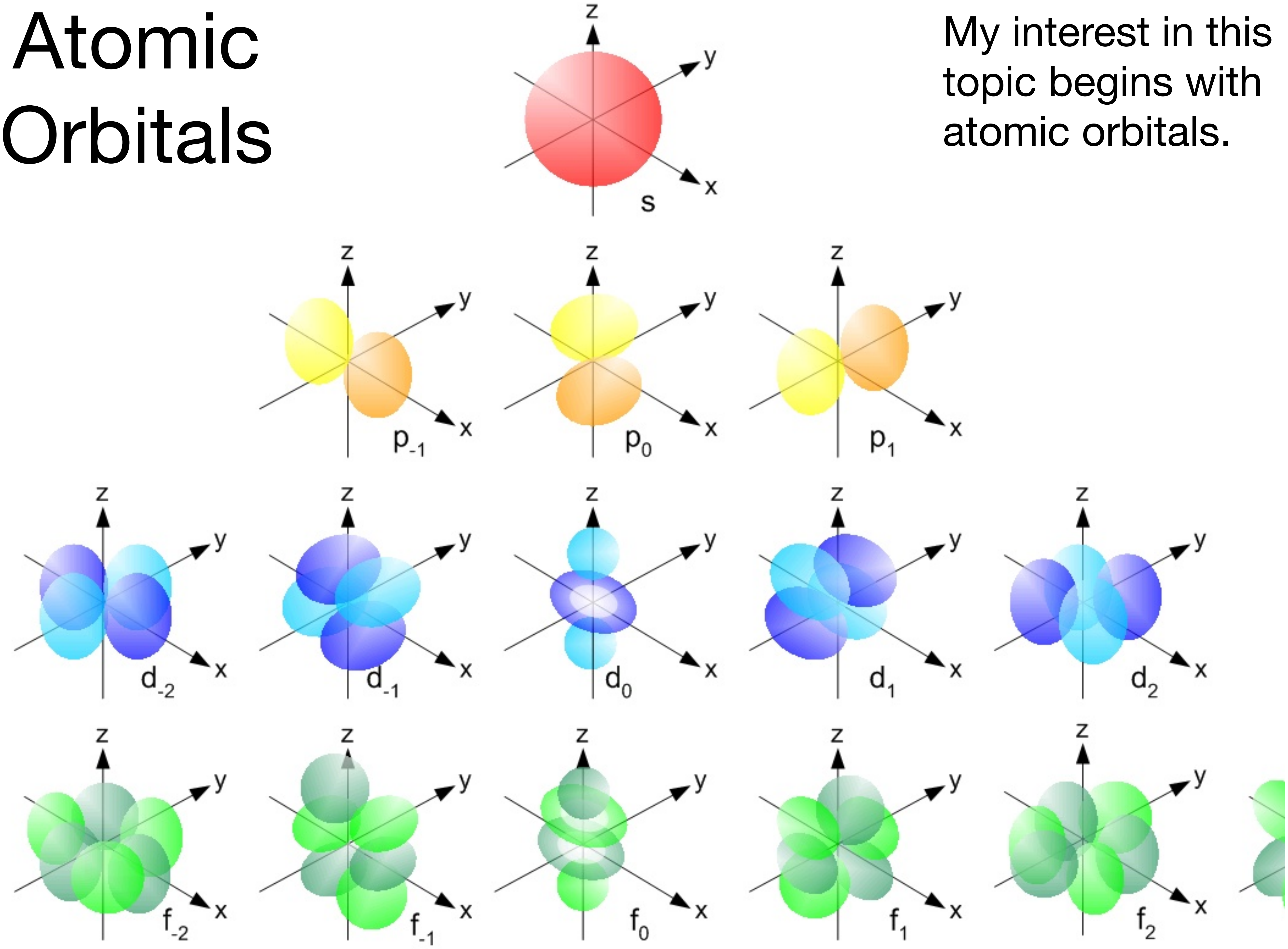
Document Clustering

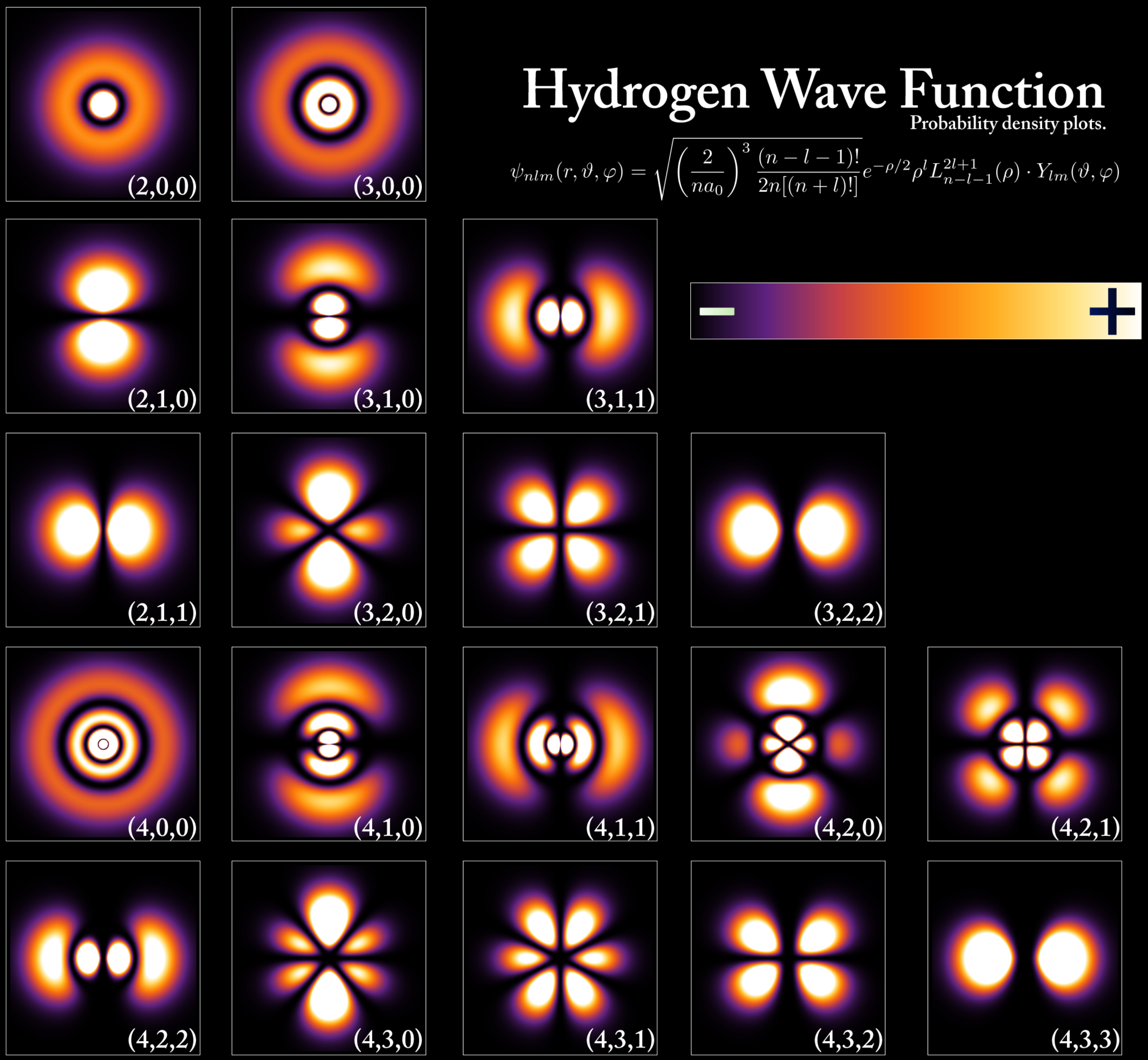
With Open Source Tools

Atomic Orbitals

My interest in this topic begins with atomic orbitals.



