

# Foundations of multidimensional data visualization

Serhiy Naumenko,  
<https://genomics.org.ua>

Унаочнювання даних

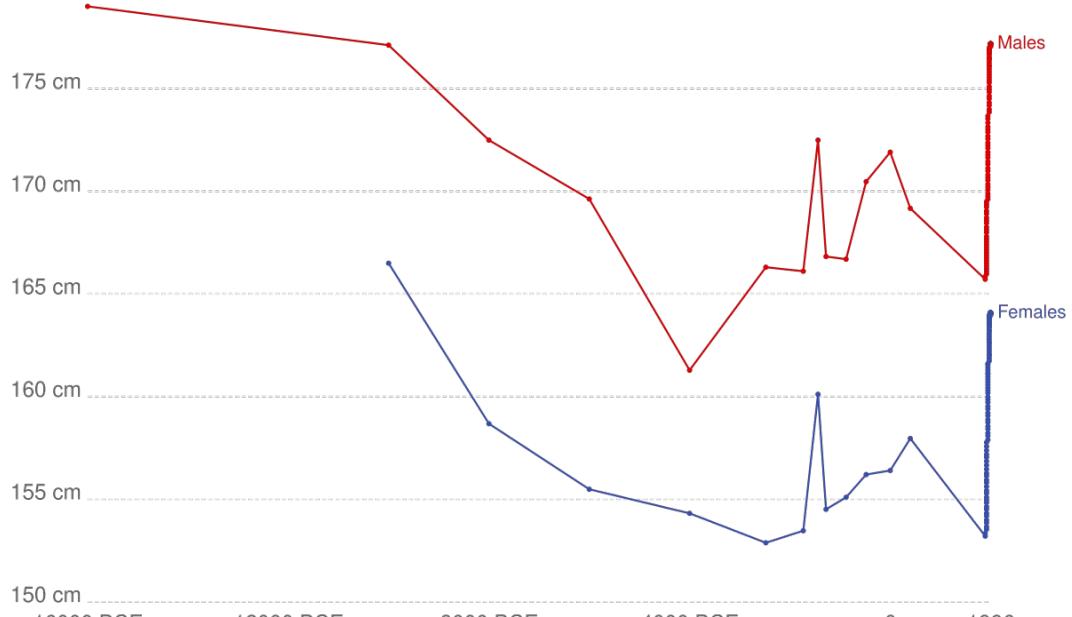
AI alert!

# Чи всі дані в біології є багатовимірні?

Time series = часові ряди

## Human heights over the long-run

Average human height in the Eastern Mediterranean from the Upper Paleolithic (before 16,000 BC) period, through to 1996.

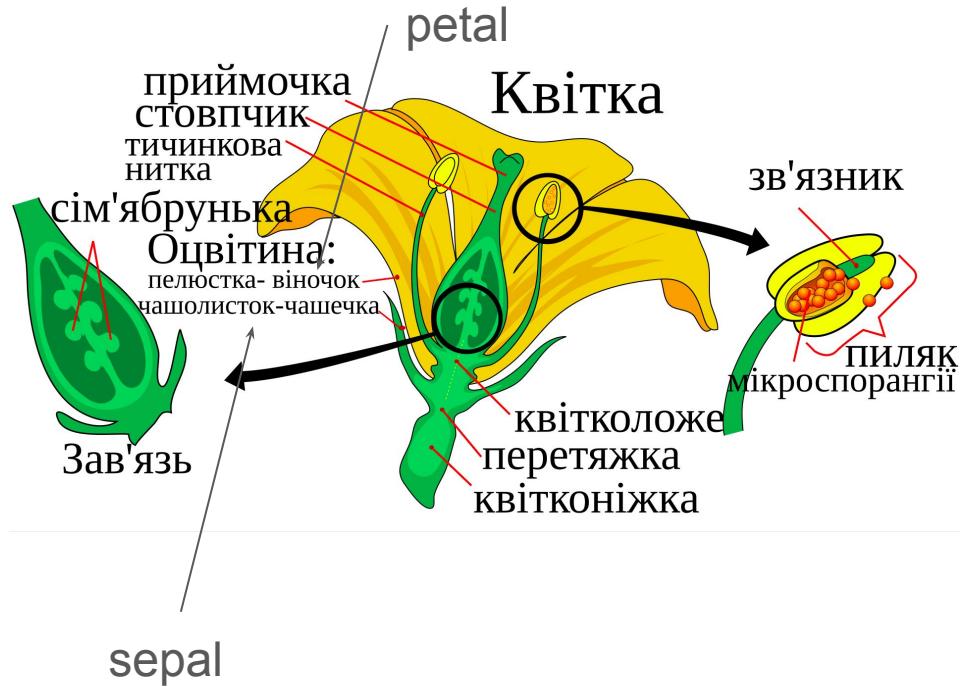


Source: Hermanussen (2003) and NCD RisC, Human Height (2017)

OurWorldInData.org/human-height/ • CC BY

[https://en.wikipedia.org/wiki/Human\\_height](https://en.wikipedia.org/wiki/Human_height)

# Iris - півник, ірис, R: iris



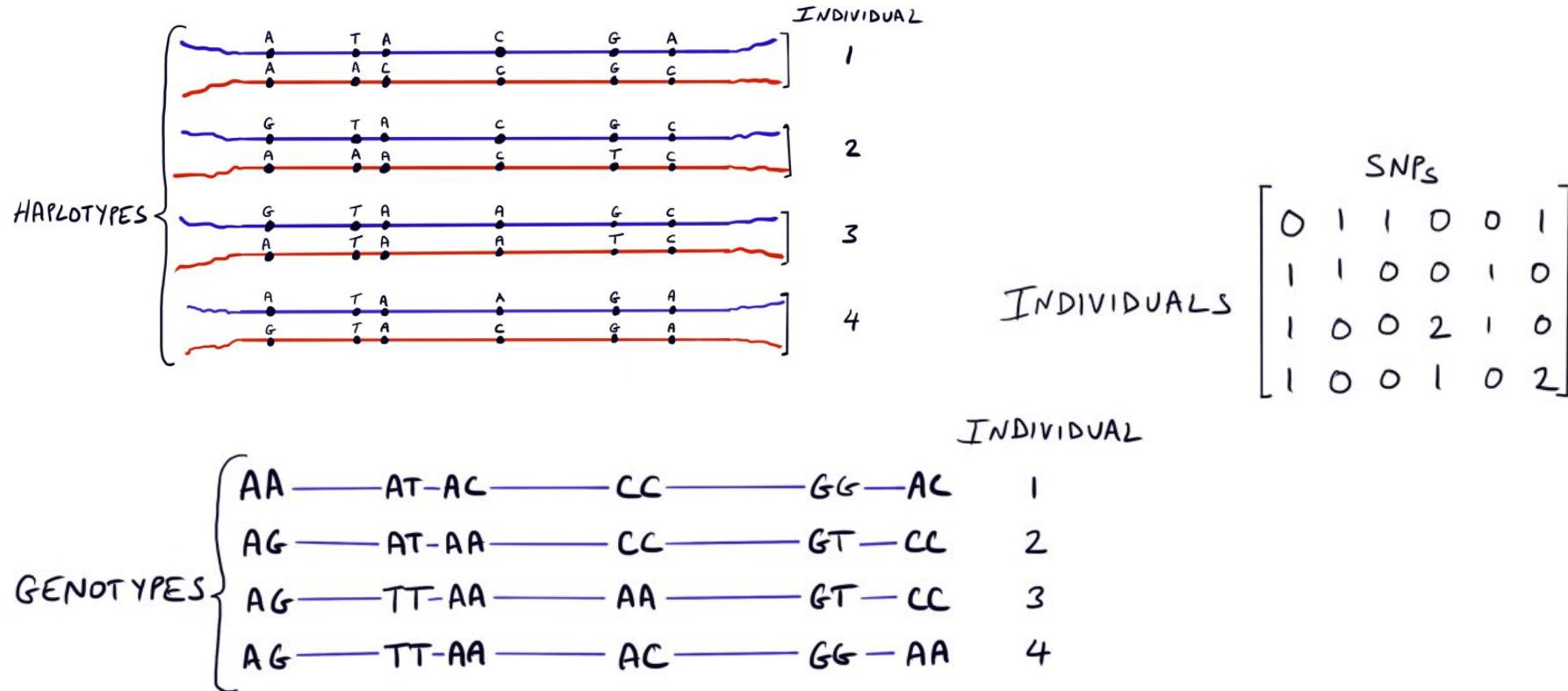
*Iris setosa*



[https://en.wikipedia.org/wiki/Iris\\_flower\\_data\\_set](https://en.wikipedia.org/wiki/Iris_flower_data_set)

<https://uk.wikipedia.org/wiki/%D0%A7%D0%B0%D1%88%D0%BE%D0%BB%D0%B8%D1%81%D1%82%D0%BE%D0%BA>

# Генотипи в популяції



# Матриця експресії

Note ML T lingo!  
Note Single Cell x 100,000 cells!

observations

	Зразок1-е	Зразок2-е	Зразок3-е	Зразок4-к	Зразок5-б	Зразок6-к
Ген1	100	105	110	50	55	54
Ген2	1,000	1,109	900	1,200	1,000	800
Ген3	3,000	3,300	3,100	5,000	4,000	4,300

features

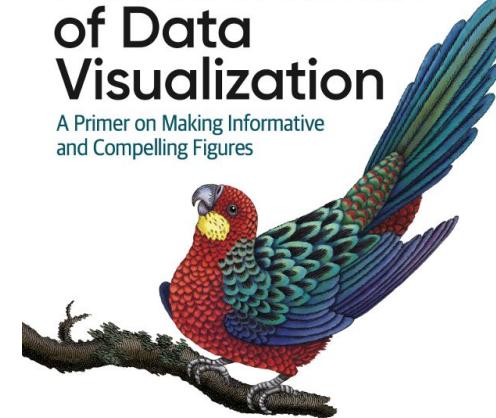
Значення експресії

Які гени (транскрипти, екзони) експресуються сильніше (слабкіше) у експерименті порівняно із контролем?

На жаль, ми можемо сприймати  
тільки 2D графіки

# Fundamentals of Data Visualization

A Primer on Making Informative  
and Compelling Figures



Claus O. Wilke

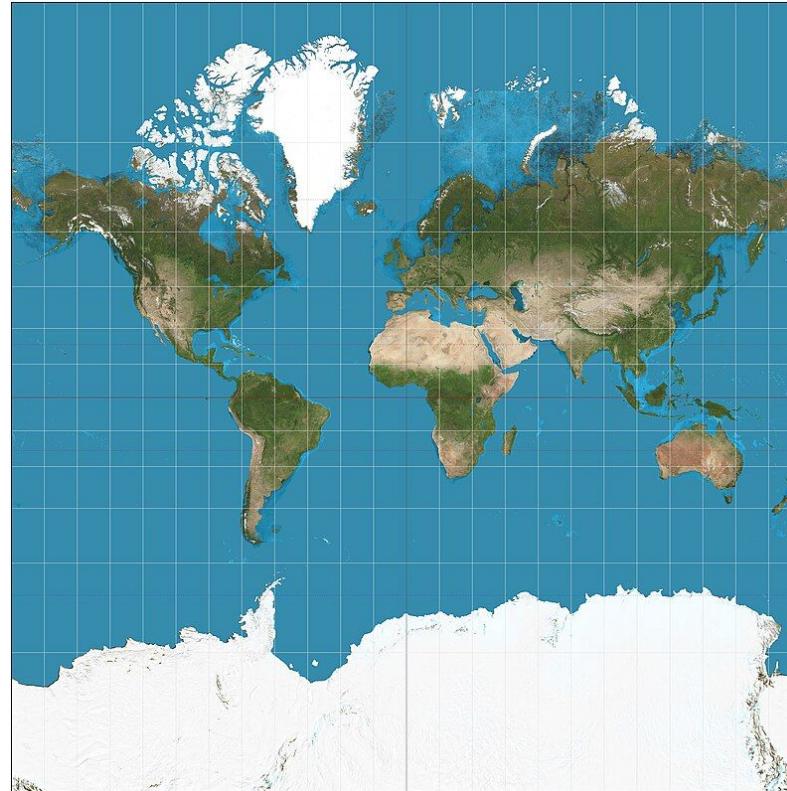
## 26 Don't go 3D

3D plots are quite popular, in particular in business presentations but also among academics. They are also almost always inappropriately used. It is rare that I see a 3D plot that couldn't be improved by turning it into a regular 2D figure. In this chapter, I will explain why 3D plots have problems, why they generally are not needed, and in what limited circumstances 3D plots may be appropriate.

# Тож необхідно проєктувати в 2D, а чи взагалі це працює?

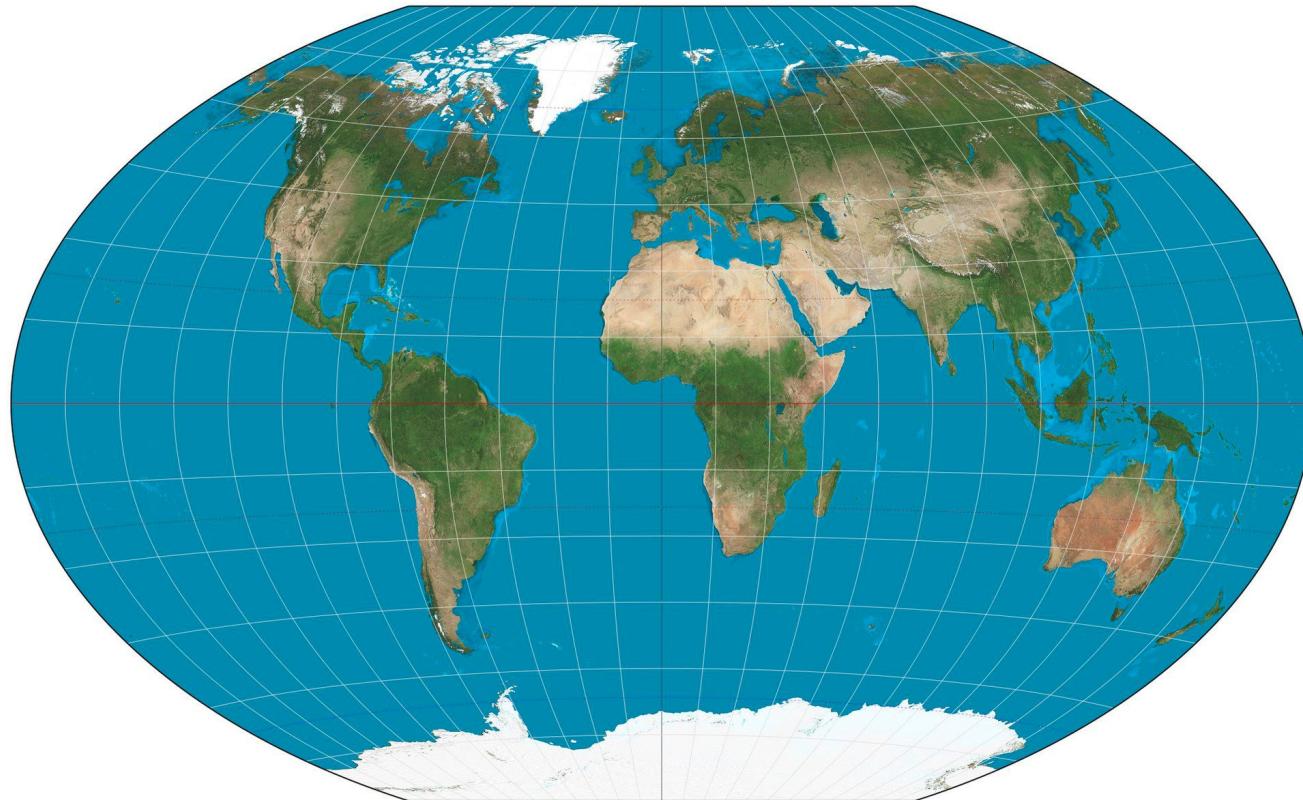
Merkator projection,  
1569,  
паралелі та  
мерідіани  
перпендикулярні,  
деформує площину  
ближче до полюсів.

