Using supervised machine learning algorithms for the building of predictive biomonitoring models.

9th Metabarcoding school

Villa de Leyva – 1-6th November 2019

Tristan Cordier

UNIVERSITÉ DE GENÈVE

FNSNF Swiss National Science Foundation

EFFECTS OF BIODIVERSITY ON ECOSYSTEM FUNCTIONING:
A CONSENSUS OF CURRENT KNOWLEDGE

D. U. Hooper,[1,16] F. S. Chapin, III,[2] J. J. Ewel,[3] A. Hector,[4] P. Inchausti,[5] S. Lavorel,[6] J. H. Lawton,[7]
D. M. Lodge,[8] M. Loreau,[9] S. Naeem,[10] B. Schmid,[4] H. Setälä,[11] A. J. Symstad,[12]
J. Vandermeer,[13] and D. A. Wardle[14,15]

Biodiversity and ecosystem stability in a decade-long grassland experiment

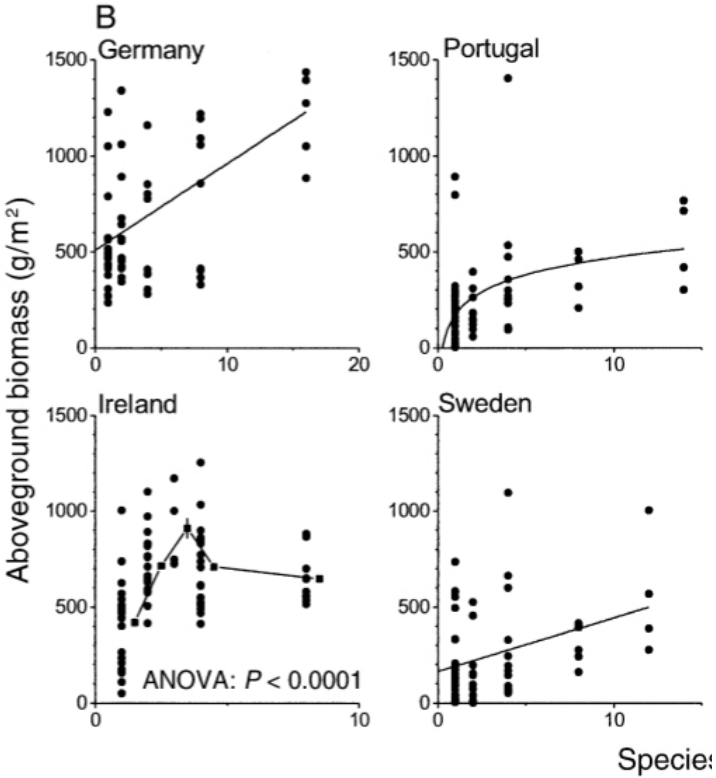
David Tilman[1], Peter B. Reich[2] & Johannes M. H. Knops[3]

REVIEW

doi:10.1038/nature11148

Biodiversity loss and its impact on humanity

Bradley J. Cardinale[1], J. Emmett Duffy[2], Andrew Gonzalez[3], David U. Hooper[4], Charles Perrings[5], Patrick Venail[1], Anita Narwani[1], Georgina M. Mace[6], David Tilman[7], David A. Wardle[8], Ann P. Kinzig[5], Gretchen C. Daily[9], Michel Loreau[10], James B. Grace[11], Anne Larigauderie[12], Diane S. Srivastava[13] & Shahid Naeem[14]

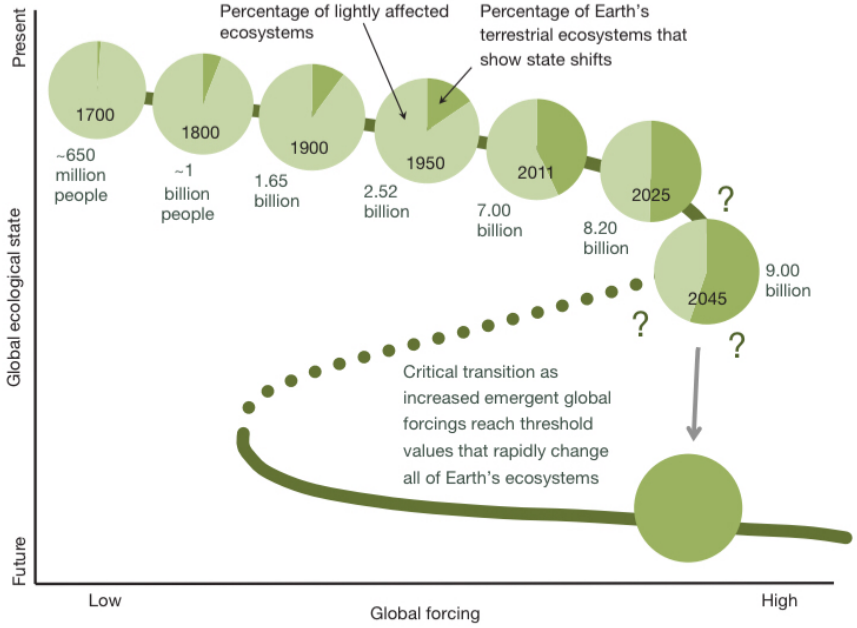


REVIEW

doi:10.1038/nature11018

Approaching a state shift in Earth's biosphere

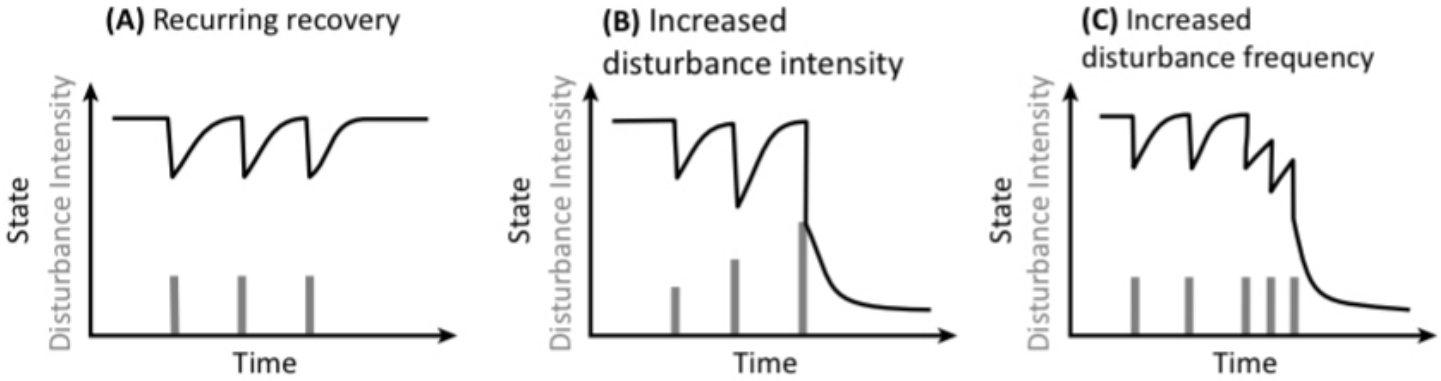
Anthony D. Barnosky[1,2,3], Elizabeth A. Hadly[4], Jordi Bascompte[5], Eric L. Berlow[6], James H. Brown[7], Mikael Fortelius[8], Wayne M. Getz[9], John Harte[9,10], Alan Hastings[11], Pablo A. Marquet[12,13,14,15], Neo D. Martinez[16], Arne Mooers[17], Peter Roopnarine[18], Geerat Vermeij[19], John W. Williams[20], Rosemary Gillespie[9], Justin Kitzes[9], Charles Marshall[1,2], Nicholas Matzke[1], David P. Mindell[21], Eloy Revilla[22] & Adam B. Smith[23]



Review

Abrupt Change in Ecological Systems: Inference and Diagnosis

Zak Ratajczak,[1,*] Stephen R. Carpenter,[2] Anthony R. Ives,[1] Christopher J. Kucharik,[3] Tanjona Ramiadantsoa,[1] M. Allison Stegner,[1] John W. Williams,[4] Jien Zhang,[1] and Monica G. Turner[1,*]



Vol 448|12 July 2007|doi:10.1038/nature05947

LETTERS

Biodiversity and ecosystem multifunctionality

Andy Hector[1] & Robert Bagchi[1,]

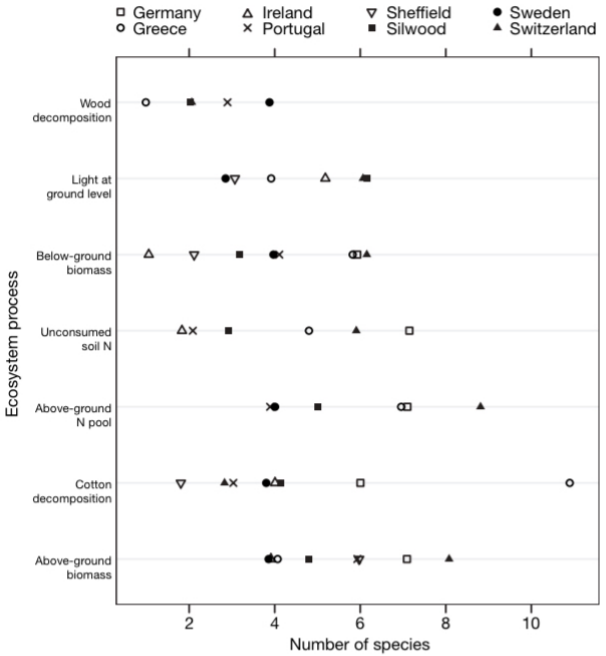


Figure 1 | Number of species with desirable effects on the suite of ecosystem processes measured in the different BIODEPTH project experiments. The number of species was identified by the AIC-based multiple regression (and species with effects with undesirable signs were then excluded).

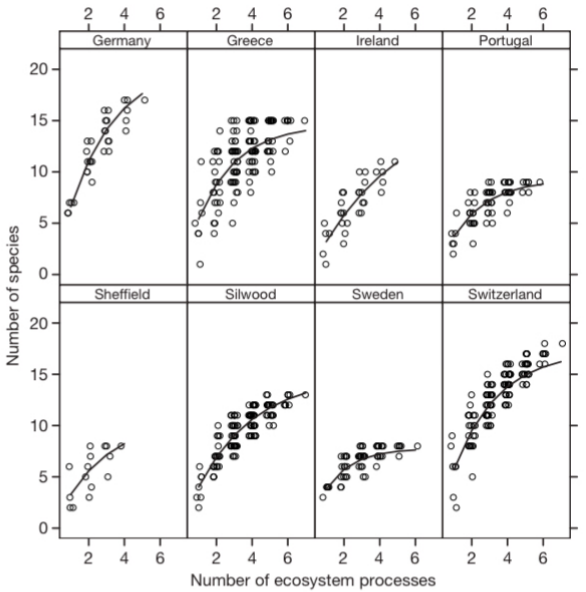
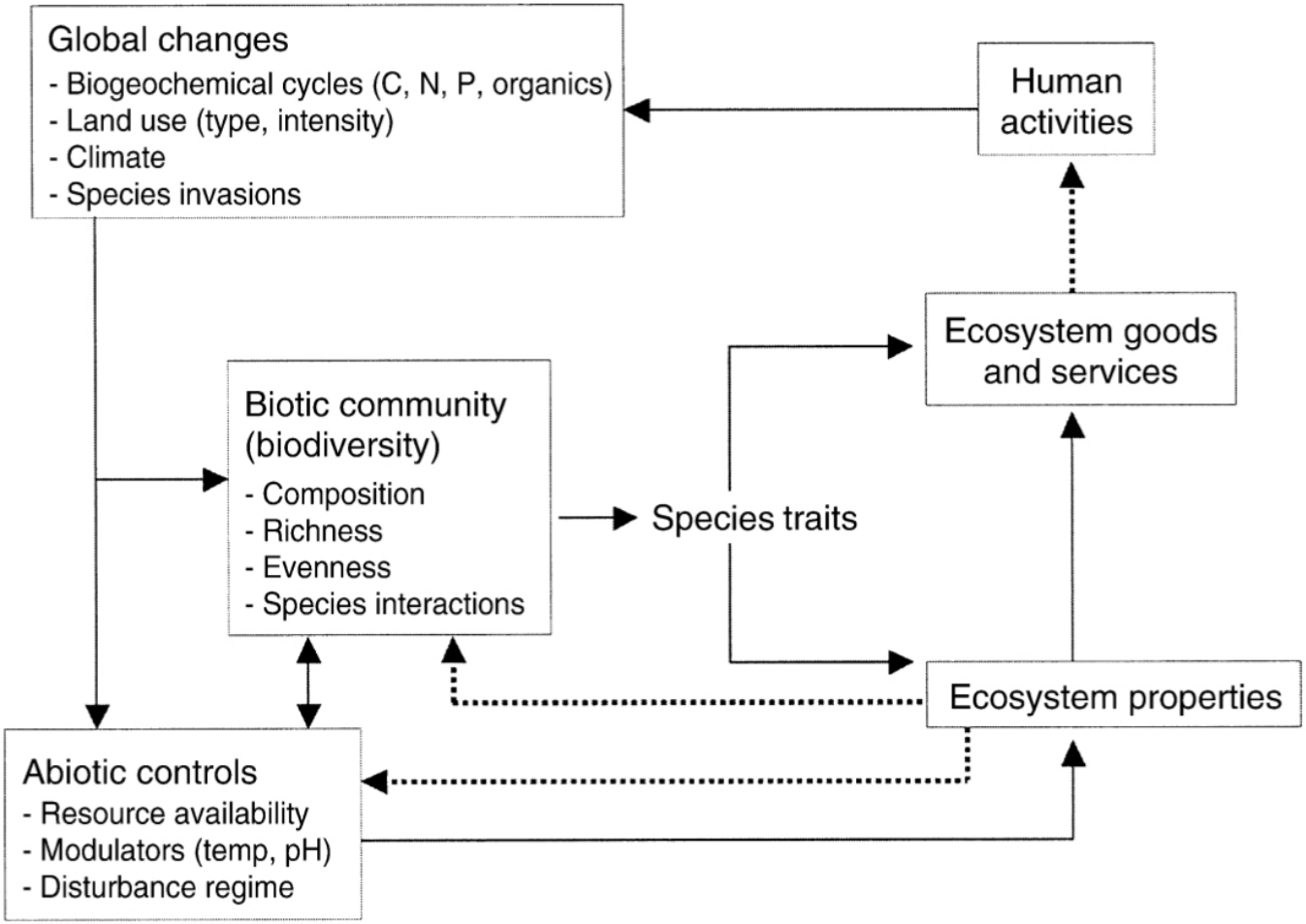


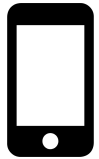
Figure 2 | Positive relationship between the range of ecosystem processes considered and the number of species that affect one or more aspect of ecosystem functioning. The points (jittered for clarity) show numbers of species required for all possible combinations of ecosystem processes. Lines are theoretical predictions from the model based on the average number of species required for a single process, $\bar{x}$, and the average overlap in the sets of species required for each pair of processes, $\bar{o}$, using equation (2).

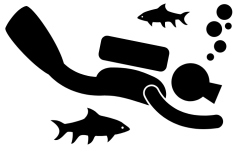


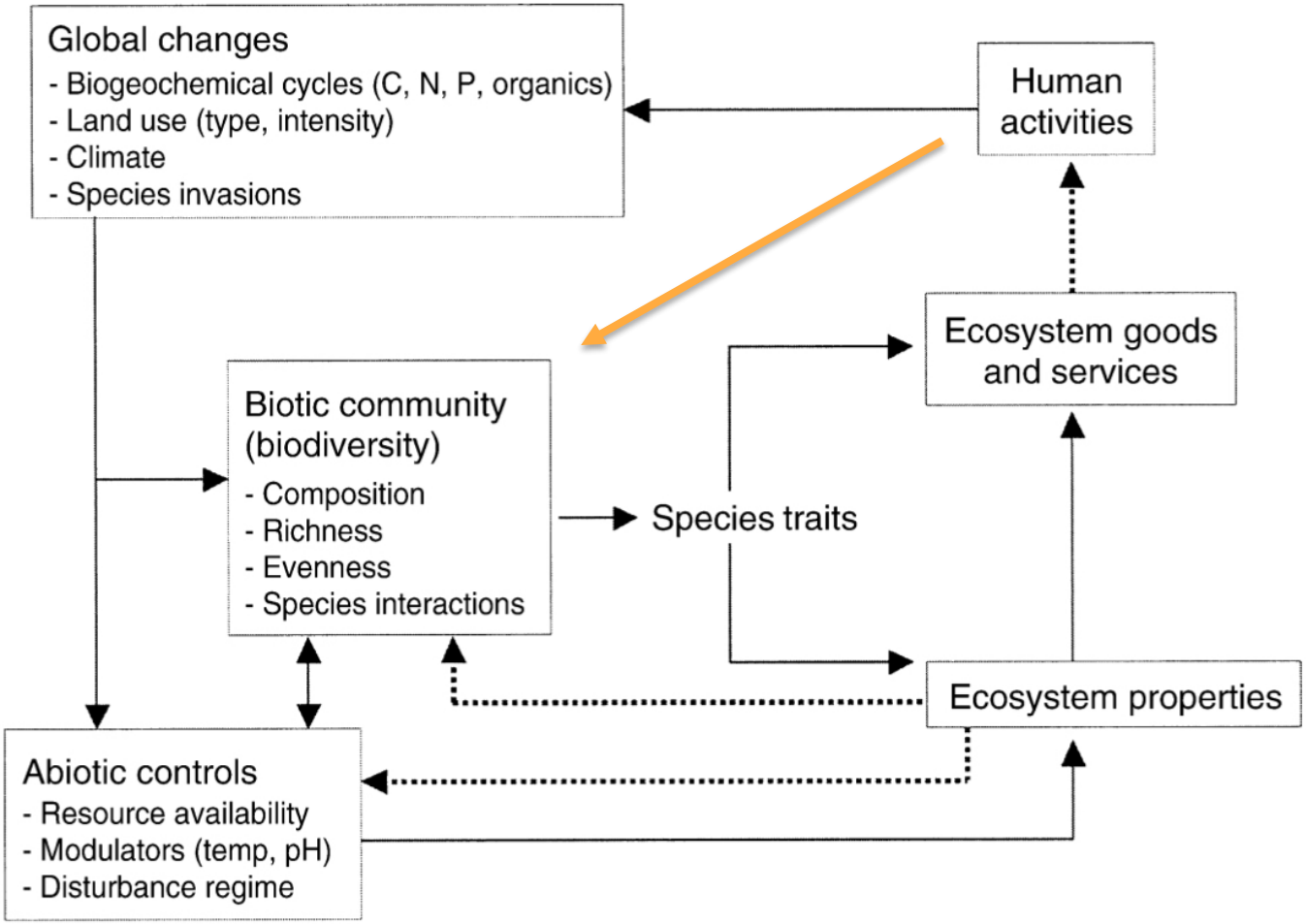
Hooper et al. 2005,

(Marine) ecosystems services

- Sustainable development !







Hooper et al. 2005,

WHAT CONSTITUTES ECOSYSTEM HEALTH?

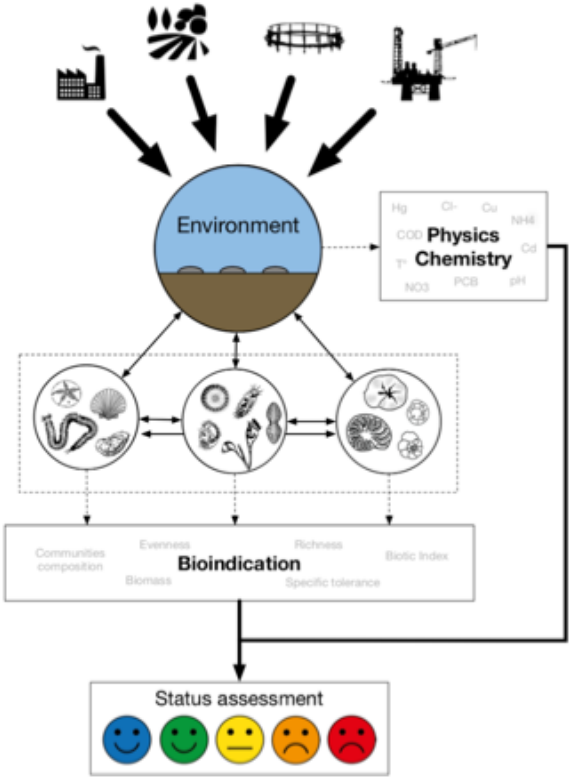
*DAVID J. RAPPORT** *Perspectives in Biology and Medicine, 33, 1 · Autumn 1989*

The Use of Metaphor in Science

In addressing the question of what constitutes ecosystem health, it is tempting to make use of metaphors drawn from human medicine. The use of metaphor, it has been argued, has not only a role in poetry, but also a legitimate place in science [1]. Its function in both areas is to stimulate associations, bringing into juxtaposition phenomena that might at first appear to have little connection but can be seen to be in some way related. In poetry the association might be quite fanciful, while in science the value of metaphor lies in pointing to phenomena in apparently different spheres that bear some structural identity.

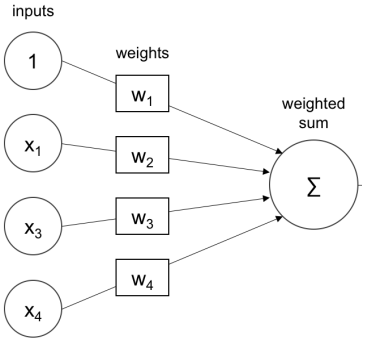
What constitutes the health of Nature—that is, what are the suitable concepts and conventions to assess the condition of environment—is a question now being raised for the earth's major ecosystems (forests, lakes, seas, etc.) and indeed the entire biosphere. Ultimately, a healthy environment is essential for a healthy human population. It does not follow, however, that the appropriate standards for the health of nature need be based solely on criteria for human health. Ecosystems have a life of their own with or without human components, and it is this life that is receiving ever more attention as situations come to light in which ecosystems have become severely damaged [2].

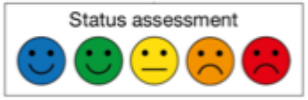
Impact assessment relies on biotic indices

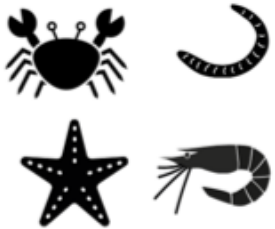


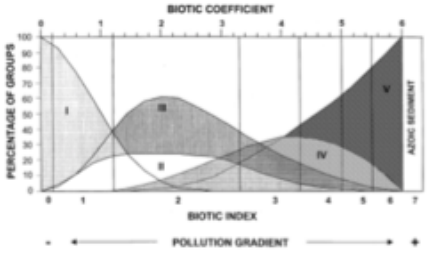
AMBI index groups (Borja 2000)

I – very sensitive species
II – indifferent species
III – tolerant species
IV – second order opportunistic species
V – first order opportunistic species









Bioindicators convey a cumulative measure of disturbances

Impact assessment relies on biotic indices

Ecological Applications, 9(2), 1999, pp. 699–713

THE INFAUNAL TROPHIC INDEX (ITI): ITS SUITABILITY FOR MARINE ENVIRONMENTAL MONITORING

DON MAURER,[1] HAI NGUYEN,[2] GEORGE ROBERTSON,[2] AND TOM GERLINGER[2]

[1]*Department of Biological Sciences, California State University, 1250 Bellflower Boulevard, Long Beach, California 90840 USA*
[2]*County Sanitation Districts of Orange County California, 10844 Ellis Avenue, Fountain Valley, California 92708-7018 USA*

Contents lists available at ScienceDirect

Marine Pollution Bulletin

journal homepage: www.elsevier.com/locate/marpolbul

A bacterial community-based index to assess the ecological status of estuarine and coastal environments

Eva Aylagas [a,*], Ángel Borja [a,*], Michael Tangherlini [b], Antonio Dell'Anno [b], Cinzia Corinaldesi [b], Craig T. Michell [c], Xabier Irigoien [a,c,d], Roberto Danovaro [b], Naiara Rodríguez-Ezpeleta [a]

[a] *AZTI - Marine Research, Herrera Kaia, Portualdea z/g – 20110 Pasaia, Gipuzkoa, Spain*
[b] *Department of Life and Environmental Sciences, Polytechnic University of Marche, Via Brecce Bianche, 60131 Ancona, Italy*
[c] *Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*
[d] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*

Biotic indices based on soil nematode communities for assessing soil quality in terrestrial ecosystems

Arantzazu Urzelai[a], Ana Jesús Hernández[b], Jesús Pastor[a,*]

[a] *Centre of Environmental Sciences, CSIC, Serrano 115 dpdo., E-28006 Madrid, Spain*
[b] *Area of Ecology, University of Alcalá de Henares, Ctra. Madrid-Barcelona km 38, E-28771 Alcalá de Henares, Spain*

Pergamon

Marine Pollution Bulletin Vol. 40, No. 12, pp. 1100–1114, 2000

Printed in Great Britain
PII: S0025-326X(00)00061-8
0025-326X/00 $ - see front matter

A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments

A. BORJA*, J. FRANCO and V. PÉREZ

Department of Oceanography and Marine Environment, Technological Institute for Fisheries and Food (AZTI), Av. Satrústegui 8, 20008 San Sebastián, Spain

Impact assessment relies on biotic indices

Ecological Applications, 9(2), 1999, pp. 699–713

THE INFAUNAL

DON MA
[1]Department
[2]Coun

Contents lists available at ScienceDirect

Ecological Indicators 18 (2012) 31–41

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

l status of

ızia Corinaldesi [b], Craig T. Michell [c],

Pollution Bulletin Vol. 40, No. 12, pp. 1100–1114, 2000

Printed in Great Britain
0025-326X/00 $ - see front matter

Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive

Sebastian Birk [a,*], Wendy Bonne [b], Angel Borja [c], Sandra Brucet [b], Anne Courrat [d], Sandra Poikane [b], Angelo Solimini [e], Wouter van de Bund [b], Nikolaos Zampoukas [b], Daniel Hering [a]

[a] *University of Duisburg-Essen, Faculty of Biology, Applied Zoology/Hydrobiology, Universitätsstraße 5, 45141 Essen, Germany*
[b] *European Commission, Joint Research Centre, Institute for Environment and Sustainability, 21027 Ispra, Italy*
[c] *AZTI-Tecnalia, Marine Research Division, Herrera Kaia, Portualdea s/n, 20110 Pasaia, Spain*
[d] *CEMAGREF - UR EPBX, 50 avenue de Verdun, 33612 Cestas, France*
[e] *Department of Public Health and Infectious Diseases, Sapienza University of Rome, Roma, Italy*

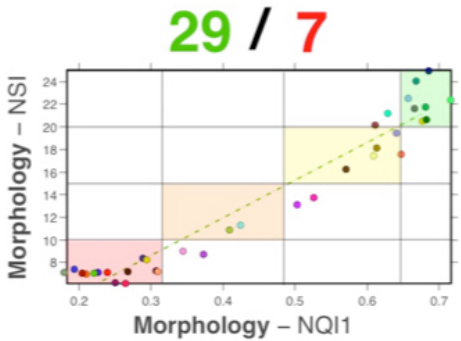
Biotic indice
assessin

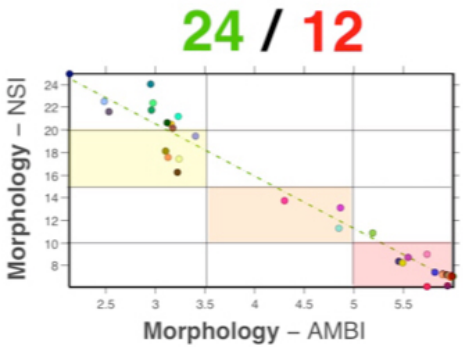
Arantzazu

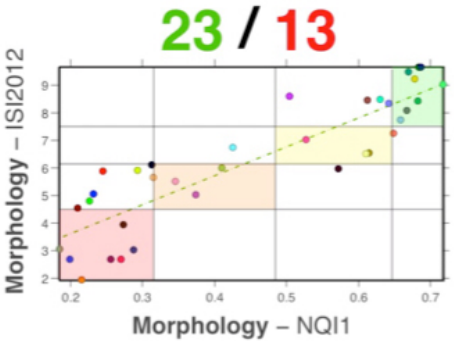
Establish the
-Bottom
Estuarine and

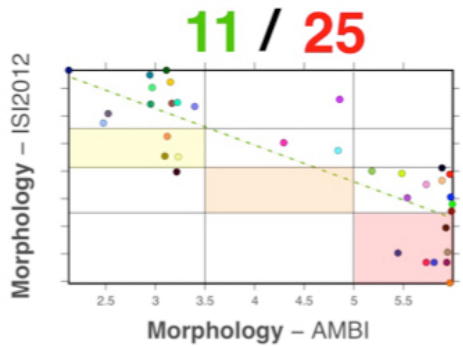
[a] *Centre of Environmental Sciences, CSIC, Serrano 115 dpdo., E-28006 Madrid, Spain*
[b] *Area of Ecology, University of Alcalá de Henares, Ctra. Madrid-Barcelona km 38, E-28771 Alcalá de Henares, Spain*

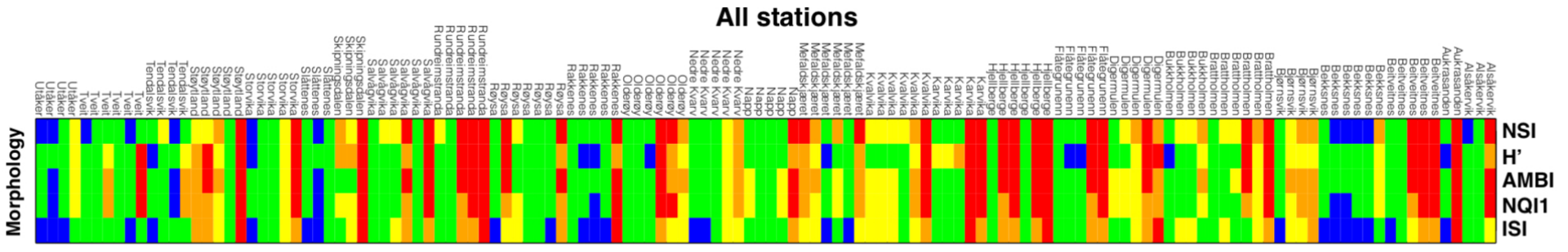
A. BORJA*, J. FRANCO and V. PÉREZ
Department of Oceanography and Marine Environment, Technological Institute for Fisheries and Food (AZTI), Av. Satrústegui 8, 20008 San Sebastián, Spain

