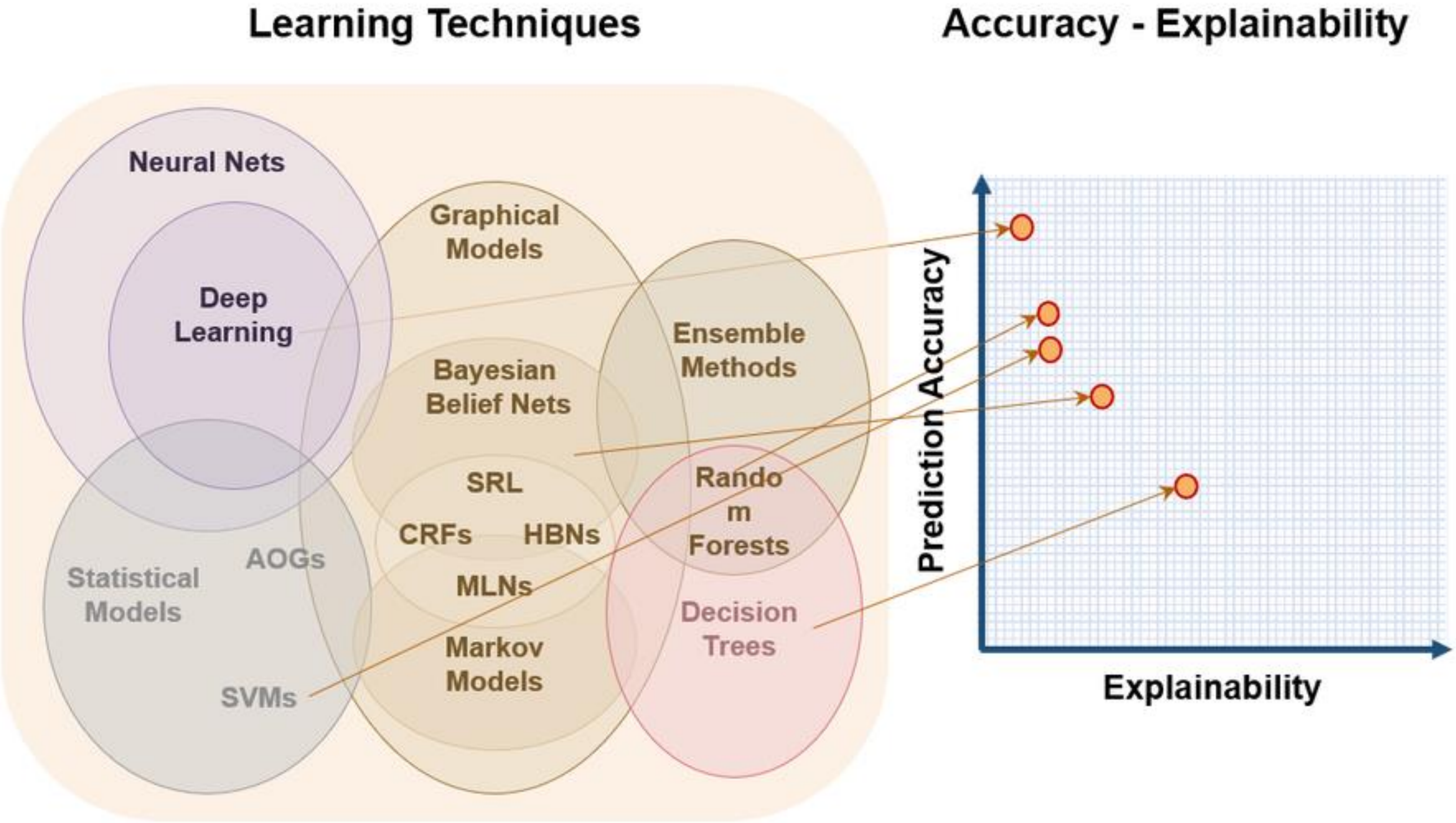
How to explain predictions of your network?

Vladyslav Kolbasin

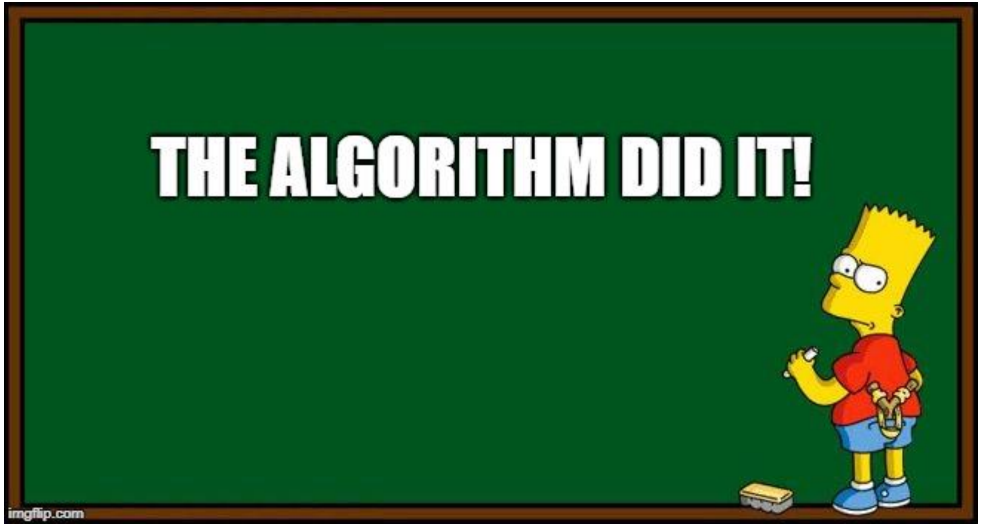
Lead Data Scientist at Globallogic
Lecturer at CMAD dep. at NTU “KhPI”

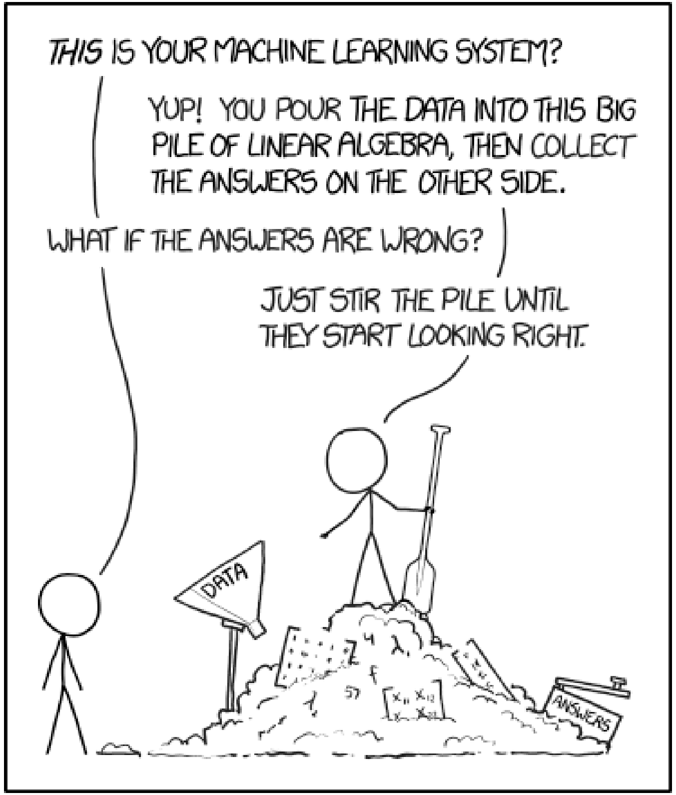
Accuracy vs Interpretability?



High Accuracy Result

Improve model

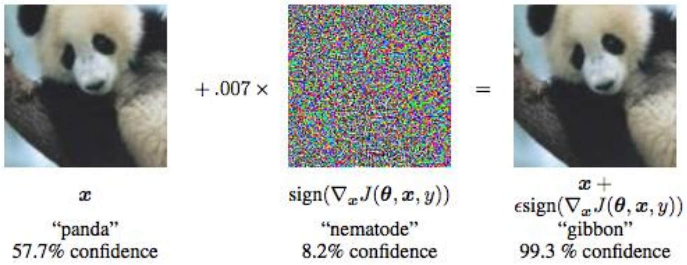


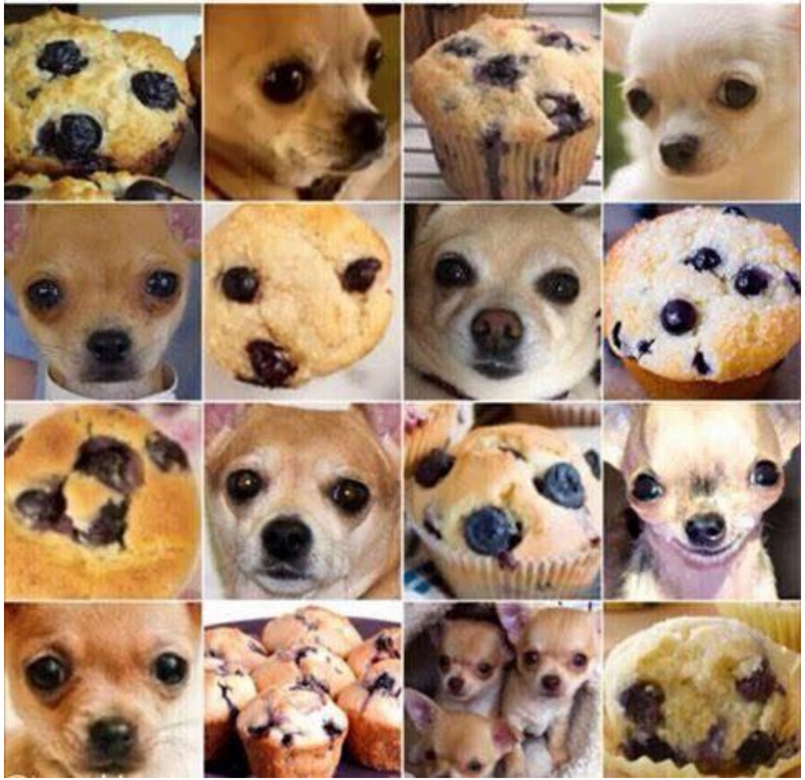


Interpretability aspects. Pragmatic

Ability to explain

- Can we trust your model?
  a. Feature analysis
  b. Model validation
- Debug model
  a. How to improve my model?
  b. Adversarial examples
- Model discovery



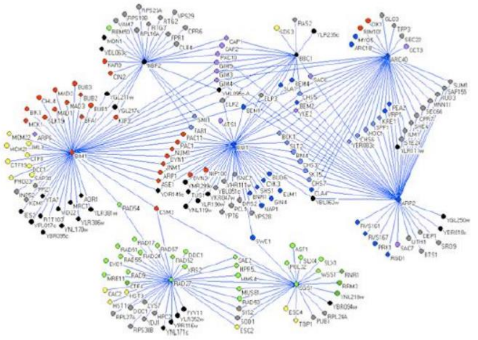


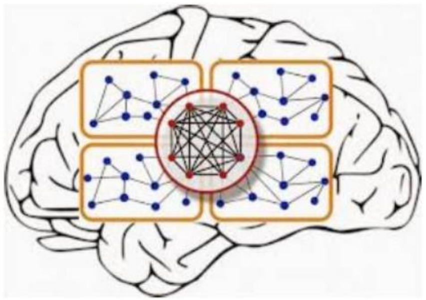
Interpretability aspects. Pragmatic

Model discovery

- Learn from ML
- Learn more in science
- Data insights

“It's not a human move. I've never seen a human play this move.” (Fan Hui)

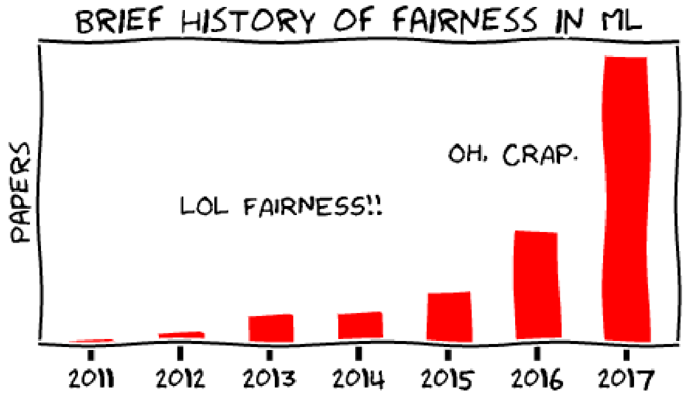




Interpretability aspects.
Philosophical, Political & Social

Right to explain

- FATE in AI:
  Fairness, Accountability
  Transparency, Ethics
- Regulatory examples:
  - Civil Rights Acts
  - Americans with Disabilities Act
  - Genetic Information Nondiscrimination Act
  - Equal Credit Opportunity Act
  - Fair Credit Reporting Act
  - Fair Housing Act
  - European Union GDPR




How do we build ML models?

