THE UNIVERSITY OF MELBOURNE

Explainability

Tim Miller

School of Computing and Information Systems
Centre for AI & Digital Ethics
The University of Melbourne
tmiller@unimelb.edu.au
@tmiller_unimelb

Learning outcomes

At the end of this module, you should be able to:

1. Motivate the need for explainability in digital applications
2. Discuss the challenges of explainable AI
3. Describe the different classifications of explainable AI approaches
4. Understand the foundational models for explainable AI:
    1. Attribution-based explainability
    2. Example-based explanation
    3. Rule-based explanation
    4. Contrastive explanation

Related reading

Required reading for this module:

- "But why?" Understanding explainable artificial intelligence. Tim Miller. *XRDS 25, 3* (Spring 2019), 20–25. https://doi.org/10.1145/3313107

Further reading for those interested:

- Principles and Practice of Explainable Machine Learning. Vaishak Belle and Ioannis Papantonis. https://arxiv.org/pdf/2009.11698.pdf
  This is an overview of explainability algorithms and research

- Interpretable machine learning. Christoph Molnar. https://christophm.github.io/interpretable-ml-book/
  A brilliant e-book on interpretable machine learning that is constantly improving

Outline

1. Why, when, and to whom explainability is important
2. The Challenges of explainable AI
3. Properties of explainable AI approaches
   a) Local vs global explainability
   b) Interpretability vs post-hoc explainability
   c) Model-agnostic vs model-specific explainability
4. Foundational methods of explainable AI:
   a) Feature attribution
   b) Rule-based explanation
   c) Example-based explanation
   d) Contrastive explanation

Motivation: why ask why?

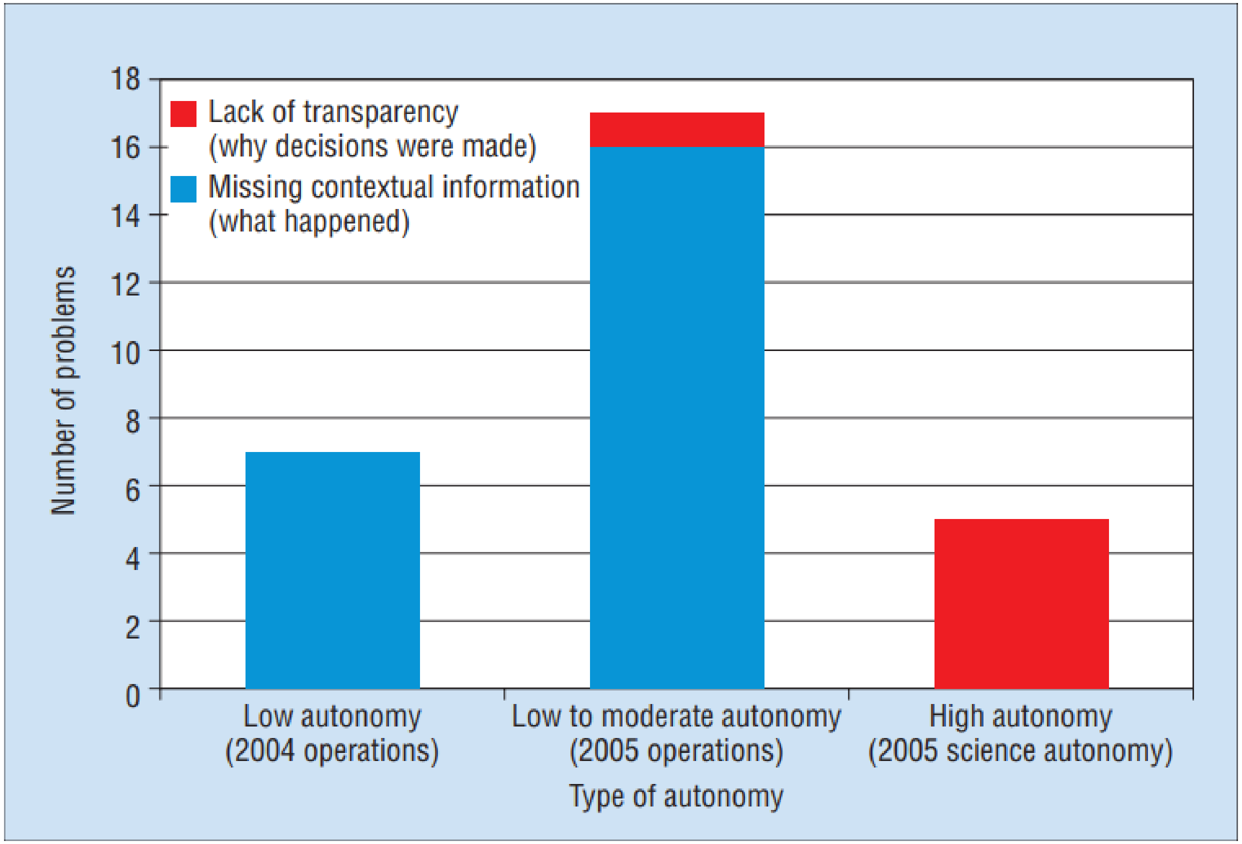
What is Explainable AI?

