DATA SOCIETY:

# Topic Modeling in NLP - TfIdf - 1

*One should look for what is and not what he thinks should be. (Albert Einstein)*

# Tf-Idf: Topic introduction

In this part of the course, we will cover the following concepts:

- The “bag-of-words” approach and when it is used
- The need for weighting terms in a corpus
- Implementation of Tf-Idf weighting on a corpus of documents

# Module completion checklist

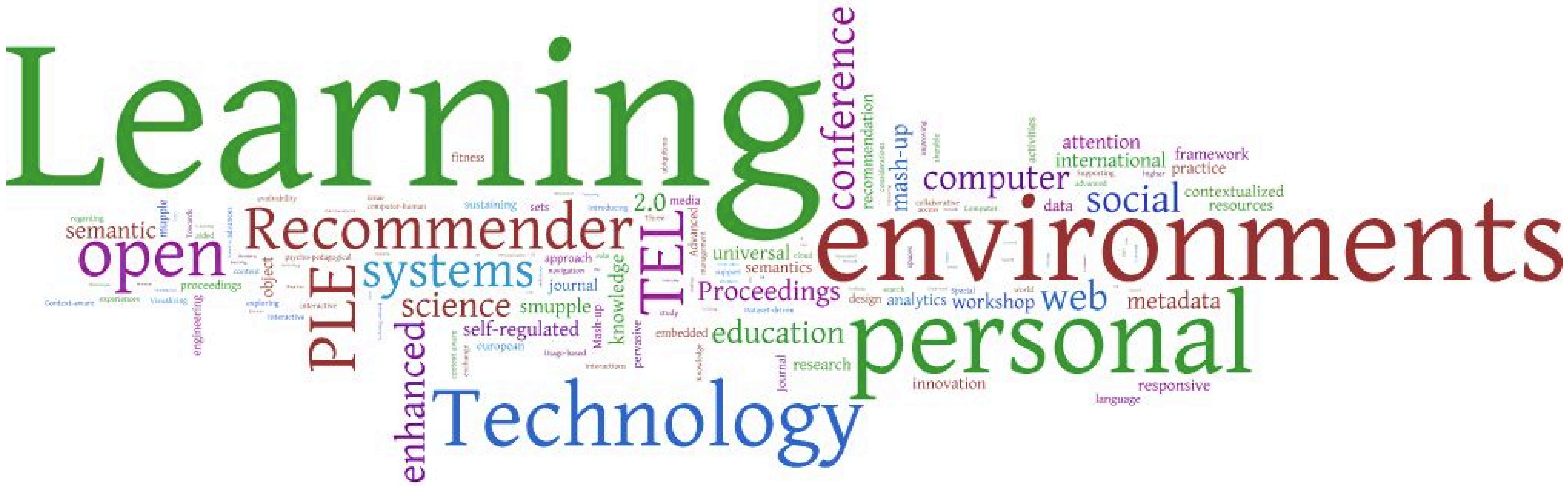
| Objective | Complete |
|---|---|
| Explain use cases for bag-of-words approach | |
| Summarize supervised vs. unsupervised learning | |

# The importance of word counts

- **TF-IDF** (term frequency-inverse document frequency) is a statistical measure that evaluates how **relevant** a specific word is to a document in a collection of documents
- This is done by multiplying two metrics: how many times a word appears **within a document**, and the inverse document frequency of the word **across a set of documents**
- In short, it's about which words are really important and stand out

# The importance of word counts

- The **word cloud** below visually represents the words with the greatest significance in a set of documents
- What **topic** would you guess this set of documents is about?
- **Share your response** in the chat box



Source: CC-BY-SA-3.0

# The math of relevance

- Algorithms need to do this with mathematical means **(counting frequency)** in Natural Language Processing (NLP) to make up for the lack of base knowledge that humans have to make a quick determination of **relevance**
- Humans also know that while certain words, like “the” and “and” are more frequent, they are never the topic of a text

