Unsupervised learning: model choice and dimensionality reduction

Maike Sonnewald[1,2]

July, 2019

[1]MIT & [2]Harvard

Introduction

Giving rise to “general intelligense

Supervised and reinforcement learning:

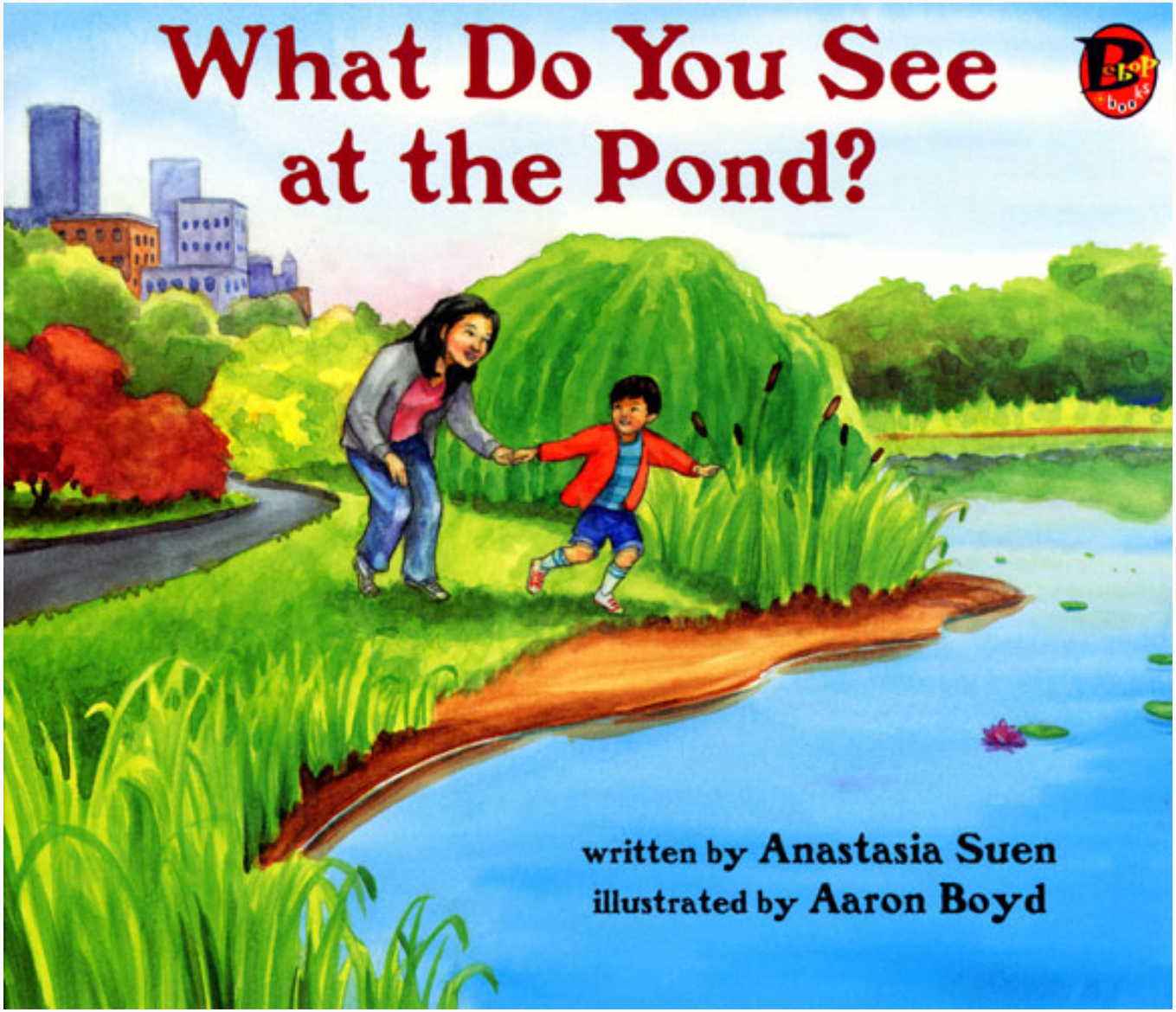
Train on target signals/rewards designed by humans.

Unsupervised learning:

Moving beyond limits of learning defined by human trainers.

Study emergent and hidden structures in data to identify groups and gain meaningful insights