Explainability (and *interpretability*) is just *understanding*.

Explainable AI is the ability for people to understand AI models and decisions.

Explanation is a mechanism to help people come to an understanding.

Why do we care about explainability?



Source: K. Stubbs et al.: Autonomy and Common Ground in Human-Robot Interaction: A Field Study, IEEE Intelligent Systems, 22(2):42-50, 2007.

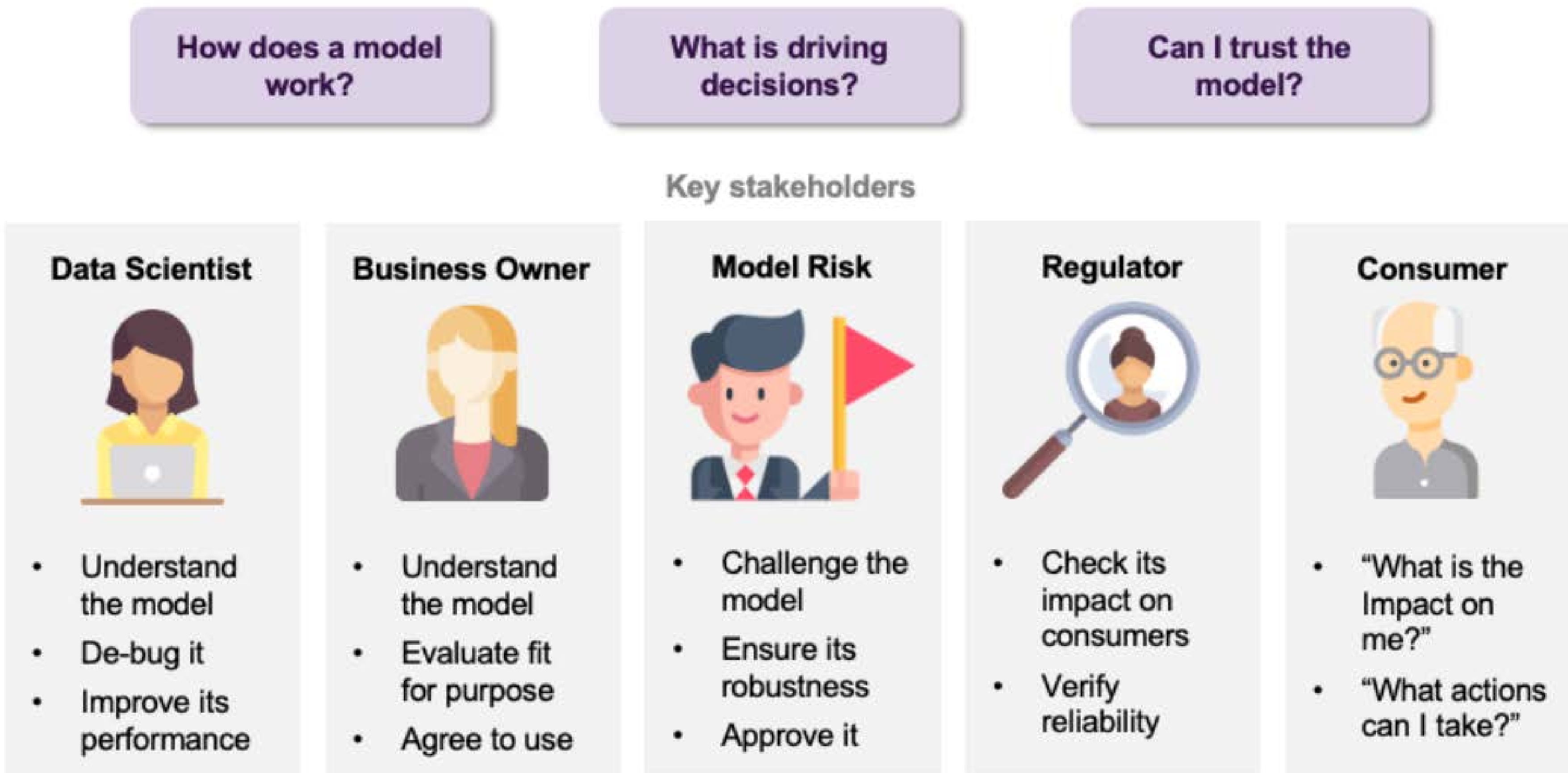
Goals of explainable AI

Why do we care about explainability?

