Introduction to dimensional reduction in R

Isabella Bicalho Frazeto

Welcome to the tutorial:

Introduction to dimensional reduction in R

The materials are at:

https://github.com/bellabf/dimensional-reduction LINK IN THE CHAT

You should have:

- 4 markdowns
- Slides

We will use:

- tidy models
- vegan and mass
- reticulate
- scikit learn + python
- UMAP

We have four sections:

- Motivation and overview
- PCA + ICA in tidymodels
- MDS non-tidyverse
- TSNE (python) e UMAP (R)

ABOUT ME

I am Isabella.

You can find me at:

@bisnotforbella

bellabf

https://www.somaquadrados.com/

GUIDELINES

This designed as a hands-on tutorial.

You are invited to ask me questions anytime.

I need feedback in order to adjust the speed, voice tone and etc.

This tutorial is in English but are free to ask me can questions in French, Spanish and Portuguese. I will try my best to translate them for everyone.

GOALS

To show how we can use R's power to do dimension reduction

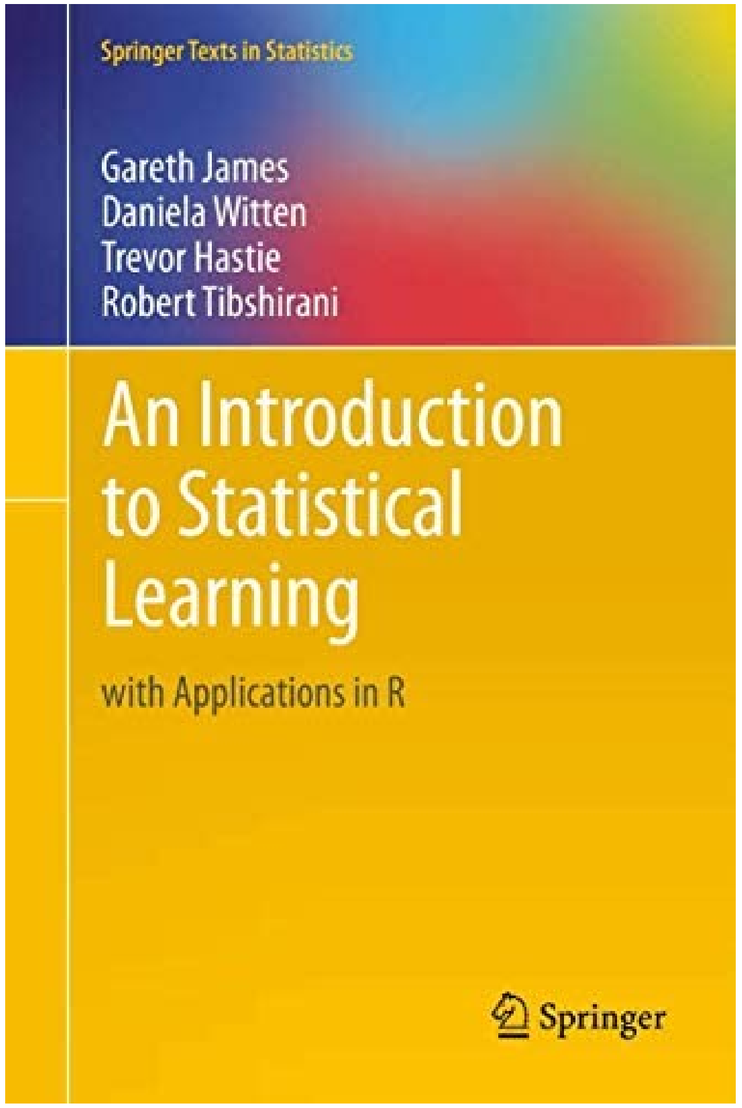
DISCLAIMERS

This is not an in-depth course on dimensional reduction
This is not an exhaustive list of reduction methods

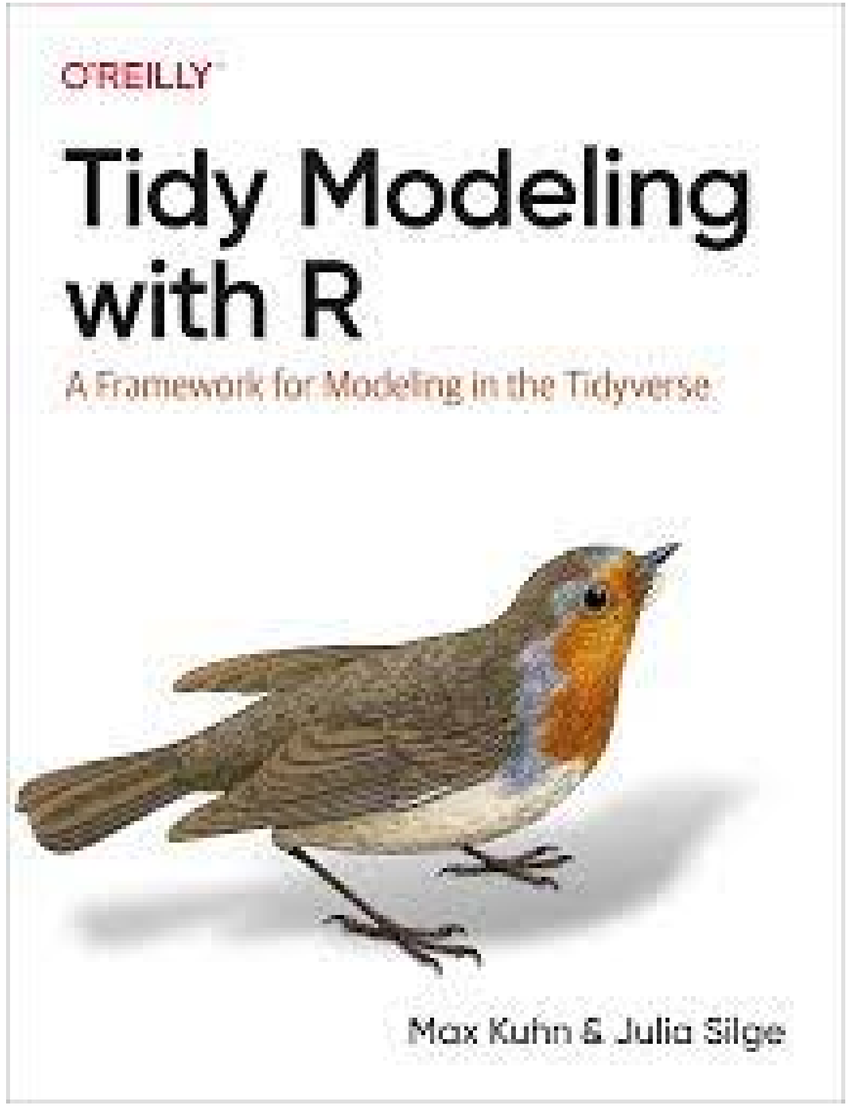
Mistakes are my own

TIME USAGE

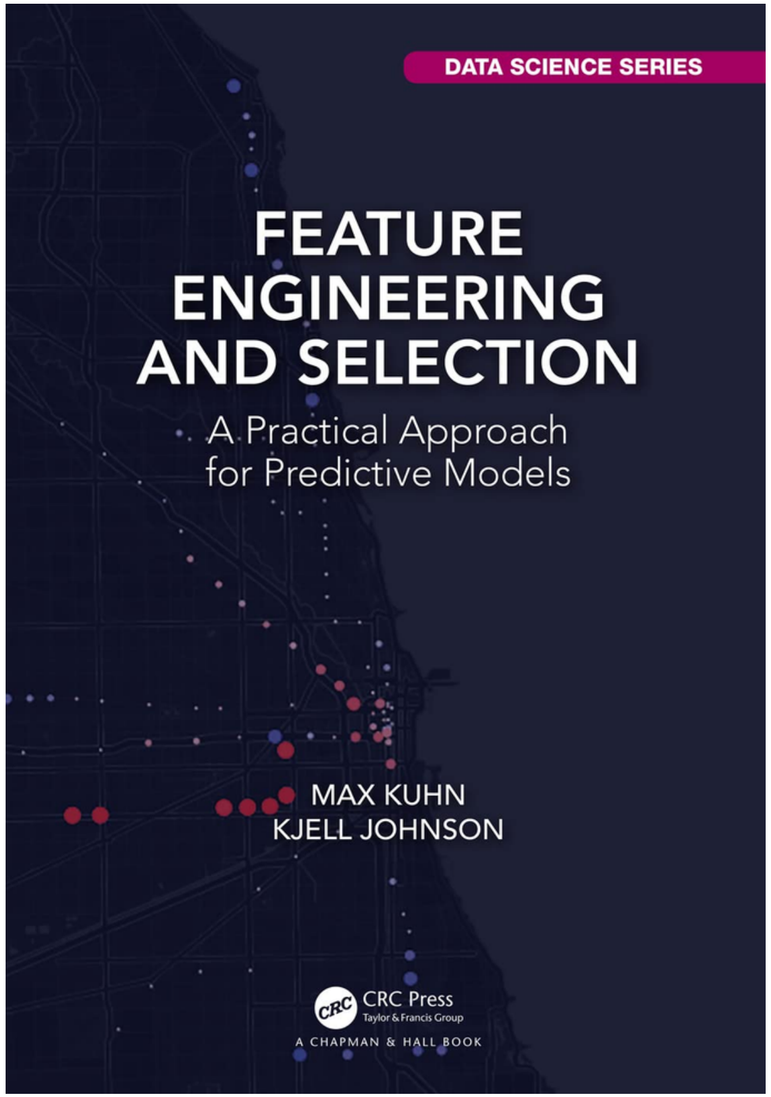
The tutorial is planned so we can have enough time to go through it at a reasonable pace.



https://www.statlearning.com/



https://www.tmwr.org/



http://www.feat.engineering/

SECTION 1

SECTION 1

what is dimension reduction?

why are we interested in it?

"The tension in working with modern data is that we often **do not fully know** which predictors are relevant to the response. What may in fact be true is that some of the available predictors (or combinations) have a meaningful association with the response. Under these conditions, the task is to try to identify the **simple or complex combination** that is relevant to the response."

http://www.feat.engineering/numeric-many-to-many.html

Dimension Reduction

Dimension Reduction

A transformation such as $M < P$

M: number of transformed predictors

P: number of predictors

N: number of observations

Dimension Reduction

A transformation such as M < P

M: number of transformed predictors

P: number of predictors

N: number of observations

If M = P, all predictors are independent, and no reduction occurs.

If a dataset P >> N is said to be high dimensional.

Why we would like to do that?

