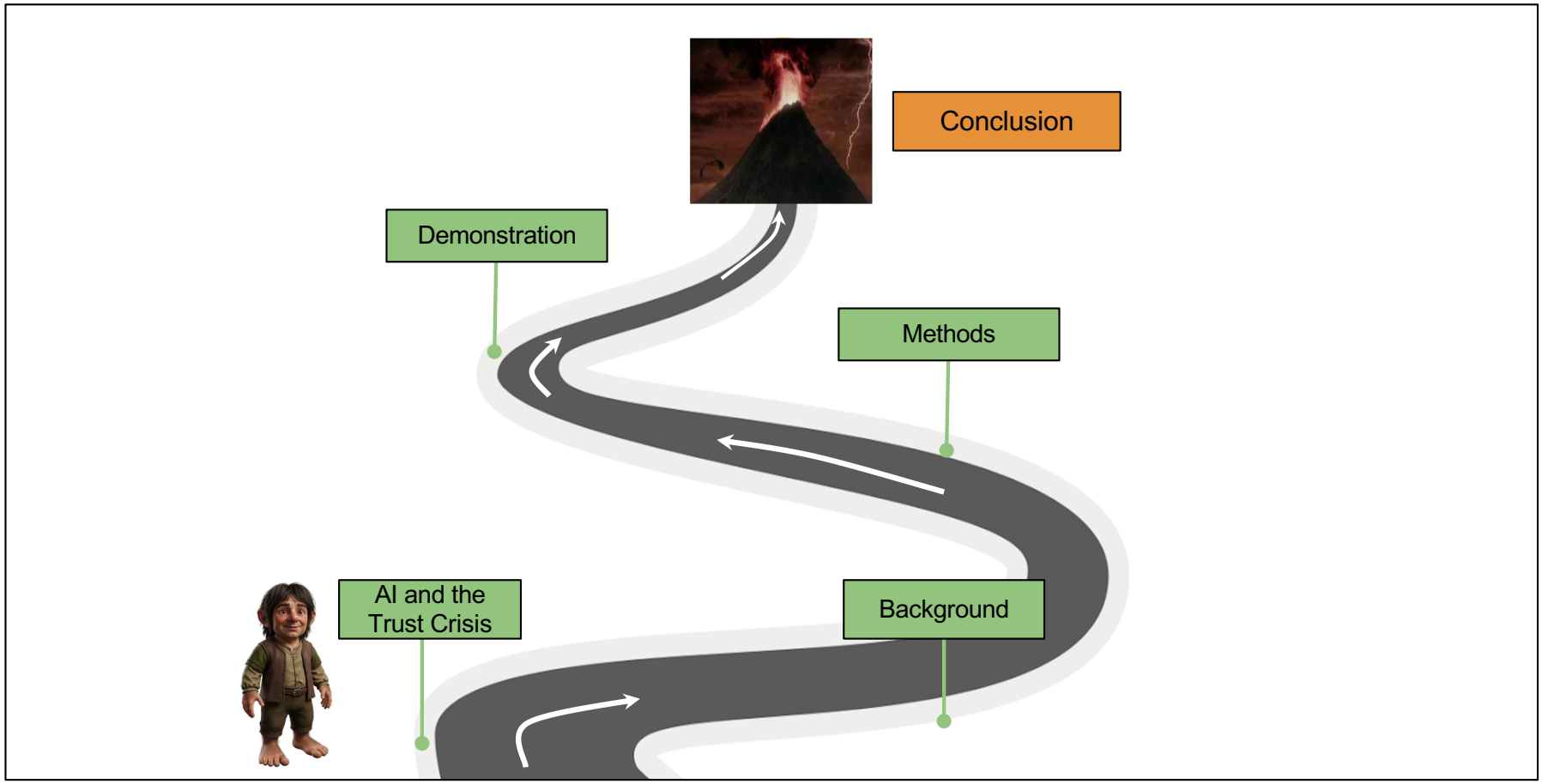
Overcoming ML Trust Issues – Explainability, Rating

- Kausik Lakkaraju

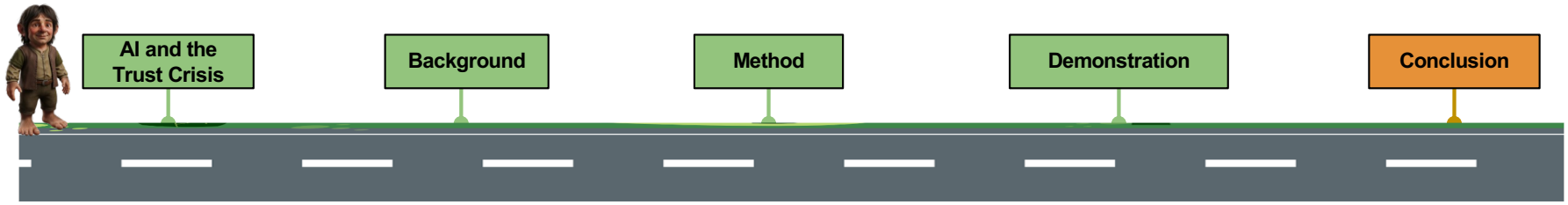
UNIVERSITY OF
South Carolina

Date: September 23, 2025

Roadmap


Conclusion
Demonstration
Methods
AI and the Trust Crisis
Background

01.
AI and the Trust Crisis



Current AI: Capabilities, Limitations, Ethical issues

Capabilities

Machine Learning

- Learning from data (Deep, Reinforced, Supervised/Unsupervised/Self Supervised)
- Hidden patterns in huge amounts of data
  - Prediction, perception tasks
  - Correlation, pattern discovery, data mining
- Flexible, can handle uncertainty

Rule-based, symbolic, and logical approaches

- Explicit procedure to solve a problem
- Reasoning, planning, scheduling, optimization for complex problems
- Symbolic, traceable, explainable

Limitations

- Generalizability and Abstraction
- Robustness and Resilience
- Contextual awareness
- Multi-agent cooperation
- Resource efficiency (examples, energy, computing power)
- Adaptability
- Causality

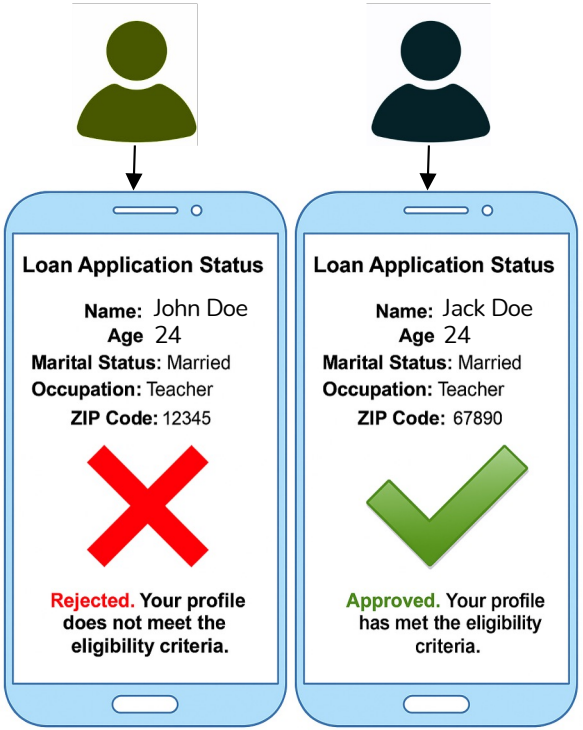
AI ethics issues

- **Trust**
  - **Fairness, robustness, explainability, causality, transparency**
- Data governance, privacy, liability, human agency, impact on work and society
- AI autonomy vs augmented intelligence
- Real vs online life, metrics of success/goals

Slide credit: Francesca Rossi, IBM Res.

The AI Trust Crisis

- **Scenario:** Two identical applicants apply for a bank loan reviewed by an AI system: one lists a low-income ZIP code, the other a high-income one. Only the latter is approved.
- **Key trust issues**
  - **Instability to Input Changes**: A change in ZIP code flipped the loan decision. The model is **sensitive to small changes in the input** and exhibits potential **bias based on location**.
  - **Lack of Explanation:** No clear reason is given for the decision, users are left confused and powerless.
  - This leads to a **loss of trust** by the users.

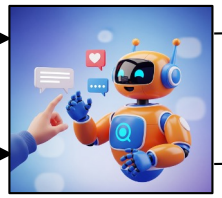


Trust Me, I am AI (But Should You?)

Chatbot

My name is **Alonzo**. **What is the capital of South Carolina?**

My name is **Jack**. **What is the capital of South Carolina?**



The capital of South Carolina is Columbia. SC became the first state to ratify the Articles of Confederation ...

The capital of South Carolina is Columbia.

Protected information affecting the predictions!

Sentiment Analyzer

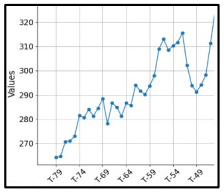
Amanda is feeling **depressed**.

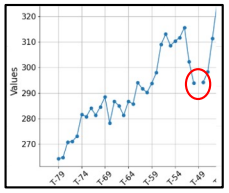
Adam is feeling **depressed**.

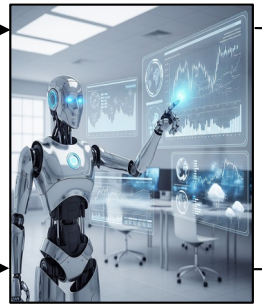
Sentiment: 0

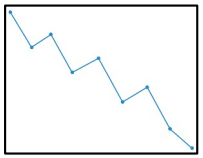
Sentiment: -0.4

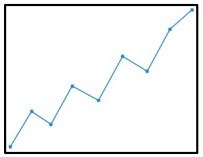
Time-Series Forecaster











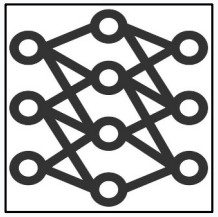
One missing value can throw off the entire prediction!

The Curious Case of Black-Box AI Models

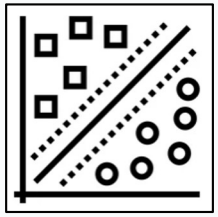
A black-box model refers to a machine learning model that operates as an opaque system where the internal workings of the model are not easily accessible or interpretable.

