Randomness and information in biological data
BIO-369

Prof. Anne-Florence Bitbol

EPFL

Lecture 13

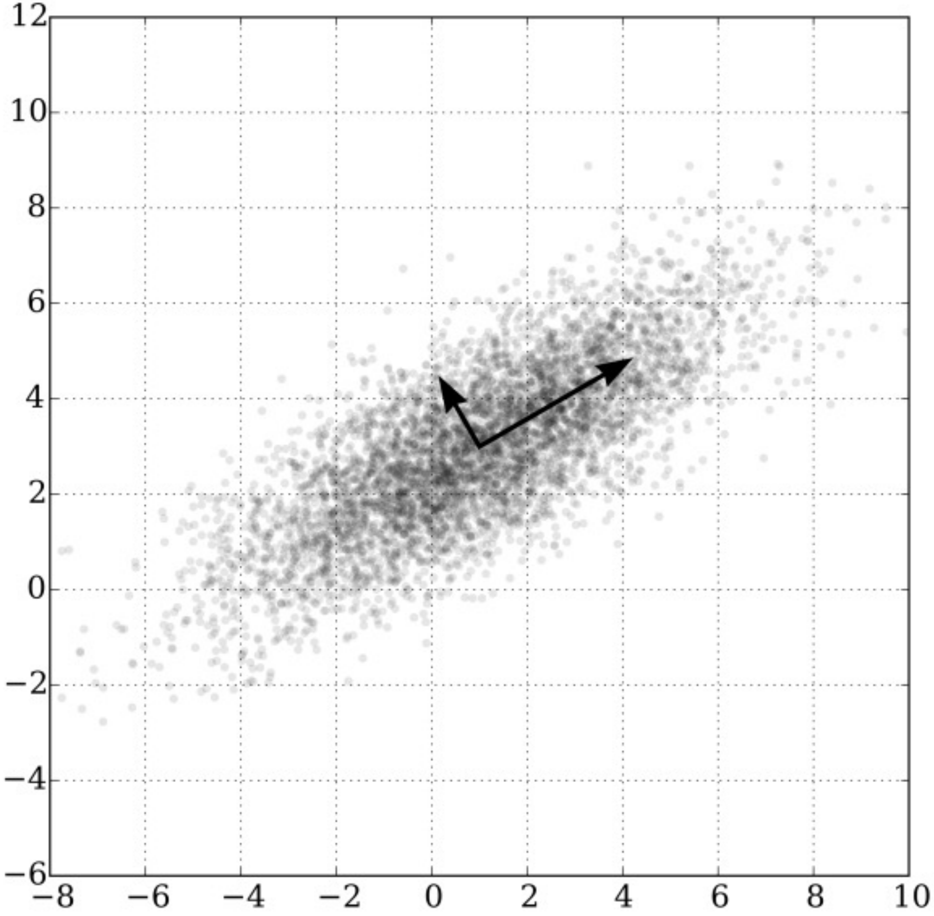
Outline of the course

II Extracting information from biological data

- 1 Quantifying randomness and information in data: entropy
  - 1.1 Notion of entropy
  - 1.2 Interpretation of entropy
  - 1.3 Entropy in neuroscience data: response of a neuron to a sensory input
- 2 Quantifying statistical dependence
  - 2.1 Covariance and correlation
  - 2.2 Mutual information
  - 2.3 Identifying coevolving sites in interacting proteins using sequence data
- 3 Inferring probability distributions from data
  - 3.1 Model selection and parameter estimation: maximum likelihood
  - 3.2 Introduction to maximum entropy inference
  - 3.3 Predicting protein structure from sequence data
- 4 Finding relevant dimensions in data: dimension reduction
  - 4.1 Principal component analysis
  - 4.2 Beyond principal component analysis

Principal component analysis

- **Two-dimensional data**



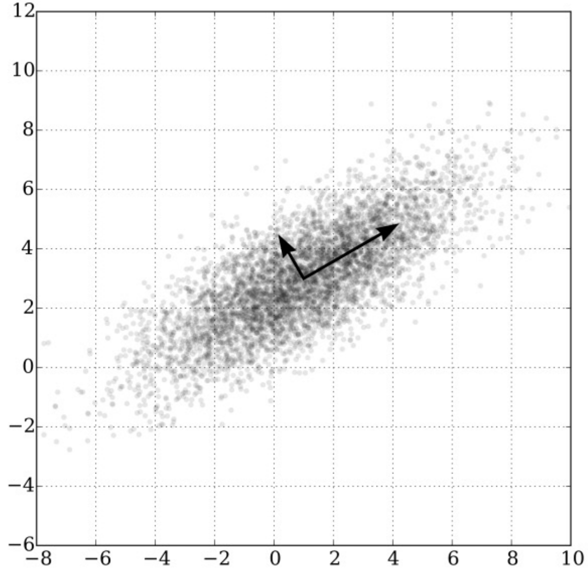
If you had to reduce this data to one dimension, what direction would you choose to project the data onto?

0% A. The x axis

0% B. The y axis

0% C. The arrow pointing to the right

0% D. The arrow pointing to the left

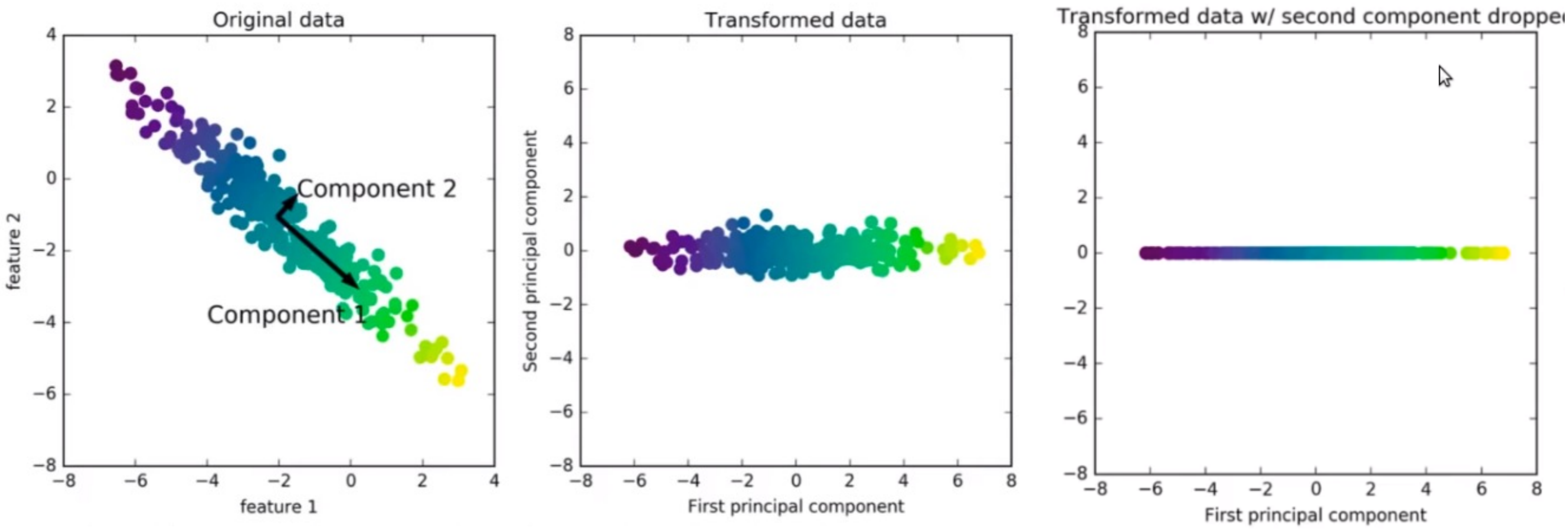


To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Principal component analysis

