COMP4131: Data Modelling and Analysis

Lecture 1: Introduction to Data Modelling and Analysis

Kian Ming Lim

University of Nottingham Ningbo China

kian-ming.lim@nottingham.edu.cn

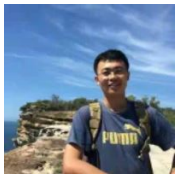
February 15, 2025

Outline

Teaching Team

Introduction to the Teaching Team - Lecturers

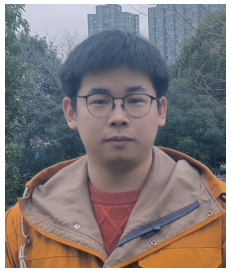
Assoc. Prof. Dr. Kian Ming Lim
PMB-424
Kian-Ming.Lim@nottingham.edu.cn
Office Hours: Monday, 09:00 - 10:00, 14:00 - 15:00
Research Profile: Link



Asst. Prof. Dr. Daokun Zhang
IAMET-229
Daokun.Zhang@nottingham.edu.cn
Office Hours: Monday, 09:00 - 11:00
Research Profile: Link

Introduction to the Teaching Team - Tutors

PhD Student Yue Yang
scxyy2@nottingham.edu.cn



PhD Student Qinglin Mao
scxqm1@nottingham.edu.cn

Overview of the Module

Overview of the Module

- To introduce the principles, techniques, and applications of data analysis and modelling.
- To enable students to recognize the most widely-used data analysis and modelling techniques and determine the appropriate technique for specific applications.
- To enable students to understand and apply computer-based data analysis and modelling techniques in practice.

Learning Outcomes

- **Knowledge and Understanding**:
  1. Understanding the capabilities, strengths and limitations of data analysis and modelling methods
  2. An appreciation of different data analysis and modelling techniques
- **Intellectual Skills**:
  1. The ability to understand complex ideas and relate them to specific situations
- **Professional Skills**:
  1. The ability to implement selected data analysis and modelling methods for real world applications
  2. The ability to evaluate data analysis and modelling techniques and select those appropriate to a given task
- **Transferable Skills**:
  1. The ability to address real problems and assess the value of their proposed solutions
  2. The ability to retrieve and analyse information from a variety of sources and produced detailed written reports on the result

Learning Activities

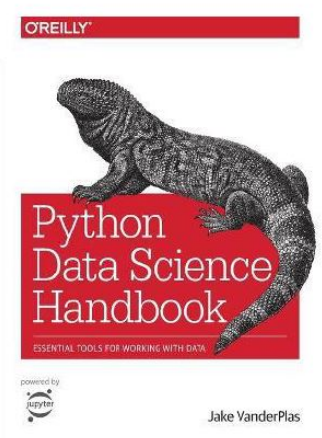
- 2 hours Lecture and 2 hours Computing / 10 weeks
  - Lecture: Monday 12.00pm - 2.00pm (DB C06)
  - Computing: Monday 4.00pm - 6.00pm (IAMET 406)
- Module managed through Moodle (for all resources)

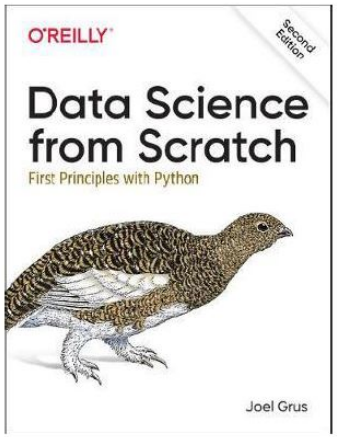
Assessment

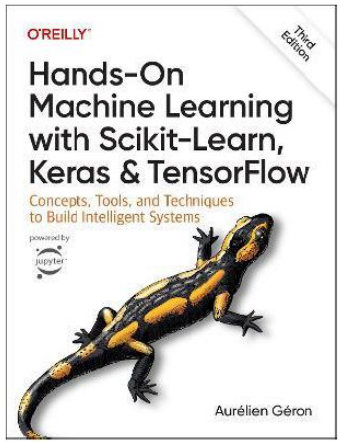
- Coursework 1 (25%)
  - Lab submission
  - You will need to complete and submit the lab. Each lab submission 2.5%.
- Coursework 2 (75%)
  - Data analysis study (code and 4000 words)
  - You will need to select a data set and use all the data modelling and analysis steps you have learnt to analyse and wrangle the data set, and create and compare models.
  - You will write your work up as an academic paper - comparing and analysing your results at every stage of the data analysis and modelling pathway.

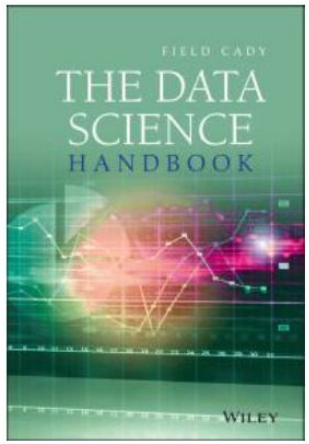
Module Resources

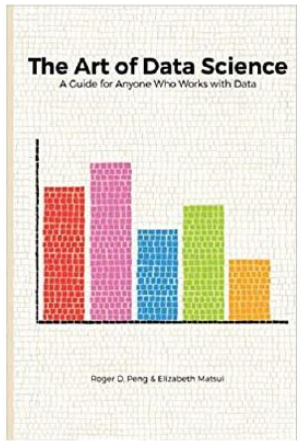
- Textbook:











- Reading list: https://rl.talis.com/3/notts/lists/F44CC47F-9CE2-1B7B-8FD5-903D6A49B5B2.html
- Programming language: Python
- Software: Jupyter Notebook (Anaconda)

Expected Workload

Activity	**Hrs by Week**	**Hours**
Lecture – deliver key material	2×10	20
Computing – putting theory into practice	2×10	20
Self-study – revise lecture material	2×10	20
Lab submission (25%) – complete the lab exercises		40
Coursework (75%) – data modelling and analysis to solve a real-world problem, and writing an academic paper.		100
Total (20 credits)		200

Schedule for the Module

Week	Lecture	Lab
1	Introduction to Data Modelling and Analysis	Jupyter Notebooks + Basic Data Manipulation
2	Data Wrangling and Pre-processing	Data Wrangling and Pre-processing
3	Data Visualisation	Visualisation
4	Analysis and Modelling	Exploratory Data Analysis
5	Unsupervised Learning	Unsupervised Machine Learning
6	Linear Regression and Logistic Regression	Linear Regression and Logistic Regression
7	Gaussian Process Regression and Classification	Gaussian Process Regression and Classification
8	Naïve Bayes and K Nearest Neighbors	Naïve Bayes and K Nearest Neighbors
9	Decision Tree and Random Forest	Decision Tree and Random Forest
10	Support Vector Machines and Kernel Methods	Support Vector Machines and Kernel Methods

Data Modelling and Analysis and your Degree

- **Core Knowledge:** DMA is key for understanding data. You'll learn to represent real-world data for computer processing.
- **Computational Thinking:** It improves your computational thinking. You'll learn to break down complex data problems. When pre-processing data for machine learning, you'll clean, normalize, and extract features. These skills are vital for creating intelligent algorithms.
- **Interdisciplinary Applications:** DMA has strong interdisciplinary connections. In today's digital age, data is everywhere, and the knowledge from this module can be applied across different fields.
- **Research Skills:** If you're interested in research, this module gives you the tools. You'll work with modern data analysis tools and techniques, which are essential for research in big data, AI, and data-driven optimization.

Data Modelling and Analysis and your Career

- Data modelling and analysis are fundamental skills for any data-oriented professional.
- The ability to perform data modelling and analysis is highly sought after in the industry. It is a key component of many data-related job roles, such as data analyst, data scientist, data engineer, business intelligence analyst, and many more.

