

The Many Faces of Multiplicity in ML

Prakhar Ganesh, Shomik Jain, Carol Long, Afaf Taik, Hsiang Hsu,
Flavio Calmon, Ashia Wilson, Kathleen Creel, Golnoosh Farnadi

Tutorial Objectives

- Highlighting the phenomenon of multiplicity in machine learning and calling attention to its growing literature.
- Discussing the implications of multiplicity for fairness and explainability in algorithmic decision-making.
- Engaging the FAccT community on when and how to address multiplicity in various practical scenarios.
- Identifying open questions and motivating future research directions on multiplicity.

Several companies are planning to partially automate their entry-level hiring pipeline. They each receive over 100 applications every week, and recruiters don't have time to review every application.

They are planning to train models and create automated hiring tools that will select the strongest applications every week for manual review.

Each of you represent a company!

You will all train your own separate models and make hiring decisions.

Let's train our own models for automated hiring



Source: dilbert.com

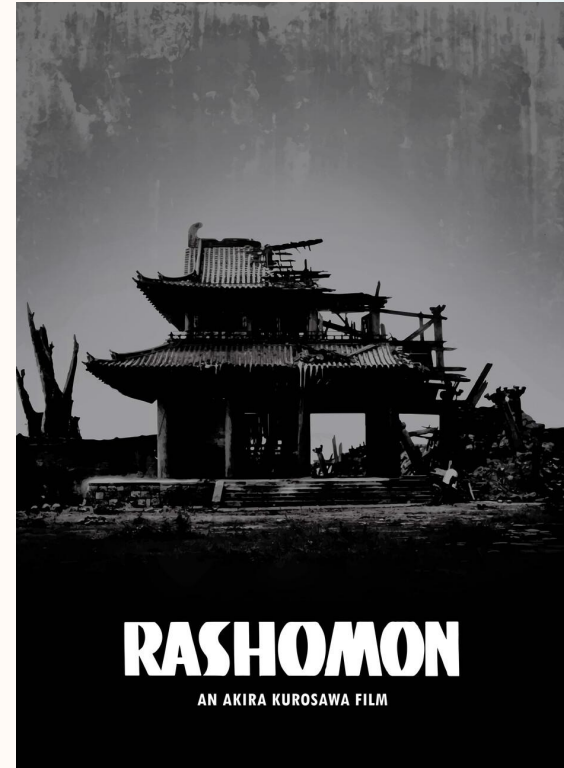


<https://huggingface.co/spaces/prakharg24/multiplicity-demo>

Rashomon Set and Multiplicity

Rashomon Effect

Based on *Rashomon* (1950)
by Akira Kurosawa



Source: Poster by Bo Kev
(<https://fineartamerica.com/featured/rashomon-bo-kev.html>)

Rashomon Effect

Rashomon effect is *“a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others...”*

Rashomon Effect

Rashomon effect is “a combination of a **difference of perspective and equally plausible accounts**, with the absence of evidence to elevate one above others...”

Rashomon Effect

Rashomon effect is *“a combination of a difference of perspective and equally plausible accounts, with the **absence of evidence to elevate one above others...**”*

Rashomon Effect

Rashomon effect is *“a combination of a difference of perspective and equally plausible accounts, with the absence of evidence to elevate one above others...”*



Davis, B., Anderson, R. & Walls, J. (2015). *Rashomon Effects: Kurosawa, Rashomon and their legacies*. Routledge.

Rashomon Effect in AI

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

*“What I call the Rashomon Effect is that there is often **a multitude of different descriptions** (equations $f(x)$) in a class of functions giving **about the same minimum error rate.**”*

Rashomon Effect in AI

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

*“What I call the Rashomon Effect is that there is often **a multitude of different descriptions** (equations $f(x)$) in a class of functions giving **about the same minimum error rate.**”*



