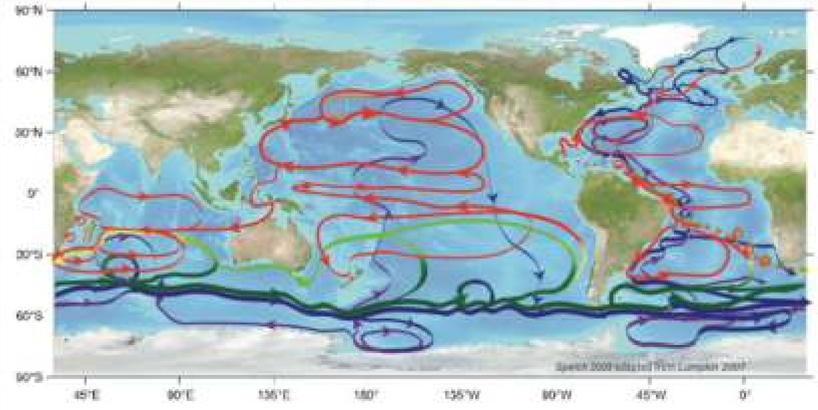
The Good, the Bad and the Ugly of unsupervised learning

Maike Sonnewald[1,2]

July, 2019

[1]MIT & [2]Harvard

CO_2 release is the fastest in the past 55 million years...

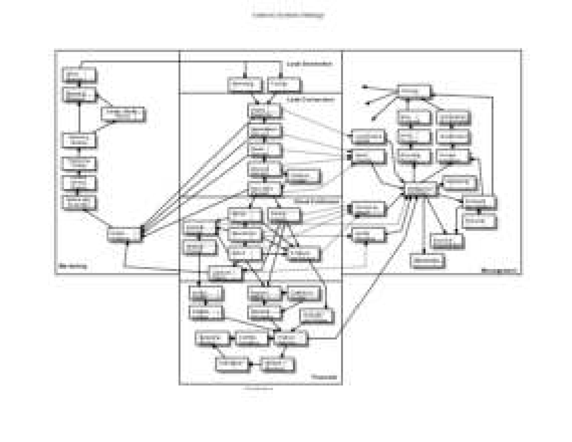


Movie by NASA

Take home:

Application of modern statistics/ML and traditional modeling/theory work can drive progress

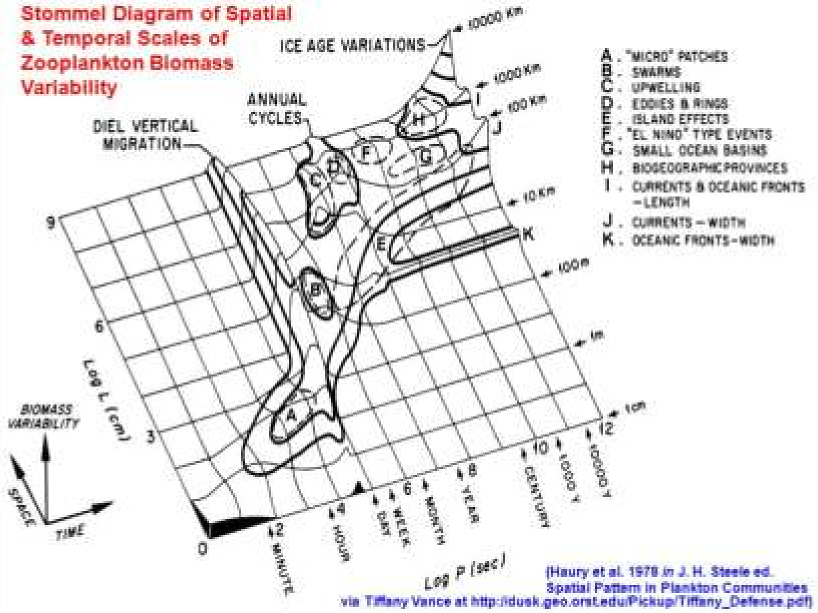
Currently models need to be complicated



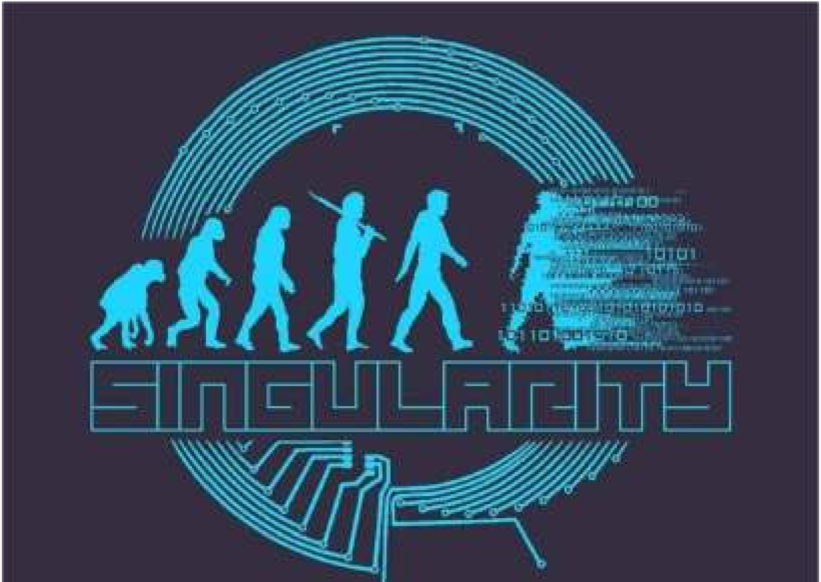
slideshare.net

Including **all** components of e.g.
an ESM at adequate resolution is unfeasible.

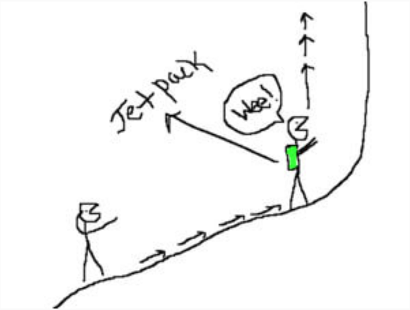
How complicated is complicated enough?



Increased computational power?

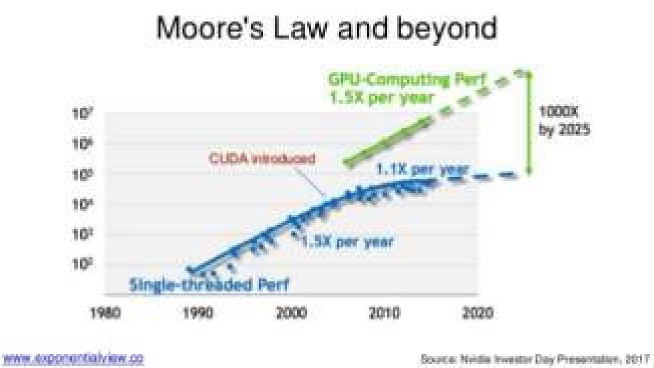


Increased computational power?



Moore's Law: Jet-Pack!

Increased computational power?



Moore Vs Rock: Nano materials/quantum computing to the resque..?

Less is more

eos.org

The dream:

ML can be used to lead the charge both in terms of supervised and unsupervised learning.

Can ML help?

- Make complicated data complex
- Allows insight to parameterize and simplify
- Create synergy between models, theory and observations

ML towards the goal of science/geoscience:
Precise and accurate understanding of the natural world.

Motivation: Why machine learning?

"*...automate those parts that can be perfectly automated*"

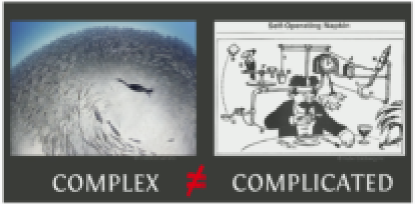
Occam's razor in Machine Learning

Motivation: Why machine learning?

