Explainable Machine Learning in Practice

Kasun Amarasinghe

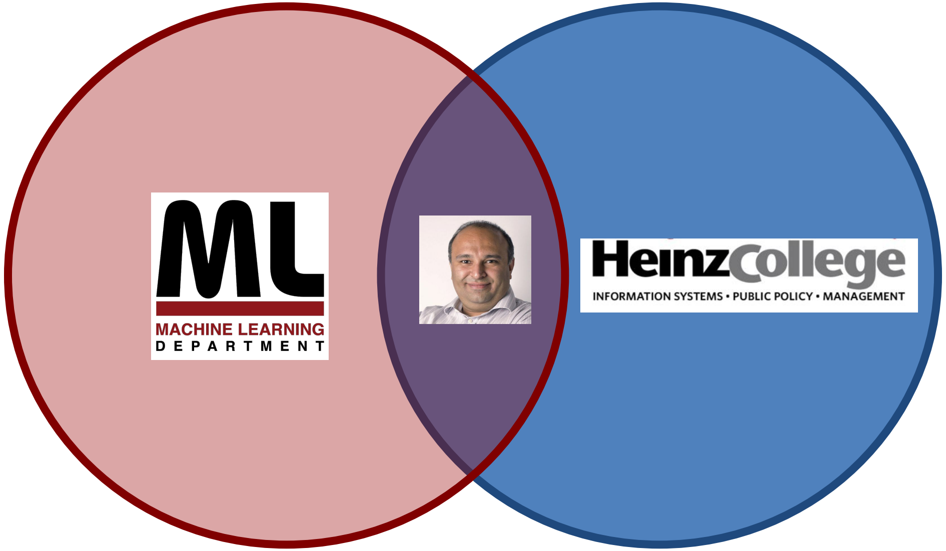
Data Science for Public Policy Team at CMU

Rayid Ghani
Professor

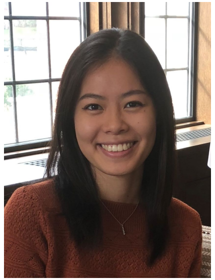
Kit Rodolfa
Senior Research Scientist

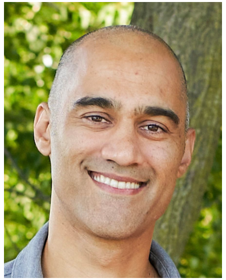
Liliana Millan
Senior Data Scientist

Alice Lai
Senior Data Scientist



And our collaborators…


Valerie Chen

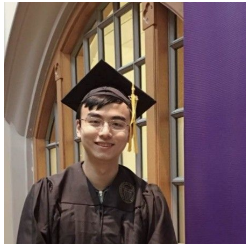

Ameet Talwalkar


Pedro Saleiro


Sergio Jesús


Vladimir Balayan


Pedro Bizarro


Wenbo Cui

Today's talk…

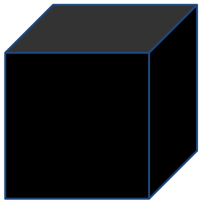
My main goal for today is to highlight the role we as practicing data scientists can play in advancing Explainable ML

- Provide an overview of Explainable ML
- Identify gaps between practical needs and the current research
- Highlight the role of practitioners in bridging those gaps

Why Explainable Machine Learning?

Human - ML Interaction

Increased Human - AI Interaction across domains

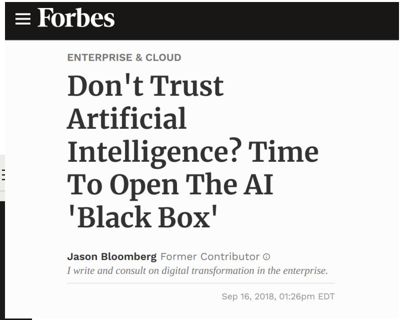


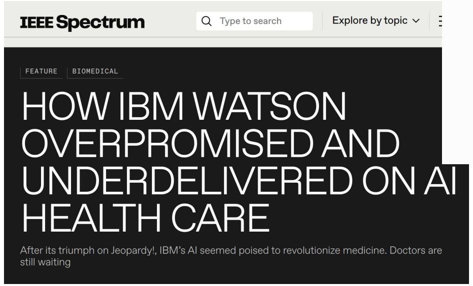
More complex the problem, more complex the model

Black-box models can surface several risks!

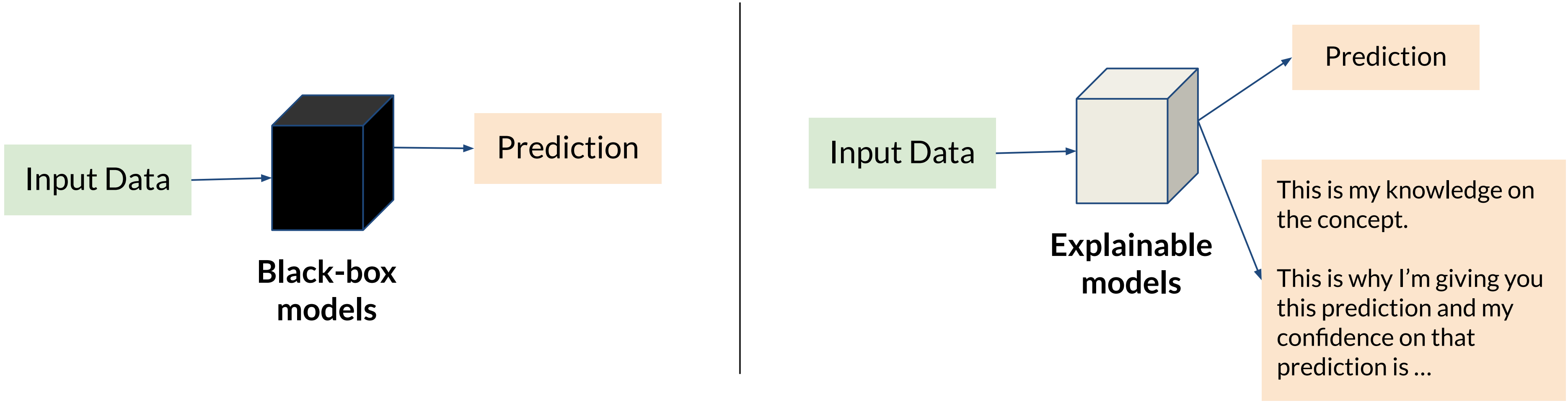
Common themes behind the need for explainability