Rashomon Set

Rashomon Effect in AI



The World

Global Income

Rashomon Effect in AI

**Loan
Applications**



The World

Equal Income

Rashomon Effect in AI

**Loan
Applications**



The World



**A Mathematical Model of
the World**

Global Income

Rashomon Effect in AI

**Loan
Applications**



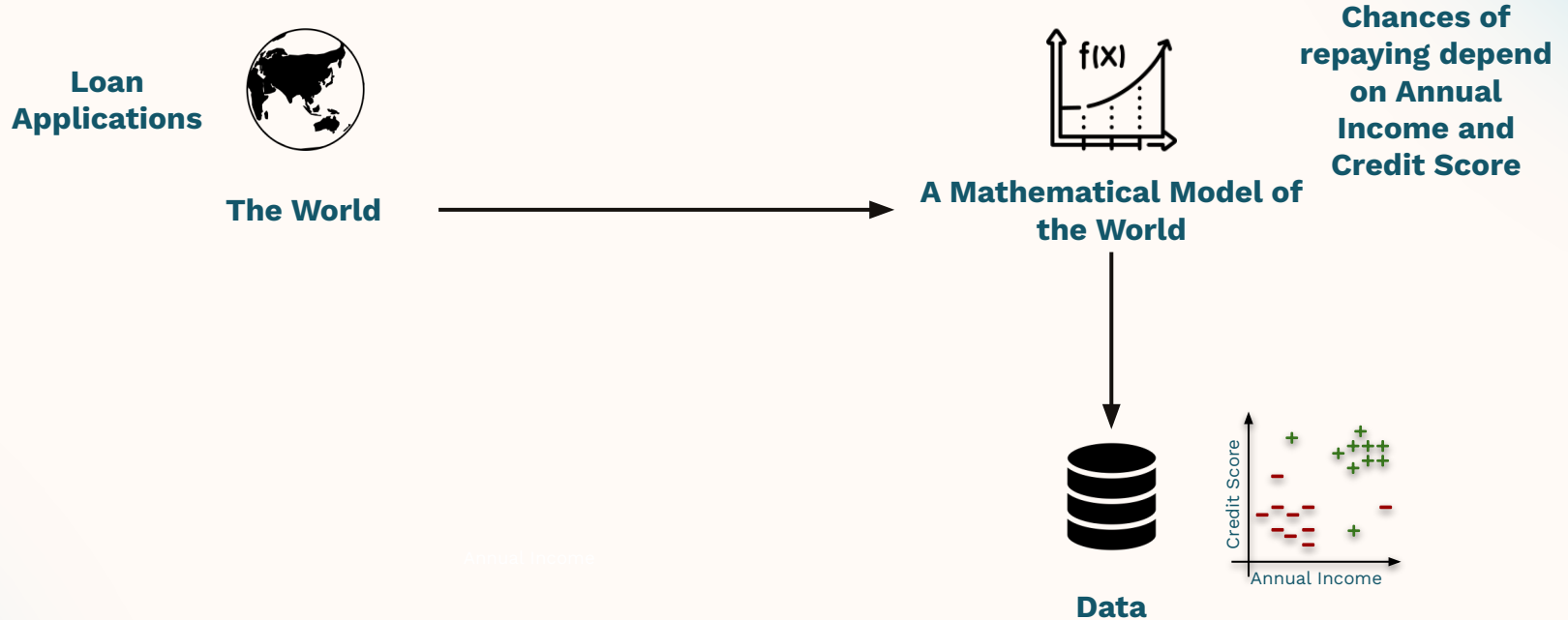
The World



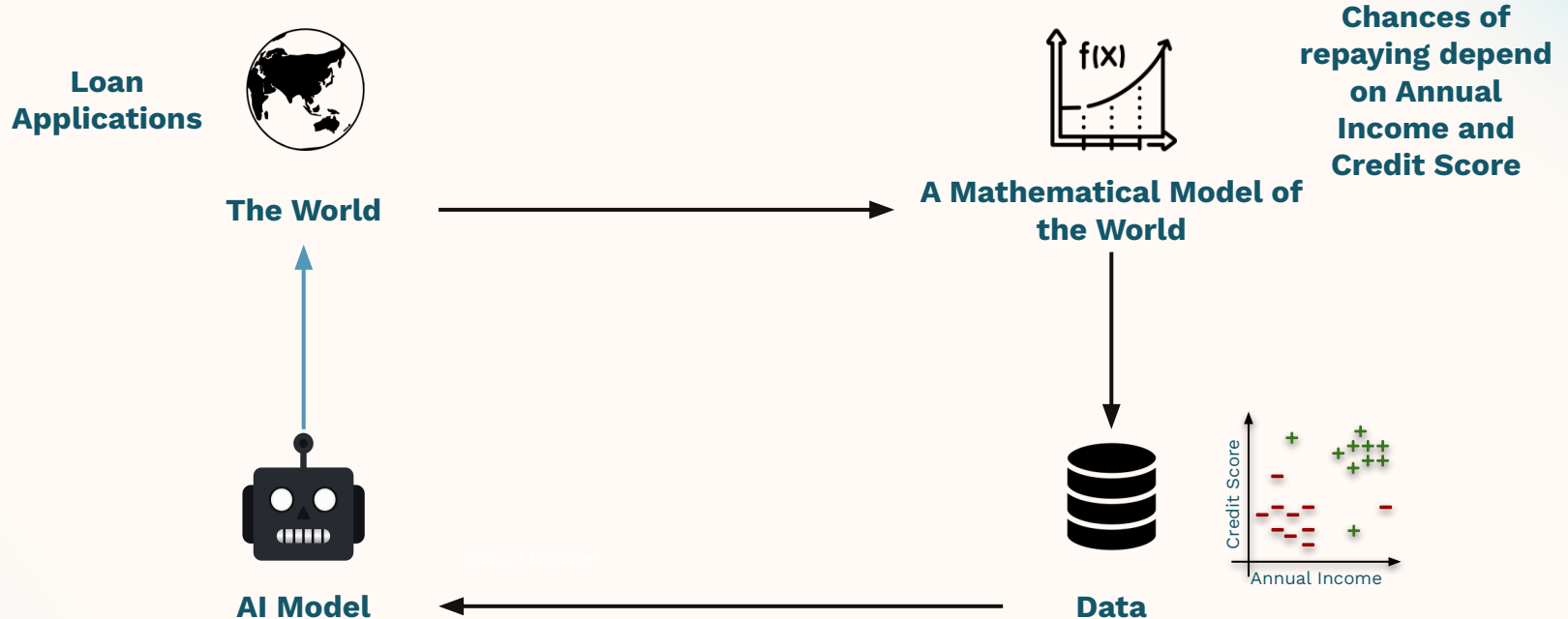
**A Mathematical Model of
the World**

**Chances of
repaying depend
on Annual
Income and
Credit Score**

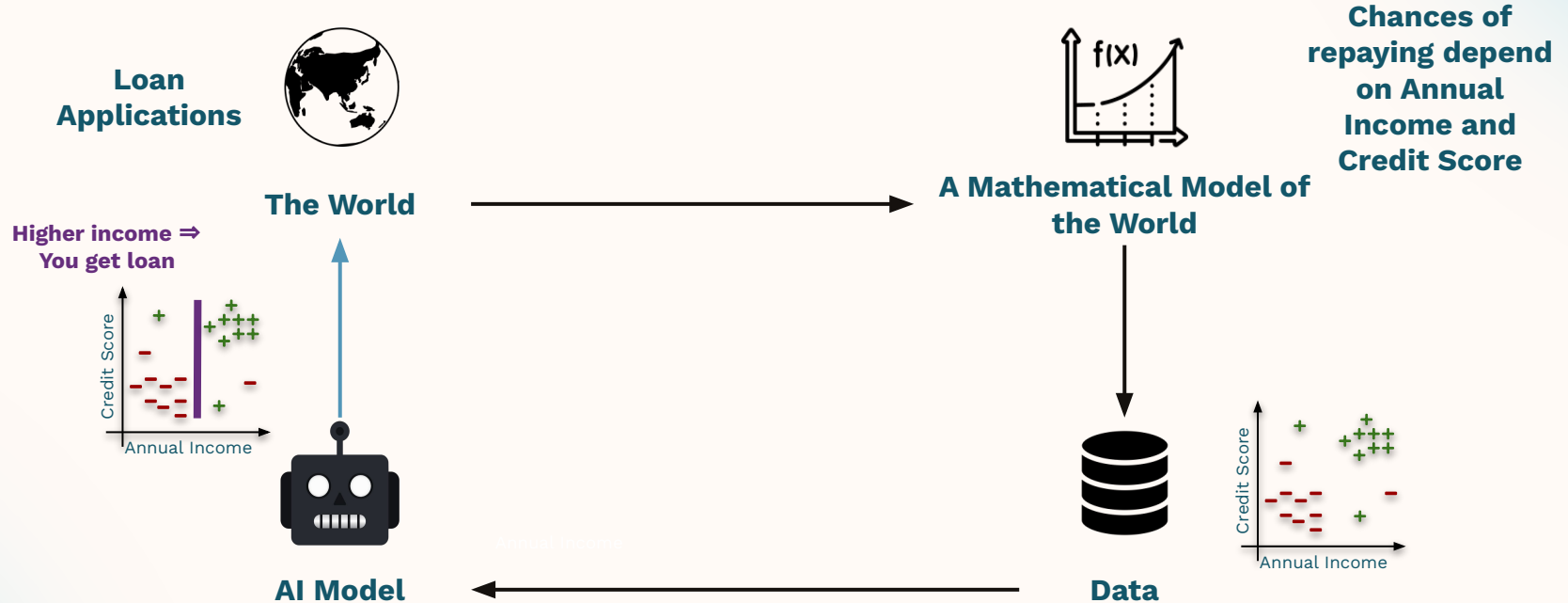
Rashomon Effect in AI



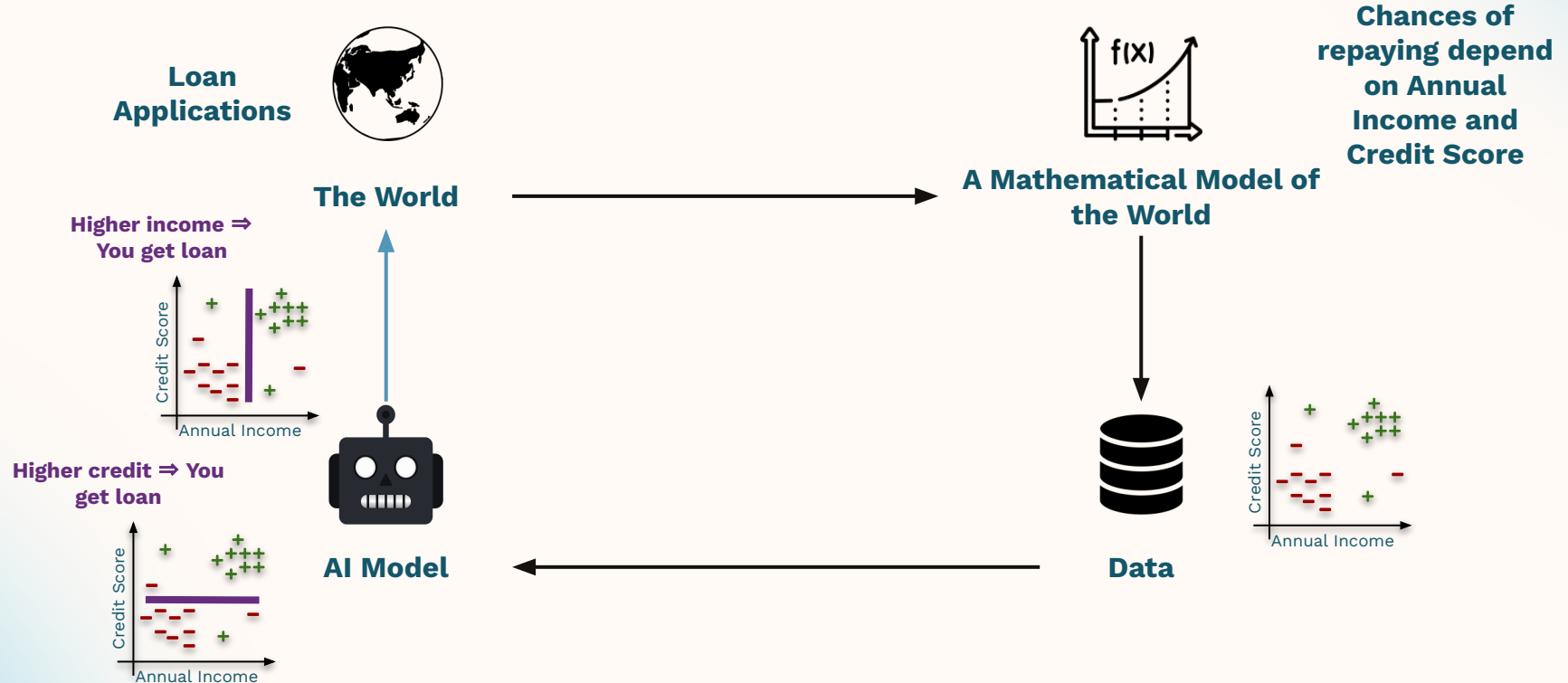
Rashomon Effect in AI



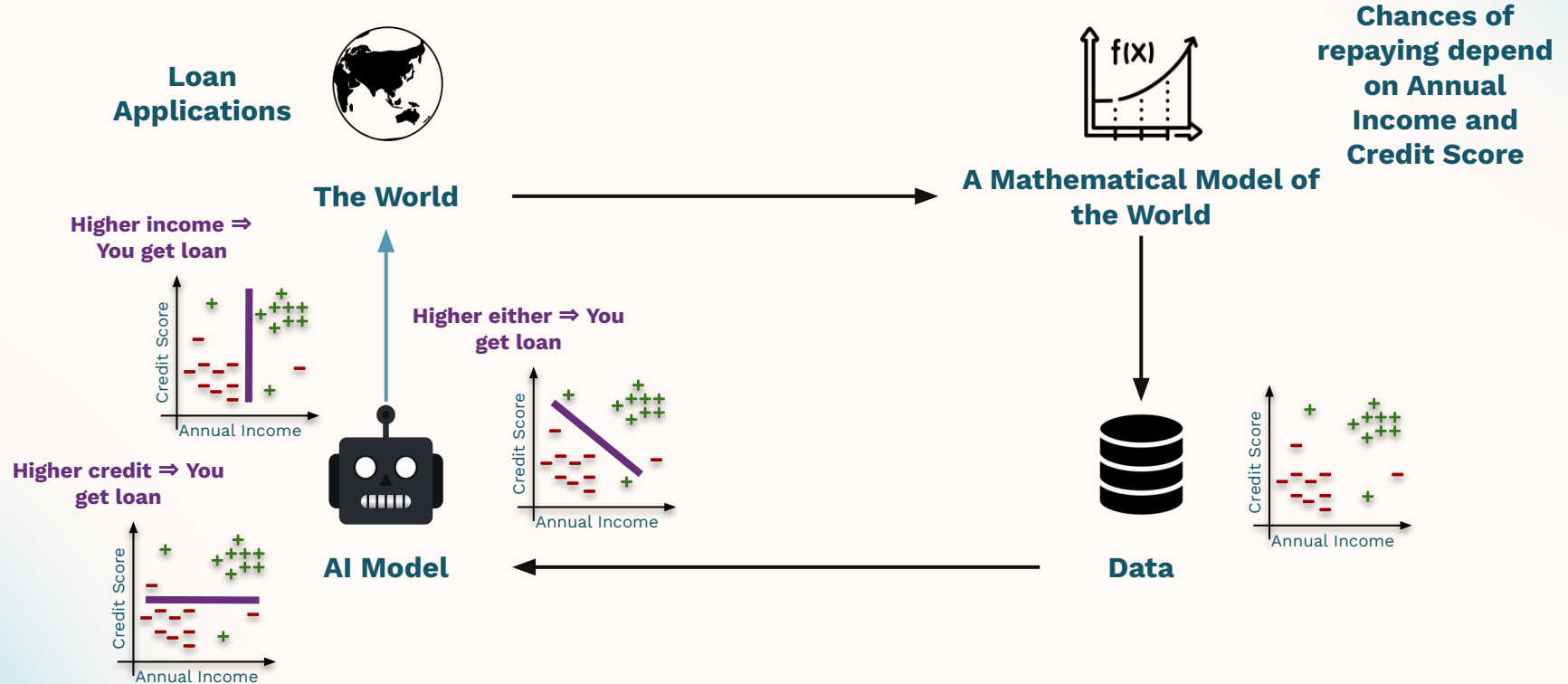
Rashomon Effect in AI



Rashomon Effect in AI



Rashomon Effect in AI

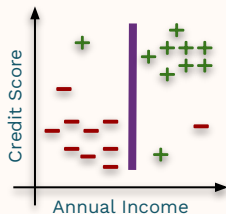


Source: All icons from Flaticon (<https://www.flaticon.com/>)

Rashomon Set

Or a set of competing models, a set of good models, ϵ -Rashomon set, ϵ -Level set, etc.

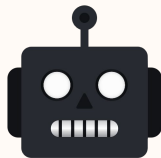
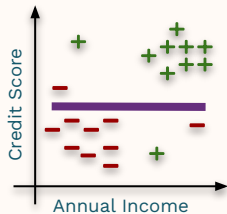
Higher income \Rightarrow You get loan



Higher either \Rightarrow You get loan



Higher credit \Rightarrow You get loan



AI Model

Rashomon Set

Or a set of competing models, a set of good models, ϵ -Rashomon set, ϵ -Level set, etc.

Def (ϵ -Level Set): Given a baseline classifier h_o and a hypothesis class \mathcal{H} , the ϵ -level set around h_o is the set of all classifiers $h \in \mathcal{H}$ with an error rate of at most $L(h_o) + \epsilon$ on the training data,

$$S_{\epsilon}(h_o) := \{h \in \mathcal{H} : L(h) \leq L(h_o) + \epsilon\}$$

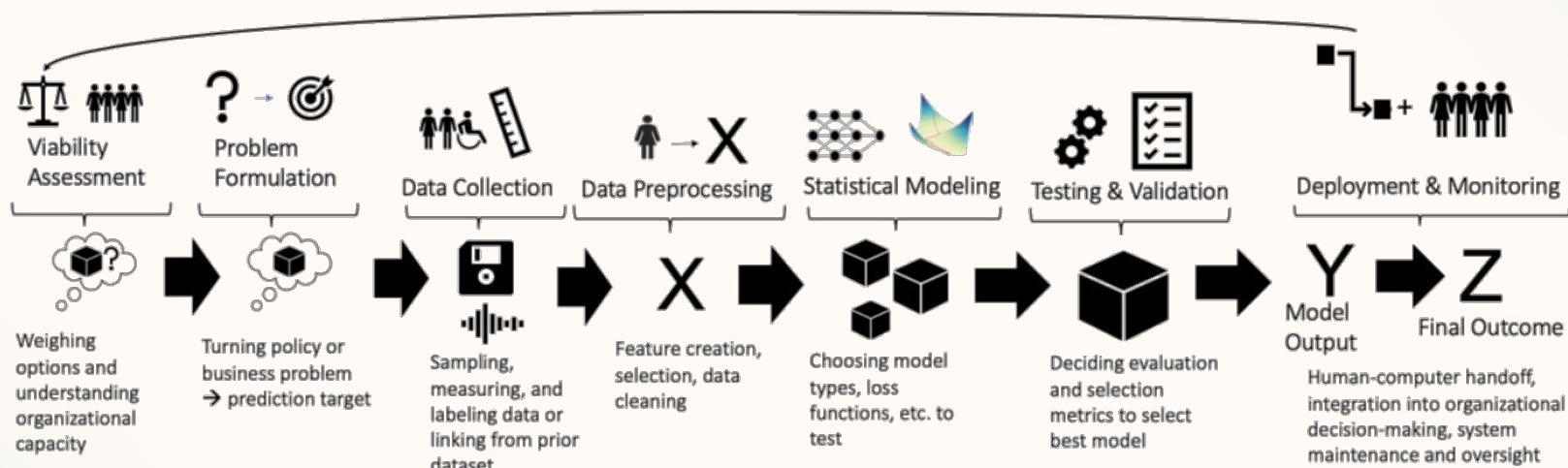
Rashomon Set

Or a set of competing models, a set of good models, ϵ -Rashomon set, ϵ -Level set, etc.

