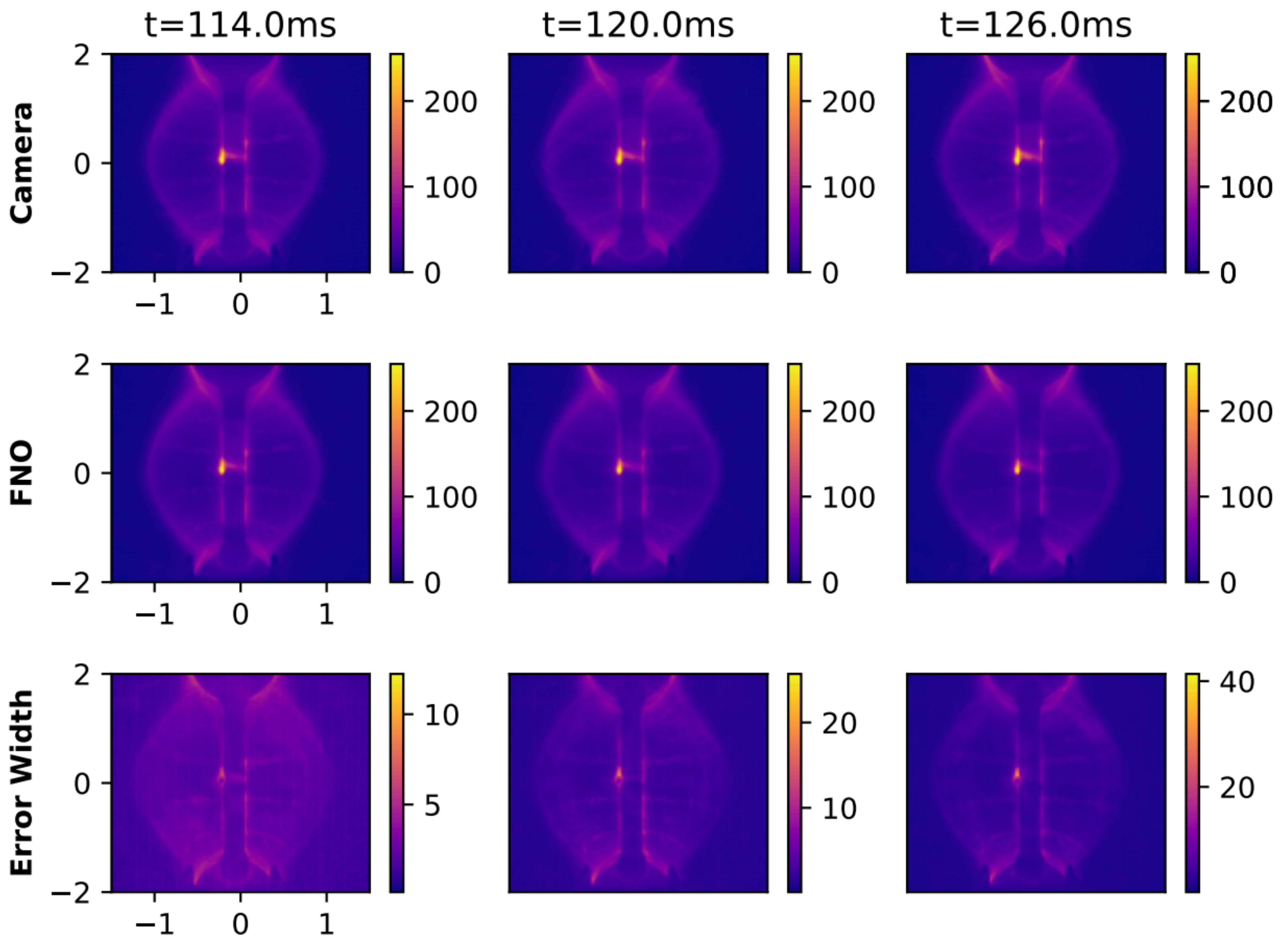
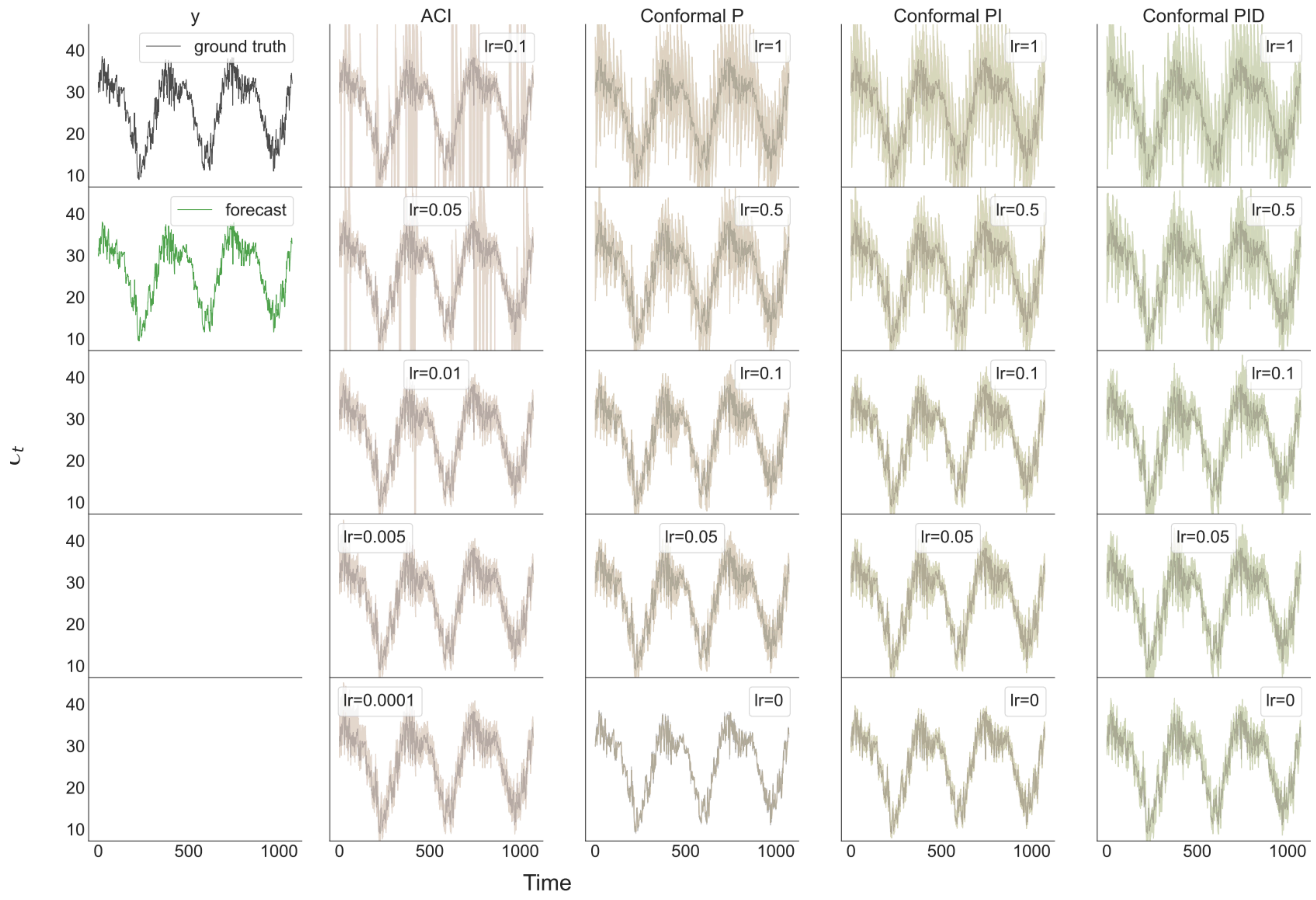


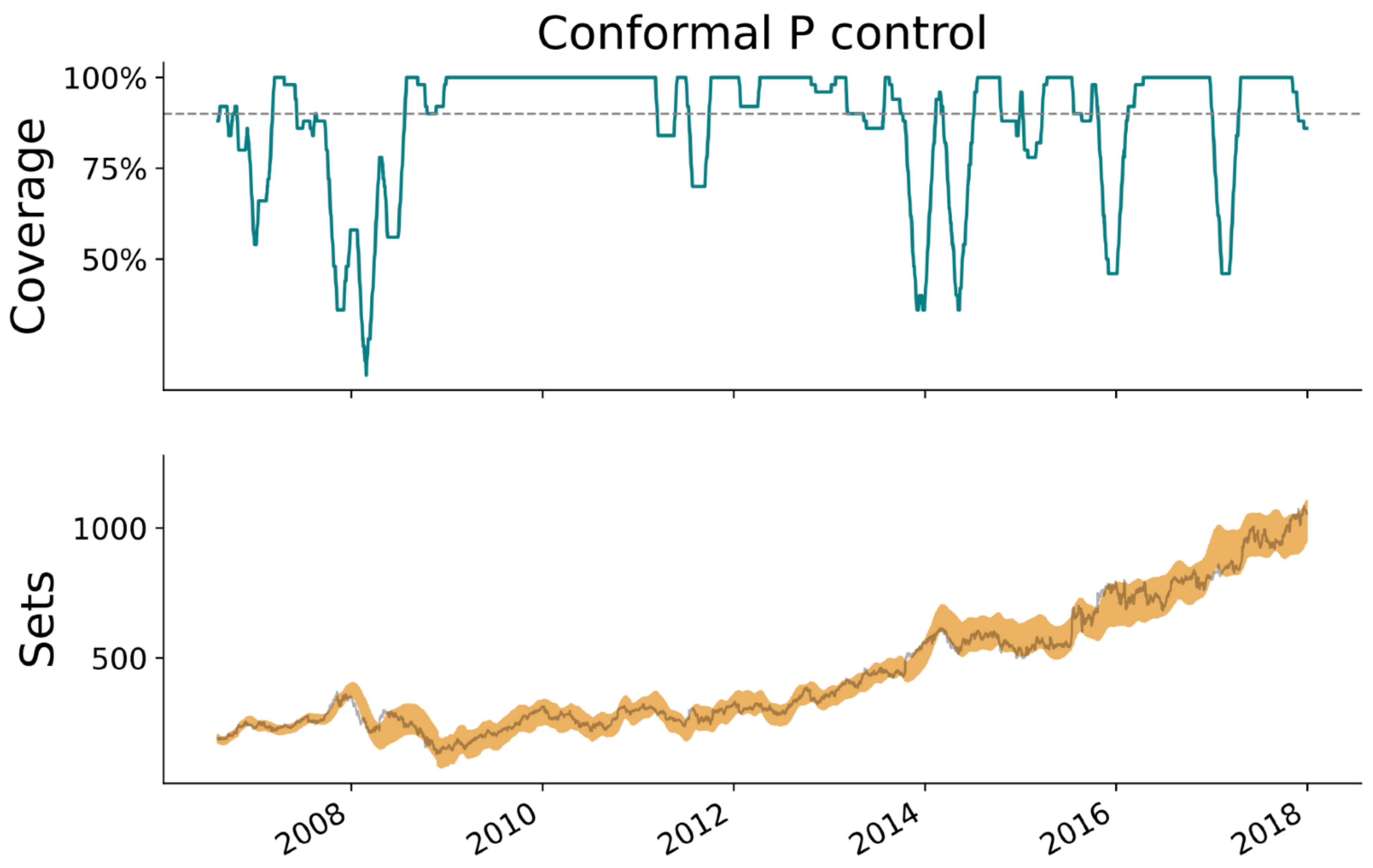
Figure 22: Camera (top), FNO (middle) and the prediction interval width obtained using CP with $\alpha = 0.5$ (bottom).

Uncertainty Quantification of Surrogate Models using Conformal Prediction, 2024

Motivation: ML for Critical Applications



Temperature in Delhi

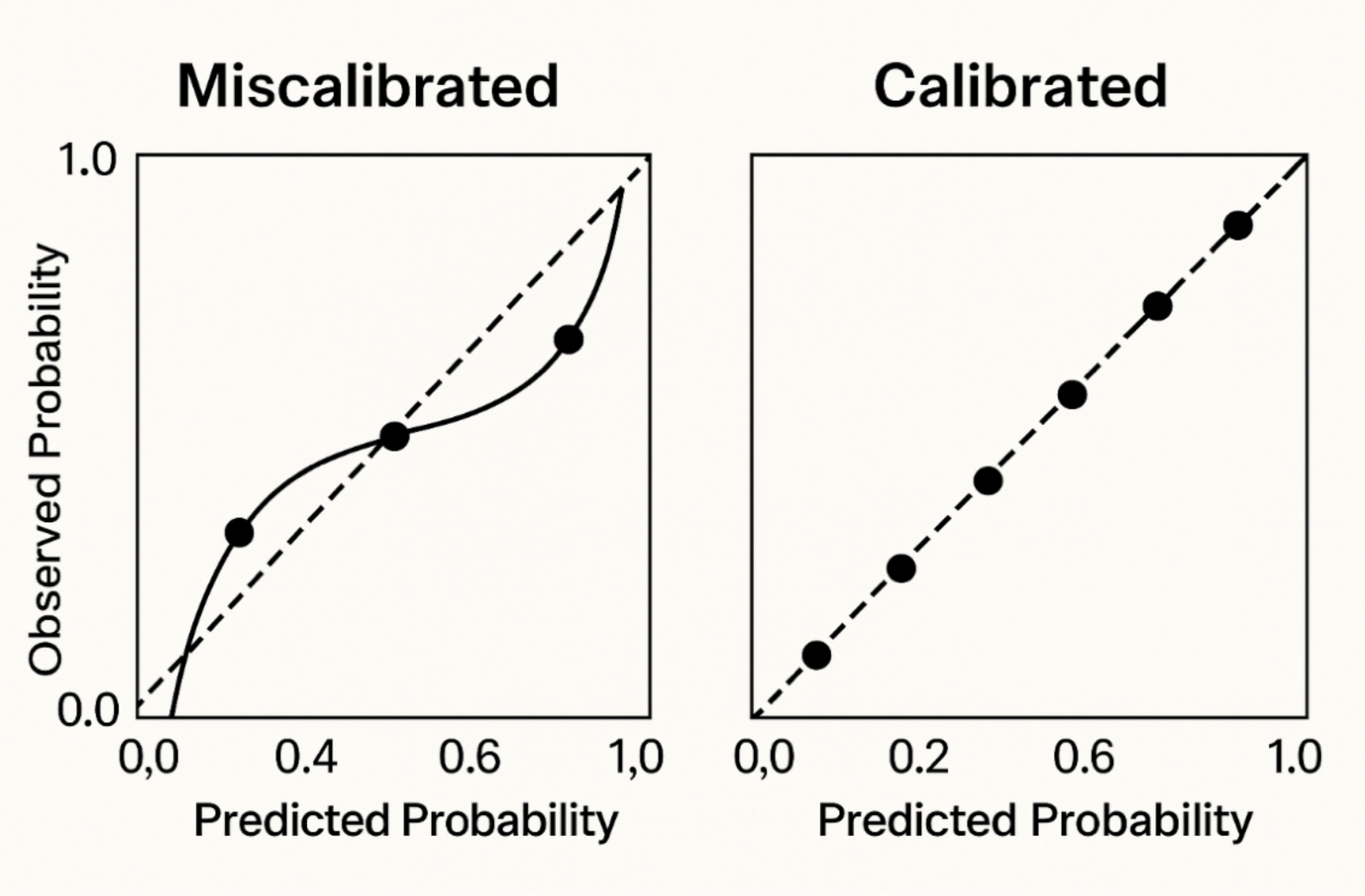


Performance of Google Stocks

Angelopoulos et al. (2023) Conformal PID Control.

Calibration of Probabilistic Forecasts

Classification Case. A predictor $f : \mathscr{X} \to [0,1]$ is **calib**



$P_{X,Y}$ if for all $p \in [0,1]$:

$$\mathbb{P}(Y = 1 \mid f(X) = p) = p$$

1-D Regression Case. A predictor $f : \mathscr{X} \to \mathscr{P}(\mathbb{R})$ where $f(x) = F_x(\,\cdot\,)$ is a CDF is **calibrated** if:

$$\mathbb{P}(Y \leq y \mid F_X(y) = p) = p \quad \forall y \in \mathbb{R}, p \in [0,1]$$

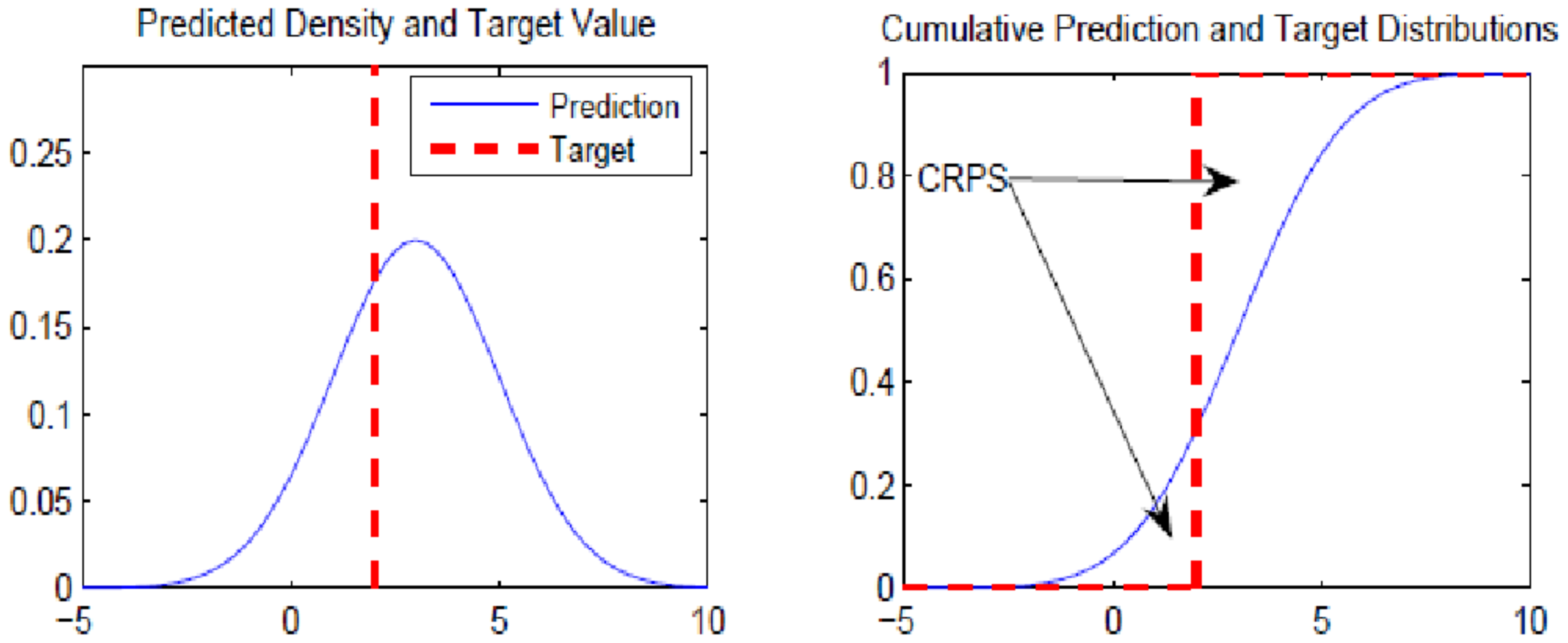
Note: A calibrated model doesn't reflect "accuracy" and can be arbitrarily bad.

Good probabilistic forecasts

"… is maximizes the sharpness of the predictive distributions subject to calibration." - Gneiting

Continuous ranked probability score (CRPS)

$$\mathrm{CRPS}(F_X, y) = \int_{-\infty}^{\infty} (F_X(z) - \mathbb{1}\{y < z\})^2 \, dz$$



Gneiting, Tilmann, and Adrian E. Raftery. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American statistical Association*, 2007

In this dissertation…

- We try to address two challenges:

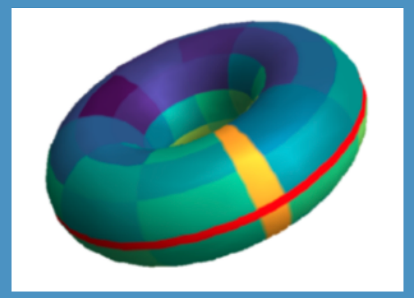
  1. How to obtain calibrated and sharp probabilistic forecasts from deep learning models?
  2. How can we use these uncertainties for better decision making?

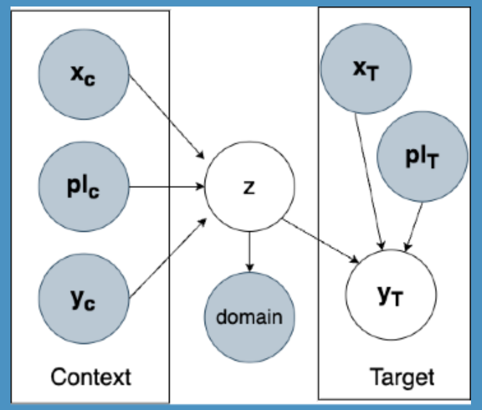
In this dissertation…

Probabilistic Modeling and Uncertainty Quantification

Leveraging structure in model design

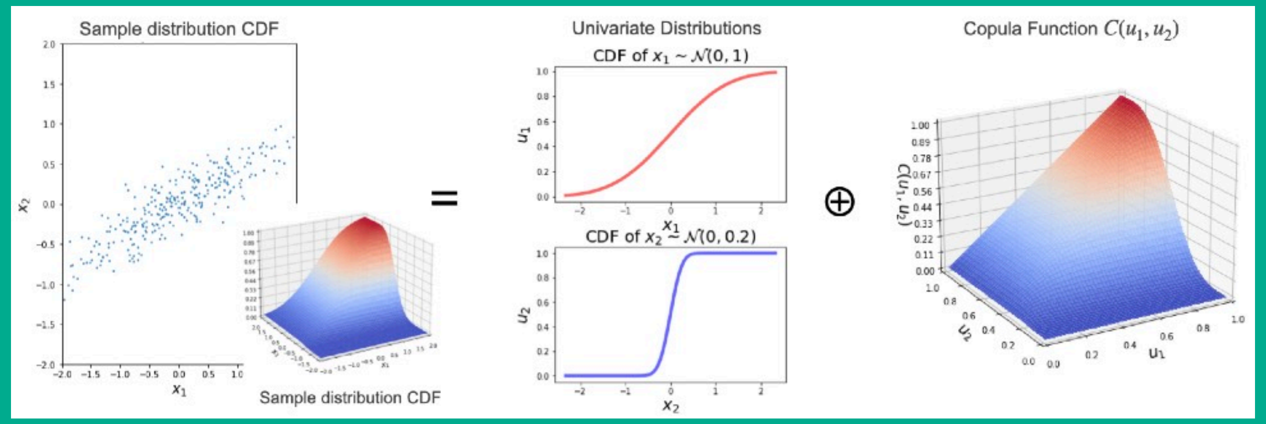
L4DC '2023, L4DC '2024





Leveraging structure in calibration

ICLR '2024, NeurIPS ‘2025

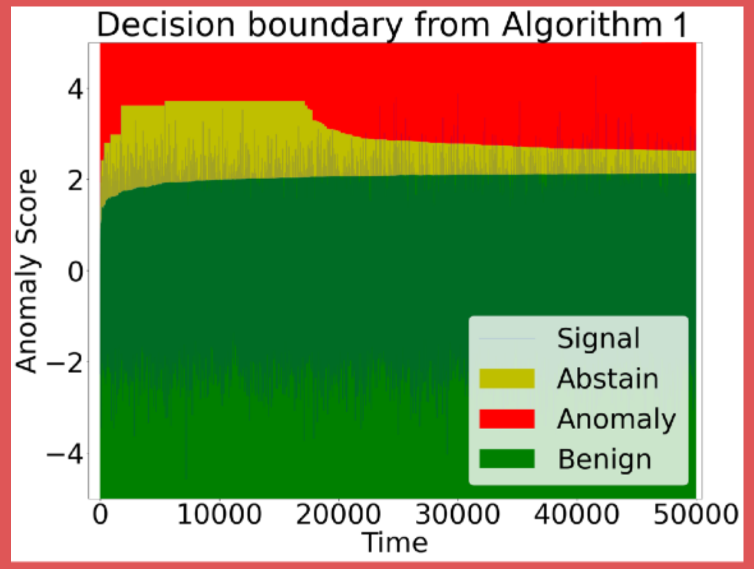


Decision making Under Uncertainty

Selective prediction

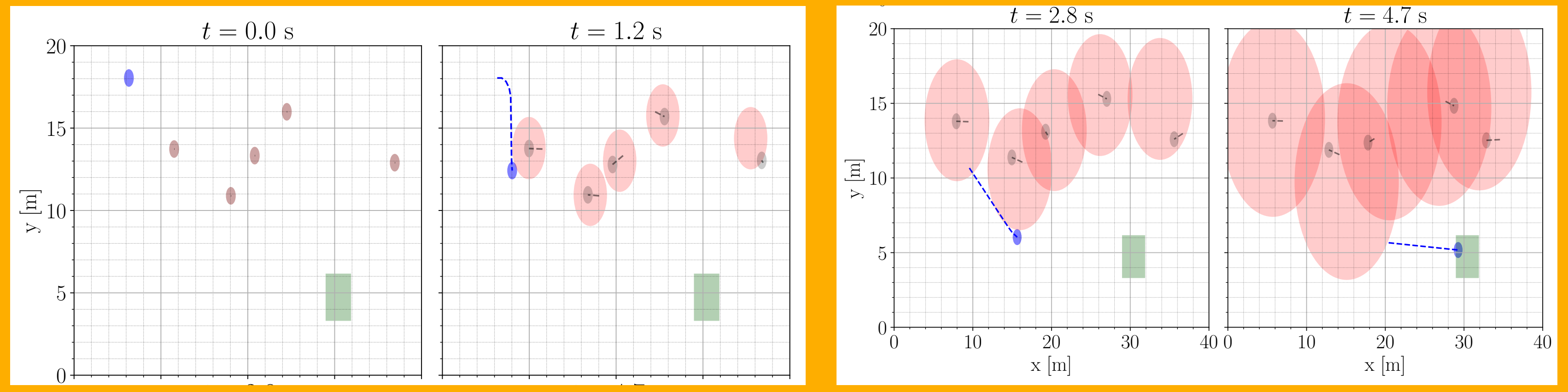
ICML ‘2024

ICLR LLM Reasoning and Planning workshop ‘2025



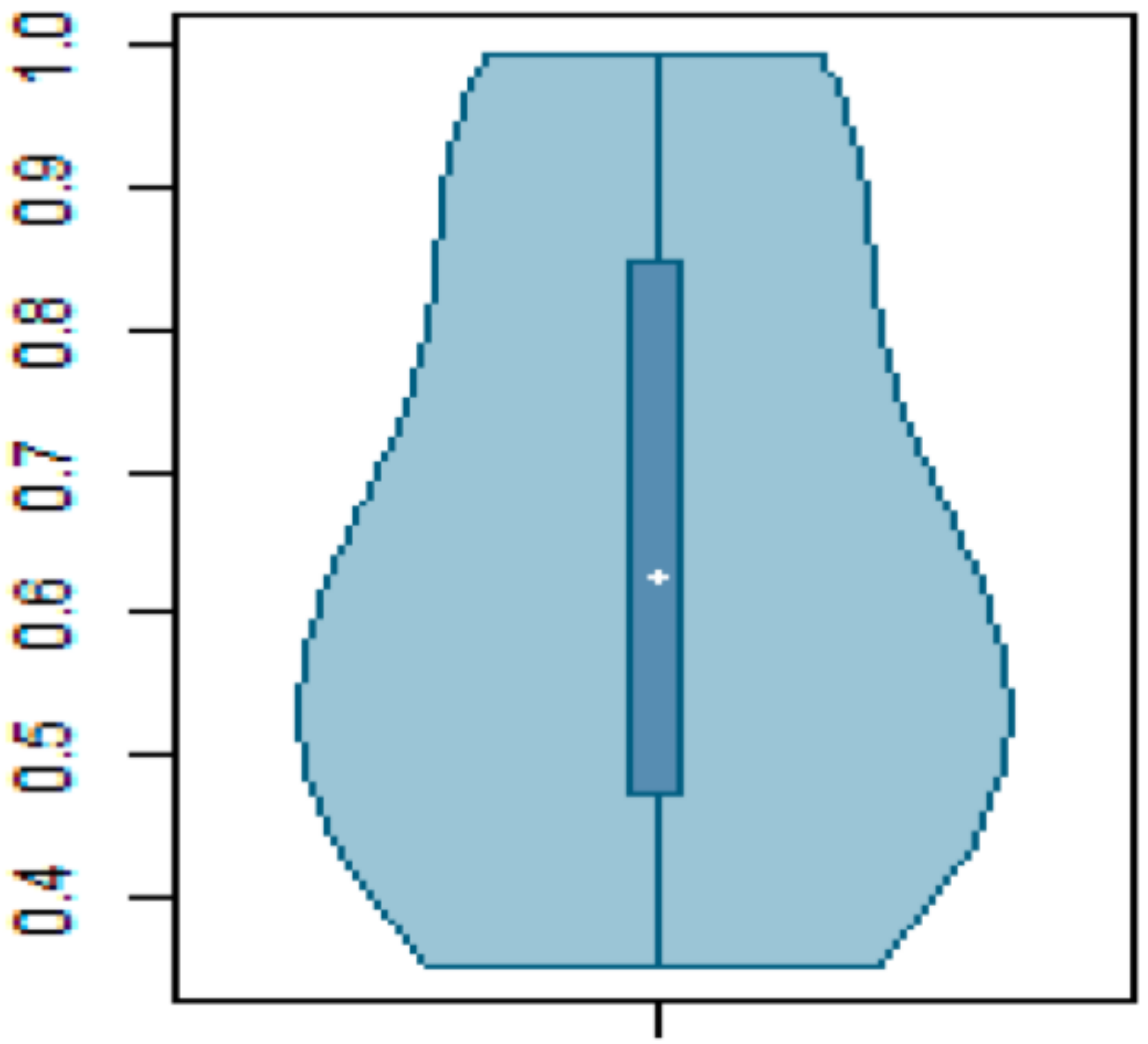
Safety constraints and pessimism

ICRA workshop '2023, ML4H ‘2025



Talk Outline

- Part I: Probabilistic Modeling and Uncertainty Quantification
  - Leveraging structure in model design
  - Leveraging structure in post-hoc calibration
- Part II: Decision making Under Uncertainty
  - Selective prediction
    - Example: No-mistake anomaly detection
  - Safety constraints and pessimistic planning
    - Example: Robot navigation
- Discussion and Conclusion

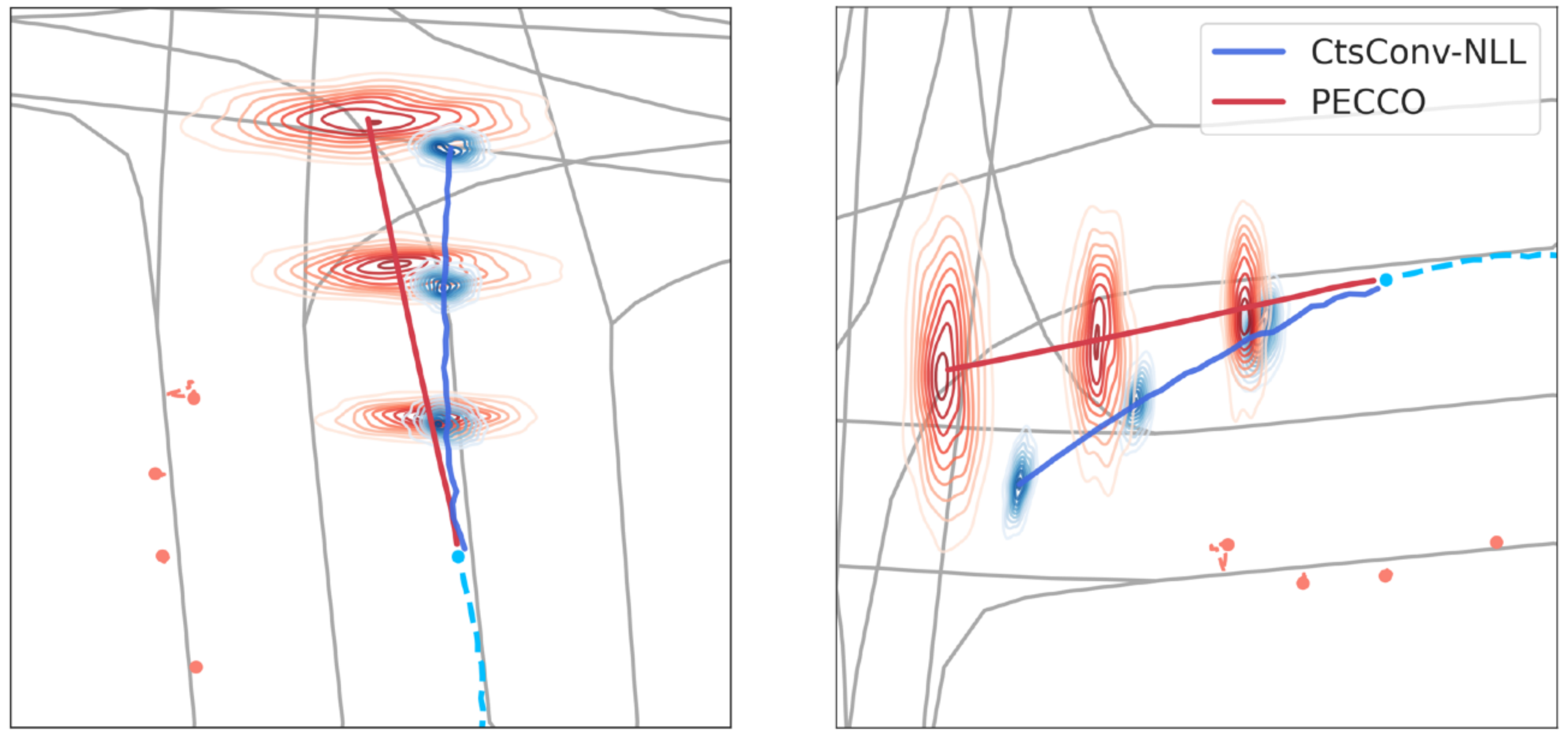


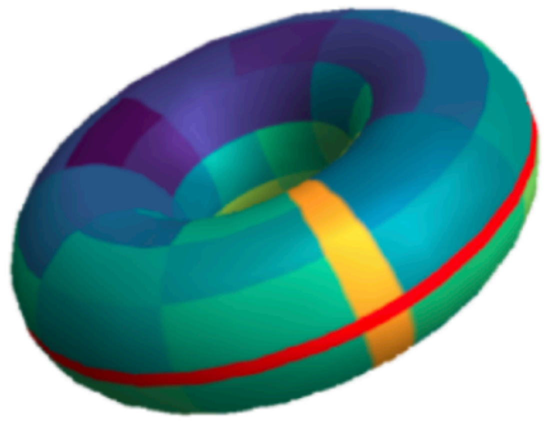
Talk Outline

- Part I: Probabilistic Modeling and Uncertainty Quantification
  - **Leveraging structure in model design**
  - Leveraging structure in post-hoc calibration
- Part II: Decision making Under Uncertainty
  - Selective prediction
    - Example: No-mistake anomaly detection
  - Safety constraints and pessimistic planning
    - Example: Robot navigation
- Discussion and Conclusion

