(Some) Steps Toward Trustworthy Machine Learning

Thomas G. Dietterich, Distinguished Professor (Emeritus)

Oregon State University, Corvallis, OR USA 97331

Collaborators:

Students: Jesse Hostetler (SRI), Majid Alkaee Taleghan (Esentire), Si Liu (Fred Hutchinson), Risheek Garrepalli (nVidia), Kim Meyer-Hall (Oregon State), Dan Hendrycks (UCB)

Faculty: Alan Fern, Jo Albers (UWyoming), Debashis Mondal

Oregon State University

Outline

- Part 0: Robust AI and Robust Human Organizations
- Part 1: Competence Modeling
  - Calibrated prediction intervals for reinforcement learning
- Part 2: Anomaly Detection
  - Open category detection with guarantees

High Reliability Human Organizations

Todd LaPorte, Gene Rochlin, and Karlene Roberts (Weick, et al., 1999)

- Preoccupation with failure
  - Fundamental belief that the system has unobserved failure modes
  - Treat anomalies and near misses as symptoms of a problem with the system
- Reluctance to simplify interpretations
  - Comprehensively understand the situation
- Sensitivity to operations
  - Maintain continuous situational awareness
- Commitment to resilience
  - Develop the capability to detect, contain, and recover from errors. Practice improvisational problem solving
- Deference to expertise
  - During a crisis, authority migrates to the person who can solve the problem, regardless of their rank

Designing AI Systems to be HROs

- Maintain Situational Awareness
  - AI methods are very good at integrating data from multiple sensors and effectors to estimate a probability distribution over states
- Detect Anomalies and Near Misses
  - Anomalies: Yes
  - Near Misses: Research needed
- Generate Candidate Explanations for Anomalies & Near Misses
  - Very little work: Research needed
- Improvise Solutions
  - Improvisational problem solving that extends or operates outside the system model

Assessment: Designing AI as an HRO

	Assessment
Situational Awareness	A mature methods
Detect Anomalies and Near Misses	B high-dimension, dynamics
Explain Anomalies and Near Misses	D only basic techniques
Improvise Solutions	F

Designing a Human + AI Team as an HRO

- Even very powerful AI systems will be surrounded by a human team
- Situational Awareness
  - AI can track the situation, but humans and AI must establish a shared mental model of the situation: Research needed
  - Humans must be aware of what version of the AI system they are using. When was it last updated/retrained? Research needed
- Detect Anomalies and Near Misses
  - AI system must understand and predict behavior of human team (and detect anomalous behavior)
  - AI and Humans must work together: interactive anomaly detection
- Generate Candidate Explanations for Anomalies & Near Misses
  - Very little work: Research needed
- Improvise Solutions
  - AI should support human improvisational problem solving: Research Needed
  - Example: mixed-initiative planning

Assessment: Human + AI HROs

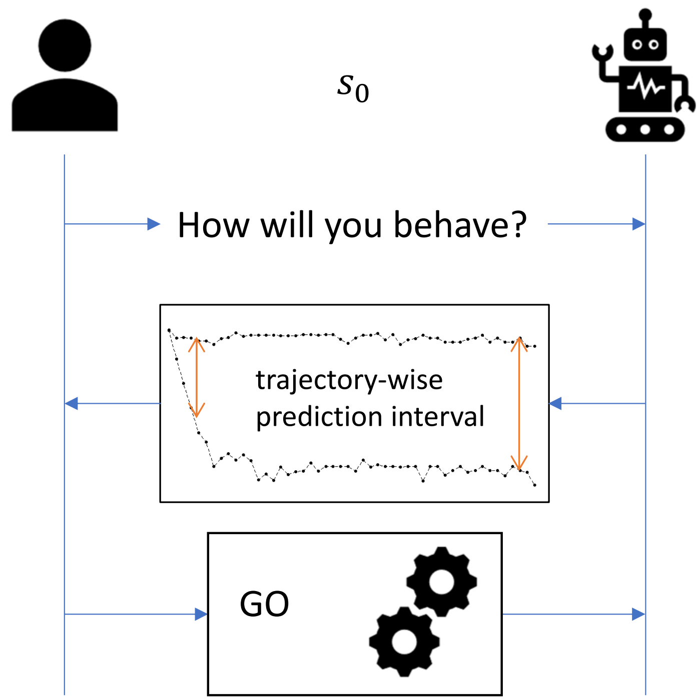
	Assessment
Situational Awareness	C poor UI, poor communication
Detect Anomalies and Near Misses	C some work on user feedback
Explain Anomalies and Near Misses	D only basic techniques
Improvise Solutions	D mixed-initiative planning

Part 1: Competence Modeling: Prospective MDP Performance Guarantees

[D & Hostetler, unpublished]

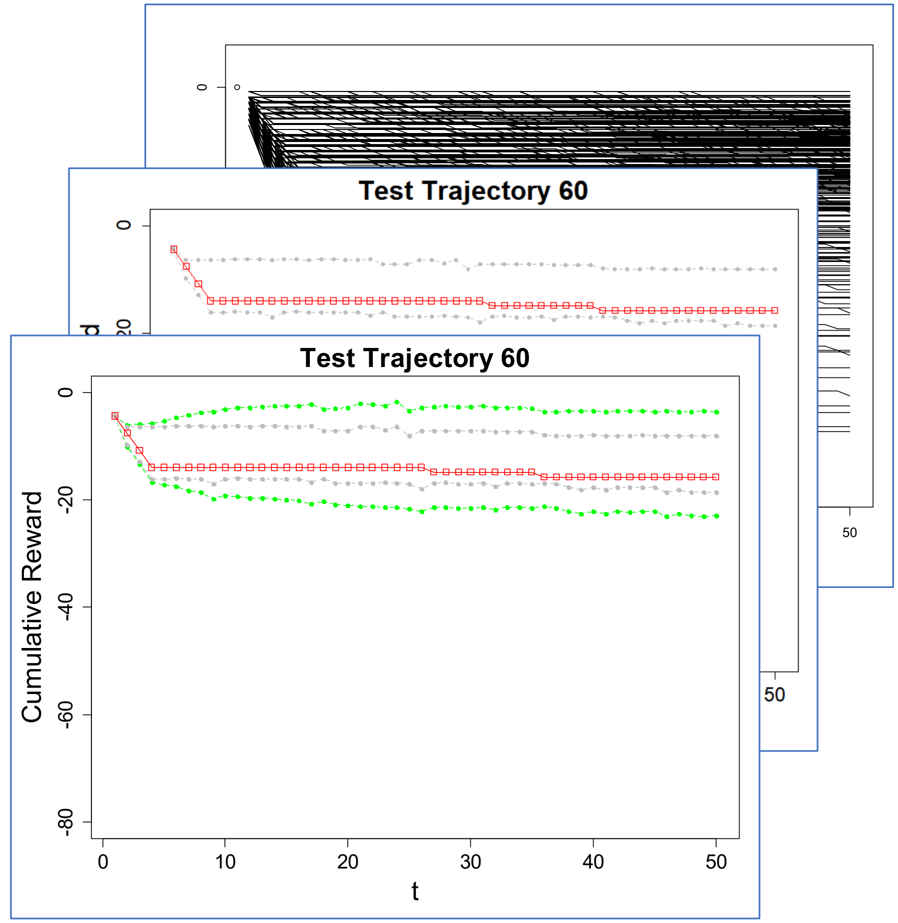
Human decision maker must decide whether to tell an AI assistant to execute policy π starting in state s_0 for h steps

AI assistant provides a trajectory-wise prediction interval that guarantees with probability $1 - \delta$ that its behavior will be inside the interval



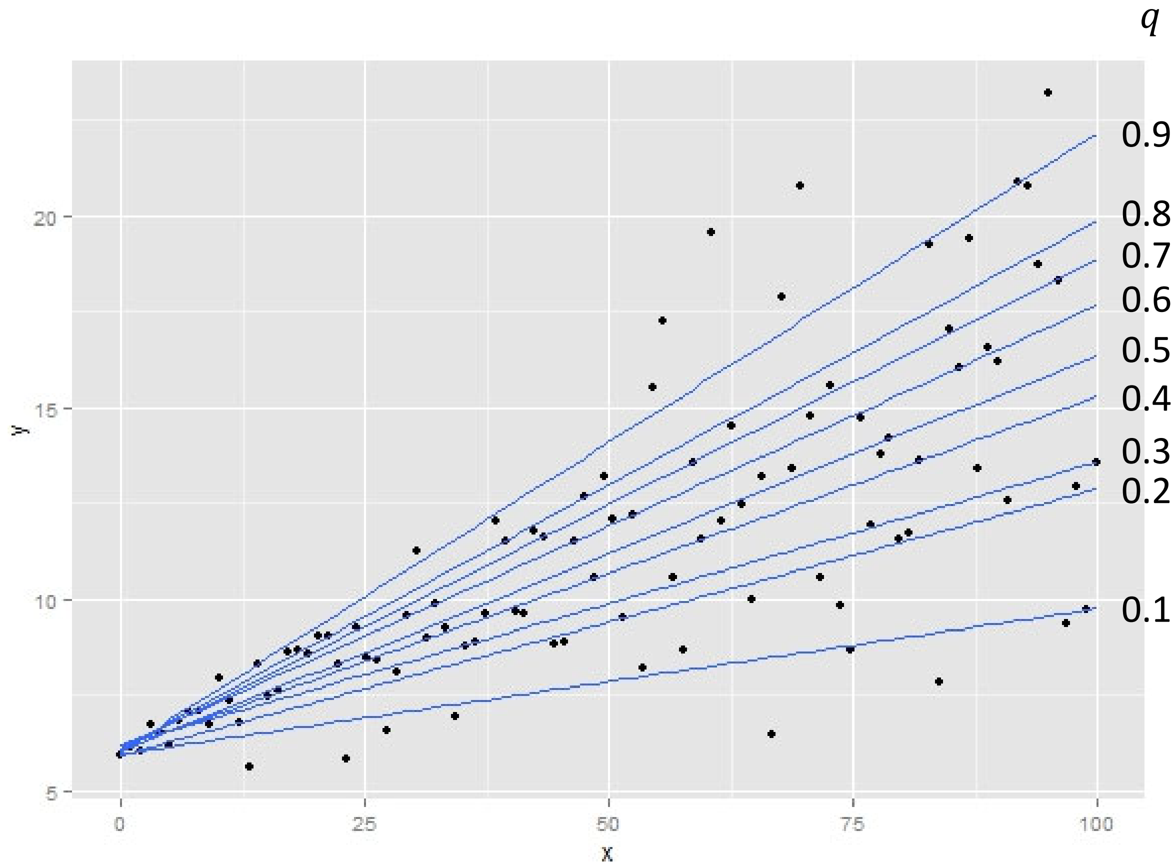
Summary of the Approach

- Repeat N times
  - Sample a starting state $s_0 \sim P_0(\cdot)$
  - Execute π for h steps to obtain a trajectory
- Apply our new technique
  - Perform quantile regression to learn two functions
    - $F_t^{-1}\left(s_0, \frac{\delta}{2}\right)$ an estimate of the $\frac{\delta}{2}$ quantile of the return at time t
    - $F_t^{-1}\left(s_0, 1-\frac{\delta}{2}\right)$ an estimate of the $1-\frac{\delta}{2}$ quantile of the return at time t
  - Adjust these to obtain valid prediction intervals using a new method, ScaledSDTrajectory