"...*automate those parts that can be perfectly automated*"

Occam's razor in Machine Learning

- Modern data science is greatly increasing the efficiency of conventional research
- Find patterns to accelerate exploration of physics
- Highlight emergence of complex interactions



Complexity: Find underlying "rules"

"Complexity: *I know it when I see it...*"

Complexity: Find underlying "rules"

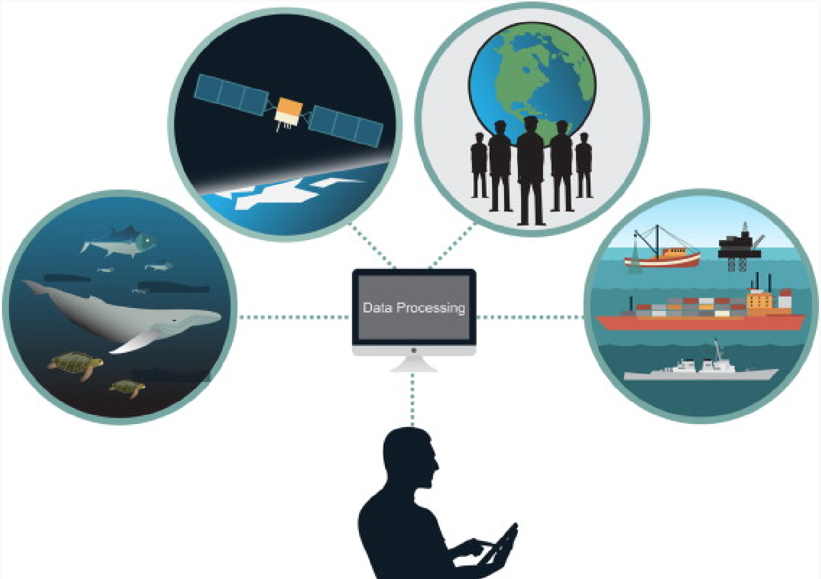
"Complexity: *I know it when I see it...*"

BOIDS Craig Reynolds (1986)

- separation: steer to avoid crowding local flockmates
- alignment: steer towards the average heading of local flockmates
- cohesion: steer to move towards the average position (center of mass) of local flockmates

We are becoming data rich in oceanography

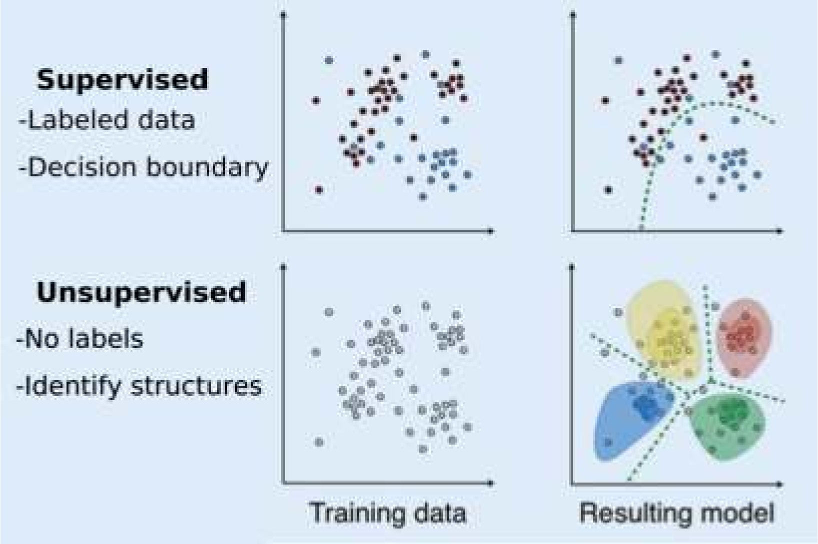
Maxwell 2015, Kim Martini and Sonnewald et al 2013



We are becoming data rich in oceanography

Maxwell 2015, Kim Martini and Sonnewald et al 2013

Unsupervised Learning: Find clusters



Langs et al. 2018

The persuasive power of numbers

"*There are three kinds of lies: lies, damned lies, and statistics.*"
Benjamin Disraeli (British prime minister)

- Bad statistics can bolster weak arguments
- Weak results can seem legitimate.
- False positives are bad.

Lies damned lies and ..ML?

Do:

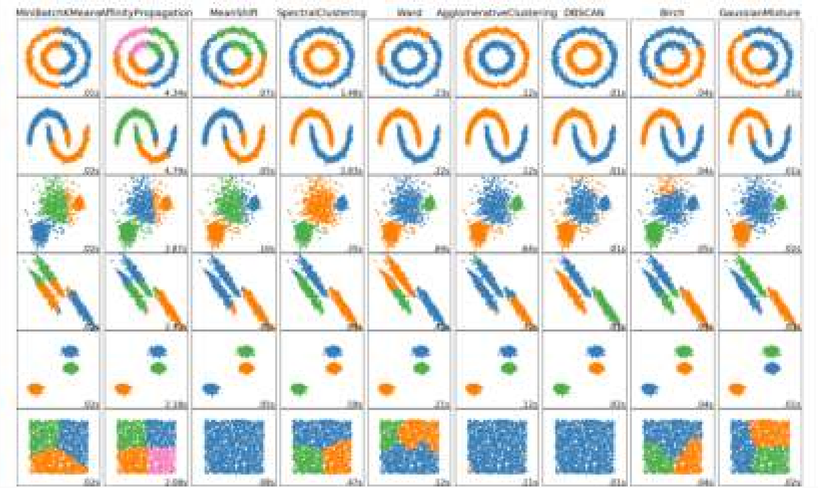
Clean data, visualize data, highlight interactions as appropriate, choose an appropriate model, check its parameters are appropriate, account for stochastic elements.

Don't:

Trust data blindly, make assume variance topography, choose model blindly, if so: Brute force the statistical robustness.

Machine Learning workflow: What model?

Keep It Simple Stupid

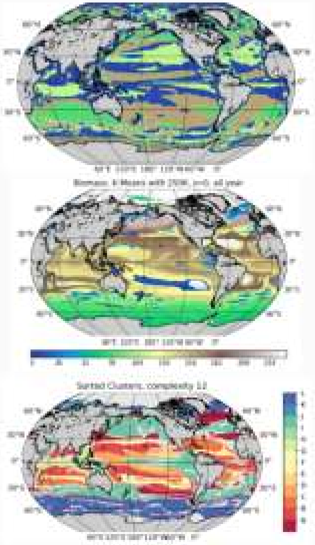


scikit-learn.org

Libraries e.g. scikit learn, dask_ml

Examples outline

- 1: The good ...Global dynamical regimes
    - Insight into global dynamical regimes!
    - Keeping it simple
- 2: The bad ... Global eco-provinces
    - Finding patterns in complicated data?
    - Avoiding false positives
- 3: The robustly ugly ... Global eco-provinces
    - Complex solution to complicated data
    - Interdisciplinarity to the rescue...



1: The Good …Global dynamical regimes

…Global dynamical regimes

Exploration - Uncover new physical phenomena?

- Identify driving features of key currents

- Analyze Big Data for climate fast

e.g. Climate Model Inter-comparison Project

Model Development - Establish largely laminar regions?

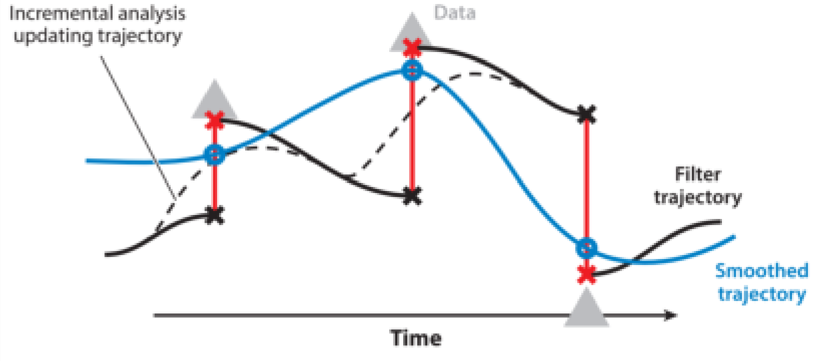
$\rightarrow$ Save computational cost

- Make inferences of key transitions

$\rightarrow$ e.g. baroclinicity

- Focus parameterization

Estimating the Circulation and Climate of the Ocean: 1992-2013



ECCO provides ocean state consistent with known physics and observations

Keeping it simple: Barotropic Vorticity equation

Momentum equations:

$$\partial_t \mathbf{u} + f\mathbf{k} \times \mathbf{u} = -\frac{1}{\rho_0}\nabla p + \frac{1}{\rho_0}\partial_z \tau + \mathbf{a} + \mathbf{b}, \partial_z p = -g\rho, \nabla \cdot \mathbf{v} = 0.$$