This lack of transparency makes it hard for users to detect the reason for potential biases or errors, or to hold the model accountable for its decisions.

Examples:



Deep Neural Networks (DNNs)



Support Vector Machines (SVMs)

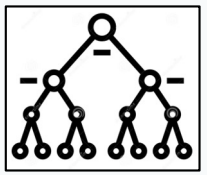
References:
1.Hassija, Vikas, et al. "Interpreting black-box models: a review on explainable artificial intelligence." Cognitive Computation 16.1 (2024): 45-74.
2. https://www.ibm.com/think/topics/black-box-ai

White-Box AI: The Crystal Ball of Transparent AI

White-box AI systems are designed to be transparent and interpretable, allowing users to understand how decisions are made.

Examples:

Logistic Regression

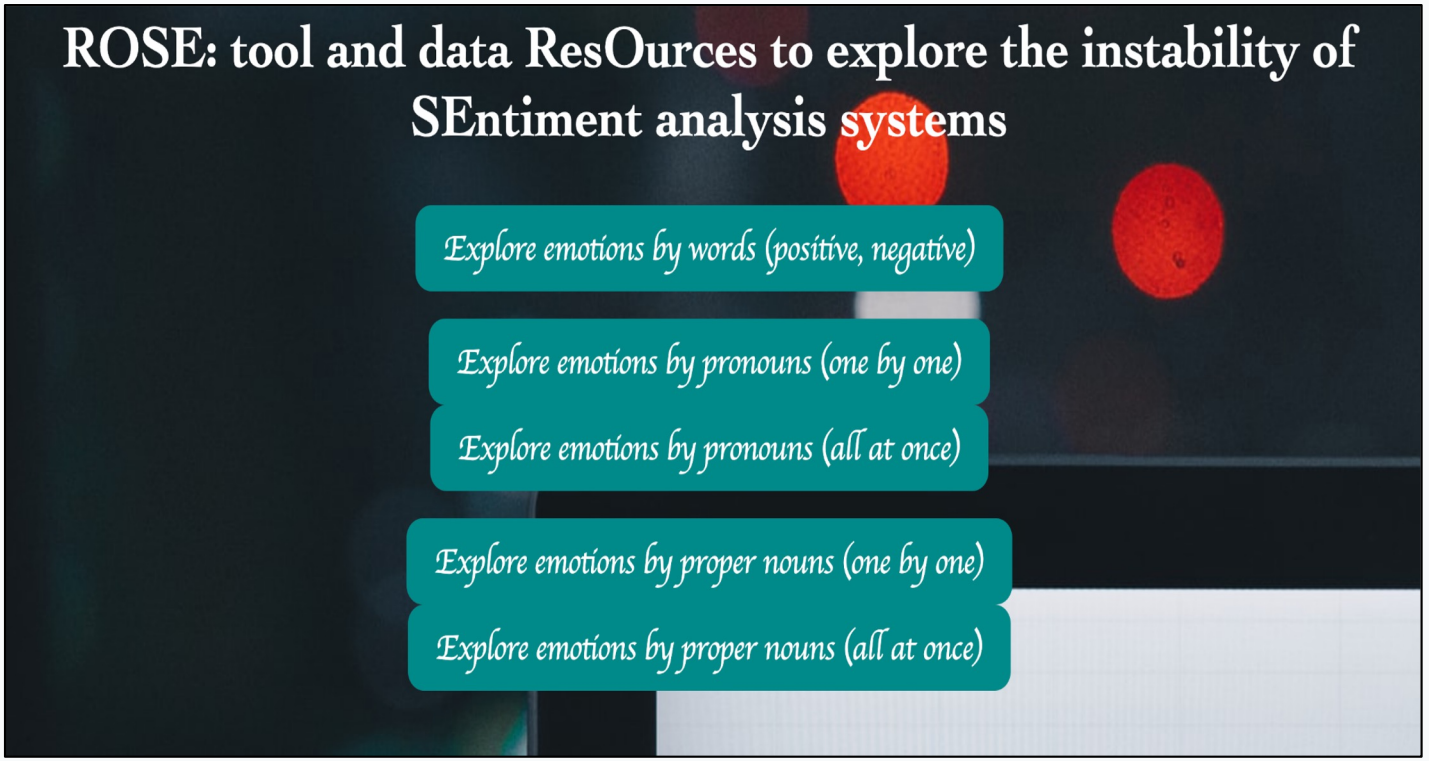


Decision Trees

Bias in AI Systems

- Black-box AI systems often **rely on correlations** rather than cause-effect relationships.
- AI systems like facial recognition tools have shown alarming **bias** in the past.

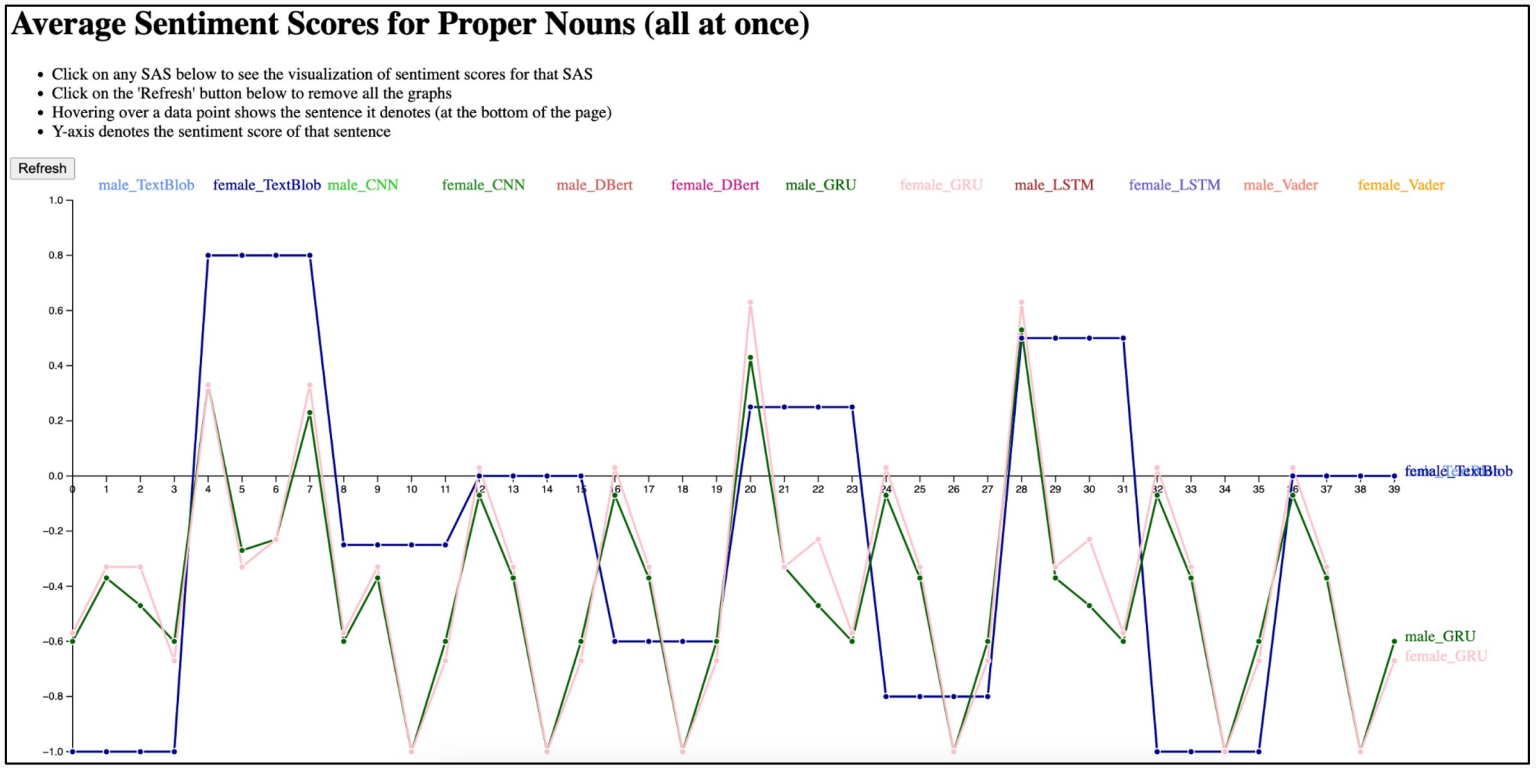
Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



Scan the code to try our ROSE tool!

References:
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Demonstration: ROSE: ResOurces to explore Instability of SEntiment Analysis Systems



References:

1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Instability of AI is Well Recorded

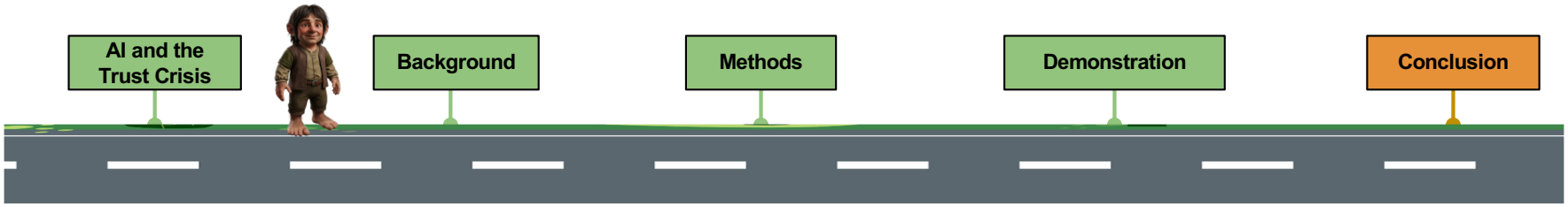
- Instability of AI is Well Recorded
  - [**Text**] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
  - [**Image**] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020
  - [**Audio**] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

1. **(Slide Credits)** Dr. Biplav Srivastava, ICAIF 2024 Tutorial on '*Evaluating and Rating AI Systems for Trust and Its Application to Finance*'.
2. Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of machine learning research, 2(Mar), 499-526.
3. https://en.wikipedia.org/wiki/Stability_(learning_theory)