Quantile Regression

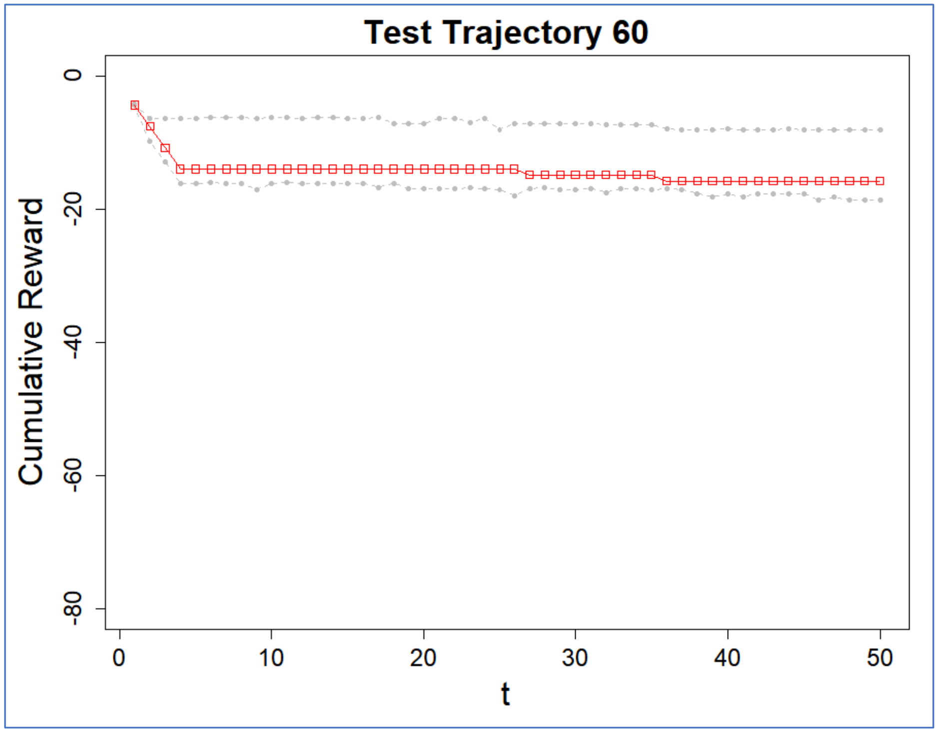
- $P(y|x)$ depends arbitrarily on x
- $F(y|x)$
  - cumulative distribution function of y at x
- $F^{-1}(q|x)$
  - the value of y such that $F(y|x) = q$
- Many algorithms for quantile regression
- We employ Quantile Random Forests (Meinshausen, 2006) to compute the $\delta/2$ and $1 - \delta/2$ quantiles



https://data.library.virginia.edu/files/qreg_fig_5.jpeg

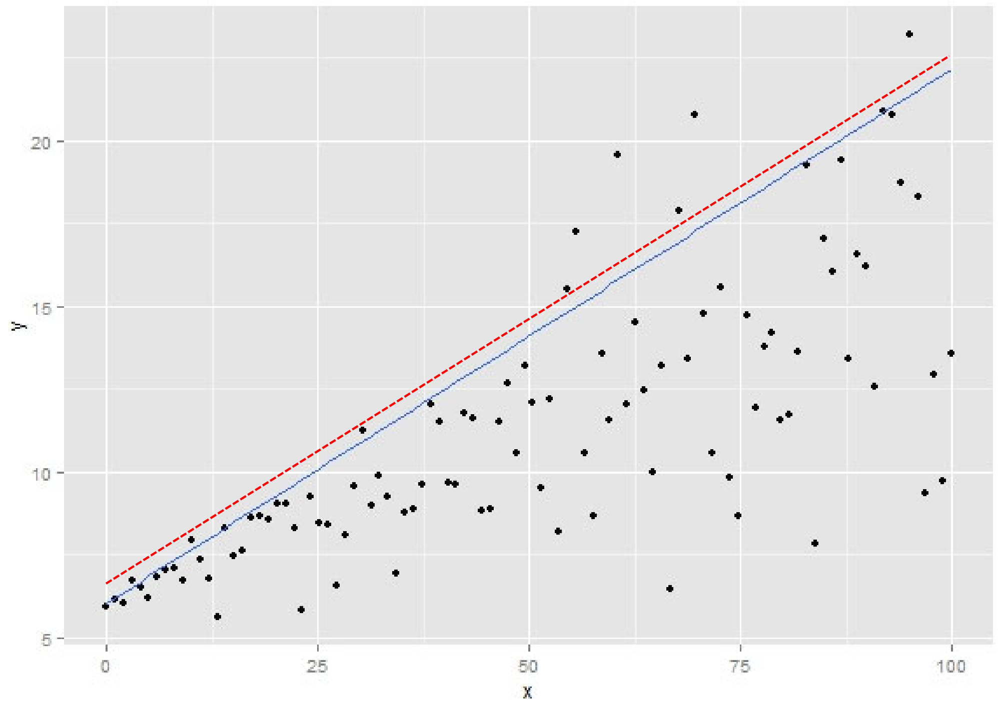
Quantile Regression for Trajectories

- Discrete time MDP with state space $\mathcal{S}$, starting state distribution P_0, and fixed policy π
- h-step trajectory τ
  - sample $s_0 \sim P_0$
  - execute π for h steps
  - collect states, actions, and rewards into τ
- Define a *behavior function* $B(\tau, t)$ to summarize the behavior of the policy at time t
  - some aspect of s_t
  - immediate reward
  - cumulative reward $r_1 + \cdots + r_{t-1}$
  - future reward $r_t + r_{t+1} + \cdots + r_{h-1}$
  - $\boldsymbol{b}(\tau) = \big(B(\tau, 1), \ldots, B(\tau, h)\big)$ is the "behavior vector" of τ
- Fit quantile regression functions for each time step
  - $F_t^{-1}(s_0, \delta/2)$ an estimate of the $\delta/2$ quantile of the return at time t
  - $F_t^{-1}(s_0, 1-\delta/2)$ an estimate of the $1-\delta/2$ quantile of the return at time t



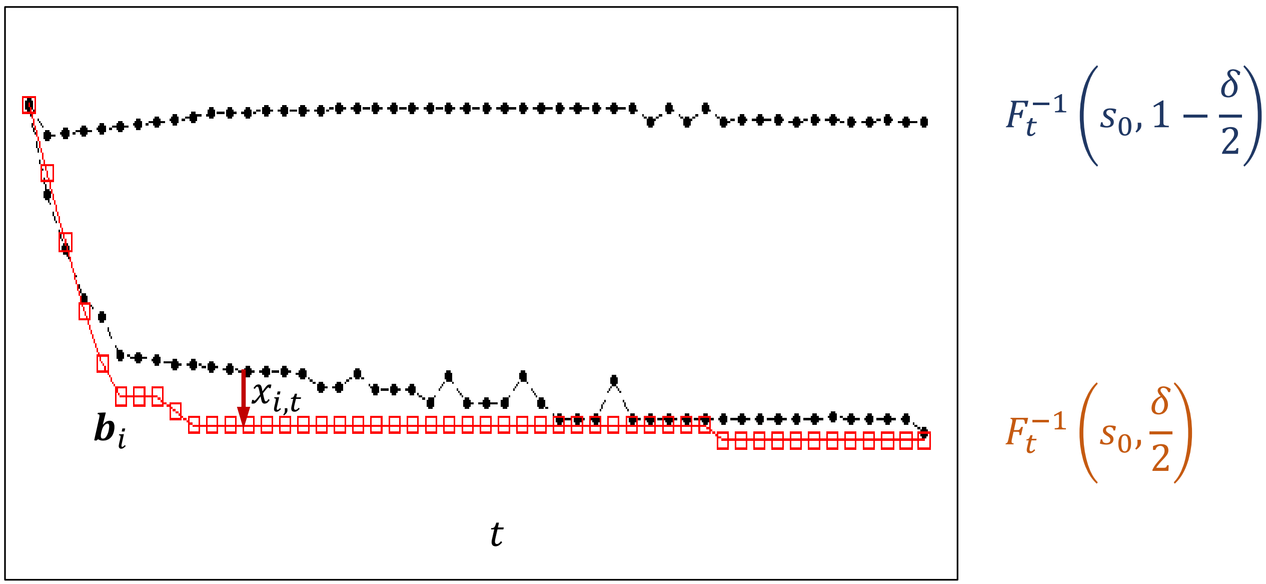
Conformal Guarantees

- Romano, Patterson & Candes (NeurIPS 2019) Conformalized Quantile Regression
- Idea: Compute the “error” between the observed values y_i and the predicted quantile $F^{-1}(x_i; q)$ and conformalize to get a “correction”
- Two data sets:
  - D_1: used for quantile regression $F^{-1}(x; q)$
  - D_2: used for conformalization
- For $(x_i, y_i) \in D_2; \quad i = 1, \ldots, n$
  - $c_i \coloneqq y_i - F^{-1}(x_i; q)$
- Sort to obtain $c_{(1)}, \ldots, c_{(n)}$
- Bound: $hi(x) \coloneqq F^{-1}(x; q) + c_{(\lceil(1-\delta)(n+1)\rceil)}$
- Let (x_{n+1}, y_{n+1}) be a new data point
  - $c_{n+1} := y_{n+1} - F^{-1}(x_{n+1}, q)$
- Claim: The c_i values are exchangeable ➔ rank of c_{n+1} will be uniformly distributed in $c_{(1)}, \ldots, c_{(n+1)}$
- Therefore, $P[y_{n+1} \le hi(x_{n+1})] \ge 1 - \delta$



