Soc306: Machine Learning with Social Data
Week 12: Interpretability

Brandon Stewart[1]

Princeton

April 14–16, 2025

[1]Huge thanks to Zenobia Chan for permission to include many of her slides here.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - fairness and racial justice

Where We've Been and Where We're Going...

- Last Week
  - fairness and racial justice
- This Week
  - interpretability
  - computing topic of choice!

Where We've Been and Where We're Going...

- Last Week
  - fairness and racial justice
- This Week
  - interpretability
  - computing topic of choice!
- Next Week
  - machine learning in practice and policy

Where We've Been and Where We're Going...

- Last Week
  - fairness and racial justice
- This Week
  - interpretability
  - computing topic of choice!
- Next Week
  - machine learning in practice and policy
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

1. Interpretability
2. Interpretability vs. Explainability
3. Interpretable Algorithmic Forensics
4. Open vs. Closed LLMs
5. Preview: (Ir)reproducible machine learning

Three Broad Categories of Values

- Privacy

  what information is exposed?
- Fairness

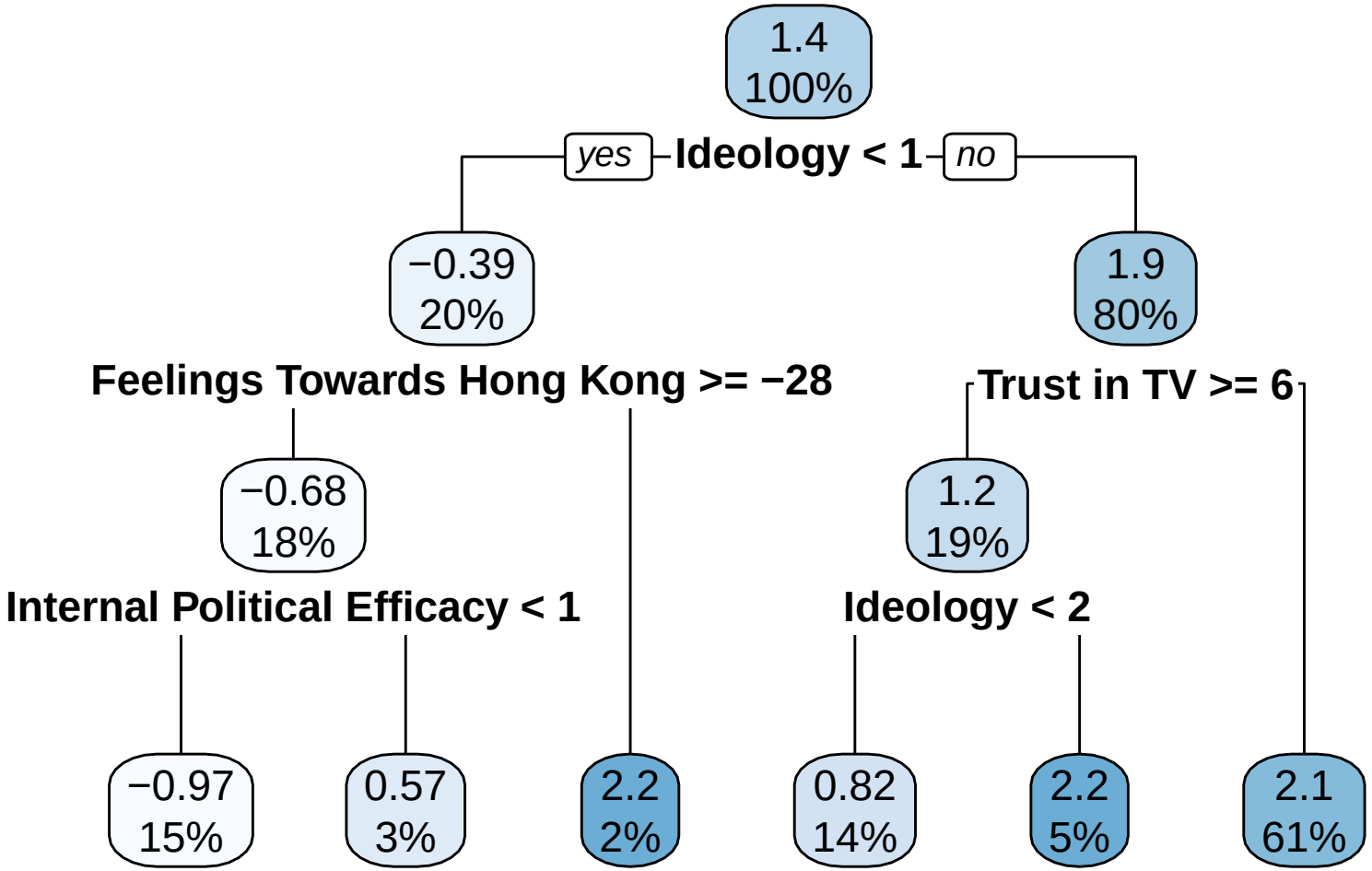
  are gains and costs distributed equitably?
- Interpretability

  do we know what the machine algorithm is doing?

What variables predict the ideology of Estonians' vote choice?

