Dimension reduction of gene expression profiles

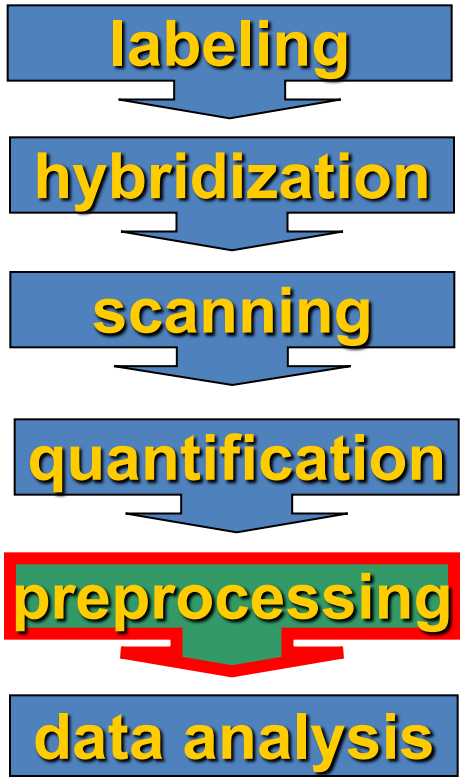
Vincent Detours

vdetours@ulb.ac.be

- IRIBHM, UNIVERSITE LIBRE DE BRUXELLES

Hierarchical clustering

The expression matrix



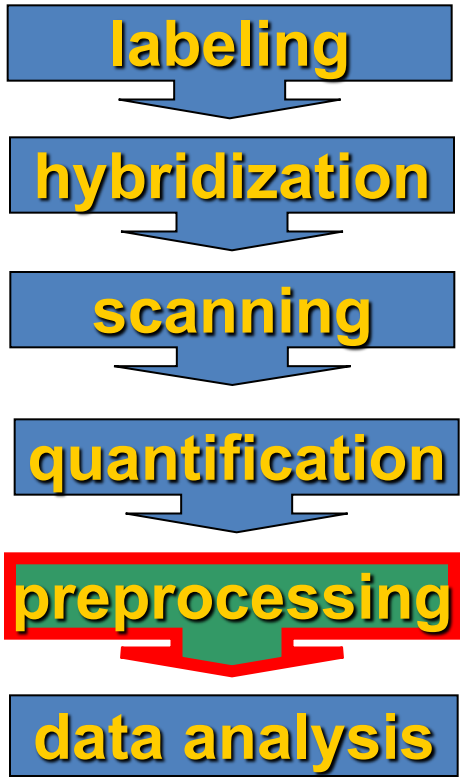
The result of all this is a matrix of expression ratios (tumor/healthy) for M genes x N tumors (e.g. 25000 x 100)

N tumors

M genes

	T1	T2	…	TN
G1	1.5	-1.0	…	0.3
G2	-0.5	3.0	…	0.2
G3	0.0	0.0	…	-0.1
…	…	…	…	…
GM	2.6	-0.2	…	1.1

The expression matrix



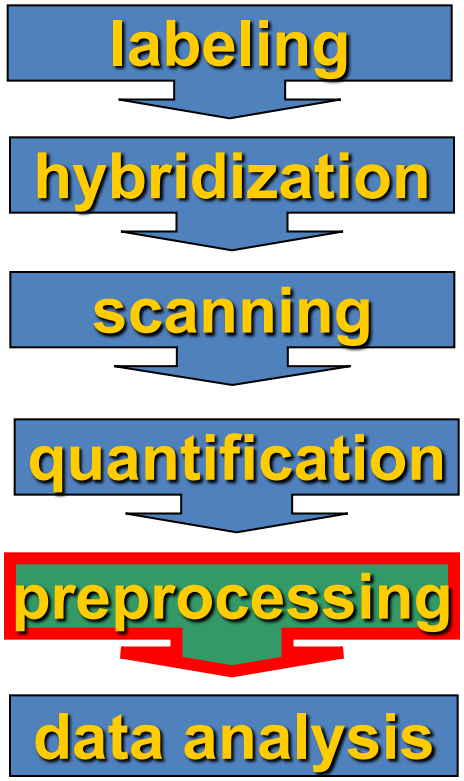
Look at the N tumors as points in the M-dimensional ‘gene space’

N tumors

M genes

	T1	T2	…	TN
G1	1.5	-1.0	…	0.3
G2	-0.5	3.0	…	0.2
G3	0.0	0.0	…	-0.1
…	…	…	…	…
GM	2.6	-0.2	…	1.1

The expression matrix



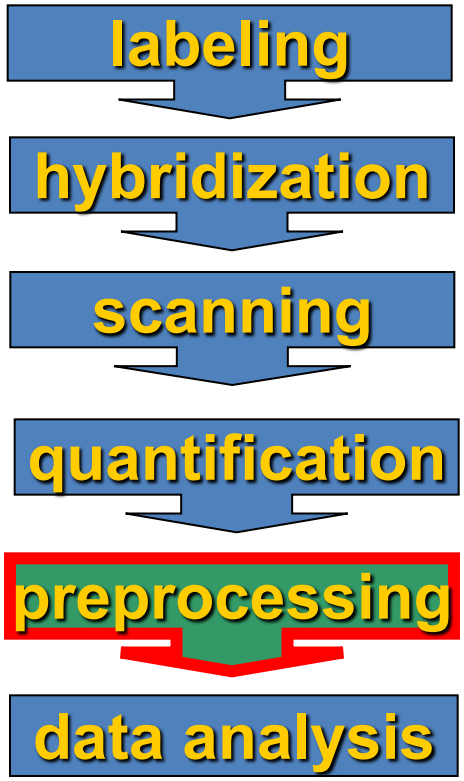
Look at the M genes as points in the N-dimensional ‘phenotype space’

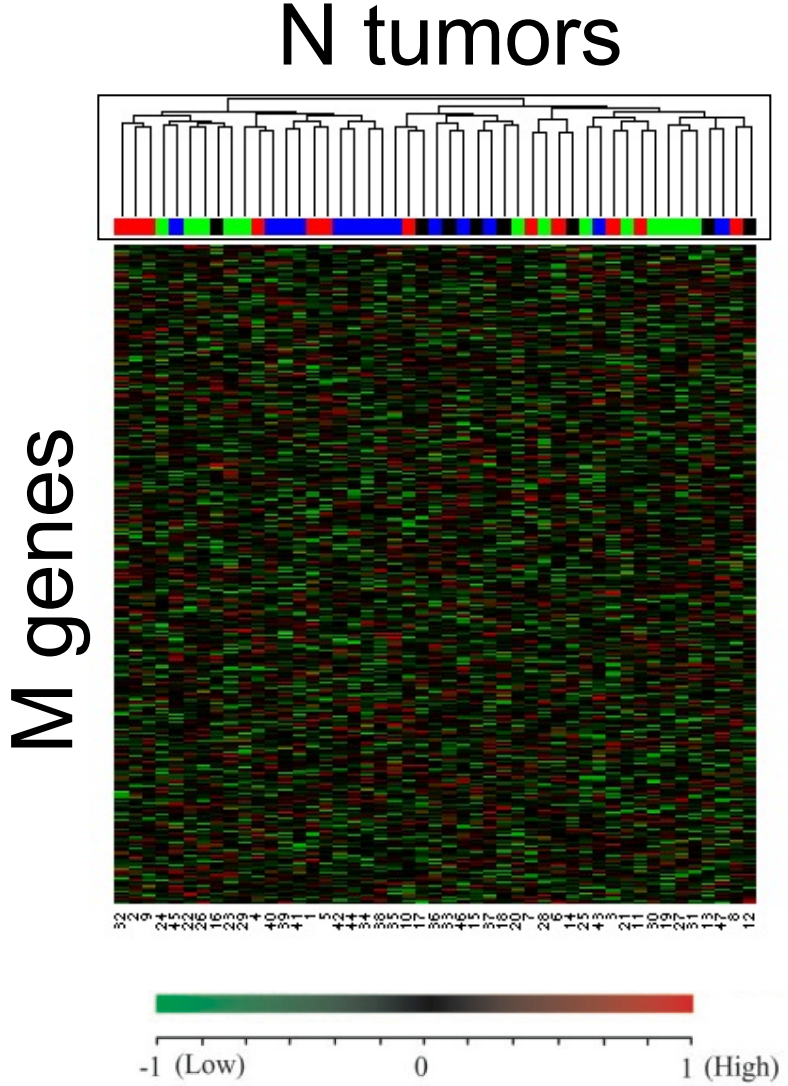
N tumors

M genes

	T1	T2	…	TN
G1	1.5	-1.0	…	0.3
G2	-0.5	3.0	…	0.2
G3	0.0	0.0	…	-0.1
…	…	…	…	…
GM	2.6	-0.2	…	1.1

The expression matrix



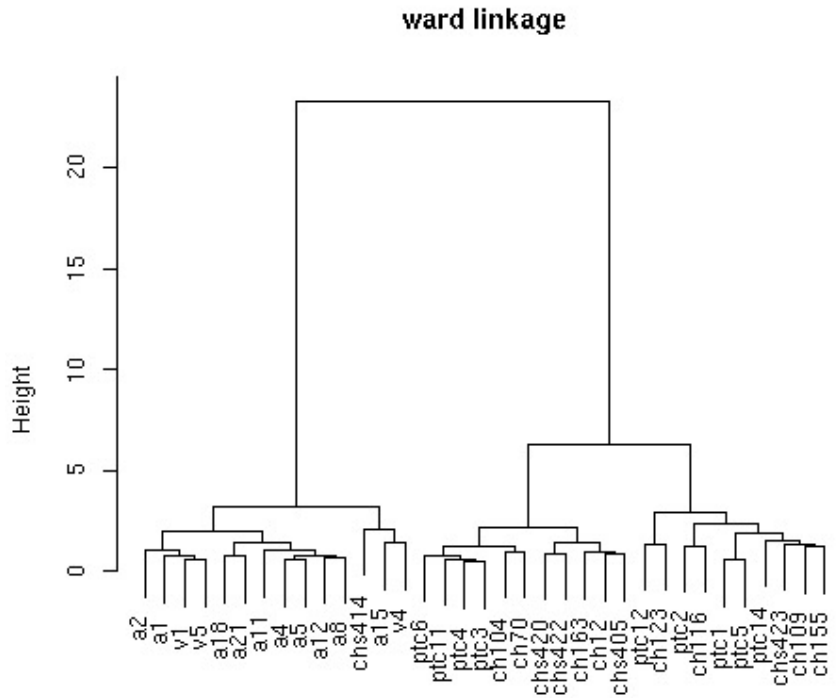


Unsupervised and supervised machine learning

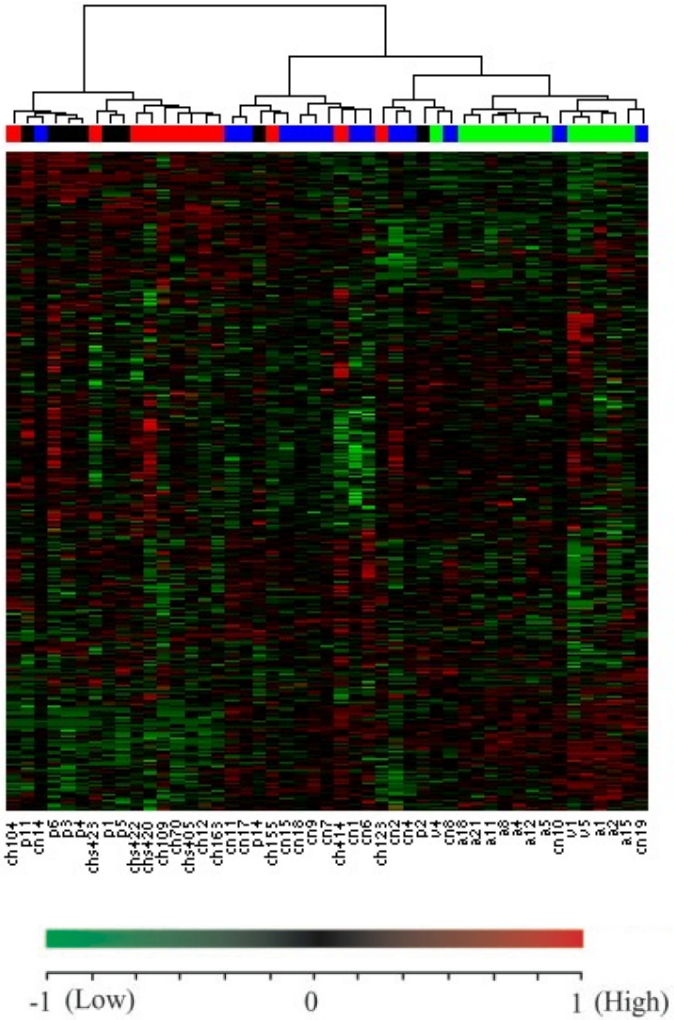
- With unsupervised learning the machine *discovers* a class structure in the data that is *not* known a priori

- With supervised learning the class structure of the data is *given* beforehand, and the machine finds gene expression patterns that *classify* tumors according to this structure

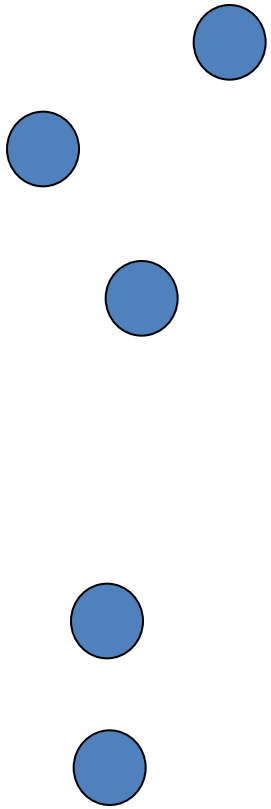
Hierarchical clustering uncovers sub-groups in the data



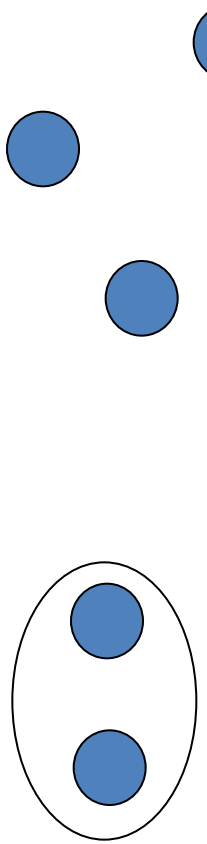
Hierarchical clustering produces dramatic displays for expression profiles



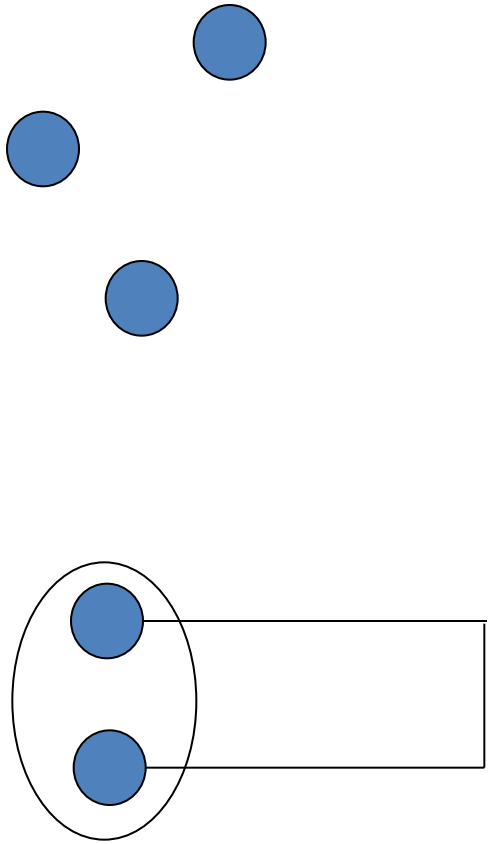
Hierarchical clustering algorithm at work



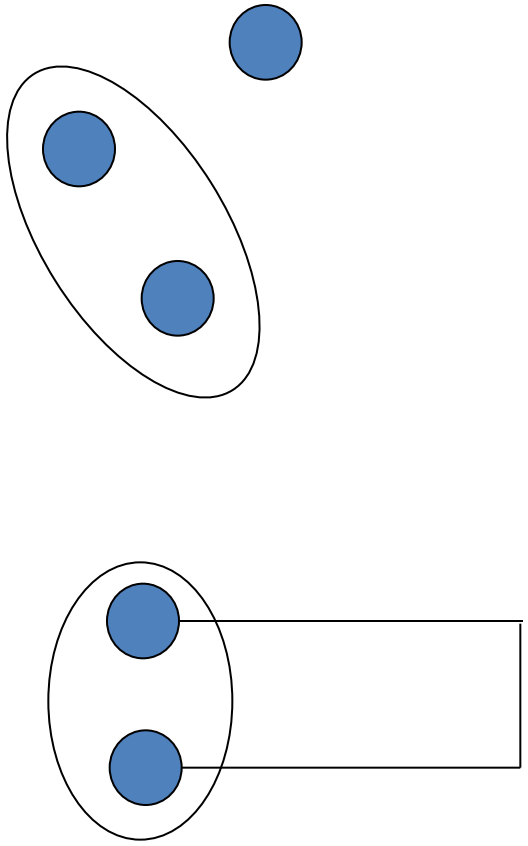
Hierarchical clustering algorithm at work



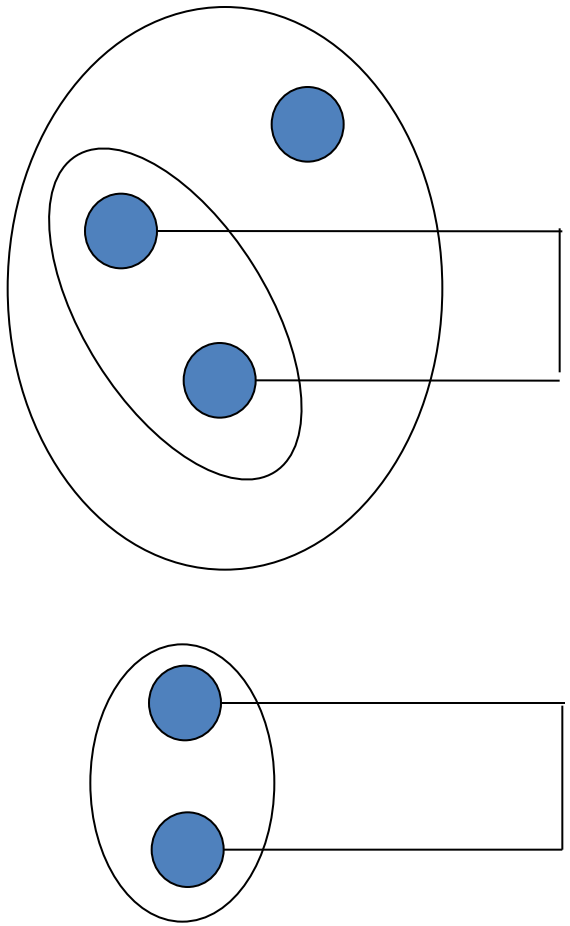
Hierarchical clustering algorithm at work



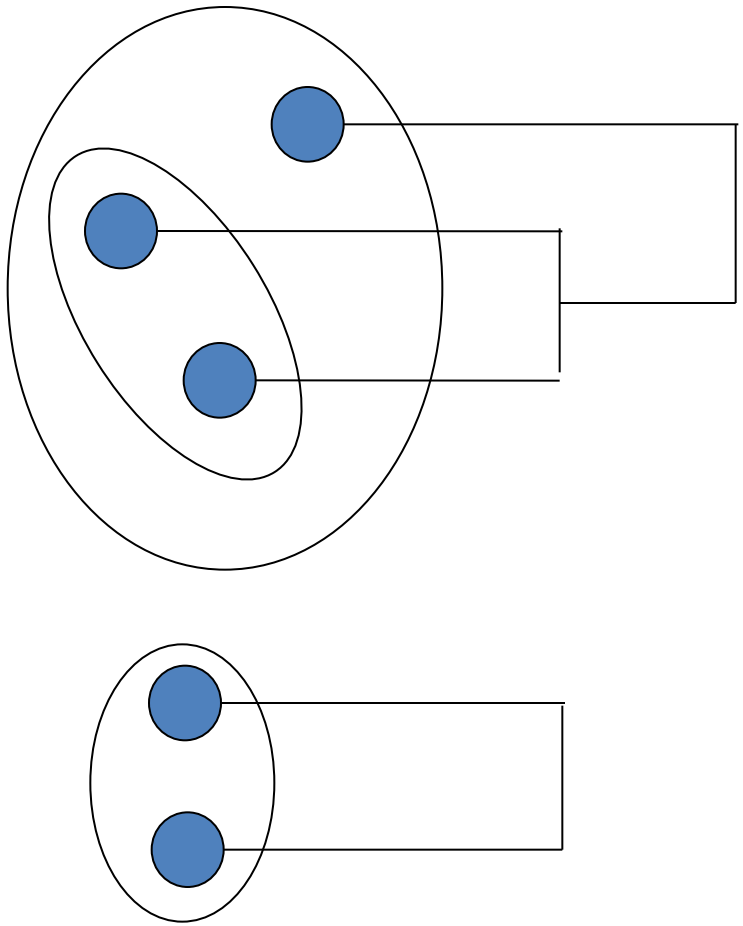
Hierarchical clustering algorithm at work



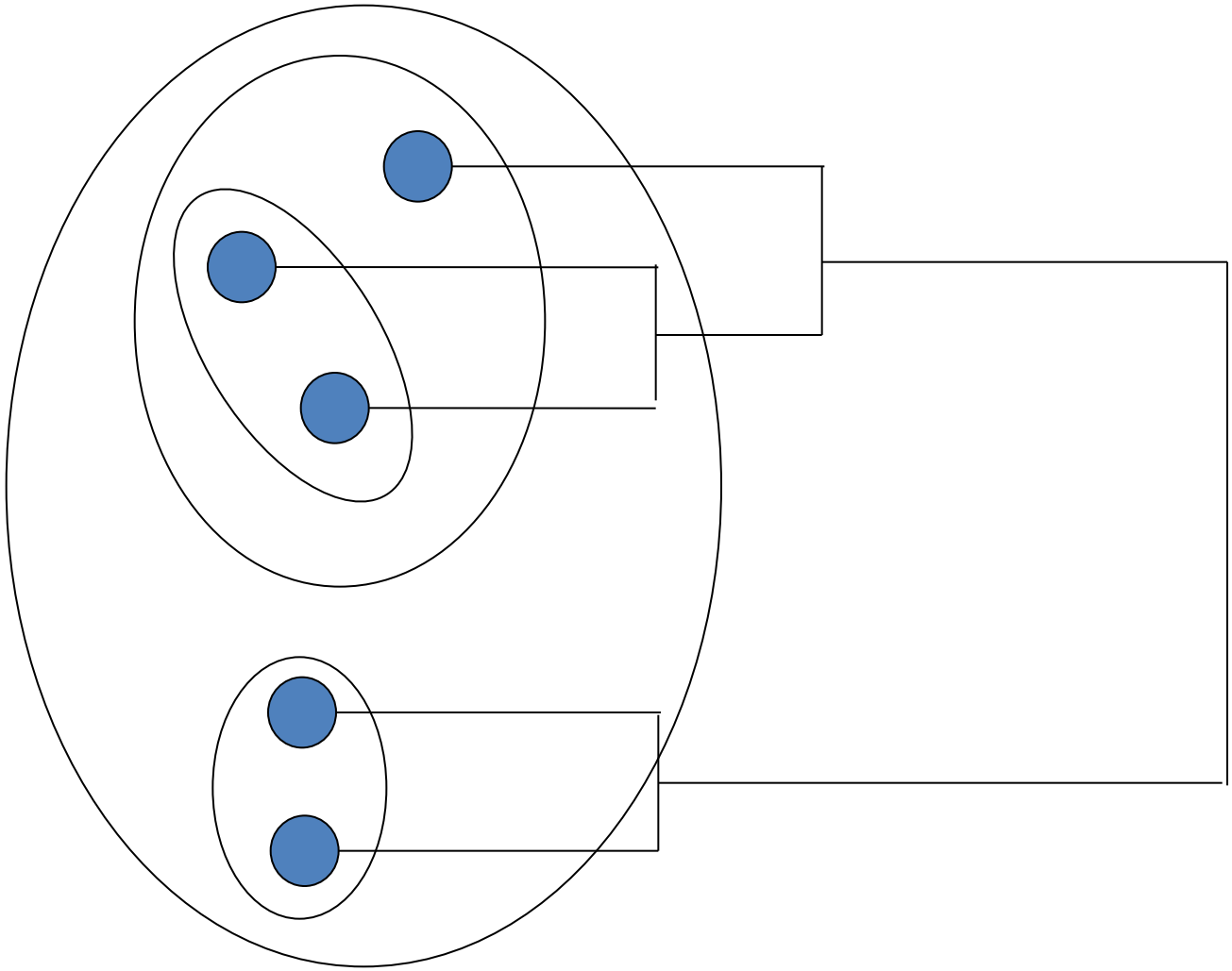
Hierarchical clustering algorithm at work



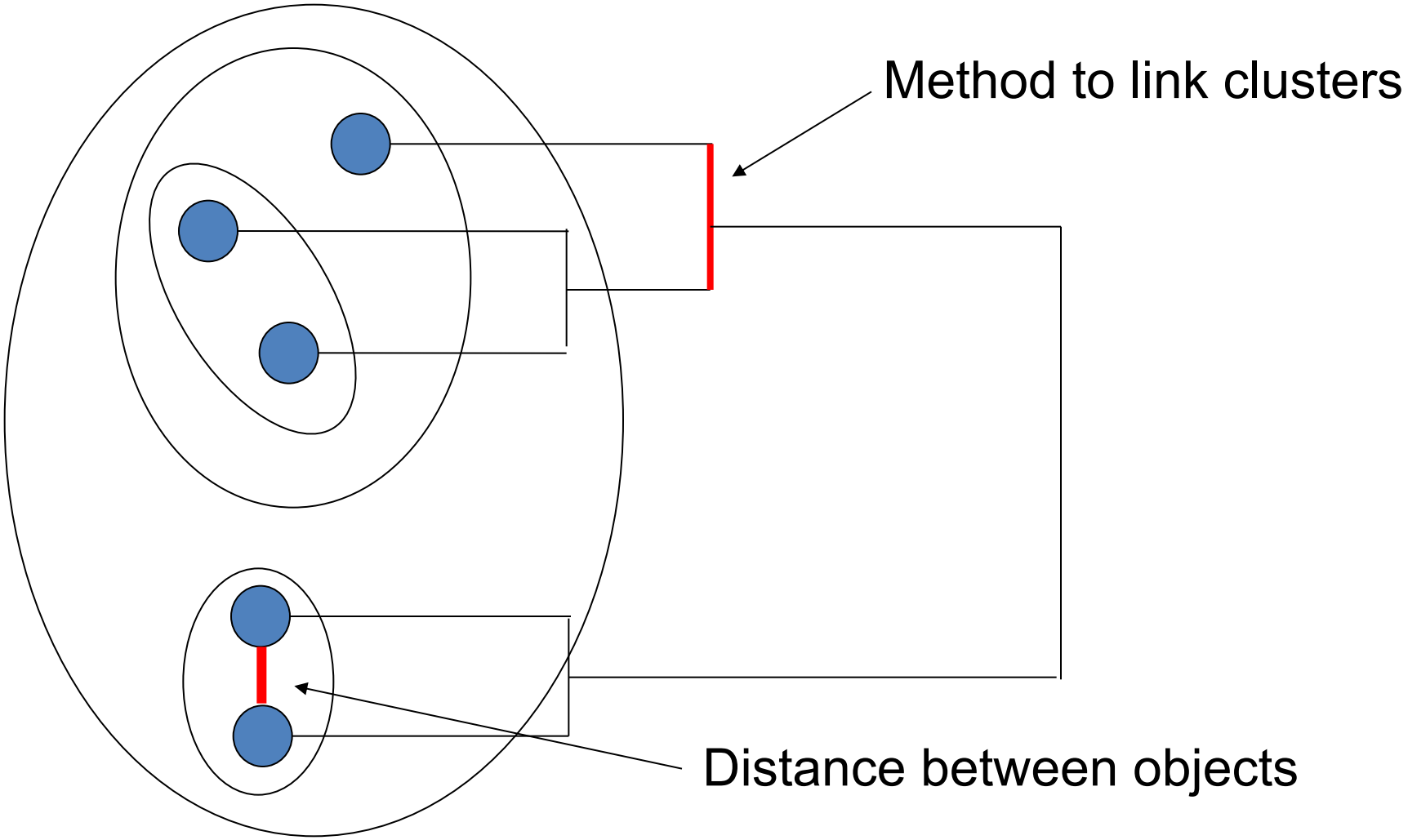
Hierarchical clustering algorithm at work



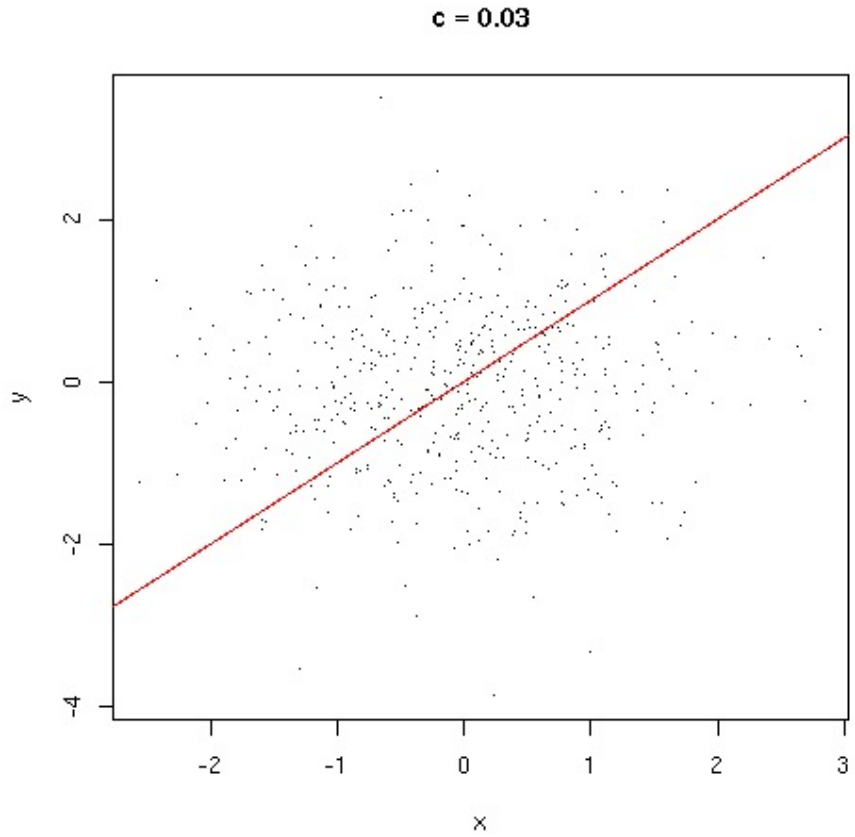
Hierarchical clustering algorithm at work



Hierarchical clustering comes in many (distance and linkage) flavors

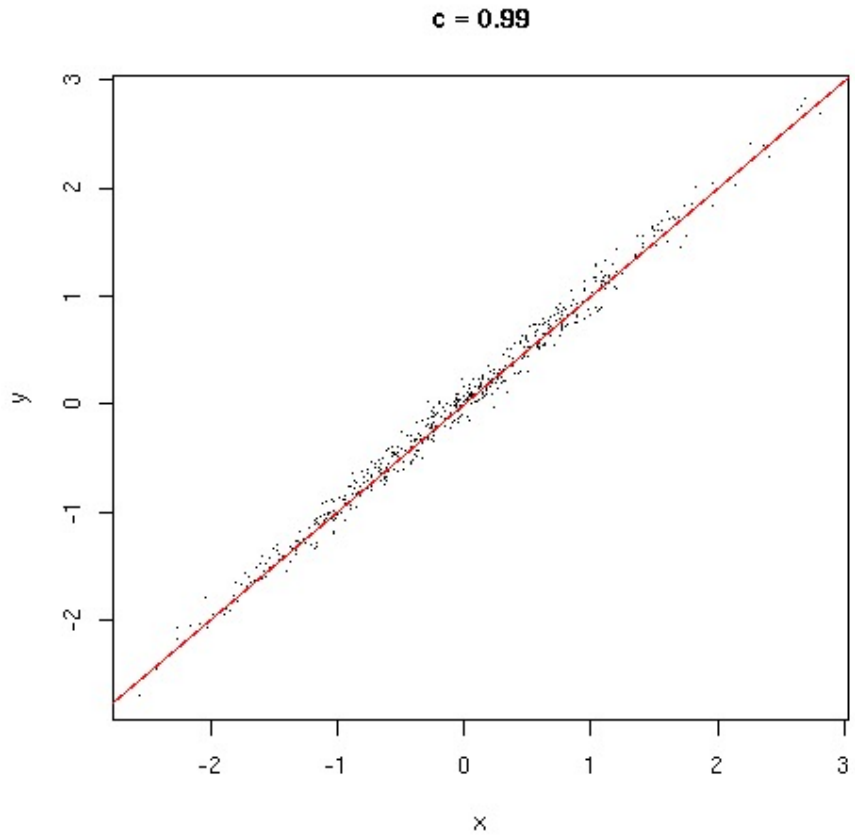


Correlation is a popular distance for clustering microarray data



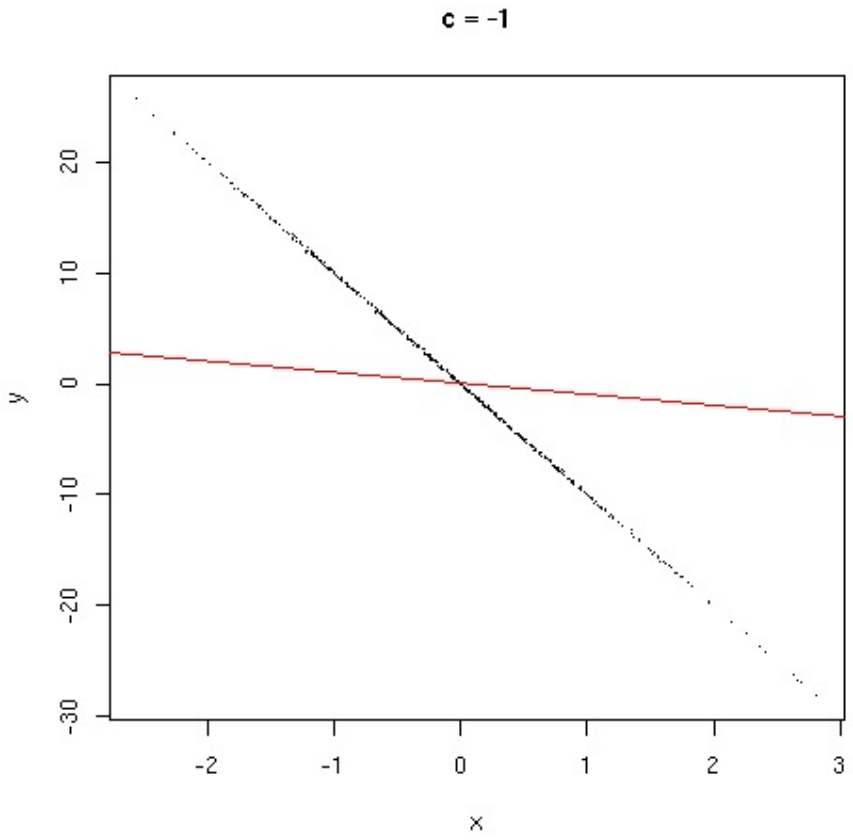
Unrelated profiles have null a correlation

Correlation is a popular distance for clustering microarray data



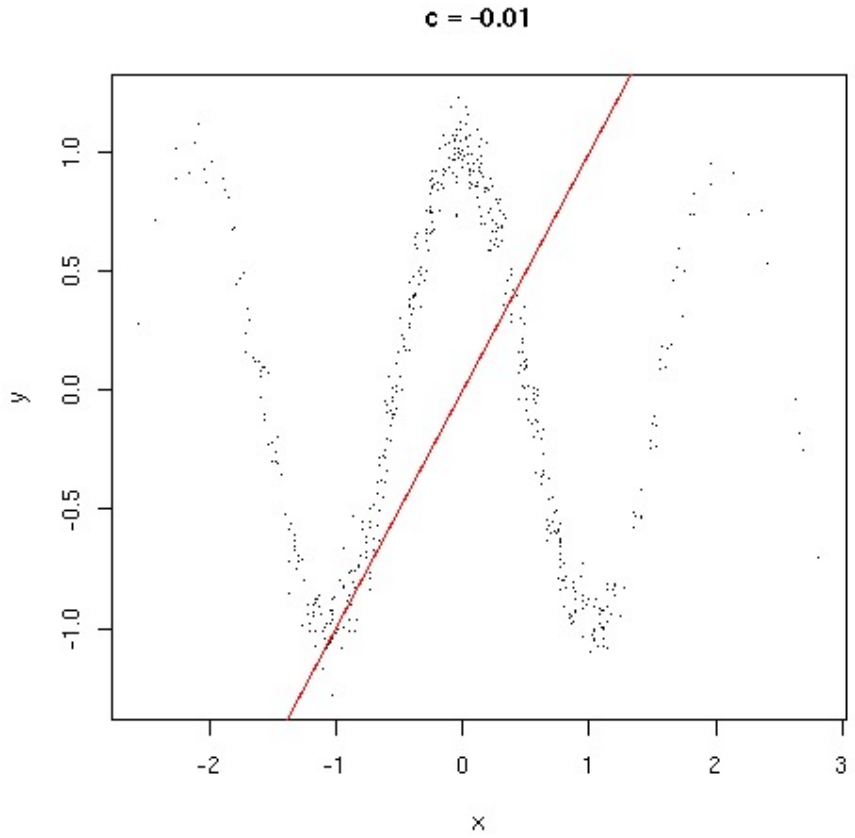
Similar profiles
have a unit correlation

Correlation is a popular distance for clustering microarray data



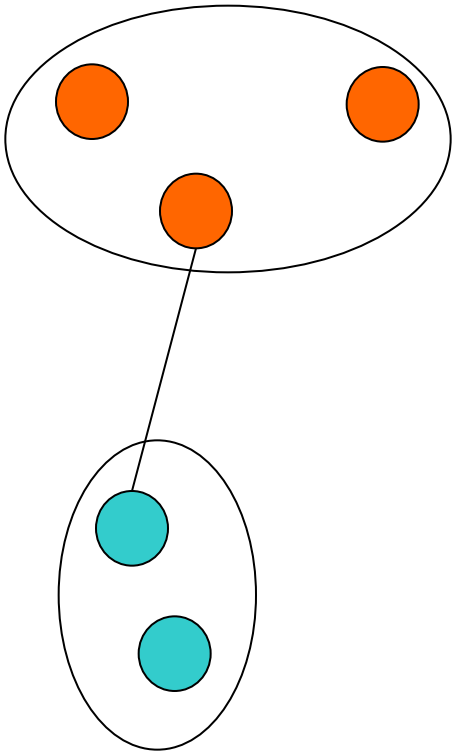
Correlation is scale invariant

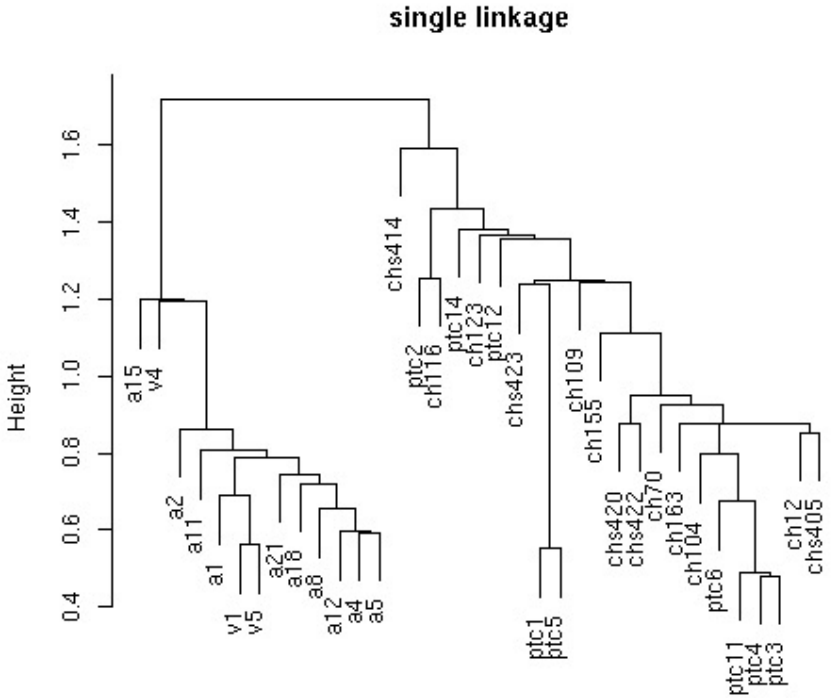
Correlation is a popular distance for clustering microarray data



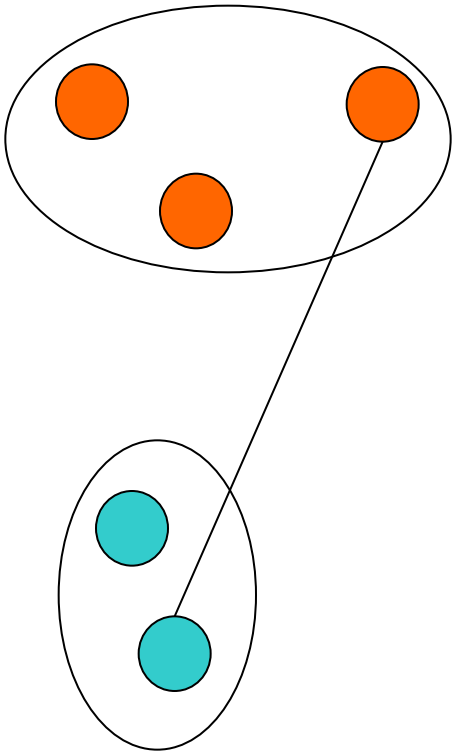
Correlation cannot detect nonlinear relations

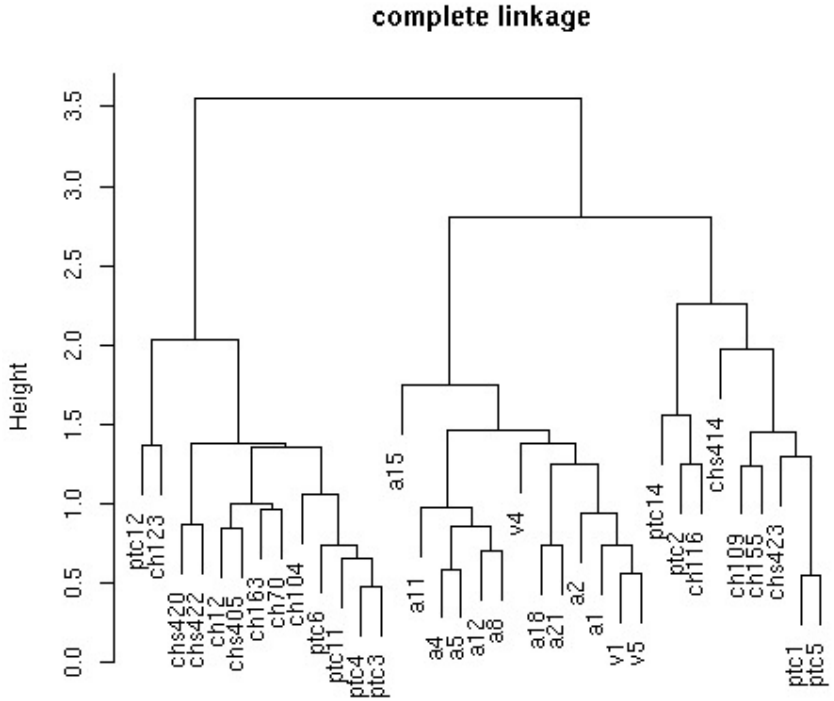
Single linkage associates clusters with smallest minimal distance



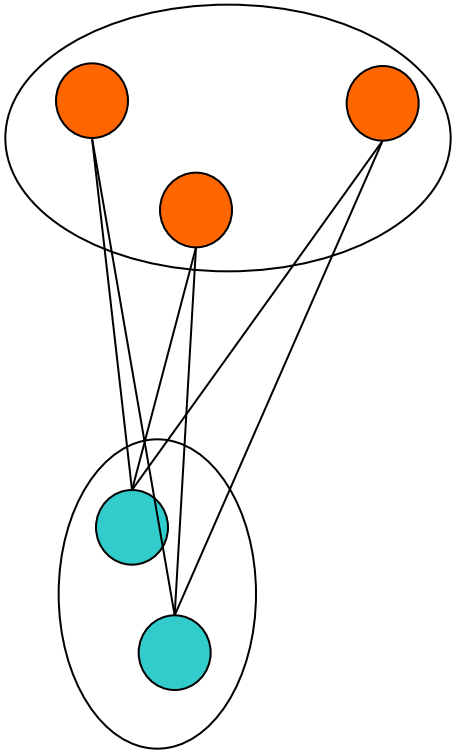


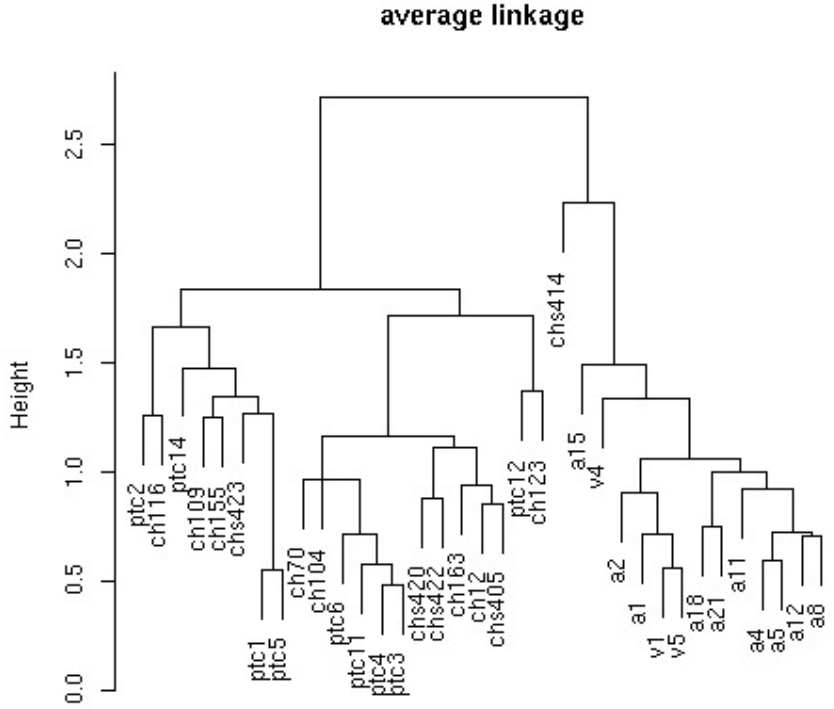
Complete linkage associates clusters with smallest maximal distance



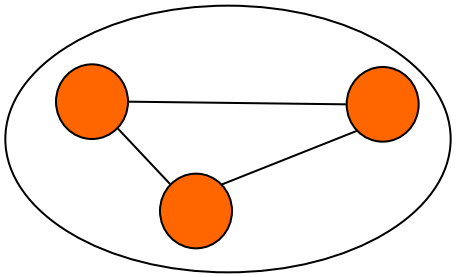


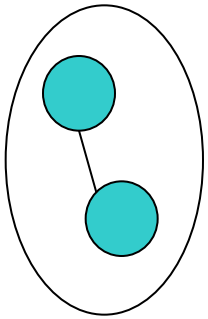
Average linkage associates clusters with smallest average distance

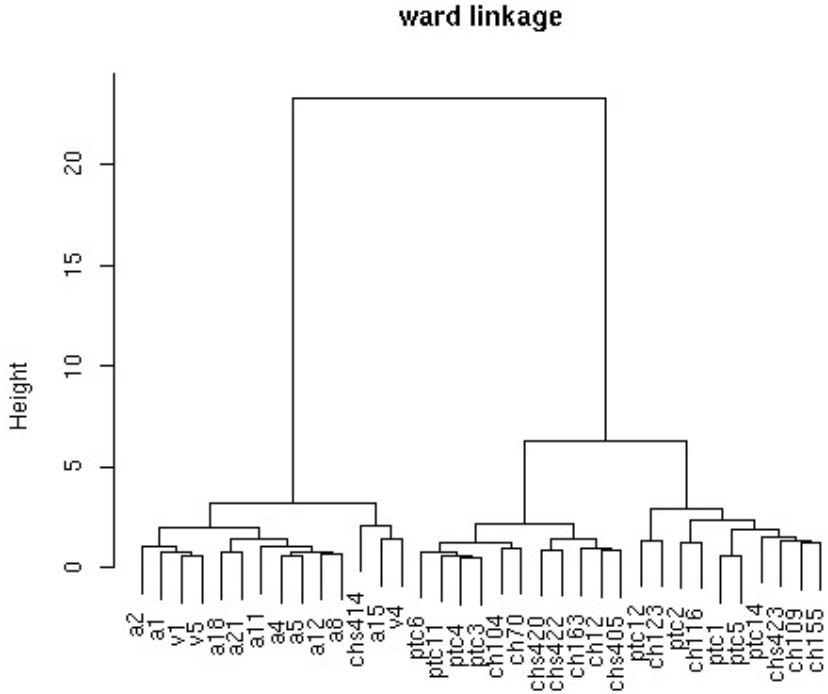




Ward linkage associates clusters as to minimizes within cluster variance





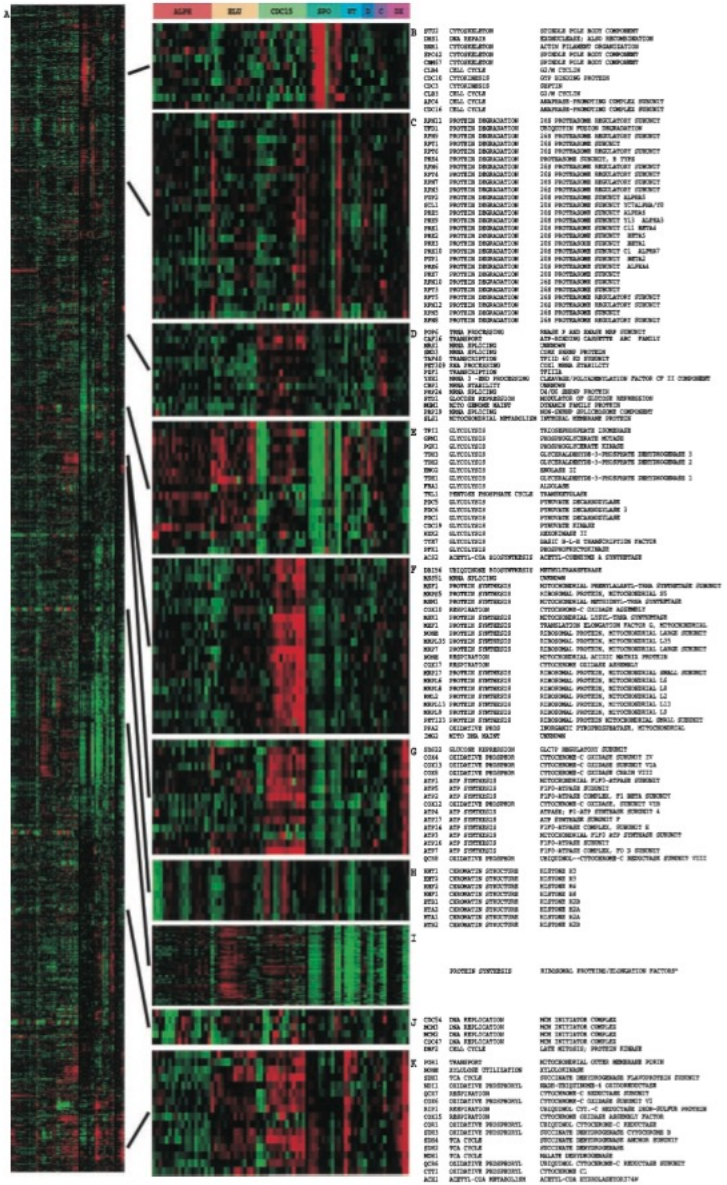


Hierarchical clustering reveals structures in expression profiles

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

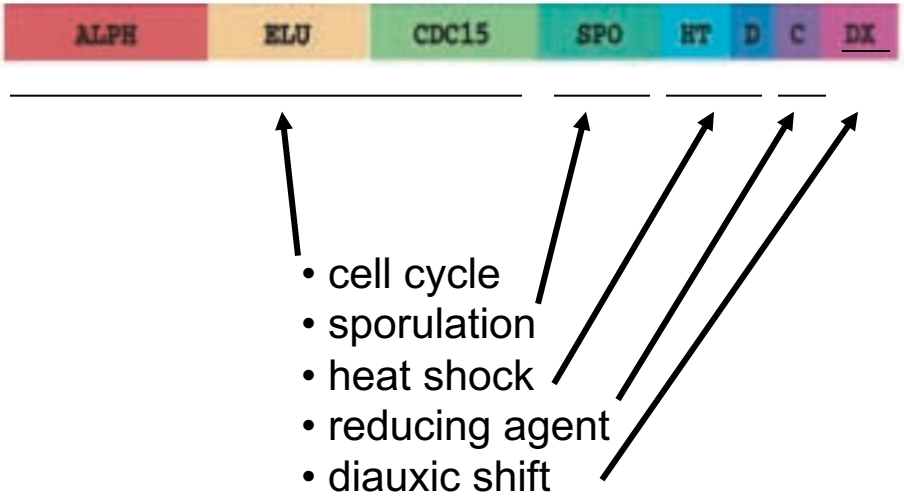
MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡



From Eisen et al., PNAS 1998

Yeast expression profiles were measured under various conditions

Two-ways hierarchical clustering was then applied to the expression matrix



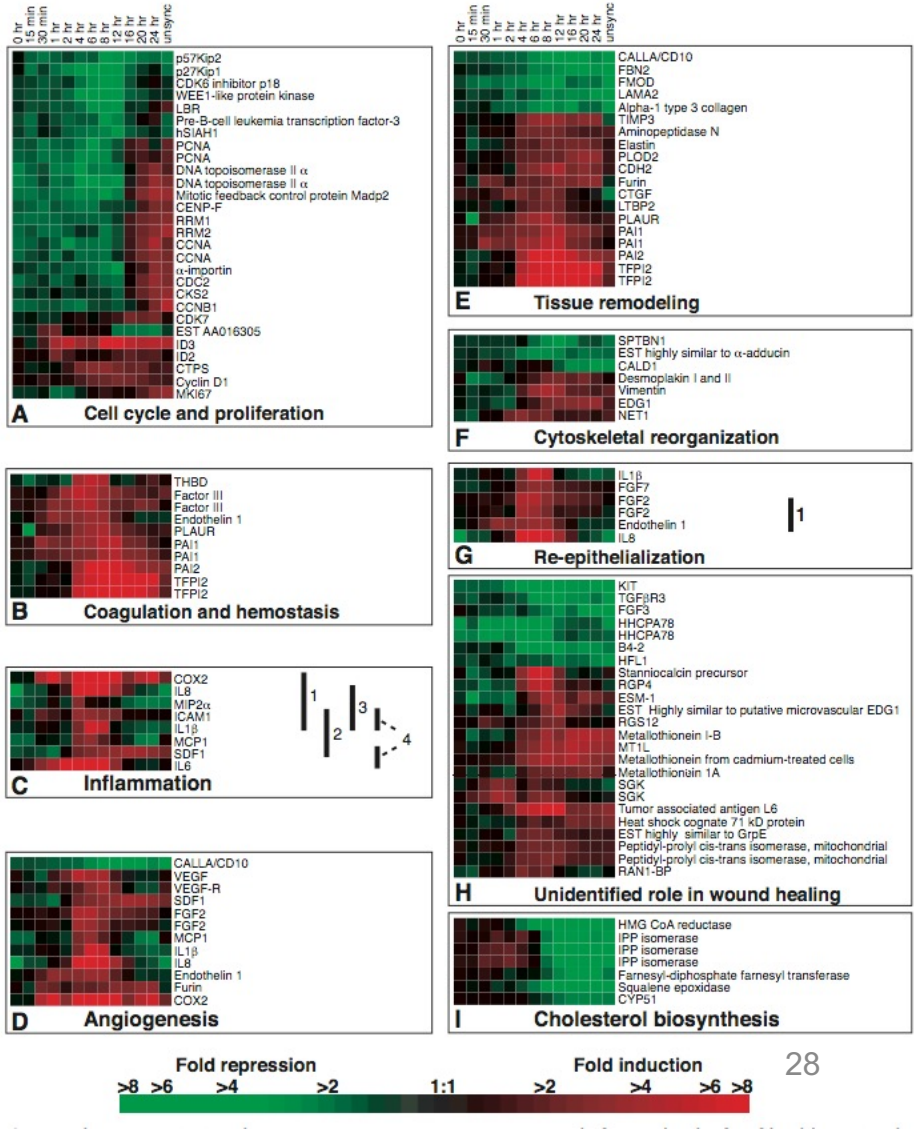
Hierarchical clustering reveals the breadth of universal phenotypes

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

Iyer et al., Science 1999

- human fibroblasts were exposed to serum
- their expression profiles revealed a response akin to wound healing
- proliferaton is only part of this complex response



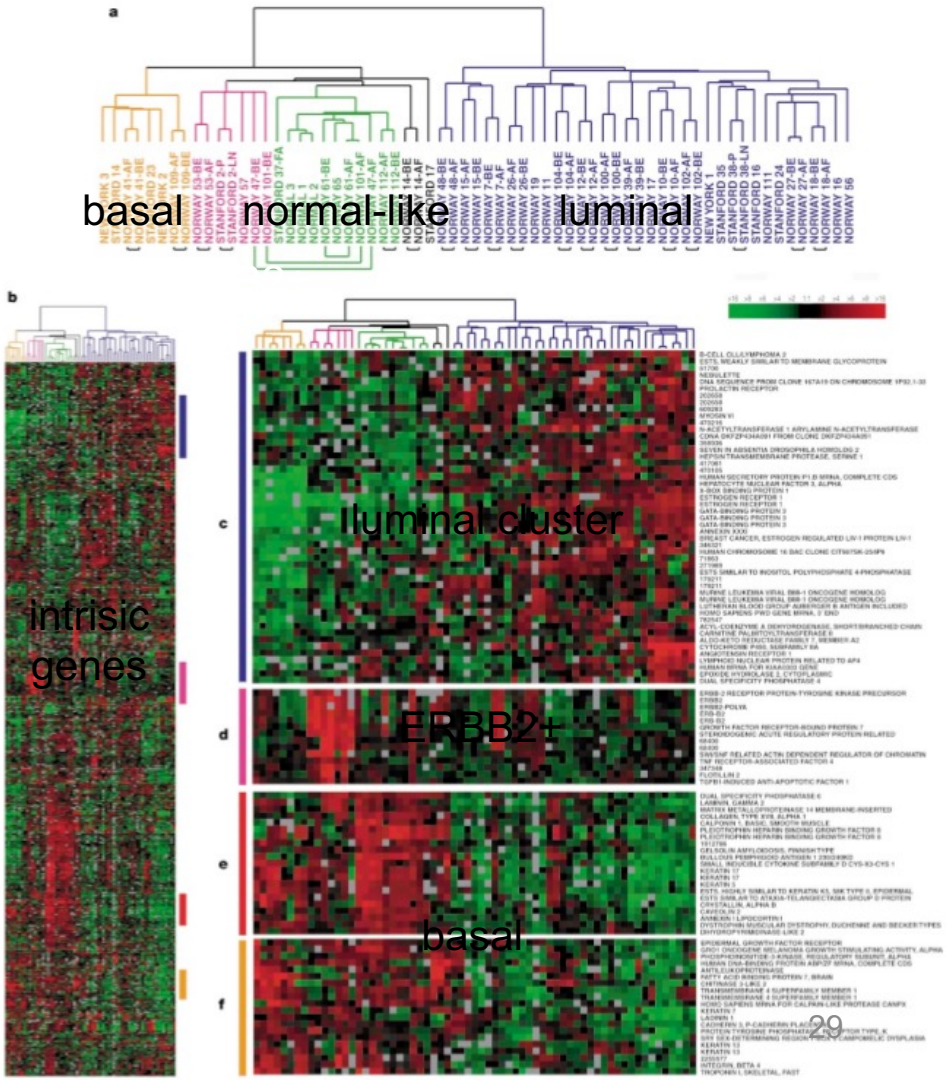
Hierarchical clustering reveals molecular cancer subtypes

letters to nature

Molecular portraits of human breast tumours

Charles M. Perou*†, Therese Sørlie†‡, Michael B. Eisen*, Matt van de Rijn§, Stefanie S. Jeffrey‖, Christian A. Rees*, Jonathan R. Pollack¶, Douglas T. Ross¶, Hilde Johnsen‡, Lars A. Akslen#, Øystein Fluge☆, Alexander Pergamenschikov*, Cheryl Williams*, Shirley X. Zhu§, Per E. Lønning**, Anne-Lise Børresen-Dale‡, Patrick O. Brown¶†† & David Botstein*

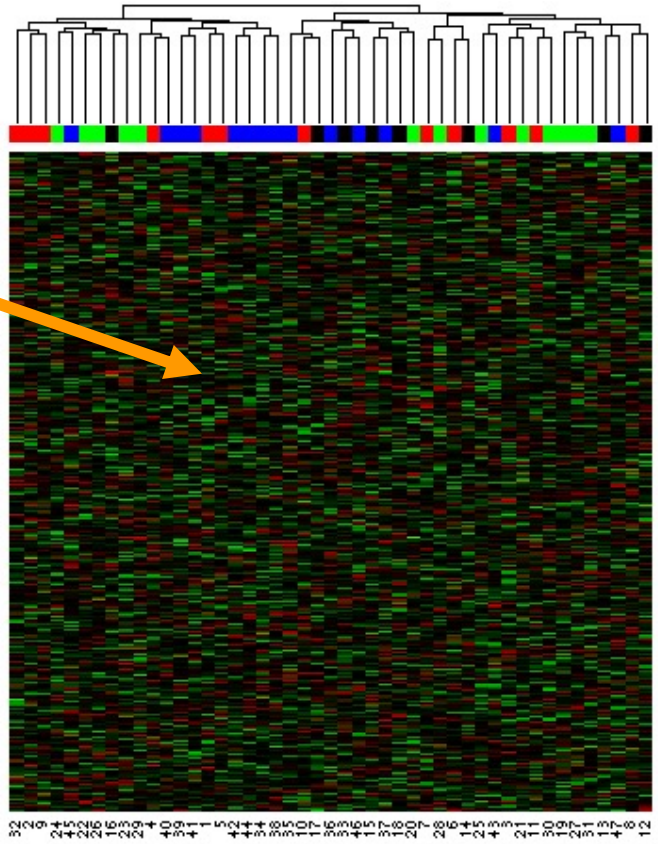
Perou et al., Nature 2000



- clustering from 'intrisic', i.e. most variable, genes
- suggests that ER$^-$ tumors are *not* one but two diseases: basal-like and ERBB2$^+$

Clustering produces clusters even when *no* clusters are there !!!

Random data



Limits of hierarchical clustering

- ✓ It imposes a discrete structure on continuous data.
- ✓ Data do not always have a hierarchical structure.
- ✓ Classes may not be mutually exclusive
- ✓ These is no standard quality measure of how relevant a clustering is.
- ✓ It is unstable: adding one sample may incur dramatic changes in clustering pattern.

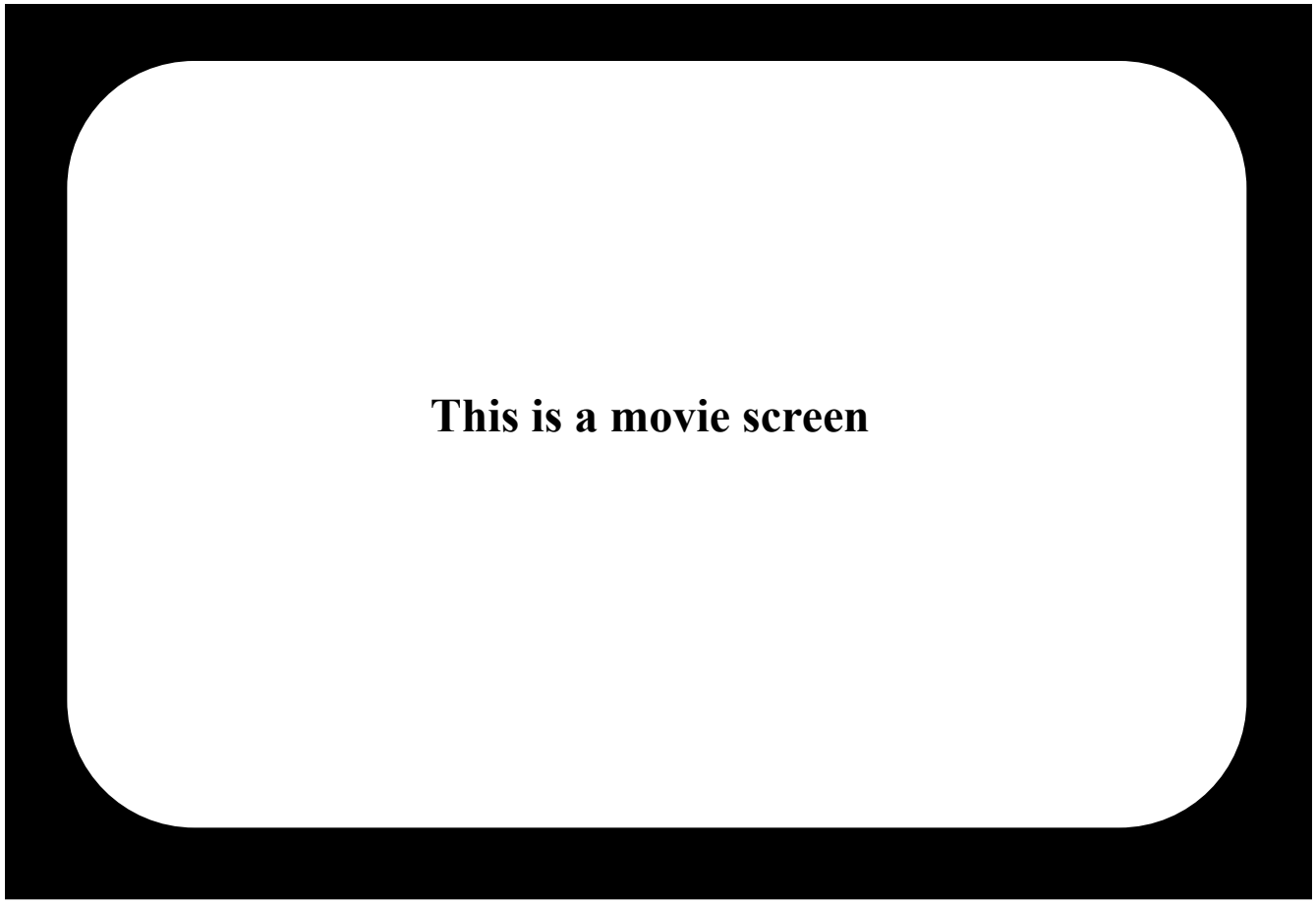
Misuses of hierarchical clustering

- ✓ It often used to (supposedly) assess signatures of classes which are known *a priori*. Supervised classification and cross-validation should be used instead.

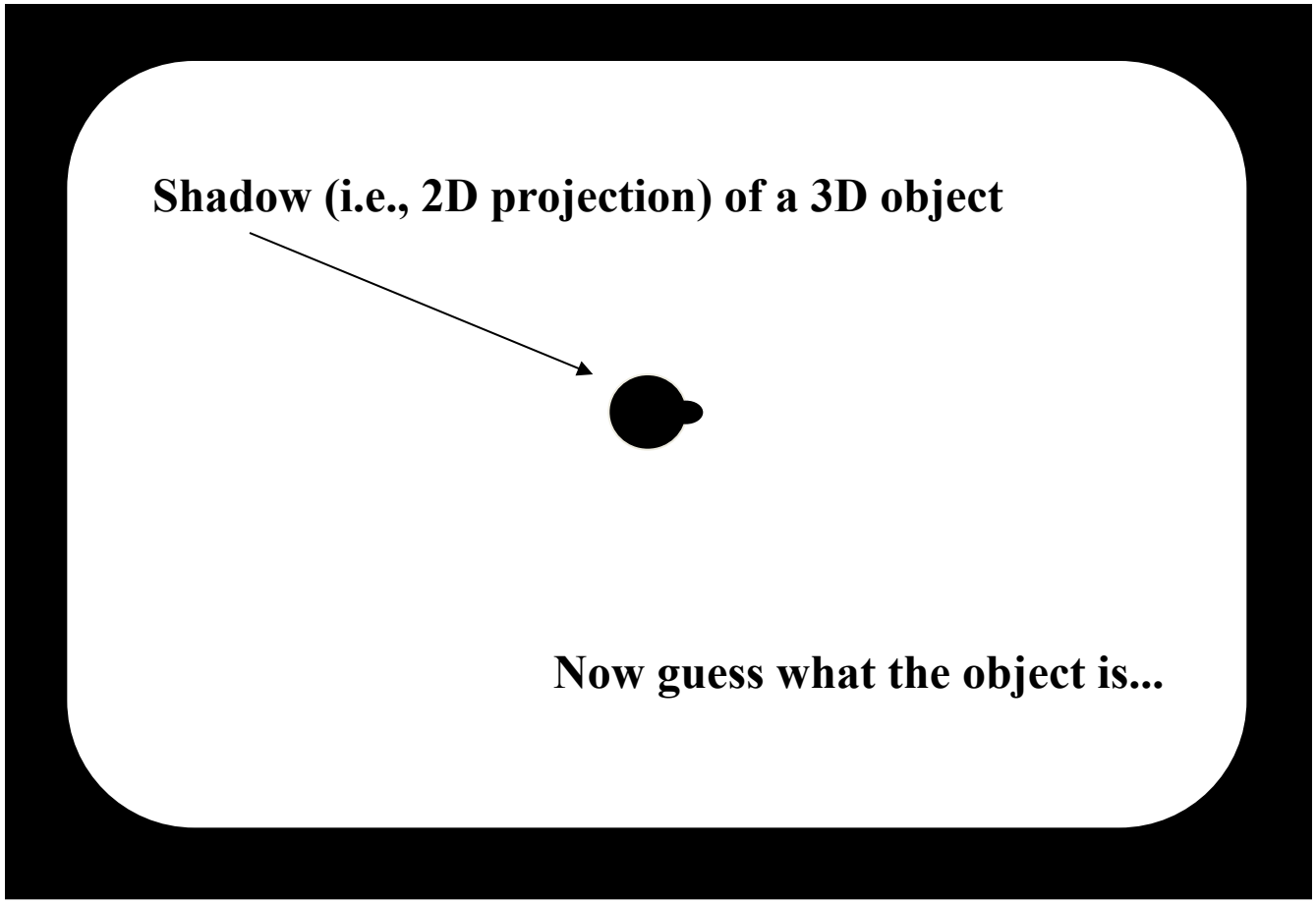
- ✓ Clustering sometime relies on a small subset of predetermined class-separating genes. Such genes may be found by over-fitting any high dimension data set.

Principal components analysis (PCA)

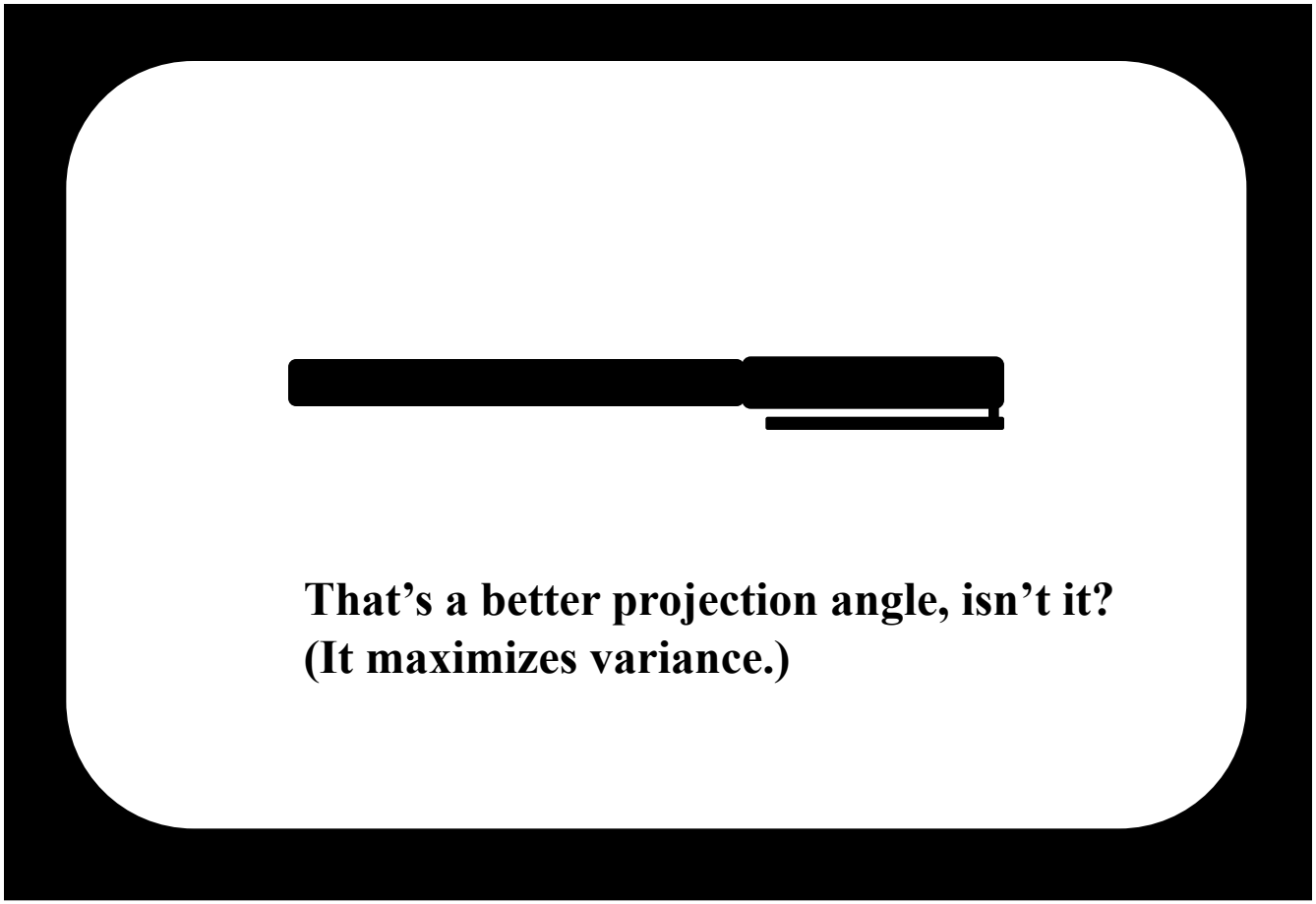
Some projections of the data are more informative than others



Some projections of the data are more informative than others

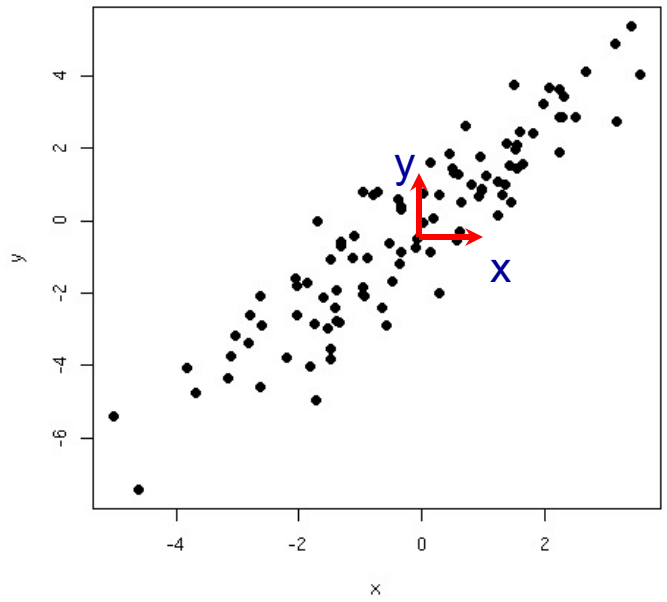


Some projections of the data are more informative than others

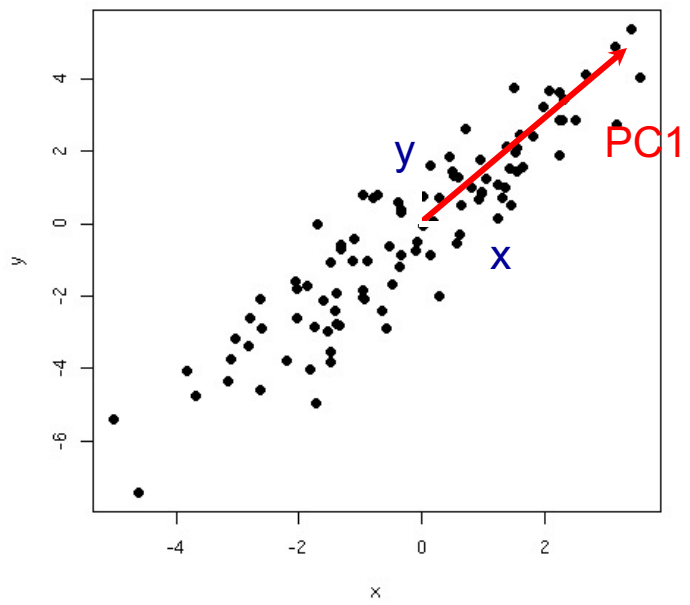


That's a better projection angle, isn't it?
(It maximizes variance.)

Principal components analysis (PCA) computes the projection of the data that explains the best its variance

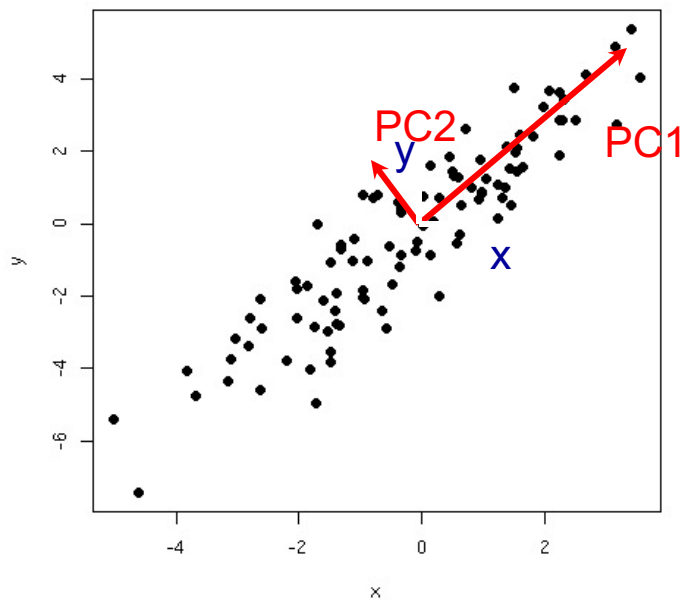


Principal components analysis (PCA) computes the projection of the data that explains the best its variance



- PCA iteratively finds basis vectors (i.e. components) that maximise data variance

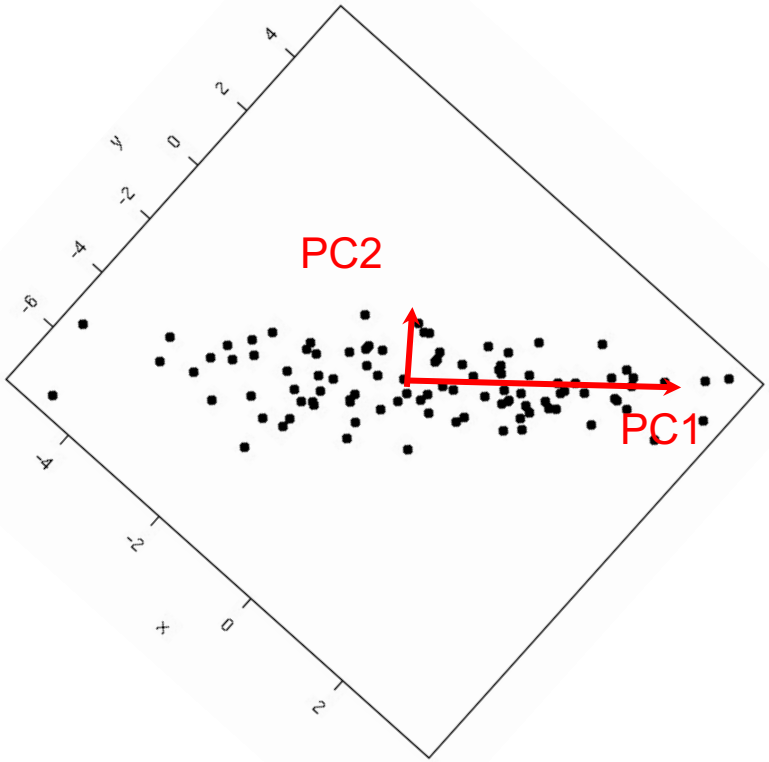
Principal components analysis (PCA) computes the projection of the data that explains the best its variance

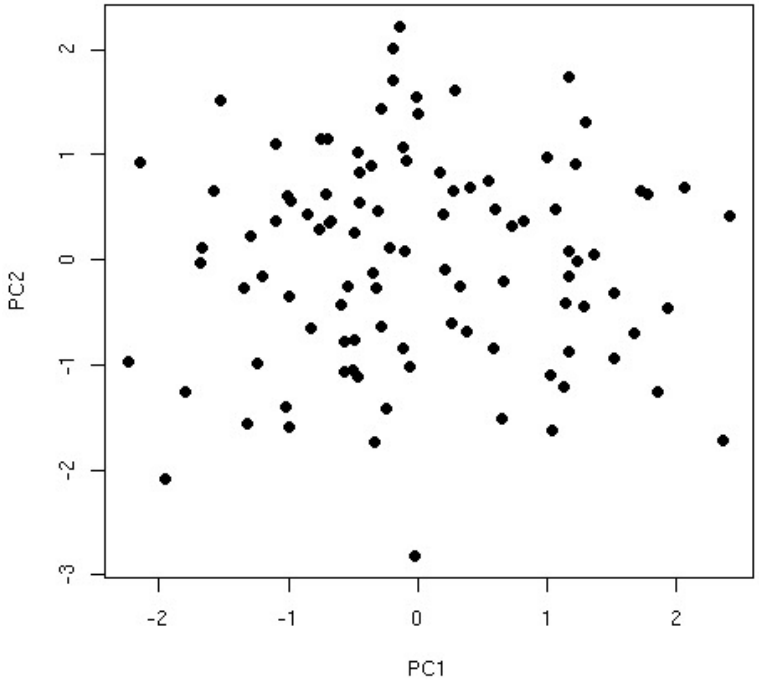


- PCA iteratively finds basis vectors (i.e. components) that maximise data variance

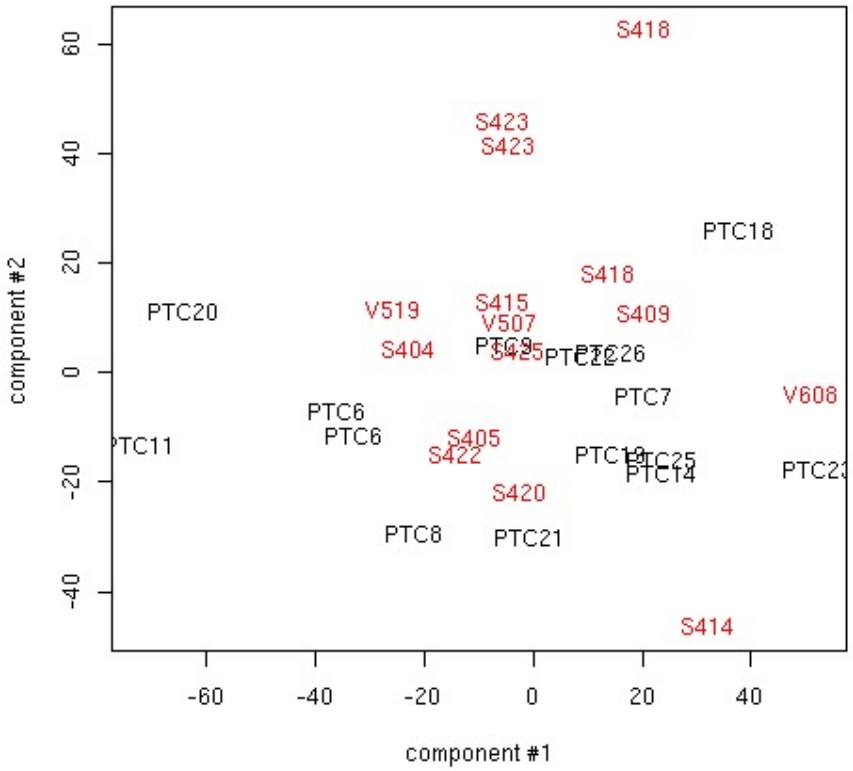
- Components are chosen orthogonal

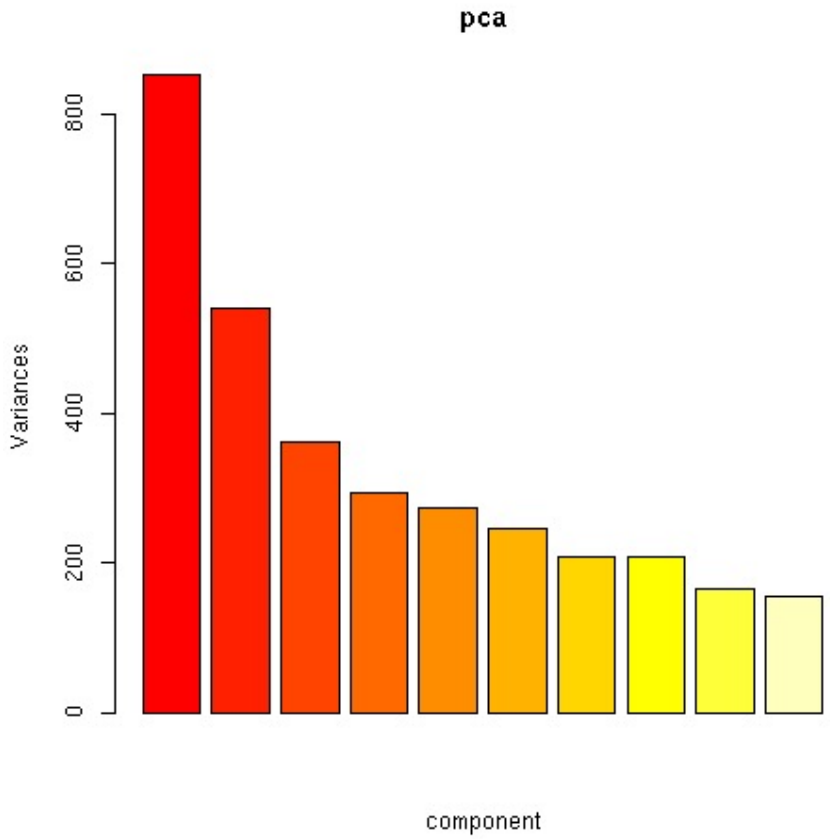
Principal components analysis (PCA) computes the projection of the data that explains the best its variance



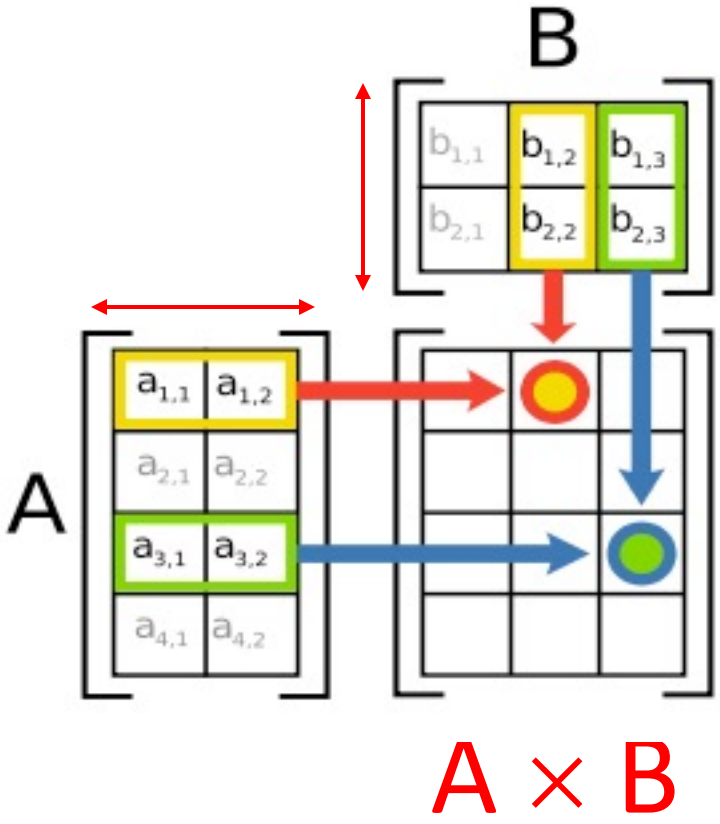


Principal components analysis (PCA) computes the projection of the data that explains the best its variance





Matrix product may represent any linear transformation



$(A \times B)_{1,2} = a_{1,1} \times b_{1,2} + a_{1,2} \times b_{2,2}$

$(A \times B)_{3,3} = a_{3,1} \times b_{1,3} + a_{3,2} \times b_{2,3}$

here the rank, k=2

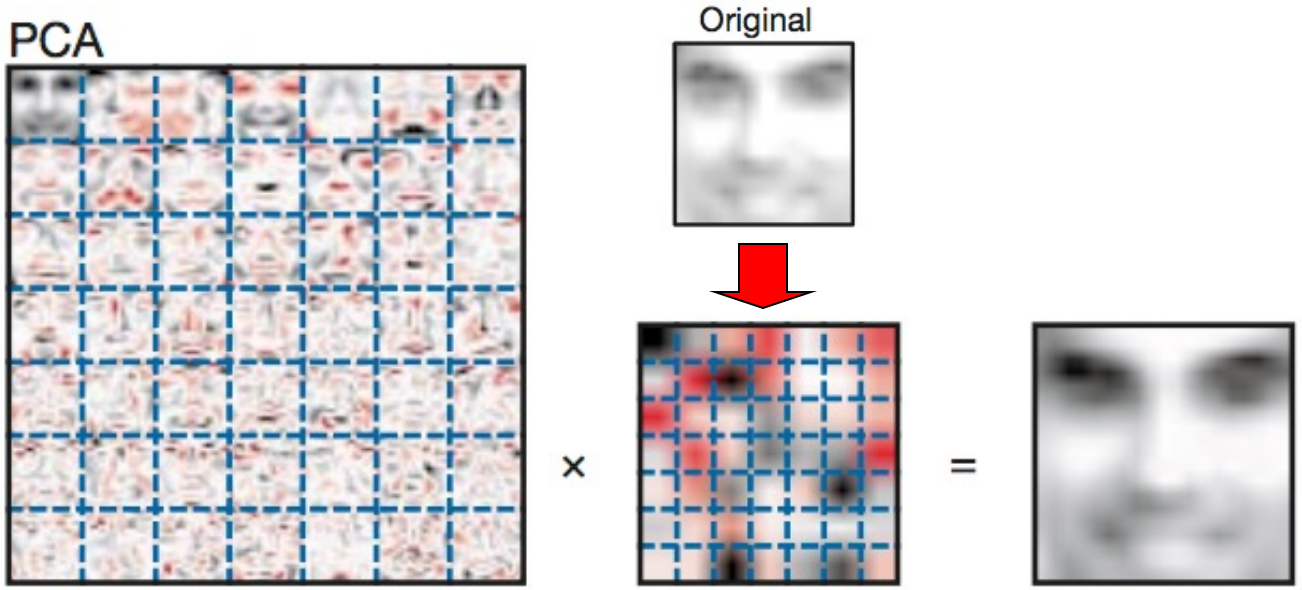
A linear transformation is a combination of rotation, translation and stretching of individual base vectors

Example of PCA application: compression of a faces database

Orthogonal ‘face’ basis

encoding

recovery

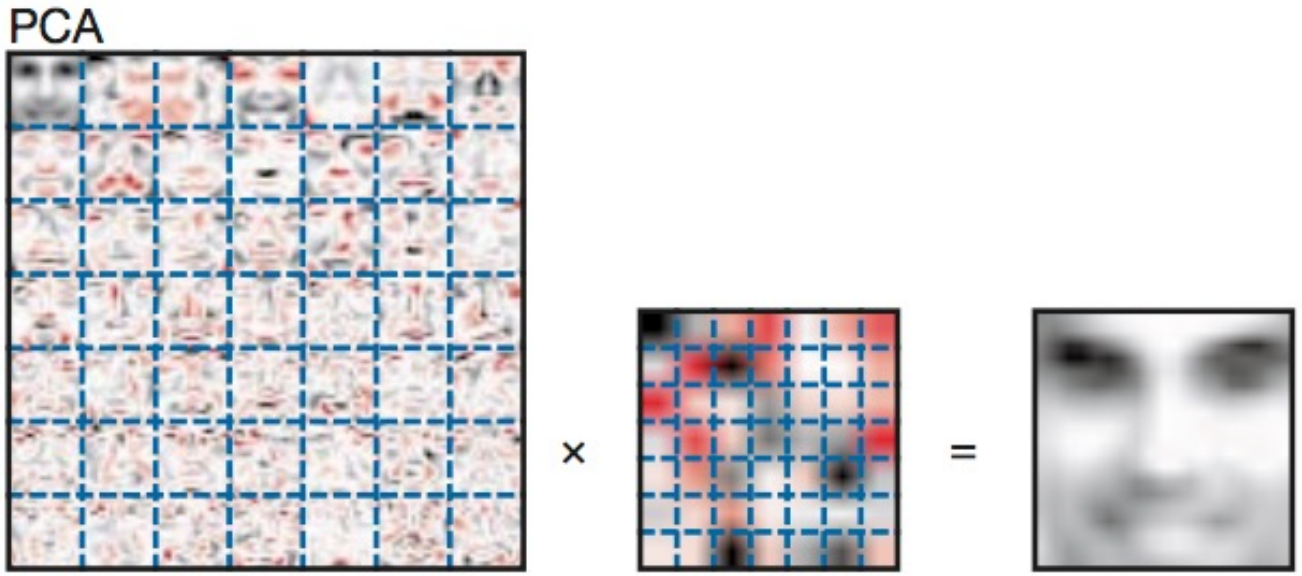


from Lee & Seung, Nature 1999

Example of PCA application: compression of a faces database

transmit 49
'face' vectors

transmit 2,500
encoding vectors



from Lee & Seung, Nature 1999

Example of PCA application: compression of a faces database

- Size of image database = 19x19x2500
- Size of PCA-compressed database = size of base vectors + size of faces encodings = 19x19x49 + 49x2500
- Compression ratio about 6.4

What works for faces works for expression matrices

$$A = W \times H$$

faces database = face basis × face encodings

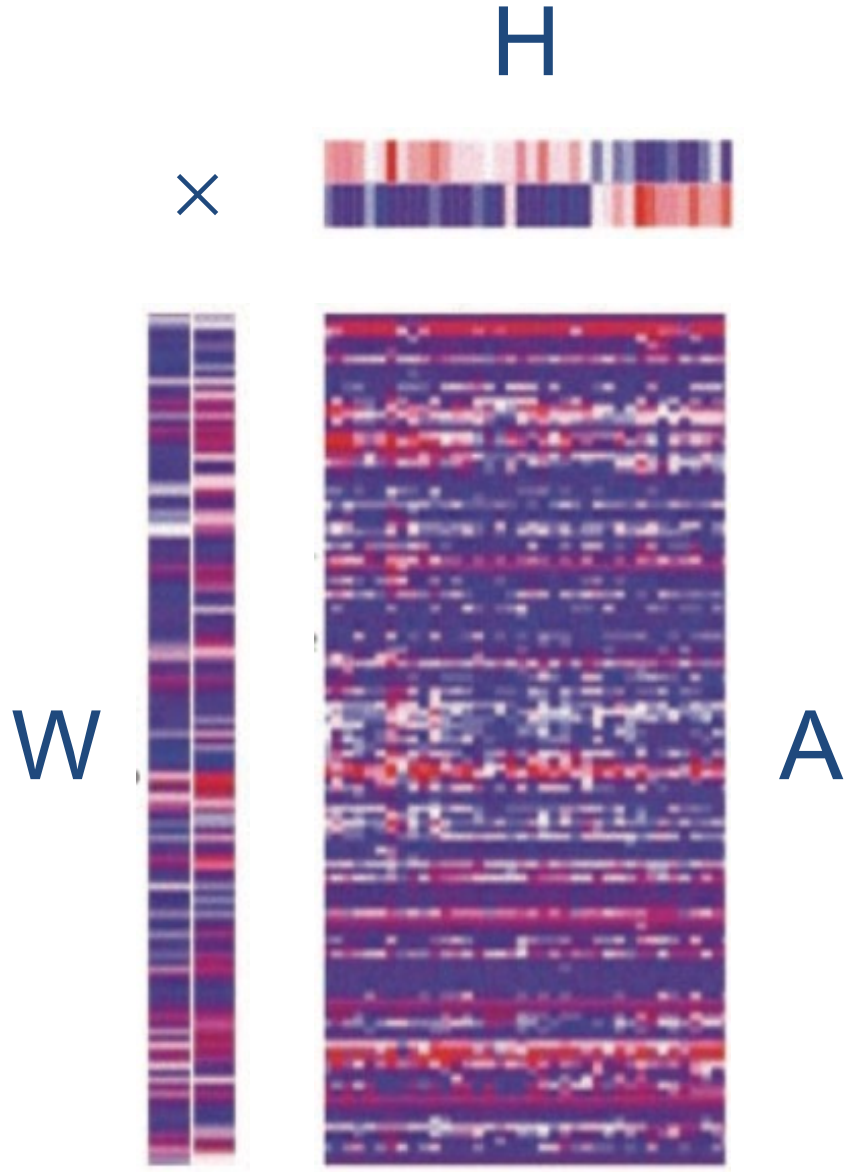
expression matrix = sample basis × sample encodings

$A \approx W \times H$

A: aproximate
expression matrix

W: sample basis

H: sample encoding
in term of k=2
'super genes'



How variable is mRNA expression among ethnic groups?

Let's have a look at the data from this paper:

Common genetic variants account for differences in gene expression among ethnic groups

Richard S Spielman[1], Laurel A Bastone[2], Joshua T Burdick[3], Michael Morley[3], Warren J Ewens[4] & Vivian G Cheung[1,3,5]

Nature Genetics, 2007

Why is the question important?

- It may help in the design of clinical trails and health policies by identifying distinct disease susceptibility and/or drug response profiles in specific populations

- It is underlying the origin and diversity of humans

- It is politically loaded by an *unthinkable* history of racism and ethnic/racial violence across the world, some of it very recent and close to home

The populations investigated by Spielman et al.

166 individuals were profiled

- 60 Caucasian Europeans
- 41 Japanese
- 41 Chinese living in Beijin, China

HapMap individuals

- 24 Chinese living in Los Angeles, USA. These are controls for the effect of environment and life-style

The data of Spielman et al.

- mRNA was extracted from Epstein-Barr virus-transformed lymphoblastoid cell lines
- Expression was measured with Affymetrix HG Focus arrays covering ~8,500 genes
- They are publicly available from www.ncbi.nlm.nih,gov/GEO, accession number GSE5859

Follow up: pervasive interethnic difference or batch effect?

A few months later Akey et al. (Nature Genet. 2007) revisited Cheung's study:

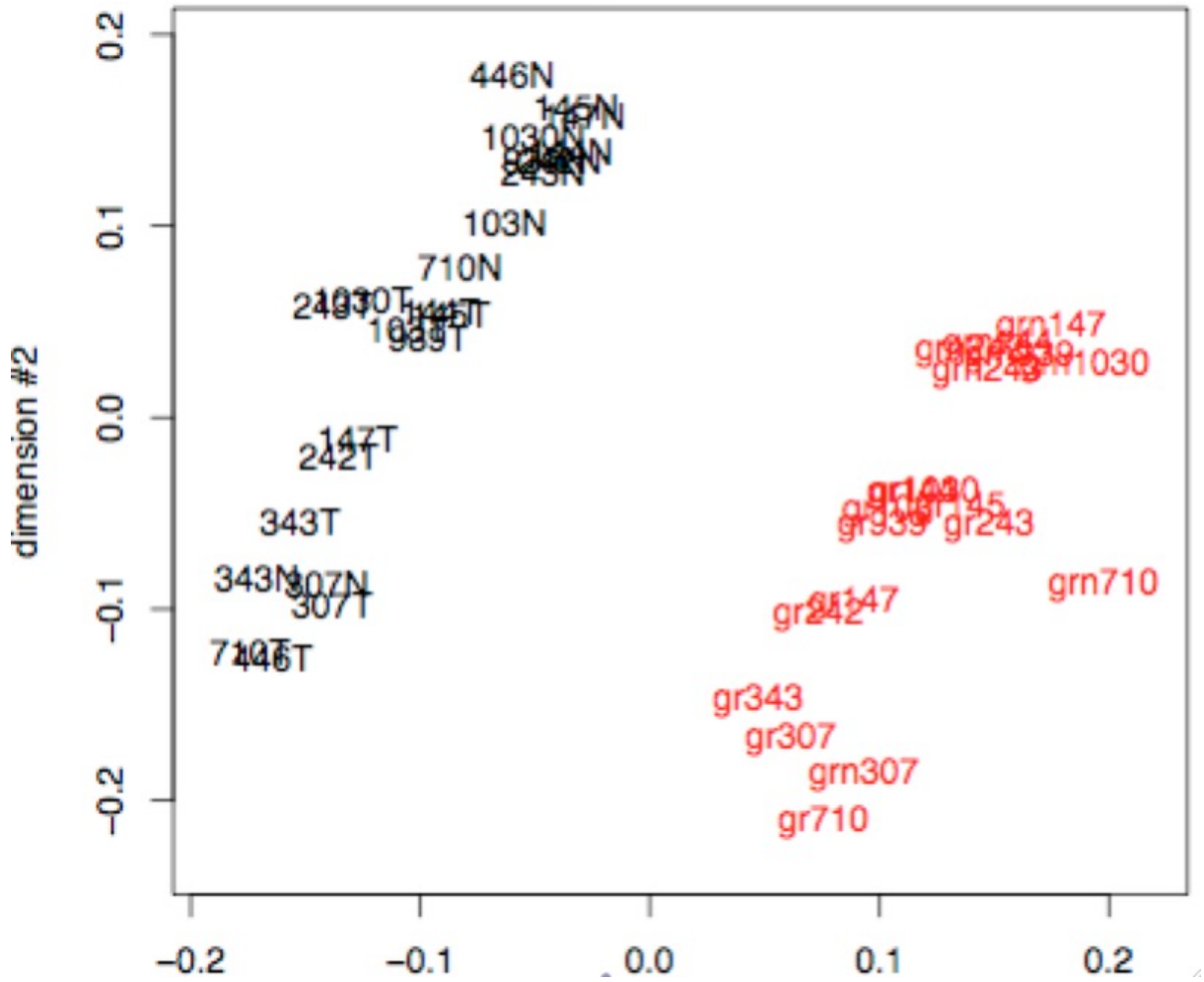
- "To explore these issues in more detail, we downloaded the raw CEL files from Gene Expression Omnibus (GSE5859) and extracted from the header line the date on which the file was created."

- "Interestingly, the arrays used to measure expression for the CEU individuals were primarily processed from 2003 to 2004, whereas the arrays used to measure expression for the ASN individuals were all processed in 2005–2006."

Follow up: pervasive interethnic difference or batch effect?

- 94% the genes associated with batch
- No differential expression associated with ethnicity after removal of batch effects
- Batch is a *confounder*: one cannot conclude whether ethnicity or batch explains Cheung's observations
- Further comparison of dates within the CEU samples shows that at least 79% of the genes are still associated with batch

Batch effects are often the *main* determinant of gene expression



Batch effects are often the *main* determinant of gene expression

- Randomize all experimental steps
- Beware of cross-study data merging
- Removing batch effects as the potential to remove key biological signal in the data or introduce more spurious signal
- […Still, Storey et al. (American J. Hum. Genet, 2007) report that 17% of the genes were differentially expressed between a group of African and Europeans]

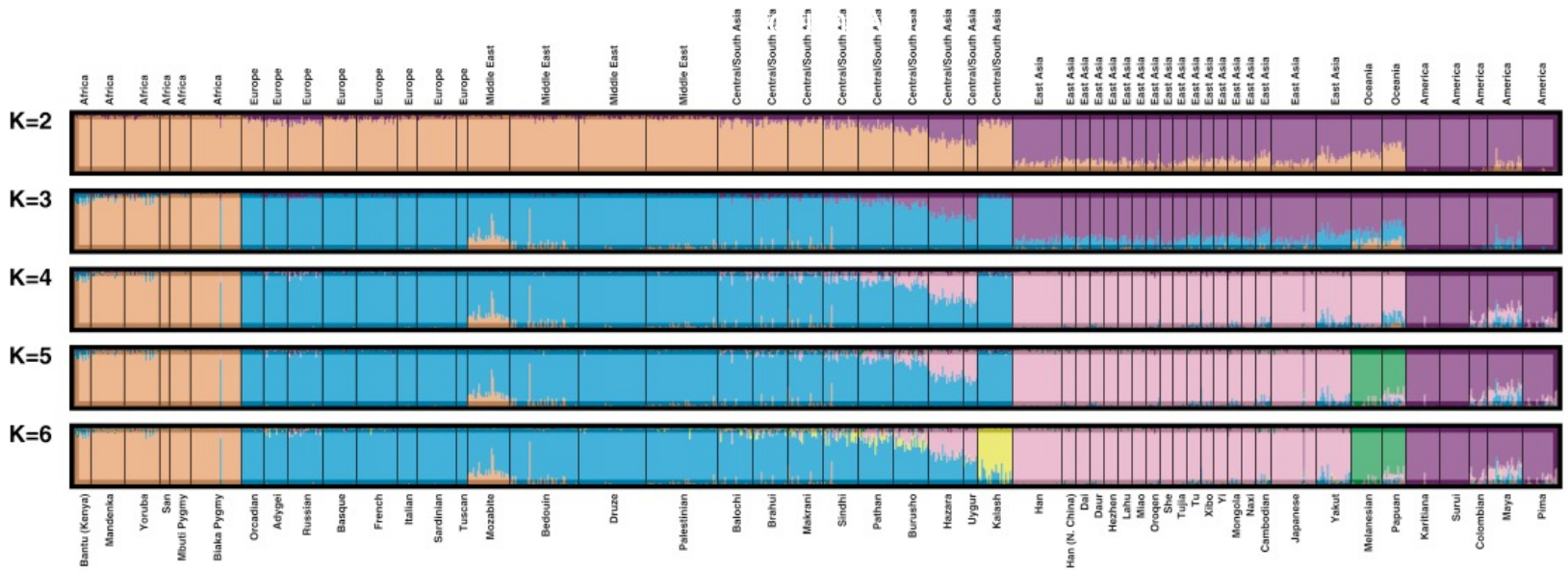
Probing the geographic heterogeinty of human populations

377 microsatellites genotyped
in 1056 individuals from
52 populations

k-mean clustering was applied to the data

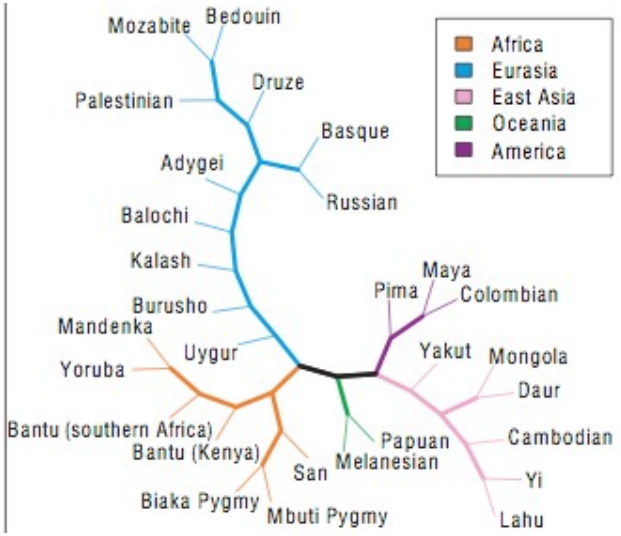
between-population accounts for 3-5% of total variance

3-5% is enough to accurately classify individuals according to their origin

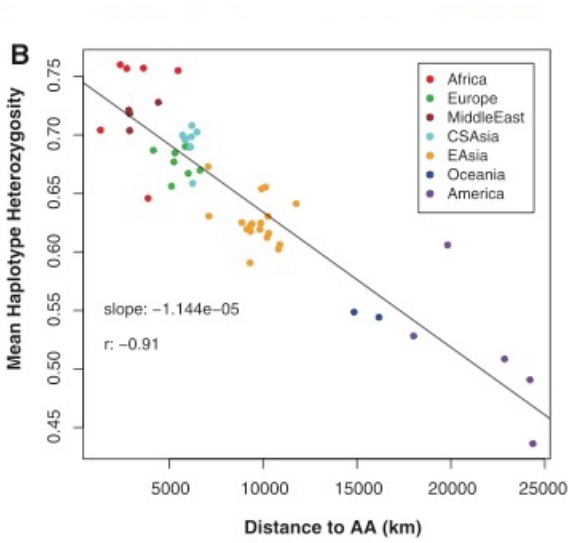


Rosenberg et al. Science 2002

Reconstructing with SNP arrays the history of human populations



Jakobsson et al., Nature 2008



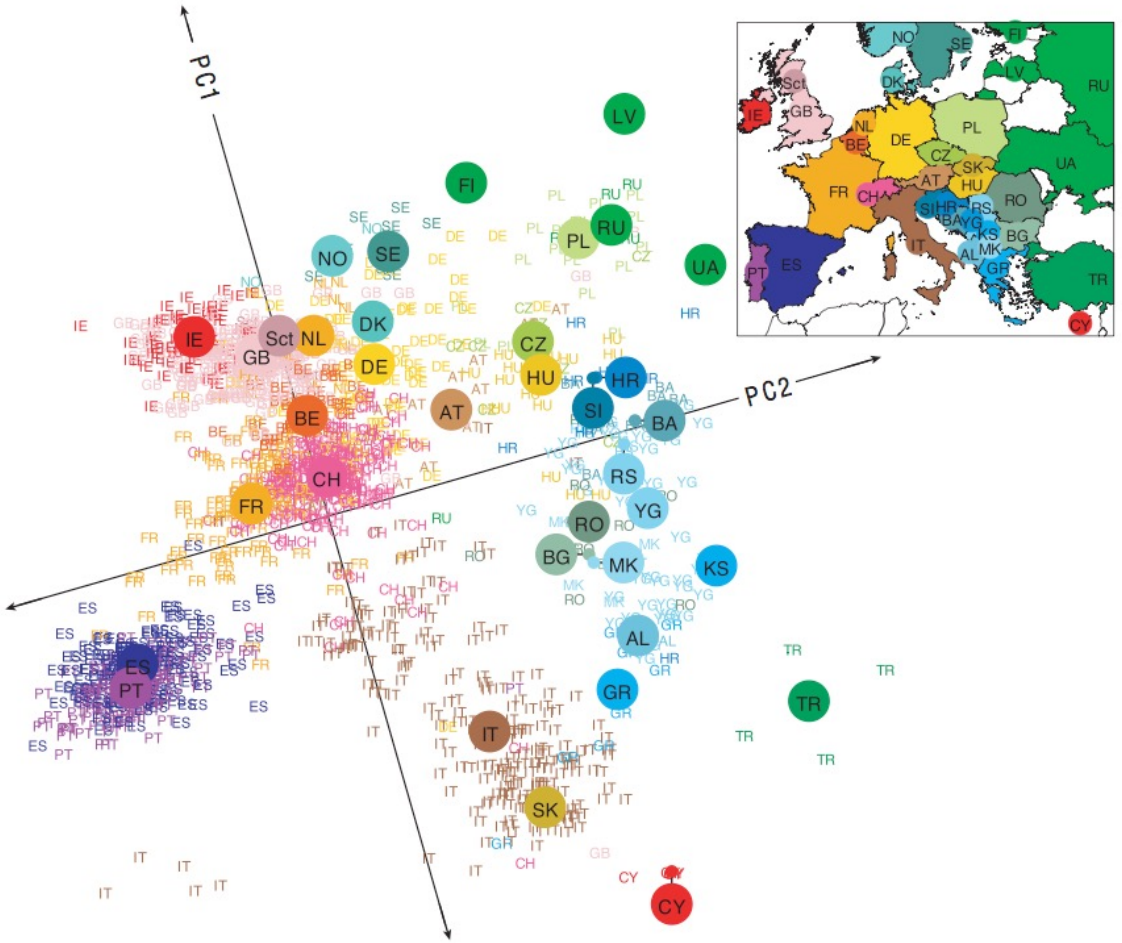
Li et al. Science 2008

Genes mirror geography within Europe

Novembre et al. Nature, 2008

- ~3200 Eurpeans were genotyped with 500K Affymetrix SNP arrays
- Individual with non European ancestry or with discordant grand-parental origins were removed
- Data were analysed with PCA

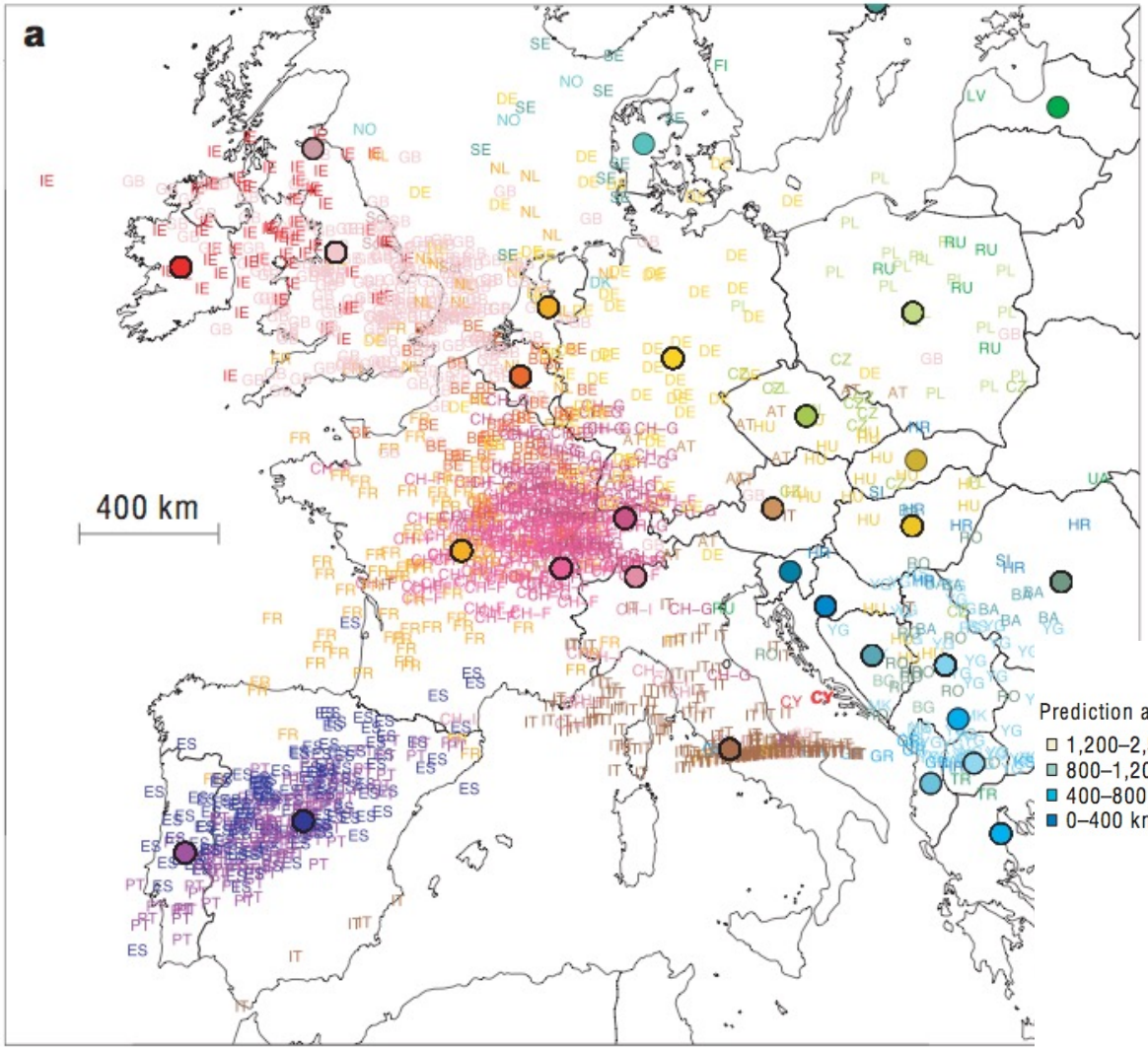
Genes mirror geography within Europe



- PC1 is aligned along a North-South axis
- PC2 is aligned along a East-West axis
- PC1 and PC2 account for 0.5% of total variance
- Thus, 99.5% of the variance is *not* related to geography

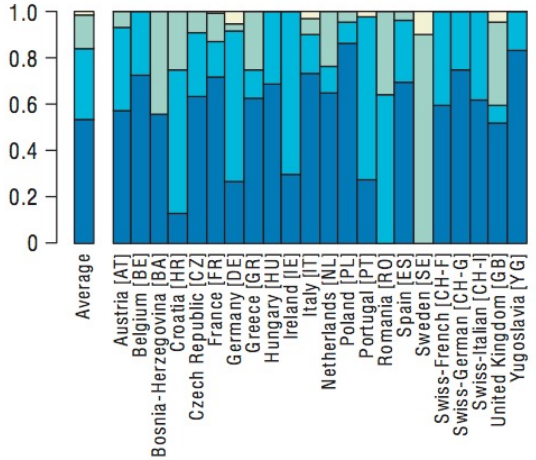
Novembre et al. Nature 2008

Genes mirror geography within Europe



Novembre et al. Nature 2008

- regional genetic differences are tiny, yet the origin of individuals can be inferred quite precisely from the genotype data alone



Discrimination always starts with a definition of the victims…

nature

www.nature.com/nature

Vol 461 | Issue no. 7265 | 8 October 2009

Genetics without borders

A UK government scheme to establish nationality through DNA testing is scientifically flawed, ethically dubious and potentially damaging to science.

So it was with understandable incredulity that researchers received a plan by the UK Border Agency to use genetics to determine nationality — specifically, the origin of asylum-seekers claiming to be from war-torn Somalia. The agency's pilot programme, which began last month, aims to determine whether some 100 individuals really are Somali nationals by checking them for the individual DNA variants known as single nucleotide polymorphisms (SNPs) in mitochondrial DNA, on the Y chromosome and elsewhere in the genome. The scheme will also use isotopic ratios of elements found in hair and fingernails — which can vary depending on a person's diet or environment — to try to establish where the migrants previously lived.

Strengths and limits of principal components analysis

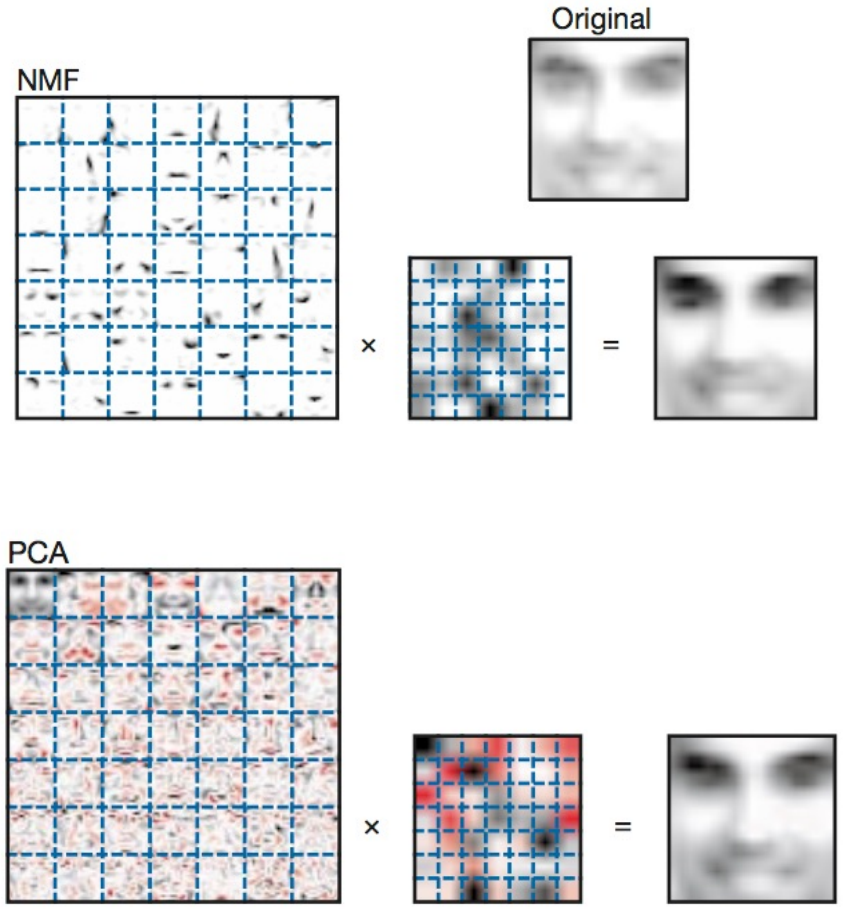
- ✓ The percentage of variance explained by the first components is a good quality measure.
- ✓ It produces continuous representations.
- ✓ Conceptual simplicity, and strong mathematical theory.

But:

- ✓ Components do not always have simple biological interpretations.
- ✓ No choice of distance measure.

Nonnegative matrix factorization (NMF)

NMF represents the data in term of nonnegative components



from Lee & Seung, Nature 1999

- The nonnegativity constrainst yields components representing *parts of objects* (here nose, eye, mouth...), which are interpretable
- NMF components are *not* orthogonal

$A \approx W \times H$

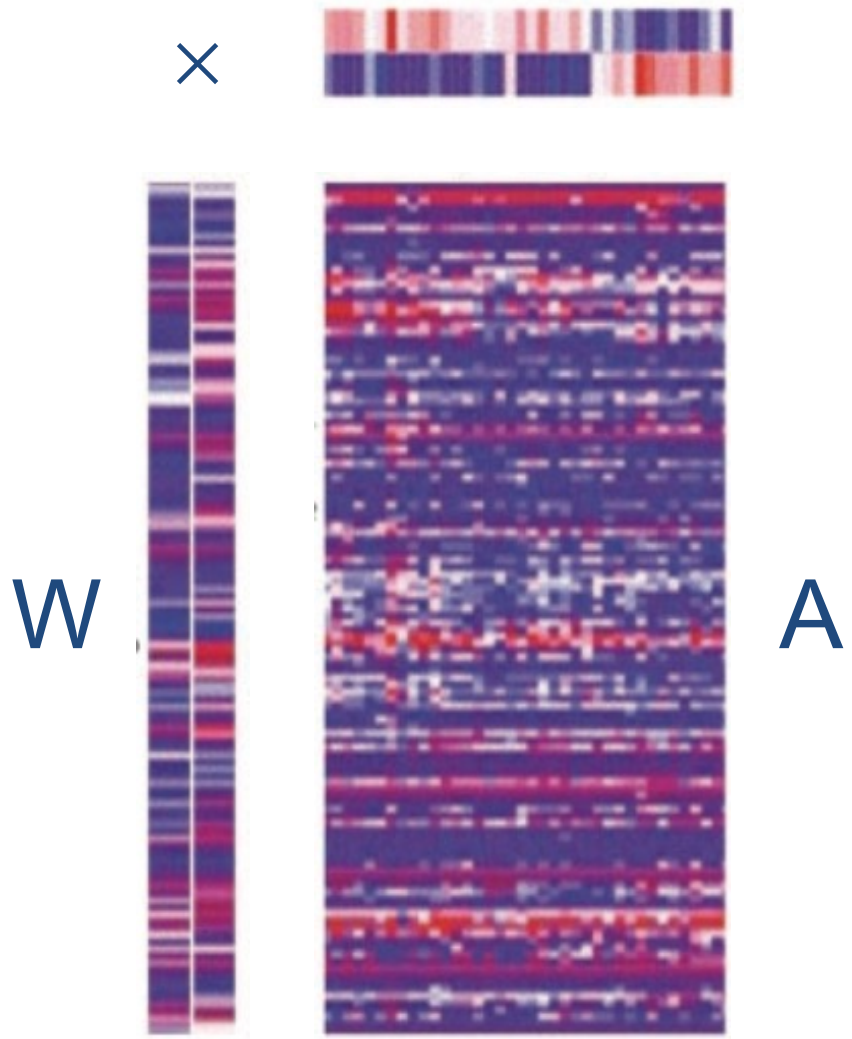
A: aproximate expression matrix

H: metagenes

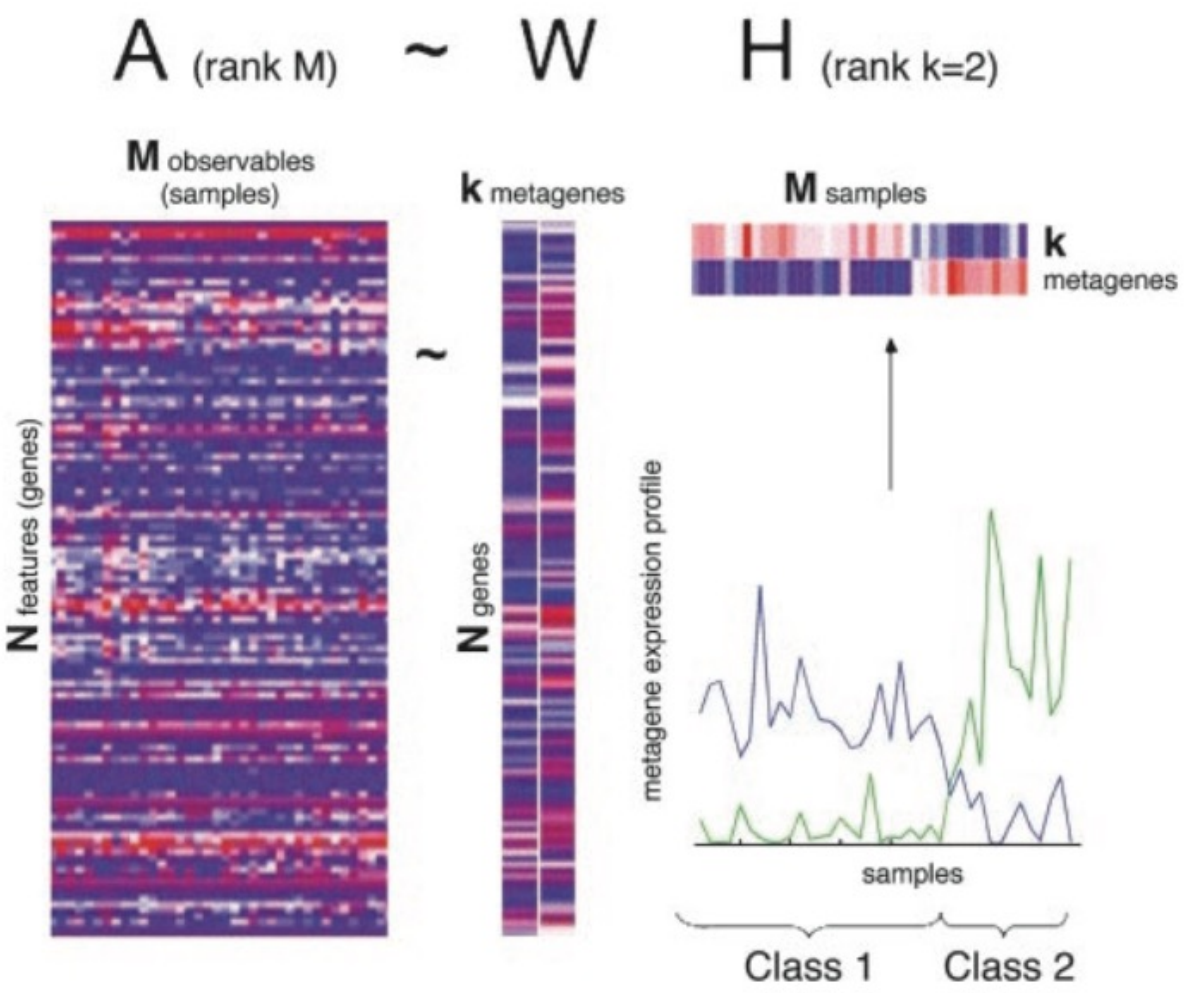
W: w_{ij} is the coefficient of gene i in metagene j

$w_{ij}>0$, $h_{ij}>0$

metagenes are *not* othogonal a priori



Metagenes split the samples in clusters, but do not force a hierarchy



k is an adjustable parameter

from Brunet et al., PNAS 2004

NMF is a stochastic algorithm

- NMF runs from different initial conditions may converges to different local mimina

- If they are k clearly defined clusters, one expect most rank k NMF runs to reveal metagenes mirroring these k clusters

Idea: use the consistency of NMF results across sets of random initial condition may be used as a measure of clustering quality

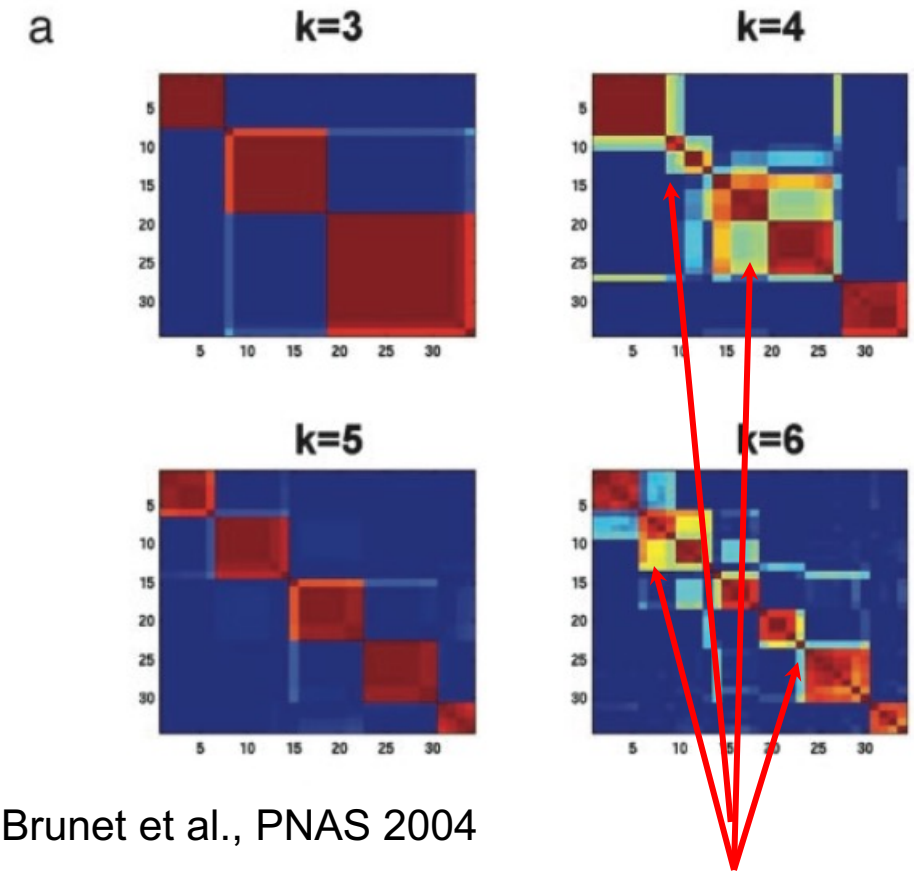
The consensus matrix represents the consistency of sample assignment to clusters

- For each run compute a *connectivity matrix* MxM matrix, C, with c_{ij}=1 if samples i and j are in the same cluster, c_{ij}=0 otherwise

- The average of connectivity matrices over many runs, |C| is called the *consensus matrix*

- One can reorder the rows and columns of |C| with average linkage hierarchical clustering

The consensus matrix represents the consistency of sample assignment to clusters



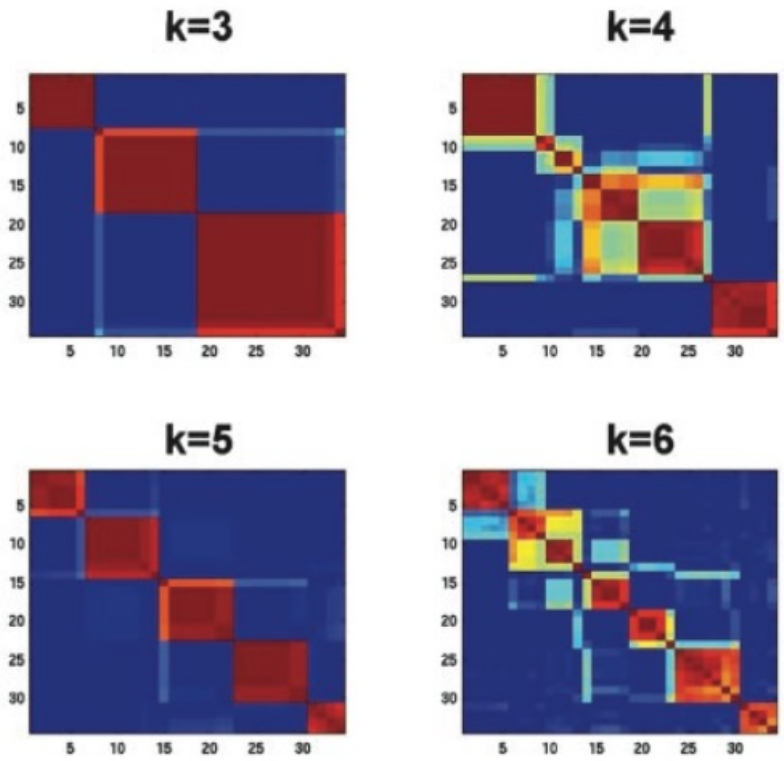
Brunet et al., PNAS 2004

- 25 classic and 9 desmoplastic medulloblastoma mRNA profiles (~6,000 genes)
- Consensus of 50 NMF runs are shown for k in {3,... 6}
- k=2 also yields clear-cut classes (not shown)

unstable assignments

The cophenetic correlation summarizes the dispersion of the consensus matrix

a

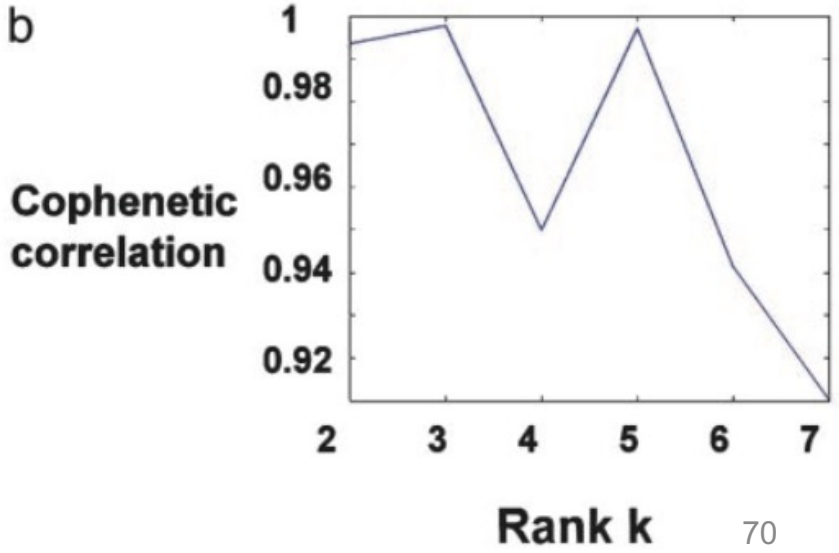


Brunet et al., PNAS 2004

$$c = \frac{\sum_{i<j}(x(i,j) - x)(t(i,j) - t)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(t(i,j) - t)^2]}}.$$

x(i, j) = distance between i and j
t(i, j) = distance between i and j in consensus matrix dendrogram
x is the average of x(i, j), t of t(i, j)

b

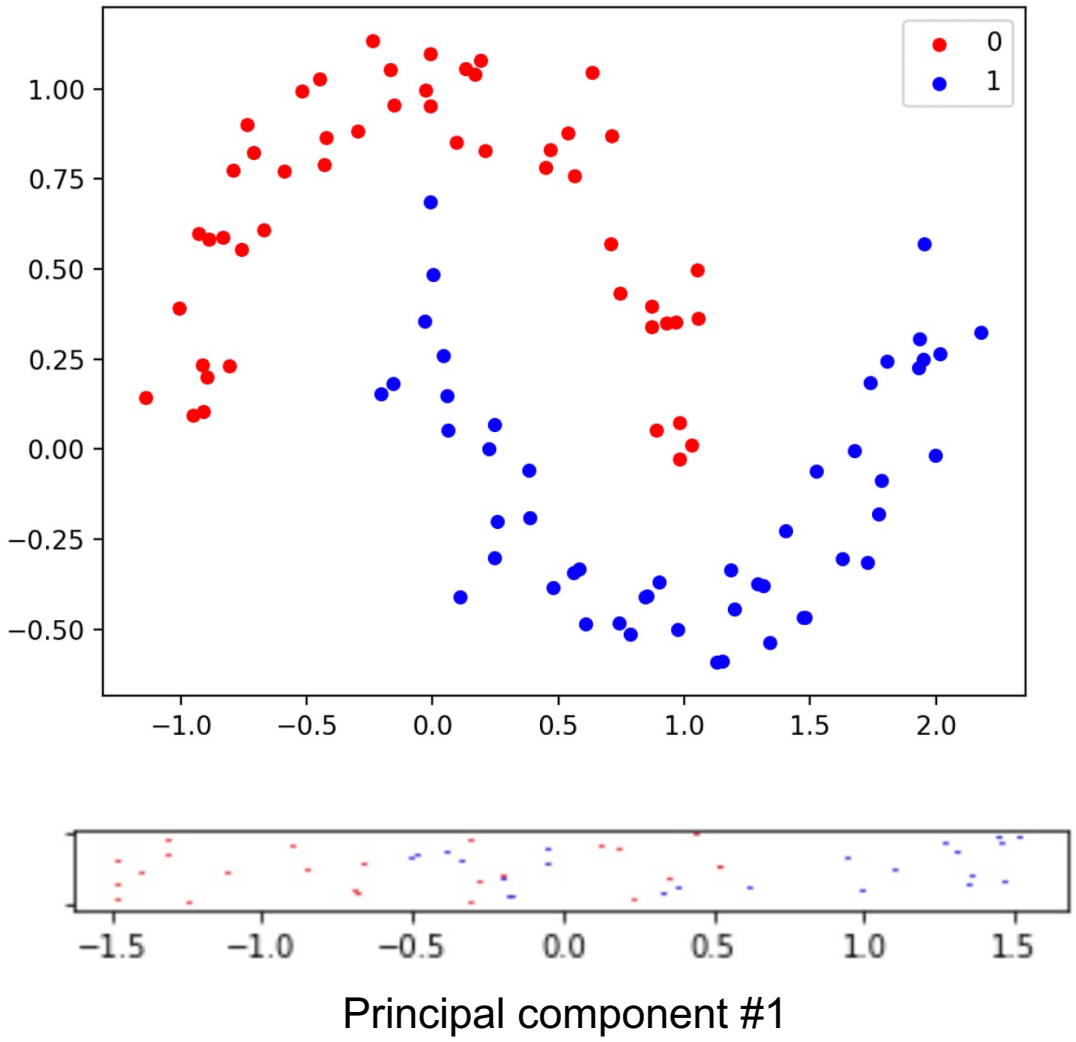


Other extentions of PCA have been proposed

- NMF uses nonnegative matrix
- ICA uses independent components
- kernel PCA maps data in higher dimension first (using SVM's "kernel trick")

Graph clustering

Non-linearly separable data



- Here, no straight lines can separate the blue class from the red one
- Thus, PCA and other linear dimension reduction methods cannot separate the two classes

Basic concept underlying graph-based clustering

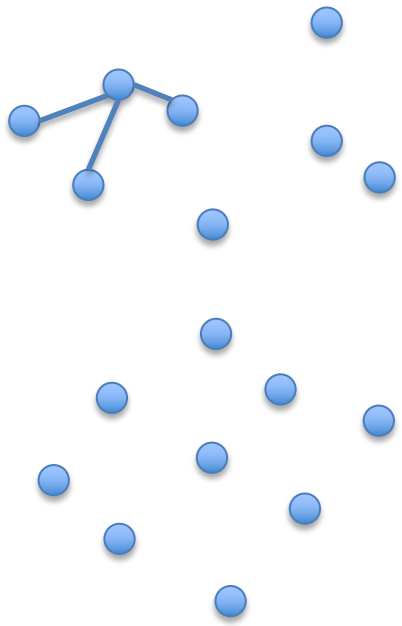
- Any dataset can be turned into a neighbours graph.
- A graph is not tied to a specific layout and, in fact, not even to a specific embedding space. Properties of a graph are independent on how it is displayed.
- Graphs can be (provided reasonable approximations) embedded in spaces of lower dimension.

Two steps of graph-based clustering

1. Turn the data into a neighbour graph, typically a k nearest neighbour graph

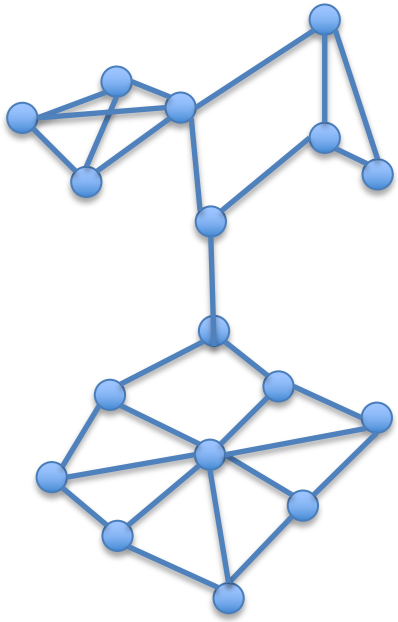
2. Layout the graph in a space of lower dimension, typically with a force directed layout algorithm

First, computes a k-nearest neighbors graph from the data



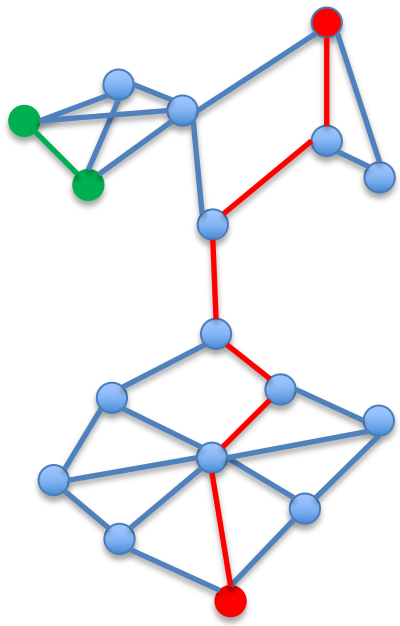
- A graph is a mathematical object made of nodes connected by edges
- A kNN graph is computed by connecting each point in the dataset to its k nearest neighbors
- Here, k = 3

First, computes a k-nearest neighbors graph from the data



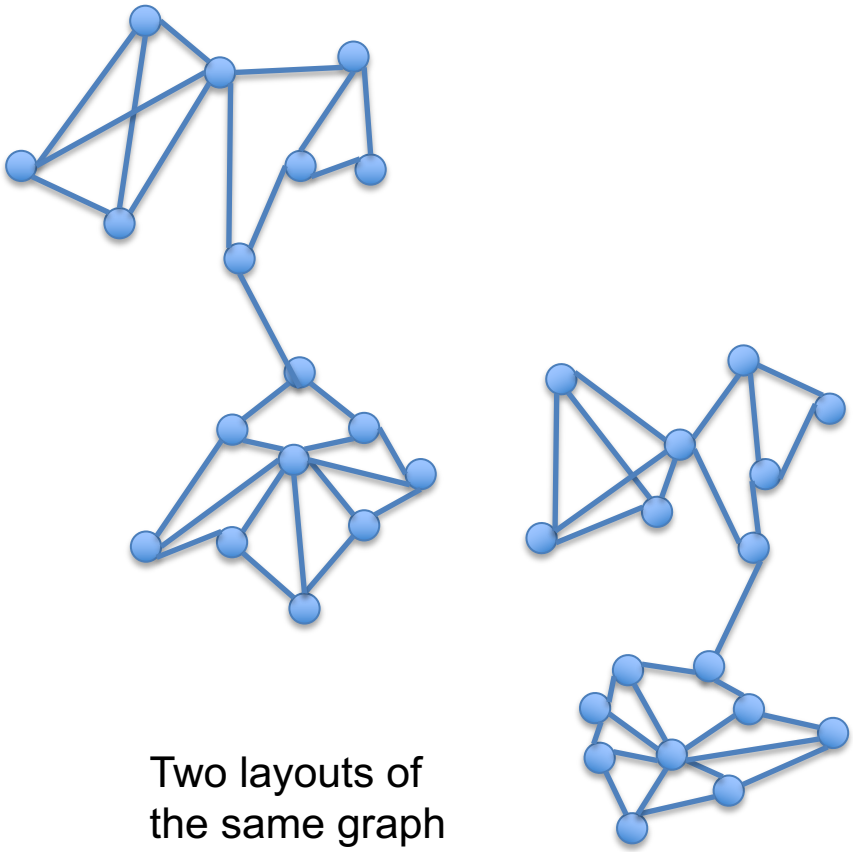
- A graph is a mathematical object made of nodes connected by edges
- A kNN graph is computed by connecting each point in the dataset to its k nearest neighbors
- Here, k = 3

First, computes a k-nearest neighbors graph from the data



- A kNN graph reflects local neighborhoods, yet it does contain information about the global structure of the data
- The distance between two points may be measured as the length of the shortest path that connect them

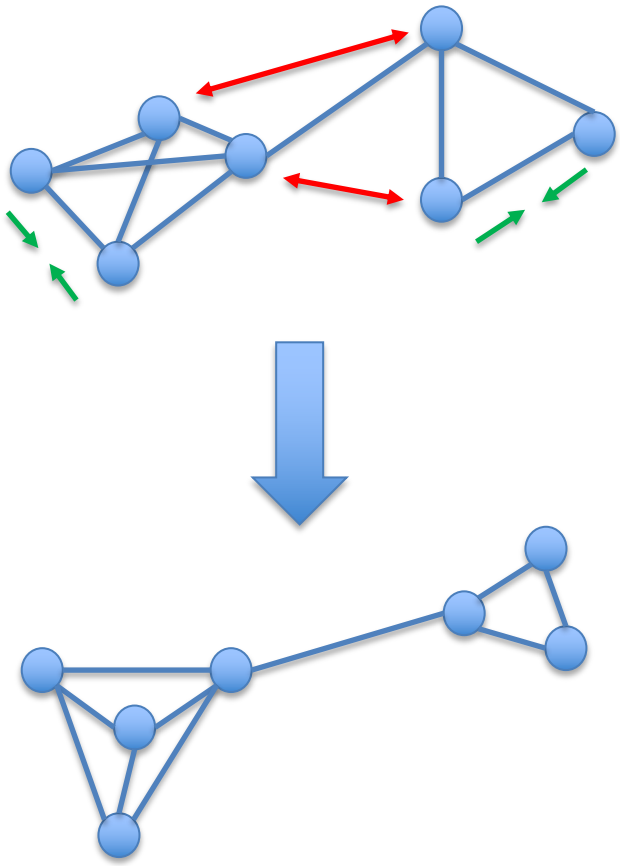
First, computes a k-nearest neighbors graph from the data



Two layouts of the same graph

- Although a kNN graph is derived of data sitting in a specific space, it graph is not tied to that space
- Many equivalent layouts exist of the same graph
- Layouts can be computed in spaces of lower dimensions than the original space

Second, compute graph layout

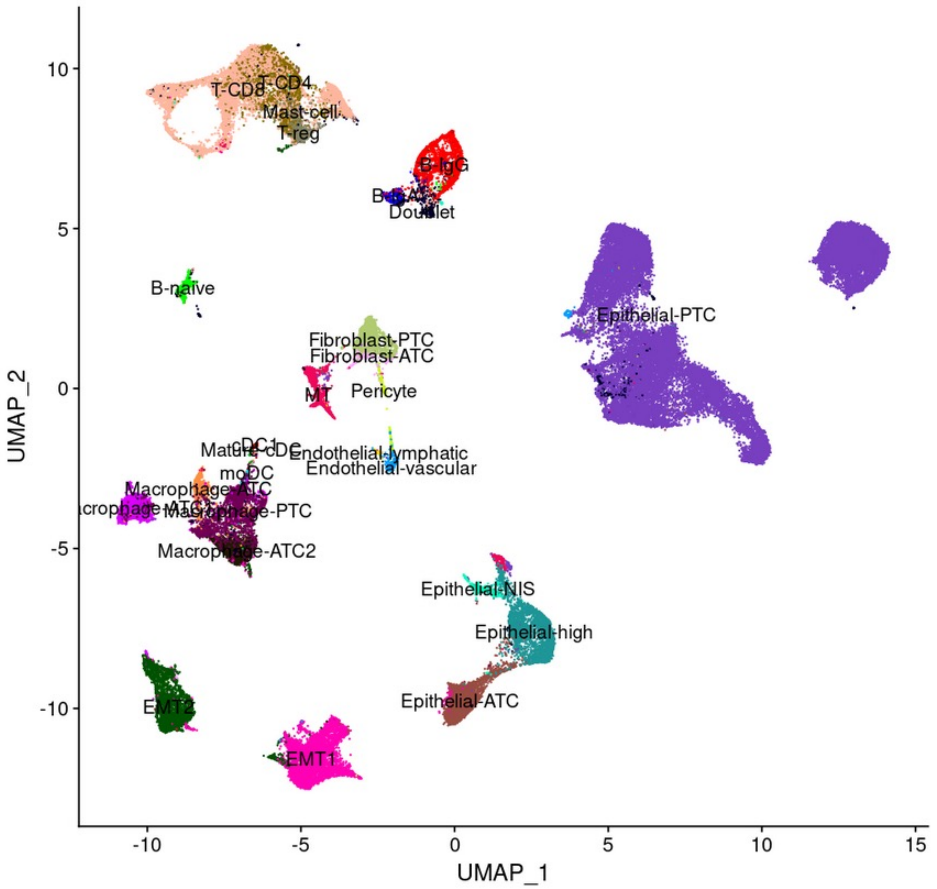


- Of course, you want to draw the graph in a low dimension space, 2 or 3D
- One approach is force directed layout:
  - Connected nodes attract each other
  - Disconnected nodes repel each other
  - 'Energy' of the layout is minimized

Different flavors of graph-based clustering

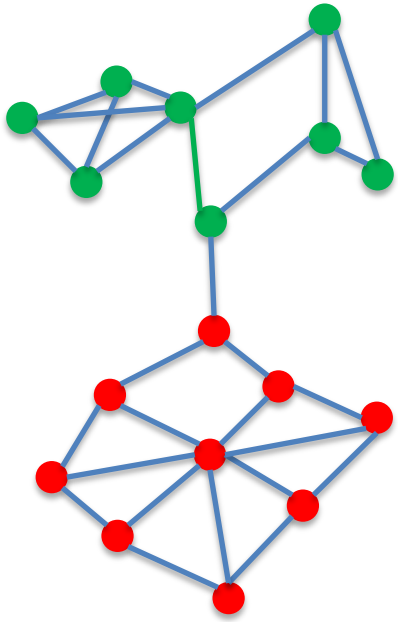
- There are many ways to build a neighbour graph (kNN is just one option)
- There are many ways to compute the layout
- Many methods have been proposed, including self organizing maps, t-SNE, UMAP, etc.

UMAP representation of 90k single cell transcriptmes



- UMAP is a popular methods in single cell transcriptomics
- Here, we see a 2D reduction of the transcriptomes of 90,000 cells from 14 thyroid cancers
- It was computed by first applying PCA to reduce the original gene expression space from 3000 dimensions to 30 dimensions, then reduce further to 2D with UMAP
- The 2D projection reveals cell types and cell states

Graph-based clustering

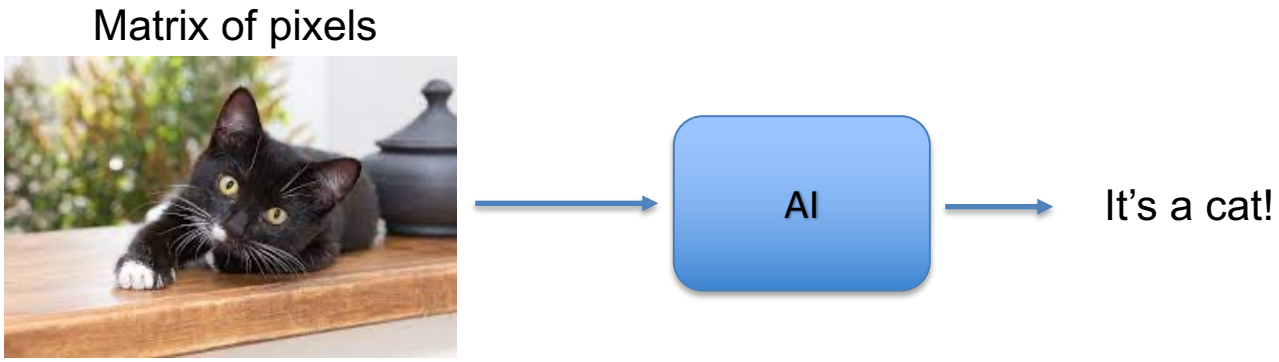


- A kNN graph connectivity reflects point density in the original embedding space
- This enables clustering via ‘community’ (a concept inspired from social network analysis) detection algorithms, such as the Lieden algorithm

Brief introduction to self supervised deep learning

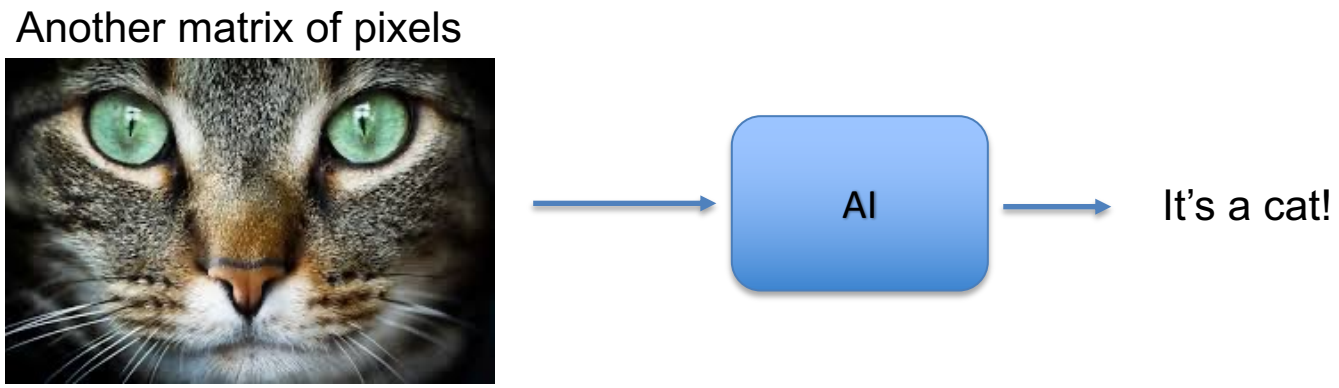
Artificial intelligence extracts semantic information from data

- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:



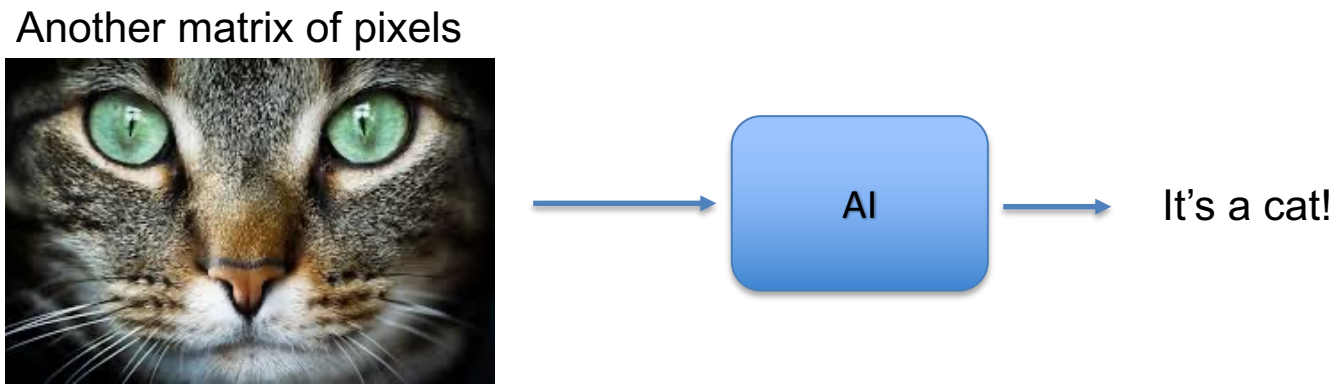
Artificial intelligence may set morphology on a quantitative basis

- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:



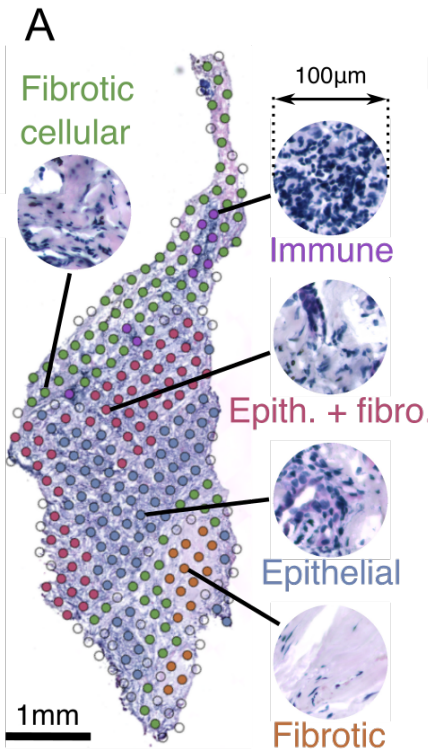
Artificial intelligence may set morphology on a quantitative basis

- Photos is a matrices of pixels reflecting the activation of an optical sensor
- AI extract high-level **semantic information** from raw images:

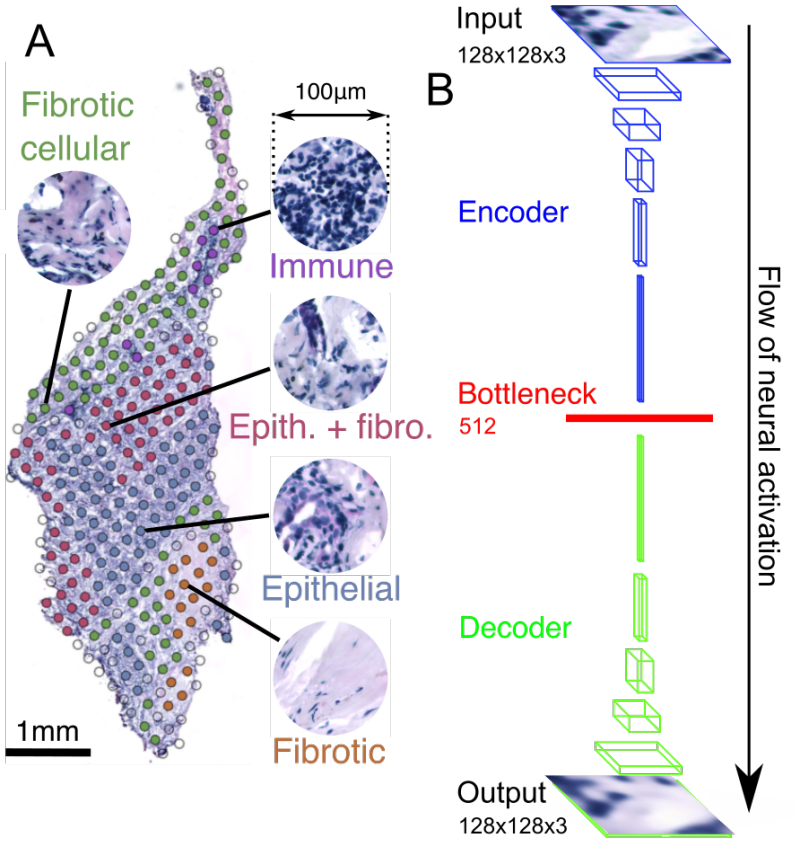


- Intermediate representations of images computed by inner layers of deep neural network can be exploited to turn images into numerical semantic representations and group them by (semantic) similarity.

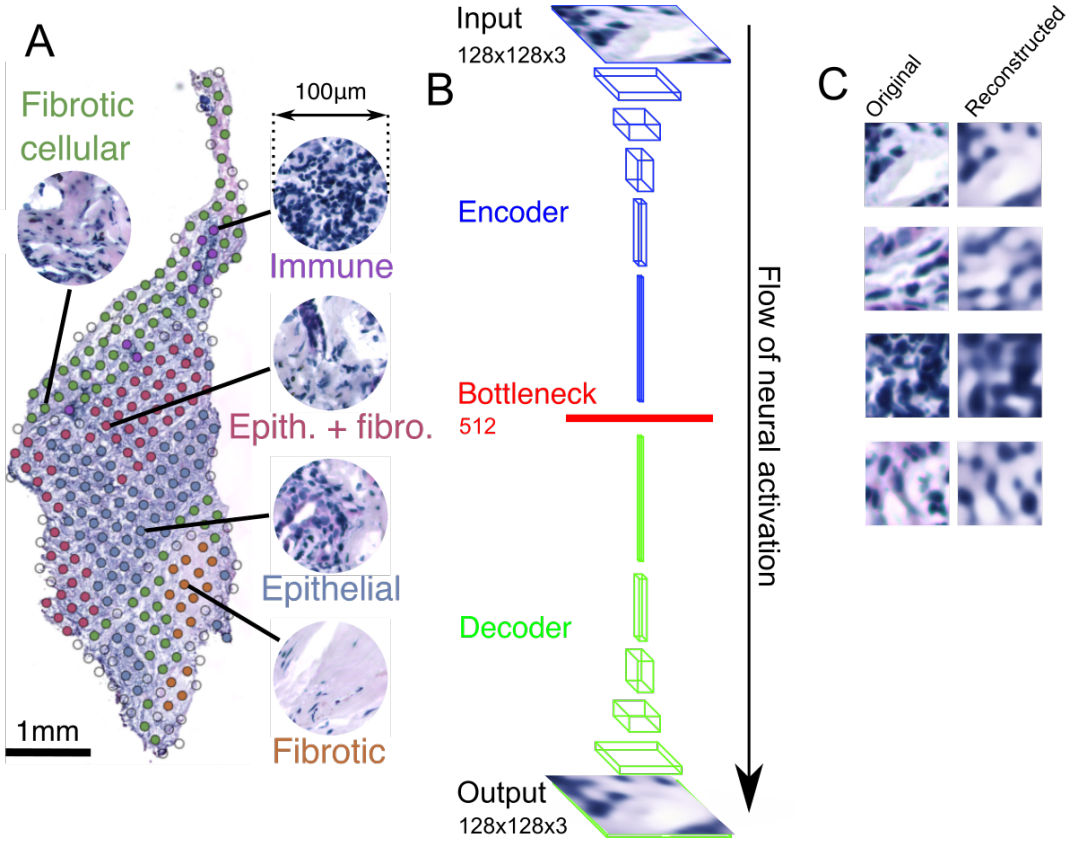
An example of self supervised AI: autoencoders



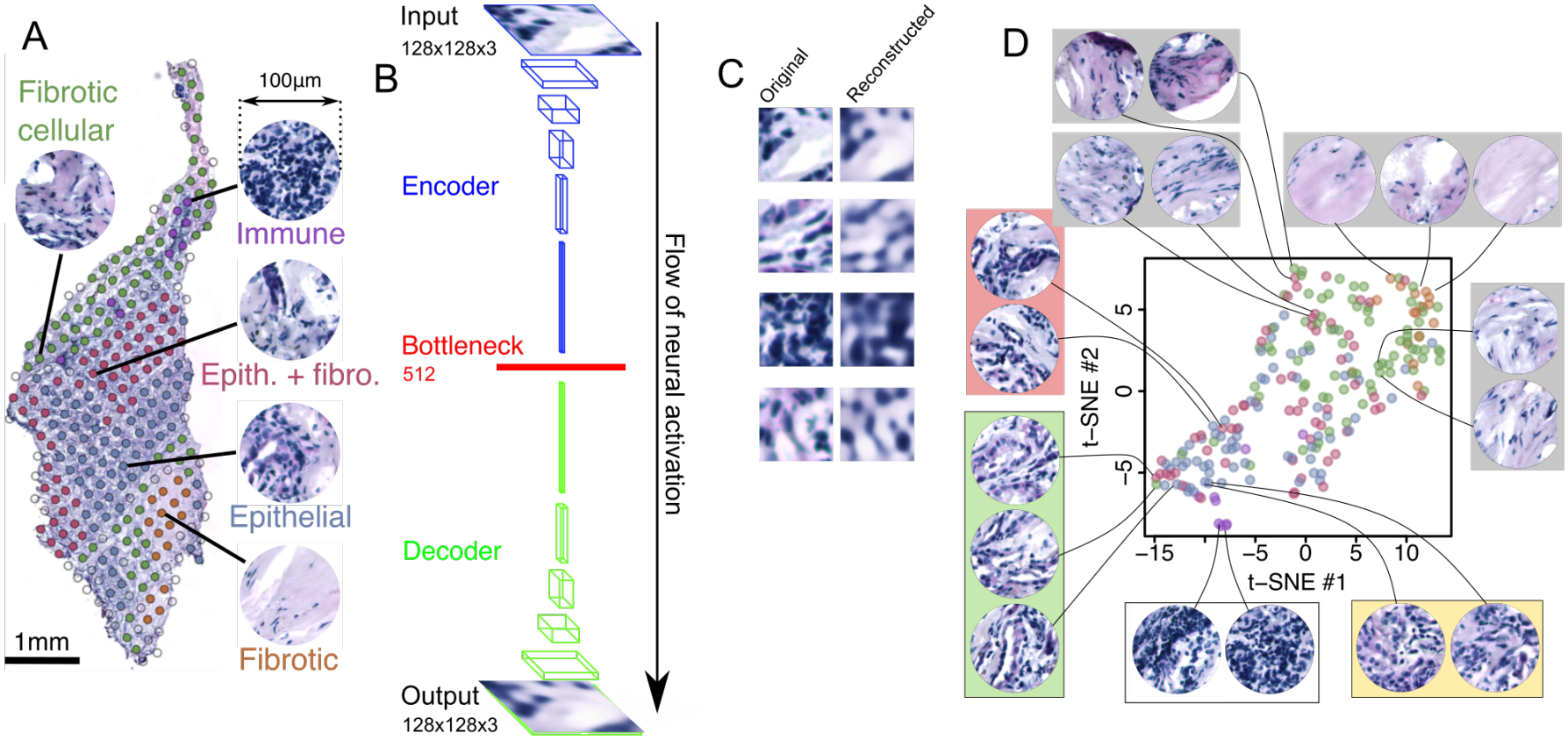
An example of self supervised AI: autoencoders



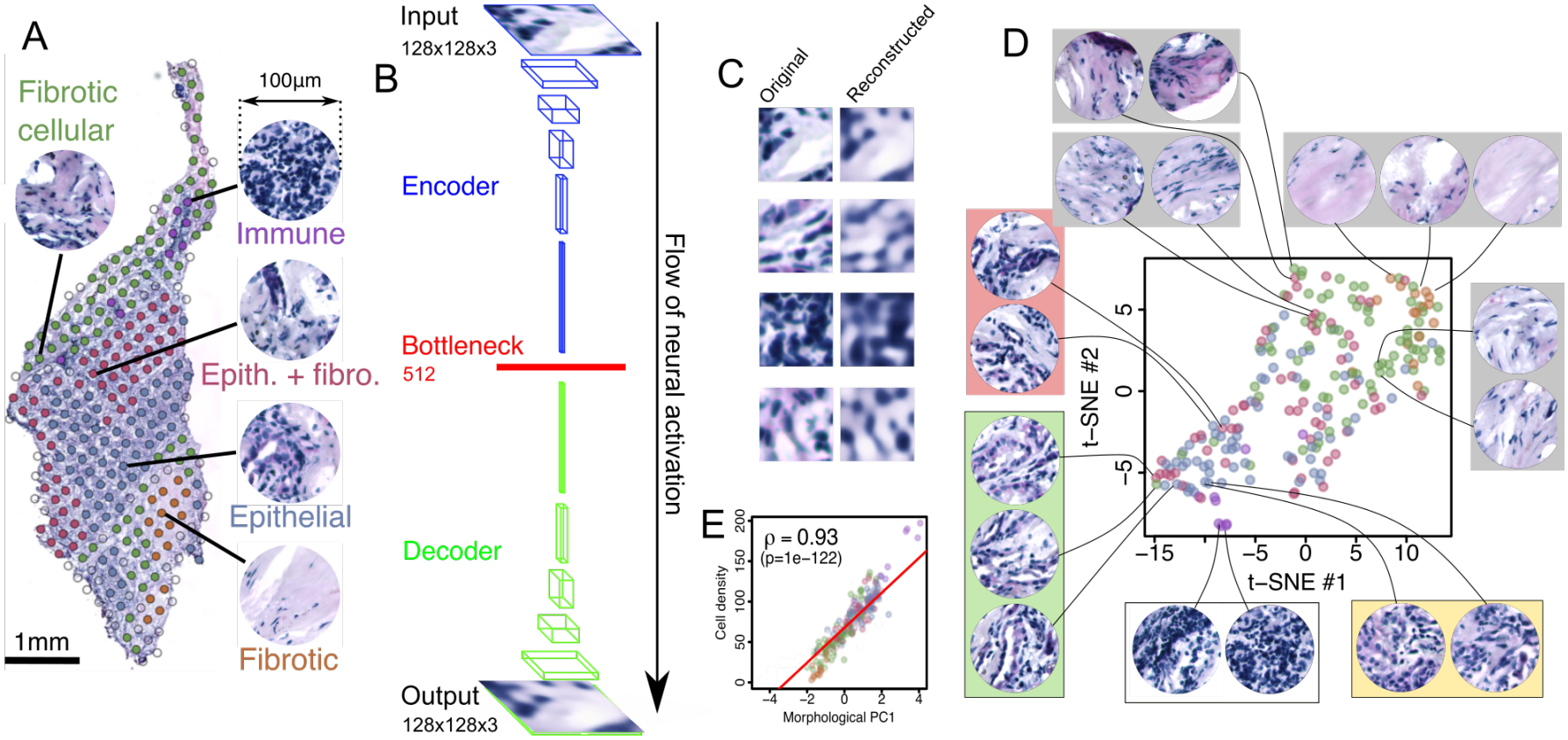
An example of self supervised AI: autoencoders



An example of self supervised AI: autoencoders



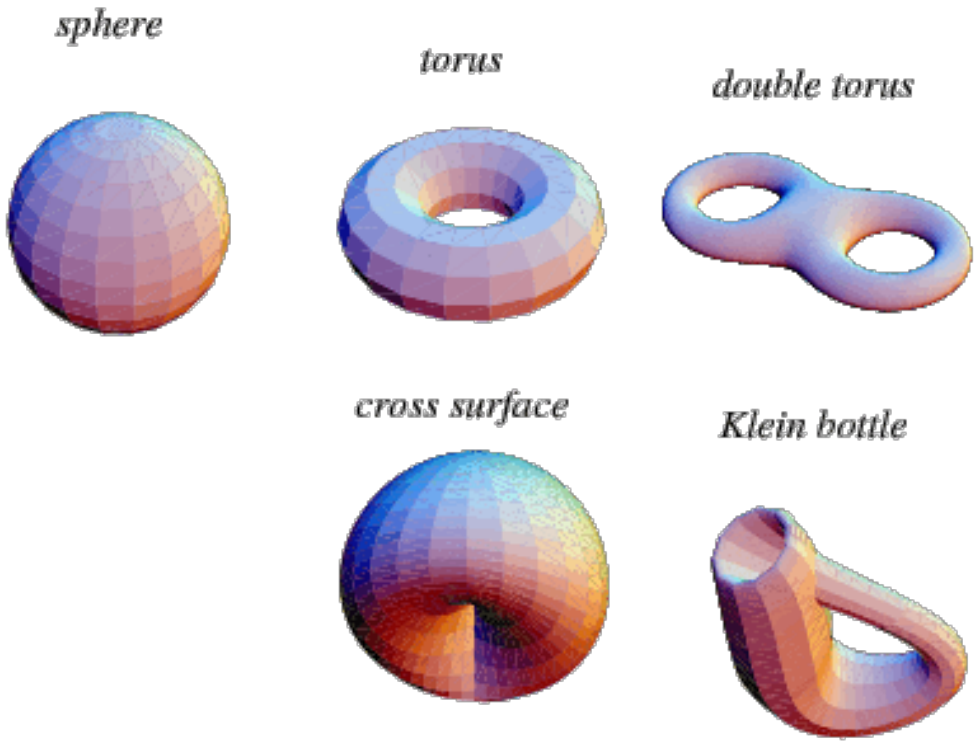
An example of self supervised AI: autoencoders



Elements of vocabulary

- The autoencoder's bottleneck layer associates a numerical vector to any image provided as input
- This vector is a numerical ***representation*** of the input image. This term, representation, is inspired by the mental representation of concepts dealt with in cognitive sciences
- The encoder defines a ***latent space***, i.e. a mathematical space where the representation of all the relevant images are found.
- The latent space is also called an ***embedding***, because it is embedded in a space of much higher dimension (the image space).
- Embedding generated by AI are though to be ***manifolds***, i.e. a topological space that locally resembles Euclidean space near each point

Examples of 2D manifolds embedded in the 3D space



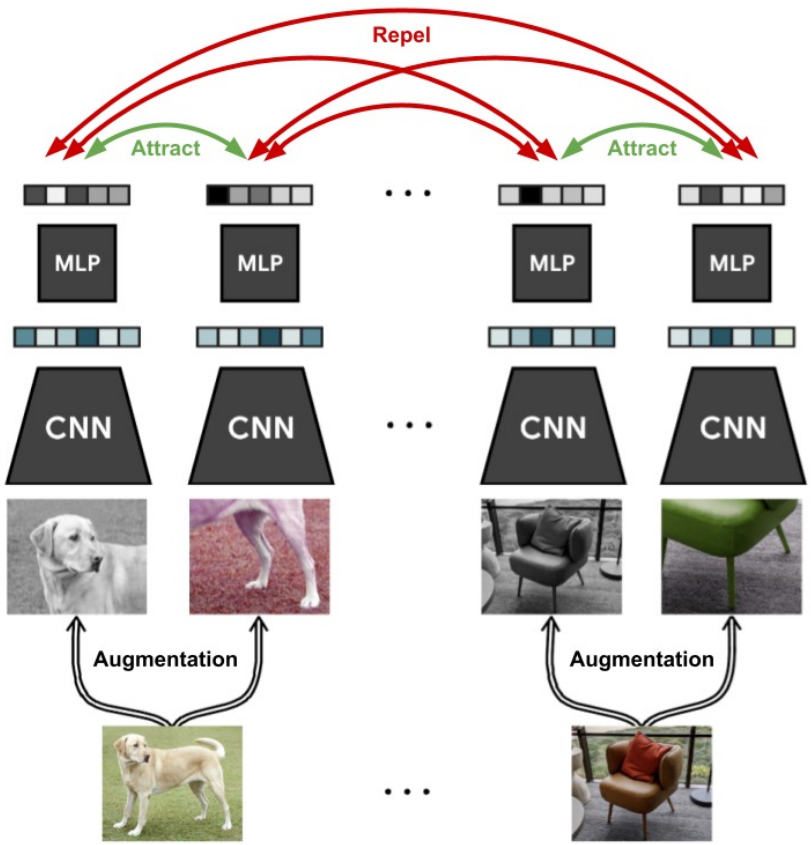
Contrastive learning

- The reconstruction objective of the autoencoder is an overkill: the exact positions of cells is morphologically irrelevant

- It is even detrimental: we need latent representations that are insensitive to technical color variations, to the orientation of the slide under the microscope, etc.

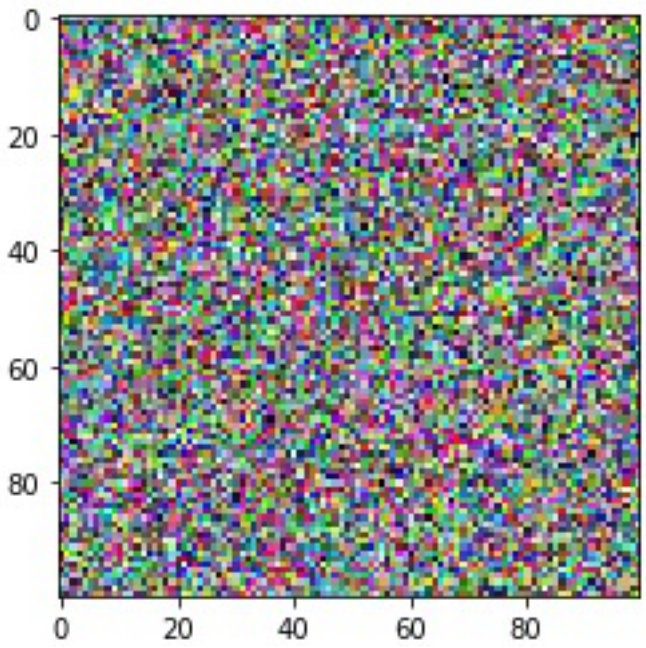
- Contrastive learning does not use a reconstruction objective and it makes it possible to train AIs invariant to specific image features

Contrastive learning



- We used the smCLR framework (Chen *et al.*, arxiv 2020)
- The AI is presented pairs of transformed images
- It is trained to guess if the two images are transformations of the same images or not
- The transformations define the invariance
- It this example, zoom/resize and color shift transformation makes the latent representations invariant to zoom level and color
- Learning is completely self-supervised because the transformations can be computed automatically

The core idea of AI



- Images sit in a very high dimensional space (here, 100x100x3 variables)
- The vast majority of pixels configurations do not exist in the real world
- Images of the real world are embedded in a small part of the image space, of muck lower dimension than the original space of the data
- This is even more true of image in sub-domains, like cats, dogs, human faces, etc.
- AI algorithms are statistical machines that learn these embeddings from data

Once the manifold underlying the data is known…

One can:

- Build classifiers by using the AI as feature encoder that is fed into regular supervised classification algorithms
- Decode point on the manifold to *generate* data in the original data space
- Denoise data by projecting them on the manifold
- Learn mappings between manifolds build from different data acquisition modalities (e.g. video+sound, transcriptome+histology, etc.)
- Etc.