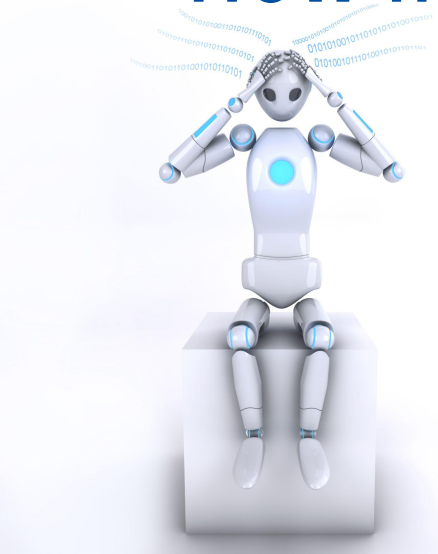
INFO901
How machines explain - concepts in XAI

inga.strumke@ntnu.no



norwegian open ai lab

NTNU

What is it

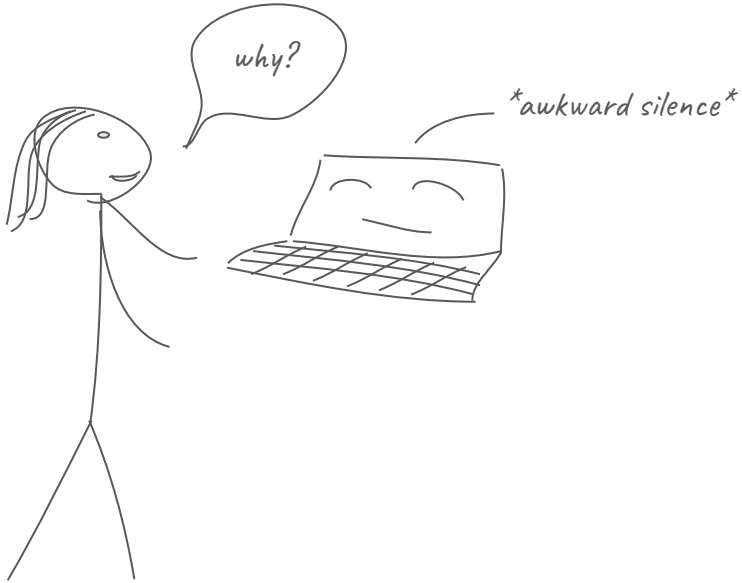
Explainability: an *active* characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions.

≠ **Interpretability:** a *passive* characteristic of a model referring to the level at which it makes sense to humans.

Understandability is imo the most essential concept in XAI.

norwegian open ai lab

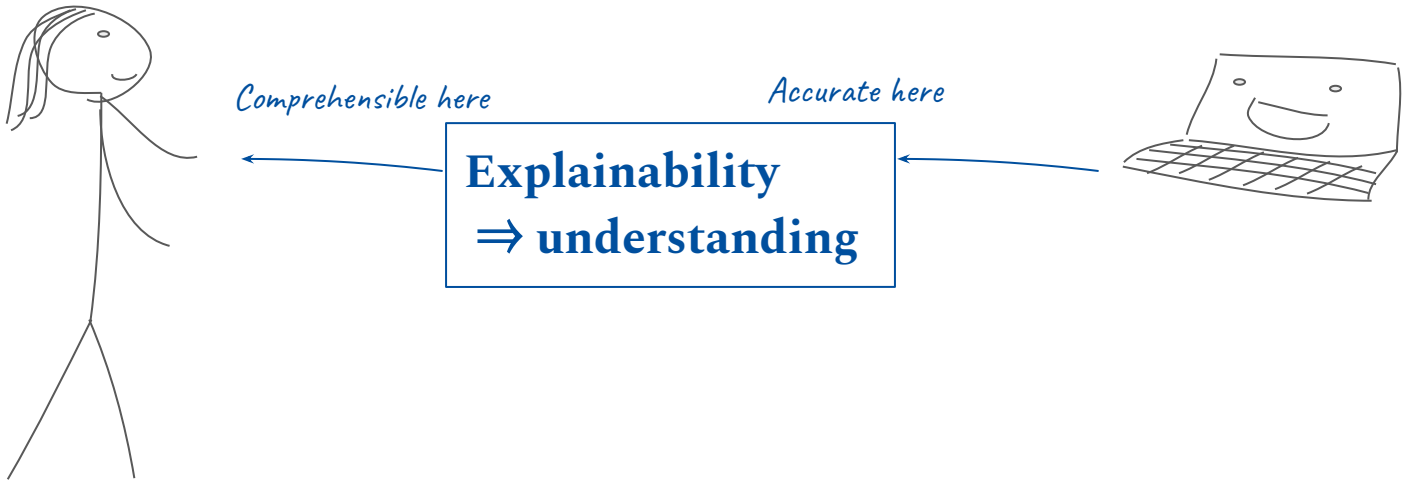
NTNU



No explainability
⇒ no understanding

norwegian open ai lab

NTNU



an *interface* between humans and the machine;

- an accurate proxy of the decision maker
- comprehensible to humans.

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

What seems to be the problem?

1) I would like...

a linear version of a highly non-linear phenomenon.
with a cherry on top, pls 😇

Machine learning models are the result of **predictive modelling** - as opposed to **explanatory modelling**, where the aim is to explain phenomena.

norwegian open ai lab

NTNU

What seems to be the problem?

2) Testing is insufficient due to curse of dimensionality

Robust Physical-World Attacks on Deep Learning Visual Classification, Kevin Eykholt et al

What seems to be the problem?

3) "Explanation" is not well-defined.