Leveraging structure in model design

Equivariance

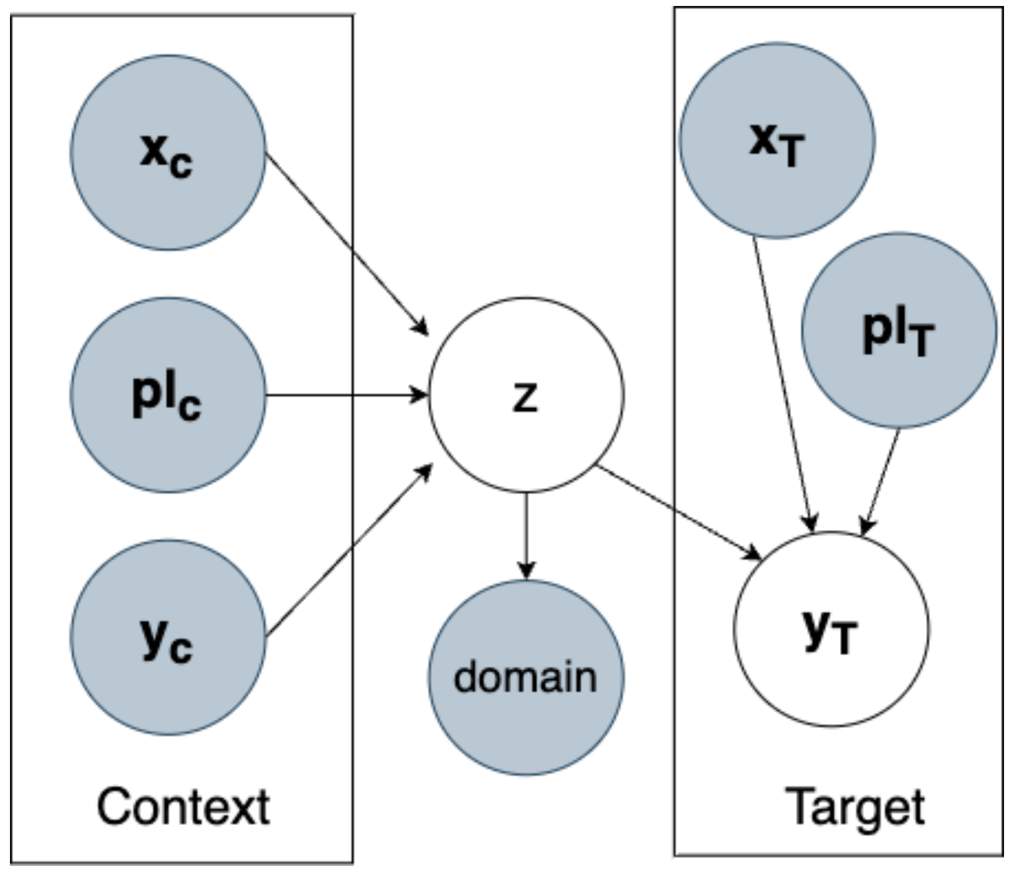


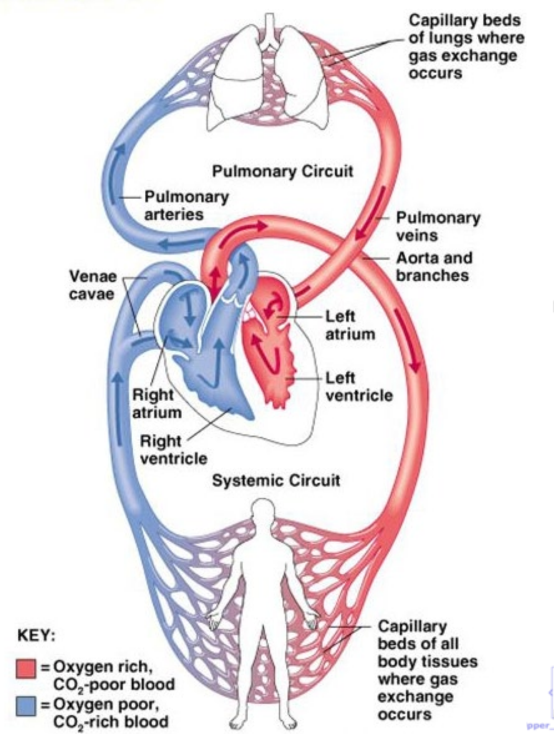


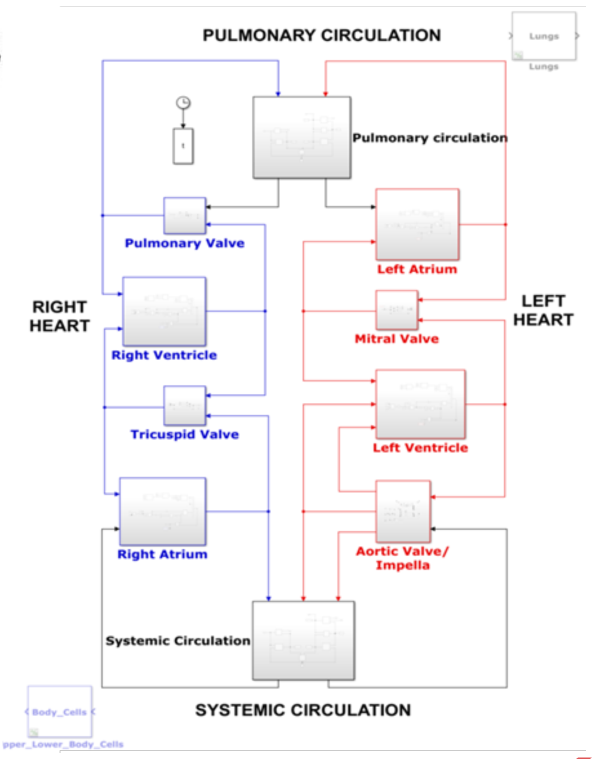
$$\mathcal{N}(\mu_t, \Sigma_t)$$

$$p_{\mu,\Sigma}(v) = p_{g\mu,\, g\Sigma g^T}(gv)$$

Domain-Adversarial

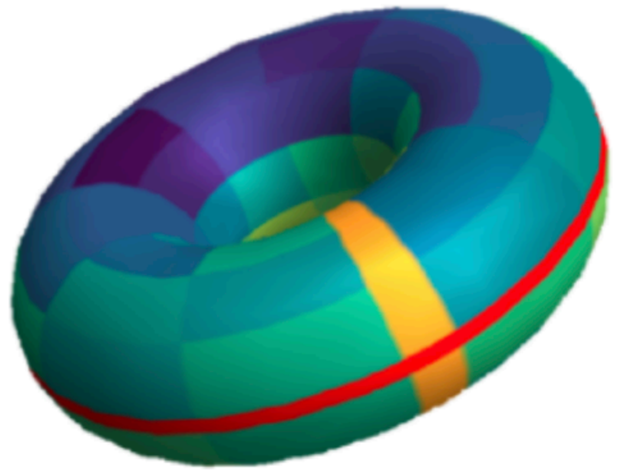






Equivariant Probabilities for Trajectory Prediction

Rotational Equivariance Definition

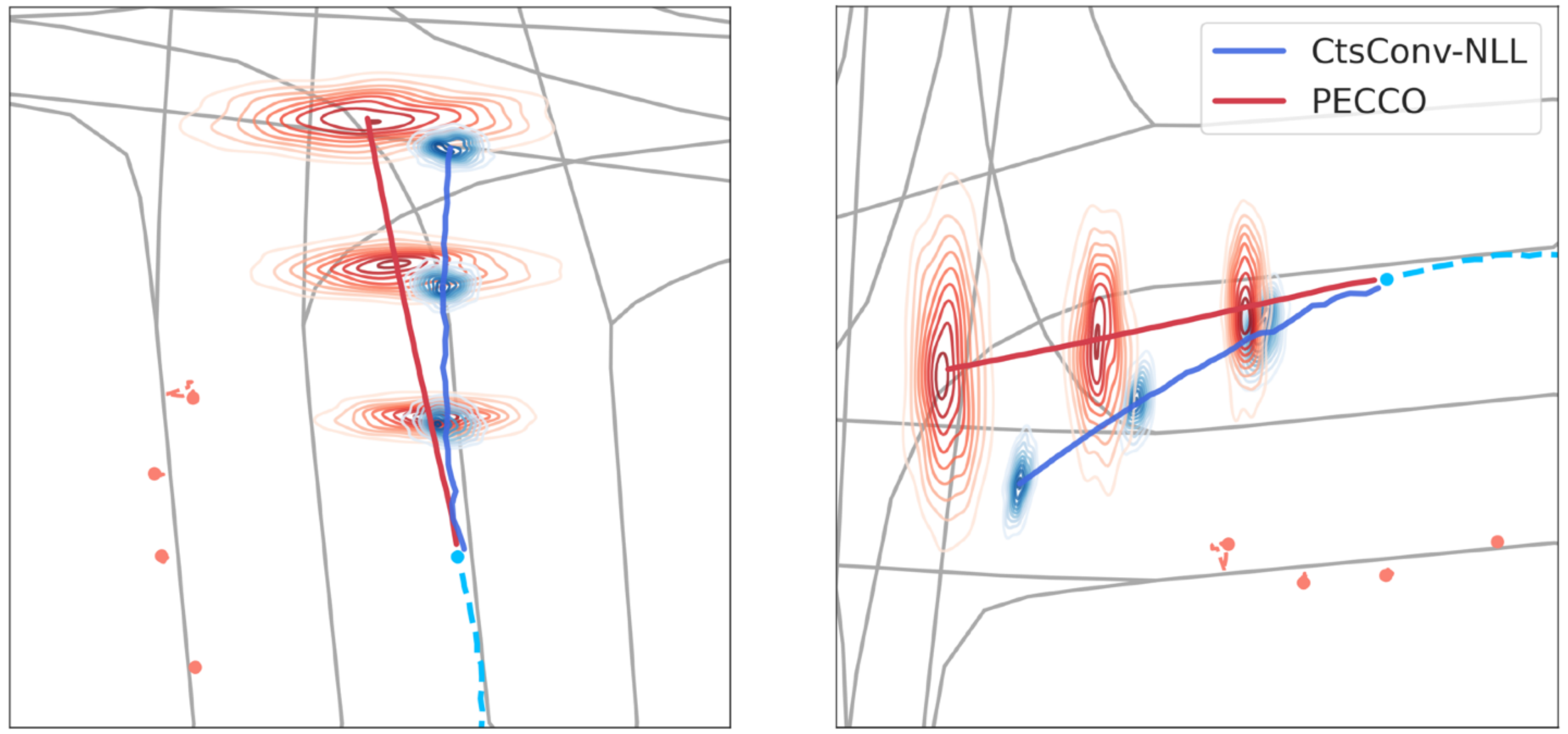


Given: trajectory $x_{1:t}$, environment covariant $\mathbf{e}$,

Learn: Probability p_θ over the next k steps of the trajectory $x_{t+1:t+k}$ as

$$p_\theta(x_{t+1:t+k} \mid x_{1:t}, \mathbf{e}) = p_\theta(g x_{t+1:t+k} \mid g x_{1:t}, g\mathbf{e})$$

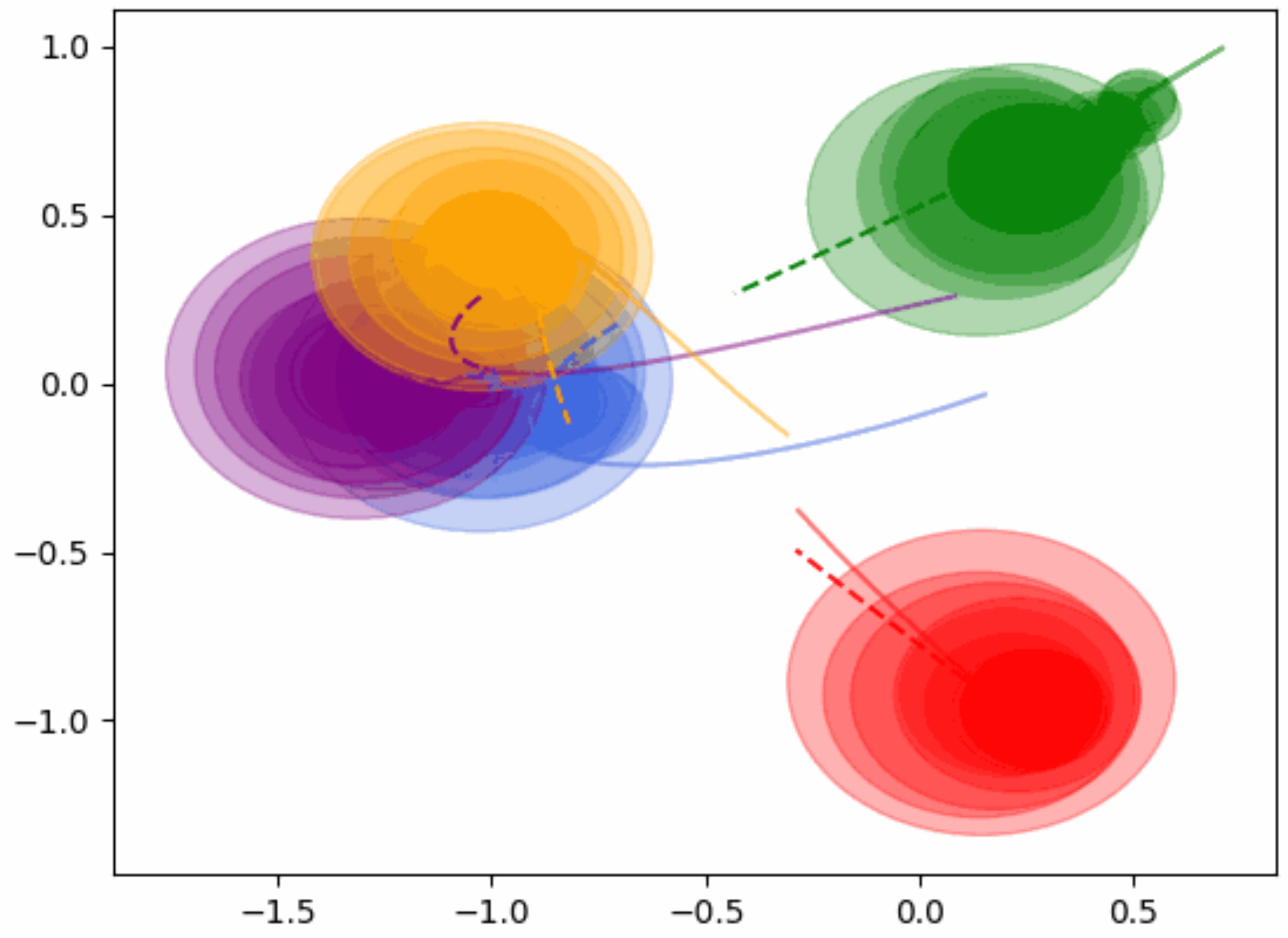
Where $g \in \mathrm{SO}(2)$: $\{\mathrm{Rot}_\theta : 0 \leq \theta < 2\pi\}$ the rotational symmetry group.



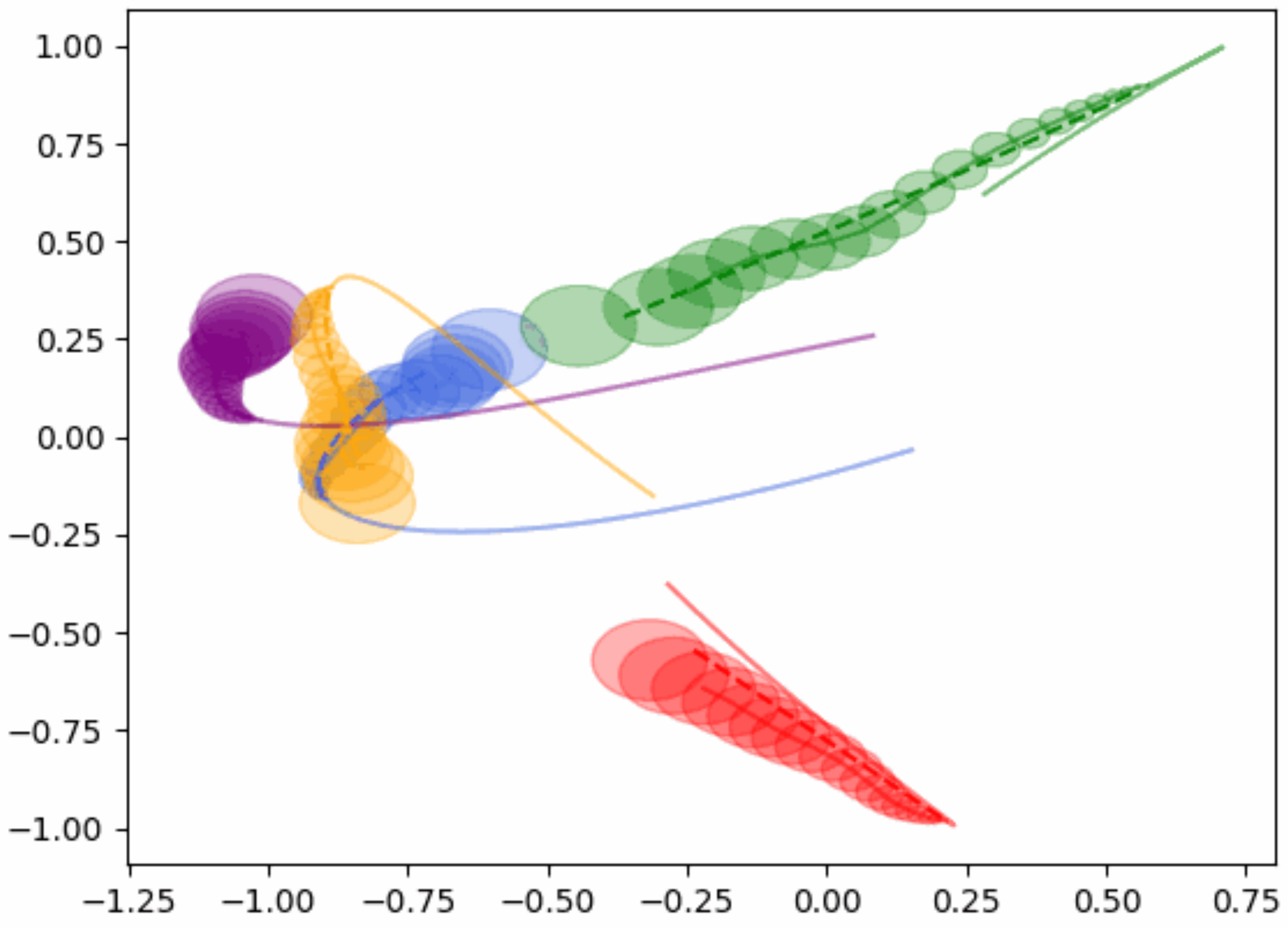
Prediction on the same scene rotated by 90 degrees.

PECCO

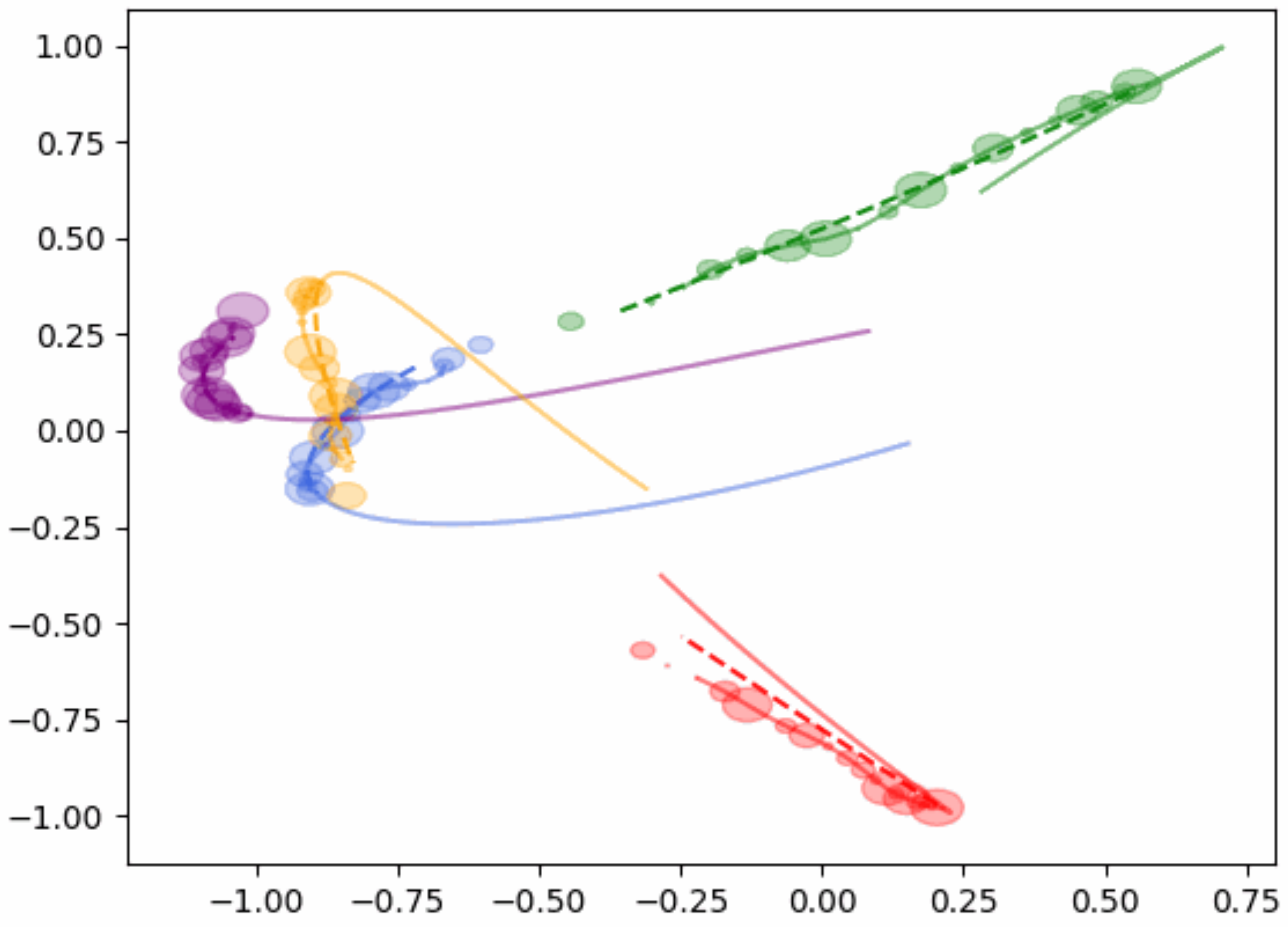
Results



LSTM



PECCO (ours)



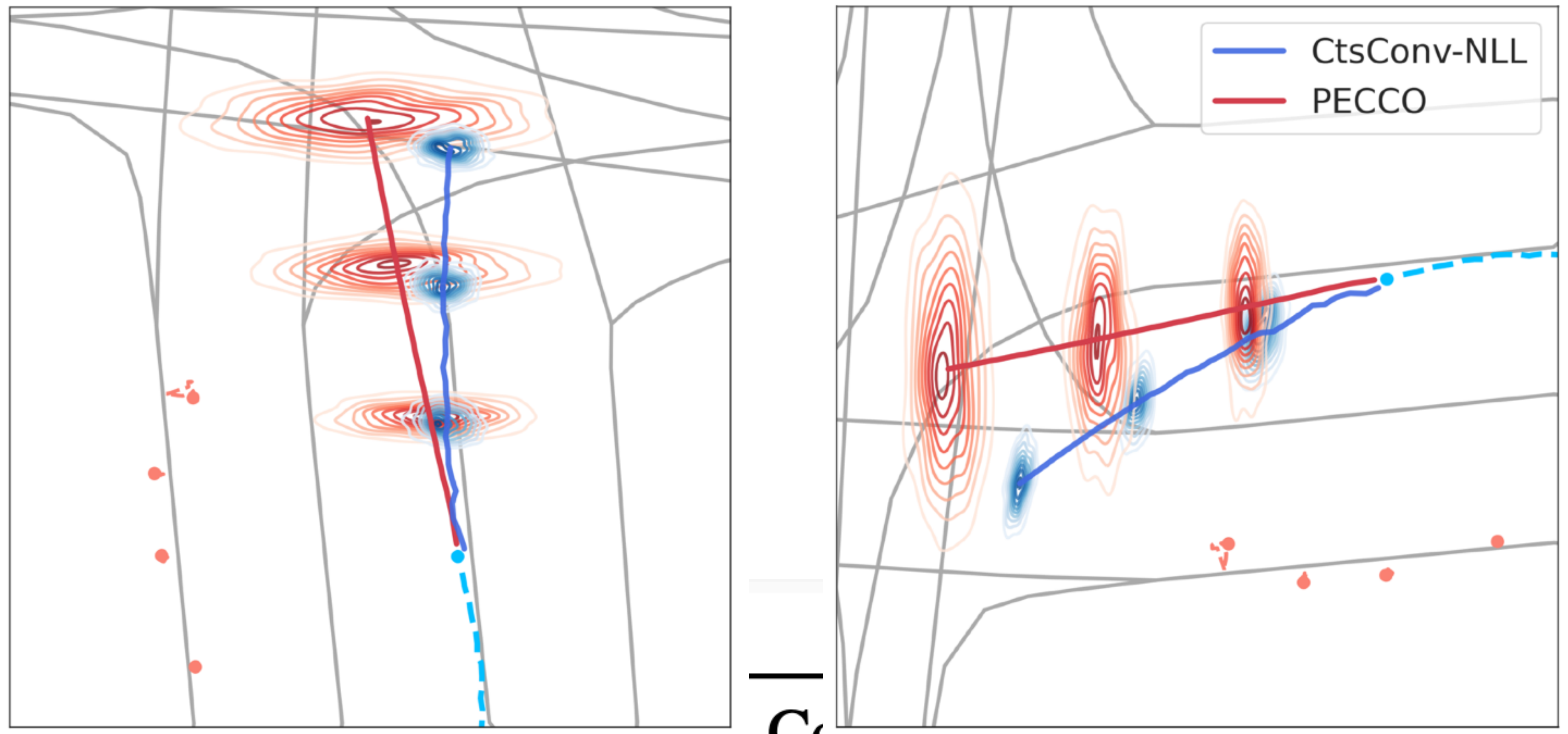
CtsConv [1]

[1] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. *Deep parametric continuous convolutional neural networks*. CVPR, 2018

PECCO

Results

Prediction on the same scene rotated by 90 degrees.



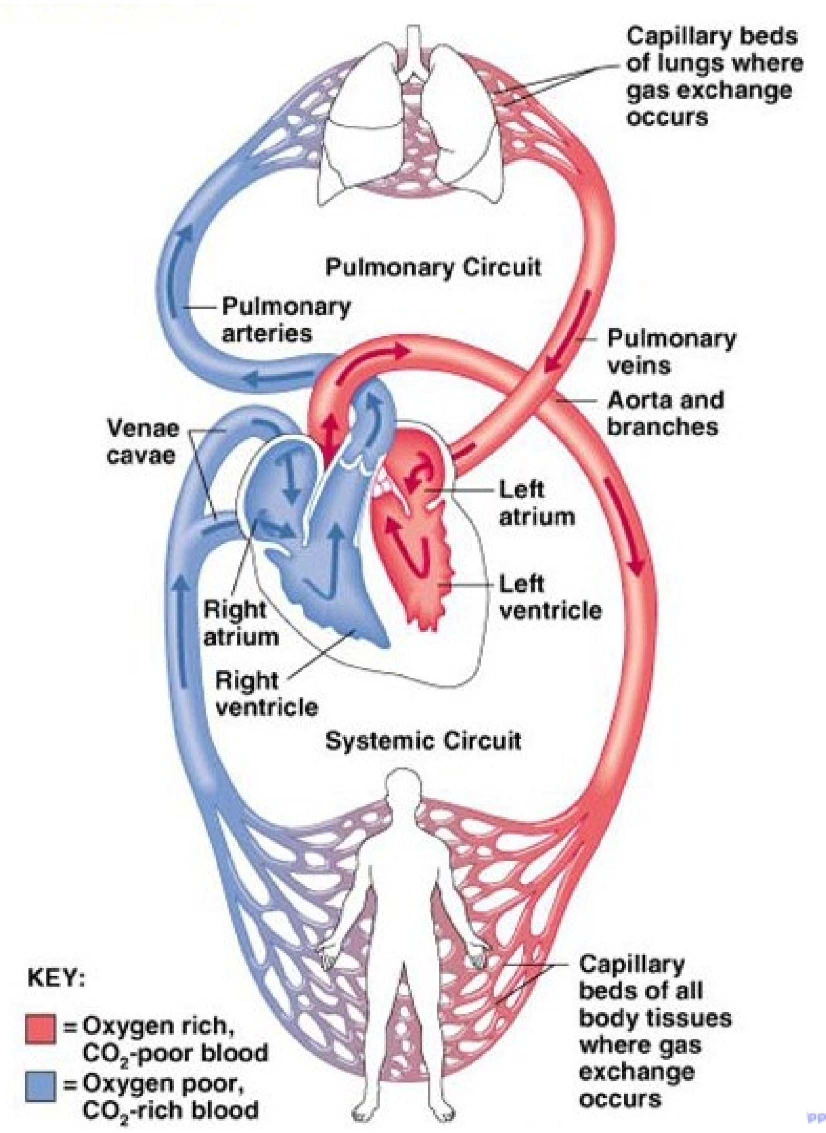
Model	minADE$_6$ ↓	minFDE$_6$ ↓	NLL↓	ES ↓	Cov@1s(%)	Cov@2s	Cov@3s
			Argoverse				
LSTM-NLL	1.64 ± .05	4.17 ± .10	3.07 ± .08	2.31 ± .54	8.8 ± 0.7	8.5 ± 0.7	7.0 ± 0.8
LSTM-NLL-aug	1.61 ± .02	4.15 ± .08	2.78 ± .03	1.99 ± .46	10.1 ± 1.5	10.5 ± 1.0	9.8 ± 1.9
CtsConv-NLL	1.74 ± .03	4.43 ± .06	29.1 ± 2.2	6.71 ± .70	6.3 ± 2.2	0.02 ± .01	0.01 ± .01
CtsConv-NLL-aug	1.66 ± .02	4.23 ± .06	11.81 ± .01	5.10 ± .35	11.9 ± 2.1	1.7 ± 0.5	0.02 ± .01
Trajectron++	1.83 ± .02	3.85 ± .07	**2.48** ± .27	3.92 ± .61	45.5 ± 5.3	37.6 ± 3.2	34.9 ± 2.5
MFP	1.53 ± .04	3.77 ± .06	3.56 ± .02	2.33 ± .21	51.3 ± 5.1	33.0 ± 4.9	8.3 ± 4.8
PECCO	**1.39** ± .02	**3.41** ± .03	4.26 ± 0.1	**1.54** ± .16	**74.9** ± 0.6	**78.6** ± 2.8	**84.5** ± 2.9
			TrajNet++				
LSTM-NLL-aug	0.85 ± .02	1.64 ± .03	2.78 ± .02	-0.28 ± .09	29.0 ± 4.3	23.2 ± 4.2	23.7± 3.9
CtsCov-NLL	1.08 ± .02	2.36 ± .09	5.33 ± .08	1.67 ± .13	43.8 ± 10.6	20.7 ± 5.2	12.2 ± 6.7
CtsCov-NLL-aug	0.92 ± .01	1.76 ± .03	6.74 ± .21	1.42 ± .11	62.1 ± 3.3	36.3 ± 4.9	34.1 ± 5.8
Trajectron++	1.14 ± .03	2.31 ± .05	2.83 ± .12	0.98 ± .17	50.2 ± 2.2	45.8 ± 3.5	32.9 ± 3.5
MFP	0.85 ± .02	1.70 ± .04	**2.20** ± .04	0.67 ± .08	79.1 ± 4.3	32.5 ± 3.1	22.8 ± 3.2
PECCO	**0.59** ± .12	**1.06** ± .17	2.37 ± .04	**-0.73** ± .10	**80.8** ± 4.5	**85.9** ± 2.3	**94.5** ± 3.0