Conformal Guarantees in $\boldsymbol{h}$ dimensions: Compute "exceedances" for each $\boldsymbol{b}_i$

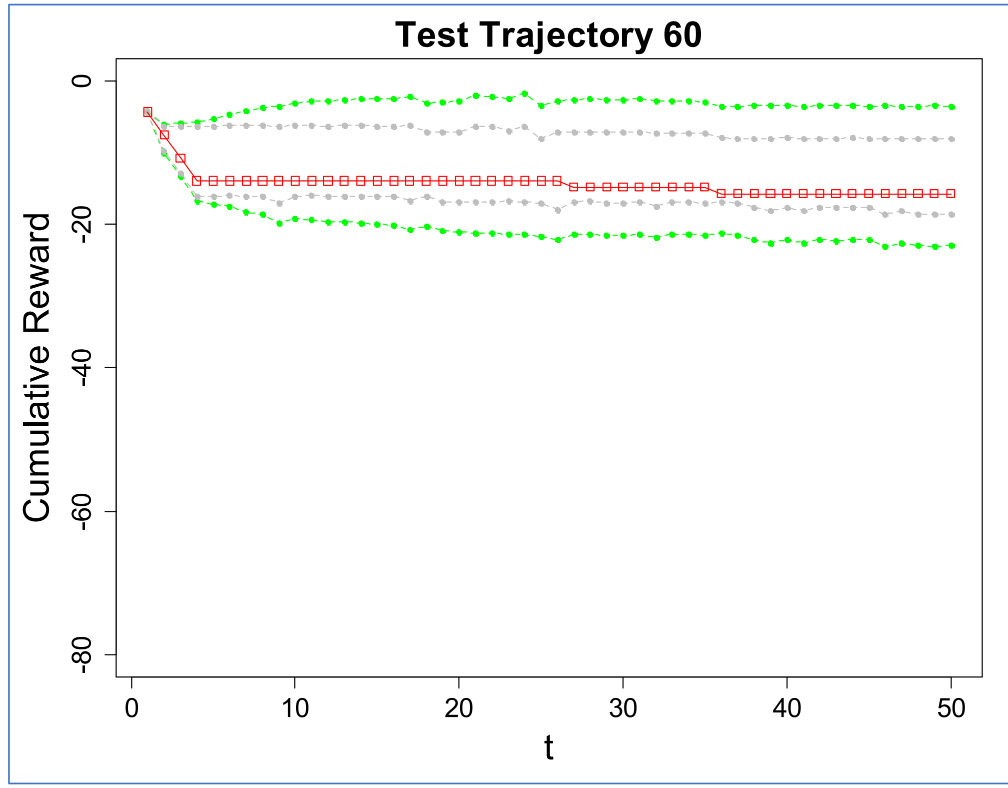
- $x_{i,t} = \max\left(0,\ F_t^{-1}\left(s_0(\tau_i), \frac{\delta}{2}\right) - b_{i,t}, b_{i,t} - F_t^{-1}\left(s_0(\tau), 1 - \frac{\delta}{2}\right)\right)$



Conformalized Quantile Regression: SCALEDSDTRAJECTORY

- Compute $\hat{\sigma}_t$ of the exceedances $x_{\cdot,t}$ at time t using small additional data set
- Rescale exceedances: $x'_{i,t} \coloneqq x_{i,t}/\hat{\sigma}_t$
- Compute c_i for each trajectory in calibration data set
  - $c_i \coloneqq \max_t x'_{i,t}$
- Compute order statistics $c_{(1)}, \ldots, c_{(n)}$
- $\beta \coloneqq c_{(\lceil(1-\delta)(n+1)\rceil)}$

$$lo_t(s_0(\tau)) \coloneqq F_t^{-1}(s_0(\tau), \delta/2) - \beta\hat{\sigma}_t$$
$$hi_t(s_0(\tau)) \coloneqq F_t^{-1}(s_0(\tau), 1-\delta/2) + \beta\hat{\sigma}_t$$

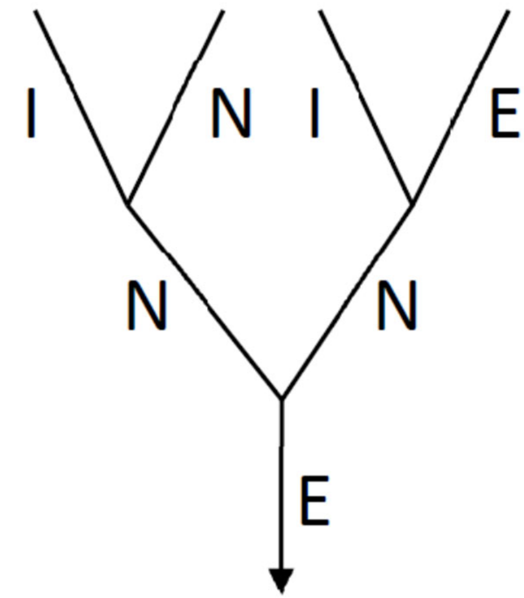


Theorem. The behavior vector $\mathbf{b}^*(\tau^*)$ will fall within the prediction interval $\left[\mathbf{lo}\left(s_0(\tau^*)\right), \mathbf{hi}\left(s_0(\tau^*)\right)\right]$ with probability $1 - \delta$, where the probability is over the choice of random starting states $s_0 \sim P_0$ and any randomness in the policy and MDP dynamics.

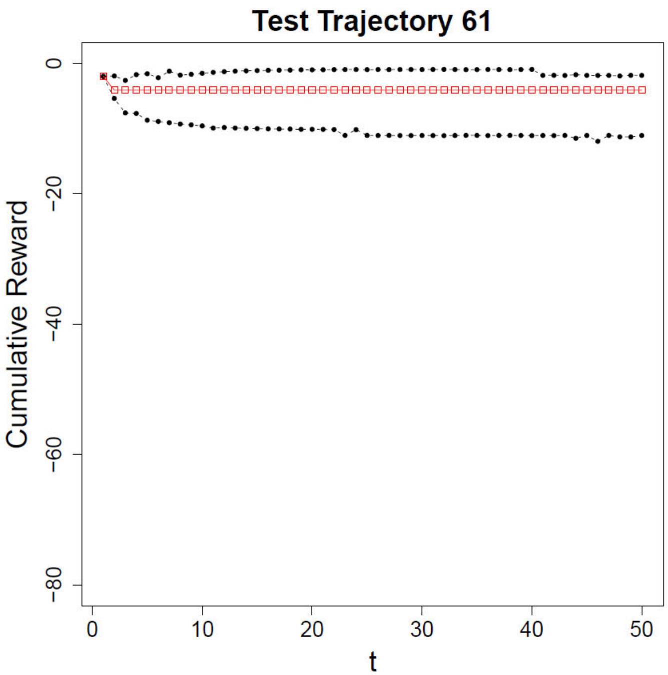
See also: Lei, Rinaldo & Wasserman (2013). Related result for general functional data

Application:
Tamarisk Invasions in River Networks

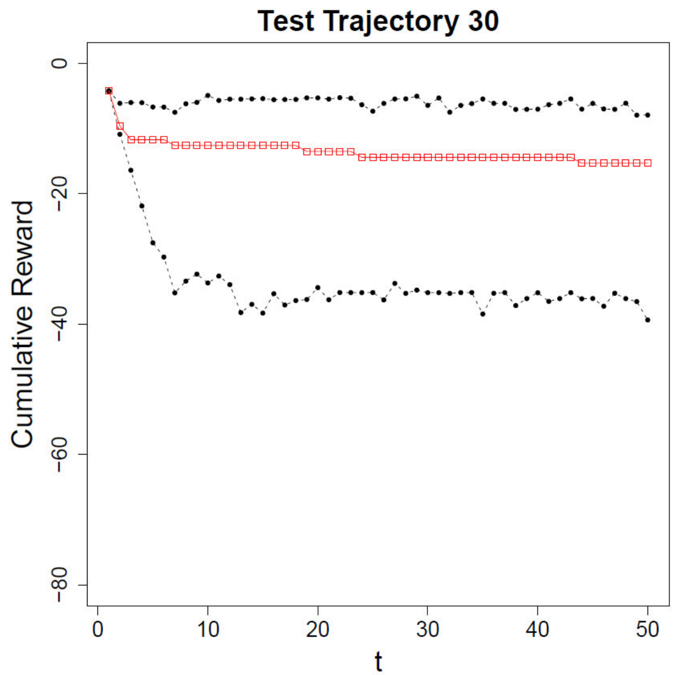
- States:
  - 7 edge river network
  - edge can be
    - I: invaded with tamarisk tree
    - N: occupied by native tree
    - E: empty
- Actions:
  - Plant native
  - Eradicate tamarisk
  - Eradicate + Plant
  - No-Op
- Budget restricts us to one action on one edge per time step
- See Hall, Albers, Alkaee-Taleghan, Dietterich (2018)



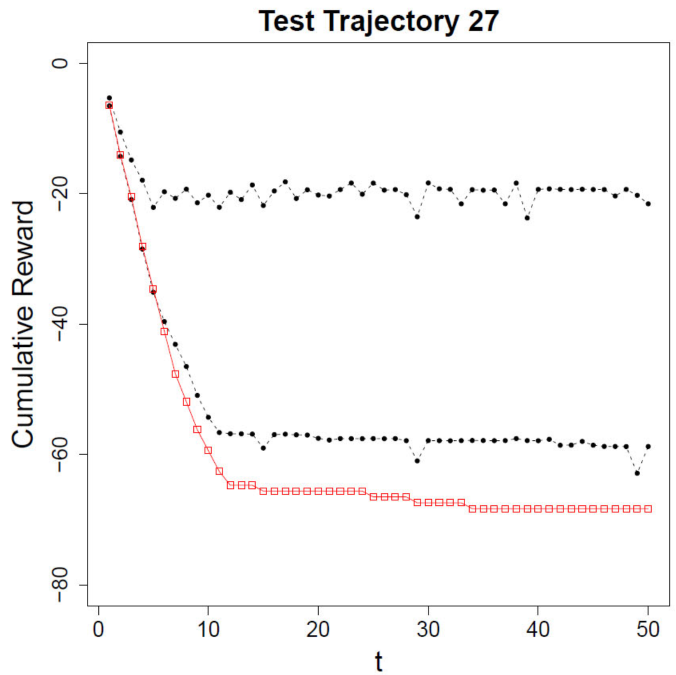
Example Prospective Intervals and Actual Trajectories



