



VISUALISATION

THEME 2: EXPLORE





HELLO!

- Dr Jocelyn Mazarura
- Theme 2: Explore (Visualisation, Data Transformation & Exploratory Data Analysis)

OUTCOMES

6.2.1. *Visualisation.* Outcomes: At the end of this study unit, the student should be able to:

- Implement the basic structure of a matplotlib plot
- Identify common plot types
- Create presentation quality statistical graphs, including appropriate use of scale, color, labels, reference markers



WHERE DOES DATA SCIENCE START?



It all starts with a question...

Context

- Will I win the election?
- What is the average 'time to repair' since a client logged a call?
- Why are flight delays longer in winter than in summer?
- What do our customers think of the new brand campaign?
- How can I increase production on my farm?
- Do my employees have a gender bias?
- Does this new drug work?

How do I answer the question with data?

"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

How do I answer the question with data?

"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

"Can we identify patterns or factors that are correlated with high employee turnover?"

"Are there specific groups of employees that are more likely to leave, and what are the characteristics of those groups?"

"Can we predict which employees are at risk of leaving in the near future?"

How do I answer the question with data?

"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

Question: "Can we identify patterns or factors that are correlated with high employee turnover?"

Model: *Logistic Regression*

Build a logistic regression model to analyze the relationship between various employee attributes (e.g., job role, salary, tenure, satisfaction) and turnover, identifying significant predictors.

How do I answer the question with data?

"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

Question: "Are there specific groups of employees that are more likely to leave, and what are the characteristics of those groups?"

Model: *Cluster Analysis*:

Utilize clustering algorithms to group employees based on their characteristics and behavior, helping to identify high-risk groups for turnover.

How do I answer the question with data?

"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

Question: "Can we predict which employees are at risk of leaving in the near future?"

Model: *Predictive Modeling*

Develop a predictive model (e.g., using random forests, gradient boosting) to forecast the likelihood of an employee leaving within a certain time frame, using features such as recent performance reviews, engagement levels, and tenure.

How do I answer the question with data?


"How can we reduce the high rate of employee turnover in our company?"

Step 1

Refine the question to one (or more) answerable with data

By addressing these data science questions, the business can gain insights into the factors influencing employee turnover, predict turnover risk, and design effective strategies to retain valuable employees.