Def (ϵ -Level Set): Given ~~a baseline classifier h_0 and~~ a hypothesis class \mathcal{H} , the ϵ -level set ~~around h_0~~ is the set of all classifiers $h \in \mathcal{H}$ with an error rate of at most ~~$L(h_0) + \epsilon$~~ on the training data,

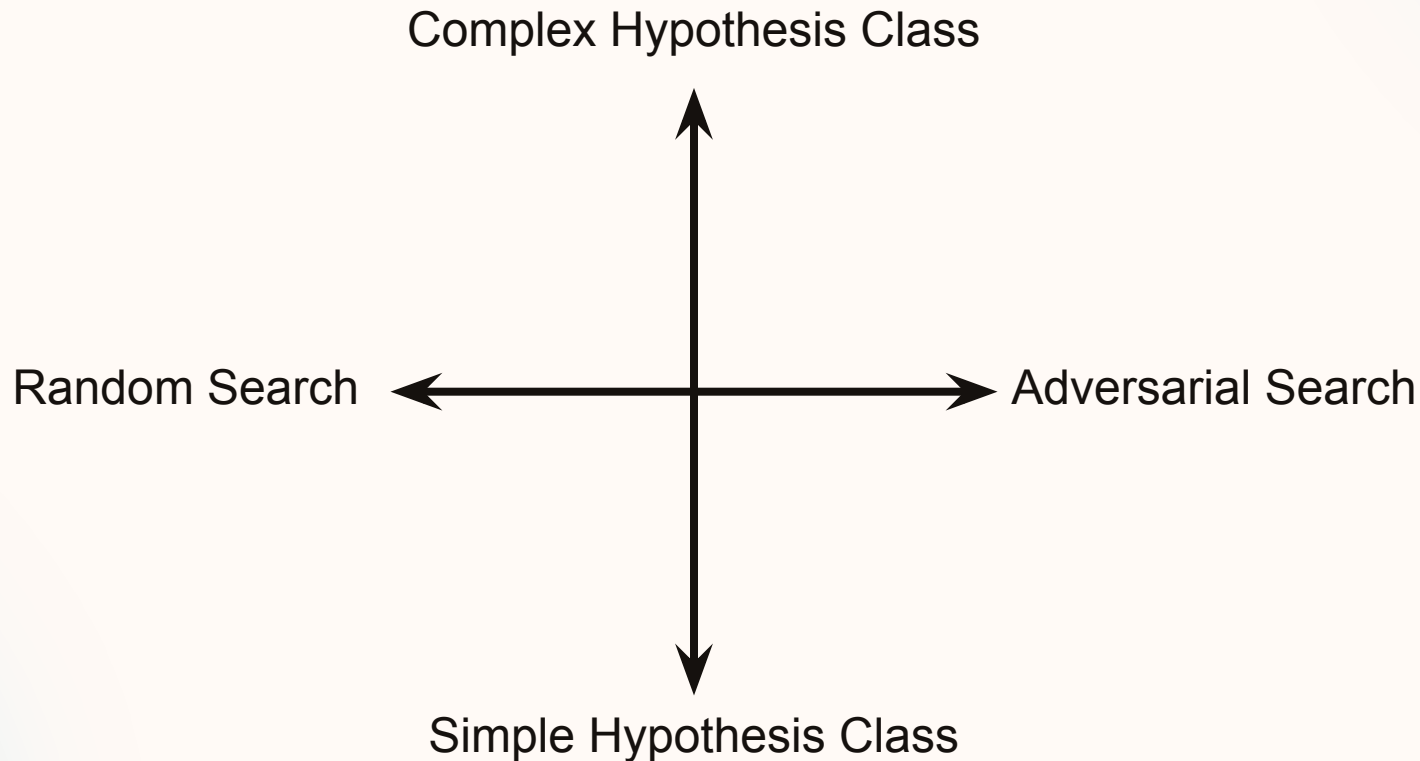
$$S_\epsilon := \{h \in \mathcal{H} : L(h) \leq \epsilon\}$$

Developer Choices and the Rashomon Effect

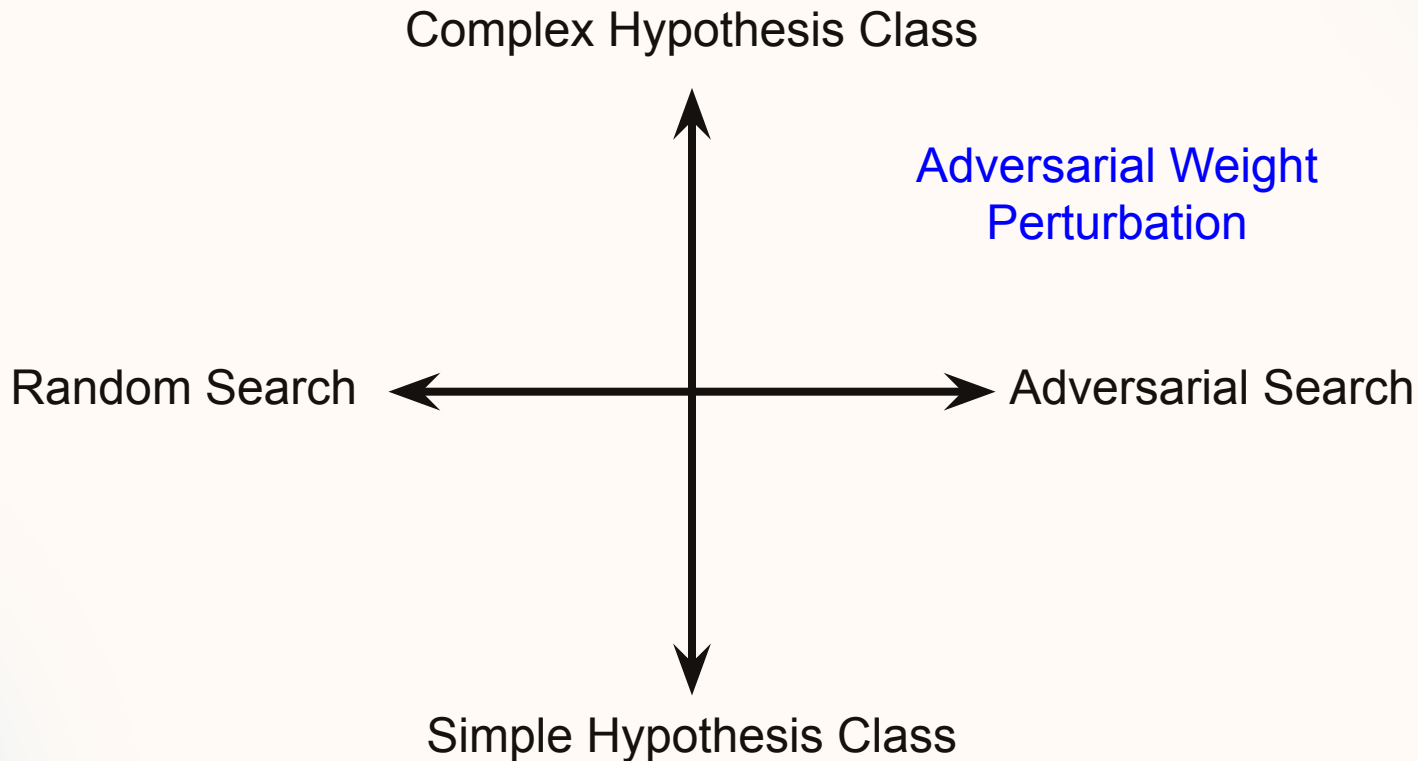


Source: Black et al. (2024)

