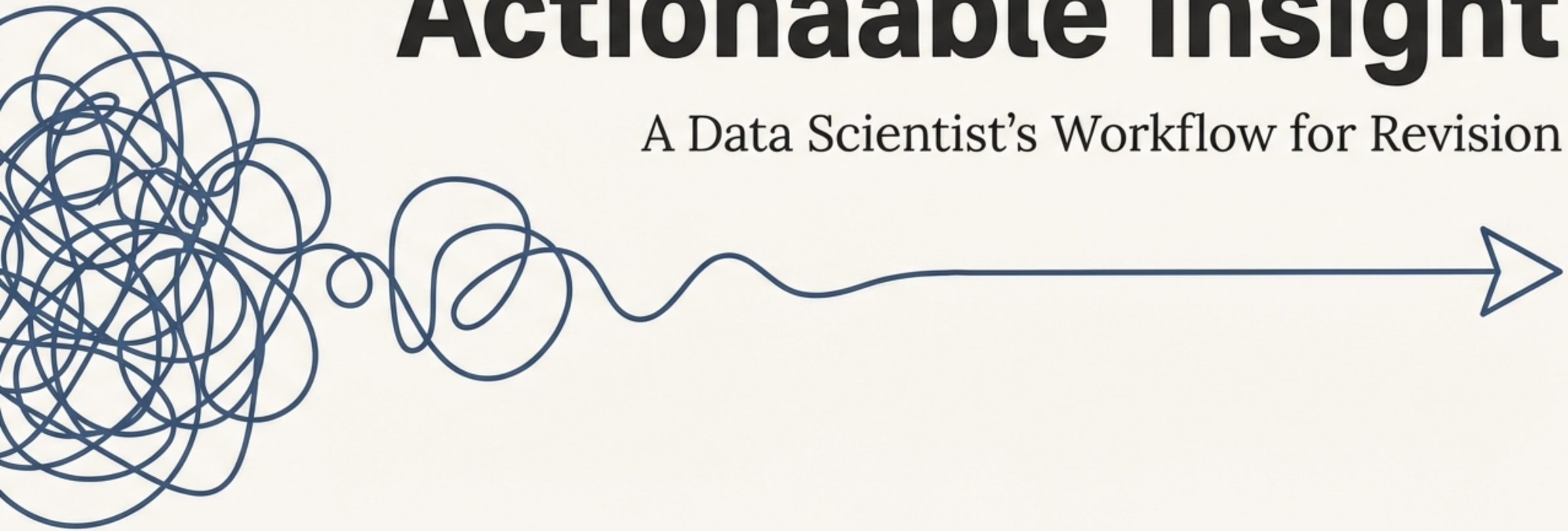
From Raw Data to Actionaable Insight

A Data Scientist's Workflow for Revision



First, What is Data Science?

Definition & Benefits

What it is
The interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

Why it matters (Benefits)
It empowers data-driven decision-making and predictions, leading to strategic advantages in areas like finance, healthcare, and retail.

How it Compares

Business Intelligence (BI)	Data Science (DS)	Data Mining
Focuses on descriptive analytics (what happened). Uses historical data. Primarily structured data.	Focuses on predictive and prescriptive analytics (what will happen and why). Uses historical and current data to forecast future outcomes. Handles structured and unstructured data.	A specific technique *within* Data Science focused on discovering patterns in large datasets. It's a step in the DS process, not the whole process itself.

Key Takeaway: Data Science is forward-looking; it builds on BI and utilises Data Mining to predict the future.

Step 1: Sourcing and Understanding the Raw Material

The Universe of Data

Key Sources

- Web Scraping (advantageous for gathering unique data)
- Streaming Data (e.g., social media feeds, IoT sensors)
- Databases
- APIs
- Public Datasets

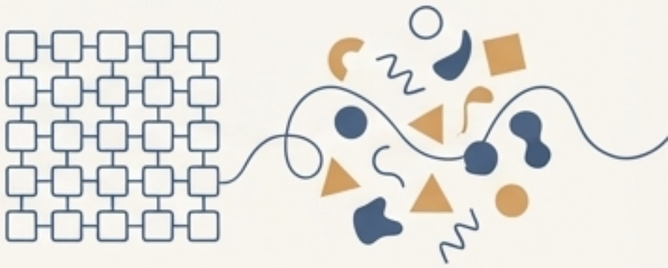
Key Formats

- CSV (comma-separated values)
- **JSON** (JavaScript Object Notation)
- XML (eXtensible Markup Language)
- Text files

