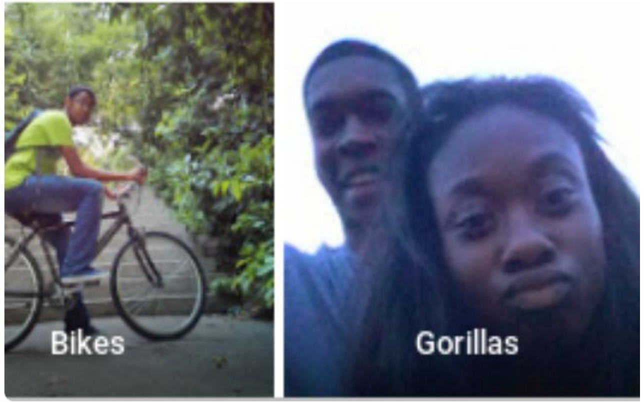
6th ENES HPC Workshop 2020

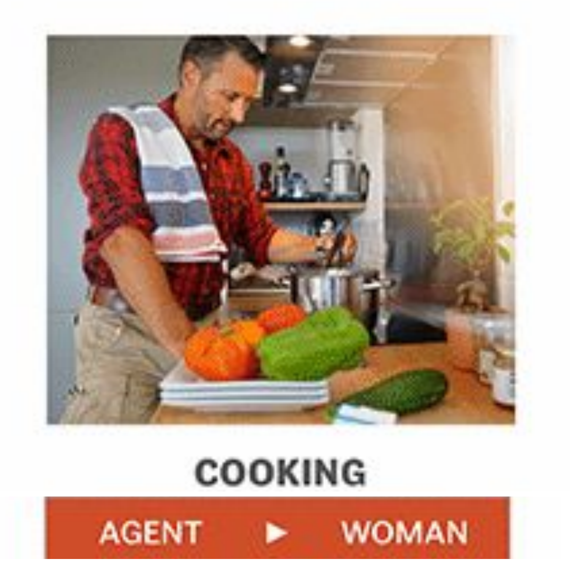
Maria Moreno de Castro
moreno@dkrz.de

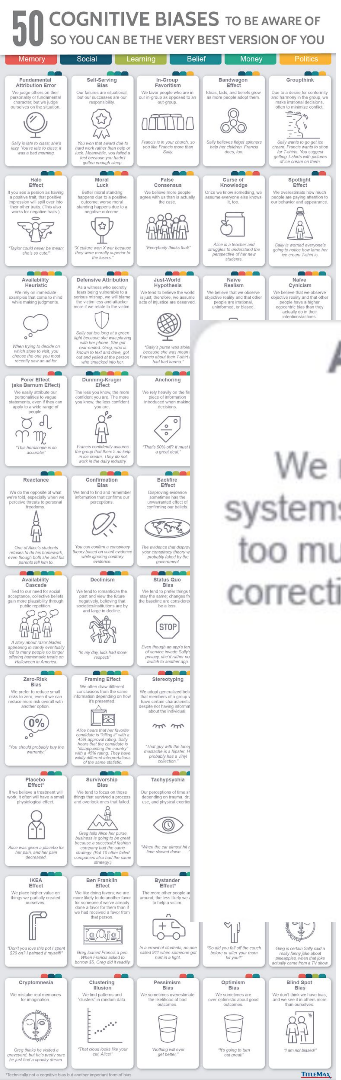
Interpretable Machine Learning

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

ML algorithms encrypt and magnify bias







50 COGNITIVE BIASES TO BE AWARE OF SO YOU CAN BE THE VERY BEST VERSION OF YOU

Memory | Social | Learning | Belief | Money | Politics

Fundamental Attribution Error
We judge others on their personality or fundamental character, but we judge ourselves on the situation.
Sally is late to class; she's lazy. You're late to class; it was a bad morning.

Self-Serving Bias
Our failures are situational, but our successes are our responsibility.
You won that award due to hard work rather than help or luck. Meanwhile, you failed a test because you hadn't gotten enough sleep.

In-Group Favoritism
We favor people who are in our in-group as opposed to an out-group.
Francis is in your church, so you like Francis more than Sally.

Bandwagon Effect
Ideas, fads, and beliefs grow as more people adopt them.
Sally believes fidget spinners help her children. Francis does, too.

Groupthink
Due to a desire for conformity and harmony in the group, we make irrational decisions, often to minimize conflict.
Sally wants to go get ice cream. Francis wants to shop for T-shirts. You suggest getting T-shirts with pictures of ice cream on them.

Halo Effect
If you see a person as having a positive trait, that positive impression will spill over into their other traits. (This also works for negative traits.)
"Taylor could never be mean; she's so cute!"

Moral Luck
Better moral standing happens due to a positive outcome; worse moral standing happens due to a negative outcome.
"X culture won X war because they were morally superior to the losers."

False Consensus
We believe more people agree with us than is actually the case.
"Everybody thinks that!"

Curse of Knowledge
Once we know something, we assume everyone else knows it, too.
Alice is a teacher and struggles to understand the perspective of her new students.

Spotlight Effect
We overestimate how much people are paying attention to our behavior and appearance.
Sally is worried everyone's going to notice how lame her ice cream T-shirt is.

Availability Heuristic
We rely on immediate examples that come to mind while making judgments.
When trying to decide on which store to visit, you choose the one you most recently saw an ad for.

Defensive Attribution
As a witness who secretly fears being vulnerable to a serious mishap, we will blame the victim less and attacker more if we relate to the victim.
Sally sat too long at a green light because she was playing with her phone. She got rear-ended. Greg, who is known to text and drive, got out and yelled at the person who smacked into her.

Just-World Hypothesis
We tend to believe the world is just; therefore, we assume acts of injustice are deserved.
"Sally's purse was stoler because she was mean t Francis about their T-shirt a had bad karma."

Naïve Realism
We believe that we observe objective reality and that other people are irrational, uninformed, or biased.

Naïve Cynicism
We believe that we observe objective reality and that other people have a higher egocentric bias than they actually do in their intentions/actions.

Forer Effect (aka Barnum Effect)
We easily attribute our personalities to vague statements, even if they can apply to a wide range of people.
"This horoscope is so accurate!"

Dunning-Kruger Effect
The less you know, the more confident you are. The more you know, the less confident you are.
Francis confidently assures the group that there's no kelp in ice cream. They do not work in the dairy industry.

Anchoring
We rely heavily on the firs piece of information introduced when making decisions.
"That's 50% off? It must b a great deal."

Reactance
We do the opposite of what we're told, especially when we perceive threats to personal freedoms.
One of Alice's students refuses to do his homework, even though both she and his parents tell him to.

Confirmation Bias
We tend to find and remember information that confirms our perceptions.
You can confirm a conspiracy theory based on scant evidence while ignoring contrary evidence.

Backfire Effect
Disproving evidence sometimes has the unwarranted effect of confirming our beliefs.
The evidence that disprov your conspiracy theory w probably faked by the government.

Availability Cascade
Tied to our need for social acceptance, collective beliefs gain more plausibility through public repetition.
A story about razor blades appearing in candy eventually led to many people no longer offering homemade treats on Halloween in America.

Declinism
We tend to romanticize the past and view the future negatively, believing that societies/institutions are by and large in decline.
"In my day, kids had more respect!"

Status Quo Bias
We tend to prefer things t stay the same; changes fro the baseline are considere be a loss.
Even though an app's terr of service invade Sally's privacy, she'd rather no switch to another app.

Zero-Risk Bias
We prefer to reduce small risks to zero, even if we can reduce more risk overall with another option.
"You should probably buy the warranty."

Framing Effect
We often draw different conclusions from the same information depending on how it's presented.
Alice hears that her favorite candidate is "killing it" with a 45% approval rating. Sally hears that the candidate is "disappointing the country" with a 45% rating. They have wildly different interpretations of the same statistic.

Stereotyping
We adopt generalized beli that members of a group w have certain characteristic despite not having informat about the individual.
"That guy with the fancy mustache is a hipster. H probably has a vinyl collection."

Placebo Effect*
If we believe a treatment will work, it often will have a small physiological effect.
Alice was given a placebo for her pain, and her pain decreased.

Survivorship Bias
We tend to focus on those things that survived a process and overlook ones that failed.
Greg tells Alice her purse business is going to be great because a successful fashion company had the same strategy. (But 10 other failed companies also had the same strategy.)

Tachypsychia
Our perceptions of time sh depending on trauma, dru use, and physical exertio
"When the car almost hit m time slowed down … ."

IKEA Effect
We place higher value on things we partially created ourselves.
"Don't you love this pot I spent $20 on? I painted it myself!"

Ben Franklin Effect
We like doing favors; we are more likely to do another favor for someone if we've already done a favor for them than if we had received a favor from that person.
Greg loaned Francis a pen. When Francis asked to borrow $5, Greg did it readily.

