Towards Responsible ML Benchmarking

NeurIPS 2021
Datasets & Benchmarks Track

Erin LeDell Ph.D.
@ledell

Agenda

- Benchmarks in Machine Learning
- Benefits & Limitations of Benchmarking
- Benchmark Frameworks
- Responsible Benchmarking

All benchmarks are wrong,
but some are useful.

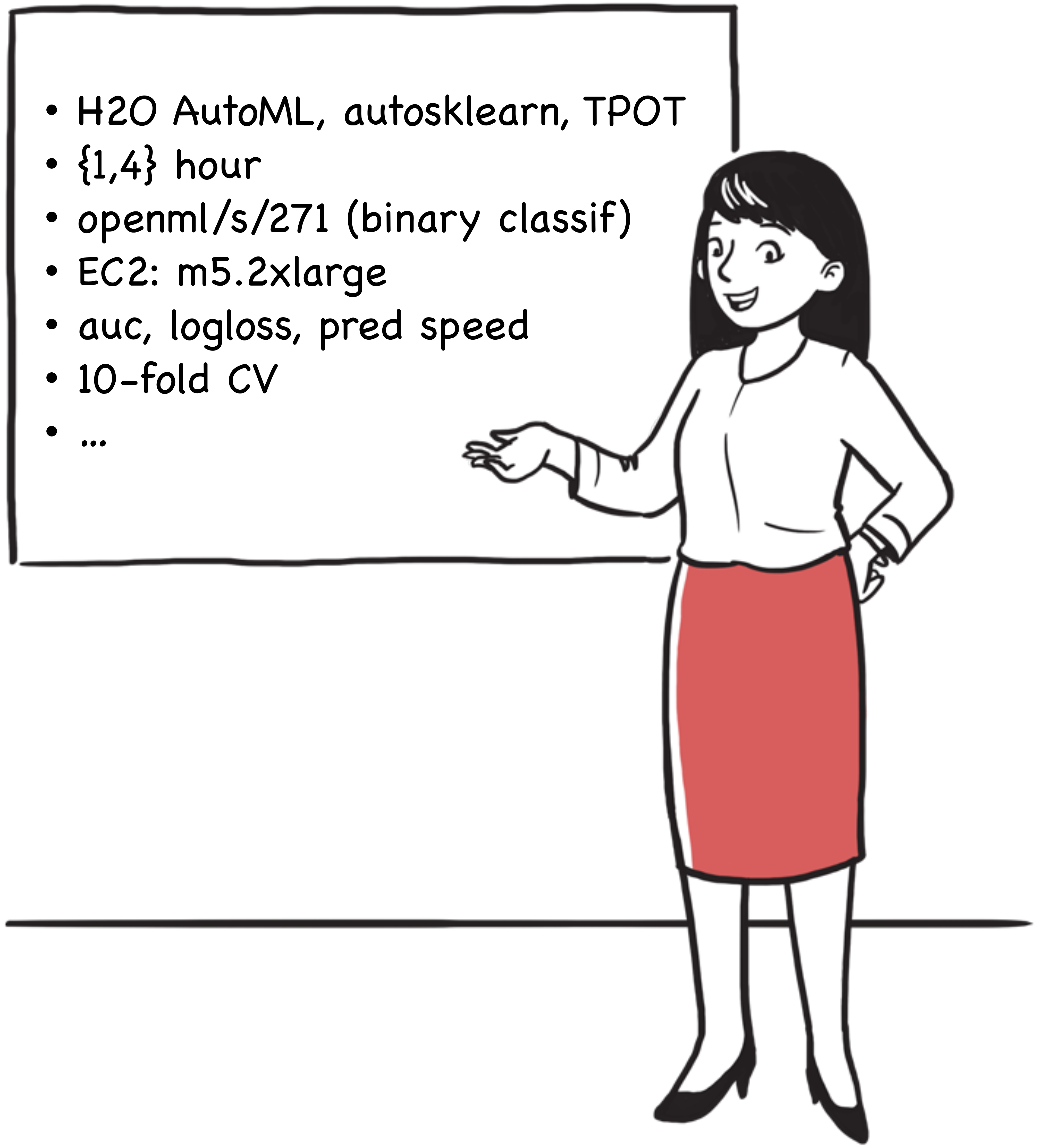
Goals of ML Benchmarking

- 📊 Empirically compare performance of machine learning algorithms, software or hardware
- 🆚 Measure improvement over baselines
- 🔍 Understand the limitations of various methods
- 🔄 Reproducible: Code + Environment
- 🇨🇭 Impartial and objective evaluation

