

Properties of Data

Module 1

1. Structured vs. Unstructured Data

- **Structured Data:**

- Well-organized (tables in databases/spreadsheets).
- Represented as matrices (rows for items, columns for properties).

- **Unstructured Data:**

- Heterogeneous nature.
- Initial structuring using models like bag-of-words.



2. Quantitative vs. Categorical Data

- **Quantitative Data:**
 - Numerical values (e.g., height, weight).
 - Direct application in mathematical models.
- **Categorical Data:**
 - Descriptive labels (e.g., gender, hair color).
 - Coding numerically; challenges in interpretation.

3. Big Data vs. Little Data

- **Big Data:**

- Analysis of massive datasets.
- Challenges: slower cycle time, visualization complexity.

- **Little Data:**

- Emphasis on relevance over volume.
- Simple models for effective decision-making.





Asking Interesting Questions from Data

- Good data scientists develop an inherent curiosity about the world around them
- Software developers are not really encouraged to ask questions, but data scientists are.
- What things might you be able to learn from a given data set?
- What do you/your people really want to know about the world?
- What will it mean to you once you find out?

The Baseball Encyclopaedia

Statistical information on the performance of Babe Ruth can be found at <http://www.baseball-reference.com>.

How can we best measure an individual player's skill or value?

How fairly do trades between teams generally work out?

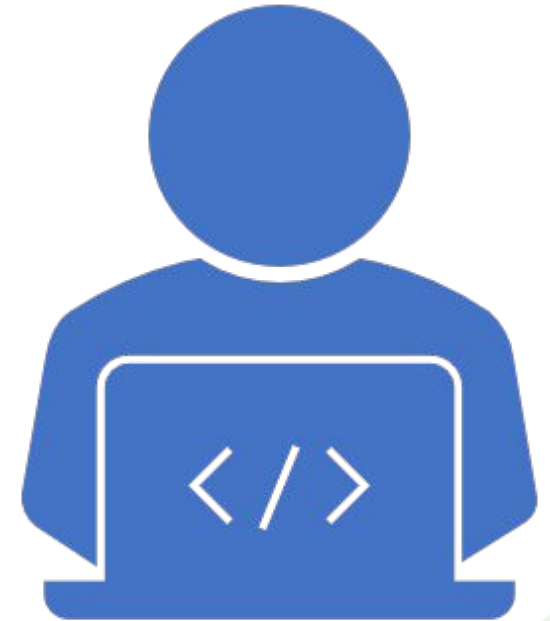
What is the general trajectory of player's performance level as they mature and age?

To what extent does batting performance correlate with position played? For example, are outfielders really better hitters than infielders?

- Do left-handed people have shorter lifespans than right-handers?
 - Handedness is not captured in most demographic data sets, but has been diligently assembled here. Indeed, analysis of this data set has been used to show that right-handed people live longer than lefties.
- How often do people return to live in the same place where they were born?
 - Locations of birth and death have been extensively recorded in this data set. Further, almost all of these people played at least part of their career far from home, thus exposing them to the wider world at a critical time in their youth.
- Do player salaries generally reflect past, present, or future performance?
- To what extent have heights and weights been increasing in the population at large?

The Internet Movie Database (IMDb)

- The Internet Movie Database (IMDb) provides crowdsourced and curated data about all aspects of the motion picture industry, at www.imdb.com.
- What are your questions?
- What a Data Scientist might ask?
- How do Hollywood movies compare to Bollywood movies, in terms of ratings, budget, and gross?
- Are American movies better received than foreign films, and how does this differ between U.S. and non-U.S. reviewers?



- What is the age distribution of actors and actresses in films?
- How much younger is the actress playing the wife, on average, than the actor playing the husband?
- Has this disparity been increasing or decreasing with time?
- ...



Checkout

- Google Ngrams: <http://books.google.com/ngrams>
- New York Taxi Records



Data Science Process

- The standard data science process involves
 1. understanding the problem,
 2. preparing the data samples,
 3. developing the model,
 4. applying the model on a dataset to see how the model may work in the real world, and
 5. deploying and maintaining the models.