- *Trust*
  Warranted trust and distrust in contracts
- *Ethics*
  Improving the ethical suitability of an application by engendering trust

Is it reasonable to hold some accountable for a decision aided by an algorithm if they cannot understanding *why* the algorithm produced its decisions?

Who cares about explainable AI? And when?



Source: V. Belle and I. Papantonis: Principles and practice of explainable machine learning, *arXiv,* 2020. https://arxiv.org/abs/2009.11698

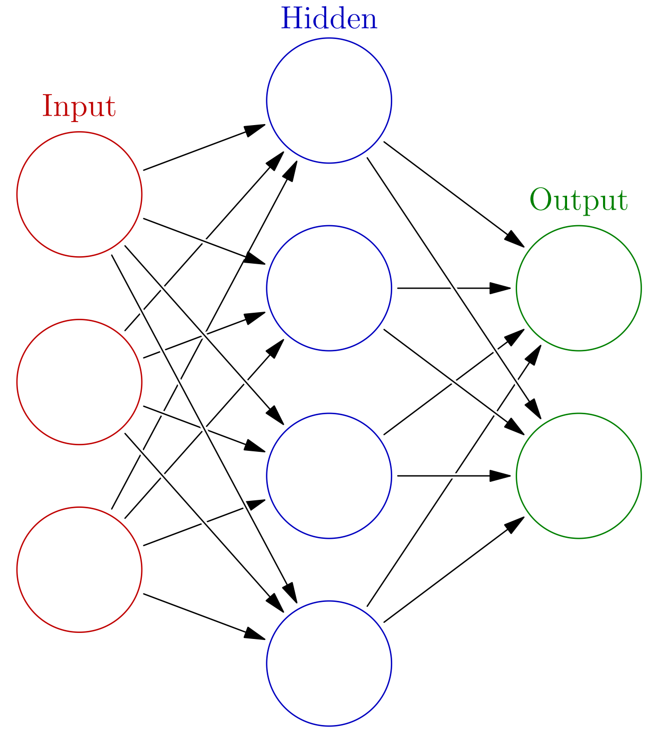
THE UNIVERSITY OF MELBOURNE

The Challenges of Explainable AI

