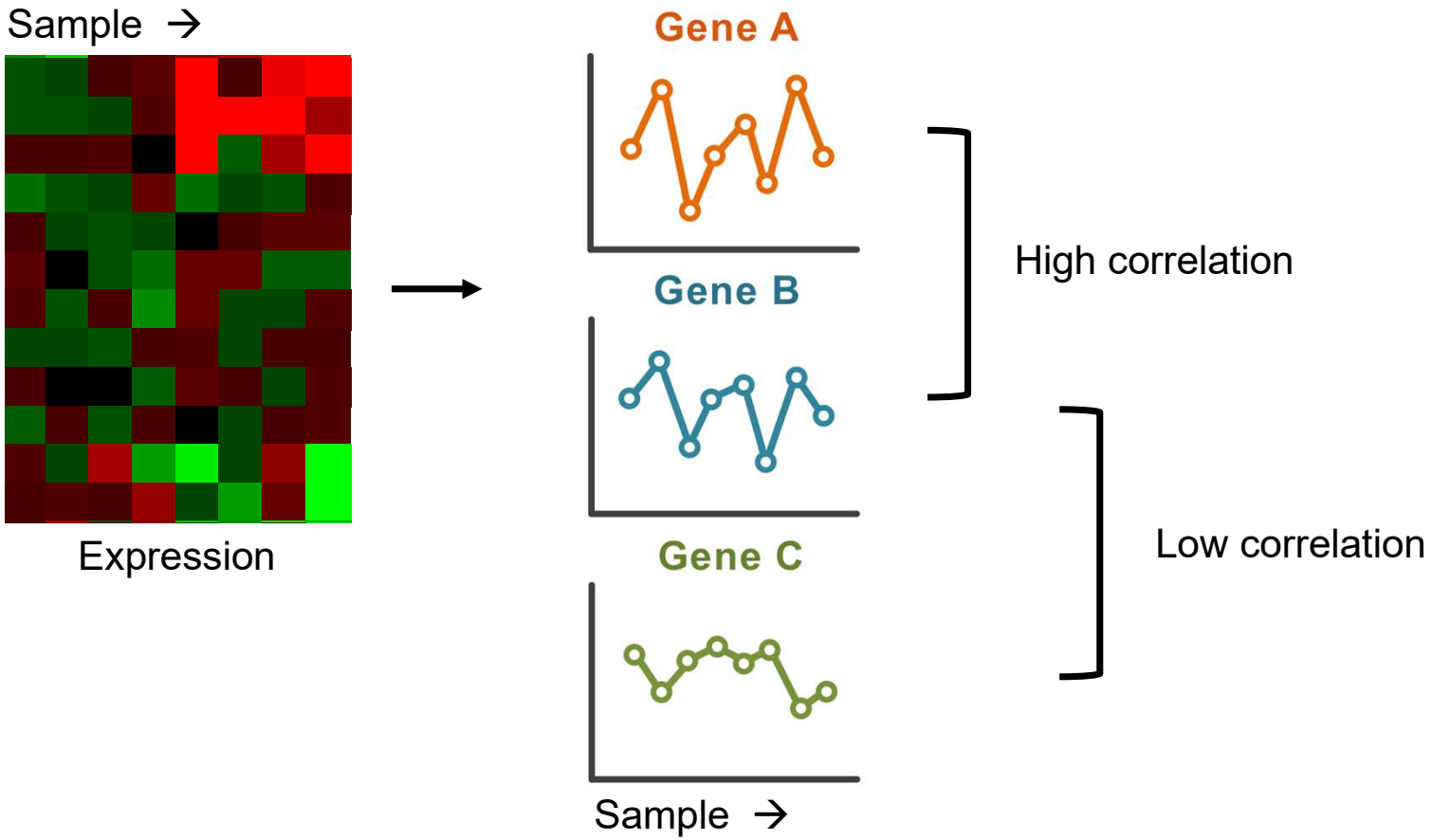
Statistic 5

-network
nw, degree centrality
-spatial
 autocorrelation, ripley k, tda, convex hull
-extra
 NMF, KL-divergence, earth movement, optimal transport,
 synthetic data: smoth

Network analysis

- Network analysis



Gene pair: correlation (Pearson correlation coefficient: PCC / Spearman correlation coefficient (SCC)

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

SCC: rank-based (non-parametric)

❖ Similarity metrics: for continuous data

➢ Pearson correlation (Centered correlation)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{l=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

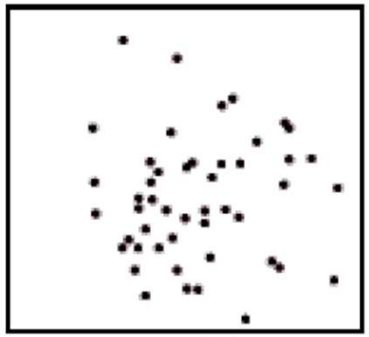
- The values of the Pearson correlation coefficient range **between -1 and +1**, with **r=1 when the two vectors are identical** (perfect correlation), r = -1 when the two vectors are exact opposites (perfect anti-correlation), and r = 0 when the two vectors are completely independent (uncorrelated or orthogonal vectors).
- The Pearson correlation coefficient is very useful **if the "shape" of the expression vector is more important than its magnitude**.

➢ Uncentered Pearson correlation

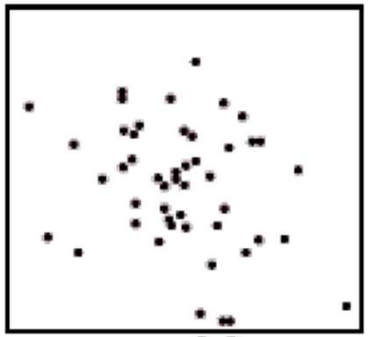
- **If the relative expression level is important**, it is better to use this.

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{l=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

❖ Co-expression vs. Pearson correlation



r=0.0

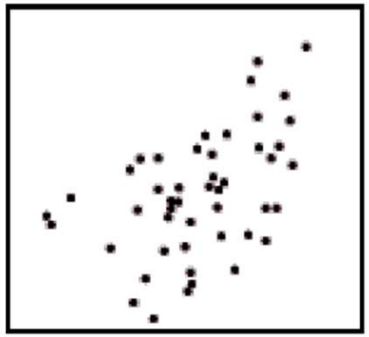


r=-0.3

X-axis:
expression of gene A

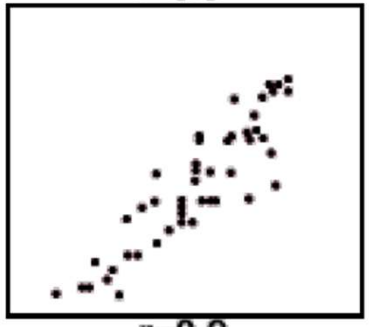
Y-axis:
expression of gene B

Each data point:
each array
(each experiment)

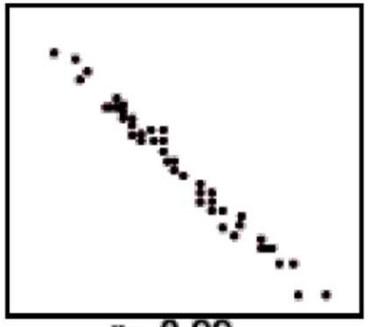


r=0.5

r=-0.7



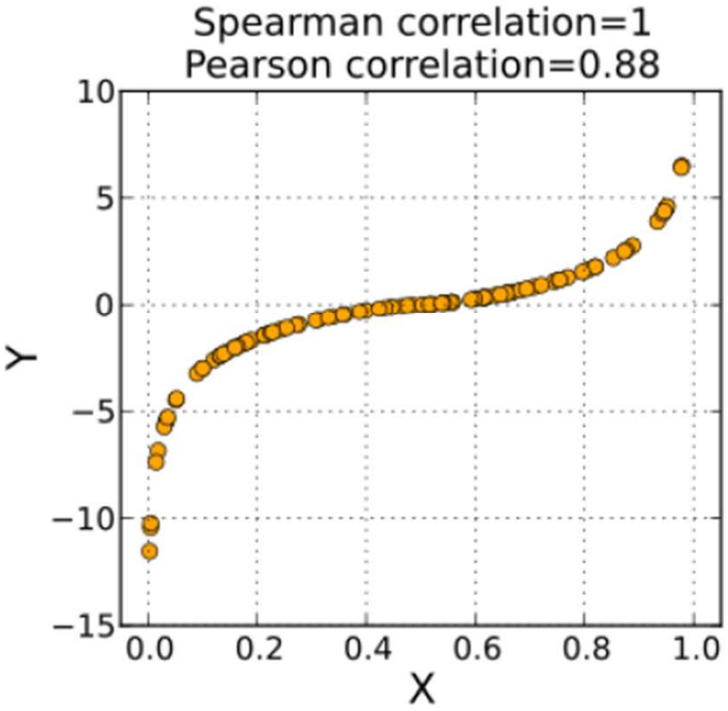
r=0.9

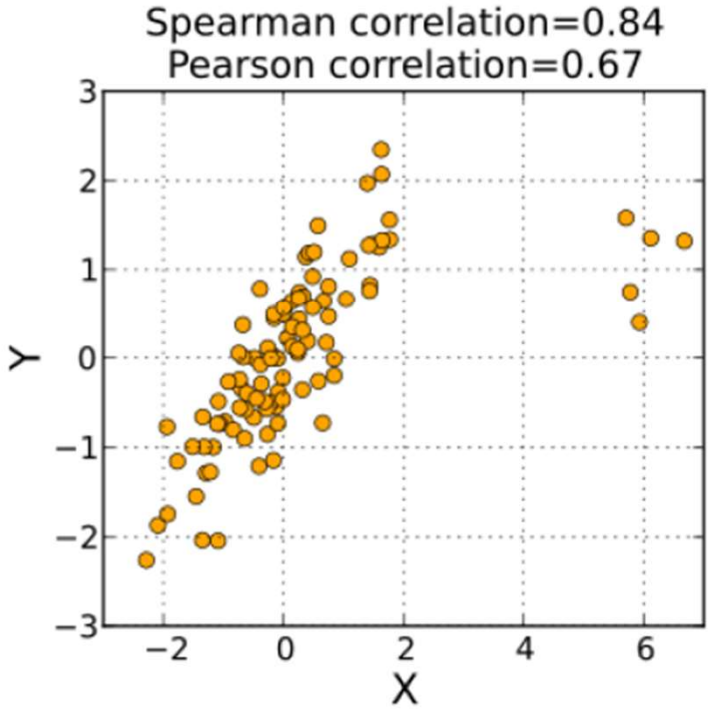


r=-0.99

- **Pearson correlation coefficient**: *assumes linear correlation*; measure positive and negative correlation separately; drawback- *sensitivity for outliers*

- **Spearman (rank) correlation coefficient**: *can detect non-linear correlation*; measure positive and negative correlation separately; rank-order based; *robust to outliers*. It is defined as the Pearson correlation coefficient between the ranked variables (convert all raw scores to ranks for coefficient calculation).