Defining Today's Data Landscape

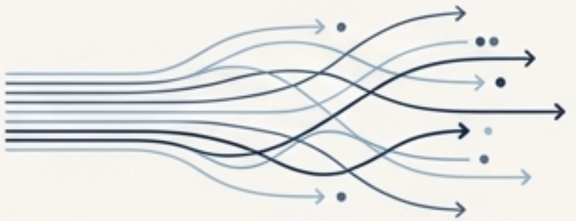
Unstructured Data: Data that doesn't have a pre-defined model or isn't organised in a pre-defined manner.

Example: The text of an email, a photograph, or a social media post.



Streaming Data: Data that is generated continuously by thousands of data sources, typically sent in small sizes.

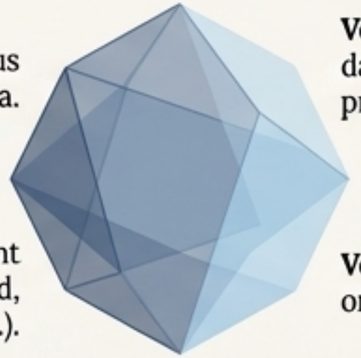
Example: Financial stock tickers, real-time user activity on a website.



Big Data: Defined by its main characteristics (the 'Vs').

Volume: Enormous scale of data.

Velocity: High speed of data generation and processing.



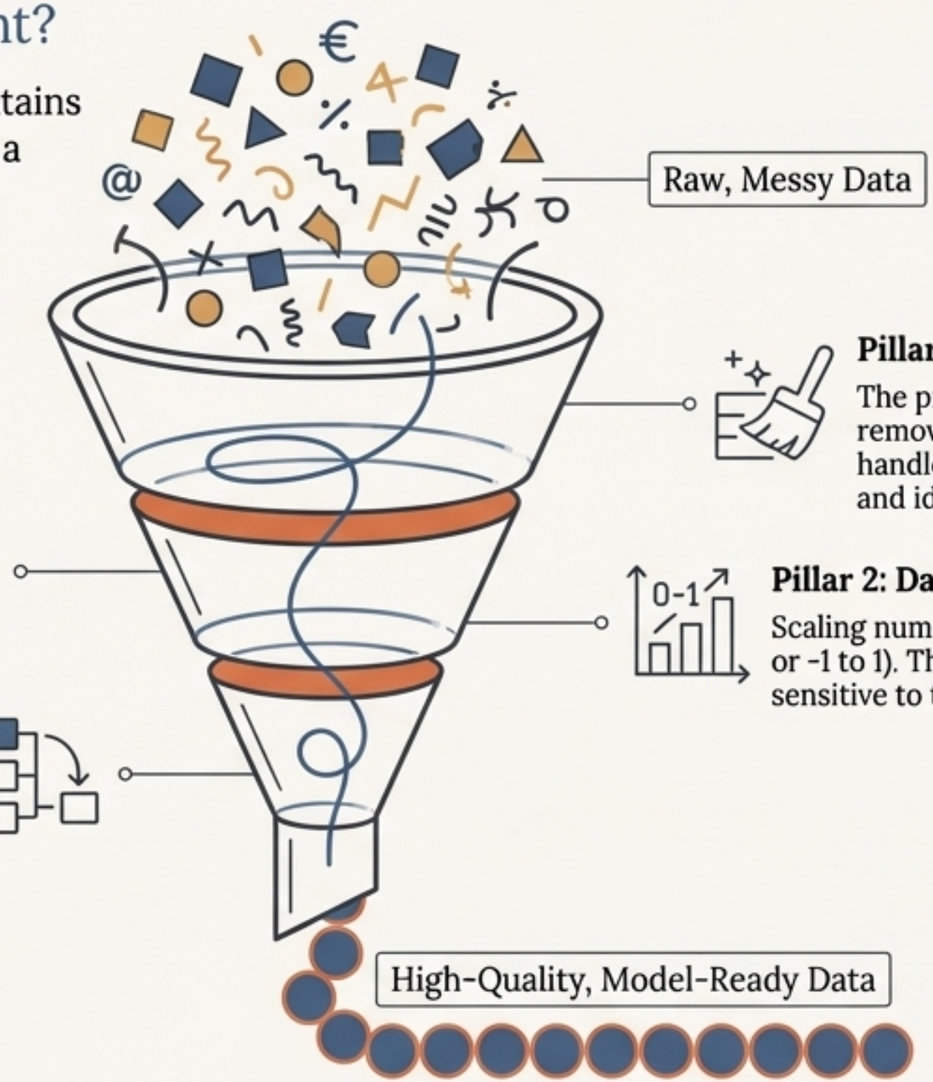
Variety: The different forms of data (structured, unstructured, etc.).

