Maria Moreno de Castro
moreno@dkrz.de

DKRZ

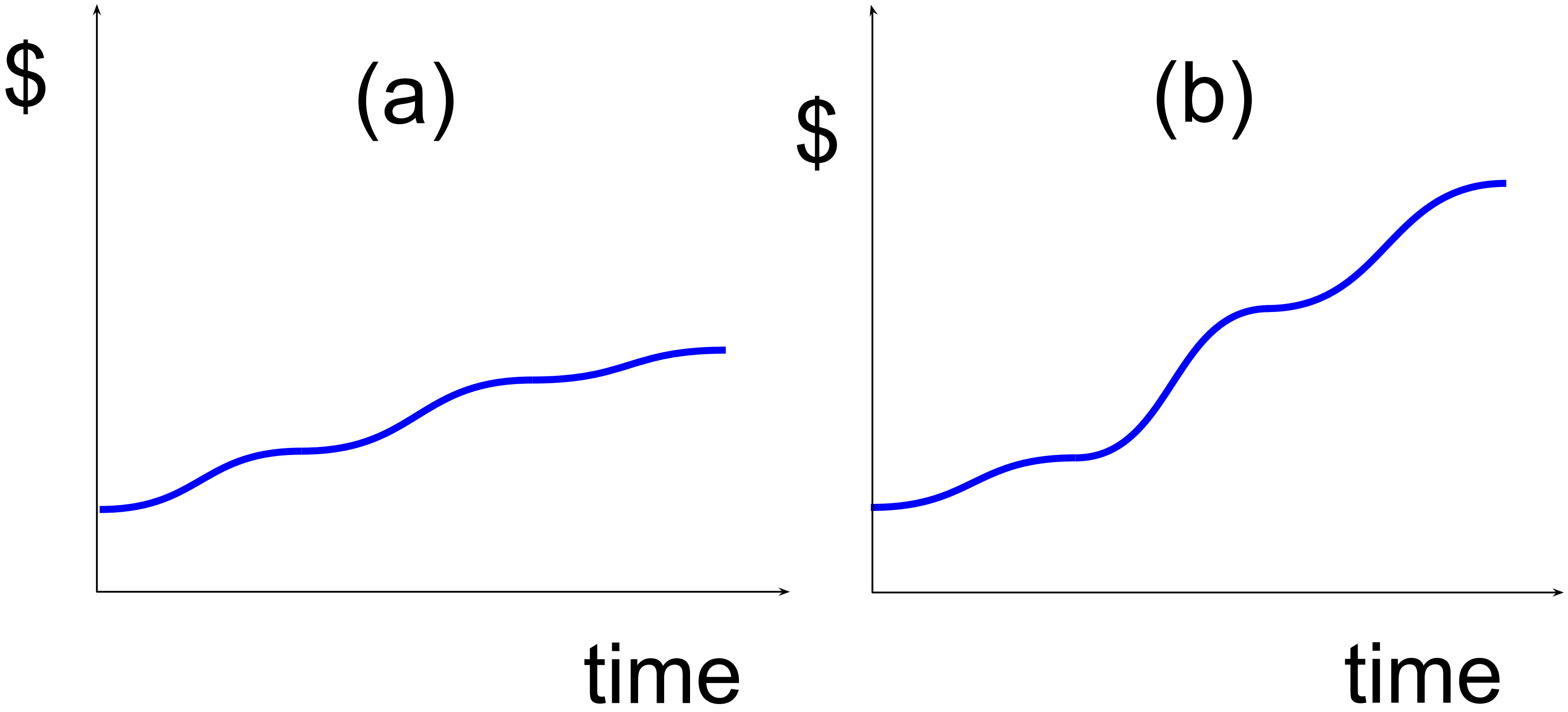
Uncertainty quantification, interpretability, and explainability

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

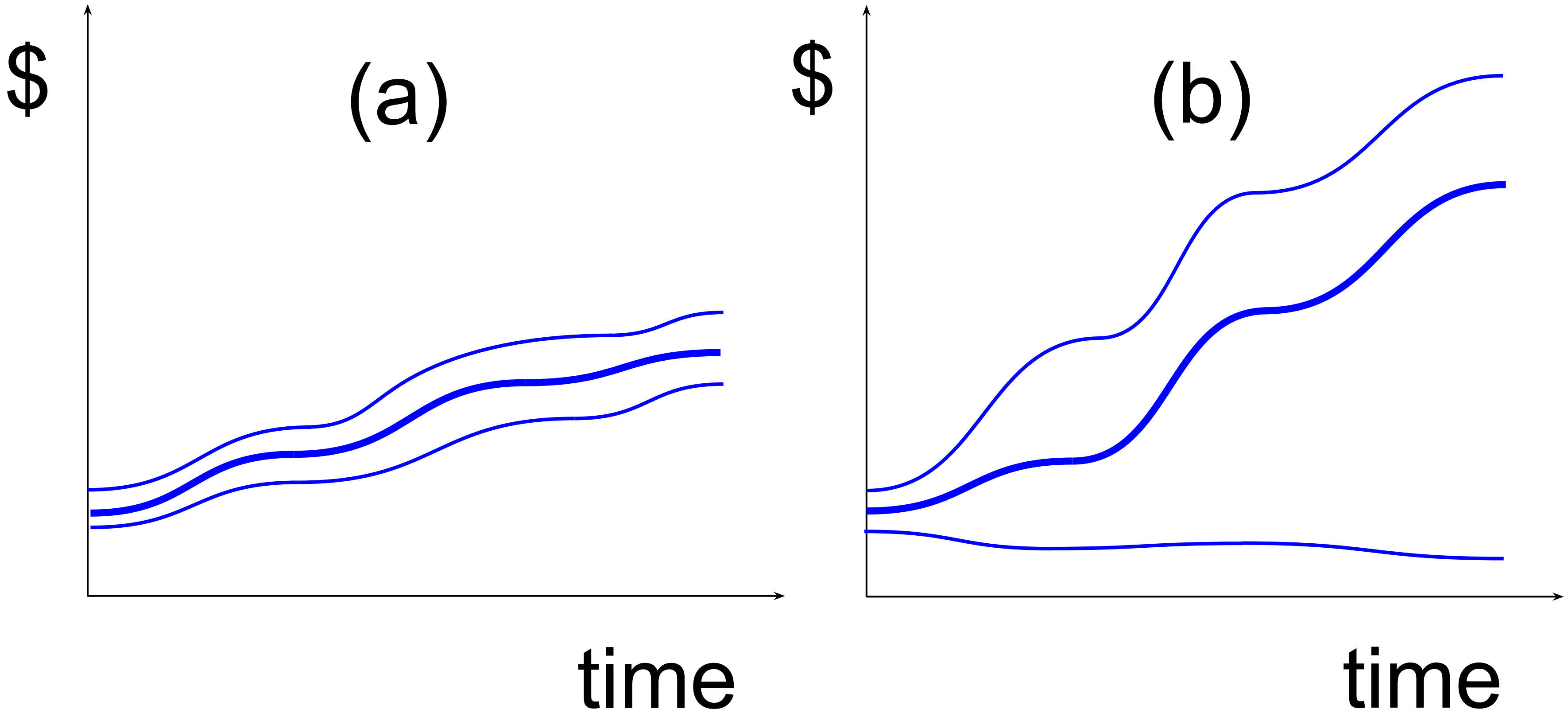
IS-ENES3 is funded by the EU Horizon 2020 Research and Innovation program under the grant agreement No 824084

uncertainty matters

Two business models, which one would you go for: (a) or (b)?



And now?



Uncertainty quantification increases the quality of the decisions

Motivation

nature

Perspective | Published: 13 February 2019

Deep learning and process understanding for data-driven Earth system science

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais & Prabhat

Nature **566**, 195–204(2019) | Cite this article

37k Accesses | **72** Citations | **320** Altmetric | Metrics

moreno@dkrz.de

Motivation

nature

Perspective | Published: 13 February 2019

Deep learning and process understanding for data-driven Earth system science

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais & Prabhat

Nature **566**, 195–204(2019) | Cite this article

37k Accesses | **72** Citations | **320** Altmetric | Metrics

Conclusions

(2) Plausibility and interpretability of inferences
Models should not only be accurate but also credible, incorporating the physics governing the Earth system.

(3) Uncertainty estimation
Models should define their confidence and credibility.

moreno@dkrz.de

Definitions

- **Black-box models**

Humans cannot understand the cause of the decisions: knowing the value of the parameters is not enough to infer what is going on and/or underlying assumptions/limitations are unknown.

- **Explainable models**

The models are still black-boxes but we use some methods (based on surrogate models) a posteriori to try to infer where/why the predictions came from.

