David Watson
16 April, 2019

UNIVERSITY OF OXFORD

Digital Ethics Lab

Oxford Internet Institute
University of Oxford

INTRODUCTION TO EXPLAINABLE AI

Clinical Challenges and Opportunities

OVERVIEW

1. What's AI?
2. What's an Explanation?
3. Why Explain?
4. Explain What?
5. Explain How?
6. How Do We Measure Explanations?

MACHINE LEARNING

Supervised Learning:

Given feature matrix $\mathbf{X}$, predict outcome Y using algorithm f.

$$f: \mathbf{X} \rightarrow Y$$

MACHINE LEARNING

Supervised Learning:

Given feature matrix $\mathbf{X}$, predict outcome Y using algorithm f.

$$f: \mathbf{X} \rightarrow Y$$

Unsupervised Learning:

Given feature matrix $\mathbf{X}$, use algorithm f to do…something.

(E.g., detect outliers, project $\mathbf{X}$ in low dimensions, cluster observations, etc.)

UBIQUITY OF ML

ML is currently used to:

- Filter spam
- Recommend movies
- Label cat pix
- Detect fraud
- Predict sports outcomes
- Read image to text
- Beat you at chess

UBIQUITY OF ML

ML is currently used to:

- Filter spam
- Recommend movies
- Label cat pix
- Detect fraud
- Predict sports outcomes
- Read image to text
- Beat you at chess

But also to:

- Recognize your face
- Detect military targets
- Predict criminal recidivism
- Screen job applicants
- Track online behavior
- Guess if you’re gay
- Tweet racist vitriol

CLINICAL ML IS ALREADY HERE

- Microsoft's InnerEye helps NHS radiologists detect cancerous tumours
- DeepMind Health has partnered with Moorfields Eye Hospital to train models to detect retinal pathologies
- Watson for Oncology is (in?)famously deployed at New York's Memorial Sloan Kettering Cancer Center

GOOD NEWS

Good news: algorithms are very good at predicting things!

GOOD NEWS & BAD NEWS

Good news: algorithms are very good at predicting things!

Bad news: algorithms are very bad at explaining things!

EXPLANATION

The deductive-nomological model (Hempel, 1965)

The explanation for some event *E* consists of two components:

1) a non-empty set of observation statements $S = \{s_1, s_2, s_3 \ldots s_n\}$; and

2) at least one law-like generalisation *L*, such that

$$(S \;\&\; L) \rightarrow E.$$

EXPLANATION

Objection 1: DN model is unnecessary

s_1: Patient A has infection x
s_2: Patient A receives treatment
L_1: 0% of untreated patients with infection x survive
L_2: 99% of treated patients with infection x survive