# TF-IDF: Topic Overview

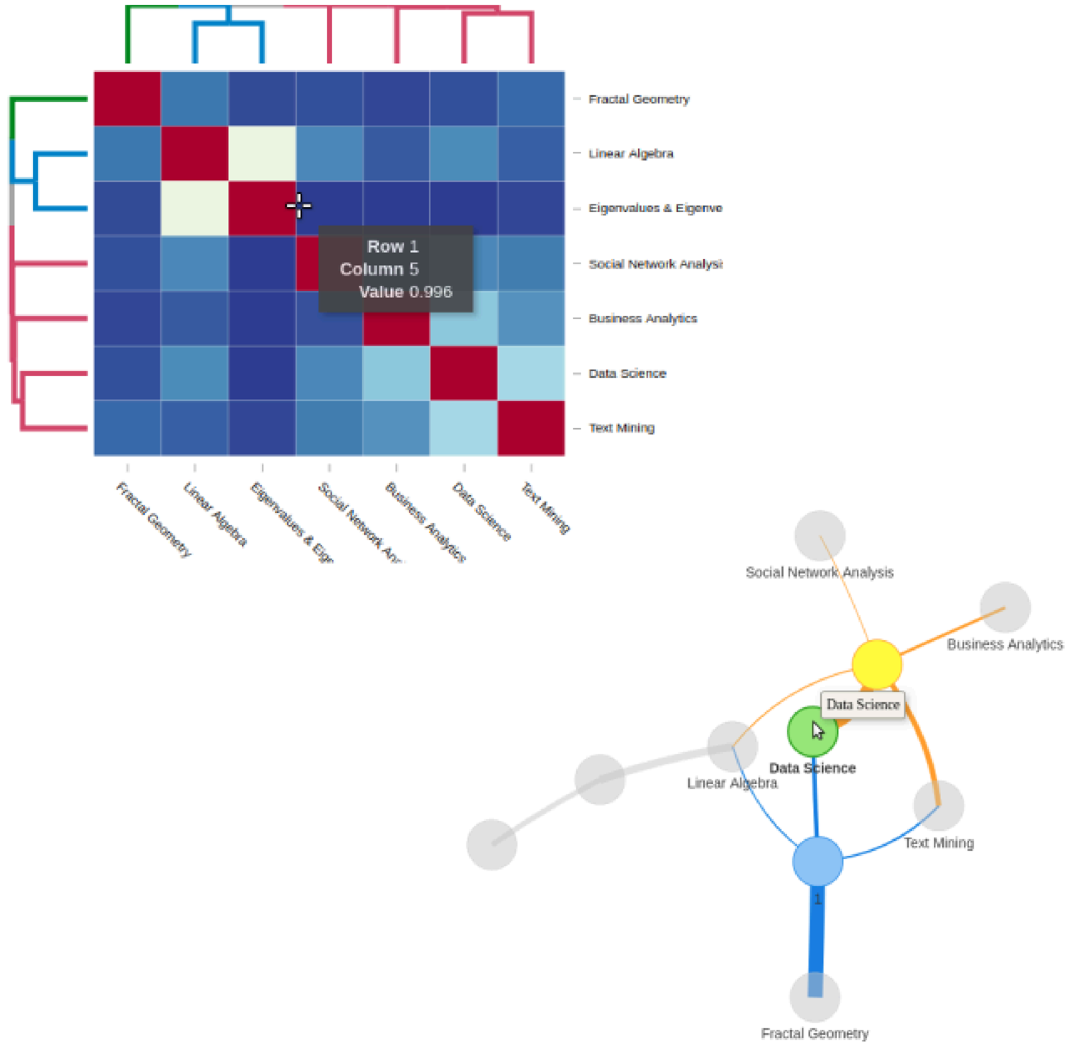
- The distribution of words in a corpus and in a language is **highly skewed right**, which indicates that few words have very high frequencies and most words have very low counts
- A Term Frequency - Inverse Document Frequency (TF-IDF) is a famous transformation of text data used to battle the skewness of the word distribution in a corpus
- The TFIDF can then be used in the Bag-of-words analysis for text mining or other NLP tasks

# `Snippet` analysis

- In order to pre-process the data for TF-IDF, the steps to be taken are:
  - **Load** the corpus, where each ‘document’ is actually one entry in `snippet`
  - **Clean** the text, removing punctuation, numbers, special characters and stop words
  - *Stem* the words to their root forms
  - **Create** a Document-Term Matrix (DTM) with counts of each word recorded for each document
- Now we will build the final, optimized matrix - a weighted **T**erm **F**requency - **I**nverse **D**ocument **F**requency (TF-IDF) by
  - **Transforming** the DTM to be a weighted TF-IDF