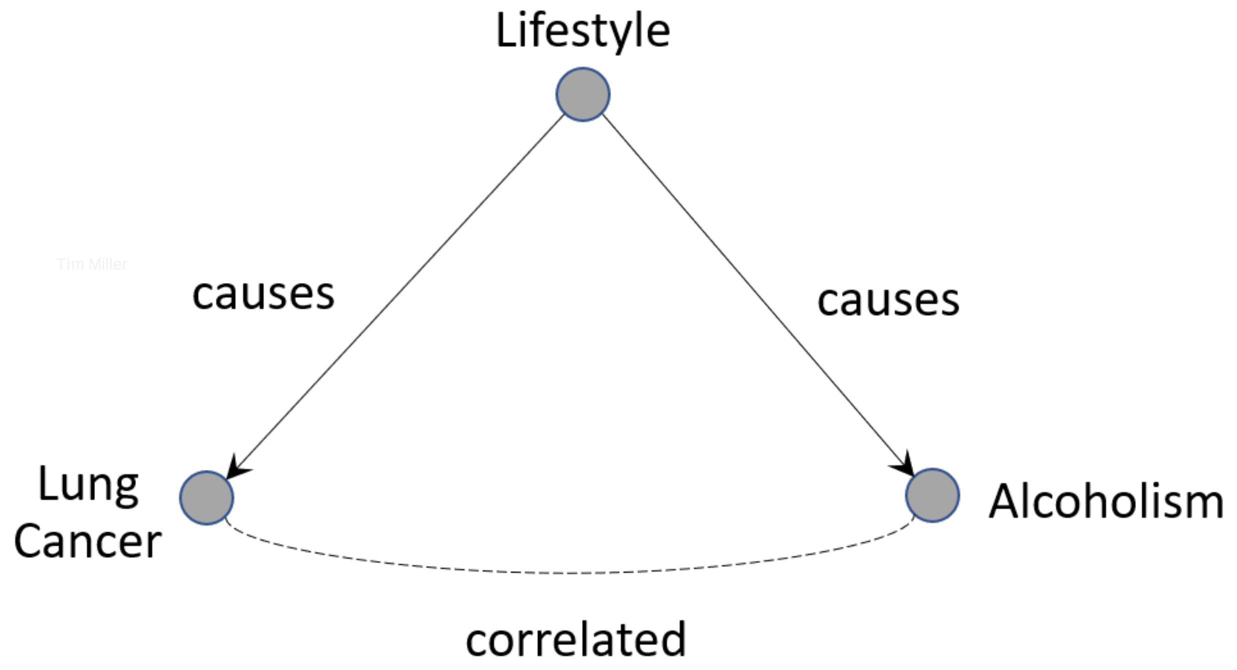
Challenge: Opacity

if Respiratory-Illness=Yes **and** Smoker=Yes **and** Age>=50
then Lung Cancer
elif Risk-LungCancer=Yes **and** Blood-Pressure>=0.3
then Lung Cancer
elif Risk-Depression=Yes **and** Past-Depression=Yes
then Depression
elif BMI>=0.3 **and** Insurance=None
and Blood-Pressure>=0.2 **then** Depression
elif Smoker=Yes **and** BMI>=0.2 **and** Age>=60
then Diabetes
elif Risk-Diabetes=Yes **and** BMI>=0.4 **and** Prob-Infections>=0.2 **then** Diabetes
Else Null

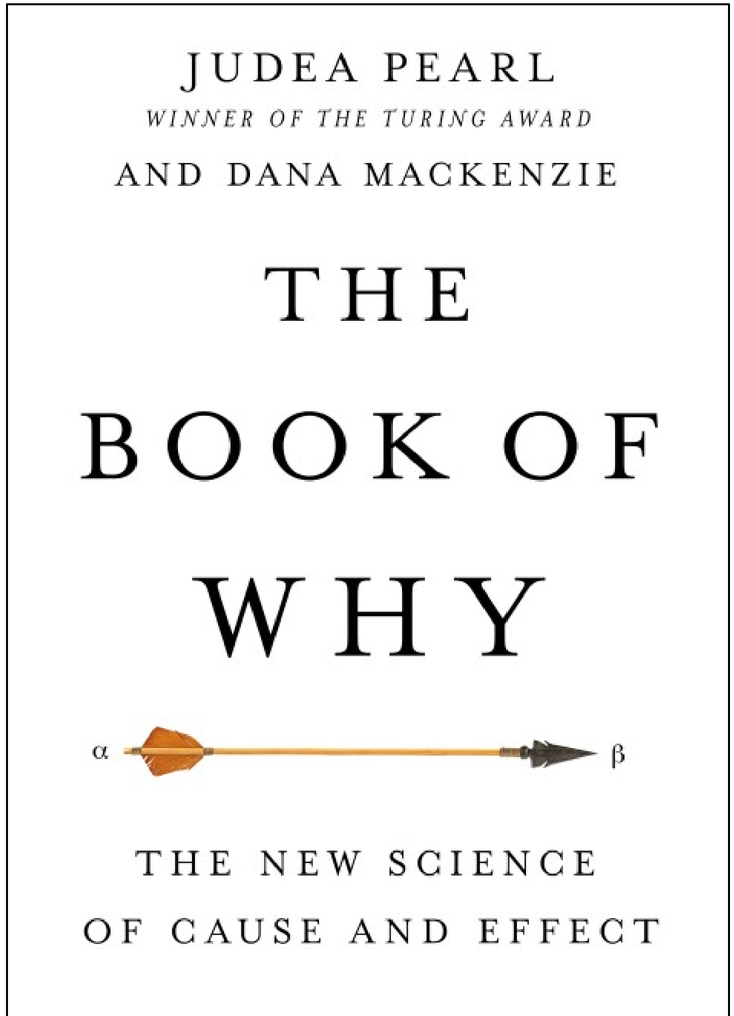
VS.



