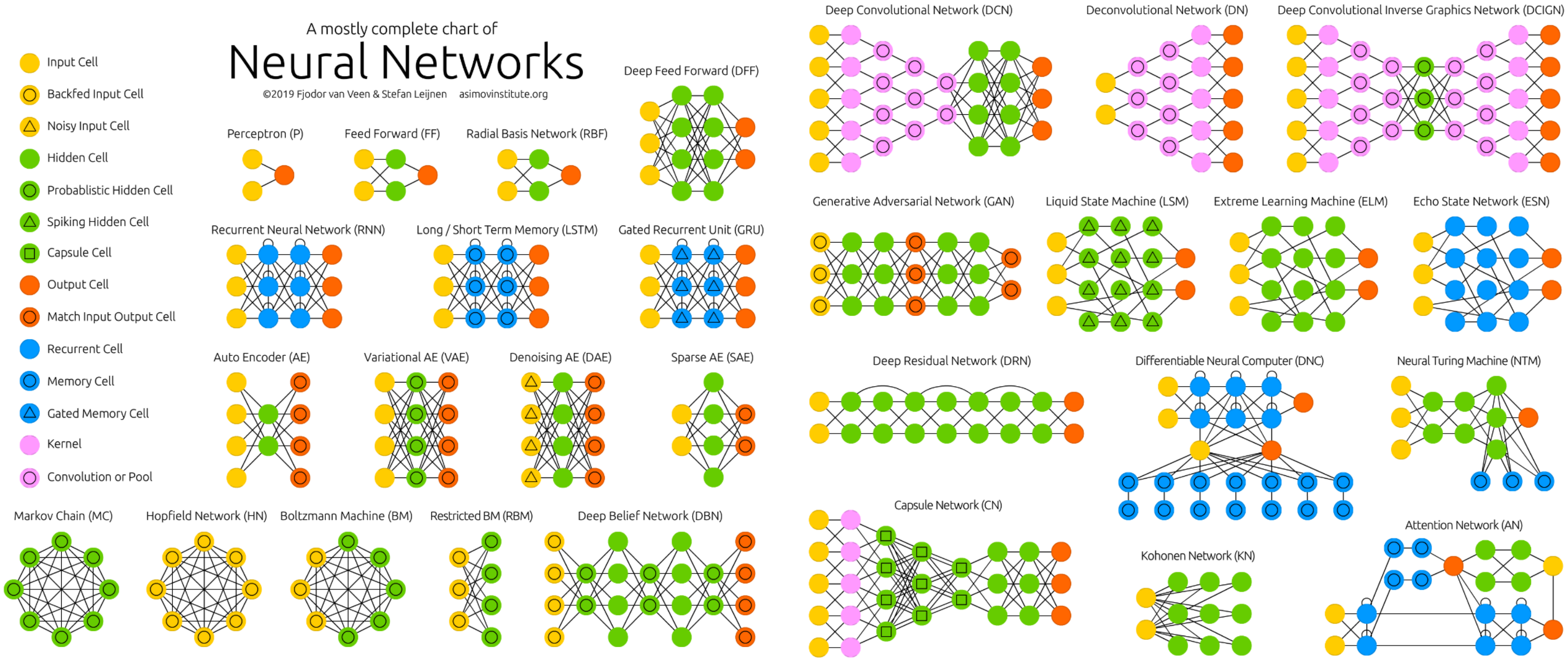
school of ai

EXPLAINABLE AI

Real applications, Trust issues, Opening up the black box



A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

"As far as we know, there is no data-set or network that is much more robust than others" Su Jiawei

There is no AI fully explainable.

