Algorithms with Calibrated Machine Learning Predictions

Ellen Vitercik, Stanford

ICML'25

Judy Hanwen Shen

Anders Wikum

Decision-making under uncertainty

In practice, many aspects of inputs are **unknown** *a priori*. E.g.:

- E.g., future traffic or demand in routing

However, we often have **rich historical data**

- ML can help predict unknown aspects of inputs
- Research area: ***Algorithms with predictions***

  [e.g., book chapter by Mitzenmacher, Vassilvitskii, ‘20]

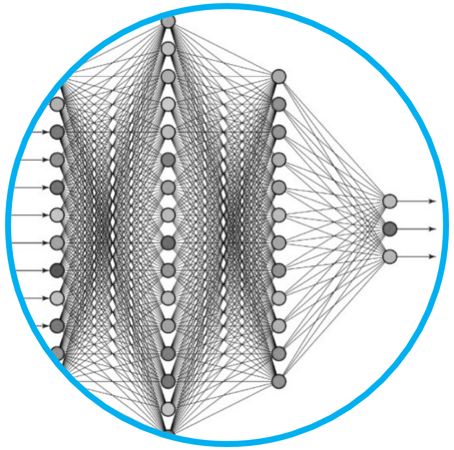
Decision-making under uncertainty

In practice, many aspects of inputs are **unknown** *a priori*. E.g.:

- E.g., future traffic or demand in routing

However, we often have **rich historical data**

- ML can help predict unknown aspects of inputs
- Research area: ***Algorithms with predictions***
  [e.g., book chapter by Mitzenmacher, Vassilvitskii, '20]



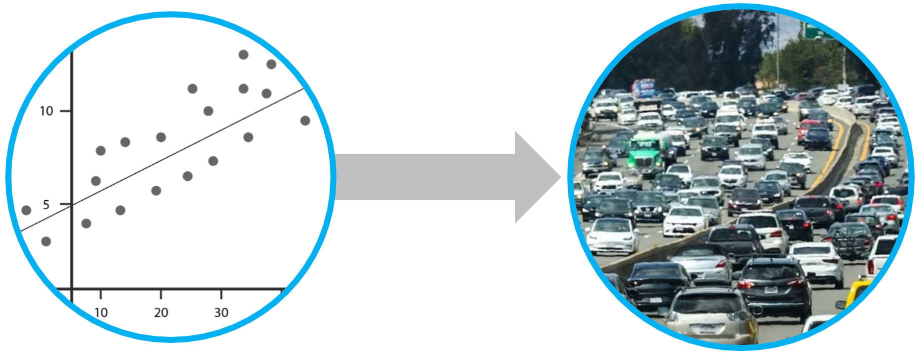
Goal: Open the ML black-box in algorithms with predictions

- ML **model selection** [Heydari, Saberi, **V**, Wikum, ICML'24; He, **V**, ICML'25]
- This talk: **uncertainty quantification** [Shen, **V**, Wikum, ICML'25]

Algorithms and prediction uncertainty

Challenge: prediction **errors can amplify** in decision-making

Don't blindly trust predictions



Insight: ML models can estimate uncertainty **automatically**

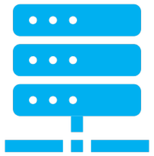
- Well-defined, statistical notion of if prediction can be trusted
- Examples: calibration and conformal predictions [Sun et al. '24]

Our contributions

Demonstrate calibration's utility through two case studies:

Ski Rental

- Prototypical online rent-or-buy decision problem
- Algorithm with guarantees that improve with accuracy and calibration error



Online Job Scheduling

- Calibrated predictions yield better schedules than prior work
  [Cho et al., '22]

Validate methods on real-world datasets

Additional related work

Probabilistic/distributional predictions

[Anand et al. '20; Gupta et al. '21; Diakonikolas et al. '21; Lin et al. '22; Cho et al. '22; Angelopoulos et al. '24; Dinitz et al. '24]