Real size of  
Africa : Greenland = 30 : 2.1

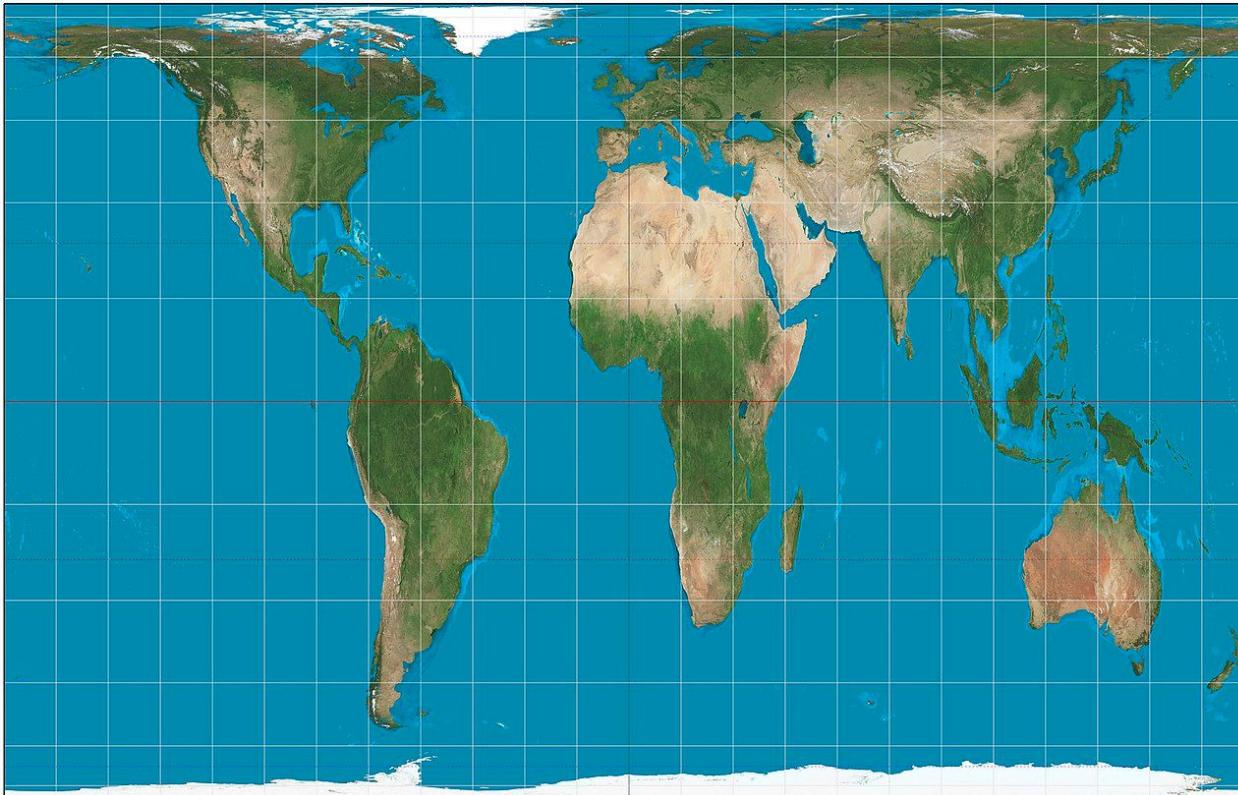


[https://en.wikipedia.org/wiki/Mercator\\_projection](https://en.wikipedia.org/wiki/Mercator_projection)

Winkel tripel projection, 1921, low distortion of area, direction, distance



Rectangular projection, saves areas, distorts shapes



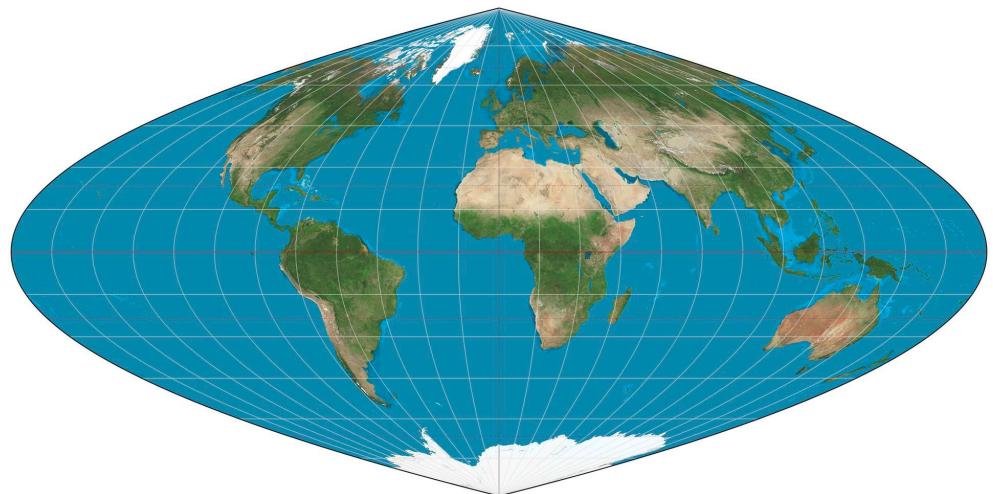
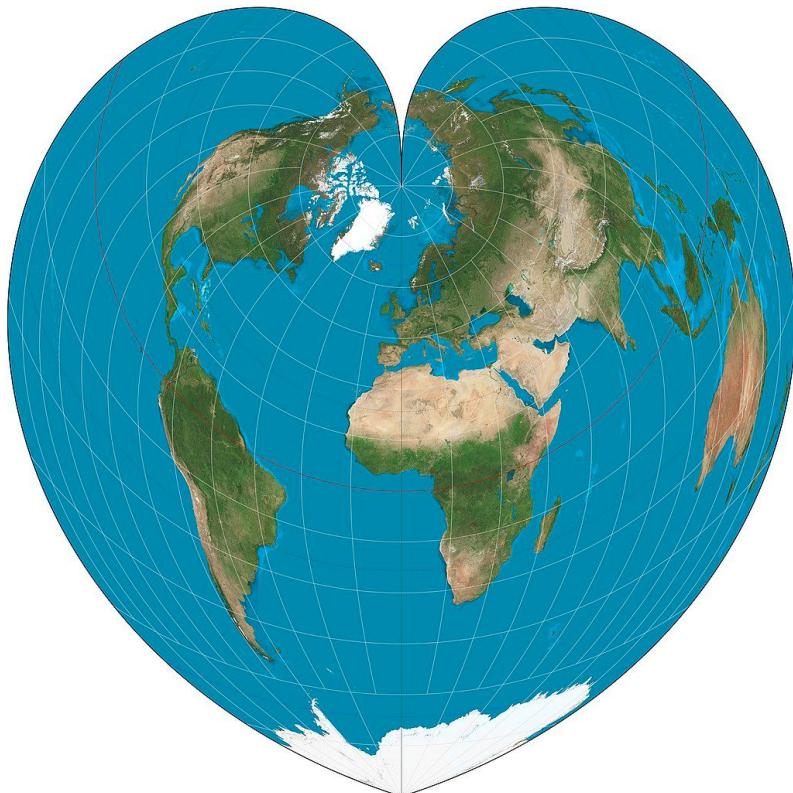
# AuthaGraph projection, polyhedral, 1999, 96 triangles



Трьоуктники - чотиригранники - прямокутники

More projections:

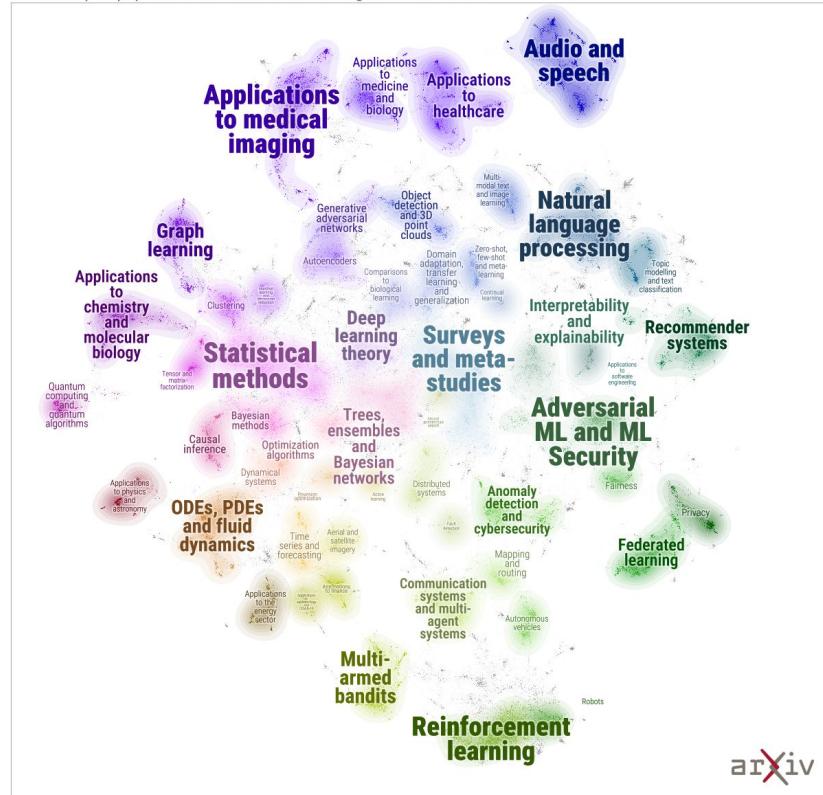
[https://en.wikipedia.org/wiki/List\\_of\\_map\\_projections](https://en.wikipedia.org/wiki/List_of_map_projections)



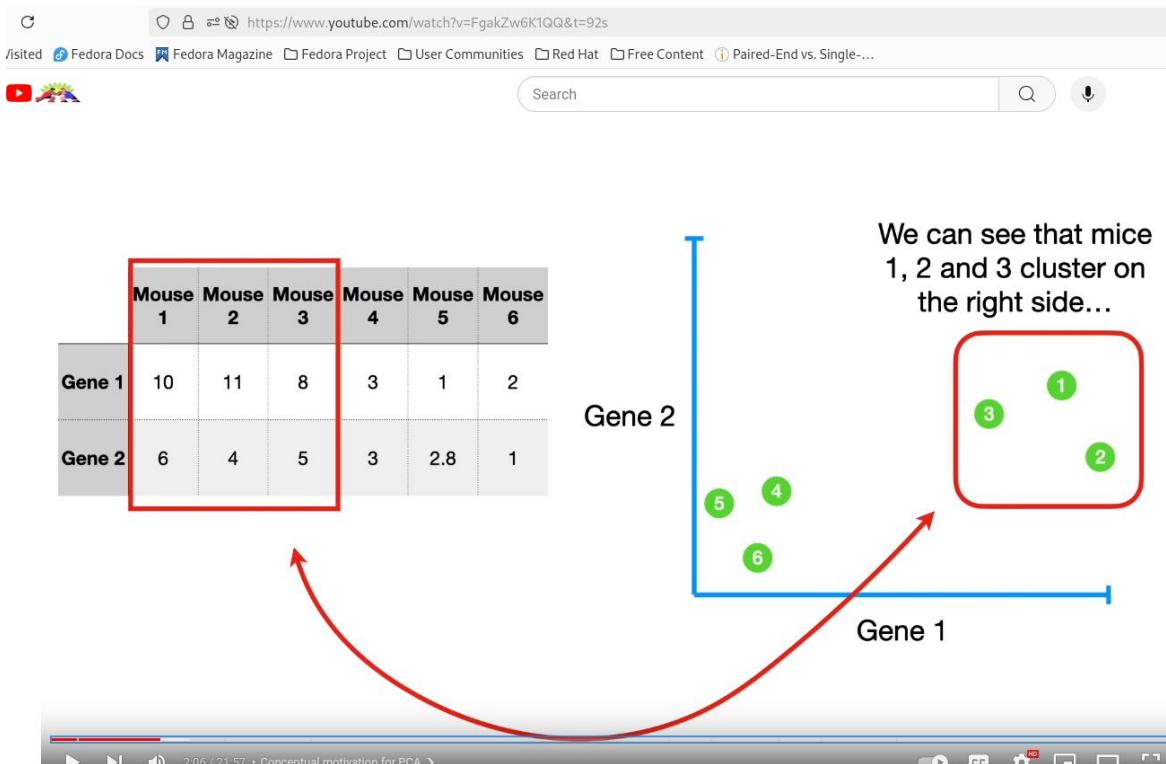
# Data map plot: статичні і динамічні візуалізації великих датасетів

<https://github.com/TuttleInstitute/datamapplot>

**ArXiv ML Landscape**  
A data map of papers from the Machine Learning section of ArXiv



# PCA - МЕТОД ГОЛОВНИХ КОМПОНЕНТ



<https://www.youtube.com/watch?v=FgakZw6K1QQ>

# PCA - метод головних компонент

https://www.youtube.com/watch?v=FgakZw6K1QQ&t=92s

/visited Fedora Docs Fedora Magazine Fedora Project User Communities Red Hat Free Content Paired-End vs. Single-...

Search

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS(distances)}$

...and we repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.

9:02 / 21:57 • Finding PC1 >

StatQuest: Principal Component Analysis (PCA), Step-by-Step



StatQuest with Josh Starmer  
1.22M subscribers

Join

Subscribe

58K

...

