Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

@fedhere

XI: Categorical Clustering
Kriging

Last Class!!!!!

Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- SQL
- Basic statistics: distributions and their moments
- Hypothesis testing: *p*-value, statistical significance
- Statistical and Systematic errors
- Goodness of fit tests
- Likelihood
- OLS
- Topics in (time) series analysis
- Visualizations
- Geospatial analysis
- Clusters

Today:

- categorical and mixed clustering
- kriging and gaussian processes
- efficient coding

Summary and Key concepts

clustering is easy, but interpreting results is tricky

Distance metrics:

- Eucledian and other Minchowski metrics
- geospacial distances
- metrics for non continuous data

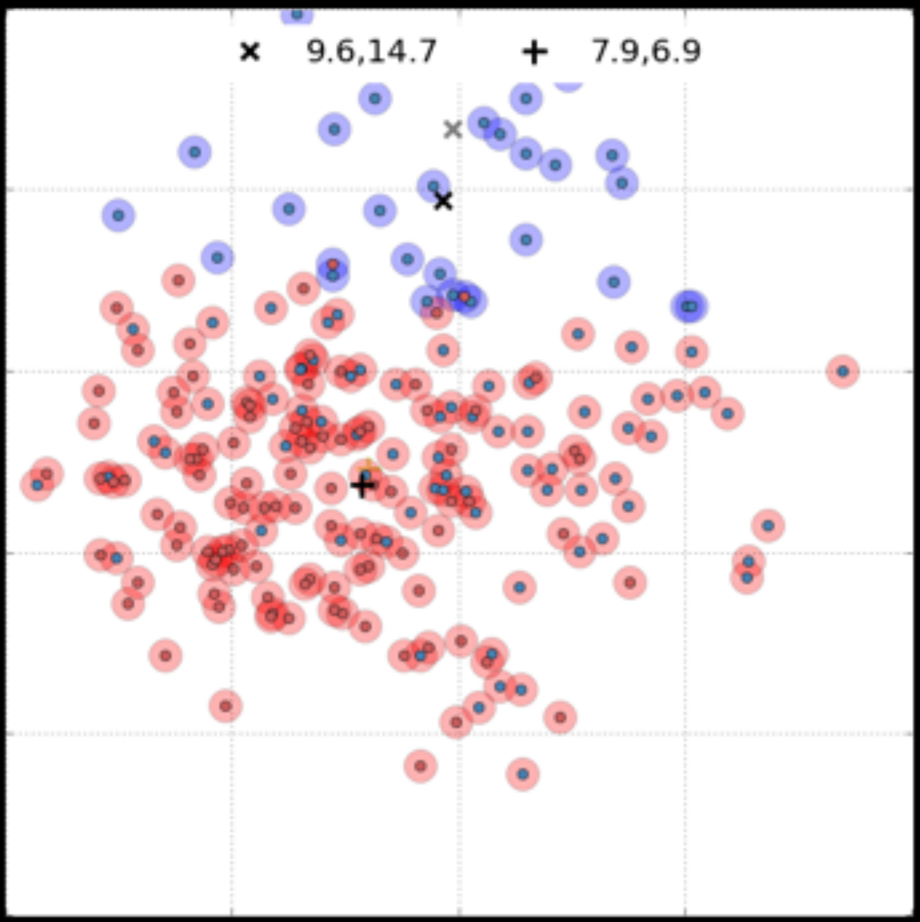
Partitioning methods: inexpensive, typically non deterministic

- Hard methods: *K-means, K-medoids*
- Soft (or fuzzy) methods: (i.e. probabilistic approach)
  - *Expectation Maximization Mixture models*

Hierarchical methods:

- divisive vs agglomerative, dendrograms

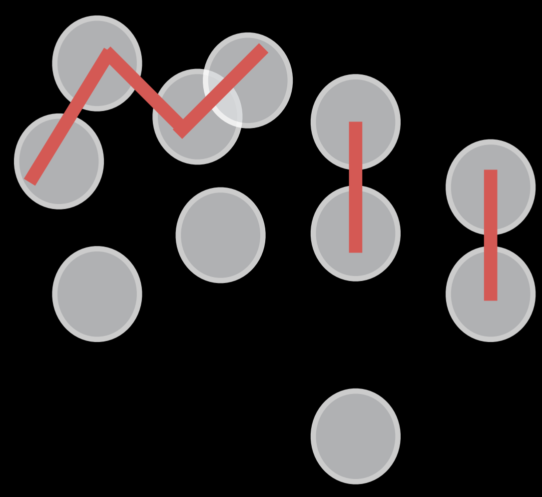
Crisp (or hard) clustering - K-means

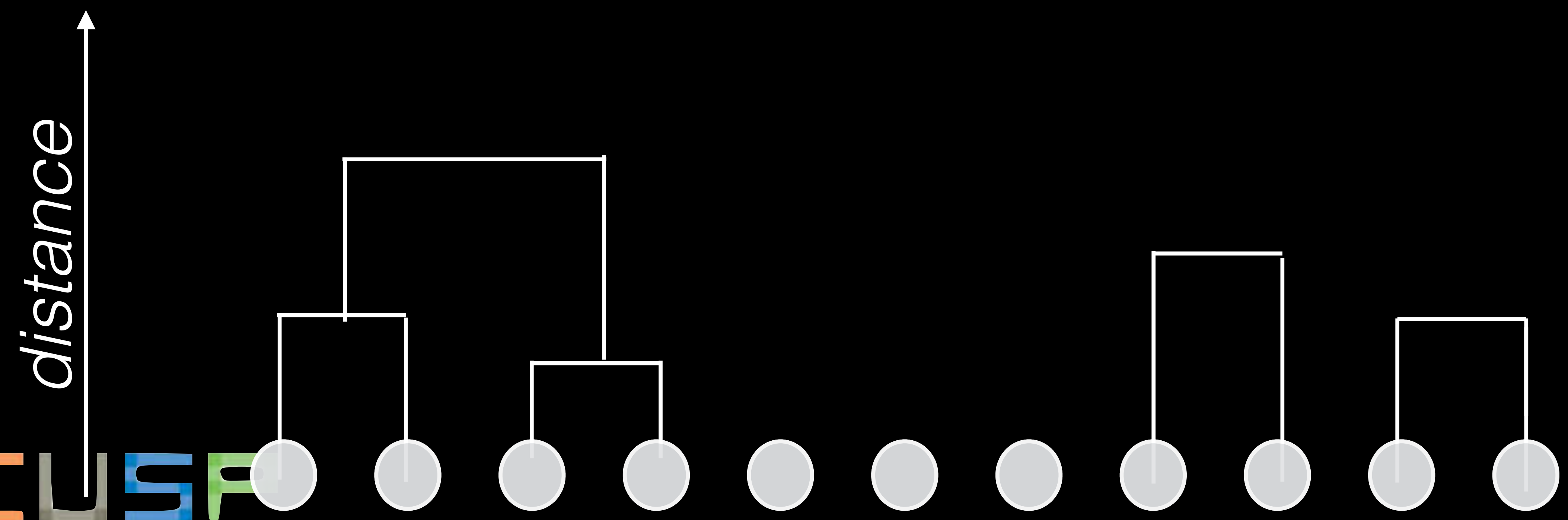


You guess the centers and assign points to clusters based on a predefined distance metric

hierarchical clustering

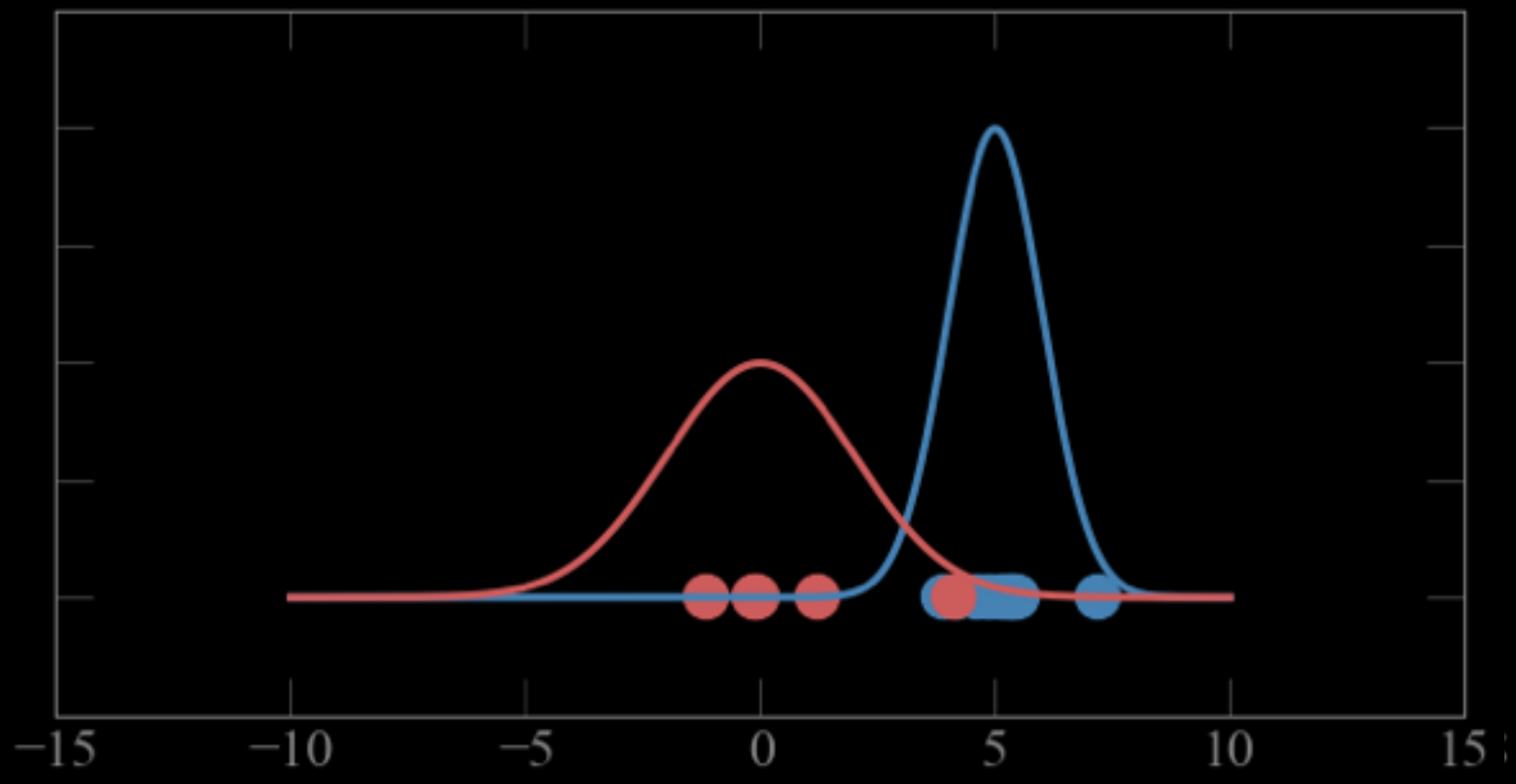
agglomerative
bottom-up





Fuzzy (or soft) clustering - Mixture models

A probabilistic way to do clustering



You adjust the parameters (μ,σ) of the gaussians iteratively based on the probability of the data coming from that gaussian

Hard Clustering:

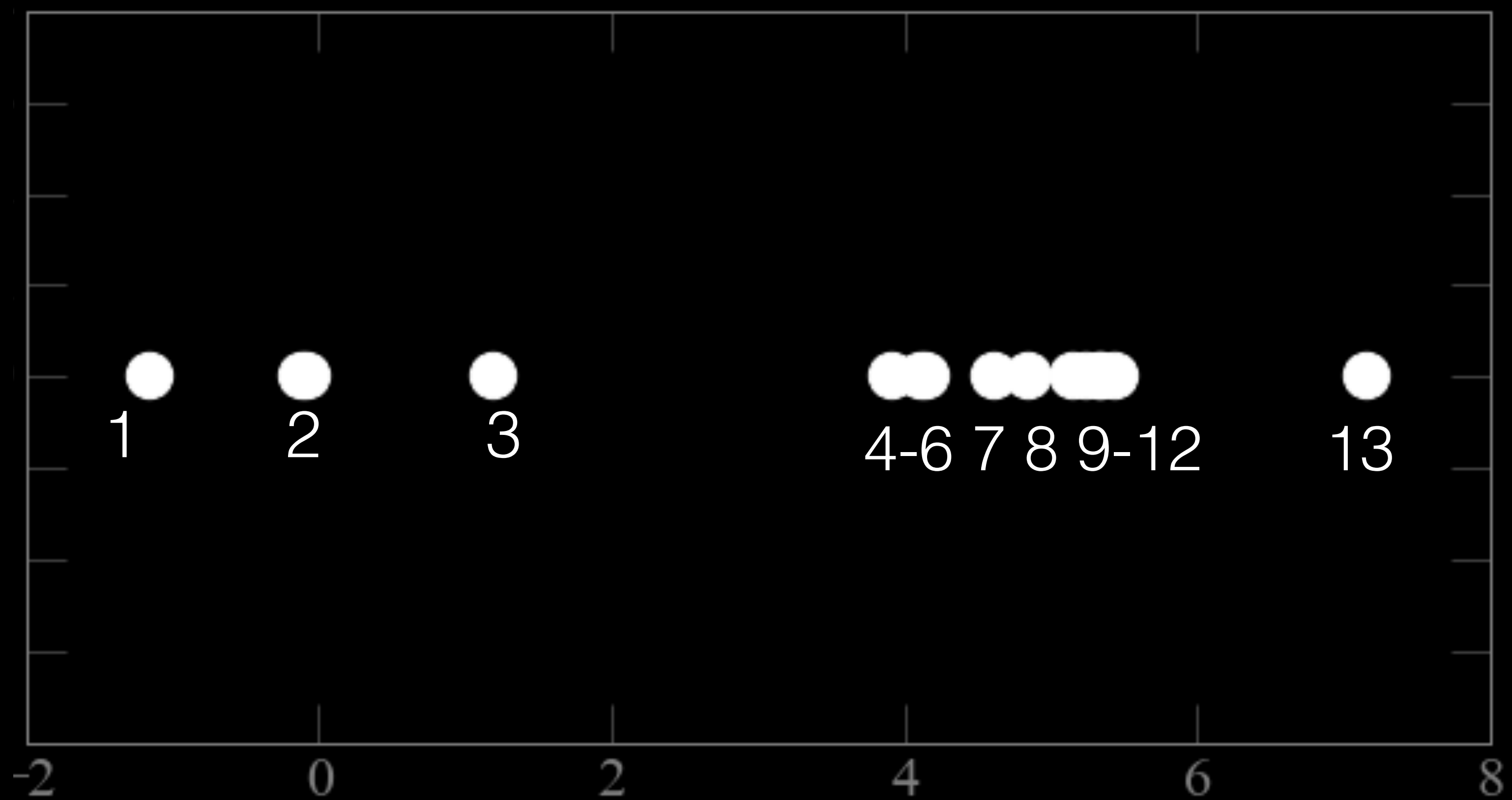
each object in the sample belongs to only 1 cluster

Soft Clustering:

to each object in the sample we assign a degree of belief that it belongs to a cluster

Mixture models

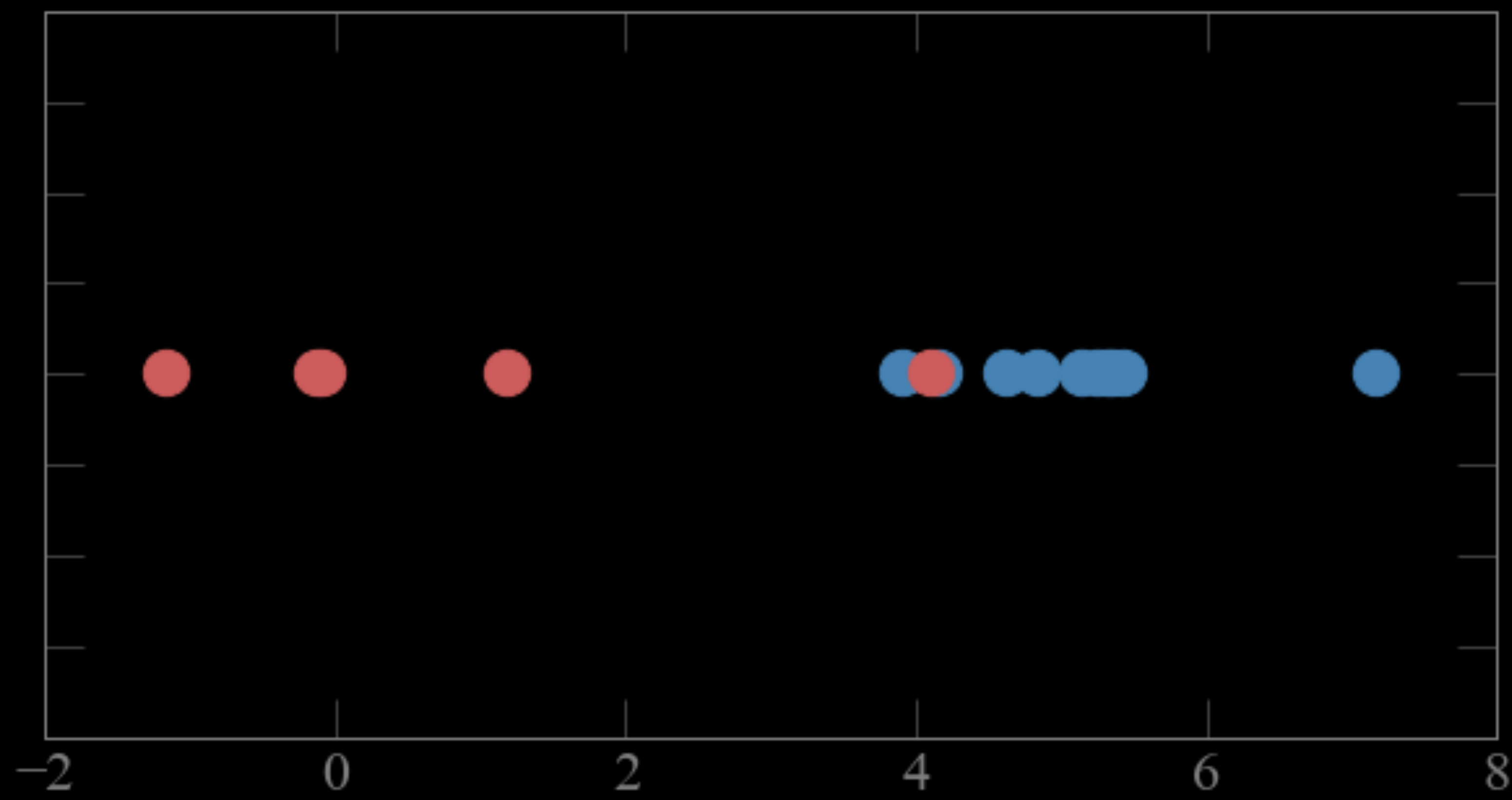
A probabilistic way to do soft clustering



these points come from 2 gaussian distribution.
which point comes from which gaussian?

Mixture models

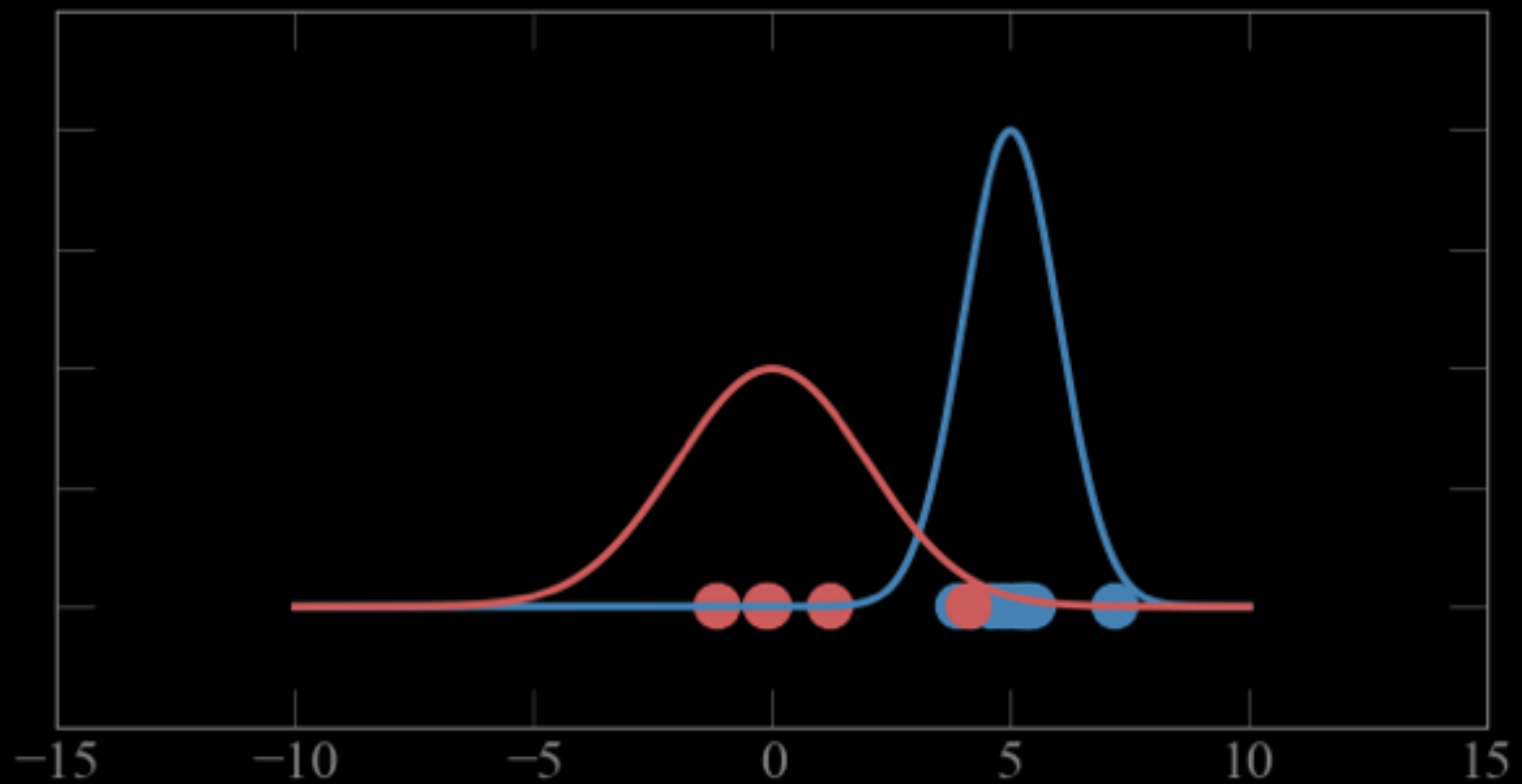
A probabilistic way to do soft clustering



if i know which point comes from which gaussian
i can solve for the parameters of the gaussian
(e.g. maximizing likelihood)

Mixture models

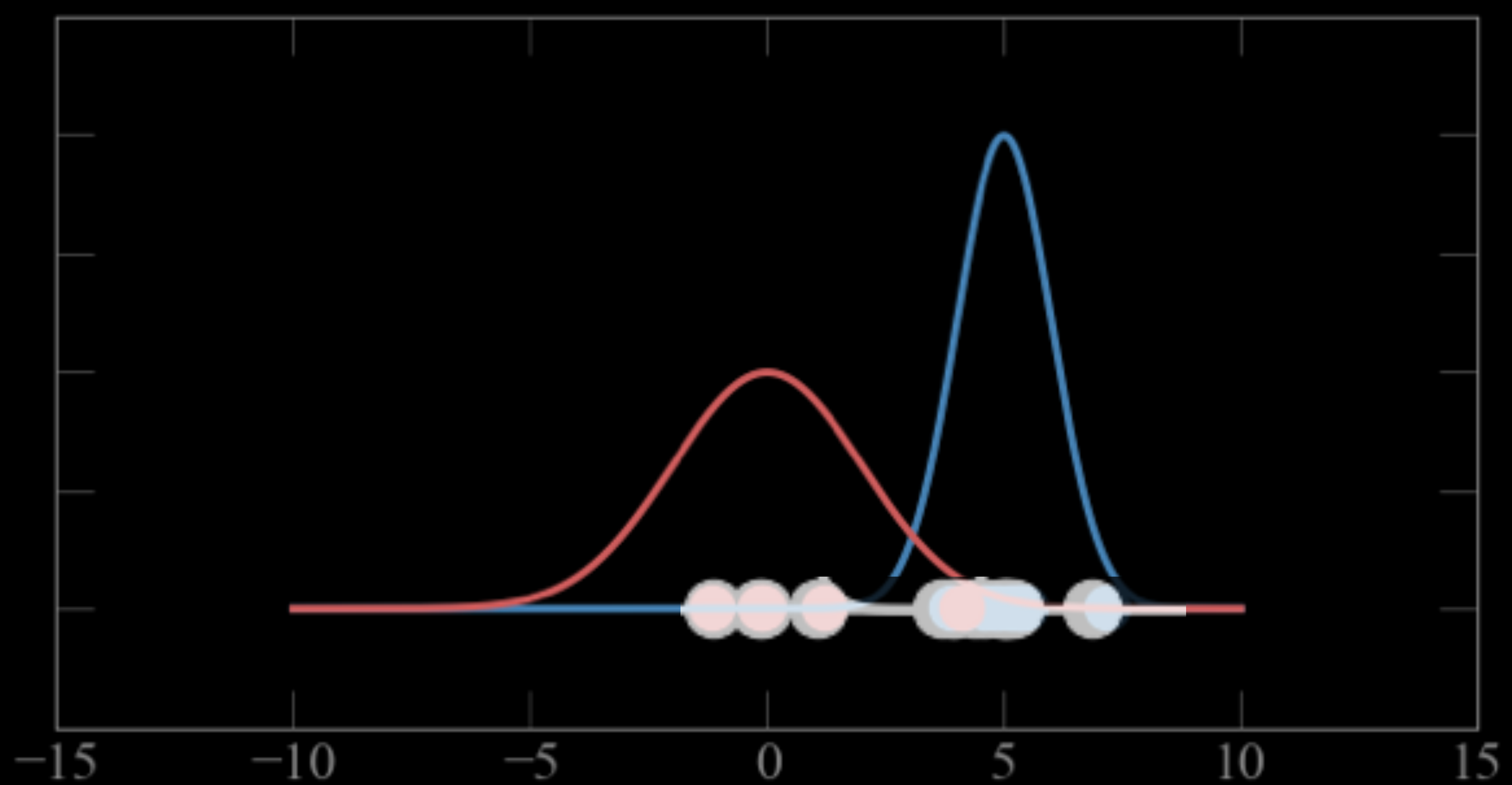
A probabilistic way to do soft clustering



if i know which the parameters (μ,σ) of the gaussians i can figure out which gaussian each point is most likely to come from (calculate probability)

Mixture models

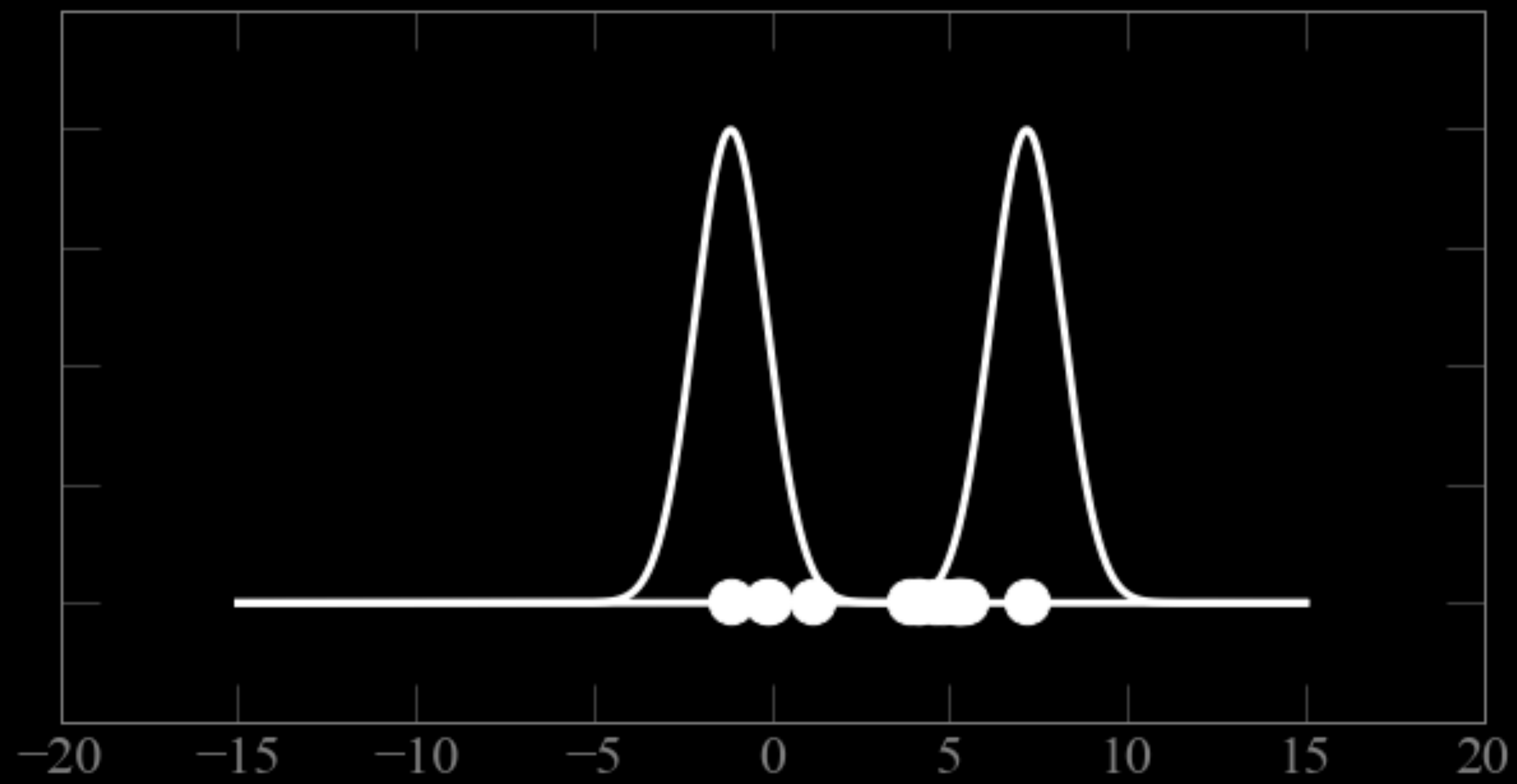
A probabilistic way to do soft clustering



if i know which the parameters (μ,σ) of the gaussians i can figure out which gaussian each point is most likely to come from (calculate probability)

EM

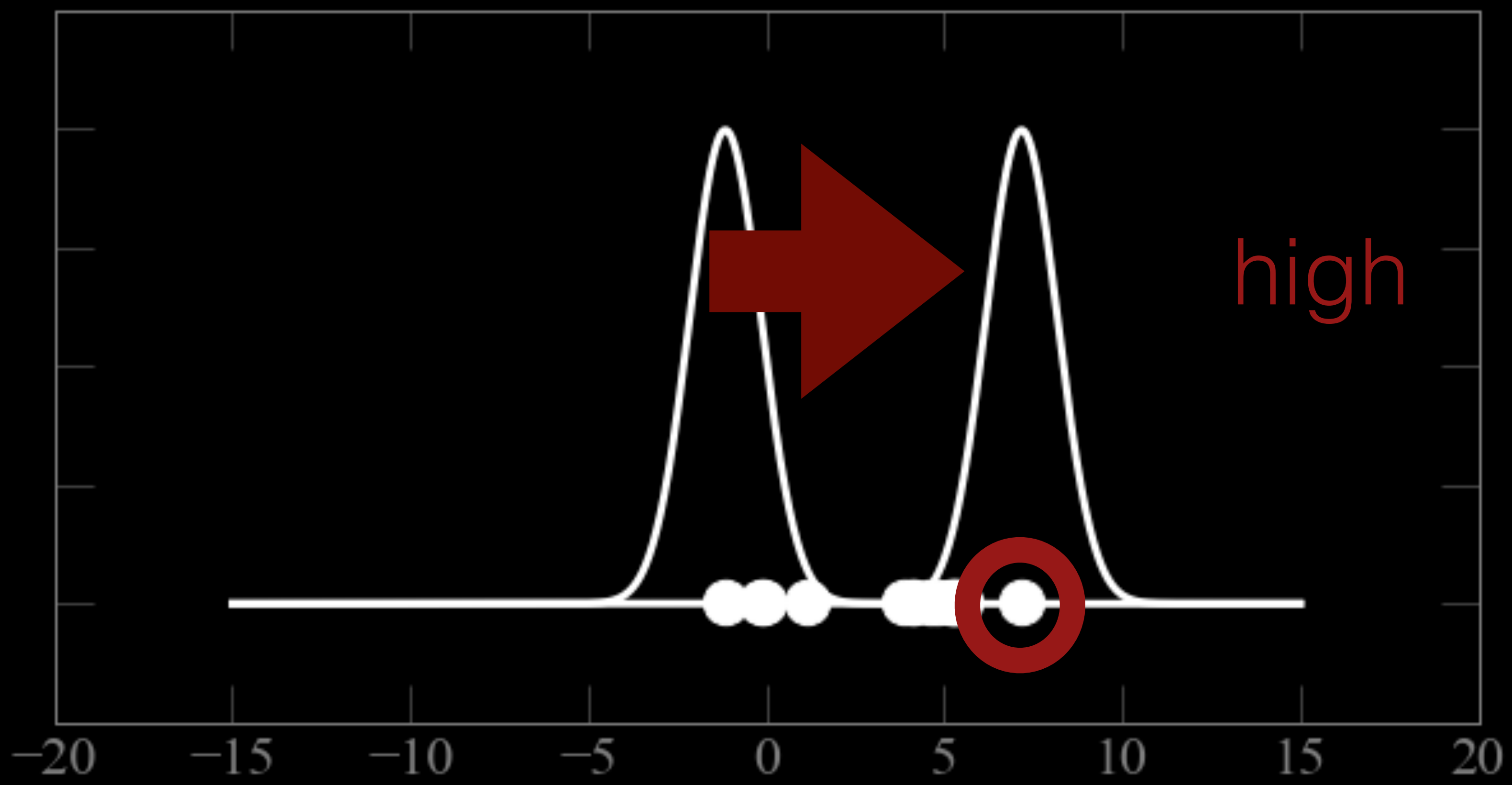
$$P(x_i|\mu_j,\sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{x_i-\mu_j}{2\sigma_j^2}\right)$$



for every point calculate the probability it comes from either gaussian

EM

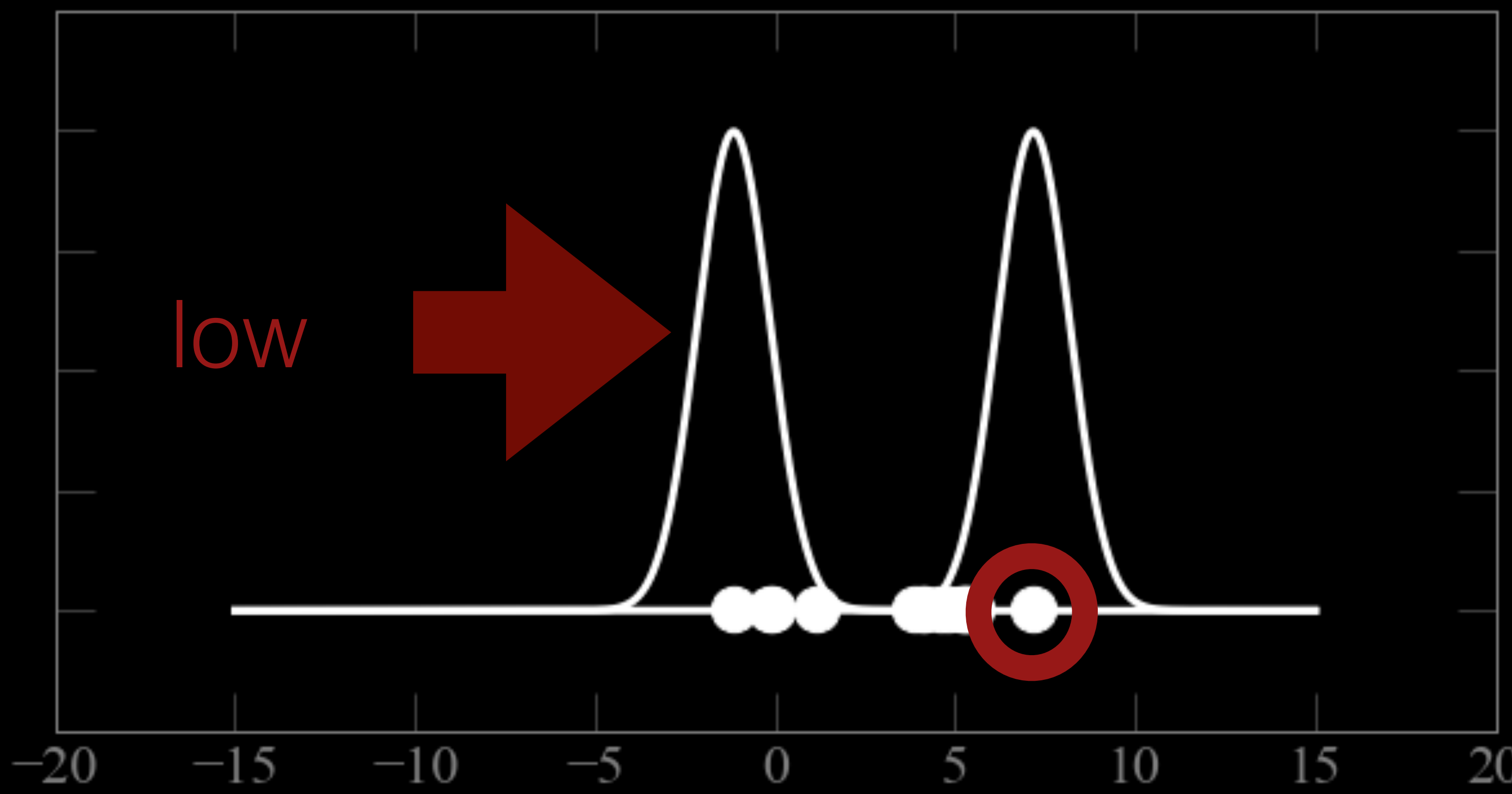
$$P(x_i \mid \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(- \frac{x_i - \mu_j}{2\sigma_j^2} \right)$$



for every point calculate the probability it comes from either gaussian

EM

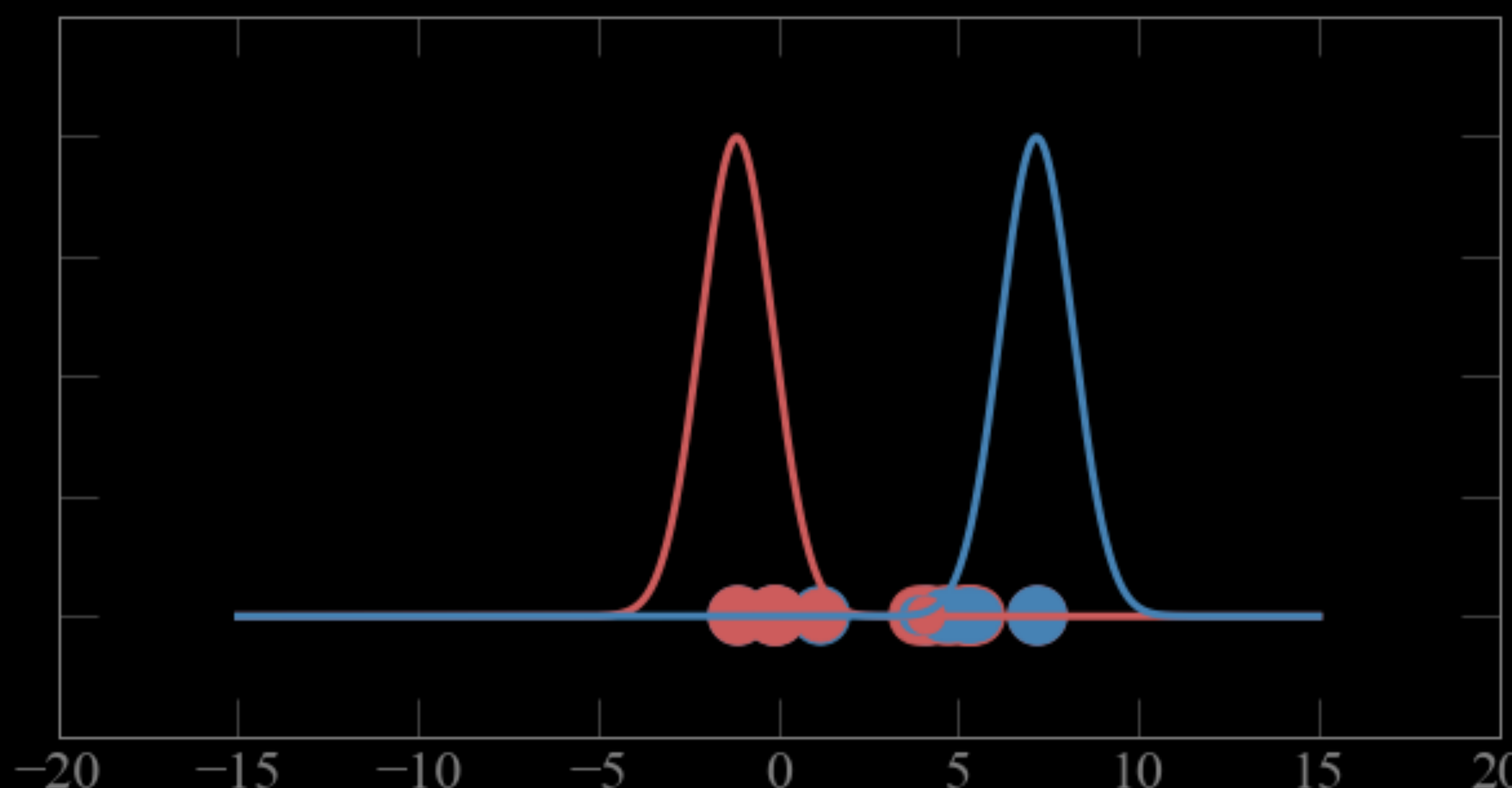
$$P(x_i | \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(- \frac{x_i - \mu_j}{2\sigma_j^2}\right)$$



for every point calculate the probability it comes from either gaussian

EM

$$P(x_i \mid \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(- \frac{x_i - \mu_j}{2\sigma_j^2} \right)$$

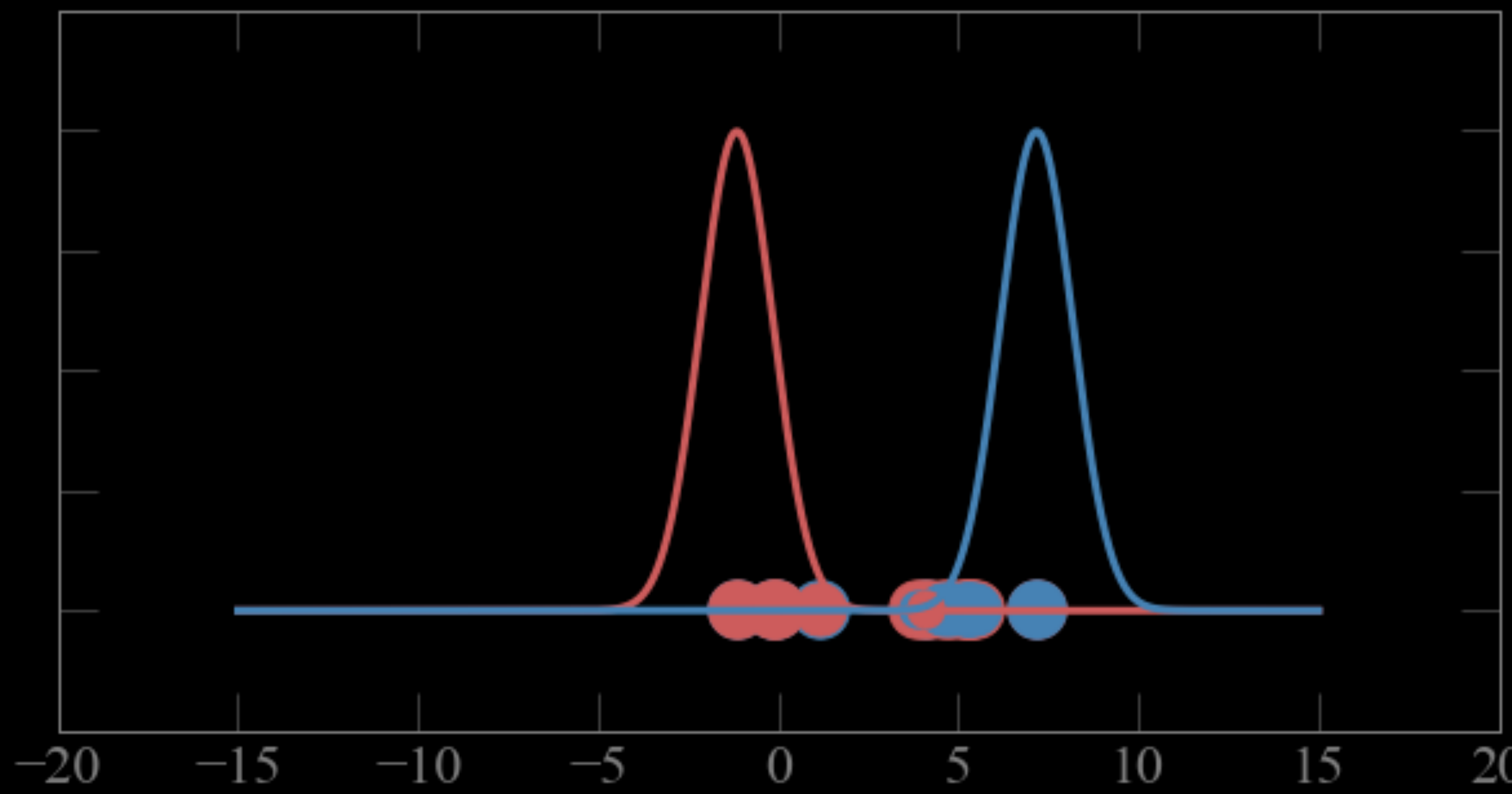


for every point calculate the probability it comes from either gaussian

EM

$$P(x_i|\mu_j,\sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{x_i-\mu_j}{2\sigma_j^2}\right)$$

$$P(\mu_1,\sigma_1|x_i) = \frac{P(x_i|\mu_1,\sigma_1)P(\mu_1,\sigma_1)}{P(x_i|\mu_1,\sigma_1)P(\mu_1,\sigma_1)+P(x_i|\mu_2,\sigma_2)P(\mu_2,\sigma_2)}$$

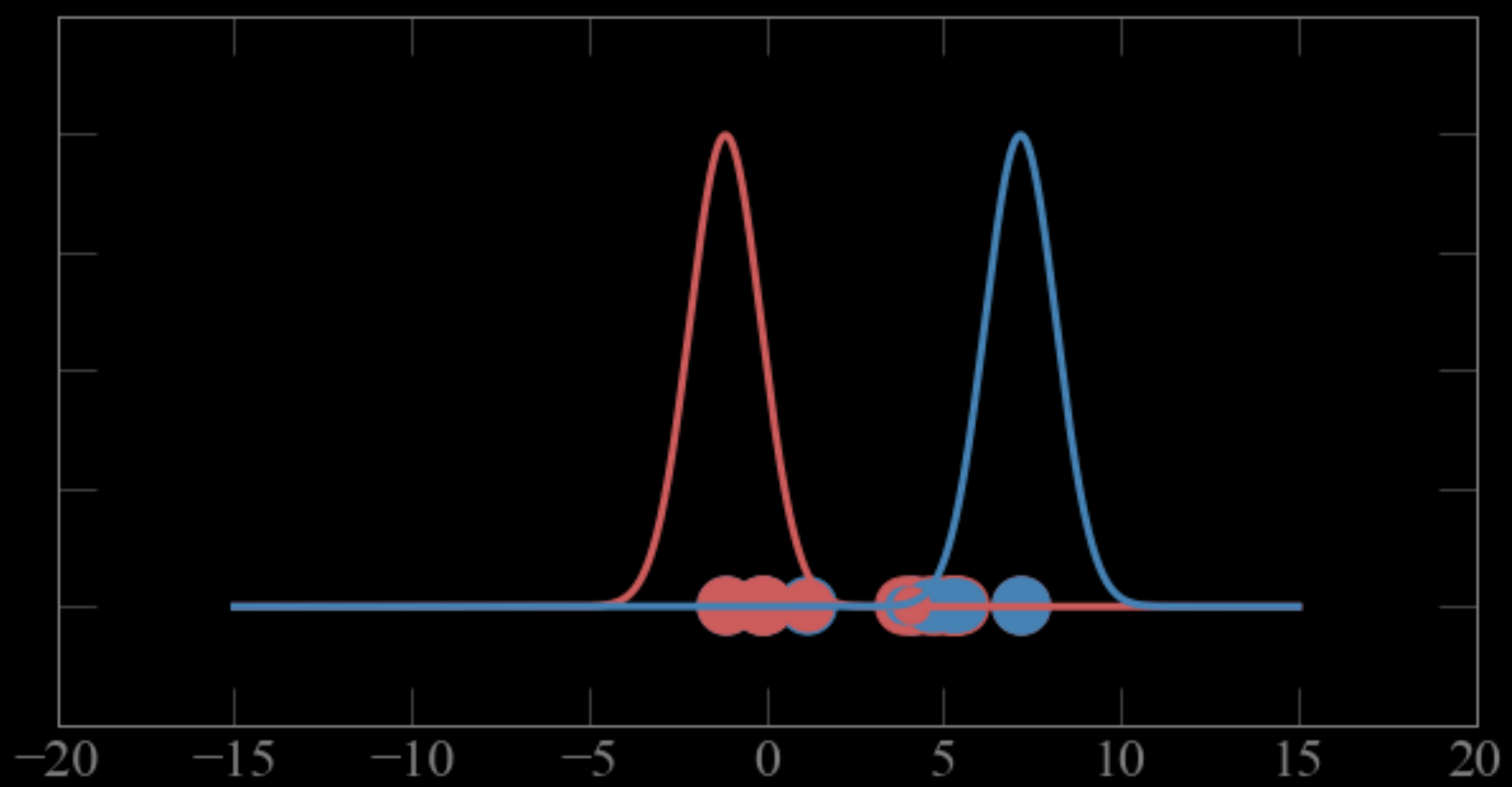


for every point calculate the probability it comes from either gaussian

EM

$$P(x_i|\mu_j,\sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(- \frac{x_i - \mu_j}{2\sigma_j^2}\right)$$

$$P(g_1|x_i) = \frac{P(x_i|g_1)P(g_1)}{P(x_i|g_1)P(g_1)+P(x_i|g_2)P(g_2)}$$

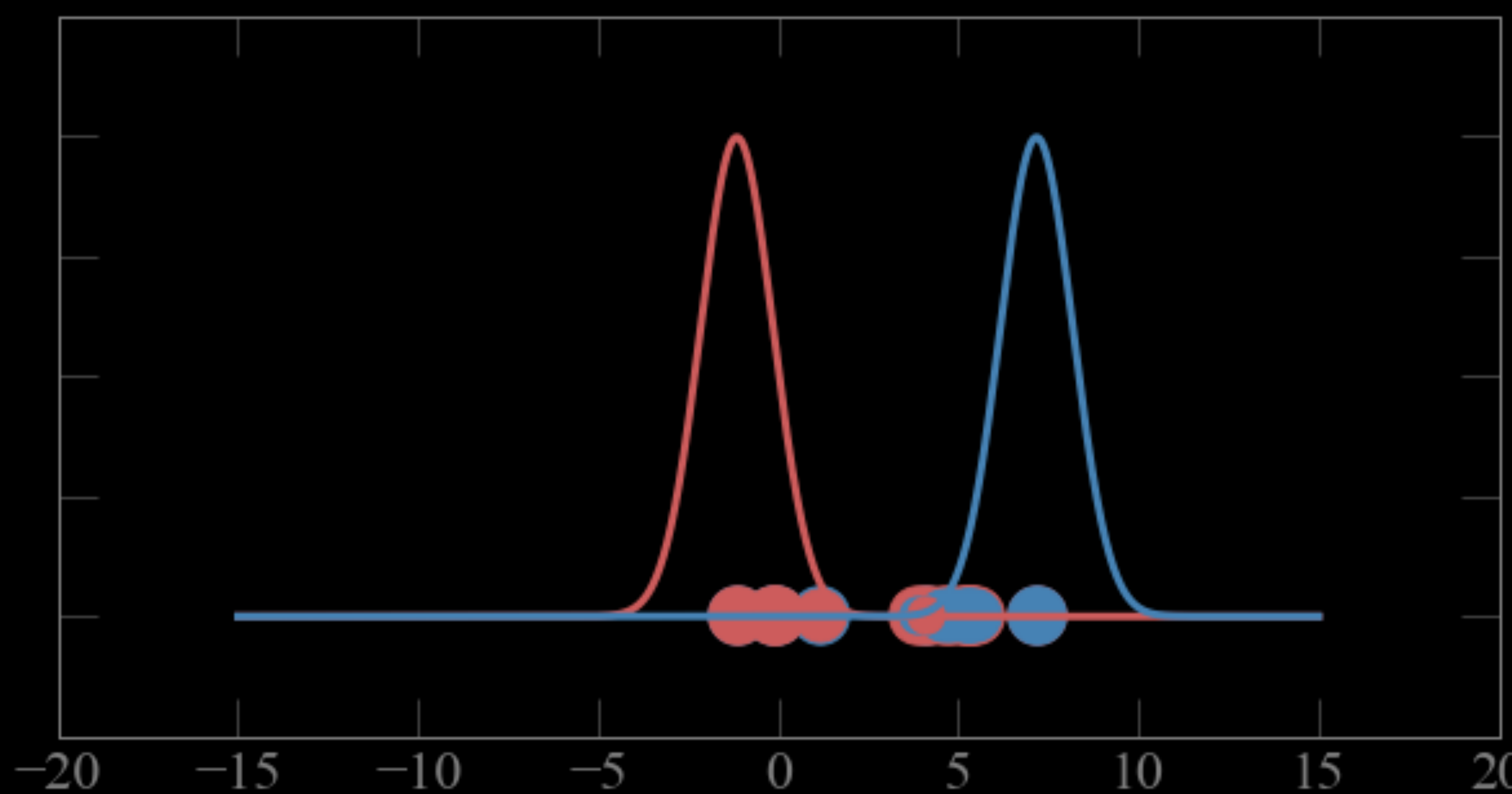


for every point calculate the probability it comes from either gaussian

EM

Bayes Theorem!

$$P(x|\alpha)P(\alpha) = P(x|\beta)P(\beta)$$

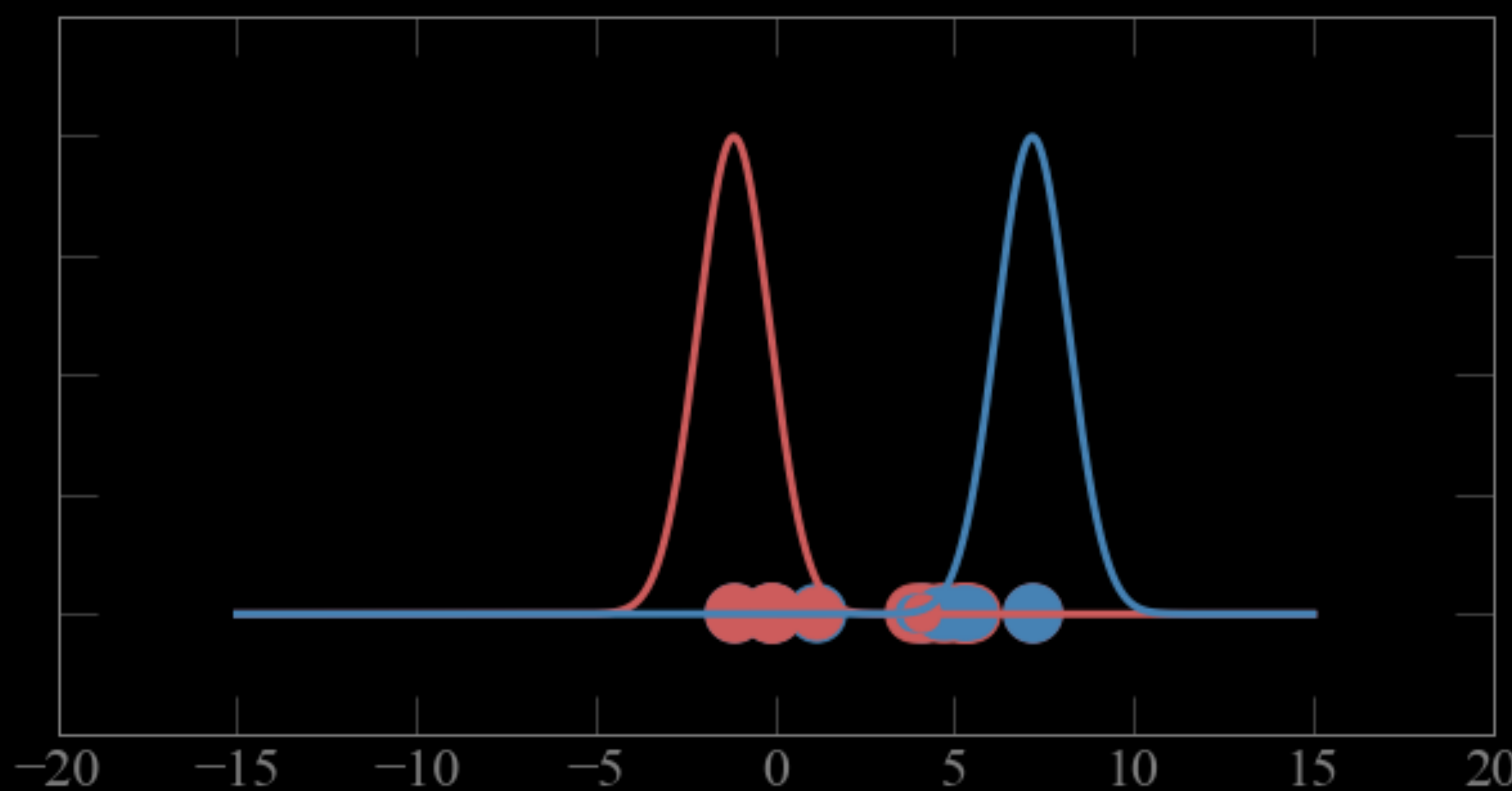


for every point calculate the probability it comes from either gaussian

EM

$$P(x_i|\mu_j,\sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(- \frac{x_i - \mu_j}{2\sigma_j^2} \right)$$

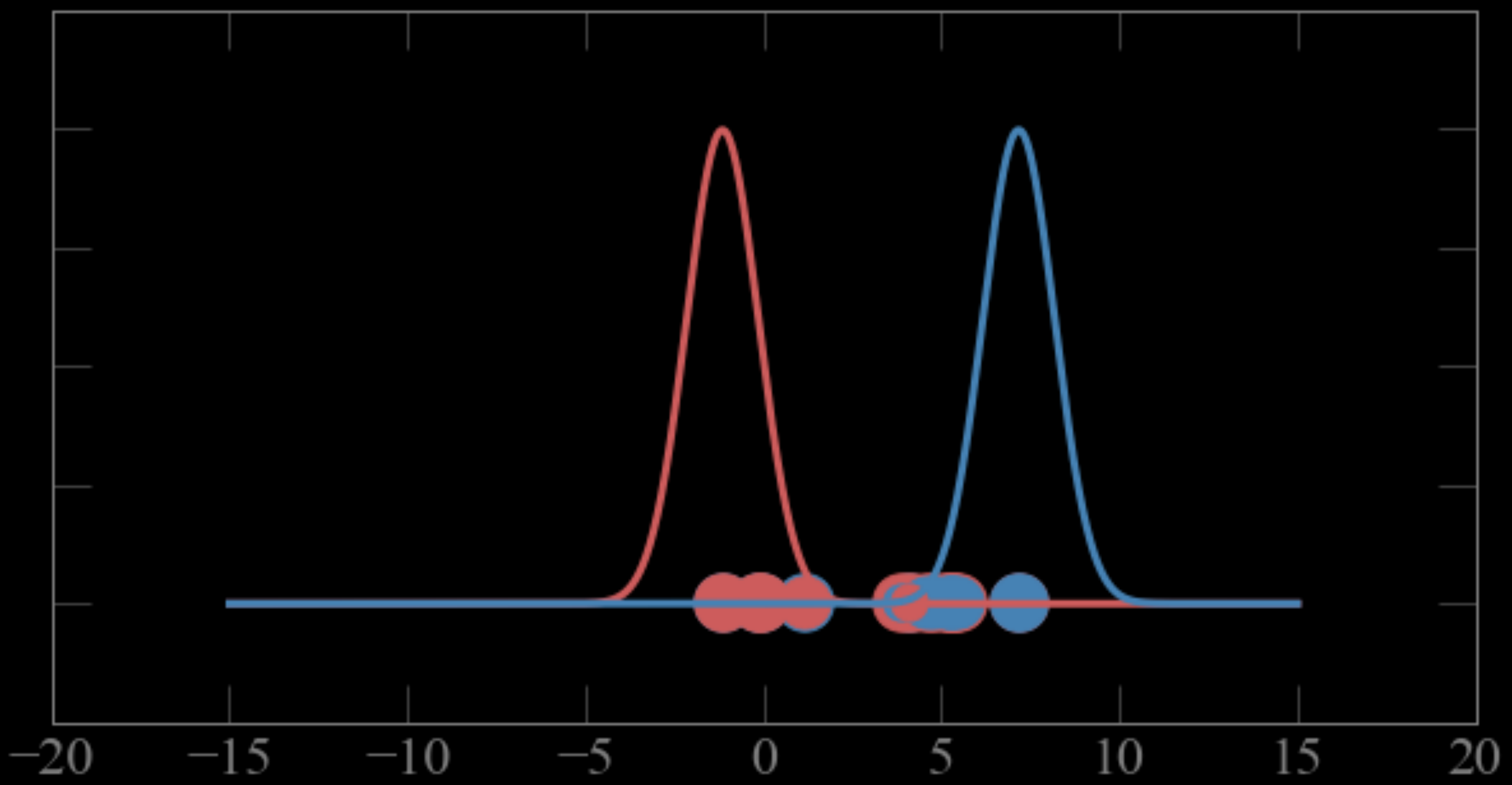
p_ji = $P(g_1|x_i) = \dfrac{P(x_i|g_1)P(g_1)}{P(x_i|g_1)P(g_1)+P(x_i|g_2)P(g_2)}$



calculate the weighted mean of the cluster, weighted by the p_ji

EM

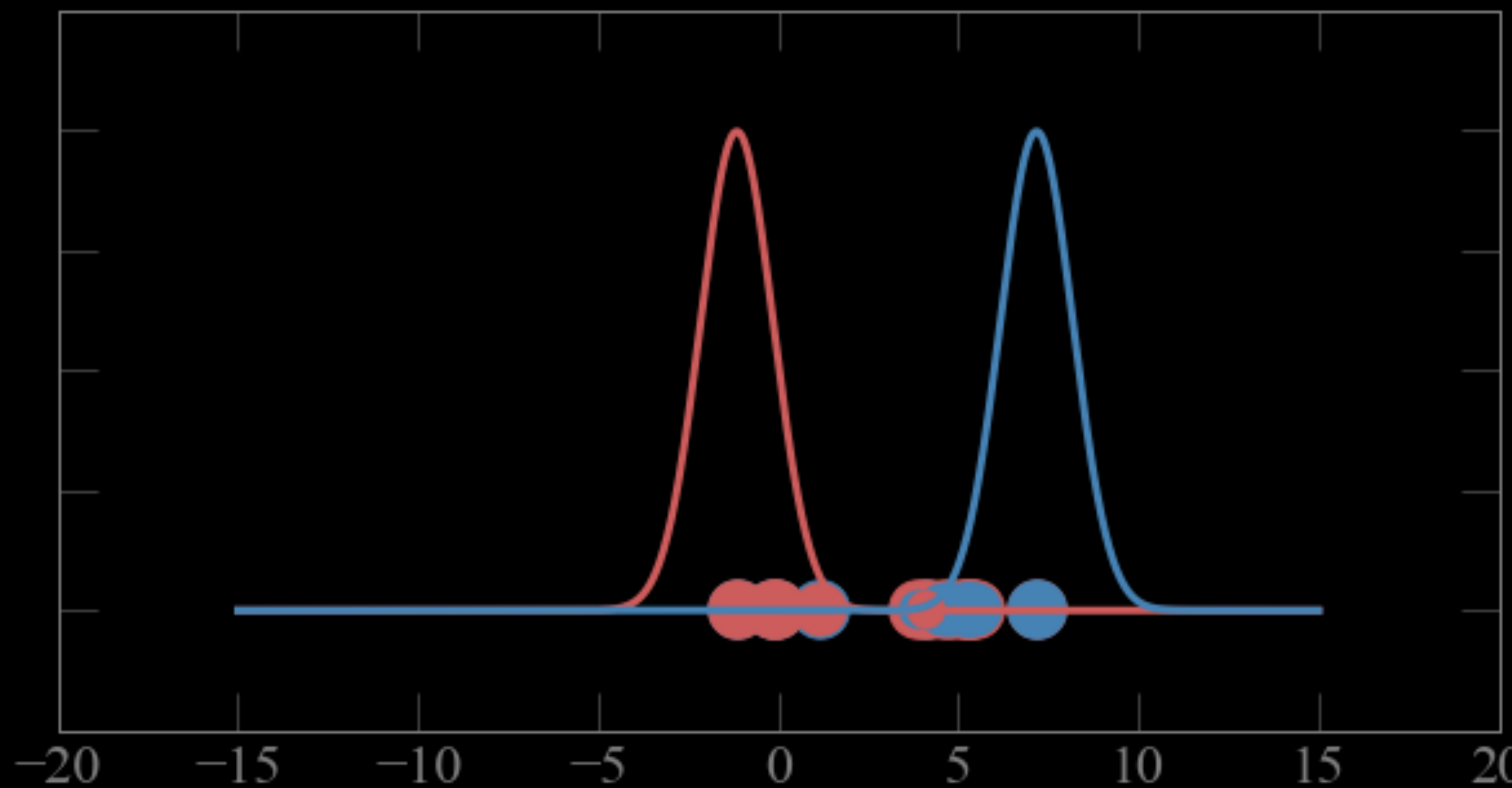
$$\mu_i = \frac{\sum_j P(g_i|x_j)x_j}{\sum_j P(g_i|x_j)}$$



calculate the weighted mean of the cluster, weighted by the p_ji

EM

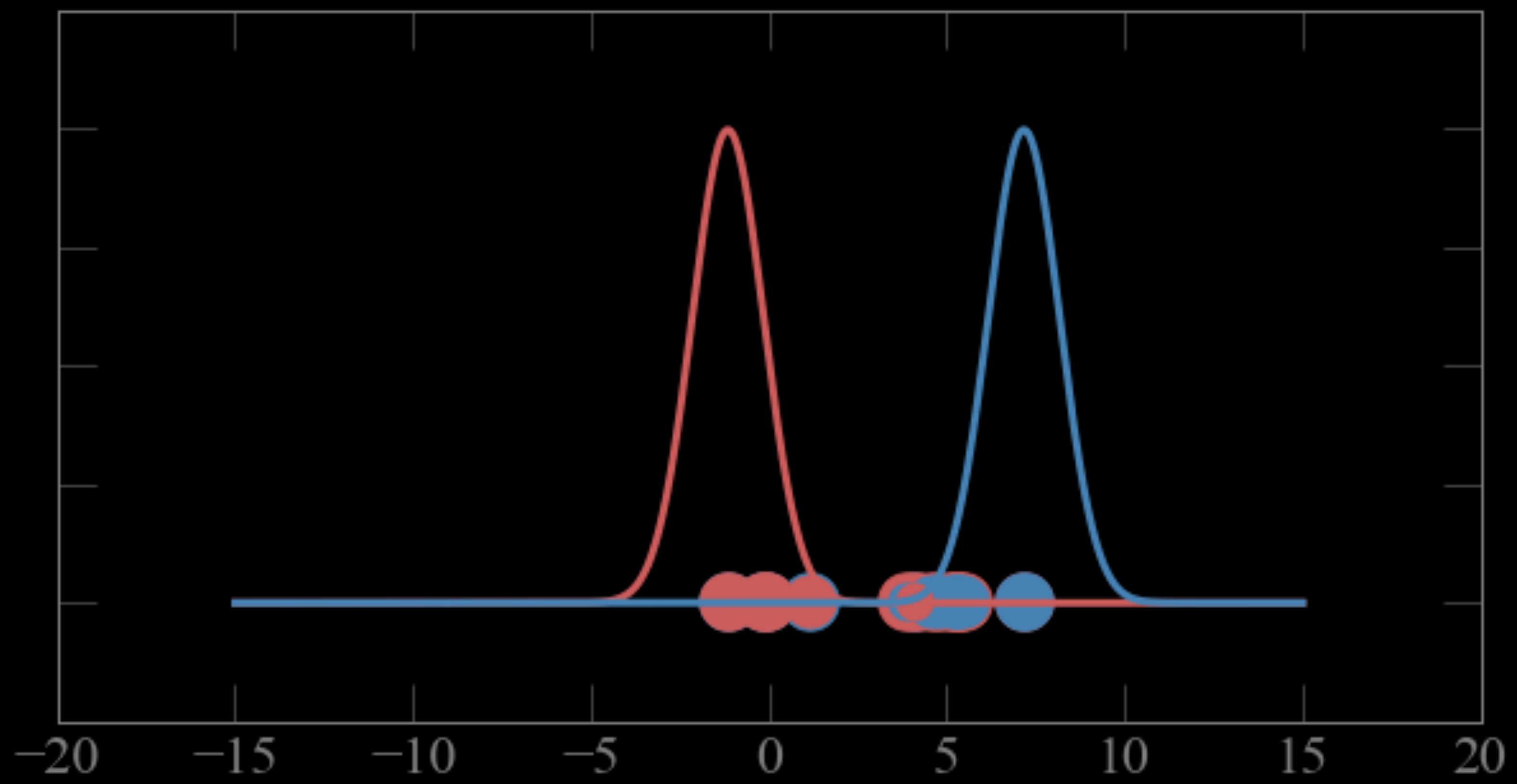
$$\mu_i = \frac{\sum_j P(g_i|x_j)x_j}{\sum_j P(g_i|x_j)} \qquad \sigma\_j = \frac{\sum_i P(g_j|x_i)(x_i - \mu_j)^2}{\sum_i P(g_j|x_i)}$$



calculate the weighted sigma of the cluster, weighted by the p_ji

EM

$$\mu_i = \frac{\sum_j P(g_i | x_j) x_j}{\sum_j P(g_i | x_j)} \qquad \sigma\_j = \frac{\sum_i P(g_j | x_i)(x_i - \mu_j)^2}{\sum_i P(g_j | x_i)}$$

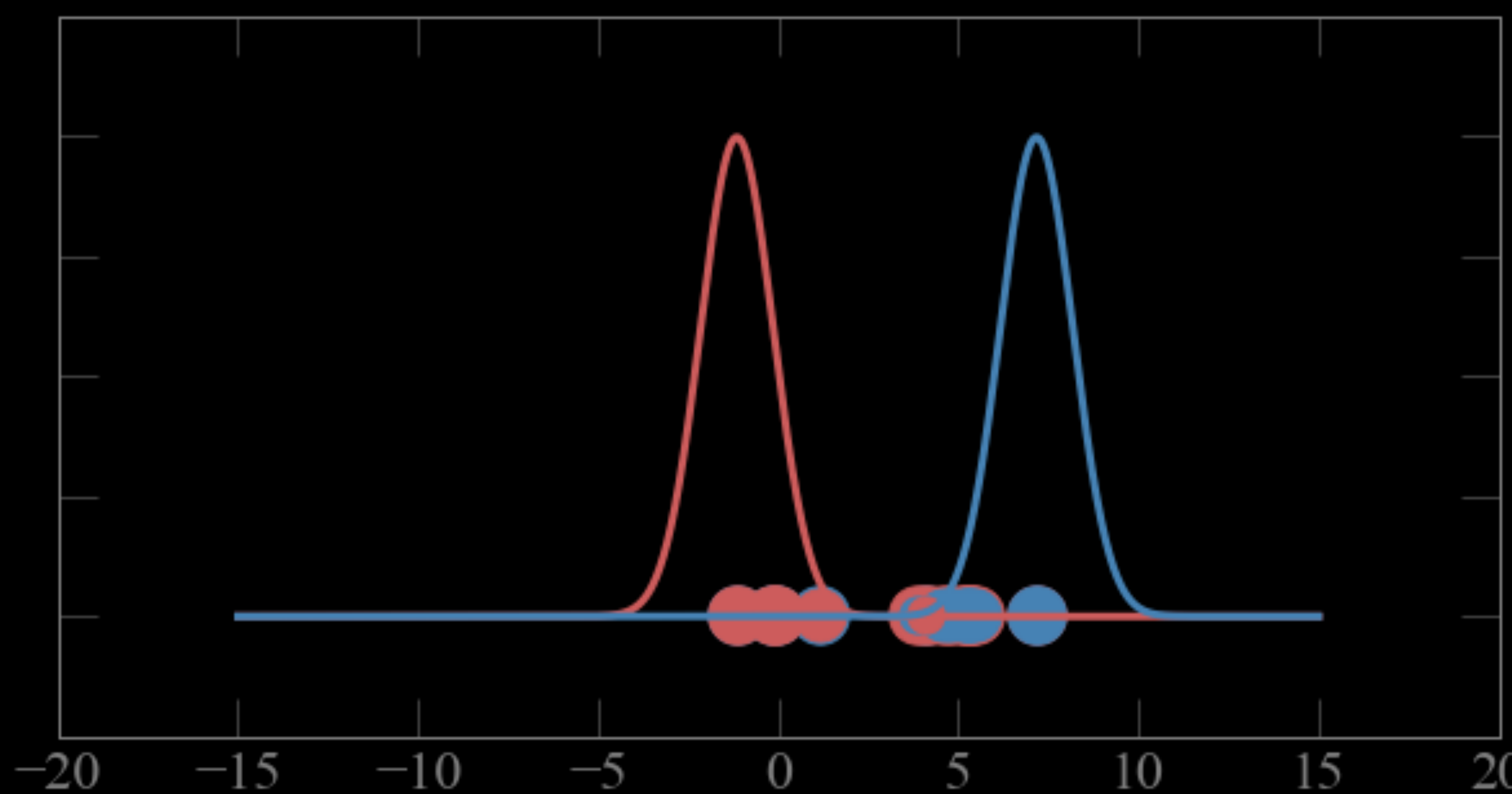


$$P(g_1|x_i) = \frac{P(x_i | g_1) P(g_1)}{P(x_i | g_1) P(g_1) + P(x_i | g_2) P(g_2)}$$

calculate the new p_ji … rinse, repeat

EM

$$\mu_i = \frac{\sum_j P(g_i|x_j)x_j}{\sum_j P(g_i|x_j)} \qquad \sigma\_j = \frac{\sum_i P(g_j|x_i)(x_i-\mu_j)^2}{\sum_i P(g_j|x_i)}$$

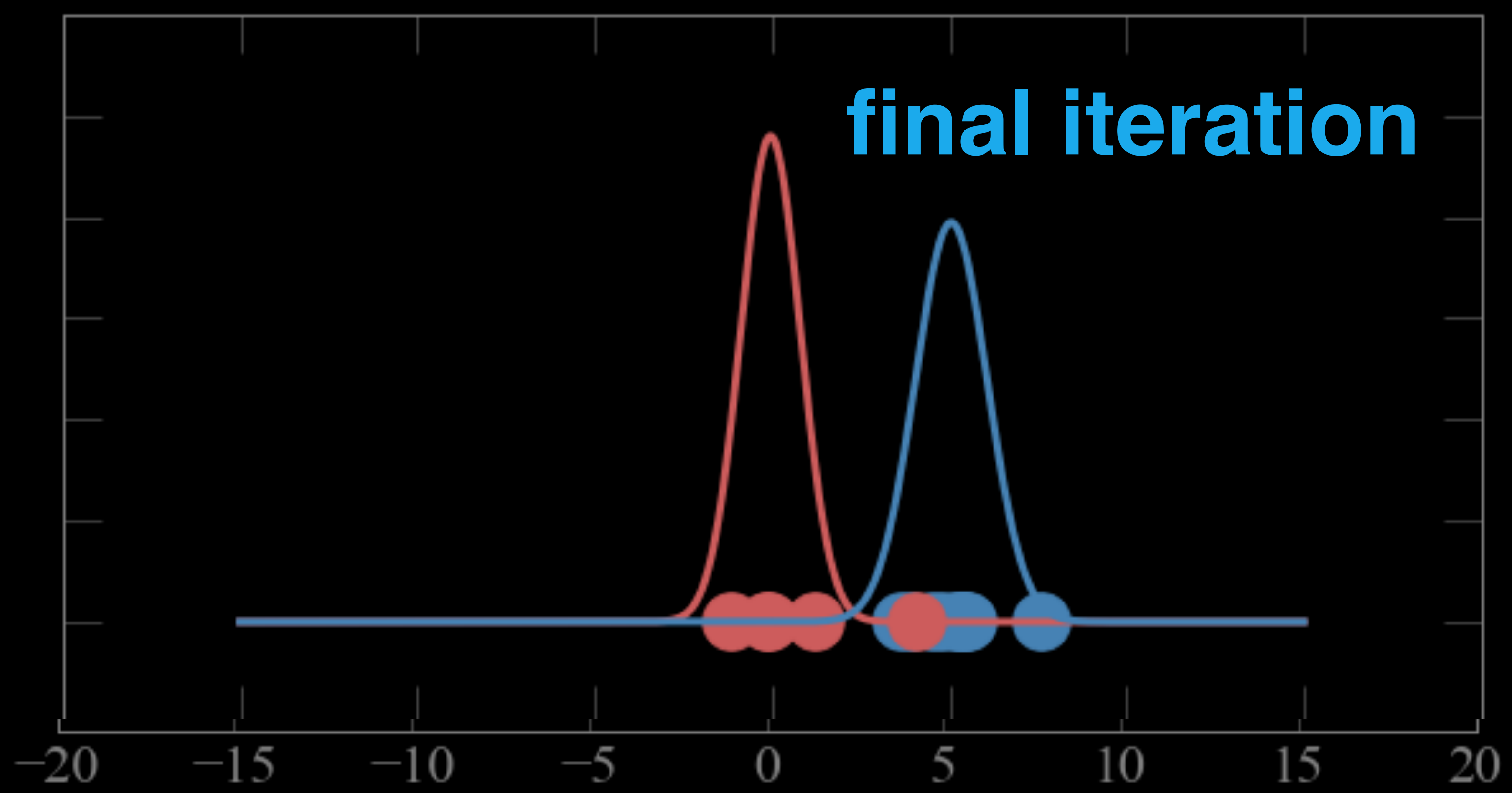


$$P(g_1|x_i) = \frac{P(x_i|g_1)P(g_1)}{P(x_i|g_1)P(g_1)+P(x_i|g_2)P(g_2)}$$

calculate the new p_ji … rinse, repeat

EM

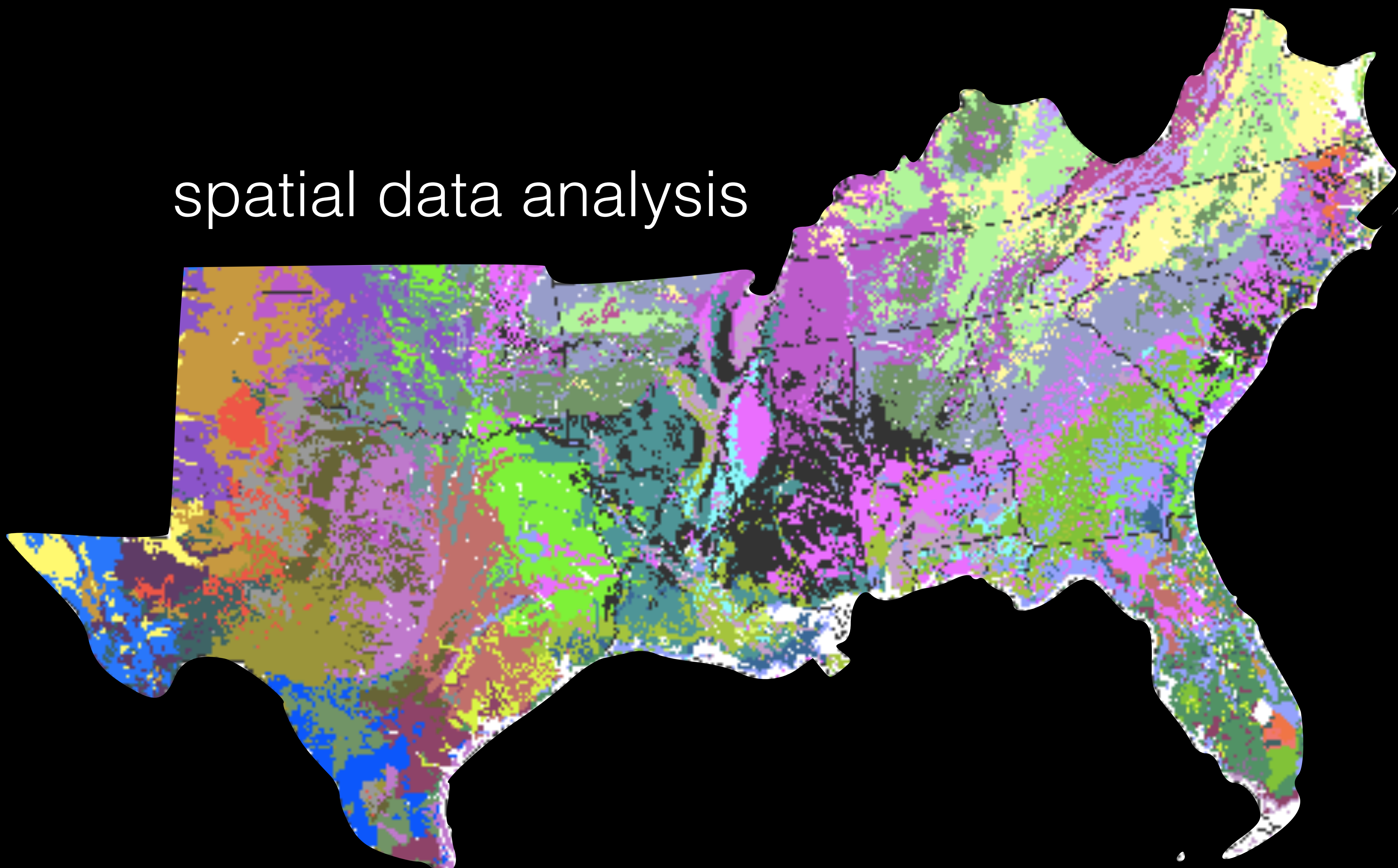
$$\mu_i = \frac{\sum_j P(g_i|x_j)x_j}{\sum_j P(g_i|x_j)} \qquad \sigma\_j = \frac{\sum_i P(g_j|x_i)(x_i - \mu_j)^2}{\sum_i P(g_j|x_i)}$$



$$P(g_1|x_i) = \frac{P(x_i|g_1)P(g_1)}{P(x_i|g_1)P(g_1)+P(x_i|g_2)P(g_2)}$$

… till it converges

spatial data analysis



A Spatial Clustering Technique for the Identification of Customizable Ecoregions

William W. Hargrove and Robert J. Luxmoore

50-year mean monthly temperature, 50-year mean monthly precipitation, elevation, total plant-available water content of soil, total organic matter in soil, and total Kjeldahl soil nitrogen

CUSP

Distance Metrics Continuous variables

Minkowski family of distances

$$D(i,j) = \sqrt[1/p]{\sum_{k=1}^{N} |x_{ik} - x_{jk}|^p}$$

N features (dimensions)

Great Circle distances: $\phi_i, \lambda_i, \phi_j, \lambda_j$

geographical latitude and longitude

$$D(i,j) = R \arccos(\sin\phi_i \cdot \sin\phi_j + \cos\phi_i \cdot \cos\phi_j \cdot \cos(\Delta\lambda))$$

Distance Metrics Continuous variables

Minkowski family of distances

$$D(i,j) = \sqrt[1/p]{\sum_{k=1}^{N} |x_{ik}-x_{jk}|^p}$$

N features (dimensions)

Distance Metrics

Binary variables

contingency table

	1	0	*sum*
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

Distance Metrics

Binary variables

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

Distance Metrics

Binary variables

contingency table

	1	0	*sum*
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	
1	7	3	
0	106	353	

Distance Metrics

Binary variables

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	sum
1	7	3	10
0	106	353	459
sum	113	356	469

Distance Metrics

Binary variables

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	sum
1	7	3	10
0	106	353	459
sum	113	356	469

IF SYMMETRIC
(same chance to appear)

$$D_{ij} = \frac{b+c}{a+b+c+d} = \frac{109}{469} = 0.23$$

Distance Metrics

Binary variables

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	sum
1	7	3	10
0	106	353	459
sum	113	356	469

IF SYMMETRIC
(same chance to appear)

$$D_{ij} = \frac{M_{i=0j=0} + M_{i=1j=1}}{M_{00}+M_{01}+M_{10}+M_{11}} = \frac{109}{469} = 0.23$$

Distance Metrics   Binary variables

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	sum
1	7	3	10
0	106	353	459
sum	113	356	469

IF ASYMMETRIC
(not same chance)

$$D_{ij} = \frac{b+c}{a+b+c} = \frac{109}{116} = 0.94$$

Distance Metrics **Binary variables**

contingency table

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

e.g.: subway station
w ESCALATOR Y/N
w ELEVATOR Y/N

ELEVATOR

ESCALATOR	1	0	sum
1	7	3	10
0	106	353	459
sum	113	356	469

IF ASYMMETRIC
(not same chance)

Jaccard similarity

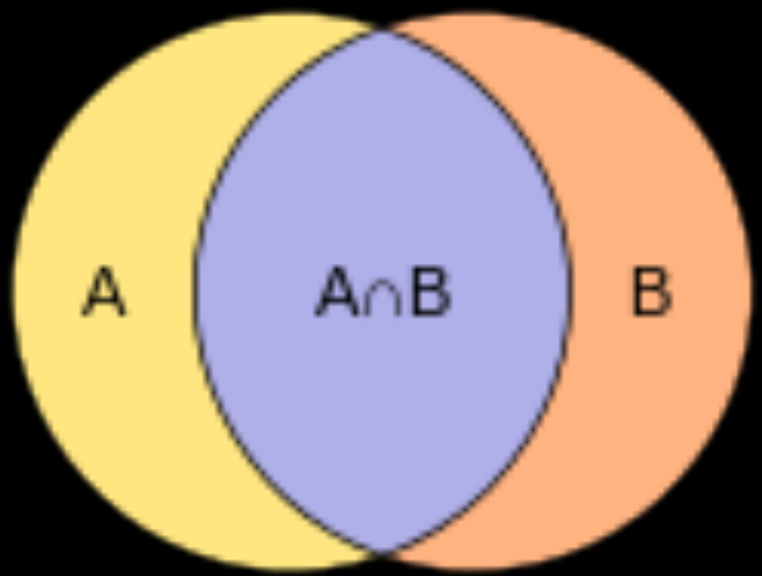
$$J_{ij} = \frac{a}{a+b+c} = \frac{7}{116} = 0.06$$

Distance Metrics   Binary variables

Uses presence/absence data

Jaccard similarity coefficient S_j

$$S_j = \frac{a}{a+b+c}$$



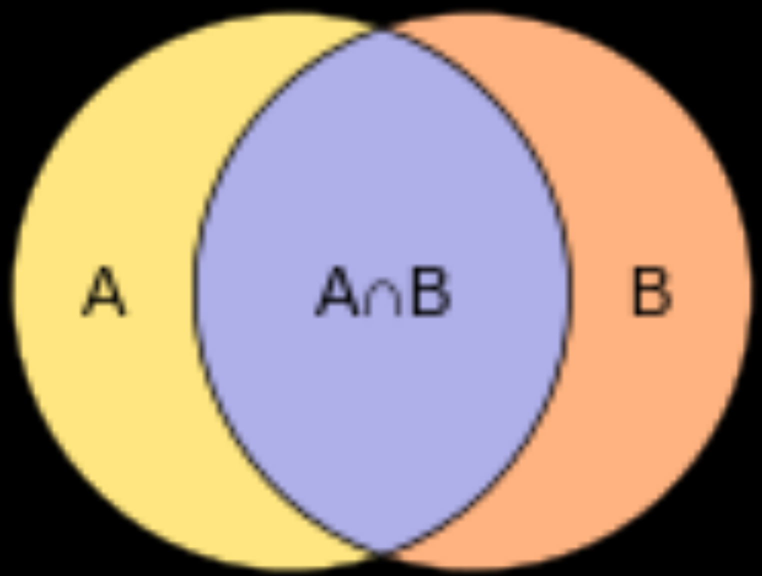
a = number of items in common,
b = number of items unique to the first set
c = number of items unique to the second set

Distance Metrics Binary variables

Uses presence/absence data

Jaccard similarity coefficient S_j

$$S_j = \frac{A \bigcap B}{A \bigcup B}$$



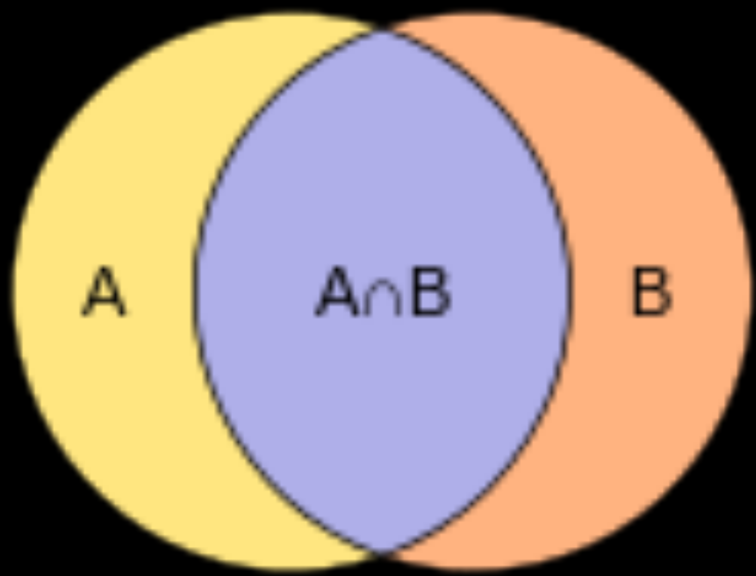
a = number of items in common,
b = number of items unique to the first set
c = number of items unique to the second set

Distance Metrics Binary variables

Uses presence/absence data

Jaccard distance
$D_j = 1 - S_j$

$$S_j = \frac{A \bigcap B}{A \bigcup B}$$



a = number of items in common,
b = number of items unique to the first set
c = number of items unique to the second set

Distance Metrics Categorical Variables

Uses presence/absence data in two samples (non exclusive)

Simple similarity coefficient

Simple Matching Method

SMC

$$S_{ij} = \frac{p-m}{p}$$

p: number of variables

m: number of matches

jupyter

https://github.com/fedhere/UInotebooks/blob/master/cluster/categorical_clustering.ipynb

CUSP

Distance Metrics Ordinal variables

Uses ranks

map range 0-1

$r_{ij} = \{1\ldots R_N\}$ -> $\mathbf{z_{ij} = \frac{r_{ij} - 1}{R_N - 1}}$

Distance Metrics vector Variables

Uses correlation coefficient! or

Pearson's correlation

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Cosine similarity

$$\cos(\theta) = \frac{A \cdot B}{||A|| \, ||B||}$$

Distance Metrics Can we think about other data??

Distance Metrics — Can we think about other data??

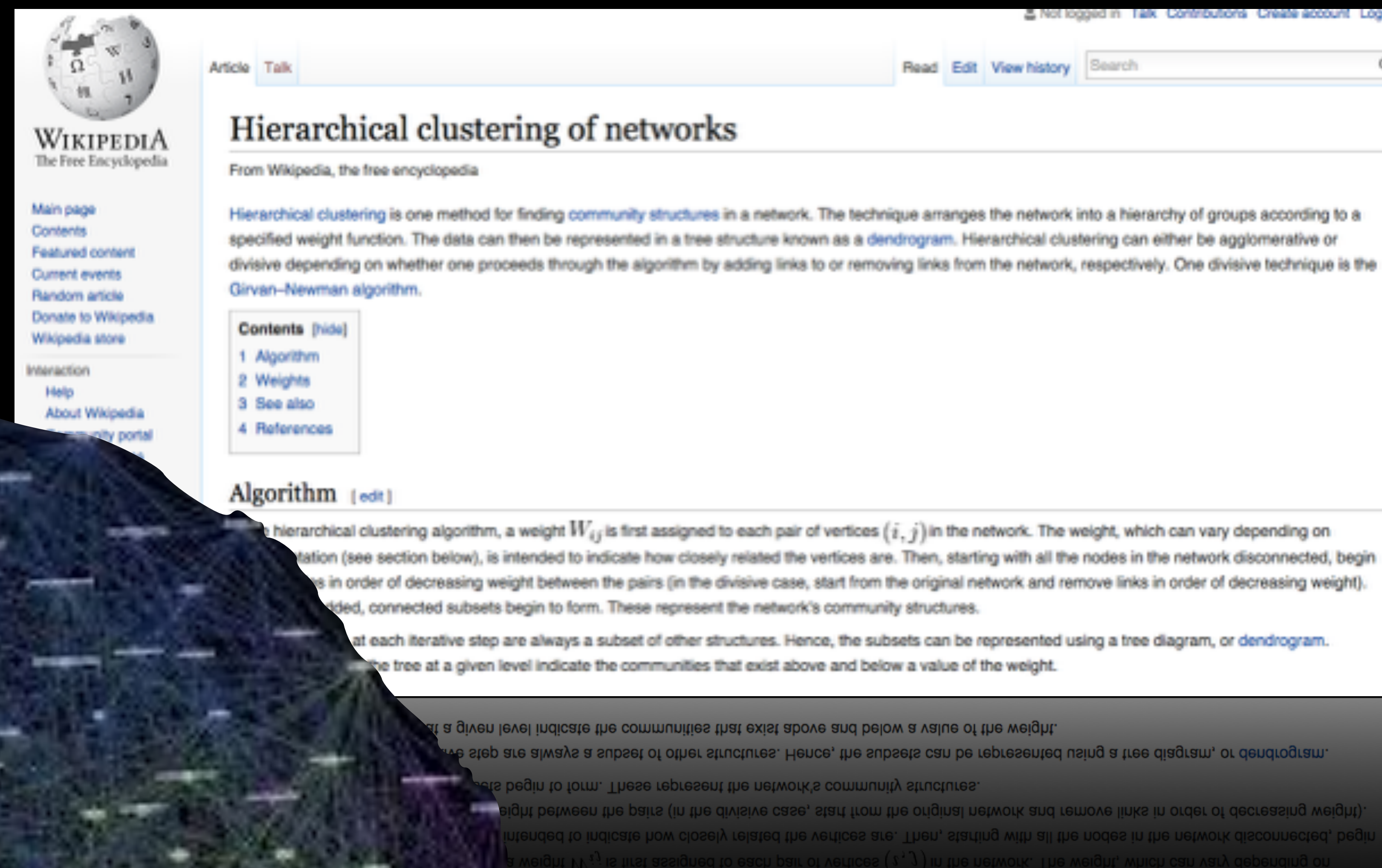
Multimedia

Distance Metrics

Can we think about other data??

Multimedia

Network



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia

Hierarchical clustering of networks

From Wikipedia, the free encyclopedia

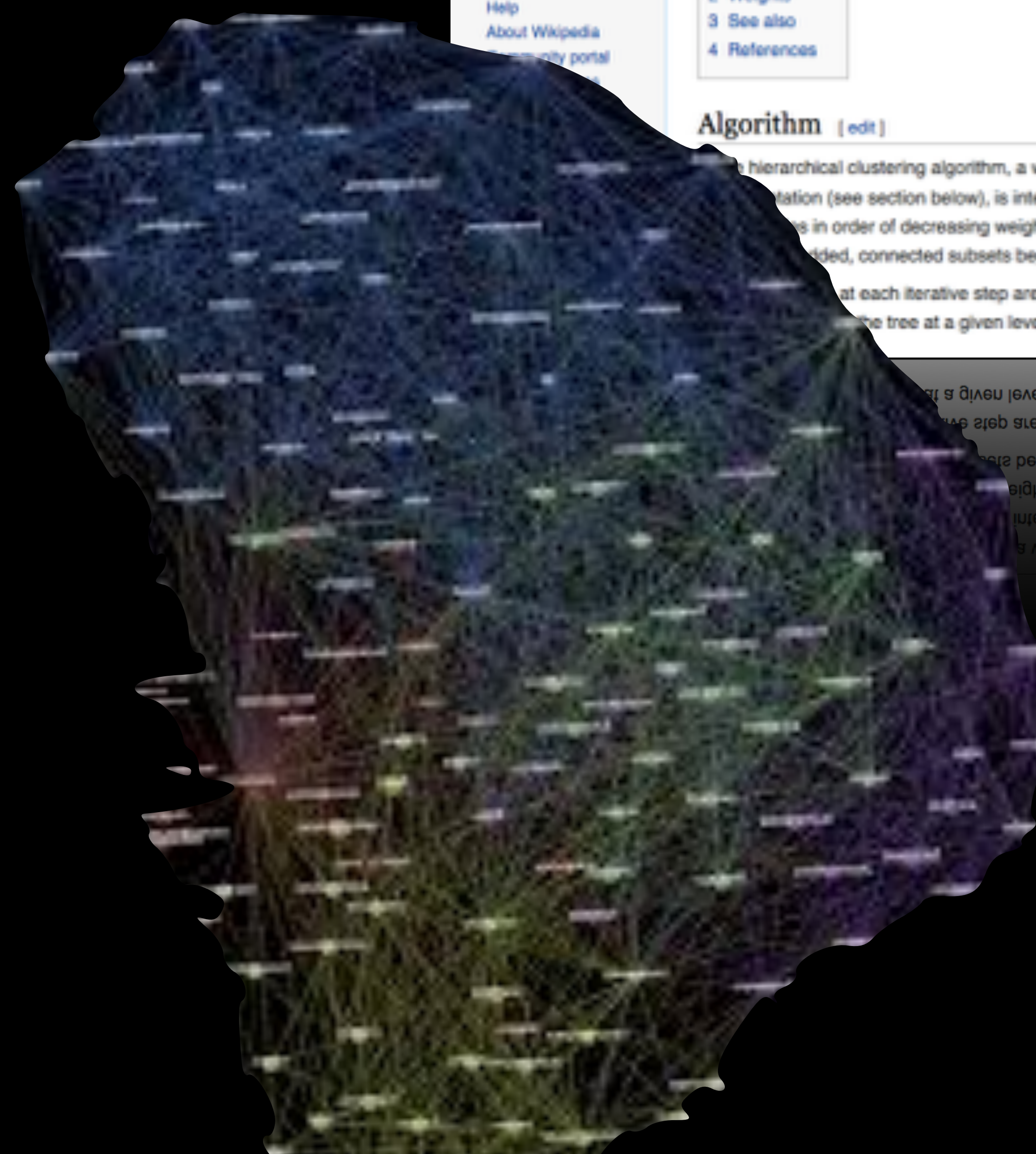
Hierarchical clustering is one method for finding community structures in a network. The technique arranges the network into a hierarchy of groups according to a specified weight function. The data can then be represented in a tree structure known as a dendrogram. Hierarchical clustering can either be agglomerative or divisive depending on whether one proceeds through the algorithm by adding links to or removing links from the network, respectively. One divisive technique is the Girvan–Newman algorithm.

Contents [hide]
1 Algorithm
2 Weights
3 See also
4 References

Algorithm [edit]

hierarchical clustering algorithm, a weight W_{ij} is first assigned to each pair of vertices (i, j) in the network. The weight, which can vary depending on ...tation (see section below), is intended to indicate how closely related the vertices are. Then, starting with all the nodes in the network disconnected, begin ...s in order of decreasing weight between the pairs (in the divisive case, start from the original network and remove links in order of decreasing weight). ...ded, connected subsets begin to form. These represent the network's community structures.

...at each iterative step are always a subset of other structures. Hence, the subsets can be represented using a tree diagram, or dendrogram. ...e tree at a given level indicate the communities that exist above and below a value of the weight.



Distance Metrics

Can we think about other data??

Multimedia

Network

Sequence

Distance Metrics MIXED variables

Hybrid dataset containing continuous, ordinal, categorical

weighted distance

$$D_w = \frac{\sum_{p=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{p=1}^{p} w_{ij}^{(f)}}$$

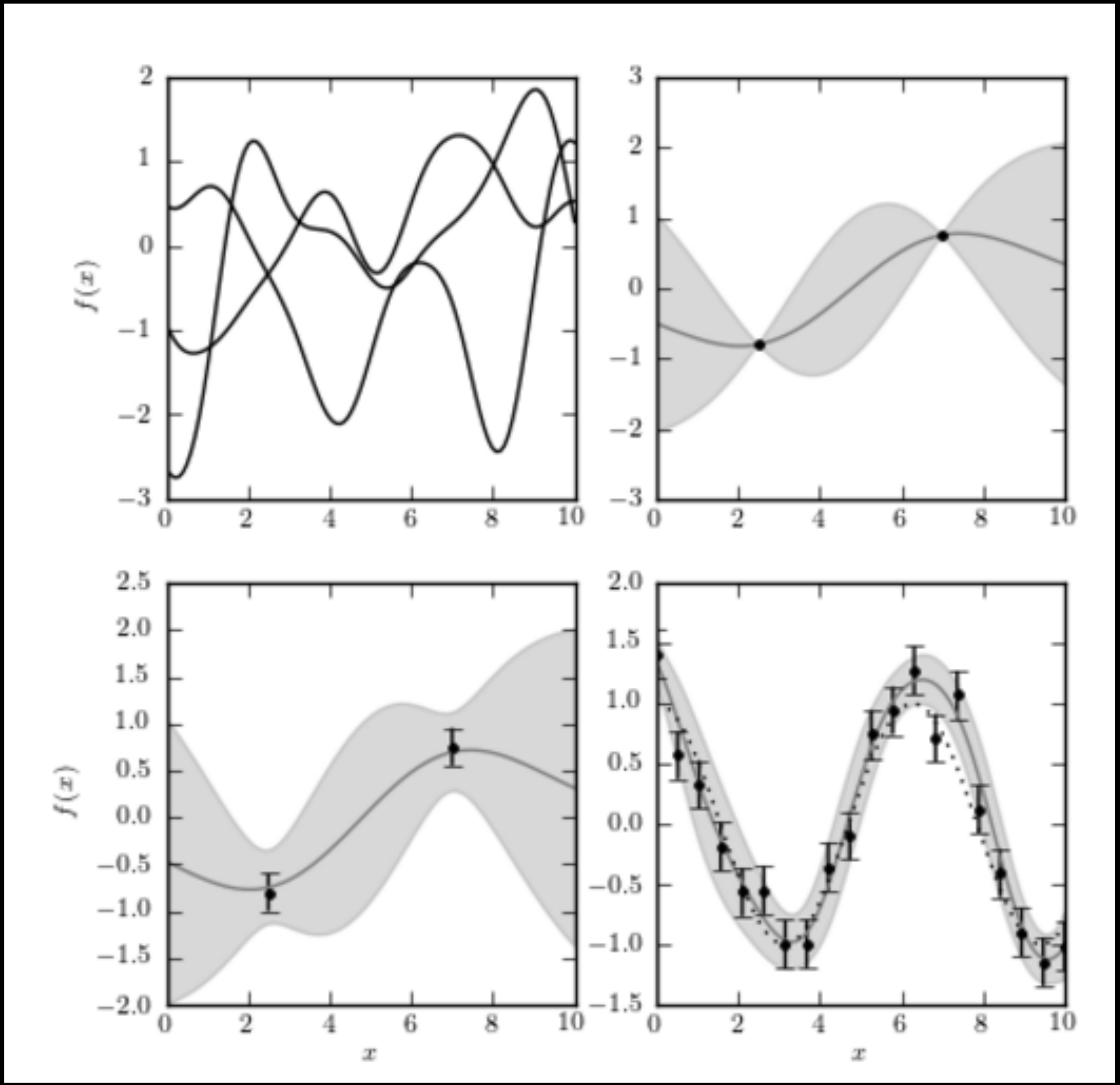
Kriging

Kriging

(1951, Danie Krieg -
geospatial statistics: evaluation of mineral sources)

Kriging
Gaussian processes
(in time domain)

every point in the support is associated with a normally distributed random variable



http://www.astroml.org/book_figures/chapter8/fig_gp_example.html

every point in the support is associated with a normally distributed random variable

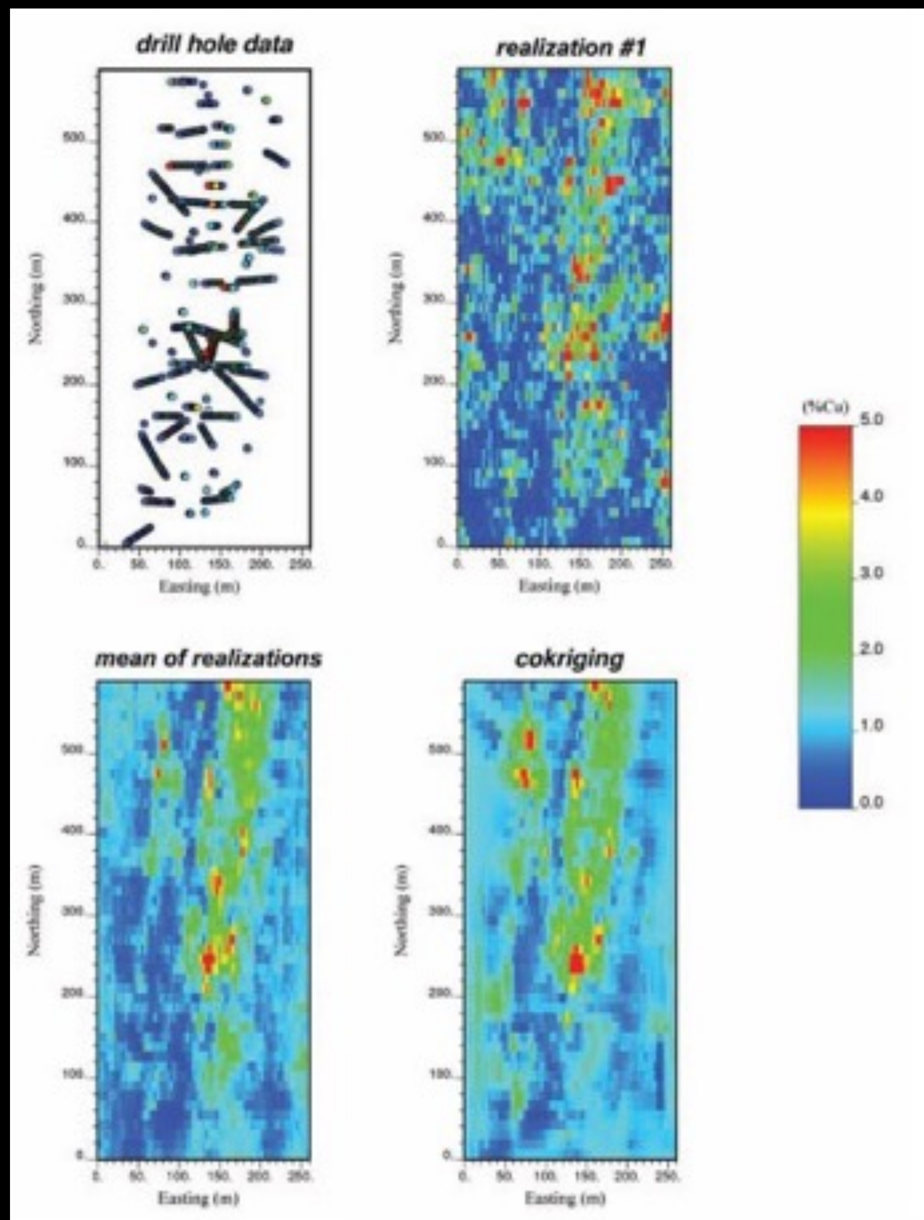
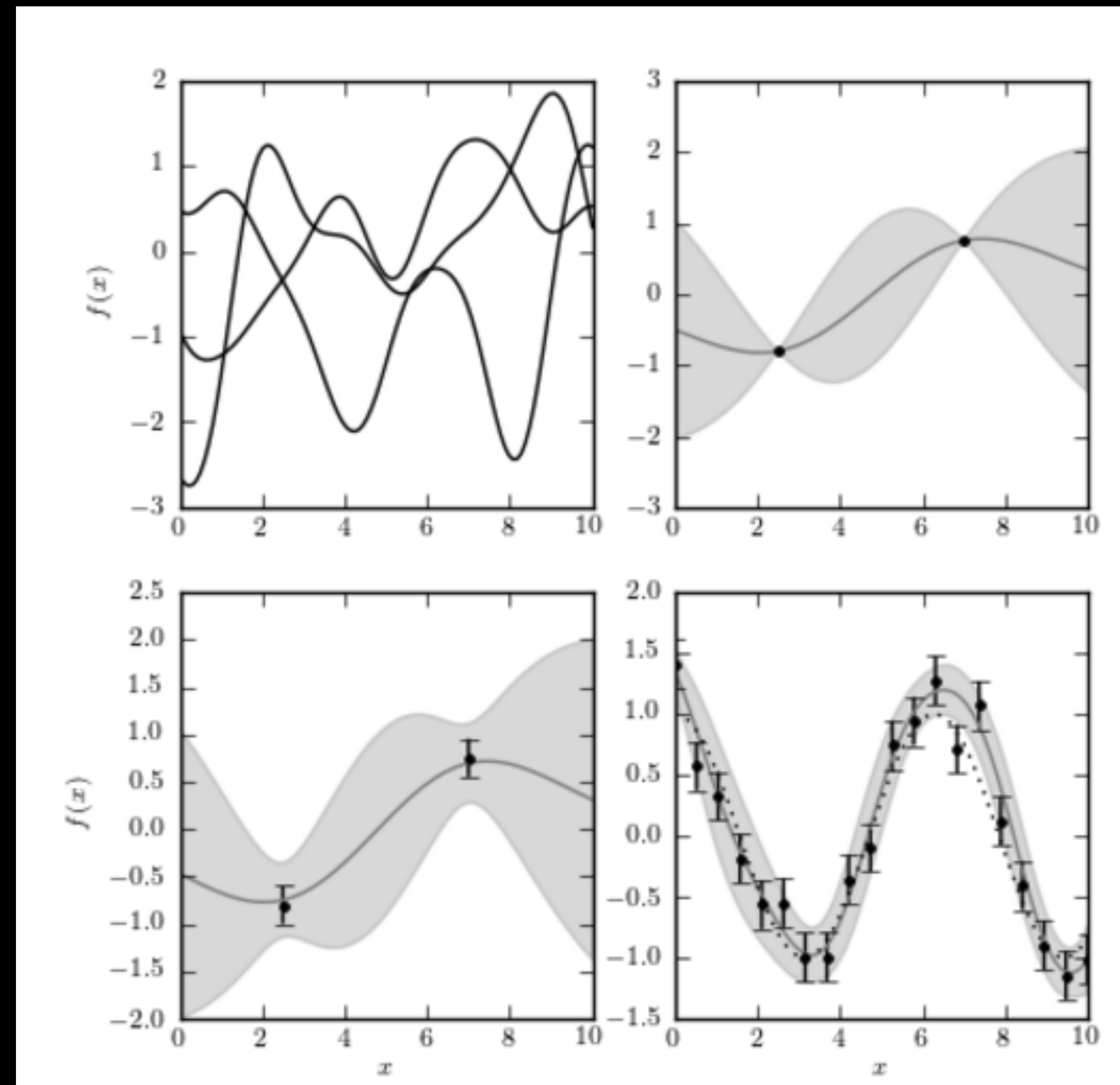


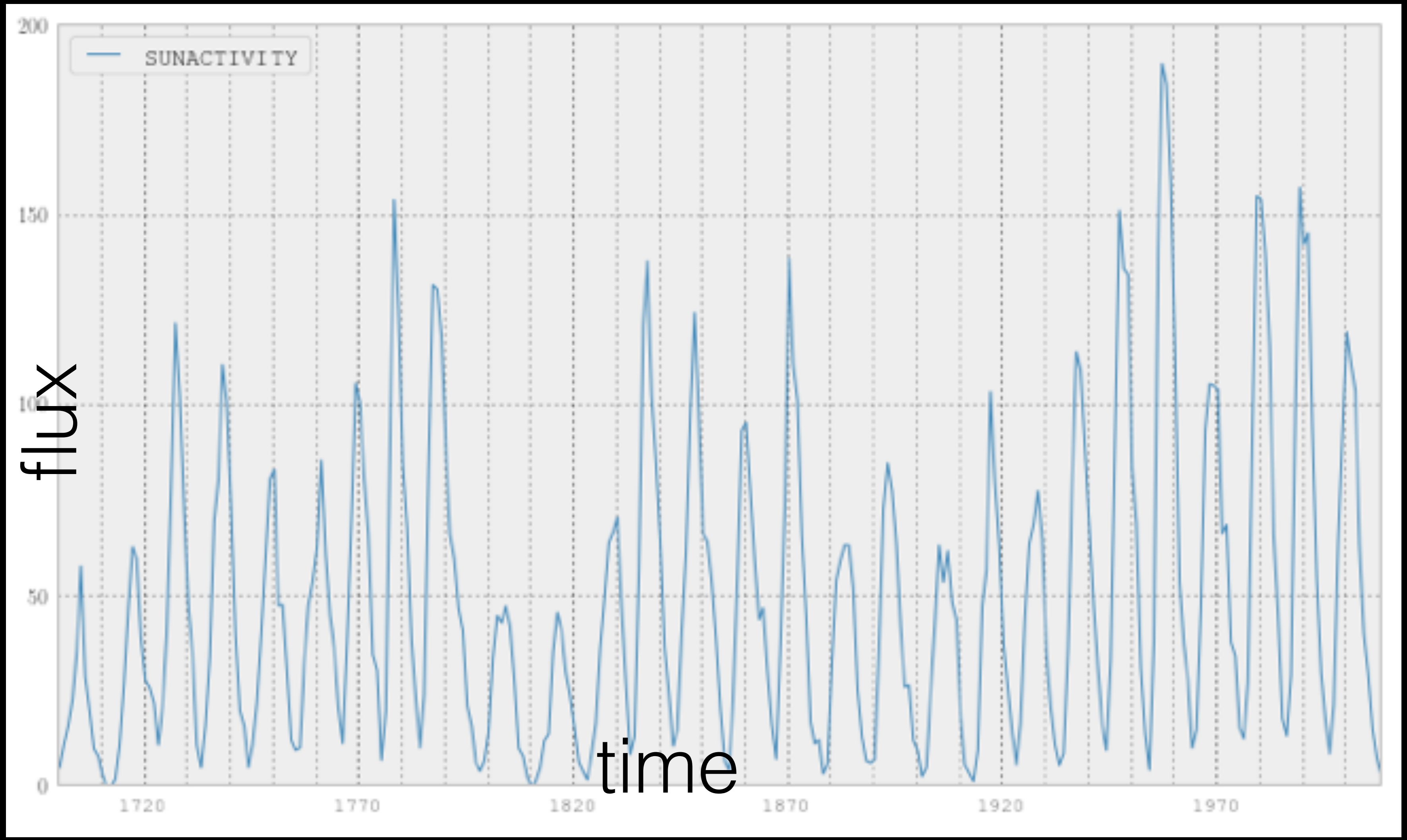
Figure 3—From top to bottom and left to right: copper grade in drill hole samples, one realization at block support, average of forty realizations, and cokriging estimate (elevation 2200 m)



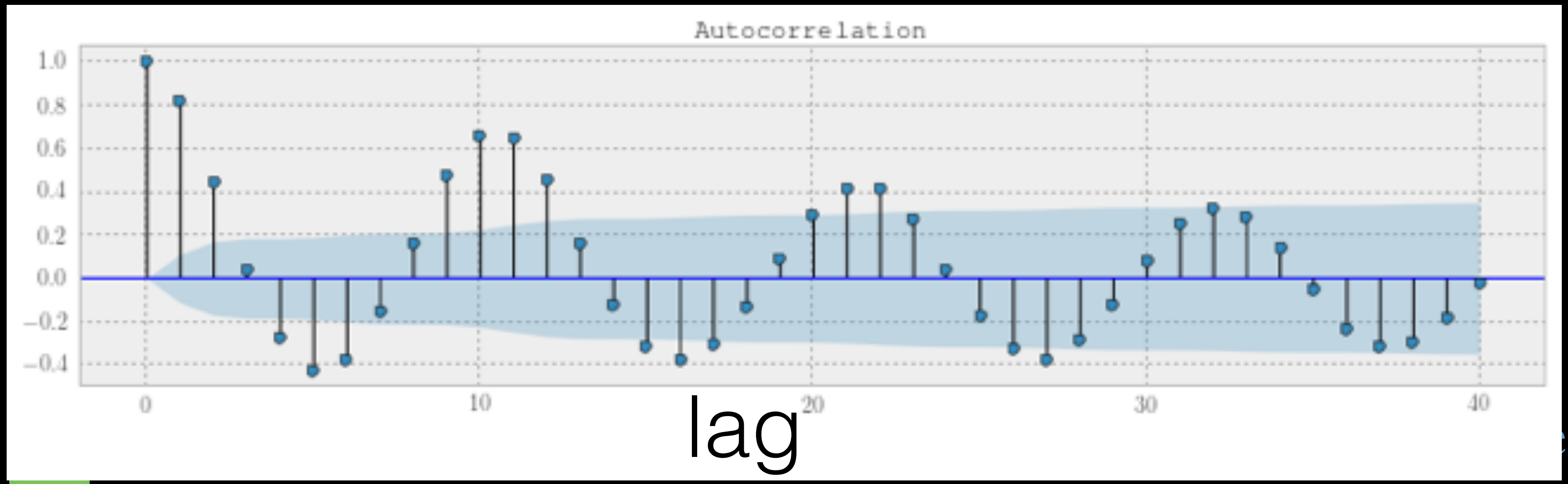
http://www.astroml.org/book_figures/chapter8/fig_gp_example.html

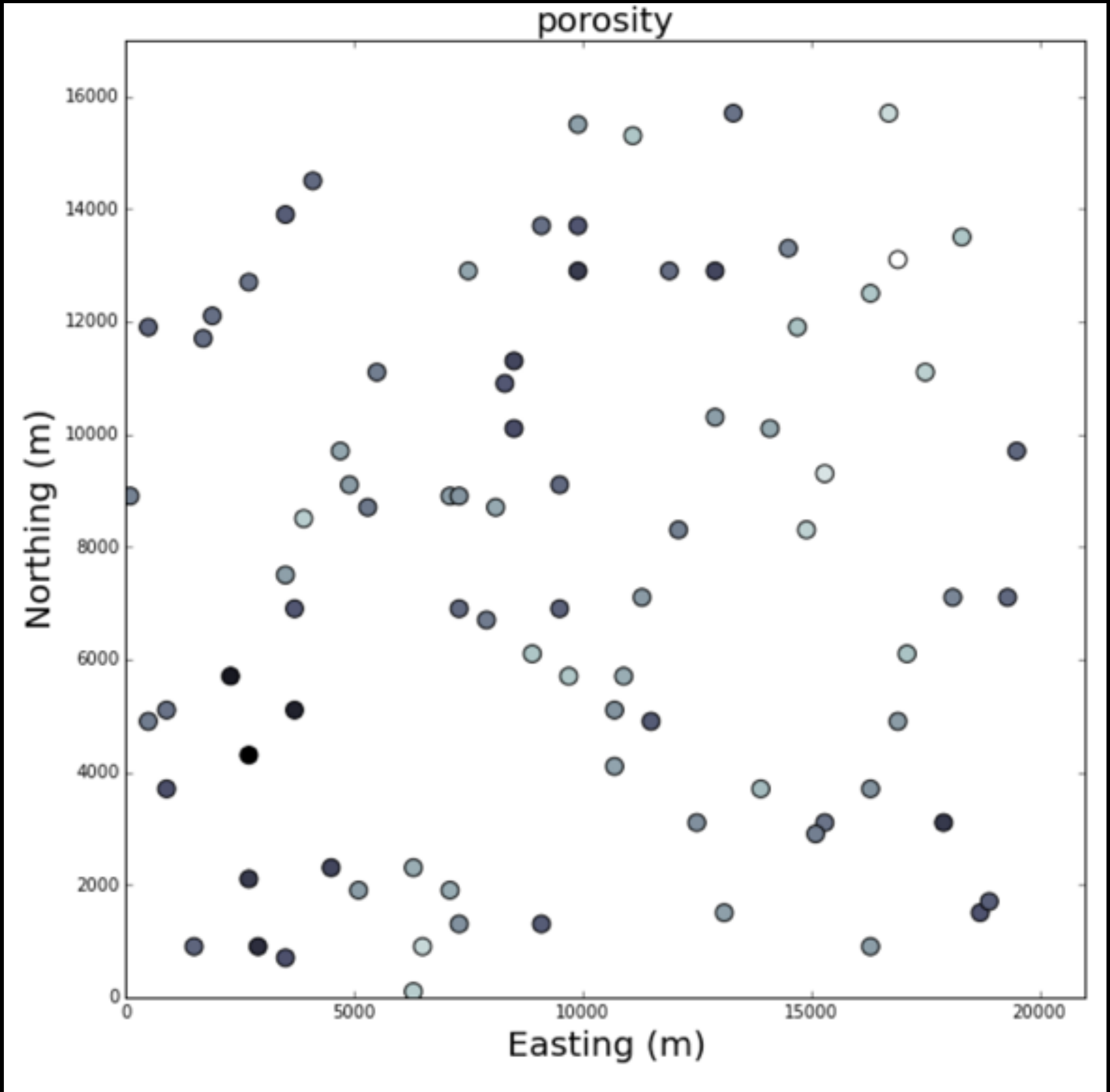
Correlation : a measure of spatial (temporal, hyperspatial) continuity

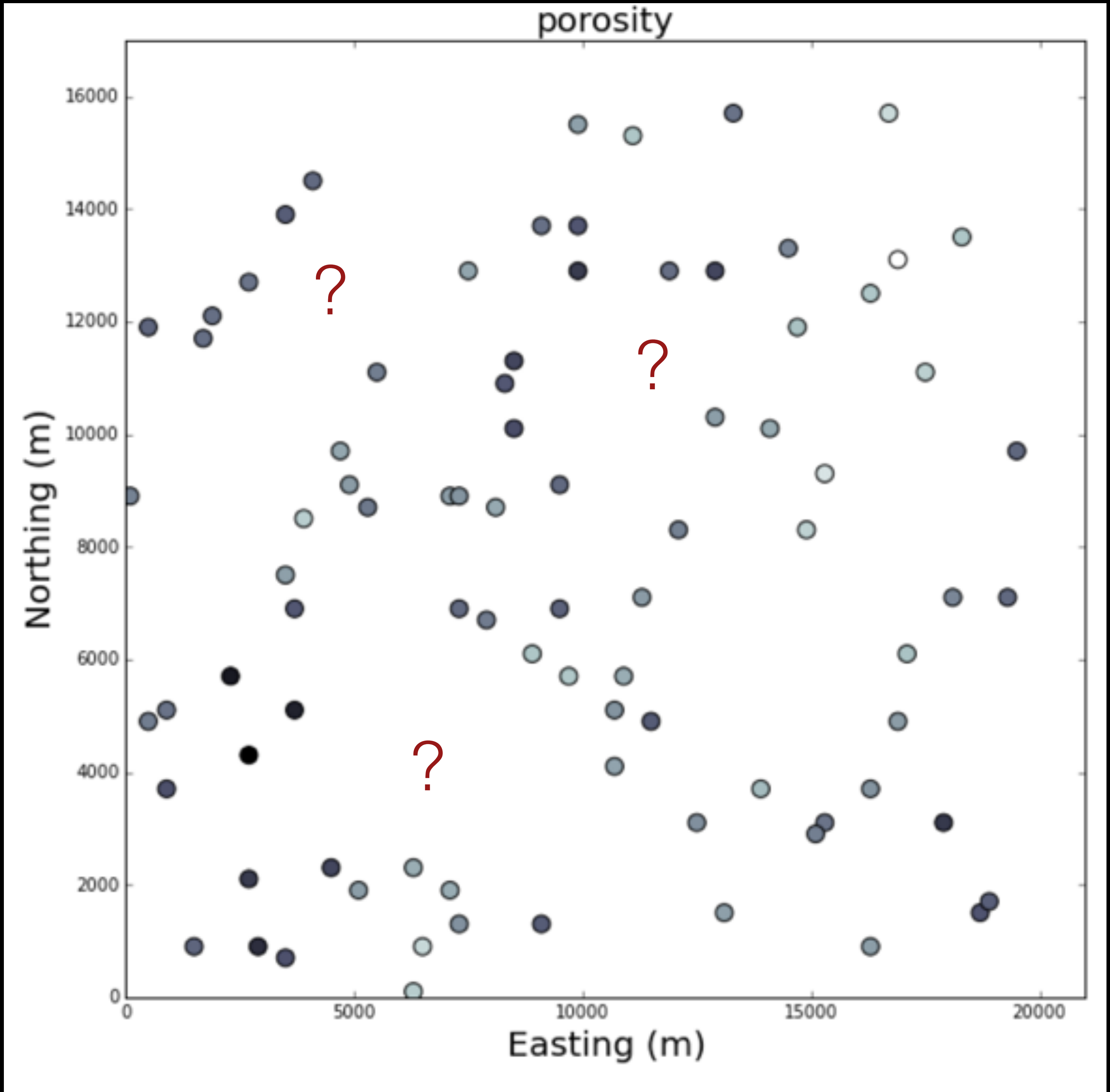
http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/tsa_arma_0.html

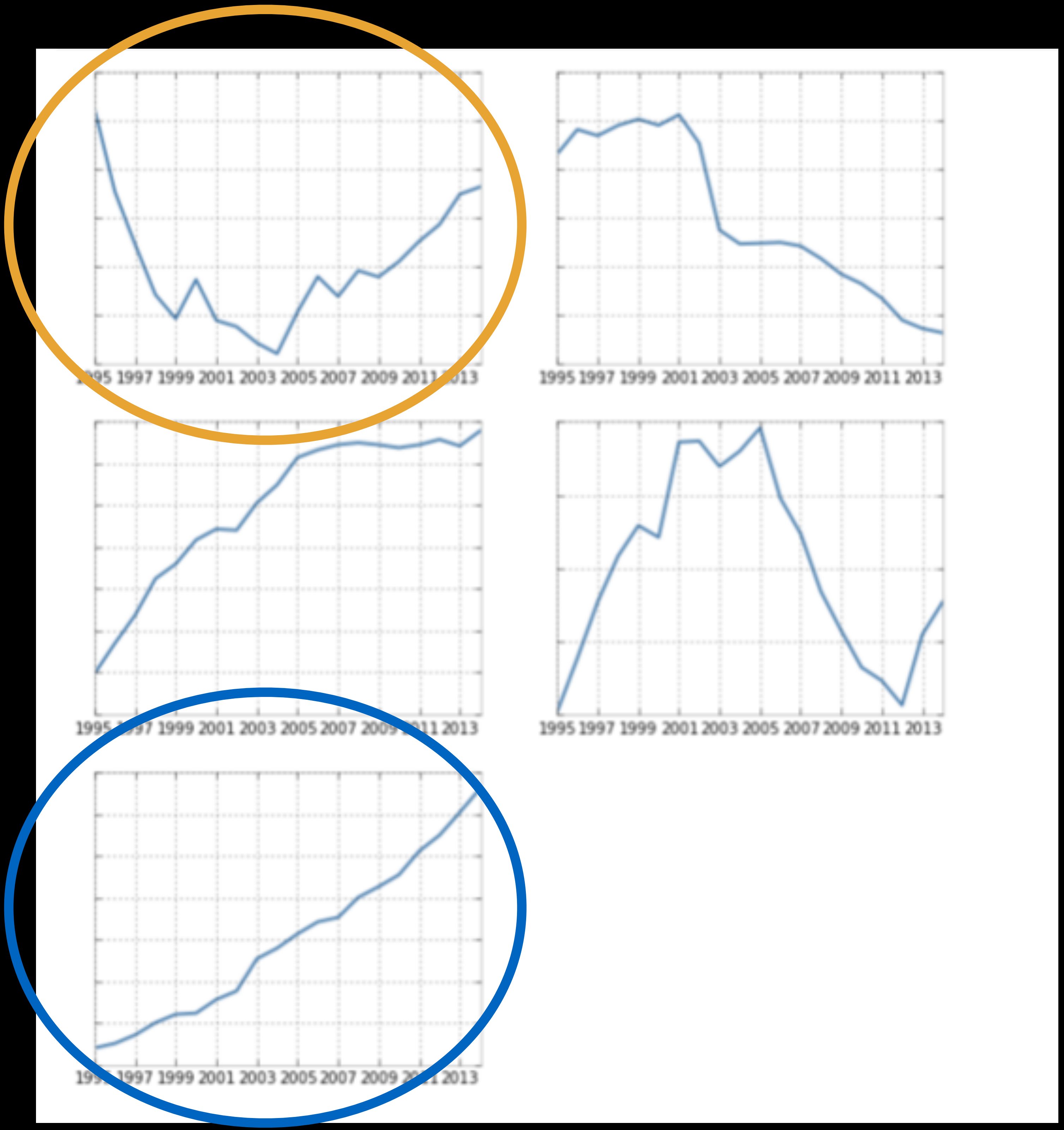


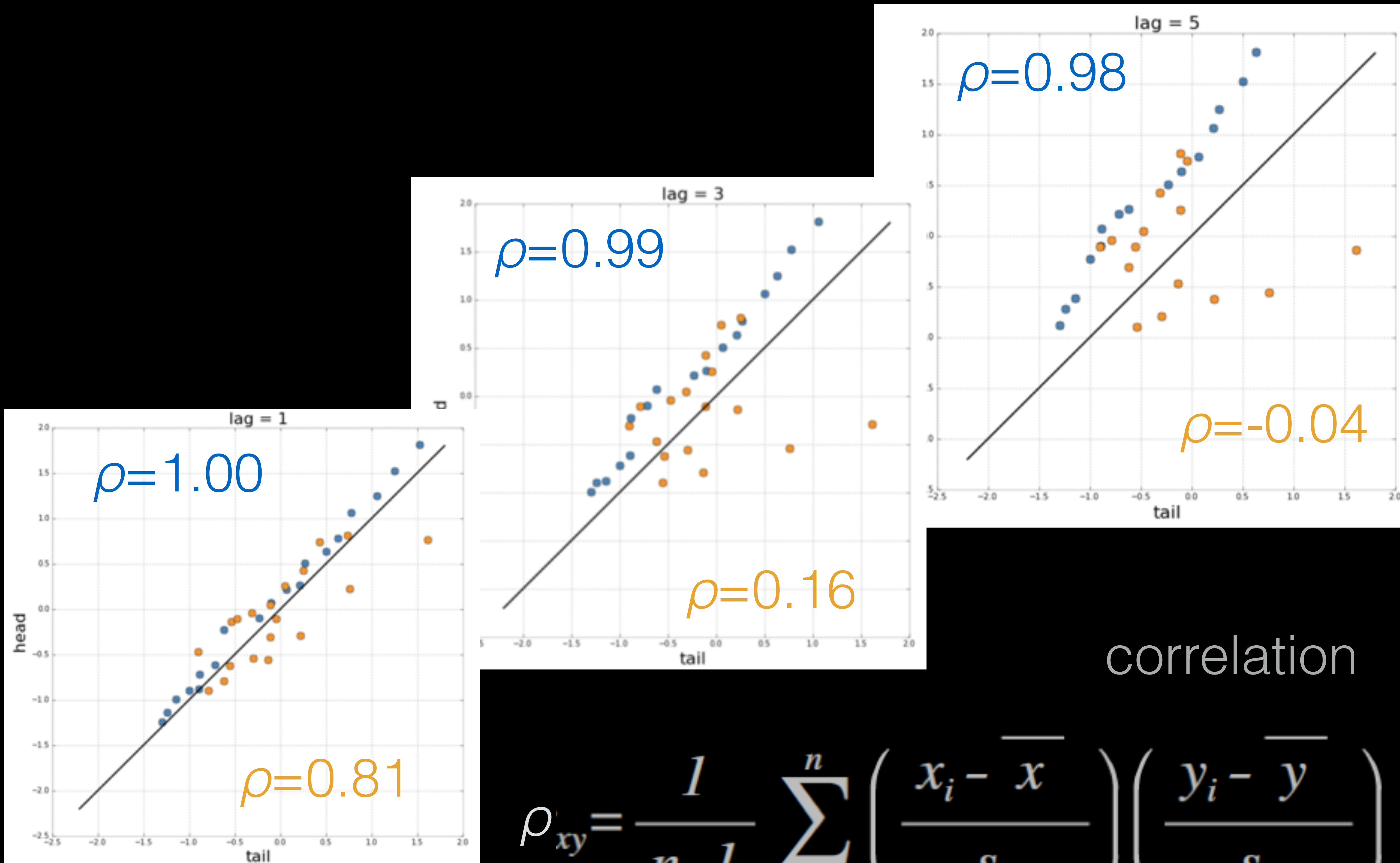
autocorrelation




porosity
16000
14000
12000
10000
8000
6000
4000
2000
0
Northing (m)
0
5000
10000
15000
20000
Easting (m)


porosity
16000
14000
12000
10000
8000
6000
4000
2000
0
Northing (m)
0
5000
10000
15000
20000
Easting (m)
?
?
?


1995 1997 1999 2001 2003 2005 2007 2009 2011 2013



correlation

$$\rho_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

correlation

$$\rho_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

covariance

$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

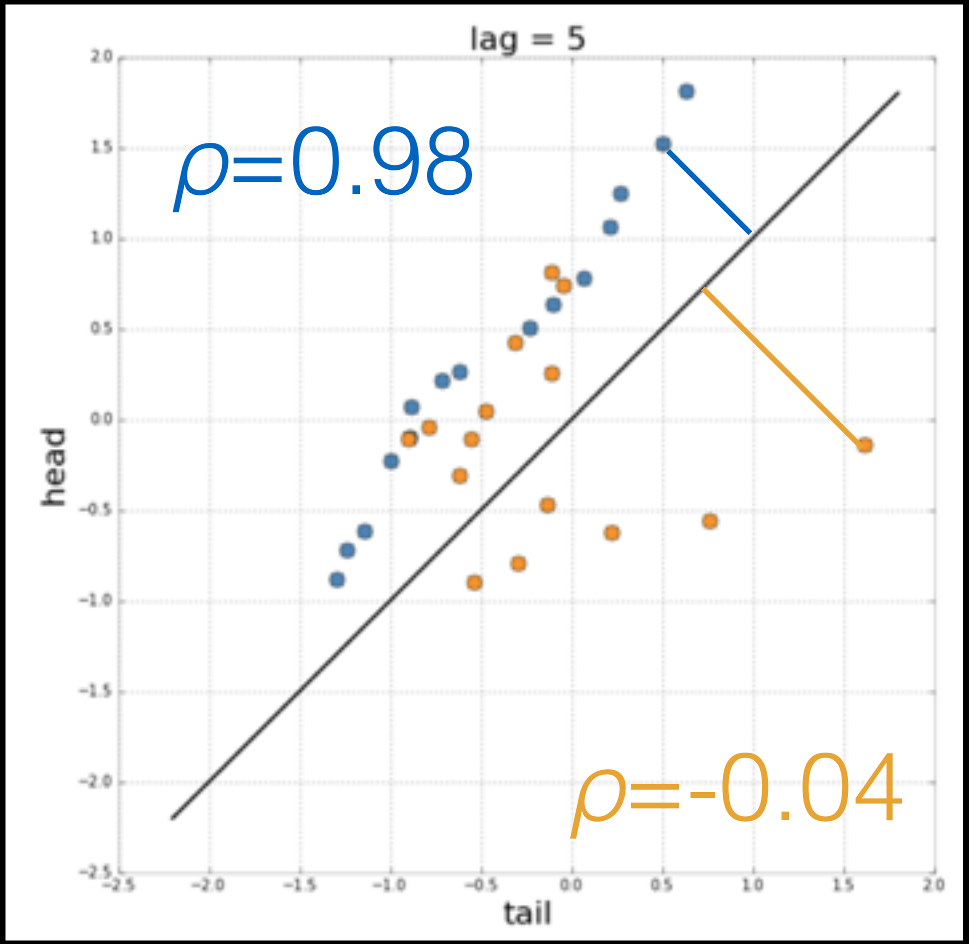
let Σ be the covariant matrix

$$\Sigma(AX) = A\,\Sigma(X)\,A^T$$

this is why we add errors in
is diagonal (independent variables)

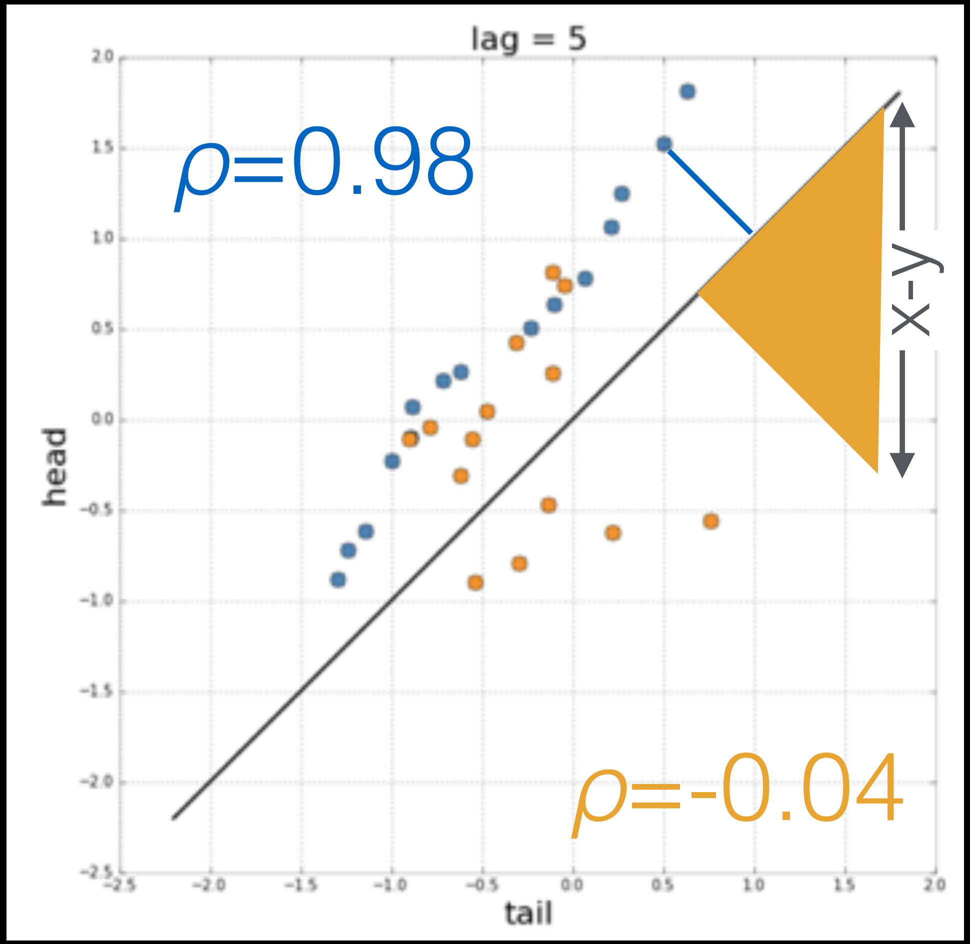
correlation $\rho_{X,Y} = \dfrac{cov(X,Y)}{\sigma_X \sigma_Y}$

covariance $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$



correlation $\rho_{X,Y} = \dfrac{cov(X,Y)}{\sigma_X \sigma_Y}$

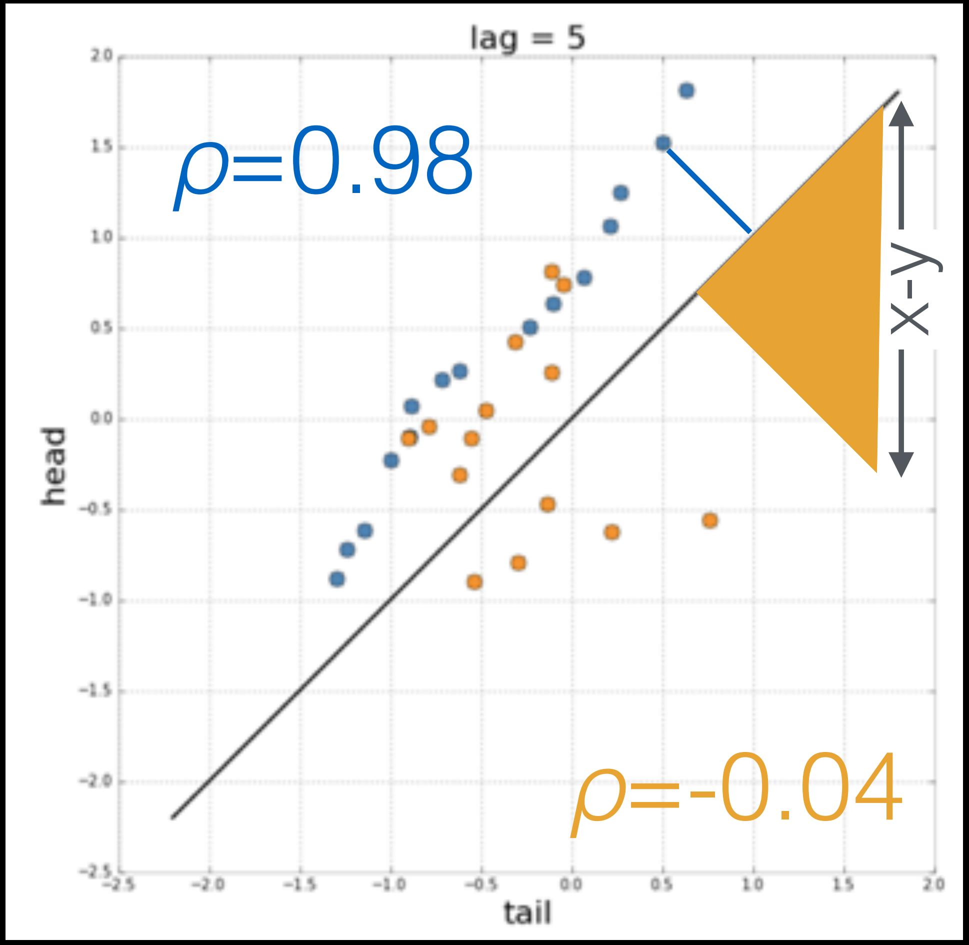
covariance $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$



correlation $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

covariance $cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$$\overline{D}^2(h) = \frac{1}{N}\sum_i^N \left(\frac{1}{\sqrt{2}}(z_x - z_{x+h})\right)^2$$



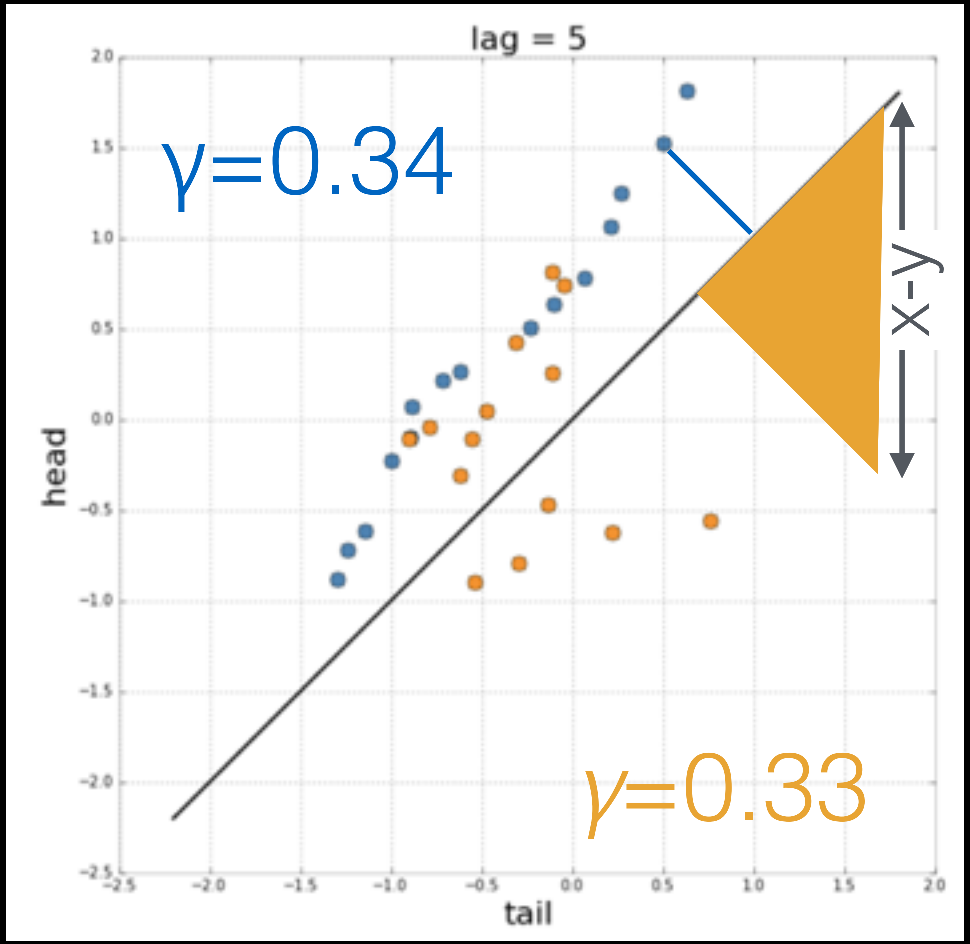
correlation $$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

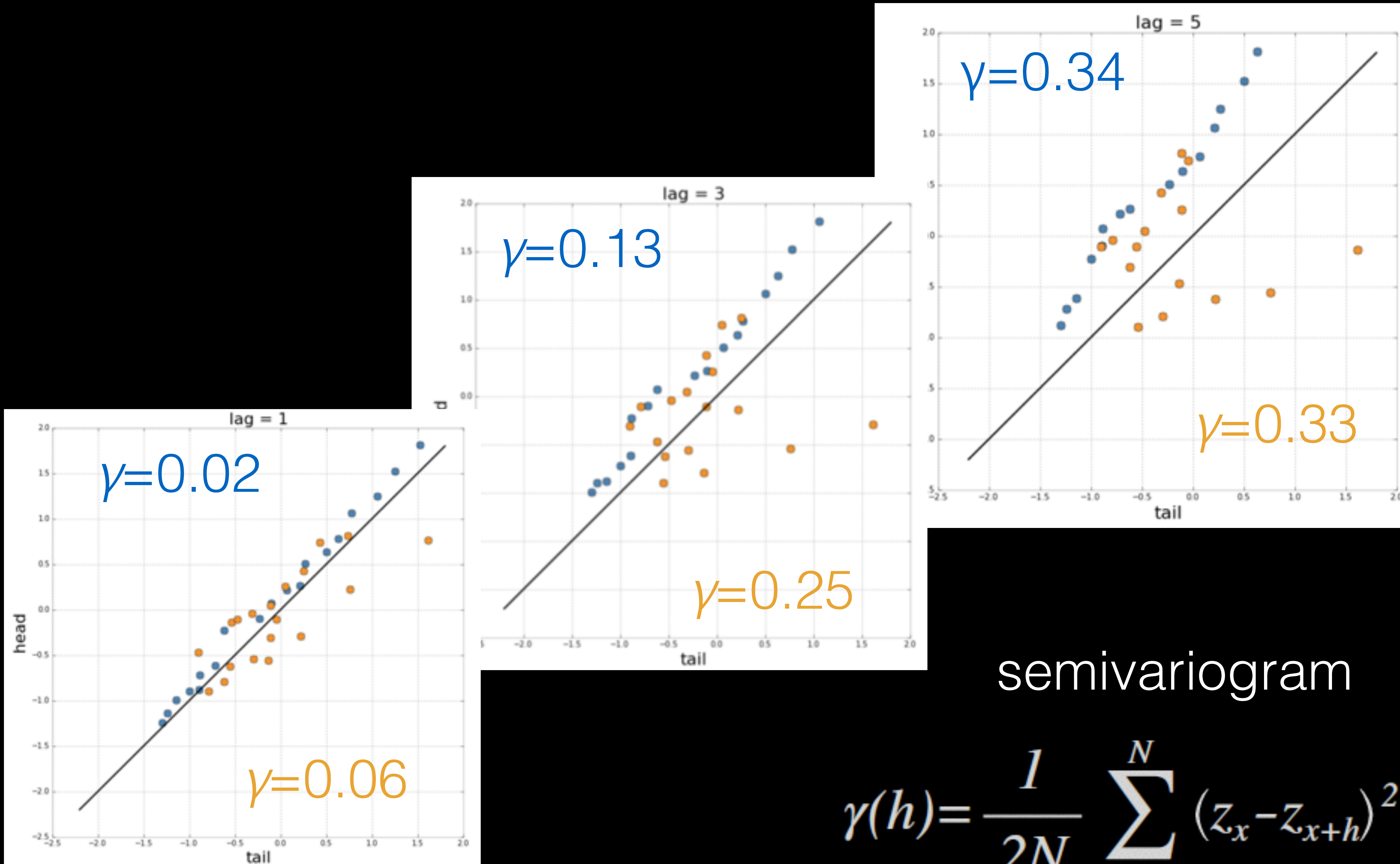
covariance $$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$$\overline{D}^2(h) = \frac{1}{N} \sum_i^N \left(\frac{1}{\sqrt{2}} (z_x - z_{x+h}) \right)^2$$

semi-variogram

$$\gamma(h) = \frac{1}{2N} \sum_i^N (z_x - z_{x+h})^2$$





semivariogram

$$\gamma(h) = \frac{1}{2N} \sum_{i}^{N} (z_x - z_{x+h})^2$$

Kriging math:

minimizes $\sigma^2{}_E\,(Z(u)\text{-}\mu(u))$ with $E[Z^E(u)\text{-}\mu(u)] = 0$

$$Z^E(u)-\mu(u) = \sum_{k=1}^{N(u)} \lambda_k(Z(u_k)-\mu(u_k))$$

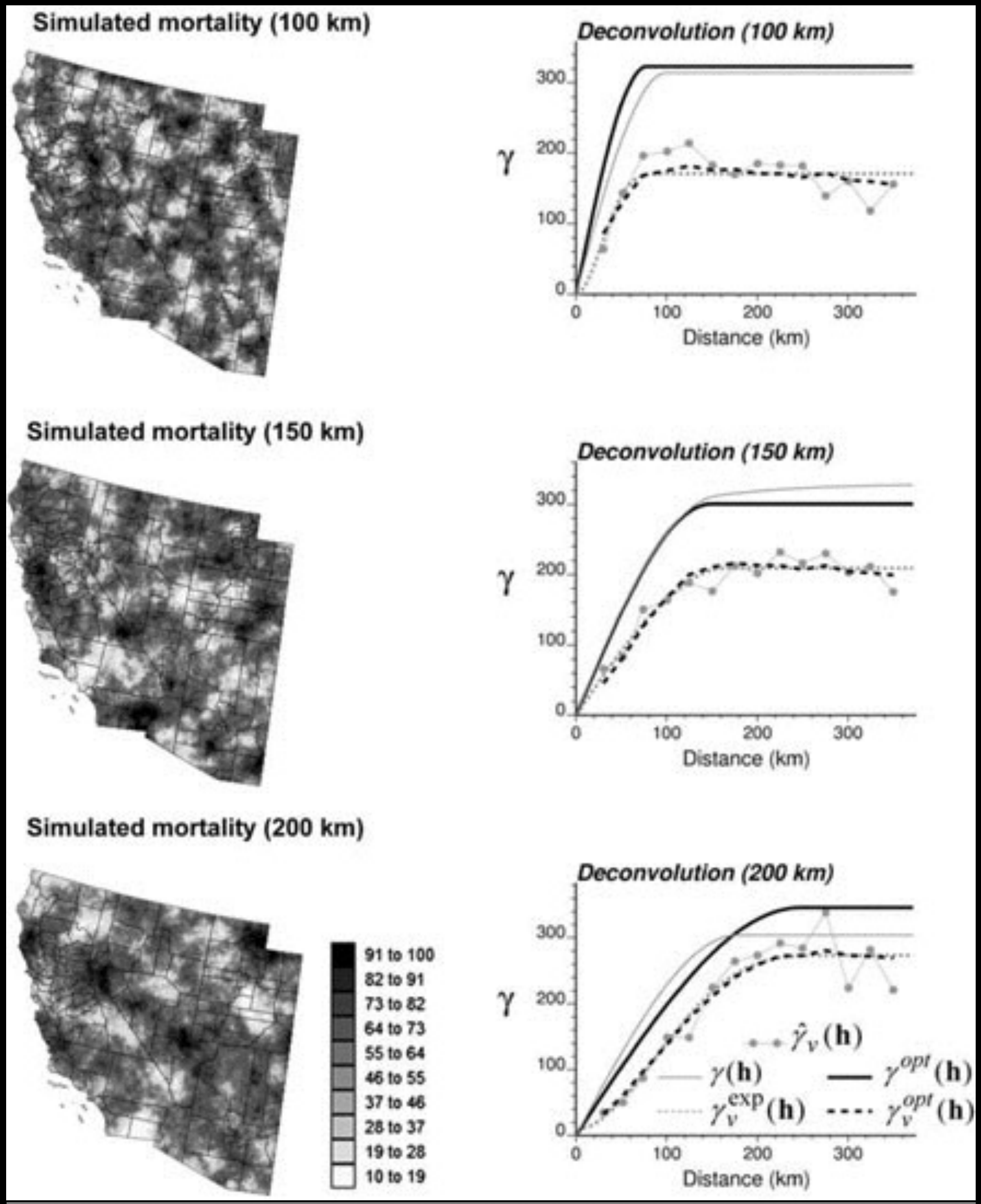
$$R(u) = Z^E(u)-\mu(u)$$

$$Cov(R(u),\ R(u+h) = E[R(u)\cdot R(u+h)]$$

$$Cov(R(u),\ R(u+h) = Cov(0,\gamma(h))$$

jupyter

XI: Categorical Clustering
Kriging



Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1697809/#B10

Kriging:

it is a form of regression (probabilistic linear regression)

it generates a family of random functions from a distribution driven by the data values, the data uncertainties, and the correlation (temporal, spatial, hyperspatial) between the data

it allows a (robust) estimate of the uncertainty in the regression

is your code optimized:

check CPU AND MEMORY usage

vectorize (slice and avoid for loops)

avoid storing information you do not need in memory

use local variables

remove all redundant calculations from inside loops

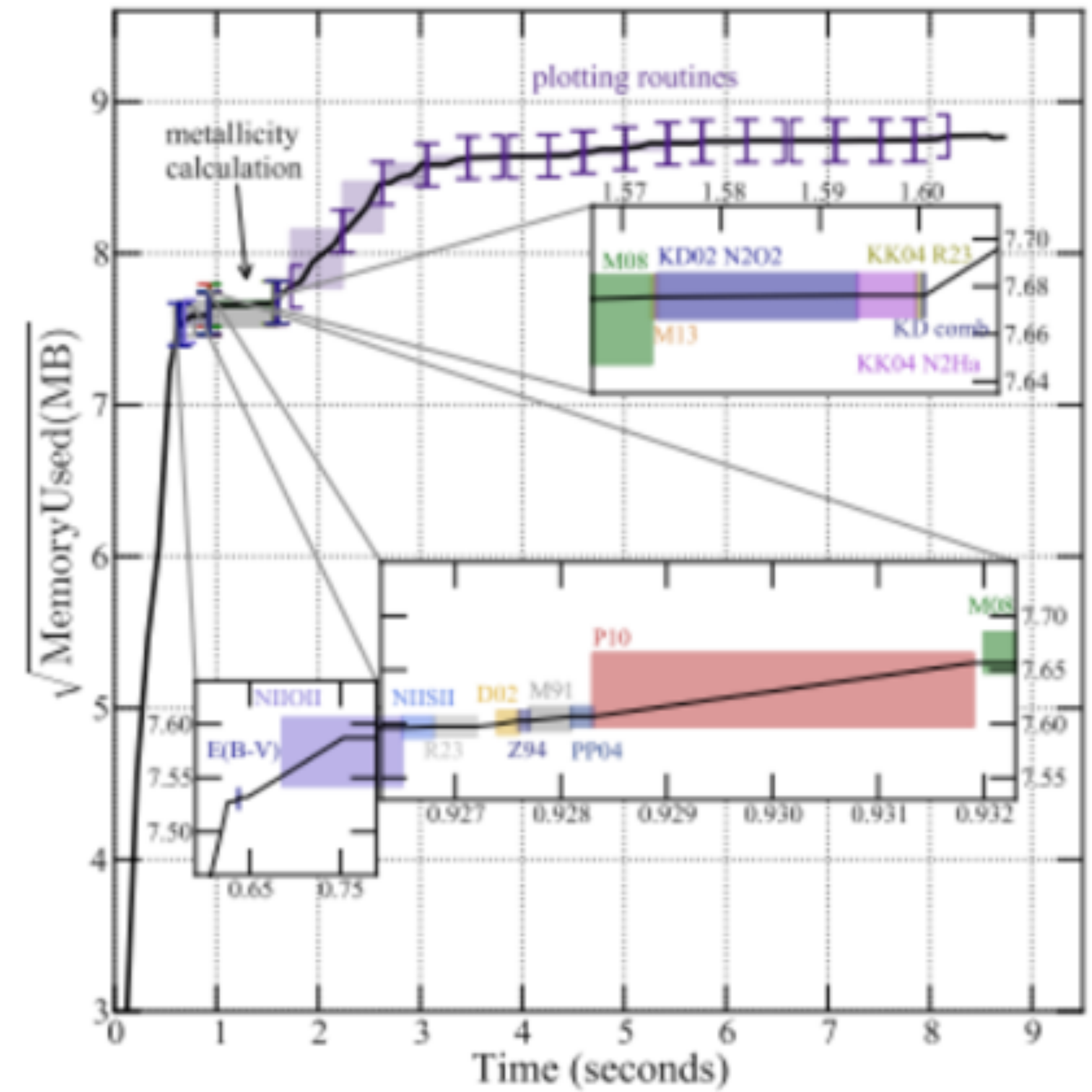


FIG. 6.— Memory usage: we plot the square root of the memory usage in Megabytes as a function of time for running our code (using N=2,000 and all default metallicity scales except the D13 *pyqz* ones) on a single set of measured emission lines (Table 2, host galaxy of SN 2008D). The square root is plotted, instead of the natural value, to enhance visibility. Three inserts show the regions where most of the metallicity scales are calculated, zoomed in, since the run time of the code is dominated by plotting routines, including the calculation of the bin size with Knuth's rule. Each function call is represented by an opening and closing bracket in the main plot, and by a shaded rectangle in the zoomed-in insets. The calculation of $N2O2$, which requires 0.25 seconds, is split be-

Kriging

Reading:

An excellent use of viz for data exploration and transition to inferential analysis

https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a#.iz1r655xo

Lee Shangqian, Daniel Sim & Clarence Ng

Homework:

- download asma discharge count by facility with SQL query
- clone and install https://github.com/bsmurphy/PyKrige locally on compute
- create a high resolution interpolated map of asthma incidence in NYC
- (or explain why you cannot…)

Distance measures for clustering:

http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/mvahtmlnode79.html

Kriging:

http://people.ku.edu/~gbohling/cpe940/Kriging.pdf

http://connor-johnson.com/2014/03/20/simple-kriging-in-python/

http://www.pykriging.com/

http://www.gaussianprocess.org/gpml/

Goovaerts P. Kriging and semivariogram deconvolution in presence of irregular geographical units. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2518693/

Efficient python coding:

https://wiki.python.org/moin/PythonSpeed/PerformanceTips