E: Patient A survives

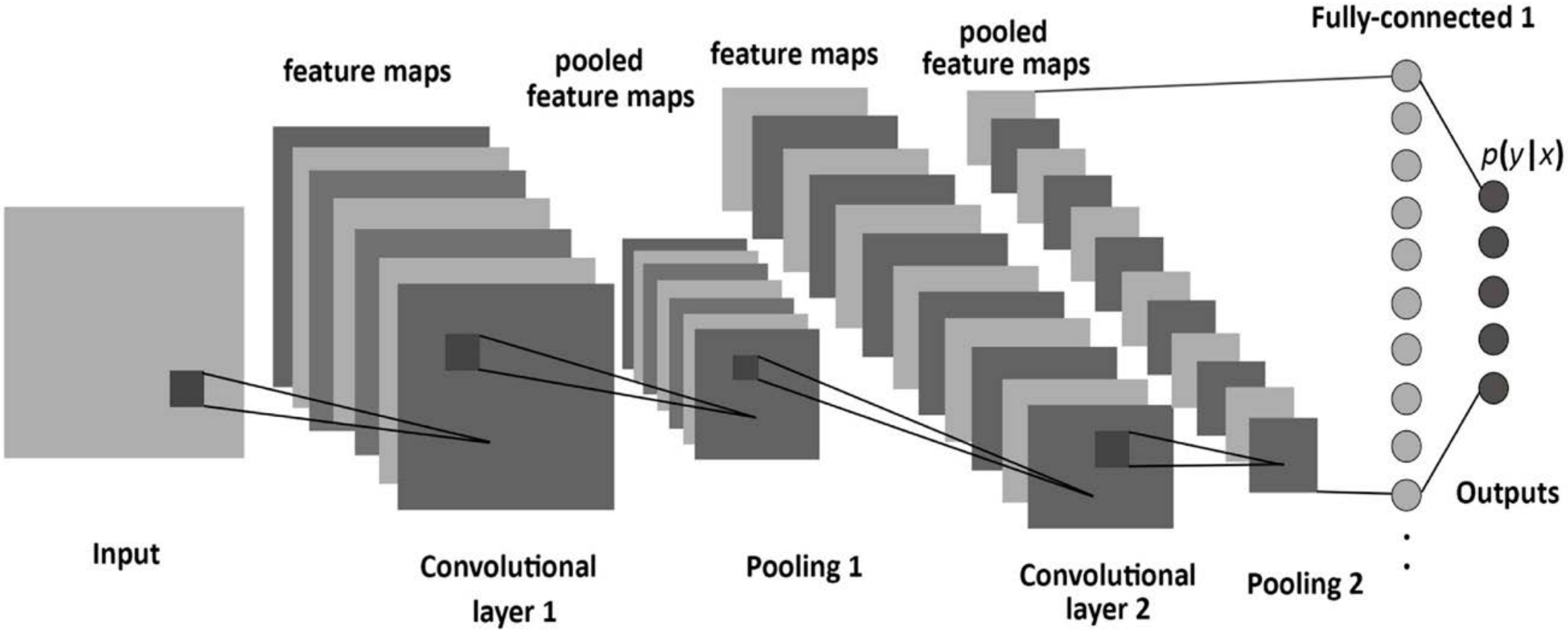
EXPLANATION

Objection 2: DN model is insufficient

S: John Jones is a male who has been taking birth control pills regularly
L: All males who take birth control pills regularly fail to get pregnant
----
E: John Jones fails to get pregnant

EXPLANATION

Objection 3: DN model fails when *L* or *S* is too complex



EXPLANATION

Miller (2017) surveys a wide array of literature on explanation and highlights four key points. Successful explanations are:

- Contrastive
- Selective
- Causal
- Social

WHY EXPLAIN?

Reason 1: To Audit

- Fairness, accountability, and transparency (FAT ML)

WHY EXPLAIN?

Reason 1: To Audit

- Fairness, accountability, and transparency (FAT ML)
- European Union’s 2018 General Data Protection Regulation (GDPR) may provide data subjects a “right to explanation” (Goodman & Flaxman, 2016)

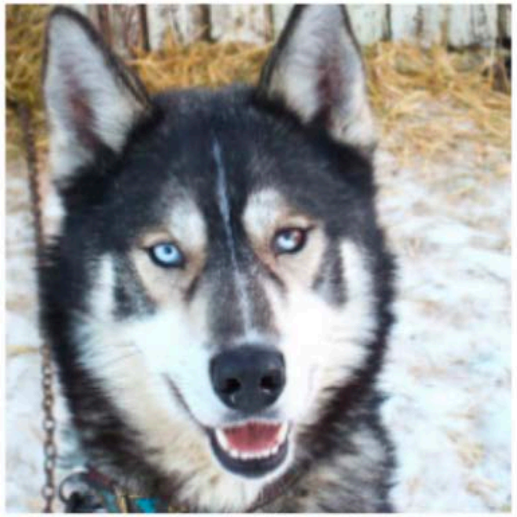
WHY EXPLAIN?

