GENERAL ASSEMBLY

DIMENSIONALITY REDUCTION

Mason Gallo, Data Scientist

AGENDA

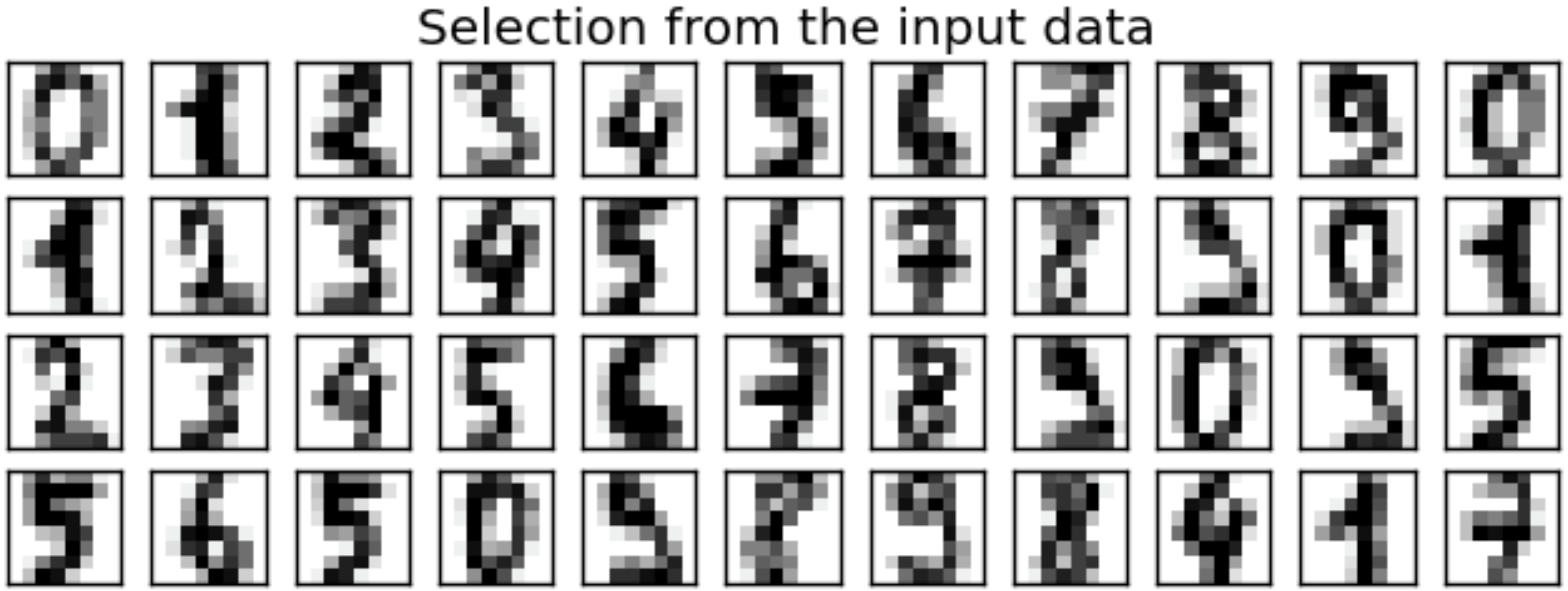
- Dimensionality reduction
- PCA
- Real world example
- Implementation

OBJECTIVES

- Dim Reduction intuition
- Understand how dim reduction is used in the real world
- Implement dim reduction in Python

MOTIVATING EXAMPLE: HANDWRITTEN DIGITS

WHAT HAPPENS WHEN WE REDUCE DIMENSIONALITY?



Now that you're familiar with this dataset, we'll try reducing its dimensionality

UNSUPERVISED LEARNING AND DIM REDUCTION

DIMENSIONALITY REDUCTION

Q: What is dimensionality reduction?

DIMENSIONALITY REDUCTION

Q: What is dimensionality reduction?

A: A set of techniques for reducing the size (in terms of features) of the dataset under examination.

DIMENSIONALITY REDUCTION

Q: What are the motivations for dimensionality reduction?

DIMENSIONALITY REDUCTION

Q: What are the motivations for dimensionality reduction?

The number of features in our dataset can be difficult to manage, or even misleading (eg, if the relationships are actually simpler than they appear).

DIMENSIONALITY REDUCTION – EXAMPLE

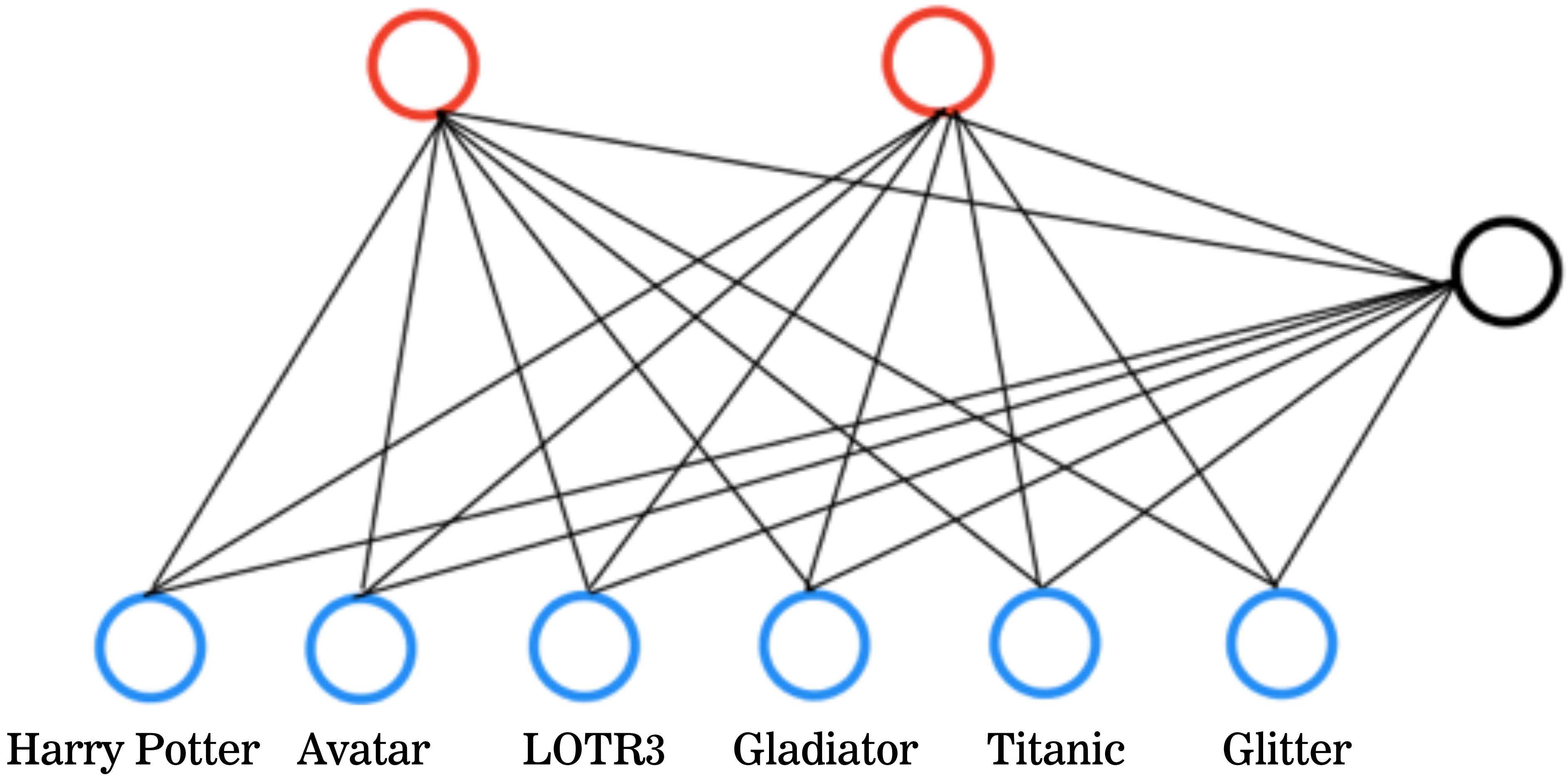
We’d like to represent a user’s taste profile by a select number of dimensions, rather than their rating of each and every movie