~~How do we build ML models?~~

How can we build ML interpretable models?

Which models are interpretable?

- Linear

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

Which models are interpretable?

- Linear

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

Is it interpretable when we have 1000 variables?

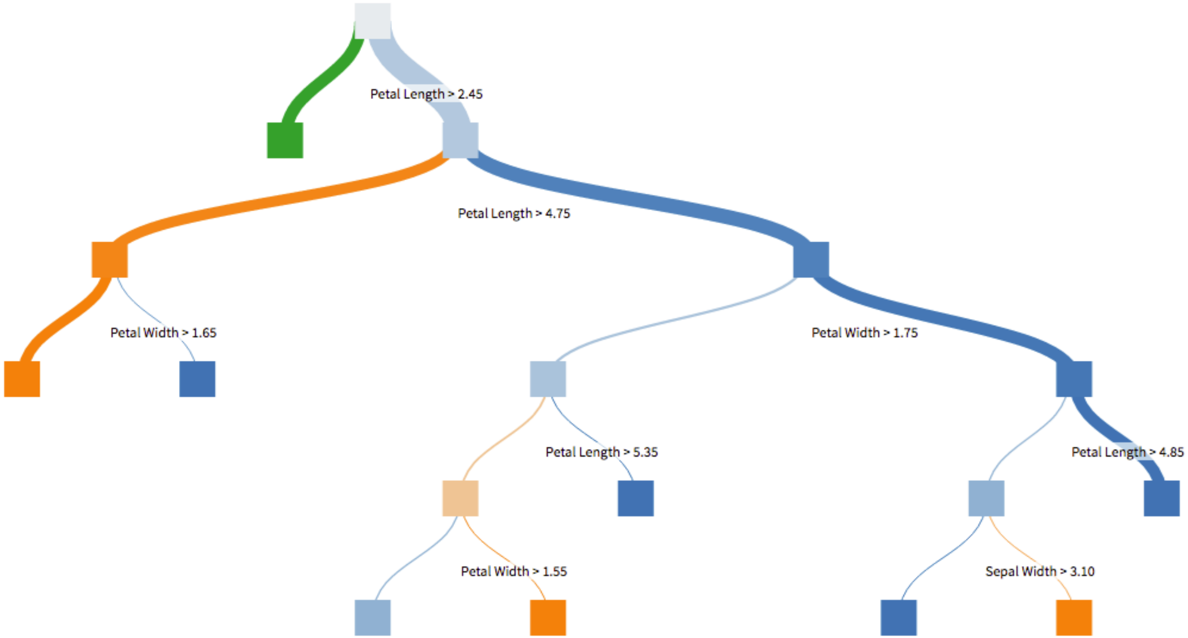
Maybe not!

We need to make it sparse…

Will it be interpretable?...

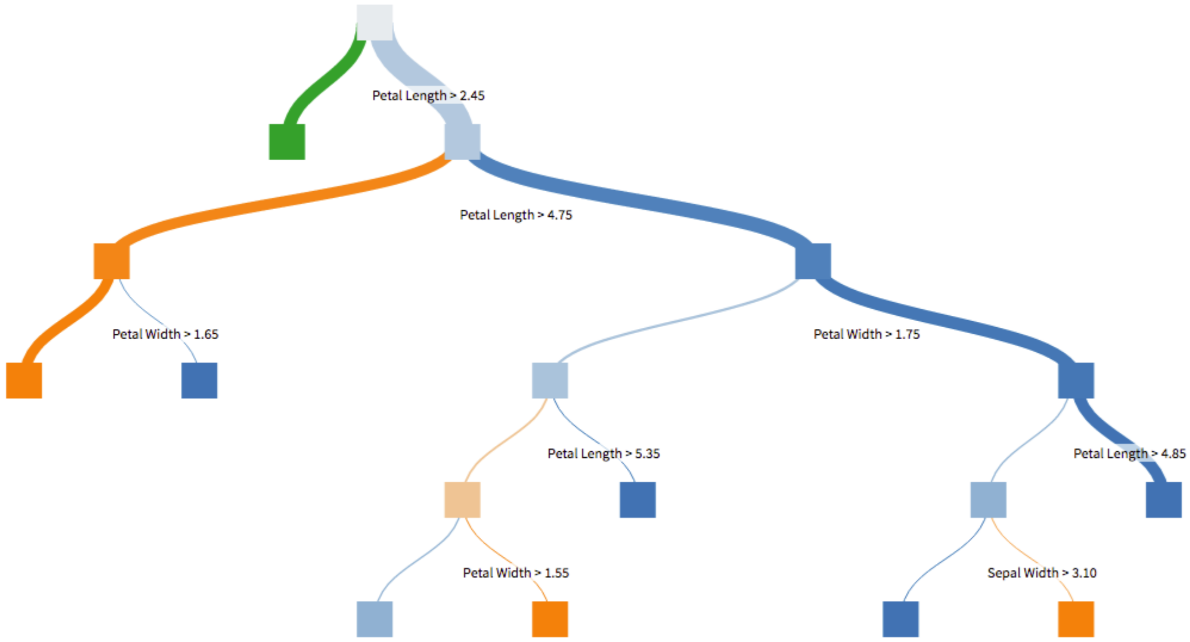
Which models are interpretable?

- Linear
- Decision trees



Which models are interpretable?

- Linear
- Decision trees



How deep tree can you interpret?

Which models are interpretable?

- Linear
- Decision trees
- Some nonlinear models

$$y = ae^{-bx} \qquad y = \frac{1}{1 + e^{-x}}$$

Which models are interpretable?

- Linear
- Decision trees
- Some nonlinear models
- Generalized additive models

$$g(y) = f_1(x_1) + f_2(x_2) + ... + f_n(x_n)$$

Which models are interpretable?

- Linear
- Decision trees
- Some nonlinear models
- Generalized additive models

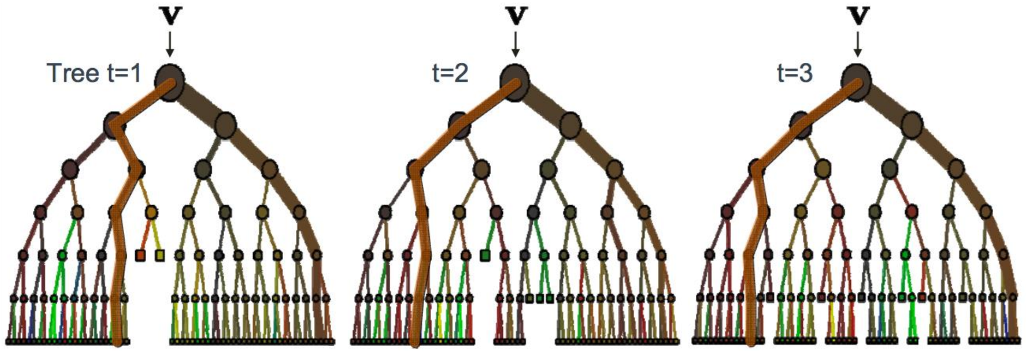
$$\hat{\log}(y) = 9.76 + 0.0063\text{RM}^2 + 8.98 \times 10^{-5}\text{AGE} - 0.19\log(\text{DIS}) + 0.096\log(\text{RAD}) - \ldots$$

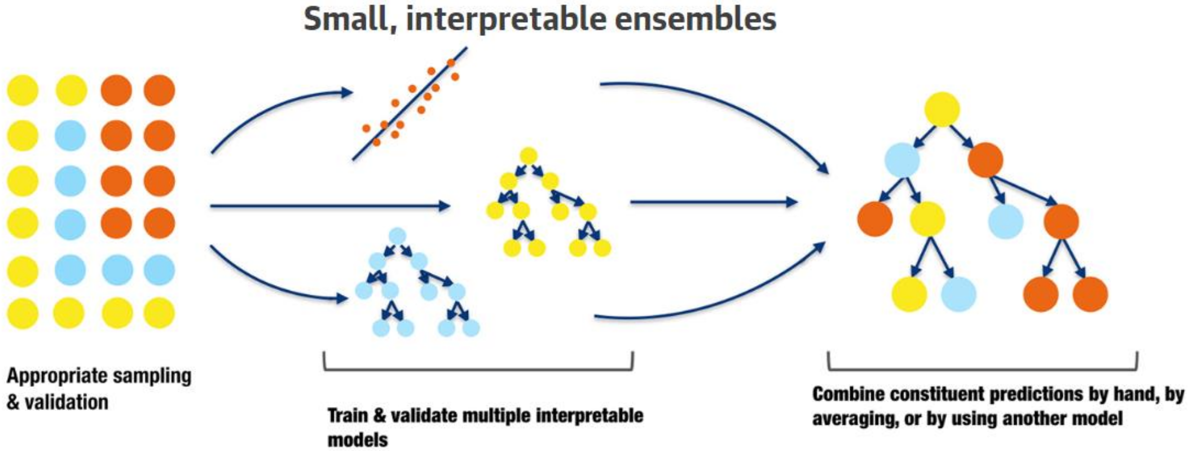
$$4.20 \times 10^{-4}\text{TAX} - 0.031\text{PTRATIO} + 0.36(\text{B} - 0.63)^2 - 0.37\log(\text{LSTAT}) - \ldots$$

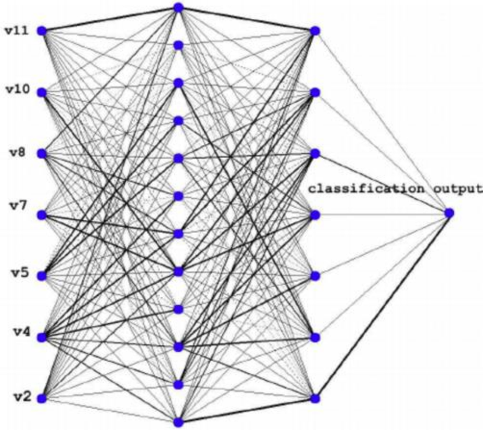
$$0.012\text{CRIM} + 8.03 \times 10^{-5}\text{ZN} + 2.41 \times 10^{-4}\text{INDUS} + 0.088\text{CHAS} - 0.0064\text{NOX}^2$$

Which models are interpretable?

