CS 4449 Data Science Capstone

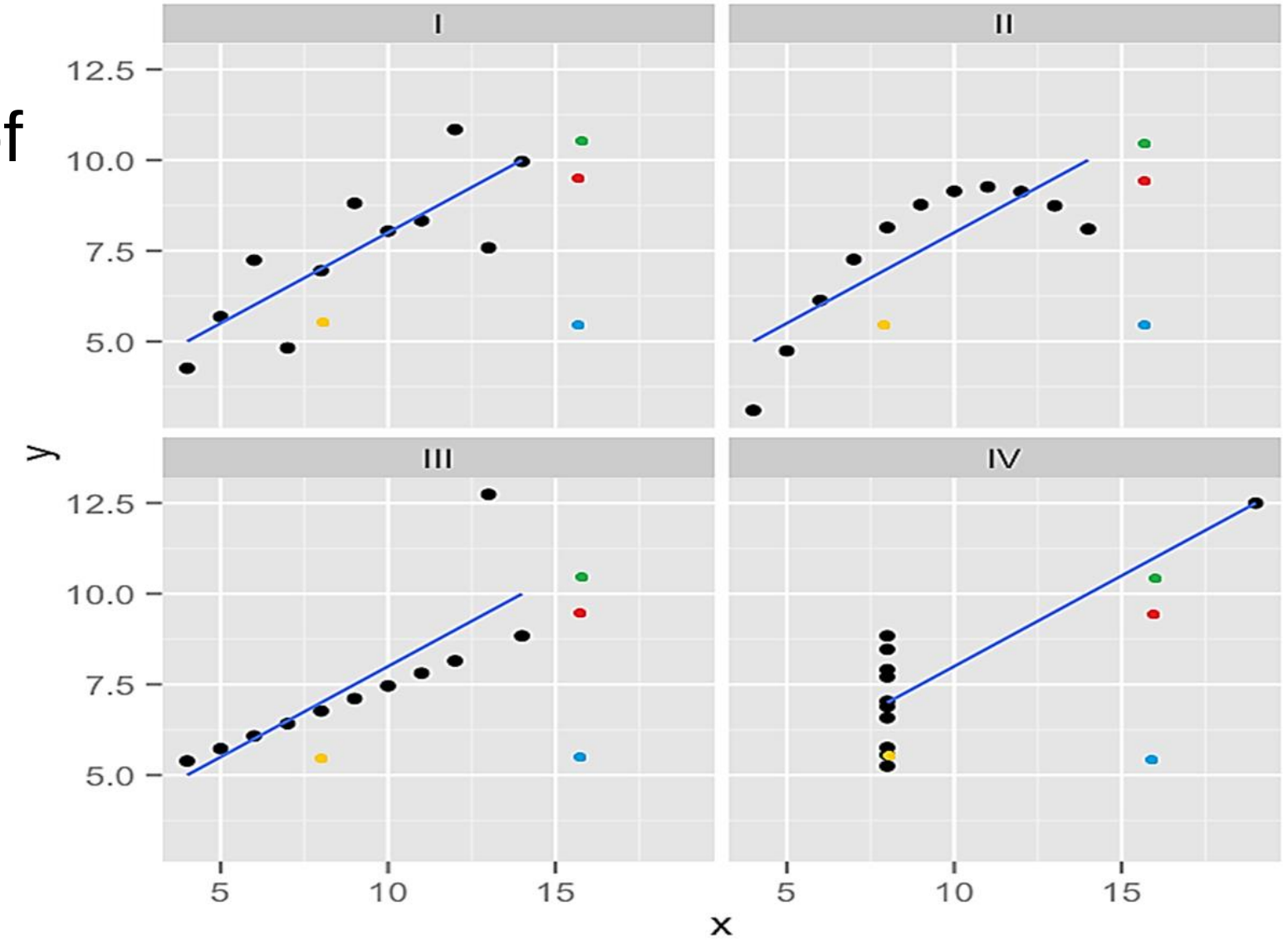
2023 Spring Quarter

Claudio Delrieux
claudio.delrieux@du.edu

Your replies to 1.1

To which dataset should we add each of the following data points:

<16, 9> in red
<16, 11> in green
<8, 6> in yellow
<16, 6> in blue



Your replies to 1.1

It is difficult to infer what is happening from looking at the dataset using just a table alone. If I have a dataset with a million rows you can’t give a good explanation of what the data looks like without doing some kind of analysis or visualization. You can get better insights from the data by putting into an easier to read form such as a graph or chart. Solely relying on the output of the statistics can lead to a misrepresentation of what the data looks like. For example, we could have never seen the outliers in datasets III and IV without looking at a graph first which is why exploratory data analysis is a crucial step in understand the data.