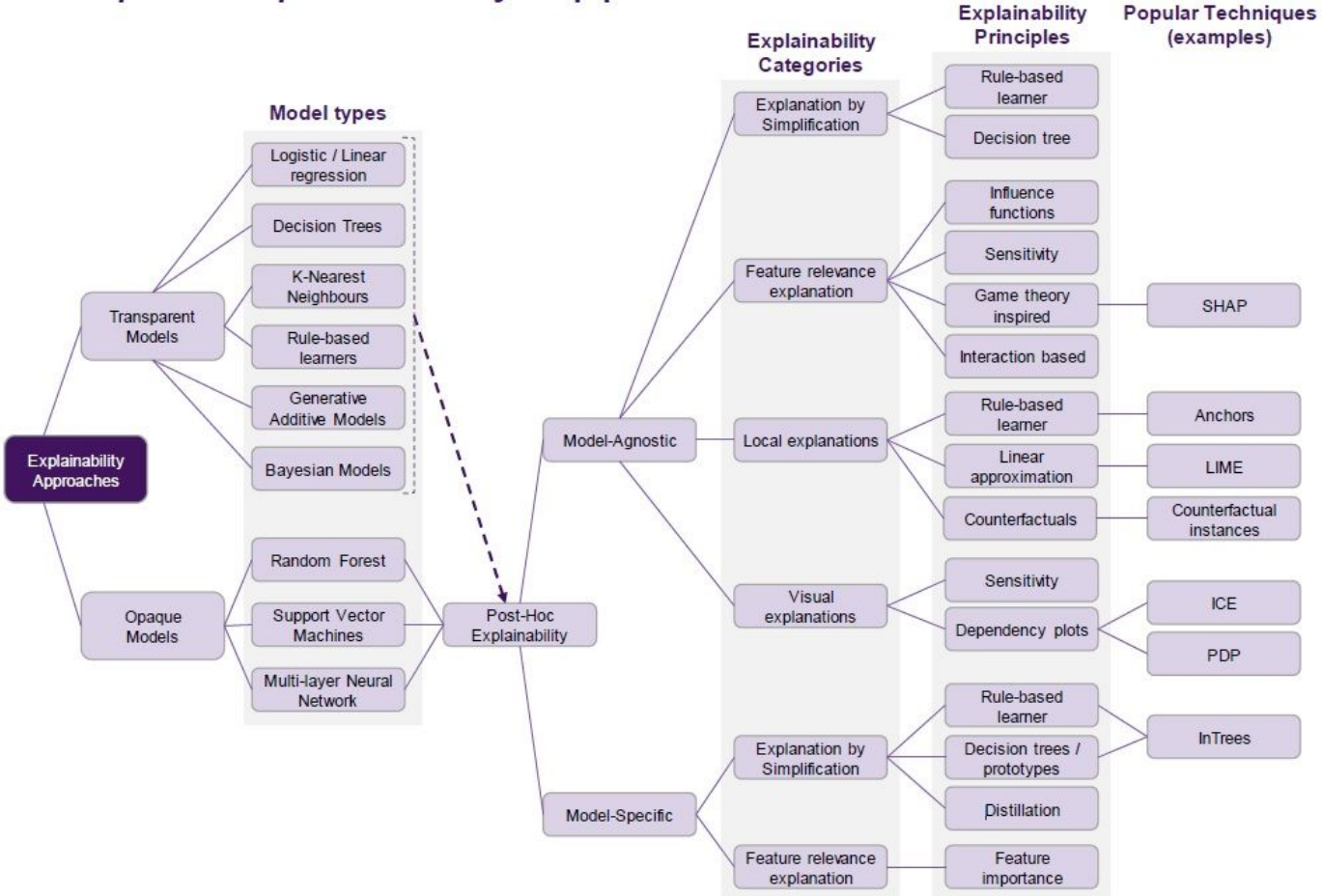
→ Different people and roles require different explanations

→ The legal status of requirements for explanations is not clear

→ The field is dynamic, full of methods accomplishing different things, and ~impossible to oversee

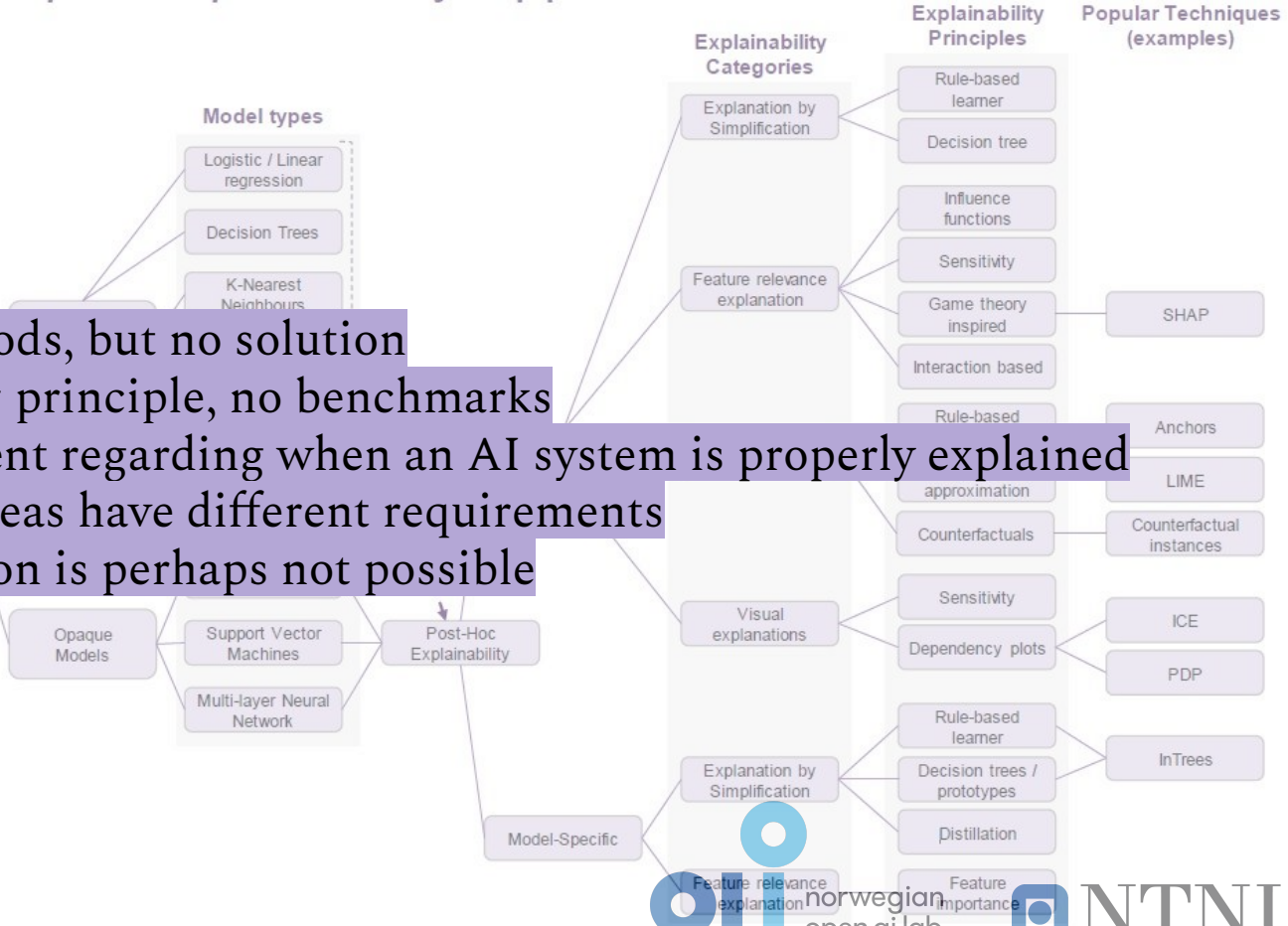
Inga Strümke - inga.strumke@ntnu.no

Map of Explainability Approaches



Inga Strümke - inga.strumke@ntnu.no

Map of Explainability Approaches



Many methods, but no solution
No unifying principle, no benchmarks
No agreement regarding when an AI system is properly explained
Different areas have different requirements
⇒ unification is perhaps not possible