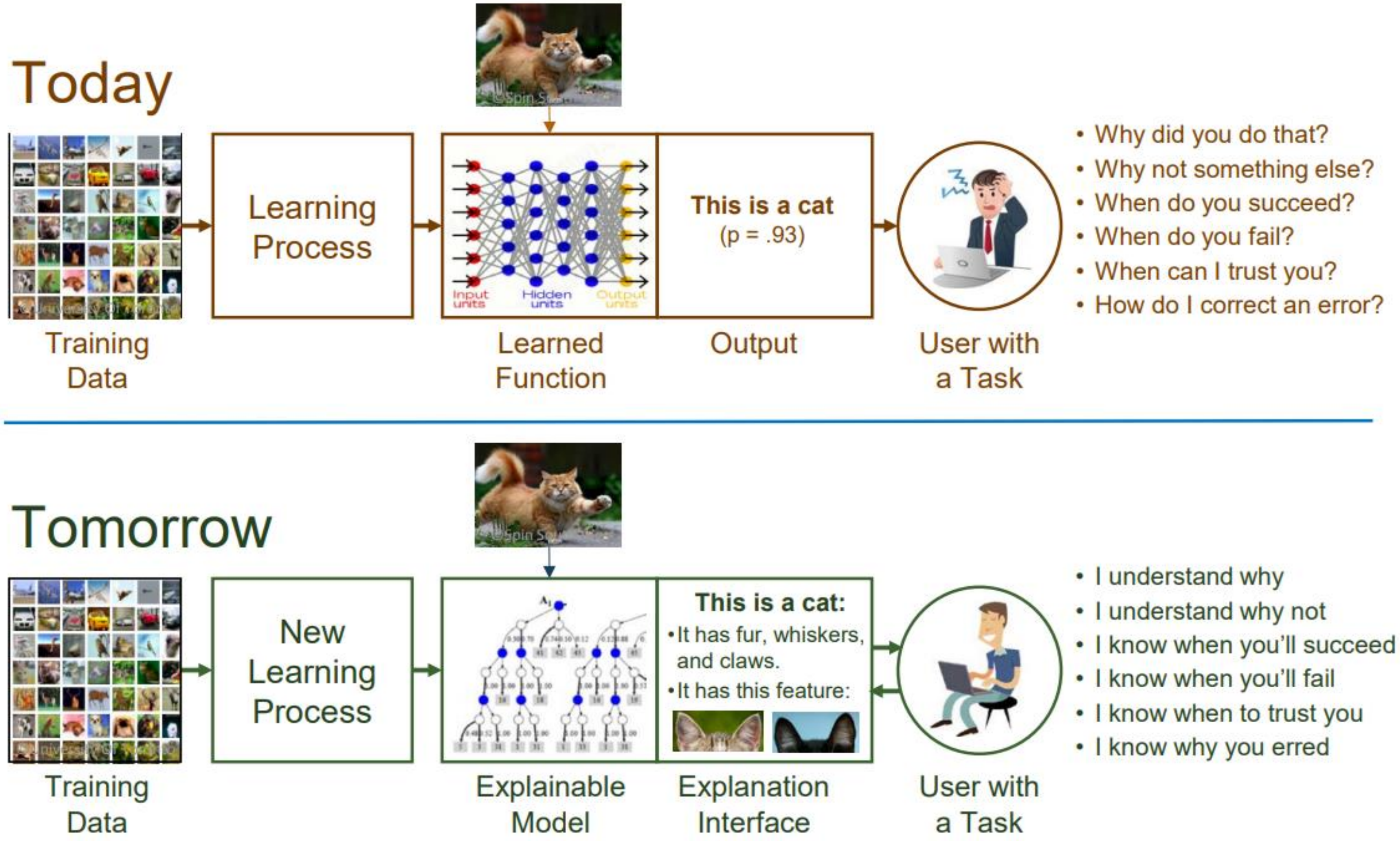
- Simple ensembles?
- RandomForest?
- Perceptron?







eXplainable Artificial Intelligence (XAI)



(C) XAI DARPA

What does ML model interpretation mean to you?

What is interpretation?

- Constraints understanding
- Algorithm understanding
- Each step argumentation
- Testing
- Behavior in nonstandard cases

What is interpretation?

Interpretation is the process of giving **explanations to <u>Humans</u>**

a. Interpretability is **NOT** about understanding all bits and bytes of the model for all data points (we cannot).
b. It's about knowing enough for your downstream tasks.

Read more:
https://christophm.github.io/interpretable-ml-book/explanation.html

Dimensions of Interpretability

Scope: Global vs Local

Type of technique: Model agnostic vs Model dependent

When to do it: Before, During or After model creation

Approach: Supervised vs Unsupervised vs RL

Explanation recipient: User, Developer, Stakeholder

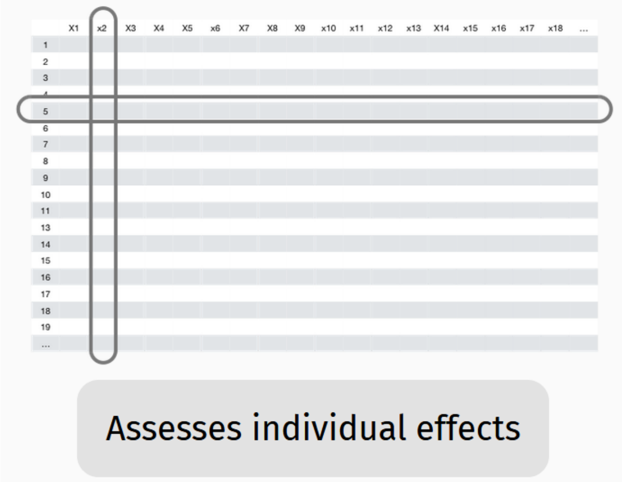
1. Scope of interpretability

Global interpretability

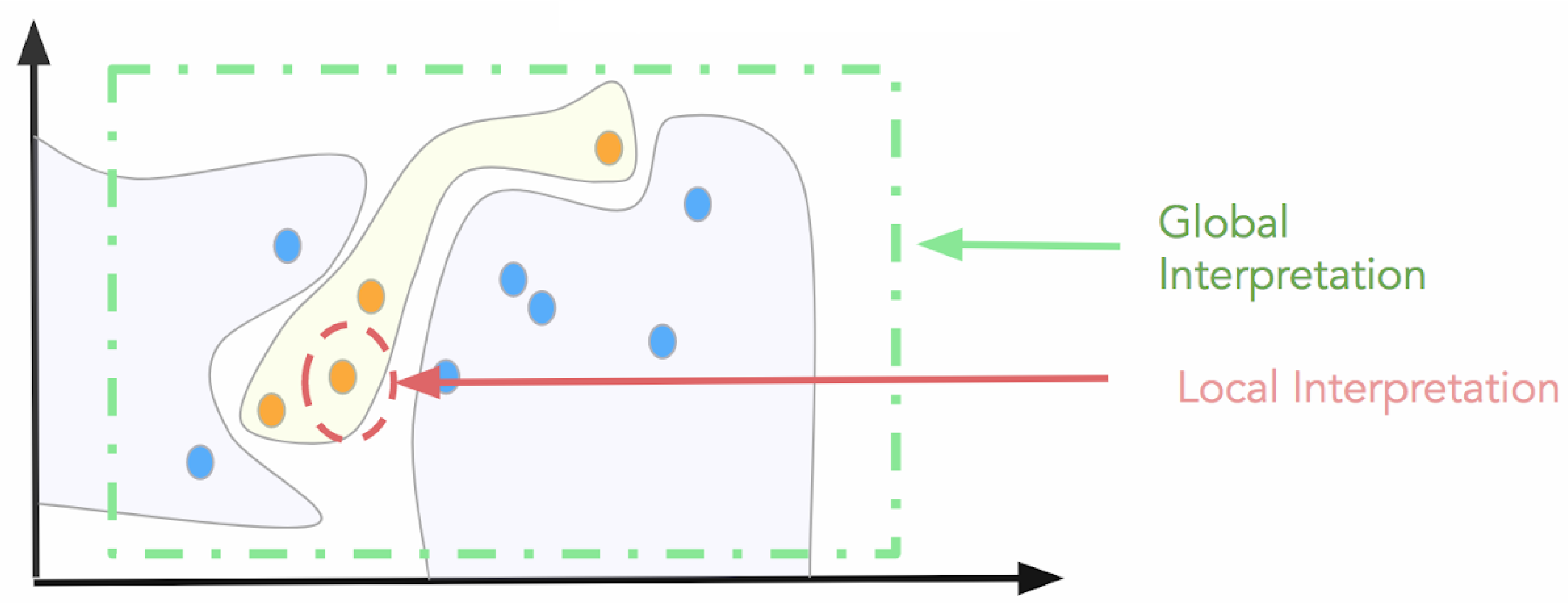
- How do features influence overall model performance?
- What is the overall relationship between features and the target?

Local interpretability

- How do our features influence individual predictions?
- What are the observation level relationships between features and the target?



1. Scope of interpretability



2. Model specific vs Model agnostic

Model specific

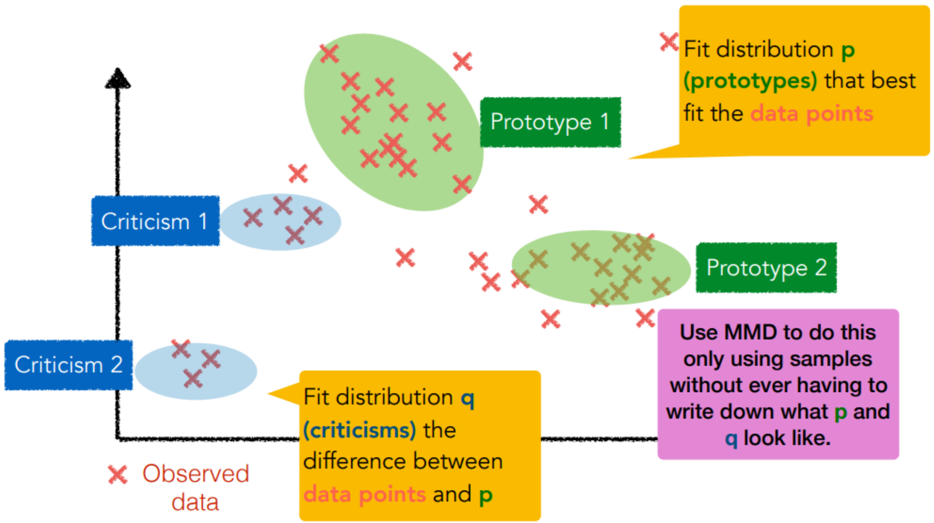
- Limited to specific ML classes
- Incorporates model-specific logic
- Examples:
  - Coefficients in linear models
  - Impurity in tree-based models
  - Attention

Model agnostic

- Can be applied to any type of ML algorithm
- Assesses inputs and outputs
- Examples:
  - Permutation-based variable importance
  - PDPs, ICE curves
  - LIME, Shapley, Breakdown

3.1. Interpretability before building a model

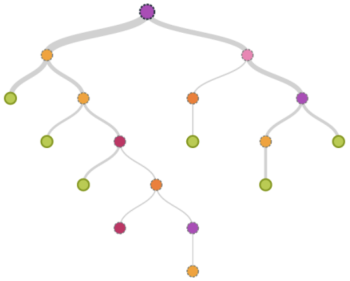
- Exploratory data analysis
- Understand distributions
- Simple feature analysis
- Clustering

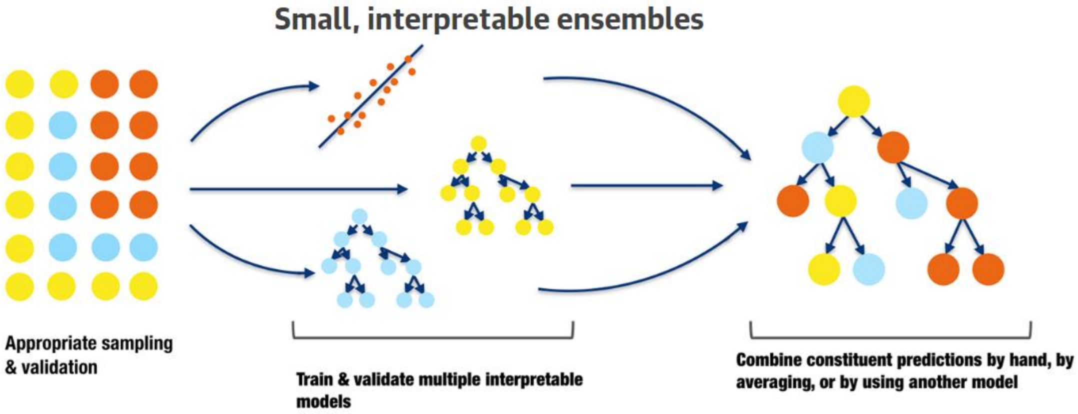


3.2. Interpretability during building a model

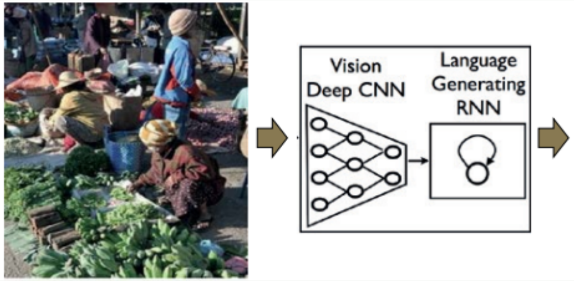
- Rule-based approaches (Decision trees)
- Linear models
- Make model sparse
- Monotonic models (monotonic constraints, e.g. in XGBoost)
- Attention in ANN

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$





Generating Image Captions

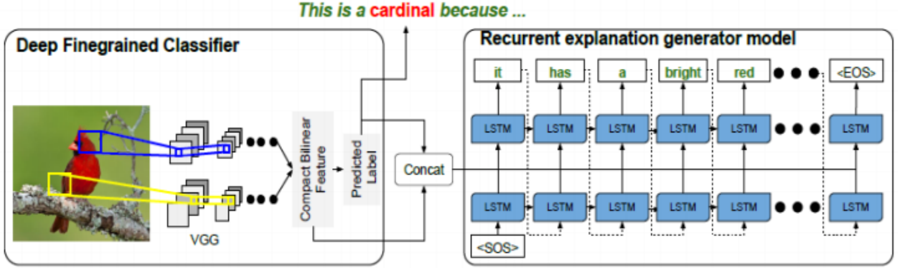


A group of people shopping at an outdoor market

There are many vegetables at the fruit stand

- A CNN is trained to recognize objects in images
- A language generating RNN is trained to translate features of the CNN into words and captions.