-Depth integrate, take curl

Barotropic Vorticity:

$$0 = \overbrace{\nabla \cdot (f\mathbf{U})}^{\text{Advection}} - \underbrace{\nabla \times (p_b \nabla H)}_{\text{Bottom Pressure Troque}} + \overbrace{\nabla \times \tau}^{\text{Wind and Bottom stress}} - \underbrace{\nabla \times \mathbf{A}}_{\text{Non-linear Torque}} + \overbrace{\nabla \times \mathbf{B}}^{\text{Lat. Visc.}}$$

e.g. Sverdrup balance:

Wind stress curl and advection balance locally

$$\nabla \cdot (f\mathbf{U}) = \nabla \times \tau$$

Keeping it simple: Barotropic Vorticity equation

Momentum equations:

$$\partial_t \mathbf{u} + f\mathbf{k} \times \mathbf{u} = -\frac{1}{\rho_0}\nabla p + \frac{1}{\rho_0}\partial_z \tau + \mathbf{a} + \mathbf{b}, \partial_z p = -g\rho, \nabla \cdot \mathbf{v} = 0.$$

-Depth integrate, take curl

Barotropic Vorticity:

$$0 = \overbrace{\nabla \cdot (f\mathbf{U})}^{\text{Advection}} - \underbrace{\nabla \times (p_b \nabla H)}_{\text{Bottom Pressure Troque}} + \overbrace{\nabla \times \tau}^{\text{Wind and Bottom stress}} - \underbrace{\nabla \times \mathbf{A}}_{\text{Non-linear Torque}} + \overbrace{\nabla \times \mathbf{B}}^{\text{Lat. Visc.}}$$

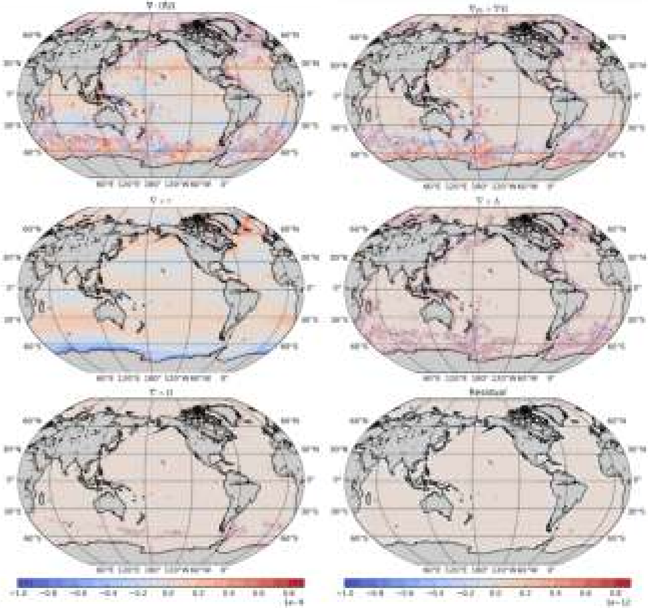
e.g. Sverdrup balance:

Wind stress curl and advection balance locally

$$\nabla \cdot (f\mathbf{U}) = \nabla \times \tau$$

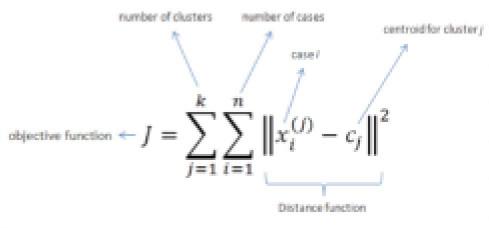
..Is this real? Is every location unique?

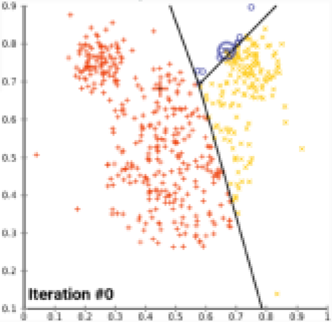
The Complicated data...



Unsupervised learning: K-Means clustering

What do we do when we don't have the answers a priori?





commons.wikimedia.org

- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Itterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to: K and initialization.

Unsupervised learning: K-Means clustering

What do we do when we don't have the answers a priori?

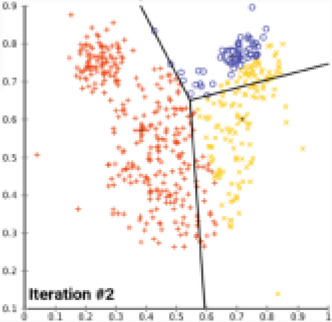
number of clusters

number of cases

case i

centroid for cluster j

objective function $\leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$

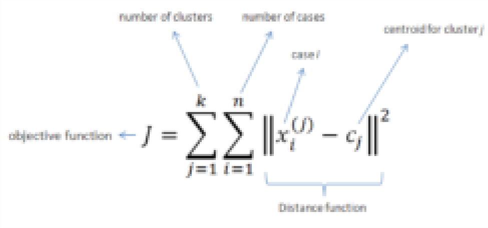


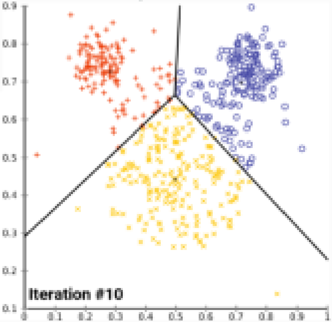
commons.wikimedia.org

- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Itterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:
K and initialization.

Unsupervised learning: K-Means clustering

What do we do when we don't have the answers a priori?





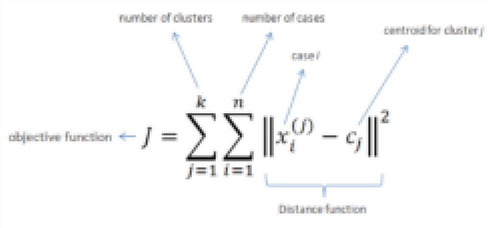
commons.wikimedia.org

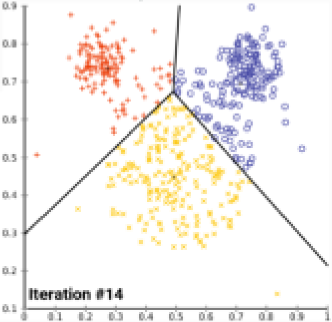
- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Itterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:

K and initialization.

Unsupervised learning: K-Means clustering

What do we do when we don't have the answers a priori?



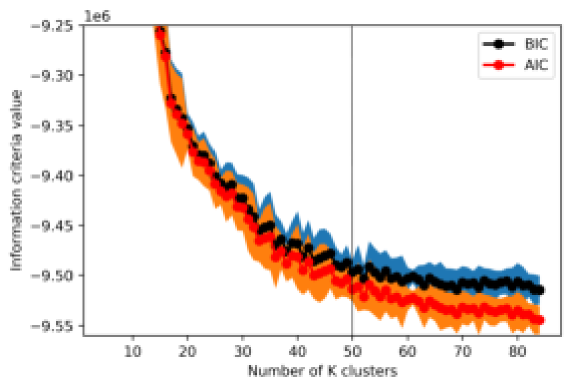


commons.wikimedia.org

- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Itterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:

K and initialization.

