

# Recent advancements in machine learning

Masahiro Ryo

Free University of Berlin

Berlin-Brandenburg Institute of Advanced Biodiversity Research



# **PART 1**

## **Introduction to Machine Learning**

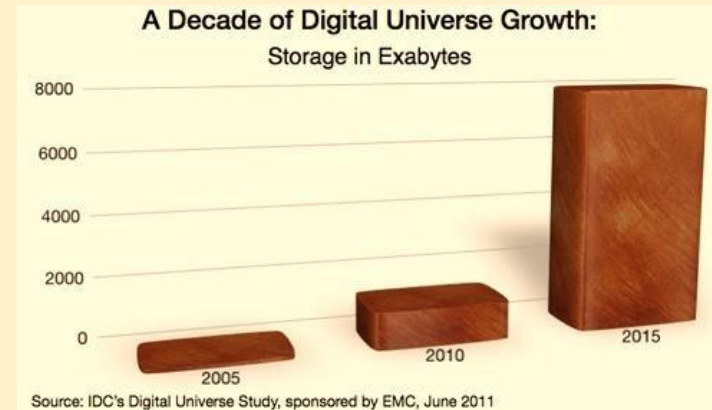
# Why Machine learning?

## 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<sup>1\*</sup>, Carly A Strasser<sup>2</sup>, Joshua J Tewksbury<sup>3</sup>, Wendy K Gram<sup>4</sup>, Amber E Budden<sup>5</sup>, Archer L Batcheller<sup>6</sup>, Clifford S Duke<sup>7</sup>, and John H Porter<sup>8</sup>



### Environmental Modelling & Software

Volume 63, January 2015, Pages 185–198



Review

## Web technologies for environmental Big Data

Claudia Vitolo<sup>a</sup>, Yehia Elkhatib<sup>b</sup>, Dominik Reusser<sup>c</sup>, Christopher J.A. Macleod<sup>d</sup>, Wouter Lavrya<sup>e</sup>



CITIZEN  
SCIENCE



Google earth



# What is Machine learning?

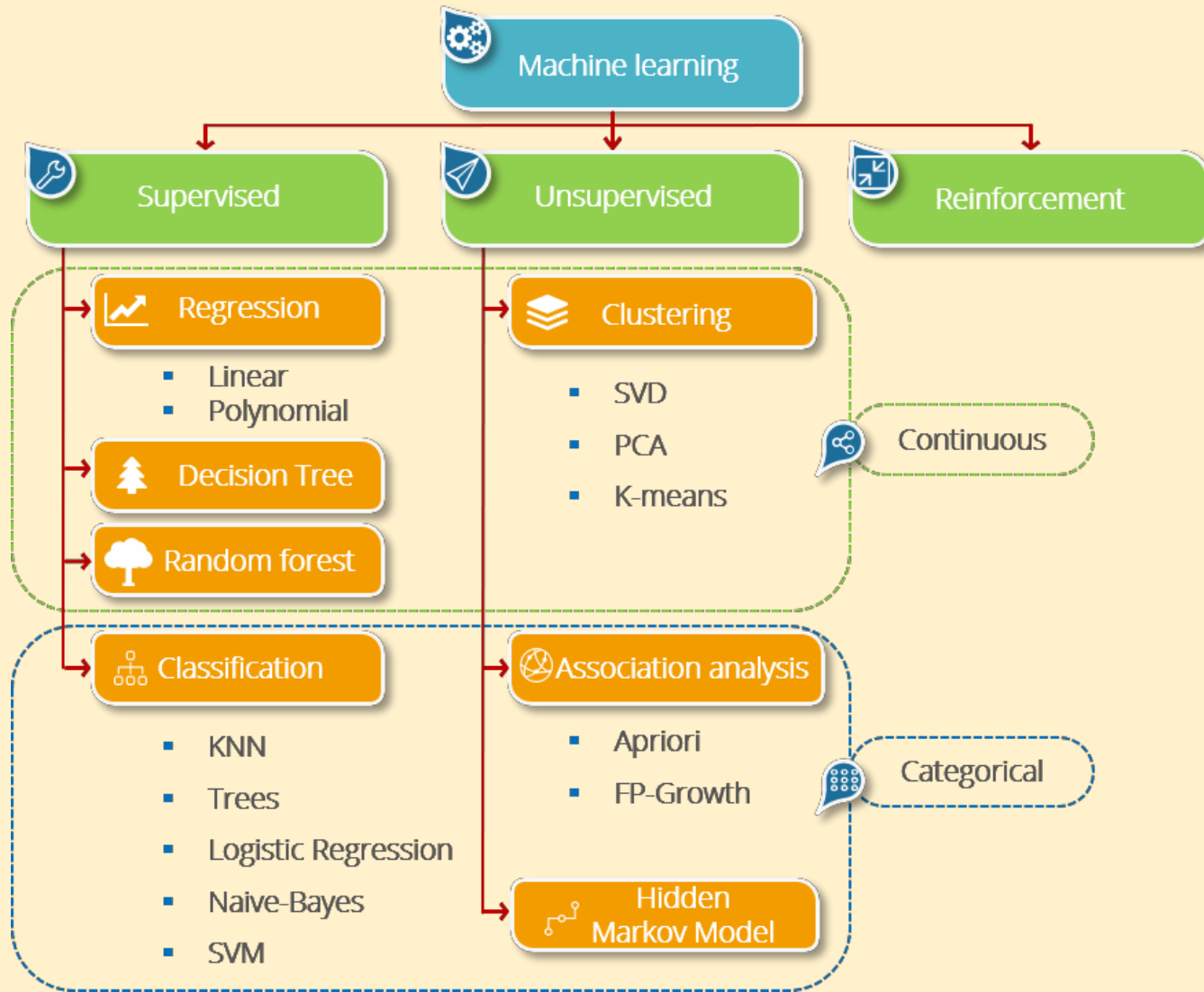
*"A field of study that gives computers the ability to learn without being explicitly programmed"*

first definition by Samuel A. (1959)

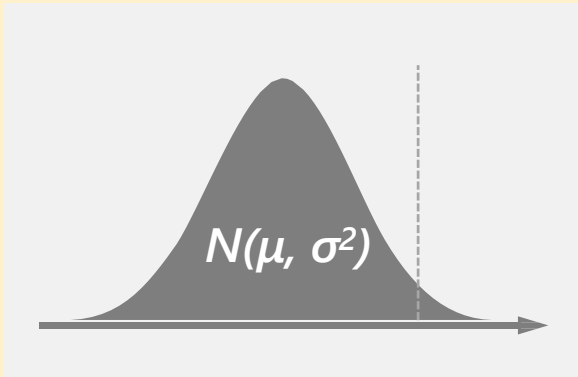
*"A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ "*