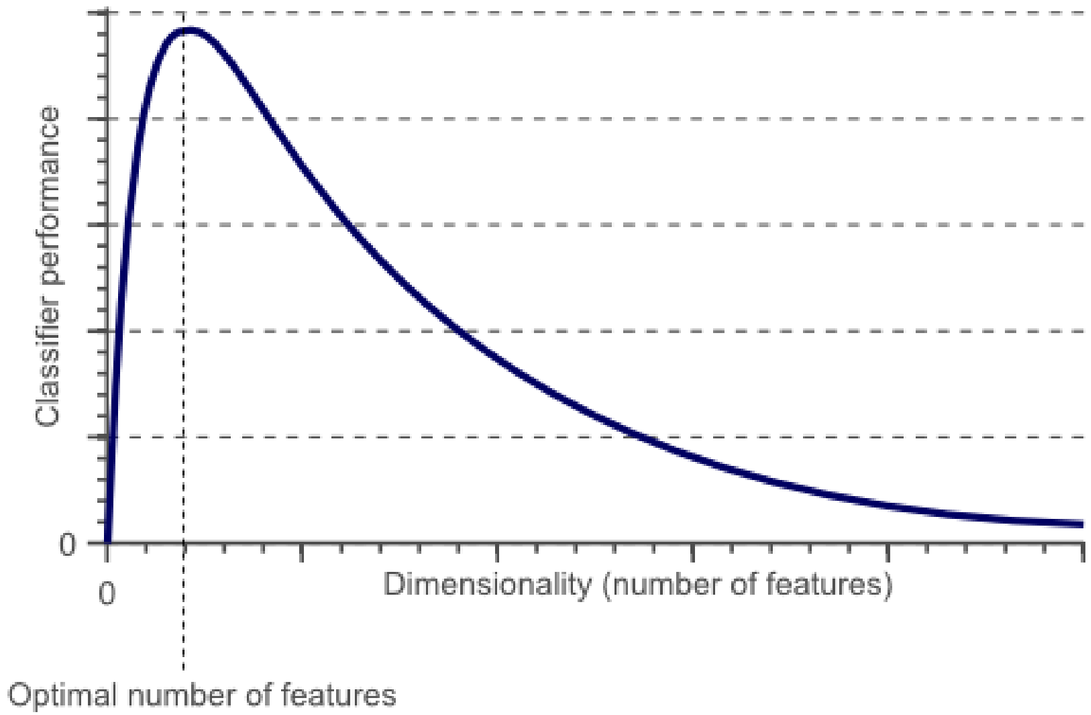
- provide tools for re-representing predictors

Why we would like to do that?

- provide tools for re-representing predictors
- predictive modeling:
  - some models (such as support vector machines and neural networks) are sensitive to irrelevant predictors
  - extra predictors might sink predictive performance
  - models like linear or logistic regression are vulnerable to correlated predictor

Curse of dimensionality

- As the dimensionality increases, space increases so fast that the data becomes too sparse.
- It makes it harder to distinguish noise from the true signal.
- Performance increases and then steadily decreases (peaking phenomenon)



https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

Why we would like to do that?

- provide tools for re-representing predictors
- predictive modeling:
  - some models (such as support vector machines and neural networks) are sensitive to irrelevant predictors
  - extra predictors might sink predictive performance
  - models like linear or logistic regression are vulnerable to correlated predictor
- keep the minimum possible set of predictors that provides acceptable results

Why we would like to do that?

- provide tools for re-representing predictors
- predictive modeling:
  - some models (such as support vector machines and neural networks) are sensitive to irrelevant predictors
  - extra predictors might sink predictive performance
  - models like linear or logistic regression are vulnerable to correlated predictor
- keep the minimum possible set of predictors that provides acceptable results
- it might reduce the cost of acquiring data

Why we would like to do that?

- provide tools for re-representing predictors
- predictive modeling:
  - some models (such as support vector machines and neural networks) are sensitive to irrelevant predictors
  - extra predictors might sink predictive performance
  - models like linear or logistic regression are vulnerable to correlated predictor
- keep the minimum possible set of predictors that provides acceptable results
- it might reduce the cost of acquiring data
- it might facilitate storage

Transformations

Transformations

Feature selection

- selecting a subset of relevant features for use in model construction

just filtering out uninformative predictors might not be enough

Transformations

Feature selection

- selecting a subset of relevant features for use in model construction

Feature extraction

- **Initial feature → transformed feature.**
- The transformed feature has to be informative

Transformations

Feature selection

- selecting a subset of relevant features for use in model construction

Feature extraction

- **Initial feature → transformed feature.**
- The transformed feature has to be informative

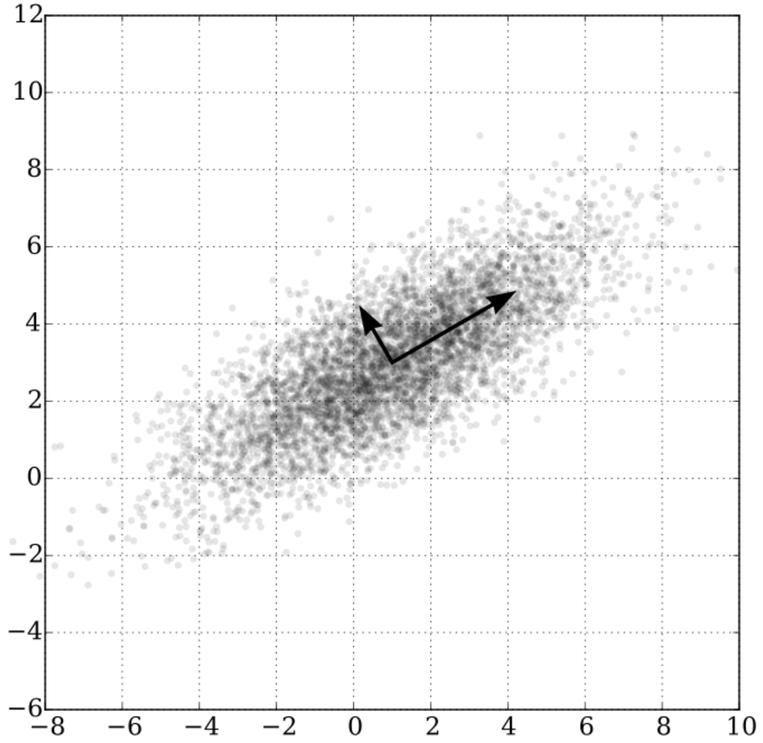
Dimensionality reduction

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties ...

Wikipedia / Jun 5

Principal Component Analysis (PCA)

- A sequence of w vectors, where the i-th vector is the direction of a line that best fits the data while being **orthogonal** to previous vectors.
- It effectively performs a change of basis in the data by minimizing the average distance between the points to the line
- Each w vector is the equivalent of the **eigenvector** of the **covariance** matrix.
- For classification, a PCA plot show clear separation we can assume that a linear classifier can be used



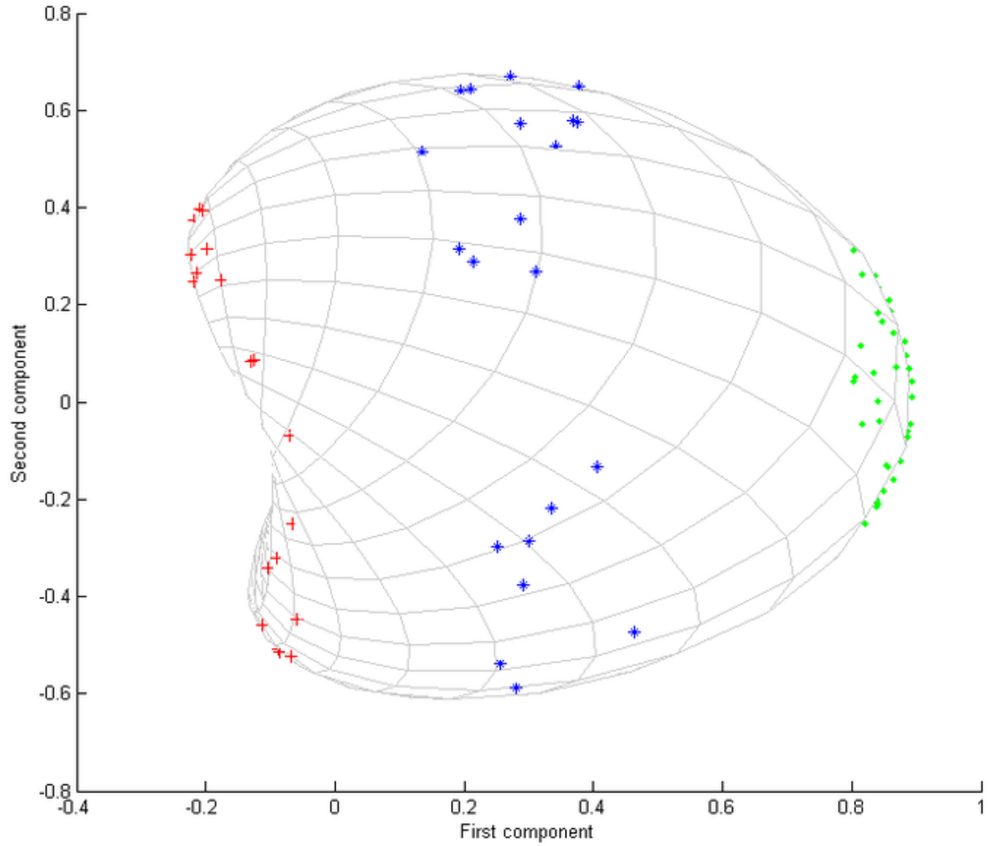
Euclidean
Orthogonal - uncorrelated
Unique analytical solution
Linear

Kernel PCA

Perform PCA reproducing kernel Hilbert space with the help of kernel basis expansion

Intuition:

1. the kernel function map our data to a higher dimensional space where data is linearly separable
2. do PCA in that space.



Independent Component Analysis

- A computational method for separating a multivariate signal into additive subcomponents
- It assumes that at most one subcomponent is a non-Gaussian signal and that the subcomponents are statistically independent of each other.

Independent Component Analysis

- A computational method for separating a multivariate signal into additive subcomponents
- It assumes that at most one subcomponent is a non-Gaussian signal and that the subcomponents are statistically independent of each other.

Cocktail party problem

Independent Component Analysis

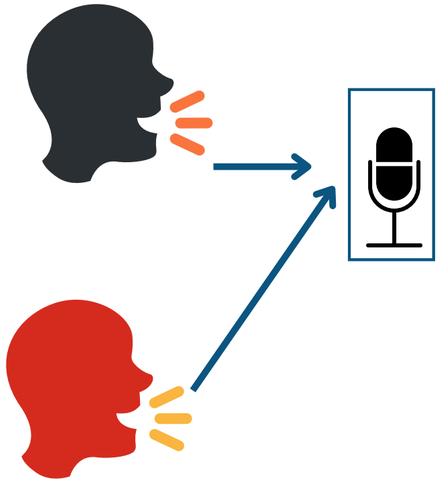
- A computational method for separating a multivariate signal into additive subcomponents
- It assumes that at most one subcomponent is a non-Gaussian signal and that the subcomponents are statistically independent of each other.