ML/Dataset Benchmarks & Frameworks

- 🖼️ Image: CIFAR-10/100, ImageNet
- 📝 NLP: GLUE, SuperGLUE
- 🗄️ Tabular: OpenML, AMLB, UCI, PMLB, benchm-ml
- 🎛️ HPO: HPOBench, bayesmark, COCO
- 💻 Hardware + Software: DAWNBench, MLPerf, DeepBench
- 🏆 Competitions: Kaggle, ChaLearn, NeurIPS, KDD Cup
- 🤷‍♂️ Ad-hoc: Hand-picked collections of datasets

What constitutes an ML Benchmark

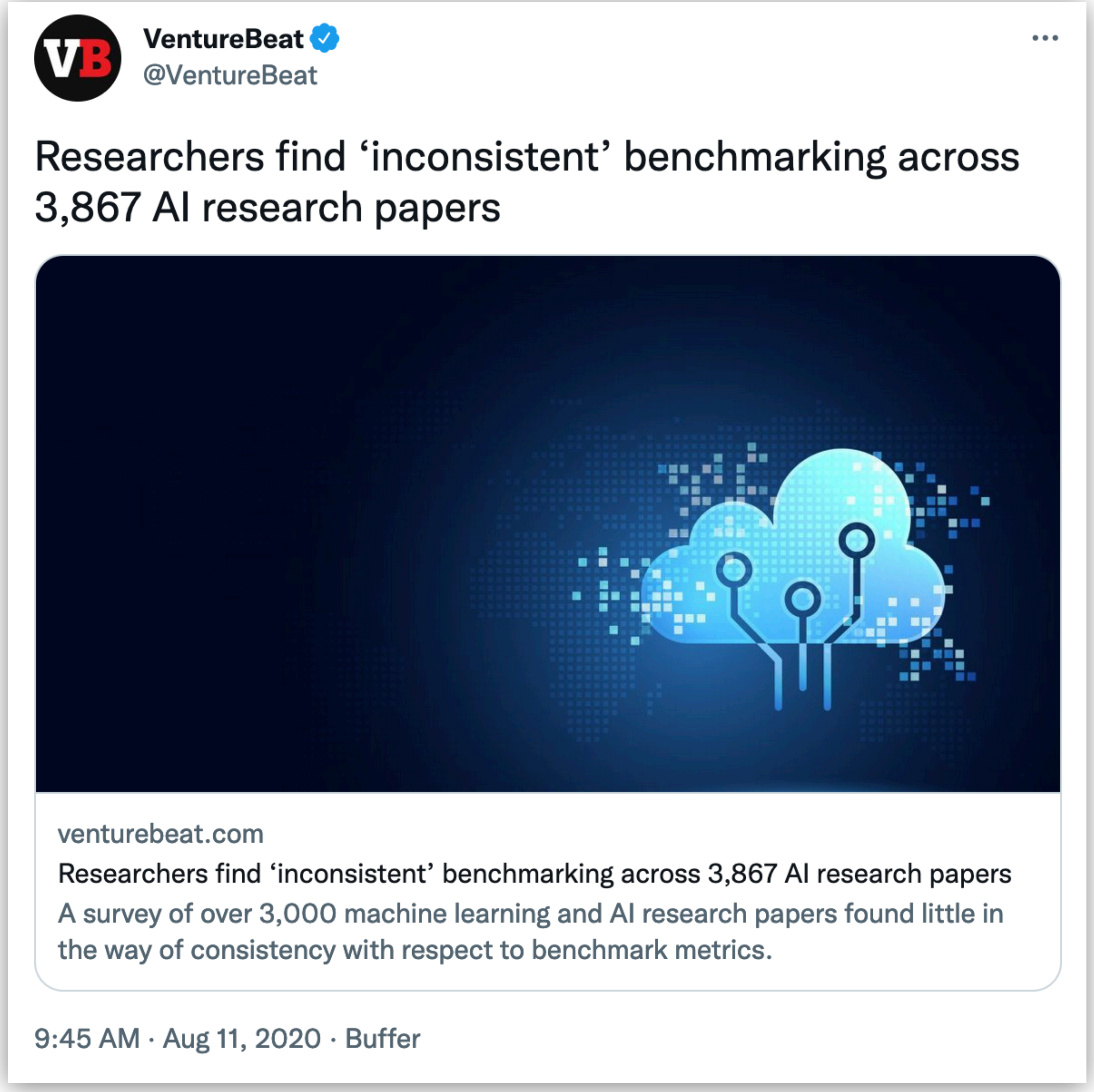


- H2O AutoML, autosklearn, TPOT
- {1,4} hour
- openml/s/271 (binary classif)
- EC2: m5.2xlarge
- auc, logloss, pred speed
- 10-fold CV
- ...