- Banking
- Insurance
- Healthcare
- other industries…

GDPR

- Transparency in collecting data
- Transparency in how decisions are made
- Transparency in how sure the results are
- Reliability of the results (not the system performance)
- Securing the privacy

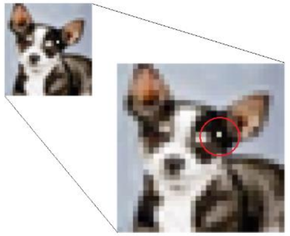
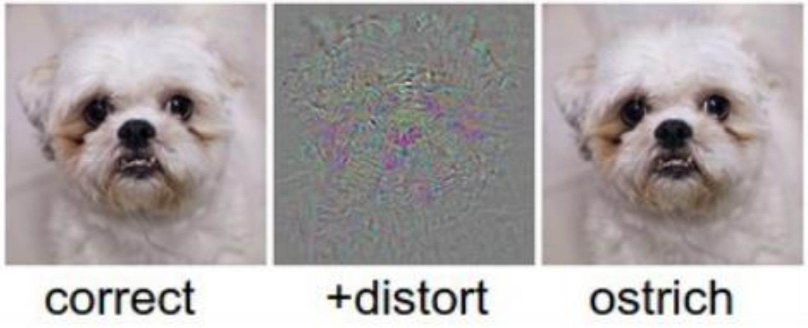
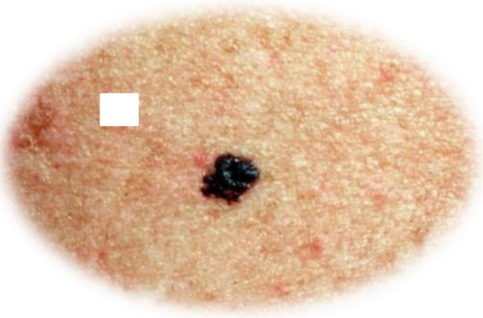


Figure.1. Adversarial perturbation:

A dog image from cifar-10 data set, can be misclassified as a cat by merely modiying one pixel.

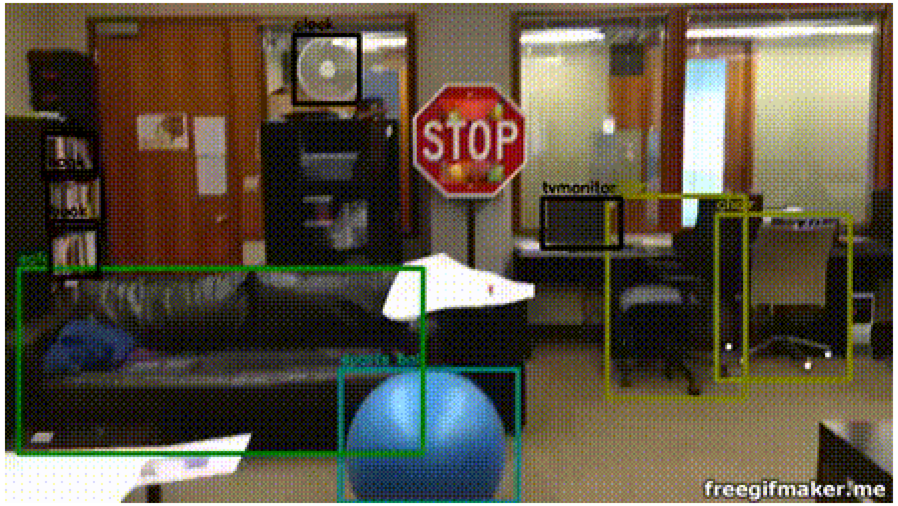




Different decision

Distinguish between elephant and cat:
Trunk, grey, tusk → elephant

A Survey on Explainable Artificial Intelligence (XAI):
Towards Medical XAI, (2019) Erico Tjoa, Cuntai Guan



