eXplainable AI (XAI)

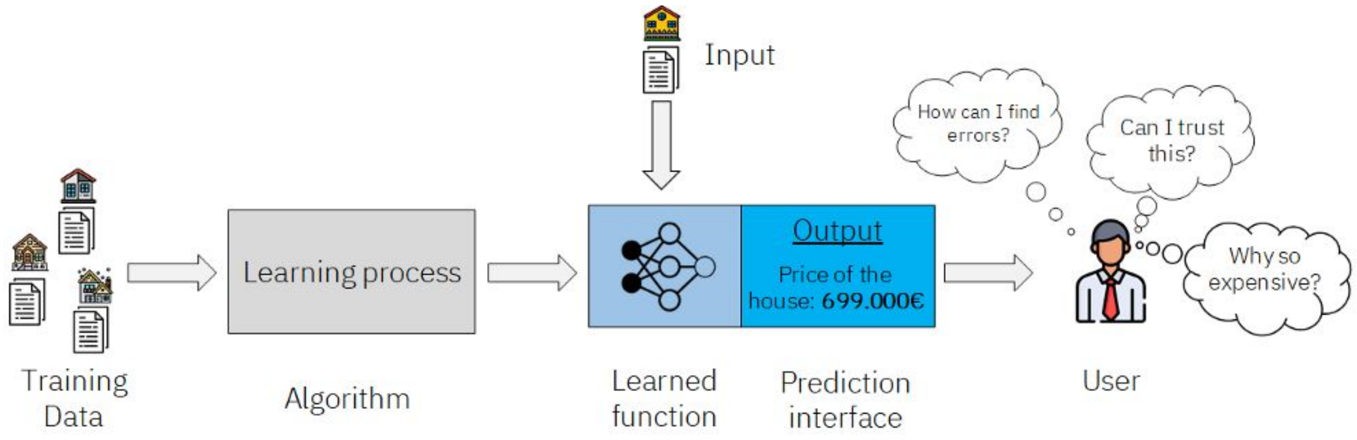
Frameworks toward interpretable systems

N. Rich Nguyen, PhD

SYS 6016

xkcd.com/1838

Black-box AI



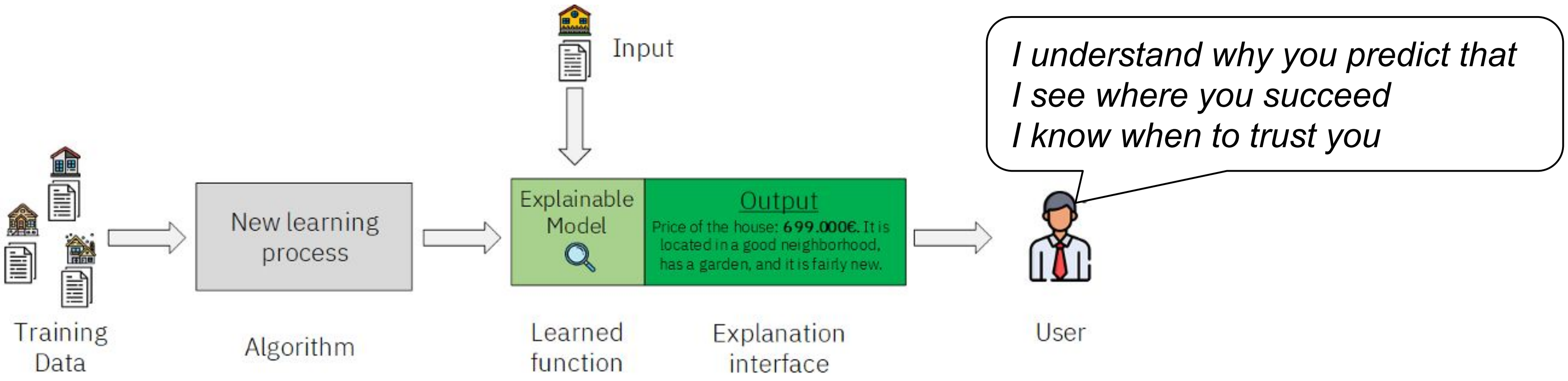
- End-users vs. AI engineers (ie. “Can I trust this AI?” vs. “How to improve the performance?”)
- Completeness vs. explainability (ie. “this image classifier uses a deep network” vs. “presenting the Inception paper as an explanation”)
- Transparency vs. bias and discrimination (exist in both training data and human society)

Why explainability? It’s the notion of trust

(a) Husky classified as wolf (b) Explanation

- Improve ML system (through human interactivity)
- Verify the model because wrong decision can be costly and dangerous (ie. self-drive car or medical diagnostic)
- Learn new intuition (ie. why AlphaZero makes such as move in Go?)
- Gain insight in science? (ie. which gene linked to cancer?)
- Address inherent bias in ML models (ie. face detection, loan approval)

Explainable AI (XAI)



Goal: Design trustworthy, reliable, and explainable AI models without sacrificing performance.