Context

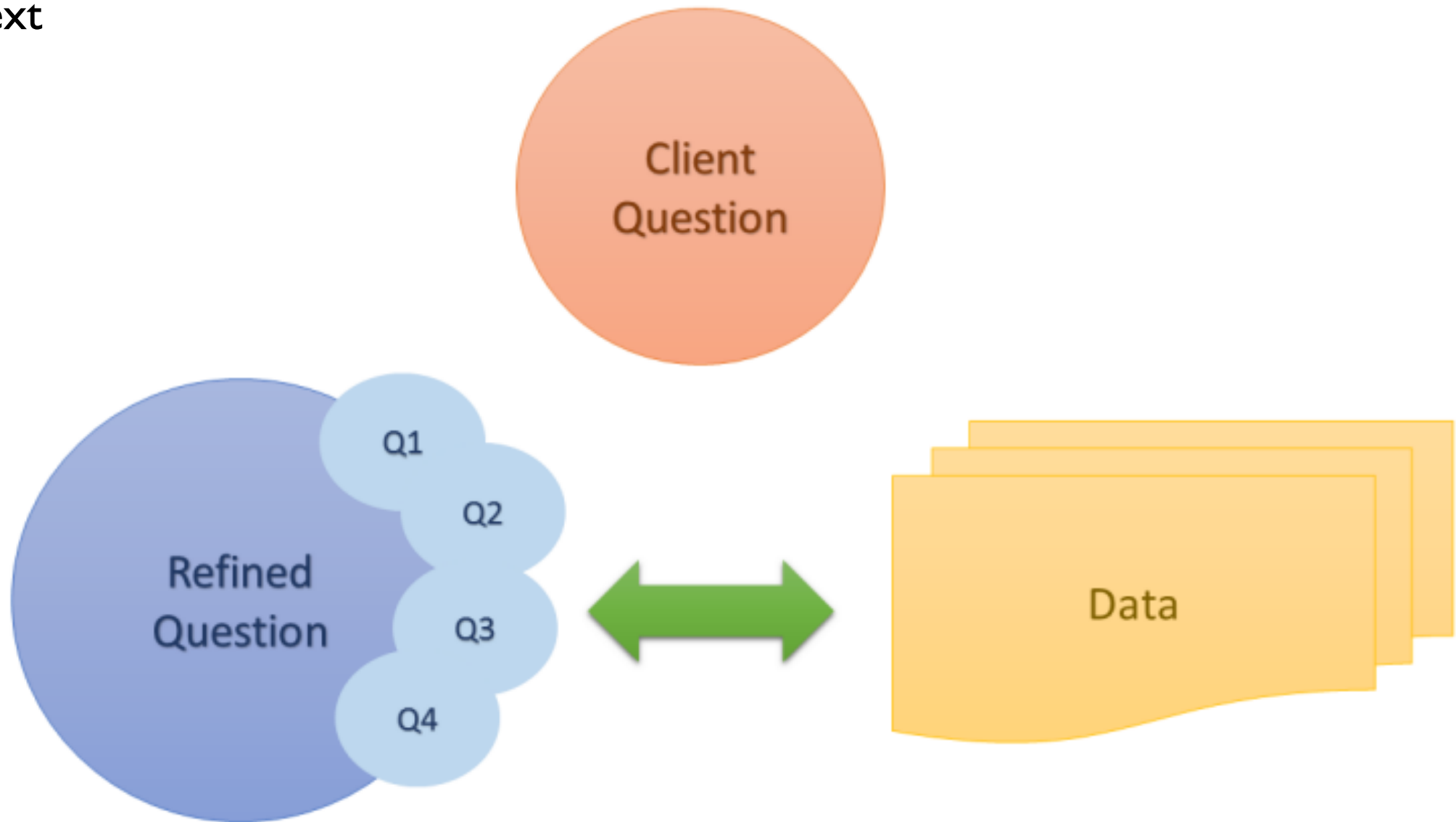


Client
Question

Context



Context



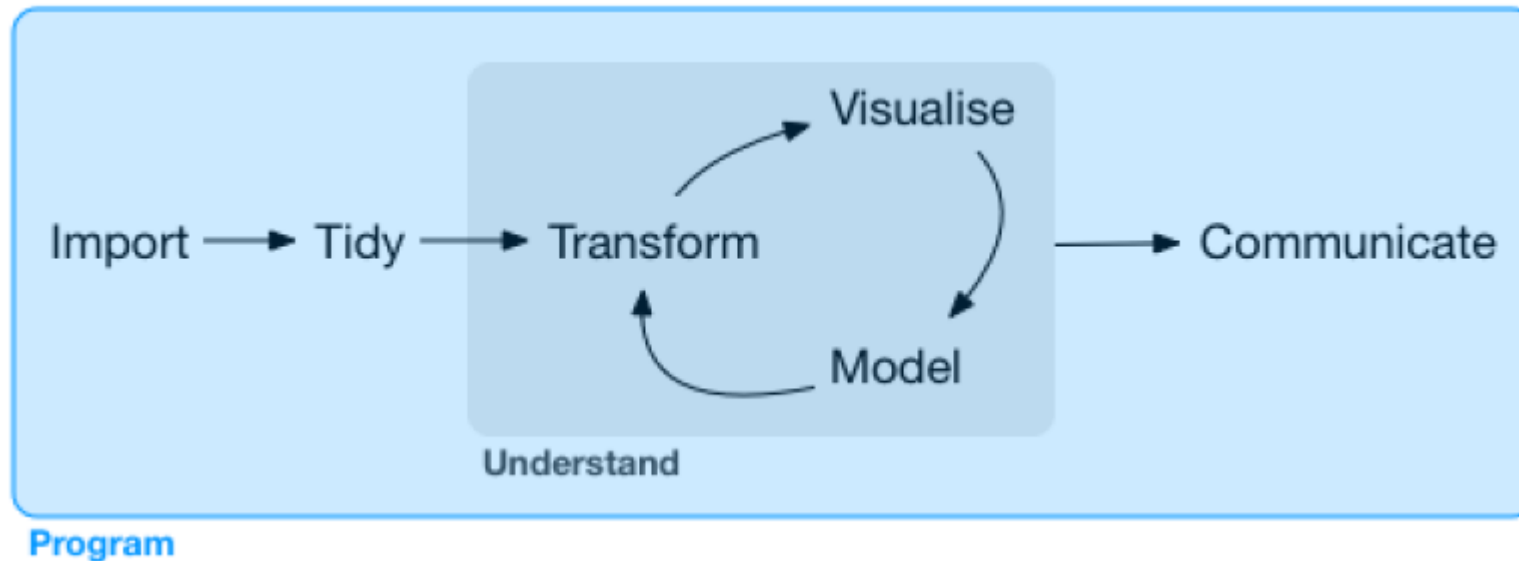


THE DATA SCIENCE WORKFLOW



Theme 1: Program.

