Introduction to eXplainable Artificial Intelligence (XAI)

Evan Krell

innovation in COmputing REsearch lab (iCORE)

Outline

Motivating XAI Examples

- Three examples where XAI revealed learned strategies
- All taken from Dr. Ebert-Uphoff's talk [5]

Murdock et al. [1]: XAI Introduction

- Framework to discuss merits of interpretation methods
- Model-based & post hoc interpretability
  (complex models typically require post hoc)

Molnar et al. [2]: XAI Pitfalls

- Using XAI methods **easy** . . . ensuring meaningful explanations **hard**

Are complex models too hard to explain?

- So many pitfalls $\rightarrow$ optimal solutions often unclear or infeasible. . .

Note: focus here is **not** on specific XAI algorithms, but rather overall challenges.

Brief, intuitive explanations of specific methods when needed.

Papers & Resources

Paper [1] Murdoch, W. James, et al. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences 116.44 (2019): 22071-22080.

Paper [2] Molnar, Christoph, et al. "General pitfalls of model-agnostic interpretation methods for machine learning models." arXiv preprint arXiv:2007.04131 (2020).

Paper [3] McGovern, Amy, et al. "Making the black box more transparent: Understanding the physical implications of machine learning." Bulletin of the American Meteorological Society 100.11 (2019)

Video Peering Inside the Black Box if Machine Learning for Earth Science

Part 1 [4] https://youtu.be/n8Lsz56_EDI (Amy McGovern)
Part 2 [5] https://youtu.be/LTD0wLyMVE8 (Imme Ebert-Uphoff)

Course [6] Ryan Lagerquist & Imme Ebert-Uphoff. AI2ES/CIRA Short Course on Explainable Artificial Intelligence for Environmental Science

https://docs.google.com/document/d/1lqpABwDl3kPe6ThE-NIDR64PimnltJEuKNkysDZuWKQ

Book [7] Christoph Molnar: Interpretable Machine Learning

christophm.github.io/interpretable-ml-book/

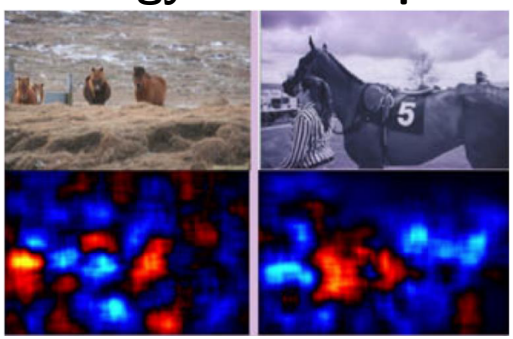
Code [8] Evan Krell: PartitionShap demo notebooks

https://github.com/conrad-blucher-institute/partitionshap-multiband-demo

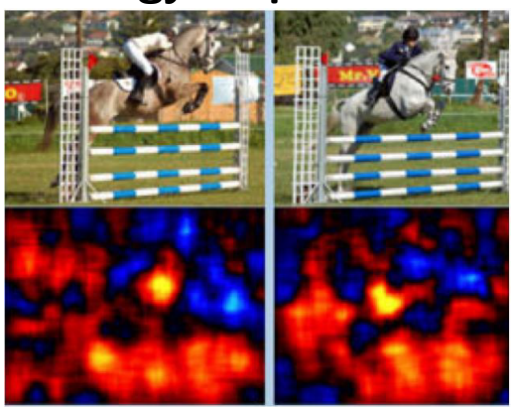
Example 1: Horse detection [5]

- Models may achieve high (test data) performance despite learning wrongly
- How? demonstrated using Layerwise Relevance Propagation (LRP)
- Lapuschkinet al. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096, doi:10.1038/s41467-019-08987-4.