- Potential errors/biases going unchecked
- Lack of trust
- Regulatory requirements (e.g., GDPR)





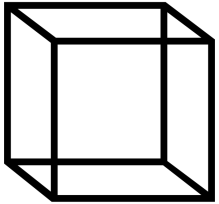
Black-Box Models vs (aspirational) Explainable ML models



Explanations can potentially give us further insight into what the model is learning

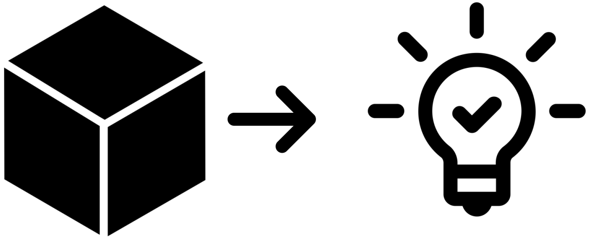
How has the research community responded?

Two main approaches



Inherently Interpretable models

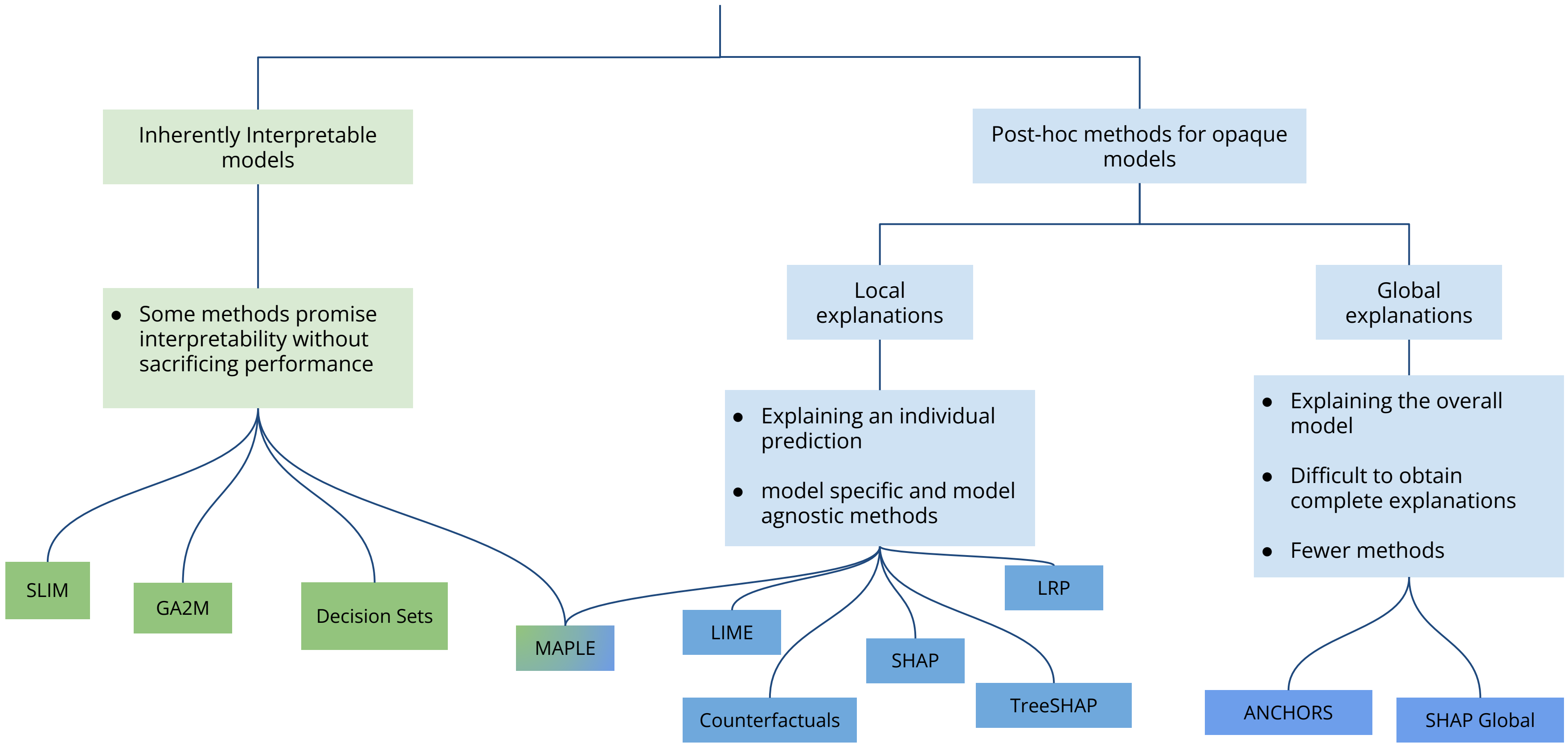
ML models that are interpretable on their own