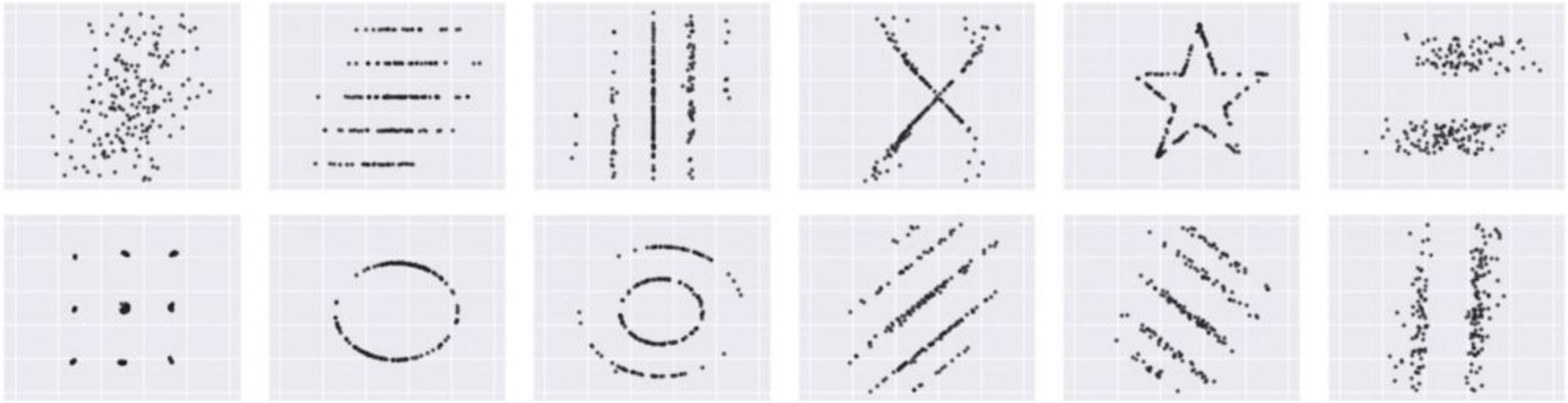
There are 3 ways to design explainability into the **training process** of a model:

1. **Pre-training** explainability: understand and describe the data
2. **In-training** explainability: develop more inherently explainable models
3. **Post-training** explainability: extract explanation from trained models

1. Pre-training explainability

Exploratory data analysis and visualization

Merely relying on statistical properties which is often not good enough



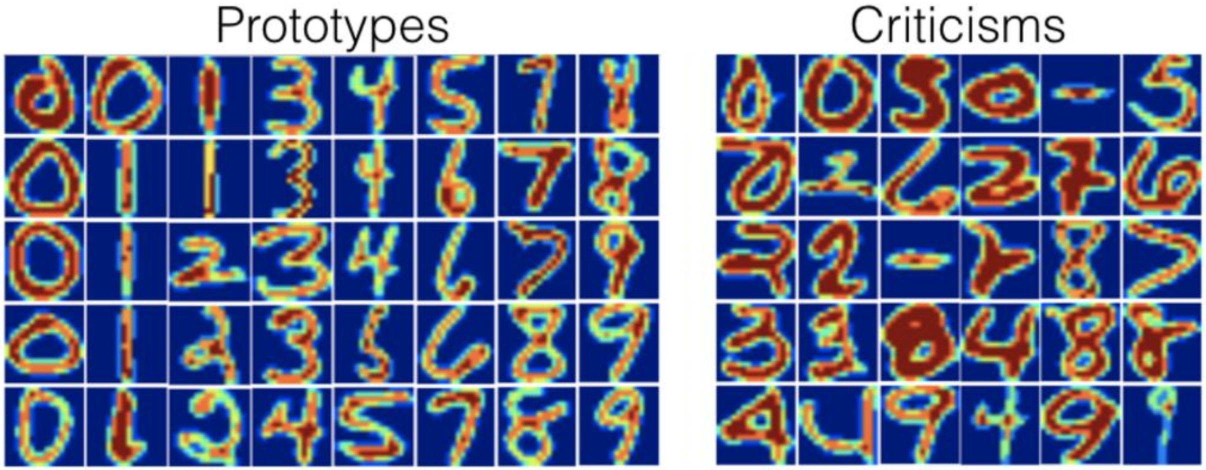
An example of datasets with **identical** mean and standard deviation, and different graphs to demonstrate the importance of visualization in exploratory data analysis