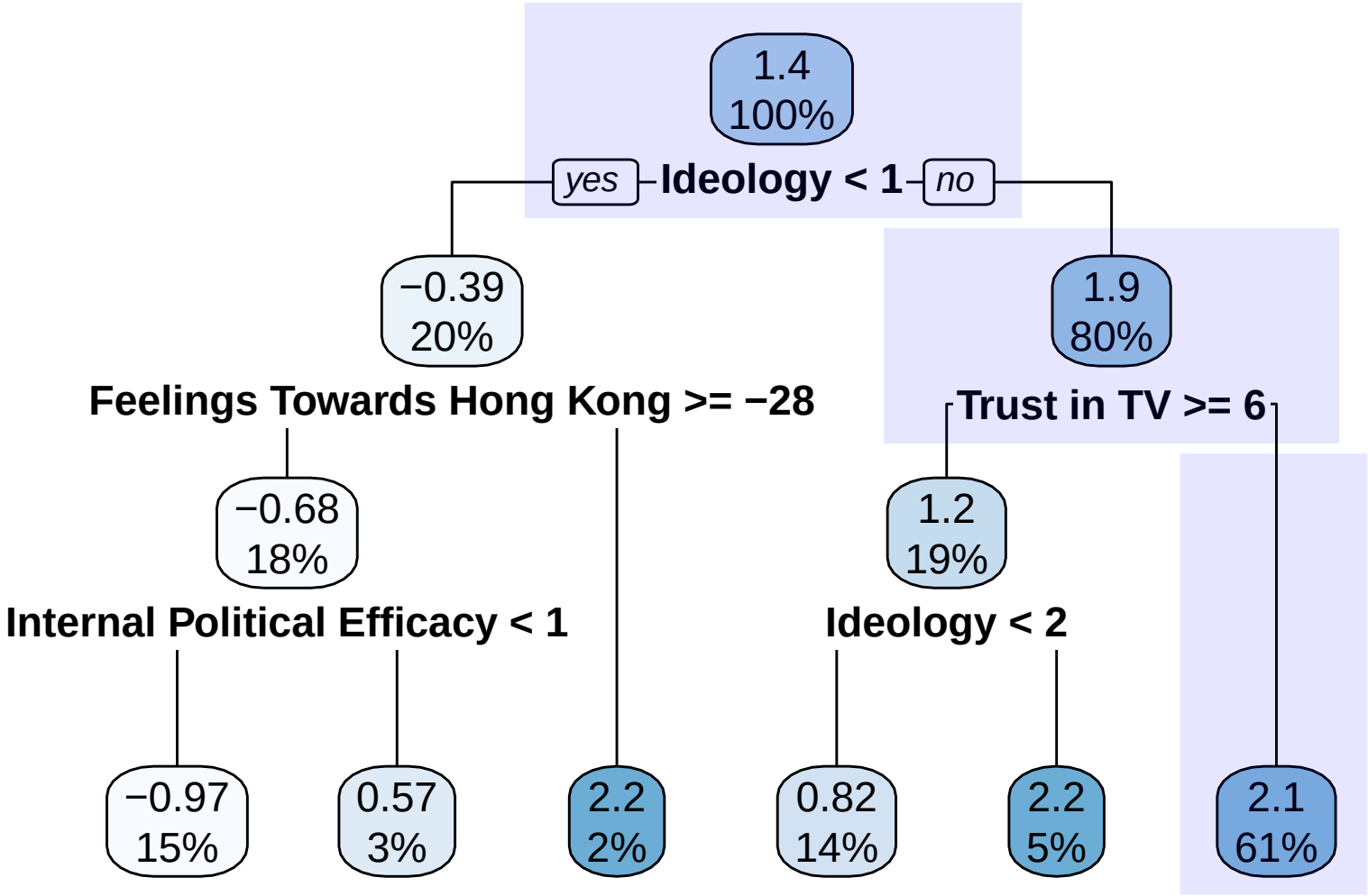
Slide Credit: Zenobia Chan

CART (Classification And Regression Trees)



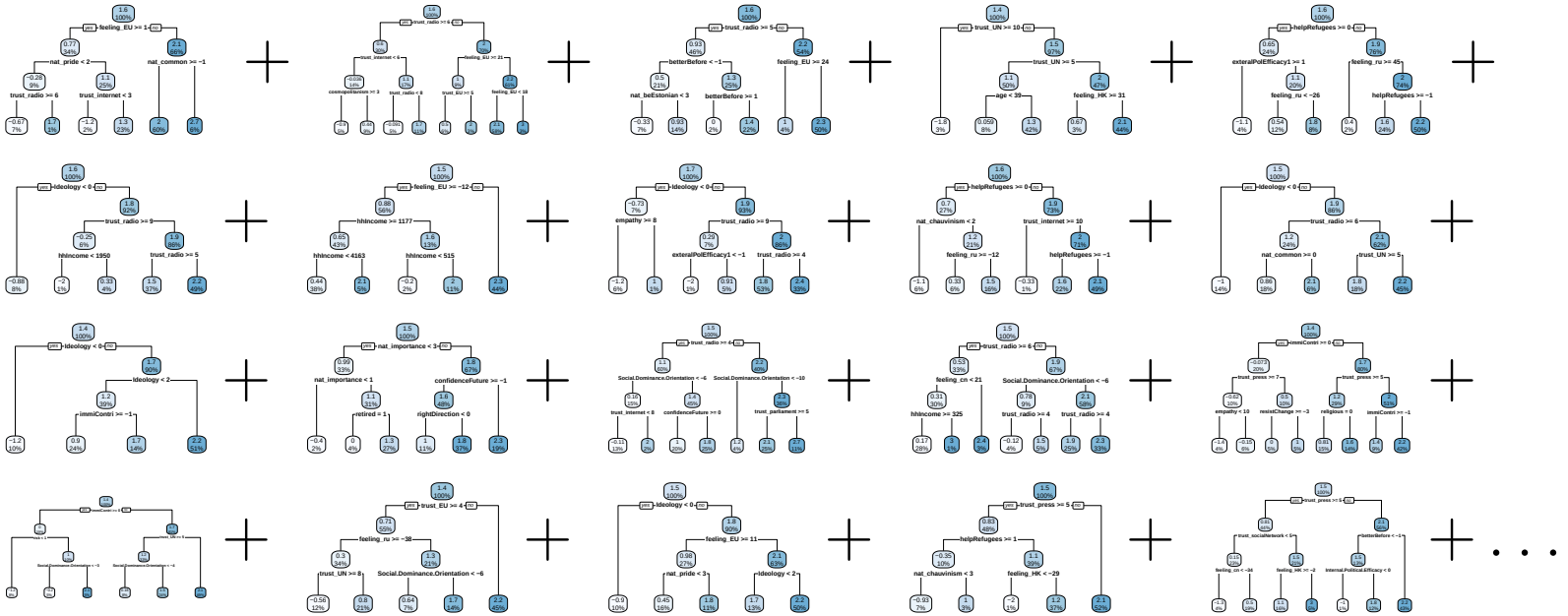
Slide Credit: Zenobia Chan

CART (Classification And Regression Trees)



Slide Credit: Zenobia Chan

From Trees to Forests

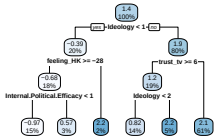


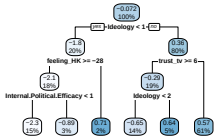
Random Forest = Trees + Bagging + Randomizing Variables

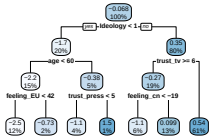
Slide Credit: Zenobia Chan

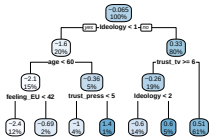
From Trees to Boosted Trees

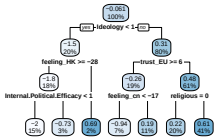
Initialize $\widehat{Y}^0 = \overline{Y}$; λ a small number

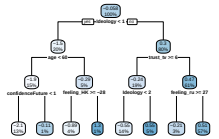
$$\widehat{Y}^1 = \widehat{Y}^0 + \lambda \times$$ 

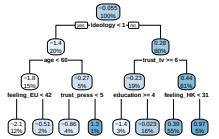
$$\widehat{Y}^2 = \widehat{Y}^1 + \lambda \times$$ 

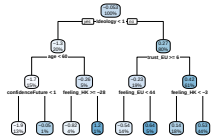
$$\widehat{Y}^3 = \widehat{Y}^2 + \lambda \times$$ 

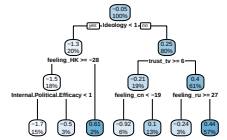
$$\widehat{Y}^4 = \widehat{Y}^3 + \lambda \times$$ 