Veracity: The uncertainty or quality of the data.

Step 2: Forging the Data for Quality

Why is Data Preprocessing Important?

Raw data is often incomplete, inconsistent, and contains errors. Preprocessing transforms this raw data into a clean, understandable format, ensuring the quality **and reliability** of any resulting insights or models. The guiding principle is “Garbage in, garbage out.”



Pillar 1: Ne*ɒss* Data

The process of detecting and correcting (and removing) corrupt or inaccurate records. This handles missing values, smooths ѕtamay noisy data, and identifies) outliers.

Pillar 1: Data Cleaning

The process of detecting and correcting (or removing) corrupt or inaccurate records. This handles missing values, smooths noisy data, and identifies outliers.

Pillar 2: Data Transformation (Normalisation)

Scaling numerical data to a common range (e.g., 0-1 or -1 to 1). This is crucial for algorithms that are sensitive to the magnitude of features.

Pillar 3: Data Reduction

Reducing the volume of data while producing the same or similar analytical results. This can involve reducing the number of records (sampling) or features (dimensionality reduction).

A Deeper Look: Investigating Anomalies with EDA

Distinguishing Key Imperfections

Noise

A random error or variance in a measured variable. It's meaningless data.

Outlier

A data point that differs significantly from other observations. It can be a valid but rare event, or an error.

The Investigator's Toolkit: Exploratory Data Analysis (EDA)

What it is

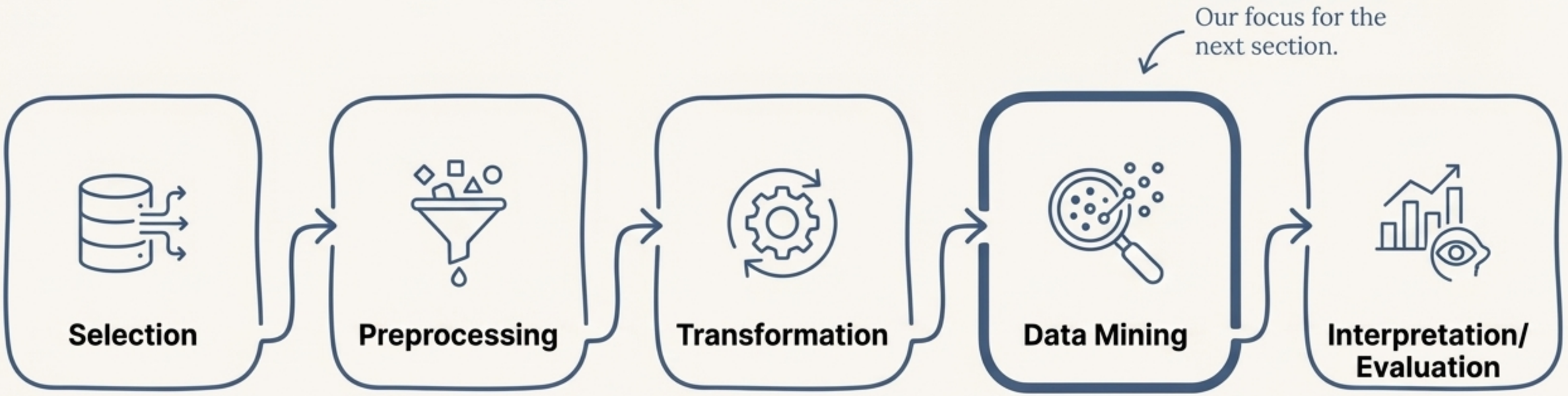
An approach to analysing data sets to summarise their main characteristics, often with visual methods. It's what a data scientist does to become familiar with the data.