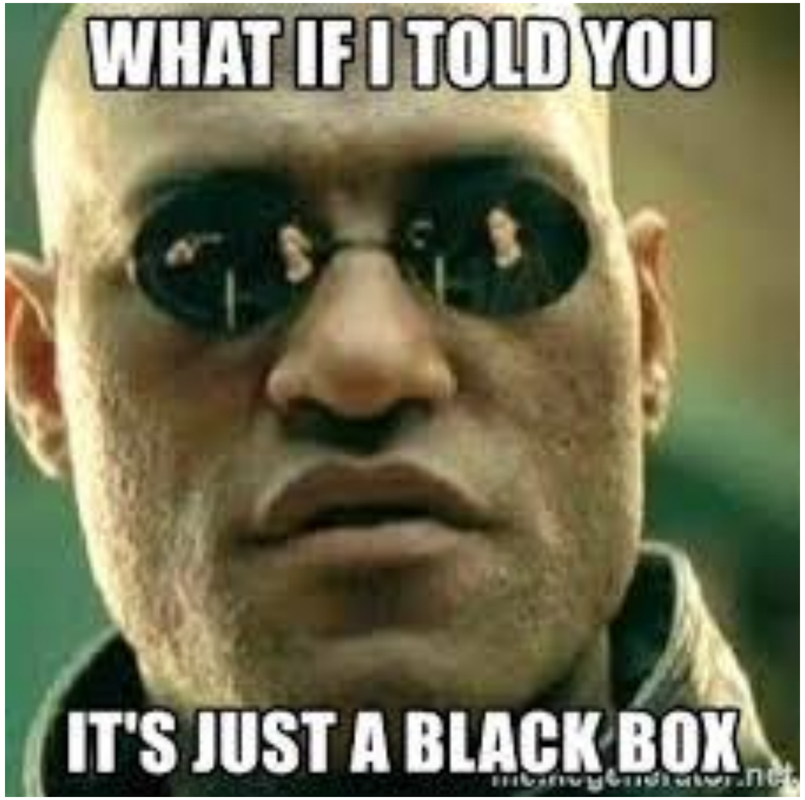
https://arxiv.org/abs/1712.08062

Note on Attacking Object Detectors with Adversarial Stickers (2017)
Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, Florian Tramer

CAN WE BUILD TRUST BASED ON THE ACCURACY?