Supervised and unsupervised learning



Supervised and unsupervised learning

Supervised

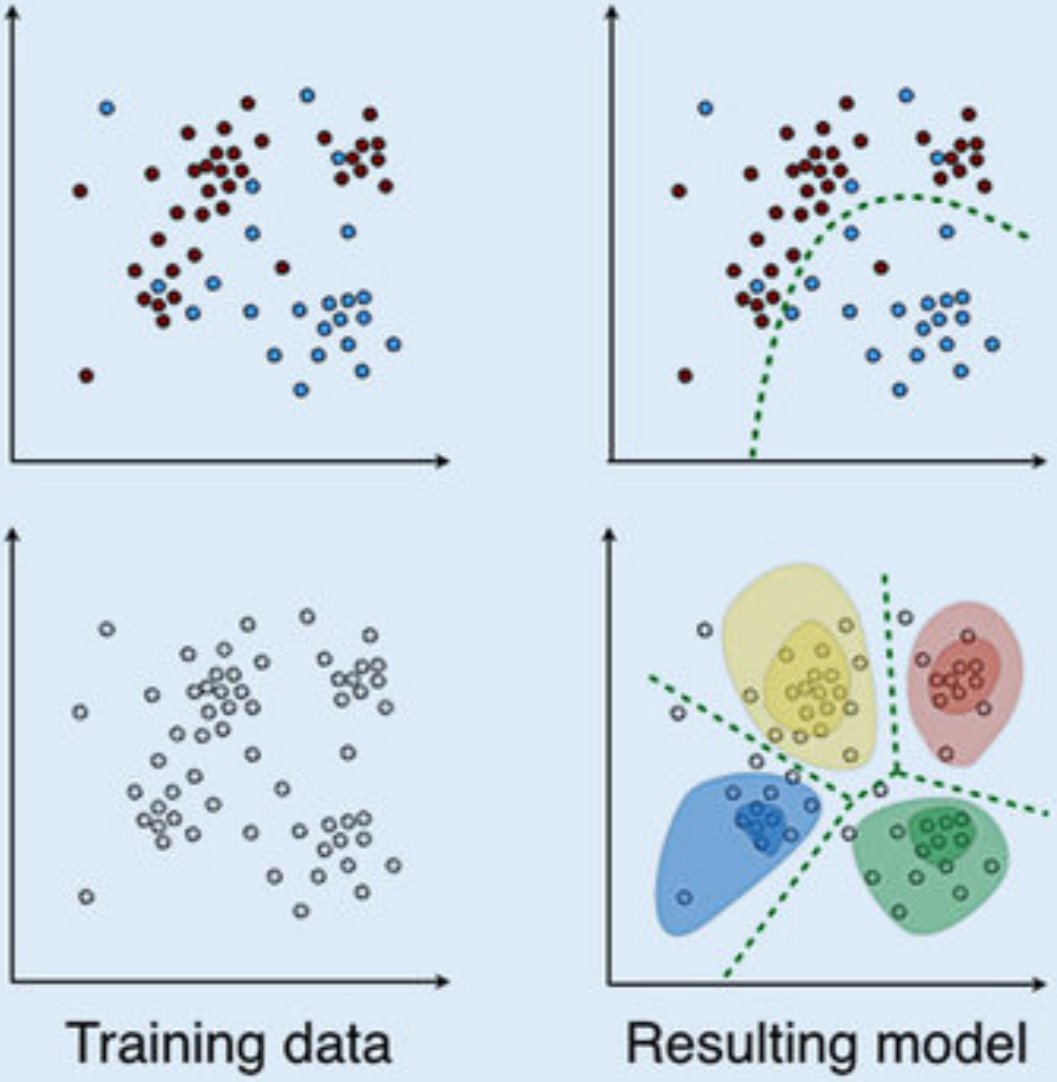
-Labeled data

-Decision boundary

Unsupervised

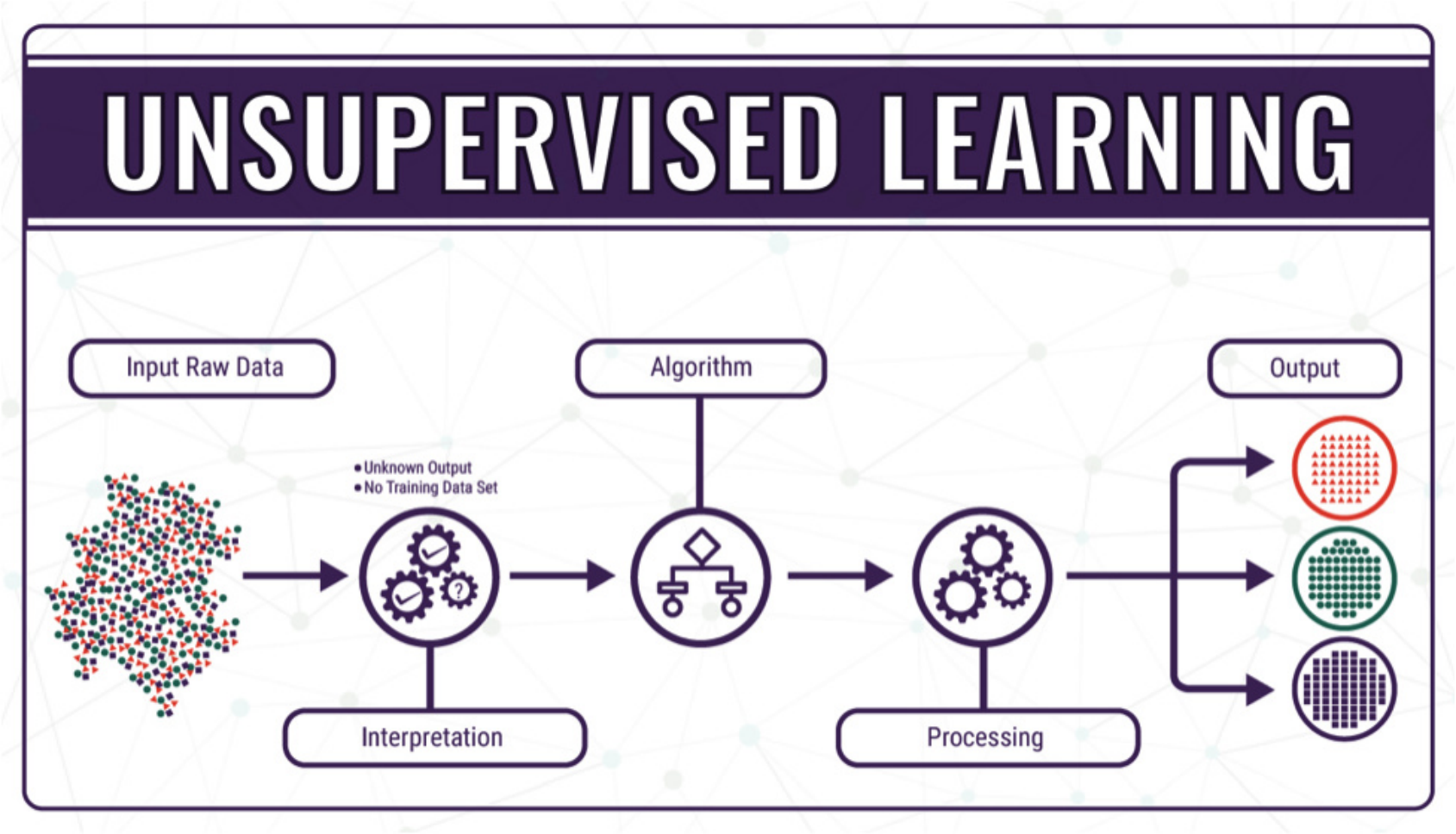
-No labels

-Identify structures



Langs et al. 2018

Outline

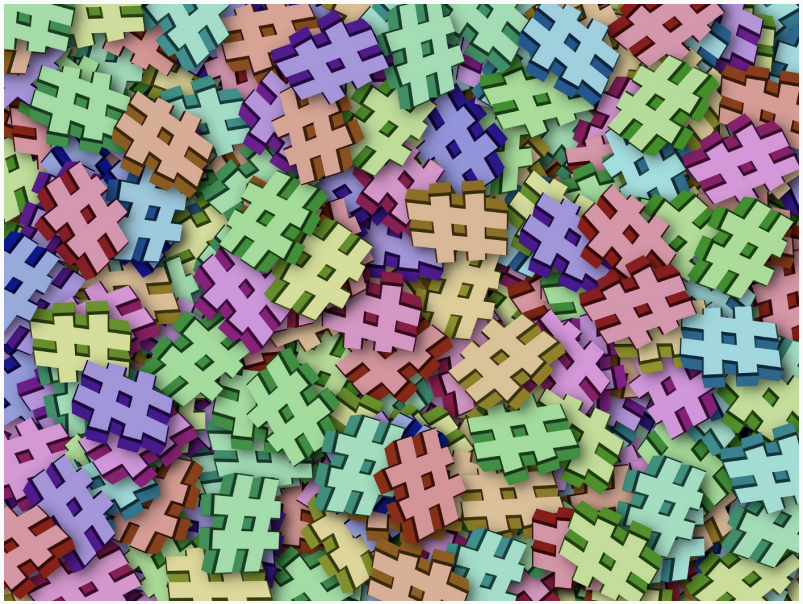


community.singularity.io

1: Input data+preparation→2: Model choice→3: Evaluation

Input data+preparation

What to do with the input data?



- With many features, using them all is difficult
- Learning is slow, and likely not robust leading to pour generalisation
- Preprocessing the data/features enhances important features
- Removes redundant features and noise

Feature Selection: select subset of meaningful features

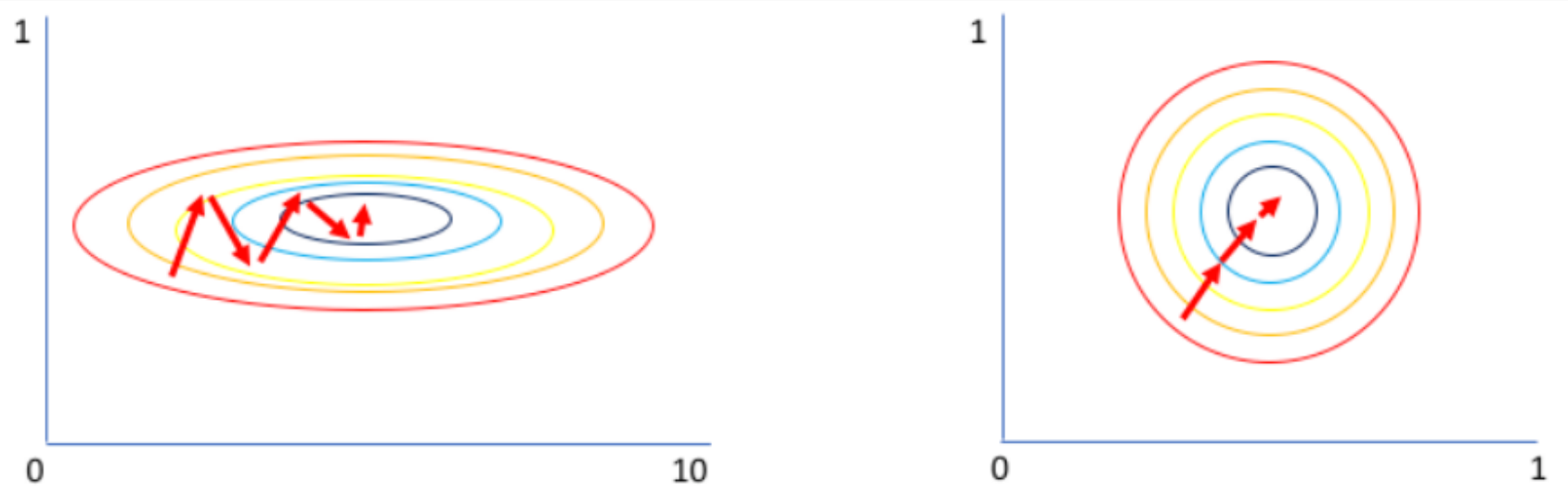
Feature Extraction: derive information building a new feature subspace

Spread of data: normalize and standardize

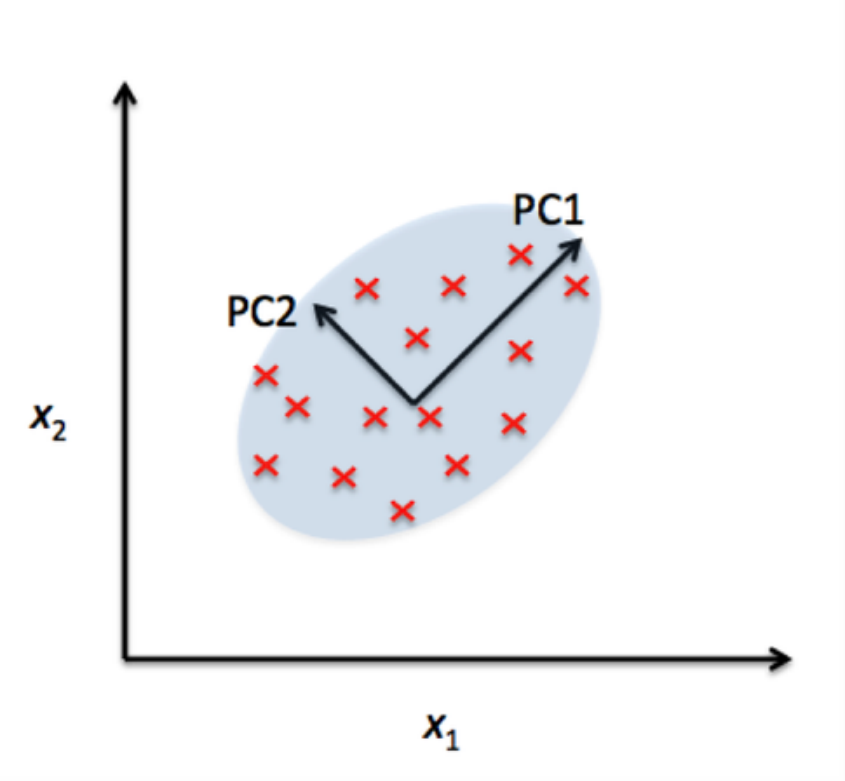
Normalization: rescaling range so values are within 0-1

Standardizing: rescaling distribution so the mean is 0 and the standard deviation is 1

- Gradients in one parameter can dominate search/update
- If comparable, both parameters contribute equally

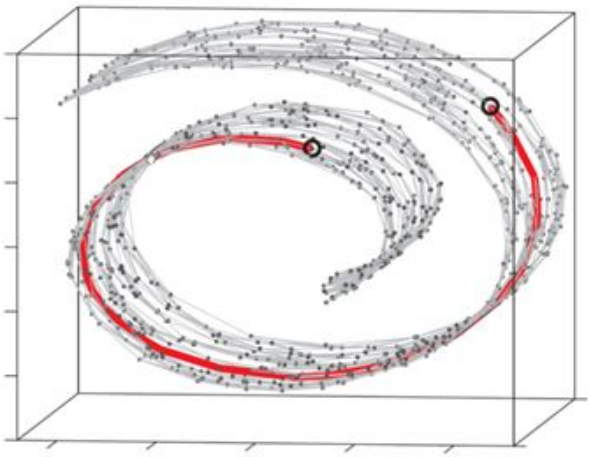


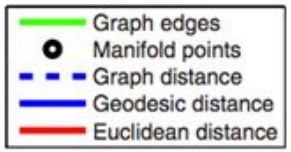
Dimensionality reduction: PCA et al.

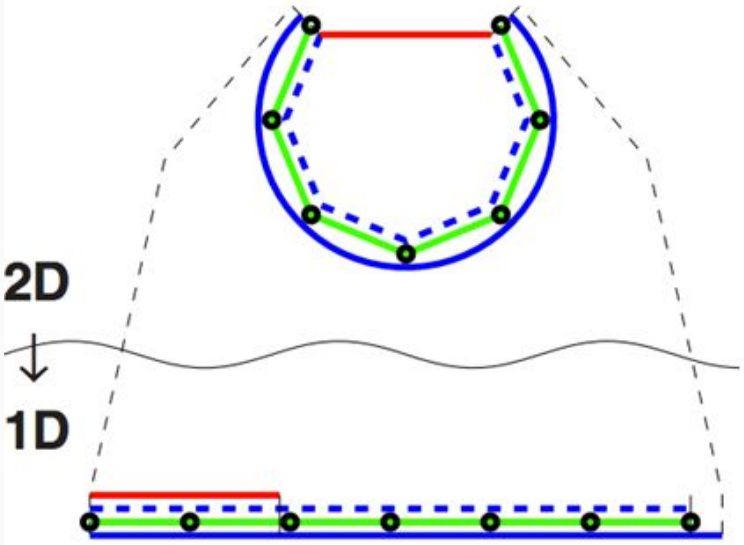


- Data often only varies in only some limited directions
- Reduce dimensions by projecting onto a low dimensional subspace conserving maximal variation

Dimensionality reduction: PCA et al.







Isomap

Approximate geodesic

$$l = \min_{\mathbf{p}(z)} \int_{z(i)}^{z(j)} \|\mathbf{J}_z \mathbf{m}(\mathbf{p}(z))\| dz$$

- Build graph with K-neighbors/ε-ball
- Weight graph with Euclidean distance
- Compute pairwise distances with Dijkstra's algorithm

Dimensionality reduction: PCA et al.