Its Role in the Workflow

- To spot anomalies like outliers and errors.
- To understand the underlying structure of the data.
- To identify important variables and relationships between them.
- To inform feature selection and model choice in later stages.

Step 3: Mapping the Path to Discovery

The Knowledge Discovery in Databases (KDD) Process

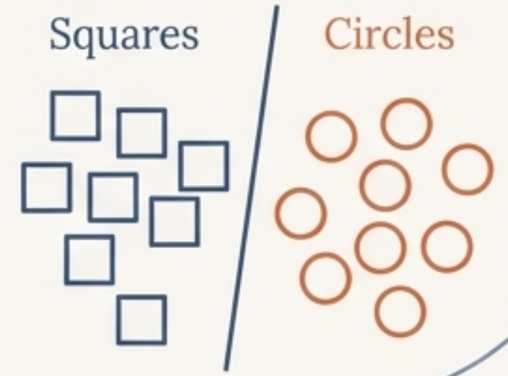


Step 4A: Uncovering Hidden Patterns in the Data

The Two Primary Goals of Unsupervised Learning

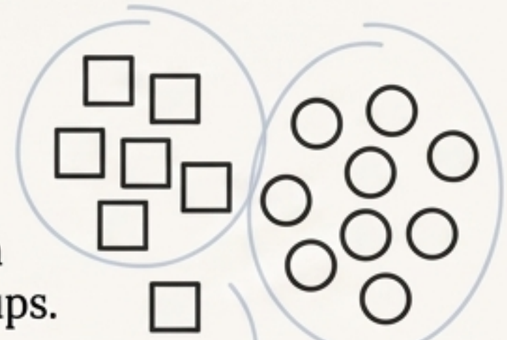
Classification (Supervised)

The task of assigning items to predefined categories. The algorithm learns from labelled data.



Clustering (Unsupervised)

The task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups. The algorithm finds the labels.

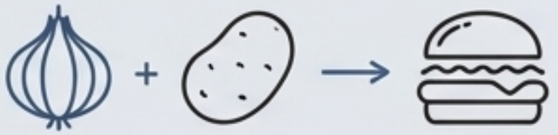


A Key Technique: Association Rule Mining

Concept: Discovering interesting relationships hidden in large datasets.

Real-World Application: Market Basket Analysis: Identifying which products customers are likely to buy together.

***Example:** If a customer buys Onions and Potatoes, they are 80% likely to also buy a Burger.*



This insight is used for store layout, promotions, and recommendations.

Association Rule Mining: The Algorithms

The Apriori Algorithm

How it works: Uses a "bottom-up" approach. It first identifies frequent individual items and extends them to larger itemsets, pruning candidates that have an infrequent subset.

Key Steps

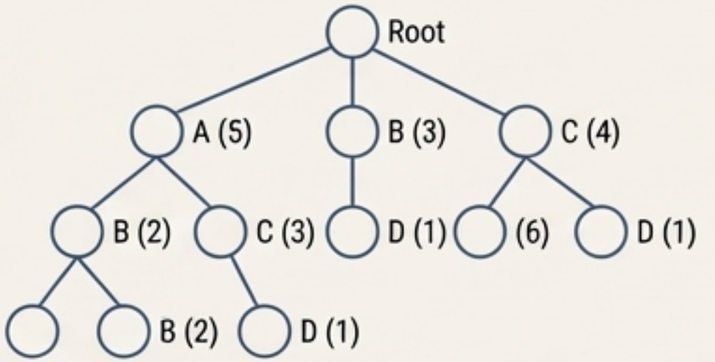
1. Set a minimum support threshold.
2. Generate a list of frequent itemsets of length 1 (C1 → L1).
3. Generate candidate itemsets of length k from frequent itemsets of length k-1.
4. Prune candidates that have an infrequent subset (the Apriori principle).
5. Repeat until no more frequent k-itemsets can be found.

The FP-Growth Algorithm

How it works: Compresses the database into a Frequent-Pattern Tree (FP-Tree) structure. It avoids the costly candidate generation step of Apriori.