Generating Visual Explanations



Researchers at UC Berkeley have recently extended this idea to generate explanations of bird classifications. The system learns to:

- Classify bird species with 85% accuracy
- Associate *image descriptions* (discriminative features of the image) with *class definitions* (image-independent discriminative features of the class)

Example Explanations

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

Limitations

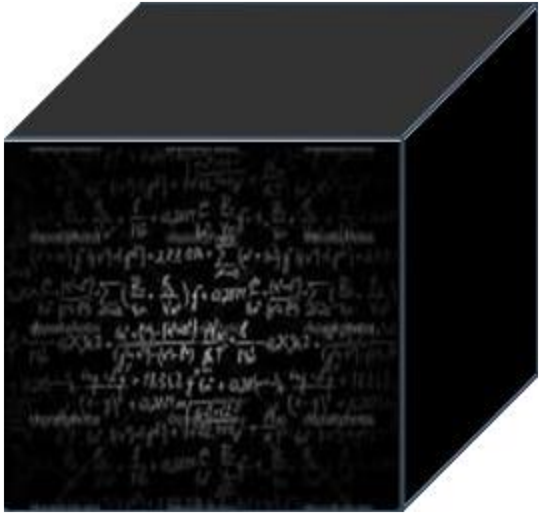
- Limited (indirect at best) explanation of internal logic
- Limited utility for understanding classification errors

Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations, arXiv:1603.08507v1 [cs.CV] 28 Mar 2016

3.3. Interpretability after building the model

Model is a **black box**: Neural networks, Support Vector Machines, Ensembles

1. Explain the **model**
2. Explain the **outcome**
3. Inspect the black box **internally**



3.3.1. Explain the model

1. Single tree approximation
   a. Prototype tree for each target class
   b. Measure tree distance, find best splits, extract tree prototypes
2. Monotonic Models (e.g. XGBoost)
3. Rule extraction for neural networks
   a. Knowledge initialization, Rule extraction, Rule refinement
   b. Dependent on the neural network
4. Agnostic explanators.

3.3.2. Explain the outcome

1. Saliency Masks
2. Sensitivity Masks
3. Conterfactual explanations
4. LIME
5. Shapley value explanations

3.3.3. Inspect the black box internally

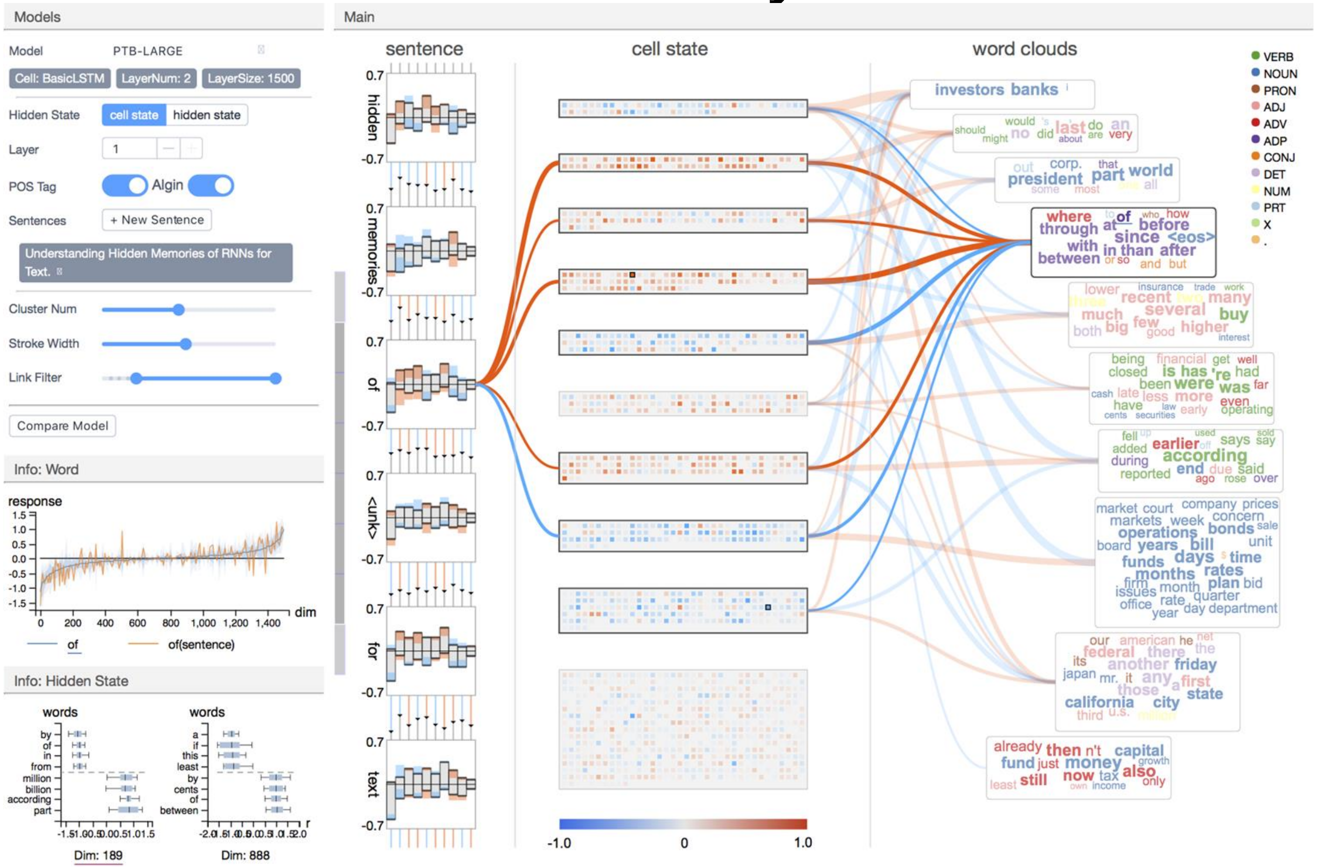
1. **Sensitivity Analysis**: Sensitivity analysis studies the correlation between the uncertainty in the output of a predictor and that one in its inputs

2. **Partial Dependence**: A partial dependence plot can show if the relationship between the target and a feature is linear, monotonic or more complex

3. Other approaches
   Lucid: GitHub Example Background

3.3.3. Inspect the black box internally

RNNVis



[Website](#) [GitHub](#)

Caveats / Discussions

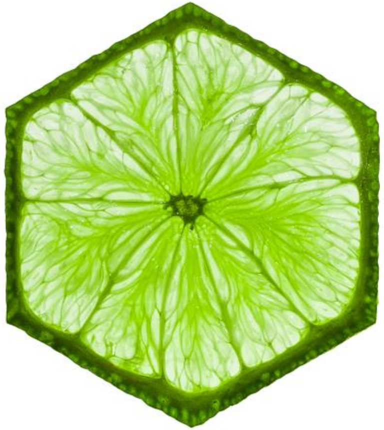
- No **formal definition** of interpretability or explanation
- No **objective measure** of how interpretable is a model
- No experiments about the **time** it takes to **understand an explanation**
- Your explanations will be as good as your **data**
- Explanations can differ according to the **purpose**
- **Conflicting** explanations?
- Can be confusing with **correlation** and **causation**
- Explanation **coverage** of a model?
- **Scalable** automatic explanations?

Practice

LIME (Local individual model-agnostic explanation)

Theory:

- LIME approximates model locally as logistic or linear model
- Repeats process many times
- Outputs features that are most important to local models



Outcome:

- Approximate reasoning
- Complex models can be interpreted

- Local
- Model agnostic
- Apply after modelling

LIME. Algorithm

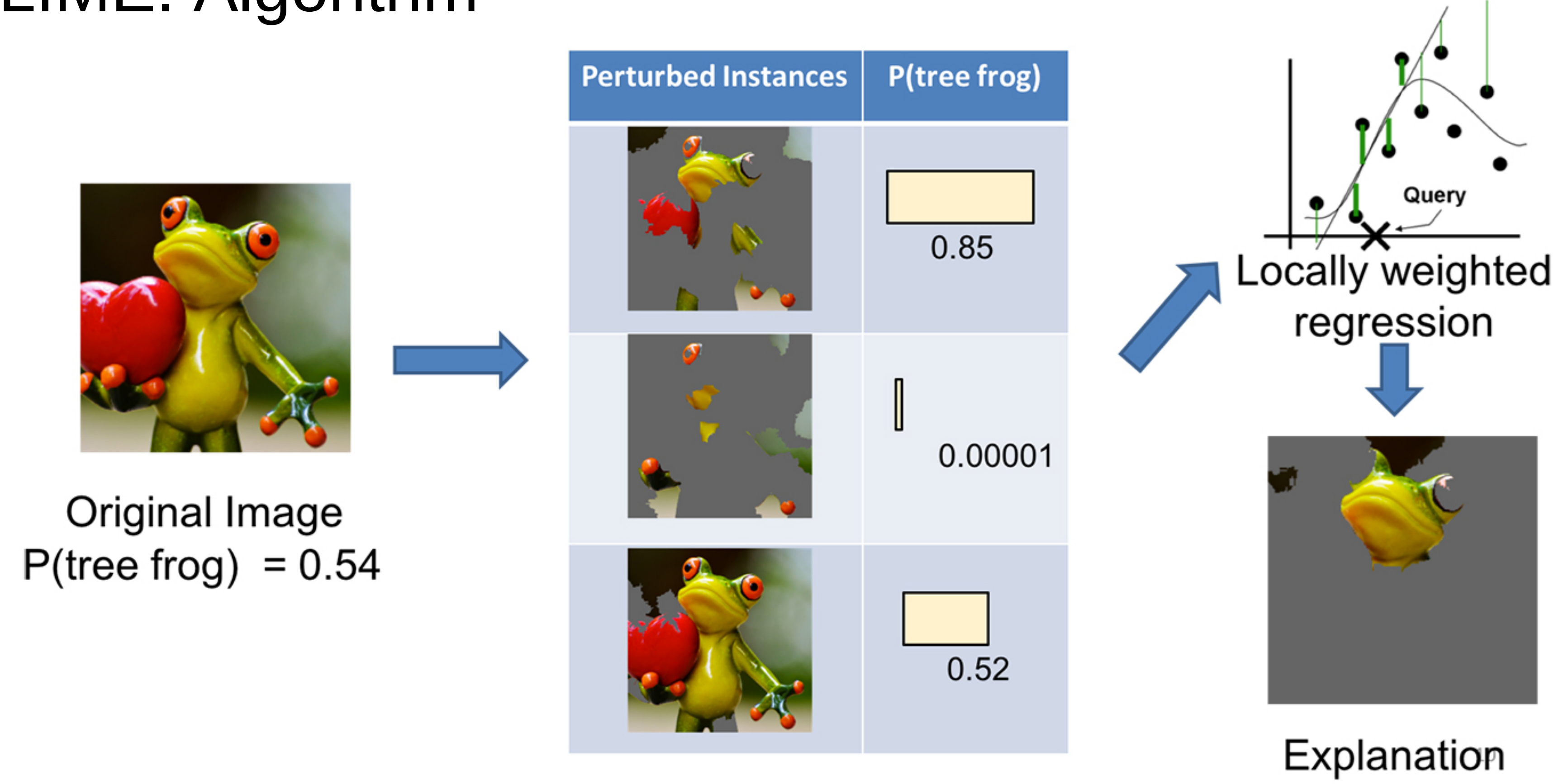
1. **Permute Data**: It take observations and create fake data for it. It permuted in different ways.
2. **Calculate distance between permutations and original observations**.
3. **Make predictions on new data using complex model**.
4. **Pick *M* features best describing the complex model outcome from the permuted data**: Then it tries different combinations of predictors i.e. m number to figure out minimum number of predictor you have that gives you maximum likelihood of the class that was predicted by the black box.
5. **Fit a simple model to the permuted data with *M* features and similarity scores as weights**.
6. **Feature weights from the simple model make explanations for the complex models local behaviour**.