Bystander Effect*
The more other people ar around, the less likely we a to help a victim.
In a crowd of students, no one called 911 when someone got hurt in a fight.

"So did you fall off the couch before or after your mom hit you?"

Greg is certain Sally said a really funny joke about pineapples, when that joke actually came from a TV show.

Cryptomnesia
We mistake real memories for imagination.
Greg thinks he visited a graveyard, but he's pretty sure he just had a spooky dream.

Clustering Illusion
We find patterns and "clusters" in random data.
"That cloud looks like your cat, Alice!"

Pessimism Bias
We sometimes overestimate the likelihood of bad outcomes.
"Nothing will ever get better."

Optimism Bias
We sometimes are over-optimistic about good outcomes.
"It's going to turn out great!"

Blind Spot Bias
We don't think we have bias, and we see it in others more than ourselves.
"I am not biased!"

*Technically not a cognitive bias but another important form of bias

TITLEMAX

IKEA Effect

We place higher value on things we partially created ourselves.

Automation Bias

We rely on automated systems, sometimes trusting too much in the automated correction of actually correct decisions.

Law of Triviality (aka "Bike-Shedding")

We give disproportionate weight to trivial issues, often while avoiding more complex issues.

Blind Spot Bias

We don't think we have bias, and we see it in others more than ourselves.

"I am not biased!"

in geosciences…

- large spread in climate prediction is often nowadays attributed to cloud radiative effects → resolving clouds seems to be the way to go to reduce epistemic uncertainty, however the ecosystem modelling community often expresses that the GCM are biased because they do not resolve phytoplankton and in general underrepresent the biological component

- geographical areas we do not have as much observational data, modelling groups,...

- use ensemble means as it everything would be normally distributed

- use deterministic solvers for differential equations with stochastic components

- our physical models are built in blocks by different people at different institutions at different times, with tons of simplifications, approximations, assumptions, and empirical parameters and to make all work together we use tuning parameters which do not have physical meaning and introduce compensation errors

ML algorithms can easily break the

the fundamental laws of physics

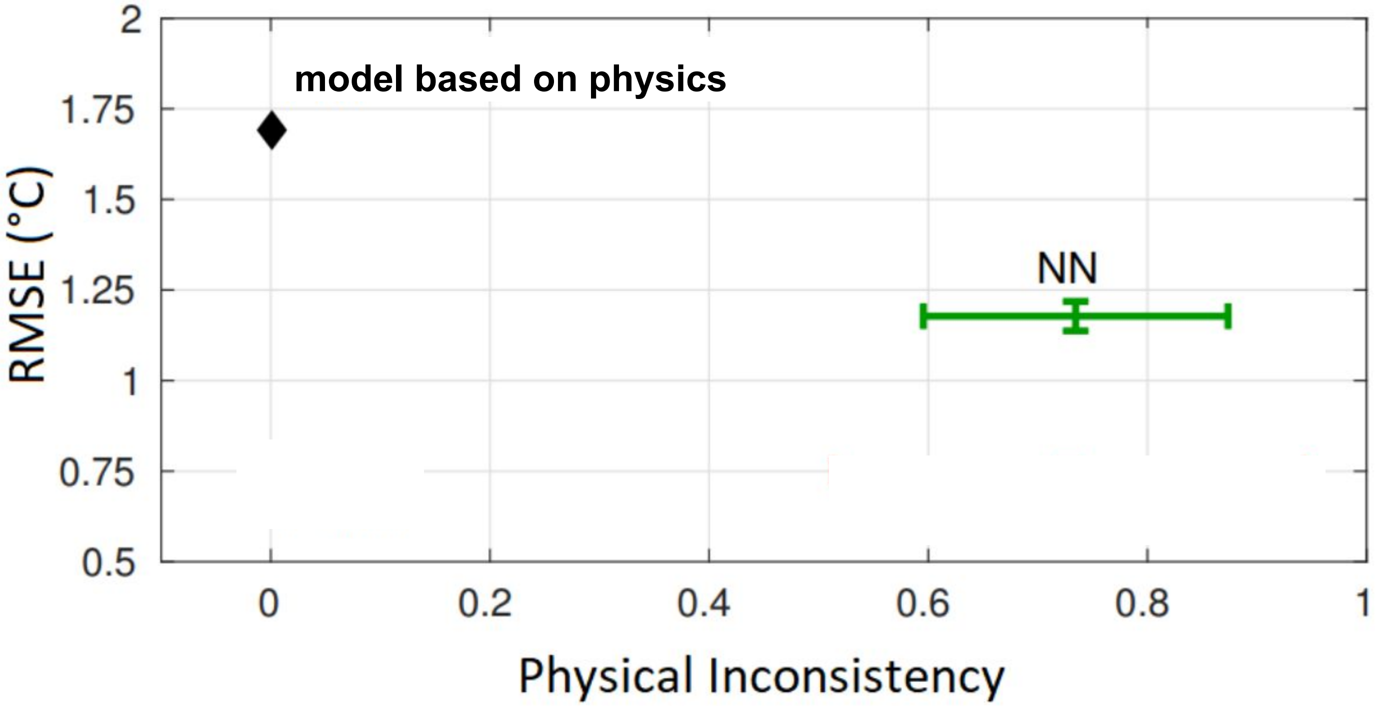
energy and mass conservations, nonnegative densities, precipitations,...

Noether's theorem explains why conservation laws exists (wikipedia)

Example:

lakes simulations

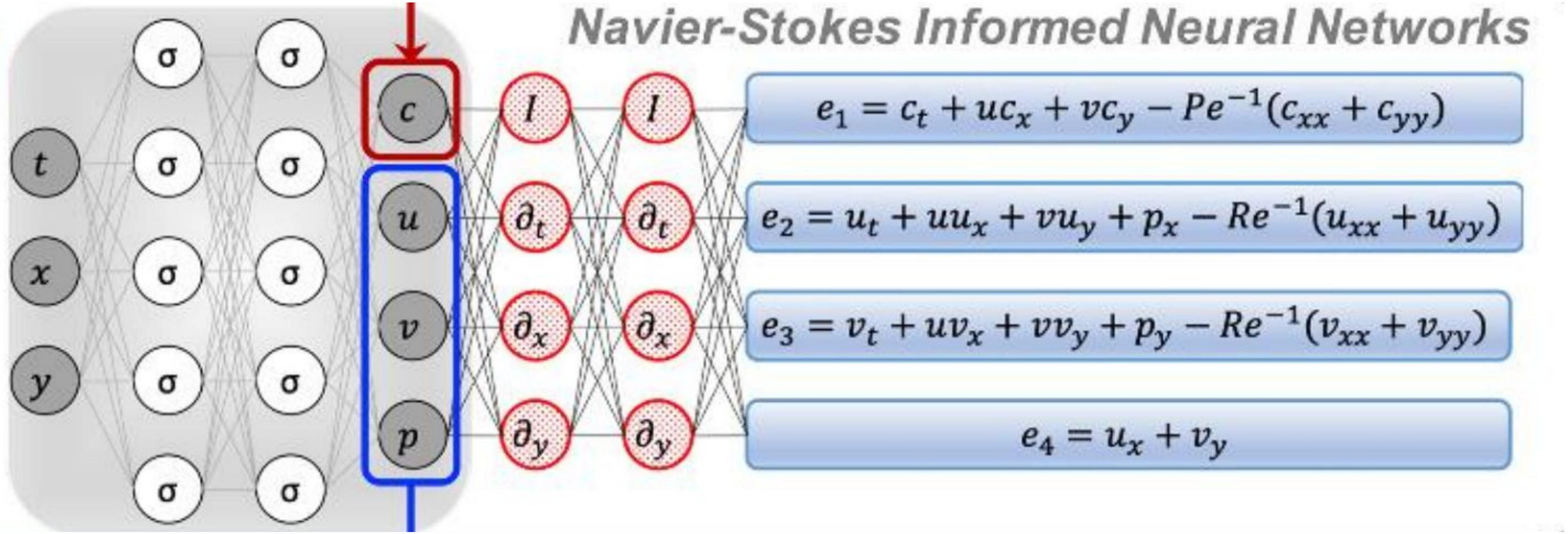
feature	prediction
Depth (m)	**Temp (°C)**



Results on Lake Mille Lacs

Example:

fluid dynamics


Navier-Stokes Informed Neural Networks
t
x
y
σ
c
u
v
p
I
∂_t
∂_x
∂_y
$e_1 = c_t + uc_x + vc_y - Pe^{-1}(c_{xx} + c_{yy})$
$e_2 = u_t + uu_x + vu_y + p_x - Re^{-1}(u_{xx} + u_{yy})$
$e_3 = v_t + uv_x + vv_y + p_y - Re^{-1}(v_{xx} + v_{yy})$
$e_4 = u_x + v_y$

- **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 so it is hard to spot when these models are biased.