Your replies to 1.1

The overarching messages I received from problem 1.1, are that:

1) descriptive statistics alone do not necessarily provide enough information about a dataset to thoroughly familiarize oneself with it and

2) The degree to which an outlier can influence and paint a deceptive picture of a dataset

Your replies to 1.1

I was super surprised to see that while the summary statistics are very similar for all datasets, the visualization show four very distinct datasets. This made me realize that I can't just rely on summary statistics.

I need to take the time to visualize the data to get full context. It is interesting how the first two datasets don't have outliers and the last two datasets have a single outlier that yank the dataset into the similar summary statistics as other datasets.

Your replies to 1.1

Outliers will not always necessarily throw off summary statistics, as evidenced by mean, variance, and correlation being relatively equal across all 4 datasets despite an outlier in both datasets III and IV.

However, they will have influence on statistical analysis tasks like regression. Visually, datasets 3 and 4 appeared visually to have a near linear relationship between x and y. However, because of the outlier point in both those datasets, the best-fit line created from a linear regression analysis is not the expected line.

(continues...)

Your replies to 1.1

Also that analysis techniques are not one-size-fit-all, such that linear regression doesn't always make sense for every set of data points (i.e., dataset II, where some kind of polynomial regression would likely produce a better-fit equation based on the parabolic nature of the scatterplot).

Your replies to 1.1

In project 1.1 the lesson is clearly to visualize everything. For humans, visualizations of data are almost always more meaningful then numbers on a page. It makes patterns in the data much more recognizable.

Your replies to 1.2

We worked on a similar text classification assignment in the Tools2 course. Whether intended or not, the message I received was the extent to which preprocessing outweighs model building in terms of time consumption.

As I recall, once the messages were transformed into word vectors, the algorithms could be run relatively quickly.

Your replies to 1.2

I haven't dealt much with unbalanced datasets before, so it's interesting to see project 2's approach to balancing the dataset, which is to randomly remove negative tweets to match the number of positive samples.

I am a bit confused why they are removing sentiment_confidence below 0.5, but I'm assuming they want to work with tweets that have high confidence. The NLTK tools took me back to the Data Science Tools 1 class, and I can see that emojis code may be a bit outdated.

(continues...)

Your replies to 1.2

New emojis are always coming out, and I think there may be a better way to handle a wide range of emojis rather than just adding them to the list in the code.

I haven't had much experience with natural language, so the transformation of tokens into feature vectors, and then using the vectorizer to transform each tweet into a bag of words was interesting.

Your replies to 1.2

I thought it was an interesting technique to balance the unbalanced class outcomes for the output variable, sentiment, since imbalanced data does affect classifier performance. I thought the approach to randomly remove Negative tweets in order to match the number of Positive tweets was an interesting way to be able to still look at true accuracy score as a meaningful performance metric along with precision, recall, and F1 score, which usually hold more weight than accuracy for imbalanced datasets.

(continues...)

Your replies to 1.2

However, I would have been interested to see what the results of this analysis would have looked like for a severely imbalanced dataset since 1000s of observations were removed in order to to balance class outcomes, or the outcome of running the analysis on various sub-samples of the negative sentiment tweets.

Your replies to 1.2

If there is a lesson it is probably to balance your dataset. An unbalanced dataset can lead to many issues including:

- a model that always predicts positive (or negative) meaning it hasn't learned anything.

- a low accuracy score if it cannot learn the patterns and is randomly guessing

- a model that cannot converge

A balanced dataset is almost always the way to go. However if you must use all you data and it is unbalanced, you will want to use class weights to aid the models learning.

