Class 5 Data Analytics Procedures

Dr. Wei Miao

UCL School of Management

October 15, 2025

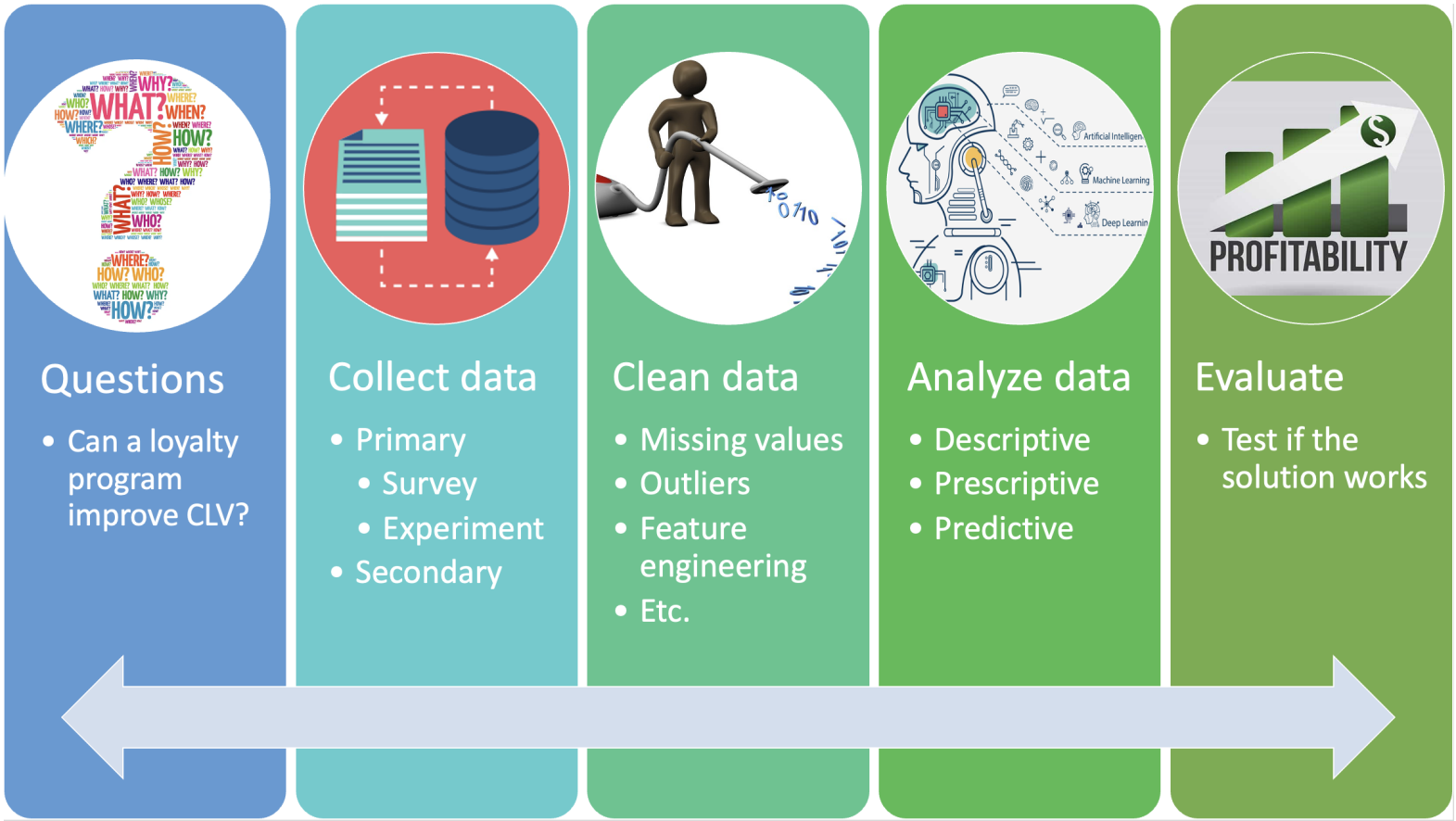
Section 1

Data Analytics Workflow

Class Objectives

- Understand the major steps to conduct data analytics. We will use the M&S case study to illustrate how to improve marketing efficiency.
- **Data collection:** Learn how to collect first-hand survey data and how to load second-hand data into R
- **Data cleaning:** Learn how to use the `dplyr` package to clean data
- **Data analysis:** Learn how to conduct descriptive analytics for the M&S case study

Overview of a Data Analytics Project



Business Objective: Our Business Question in Weeks 3 - 5

Our project for M&S in Weeks 3-5: Help M&S to **improve marketing efficiency** by improving its **ROI** on its targeted marketing offers. The project will involve data collection, data cleaning, and data analysis, including both **descriptive** and **prescriptive** analytics, to identify the most profitable customers and develop a personalised marketing targeting strategy.

Business Objective: Example Dissertation Projects in Term 3: CLV and CRM

Customer Lifetime Value and Customer Relationship Management Projects

- "At ***Lebara***, as in many subscription-based companies such as Netflix, Customer Lifetime Value (CLTV) is a crucial metric that informs our approach to customer acquisition, retention, and investment strategies. CLTV enables us to understand the long-term revenue potential of each customer, guiding how we allocate resources to attract and retain valuable subscribers. Alongside CLTV, accurately predicting customer churn is equally important, as it allows us to proactively address retention risks and optimise customer engagement efforts. Until now, our approach has relied on a static, rules-based model that provides insights at a cohort level, offering only a broad view rather than individual predictions. To improve accuracy, we aim to move beyond this cohort-based approach by building a tenure prediction model that will feed into and enhance our CLTV model. This project will focus on developing a machine learning-driven tenure prediction model capable of estimating each customer's likely duration with Lebara. By forecasting tenure at the individual level, this model will ultimately allow us to create a more dynamic and precise CLTV model, supporting data-informed decisions on customer investment, targeted retention strategies, and overall customer experience enhancements."