- Set of methods/implementations to be evaluated
- Task being generalized
- Dataset or collection of datasets
- Hardware environment(s)
- Set of evaluation metrics
- Interpretation of results

Limitations

ML Benchmark Limitatations



VentureBeat
@VentureBeat

Researchers find 'inconsistent' benchmarking across 3,867 AI research papers

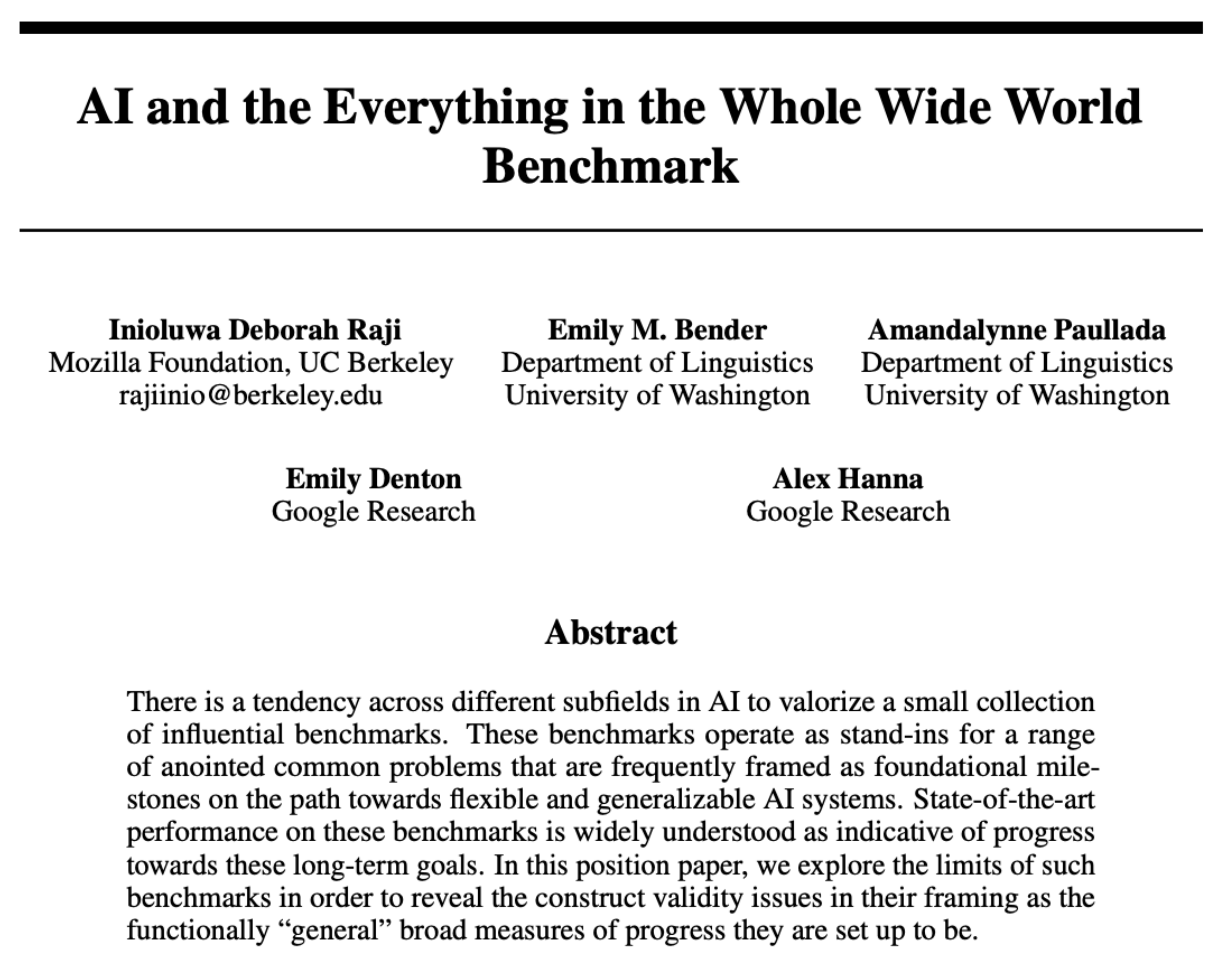
venturebeat.com
Researchers find 'inconsistent' benchmarking across 3,867 AI research papers
A survey of over 3,000 machine learning and AI research papers found little in the way of consistency with respect to benchmark metrics.