Data science process frameworks by Cross Industry Standard Process for Data Mining (CRISP-DM)

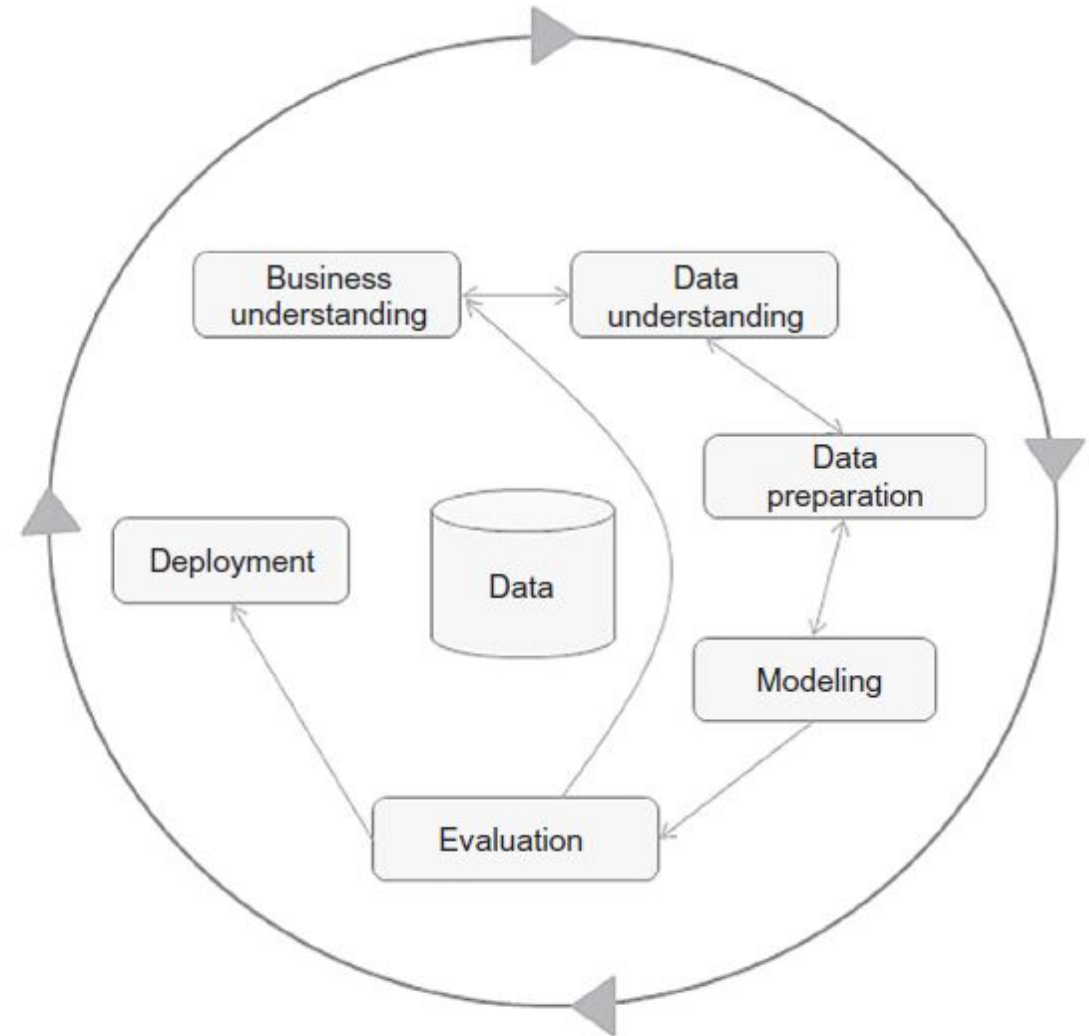


FIGURE 2.1
CRISP data mining framework.

The data science process presented in Fig. 2.2 is a generic set of steps that is problem, algorithm, and data science tool agnostic.

