Data Science: Introduction

MIE223

Winter 2025

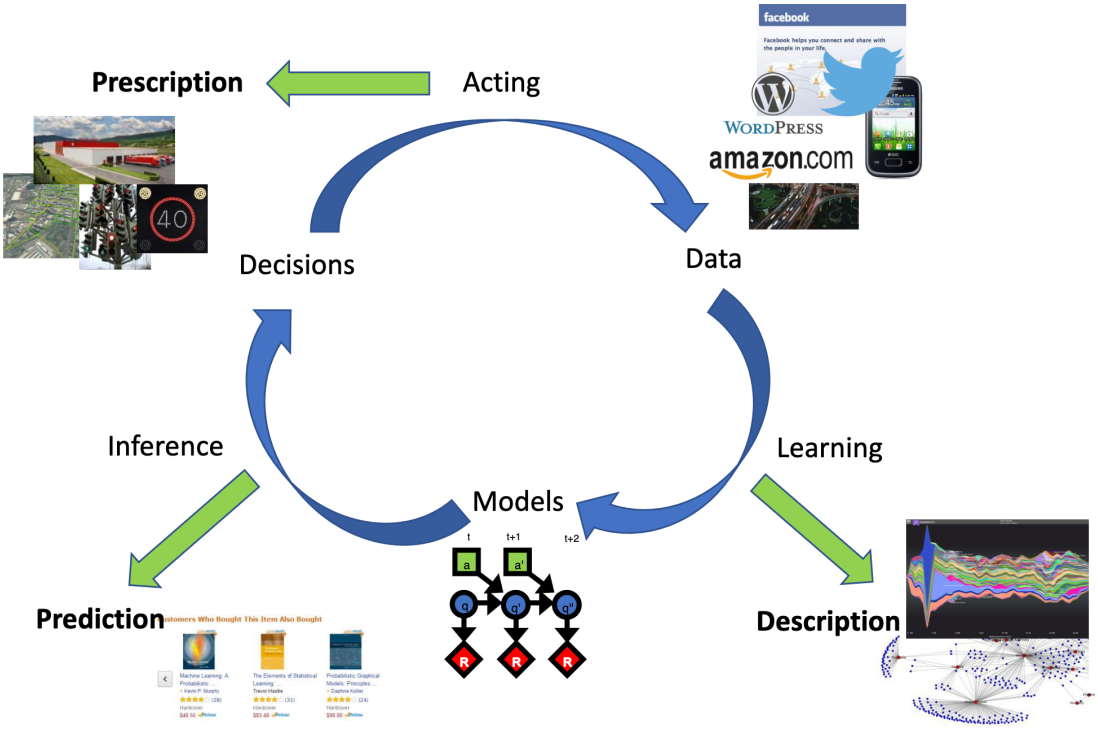
1 Introduction

1.1 Data Science

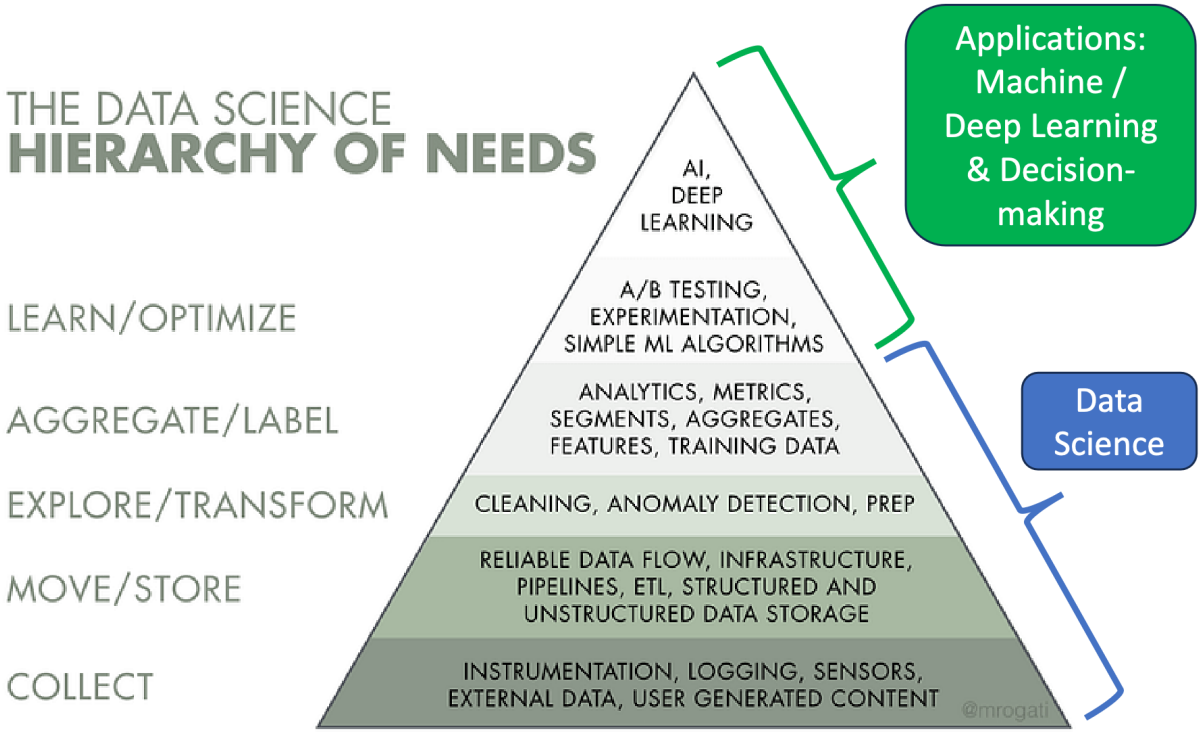
We use data to make data-driven decisions.

