Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Varya

Mountain hydrology literature discussion series
23.01.23

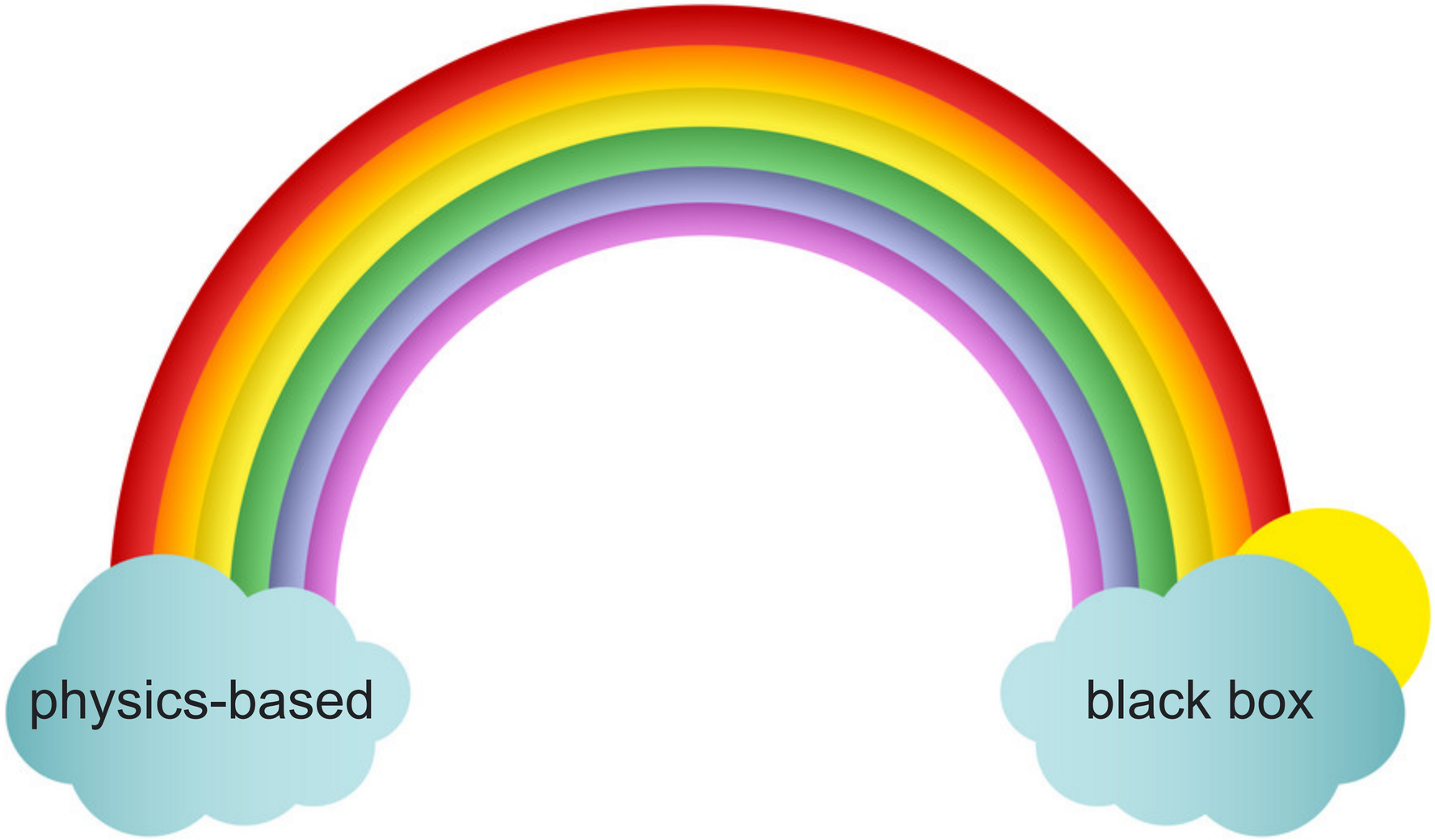
… about the author

*… Dr. Shmueli's research focuses on statistical and machine learning methodology with applications in information systems and healthcare, and an emphasis on human behavior **