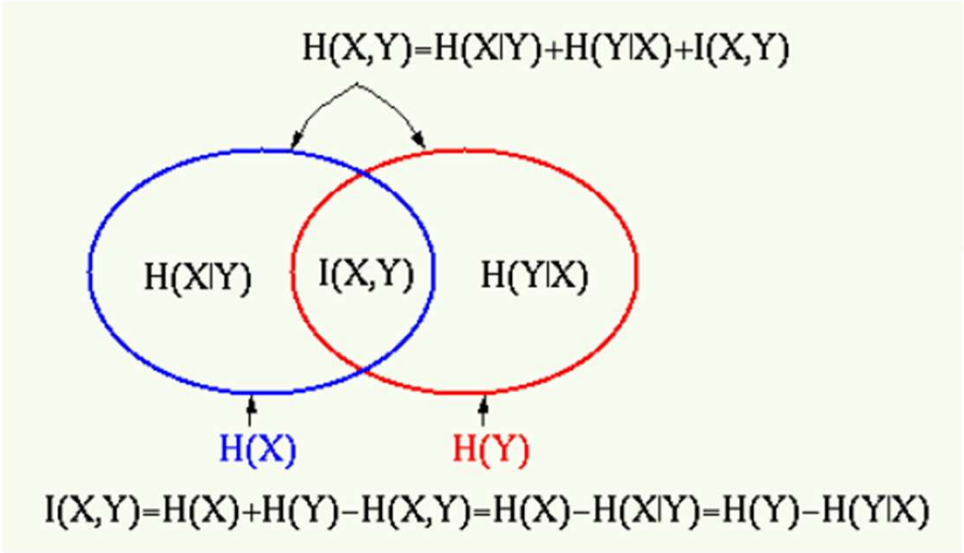




❖ Similarity metrics: for binary data

➢ Mutual Information (MI)

- Mutual information is **entropy based** and **non-parametric** measure.
- It is a measure of the **mutual dependence between the two random variables**.
- It quantifies **the amount of information (at bits) obtained about one random variable through the other random variable**.



- Conditional entropy: H(X|Y), H(Y|X)
  Joint entropy: H(X, Y)
- Marginal entropy: H(X), H(Y)

$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X,Y),$$
$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)},$$

- *I(X, Y)* increases when marginal entropy increases and joint entropy decrease

<Summary>

- **Pearson correlation coefficient [-1, 1]**: assumes linear correlation; measure positive and negative correlation separately; drawback- sensitivity for outliers
- **Spearman correlation coefficient [-1, 1]**: can detect non-linear correlation; measure positive and negative correlation separately; rank-order based; robust to outliers.
- **Mutual information [0, ∞]**: counts complexity of each expression profile; robust for sparse matrix; no directionality (positive/negative) of correlation

❖ Network Biology

"Molecular network" is the core component of Systems Biology.

Definition:

The study of complex biological systems through interactions between bio-molecules or other system components.

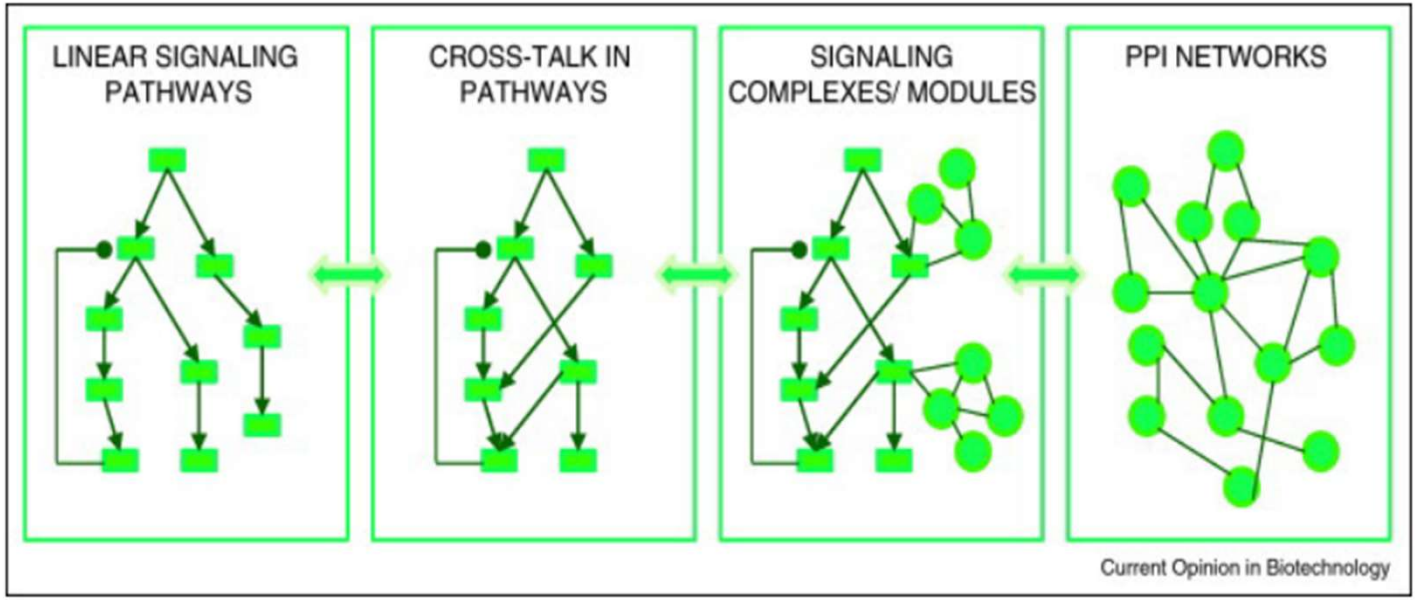
Questions:

- How to construct molecular networks.
- How to generate hypothesis from the molecular network.

❖ Types of biomolecular networks

Type	Directionality	Node	edge
Protein-protein interaction network	Undirectional	Protein	Physical interaction
Gene co-function network	Undirectional	Gene	Functional association
TF-target regulatory network	From TF to genes to be regulated	TF and target gene	Regulation
miRNA-target regulatory network	From miRNA to genes to be regulated	miRNA and target gene	Regulation
Drug-target network	From drug to proteins to be inhibited	Drug and target protein	Inhibition

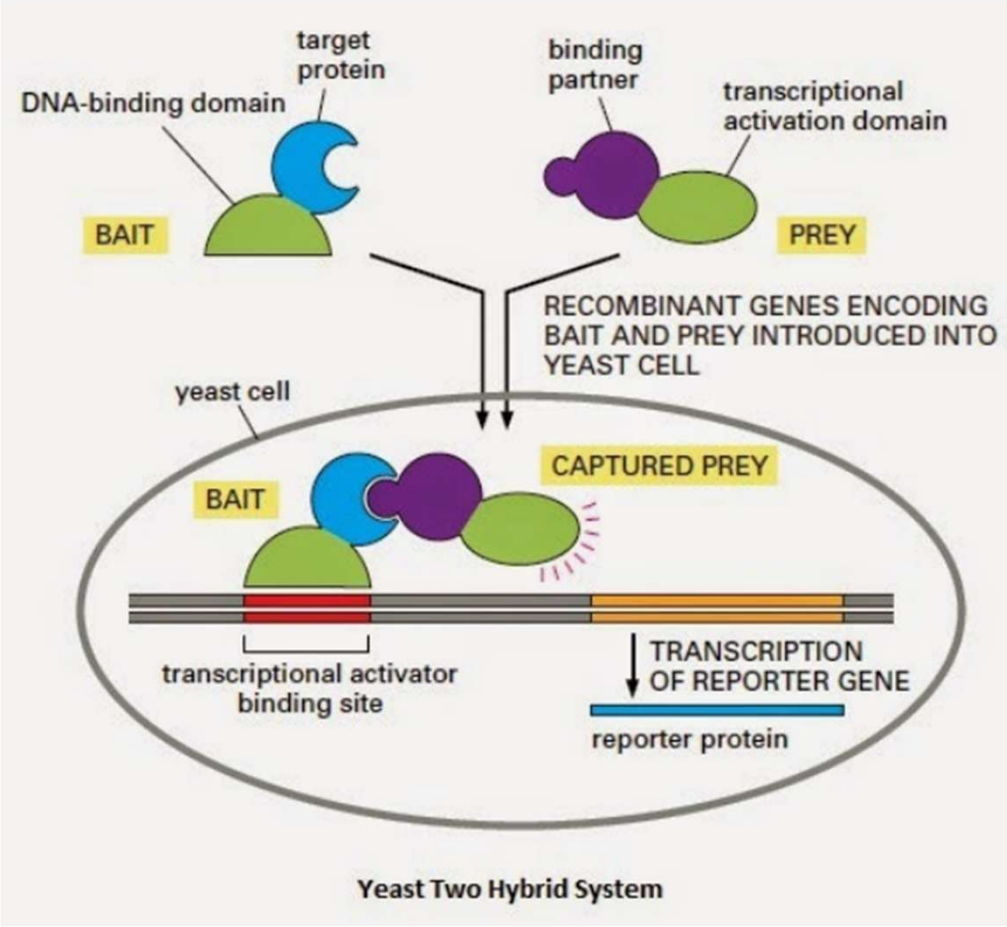
- **Paradigm shift: from linear pathways to complex networks**