Post-hoc Explainable ML methods

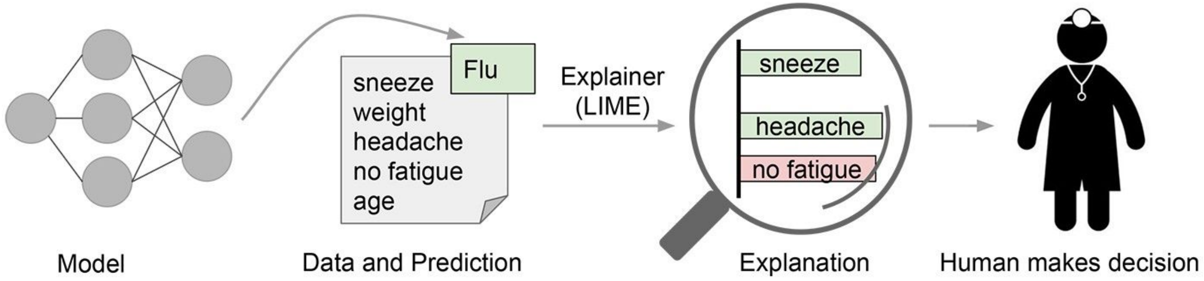
The learning algorithm is not tampered with, a post-hoc method is used to probe the trained model for extracting an explanation

(Some) Existing work in Explainable ML



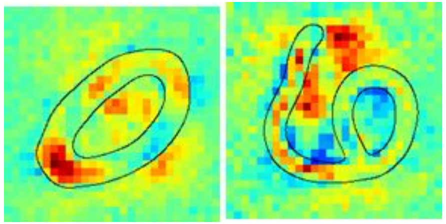
Feature attribution type explanations

- Feature attribution:
  - Assigning an "importance" to each input feature that quantifies its contribution to a prediction
  - Most popular explanation type



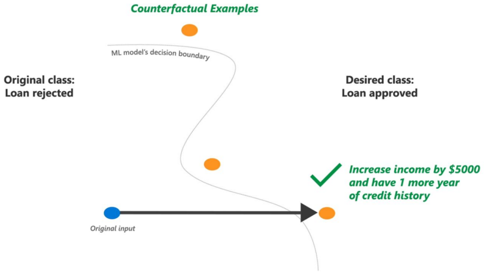
Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier

Other types of explanations..



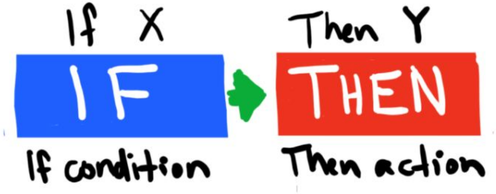
Heatmaps

Mainly used in image classification



Counterfactuals

"What's the smallest change in data that would produce a different outcome?"



IF-THEN type rules

Mostly used for global explanations

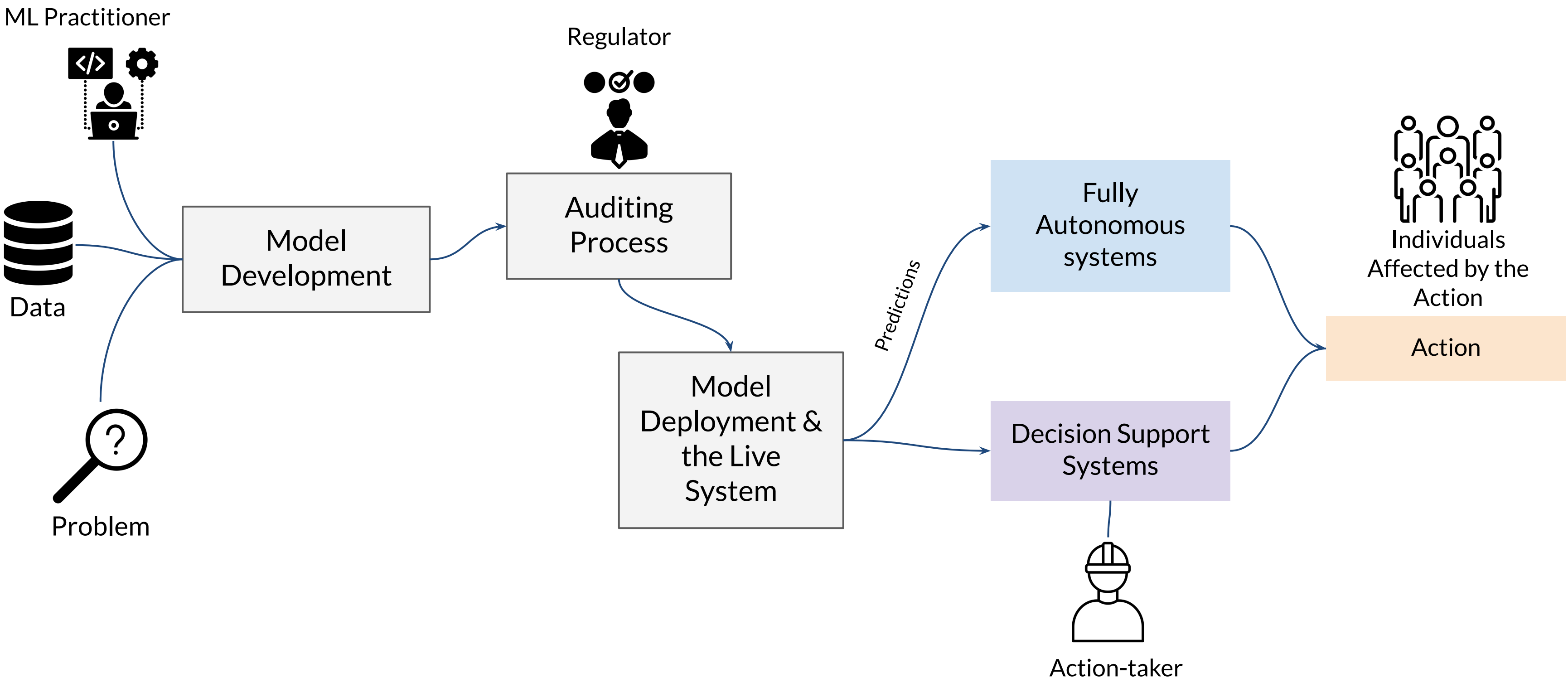
Montavon, Grégoire,et al.. "Layer-wise relevance propagation: an overview." *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209.
Mothilal, Ramaravind K., et al.. "Explaining machine learning classifiers through diverse counterfactual explanations." *In Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020.
Ribeiro, Marco et al.. "Anchors: High-precision model-agnostic explanations." *In Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1. 2018.