Choosing K: Information criteria

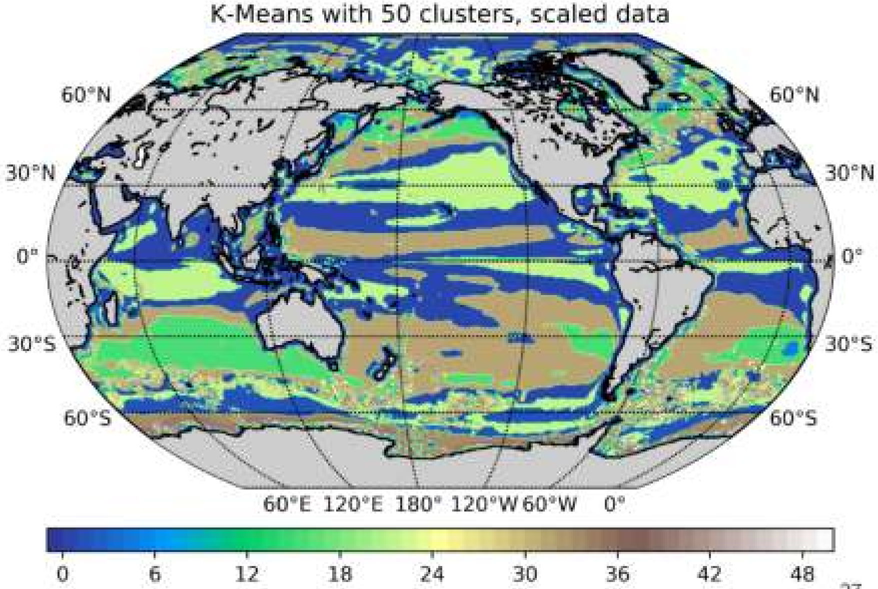


$$\mathrm{BIC} = K \ln(n) - 2 \ln(\mathcal{L}),$$

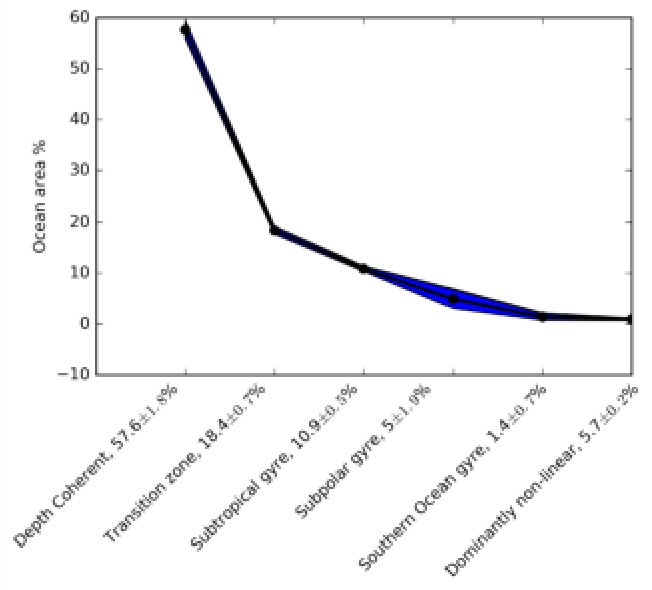
where n is the number of datapoints and $\mathcal{L}$ is the likelihood:

$$\mathcal{L} = \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\zeta_i - \hat{\zeta}_i)^2}{2\sigma^2}\right).$$

Dynamical regions

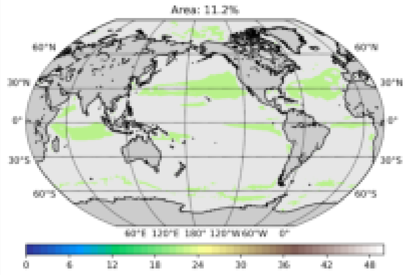


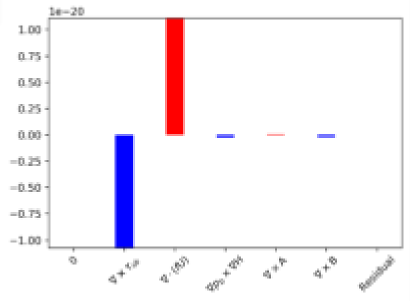
Percentage of area accounted for



Sverdrup balance: Subtropical Gyre

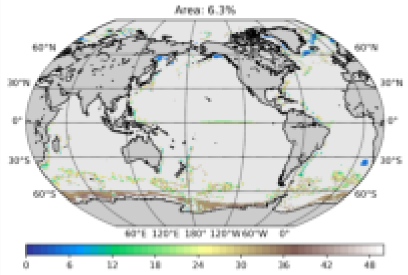
11.2%

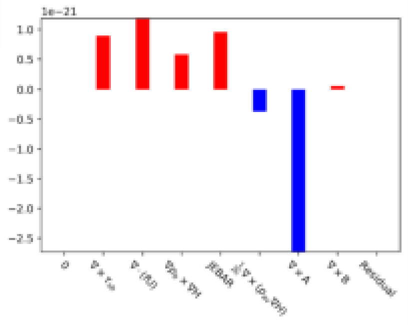




Present globally!

Dominantly non-linear
6.3%

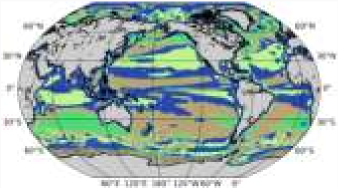




Only here does the linearity approximation break down.

Ocean globally organizes into distinct 6 regimes
→ Use this to analyze Big Data **fast**

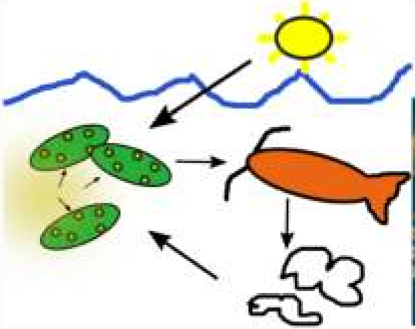
- Objectively describe 3D ocean using k-means
- Regions comply with theory
- Use IC and check for degeneracy

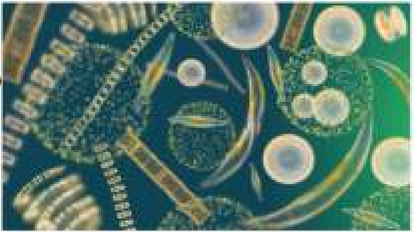


Sonnewald, Wunsch and Heimbach (2019) and **Sonnewald et al. (Nature Com. resubmission)**

2: The Bad …Global ecological provinces

Ocean ecology





It's complicated

2: The Bad …Global ecological provinces

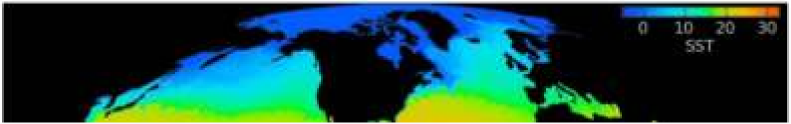
SIMONS FOUNDATION

PHYSICAL/BIOGEOCHEMICAL/ECOSYSTEM MODEL

Physics:
e.g. velocity, mixing, temperature

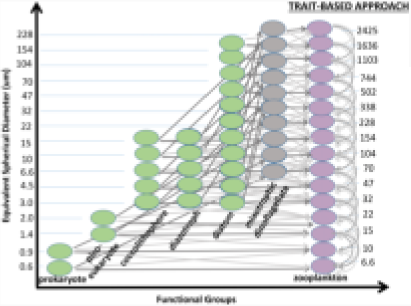
Biogeochemistry:
e.g. carbon, nutrients, DOM, detritus

Ecosystem:
e.g. phytoplankton (C, Chl), zooplankton



Dutkiewicz et al., 2015

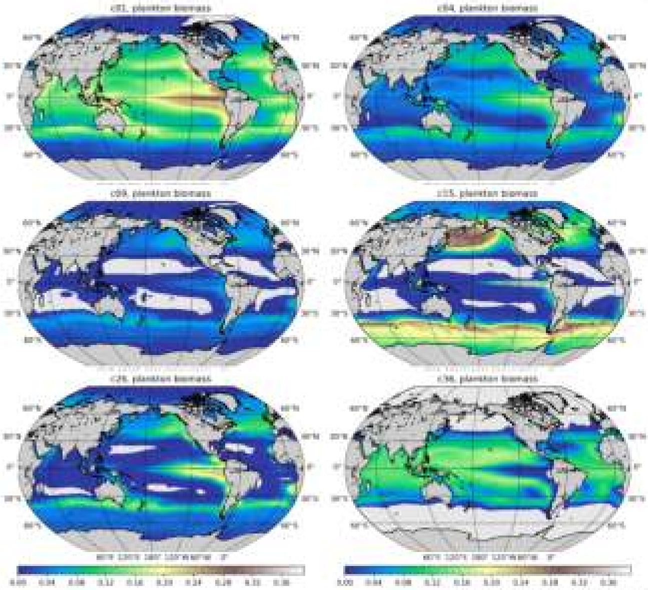
Biogoechemical ecosystem model: Empirical modeling approach



