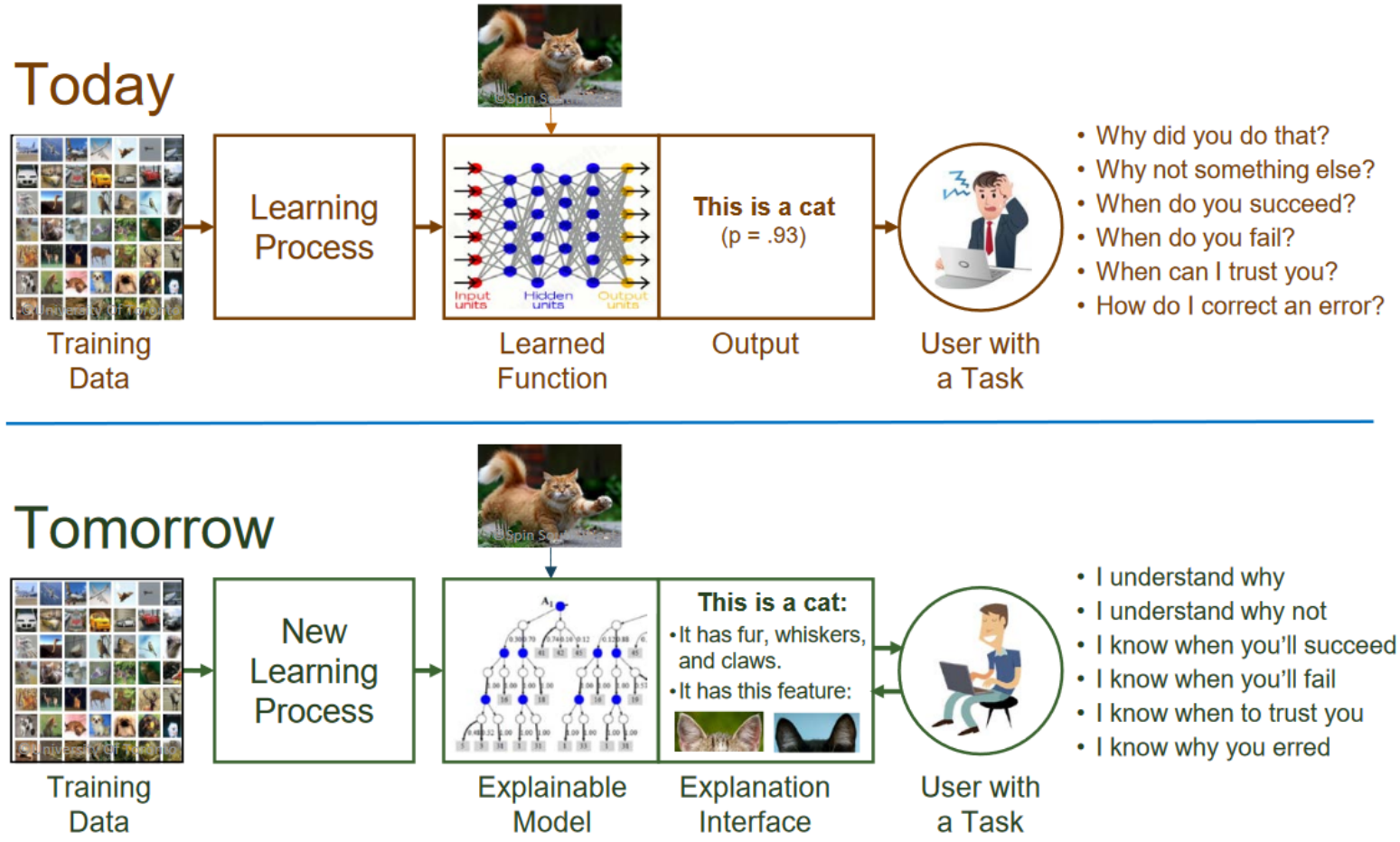
Explainable DNNs: what has your network really learnt?

Catherine Easdon

Motivation



[DARPA]

Motivation

DNNs are being deployed in contexts where the trustworthiness of the model really matters!