Learning prediction reliability online

[Khodak et al. '22]

Sun et al. '24: Algorithms with **conformal** ML predictions

- We show calibration has key advantages over conformal methods:
- Especially helpful when predictions have high variance

Outline

1. Introduction
2. **Background**
3. Case study 1: Ski Rental
4. Case study 2: Scheduling
5. Conclusions and future directions

Calibration (binary target)

- Random variables (X, Y) with support $\mathcal{X} \times \{0,1\}$
- $f: \mathcal{X} \to [0,1]$ is **calibrated** if $\mathbb{P}[\, Y = 1 \mid f(X) = p \,] = p$
  - E.g., rain prediction: weather is rainy on 50% of days where $f(X) = .5$
- Let $T(X) = \mathbb{P}[\, Y = 1 \mid f(X) \,]$ *(equals $f(X)$ if perfectly calibrated)*

$$\underbrace{\mathbb{E}\left[\left(Y - f(X)\right)^2\right]}_{\ell_2 \text{ error}} = \underbrace{\mathrm{Var}(Y)}_{\text{Uncertainty}} - \underbrace{\mathrm{Var}\left(T(X)\right)}_{\text{Sharpness}} + \underbrace{\mathbb{E}\left[\left(T(X) - f(X)\right)^2\right]}_{\text{Calibration error}}$$

- Calibrated, **unsharp**: $f(X) = \mathbb{P}[Y = 1]$ for all X
- Calibrated, **sharp**: $f(X) = \mathbb{P}[Y = 1 \mid X]$ for all X

Outline

1. Introduction
2. Background
3. **Case study 1: Ski Rental**
4. Case study 2: Scheduling
5. Conclusions and future directions

Ski rental problem

- Prototypical online **rent-or-buy** decision-making problem
- Skier will ski for some **unknown** number of days $Z \in \mathbb{R}_+$
- Each day, decide to rent skis for \$1 or buy for one-time cost $\$b$
- **Goal:** Minimize total skiing cost
- Worst-case "breakeven" strategy:
  Rent for b days, and if still want to ski, buy [Karlin et al. '01]

$$\text{Competitive ratio (CR)} := \frac{\text{ALG}}{\text{OPT}} = \frac{\text{amount algorithm pays}}{\min\{Z, b\}} \leq 2$$

Ski rental with predictions: Prior work

Algorithm with prediction of $I(Z > b)$ [Kumar et al. '18]:

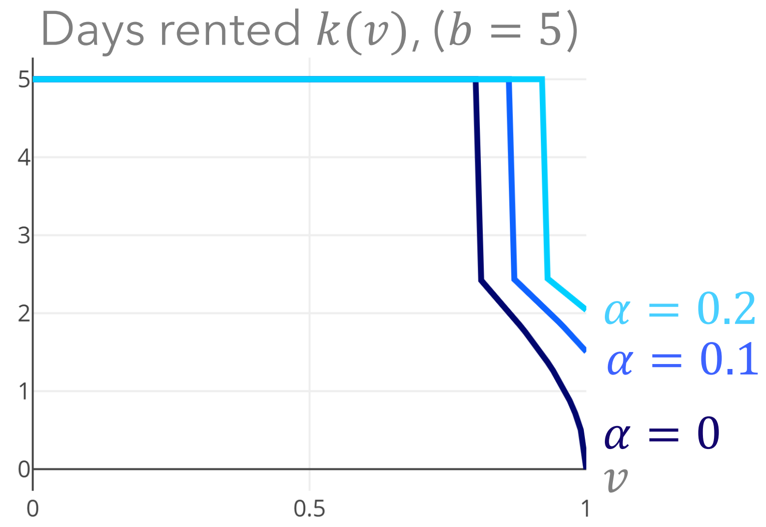
- Uses "trust" parameter $\lambda \in [0,1]$
  - $\lambda = 0$: fully trust predictor
  - $\lambda = 1$: don't trust predictor at all
- *Consistency guarantee*: Perfect prediction yields $\mathrm{CR} \leq 1 + \lambda$
- *Robustness guarantee*: Any prediction yields $\mathrm{CR} \leq 1 + \frac{1}{\lambda}$