Visualizing high dimensional data with parallel-coordinate plots, PCA, t-SNE

Dataset Summarization

To summarize a dataset often means to seek a minimal subset of representative samples (aka **prototypes**) that provide a condensed view of it.

Prototypes are usually not adequate, we need **criticisms** too. A criticism is an often (relatively) rare data point that is not well described by the set of prototypes.



2. In-training explainability

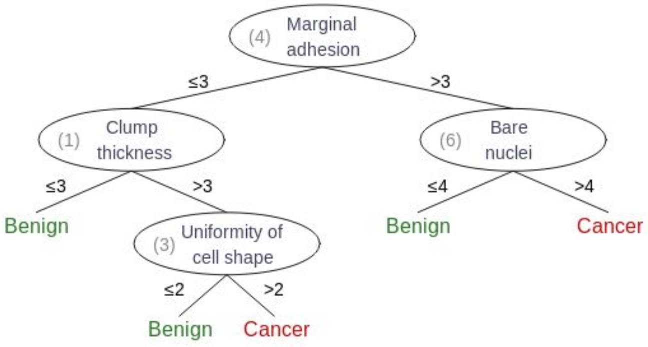
Inherently explainable models

Adopt model from a specific family:

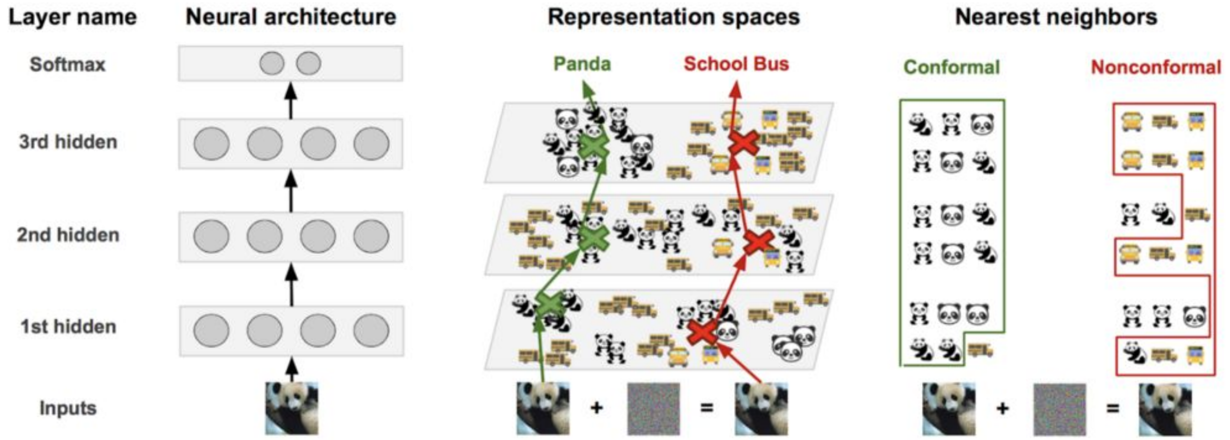
- Generalized Linear Models
- Decision Trees
- Tree-based Ensemble Models
- Generalized Additive Models

It could mean **lower** predictive performance in large and complex dataset

Explainability should not (does not) mean lower performance.



Hybrid Architectures



- The deep-kNN approach proposes to use K-nearest neighbor (kNN) **inference** on the hidden representation of training dataset
- **Interpretability** of each layer outcome is provided by the **nearest neighbors**
- **Robustness** stems from detecting non-conformal predictions from nearest neighbor labels found for out-of-distribution inputs (e.g. an **adversarial** panda)
- Requires storing hidden representation of the entire training dataset

Papernot, Nicolas, and Patrick McDaniel. "Deep k-nearest neighbors." *arXiv preprint arXiv:1803.04765* (2018).