Share

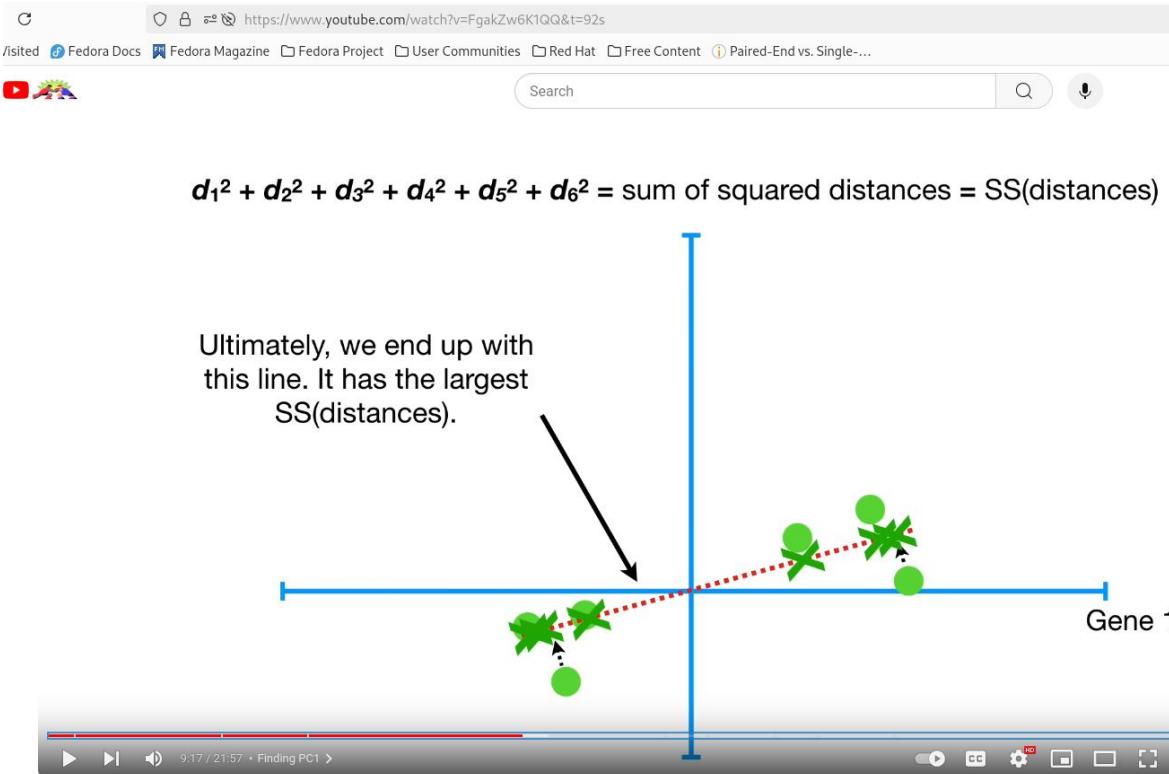
Thanks

Clip

Save

...

# PCA - метод головних компонент



StatQuest: Principal Component Analysis (PCA), Step-by-Step



StatQuest with Josh Starmer  
Join [Subscribe](#)

58K [Share](#) [Thanks](#) [Clip](#) [Save](#) ...

# PCA - МЕТОД ГОЛОВНИХ КОМПОНЕНТ

https://www.youtube.com/watch?v=FgakZw6K1QQ&t=92s

/visited Fedora Docs Fedora Magazine Fedora Project User Communities Red Hat Free Content Paired-End vs. Single-...

Search

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(distances)$

Eigenvalue Measure Of variation

$\frac{\text{SS}(distances for PC1)}{n - 1} = \text{Eigenvalue for PC1}$

Also, while I'm at it, PCA calls the average of the SS(distances) for the best fit line the **Eigenvalue for PC1...**

PC1, PC2 - linear combinations of variables

PC1: 4:1  
PC2: -1:4

12:42 / 21:57 • Singular vector/value, Eigenvector/value and loading scores defined >

StatQuest: Principal Component Analysis (PCA), Step-by-Step

StatQuest with Josh Starmer 1.22M subscribers

Join Subscribe

58K

Share

Thanks

Clip

Save

...

# PCA - МЕТОД ГОЛОВНИХ КОМПОНЕНТ

https://www.youtube.com/watch?v=FgakZw6K1QQ&t=92s

/visited Fedora Docs Fedora Magazine Fedora Project User Communities Red Hat Free Content Paired-End vs. Single-...

Search

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(distances)$

$\frac{\text{SS}(distances for PC2)}{n - 1} = \text{Eigenvalue for PC2}$

Lastly, the **Eigenvalue for PC2** is the average of the sum of squares of the distances between the projected points and the origin.

14:07 / 21:57 • Finding PC2 >

StatQuest: Principal Component Analysis (PCA), Step-by-Step

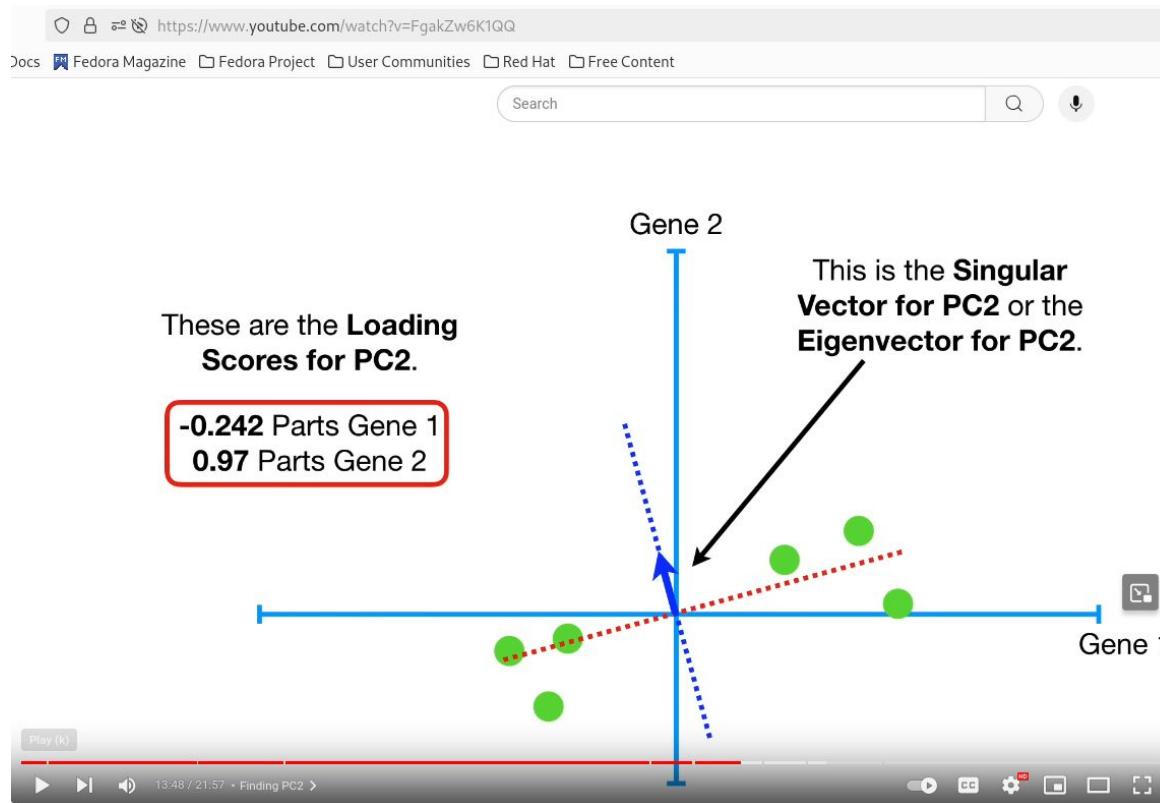
StatQuest with Josh Starmer 1.22M subscribers

Join Subscribe

58K

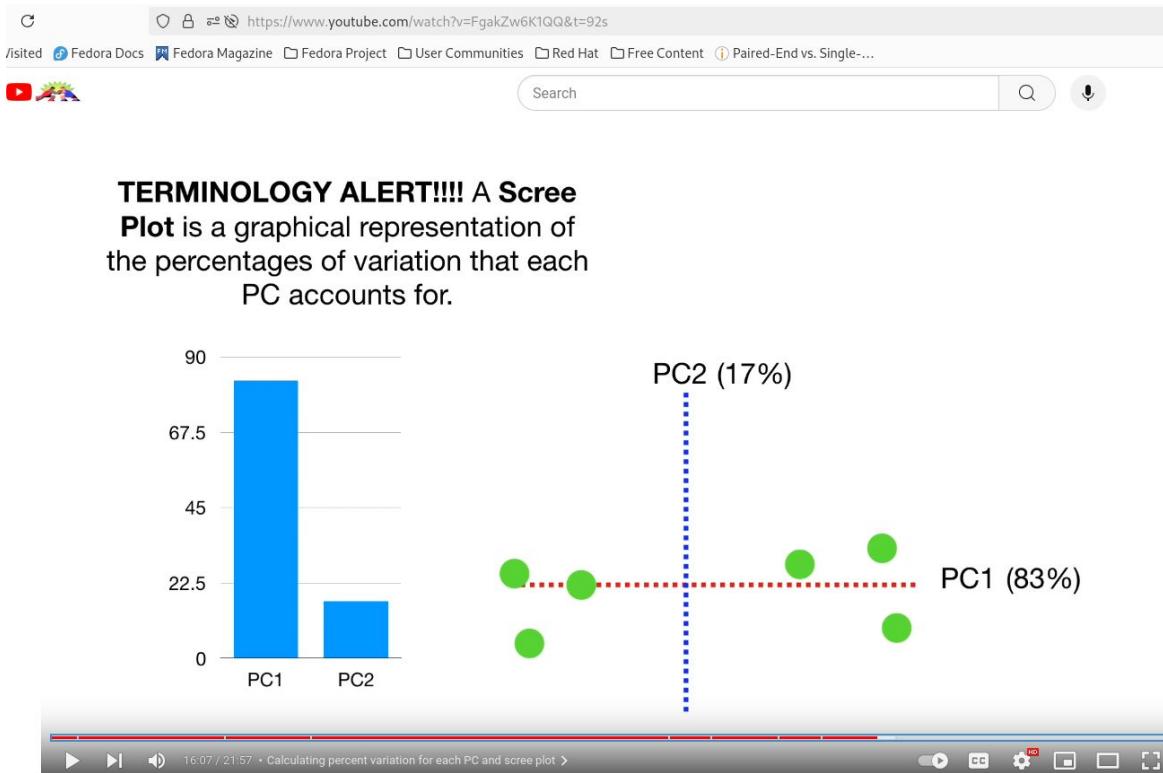
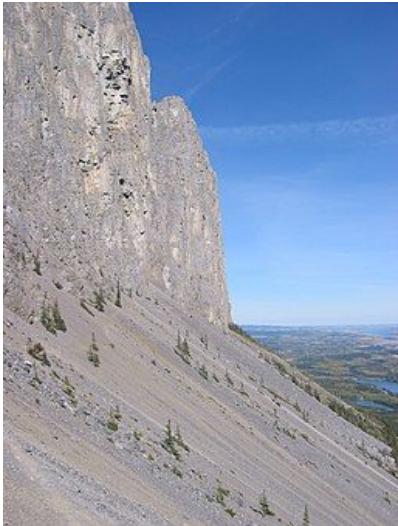
Share Thanks Clip Save ...

# Метод головних компонент



Loadings -  
навантаження

# PCA - метод головних компонент - осиповий графік



StatQuest: Principal Component Analysis (PCA), Step-by-Step



StatQuest with Josh Starmer  
1.22M subscribers

Join

Subscribe

58K

Clip

Thanks

Save

19

# Проблеми і переваги РСА

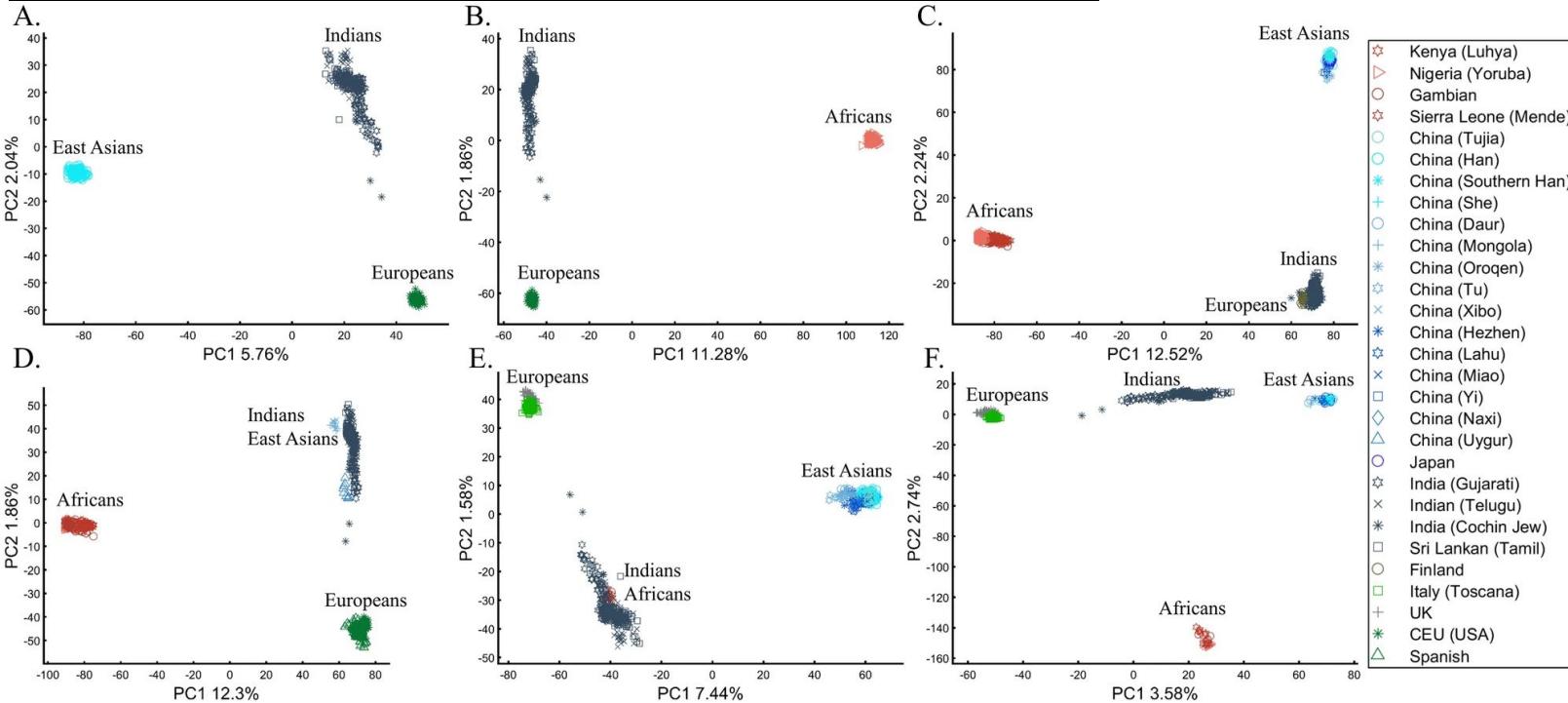
Переваги:

- Швидкий розрахунок
- Легко інтерпретувати

Недоліки:

- Зберігає глобальну, а не локальну структуру даних
- Тільки лінійні залежності
- Чутливий до викидів (outliers)

# Проблеми із PCA в популяційній генетиці



- <https://www.nature.com/articles/s41598-022-14395-4>, fig.5, Scientific Reports, 2022
- Змінюючи розміри популяції отримуємо різні PCA!