DIMENSIONALITY REDUCTION – EXAMPLE

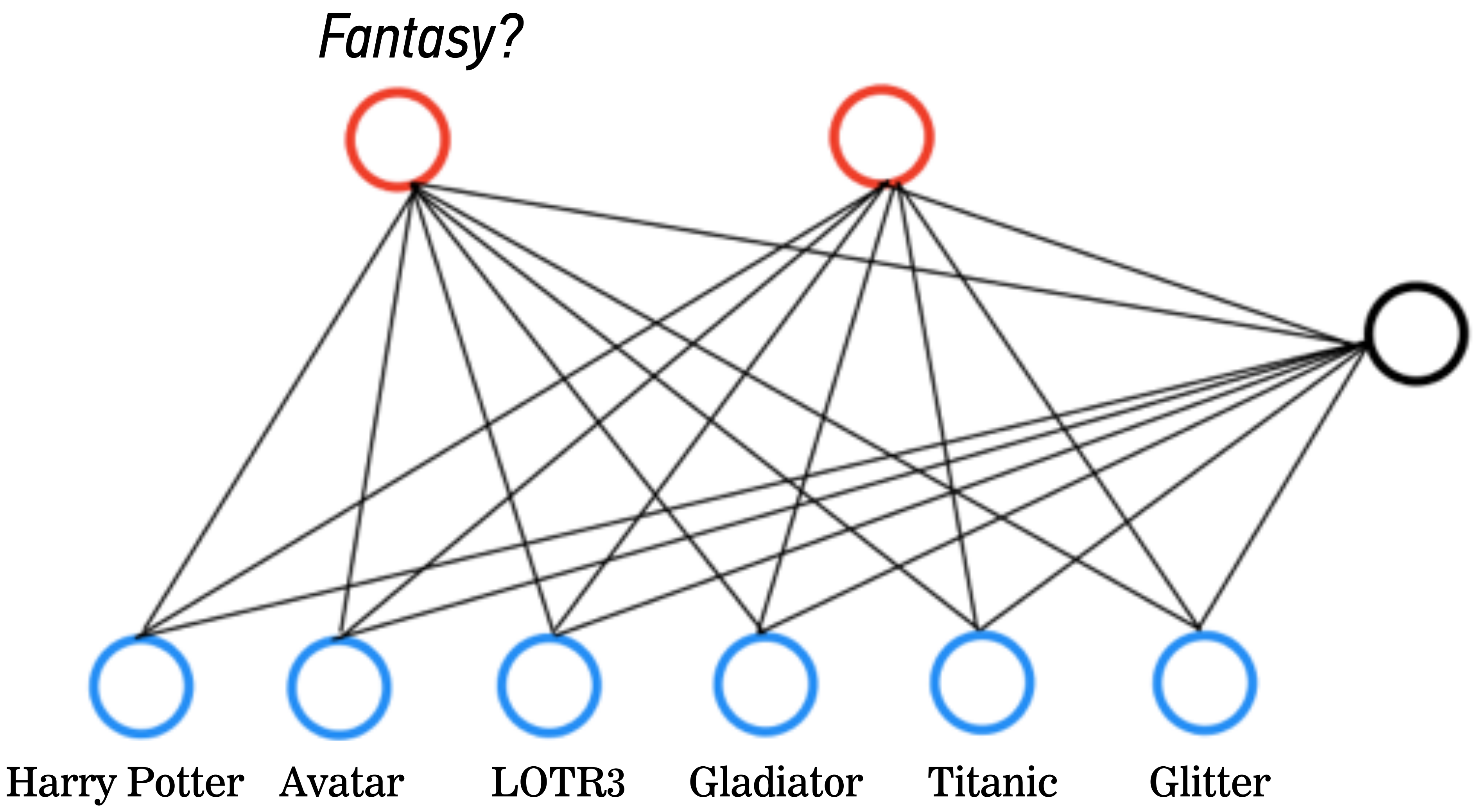
We'd like to represent a user's taste profile by a select number of dimensions, rather than their rating of each and every movie