Data Science vs Machine Learning

- Coursera
- https://www.coursera.org/articles/data-science-vs-machine-learning

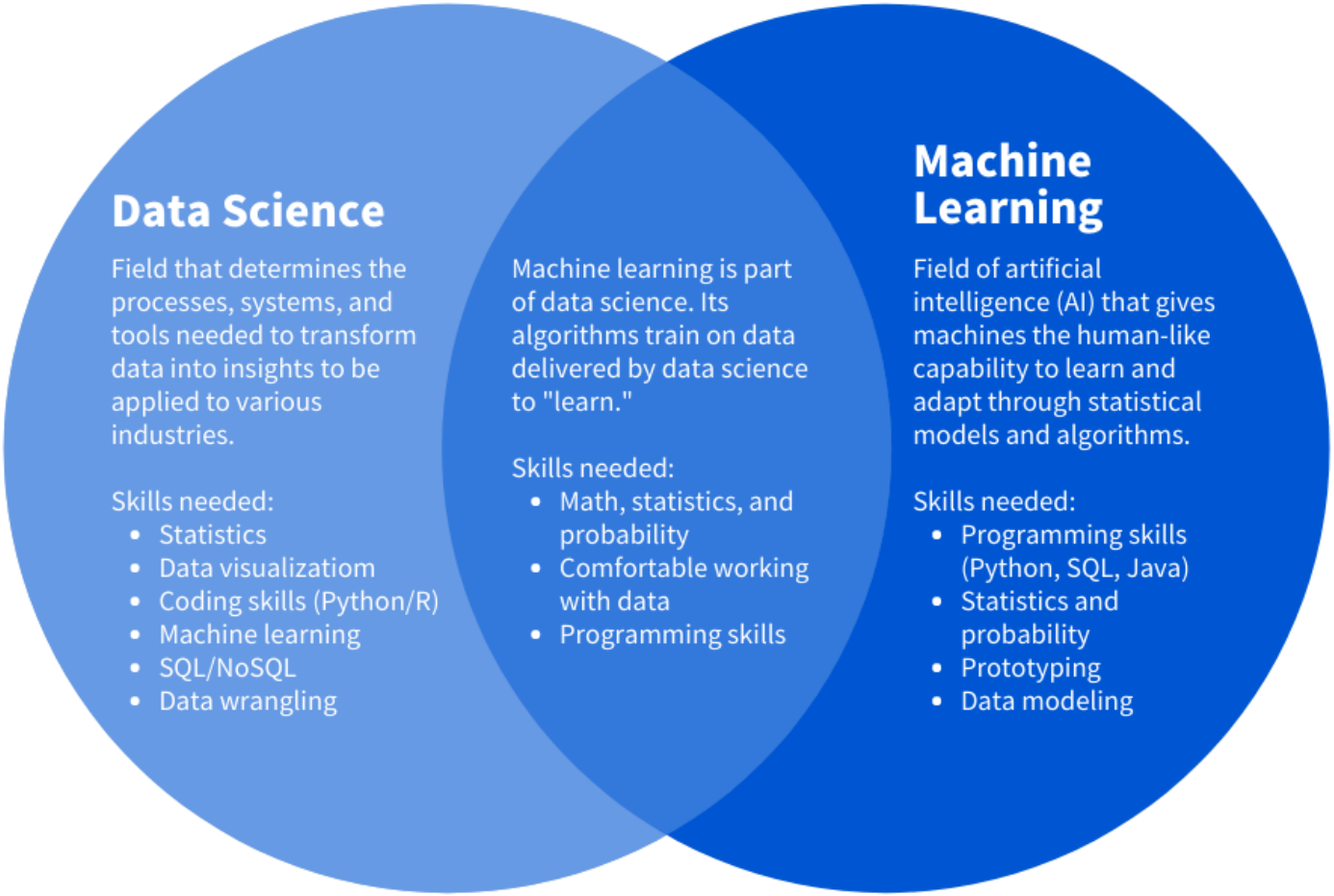


Image Source: Data science vs. machine learning: What's the difference?

Communication, Attendance, Academic Integrity

Module Communication and Q&A

- All module communication will be performed via the "Announcement" forum on Moodle.
  - Please check the forum regularly for important updates.
- If you have a question about the module, please post it on the "Q&A" forum on Moodle.
  - We will respond to your question as soon as possible.
  - If you have a question about the module, it is likely that other students have the same question. Therefore, please post your question on the forum rather than emailing the teaching team directly.
- If you have a question about your personal circumstances, please email the teaching team directly.

Attendance

- Attendance is compulsory for all lectures and labs.
- Attendance monitoring is performed by the University - the teaching team does not mark attendance, nor have the ability to change your attendance record.
- If you are unable to attend, you must obtain an authorised absence via the University's "Extenuating Circumstances" procedure.
- Please attend the lab session on your timetable.
  - Lab groups are organized by the University timetabling team. The teaching team cannot change your assigned group.

Academic Integrity and Practical Advice

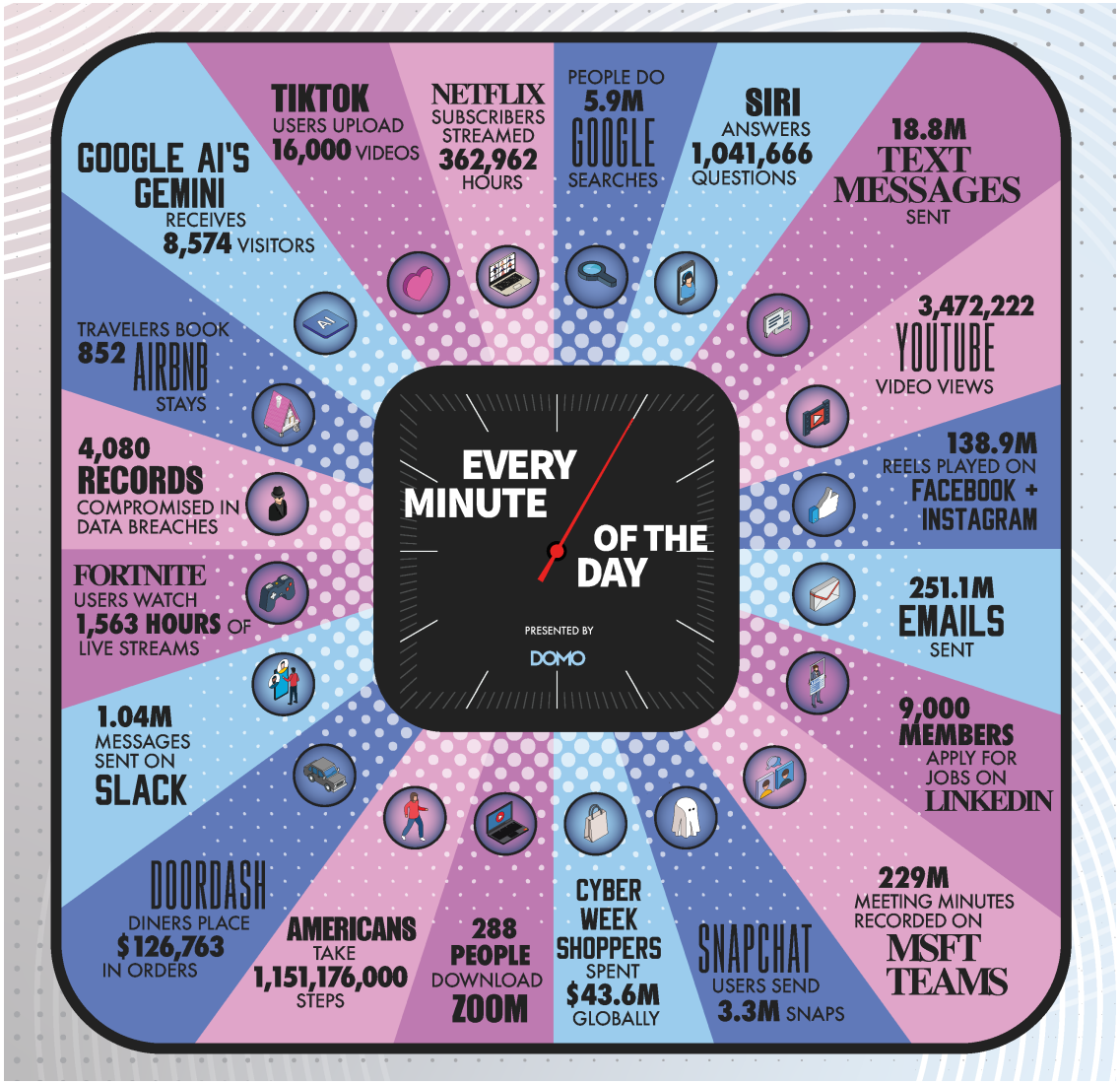
- You are expected to complete all module work independently.
- Please be familiar with the University's Academic Misconduct policy. `https://www.nottingham.edu.cn/en/academic-services/academic-misconduct/academic-misconduct.aspx`
- We do check every submission for plagiarism. Every year, students are caught plagiarising and are penalised accordingly. You do not want to be one of these students.
- Our practical advice are:
  - If you are unsure about what constitutes plagiarism, please ask the teaching team.
  - Do not copy code from the internet without referencing it.
  - Do not share your code with other students.
  - Do not share your code on public repositories (e.g. GitHub).
  - Be cautious of your dorm-mates and friends asking for your code.