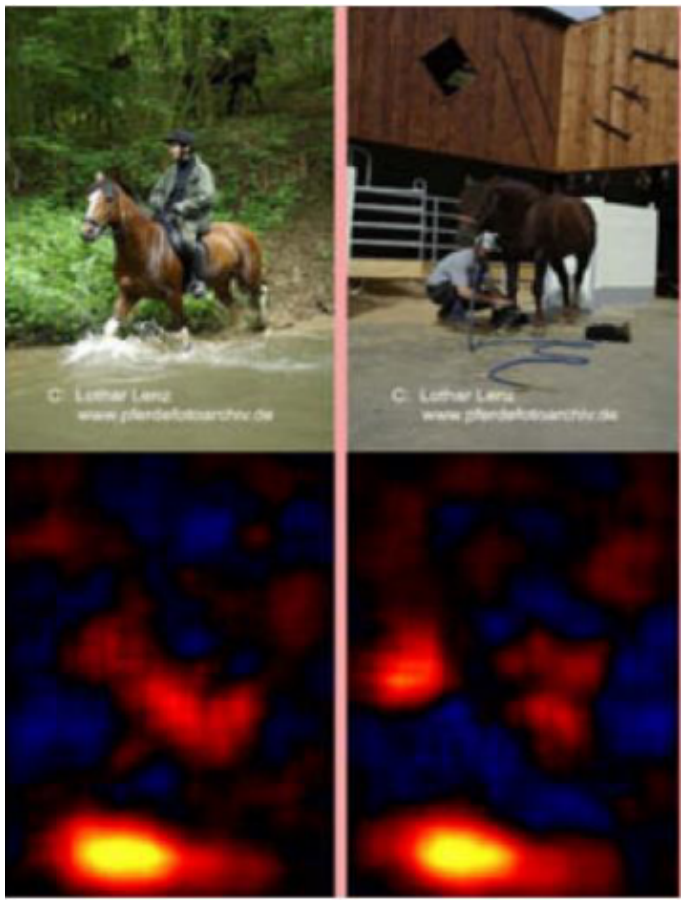
Strategy 1: horse parts



Strategy 2: poles



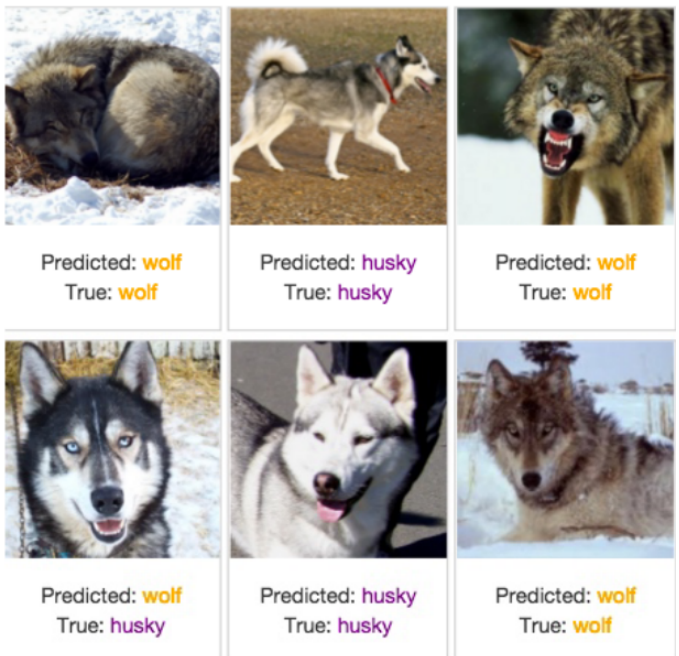
Strategy 3: photographer's watermark



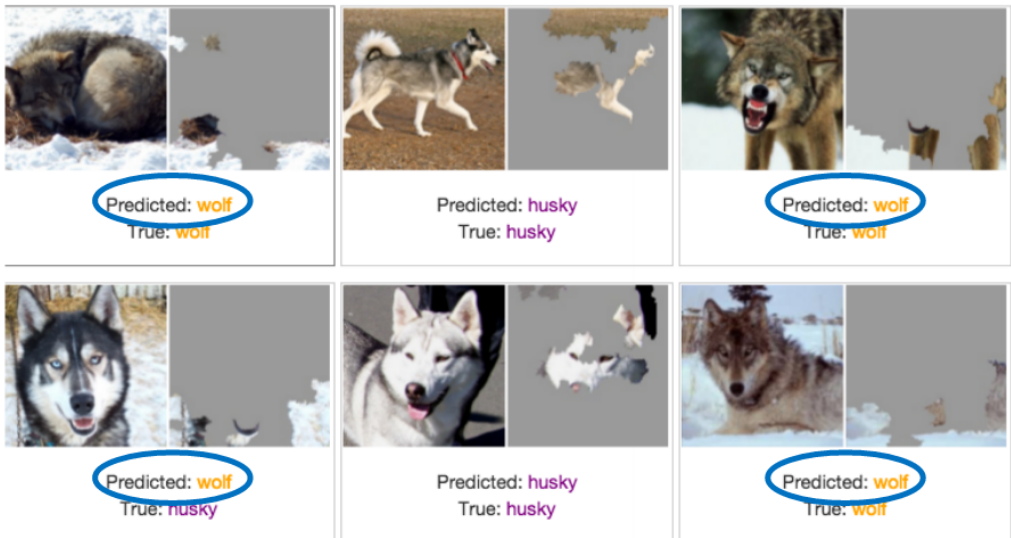
Example 2: Wolf or husky [5]

- Hard problem, but achieved good performance
- But is it really a wolf vs husky detector?
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Results look good



But actually a snow detector?