$s_0 =$ EEENENI



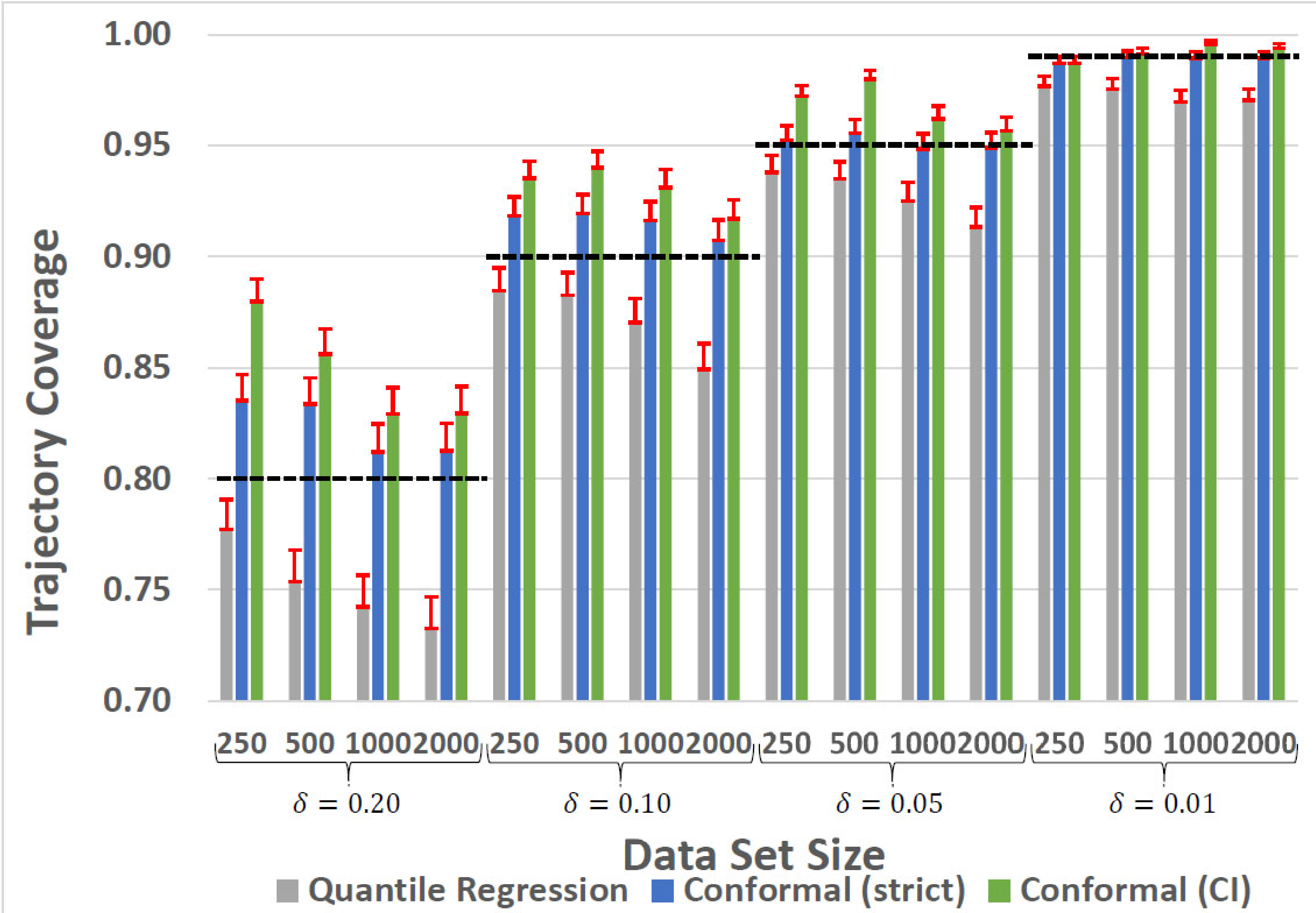
$s_0 =$ EIEINIE



$s_0 =$ IEIIIEI

Tamarisk Prediction Interval Coverage

Raw QR: 0/16
Strict: 16/16
CI: 16/16

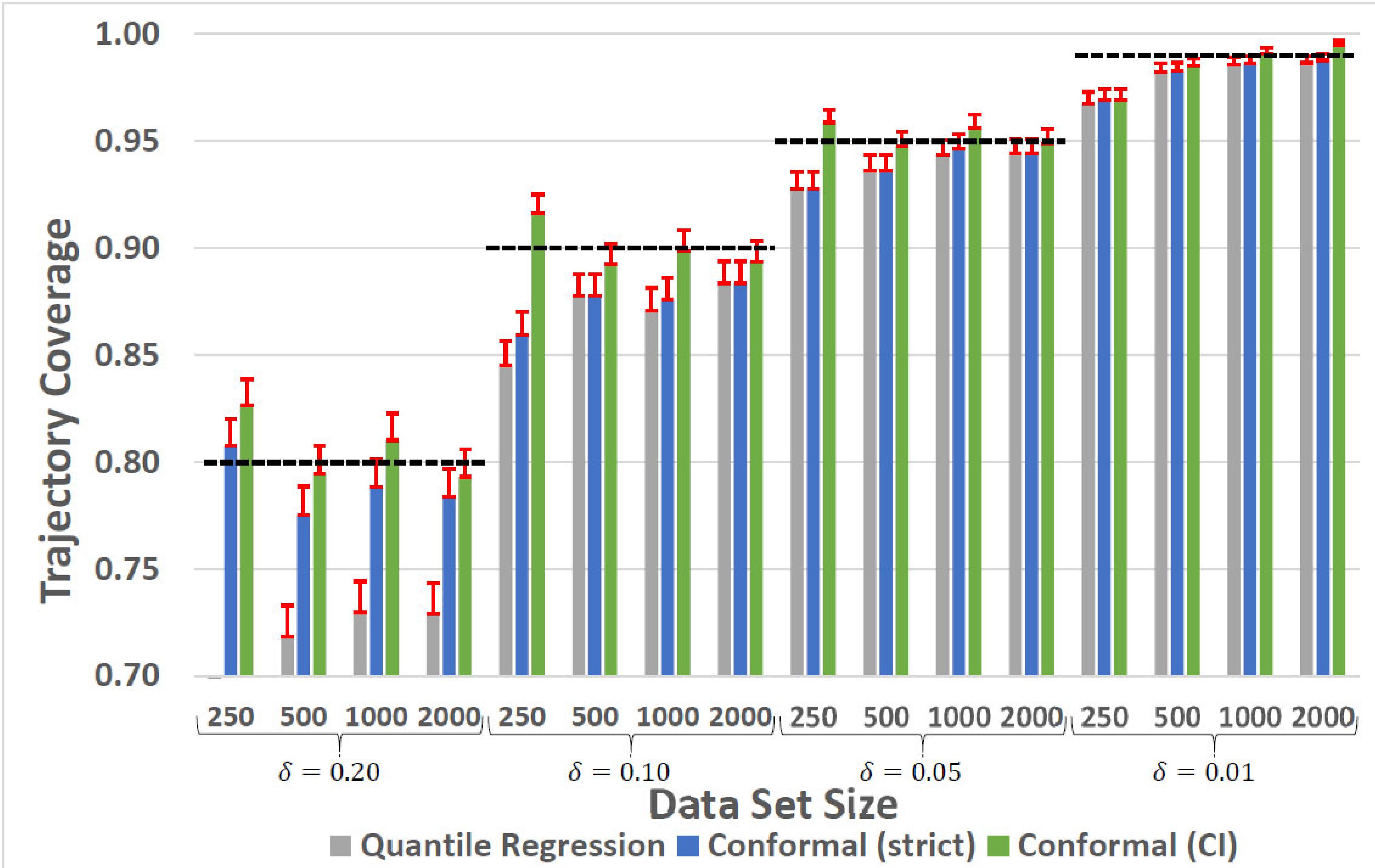


MDP 2: Starcraft Battles

- Reinforcement Learning Scenario
  - StarCraft battle
  - Red forces will be receiving an unknown number of reinforcements at $t = 14$
  - Blue forces receive rewards for winning the battle and for destroying Red units; negative rewards for losing Blue units
  - Value of the starting state is the sum of future rewards
- Goal: Provide probabilistic guarantee on the total Blue Team reward

Starcraft Prediction Interval Coverage

Raw QR: 2/16
Strict: 5/16
CI: 14/16



Careful Interpretation of Prediction Intervals

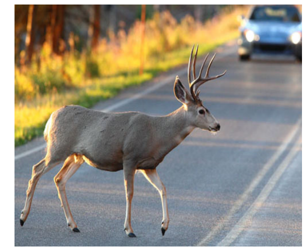
- The 80% guarantee says that over all queries x_q drawn from the same distribution as the training trajectories, 80% of the time, the true r_q will lie within the prediction interval
- It is not a pointwise guarantee
- Theorem: A pointwise guarantee is impossible
  - Barber, Candès, Ramdas, Tibshirani (arXiv 1903.04684)

Part 2: Runtime Open Category Detection

[Liu, Garrepalli, D, Fern: ICML 2018]

- Training data $\{(x_i, y_i)\}$ for $y_i \in \{1, \dots, K\}$ known categories
- Test data $\{(x_j, y_j)\}$ for $y_j \in \{1, \dots, K, K+1, \dots, K+U\}$ with U unknown classes
- ML system should detect the queries that belong to novel categories

