

# **MachineLearnAthon - Interpretability in Machine Learning- Introduction**

Radwa El Shawi

**MachineLearnAthon**  
**A project Co-funded by the Erasmus+ programme of the European Union**

10.10.2024

# Learning outcomes of today

After completing this micro-lecture, you are able to....

- Understand the concept of interpretability in machine learning and its significance.
- Differentiate between model interpretability and model accuracy.
- Identify key approaches to explaining machine learning model

# Agenda for today

- Introduction to Interpretability
- Model Accuracy vs Interpretability
- Key Approaches to Explainability

# What is interpretability

Interpretability is the degree to which an observer can understand the cause of a decision

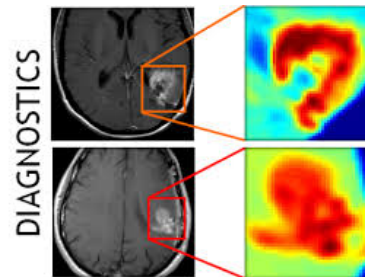
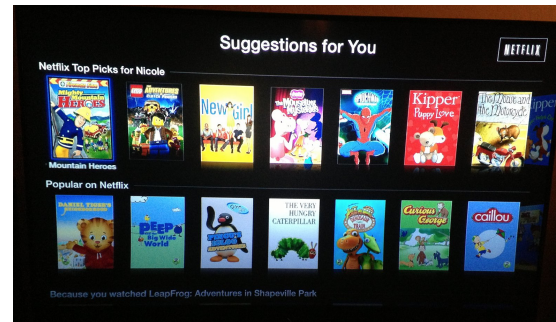


Miller, T. (2019). *Explanation in artificial intelligence: Insights from the social sciences*. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

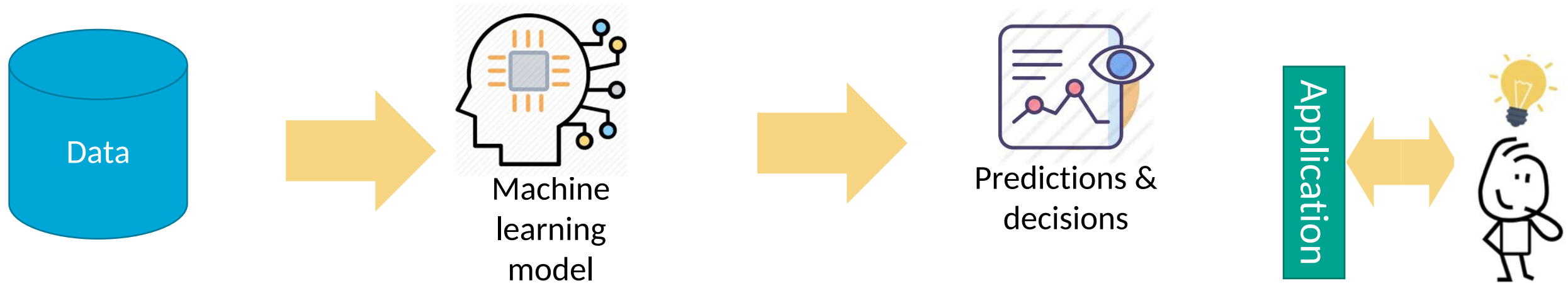
# Machine Learning applications becoming pervasive...



