Recent advancements in machine learning relevant to ecological science

Masahiro Ryo

Free University of Berlin

Berlin-Brandenburg Institute of Advanced Biodiversity Research

Seminar @Helmholtz-Zentrum für Umweltforschung UFZ, Germany. 2018.02.12.

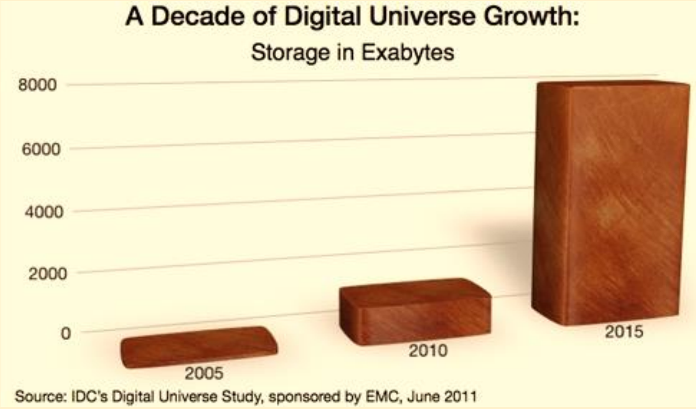
BBIB

Berlin-Brandenburg Institute of Advanced Biodiversity Research

PART 1

Introduction to Machine Learning

Why Machine learning?



Biology: The big challenges of big data

Vivien Marx

Nature **498**, 255–260 (13 June 2013) | doi:10.1038/498255a
Published online 12 June 2013

CONCEPTS AND QUESTIONS

Hampton et al. (**2013**)

Big data and the future of ecology

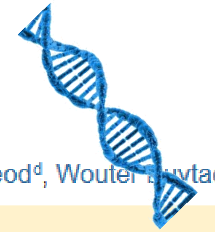
Stephanie E Hampton[1*], Carly A Strasser[2], Joshua J Tewksbury[3], Wendy K Gram[4], Amber E Budden[5], Archer L Batcheller[6], Clifford S Duke[7], and John H Porter[8]

ELSEVIER
Environmental Modelling & Software
Volume 63, January 2015, Pages 185–198

Review

Web technologies for environmental Big Data

Claudia Vitolo[a], Yehia Elkhatib[b], Dominik Reusser[c], Christopher J.A. Macleod[d], Wouter ...ytae



What is Machine learning?

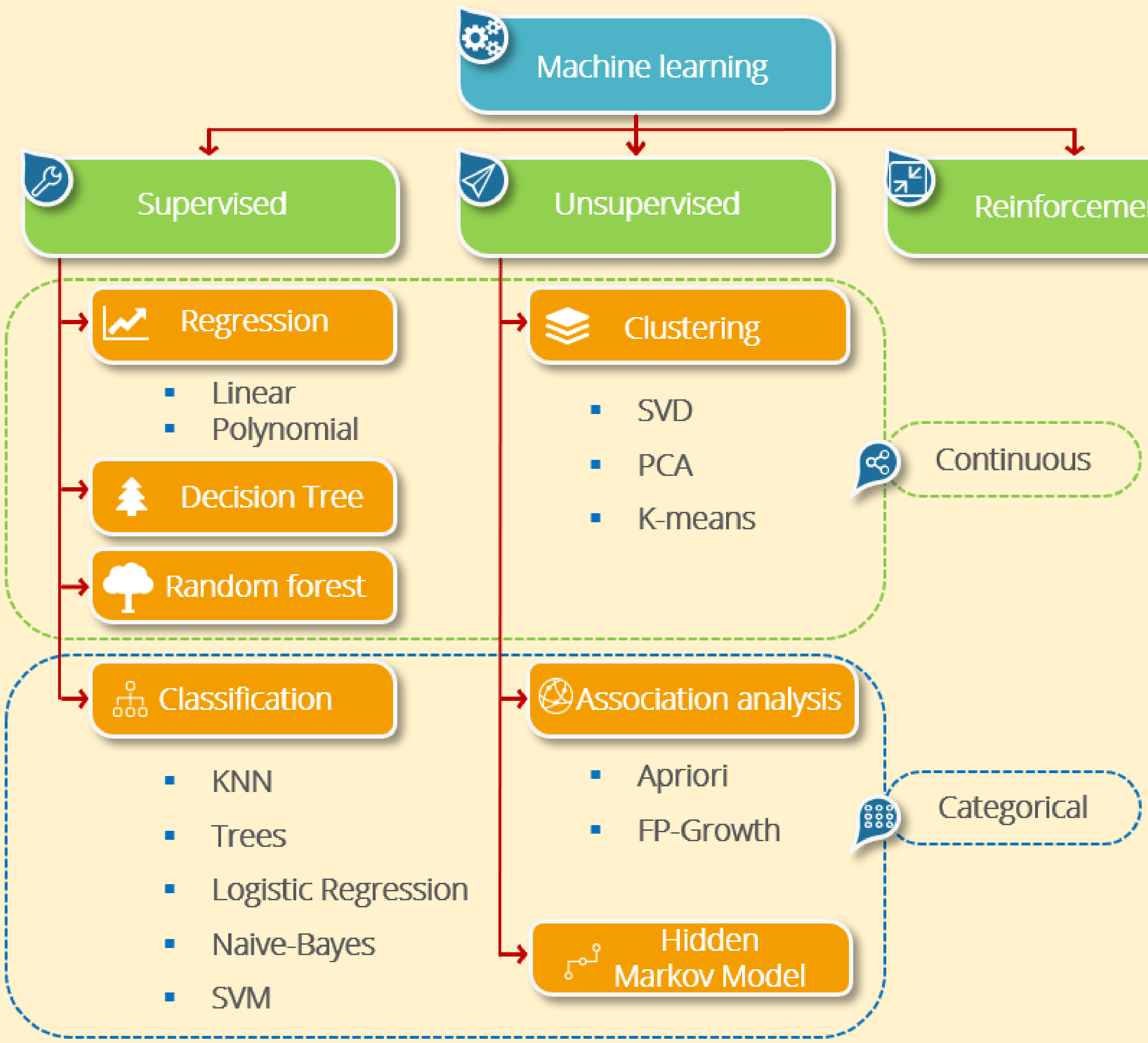
"A field of study that gives computers the ability to learn without being explicitly programmed"

first definition by Samuel A. (1959)

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

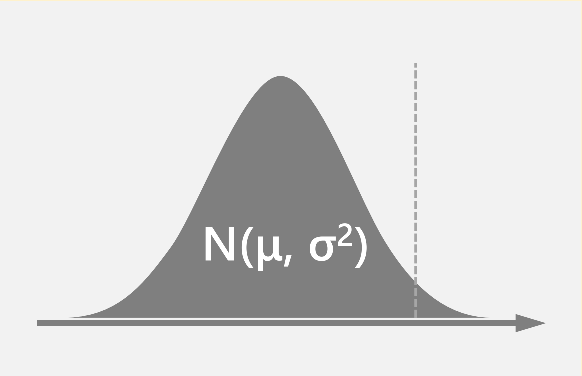
popular definition by Mitchell T.M. (1997)

What Machine learning does?



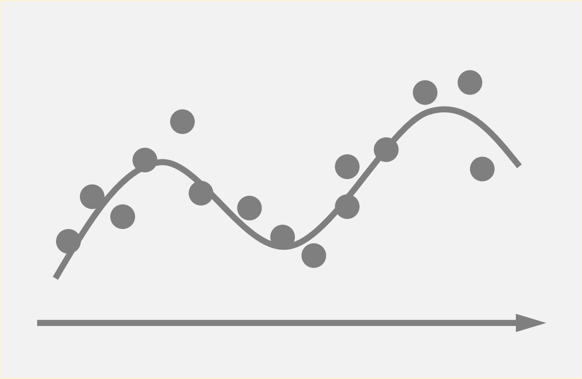
https://www.edureka.co/blog/what-is-machine-learning/

Statistics vs Machine learning?



Statistics

- Hypothesis-testing, theory-driven
- Strict assumptions
  (e.g. Linearity, normality, additivity)
- Probability



Machine learning

- Information-searching, data-driven
  (missing data)
- Loose assumptions
  (e.g. non-linearity, non-normality, non-additivity)
- Predictability

Breiman, 2001

When Machine learning > Statistics?