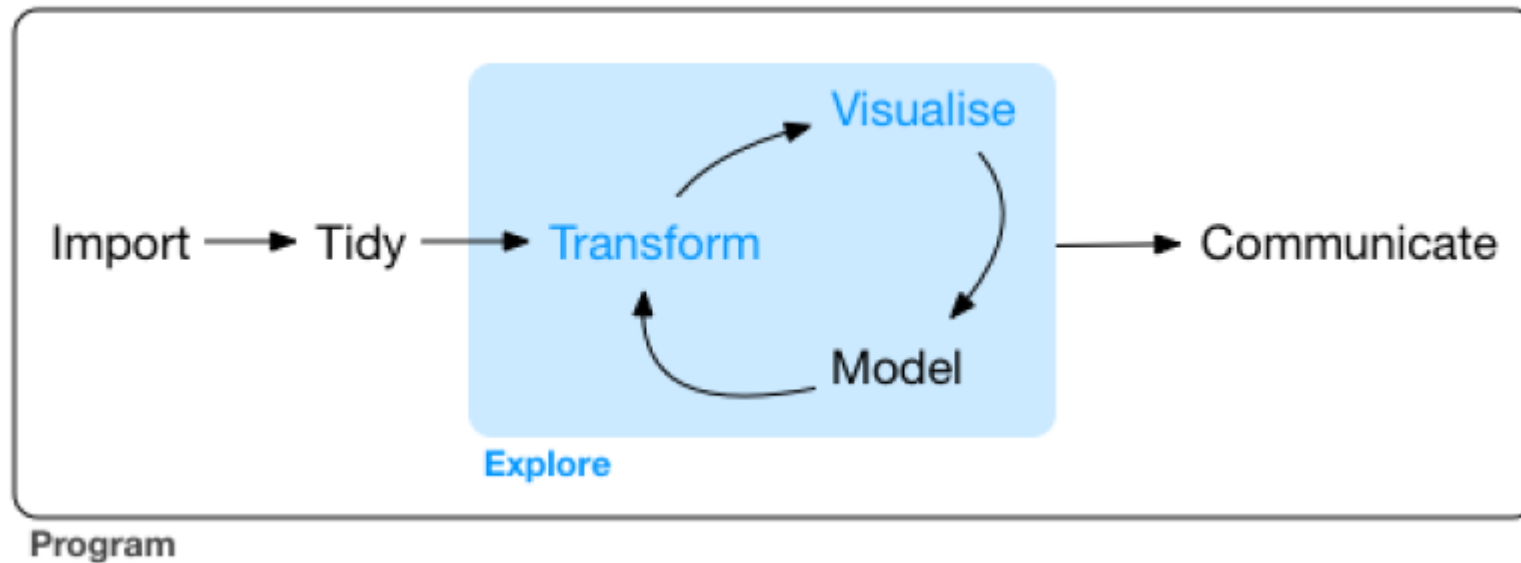
Fundamental programming skills



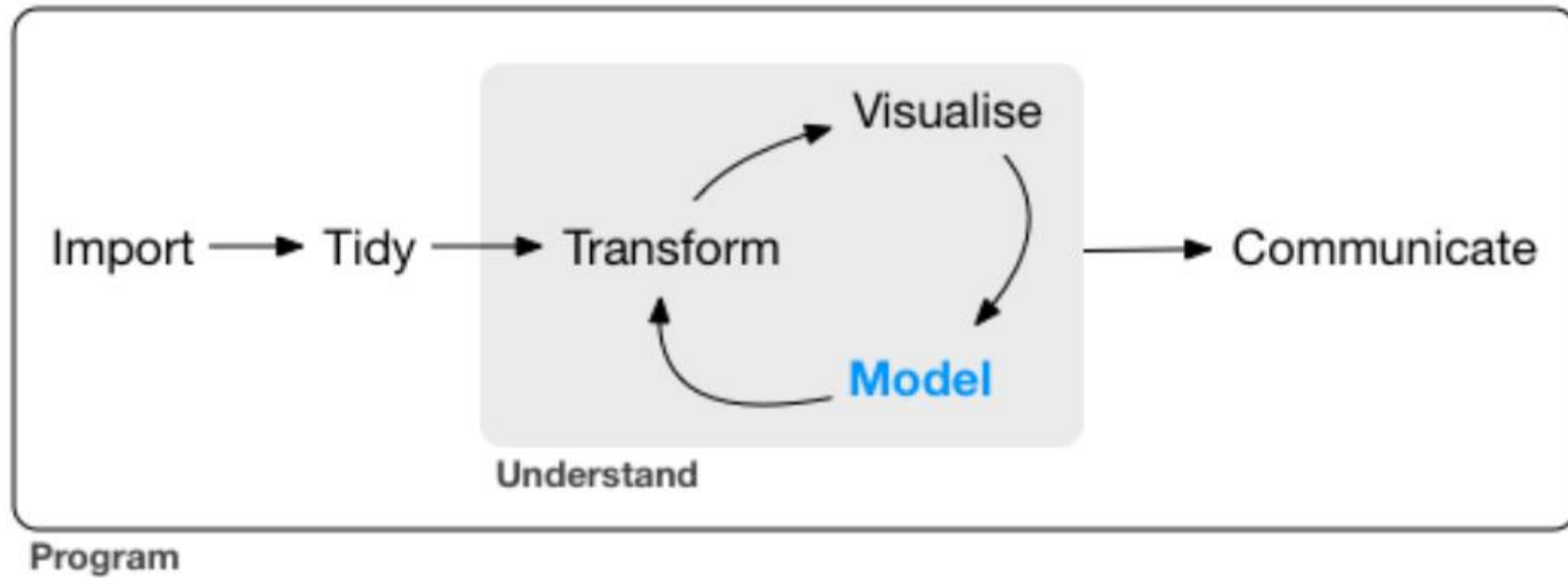
The Data Science Workflow

Theme 2: Explore.

Basic tools of data exploration.



The Data Science Workflow



