Philosophy / Discussion Topics

Data Science and AI for Neuroscience Summer School
Tara Chari and Lior Pachter
July 11, 2022

What is Exploratory Data Analysis (EDA) ?

1977 *Exploratory Data Analysis* - John Tukey

Alternative to 'confirmatory' data analysis → Allow data to generate hypotheses

Can be confounding to generate and test hypotheses on same data

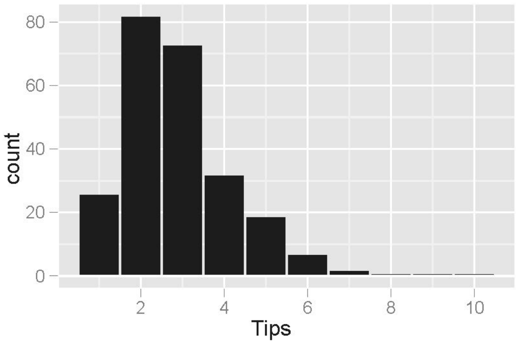
Objectives

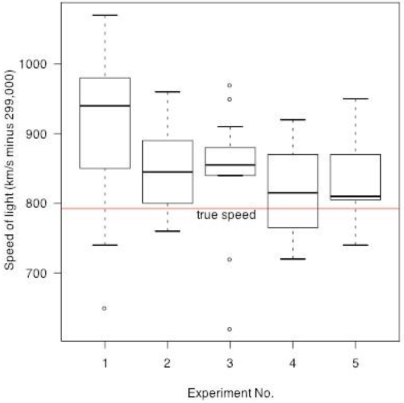
- Enable unexpected discoveries in the data
- Suggest hypotheses about the causes of observed phenomena
- **Assess assumptions** on which statistical inference will be based
- Support the **selection of appropriate statistical tools** and techniques
- Provide a basis for further data collection through surveys or experiments

Common Metrics and Methods for EDA

Visual techniques

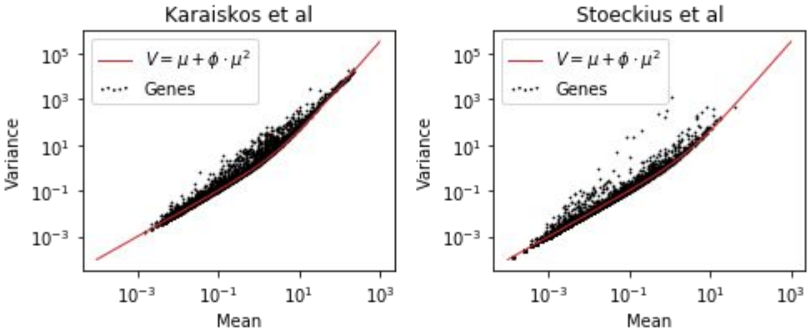
- Box plots
- Histograms
- Scatterplots on various features





Statistical metrics

- Max, min, median, quartiles (mean, std dev)
- Covariance, correlations, autocorrelation
- Compare distribution properties to assumptions



Svensson, 2017

Common Metrics and Methods for EDA

Dimensionality reduction (Unsupervised or Supervised)

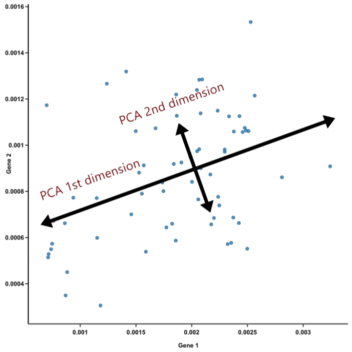
- Find patterns/features in high dimensional data, determine separation between labeled data
- Remove noise (what is biological, what is technical ...)

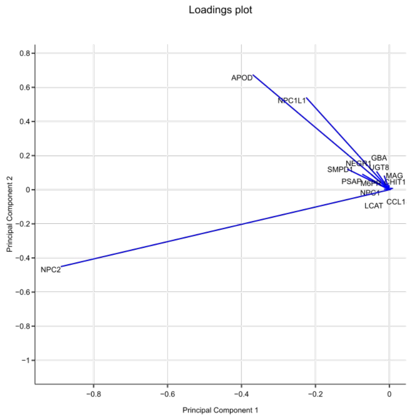
PCA - weighted sum of features (in each principal component), maximizing variance captured

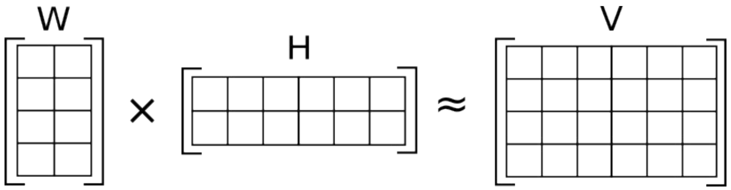
T = XW where W transforms X to new coordinate system (to T)

