# Introduction to Statistical Machine Learning

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

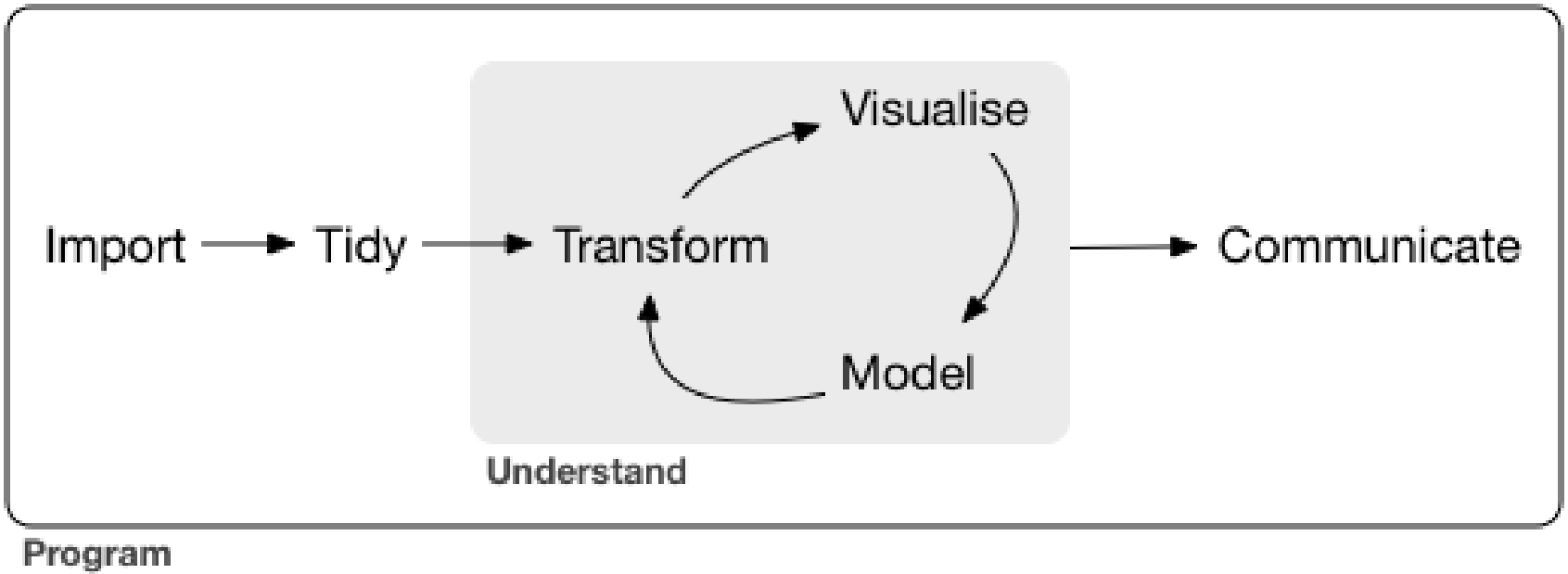
# Outline

- Unsupervised vs. supervised learning
- Unsupervised methods: hierarchical and k-means clustering
- Dimension reduction methods: principal components analysis
- Supervised methods: classification – logistic regression and k-nearest neighbours
- Regularisation methods: ridge, LASSO, and elastic net
- Smoothing methods: splines and generalized additive models
- Tree-based methods: classification/regression trees and random forests
- Extensions to tree-based methods: Bayesian additive regression trees (BART) and generalized additive models for location, scale and shape (GAMLSS)