Atomic Orbitals



Hydrogen Wave Function

Probability density plots.

$$\psi_{nlm}(r,\vartheta,\varphi) = \sqrt{\left(\frac{2}{na_0}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]}} e^{-\rho/2}\rho^l L_{n-l-1}^{2l+1}(\rho) \cdot Y_{lm}(\vartheta,\varphi)$$

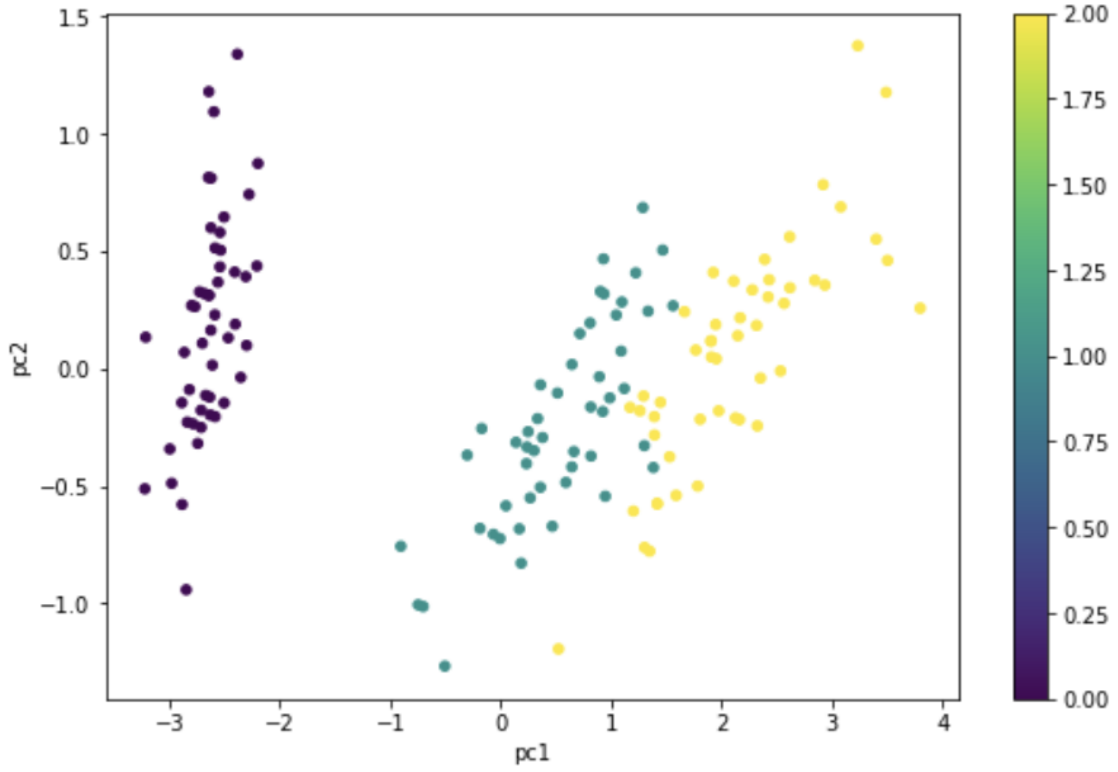
- The atomic orbitals as we understand them are **eigenvector** solutions to the Schrödinger Equation.

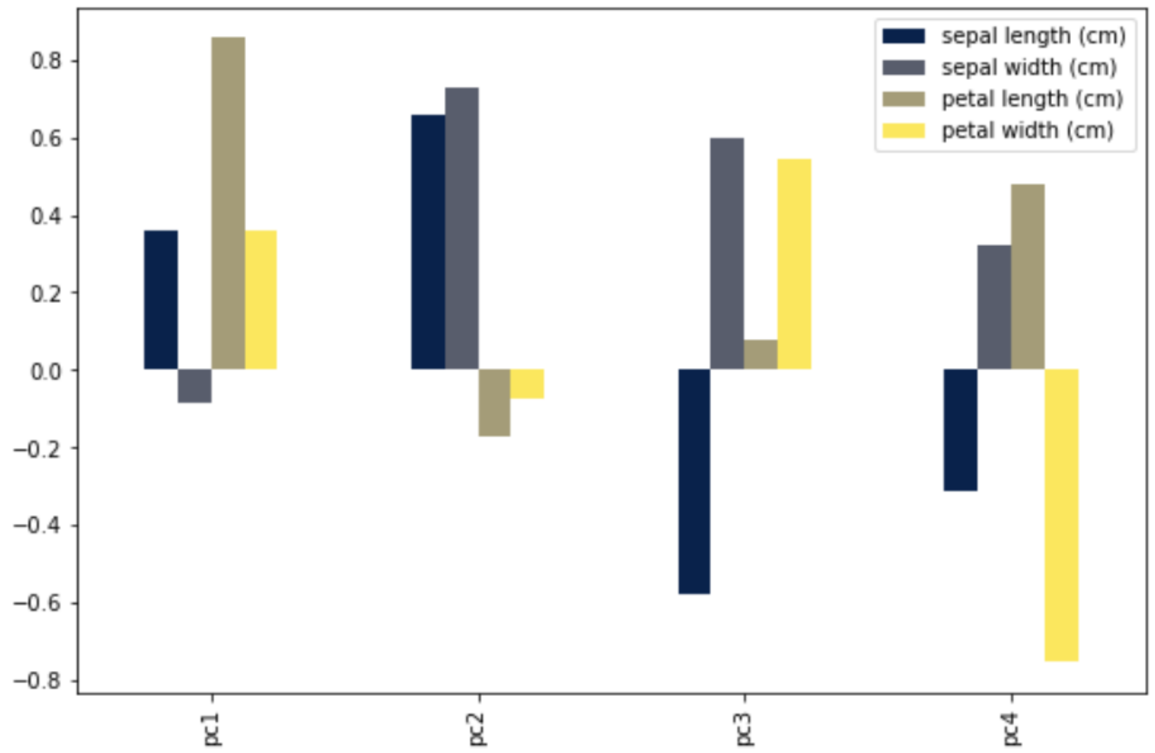
- The word "eigen" meaning "self-essential", these orbitals describe the essential nature of electron behavior.