- **Interpretable models or Glass-box models**

Humans can understand the cause of a decision: knowing the value of the parameters helps and the underlying assumptions/limitations are known. Examples: linear models, logistic regression, decision trees, naive Bayes, and k-nearest neighbors.

moreno@dkrz.de

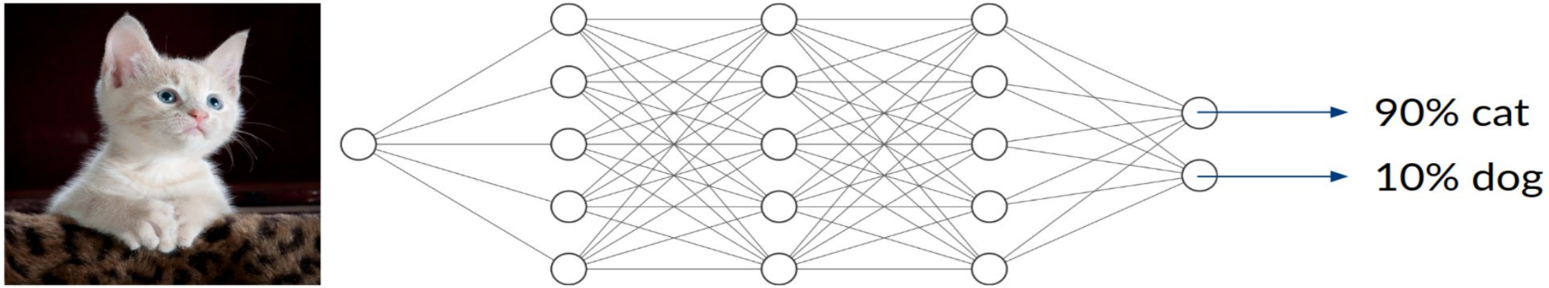
Fundamental problems (I):
algorithms are designed for interpolation, not extrapolation

moreno@dkrz.de

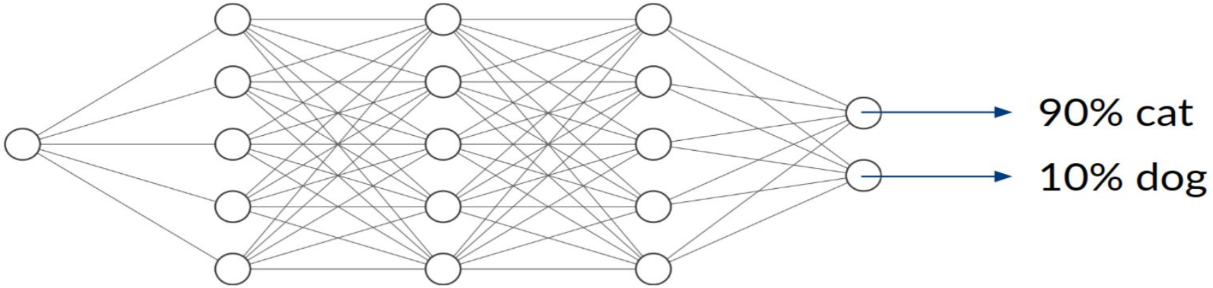
classical example:

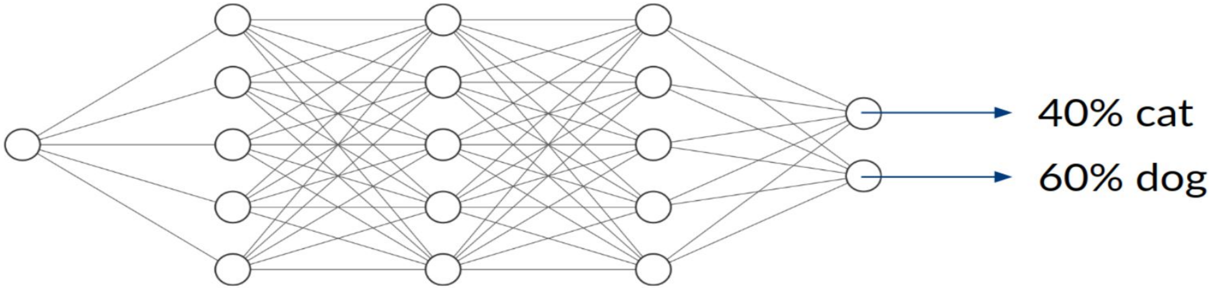
we develop an algorithm that distinguishes pictures of dogs and cats by exposing it to many labelled pictures of dogs and cats and let it find what are the main features

moreno@dkrz.de



moreno@dkrz.de



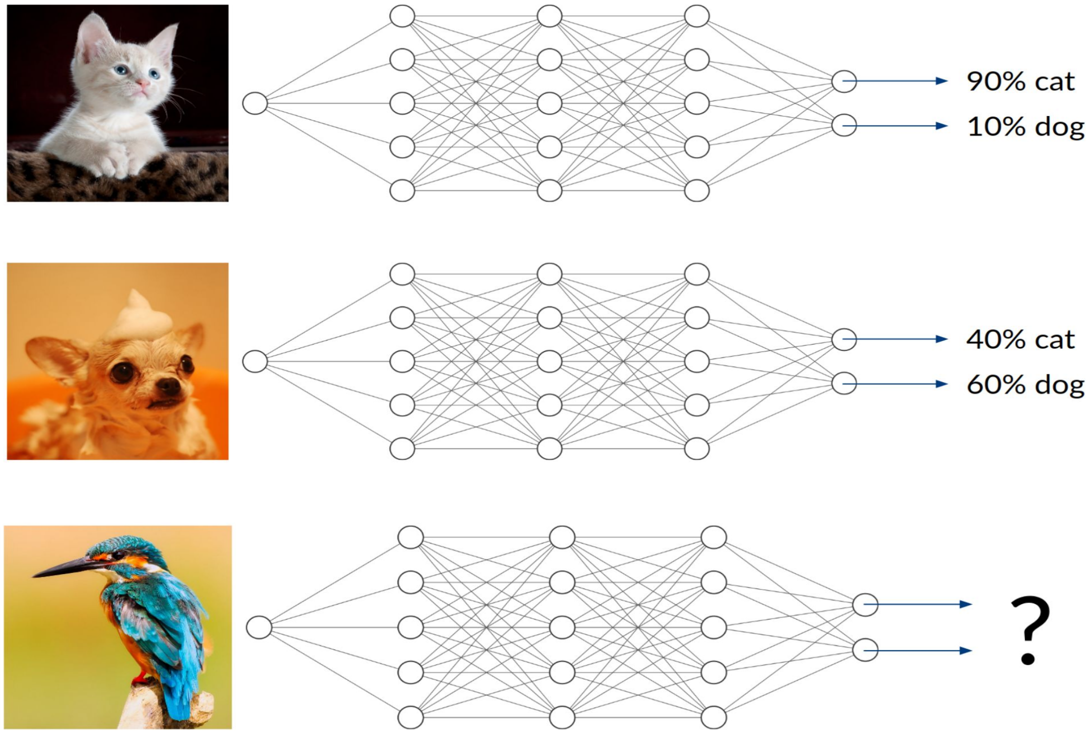


The dog in this picture looks a bit like a cat

moreno@dkrz.de

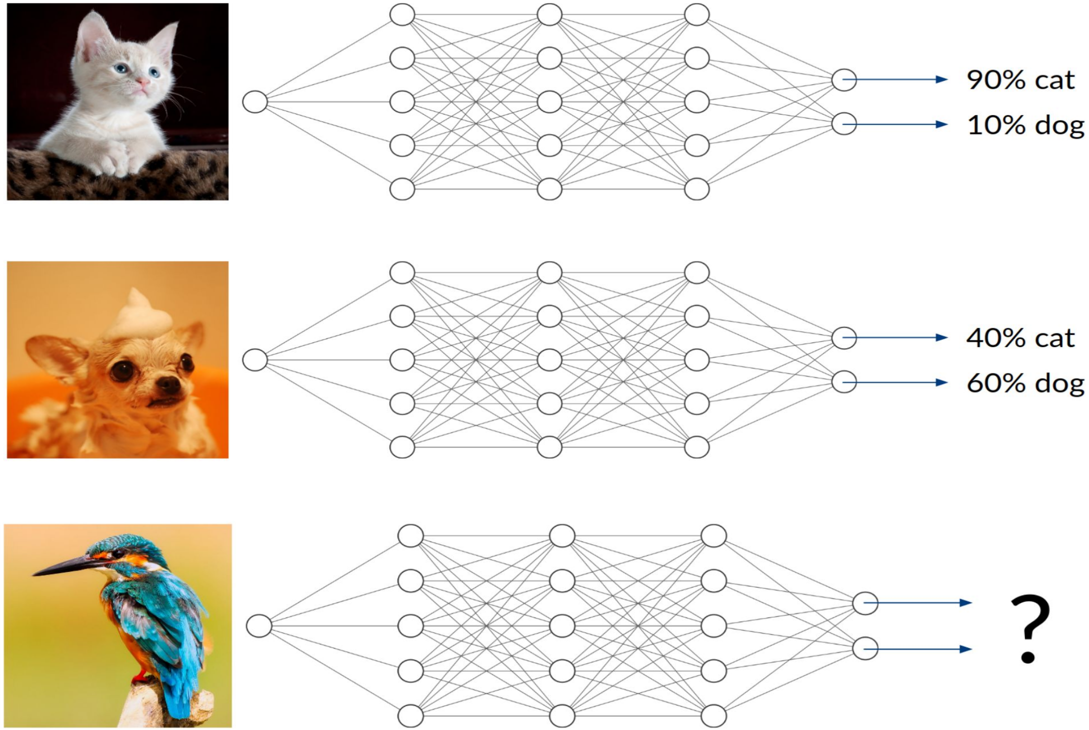
what if we show to the model a picture of something that it has never seen before?

moreno@dkrz.de



zero guarantee of a meaningful result (it can also be 40/60, for instance), but the algorithm always seem to be very confident!

moreno@dkrz.de



solution

± something

i.e., error bars,
confidence
intervals,...

moreno@dkrz.de

solutions: conformal predictors

Maria Navarro: Quantifying uncertainty in Machine Learning predictions | PyData...