*popular definition by Mitchell T.M. (1997)*

# What Machine learning does?

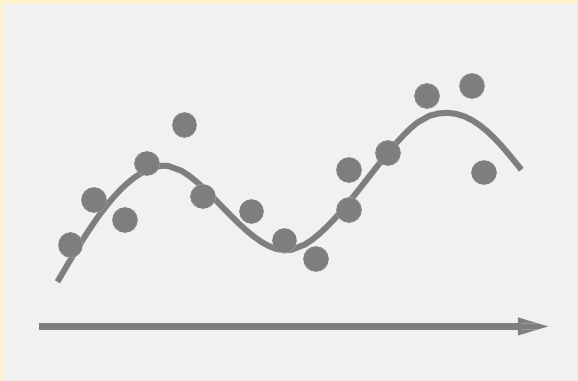


# Statistics vs Machine learning?



## 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

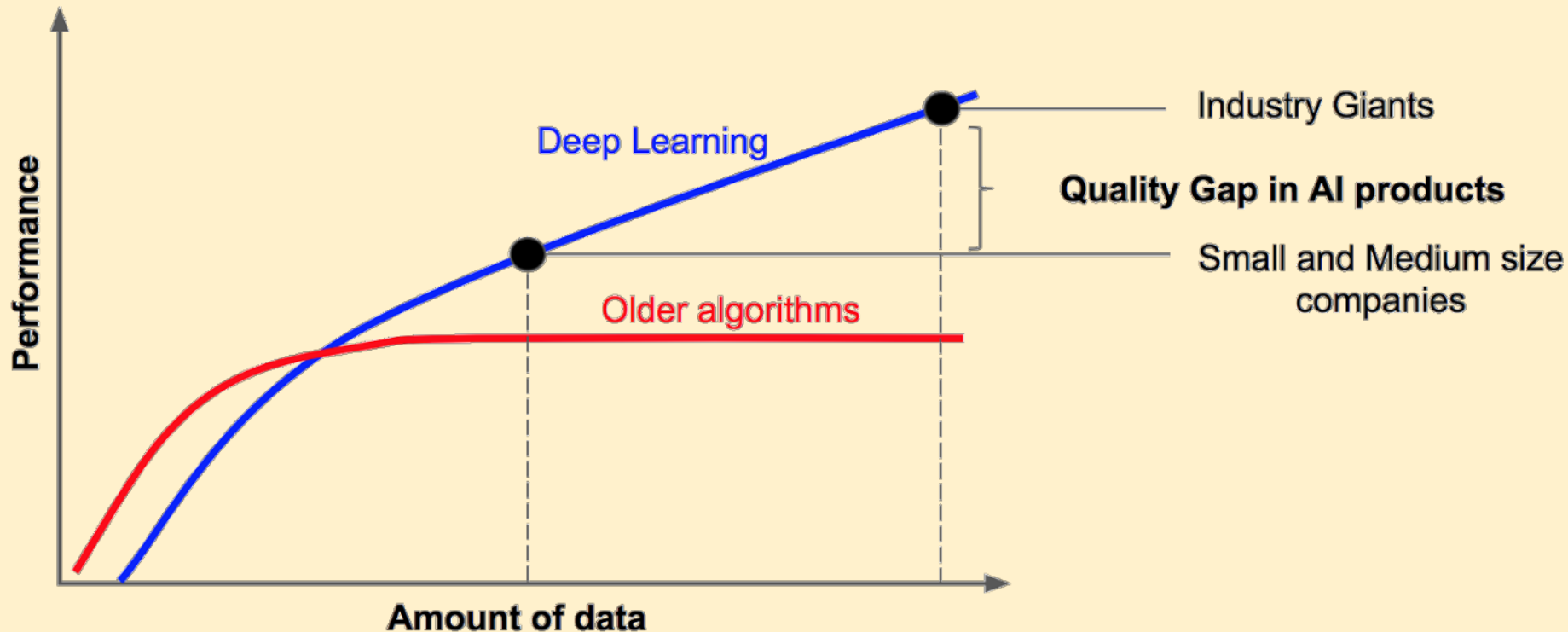
Small-data

## **Interpretability**

Pattern detection & visualization

Integration with statistics

# Predictability: The power of big data



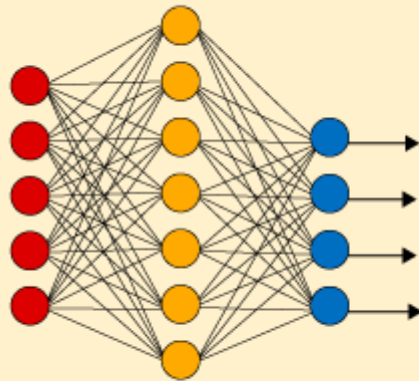
So many DL algorithms have been proposed since 2006

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

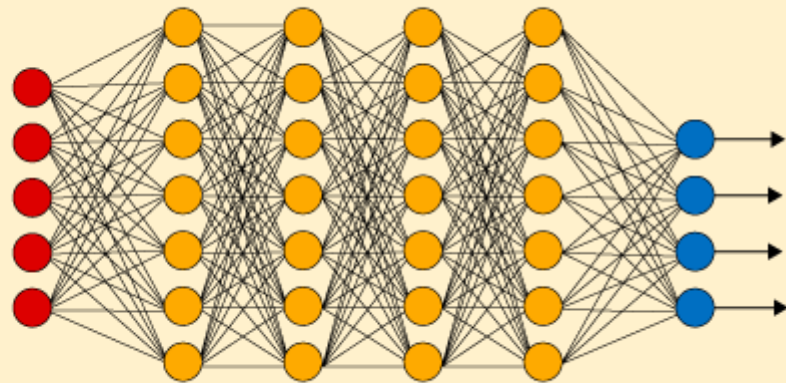
<https://www.slideshare.net/ExtractConf>

# Predictability: The power of big data

**Simple Neural Network**

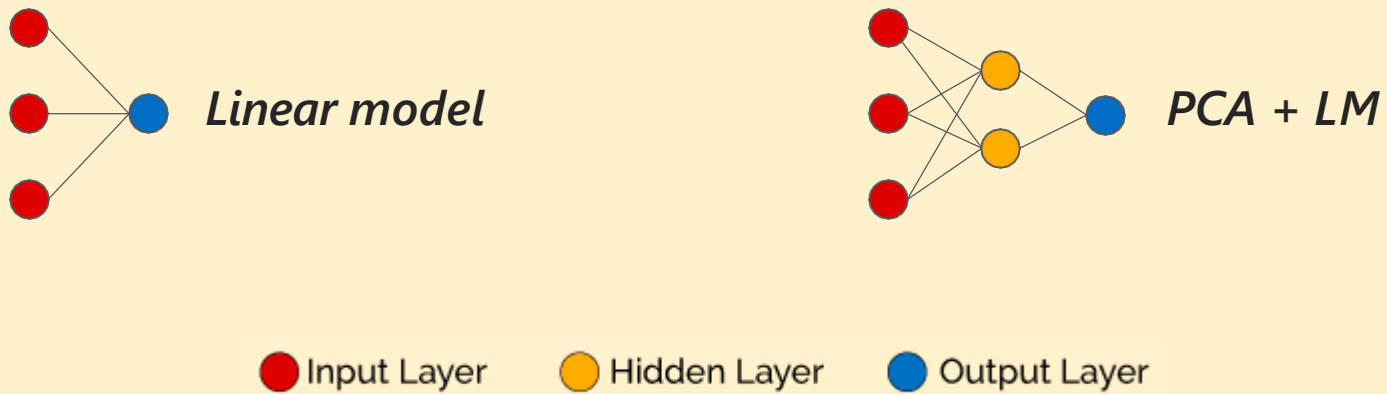


**Deep Learning Neural Network**



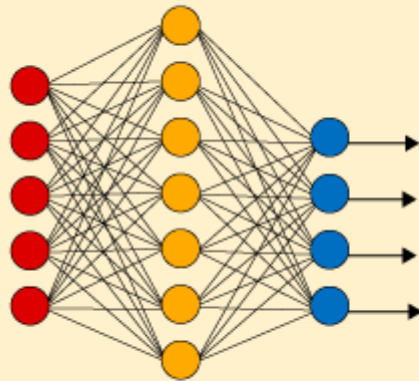
● Input Layer    ● Hidden Layer    ● Output Layer