*… Dr. Shmueli teaches courses on data mining, forecasting analytics, interactive visualization, research methods, and other business analytics topics **

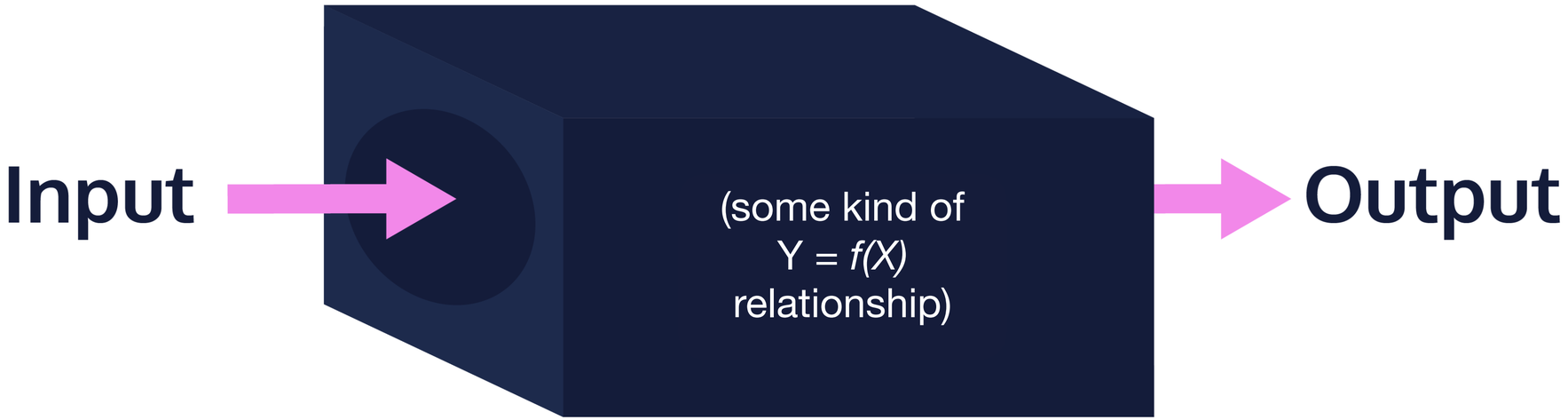
** https://www.galitshmueli.com/*

let’s start with some concepts…

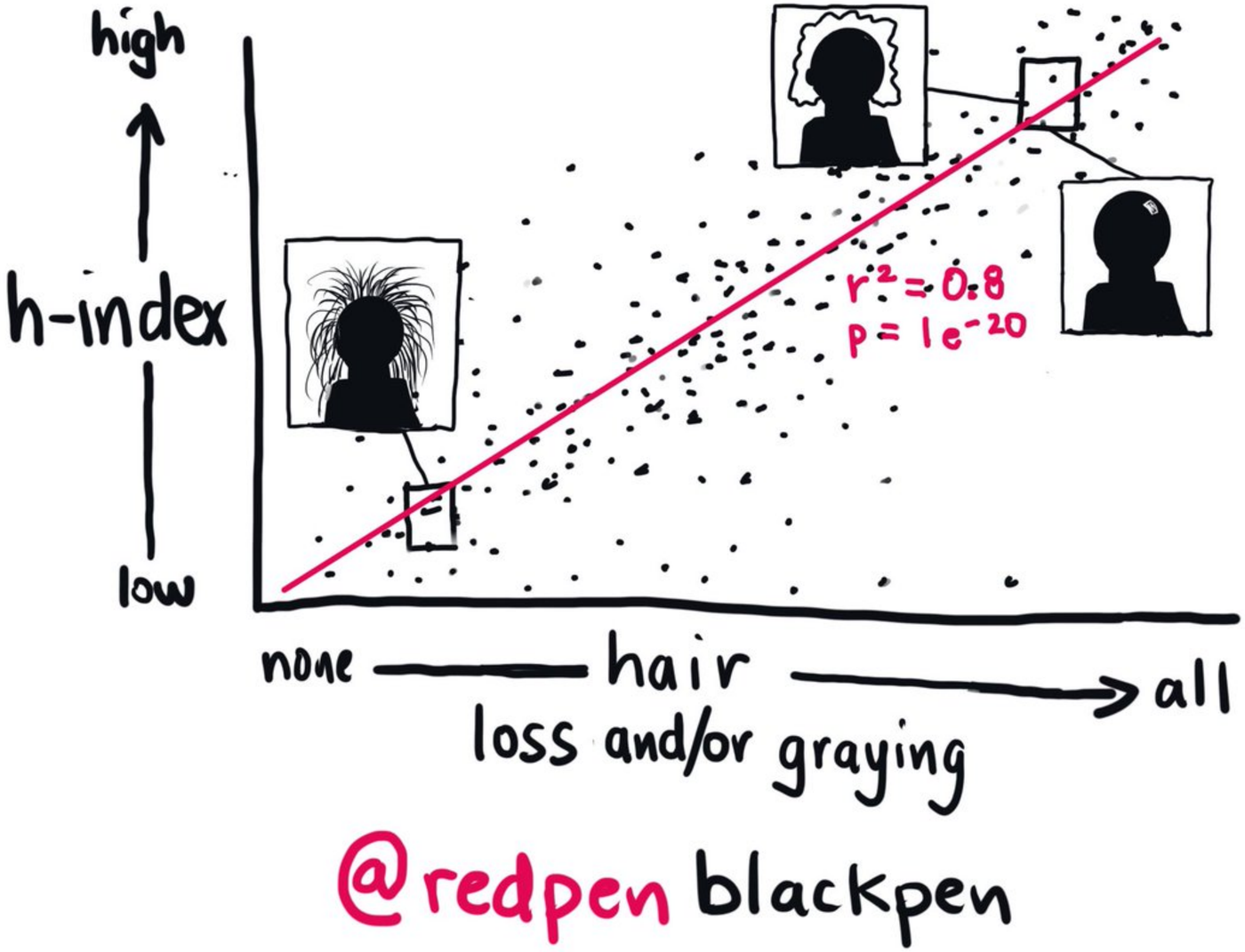


black box models*

* (from describing-processes-goal stand point)



… for example

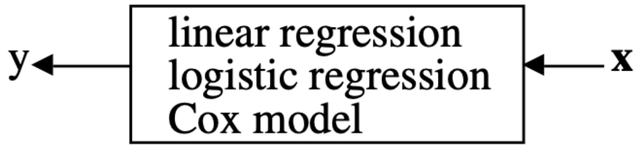

high
h-index
low
$r^2 = 0.8$
$p = 1e^{-20}$
none
hair loss and/or graying