NMF (Nonnegative Matrix Factorization): X = WH, W coefficients on row variables, H coefficients on columns

- Can represent gene 'modules', weighted contributions of genes to each 'module'
- Can 'cluster' column variables

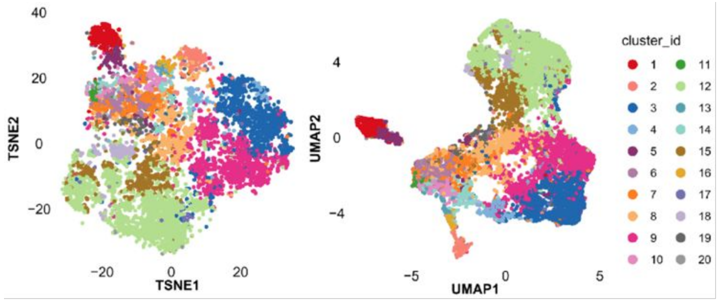






Avoiding Circular Analysis

Confounding 'exploratory' with 'all-in-one'

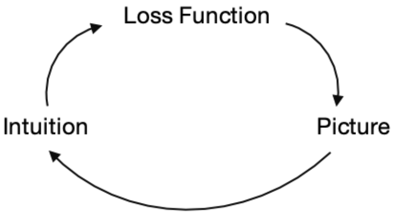


Nowicka et al. 2019

$$KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

or

$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + (1 - p_{ij}(X)) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

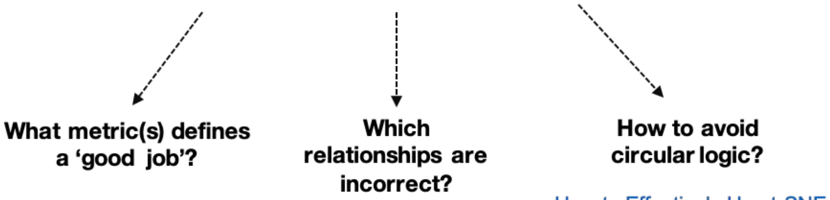

Loss Function

Picture

Intuition

SNE Paper (Hinton & Roweis 2002):

"… placed similar objects nearby in a low-dimensional space while keeping dissimilar objects well separated"

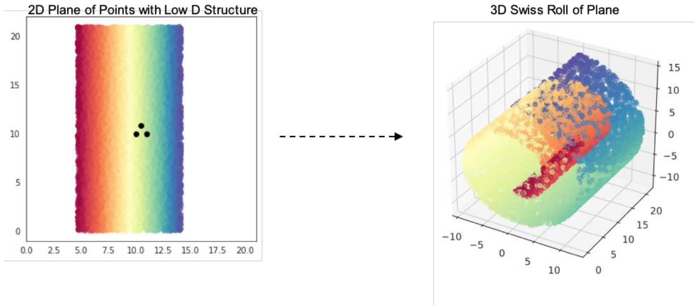

What metric(s) defines a 'good job'?

Which relationships are incorrect?

How to avoid circular logic?