1.2 Data Analytics

- Description
- Prediction
- Prescription



1.3 Data Science Hierarchy of Needs



https://towardsdatascience.com/the-data-science-pyramid-8a018013c490

2 The Science of Data

2.1 Data are Observations of a Generative Process

- Data are just the observables of a generative process
  - We observe election votes of a population
  - We observe Gallup poll survey of a population's voting disposition
  - But what process generated both?
- Example: 2016 US Presidential Election
  - Gallup Poll Surveys predicted a landslide for Hillary Clinton
  - We know how the election turned out
  - Why was there a discrepancy?
    - Sample bias in surveys
    - Human behavioural biases in revealing their true preferences
- We care about the (latent) variables that generate the data
  - We can use understanding of this generative process for prediction
  - Aside: critically connected to modern perspective of Generative AI

2.2 Methodology of Data Science

- Start: Real-world task or problem (business needs, societal or scientific problem)
  - E.g, investigate sales performance, transit access and equity, pollution impacts on society
- Develop research questions (RQs) and hypotheses
  - Can we predict seasonal trends in perishable demands of food to reduce waste & sales loss?
  - Is public transit access equitably allocated among high and low income areas of a city?
  - Does increased pollution lead to increased mortality rates?
- Collect relevant data to test the hypotheses and answer the RQs
  - Need to collect data or extract it from existing sources (Data Engineering, ETL)
  - Need data for RQ that is representative of downstream use (performance, fairness & bias)
- Clean the data!
- Perform exploratory analysis on the data (feature analysis and visualization)
  - Consider revisiting the data collection and cleaning process based on this data
- Evaluate the original hypotheses and RQs w.r.t. the data, iterate as needed
- Finish: Report back to stakeholders (decision-making)

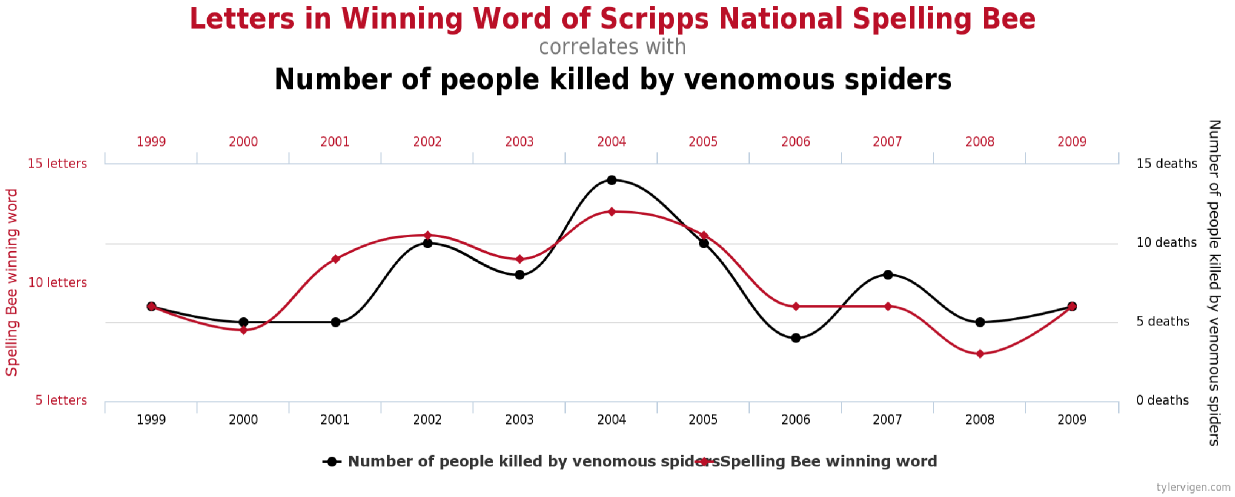
2.3 Data as a Science

Data Science starts with a task and questions e.g. does increased pollution lead to higher mortality?