Examples: Random forest, NN,...

- **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 so it is hard to spot when these models are biased.

Examples: Random forest, NN,...

- **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.

- **Explainable models**

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

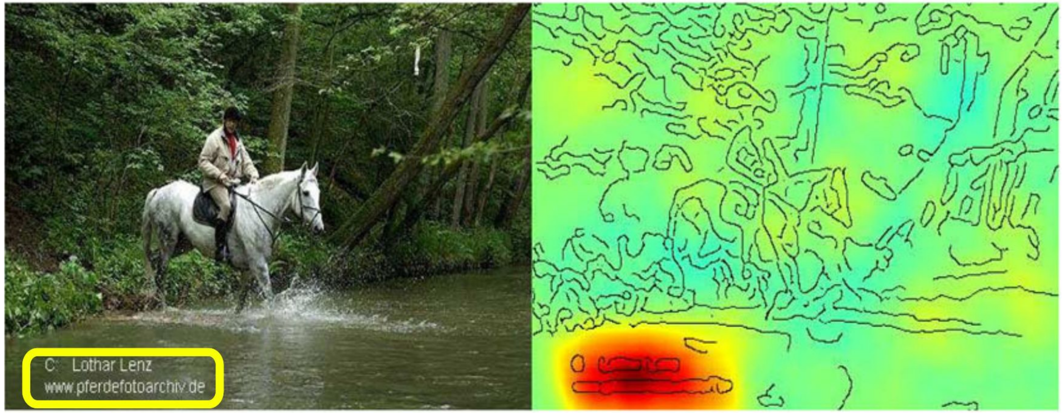
- **Explainable models**

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

- sensitivity analysis based on observing the effect of perturbations on model components
- identify what features or feature values contributed the most to the predictions (often presented in saliency maps)
- for every model, many available libraries (skitlearn, Tensorflow,... or extensions)

✓ essential ✓ feasible ✓ add scientific value

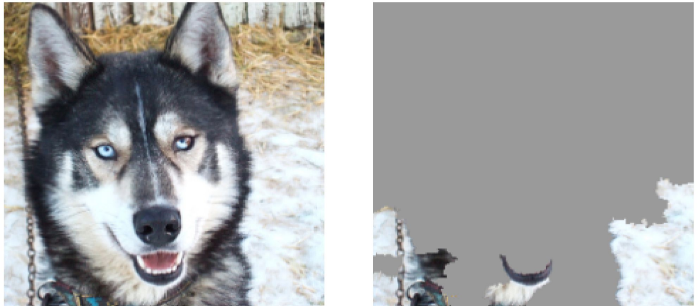
Layer-wise
Relevance Propagation (LRP)



single prediction, run again a back propagation that **tracks** what was activated and how much (nice but **costly**)

https://github.com/atulshanbhag/Layerwise-Relevance-Propagation

Local Interpretable Model-agnostic Explanation (LIME)



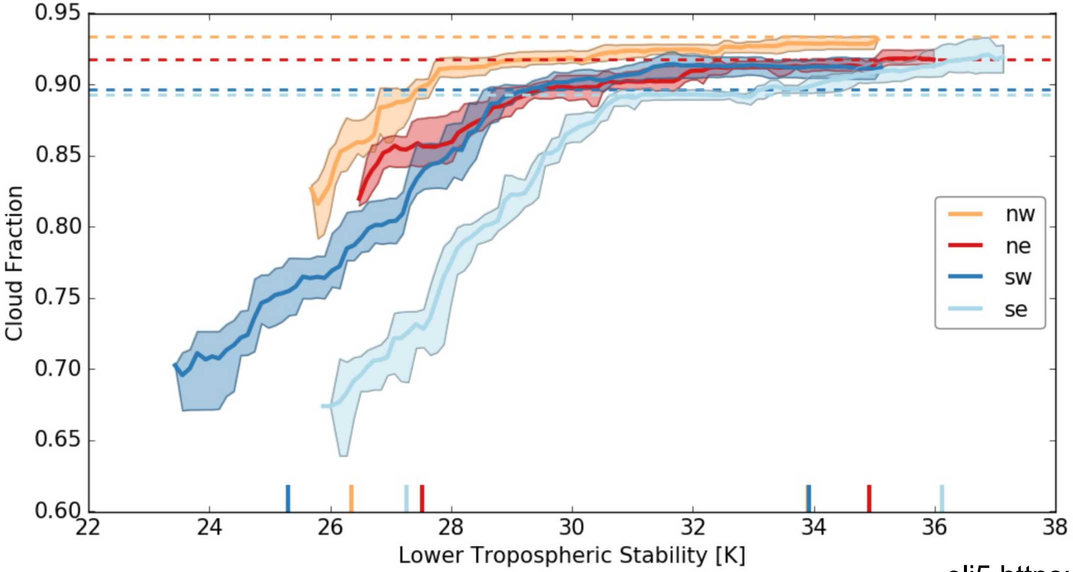
single prediction, run an interpretable model with **the black-box prediction as target** (**superpixels**)

https://github.com/marcotcr/lime, for time series: https://github.com/emanuel-metzenthin/Lime-For-Time

Feature importance

Permutation importance: for any supervised model, just **shuffle** a feature of the validation data and compare performance (what **features values** are important?)

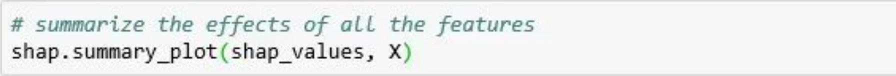
Partial dependence plots: 1. select feature 2. define a grid 3. per grid value: a) **replace feature with grid value** and b) average prediction 4. draw curve. (**cancelation effects? how to know the direction of the effect?**)



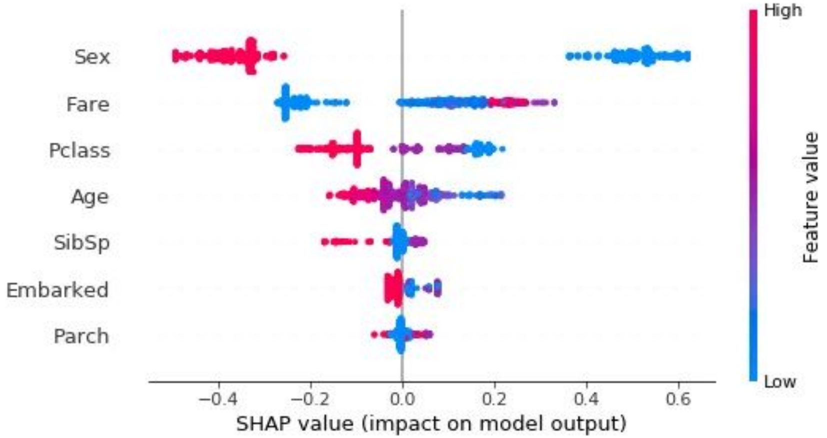
Fuchs et al. 2018 (ACP)

eli5 https://eli5.readthedocs.io/en/latest/,
PDPBox https://pdpbox.readthedocs.io/en/latest/,...

Shapley Additive exPlanation (SHAP)