**Figure 5.** Studying the origin of Indians using PCA. (A) Replicating Reich et al.'s<sup>45</sup> results using  $n_{Eu}=99$ ;  $n_{As}=146$ ;  $n_{Ind}=321$ . Generating alternative PCA scenarios using: (B)  $n_{Af}=178$ ;  $n_{Eu}=99$ ;  $n_{Ind}=321$ , (C)  $n_{Af}=400$ ;  $n_{Eu}=40$ ;  $n_{As}=100$ ;  $n_{Ind}=321$ , (D)  $n_{Af}=477$ ;  $n_{Eu}=253$ ;  $n_{As}=23$ ;  $n_{Ind}=321$ , (E)  $n_{Af}=25$ ;  $n_{Eu}=220$ ;  $n_{As}=490$ ;  $n_{Ind}=320$ , and (F)  $n_{Af}=30$ ;  $n_{Eu}=200$ ;  $n_{As}=50$ ;  $n_{Ind}=320$ .

# Штучні коливання в PCA (проблеми в нейронауках)

**PNAS**

RESEARCH ARTICLE

NEUROSCIENCE  
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

## Phantom oscillations in principal component analysis

Maxwell Shinn<sup>a,1</sup> 

Edited by Michael Goldberg, Columbia University, New York, NY; received July 10, 2023; accepted October 18, 2023

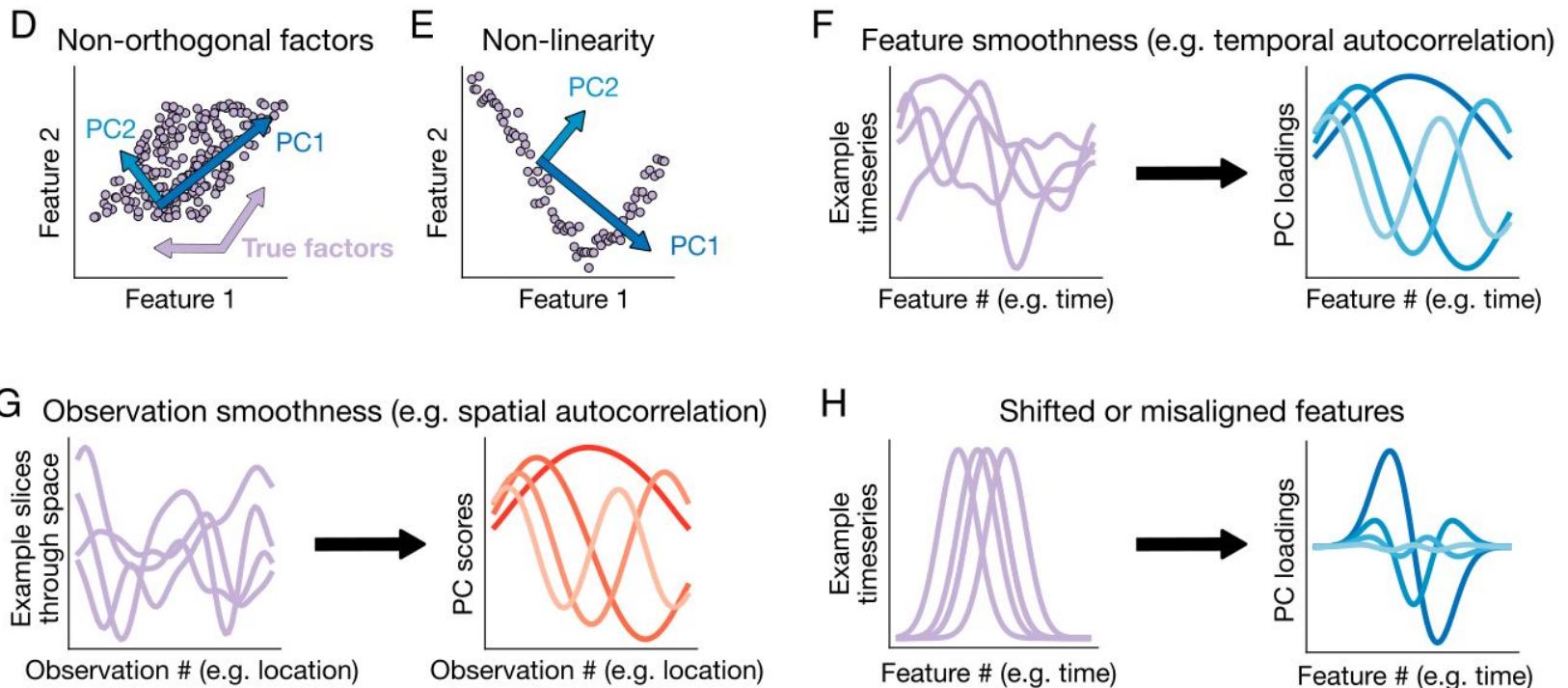
Principal component analysis (PCA) is a dimensionality reduction method that is known for being simple and easy to interpret. Principal components are often interpreted as low-dimensional patterns in high-dimensional space. However, this simple interpretation fails for timeseries, spatial maps, and other continuous data. In these cases, nonoscillatory data may have oscillatory principal components. Here, we show that two common properties of data cause oscillatory principal components: **smoothness** and shifts in time or space. These two properties implicate almost all neuroscience data. We show how the oscillations produced by PCA, which we call “phantom oscillations,” impact data analysis. We also show that traditional cross-validation does not detect phantom oscillations, so we suggest procedures that do. Our findings are supported by a collection of mathematical proofs. Collectively, our work demonstrates that patterns which emerge from high-dimensional data analysis may not faithfully represent the underlying data.

### Significance

Dimensionality reduction methods like PCA simplify high-dimensional data into a smaller set of representative features. However, PCA's dimensionality reduction is based on the assumption that the data can be represented by a few linear combinations of the original variables. This assumption is often violated in real-world data, leading to artifacts called “phantom oscillations.” These oscillations are caused by two common properties of data: smoothness and shifts in time or space. Smoothness means that the data points are close together, while shifts mean that the data points change over time or space. When these properties are present, PCA finds oscillatory principal components even though the data is non-oscillatory. This can lead to inaccurate results in data analysis, such as incorrect conclusions about the underlying patterns in the data. To address this issue, the authors propose new methods for detecting and correcting for phantom oscillations in PCA. Their findings have important implications for the use of PCA in fields like neuroscience, where data often exhibits complex, non-oscillatory patterns.

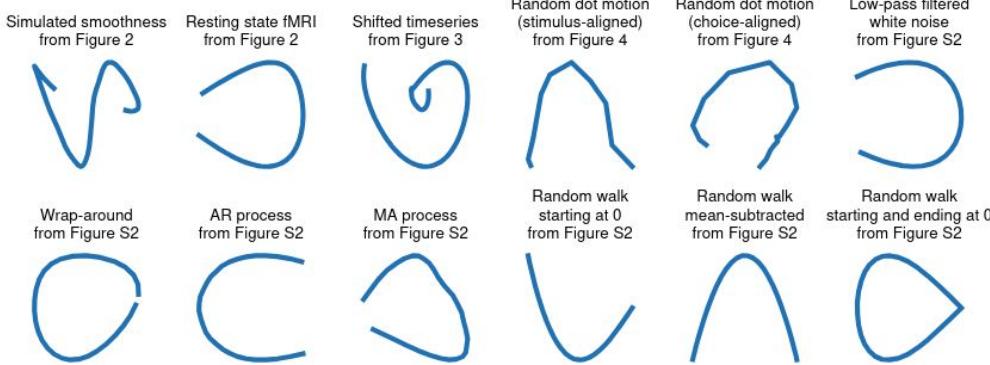
<https://www.pnas.org/doi/epub/10.1073/pnas.2311420120>

## Biases in PCA

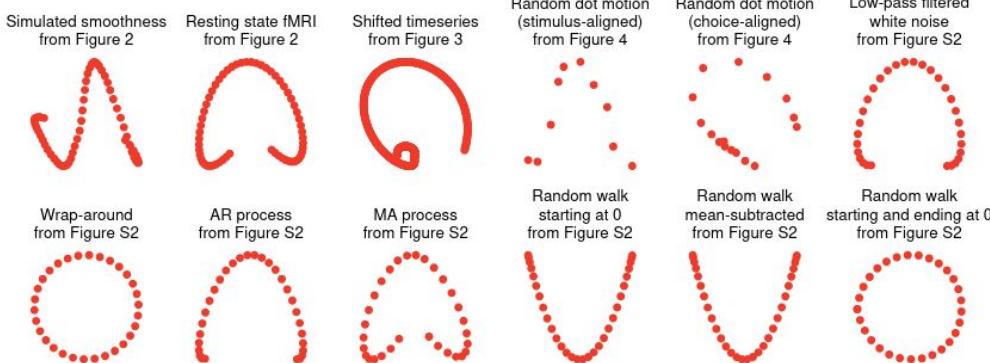


**Fig. 1.** Summary of biases in PCA (A) Schematic including terminology we use throughout. (B) Summary of how to compute principal components (PCs). (C) Summary of how to compute scores. (D–H) Illustrations of (D) nonorthogonality bias, (E) nonlinearity bias, (F) smooth features leading to oscillatory loadings, (G) smooth observations leading to oscillatory scores, and (H) time-shifted features leading to oscillatory loadings.

### PC components - PC1 vs PC2



### PC scores (transposed matrix) - PC1 vs PC2

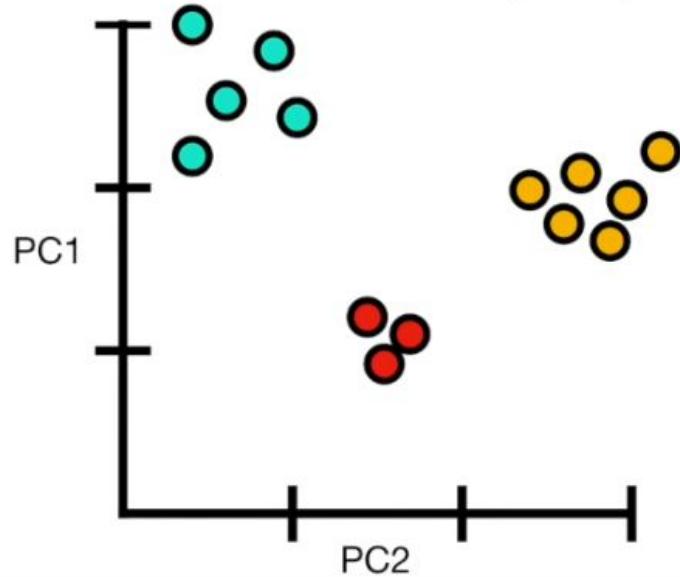


**Fig. S3. Phantom oscillations have U-shaped PC plots.** PC1 and PC2 can be plotted against each other to find population trajectories or to visualize variation across the population. For each dataset analyzed, we show the PC1 loadings on the  $x$  axis plotted against the PC2 loadings on the  $y$  axis. When plotted this way, phantom oscillations often appear as U-shaped plots. When PCA is computed on the transposed data matrix, phantom oscillations appear as a U-shape in the first two PC scores. This is sometimes known as the "horseshoe effect". Higher components plotted against each other show Lissajous-like trajectory patterns.

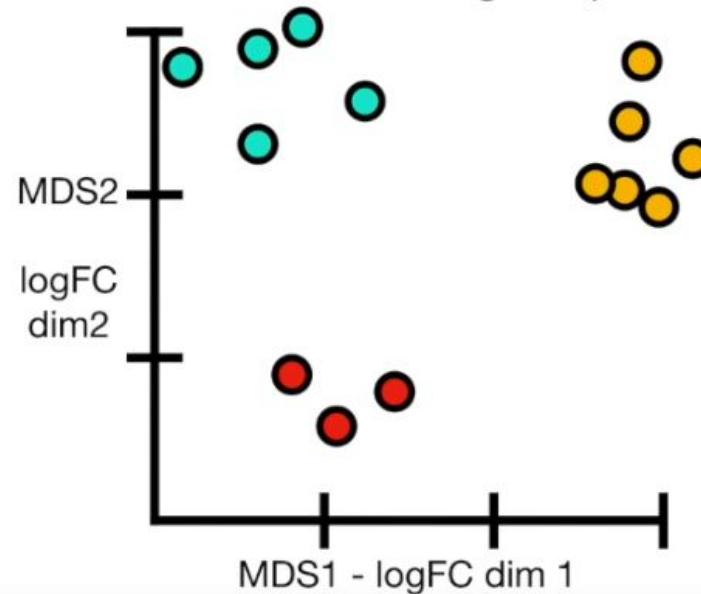
Покращення PCA: замість пошуку кореляцій між векторами будемо працювати із відстанями між точками і намагатись зберігти відстані

MDS (multidimensional scaling) може використовувати різні метрики

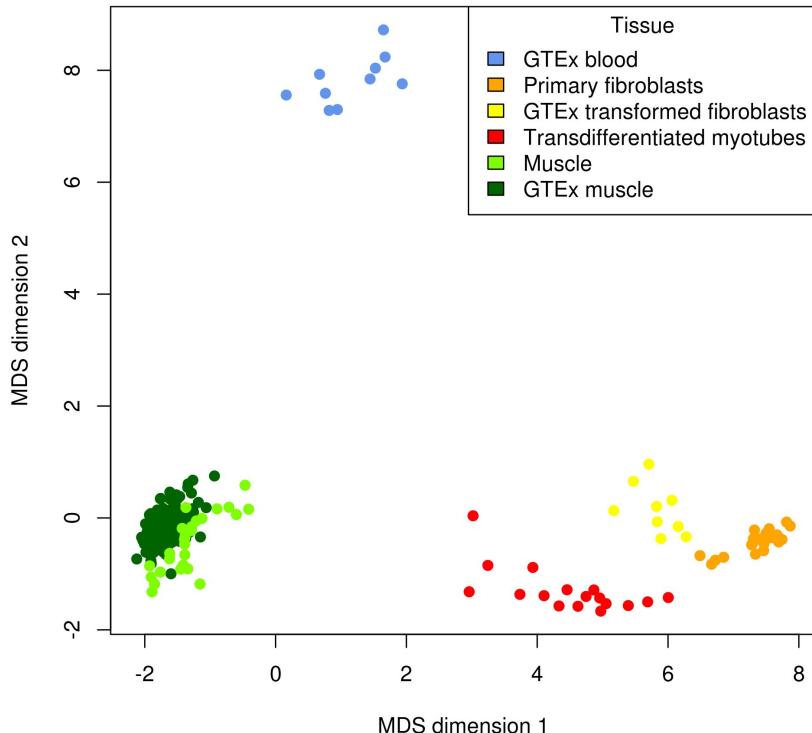
PCA creates plots based on correlations among samples.