Why it's more efficient

- It scans the database only twice.
- It avoids generating a large number of candidate sets.
- This makes it significantly faster for large datasets.

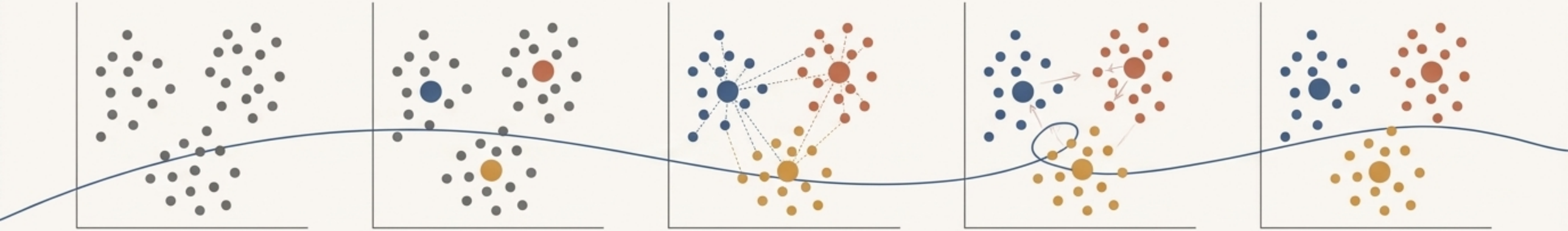


Note: For revision, practice numerical examples involving the calculation of frequent itemsets and association rules using these steps with given minimum support and confidence values.

Clustering in Practice: The K-Means Algorithm

****Core Concept*:** An iterative algorithm that partitions a dataset into a pre-defined number (K) of distinct, non-overlapping clusters.

How It Works: A Visual Step-by-Step Guide



Step 1: Choose K

The user specifies the number of clusters to find (e.g., K=3).

Step 2: Initialise Centroids

Randomly select K points as initial cluster centres (centroids).

Step 3: Assign Points

Assign each data point to the nearest centroid, forming K initial clusters.

Step 4: Update Centroids

Recalculate the centre of each cluster by taking the mean of all its points.

Step 5: Repeat & Converge

Repeat Steps 3 and 4 until the centroids no longer move significantly.

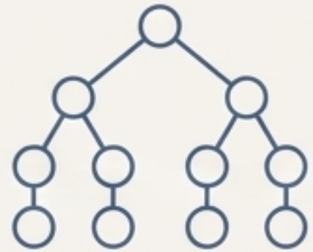
Step 4B: Making Predictions with Supervised Learning

Core Concept: Machine Learning (ML)

The study of computer algorithms that can improve automatically through experience and by the use of data. It is a subset of AI.

- Supervised (labelled data), Unsupervised (unlabelled data), Reinforcement (learning through rewards/penalties). This section focuses on Supervised.

Key Classification Algorithms



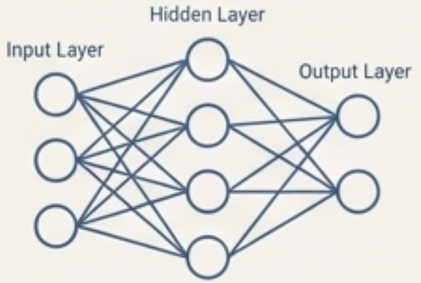
Decision Tree

A flowchart-like structure where each internal node represents a 'test' on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. It's highly interpretable.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Naïve Bayes

A probabilistic classifier based on applying Bayes' theorem with a 'naïve' assumption of conditional independence between features. It's fast, simple, and works well for text classification.