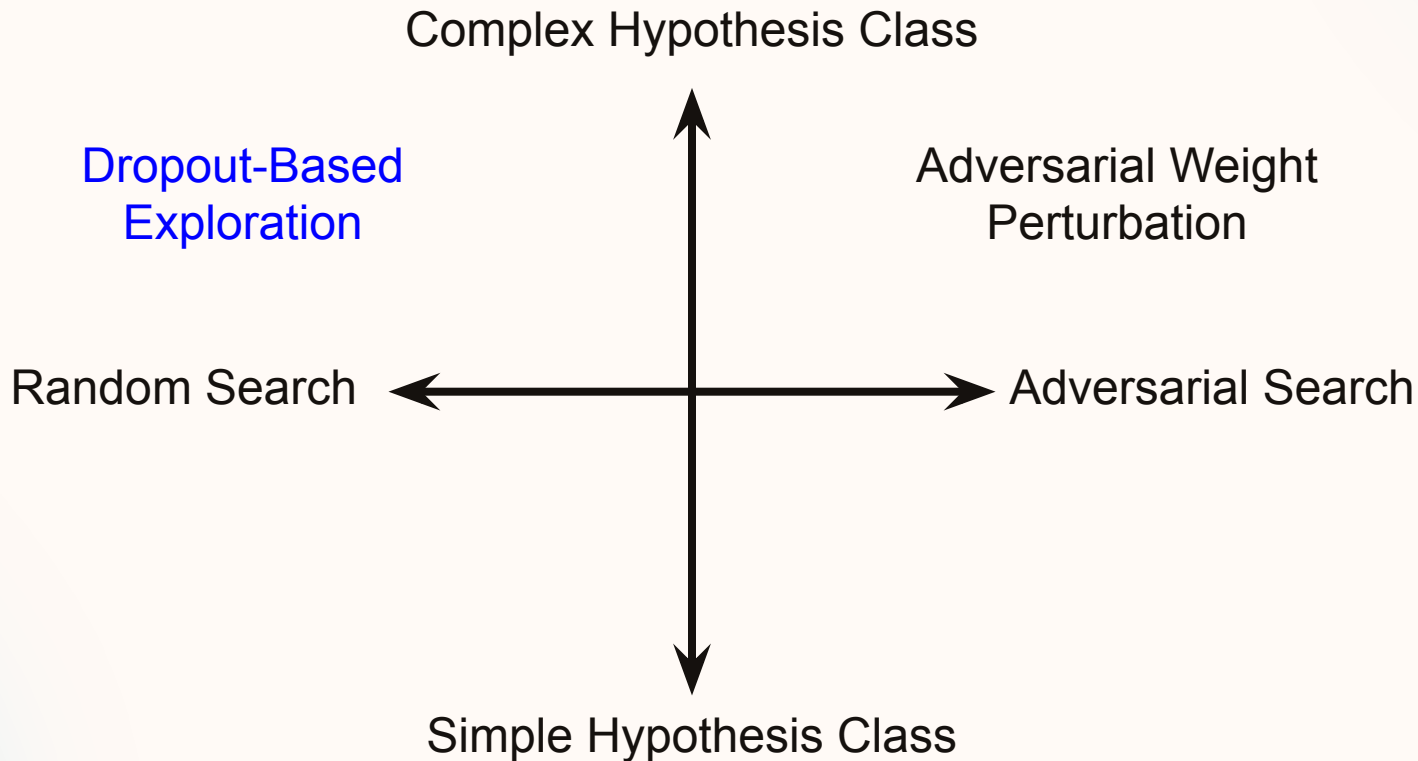
Exploring The Rashomon Set



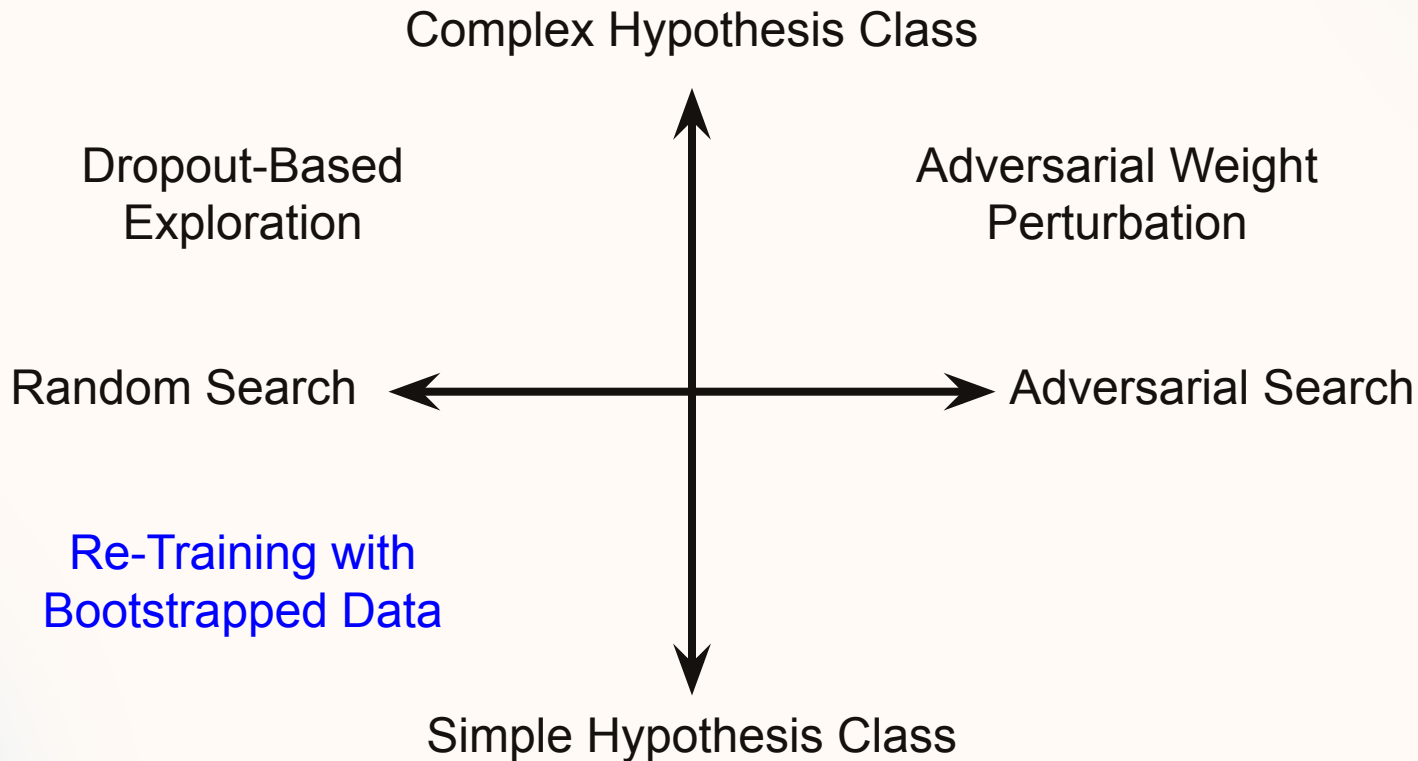
Exploring The Rashomon Set



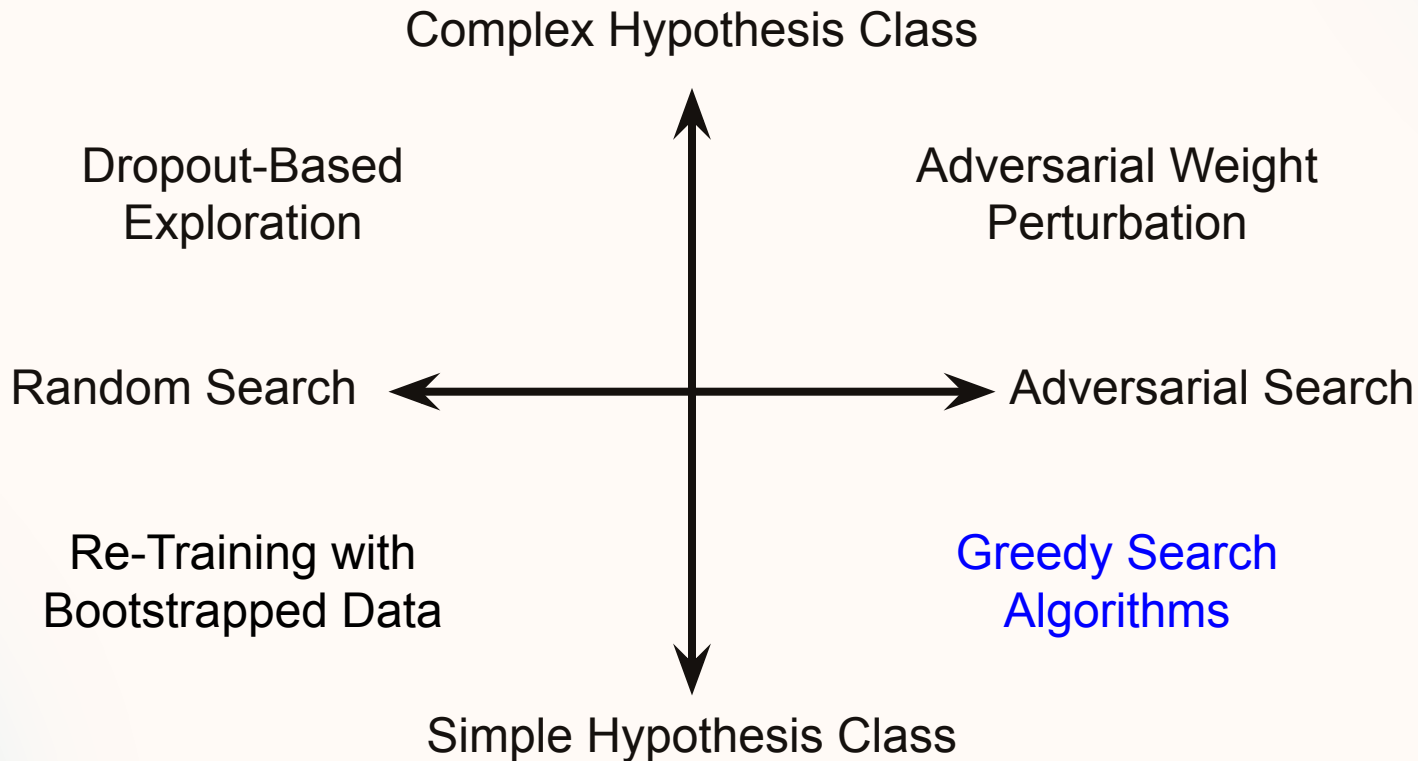
Exploring The Rashomon Set



Exploring The Rashomon Set



Exploring The Rashomon Set



Predictive Multiplicity

*“we define predictive multiplicity as the ability of a prediction problem to admit **competing models** that assign **conflicting predictions**.”*



Rashomon Set

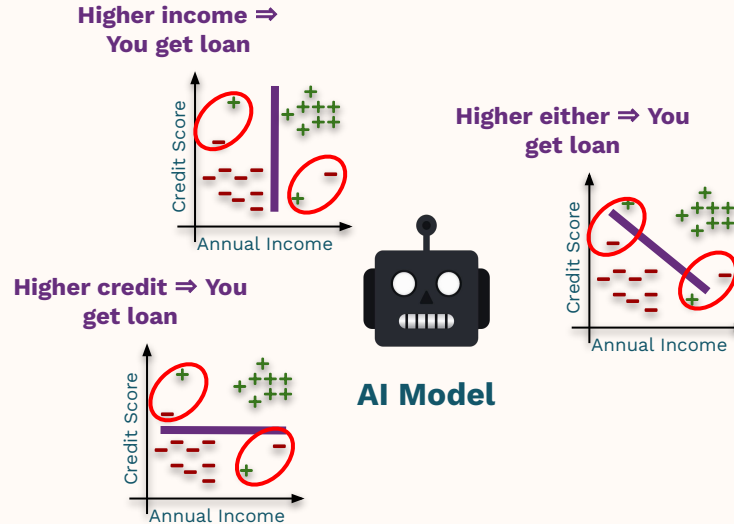


Predictive Multiplicity

Predictive Multiplicity

Definition (Predictive Multiplicity): Given a baseline classifier h_0 , a prediction problem exhibits predictive multiplicity over the ε -level set $S_\varepsilon(h_0)$ if there exists a model $h \in S_\varepsilon(h_0)$ such that $h(\mathbf{x}_i) \neq h_0(\mathbf{x}_i)$ for some \mathbf{x}_i in the training set.

Predictive Multiplicity & Conflicting Outcomes



The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

Allocation Multiplicity: multiplicity in predictions \Rightarrow multiplicity in allocation outcomes

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

Allocation Multiplicity: multiplicity in predictions => multiplicity in allocation outcomes

Procedural Multiplicity: multiplicity in model internals

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

Allocation Multiplicity: multiplicity in predictions => multiplicity in allocation outcomes

Procedural Multiplicity: multiplicity in model internals

Dataset Multiplicity: multiplicity in training data => multiplicity in predictions

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

Allocation Multiplicity: multiplicity in predictions => multiplicity in allocation outcomes

Procedural Multiplicity: multiplicity in model internals

Dataset Multiplicity: multiplicity in training data => multiplicity in predictions

Multi-Target Multiplicity: multiplicity in target variables => multiplicity in allocations

The Many Forms of Multiplicity

Predictive Multiplicity for Binary Classification: multiplicity in prediction classes

Predictive Multiplicity for Probabilistic Classification: multiplicity in prediction probabilities

Allocation Multiplicity: multiplicity in predictions \Rightarrow multiplicity in allocation outcomes

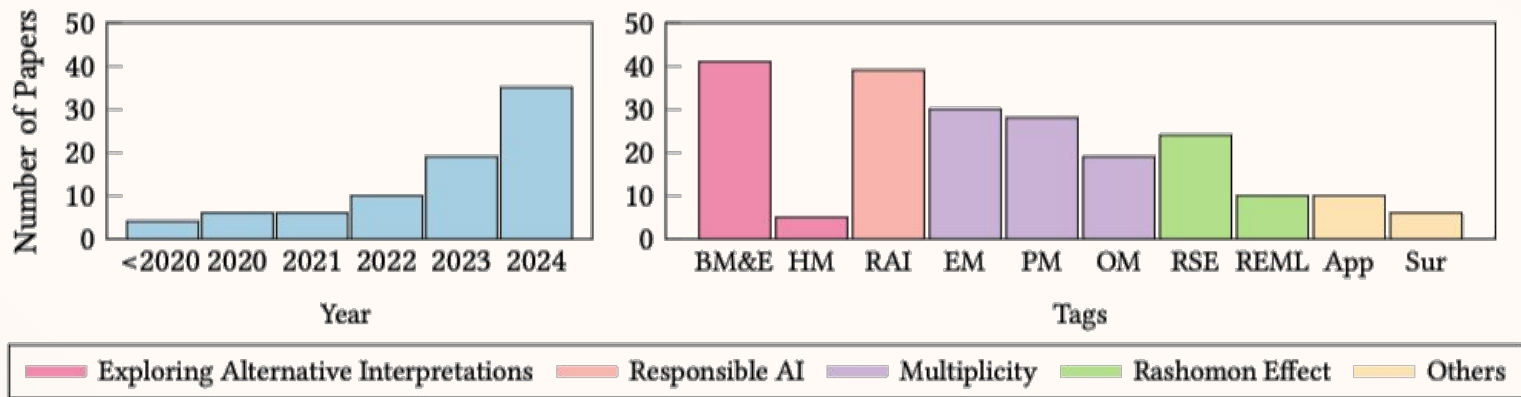
Procedural Multiplicity: multiplicity in model internals

Dataset Multiplicity: multiplicity in training data \Rightarrow multiplicity in predictions

Multi-Target Multiplicity: multiplicity in target variables \Rightarrow multiplicity in allocations

Explanation Multiplicity: multiplicity in model internals \Rightarrow multiplicity in explanations

Growing Literature About Multiplicity



BM&E: Better Models and Ensembles; **HE:** Hacking Metrics; **RAI:** Responsible AI; **PM:** Predictive Multiplicity; **EM:** Explanation Multiplicity; **OM:** Other Multiplicity; **RSE:** Rashomon Set Exploration; **REML:** Rashomon Effect in ML; **App:** Application; **Sur:** Survey.