Cocktail party problem

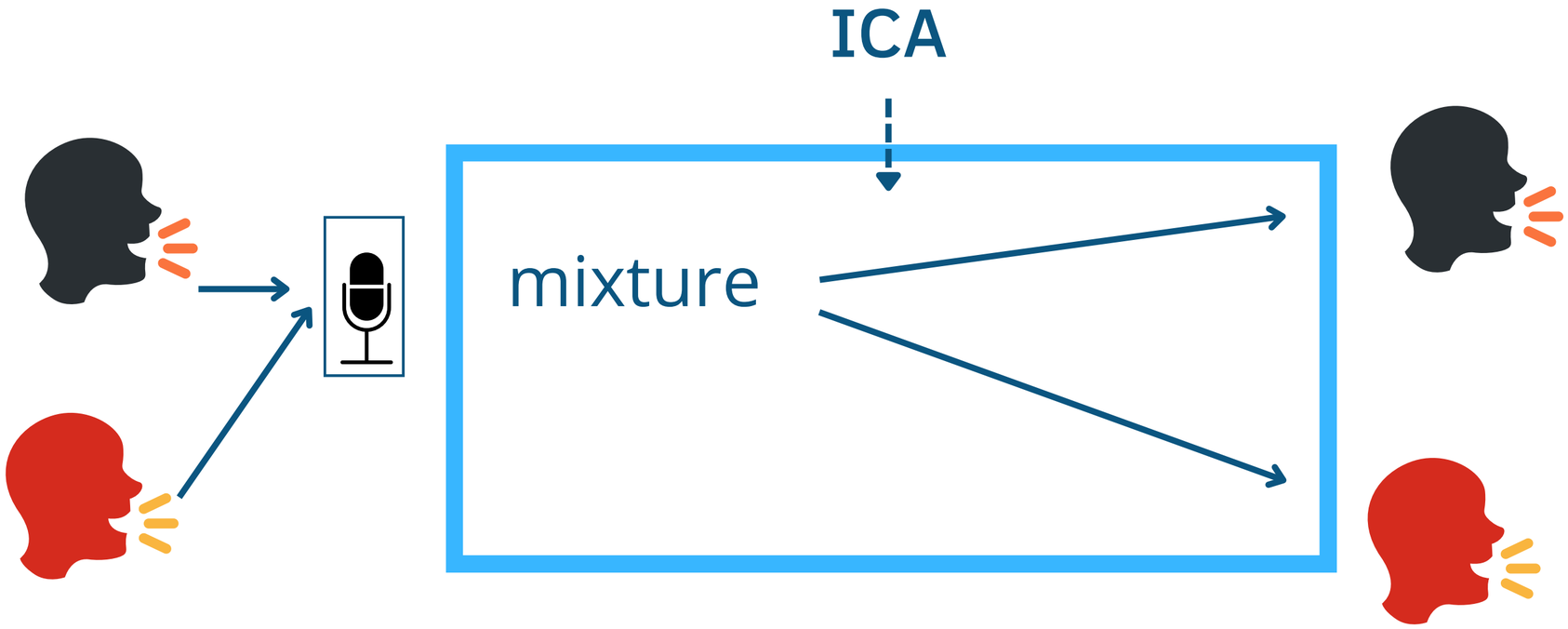
Independent Component Analysis

- A computational method for separating a multivariate signal into additive subcomponents
- It assumes that at most one subcomponent is a non-Gaussian signal and that the subcomponents are statistically independent of each other.

Cocktail party problem



How do you identify different people talking?


ICA
mixture

ICA

Additively separable components

separates information

"non-gaussianity"

non-orthogonal

PCA

successive approximations

compress information

variance

orthogonal

Multidimensional Scaling (MDS)

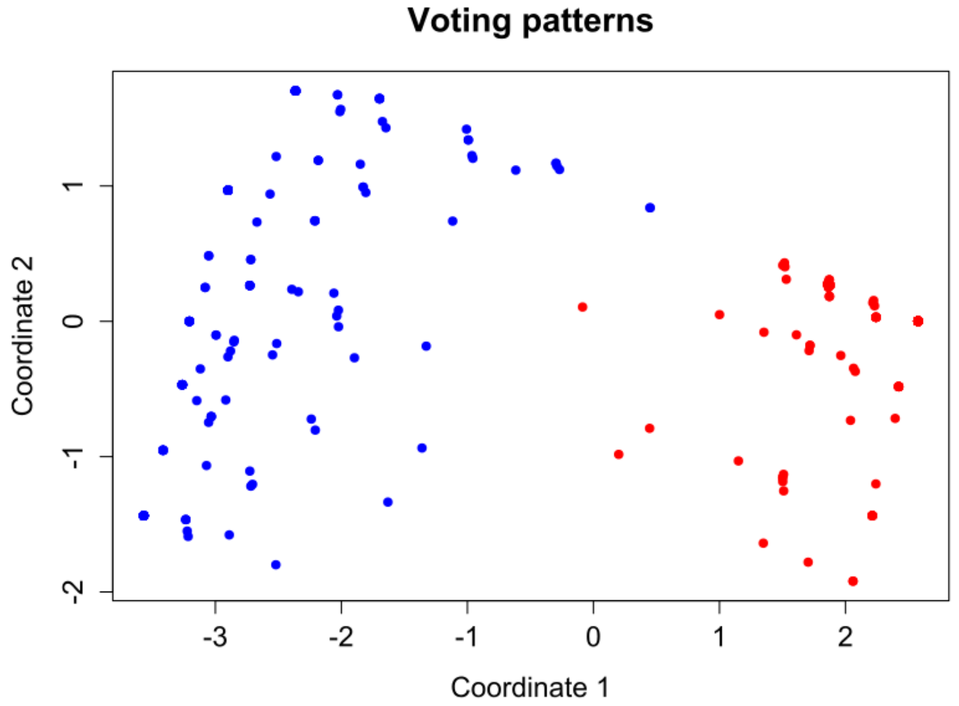
- Non-linear dimensionality reduction
- Preserves pairwise distance
- Metric and non-metric solutions
- Several loss functions

If metric, it has a couple of different aliases:
Principal Coordinate Analysis, Classical MDS

Advantages:

- non-metric distance
- tolerance for missing data
- asymmetrical distance matrix.

Support non-metric distances
At least $O(N^2)$



Metric

$$\text{Stress}_D(x_1, x_2, \ldots, x_N) = \sqrt{\sum_{i \neq j = 1,\ldots,N} \left(d_{ij} - \|x_i - x_j\|\right)^2}.$$

residual sum of squares.

parametric in nature

if euclidean, an exact solution exists (eigenvalues and vectors)

Procedure

1. Assign a number of points to coordinates in n-dimensional space
2. Calculate distances for all pairs of points (similarity matrix)
3. Compare the similarity matrix with the original input matrix by evaluating the stress function.
4. Adjust coordinates, if necessary, to minimize stress.

Non-metric

$$\text{Stress} = \sqrt{\frac{\sum \left(f(x) - d\right)^2}{\sum d^2}}.$$

f(x) is a isotonic regression

non-parametric

Procedure

1) optimal monotonic transformation of the proximity

2) arranged the points so that their distance closely matches the scaled proximity

Ecology

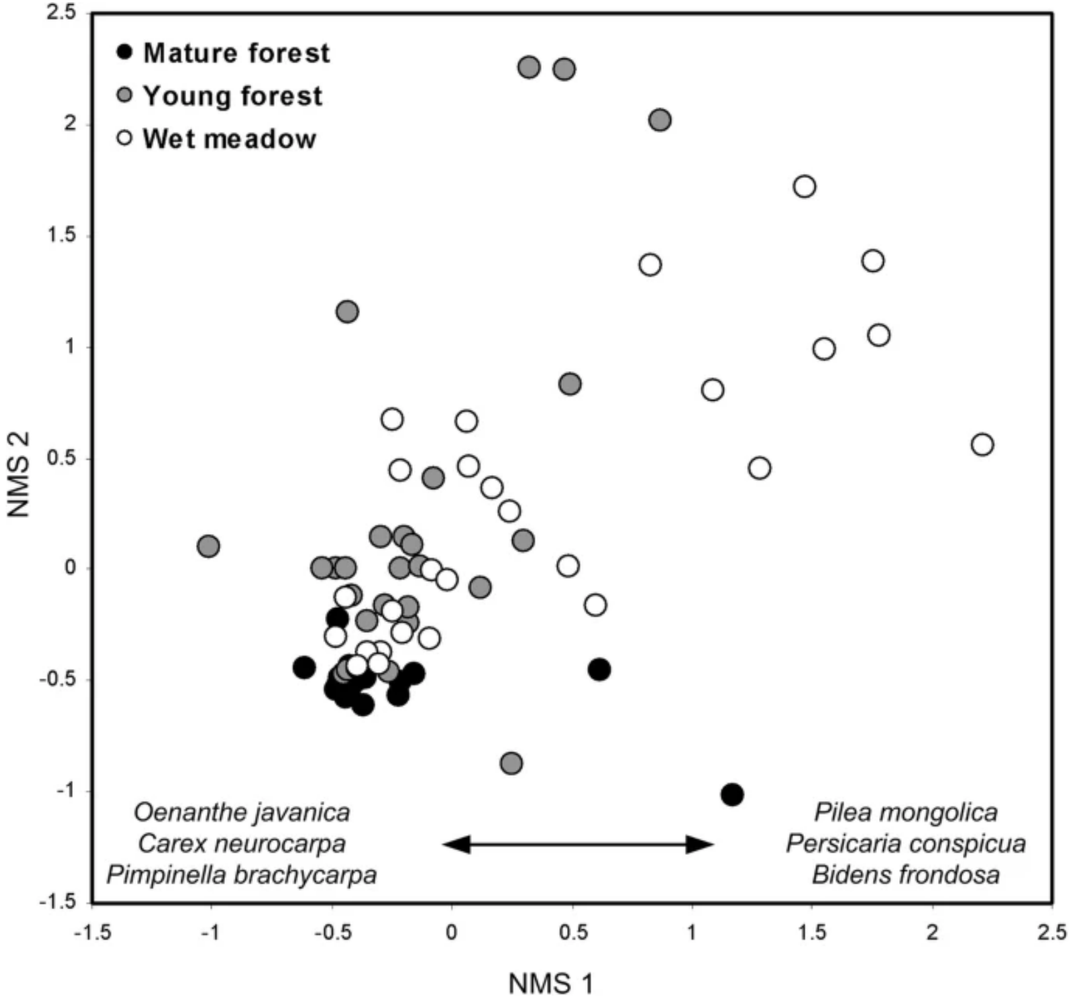
"We sought to understand the relationship between the seed bank and vegetation in abandoned rice paddies in South Korea, in order to guide management of these sites"

Floristic composition and species richness of soil seed bank in three abandoned rice paddies along a seral gradient in...

Background We sought to understand the relationship between the seed bank and vegetation in abandoned rice paddies in South Korea, in order to guide management of these sites. We investigated the floristic composition and species richness of the soil seed bank and ground vegetation in former paddies along thr...

