Data Science Seminar

Lesson 2

- Update of MSc topics
- Guidelines for Master's Dissertations at ISA
- Structure of a Data Science dissertation/report
- Best practices of formal and informal planning

U LISBOA | UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR DE AGRONOMIA

ISA

- *Update of MSc themes*
- *Guidelines for Master's Dissertations at ISA*
- *Structure of a Data Science dissertation/report*
- *Best practices of formal and informal planning*

- ***Update of MSc themes***
- *Guidelines for Master's Dissertations at ISA*
- *Structure of a Data dissertation/report*
- *Best practices of formal and informal planning*

Thesis themes and supervisors

Name	Topic	Status
Alícia Gouveia	Applications of Data Science on Biodiversity: GBIF and invasive species	±
Ana Moreira	Visualization tool for residuals and recycling data / Sustainable Water Use in Agriculture with IoT and Data Analytics (both with Sonae)	±
Damião de Goes		±
Diogo Simão	Topic 1?	✗
Dominic Welsh	Continuous change detection algorithm for Portugal land use changes	✓
Emmanuel Rivera	Remote detection to monitor rice growth in mangrove swamp rice crops	✓
Inês Schwartz	Database development with applications on soil microbiology	✓
Maria Dolgaya	Internship task: Mapping material stocks in buildings and modelling construction waste flows for a circular built environment	✓
Mariana Coelho	Predicting adaptations of orchids to climate change	✓
Maria Navalho	Optimization of products inventory management	±
Miguel Ferreira		✗
Rafael Rodrigues	Digital technologies in olive farming	✓
Rubén Torrado		✗
Sofia Rodrigues	Topic 1 or 7?	±

✓ Defined

± Under progress

✗ Not defined

- *Update of MSc themes*
- ***Guidelines for Master's Dissertations at ISA***
- *Structure of a Data Science dissertation/report*
- *Best practices of formal and informal planning*

Guidelines for Master's Dissertations at ISA

From “Normas-para-a-elaboracao-da-dissertacao-de-mestrado_NOVO.pdf”

a) Cannot exceed 80 main pages, A4 format, font Arial (or similar) size 10 or 11, single-spaced, with 2.5cm margins. Additional supplementary documentation may be added in the form of annexes with no more that 120 pages in total;

b) It must include summaries in Portuguese and another official language of the European Union of up to 300 words each, up to 5 key words in Portuguese and another official language of the European Union, and indexes;

c) When the work is written in a foreign language, it must be accompanied by a more developed summary in Portuguese, with a length between 1,200 and 1,500 words;

d) The cover must include the name of the University of Lisbon and the Instituto Superior de Agronomy, with their respective logos, the title of the work, the name of the student, the name of the supervisors, the name of the master's programme and, if applicable, the area of specialisation, the type of work being presented (dissertation, project work or internship report), the year of completion and, in the case of degrees awarded in associations, the awarded in association, the identification of the partner institutions.

Guidelines for Master's Dissertations at ISA

The general structure of the dissertation/report should be as follows:

a) Cover

b) Acknowledgements (optional);

c) Abstracts and keywords (two languages);

d) Table of contents;

e) List of tables, figures and abbreviations;

f) Main text (not to exceed 80 main pages);

g) Bibliographical references;

h) Appendices (optional).

- *Update of MSc themes*
- *Guidelines for Master's Dissertations at ISA*
- ***Structure of a Data Science dissertation/report***
- *Best practices of formal and informal planning*

Structure of a Data Science report

Two broad types of final MSc reports	**Types of master thesis (ISA's regulation)**
1. Professional activity reports	• Internship report
2. Scientific technical reports	• Dissertation • Project work

Structure of a Data Science report

Scientific Technical report

- It is the first presentation of an original result, previously untested, or an original thematic synthesis;
- It can be repeated experimentally by other researchers;
- It can be a topic of immediate application or provide purely academic outputs;
- It is in a form that is ± accessible to the scientific and technical community (journal, book, database or application);
- It is subject to public evaluation by a specialized jury
- When published, it is screened by professionals in the respective scientific field (referees, editorial boards).