all
@redpen blackpen

Statistical modeling: The two cultures

There are two goals in analyzing the data:

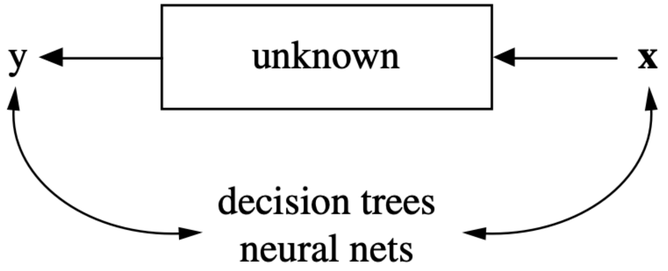
Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Data modeling culture



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

Algorithmic modeling culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

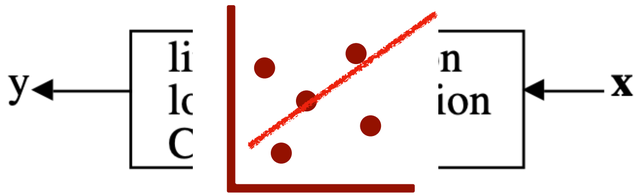
Breiman, Leo. "Statistical modeling: The two cultures." Statistical Science 16.3 (2001): 199-231.