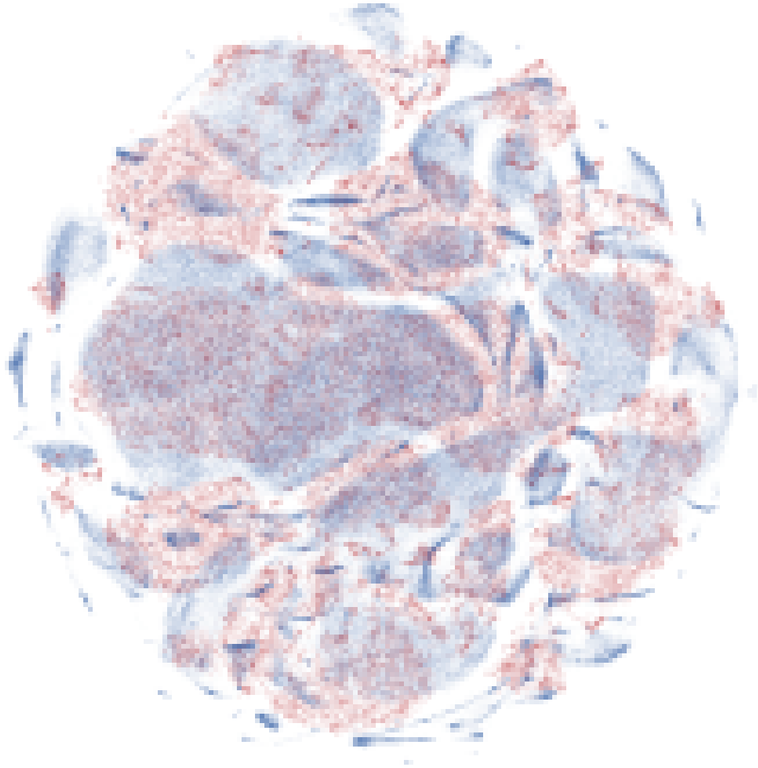
BioMed Central / Jul 3, 2018

https://doi.org/10.1186/s41610-018-0074-3



t-SNE

- It converts high-dimensional Euclidean distances between data points → similar objects are assigned a higher probability while dissimilar points are assigned a lower one
- Subsequently, it defines a t-student distribution in the low-dimensional map
- It tries to minimize the difference between the two probability distributions



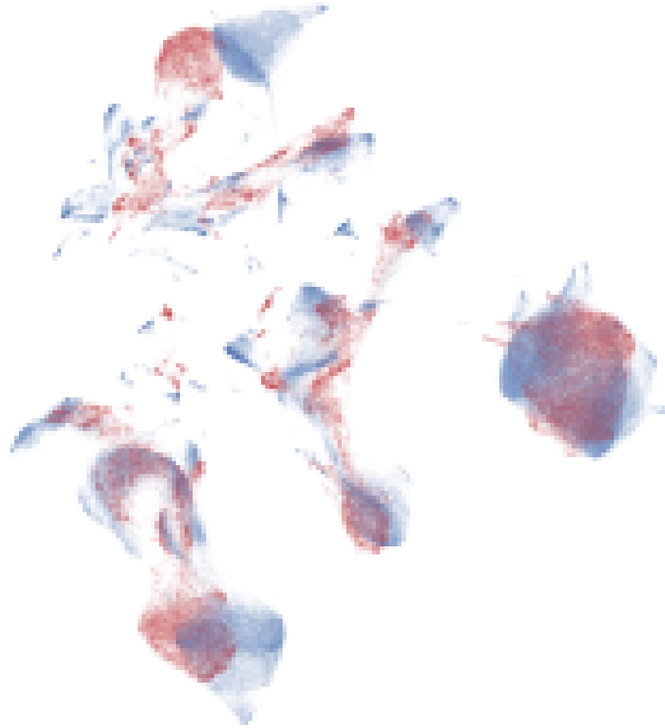
https://arxiv.org/pdf/1802.03426.pdf

Euclidean distance, but can be changed
t-SNE is a non-linear
does not preserve global structure well

Uniform manifold approximation and projection (U-MAP)

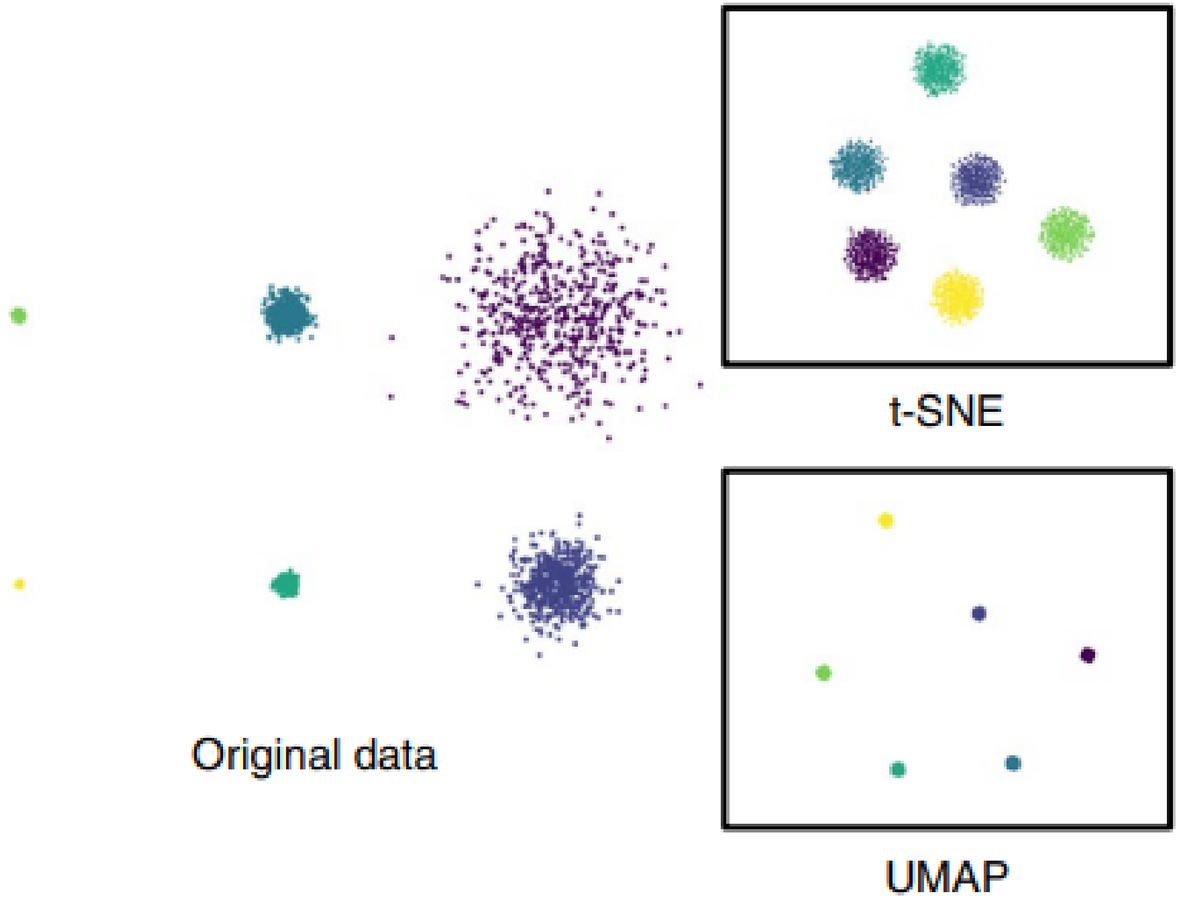
- It constructs a fuzzy topological representation of high-dimensional data by changing the low-dimensional embedding until it becomes more similar to the original data
- It is less computational expensive while still preserving more of the global structure compared to t-SNE

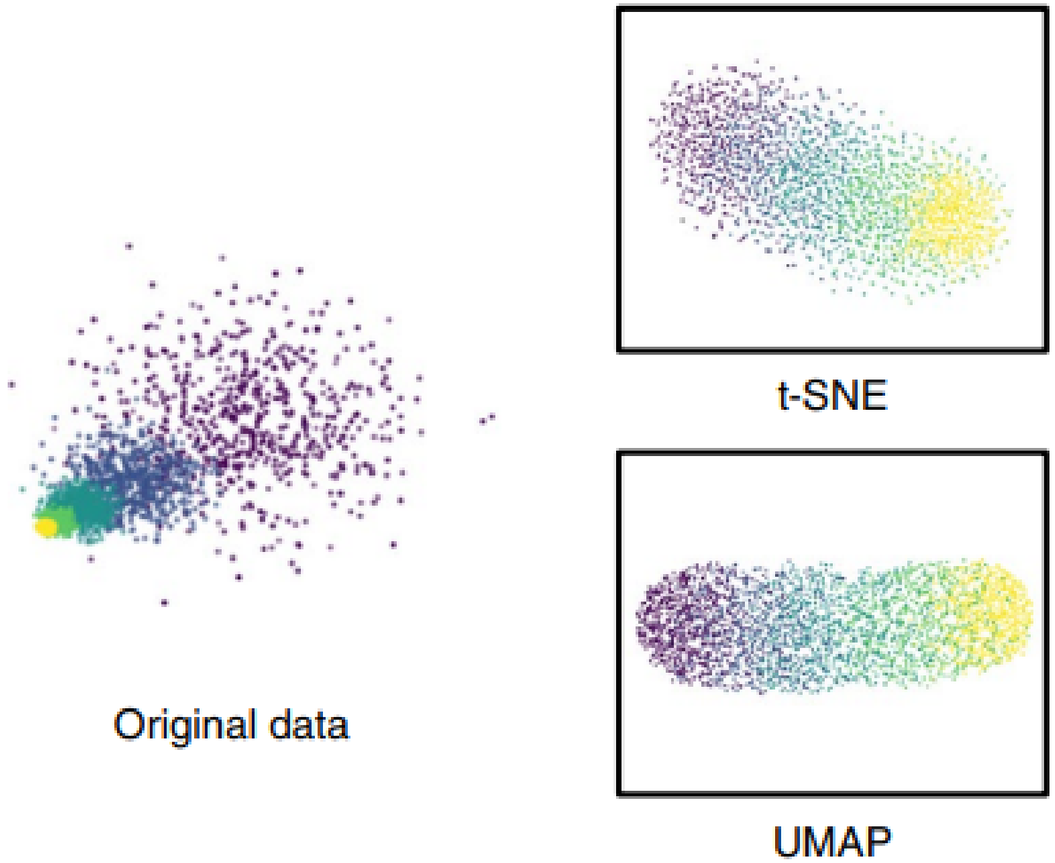
- Assumptions:
  - The data is uniformly distributed on the Riemannian manifold.
  - The Riemannian metric is locally constant (or can be approximated as such).
  - The manifold is locally connected.