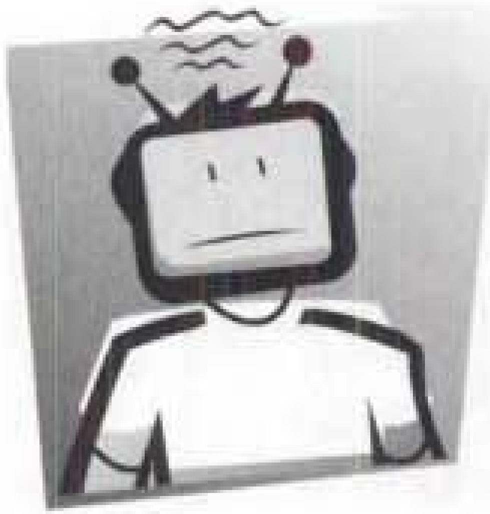
Challenge: Casuality



Challenge: Casuality



Challenge:
The human problem



Homo logicus

wants control—
accepts complexity
as trade-off

Homo sapiens

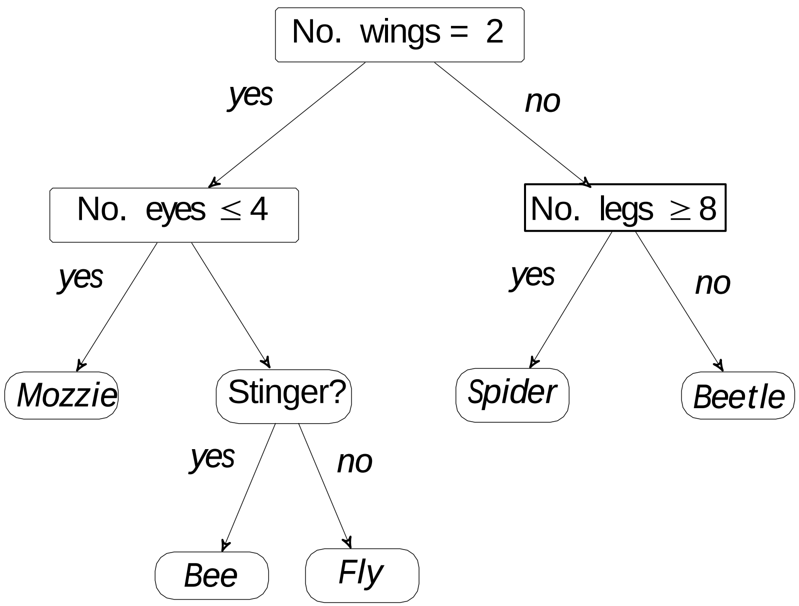
wants simplicity—
accepts less control
as trade-off

Properties of explainable AI approaches

Global vs. local

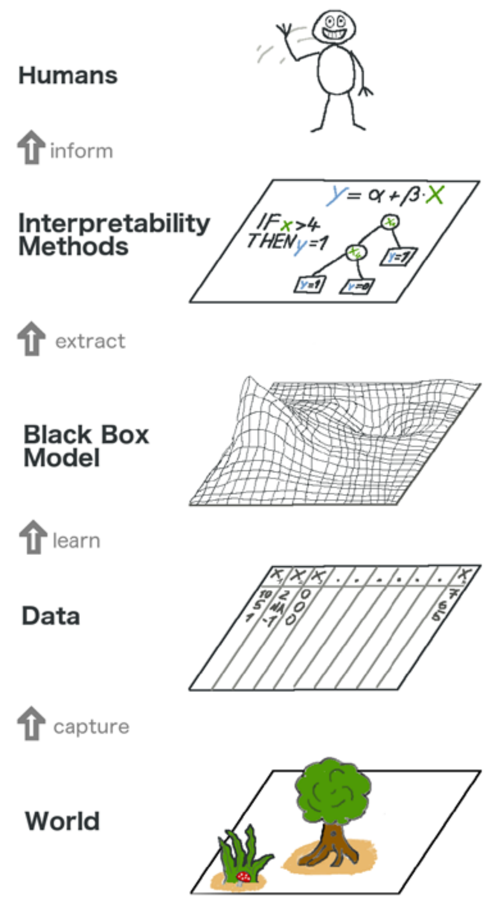
vs.