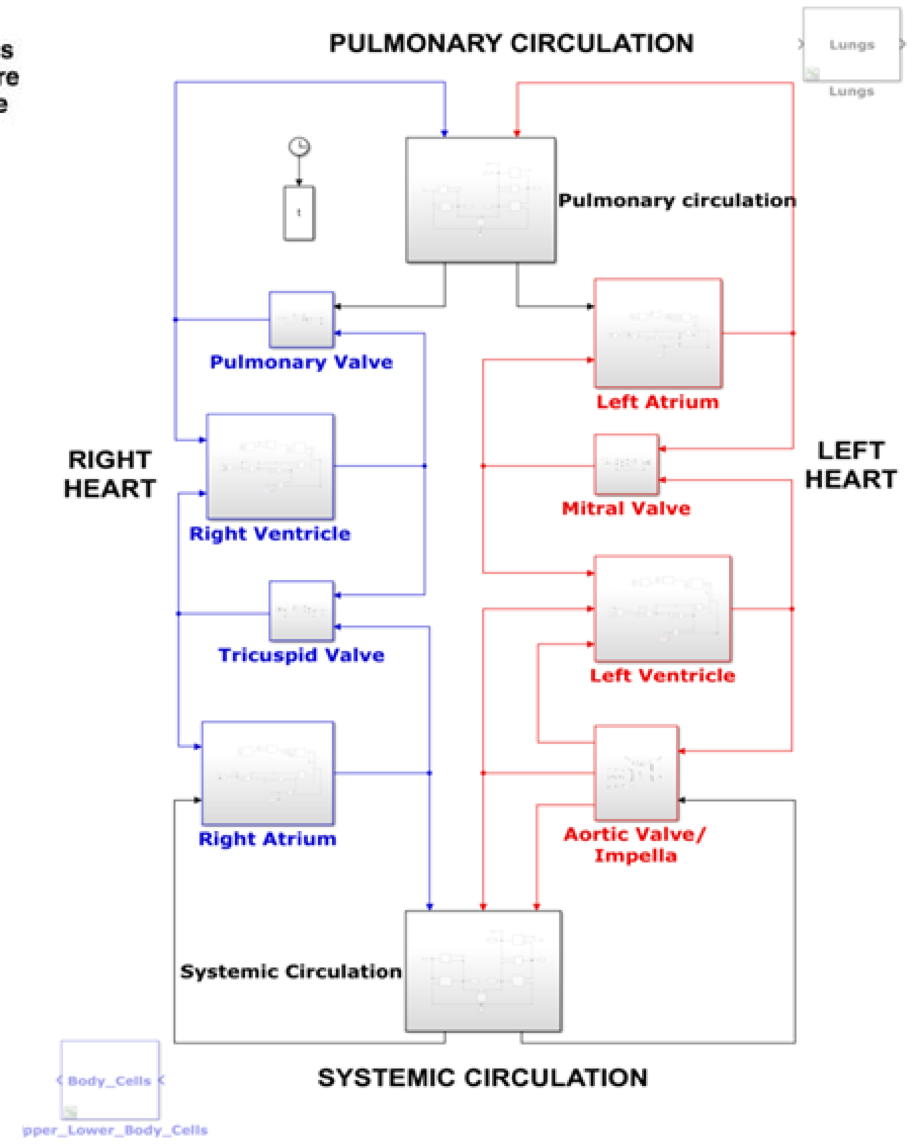
Domain-Adversarial Neural Process

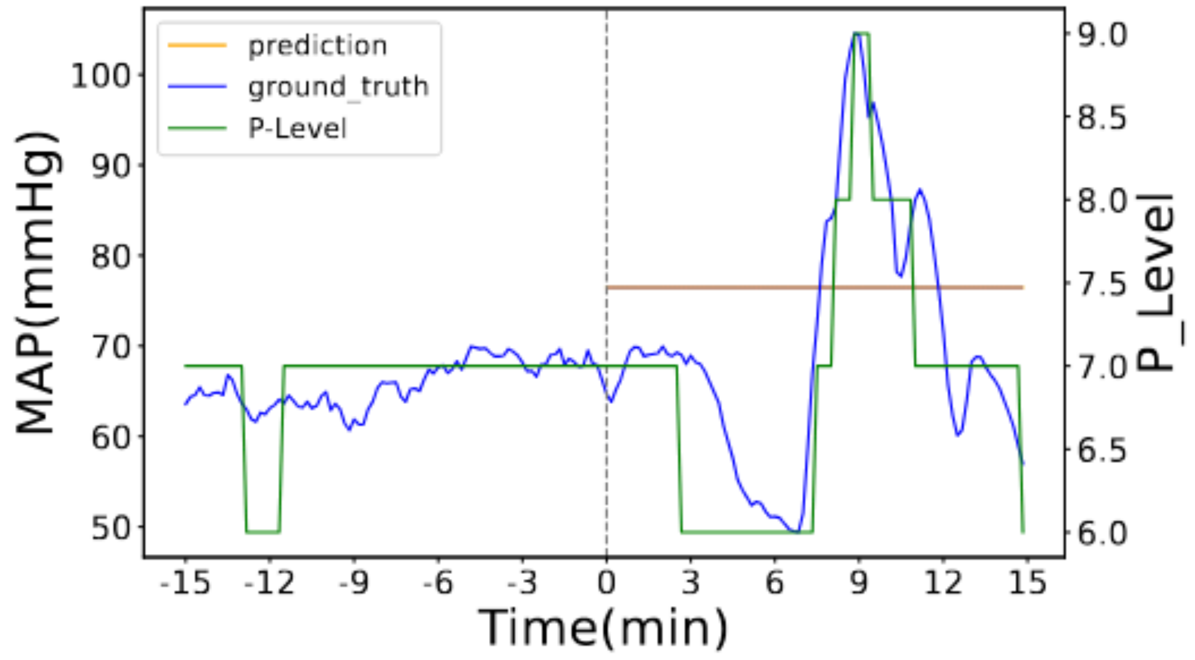
Results

Method	MAE (mmHg) ↓	MAE (inc) ↓	MAE (dec) ↓	MAE (stat) ↓	Trend Acc ↑
MLP	7.97 ± .26	9.04 ± .68	10.96 ± .61	6.78 ± .43	0.57 ± .03
CLMU	6.93 ± .11	8.65 ± .56	8.47 ± .24	5.51 ± .04	0.65 ± .01
NP direct transfer	7.36 ± .91	9.72 ± 1.23	8.79 ± 1.06	6.25 ± .95	0.64 ± .00
NP no sim	8.68 ± .06	**6.90** ± .01	15.34 ± .02	7.63 ± .01	0.52 ± .00
DANP (ours)	**6.65** ± .13	**6.94** ± .10	**8.46** ± .17	**5.36** ± .09	**0.70** ± .01

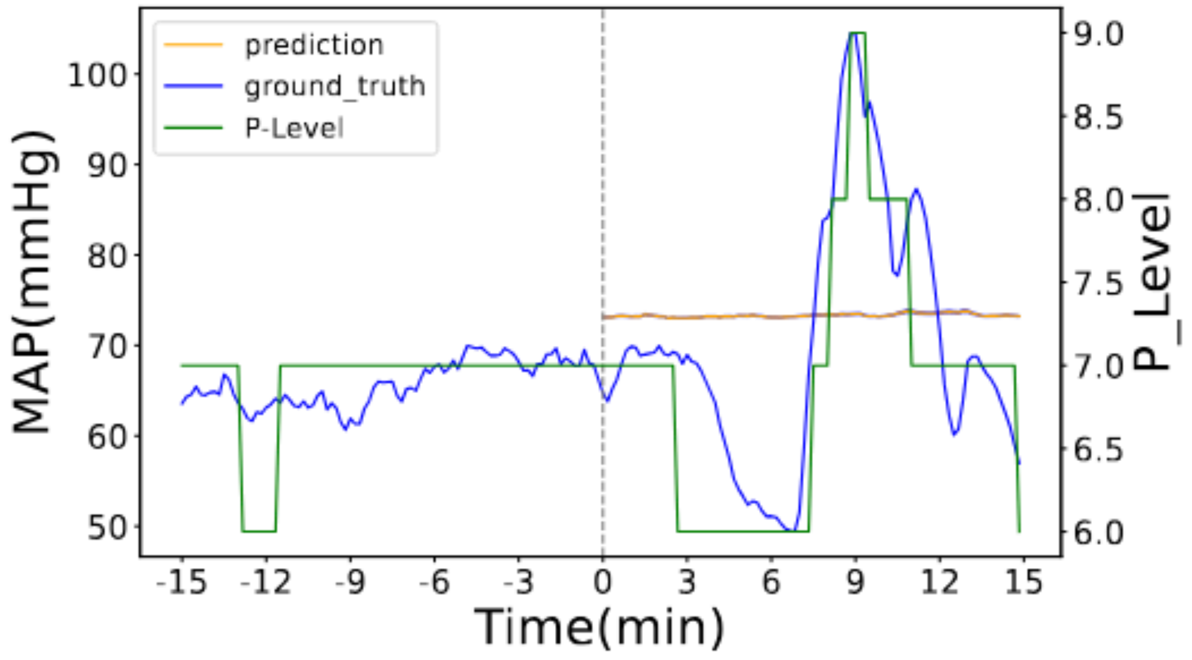
Table 3.1. Empirical results in terms of Mean Average Error (MAE) for data with increasing (inc), decreasing (dec), stationary (stat) trends, and trend prediction accuracy. DANP achieves significantly lower performs significantly better on trending data compared to baselines.



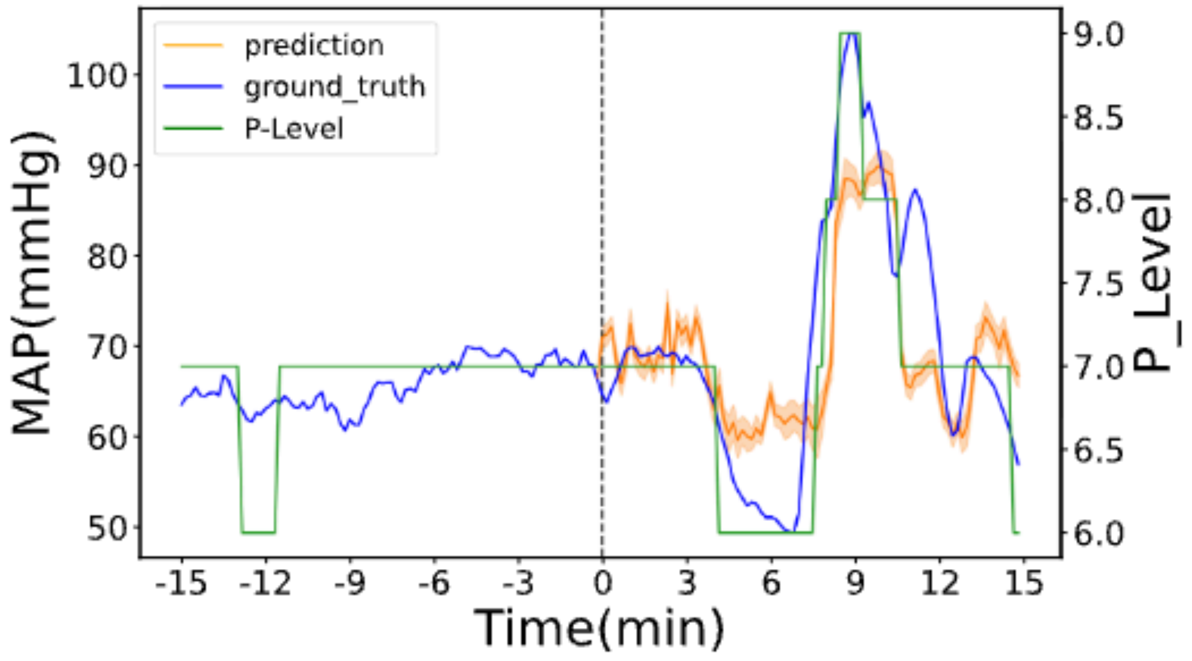




(a) NP direct transfer



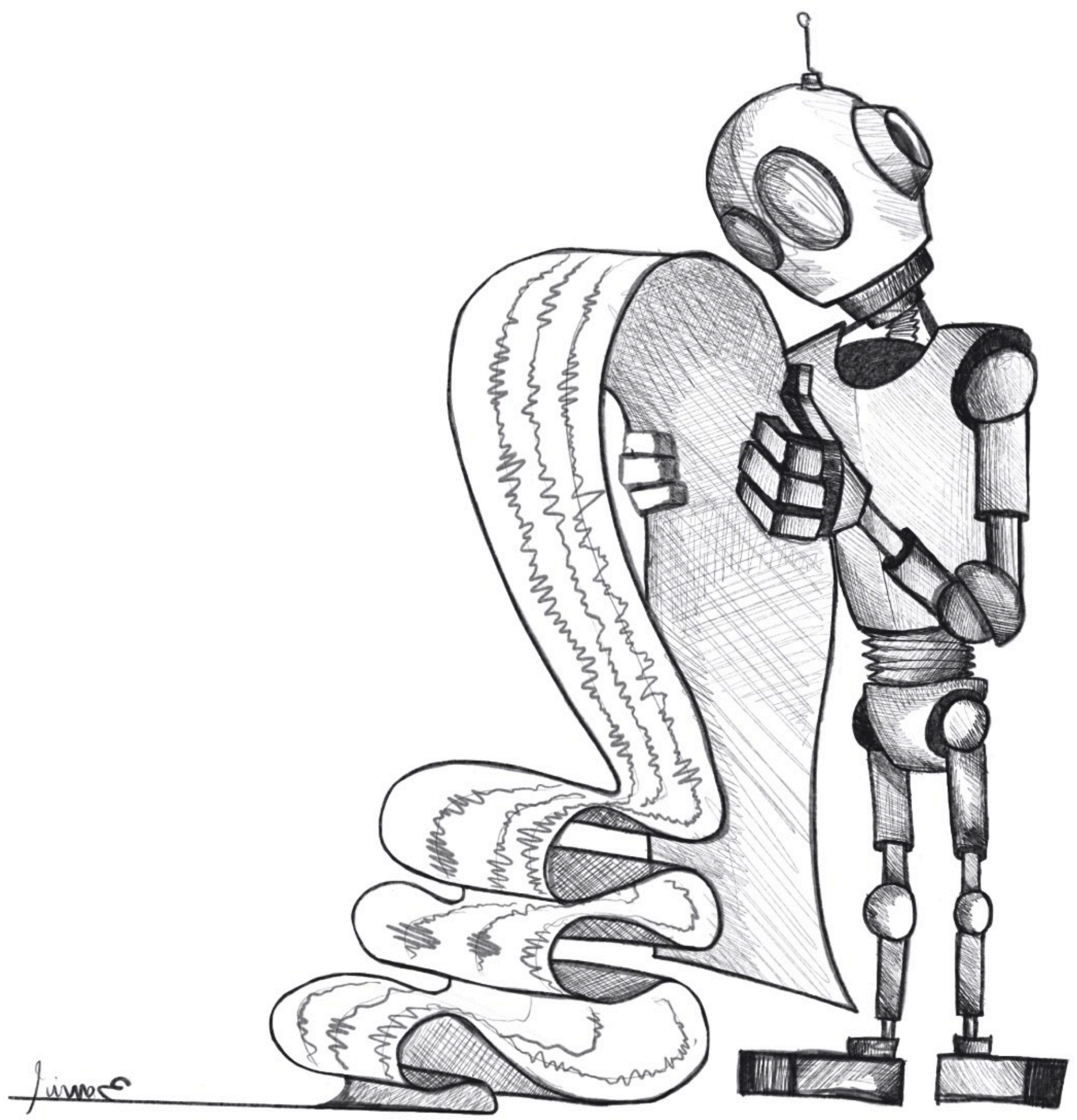
(b) CLMU



(c) DANP (ours)

Takeaway of this section

- Probabilistic models trained directly with NLL or CRPS is over-confident
- Incorporating structure (equivariance / domain knowledge) improves calibration
- For time-series data, it is very hard to calibrate the forecasts consistently just through model training.



Talk Outline

- Part I: Probabilistic Modeling and Uncertainty Quantification
  - Leveraging structure in model design
  - **Leveraging structure in post-hoc calibration**
- Part II: Decision making Under Uncertainty
  - Selective prediction
    - Example: No-mistake anomaly detection
  - Safety constraints and pessimistic planning
    - Example: Robot navigation
- Discussion and Conclusion

Quick intro of Conformal Prediction

Prediction sets

Figure 1: Prediction set examples on Imagenet. *We show three progressively more difficult examples of the class* ***fox squirrel*** *and the prediction sets (i.e., $\mathcal{C}(X_{\text{test}})$) generated by conformal prediction.*

Quick intro of Conformal Prediction

Split conformal prediction

- Nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $s(x, y) = \|y - \hat{f}(x)\|$)
- Calibration dataset $D_{cal} \sim \mathcal{D}^n$, confidence level $1 - \alpha$

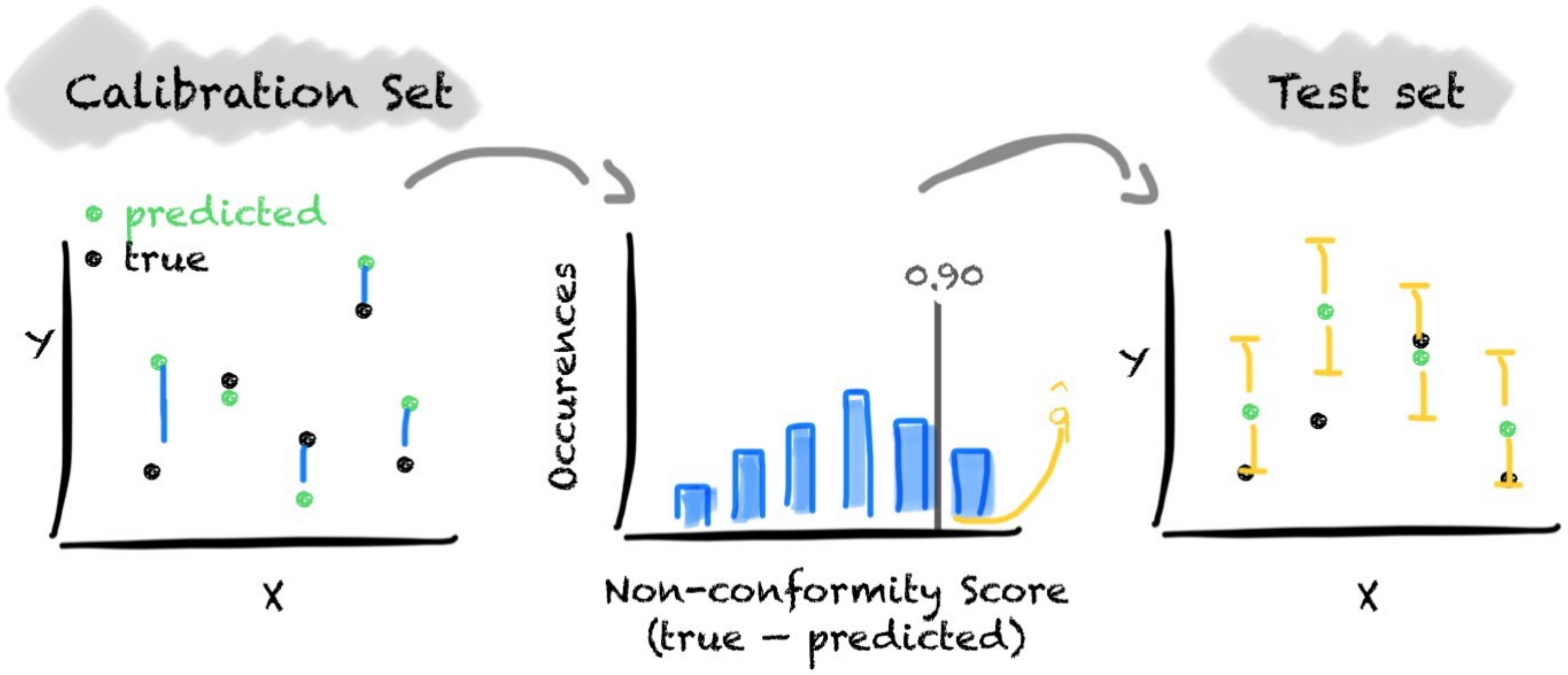
Algorithm 20 SplitConformal(D, s, α)

Let τ be the smallest value such that:

$$\sum_{i=1}^{n} \mathbb{1}[s(x_i, y_i) \leq \hat{q}] \geq (1 - \alpha)(n + 1)$$

i.e. $\hat{q}$ is an empirical $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of D.
Output the function:

$$\Gamma(x) = \{\hat{y} : s(x, \hat{y}) \leq \hat{q}\}$$



Quick intro of Conformal Prediction

Marginal Coverage Guarantees

- Nonconformity score function $s : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ (e.g. $s(x, y) = \|y - \hat{f}(x)\|$)

- For a new sample $(X_{test}, Y_{test}) \sim \mathscr{D}$, we have

$$1 - \alpha \leq \mathbb{P}(Y_{\text{test}} \in \Gamma(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1}$$

$D_{cal} \sim \mathscr{D}^n, (X_{test}, Y_{test}) \sim \mathscr{D}$

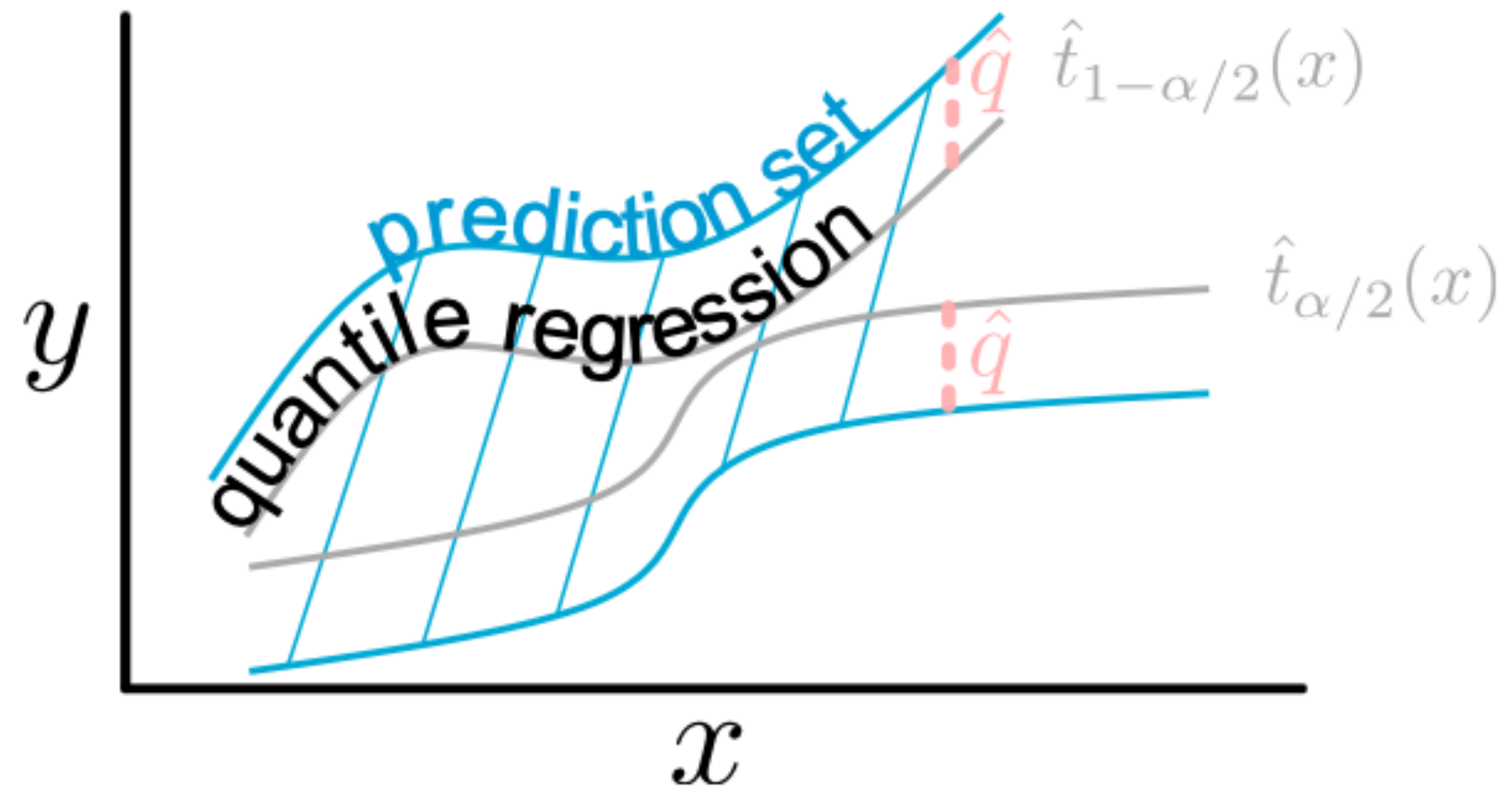
Quick intro of Conformal Prediction

Nonconformity Score

- Nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $s(x, y) = \|y - \hat{f}(x)\|$)

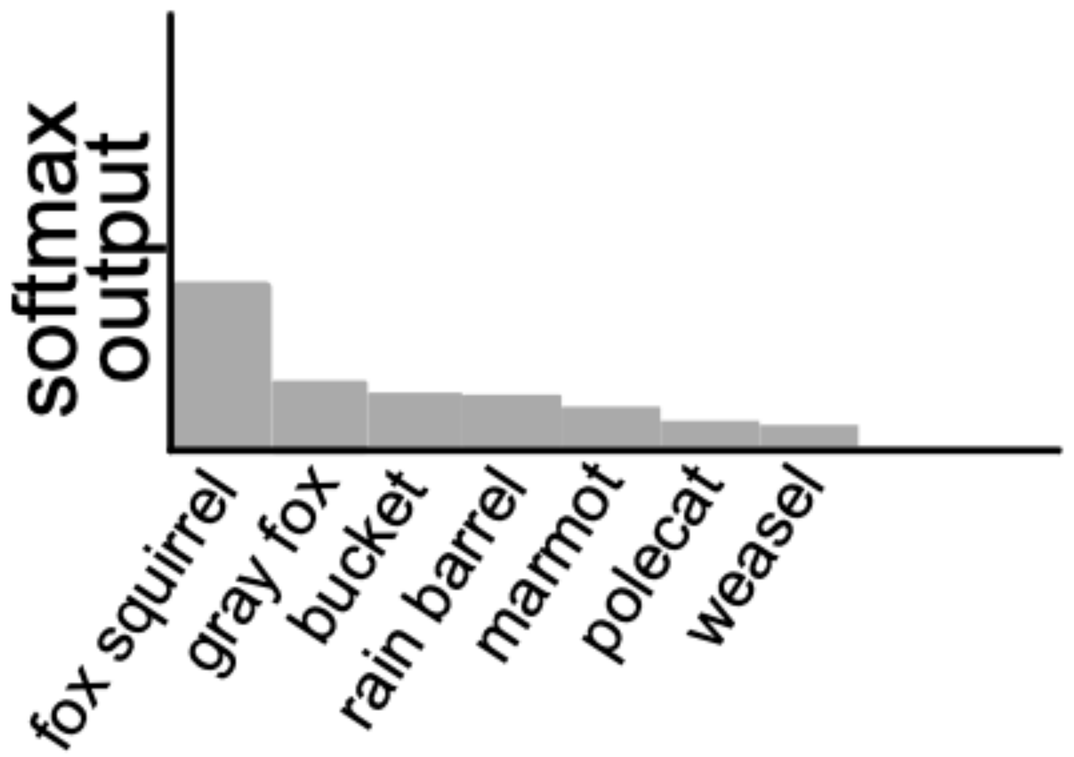
For Quantile Regression

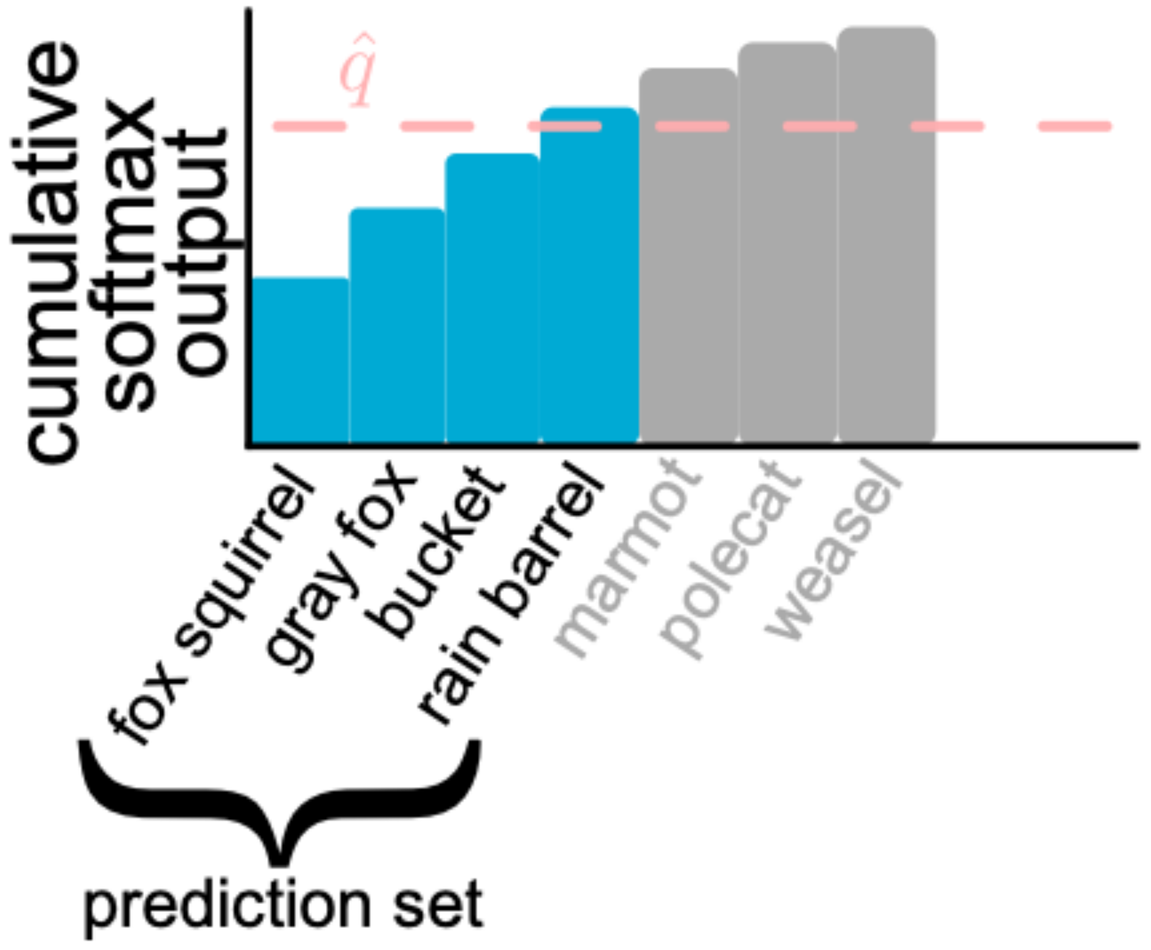
$$s(x, y) = \max\left\{\hat{t}_{\alpha/2}(x) - y, y - \hat{t}_{1-\alpha/2}(x)\right\}$$



For Multi-class Classification

$$s(x, y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}, \text{ where } y = \pi_k(x)$$





Quick intro of Conformal Prediction

Round Down

+ Model and distribution agnostic

+ Powerful results that normal ML cannot easily warrant

  Finite sample guarantees

  Group-conditional calibration for fairness

  Multi-calibration

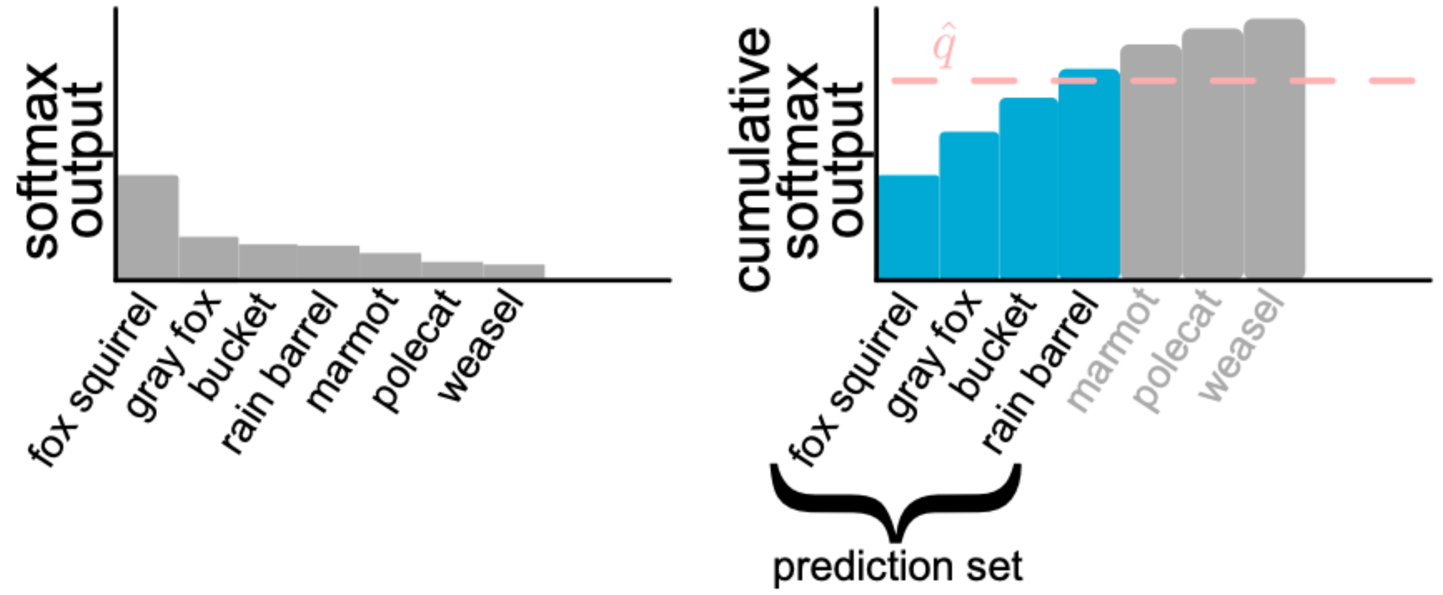
  Distribution shifts

  ……

- usefulness highly depends on the underlying model and score design.
- Marginal guarantees only.