- Fields where techniques advance faster than theories (no theoretical hypothesis/prediction possible)
- Exploratory study with many predictors
- Data synthesis with many missing values
- Prediction is more important than explanation
- Nonlinear, non-additive modelling is preferred
- Unexpected outcomes wanted (cf. hypothesis generation)

PART 2

Recent trends in machine learning

In a nutshell

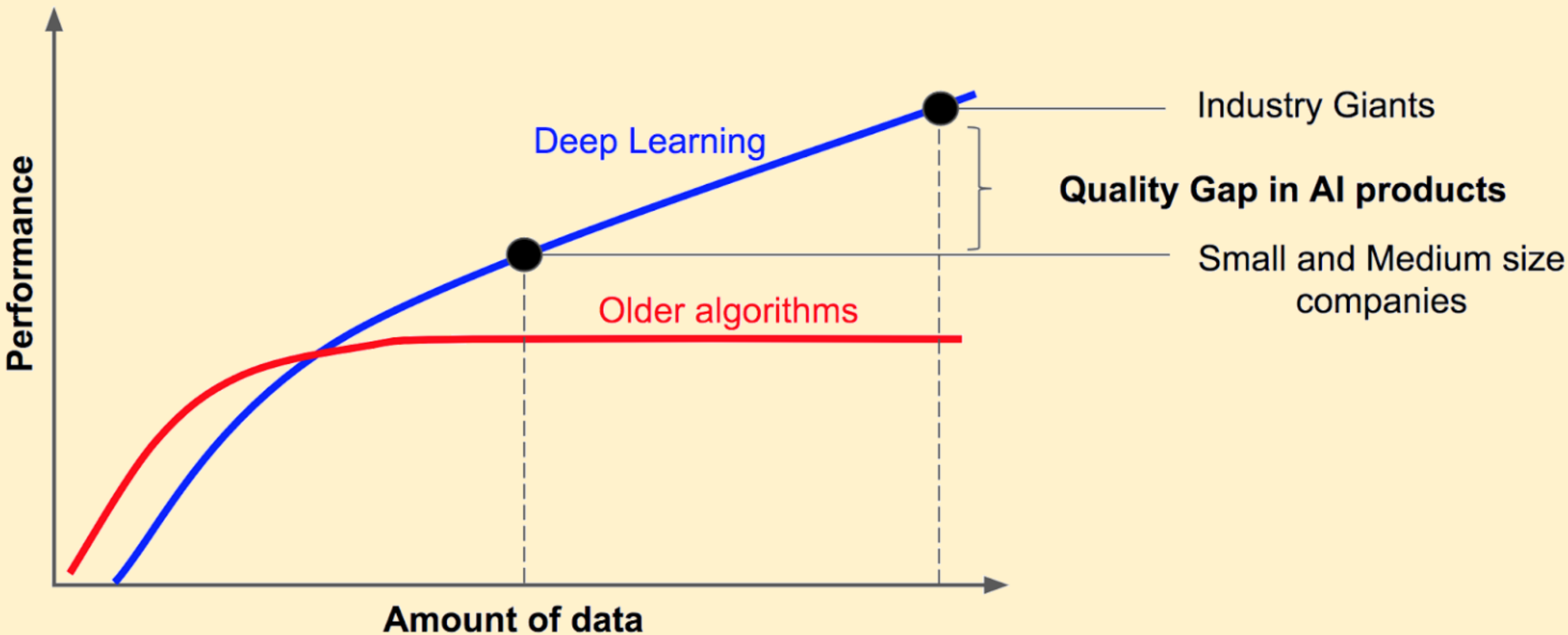
Predictability

- Big-data
- Small-data

Interpretability

- Pattern detection & visualization
- Integration with statistics

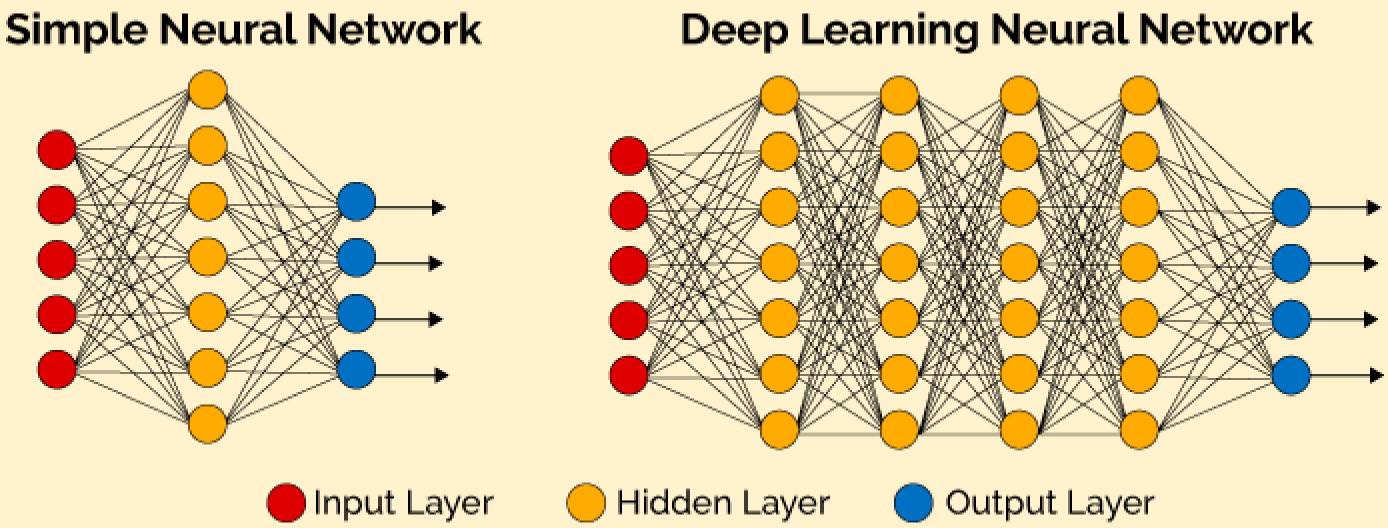
Predictability: The power of big data



So many DL algorithms have been proposed since 2006

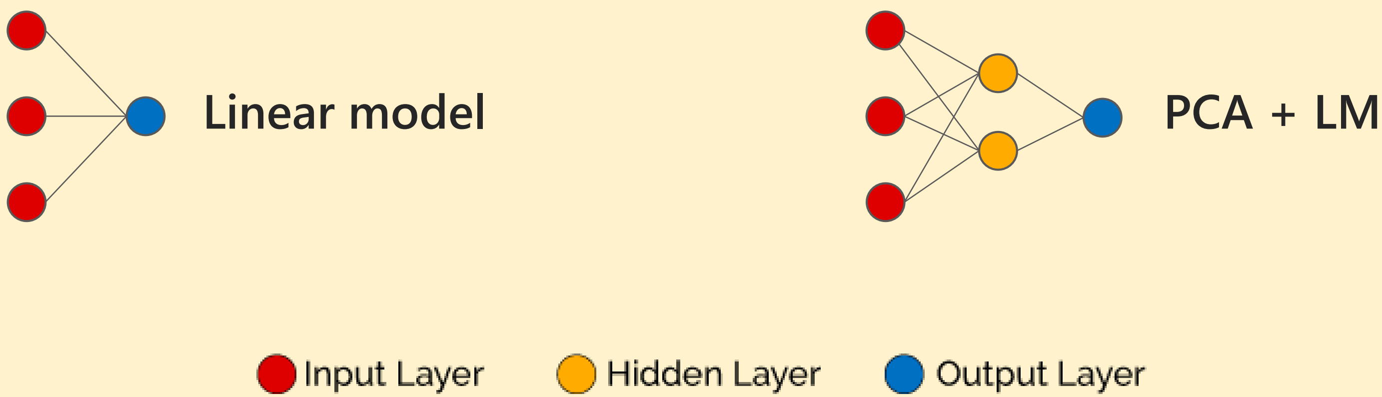
See) https://arxiv.org/pdf/1807.08169.pdf