# Predictability: The power of big data

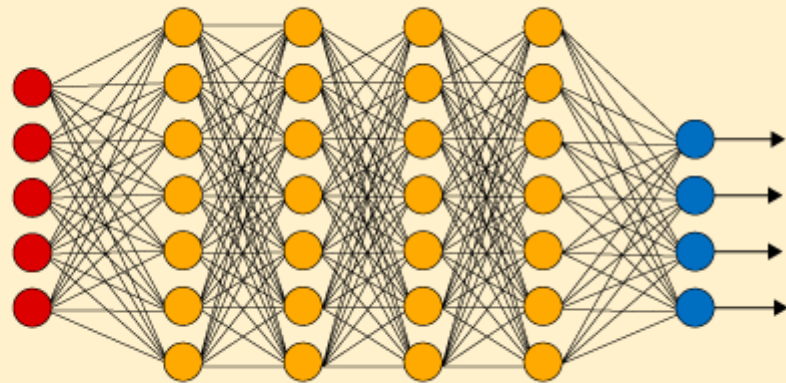


# Predictability: The power of big data

**Simple Neural Network**

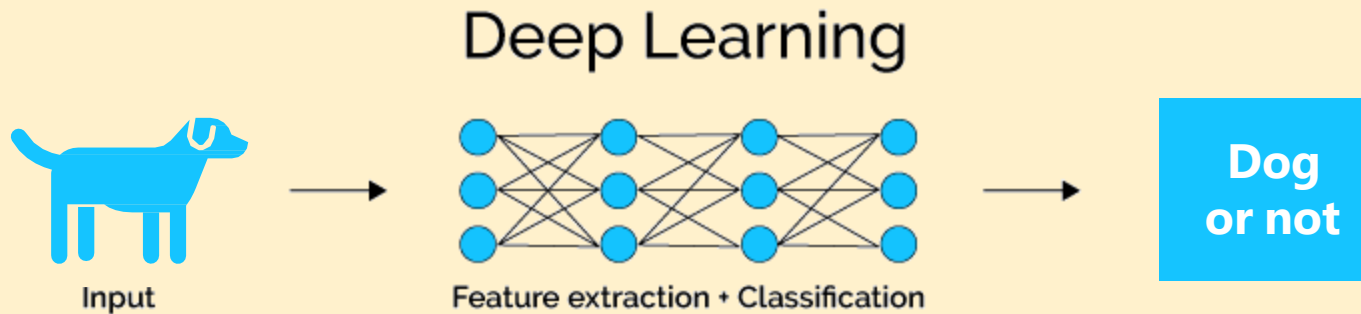
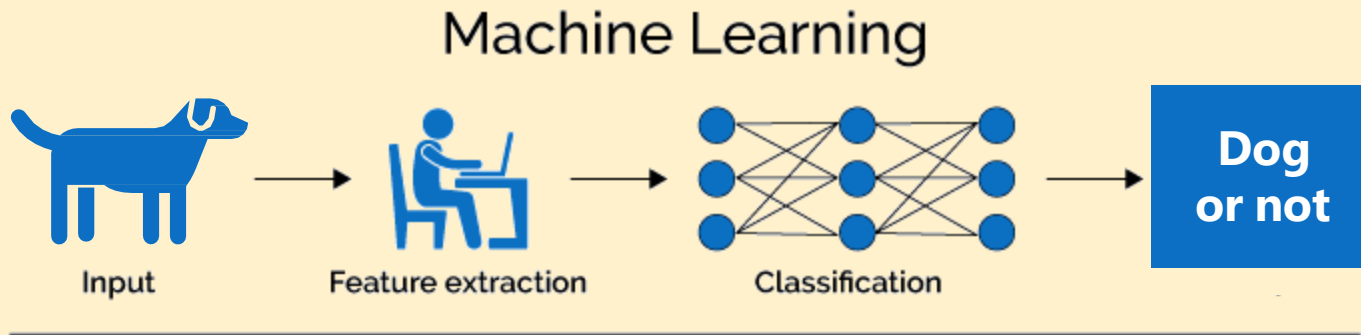


**Deep Learning Neural Network**



● Input Layer    ● Hidden Layer    ● Output Layer

# Predictability: The power of big data



# Predictability: The power of big data

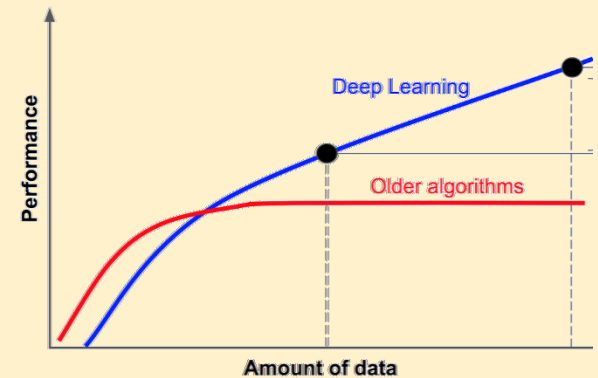
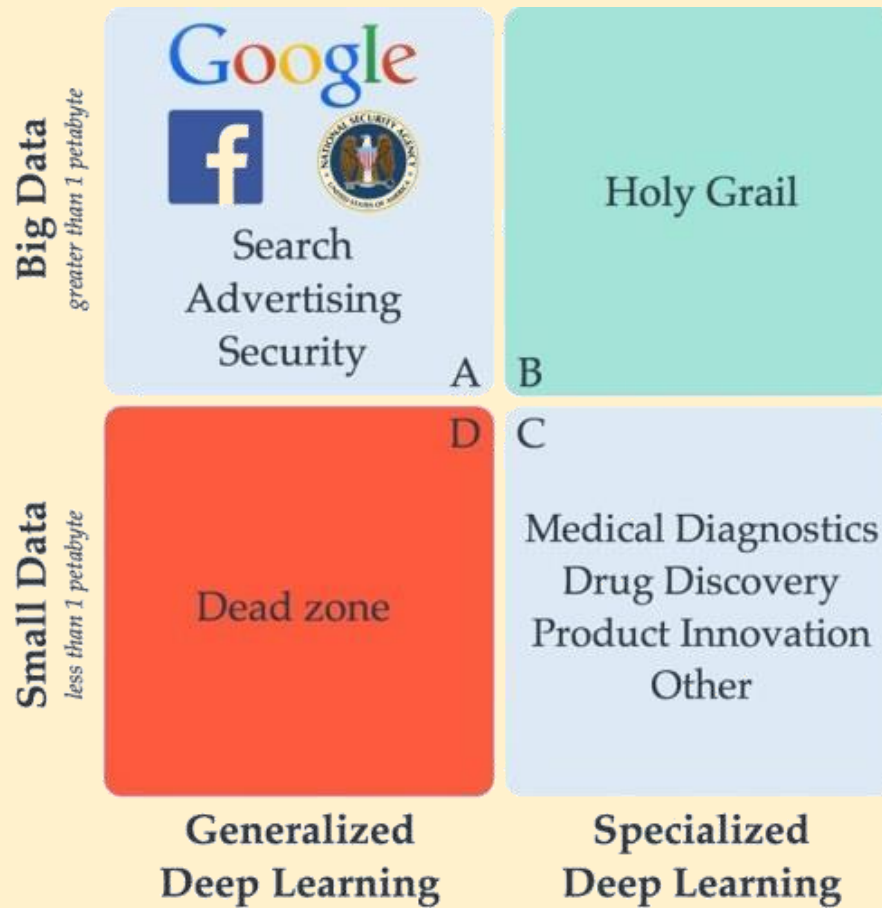
chihuahua or muffin