Some of existing approaches in each category

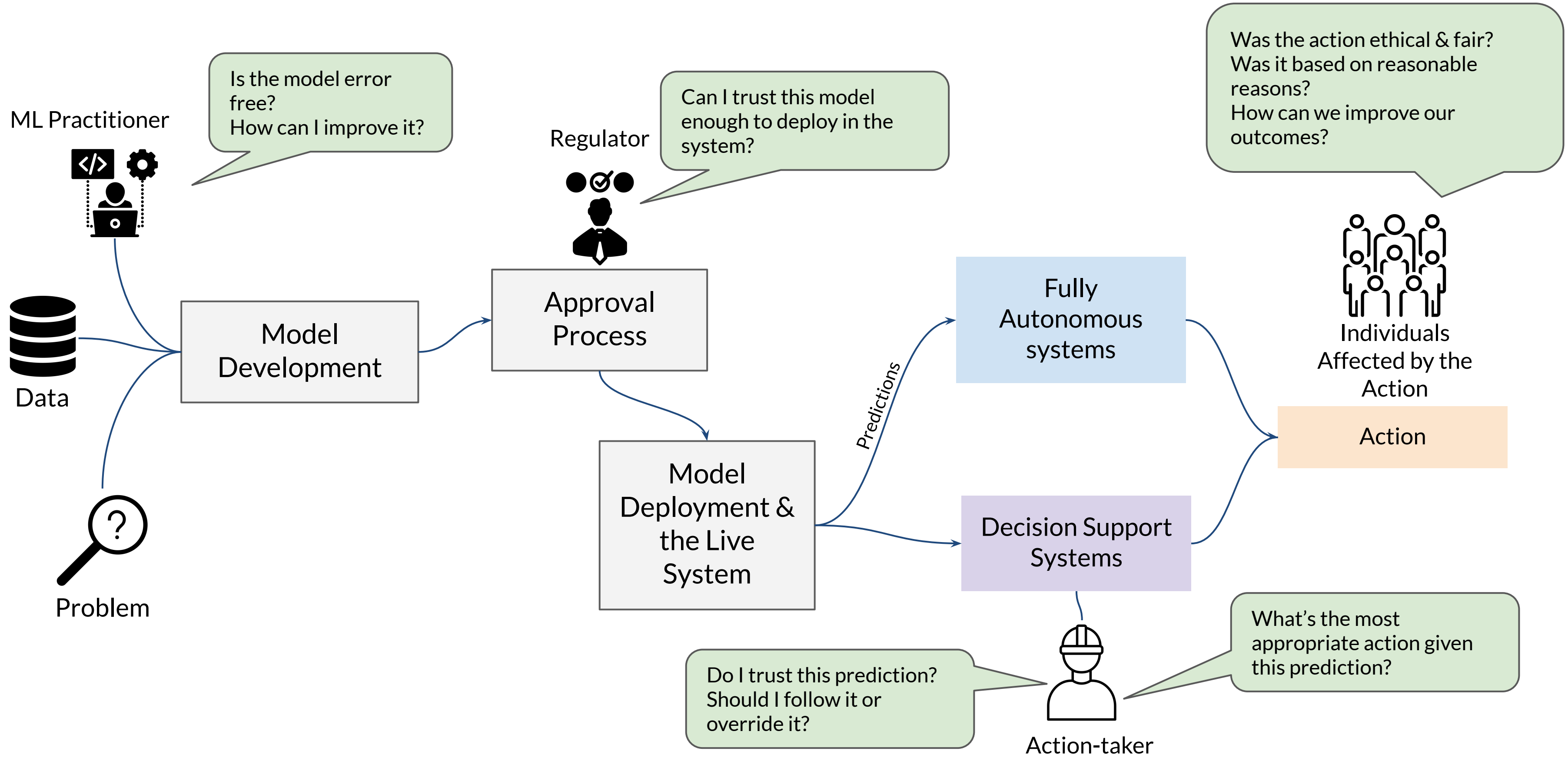
	Inherently Explainable Models	Post-hoc Methods	
	Super-sparse Linear Models Generalized Additive Models Generalized Linear Models Rule-lists Shallow Decision Trees	Feature Attribution based explanations Example based explanations Counterfactual / Contrastive Explanations	**Local**
		Rule based summaries of the model Distilling to a surrogate model	**Global**

We have all these methods, how and when to use them?

Let’s look at the different human - ML interactions



Different ways in which explainable ML can help...



Method capabilities versus needs

★☆☆: Potentially applicable methods exist. Efficacy not demonstrated with any evaluation

★★☆: Some evidence exist, but real world efficacy is not validated

★★★: Real world efficacy of the methods empirically validated

Use-case	Post-hoc Local	Post-hoc Global	Interpretable Models
Model debugging	★★☆	★★☆	★★☆
User trust	★☆☆	★☆☆	★☆☆
Improving decision making system performance	★☆☆	N/A	★☆☆
Improving interventions	★☆☆	N/A	★☆☆
Recourse	★★☆	N/A	★★☆

We couldn't find a well-designed empirical study that verified utility of methods for any use-case, and thus, practitioners have little to no information on when or how to use these methods!

Amarasinghe, K., Rodolfa, K., Lamba, H., & Ghani, R. (2020). Explainable machine learning for public policy: Use cases, gaps, and research directions. *arXiv preprint arXiv:2010.14374*.

So, what can we (data scientists) do bridge this gap?

We need to evaluate methods on real use-cases!

