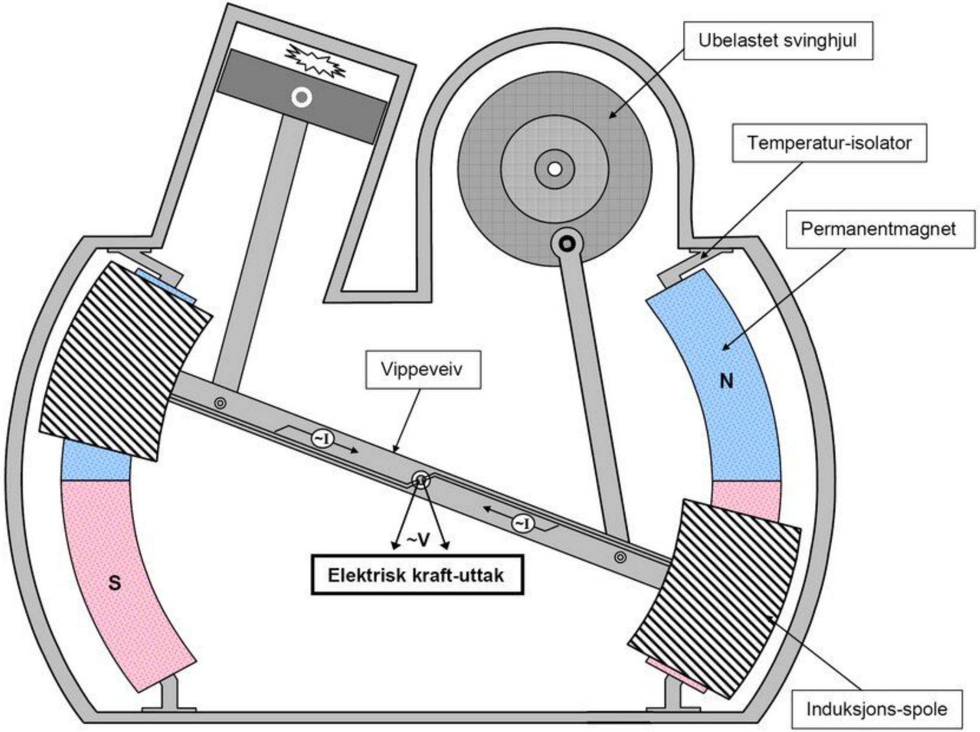
Explaining AI
Why, why not and how

Inga Strümke

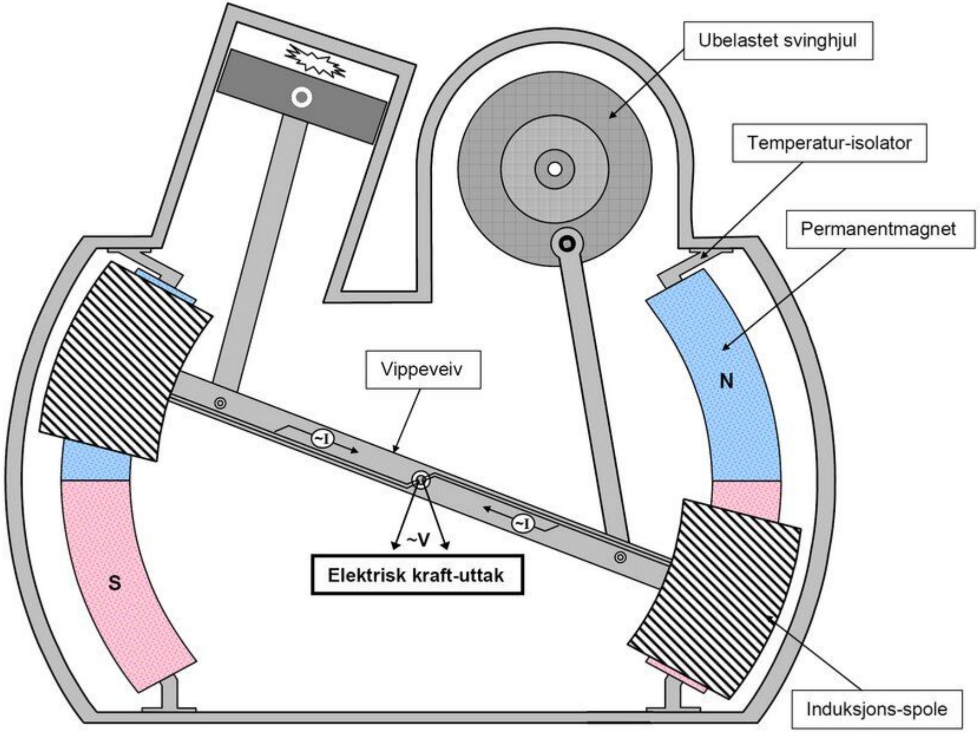
The engine: An explanation



Courtesy of Teknisk Ukeblad

The engine: An explanation

(relax)



Courtesy of Teknisk Ukeblad

Today's menu

➔ **Why?**
3 perspectives and needs for explanations.

➔ **Why not?**
Mathematics, we have a problem.

➔ **How?**
Many directions, methods and open questions

Let’s start with some basic understanding

AI = Artificial intelligence.

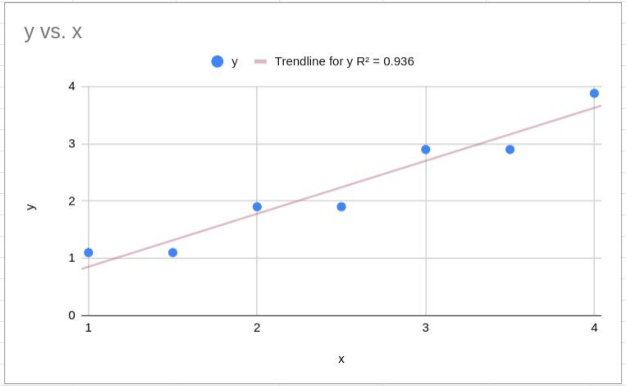
XAI = eXplainable AI

In the present discussion, AI means machine learning

Let’s start with some basic understanding

Machine learning: A program that learns from data (i.e. from experience)

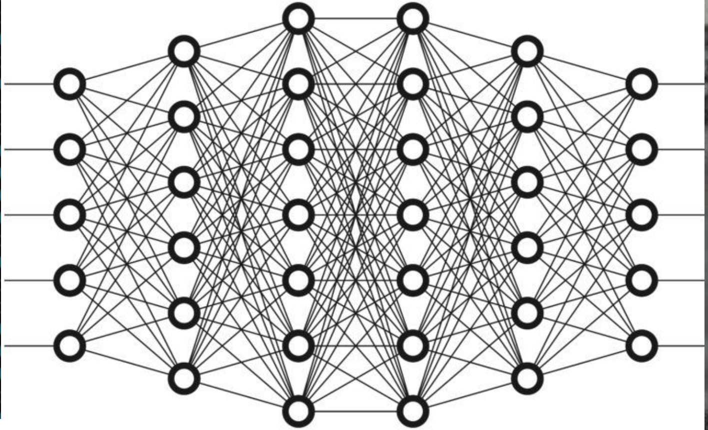
x	y
1	1.1
1.5	1.1
2	1.9
2.5	1.9
3	2.9
3.5	2.9
4	3.88



DIY ML in Excel!