$$\widehat{Y}^5 = \widehat{Y}^4 + \lambda \times$$ 

$$\widehat{Y}^6 = \widehat{Y}^5 + \lambda \times$$ 

$$\widehat{Y}^7 = \widehat{Y}^6 + \lambda \times$$ 

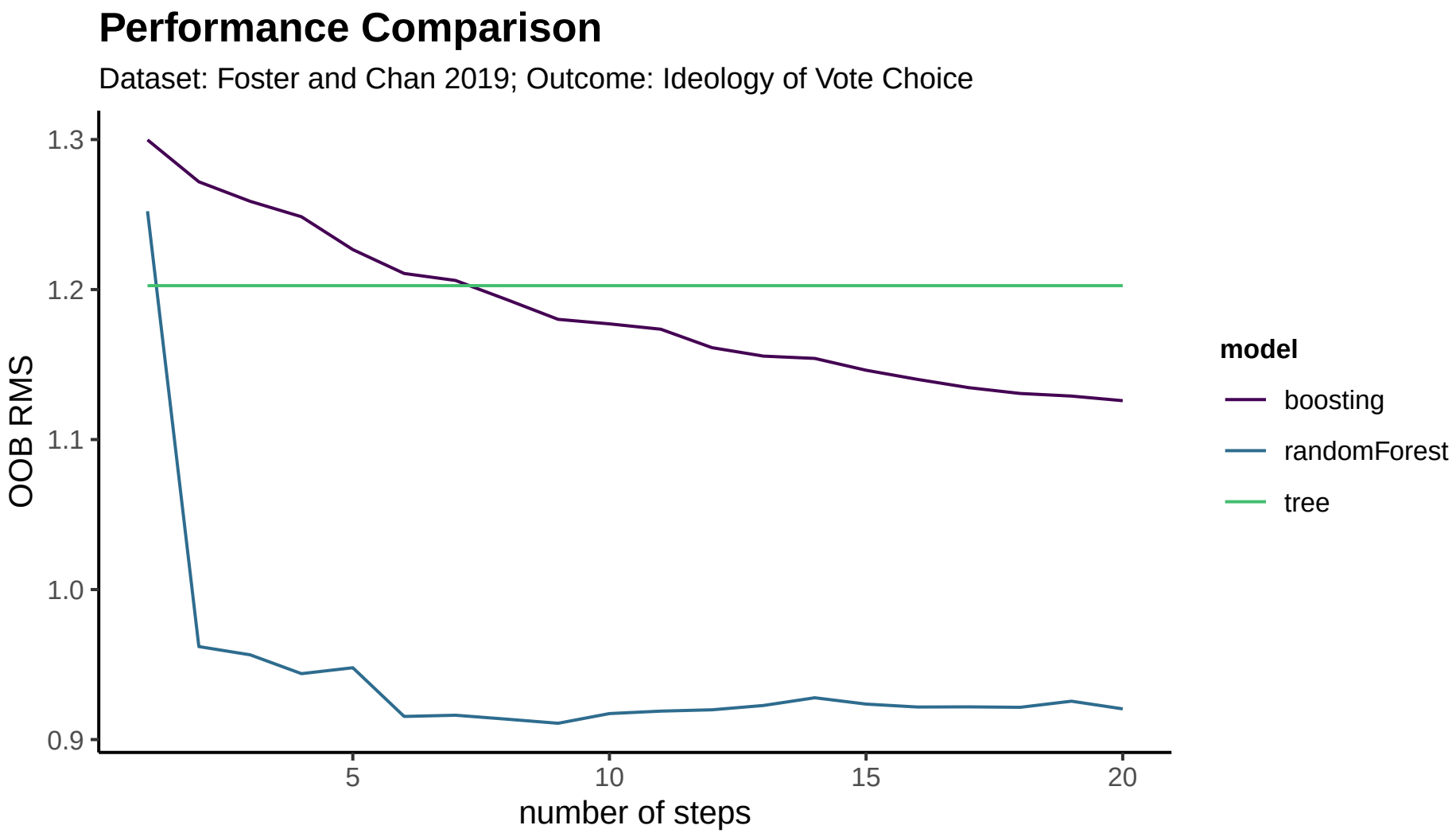
$$\widehat{Y}^8 = \widehat{Y}^7 + \lambda \times$$ 

$$\widehat{Y}^9 = \widehat{Y}^8 + \lambda \times$$ 

$$\ldots$$

Slide Credit: Zenobia Chan

MSE Goes Down...



...but so does interpretability!

Slide Credit: Zenobia Chan

Why do we need interpretability?

Why do we need interpretability?

1. Legal requirements

Why do we need interpretability?

1. Legal requirements

Why do we need interpretability?

1. Legal requirements
2. Build trust for human decision makers

Why do we need interpretability?

1. Legal requirements
2. Build trust for human decision makers
3. Anticipate behavior in new situations

Why do we need interpretability?

1. Legal requirements
2. Build trust for human decision makers
3. Anticipate behavior in new situations
4. Help to monitor for other value problems: safety, ethics, privacy etc.

Why do we need interpretability?

1. Legal requirements
2. Build trust for human decision makers
3. Anticipate behavior in new situations
4. Help to monitor for other value problems: safety, ethics, privacy etc.

$\rightsquigarrow$ Incompleteness in the problem formalization

Slide Credit: Zenobia Chan

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Use cases

1. Scientific understanding

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Use cases

1. Scientific understanding
2. Safety

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Use cases

1. Scientific understanding
2. Safety
3. Ethics

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Use cases

1. Scientific understanding
2. Safety
3. Ethics
4. Mismatched objectives

Interpretability can assist in qualitatively ascertaining other important desiderata of ML systems

Fairness, privacy, reliability, robustness, causality, usability, trust...

Use cases

1. Scientific understanding
2. Safety
3. Ethics
4. Mismatched objectives
5. Multi-objective trade-offs

Slide Credit: Zenobia Chan

How should we evaluate interpretability? (Doshi-Velez and Kim 2017)

more specific and costly

less

Evaluation		End Goal (s)	Real Human Involved	Task
Application-grounded	Whether it results in better identification of errors, new facts, or less discrimination	Yes	Yes	Real
Human-grounded	Whether humans and models yield similar results	No	Yes	Simple
Functionality-grounded	Whether the models are known to be interpretable	No	No	Proxy

Slide Credit: Zenobia Chan

How should we evaluate interpretability? (Doshi-Velez and Kim 2017)

more specific and costly less

Evaluation		End Goal (s)	Real Human Involved	Task
Application-grounded	Whether it results in better identification of errors, new facts, or less discrimination	Yes	Yes	Real
Human-grounded	Whether humans and models yield similar results	No	Yes	Simple
Functionality-grounded	Whether the models are known to be interpretable	No	No	Proxy

Slide Credit: Zenobia Chan

How should we evaluate interpretability? (Doshi-Velez and Kim 2017)

more specific and costly less

Evaluation		End Goal (s)	Real Human Involved	Task
Application-grounded	Whether it results in better identification of errors, new facts, or less discrimination	Yes	Yes	Real
Human-grounded	Whether humans and models yield similar results	No	Yes	Simple
Functionality-grounded	Whether the models are known to be interpretable	No	No	Proxy

Slide Credit: Zenobia Chan

How should we evaluate interpretability? (Doshi-Velez and Kim 2017)

more specific and costly / less

Evaluation		End Goal (s)	Real Human Involved	Task
Application-grounded	Whether it results in better identification of errors, new facts, or less discrimination	Yes	Yes	Real
Human-grounded	Whether humans and models yield similar results	No	Yes	Simple
Functionality-grounded	Whether the models are known to be interpretable	No	No	Proxy

Slide Credit: Zenobia Chan

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

Definitions

Black-box model: a function that is too complicated for human comprehension; and/or proprietary

Definitions

Black-box model: a function that is too complicated for human comprehension; and/or proprietary

Interpretability

- Domain-specific notion $\implies$ No all-purpose definition
- Common characteristics of interpretable models:
  1. Constrained in model form s.t. it is useful
  2. Obeys structural knowledge of the domain

Definitions

Black-box model: a function that is too complicated for human comprehension; and/or proprietary

Interpretability

- Domain-specific notion $\implies$ No all-purpose definition
- Common characteristics of interpretable models:
  1. Constrained in model form s.t. it is useful
  2. Obeys structural knowledge of the domain