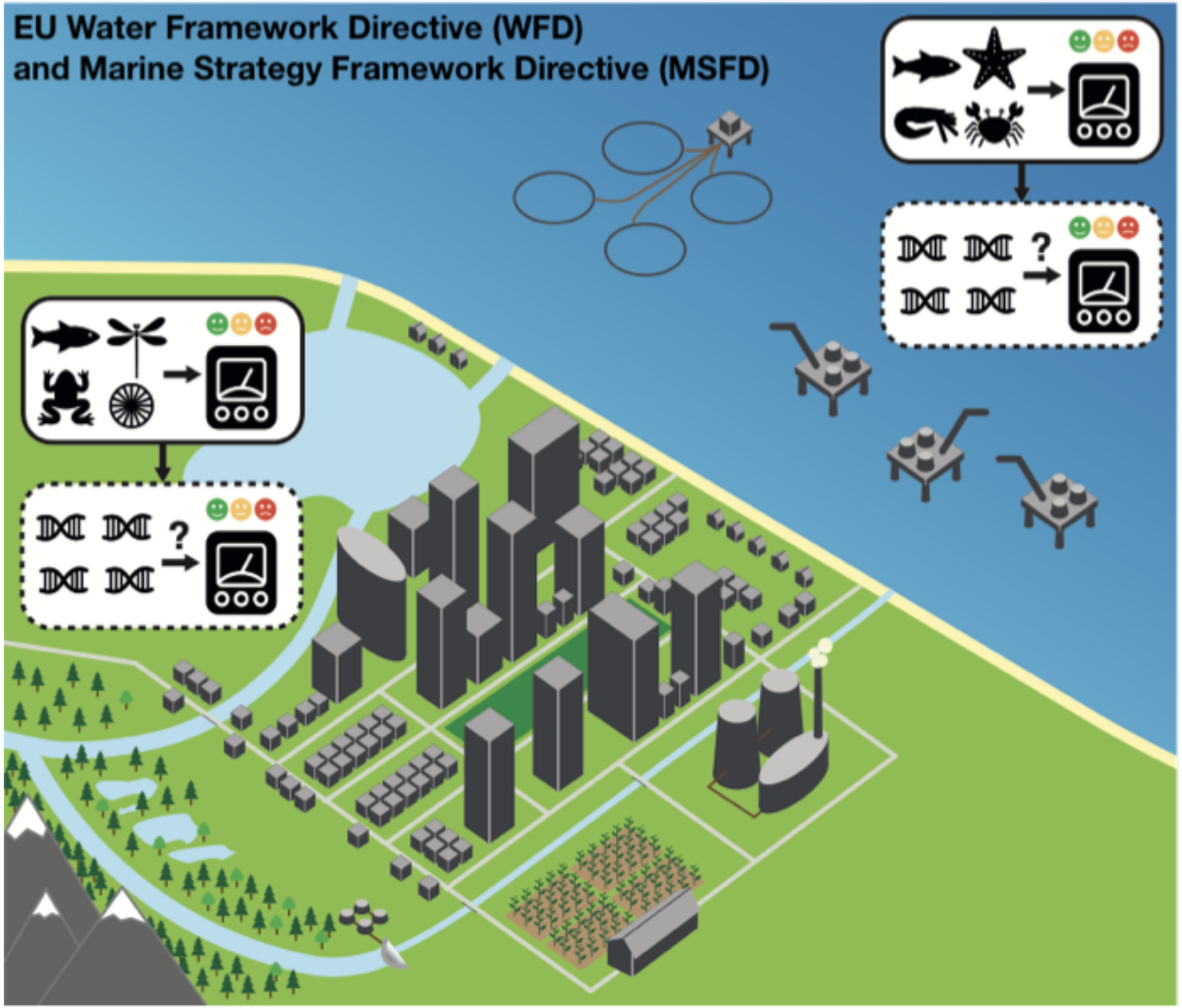

29 / 7
Morphology – NSI
Morphology – NQI1


24 / 12
Morphology – NSI
Morphology – AMBI


23 / 13
Morphology – ISI2012
Morphology – NQI1


11 / 25
Morphology – ISI2012
Morphology – AMBI


All stations
Morphology
NSI
H'
AMBI
NQI1
ISI



- Time consuming
- Only morphosp.
- Taxonomic skills

- Faster
- Full community
- Standardized
- Automatized

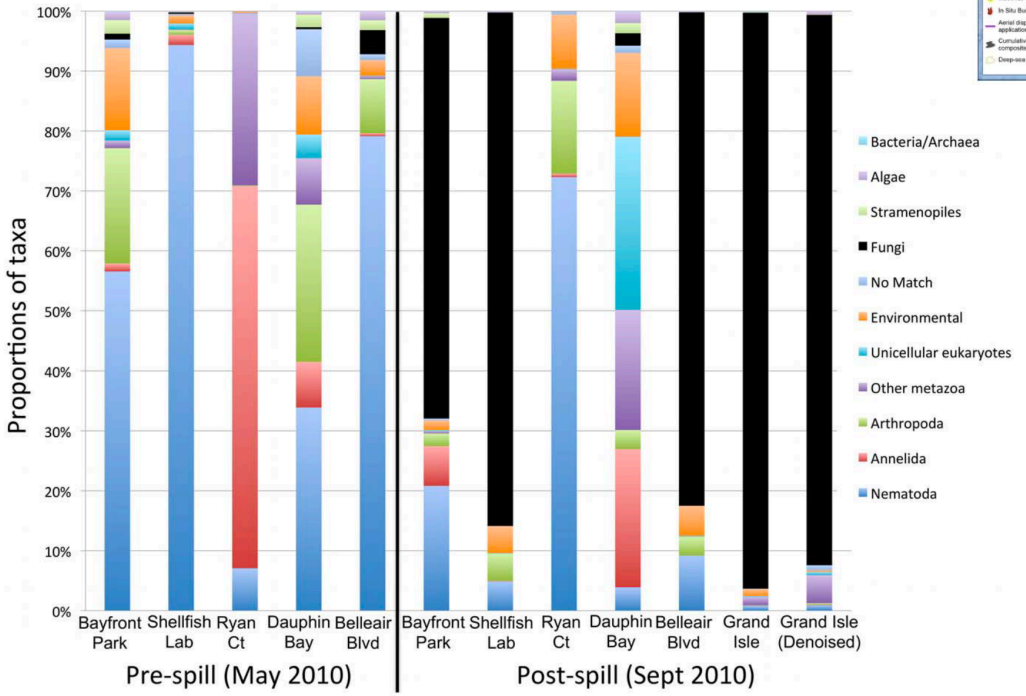
Cordier et al., 2019

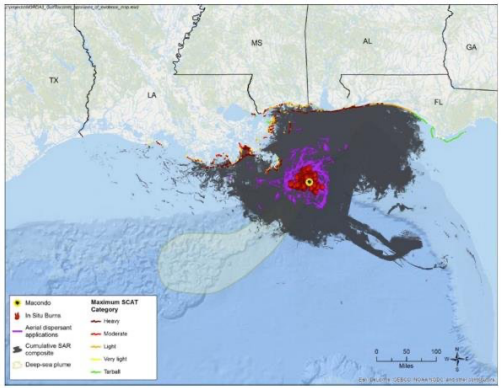
OPEN ACCESS Freely available online

PLoS one

Dramatic Shifts in Benthic Microbial Eukaryote Communities following the Deepwater Horizon Oil Spill

Holly M. Bik[1,2]*, Kenneth M. Halanych[3], Jyotsna Sharma[4], W. Kelley Thomas[1]





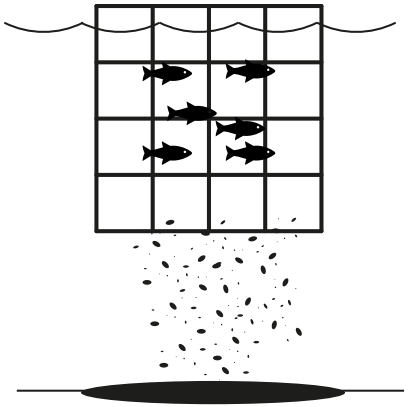
SCIENTIFIC REPORTS

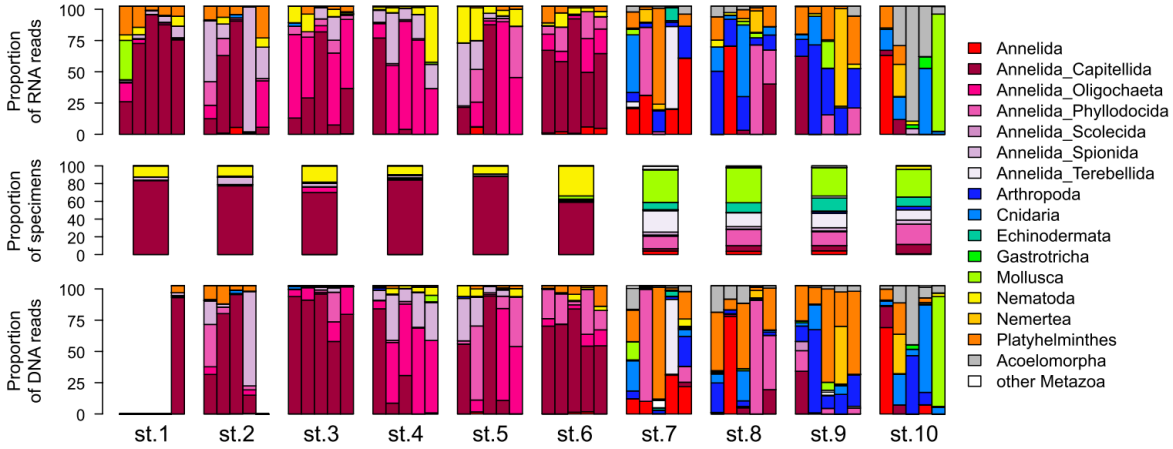
OPEN

High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems

Received: 01 April 2015
Accepted: 12 August 2015
Published: 10 September 2015

Franck Lejzerowicz[1,†], Philippe Esling[1,2], Loïc Pillet[1,3], Thomas A. Wilding[4], Kenneth D. Black[4] & Jan Pawlowski[1]

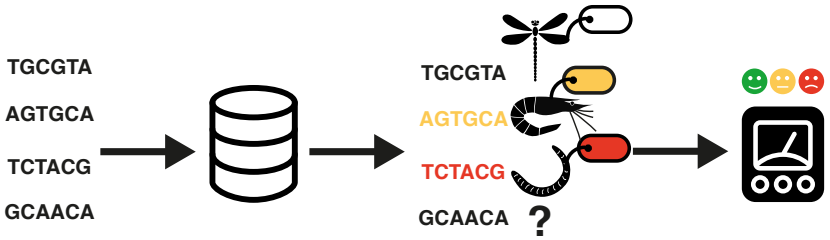


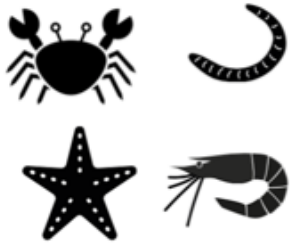


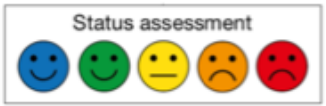
A

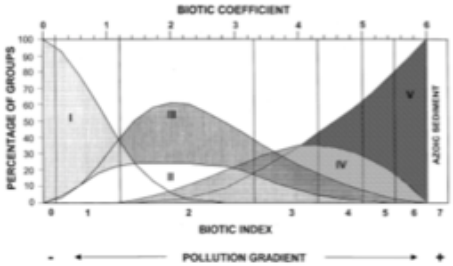
Taxonomy-based approaches

- Screening species / screening bioindicators





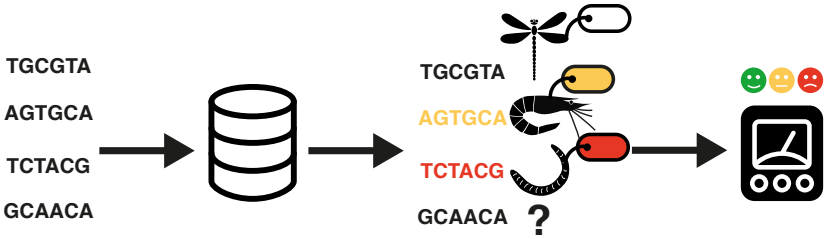


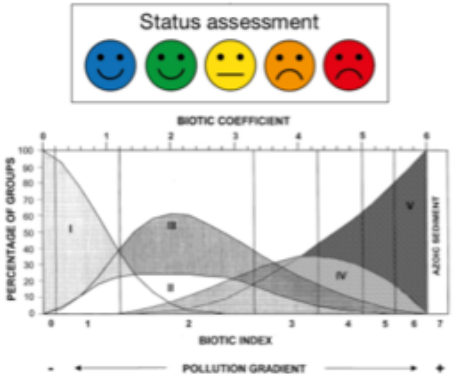


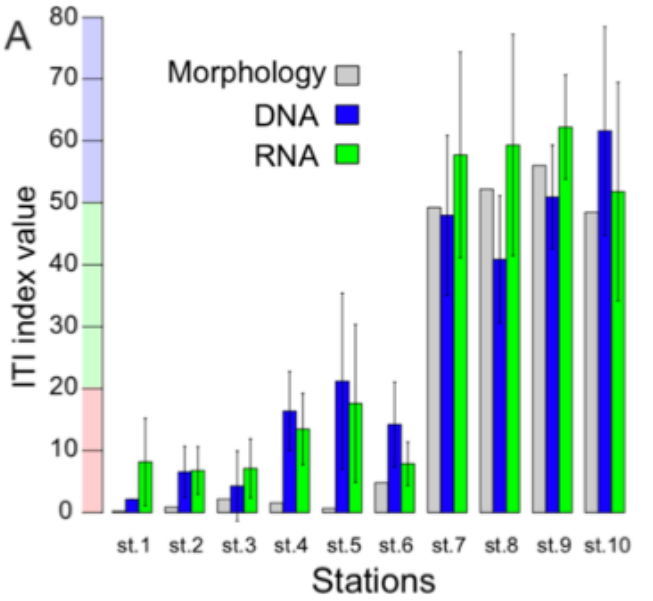
A

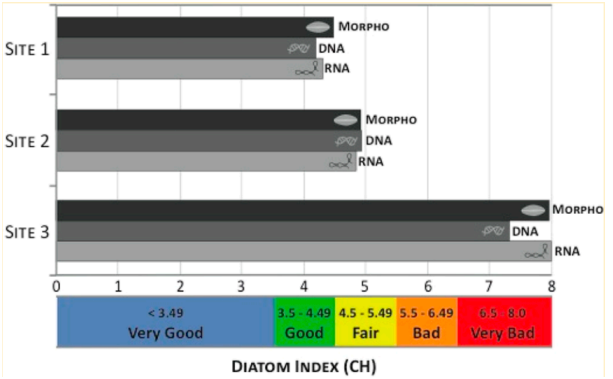
Taxonomy-based approaches

- Screening species / screening bioindicators

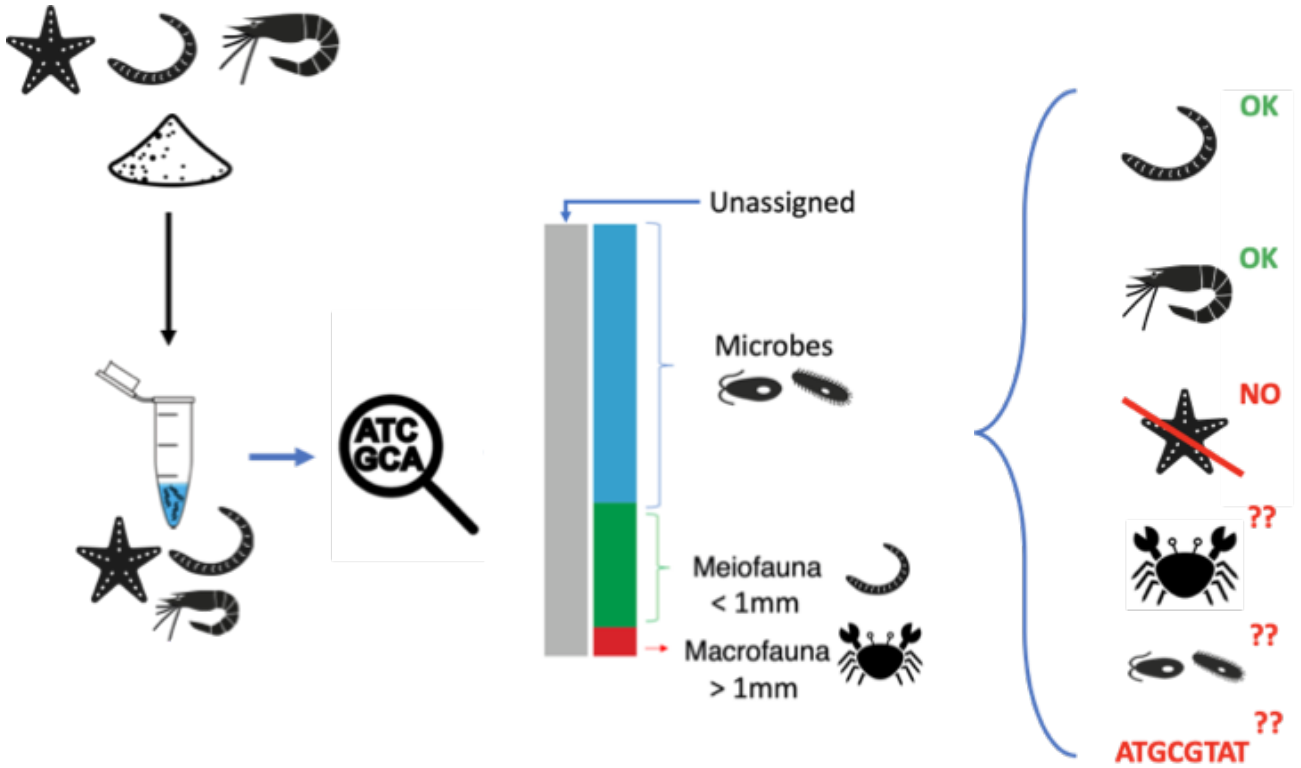


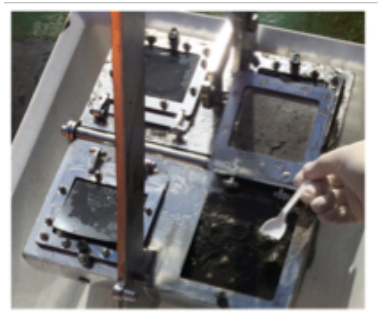




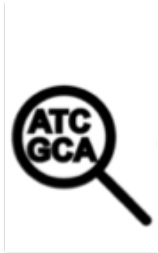


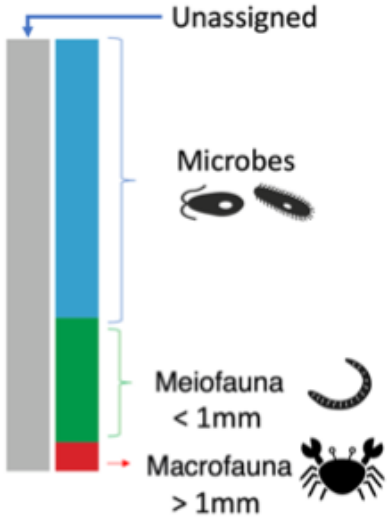
Sampling eDNA : What you actually get

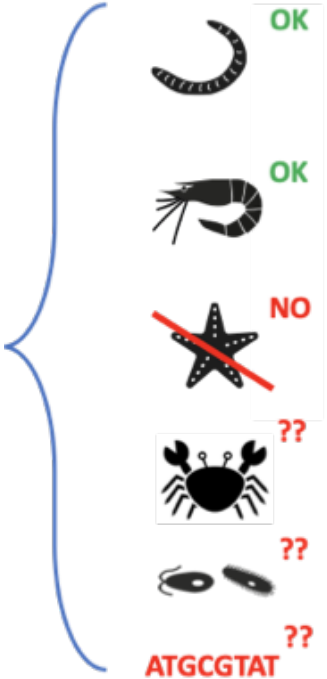




Sampling eDNA : What you actually get







Reference DB? → Filling it!

Quantification? → Correction factor

Quest for the perfect PCR primers..

WHY MICROBES?

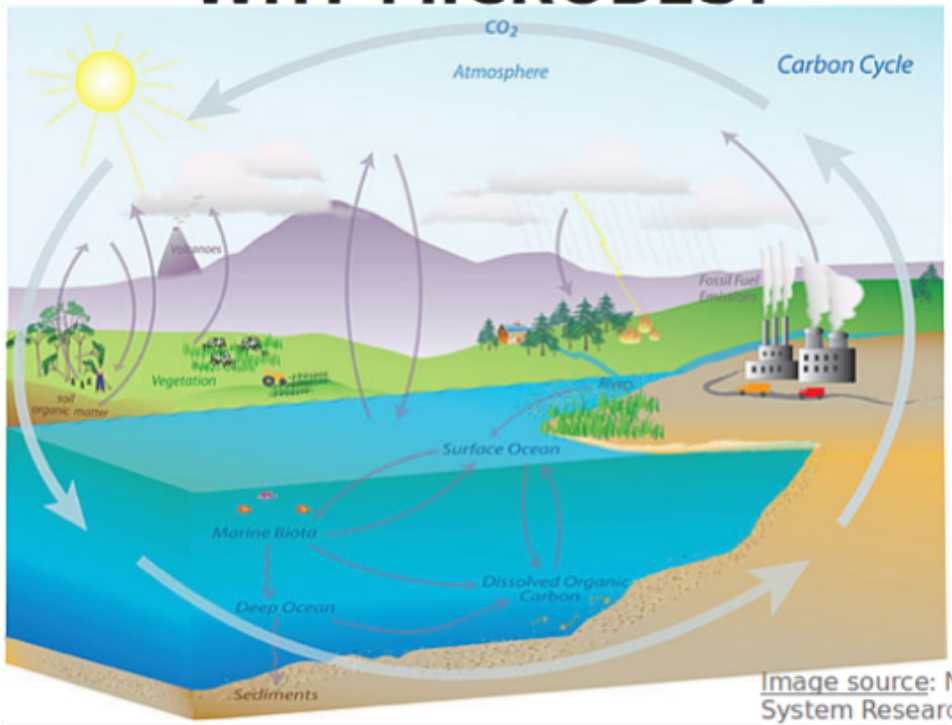


Image source: NOAA Earth System Research Laboratory, Dept. of Commerce

- Microbial life dominates
- Drivers of **global biogeochemical cycles**
- **Sensitive and responsive** to changing conditions

OPEN

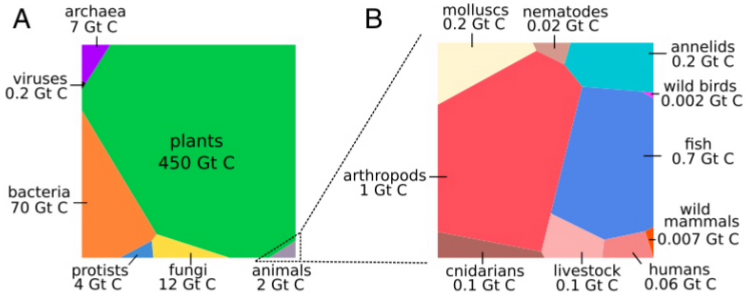
Scientists' warning to humanity: microorganisms and climate change

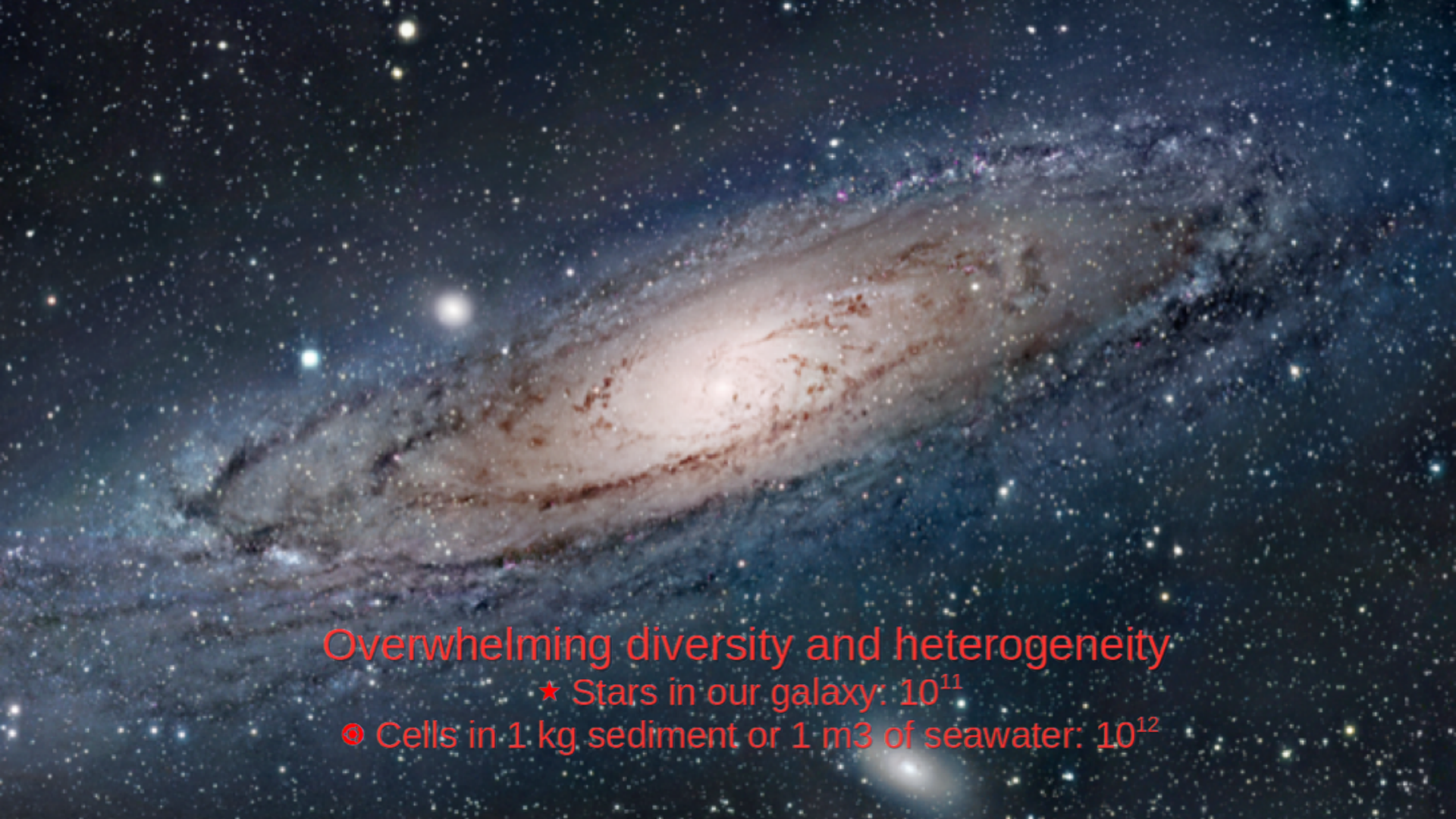
Ricardo Cavicchioli[1], William J. Ripple[2], Kenneth N. Timmis[3], Farooq Azam[4], Lars R. Bakken[5], Matthew Baylis[6], Michael J. Behrenfeld[7], Antje Boetius[8,9], Philip W. Boyd[10], Aimée T. Classen[11], Thomas W. Crowther[12], Roberto Danovaro[13,14], Christine M. Foreman[15], Jef Huisman[16], David A. Hutchins[17], Janet K. Jansson[18], David M. Karl[19], Britt Koskella[20], David B. Mark Welch[21], Jennifer B. H. Martiny[22], Mary Ann Moran[23], Victoria J. Orphan[24], David S. Reay[25], Justin V. Remais[26], Virginia I. Rich[27], Brajesh K. Singh[28], Lisa Y. Stein[29], Frank J. Stewart[30], Matthew B. Sullivan[31], Madeleine J. H. van Oppen[32,33], Scott C. Weaver[34], Eric A. Webb[17] and Nicole S. Webster[33,35]

PNAS

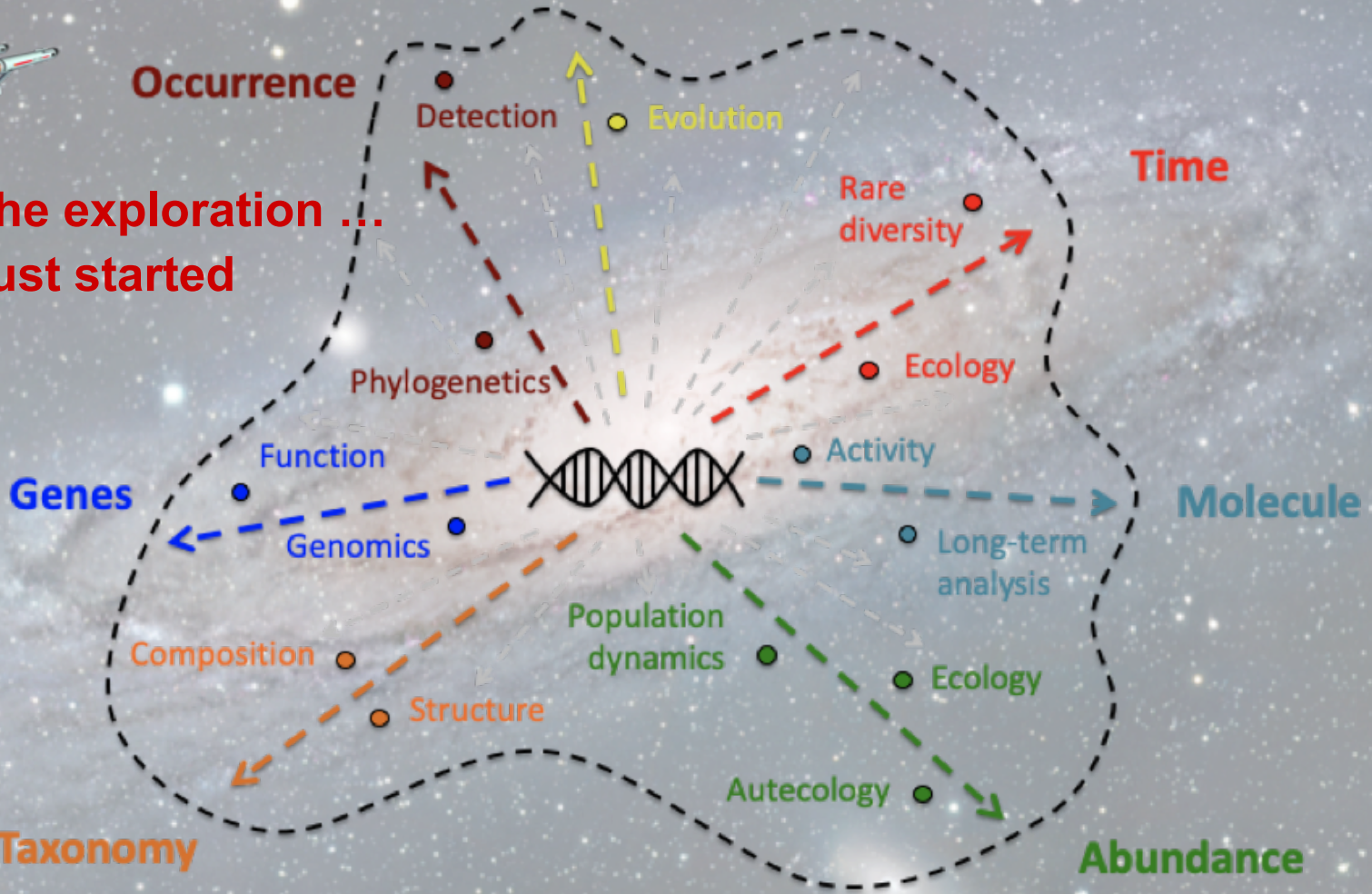
The biomass distribution on Earth

Yinon M. Bar-On[a], Rob Phillips[b,c], and Ron Milo[a,1]




Overwhelming diversity and heterogeneity
★ Stars in our galaxy: 10^{11}
◎ Cells in 1 kg sediment or 1 m3 of seawater: 10^{12}

... and the exploration ... just started

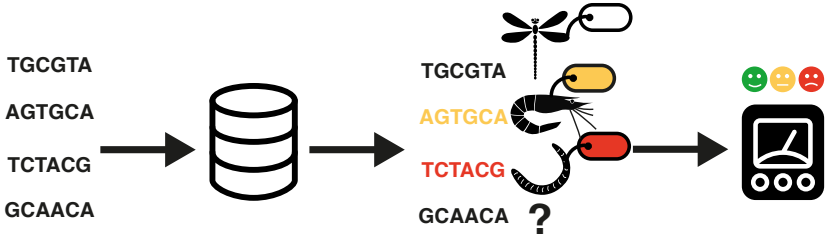


Courtesy Franck Lejzerowicz

A

Taxonomy-based approaches

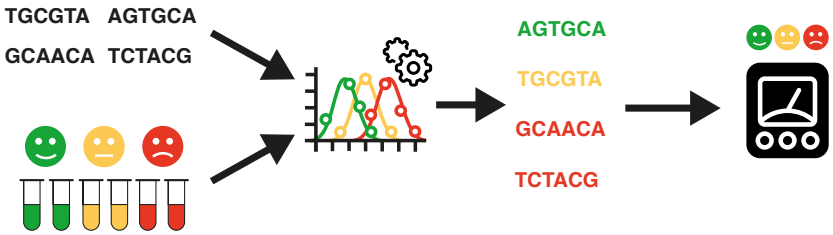
- Screening species / screening bioindicators