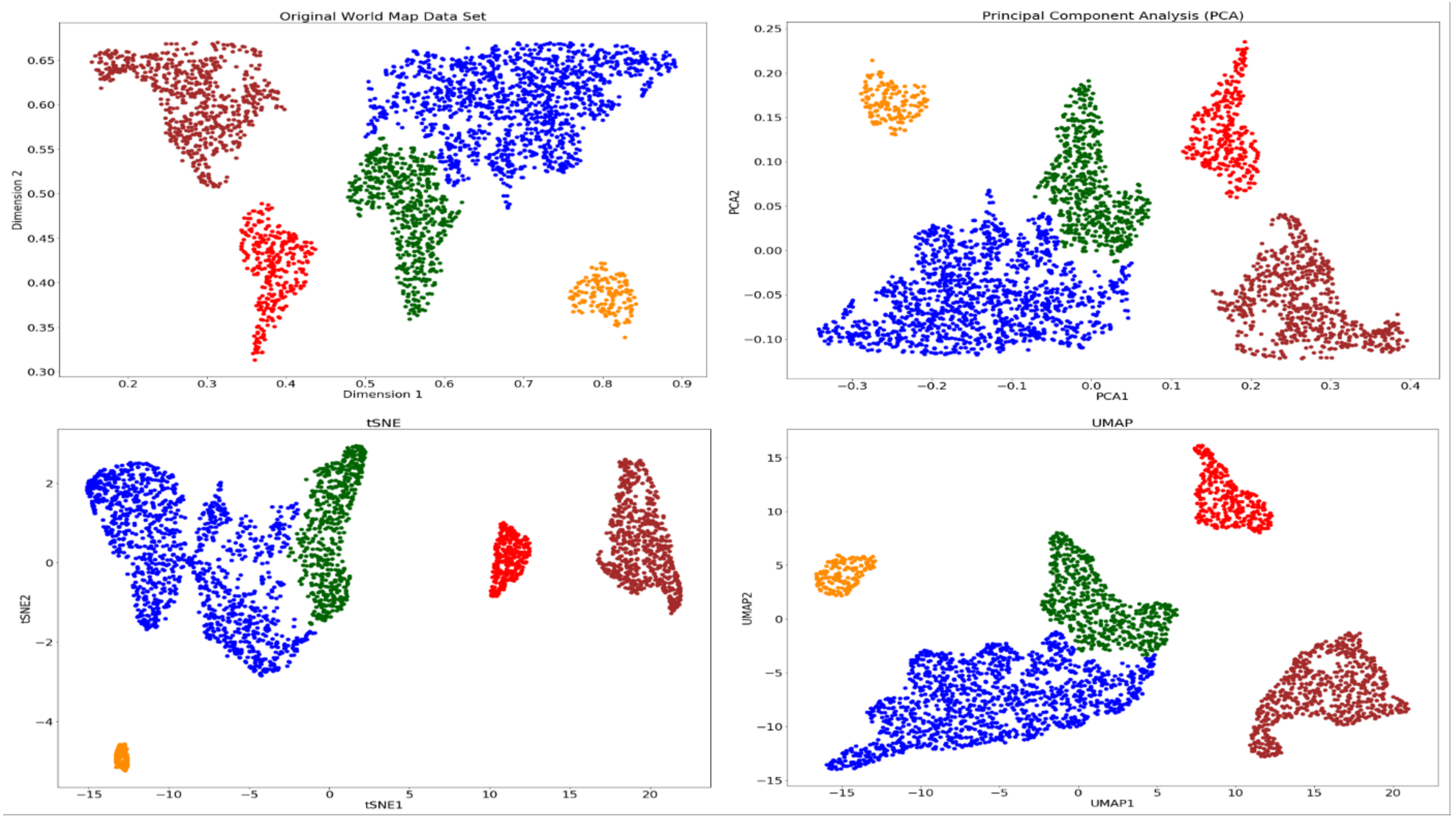
https://arxiv.org/pdf/1802.03426.pdf

geometry based
non-linear
preserve global structure





doi:10.1038/s41587-020-00801-7



https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Molecular Biology

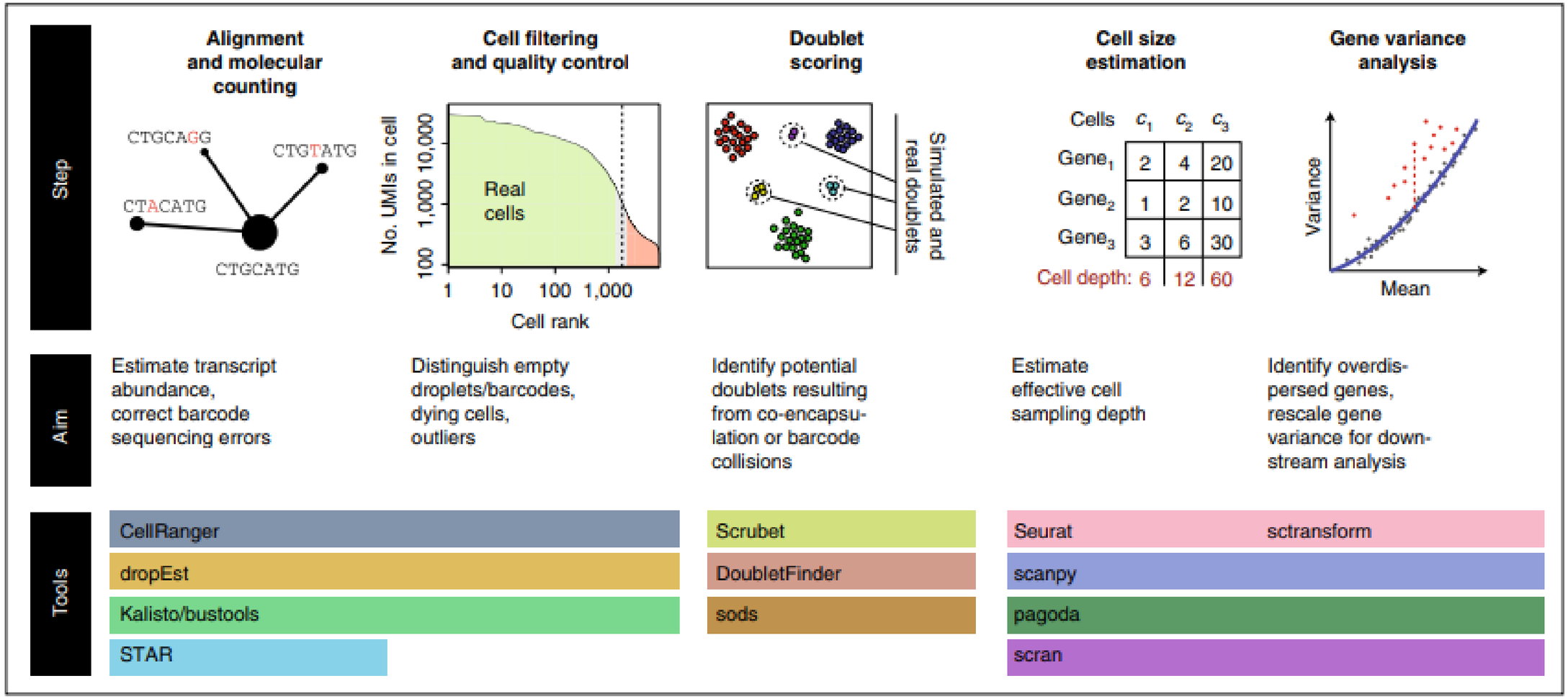


Fig. 1 | Key preprocessing steps in single-cell RNA-seq analysis. The diagram shows major steps involved in analysis of individual scRNA-seq datasets, along with the relevant software tools (see Box 1 and Luecken et al.[86] for other practical considerations).

10.1038/s41592-021-01171-x

Molecular Biology

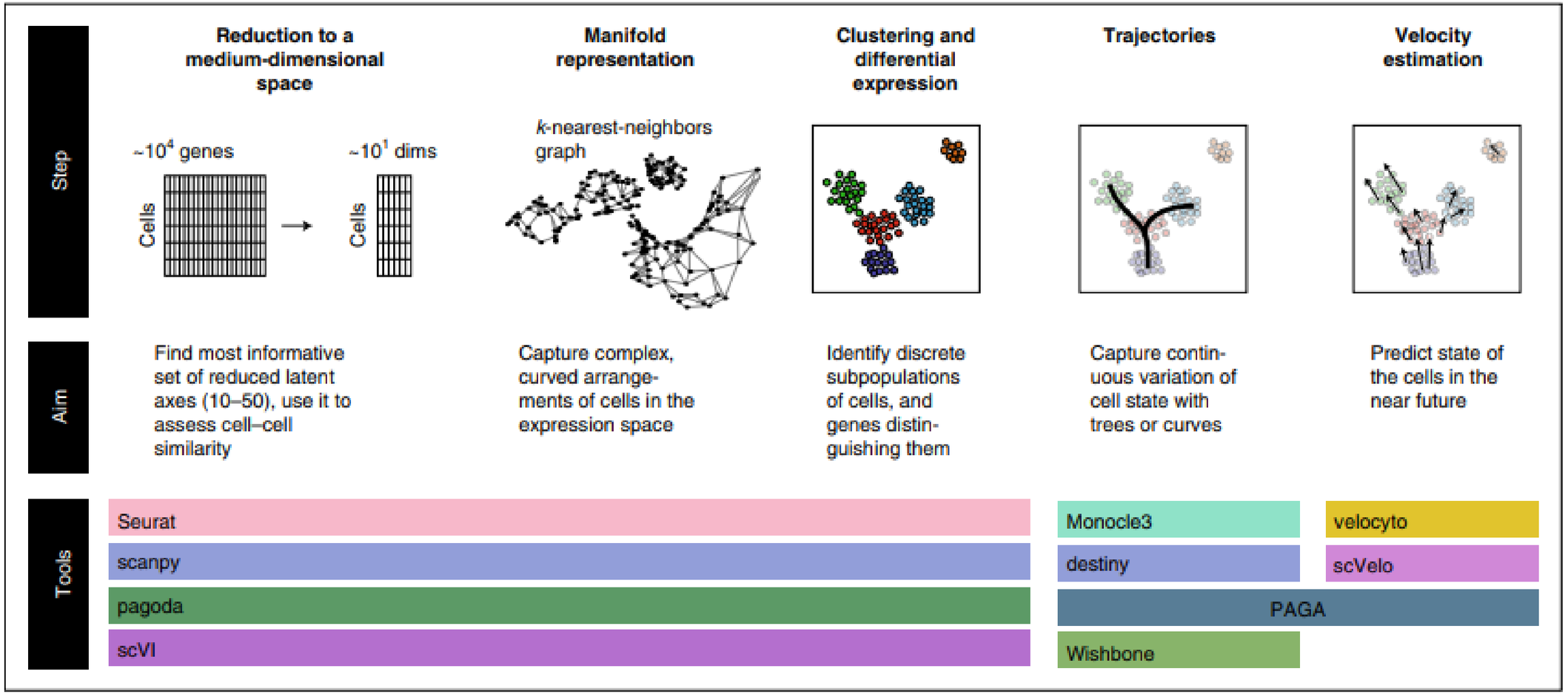
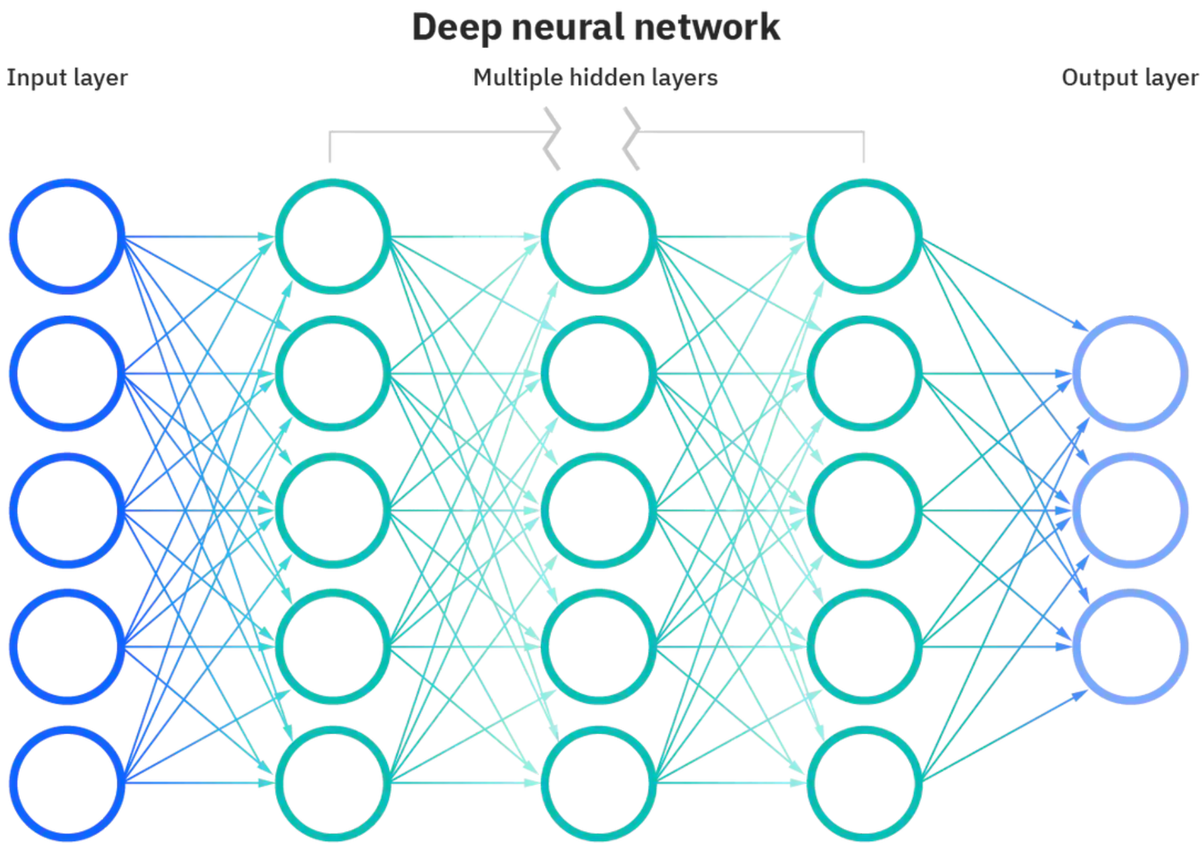
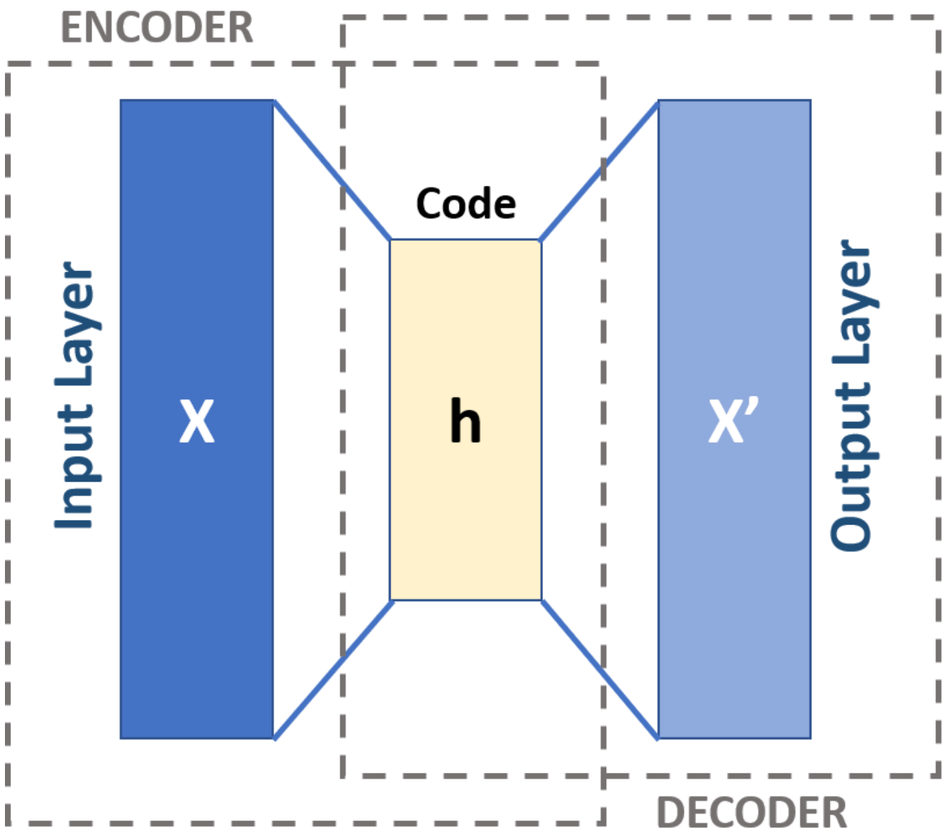