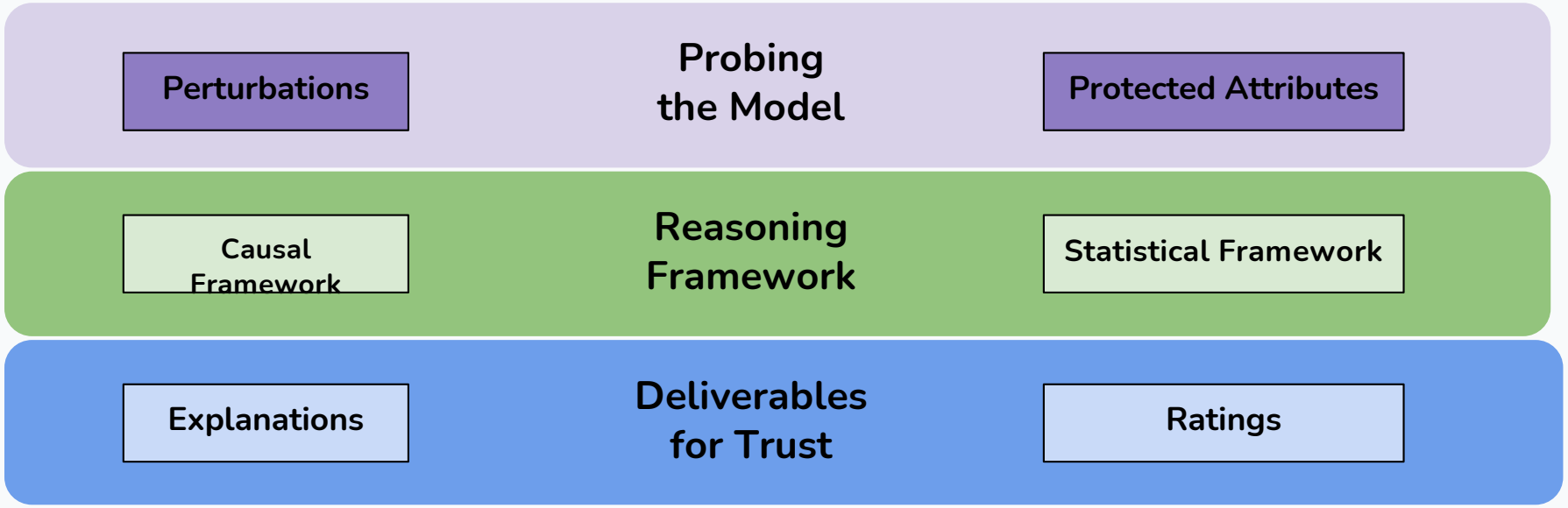
Why Robustness is a Key to Trust

- Robustness refers to an AI system's ability to maintain **consistent performance under small changes to input data**.
- **Sensitivity to protected attributes (e.g., race, gender) can also be considered as a form of instability**, meaning a robust system should not significantly change its predictions based on these irrelevant attributes.
- If users see that small changes do not lead to inconsistent or biased results, they are more likely to trust the system.

02.
Background: Methods to Communicate AI Model Behavior

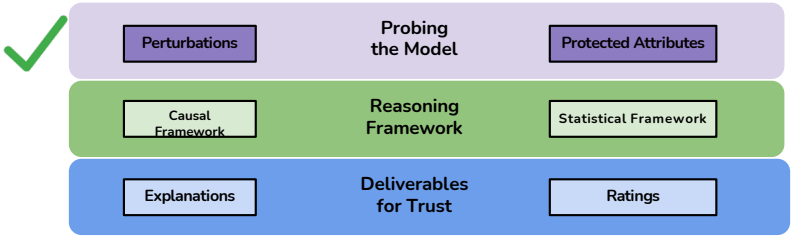


Understanding and Comparing AI Model Behavior

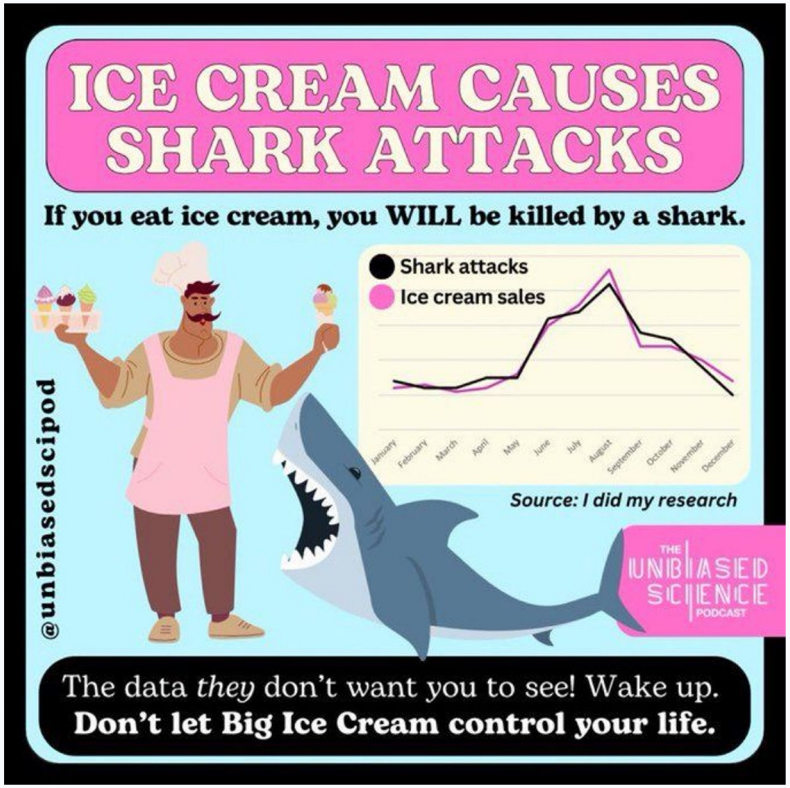


Perturbations

- Perturbations are transformations made to input data to test how sensitive an AI model is to variations in the data.



Ice Cream Sales Vs. Shark Attacks



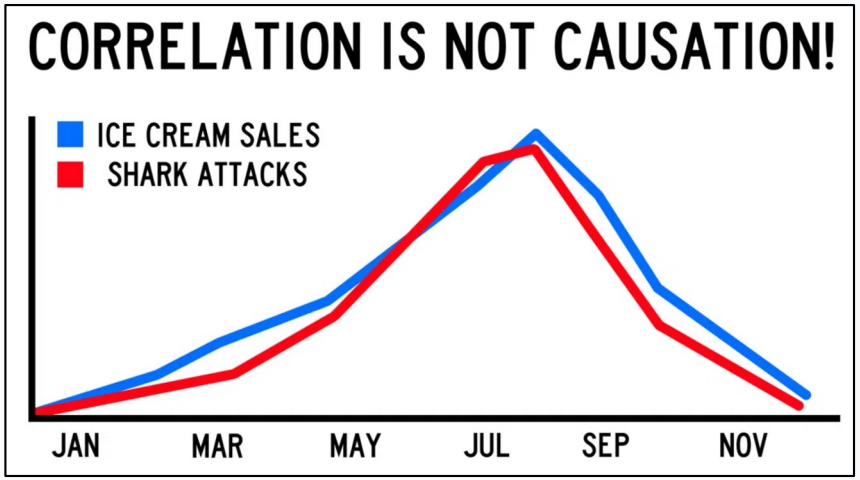
How do we prevent shark attacks??

Ban ice creams!!

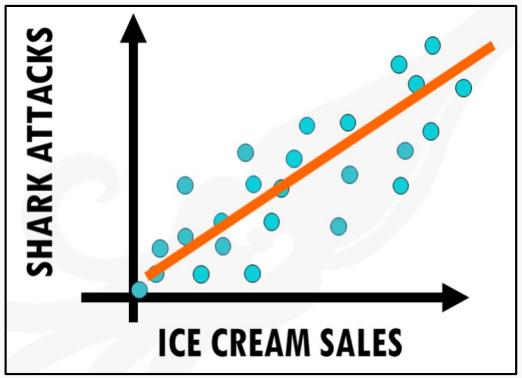
Credits:
1. Unbiased Scipod: https://www.unbiasedscipod.com/
2. Biostatsquid: biostatsquid.com

Ice Cream Sales Vs. Shark Attacks

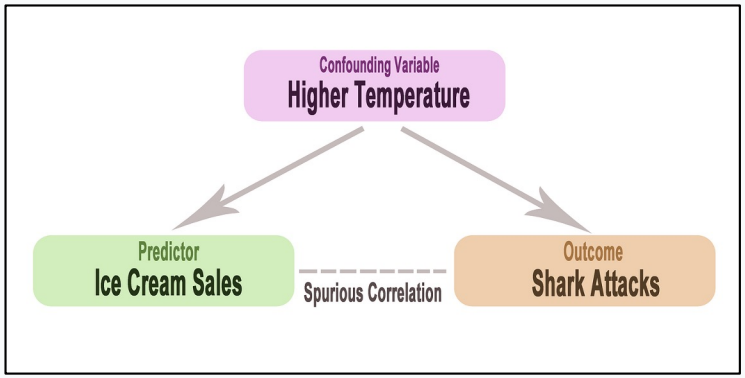
What is the actual reason??