- Data is a Science
  - Science is the Scientific Method (the "why")
  - Engineering is the "how"
- Recap: Scientific Method
  1. Ask a Question
  2. Make a Falsifiable Hypothesis
  3. Design an Experiment to test the hypothesis
  4. If Hypothesis is falsified, go to step 2
- Be careful, details matter and many things can go wrong!
  - E.g., statistical fallacies arise from multiple hypothesis testing
    - The more datasets you compare, the more likely that two randomly correlate
    - Can mitigate this with a Bonferroni Correction

2.4 Spurious Correlations and Multiple Hypothesis Testing

Note 1. Spurious: comparing so many datapoints that it is bound to correlate. Correlation is **not** conducive of causation.



3 Data Comes in All Shapes and Forms

Need proficiency in how to leverage structure of data and understanding of nuances of handling each data type

3.1 Tabular Data (i.e., Pandas)

Note 2. NaN: Not a Number.

- Passengers on the Titanic

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

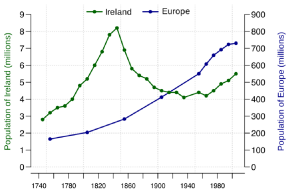
-

3.2 Text Data and Analytics

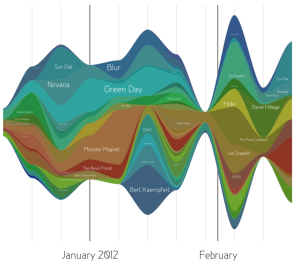
- 75% of Data most IEs work with is text!
- Sources:
  - Customer Feedback
  - Reviews
  - Blogs
  - Social Media
  - Academic Publication
  - Healthcare / records
  - News

3.3 Time Series Data

Due to the famine and ex-immigration, the population in Ireland decreased while Europe's increased.



(a) Population of Europe and Ireland vs. Year (1740-1990)

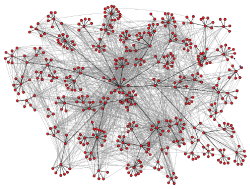


(b) Stream Graph of a Last.fm user's Time Spent Listening to Different Artists

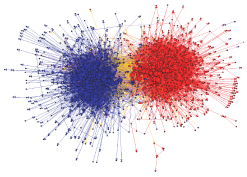
Figure 1: Time Series Data

3.4 Network / Graph Data

Communication is aligned with management structure.
Political blogs live in their own echo chambers.



(a) HP Labs Manager and Email Network



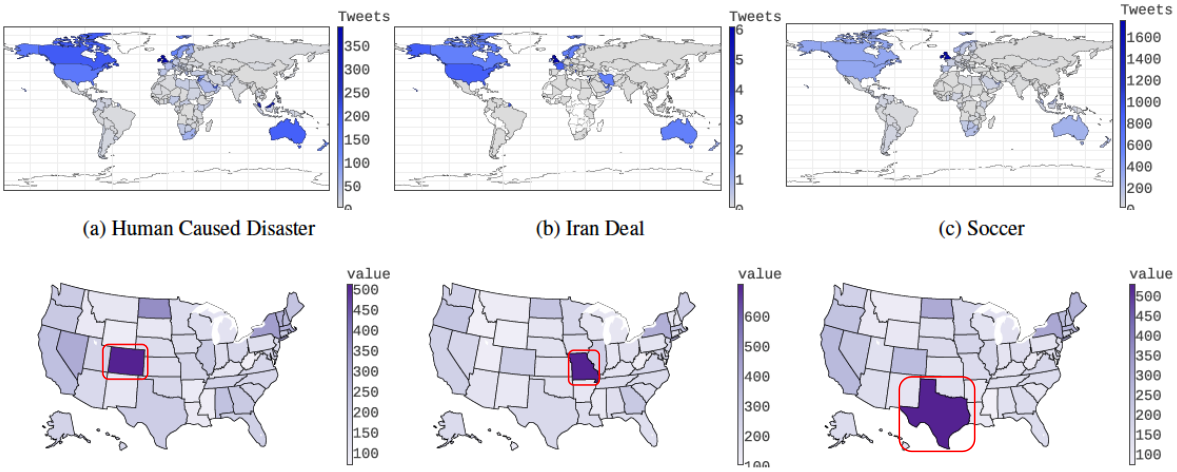
(b) Link Structure of Political Blogs

Figure 2: Network / Graph Data

3.5 Knowledge Graph Networks: Ingredients

Edges connect combined ingredients. There becomes two main clusters of ingredients that are commonly used together. The two clusters are sweet and savoury ingredients.