Source: Ganesh et al. (2025)

Implications of Multiplicity for Fairness

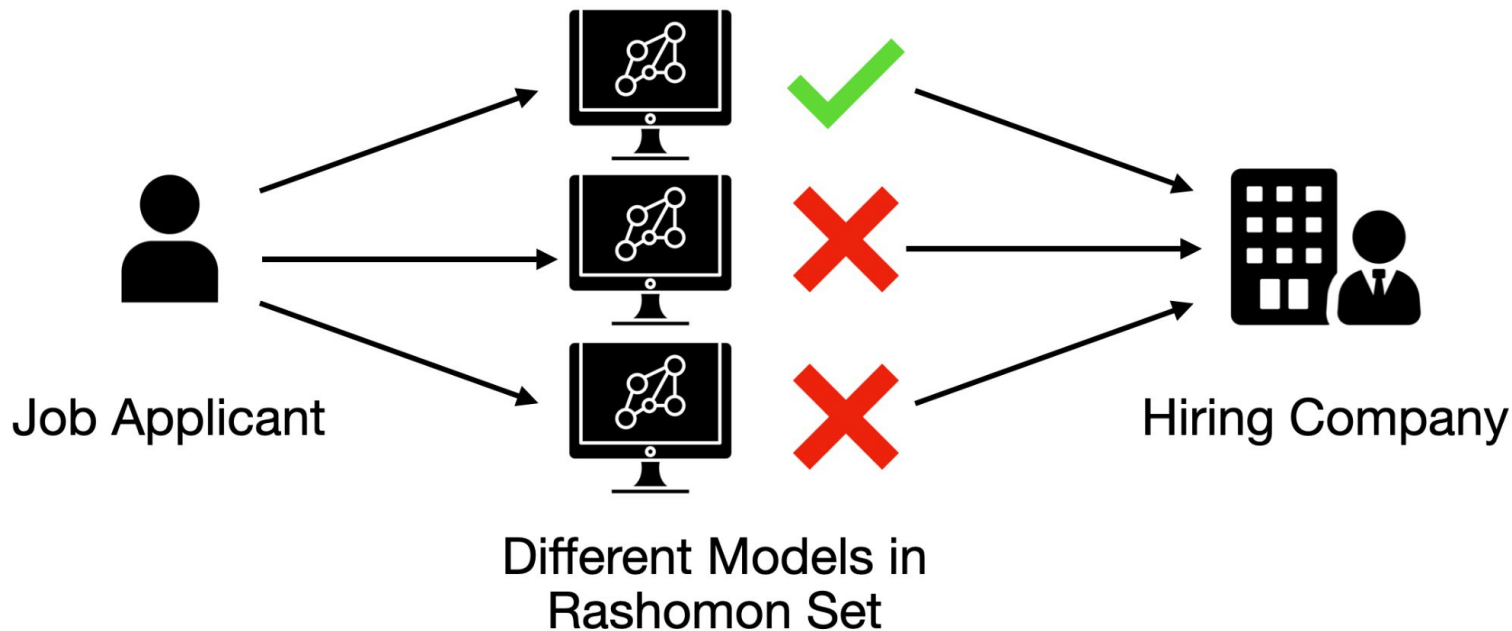
Implications of Multiplicity for Fairness

Different fairness concerns => Different multiplicity interventions

Fairness Concern	Multiplicity Intervention
Conflicting Outcomes	Combining Models
Outcome Homogenization	Randomizing Among Models
Group Fairness	Secondary Criteria to Choose Models

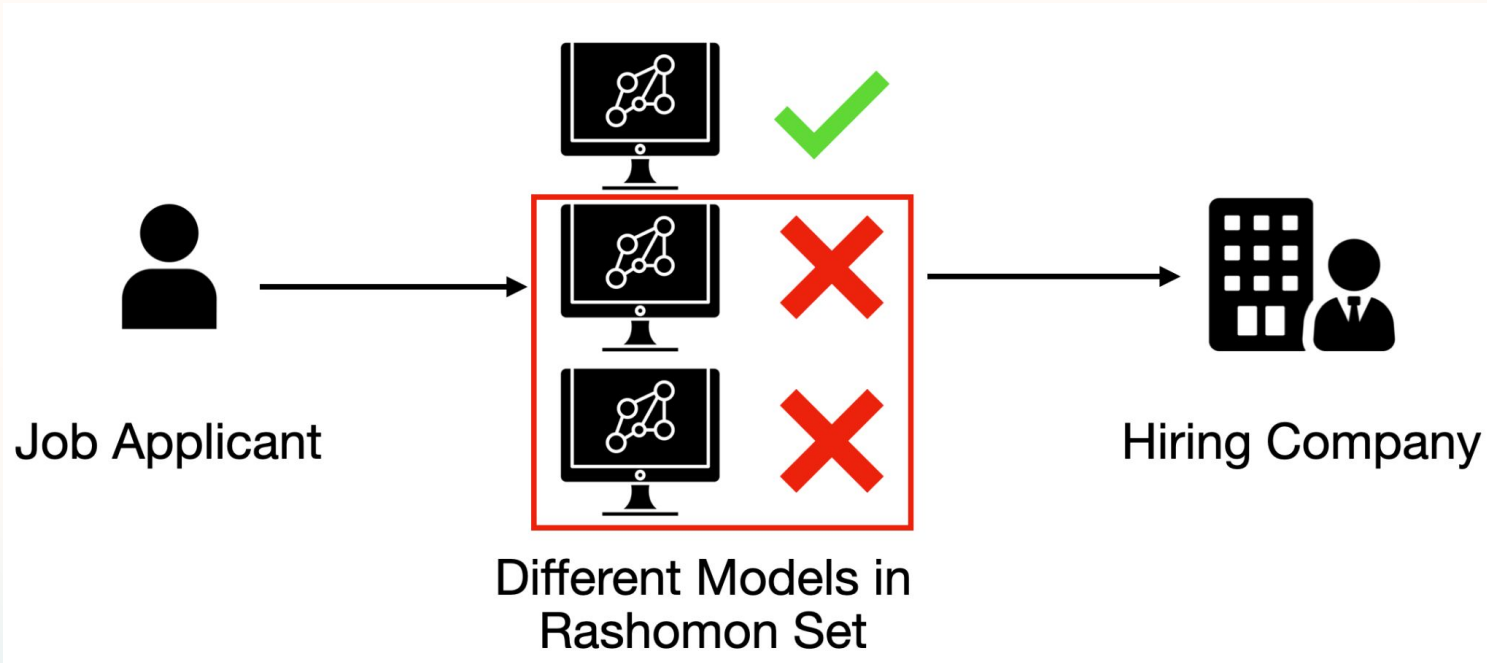
Conflicting Outcomes

A job applicant may be accepted or rejected depending on the model chosen

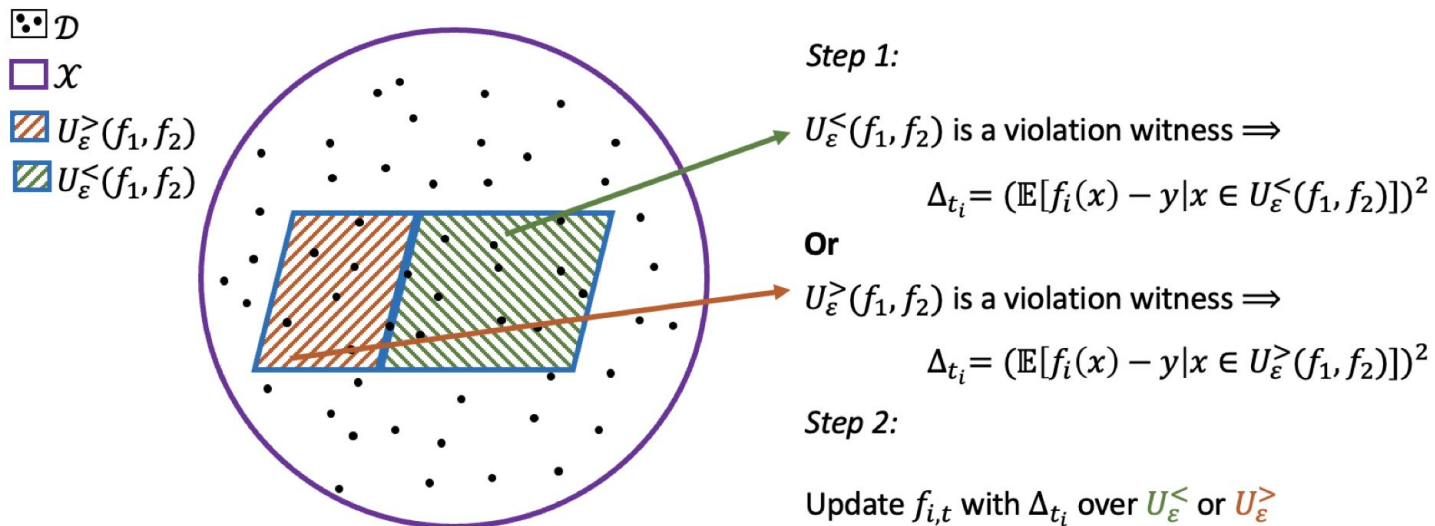


Ensemble of Rashomon Models

Is the ensemble of Rashomon models more fair than using a single model?