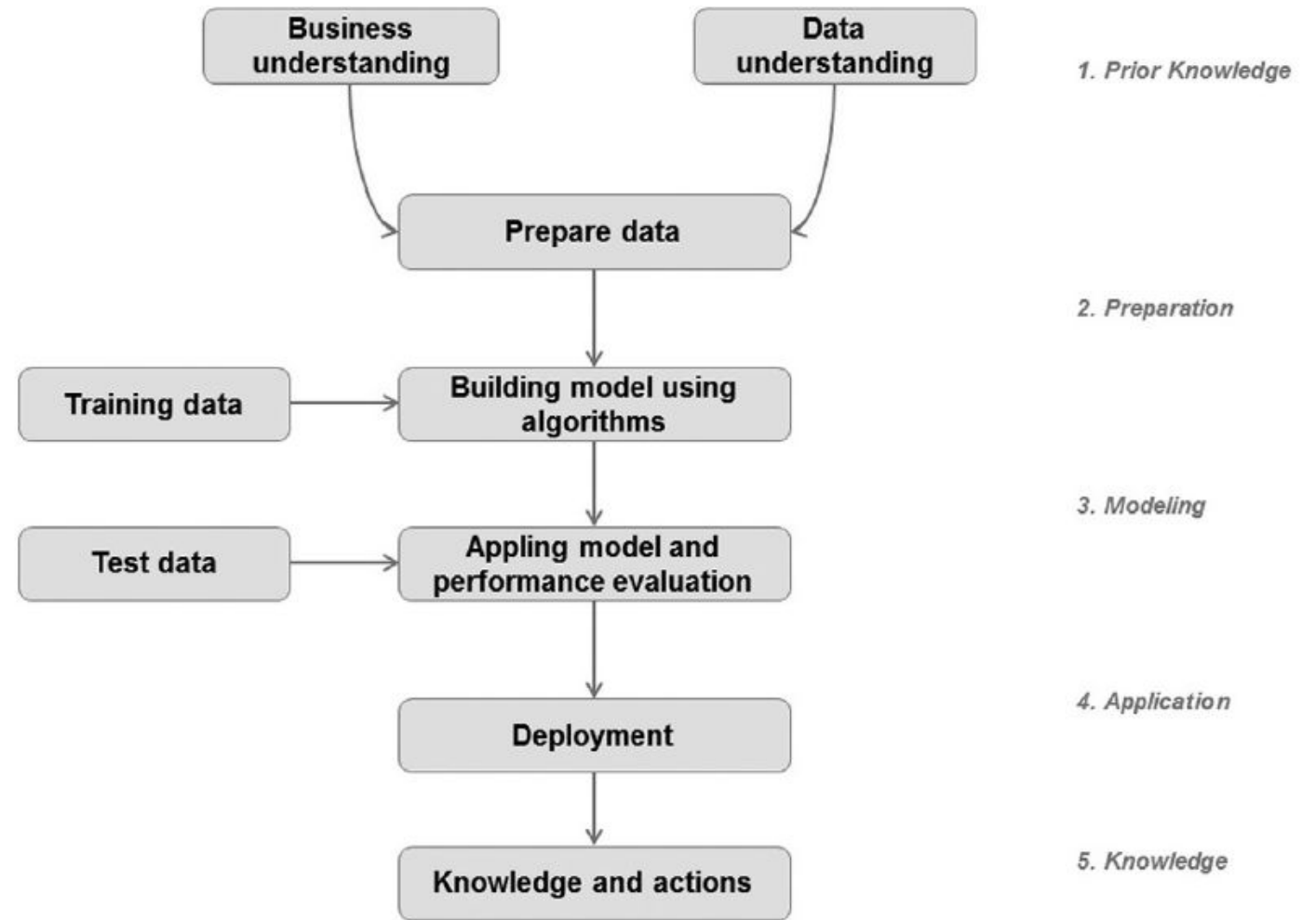



FIGURE 2.2
Data science process.

Step 1: Framing the data science question and understanding the context.



