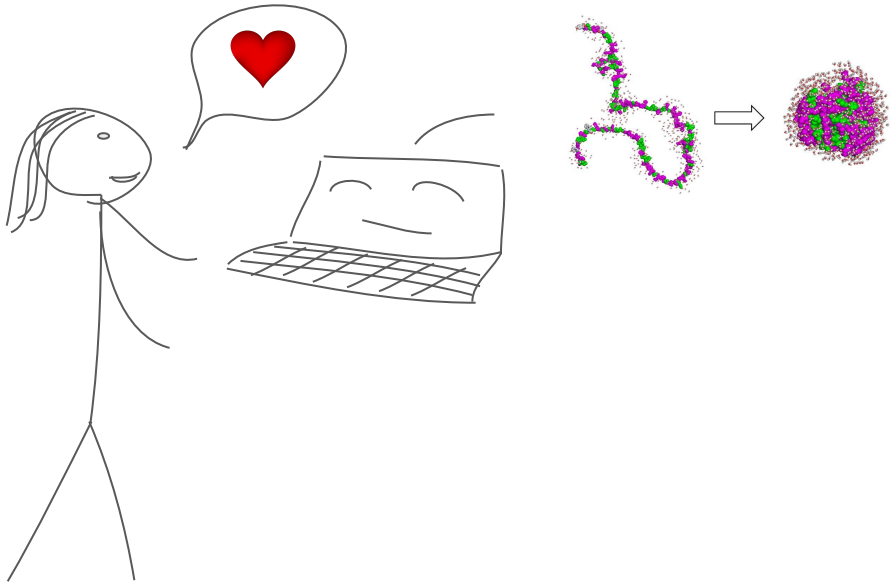
XAI.zip

Inga Strümke

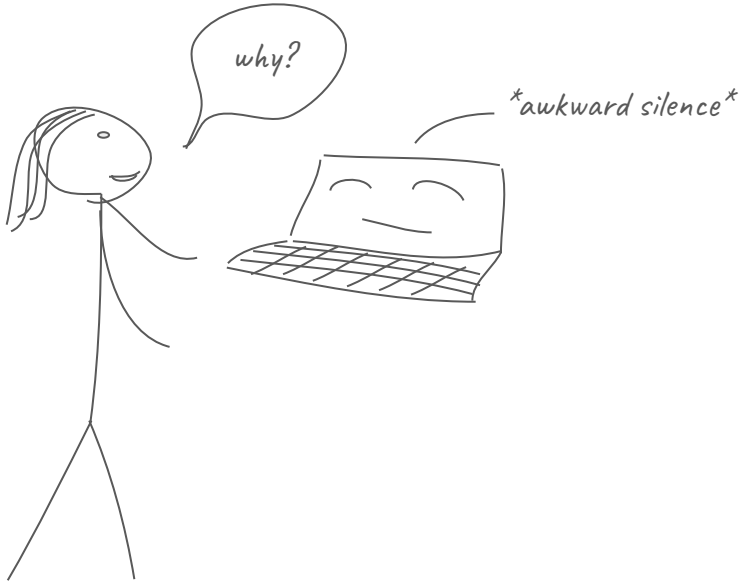
norwegian
open ai lab

NTNU



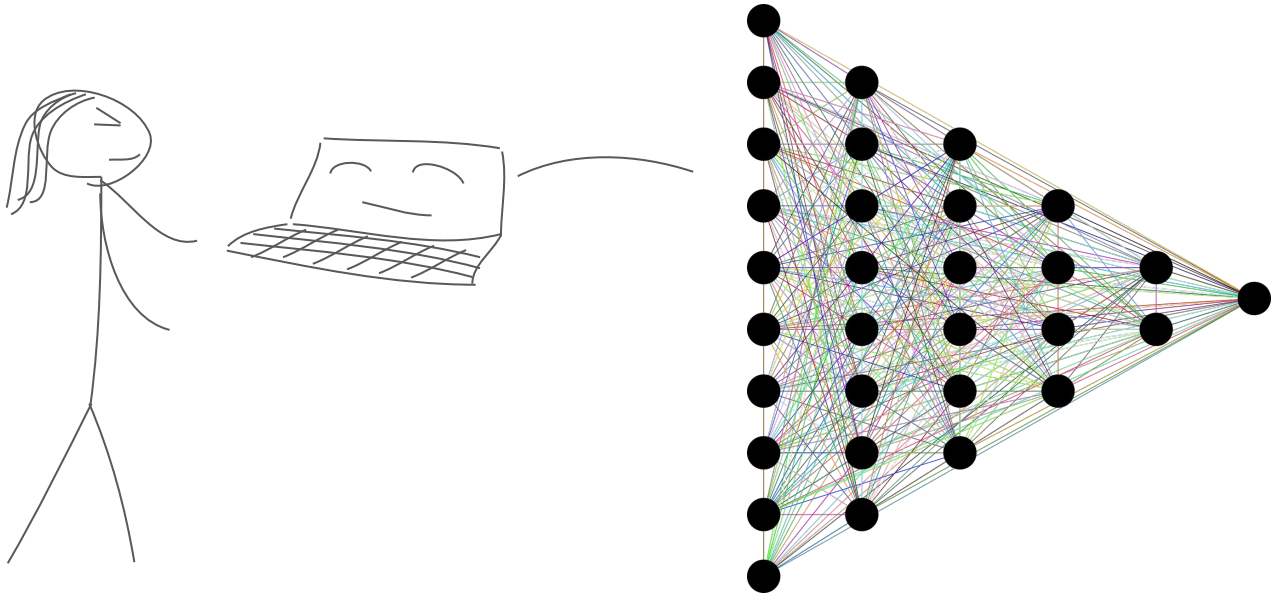
norwegian open ai lab

NTNU



norwegian open ai lab

NTNU



Transparent, traceable,
but not *interpretable*

norwegian
open ai lab

NTNU

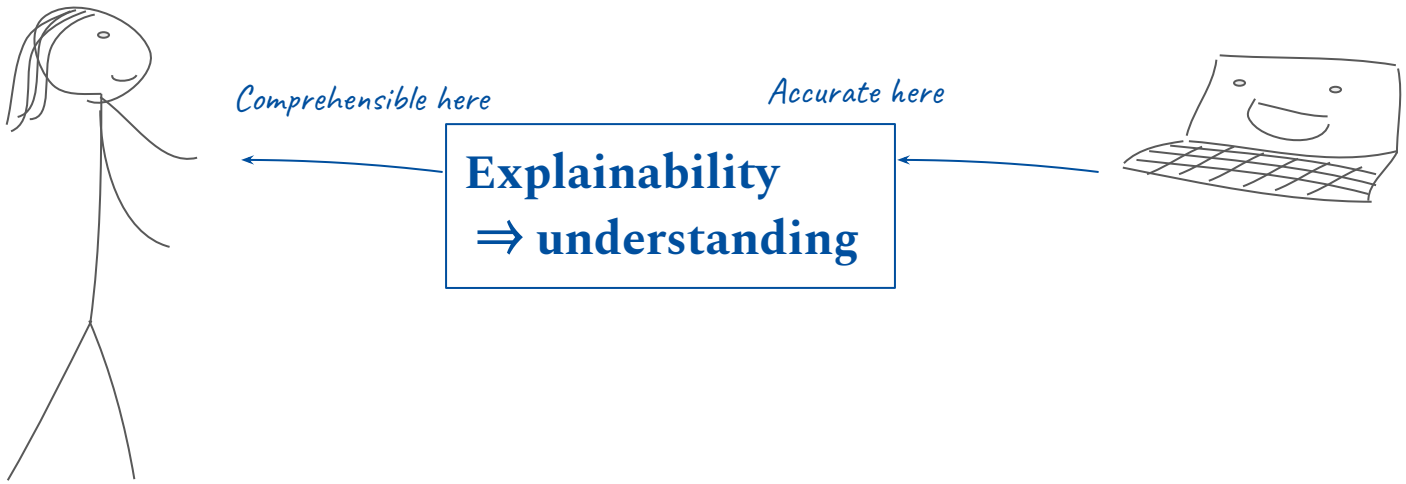
What is it

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

≠ **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.

Explainable AI (XAI) is a set of tools and frameworks to help humans understand predictions made by AI systems.

Understandability is imo the most essential concept in XAI.

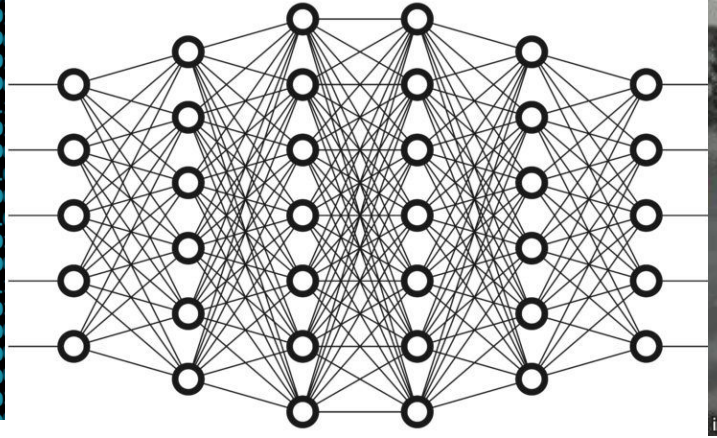


an *interface* between humans and the machine;

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

What seems to be the problem?

We do ML because we don't have a choice
We have data + goal, but no sufficiently complex model of the system/dynamics