labradoodle or fried chicken



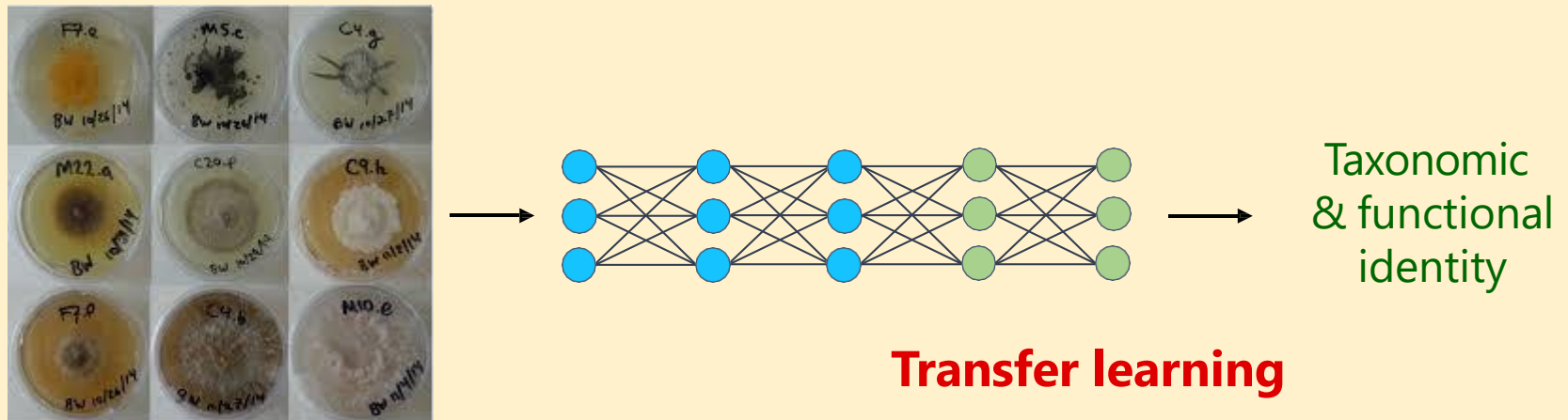
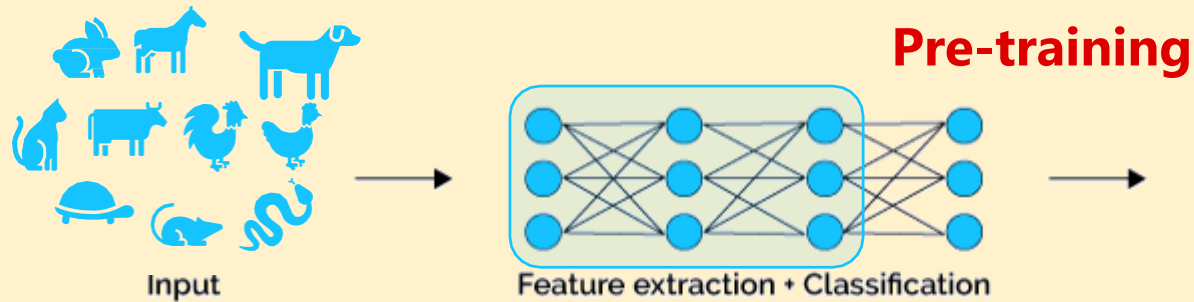


# Predictability: Small data

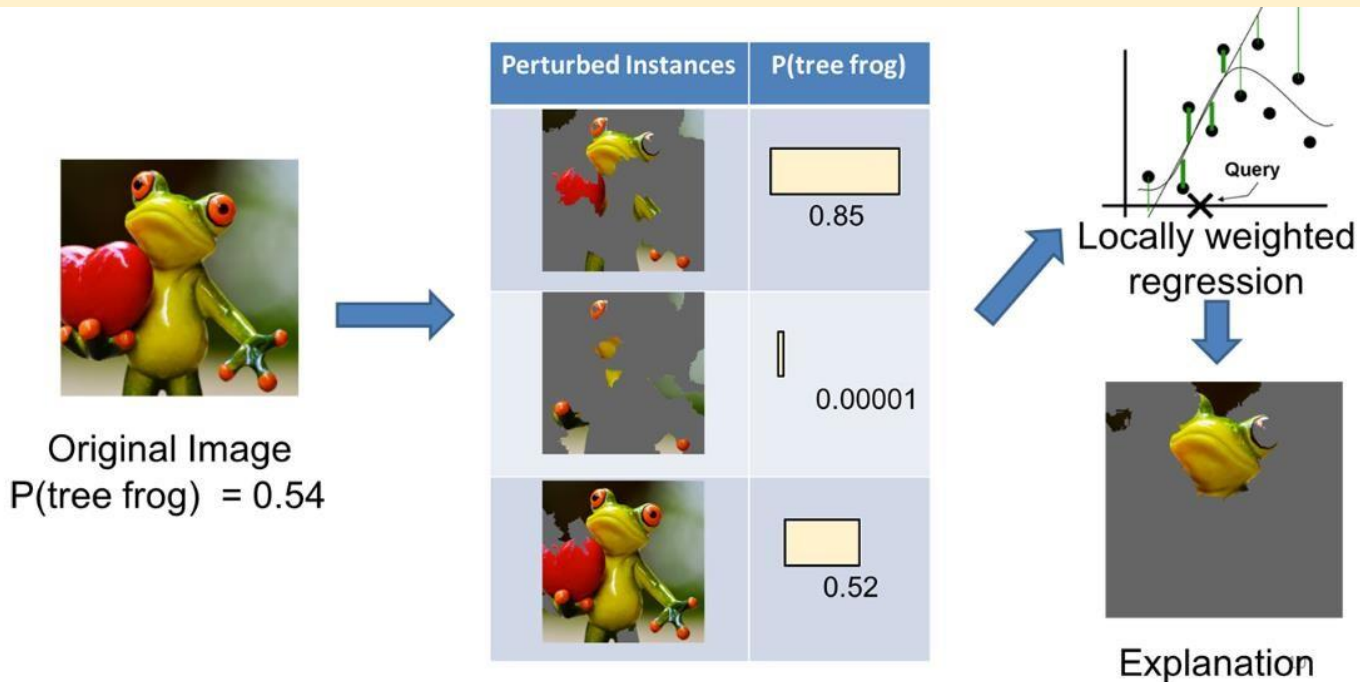
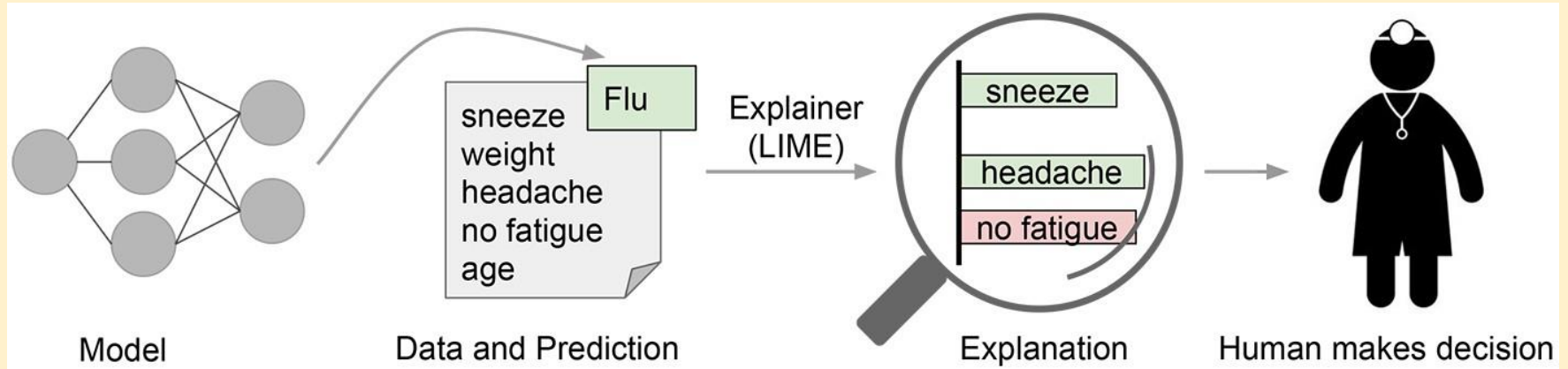