Intrinsic vs. post-hoc

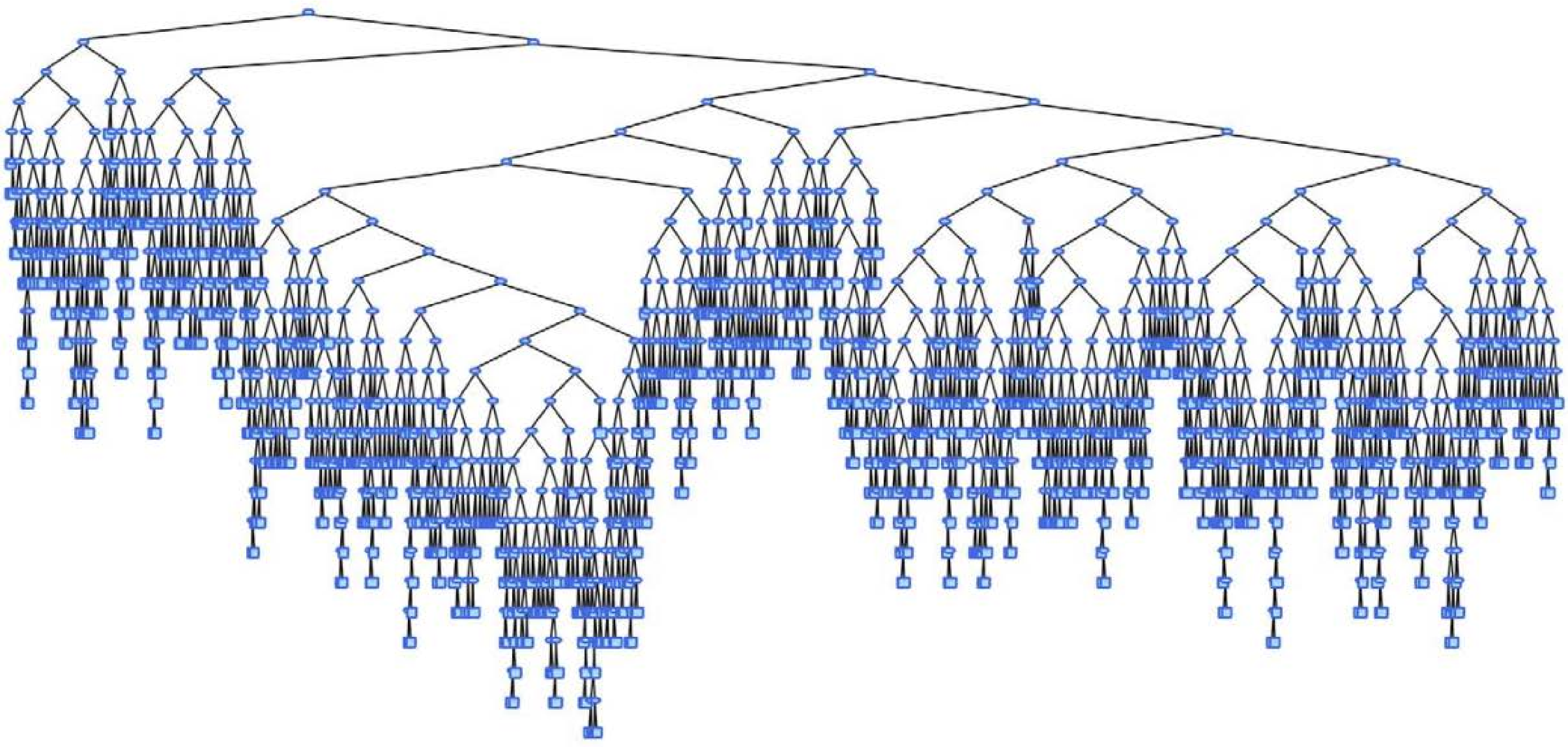


$$f(x) = 2.4x_1 + 0.12x_2 + \ldots 1.1x_n$$

vs.



Intrinsic *and* post-hoc



Source: Stiglic G, Kocbek S, Pernek I, Kokol P: Comprehensive Decision Tree Models in Bioinformatics. PLoS ONE 7(3): e33812, 2012.

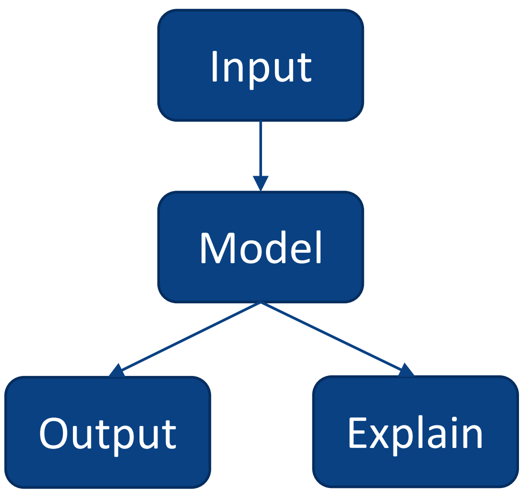
Model-agnostic vs model-specific

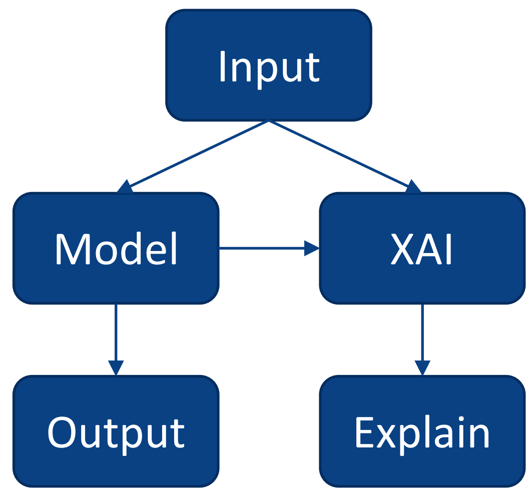
Model specific:

- Uses inner workings and properties of models to derive explainability mechanisms

Model agnostic:

- Uses only inputs and outputs to derive explainability mechanisms





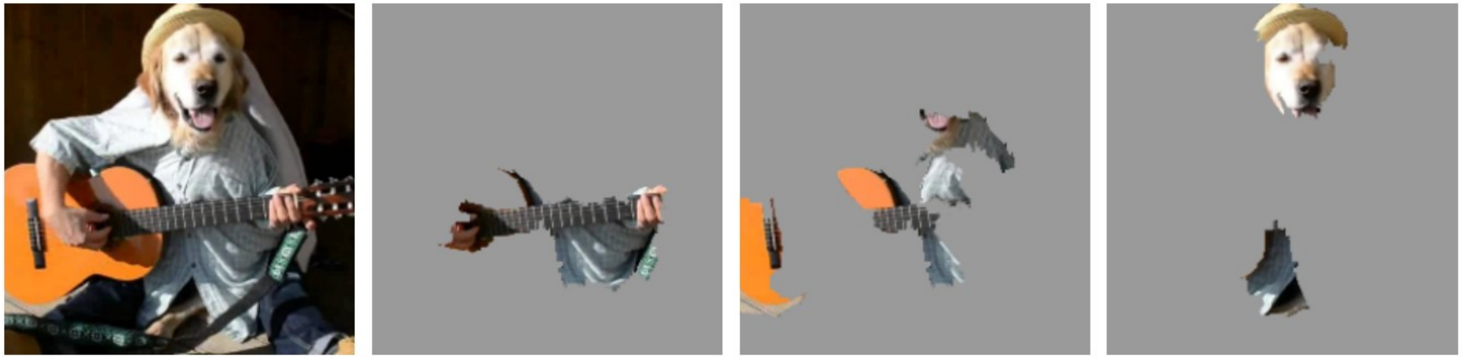
THE UNIVERSITY OF MELBOURNE

Foundational methods in explainable AI

Attribution-based explanations

Source: T. Miller: "But why?" Understanding explainable artificial intelligence. *XRDS 25, 3* (Spring 2019), 20–25. https://doi.org/10.1145/3313107