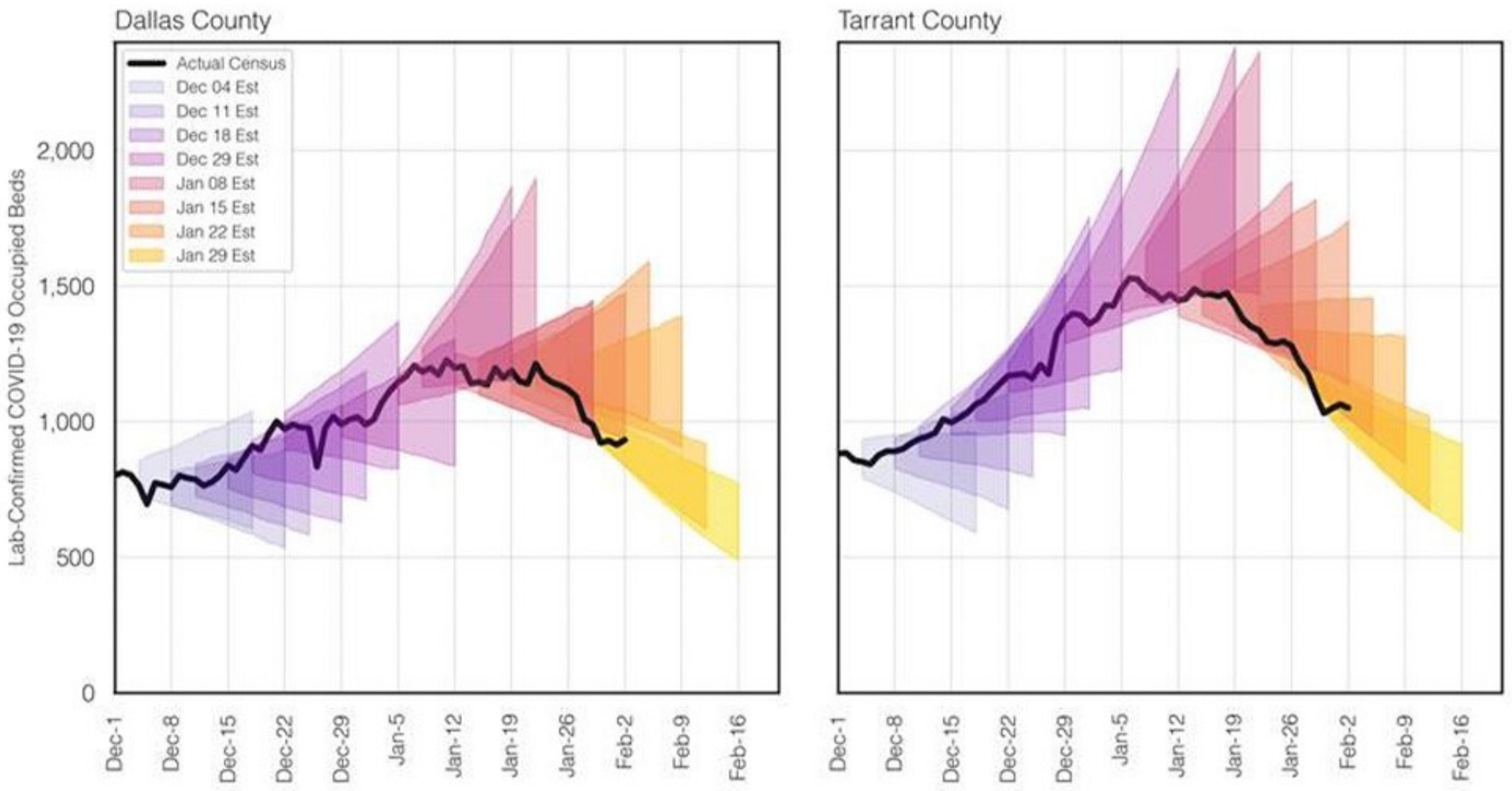
Multi-class Classification

$$s(x,y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}, \text{ where } y = \pi_k(x)$$

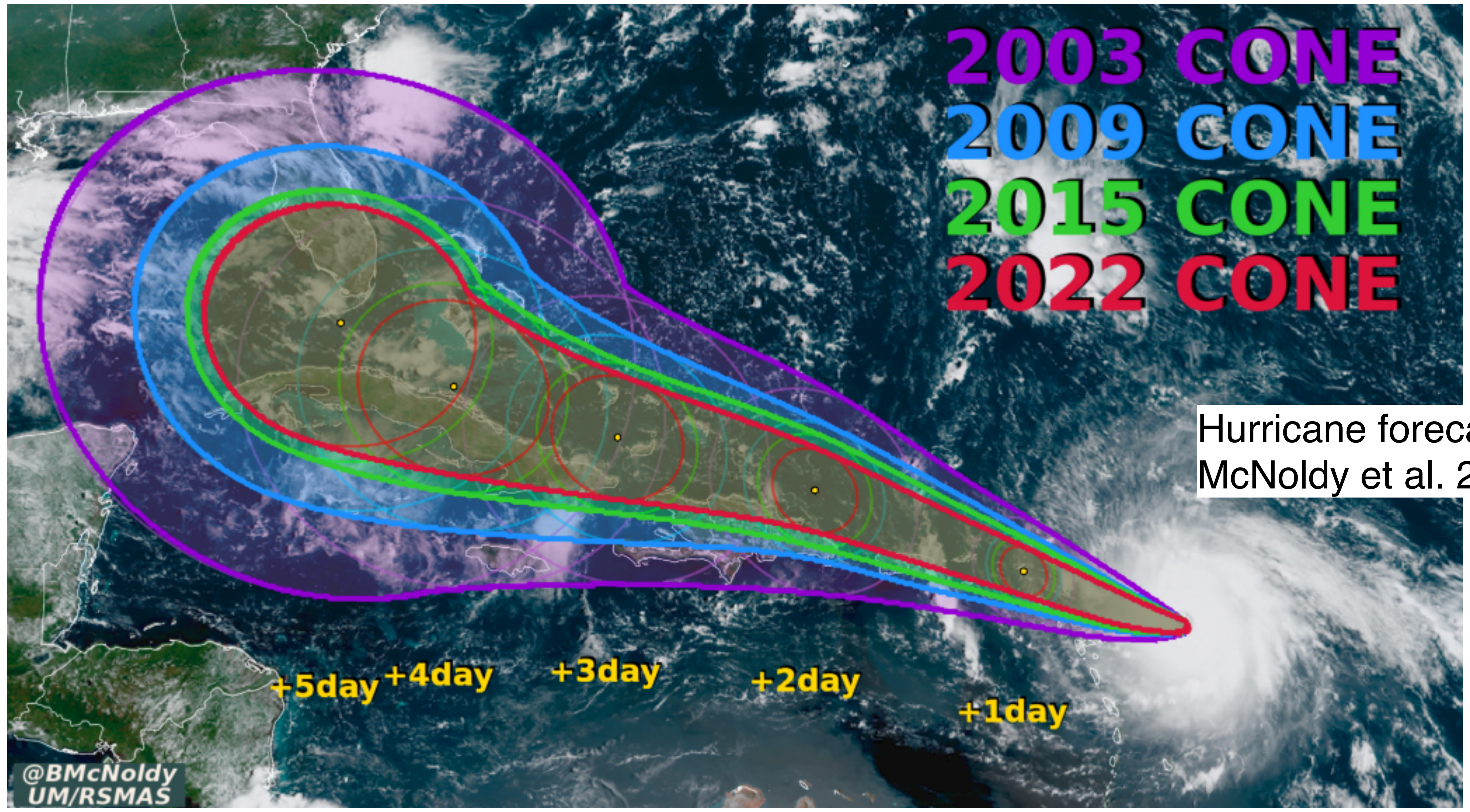


Back to Sophia's Work!

CP Work 1, Copula Conformal Prediction [ICLR2024]



Covid Forecasts. Patrick McGee / UT Southwestern 2021



Hurricane forecasting
McNoldy et al. 2022

Dataset $\mathscr{D} = \{(\mathbf{x}_{1:t}^{(i)}, \mathbf{y}_{t+1:t+k}^{(i)})\}_{i=1}^{n}$

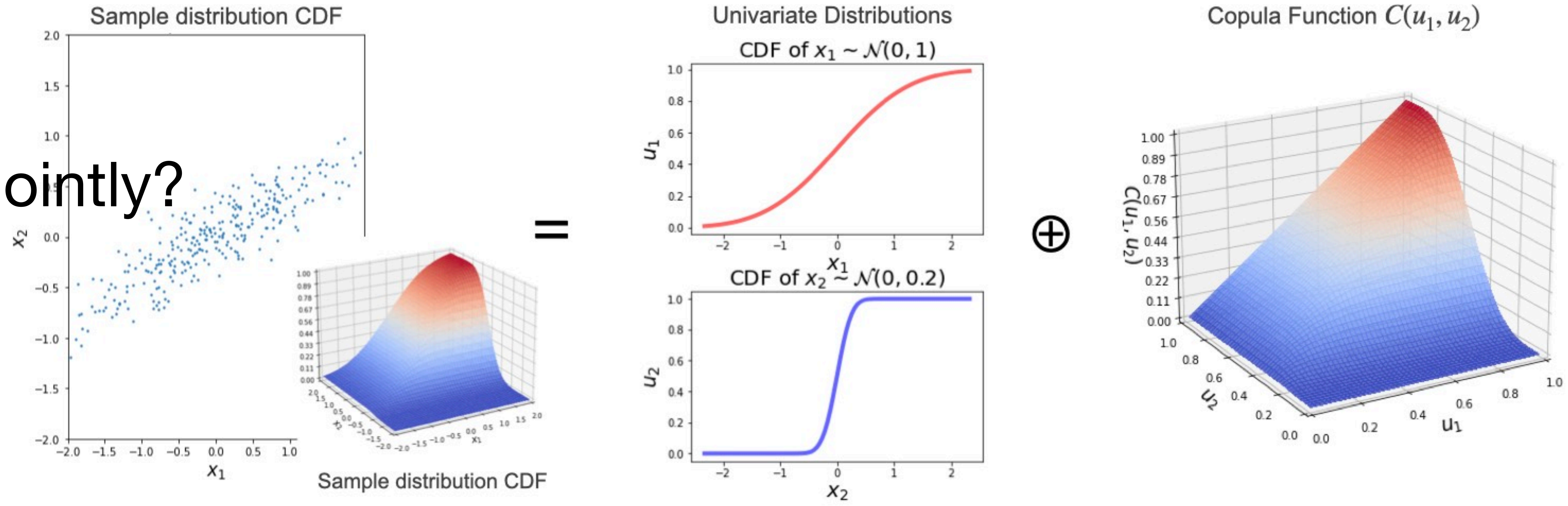
Goal: "Cone of uncertainty" valid for all time steps of $\mathbf{y}$

$$\mathbb{P}[\,\forall h \in \{1,\ldots,k\}, \mathbf{y}_{t+h} \in \Gamma_h^{1-\alpha}\,] \geq 1-\alpha$$

Copula Conformal Prediction for Time Series

How can we model the distributions jointly?

Idea: **Copulas**



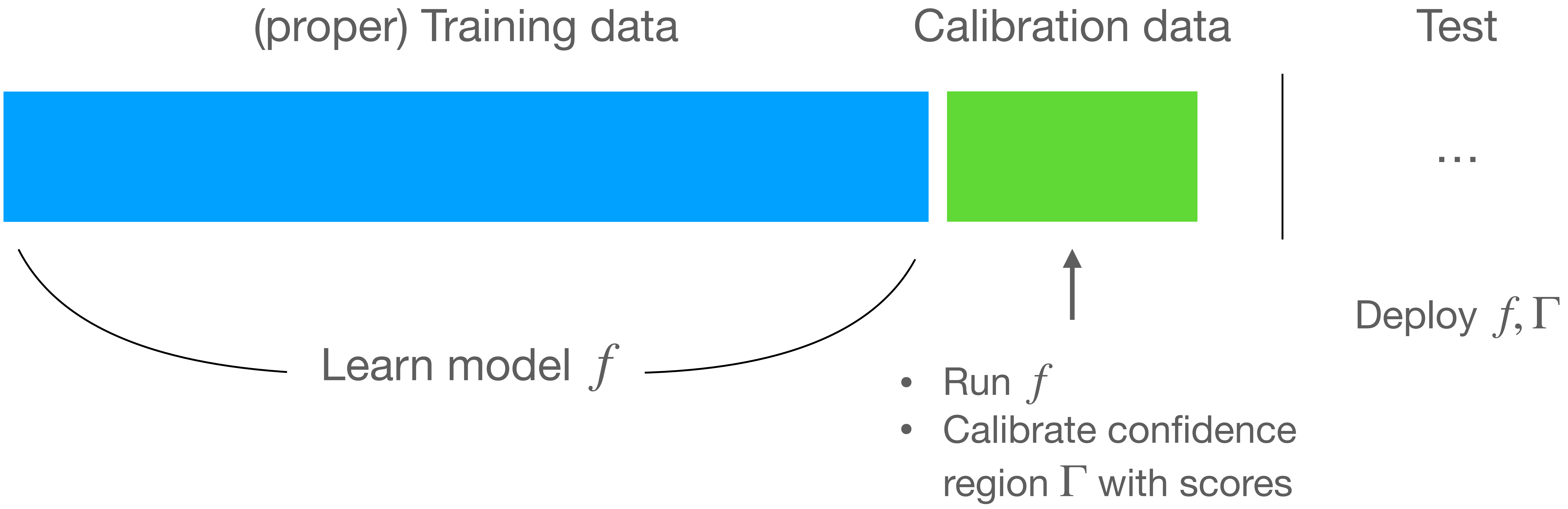
A copula is a function that synthesizes multiple CDFs to a joint CDF

$$C(u_1, \cdots, u_k) = \mathbb{P}(U_1 \leq u_1, \cdots, U_k \leq u_k)$$

$$F(x_1, \cdots, x_k) = C(F_1(x_1), \cdots, F_k(x_k)) \text{ (Sklar's theorem)}$$

For joint coverage guarantees, we only have to calibrate for the Copula.

Conformal Prediction (original algorithm)



Copula Conformal Prediction

We prove that it also has finite-sample validity guarantee 📢

(proper) Training data

Per time step Calibration

Copula Calibration

Test

...

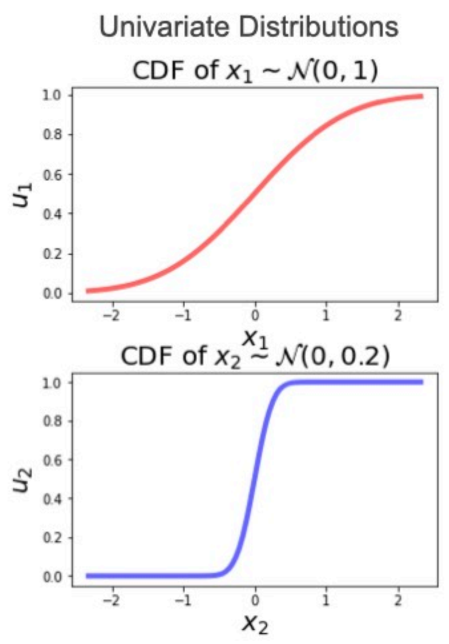
Learn model f

Calibrate $\hat{F}_1, \ldots, \hat{F}_h$

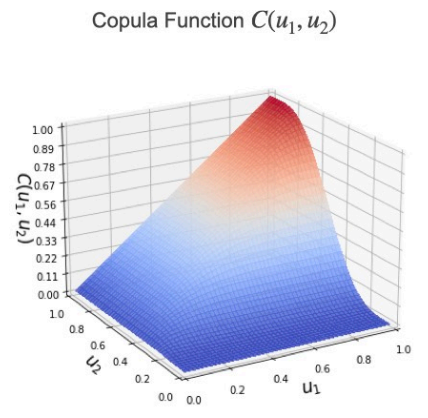
Calibrate Copula $C(\hat{F}_1(x_1), \cdots, \hat{F}_k(x_k))$

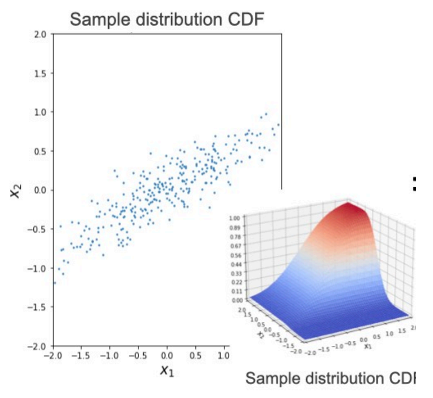
Use C to find appropriate Γ

Deploy f, Γ



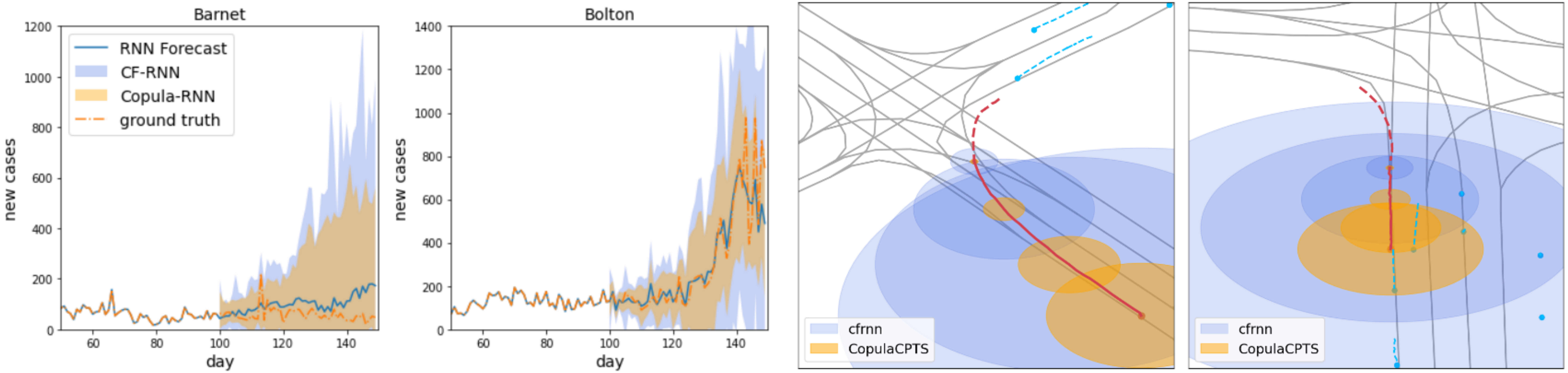
⊕





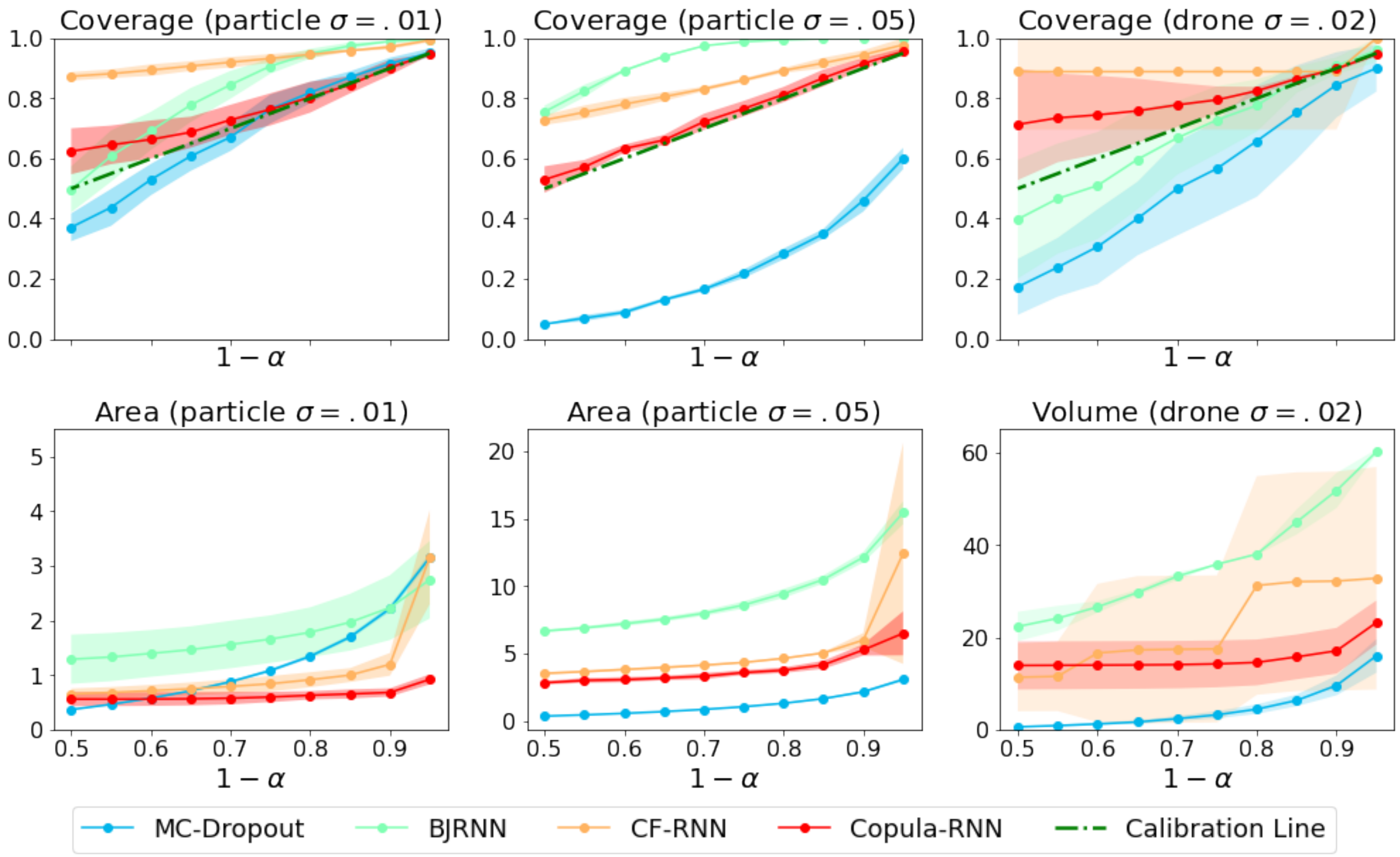
Copula Conformal Prediction

Results - examples



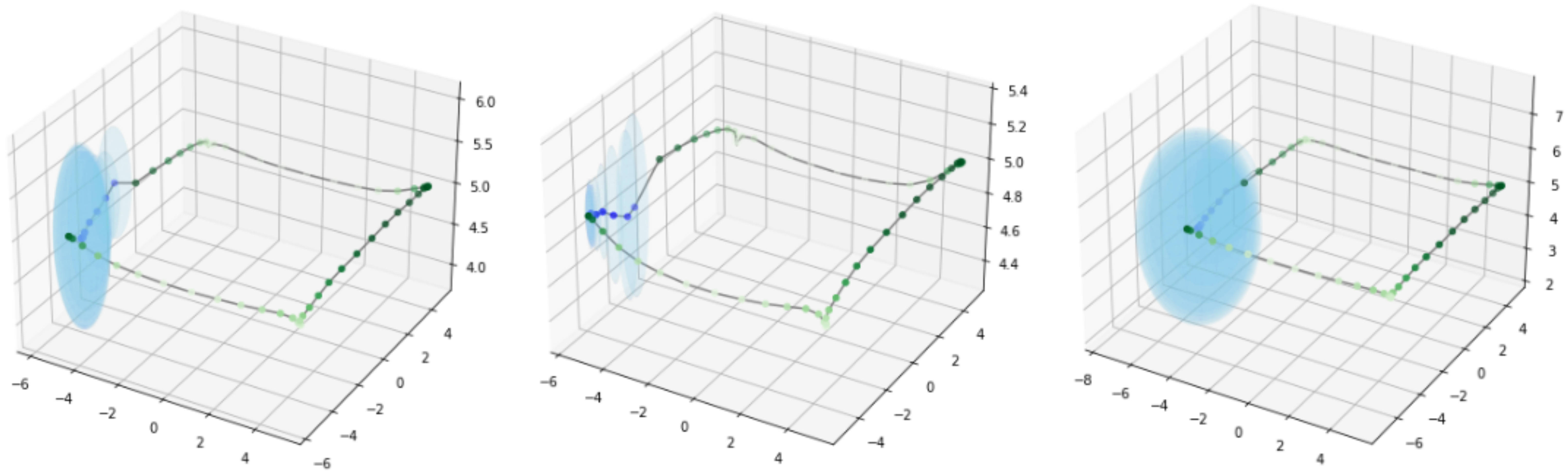
Covid-19 Cases in UK / Argoverse autonomous vehicles datasets. 90% confidence interval

Results: calibration and sharpness



Copula Conformal Prediction

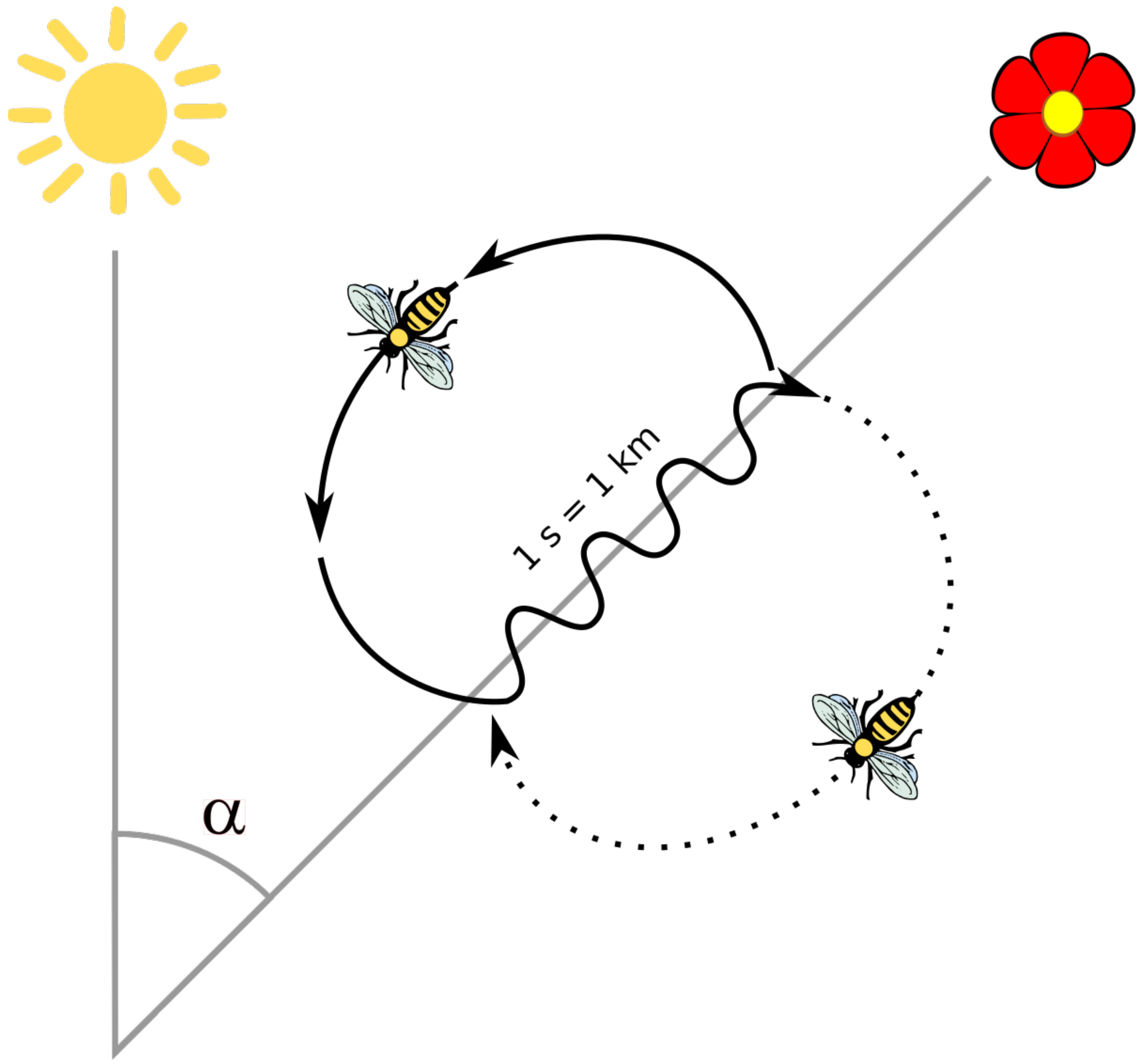
Results - examples



(a) Copula-EncDec (b) MC Dropout (c) CF-RNN

Figure 4.7. 99% Confidence region produced by three methods for the drone dataset. Copula methods (a) produces a more consistent, expanding cone of uncertainty compared to MC-Dropout (b) sharper one compared to CF-RNN (c).

Work 2: Adapting to Change Points [Neurips2025]



Conformal Prediction with Change Points

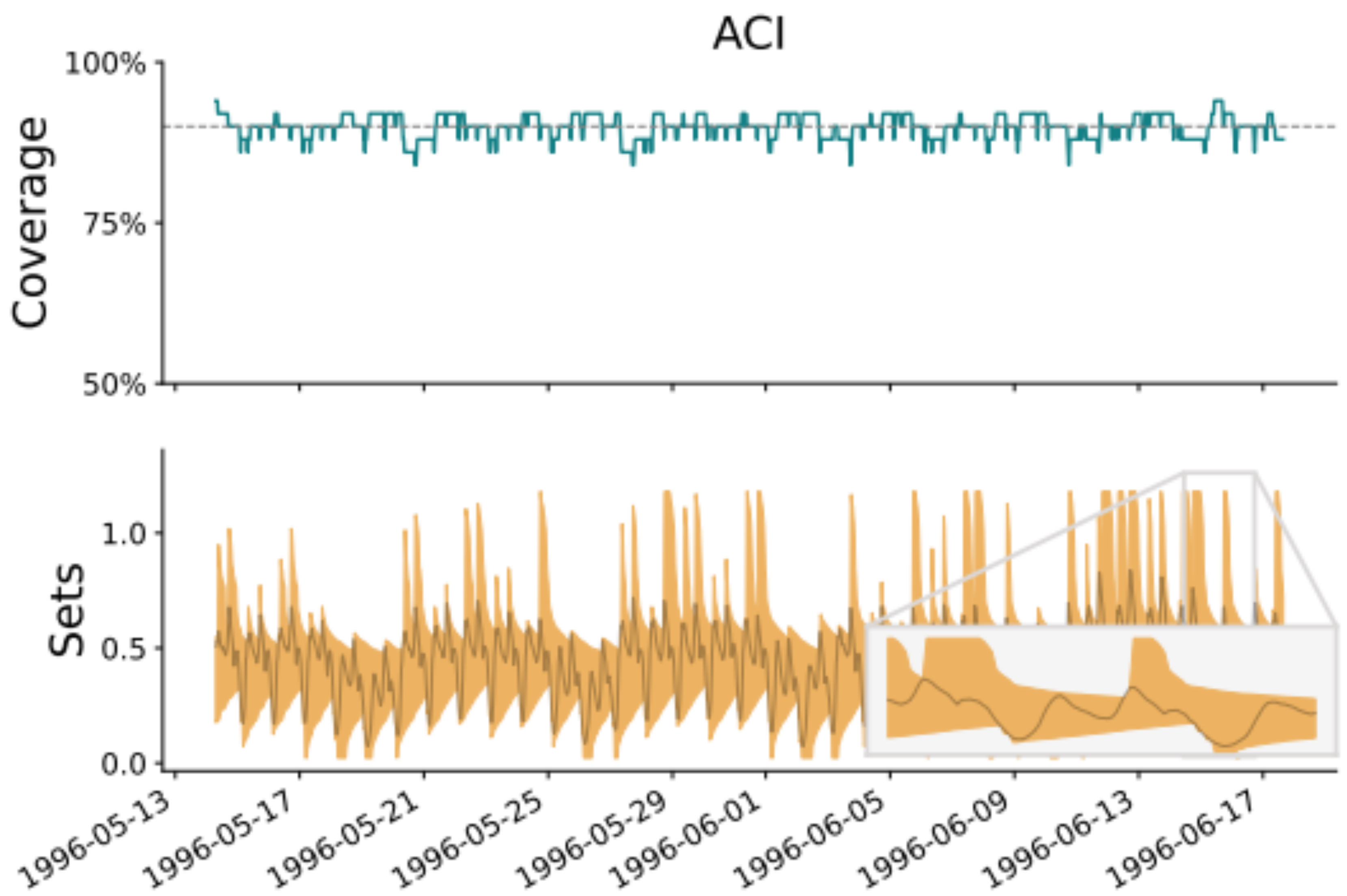
Setup / Baselines

- Observe a data stream $\{(x_t, y_t)\}_{t \in \mathbb{N}^+}$
- Perhaps $(x_t, y_t) \sim P_t$ with P_t varying across time
- At time t, want to use past data along with x_t to form prediction set Γ_t for y_t

Adaptive Conformal Inference

(Online estimation)

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \mathbf{1}[y_t \notin \Gamma_t])$$



Conformal Prediction with Change Points

Baselines / Context

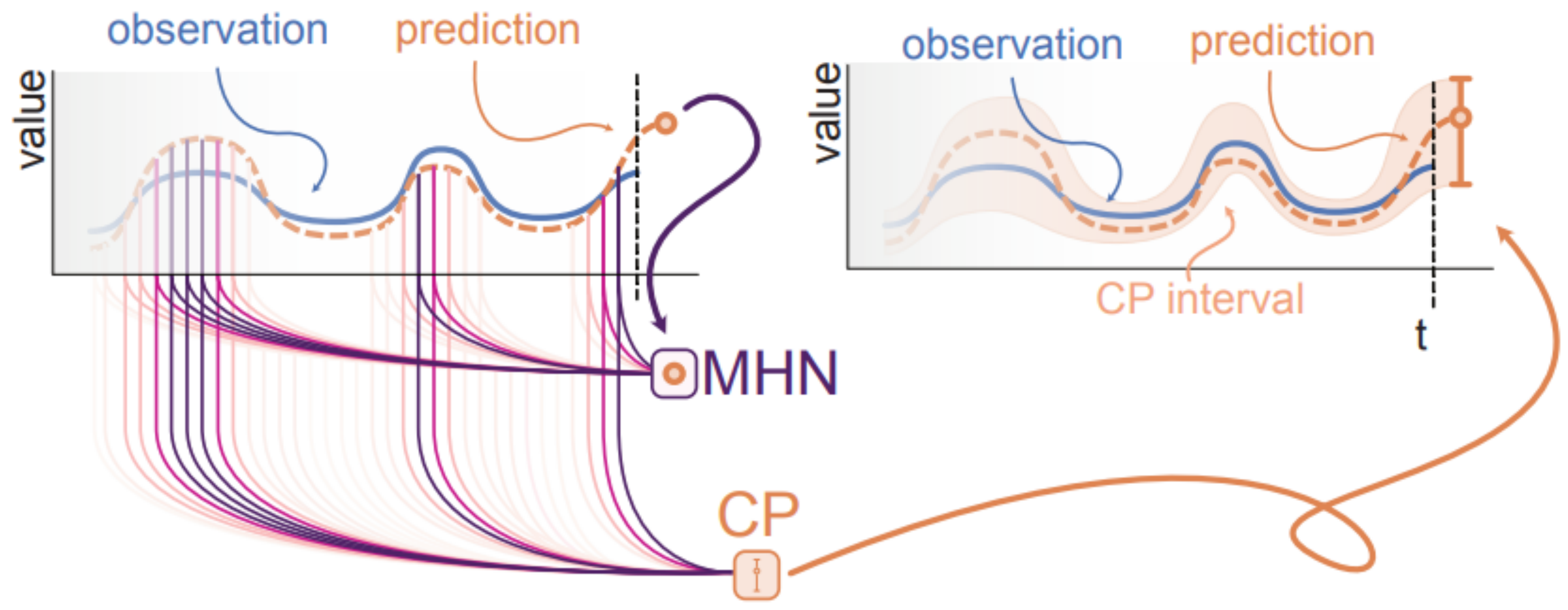
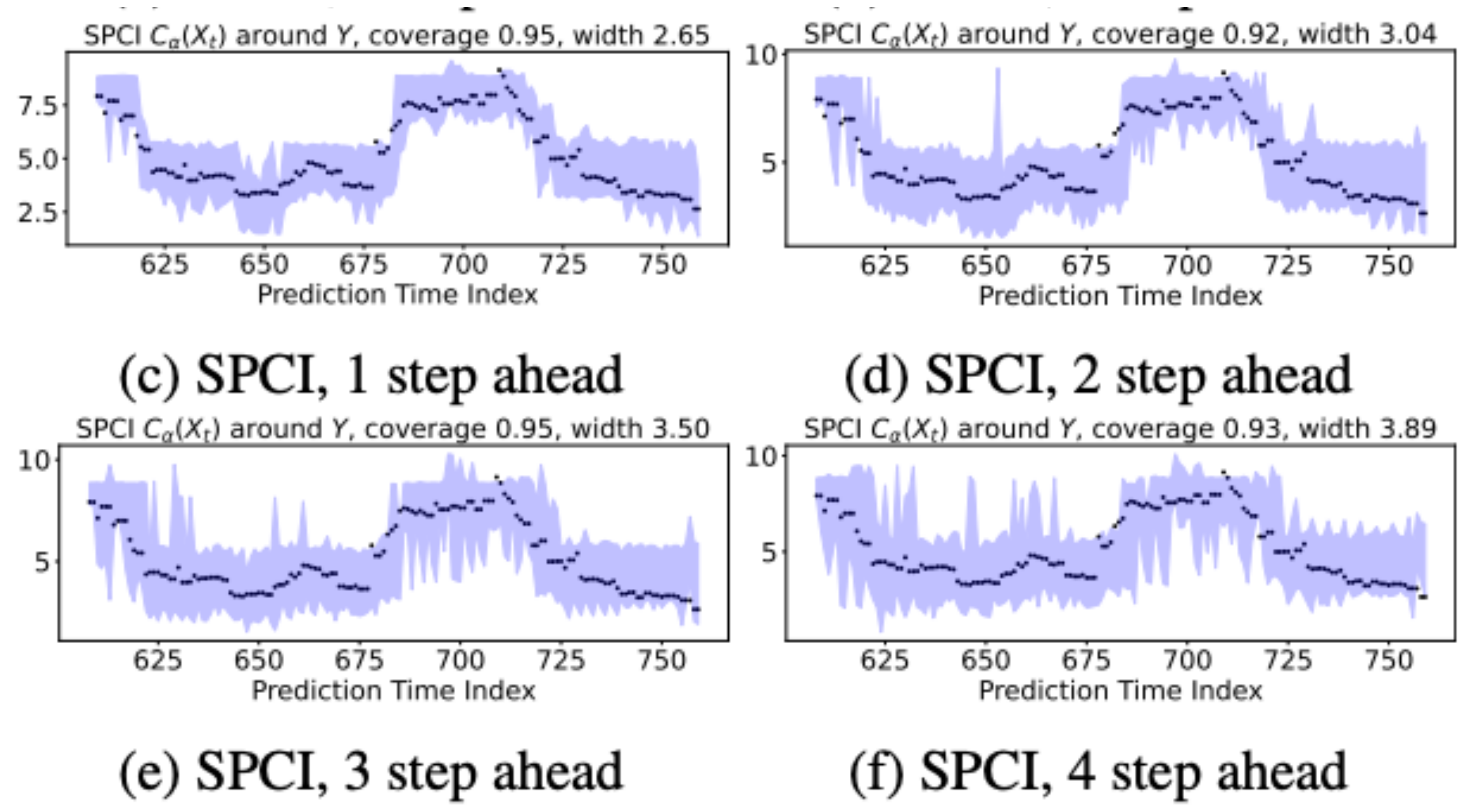


Figure 1: Schematic illustration of HopCPT. The Modern Hopfield Network (MHN) identifies regimes similar to the current one and up-weights them (colored lines). The weighted information enriches the conformal prediction (CP) procedure so that prediction intervals can be derived.