Our goal: leverage **calibration** to encode trust/uncertainty

Algorithm with calibrated prediction

- $\mathcal{X}$ = skier features
- Predictor $f(X)$ of target $Y = I(Z > b)$
  - Max calibration error $\alpha = \max_{v \in R(f)} |v - \mathbb{P}[Y = 1 \mid f(X) = v]|$
- **Algorithm:** given prediction $f(X) = v$ rent for $k(v)$ days

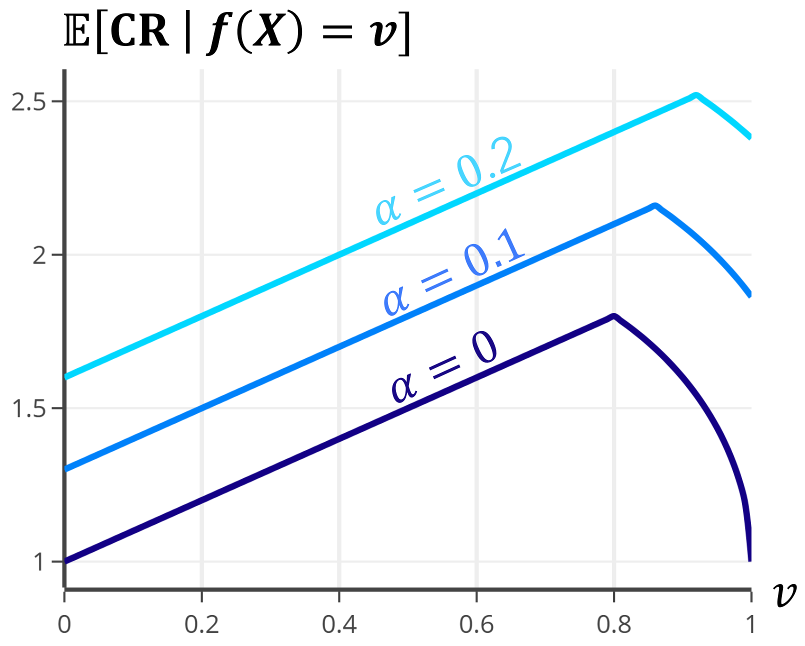
$$k(v) = \begin{cases} b, & v \leq \dfrac{4 + 3\alpha}{5} \\ b\sqrt{\dfrac{1 - v + \alpha}{v + \alpha}}, & \text{else} \end{cases}$$



Ski rental: Main results

Prediction-wise bound:

$$\mathbb{E}[\mathrm{CR} \mid f(X) = v] \le 1 + 2\alpha + \min\left(v + \alpha, 2\sqrt{(v+\alpha)(1 - v + \alpha)}\right)$$



Ski rental: Main results

Prediction-wise bound:

$$\mathbb{E}[\mathrm{CR} \mid f(X) = v] \leq 1 + 2\alpha + \min\left(v + \alpha, 2\sqrt{(v+\alpha)(1 - v + \alpha)}\right)$$

Lower bound: $\forall v$, exists distribution & calibrated predictor s.t.

$$\mathbb{E}[\mathrm{CR} \mid f(X) = v] \geq 1 + \min\left(v, 2\sqrt{v(1-v)}\right)$$

Global bound:

$$\mathbb{E}[\mathrm{CR}] \leq 1 + 3\alpha + \min\left(\mathbb{P}[Z > b], 2\sqrt{\mathrm{MSE}(f) + 3\alpha}\right)$$

✅ Lower MSE and calibration error lead to near-optimal CR

Outline

1. Introduction
2. Background
3. Case study 1: Ski Rental
4. **Case study 2: Scheduling**
5. Conclusions and future directions