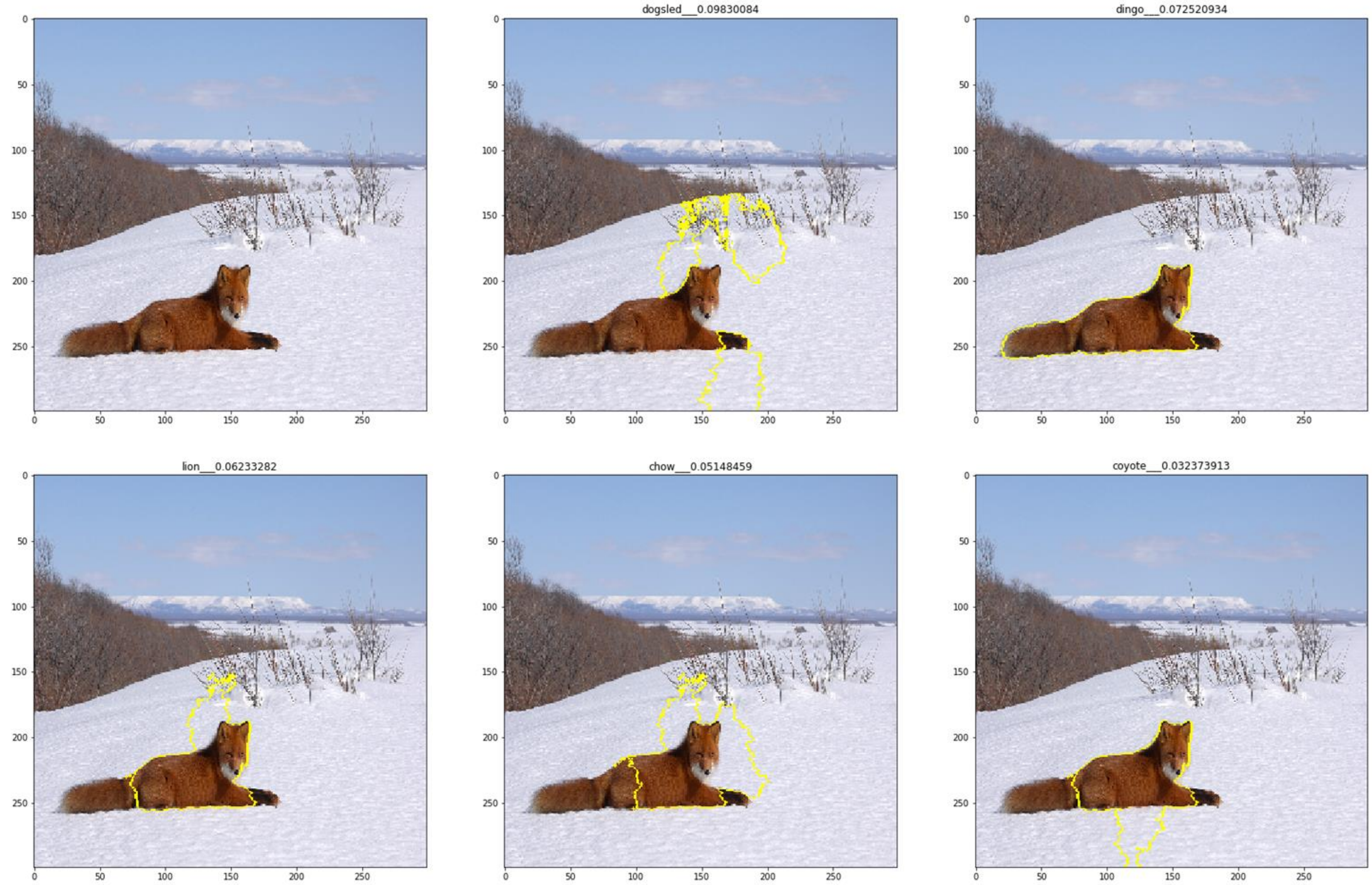
LIME. Algorithm

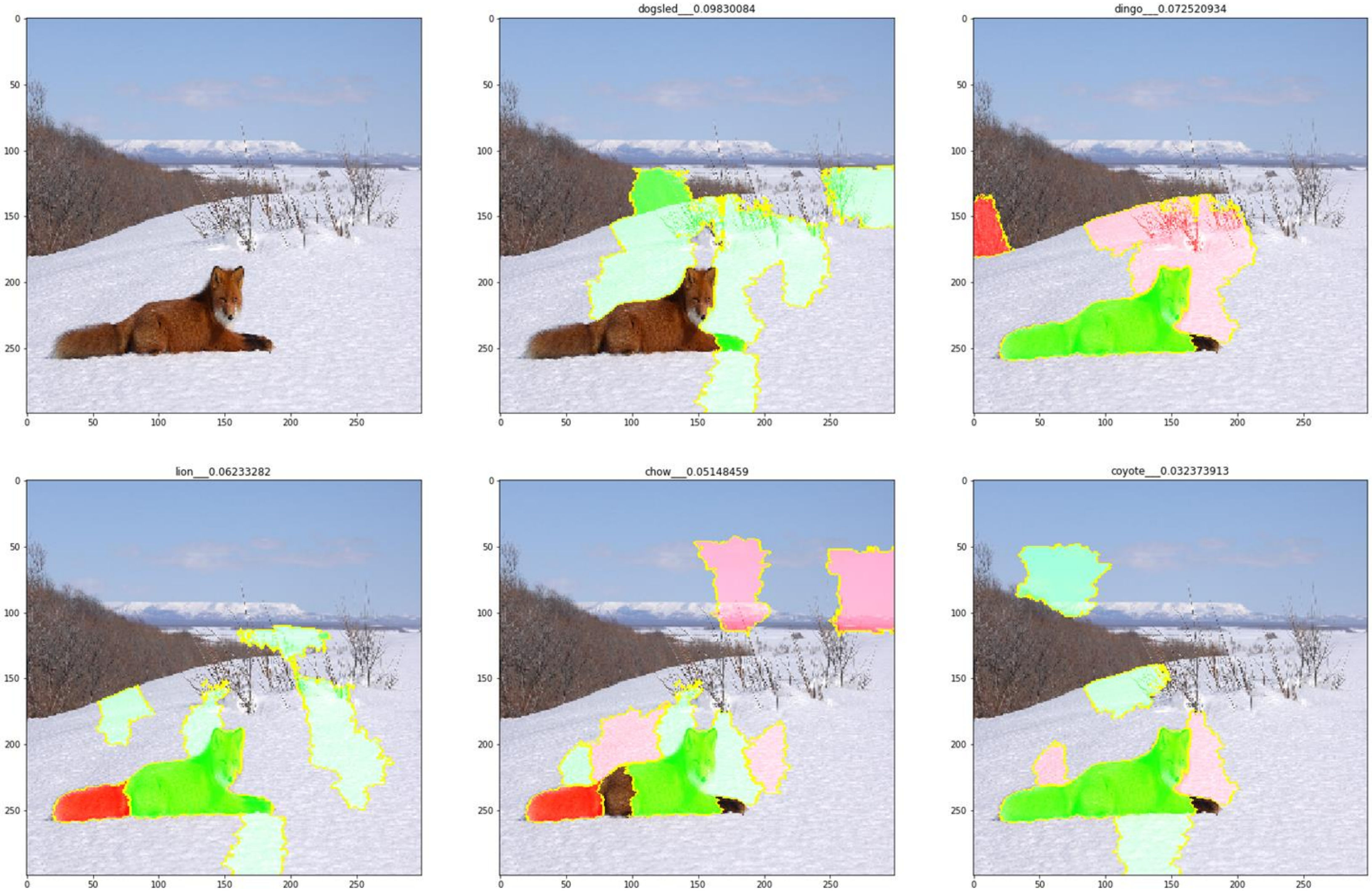
Original Image

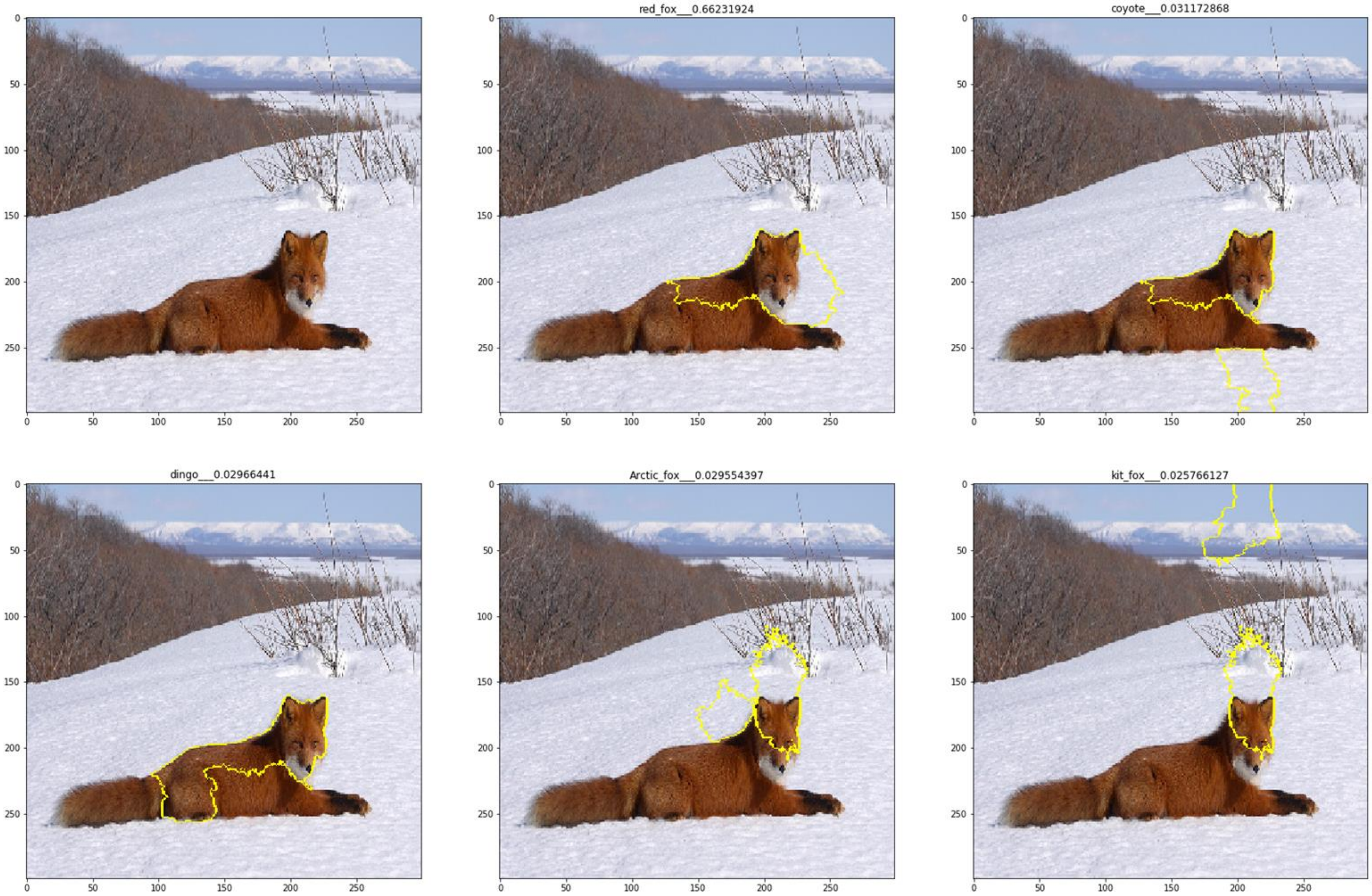
Interpretable Components

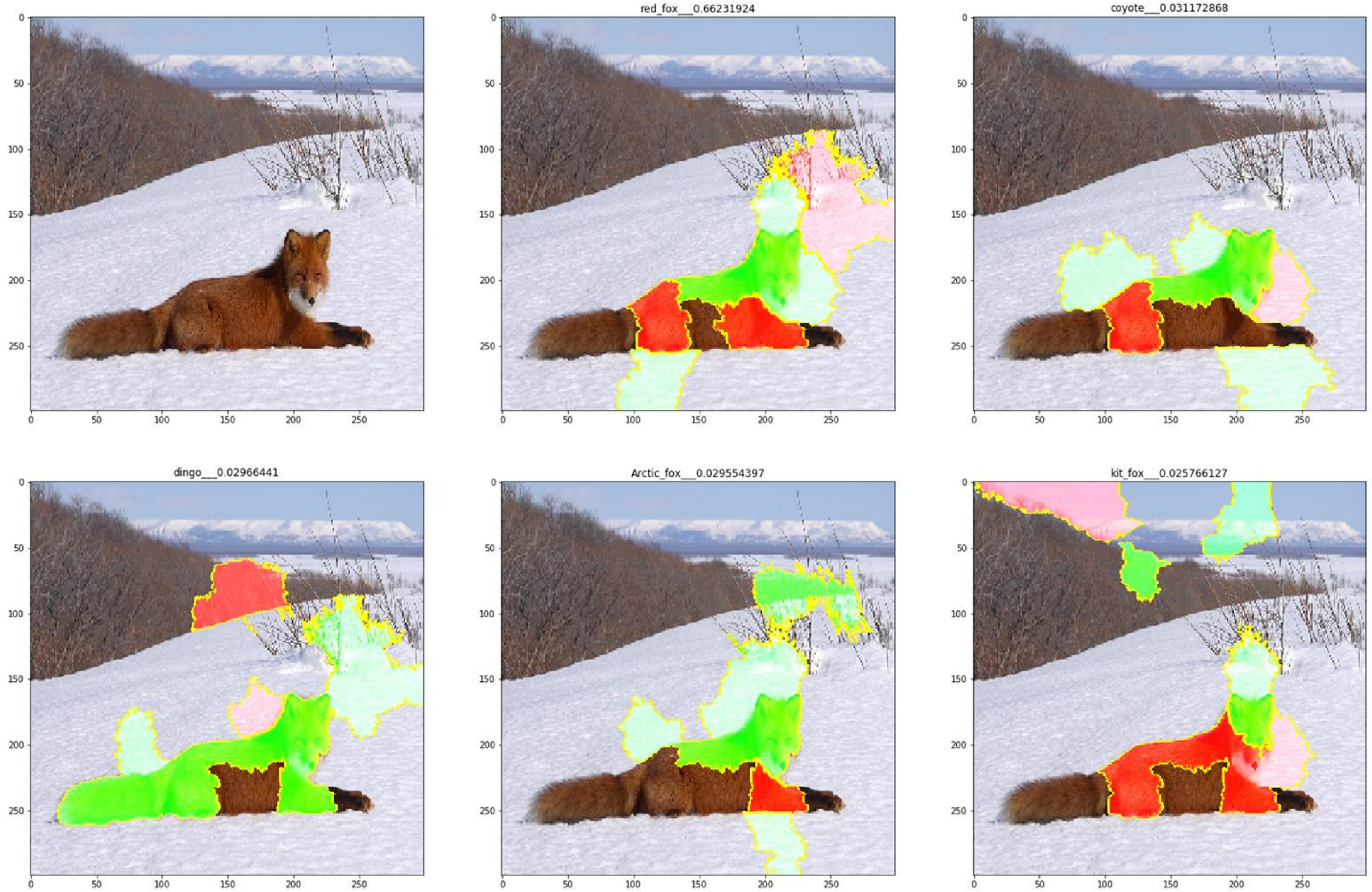
LIME. Algorithm




dogsled___0.09830084
dingo___0.072520934
lion___0.06233282
chow___0.05148459
coyote___0.032373913


dogsled___0.09830084
dingo___0.072520934
lion___0.06233282
chow___0.05148459
coyote___0.032373913


red_fox___0.66231924
coyote___0.031172868
dingo___0.02966441
Arctic_fox___0.029554397