Uses a **Modern Hopfield Network** to learn temporal patterns

A. Auer, M. Gauch, D. Klotz, and S. Hochreiter. *Conformal prediction for time series with modern hopfield networks*. NeurIPS 2023.



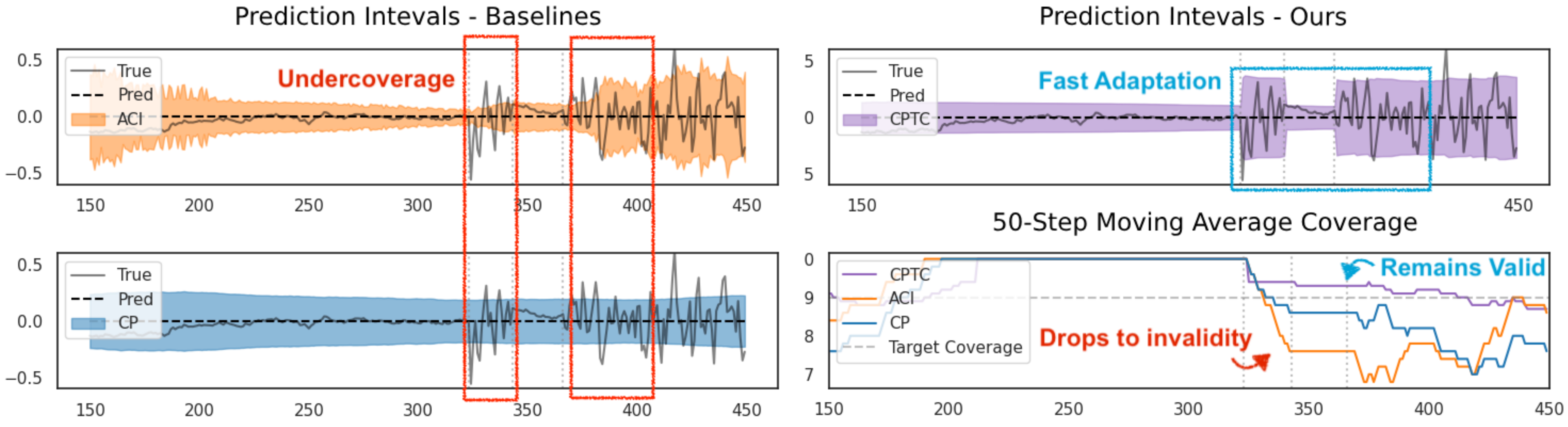
Uses **Quantile Random Forest** to learn temporal patterns

C. Xu and Y. Xie. Sequential predictive conformal inference for time series. ICML 2023

Conformal Prediction with Change Points

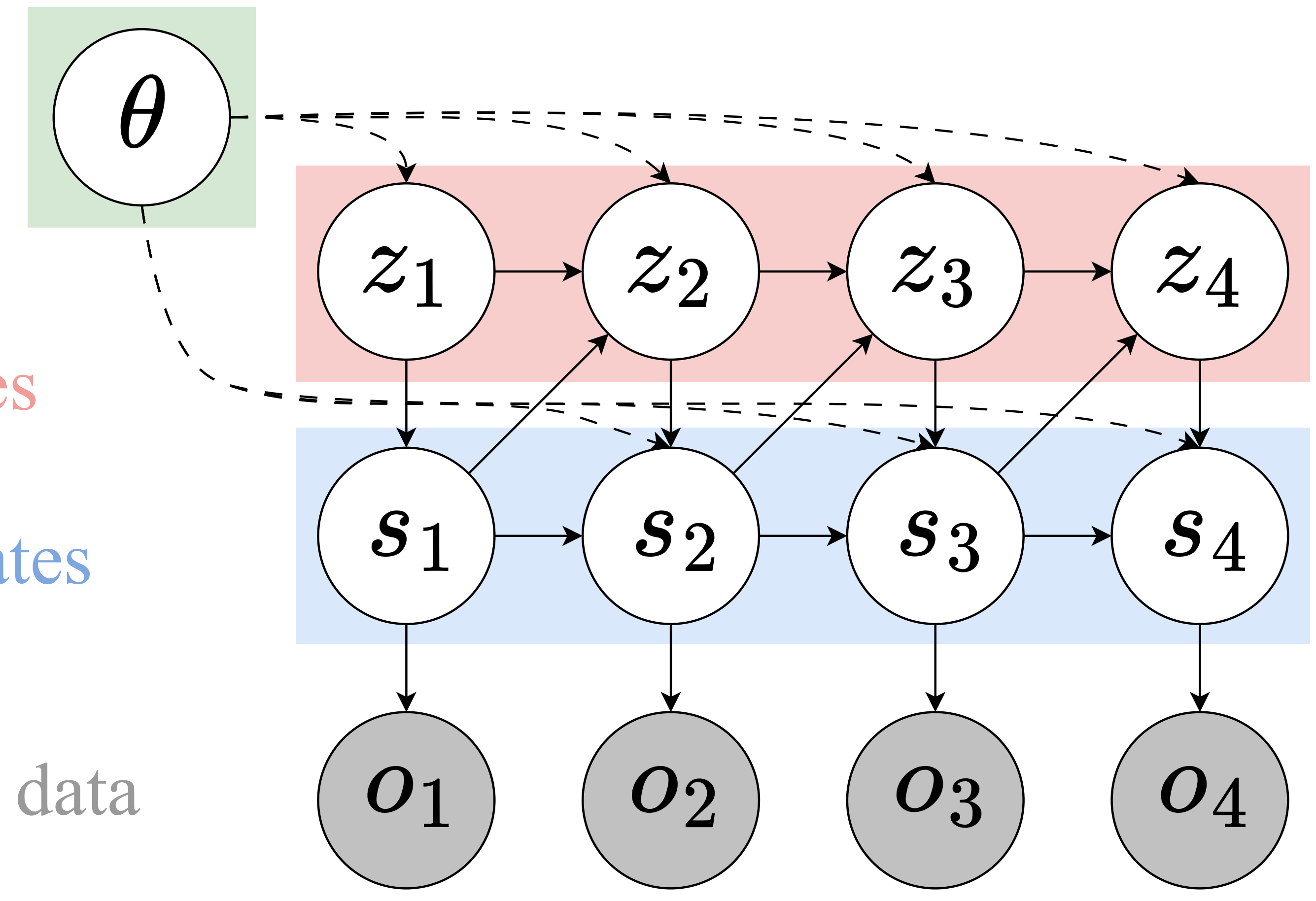
Motivation

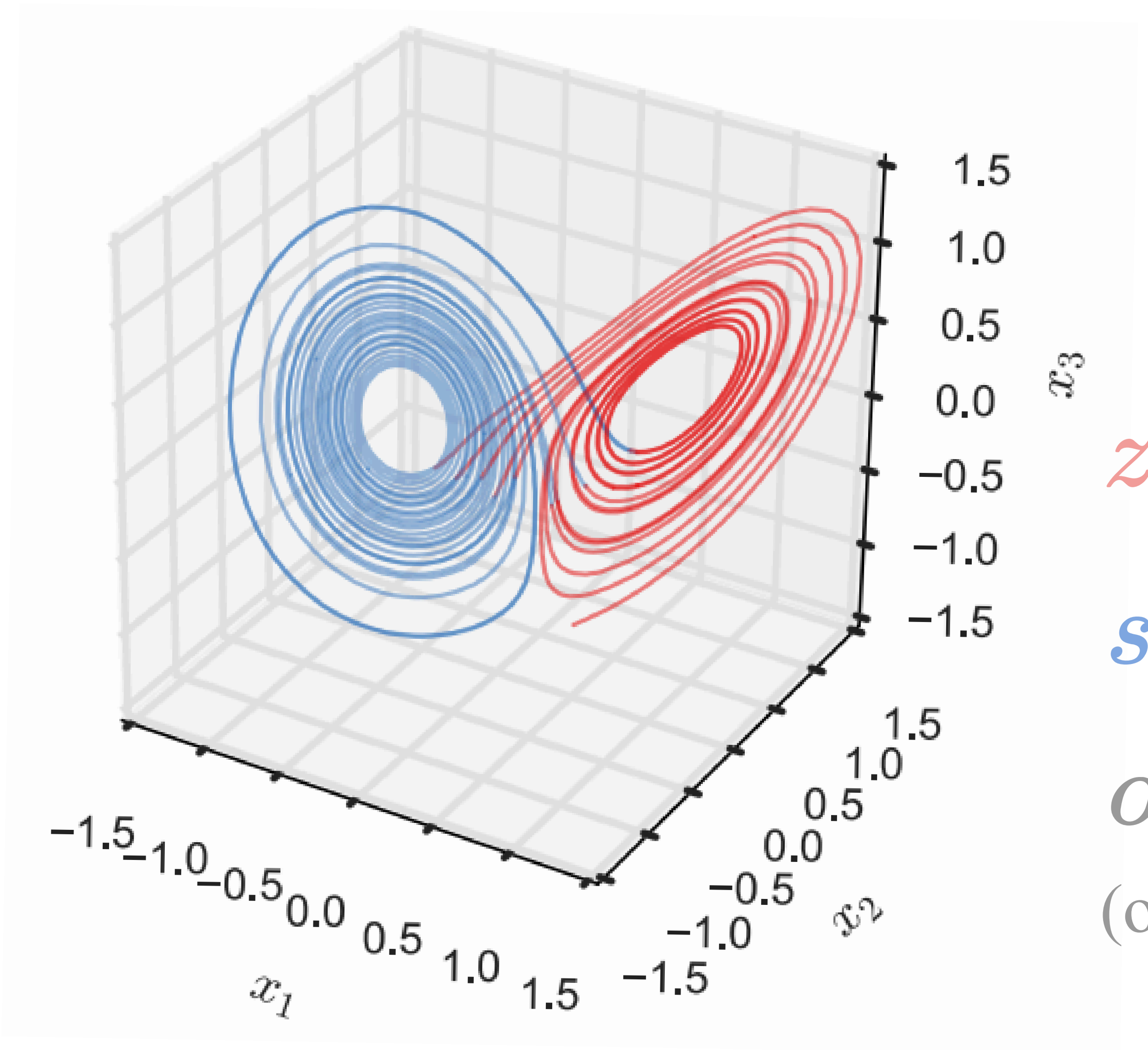
- Baseline *react*, or use *regression* at test time to learn correlations.
- What happens when we can *anticipate* distribution shifts?



Conformal Prediction with Change Points

Switching Dynamical Systems





$z_t \in \{L, R\}$

$s_t \in \mathbb{R}^3$

$o_t \in \mathbb{R}^d$

(observation dimension)

Gives us $P(y_t \mid x_{0:t}) = \sum_{z \in \mathcal{Z}} P(y_t \mid x_{0:t}, z_t = z) P(z_t = z \mid x_{0:t})$

Conformal Prediction with Change Points

Switching Dynamical Systems

Decompose coverage goal into

$$\sum_{z \in \mathcal{Z}} P(y_t \in \Gamma_{z,t} \mid x_{0:t}, z_t = z) \cdot P(z_t = z \mid x_{0:t}) \geq 1 - \alpha$$

Conformal Prediction with Change Points

Switching Dynamical Systems

Decompose coverage goal into

$$\sum_{z \in \mathcal{Z}} \boxed{P(y_t \in \Gamma_{z,t} \mid x_{0:t}, z_t = z)} \cdot P(z_t = z \mid x_{0:t}) \geq 1 - \alpha$$

We can track **state-specific uncertainty**

Conformal Prediction with Change Points

Switching Dynamical Systems

Decompose coverage goal into

$$\sum_{z \in \mathcal{Z}} \boxed{P(y_t \in \Gamma_{z,t} \mid x_{0:t}, z_t = z)} \cdot \boxed{P(z_t = z \mid x_{0:t})} \geq 1 - \alpha$$

We can track **state-specific uncertainty**

then **combine** them based on state probability

Conformal Prediction with Change Points

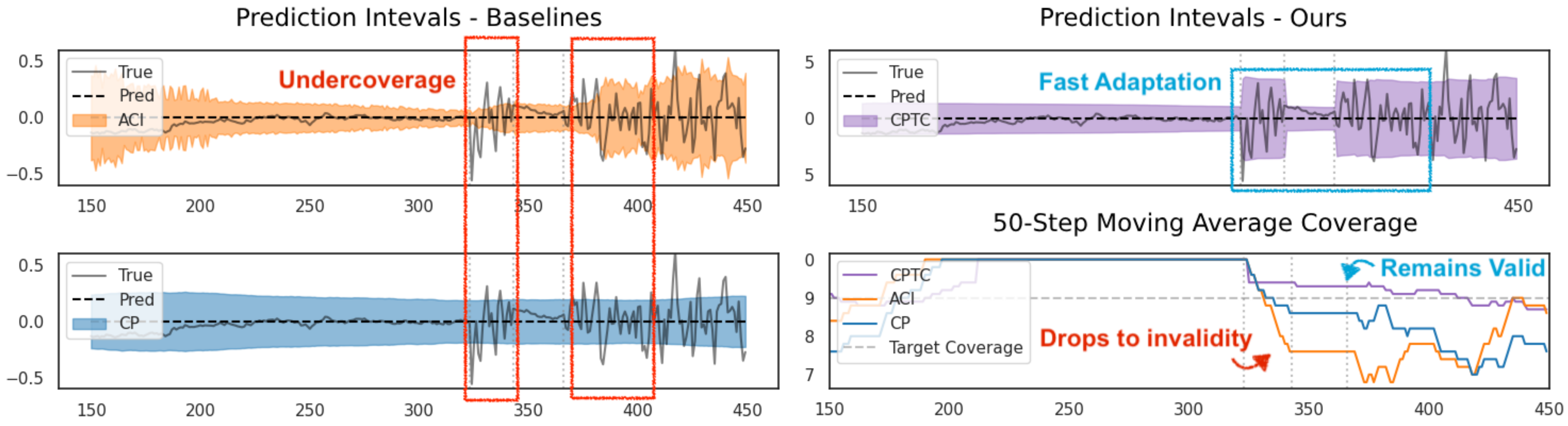
Switching Dynamical Systems

Decompose coverage goal into

$$\sum_{z \in \mathcal{Z}} \boxed{P(y_t \in \Gamma_{z,t} \mid x_{0:t}, z_t = z)} \cdot \boxed{P(z_t = z \mid x_{0:t})} \geq 1 - \alpha$$

Slow updates

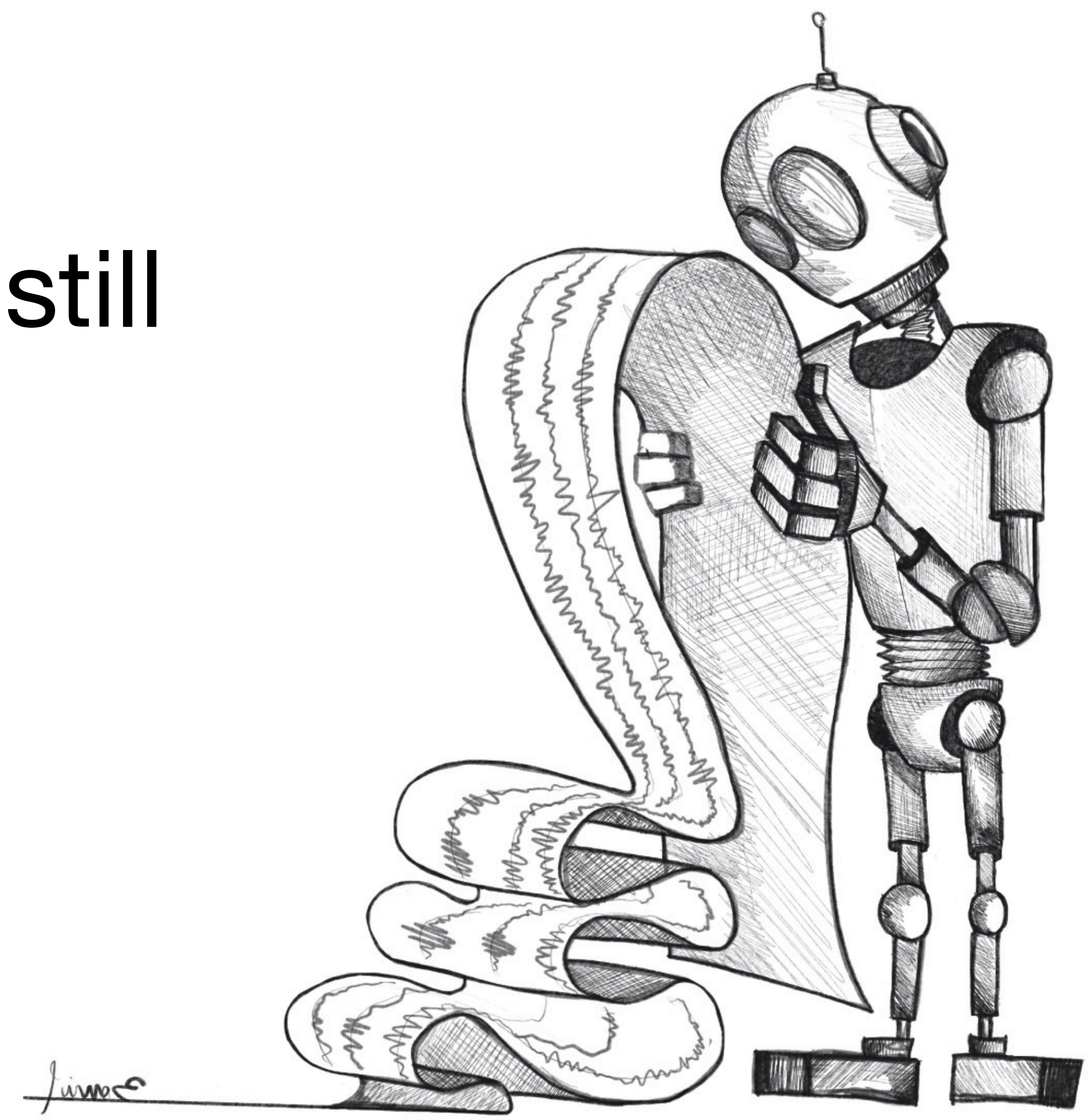
Fast updates



Conformal Prediction with Change Points

Summary of Theoretical Results

- We have finite sample validity guarantee if If noise is stationary.
- Without any assumption, we achieve a finite-sample miscoverage bound (Decays at $\mathcal{O}(1/T)$)
- Robust to state prediction errors. Finite sample bound still holds!
- Faster adaptation if state prediction is correct.



Conformal Prediction with Change Points

Results

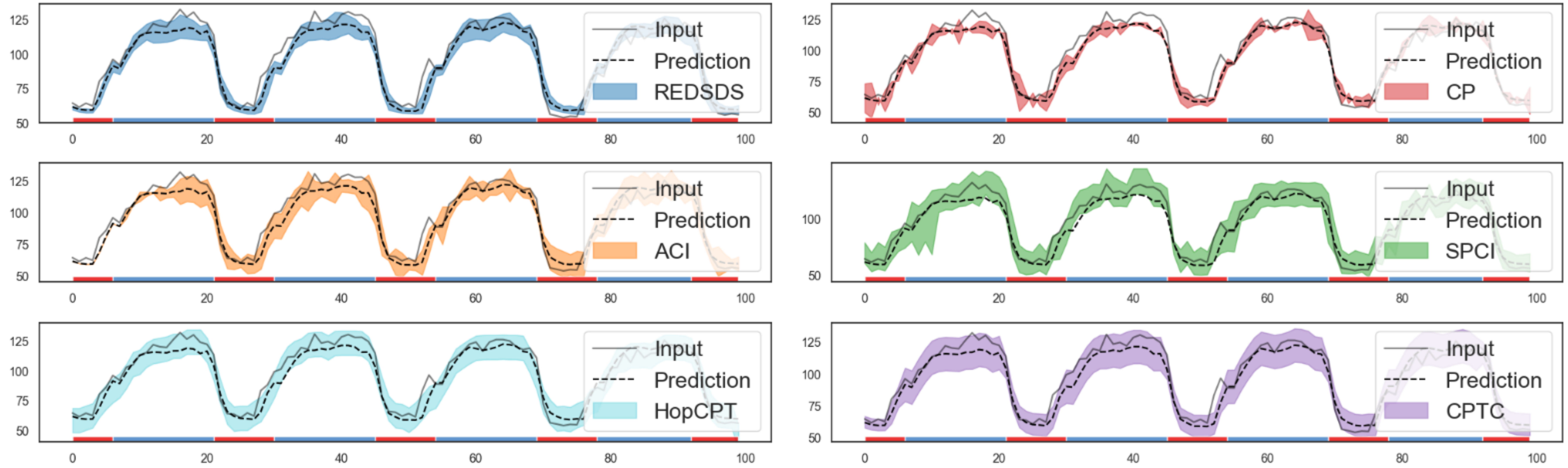
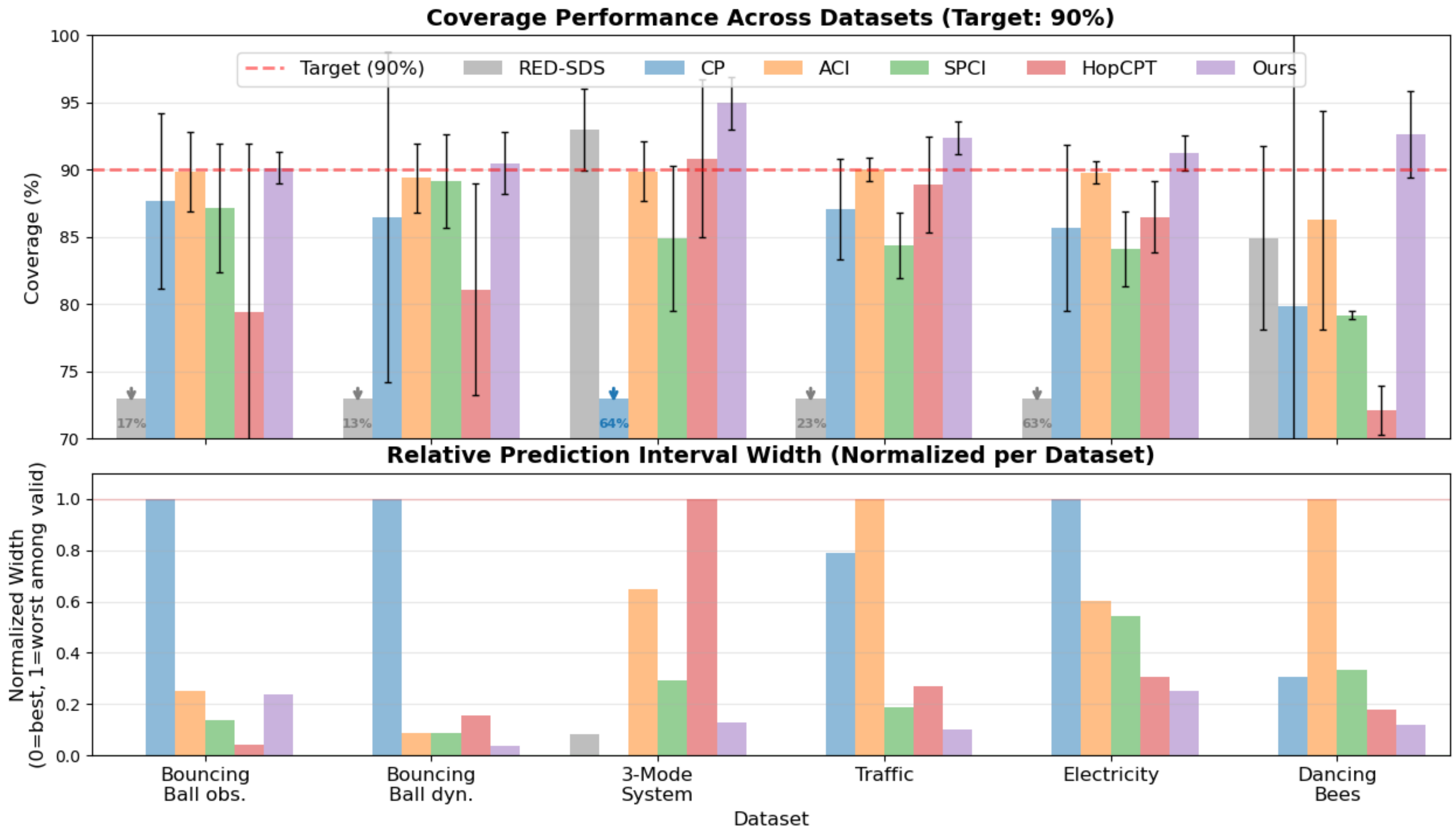


Figure 5.3. Visualization of prediction intervals on the Electricity hourly demand dataset. The red and blue bars in the bottom reflects the underlying switching state of day and night. Our method (purple) adapts to different levels of volatility between day and night, and achieves stabler coverage over time, whereas ACI (yellow) over-covers during the night and under-covers at change points.

Conformal Prediction with Change Points

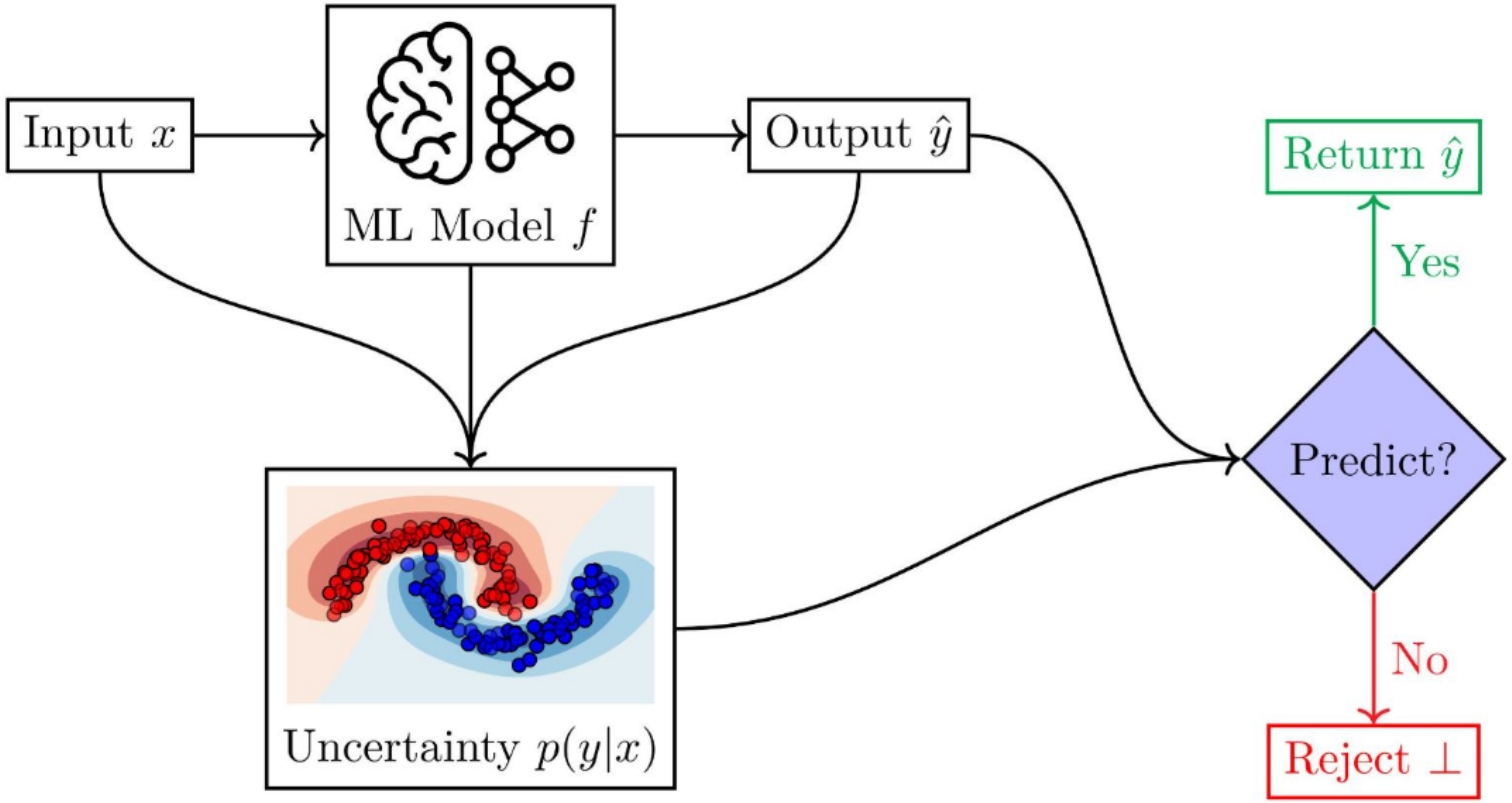
Results



Talk Outline

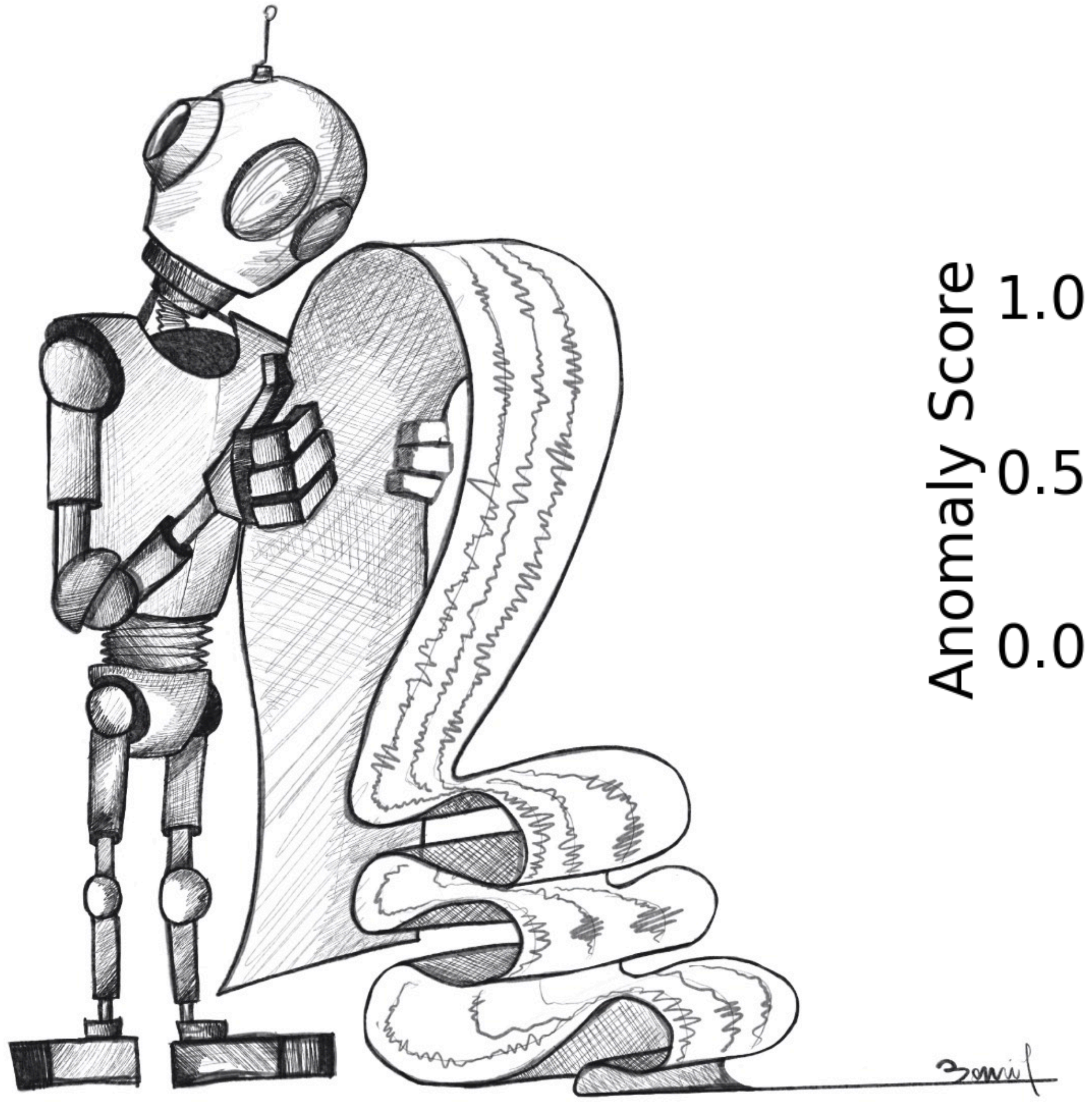
- Part I: Probabilistic Modeling and Uncertainty Quantification
  - Leveraging structure in model design
  - Leveraging structure in post-hoc calibration
- Part II: Decision making Under Uncertainty
  - **Selective prediction**
    - Example: no-mistake anomaly detection
  - Safety constraints and pessimistic planning
    - Example: Robot navigation
- Discussion and Conclusion