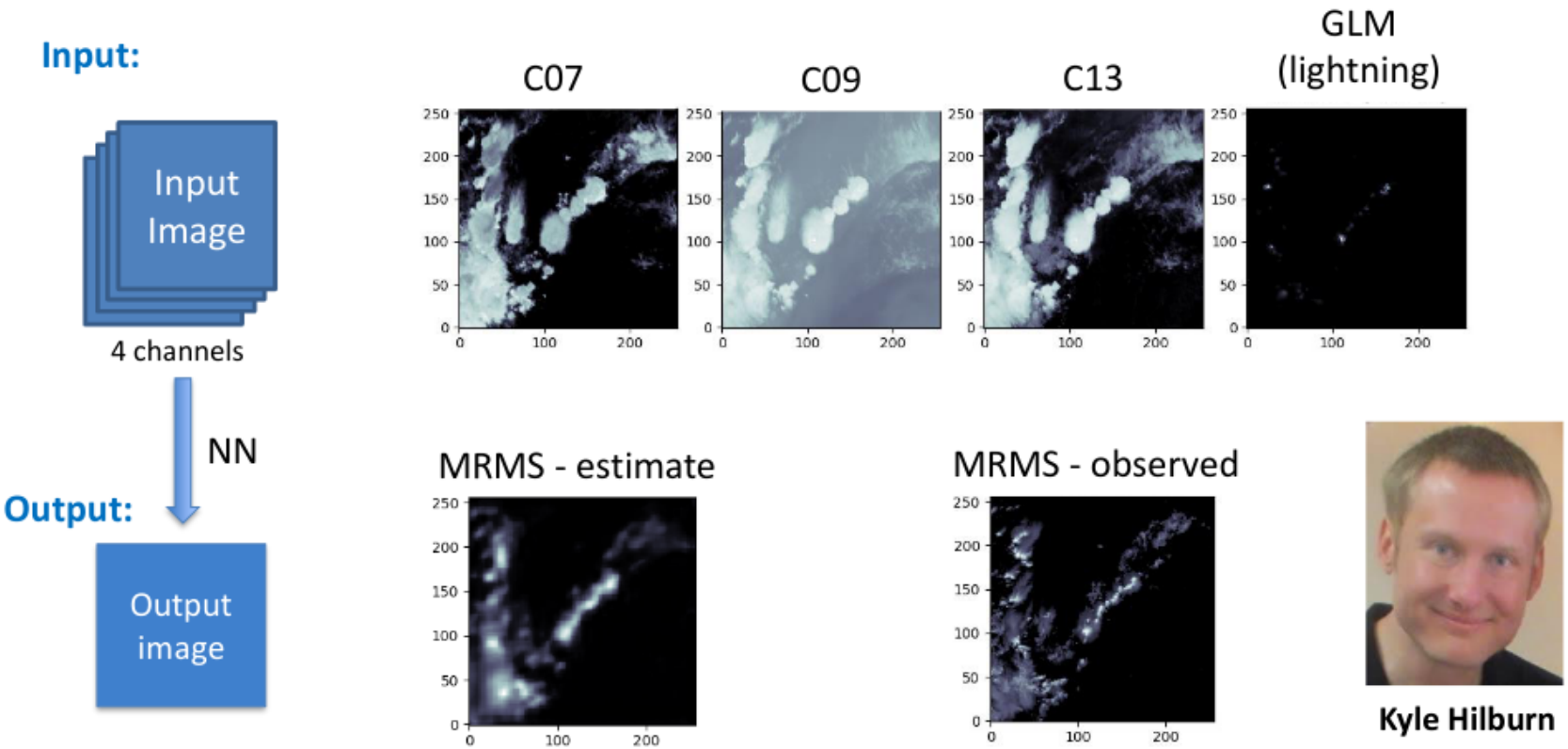


Example 3: Generating synthetic radar images from GOES imagery (1/2)

Slide exactly a screenshot from [5]

Application 2:
Generating synthetic radar images from GOES imagery

Input: GOES Channels C07, C09, C13, GLM. Output: MRMS (radar).

