THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

ENGG7811 | Research Methods
Lecture 7: Comprehensive Data Handling in Research and Study

Dr. Lu Zhang

Introduction

What is data handling?

Lifecycle of data from, collection, through to storage, analysis, interpretation, and management.

Foundational skill that will not only support your current studies but will also be invaluable in your future professional work or research.

Undergraduates: Well-prepared for select, design, and execute a successful final thesis project

Postgraduates: Deepen the understanding of data handling that sets a strong foundation for future study, research or profession work.

Importance of Data Handling

Data is the backbone of research: it forms the basis upon which we draw conclusions, make decisions, and contribute to knowledge. Effective data handling ensures that the insights we derive are accurate, reliable, and meaningful.

Data Quality: Ensuring that data is accurate, complete, and reliable

Reproducibility: Allow other researchers to replicate studies and verify results.

Efficiency: minimizing errors, reducing redundant efforts, and accelerating the research workflow.

Ethics: Adhering to data handling standards and regulations is essential for ethical research practices.

Key Components

Data Collection

Data Storage

Data Processing

Data Analysis

Data Interpretation and reporting

Data Management

Manage research or study data

Before you start

Be aware of your data management responsibilities before you start a project.

- Use the Research Data Manager
- Create a data management plan
- Manage research data for ethics approval (if needed)

During research

Create a system that enables you to collect, record and interpret research data, while keeping it secure.

- Manage research data
- Develop research data skills
- Store and secure your project's data

Project Management tools in UQ

UQ Research Data Manager (UQRDM)

The UQRDM allows you to organise, store, and share all your project's research or laboratory data, including working documents, images, observations, and data in any format.

Digital Research Notebooks (DRNs)

DRNs are a UQ wide digital solution that replaces paper-based research notebooks and lab notebooks.

Data Management Plan (DMP)

A data management plan (DMP) is a tool for maximising the efficiency and quality of your research and its data.

Research Data Manager (RDM)

UQRDM will:

•provide you with storage for your project's research data

•allow you to share and collaborate with colleagues

•enable you to plan and manage your project's research data.

•Publish your datasets to the institutional repository,

UQRDM can be used to:

•organise all the relevant data and documentation relating to your project

•ensure that you are compliant with institutional policies relating to data management

•set up a data management plan

Data Management Plan (DMP)

A DMP is a tool to better manage your own research data. Consider creating a plan at **the beginning of your research project** for the best results and use it in conjunction with the UQ Research Data Manager.

It is a document that can be used to:

- reduce the risk of data loss
- avoid ethics or privacy breaches
- ensure compliance with legislation, policies, grant requirements, and discipline best practice.

Before you start

Metadata

Metadata is data that provides information about other data. It helps to describe, explain, or contextualize the primary data, making it easier to understand, manage, and use. In essence, metadata is "data about data" and is crucial for organizing information efficiently

- discovery (title, keywords, project description)
- evaluation (methods, dates)
- re-use (information on variables, software or hardware required, access and reuse conditions)

What a Data Management Plan (DMP) Looks Like

Data Description

Metadata: What kinds of data will be collected or generated

Data Collection and Organization

Collection Methods: How data will be gathered (e.g., surveys, experiments).

Organization: How data will be structured and labelled to ensure consistency and ease of use.

Data Storage and Backup

Storage Locations: Where the data will be stored (e.g., local servers, cloud storage).

Backup Procedures: How data will be backed up and how often.

Data Security and Privacy

Access Control: Who will have access to the data and how access will be managed.

Confidentiality: Measures to protect sensitive or personal data.

Creating a comprehensive DMP helps ensure that data is handled systematically, supporting the integrity and usability of the data throughout its lifecycle.

Data Sharing and Accessibility	Data Preservation and Archiving	Roles and Responsibilities	Ethics and Compliance	Review and Update
• **Sharing Plan**: How and with whom data will be shared. • **Access**: How others will access the data	• **Long-Term Storage**: Plans for long-term storage	• **Team Members**: Who is responsible for various aspects of data management.	• **Regulatory Requirements**: Compliance with legal and ethical standards related to data handling.	• **Plan Review**: How often the DMP will be reviewed and updated.