Let’s start with some basic understanding

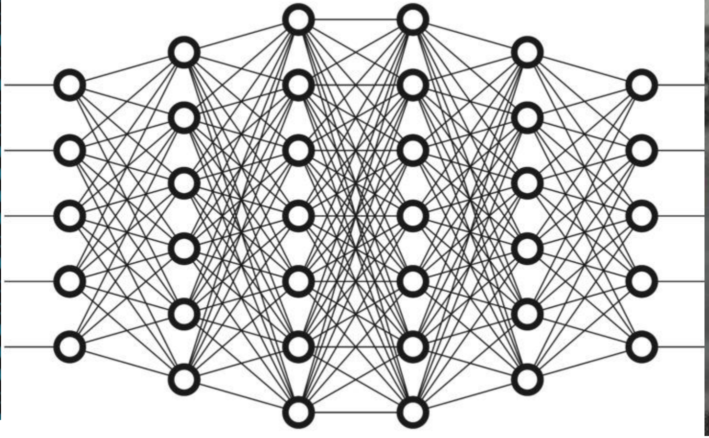
Machine learning: A program that learns from data (i.e. from experience)



Inga Strümke - inga.strumke@ntnu.no

Let’s start with some basic understanding

“Artificial intuition”



Inga Strümke - inga.strumke@ntnu.no

Part 1: Legal requirements

In the EU (and parts of the US) customers are protected by regulations

GDPR:

"meaningful information about the logic of processing"

The end user doesn't care about your model;

they care about how they can affect the outcome

⇒ *A convincing story about how the relevant part of the world works*

Inga Strümke - inga.strumke@ntnu.no

Note: The legal status isn't clear yet

GDPR:

"an explanation of the decision reached after [algorithmic] assessment"

However, it's not specified what such an explanation entails

THE RIGHT TO EXPLANATION, EXPLAINED

Margot E. Kaminski†

DOI: https://doi.org/10.15779/Z38TD9N83H
© 2019 Margot E. Kaminski.
Associate Professor of Law, University of Colorado Law School.

Problem? Yes. But: New EU proposal for AI regulation released April 2021.

Part 1: Legal requirements

Stakeholder / business leader must understand in order to evaluate risk and defend decisions. Risk is a central aspect in business decisions.

New EU proposal for AI regulation also has a ***risk based approach.***

Question: Is the system ***trustworthy?***

84%

Of respondents think that AI based decisions need to be explainable in order to be trusted (PwC CEO Survey 2019)

The traditional approach is testing…

Inga Strümke - inga.strumke@ntnu.no

Regarding testing:

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

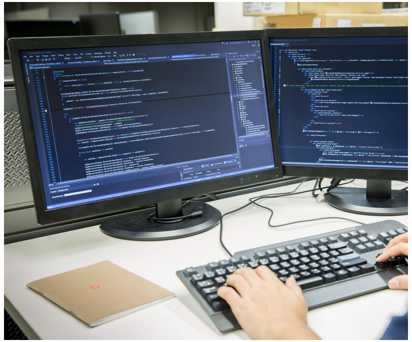
(How can this happen?? Stay tuned)

Inga Strümke - inga.strumke@ntnu.no

Part 2: Researchers and developers

Wants to know

➔ *Which part of the* ***world*** *has my* ***model*** *understood?*



Inga Strümke - inga.strumke@ntnu.no

What, 'which part'? My model is perfect.

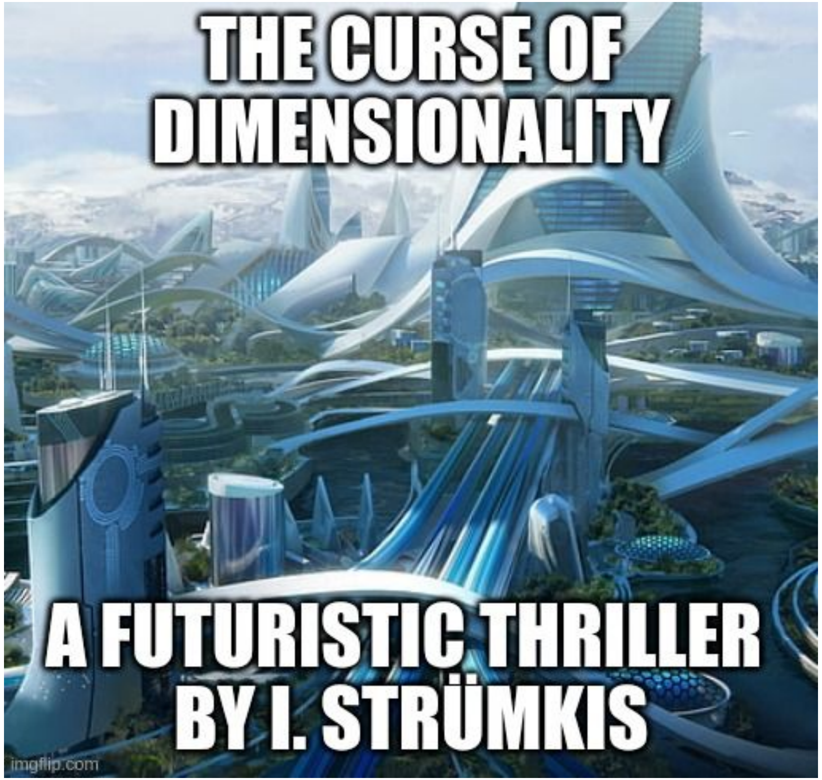
"any two optimization algorithms are equivalent when their performance is averaged across all possible problems"

-No free lunch theorems, Wolpert and Macready (2005)

Your model is never perfect. But it can have a high accuracy where you tested it.