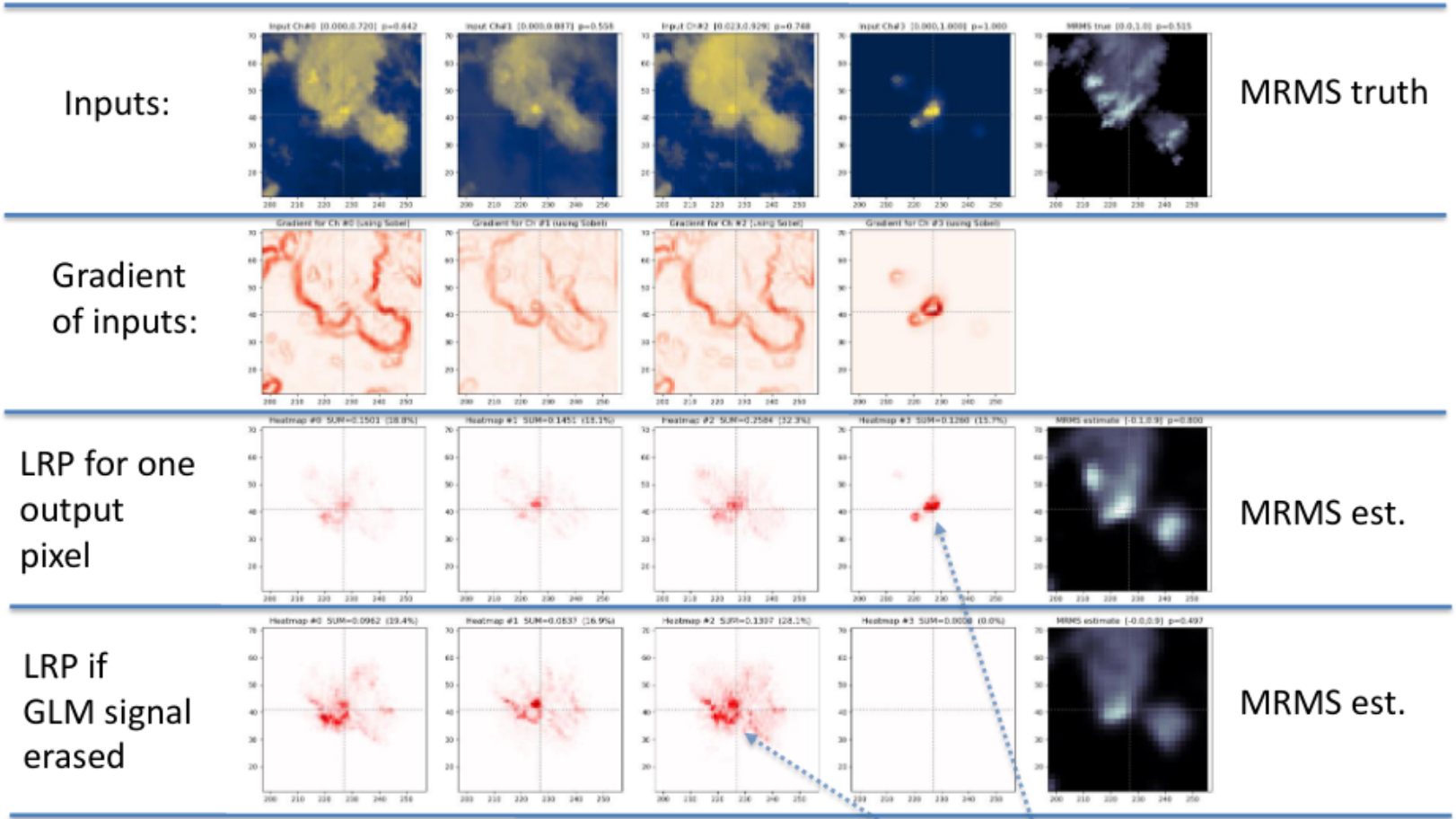


Motivation: GOES imagery is available in all of CONUS, but MRMS is not.