Reason 1: To Audit

- Fairness, accountability, and transparency (FAT ML)
- European Union’s 2018 General Data Protection Regulation (GDPR) may provide data subjects a “right to explanation” (Goodman & Flaxman, 2016)
- Or maybe not (Wachter et al., 2017)

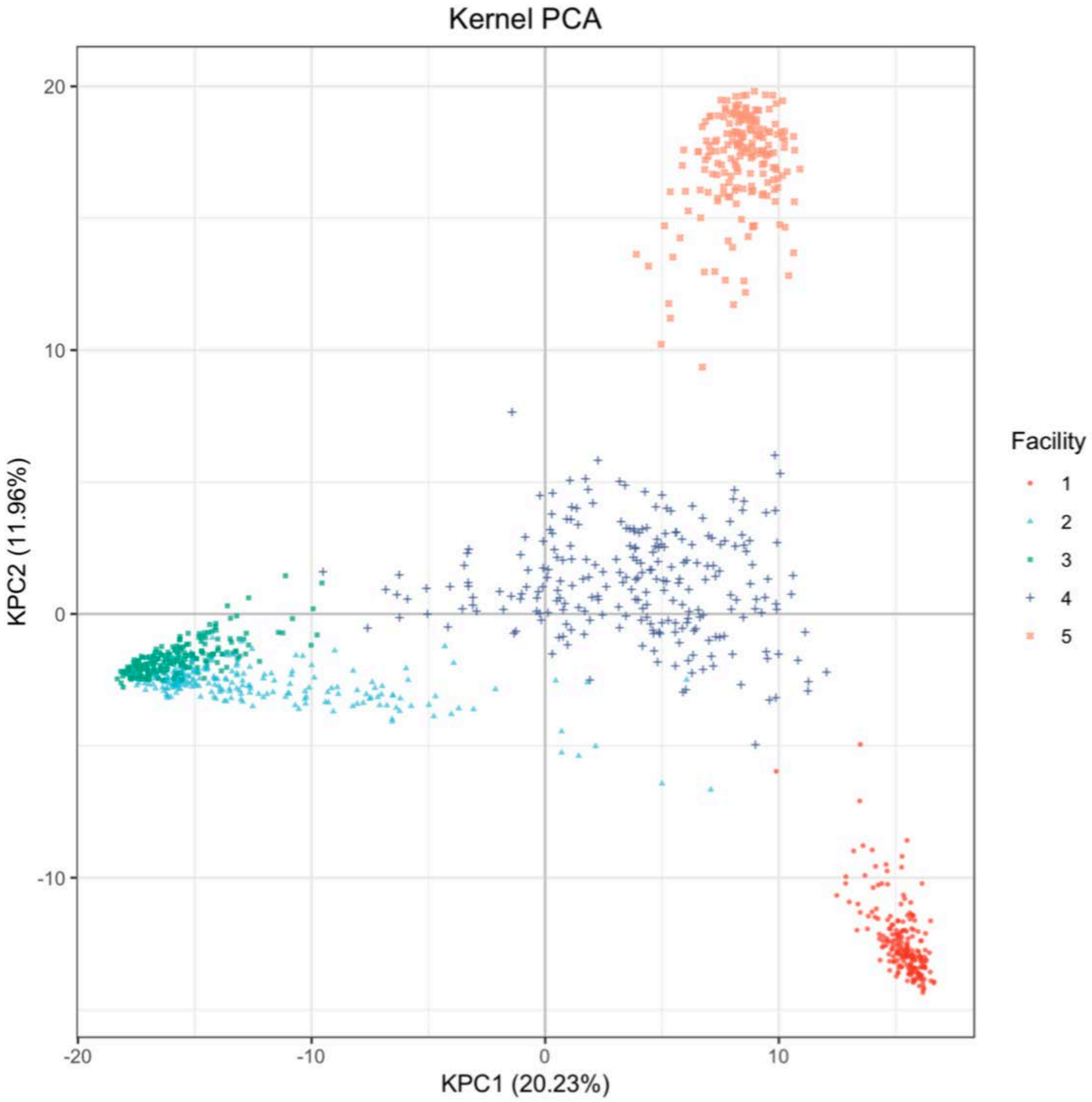
WHY EXPLAIN?