Known Classes

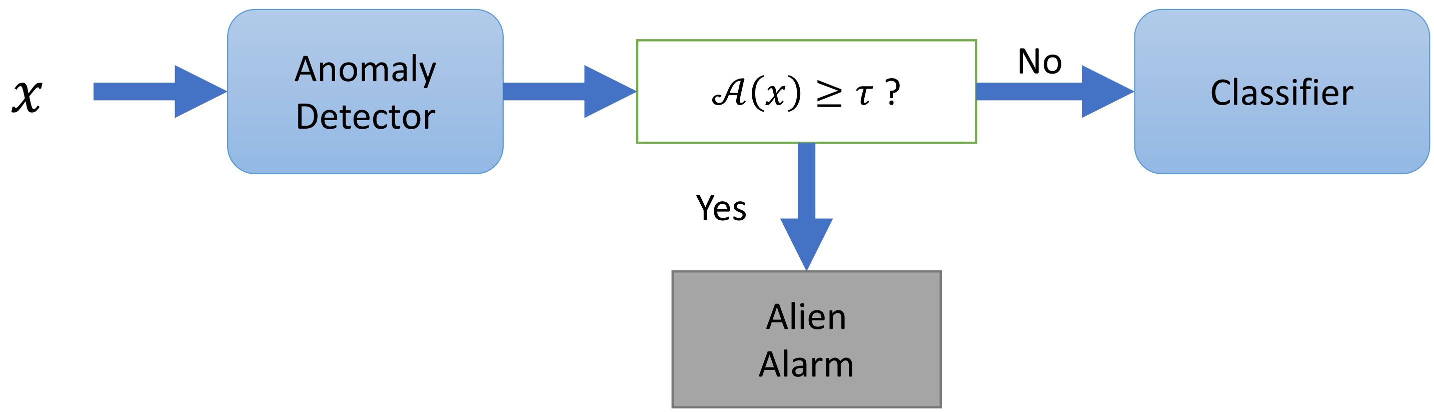


Technology

Volvo's driverless cars 'confused' by kangaroos

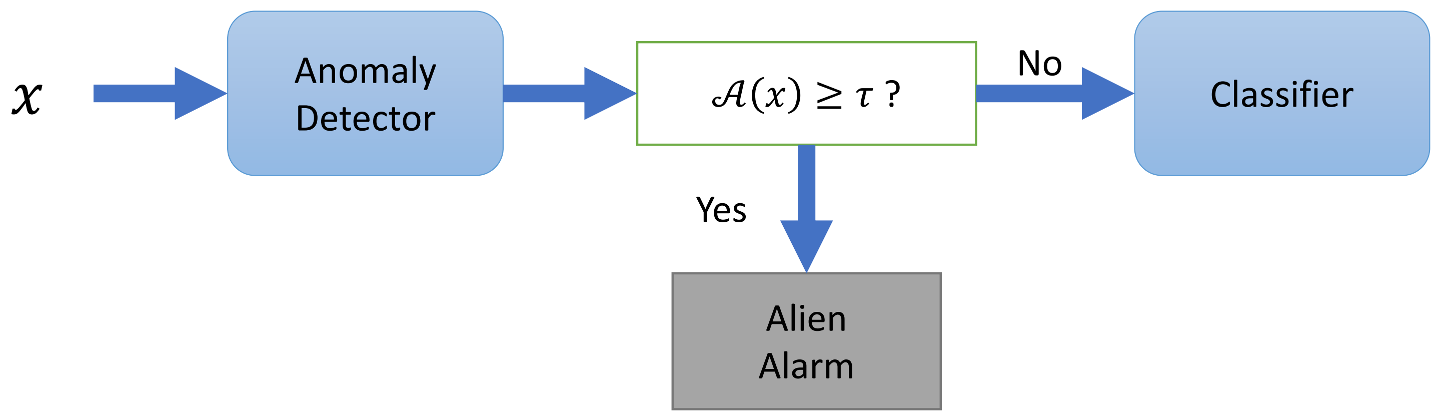
Novel Classes

Method:
Reject Aliens Using Anomaly Detection



We will assume that a (good) anomaly detector $\mathcal{A}$ has been trained

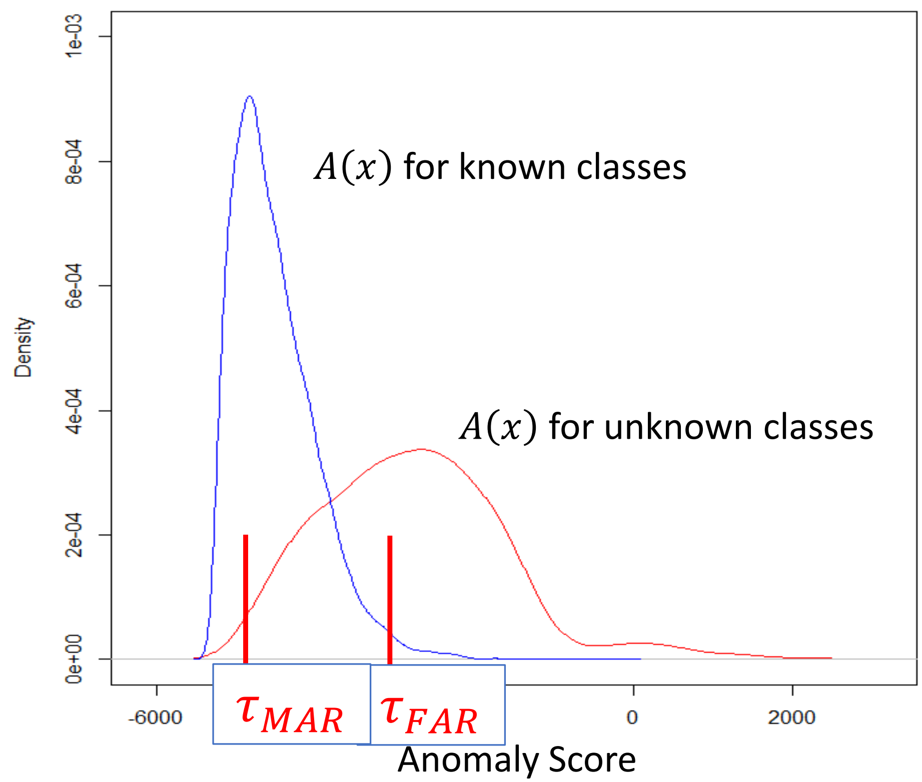
Question:
How to set τ without labeled data?



Setting τ to control false alarms / missed alarms

- To achieve False Alarm Rate of η, set τ to the $1-\eta$ quantile of the $A(x)$ distribution for known classes
- Is there a way to control the Missed Alarm Rate to be no more than η? We need to estimate the η quantile of the $\mathcal{A}(x)$ distribution for the unknown classes
- We have no labeled data for the unknown classes, that is why they are unknown!

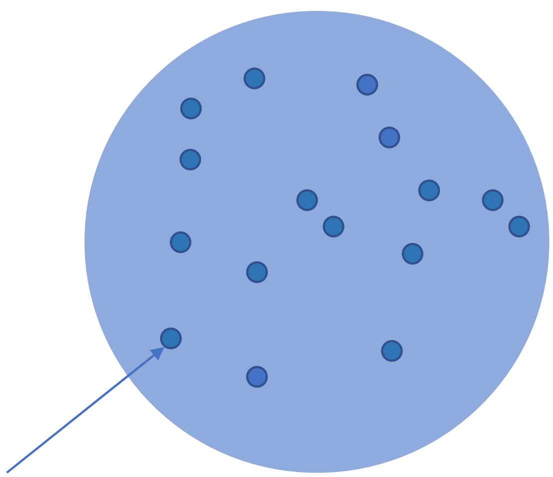
Typical Anomaly Score Plot



Idea: Use Unlabeled Data that Contains Novel Class Examples

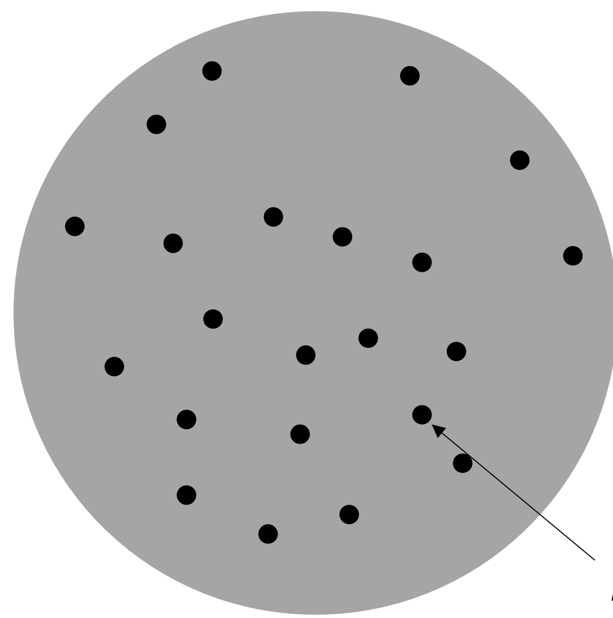
Nominal Distribution

$$D_0$$



Mixture Distribution

$$D_m = (1-\alpha)D_0 + \alpha D_a$$



Where
D_α = Alien Distribution
α = Proportion of Aliens

Notation:

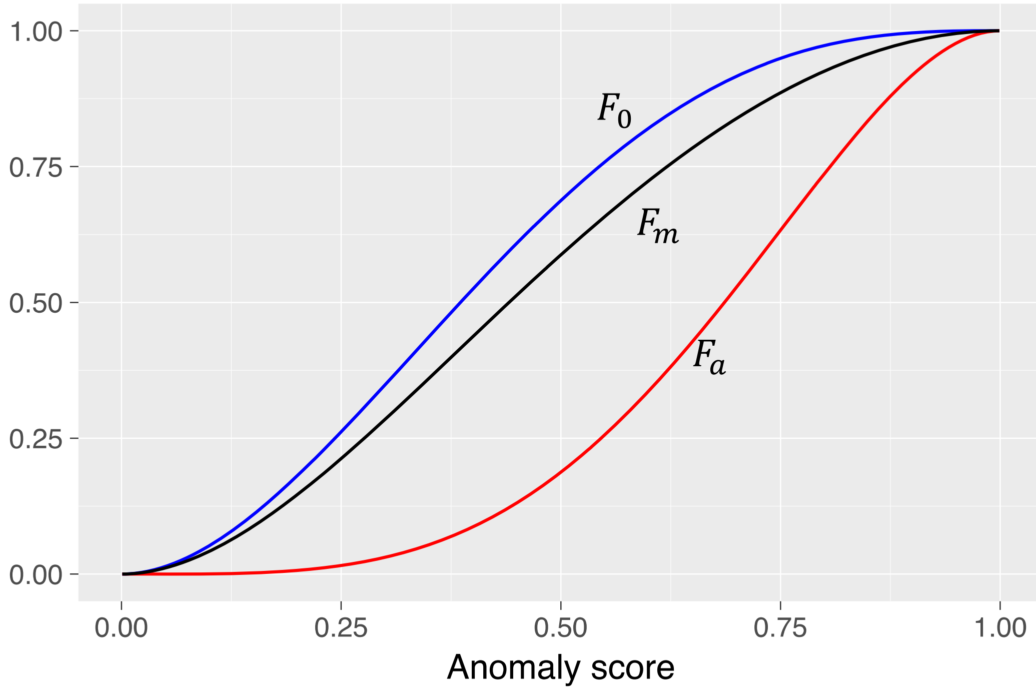
$$\text{Let } F_0(x) = \text{CDF of } \mathcal{A}(D_0)$$
$$F_m(x) = \text{CDF of } \mathcal{A}(D_m)$$
$$F_a(x) = \text{CDF of } \mathcal{A}(D_a)$$