Statistical modeling: The two cultures

There are two goals in analyzing the data:

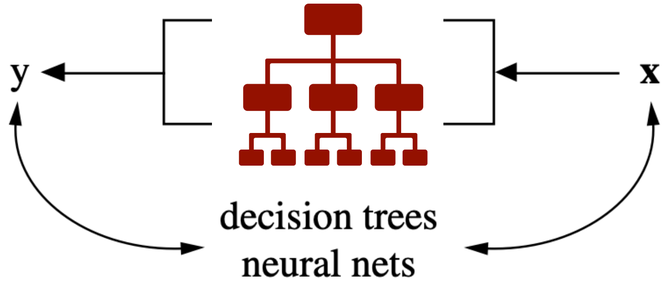
Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

Data modeling culture



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

Algorithmic modeling culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Breiman, Leo. "Statistical modeling: The two cultures." Statistical Science 16.3 (2001): 199-231.

… in other words:

To explain or to predict?

- **Explanatory modelling:** refers here to the application of statistical models to data for testing causal hypotheses about theoretical constructs.
  In such models, a set of underlying factors that are measured by variables X are assumed to cause an underlying effect, measured by variable Y

- **Predictive modelling:** the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations, including temporal forecasting (= generating values for new/future observations)

… more about predictive modeling:

what else can we say about it?

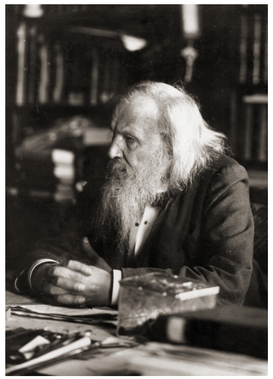
- large and rich datasets often contain complex relationships and patterns that are hard to hypothesize, especially given theories that exclude newly measurable concepts
- capturing underlying complex patterns and relationships, predictive modeling can suggest improvements to existing explanatory models
- predictive models are advantageous in terms of negative empiricism: a model either predicts accurately or it does not, and this can be observed.
- developing and testing new theories and ideas

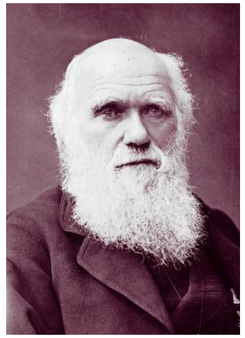
- operational use 🪄

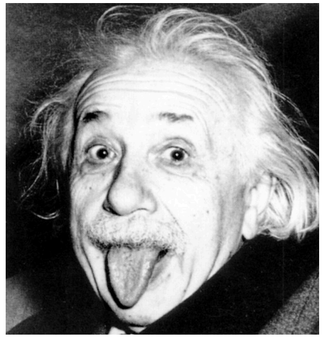
… for example:

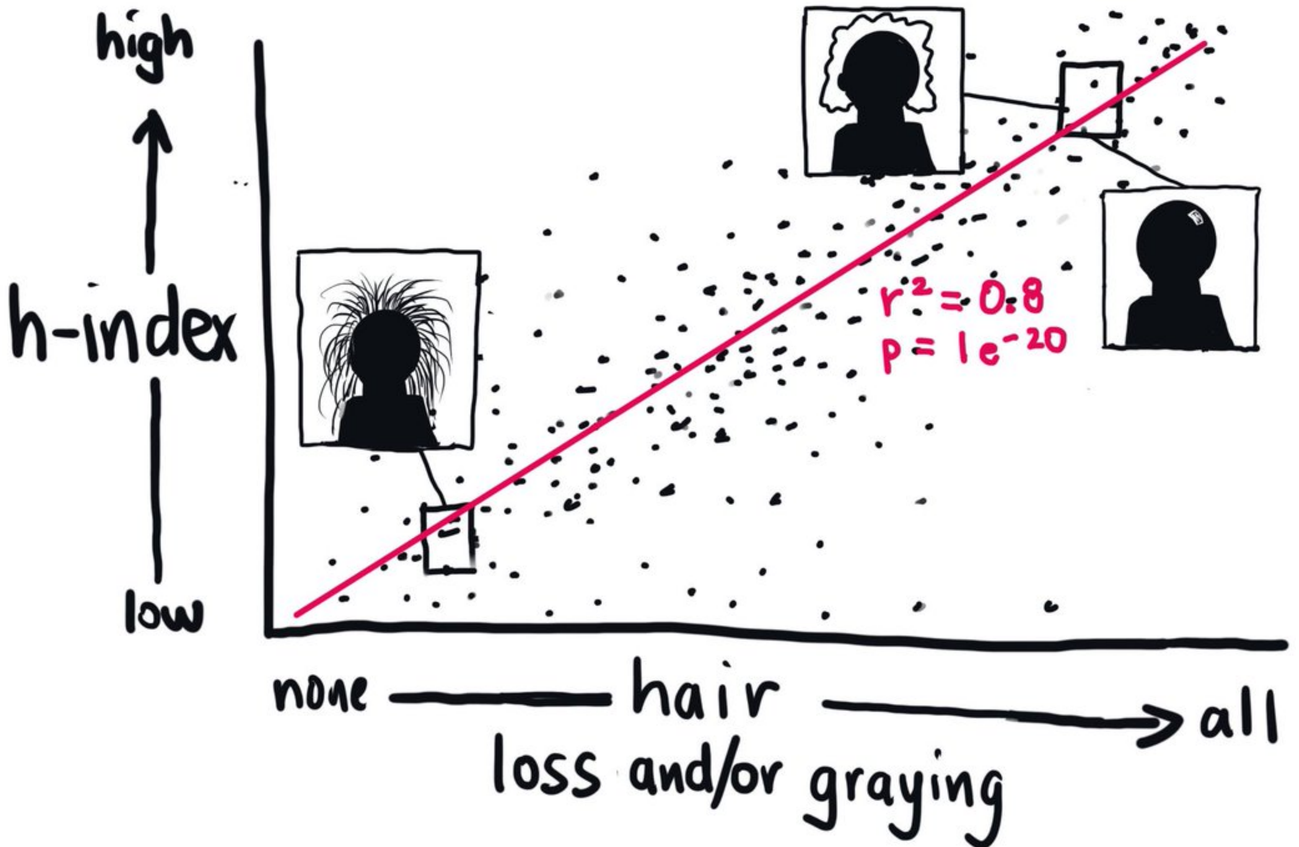
We can look at the population of scientists, collect list of features (X), create some metric (Y) and hypothesize, that one is caused by the other

Can we use feature X to predict this metric Y?





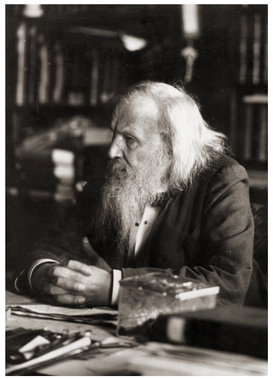


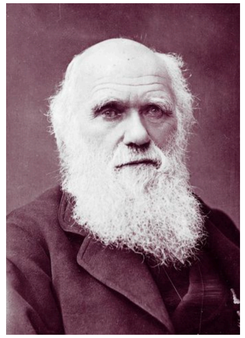


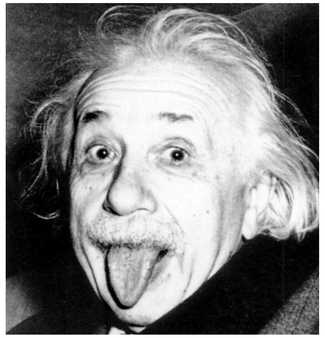
… for example:

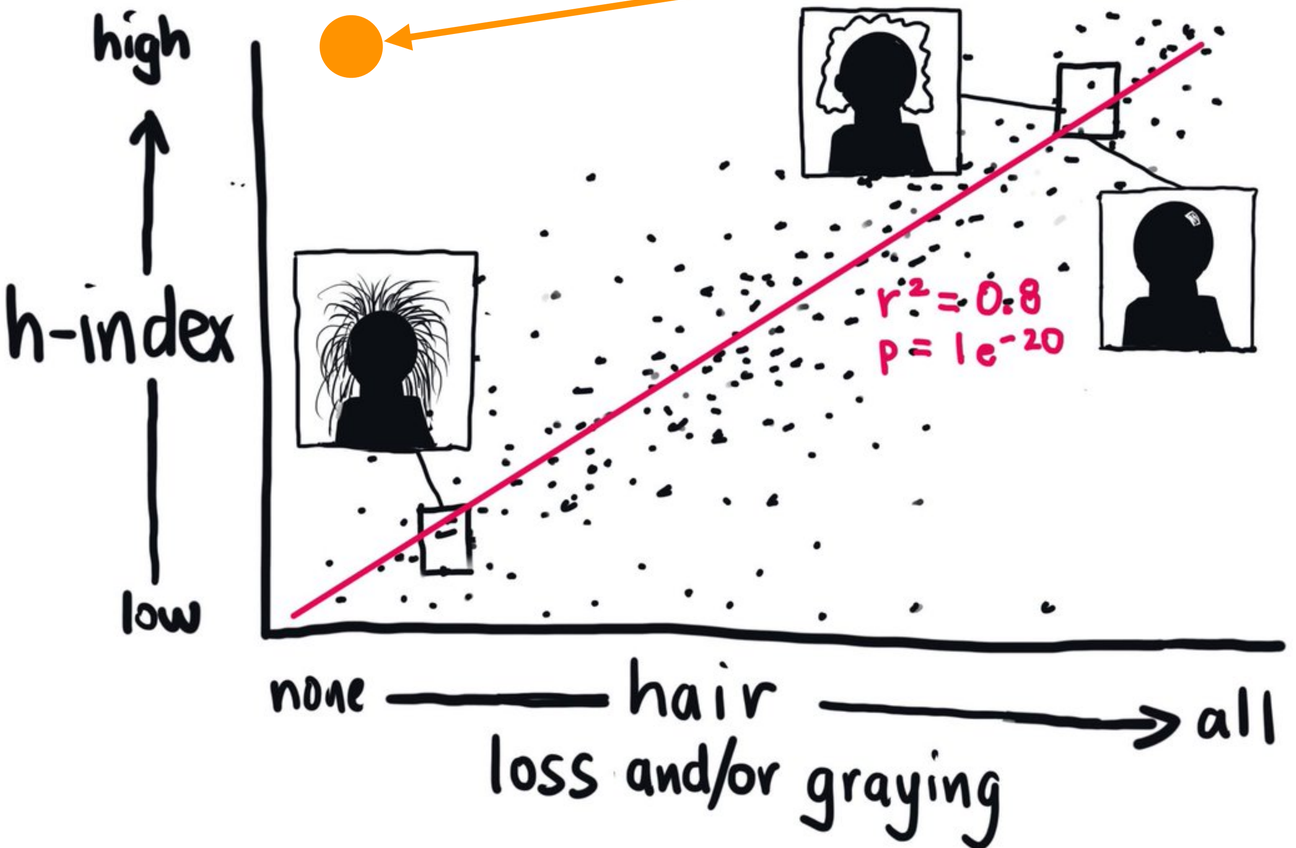
We can look at the population of scientists, collect list of features (X), create some metric (Y) and hypothesize, that one is caused by the other

Can we use feature X to predict this metric Y?









… another example: election polls

"There is a subtle, but important, difference between reflecting current public sentiment and predicting the results of an election. Surveys [election polls] have focused largely on the former—in other words, on providing a current snapshot of voting preferences, even when asking about voting preference as if elections were carried out on the day of the survey. In that regard, surveys are accurately describing current opinions of the electorate. However, the public perception is often focused on projecting the survey results forward in time to election day, which is eventually used to evaluate the performance of election surveys."

Blog by Prof. Galit Shmueli
http://www.bzst.com/search/label/Explaining%20vs.%20Predicting

“a priori determination of the main study goal as either explanatory or predictive is essential to conducting adequate modeling.”

So what is the difference?

Causation - Association

underlying causal function vs association between X and Y

Theory - Data

hypothesis-driven vs data-driven

Retrospective - Prospective

testing already existing hypothesis vs predicting new observations

Bias - Variance

minimizing bias vs minimizing combination of bias and variance (often sacrificing theoretical accuracy)

Data collection and visualization (EDA*)

(There are a lot of details, but let's focus only on some)

Data: Data driven modelling requires more data (relative to hypothesis driven)

Visualization: exploratory vs confirmatory visualization

* EDA = Exploratory data analysis

… and visualization (EDA*)

Visualization: exploratory vs confirmatory visualization

"Visualizations can be used to explore data, to confirm a hypothesis, or to manipulate a viewer...In exploratory visualization the user does not necessarily know what he is looking for. In a confirmatory visualization, the user has a hypothesis that needs to be tested. This scenario is more stable and predictable."

* EDA = Exploratory data analysis

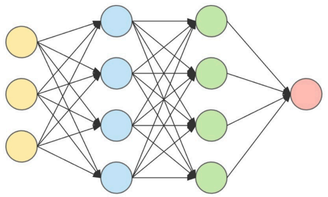
Choice of methods:

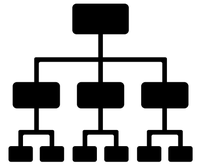
- **explanatory** modeling requires interpretable statistical models that are easily linked to the underlying theoretical model.
- popularity of statistical models, and especially regression-type methods

data modeling

- **in predictive** modeling modeling he top priority is generating predictions (*f* is unknown)
- model might not shed light on an underlying causal mechanism, but it can capture complicated associations, leading to accurate predictions
- model transparency is of secondary importance: “Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why” (Breiman, 2001b).

algorithmic modelling





Concluding remarks: what is important

explanatory power ≠ predictive power of the model

a priori determination of the main study goal → study design should differ

Predictive modelling vs scientific research: research and practice gap?

Should the model be able to do both?

How is it all relevant to MH research?

Thoughts?

- Modelling principal: crap in crap out?
- My own work trying to find controlling parameters of the flood vs debris flow (dominated) catchment
- Are physics based models “better”?