Correlation

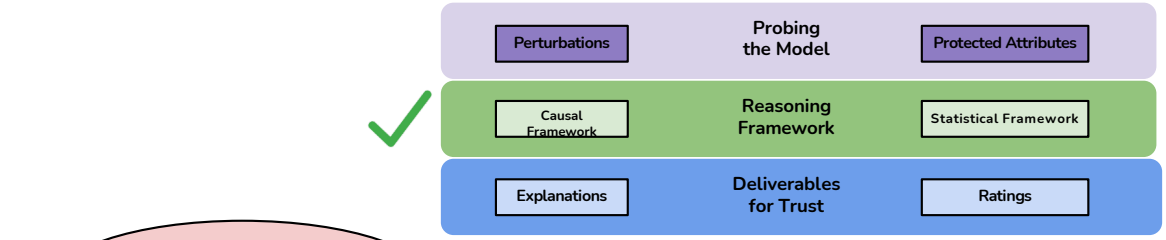


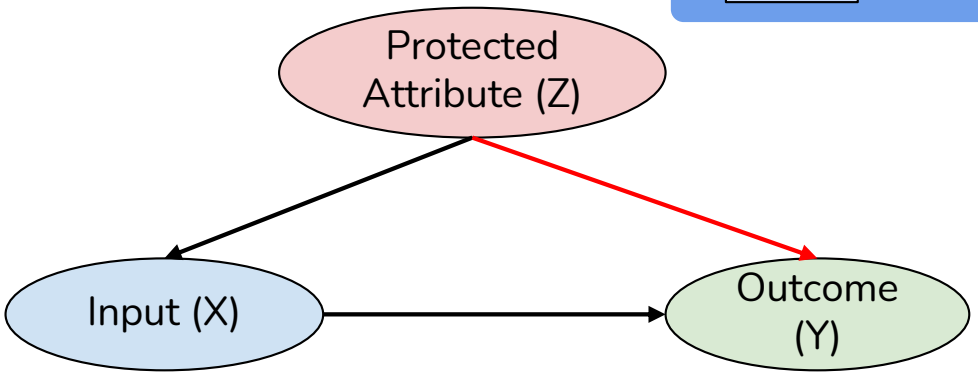
Causation



Credits:
1. https://swflreia.com/2017/05/17/ice-cream-sales-cause-shark-attacks/
2. https://vivdas.medium.com/confounding-variable-and-spurious-correlation-key-challenge-in-making-causal-inference-4e33d8ba60c2

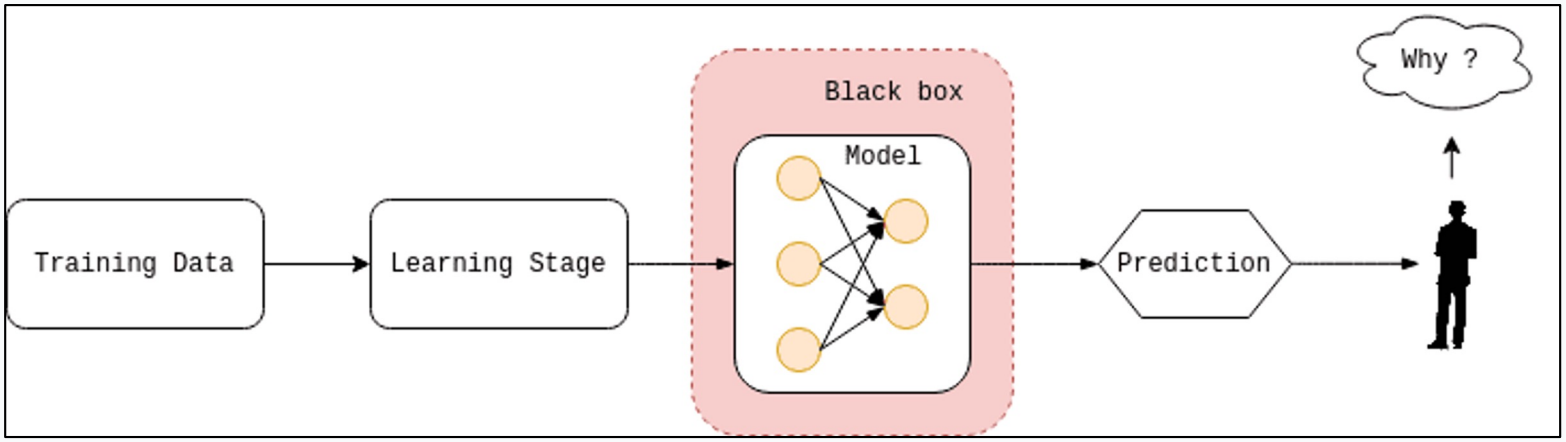
Causality





- Causality is the **science of cause and effect**.
  - A causal model describes **how variables influence each other**, not just how they correlate.
- A causal graph (like the one shown) uses **arrows to represent direct cause-effect relationships** between variables.
- When a **variable *Z* affects both *X* and *Y***, it acts as a ***confounder***, making it harder to isolate the true effect of X on Y.

What is an Explanation?



- Explanation is the process of communicating an AI model's behavior to users in a way that makes it transparent and interpretable, thereby promoting user trust.

References:
1. Papastratis, I. (2021). Introduction to Explainable Artificial Intelligence (XAI). Https://Theaisummer. Com/. Retrieved from https://theaisummer.com/xai/

What is an Explanation?

Bluster
(Recommender System)

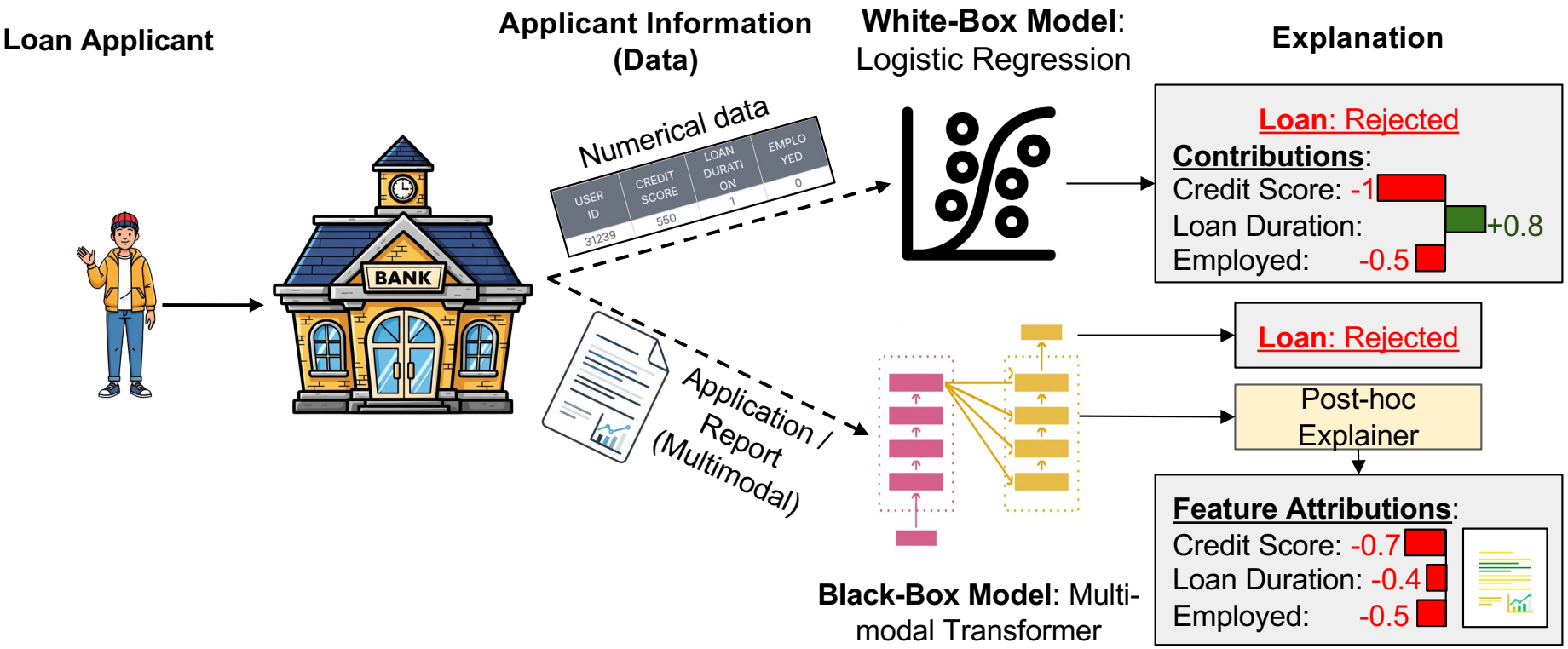
Recommendation:
Approve partial loan
Explanation:
Applicant has stable income but high debt-to-income ratio.

Decision Maker

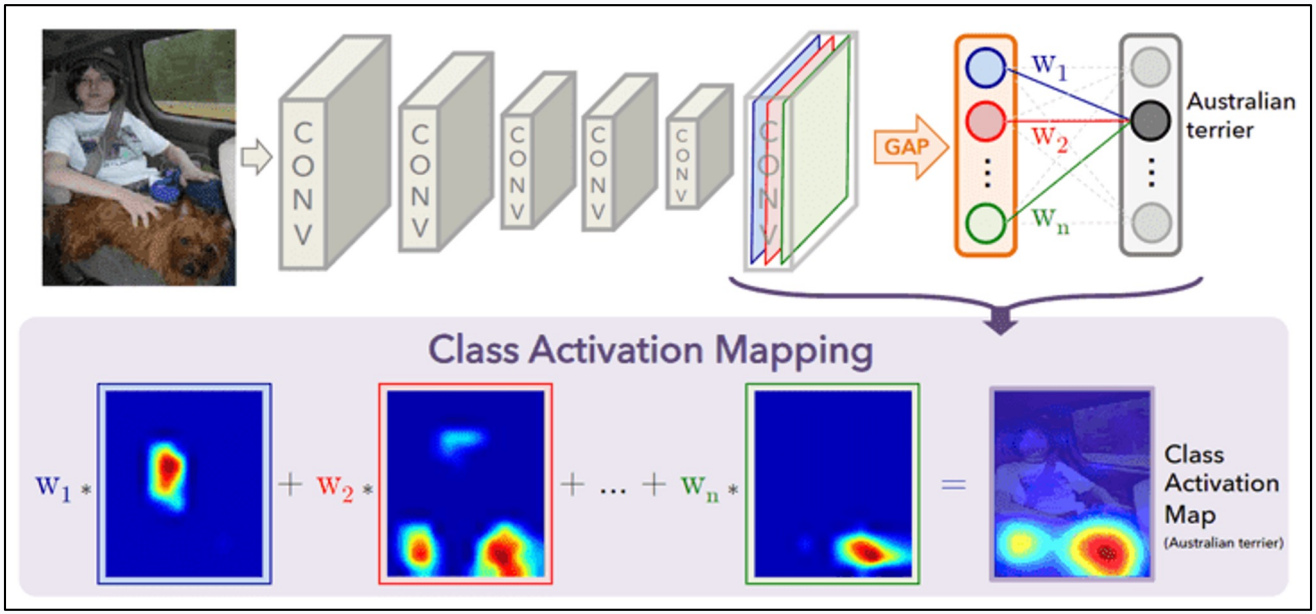
- Explanation ensures that AI systems are transparent, and accountable to users.