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

XAI for people in a hurry

Information →

→ Predictions

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

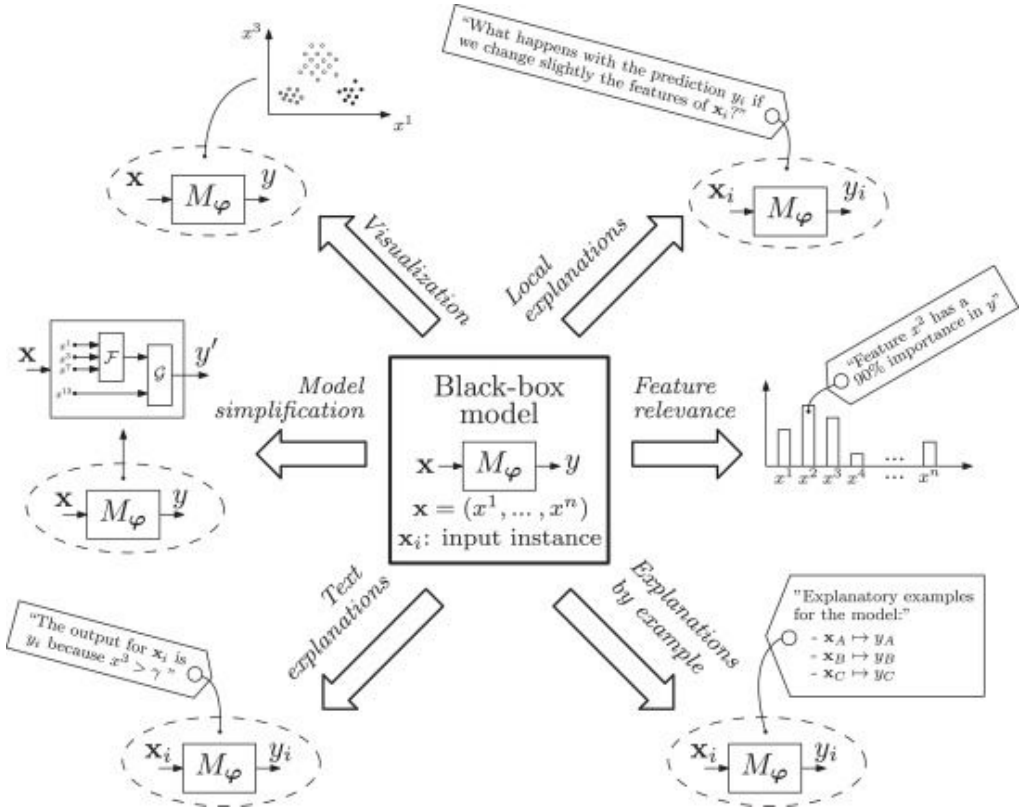
XAI for people in a hurry

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Option 1: What kind of explanation am I after



Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Option 2: What will I do with the model to get the explanation

Three options:

Poke the box systematically and see what happens
Extrinsic explanations

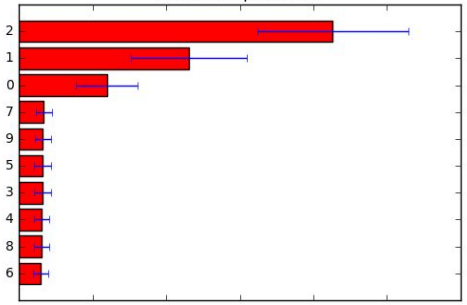
Approximate the box (locally) by an interpretable model
Surrogate model explanations

Smash the box and look inside
Intrinsic explanations

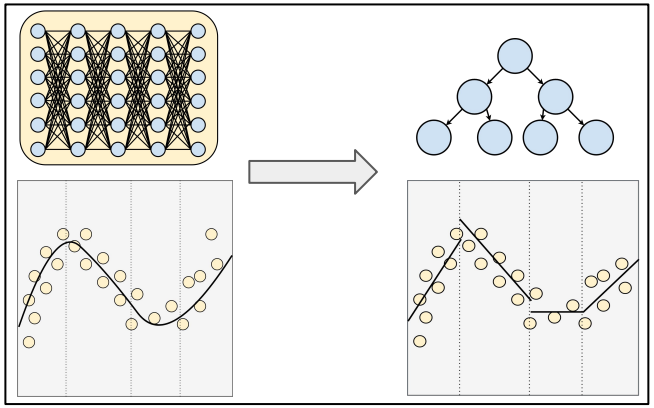
norwegian open ai lab

NTNU

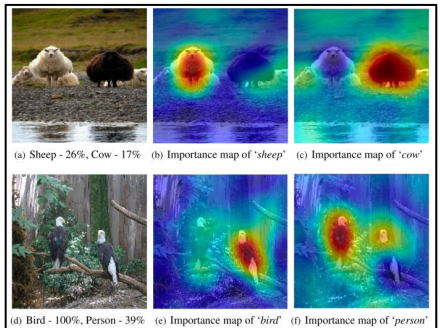
Poke the box systematically and see what happens
Extrinsic explanations



Approximate the box (locally) by an interpretable model
Surrogate model explanations



Smash the box and look inside
Intrinsic explanations


(a) Sheep - 26%, Cow - 17%
(b) Importance map of '*sheep*'
(c) Importance map of '*cow*'
(d) Bird - 100%, Person - 39%
(e) Importance map of '*bird*'
(f) Importance map of '*person*'

Inga Strümke - inga.strumke@ntnu.no

Poke the box like a boss

SHAP: (most?) widely used extrinsic explanation.
⇒ Feature attribution

Based on a solution concept from game theory:

The Shapley decomposition.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right)$$

Intuitive example: Sharing a cab bill

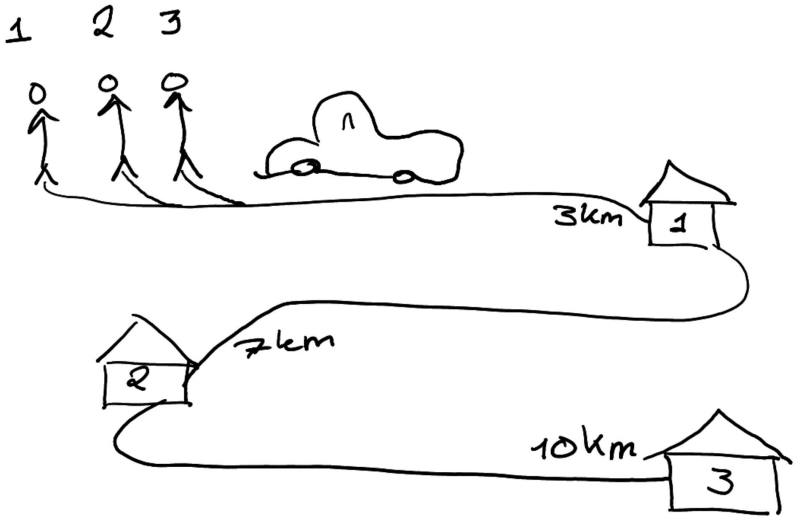
We need the characteristic function values and then we can calculate
the Shapley decomposition