# Predictability: Small data



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



# In a nutshell

## **Predictability**

Big-data: *Deep learning*

Small-data: *Transfer learning*

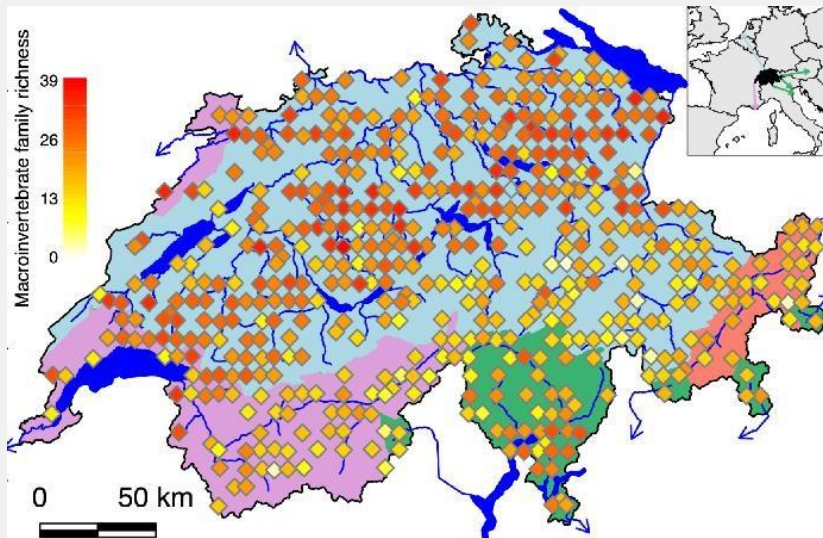
## **Interpretability**

Interaction detection

Integration with statistics

# Nonlinear interactions explains diversity pattern

? *What are the most important abiotic interactions?*

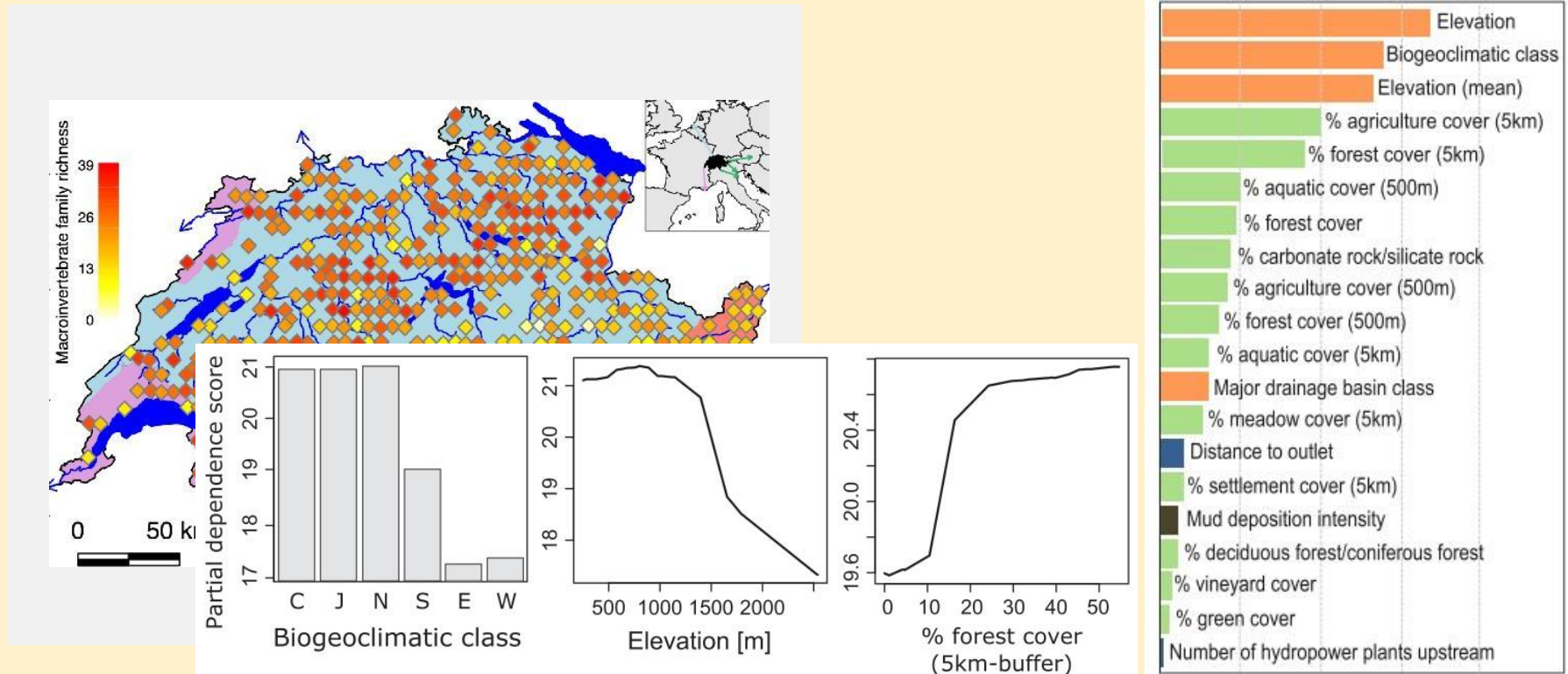


## *Macroinvertebrate diversity in Swiss rivers (n = 518)*

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

# Nonlinear interactions explains diversity pattern

? *What are the most important abiotic interactions?*



# Nonlinear interactions explains diversity pattern

*Variable selection*



*Testing all 3-way combinations*