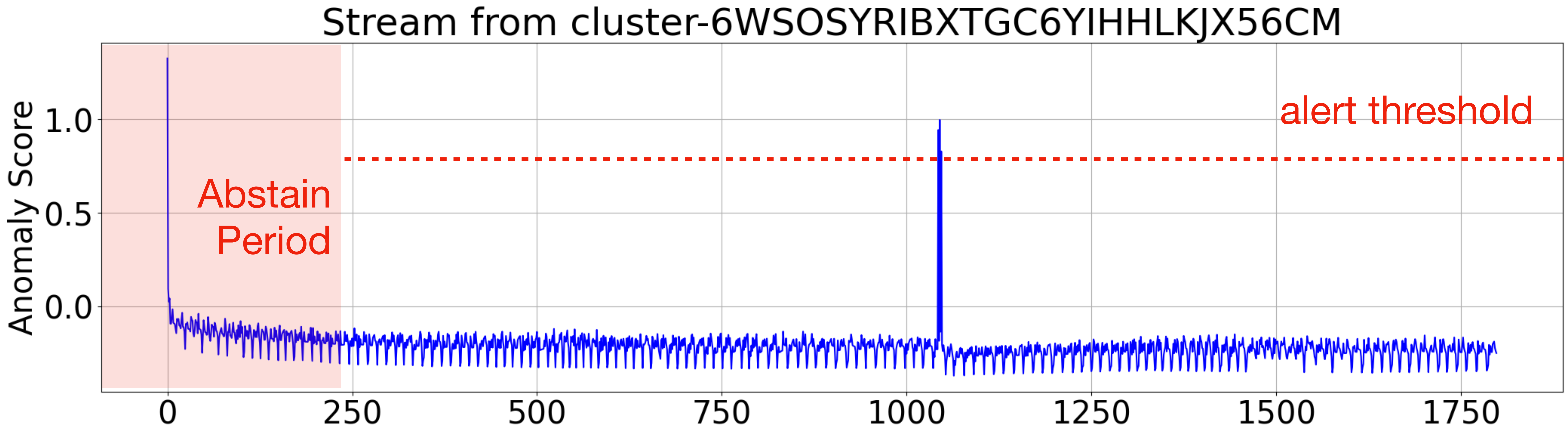
Selective Prediction



In the context of Anomaly Detection…

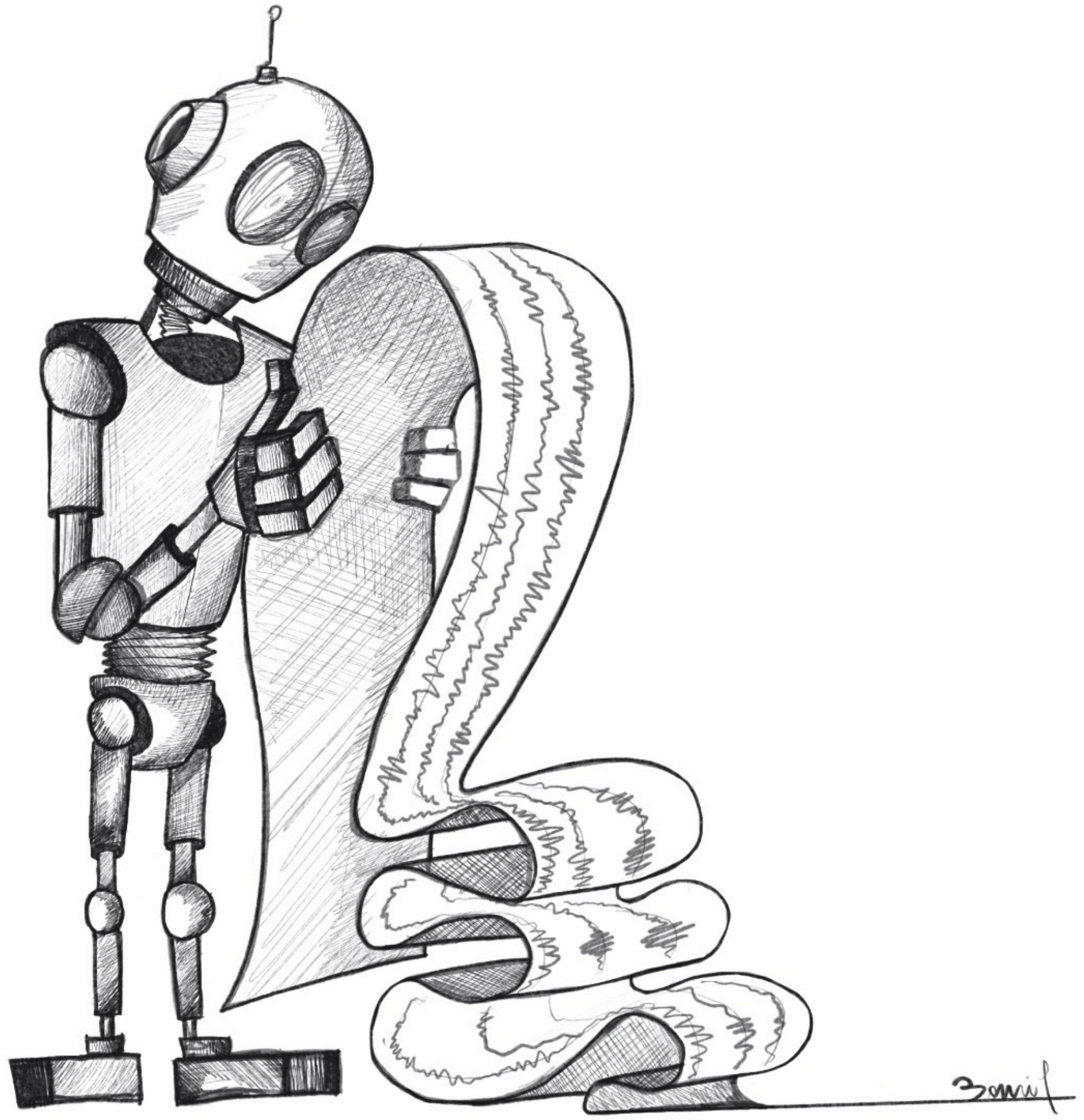
- a probabilistic model outputs an "anomaly score"
- The threshold is usually determined by heuristics and hard to calibrate

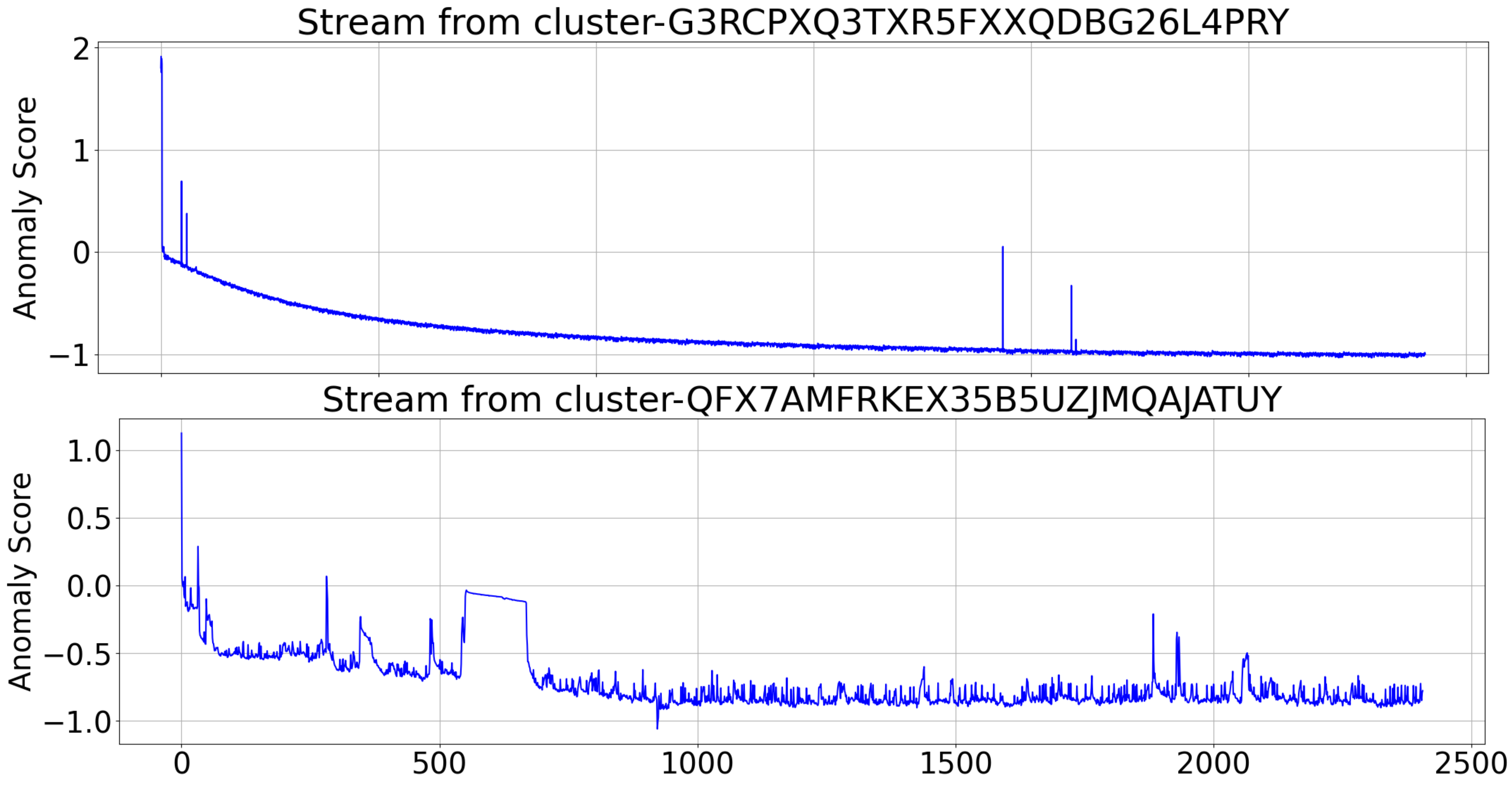




Problem Formulation

- We want: Low abstains, low false positive, high power
- identify shifts and drifts, and adapt the threshold to data?

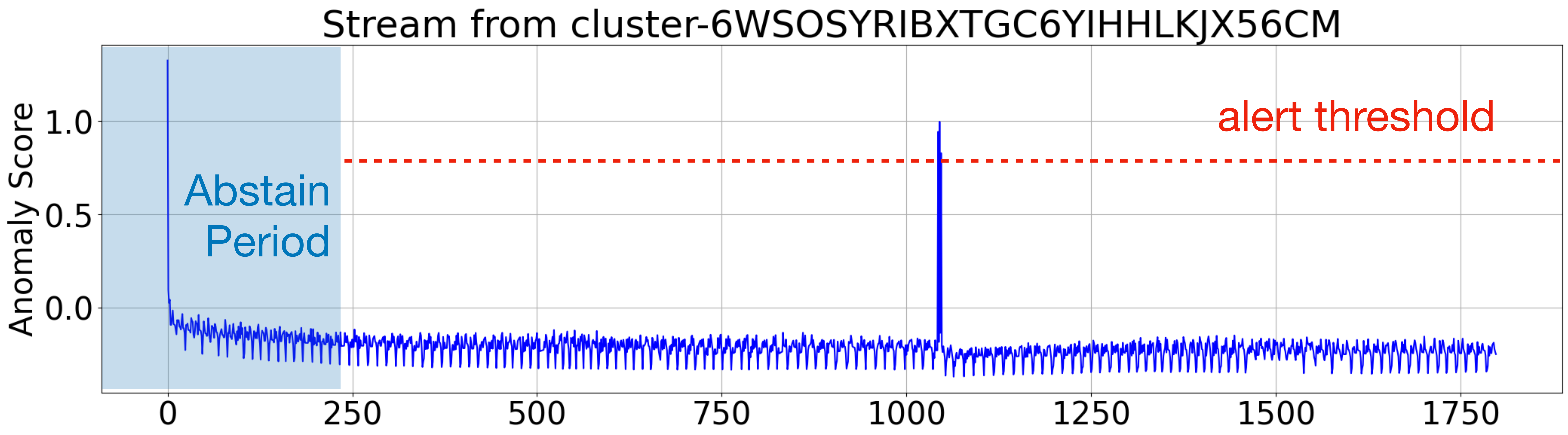




Problem Formulation

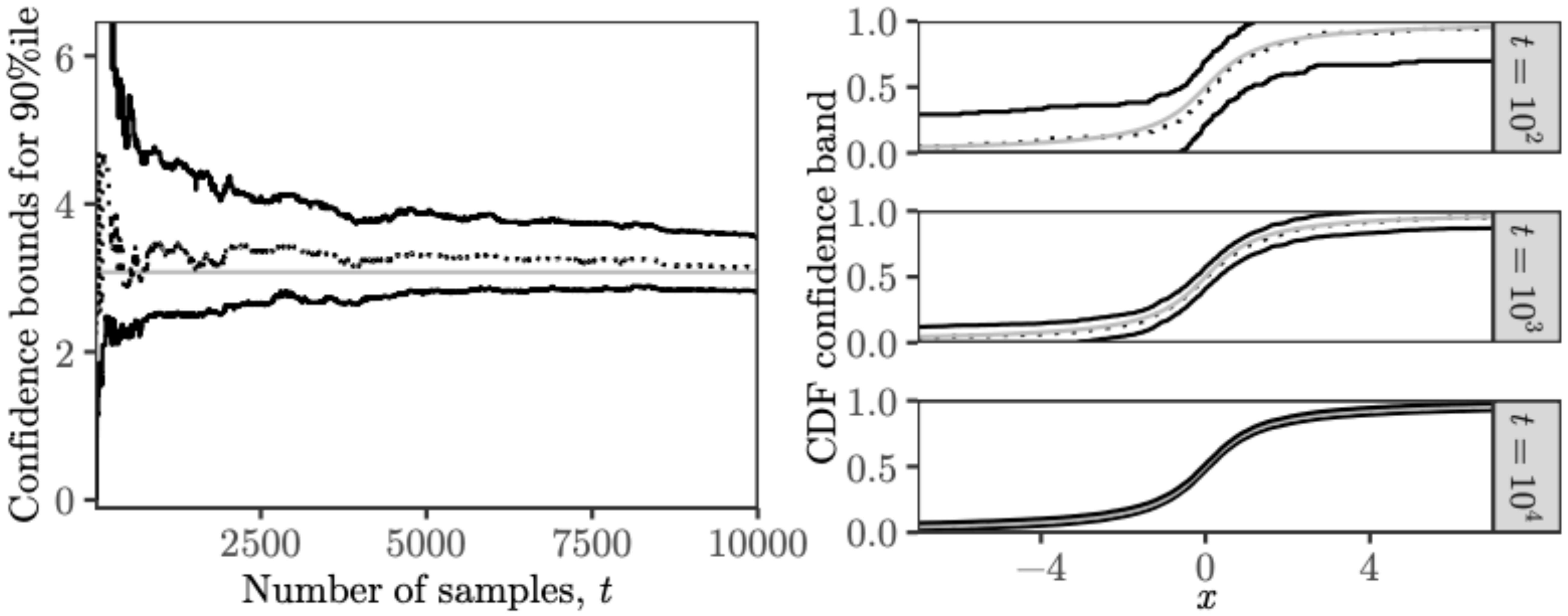
- At every time t, outputs $\hat{y}_t = \mathscr{A}(x_1, \ldots, x_t)$. $\hat{y}_t \in \{0,1,*\,(\text{abstain})\}$
- Minimize regret:

$$\text{Reg}_T\left(\mathscr{A}; (x_t^{(i)})_{t=1}^T\right) = \overbrace{c_1 \cdot \sum_{t=1}^{T} \mathbf{1}(\hat{y}_t = *)}^{\#\text{ Abstains}} + \overbrace{c_2 \cdot \sum_{t=1}^{T} \mathbf{1}(\hat{y}_t \neq y_t, \hat{y}_t \neq *)}^{\#\text{ Mistakes}}$$



Confidence Sequences

Confidence sequences are time-indexed confidence intervals for estimating statistics of i.i.d samples.



$$\mathbb{P}\left(\forall t, p \in (0,1) : \hat{Q}_t(p - u_t) \le Q(p) \le \hat{Q}_t(p + u_t)\right) \ge 1 - \alpha$$

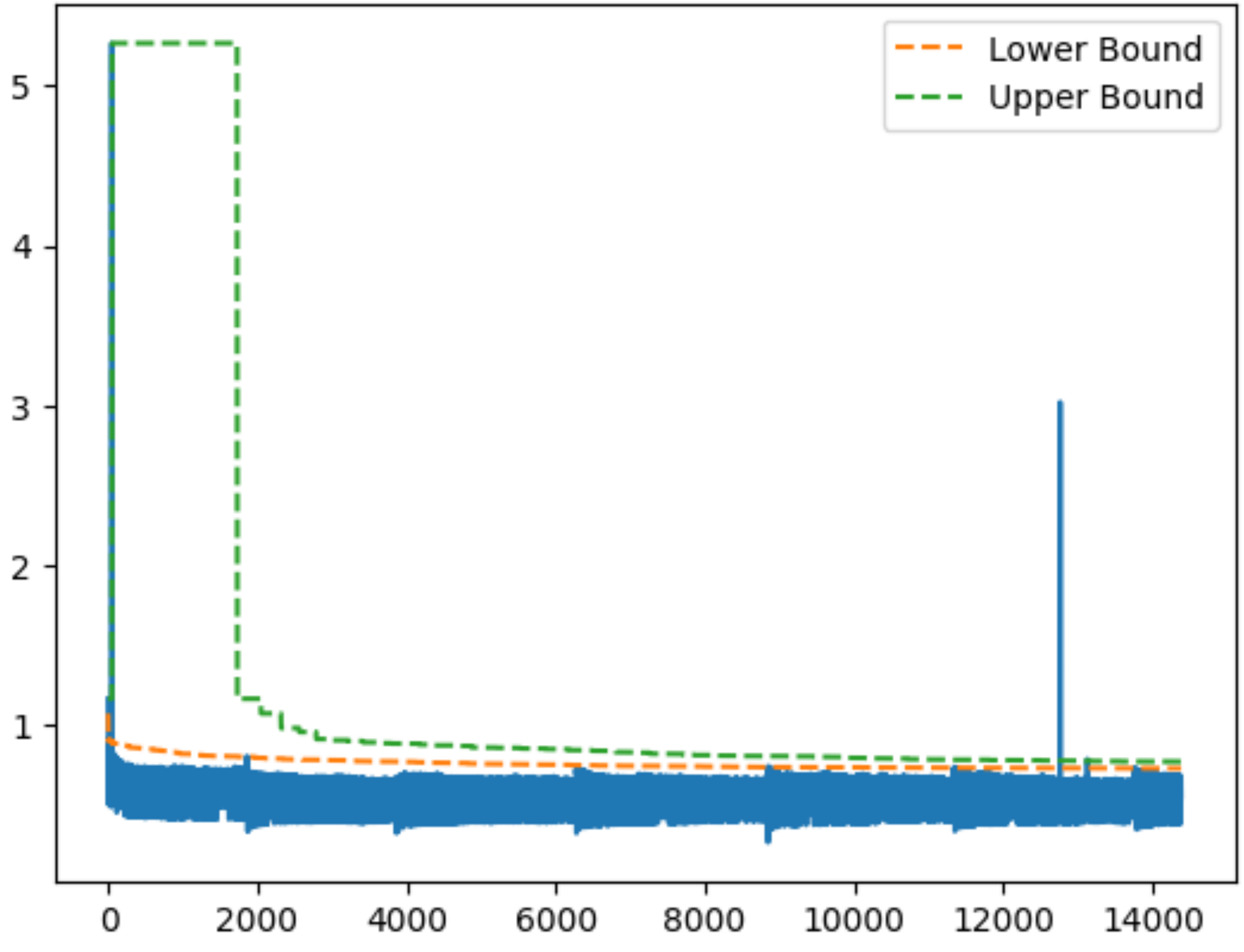
$$u_0 = 1,\ u_t = 0.85\sqrt{t^{-1}[\log\log(et) + 0.8\log(1612/\alpha)]}$$

[1] Sequential estimation of quantiles with applications to A/B testing and best-arm identification. Howard & Ramdas, 2022

Easy Algorithm for i.i.d. Case

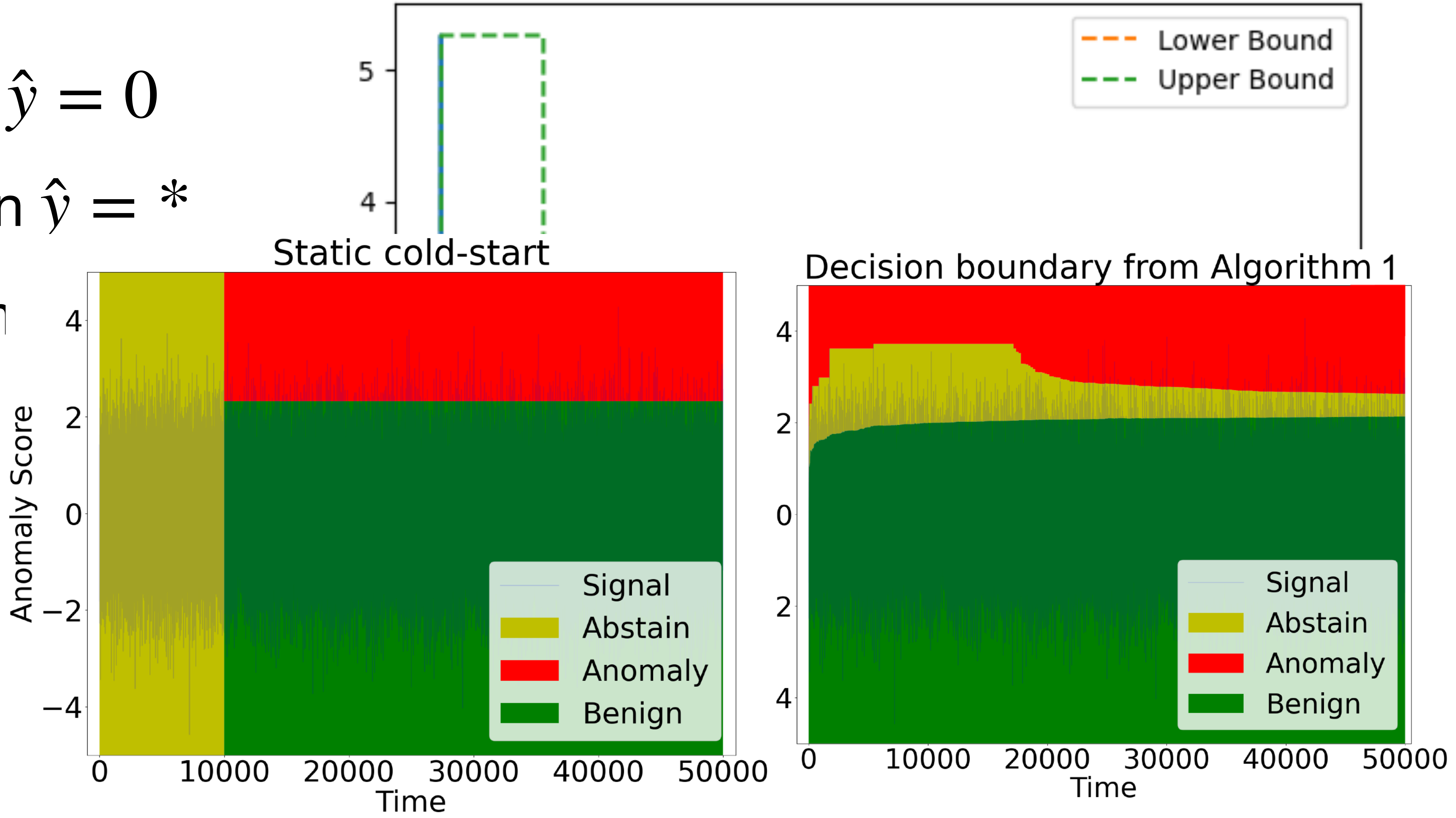
- $x_t < lb$: Benign data $\hat{y} = 0$
- $lb \leq x_t \leq ub$: abstain $\hat{y} = *$
- $x_t > ub$: Report anomaly $\hat{y} = 1$

- $\mathcal{O}(\sqrt{T})$ abstains
- 0 mistakes w.h.p



Easy Algorithm for i.i.d. Case

- $x_t < lb$: Benign data $\hat{y} = 0$
- $lb \leq x_t \leq ub$: abstain $\hat{y} = *$
- $x_t > ub$: Report anon

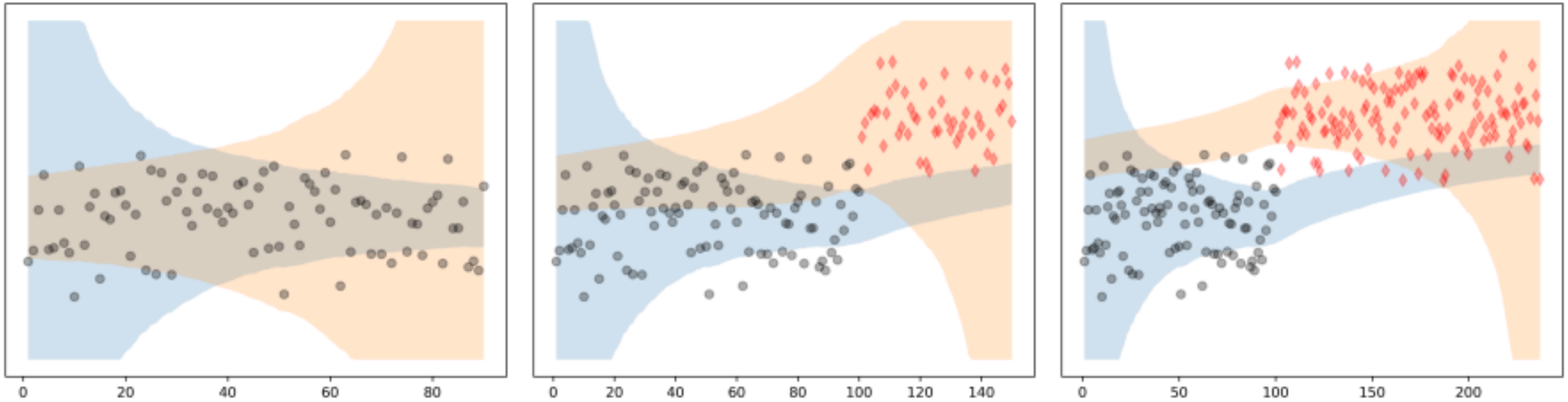


- $\mathcal{O}(\sqrt{T})$ abstains
- 0 mistakes w.h.p.

Confidence Sequences can detect shifts!

For any statistics θ (e.g. 99% quantile):

If we can construct a CS for θ $\Rightarrow$ we can detect changes in θ $\Rightarrow$ when forward and backward CS disagrees.



$$D(\Delta_k, \alpha) = \mathbb{E}[(\tau - \tau_c)] = \mathcal{O}(\frac{\log\log(1-\Delta)}{\Delta^2}) \;\; w.h.p. \text{ where } \Delta = d(\theta_1, \theta_2)$$

Detected change time   True change time

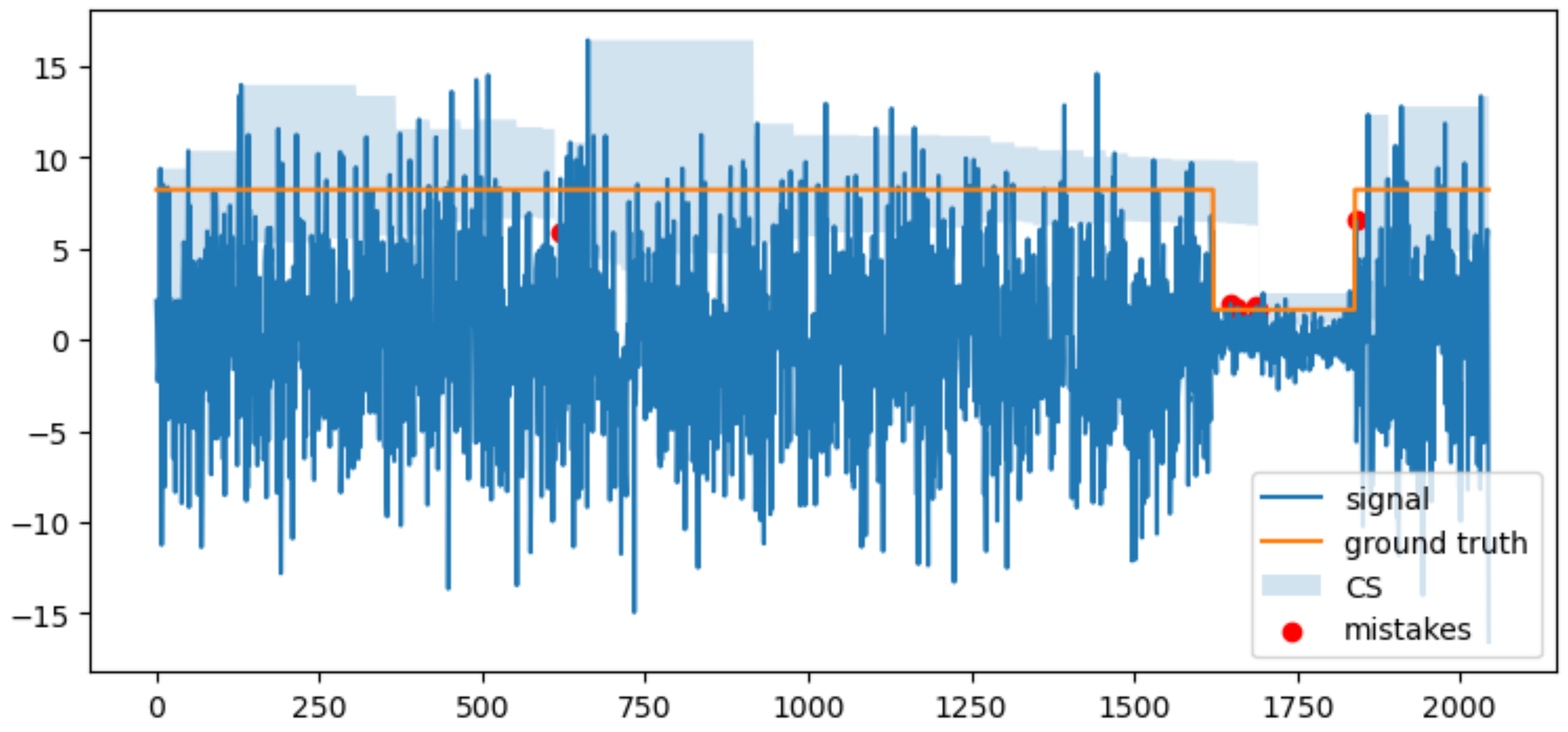
Piecewise Stationary Stream

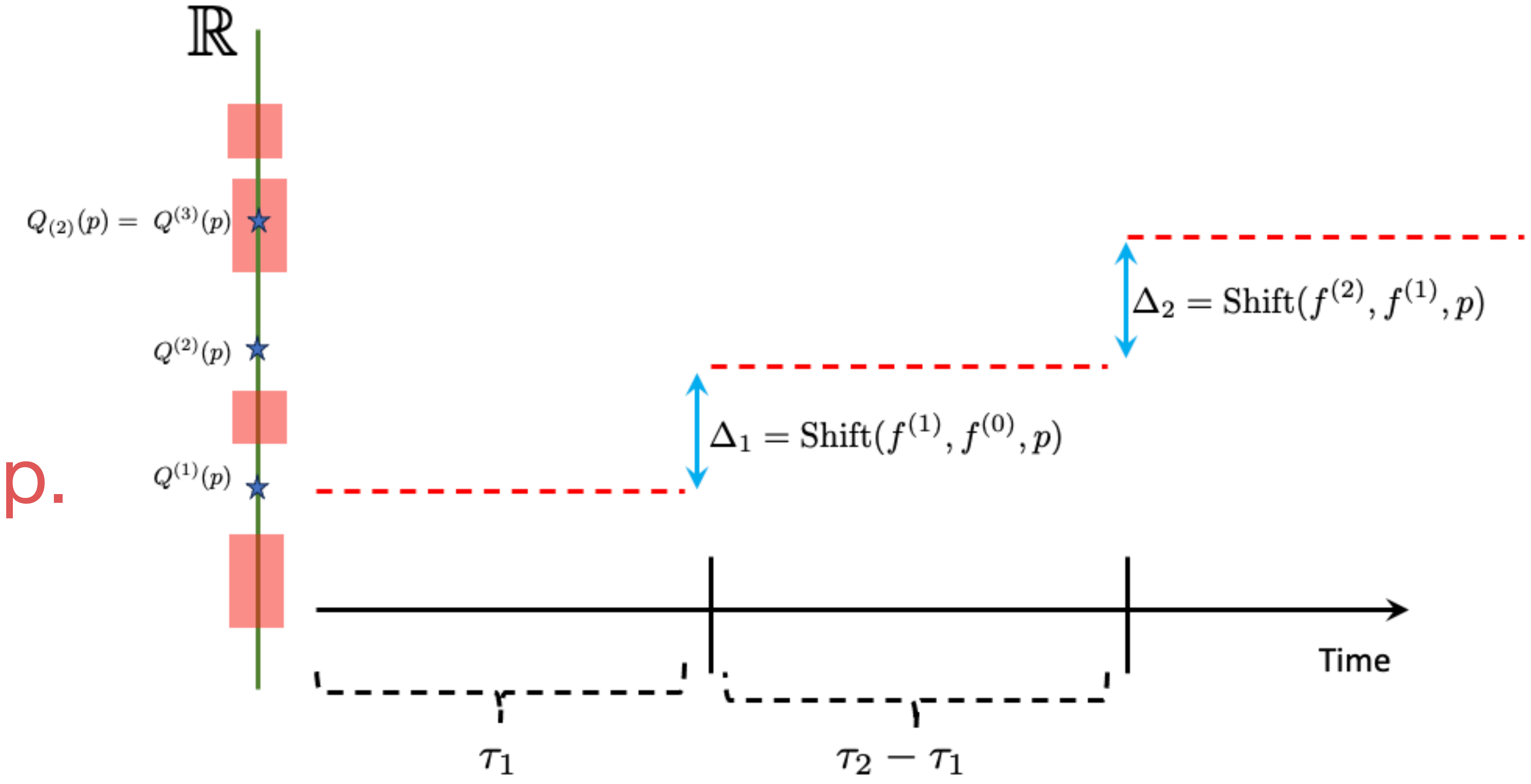
H_T = number of changes until T,

$D(\Delta_k, \alpha)$ = Detection delay $= \tilde{\mathcal{O}}(\frac{1}{(\Delta_k)^2})$

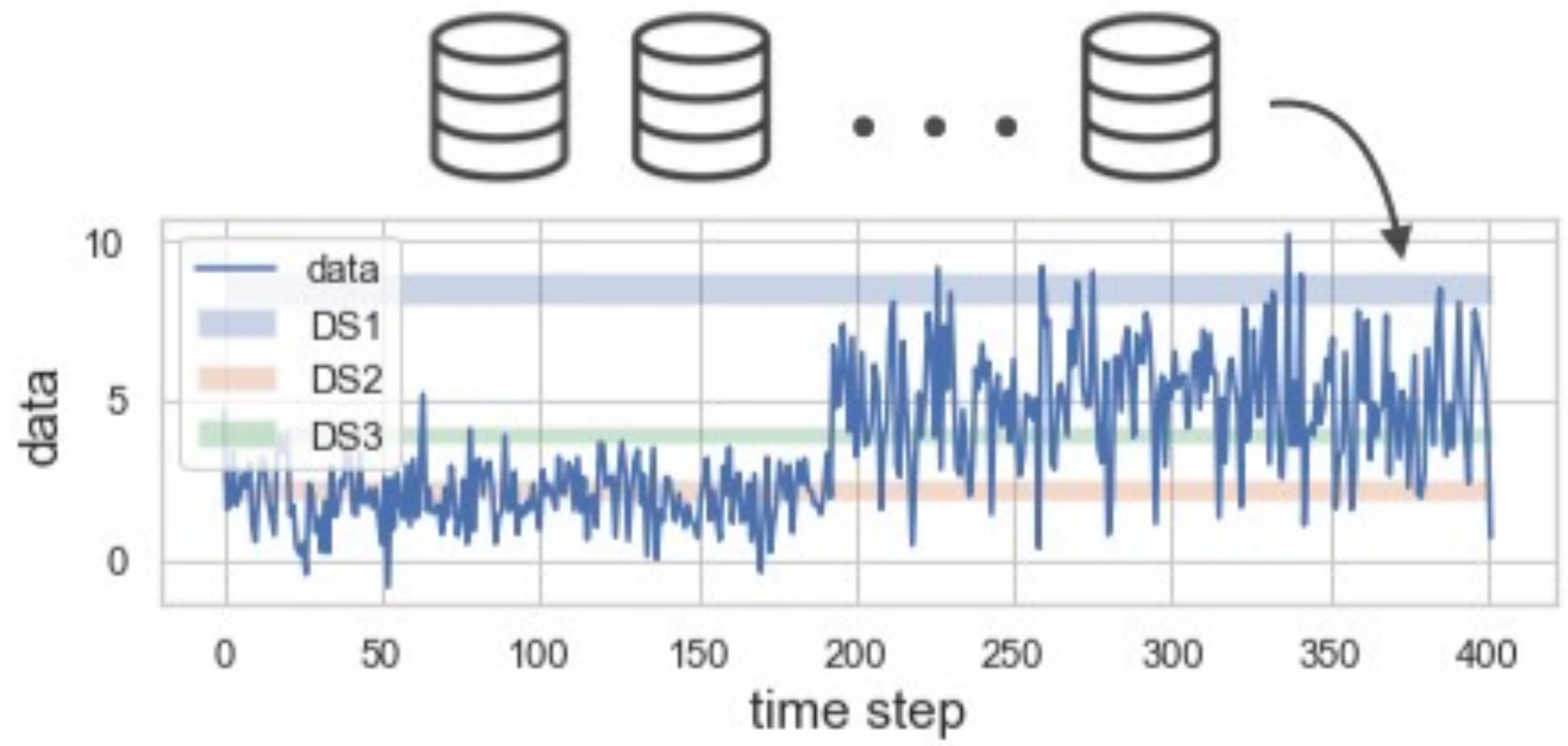
abstains $\leq \mathcal{O}(\sqrt{T}) + \sum_{k=1}^{H_T} D(\Delta_k, \alpha)$

Mistakes (FP + FN) $\leq \sum_{k=1}^{H_T} D(\Delta_k, \alpha)$ w.h.p.