3.6 Geographical Data (Choropleths)

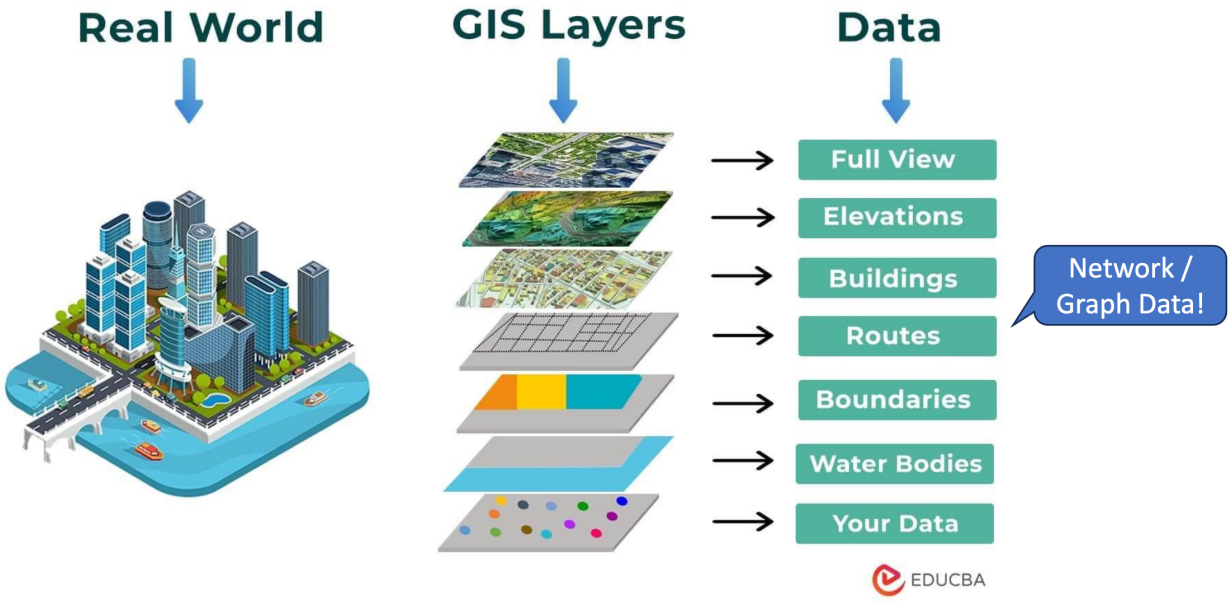


(a) Human Caused Disaster (b) Iran Deal (c) Soccer

(d) Health Epidemics (e) Social Issues (f) Space

Figure 3: Per capita tweet (X) frequency across different international and U.S. locations for different topics

3.7 Geographical (GIS) Data



4 Content Highlights

4.1 Data Cleaning and Exploratory Data Analysis

- Data Cleaning
  - Data summary
    - Simple column analysis (basic statistics) and univariate visualization (histograms)
    - Correcting (mixed) types, validating ranges, checking for outliers
  - Dealing with missing data
    - Types of missingness, imputation
  - Data transformation
    - Smoothing, aggregation, generalization, derived features, normalization (e.g. [min,max] or z-score)
- Exploratory Data Analysis
  - Feature Analysis
    - Correlation, (Pointwise) Mutual Information, Cross-Entropy
  - Visualization
    - Visualizations for discrete and continuous data
    - Visualizations for univariate (histogram), bivariate (e.g., scatterplot), and multivariate data (bubble plot)
    - Custom and interactive visualization!

4.2 Working with Structured Data

- Text data and natural language processing
  - The NLP pipeline and token processing
    - Beyond tokens to phrases (easier to understand)
  - Sentiment (polarity and beyond!)
  - The power of PMI!
- Time series data
  - Smoothing
  - Autocorrelation and cross-correlation
- Network data (Knowledge graph and social networks)
  - Centrality measures
  - Betweenness clustering
- Geographical data
  - Shapefiles and chloropleths
  - Vector vs. raster data
  - Coordinate Reference Systems

4.3 Data Science in Practice

- Advanced Data Science and Fallacies
  - Be careful... a lot can go wrong!
- Bayesian Data Science
  - How do we model complex generative data processes
  - And infer latent properties of those models (especially with “small” data)
- Privacy and Anonymity
  - Releasing data is help to society and business
  - But we need to maintain privacy obligations for this data
- Fairness and Bias (not tested)
  - Data is biased
  - How do we ensure fairness in the use of the data
  - We cannot have it all, so we need to make our fairness (parity) decisions carefully