Unsupervised Learning and Eigenvectors

- Here on the Iris dataset

- No doubt you are familiar with PCA's use to plot Iris in two dimensions, but what about the plot on the right?



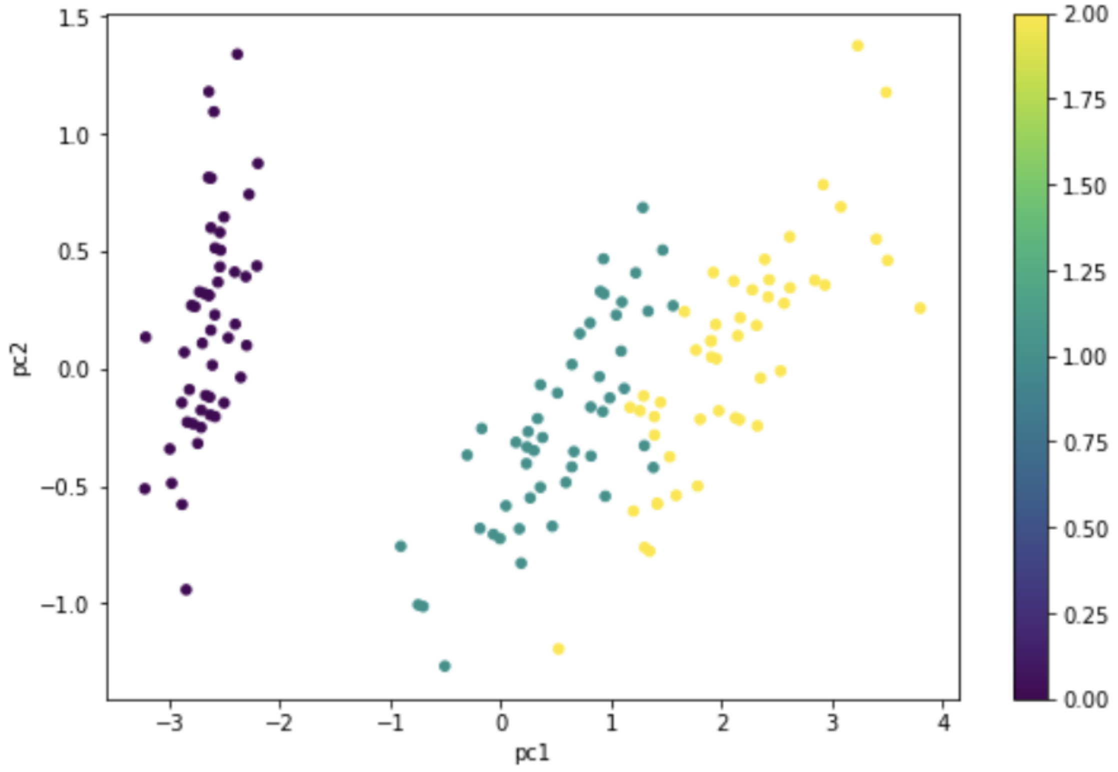


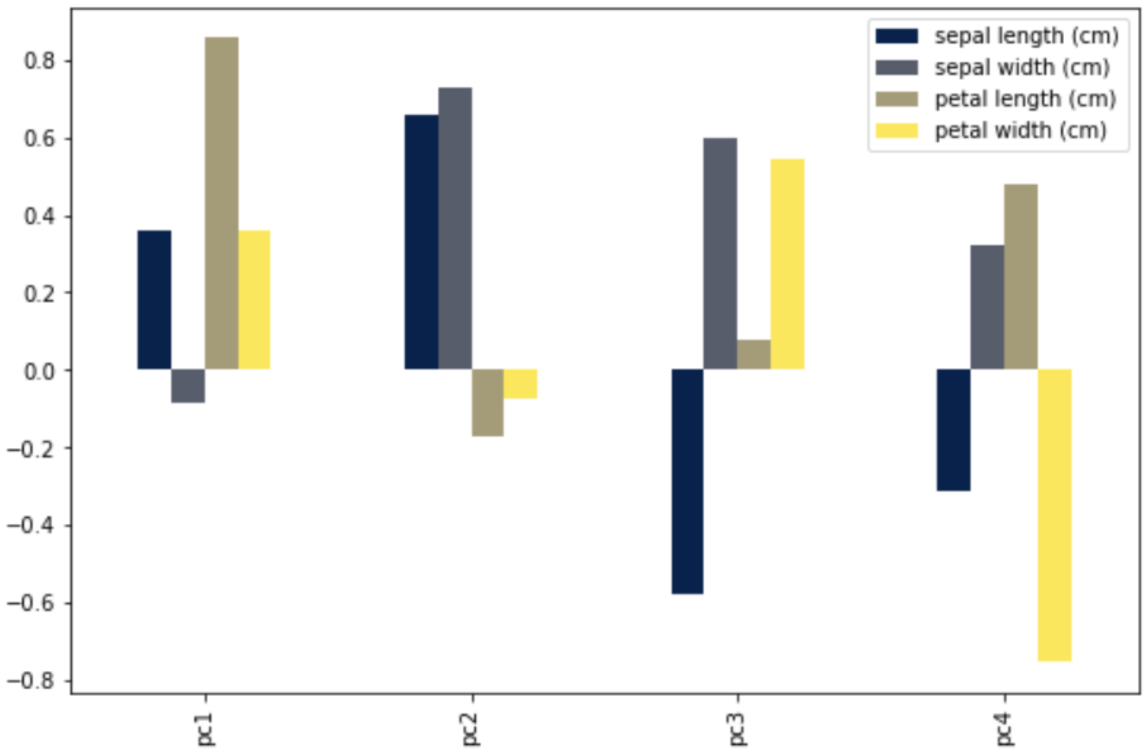
Unsupervised Learning and Eigenvectors

- PCA is typically solved via a matrix factorization to so that if the data is given by a matrix *X*

$$X = U\Sigma V^T$$

- The left makes use of U
  - a reduced dimensional representation
- The right is a plot of V
  - the eigenvectors of the covariance in X
- Furthermore, U is an embedding of X using the eigenvectors as axes.