References:
1. Papastratis, I. (2021). Introduction to Explainable Artificial Intelligence (XAI). Https://Theaisummer. Com/. Retrieved from https://theaisummer.com/xai/
2. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

White-Box Vs. Black-Box Explanations

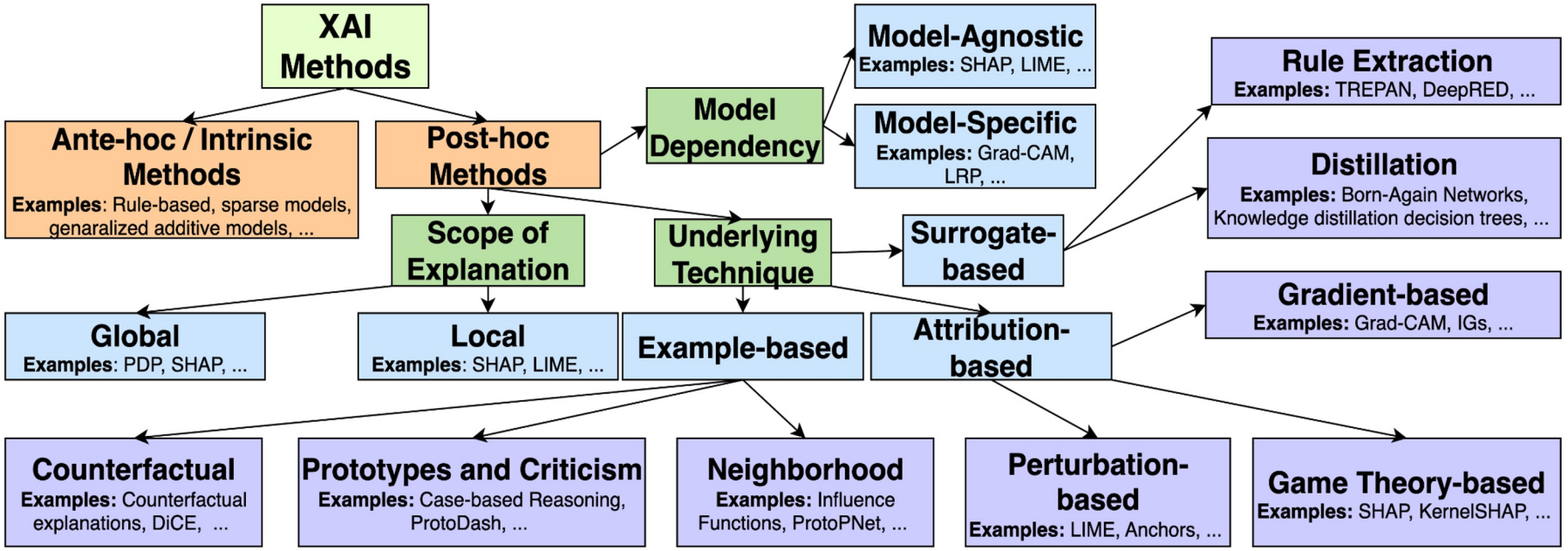


Post-hoc Explainers



- **Post-hoc explainability** refers to methods that are applied to explain the decisions of a model after it has made its predictions.

Classification of XAI Methods



Black-Box Vs. White-Box: Accuracy Vs. Interpretability

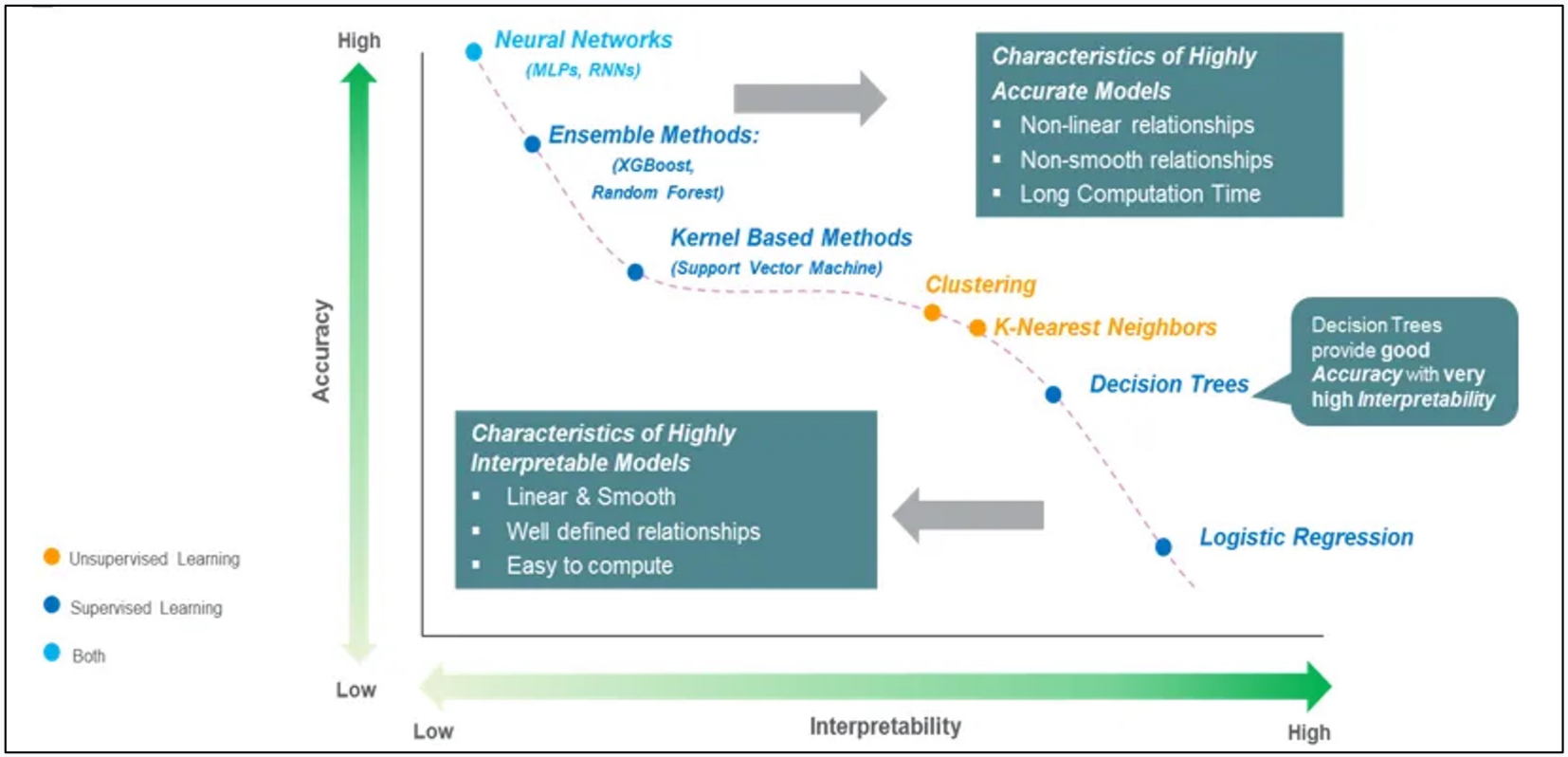


Image Source: https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062

Problem with Current Explainable AI (XAI) Methods: Example Scenario

Bluster

Recommendation: Approve partial loan
Explanation: Applicant has stable income but high debt-to-income ratio.

Decision Maker

Recommendation	Supporting Explanation	Refuting Explanation
Partial loan	Stable job, moderate credit score	High debt-to-income, limited savings
Full loan	No recent defaults	High credit utilization

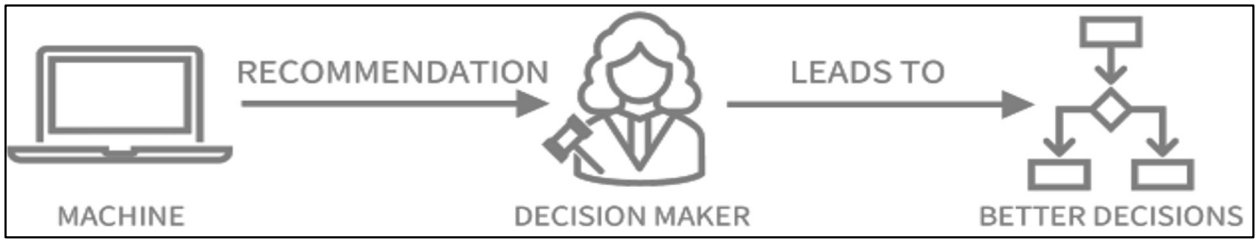
Bluster tells you what is the right decision and also explains why he is right.

Prudence, on the other hand, asks you what you want to do and provides evidence for and against your proposed decision.

Prudence

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Problem with Current Explainable AI (XAI) Methods

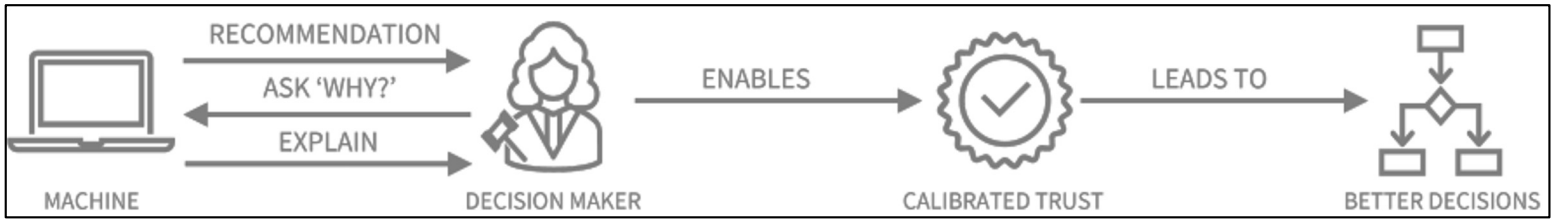


Recommendation-based Decision Support: System gives recommendation without explanation. Assumes that the decision maker considers the recommendation carefully.

But do people consider the recommendation carefully?