TRUST YOUR GUT
AND LET FATE BE YOUR GUIDE
quickmeme.com

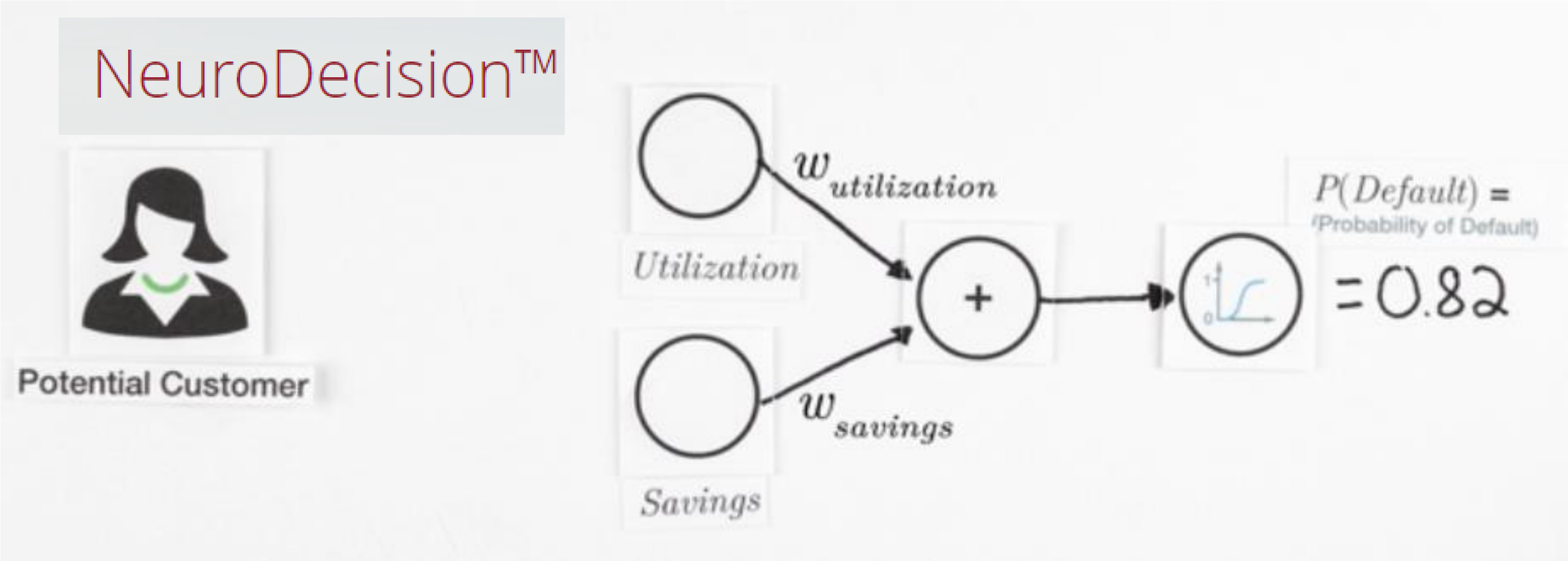
DATA ALONE IS NOT ENOUGH

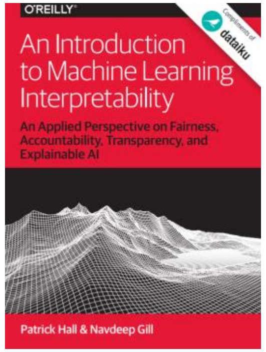

WHAT IF I TOLD YOU
IT'S JUST A BLACK BOX

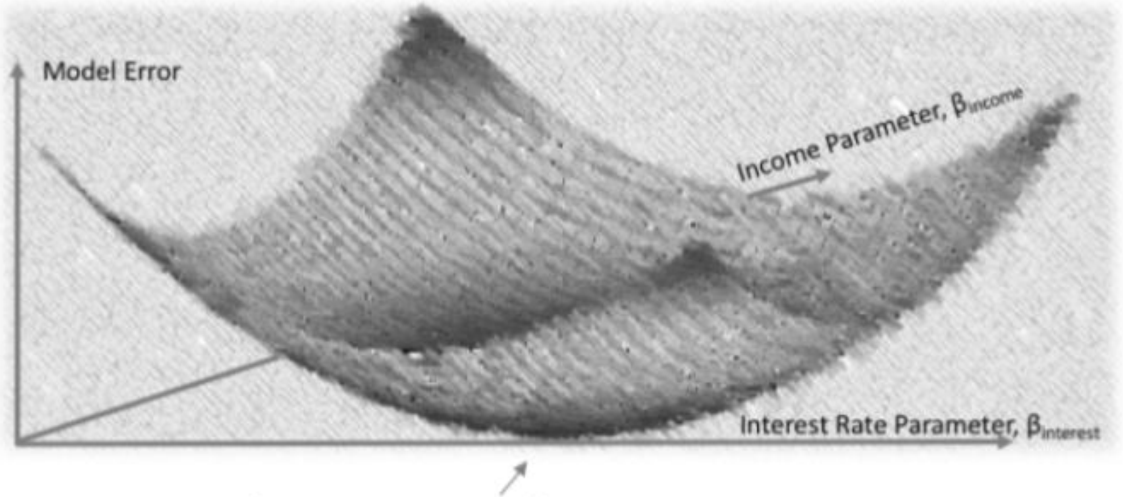
	INTERPRETABLE	ACCURATE
COMPLEX MODEL	✗	✓
SIMPLE MODEL	✓	✗

HOW COMPLEX IS THE SYSTEM?