$$D_m = (1-\alpha)D_0 + \alpha D_a$$

implies that

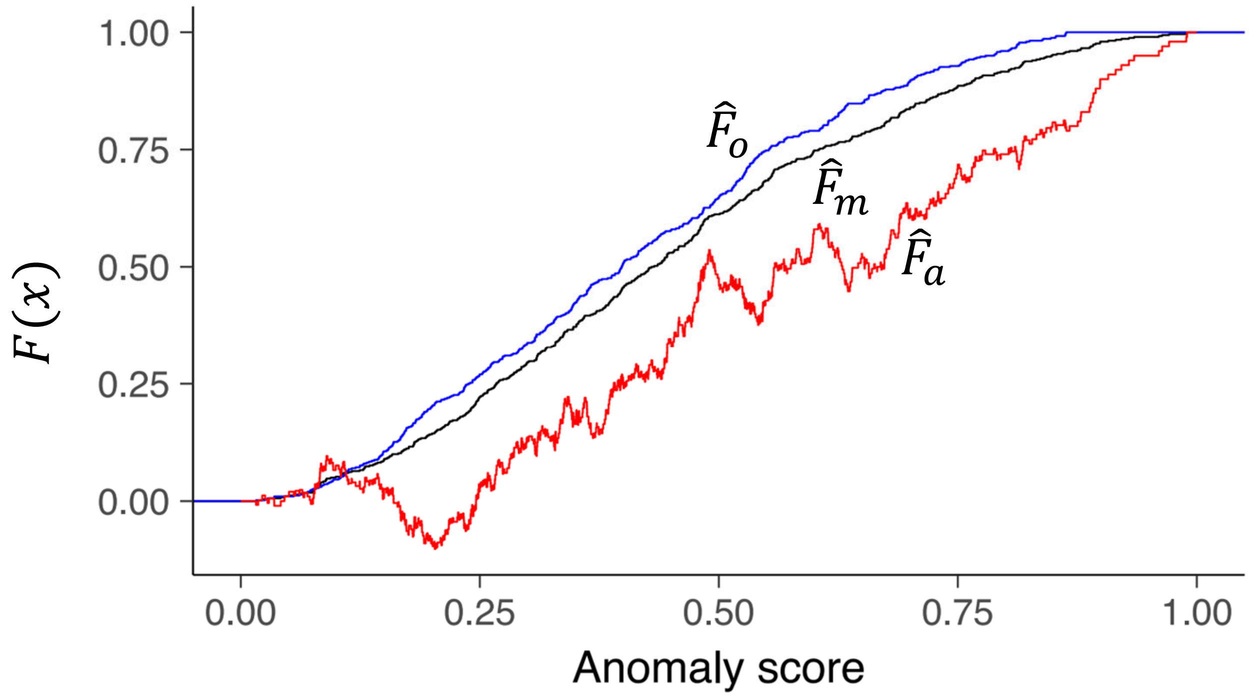
$$F_m(x) = (1-\alpha)F_0(x) + \alpha F_a(x)$$

CDFs of Nominal, Mixture, and Alien Anomaly Scores



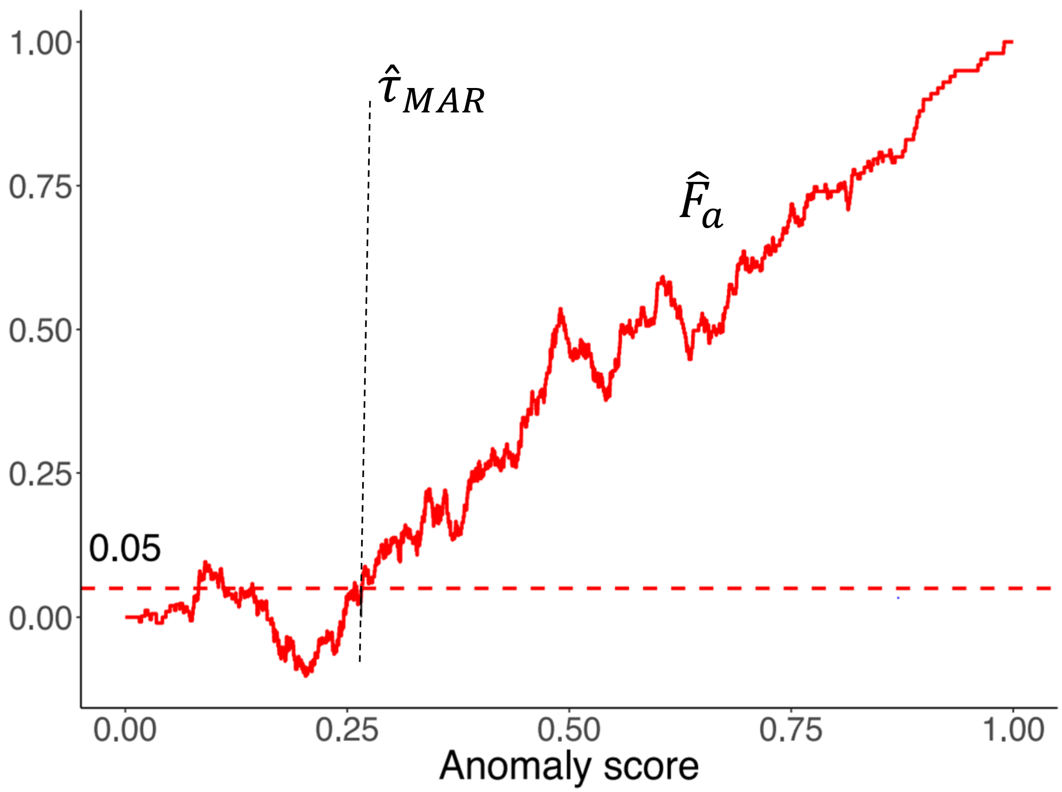
$$F_a(x) = \frac{F_m(x) - (1-\alpha)F_0(x)}{\alpha}$$

We Only Have The Empirical CDFs



$$\hat{F}_a(x) = \frac{\hat{F}_m(x) - (1-\alpha)\hat{F}_0(x)}{\alpha}$$

Choosing the estimate $\hat{\tau}_{MAR}$



EstimateTau(S_0, S_m, MAR, α):

- Anomaly scores of S_0: $x_1, x_2, \cdots, x_k$
- Anomaly scores of S_m: $y_1, y_2, \cdots, y_m$

$$\hat{\boldsymbol{\tau}}_{MAR} = \textbf{max}\{\boldsymbol{u} \in \mathcal{A}(\boldsymbol{S}) : \widehat{\boldsymbol{F}}_{\boldsymbol{a}}(\boldsymbol{u}) \leq \boldsymbol{MAR}\},$$

where

$$S = \{x_1, x_2, \cdots, x_k, y_1, y_2, \cdots, y_m\}.$$

Theorem 1 (Finite Sample Guarantee)

Algorithm 1 will return a threshold $\hat{\tau}_q$ that achieves an alien detection rate of at least $1 - (MAR + \epsilon)$ with probability $1 - \delta$

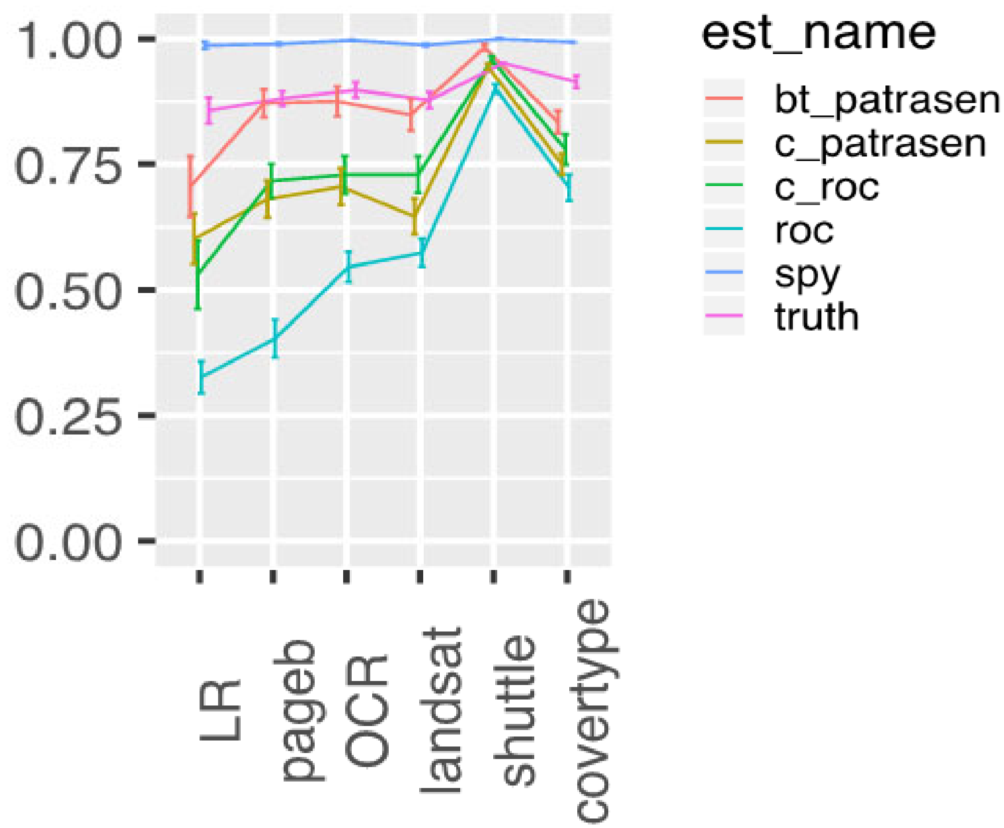
$$n \geq \frac{1}{2} \ln \frac{2}{1 - \sqrt{1-\delta}} \left(\frac{1}{\epsilon}\right)^2 \left(\frac{2-\alpha}{\alpha}\right)^2,$$

Assume F_0 and F_a continuous with convex support. $|S_0| = |S_m| = n$
For any ϵ and $\delta \in (0, 1)$.