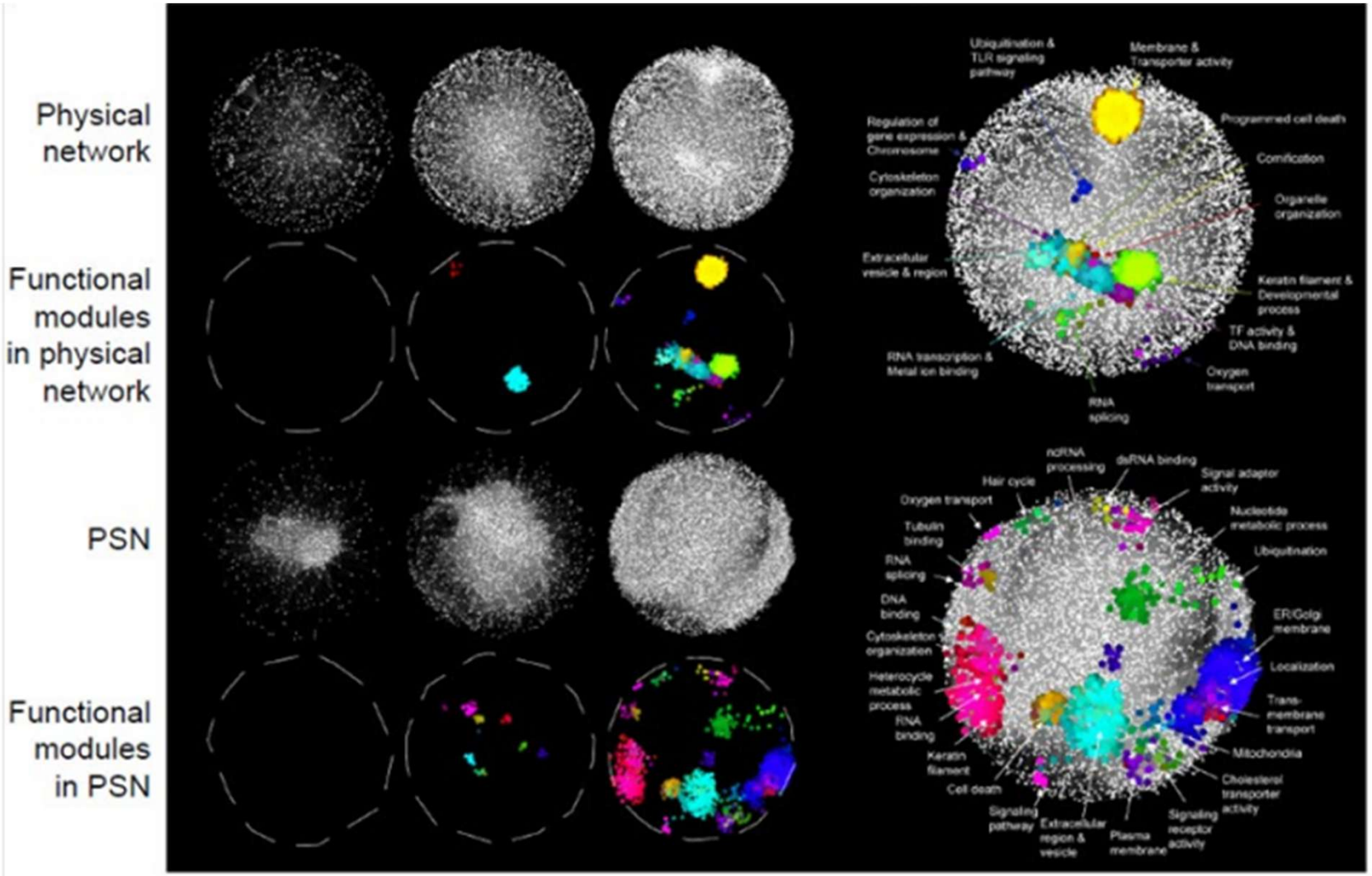
From linear pathways to protein protein interaction (PPI) networks. Illustration of the relation between linear signaling pathways, cross-talk in pathways, signaling complexes/modules, and PPI networks. (Current Biology 23:305)

❖ Experimental determination of PPI

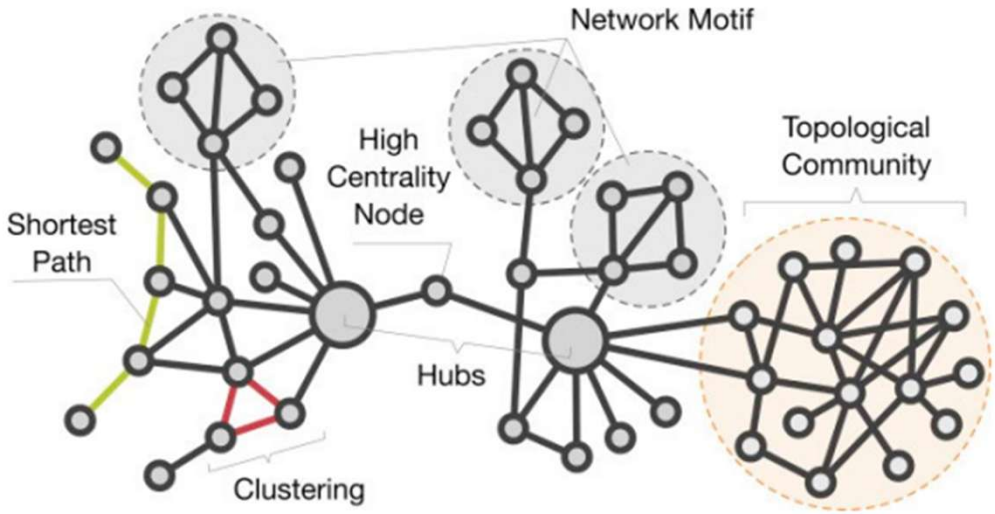
- **Yeast two-hybrid (Y2H):**
- Invented by Stanley Fields and Ok-Kyu Song (*Nature 340:245, 1989)*
- Originally use Gal4 transcriptional activator involved in galactose utilization in yeast



Yeast Two Hybrid System


Physical network
Functional modules in physical network
PSN
Functional modules in PSN
Ubiquitination & TLR signaling pathway
Membrane & Transporter activity
Regulation of gene expression & Chromosome
Programmed cell death
Cytoskeleton organization
Cornification
Organelle organization
Extracellular vesicle & region
Keratin filament & Developmental process
TF activity & DNA binding
RNA transcription & Metal ion binding
Oxygen transport
RNA splicing
ncRNA processing
dsRNA binding
Signal adaptor activity
Hair cycle
Oxygen transport
Nucleotide metabolic process
Tubulin binding
Ubiquitination
RNA splicing
DNA binding
ER/Golgi membrane
Cytoskeleton organization
Localization
Heterocycle metabolic process
Trans-membrane transport
RNA binding
Mitochondria
Keratin filament
Cholesterol transporter activity
Cell death
Signaling pathway
Extracellular region & vesicle
Plasma membrane
Signaling receptor activity

❖ Basic topological characteristics of networks



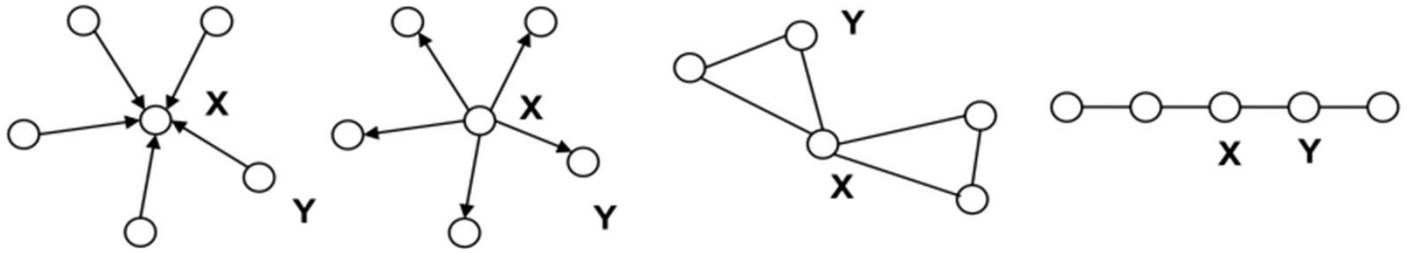
- The **degree** of a node is the number of links attached to it, i.e., the number of direct neighbors.
- A **path** between two nodes is a sequence of links connecting the two. The minimum number of links needed to connect the two is called '**shortest path length**'.
- **Centrality measures** exist for both nodes and for links and quantify their topological importance within the network. There are different types of centrality measures, e.g., the '**degree centrality**' (simply given by the degree) or '**betweenness centrality**' (quantifying how many shortest paths of the full network cross through a certain node).
- **Clustering** describes a tendency observed in many biological (and other) networks that two neighbors of a node are often also connected to each other, thus forming a triangle.
- **Motifs** are small recurrent subgraphs in a network that occur particularly frequently.
- **Network communities** are groups of tightly interconnected nodes that have more connections among themselves than to the rest of the network.

(*Curr. Opin. Syst. Biol. 3:88, 2017*)

- ***Node centrality: Depends on Application***
  - Influence: Which social network nodes should I pick to advertise/spread a product/opinion?
  - Resilience: Which node(s) should I attack to disconnect the network?

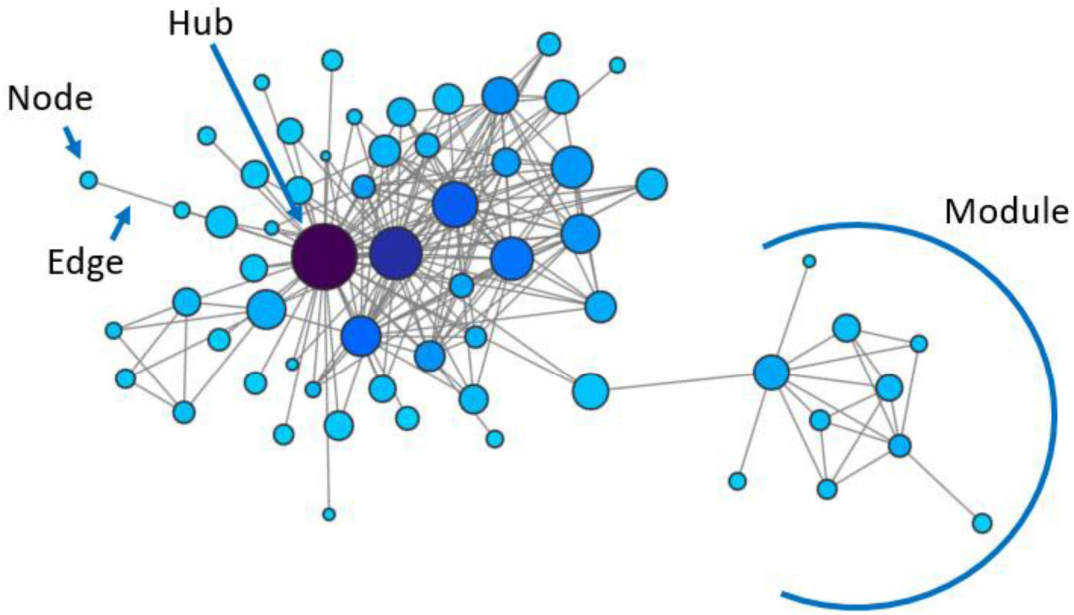
- ***Centrality: Importance based on network position.***

  In each of the following networks, X has higher centrality than Y according to a particular measure