Explainability

- Understanding how a model works
- NOT understanding how the world works

Slide Credit: Zenobia Chan

Key Issues with Explainable ML

Key Issues with Explainable ML

1. There's no necessary trade-off between accuracy

Key Issues with Explainable ML

1. There's no necessary trade-off between accuracy
2. Explanations from explainable ML models are not faithful to what the original model computes

Key Issues with Explainable ML

1. There's no necessary trade-off between accuracy
2. Explanations from explainable ML models are not faithful to what the original model computes
3. Explanations are sometimes not sensible and lack details for understanding the black box

Key Issues with Explainable ML

1. There's no necessary trade-off between accuracy
2. Explanations from explainable ML models are not faithful to what the original model computes
3. Explanations are sometimes not sensible and lack details for understanding the black box
4. Black-box models make it difficult to combine with information additional to what is in the dataset

Key Issues with Explainable ML

1. There's no necessary trade-off between accuracy
2. Explanations from explainable ML models are not faithful to what the original model computes
3. Explanations are sometimes not sensible and lack details for understanding the black box
4. Black-box models make it difficult to combine with information additional to what is in the dataset
5. Black-box models are prone to errors in training data

Slide Credit: Zenobia Chan

Fig. 2 | Saliency does not explain anything except where the network is looking. We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofan Chen, Duke University

Key Issues with Interpretable ML

1. Black-box models allow firms to profit

Key Issues with Interpretable ML

1. Black-box models allow firms to profit
2. Interpretable models create opportunities for gaming the system

Key Issues with Interpretable ML

1. Black-box models allow firms to profit
2. Interpretable models create opportunities for gaming the system
3. Interpretable models can be difficult and costly to create
   - High-stake decisions:
     Costs of analysts + computation $<$ Costs of wrong decision

Key Issues with Interpretable ML

1. Black-box models allow firms to profit
2. Interpretable models create opportunities for gaming the system
3. Interpretable models can be difficult and costly to create
   - High-stake decisions:
     Costs of analysts + computation $<$ Costs of wrong decision
4. Black-box models seem to uncover "hidden patterns"

Slide Credit: Zenobia Chan

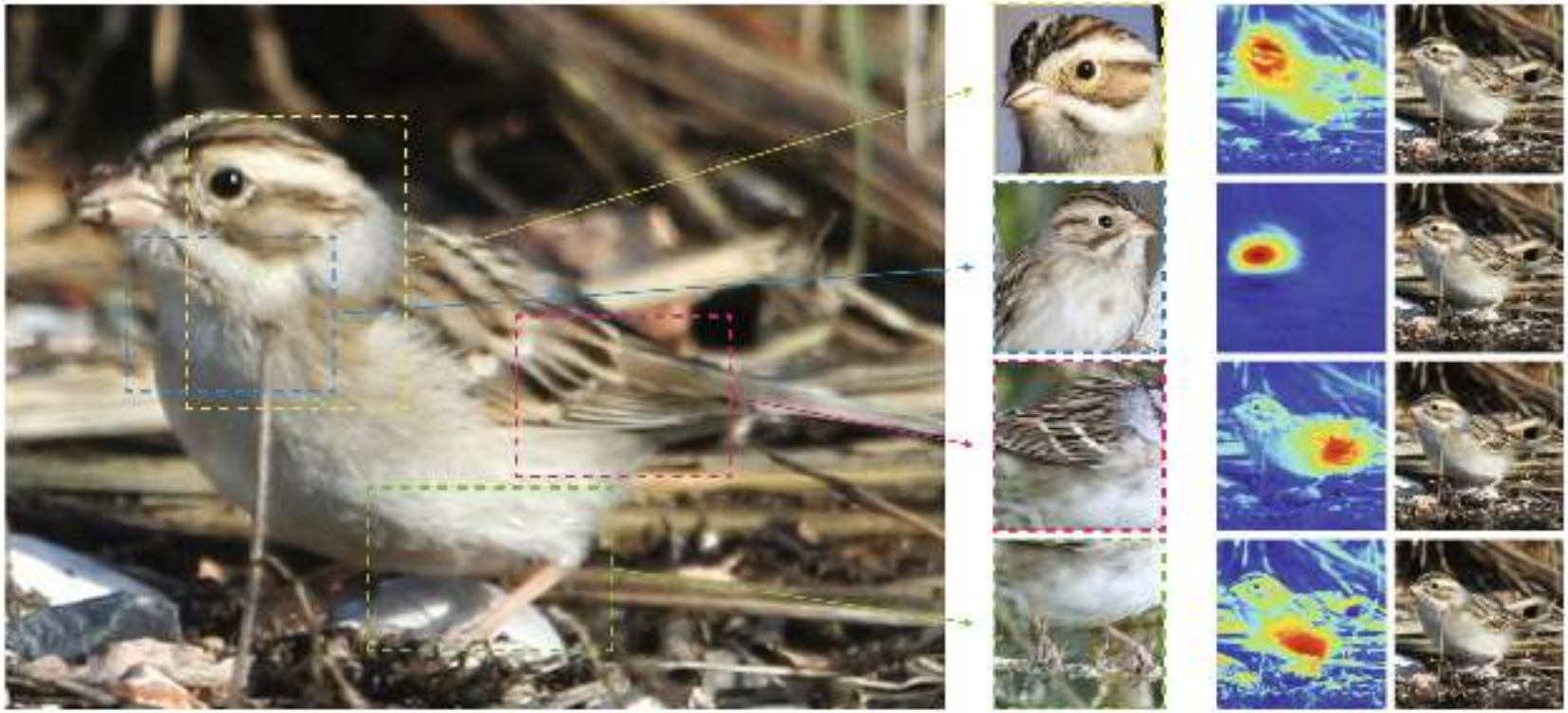


Fig. 3 | Image from the authors of ref. [18], indicating that parts of the test image on the left are similar to prototypical parts of training examples. The test image to be classified is on the left, the most similar prototypes are in the middle column, and the heatmaps that show which part of the test image is similar to the prototype are on the right. We included copies of the test image on the right so that it is easier to see to what part of the bird the heatmaps are referring. The similarities of the prototypes to the test image are what determine the predicted class label of the image. Here, the image is predicted to be a clay-coloured sparrow. The top prototype seems to be comparing the bird's head to a prototypical head of a clay-coloured sparrow, the second prototype considers the throat of the bird, the third looks at feathers, and the last seems to consider the abdomen and leg. Credit: Image constructed by Alina Barnett, Duke University

TECHNOLOGY

We Need Transparency in Algorithms, But Too Much Can Backfire

by Kartik Hosanagar and Vivian Jair

July 23, 2018 **UPDATED** July 25, 2018

A student revolt...