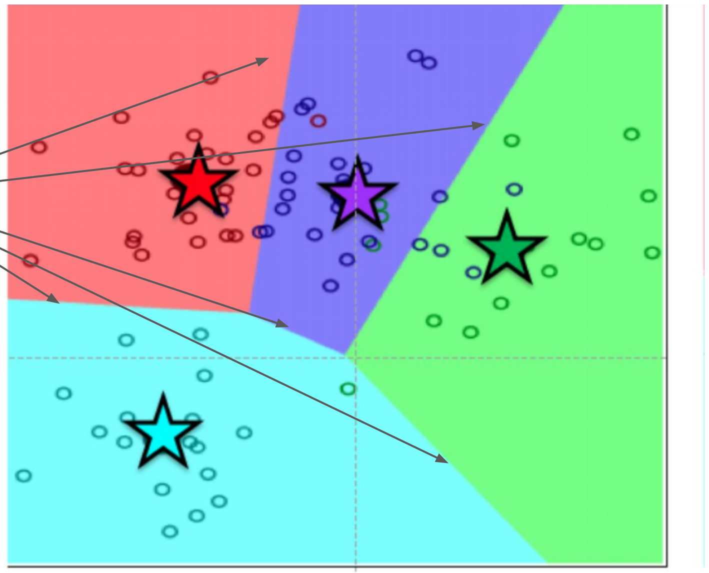
PyData • 1.3K views • 6 months ago

It produces a prediction region around the prediction that is agnostic about the noise distribution
For classification or regression and suitable for online assimilations
Assumption: samples are exchangeable
Library: nonconformist extension scikit-learn

moreno@dkrz.de

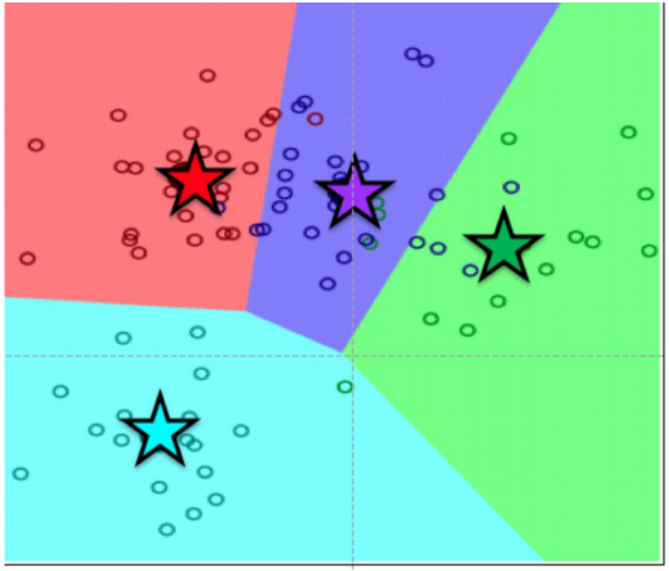
another classification example

DECISION
BOUNDARIES

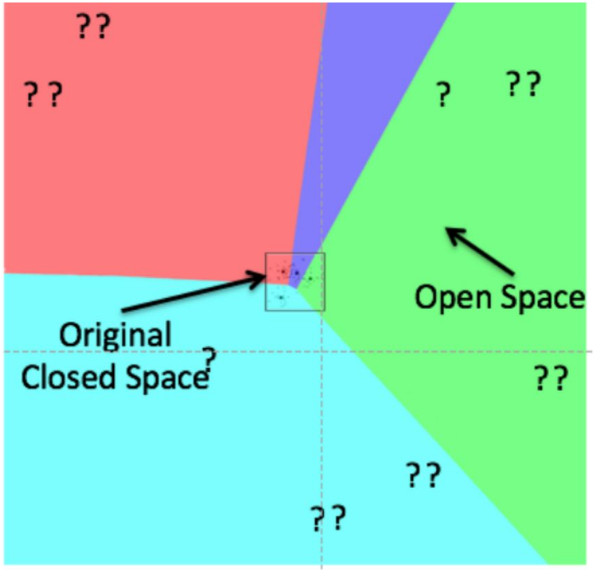


(a) Example four-class model from closed set point of view.

moreno@dkrz.de

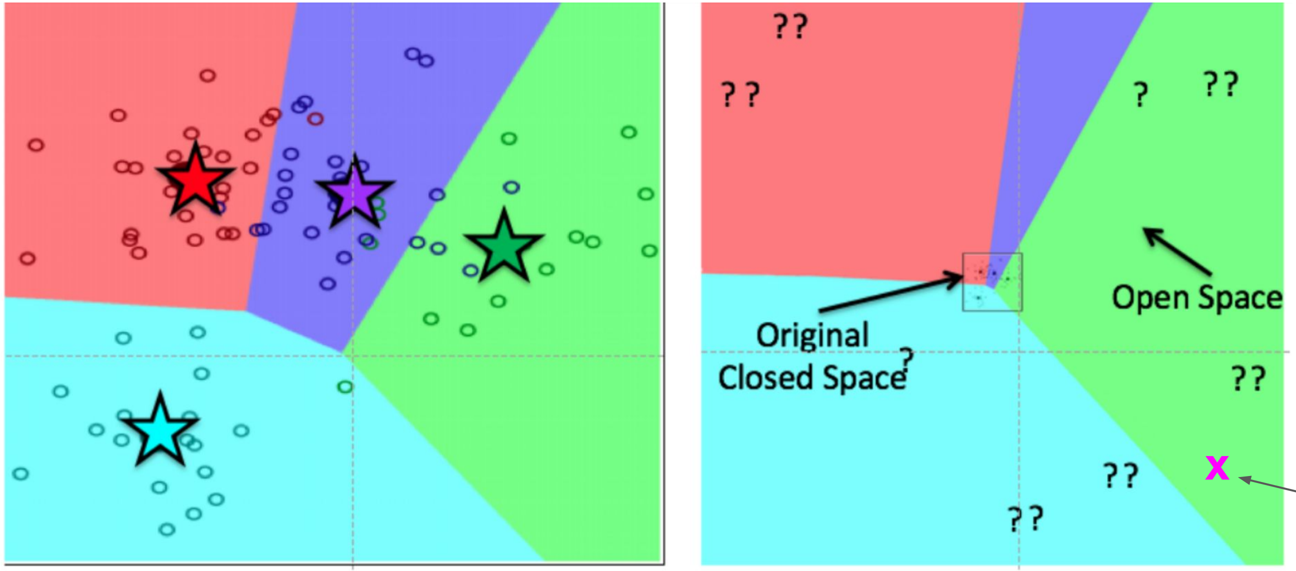


(a) Example four-class model from closed set point of view.



(b) Zooming out to show some open space.

moreno@dkrz.de



(a) Example four-class model from closed set point of view.