Harry Potter Avatar LOTR3 Gladiator Titanic Glitter

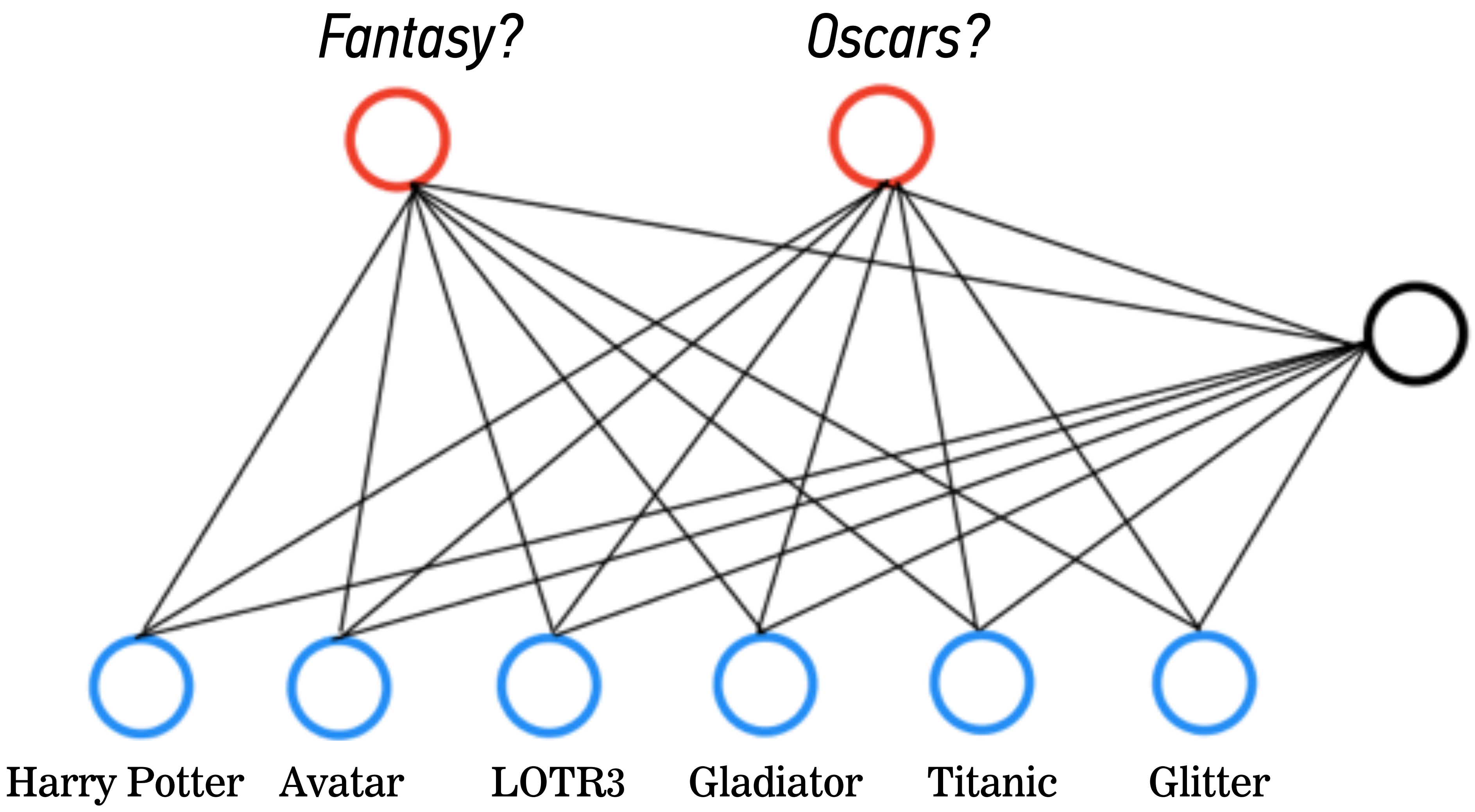
DIMENSIONALITY REDUCTION – EXAMPLE



DIMENSIONALITY REDUCTION – EXAMPLE



DIMENSIONALITY REDUCTION – EXAMPLE



DIMENSIONALITY REDUCTION

Q: What is the goal of dimensionality reduction?

DIMENSIONALITY REDUCTION

Q: What is the goal of dimensionality reduction?

- *reduce computational expense*
- *reduce susceptibility to overfitting*
- *reduce noise in the dataset*
- *enhance our intuition*

DIMENSIONALITY REDUCTION

The goal of feature extraction is to create a new set of coordinates that simplify the representation of the data.

DIMENSIONALITY REDUCTION

Q: What are some applications of dimensionality reduction?

DIMENSIONALITY REDUCTION

Q: What are some applications of dimensionality reduction?

- *document clustering*
- *image recognition/computer vision*
- *recommender systems*

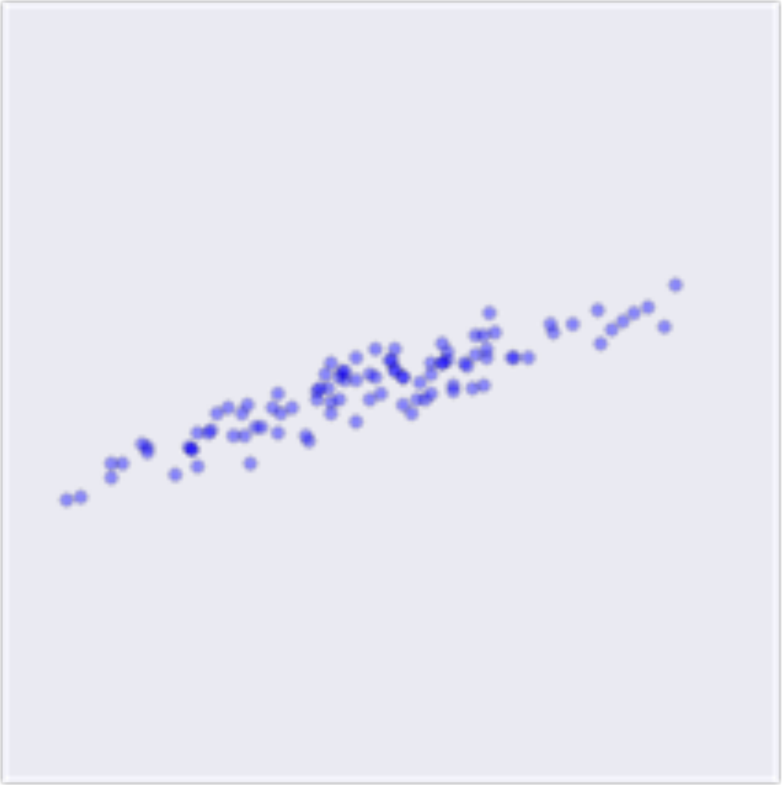
DIMENSIONALITY REDUCTION

PRINCIPAL COMPONENTS ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

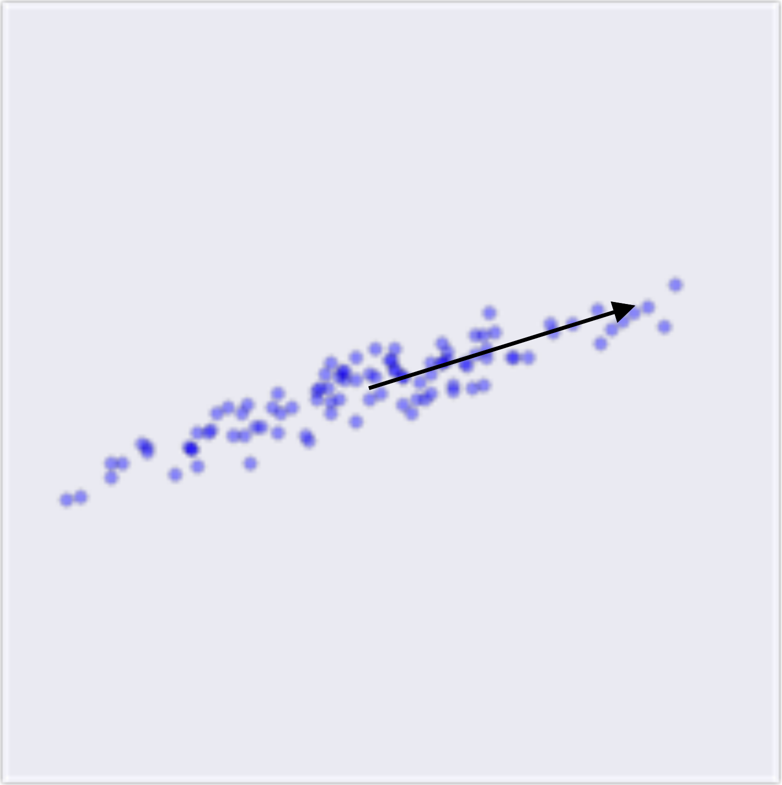
Principal component analysis is a dimension reduction technique that can be used on a matrix of any dimensions.