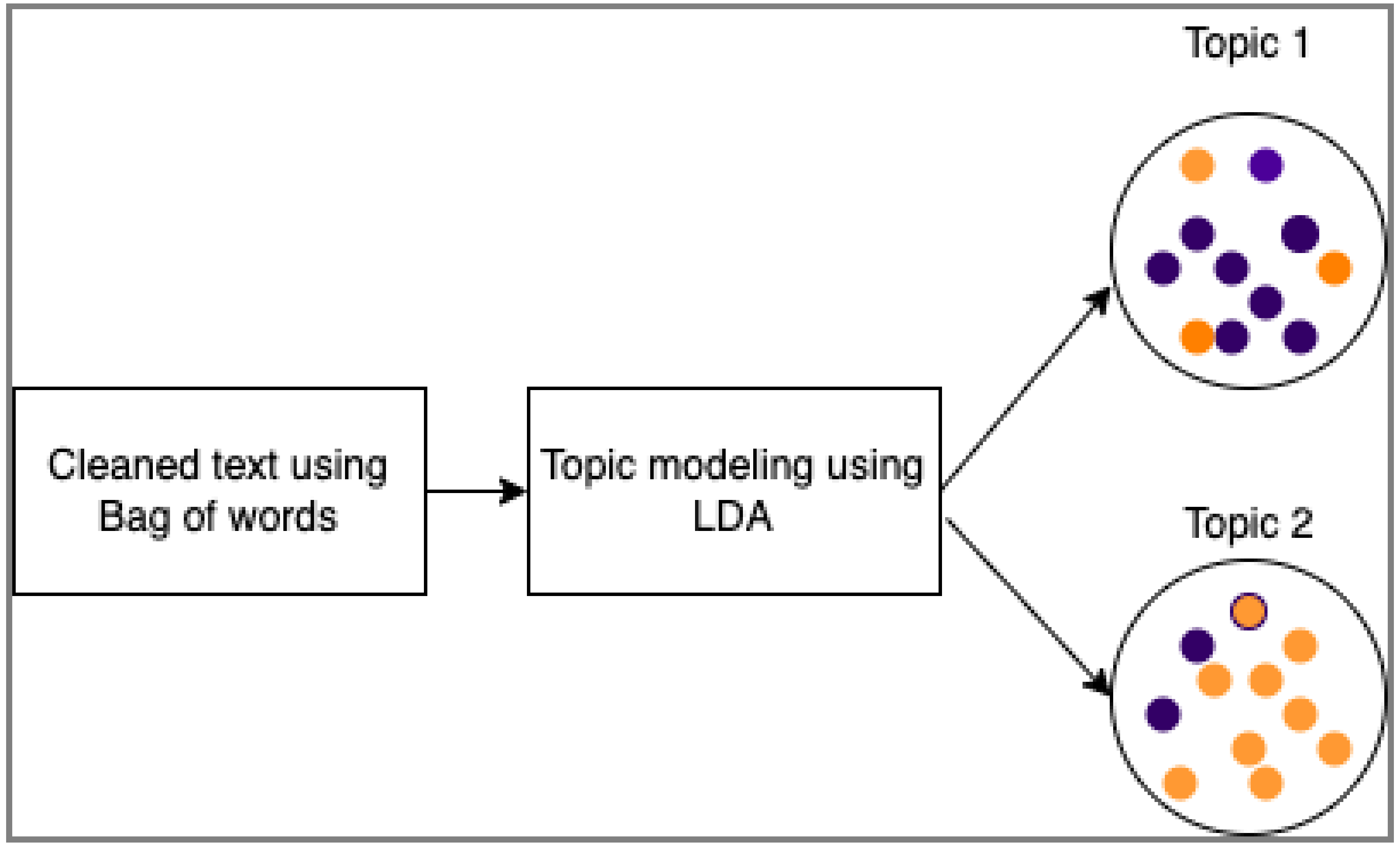
# “Bag-of-words” analysis: use cases

- What can be done with such a seemingly *crude* approach?
- Quite a few things, actually! They include:
  - Topic modeling
  - Word and document similarity query processing
  - Word and document clustering
  - Sentiment analysis
  - Automated document summarization