PRIOR KNOWLEDGE

- Prior knowledge refers to information that is already known about a subject.
 - The prior knowledge step in the data science process helps to define
 - what problem is being solved
 - how it fits in the business context, and
 - What data is needed to solve the problem.
- 

PRIOR KNOWLEDGE

Objective


- The data science process starts with a need for analysis, a question, or a business objective
- Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm.

Subject Area

- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes
- It is essential to know the subject matter, the context, and the business process generating the data.



PRIOR KNOWLEDGE

- Data
 - Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process
 - factors to consider: quality of the data, quantity of data, availability of data, gaps in data, does lack of data compel the practitioner to change the business question, etc.
- 

PRIOR KNOWLEDGE

- Terminology

- A **dataset** (example set) is a collection of data with a defined structure.
- A **data point** (record, object or example) is a single instance in the dataset.
- An **attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset.
- A **label** (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes
- **Identifiers** are special attributes that are used for locating or providing context to individual records.

- Dataset
- Data point
- Attribute
- Label
- Identifier

Roll No	Series (50)	Series (50)	Tutorial (10)	Assignment (5)	Attendance (%)	Attendance (10)	Total (50)
1	18	38	9	4	79	8	35
2	20	33	10	4	100	10	37
3	41	50	10	4	90	10	47
4	22	39	9	5	95	10	39
5	21	30	8	4	98	10	35
6	29	41	9	4	98	10	41
7	37	50	8	5	98	10	45
8	48	50	9	4	98	10	48
9	37	50	9	4	93	10	45
10	10	32	10	5	83	9	35
11	20	33	9	5	81	9	36
12	44	50	9	4	100	10	47
13	40	50	8	5	100	10	46
14	26	37	8	5	93	10	39
15	23	46	10	5	88	9	41
16	15	45	9	5	88	9	38
17	46	50	8	4	93	10	46

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

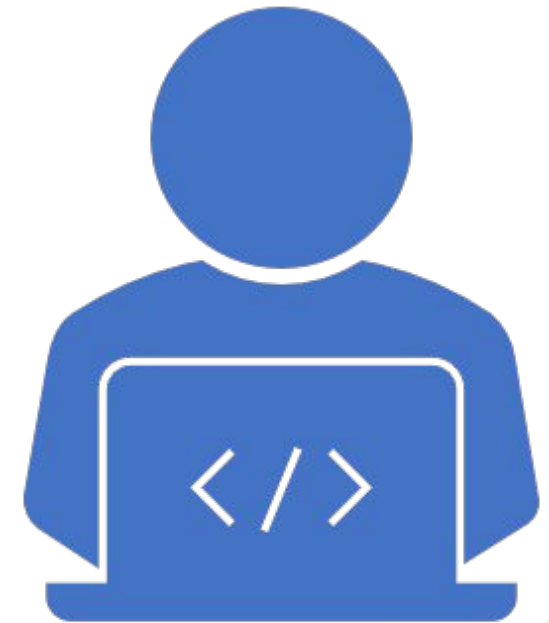


Causation Versus Correlation

- Suppose the business question is inverted
- The correlation between the input and output attributes doesn't guarantee causation.
- Hence, it is important to frame the data science question correctly using the existing domain and data knowledge.

DATA PREPARATION

- Most time-consuming part of the process.
- It is extremely rare that datasets are available in the form required by the data science algorithms.
- Most of the data science algorithms would require data to be structured in a tabular format with records in the rows and attributes in the columns.
- If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.