Business Objective: Example Dissertation Projects in Term 3: Predictive Analytics

Predictive Analytics / Machine Learning Projects

- ***Burberry*** provides relevant **product recommendations** on Burberry.com to facilitate in-session **product exploration** and to create a more personalised user experience. This project is to develop a new **product recommendation system** that tailors suggestions to individual users based on their product selections and preferences.
- The ***AXA*** project will explore **fraud detection approaches** using **unsupervised ML** including models such as isolation forests. The candidate will develop an understanding of the business problem and our data, formulating hypotheses and testing them. They will build, evaluate, and interpret their ML models.
- At ***Waitrose***, it's crucial to balance product availability with minimising waste by understanding sales rates. Factors like product shelf life, varying sales velocities, promotions, and unexpected trends make it challenging to find a one-size-fits-all solution. Current manual forecasting introduces inaccuracies and process delays. We aim to develop a **machine learning algorithm** to generate daily **product forecasts**, integrate with our stock management system, and enable automated, accurate forecasting.

Business Objective: Example Dissertation Projects in Term 3: Causal Inference

- At ***The Economist***, we aim to build models that predict the likelihood of an experiment's success without needing to run A/B tests. Currently, every change to our website and apps goes through an A/B test to assess its impact on key metrics such as Customer Lifetime Value (CLTV). However, the time it takes for an A/B test to reach statistical significance limits our ability to quickly iterate on new ideas. More businesses are now exploring Causal Inference modelling in this context. We have accumulated data from past experiments and are keen to start building our own models.

Section 2

Data Collection

Types of Data by Source

- **Primary Data:** Data that are generated by the data analyst through surveys, interviews, and experiments, which are specially designed for understanding and solving the research problem at hand.
- **Secondary Data:** Existing data generated by the company's or consumer's past activities, as part of organisational record-keeping.

Basis for Comparison	Primary Data	Secondary Data
Meaning	Primary data refers to the first-hand data gathered by the analyst.	Secondary data means data collected by someone else earlier (usually by the company).
Data	New data	Historical data in the past
Source	Surveys, observations, experiments, questionnaires, personal interviews, etc.	Company databases, government publications, websites, books, journal articles, internal records, etc.
Cost	Expensive; very involved and costly	Economical; quick and easy
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's needs.

Types of Data by Structure

- We often consider two dimensions of a dataset: units and time.
  - **Cross-sectional data**: data collected at a single point in time.
  - **Time-series data**: data collected over time for a single individual or entity.
  - **Longitudinal data**: data collected over time that may not contain the same individuals at each time point.
  - **Panel data**: data collected over time that contain the same individuals.

Cross-sectional Data

- Many datasets we will use in this course are cross-sectional data, which contain observations on multiple subjects (e.g., customers) at a single point in time.
- For example, the following table shows data for 5 customers at a specific point in time.

CustomerID	Age	Income	Last Purchase Date
CUST001	34	58000	2023-10-15
CUST002	45	92000	2023-10-12
CUST003	28	45000	2023-10-16
CUST004	52	120000	2023-10-11
CUST005	21	35000	2023-10-15

Time-series Data

- Time-series data track a single subject (e.g., a customer) over multiple time periods.
- For example, the following table shows monthly spending data for a single customer over 6 months.

Month	Spending
Jan 2023	200
Feb 2023	250
Mar 2023	300
Apr 2023	220
May 2023	280
Jun 2023	350

Longitudinal Data

- Longitudinal data track multiple subjects over time, but the subjects may differ at each time point. This is common in survey data where different respondents are surveyed at different times. Sometimes, longitudinal data are also called **repeated cross-sectional data**.
- For example, the following table shows survey data collected from different customers in 2022 and 2023.

Year	CustomerID	Satisfaction Score
2022	CUST001	8
2022	CUST002	7
2022	CUST003	9
2023	CUST004	6
2023	CUST005	8

Panel Data

- Panel data track the same subjects over multiple time periods, allowing analysis of changes within individuals over time. This is the ideal data structure for understanding individual-level dynamics.
- Panel data can be unbalanced (different number of observations per subject) or balanced (same number of observations per subject).
- For example, the following table shows panel data for 3 customers over several time periods.

CustomerID	Year	Spending
CUST001	2022	500
CUST001	2023	600
CUST002	2022	700
CUST002	2023	800
CUST003	2022	400
CUST003	2023	450

Section 3

Marketing Survey

Primary Data: Marketing Surveys

- A marketing survey is often the easiest and most cost-effective way to collect primary data. We often collect the following variables:
  - purchase intention: how likely a customer will buy a product, helps to predict sales
  - willingness to pay: how much a customer is willing to pay for a product, helps to set the optimal price
  - shopping basket: what products a customer usually buys, helps to cross-sell
  - share of wallet: how much a customer spends on a product category, helps to identify the high-potential customers for market penetration
  - demographics: gender, age, income, education, etc., helps to segment customers
- Let's see an example in Mentimeter of how to design a marketing survey!
- You can design surveys to collect data for your Term 1 projects or Term 3 dissertation.
  - The quick start guide on how to conduct market research surveys

Limitations of Marketing Surveys

- **Hawthorne Effect** and **Response Bias**: Participants may answer in ways they think are socially desirable or expected, rather than their true feelings or behaviour.
- **Sampling Bias**: The sample may not be representative of the customer population.
- **Fatigue**: Long surveys may lead to respondent fatigue, causing rushed or careless answers toward the end. Ask a reasonable number of questions and place important questions earlier on.