<https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long>



# Why Explainability

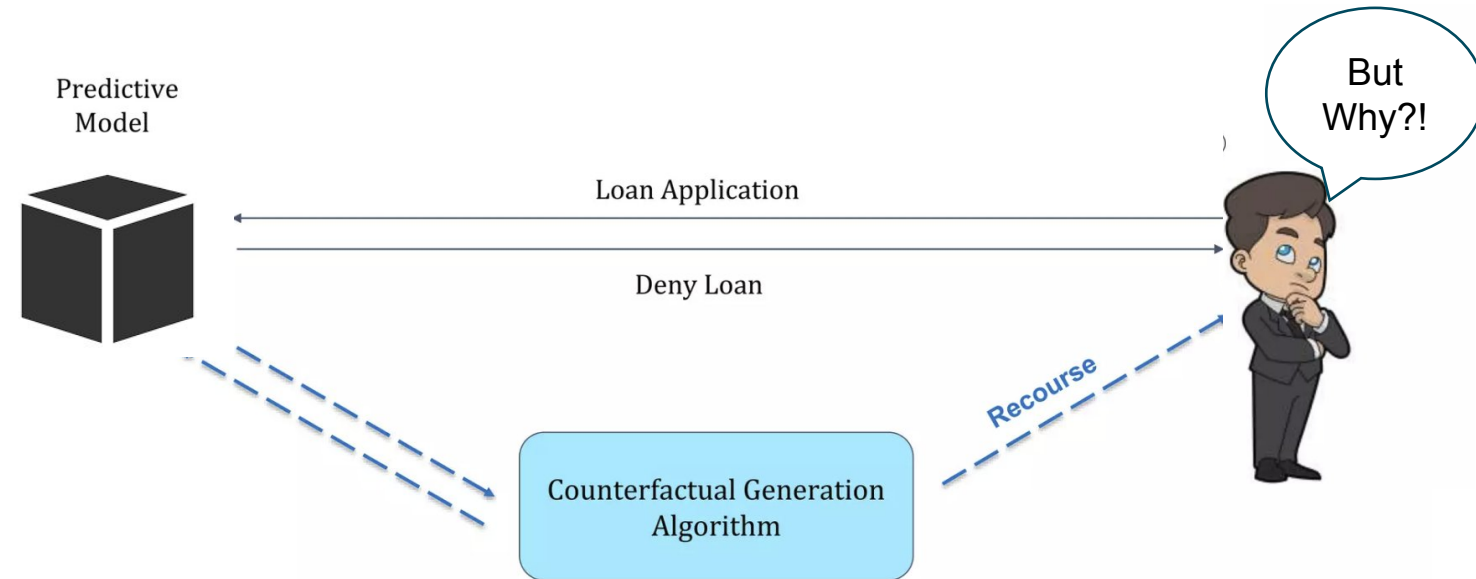


## Trust Challenge

- Why is the model making these predictions?
- Is the model making correct predictions for the right reasons?

# Why is the model making these predictions?

- As ML models are increasingly deployed to make high-stakes decisions (e.g., loan applications), it becomes important to provide recourse to affected individuals.



**Recourse:** Increase your salary by 5K & pay your credit card bills on time for next 3 months

## Counterfactual Explanations

What features need to be changed and by how much to flip a model's prediction?

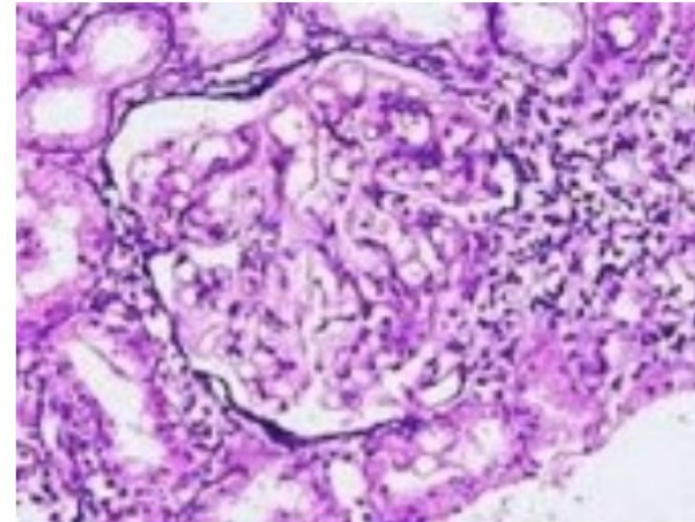
# Why Explainability: Verify the ML Model / System

Wrong decisions can be costly and dangerous

*“Autonomous car crashes, because it wrongly recognizes ...”*




*“AI medical diagnosis system misclassifies patient’s disease ...”*




Samek, W., & Binder, A. (2018). *Interpretable machine learning* [Image]. MICCAI'18 Tutorial on Interpretable Machine Learning, Fraunhofer HHI & Singapore University of Technology and Design (SUTD). Available at [http://heatmapping.org/slides/2018\\_MICCAI.pdf](http://heatmapping.org/slides/2018_MICCAI.pdf)