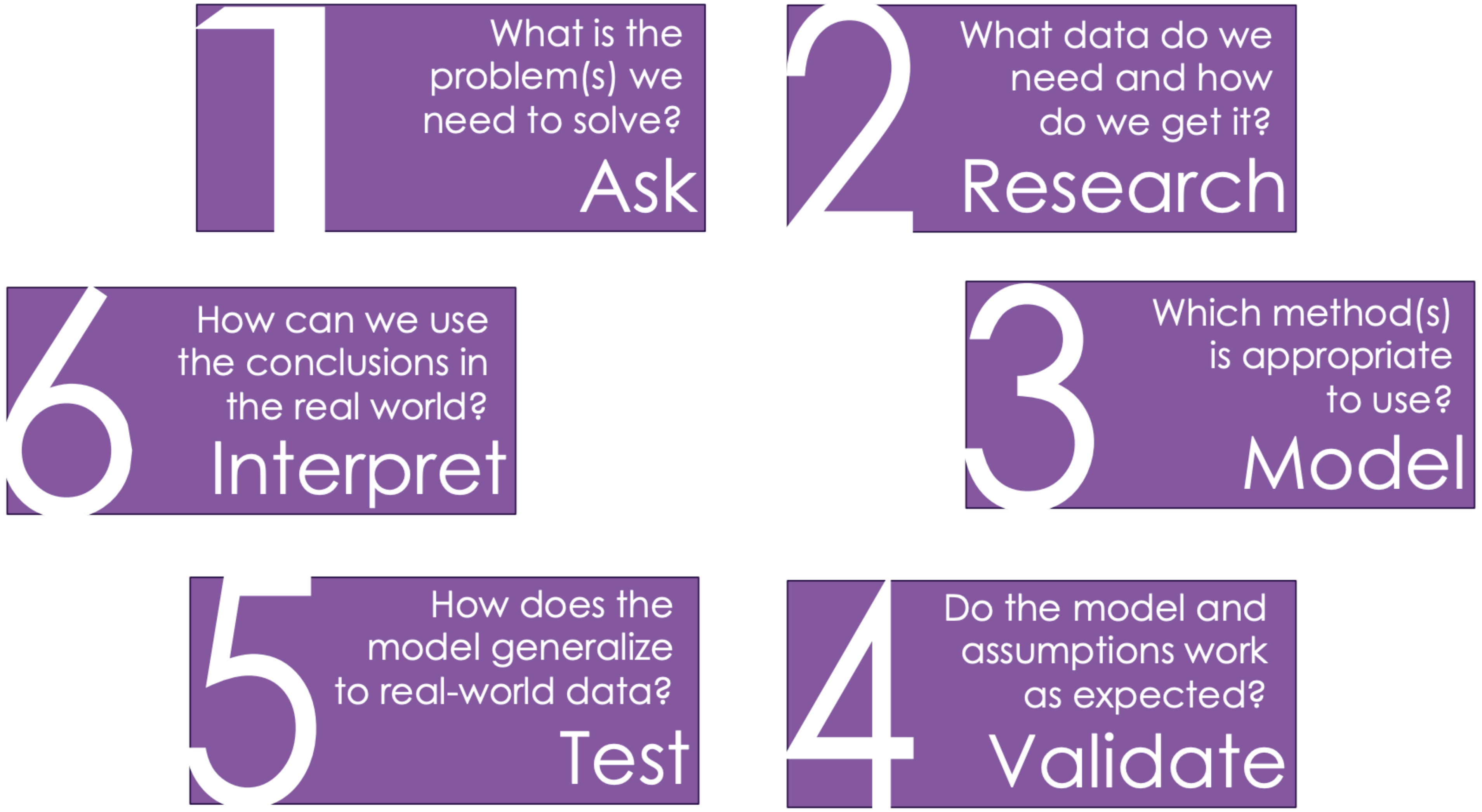
# “Bag-of-words” analysis: `snippet`

- We are going to dive deeper into one of these use cases with **topic modeling**
- We are going to implement a very popular method called **Latent Dirichlet Allocation (LDA)**
- **This will help us understand broad topics within our corpus of documents**



# Data exploration

- A data scientist must be able to **explore** data to generate a hypothesis



1 What is the problem(s) we need to solve? Ask

2 What data do we need and how do we get it? Research

3 Which method(s) is appropriate to use? Model

4 Do the model and assumptions work as expected? Validate

5 How does the model generalize to real-world data? Test

6 How can we use the conclusions in the real world? Interpret

**What stage of the Data Science Control Cycle (DSCC) would this goal fit into?**

# DSCC: modeling

- **Have you encountered any type of text analysis models?**
- For text data, frequently the first step in model building is to use **unsupervised learning**