Reconciliation of Rashomon Models



Algorithmic Monoculture

Decision-makers using the same or similar algorithm

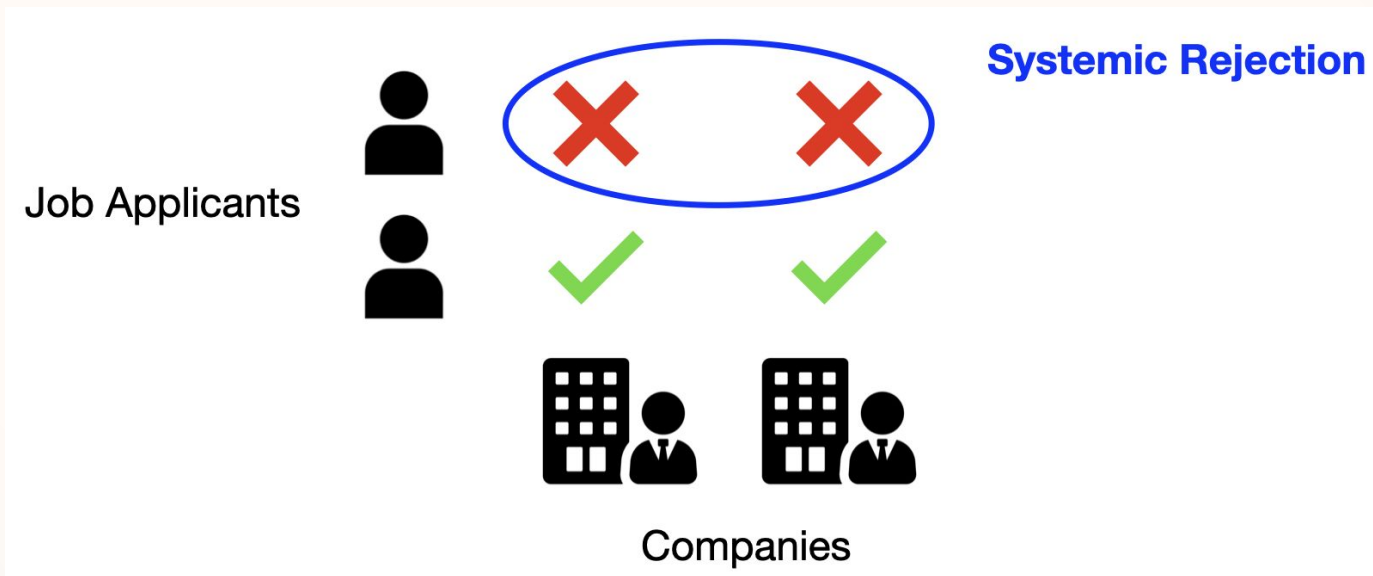


Kleinberg and Raghavan. "Algorithmic Monoculture and Social Welfare" (PNAS 2021)

Bommasani et al. "Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?" (NeurIPS 2022)

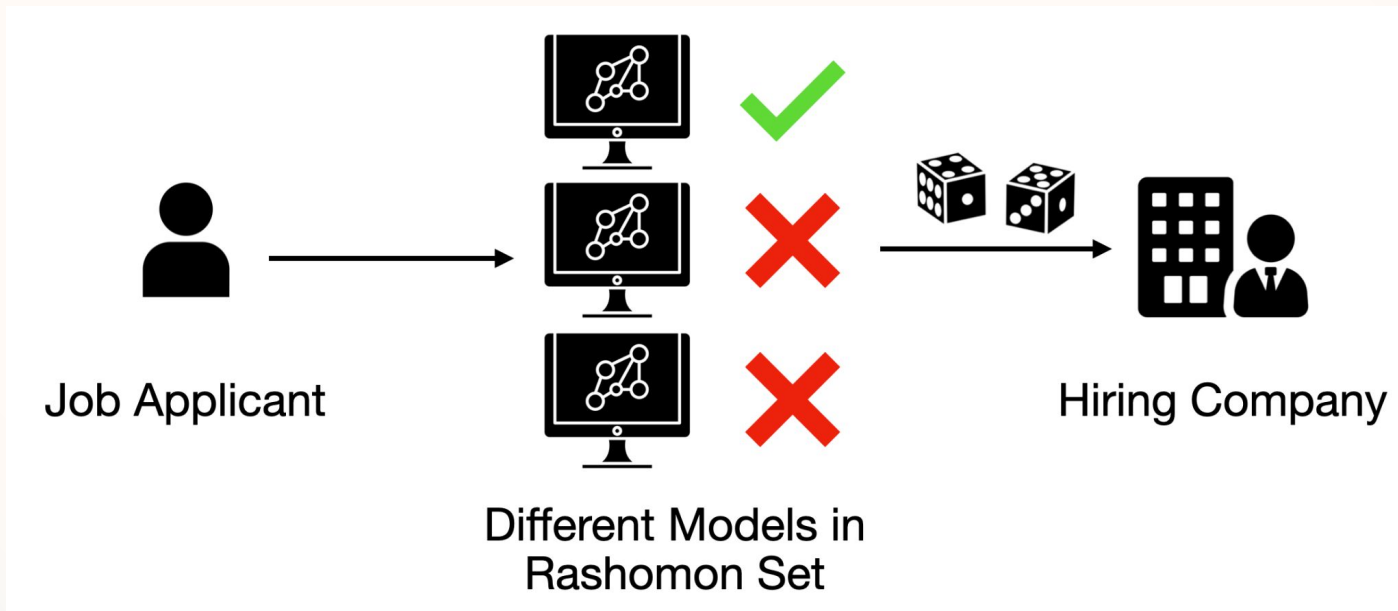
Outcome Homogenization

When individuals receive negative outcomes from all decision-makers



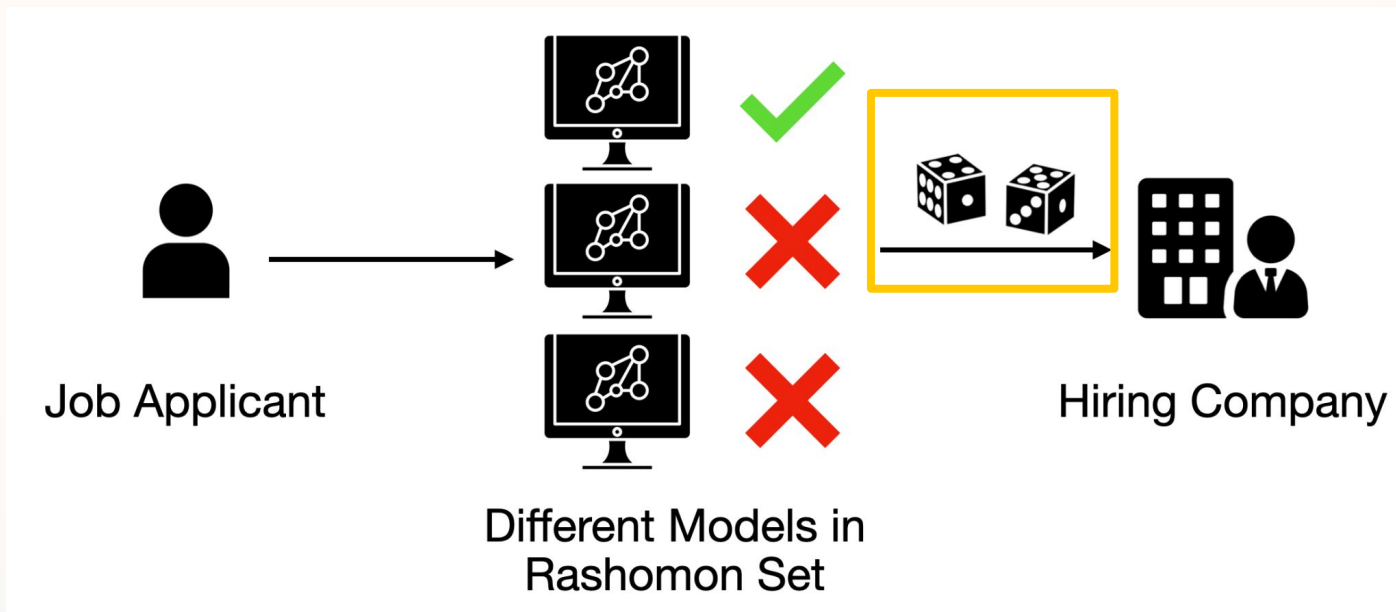
Randomly Choosing Among Rashomon Models

Reducing Outcome Homogenization Without Compromising Accuracy



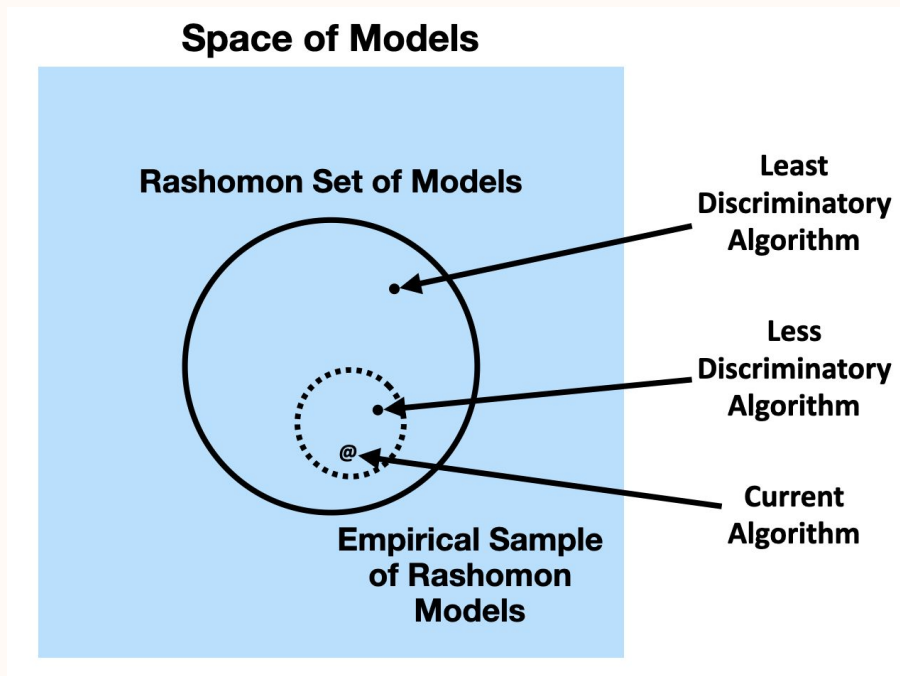
Randomly Choosing Among Rashomon Models

Reducing Outcome Homogenization Without Compromising Accuracy

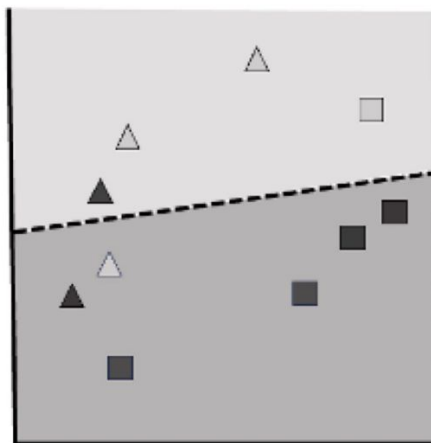
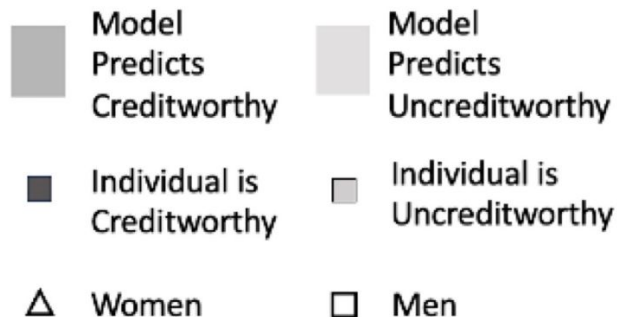