kit_fox___0.025766127


red_fox___0.66231924
coyote___0.031172868
dingo___0.02966441
Arctic_fox___0.029554397
kit_fox___0.025766127

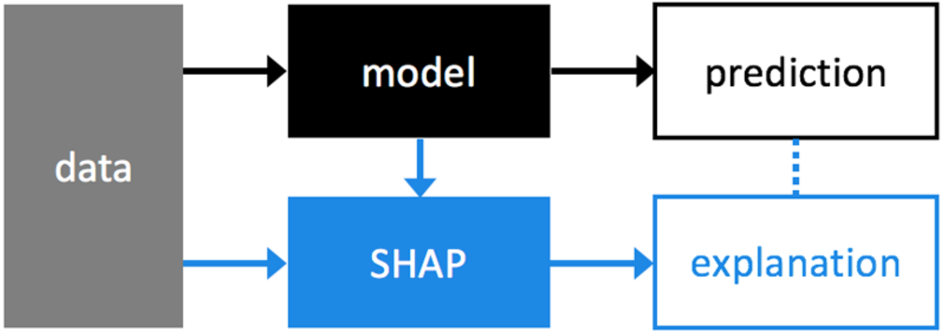
LIME. Overview

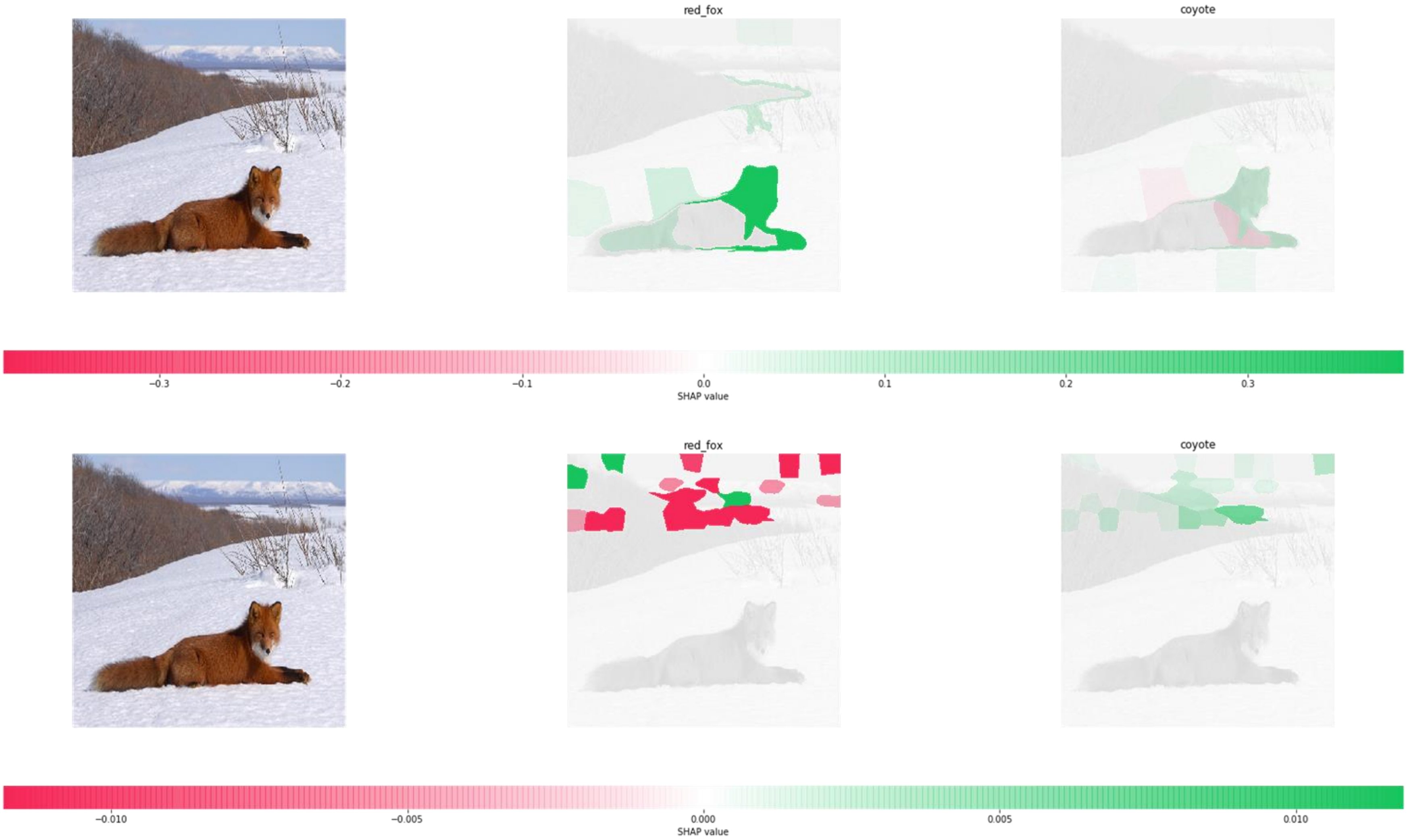
- Can be applied to any type of models
  (Images, Texts, Table structured data)
- Local, model-agnostic tool
- Linear modelling
- Intelligently select multiple local explanations for global explainability
- Many extensions (aLIME, SP-LIME)

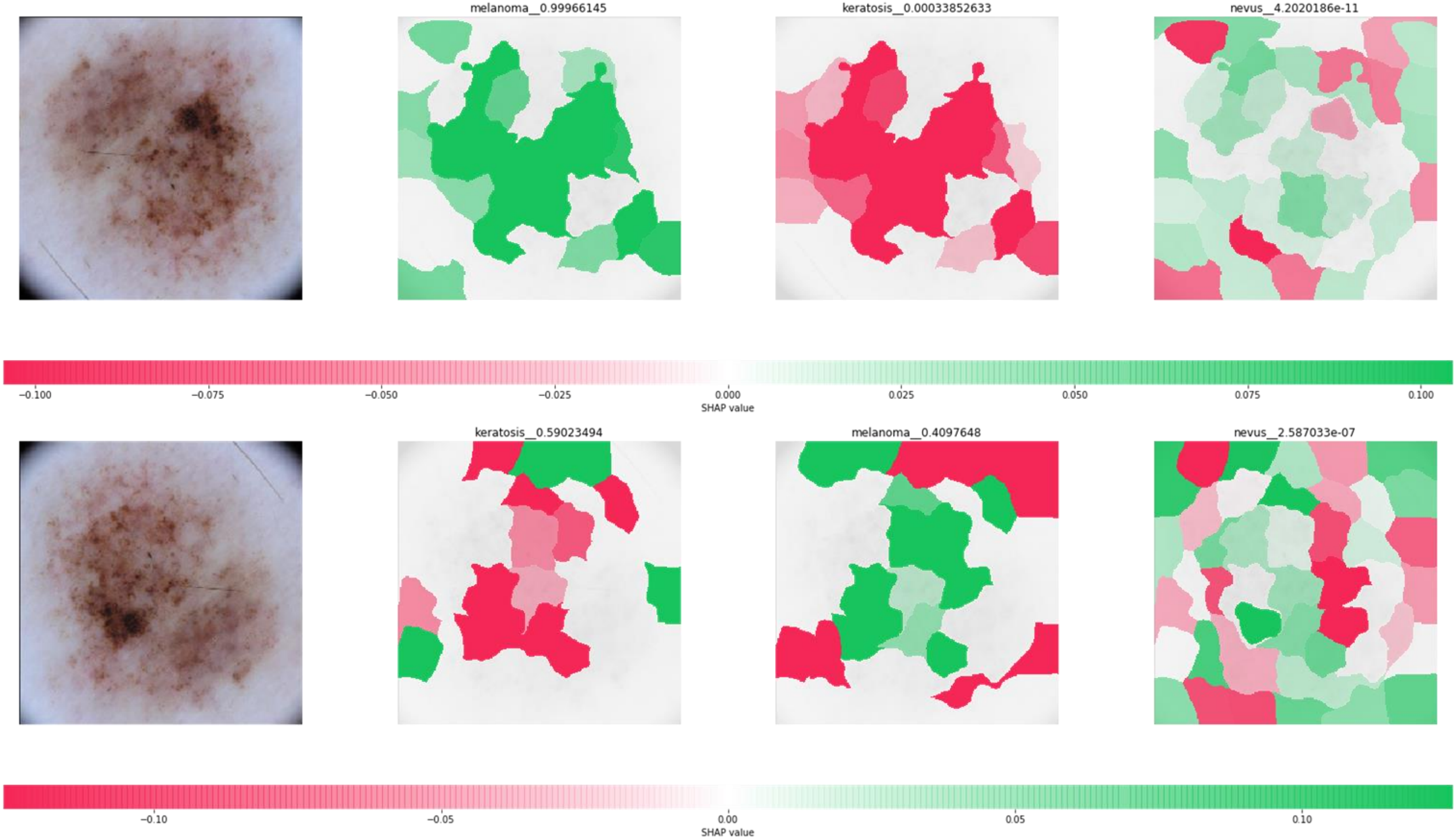
SHapley Additive exPlanations (SHAP)