Reason 2: To Validate



(a) Husky classified as wolf

(b) Explanation

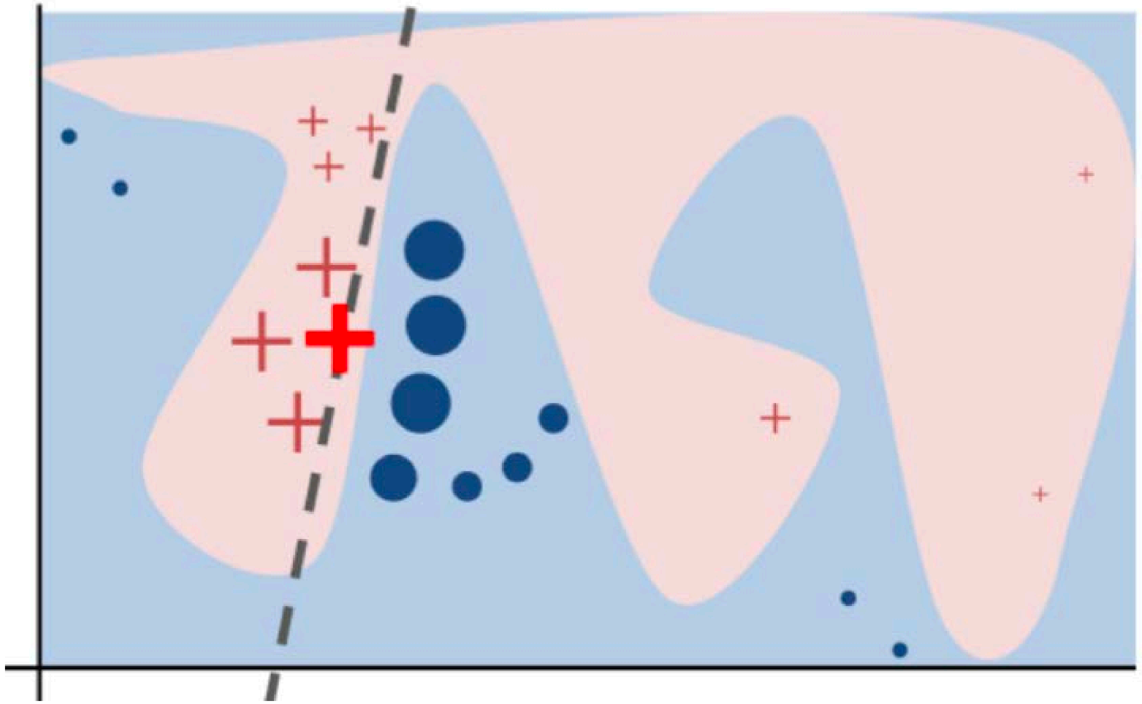


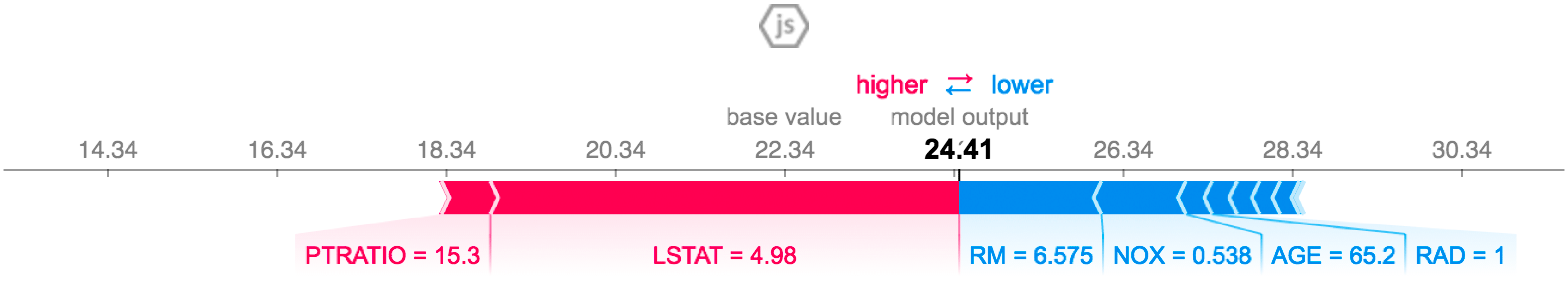
WHY EXPLAIN?