PRINCIPAL COMPONENT ANALYSIS



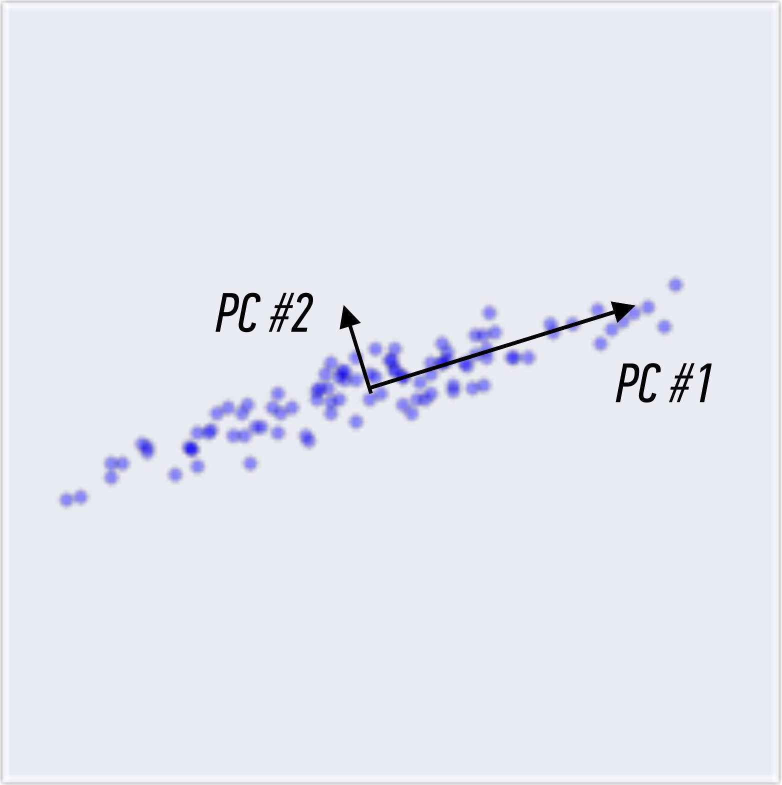
Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs

PRINCIPAL COMPONENT ANALYSIS



Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs

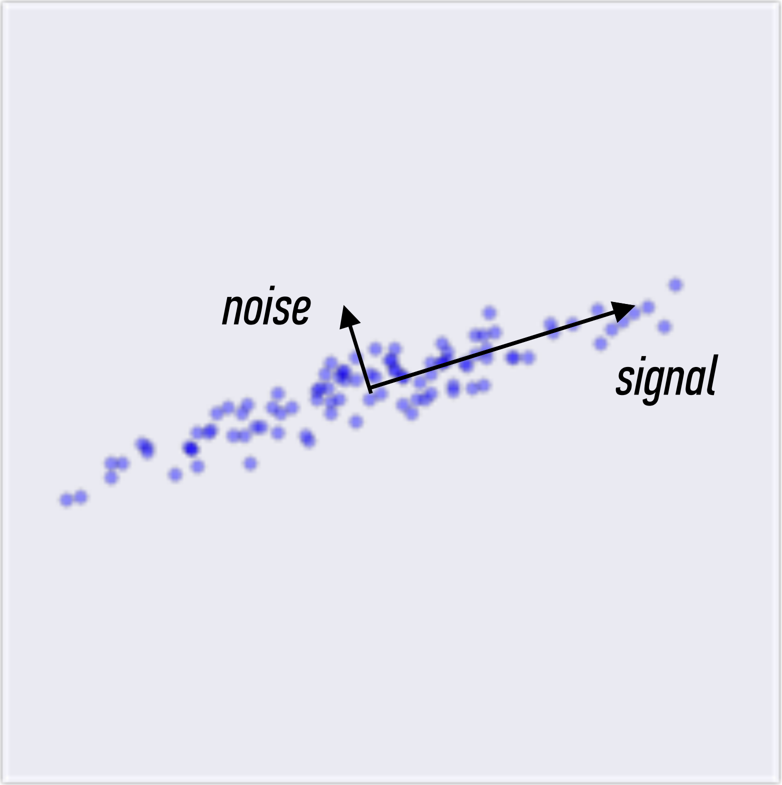
PRINCIPAL COMPONENT ANALYSIS



Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs

It can be seen as a transformation to a new orthogonal basis, ordered by variance

PRINCIPAL COMPONENT ANALYSIS



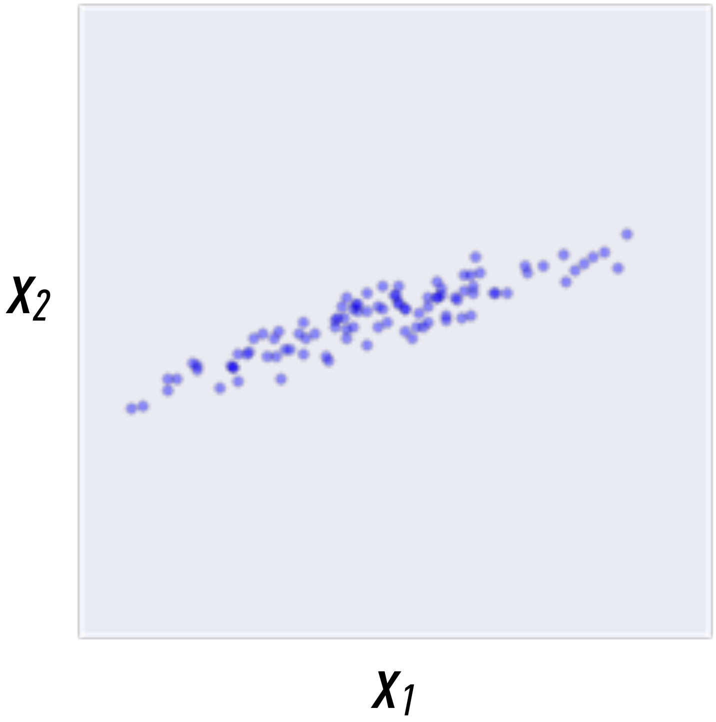
Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs

It can be seen as a transformation to a new orthogonal basis, ordered by variance

The idea is that the first principal components contain the most information, while the latter ones contain noise

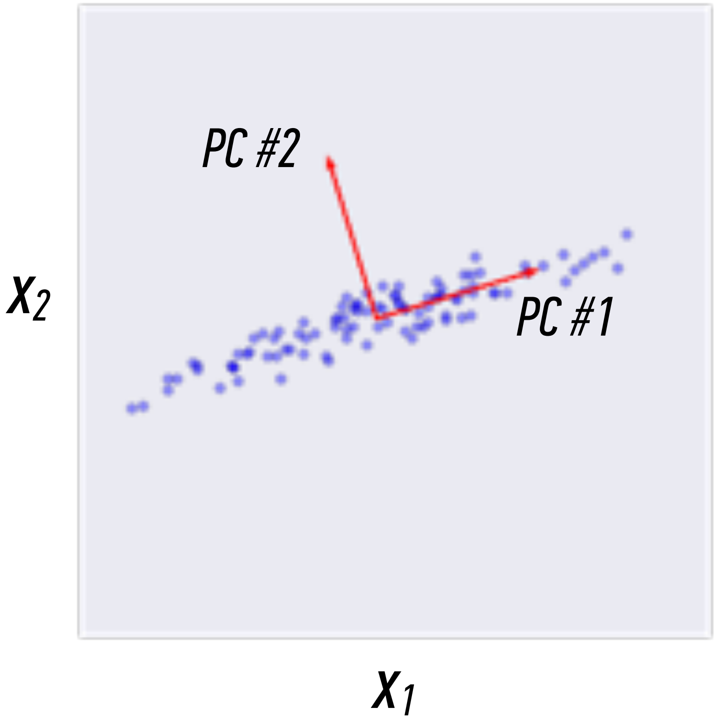
PRINCIPAL COMPONENT ANALYSIS

- *Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs*



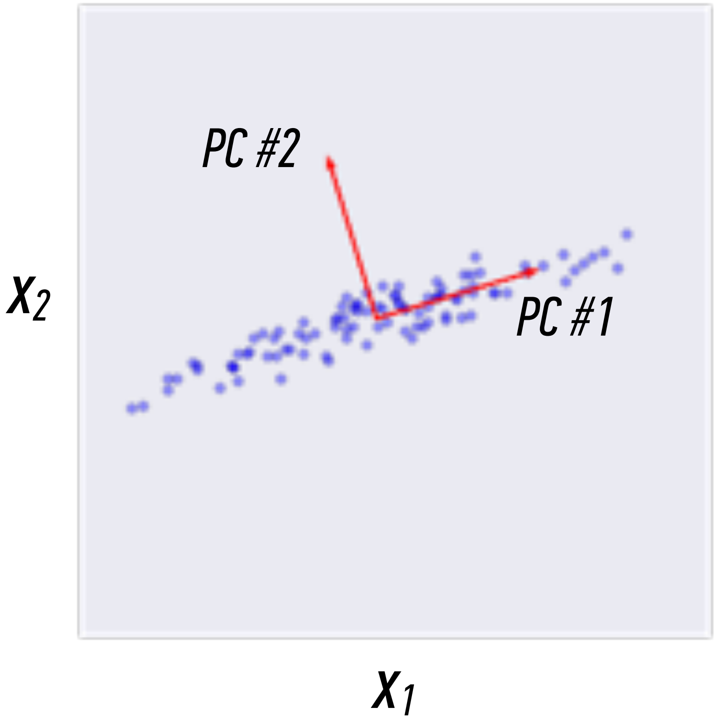
PRINCIPAL COMPONENT ANALYSIS

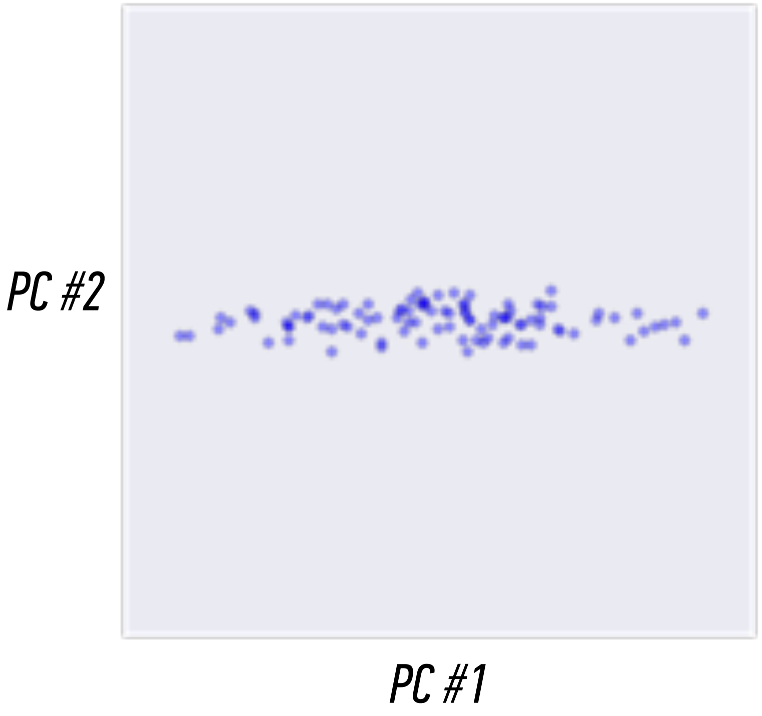
- *Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs*



PRINCIPAL COMPONENT ANALYSIS

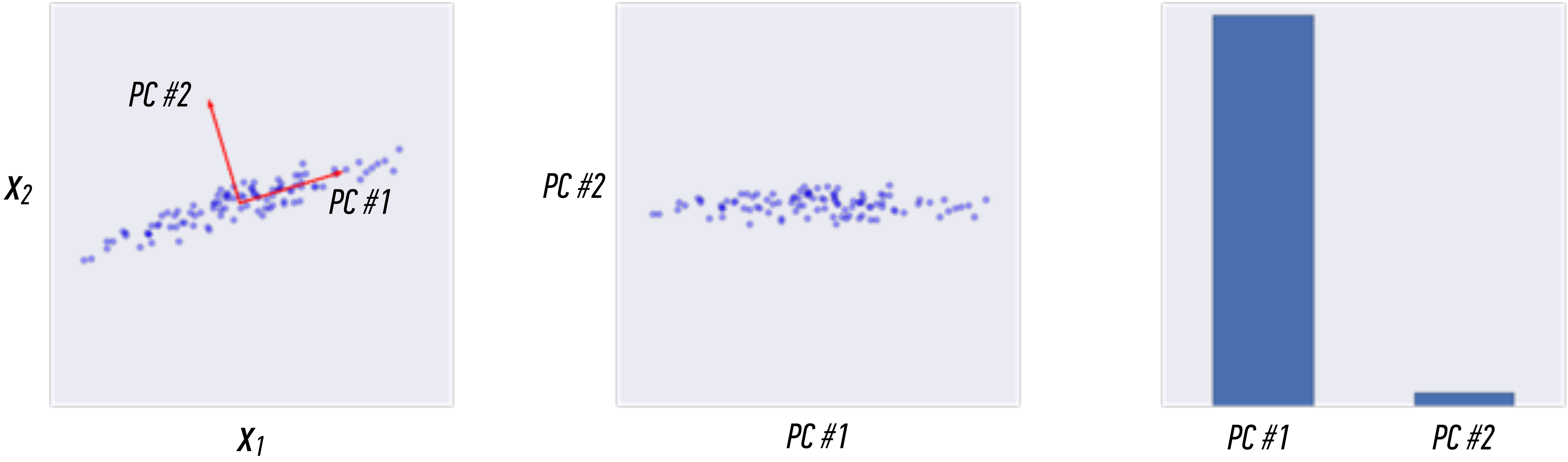
- *Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs*
- *It can be seen as a transformation to a new orthogonal basis*