MDS and PCoA create plots based on distances among samples

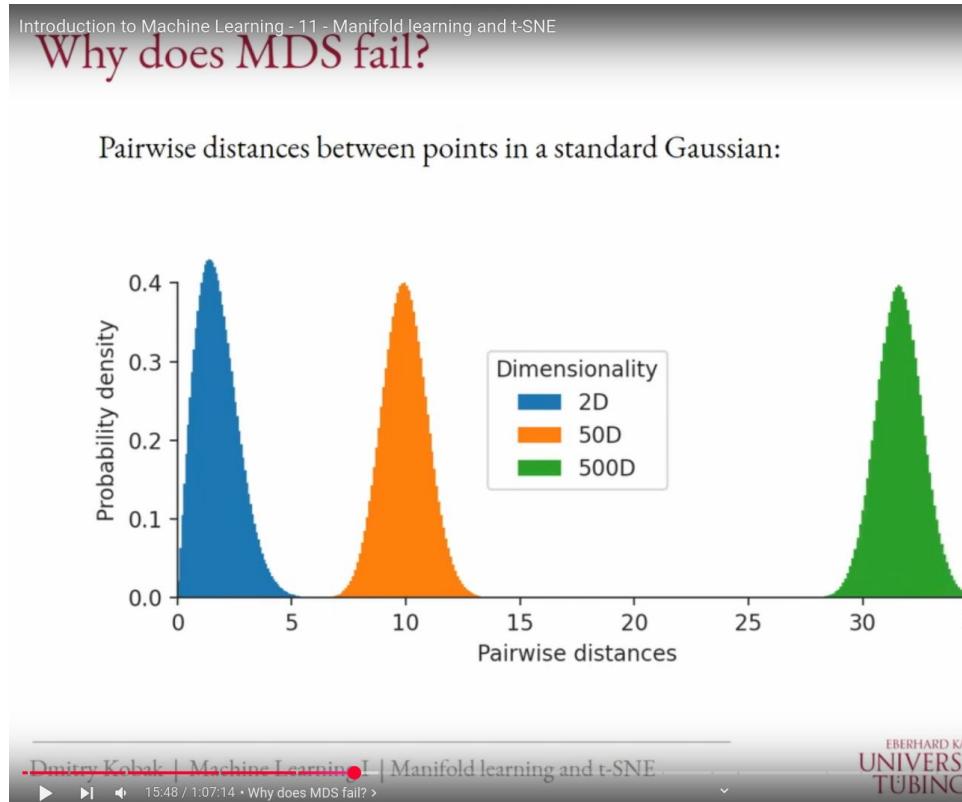


# Приклад зі статті: bulk RNA-seq data, plotMDS /edgeR



Недолік MDS - не можна працювати із великими датасетами (>500 зразків)  
Bulk RNA-seq: PCA, MDS, SingleCell RNA-seq: tSNE, UMAP

# Чому важко зберігти відстані - див. лекцію Кобяка у посиланнях

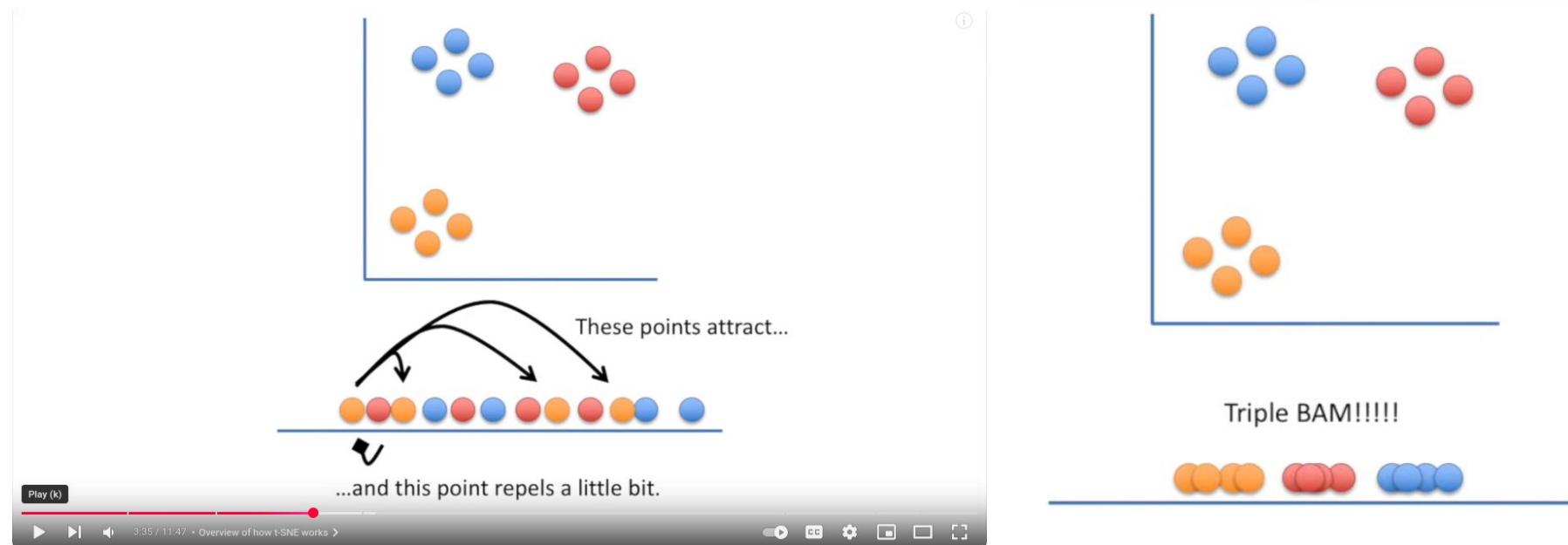


tSNE - нелінійний спосіб зниження розмірності даних

“Будемо зберігати найближчих сусідів замість зберігання відстаней”

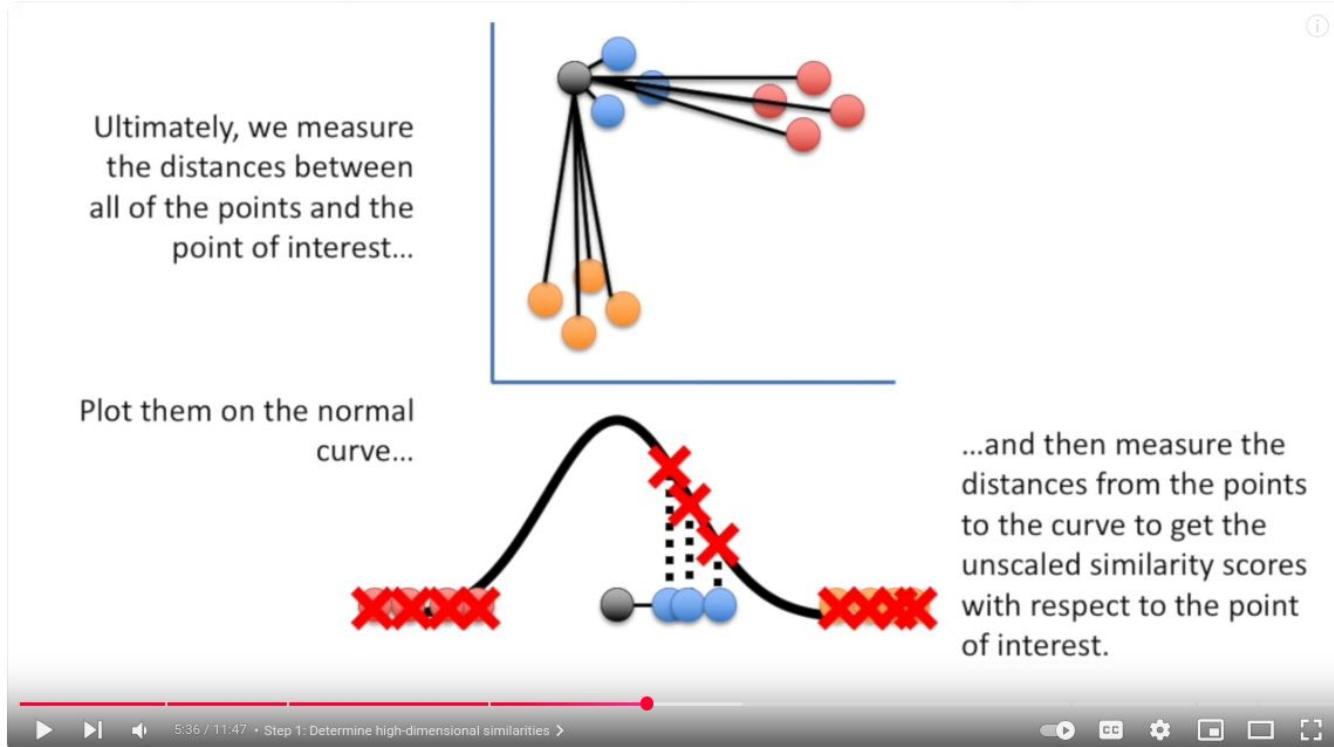
t-Distributed Stochastic Neighbor Embedding

Т-розподілене вкладення стохастичної близькості



How tSNE works: <https://www.youtube.com/watch?v=NEaUSP4YerM>

# Моделювання сусідів; схожість точок у мультивимірному просторі (p) - gaussian kernel - використовуємо нормальний розподіл



# Perplexity ~ перплексивність і sigma (стандартне відхилення для нормального розподілу)

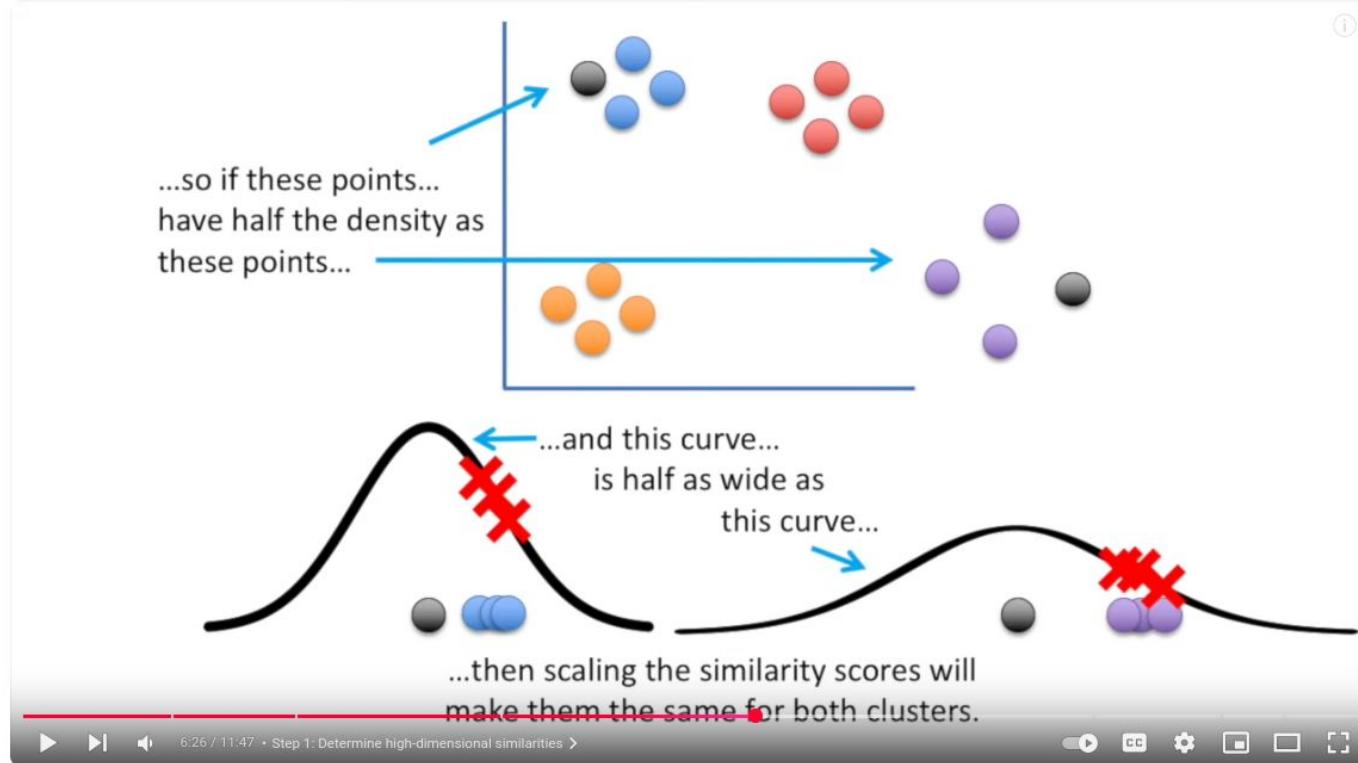
- Перплексивність ~ кількість сусідів
- Sigma вираховується адаптивно (дляожної точки даних свій розподіл), щоб досягти заданої перплексивності
- “Пропустити через Гаусове ядро”
- «подібність точки даних  $x_j$  до точки даних  $x_i$  — це умовна ймовірність,  $p(j|i)$ , що  $x_i$  вибрал би  $x_j$  як свого сусіда, якби сусіди були обрані пропорційно їх гаусовій [густині ймовірності](#) з центром в  $x_i$ .»<sup>[1]</sup>