- The first step of bridging the gap between research and practice would be to evaluate existing methods on real-world use-cases

- This presents a great opportunity for practitioners to highlight specific gaps that exist between real-world needs and explainable ML methods

Evaluating Explainable ML Methods

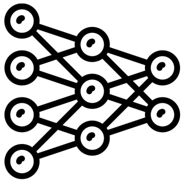
How do we evaluate explainable ML?

- Compared to method development, research into evaluation of explainable ML has lagged
- Evaluation of explainable ML is multi-faceted
  - Intrinsic qualities of the explanation
  - Ability to improve human-ML collaboration
- Doshi-Velez and Kim captured this spectrum in a three-staged framework

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Functionally-grounded evaluation

- Evaluating the intrinsic qualities of the artifact (i.e., the explanation)
- No users involved in the evaluation



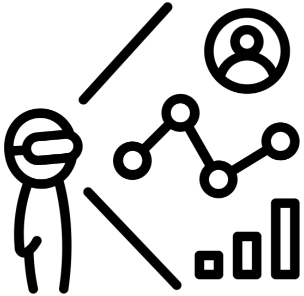
Fidelity to the underlying model

Completeness of the explanation

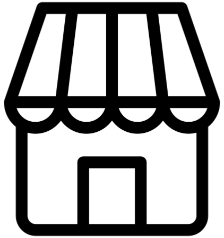
Human-Friendliness

Human-grounded Evaluation

User studies are conducted, but with simple/proxy tasks and typically users in research settings



Forward Simulation has been the most popular proxy task

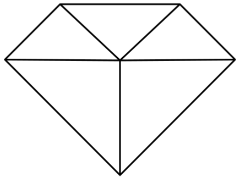


Mechanical Turk, Prolific are popular proxy user bases

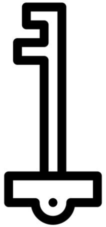
Application-grounded Evaluation

User study with real-world users performing the real task

Entail significant logistical challenges



Very rare in the literature



These types of studies are necessary to evaluate real world efficacy

Data Scientists can lead the way in designing and conducting application grounded evaluations

Some Common Pitfalls

Using proxy tasks

- Performance on a proxy task is used as a metric of explanation quality
  - Forward simulation

- Tasks used in these settings are not tasks humans would face in the real world

- Unlikely that the performance on the proxy task is predictive of real world efficacy
  - Can overestimate capabilities
  - Quantified by Bucinca and colleagues

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Proceedings of the 25th ACM IUI '20. New York, NY, USA,

Using subjective measures as metrics of explanation quality

- User reported quality measures are commonly used to assess the explanations
  - User experience
  - Trust
  - Preference

- Captures what the users think of the explanation, not the objective task performance
  - Humans can be mislead with explanations (Lakkaraju et al. 2020)
  - User preference doesn’t correlate with task performance (Poursabzi-Sangdeh et al. 2021, Bucinca et al. 2020)

Experimental Design Flaws

- There exists a few experiments where real users of a system are performing the real task
- Unfortunately, there are flaws in the experimental setup that limits the conclusions we can draw
- Let’s look at one…

Assisting anesthesiologists detect hypoxemia in surgery

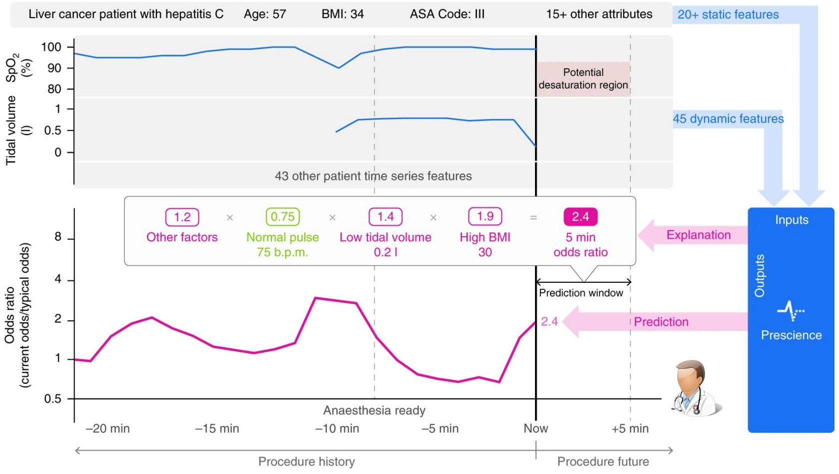
- **User:** Anesthesiologists
- **Task:** Given the data for the last 20 mins for the surgery, predict the risk of hypoxemia in the next 5 minutes
- ML system is named "Prescience"
  - ML model prediction
  - SHAP explanation
- Research Q: Can Prescience improve Anesthesiologists' decisions?

Article | Published: 10 October 2018

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim & Su-In Lee

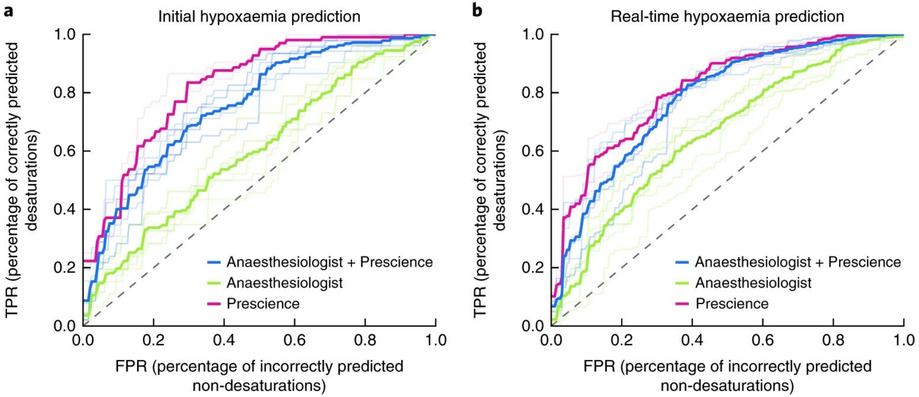
Nature Biomedical Engineering **2**, 749–760 (2018) | Cite this article



Assisting anesthesiologists detect hypoxemia in surgery

- Compare performance:
  - Anesthesiologists alone
  - Anesthesiologist + Prescience
  - Prescience

- Anaesthesiologists made better decisions assisted by Prescience



Do we attribute the performance to the explanation? Or to the ML prediction?