Less Discriminatory Algorithms

There may exist equally accurate models that improve on a given fairness metric

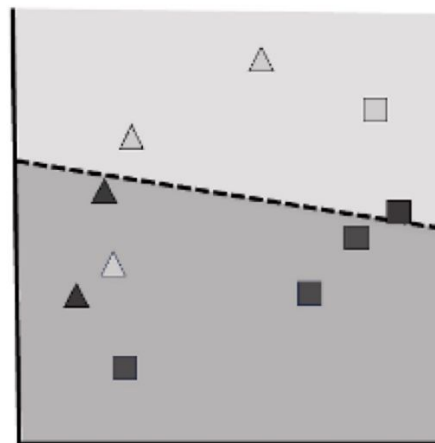


Choosing a Less Discriminatory Algorithm



□ Men's Selection Rate: 80%

△ Women's Selection Rate: 40%



□ Men's Selection Rate: 60%

△ Women's Selection Rate: 60%

Overall Accuracy: 80%

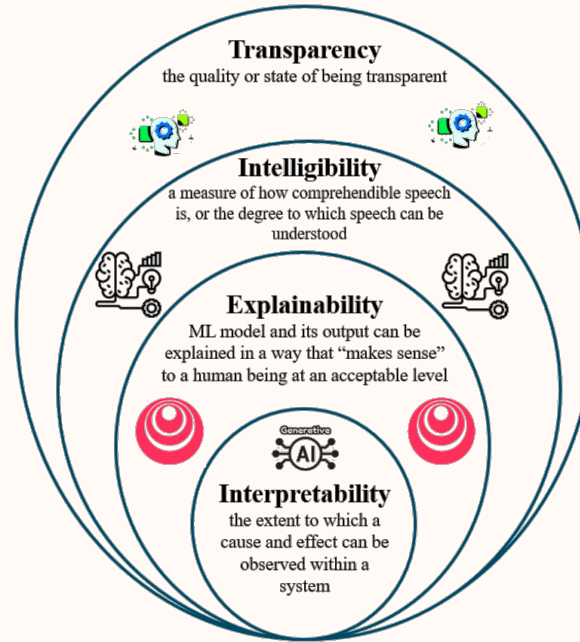
Implications of Multiplicity for Fairness

Different fairness concerns => Different multiplicity interventions

Fairness Concern	Multiplicity Intervention
Conflicting Outcomes	Combining Models
Outcome Homogenization	Randomizing Among Models
Group Fairness	Secondary Criteria to Choose Models

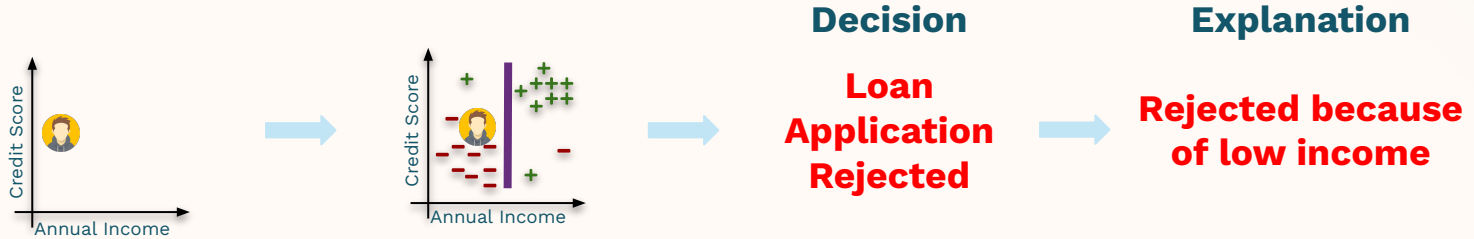
Implications of Multiplicity for Explainability

What is an Explanation?

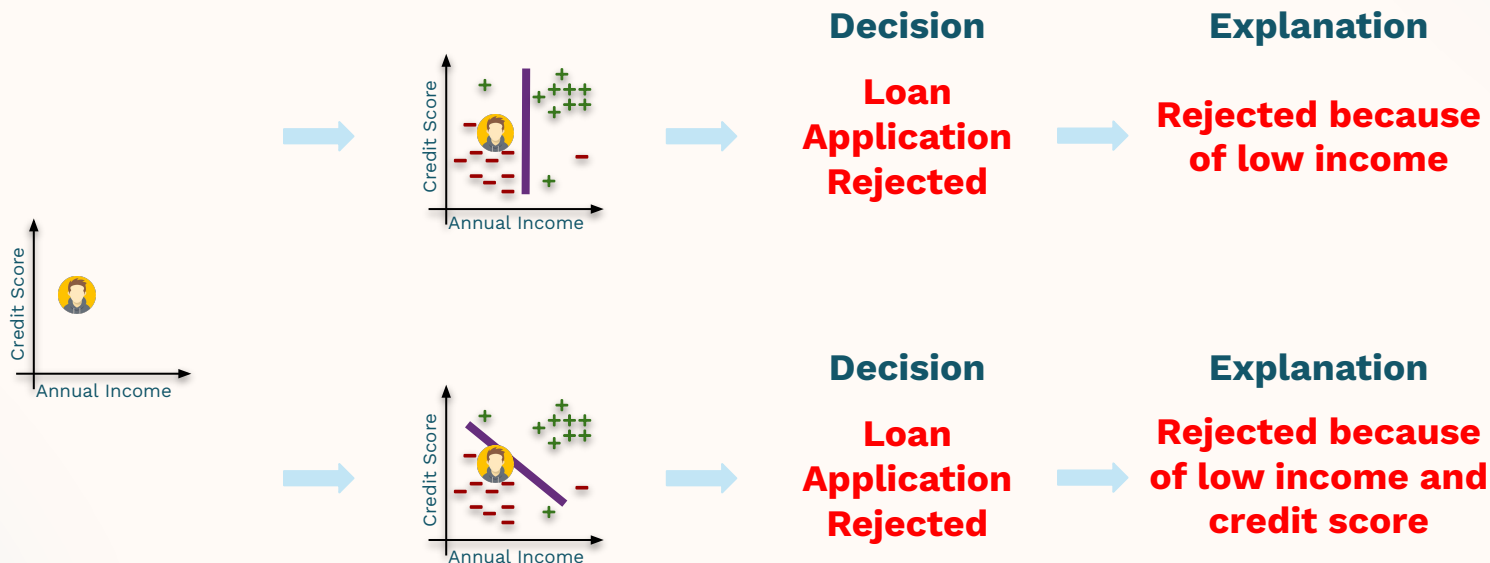


Source: Shafik et al. (2024)

What is an Explanation?



Explanation Multiplicity



Implications of Multiplicity for Explainability

Fairness Explainability Concern	Multiplicity Intervention
Multiple Outcomes Explanations	Combining Models
Outcome Homogenization	Randomizing Among Models
Group Fairness Interpretability	Secondary Criteria to Choose Models

Explanation Multiplicity

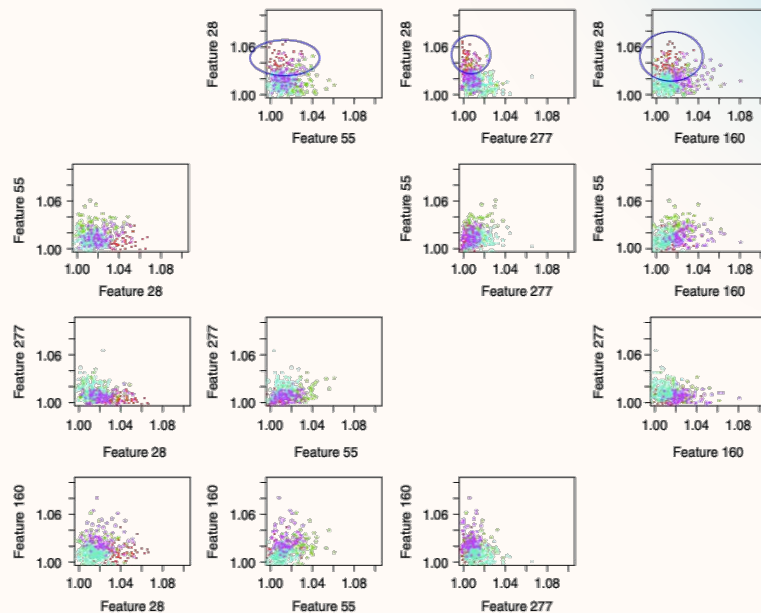
Attribution-based Explanations

Recourse/Counterfactuals

Explanation Multiplicity

Attribution-based Explanations

Recourse/Counterfactuals



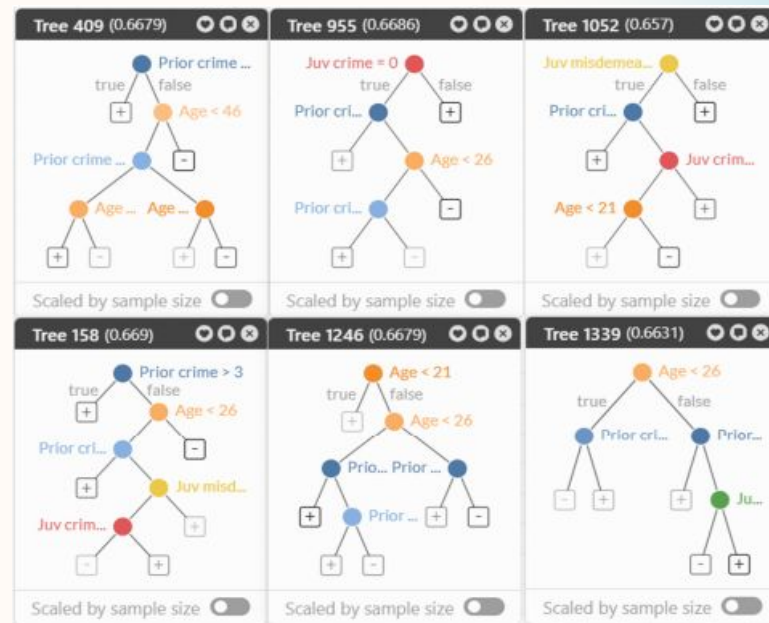
Variable Importance Clouds

Source: Dong et al., 2020

Explanation Multiplicity

Attribution-based Explanations

Recourse/Counterfactuals



Visualizing the Rashomon Set

Source: Rudin et al., 2024

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., ... & Boner, Z. (2024, July). Position: amazing things come from having many good models. In Proceedings of the 41st International Conference on Machine Learning (pp. 42783-42795).

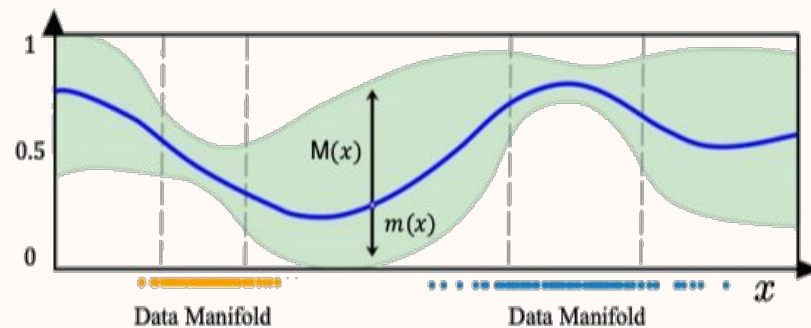
Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., & Rudin, C. (2022). Exploring the whole rashomon set of sparse decision trees. Advances in neural information processing systems, 35, 14071-14084.

Wang, Z. J., Zhong, C., Xin, R., Takagi, T., Chen, Z., Chau, D. H., ... & Seltzer, M. (2022, October). Timbertrek: Exploring and curating sparse decision trees with interactive visualization. In 2022 IEEE Visualization and Visual Analytics (VIS) (pp. 60-64). IEEE.

