Statistically reinforced machine learning for nonlinear interactions of factors & hierarchically nested spatial patterns

Masahiro Ryo & Matthias C. Rillig

Free University of Berlin

Berlin-Brandenburg Institute of Advanced Biodiversity Research

R2.1: Understanding species distribution, population dynamics and phenology by machine learning
10th International Conference on Ecological Informatics 2018, Jena, Germany. 2018.09.20.

Recipe

#1 Statistics + Machine learning

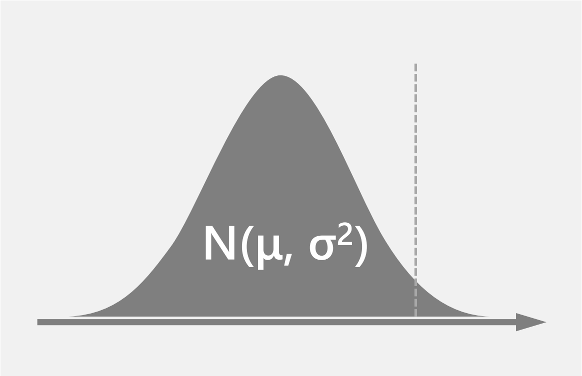
#2 Variable interactions detection

#3 Multiscale autocorrelation

cf. Ryo & Rillig (2017)

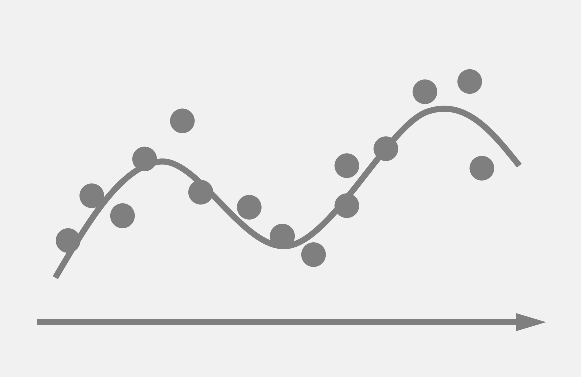
#1 Statistically reinforced machine learning?

= Statistics + Machine learning



Statistics

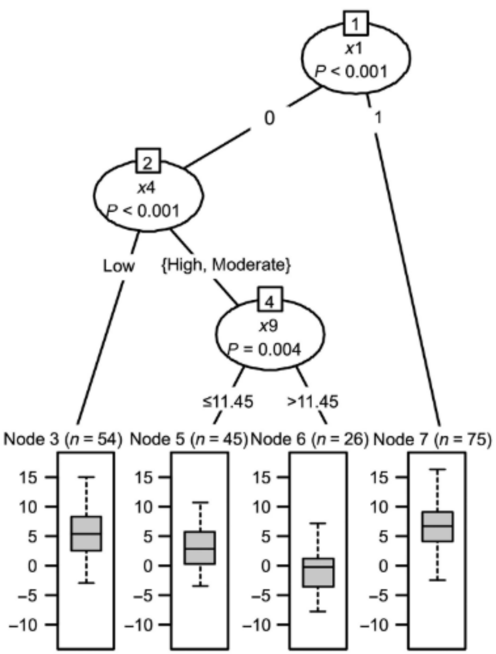
- Hypothesis-testing, theory-driven
- Some strong assumptions
  (e.g. Linearity, normality, additivity)
- Probability