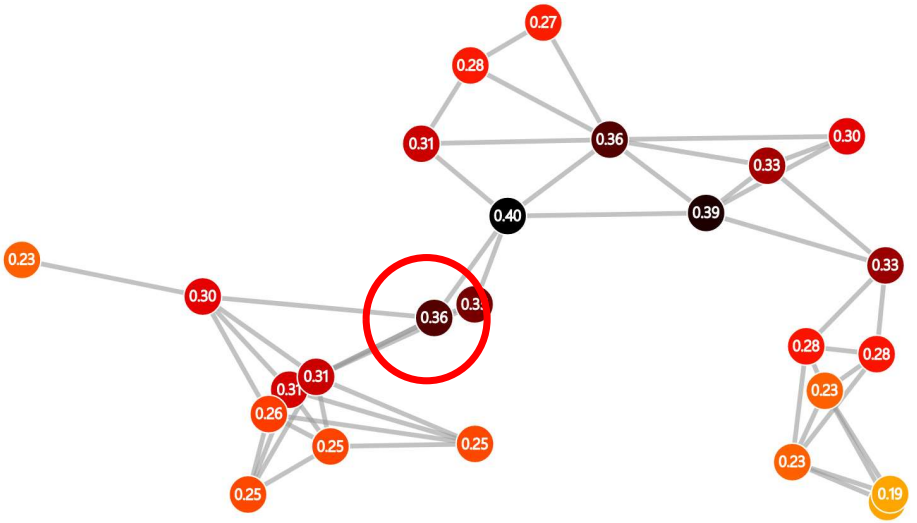


indegree outdegree betweenness closeness

- Network analysis



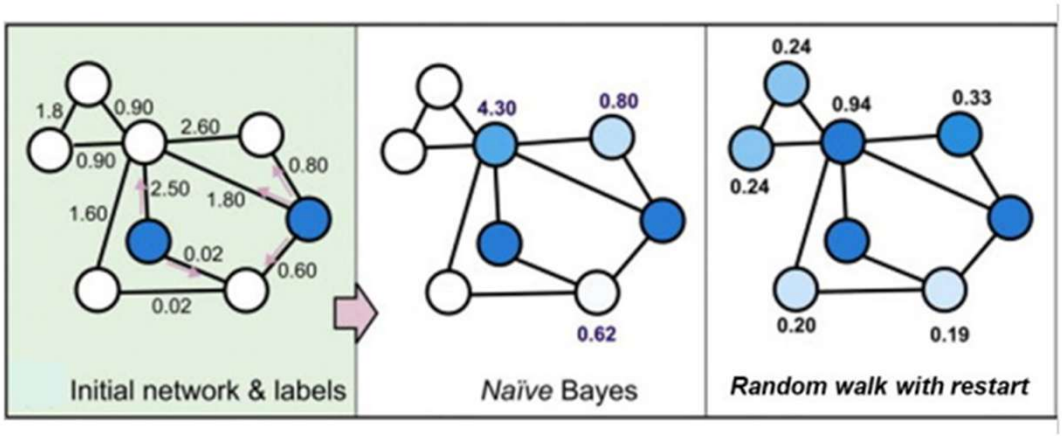
Centrality by node degree



Centrality by betweenness

-Node degree: number of edges for each node → highest
-Betweenness: find shortest path for each node pair → sort by how many shortest paths pass each node → highest

❖ Network propagation algorithms



- ***Naïve Bayes***: Assigns **scores to direct neighbors** with accounting for edge weight
- ***Random walk with restart (RWR):*** a popular **network diffusion method** which assigns scores to all nodes including indirect neighbors by diffusing initial scores through the entire network (there are more network diffusion methods such as heat diffusion, Gaussian smoothing and etc.)

❖ Random Walks (on Graphs)

- Start from a node chosen uniformly at random with probability $\frac{1}{n}$.
- Pick one of the outgoing edges uniformly at random
- Move to the destination of the edge
- Repeat.

➢ What is the probability p_i^t of being at node i after t steps?

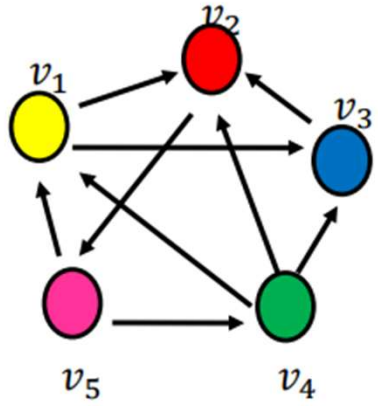
$$p_1^0 = \frac{1}{5} \qquad p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^0 = \frac{1}{5} \qquad p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^0 = \frac{1}{5} \qquad p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^0 = \frac{1}{5} \qquad p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^0 = \frac{1}{5} \qquad p_5^t = p_2^{t-1}$$



➢ Stationary distribution

- After many steps ($t \to \infty$) the probabilities **converge** (updating the probabilities does not change the numbers)
- The converged probabilities define the **stationary distribution** of a random walk π
- The probability π_i is the **fraction of times that we visited node i as $t \to \infty$**
- **Markov Chain Theory:** The random walk converges to a unique stationary distribution independent of the initial vector if the graph is strongly connected, and not bipartite.

Spatial analysis

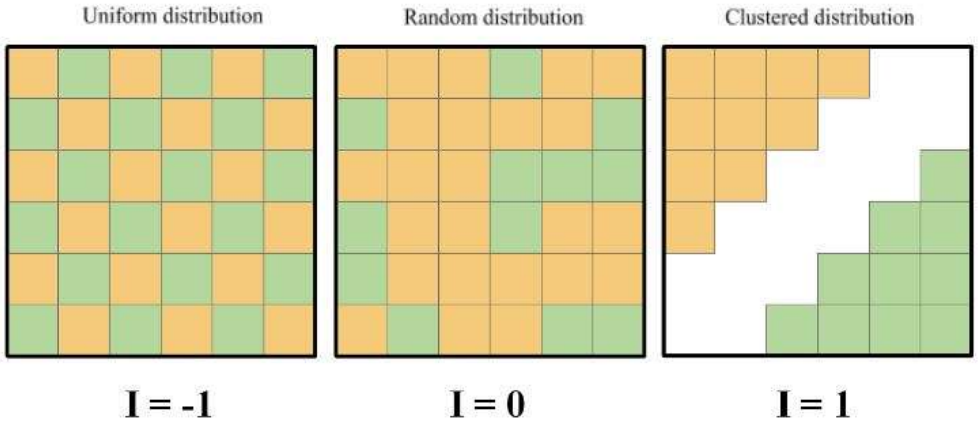
Moran's I

Spatial autocorrelation (colocalization)

-1: perfect symetric

0: random

1: clusterized



Global Moran's *I* is a measure of the overall clustering of the spatial data. It is defined as

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

where

- N is the number of spatial units indexed by i and j;
- x is the variable of interest;
- $\bar{x}$ is the mean of x;
- w_{ij} are the elements of a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$);
- and W is the sum of all w_{ij} (i.e. $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$).

Ripley's K function

$$\widehat{K}(t) = \lambda^{-1} \sum_{i \neq j} \frac{I(d_{ij} < t)}{n},$$

where d_{ij} is the Euclidean distance between the i^{th} and j^{th} points in a data set of n points, t is the search radius, λ is the average density of points (generally estimated as n/A, where A is the area of the region containing all points) and I is the indicator function (i.e. 1 if its operand is true, 0 otherwise).[3] In 2 dimensions, if the points are approximately homogeneous, $\widehat{K}(t)$ should be approximately equal to πt^2.

→ Colocalization pattern within radius "t"

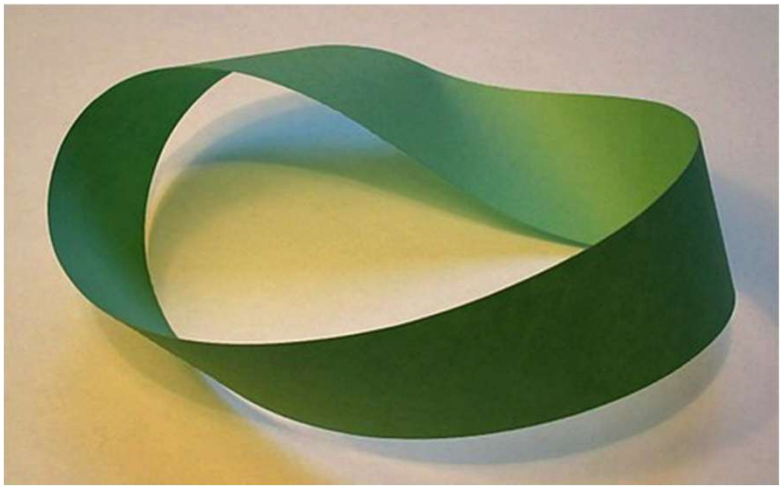
-K(t) > πr^2 : clusterized

-K(t) < πr^2 : dispersed

-K(t) > πr^2 : random

More than a slight detour into pure mathematics ….

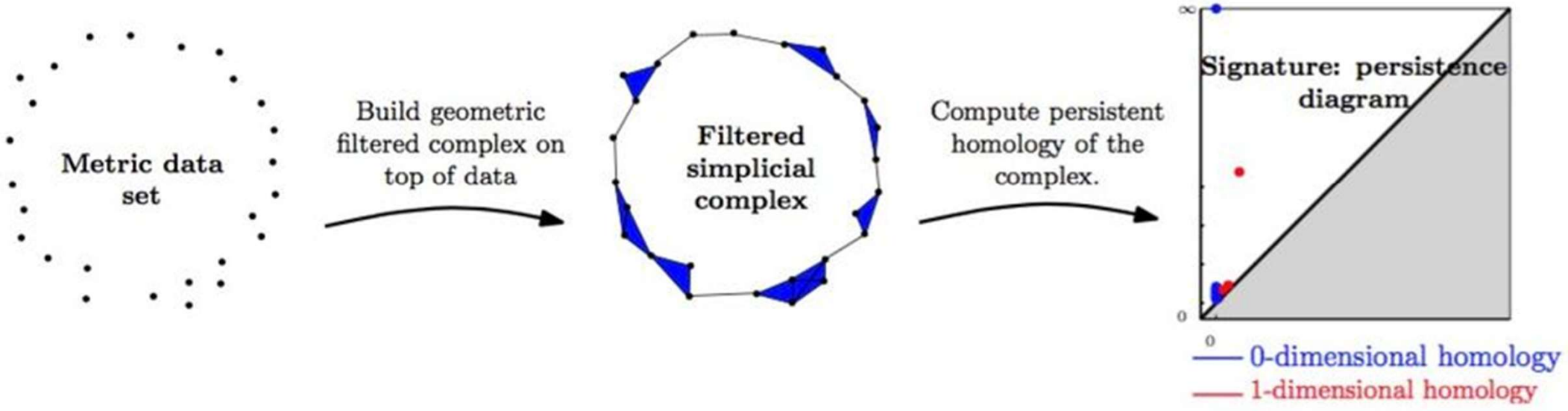
- Topology is concerned with geometric properties of objects
- Topological data analysis (TDA) applies these methods to noisy data



Persistent homology

- Compute topological features using a Rips complex
- Summarize using a persistence diagram