ML uses complex, non-linear correlations to optimize



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

Why so unexplainable?

Explainable modelling

Goal:
Figure out how stuff works and *why* things happen

Machine learning

Predictive modelling

Goal:
Foresee what will happen, use past knowledge to estimate

norwegian open ai lab

NTNU

What seems to be the problem?

2) Testing is mostly infeasible due to curse of dimensionality

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

norwegian open ai lab

NTNU

The curse of dimensionality

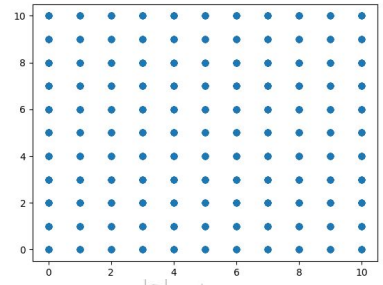
Du har et problem du vil løse med maskinlæring, og samler inn et datasett.

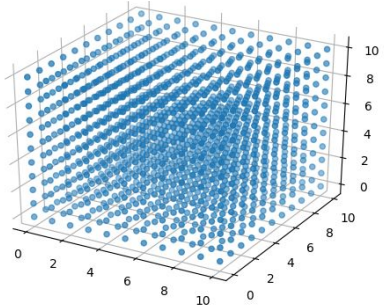
Si: variabelområde = [0, 10], ett datapunkt per heltall

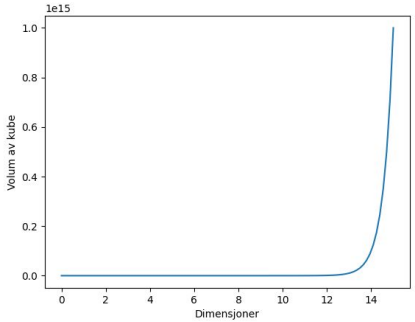
To variable (dimensjoner): $10^2 = 100$ datapunkt

Tre variable: $10^3 = 1000$ datapunkt

13 variable: 10^{13} = ti tusen milliarder datapunkt. Lykke til.