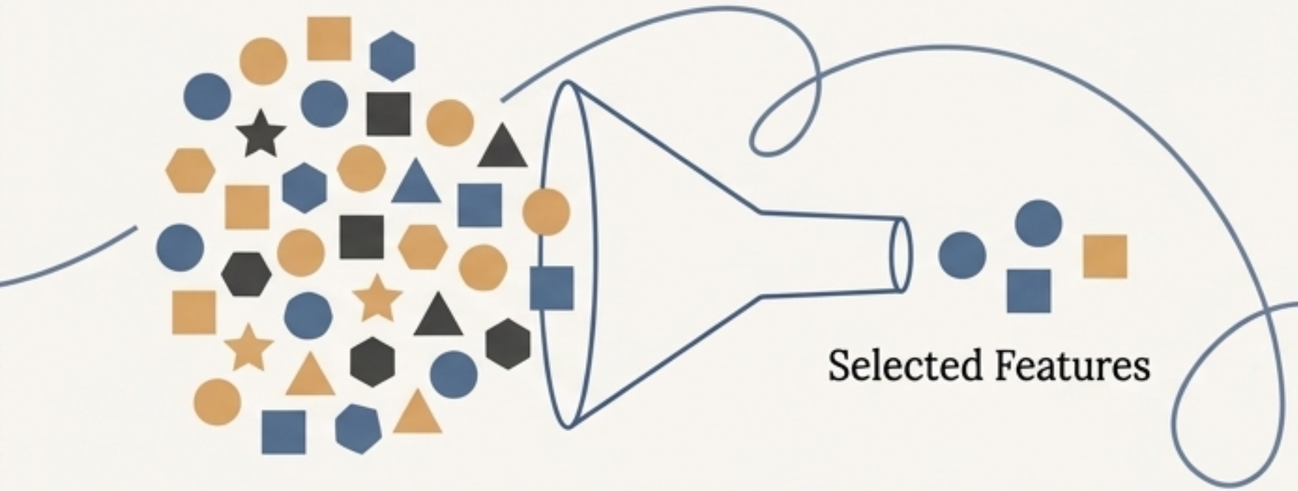
Neural Network

A computational model inspired by the structure and function of biological neural networks. It consists of interconnected layers of nodes (neurons) that process information.

The Art of Feature Engineering and Model Fitting

Section 1: Feature Selection

The process of selecting a subset of relevant features (variables, predictors) for use in model construction.



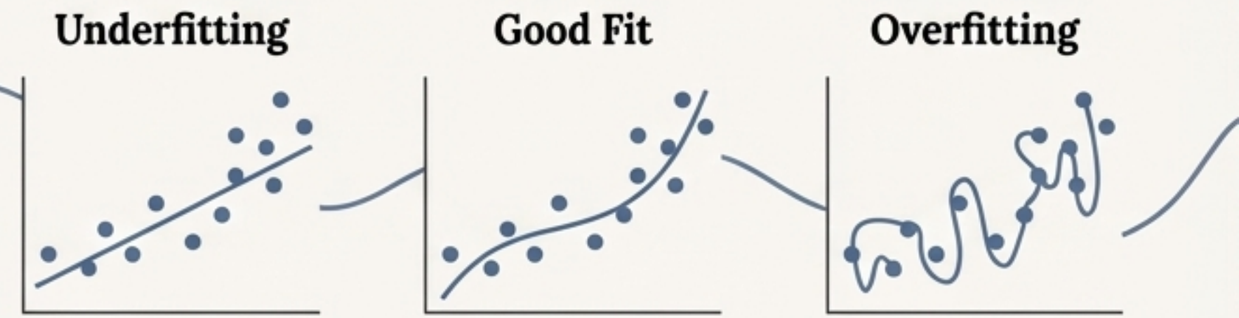
- Simplifies models, making them easier to interpret.
- Reduces training times.
- Mitigates the 'curse of dimensionality.'
- Improves model accuracy by removing irrelevant or redundant features.

Briefly, methods include Filter methods, Wrapper methods, and Embedded methods.

Section 2: Model Fitting

The process of training a machine learning model on a dataset (the 'training set') to learn the relationships between features and the target outcome.

Goal: The model should generalise well to new, unseen data. This involves finding the right balance between underfitting (too simple) and overfitting (too complex).



A straight line failing to capture the pattern in a set of curved data points.

A smooth curve that accurately represents the trend of the data points.

A frantic, squiggly line that passes through every single data point, including noise.