- Self-driving cars [R1]
- Healthcare [R2, R3]
- Criminal justice [R4]
- Finance [R5]
- Warfare [R6, R7]
- Social media (societal and political impact) [R8, R9]

Motivation

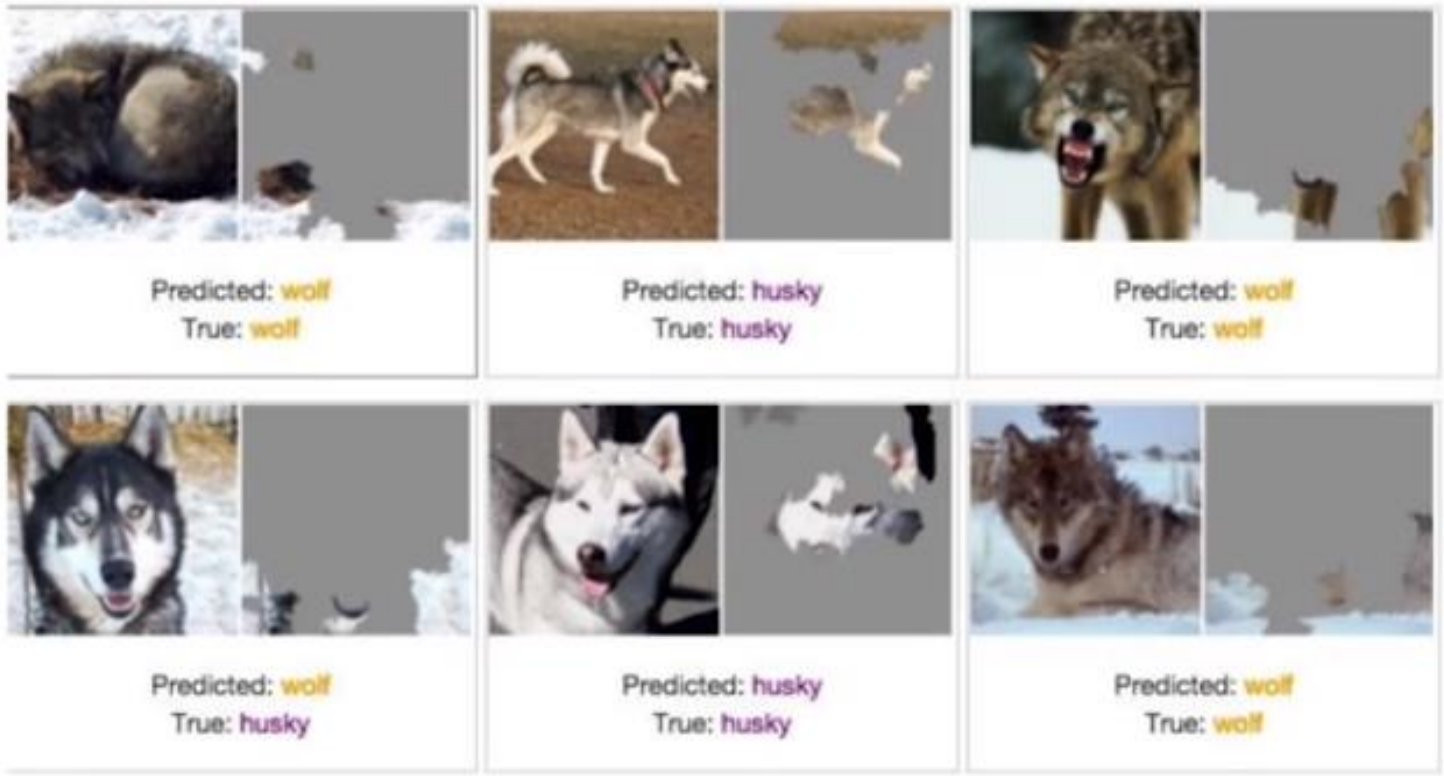
1. Articles 13 and 14 – Right to be informed

Given the core principle of transparency underpinning the GDPR, controllers must ensure they explain clearly and simply to individuals how the profiling or automated decision-making process works.