An Insight into Data Modelling and Analysis

How Much Data Is Collected Every Minute of the Day?

Some stats as of late 2024:

- 5.52 billion people – approximately 67.5% of the global population – are online
- Total amount of data created, captured, copied, and consumed globally is expected to reach 149 zettabytes by the end of 2024, with projections surpassing 394 zettabytes by 2028.



https://www.domo.com/learn/infographic/data-never-sleeps-12

DMA as a Quest for Efficient Planning

Some of the earliest records of data collection and analysis:

- Sumerian Cuneiform Clay Tablets Recording Labour Workforce Data 4000 BC
- To plan food requirements for each member of the population
- Record beer given to workers as part of their daily rations

Sources: 1 2 3

Sumerian Cuneiform Tablet Account of workers.

Sumerian Cuneiform Tablet Record of Beer.

DMA as a Quest for Survival: Data Visualisation

- Florence Nightingale the Statistician who knew how to communicate data
- Source: here

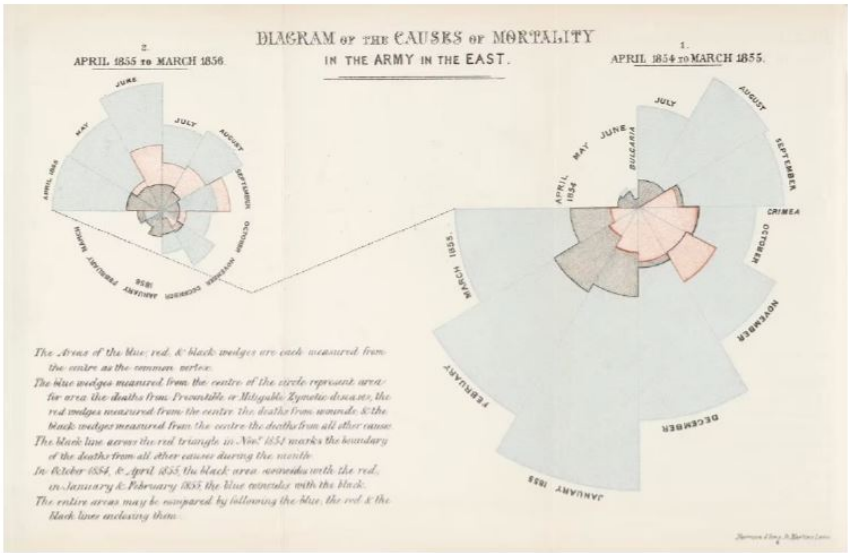


Image Source: Diagram originally from Florence Nightingale's book, Notes on matters affecting the health, efficiency, and hospital administration of the British Army. https://wellcomecollection.org/works/jxwtskzc/items
Video: Florence Nightingale - Joy of Stats

Interactive Visualisation: The Joy of Stats

- Hans Rosling
- Gapminder Tools: Interactive Bubbles Chart

Video Source: Hans Rosling: The Joy of Stats (YouTube Link)

Ubiquitous Computing – Everywhere and Anywhere Data

- Sensemaking environment of connected devices, interactive visualizations, and multiple temporal and contextually varying data sources and dependencies.
- Tools for augmenting and transforming our cognitive activities

Video Source: SDF Farm Management (YouTube Link)

ACTIVITY . THINK-PAIR-SHARE

- Pair up with the person next to you and discuss the challenges and difficulties that will be faced, as shown in the video on the previous slide.
- Share your answers with us!

Ubiquitous Computing – Everywhere and Anywhere Data

- Is the farmer better informed to carry out their role than your doctor?
- What are the ethical issues of transferring this model to people?
- The Reality: link