Ethics and research data management

Ethics and data management are an integral part of good governance in research.

When applying for ethics approval, consider your responsibilities for research data according to:

- Australian Code for the Responsible Conduct of Research

Researchers should explain the processes used to protect research participants

If you will be **sharing or re-using** your project's data in the future, ensure you have appropriate consent and ethics approval.

Journal Activity (15 minutes)

Talk with the people beside you, in your study or research

Do you have any data management strategy before study?

- Data Management Strategy: How do you plan and organize data before starting a study? What are the key considerations?
- Data Storage: Where and how do you store your data? What tools or platforms do you use?
- Naming Conventions: Do you use naming conventions for your study? Why are they important? What is your naming convention structure?
- Backup Plans: What are your strategies for backing up your data? How often do you back it up?
- Remote Access: Can you access your data remotely if your primary device is unavailable? How do you manage remote access?

Data Collection and Acquisition

Overview

Data collection is the systematic process of gathering information to address research questions, validate hypotheses, or derive conclusions.

Importance

Accurate and reliable data are crucial for ensuring valid research outcomes

Objective

The goal of this section is to explore the various methods of data collection, address the challenges specific to research, and discuss best practices to optimize data quality.

Data Type

Study

- Class notes
- Reports
- Assignments
- Programming code

Research

Quantitative

Numerical data for statistical analysis

Qualitative

Descriptive data for understanding concepts and themes

Primary data

Data collected directly from original sources

Sensitive data

- Human involved
- Human health and personal data, including information about secret practices
- Data that may place vulnerable species at risk.

Standardization of Data Collection Processes (SOPs)

Standard Operating Procedures for data collection are detailed, written instructions to ensure that data collection is carried out consistently and accurately.

Define

Purpose: Clearly state the objectives of data collection.

Design

Design Data Collection Tools

- Instruments: Develop or select tools (e.g., surveys, measurement devices) that are validated and reliable.
- Formats: Ensure data collection instruments are in standardized formats

Develop

Develop Protocols

- Procedure Steps: Outline step-by-step instructions for how data should be collected.

Ensure

Ensure Consistency

- Data Collection Conditions: Standardize environmental conditions and procedures under which data is collected (e.g., time of day, setting).

Standardization of Data Collection Processes (SOPs)

Documentation

Recording: Detail how data should be recorded and stored.

Logs: Maintain logs of data collection activities, including dates, times, and person involved.

Quality Control

Checks: Implement quality control measures, such as regular checks or calibrations of instruments.

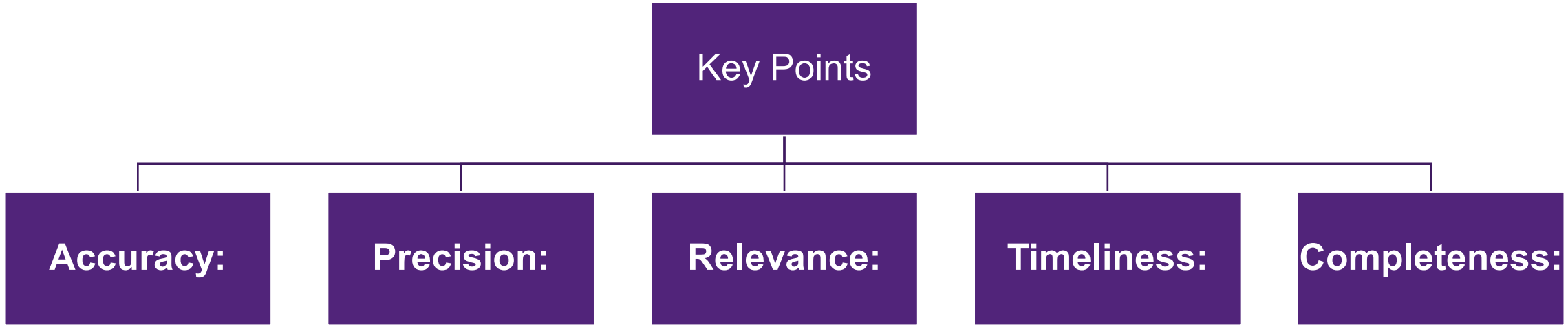
Error Handling: Develop protocols for identifying and addressing errors.