Persistent homology → arrangement of coordinates



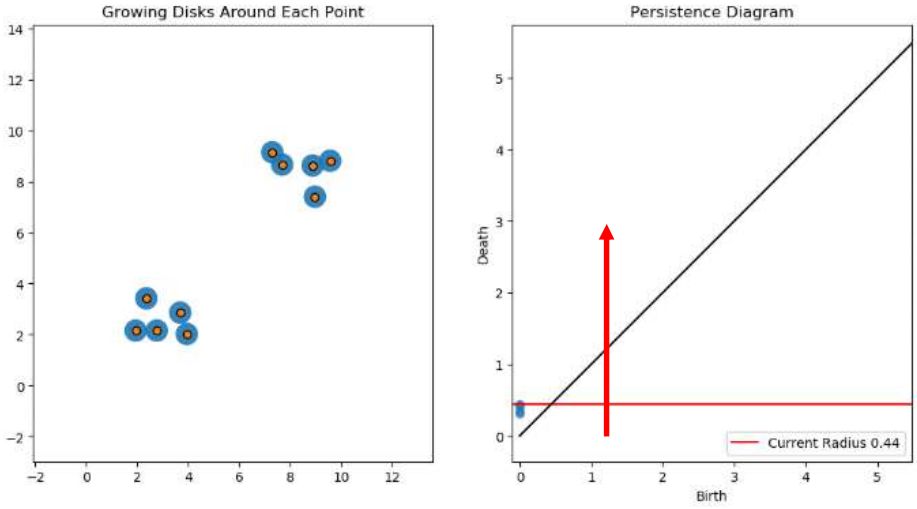
Persistent homology

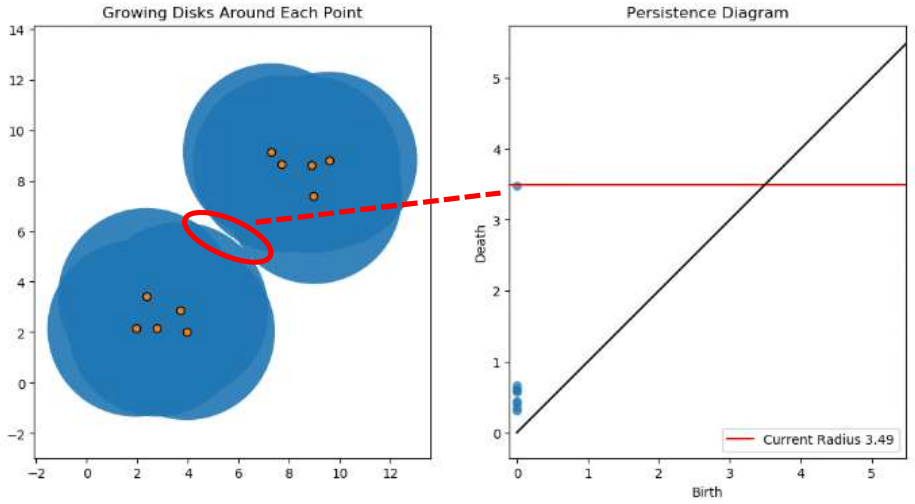
ds = pairwise(Euclidean(), d, dims=2);

ripserer(ds, dim_max=0)
# calculate the rips complex
→ Persistent diagram

Born: threshold =0;
all the connected component

Birth: new homology
Death: disappearing homology
Persistent: maintains homology for long time



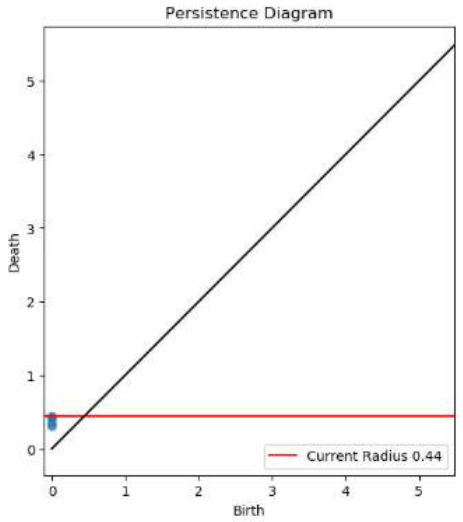


Euclidian space → grow radius threshold → overlap: death → sample color (identity)
New death

Persistent homology

```
julia> ripsH0
545×2 Matrix{Float64}:
 0.0    0.000750268
 0.0    0.000766613
 0.0    0.000845232
 0.0    0.000845232
 0.0    0.000862439
 0.0    0.00093351
 0.0    0.00103075
 0.0    0.0010748
 0.0    0.00109301
 0.0    0.00109469
 0.0    0.00109469
 0.0    0.00109762
 0.0    0.00112489
 0.0    0.00113586
 0.0    0.00114552
 0.0    0.00114712
 0.0    0.00114712
 0.0    0.00115152
 0.0    0.00115986
 0.0    0.00116617
 0.0    0.00117871
 0.0    0.00118531
 ⋮
 0.0    0.0069288
 0.0    0.00695215
 0.0    0.00695261
 0.0    0.0069745
```

Birth death

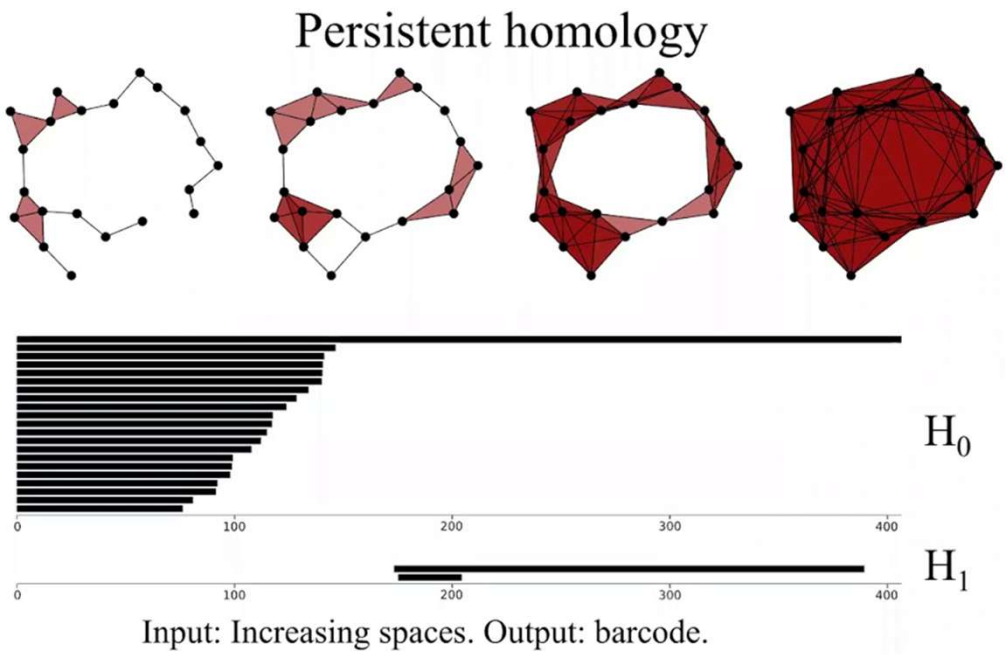


ripsH0[l,:] = [rips[1][l].birth, rips[1][l].death];

High-order persistent homology

-H1 persistent homology
Birth: loop: closed topology
Death: when the loop disappear

-H2: sphere with hole inside



Input: Increasing spaces. Output: barcode.

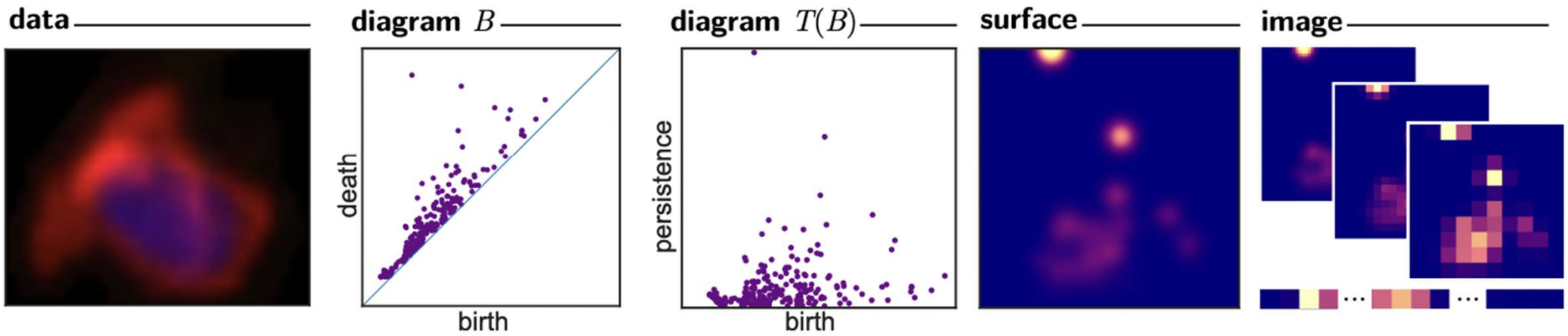
Persistent images allows for ML applications

- Convert the persistence diagram into a vector
- This vector can be plugged into traditional ML approaches for classification

Journal of Machine Learning Research 18 (2017) 1-35 Submitted 7/16; Published 2/17

Persistence Images: A Stable Vector Representation of Persistent Homology

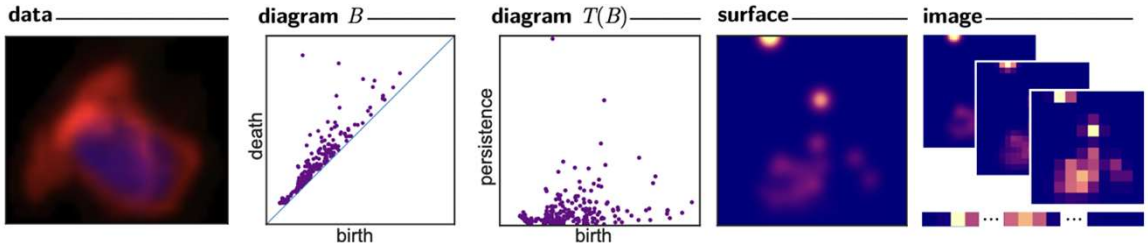
Henry Adams ADAMS@MATH.COLOSTATE.EDU
Tegan Emerson EMERSON@MATH.COLOSTATE.EDU