Сусіди ~30 - висока схожість



Чужинці - низька схожість

# Масштабуємо для збереження локальної структури



StatQuest: t-SNE, Clearly Explained



StatQuest with Josh Starmer  
1.35M subscribers

Join

Subscribe

11K



Share

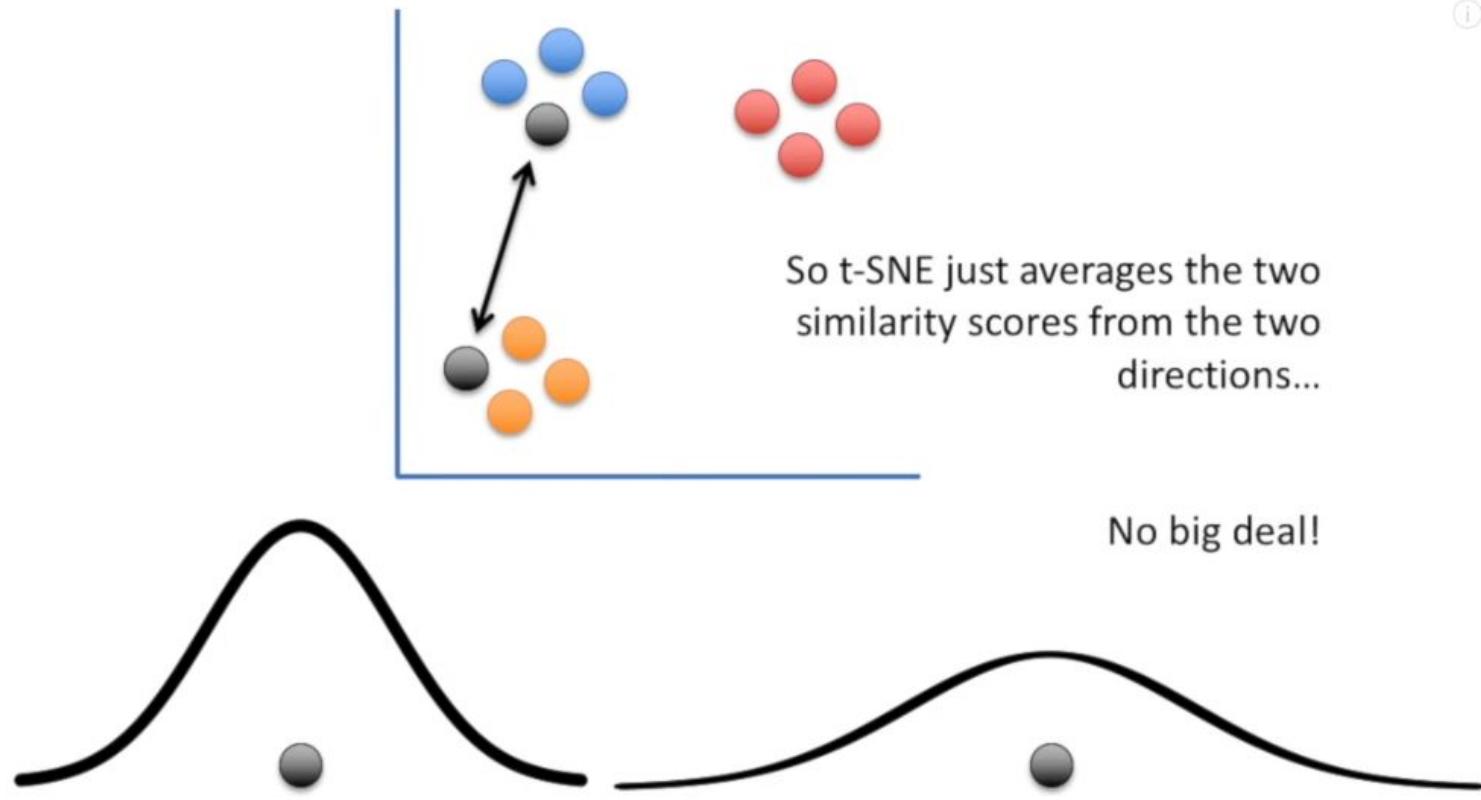
Thanks

Clip

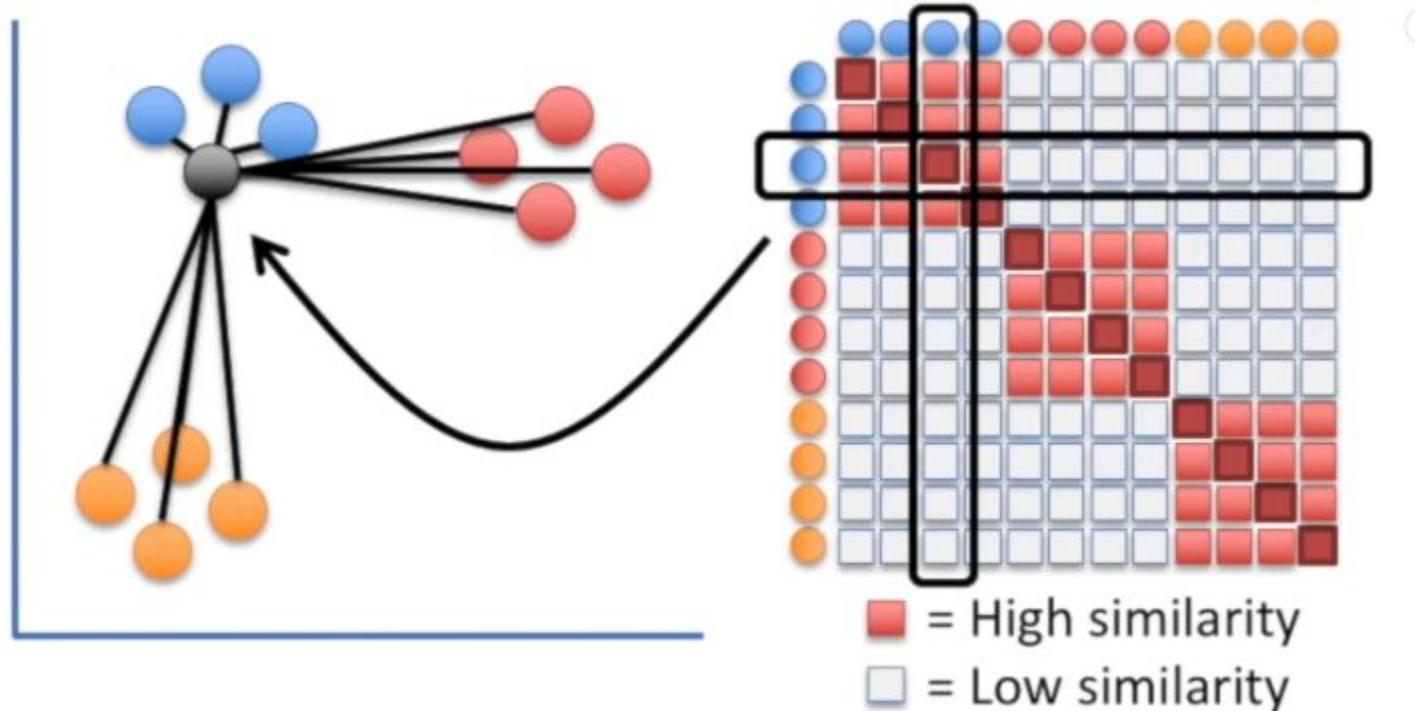
Save

...

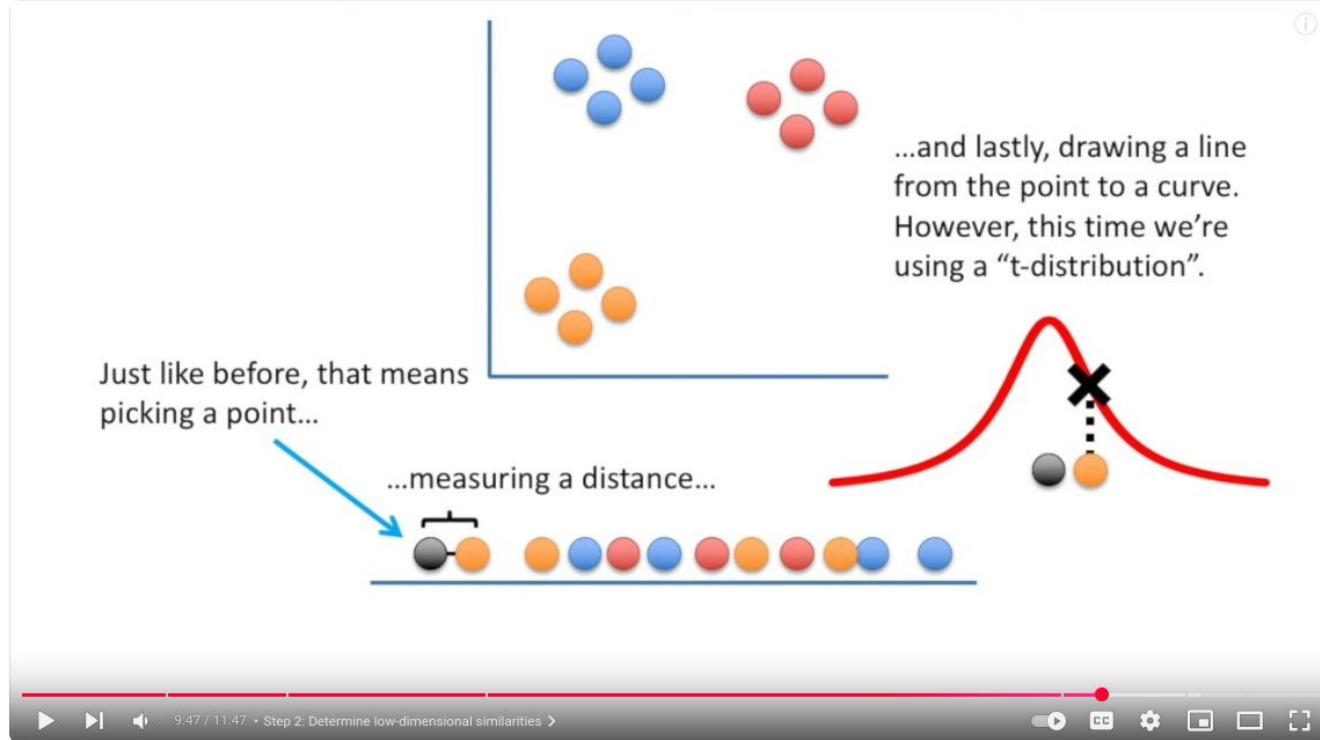
# Міра подібності не є симетричною



# Матриця р-значень подібності точок даних (не плутати із p-value у статистиці!)



# Моделювання схожості у 2D - t-розподіл



StatQuest: t-SNE, Clearly Explained



StatQuest with Josh Starmer  
1.35M subscribers

Join

Subscribe

11K



Share



Thanks



Clip



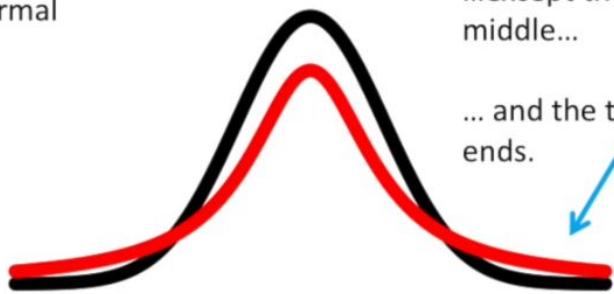
Save

...

# Проблема скучення вирішується за допомогою Т

A “t-distribution” ...

...is a lot like a normal distribution...



The “t-distribution” is the “t” in t-SNE.

...except the “t” isn’t as tall in the middle...

... and the tails are taller on the ends.

Якщо сусідство дуже густе, то з заданої перплексивності багато точок отримують низькі значення схожості за нормального розподілу



StatQuest: t-SNE, Clearly Explained

 StatQuest with Josh Starmer   
1.35M subscribers

Join

Subscribe

11K



Share

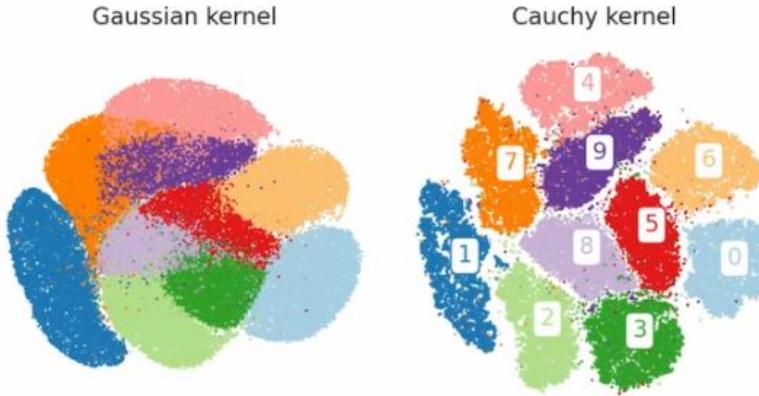
Thanks

Clip

Save

...

# Проблема скучення - близькі точки розташовані далеко у 2D за використання Гаусового ядра



Even heavier-tailed kernels can bring out even finer cl

Dmitry Kobak | Machine Learning I | Manifold learning and t-SNE  
47:15 / 1:07:14 • Low-dimensional similarity kernel >

Introduction to Machine Learning - 11 - Manifold learning and t-SNE



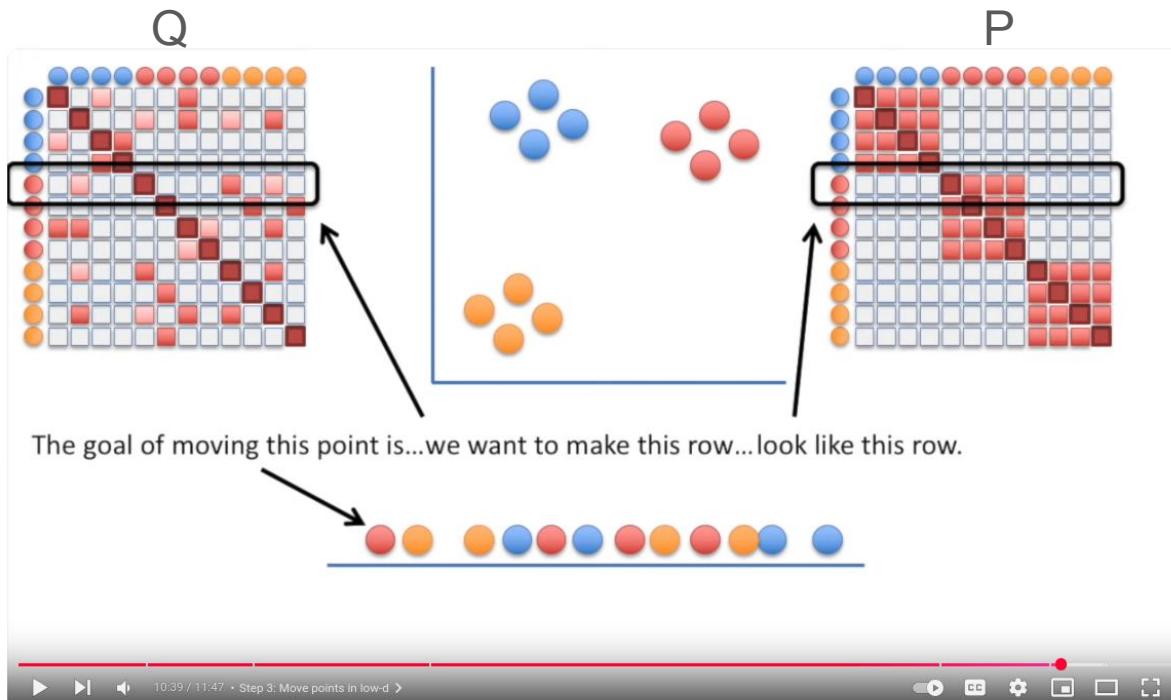
Tübingen Machine Learning  
42.4K subscribers

Subscribe

# tSNE - мінімізація міри відмінності Q і P

Координати точок  $y_i$  при відображення визначаються шляхом мінімізації (несиметричної) відмінності по мірі Кульбака-Лейблера розподілу  $Q$  від розподілу  $P$ , тобто:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$



StatQuest: t-SNE, Clearly Explained



StatQuest with Josh Starmer  
1.35M subscribers

Join

Subscribe

11K

Share

Thanks

Clip

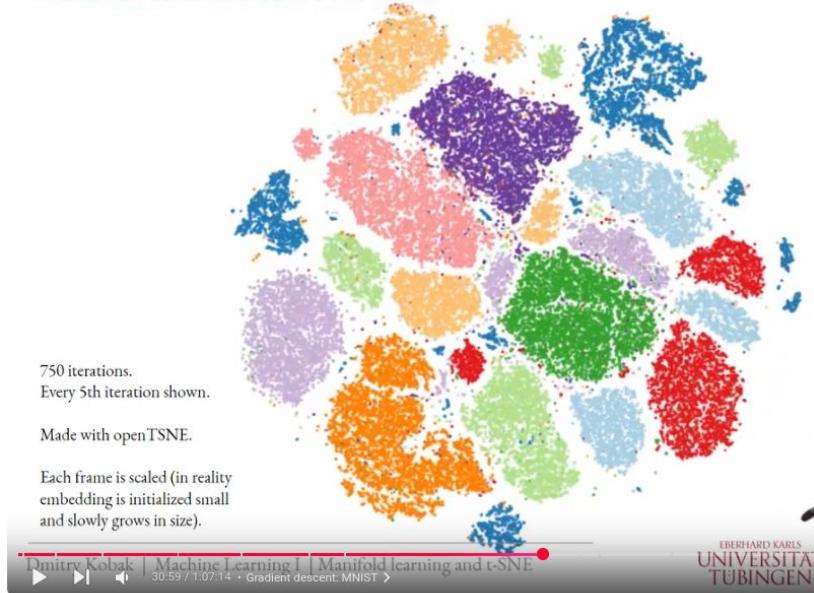
Save

37

# tSNE algorithm

- Метод градієнтного спуску
- Сили тяжіння: притягання і відштовхування
- Налаштування exaggeration (перебільшення) - на перших ітераціях сила тяжіння має більший коефіцієнт ніж відштовхування
- Оптимізації алгоритмів: kNN-графи

Gradient descent: MNIST



Introduction to Machine Learning - 11 - Manifold learning and t-SNE



Tübingen Machine Learning  
42.4K subscribers

Subscribe

1.3K

# tSNE: Laurens van der Maaten (Netherlands) and Geoffrey Hinton (Canada)



Laurens van der  
Maaten

Research scientist in  
artificial intelligence.