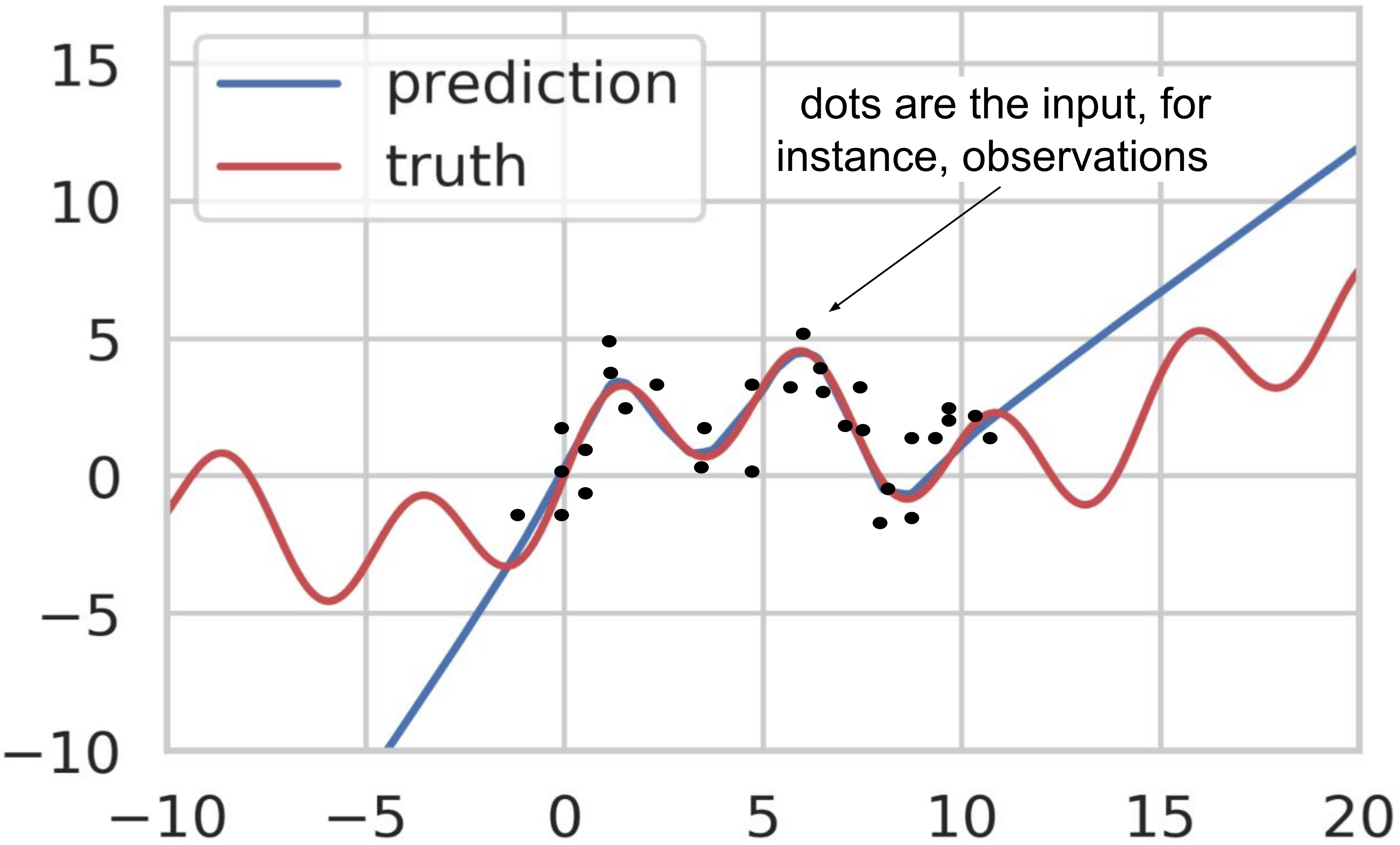
(b) Zooming out to show some open space.

a new input here will get a prediction too, even if the algorithm never saw anything near this position before

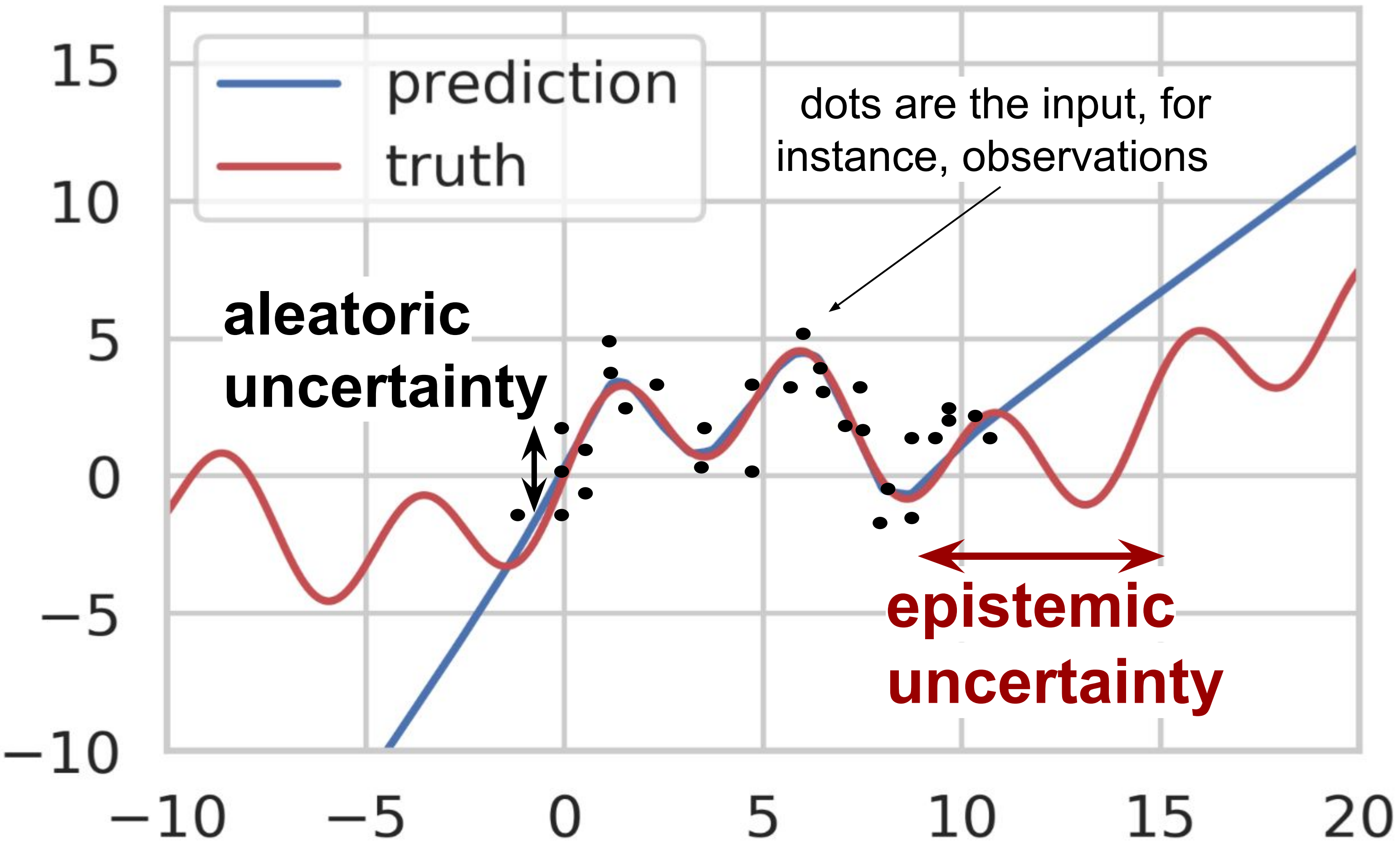
moreno@dkrz.de

a regression example

moreno@dkrz.de



moreno@dkrz.de



moreno@dkrz.de

Two types of uncertainty

Aleatoric: “what is the next outcome of tossing a coin?” it does not reduce with more input data, it is the noise in the data.

Epistemic: “How much do I believe the coin is fair?” it is related to the model’s belief after seeing the sample, it does reduce when having more data.

moreno@dkrz.de

solutions: Gaussian Processes, Monte Carlo dropout, deep ensembles, dropout ensembles, and quantile regression

Florian Wilhelm: Are you sure about that?! Uncertainty Quantification in AI | PyData...