# What is Statistical Learning?

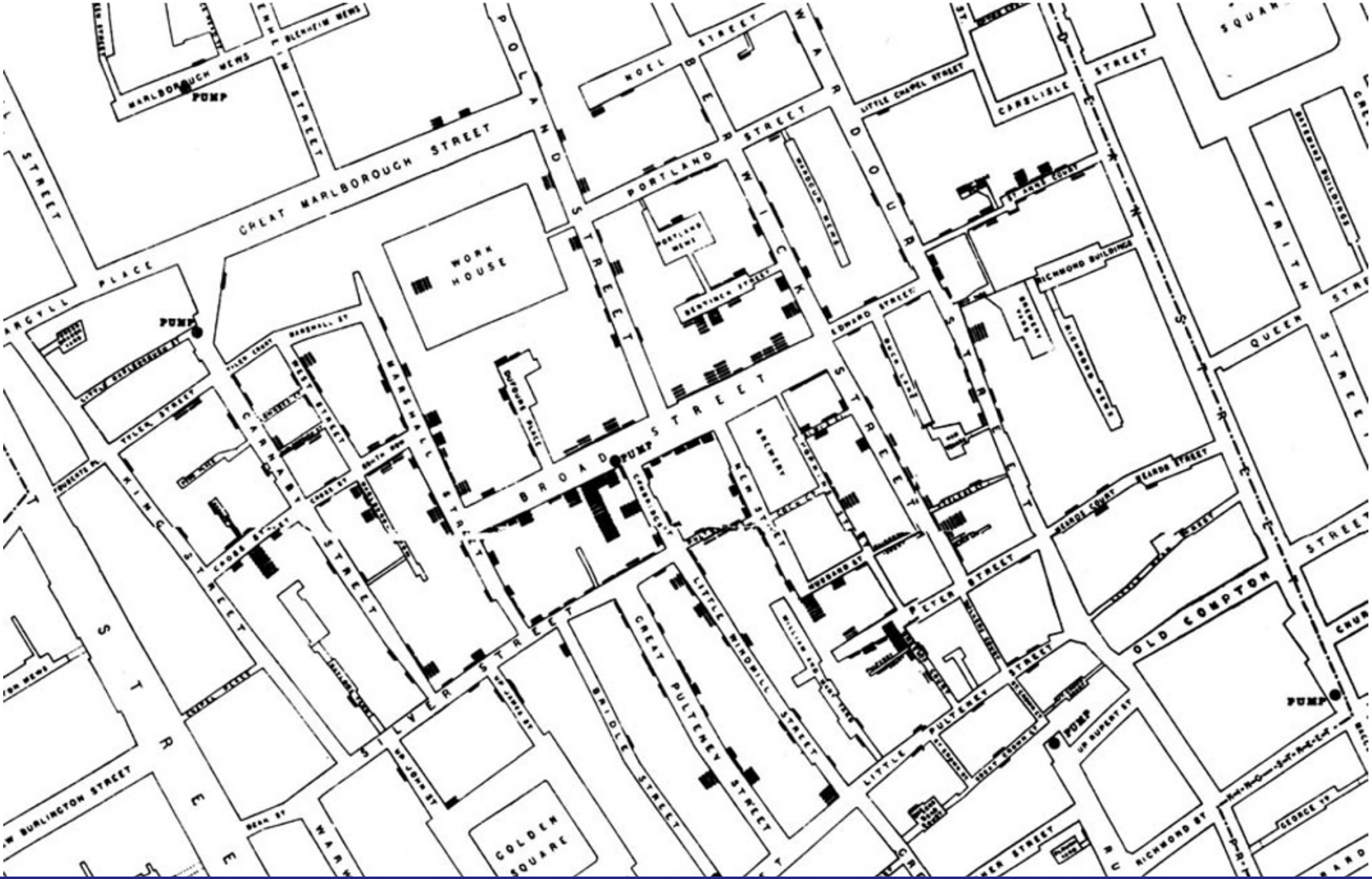
# Statistical Learning

- The data science process
- Prediction and inference



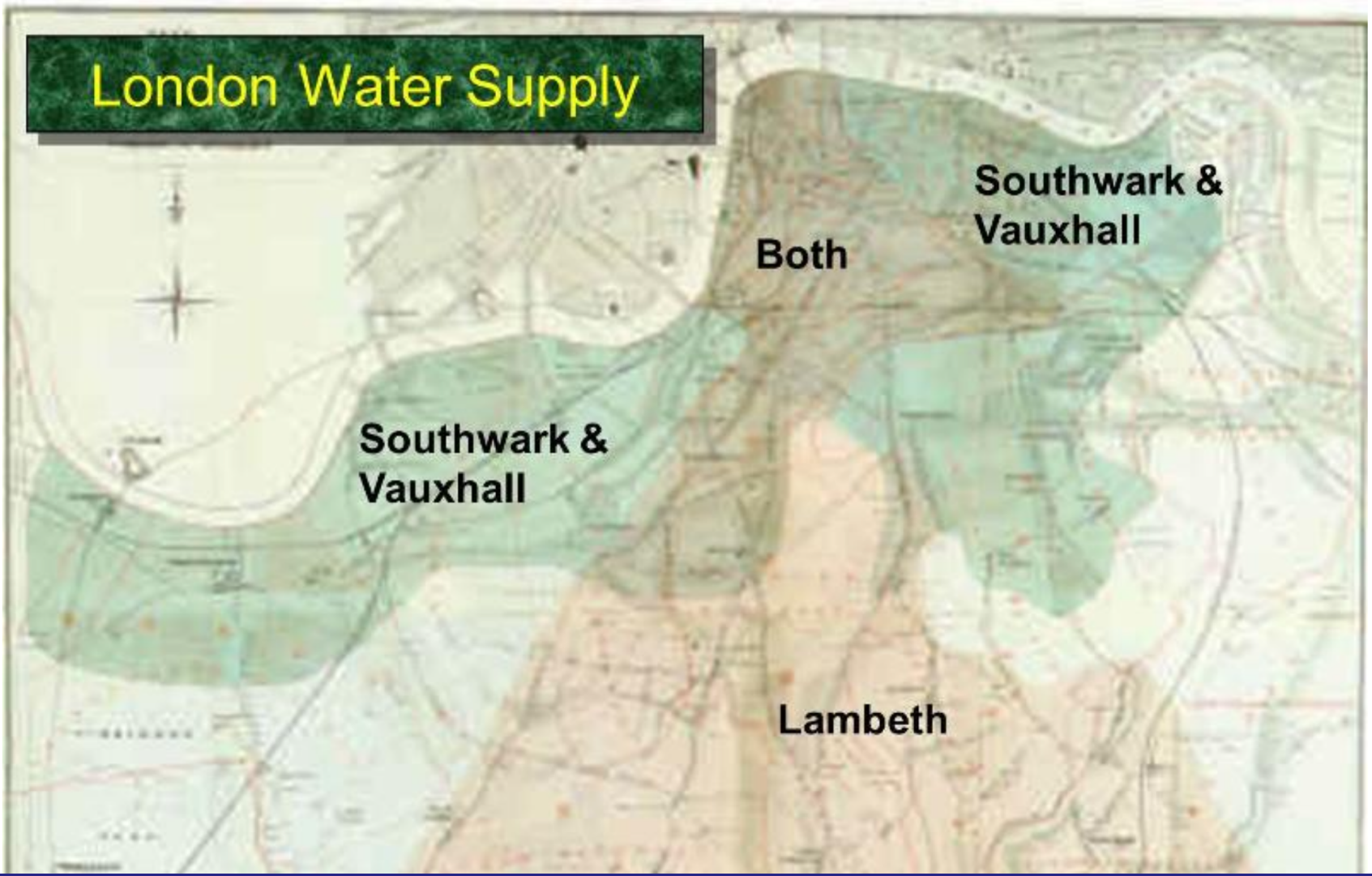

- Many times the right visualisation of the data can solve the problem

# Example: 1854 Broad St. Cholera Outbreak



# Example: 1854 Broad St. Cholera Outbreak

# Example: 1854 Broad St. Cholera Outbreak



# Example: 1854 Broad St. Cholera Outbreak

TABLE IX.

| | Number of houses. | Deaths from Cholera. | Deaths in each 10,000 houses. |
|---|---|---|---|
| Southwark and Vauxhall Company | 40,046 | 1,263 | 315 |
| Lambeth Company . . . . | 26,107 | 98 | 37 |
| Rest of London . . . . | 256,423 | 1,422 | 59 |

# Example 1: Wage Data

- The data comes from a survey of $n = 3000$ males wages in a region of the US.
- The response is wage, predictors are age, education level, and the year the wage was earned.
- Plots of wage versus the predictors show that wage increases with education and year, but for age wage increases for ages up to 60, but then decreases.
- Linear regression could be used here (probably with a transformation of age).
- But we will look at more advanced approaches.

# Example 2: Stock Market Data

- 1250 observations of daily percentage returns for S&P 500 stock index between 2001 and 2005.
- Here the response is categorical, whether the index goes up or down. Predictors are percentage returns on the stock market for the 5 previous days and volume of shares traded (number of daily shares traded in billions).
- This is an example of a **classification** problem.
- A model that could accurately predict the direction in which the market will move would be very useful!
- Linear regression is not suitable for problems with categorical responses, as assumptions are violated.
- Logistic regression could be used instead.

# Example 3: Gene Expression Data

- The NCI60 data set consists of 6,830 gene expression measurements for each of 64 cancer cell lines. There is no response or output variable.
- Interest is in finding groups or clusters among the 64 cell lines bases on the gene measurements.
- Techniques include principal components analysis and clustering algorithms such as $k$-means.
- Note also this is an example of wide data $n << p$, where there are $n = 64$ cases and over $p = 6{,}830$ variables. Typically data for regression problems has $n > p$.

# Statistical Learning

- "Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).
- Statistical learning is machine learning with its roots in statistical concepts and probability.
- Data mining is a related field: here the focus is on finding interesting nuggets in the data.
- Statistical machine learning focuses on prediction.
- Problems such as the stock market and wage data examples are often termed *supervised learning* problems, so-called because you know the response output for the given data.
- Problems such as the gene expression data are often termed *unsupervised learning* problems, because there is no response, i.e. you do not know the groups or classes for the given data.