# Can I trust the Model based on Accuracy?

| Patient ID |     | Has Asthma |     | Risk of Death |
|------------|-----|------------|-----|---------------|
| 84         | ... | Yes        | ... | 5%            |
| 85         | ... | Yes        | ... | 6%            |
| 86         | ... | No         | ... | 12%           |
| 87         | ... | No         | ... | 15%           |
| ...        | ... | ...        | ... | ...           |

Feature Importance (Higher risk of death): Low  High

Feature Importance (Lower risk of death): Low  High

**With Context:** Patients with asthma have a lower risk of death from pneumonia because they receive more intensive care.

## Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana  
Microsoft Research  
rcaruana@microsoft.com

Yin Lou  
LinkedIn Corporation  
ylou@linkedin.com

Johannes Gehrke  
Microsoft  
johannes@microsoft.com

Paul Koch  
Microsoft Research  
paulkoch@microsoft.com

Marc Sturm  
NewYork-Presbyterian Hospital  
mas9161@nyp.org

Noémie Elhadad  
Columbia University  
noemie.elhadad@columbia.edu

### ABSTRACT

In machine learning often a tradeoff must be made between

the application of machine learning to important problems in healthcare such as predicting pneumonia risk. In the study, the goal was to predict the probability of death (POD) for

[1] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). <https://doi.org/10.1145/2783258.2788613>

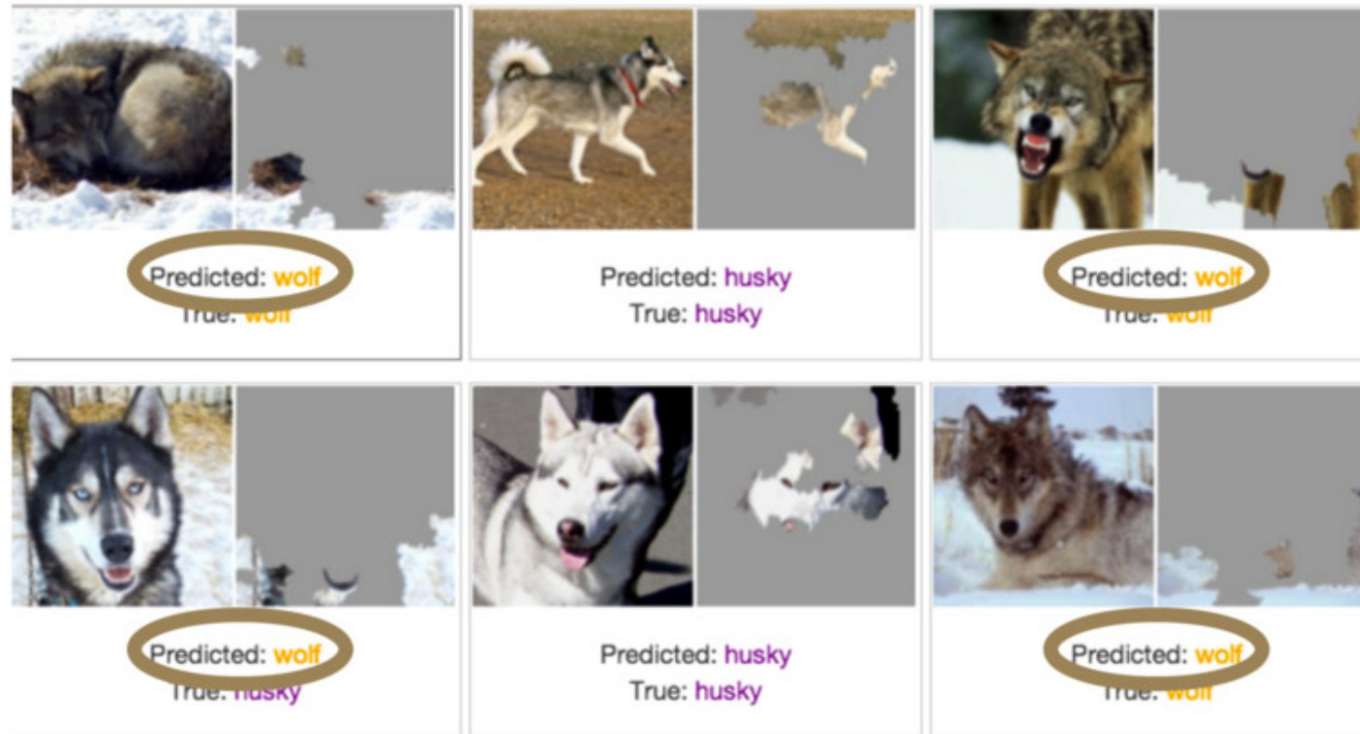
# Can I trust the Model based on Accuracy?