Data Exploration

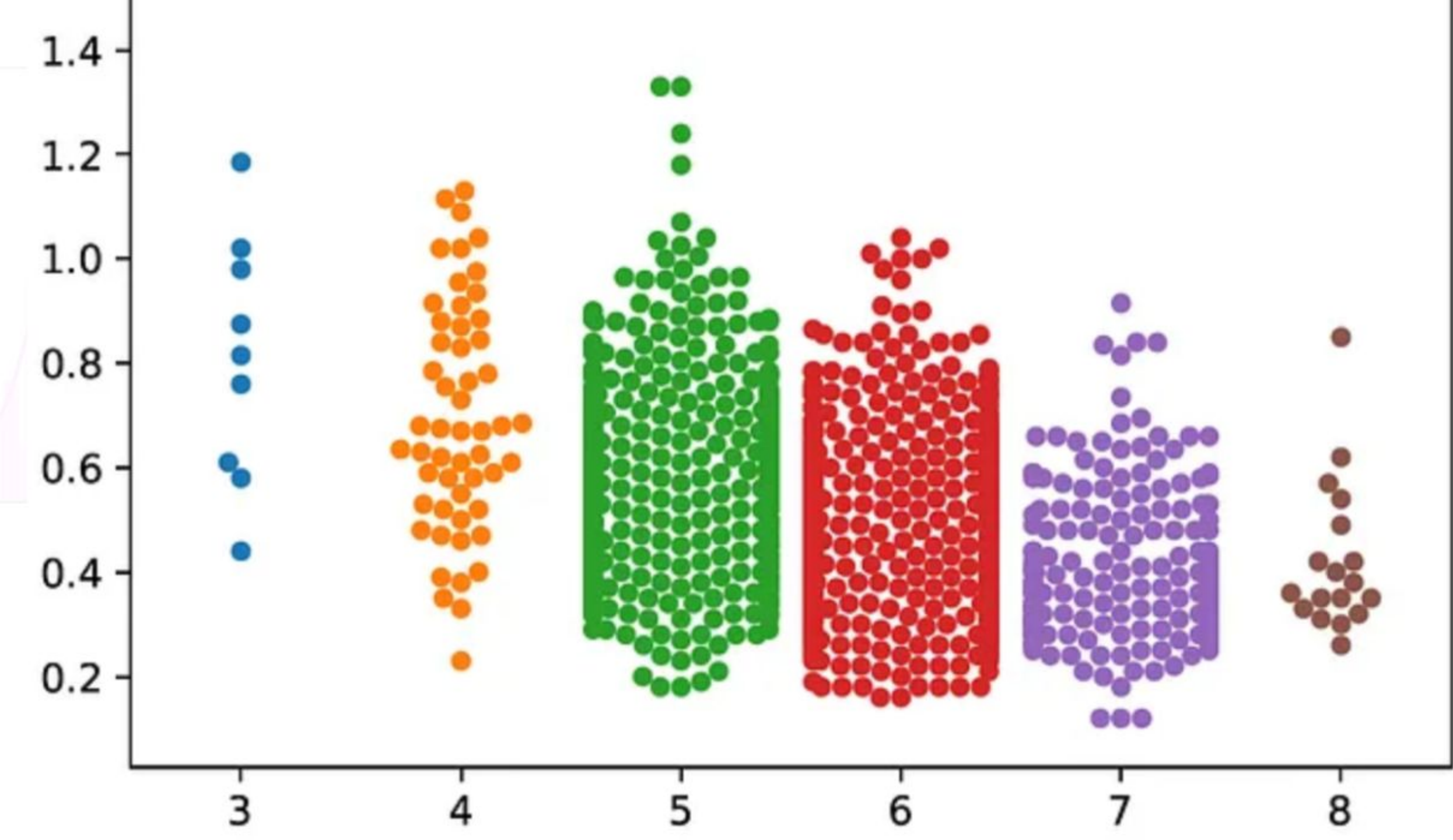
Data exploration, also known as **exploratory data analysis**, provides a set of simple tools to achieve basic understanding of the data.

Involves computing descriptive statistics and visualization of data

They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset.

Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data.

A visual plot of data points provides an instant grasp of all the data points condensed into one chart.