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Shapley cab sharing



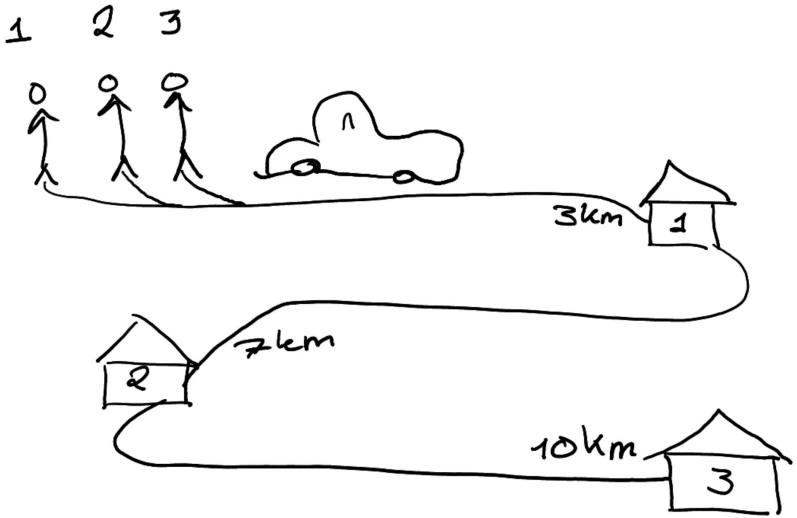
$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

The ***characteristic function*** values fully describe the problem

norwegian open ai lab

NTNU

Shapley cab sharing



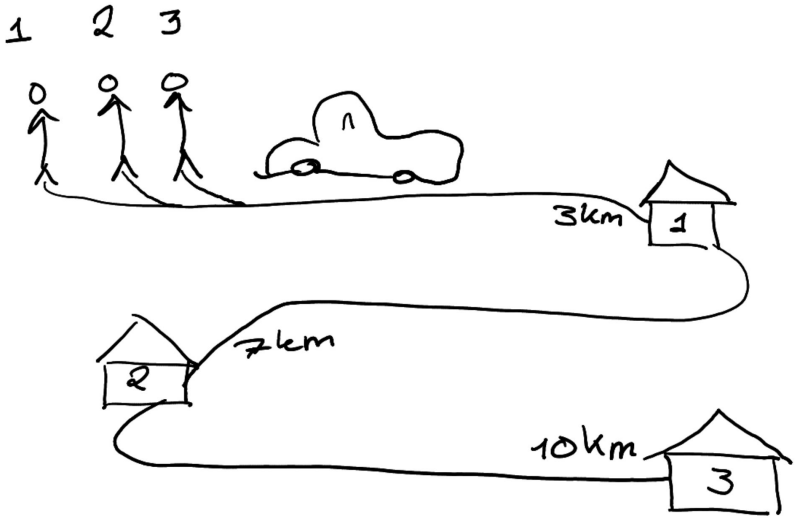
$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