Reason 3: To Discover

EXPLAIN WHAT?

Global vs. Local





EXPLAIN WHAT?

Contrastive Counterfactuals

EXPLAIN WHAT?

Model-Specific

DeepLift
(Shrikumar et al., 2017)

RF permutations
(Breiman, 2001)

Fixed-X knockoffs
(Barber & Candès, 2015)

Model-Agnostic

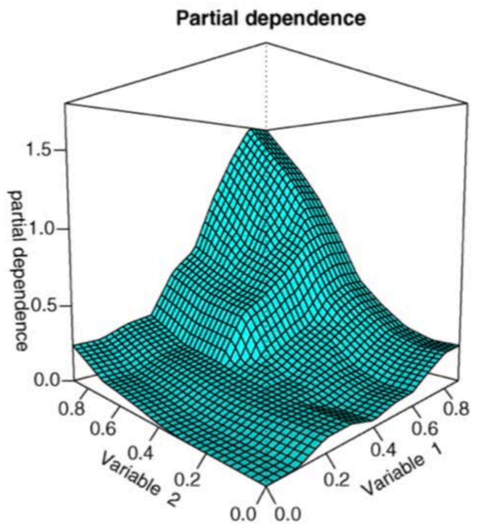
LIME
(Ribeiro et al., 2016)

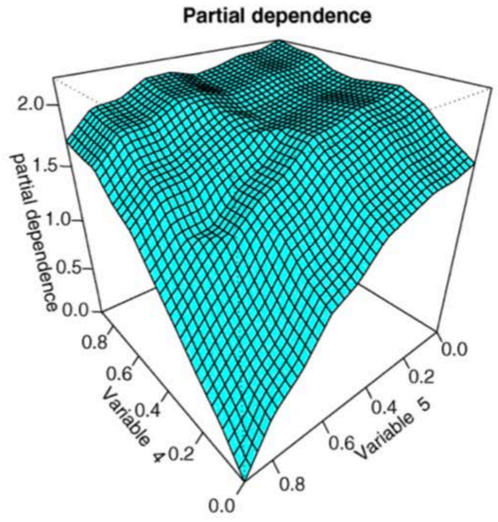
SHAP
(Lundberg & Lee, 2017)

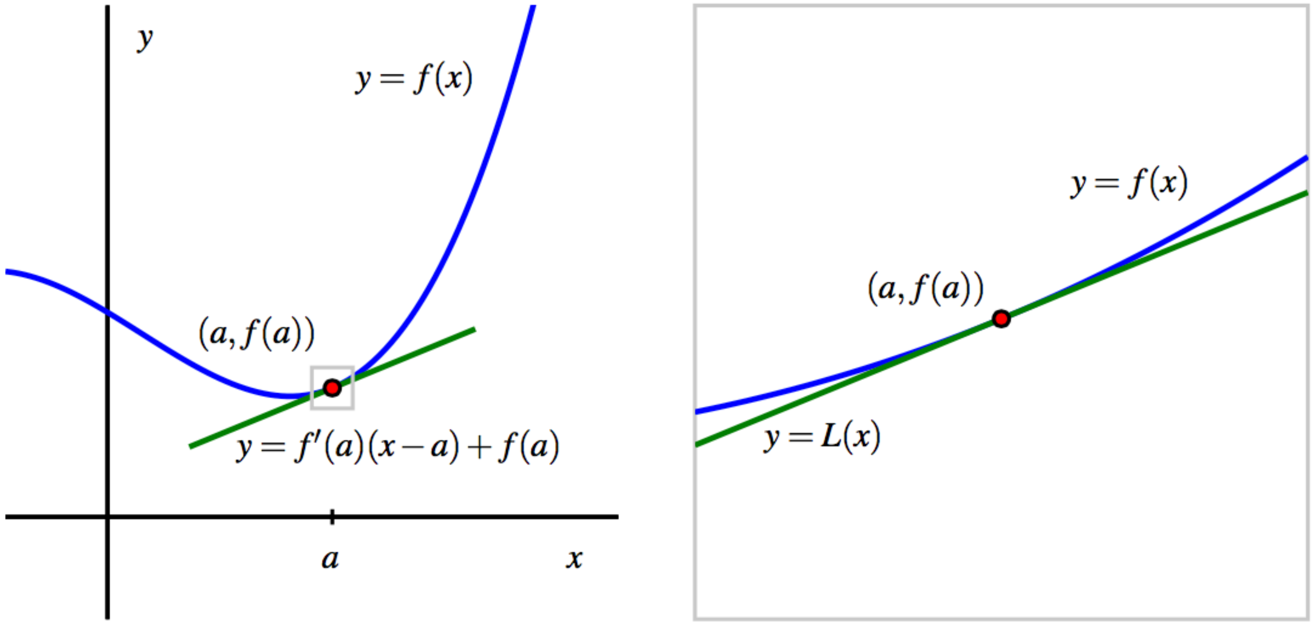
SBRL
(Yang et al., 2017)