How to Effectively Use t-SNE

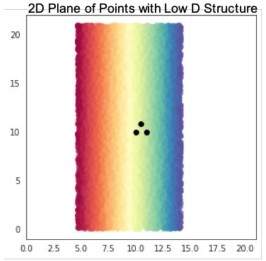
Avoiding Circular Analysis

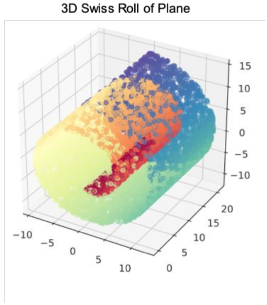


2D Plane of Points with Low D Structure

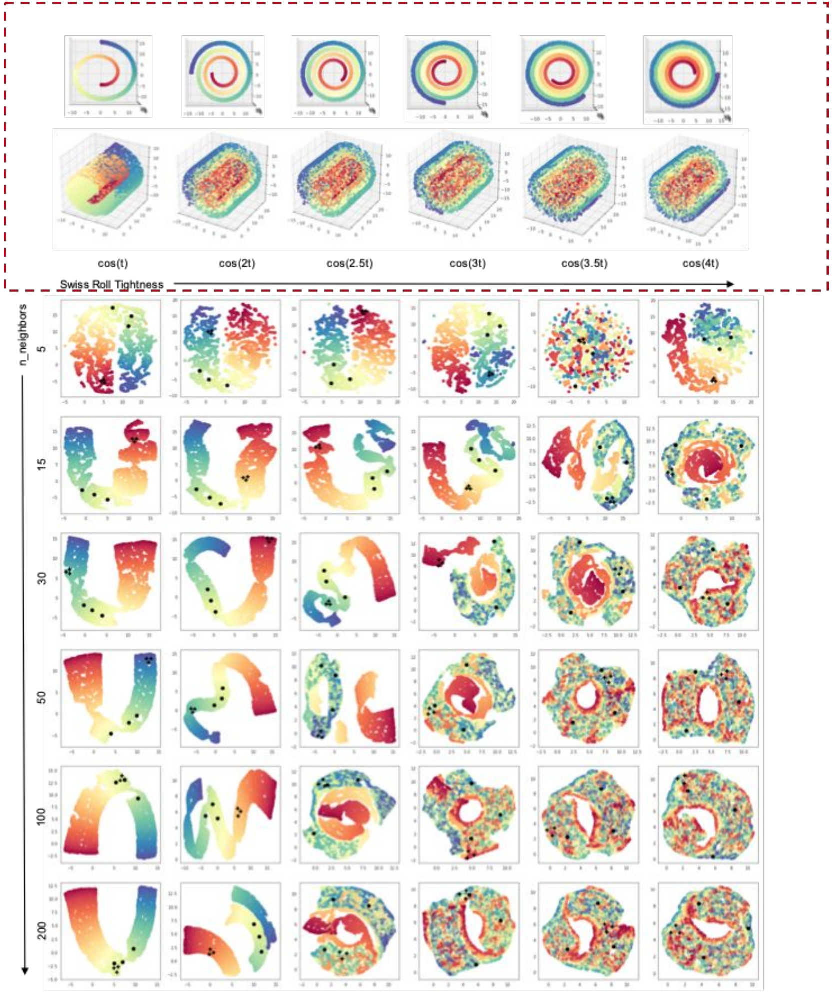
3D Swiss Roll of Plane

Avoiding Circular Analysis

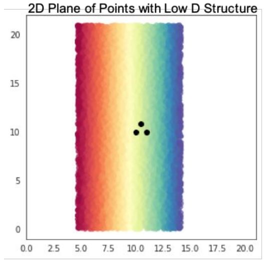




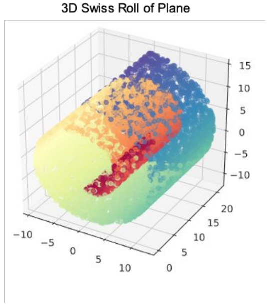
Embed in 2D (with UMAP)