Data Exploration

Data Quality

Missing Values

Data Types and
Conversion

Transformation - normalized

Outliers –
anomalies

Feature
Selection

Data Sampling

MODELING

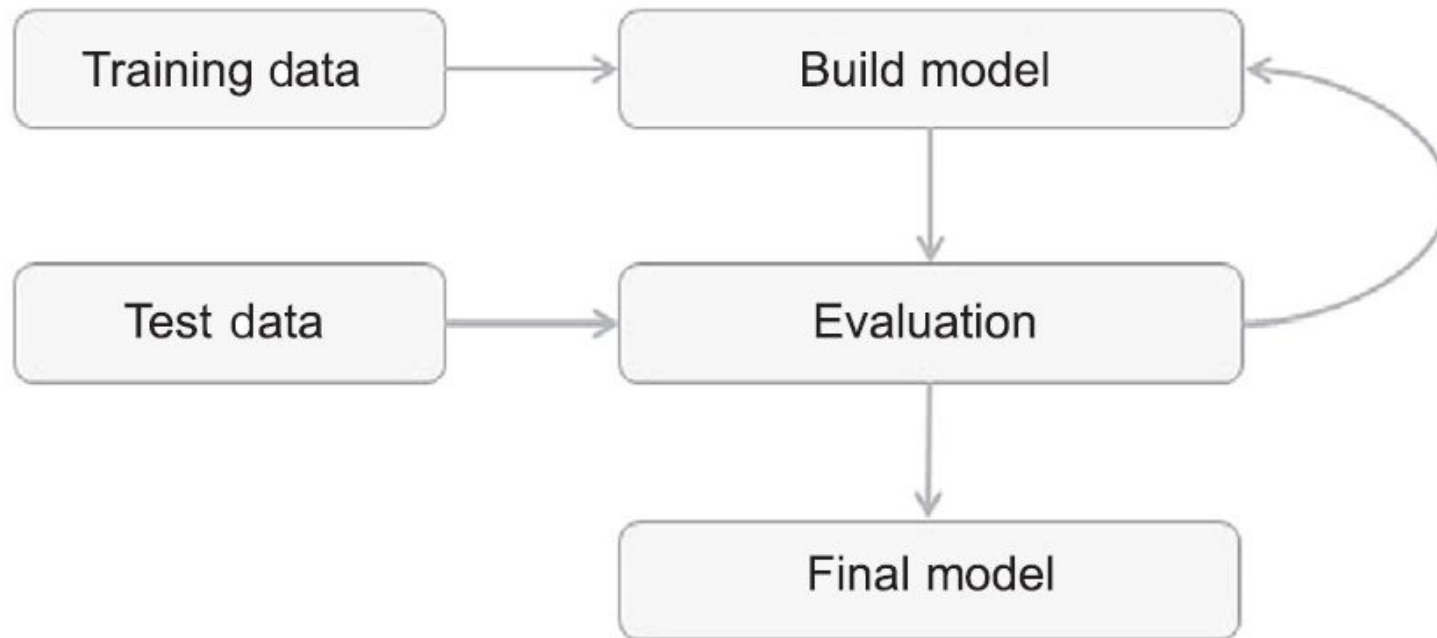


FIGURE 2.4
Modeling steps.

Training and Testing Datasets

Table 2.3 Training Dataset

Borrower	Credit Score (X)	Interest Rate (Y) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40
07	750	5.90
10	825	5.70

Scatterplot of the entire example dataset with the training and test datasets marked