My thoughts on your replies to 1.2

This shows how "open ended" data analysis may be sometimes, we perceive that there are different interpretations that may guide the analysis tasks into different aims.

Thanks so much for your replies!

Midterm presentation

Your presentation will be about 7 minutes plus 2 minutes for questions.

A 3 / 4 page report including dataset description, data preparation and analysis, results, and discussion, together with a documented github/notebook, to be delivered within the week after presentation.

Midterm presentation

Present what is the problem, why did you considered it, what is its importance, how did you approached it, and the challenges you found.

ABCD approach.

A is **a**ssessing globally the understanding of the project, the requirements and the expected results.

B is find the **b**est action plan, what kind of data do we have, what data best describes the problem.

Midterm presentation

C is choosing an approach. What kind of model (i.e.,classification, regression, etc.), what kind of metrics you would use to measure the outcome of this project, what is the success criteria.

D is for developing (hands on!).

Explain and justify your choices. What other alternatives were feasible.

Midterm presentation

It is likely that your final model will not be 100% accurate, then include error analysis and try to provide hints therof.

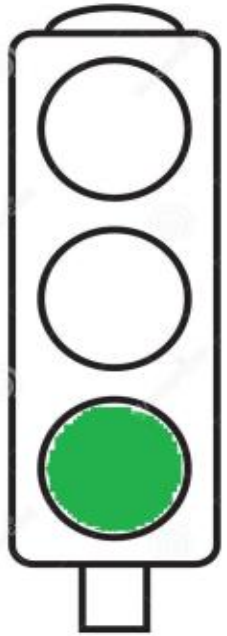
Also (very important!) include a time break-out (% of the time devoted to data preparation, model building, etc.).

Midterm presentation

Please start with a short pitch (less than a minute) that should be easily understood, as if you were addressing laypersons without prior exposure to the class material.

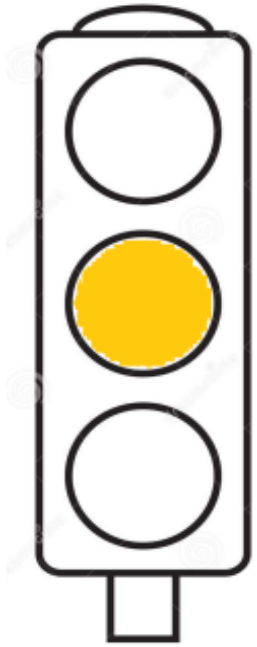
If by the presentation day you still don’t have everything settled, then focus your presentation as a progress report and explain what was done and what was left and will be done.

Midterm presentation



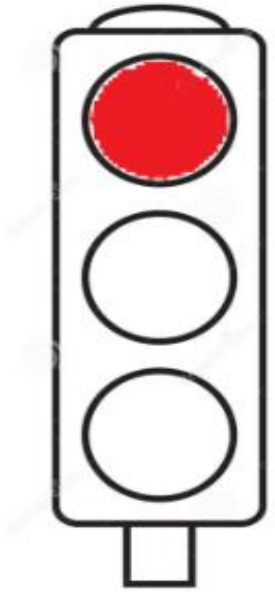
	Mandatory	Optional
Data preparation	• Describe how the features are distributed. • Describe if the records group around a given tendency.	• Present a statistical description of features and/or records.
Data analysis	• Apply at least two different analysis techniques and compare their evaluation parameters.	• Apply three or more analysis techniques.
Data visualization	• Use visualization techniques to show the results.	• Use actionable visualization techniques to explore the results.

Midterm presentation



	Mandatory	Optional
Data preparation	• Provide an overall description of the dataset and the salient features.	• Describe how the features are distributed. • Describe if the records group around a given tendency.
Data analysis	• Apply one analysis technique and discuss the evaluation parameters found.	• Apply two or more analysis techniques and compare the results.
Data visualization	• Use visualization techniques to show the results.	• Use actionable visualization techniques to explore the results.