Online job scheduling problem

1 machine to process n unit-length jobs

Each job i has **unknown** high ($y_i = 1$) or low ($y_i = 0$) **priority**

- Processing a θ-fraction of a job reveals its priority

Jobs can be stored after partial processing

Objective: Minimize weighted sum of completion times

$$\sum C_i \cdot w_{y_i}$$

Cost per unit delay, with $w_1 > w_0 > 0$

Completion time of job i

Online scheduling with predictions

$\mathcal{X}$ = job features

Predictor $f(X)$ of target $Y = I$(job is high priority)

Scheduling strategies:

- **Preemptive:** Start new job if discover current is low-priority
- **Non-preemptive:** Always process opened jobs to completion

β-threshold rule [Cho et al. '22]

Input: probabilities p_i that job i is high priority

What if predictions are calibrated?
- Cho et al. '22: Specific calibrated, unsharp predictor
- **This paper:** Arbitrary calibrated, **sharp** predictor

1. $\beta \leftarrow \frac{\theta}{1-\theta} \cdot \frac{w_1}{w_1 - w_0}$
2. Order probabilities $p_{i_1} \geq \cdots \geq p_{i_n}$
3. $m \leftarrow |\{i\colon p_i > \beta\}|$
4. Run jobs $i_1, \ldots, i_m$ preemptively, in order
5. Complete remaining jobs non-preemptively, in order

Importance of predictor sharpness

Input: probabilities p_i that job i is high priority

What if predictions are calibrated?
- Cho et al. '22: Specific calibrated, unsharp predictor
- **This paper:** Arbitrary calibrated, **sharp** predictor

Key insight: **interchanges** are the primary source of **regret**

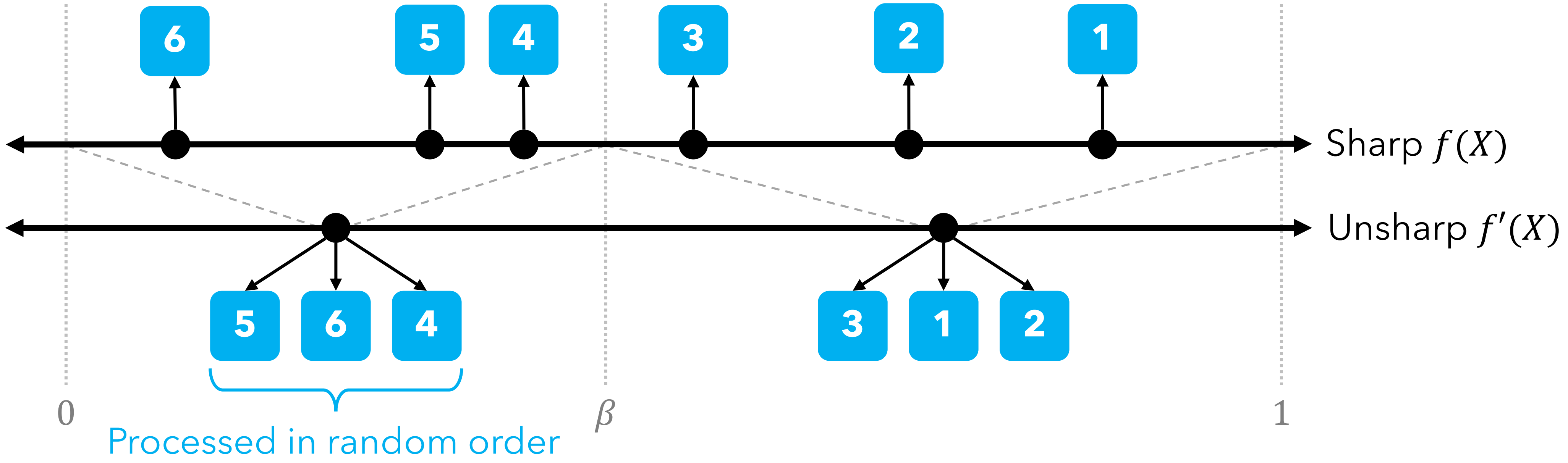
Weighted sum of completion times compared to optimal in hindsight