[R10]

High validation accuracy is not enough

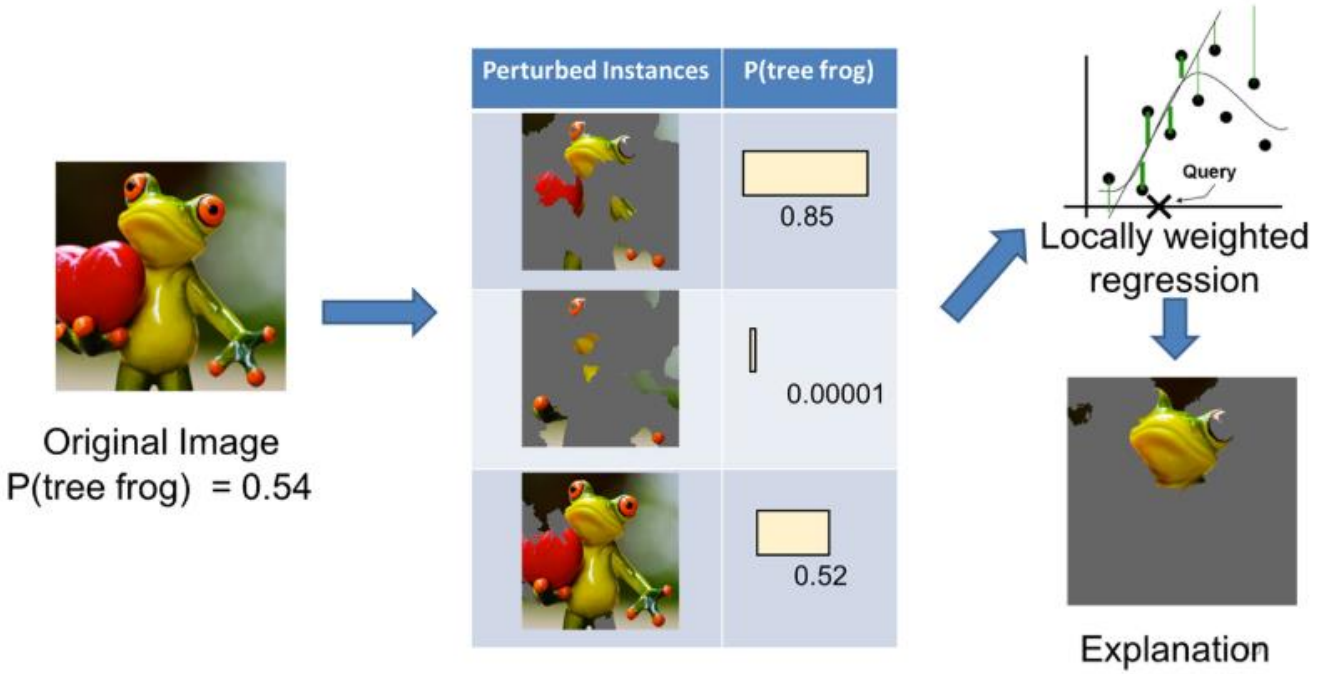
We need proof that the network has learnt **relevant** features – this helps us assess the network's susceptibility to **adversarial inputs** too



[R11]

LIME

Local
Interpretable
Model-agnostic
Explanations

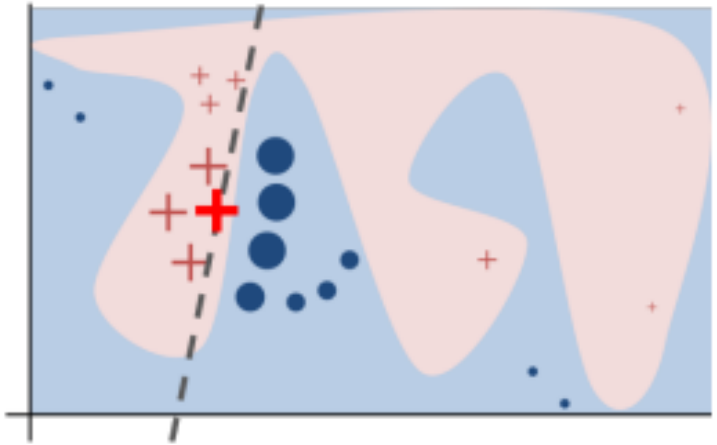


[R11, R12]