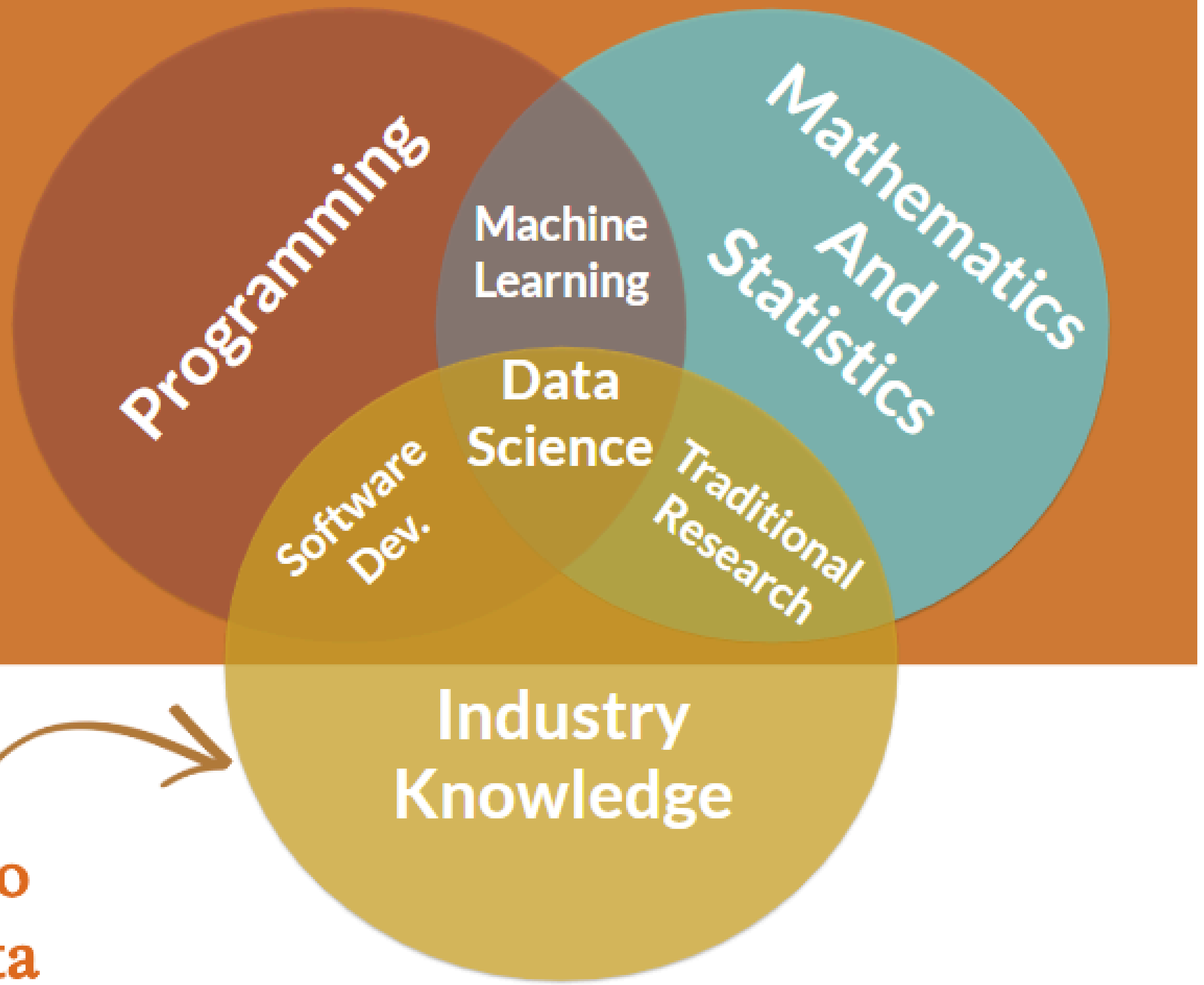
- Model, by definition, is a replica of a real thing
- Using a ship to imitate a train won’t cut it
- Select a **model that suits your problem/data** or **simulates the real-life situation** in the closest possible way

# Module completion checklist

| Objective | Complete |
|---|---|
| Explain use cases for bag-of-words approach | ✔ |
| Summarize supervised vs. unsupervised learning | |

# What is data science?

- Data science applies the scientific method to analyzing data
- It lies at the intersection of several disciplines
- It draws on industry knowledge that makes the analysis of Big Data possible