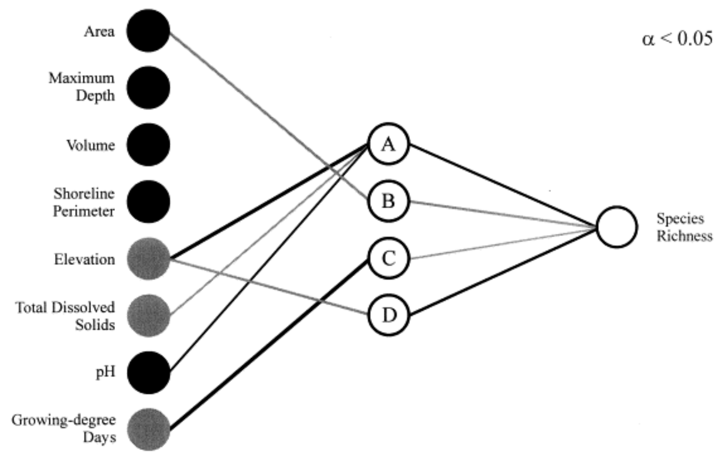


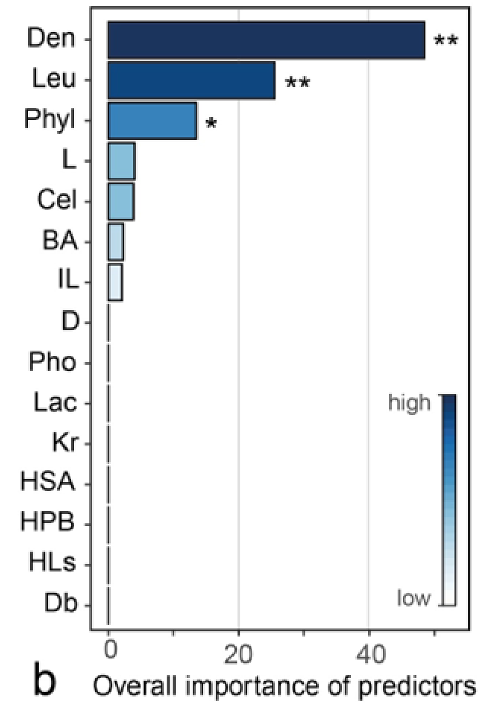
Machine learning

- Information-searching, data-driven
- No assumptions (nonparametric)
- Predictability

#1 Statistically reinforced machine learning?





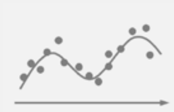


- Permuting **X**, building a model, evaluating the reduction in accuracy
- After repeating this, evaluate if the reduction is significant or not

cf. Ryo & Rillig (2017)

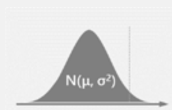
#1 Statistically reinforced machine learning?

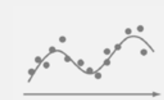
High predictability & model-free hypothesis test

 **Prediction** with 

p-value Variable selection

Using only useful info. increases model performance



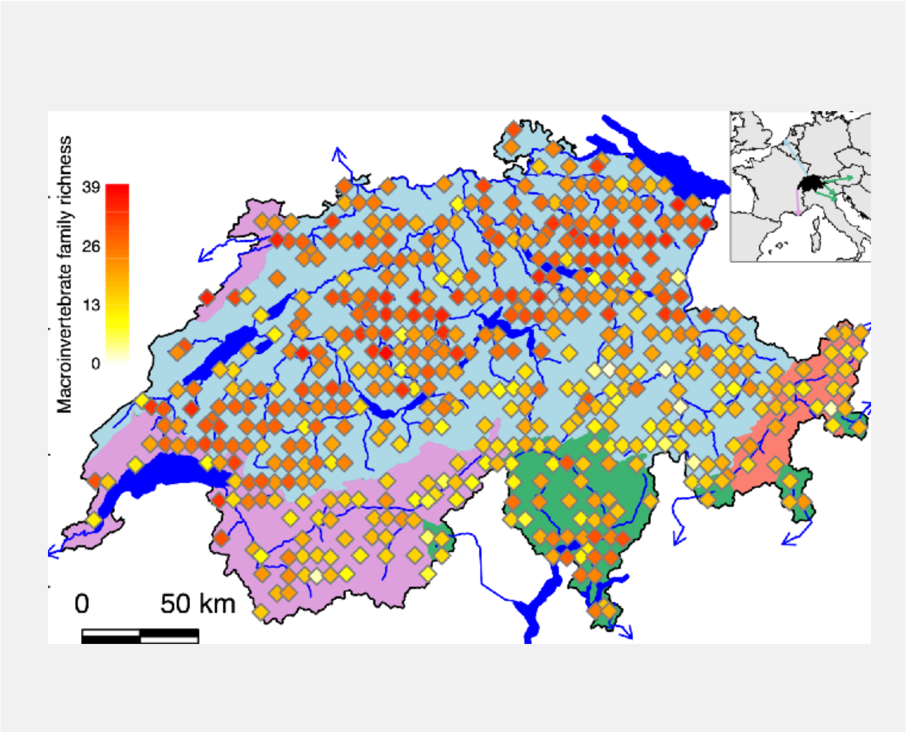
Hypothesis-testing with  Information-searching

Discovering nonlinearity & interactive effect without *a priori* assumption

cf. Ryo et al. (2018)