LIME

Original Image

Interpretable Components

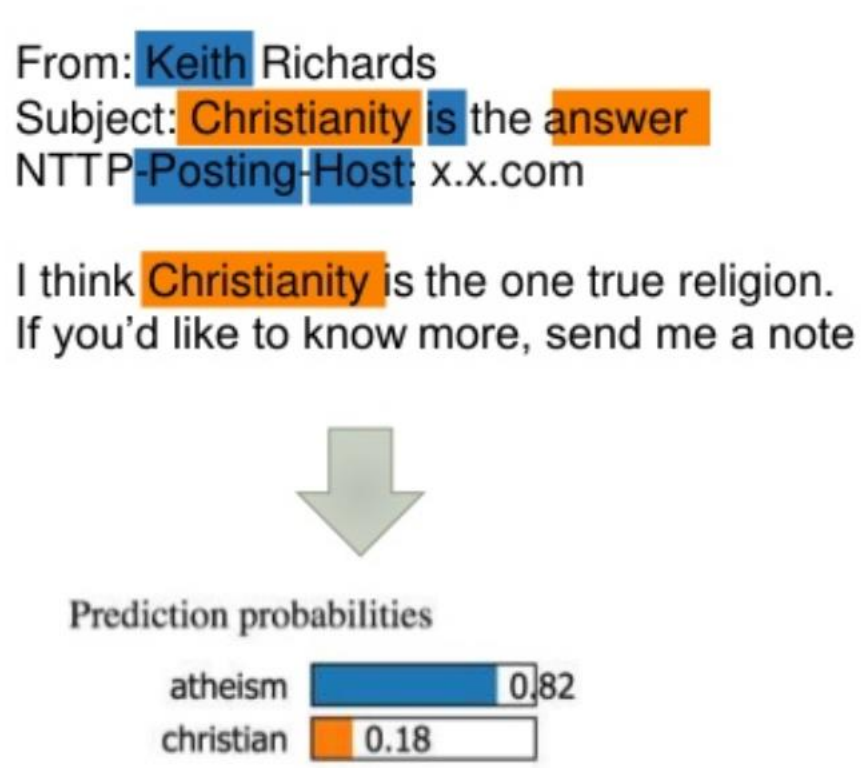


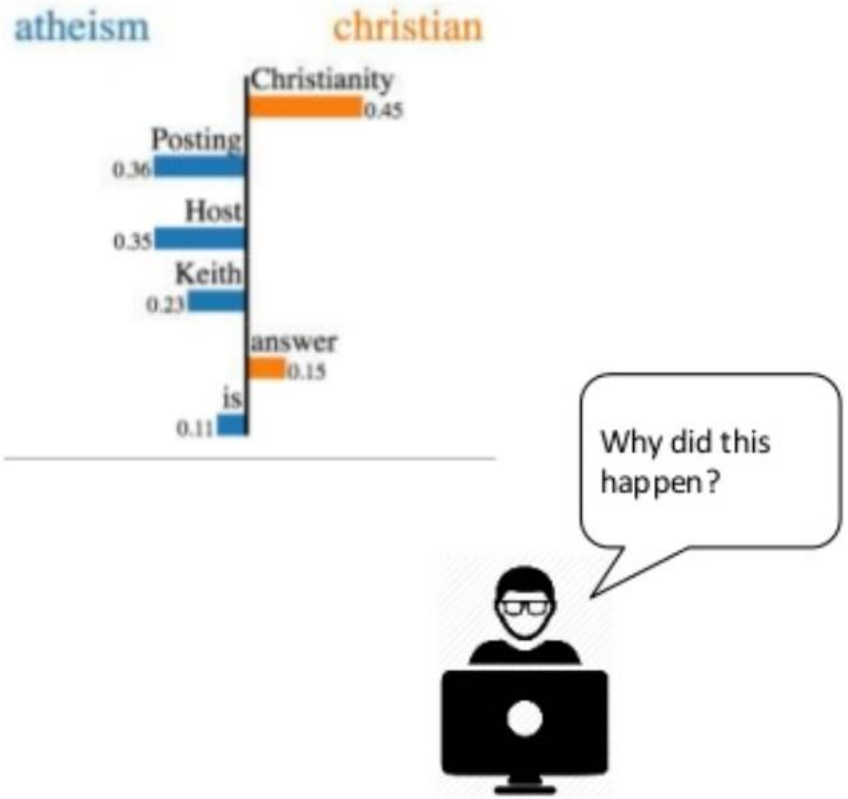
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

***L*(f, g, π_x)** : g's error in approximating f in the locality π_x
x : input being explained
f : model being explained (which is a black box to us)
g : an explanation model from the class of interpretable models G
π_x : measure of proximity to x
Ω : measure of the complexity of the explanation

[R11, R12]

LIME

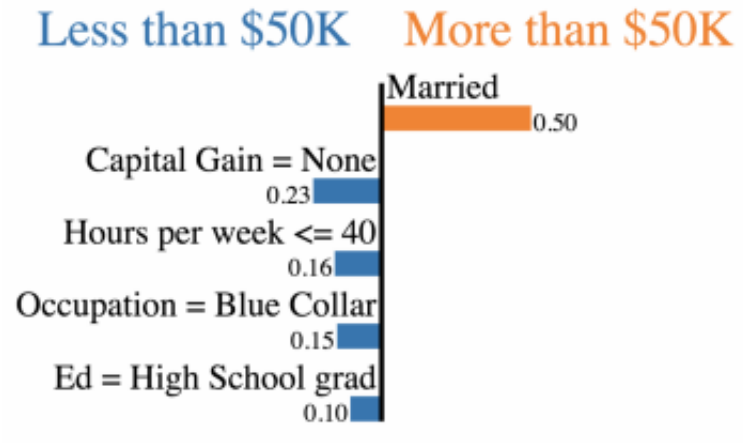

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com
I think Christianity is the one true religion.
If you'd like to know more, send me a note
Prediction probabilities
atheism 0.82
christian 0.18


atheism
christian
Christianity 0.45
Posting 0.36
Host 0.35
Keith 0.23
answer 0.15
is 0.11
Why did this happen?

[R12]

Improving LIME

- What makes a good explanation?
  - Vanilla LIME: explanations don't generalise well
  - Anchor LIME: identify a set of constraints
  - Ultimately: convert constraints into a decision tree. Create a simple *program* as an explanation – *decompile* the model locally

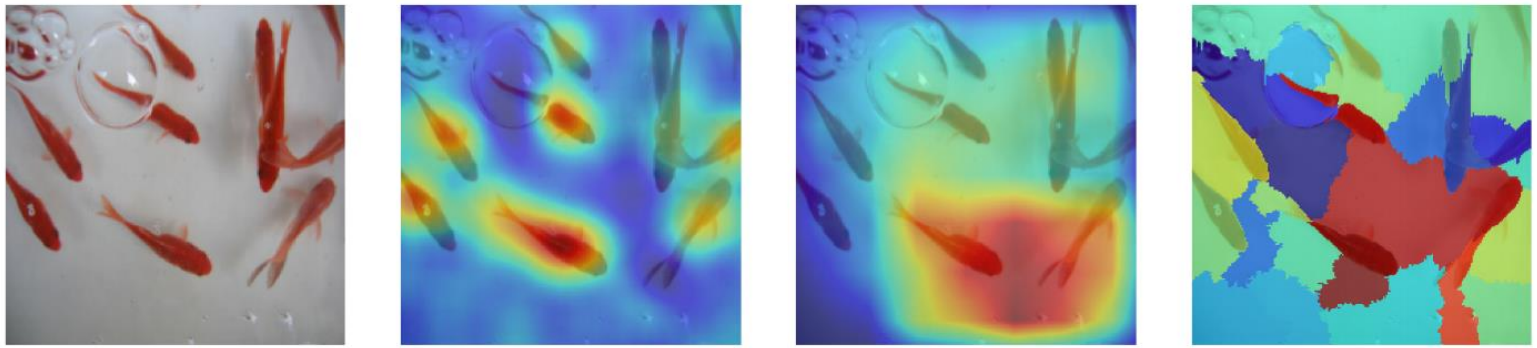


IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 $<$ Age $\leq$ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary $>$ $50K

[R13, R14]

Improving LIME

- Interpretable components
  - Finding sub-regions with similar colours has limitations
  - **RISE: Randomized Input Sampling for Explanation of black-box models** uses random pixel masks instead of components



(a) Input (b) RISE (ours) (c) GradCAM (d) LIME

[R15]

RISE: Randomized Input Sampling for Explanation

- Generates an importance map over all pixels
- Causal metrics distinguish between insertion and deletion
- Better performance than LIME on ImageNet

$$S(\lambda) = \mathbb{E}_M \left[f(I \odot M) \mid M(\lambda) = 1 \right]$$

region λ is preserved

1) Continuous masks instead of binary.
2) Monte Carlo approximation of expectation.

$$S(\lambda) \overset{MC}{\approx} \mathbb{E}_M \left[f\left(\right) \right]$$

always preserved

Table 1: Comparative evaluation in terms of deletion (lower is better) and insertion (higher is better) scores on ImageNet dataset. Except for Grad-CAM, the rest are black-box explanation models.

Method	ResNet50		VGG16	
	Deletion	Insertion	Deletion	Insertion
Grad-CAM [23]	0.1232	0.6766	0.1087	0.6149
Sliding window [31]	0.1421	0.6618	0.1158	0.5917
LIME [21]	0.1217	0.6940	0.1014	0.6167
RISE (ours)	**0.1076 ± 0.0005**	**0.7267 ± 0.0006**	**0.0980 ± 0.0025**	**0.6663 ± 0.0014**

[R15, R16]

Upcoming research

Currently under review for ICML 2019 (https://openreview.net/group?id=ICLR.cc/2019/Conference):

- *Identifying Bias in AI using Simulation* – testing face detection systems for racial bias using (highly realistic) CGI
- *Explaining Image Classifiers by Counterfactual Generation* – saliency detection should be conducted with **plausible alternative values in the context of the surrounding region**
- *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness* and *What a difference a pixel makes: An empirical examination of features used by CNNs for categorisation* both found that CNNs learn non-shape features (such as a single predictive pixel amongst 50176 pixels!)

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Further Reading

LIME is just one of many existing techniques and the research field is rapidly expanding. The following provide good surveys from contrasting perspectives:

- *What do we need to build explainable AI systems for the medical domain?*, Holzinger et al. (2017)
- *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, Adadi & Berrada (2018)
- *DARPA Explainable AI (XAI) Program Update*, Gunning (2017)

Conclusion

It is possible to explain your DNN's output, and you should

It is our responsibility as computer scientists to develop **explainable** deep learning systems to ensure outcomes are **accurate, ethical,** and **fair**.

Conclusion

It is possible to explain your DNN's output, and you should

You can start explaining your DNNs using LIME with a few lines of Python: https://github.com/marcotcr/lime

References I

1. *Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car*, Bojarski et al. (2017). https://arxiv.org/abs/1704.07911
2. *What do we need to build explainable AI systems for the medical domain?,* Holzinger et al. (2017). https://arxiv.org/abs/1712.09923
3. *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*, Caruana et al. (2015). https://dl.acm.org/citation.cfm?id=2788613
4. *How We Analyzed the COMPAS Recidivism Algorithm*, Larson et al. (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Subsequent rebuttal by Northpointe: *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, Dieterich, Mendoza & Brennan (2016). http://www.equivant.com/blog/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity. See also *Fairer and more accurate, but for whom?*, Chouldechova & G'Sell (2017). https://arxiv.org/abs/1707.00046
5. *Deep Learning in Finance Summit London (2018).* Companies presenting talks about their use of deep learning included Lloyds, Liverpool Victoria, Prudential, and Experian. https://www.re-work.co/events/deep-learning-in-finance-summit-london-2018
6. *Anomaly detection for advanced military aircraft using neural networks*, Brotherton & Johnson (2001). https://ieeexplore.ieee.org/abstract/document/931329
7. *Weakly Supervised Learning Based on Coupled Convolutional Neural Networks for Aircraft Detection*, Zhang et al. (2016). https://ieeexplore.ieee.org/abstract/document/7485835
8. *'Fiction is outperforming reality': how YouTube's algorithm distorts truth*, Paul Lewis (2018). https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth
9. *Using Deep Learning at Scale in Twitter's Timelines*, Koumchatzky & Andryeyev (2017). https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

References II

10. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, Official Journal of the European Union (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
11. *Explaining Black-Box Machine Learning Predictions*, Sameer Singh (2017). https://www.youtube.com/watch?v=TBJqgvXYhfo
12. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, Ribeiro, Singh, & Guestrin (2016). https://arxiv.org/abs/1602.04938
13. *Anchors: High-Precision Model-Agnostic Explanations,* Ribeiro, Singh, & Guestrin (2018). https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982
14. *Programs as Black-Box Explanations,* Singh, Ribeiro, & Guestrin (2016). https://arxiv.org/pdf/1611.07579.pdf
15. *RISE: Randomized Input Sampling for Explanation of Black-Box Models*, Pesuik, Das, & Saenko (2018). https://arxiv.org/abs/1806.07421
16. *Title image: RISE BMVC Slides*, Petsiuk, Das, & Saenko (2018). http://cs-people.bu.edu/vpetsiuk/rise/ (see 'Slides')