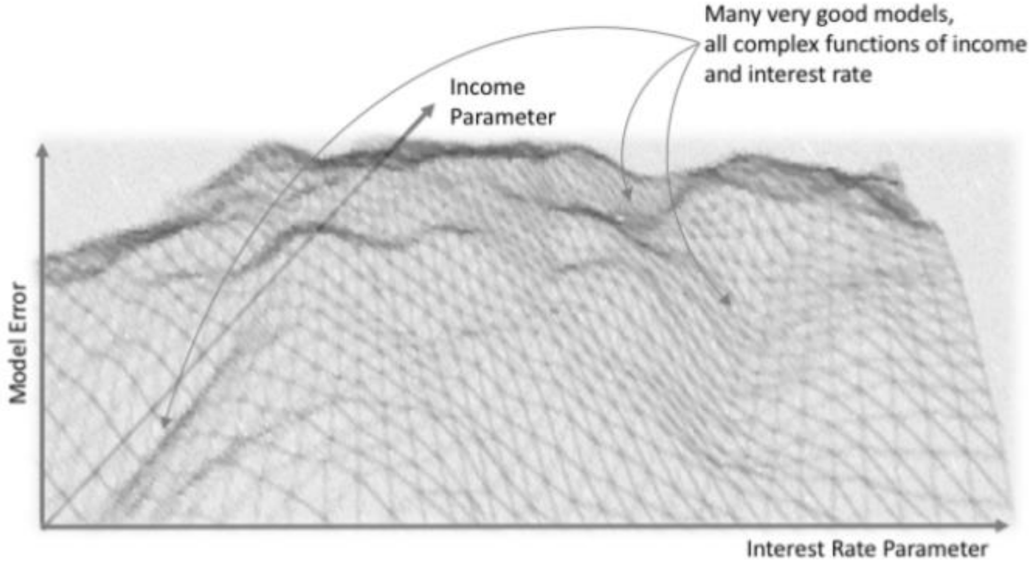
- Number of the neurons
- Vapnik–Chervonenkis (VC) dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a statistical classification algorithm.


NeuroDecision™
Potential Customer
Utilization
Savings
$w_{utilization}$
$w_{savings}$
+
$P(Default)$ =
(Probability of Default)
=0.82





One best model: $f(Income, Interest\ Rate) \sim \beta_{income} * Income + \beta_{interest} * Interest\ Rate$



Academics FAT* aca- demics (meaning fairness, accountability, and transparency in multi- ple artificial intelligence, machine learning, computer science, legal, social science, and policy applications)

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Defense Advanced Research Projects Agency (DARPA).

Military researchers

XAI

INTERPRETABILITY

Loosely defined but…

- Directly transparent “white-box” models
- Explanation of “black-box” models to enhance transparency
- Debugging models to increase trust
- Ensuring fairness in algorithmic decision-making
- Model documentation