Midterm presentation



	Mandatory	Optional
Data preparation	• Provide an overall description of the dataset.	• Describe how the features and records are distributed.
Data analysis	• Apply one analysis technique and discuss the evaluation parameters found.	• Apply two or more analysis techniques and compare the results.
Data visualization		• Use visualization techniques to show the results.

Midterm presentation

Questions?

What other analysis techniques?

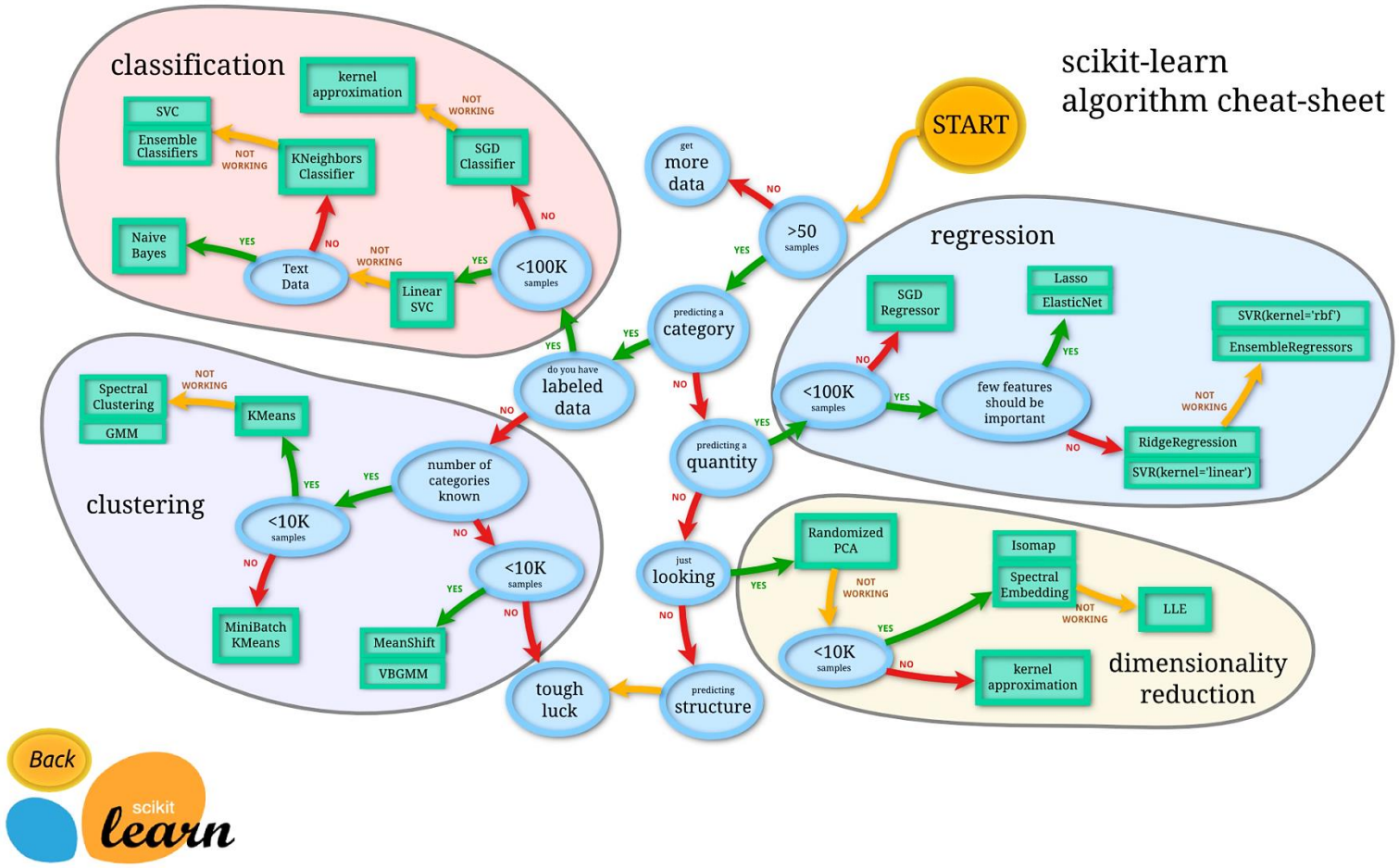
In the vids & handouts we stress the most frequently required data analysis techniques (classification, clustering and regression).