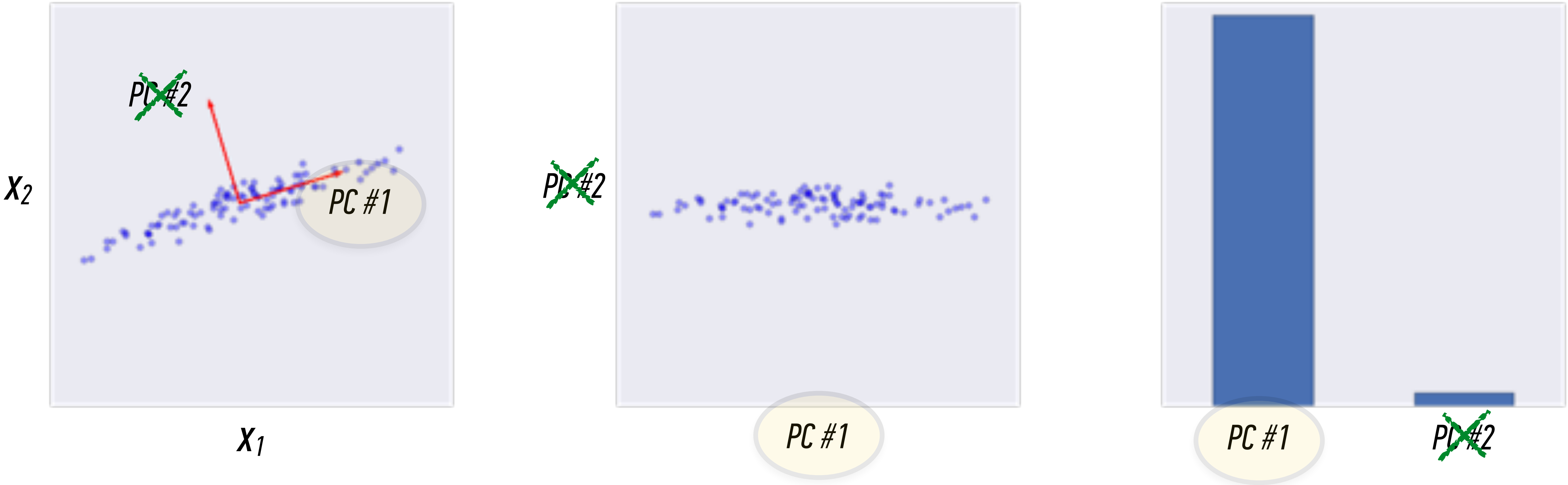
PRINCIPAL COMPONENT ANALYSIS

- *Principal Component Analysis (PCA) seeks the dimensions in which the most variance occurs*
- *It can be seen as a transformation to a new orthogonal basis*
- *The principal components are ordered by the size of their variance*