Ethical Considerations

Consent: Ensure procedures for obtaining informed consent are standardized and documented.

Confidentiality: Outline how to handle sensitive or personal information to ensure privacy and confidentiality.

Importance of Data Quality in Research



Data Collection Methods

- **Experimental Data:**
  - Data collected from controlled experiments.
- **Field Data:**
  - Data gathered from real-world environments.
- **Simulation Data:**
  - Data generated from computer simulations.
- **Sensor Data:**
  - Data collected using sensors.
- **Survey Data:**
  - Data obtained from surveys

Experimental Data Collection

- **Controlled Experiments:**
  - Conducting experiments in a controlled environment allows for precise manipulation of variables.

- **Laboratory-Based Data Collection:**
  - Labs are designed to minimize external influences and ensure the repeatability of experiments.
  - **Considerations:**
    - Calibration of equipment is essential to ensure accuracy.
    - Safety protocols must be followed to protect researchers and the integrity of the data.

Field Data Collection

- **Real-World Data Collection:**
  - Collecting data in the field provides real-world insights that laboratory conditions might not reveal.

- **Challenges:**
  - Environmental factors like weather can affect the accuracy and feasibility of data collection.
  - Requires robust planning and often the use of portable equipment.

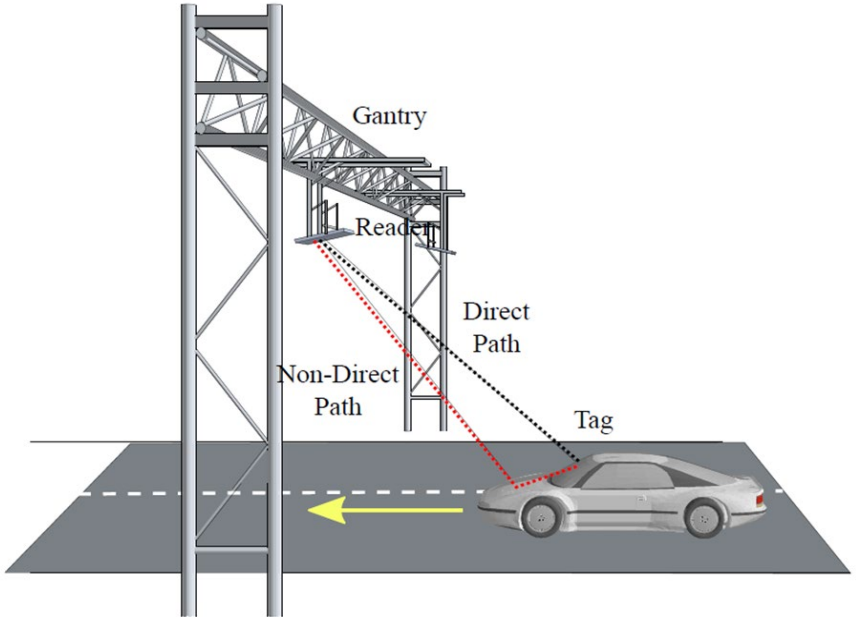
Simulation Data Collection

- **Computer Simulations:**
  - Simulations are used to model complex systems and predict behaviour under various conditions.

- **Advantages:**
  - Allows for testing scenarios that might be impractical or unsafe in real life.
  - Provides the ability to run multiple iterations quickly and analyse various outcomes.