Video Source: https://www.youtube.com/watch?v=aThi1z0I_L4

Data Modelling and Analysis Pipeline

Data Modelling and Analysis Pipeline

- The building block of Data Science

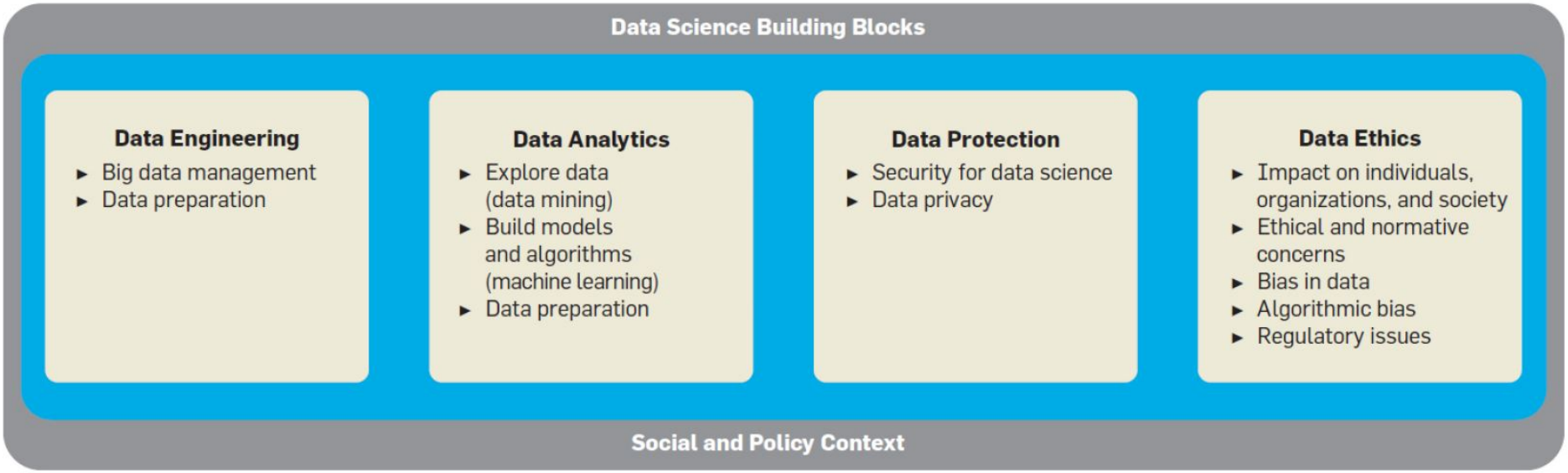
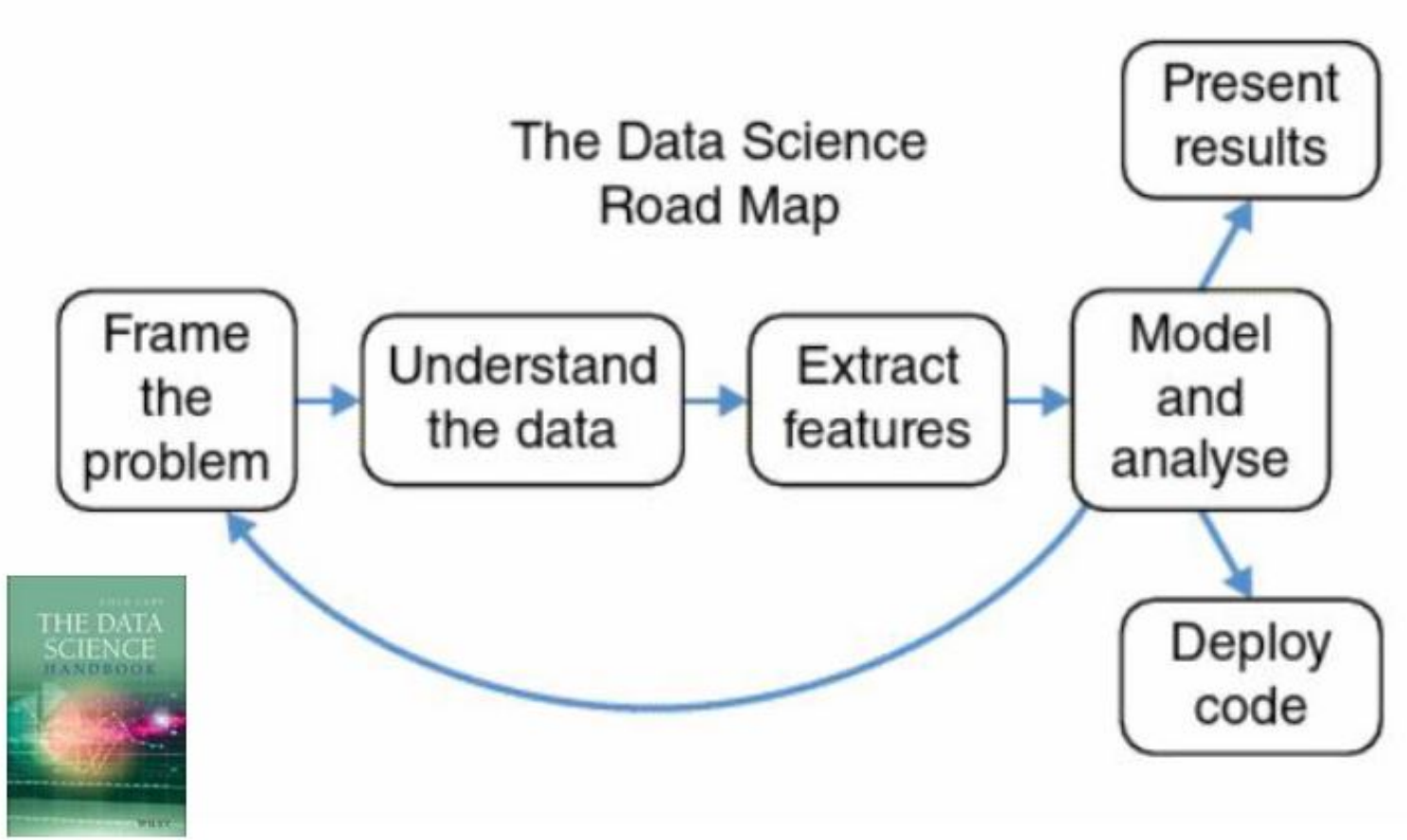