Structure of a Data Science report

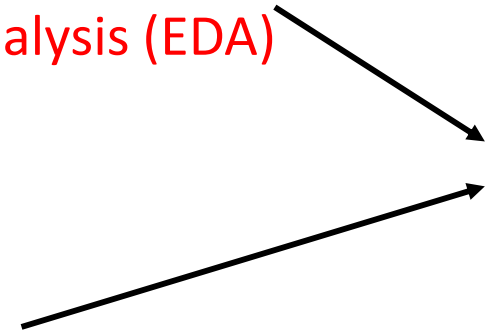
ChatGPT: “What should be the structure of a Data Science report?”

1. Title Page
2. Table of Contents
3. Executive Summary
4. Introduction
5. Data Collection and Description
6. Exploratory Data Analysis (EDA)
7. Methodology
8. Results
9. Discussion
10. Recommendations
11. Conclusion
12. References
13. Appendices

Structure of a Data Science report

ChatGPT: “What should be the structure of a Data Science report?”

1. Title Page
2. Table of Contents
3. Executive Summary
4. Introduction
5. Data Collection and Description
6. Exploratory Data Analysis (EDA)
7. Methodology
8. Results
9. Discussion
10. Recommendations
11. Conclusion
12. References
13. Appendices



Main difference from a typical scientific technical report.

Structure of a Data Science report

ChatGPT: “What should be the structure of a Data Science report?”

1. Title Page
2. Table of Contents
3. Executive Summary
4. Introduction
5. Data Collection and Description → Should be moved to Methodology
6. Exploratory Data Analysis (EDA) → Should be moved to Results
7. Methodology
8. Results
9. Discussion
10. Recommendations → Should be moved to the end of the Discussion
11. Conclusion
12. References
13. Appendices

Structure of a Data Science report

Summary

- A **concise** summary of your work.
- After reading it, even a non-technical stakeholder, will understand the **context** and **relevance** of the work, the overall **approach** to deal with the problem, the **key findings** and **recommendations**
- Also important to **clear up ideas** about the work in an early writing stage.

Structure of a Data Science report

Summary

concise

context **relevance** **approach**

key findings **recommendations**

clear up ideas

Structure of a Data Science report

Introduction

- Describes the **state of the art** on the topic, defines the **problems** or **questions** to be addressed and the **relevance** of the work;
- It helps to **consolidate ideas and knowledge** about the thesis topic;
- Contributes to disseminate **established knowledge** and ideas on that topic, including **strengths** and **weaknesses**;
- Compile and examine the **current state of knowledge** on a given topic based on previous studies that has already been recognised;
- To **contextualise the research** within the knowledge and work carried out on the topic;
- Identifies **knowledge gaps** that need to be filled.
- Allows the **identification of the topics** to be investigated;

Structure of a Data Science report

Introduction

state of the art

problems

questions

relevance

consolidate ideas and knowledge

established knowledge

strengths

weaknesses

current state of knowledge

contextualise the research

knowledge gaps

identification of the topics

Structure of a Data Science report

Methodology

Data Collection and Description

- Description of the **data sources** and **data collection** methods.
- Information about the **dataset(s) characteristics**.
- Data preprocessing steps, including **cleaning**, handling **missing data**, and **data transformation**.

Data analysis

- Detailed explanation of the **analytical techniques**, algorithms and models used.
- **Justification** for the chosen methods.

Structure of a Data Science report

Methodology

data sources data collection

dataset(s) characteristics.

cleaning missing data data transformation.

analytical techniques

Justification

Structure of a Data Science report

Results

Exploratory Data Analysis (can be part of the Results)

- Preliminary understanding of the data through **visualization** and **statistical summaries**

Main outputs

- Presentation of the **main findings** and insights.
- Use **visuals** (charts, graphs, tables) that support your points.
- Include any statistical analysis or machine learning model **performance metrics**.

- Important: this section is **not intended to interpret** results

Structure of a Data Science report

Results

visualization statistical

summaries

main findings

visuals

performance metrics.

not intended to interpret

Structure of a Data Science report

Discussion

- Provide an **interpretation** of the results in the context of the problem.
- Address any **unexpected findings or challenges** encountered.
- Discuss the **implications** of the results and their **relevance** to the project objectives.

Recommendations

- Propose and prioritize **actions or decisions** based on the analysis.
- Justify your recommendations with data-driven insights.

Conclusion (can also be part of the discussion)

- Summarize the **key points** of the report.
- Emphasize the main **take-home messages** and their significance.

Structure of a Data Science report

Discussion

interpretation

unexpected findings or challenges

implications

relevance

actions or decisions

key points

take-home messages

Structure of a Data Science report

References

- Cite any external sources, **books**, research **papers** or **datasets** used.
- Follow a **consistent citation style** (e.g., APA, MLA, or a style relevant to your field).

Appendices

- Supplementary information such as **code**, **additional charts**, or too **detailed explanations**.
- It ensures that the main report remains concise and accessible, moving technical details to the appendices.

Structure of a Data Science report

References

books papers datasets

consistent citation style

Appendices

code additional charts detailed explanations.

Structure of a Data Science report

Writing best practices

- Clear, short and concise sentences, without jargon, unnecessary details or redundancies
- Use active voice (not consensual)
- Use transitions and connectors (e.g. additionally, also, moreover, …) to link your sentences and paragraphs, helping the text to flow better.
- Use headings and subheadings to organize the text into sections and subsections

Structure of a Data Science report

Writing best practices (cont.)

- Final review for any errors, inconsistencies or gaps in content, language or format. Use review tools:
  - Spelling, grammar, punctuation or syntax errors – Grammarly; Hemingway; ProWritingAid
  - Content, structure or logic problems - CoSchedule Headline Analyzer; Readable; Yoast SEO

Structure of a Data Science report

Writing the introductory chapter: objectives

The purpose of an introduction (Cícero, 55 B.C.) should be to:

- "*Attract the hearer or reader straight away*" - **advertising**;
- "*State the whole of the matter that is to be put forward*" – **summary**
- "*Approach to the case and a preparation of the ground*" – **context-setting**.

Structure of a Data Science report

Writing the introductory chapter: recommendations

- The introduction is the entrance hall of your work: it has to impress your guests!
- Should be a continuous and organic document: avoid waiting for the deadline to start working on it
- Don't assume it's closed after you've started writing another chapter
- Don't wait until it's finished before moving on to other activities or writing new chapters
- Make sure you have included all relevant and recent sources in the field;
- Be careful when selecting sources of information: give preference to scientific articles that are peer-reviewed; prioritize taking into account the impact factor of journals, for example.

Structure of a Data Science report

Writing the introductory chapter: recommendations

Exercise

https://www.menti.com/alhodv9aagge

Exercise: order the following paper from EPJ Data Science journal

A. Here we investigate the performance of 538 students within a novel dataset collected as part of the Copenhagen Network Study (CNS), with data collection ongoing for more than two years [12]. Due to the scale of the CNS, and the inclusion of directly observed data from smartphones in place of self-reports, we are able to mitigate some of the limitations encountered in existing 'traditional' studies. The strength of the CNS data is the high-resolution multi-channel measures for social interactions, including person-to-person proximity (using Bluetooth scans), calls and text messages, activity on online social networks (Facebook), and mobility traces.

B. The aim of our study was to better understand the impact of individual and network factors on our ability to distinguish between groups of students based on their performance. That is, we wanted to identify the ways in which low performers are significantly different from high performers and vice versa. We divide this goal into three specific objectives: (i) Identify individual and network factors that correlate with students' performances; (ii) Analyze the importance of different sets of features for supervised learning models to classify students as low, moderate, or high performers; (iii) Investigate significant differences among performance groups for the most important individual and network features.

C. Since research on academic achievement began to emerge as a field in the 1960s, it has guided educational policies on admissions and dropout prevention [1]. Although much of the literature has focused on higher education, the knowledge obtained on behavioral phenomena observed in colleges and universities can potentially guide research on student behavior in primary and secondary schools. A number of behavioral patterns have been linked to academic performance, such as time allocation [2], active social ties [3], sleep duration and sleep quality [4], or participation in sport activity [5]. Most of the existing studies, however, suffer from biases and limitations often associated with surveys and self-reports [6, 7], particularly when measuring social networks [8–11].

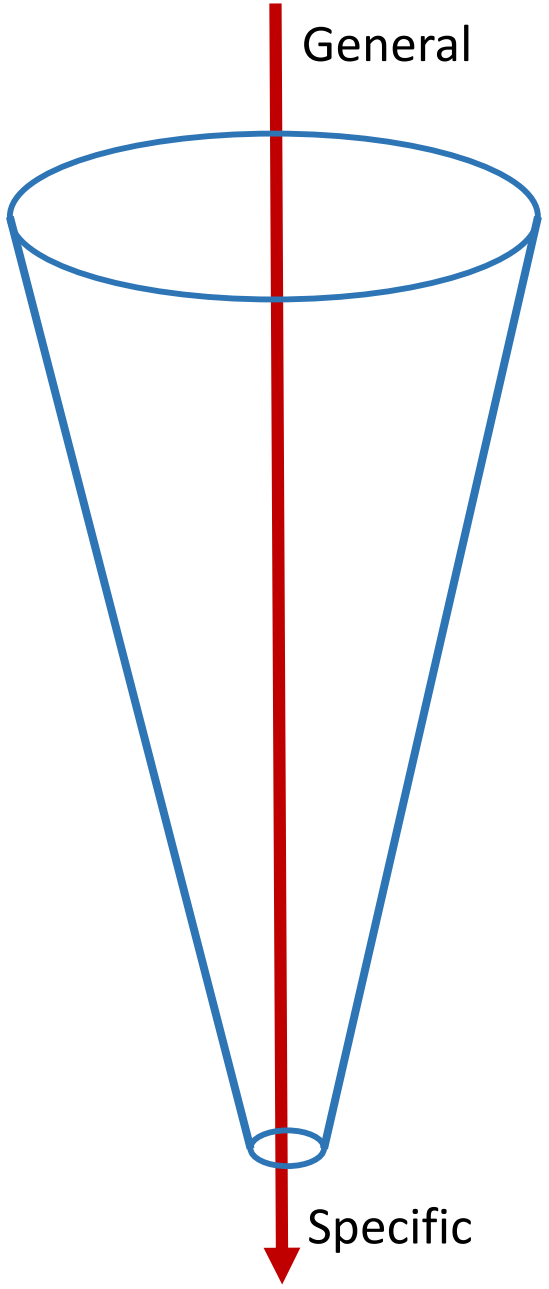
Structure of a Data Science report

Writing the introductory chapter: structure

Part 1: Define a research territory (Context). Start with sentences that define the broadest possible context for the subject of the study to be carried out (captivating the largest number of potential readers). Then focus the text on more specific areas.

Part 2: Establish a niche within the research area. Identify a concrete problem in which there are gaps in knowledge or alternative theoretical models. End with the central question that will be investigated.

Part 3: Filling the niche (how you will fill the information gap). Show how the work will fill the niche identified. Describe the approach that will be adopted to answer the central question and show how the answer helps to solve the open problem that has been identified. How the data and analysis can answer the central question being investigated.