Predictability: The power of big data



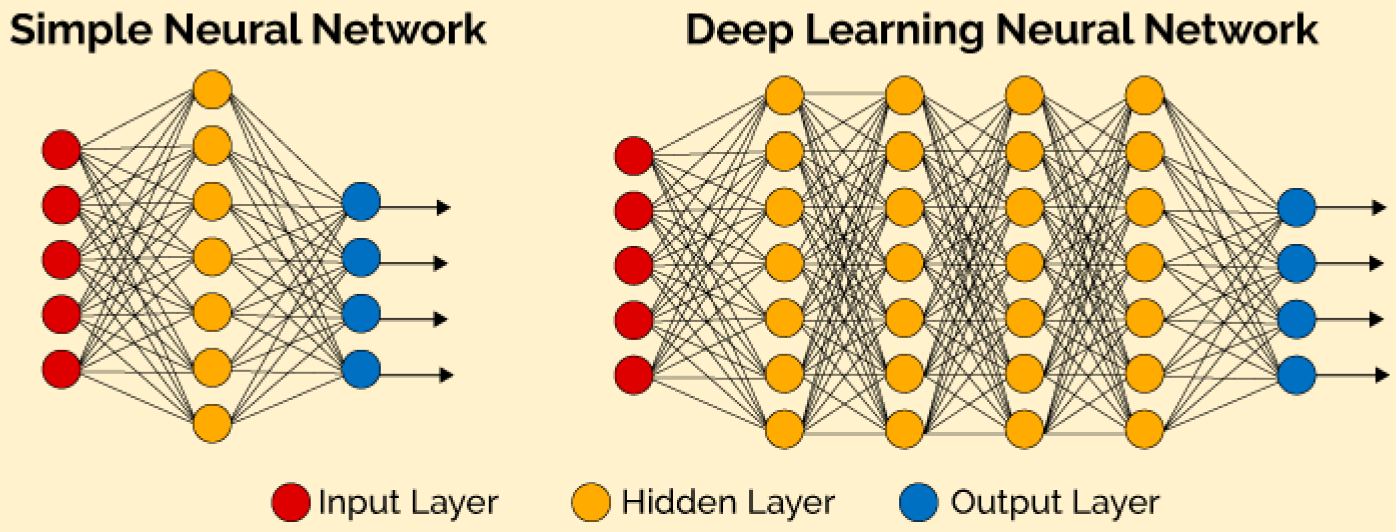
https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: The power of big data



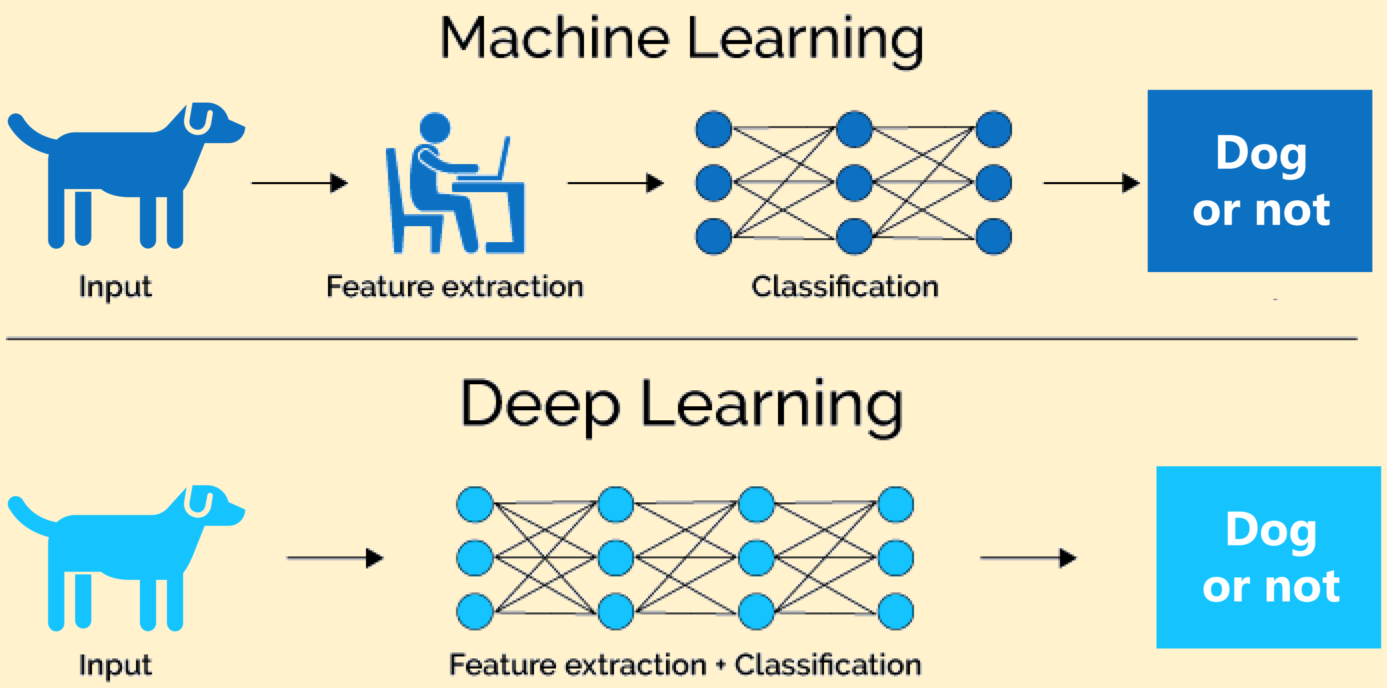
https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: The power of big data



https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: The power of big data

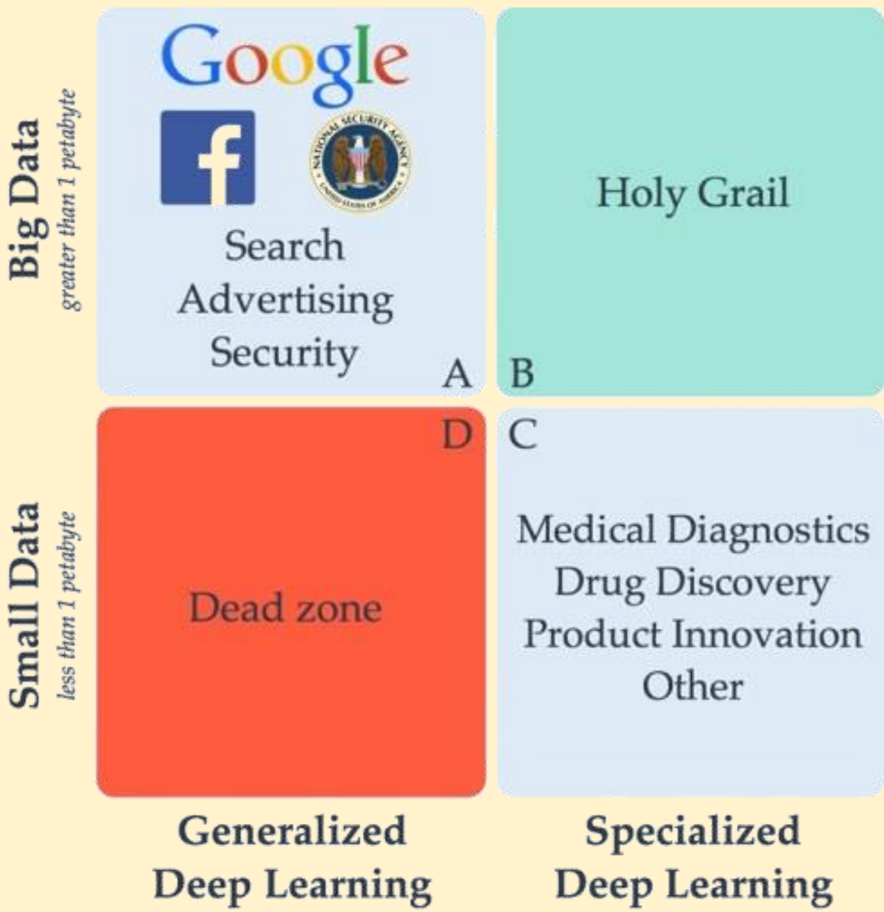


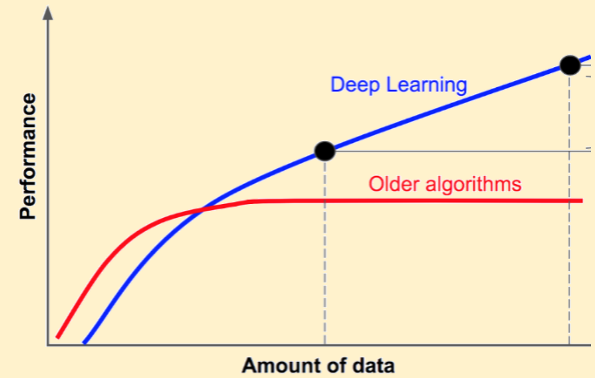
https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: The power of big data