norwegian open ai lab

NTNU

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

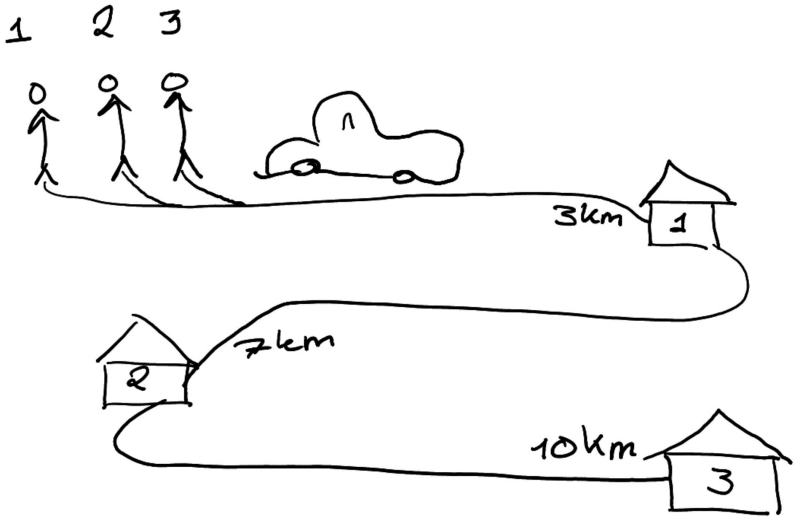
N = total passengers

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

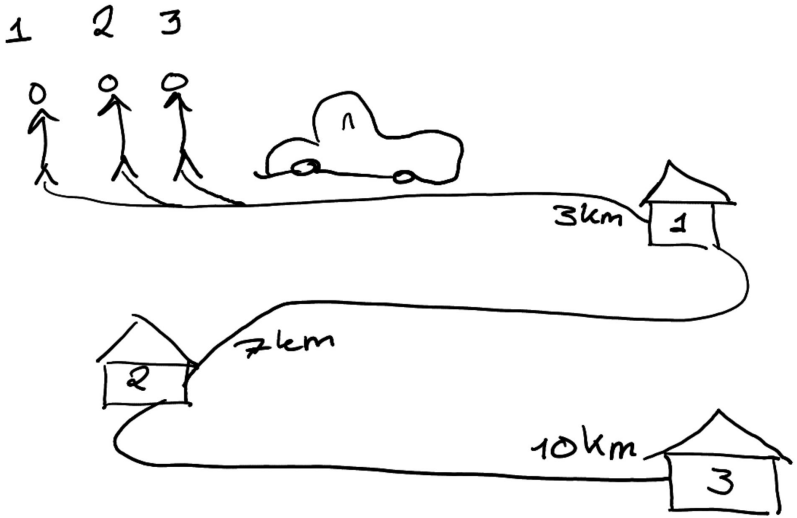
N = total passengers
S = subsets of N

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

subsets S of N
excluding passenger i

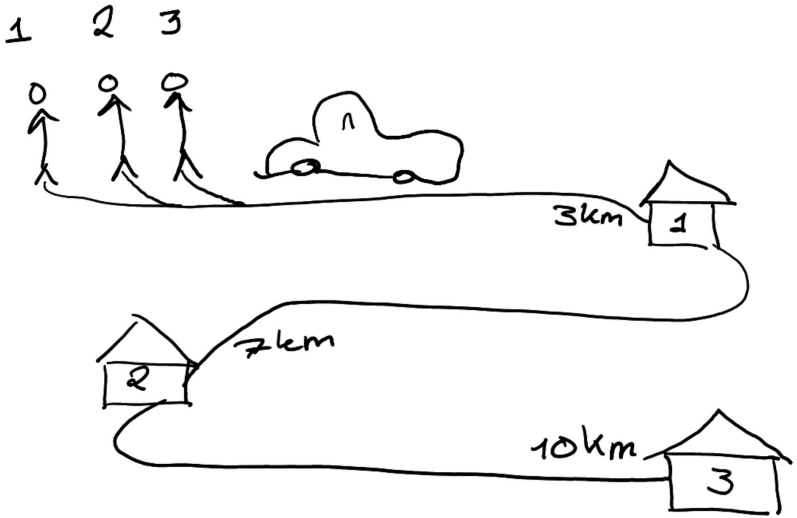
N = total passengers
S = subsets of N

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

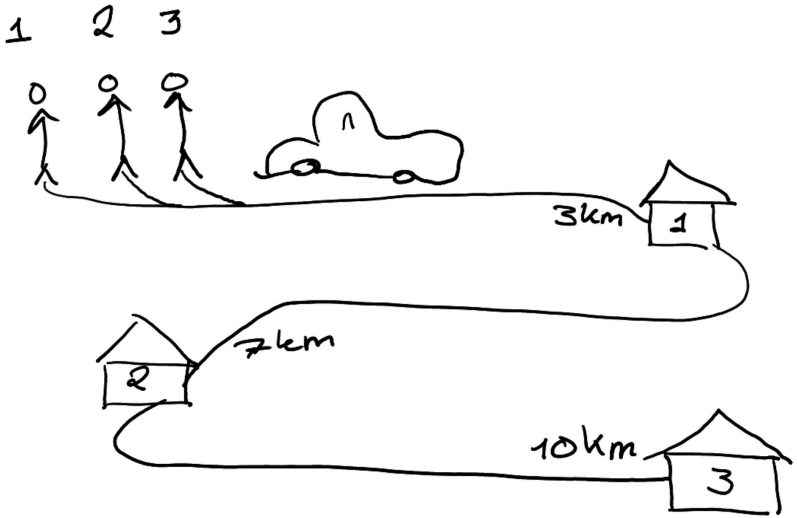
$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

Sets excluding passenger 1: $\{\}, \{2\}, \{3\}, \{2,3\}$

norwegian open ai lab

NTNU

Shapley cab sharing



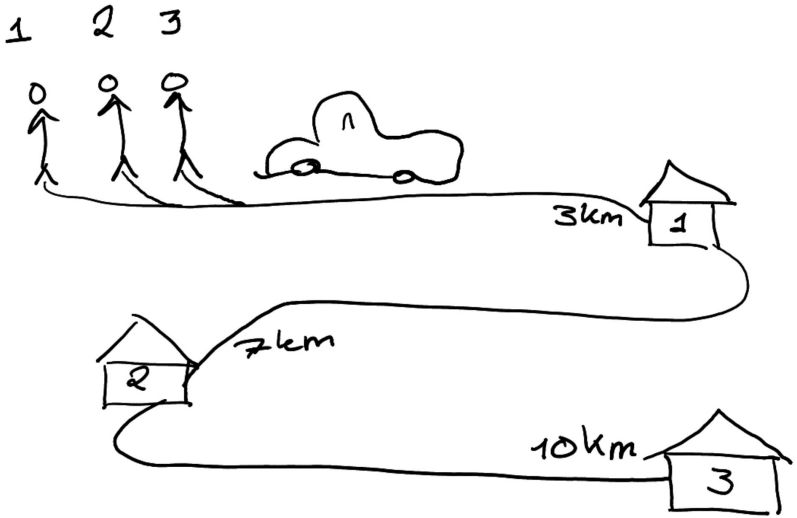
$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

Sets excluding passenger 1: $\{\}, \{2\}, \{3\}, \{2,3\}$

$$\phi_1 = \frac{1(3-0-1)!}{3!}\left(v(\{1\}) - v(\{\})\right) + \frac{1(3-1-1)!}{3!}\left(v(\{1,2\}) - v(\{2\})\right) + \frac{1(3-1-1)!}{3!}\left(v(\{1,3\}) - v(\{3\})\right) + \frac{1(3-2-1)!}{3!}\left(v(\{1,2,3\} - v(\{2,3\})\right) = 1$$

Shapley cab sharing

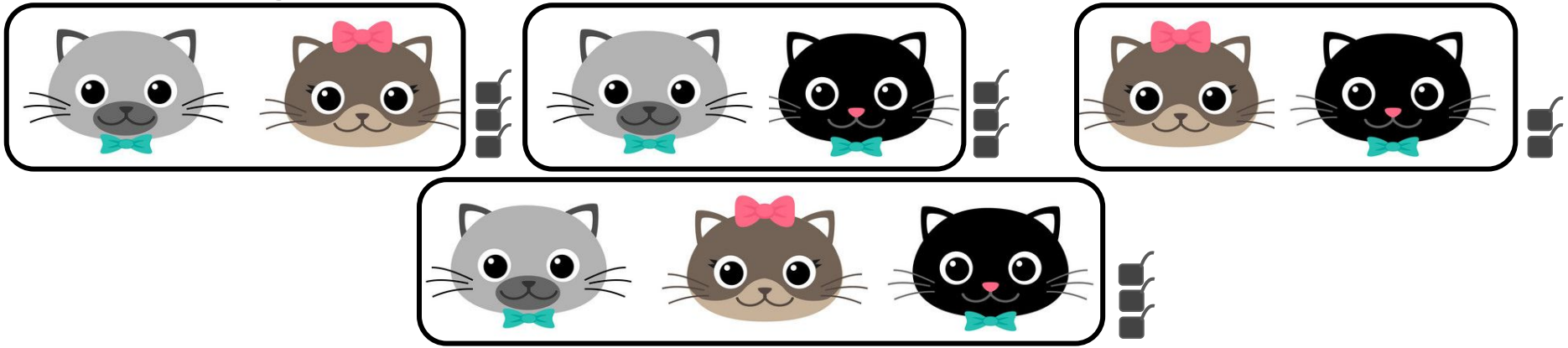


True story: Den unique and **fair** way to distribute the total cost of the journey among the passengers is when
passenger 1 pays for 1 km
passenger 2 pays for 3 km og
passenger 3 pays for 6 km

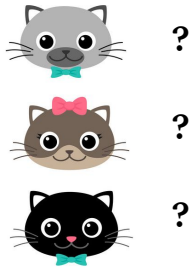
$$\phi_1 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{2,3\})\right) + \frac{1}{6}\left(v(\{1,2\}) - v(\{2\})\right) + \frac{1}{6}\left(v(\{1,3\}) - v(\{3\})\right) + \frac{1}{3}\left(v(\{1\}) - v(\emptyset)\right) = 1$$
$$\phi_2 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{1,3\})\right) + \frac{1}{6}\left(v(\{1,2\}) - v(\{1\})\right) + \frac{1}{6}\left(v(\{2,3\}) - v(\{3\})\right) + \frac{1}{3}\left(v(\{2\}) - v(\emptyset)\right) = 3$$
$$\phi_3 = \frac{1}{3}\left(v(\{1,2,3\} - v(\{1,2\})\right) + \frac{1}{6}\left(v(\{1,3\}) - v(\{1\})\right) + \frac{1}{6}\left(v(\{2,3\}) - v(\{2\})\right) + \frac{1}{3}\left(v(\{3\}) - v(\emptyset)\right) = 6$$