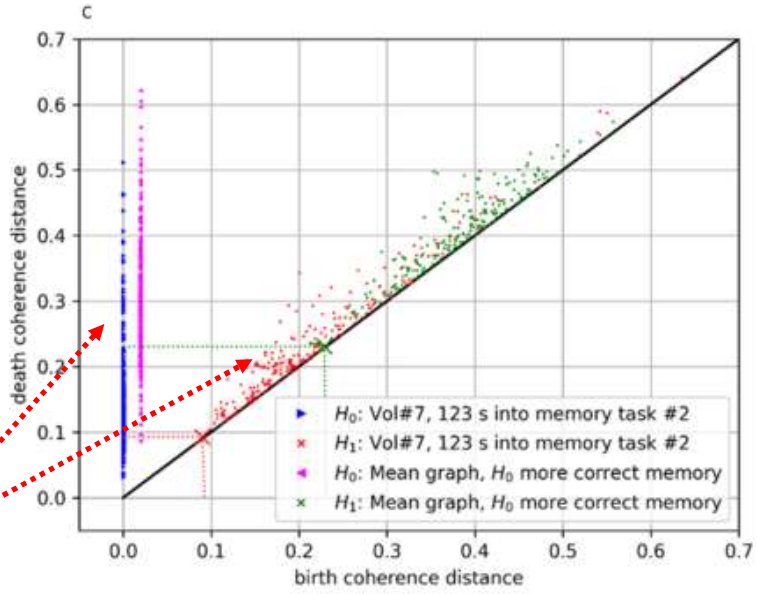
Persistent images allows for ML applications

Persistent diagram → persistent image
(Dimension reduced matrix)

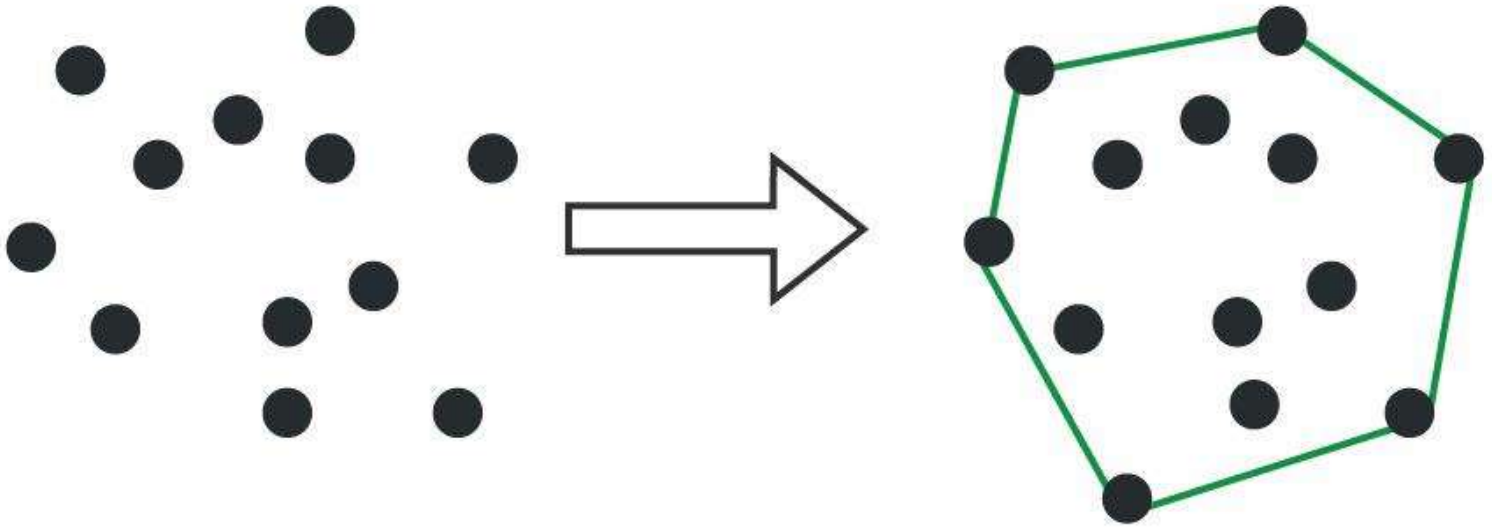
Persistence = death - birth



T(B) → gaussian smoothing
H0: birth → 0 → persistent image: 1 dimension
H1: persistent image: 2 dimension



Convex Hull



-Make a polygon (smallest convex set covers every data point)

Extra

- cNMF

(consensus non-negative matrix factorization)

*Unsupervised approach
NMF: Decomposition method
Make W,H be close to X
Gradient descending

$$X = WH$$

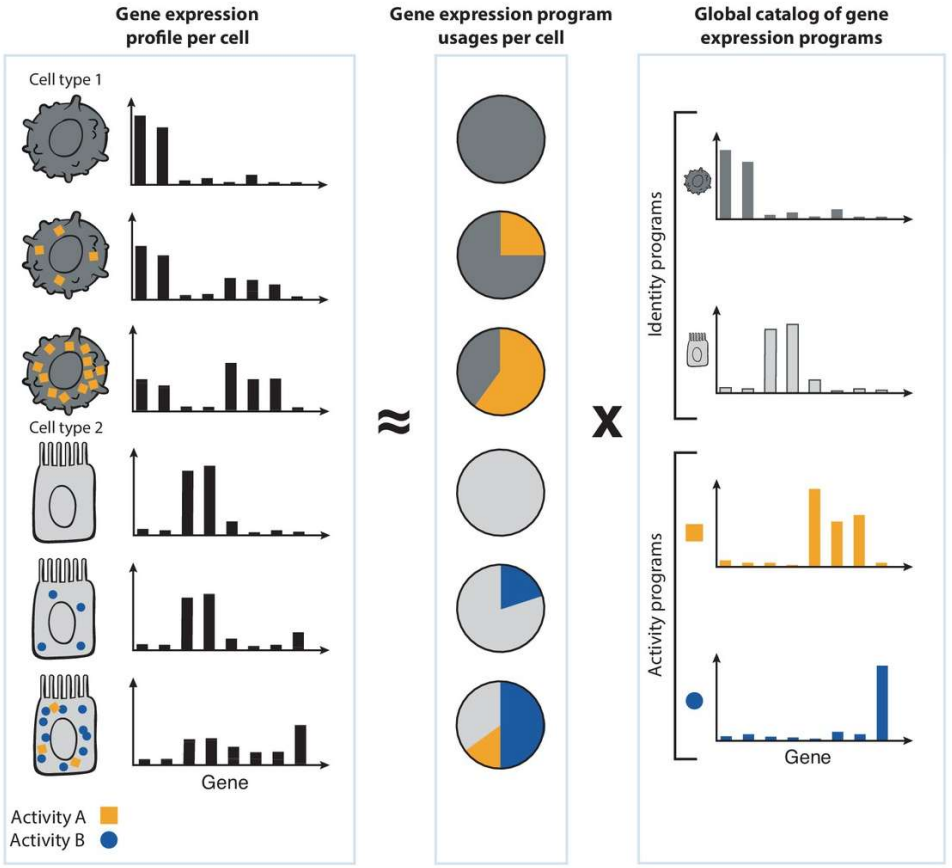
$$H := H - \eta_H \circ \nabla_H \|X - WH\|_F^2$$

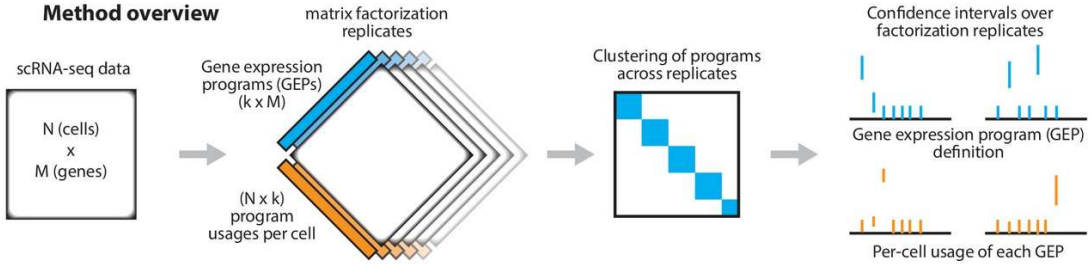
$$W := W - \eta_W \circ \nabla_W \|X - WH\|_F^2$$

$$\therefore H := H \circ \frac{W^T X}{W^T W H}$$

$$\text{식}(36) \Rightarrow W := W + \frac{W}{WHH^T} \circ (XH^T - WHH^T)$$

$$= W + W \circ \frac{XH^T}{WHH^T} - W \circ \frac{WHH^T}{WHH^T} = W \circ \frac{XH^T}{WHH^T}$$





- cNMF

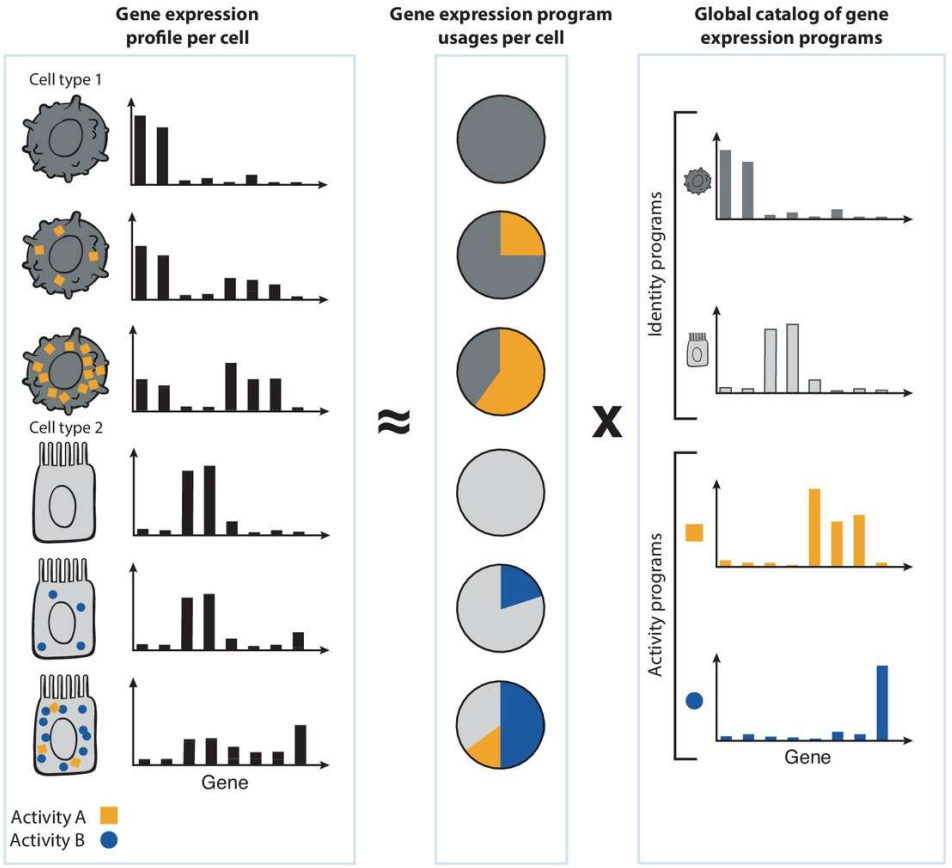
X = WH

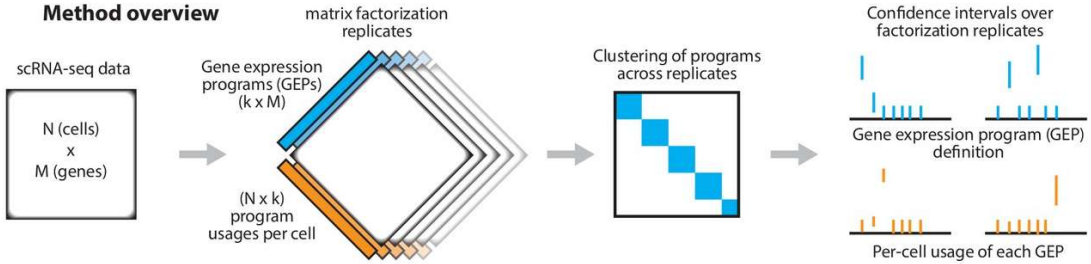
-X: gene expression (N x M)
N: cell, M: gene