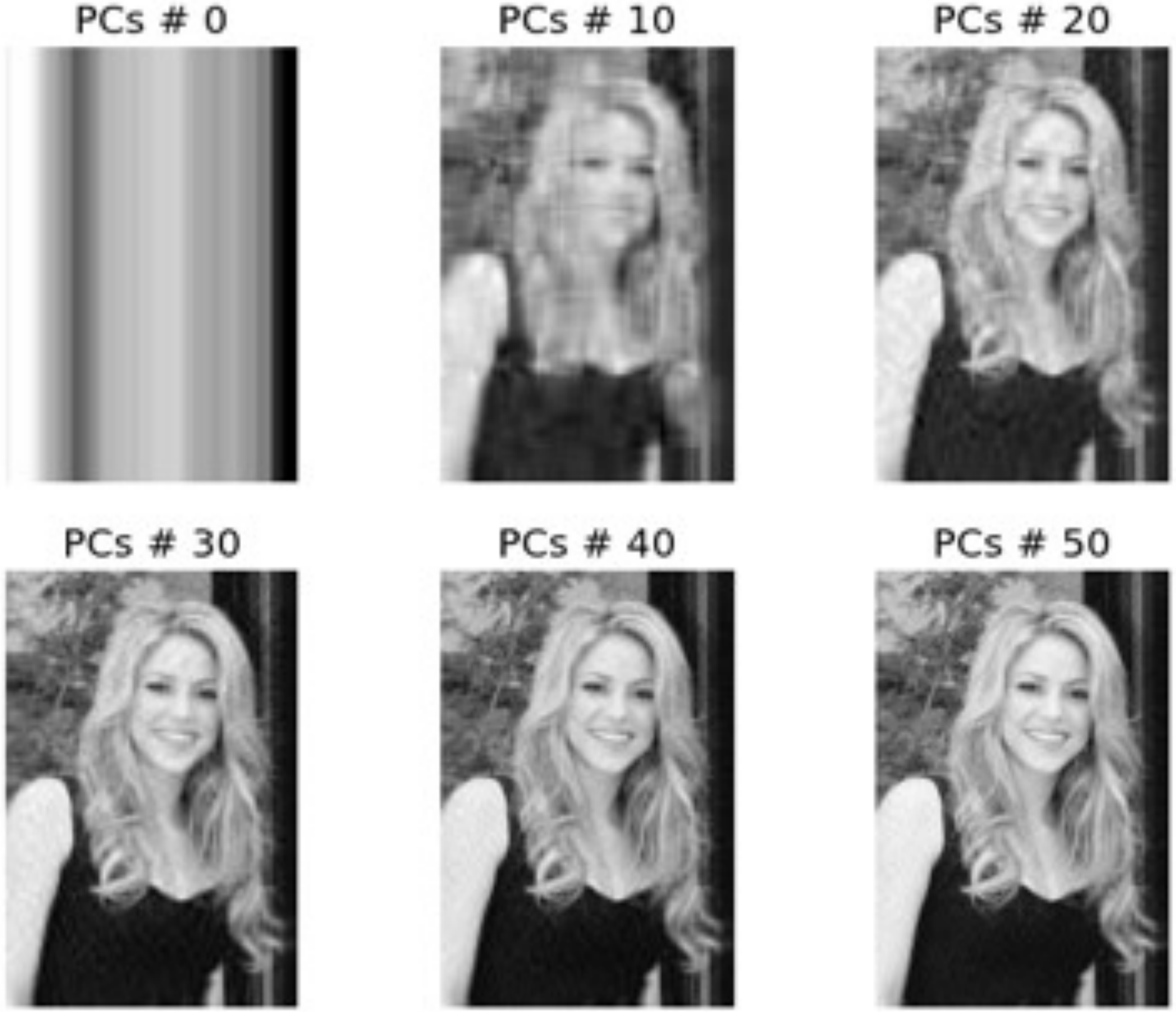


PRINCIPAL COMPONENT ANALYSIS

We can now **reduce the dimension** *by only looking at the first few principal components that explain the most variance*



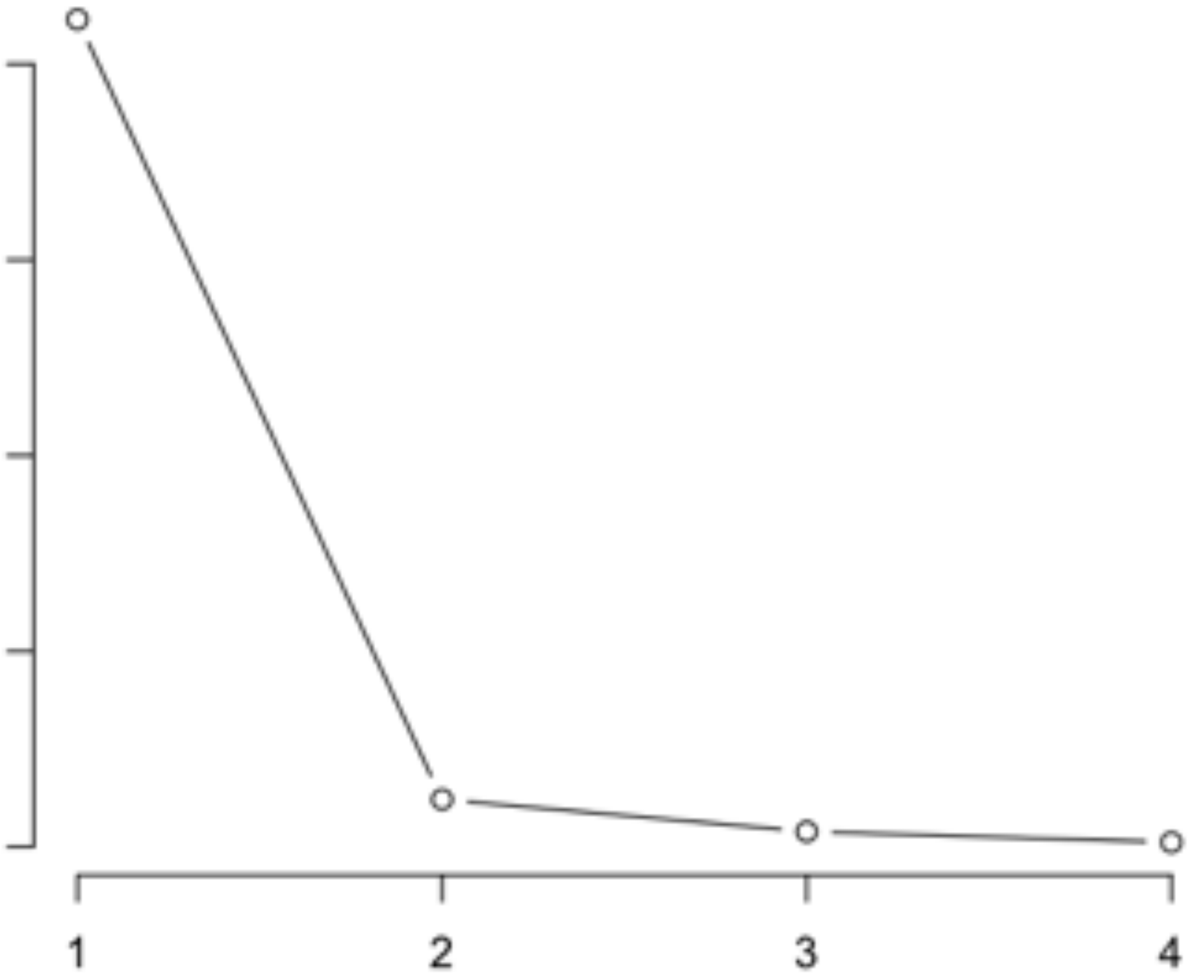
DIMENSIONALITY REDUCTION



source: http://glowingpython.blogspot.it/2011/07/pca-and-image-compression-with-numpy.html