9:45 AM · Aug 11, 2020 · Buffer

- 3,867 papers from 'Papers with Code'
- "Accuracy" is top metric
- 77% papers reported only one metric
- 14.4% two top-level metrics
- 6% had three metrics

https://arxiv.org/abs/2008.02577

ML Benchmark Over-generalization



AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation, UC Berkeley
rajiinio@berkeley.edu

Emily M. Bender
Department of Linguistics
University of Washington

Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

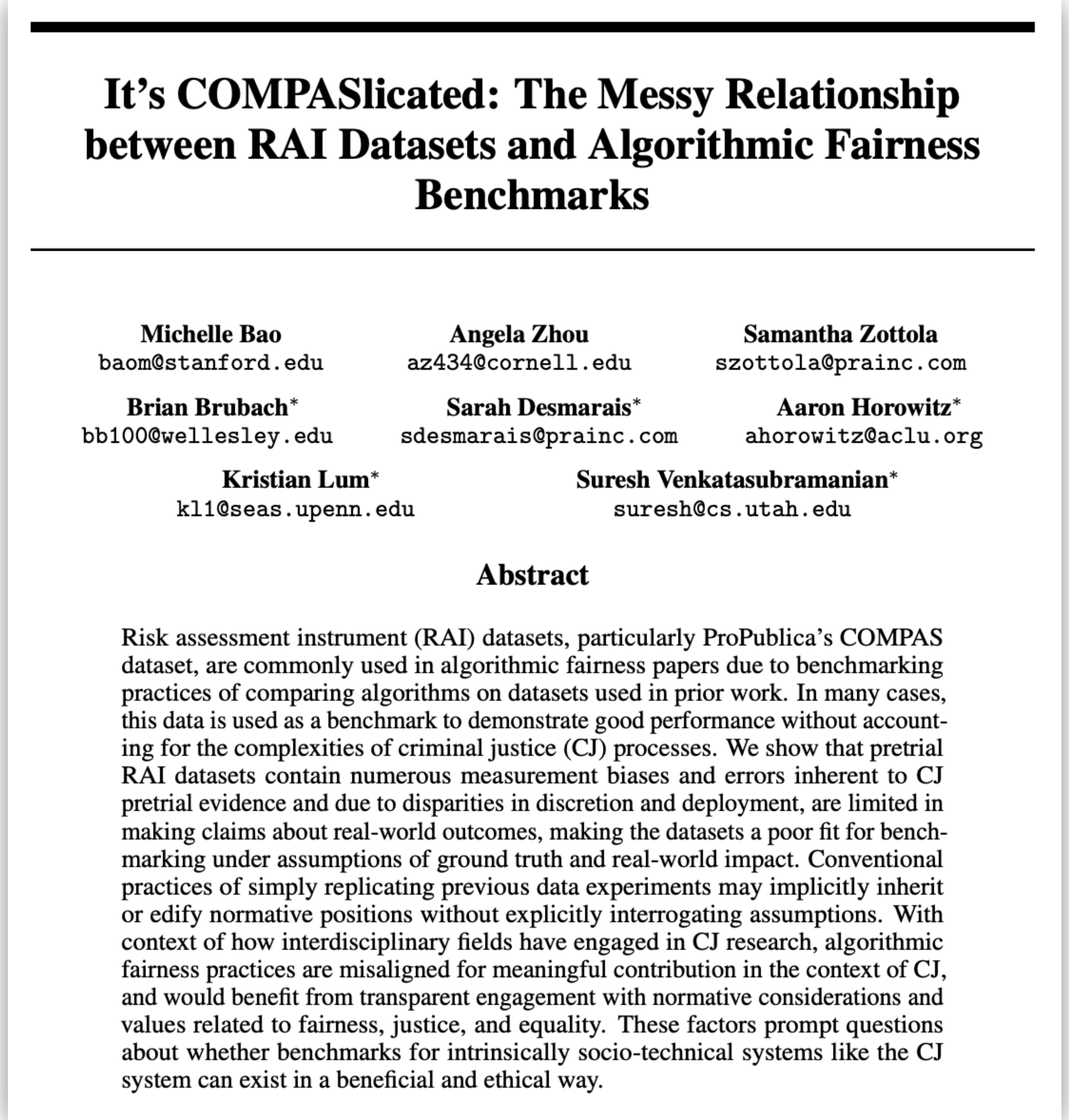
Abstract

There is a tendency across different subfields in AI to valorize a small collection of influential benchmarks. These benchmarks operate as stand-ins for a range of anointed common problems that are frequently framed as foundational milestones on the path towards flexible and generalizable AI systems. State-of-the-art performance on these benchmarks is widely understood as indicative of progress towards these long-term goals. In this position paper, we explore the limits of such benchmarks in order to reveal the construct validity issues in their framing as the functionally "general" broad measures of progress they are set up to be.