Inga Strümke - inga.strumke@ntnu.no

The curse of dimensionality



Inga Strümke - inga.strumke@ntnu.no

The curse of dimensionality

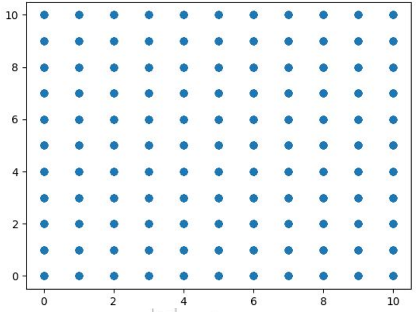
You want to solve some problem using machine learning, and collect a data set...

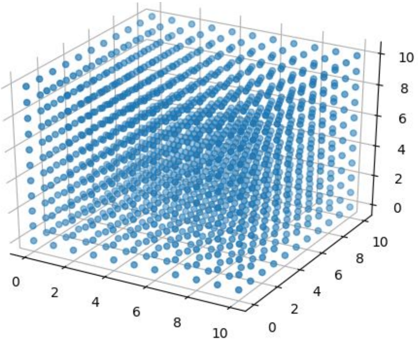
Say... feature values in the range [0, 10], one data point per integer

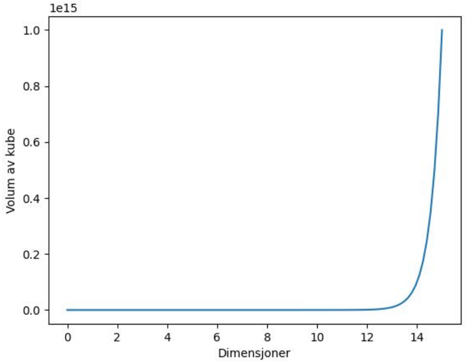
Two features (dimensions): 10^2 = 100 data points

Three features: 10^3 = 1000 data points

13 features: 10^{13} = ten thousand billion data points. Good luck.







Inga Strümke - inga.strumke@ntnu.no

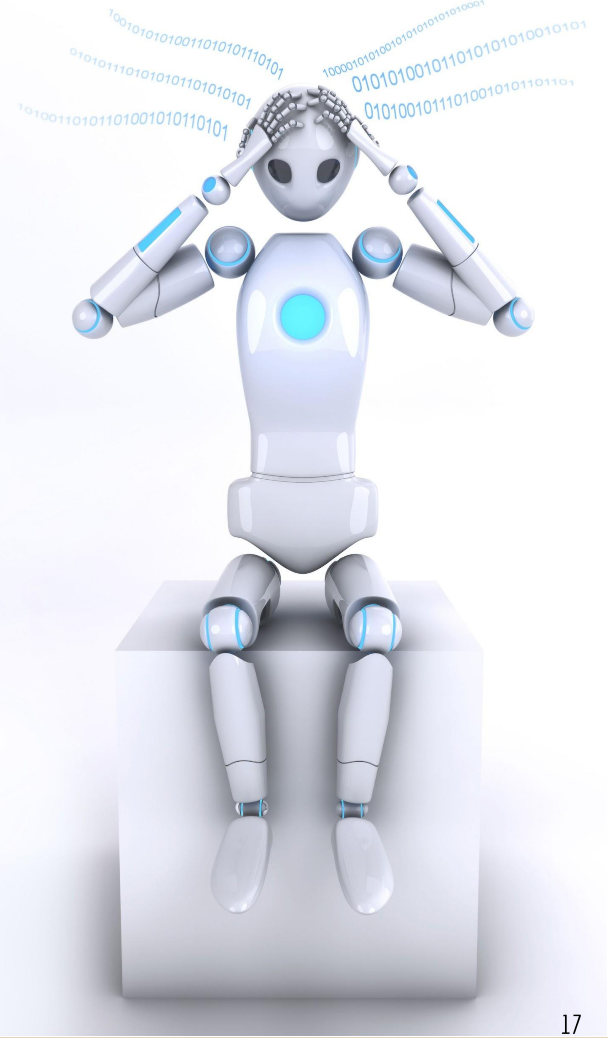
What to do?

Explainable AI (XAI) is a fairly young and active field of research. Nobody knows exactly what to do. But:

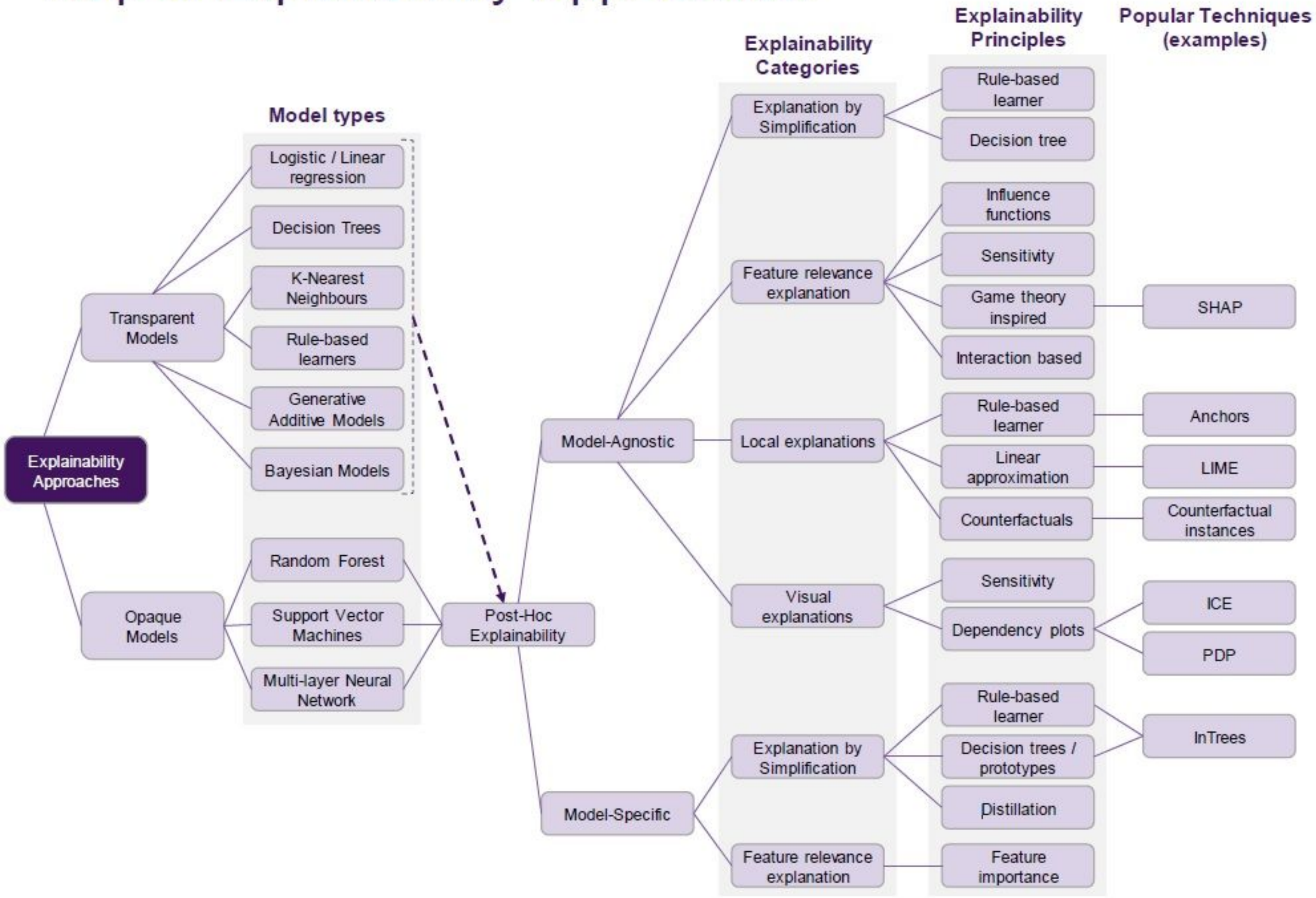
1. *Accept that human intuition won't cut it. We can't understand complex models by looking at them.*