- Low priority job (partially) processed before high priority job
- **Sharp predictors** lead to **fewer interchanges**

Importance of predictor sharpness

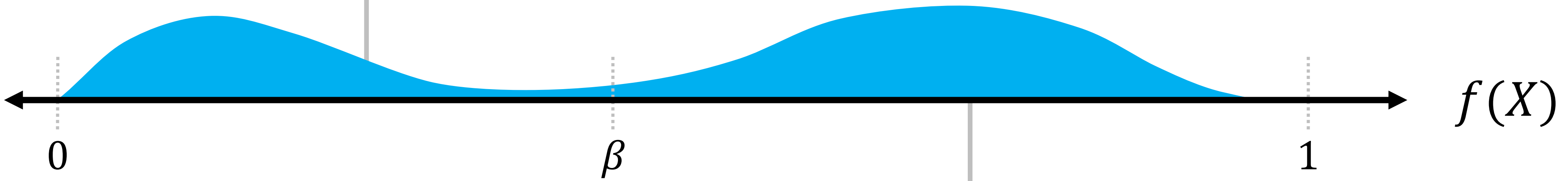
Key insight: **interchanges** are the primary source of **regret**

- Low priority job (partially) processed before high priority job
- **Sharp predictors** lead to **fewer interchanges**



Scheduling: Main result

$$\kappa_1 := \mathrm{Var}(\, f(X) \mid f(X) \le \beta \,) \cdot \mathbb{P}[f(X) \le \beta]^2$$



$$\kappa_2 := \mathrm{Var}(\, f(X) \mid f(X) > \beta \,) \cdot \mathbb{P}[f(X) > \beta]^2$$

Thm: Let $\mathbb{P}[\, f(X) > \beta \mid Y = 0 \,] = \epsilon_0,\ \mathbb{P}[\, f(X) < \beta \mid Y = 1 \,] = \epsilon_1$

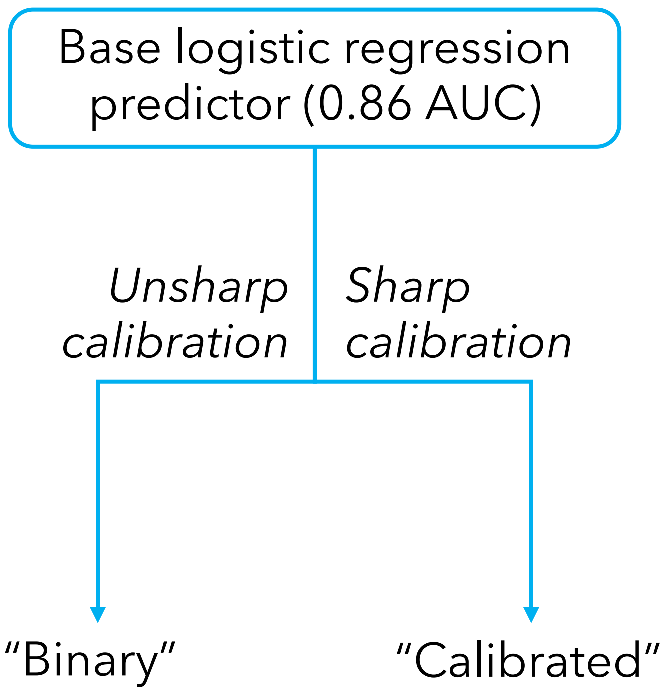
$$\frac{\mathbb{E}[\#\text{ interchanges}]}{\binom{n}{2}} \le \mathrm{Var}(Y)(\epsilon_0 + \epsilon_1) - \kappa_1 - \kappa_2$$

(Eventual) corollary: bound on algorithm's regret (see paper)