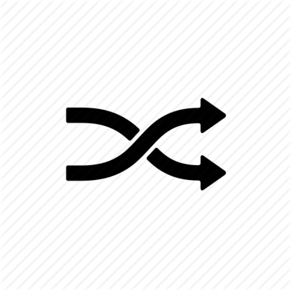
3. Post-training Explainability

Permutation Feature Importance

Model-agnostic explanation: treat model as *a black box* and does not inspect internal model parameters

Given a trained model with *n* features and a validation set.

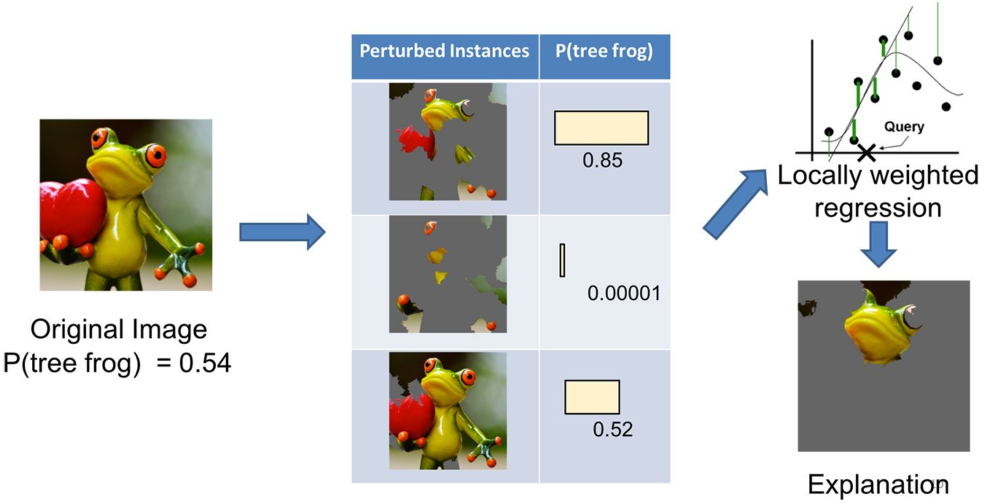
- Compute the **baseline** prediction score on the validation set
- For each feature:
    - Modify the validation set by **shuffling** the values of that feature
    - Compute the prediction score of the model on the **modified** set
    - Compute the **decrease** in model accuracy to the baseline
- The **rank** across the features according to the **reduction** of their score

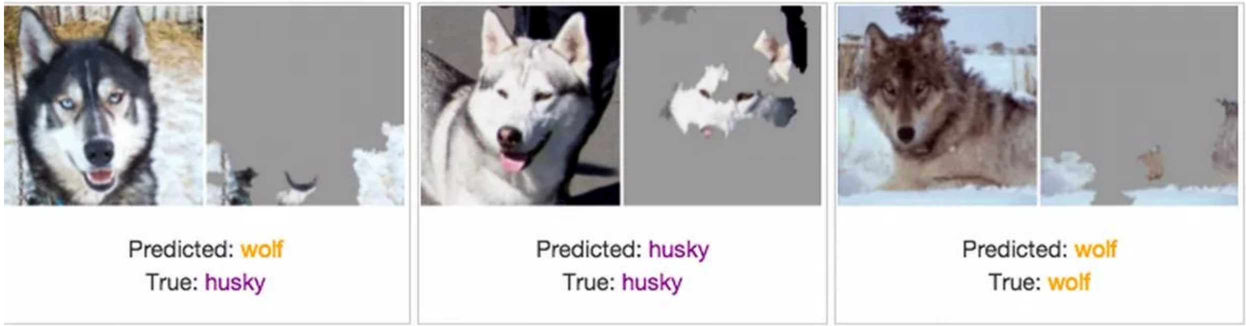


LIME: Local Interpretable Model-Agnostic Explanations

LIME explains the predictions of any classifier by learning an interpretable model **locally** around the prediction.

Work by **perturbing** the original image and **learning** a linear **regression** on the received dataset, so we can understand how it behaves in a small local environment.