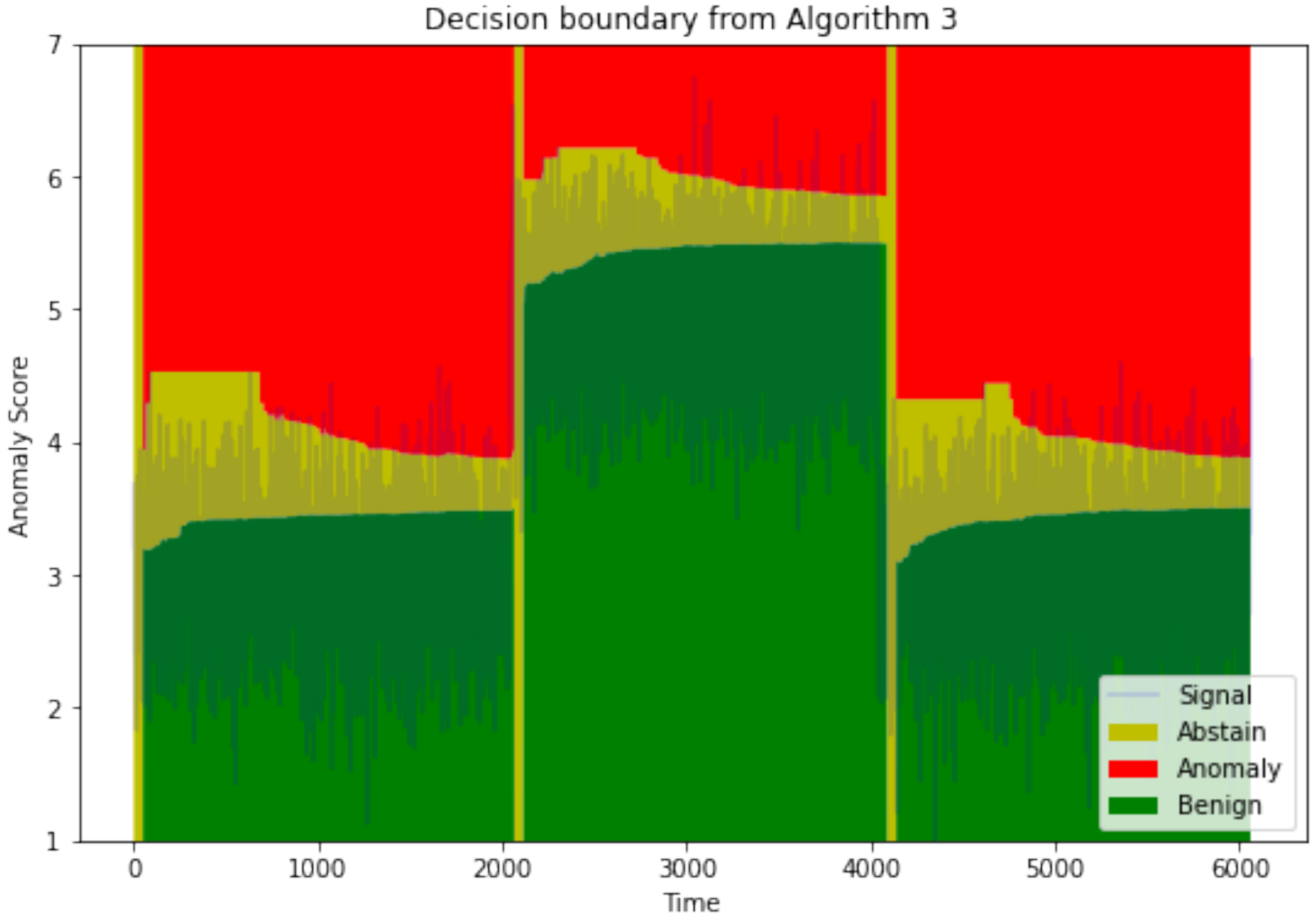


Streams with offline data



abstains $\leq \mathcal{O}(\sqrt{N+T} - \sqrt{N})$ + detection delays (N = size of offline data)

Bounded degradation if offline dataset is arbitrary.

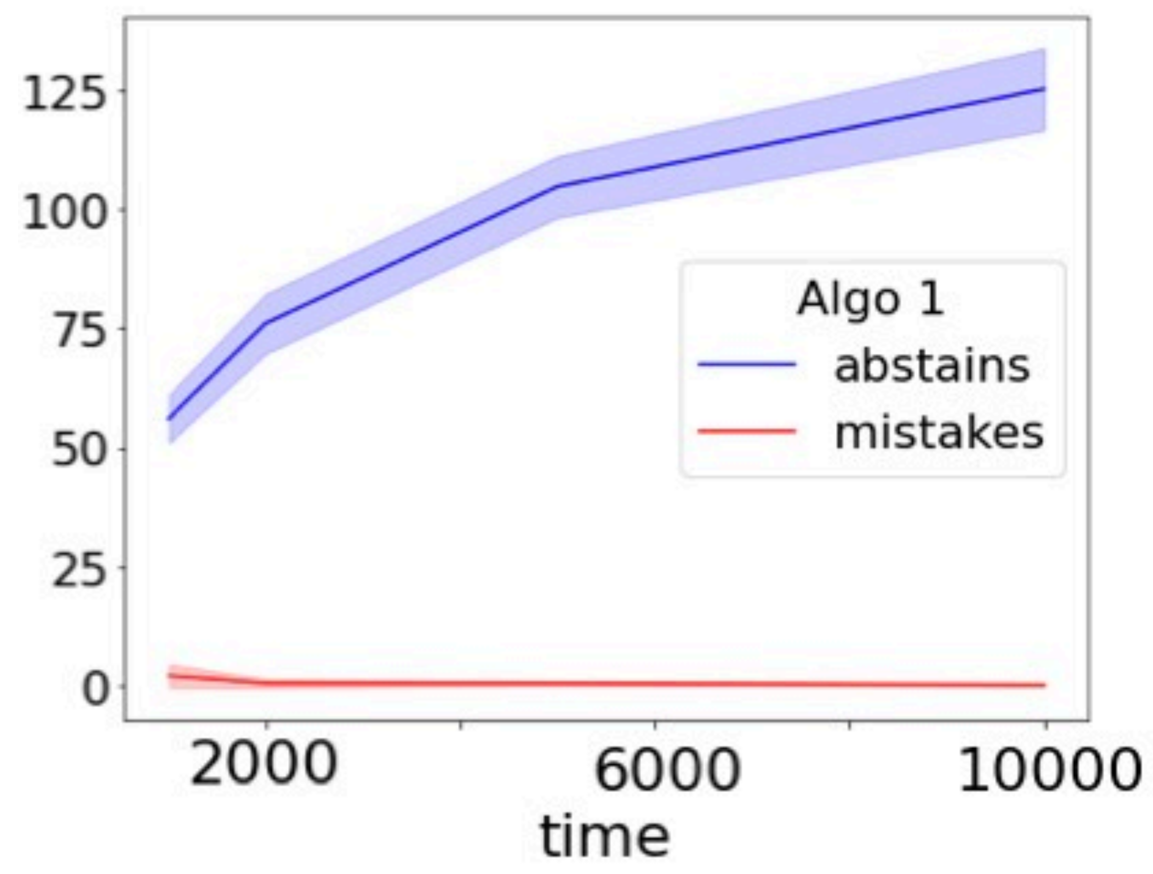


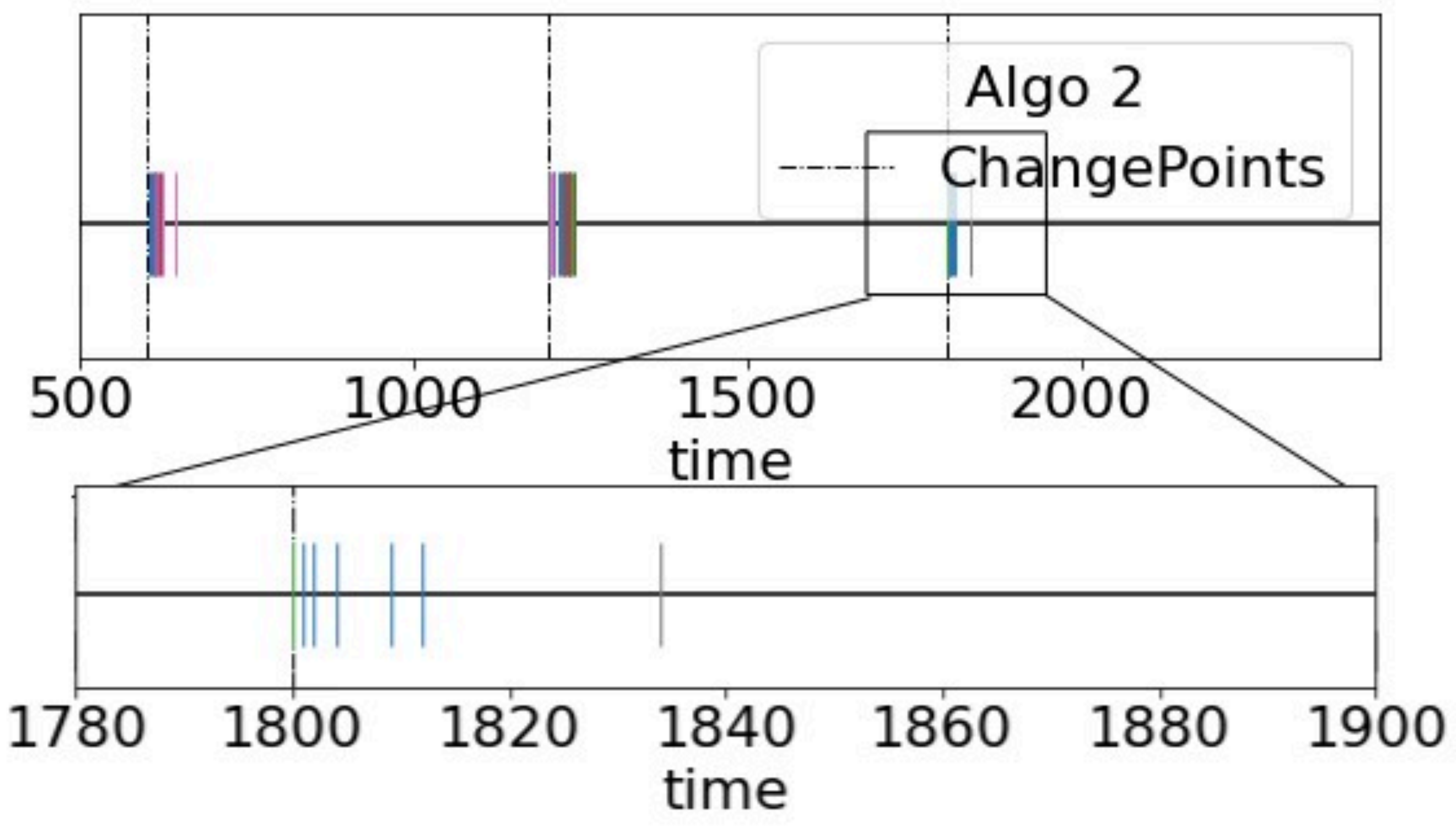
Experiments

Synthetic - Normal and Pareto distributions

shift	data		Ours	$\tau^{30\%}$	DSpot	EQ
x	x	Abs. %	12.1 $\pm$ 1.4	30	15	0
		FP+FN	**0** $\pm$ 0	5.3 $\pm$ 3.7	9.1 $\pm$ 3.6	3.9 $\pm$ 2.1
✓	x	Abs. %	32.7 $\pm$ 9.1	30	15	0
		FP+FN	**5.2** $\pm$ 1.3	195 $\pm$ 71	225 $\pm$ 95	219 $\pm$ 71
x	✓	Abs. %	9.3 $\pm$ 1.2	30	15	0
		FP+FN	**0** $\pm$ 0	5.3 $\pm$ 3.7	9.1 $\pm$ 3.6	3.9 $\pm$ 2.1
✓	✓	Abs. %	18.9 $\pm$ 4.1	30	15	0
		FP+FN	**2.0** $\pm$ 1.5	155 $\pm$ 69	173 $\pm$ 64	160 $\pm$ 64

Table 1: Synthetic dataset results. We used Algorithm 1, 3, 4, 5 as "ours" for the four settings respectively. Compared to baselines, we achieve significant less mistakes (FP+FN) with low abstain rate, especially in settings with shift.



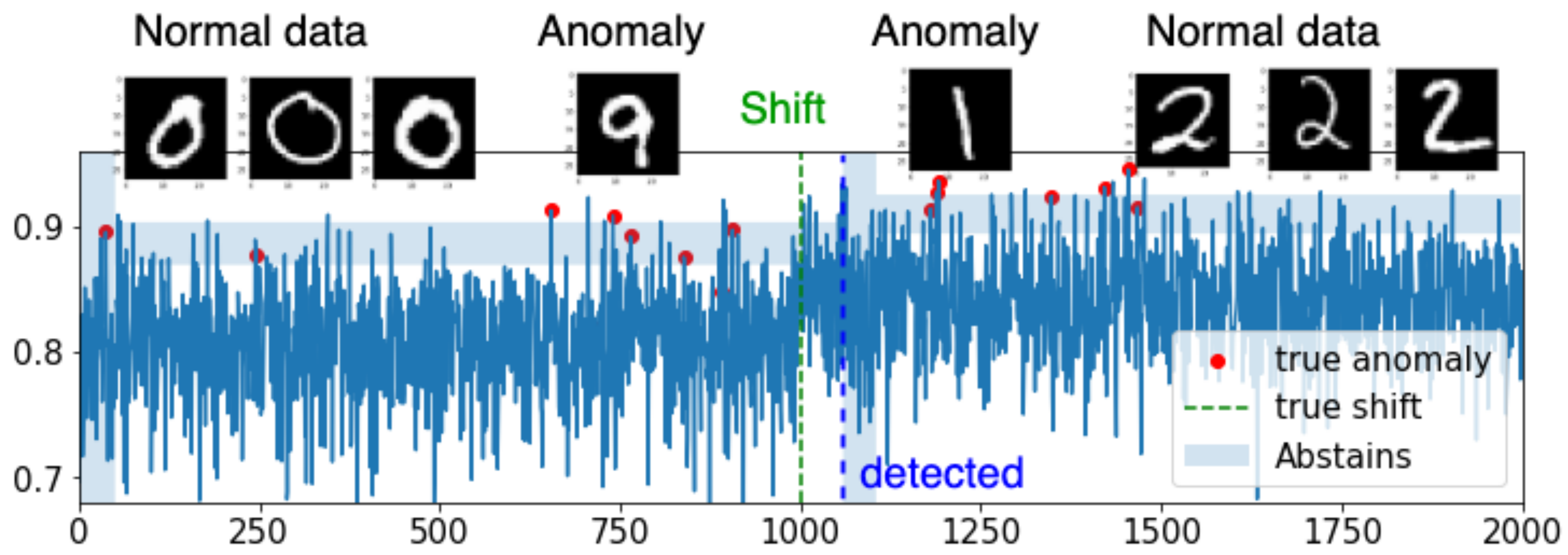


Abstains grows at $\mathcal{O}(\sqrt{T})$

Mistakes cluster at change points

Experiments

MNIST



Static	abstains	FP+FN	Shift	abstains	FP+FN
$\tau^{30\%}$	327 $\pm$ 0	12.7 $\pm$ 0.9	$\tau^{30\%}$	670 $\pm$ 0	91 $\pm$ 8.0
DSpot	150 $\pm$ 0	78.1 $\pm$ 2.5	DSpot	150 $\pm$ 0	129.5 $\pm$ 10.2
IF+A1	172 $\pm$ 16	17.6 $\pm$ 5.3	IF+A3	190 $\pm$ 39	75.7 $\pm$ 18.9
NN+A1	133 $\pm$ 4	**3.9** $\pm$ 0.7	NN+A3	339 $\pm$ 12	**9.3** $\pm$ 2.1
NN+A5	89 $\pm$ 6	**2.1** $\pm$ 0.2	NN+A5	210 $\pm$ 8	**8.8** $\pm$ 2.3

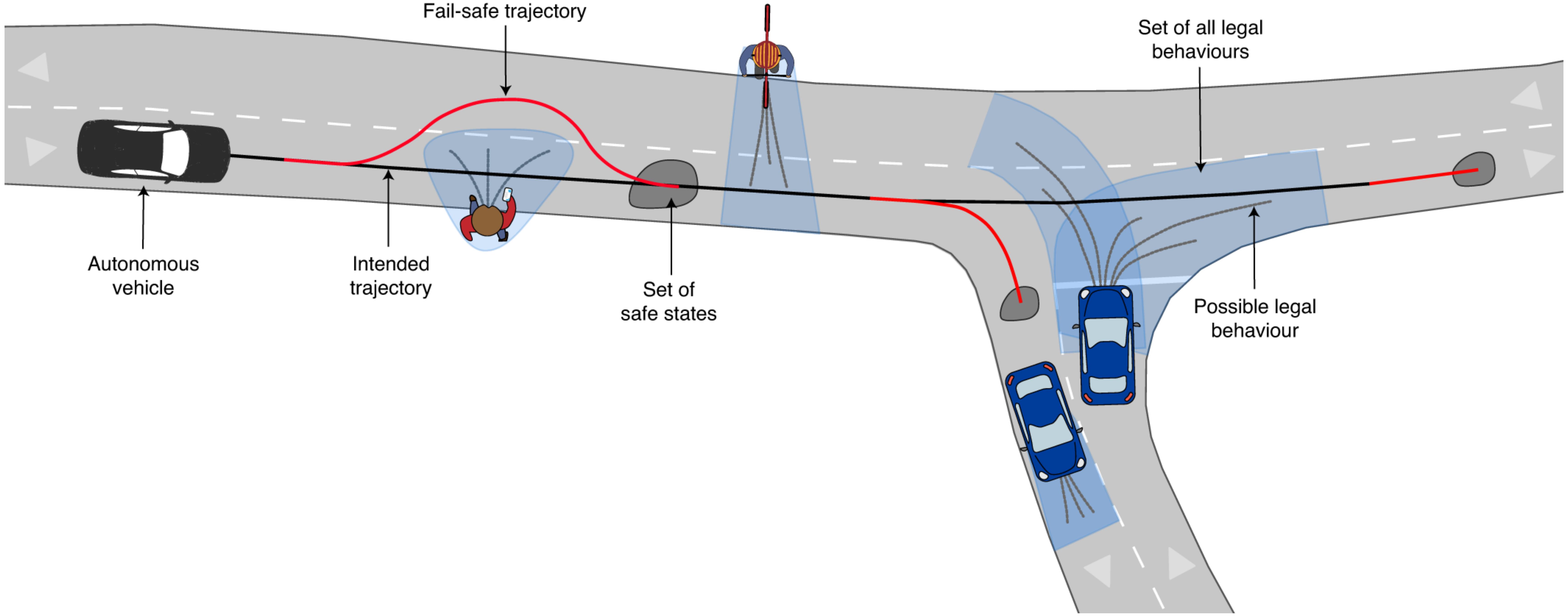
Table 2: One class MNIST result. We applied our algorithms (A as shorthand) to anomaly scores generated by Isolation Forest (IF) and neural networks (NN) and achieve much lower mistakes with moderate number of abstains.

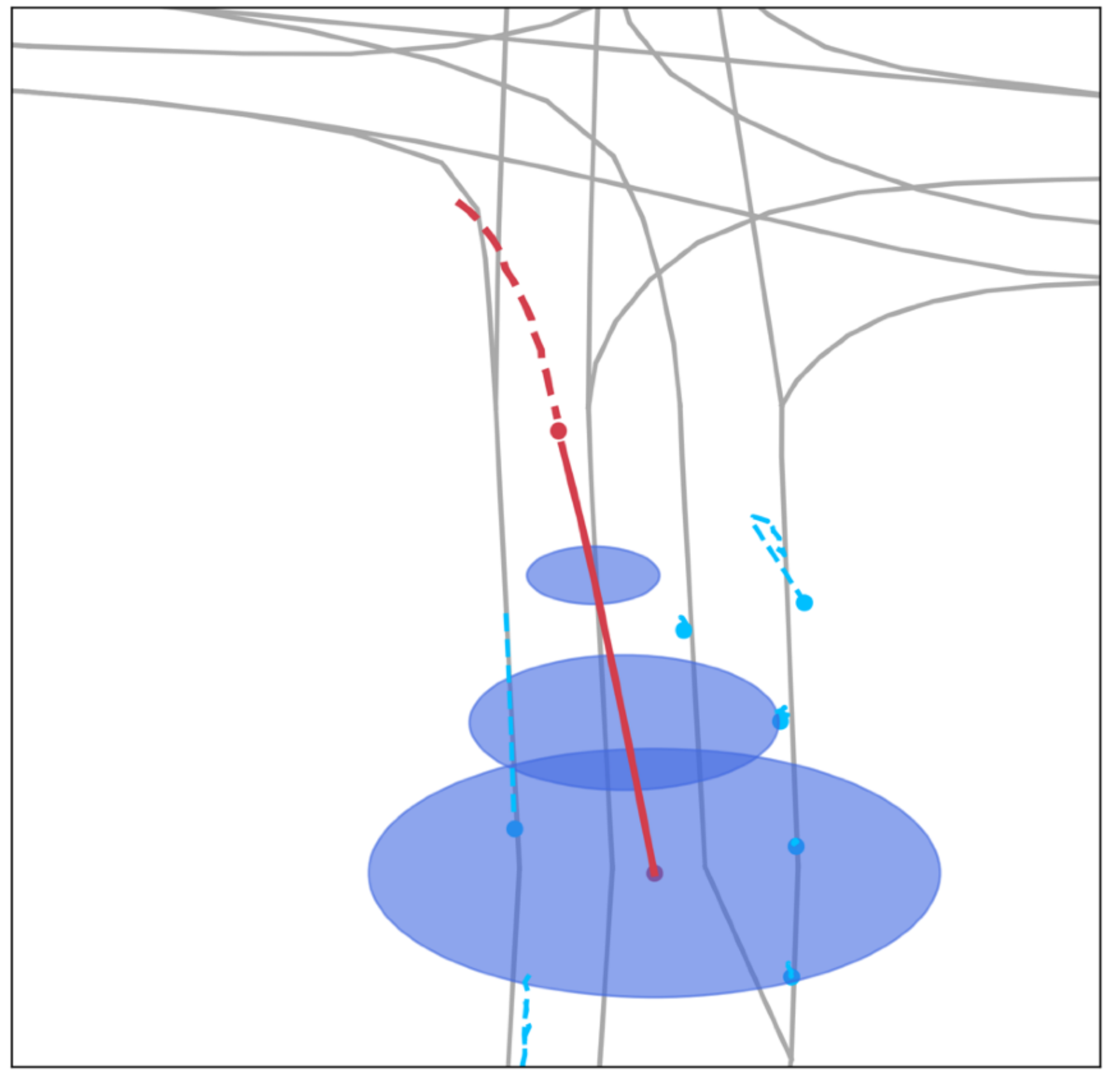
Talk Outline

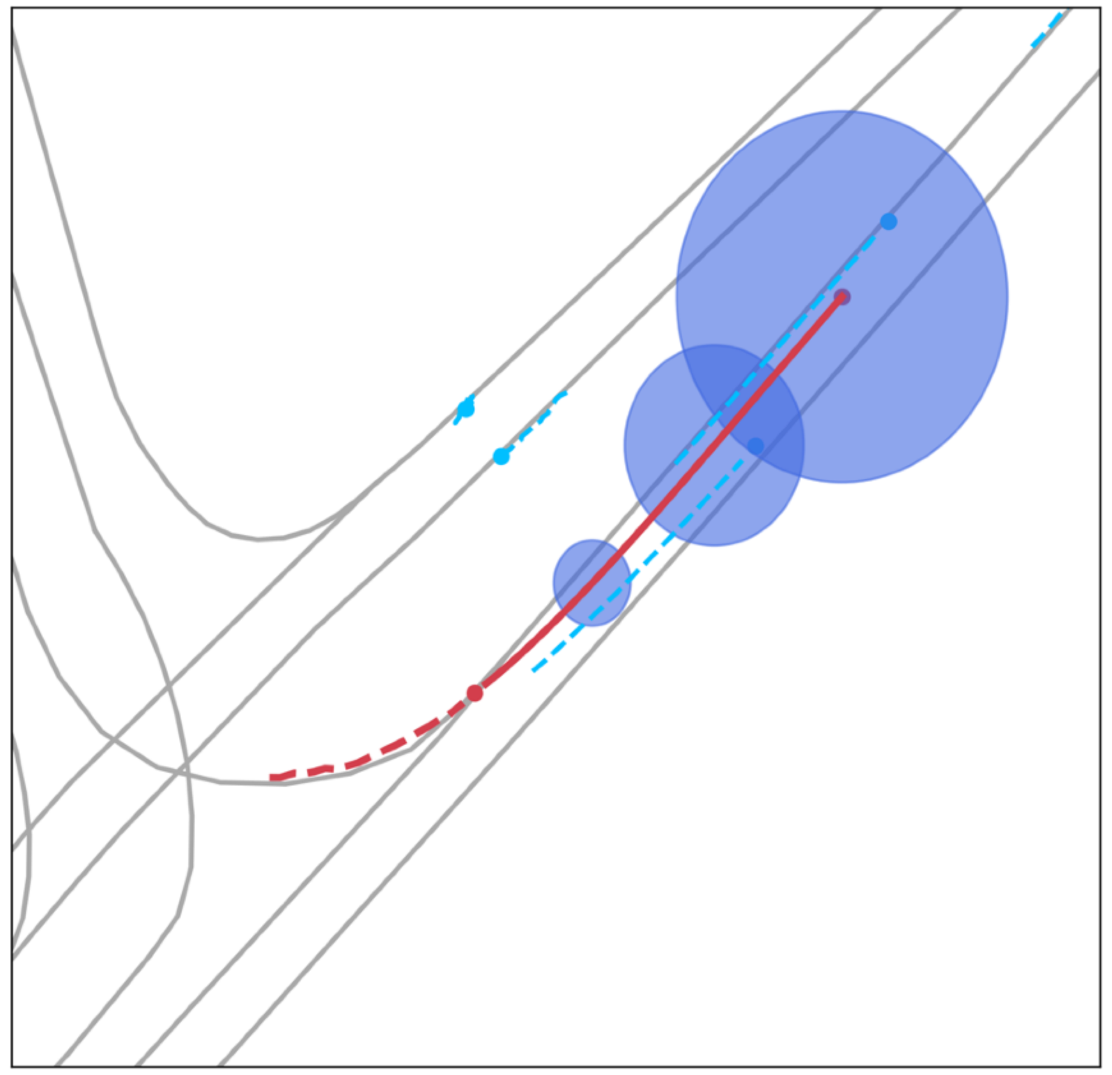
- Part I: Probabilistic Modeling and Uncertainty Quantification
  - Leveraging structure in model design
  - Leveraging structure in post-hoc calibration
- Part II: Decision making Under Uncertainty
  - Selective prediction
    - Example: No-mistake anomaly detection
  - **Safety constraints and pessimistic planning**
    - Example: Robot navigation
- Discussion and Conclusion

Planning using conformal prediction

- Encode confidence regions as dynamic obstacles $\mathcal{O}_t$
- Model Uncertainty Propagation using CP.