- Email
- Twitter
- Google Scholar
- LinkedIn
- Github

## t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. The technique can be implemented via Barnes-Hut approximations, allowing it to be applied on large real-world datasets. We applied it on data sets with up to 30 million examples. The technique and its variants are introduced in the following papers:

- L.J.P. van der Maaten. **Accelerating t-SNE using Tree-Based Algorithms.** *Journal of Machine Learning Research* 15(Oct):3221–3245, 2014. [PDF](#) [Supplemental material]
- L.J.P. van der Maaten and G.E. Hinton. **Visualizing Non-Metric Similarities in Multiple Maps.** *Machine Learning* 87(1):33–55, 2012. [PDF](#)
- L.J.P. van der Maaten. **Learning a Parametric Embedding by Preserving Local Structure.** In *Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP* 5:384–391, 2009. [PDF](#)
- L.J.P. van der Maaten and G.E. Hinton. **Visualizing High-Dimensional Data Using t-SNE.** *Journal of Machine Learning Research* 9(Nov):2579–2605, 2008. [PDF](#) [Supplemental material] [Talk]

An accessible introduction to t-SNE and its variants is given in this [Google Techtalk](#).

<https://lvdmaaten.github.io/tsne/>

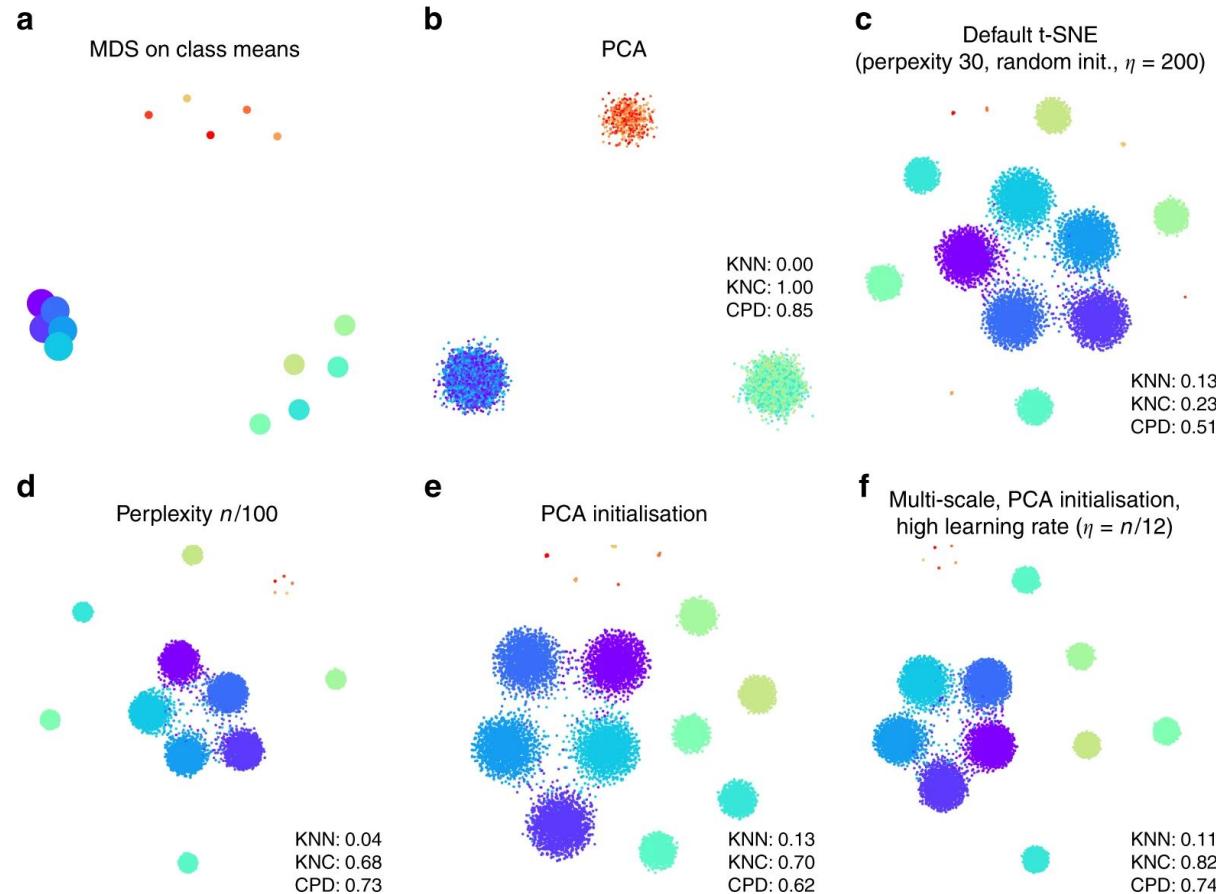
# PCA vs tSNE

Feature	PCA	t-SNE
Type of Dimensionality Reduction	Linear dimensionality reduction technique	Non-linear dimensionality reduction technique
Structure Preservation	Preserves global structure of data	Preserves local structure (clusters) of data
Effectiveness	Works well for global patterns but may not capture local clusters effectively	One of the best techniques for visualizing local clusters
Hyperparameters	Has fewer hyperparameters	Involves hyperparameters such as learning rate and number of steps
Sensitivity to Outliers	Sensitive to outliers	More robust to outliers
Algorithm Type	Deterministic algorithm i.e it produces the same result every time	Non-deterministic algorithm as results may vary due to randomness
Transformation Method	Transforms data into a new coordinate system to maximize variance	Minimizes the distance between points in a Gaussian probability distribution
Variance Preservation Control	Allows control over variance preservation using eigenvalues	Preserves distances rather than variance and is controlled by hyperparameters
Computational Efficiency	Computationally efficient especially for large datasets	Computationally expensive for large datasets
Primary Use	Can be used for dimensionality reduction and visualization	Primarily designed for data visualization and exploratory analysis
Data Separability	Works well for linearly separable datasets	Better suited for non-linearly separable datasets
Sensitivity to Data Ordering	Can be sensitive to the ordering of data points	Less sensitive to data ordering

практика!

# tSNE in Single Cell transcriptomics - synthetic dataset:

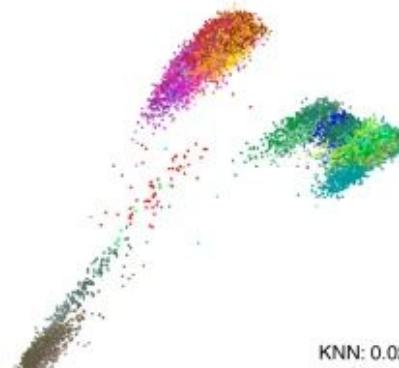
<https://www.nature.com/articles/s41467-019-13056-x>



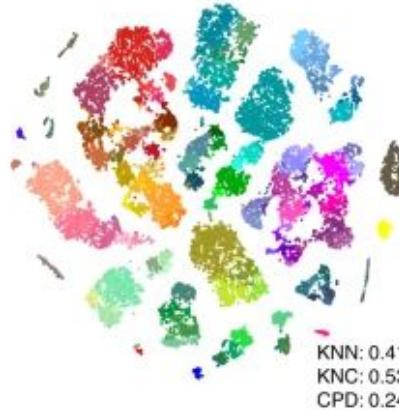
**a**  
MDS on class means



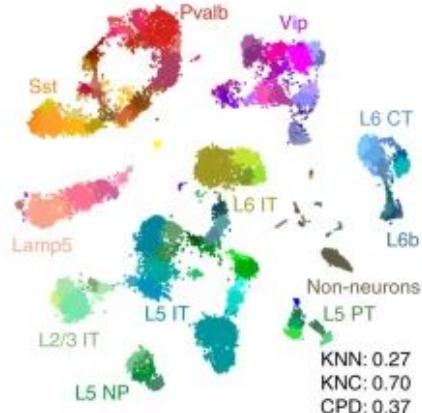
**b**  
PCA



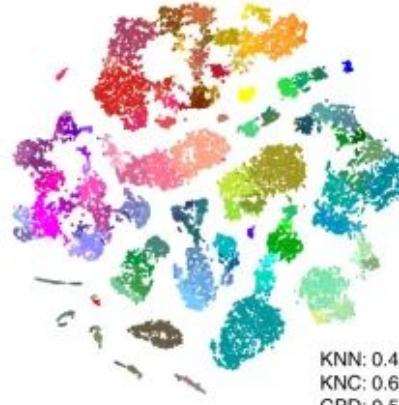
**c**  
Default t-SNE  
(perplexity 30, random init.,  $\eta = 200$ )



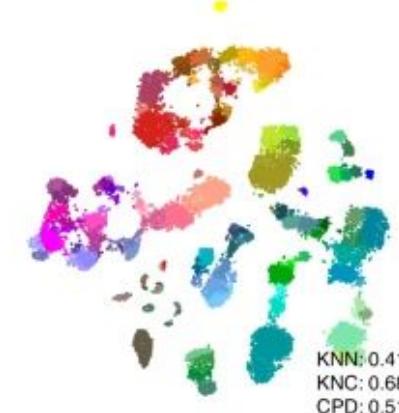
**d**  
Perplexity  $n/100$



**e**  
PCA initialisation



**f**  
Multi-scale, PCA initialisation,  
high learning rate ( $\eta = n/12$ )



## The art of using t-SNE for single-cell transcriptomics

[Dmitry Kobak](#)  & [Philipp Berens](#) 

[Nature Communications](#) 10, Article number: 5416 (2019) | [Cite this article](#)

100k Accesses | 582 Citations | 163 Altmetric | [Metrics](#)

## The specious art of single-cell genomics

Tara Chari, Lior Pachter 

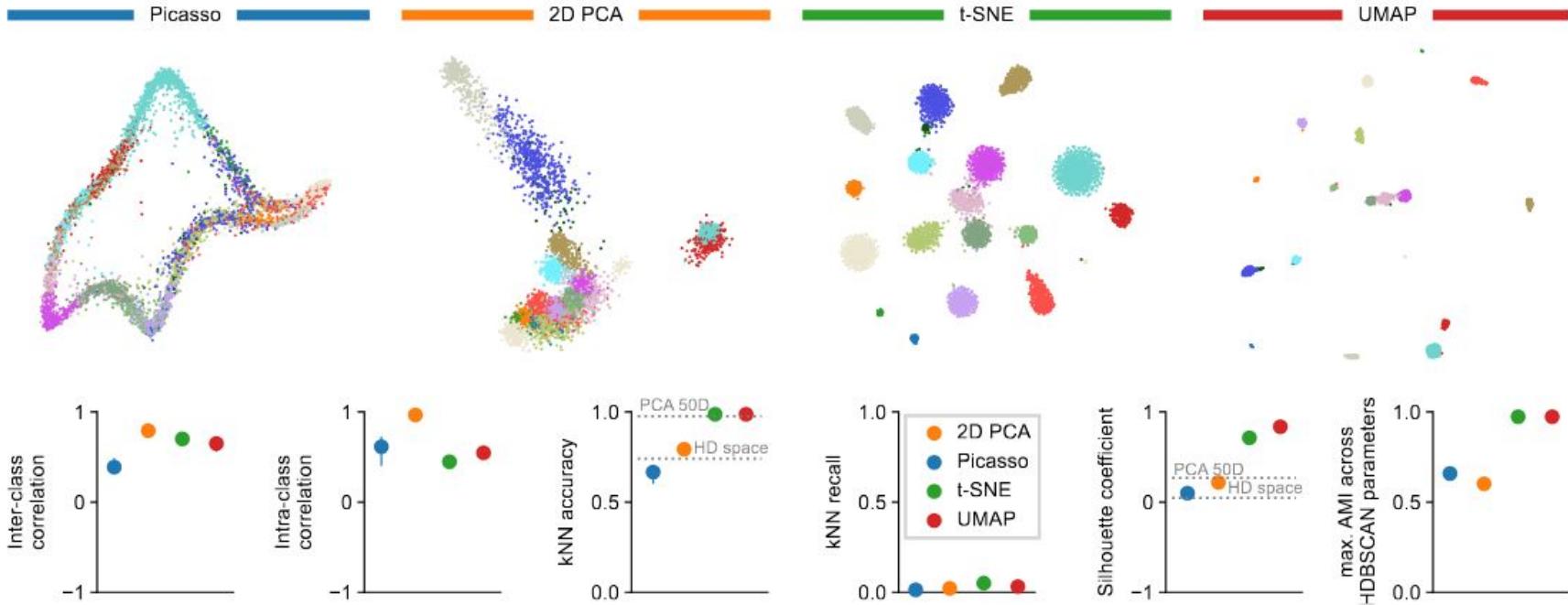
Published: August 17, 2023 • <https://doi.org/10.1371/journal.pcbi.1011288>

## The art of seeing the elephant in the room: 2D embeddings of single-cell data do make sense

Jan Lause, Philipp Berens, Dmitry Kobak 

Published: October 2, 2024 • <https://doi.org/10.1371/journal.pcbi.1012403>

[See the preprint](#)

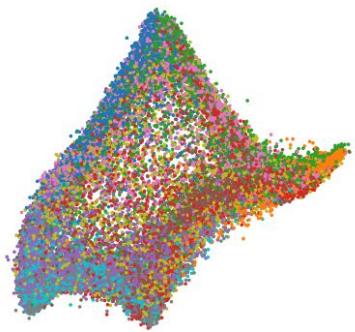


Picasso

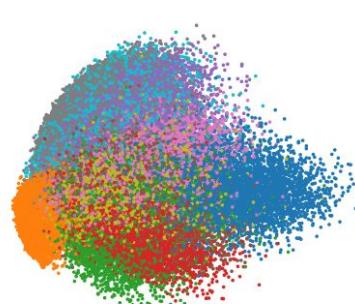
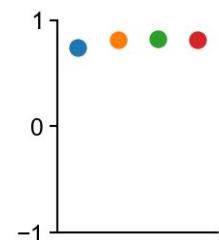
2D PCA

t-SNE

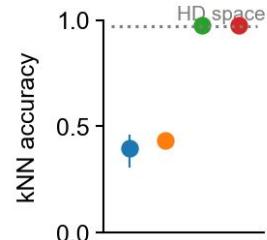
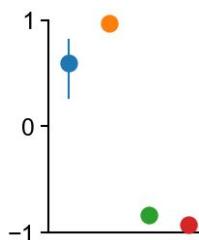
UMAP