https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: Small data

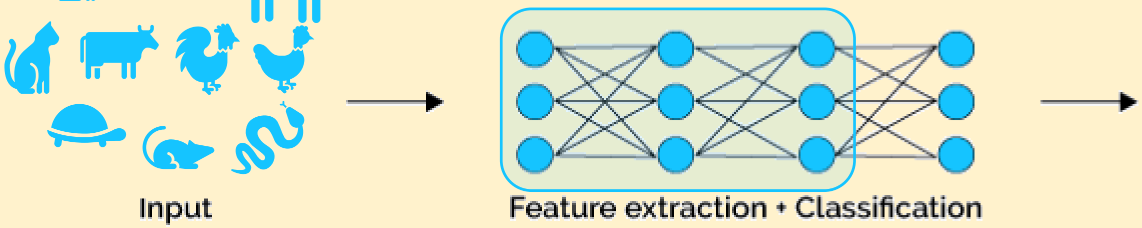


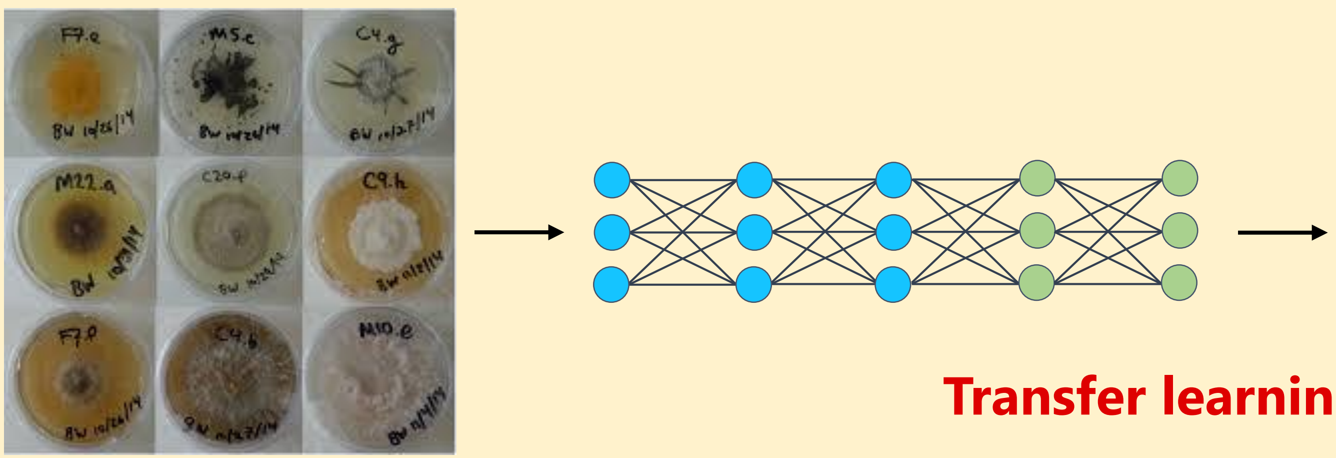


https://www.xenonstack.com/blog/data-science/log-analytics-deep-machine-learning-ai/

Predictability: Small data

Pre-training

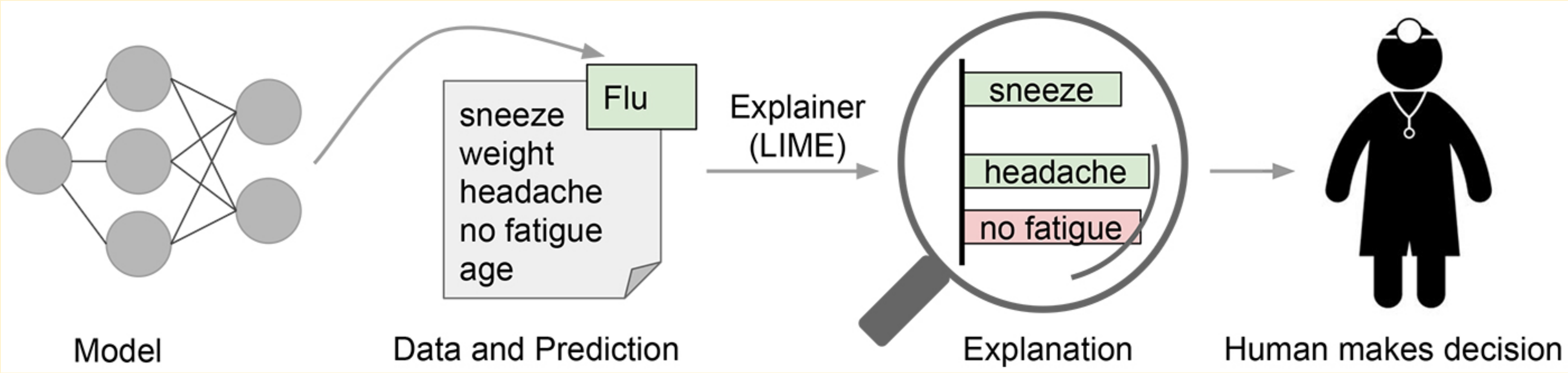


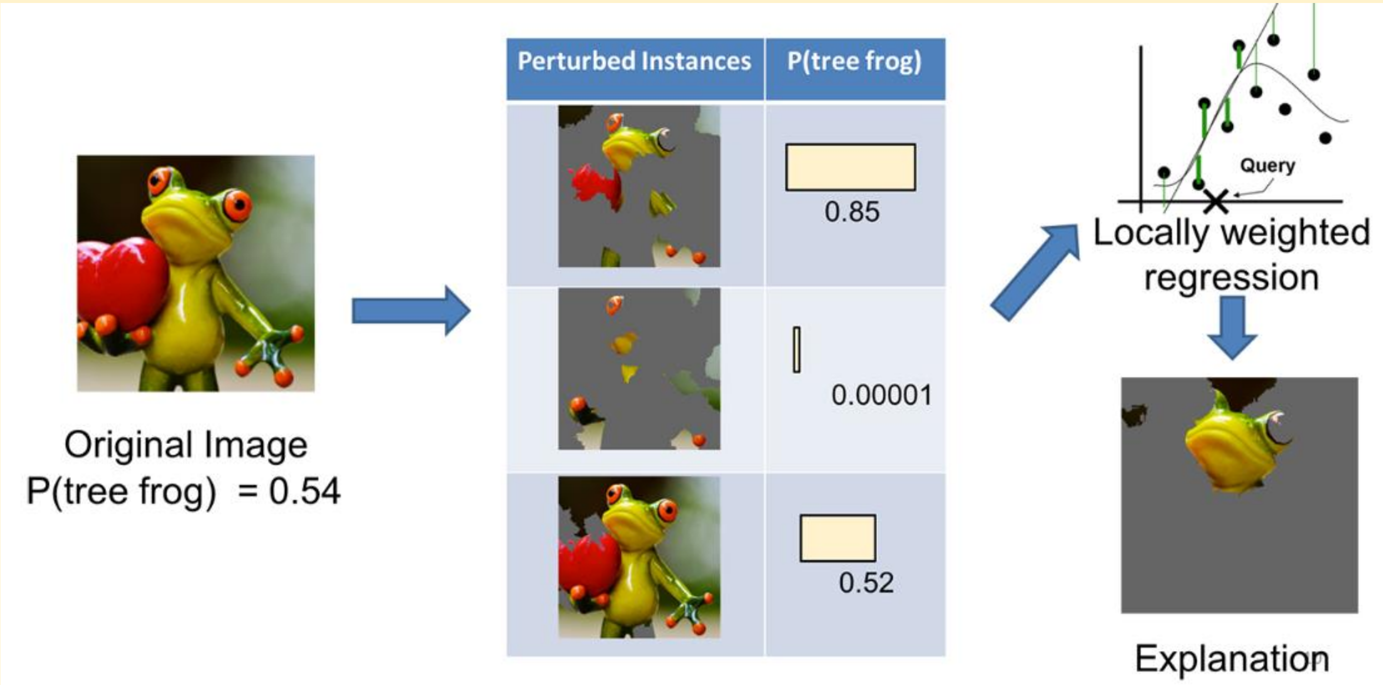


Taxonomic
& functional
identity

Transfer learning

Opening a black-box: LIME (Local Interpretable Model-Agnostic Explanations)





In a nutshell

Predictability

Big-data: *Deep learning*

Small-data: *Transfer learning*

Interpretability

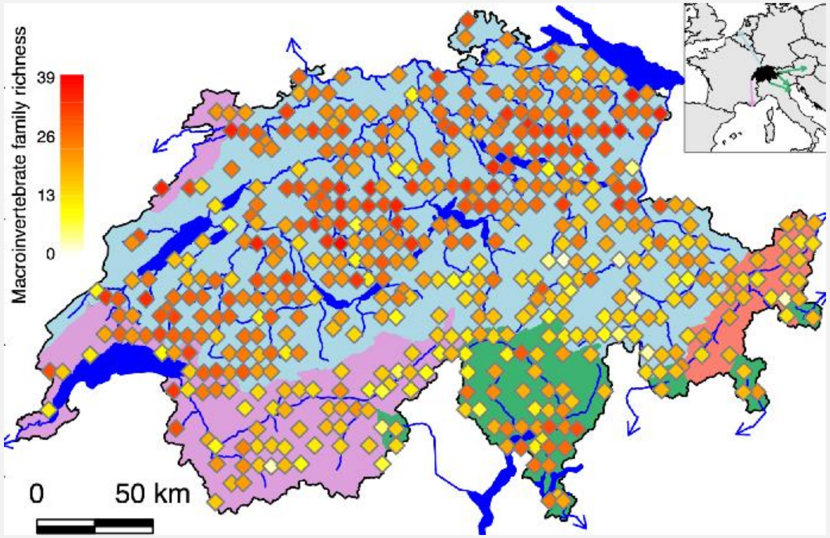
Interaction detection

Integration with statistics

cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

? What are the most important abiotic interactions?