Image Source: Data Science by Ozsu

Data Modelling and Analysis Pipeline

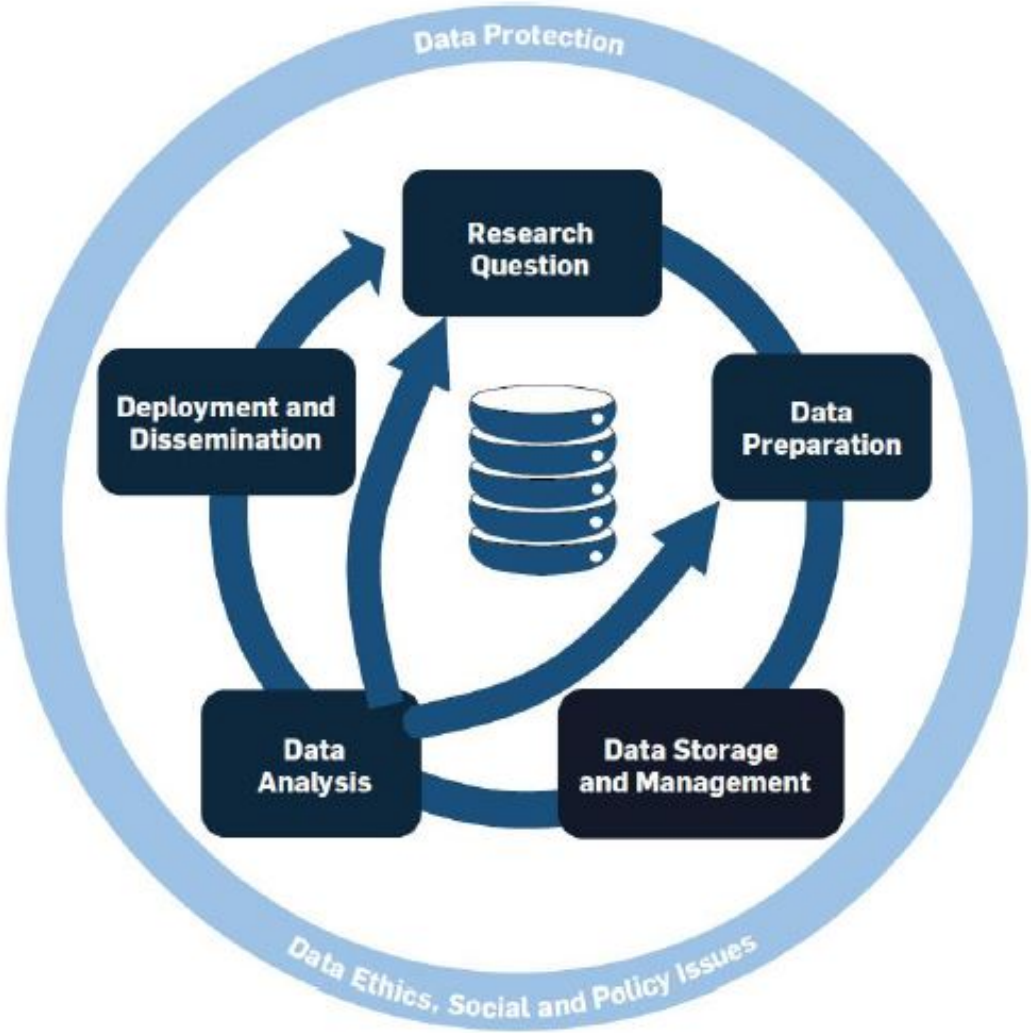
- Understand how to apply the Data Science investigation cycle



From: The Data Science Handbook, Field Cady

Data Modelling and Analysis Pipeline

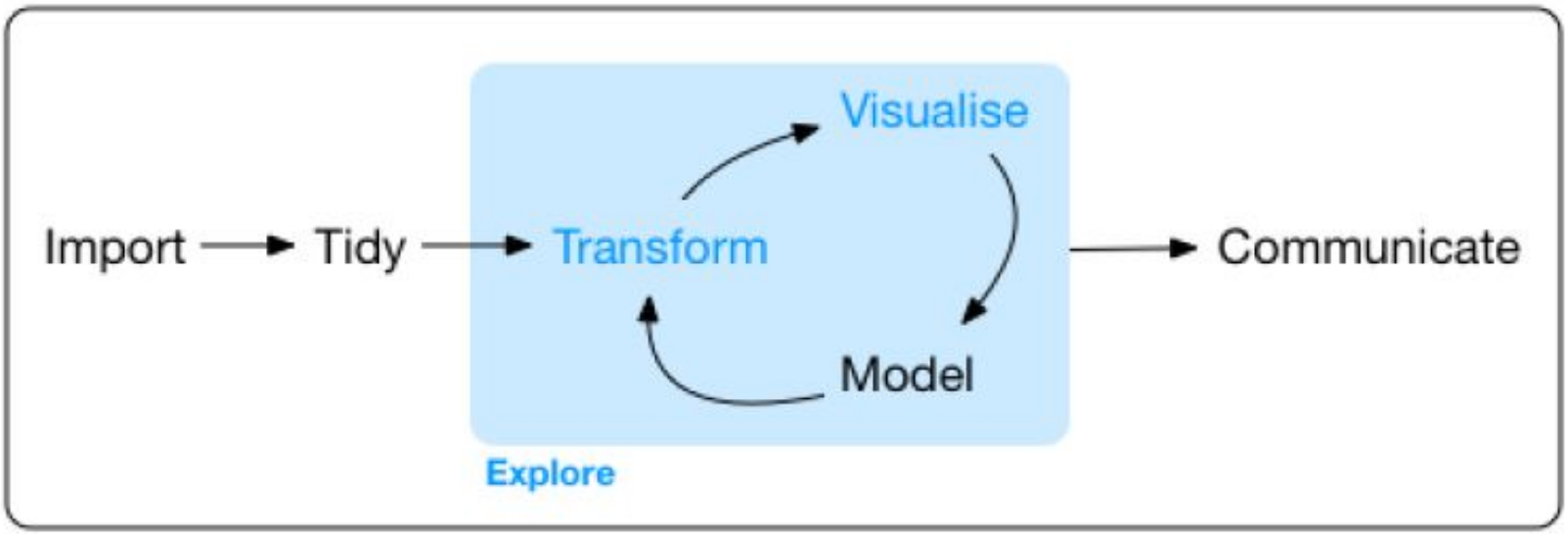
- Understand how to apply the Data Science investigation cycle