Performance on a few datasets is often over-generalized to claim dominance on a vague, broad task like "language understanding" 🦜

https://arxiv.org/abs/2111.15366

ML Benchmark Misalignment



It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks

Michelle Bao baom@stanford.edu

Angela Zhou az434@cornell.edu

Samantha Zottola szottola@prainc.com

Brian Brubach* bb100@wellesley.edu

Sarah Desmarais* sdesmarais@prainc.com

Aaron Horowitz* ahorowitz@aclu.org

Kristian Lum* kl1@seas.upenn.edu

Suresh Venkatasubramanian* suresh@cs.utah.edu

Abstract

Risk assessment instrument (RAI) datasets, particularly ProPublica's COMPAS dataset, are commonly used in algorithmic fairness papers due to benchmarking practices of comparing algorithms on datasets used in prior work. In many cases, this data is used as a benchmark to demonstrate good performance without accounting for the complexities of criminal justice (CJ) processes. We show that pretrial RAI datasets contain numerous measurement biases and errors inherent to CJ pretrial evidence and due to disparities in discretion and deployment, are limited in making claims about real-world outcomes, making the datasets a poor fit for benchmarking under assumptions of ground truth and real-world impact. Conventional practices of simply replicating previous data experiments may implicitly inherit or edify normative positions without explicitly interrogating assumptions. With context of how interdisciplinary fields have engaged in CJ research, algorithmic fairness practices are misaligned for meaningful contribution in the context of CJ, and would benefit from transparent engagement with normative considerations and values related to fairness, justice, and equality. These factors prompt questions about whether benchmarks for intrinsically socio-technical systems like the CJ system can exist in a beneficial and ethical way.

Sometimes benchmarks are incapable of measuring what you're actually trying to measure. ⚖️ 🙅‍♀️ ⛔

https://arxiv.org/abs/2106.05498

ML Benchmarking Mistakes

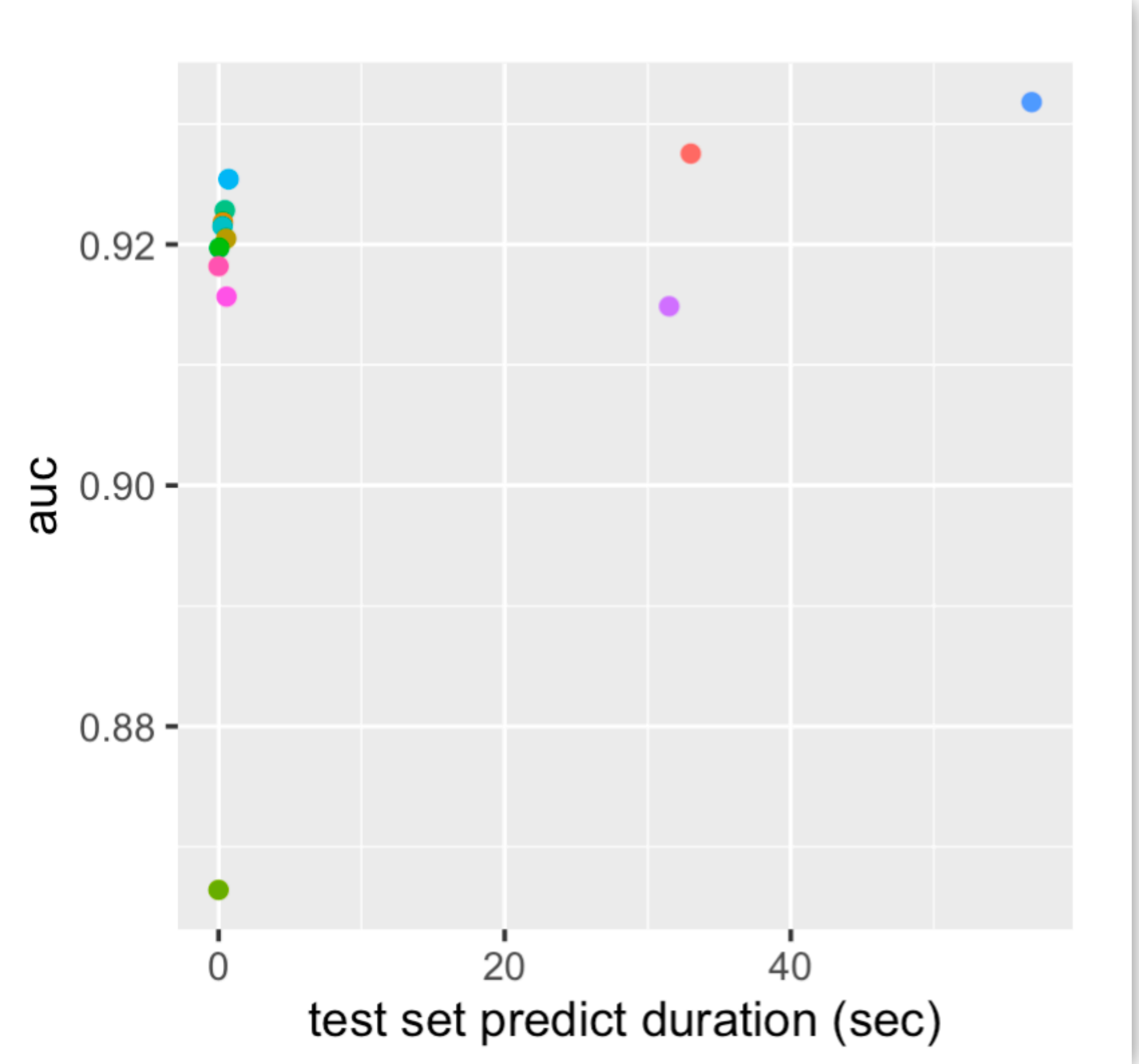
- Not enough datasets, not enough diversity among the datasets and/or the data is biased ❌
- Benchmark has mistakes or is unfair:
  - Authors are experts at using their own tool/method but make mistakes using others ❌
  - Tuning some algorithms more than others ❌
  - Insufficient or inappropriate metrics used ❌
  - Insufficient range of hardware ❌
  - Over-generalization of results ❌