Sensor Data Collection

- **Sensor Technology:**
  - Sensors convert physical phenomena into digital data, which can be used for monitoring and control purposes.
- **Data Integrity:**
  - Ensuring sensors are calibrated and maintained is crucial for accurate data collection.
  - Data from sensors often requires preprocessing to remove noise and correct errors.



Survey Data Collection

- **Surveys and Questionnaires:**
  - Surveys are useful for gathering subjective data, such as user feedback or expert opinions.
- **Design Considerations:**
  - Ensure the survey is well-structured with clear and unbiased questions.
  - Sampling must be representative of the population to ensure valid results.

Data Collection Best Practices

- **Planning and Preparation:**
  - Detailed planning ensures that all necessary data is collected accurately and efficiently.
  - **Steps:**
    1. Define objectives clearly.
    2. Choose the appropriate data collection methods.
    3. Ensure all tools and instruments are ready and calibrated.
- **Documentation:**
  - **Importance:** Keep detailed records of how data was collected, including environmental conditions, equipment settings, and any other important conditions.

Ethical Considerations in Research Involving Human Participants

Ethics in research is crucial to ensure the integrity of the research process and the protection of participants' rights.

When conducting research involving human participants, adherence to ethical standards is essential to uphold respect, privacy, and safety.

When the data **collection or share** involves Human Participants, it needs to be considered:

Ethical Approval

Ethical Approval Process

1.Preparing the Ethics Application:

- **Research Proposal**: Develop a detailed research proposal outlining the study's objectives, methodology, participant recruitment, and data handling procedures.
- **Informed Consent Documents**: Prepare clear and comprehensive informed consent forms that detail the study's purpose, procedures, risks, and benefits.

2. Submission to Human Research Ethics Committee (HREC):

- **Review Process**: Submit the research proposal and informed consent documents to a Human Research Ethics Committee (HREC) for review. The HREC will assess the study's ethical considerations and provide feedback or approval.
- **Ethics Approval**: Obtain formal ethics approval before commencing the research. Approval ensures that the study meets ethical standards and safeguards participants' rights.

Ethical Approval Process

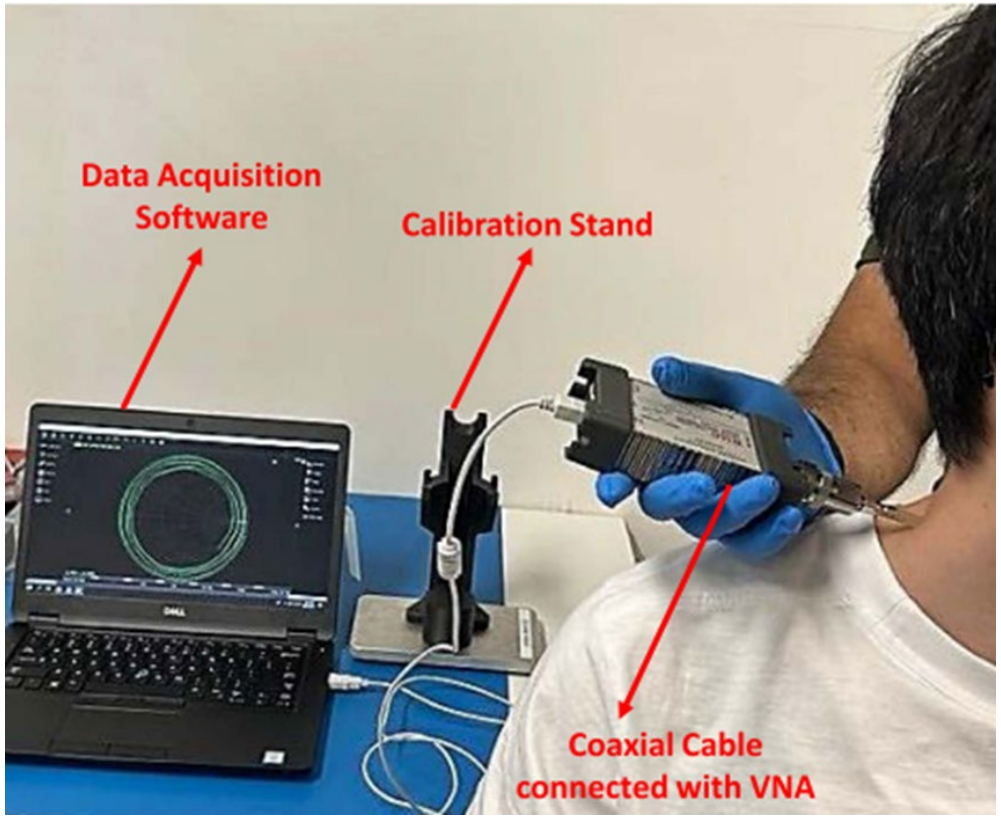
3. Conducting the Research:

- **Compliance**: Follow the approved research protocol and adhere to ethical guidelines throughout the study. Monitor and address any issues that arise during the research.
- **Ongoing Consent**: Maintain ongoing communication with participants, ensuring they are aware of any new information or changes in the study.

4. Reporting and Accountability:

- **Reporting to HREC**: Provide progress reports and final results to the HREC as required. Report any adverse events or ethical issues encountered during the study.
- **Publication and Dissemination**: Publish and share research findings transparently, ensuring that participants' confidentiality is maintained, and their contributions are acknowledged.

Case study



1. Looking for volunteers for skin data collection
2. Data is collected and stored in the experiment folder of his laptop
3. Data Types: Raw data will include measurements from S11 reflection data, Vector Network Analyzer (VNA) readings, and skin response properties.
4. Processing: Apply statistical filters and remove outliers to ensure data accuracy.
5. Statistical: Assess significant contrasts between different skin conditions, regions, and frequencies.
6. Frequency Analysis: Determine which frequencies offer the most significant contrast for skin detection and property analysis.
7. The results are stored in the safe personal computer.
8. He shared the data to his friend to help her if the data can be used for her business,

Journal Activity (15 minutes)

These are the procedures of research on Skin Cancer Detection

Based on today's content, please list the missing steps and incorrect procedures, and how to correct it?

Practical Examples of Data Collection From Dr. Syed from his Skin Cancer Project

This project involves human participants and must adhere to ethical standards as outlined by the relevant ethics committee.

1. **Make a Data Management Plan before the project start.**
2. **Ethical Considerations and Approval before data collection**
3. **Data Management:**
   - Data Storage: All collected data will be securely stored on the research drive, which is managed in accordance with data protection regulations.
   - Data Types: Raw data will include measurements from S11 data, reflection data, Vector Network Analyzer (VNA) readings, and skin response properties.
4. **Data Analysis:**
   - Methods: Apply statistical filters and remove outliers to ensure data accuracy.
5. **Analysis Focus:**
   - Statistical : Assess significant contrasts between different skin conditions, regions, and frequencies.
   - Frequency Analysis: Determine which frequencies offer the most significant contrast for skin detection and property analysis.
6. **Store all the data and documentation in the UQ Research Data Manager**
7. **Reporting to HREC**

Data Storage and Management

Overview

Effective data storage and management are crucial for organizing, retrieving, and safeguarding data throughout the research process

Importance

Proper storage and management ensure data integrity, accessibility, and security, which are essential for reproducibility and long-term research success.

Objective

The goal of this section is to explore the various methods of data storage and management methods

Types of Data Storage Solutions

Local Storage:

- Storing data on physical devices such as hard drives or SSDs.
  - **Advantages:** Immediate access, no dependency on internet connectivity.
  - **Disadvantages:** Limited capacity, risk of data loss if hardware fails.

Network Storage:

- Data is stored on a network-attached storage (NAS) or storage area network (SAN).
  - **Advantages:** Centralized access, better management and backup options.
  - **Disadvantages:** Requires network infrastructure, potential for bottlenecks.