Example 3: Generating synthetic radar images from GOES imagery (2/2)

Slide exactly a screenshot from [5]

Question: How does NN know when to create **large** MRMS estimates?
Method: Select examples where MRMS estimate is high. Where is NN looking (LRP)?



LRP yields 2 strategies for creating large MRMS estimates:

Strategy 1: Presence of lightning triggers high MRMS values. **Lightning** = strongest trigger.

Strategy 2: In no lightning NN focuses on locations with strong gradients: **cloud boundaries**.

Machine Learning Models

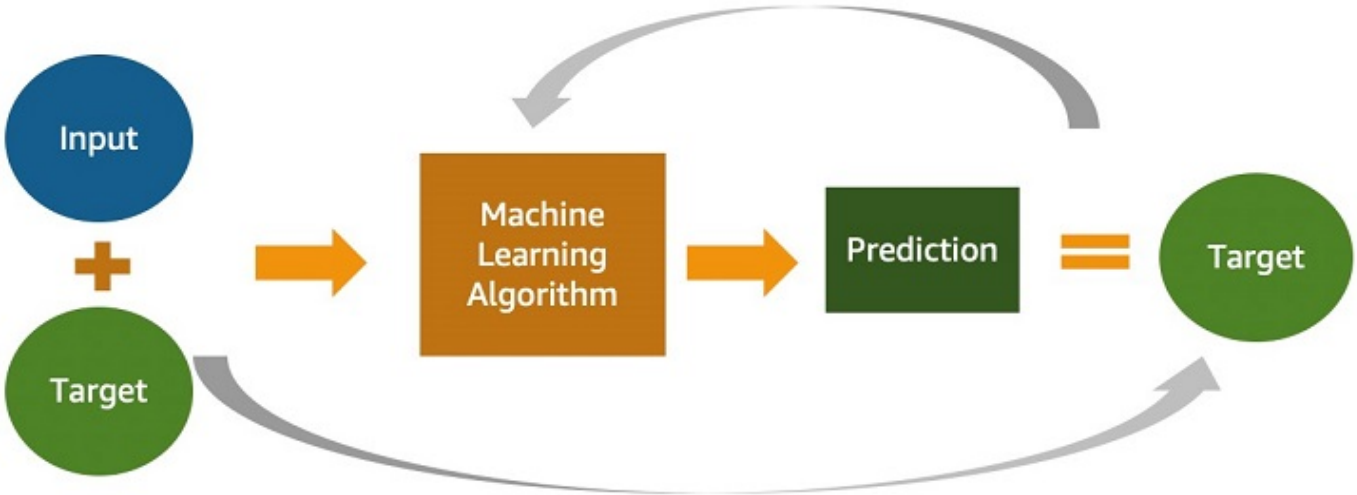


Image: aws.amazon.com/blogs/big-data/create-train-and-deploy-machine-learning-models-in-amazon-redshift-using-sql-with-amazon-redshift-ml/

Components

- **Input:** features from some data-generating process $\rightarrow$ used as predictors
- **Prediction:** variable related to the data such that a given instance of data features predicts this variable (hopefully)
- **Model:** learned **association** (don't assume causation) between data & predictions that might reflect real association in data-generating process

Model Interpretation

"We define interpretable machine learning as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model." - [1]

- Goal: learned relationships reflect real-world
- Impediments: training data biases, insufficient learning algorithm, ...
- **Horse:** learned association between watermark & horse label
- **Wolf:** learned association between snow & wolf label

Interpretation uses

- **Model evaluation**
  To predict wolf, it is only looking at snow!
- **Model repair**
  So, ensure training data has diverse backgrounds
- **Build trust**
  Now XAI output shows looking at wolf characteristics
  $\rightarrow$ more confident that model will work on future images

XAI method menagerie

A selection of XAI methods applied to an image classification model:

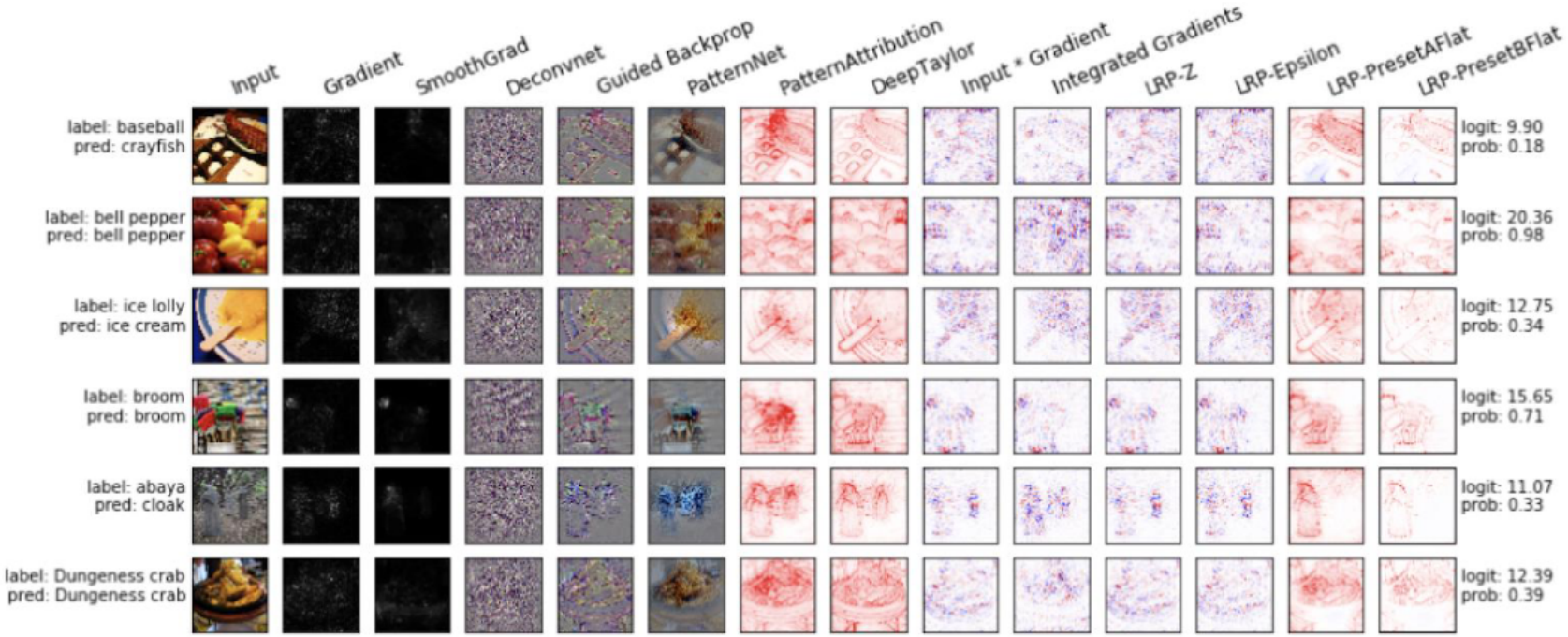


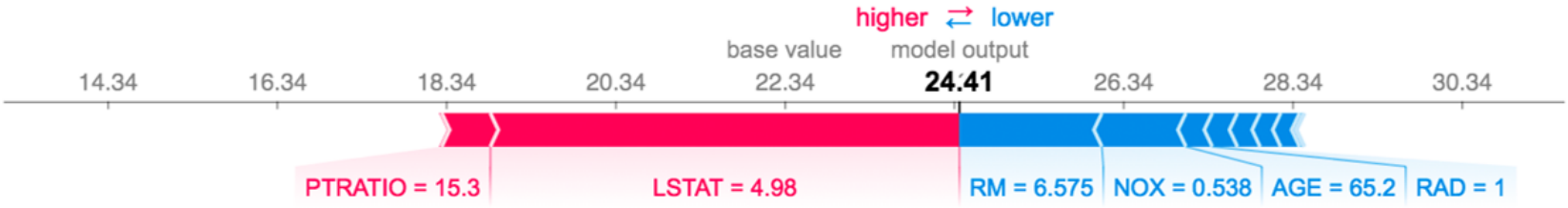
Image: pypi.org/project/innvestigate/

- Large number of XAI methods exist
  - Novel methods proposed regularly
  - Often easy to apply using software packages
  - Generally variants on a smaller set of classes:
    gradient-based, permutation-based, optimization-based, ...
- But which to choose? And how to know that the explanation is correct?