Fig. 2 | Key analysis steps in single-cell RNA-seq analysis. The diagram shows major steps involved in analysis of individual scRNA-seq datasets, along with the relevant software tools (see Box 1 and Luecken et al.[86] for other practical considerations).

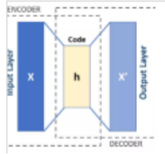
10.1038/s41592-021-01171-x





What are Neural Networks?

Learn about neural networks that allow programs to recognize patterns and solve common problems in artificial intelligence, machine learning and deep...



Papers with Code - AutoEncoder Explained

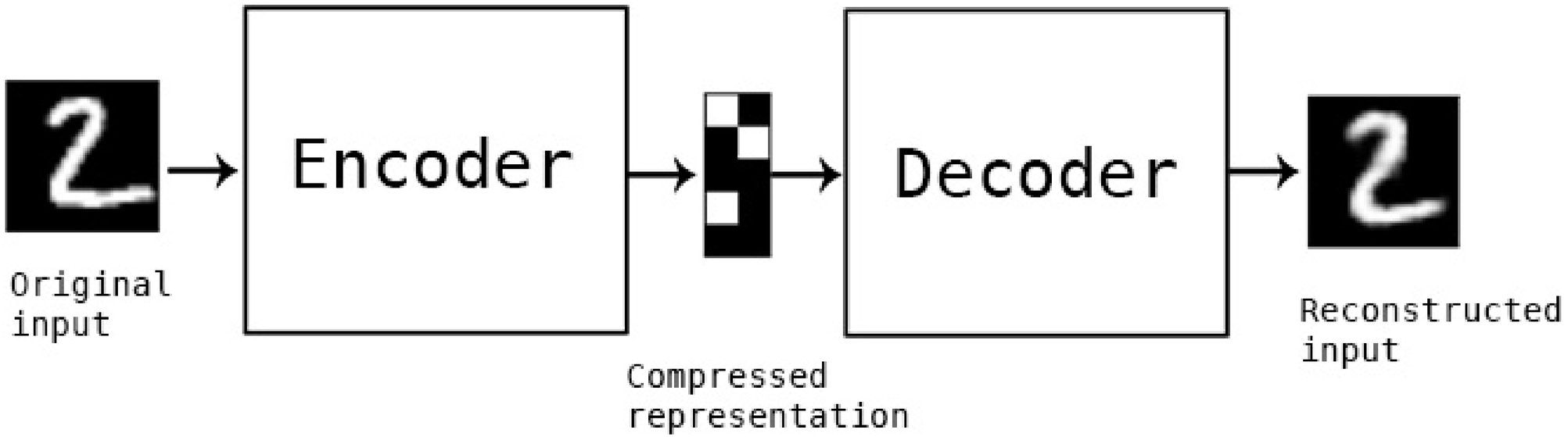
An Autoencoder is a bottleneck...

paperswithcode.com

Computer Vision

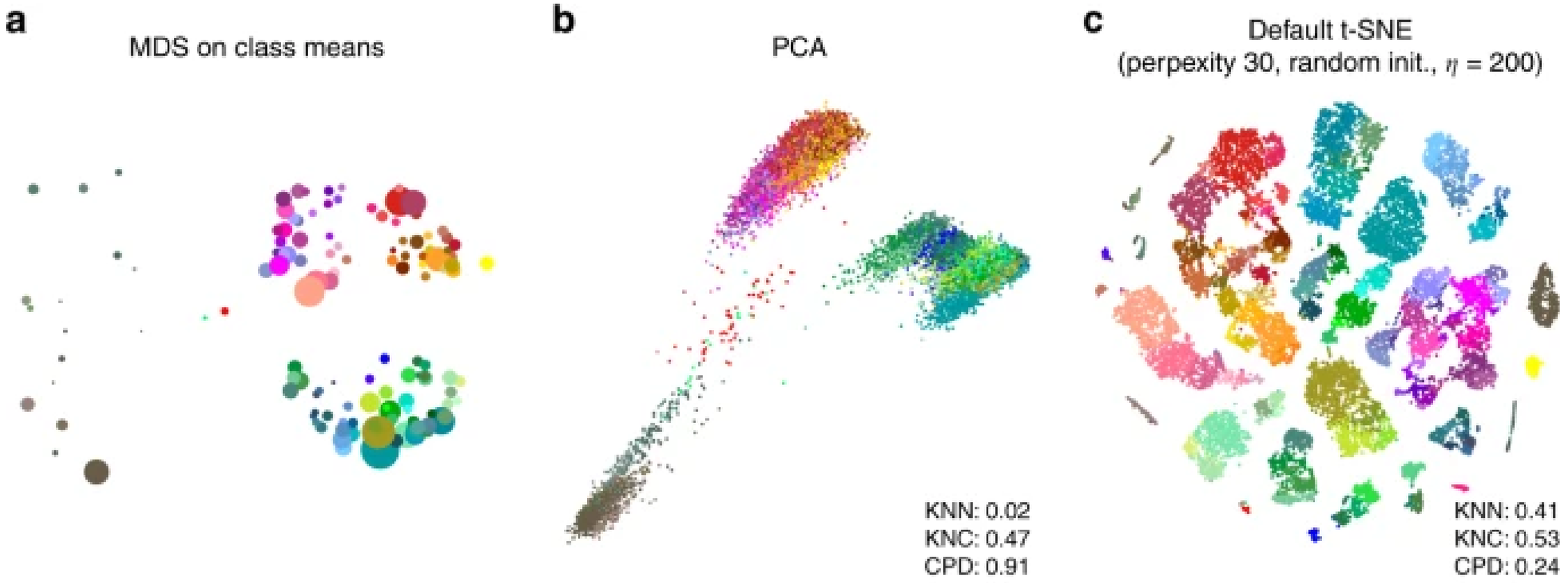
https://blog.keras.io/building-autoencoders-in-keras.html