From: Data Science – A Systematic Treatment, Ozsu, Comms ACM 2023

Data Modelling and Analysis Pipeline

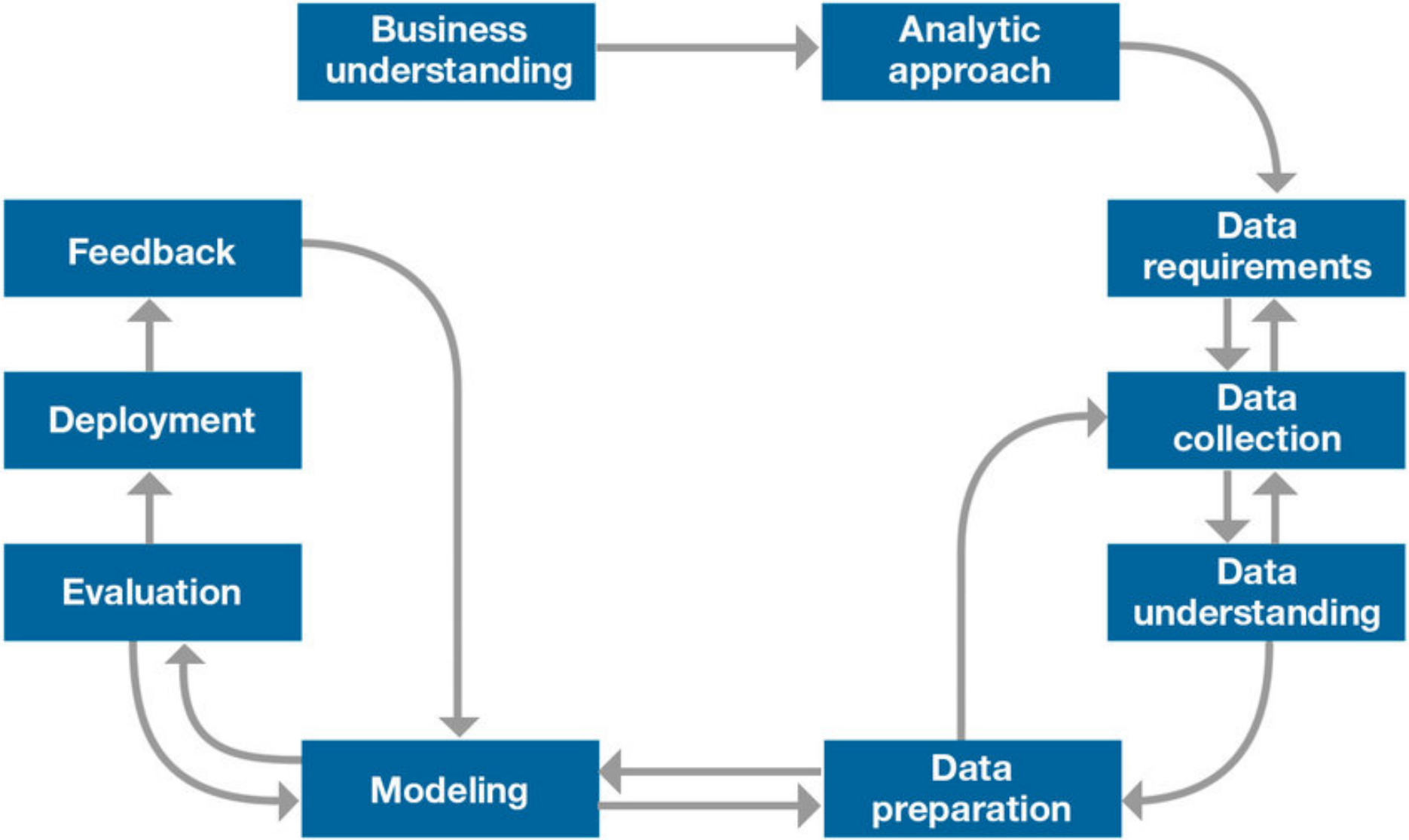
- Understand how to apply the Data Science investigation cycle



From: R for Data Science, Wickham and Grolemund

Data Modelling and Analysis Pipeline

- IBM



Data Modelling and Analysis Pipeline

- Google Cloud
- https://cloud.google.com/blog/topics/developers-practitioners/intro-data-science-google-cloud