Too much transparency can indeed backfire

Enter Kizilcec (2016)

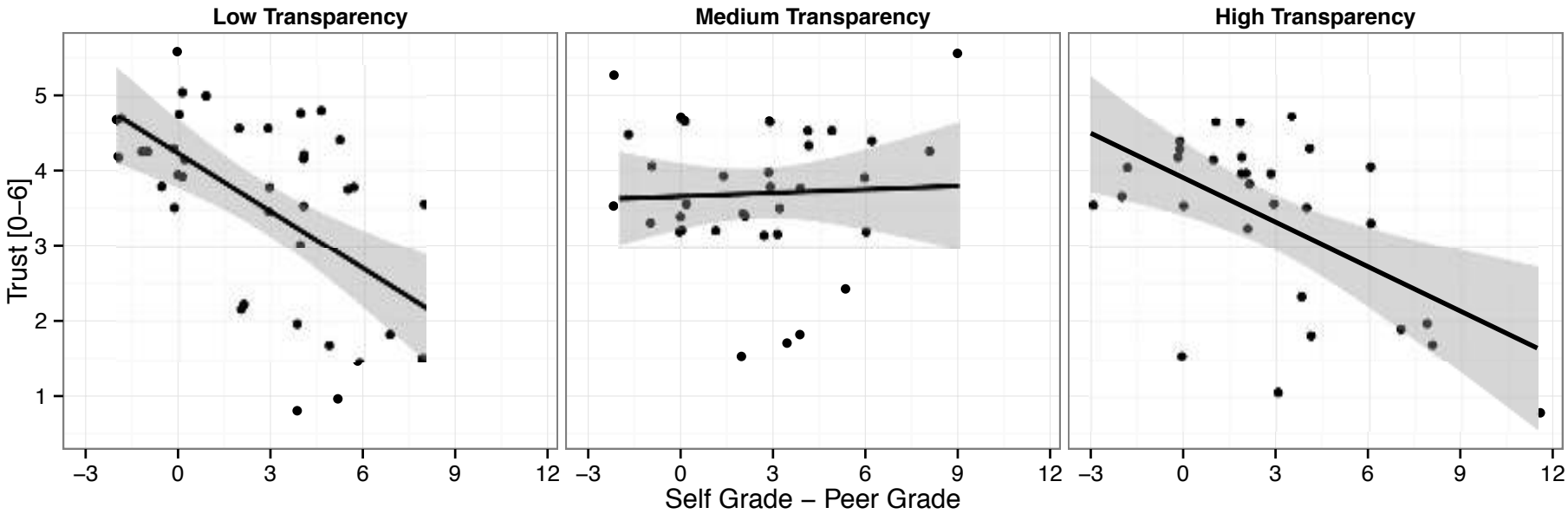


Figure 2. Trust examined as a function of the difference in expected (self) grade and received (peer) grade in each randomly assigned transparency condition. Points are jittered for presentation; OLS regression lines with 95% confidence bounds.

Introduction to Interpretable AI in Forensics

- Topic: Importance of interpretable AI in criminal justice.

Introduction to Interpretable AI in Forensics

- Topic: Importance of interpretable AI in criminal justice.
- Context: Increasing use of black box AI which lacks transparency.

Introduction to Interpretable AI in Forensics

- Topic: Importance of interpretable AI in criminal justice.
- Context: Increasing use of black box AI which lacks transparency.
- Issue: Such approaches can obscure forensic evidence and violate constitutional rights.

Problems with Black Box AI

- Definition: Black box AI are systems whose workings are not understandable.

Problems with Black Box AI

- Definition: Black box AI are systems whose workings are not understandable.
- Examples: DNA mixture interpretation, facial recognition, risk assessments.

Problems with Black Box AI

- Definition: Black box AI are systems whose workings are not understandable.
- Examples: DNA mixture interpretation, facial recognition, risk assessments.
- Consequences: Risks to public safety and legal rights due to non-transparency.

Advantages of Glass Box AI

- Definition: Glass box AI is designed to be transparent and interpretable.

Advantages of Glass Box AI

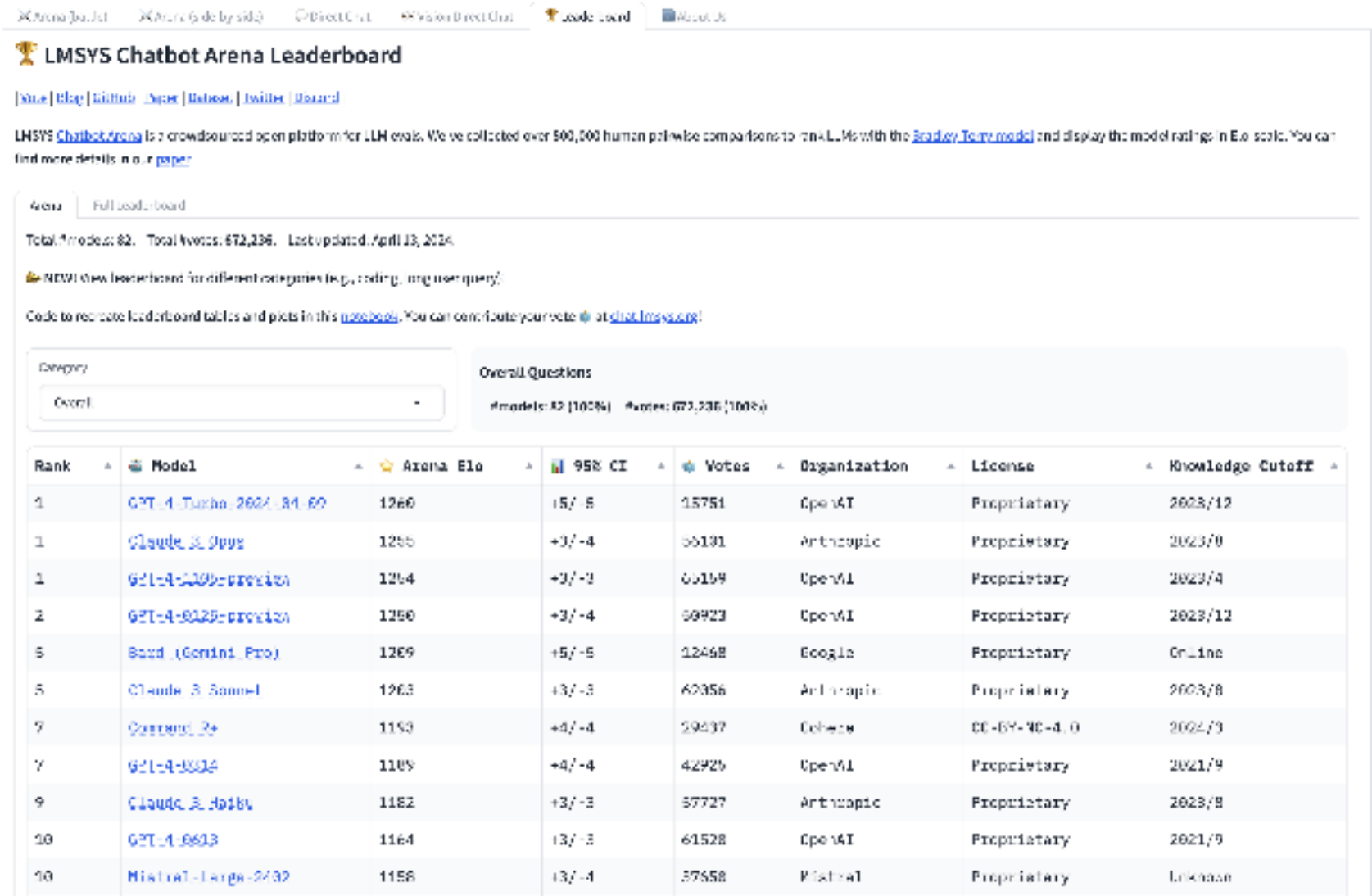
- Definition: Glass box AI is designed to be transparent and interpretable.
- Research Findings: Shows higher accuracy than black box AI.