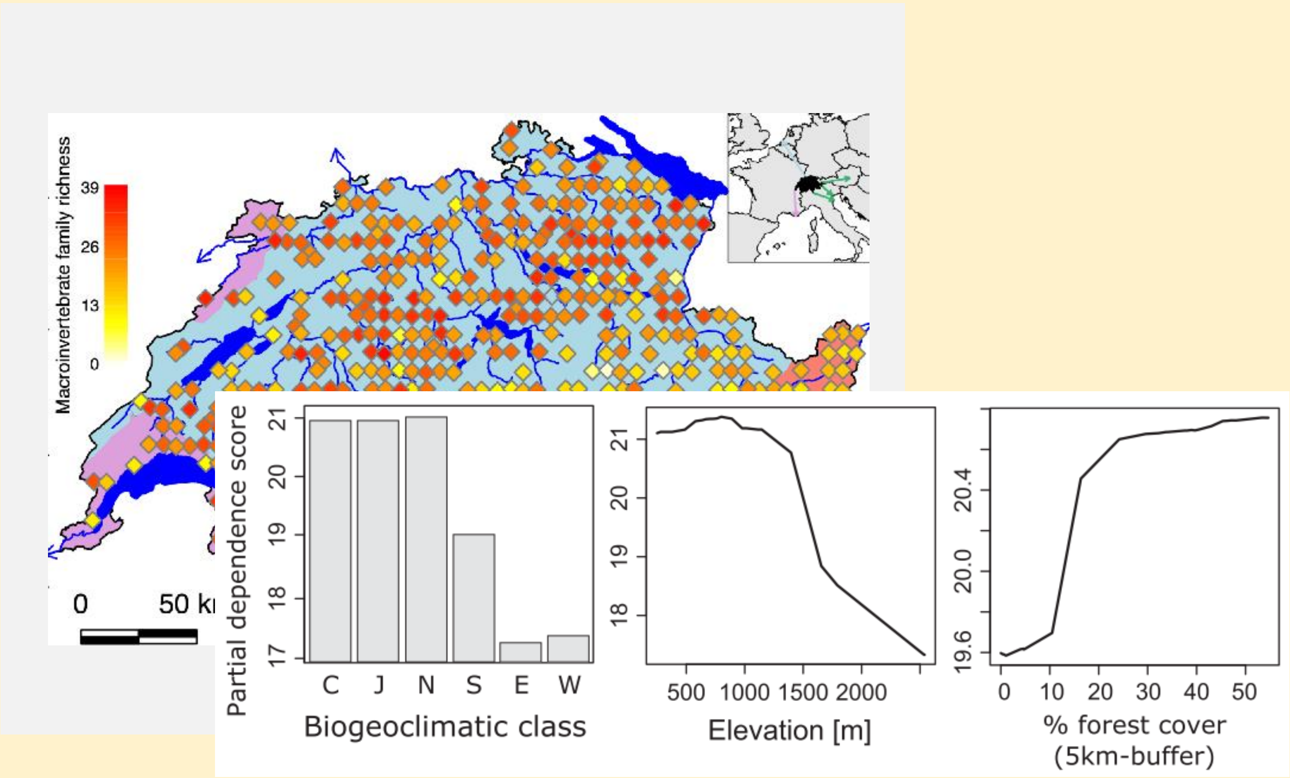
Macroinvertebrate diversity in Swiss rivers (n = 518)

- Family richness (α-diversity)
- **70** abiotic factors
- Nonlinear interactions of abiotic factors are often fully neglected at the regional scale

cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

? What are the most important abiotic interactions?



cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

Variable selection

Testing all 3-way combinations

Finding imporant interactions

Random Forest testing significance of each predictor

- **70** factors
- **2415** of 2-way interactions
- **54740** of 3-way interactions

- **20** factors
- **190** of 2-way interactions
- **1140** of 3-way interactions

cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

Variable selection

↓

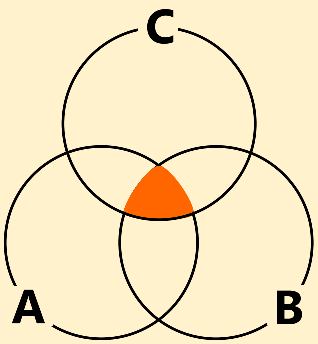
Testing all 3-way combinations

↓

Finding imporant interactions

Mutual Information Theory

cf. Kelly & Okada (2012)