2. *Methods. New methods are developed all the time.*

Let's have a look at methods



Map of Explainability Approaches

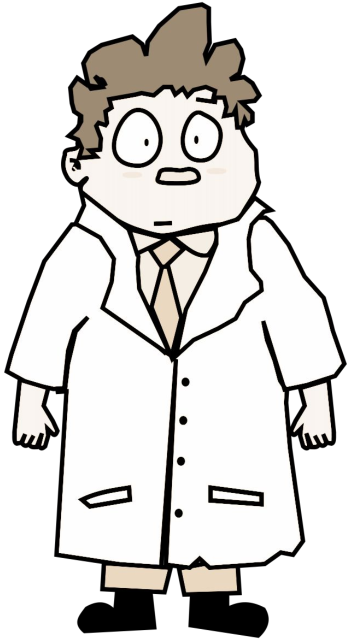


A mystery box fell from the sky

A mystery box fell from the sky

Information ⟶  ⟶ Predictions

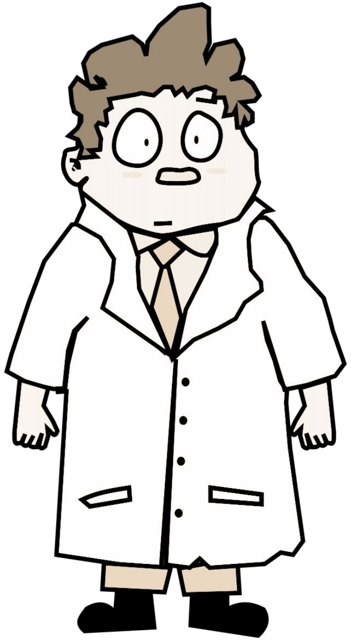
A mystery box fell from the sky



Scientist

Inga Strümke - inga.strumke@ntnu.no

What do do with the mystery box?



Scientist

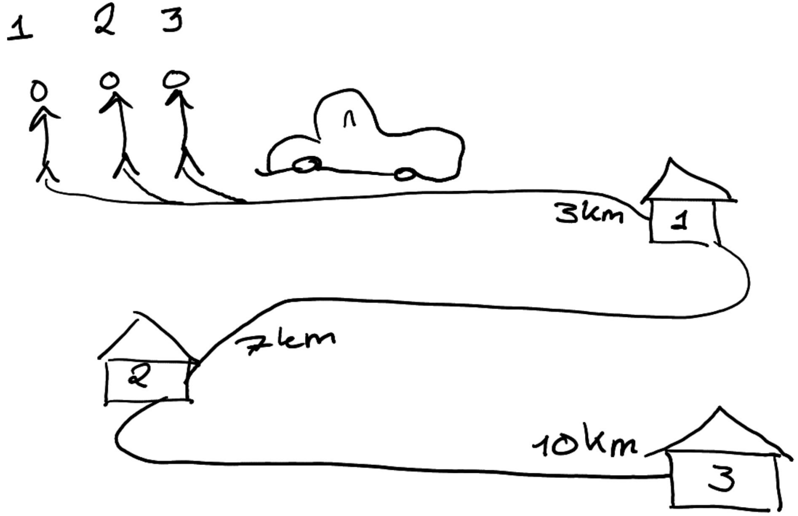
- ➔ Poke from outside
- ➔ Look inside

Poke the box like a boss

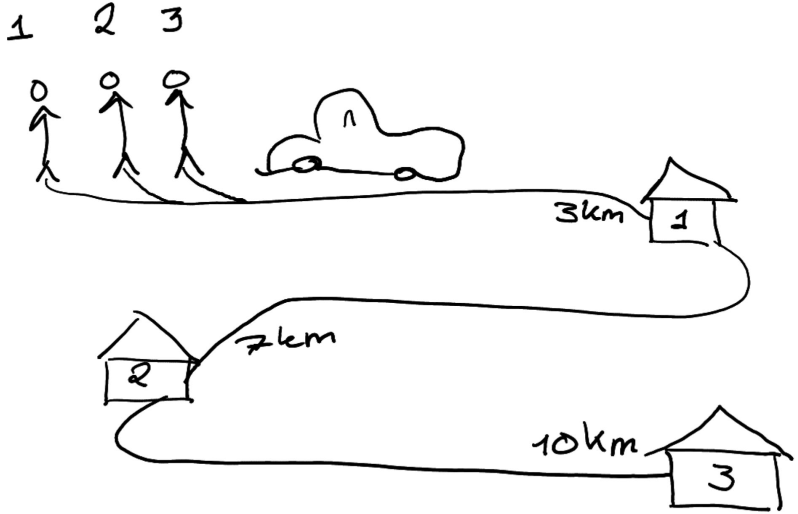
==> Feature attribution

The Shapley decomposition: $$\phi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right)$$