Here we will present a brief introduction to other DS models and techniques that may arise in several contexts.

Dimensionality reduction

For many, DR together with the former methods are the “DS big four”.

See for instance the SciKitLearn cheat-sheet:



Dimensionality Reduction

The curse of dimensionality:

In feature spaces with many dimensions, the volume of space grows exponentially, which implies that data density shrinks significantly.

This compromises the efficiency, stability and robustness of the learning techniques.

Dimensionality Reduction

Also, statistical significance weakens, and the training dataset required to cover all the possible cases is huge.

Distance-based methods (K-NN, K-means, etc.) get significantly biased (having many dimensions obscures the meaning of large differences in any of them).

Dimensionality Reduction

Last and foremost: many many attributes raises the likelihood of having irrelevant or redundant features, and missing or noisy data, all of which renders less stable models with high variance, and overfitting is difficult to avoid.

Thus, DR tries to avoid all these issues without loosing the representativeness of the feature set.

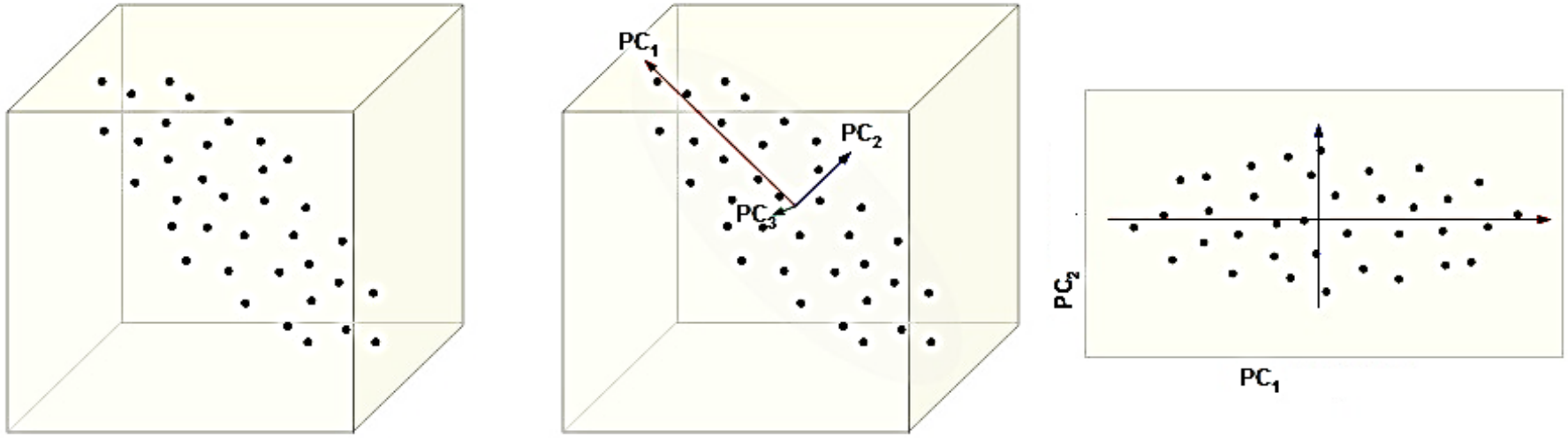
Feature selection

Retain only the features considered relevant under certain criteria, expecting to do away irrelevant and redundant features.

You can progress by means of “forward aggregation” (f.e., using a decision tree to add the most significant features one by one), or by “backwards elimination”.