Trait	Size-based	Functionality-based
Maximum growth rate	Yes	Yes
Nutrient uptake half saturation	Yes	No
Sinking rate	Yes	No
Light Absorption	Yes	Yes
Palatability	Yes	Yes
Grazing rate	Yes	No
Other mortality	No	No
Carbon quota	Yes	No
Stoichiometry	No	Yes

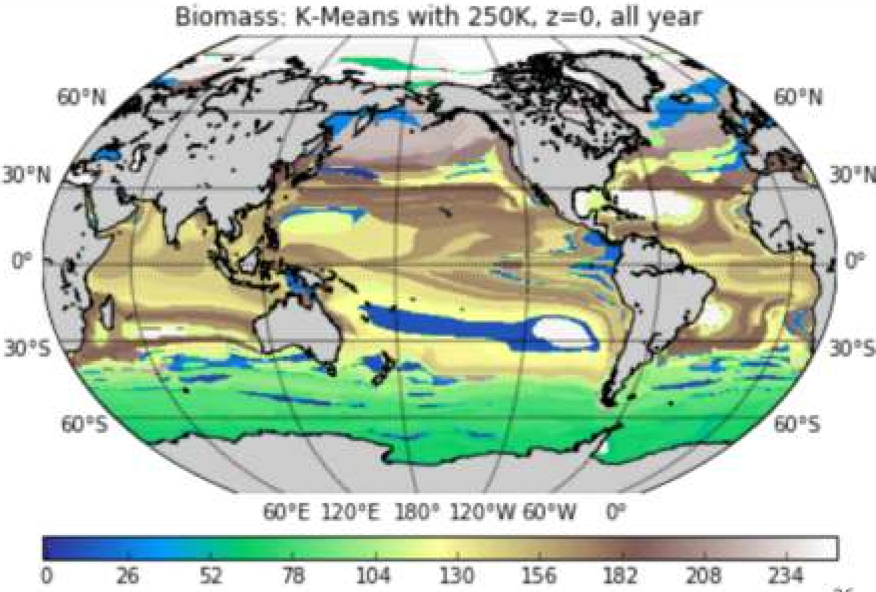
Dutkiewicz et al., 2015

51 species (biomass) and 4 nutrients (concentration)

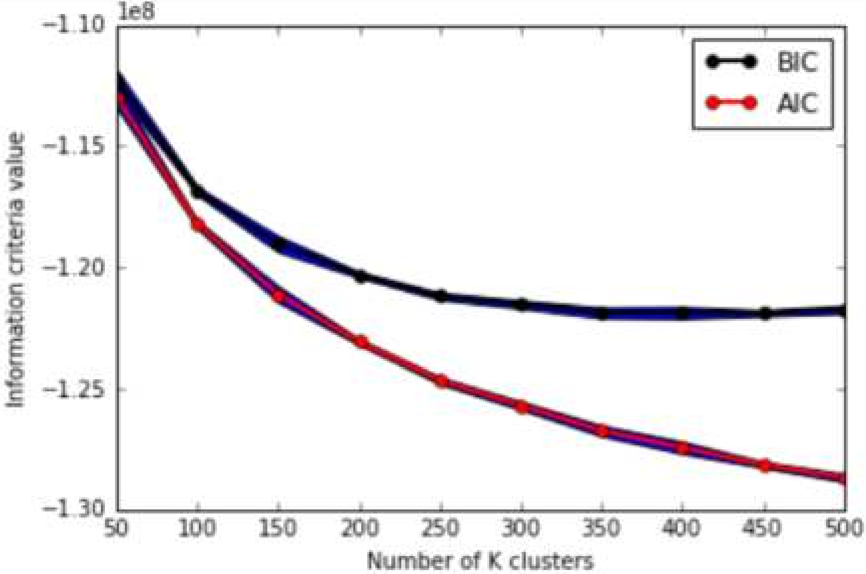


Sonnewald, Dutkiewicz and Hill. in prep

cbiomes K-Means

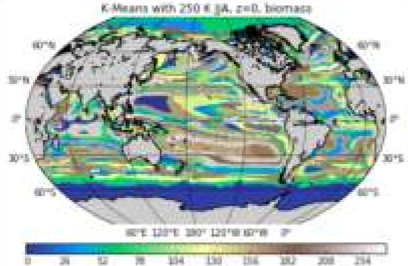


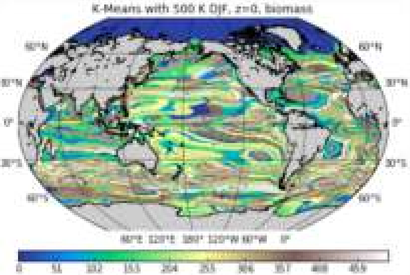
IC: cbiomes K-Means



AIC penalizes more complex models, i.e., models with additional

cbiomes K-Means: Summer/Winter



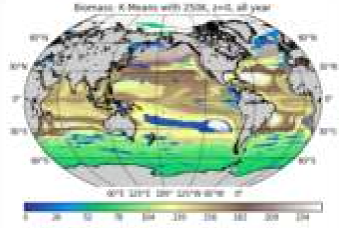


2: The Bad ...Global ecological provinces

The data is clearly not like the BV

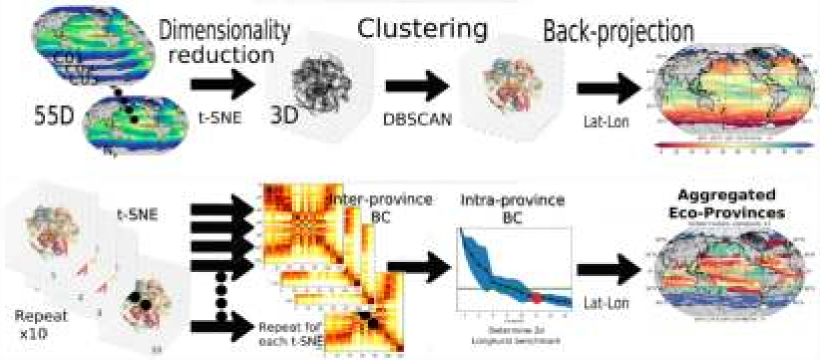
→ We do **not** have round ND distributions

- The yearly "provinces" looked reasonable
- Winter/Summer split was odd
- AIC+BIC confirmed suspicion



3: The (robustly) Ugly …Global ecological provinces

Clustering Eco-Provinces: Workflow



Learning: t-SNE to understand the topology of the data

Given a set of N high-dimensional objects x_1 ,..., x_N, the t-Statistic Neighbourhood Embedding minimize Kullbach-Leibner distance between the likelyhood of association between a low dimentional rendition and the high dimentional data.

- If x_i it the i-th object in the N dim space and y_i is the i-th object in the low-dim space:

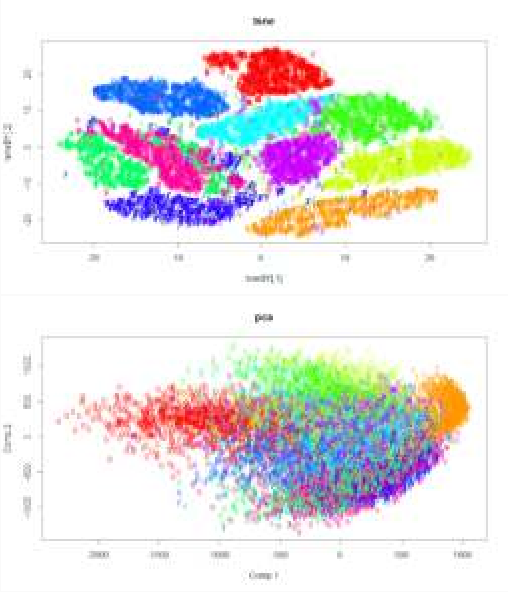
$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)},$$

and the same for a reduced dimensional set:

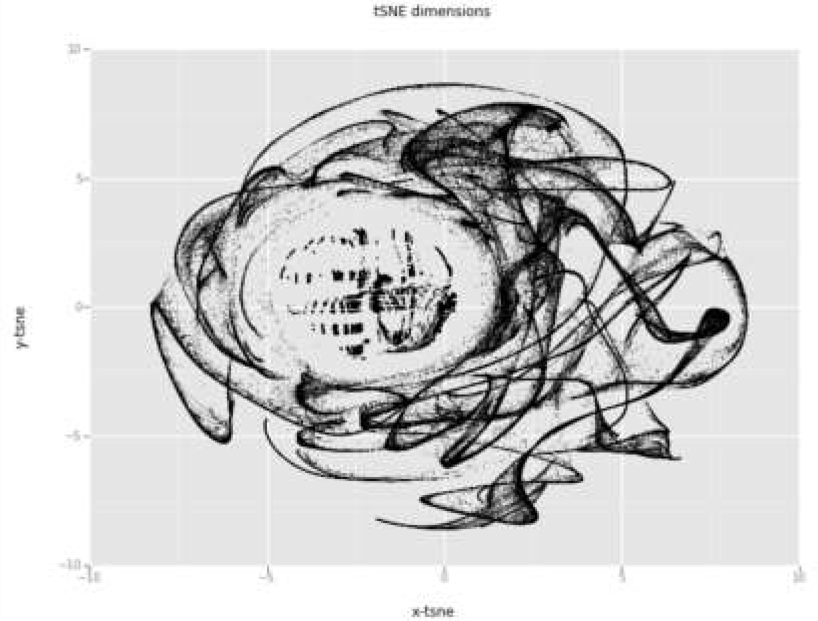
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}.$$

This is done as: $KL(P||Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

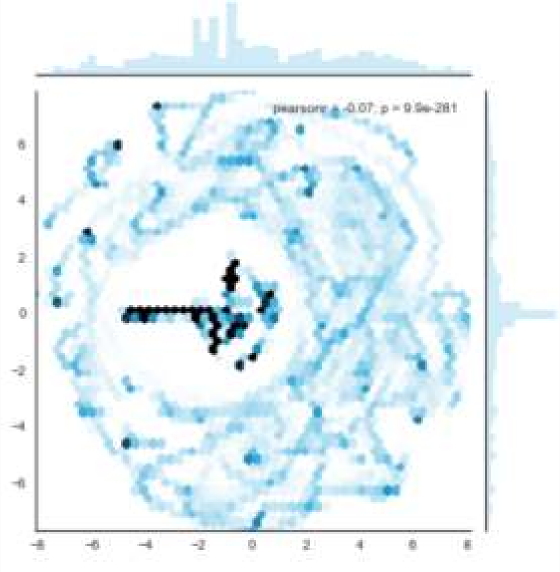
Learning: t-SNE vs PCA



Learning: What model?

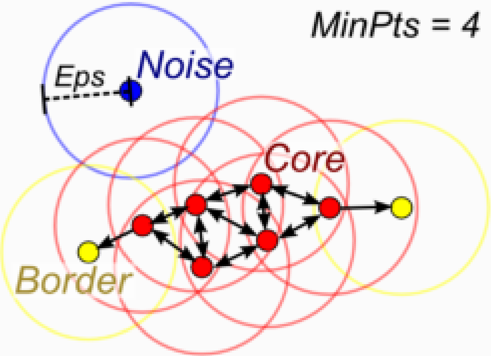


Learning: Look at the data!



Learning: Look at the data!

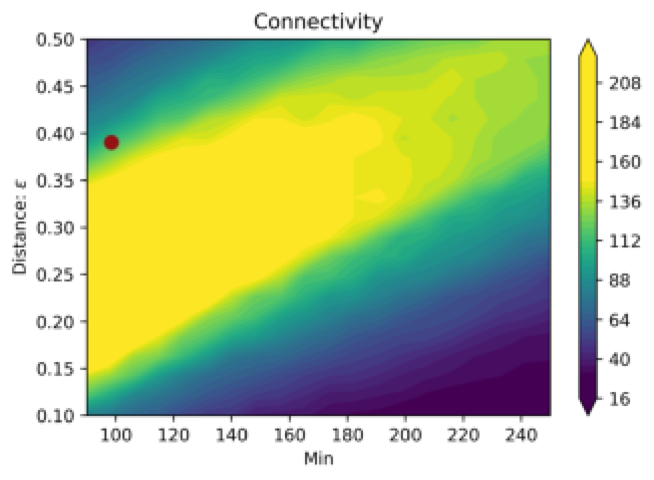
Unsupervised learning: DBSCAN



stackexchange.org

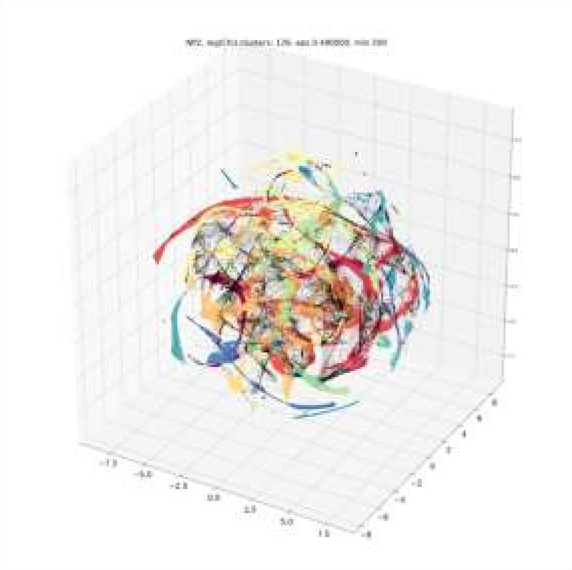
- Set: Eps and MinPts.
- Note: Not stochastic.
- Global. Need to preprocess data.

Unsupervised learning: DBSCAN

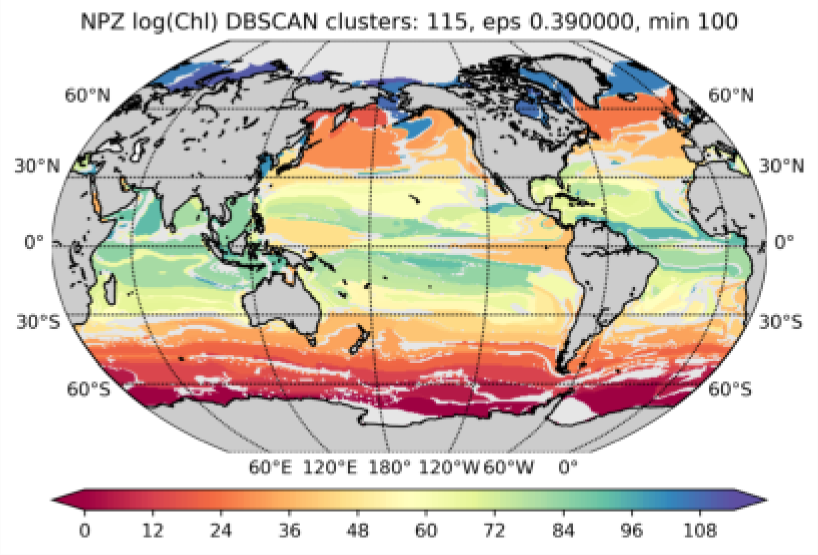


2D “elbow” check in connectedness

Unsupervised learning: DBSCAN

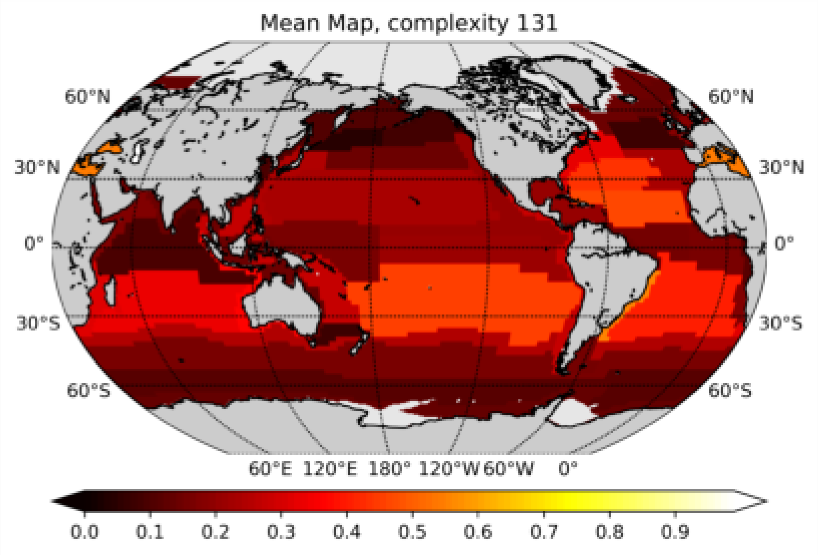


Biogeography: Clustering Eco-Provinces



Do we need all provinces? Are we doing better than Longhurst?