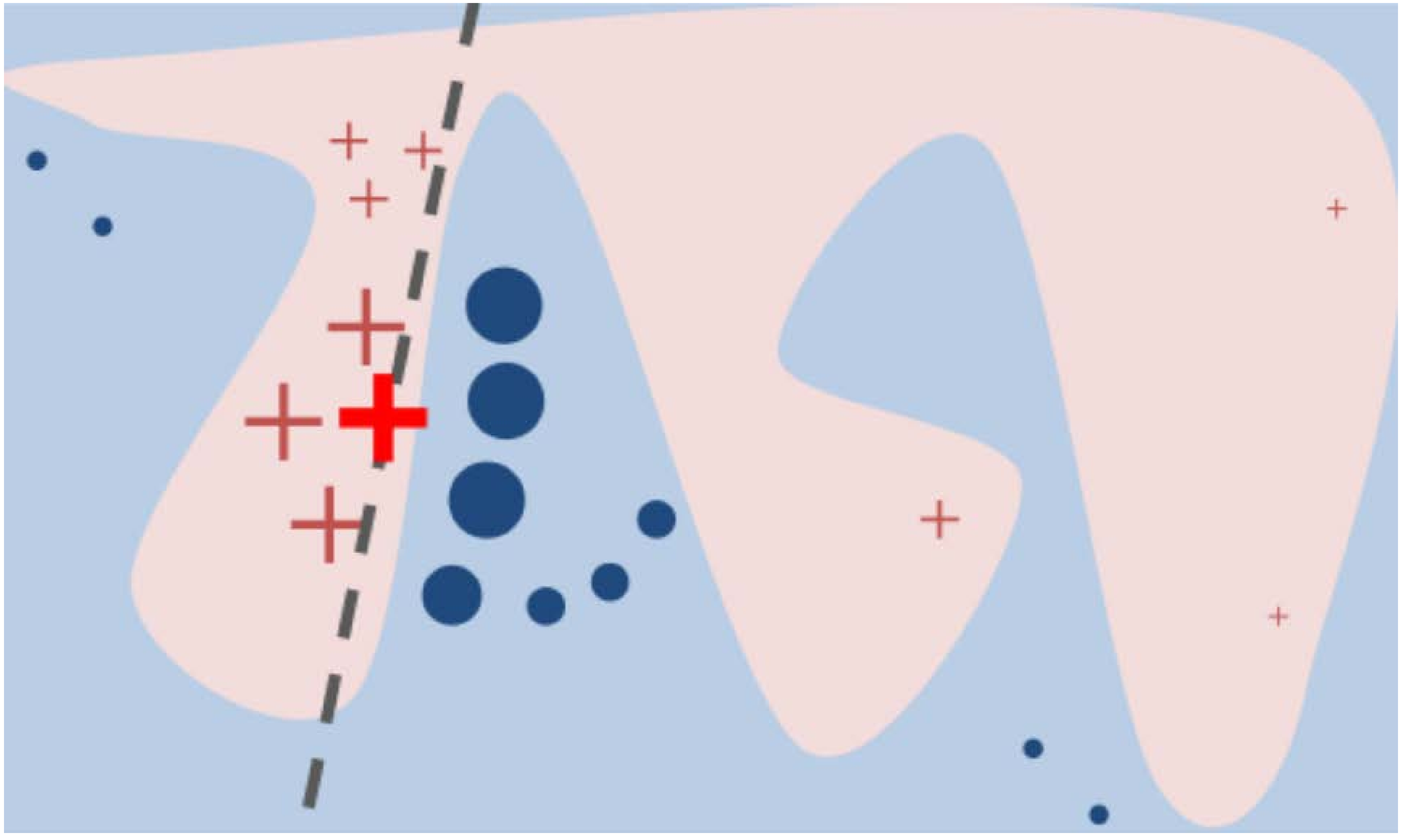
Attribution-based explanations



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Source: Ribeiro et al.: Why should I trust you?: Explaining the predictions of any classifier. In SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

LIME: Local Interpretable, Model-agnostic Explanations



Source: Ribeiro et al.: Why should I trust you?: Explaining the predictions of any classifier. In SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

Example-based explanation: Prototypes

Source: Kim et al.: Examples are not enough, learn to criticize! Criticism for interpretability. In NeurIPS. 2016.

Rule-based explanation

- Extract rules post-hoc or learn interpretable rules directly

if Respiratory-Illness=Yes **and** Smoker=Yes **and** Age>=50
then Lung Cancer
elif Risk-LungCancer=Yes **and** Blood-Pressure>=0.3
then Lung Cancer
elif Risk-Depression=Yes **and** Past-Depression=Yes
then Depression
elif BMI>=0.3 **and** Insurance=None **and** Blood-Pressure>=0.2
then Depression
elif Smoker=Yes **and** BMI>=0.2 **and** Age>=60
then Diabetes
elif Risk-Diabetes=Yes **and** BMI>=0.4 **and** Prob-Infections>=0.2
then Diabetes
Else Null

Contrastive explanation

"The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case."

D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin. 107 (1) (1990) *65–81.*

Contrastive Explanation — The Difference Condition

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *Knowledge Engineering Review, (in print).* https://arxiv.org/abs/1811.03163

Contrastive Explanation — The Difference Condition

Why is it a fly?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *Knowledge Engineering Review, (in print).* https://arxiv.org/abs/1811.03163

Contrastive Explanation — The Difference Condition

Why is it a fly rather than a beetle?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	2	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	5	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *Knowledge Engineering Review, (in print).* https://arxiv.org/abs/1811.03163

Contrastive Explanation — The Difference Condition

Why is it a fly rather than a beetle?

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✘	8	✘	0
Beetle	6	✘	**2**	✔	2
Bee	6	✔	5	✔	4
Fly	6	✘	**5**	✔	2

T. Miller. Contrastive Explanation: A Structural-Model Approach, *Knowledge Engineering Review, (in print).* https://arxiv.org/abs/1811.03163

Explainability: summary

Explainability

- Different people with different explainability needs
- Trust and ethics
- Human and technical challenges
  - Opacity
  - Causality
  - Human interpretation

Explainability approaches

- Classifications
  - Local vs global
  - Interpretable vs post-hoc
  - Model-agnostic vs model specific
- Key approaches
  - Attribution
  - Rules
  - Examples
  - Contrasts

Thank you