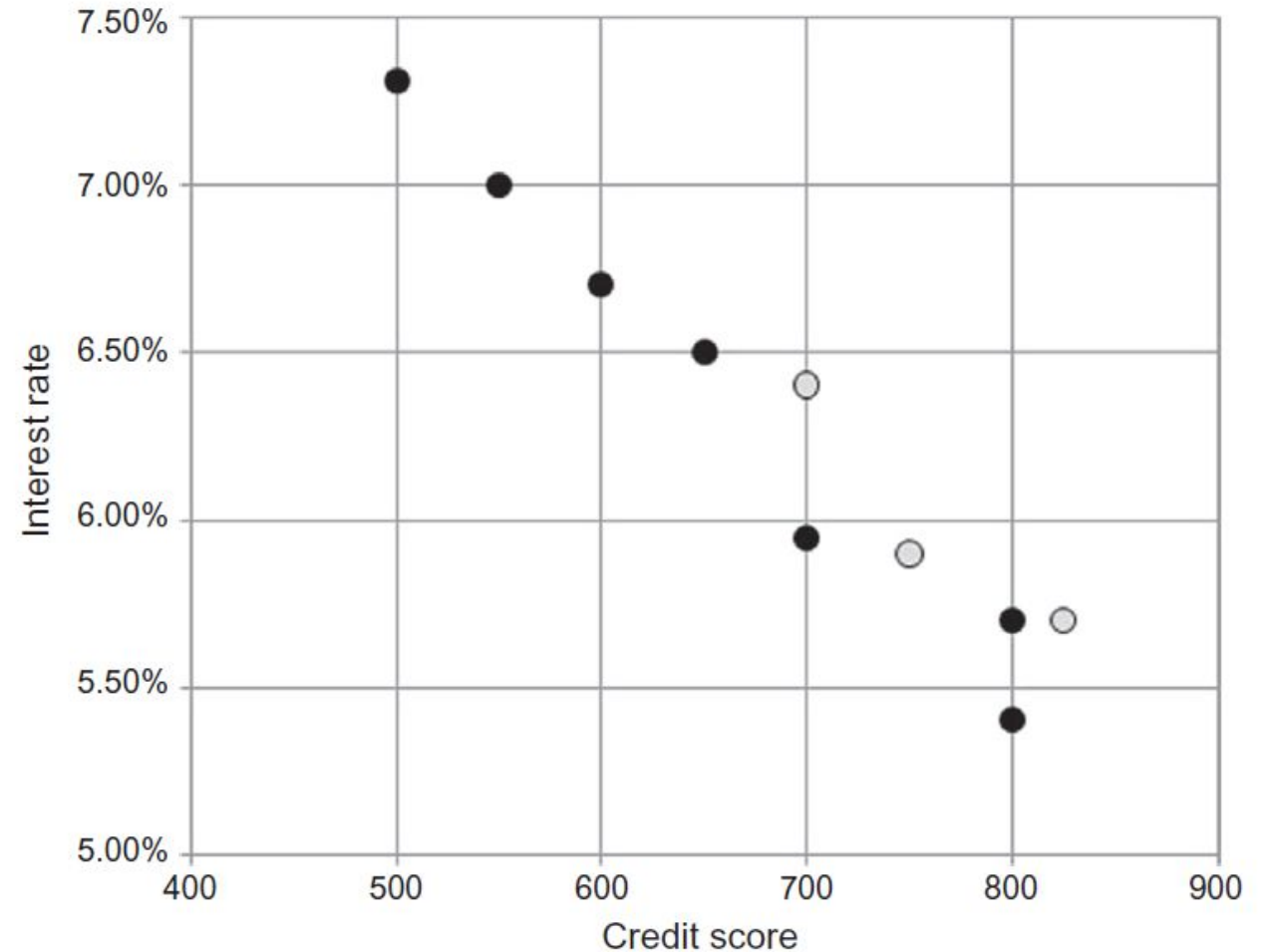


FIGURE 2.5

Scatterplot of training and test data.

Learning Algorithms

- Classification
 - decision trees, rule induction, neural networks, Bayesian models, k-NN
- Classification and regression tree (CART)
- CHi-squared Automatic Interaction Detector (CHAID)
- Simple linear regression technique

$$y = a * x + b$$

Evaluation of the Model

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (X)	Interest Rate (Y) (%)	Model Predicted (Y) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

Ensemble Modeling

At the
end of
the
modeling
stage

Analyzed the business question;

sourced the data relevant to answer the question;

selected a data science technique to answer the question;

picked a data science algorithm and prepared the data to suit the algorithm;

split the data into training and test datasets;

built a generalized model from the training dataset; and

validated the model against the test

APPLICATION

Production Readiness

Technical Integration

Response Time

Model Refresh

Assimilation