Only 1 mistake!



Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, November 18). "Why should I trust you?": Explaining the predictions of any classifier [Slide 3]. FATML 2016. Speaker Deck.  
<https://speakerdeck.com/fatml/why-should-i-trust-you-explaining-the-predictions-of-any-classifier?slide=3>

# Can I trust the Model based on Accuracy?



We've built a great snow detector... ☹️

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, November 18). "Why should I trust you?": Explaining the predictions of any classifier [Slide 21]. FATML 2016. Speaker Deck.  
<https://speakerdeck.com/fatml/why-should-i-trust-you-explaining-the-predictions-of-any-classifier?slide=3>

# Some properties of Interpretations

- **Faithfulness** - providing explanations accurately representing the true reasoning behind the model's final decision.
- **Understandable** – Can I put it in terms that end users without in-depth knowledge of the system can understand?
- **Stability** – Do similar instances have similar interpretations?

# Evaluating Interpretability

- **Human evaluation** – Set up a Mechanical Turk task and ask non- experts to judge the explanations
- **Functional evaluation** – Design metrics that directly test properties of your explanation.

# Legal Issues-GDPR



# Approaches to Explaining AI

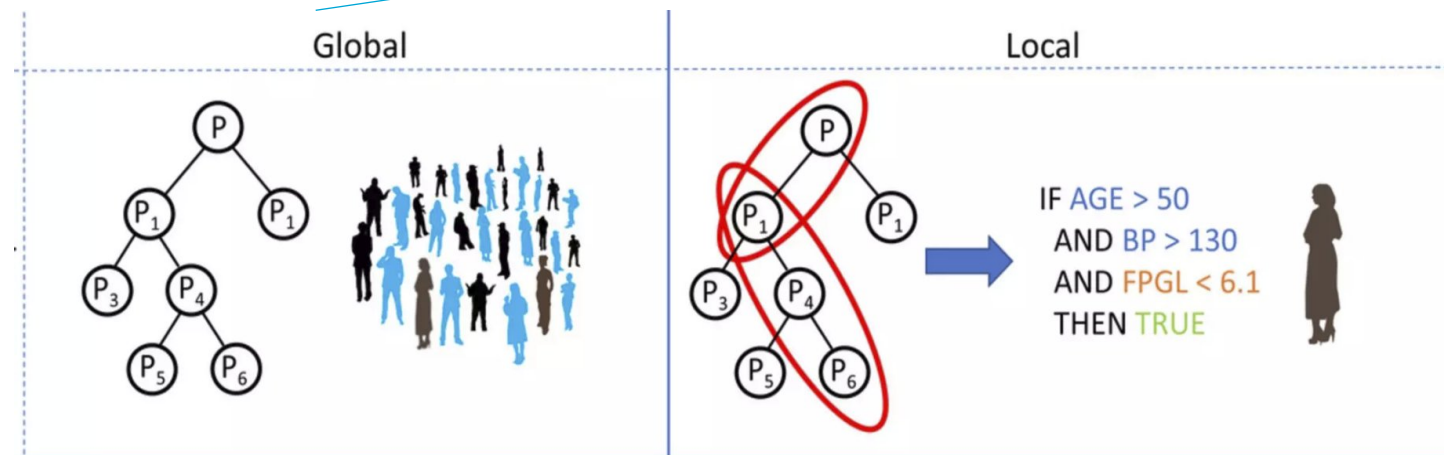
**Transparent**  
(how the model functions internally)

**Explainable AI**

**Black box**  
(how the model behaves)

**Global**

**Local**



*The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.*

This material is licenced under CC BY-NC-ND 4.0  
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>).