- **Two-dimensional data**

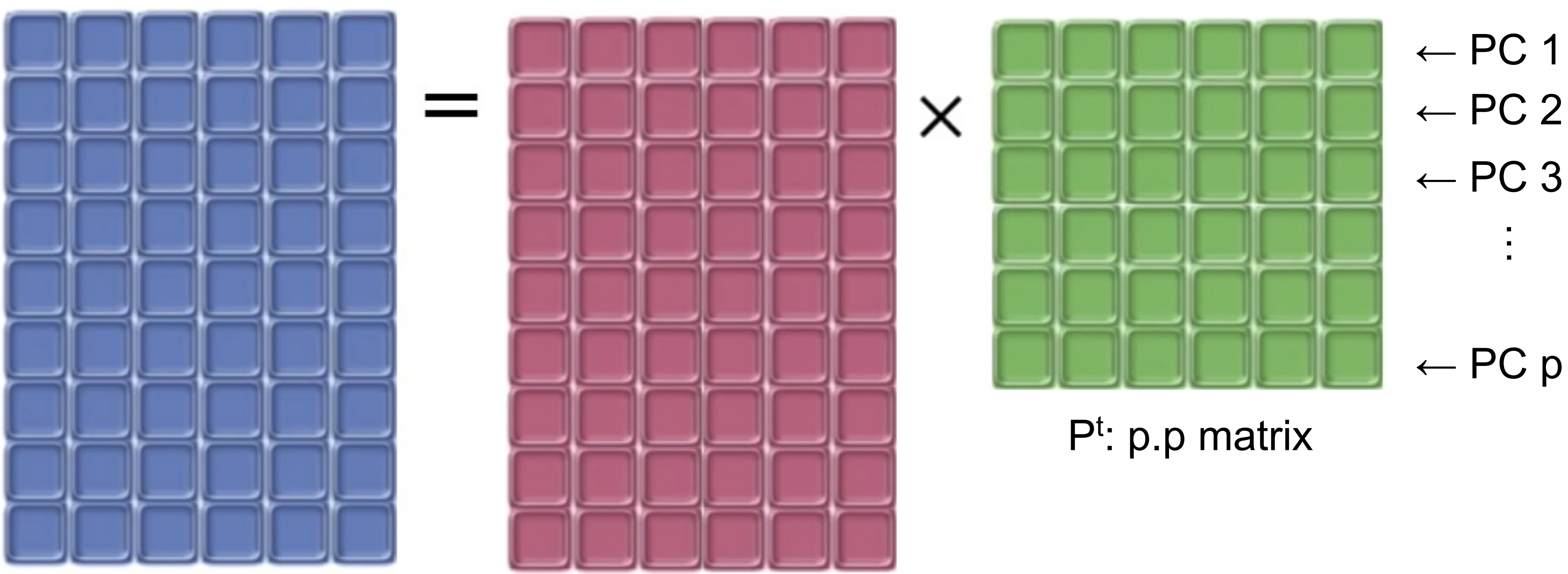


Principal component analysis

- **Matrix factorization**

X: n.p data matrix
Each column = a feature; p(=6) features
Each row = a measurement; n(=9) measurements

Change basis:
$Y = X P \quad \Rightarrow \quad X = Y P^t$



X: n.p matrix

Y: n.p matrix

P^t: p.p matrix

To make a d-dimension approximation of the data $X = Y P^t$, should we focus on:

A. The first d rows of P^t
B. The first d columns of P^t
C. The first d rows of Y
D. The first d columns of Y
E. The first d rows of P^t and the first d columns of Y
F. The first d columns of P^t and the first d rows of Y

To answer, please:

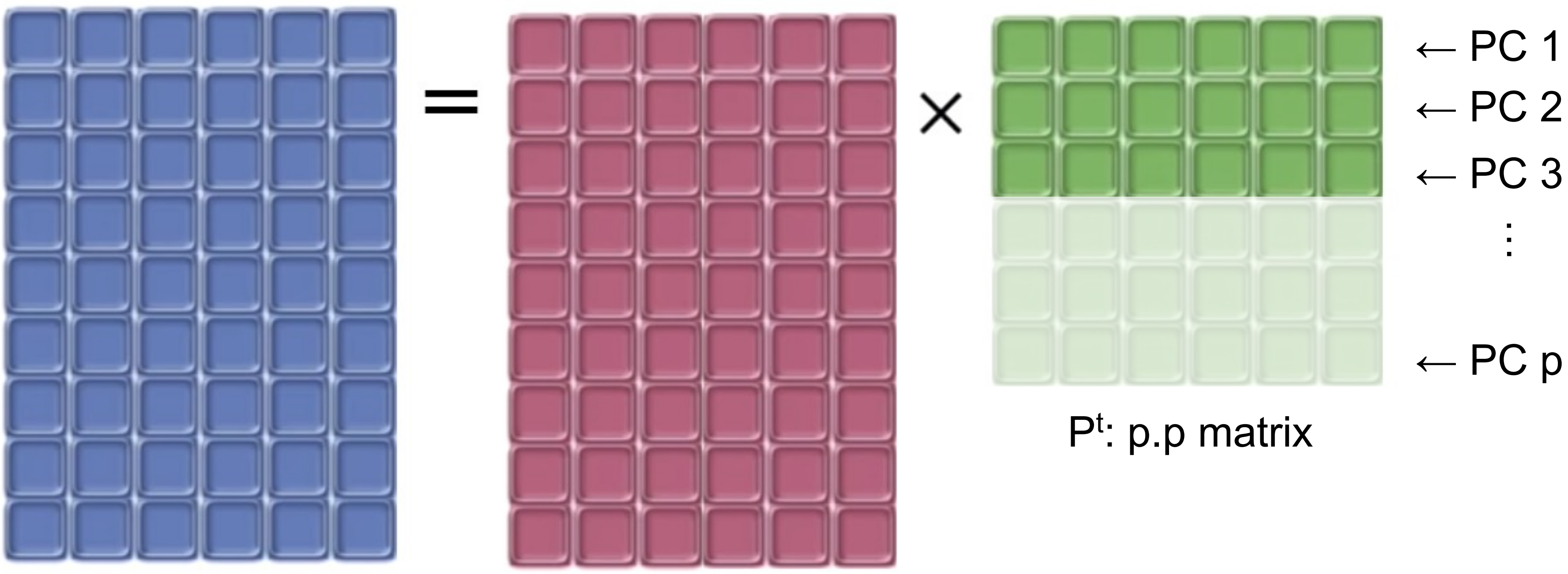
- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Principal component analysis

- **Matrix factorization**

X: n.p data matrix
Each column = a feature; p(=6) features
Each row = a measurement; n(=9) measurements