- Additive feature attribution method
- A method from coalitional game theory
- Tells us how to fairly distribute the ‘payout’ among contributors
- Explainers:
  - TreeExplainer
  - DeepExplainer
  - GradientExplainer
  - KernelExplainer

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$




red_fox
coyote
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value
red_fox
coyote
−0.010
−0.005
0.000
0.005
0.010
SHAP value

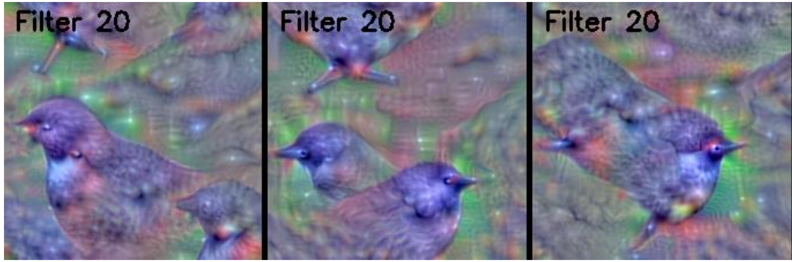

melanoma__0.99966145
keratosis__0.00033852633
nevus__4.2020186e-11
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
SHAP value
keratosis__0.59023494
melanoma__0.4097648
nevus__2.587033e-07
−0.10
−0.05
0.00
0.05
0.10
SHAP value

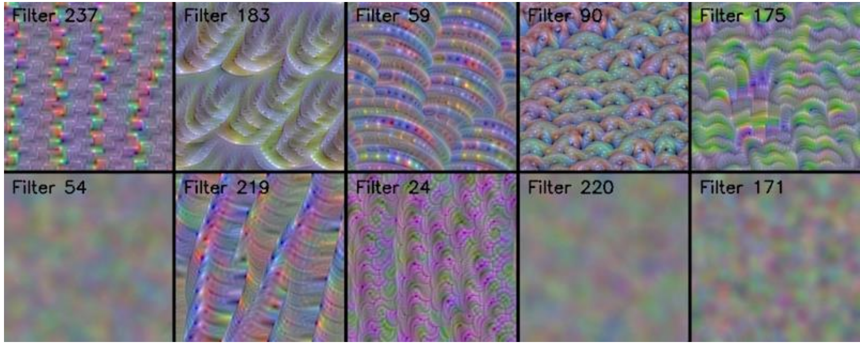
nevus__0.74570256
melanoma__0.1733578
keratosis__0.080939665
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
SHAP value
melanoma__0.7642409
nevus__0.2110382
keratosis__0.024720911
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
SHAP value

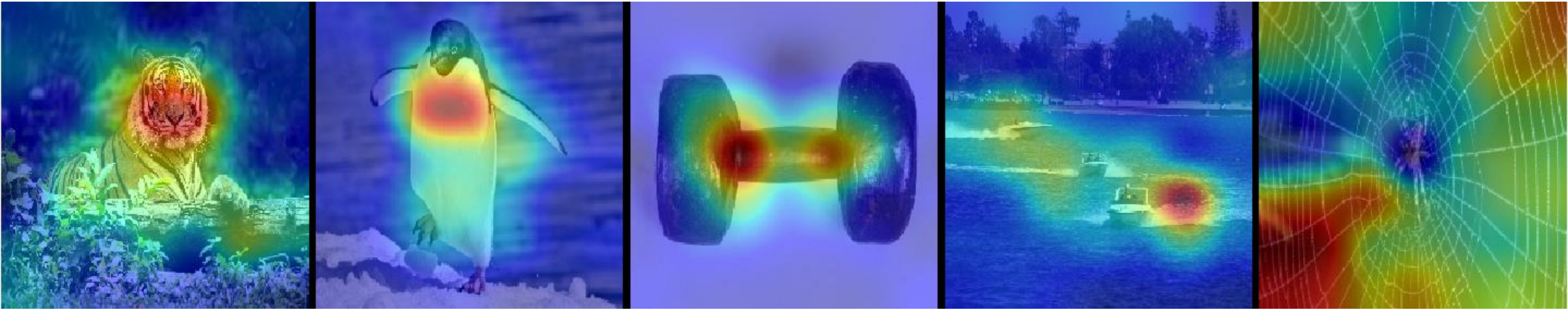
Keras-Vis

High-level toolkit for visualizing and debugging your trained keras neural network models.