EXPLAIN HOW?

Feature Importance

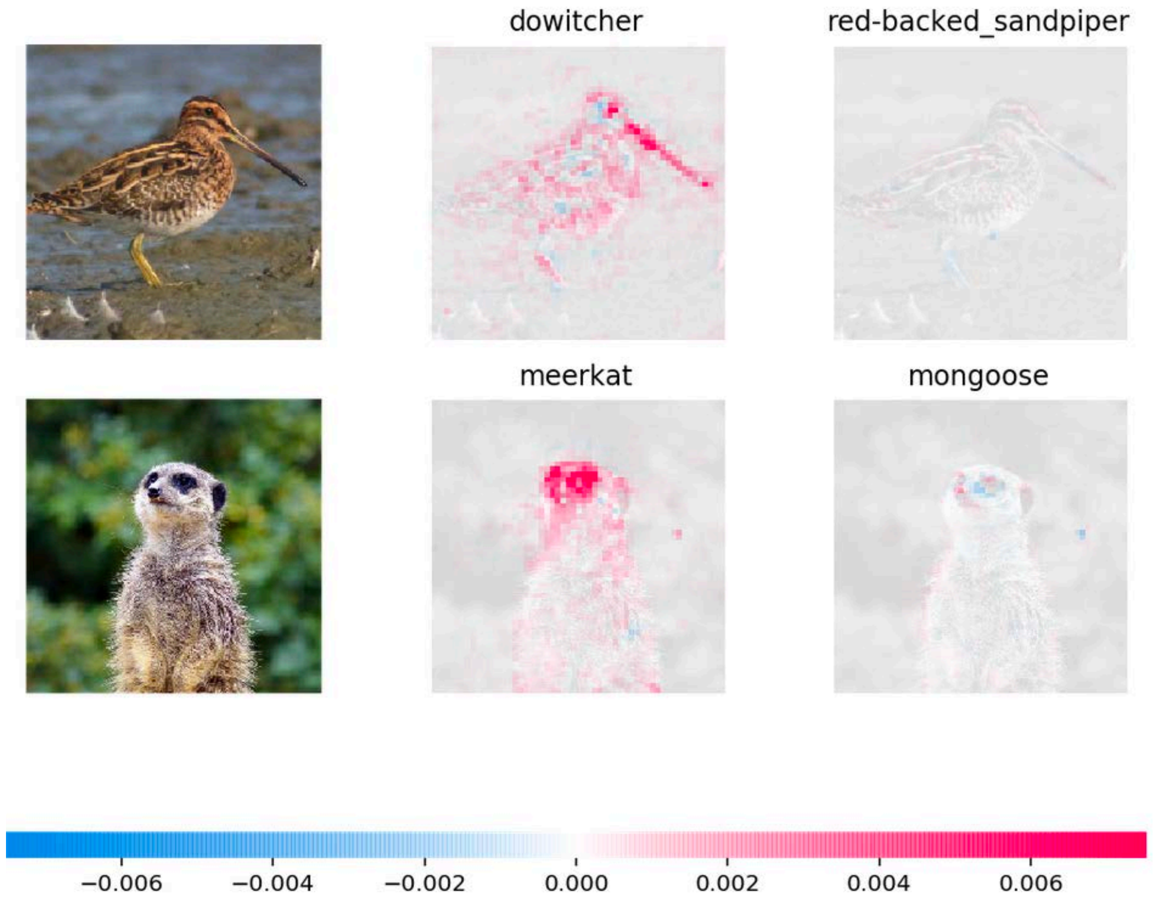






EXPLAIN HOW?