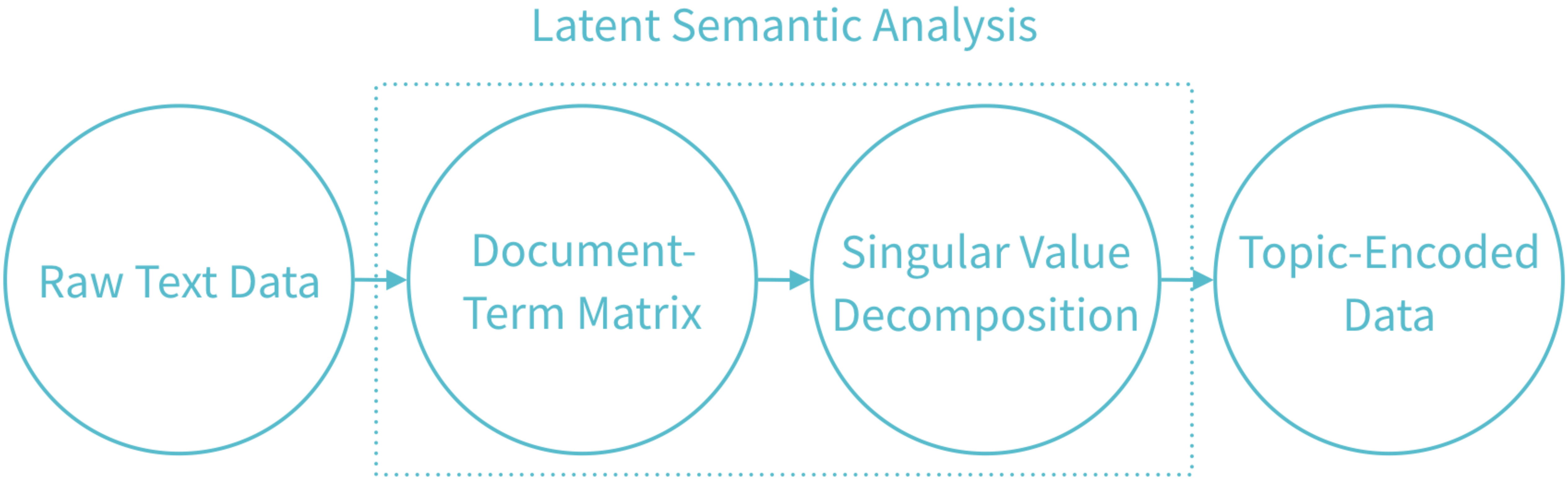
Vocabulary

- **Natural language processing -** refers to a family of techniques used to derive meaning from text data.
- A **document** refers to some collection of words and represents the instances or "rows" of our dataset.
- A **body** is a collection of documents and is our entire data set.
- A **dictionary** is the set of all words that appear in at least one document in our body.
- A **topic** is a collection of words that co-occur.
- The word **latent** means hidden. In this context, we are referring to features that are "hidden" in the data. That they are hidden refers to the fact that they can not be directly measured. These latent features are essential to the data, but are not the original features of the data set.

Latent Semantic Analysis

- an unsupervised, natural language processing technique
- aims to create representations of the documents in a body based on the topics inherent to that body
- consists of two steps:
  - creating a document-term matrix
  - dimensionality reduction via a singular value decomposition

Latent Semantic Analysis

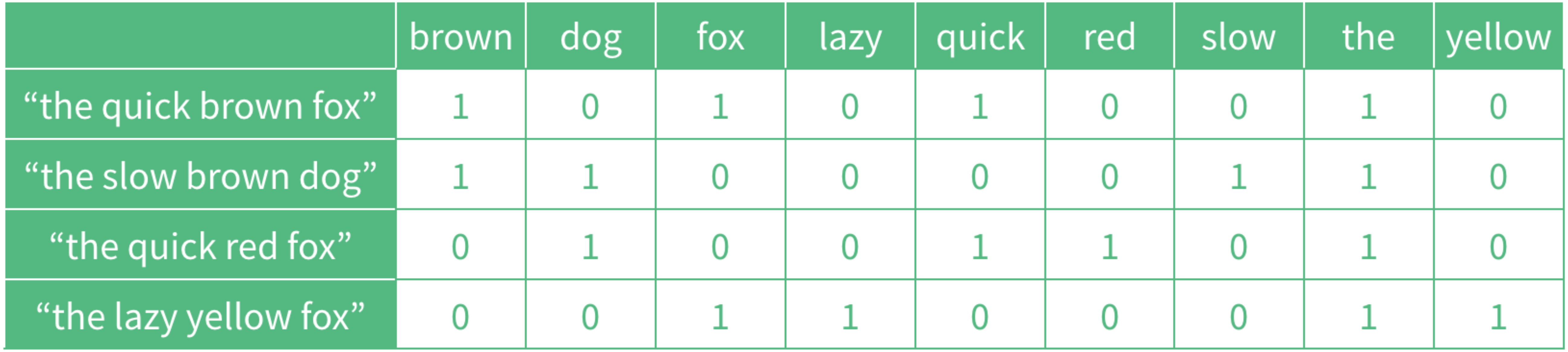


Document-Term Matrix

- A basic idea of a Document-Term Matrix is that documents can be represented as points in Euclidean space aka vectors.

- Here is an example of a document-term matrix.