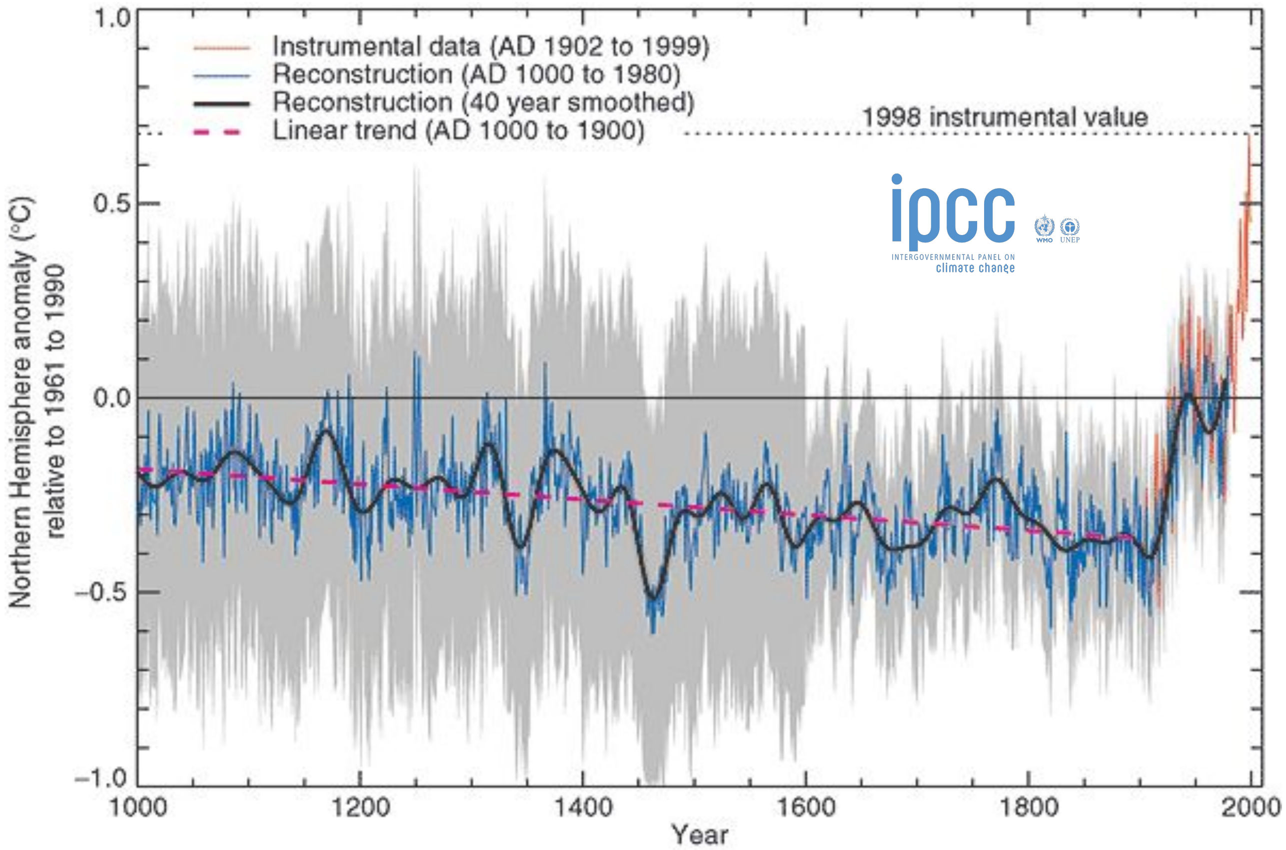
PyData • 162 views • 1 month ago

moreno@dkrz.de

Actually, there is a 3rd type of uncertainty:

Distribution shift: “Am I still flipping the same coin?” it is related to changes of the underlying quantity of interest, we assume that training and test data are i.i.d. from the same distribution but data drifta in time, or the labeller changed.

moreno@dkrz.de

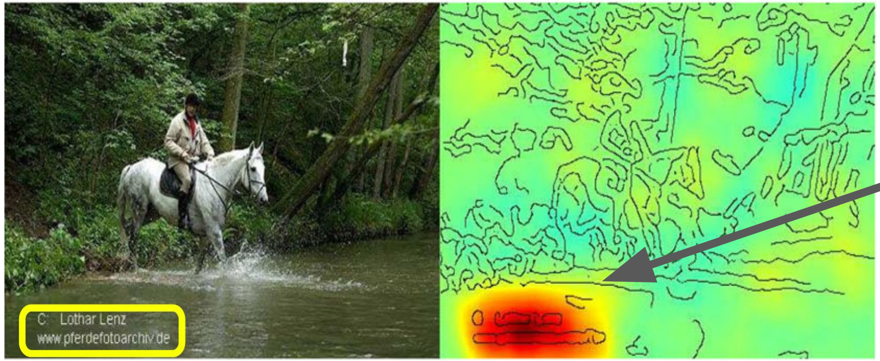


many problems in geoscience are not stationary

- training data are not longer representative if the system has changed
- the accuracy of the trained model definitely decreased under data shift

Fundamental problems (II): algorithms relying on spurious correlations (leakage)

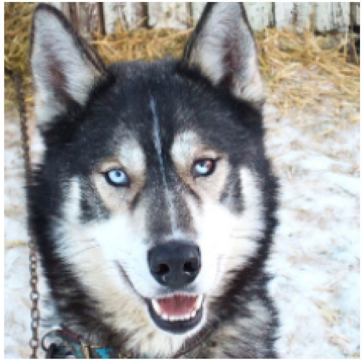
moreno@dkrz.de



Pixels area that the algorithm took as most relevant for the decision

Horse classified as a horse because the model learnt to read the image caption

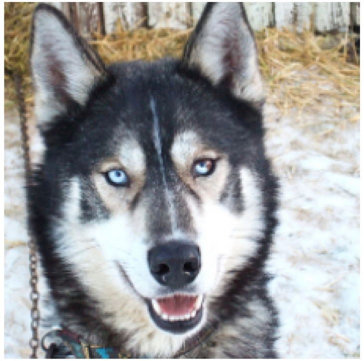
moreno@dkrz.de



(a) Husky classified as wolf

(b) Pixels area that the algorithm took as most relevant for the decision

moreno@dkrz.de



(a) Husky classified as wolf

(b) Pixels area that the algorithm took as most relevant for the decision

The algorithm was developed to distinguish wolves from huskies by exposing it to pictures of wolves and huskies but it just become an accurate snow identifier

moreno@dkrz.de

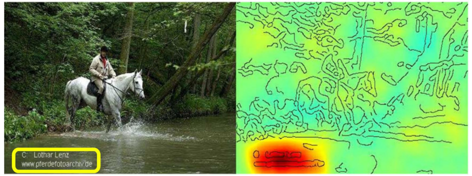
solution: Explainable Artificial Intelligence (XAI)

To explain black boxes decisions a posteriori in order to gain insights into the algorithm presumptions, biases, and reasoning.

XAI helps to determine “saliency”: to figure it out what part of the image was considered relevant

XAI also possible for time series and tabelled data, not only for images, there are many libraries

Layer-wise Relevance Propagation (LRP)

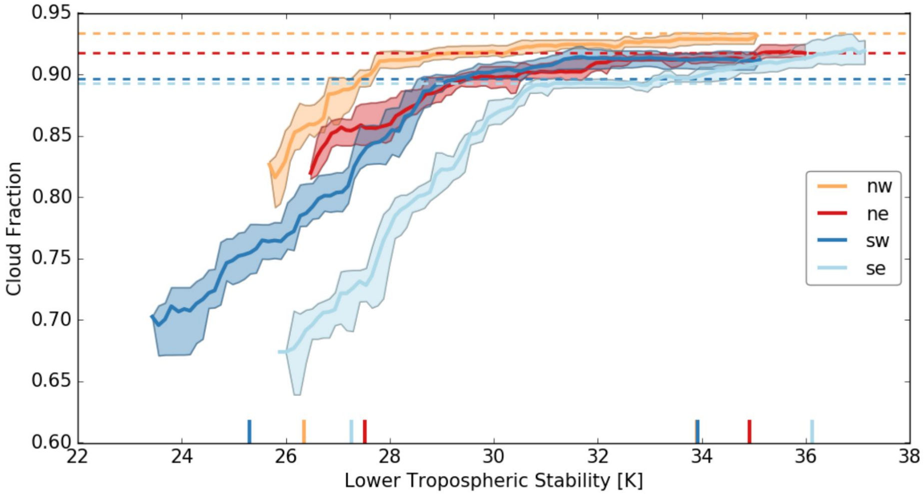


Local Interpretable Model-agnostic Explanation (LIME)

moreno@dkrz.de

more solutions to leakage:

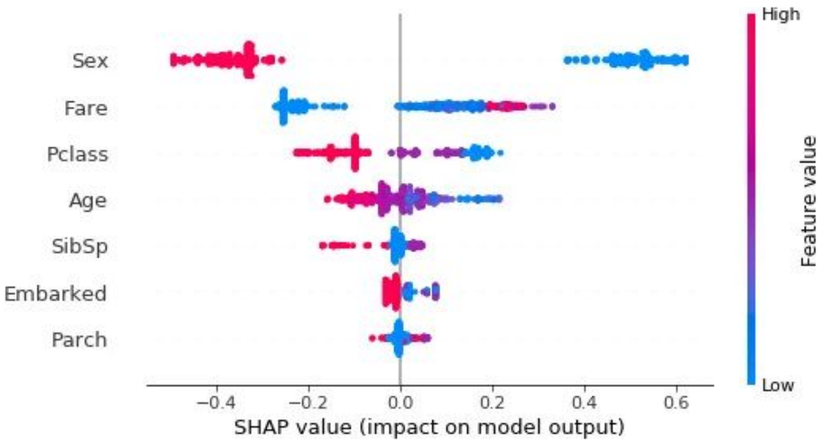
Partial Dependence Plot



Fuchs et al. 2018 (ACP)

Shapley values

```
# summarize the effects of all the features
shap.summary_plot(shap_values, X)
```



They are just sensitivity analysis
Easy to implement, many libraries: eli5, PDPBox,...

moreno@dkrz.de

warning

XAI techniques are not the ultimate solution: they rely on surrogate models, which bring their own assumptions, limitations, and are also error-prone, an interpretable model is always more trustable

RESEARCH-ARTICLE FREE ACCESS

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Authors: Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, Himabindu Lakkaraju