- 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:

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

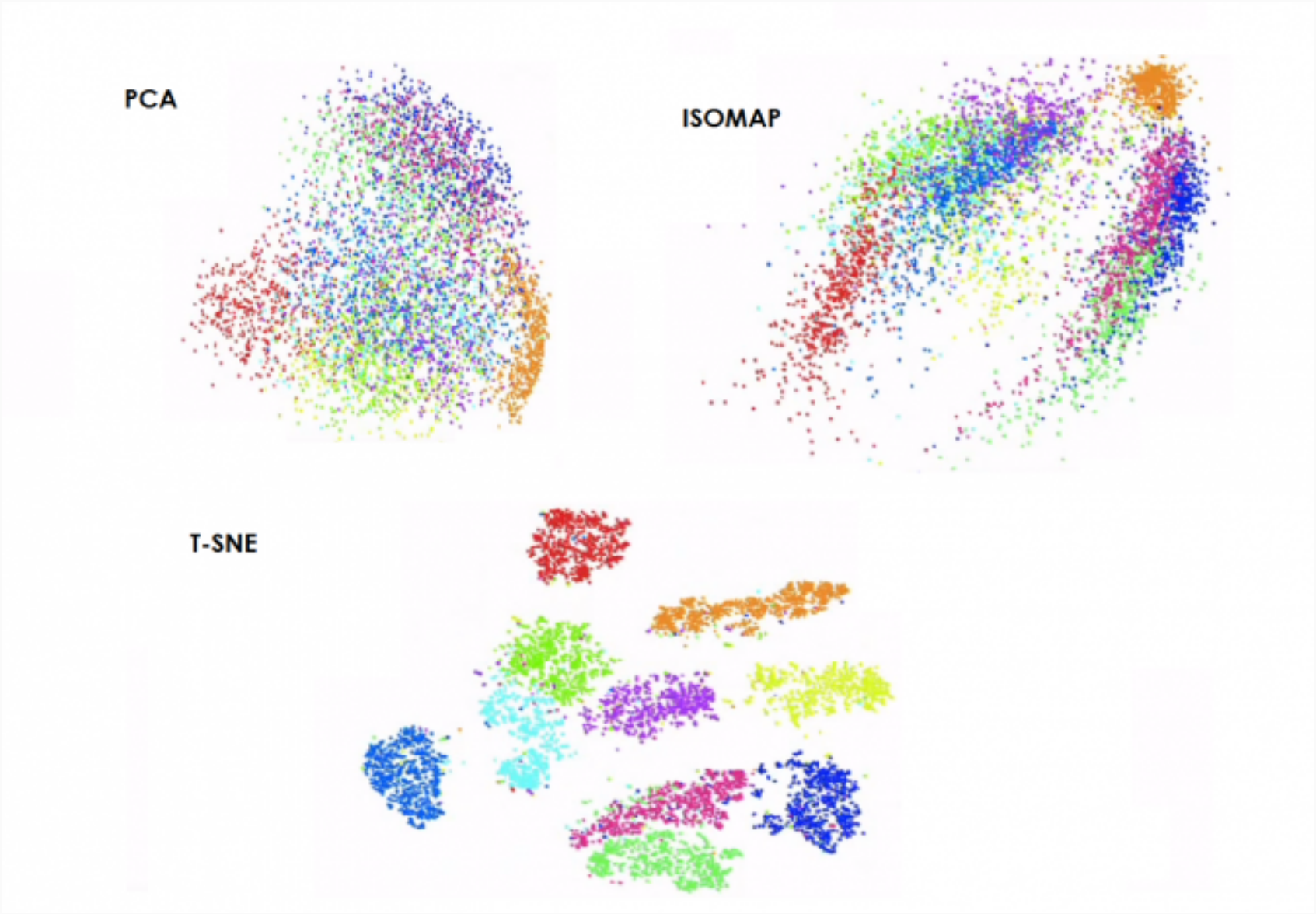
and the same for a reduced dimensional set:

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

This is done as:

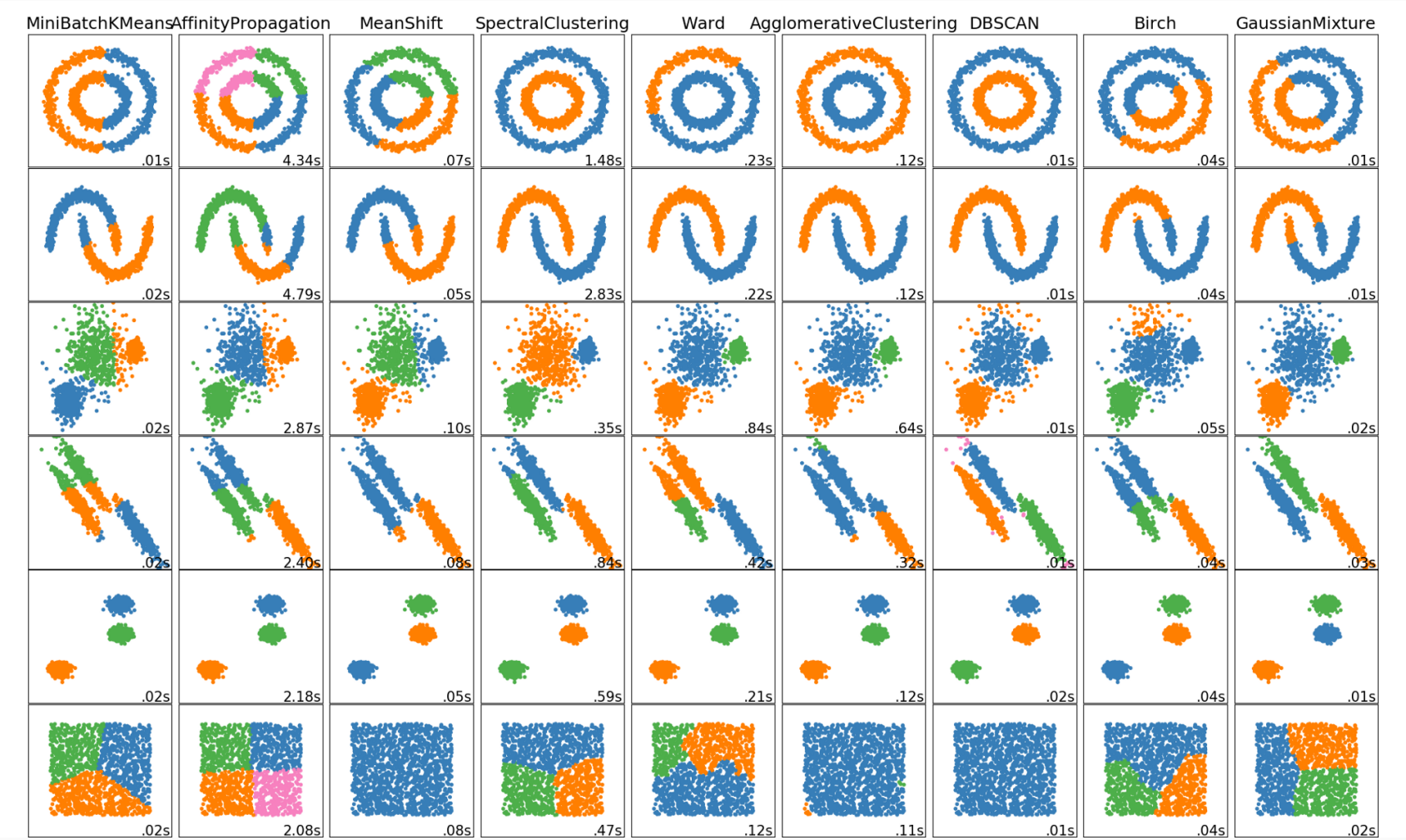
$$KL(P||Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

Visualization



Model Choice

Types of models



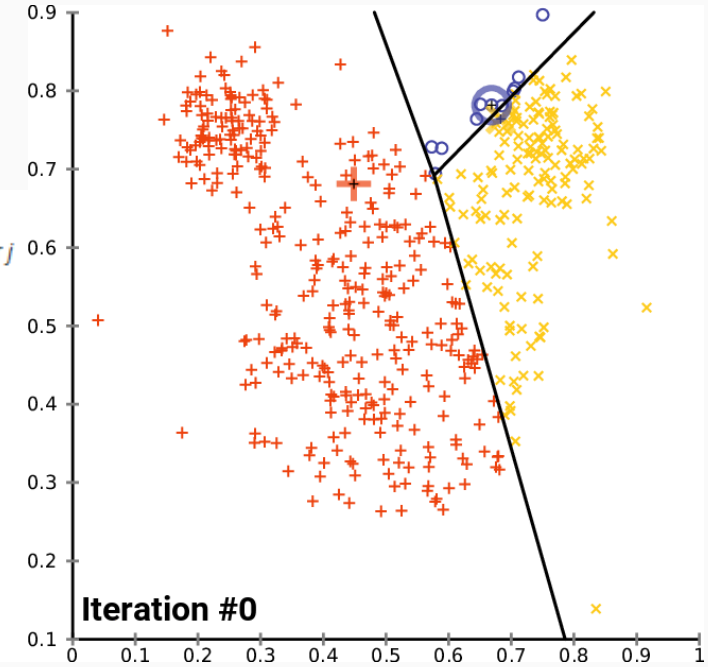
scikit Learn

Unsupervised learning: K-Means clustering

number of clusters
number of cases
centroid for cluster j
case i

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

Distance function



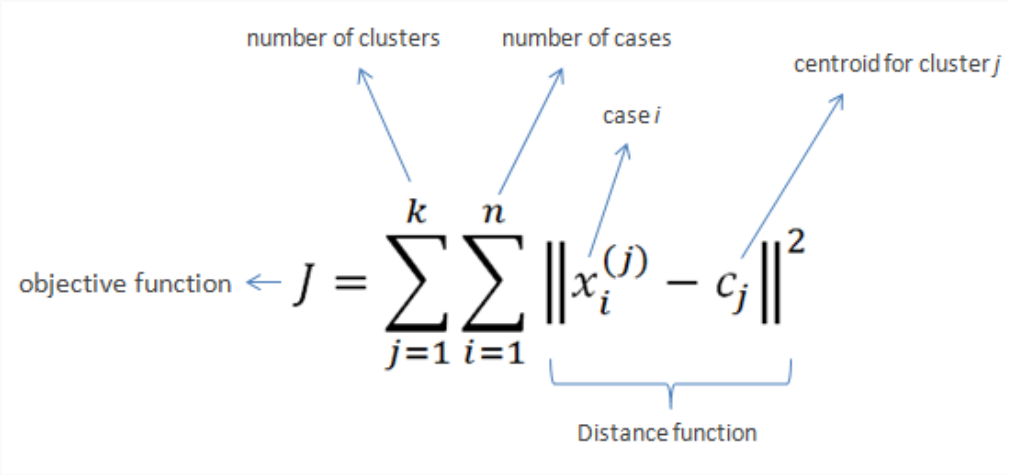
commons.wikimedia.org

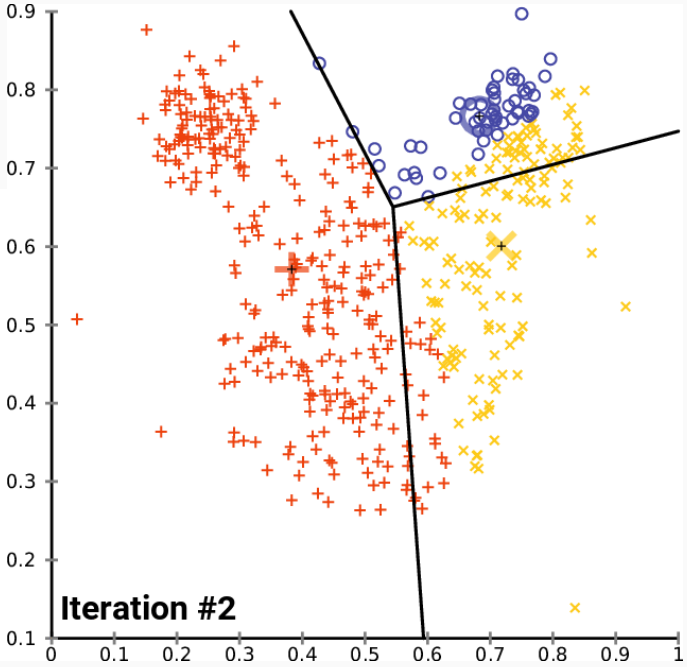
- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Iterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:

K and initialization.