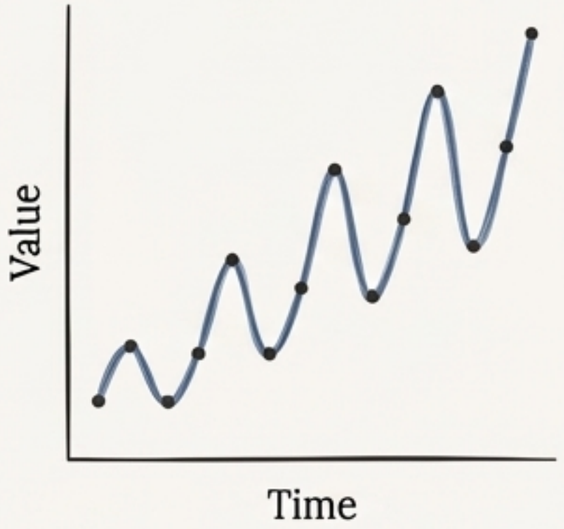
Statistical Inference: We use a **sample** (the training data) to infer properties about an entire **population** (all possible data).

Step 5: Advanced Frontiers - Analysing Sequential Data

Application 1: Time Series Analysis

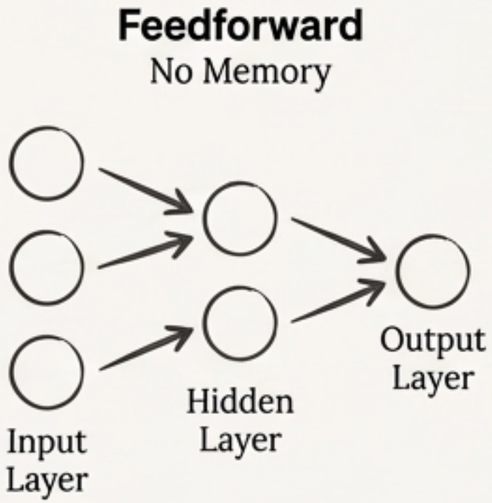
Definition: A series of data points indexed in time order. It involves analysing statistical characteristics of the data to extract meaningful statistics and other characteristics.

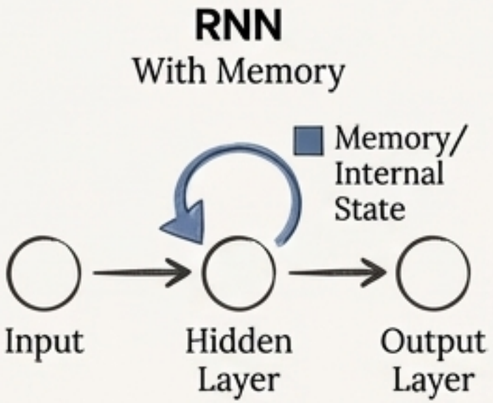
Examples: Stock prices over time, daily weather records, monthly sales figures.



Application 2: Recurrent Neural Networks (RNNs)

What it is: A class of neural networks designed to work with sequence data. They have 'memory' because information can persist through loops in the network.





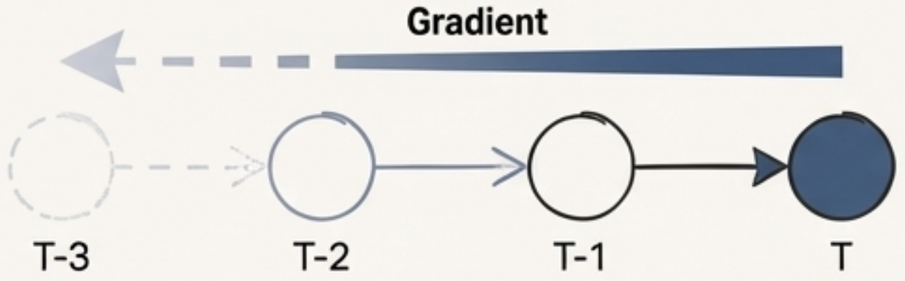
Connections between nodes form a directed cycle, allowing it to exhibit temporal dynamic behaviour. It can use its internal state (memory) to process sequences of inputs.