Shapley value example: Cab sharing



Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$

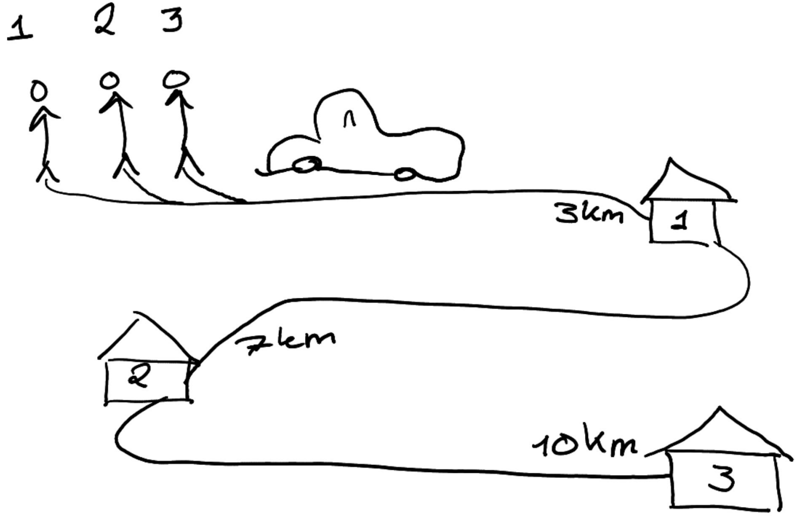
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$

$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$

$$v(\{1,2,3\}) = 10$$

Characteristic function values

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$

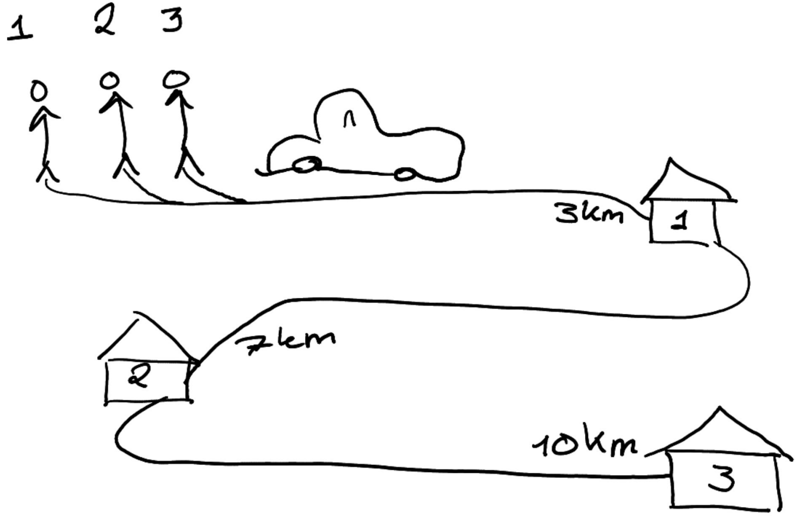
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$

$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$

$$v(\{1,2,3\}) = 10$$

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

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$

$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$

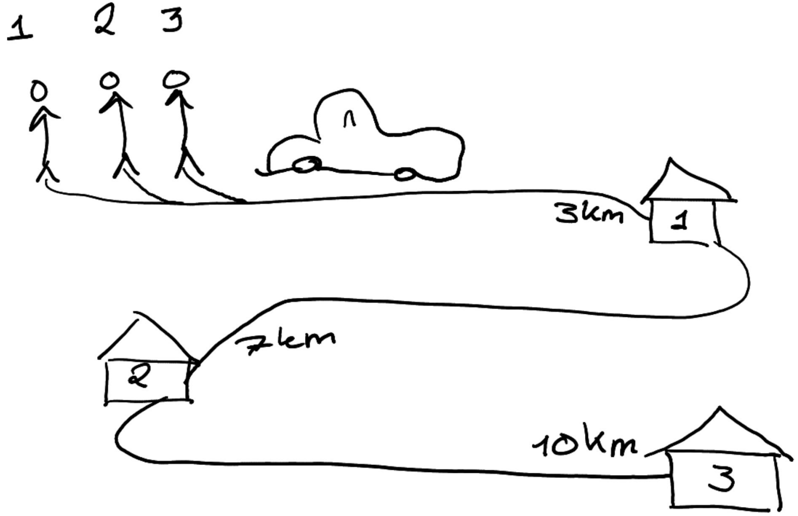
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$