Unsupervised learning: K-Means clustering

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





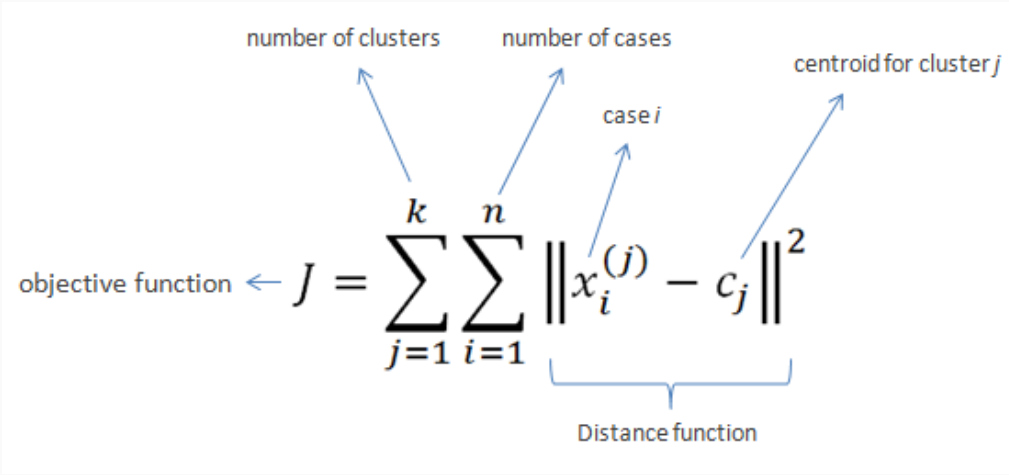
commons.wikimedia.org

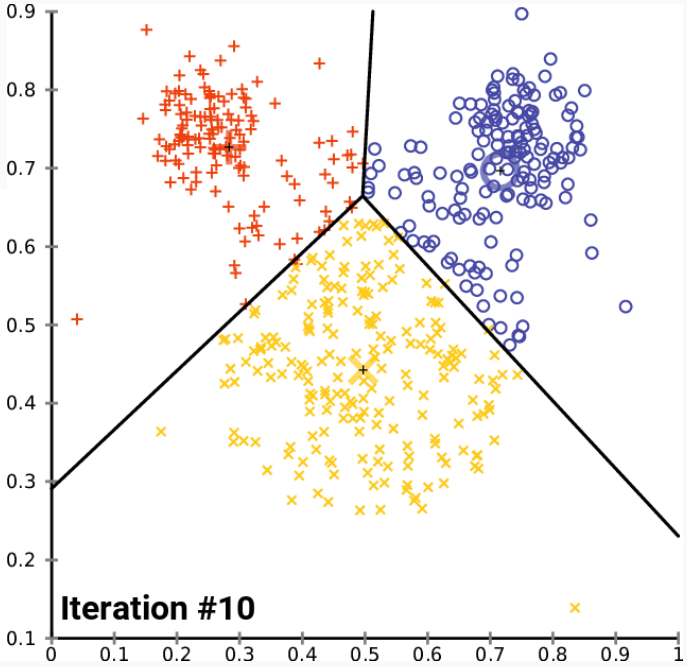
- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Iterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:

K and initialization.

Unsupervised learning: K-Means clustering

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





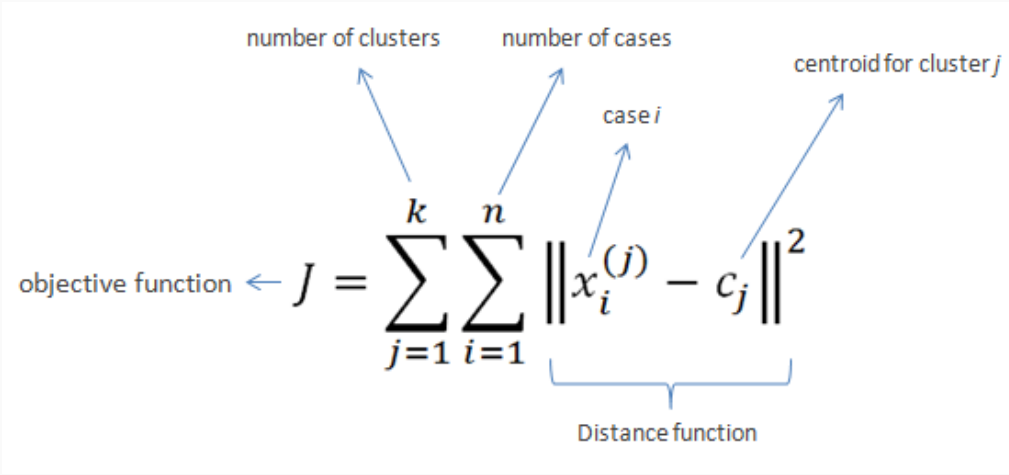
commons.wikimedia.org

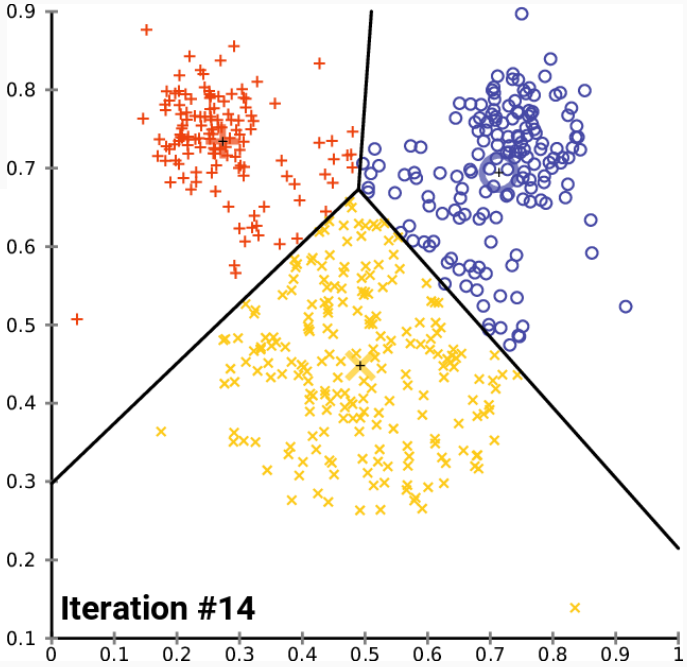
- Assign: Cluster w mean minimizing least squared Euclidean dist.
- Iterate: Calculate the new means.
- NB! NP-hard. Not global. Need to treat data. Sensitive to:

K and initialization.

Unsupervised learning: K-Means clustering

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



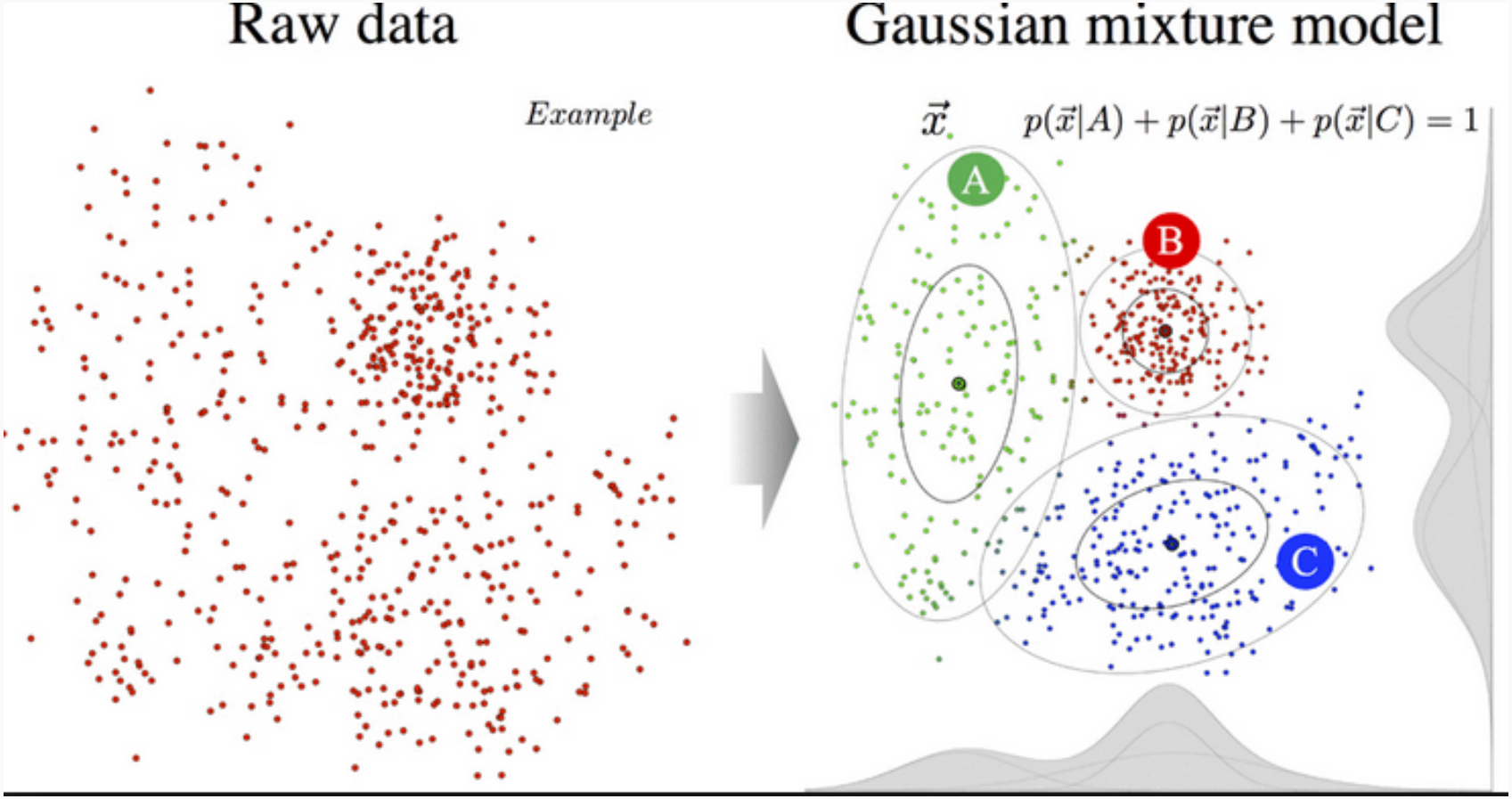


commons.wikimedia.org

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

K and initialization.