Explanation Multiplicity

Attribution-based Explanations

Recourse/Counterfactuals

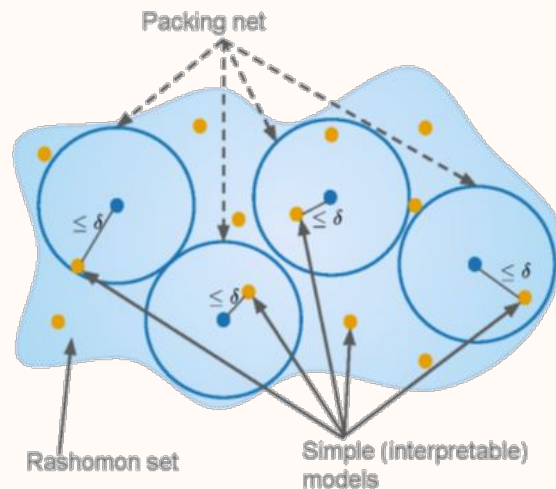


Source: Hamman et al., 2023

Simpler Models and Multiplicity

Larger Rashomon ratios lead to better chances of finding simpler models

$$\text{Rashomon Ratio} := \frac{\text{Volume of Rashomon Set}}{\text{Volume of Hypothesis Space}}$$



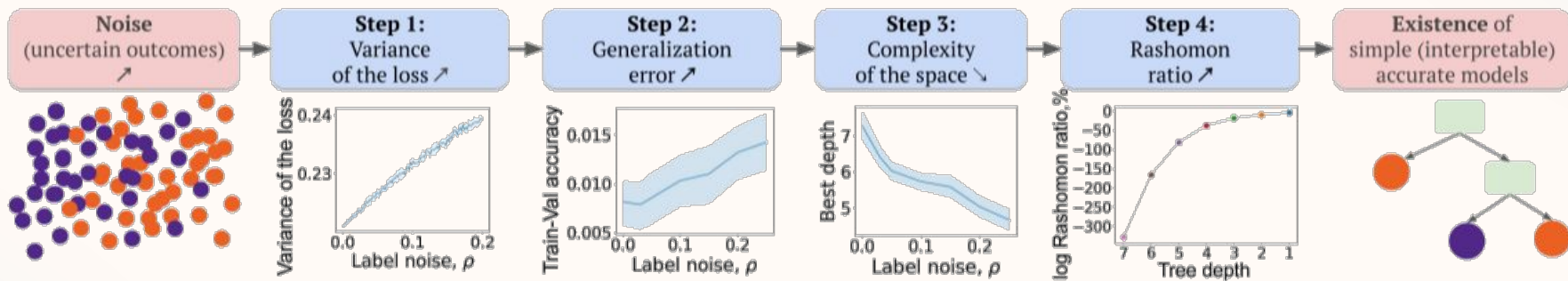
Source: Rudin et al., 2024

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., ... & Boner, Z. (2024, July). Position: amazing things come from having many good models. In Proceedings of the 41st International Conference on Machine Learning (pp. 42783-42795).

Semenova, L., Rudin, C., & Parr, R. (2022, June). On the existence of simpler machine learning models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1827-1858).

Simpler Models and Multiplicity

Noisier settings lead to larger Rashomon ratios



Source: Rudin et al., 2024

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., ... & Boner, Z. (2024, July). Position: amazing things come from having many good models. In Proceedings of the 41st International Conference on Machine Learning (pp. 42783-42795).

Semenova, L., Chen, H., Parr, R., & Rudin, C. (2023). A path to simpler models starts with noise. Advances in neural information processing systems, 36, 3362-3401.

Discussion

How should decision-makers address multiplicity?

Fairness Concern	Multiplicity Intervention
Conflicting Outcomes	Combining Models
Outcome Homogenization	Randomizing Among Models
Group Fairness	Secondary Criteria to Choose Models

Explainability Concern	Multiplicity Intervention
Multiple Explanations	Combining Models
Interpretability	Secondary Criteria to Choose Models

How should decision-makers address multiplicity?

Domain-Specific Considerations

- **Factual vs Normative Decisions:** Whether there exists a single ground truth for the target variable
- **Single-Shot vs Multi-Shot Settings:** Whether individuals may receive the resource or opportunity from another decision-maker at a later time
- **Low vs High Aleatoric Uncertainty:** Whether there is variability in the target variable for the same inputs
- **Comparative vs Individual Decisions:** Whether decision-makers are evaluating individuals separately or collectively

Scenario 1: Hiring

A large company is planning to partially automate their entry-level hiring pipeline. They receive over 100 applications each week, and recruiters don't have time to review every application. They are planning to train a model that will select the top 25 applicants every week for manual review. The company's competitors have recently adopted similar automated hiring tools.

Scenario 1: Hiring

When developing their model, the company notices there is predictive multiplicity, so they search for 100 models in the Rashomon Set. All models satisfy the 80% rule in demographic parity.

Imagine you are an applicant for the company. How would you prefer they address multiplicity?

1. Use an ensemble model that averages predictions
2. Randomize among individuals that would be in the top 25 applicants under any Rashomon model
3. Use the model with the best demographic parity on race & gender



Slido: #2634641

Scenario 2: Tax Audit

A government tax agency is planning to use an AI system to detect potential tax fraud. Everyone flagged by the AI will be audited, and audits will occur solely based on AI flags. However, taxpayers will only face penalties if a human investigator confirms the fraud. The tax agency checks that the model works well by seeing how often it agrees with human experts' choices of who to audit. The AI and the human experts agree 85% of the time.

Scenario 2: Tax Audit

When developing their model, the government notices there is predictive multiplicity, so they search for 100 models in the Rashomon Set. Imagine you are a taxpayer. How would you prefer they address multiplicity?

1. Use an ensemble model that averages predictions, but only flag individuals with high risk
2. Randomize so the chance of being flagged depends on average risk score
3. Use the model with the most similar false positive rates across income levels



Slido: #2634642

Scenario 3: Loan Application

A bank is planning to use an AI system to automate the process of small loan applications. All small loan applications will require filling a structured form, and the loan application decision will be made in an automated manner by an AI system. If the application is rejected, the bank also wants the system to provide recourse to the applicant.

Scenario 3: Loan Application

When developing their model, the bank notices explanation multiplicity. They search for 100 models in the Rashomon set and find that even when an application is rejected by them all, the suggested recourse differs between them. Imagine you are an applicant who was rejected. How would you prefer the bank addresses multiplicity?

1. Use the model with the easiest recourse for the candidate
2. Use the model with the most 'robust' recourse. Same as finding recourse that flips the decisions for most models
3. Use the model with the most 'diverse' recourse options



Slido: #2634643

Scenario 4: Recognizing Fault Points

A company that manufactures duplex steel is planning to use an AI system to study the causes of heating silver faults in their manufacturing pipeline. These faults can occur at various points in the process, and the company wants to use AI systems to recognize which parts of the pipeline need to be improved or replaced.

Scenario 4: Recognizing Fault Points

When developing their model, the company notices multiplicity: different models recognize different parts of the pipeline to be faulty even though they have similar overall accuracy. Imagine you are responsible for maintenance. How would you address multiplicity?

1. Use majority-voting to find the most likely failure point
2. Cluster models based on their recognized failure points, and then test each possibility in a random order
3. Only consider the failure points recognized by the model that had the best recall, i.e., recognized the most faults



Slido: #2634644

Scenario 5: Diagnosis Support

A hospital is planning to use an AI system to help doctors with preliminary diagnosis. Medical records of patients are fed to the AI system, and a diagnosis along with an explanation is shown to the doctor. The doctor will study the explanation provided by the AI system to judge its reliability, and to guide their own diagnosis. The final diagnosis will be provided by the doctor.

Scenario 5: Diagnosis Support

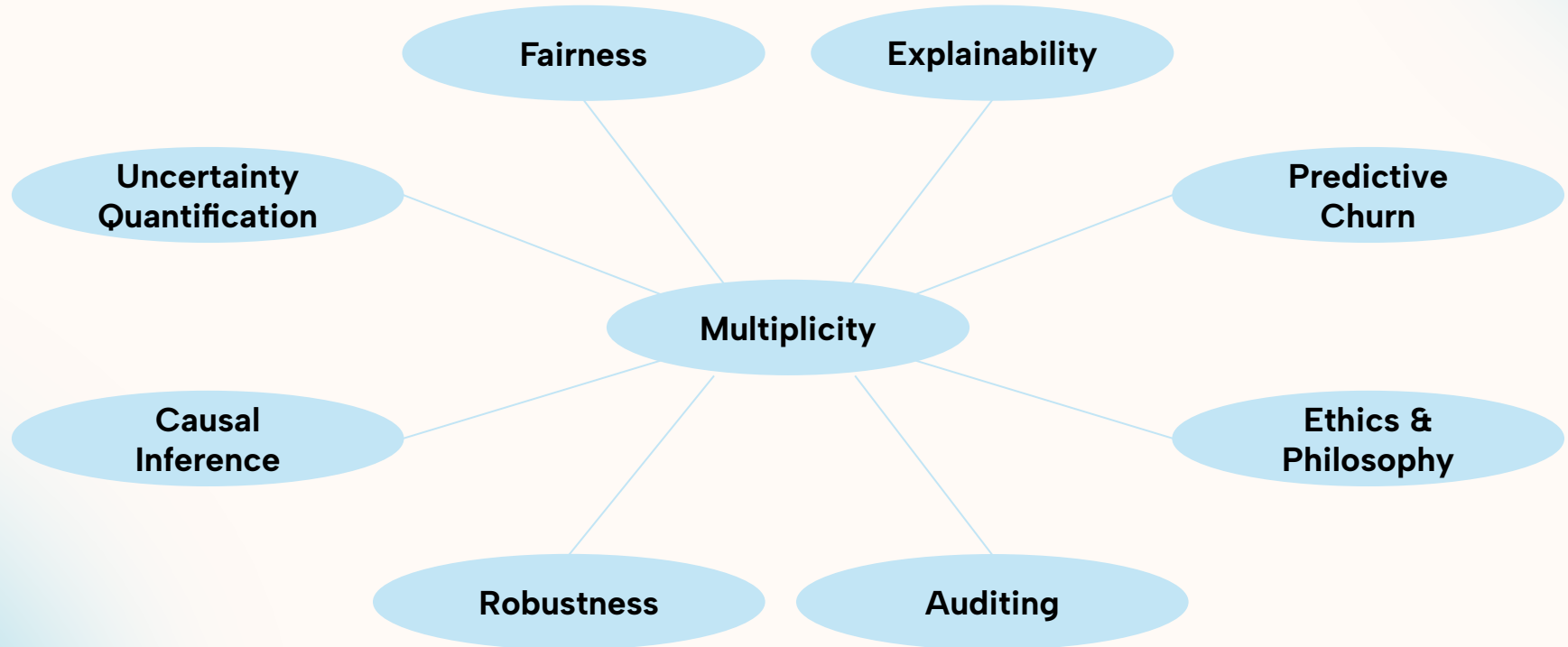
When developing their model, the hospital notices there is explanation multiplicity even when two models agree on their diagnosis. Imagine you are a doctor. How would you prefer the AI system developers address multiplicity?

1. Aggregate explanations to find the most common explanation
2. Provide all explanations under a sample of Rashomon models
3. Only provide a diagnosis if majority of explanations are aligned, otherwise do not provide any diagnosis or explanations



Slido: #2634645

The Many Faces of Multiplicity



Directions for Future Work

- Methods to Explore the Rashomon Set
- Normative Implications of Multiplicity
- Mitigating Underspecification
- Replicability in ML Research
- Multiplicity in Generative AI

Multiplicity in Generative AI

- Prompt Multiplicity
- Multiplicity in Reward Models
- Generative Monoculture
- Pluralistic Alignment

The Many Faces of Multiplicity in ML

There is a growing debate about the implications of multiplicity for algorithmic decision-making.

With this tutorial, we hope to have raised awareness of different perspectives on multiplicity and their connection to broader discussions in the FAccT community.



Prakhar Ganesh



Shomik Jain



Carol Long



Afaf Taik



Hsiang Hsu



Flavio Calmon



Ashia Wilson



Kathleen Creel



Golnoosh Farnadi

Feedback Form



Connect with Us

Find Slides and Demo
Keep track of Future Events
Multiplicity Reading List