ML Benchmark Harms

- Strongly rewards R&D for methods that overfit to the popular ML benchmarks
- Benchmarks influence dev/methods in the same way that hardware availability does over time (see "Hardware Lottery" paper by Sara Hooker)
- If the benchmark is limited in scope, it encourages methods that optimize what's being measured
- Can lead to development of complex, impractical algorithms/models – The "Kaggle Problem"

The “Kaggle Problem” (SOTA-chasing)

AutoML tools overfit to benchmark



- First AMLB looked only at accuracy metrics, given a fixed runtime
- New, highly complex, AutoML methods were then developed to maximize accuracy at all costs (top right) a la “Kaggle”

Benchmark Frameworks

Benchmarking Framework

- 💻 Software to perform ML benchmarks
- 📦 Should define collections of datasets as well, but this can be decoupled from the framework
- ⚖️ Evaluation metrics & analysis
- ➕ Extensible (new methods, data, metrics)
- 🐳 Containerization is useful for reproducibility
- 🔄 Easy to run on public cloud for reproducibility

Please benchmark
(and review papers)
responsibly!

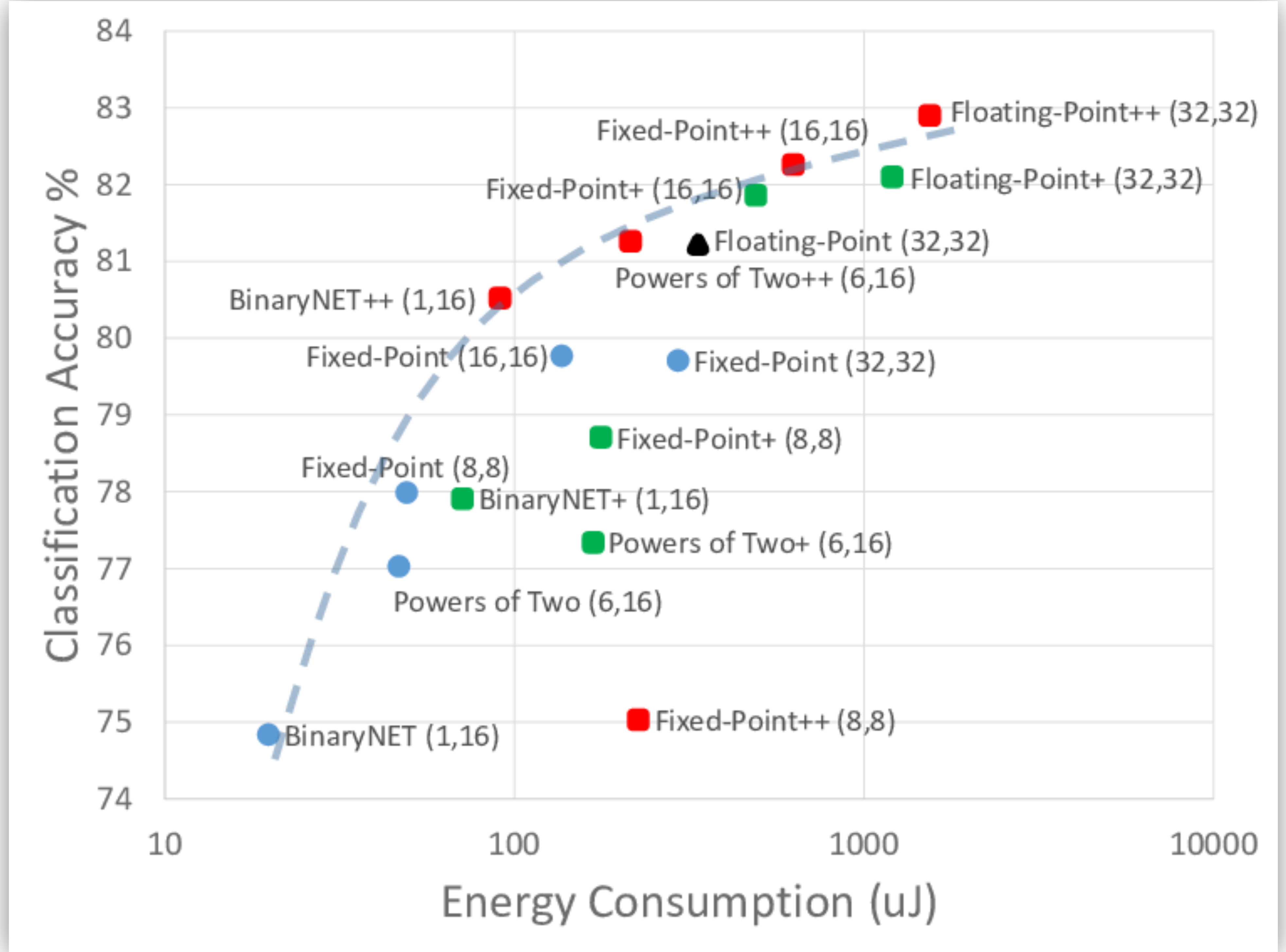
Benchmark Checklist

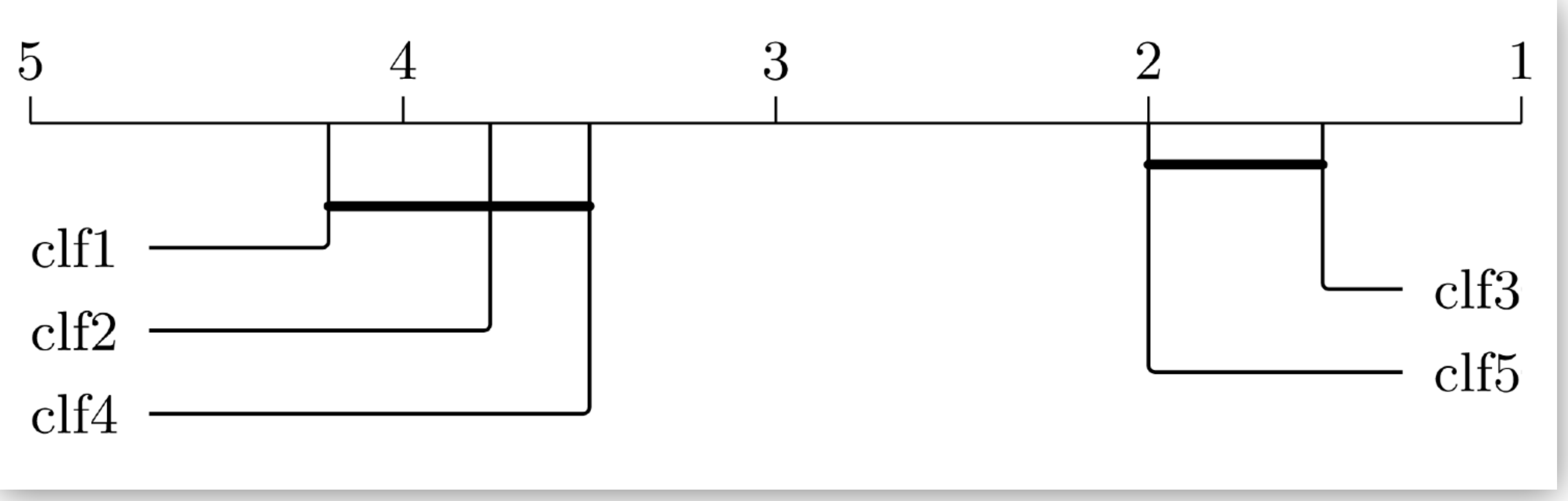
- ✅ Benchmarks as software (code + environment) for full reproducibility and accessibility
- ✅ Measure as much as you can in addition to accuracy (inference speed, cost, energy use, etc.)
- ✅ Consider fairness/bias analysis
- ✅ Invite the public to analyze the results data
- ✅ Invite the public to extend your benchmark
- ✅ Present results in context, do not over-generalize