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Problem with Current Explainable AI (XAI) Methods

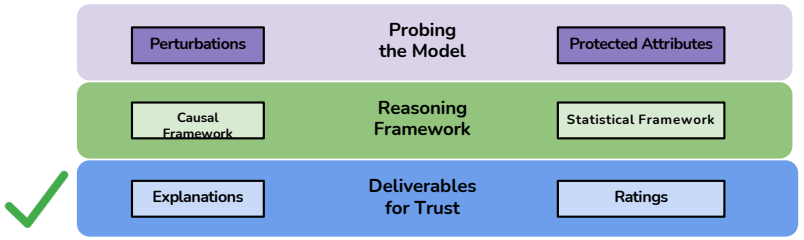


XAI for Decision Support: System gives recommendations with explanation / interpretable model. Assumes that distrust can be mitigated through explanation.

But do people pay careful attention to the explanation?

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Problem with Current Explainable AI (XAI) Methods



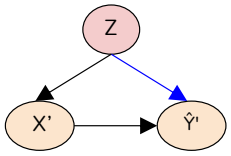
<u>Cognitive Forcing Method</u>: Gives explanation but not the actual recommendation and the decision maker is forced to engage with this explanatory information.

The method is still recommendation-driven as it 'explains' just the machine decision.

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Rating AI Models: Choose Your AI Model Like You Choose Your Peanut Butter!

Input: Test data and models.
Hypothesis: *Attribute X* influences the models' *Predictions Ŷ.*



Rating Method

Output: Rating is given to different binary classifiers.

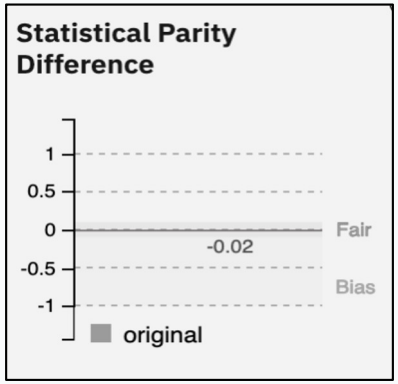
Model Rating	**Raw Scores**		
Logistic (Interpretable)	0	1	
Random (Non-Interpretable)	1.6	2	
Random Forest (Non-Interpretable)	3.9	3	
Biased (Interpretable)			4.6
4			

Image Credits: https://slideplayer.com/slide/8155169/

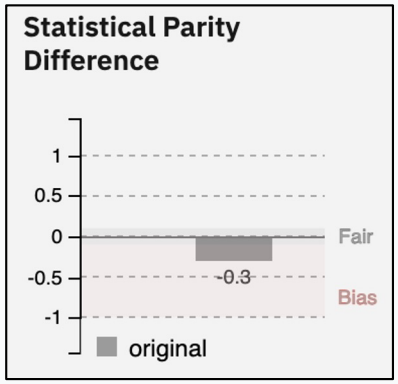
Rating Vs. Global Explanation: An Example

- A global explanation provides an overview of how features generally influence a model's predictions across the entire dataset.

Example Scenario:
ABC Bank customers' data with the following attributes:
Input attributes: Gender, Age, Credit amount
Following is the dependent variable that the model predicts:
Output: Risky customer (1) / not risky (0)



With respect to **'Gender'**



With respect to **'Age'**

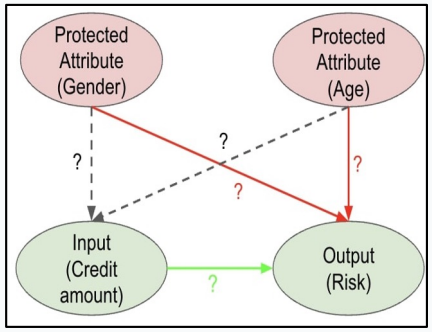
- Statistical Parity Difference is computed as the **difference of the rate of favorable outcomes received by the unprivileged group to the privileged group**.
- The ideal value of this metric is 0. **Value > 0 denotes higher benefit for privileged group (male or old)**.

References:
1. Bellamy, Rachel KE, et al. "AI Fairness 360: an extensible toolkit for detecting." Understanding, and Mitigating Unwanted Algorithmic Bias 2 (2018).

Rating Vs. Global Explanation: An Example

- The rating method tests hypotheses based on the system's causal structure to isolate and measure each attribute's effect on outcomes.

Example Scenario:
ABC Bank customers' data with the following attributes:
Input attributes: Gender, Age, Credit amount
Following is the dependent variable that the model predicts:
Output: Risky customer (1) / not risky (0)



The partial order is (lower scores are desirable):

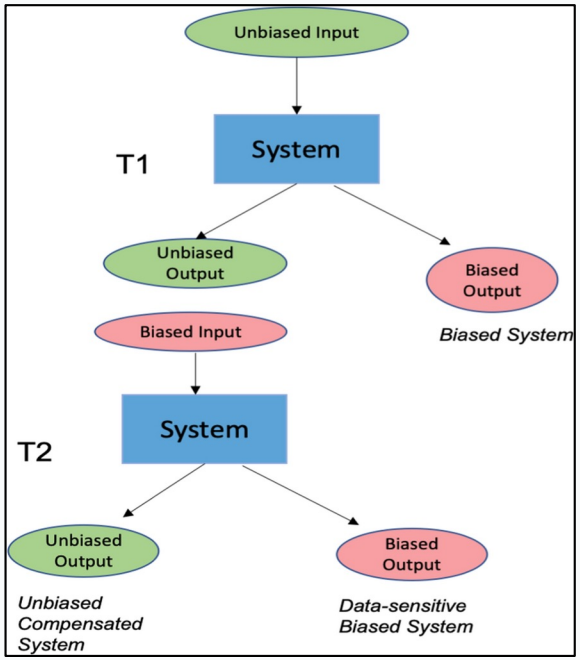
	Logistic Regression	Random
Input_0 (Credit amount=0; Protected Var: Gender)	51.58	3.15
Input_1 (Credit amount=1; Protected Var: Gender)	7.58	22.73
Input_2 (Credit amount=2; Protected Var: Gender)	83.33	0.00
Input_3 (Credit amount=0; Protected Var: Age)	48.87	45.95
Input_4 (Credit amount=1; Protected Var: Age)	19.70	39.39
Input_5 (Credit amount=2; Protected Var: Age)	41.67	0.00

The final ratings with respect to ['Gender', 'Age'] (lower ratings are desirable):
Input_0 (Credit amount=0; Protected Var: Gender)
{'Random': 1, 'Logistic Regression': 2}
Input_1 (Credit amount=1; Protected Var: Gender)
{'Logistic Regression': 1, 'Random': 2}
Input_2 (Credit amount=2; Protected Var: Gender)
{'Random': 1, 'Logistic Regression': 2}
Input_3 (Credit amount=0; Protected Var: Age)
{'Random': 1, 'Logistic Regression': 2}
Input_4 (Credit amount=1; Protected Var: Age)
{'Logistic Regression': 1, 'Random': 2}
Input_5 (Credit amount=2; Protected Var: Age)
{'Random': 1, 'Logistic Regression': 2}

Rating AI Systems: Statistical Approaches

- Many statistical based approaches were used before to assess the trustworthiness of AI systems such as machine translators and chatbots.

More papers on rating can be found here!

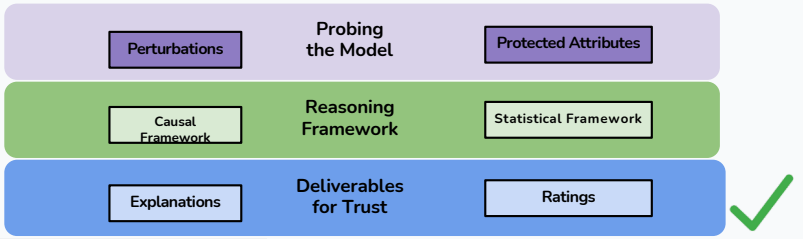


[Machine Translators] Srivastava, B.; and Rossi, F. 2020. Rating AI Systems for Bias to Promote Trustable Applications. In IBM Journal of Research and Development.

[**Chatbots**] Srivastava, B., Rossi, F., Usmani, S., & Bernagozzi, M. (2020). Personalized chatbot trustworthiness ratings. IEEE Transactions on Technology and Society, 1(4), 184-192.

[**Composite Services**] Srivastava, B., & Rossi, F. (2018, December). Towards composable bias rating of AI services. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 284-289).

Rating AI Systems through a Causal Lens



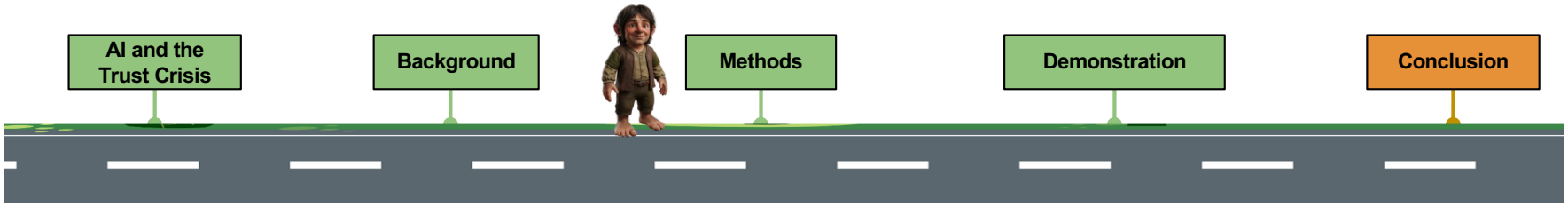
[MM-Time-Series Forecasting Models] Lakkaraju, K., Kaur, R., Zeng, Z., Zehtabi, P., Patra, S., Srivastava, B., & Valtorta, M. (2024). Rating Multi-Modal Time-Series Forecasting Models (MM-TSFM) for Robustness Through a Causal Lens. arXiv preprint arXiv:2406.12908.

[**Sentiment Analysis Systems + Translators (Composite)**] Lakkaraju, K., Gupta, A., Srivastava, B., Valtorta, M., & Wu, D. (2023, November). The Effect of Human v/s Synthetic Test Data and Round-tripping on Assessment of Sentiment Analysis Systems for Bias. In 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) (pp. 380-389). IEEE.