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



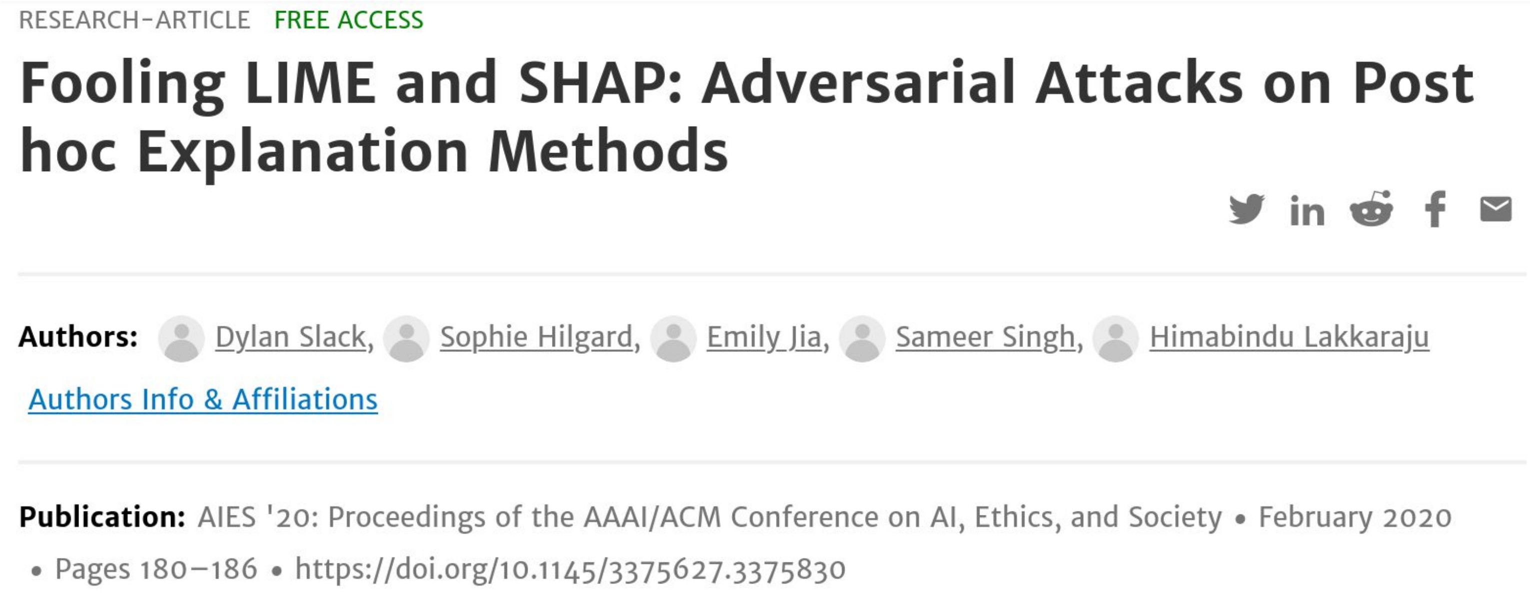
1. single prediction, based on game theory, a prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction of that instance is the “payout”, in a linear world is easy, just adding the contributions, but here **coalitions, synergies, direction of the effect**,... are considered.

2. do it for all of the prediction to get the global shapley values (**costly**)

https://github.com/slundberg/shap

warning

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



RESEARCH-ARTICLE FREE ACCESS

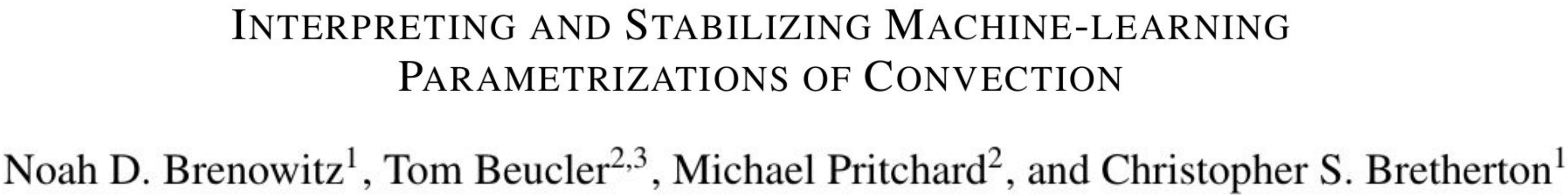
Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

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

Parameterization schemes



Interpreting and Stabilizing Machine-learning Parametrizations of Convection

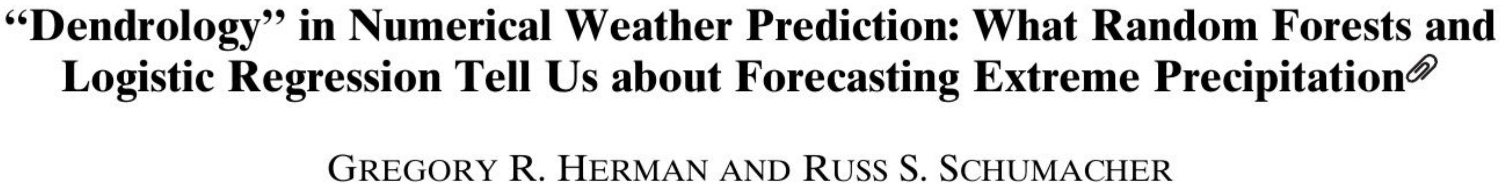
Noah D. Brenowitz[1], Tom Beucler[2,3], Michael Pritchard[2], and Christopher S. Bretherton[1]

Predicting the behavior of NNs is tied to the difficult problem of interpreting NN emulators of physical processes. [...] How can we tailor ML interpretability techniques [...] for the particular purpose of interpreting NN parameterizations of convection?

1) Partial dependence plots to tests the nonlinear sensitivity of a single ML parametrization to systematic changes in its inputs

2) Linear Response Functions (LRF) or saliency map extend the analysis to the full input space of a parameterization. In a previous paper they computed LRFs to analyze what was causing their NN parameterizations to produce unstable dynamics when coupled to a GCM → reducing the potential for spurious causality by ablating both the upper atmospheric temperature and humidity from the input features of an NN parameterization results in a stable scheme [...] which demonstrates that ML interpretability techniques have already significantly aided the development of ML parameterizations.

https://arxiv.org/pdf/2003.06549v1.pdf



“Dendrology” in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation

GREGORY R. HERMAN AND RUSS S. SCHUMACHER

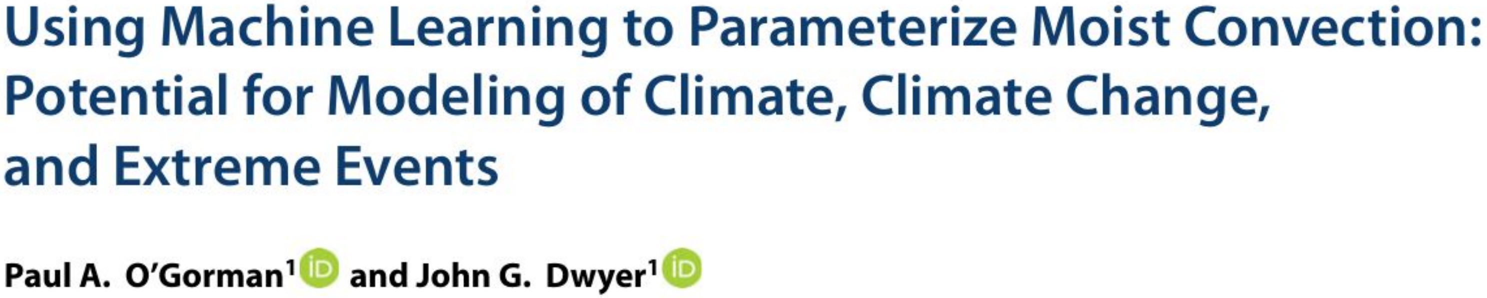
Feature importance for Random Forest or Gini importance:

the number of splits based on the feature summed over the forest [...], it is normalized, an importance of one then indicates that all decision nodes in every tree of the forest split on the corresponding feature, while an importance of zero indicates that no decision node splits based on that feature.

Permutation accuracy importance:

for each predictive feature, the feature value for each sample used to construct a given tree is permuted to a different sample’s value. Importance is determined by the decline in the model’s predictive performance when replacing the true values with the permuted ones.

https://journals.ametsoc.org/doi/10.1175/MWR-D-17-0307.1



Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events

Paul A. O'Gorman[1] and John G. Dwyer[1]

1. use the Random Forest parameterization to generate a Linear Response Function for the response of convective precipitation to small perturbations in temperature, specific humidity, and surface pressure.

2. use the concept of feature importance which seeks to measure the importance of the different input features (here temperature and humidity at different levels and surface pressure [...] for both the occurrence and strength of convection. [...] is implemented in RandomForestRegressor class of scikit-learn.

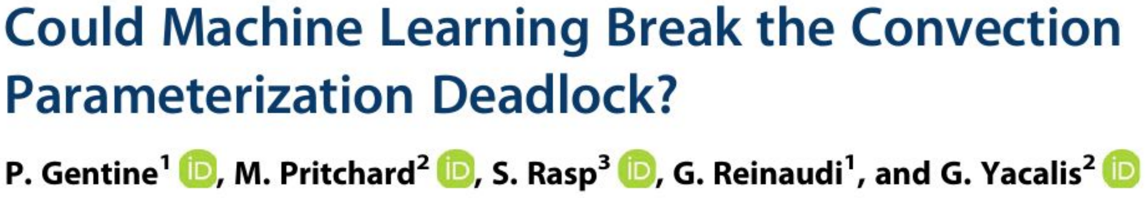
   [...] the feature importance is a useful additional diagnostic for the interaction of convection with the large-scale environment.