Cloud Storage:

- Data is stored on remote servers accessed via the internet.
  - **Advantages:** Scalability, remote access, integrated backup and recovery options.
  - **Disadvantages:** Security concerns, ongoing cost.

Data Management Strategies

Organizing Data

- Structuring data in a logical manner

Data Backup and Recovery

- Implementing regular backups and establishing recovery procedures to protect against data loss.

Data Access Control:

- Setting permissions and access levels to ensure only authorized person can access sensitive data.

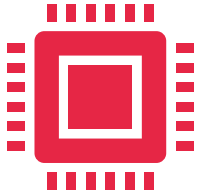
Manage my code and data platform I use

Project management

Research Data Manager

Code management

GitHub



Document management

UQ Research Drive(Server)

NAS (Server)

OneDrive (Cloud)

laptop (Local)

Introduction of GitHub

What is GitHub?

- Web-based platform for code management and collaboration

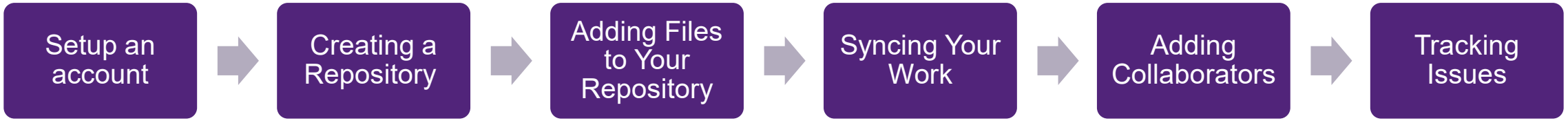
Why GitHub Matters?

- Significance of collaboration in software development
- Importance of version control and organized code management

For ENGG7811

- Recommend for both study and research
- Manage, track, and develop your code both local and online

Demonstration of GitHub



Journal Activity (20 Minutes)

Talk with the people beside you,

- Named the key features of GitHub, and which feature is best suit for your study, why ?
- Setup a GitHub account, and create a repository for ENGG 7811
- Show you are able to demonstrate the GitHub synchronization

Data Analysis and Modelling

Overview

Data analysis involves examining data to extract meaningful insights, while modelling involves creating representations to understand or predict phenomena.

Importance

Both are crucial for interpreting data, making informed decisions, and developing engineering solutions.

Objective:

This section will cover various techniques and approaches for data analysis and modelling.

Standardization of Analysis Pipelines

The sequence of steps or workflows used to process and analyse data. Standardizing these pipelines ensures consistency and reproducibility in data analysis.

Define

Define Analysis Objectives

- Goals: Clearly outline the objectives of the data analysis.
- Requirements: Identify the specific requirements or hypotheses to be tested.

Develop

Develop Analysis Protocols

- Workflow: Create a standardized workflow for data processing and analysis, including data cleaning, transformation, and statistical analysis.
- Tools and Software: Specify the tools, software, and versions to be used, and ensure they are consistently applied.

Implement

Implement Data Cleaning Procedures

- Cleaning Steps: Define procedures for handling missing data, outliers, and inconsistencies.
- Documentation: Document the cleaning steps and criteria used.

Establish

Establish Analytical Methods

- Techniques: Standardize the statistical methods or algorithms used for analysis.
- Parameters: Define parameters and settings for analytical methods.

Standardization of Analysis Pipelines

Automate and Document

Automation: Where possible, automate repetitive tasks in the analysis pipeline to reduce errors and improve efficiency.

Documentation: Maintain detailed records of the analysis process, including code, settings, and decisions made.

Validation and Testing

Verification: Regularly verify and validate the analysis pipeline to ensure accuracy and reliability.

Reproducibility: Test the pipeline to confirm that it produces consistent results with different data sets.

Review and Update

Periodic Reviews: Regularly review and update the analysis pipeline to incorporate new techniques or address identified issues.