Authors Info & Affiliations

Publication: AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society • February 2020 • Pages 180–186 • https://doi.org/10.1145/3375627.3375830

moreno@dkrz.de

Fundamental problems (III): we are too optimistic (accuracy is not enough)

moreno@dkrz.de

Performance metrics to evaluate the algorithm skills

how often

(accuracy or true positives, precision, ROC AUC, confusion matrix,... used in classification)

or

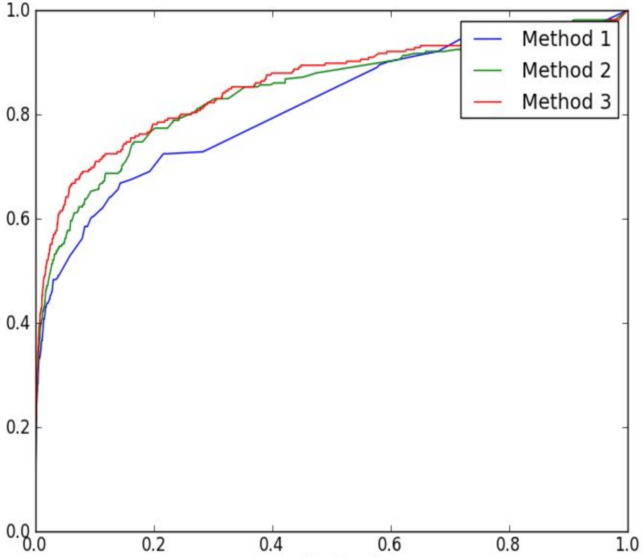
how well

(R^2, RMSE, log loss,... used in regression)

the predictions matched the correct target during the testing/validation phase.

Receiver Operating Characteristic curve (ROC)

TRUE POSITIVES (SENSITIVITY)



FALSE POSITIVES (1 - SPECIFICITY)

Libraries: sklearn.metrics, tf.keras.metrics,...

moreno@dkrz.de

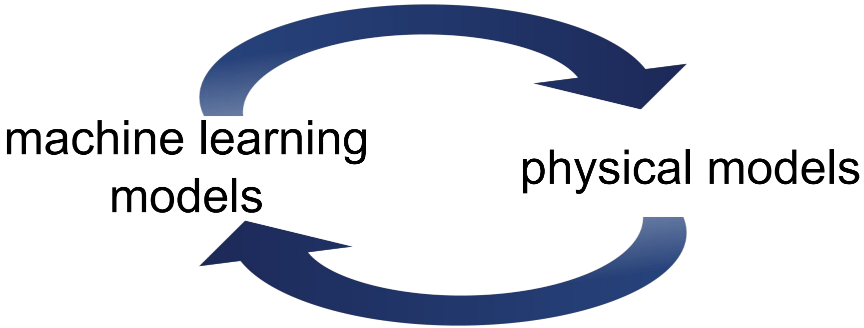
but to optimise your algorithm to achieve high accuracy is not enough, it might be more relevant **why** the model was correct than how much correct it is (remember the husky example, the model was very accurate, but in predicting snow!)

“We do not want a correct model, we want understanding”

(Doshi-Velez and Kim 2017)

moreno@dkrz.de

Best practices (I): Hybrid models



moreno@dkrz.de

machine learning models

physical models

Lightweighting/simplifying/speeding up physical models

- improve parametrizations
- analysis of model-observations mismatch
- emulation

machine learning models

physical models

Domain knowledge can guide/optimize the pure data-driven methods

- avoid inconsistencies
- design the architecture
- constrain the cost (or reward) function
- physically based data augmentation: expansion of the data set for undersampled regions

Example: lakes simulations to predict temperature from depth measurements

feature prediction

Depth (m)	Temp (°C)

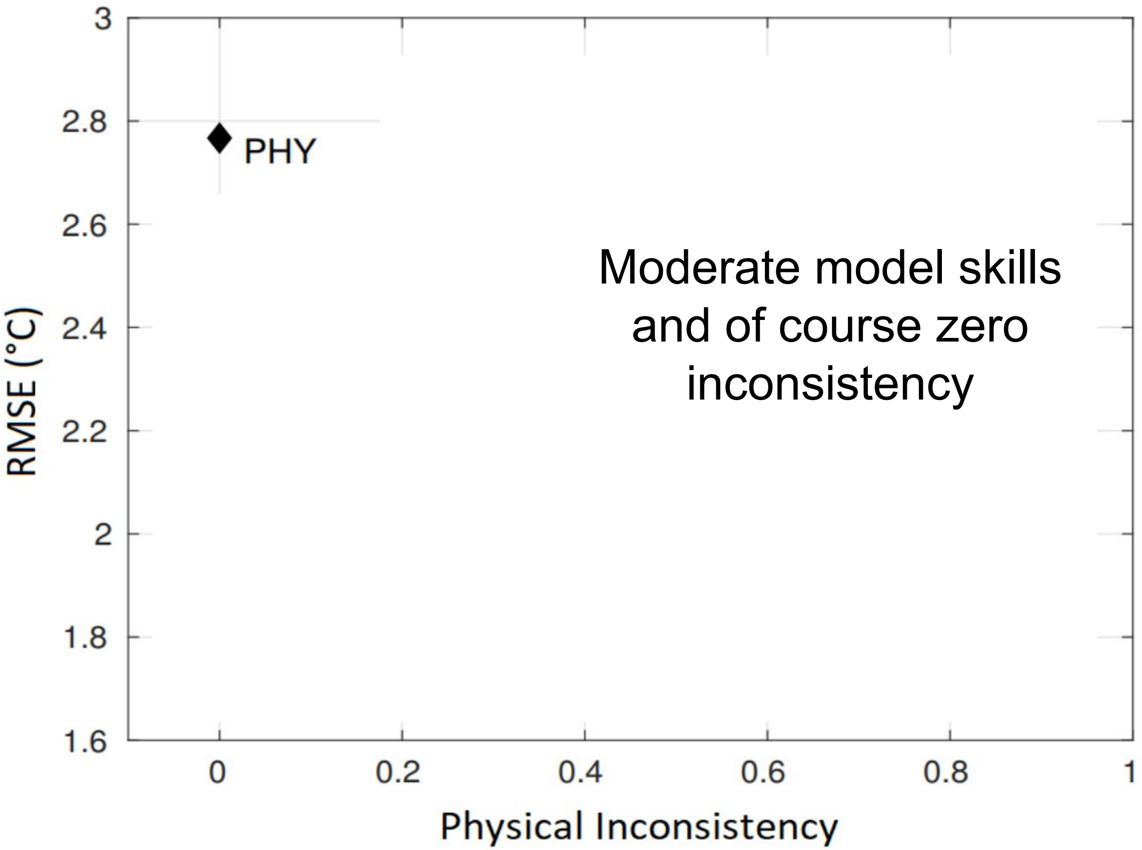
Physical model

example: $Temp_{d+1} = Temp_{d} + sun_{d} - wind_{d} - upwelling_{d}$

given that we measured $T_{d=surface} = 15°C$

feature prediction

Depth (m)	Temp (°C)



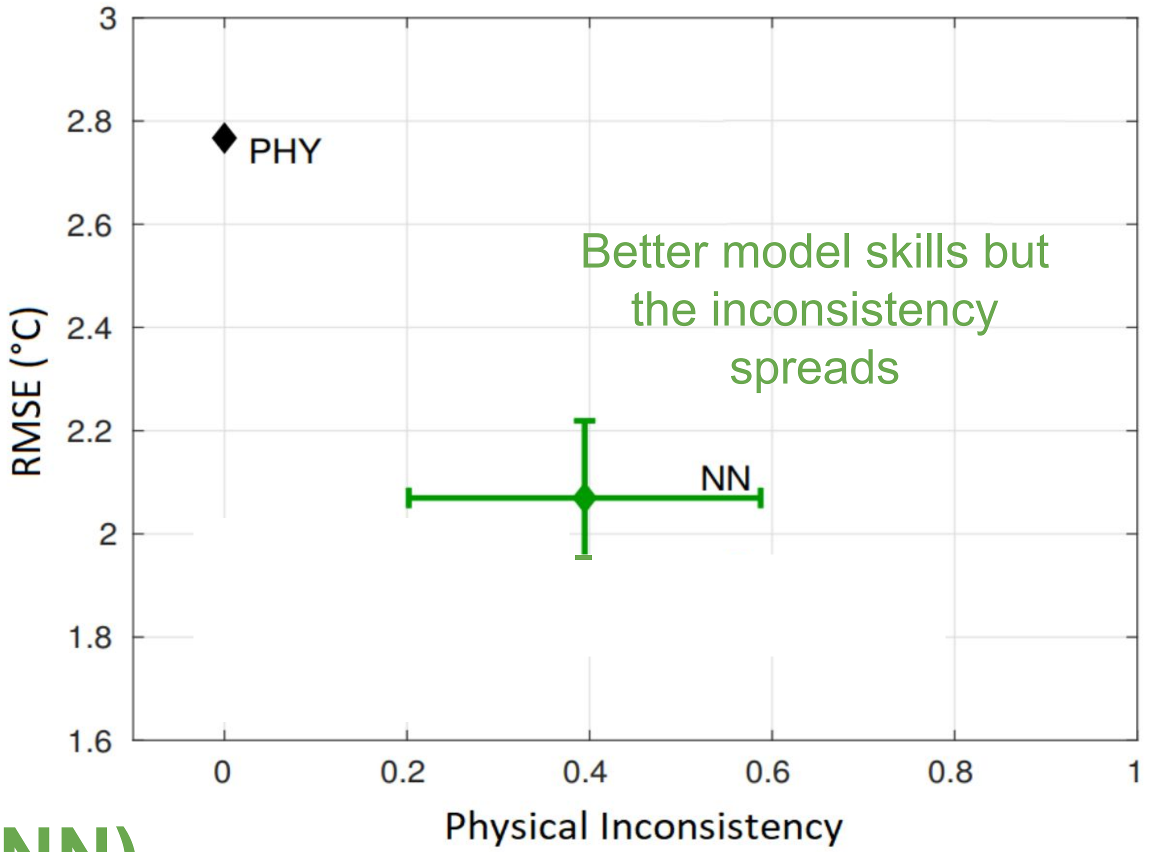
Moderate model skills and of course zero inconsistency