Local Linear Approximations



EXPLAIN HOW?

Saliency Maps

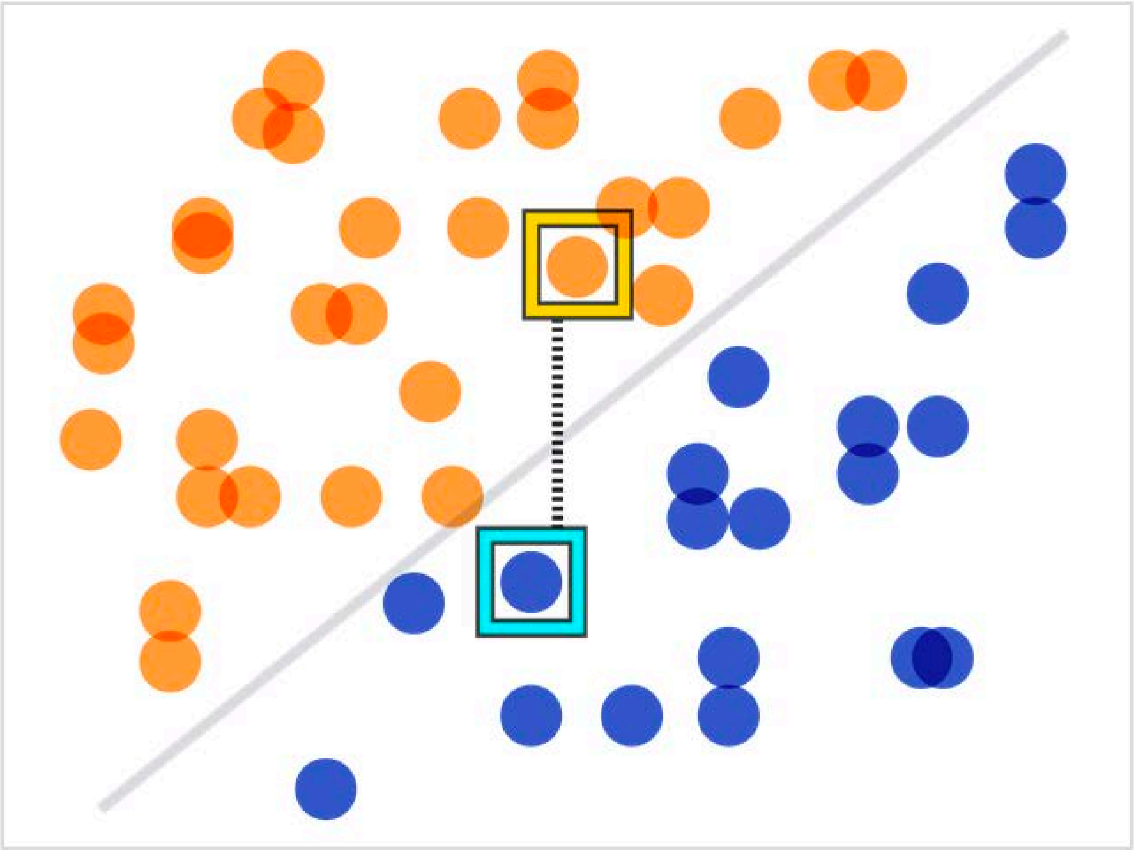
if hemiplegia **and** age $>$ 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age $>$ 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age $\leq$ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

EXPLAIN HOW?

Rule Lists

EXPLAIN HOW?