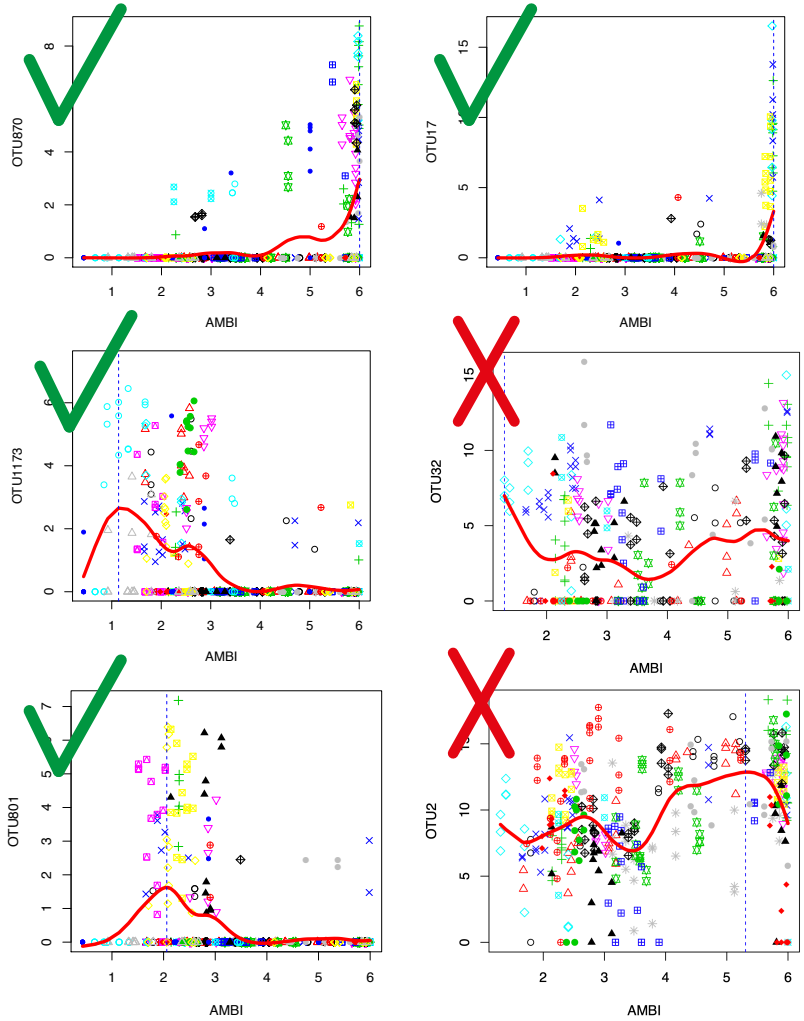


B

De novo approaches

- Biondicators discovery / supervised learning

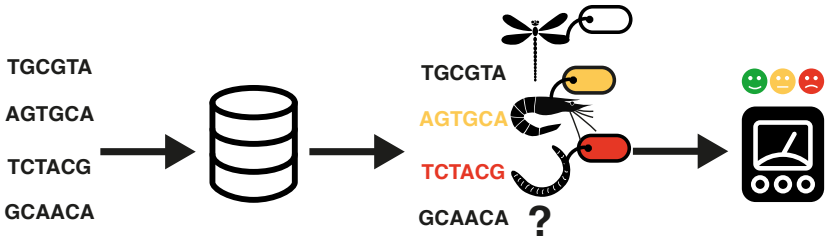




A

Taxonomy-based approaches

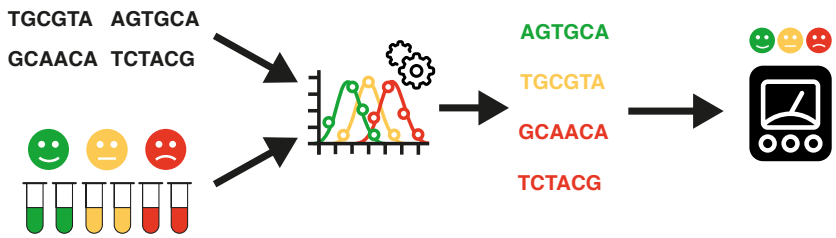
- Screening species / screening bioindicators

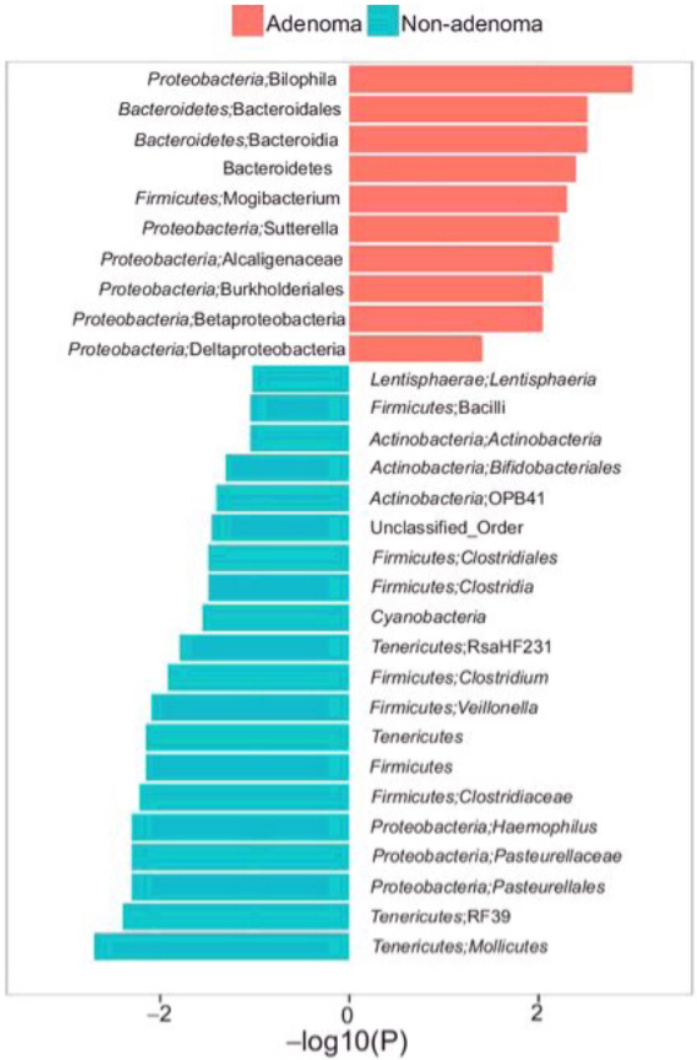


B

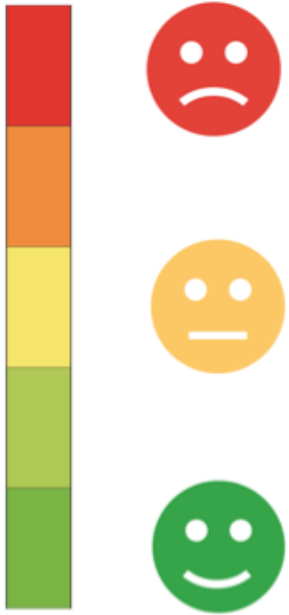
De novo approaches

- Biondicators discovery / supervised learning

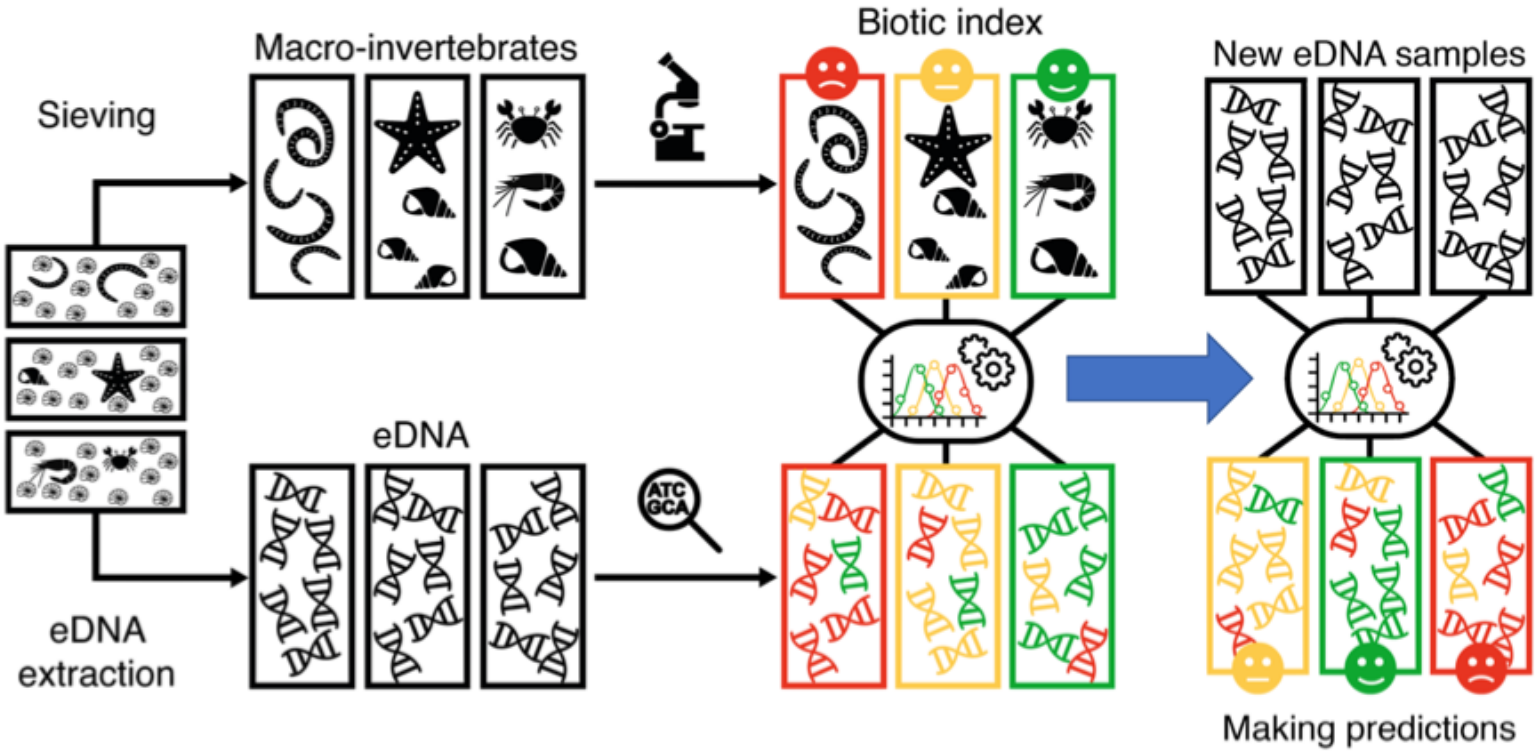




Biomonitoring: another classification problem → Machine Learning !

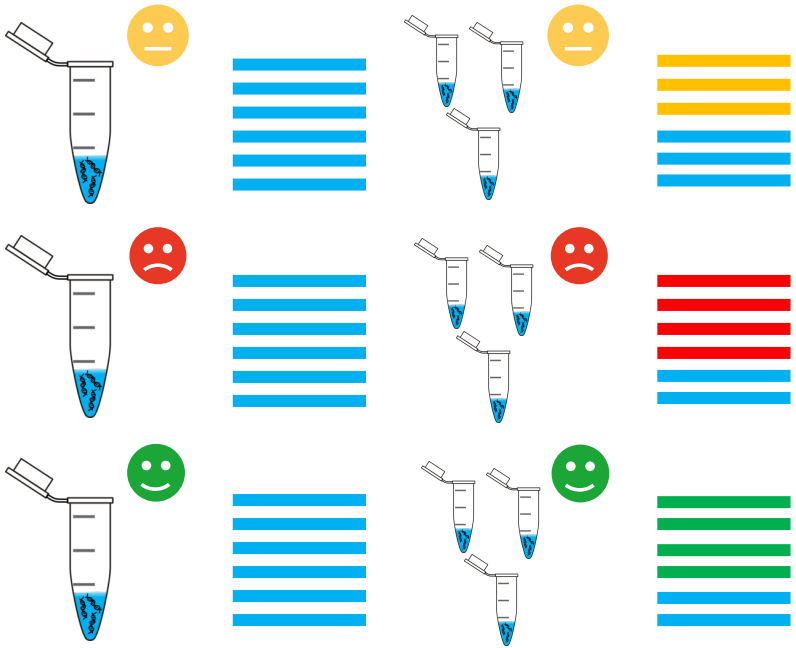


Biomonitoring: another classification problem → Machine Learning !

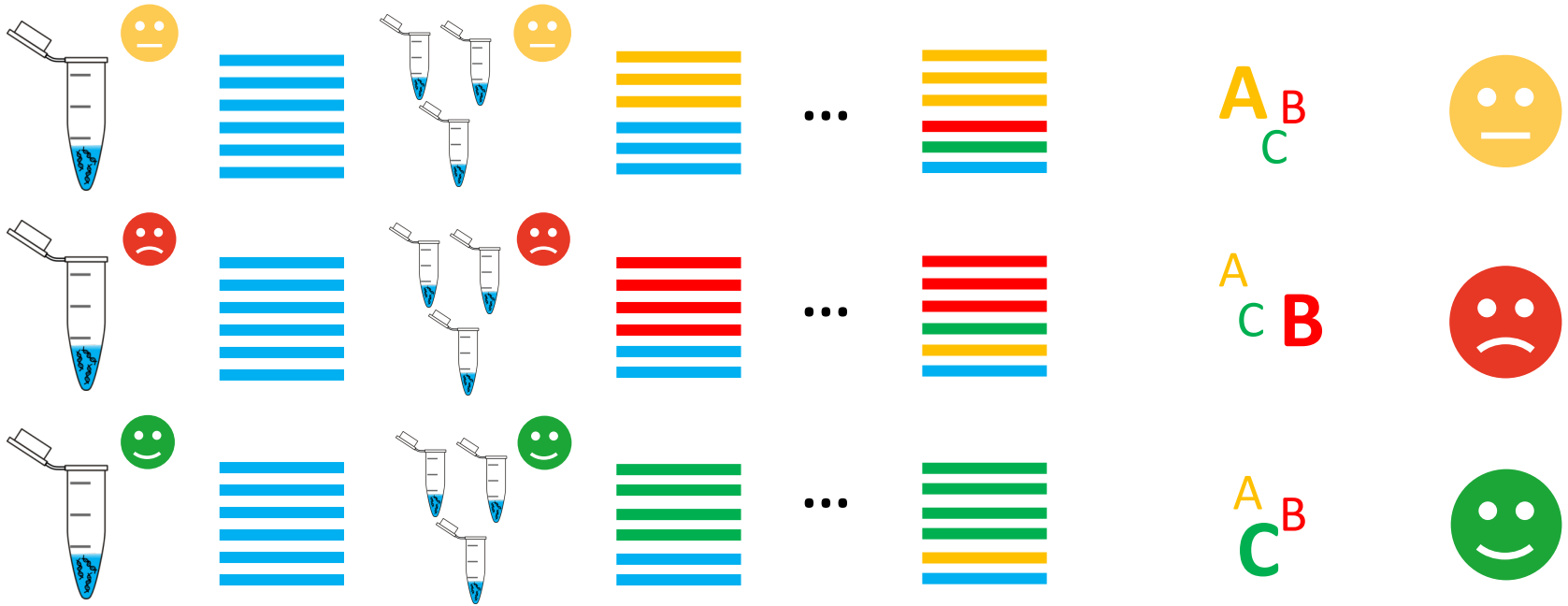


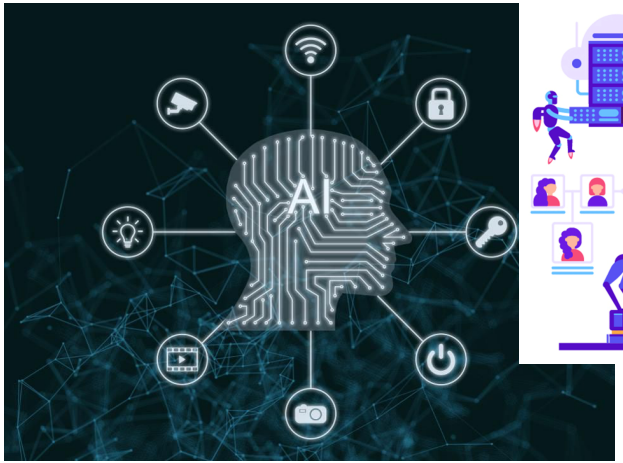
Cordier et al., 2017

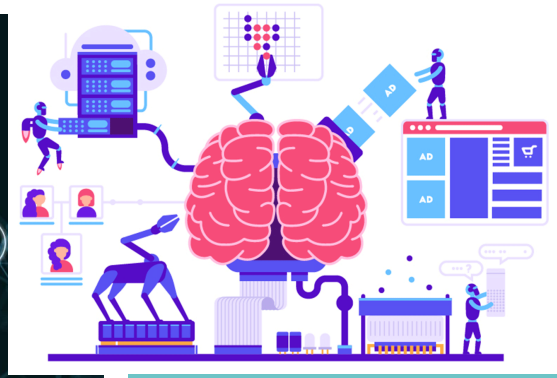
Biomonitoring: another classification problem → Machine Learning !



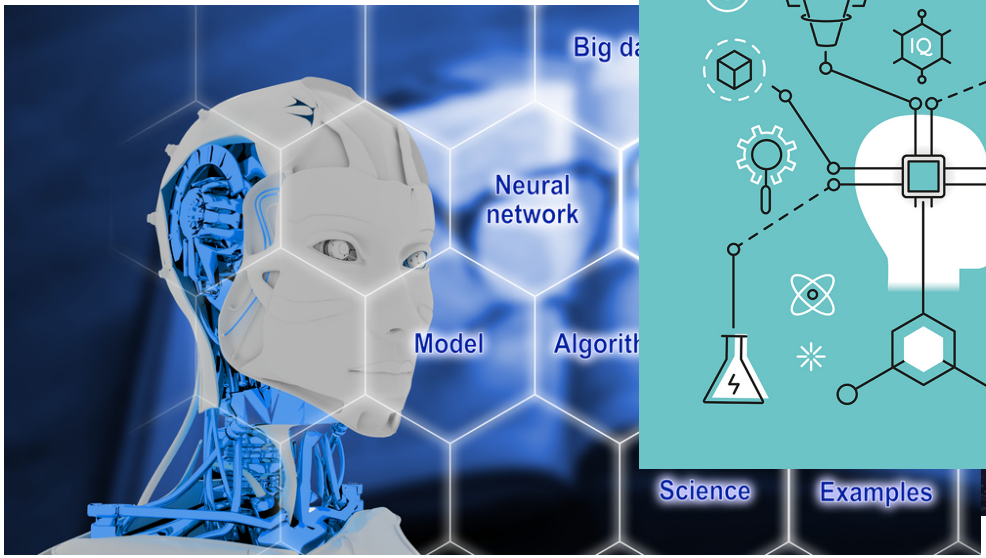
Biomonitoring: another classification problem → Machine Learning !

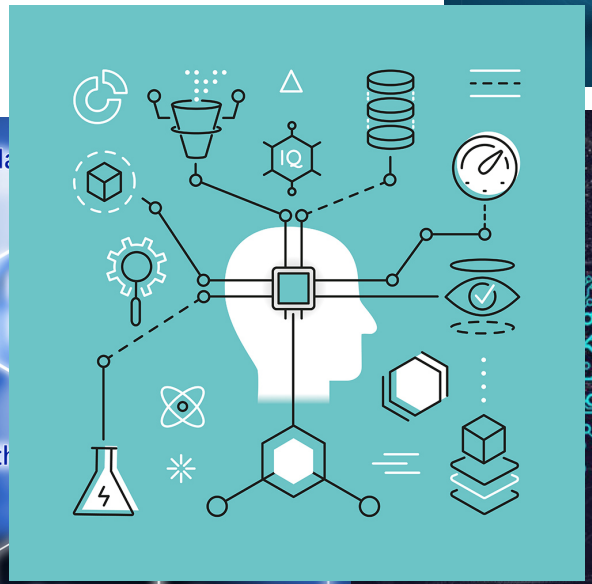


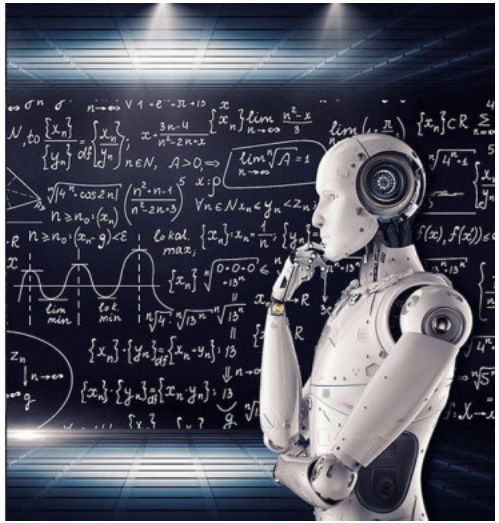

AI

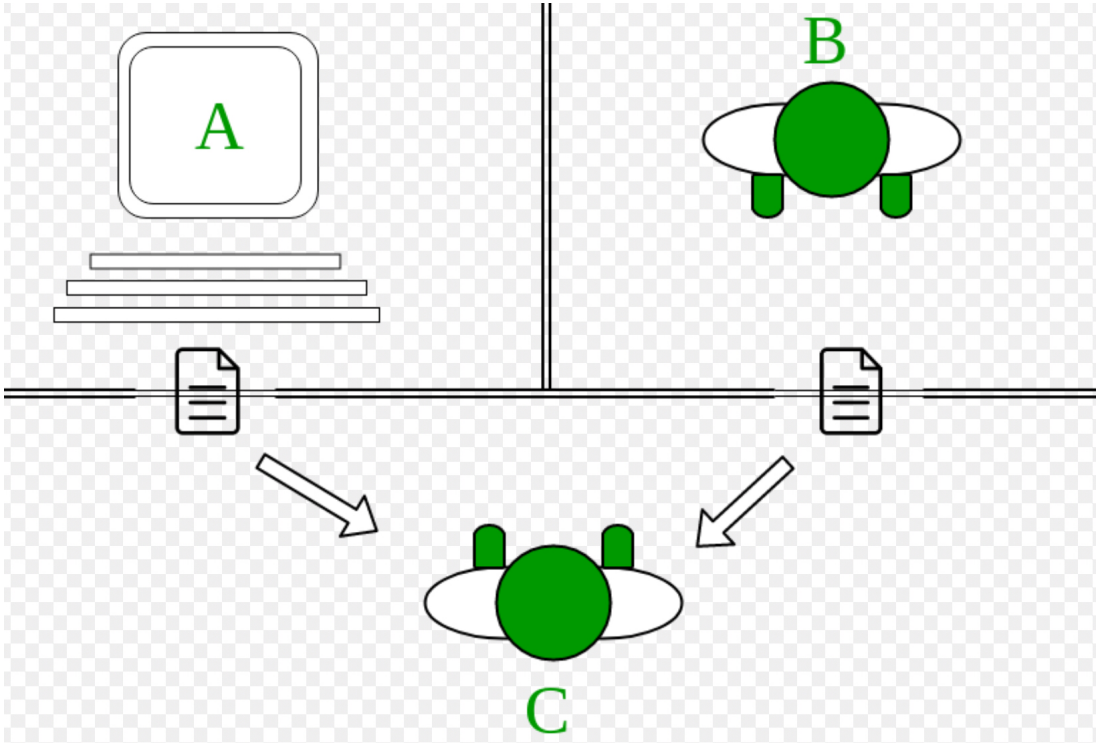

AD
AD
AD

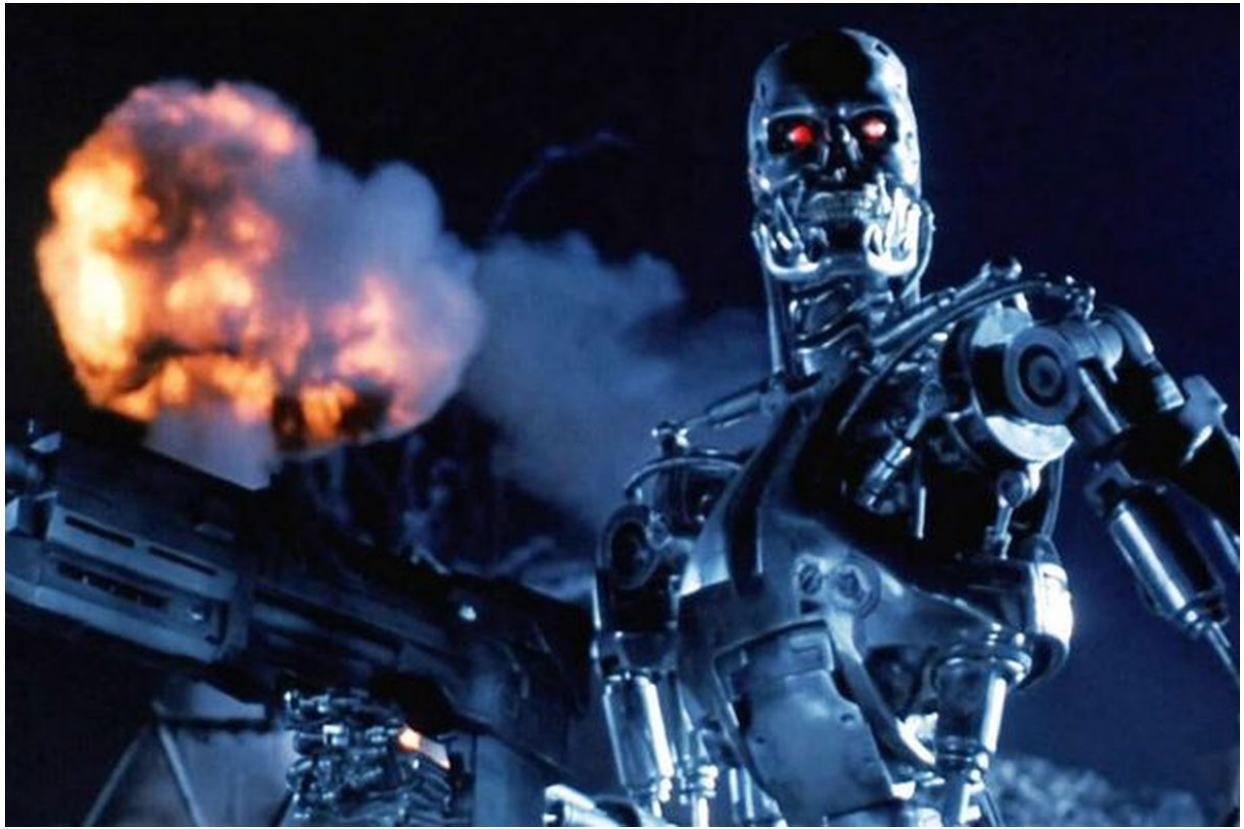

edureka!
What is
Machine Learning?