$$v(\{1,2,3\}) = 10$$

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

N = total passengers

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$

$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$

$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$

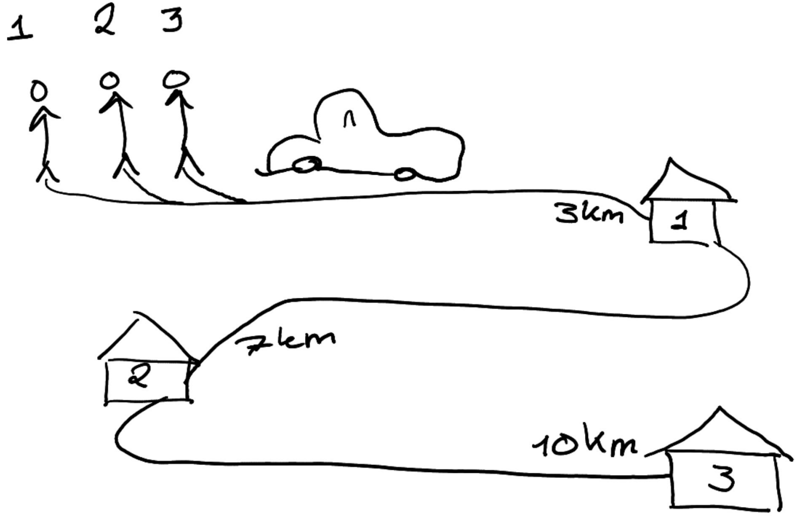
$$v(\{1,2,3\}) = 10$$

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

N = total passengers

S = subsets of N

Shapley cab sharing



$$v(\{\}) = 0 \quad \text{(no passengers costs nothing)}$$

$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$

$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$

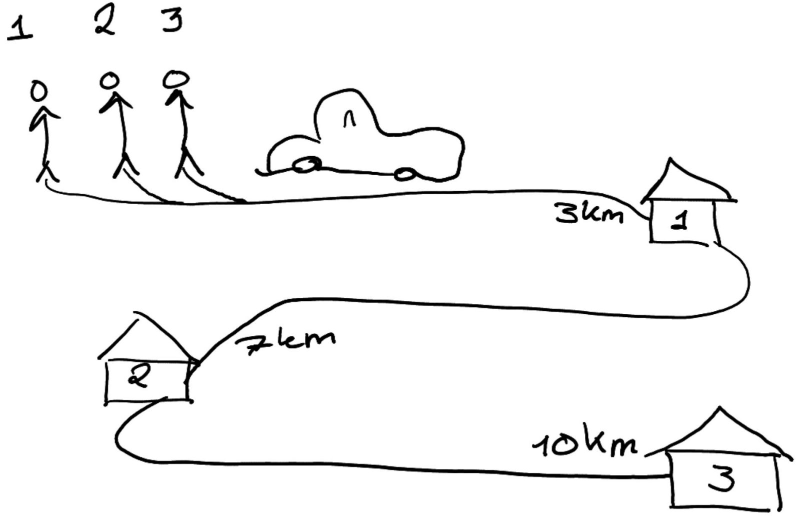
$$v(\{1,2,3\}) = 10$$

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

subsets S of N
excluding passenger i

N = total passengers
S = subsets of N

Shapley cab sharing

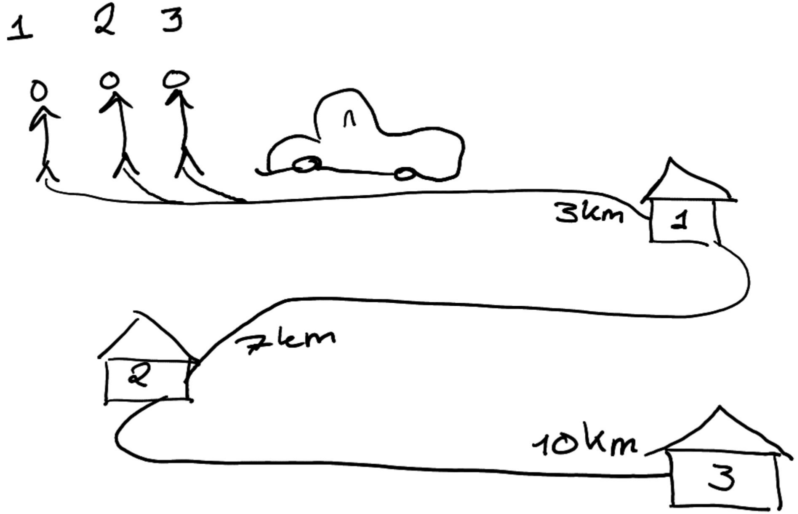


$$v(\{\}) = 0 \quad (\text{no passengers costs nothing})$$
$$v(\{1\}) = 3, \quad v(\{2\}) = 7, \quad v(\{3\}) = 10$$
$$v(\{1,2\}) = 7, \quad v(\{1,3\}) = 10, \quad v(\{2,3\}) = 10$$
$$v(\{1,2,3\}) = 10$$

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

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

Shapley cab sharing



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

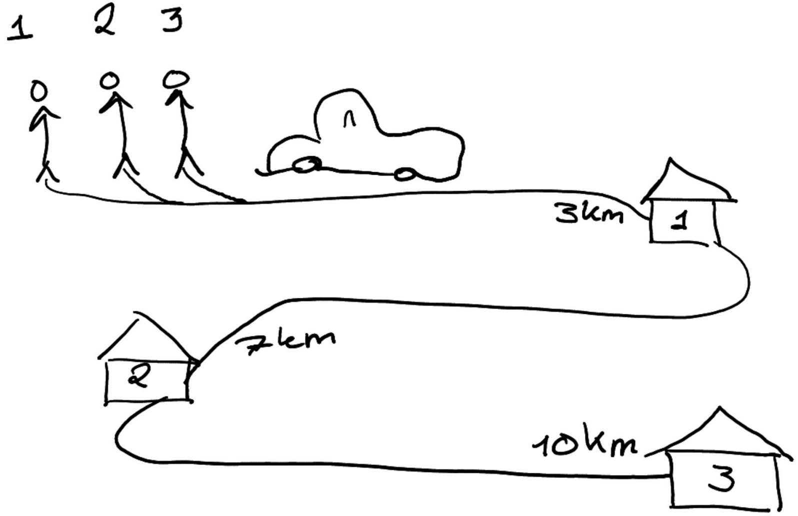
$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left(v(S \cup \{i\}) - v(S)\right), \quad i = 1, \ldots, N$$

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

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

Inga Strümke - inga.strumke@ntnu.no

Shapley cab sharing



True story: The fair and unique way to distribute the cost of the journey (at a price of 1NOK per km), is when
passenger 1 pays 1,
passenger 2 pays 3 and
passenger 3 pays 6.

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

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

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

Shapley values for machine learning...

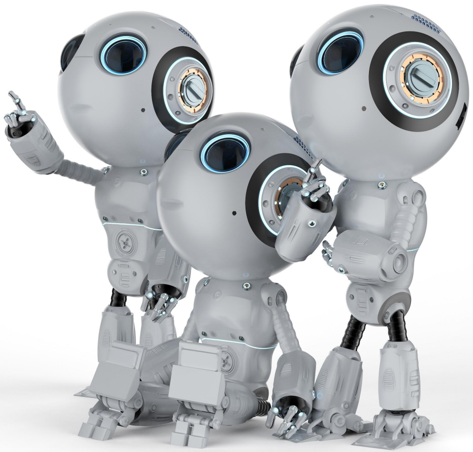
Shapley values do dependence attribution. From game theory to ML