Advantages of Glass Box AI

- Definition: Glass box AI is designed to be transparent and interpretable.
- Research Findings: Shows higher accuracy than black box AI.
- Implications: Supports due process and safeguards defendant's rights.

Key Premise

The Black Box Performance Myth



LMSYS Chatbot Arena Leaderboard

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	55131	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	58923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	60356	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29637	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42926	OpenAI	Proprietary	2021/9
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/8
10	GPT-4-0613	1164	+3/-3	61528	OpenAI	Proprietary	2021/9
10	Mistral-Large-2402	1158	+3/-4	37558	Mistral	Proprietary	Unknown

Legislative and Judicial Response

- Current Trends: Absence of required transparency in AI use in courts.

Legislative and Judicial Response

- Current Trends: Absence of required transparency in AI use in courts.
- Proposed Changes: Legislation to mandate interpretable AI in investigations.

Legislative and Judicial Response

- Current Trends: Absence of required transparency in AI use in courts.
- Proposed Changes: Legislation to mandate interpretable AI in investigations.
- Judicial Rulings: Enforcing rights to interpretable forensic AI.

Case Studies and Legal Implications

- Case Studies: Instances where black box AI was legally challenged.

Case Studies and Legal Implications

- Case Studies: Instances where black box AI was legally challenged.
- Legal Considerations: Need for scrutinizable AI systems.

Case Studies and Legal Implications

- Case Studies: Instances where black box AI was legally challenged.
- Legal Considerations: Need for scrutinizable AI systems.
- Future Outlook: Potential for action to promote interpretable AI in forensics.

Conclusion

- Summary: Importance of shifting from black box to glass box AI in justice.

Conclusion

- Summary: Importance of shifting from black box to glass box AI in justice.
- Call to Action: Advocacy for policies that prioritize transparency.

Conclusion

- Summary: Importance of shifting from black box to glass box AI in justice.
- Call to Action: Advocacy for policies that prioritize transparency.
- Goal: Ensuring AI aligns with constitutional protections and public safety.

Leakage and the Reproducibility Crisis in ML-based Science

Sayash Kapoor [1] Arvind Narayanan [1]

Abstract

The use of machine learning (ML) methods for prediction and forecasting has become widespread across the quantitative sciences. However, there are many known methodological pitfalls, including data leakage, in ML-based science. In this paper, we systematically investigate reproducibility issues in ML-based science. We show that data leakage is indeed a widespread problem and has led to severe reproducibility failures. Specifically, through a survey of literature in research communities that adopted ML methods, we find 17 fields where errors have been found, collectively affecting 329 papers and in some cases leading to wildly overoptimistic conclusions. Based on our survey, we present a fine-grained taxonomy of 8 types of leakage that range from textbook errors to open research problems.

We argue for fundamental methodological changes to ML-based science so that cases of leakage can be caught before publication. To that end, we propose model info sheets for reporting scientific claims based on ML models that would address all types of leakage identified in our survey. To investigate the impact of reproducibility errors and the efficacy of model info sheets, we undertake a reproducibility study in a field where complex ML models are believed to vastly outperform older statistical models such as Logistic Regression (LR): civil war prediction. We find that all papers claiming the superior performance of complex ML models compared to LR models fail to reproduce due to data leakage, and complex ML models don't perform substantively better than decades-old LR models. While none of these errors could have been caught by reading the papers, model info sheets would enable the detection of leakage in each case.

[1]Department of Computer Science and Center for Information Technology Policy, Princeton University. Correspondence to: Sayash Kapoor <sayashk@princeton.edu>.

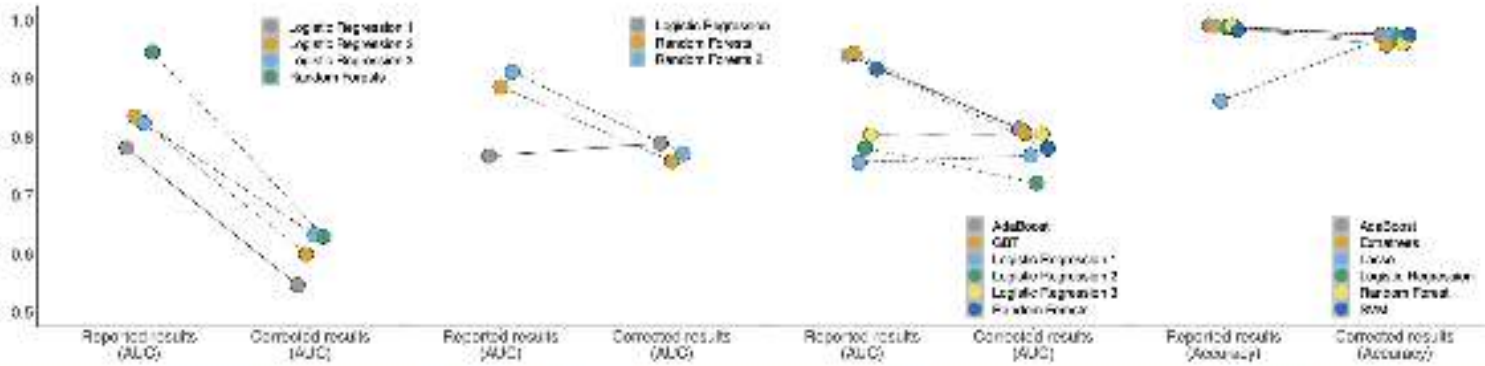
1. Overview

There has been a marked shift towards the paradigm of predictive modeling across quantitative science fields. This shift has been facilitated by the widespread use of machine learning (ML) methods. However, pitfalls in using ML methods have led to exaggerated claims about their performance. Such errors can lead to a feedback loop of overoptimism about the paradigm of prediction – especially as non-replicable publications tend to be cited more often than replicable ones (Serra-Garcia & Gneezy, 2021). It is therefore important to examine the reproducibility of findings in communities adopting ML methods.

Scope. We focus on reproducibility issues in ML-based science, which involves making a scientific claim using the performance of an ML model as evidence. There is a much better known reproducibility crisis in research that uses traditional statistical methods (Open Science Collaboration, 2015). We also situate our work in contrast to other ML domains, such as methods research (creating and improving widely-applicable ML methods), ethics research (studying the ethical implications of ML methods), engineering applications (building or improving a product or service), and modeling contests (improving predictive performance on a fixed dataset created by an independent third party). Investigating the validity of claims in all of these areas is important, and there is ongoing work to address reproducibility issues in these domains (Hullman et al., 2022; Pineau et al., 2020; Erik Gundersen, 2021; Bell & Kampman, 2021).