Difference between Statistics and Machine Learning

Statistics

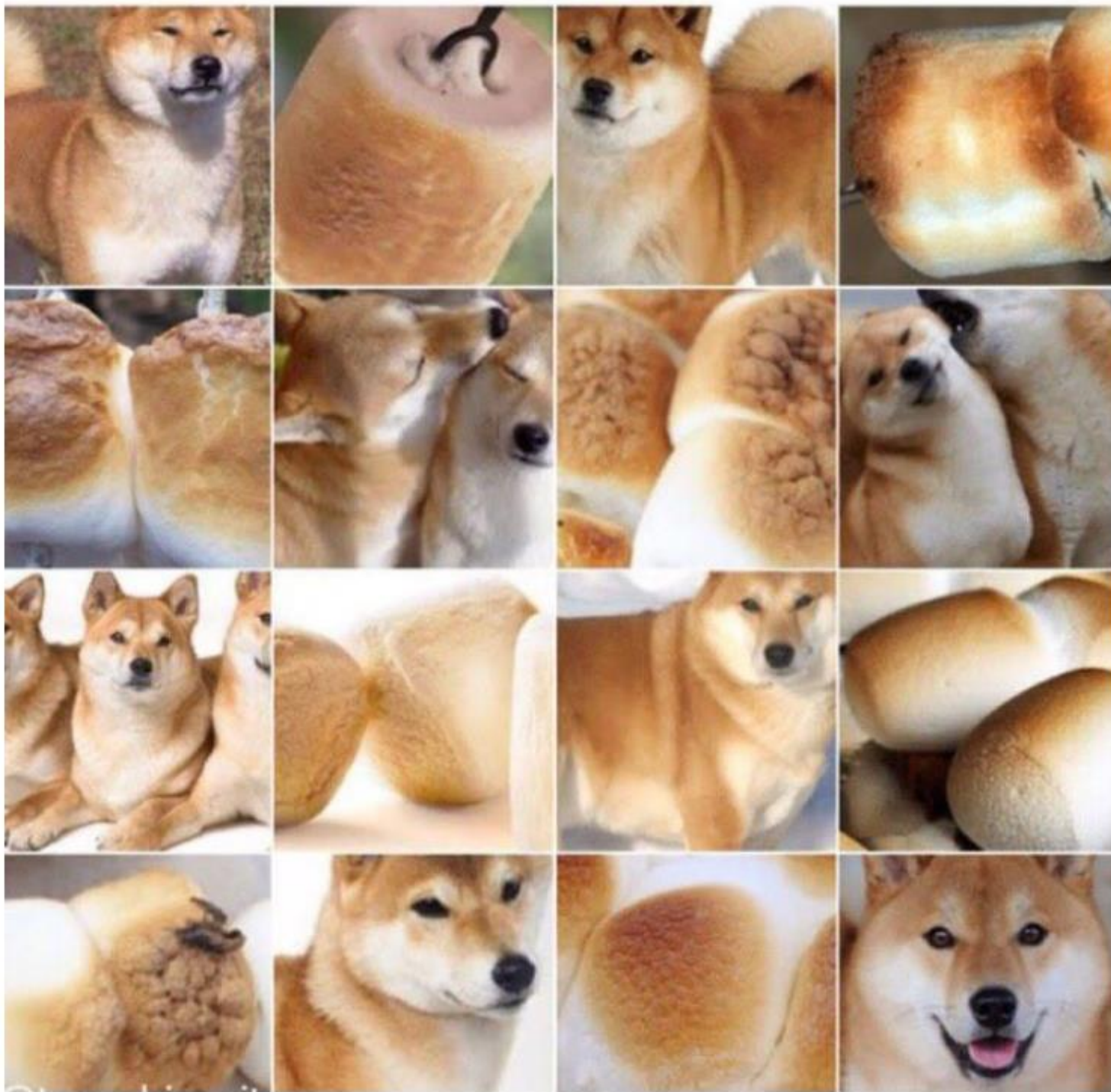
- Subfield of Mathematics
- Statistical model exists regardless of computers
- Inference
- Understanding of how data was generated