Avoiding Circular Analysis

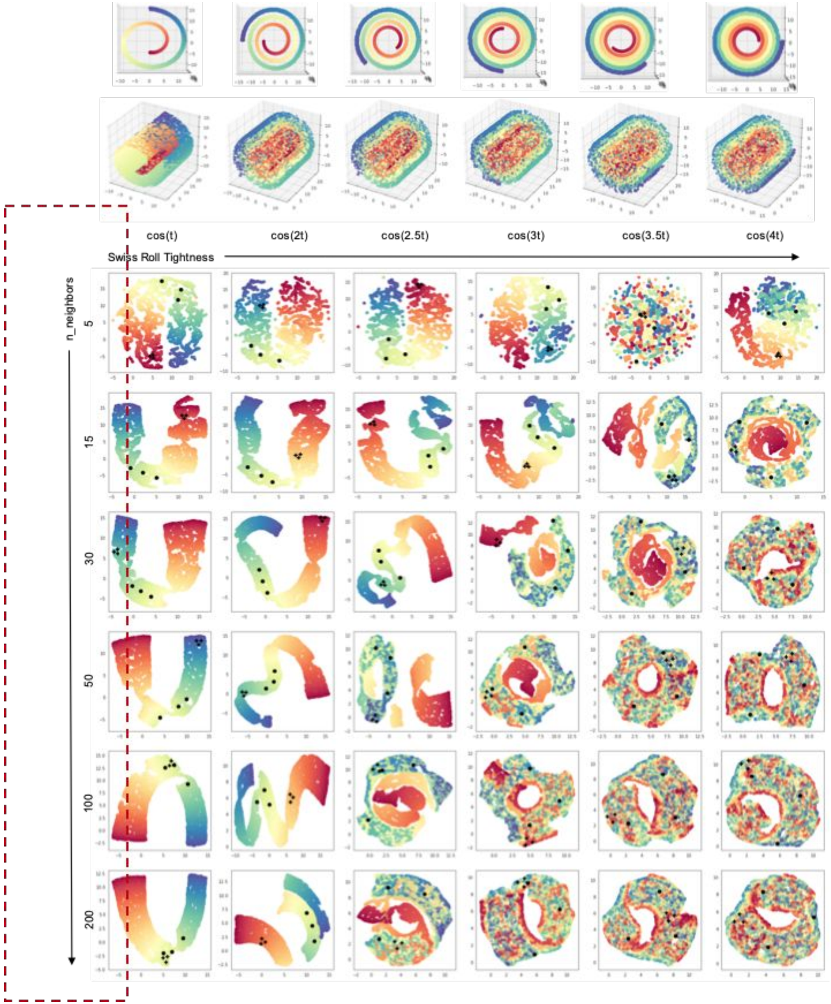


2D Plane of Points with Low D Structure

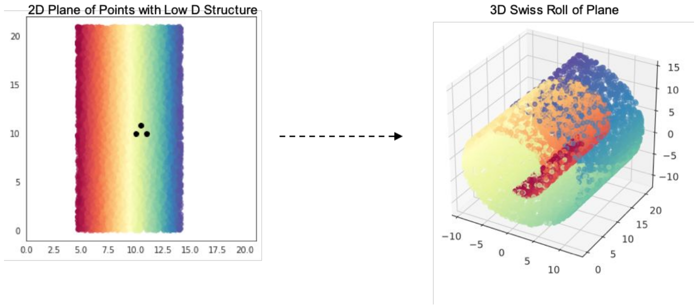


3D Swiss Roll of Plane

Embed in 2D (with UMAP)



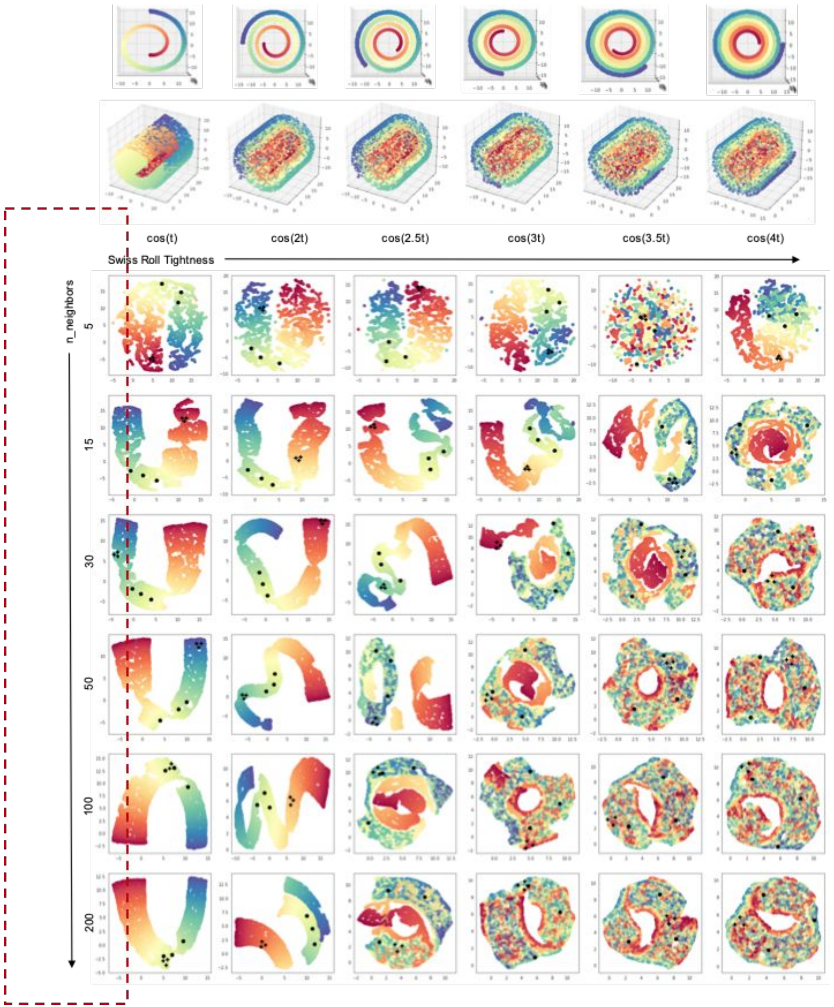
Avoiding Circular Analysis



- Use Euclidean distance by default to build neighbor graph

- Hard to say what metric is good/optimal (will always have poor neighborhood recapitulation)

- Same graph often fed to clustering algorithms in Scanpy and Seurat, thus the embedding does not provide an 'orthogonal' check



Questions for you:

- Have you used dimensionality reduction in your analyses, and for what purposes? How do you decide the number of dimensions to use?

- Have you normalized/pre-processed your data? How did you choose the transformations to apply?
    - What data type do you usually work with?

- What are the main metrics you use to assess data quality?