The experiment does not isolate the incremental impact of the explanation!

We need well designed experiments to evaluate explainable ML methods…

We attempted to design and conduct one…

Desiderata for robust application-grounded evaluation

- A real task
  - With performance metrics that capture operational goals
- Real data
  - Reflects the nuances and complexities of the deployment context
- Real users
  - who perform the task in the real world
- A robust inference strategy
  - appropriate hypotheses and experimental conditions

We started with a previous study

- **User:** Fraud analyst
- **Task:** Detect fraudulent e-commerce credit card transactions
- **Data:** Historical transactions from one merchant
- Their setting evaluated the appropriate hypotheses

How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations

Sérgio Jesus
Feedzai, DCC-FCUP
Universidade do Porto
sergio.jesus@feedzai.com

Catarina Belém
Feedzai
catarina.belem@feedzai.com

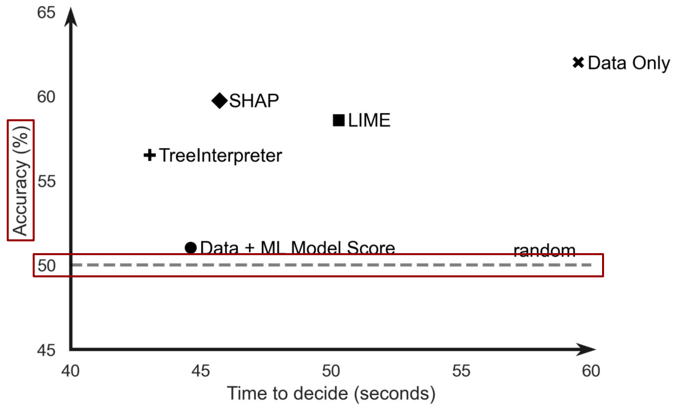
Vladimir Balayan
Feedzai
vladimir.balayan@feedzai.com

João Bento
Feedzai
joao.bento@feedzai.com

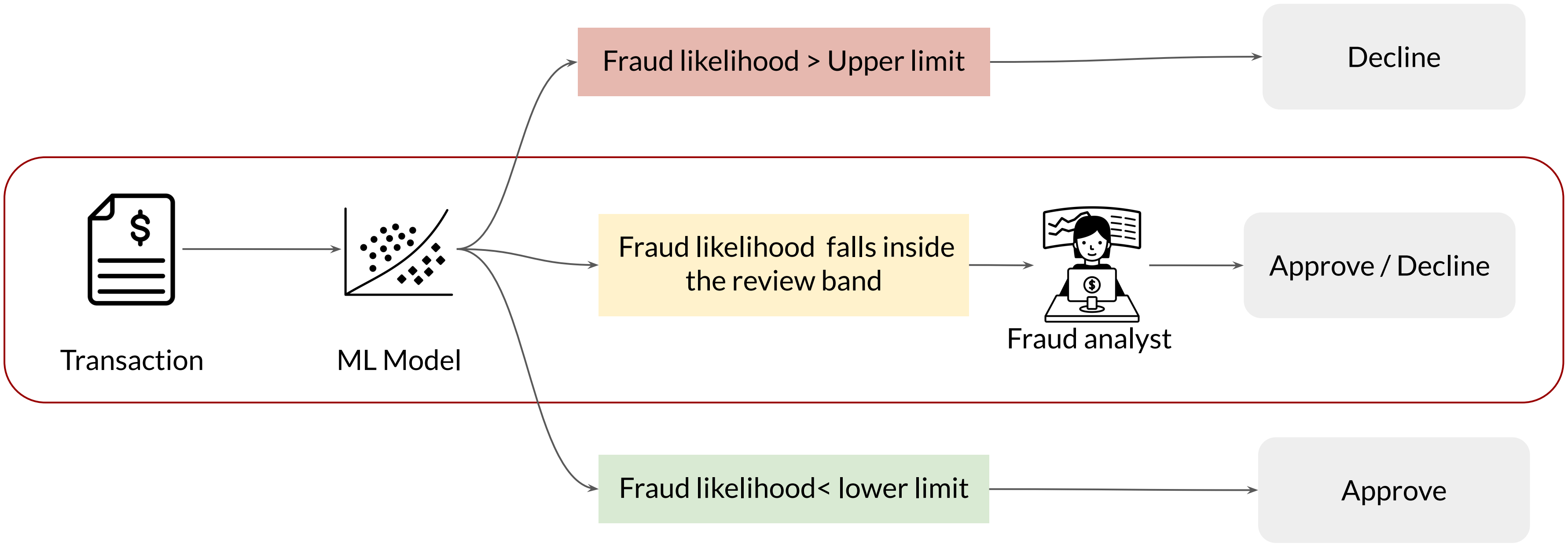
Pedro Saleiro
Feedzai
pedro.saleiro@feedzai.com