#2 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)

#2 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)

#2 Nonlinear interactions explains diversity pattern

Variable selection

↓

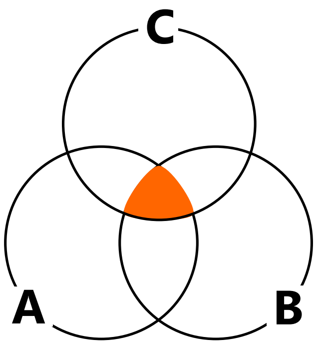
Testing all 3-way combinations

↓

Finding imporant interactions

Mutual Information Theory

cf. Kelly & Okada (2012)



Interaction importance

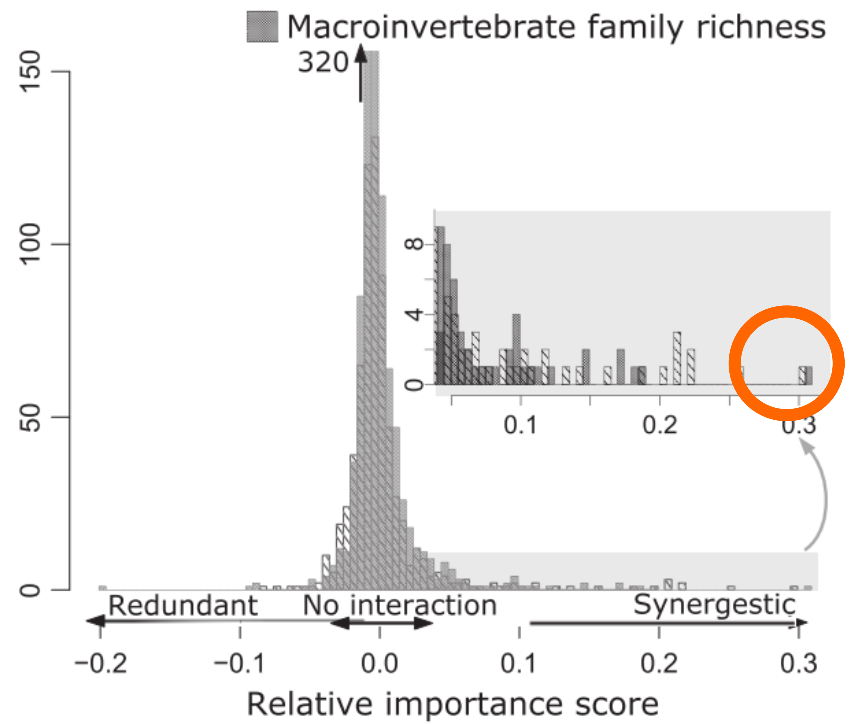
$\boldsymbol{I(A \cap B \cap C)}$

#2 Nonlinear interactions explains diversity pattern