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018MS001351

Applications of Deep Learning to Ocean Data Inference and Subgrid Parameterization

Thomas Bolton[1] and Laure Zanna[1]

Machine Learning for Stochastic Parameterization:
Generative Adversarial Networks in the Lorenz ’96 Model

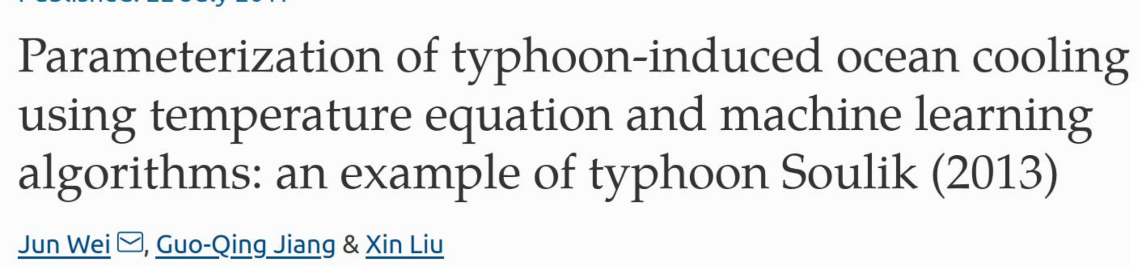


Could Machine Learning Break the Convection Parameterization Deadlock?

P. Gentine[1], M. Pritchard[2], S. Rasp[3], G. Reinaudi[1], and G. Yacalis[2]

Deep learning to represent subgrid processes in climate models

Stephan Rasp[a,b,1], Michael S. Pritchard[b], and Pierre Gentine[c,d]



Published: 22 July 2017

Parameterization of typhoon-induced ocean cooling using temperature equation and machine learning algorithms: an example of typhoon Soulik (2013)

Jun Wei, Guo-Qing Jiang & Xin Liu

Research Letter | Open Access

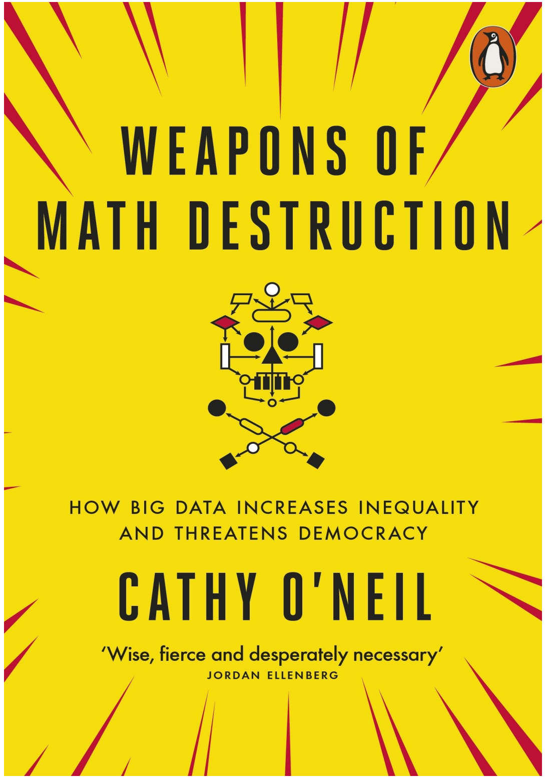
A Deep Learning Algorithm of Neural Network for the Parameterization of Typhoon-Ocean Feedback in Typhoon Forecast Models

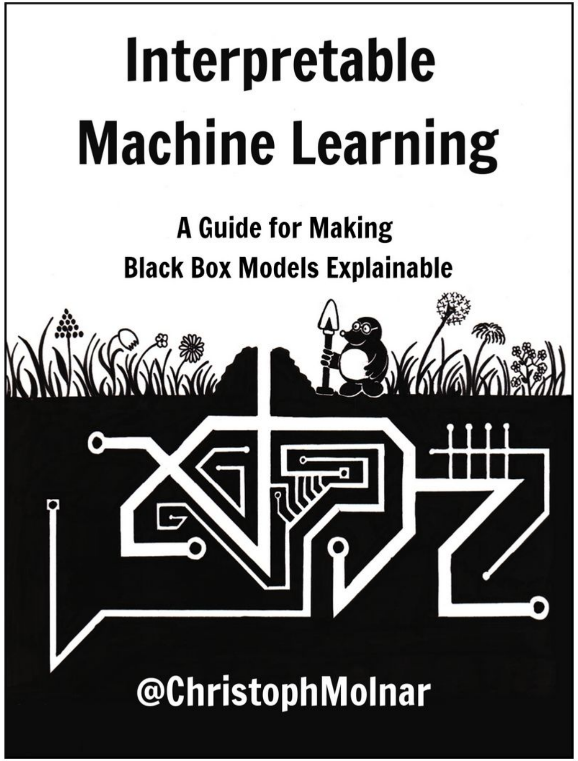
Guo-Qing Jiang, Jing Xu, Jun Wei

First published:26 March 2018 | https://doi.org/10.1002/2018GL077004 | Citations: 18

References

- Bias
  - Google face recognition: https://twitter.com/jackyalcine/status/615329515909156865 and https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/
  - Man in kitchen: https://jyzhao.net/ and https://searchenterpriseai.techtarget.com/feature/Big-data-throws-big-biases-into-machine-learning-data-sets
  - 50 types https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world/
- Physics-guided neural networks
  - Lakes simulation: Karpatne et al 2018 https://arxiv.org/pdf/1710.11431.pdf and https://towardsdatascience.com/physics-guided-neural-networks-pgnns-8fe9dbad9414
  - Navier-Stokes: Raissi et al 2020 https://www.sciencedirect.com/science/article/pii/S0021999118307125
- Uncertainty
  - Epistemic: http://yingzhenli.net/home/pdf/epistemic_uncertainty_neurips_bdl2019.pdf
  - Under data shift: Ovadia et al 2020 https://papers.nips.cc/paper/9547-can-you-trust-your-models-uncertainty-evaluating-predictive-uncertainty-under-dataset-shift.pdf
- XAI
  - 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


WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
'Wise, fierce and desperately necessary'
JORDAN ELLENBERG


Interpretable Machine Learning
A Guide for Making Black Box Models Explainable
@ChristophMolnar

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

BONUS TRACK SLIDE

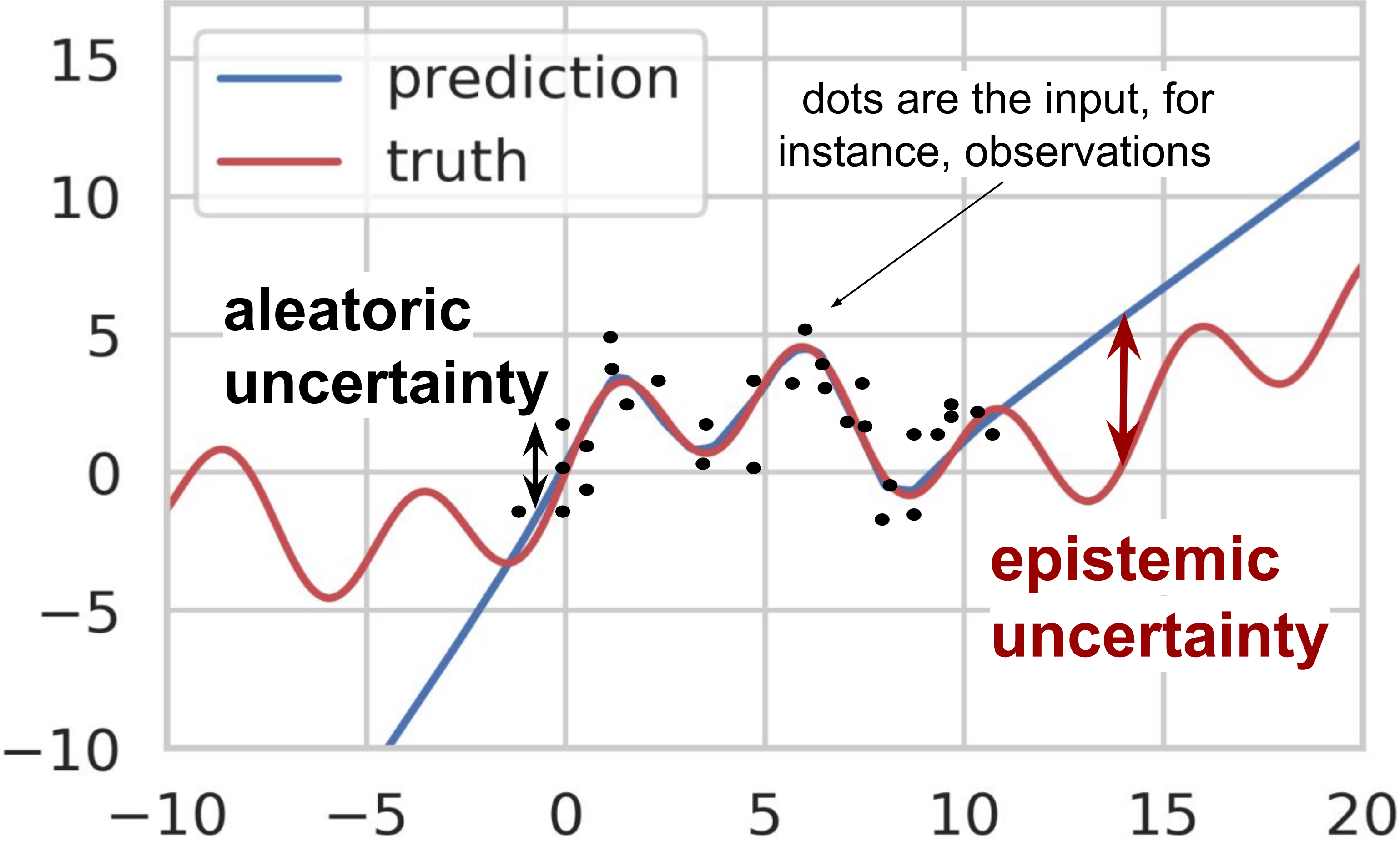
Two types of uncertainty

Aleatoric: “what is the next outcome of tossing a coin?” related to an individual experiment outcome, it is non-reducible 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.