Pedro Bizarro
Feedzai
pedro.bizarro@feedzai.com

João Gama
LIAAD, INESCTEC
Universidade do Porto
jgama@fep.up.pt

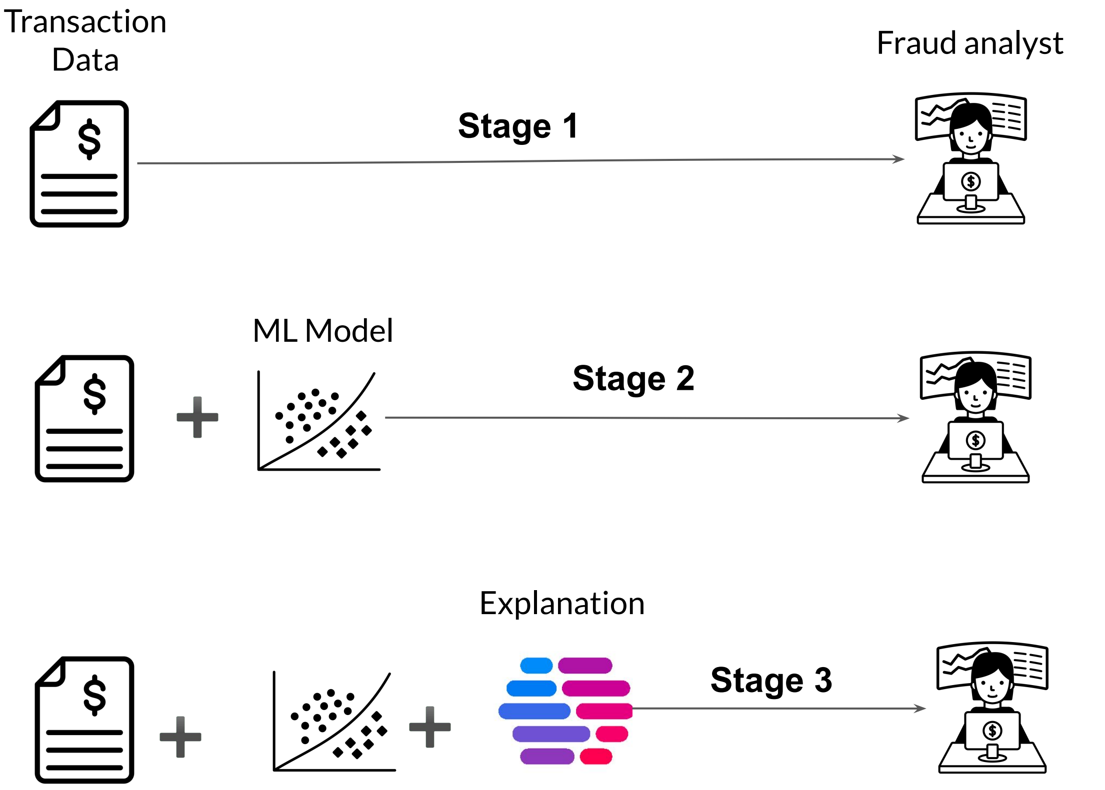


The fraud detection context



A human analyst reviews transactions that the model is uncertain about

Our unit of randomization was transactions



Possible Decisions

- Decline
- Approve
- Suspicious

Post-decision questions

Did you have enough information to make the decision?

How confident are your about your decision?

Designing the Performance Metric

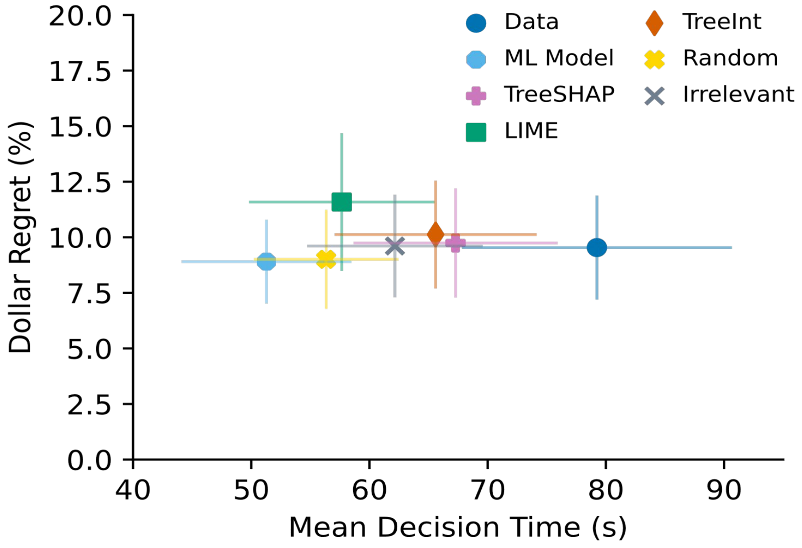
- A metric that captures operational objectives
  - TP vs FP tradeoff
  - Revenue generated by the transaction

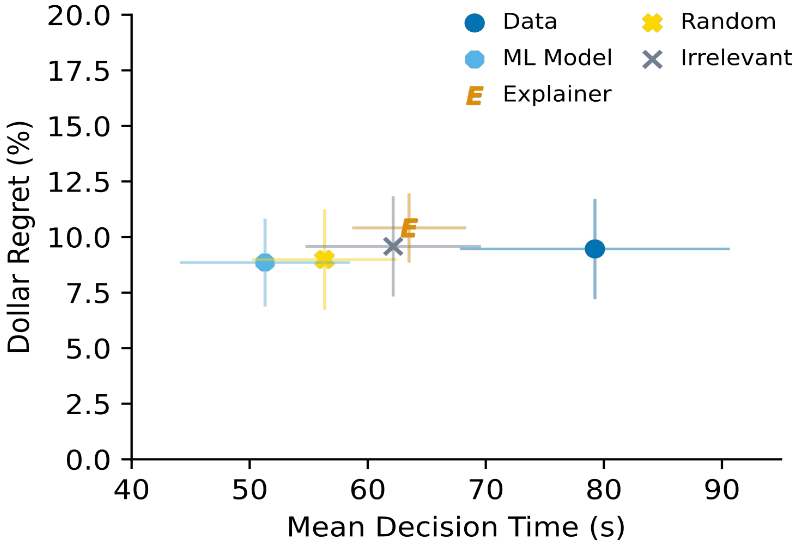
- We assume that the merchant's main objective is to maximize long term and short term revenue
  - Ideally, this should be profit, but we didn't have that data

- Percent Dollar Regret (PDR)

$$PDR = 1 - \frac{\text{Realized Revenue}}{\text{Possible Revenue}}$$

What we found





ML Model improved decisions, but the explanations did not!