Feature extraction

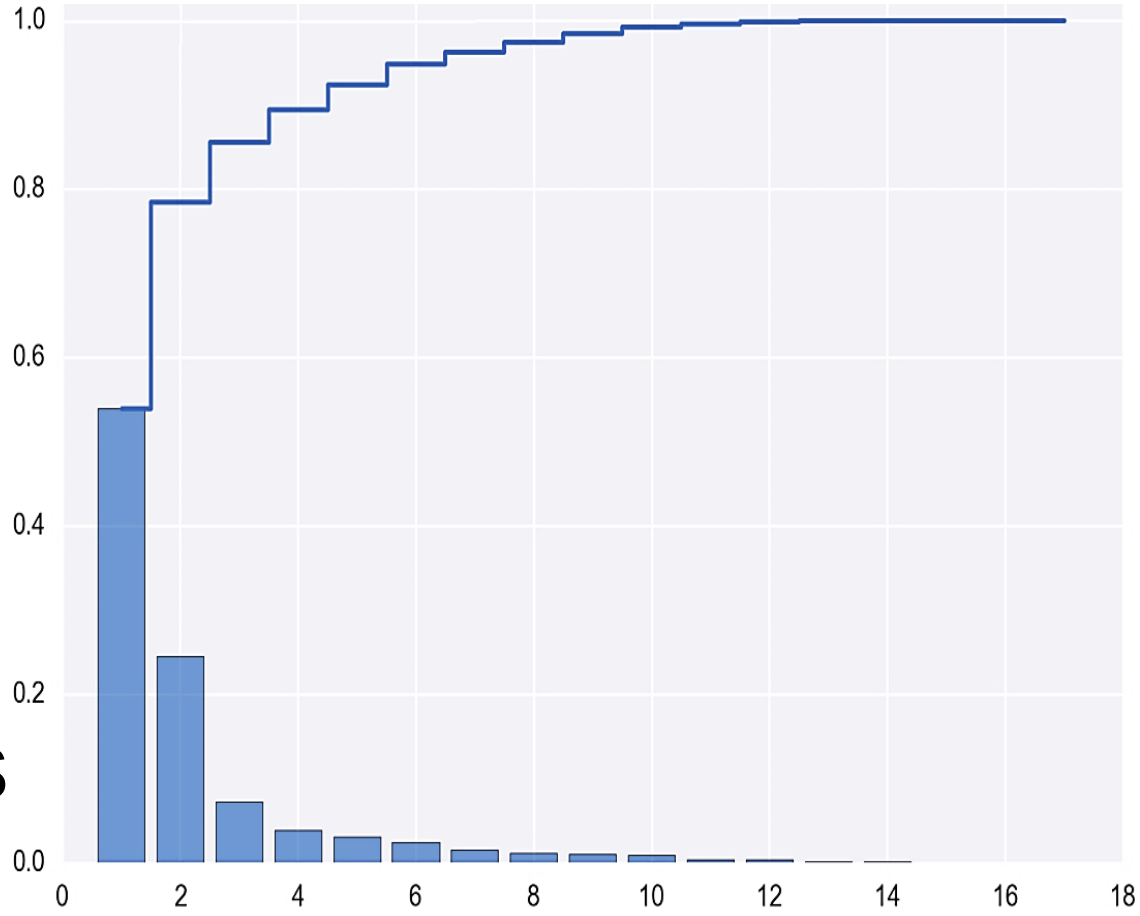
Linear FE models are based on the idea that the dataset can be projected into a subspace that captures most of its variance. PCA and LDA are the most widely known examples.



Feature extraction

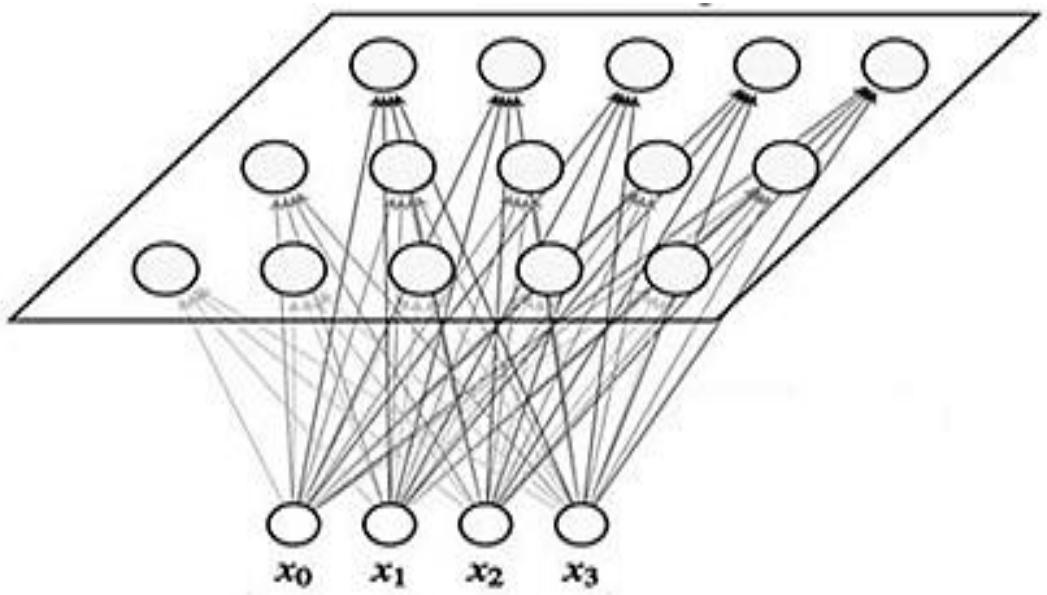
The PCs are the eigenvectors of the covariance matrix of the dataset, ordered by their respective eigenvalues.