Change basis:
$Y = X P \quad \Rightarrow \quad X = Y P^t$



X: n.p matrix

Y: n.p matrix

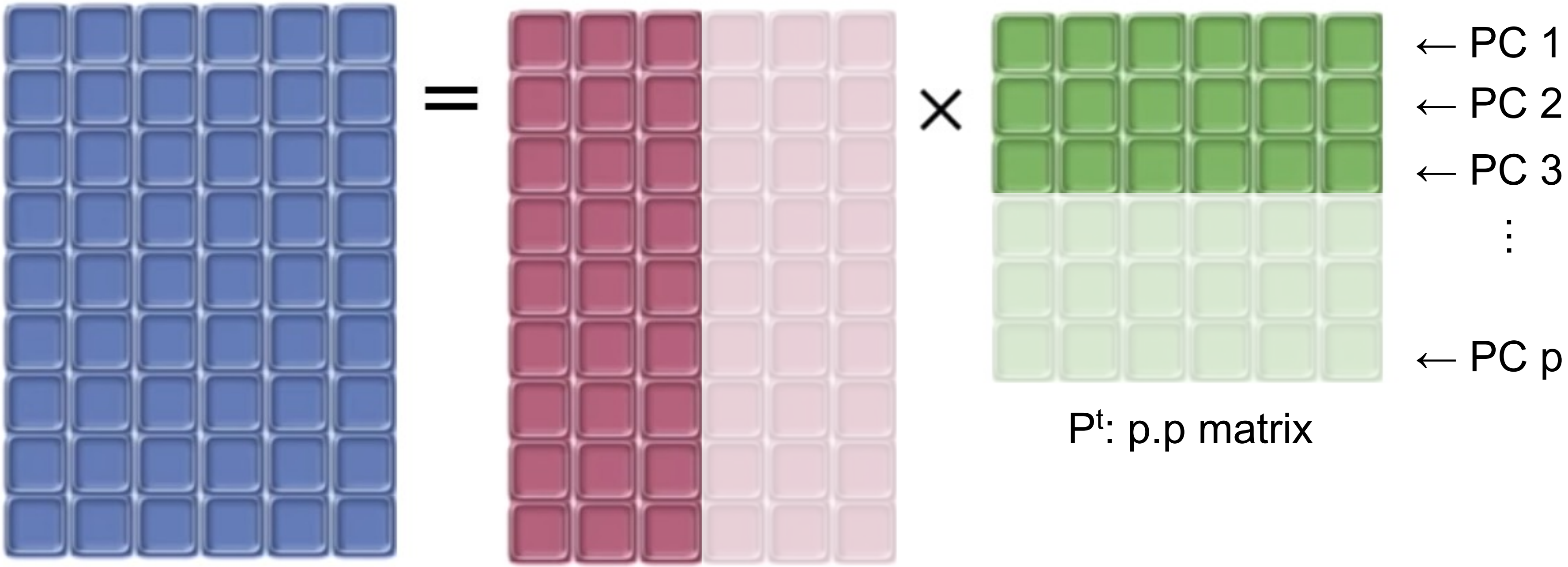
P^t: p.p matrix

Principal component analysis

Matrix factorization

X: n.p data matrix
Each column = a feature; p(=6) features
Each row = a measurement; n(=9) measurements

Change basis:
$Y = X P \quad => \quad X = Y P^t$



X: n.p matrix

Y: n.p matrix

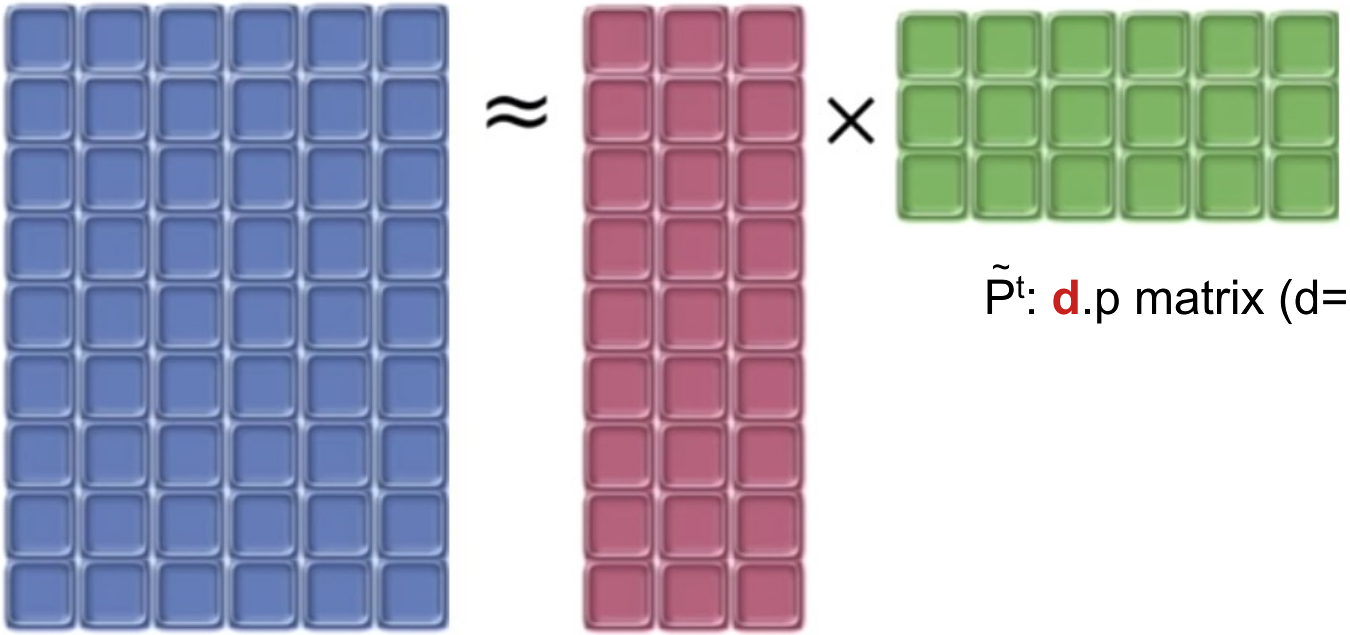
P^t: p.p matrix

Principal component analysis

- **Matrix factorization**

X: n.p data matrix
Each column = a feature; p(=6) features
Each row = a measurement; n(=9) measurements

Change basis:
$Y = X P \quad \Rightarrow \quad X = Y P^t \approx \tilde{Y} \tilde{P}^t$



← PC 1
← PC 2
← PC 3

$\tilde{P}^t$: **d**.p matrix (d=3)