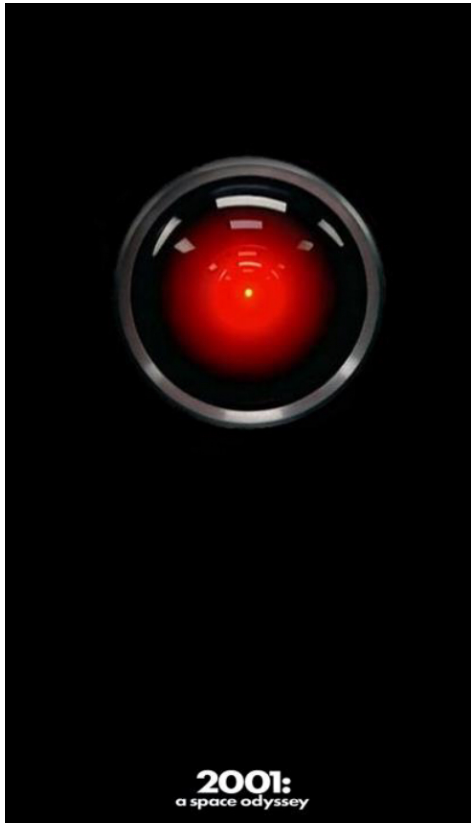

Big da
Neural
network
Model
Algorith
Science
Examples


IQ

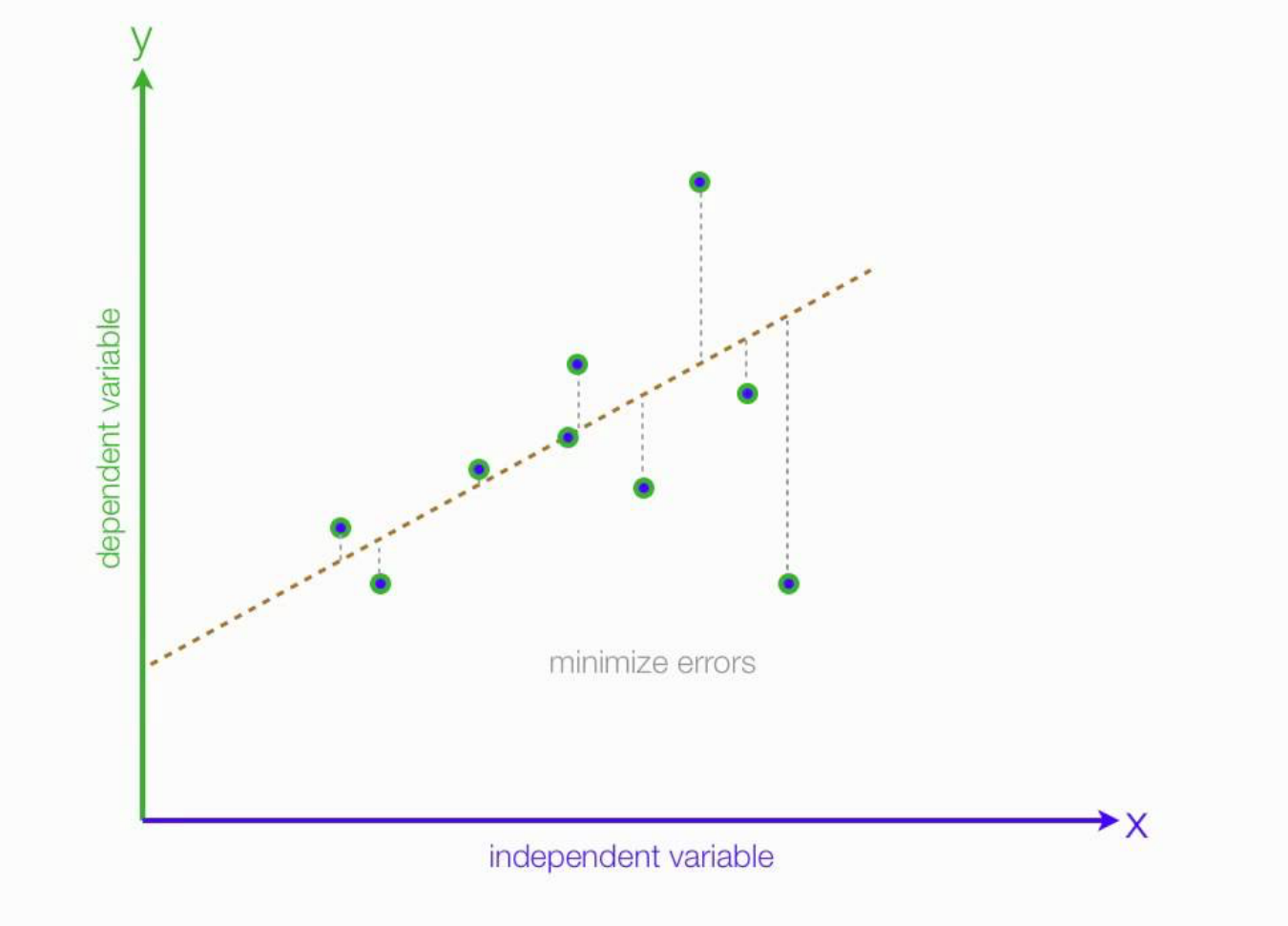



A
B
C




2001:
a space odyssey

Association between things



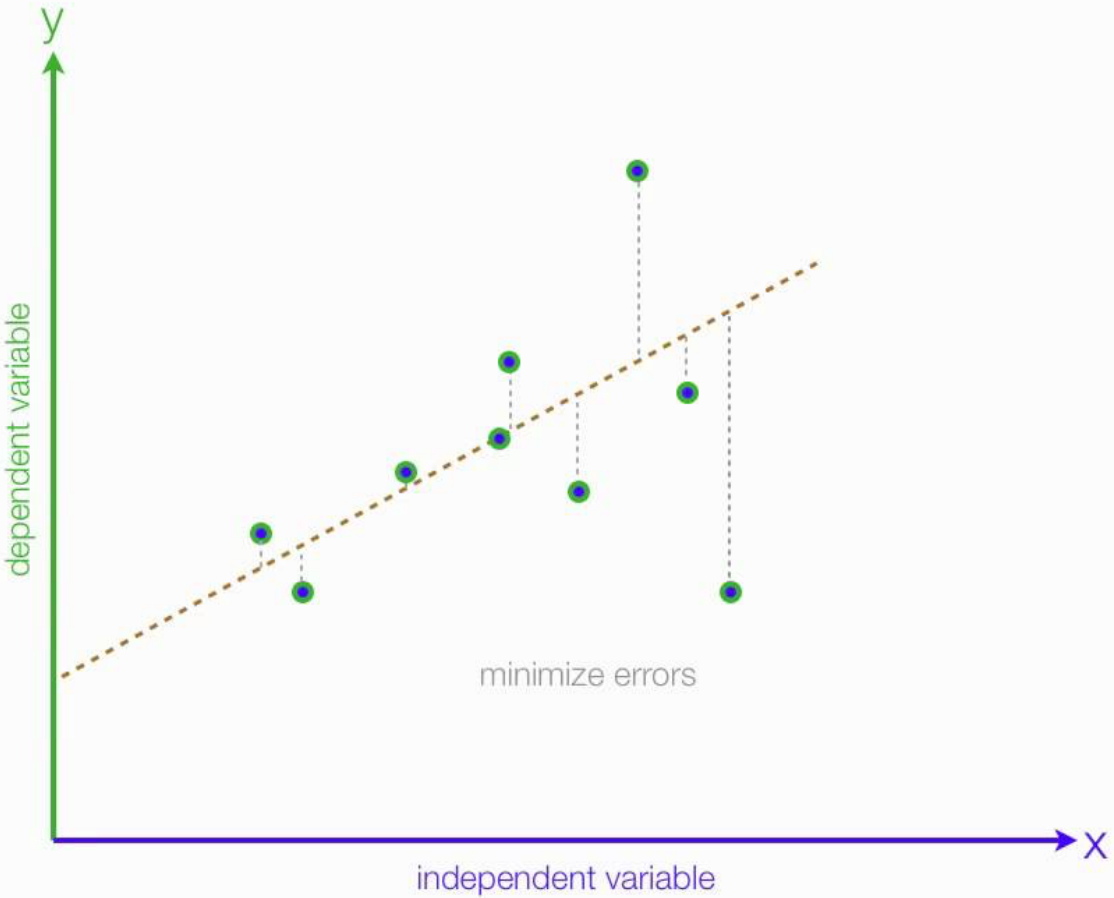
Statistical Science
2001, Vol. 16, No. 3, 199–231

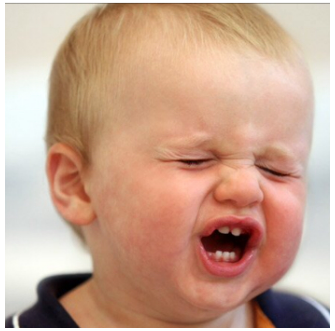
Statistical Modeling: The Two Cultures

Leo Breiman

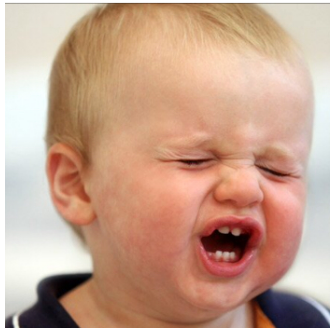
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

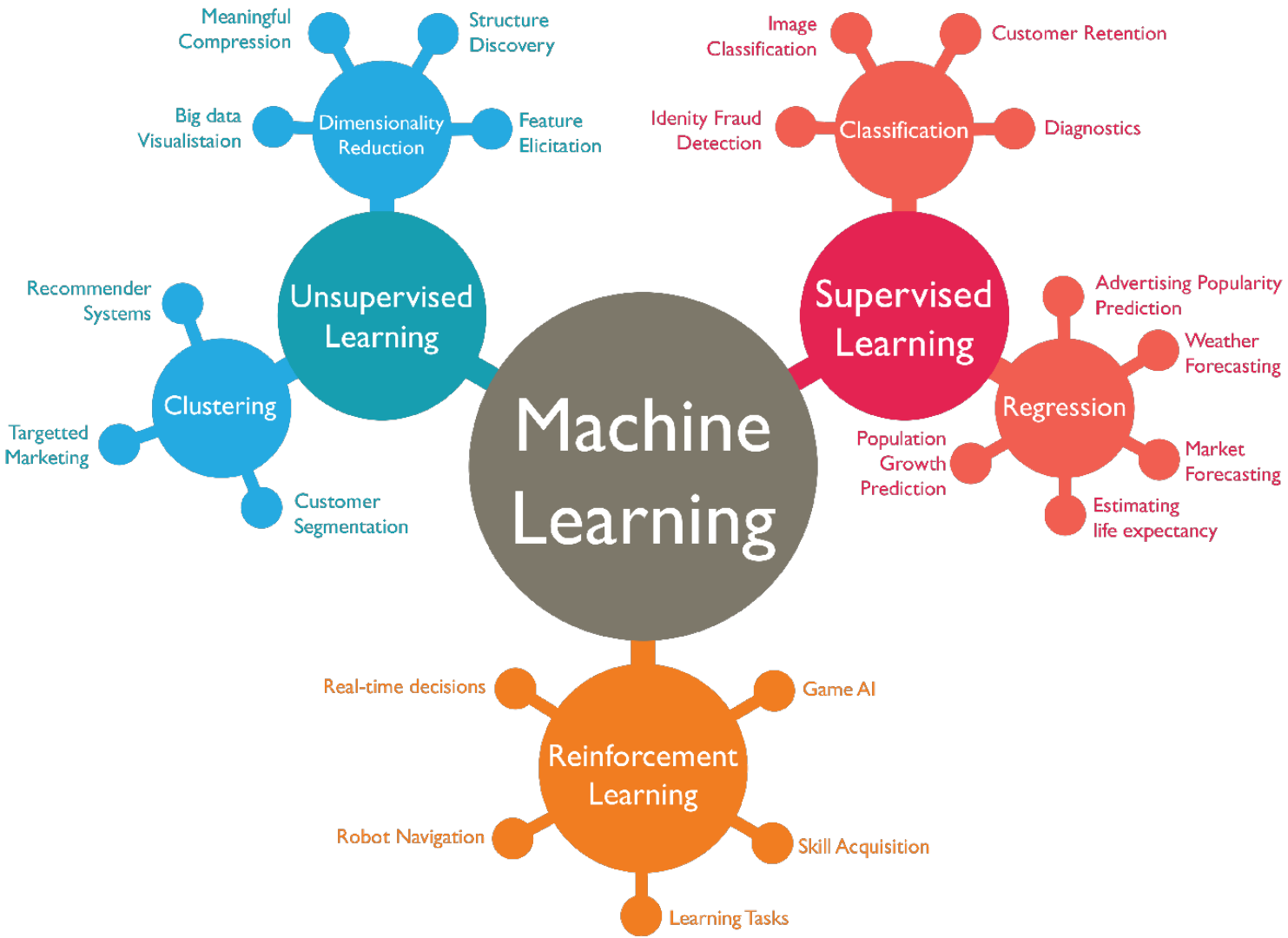
Association between things

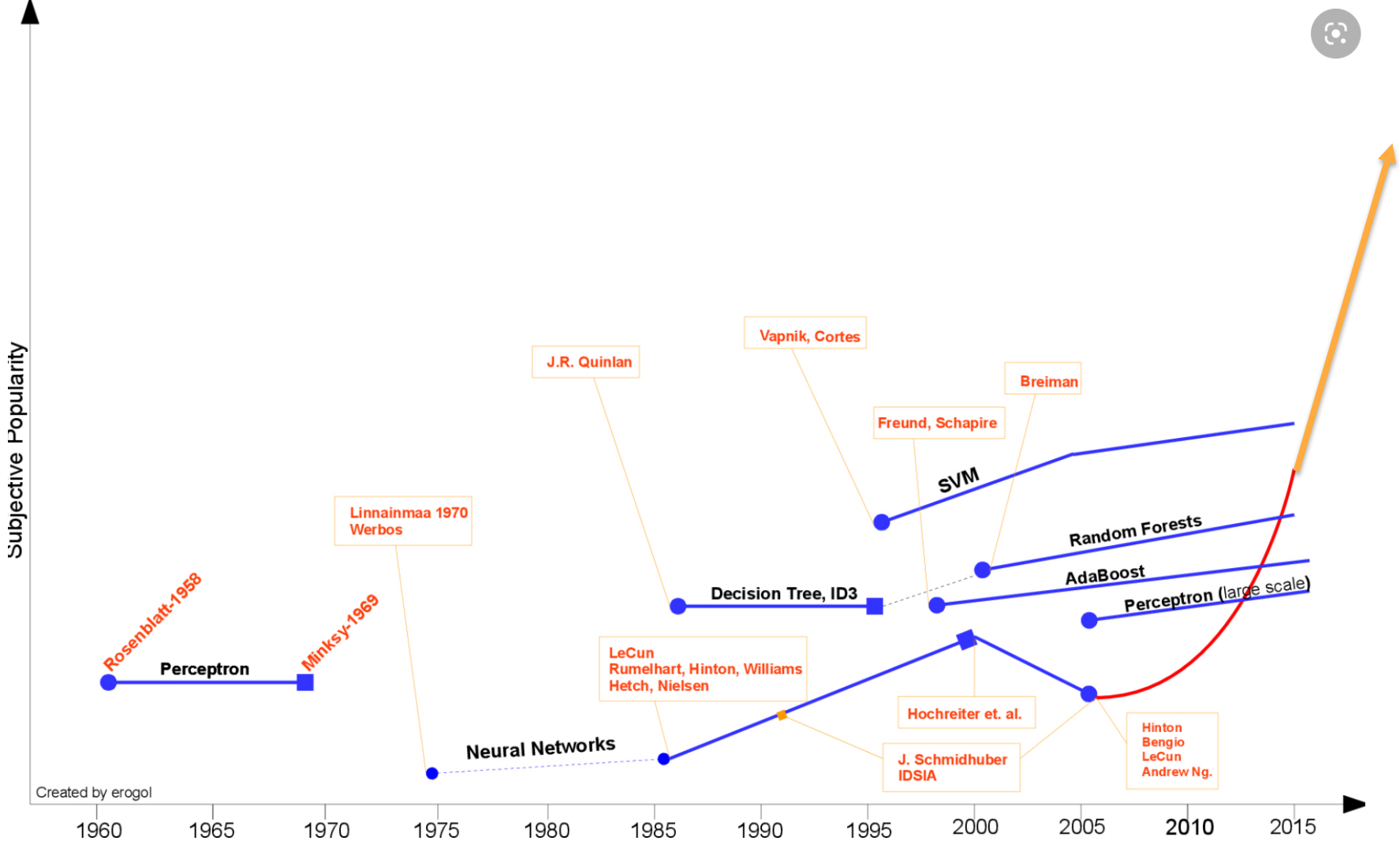




Association between things

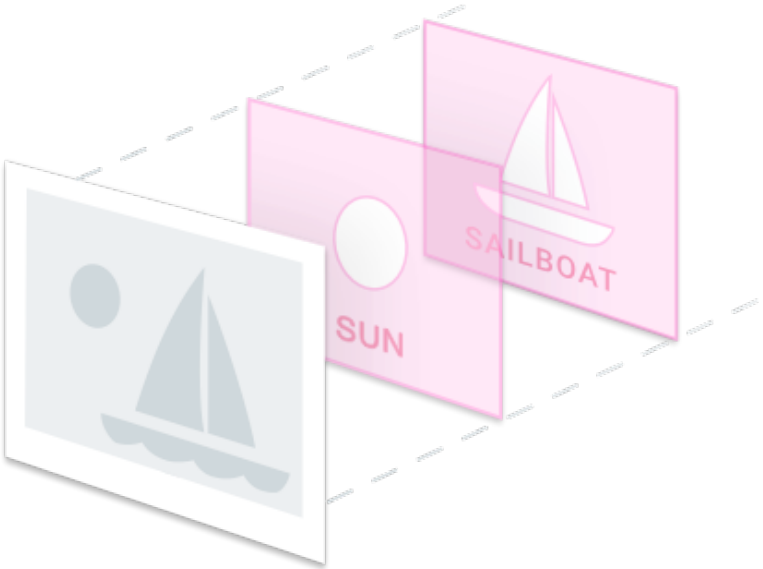



Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks


Subjective Popularity
Rosenblatt-1958
Perceptron
Minksy-1969
Linnainmaa 1970
Werbos
Neural Networks
J.R. Quinlan
Decision Tree, ID3
Vapnik, Cortes
Freund, Schapire
Breiman
SVM
Random Forests
AdaBoost
Perceptron (large scale)
LeCun
Rumelhart, Hinton, Williams
Hetch, Nielsen
Hochreiter et. al.
J. Schmidhuber
IDSIA
Hinton
Bengio
LeCun
Andrew Ng.
Created by erogol
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015

Machine learning applications

- Widely used in classification problems
  - Computer vision tasks



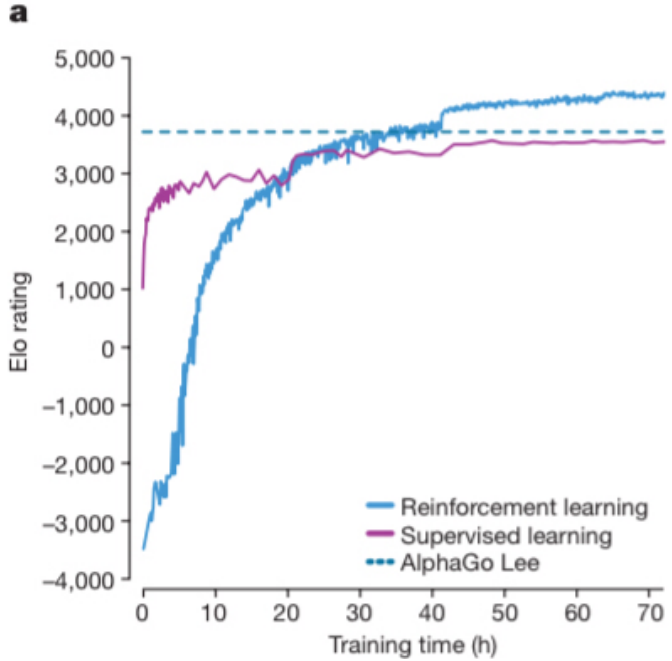
Machine learning applications

ARTICLE

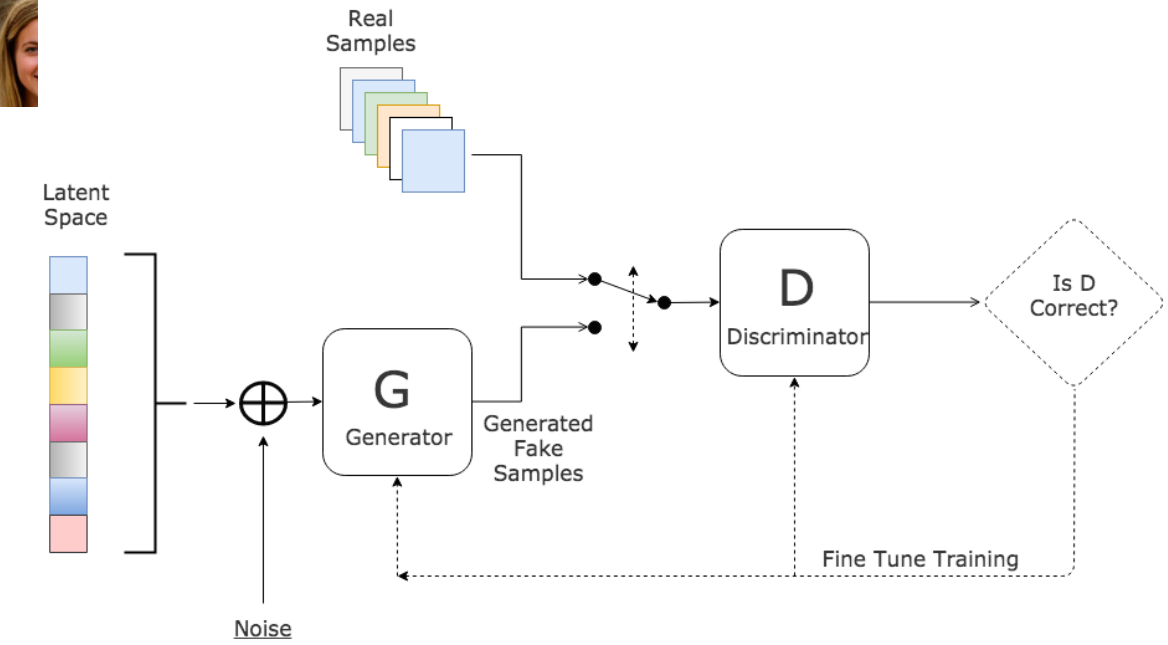
doi:10.1038/nature24270

Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

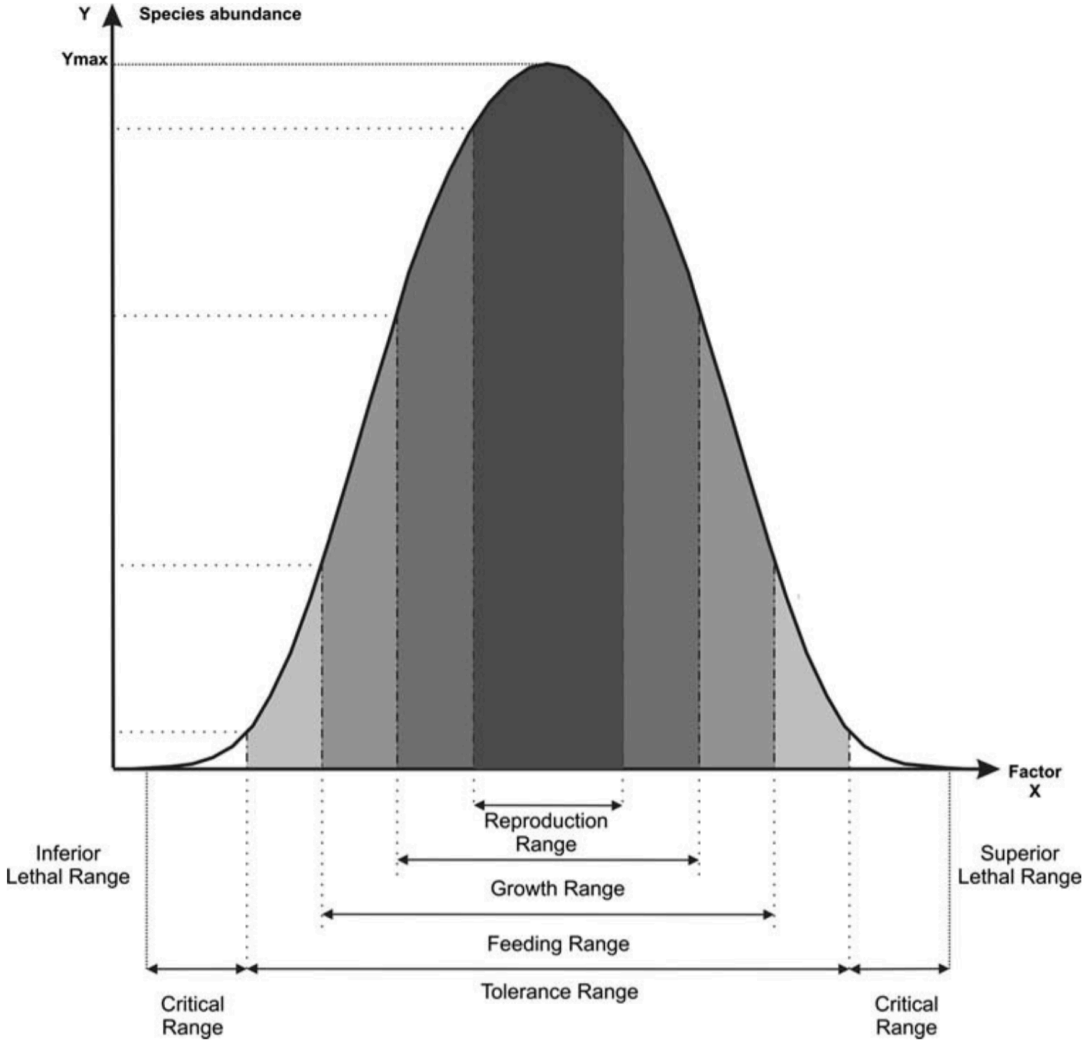


Generative Adversarial Network

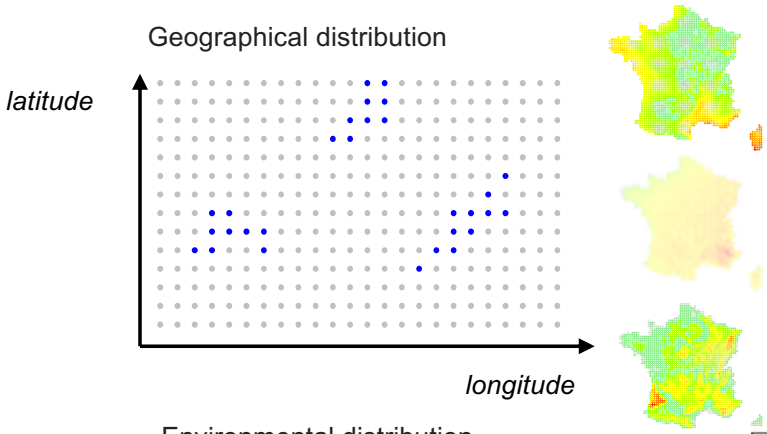


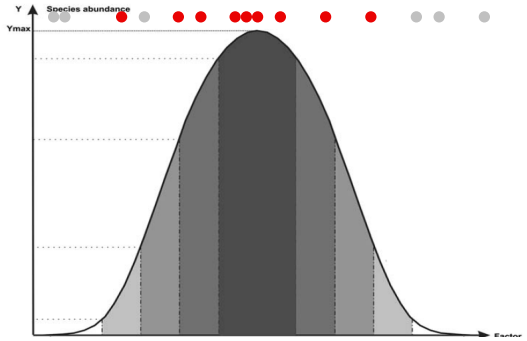
Ecological niche?

- Define the full set of environmental parameters (abiotic and biotic) that allow a species to live and reproduce. (Hutchinson 1957)

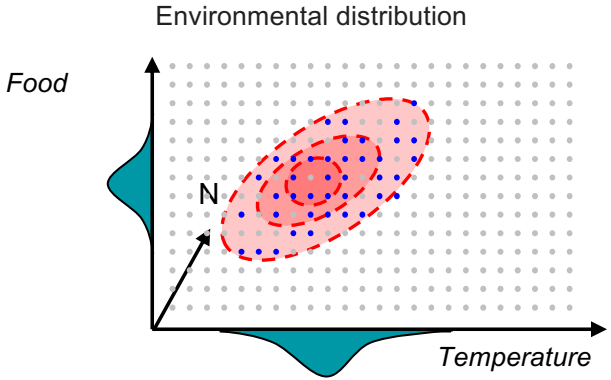


Niche modeling

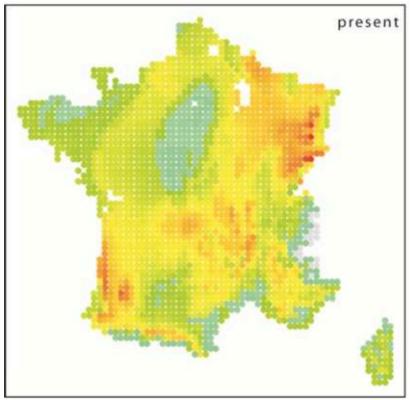


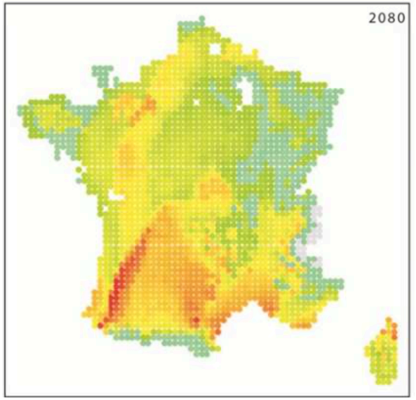


Response to each environmental variables

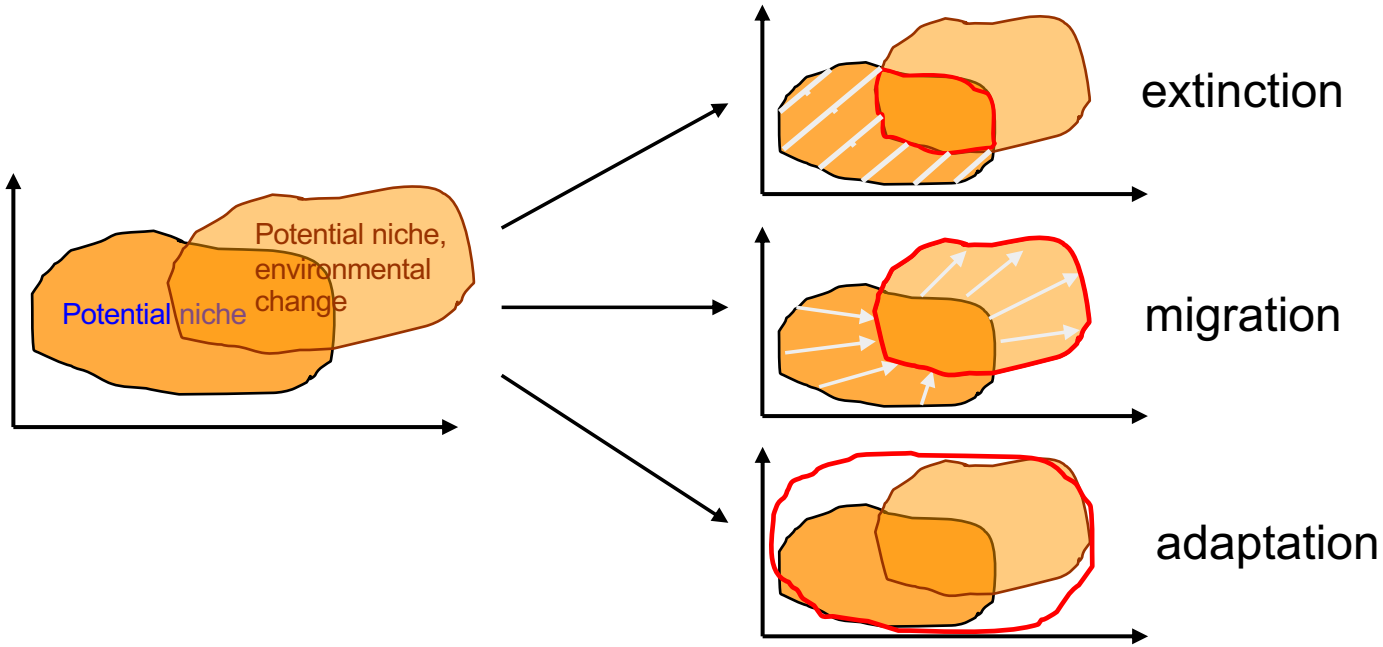


N dimensions : biotic layers, salinity…



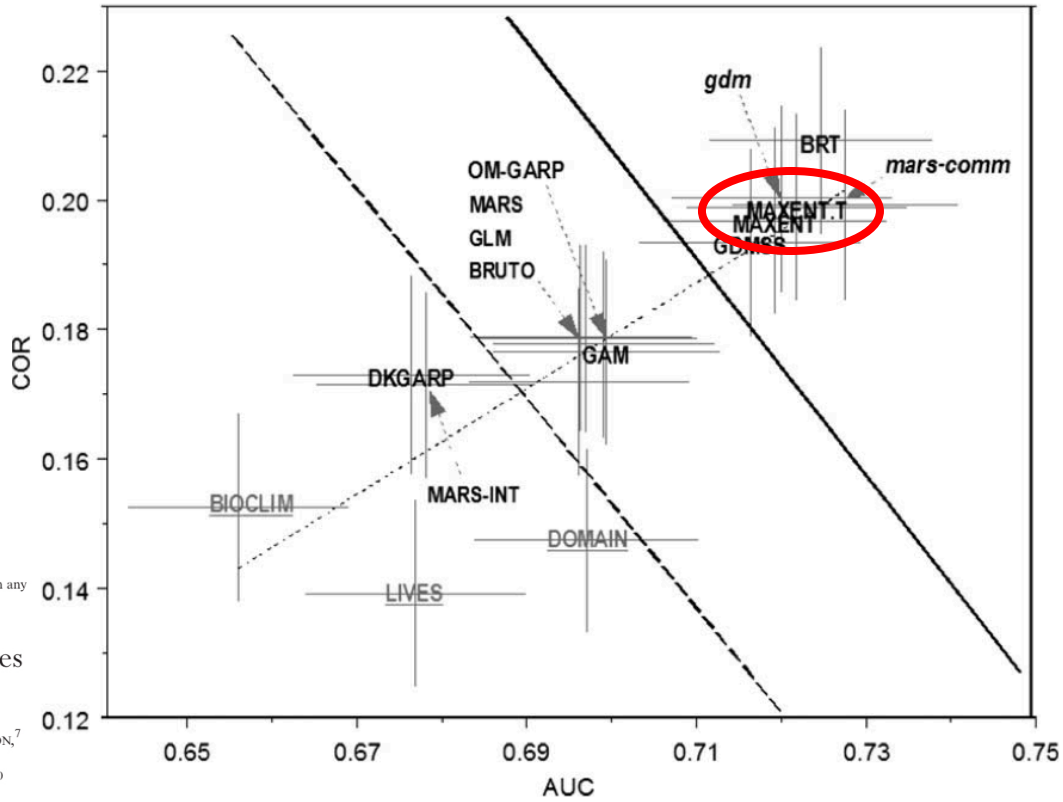


Which response to changes?