*Finding important 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

# Nonlinear interactions explains diversity pattern

*Variable selection*



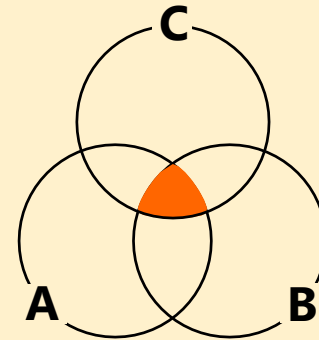
*Testing all 3-way combinations*



*Finding important interactions*

## Mutual Information Theory

cf. Kelly & Okada (2012)



Interaction importance  
 $I(A \cap B \cap C)$

# Nonlinear interactions explains diversity pattern

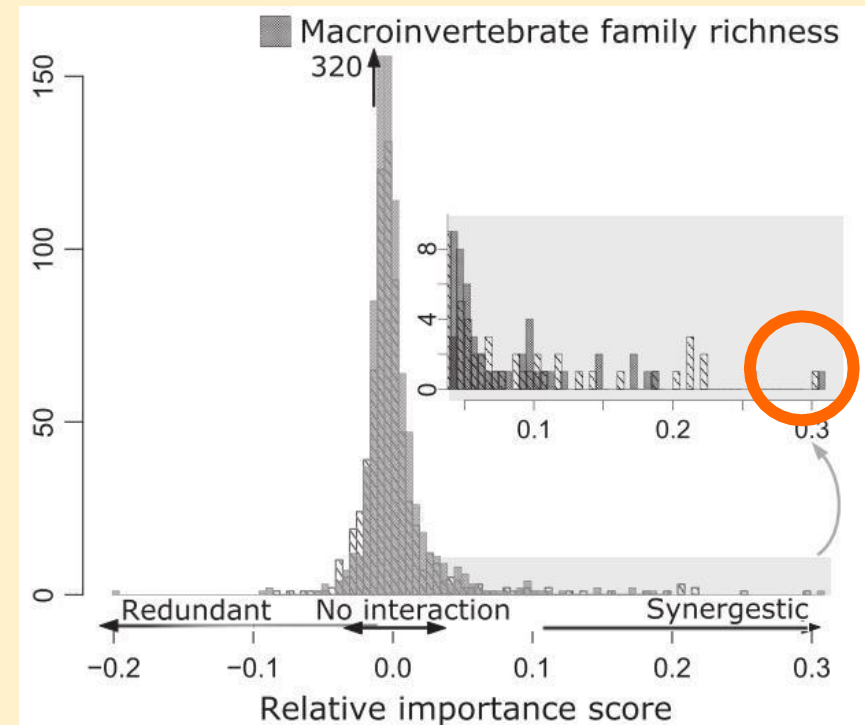
*Variable selection*



*Testing all 3-way combinations*



*Finding important interactions*

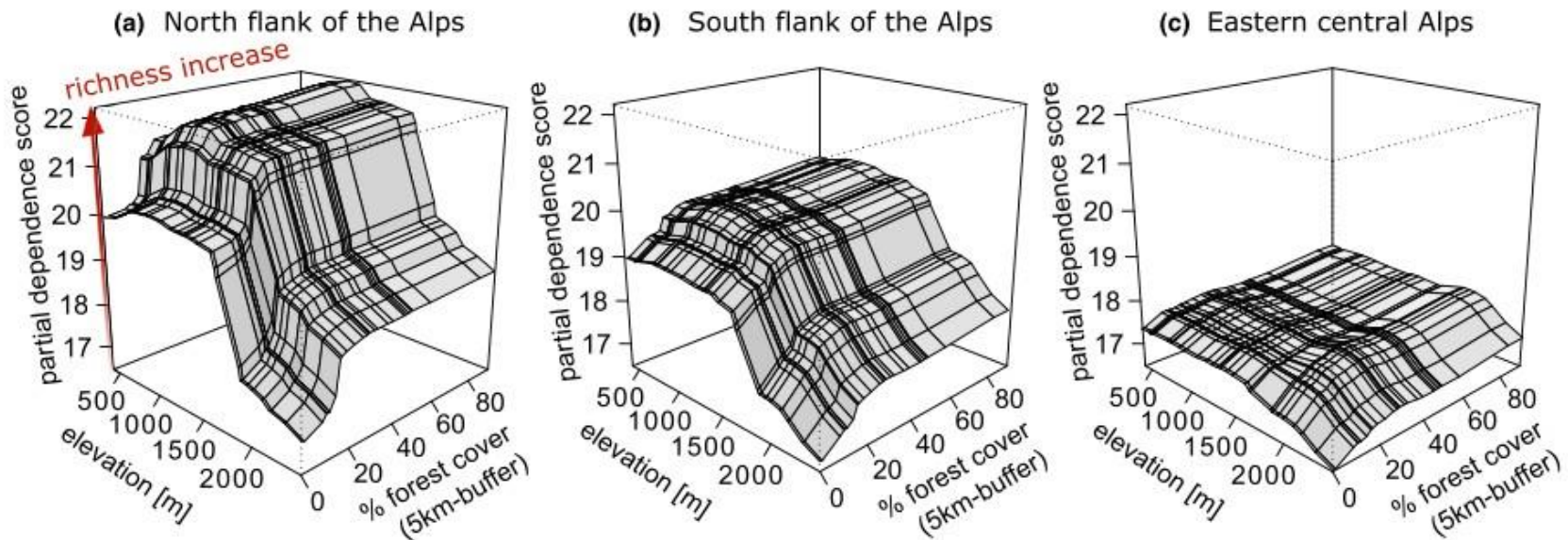




# Nonlinear interactions explains diversity pattern



*Elevation × Forest coverage × Geographic region*



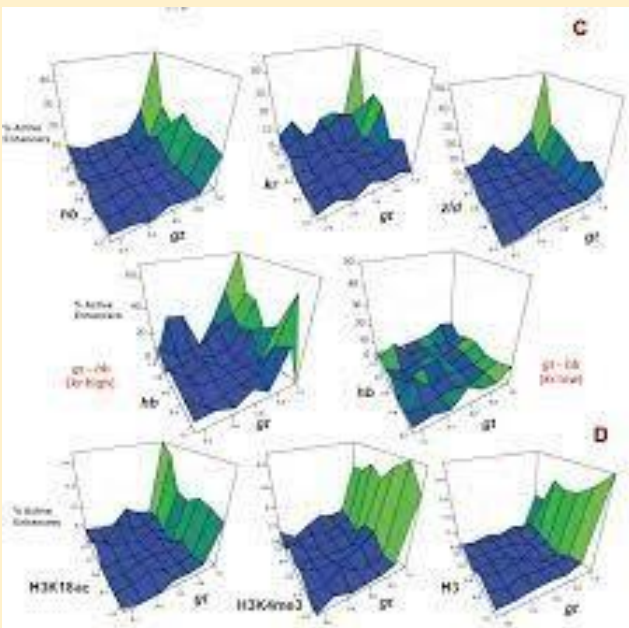
# More crazy interaction detection... but is it accurate?