BONUS TRACK SLIDE

ML are designed for interpolation, not extrapolation



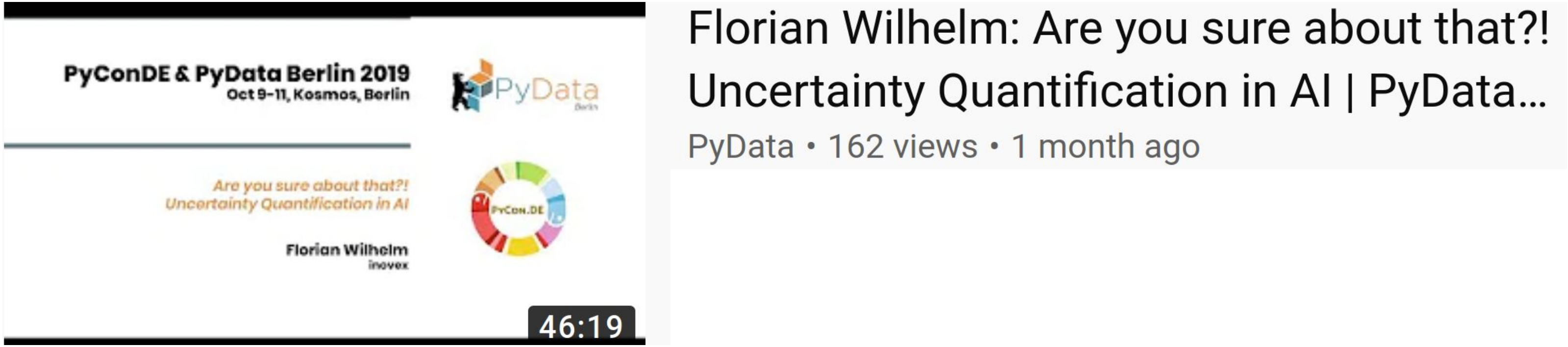
dots are the input, for instance, observations

aleatoric uncertainty

epistemic uncertainty

BONUS TRACK SLIDE

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


PyConDE & PyData Berlin 2019
Oct 9-11, Kosmos, Berlin
PyData
Are you sure about that?!
Uncertainty Quantification in AI
Florian Wilhelm
inovex
PyCon.DE
46:19

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

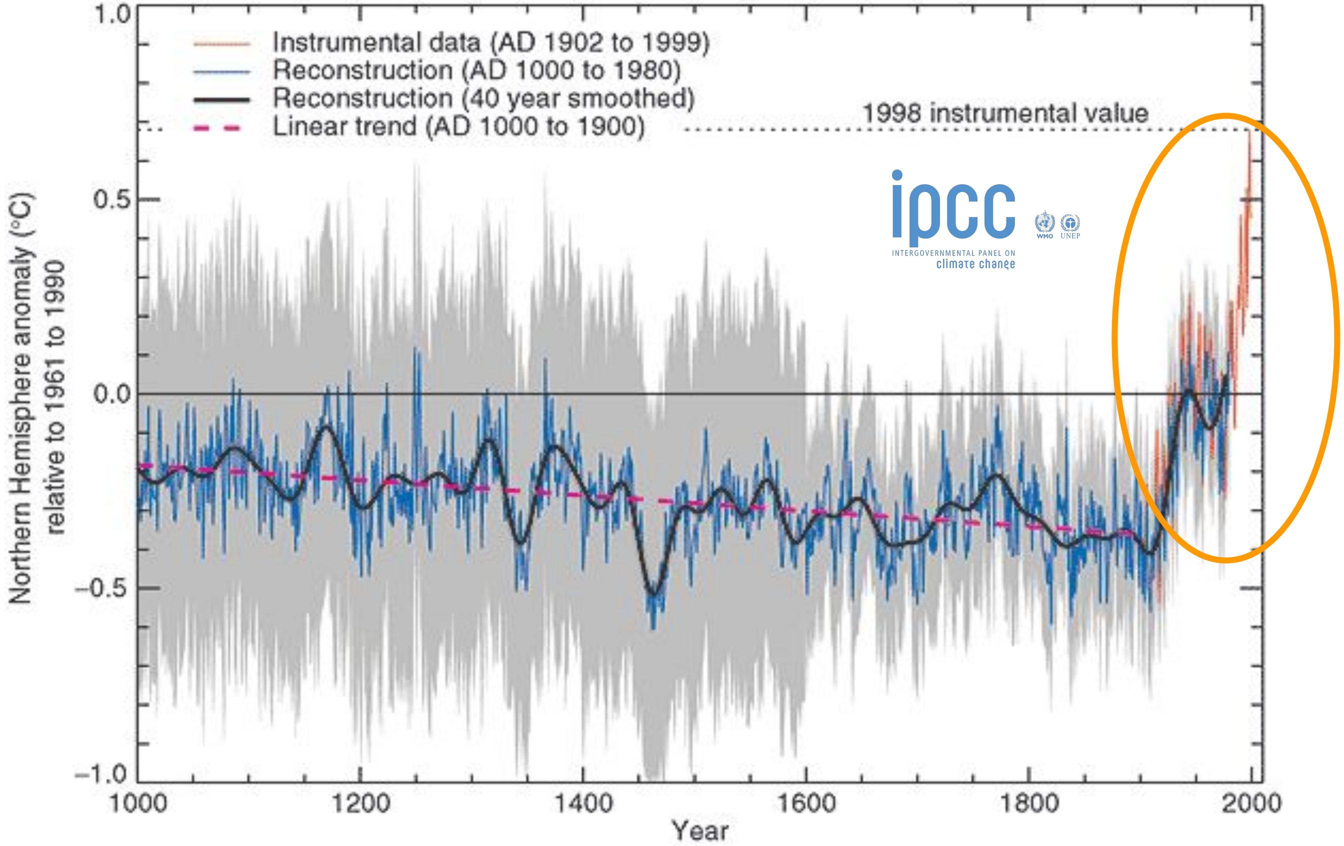
PyData • 162 views • 1 month ago

BONUS TRACK SLIDE

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 application data are i.i.d. but data drifts in time, we apply the model to data from a different location, the labeller changed,...