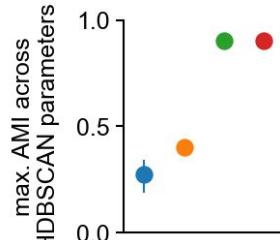
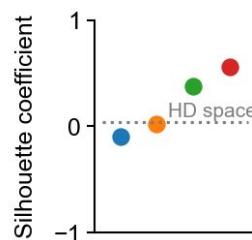
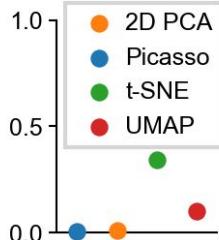
Inter-class correlation



Intra-class correlation

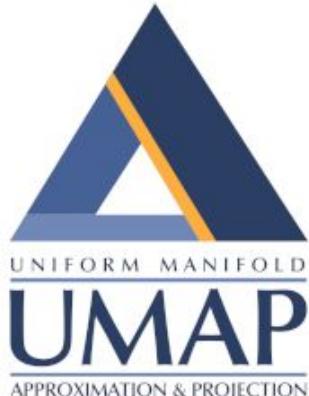


kNN recall



# UMAP = Uniform Manifold Approximation and Projection

- (Намагається) Зберігає локальну і глобальну структуру
- Метод оснований на топології, теорії многовидів
- Алгоритм досить подібний до tSNE, але:
  - Ініціалізація стандартна (спектральний алгоритм)
  - Швидко працює
  - Більш стабільний
  - Дуже детально розроблений



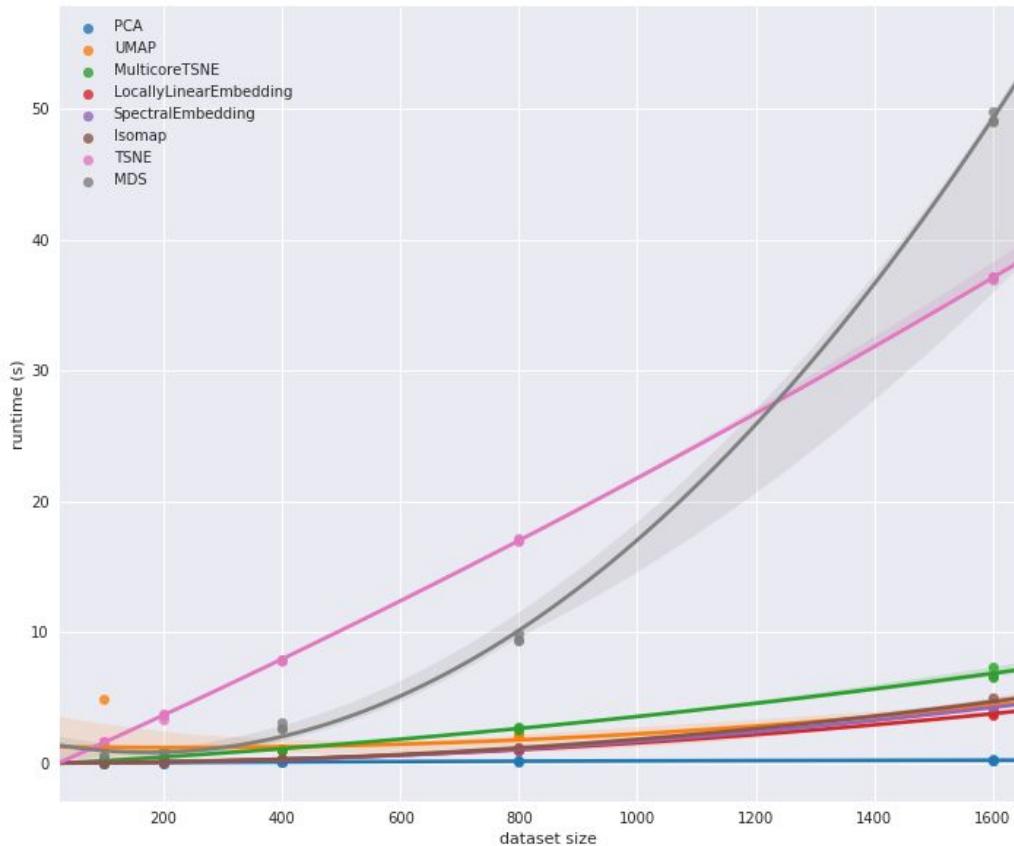
<https://umap-learn.readthedocs.io/en/latest/index.html>

[https://umap-learn.readthedocs.io/en/latest/how\\_umap\\_works.html](https://umap-learn.readthedocs.io/en/latest/how_umap_works.html)

# Параметри UMAP

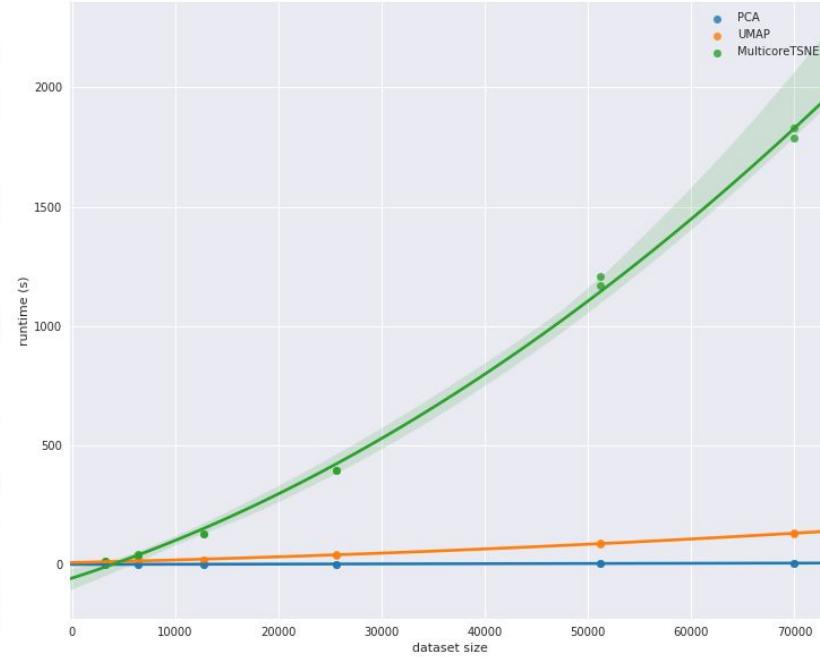
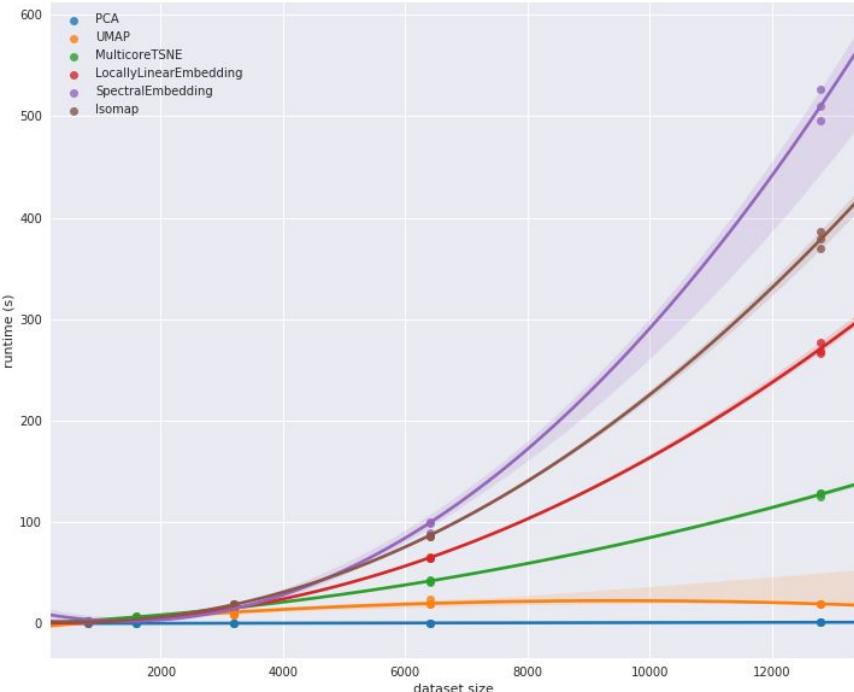
- <https://umap-learn.readthedocs.io/en/latest/parameters.html>
- Дозволяють налаштовувати спiввiдношення високомiрного простору i двомiрного
- n\_neighbors = 15 # сусiди в високомiрному просторi (як перплексивнiсть в t-SNE)
- min\_dist = 0.1 # мiнiмальна вiдстань в 2D

На великих датасетах UMAP працює майже так же швидко, як і PCA

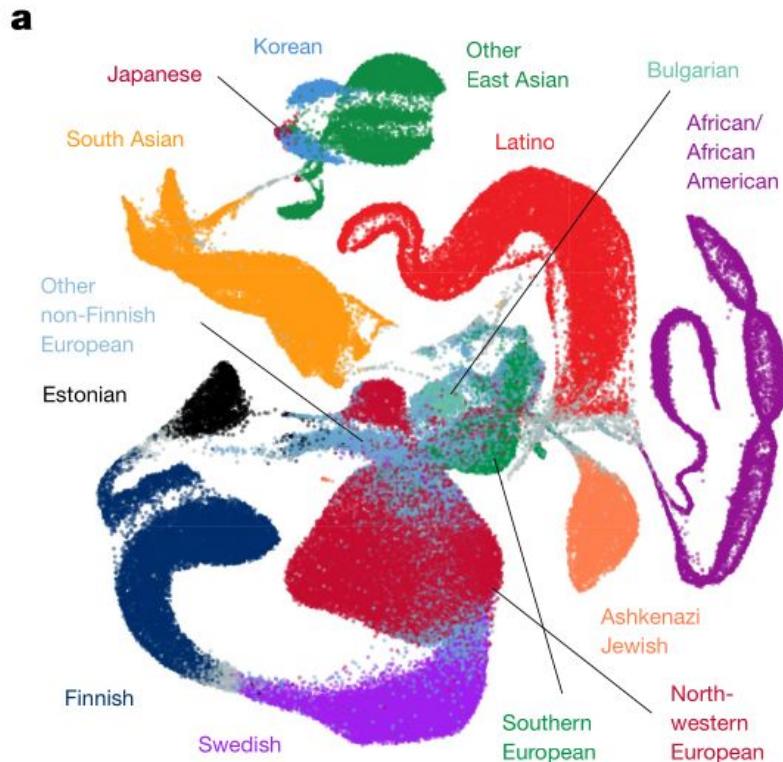


<https://umap-learn.readthedocs.io/en/latest/benchmarking.html>

# На великих датасетах УМАР працює майже так же швидко, як і РСА



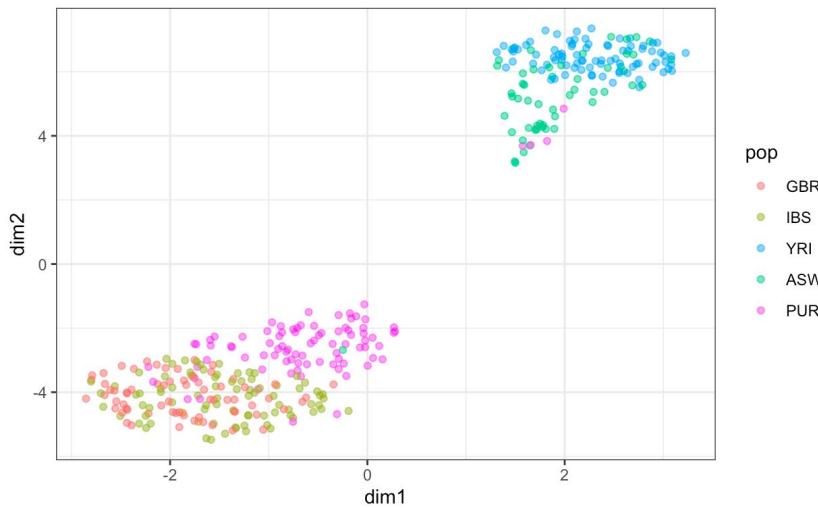
# UMAP - gnomad dataset v2, 2020, Fig1, 141k genomes + exomes



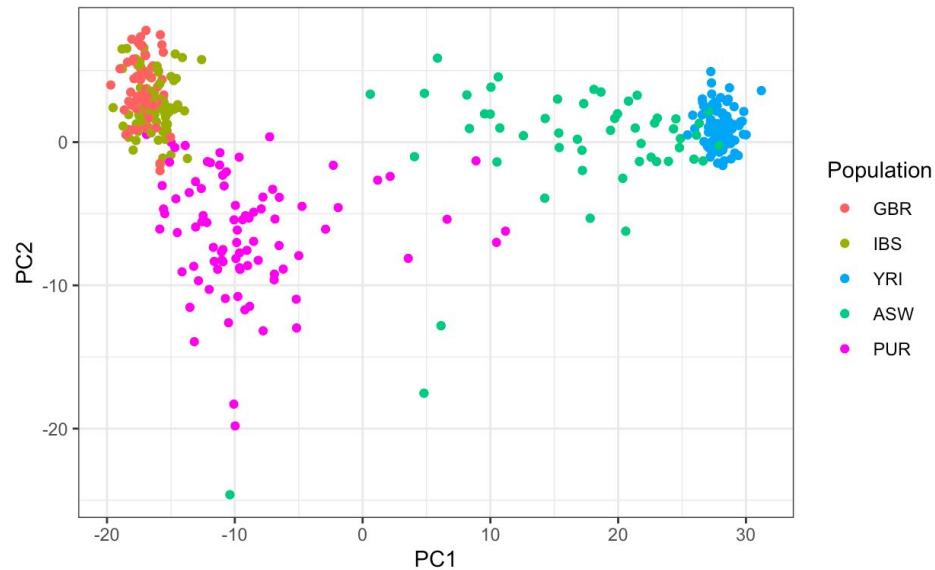
<https://www.nature.com/articles/s41586-020-2308-7>

# Треба завжди обережно ставитись до інтерпретації

УМАР



PCA



<https://simplystatistics.org/posts/2024-12-23-biologists-stop-including-umap-plots-in-your-papers/>

## PCA

# Посилання

- [PCA - вікіпедія](#)
- [PCA step-by-step by Josh Starmer](#)
- [PCA із прикладами в R із книги “Modern Statistics for Modern biology” by Susan Holmes and Wolfgang Huber](#)
- [PCA-based findings in population genetics studies are highly biased and must be reevaluated](#)
- [Phantom oscillations in PCA](#)

## tSNE

- [StatQuest: t-SNE, Clearly Explained](#)
- [tSNE in Ukrainian](#)
- [tSNE by the author](#)
- [The art of using t-SNE for single-cell transcriptomics](#)
- [The specious art of single-cell genomics](#)
- [The art of seeing elephant in the room](#)
- [Quick and easy t-SNE analysis in R](#)
- [Introduction to Machine Learning - 11 - Manifold learning and t-SNE](#)

## UMAP

- [StatQuest: UMAP](#)
- [UMAP by the author](#)
- [t-SNE vs UMAP](#)
- [Running UMAP for data visualization in R](#)
- [Biologists, stop putting UMAP plots in your papers](#)
- [Open Syllabus Galaxy - UMAP](#)