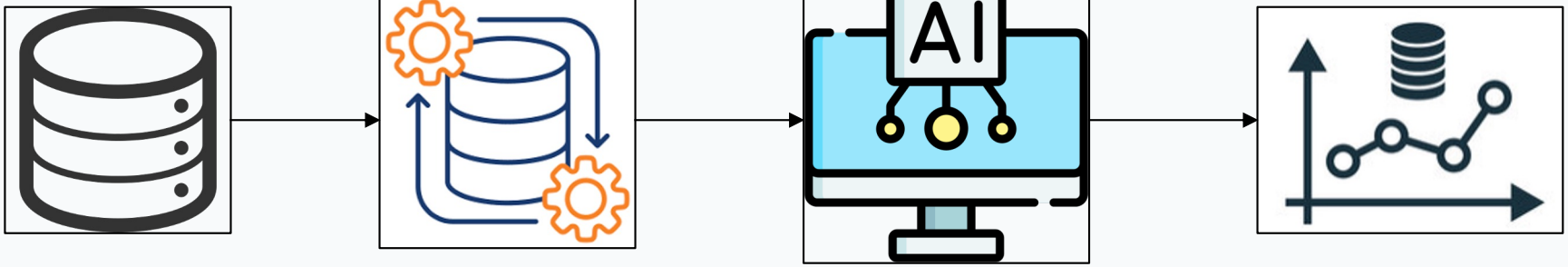
[**Sentiment Analysis Systems**] Lakkaraju, K., Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.

More papers on rating can be found here!

03.
Methods


AI and the Trust Crisis
Background
Methods
Demonstration
Conclusion

Rating Workflow: Input to Predictions



Obtaining the test data: Can be text-only, image-only, numerical, or multi-modal

Preprocessing the data: Data must be prepared for AI systems, especially when applying perturbations.

Setting up the test system: A pre-trained AI system that can be unimodal system or multi-modal.

Obtaining the predictions: Modality may vary based on the modality of the data and the AI system.

Example Scenario: Credit Risk Assessment

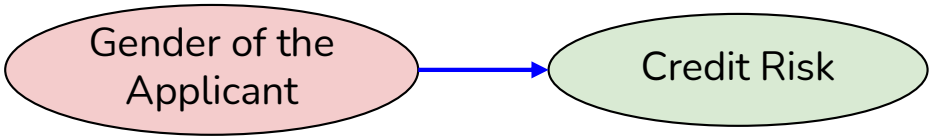
Bank XYZ uses an **a Random Forest model to predict credit risk and decide who gets a loan based on user attributes**.

But these models can behave strangely. **Small changes in user attributes** like income or age can lead to **big shifts in predictions**, and they often provide **no explanation** for why a loan is approved or rejected.

We want to test **how robust these models really are.**

Image Credits: ChatGPT

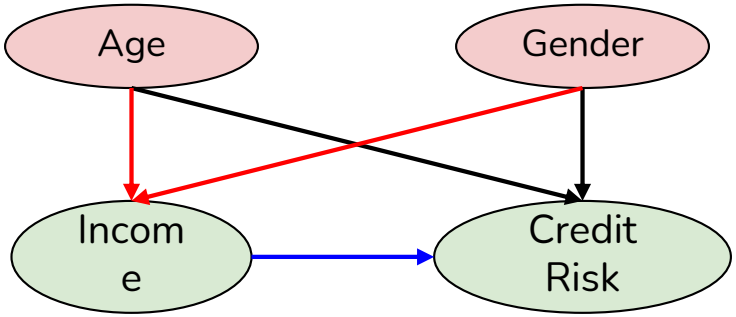
RQ-1: Can we quantify the effect of protected attributes on the model's outcome, indicating statistical bias?



Key Idea:

- Use **student's t-test** to estimate if the **average predicted risk is significantly different between the groups.**
- For example, is there a statistically significant **difference in the predicted risk between male and female applicants**?

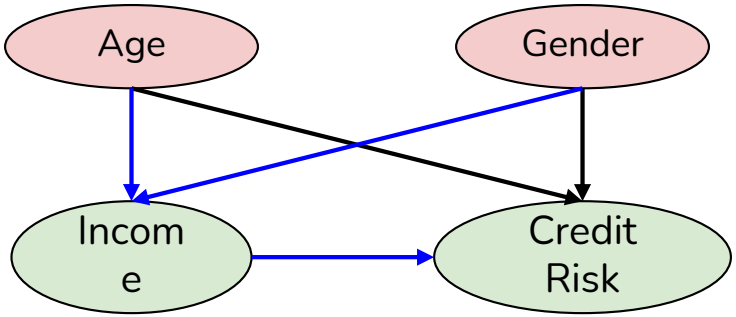
RQ-2: Can we estimate the causal effect of changing an input feature on model predictions?



Key Idea:

- Apply a **treatment** to an **input feature (e.g., income)** and observe how the **predicted credit risk changes.**
- For example, what happens to the **predicted credit risk** if the **applicants' income changes from $5,000 to $10,000**?

RQ-3: Can we quantify the effect of protected attributes on the *relationship between input and* model *outcomes*, indicating confounding bias?



Key Idea:

- Measure the causal effect of change in **input (e.g., change in income) before and after adjusting for protected attributes (i.e., deconfounding)**.
- The difference reveals **how much gender and age were confounding the model's predictions**.

RQ-4: Can rating method complement existing XAI methods and work in tandem with it to provide a holistic explanation to various stakeholders?

Key Idea:

- Current XAI methods are typically one-off and are designed more for AI developers than for other stakeholders involved.
- The idea is to combine traditional XAI methods with rating to provide a holistic view of an AI model to different stakeholders.

Example Scenario: Holistic XAI

- **Stakeholder: Jack (applicant; individual)**.
- **Jack was classified as a bad risk** and wants to understand why, and **what changes he could make to his application to be considered a good risk**.

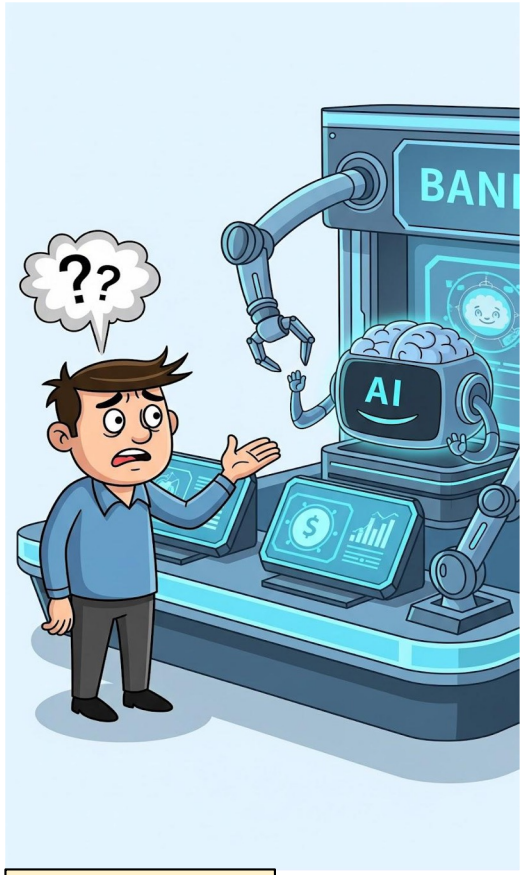


Image Credits: ChatGPT

Example Scenario : Credit Risk Classification

Q1: *On my data instance, I have observed the AI model used by the* ***bank to be biased with respect to age, personal status, and gender****, especially in how it uses credit amount to predict risk. How can I rate this model for expected behavior?*

Approach: We assess the impact of **confounders** on the **relationship between credit amount and predicted risk (method from RQ-3)**. Rating compares the tested model against random and biased baselines.

Explanation:

Model	Raw Scores	Rating
Random Forest	13.735	3
Random	3.020	1
Biased	10.435	2

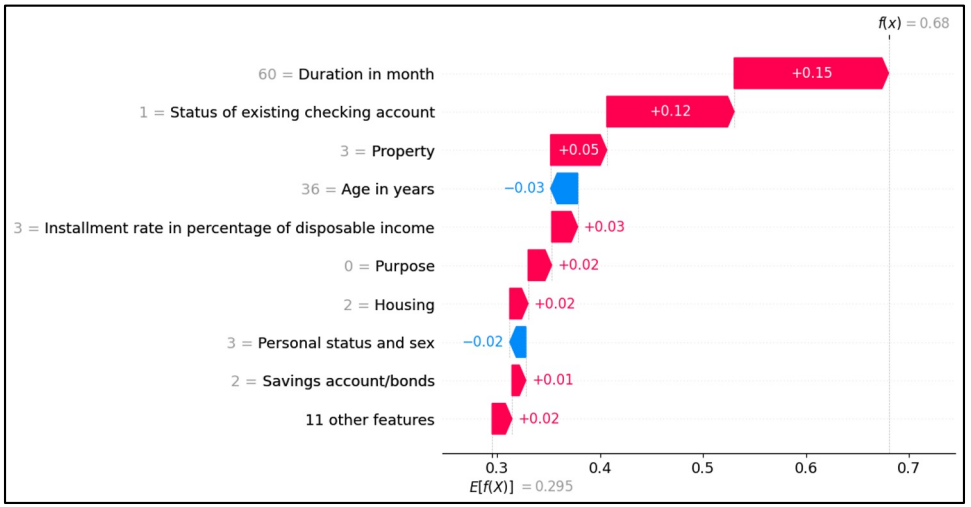
Random Forest model used by the bank exhibits greater bias than both a randomly predicting model and an average biased baseline, indicating relatively high group-level disparities.

Example Scenario : Credit Risk Classification

Q2: I want to investigate how the **protected features, along with other features, contributed to his loan rejection?**

Approach: SHAP values were used to explain the prediction by attributing importance to each feature. Rating does not provide local explanations.