N all players --> features

i player --> feature

S coalition of players --> set of features

v game --> model



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 $v(N)$

Shapley values for machine learning…

Various libraries: SHAP, SAGE, …

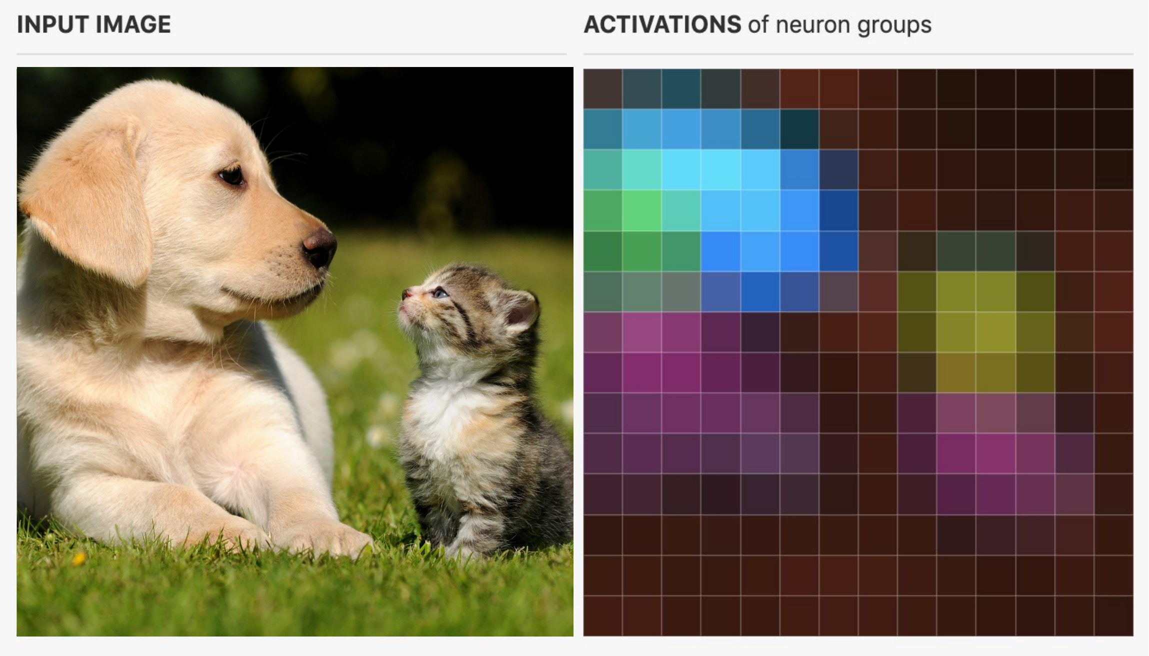
Model prediction drivers *Model loss drivers*

Location	Trysil
Sqm	130 m²
Elevation	898 m
Built	1976
Distance to ski	6.2 km
Distance to road	18 meters
Distance to sea	175 km
Distance to lake	10 km
Neighbours	776 w/1000 m
Near neighbour	31 m
Population	76 w/5 km

PwC - Responsible AI - Inga Strümke

Intrinsic explanations

What does the model focus on?

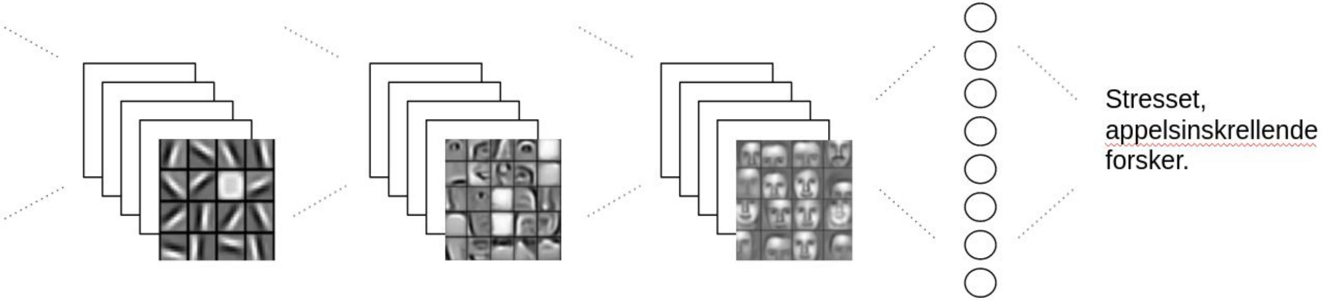


Intrinsic explanations

What do the different parts of the model focus on?

=

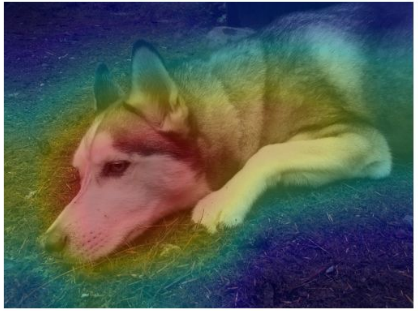
(178,99,42)	(146,66,136)	(104,119,176)	(86,198,33)
(129,68,48)	(222,55,137)	(137,9,190)	(206,84,95)
(140,23,141)	(159,167,168)	(48,86,97)	(21,113,134)
(84,98,176)	(109,31,27)	(195,53,65)	(126,101,215)
(140,23,197)	(216,99,32)	(75,48,129)	(77,96,18)
(86,215,221)	(97,190,33)	(175,34,92)	(124,61,78)
(191,206,119)	(107,101,71)	(74,13,220)	(208,23,162)
(130,216,182)	(112,90,6)	(86,179,164)	(19,146,140)
(214,208,178)	(158,101,90)	(154,172,195)	(168,104,143)
(14,117,66)	(19,192,89)	(188,220,162)	(210,177,182)
(137,47,107)	(95,1,55)	(113,175,10)	(189,94,119)
(48,70,191)	(214,158,121)	(37,86,2)	(79,84,164)
(188,204,30)	(122,77,107)	(185,168,170)	(44,181,138)
(178,209,31)	(132,208,40)	(168,138,78)	(73,29,65)
(140,96,139)	(110,189,153)	(20,7,99)	(131,171,63)
(89,43,177)	(51,75,91)	(100,130,125)	(118,44,173)
(7,155,63)	(87,129,105)	(179,82,94)	(123,12,125)
(199,76,77)	(110,146,63)	(109,223,120)	(197,115,224)
(163,112,99)	(53,25,66)	(103,78,96)	(52,132,153)
(160,29,79)	(105,117,134)	(220,183,98)	(47,174,152)