Tools for modelling

- BIOCLIM
- GLM
- GAM
- BRT
- **MAXENT** (Phillips, 2004)
- BIOMOD2 R package (Thuillier, 2009)



Elith et al., 2006

Ecological Monographs, 89(3), 2019, e01370

A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels

Anna Norberg [1,34] Nerea Abrego,[2,3] F. Guillaume Blanchet,[4] Frederick R. Adler,[5,6] Barbara J. Anderson,[7] Jani Anttila,[1] Miguel B. Araújo,[8,9,10] Tad Dallas,[1] David Dunson,[11] Jane Elith,[12] Scott D. Foster,[13] Richard Fox,[14] Janet Franklin,[15] William Godsoe,[16] Antoine Guisan,[17,18] Bob O'Hara,[19] Nicole A. Hill,[20] Robert D. Holt,[21] Francis K. C. Hui,[22] Magne Husby,[23,24] John Atle Kålås,[25] Aleksi Lehikoinen,[26] Miska Luoto,[27] Heidi K. Mod,[18] Graeme Newell,[28] Ian Renner,[29] Tomas Roslin [3,30] Janne Soininen [27] Wilfried Thuiller,[31] Jarno Vanhatalo,[1] David Warton,[32] Matt White,[28] Niklaus E. Zimmermann,[33] Dominique Gravel,[4] and Otso Ovaskainen [1,2]

ML in genomics → Funct. annotation tax.classification

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences

Morgan G I Langille[1,14], Jesse Zaneveld[2,14], J Gregory Caporaso[3,4], Daniel McDonald[5,6], Dan Knights[7,8], Joshua A Reyes[9], Jose C Clemente[10], Deron E Burkepile[11], Rebecca L Vega Thurber[2], Rob Knight[10,12], Robert G Beiko[1] & Curtis Huttenhower[9,13]

Opinion

Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks

David A. Bohan,[1,*] Corinne Vacher,[2] Alireza Tamaddoni-Nezhad,[3] Alan Raybould,[4] Alex J. Dumbrell,[5] and Guy Woodward[6]

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy[▽]†

Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]*

Center for Microbial Ecology[1] and Department of Microbiology and Molecular Genetics,[2] Michigan State University, East Lansing, Michigan 48824

From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer

Aaron Weimann,[a,b,c] Kyra Mooren,[a,c] Jeremy Frank,[d] Phillip B. Pope,[d] Andreas Bremges,[a,b] Alice C. McHardy[a,b,c]

Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany[a]; German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany[b]; Department for Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany[c]; Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway[d]

IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences

Adithya Murali[1], Aniruddha Bhargava[2] and Erik S. Wright[3*]

ML in genomics → Medical research

MEDICAL RESEARCH

Machine learning classifies cancer

Brain tumours are often classified by visual assessment of tumour cells, yet such diagnoses can vary depending on the observer. Machine-learning methods to spot molecular patterns could improve cancer diagnosis. SEE ARTICLE P.469

DEREK WONG & STEPHEN YIP

Accurate diagnosis is essential for appropriate disease treatment. A core technique used to diagnose brain cancer today is the microscope-based analysis of tumour samples on glass slides, termed histology. However, this requires the appraisal of subtle cellular alterations, which in some cases may lead to different classifications for a given sample by different individuals. Nowadays, technological developments enable vast amounts of molecular data to be obtained and assessed for a tumour without the need for such subjective diagnostics. Machine-based-learning approaches are being developed to aid the diagnosis of clinical samples, and on page 469, Capper *et al.*[1] report such a method for classifying brain tumours on the basis of molecular patterns.

In 1926, a publication entitled *A Classification of the Tumors of the Glioma Group on a Histo-Genetic Basis with a Correlated Study of Prognosis*[2] by neurosurgeons Percival Bailey and Harvey Cushing provided early insight into the development, cellular characteristics and clinical consequences of glioma, a type of

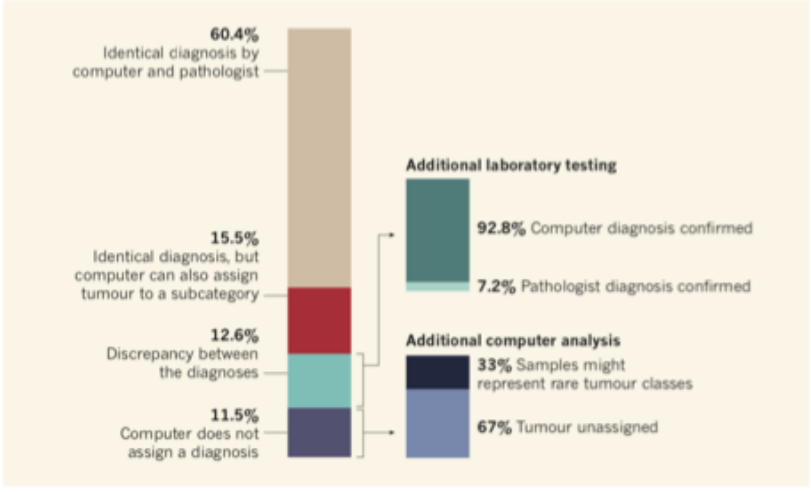
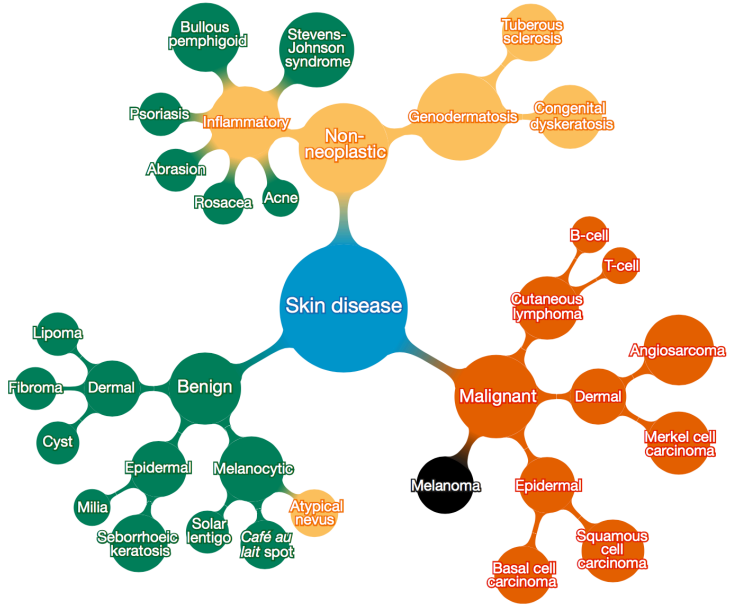


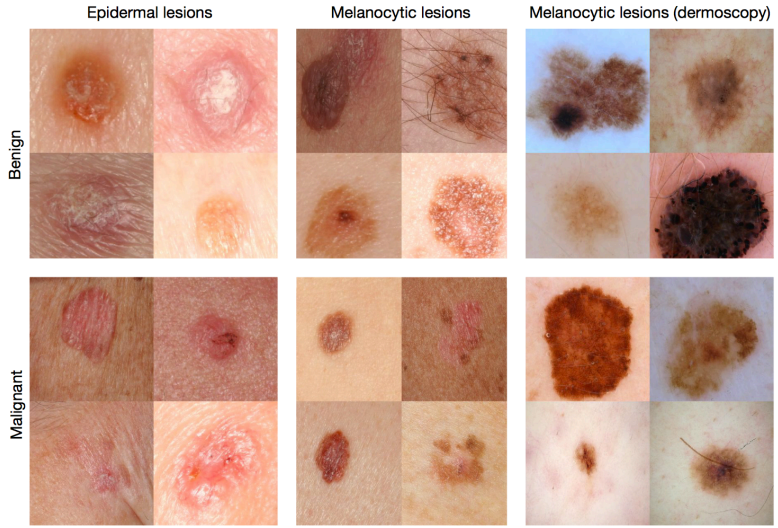
Figure 1 | Tumour classification using a machine-learning approach. Capper *et al.*[1] used a machine-learning approach to classify brain tumours on the basis of genome-wide patterns of a type of DNA alteration called methylation. The computer was trained using methylation data for tumour samples that had been diagnosed by pathologists using standard microscopy-based analysis or analysis of selected genes. After training, the computer was given 1,104 test cases. The authors compared the diagnoses made by the computer and by the pathologists. Although the machine was unable to diagnose all specimens, of the specimens that it classified, the machine-based diagnosis was more accurate or could assign tumours to more-specific subcategories than the classifications made by the pathologists.

ML in genomics → Medical research

- Widely used in classification problems
  - Cancer detection



Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.4 ± 0.8%
CNN - PA	**72.1 ± 0.9%**



Esteva et al., 2017

ML in genomics → Marine research

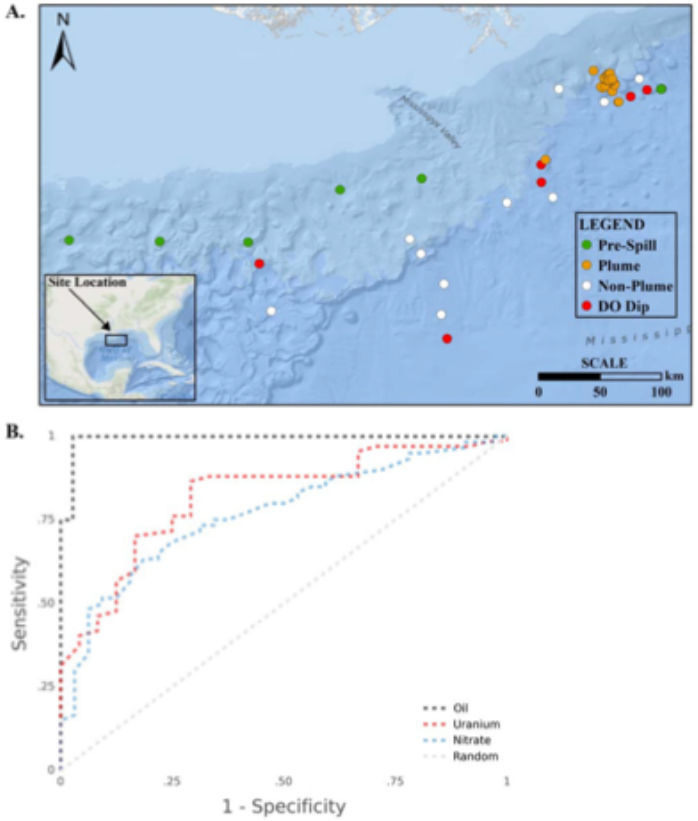
→ Impact assessment - biomonitoring

Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors

Mark B. Smith,[a] Andrea M. Rocha,[b] Chris S. Smillie,[c] Scott W. Olesen,[d] Charles Paradis,[e] Liyou Wu,[f] James H. Campbell,[b,m] Julian L. Fortney,[g] Tonia L. Mehlhorn,[h] Kenneth A. Lowe,[h] Jennifer E. Earles,[h] Jana Phillips,[h] Steve M. Techtmann,[g] Dominique C. Joyner,[g] Dwayne A. Elias,[b] Kathryn L. Bailey,[b] Richard A. Hurt, Jr.,[b] Sarah P. Preheim,[d] Matthew C. Sanders,[d] Joy Yang,[c] Marcella A. Mueller,[h] Scott Brooks,[h] David B. Watson,[h] Ping Zhang,[f] Zhili He,[f] Eric A. Dubinsky,[i] Paul D. Adams,[i,l] Adam P. Arkin,[i,l] Matthew W. Fields,[j] Jizhong Zhou,[f] Eric J. Alm,[a,c,d] Terry C. Hazen[b,e,g,k]

Here we show that statistical analysis of **DNA from natural microbial communities can be used to accurately identify environmental contaminants**, including uranium and nitrate at a nuclear waste site. In addition to contamination, **sequence data from the 16S rRNA gene alone can quantitatively predict a rich catalogue of 26 geochemical features** collected from 93 wells with highly differing geochemistry characteristics.

Smith et al., mBio, 2015

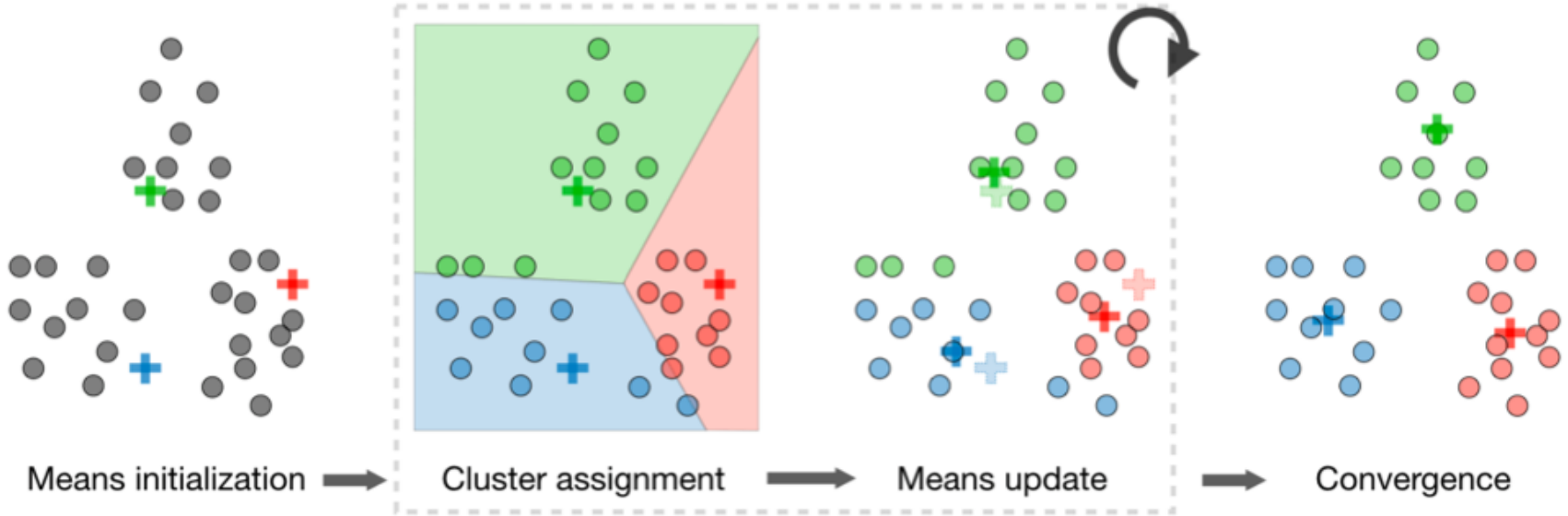


What tool do I need?

	Unsupervised	Supervised
Continuous	● Clustering - dimensionality reduction ○ PCA ○ NMDS ○ K-means	● Regression (linear or not) ○ Linear or polynomial models ○ Regression trees (random forest) ○ Neural networks
Categorical	● Association analysis ○ Correspondence analysis ○ Factorial analysis	● Classification ○ Random forest ○ Neural networks ○ Logistic regression ○ SVMs ○ ...

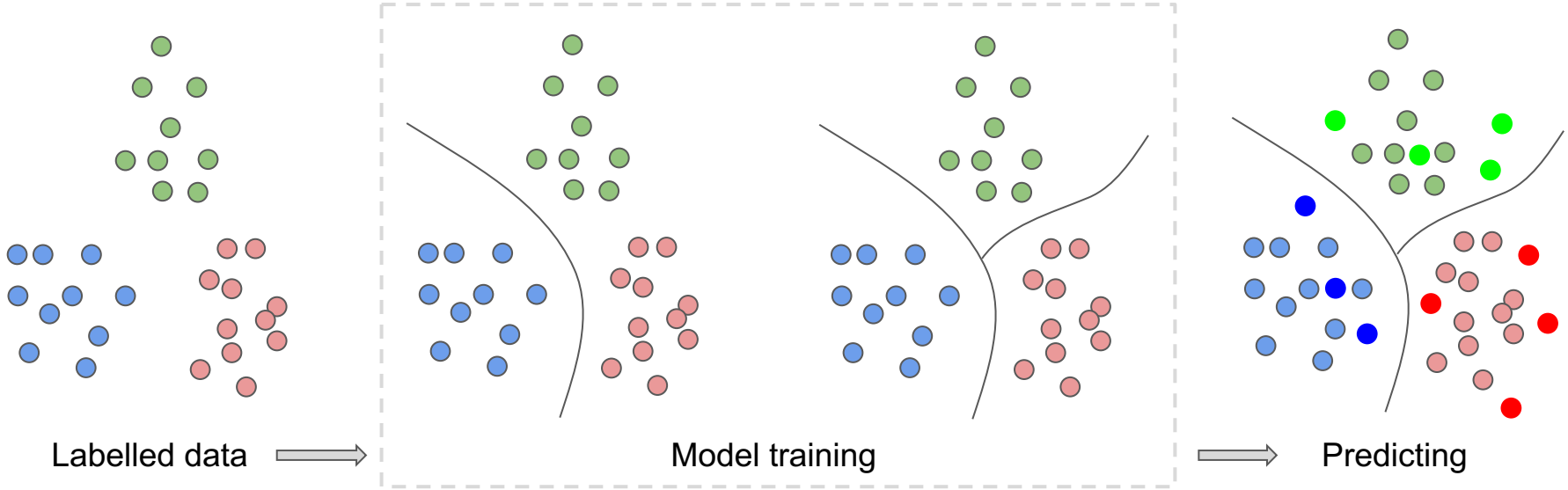
Type of problems in machine learning

Unsupervised: the solution is not known. — **Goal**: Find hidden patterns

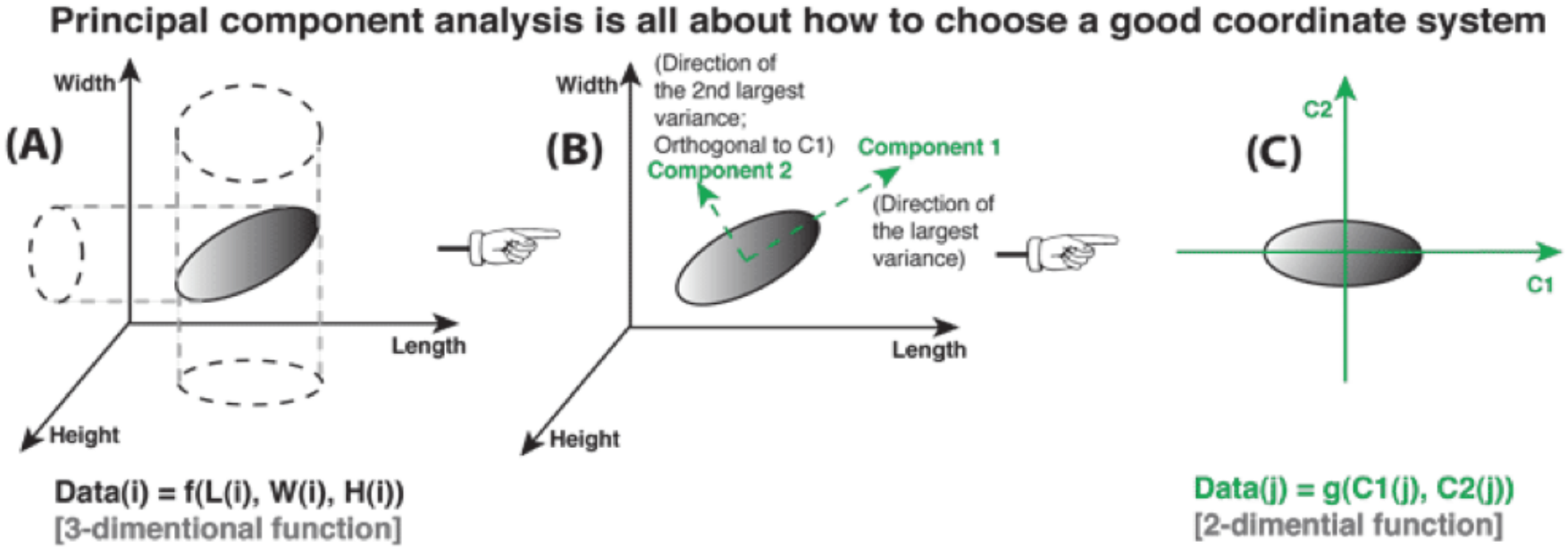


Type of problems in machine learning

Supervised: the solution is known. — **Goal**: Make predictions on new data

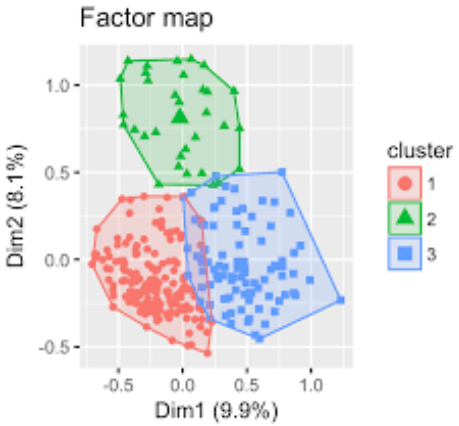


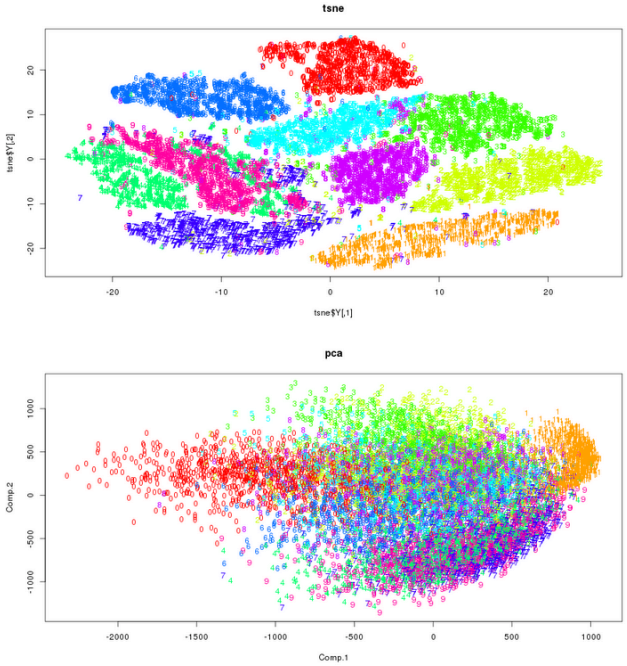
Unsupervised - Dimensionality reduction - PCA

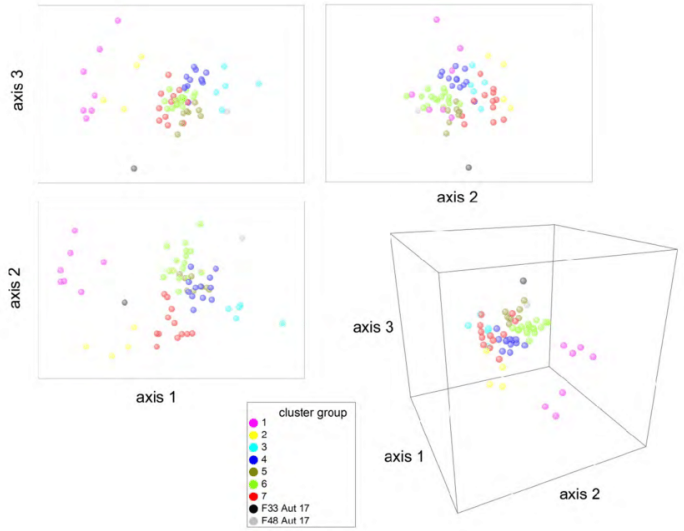


- **Projection of n-dimensional data** on new coordinate axis using eigen-decompositio
- Uses actual OTU abundances with only linear transformation

Unsupervised - Dimensionality reduction - PCA



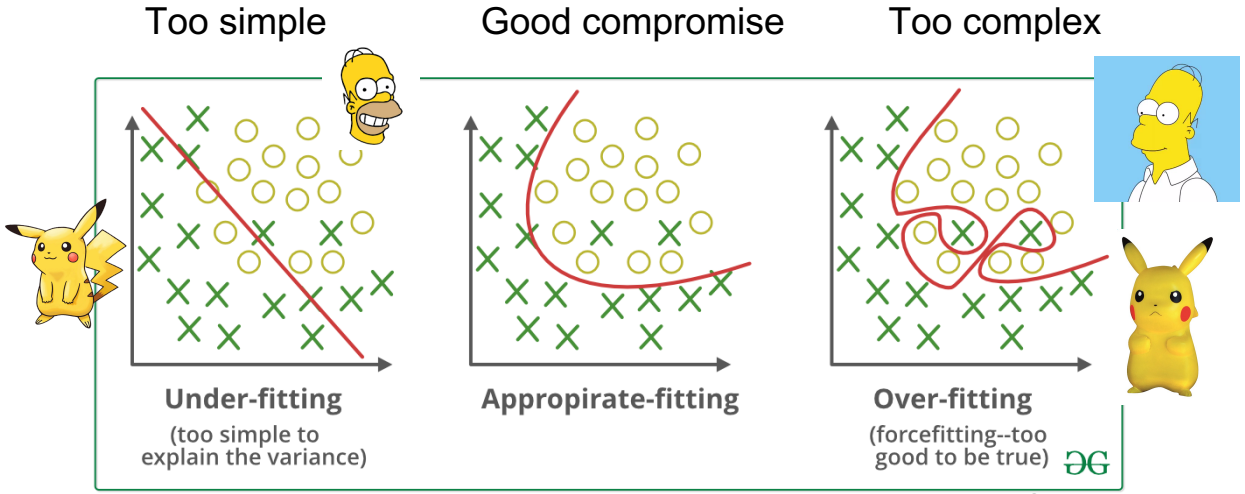


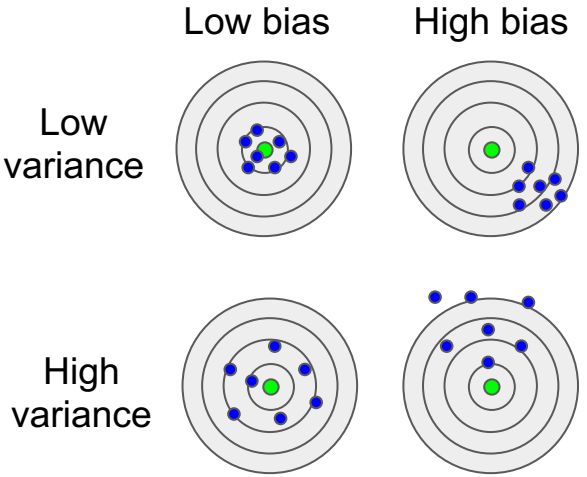


- **Projection of n-dimensional data** on new coordinate axis using eigen-decompositio
- Uses actual OTU abundances with only linear transformation

Supervised - predictive model training

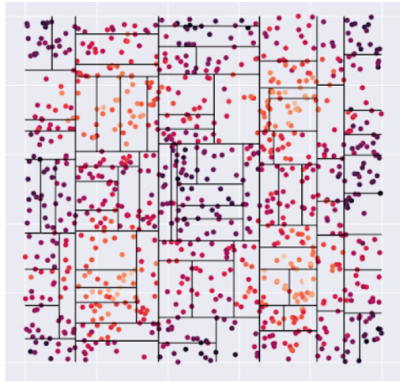
- Inferring a ***function*** from ***labelled training data*** (subset of of the real world!)
- The aim is to ***minimize*** the ***error*** of predictions in the real world.

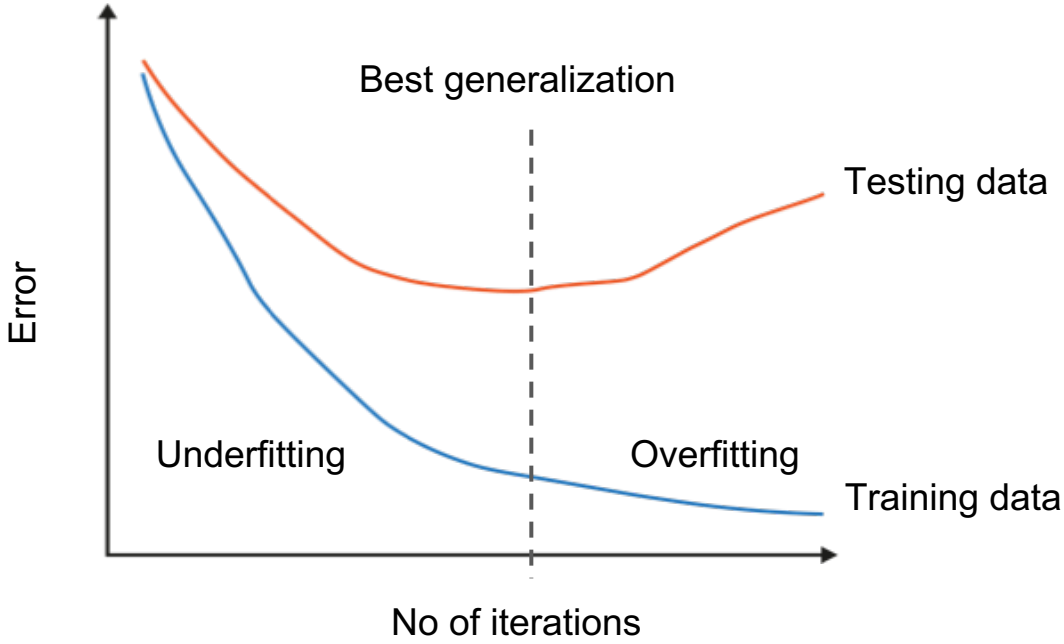


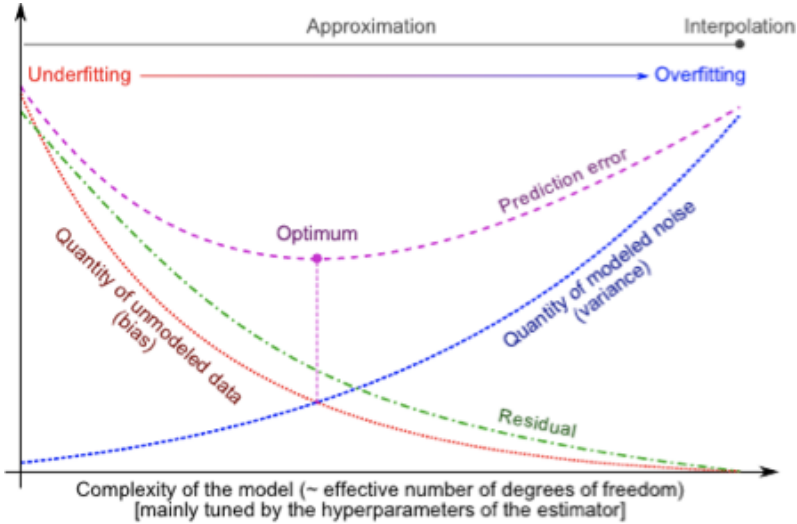


Supervised - predictive model training

- The bias-variance trade-off

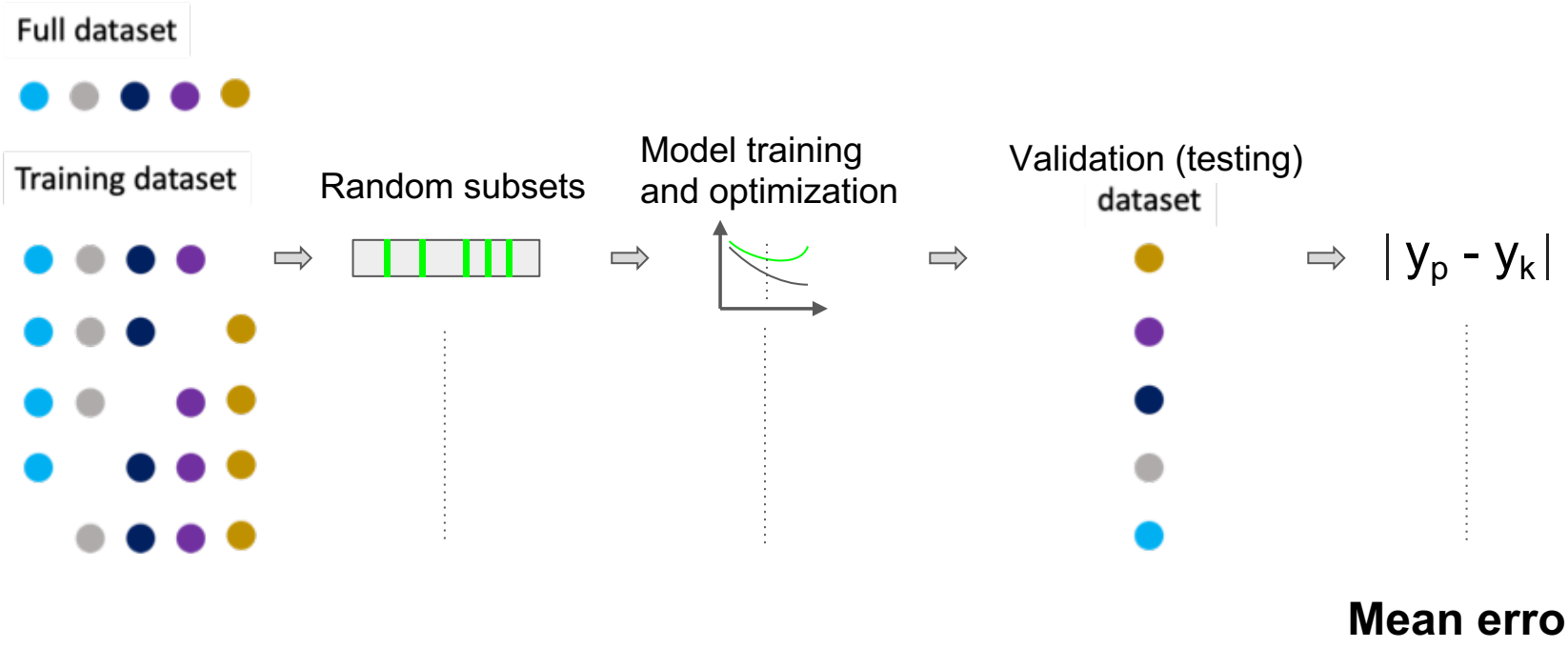






Supervised - predictive model training

- Using cross-validation to measure ***generalization error***

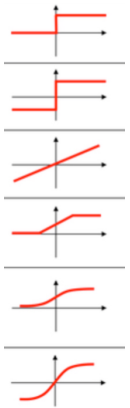


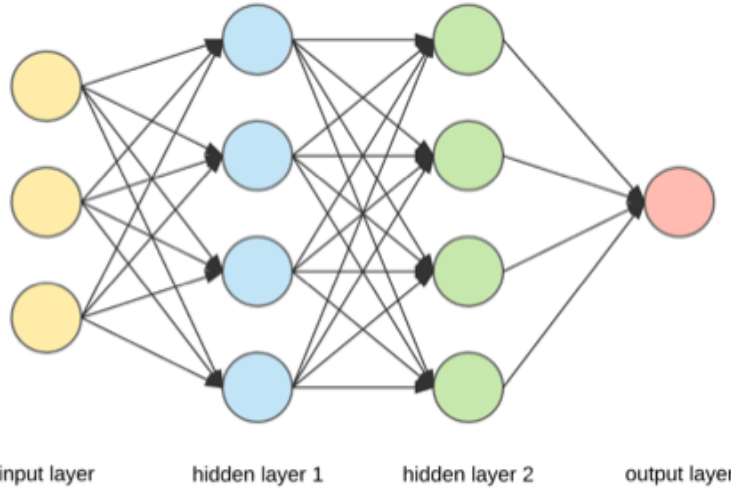
neural networks

Randomly initiate weights for several nodes in several layers of classifiers → "predict" training data and compare to ground truth (using a *loss function*) → backwards pass to try to correct (complex…)

Multiple parameters to play with :

- Number of layers / neurons
- Learning rate
- Activation function (sigmoid, linear...)