Longhurst: Expert assessment



NPZ complexity: Bray-Curtis dissimilarity

How similar are the identified clusters?

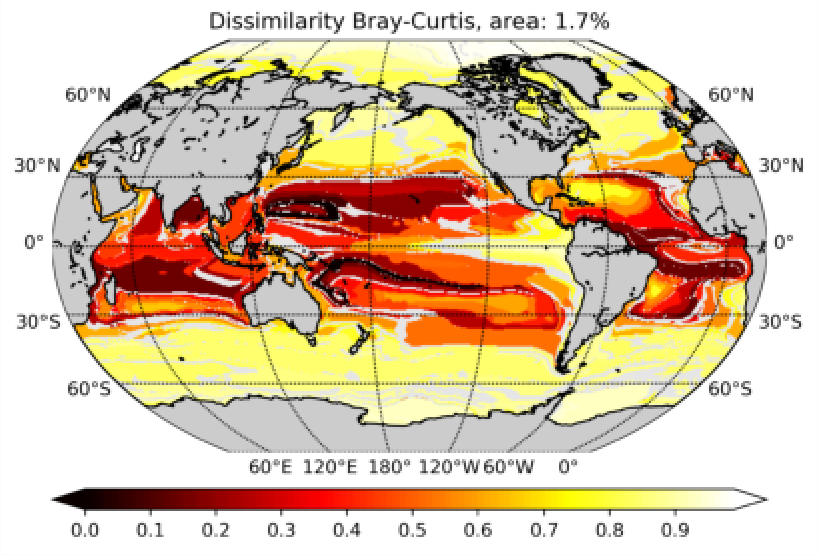
$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

C_{ij} is the minimum value present where similarities exist:

$$C_{ij} = \sum_{c51}^{c01} min(\mathrm{biomass}_i, \mathrm{biomass}_j)$$

S_i is the total across plankton: $S_i = \sum_{c51}^{c01}(\mathrm{biomass}_i)$

Inter-province: Bray-Curtis dissimilarity

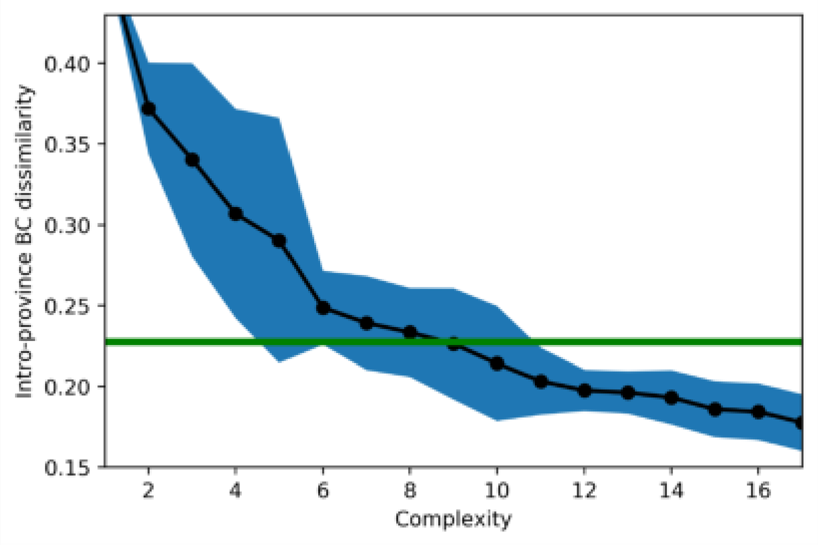


Many of eco-provinces are very similar

NPZ complexity: Connectivity

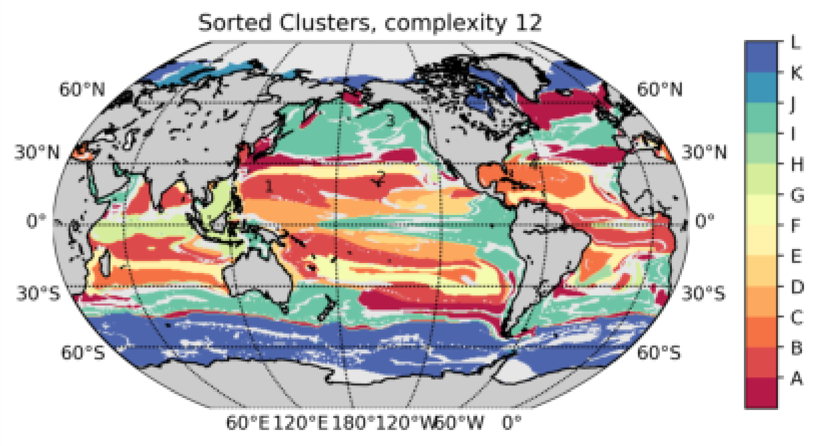
- Each eco-province connected to every other
- Facebook allegory
- Graph theory allows sorting

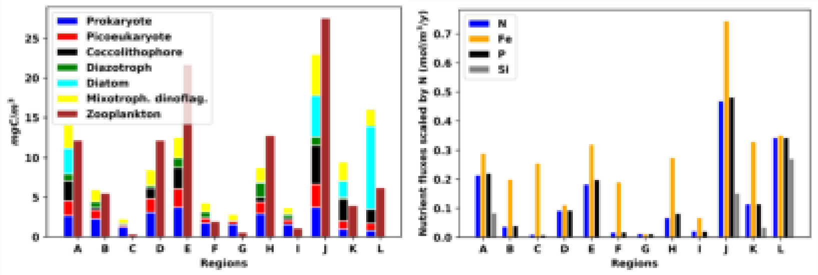
Intra-province: Bray-Curtis dissimilarity



A minimum complexity of 12 is recommended

Aggregated Ecological Provinces



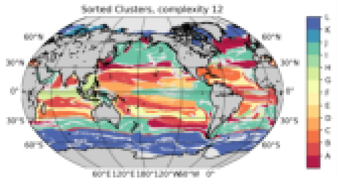


- Similar biomass/chl but different community structure
- Biomass is poor predictor of zooplankton: Trophic cascades?

3: The (robustly)Ugly ...Global ecological provinces

Visualizing the data in 3D allowed model selection
→ Knowledge of topology avoids brute-force

- Robust eco-provinces
- Improvements over Longhurst
- Aggregation allowed wider application



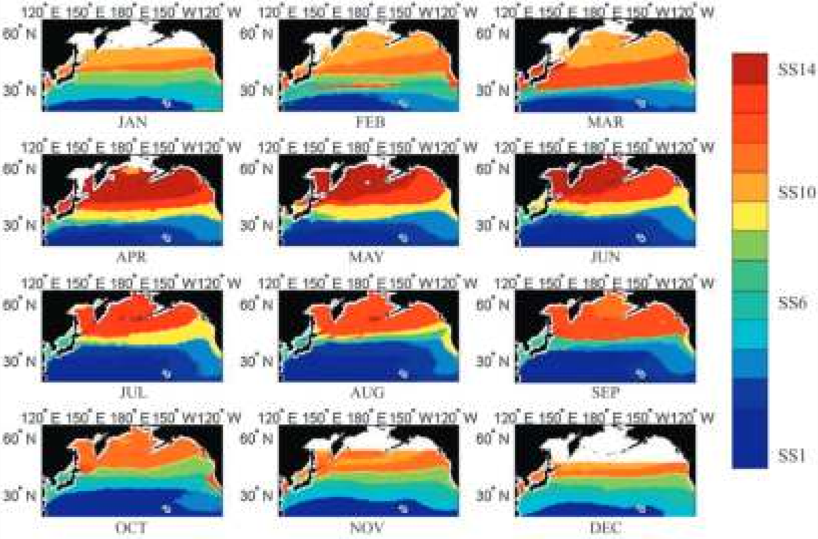
The value of provinces/seascapes

How complicated is complicated “enough”?

The value of provinces/seascapes

How complicated is complicated “enough”? What does conservation work require?