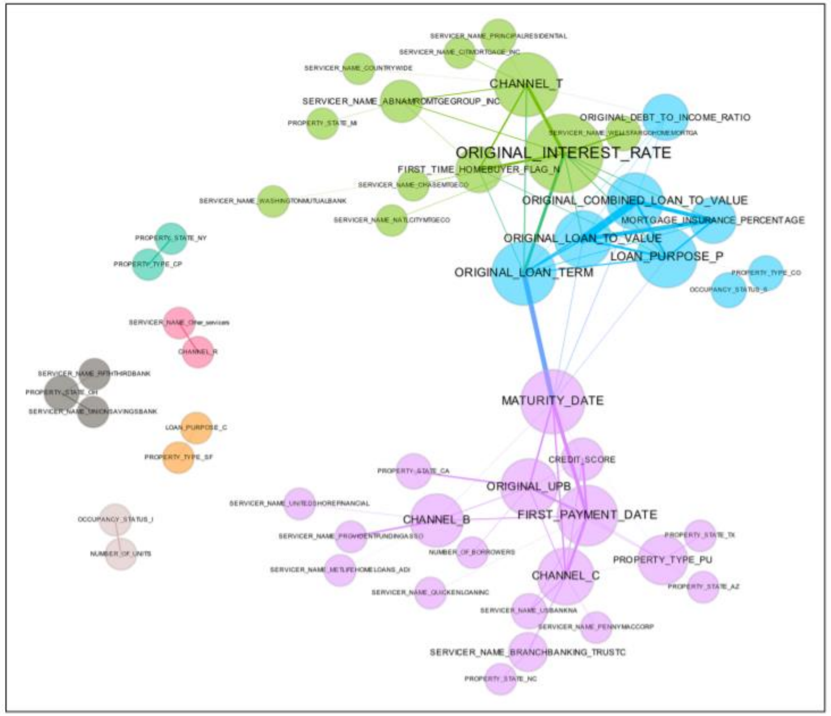
COMMON INTERPRETABILITY METHODS

COMMON INTERPRETABILITY METHODS

1. Seeing and understanding the data

COMMON INTERPRETABILITY METHODS

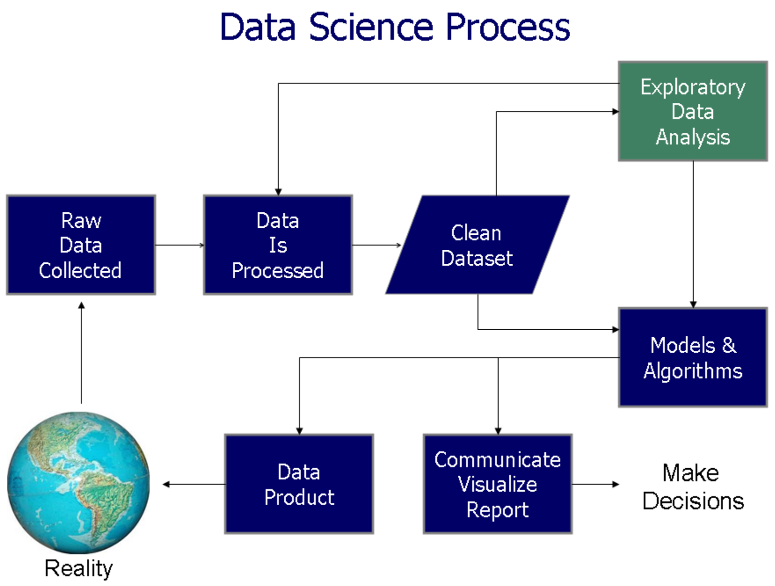
1. Seeing and understanding the data (feature extraction, dimension reduction, ploting projection on a dimension, graph visualization, creating decision trees)



H2o.ai

COMMON INTERPRETABILITY METHODS

1. Seeing and understanding the data (feature extraction, dimension reduction, ploting projection on a dimension, graph visualization, creating decision trees)