BONUS TRACK SLIDE



many problems in geoscience are non stationary

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

BONUS TRACK SLIDE

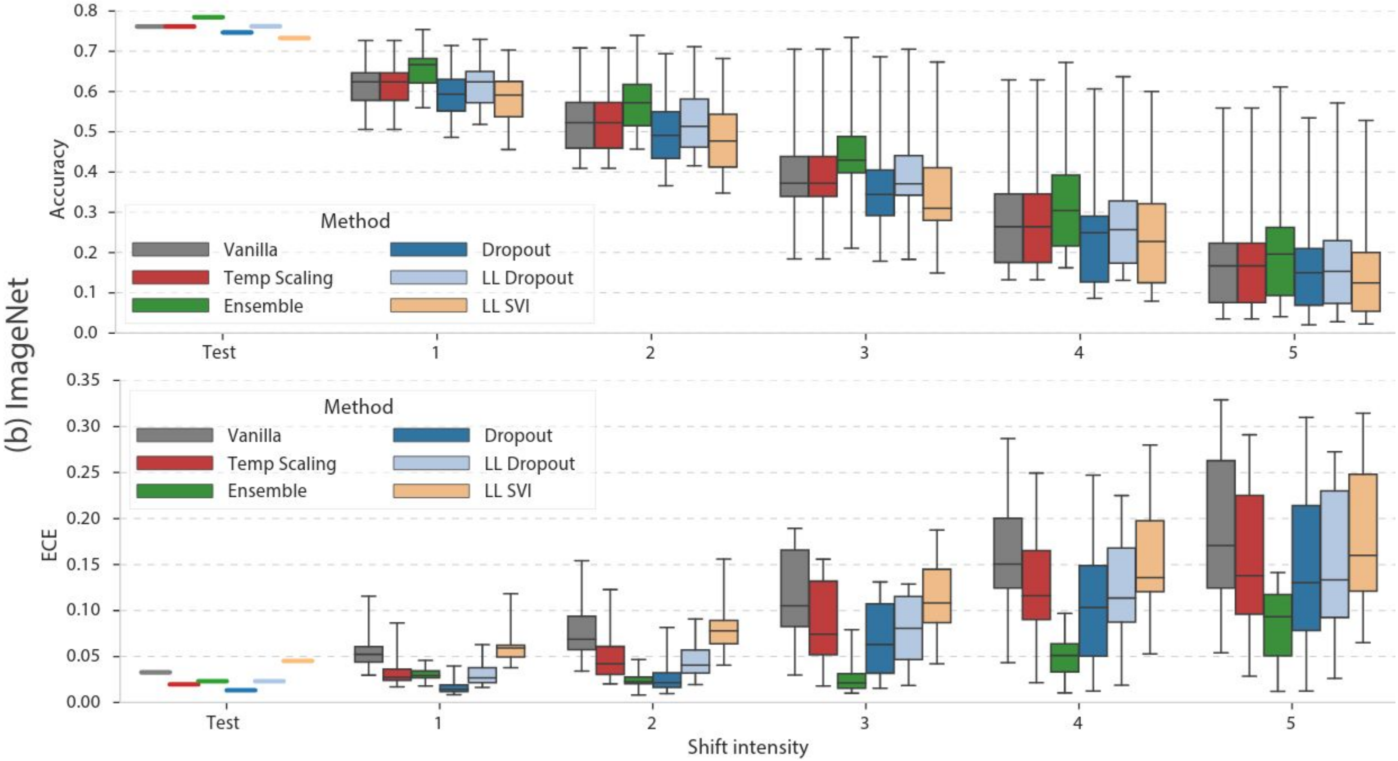
Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift

Yaniv Ovadia[*]
Google Research

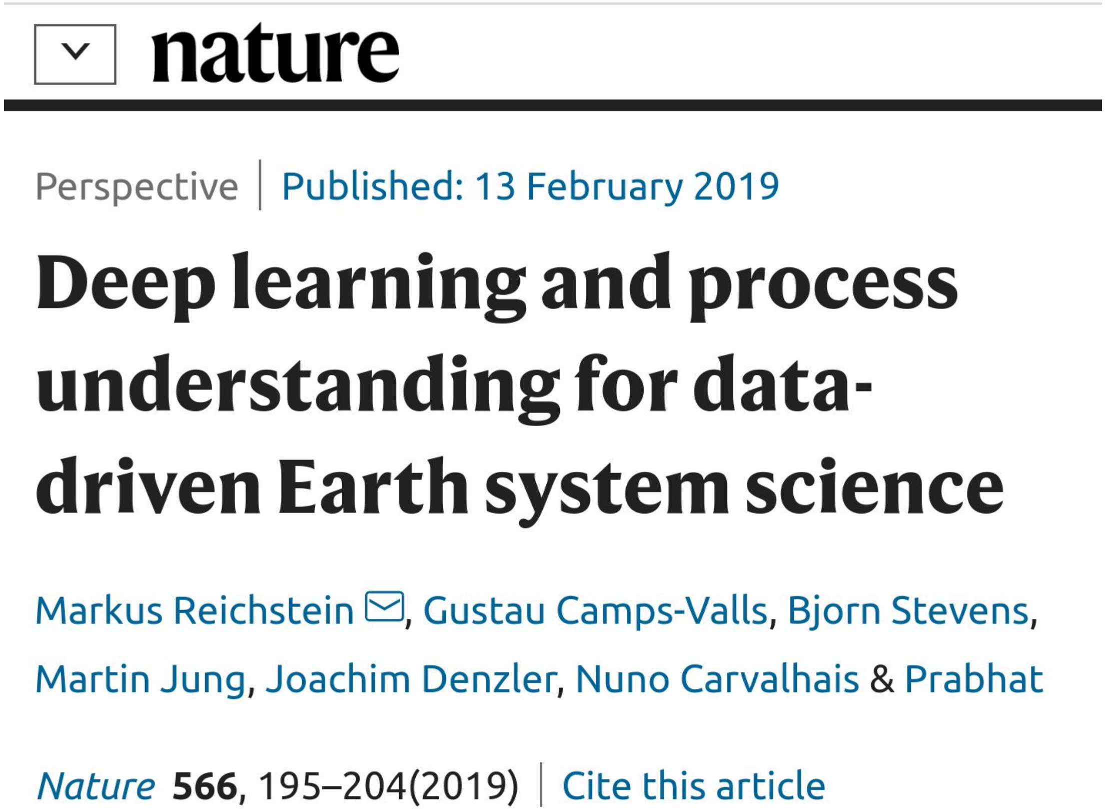
Emily Fertig[*†]
Google Research

Jie Ren[†]
Google Research

et al



BONUS TRACK SLIDE



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

BONUS TRACK SLIDE

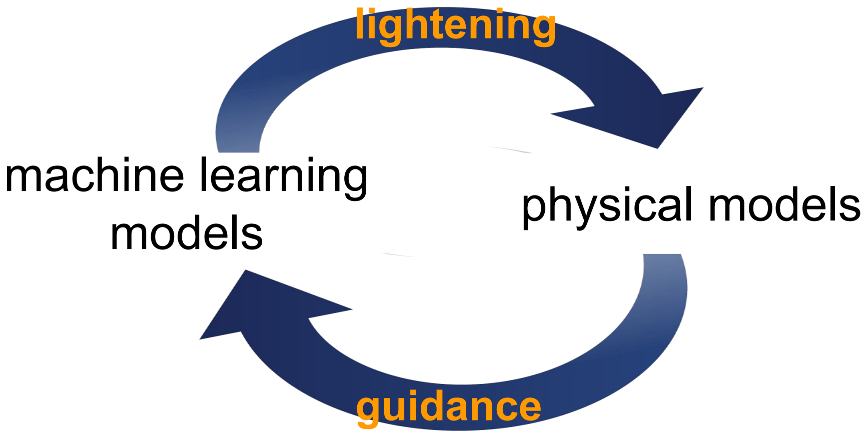
Conclusions

(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.

BONUS TRACK SLIDE

Best practices (I): Hybrid models



lightening

machine learning models

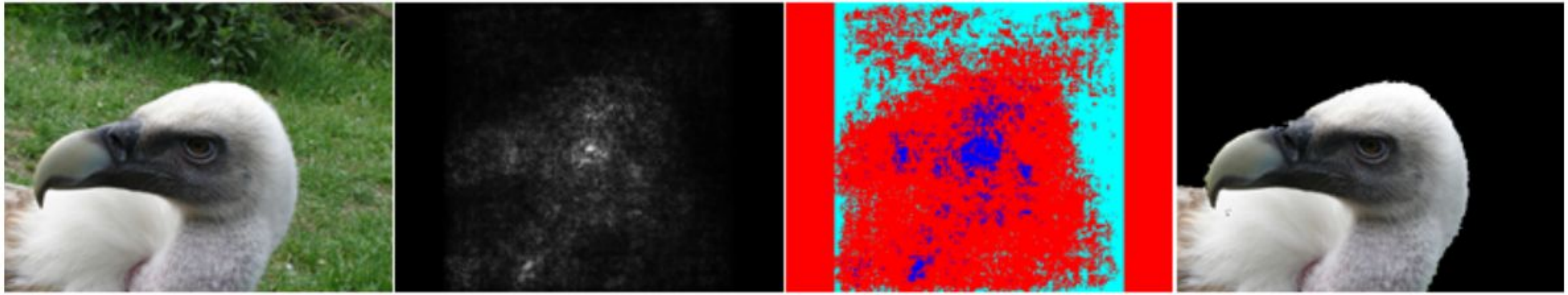
physical models

guidance

BONUS TRACK SLIDE

Best practices (II): put your model on diet

Select samples using domain knowledge (lakes, Navier-Stokes examples) , use XAI to identify and remove background, spurious correlations (leakage),...



Simonyan et al 2014
https://arxiv.org/pdf/1312.6034v2.pdf

BONUS TRACK SLIDE

Best practices (III): accuracy is not enough to evaluate the model skills

… the model was very accurate, but in classifying grass (cows example) or it allowed for denser water up (lakes example)...