This allows to establish a tradeoff between dimensions and the retained variance.



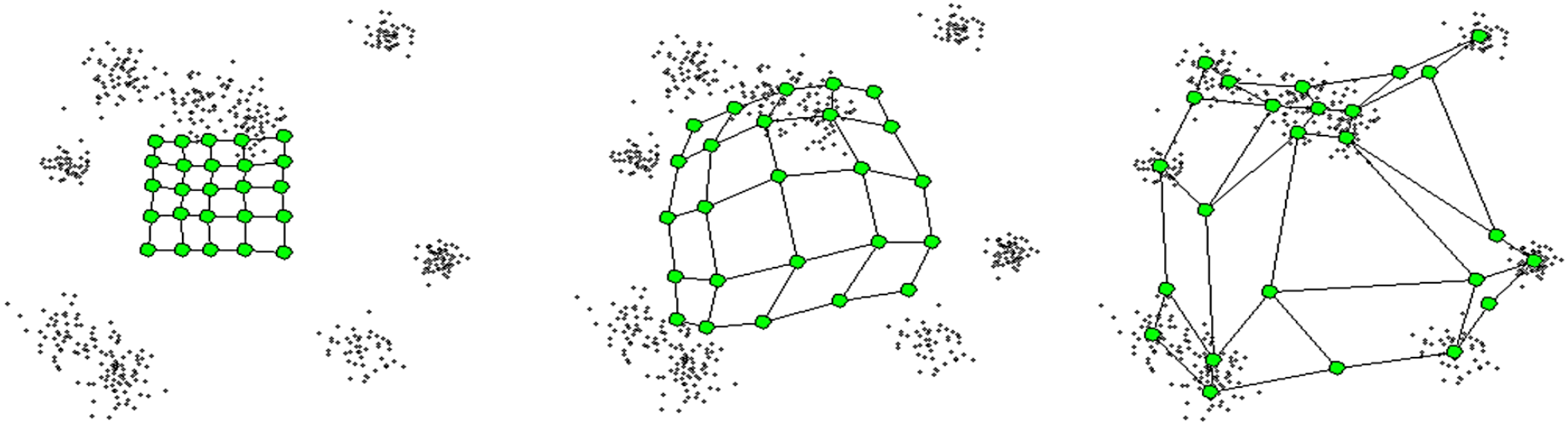
Nonlinear methods: SOMs

Self-organized maps are a regular grid of perceptrons, each fully connected to each feature of the dataset. After training, each data is mapped to the output of any of these perceptrons.



Nonlinear methods: SOMs

Training consists on randomly iterating the dataset. For each data, locate the perceptron most likely to get activated, apply the perceptron learning rule to it, and partially to its neighbors.



Advanced methods

T-SNE and UMAP are contemporary and advanced visualization techniques. They both apply elaborate mathematical techniques to perform a nonlinear projection of the feature space into a 2D or 3D space in which data visualization is meaningful.

https://distill.pub/2016/misread-tsne

https://umap-learn.readthedocs.io/en/latest/

Questions please?