Machine Learning

- Subfield of Computer Science
- Train computers to learn by analysing data
- Optimisation and prediction
- Makes no underlying assumptions about data

Supervised Learning

- Labeled data
- Classification
 - Applications
 - Document classification and email spam filtering
 - Image classification and handwriting recognition
 - Face detection and recognition
- Regression
 - Applications
 - Predict stock market price
 - Predict age of a viewer watching video on YouTube
 - Predict temperature

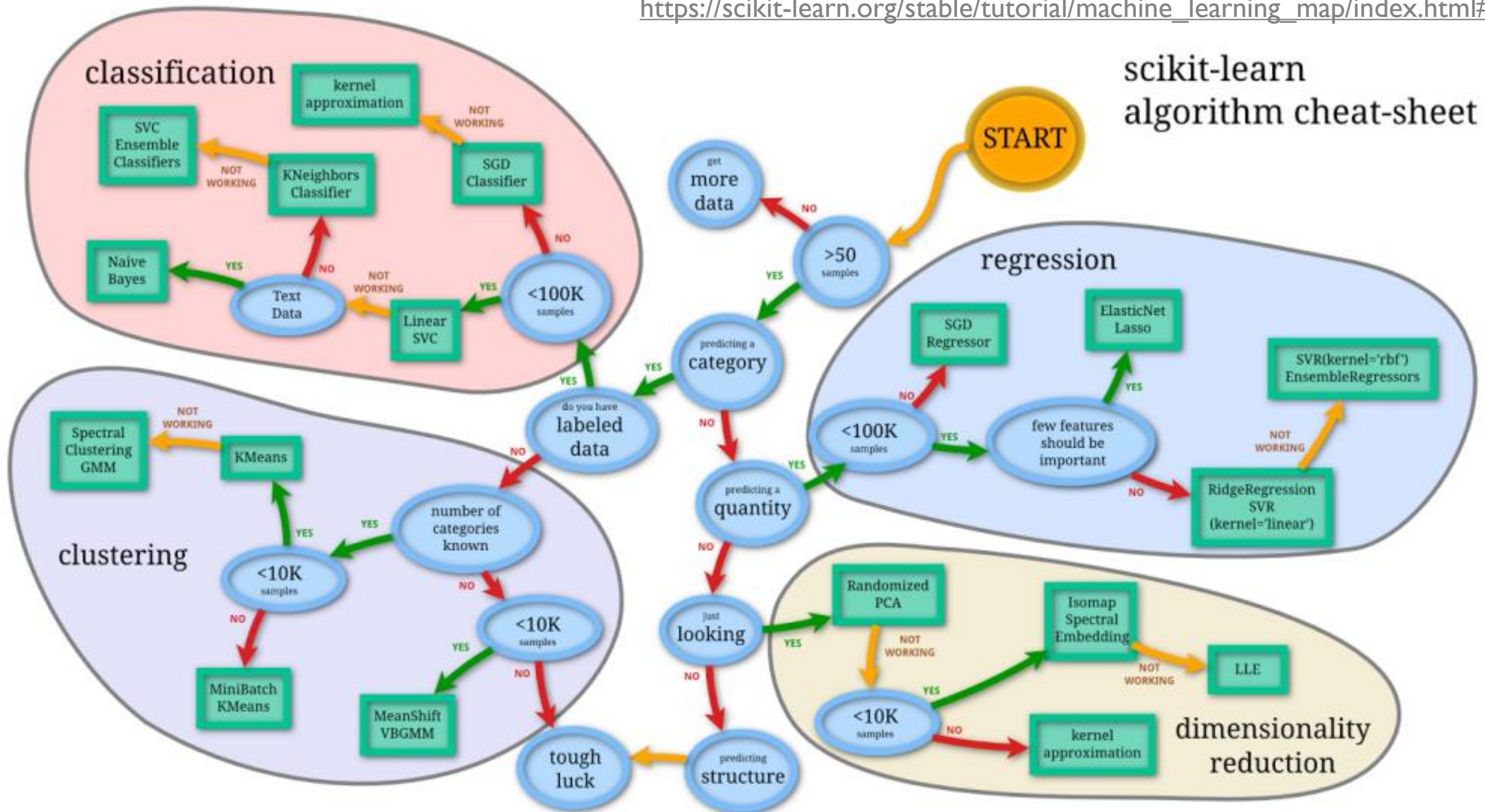




Unsupervised Learning

- Unlabeled data
- Discover clusters
 - Applications
 - Cluster users into groups, based on purchasing or web-surfing behavior.