The data size n required grows in $O(\frac{1}{\epsilon^2\ \alpha^2} \log \frac{1}{\delta})$

Estimating the mixing proportion α

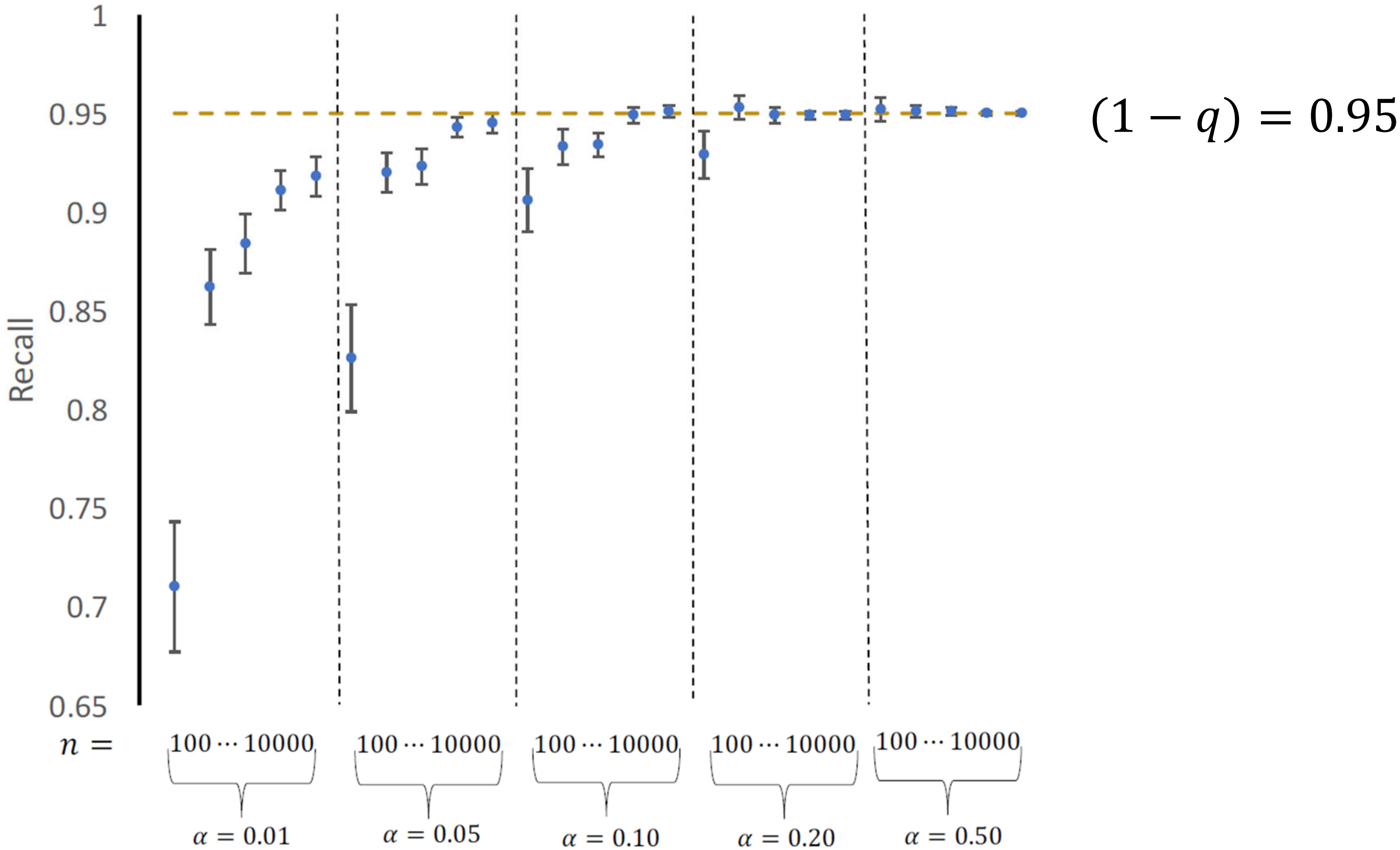
- The mixing proportion is not identifiable in general
- However, under reasonable assumptions, we can obtain an estimate α_0 guaranteed with high probability to be a lower bound on α
- Comparison of five estimators
  - bt_patrasen comes closest to achieve the target recall of 0.95 on six datasets
- Liu, Mondal, Dietterich (under review)



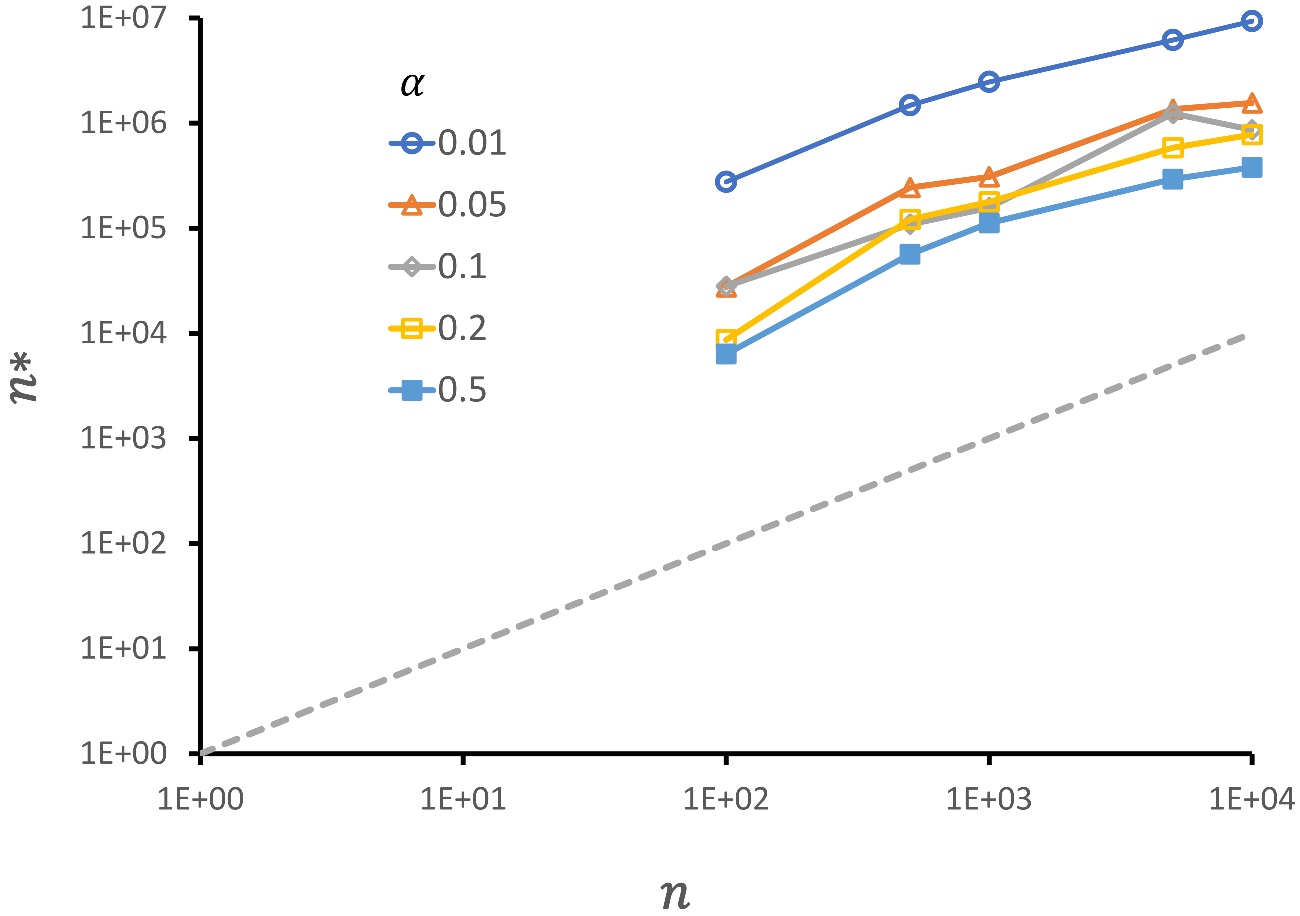
Three Experimental Questions

1. How accurate is our estimate of τ_{MAR}?
2. How loose is the bound on n?
3. How good are Recall and FAR in practice?

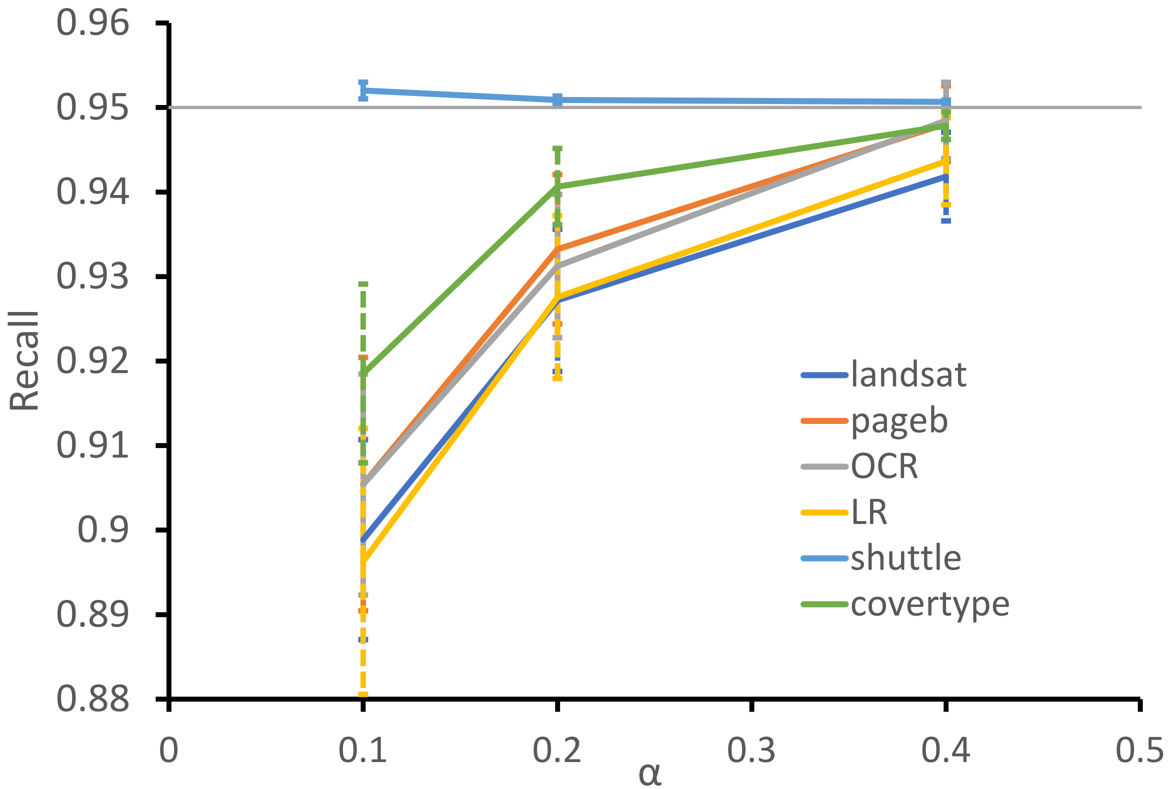
Q1: How accurate is our estimate of τ_q?