Variable selection

Testing all 3-way combinations

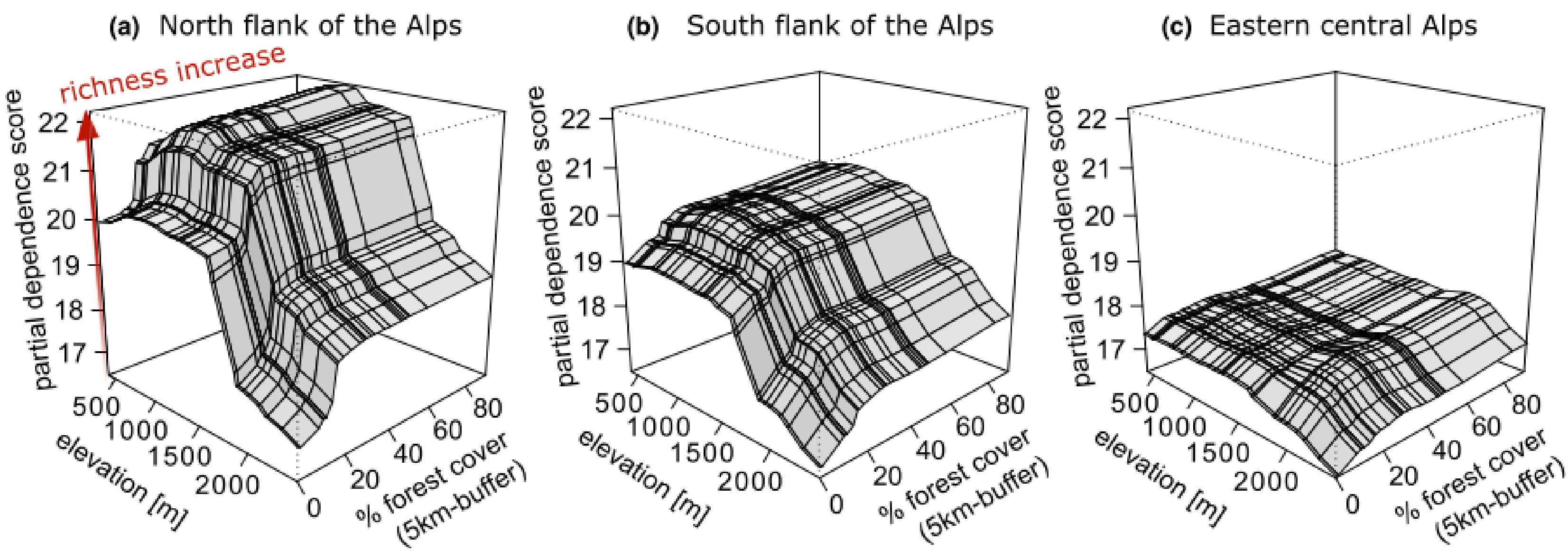
Finding imporant interactions



cf. Ryo et al. (2018)

#2 Nonlinear interactions explains diversity pattern

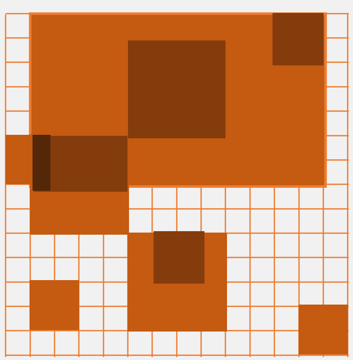
Elevation × Forest coverage × Geographic region



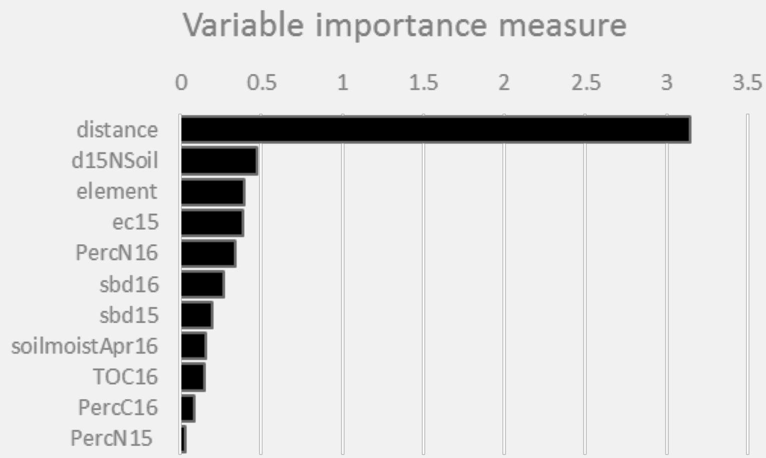
cf. Ryo et al. (in review)

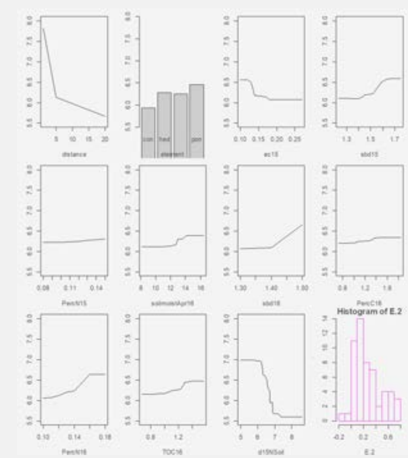
#3 Multiscale spatial autocorrelation

? Spatial autocorrelation in machine learning?