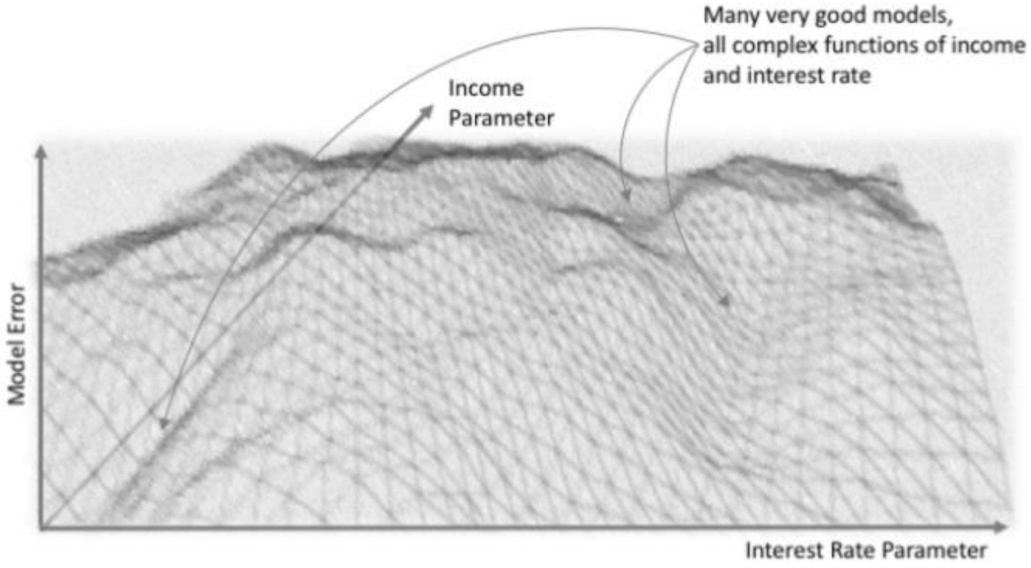
Wikipedia

COMMON INTERPRETABILITY METHODS

2. Start with old ML techniques

COMMON INTERPRETABILITY METHODS

3. Look for the global explanations



COMMON INTERPRETABILITY METHODS

4. Find out which variables are important

…and simplify the structure

(PCA, autoencoders or other dimension reduction methods)

COMMON INTERPRETABILITY METHODS

5. Provide reasoning as an output

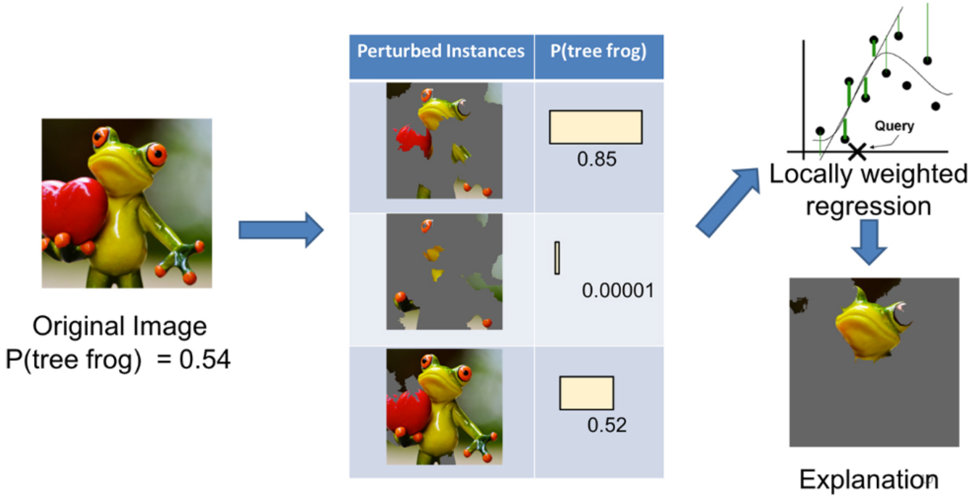
…try to reflect in the output parameters why this decision was given

COMMON INTERPRETABILITY METHODS

6. Look at local reasoning

LIME

- **Locally**: Every complex model is linear on a local scale
- **Interpretable**: Representation that can be interpreted by humans
- **Model-agnostic**: Applied to any black box machine learning model
- **Explanations**: A statement that explains individual predictions.



SENSITIVITY ANALYSIS: TESTING MODELS FOR STABILITY AND TRUSTWORTHINESS

- simply try a random data attacks!

AUTOMATED TESTING OF INTERPRETABILITY

- Generate data with simulations
- Show accuracy increase with learning from more and better examples
- Show accuracy with respect to the random attacks
- Show that different parameter setup for the architecture reduces the accuracy

FAIRNESS


001010101101010
0101010100001010101
0101010101010101010101
00001010101010101000010101

MODEL DOCUMENTATION

Model documentation is required in some industries but represents a best practice for all. Documentation should include essential information about machine learning models including:

- The creation date and creator of the model
- The model's intended business purpose
- A description of the input dataset
- Description of the algorithm(s) used for: Data preparation and model training
- Final model tuning parameters
- Model validation steps
- Results from explanatory techniques
- Results from disparate impact analysis
- Results from sensitivity analysis
- Who to contact when a model causes problems
- Ideas about how to fix any potential problems

LEARNING FROM THE MODEL

"It's not a human move. I've never seen a human play this move." (Fan Hui, 2016).

MAYBE…

Non-explainable → Explained → Non-explainable again