 - Cluster clients into risk clusters, based on behavior and demographics.
- Discover latent factors
 - What drives the clusters?
 - Latent - can't observe the factors
- Discover graph structure

scikit-learn algorithm cheat-sheet





MATPLOTLIB



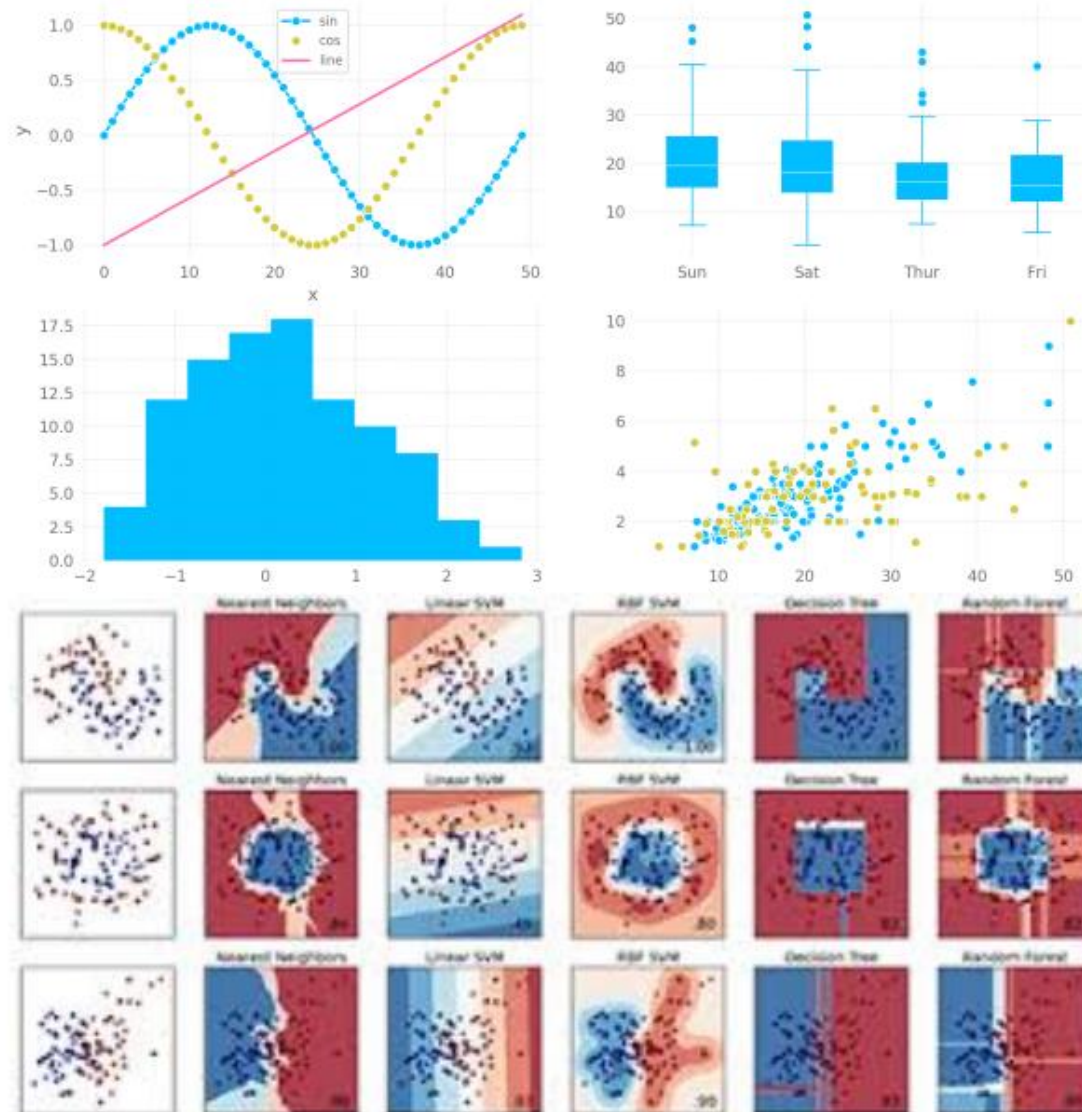
OUTCOMES

6.2.1. *Visualisation.* Outcomes: At the end of this study unit, the student should be able to:

- Implement the basic structure of a matplotlib plot
- Identify common plot types
- Create presentation quality statistical graphs, including appropriate use of scale, color, labels, reference markers

Visualisation

Matplotlib implements the grammar of graphics



Context

Context

Do cars with big engines use more fuel than cars with small engines?