Feedback Mechanism: Implement a system for incorporating feedback and improving the pipeline based on experience and new developments.

Reporting and Interpretation

Standard Reports: Develop standardized formats for reporting results.

Interpretation Guidelines: Provide guidelines for interpreting and presenting findings consistently.

Analysis

- Summarizes data using statistics such as mean and standard deviation.
- Uses visualization and statistical techniques to explore data patterns and relationships.
- Draws conclusions about a population based on sample data using hypothesis testing and confidence intervals.

Modelling

A fundamental process in research and data analysis that involves creating mathematical, statistical, or computational representations of real-world systems or phenomena. These models help researchers understand, predict, and make decisions based on data. The goal is to capture essential features and relationships within the data to make predictions or gain insights.

Model Formulation:
Defining the structure and relationships within the model based on theoretical understanding

Data Collection:
Gathering data that will be used to fit the model and validate its accuracy.

Model Training:
Using algorithms to estimate model parameters from the data.

Model Evaluation:
Assessing the model's performance using metrics and validation techniques.

Prediction and Analysis:
Using the model to make predictions or analyse new data.

Types of Models

Statistical Models	• **Purpose**: To identify relationships between variables and make predictions based on statistical methods.
Machine Learning Models	• **Purpose**: To learn patterns from data and make predictions or decisions without explicit programming.
Mathematical Models	• **Purpose**: To represent systems and processes using mathematical equations.
Simulation Models	• **Purpose**: To mimic the behavior of real-world systems over time to understand their dynamics and outcomes.

Model Validation and Evaluation

Ensuring Model Accuracy

Learning from data

- **Cross-Validation:**
  - Splits data into training and testing sets to evaluate model performance.

- **Performance Metrics:**
  - Measures model accuracy using metrics such as RMSE, MAE, and F1-score.

Model Validation and Evaluation

Cross-Validation

- **Model Evaluation**: To assess how well a model generalizes to an independent dataset that it hasn't been trained on.
- Helps in balancing the model's ability to generalize versus its performance on training data.

Performance Metrics

- **Model Assessment**: To quantify how well a model performs and how accurately it makes predictions.
- **Comparison**: To compare the performance of different models or algorithms.

Case Studies in Modelling

Real-World Applications and Case Studies from Dr Awal from his Heart rate analysis using Machine Learning

Electrocardiogram (Ecg data), from MIT-BIH database

- SEGMENT data
- label the data
- feature extraction and feature selection
- training the data

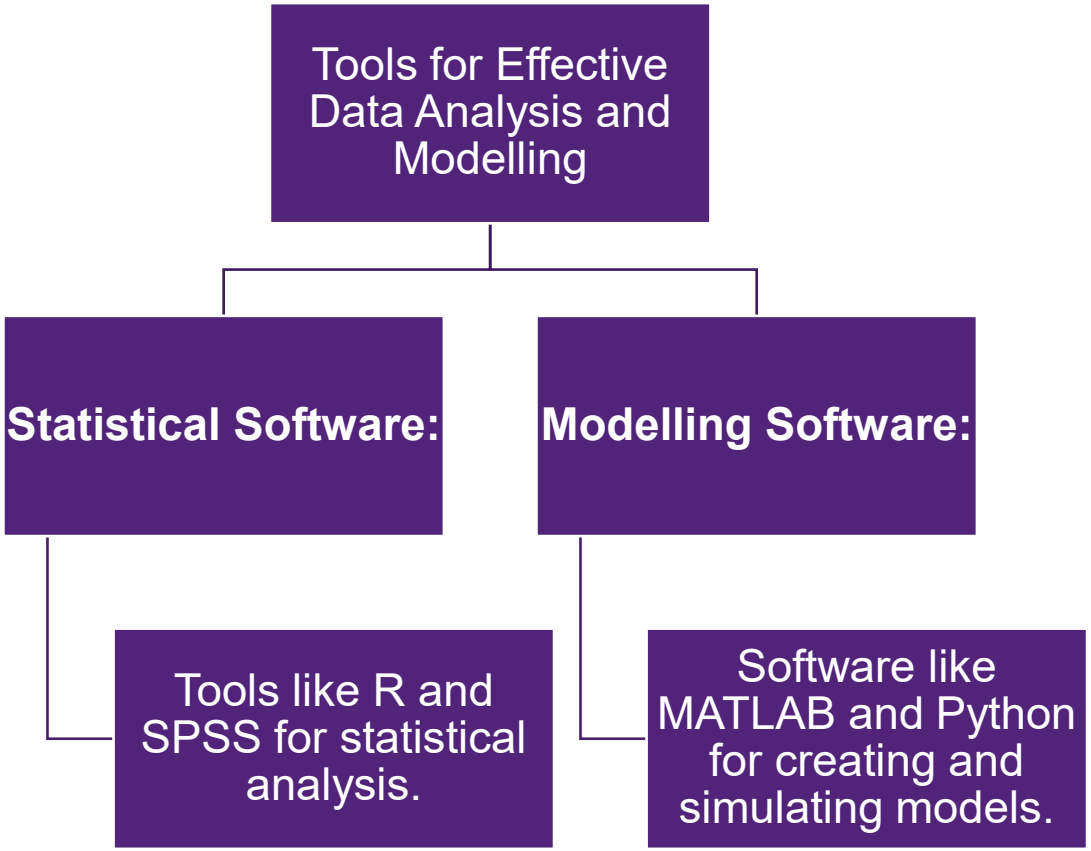
machine learning or deep learning

Python & MATLAB

Figure: ECG from MIT-BIH

- Make model, and evaluate the model, verify the model
- Accuracy and sensitivity all how well your model fitting the problem.

Data Analysis Tools and Technologies



Data Interpretation and Reporting

Data interpretation involves deriving meaning from analysed data, while reporting findings

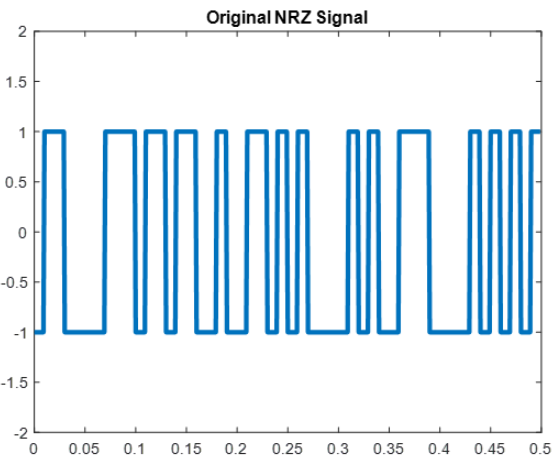
Importance:

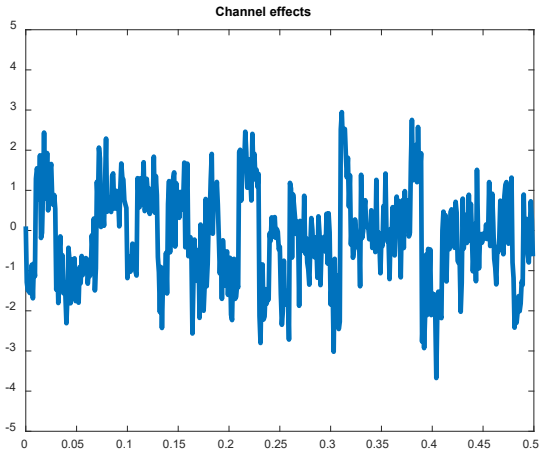
Effective interpretation and reporting are crucial for making informed decisions and ensuring that research findings are understood

Data Visualization for Interpretation

Charts and Graphs:

- Using bar charts, line graphs, and scatter plots to illustrate data trends and relationships.





Heatmaps and Contour Plots:

Visualizing data density and patterns using heatmaps and contour plots.