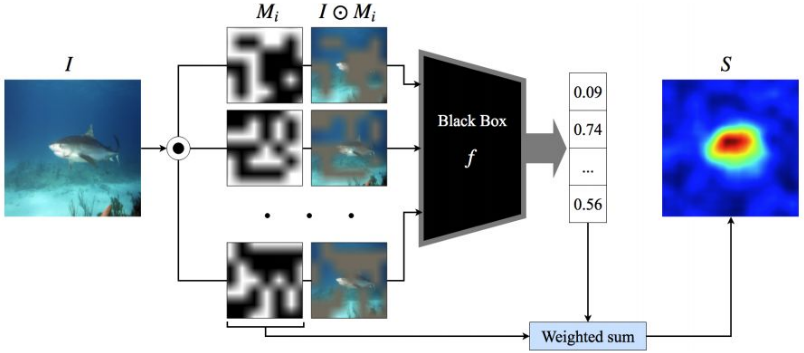


Ribeiro, Marco et al "Why should I trust you?" Explaining the predictions of any classifier." *ACM SIGKDD* 2016.

RISE: Randomized Input Sampling for Explanation

In RISE, input image is element-wise multiplied with random masks, then the masked images are fed to the base model.

The heat map is a linear combination of the masks where the weights come from the score of the target class responding to the respective masked inputs.

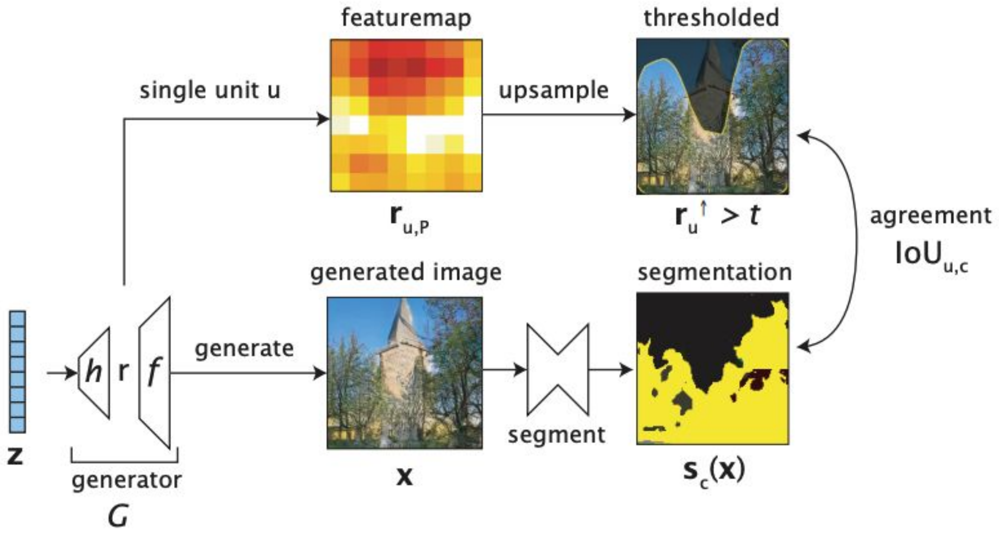


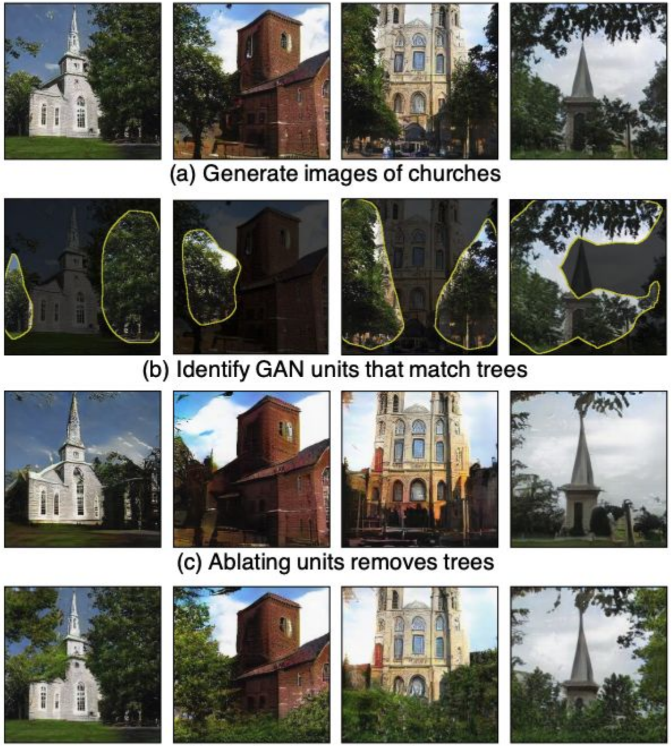
(a) "A horse and carriage on a city street." (b) "A horse..." (c) "A horse and carriage..." (d) "White..."

Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *arXiv preprint arXiv:1806.07421* (2018).

GAN Dissection

Given a pre-trained GAN model, first identify a set of interpretable units, whose featuremap is highly correlated to the region of an object class across different images.





(a) Generate images of churches

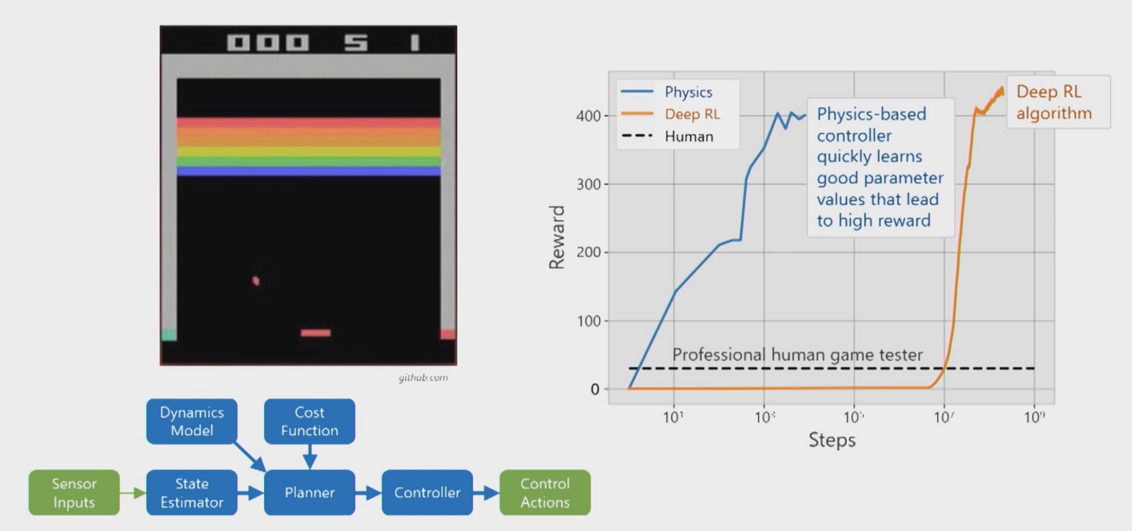
(b) Identify GAN units that match trees

(c) Ablating units removes trees

(d) Activating units adds trees

Bau, David, et al. "Gan dissection: Visualizing and understanding generative adversarial networks." *arXiv preprint arXiv:1811.10597* (2018).

Differentiable Physics Engine for Deep RL

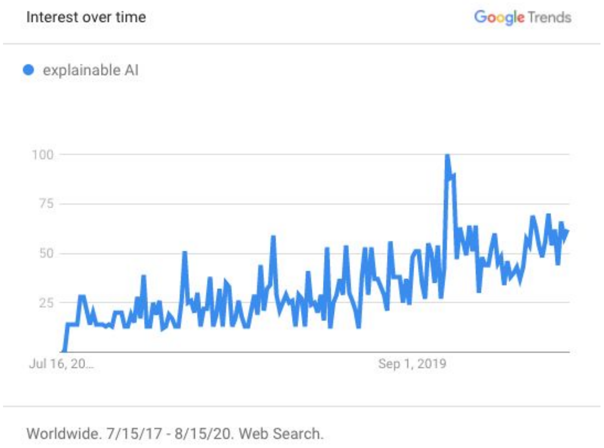


- Differentiable physics engine that can be integrated as a module in deep neural networks for end-to-end learning
- Perform backpropagation analytically through a physical simulator defined via a linear problem.

de Avila Belbute-Peres, Filipe, et al. "End-to-end differentiable physics for learning and control." *NIPS* 2018.

Summary

- Users prefer explanation, explanations generate **trust**!
- Explainable AI (XAI) is of importance for trustworthy and reliable AI systems
- **Pre-training** explainability aims to understand and describe the data
- **In-training** explainability aims to develop more inherently explainable models
- **Post-training** explainability aims to extract explanations from trained models
- Promising approaches to develop XAI **best practices**



Acknowledgements

Slides contain figures from Bahador Khaleghi (Facebook), David Aha (Naval Research Lab), and various researchers reproduced only for educational purposes.