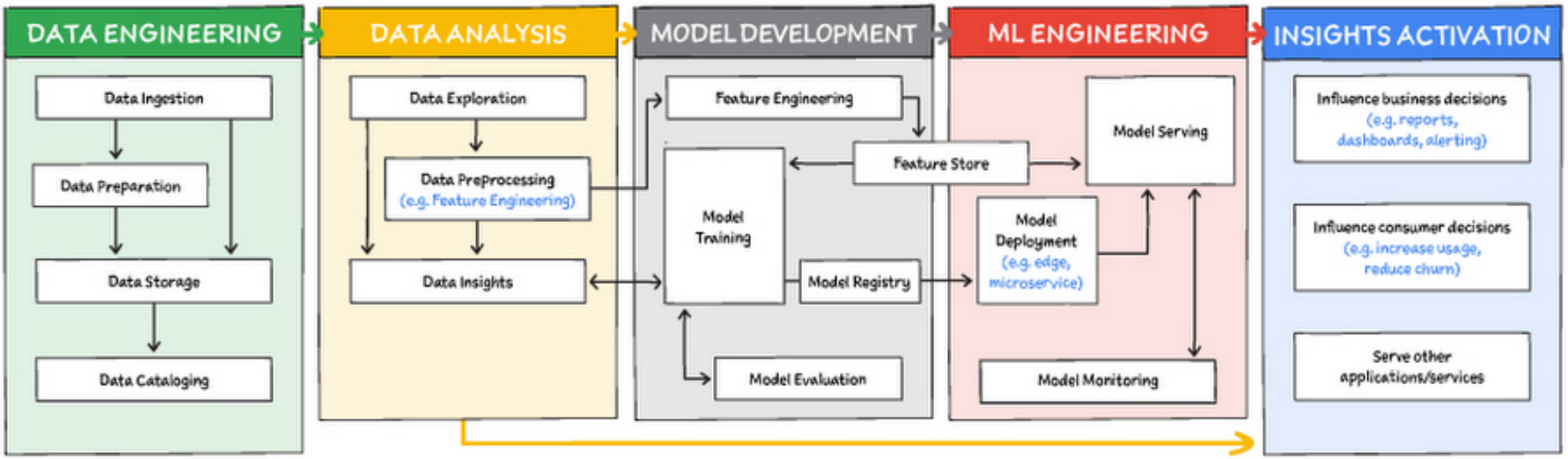


Image Source: Intro to Data Science on Google Cloud

Data Engineering

Objective: Prepare and organize raw data for downstream analysis and model development.

1. **Data Ingestion:**
   - Collect data from various sources (databases, APIs, streaming data, etc.).
   - Formats: CSV, JSON, databases, etc.
2. **Data Preparation:**
   - Cleaning, deduplication, and integration of datasets.
   - Handling missing values and ensuring data consistency.
3. **Data Storage:**
   - Centralized storage using data lakes, warehouses, or cloud storage.
   - Examples: AWS S3, Google BigQuery.
4. **Data Cataloging:**
   - Metadata management for better discoverability.
   - Tools: Data catalog services to enable seamless data discovery.

Data engineering ensures data quality and accessibility to support reliable analysis and modeling.

Data Analysis

Objective: Extract insights and prepare data for modeling.

1. **Data Exploration:**
   - Perform exploratory data analysis (EDA) using visualizations and summary statistics.
   - Tools: Python (pandas, matplotlib), SQL, Tableau, etc.
2. **Data Preprocessing (Feature Engineering):**
   - Transform raw data into meaningful features for modeling.
   - Techniques: normalization, encoding categorical variables, creating derived features.
3. **Data Insights:**
   - Generate actionable insights to understand trends and patterns.
   - Example: Identifying correlation, seasonality, or data anomalies.

Data analysis bridges the gap between raw data and useful insights, setting the stage for successful model development.

Model Development

Objective: Train, evaluate, and optimize machine learning models.

1. **Feature Engineering:**
   - Improve model performance by selecting, transforming, and creating features.
2. **Model Training:**
   - Train machine learning models using algorithms such as linear regression, decision trees, neural networks, etc.
   - Tools: Scikit-learn, TensorFlow, PyTorch.
3. **Model Evaluation:**
   - Assess the model's performance using metrics (e.g., accuracy, precision, recall, F1-score).
   - Perform cross-validation to validate the model's robustness.
4. **Model Registry:**
   - Version control and management of models to ensure traceability.

Model development transforms features into actionable predictive capabilities through model training and evaluation.

ML Engineering

Objective: Deploy and monitor machine learning models for real-world applications.

1. **Model Deployment:**
   - Deploy models as APIs, microservices, or edge devices.
   - Deployment types: Batch, real-time, and on-device.
2. **Model Serving:**
   - Serve predictions to applications or dashboards.
   - Infrastructure tools: Kubernetes, Docker, AWS Lambda.
3. **Model Monitoring:**
   - Track model performance in production (accuracy drift, latency).
   - Monitor data for changes that can affect model performance.

ML engineering ensures that models are effectively integrated into business workflows and remain accurate over time.

Insights Activation

Objective: Leverage model outputs and data insights to support decision-making.

1. **Influence Business Decisions:**
   - Use insights to generate reports, dashboards, and alerts.
   - Example: Dashboards showing real-time KPIs for decision-making.
2. **Influence Consumer Decisions:**
   - Personalize recommendations, offers, and experiences.
   - Example: Reducing customer churn or increasing user engagement.
3. **Serve Other Applications/Services:**
   - Integrate insights with external systems (e.g., customer service platforms, fraud detection).

Insights activation drives business value by turning model outputs into meaningful actions that improve processes and user outcomes.