Distribution

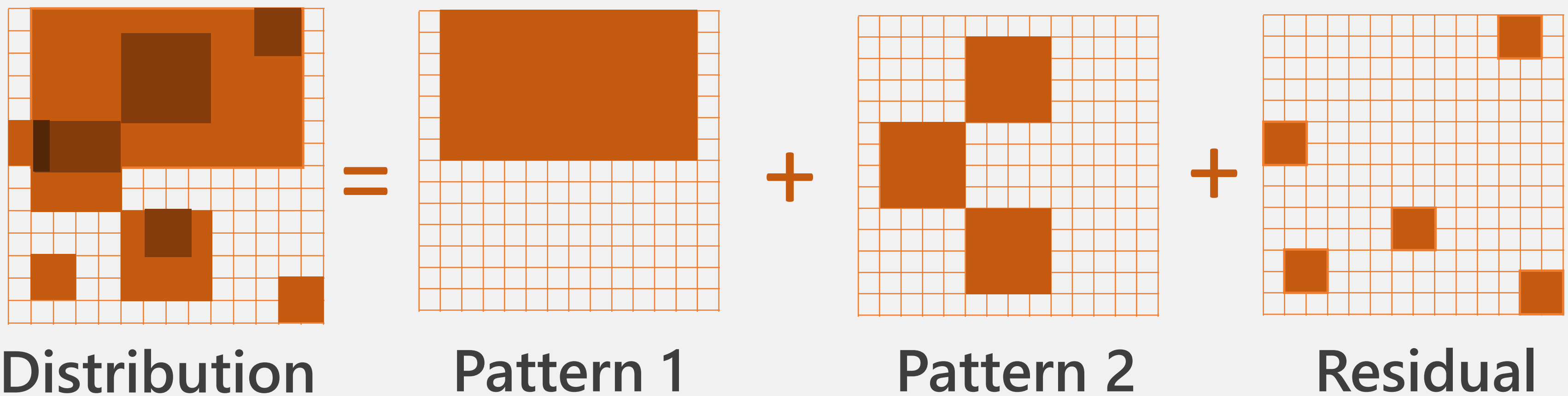




cf. Ryo et al. (in review)

#3 Multiscale spatial autocorrelation

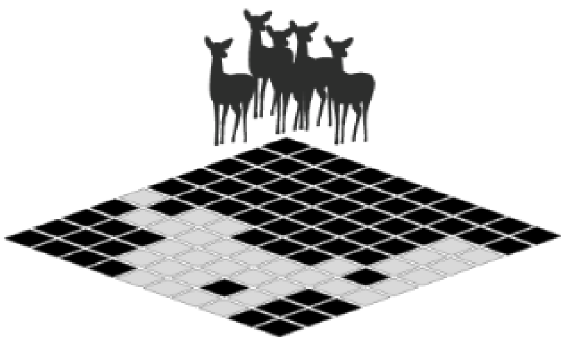
? Spatial autocorrelation in machine learning?

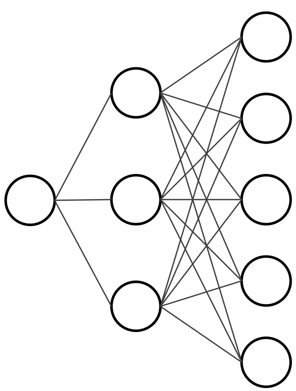


cf. Ryo et al. (in review)