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

Doshi-Velez and Kim, 2017
Towards A Rigorous Science of Interpretable Machine Learning
https://arxiv.org/abs/1702.08608

moreno@dkrz.de

BONUS TRACK SLIDE

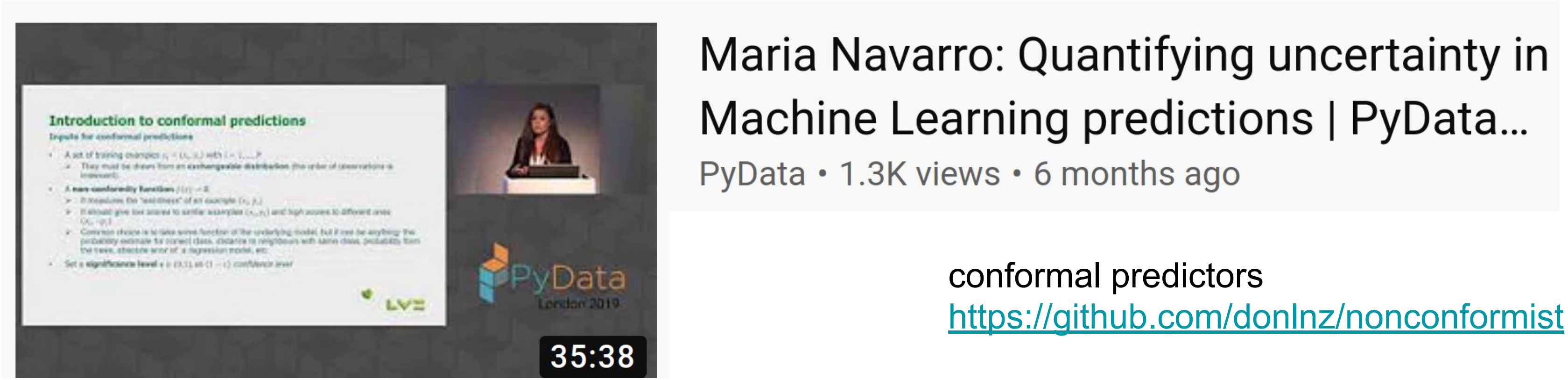
Best practices (III): call a human!

Calculate confidence intervals with uncertainty quantification techniques:

- Conformal Predictors
- MC dropouts
- Deep Ensembles
- Quantile regression
- Gaussian processes (also to propagate distributed input data: https://presentations.copernicus.org/EGU2020/EGU2020-14677_presentation.pdf)

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

BONUS TRACK SLIDE



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

PyData • 1.3K views • 6 months ago

conformal predictors
https://github.com/donlnz/nonconformist

PyConDE & PyData Berlin 2019
Oct 9-11, Kosmos, Berlin
PyData
Berlin
Are you sure about that?!
Uncertainty Quantification in AI
Florian Wilhelm
inovex
PyCon.DE
46:19

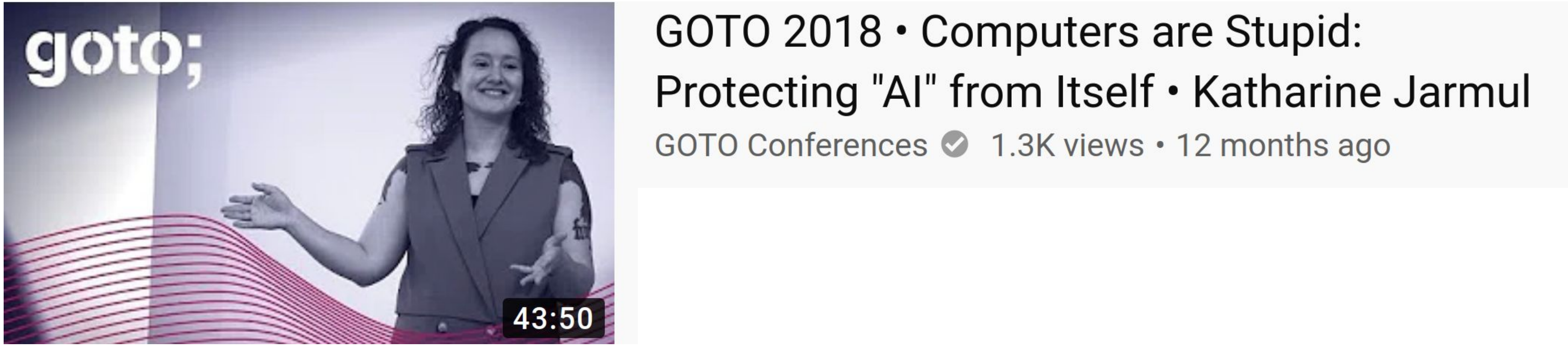
Florian Wilhelm: Are you sure about that?! Uncertainty Quantification in AI | PyData...

PyData • 162 views • 1 month ago

BONUS TRACK SLIDE

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

PyData • 3K views • 6 months ago

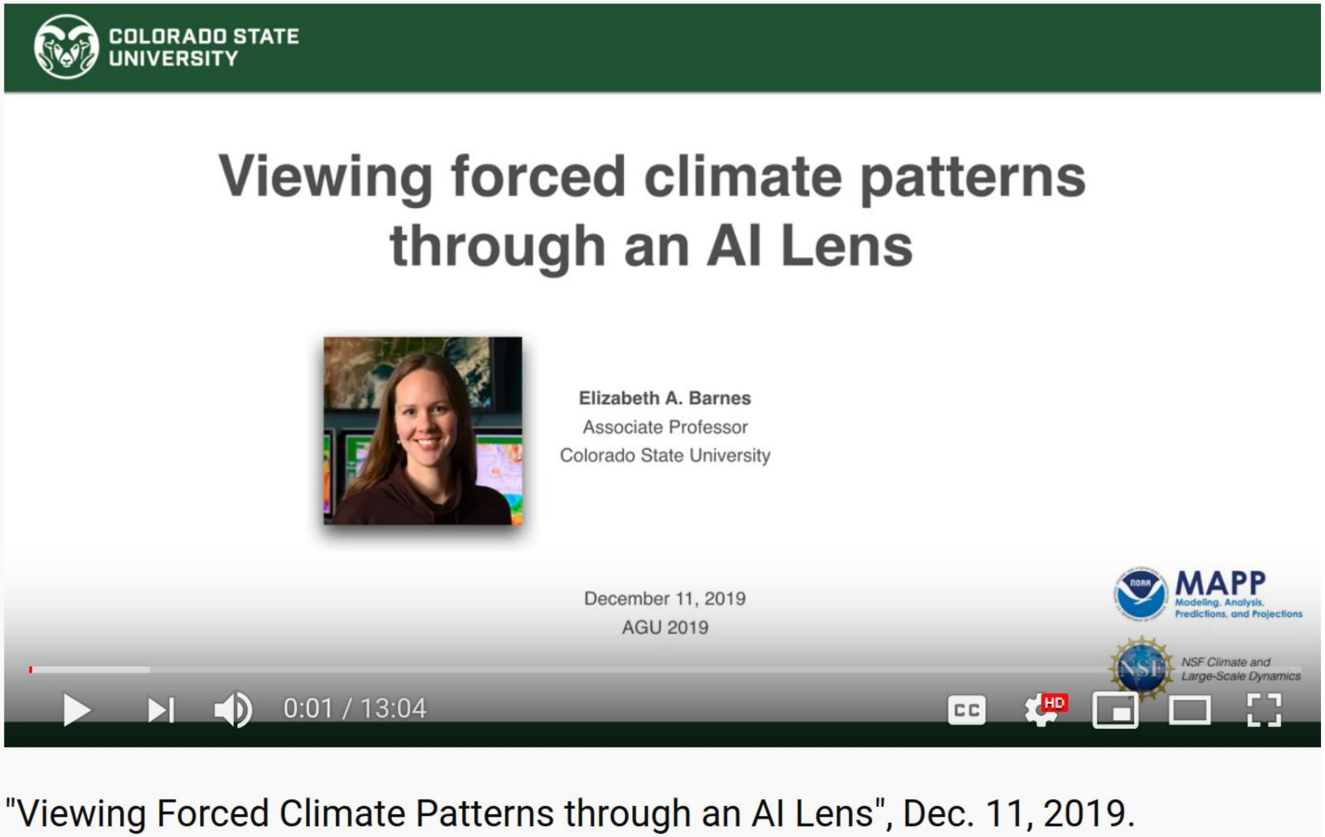


GOTO 2018 • Computers are Stupid: Protecting "AI" from Itself • Katharine Jarmul

GOTO Conferences 1.3K views • 12 months ago

BONUS TRACK SLIDE

Explainability added value



COLORADO STATE UNIVERSITY

Viewing forced climate patterns through an AI Lens

Elizabeth A. Barnes
Associate Professor
Colorado State University

December 11, 2019
AGU 2019

MAPP
Modeling, Analysis, Predictions, and Projections
NSF Climate and Large-Scale Dynamics
0:01 / 13:04
CC
HD

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