Physical model

example: $\text{Temp}_{d+1} = \text{Temp}_{d} + \text{sun}_{d} - \text{wind}_{d} - \text{upwelling}_{d}$
given that we measured $T_{d=\text{surface}} = 15°C$

feature prediction

Depth (m)	Temp (°C)



Neural Network (NN)

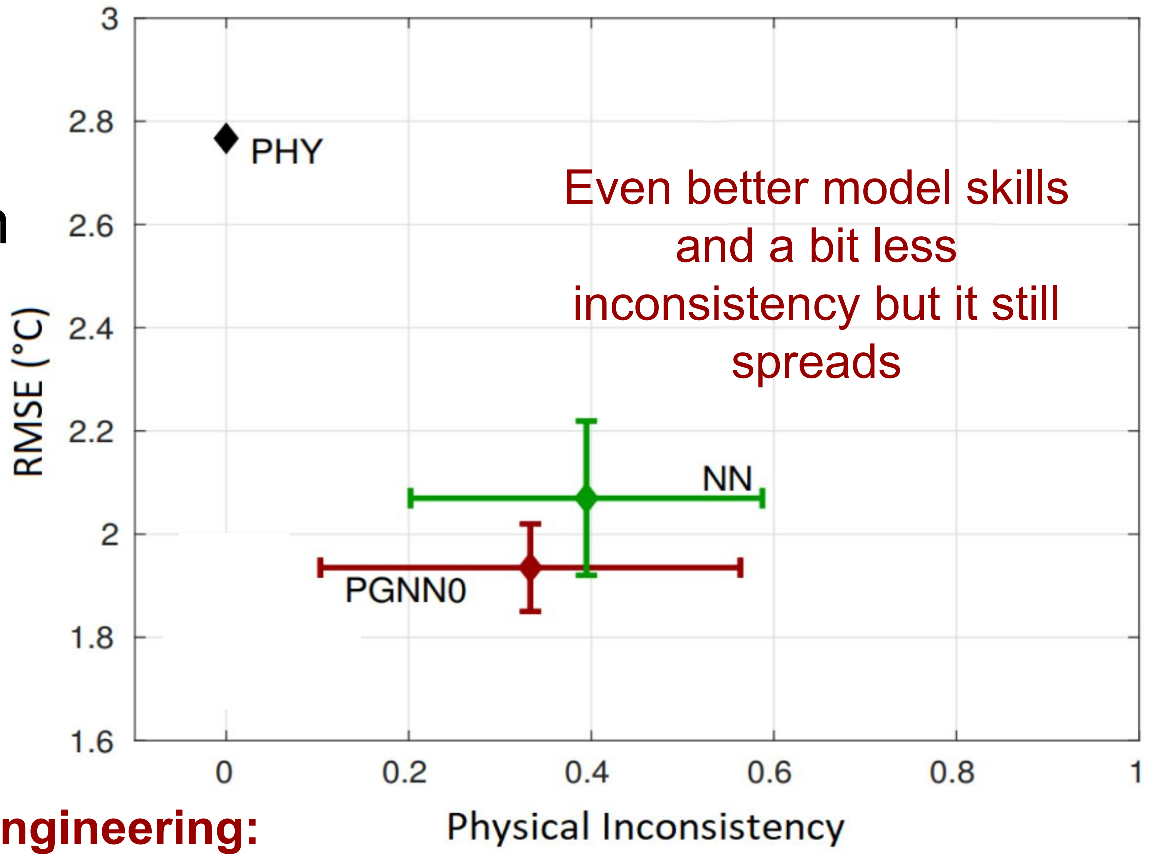
might allow negative densities and other inconsistencies (conservation laws)!

features prediction

Depth (m)	Density (g/L)	Temp (°C)

data augmentation/feature engineering: include new features driven by physical knowledge and then run the NN

features		prediction
Depth (m)	**Density (g/L)**	**Temp (°C)**



data augmentation/feature engineering: include new features driven by physical knowledge and then run the NN

features prediction

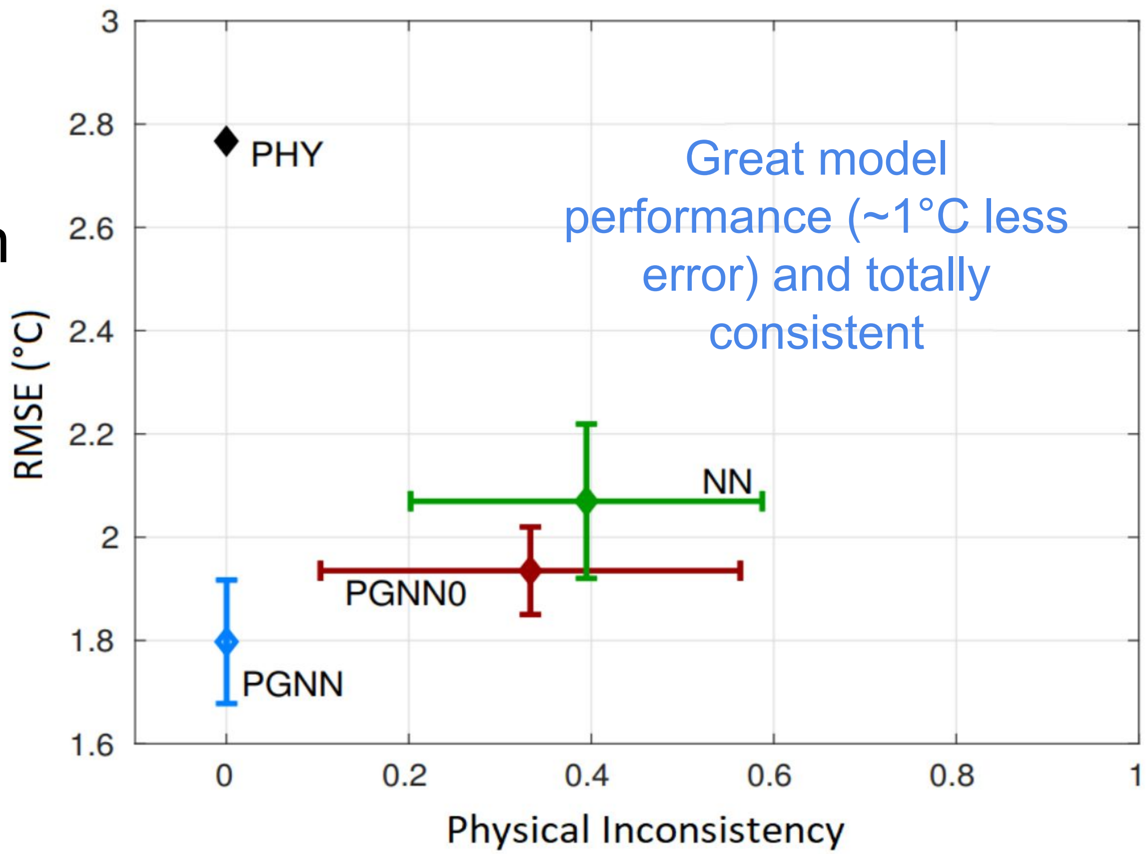
Depth (m)	Density (g/L)	Temp (°C)
		✓
		✗

physically driven feature + NN + constrain loss function: denser water must be deeper

Totally consistent and high model skills!

features prediction

Depth (m)	Density (g/L)	Temp (°C)
		✓
		✗



physically driven feature + NN + constrain loss function: denser water must be deeper

Best practices (II):
put your model on diet

moreno@dkrz.de

Put your model on diet before the training to prevent leakage

- identify and remove snow (see LIME example), captions (see LRP example),...
- most neural networks are over-parameterized. Many trained weights have little impact on overall accuracy and can be removed, it is called pruning, use techniques like MC dropouts

moreno@dkrz.de

Best practices (III):
call a human!

moreno@dkrz.de

Calculate the confidence with uncertainty quantification techniques (see previous slides)

- conformal predictors
- MC dropouts
- Deep Ensembles
- Quantile regression
- ...

and implement fallbacks if the confidence of the prediction is low.

moreno@dkrz.de

Just great!