- Here, each document is a simple statement describing the nature of a canine and defines the rows of our matrix. The dictionary defines the columns of our matrix.



	brown	dog	fox	lazy	quick	red	slow	the	yellow
“the quick brown fox”	1	0	1	0	1	0	0	1	0
“the slow brown dog”	1	1	0	0	0	0	1	1	0
“the quick red fox”	0	1	0	0	1	1	0	1	0
“the lazy yellow fox”	0	0	1	1	0	0	0	1	1

Documents as Vectors

- According to this Document-Term matrix:
    - "the quick brown fox"=(1,0,1,0,1,0,0,1,0)
    - "the slow brown dog"=(1,1,0,0,0,0,1,1,0)
    - "the quick red dog"=(0,1,0,0,1,1,0,1,0)
    - "the lazy yellow fox"=(0,0,1,1,0,0,0,1,1)
- These vectors "embed" these strings in **document space**

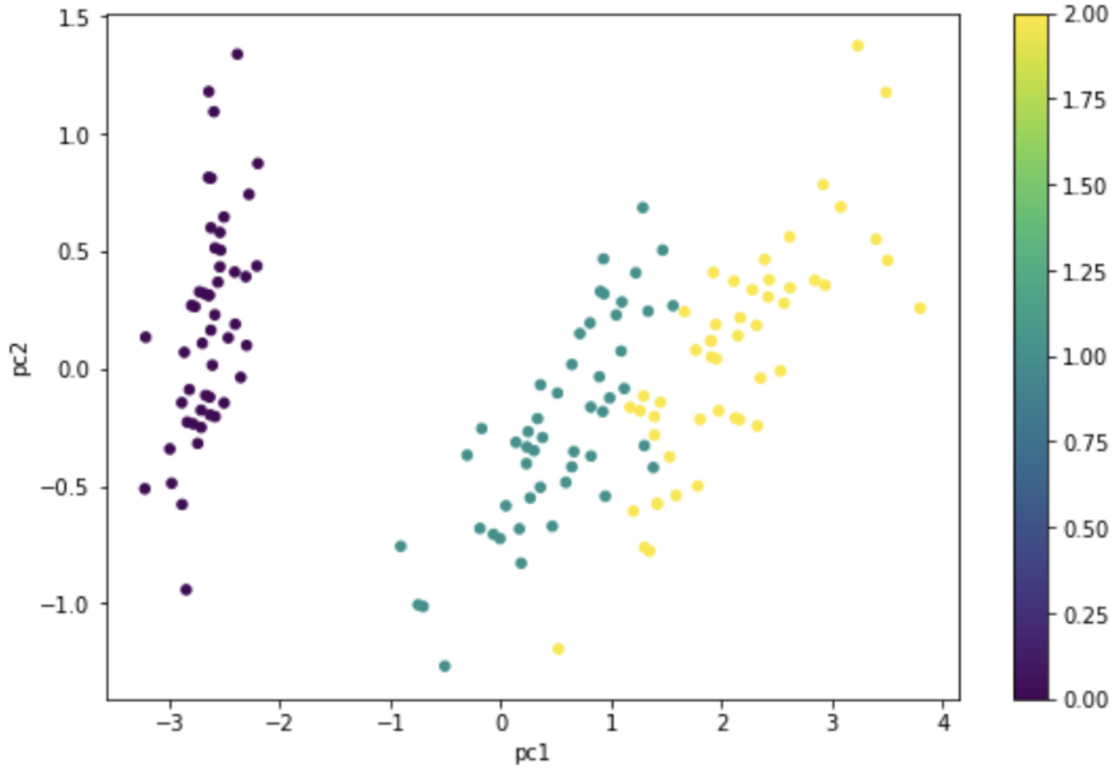
Latent Semantic Analysis

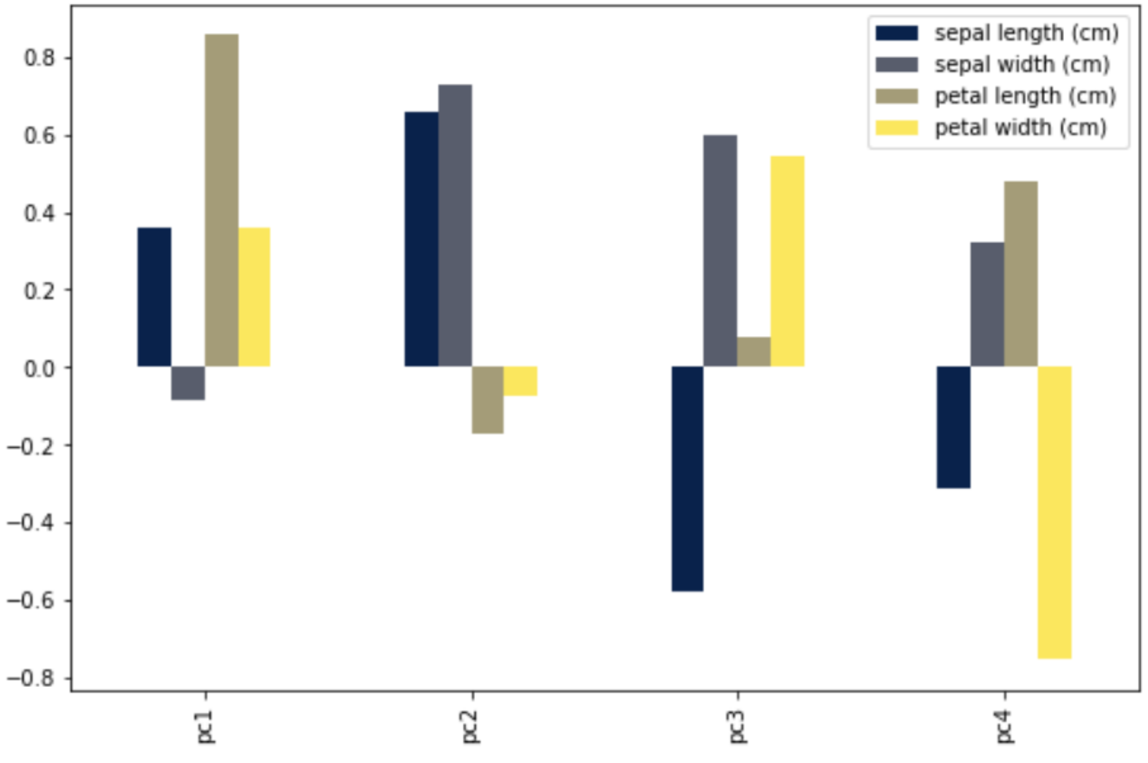
- Similar to PCA, if the encoded text data is given by a matrix *X*

$$X = U\Sigma V^T$$

- U is a reduced dimensional representation, the **documents** embedded in **topic space**

- V is the eigenvectors of the covariance in the **body** of documents





SVD

- Finding the vectors generated by an SVD is analogous to finding the eigenvectors of the covariance matrix of the data. Without going to deeply into the meaning of this suffice it to say that these vectors are deterministic and will not change with subsequent fiitting of the model.

- In other words, as data scientists we have very little control of the output of an SVD. We can choose a number of output vectors, but the number of vectors returned has no bearing on the output. We can not tune the model in the conventional sense by, that is, by adjusting hyperparameters. The only hyperparameter to be tuned has no impact on any one individual vector returned.

Tuning a Latent Semantic Analysis

- The word "eigen" is German for "own" or "inherent". Each eigenvector returned is in some way "of" or "inherent to" the underlying Document-Term Matrix. The LSA is based upon these eigenvectors, but the SVD being deterministic, there is no way to directly alter them.

- What can be done is to alter the Document-Term Matrix itself. One popular method for doing this is to use the term frequency-inverse document frequency algorithm in the preparation of the Document-Term Matrix rather than a simple count.

TFIDF

- A simple count collects term frequency, the number of times a term appears in a document. A TFIDF weighs this term frequency by the inverse of the document frequency, that is, the number of documents in which the term appears.

- Without going into the details of how a TFIDF is calculated we can note conceptually that will help to reflect the importance of a term to a document in the body. If we only use term frequency to measure the importance, it is very easy to over-emphasize terms that appear very often but carry little information about the document, e.g. “a”, “the”, and “of”. If a term appears very often across the body, it means it doesn’t carry special information about a particular document. Inverse document frequency is a numerical measure of how much information a term provides.

- Here, we will look at using TFIDF to generate the Document-Term Matrix and how it effects the LSA.