Interaction importance

***I*(A∩B ∩C)**

cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

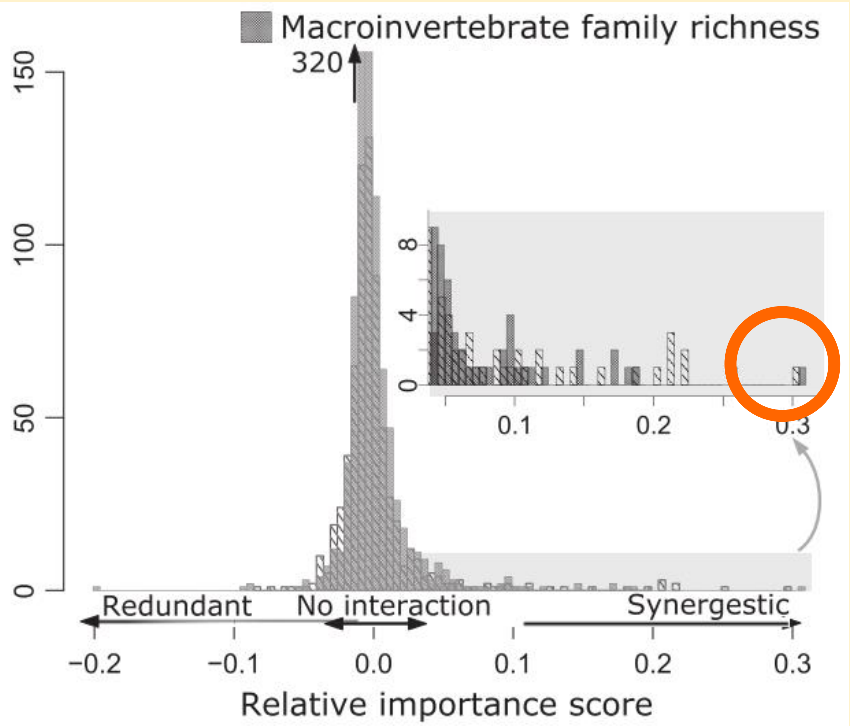
Variable selection

↓

Testing all 3-way combinations

↓

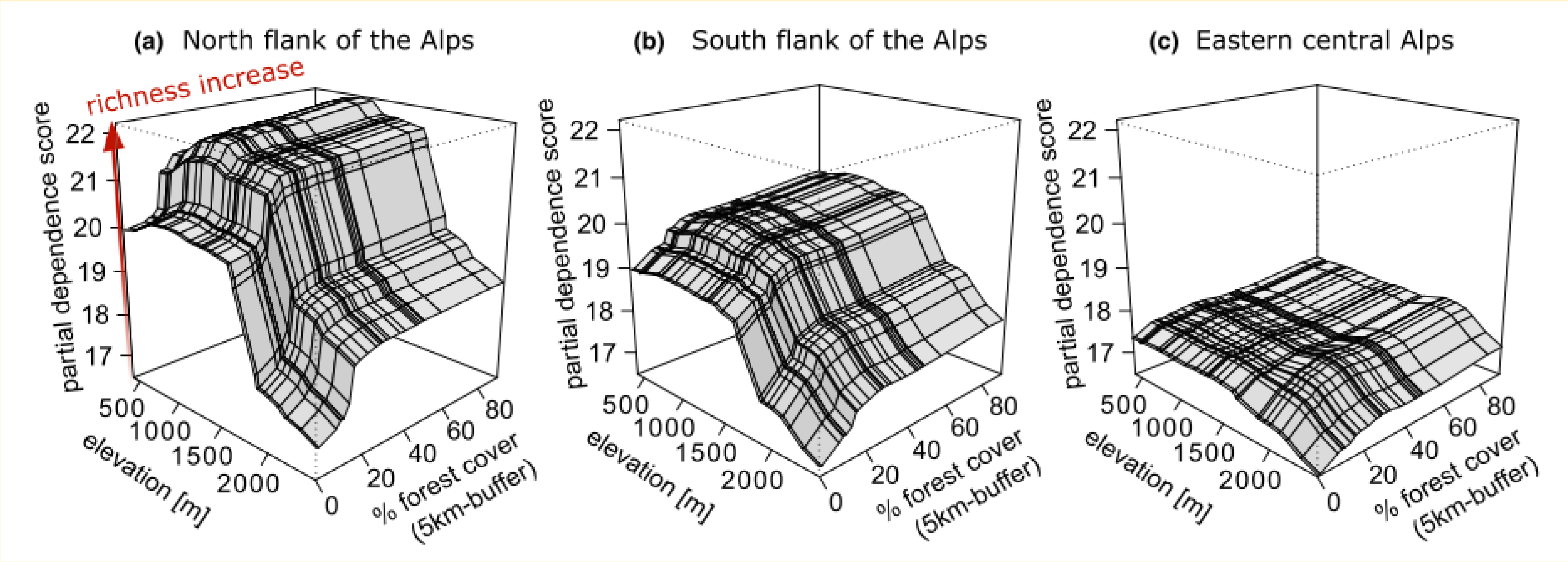
Finding imporant interactions



cf. Ryo et al. (2018)

Nonlinear interactions explains diversity pattern

Elevation ✖ Forest coverage ✖ Geographic region

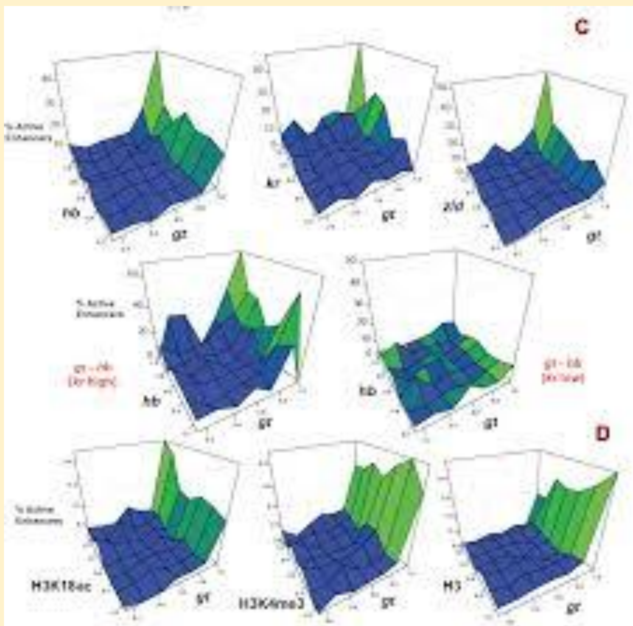


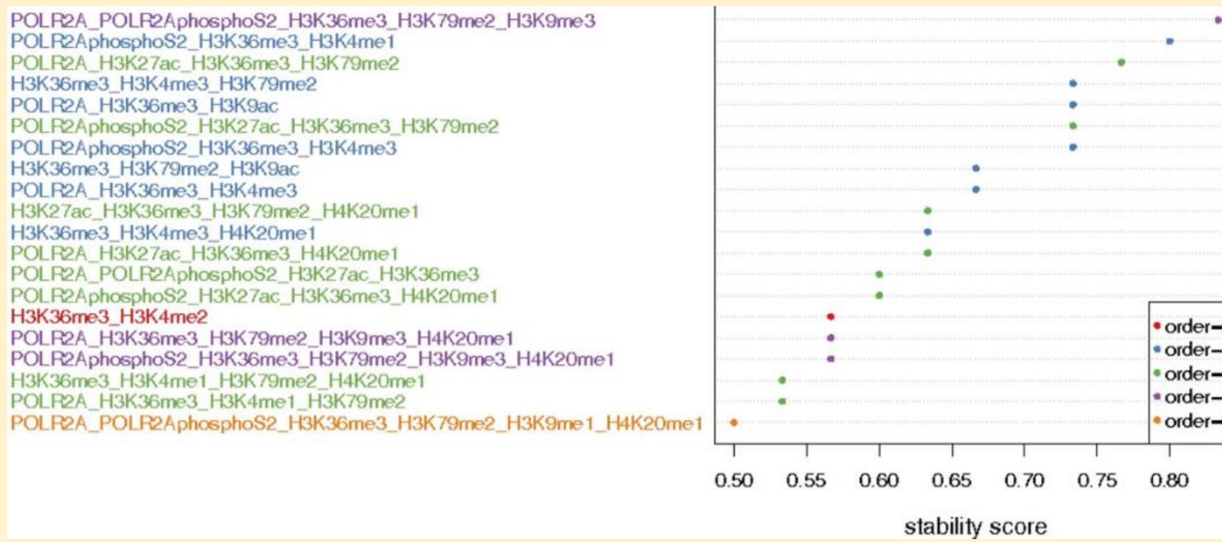
More crazy interaction detection... but is it accurate?

Iterative random forests to discover predictive and stable high-order interactions

Sumanta Basu, Karl Kumbier, James B. Brown, and Bin Yu

PNAS February 20, 2018 115 (8) 1943-1948; published ahead of print January 19, 2018





Do little interactions get lost in dark random forests?

March 2016 · BMC Bioinformatics 17(1):145
DOI: 10.1186/s12859-016-0995-8

Marvin N. Wright · Andreas Ziegler · Inke König

In a nutshell

Predictability

Big-data: *Deep learning*

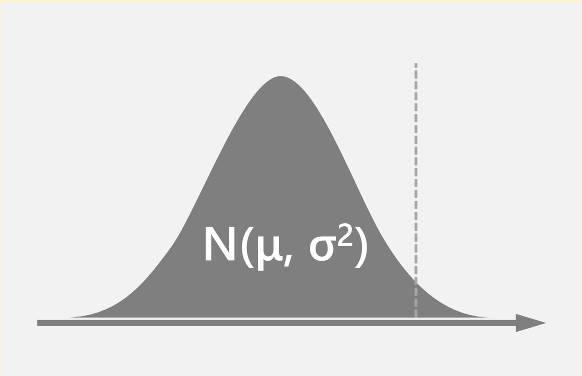
Small-data: *Transfer learning*

Interpretability

Interaction detection: *controversial but needed*

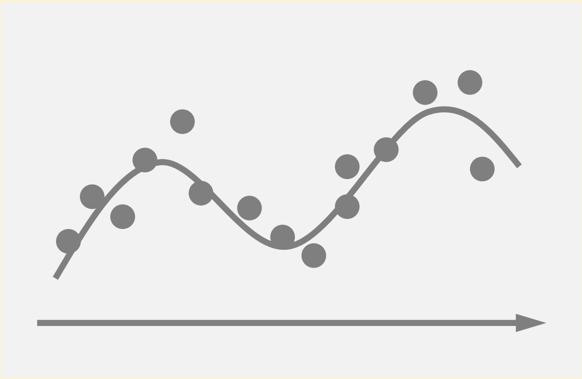
Integration with statistics

Statistics vs Machine learning?