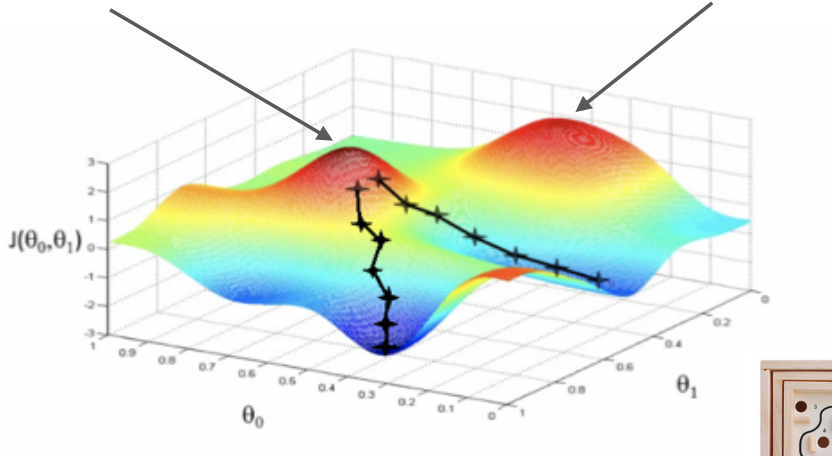


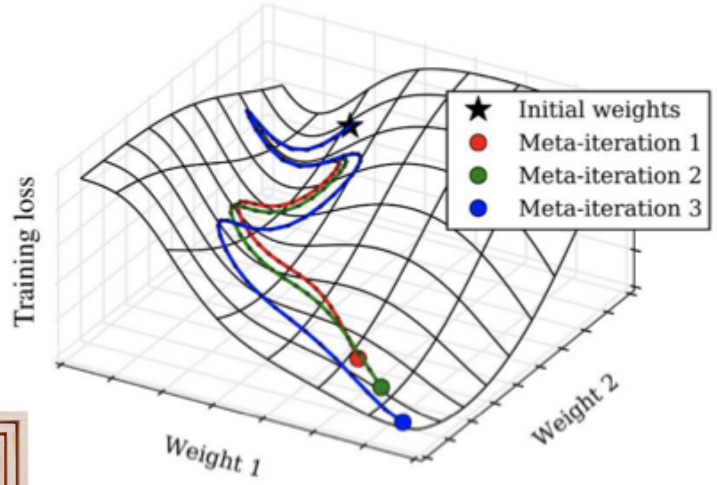
Supervised - predictive model training

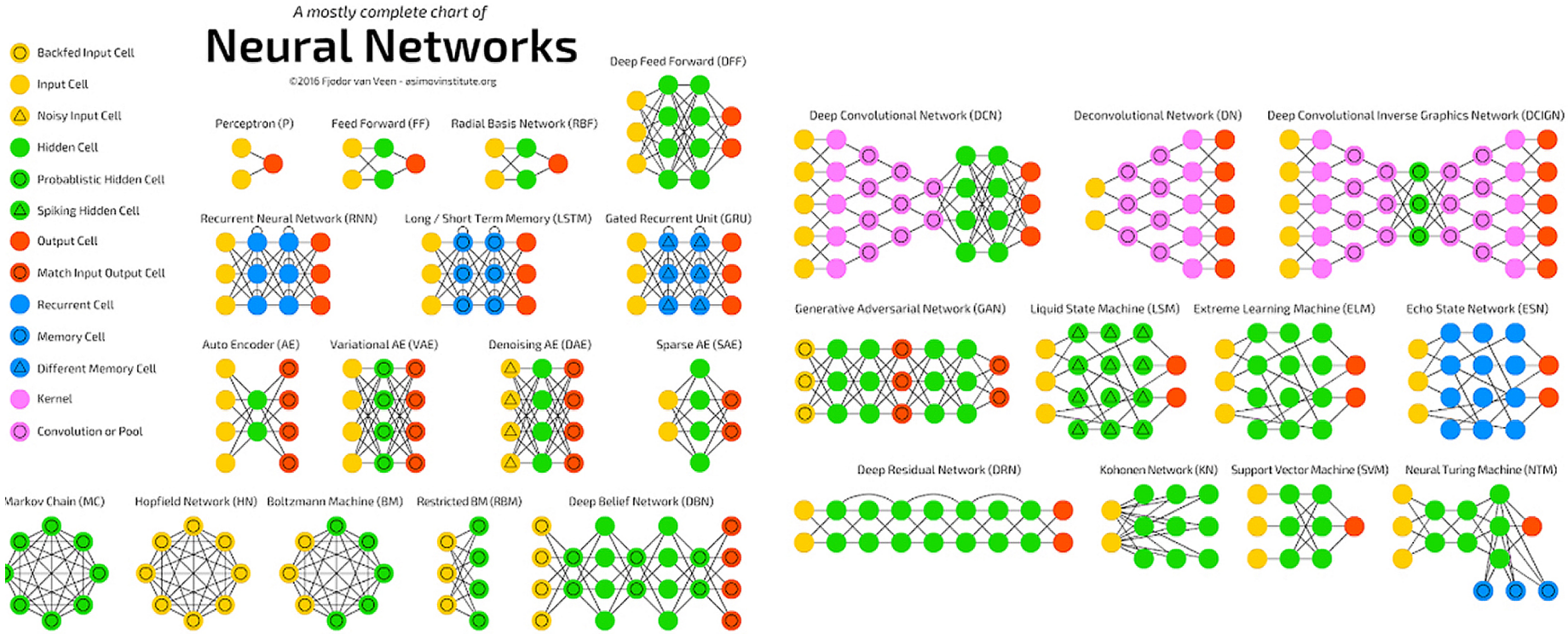
- Hill-climbing vs gradient descent for model optimization

Global optima

Local optima



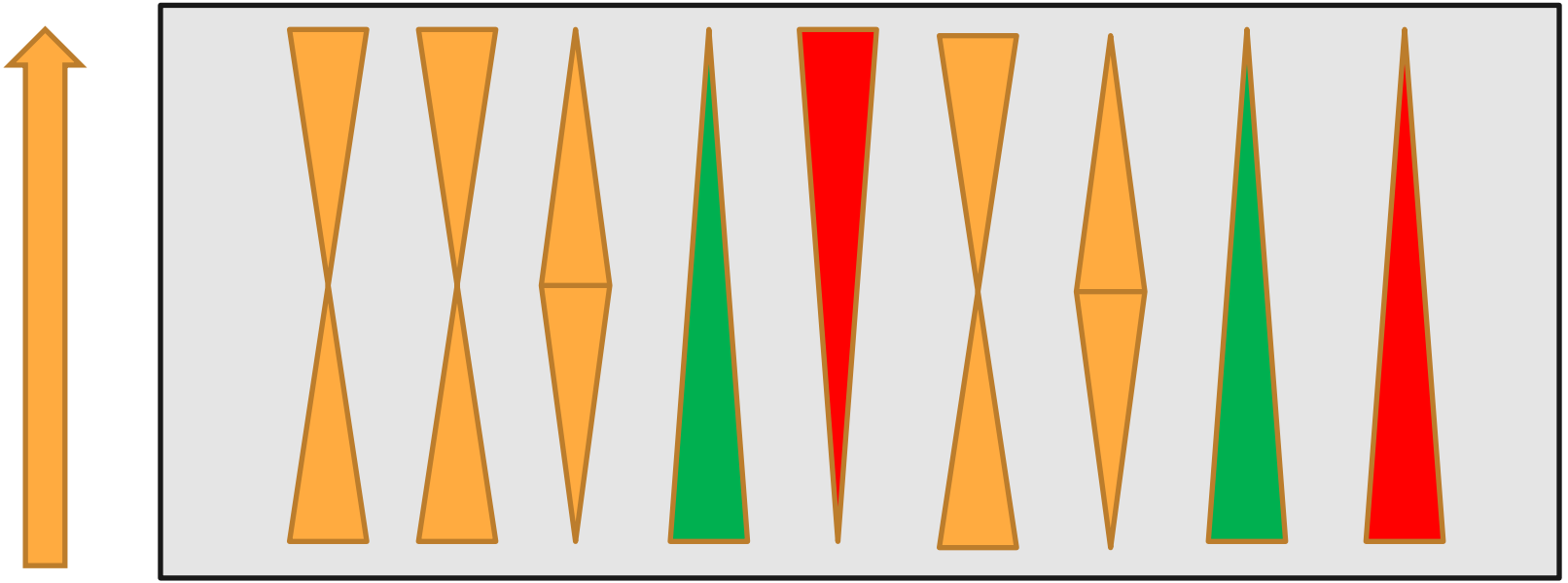


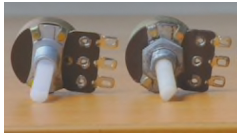


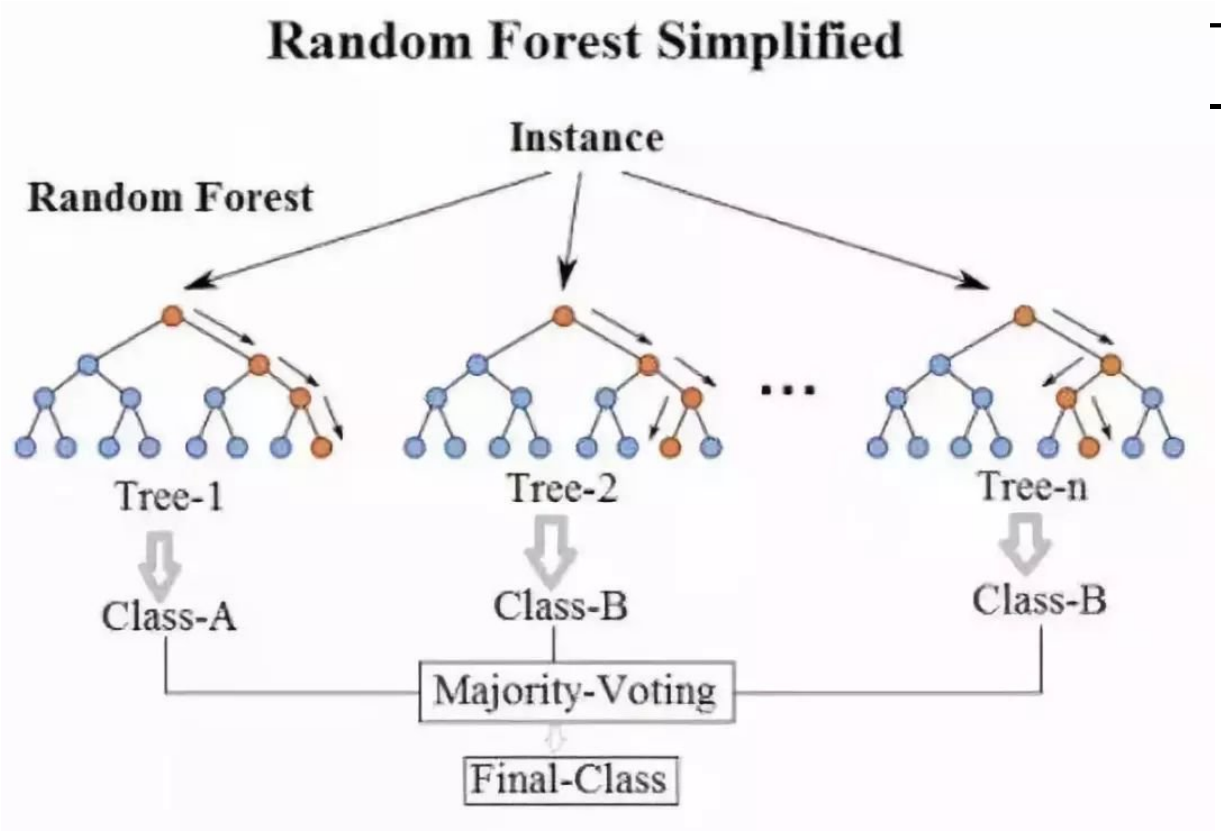
Millions to billions of samples with thousands of variables

eDNA : thousand of samples with millions of variables (OTUs)

eDNA : thousand of samples with millions of variables (OTUs)
—> curse of dimentionality

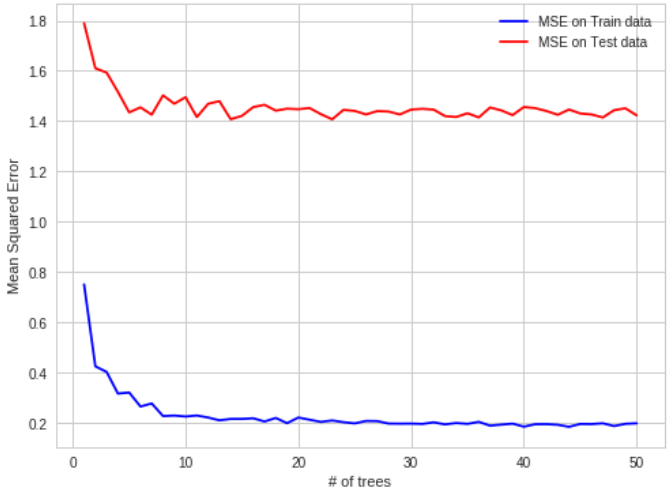


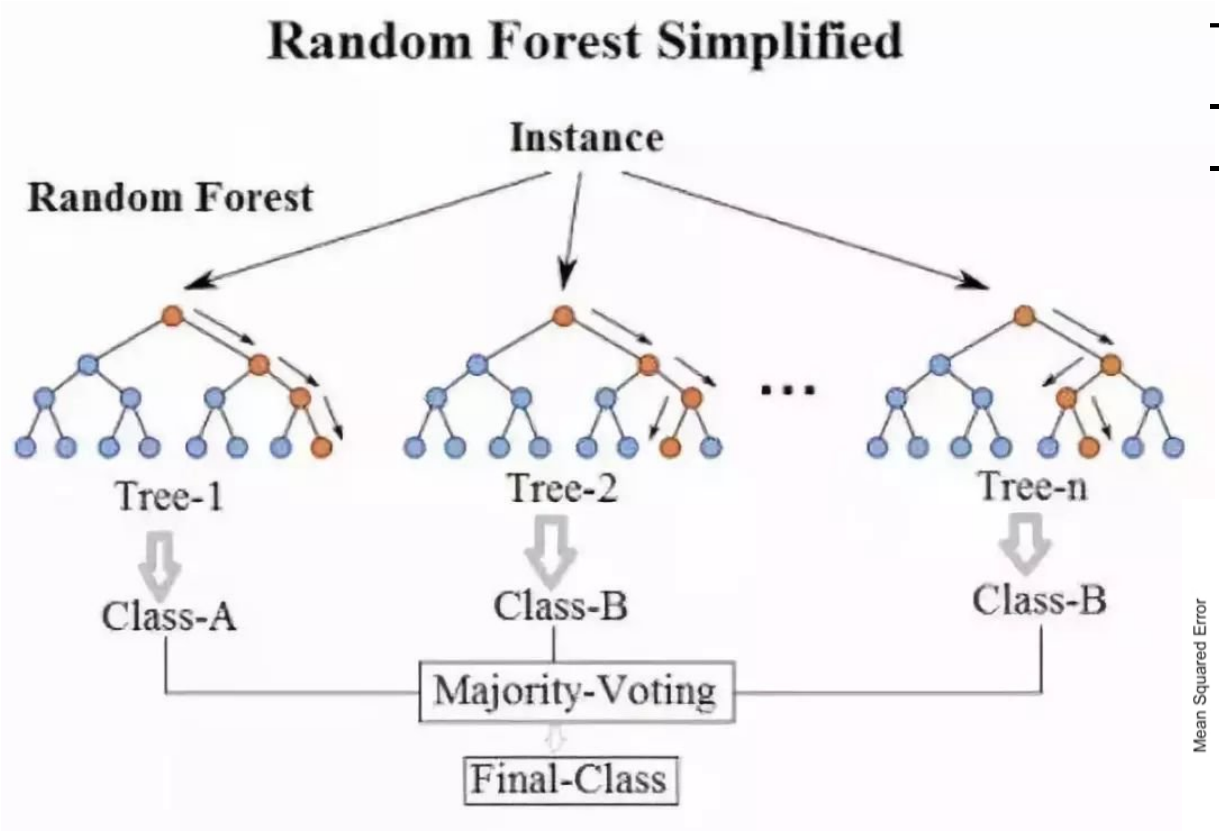




Two parameters:

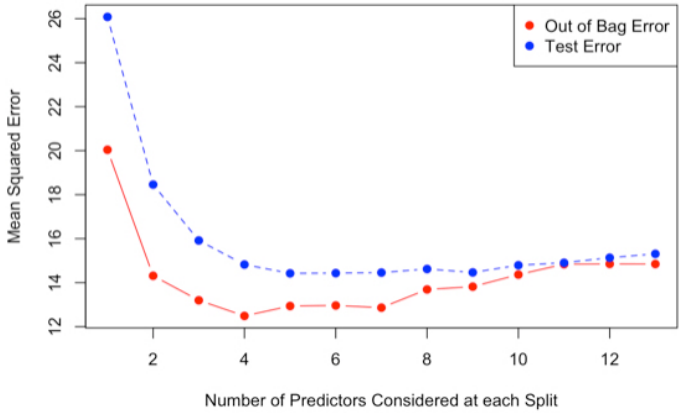
- Number of trees



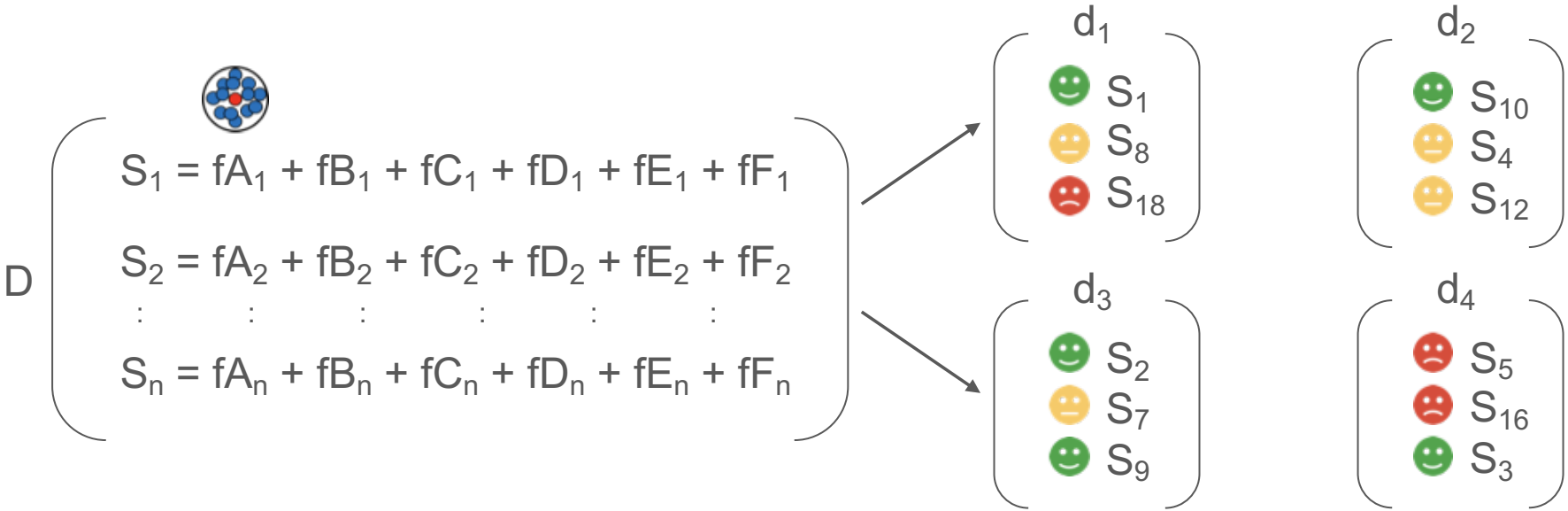


Two parameters:

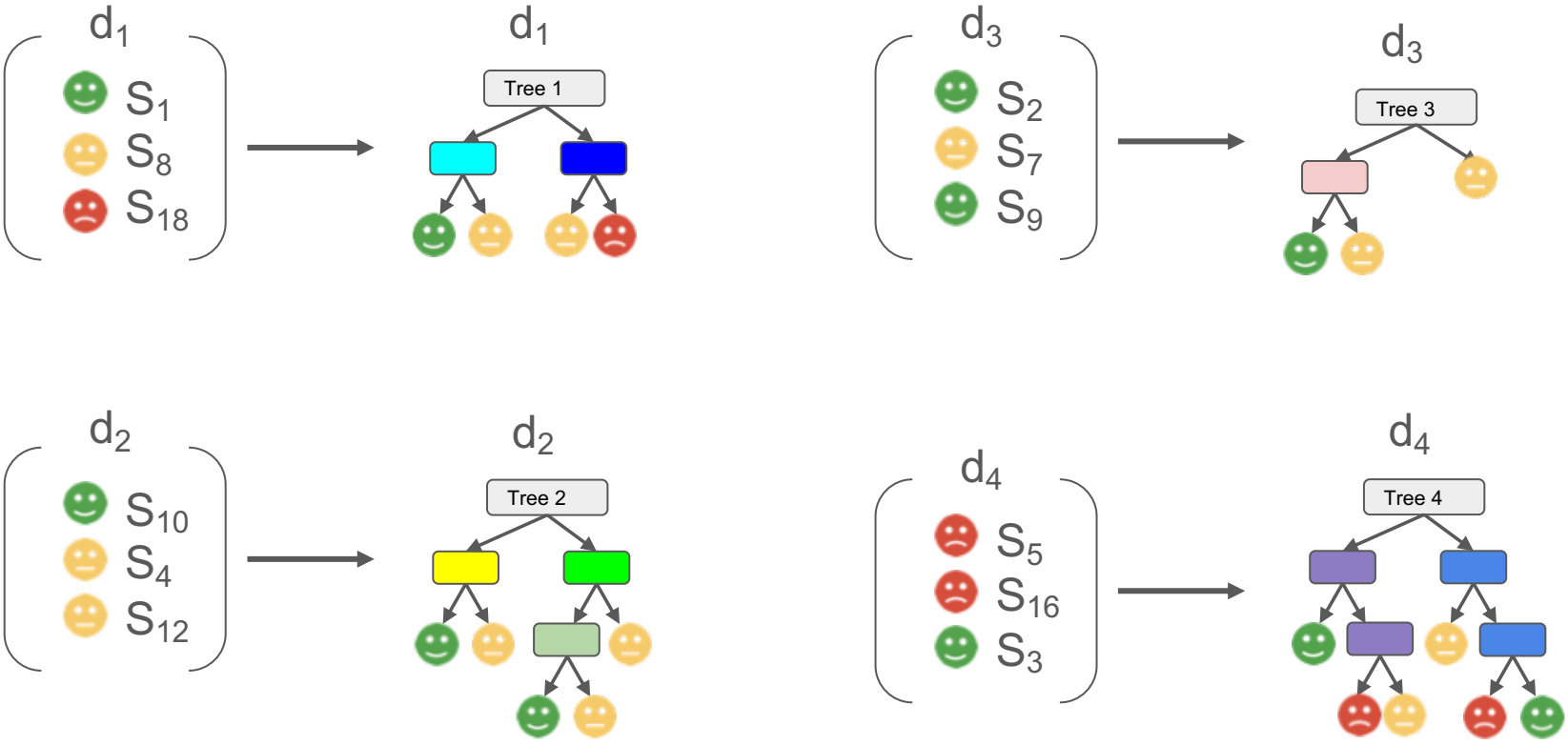
- Number of trees
- Number of features at each split

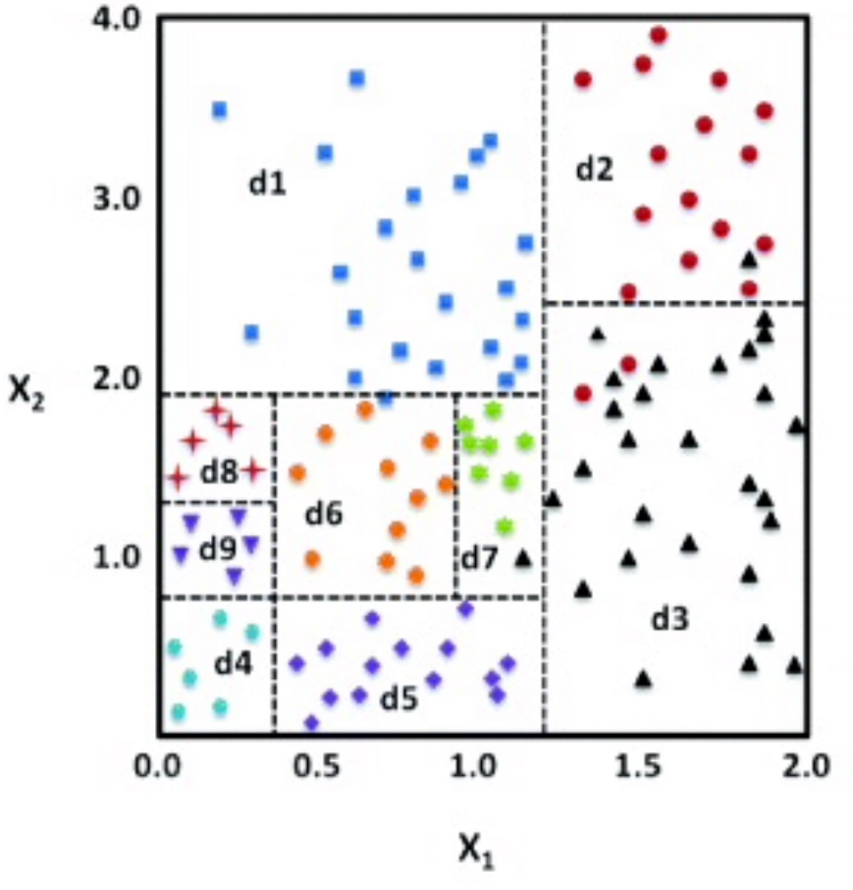


Random forest

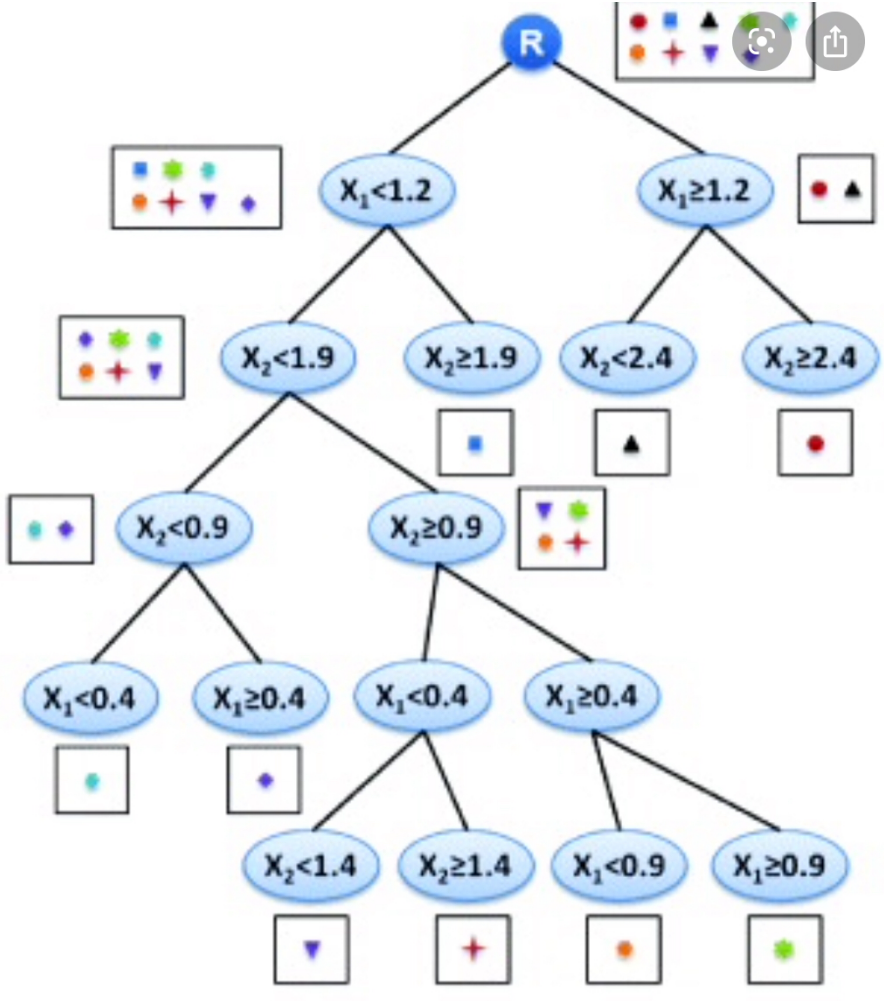


Random forest - training




4.0
3.0
2.0
1.0
0.0
0.5
1.0
1.5
2.0
X2
X1
d1
d2
d3
d4
d5
d6
d7
d8
d9

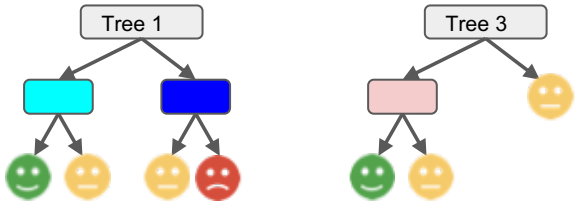
(a)