X: n.p matrix

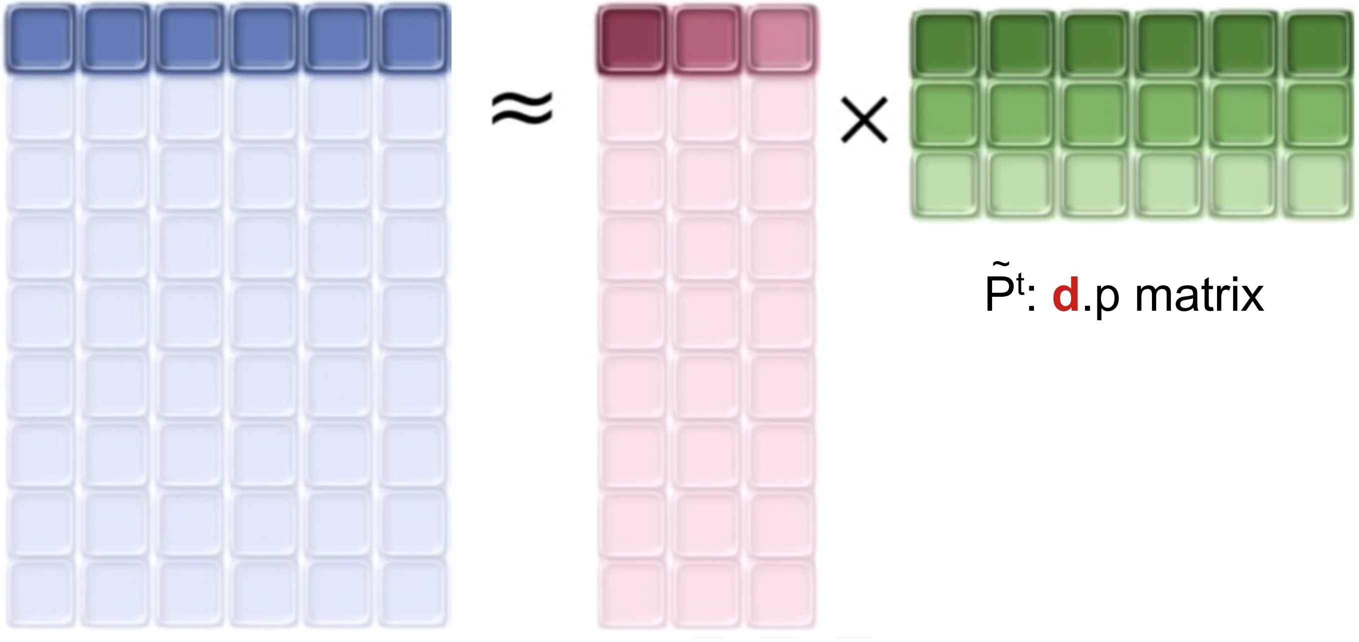
$\tilde{Y}$: n.**d** matrix: coordinates of the data on the **d** (=3) top PCs

Principal component analysis

Matrix factorization

X: n.p data matrix
Each column = a feature; p(=6) features
Each row = a measurement; n(=9) measurements

Change basis:
$Y = X P \quad => \quad X = Y P^t \approx \tilde{Y} \tilde{P}^t$



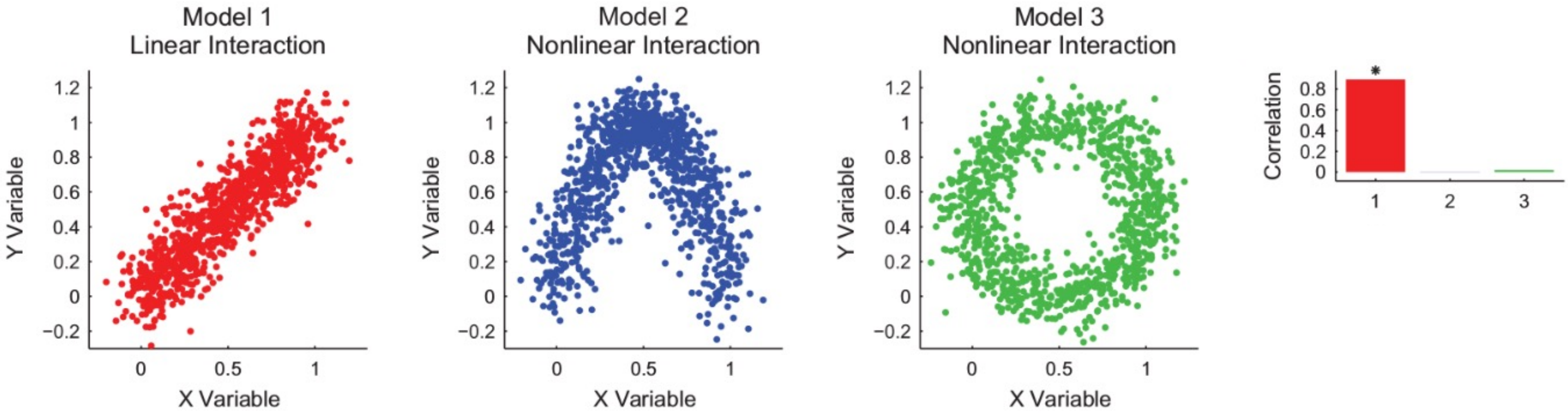
← PC 1
← PC 2
← PC 3

$\tilde{P}^t$: **d**.p matrix

X: n.p matrix

$\tilde{Y}$: n.**d** matrix : coordinates of the data on the **d** (=3) top PCs

Reminder



Correlation between random variables X and Y
Different draws are performed, yielding values x and y, and the correlation and mutual information are estimated

Some nonlinear forms of statistical dependence are missed by correlation

Do you think PCA can well reduce to one dimension the data:

0% A. Of no model

0% B. Of model 1 but not others

0% C. Of models 1 and 2 but not 3

0% D. Of models 1 and 3 but not 2

0% E. Of models 2 and 3 but not 1

0% F. Of all models

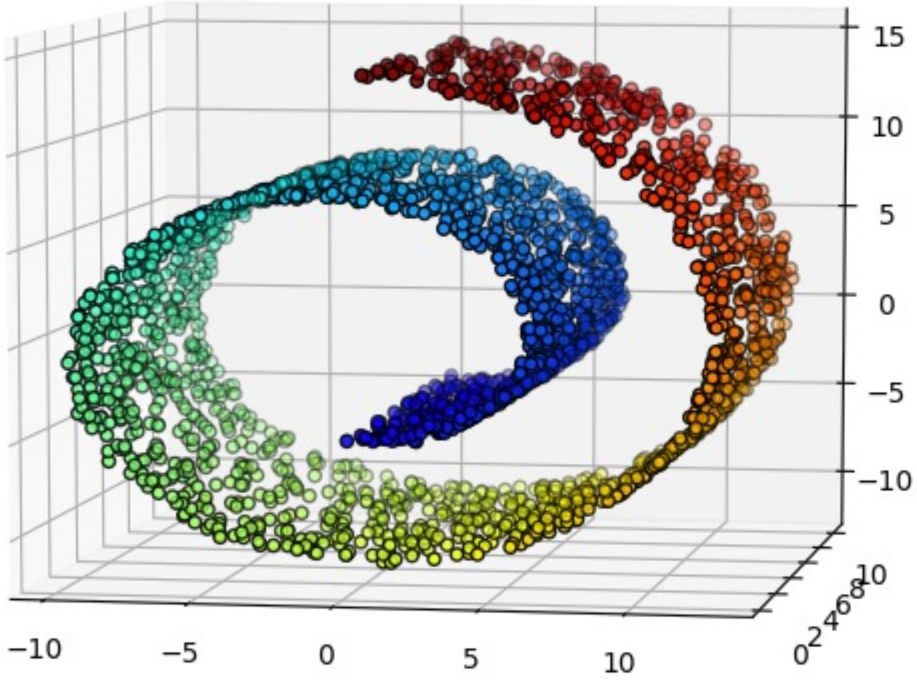
To answer, please:

- Connect to http://ttpoll.eu
- Enter the session ID **bio369**
- Select your answer

Swiss roll

- **Applying PCA to the Swiss roll dataset**

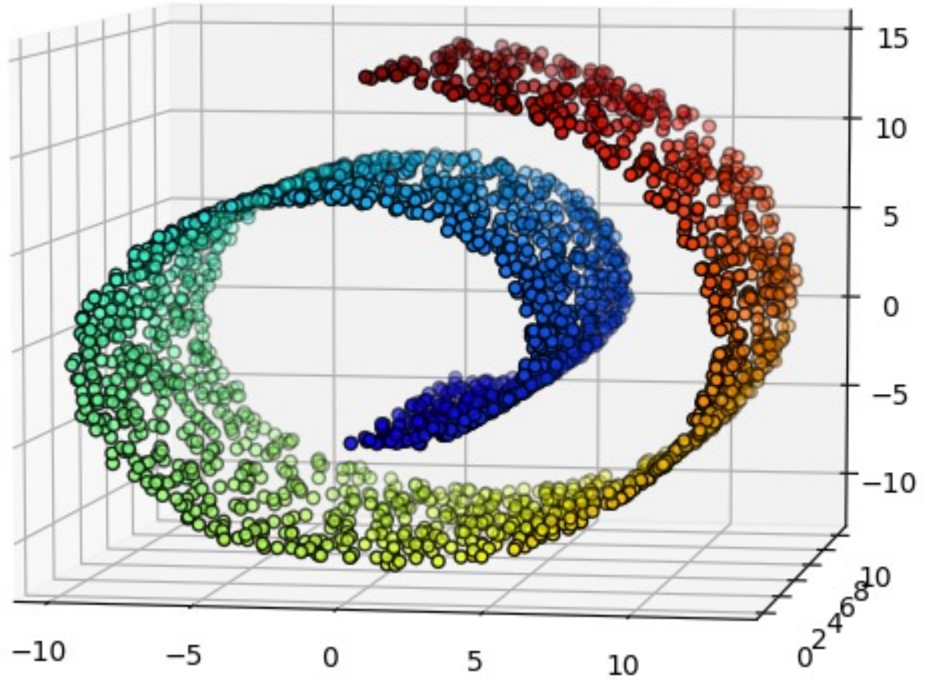
2D surface rolled in 3D space



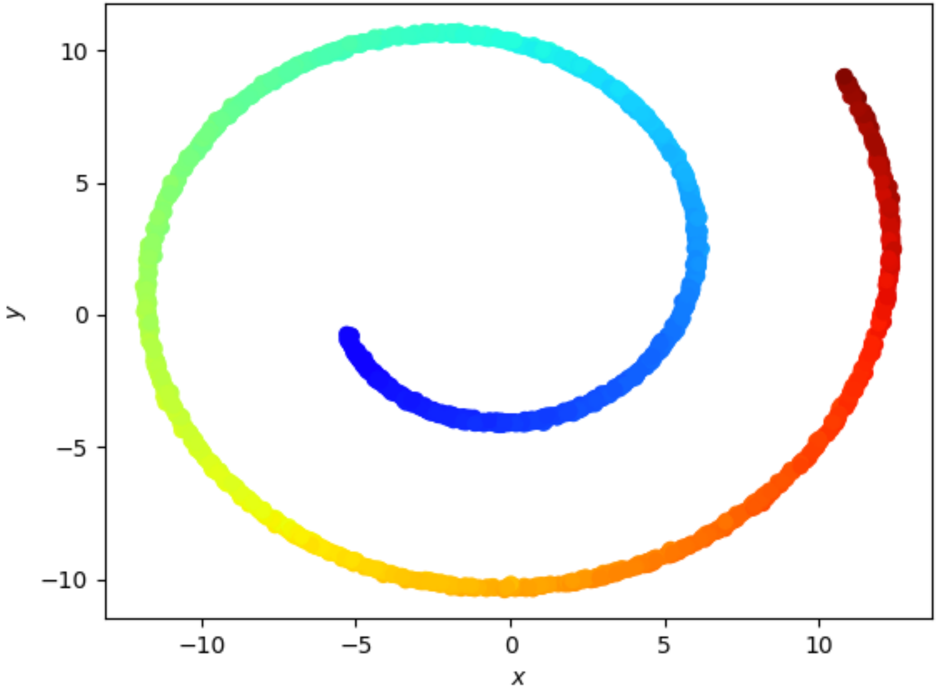
Swiss roll

- **Applying PCA to the Swiss roll dataset**

2D surface rolled in 3D space



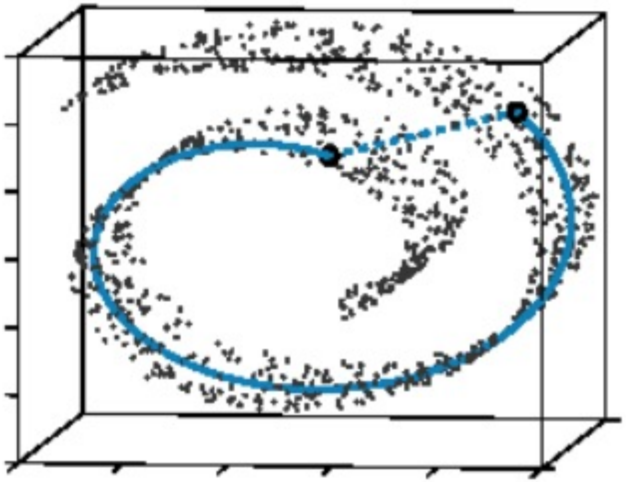
PCA → 2 top PCs:



Swiss roll

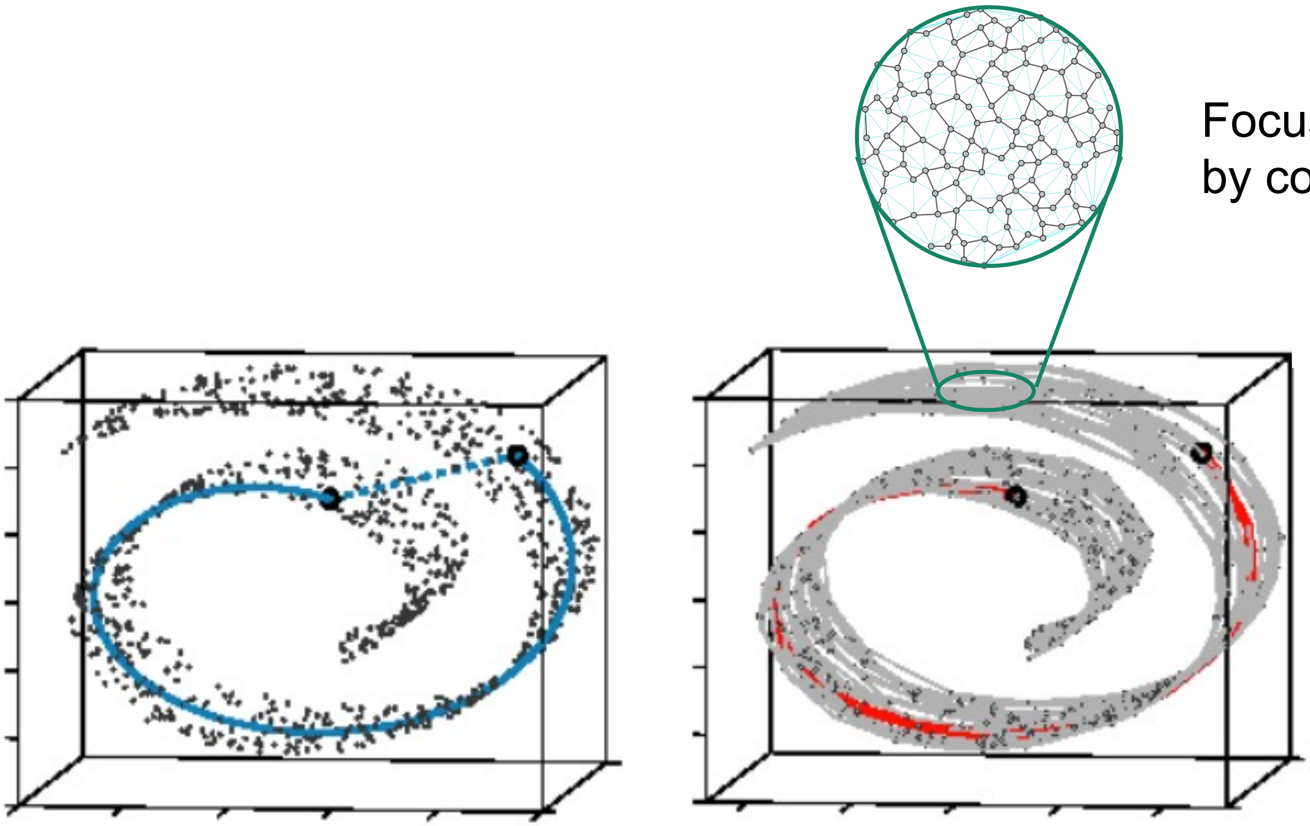
- **Nonlinear dimension reduction methods based on neighbor graphs**

Shortest distance
vs. shortest distance along
the 2D structure



Swiss roll

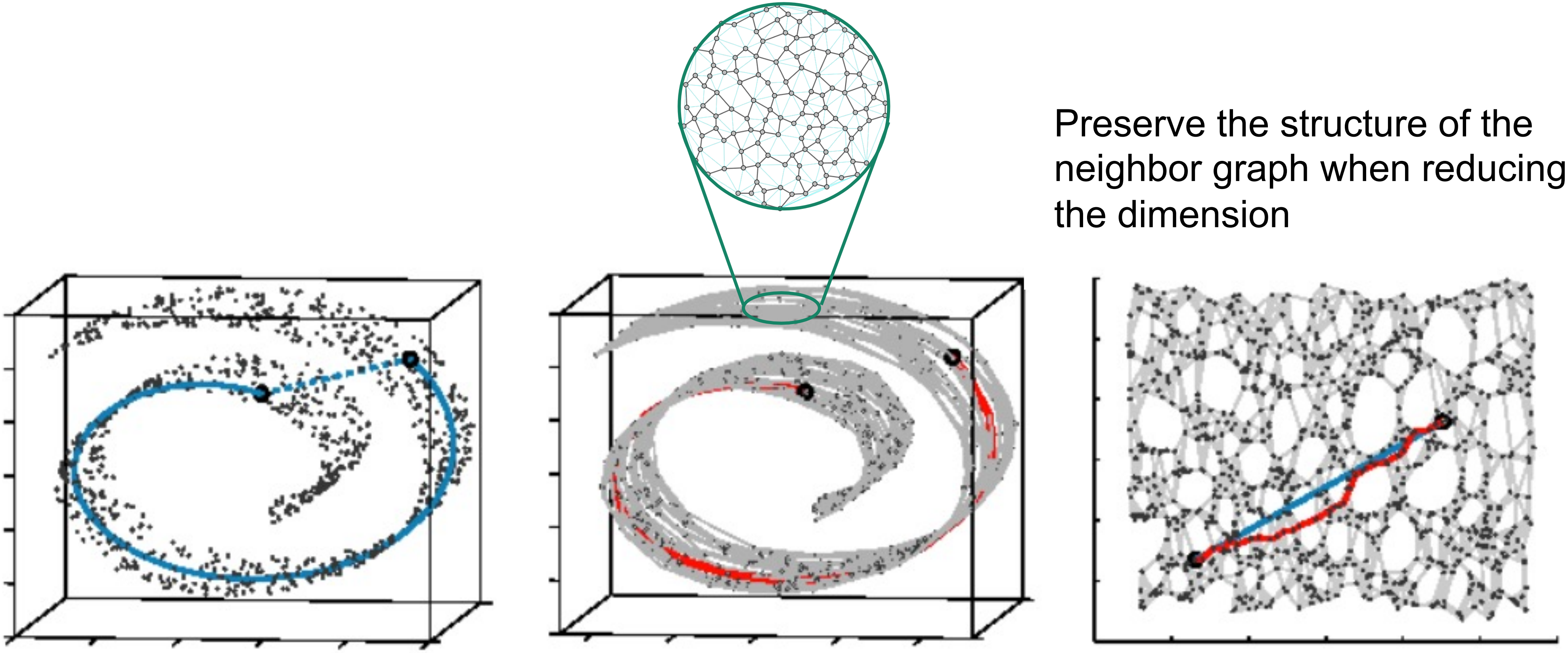
- **Nonlinear dimension reduction methods based on neighbor graphs**



Focus on (k) nearest neighbors
by constructing a neighbor graph

Swiss roll

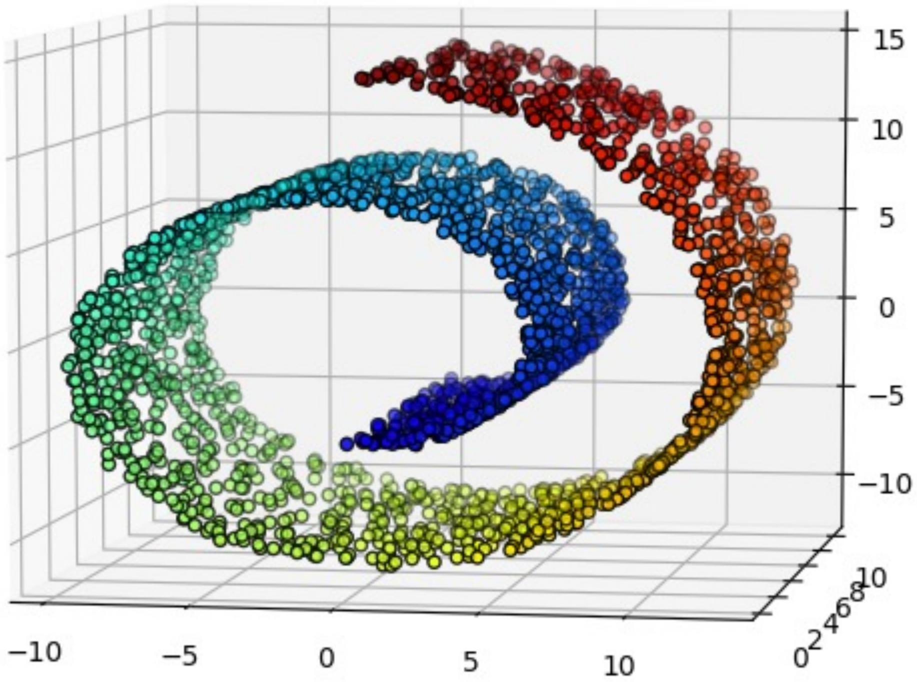
- **Nonlinear dimension reduction methods based on neighbor graphs**