Statistics

- Hypothesis-testing, theory-driven
- Strict assumptions
  (e.g. Linearity, normality, additivity)
- Probability



Machine learning

- Information-searching, data-driven
  (missing data)
- Loose assumptions
  (e.g. non-linearity, non-normality, non-additivity)
- Predictability

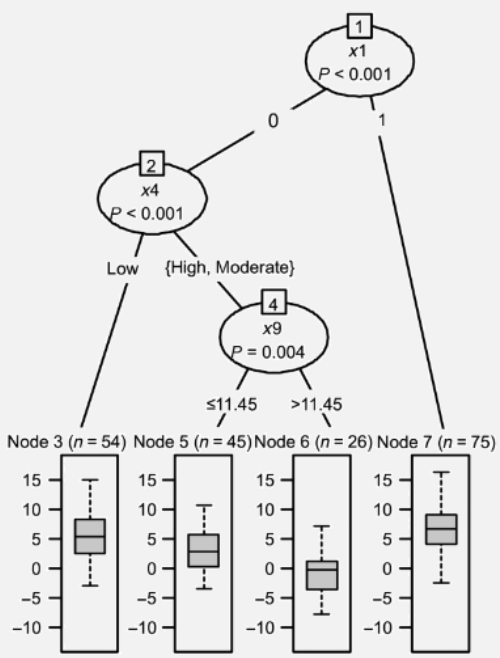
Breiman, 2001

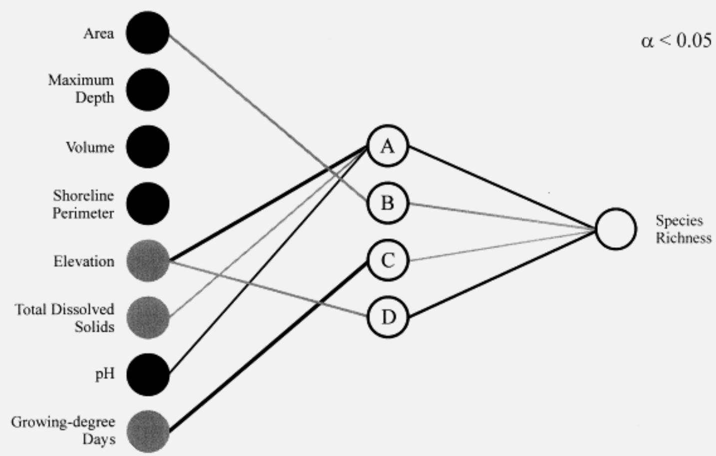
Statistically reinforced machine learning

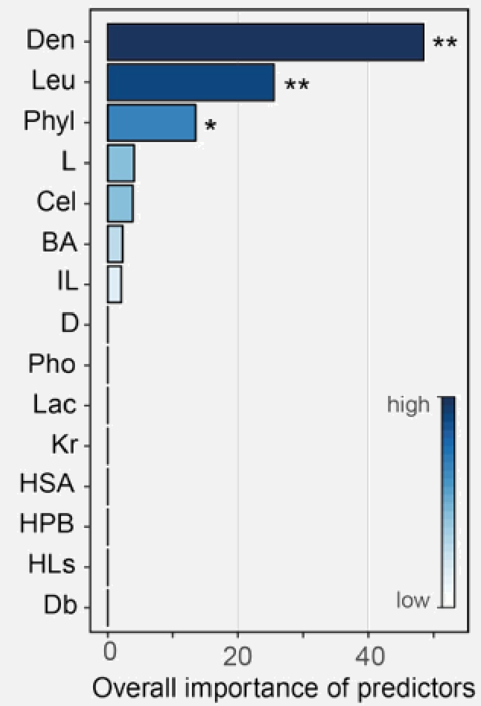
EMERGING TECHNOLOGIES

Statistically reinforced machine learning for nonlinear patterns and variable interactions

Masahiro Ryo[1,2,†] and Matthias C. Rillig[1,2]



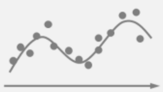
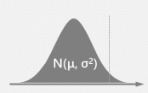




cf. Ryo & Rillig (2017)

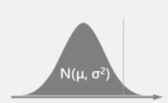
Statistically reinforced machine learning

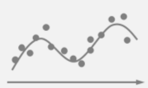
High predictability & model-free hypothesis test