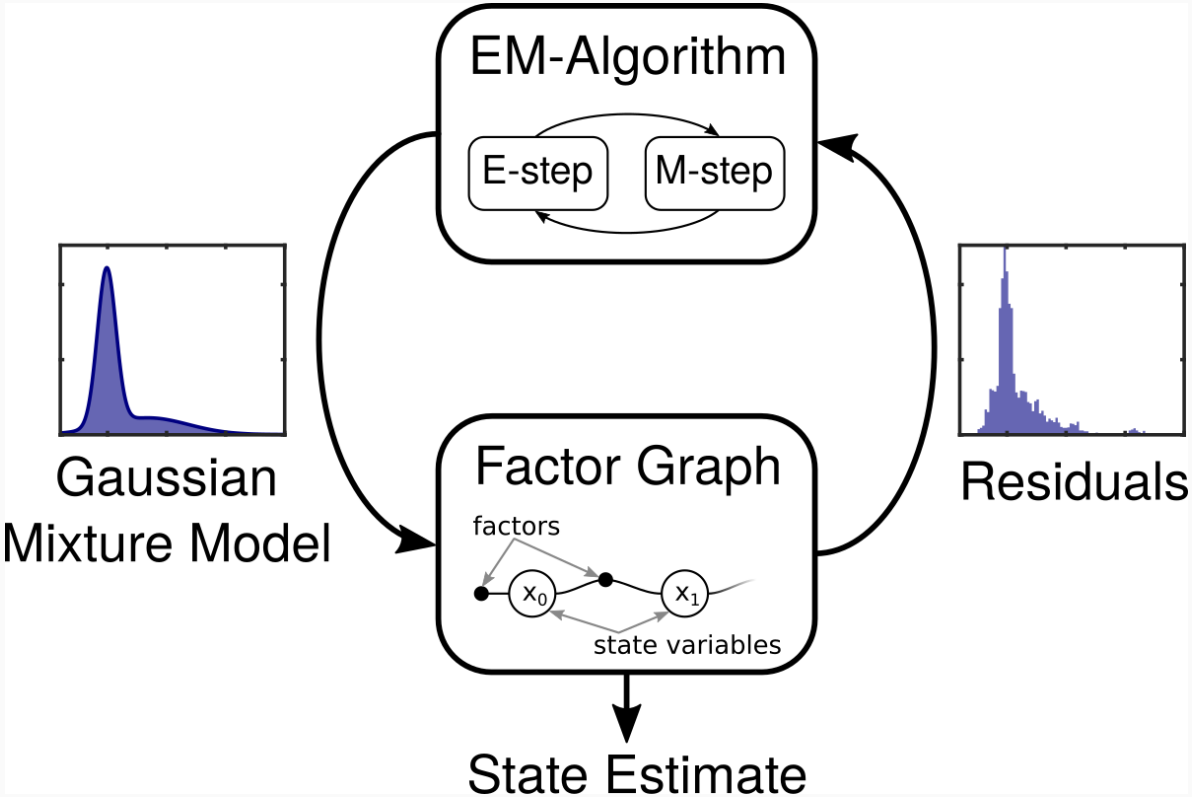
GMM



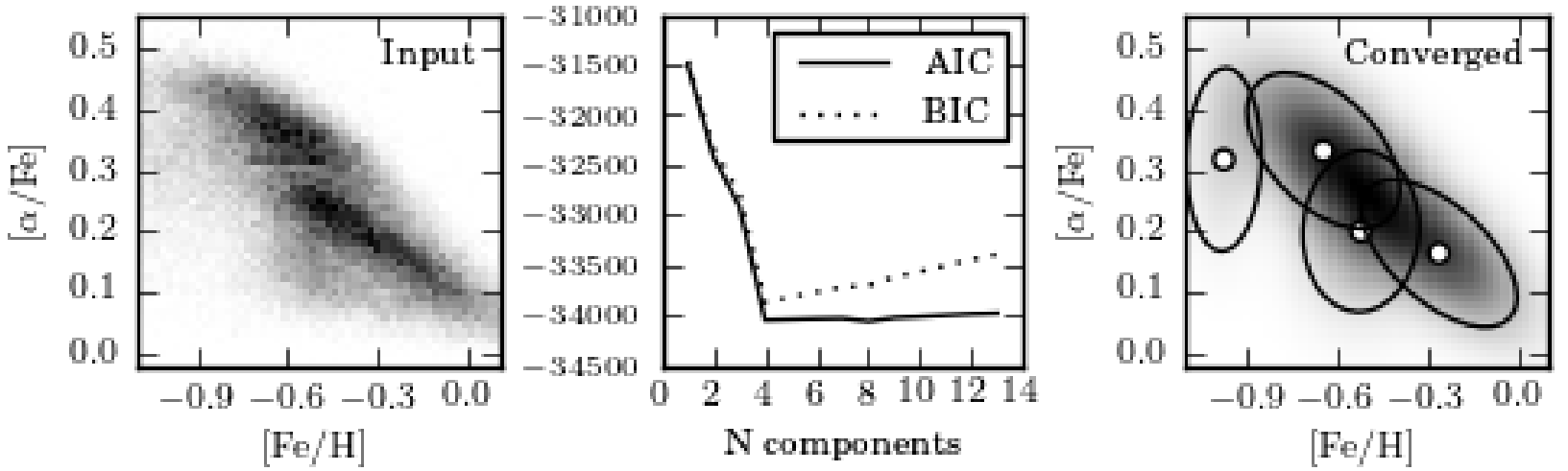
Pham, 2016

GMM + EM

Iterative maximum likelihood method for latent variables



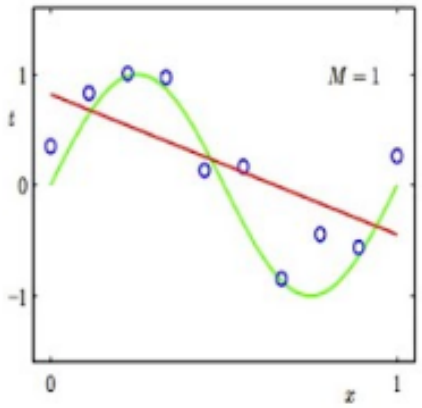
Don’t forget the IC

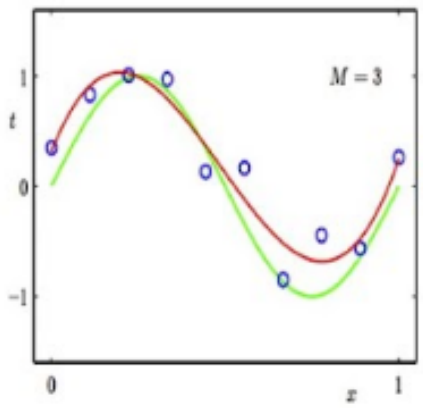


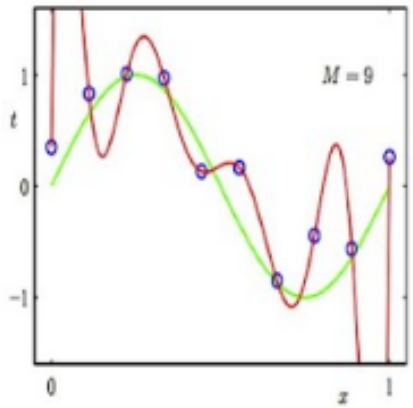
astroML

Under and overfitting

Regression:



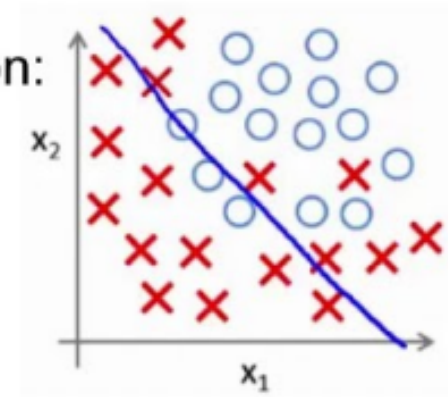


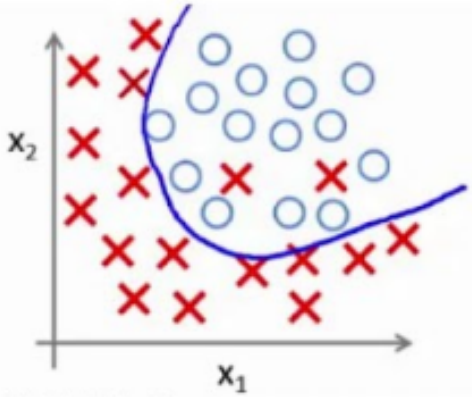


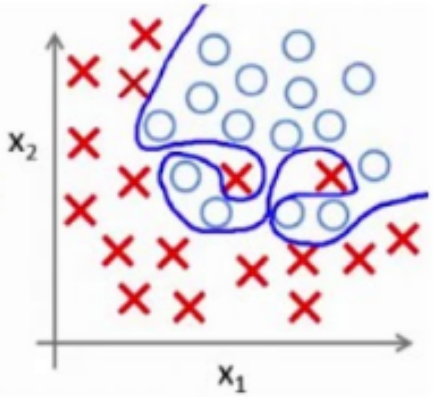
predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:





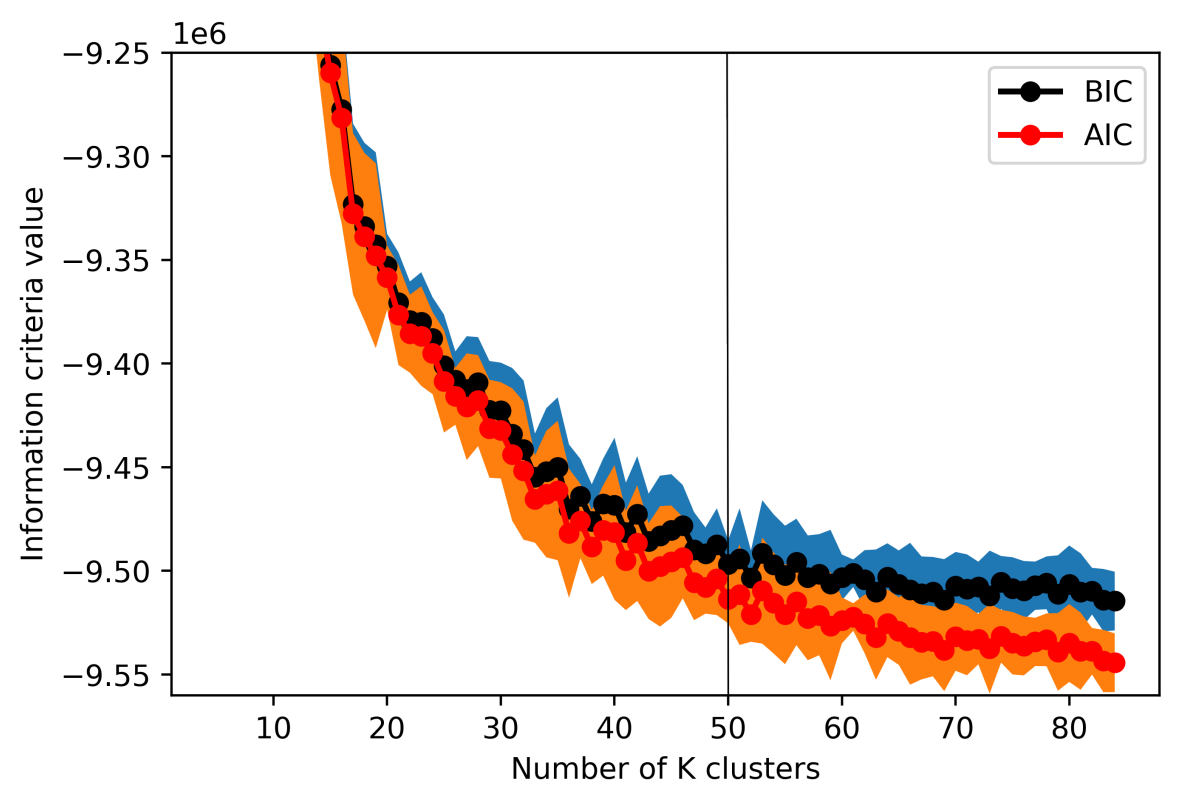


What is under and overfitting?