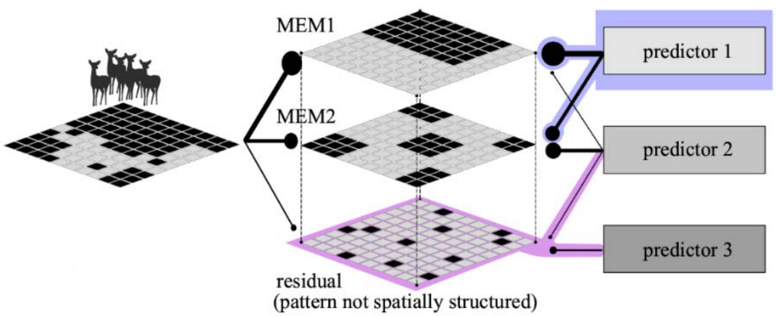
#3 Multiscale spatial autocorrelation

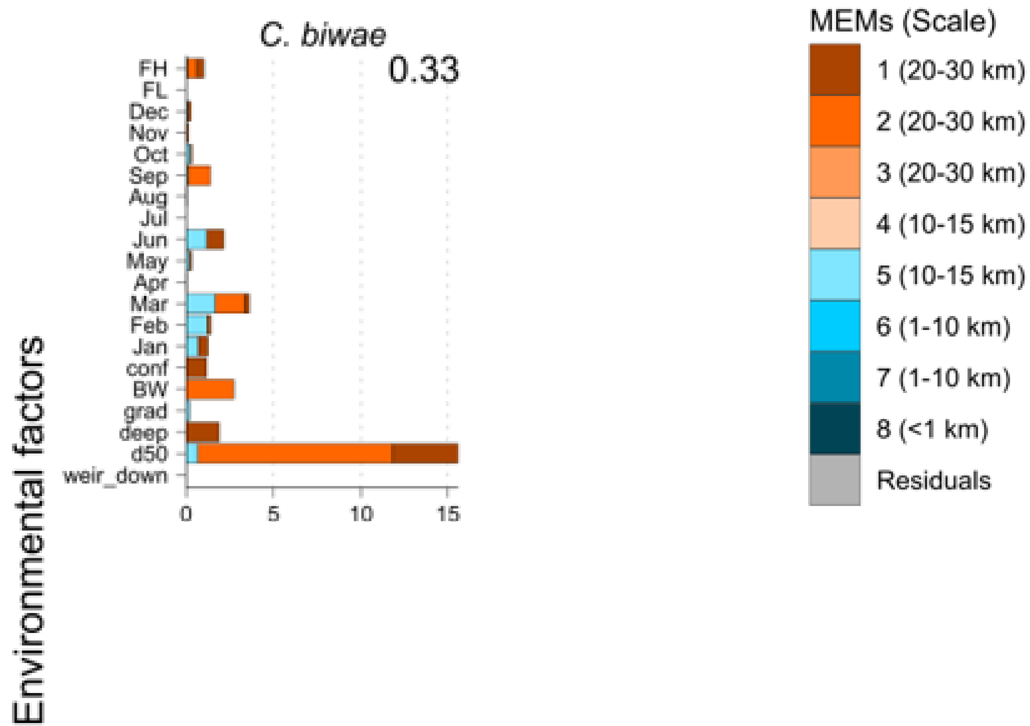
Decomposition to patterns and then regress them ☺





#3 Multiscale spatial autocorrelation





Take-home messages

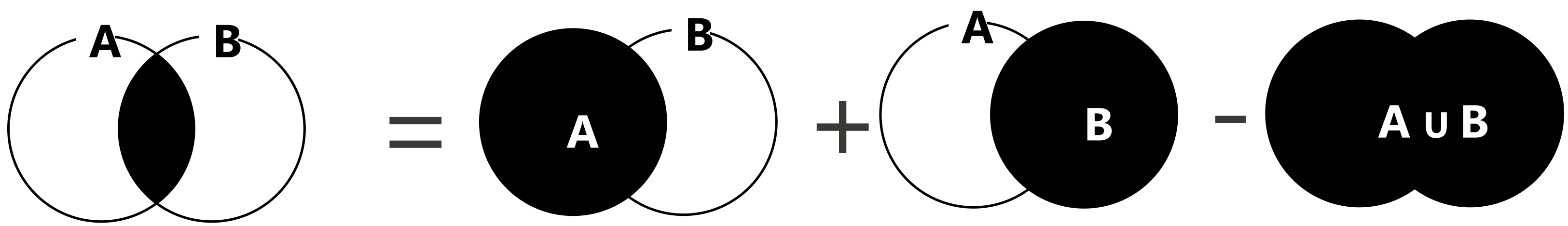
ML can better support ecological studies by offering:

1. Statistical summary for more flexible hypothesis-testing
2. Nonlinear variable interactions discovery
3. Multiscale variable importance with hierarchical structure

- Consultations
- Collaborations
- ML workshops

Masahiro Ryo
https://masahiroryo.jimdo.com
masahiroryo@gmail.com

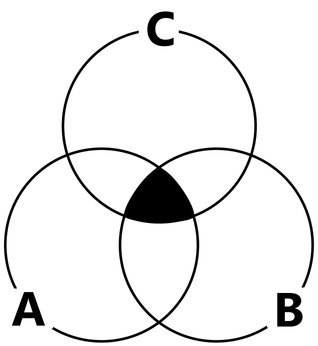
Mutual information theory



Interaction importance **I(A∩B)** = Importance **I(A)** + Importance **I(B)** - Joint importance **I(A∪B)**

Kelly & Okada (2012) Variable interaction measures with random forest classifiers

Mutual information theory



Interaction importance
I(A∩B ∩C)