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

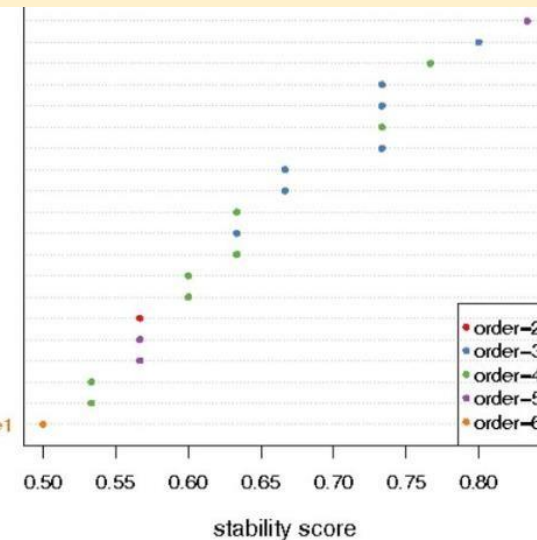


Sumanta Basu, Karl Kumbier, James B. Brown, and Bin Yu

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



POLR2A\_POLR2AphosphoS2\_H3K36me3\_H3K79me2\_H3K9me3  
 POLR2AphosphoS2\_H3K36me3\_H3K4me1  
 POLR2A\_H3K27ac\_H3K36me3\_H3K79me2  
 H3K36me3\_H3K4me3\_H3K79me2  
 POLR2A\_H3K36me3\_H3K9ac  
 POLR2AphosphoS2\_H3K27ac\_H3K36me3\_H3K79me2  
 POLR2AphosphoS2\_H3K36me3\_H3K4me3  
 H3K36me3\_H3K79me2\_H3K9ac  
 POLR2A\_H3K36me3\_H3K4me3  
 H3K27ac\_H3K36me3\_H3K79me2\_H4K20me1  
 H3K36me3\_H3K4me3\_H4K20me1  
 POLR2A\_H3K27ac\_H3K36me3\_H4K20me1  
 POLR2A\_POLR2AphosphoS2\_H3K27ac\_H3K36me3  
 POLR2AphosphoS2\_H3K27ac\_H3K36me3\_H4K20me1  
 H3K36me3\_H3K4me2  
 POLR2A\_H3K36me3\_H3K79me2\_H3K9me3\_H4K20me1  
 POLR2AphosphoS2\_H3K36me3\_H3K79me2\_H3K9me3\_H4K20me1  
 H3K36me3\_H3K4me1\_H3K79me2\_H4K20me1  
 POLR2A\_H3K36me3\_H3K4me1\_H3K79me2  
 POLR2A\_POLR2AphosphoS2\_H3K36me3\_H3K79me2\_H3K9me1\_H4K20me1



### Do little interactions get lost in dark random forests?

March 2016 · BMC Bioinformatics 17(1):145

DOI: 10.1186/s12859-016-0995-8

License · CC BY 4.0

Marvin N. Wright · Andreas Ziegler · Inke König

# In a nutshell

## **Predictability**

Big-data: *Deep learning*

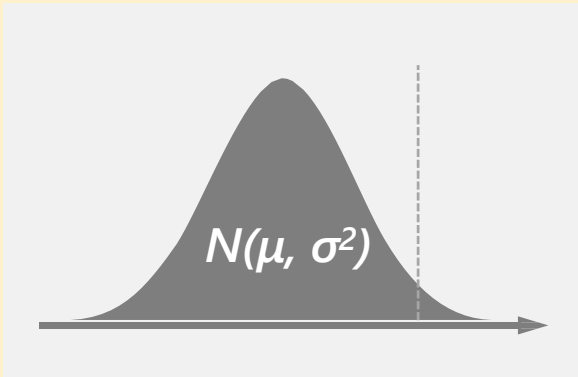
Small-data: *Transfer learning*

## **Interpretability**

Interaction detection: *controversial but needed*

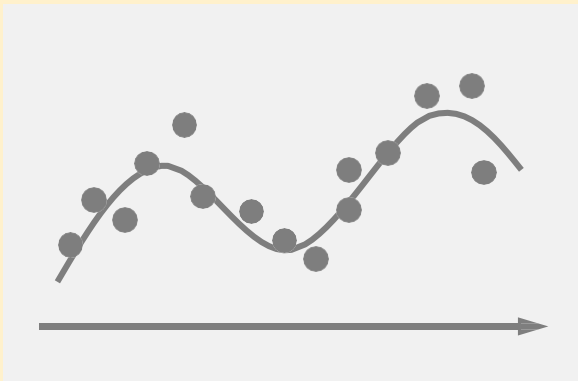
Integration with statistics

# 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**

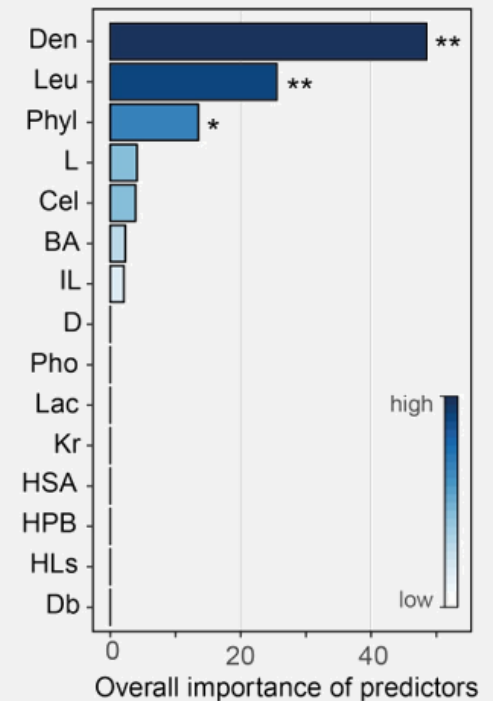
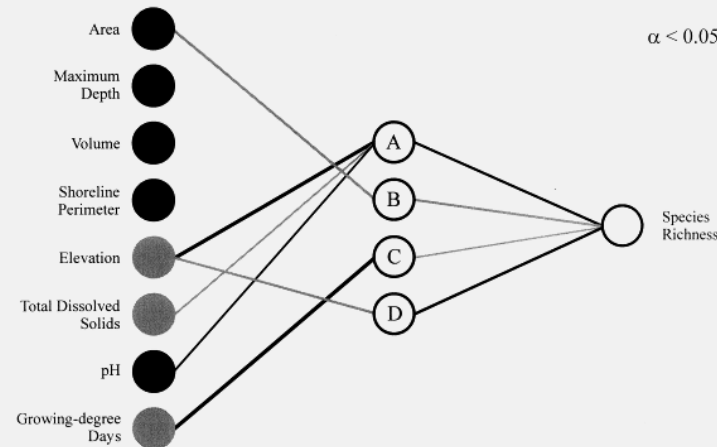
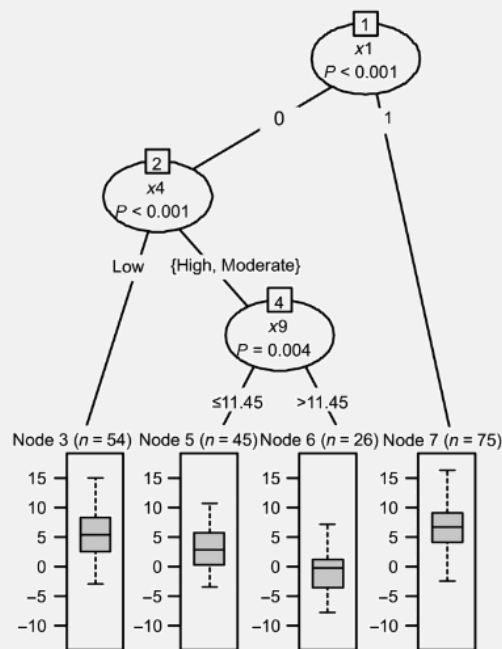
Breiman, 2001