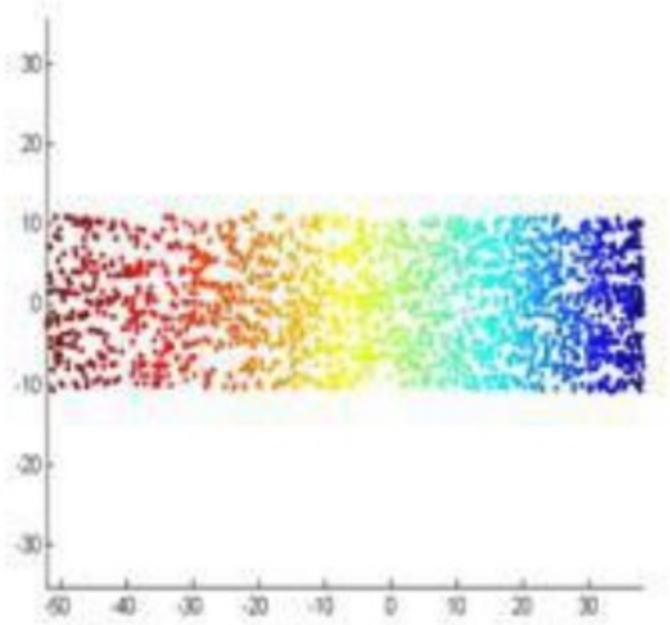


Preserve the structure of the neighbor graph when reducing the dimension

Swiss roll

- **Nonlinear dimension reduction methods based on neighbor graphs**

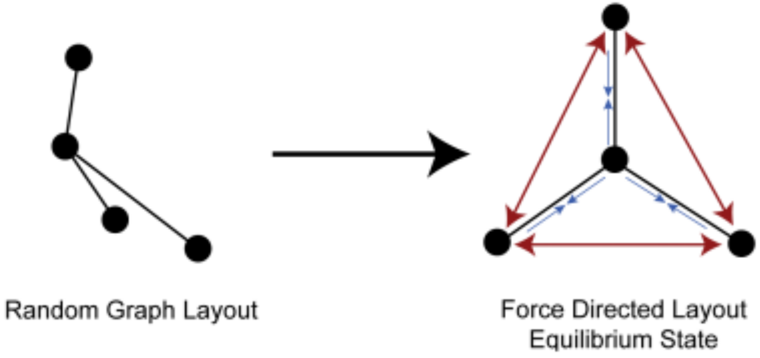




Preserve the structure of the neighbor graph when reducing the dimension

How is this transformation done in practice?

- **Force directed graph layout (t-SNE, UMAP)**



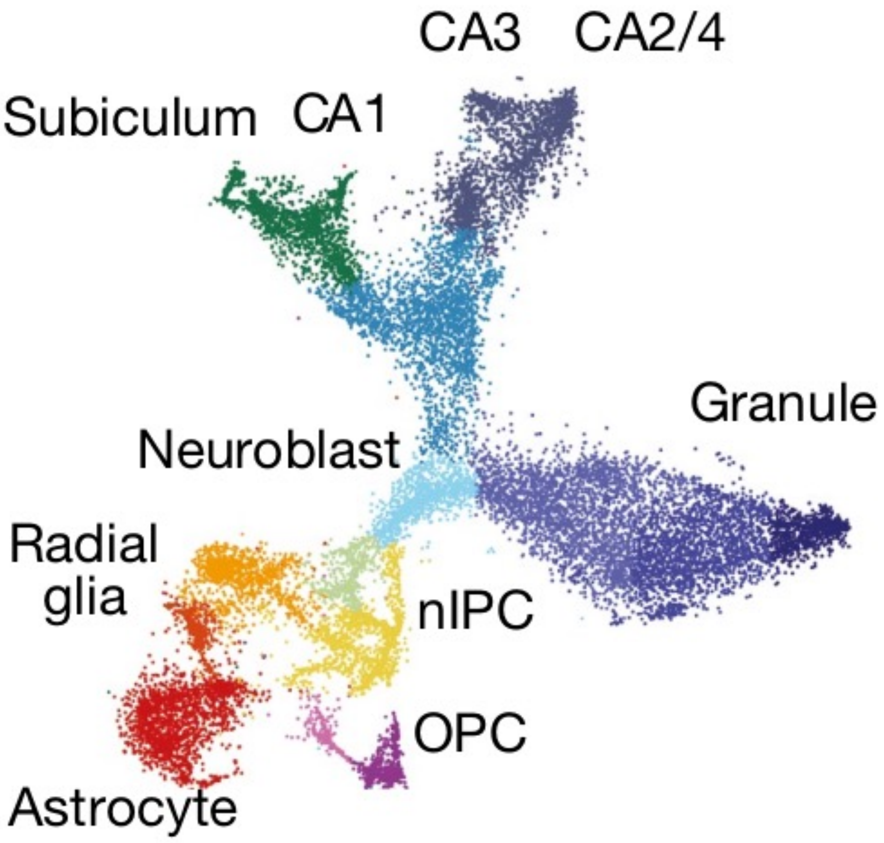
Attraction along the edges (springs)
Repulsion between vertices otherwise

→ Preserve the structure of the neighbor graph when reducing the dimension

Applications

- **Single-cell RNA sequencing (sc-RNA) and cell types**

La Manno et al 2018

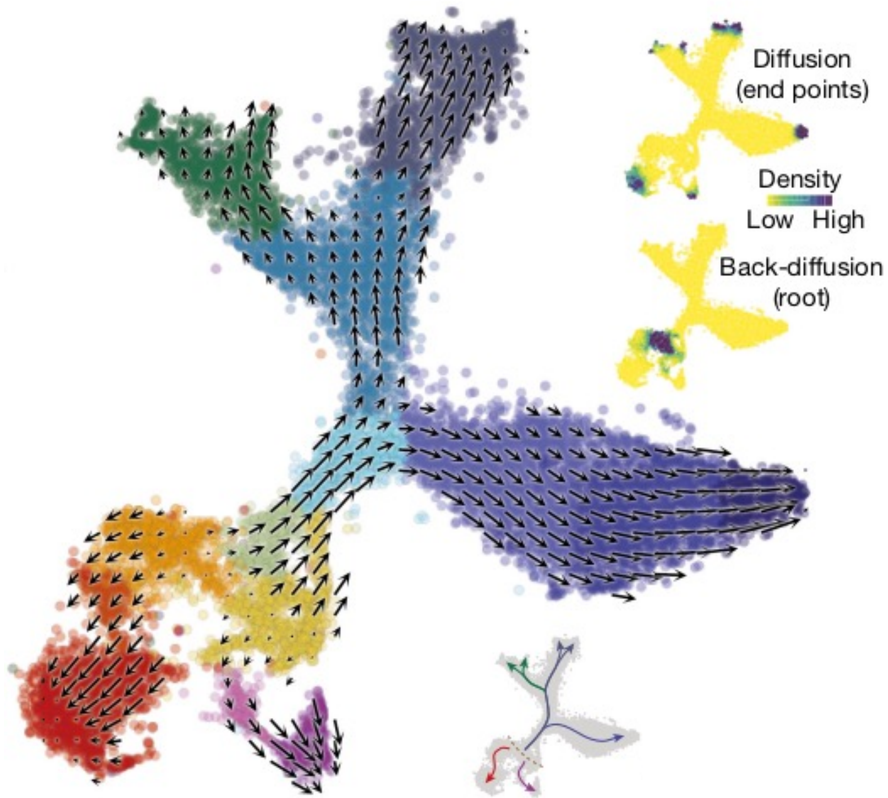


t-SNE plot of developing mouse hippocampus cells (18,213 cells), showing major transient and mature subpopulations