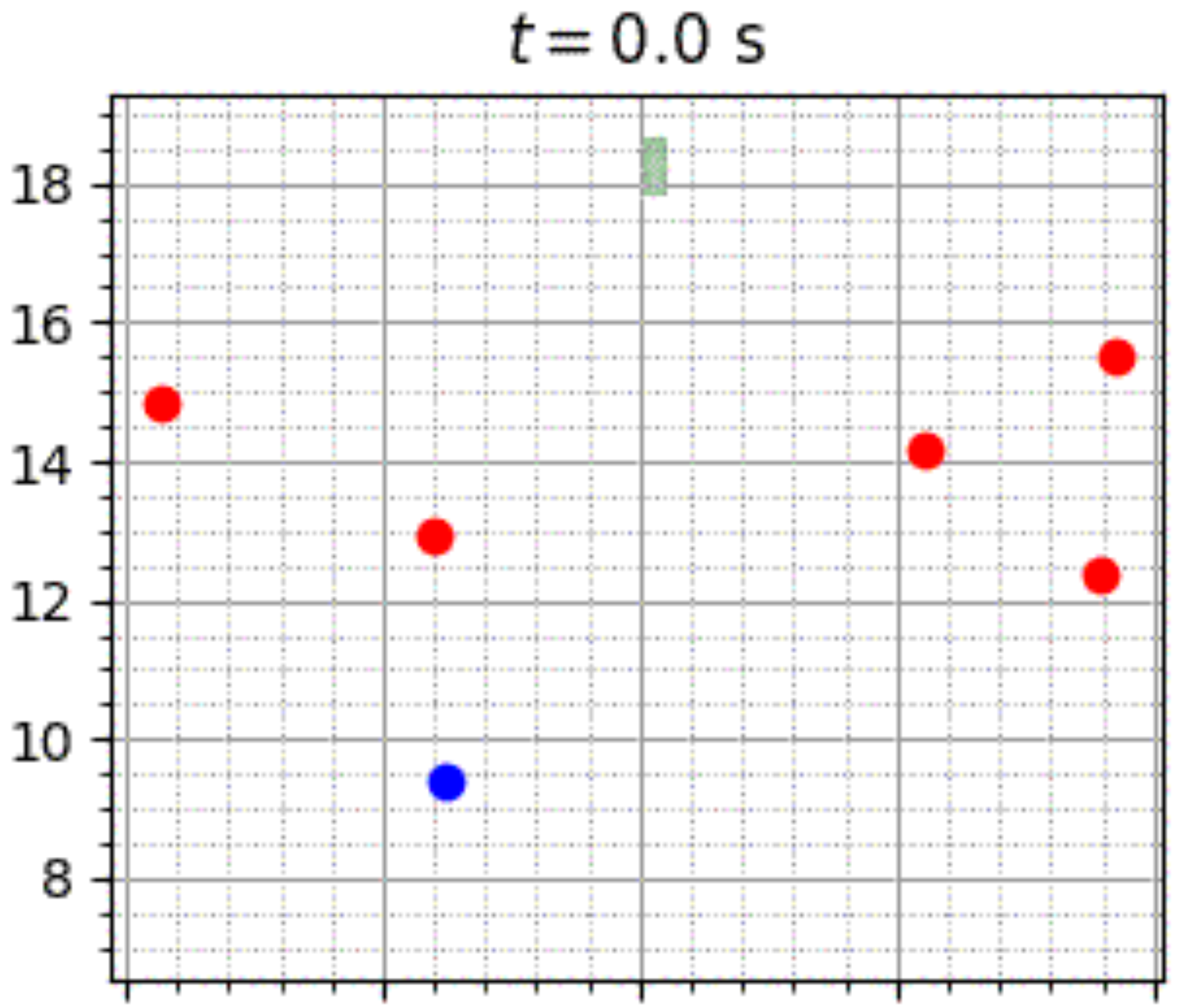




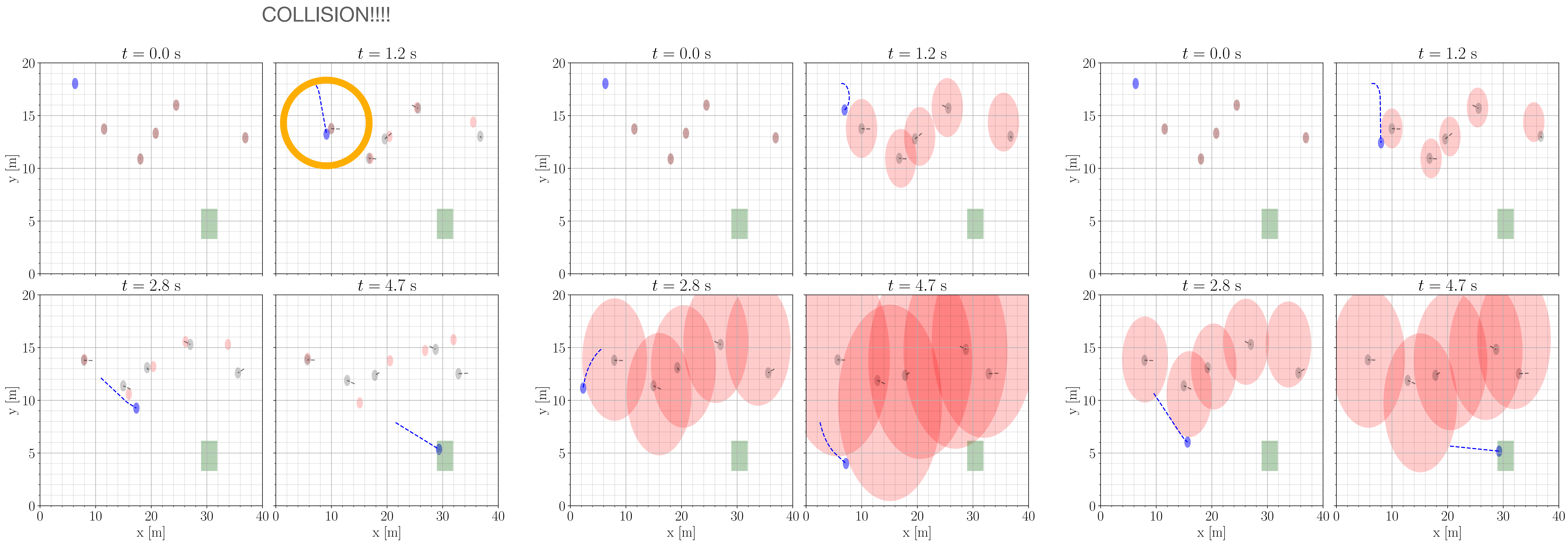
Planning using conformal prediction

- Encode confidence regions as dynamic obstacles $\mathcal{O}_t$
- Model Uncertainty Propagation using CP.

$$
\begin{aligned}
\min_{s_{1:N+1}, a_{1:N}} \quad & J(s_{1:N+1}, a_{1:N}), \\
\text{s.t.} \quad & s_{t+1} = f(s_t, a_t) \qquad t \in \{1, \ldots, N\}, \\
& s_1 = s_{\text{init}}, \quad s_{N+1} \in \mathcal{S}_{\text{final}}, \\
& a_t \in \mathcal{A},\ s_t \in \mathcal{S},\ \boxed{g(s_t) \notin \mathcal{O}_t} \quad t \in \{1, \ldots, N\}
\end{aligned}
$$



Planning using conformal prediction

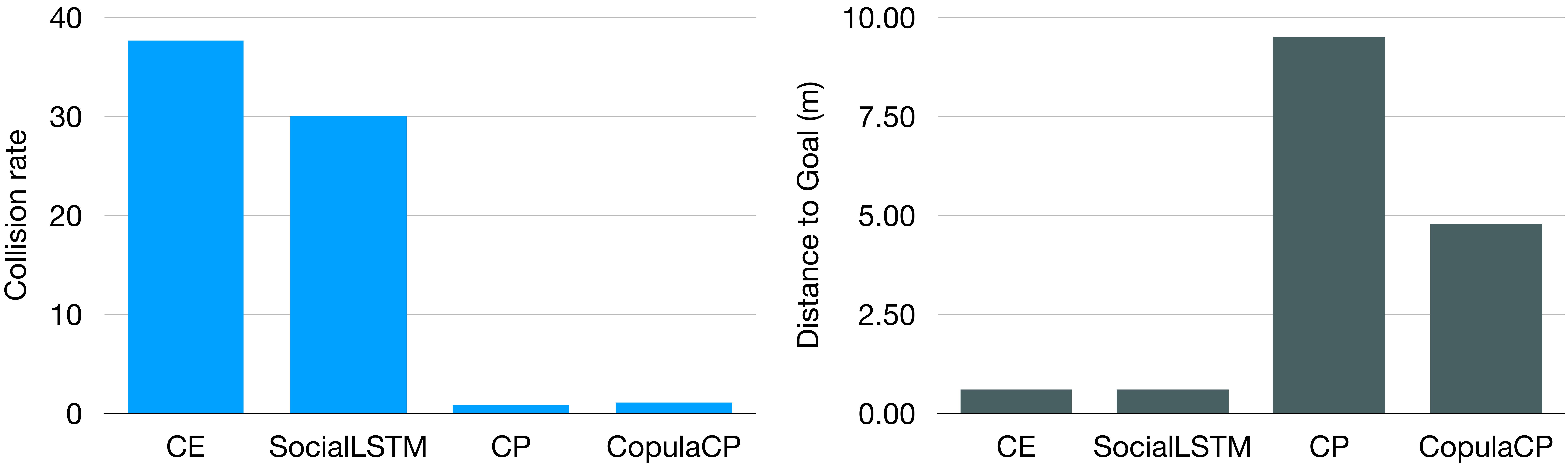


(a) Certainty Equivalence

(b) Union Bounding

(c) Copula CP

Ensures safety and promotes robustness



Talk Outline

- Part I: Probabilistic Modeling and Uncertainty Quantification
  - Leveraging structure in model design
  - Leveraging structure in post-hoc calibration
- Part II: Decision making Under Uncertainty
  - Selective prediction
    - Example: No-mistake anomaly detection
  - Safety constraints and pessimistic planning
    - Example: Robot navigation
- **Discussion and Conclusion**

Summary and Conclusion

- We introduced *methods to quantify uncertainty* for deep learning-based time series models.
  - We can leverage structure within the data, such as equivariance and distributional knowledge, to achieve more calibrated probabilistic predictions.
  - Conformal Prediction simplifies the UQ problem, producing calibrated uncertainty sets.
  - We can leverage structure for post-hoc calibration, such as temporal correlation or state-space information, to achieve sharper intervals.

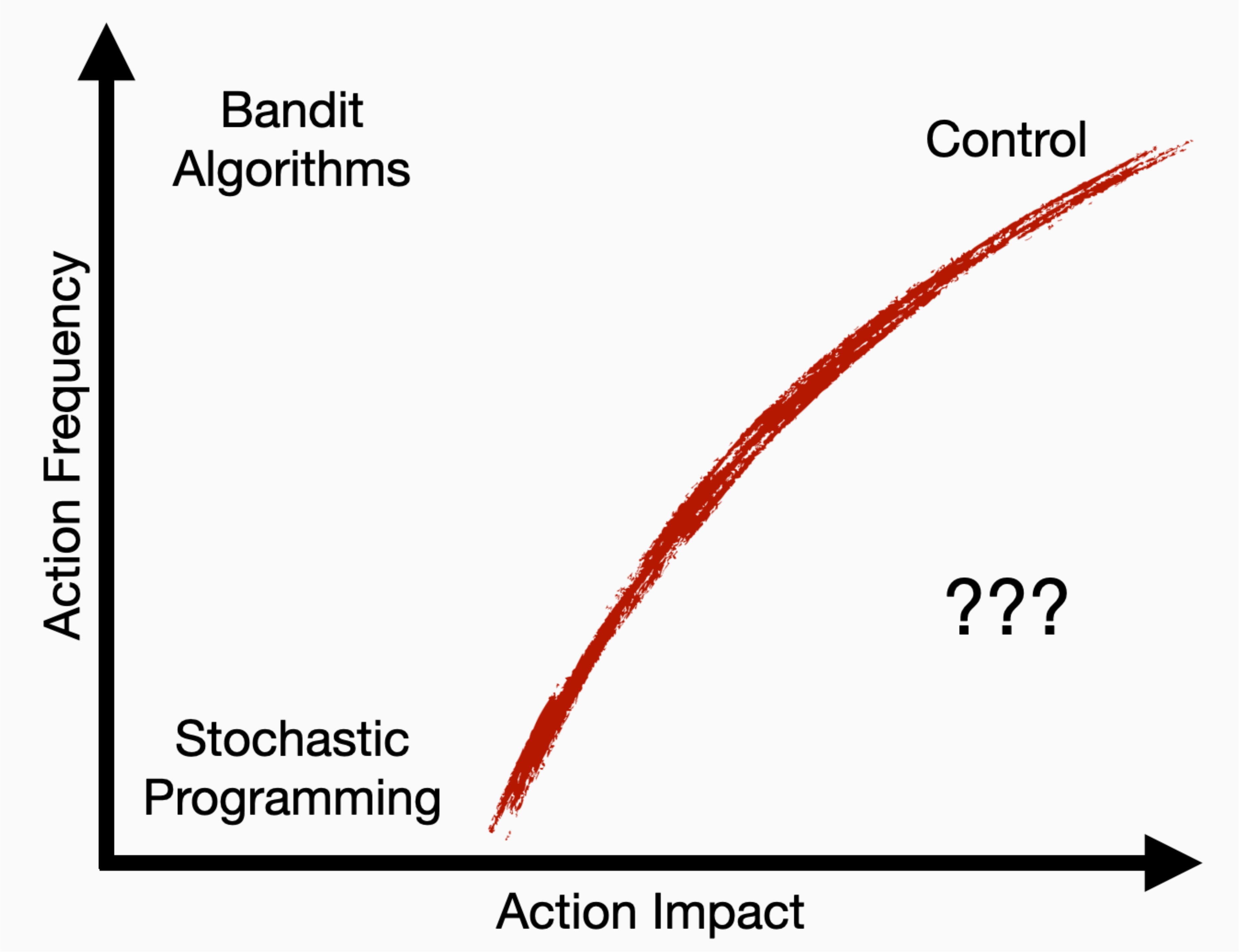
Summary and Conclusion

- We explored principled methodologies to *make decisions under uncertainty*.
  - Selective prediction allows for abstains, adding flexibility to the decision-making framework.
  - With tools like confidence sequences, we can achieve anytime guarantees on mistakes and abstentions.
  - Alternatively, using predictive uncertainty as (hard or soft) constraints for planning can help steer decisions towards safe regions.

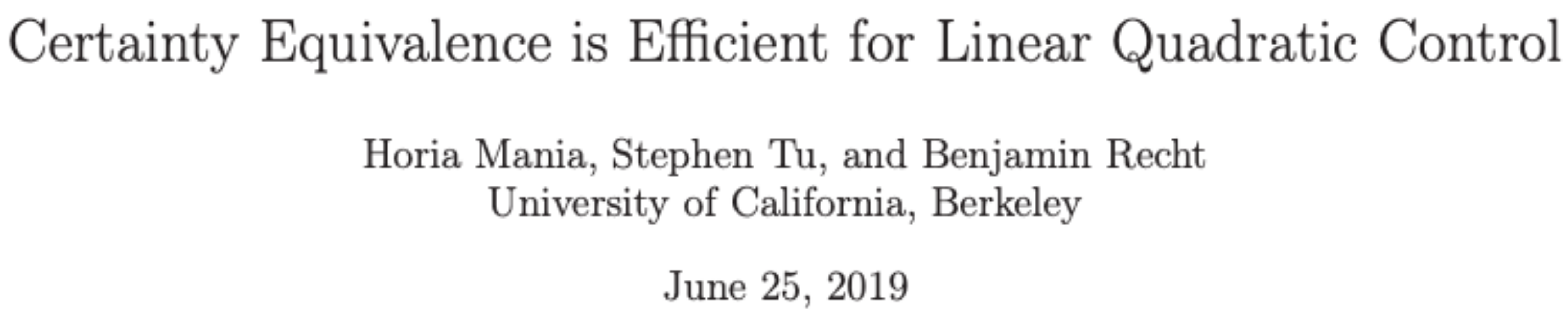
Discussion and Future Work

When is UQ helpful for Decision Making?

- In high frequency settings, UQ brings little utility.
- Greedy or Certainty equivalent / mean-field solutions are enough.
  - state estimation, timely system feedback, and recourse handles uncertainty for you.



Ben Recht (2024), *Purpose Driven Uncertainty Quantification*



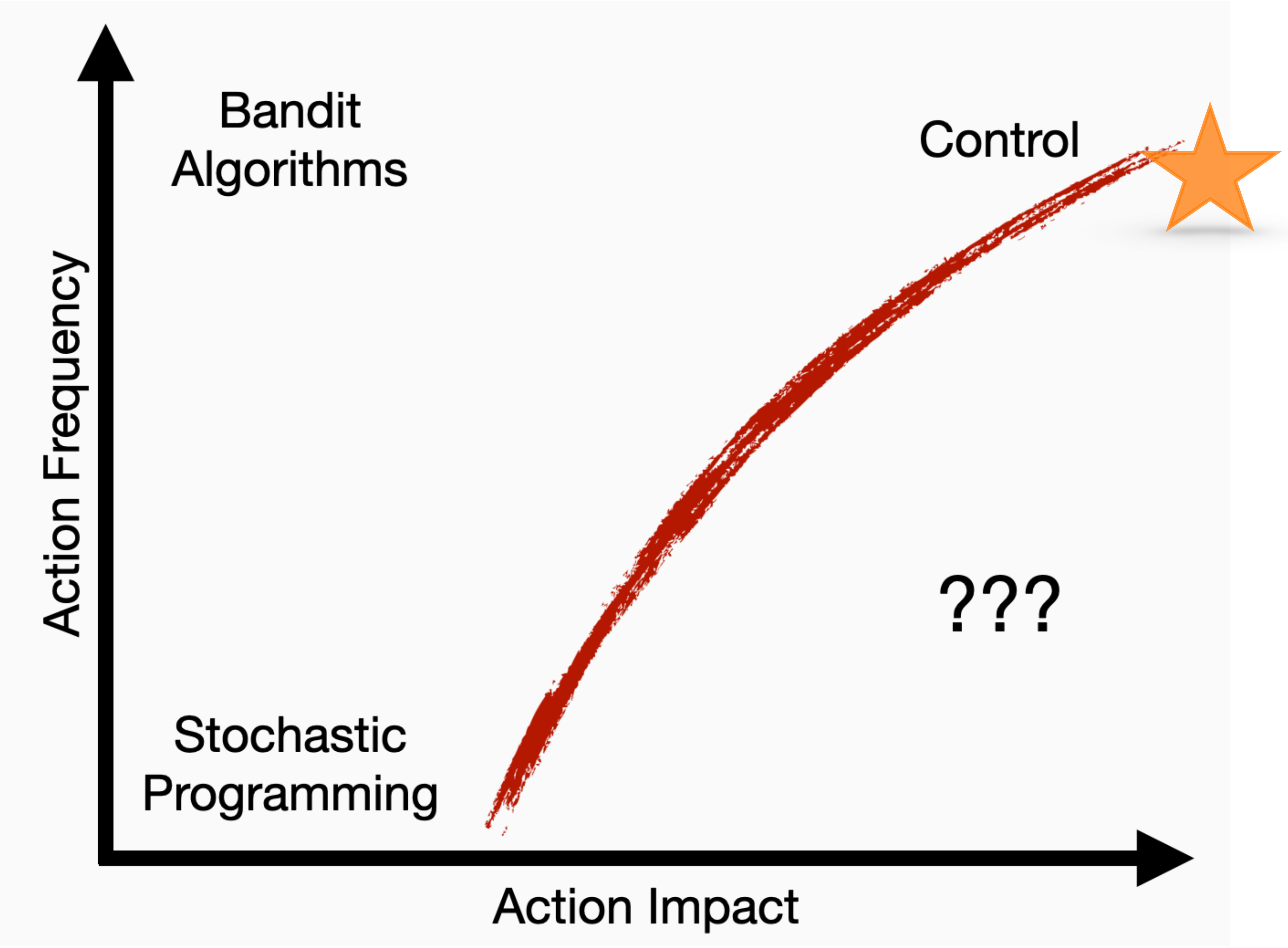
When is UQ helpful for Decision Making?

★ Mission-critical systems (plane, rockets, nuclear plants, medical robots)

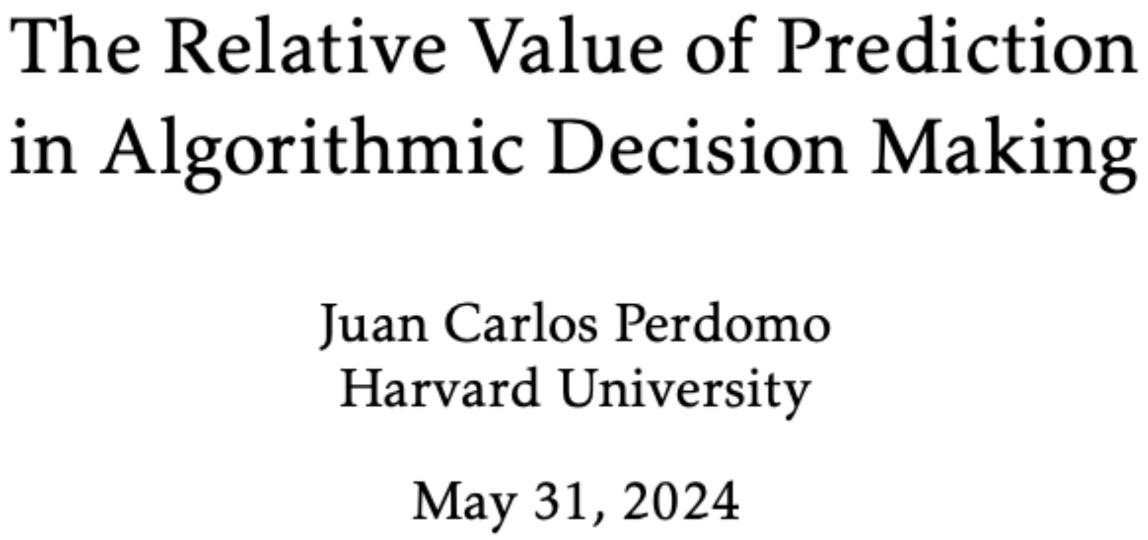
- UQ for conservatism

??? Economic policy, medical diagnosis, LLM alignment.

- UQ as a detailed evaluation of how accurate or trustworthy the model is.

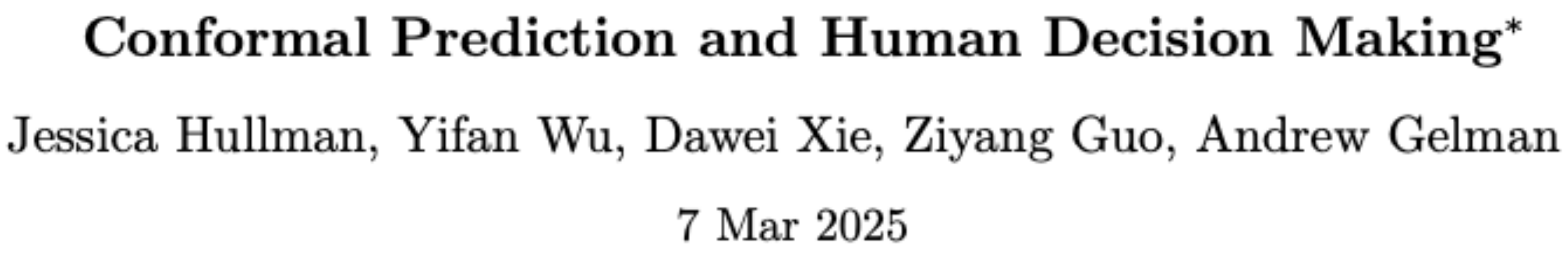


Ben Recht (2024). *Purpose Driven Quantification*


The Relative Value of Prediction in Algorithmic Decision Making

Juan Carlos Perdomo
Harvard University

May 31, 2024


Conformal Prediction and Human Decision Making*

Jessica Hullman, Yifan Wu, Dawei Xie, Ziyang Guo, Andrew Gelman

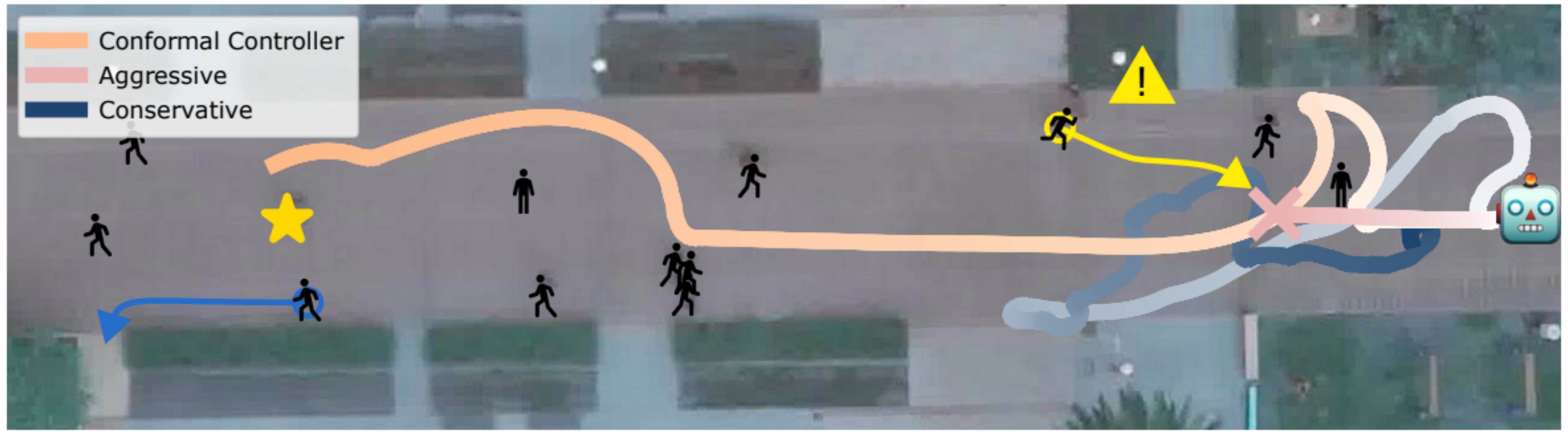
7 Mar 2025

What next?

From prediction only to optimizing-for-decision

Conformal Decision Theory:
Safe Autonomous Decisions from Imperfect Predictions

Jordan Lekeufack[1,*] Anastasios N. Angelopoulos[2,*] Andrea Bajcsy[3,*] Michael I. Jordan[1,2,**] Jitendra Malik[2,**]



Directly calibrate for risk and utility.

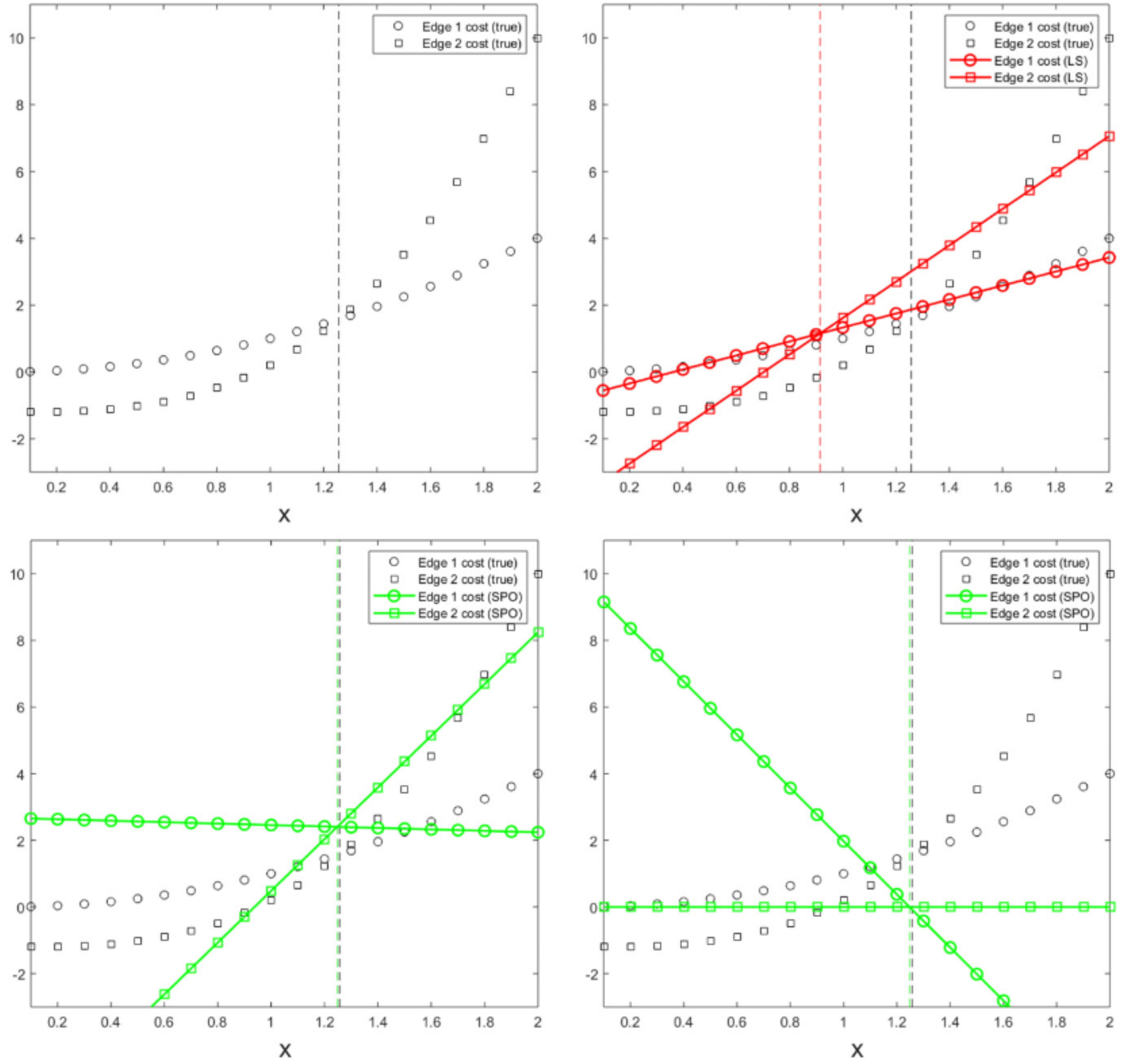
Smart "Predict, then Optimize"

Adam N. Elmachtoub,[a] Paul Grigas[b]

[a] Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York, New York 10027; [b] Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, California 94720

Contact: adam@ieor.columbia.edu, https://orcid.org/0000-0003-0729-4999 (ANE); pgrigas@berkeley.edu, https://orcid.org/0000-0002-5617-1058 (PG)

Figure 3. Illustrative Example



What next?

Beyond predict-then-optimize

- Omni-prediction: optimizing for multiple down-stream decision tasks.

Omnipredictors

Parikshit Gopalan* Adam Tauman Kalai† Omer Reingold‡
VMware Research Microsoft Research Stanford University
Vatsal Sharan§ Udi Wieder¶
USC VMware Research

Robust Decision Making with Partially Calibrated Forecasts

Shayan Kiyani[1], Hamed Hassani[1], George Pappas[1], and Aaron Roth[1]
[1]University of Pennsylvania

October 28, 2025

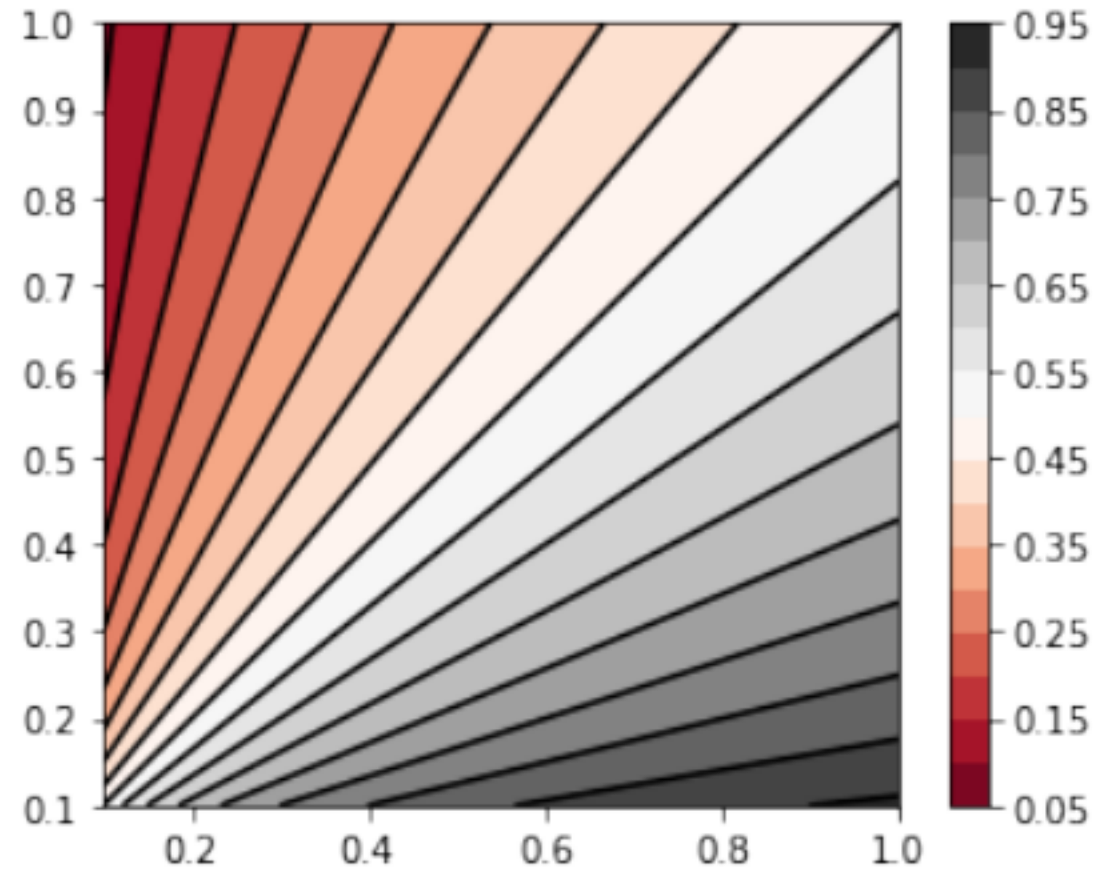


Figure 1: Binary classification with target function $\Pr[y = 1|x] = \frac{x_1}{x_1+x_2}$ for $x \in [0.1, 1]^2$. As can be seen from the level sets, the direction of the optimal linear classifier varies depending on the cost of false positives and negatives. This example is learned to near optimal loss for any loss with fixed costs of false-positives and false-negatives by an omnipredictor for the class $\mathcal{C} = \{x_1, x_2\}$.

Methodology closely related to group fairness and multi-calibration.

Thank You!

Acknowledgments

Prof. Rose Yu
UCSD

Prof. Robin Walters
Northeastern University

Prof. Sylvia Herbert
UCSD

Sander Tonkens
UCSD

Jinxi Li
Hong Kong Polytechnic U

Murali Narayanaswamy
AWS

Abishek Sankararaman
AWS

Sonia Fereidooni
UCSD, now AWS

Aysin Tumay
UCSD

Elise Jortberg
Johnson & Johnson

Zihao Zhou
UCSD

Acknowledgments

Thank You!

Backups

Conformal Prediction

the good and the bad

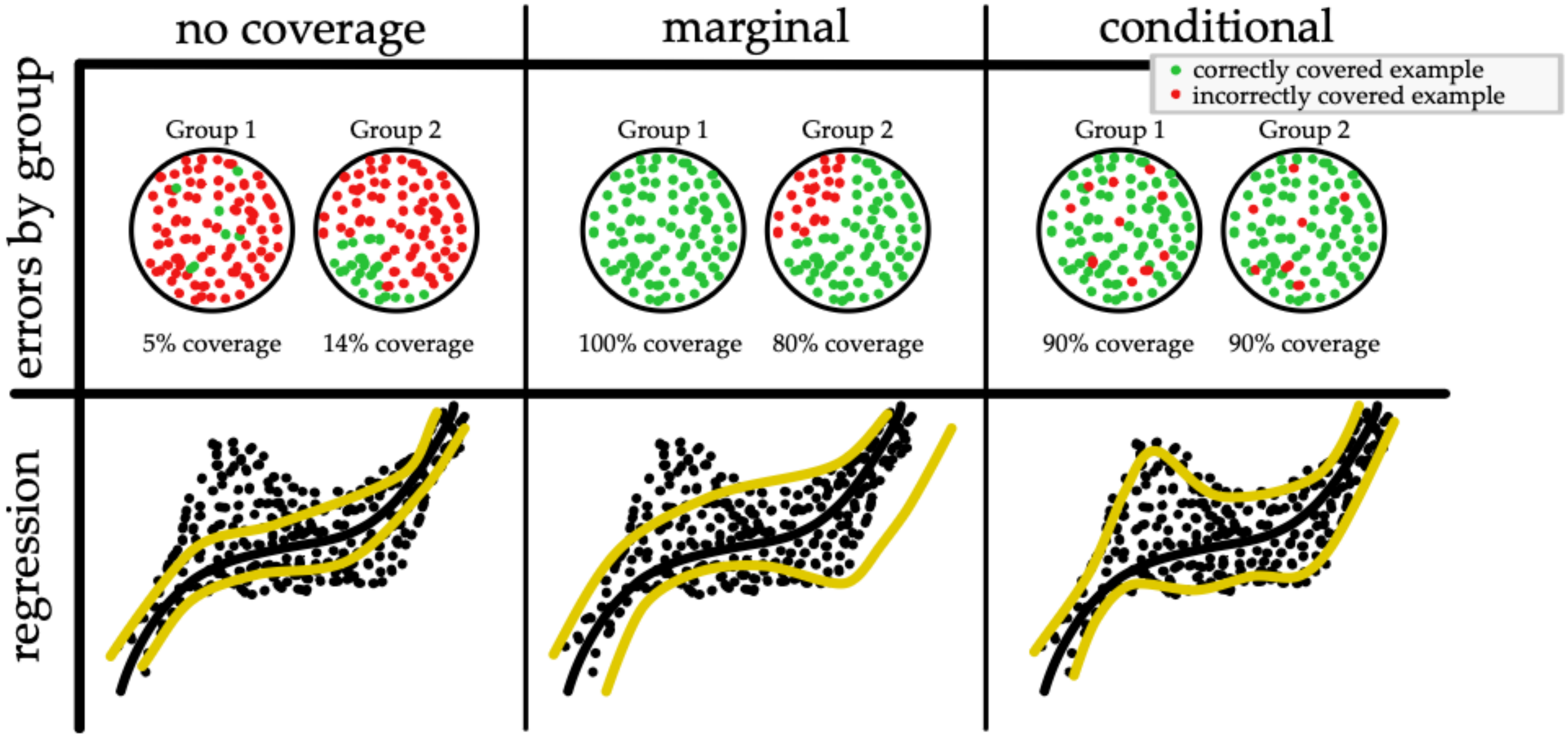


Figure 10: Prediction sets with various notions of coverage: *no coverage, marginal coverage, or conditional coverage (at a level of 90%). In the marginal case, all the errors happen in the same groups and regions in X-space. Conditional coverage disallows this behavior, and errors are evenly distributed.*