Too simple: inflexible learning due to too few/wrong features or too strict regularization $\rightarrow$ little variance but more bias

Too complex: more prediction variance

Information Criteria



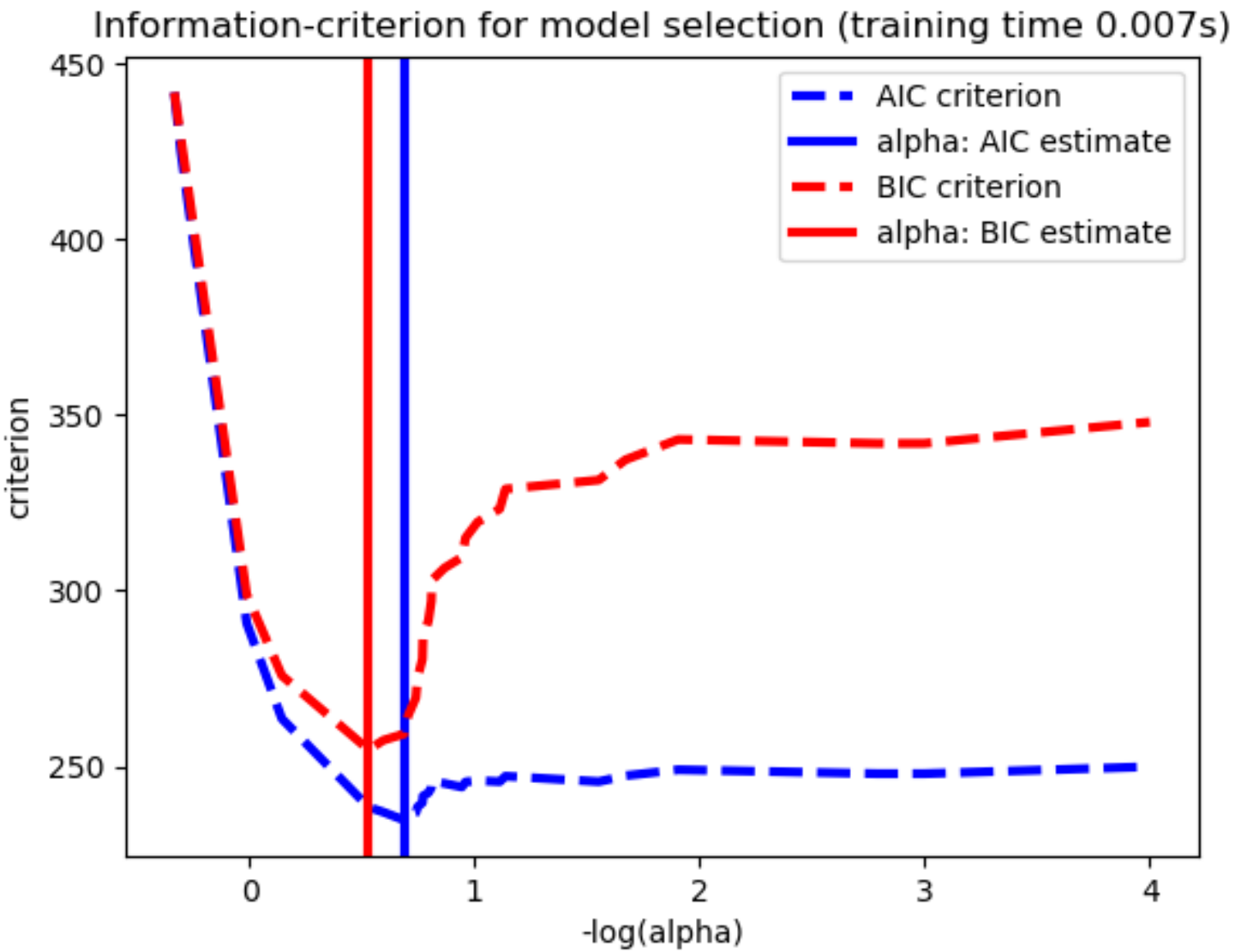
$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

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

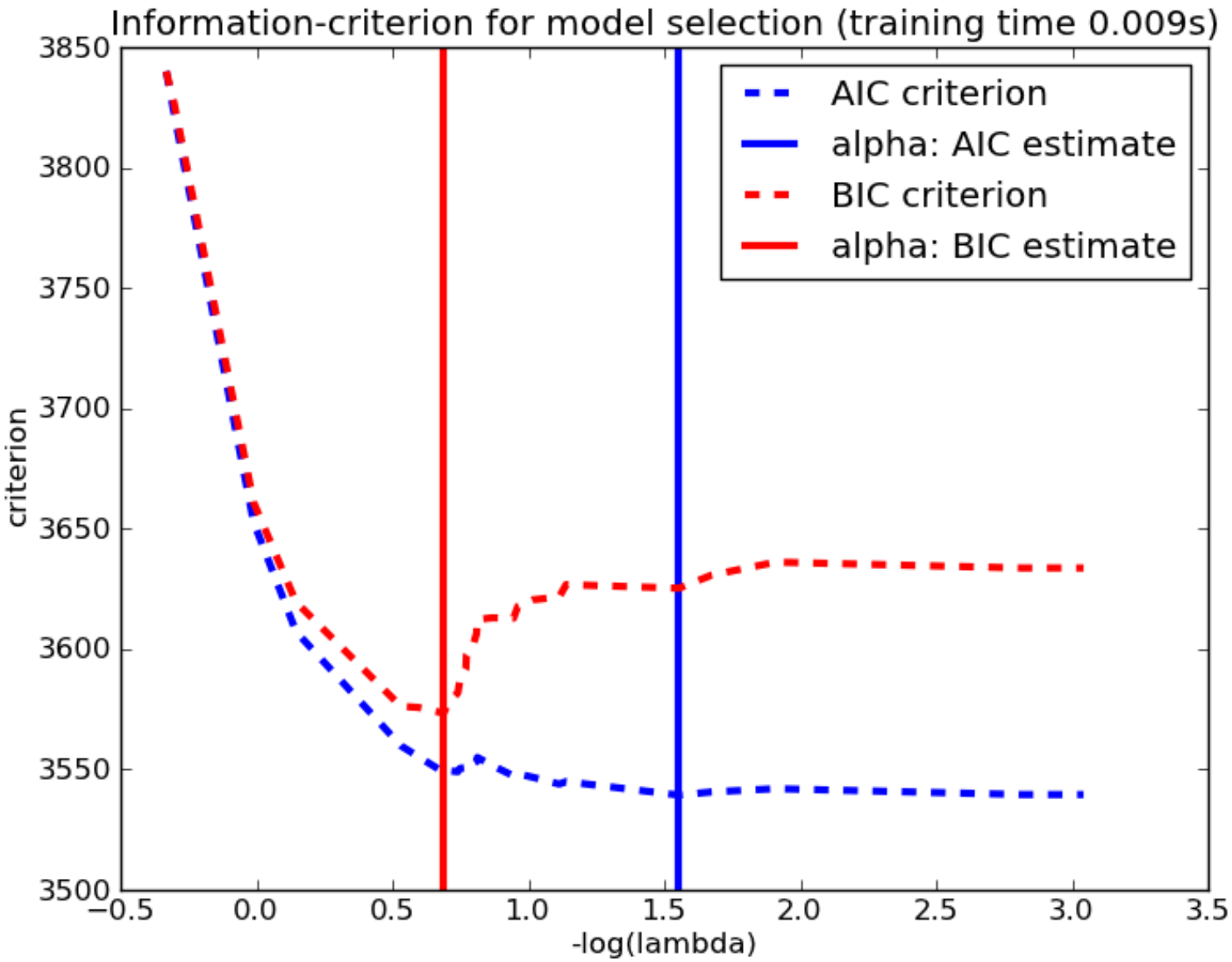
where n is the number of datapoints and $\mathcal{L}$ is the likelihood:

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

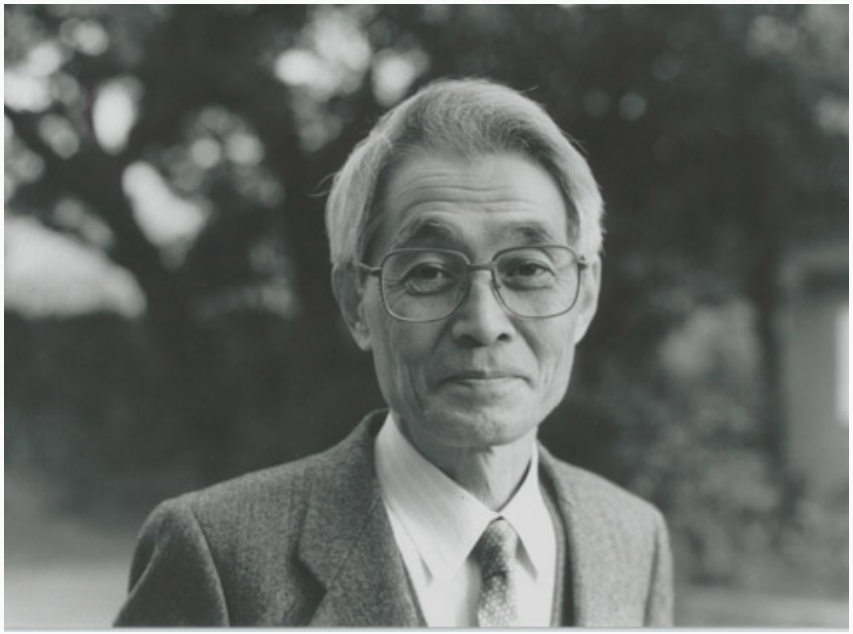
Information Criteria: Note



Information Criteria: Note



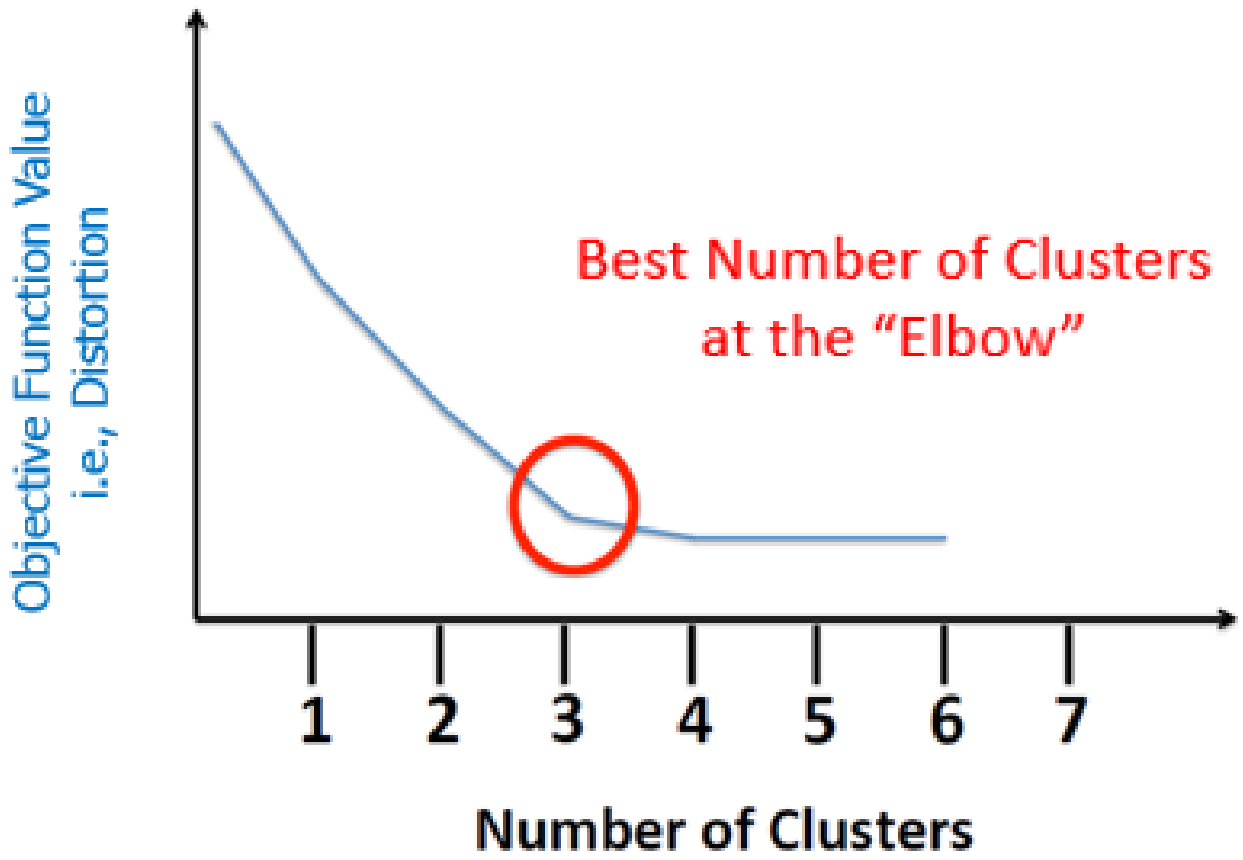
Information Criteria: G. Schwartz and H. Akaike



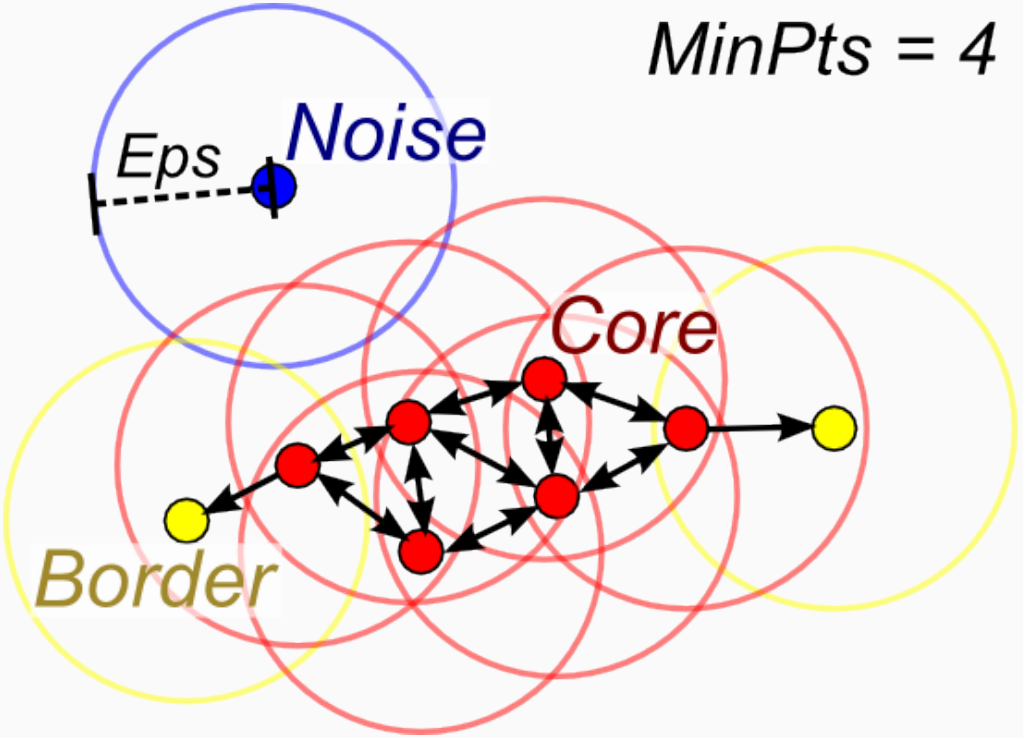
...now forms the basis of a paradigm for the foundations of statistics; as well, it is widely used for statistical inference.