Computer Vision



https://blog.keras.io/building-autoencoders-in-keras.html

Finally



https://doi.org/10.1038/s41467-019-13056-x

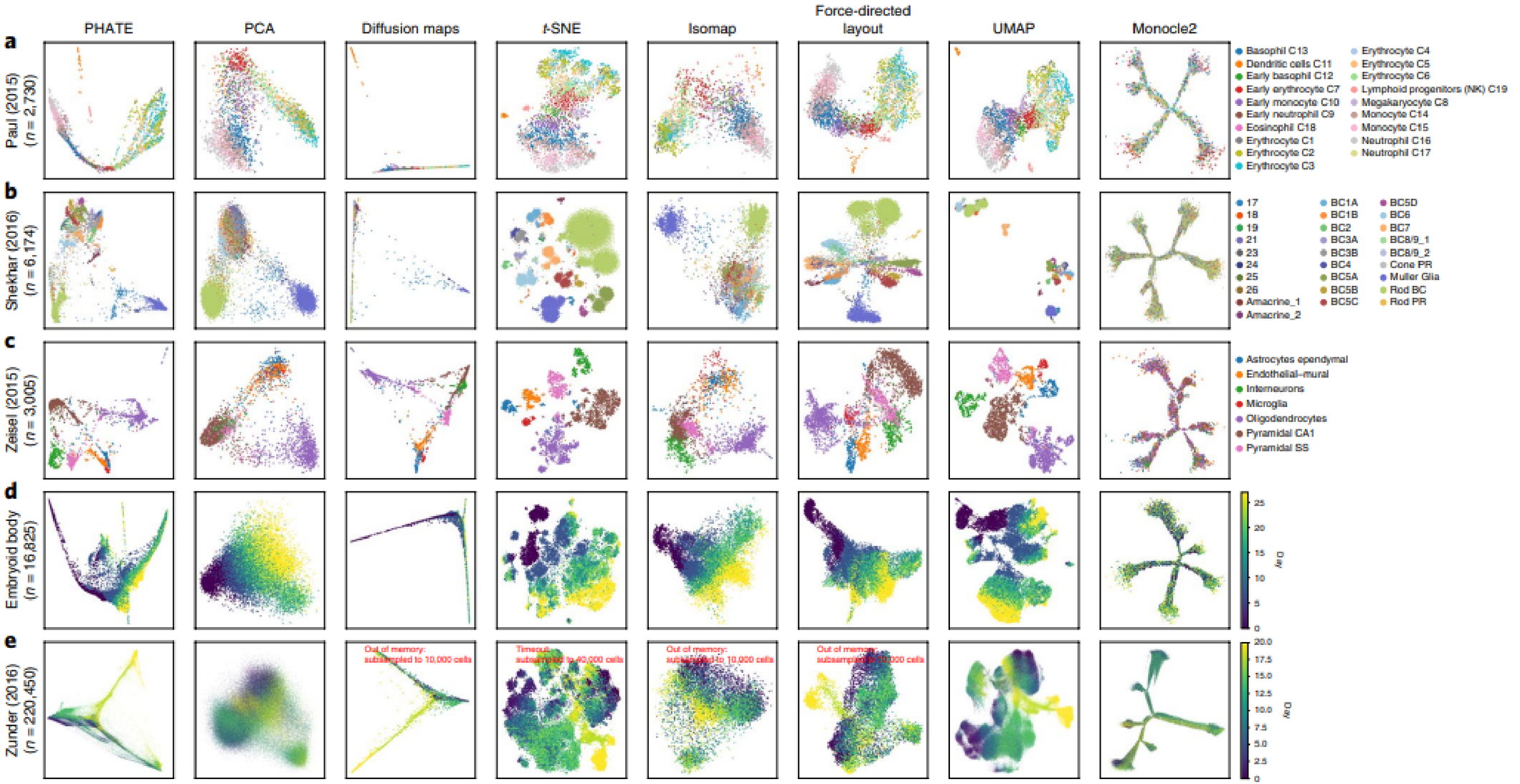


Fig. 5 | Comparison of PHATE to other visualization methods on biological datasets. Columns represent different visualization methods, rows

https://doi.org/10.1038/s41587-019-0336-3

Some takeaways

- Dimension reduction is hard
- Cross-validation might help
- As novel methodologies arise, so do novel data science techniques: we must keep updated with the field
- There is no superiority of one methodology (no free lunch theorem)
- Dimension reduction is context-dependent: it should fit both the (biological) question at hand and the structure of the data itself
- Benchmark plays a role in that case

QUESTIONS?

DATA

BREAK?

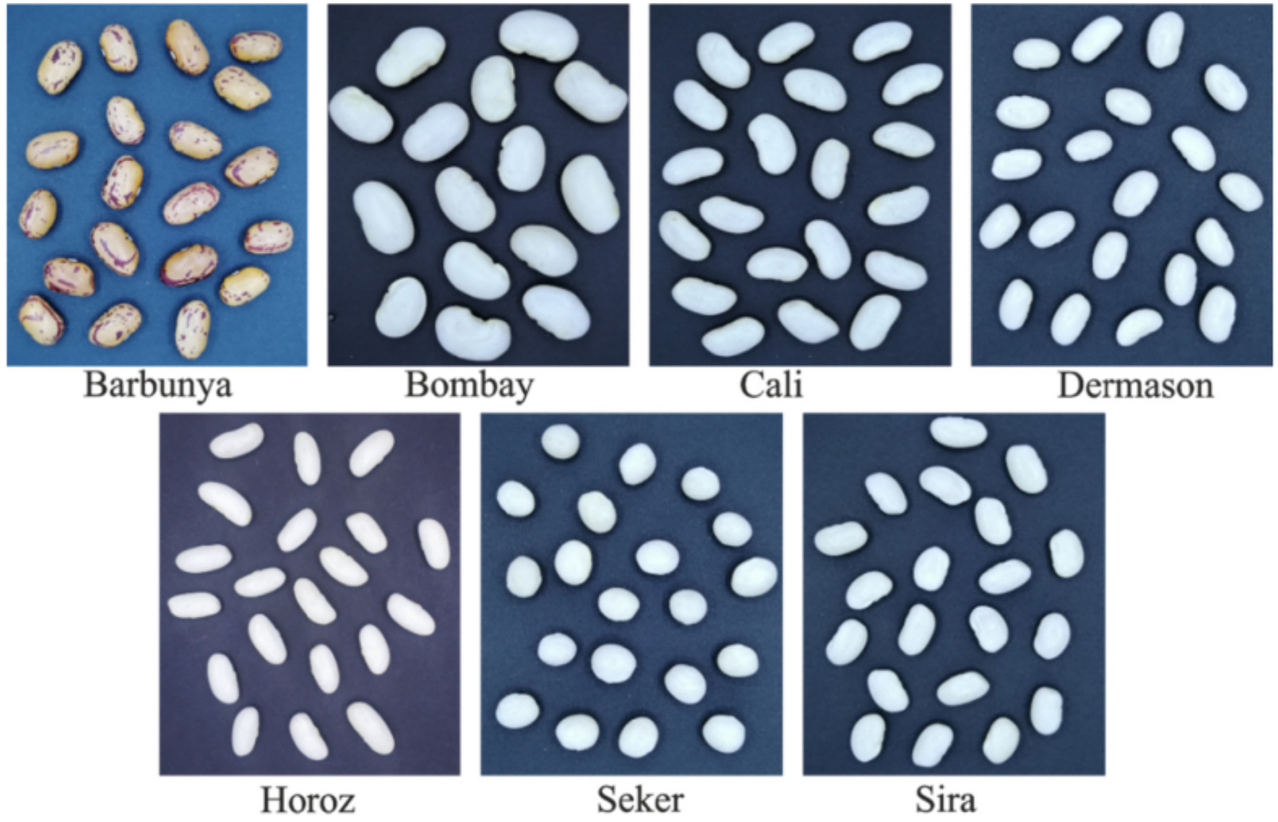
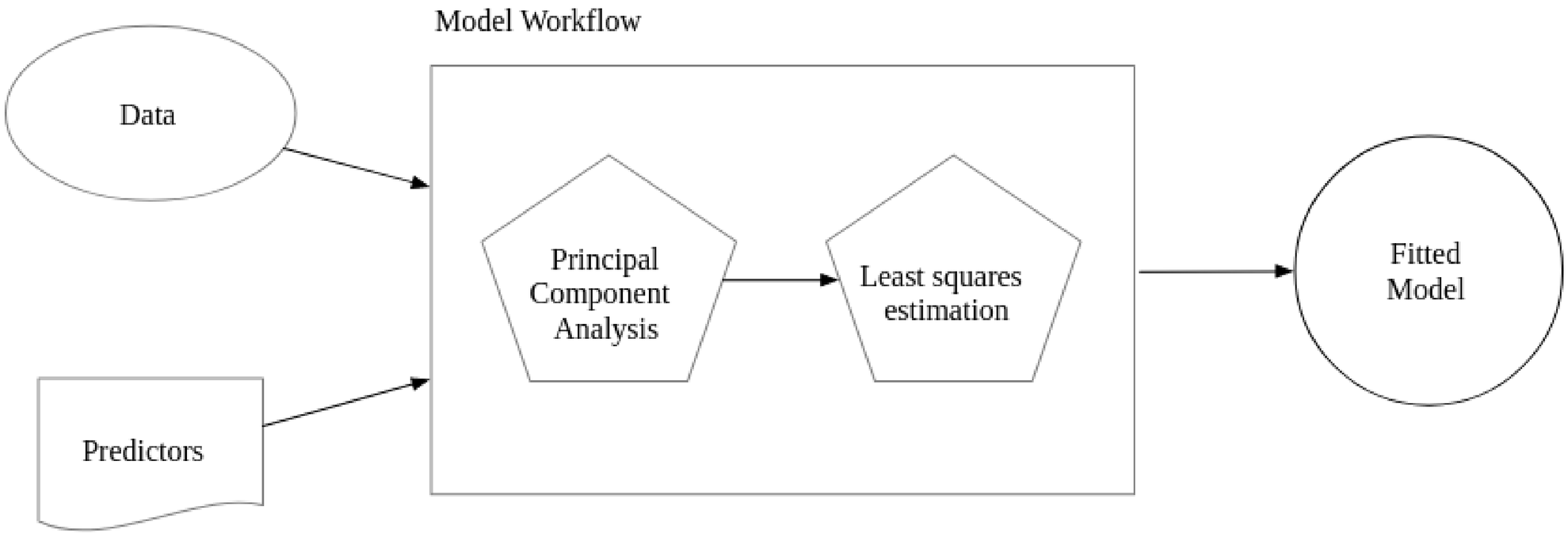


Fig. 3. Sample of taken dry bean images.

area, perimeter, major axis length, minor axis length, aspect ratio, eccentricity, convex area, equiv diameter, extent, solidity, roundness, compactness, shape factor 1, shape factor 2, shape factor 3, and shape factor 4