Explanation:



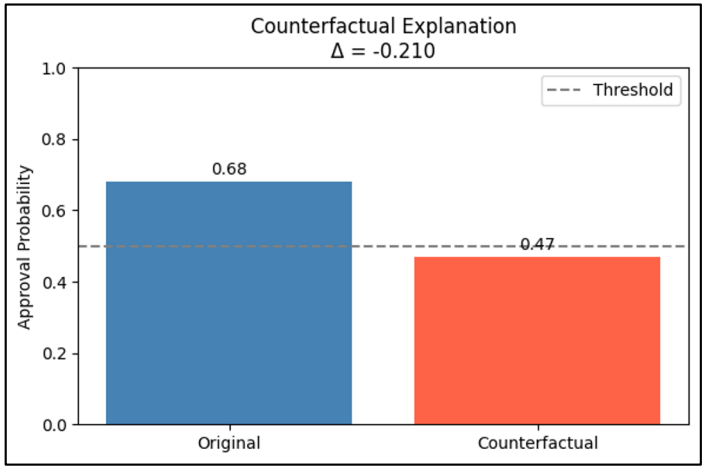
Age and gender pushed the prediction toward good risk, while duration and account status pushed it toward bad risk.

Example Scenario : Credit Risk Classification

Q3: I want to find the minimal change that could be made to the top-2 contributing features to flip his prediction. Let's start with ***Loan duration in months.***

Approach: **Counterfactual explanation** was used to identify minimal change in duration needed to flip the prediction.

Explanation:



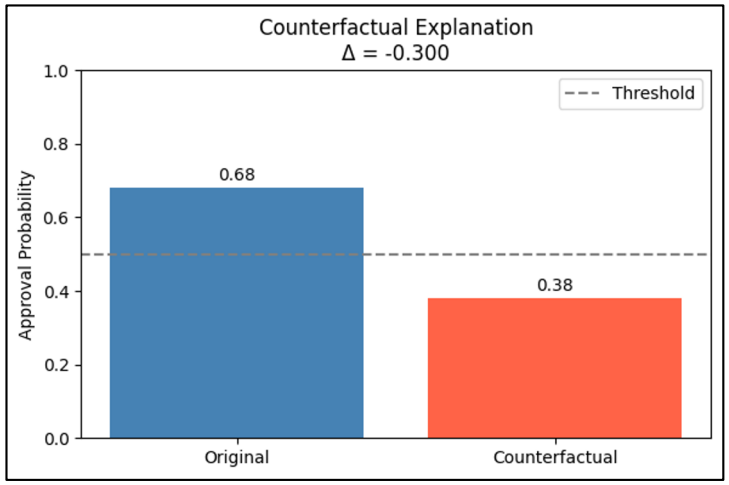
Potential action 1: Reducing the loan duration to 6 months from 60 months would decrease your rejection chance from 68% to 47%, and the bank would approve your loan.

Example Scenario : Credit Risk Classification

Q4: What is the smallest change I could make to my ***checking account balance*** to get my loan approved?

Approach: **Counterfactual explanation** was used to identify minimal change in duration needed to flip the prediction.

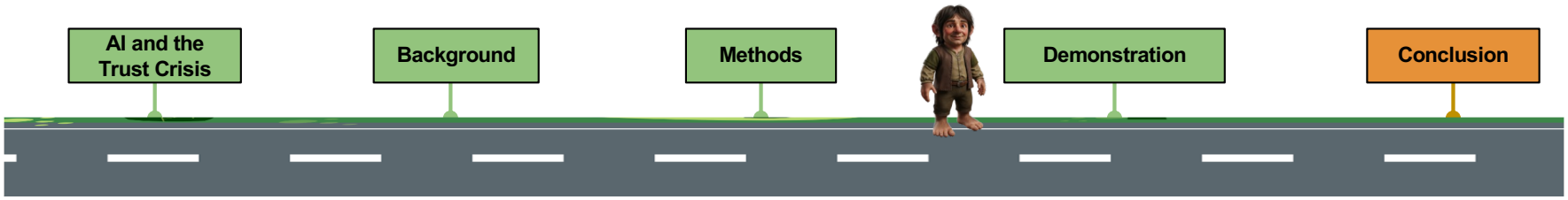
Explanation:



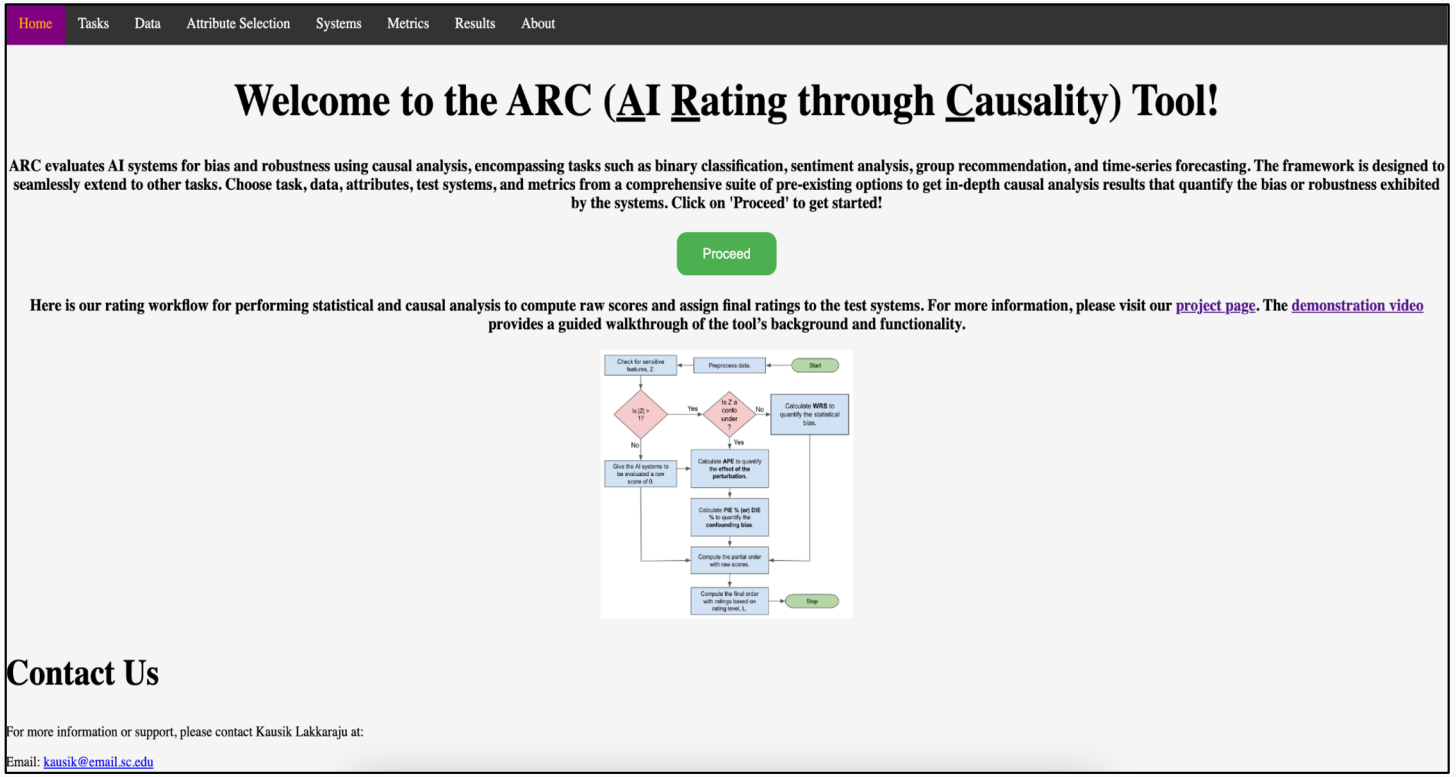
Potential action-2: Raising your balance to at least 200 DM would decrease your rejection chance from 68% to 38%, and the bank would approve your loan.

04.

ARC Tool Demonstration

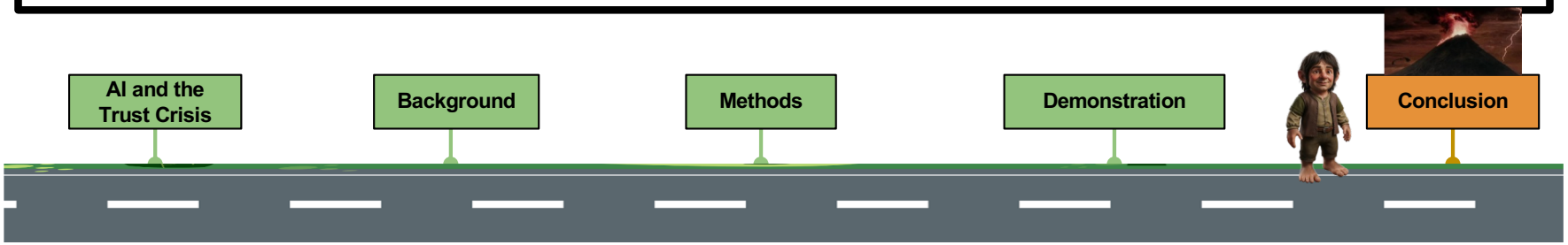


ARC: AI Rating through Causality



Scan the code try our ARC tool!

Conclusion


AI and the Trust Crisis
Background
Methods
Demonstration
Conclusion

Summary

- **We discussed**
  - Importance of building trustworthy AI systems.
  - Blackbox and whitebox settings.
  - Introduced explanations and rating to promote trust.
  - Discussed data and methods used for rating AI systems
  - Introduced causality-based rating.
  - Demonstrated ARC tool
- **Perspective**
  - Assessment and rating is an important approach to build trust with AI users
  - More to be done with foundation models (FMs) and multi-modal data
  - Scaled to other types of AIs and domains.

Glossary

In the Context of AI

- **Black-Box AI Models:** Models whose decision process is hidden or too complex to understand (e.g., deep neural networks).
- **White-Box AI Models:** Models whose decision process is transparent and easily understood (e.g., decision trees, linear regression).
- **Bias:** A systematic error in data or models that leads to unfair or skewed outcomes.
- **Instability**: When small changes in input or data lead to big changes in model output.
- **Perturbations:** Small modifications to input data used to test or probe a model's behavior.

In the Context of AI

- **Robustness:** The ability of a model to maintain performance even when faced with noisy inputs.
- **Causality:** Understanding not just correlations but true cause-and-effect relationships between different attributes in a system.
- **Explanations:** Process of describing why a model made a particular prediction.
- **Interpretability:** How easily humans can understand the reasoning behind a model's decisions.
- **Ratings:** Scores assigned to models based on quantified measures of their robustness (using causal and statistical methods), which are then used to rank and compare the models.

Q&A

Our rating project page