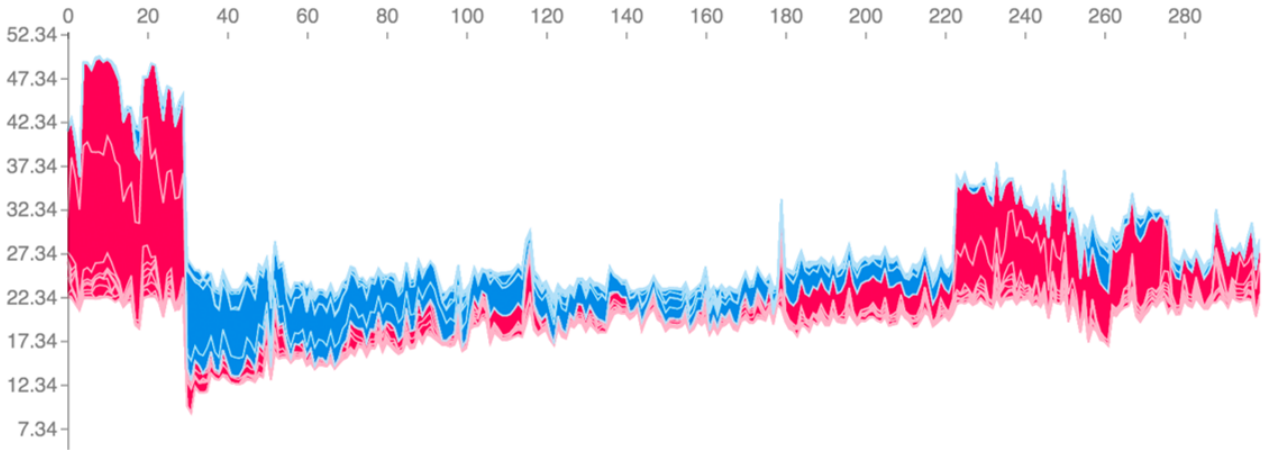
"considerable confusion about the notion of interpretability...
standard of evaluation varies considerably ...
challenging both for researchers in the field to measure progress
and for prospective users to select suitable methods" - [1]

Explanations: local & global

- **Local:** Explains a single (predictors, prediction) instance
- Example: SHAP output → contribution of each feature to specific output



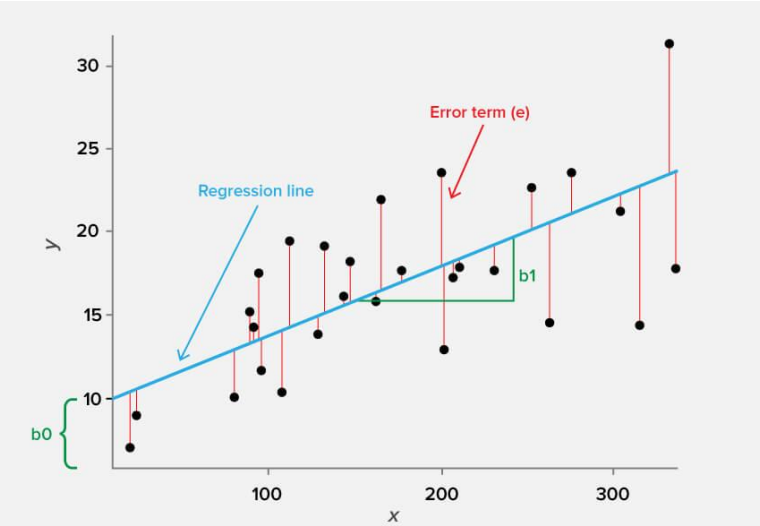
- **Global:** Overall patterns used for prediction by the model
  - Example: many SHAP local explanations combined



"Important scores at the prediction level can offer much more information than feature importance scores at the dataset level. This is a result of heterogeneity in a nonlinear model: The importance of a feature can vary for different examples as a result of interactions with other features." - [1]

Model-based interpretation (linear regression example)

Linear regression



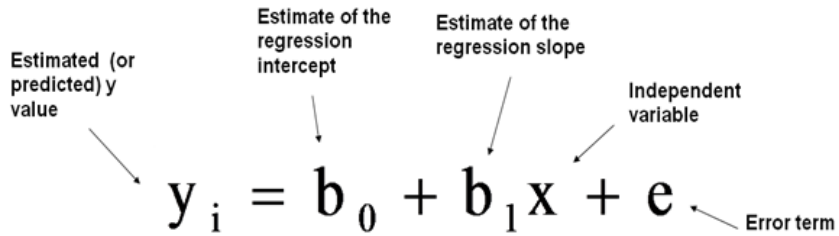


Image: datascience.foundation/sciencewhitepaper/understanding-linear-regression-with-python-practical-guide-2

Weights = explanations . . . ?