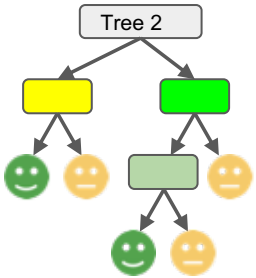

R
X1<1.2
X1≥1.2
X2<1.9
X2≥1.9
X2<2.4
X2≥2.4
X2<0.9
X2≥0.9
X1<0.4
X1≥0.4
X1<0.4
X1≥0.4
X2<1.4
X2≥1.4
X1<0.9
X1≥0.9

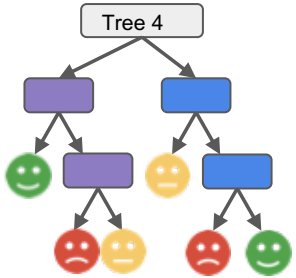
(b)

Random forest - predicting

$fA_p + fB_p + fC_p + fD_p + fE_p + fF_p$

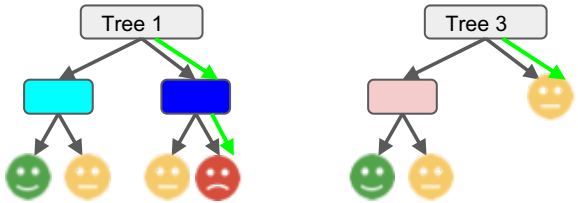


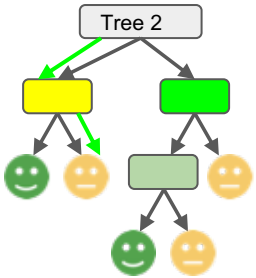


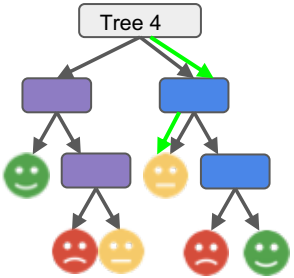


Random forest - predicting

$fA_p + fB_p + fC_p + fD_p + fE_p + fF_p$

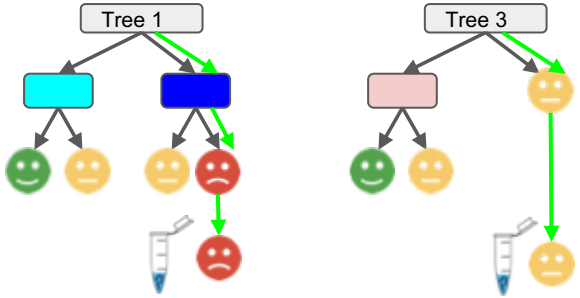


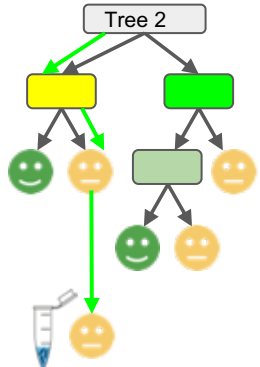


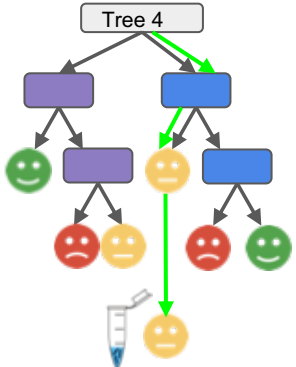


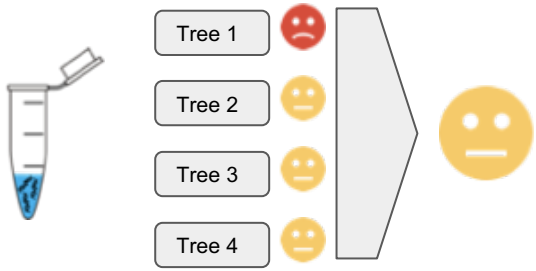
Random forest - predicting

$$fA_p + fB_p + fC_p + fD_p + fE_p + fF_p$$









Collecting a *good* training set

- Know your system
  - what is my reference? concentration of pollutant? Natural / anthropised? An index?
  - Does it make sense to make so much effort? Sometimes easier to measure the pollutant

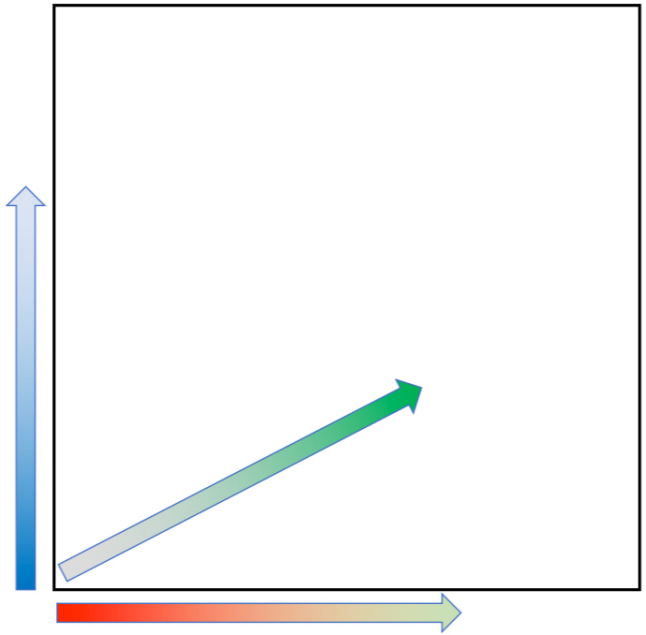
- Cover the gradient in a balanced effort and account for counfounding factors (season, landscape attributes, geography…)
  - Your sampling, molecular and metadata collection protocols cannot change afterwards!
  - Have backups! Samples, eDNA and data…

The SML compensates for the variations of benthic communities due to the biogeography, seasonality, or bottom features

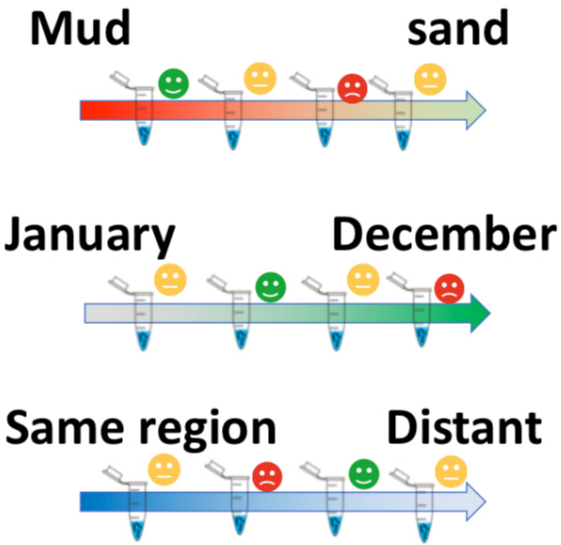
Mud sand

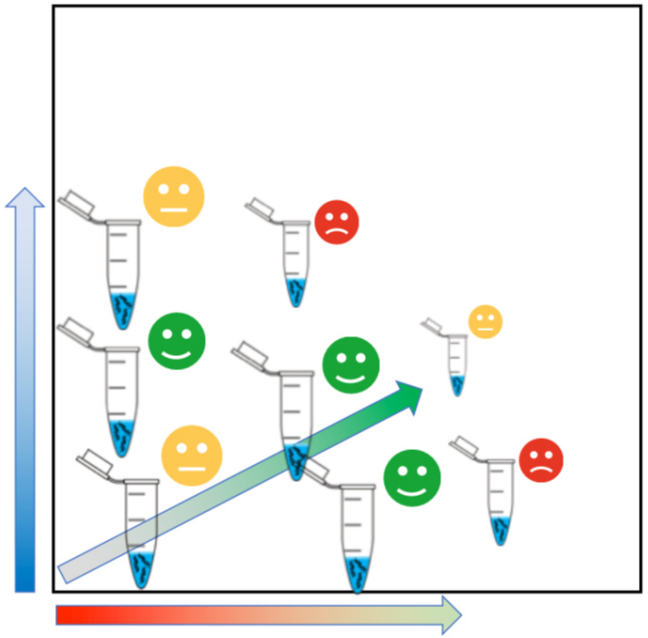
January December

Same region Distant

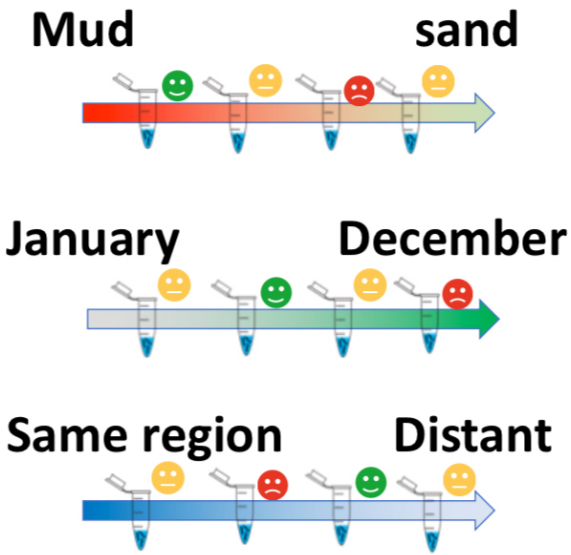


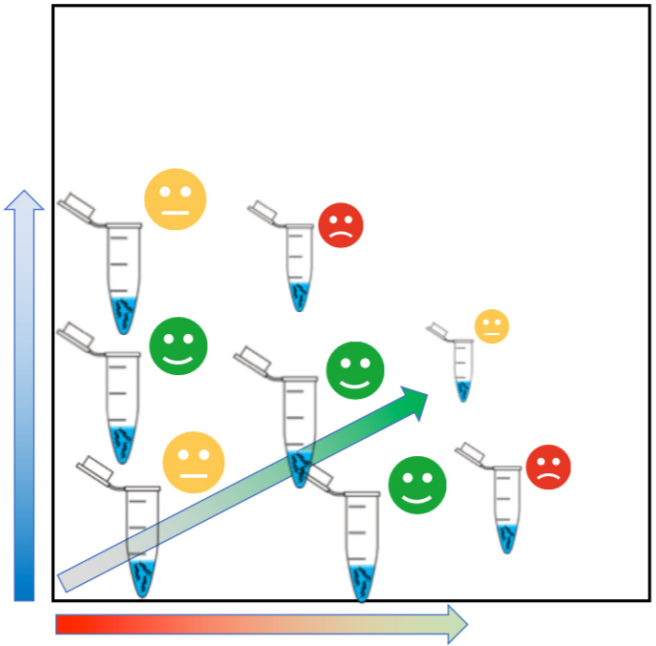
The SML compensates for the variations of benthic communities due to the biogeography, seasonality, or bottom features



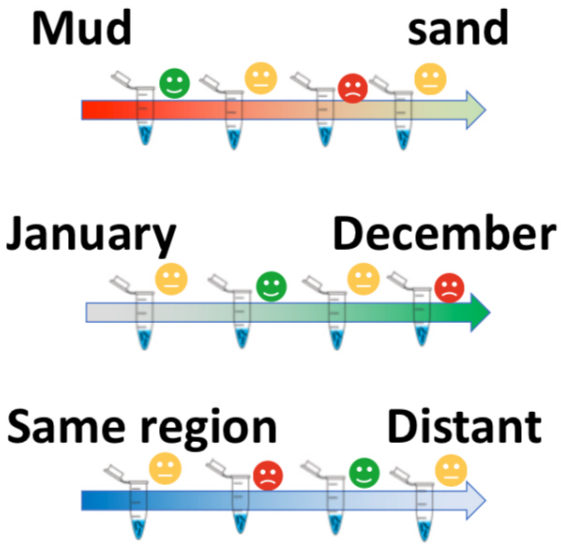


The SML compensates for the variations of benthic communities due to the biogeography, seasonality, or bottom features





However, the SML requires a good range of sampling covering different regions, bottom conditions (and seasons)



Available online at www.sciencedirect.com

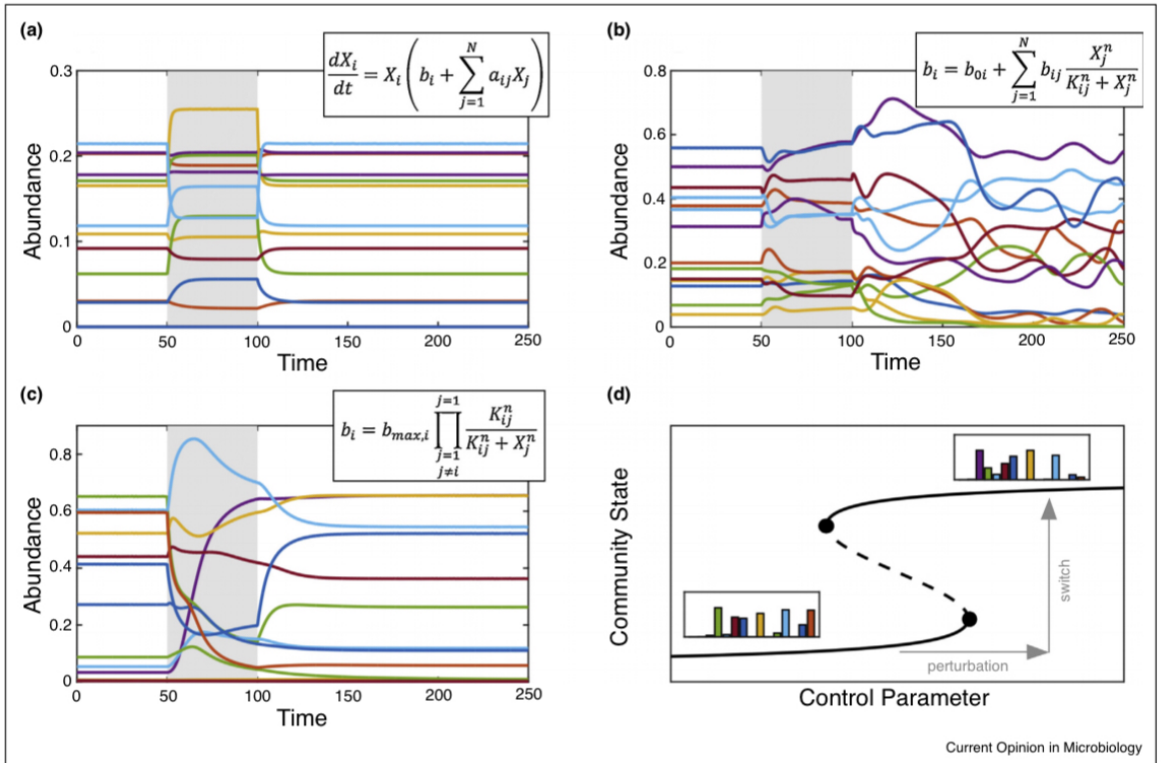
ScienceDirect

Current Opinion in Microbiology

Microbial communities as dynamical systems

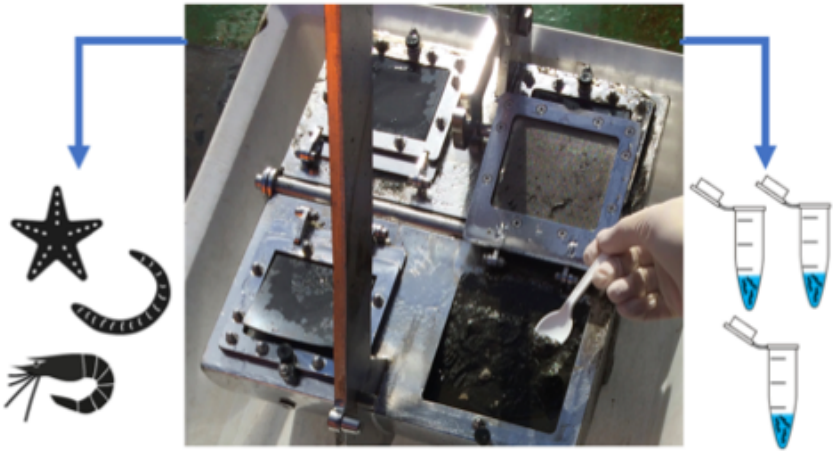
Didier Gonze[1,2], Katharine Z Coyte[3,4], Leo Lahti[5,6,7] and Karoline Faust[5]

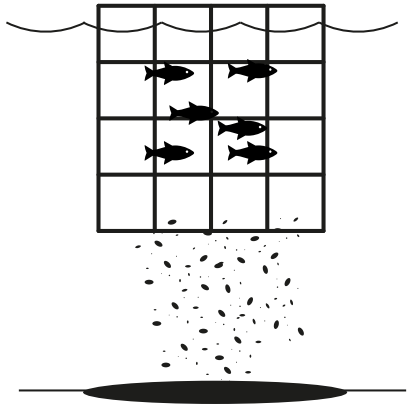
Figure 3



Testing the approach in a marine system

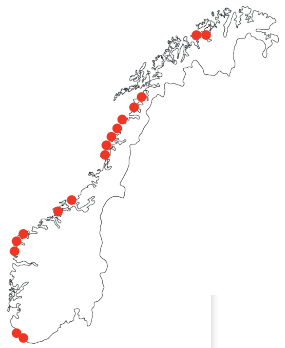
FHF FISKERI- OG HAVBRUKSNÆRINGENS FORSKNINGSFOND

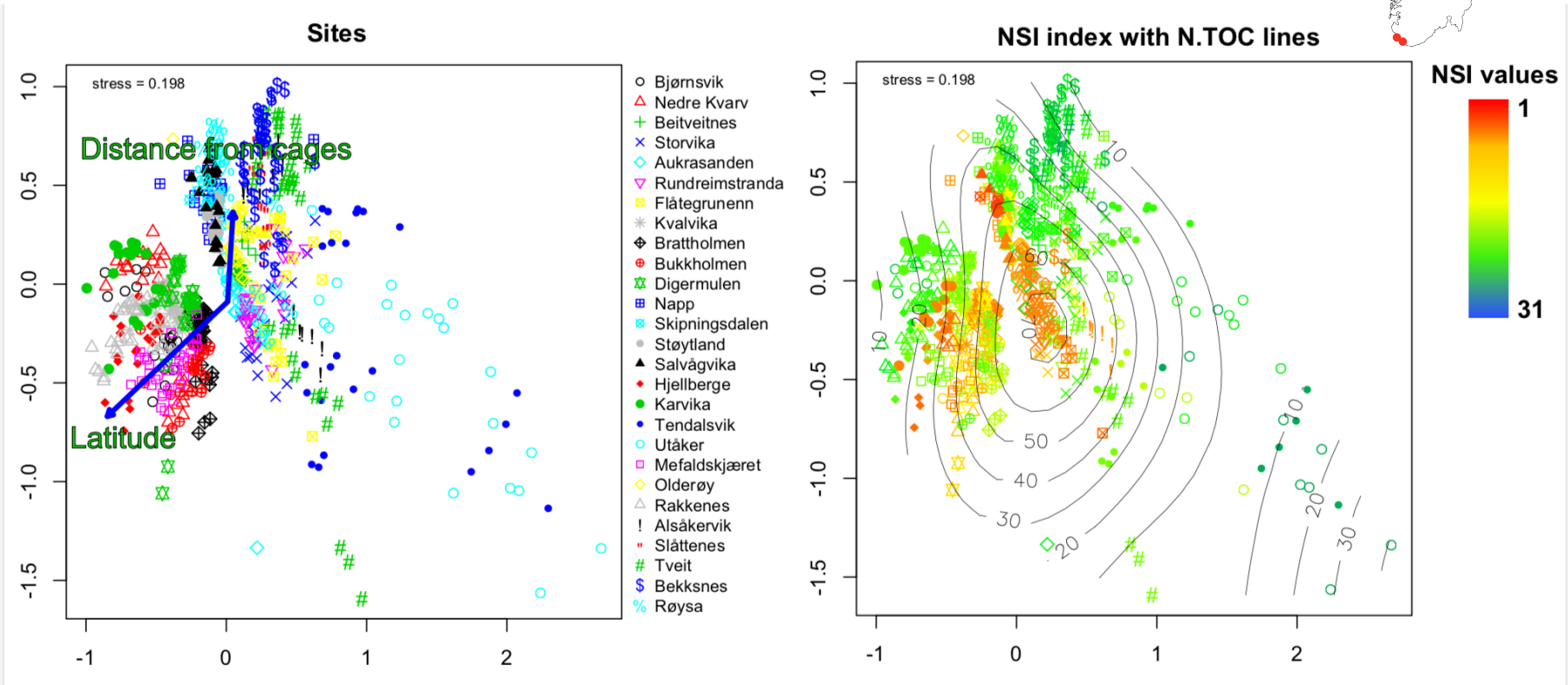




FARMED RESPONSIBLY asc CERTIFIED ASC-AQUA.ORG ™

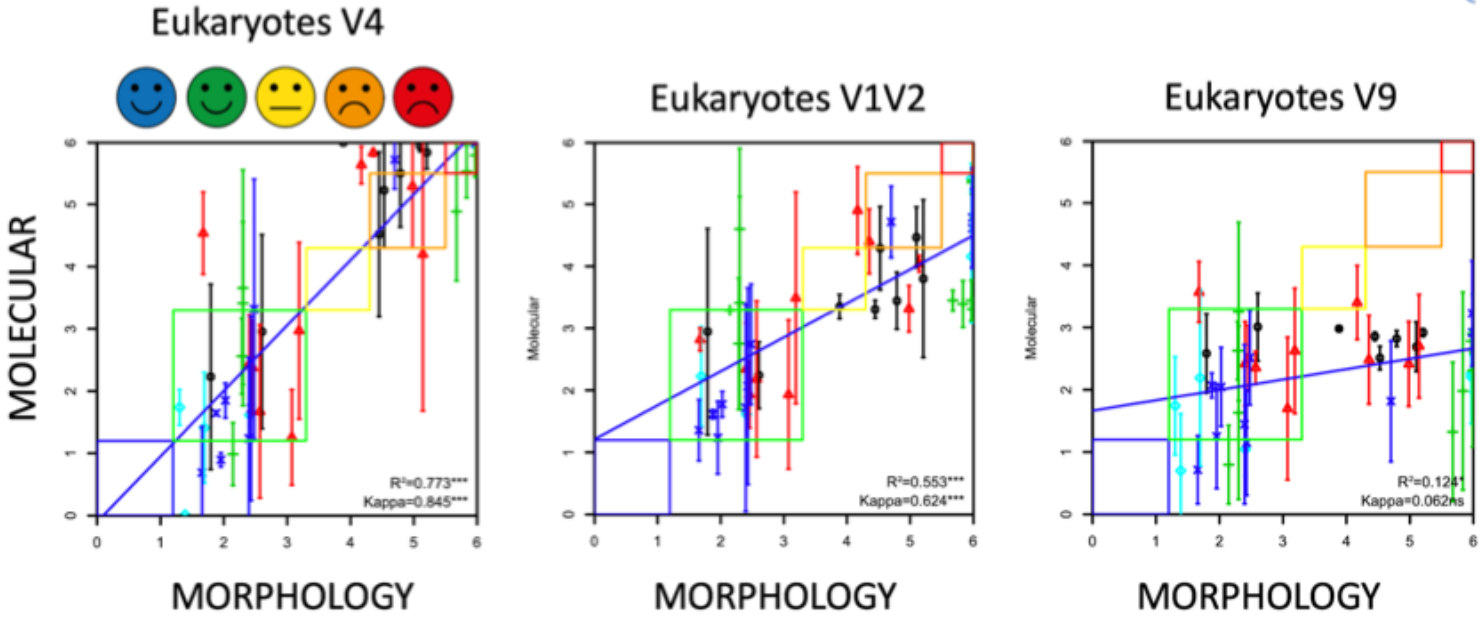
Controlling for other factors





Taxonomy-based approach

eDNA-taxonomy

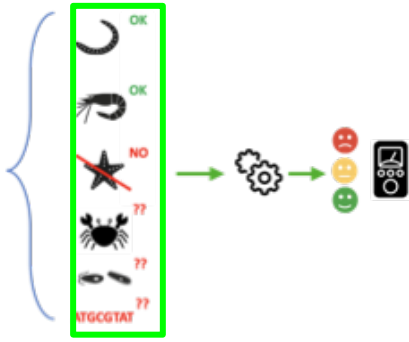


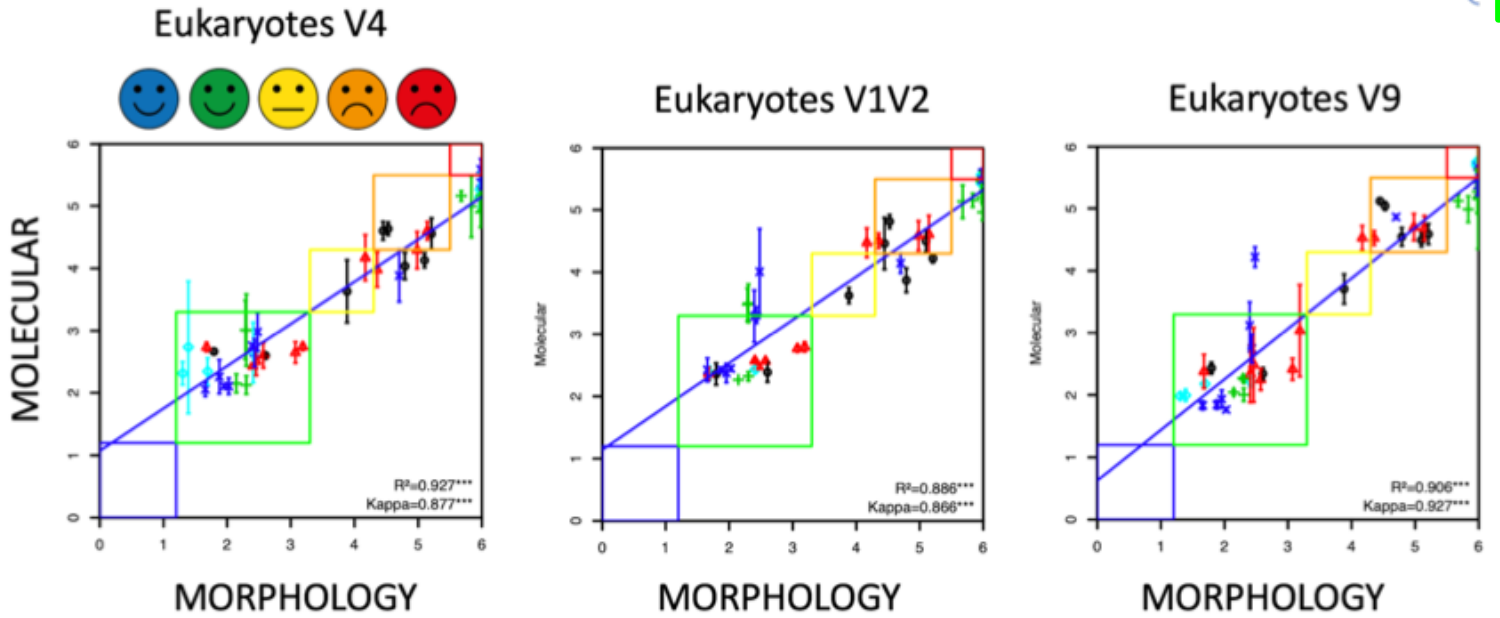
10-20% of the data

Cordier et al., 2018

De novo approach

eDNA-ML





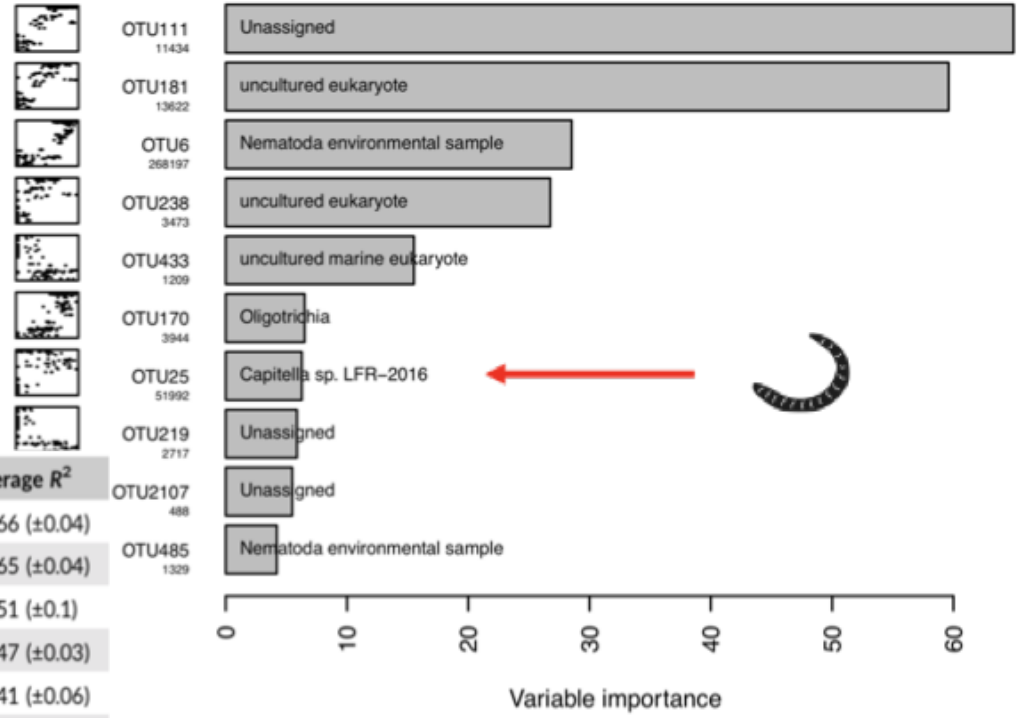
100% of the data

Cordier et al., 2018

Identifying new bioindicators

Best ones are microbial!