Mice caught alone

Mice caught in coalitions



Shapley value per cat:



$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N-|S|-1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

Inga Strümke - inga.strumke@ntnu.no

Shapley values for machine learning

Shapley values do dependence attribution. From game theory to ML:

N: all players → all data features

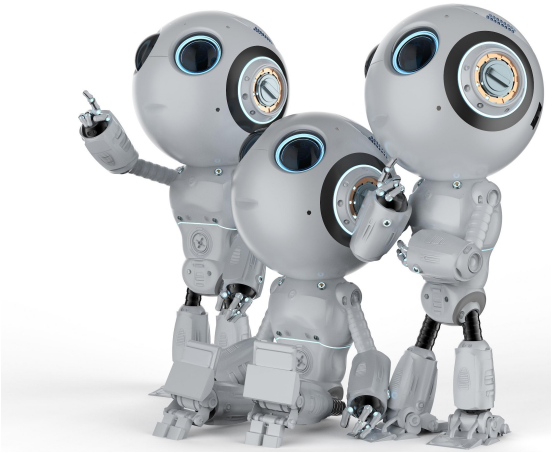
i: player index → feature index

S: coalition of players → set of features

v: game → model

φ: Shapley value → Feature attribution

The Shapley value takes as input a set function $v: 2^N \to R$ and produces attributions φ_i for each player $i \in N$, that add up to $v(N)$.



norwegian open ai lab

NTNU

Inga Strümke - inga.strumke@ntnu.no

Most used Shapley based library: SHAP (python and R)

How would you do this for a machine learning model?
Or: *What is the characteristic function for this problem?*

ML models expect a specific input shape ⇒ Can't just remove input features!

norwegian open ai lab

NTNU

Most used Shapley based library: SHAP (python and R)

ML models expect a specific input shape ⇒ Can't just remove input features!

Instead, SHAP **approximates** the model output via $v(\mathcal{S}) = E[f(\mathbf{x})|\mathbf{x}_{\mathcal{S}} = \mathbf{x}^*_{\mathcal{S}}]$

"Missing features" are sampled from the data set

norwegian open ai lab

NTNU

SHAP

ML models expect a specific input shape ⇒ Can't just remove input features!

Instead, SHAP **approximates** the model output via $v(\mathcal{S}) = E[f(\mathbf{x})|\mathbf{x}_\mathcal{S} = \mathbf{x}^*_\mathcal{S}]$

"Missing features" are sampled from the data set

2017: KernelSHAP (marginal) by *Lundberg et al*

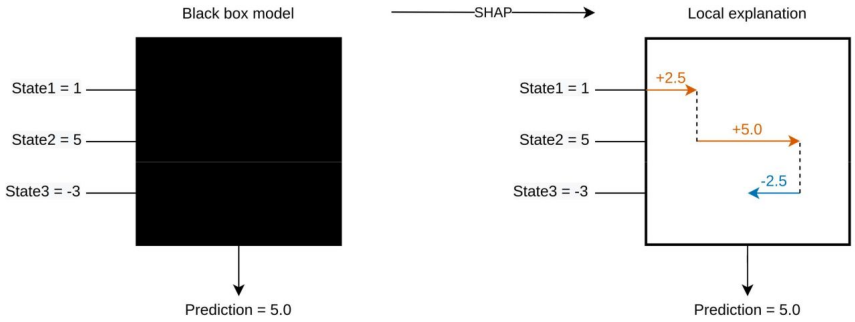
2019: Conditional SHAP *by Aas et al*

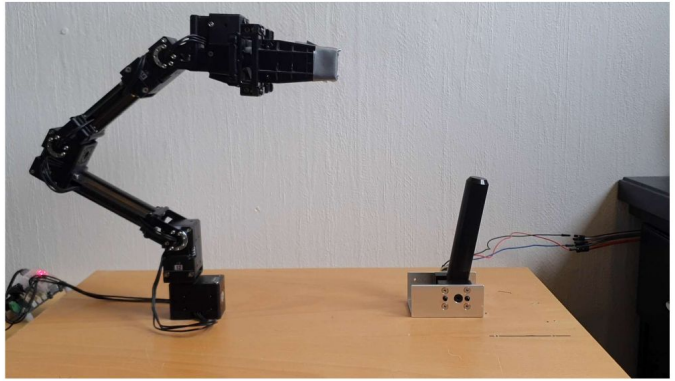
2020: CausalSHAP (interventional) *by Heskes et al*

norwegian open ai lab

NTNU

SHAP for robotic manipulator





arXiv.org > cs > arXiv:2111.02936

Computer Science > Robotics

[Submitted on 4 Nov 2021]

Causal versus Marginal Shapley Values for Robotic Lever Manipulation Controlled using Deep Reinforcement Learning

Sindre Benjamin Remman, Inga Strümke, Anastasios M. Lekkas

Surrogate model explanations

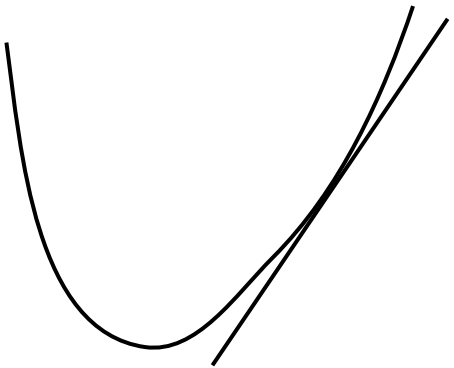
Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

Approximate the full model in a neighborhood

Think: Complex function with a local approximation



Behave similarly *locally*

Do not have the same properties *globally* (derivative, …)

Here (in two dimensions) the neighborhood is easy to identify

In 13 dimensions, sampling 3 points from the neighborhood in each direction would have required >1.5 million samples

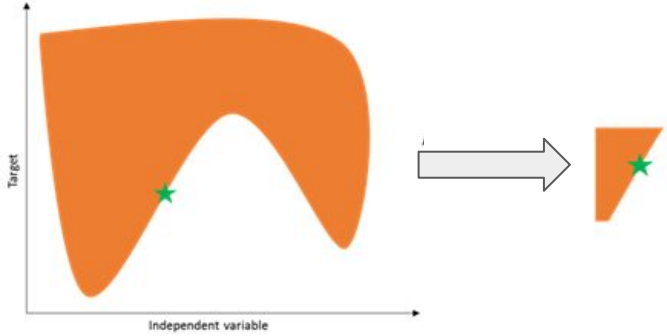
norwegian open ai lab

NTNU

LIME: Local interpretable model-agnostic explanations

Creates an interpretable model locally around the instance, agnostic of full model

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



Find g from a class of interpretable models G that minimises:

Loss $\mathcal{L}$ ← how bad g is at approximating f in the neighborhood π_x

The complexity $\Omega(g)$ of the interpretable model

Inga Strümke - inga.strumke@ntnu.no

norwegian open ai lab

NTNU

LIME: Local interpretable model-agnostic explanations

It's not necessary to use LIME - but it's fast

LIME builds a new local model per instance to explain

```
import lime.lime_tabular

# instantiate explainer object
explainer_lime = lime.lime_tabular.LimeTabularExplainer(X_train,
                                                        feature_names=features,
                                                        verbose=True, mode='regression')
```

Pro:

Con:

norwegian open ai lab

NTNU

LIME: Local interpretable model-agnostic explanations

It's not necessary to use LIME - but it's fast

LIME builds a new local model per instance to explain

```
import lime.lime_tabular

# instantiate explainer object
explainer_lime = lime.lime_tabular.LimeTabularExplainer(X_train,
                                                        feature_names=features,
                                                        verbose=True, mode='regression')
```

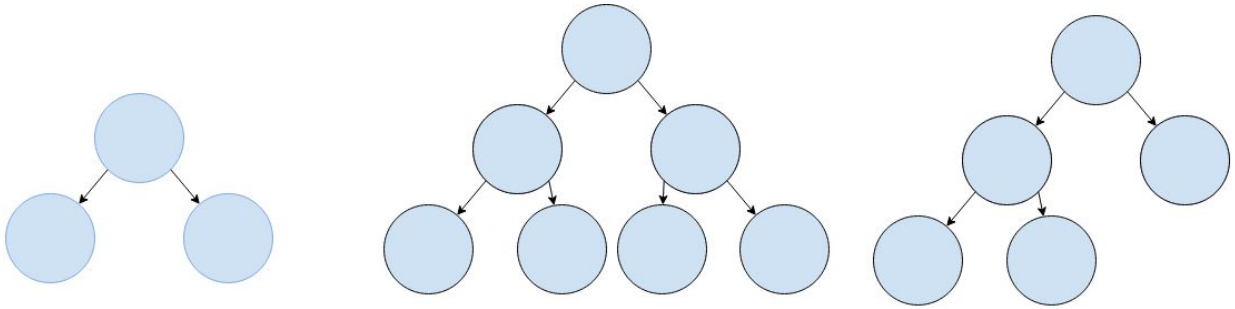
Pro: easy peasy <3

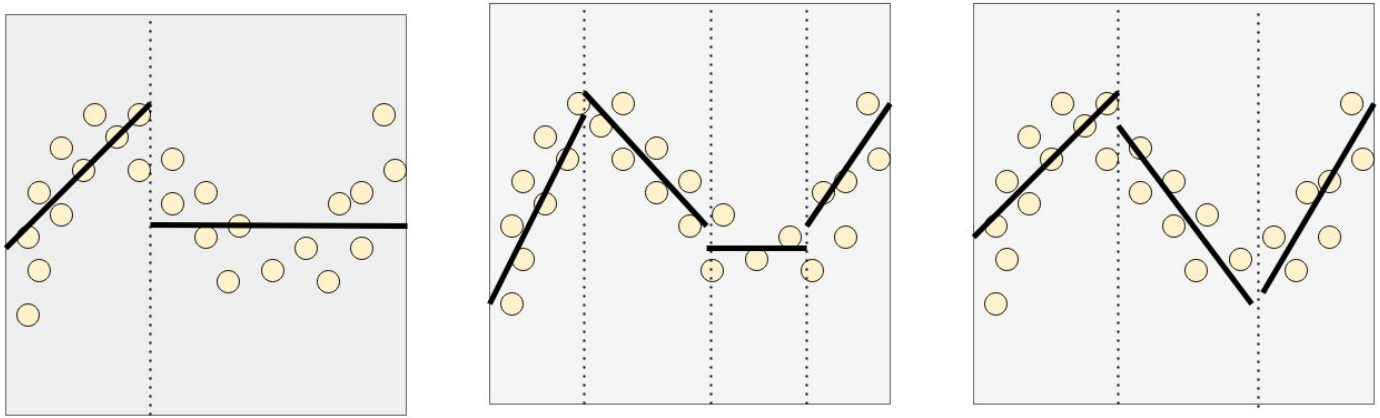
Con: a new black box?

norwegian open ai lab

NTNU

Or: one model, many functions
(Linear) Model Trees:

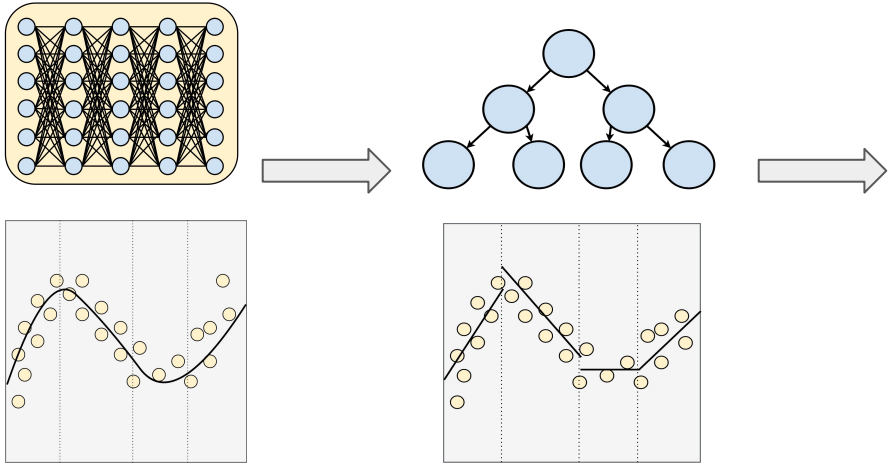


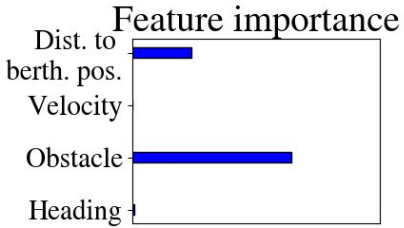


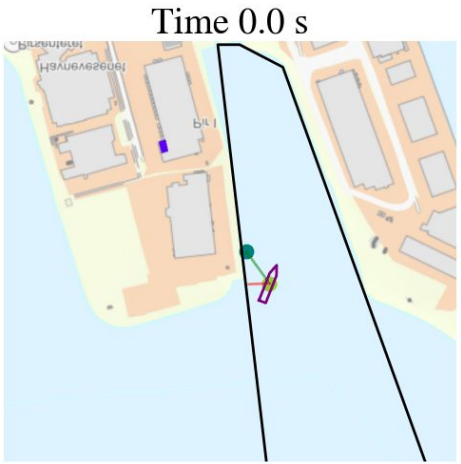
open ai lab

NTNU

LMT surrogate model for explaining an autonomous surface vessel







Thanks to the structure of the LMT, feature importances can be extracted realtime

Had the LMT been even simpler, it would have been directly interpretable

Explaining a Deep Reinforcement Learning Docking Agent Using Linear Model Trees with User Adapted Visualization

by Vilde B. Gjærum [1,*], Inga Strümke [1], Ole Andreas Alsos [2] and Anastasios M. Lekkas [1,3]

Inga Strümke - inga.strumke@ntnu.no

Intrinsic explanations

norwegian open ai lab

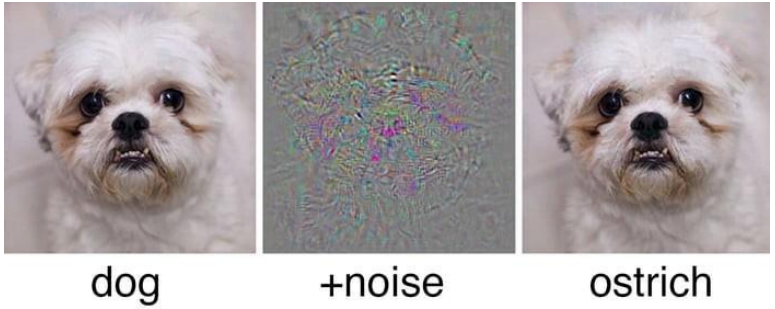
NTNU

Activation maps

Based on the model's components. For a neural network, these are nodes, with

1. activation functions
2. weights

most informative are often the weights' gradients, i.e. how much does the NN output change as function of the *change* of the weights?



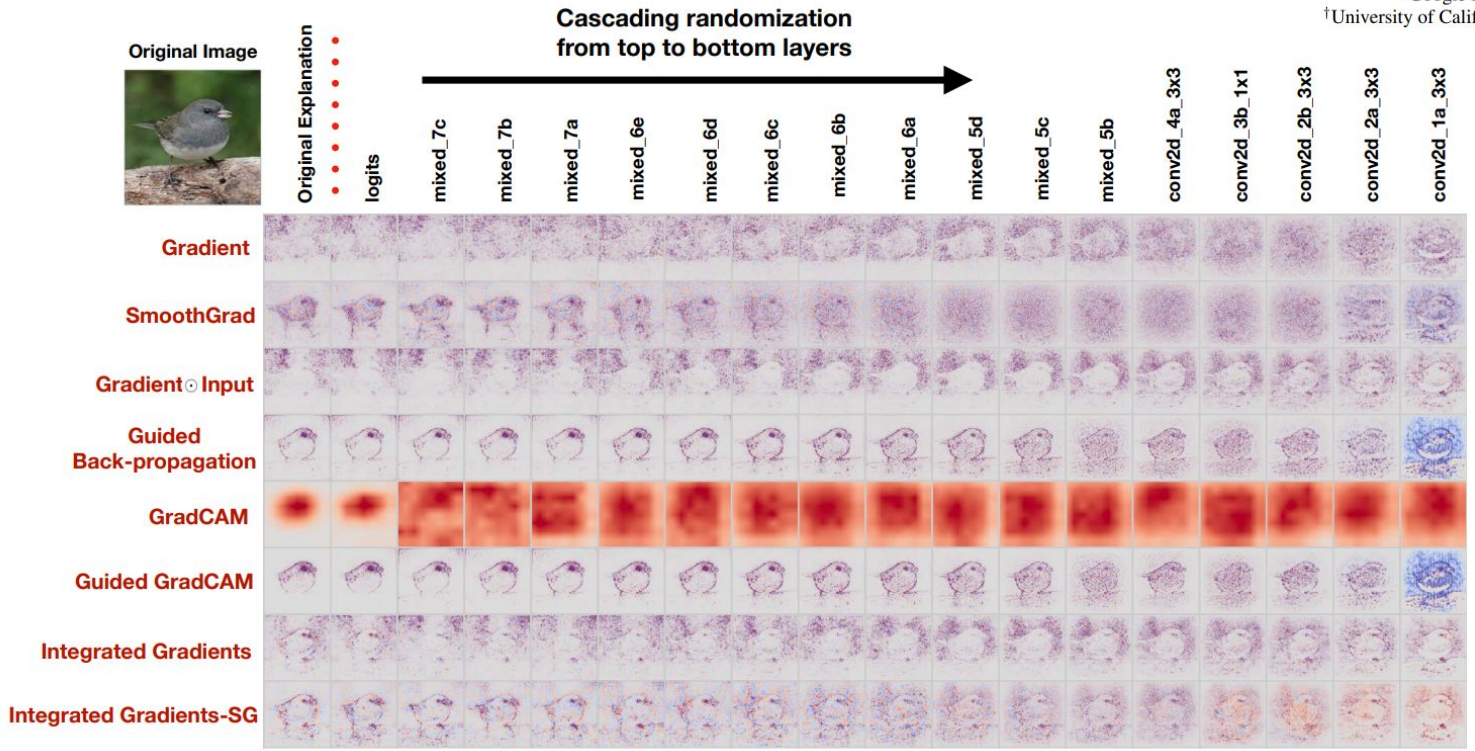
norwegian open ai lab

NTNU

Activation maps

Sanity Checks for Saliency Maps

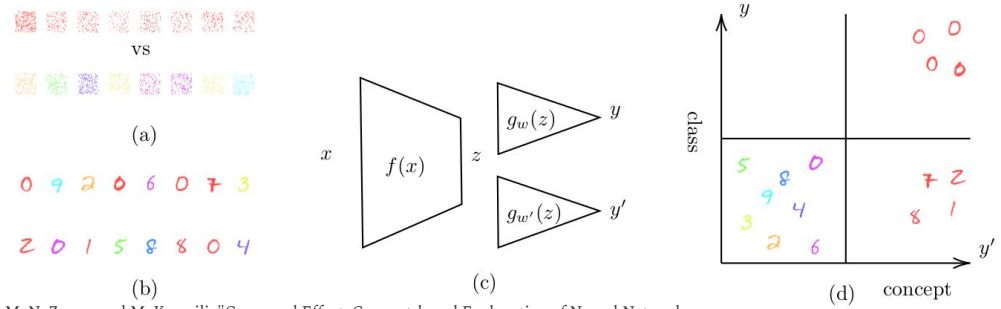
Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley



open ai lab

NTNU

Concepts



M. N. Zaeem and M. Komeili, "Cause and Effect: Concept-based Explanation of Neural Networks

Acquisition of Chess Knowledge in AlphaZero

Thomas McGrath[1,+], Andrei Kapishnikov[2,+], Nenad Tomašev[1], Adam Pearce[2], Demis Hassabis[1], Been Kim[2], Ulrich Paquet[1], and Vladimir Kramnik[3]

[1]DeepMind
[2]Google Brain
[3]World Chess Champion, 2000–2007*

norwegian open ai lab

NTNU

Conclusion: Lots to do!
Thank you

inga.strumke@ntnu.no

norwegian open ai lab

NTNU