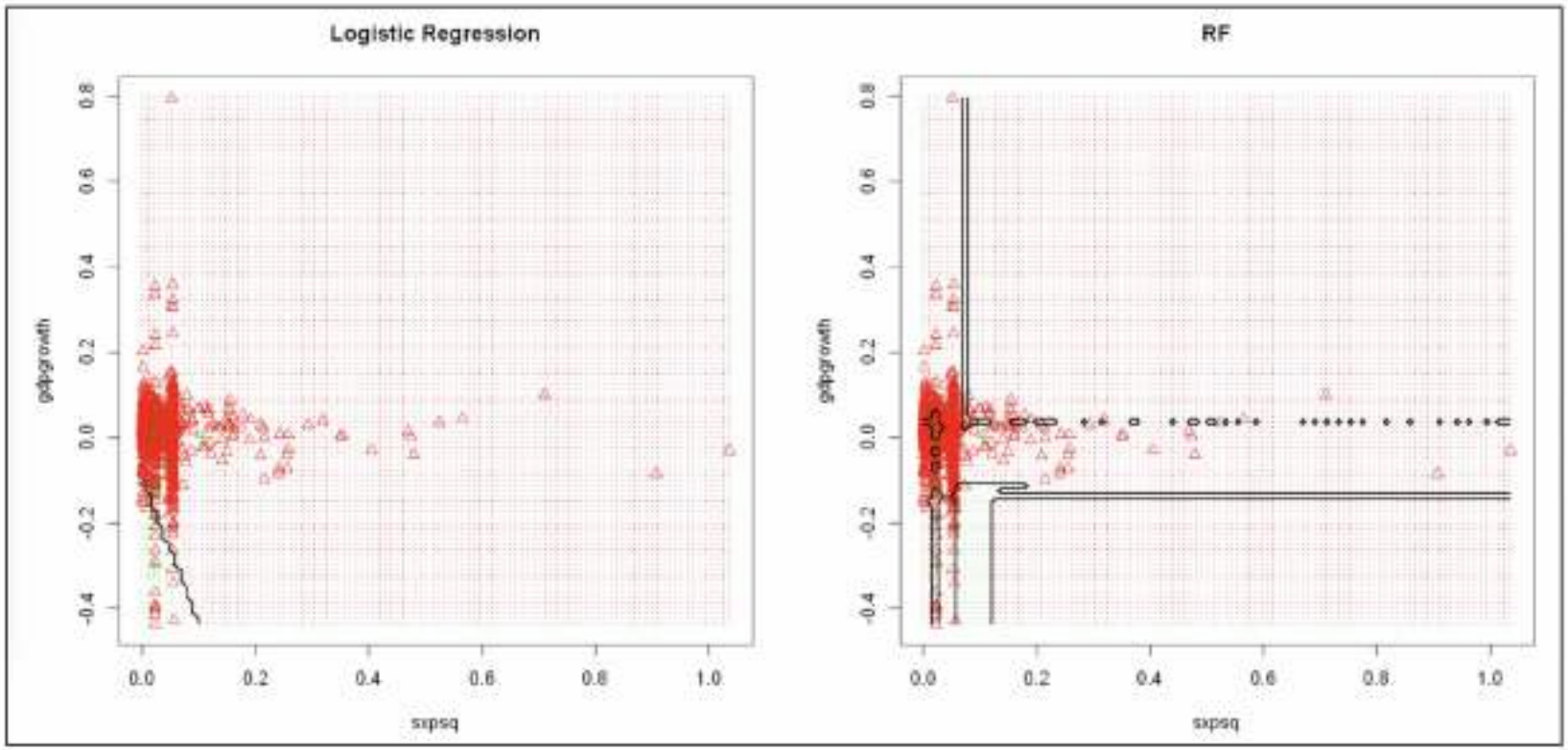
We define a research finding as reproducible if the code and data used to obtain the finding are available and the data is correctly analyzed (Hofman et al., 2021a; Leek & Peng, 2015; Pineau et al., 2020). This is a broader definition than computational reproducibility — when the results in a paper can be replicated using the exact code and dataset provided by the authors (see Appendix A).

Leakage. Data leakage has long been recognized as a leading cause of errors in ML applications (Nisbet et al., 2009). In formative work on leakage, Kaufman et al. (2012) provide an overview of different types of errors and give several recommendations for mitigating these errors. Since this paper was published, the ML community has investigated leakage in several engineering applications and modeling competitions (Fraser, 2016; Ghani et al., 2020; Becker, 2018; Brownlee, 2018; Collins Thompson).



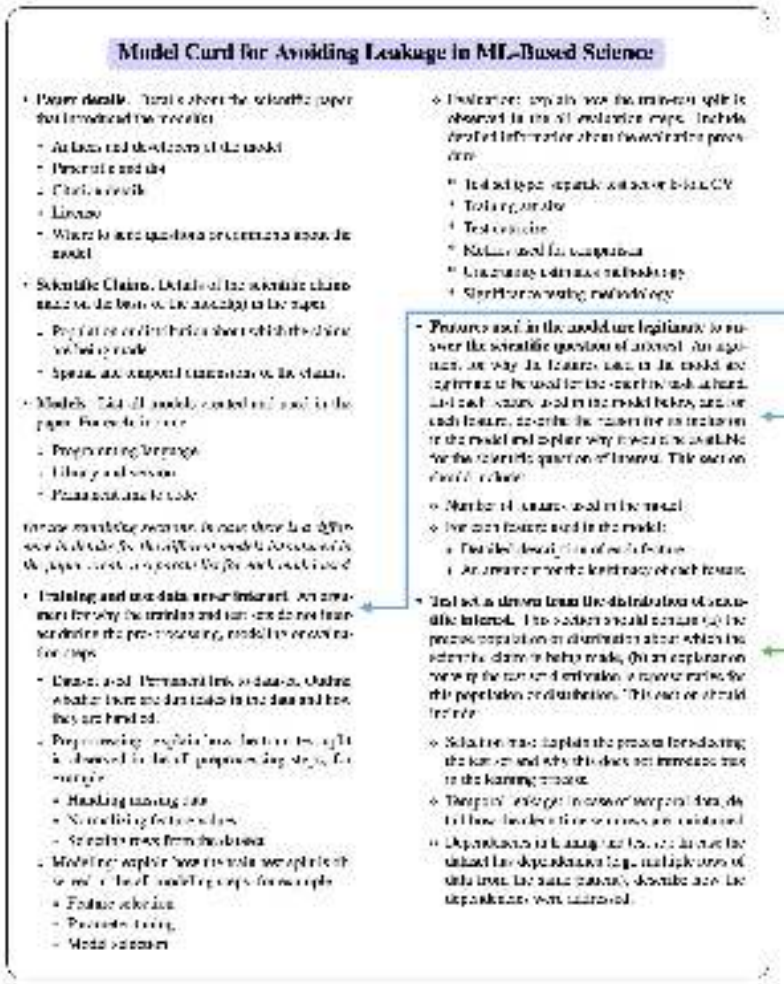
Paper	Muchlinski et al.	Colaresi and Mahmood	Wang	Kaufman et al.
Claim	Random Forests model drastically outperforms Logistic regression models	Random Forests models drastically outperform Logistic regression model	Adaboost and Gradient Boosted Trees (GBT) drastically outperform other models	Adaboost outperforms other models
Error	**[L1.2] Pre-proc. on train-test** (Incorrect imputation)	**[L1.2] Pre-proc. on train-test** (Incorrect reuse of an imputed dataset)	**[L1.2] Pre-proc. on train-test.** (Incorrect reuse of an imputed dataset) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)	**[L2] Illegitimate features** (Data leakage due to proxy variables) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)
Impact	Random Forests perform no better than Logistic Regression	Random Forests perform no better than Logistic Regression	Difference in AUC between Adaboost and Logistic Regression drops from 0.14 to 0.01	Adaboost no longer outperforms Logistic Regression. None of the models outperform a baseline model that predicts the outcome of the previous year
Discussion	Impact of the incorrect imputation is severe since 96% of the out-of-sample dataset is missing and is filled in using the incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Use several proxy variables for the outcome as predictors (e.g., *colwars*, *covwars*, *sdwars*, all proxies for civil war), leading to near perfect accuracy

Figure 1. A comparison of reported and corrected results in civil war prediction papers published in top political science journals. The main findings of each of these papers are invalid due to various forms of data leakage: Muchlinski et al. (2016) impute the training and test data together, Colaresi & Mahmood (2017) and Wang (2019) incorrectly reuse an imputed dataset, and Kaufman et al. (2019) use proxies for the target variable which causes data leakage. The use of model info sheets (Section 3) would detect leakage in every paper. When we correct these errors, complex ML models (such as Adaboost and Random Forests) do not perform substantively better than decades-old Logistic Regression models for civil war prediction in each case. Each column in the table outlines the impact of leakage on the results of a paper. The figure above each column shows the difference in performance that results from fixing leakage issues.