Q2: How loose is the bound on n?

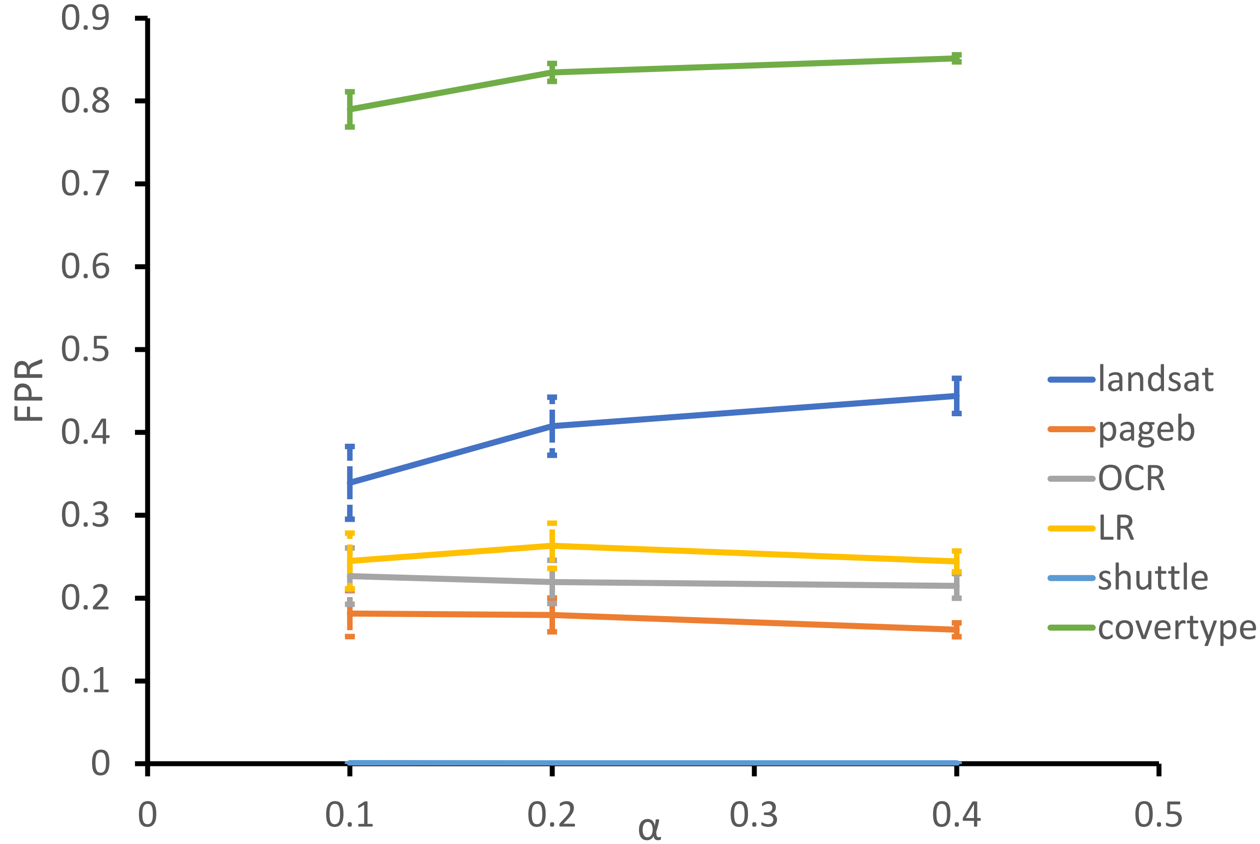


Q3: How good are Recall and FPR in practice? UCI Datasets

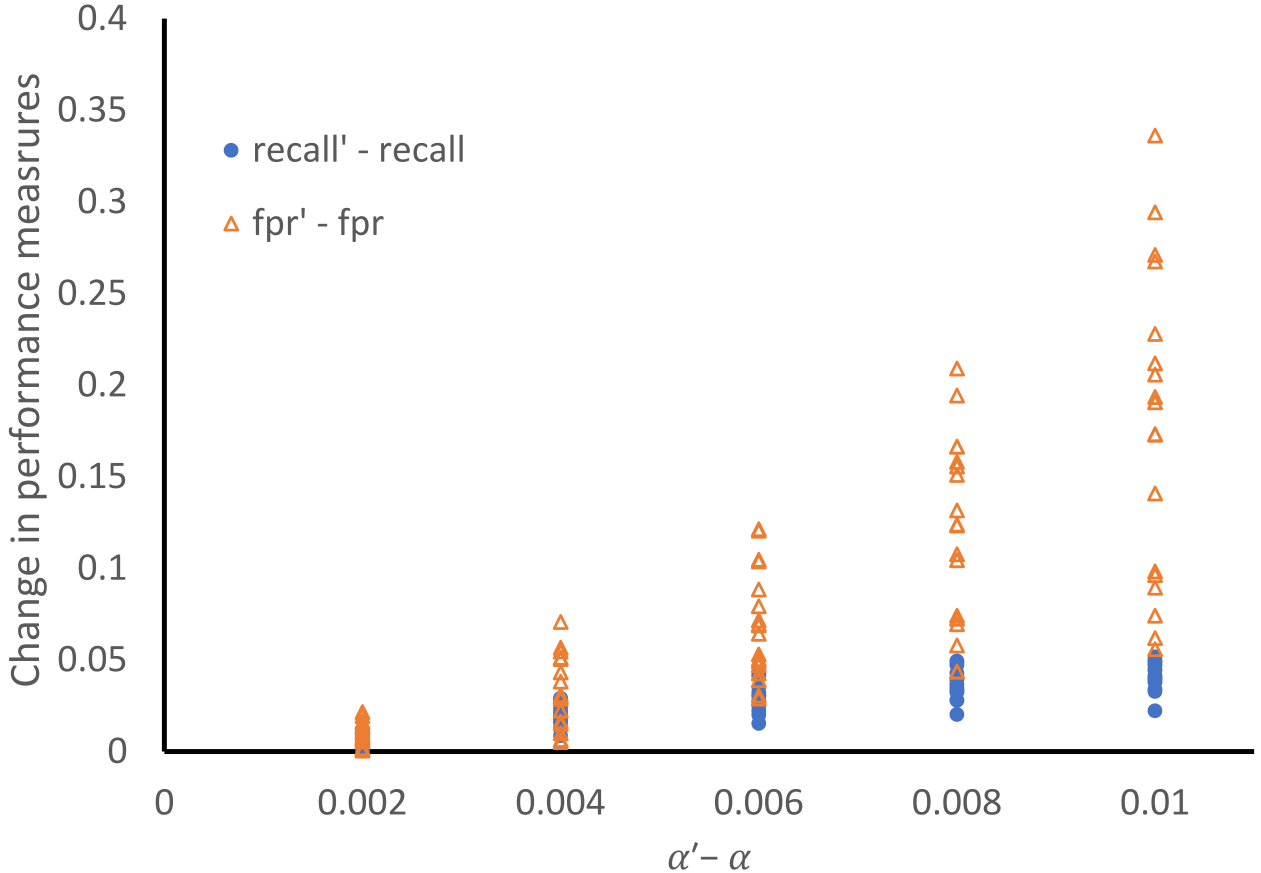


$(1 - q) = 0.95$

Q3: How good are Recall and FPR in practice? UCI Datasets



Q4: What is the impact of using $\alpha' > \alpha$?



Concluding Remarks

- Robust AI and High-Reliability Organizations
  - Competence modeling for HRO teamwork
  - Anomaly Detection
- Competence Modeling
  - Calibrated prediction intervals for reinforcement learning
    - Quantile regression (value function approximation) to predict bounds on reward
    - Conformalization to obtain tight probabilistic guarantees
- Anomaly Detection
  - Open category detection with guarantees
    - Theoretical guarantees on missed alarm rate for novel-class queries
    - Practical algorithms for estimating novelty proportion and setting alarm threshold

Acknowledgments

- National Science Foundation
- DARPA
- Gift from Huawei, Inc.

- Thank you to Kiri Wagstaff (and anonymous reviewers) for feedback on the prediction intervals work

Bibliography

- Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *ArXiv*, *1903.04684*, 1–34. http://arxiv.org/abs/1903.04684
- Dietterich, T. G. (2019). Robust artificial intelligence and robust human organizations. *Frontiers in Computer Science*, *13*(1), 1–3.
- Dietterich, T. G. (2018). Robust artificial intelligence and robust human organizations. https://arXiv.org/abs/1811.10840
- Hall, K. M., Albers, H. J., Alkaee Taleghan, M., & Dietterich, T. G. (2018). Optimal Spatial-Dynamic Management of Stochastic Species Invasions. Environmental and Resource Economics, 70(2), 403–427. https://doi.org/10.1007/s10640-017-0127-6
- Lei, J., Rinaldo, A., & Wasserman, L. (2013). A Conformal Prediction Approach to Explore Functional Data. Annals of Mathematics and Artificial Intelligence, 74(1), 23–43. http://arxiv.org/abs/1302.6452
- Liu, S., Garrepalli, R., Dietterich, T. G., Fern, A., & Hendrycks, D. (2018). Open Category Detection with PAC Guarantees. Proceedings of the 35th International Conference on Machine Learning, PMLR, 80, 3169–3178.

Bibliography (2)

- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.
- Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized Quantile Regression. http://arxiv.org/abs/1905.03222; NeurIPS 2019
- Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko, Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft II: A new challenge for reinforcement learning. http://arXiv.org/abs/1708.04782, 2017.
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.
- Weick, K., Sutcliffe, K., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.), *Research in Organizational Behavior* (Vol. 1, pp. 81–123). Jai Press.