Context

Do cars with big engines use more fuel than cars with small engines?

Question

What does the relationship between engine size and fuel efficiency look like? Is it positive? Negative? Linear? Nonlinear?

Dataset

The dataset is a text book dataset from the UCI Machine Learning Library (<http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/>) Two of the variables in the dataset:

- *displacement*, a car's engine size, in litres
- *mpg*, a car's fuel efficiency on the highway, in miles per gallon (mpg). A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance.

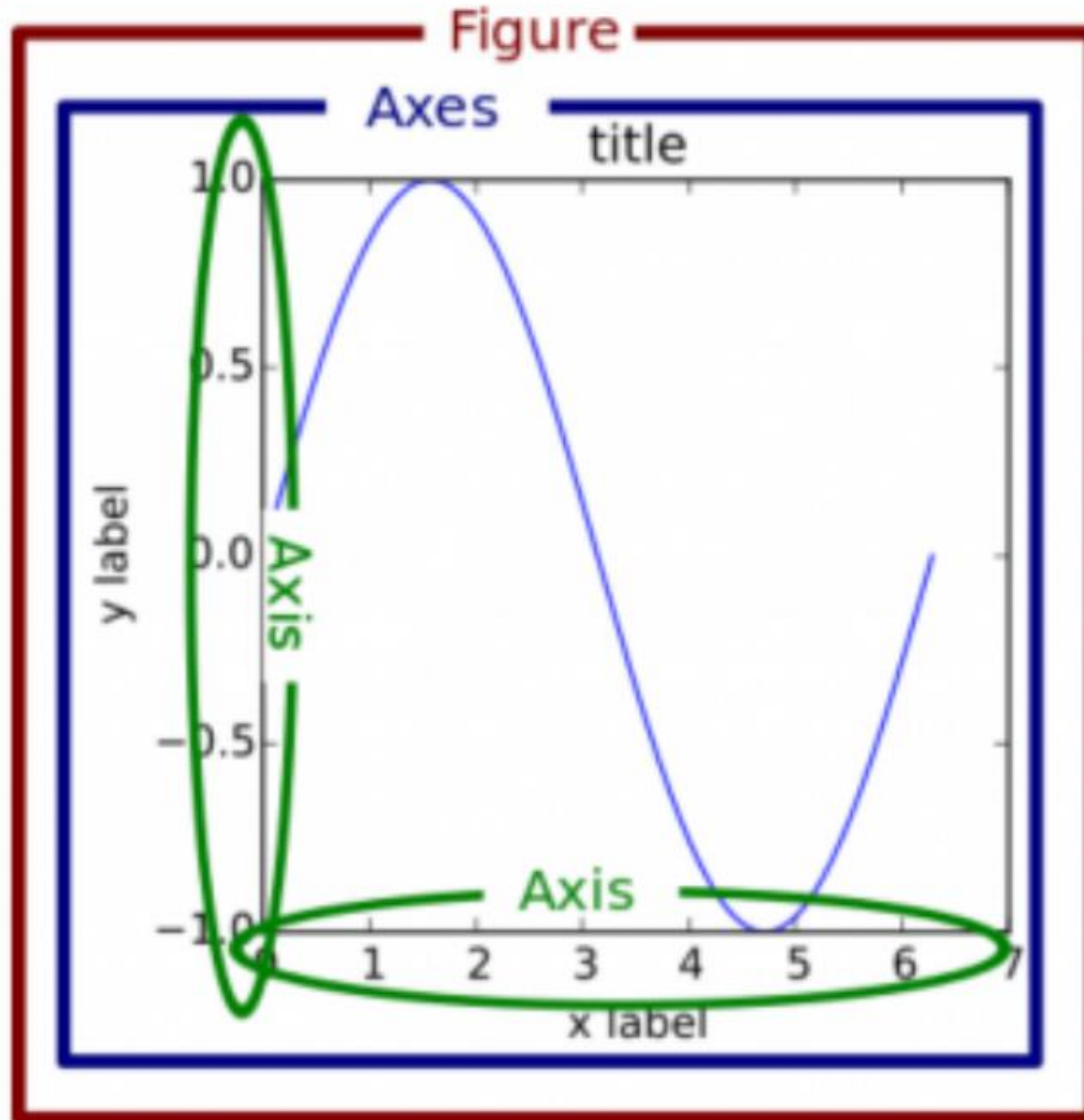
We will explore these two variables

Class example

Jupyter notebook:

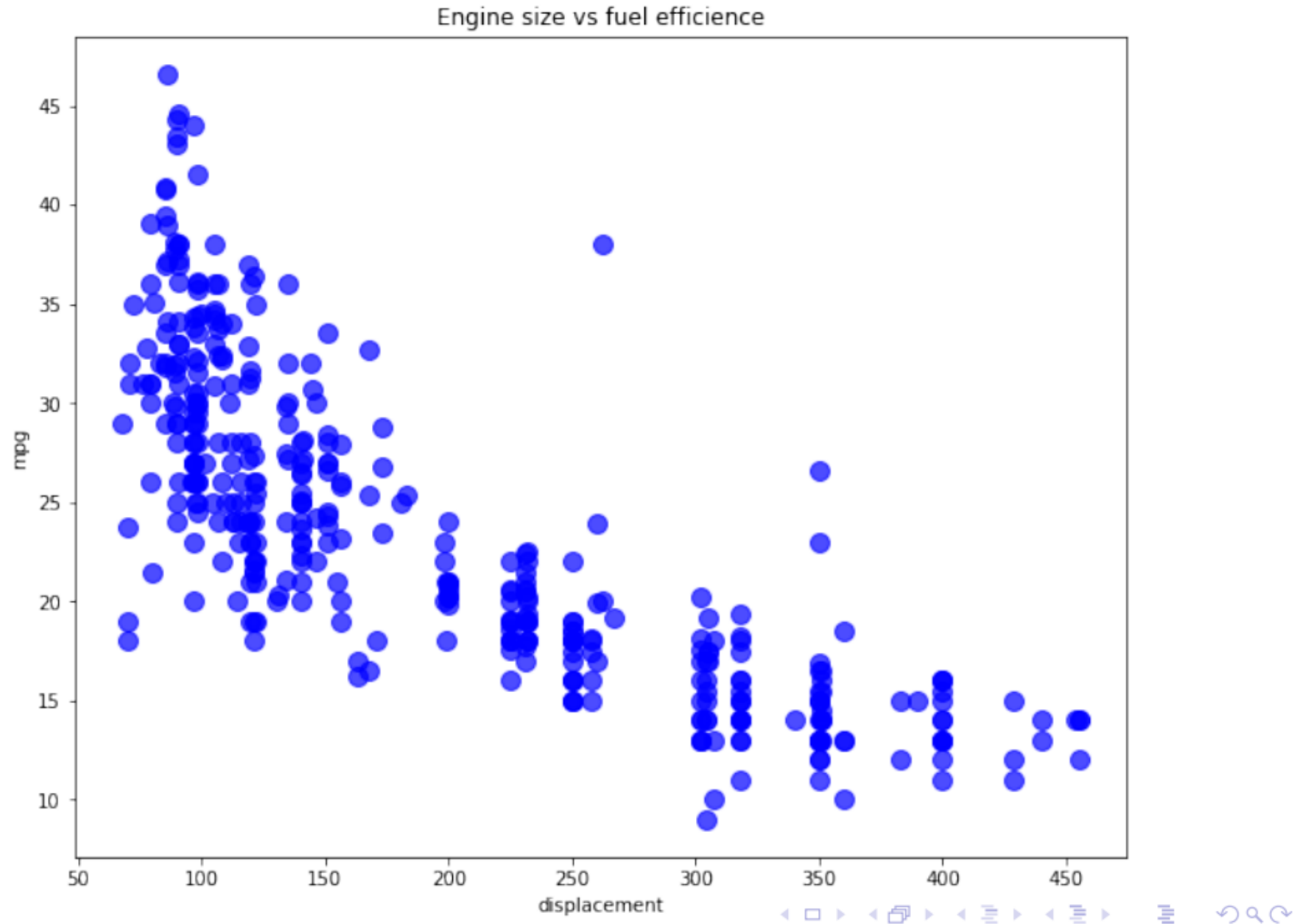
Explore Auto MPG Dataset

The anatomy of a pyplot



Objects: variables that contain data and functions (methods) that can be used to manipulate the data. eg. lists are an example of python objects.
``my_list.extend('5')``, the ``extend()`` is the method.

Explore the plot



Different types of plots

Continuous data:

- scatterplots
- histograms
- boxplots

Categorical data:

- pie charts
- categorical scatterplots
- bar graphs