Benchmark Analysis

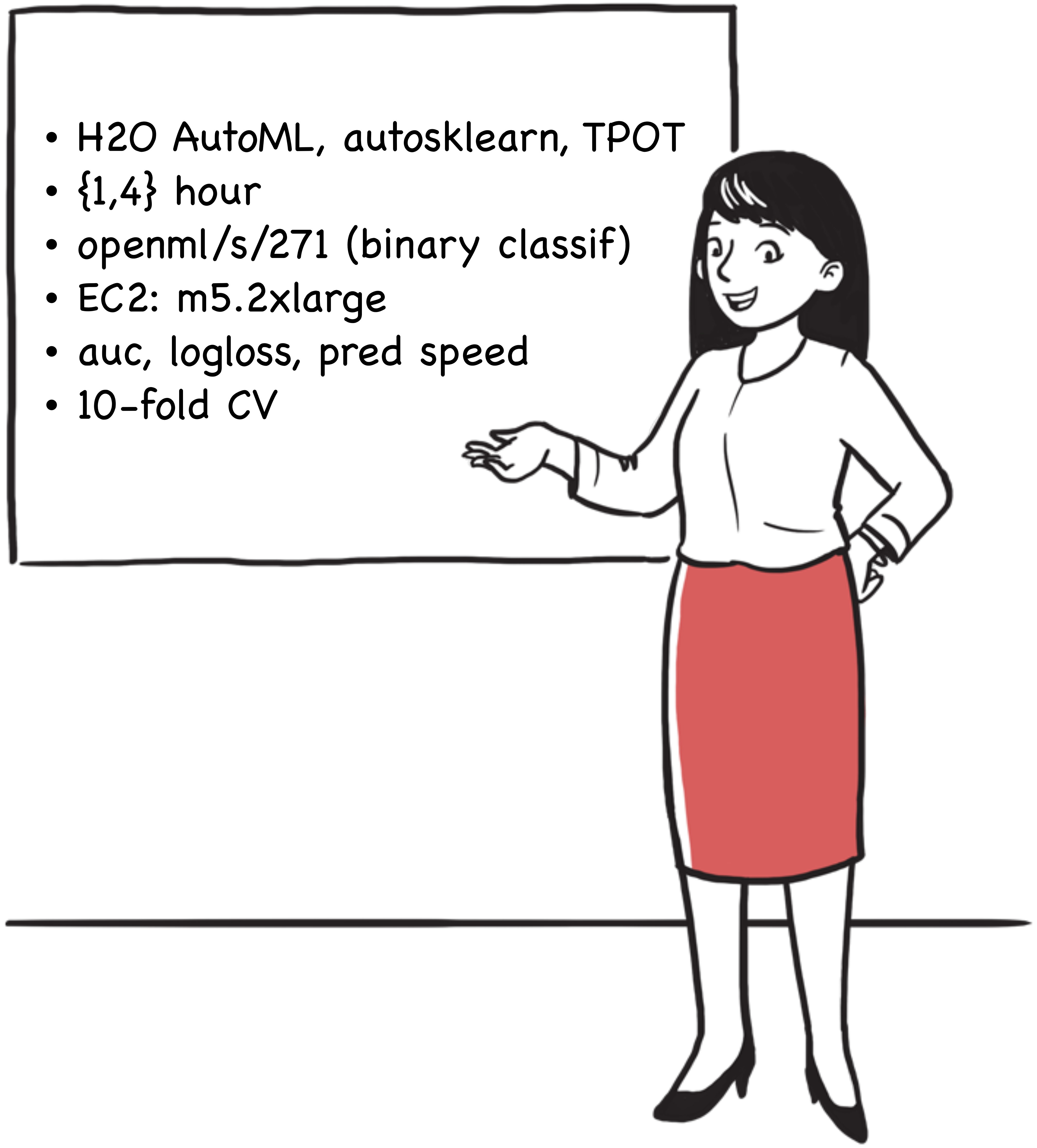
Rather than simply providing a rank of methods based on a single metric, use:

- Pareto frontier
- Critical difference diagrams
- Explainability methods





Benchmark Cards ?



- H2O AutoML, autosklearn, TPOT
- {1,4} hour
- openml/s/271 (binary classif)
- EC2: m5.2xlarge
- auc, logloss, pred speed
- 10-fold CV

- Datasheets for Datasets
- Model Cards
- Benchmark Cards ?
  - Methods {A, …., Z}
  - Task being evaluated
  - Datasets representing that task
  - Software & hardware
  - Evaluation metrics
  - Methods/viz to compare metrics
  - Fairness evaluation

ML Benchmark Re-framing

"Benchmarking, appropriately deployed, is not about winning a contest but more about surveying a landscape—the more we can re-frame, contextualize and appropriately scope these datasets, the more useful they will become as an informative dimension to more impactful algorithmic development and alternative evaluation methods."

AI and the Everything in the Whole Wide World Benchmark

FIN