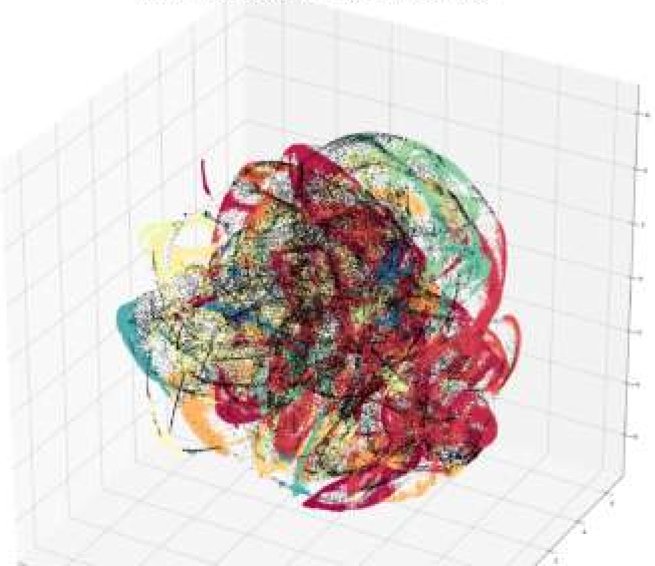
Seascapes



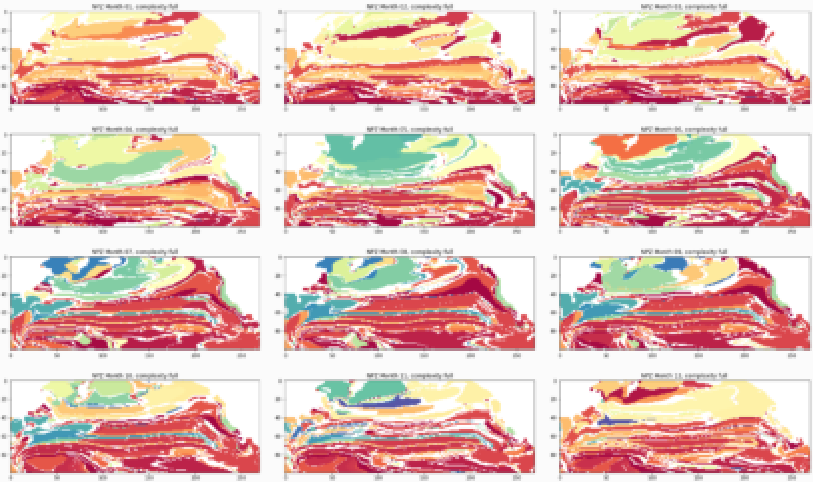
Kavenaugh et al. 2016

SOM+hierarchical clustering: SST, PAR and Chl a

Seascapes+ecology?



Seascapes+ecology



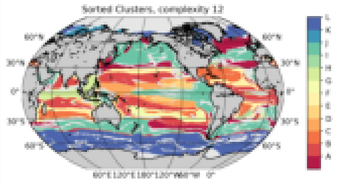
Similarities when looking at ecology!

Summary

Summary

The singularity may not be coming; Make complicated models more complex?

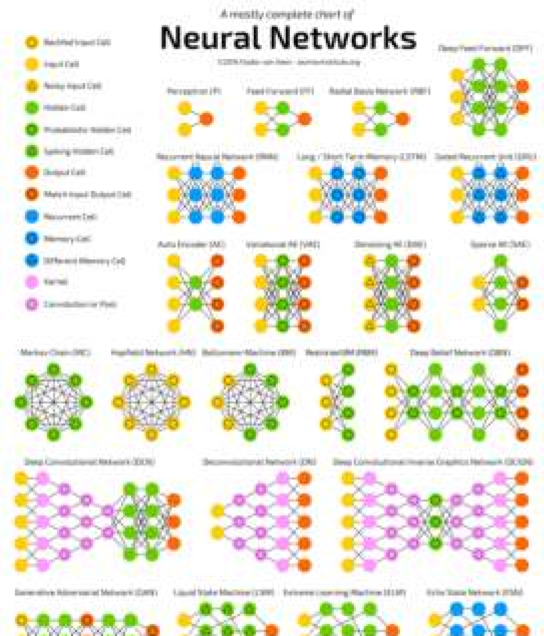
- To allow unsupervised ML to help:
    - Preprocess data as appropriate
    - Start with simple model
    - Carefully estimate parameters
    - Assess with system insight
- System insight to simplify the complicated components:
    - Ecosystem models?
    - Bathymetric interactions
    - 3D ocean insight



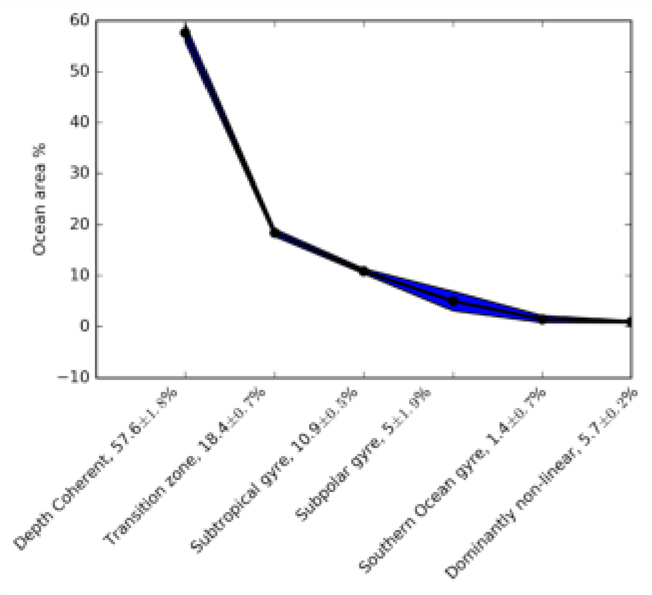
Modeling+Theory+Observations = Understanding of natural world

Summary II

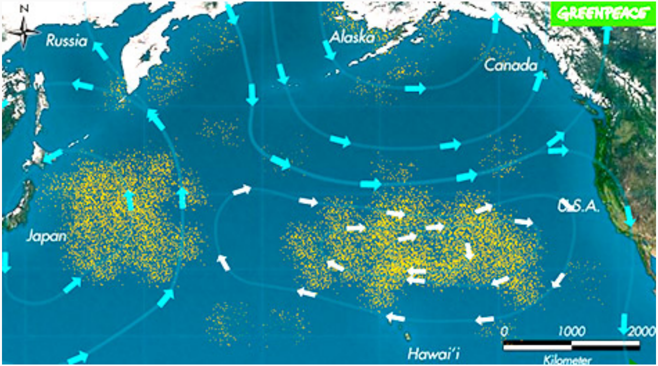
Neural Nets



Percentage of area accounted for



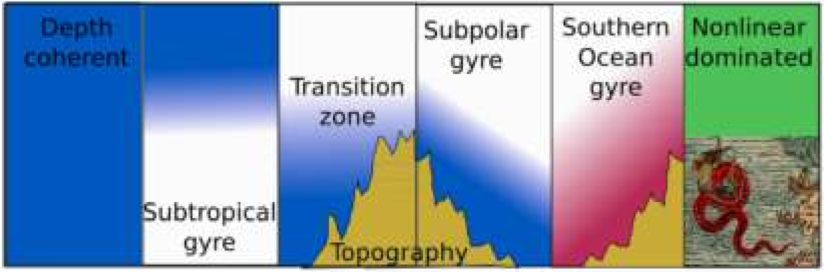
The great garbage patch



Garbage follows currents. Ecology is not this simple...

Good to know if you want to clean it up with food webs in mind!

Interpretation: Cartoon



Implications: Gyre $\Leftrightarrow$ overturning?

kMeans_50_NAtlFit.png

Non-linear terms

$$\nabla \times \mathbf{A} = \nabla \times \left[\int_{-H}^{\eta} \nabla \cdot (\mathbf{uu}) \mathrm{d}z \right] + [w\zeta]_{z=H}^{z=\eta} + [\nabla w \times \mathbf{u}]_{z=H}^{z=\eta} \,, \quad (1)$$

where **uu** is a second order tensor. The RHS of equation (1) represents the curl of the vertically integrated momentum flux divergence, the non-linear contribution to vortex tube stretching and the conversion of vertical shear to barotropic vorticity.