Wikipedia

Information Criteria: Elbows



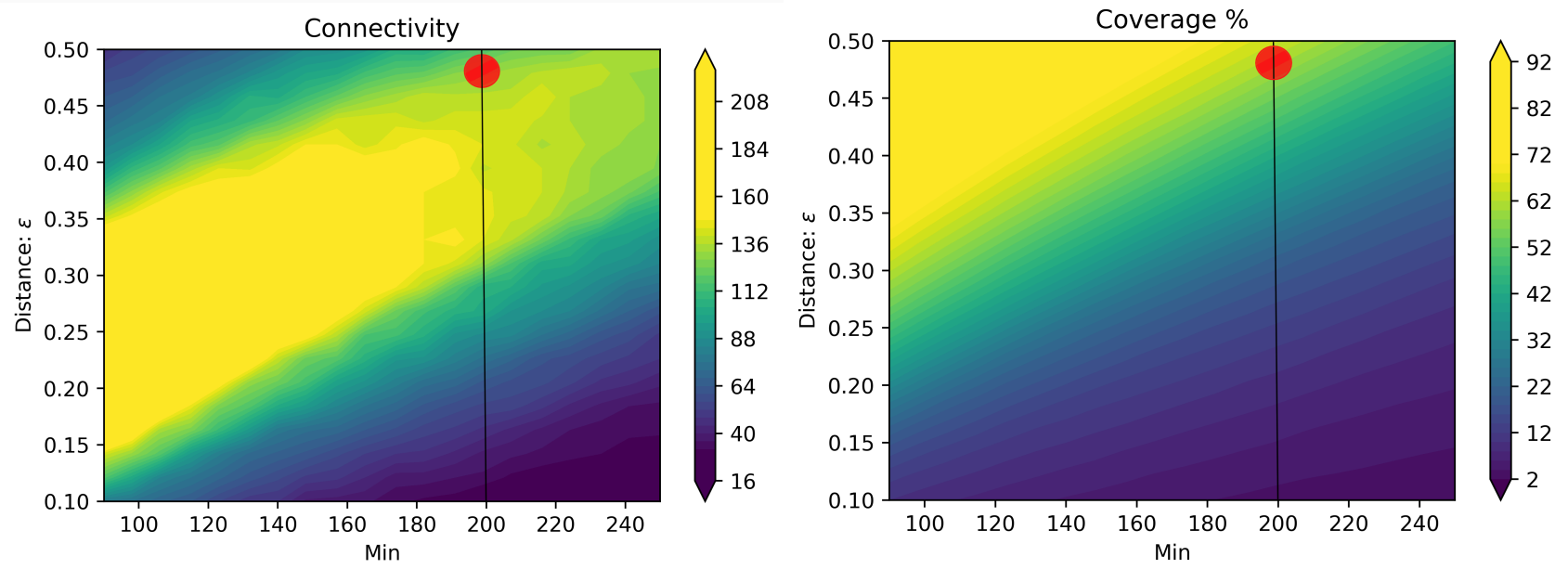
Other parameters: DBSCAN



stackexchange.org

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

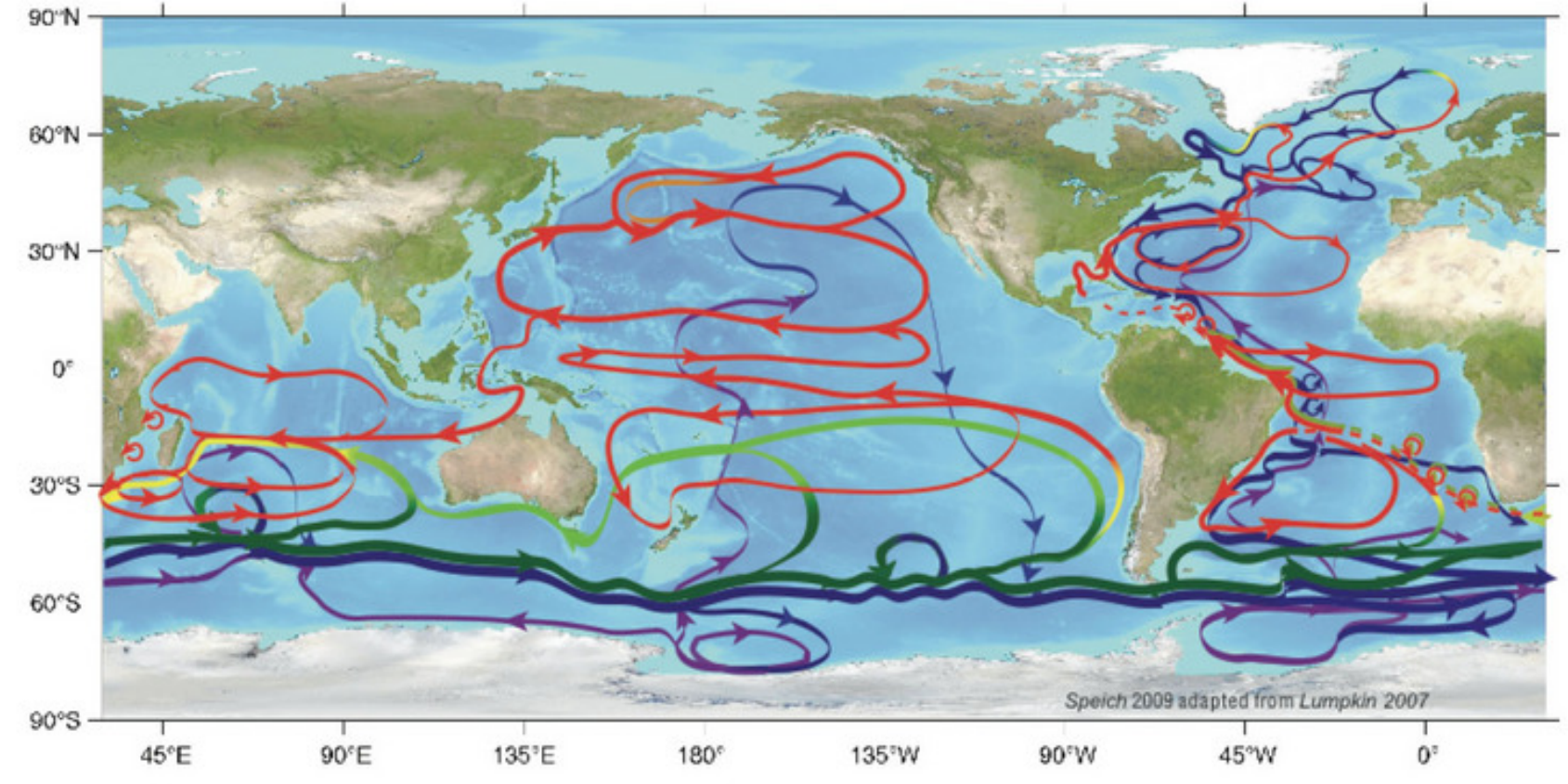
Unsupervised learning: DBSCAN



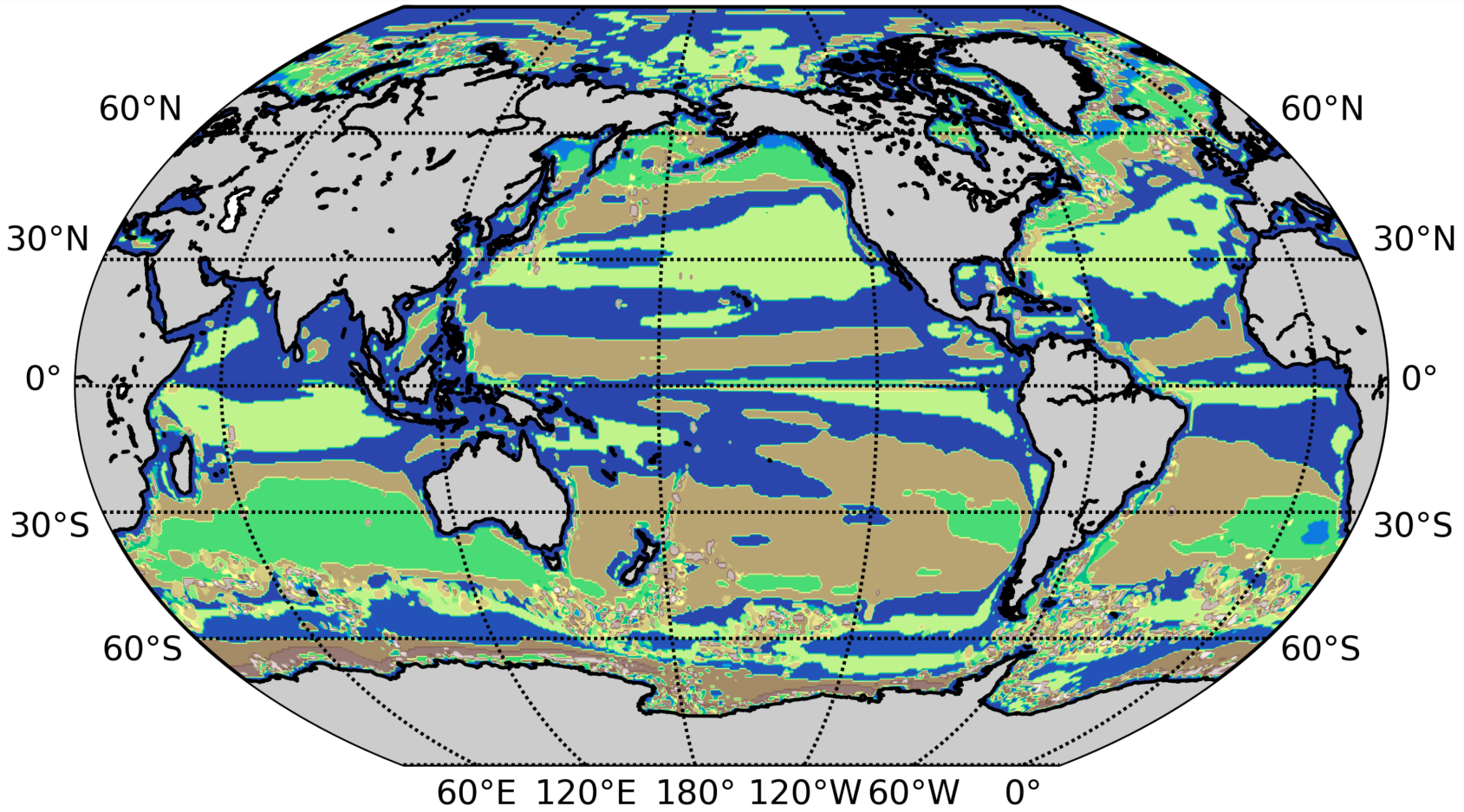
2D "elbow" check in connectedness+cover

Evaluation with domain knowledge

Global ocean dynamical regimes

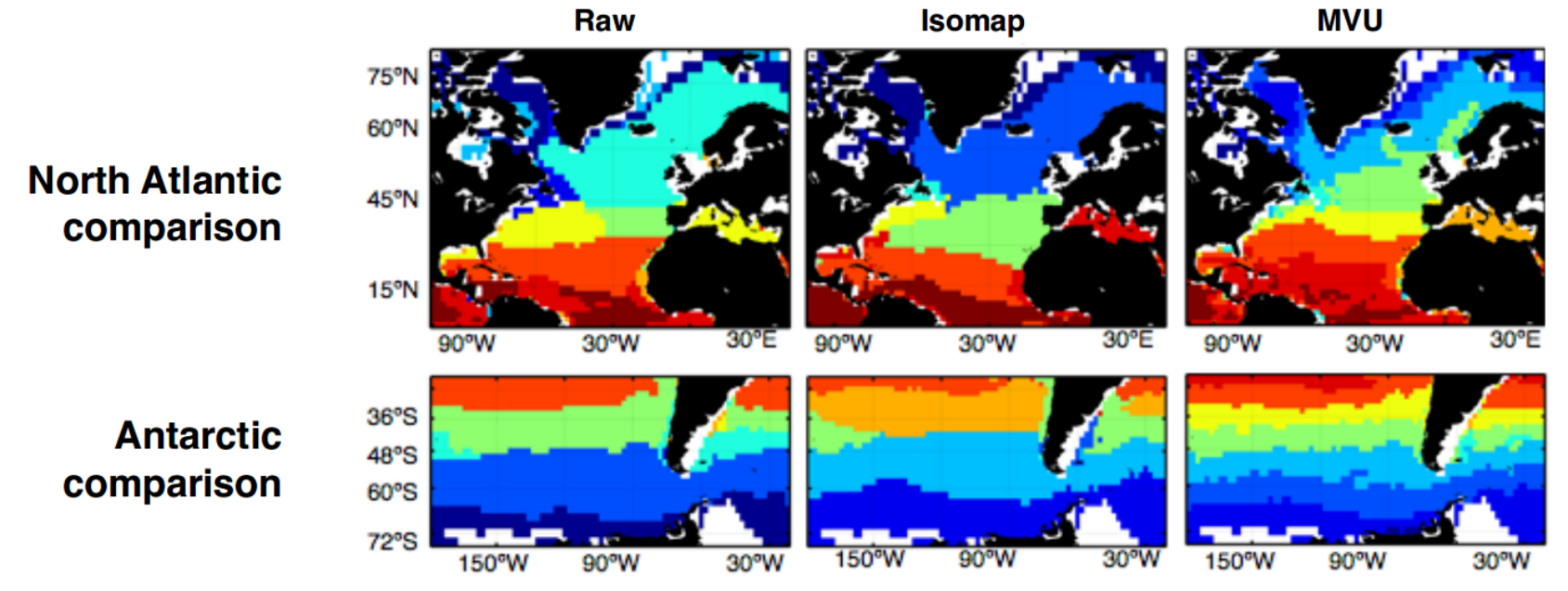


Global ocean pathways



Timely?

Brute force model search?



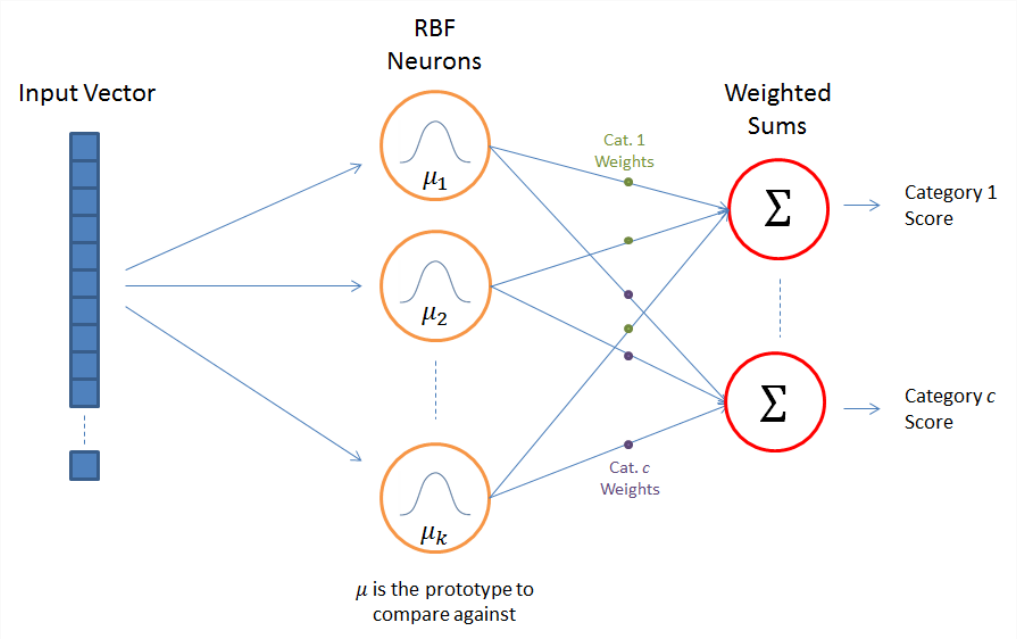
Lewis IEEE, 2008

t-Statistic Neighbourhood Embedding

-L.J.P. van der Maaten. **Accelerating t-SNE using Tree-Based Algorithms**. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.
-L.J.P. van der Maaten and G.E. Hinton. Visualizing Non-Metric Similarities in Multiple Maps. Machine Learning 87(1):33-55, 2012.
-L.J.P. van der Maaten. Learning a Parametric Embedding by Preserving Local Structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP 5:384-391, 2009.
-L.J.P. van der Maaten and G.E. Hinton. **Visualizing High-Dimensional Data Using t-SNE**. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

Supervised+unsupervised

E.g. Radial Basis Functions have a "basis" defined by k-Means:



McCormick

Variances and centres chosen using k-Means