-W: program usage (activity): N x k
k: number of program

-H: gene expression program: weight of each gene
k x M

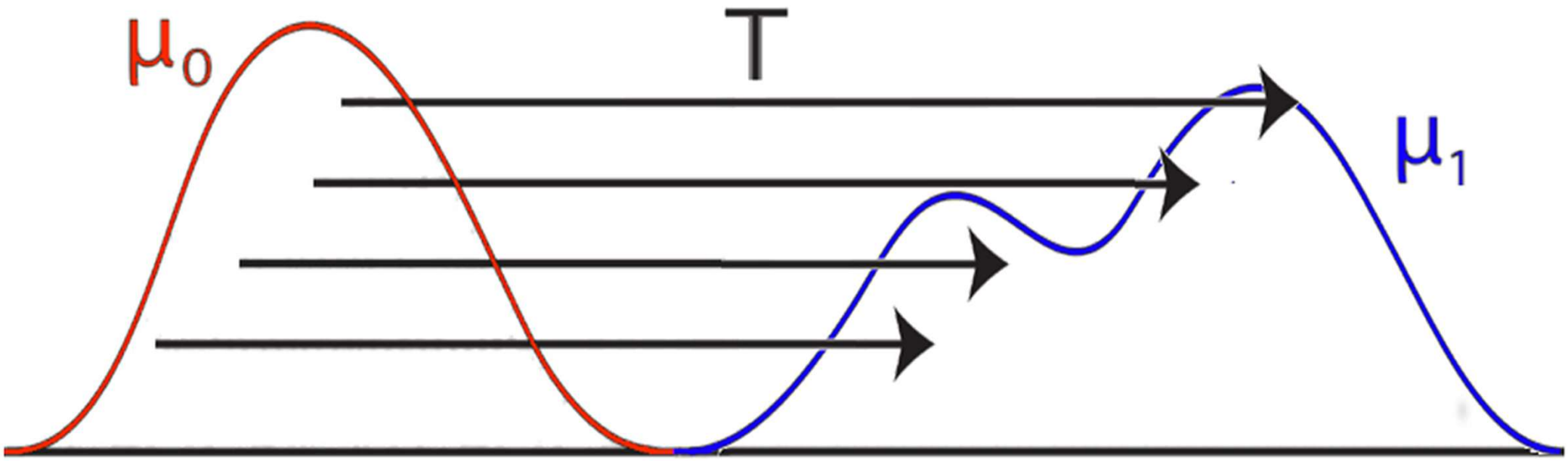
Consensus → robustness
Take median value of each gene



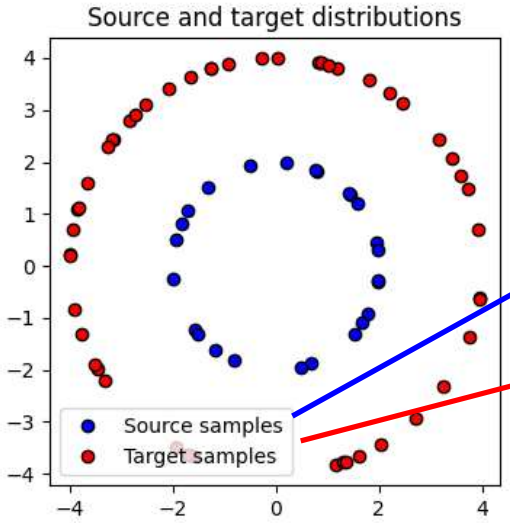


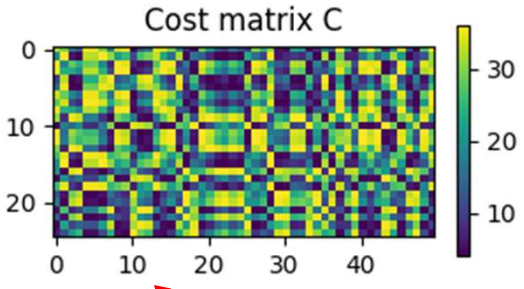
Optimal transport

-Transport plan γ(x,y): measure the distance between two distributions
→ define how to transport each sample point
-Cost function c(x,y): cost for x → y

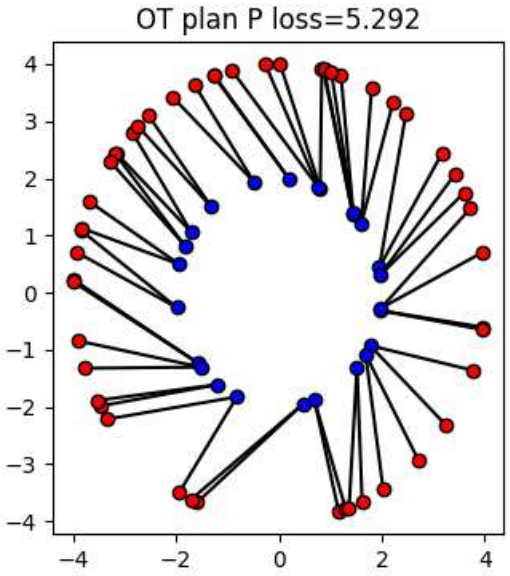


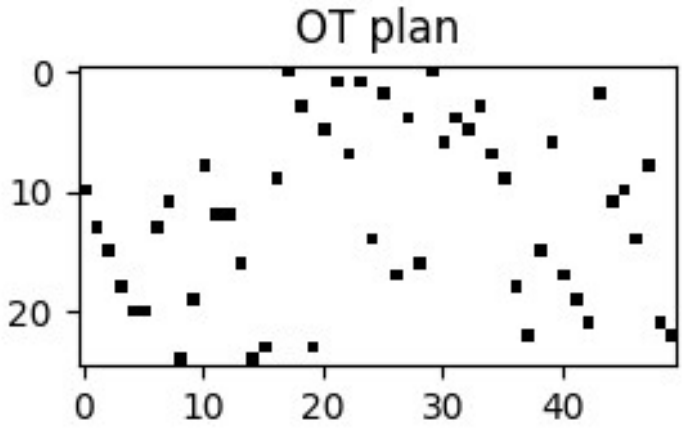
N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015





-Cost function: distance
→ User can define!





Transport plan (for optimal transport)

-KL-divergence
-Earth mover’s distance

→ How to move one distribution to another

KL-divergence (Kullback-Leibler)

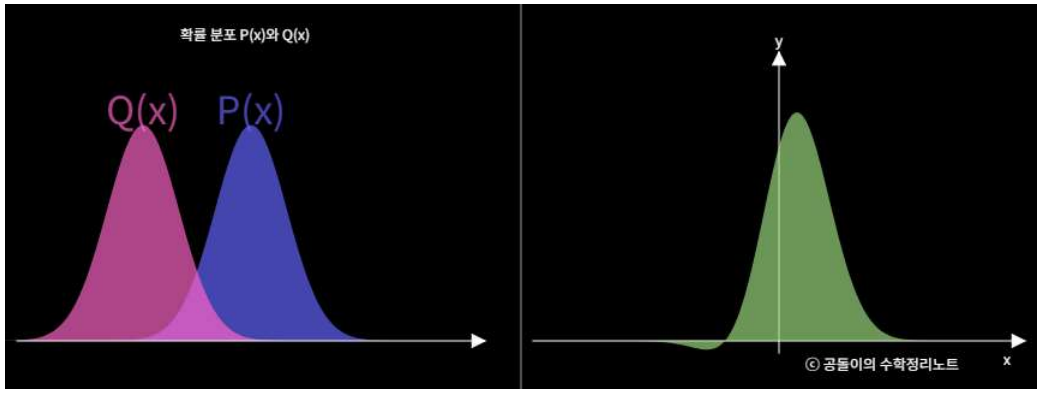
$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

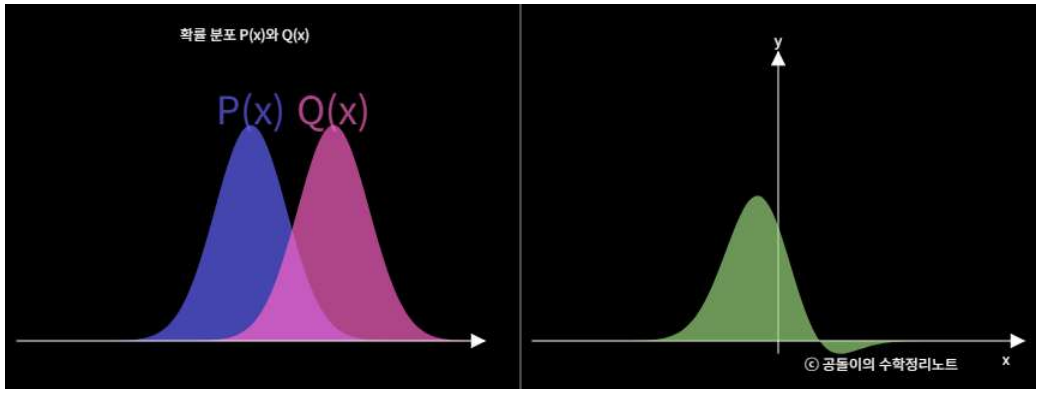
-Difference between two probability distributions