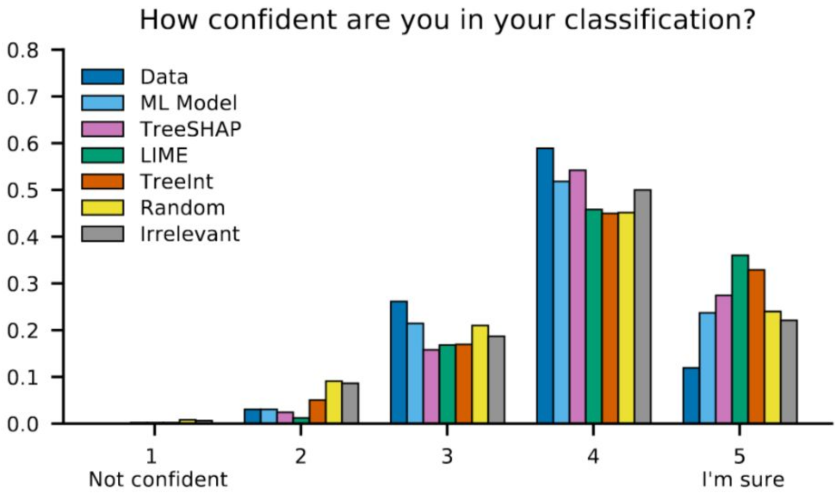
What we found

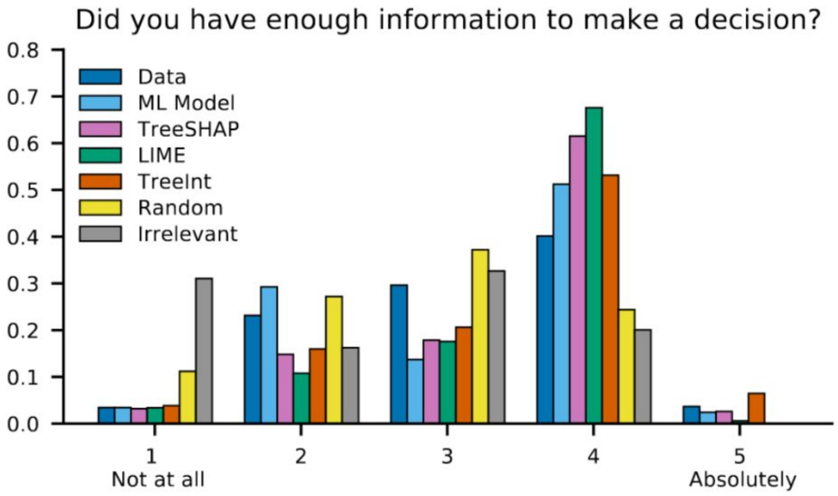
Table 2: Performance summary across the experiment arms

Arm	PDR	Time	Acc	FPR	TPR	Prec	Appr.	Decl.	Escl.
Data Only	9.5	79.2	76.6	18.7	48.6	30.7	71.7	22.4	5.9
Model	8.9	51.3	82.2	12	49.3	42	80.8	17.0	2.2
TreeSHAP	9.7	67.3	81.9	10.6	38.4	38.4	83.1	13.7	3.2
TreeInterpreter	10	65.6	80.9	12.1	40.5	37	81.7	15.5	2.8
LIME	11.6	57.7	83.2	8	38.7	43.3	85.2	12.2	2.6
Random Exp.	9	56.3	82.7	10	38.7	42	85.5	13.3	1.2
Irrelevant Exp.	9.7	62.2	81.2	9.5	29.9	36.5	85.8	12.6	1.6

Escalation rates did not change with explanations, and ad-hoc methods performed similarly to "real" explanations

What we found





However, their confidence, and perceived sense of information goes up with explanations!

(Hopeful) Takeaways / Summary

Explainable ML has the potential of enhancing human-ML collaboration and helping achieve better operational outcomes

However, we need a more practice centered approach:

We need more application grounded evaluation studies for explainable ML

We need studies that capture the nuances of the use-case, and practioners are better suited to understand those nuances

We need to let use-cases inform method development rather and move beyond general-purpose explanation methods

Thank you!

Kasun Amarasinghe
Postdoctoral Researcher
Machine Learning Dept. & Heinz College of Public Policy
Carnegie Mellon University
amarasinghek@cmu.edu, amarasinghe.kas@gmail.com, dssg@cmu.edu

All the icons: *thenounproject.com*

Additional Slides

Defining the specifics of how the system would be used

- What decision is made based on the ML system?
- Who is going to make that decision?
- How would you use the explanations to make that decision?
- What is your measure of success?

Current Evaluation Studies

- Most evaluations focus on the artifact and limited to functionally grounded evaluations

- The most common type of user study is human-grounded
  - Proxy tasks
  - Proxy users

- Three main shortcomings of existing user-studies

Some hypotheses we tested

- Model score improves analyst compared to “data only”
- Explanation improves analyst performance compared to data + ML score
- Explanation impact is different based on which post-hoc explainer is used.
- Explanations generated from an ad-hoc method would be worse compared to those generated by “bona fide” explanation methods.

Mapping the confusion matrix to the application

True Positive:

- Fraudulent transaction declined
- Zero contribution to revenue
- Weight → **0**

True Negative:

- Legitimate transaction approved
- Transaction value is revenue
- Weight → **$trx + ƛ**

False Negative:

- Fraudulent transaction approved
- Lose the item
- return the money + surcharge
- Weight → **- α * $trx**

False Positive:

- Legitimate transaction declined
- Could lose the transaction
- Could lose the customer
- Weight → **(1 - β) * $trx + (1 - δ) * ƛ**

α → item cost as a fraction of sale price + surcharge %
β → Probability of losing the sale
δ → probability of losing the customer
ƛ → long term worth of the customer