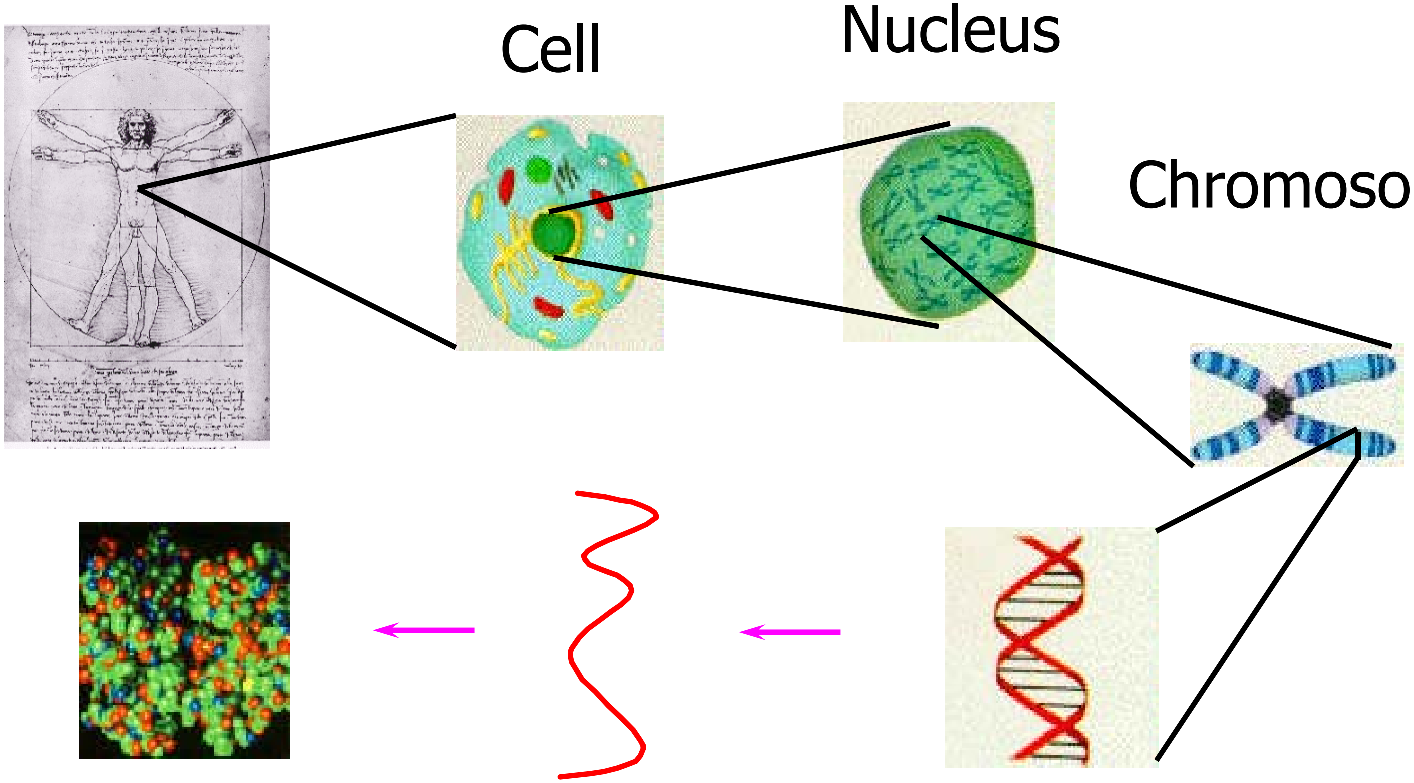
Microarrays

- Technology behind microarrays
- Data analysis approaches
- Clustering microarray data

Molecular biology overview



Cell

Nucleus

Chromosome

Protein

Gene (mRNA), single strand

Gene (DNA)

Graphics courtesy of the National Human Genome Research Institute

Gene expression

- Cells are different because of **differential gene expression.**
- About 40% of human genes are expressed at any one time.
- Gene is expressed by **transcribing** DNA into single-stranded mRNA
- mRNA is later **translated** into a protein
- Microarrays measure the level of mRNA expression

Basic idea

- mRNA expression represents dynamic aspects of cell
- mRNA expression can be measured with latest technology
- mRNA is isolated and labeled using a fluorescent material
- mRNA is hybridized to the target; level of hybridization corresponds to light emission which is measured with a laser
- Higher concentration → more hybridization → more mRNA

A demonstration

- DNA microarray animation by A. Malcolm Campbell.

http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

Experimental conditions

- Different tissues
- Different developmental stages
- Different disease states
- Different treatments

Background papers

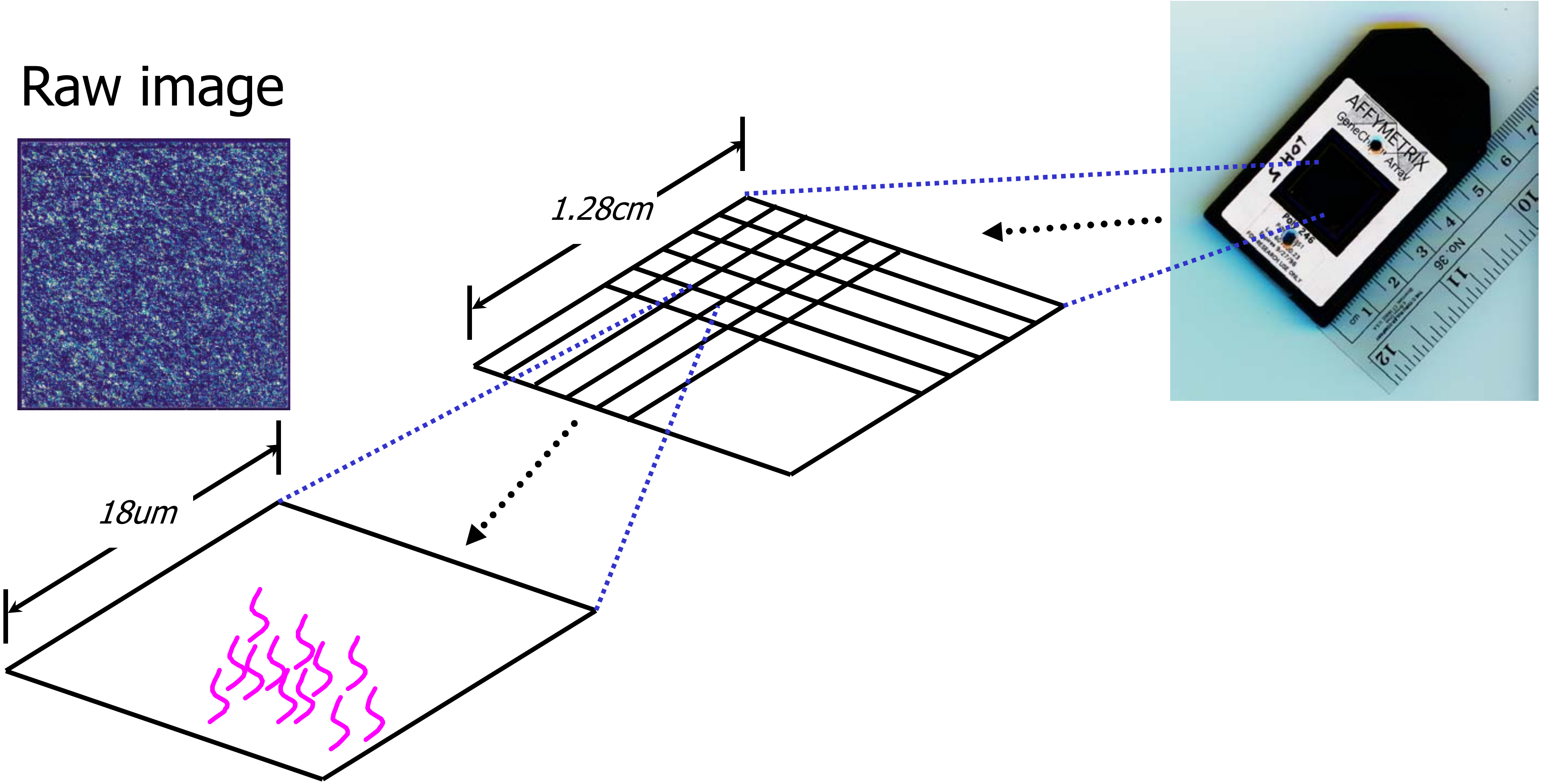
- Background paper 1
- Background paper 2
- Background paper 3

Microarray types

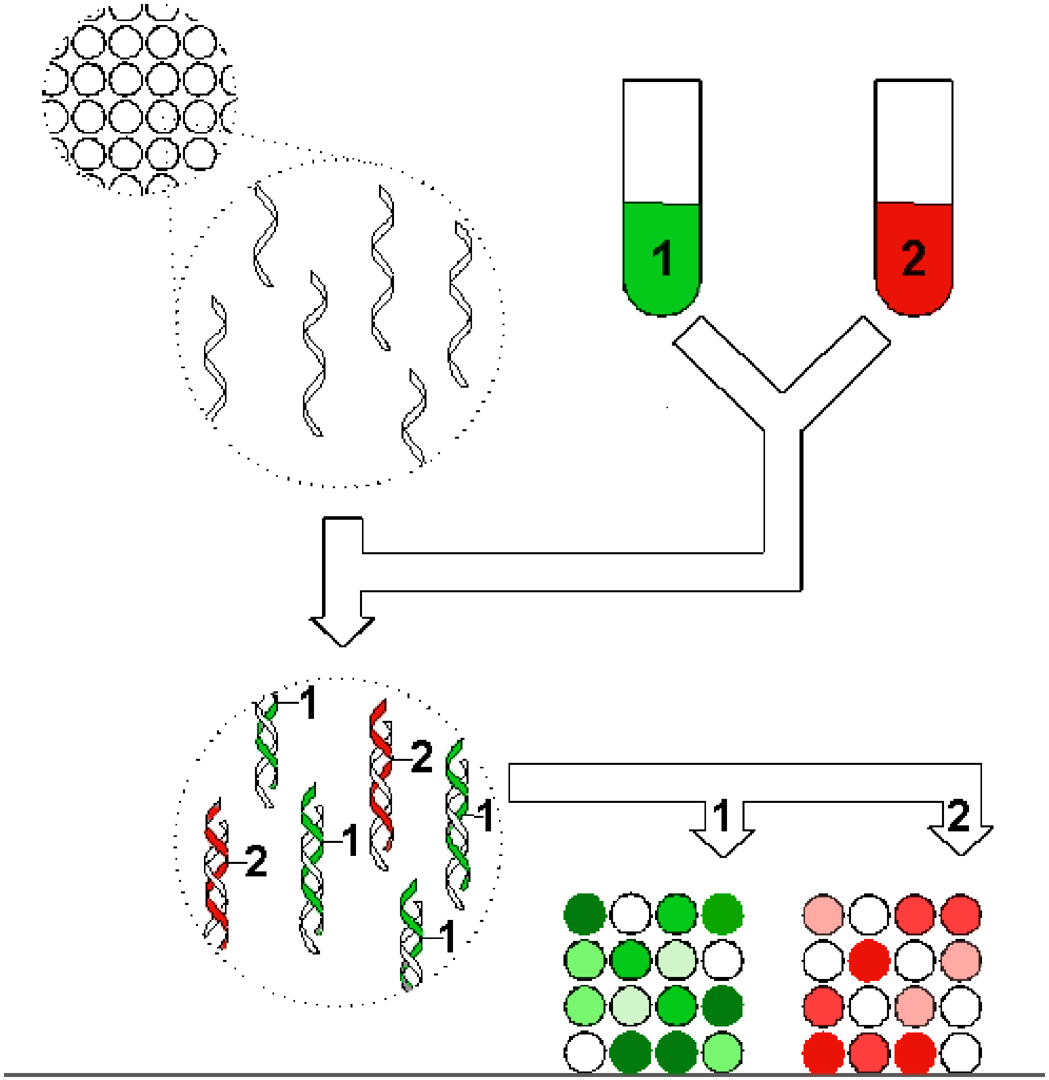
The main types of gene expression microarrays:

- Short oligonucleotide arrays (Affymetrix)
- cDNA or spotted arrays (Brown lab)
- Long oligonucleotide arrays (Agilent Inkjet)
- Fiber-optic arrays
- ...

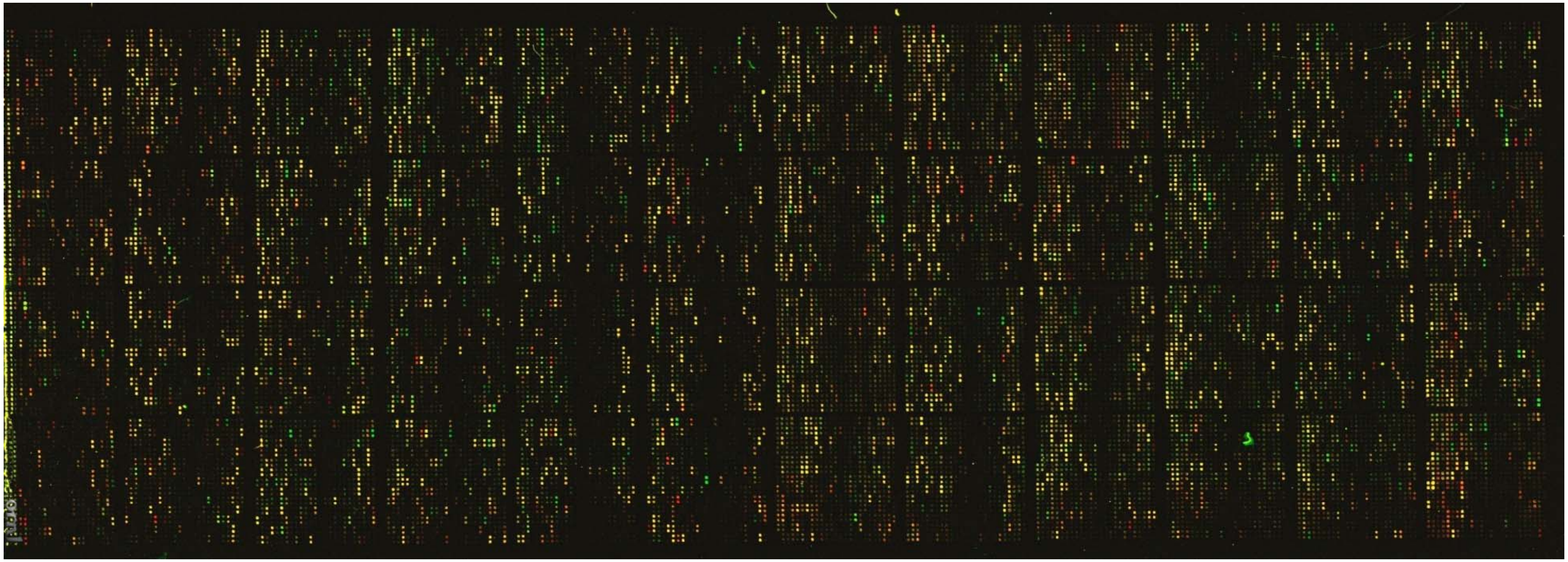
Affymetrix chips



Competitive hybridization

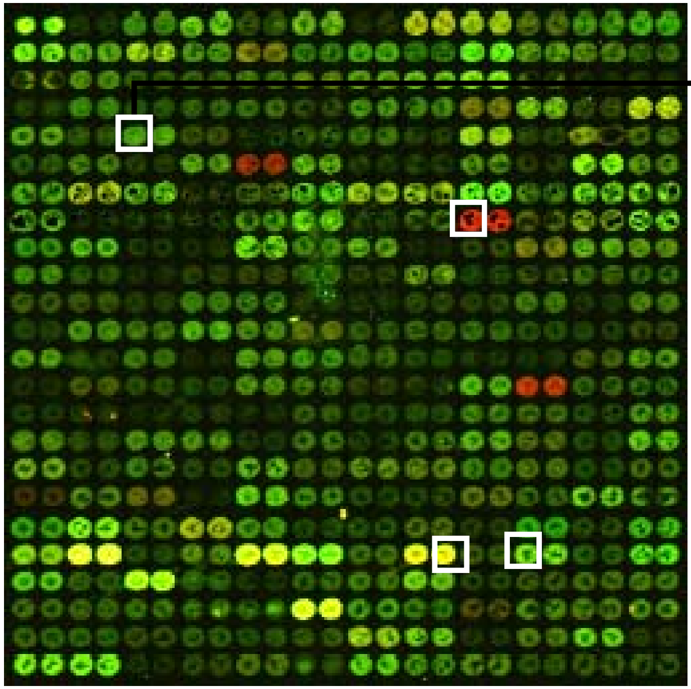


Microarray image data



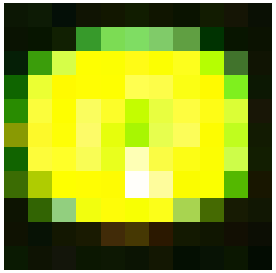
mouse heart versus liver hybridization

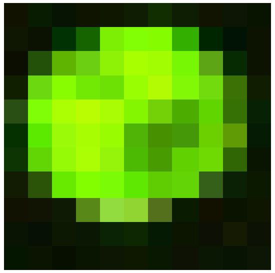
More images



Gene GTF4

Upregulated





Downregulated

🟩 Reference cDNA

🟥 Experimental cDNA

Characteristics of microarray data

- Extremely high dimensionality
  - Experiment = (gene$_1$, gene$_2$, …, gene$_N$)
  - Gene = (experiment$_1$, experiment$_2$, …, experiment$_M$)
  - N is often on the order of 10^4
  - M is often on the order of 10^1
- Noisy data
  - Normalization and thresholding are important
- Missing data
  - For some experiments a given gene may have failed to hybridize

Microarray data

GENE_NAME	alpha 0	alpha 7	alpha 14	alpha 21	alpha 28	alpha 35	alpha 42
YBR166C	0.33	-0.17	0.04	-0.07	-0.09	-0.12	-0.03
YOR357C	-0.64	-0.38	-0.32	-0.29	-0.22	-0.01	-0.32
YLR292C	-0.23	0.19	-0.36	0.14	-0.4	0.16	-0.09
YGL112C	-0.69	-0.89	-0.74	-0.56	-0.64	-0.18	-0.42
YIL118W	0.04	0.01	-0.81		-0.3	0.49	0.08
YDL120W	0.11	0.32	0.03	0.32	0.03	-0.12	0.01

Missing Value!

Data mining challenges

- Too few experiments (samples), usually < 100
- Too many columns (genes), usually > 1,000
- Too many columns lead to false positives
- For exploration, a large set of all relevant genes is desired
- For diagnostics or identification of therapeutic targets, the smallest set of genes is needed
- Model needs to be explainable to biologists

Data processing

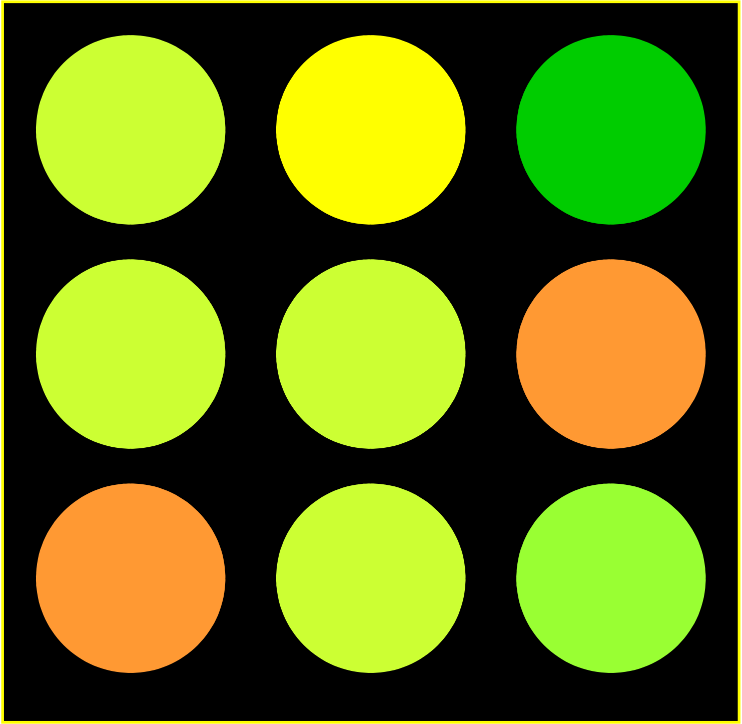
- Gridding
  - Identifying spot locations
- Segmentation
  - Identifying foreground and background
- Removal of outliers
- Absolute measurements
  - cDNA microarray
    - Intensity level of red and green channels

Data normalization

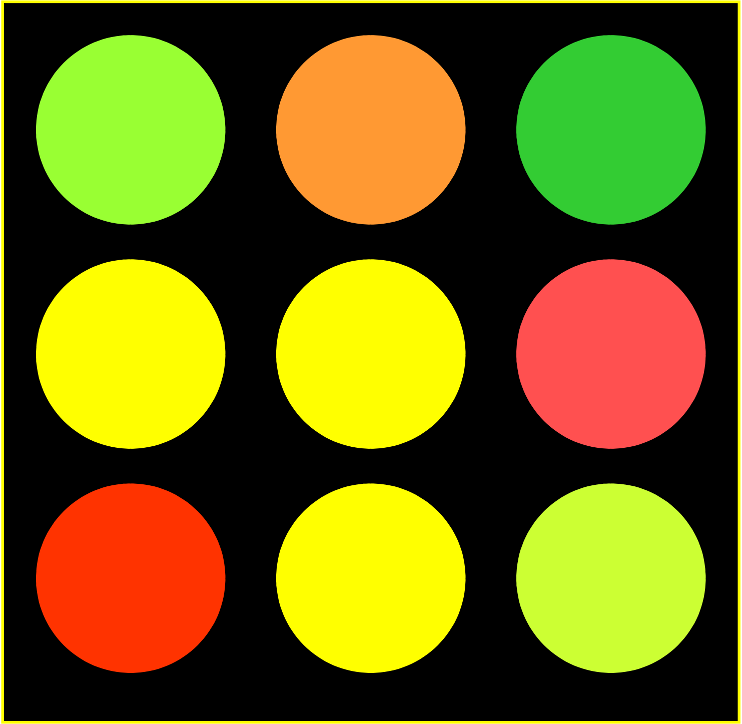
- Normalize data to correct for variances
  - Dye bias
  - Location bias
  - Intensity bias
  - Pin bias
  - Slide bias
- Control vs. non-control spots
  - Maintenance genes

Data normalization

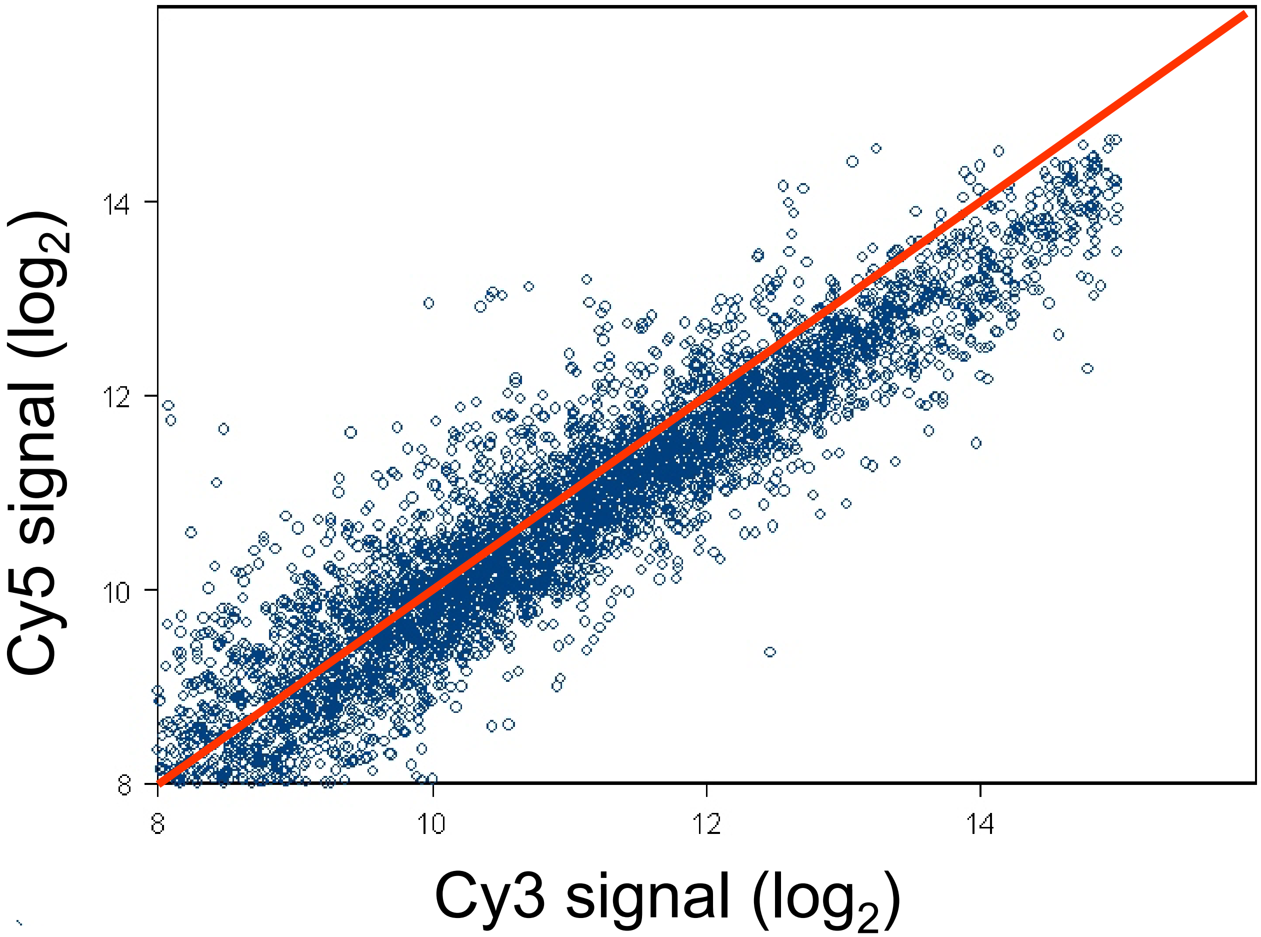
Uncalibrated, red light under detected



Calibrated, red and green equally detected



Normalization



Data analysis

- What kinds of questions do we want to ask?
  - Clustering
    - What genes have similar function?
    - Can we subdivide experiments or genes into meaningful classes?
  - Classification
    - Can we correctly classify an unknown experiment or gene into a known class?
    - Can we make better treatment decisions for a cancer patient based on gene expression profile?

Clustering goals

- Find natural classes in the data
- Identify new classes / gene correlations
- Refine existing taxonomies
- Support biological analysis / discovery
- Different Methods
  - Hierarchical clustering, SOM's, k-means, etc

Clustering techniques

- Distance measures
  - Euclidean: $\sqrt{}\ \Sigma\ (x_i - y_i)^2$
  - Vector angle: cosine of angle = $x.y\ /\ \sqrt{}\ (x.x)\ \sqrt{}\ (y.y)$
  - Pearson correlation
    - Subtract mean values and then compute vector angle
    - $(x-\underline{x}).(y-\underline{y})\ /\ \sqrt{}\ ((x-\underline{x}).(x-\underline{x}))\ \sqrt{}\ ((y-\underline{y}).(y-\underline{y}))$
    - Pearson correlation treats the vectors as if they were the same (unit) length, therefore it is insensitive to the amplitude of changes that may be seen in the expression profiles.

K-means clustering

- Randomly assign k points to k clusters
- Iterate
  - Assign each point to its nearest cluster (use centroid of clusters to compute distance)
  - After all points are assigned to clusters, compute new centroids of the clusters and re-assign all the points to the cluster of the closest centroid.

K-means demo

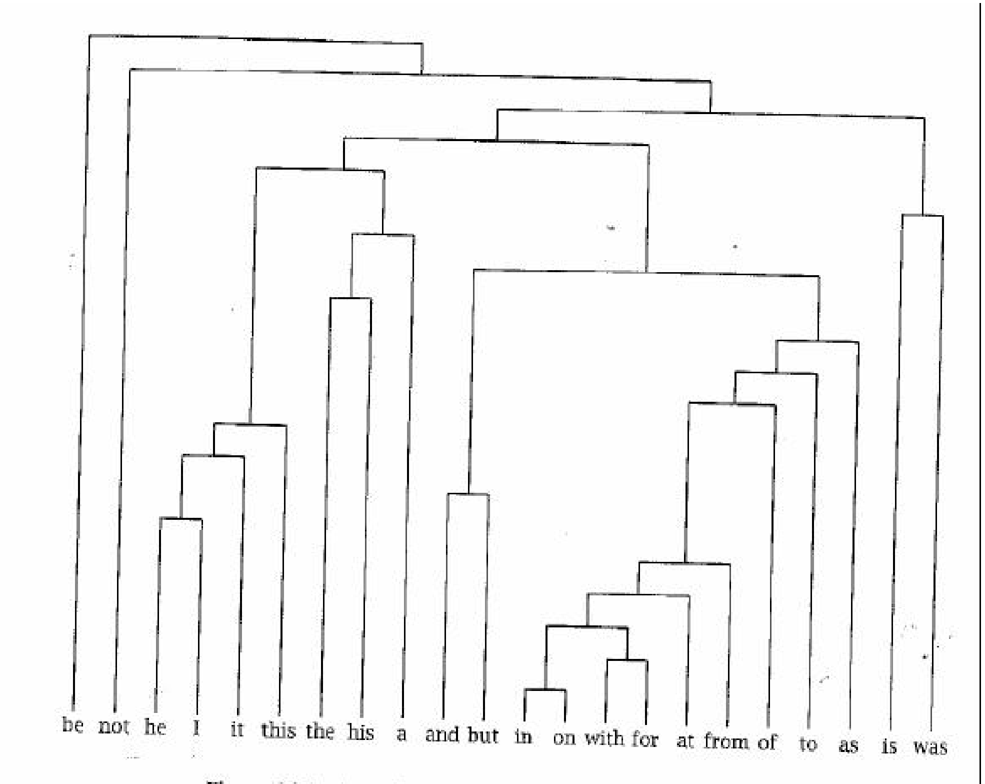
- K-means applet

Hierarchical clustering

- Techniques similar to construction of phylogenetic trees.
- A distance matrix for all genes are constructed based on distances between their expression profiles.
- Neighbor-joining or UPGMA can be applied on this matrix to get a hierarchical cluster.
- Single-linkage, complete-linkage, average-linkage clustering

Hierarchical clustering

- Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. A hierarchical clustering is often represented as a dendrogram.

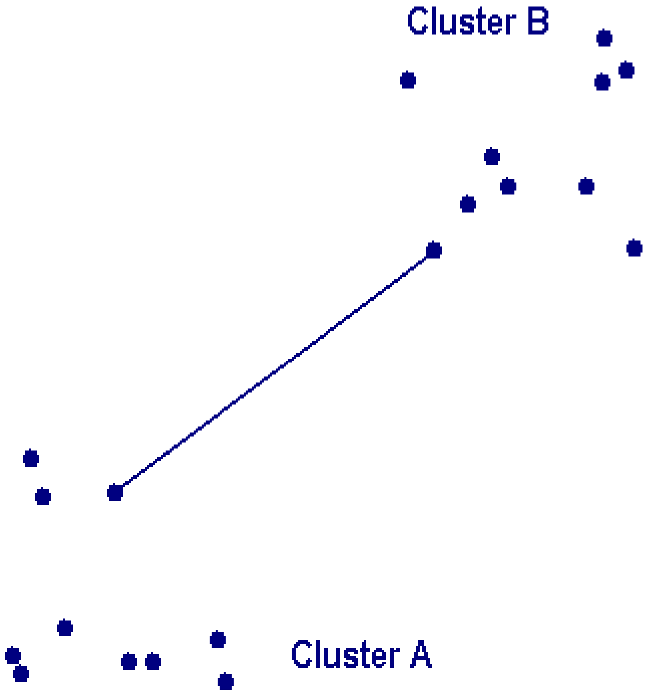


A hierarchical clustering of most frequently used English words.

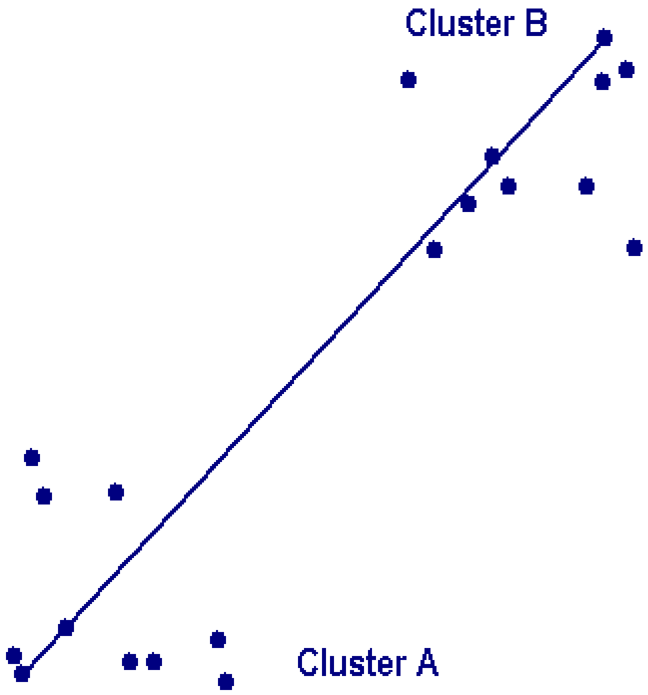
Hierarchical clustering

- In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest **maximum** pairwise distance).
- In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest **minimum** pairwise distance).

Inter-group distances



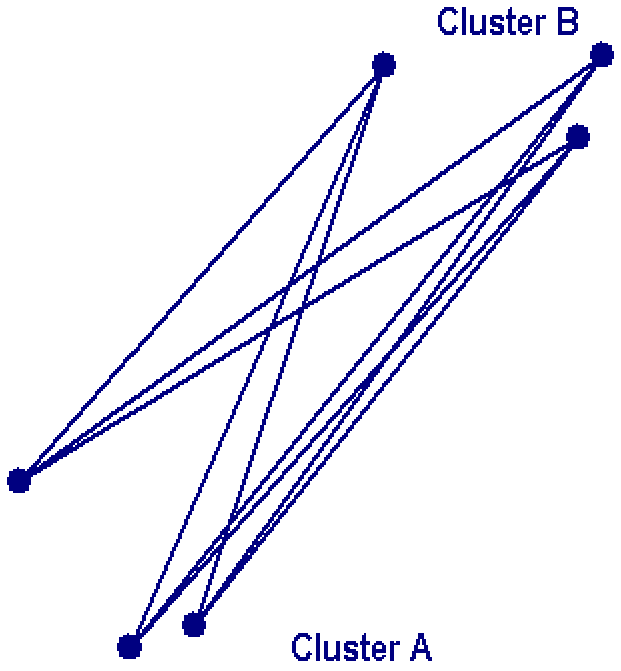
single-linkage



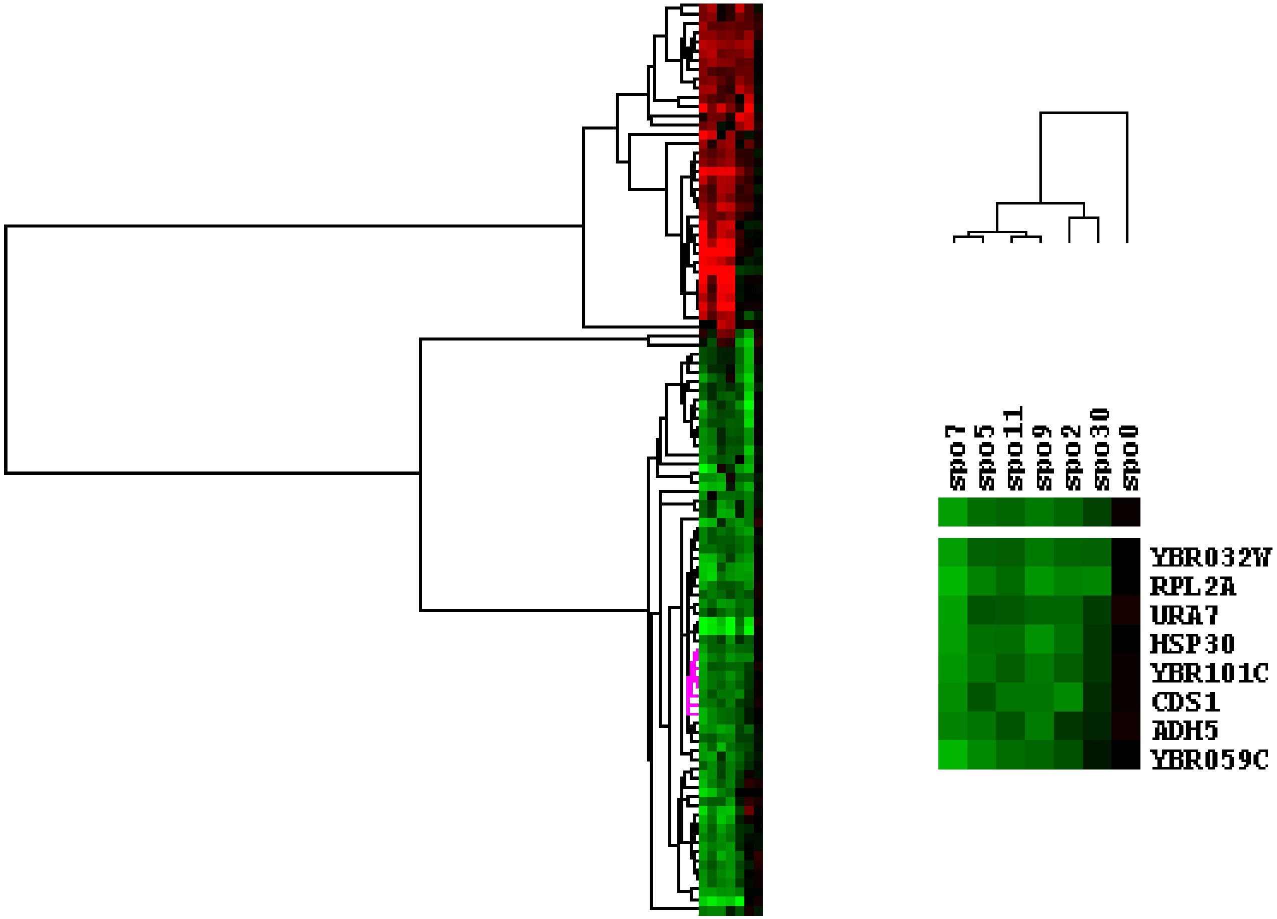
complete-linkage

Average-linkage

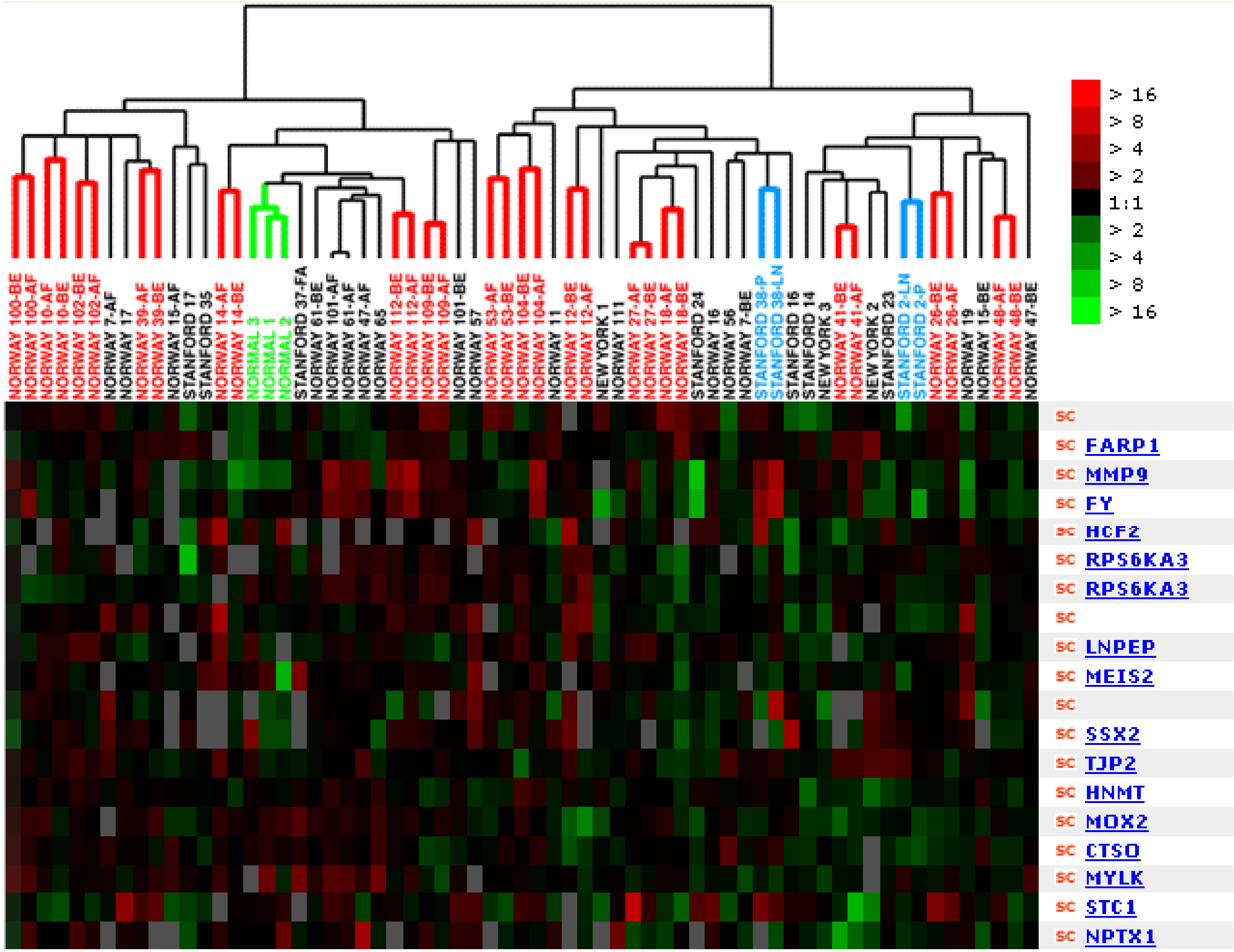
- UPGMA and neighbor-joining considers all cluster members when updating the distance matrix



Hierarchical Clustering



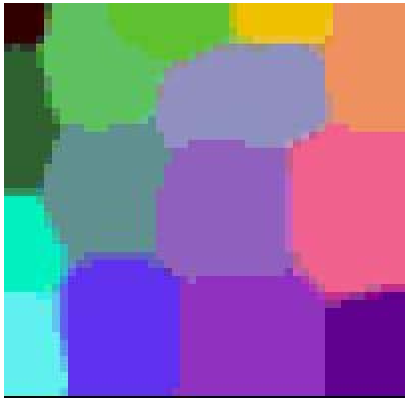
Hierarchical Clustering



Perou, Charles M., et al. *Nature*, 406, 747-752 , 2000.

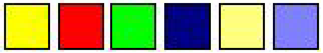
Self organizing maps (SOM)

- Self Organizing Maps (SOM) by **Teuvo Kohonen** is a data visualization technique which helps to understand high dimensional data by reducing the dimensions of data to a **map**.
- The problem that data visualization attempts to solve is that humans simply cannot visualize high dimensional data as is, so techniques are created to help us understand this high dimensional data.
- The way SOMs go about reducing dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together.



Components of SOMs: sample data

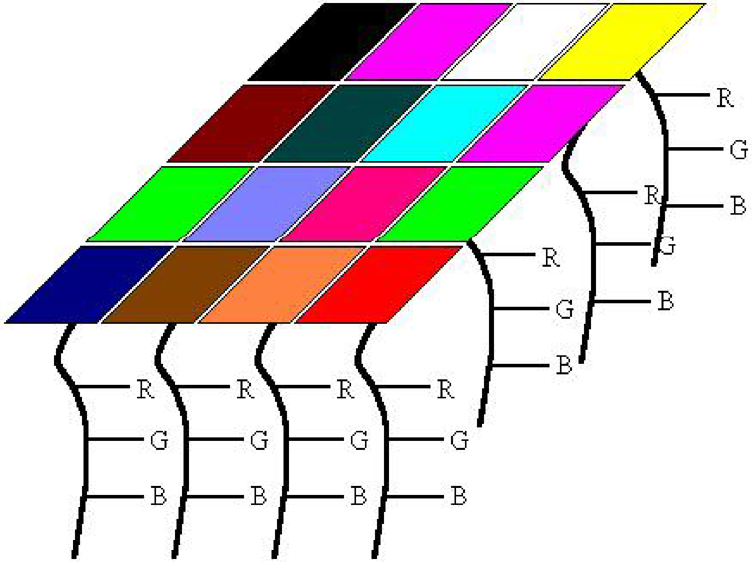
- The sample data that we need to cluster (or analyze) represented by n-dimensional vectors
- Examples:
  - colors. The vector representation is 3-dimensional: (r,g,b)



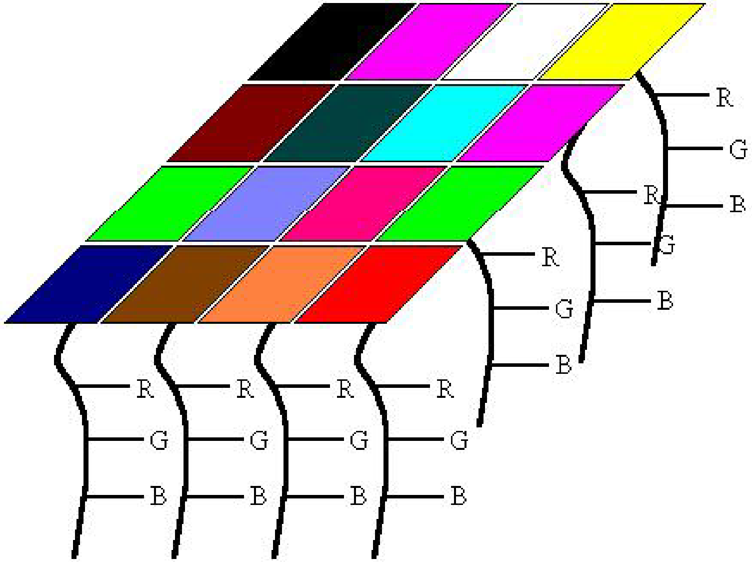
  - people. We may want to characterize 400 students in CEng. Are there different groups of students, etc. Example representation: 100 dimensional vector = (age, gender, height, weight, hair color, eye color, CGPA, etc.)

Components of SOMs: the map

- Each pixel on the map is associated with an n-dimensional vector, and a pixel location value (x,y). The number of pixels on the map may not be equal to the number of sample data you want to cluster. The n-dimensional vectors of the pixels may be initialized with random values.



Components of SOMs: the map



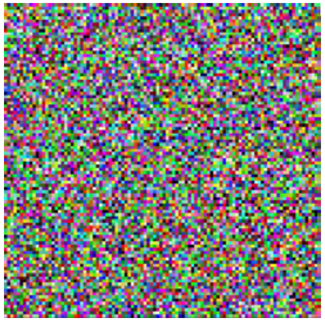
- The pixels and the associated vectors on the map are sometimes called “weight vectors” or “neurons” because SOMs are closely related to neural networks.

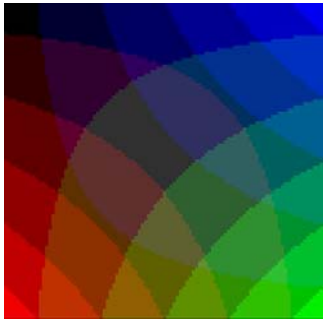
SOMs: the algorithm

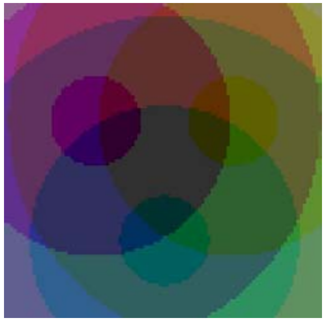
- initialize the map
- *for* **t** *from* 0 *to* 1
    - randomly select a sample
    - get the best matching pixel to the selected sample
    - update the values of the best pixel and its neighbors
    - increase **t** a small amount
- *end for*

Initializing the map

- Assume you are clustering the 400 students in CEng.
- You may initialize a map of size 500x500 (250K pixels) with completely random values (i.e. random people). Or if you have some information about groups of people *a priori*, you may use this to initialize the map.







Finding the best matching pixel

- After selecting a random student (or color) from the set that you want to cluster, you find the best matching pixel to this sample.
- Euclidian distance may be used to compute the distance between n-dimensional vectors.
  - I.e., you select the closest pixel using the following equation:
    - best_pixel = argmin $\sqrt{\sum_{i=1}^{n}(x_p - x_{sample})^2}$

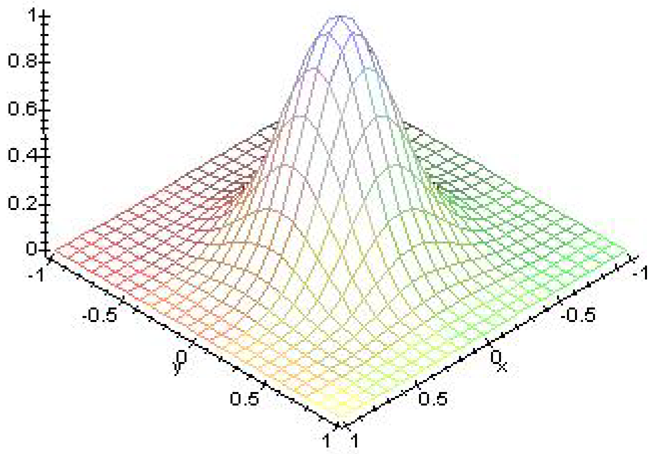
for all p $\in$ map

Updating the pixel values

- The best matching pixel and its neighbors are allowed to update themselves to resemble the selected sample
  - new vector of a pixel is computed as

    current_pixel_value*(t)+sample_value*(1-t)
  - in other words, in early iterations when t is close to 0, the pixel directly copies the properties of the randomly selected sample, but in subsequent iterations the allowed amount of changes decreases.
  - Similarly for the neighbors of the best pixel, as the distance of the neighbor increases, they are allowed to update themselves in a smaller amount.

Updating the pixel values



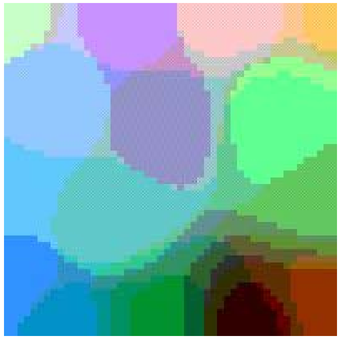
- A Gaussian function can be used to determine the neighbors and the amount of update allowed in each iteration. The height of the peak of the Gaussian will decrease and base of the peak will shrink as time (**t**) progresses.

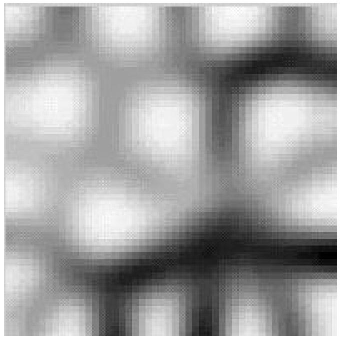
Why do similar objects end up in near-by locations on the map?

- Because a randomly selected sample, A, influences the neighboring samples to become similar the itself at a certain level.
- At the following iterations when another sample, B, is selected randomly and it is similar to A. We have a greater chance of obtaining B’s best pixel on the map closer to A’s best pixel, because those pixels around A’s best pixel are updated to resemble A, if B is similar to A, its best pixel may be found in the same neighborhood.

How to visualize similarities between high-dimensional vectors?

- Colors are easy to visualize, but how do we visualize similarities between students?
- The SOM may show how similar a pixel is to its neighbors (dark color: not similar, light color: similar). White blobs in the map will represent groups of similar people. Their properties can be analyzed by inspecting the vectors at those pixels.





SOM demo

- SOM applet