 **Prediction** with 

p-value Variable selection

Using only useful info. increases model performance

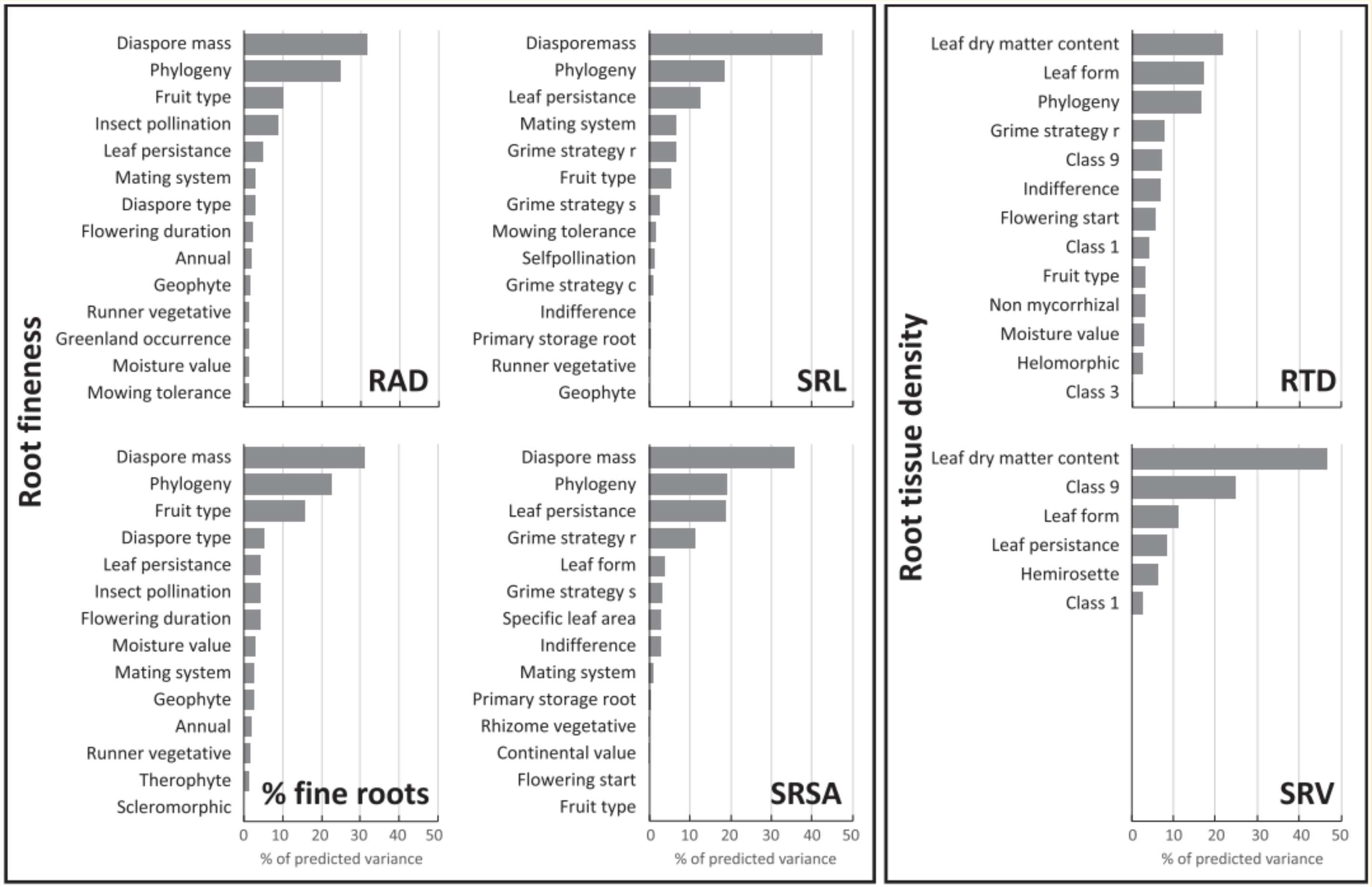


Hypothesis-testing with  Information-searching

Discovering nonlinearity & interactive effect without *a priori* assumption

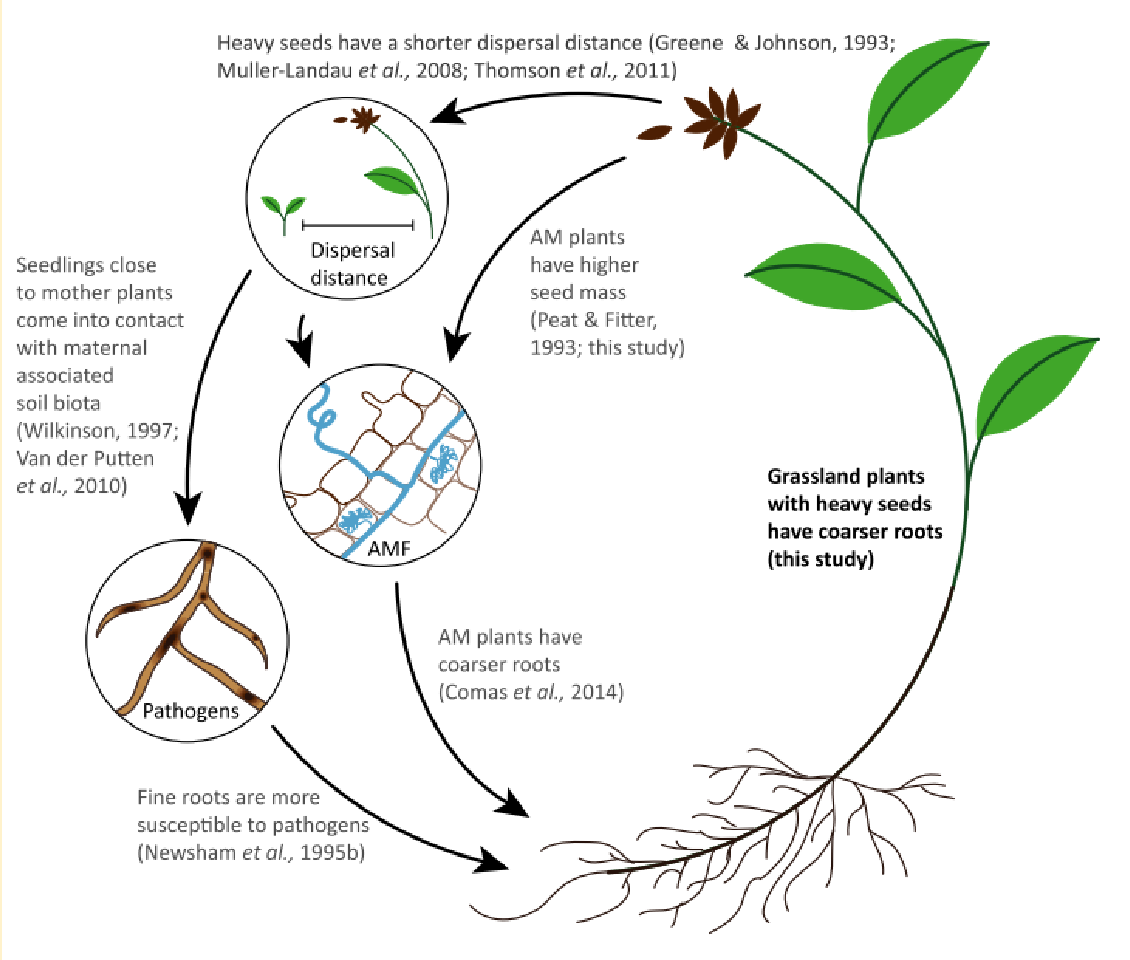
Root traits are more than analogues of leaf traits: the case for diaspore mass

Joana Bergmann[1,2], Masahiro Ryo[1,2], Daniel Prati[3], Stefan Hempel[1,2] and Matthias C. Rillig[1,2]

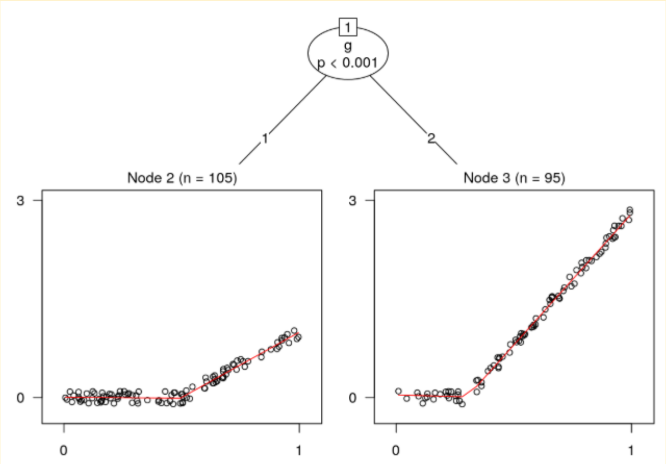


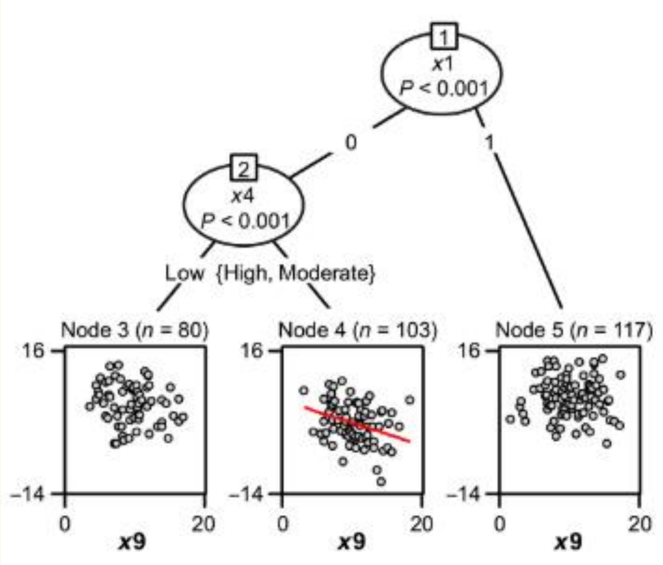
Root traits are more than analogues of leaf traits: the case for diaspore mass

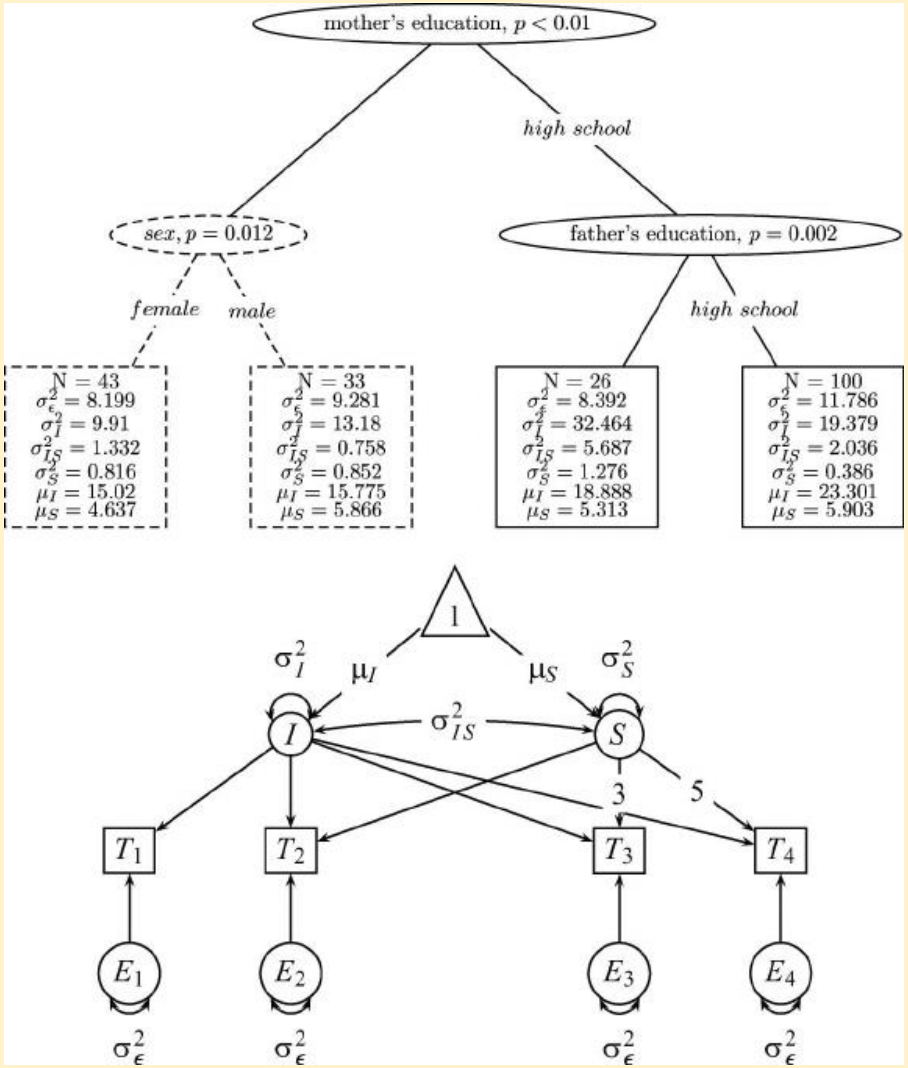
Joana Bergmann[1,2], Masahiro Ryo[1,2], Daniel Prati[3], Stefan Hempel[1,2] and Matthias C. Rillig[1,2]



Parameter estimates of stats model with ML?







SUMMARY

PART1: What is machine learning?

- Like statistics: regression, classification, clustering
- No *a priori* assumptions on data structure
- Nonlinear, non-additive modeling
- OK with Missing values

PART2: What are the recent advances?

- Along predictability-interpretability tradeoff
- Predictability: deep learning with big-data & small-data
- Interpretability: interaction detection & mix w/ statistics

Acknowledgement

Bridging in Biodiversity Science - BIBS

VERBUNDPROJEKT

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung