

Explainable AI



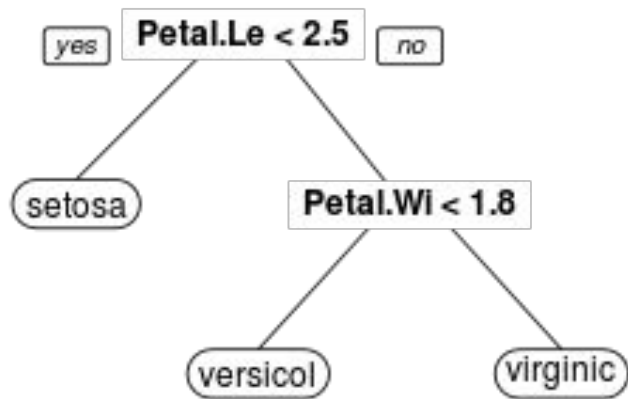
How to explain the decision of a machine learning algorithm?

Christoph Molnar

The problem

- Opaque decision making by machine learning algorithms
- Examples: AI doctor, self-driving cars, creditworthiness
- Trust as a user: Should I start severe therapy, because the machine said so?
- Debugging as a practitioner: Why did the algorithm miss-classify sample X? Did it learn generalizable features?

Keep it simple



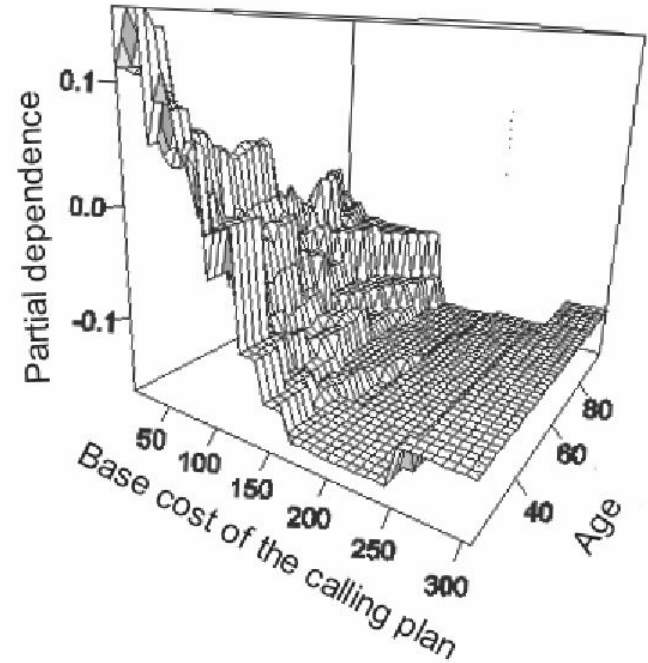
if irregular **then** *malignant* (80%)
else if spiculated **then** *malignant* (95%)
else if age > 57 **and** ill-defined **then** *malignant* (70%)
else if lobular **and** low density **then** *benign* (57%)
else *benign* (88%)

Letham, B., Rudin, C., Tyler, M., McCormick, H., & Madigan, D. (2012). Building Interpretable Classifiers with Rules using Bayesian Analysis.

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots$$

Analyse the system

Photosynthesis
Phages..Prophages..Transposable.elements
Membrane.Transport
Protein.Metabolism
Nitrogen.Metabolism
Dormancy.and.Sporulation
Cofactors..Vitamins..Prosthetic.Groups..Pigments
Stress.Response
Cell.Wall.and.Capsule
DNA.Metabolism
Carbohydrates
Metabolism.of.Aromatic.Compounds
Phosphorus.Metabolism
Motility.and.Chemotaxis
Virulence
Sulfur.Metabolism
Amino.Acids.and.Derivatives
Nucleosides.and.Nucleotides
Potassium.metabolism
Secondary.Metabolism
Regulation.and.Cell.signaling
RNA.Metabolism
Respiration
Cell.Division.and.Cell.Cycle
Fatty.Acids..Lipids..and.Isoprenoids
Plasmids



Explain single decisions

Local Interpretable Model-agnostic Explanations (LIME) [1]



(a) Original Image

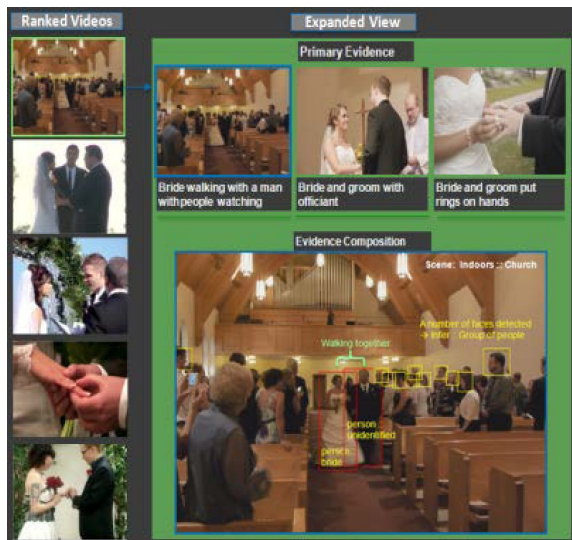
(b) Explaining *Electric guitar*

(c) Explaining *Acoustic guitar*

(d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Let algorithm explain



[2]



[3]

*This is a **White Pelican** because...*



Description: this bird is white and black in color with a long curved beak and white eye rings.

Explanation: this is a large white bird with a **long neck** and a **large orange beak**.

[4]

References

- [1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. Retrieved from <http://arxiv.org/abs/1602.04938>
- [2] Cheng, H., Liu, J., Chakraborty, I., Chen, G., Liu, Q., Elhoseiny, M., ... Curtis, J. (n.d.). Multimedia Event Detection and Recounting.
- [3] <https://www.youtube.com/watch?v=vXcuLEBwXsQ>
- [4] Hendricks, L. A., & Donahue, J. (n.d.). Generating Visual Explanations.
- [5] DARPA-BAA-16-53. (2016). Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency, 1. Retrieved from <http://www.darpa.mil/program/explainable-artificial-intelligence>