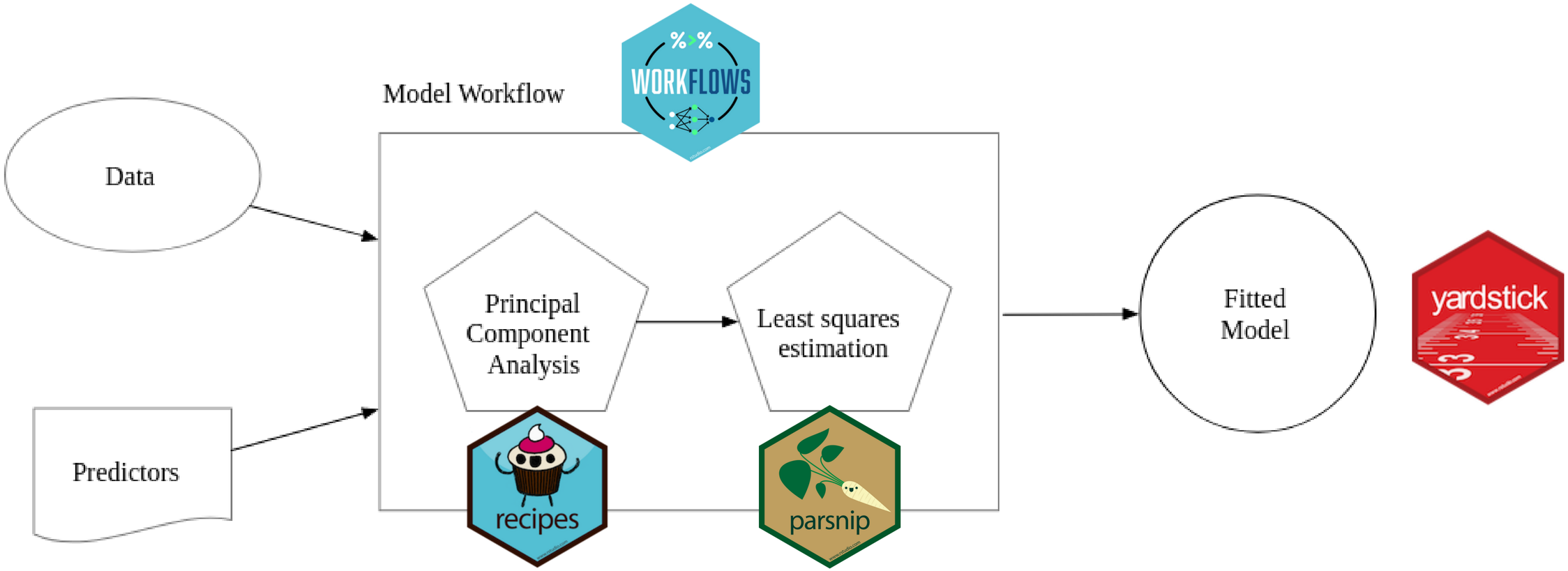
TIDY MODELS

Tidymodels



https://www.tmwr.org/workflows.html

Tidymodels



https://www.tmwr.org/workflows.html

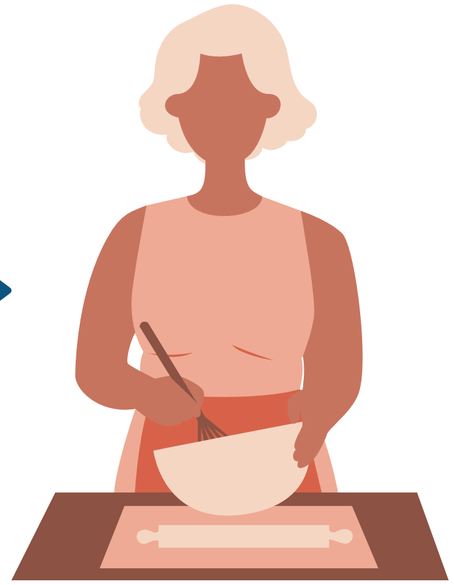
BREAK

SECTION 2

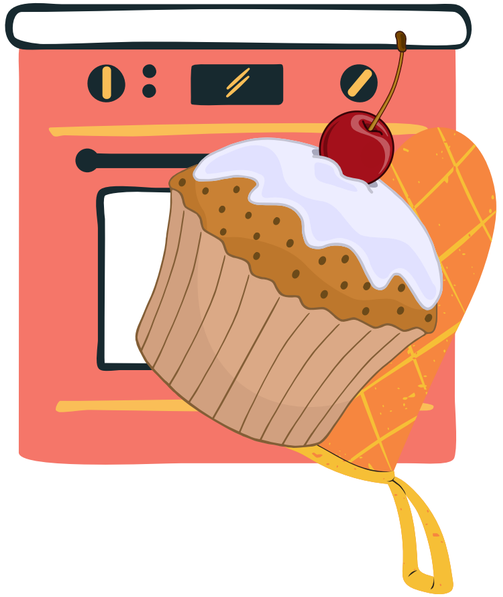
PCA, ICA

GOAL: PCA + ICA in tidymodels

recipe

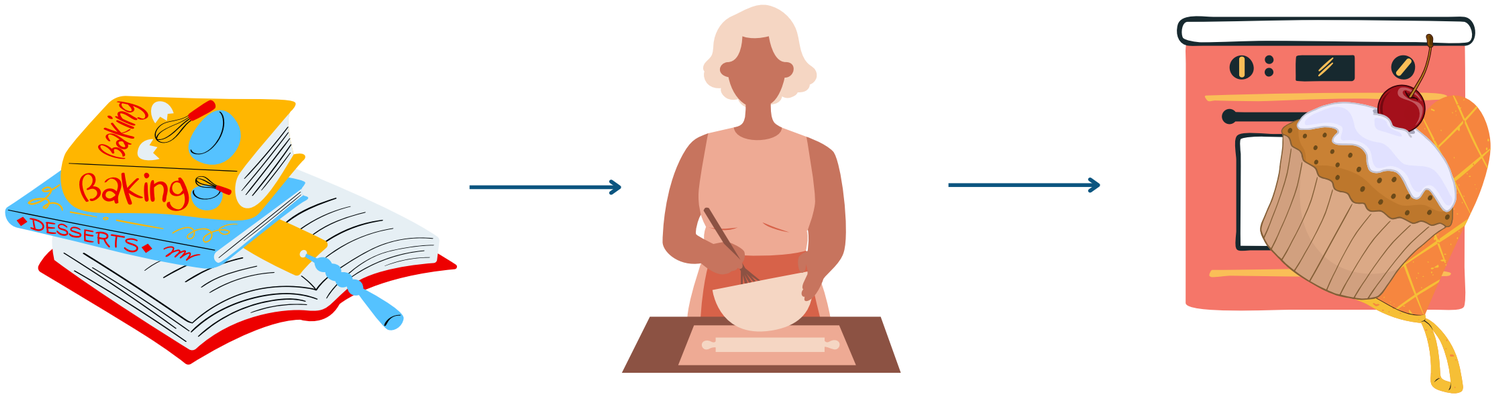


prep



bake

GOAL: PCA + ICA in tidymodels



recipe() → prep() → bake()

Defines the preprocessing

(returns a recipe)

Calculates statistics from the training set

Analogous to fit()

(returns a recipe)

Applies the preprocessing to data sets

Analogous to predict()

(returns a tibble)

https://www.tmwr.org/dimensionality.html

BREAK

SECTION 3

Multdimensional Scaling

DATA

These algorithms are more computationally expensive, so we are going to use another dataset

DATA

These algorithms are more computationally expensive, so we are going to use another dataset

Trust me on this one :)

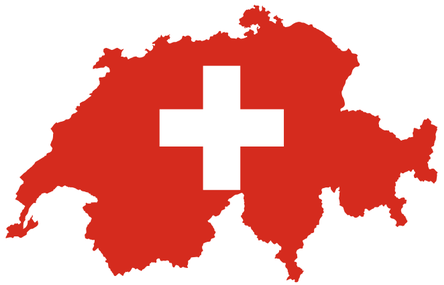
Swiss Fertility and Socioeconomic Indicators (1888) Data

Description

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Usage

```
swiss
```



Format

A data frame with 47 observations on 6 variables, *each* of which is in percent, i.e., in $[0, 100]$.

[,1] Fertility	I_g, 'common standardized fertility measure'
[,2] Agriculture	% of males involved in agriculture as occupation
[,3] Examination	% draftees receiving highest mark on army examination
[,4] Education	% education beyond primary school for draftees.
[,5] Catholic	% 'catholic' (as opposed to 'protestant').
[,6] Infant.Mortality	live births who live less than 1 year.

All variables but 'Fertility' give proportions of the population.

Details

(paraphrasing Mosteller and Tukey):

Switzerland, in 1888, was entering a period known as the *demographic transition*; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.

The data collected are for 47 French-speaking "provinces" at about 1888.

BREAK

SECTION 4

T-SNE, UMAP

SECTION 4

what if I am interested in something that was developed in python?

vs

+

why choose when you can combine both

Type Conversions

When calling into Python, R data types are automatically converted to their equivalent Python types. When values are returned from Python to R they are converted back to R types. Types are converted as follows:

R	Python	Examples
Single-element vector	Scalar	`1`, `1L`, `TRUE`, `"foo"`
Multi-element vector	List	`c(1.0, 2.0, 3.0)`, `c(1L, 2L, 3L)`
List of multiple types	Tuple	`list(1L, TRUE, "foo")`
Named list	Dict	`list(a = 1L, b = 2.0)`, `dict(x = x_data)`
Matrix/Array	NumPy ndarray	`matrix(c(1,2,3,4), nrow = 2, ncol = 2)`
Data Frame	Pandas DataFrame	`data.frame(x = c(1,2,3), y = c("a", "b", "c"))`
Function	Python function	`function(x) x + 1`
Raw	Python bytearray	`as.raw(c(1:10))`
NULL, TRUE, FALSE	None, True, False	`NULL`, `TRUE`, `FALSE`

If a Python object of a custom class is returned then an R reference to that object is returned. You can call methods and access properties of the object just as if it was an instance of an R reference class.

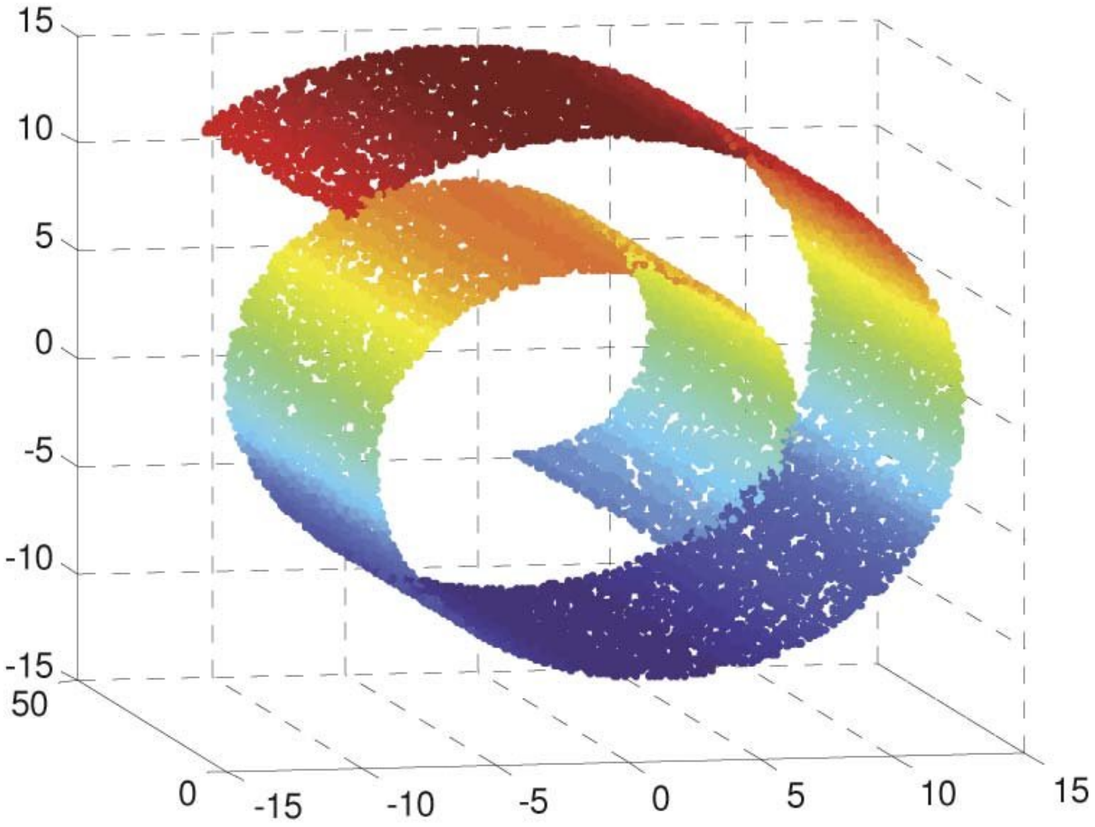
DATA

Yes, we are changing it again

DATA

Yes, we are changing it again

If we are using non-linear methods, we should use a data fit into it :)


15
10
5
0
-5
-10
-15
50
0
-15
-10
-5
0
5
10
15

Wrap-up

Thank you so much!

Do not hesitate to contact me if you have any more questions.