- Saliency maps
- Class activation maps
- Activation maximization





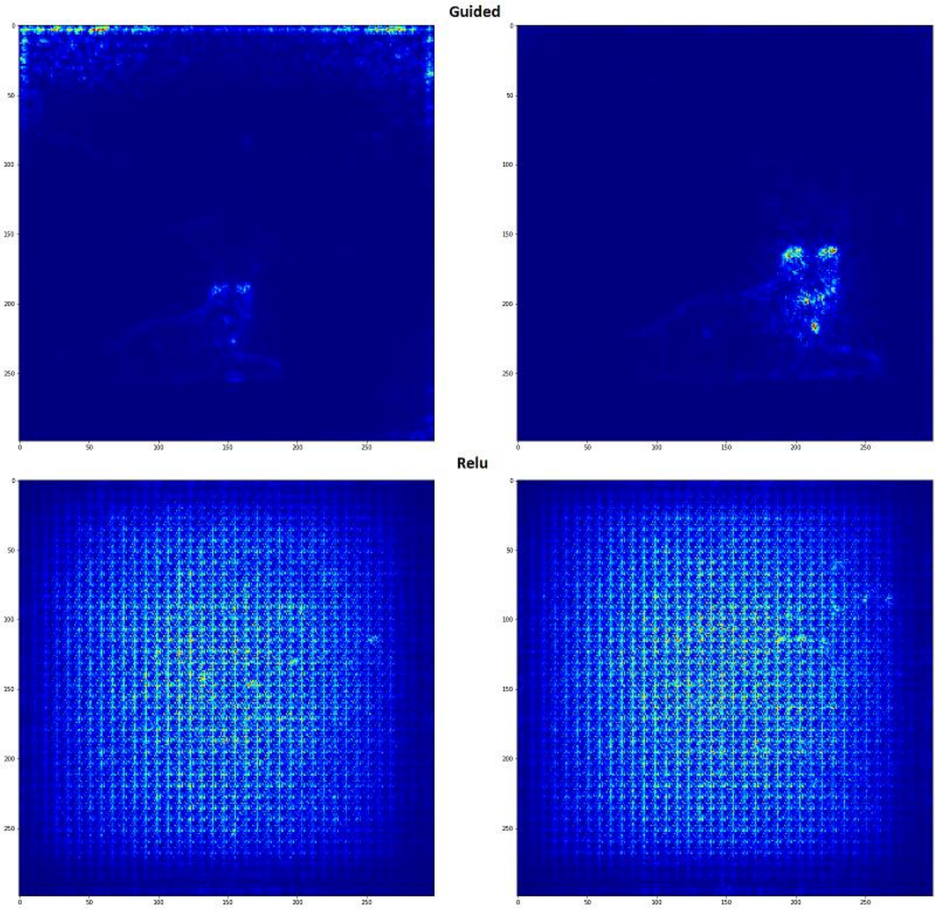


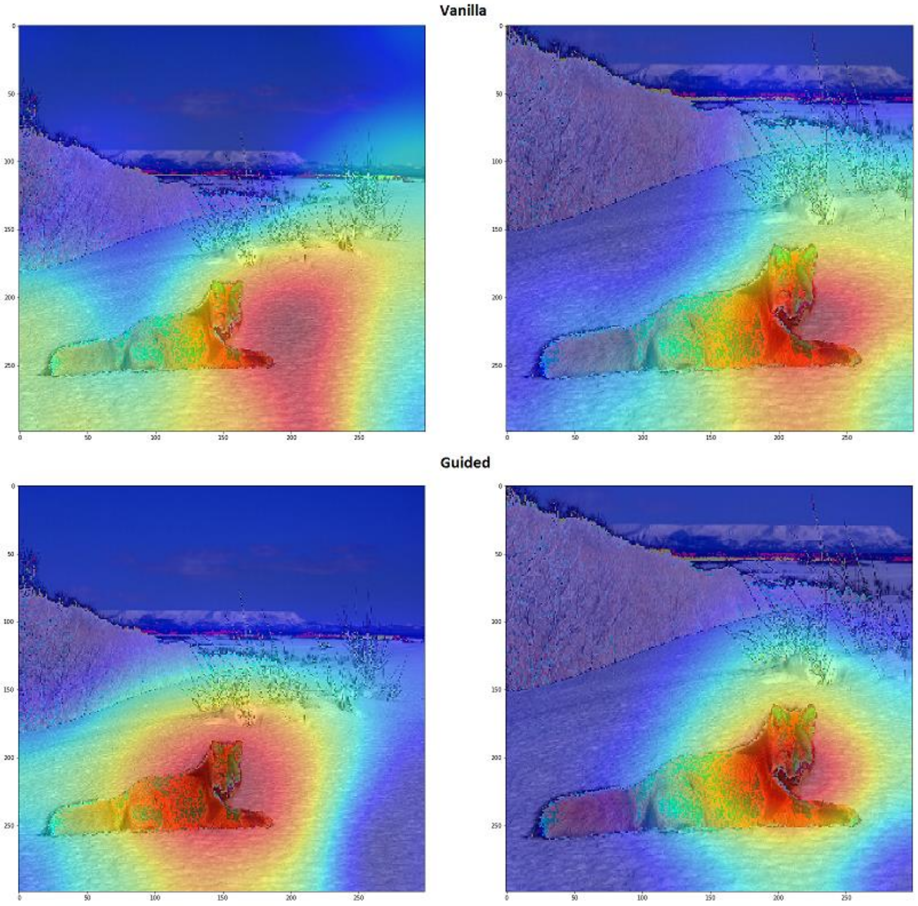
Saliency maps

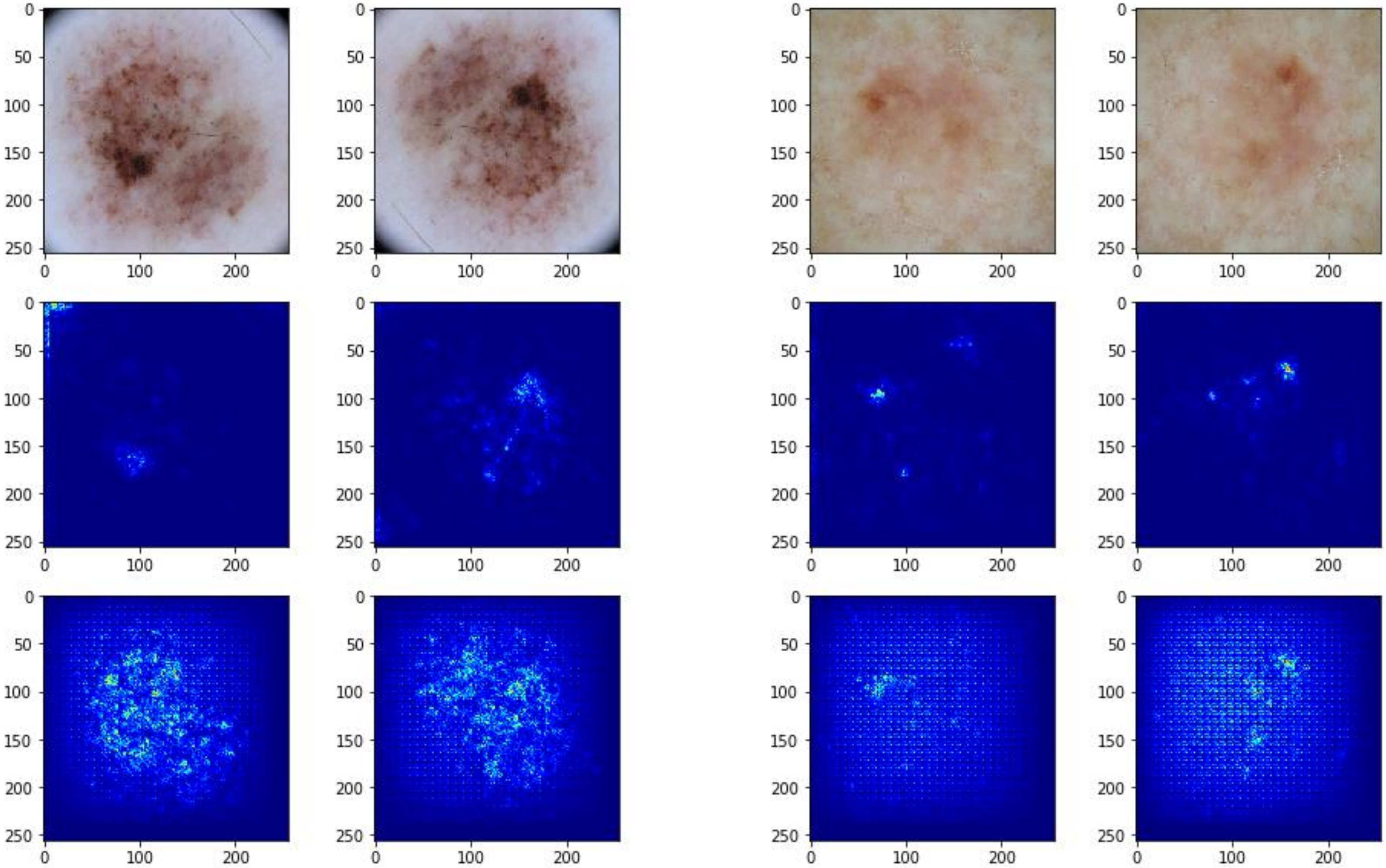
- Gives features in the input space that mattered for the classification:

$$\arg\max_{I} S_c(I) - \lambda \|I\|_2^2,$$

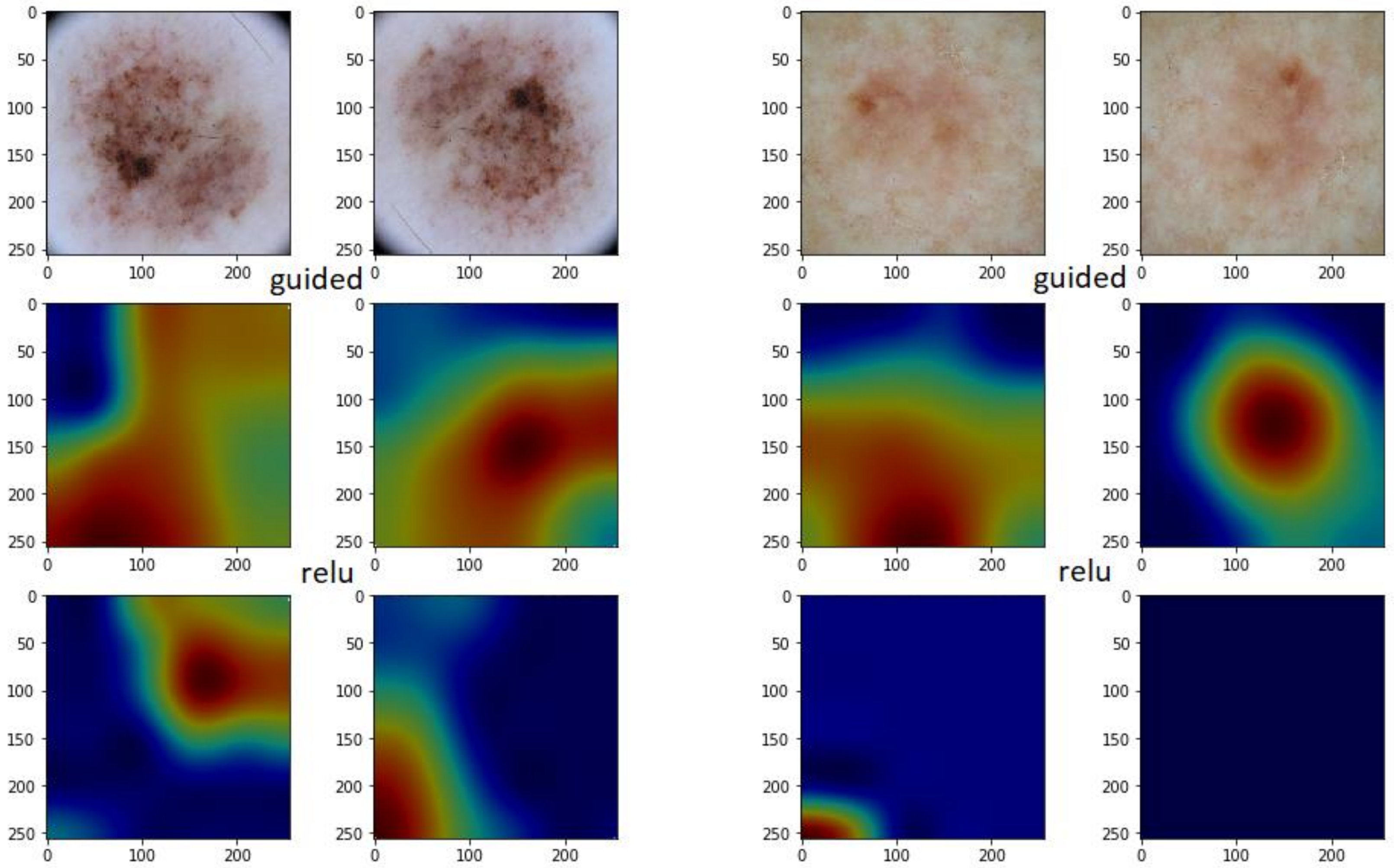
- Guided / rectified saliency. Grad-CAM
  - [Striving for Simplicity: The All Convolutional Net]()
  - The same as Saliency maps, but we use latest Convolution layer instead of output layer


Guided
Relu


Vanilla
Guided



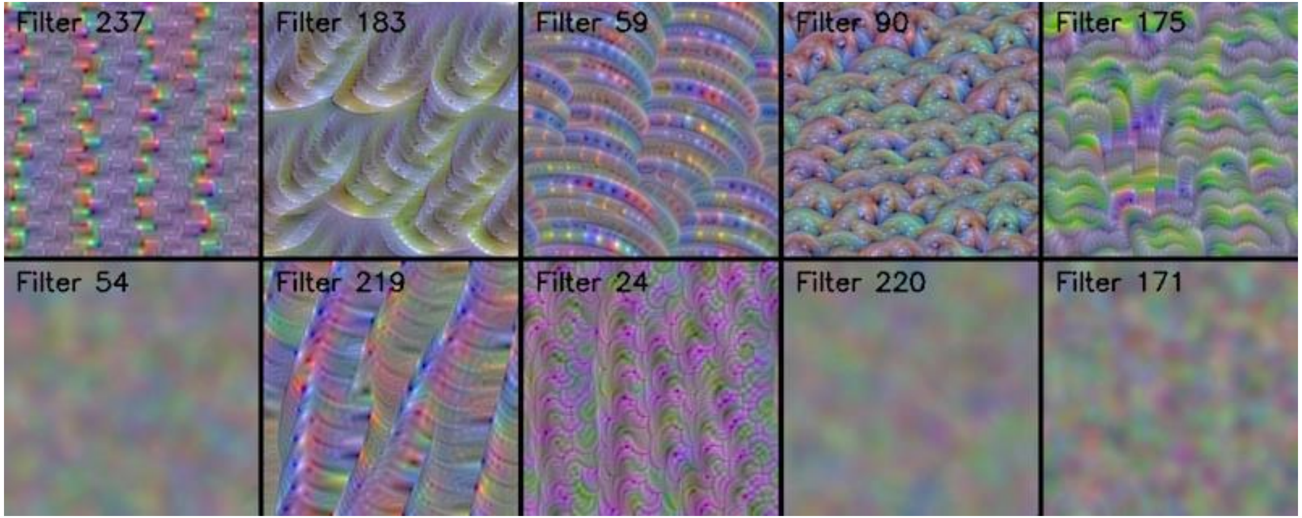
0
50
100
150
200
250
0
100
200


guided
guided
relu
relu

Keras-Vis. Activation maps

- Idea: generate an input image that maximizes the filter output activations. i.e., we compute

$$\frac{\partial ActivationMaximizationLoss}{\partial input}$$



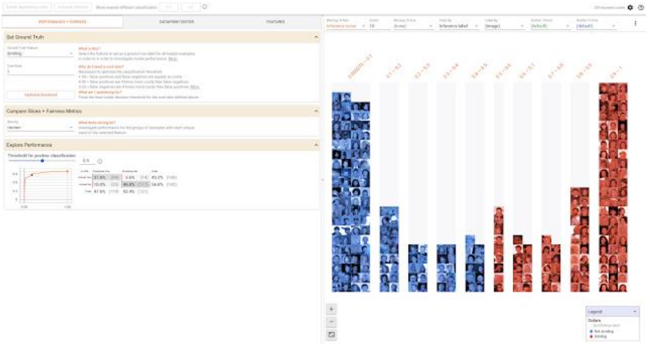
Conclusions

- Use techniques for additional model validation/exploration
- There appear a lot of new tools for model agnostic exploration
- High interest in the world (conferences)

New tools

Big companies propose new solutions for detecting models bias:

- Google "What-if" https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html
- IBM "Fairness 360 Kit" https://www.bbc.com/news/technology-45561955



Libraries

1. **LIME** - https://github.com/marcotcr/lime https://arxiv.org/pdf/1602.04938v1.pdf
2. **Keras-Vis** - https://github.com/raghakot/keras-vis
3. **SHapley Additive exPlanations** - https://github.com/slundberg/shap https://christophm.github.io/interpretable-ml-book/shapley.html
4. **Lucid** - https://github.com/tensorflow/lucid https://distill.pub/2018/building-blocks/
5. **“What If…”** tool - https://pair-code.github.io/what-if-tool/

Links

1. “Interpretable Machine Learning” book
   https://christophm.github.io/interpretable-ml-book/
2. A Survey Of Methods For Explaining Black Box Models
   https://arxiv.org/pdf/1802.01933.pdf
3. Techniques for Interpretable Machine Learning
   https://arxiv.org/pdf/1808.00033.pdf
4. Interpretable Machine Learning, ICML presentation
   https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf
5. https://www.kaggle.com/dansbecker/advanced-uses-of-shap-values
6. https://github.com/lopusz/awesome-interpretable-machine-learning

Questions?

Vladyslav Kolbasin
vladyslav.kolbasin@globallogic.com