Industry knowledge is essential to knowing what to look for when exploring data

# What can data science do?

Data science can be applied across any industry

Empowers management to make better decisions

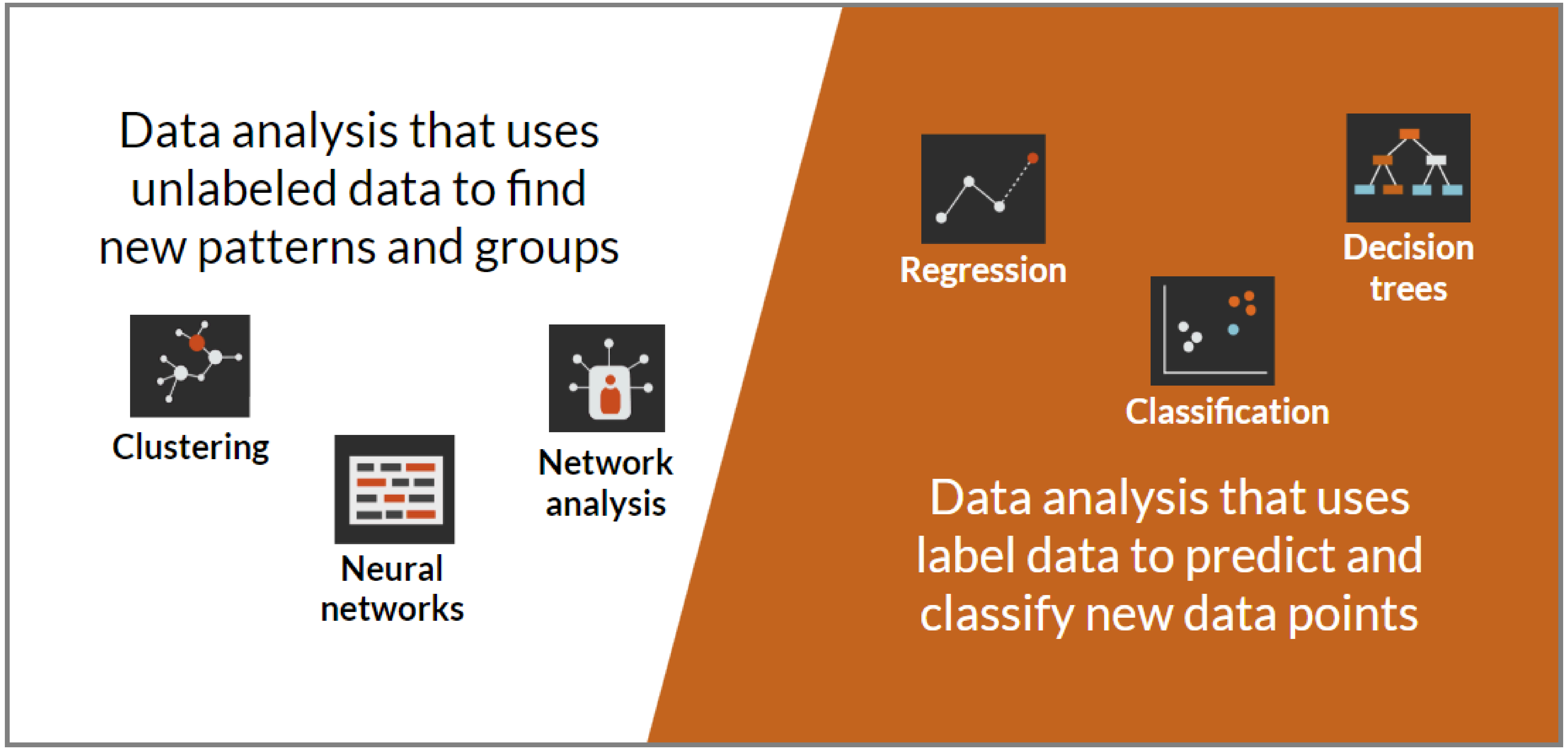
Increases operational efficiency and investment from staff

Increases accountability and validates decisions

Identifies new opportunities to stay competitive

# Unsupervised vs supervised

- What types of text analysis would fall under unsupervised learning?
- What types of text analysis would fall under supervised learning?



Data analysis that uses unlabeled data to find new patterns and groups

Data analysis that uses label data to predict and classify new data points

# Text analysis: unsupervised text analysis

- How does text analysis fall into the category of **unsupervised learning**?
  - In topic modeling, **topics** are formed from **unlabeled** data
  - **It involves weighing and clustering documents into “topics”**
  - **Clustering is one of the best known unsupervised techniques**
- We will learn how to **transform our DTM to a TF-IDF weighted matrix**

# Knowledge check

# Module completion checklist

| Objective | Complete |
|---|---|
| Explain use cases for bag-of-words approach | ✔ |
| Summarize supervised vs. unsupervised learning | ✔ |

# Congratulations on completing this module!