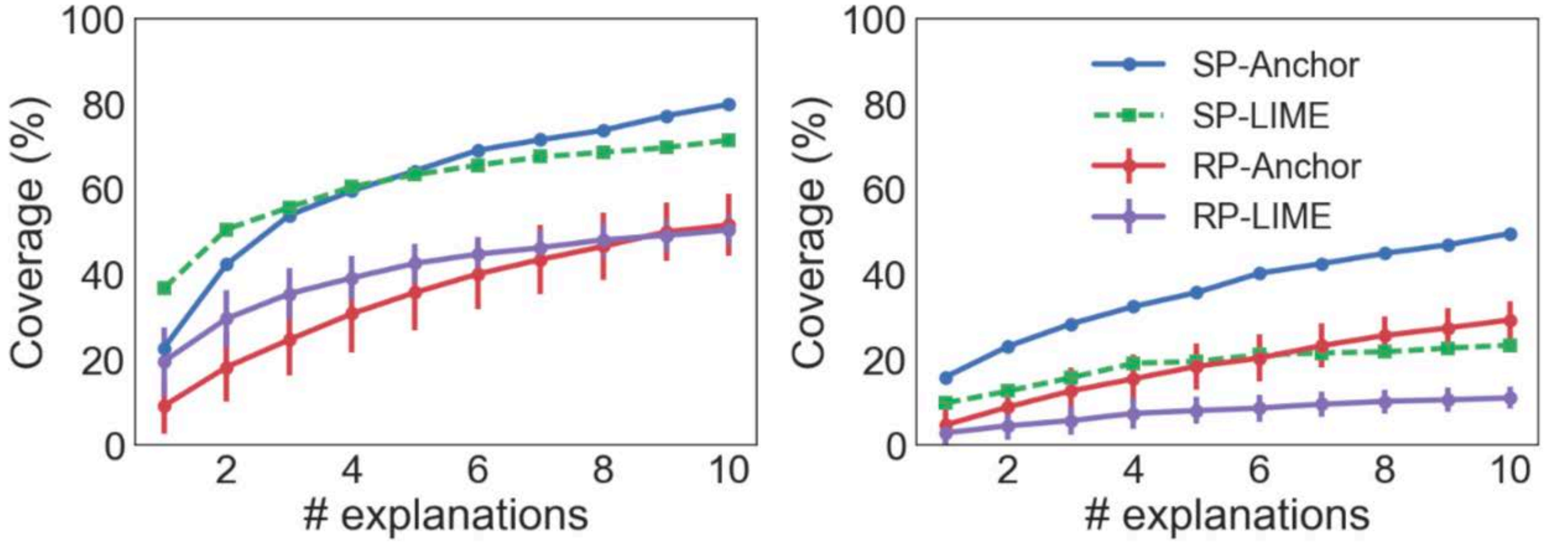
Counterfactuals



MEASURING EXPLANATIONS

"[T]he task of interpretation appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable." (Lipton, 2017, p. 1)

"Unfortunately, there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking." (Doshi-Velez & Kim, 2017, p. 1)



(a) *adult* dataset

(b) *rcdv* dataset

HUMANS VS. MATH

Fidelity & Sparsity

CONCLUSION

CONCLUSION

- Explanations are thoroughly context-dependent: who's the audience? What's the goal?
- Tradeoffs between fidelity to the target model and explanatory parsimony are inevitable
- Explanations are a process, not a product

REFERENCES

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, *43*(5), 2055–2085.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 1–33.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning, (Ml), 1–13. Retrieved from http://arxiv.org/abs/1702.08608

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, *9*(3), 1350–1371.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. Retrieved from http://arxiv.org/abs/1606.03490

Lundberg, S. M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *arXiv* preprint 1706.07269.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). New York, NY, USA: ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Retrieved from http://arxiv.org/abs/1704.02685.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887.

Yang, H., Rudin, C., & Seltzer, M. (2016). Scalable Bayesian Rule Lists.

THANKS!

For questions, complaints, and/or readings, email me:

david.watson@oii.ox.ac.uk