# Statistically reinforced machine learning

## EMERGING TECHNOLOGIES

### Statistically reinforced machine learning for nonlinear patterns and variable interactions

MASAHIRO RYO<sup>1,2,†</sup> AND MATTHIAS C. RILLIG<sup>1,2</sup>



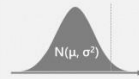
# Statistically reinforced machine learning



*High predictability & model-free hypothesis test*

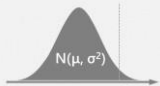


**Prediction** with



p-value Variable selection

*Using only useful info. increases model performance*



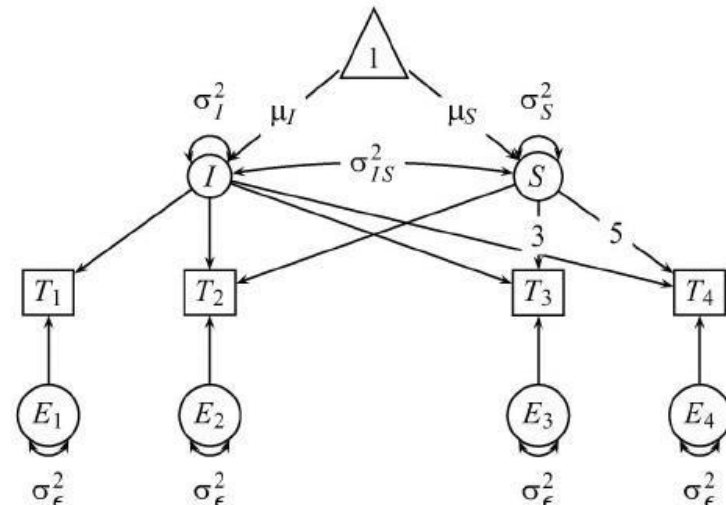
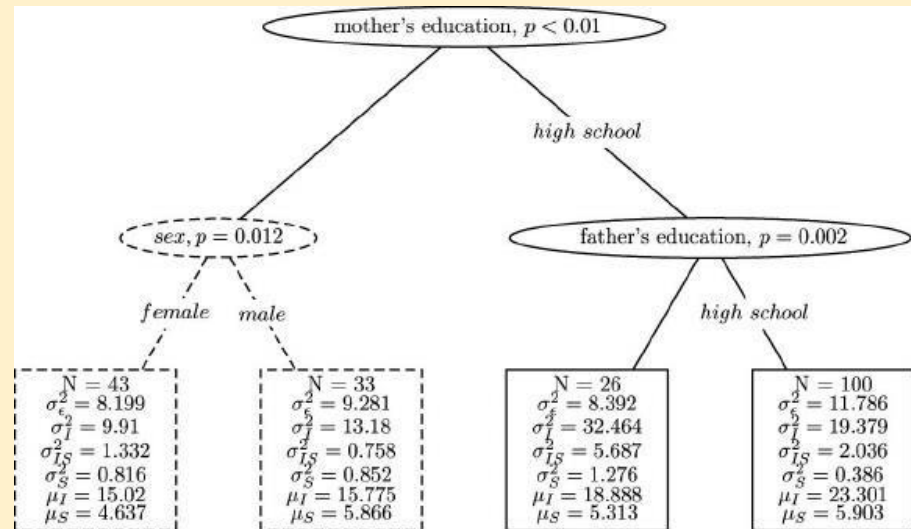
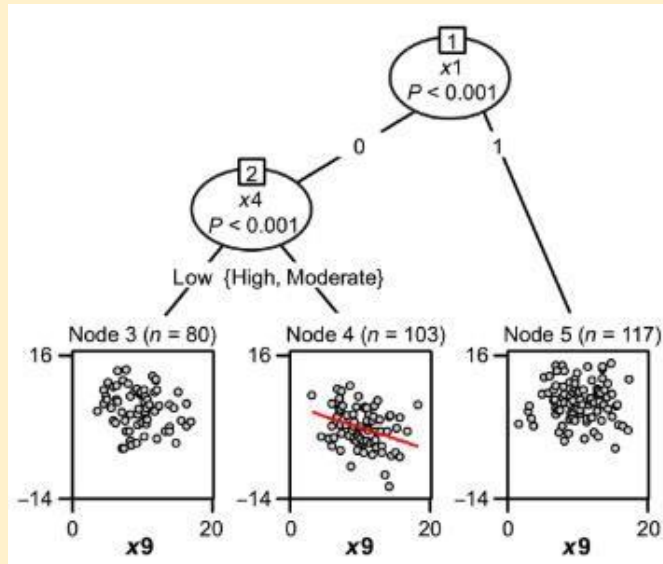
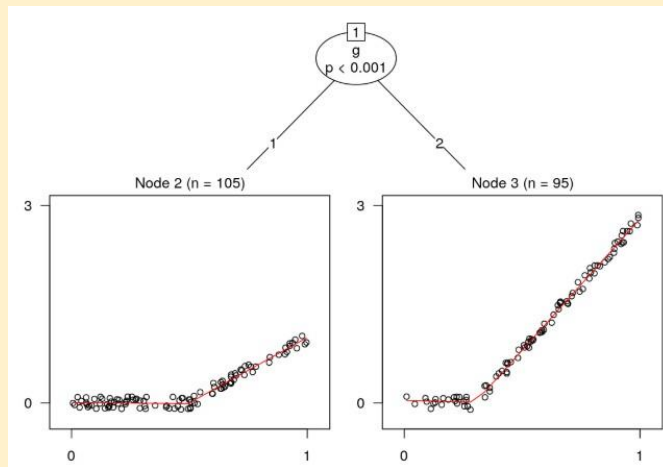
**Hypothesis-testing** with



Information-searching

*Discovering nonlinearity & interactive effect  
without a priori assumption*

# 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



Matthias C. Rillig,  
the group members,  
Boyan Angelov  
... & Susana Martinez,  
RSG Luxemburg