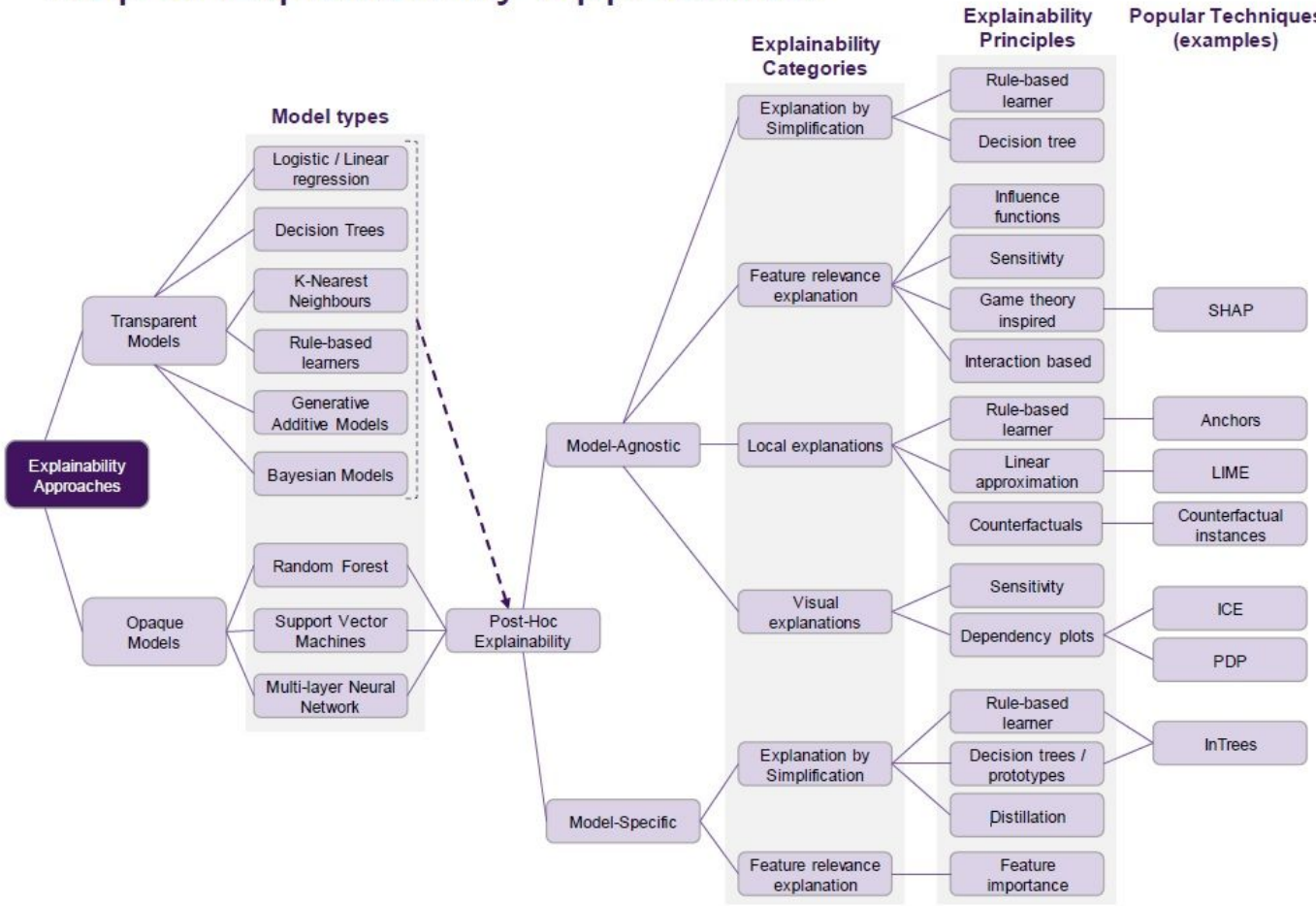




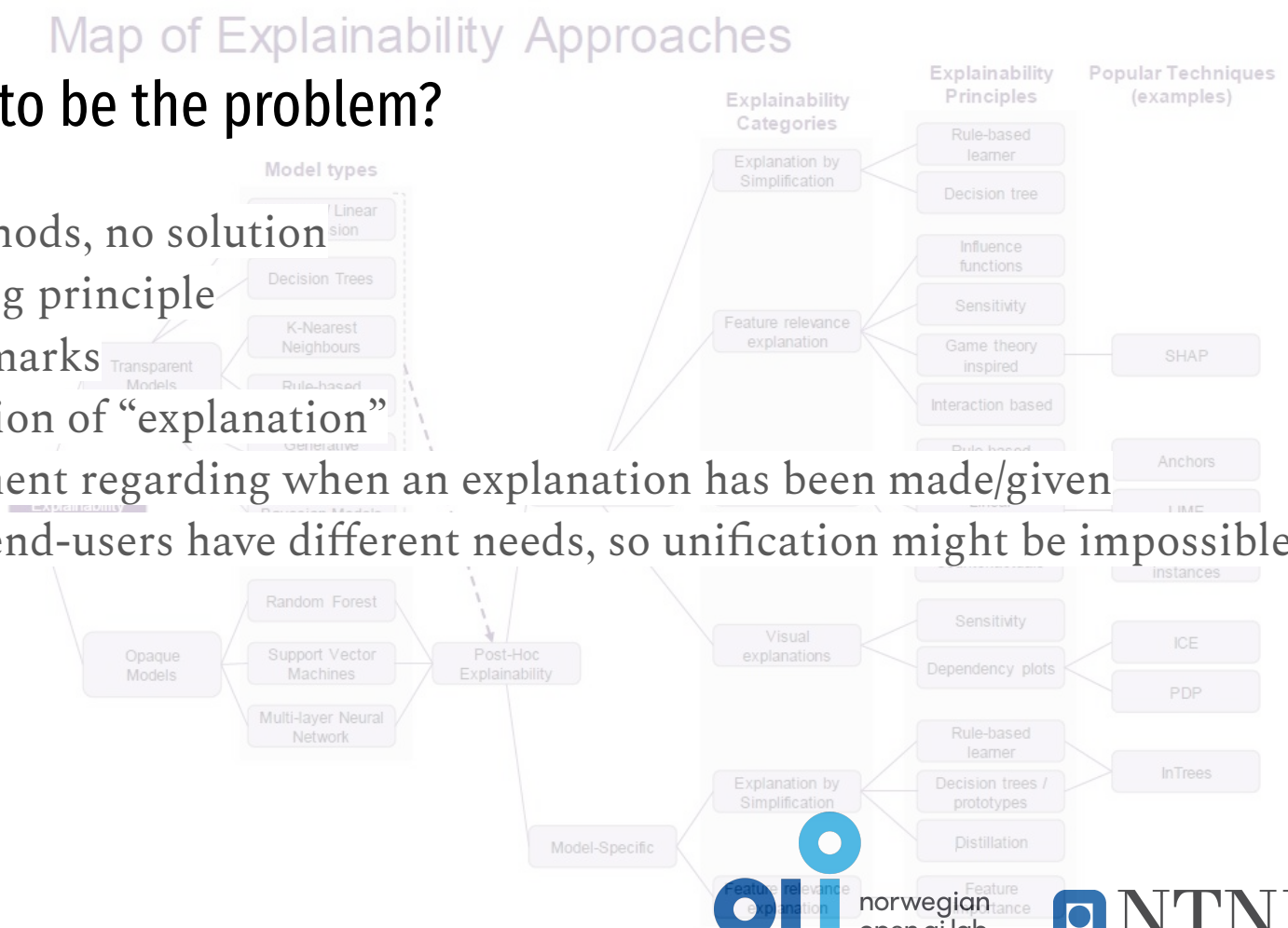
norwegian open ai lab

NTNU

Map of Explainability Approaches



What seems to be the problem?



3) Many methods, no solution
   No unifying principle
   No benchmarks
   No definition of “explanation”
   No agreement regarding when an explanation has been made/given
   Different end-users have different needs, so unification might be impossible

norwegian open ai lab

NTNU

XAI.zip

Three options:

Build an interpretable model that behaves like the box
Surrogate model explanations

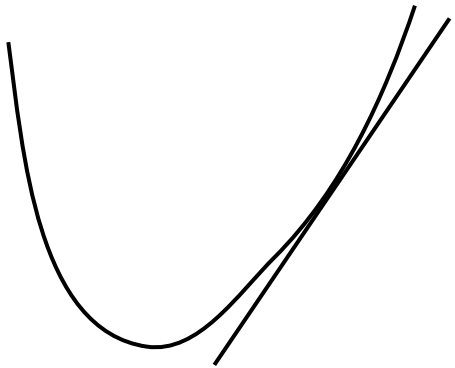
Poke the box systematically and see what happens
Extrinsic explanations

Smash the box and look inside
Intrinsic explanations

Surrogate model explanations

An interpretable model in a neighborhood

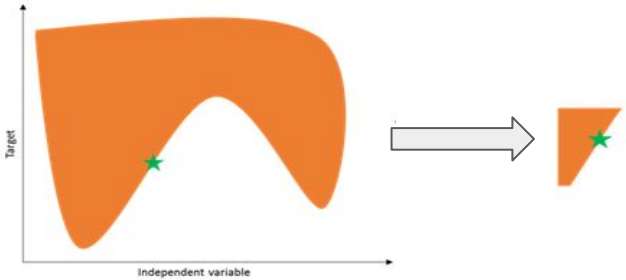
~ the earth is locally flat



Surrogate model explanations

LIME: An interpretable model per neighborhood

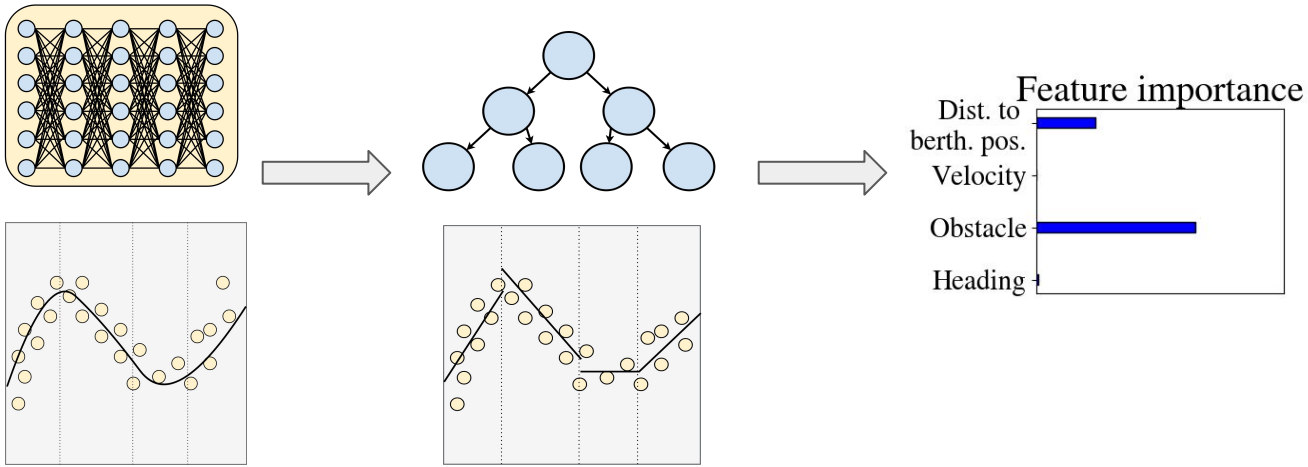
$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

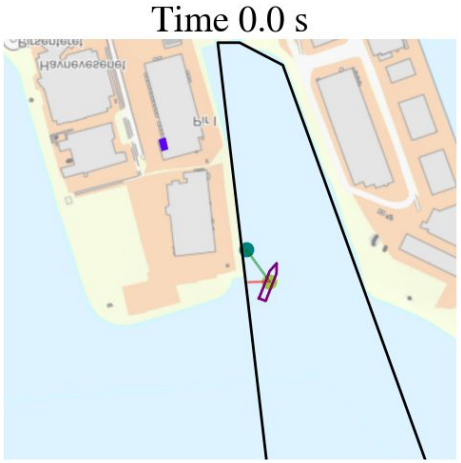


norwegian open ai lab

NTNU

Surrogate model explanations





Linear Model Tree: One linear model in each root node

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

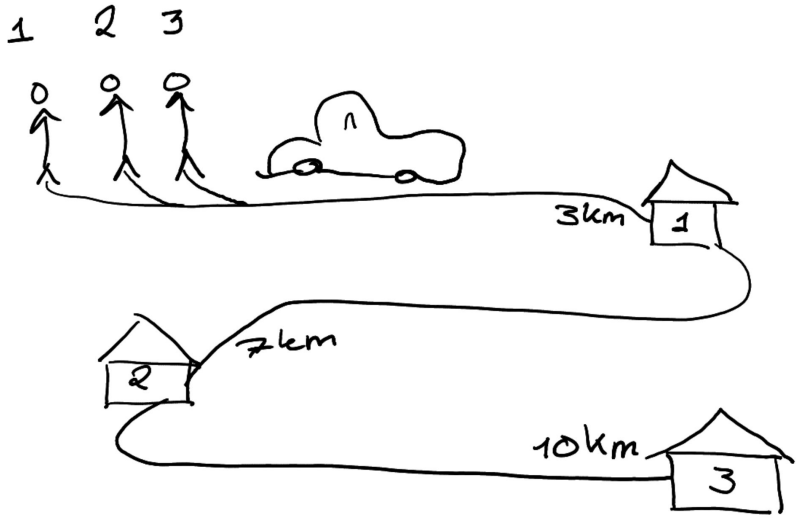
by Vilde B. Gjærum [1,*], Inga Strümke [1], Ole Andreas Alsos [2] and Anastasios M. Lekkas [1,3]

Extrinsic explanations

Feature importances ~ sensitivity analysis

The Shapley decomposition

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



Decomposes gain among players (features)

Need to know **characteristic function values**

norwegian open ai lab

NTNU

Extrinsic explanations

SHAP: Shapley additive explanations: widely used library
indicates how much a feature contributes to model *outcome*, by estimating

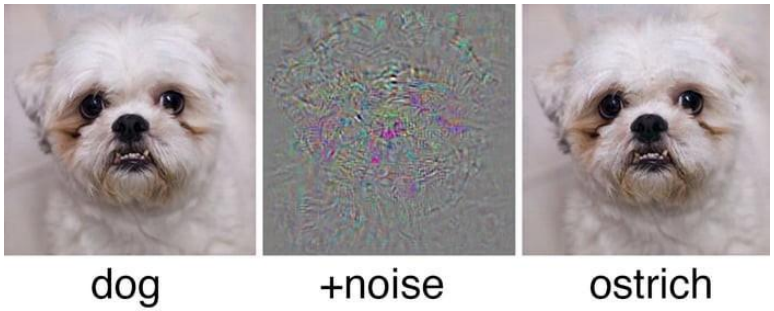
$$v(\mathcal{S}) = E[f(\mathbf{x})|\mathbf{x}_{\mathcal{S}} = \mathbf{x}^*_{\mathcal{S}}]$$

Intrinsic explanations

Activation maps based on model components

1. activation functions
2. weights

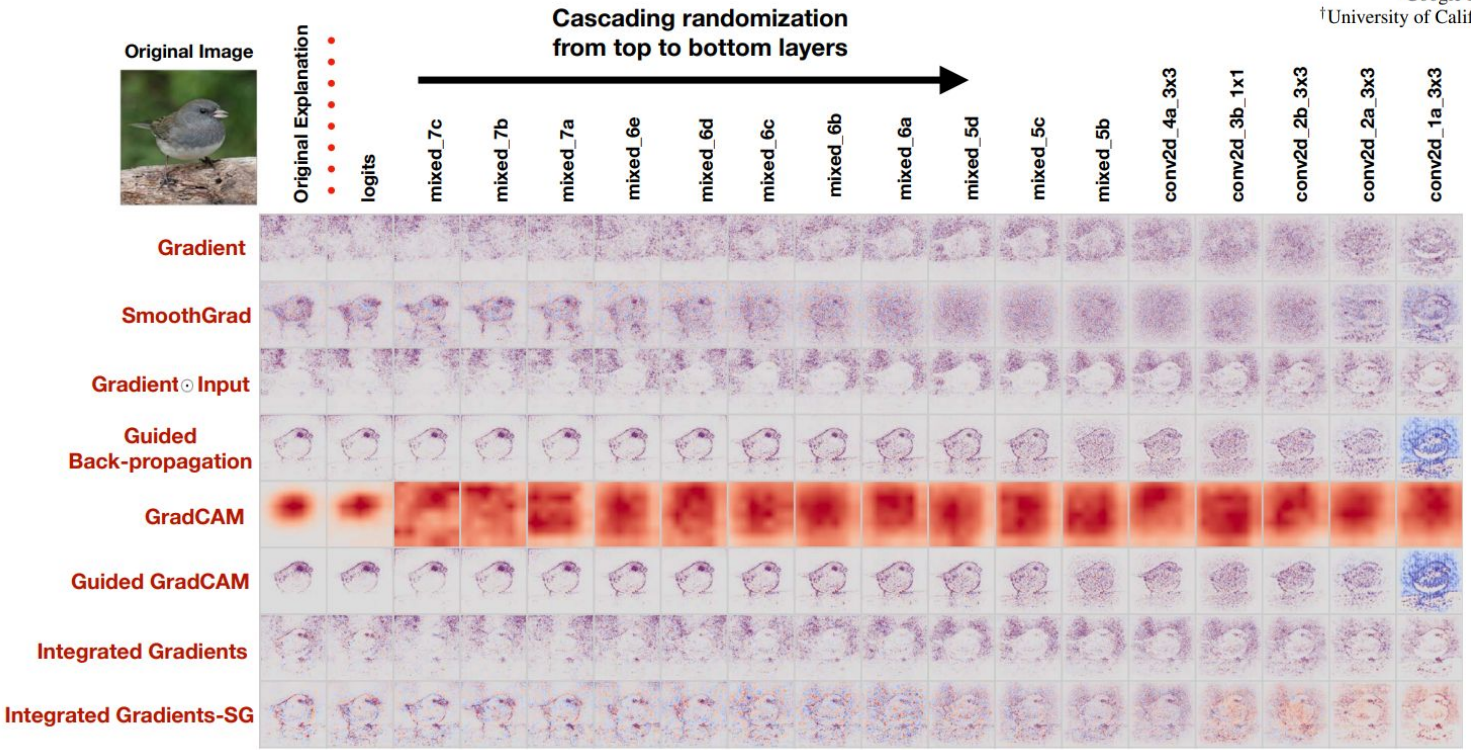
most informative are often the weights' gradients, i.e. how much does the NN output change as function of the *change* of the weights?



Intrinsic explanations

Sanity Checks for Saliency Maps

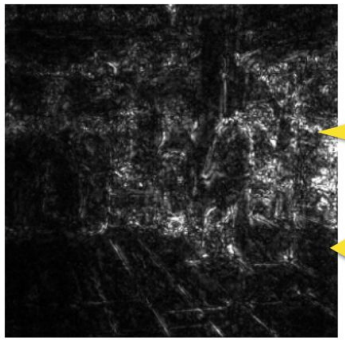
Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

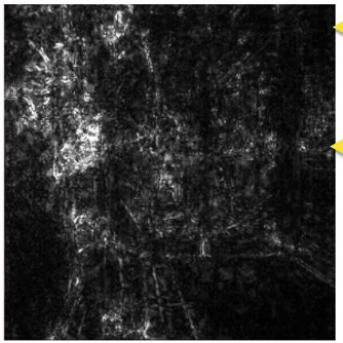


open ai lab

NTNU

Let's look for *concepts*





Were there more pixels on the cash machine than on the person?

Did the 'human' concept matter? Did the 'glasses' or 'paper' matter?

Which concept mattered more?

Is this true for all other cash machine predictions?

The concepts in the images (human, cash machine, sliding door)

1. can't be represented as pixels, and
2. are not explicitly present in the input data.

[1] Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), PMLR 80:2668-2677

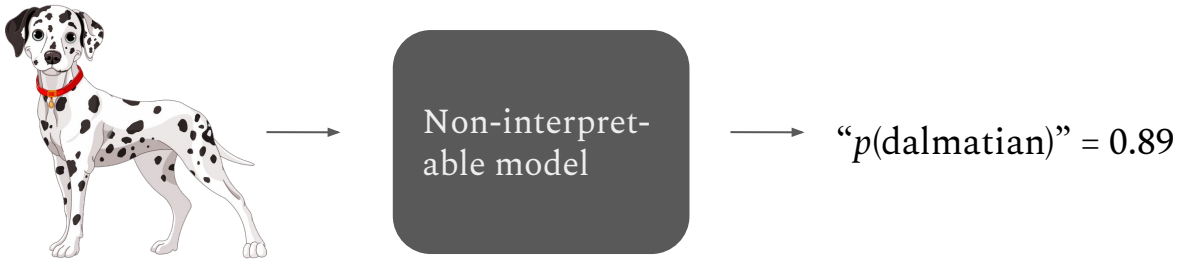
norwegian open ai lab

NTNU

Let's look for *concepts*

Q: How much did a **concept** (anything from ethnicity to shapes, patterns, objects) contribute to the model prediction?

Example: How much did the concept "**dotted**" contribute to the following classification, *even if* the concept was not a feature of the training data?

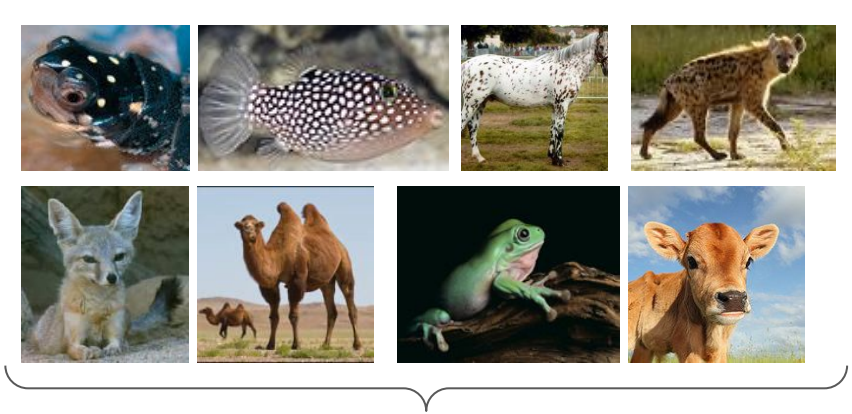


norwegian open ai lab

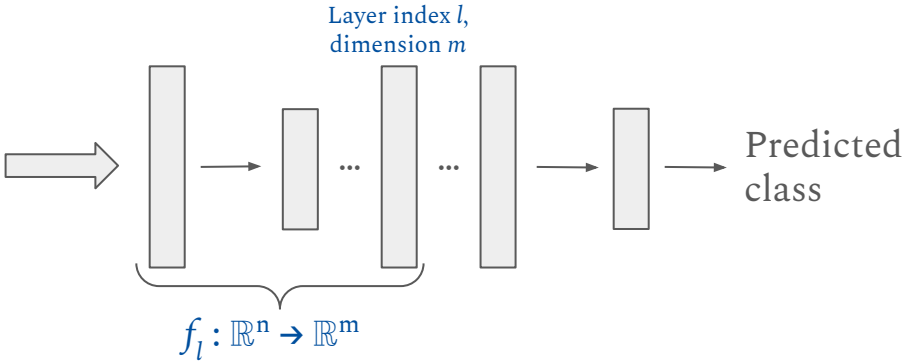
NTNU

Let's look for concepts

A: We can find out by doing concept detection, i.e. look for **concept activation vectors** inside the neural network model, by testing for the concept:



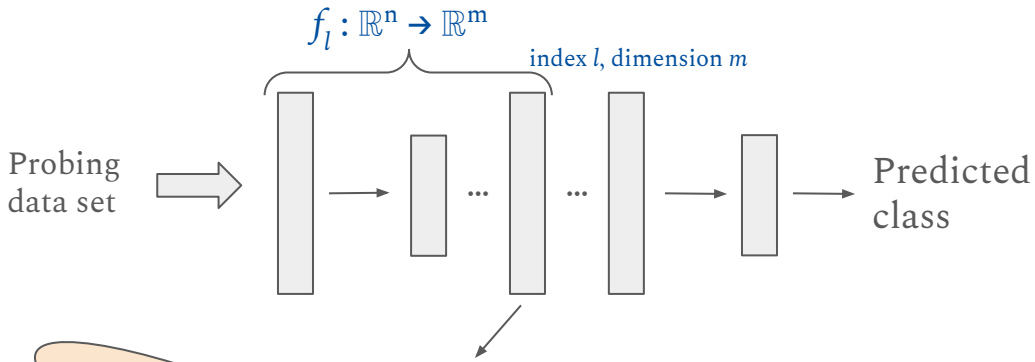
Test data set representing the concept, *and* random images

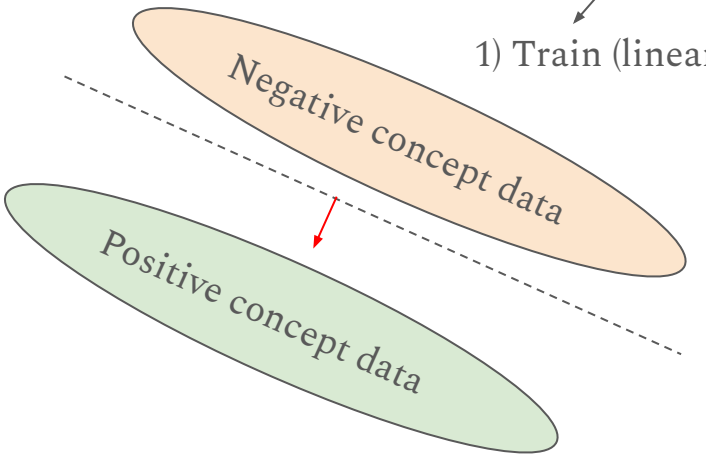


norwegian open ai lab

NTNU

Let's look for concepts





2) The vector orthogonal to the decision boundary is the concept activation vector v^l_C for layer l and concept C.

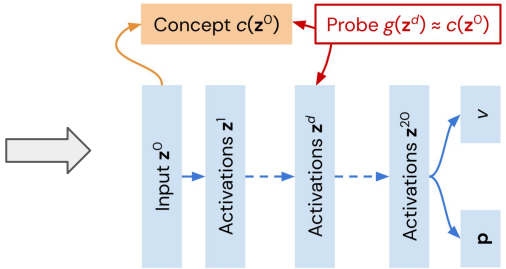
norwegian open ai lab

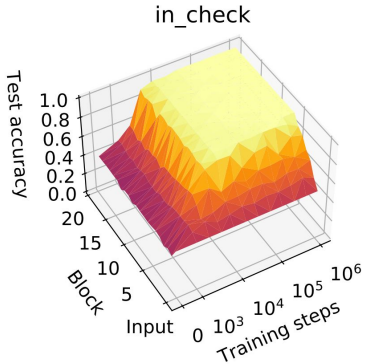
NTNU

Explanations with concepts

Personal favourite application: detection of learned concepts in AlphaZero [1]