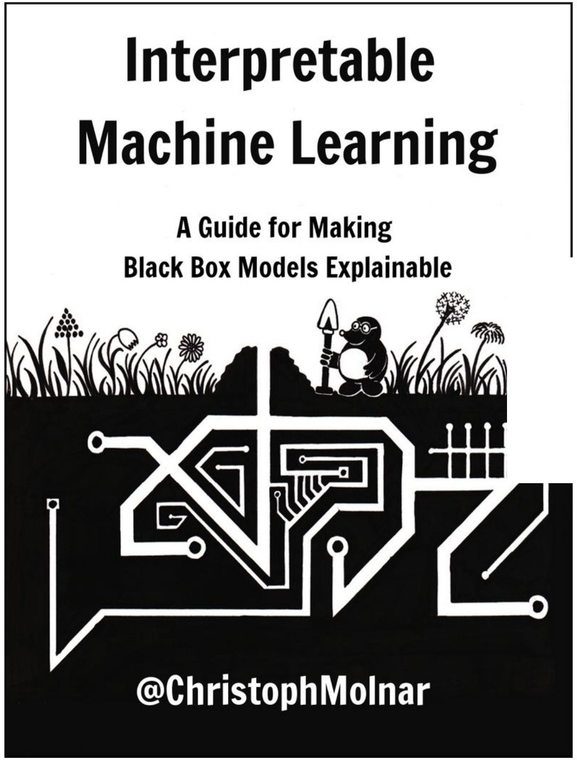
Vincent Warmerdam: How to Constrain Artificial Stupidity | PyData London 2019

PyData • 3K views • 6 months ago

GOTO 2018 • Computers are Stupid: Protecting "AI" from Itself • Katharine Jarmul

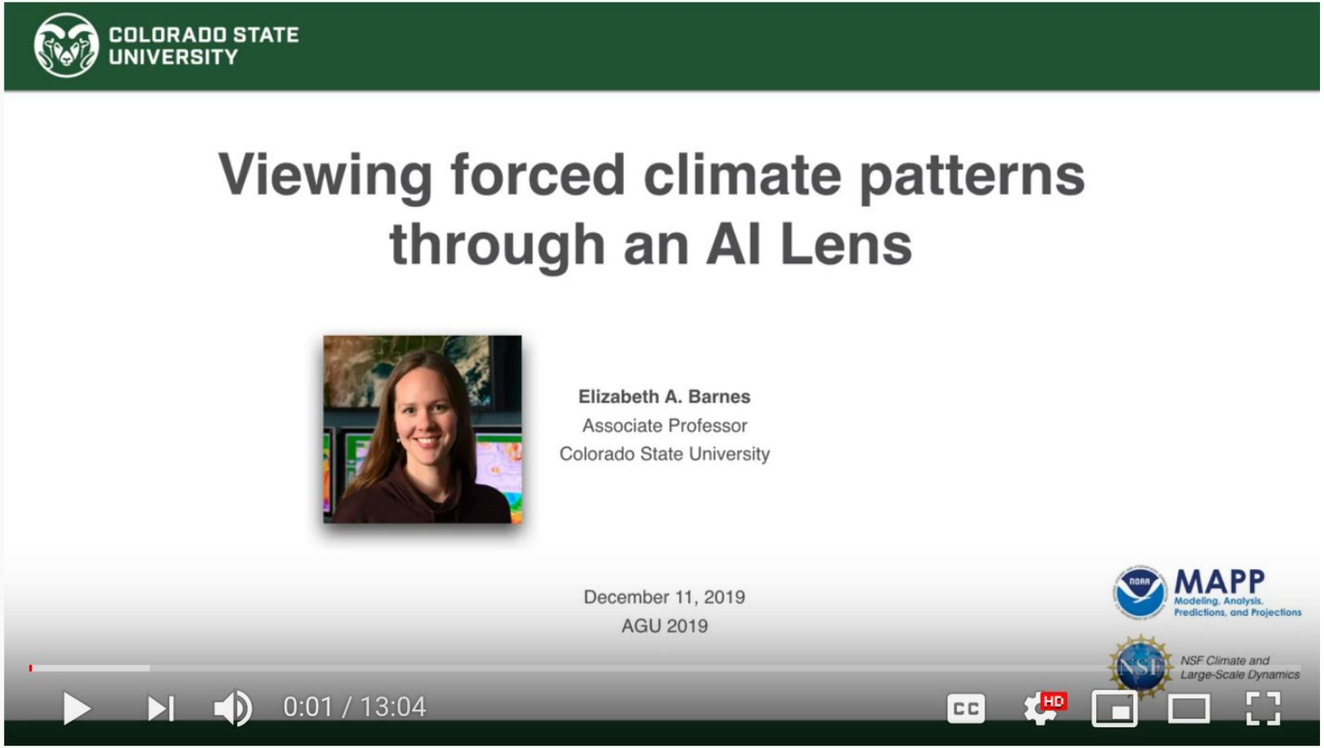
GOTO Conferences 1.3K views • 12 months ago

moreno@dkrz.de



moreno@dkrz.de

Explainability added vaue



"Viewing Forced Climate Patterns through an AI Lens", Dec. 11, 2019.

moreno@dkrz.de

References

- Cat, dog, and bird and UQ: https://www.inovex.de/blog/uncertainty-quantification-deep-learning/
- ROC curve: https://stats.stackexchange.com/questions/15127/evaluating-and-combining-methods-based-on-roc-and-pr-curves
- Phase space https://vast.uccs.edu/~tboult/PAPERS/Learning_and_the_Unknown_Surveying_Steps_Toward_Open_World_Recognition_AAAI19.pdf
- Epistemic uncertaity http://yingzhenli.net/home/pdf/epistemic_uncertainty_neurips_bdl2019.pdf
- UQ under data shift http://bayesiandeeplearning.org/2019/slides/Jasper%20Snoek.pdf
- Horse and LRP: https://www.nature.com/articles/s41467-019-08987-4
- Husky vs Wolf and LIME: https://arxiv.org/pdf/1602.04938.pdf
- Feature importance, partial dependence plots, and individual conditional expectation https://www.kaggle.com/learn/machine-learning-explainability
- Physics-guided neural networks : https://arxiv.org/pdf/1710.11431.pdf and https://towardsdatascience.com/physics-guided-neural-networks-pgnns-8fe9dbad9414

Libraries

conformal predictors https://github.com/donlnz/nonconformist
eli5 https://eli5.readthedocs.io/en/latest/
PDPBox https://pdpbox.readthedocs.io/en/latest/
SHAP https://github.com/slundberg/shap
LIME https://github.com/marcotcr/lime
LRP https://github.com/atulshanbhag/Layerwise-Relevance-Propagation

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

IS-ENES3 is funded by the EU Hor
Research and Innovation program
grant agreement No 824084

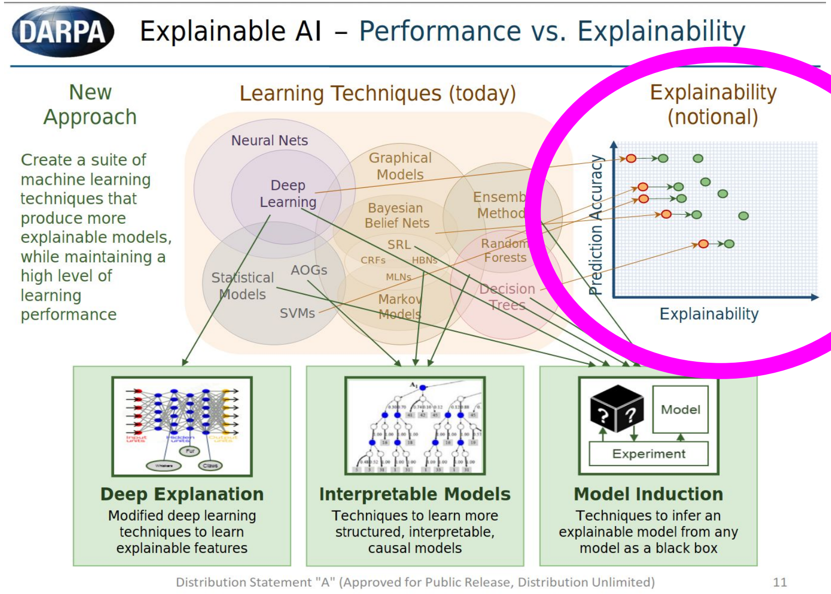
moreno@dkrz.de

Bonus track: do we lose performance?

Please Stop Doing "Explainable" ML - Cynthia Rudin

The Berkman Klein Center for Internet & Society •

1.3K views • 8 months ago



https://www.cc.gatech.edu/~alanwags/DLAI2016/%28Gunning%29%20IJCAI-16%20DLAI%20WS.pdf