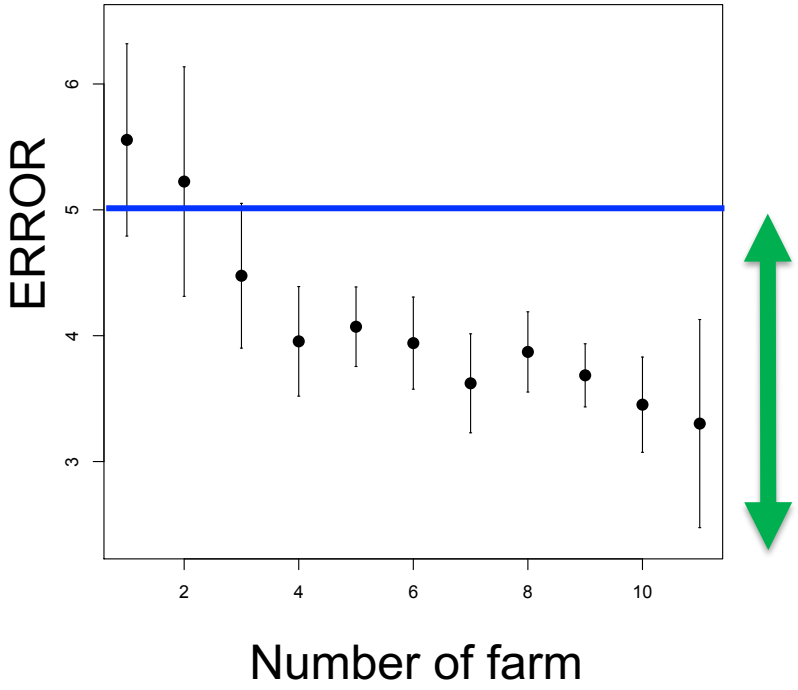
Marker	Taxonomic subgroups	OTUs	Percentage	Average R^2
Bacteria	Gemmatimonadetes	29	0.59	0.866 (±0.04)
Bacteria	Proteobacteria	1,439	40.82	0.865 (±0.04)
Bacteria	Planctomycetes	253	4.18	0.851 (±0.1)
Bacteria	Actinobacteria	170	8.86	0.847 (±0.03)
Bacteria	Verrucomicrobia	108	2.03	0.841 (±0.06)
Bacteria	Acidobacteria	180	4.8	0.822 (±0.07)
Eukaryotes V4	Apicomplexa	20	0.36	0.818 (±0.09)
Eukaryotes V1V2	Apicomplexa	28	0.27	0.78 (±0.11)
Metazoa V1V2	Nematoda	157	17.28	0.766 (±0.06)
Bacteria	Chlorobi	7	0.23	0.766 (±0.09)
Eukaryotes V9	Charophyta	13	0.15	0.718 (±0.09)
Bacteria	Nitrospirae	9	0.62	0.717 (±0.11)
Metazoa V9	Nematoda	18	6.16	0.694 (±0.12)

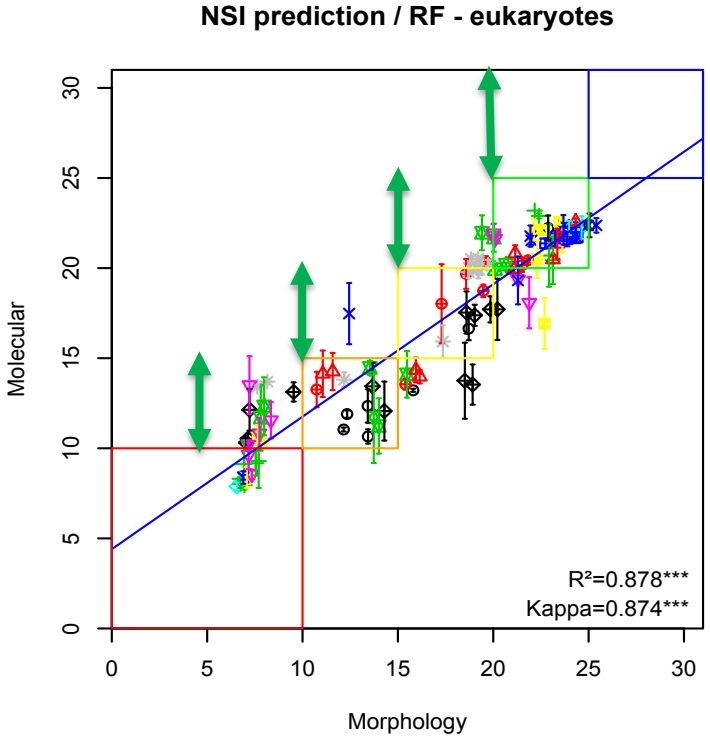


eDNA-based biomonitoring should focus on microbial communities

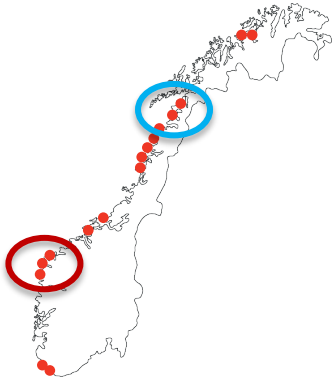
Cordier et al., 2018

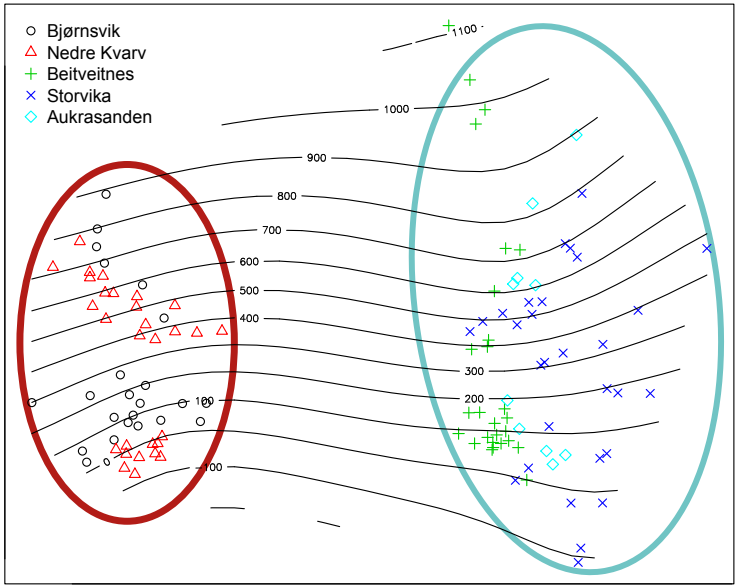
How many farms do we need ?

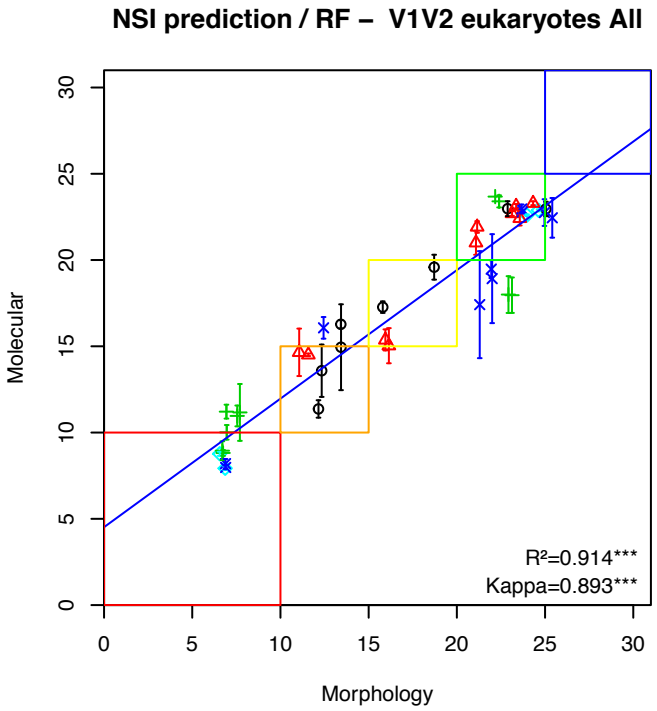




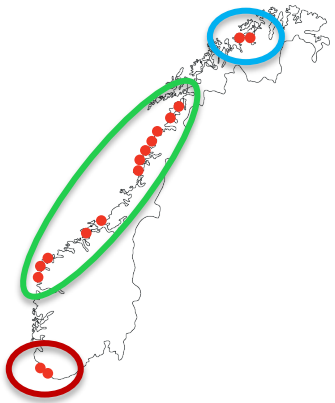
ML can compensates for geography…

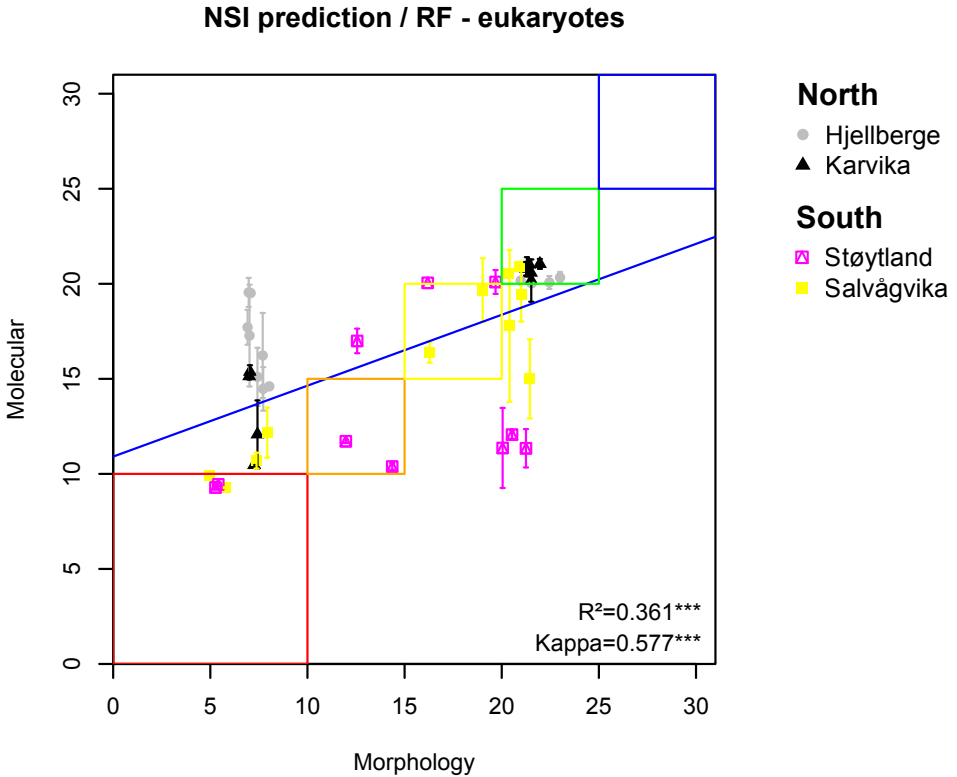


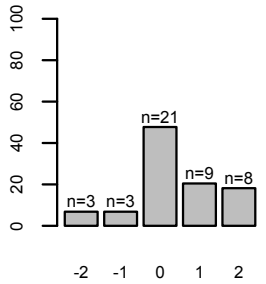


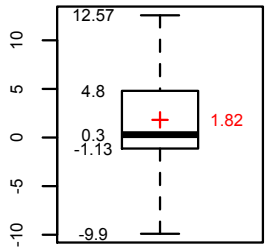


… to some extent

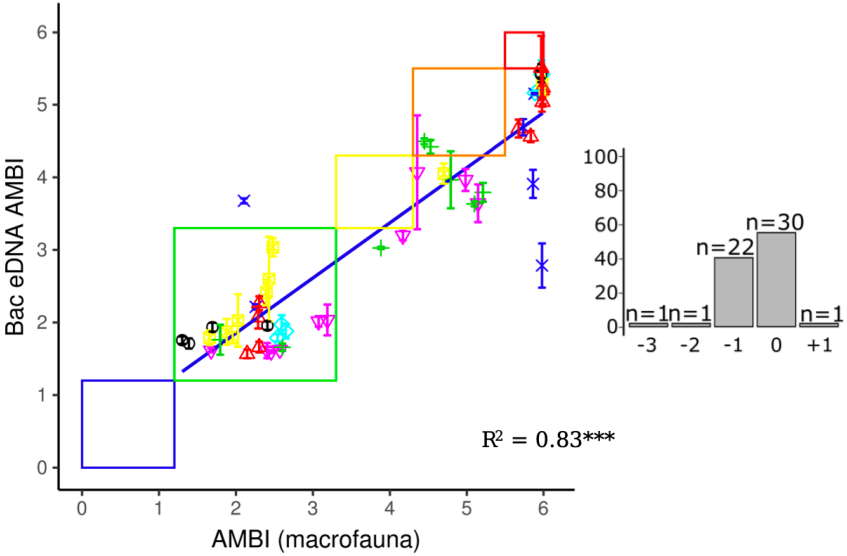


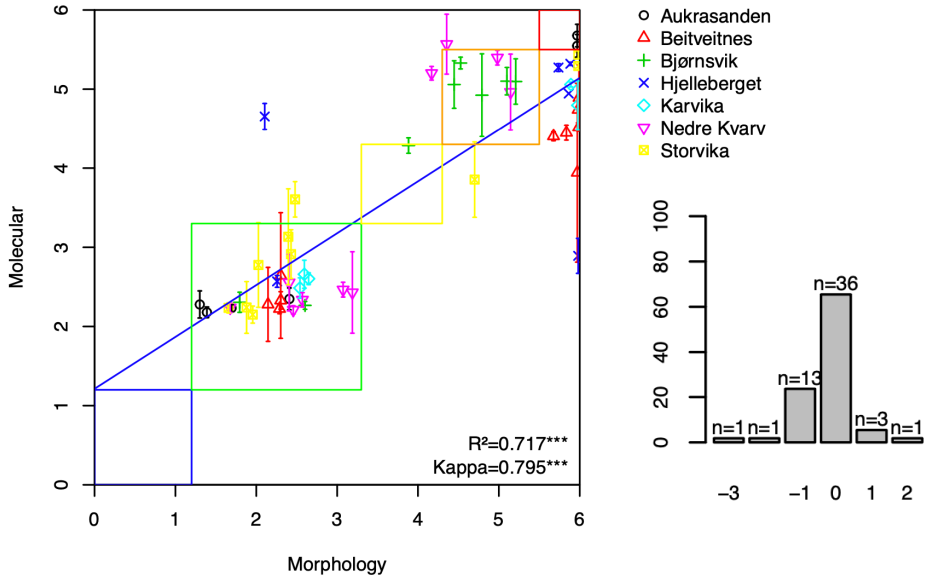


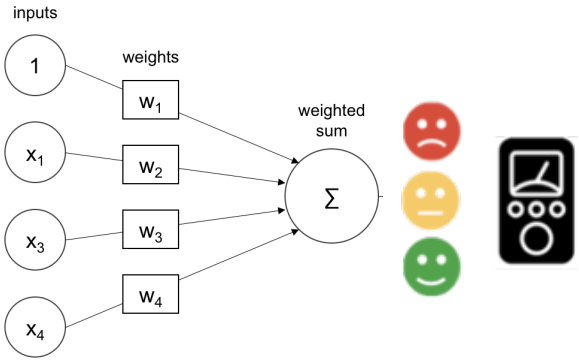




→ **Increasing the spatial coverage of the training set might resolve this**



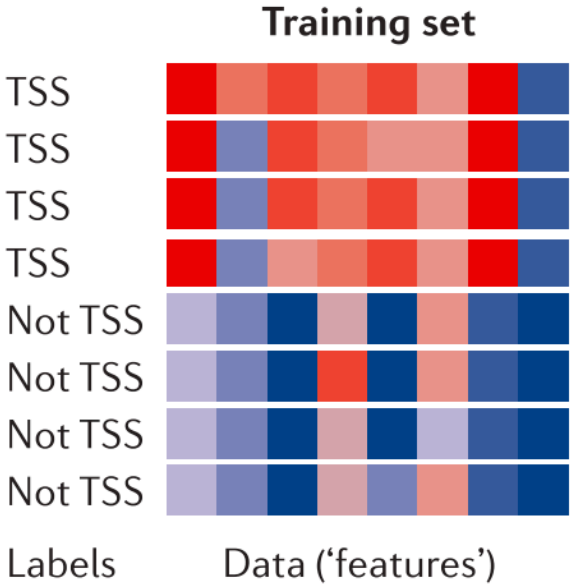




Frühe et al., *submit*.

OTU-based indices VS Supervised Machine Learning

OTU_ID	Weight	Taxa
OTU252	5	Gastrotricha 1
OTU343	2	Nematoda 1
OTU665	4	????
OTU292	3	Bacteria 1
OTU305	1	Nematoda 2
OTU152	2	????
OTU259	4	Bacteria 2
OTU149	2	Fungi 1
OTU1621	3	Fungi 2
OTU1382	2	Bacteria 3
OTU1648	5	???



+ Easily interpretable
+ Fit the current standards and regulations
— Less accurate
— One-by-one approach

+ More accurate
+ Can identify meaningfull co-occurrences
— Less easily intepretable
— Does not (yet) fit current standards

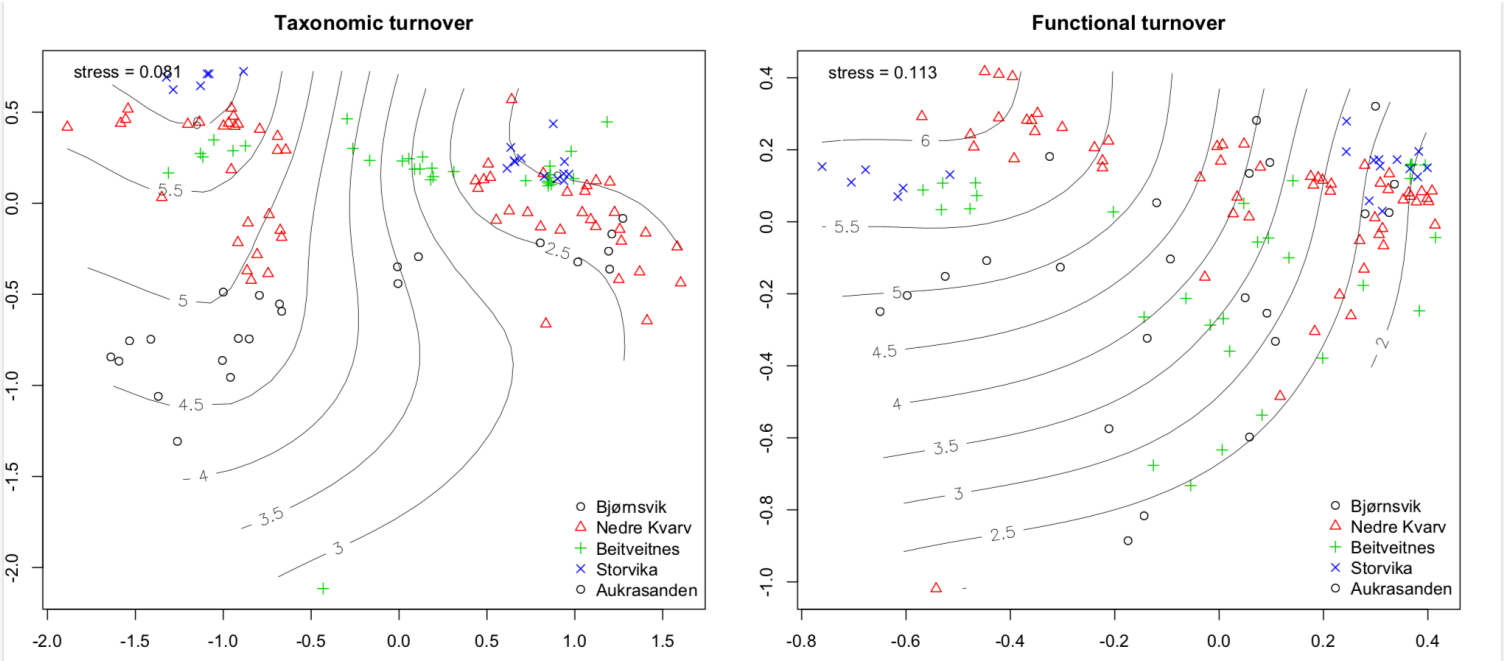
Who is there vs what they do?

Bioinformatics, 31(17), 2015, 2882–2884
doi: 10.1093/bioinformatics/btv287
Advance Access Publication Date: 7 May 2015
Applications Note

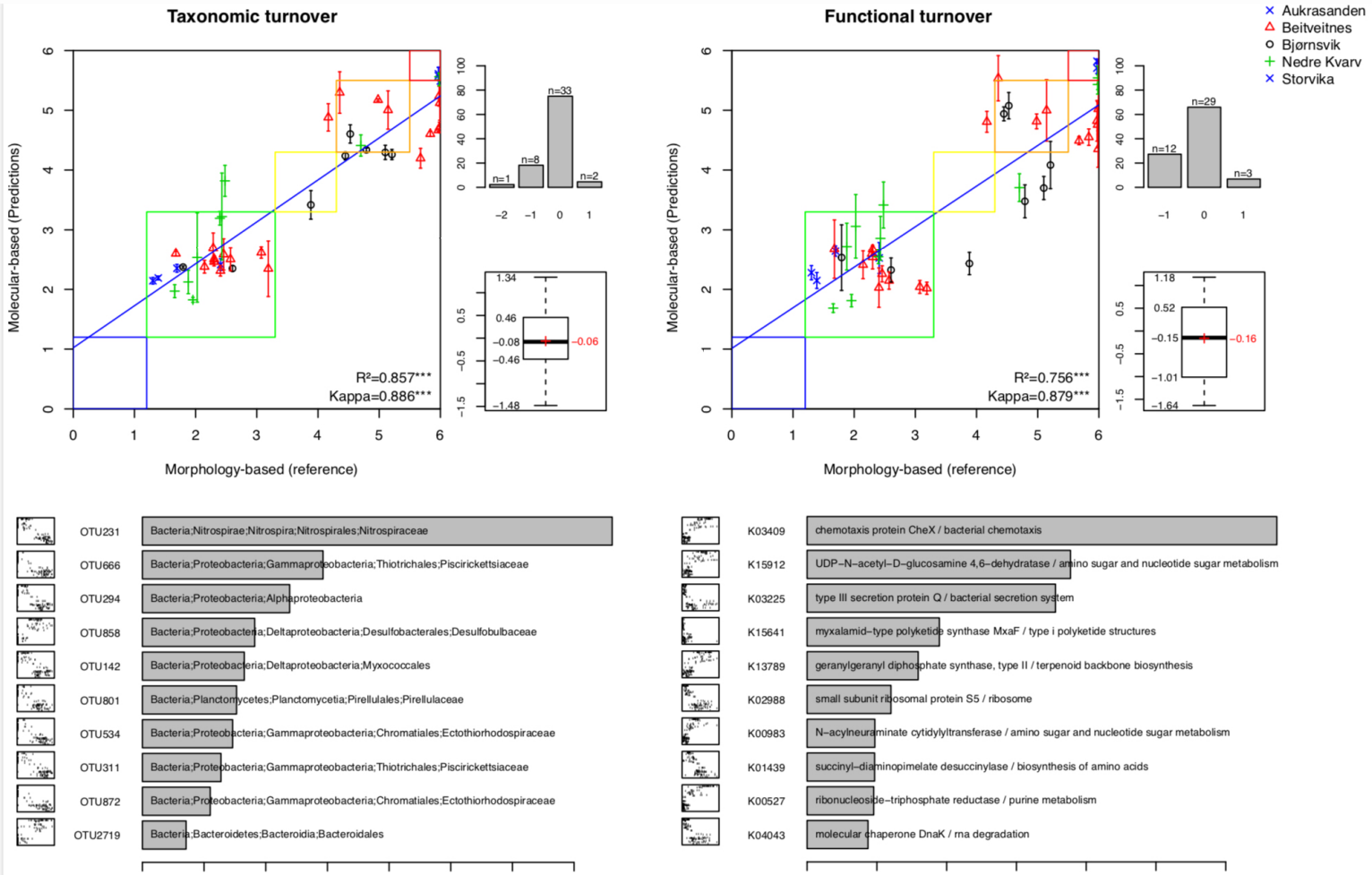
Sequence analysis

Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data

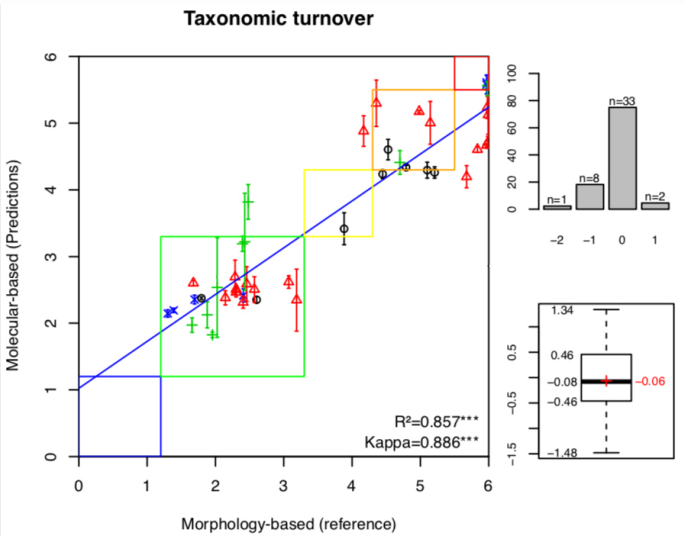
Kathrin P. Aßhauer[1,*], Bernd Wemheuer[2], Rolf Daniel[2] and Peter Meinicke[1]

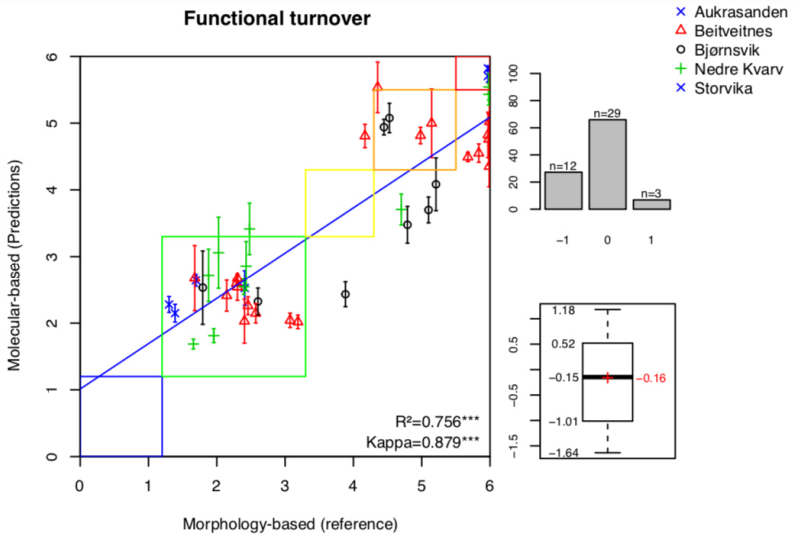


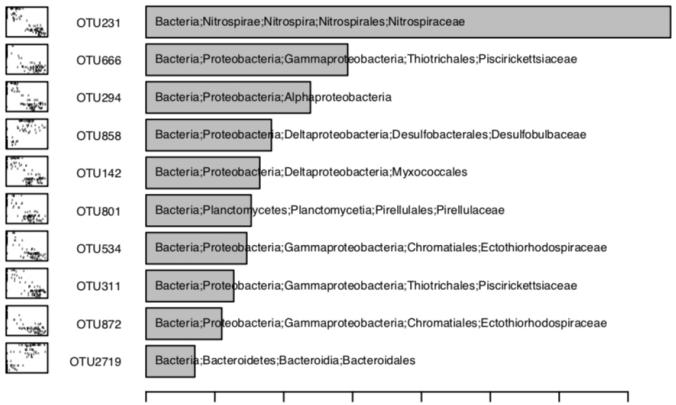
Cordier 2019, environmental DNA, in press

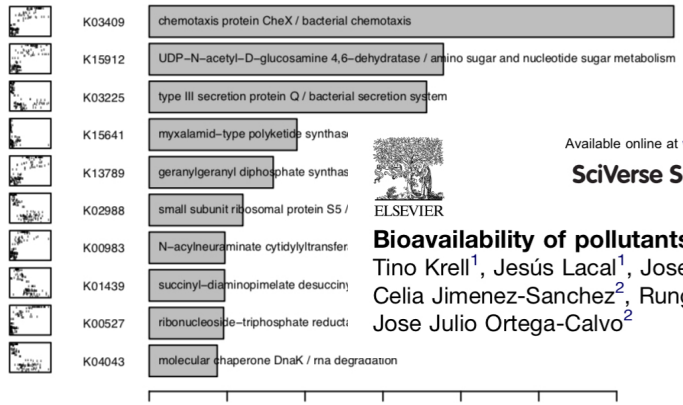


Cordier 2019, environmental DNA, in press









Available online at www.sciencedirect.com
SciVerse ScienceDirect
Current Opinion in Biotechnology

Bioavailability of pollutants and chemotaxis

Tino Krell[1], Jesús Lacal[1], Jose Antonio Reyes-Darias[1], Celia Jimenez-Sanchez[2], Rungroch Sungthong[2] and Jose Julio Ortega-Calvo[2]

Cordier 2019, environmental DNA, in press

Untitled.R | Untitled6* | Untitled4* | Untitled5* | dada2_table_FHF.R | metaB_tutorial.R*

Source on Save | Run | Source

```
1  ##########################################################################################################
2
3  ### 9th Metabarcoding school - Colombia
4  ### AI/ML analysis of metabarcoding data for biomonitoring - 1st-6th November 2019
5  # Tristan Cordier
6
7  ########################################## Setting the stage ##############################################
8
9  # setting the working directory
10 #setwd("/path/to/the/folder/metaB_ml")
11 setwd("~/Desktop/POGO_ML_workshop_genomics/Metabarcoding school/")
12
13 # libraries we need
14 library('vegan')
15 library('matrixStats')
16 library('ranger')
17 library('irr')
18 ## for parralel CPU computation
19 library('doMC')
20 registerDoMC(cores = detectCores())
21
22
```

68:10 'Playing' and manipulating the data | R Script

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
met	629 obs. of 15 variables
otu	629 obs. of 4903 variables
otu_raw	629 obs. of 4903 variables
taxo	Large matrix (9806 elements, 662.9 Kb)
Values	
row_S	Named num [1:629] 140042 150430 74820 42 6…
s	"samples_names"
seq_depth_cuto…	10000
Functions	
bar_plot	function (otu_table, comp_, aggreg = F, …
plot_ml	Large function (989.1 Kb)
sml_compo	Large function (785.3 Kb)

Console | Jobs

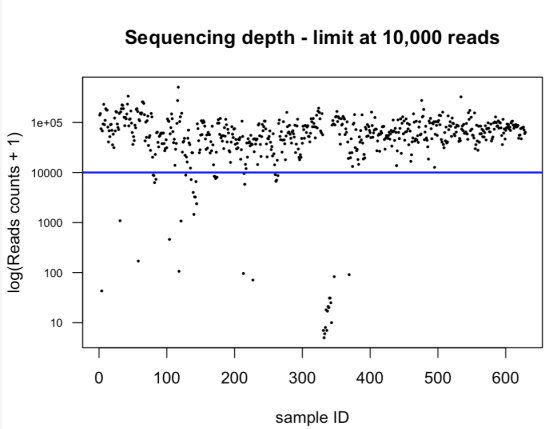
~/Desktop/POGO_ML_workshop_genomics/Metabarcoding school/

```
> # and do they match exactly?
> table(gsub("-",".",met[,s]) == rownames(otu))

TRUE
 629
>
> ### keep the dataset at this point for coming back to it
> otu_raw <- otu
>
> ### sum of samples
> row_S <- rowSums(otu_raw)
> plot(row_S, main="Sequencing depth - limit at 10,000 reads", ylab = "Reads counts", xlab = "sample ID",
 pch=16, cex=.4)
> abline(h=10000, col="blue", lwd=2)
> plot(log(row_S+1), main="Sequencing depth - limit at 10,000 reads", ylab = "log(Reads counts + 1)", xlab
 = "sample ID", yaxt="n", pch=16, cex=.4)
> axis(2, at=log(c(10,100,1000,10000,100000,1000000)),labels=c(10,100,1000,10000,100000,1000000),cex.axis=
0.7, las=2, ylim=log(c(10,1000000)))
> abline(h=log(10000+1), col="blue", lwd=2)
>
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



swiss network for international studies

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

FHF
FISKERI- OG HAVBRUKSNÆRINGENS FORSKNINGSFOND

UNIVERSITÉ DE GENÈVE

Department of Genetics and Evolution
Molecular Systematics and Environmental Genomics
•Jan Pawlowski
•Maria Holzmann
•Laure Apothéloz Perret Gentil
•Emanuella Reo
•Florian Mauffrey
•Kristina Cermacova
•Sofia Caetano Wyler
•Léo Charvoz
•Jade Boutten
•Brian Bourrat

University of Kaiserslautern, Germany
Ecology group
•Thorsten Stoeck
•Dominik Forster
•Larissa Frühe

AZTI-tecnalia, Spain
Coastal monitoring group
•Anders Lanzén
•Angel Borja
•Naiara Rodríguez-Ezpeleta