Concept names	Description
pawn_fork [m|o]	True if a pawn is attacking two pieces of higher value (knight, bishop, rook, queen, or king) and is not pinned.
knight_fork [m|o]	True if a knight is attacking two pieces of higher value (rook, queen, or king) and is not pinned.
bishop_fork [m|o]	True if a bishop is attacking two pieces of higher value (rook, queen, or king) and is not pinned.
rook_fork [m|o]	True if a rook is attacking two pieces of higher value (queen, or king) and is not pinned.
has_pinned_pawn [m|o]	True if the side has a pawn that is pinned to the king of the same colour.
has_pinned_knight [m|o]	True if the side has a knight that is pinned to the king of the same colour.
has_pinned_bishop [m|o]	True if the side has a bishop that is pinned to the king of the same colour.
has_pinned_rook [m|o]	True if the side has a rook that is pinned to the king of the same colour.
has_pinned_queen [m|o]	True if the side has a queen that is pinned to the king of the same colour.
material [m|o|diff]	Material calculated as (♯♙) + 3 * (♯♘) + 3 * (♯♗) + 5 * (♯♖) + 9 * (♯♕)
num_pieces [m|o|diff]	Number of pieces that a side has.
in_check	True if the side that makes a turn is in check.
has_bishop_pair [m|o]	True if the side has a pair of bishops.
has_connected_rooks [m|o]	True if the side has connected rooks.
has_control_of_open_file [m|o]	True if the side controls an open file (with the rooks, queen)
has_mate_threat	True if the opponent could mate the current side in a single move if the turn was passed to the opponent.
has_check_move [m|o]	True if the side can check the opponent's King.
can_capture_queen [m|o]	True if the side can capture the opponent's queen.
num_king_attacked_squares [m|o|diff]	The number of squares around the opponent's king that the playing side attacks. Can include occupied squares.
has_contested_open_file	True if an open file is occupied simultaneously by a rook and/or queen of both colours.
has_right_bc_ha_promotion [m|o]	True if 1) the side has a passed pawn on a or h files and 2) the side has a bishop that is of the colour of the promotion square of that pawn.
num_scb_pawns_same_side [m|o|diff]	The number of own pawns that occupy squares of the same colour as the colour of own bishop. Applicable only when the side has a single bishop.
num_ocb_pawns_same_side [m|o|diff]	The number of own pawns that occupy squares of the opposite colour to that of own bishop. Applicable only when the side has a single bishop.
num_scb_pawns_other_side [m|o|diff]	The number of opponent's pawns that occupy the squares of the same colour as the colour of own bishop. Applicable only when the side has a single bishop.
num_ocb_pawns_other_side [m|o|diff]	The number of opponent's pawns that occupy the squares of the opposite colour to the colour of own bishop. Applicable only when the side has a single bishop.
capture_possible_on_{sq} [m|o] sq=[d1|d2|d3|e1|e2|e3|g5|b5]	True is the side can capture a piece on the given square. The squares are named as if the side were playing White.
capture_happens_next_move_... ...on_{sq} sq=[d1|d2|d3|e1|e2|e3|g5|b5]	True if the capture of a piece on the given square had happened according to the game data. The squares are named as if the side were playing White.





(c) Is the playing side in check?

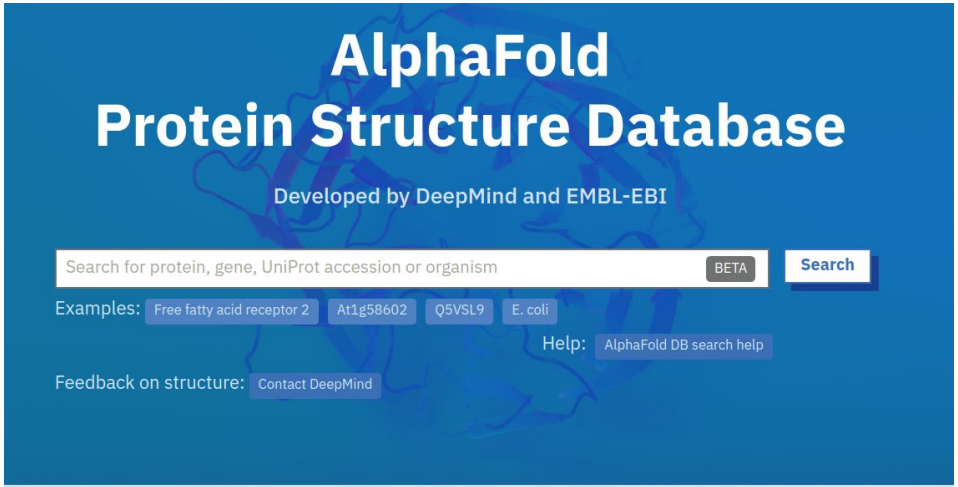
[1] Acquisition of chess knowledge in AlphaZero, DOI:10.1073/pnas.2206625119
Longer (nicer :)) version: ArXiv:2111.09259

norwegian open ai lab

NTNU

(possible fourth problem)

Machines might model non-human concepts

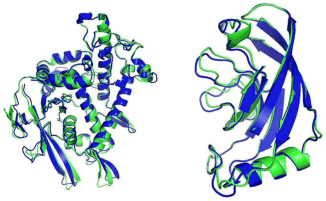


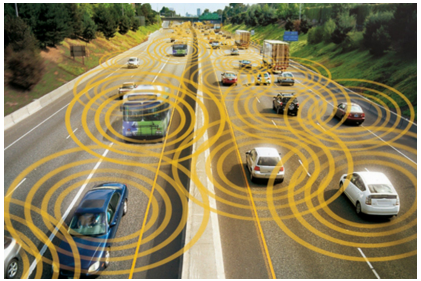
AlphaFold DB provides open access to protein structure predictions for the human proteome and other key proteins of interest, to accelerate scientific research.

norwegian open ai lab

NTNU

XAI <3





Fundamental research

Applications

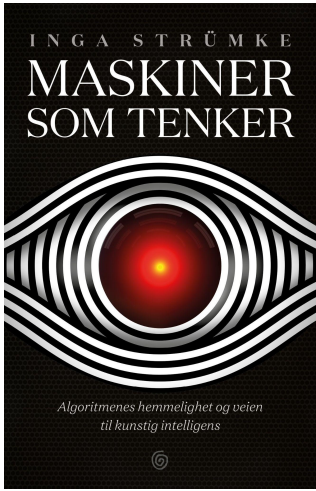
norwegian open ai lab

NTNU

Thank you

Inga Strümke

PS: pls order book ♥



norwegian open ai lab

NTNU