Applications

- **RNA velocity and cell differentiation trajectories**

La Manno et al 2018



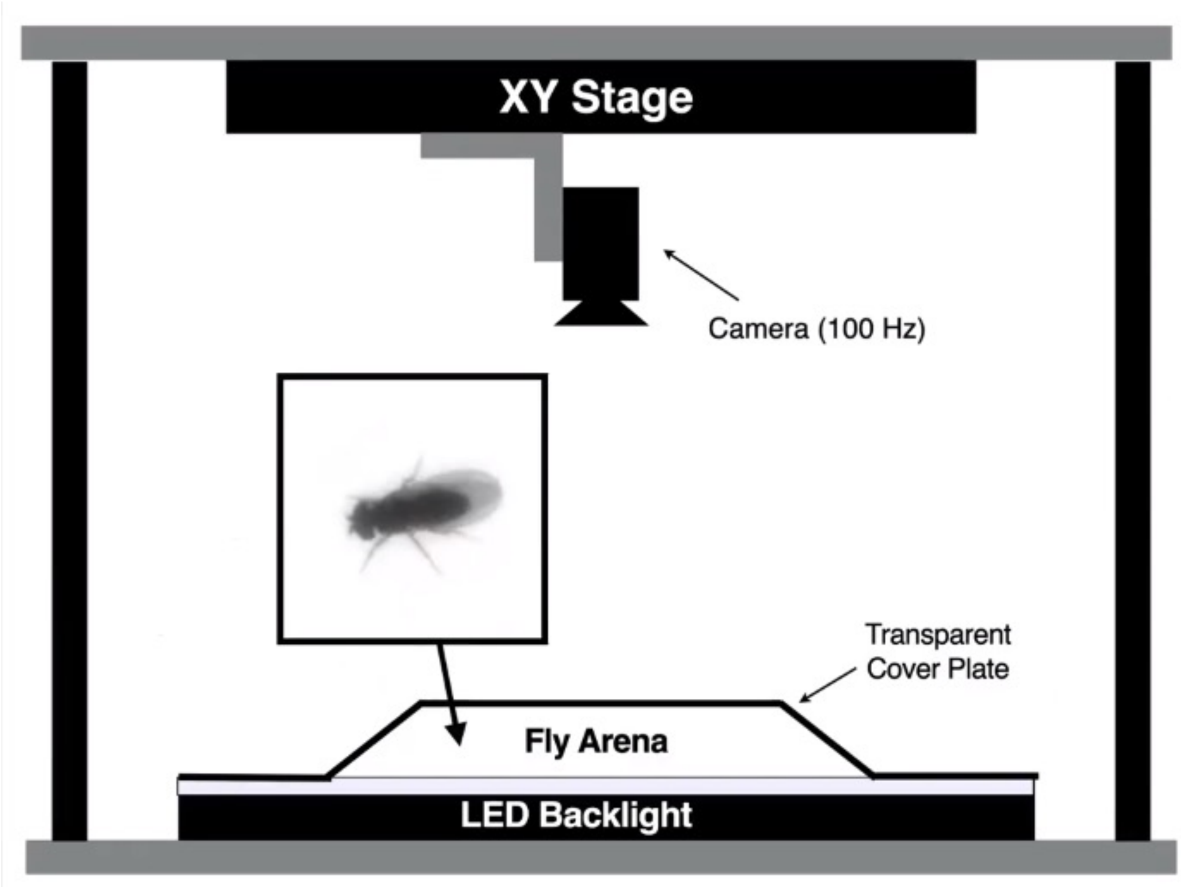
Unspliced vs. spliced RNA → time evolution of gene expression: RNA velocity

Velocity field (arrows) projected onto the t-SNE plot
Top inset, differentiation endpoints (mature cell types) and root (progenitor cells)
Bottom inset, summary schematic of the RNA velocity field

Applications

- **Animal behavior**

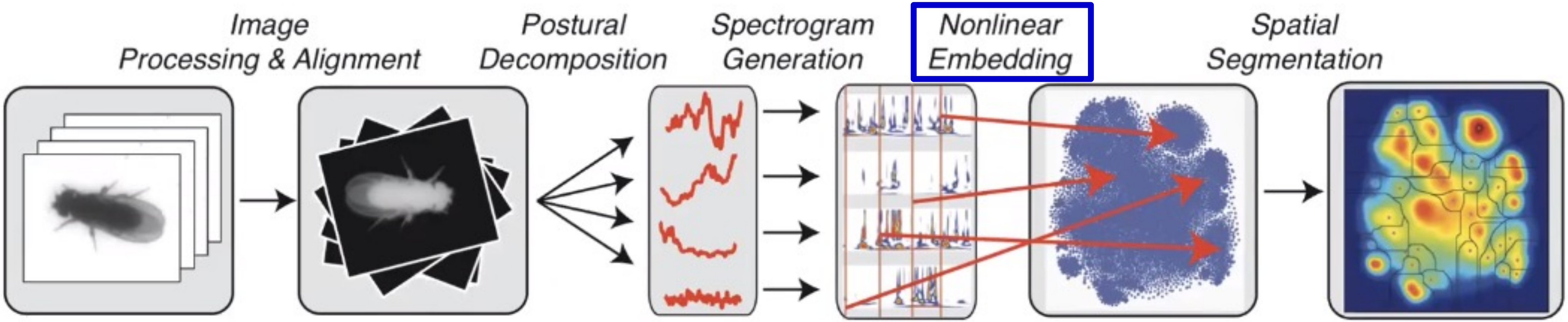
Berman et al 2014



Applications

- **Animal behavior**

Berman et al 2014



https://mediaspace.epfl.ch/media/Gordon%20Berman%2C%20Emory%20University%20-%20Measuring%20the%20hidden%20dynamics%20of%20animal%20behavior/0_opd7gmv3
21:10 → 21:55

Announcements

- Next week: exam preparation & review session
    - Problem class on May 26 at 3:15pm: working on the mock exam
    - Lecture on May 28 at 10:15pm: review session

- Extra problem session during exam preparation period – week of June 3, please answer the poll at https://forms.gle/CotgTneH5krGk6kA8

- Please fill in the **in-depth evaluation of the class** (closes on June 8 at midnight) – you should have received an email about it

Thank you very much!