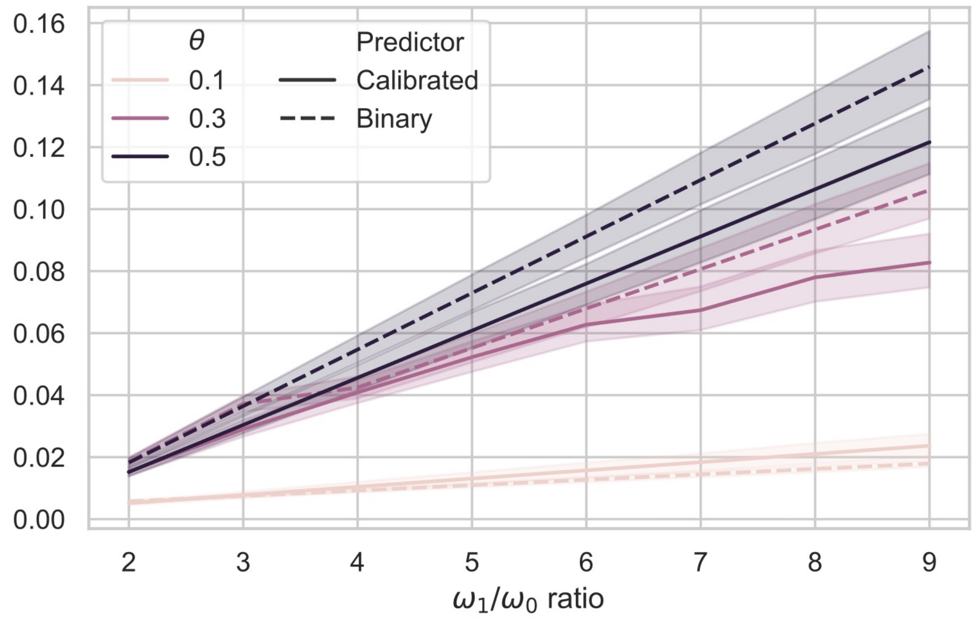
Experiments: Sepsis triage

ML for predicting **sepsis onset** to improve early detection

Dataset of 110,204 hospital admissions



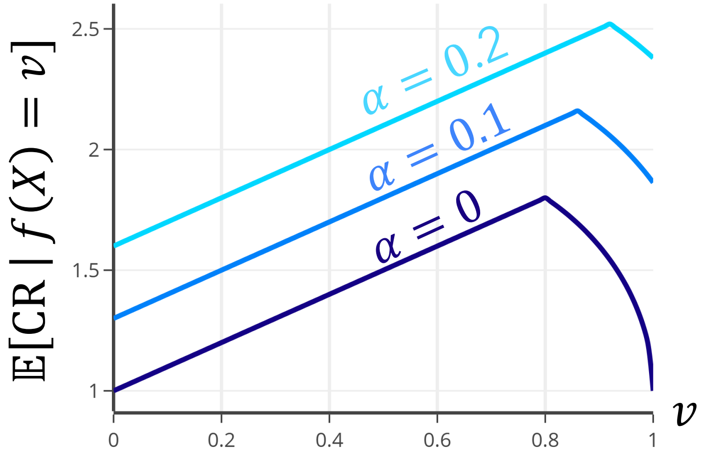
Expected regret per job*



*Scheduling $n = 100$ patient reviews

Conclusions

- Calibration: valuable tool for *algorithms with predictions*
- Case studies: ski rental and job scheduling
- Performance guarantees improve with calibration error
- Validated methods on real-world datasets



Future directions

Further open the ML black box of *algorithms with predictions*

- Analyze how decisions depend on **predictor properties**
  - MSE, calibration, sharpness, and inherent uncertainty
  - False positive/negative rate [see also, e.g., Anand et al. '20]
- Ultimate goal: Guide **model choice** and **training** for decision tasks

Algorithms with Calibrated Machine Learning Predictions

Ellen Vitercik, Stanford

ICML'25

Judy Hanwen Shen

Anders Wikum