- Each feature has a weight
- Magnitude of weight $\rightarrow$ feature importance

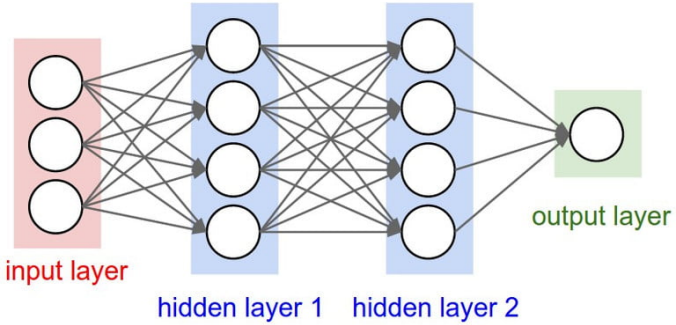
If a simple model works

$\rightarrow$ use it

Weights + data = explanations

- Linear model = $1000x_1 + 10x_2 + 1x_3$
  Which feature is most important?
- Average data x = (1, 1, 1000000000000000)
  Does this change your answer?

Explaining complex models with post hoc XAI

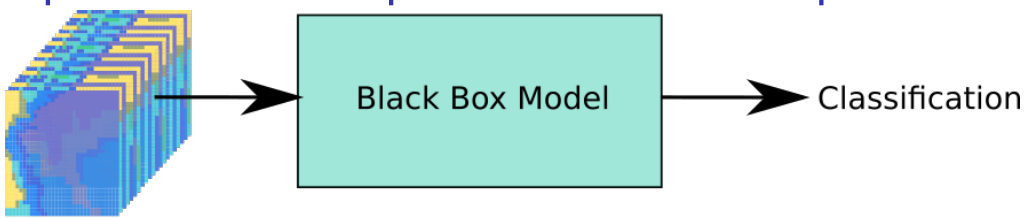


Complex models

- Learn arbitrary, nonlinear relationships
- Complex relationships → complex models

Image: digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

Learns very complex relationships → hard to interpret



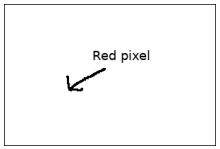
Given already trained model → post hoc XAI

- Methods probe the model in various ways to reveal descriptions of the relationship between input data and predictions
- Simple example:
  - is "age" an important feature?
  - repeatedly replace "age" with random values
  - prediction is ∼ unchanged → low importance score
  - But dependencies, autocorrelation, feature interactions → **harder**

To clarify: can (and do) apply post hoc methods to simpler models too. Maybe modified to take advantage of known structures. TreeSHAP is a faster implementation of SHAP for tree-based models.

Some challenges of post hoc XAI

Single feature `red pixel`



Entire image

Individual features not semantically meaningful

- CNN $\rightarrow$ pixels into *edges*, *textures*, *shapes*, ... $\rightarrow$ semantic meaning
- data are high dimensional and complex ... methods must deal with the challenge that individual features are not semantically meaningful - [1]

Dependency breaks many technique assumptions

- Permutation-based XAI example:
  Original data: (age = 50, salary = 100K), (age = 1, salary = 0)

  Permuted data: (age = 50, salary = 0), (age = 1, salary = 100K)

Correlations make it hard to assign importance

Which pixel is important to classify the image as *forest*?

Computational complexity & large number of features

- Some method's optimal solutions have combinatorial complexity
- Instead, methods approximate with sampling
- Goal: multiple runs $\rightarrow \sim$ same explanation

Firetruck image: jdpower.com/cars/shopping-guides/what-are-the-dimensions-of-firetrucks

Post hoc explanation evaluation

User's perspective

- Confirmation bias: correct method might not be what *seems reasonable*
- Triangulate: multiple XAI methods $\rightarrow$ see where they agree (see [3], [6])
- More rigour: apply sanity checks & significance testing
- Also make sure that the XAI methods make sense for the task
  Most will produce output with your model, but might not be appropriate
  See [3] for a table of (methods, tasks) for earth science audience!

XAI researcher's perspective

- Proposed novel XAI method $\rightarrow$ is it any good?
- Difficult to measure corrective of an explanation ([1]: *descriptive accuracy*)
- Triangulation method?
  - IF novel output similar to others $\rightarrow$ why do we need the novel method?
  - ELSE disagrees:
    - Because it reveals information missed by others?
    - Because it is wrong?