Overcoming the Limits of Memory: The LSTM Solution

The Challenge: The Vanishing Gradient Problem

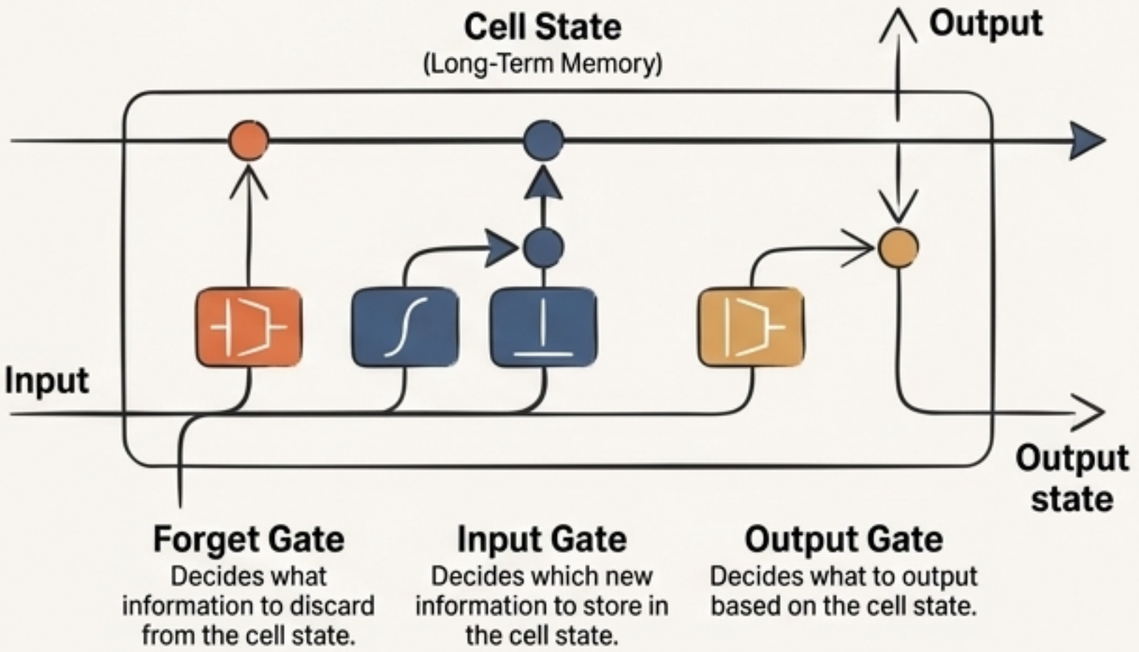
What it is: In deep networks like RNNs, as the gradient is back-propagated over many time steps, it can shrink exponentially and "vanish".

The Consequence: The network struggles to learn and tune the parameters of earlier layers, preventing it from capturing long-range dependencies in the data.



The Solution: Long Short-Term Memory (LSTM) Networks

How LSTMs address the problem: LSTMs are a special kind of RNN that use a "gating" mechanism to regulate the flow of information. These gates can learn which information is important to keep or throw away, allowing the network to maintain a constant gradient over long sequences.



Advanced Frontiers: Deriving Meaning from Vision and Text

Domain 1: Computer Vision

Definition: A field of AI that trains computers to interpret and understand the visual world.

Real-World Applications:

- Self-driving cars (object detection)
- Medical imaging analysis
- Facial recognition

Domain 2: Natural Language Processing (NLP)

Definition: A field of AI that gives computers the ability to read, understand, and derive meaning from human language.

The NLP Toolkit - Key Techniques

- **Tokenization:** The process of breaking down a stream of text into words, phrases, symbols, or other meaningful elements called tokens. (*Types*: word, sentence).
- **Text Lemmatization:** The process of reducing the different inflected forms of a word to a single, canonical form (the "lemma"). *Example*: '*studies*,' '*studying*' → '*study*."
- **Sentiment Analysis:** The use of NLP to systematically identify, extract, quantify, and study affective states and subjective information. *Example: Classifying a product review as 'positive,' 'negative,' or 'neutral.'*

The Payoff: From Insight to Impact

Bringing it all together: The data science workflow is a systematic process for making decisions and predictions. By sourcing, cleaning, and modelling data, we move from raw information to structured knowledge that can guide strategy.

How Decisions are Made

Descriptive & Diagnostic: EDA and unsupervised learning help us understand *what happened* and *why*.

Predictive & Prescriptive: Supervised learning and advanced models help us forecast *what will happen* and determine the *best course of action*.

A Final Challenge

Can Data Science Predict the Stock Market?

While it can identify trends, correlations, and anomalies using Time Series Analysis and Sentiment Analysis on news, the market is influenced by countless unpredictable, often irrational, human factors. It remains one of the most complex challenges, highlighting both the power and the limitations of the field.