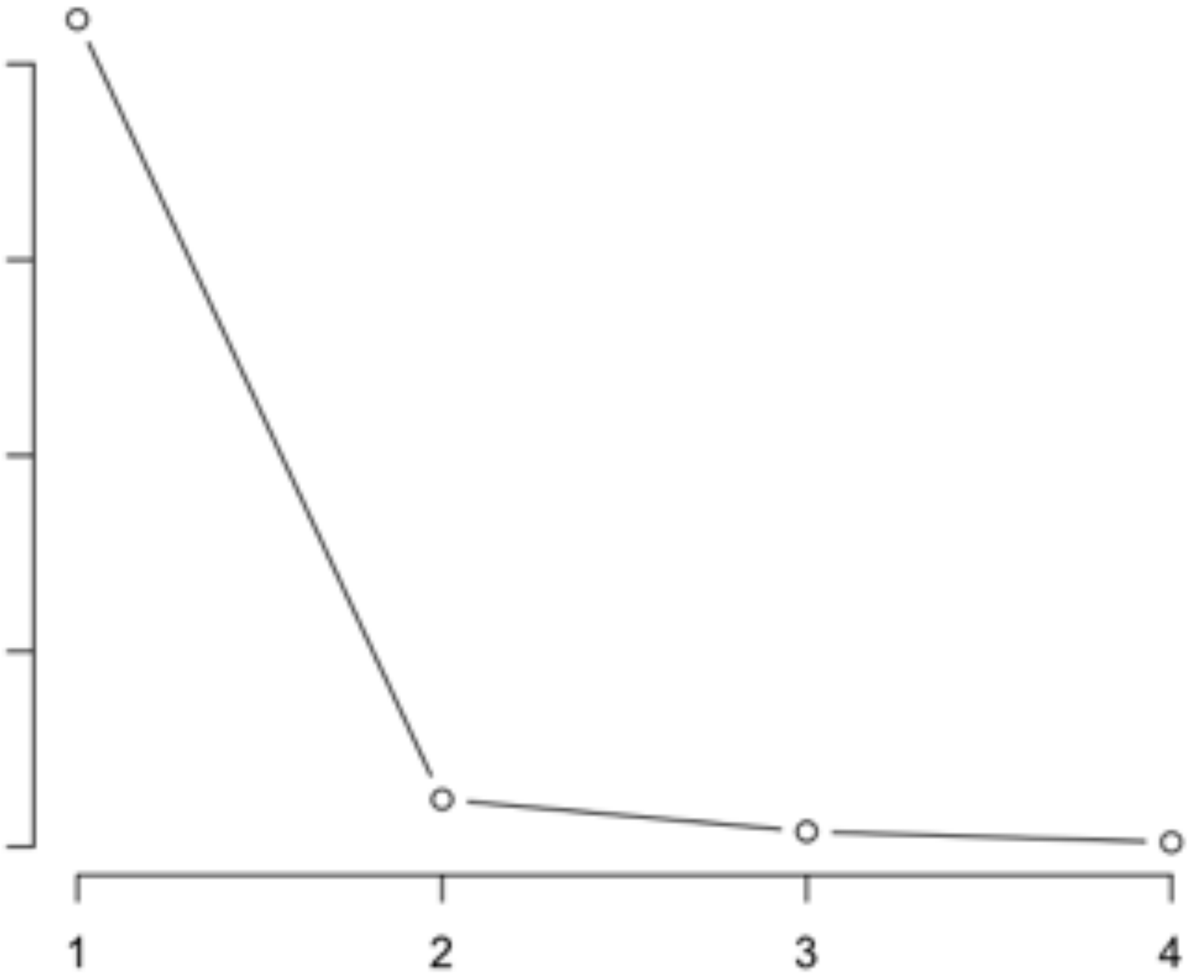
PRINCIPAL COMPONENT ANALYSIS

Principal components of Iris dataset



PRINCIPAL COMPONENT ANALYSIS

Principal components of Iris dataset



NOTE

Looking at this plot also gives you an idea of how many principal components to keep.

Apply the ***elbow*** *test*: keep only those pc’s that appear to the left of the elbow in the graph.

DIMENSIONALITY REDUCTION

VISUALIZATION

VISUALIZING PCA

http://setosa.io/ev/principal-component-analysis/

REAL WORLD BIG PICTURE

HOW I USE DIM REDUCTION

Understand latent variables for story-telling

WHAT ARE LATENT VARIABLES?

Non-measurable themes

Usually groups of variables rolled up into a single category

Ex: square foot of house and number of rooms —> house size

THESE LATENT VARIABLES ARE THE PRINCIPAL COMPONENTS

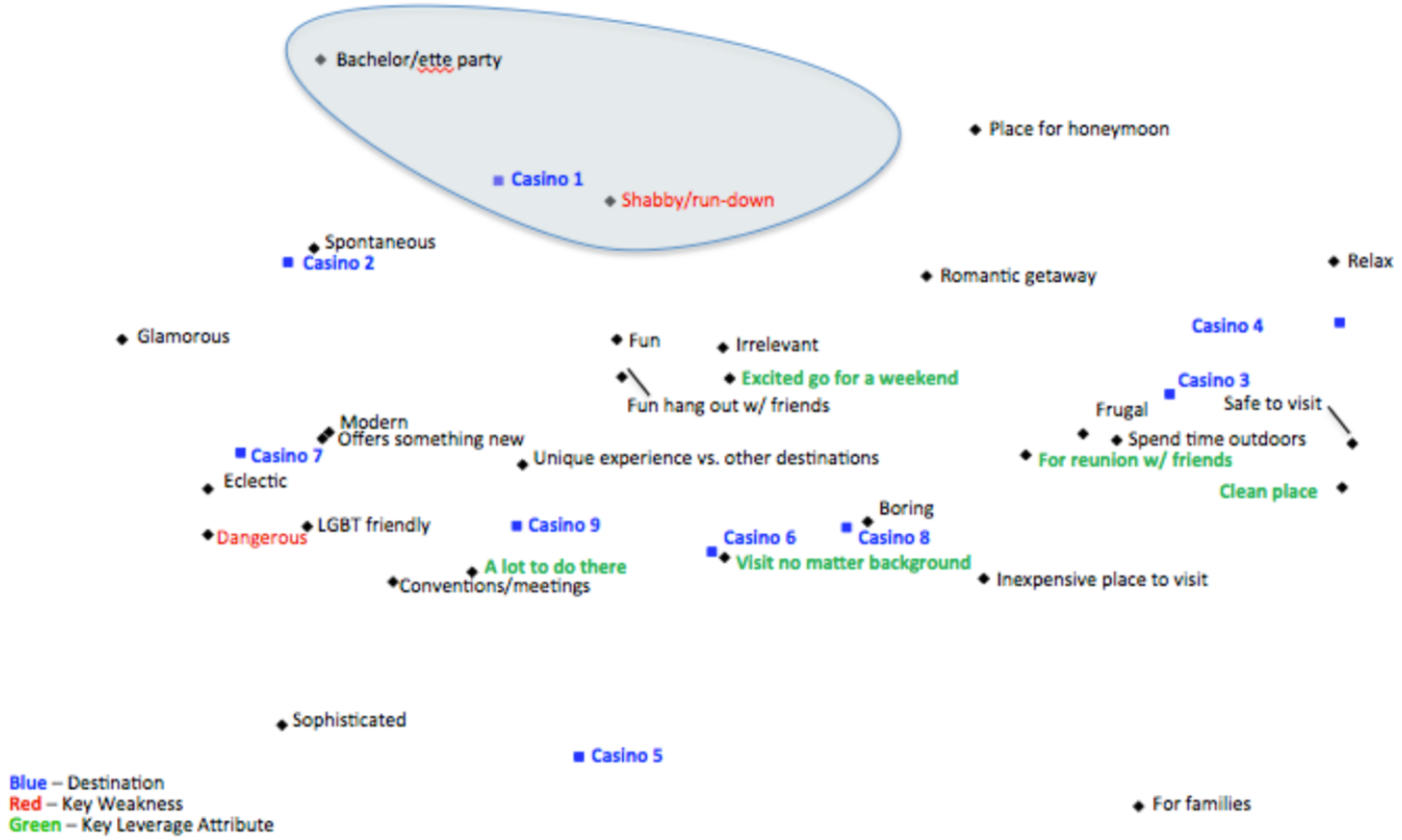
HOW I USE DIM REDUCTION

Understand latent variables for story-telling

Visualize multi-dimensional data

HOW I USE DIM REDUCTION

Visualize multi-dimensional data



LET’S CODE!