Logistic Regression
RF
gdpgrowth
sxpsq
0.0
0.2
0.4
0.6
0.8
1.0
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8

Field	Paper	Number of papers reviewed	Number of papers with pitfalls	**[L1.1] No test set**	**[L1.2] Pre-proc. on train-test**	**[L1.3] Feature sel. on train-test**	**[L1.4] Duplicates**	**[L2] Illegitimate features**	**[L3.1] Temporal leakage**	**[L3.2] Non-ind. b/w train-test**	**[L3.3] Sampling bias**	Comput. reproducibility issues	Data quality issues	Metric choice issues	Standard dataset used?
Medicine	Bouwmeester et al. (2012)	71	27	○									○		
Neuroimaging	Whelan & Garavan (2014)	–	14	○		○									
Autism Diagnostics	Bone et al. (2015)	–	3				○			▫			○	▫	○
Bioinformatics	Blagus & Lusa (2015)	–	6		○										
Nutrition Research	Ivanescu et al. (2016)	–	4	○									○	○	
Software Eng.	Tu et al. (2018)	58	11					○			○		○		○
Toxicology	Alves et al. (2019)	–	1				○				○		○		
Satellite Imaging	Nalepa et al. (2019)	17	17						○				○		▫
Tractography	Poulin et al. (2019)	4	2	○							○		○	▫	○
Clinical Epidem.	Christodoulou et al. (2019)	71	48			▫							○		
Brain-computer Int.	Nakanishi et al. (2020)	–	1	○											○
Histopathology	Oner et al. (2020)	–	1						○						
Neuropsychiatry	Poldrack et al. (2020)	100	53	○	○								○	▫	
Medicine	Vandewiele et al. (2021)	24	21			▫			○	○	○		○		▫
Radiology	Roberts et al. (2021)	62	62	○			○			▫	○				○
IT Operations	Lyu et al. (2021)	9	3					○							○
Medicine	Filho et al. (2021)	–	1				○								
Neuropsychiatry	Shim et al. (2021)	–	1			▫					○				
Genomics	Barnett et al. (2022)	41	23			▫							○		
Computer Security	Arp et al. (2022)	30	30	○	○	○		○		○	○		○	○	

Table 1. Survey of 20 papers that identify pitfalls in the adoption of ML methods across 17 fields, collectively affecting 329 papers. In each field, papers adopting ML methods suffer from data leakage. The column headings for types of data leakage, shown in bold, are based on our taxonomy of data leakage. We also highlight other issues that are reported in the papers, including issues with computational reproducibility (the availability of code, data, and computing environment to reproduce the exact results reported in the paper), data quality (for example, small size or large amounts of missing data), metric choice (using incorrect metrics for the task at hand, for example, using accuracy for measuring model performance in the presence of heavy class imbalance), and standard dataset use, where issues are found despite the use of standard datasets in a field.



Model Cards Template

1. Train and test sets never interact
2. Features used in the model are legitimate to answer the scientific question of interest
3. Test set is drawn from the distribution of scientific interest

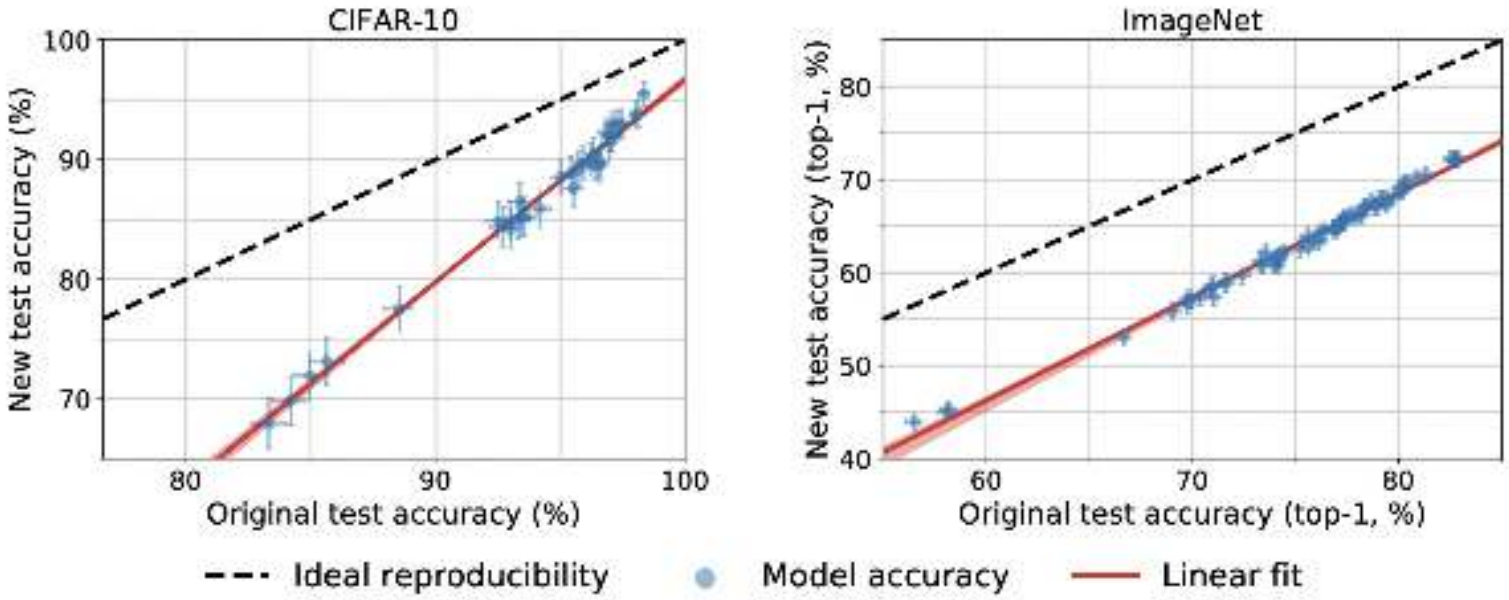


Image from: "Do ImageNet Classifiers Generalize to ImageNet?", Recht et al., ICML 2019

This Week in Review

This Week in Review

- Interpretability

This Week in Review

- Interpretability

Next Week: Machine Learning in Practice and Policy

- Sayash Kapoor and Arvind Narayanan. 2023. "Leakage and the reproducibility crisis in machine-learning-based science" *Patterns*
- Johnson and Zhang. 2022. "What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy" *FAccT '22*.
- Wang, Kapoor, Barocas, and Narayanan. 2024. "Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy" *ACM Journal of Responsible Computing*