frontiers in Water

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/frwa.2020.00008

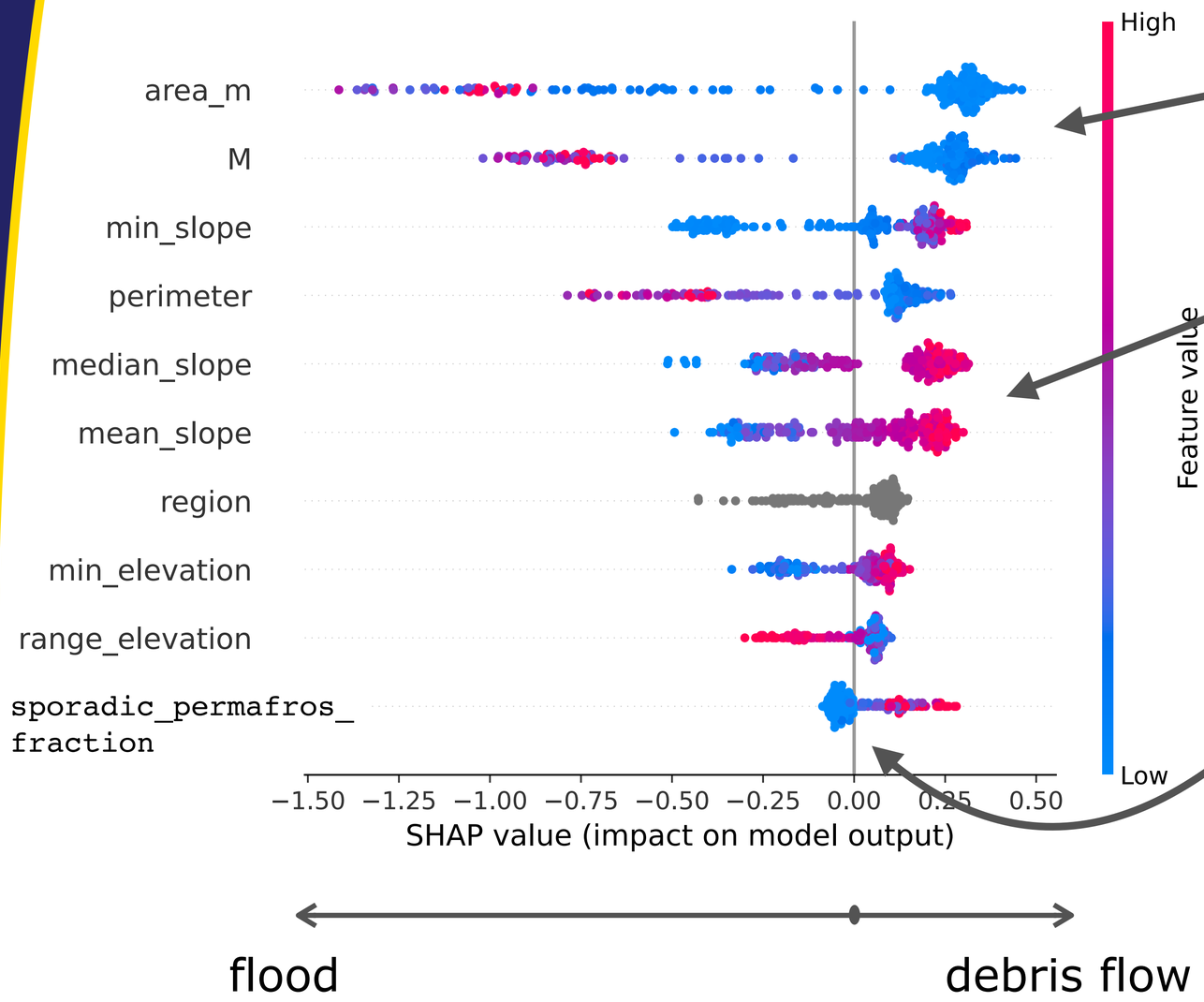
Check for updates

Machine Learning vs. Physics-Based Modeling for Real-Time Irrigation Management

Silvio J. Gumiere[1], Matteo Camporese[2], Anna Botto[2], Jonathan A. Lafond[1], Claudio Paniconi[3], Jacques Gallichand[1] and Alain N. Rousseau[3]*

[1] Department of Soils and Agri-Food Engineering, Laval University, Quebec, QC, Canada, [2] Civil, Environmental, and Architectural Engineering, University of Padova, Padua, Italy, [3] INRS-ETE, Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Quebec, QC, Canada

Why does Catboost model make this predictions?



small areas, small Melton ratio make it more likely to be classified as “debris flow”(1)

high mean and median slope bring the model result closer to the “debris flow” class (1)

low fraction of sporadic permafrost drags the model towards “flood” (0) class

Morphometric + climate