Intrinsic explanations

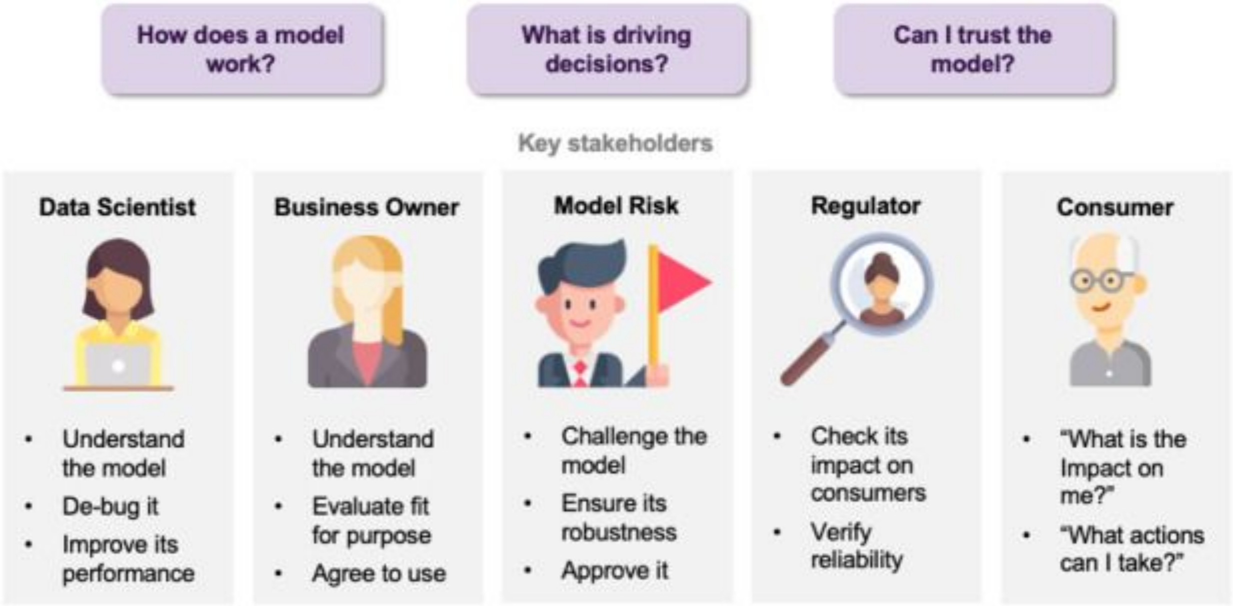
A convincing story and the truth are not necessarily the same thing

(this holds for all aspects in life)

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead: https://arxiv.org/pdf/1811.10154.pdf

The story these days



… can you spot what's missing?

Principles and Practice of Explainable Machine Learning
Vaishak Belle and Ioannis Papantonis

Part 3: Ethics

I cannot stress enough how important ethical development is in AI.

Think

$$\text{AI impact} \propto \text{medicine} + \text{atomic bomb}$$

Facebook language predicts depression in medical records

Johannes C. Eichstaedt[a,1,2], Robert J. Smith[b,1], Raina M. Merchant[b,c], Lyle H. Ungar[a,b], Patrick Crutchley[a,b], Daniel Preoţiuc-Pietro[a], David A. Asch[b,d], and H. Andrew Schwartz[e]

[a]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104; [b]Penn Medicine Center for Digital Health, University of Pennsylvania, Philadelphia, PA 19104; [c]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, PA 19104; [d]The Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center, Philadelphia, PA 19104; and [e]Computer Science Department, Stony Brook University, Stony Brook, NY 11794

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved September 11, 2018 (received for review February 26, 2018)

Depression, the most prevalent mental illness, is underdiagnosed and undertreated, highlighting the need to extend the scope of current screening methods. Here, we use language from Facebook posts of consenting individuals to predict depression recorded in electronic medical records. We accessed the history of Facebook statuses posted by 683 patients visiting a large urban academic emergency department, 114 of whom had a diagnosis of depression in their medical records. Using only the language preceding their first documentation of a diagnosis of depression, we could identify depressed patients with fair accuracy [area under the curve (AUC) = 0.69], approximately matching the accuracy of screening surveys benchmarked against medical records. Restricting Facebook data to only the 6 months immediately preceding the first documented diagnosis of depression yielded a higher prediction accuracy (AUC = 0.72) for those users who had sufficient Facebook data. Significant prediction of future depression status was possible as far as 3 months before its first documentation. We found that language predictors of depression include emotional (sadness), interpersonal (loneliness, hostility), and cognitive (preoccupation with the self, rumination) processes. Unobtrusive depression assessment through social media of consenting individuals may become feasible as a scalable complement to existing screening and monitoring procedures.

the diagnosis of depression, which prior research has shown is feasible with moderate accuracy (15). Of the patients enrolled in the study, 114 had a diagnosis of depression in their medical records. For these patients, we determined the date at which the first documentation of a diagnosis of depression was recorded in the EMR of the hospital system. We analyzed the Facebook data generated by each user before this date. We sought to simulate a realistic screening scenario, and so, for each of these 114 patients, we identified 5 random control patients without a diagnosis of depression in the EMR, examining only the Facebook data they created before the corresponding depressed patient's first date of a recorded diagnosis of depression. This allowed us to compare depressed and control patients' data across the same time span and to model the prevalence of depression in the larger population (~16.7%).

Results

Prediction of Depression. To predict the future diagnosis of depression in the medical record, we built a prediction model by using the textual content of the Facebook posts, post length, frequency of posting, temporal posting patterns, and demographics (*Materials and Methods*). We then evaluated the performance of this model by comparing the probability of depression estimated by our algorithm

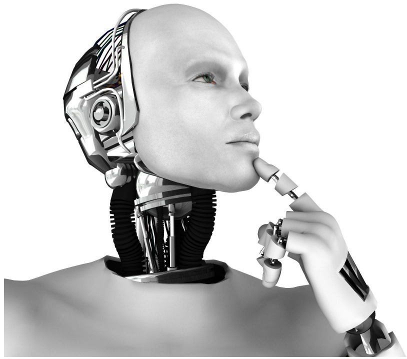
The full picture?

We don't even know what we expect from an explanation

Explanation for human-beings:

Not just the model but its impact

(depends on environment and context)



Super secret figure removed after presentation :)

Thank you!

inga.strumke@ntnu.no

Discussion points:

1. I said that the discussion on XAI is about machine learning. Does it have to be?
    a. Can non-learning approaches to AI require explanations?
2. What are heat maps?
    a. What are their weaknesses?
    b. Do they match the way humans would explain a visual decision?
3. What is a feature importance ranking?
    a. How could a bank use feature ranking to tell you why you got a loan?
    b. Would you accept such an explanation, and why (not)?
    c. What weaknesses does this have when used as an explanation?