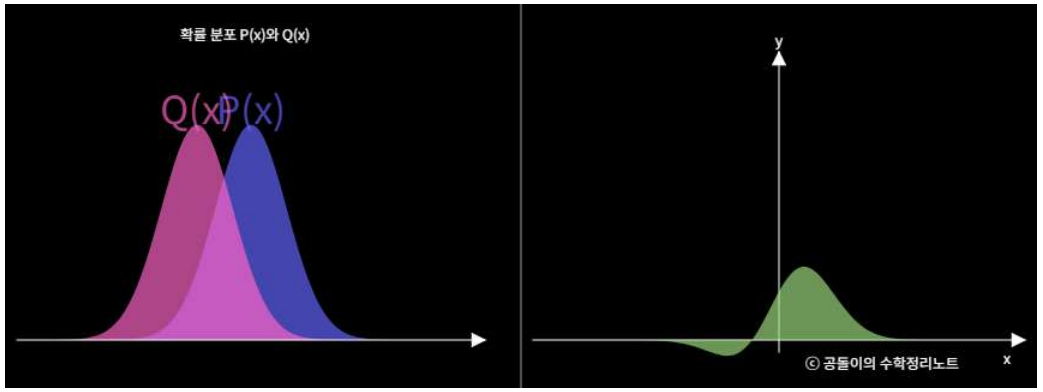
$KL(p\|q) = H(p,q) - H(p)$ -Cross entropy(p,q) – entropy(p)

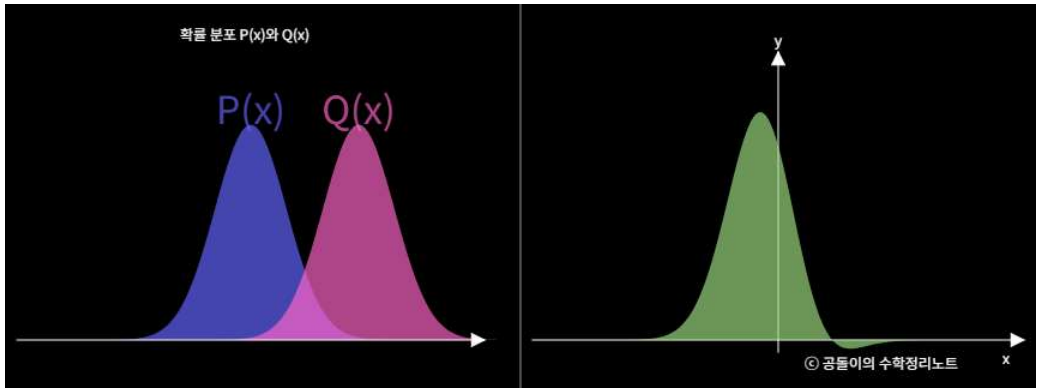
- $KL(p|q) \geq 0$
- $KL(p|q) \neq KL(q|p)$

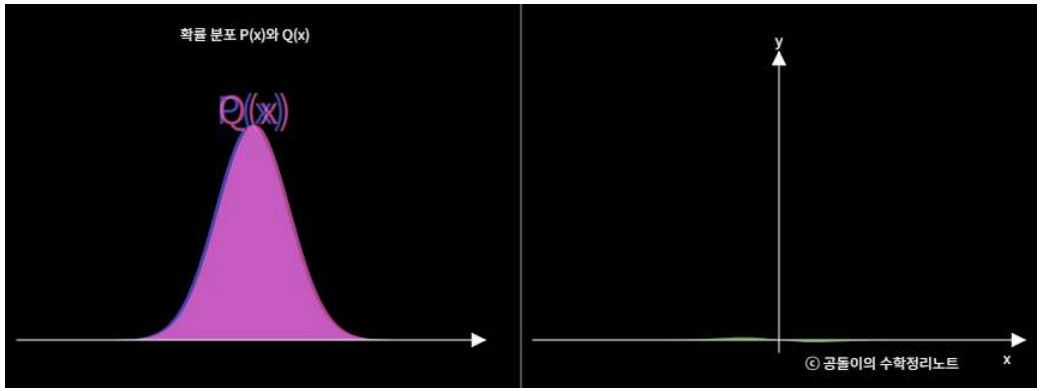
-It is not a “distance” → not symmetric











-Perfect overlap → KL = 0

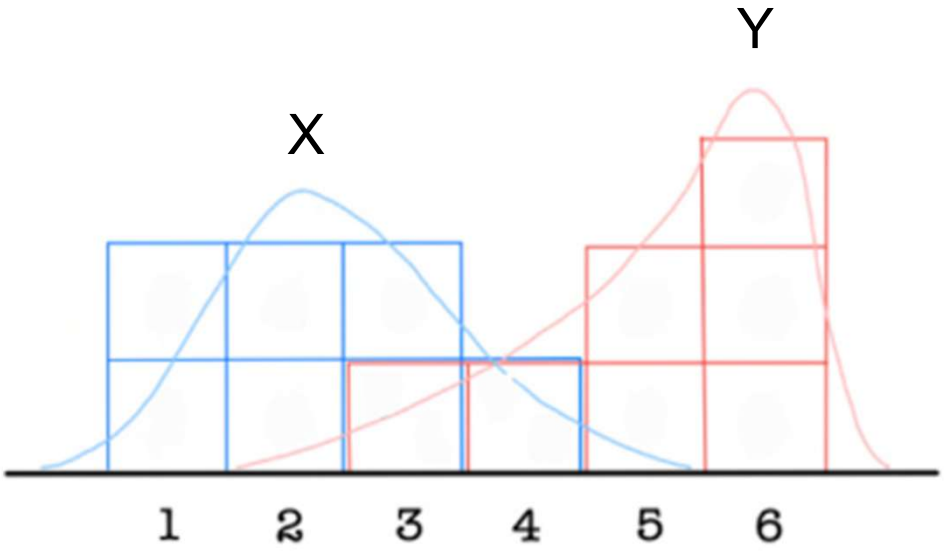
Earth mover's distance (Wasserstein Distance)

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

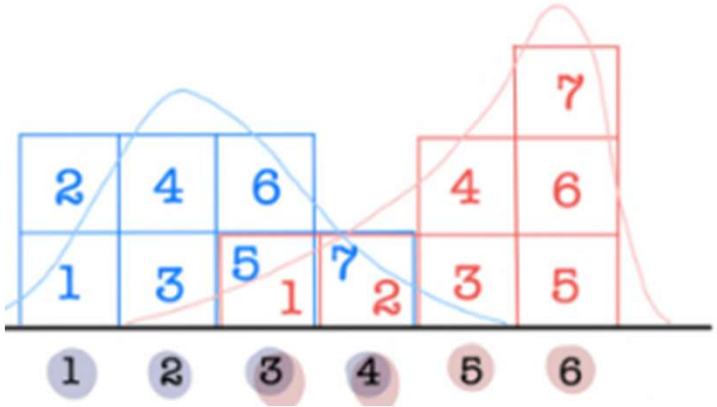
- $\Pi(P,Q)$ is the set of all joint distributions $\gamma(x,y)$ with marginals P and Q respectively. Intuitively, $\gamma(x,y)$ indicates how much 'mass' must be transported from x to y (i.e., $\|x-y\|$) in order to transform P into Q. So, $W(P,Q)$ is the cost of the optimal transport plan.

$$\mathrm{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}d_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}}$$

-γ: joint distribution

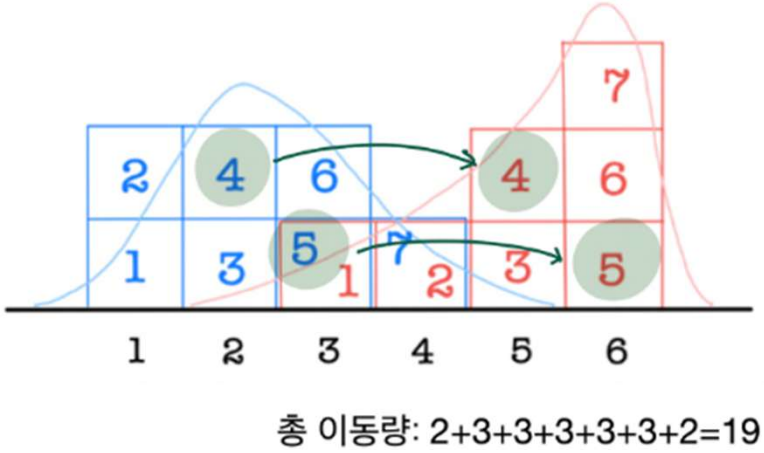


-γ distribution1 → cost:19



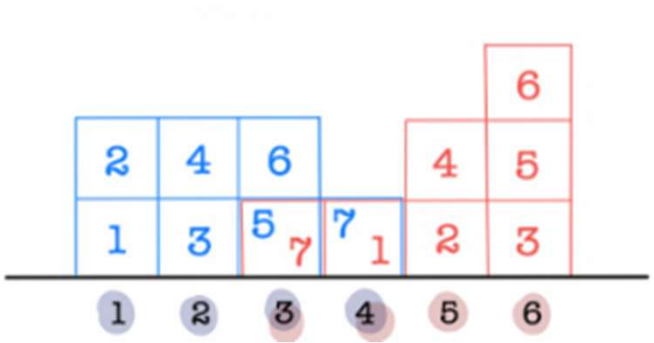
γ

x \ y	3	4	5	6
1	1	1	0	0
2	0	0	2	0
3	0	0	0	2
4	0	0	0	1



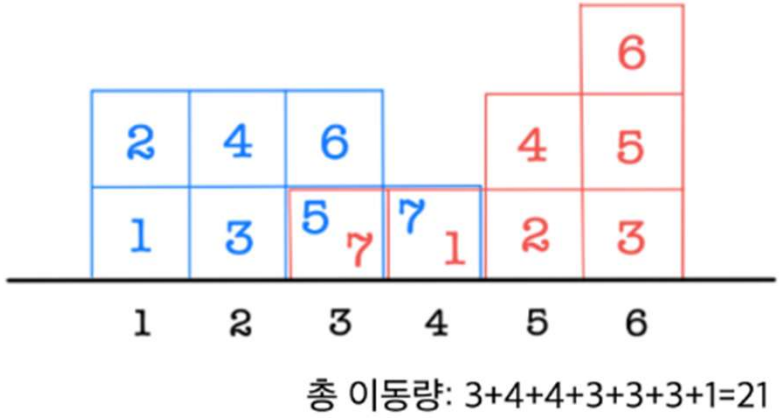
총 이동량: 2+3+3+3+3+3+2=19

-γ distribution2 → cost: 21



γ

x \ y	3	4	5	6
1	0	1	1	0
2	0	0	1	1
3	0	0	0	2
4	1	0	0	0



총 이동량: 3+4+4+3+3+3+1=21

→ Find the minimum cost transport plan

Synthetic data (SMOTE: Synthetic Minority Over-sampling Technique)

-Artificial data (does not exist in the real data) → is not same as random sampling
→ Generate from “real” distribution

1. **Find k nearest neighbors** (typically k=5) in the minority class.
2. **Pick one neighbor at random**.
3. **Create a synthetic point** along the line between the original sample and that neighbor:

$$x_{\text{new}} = x_i + \delta \cdot (x_{\text{neighbor}} - x_i)$$

where δ is a random number in $[0, 1]$.

This produces new samples that represent plausible but unseen data points.