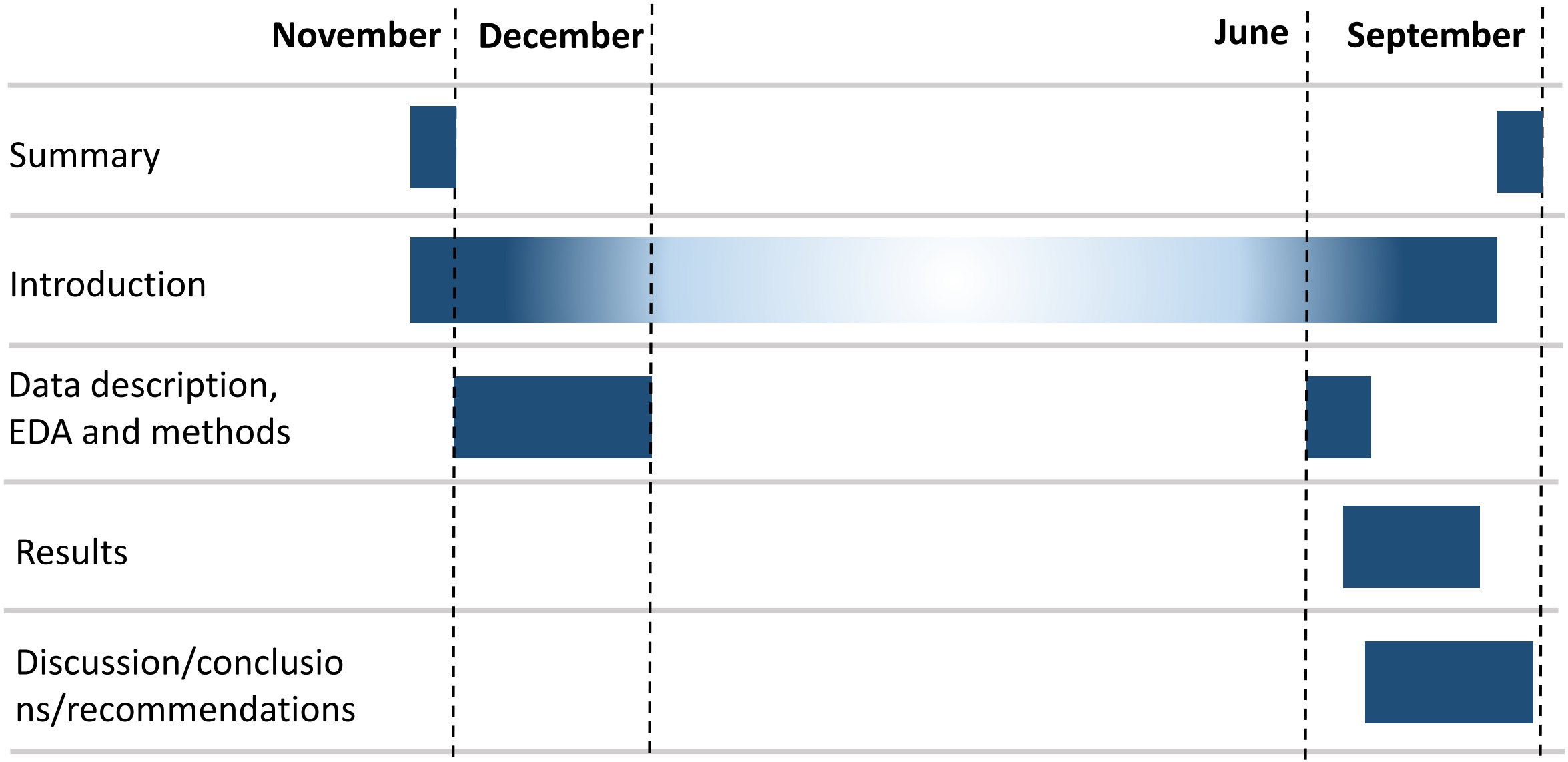
* Swales JM (1990) Genre analysis: English in academic and research settings. Cambridge University Press, Cambridge, UK

Structure of a Data Science report

Writing the introductory chapter: structure

General structure of an introduction	Citation/Doominant types
1. Introduce the general theme of the work; optionally include a sentence at the end that summarizes the work you intend to do. Usually 1 paragraph.	Compulsory; high impact books or review articles
2. Theoretical framework more directly related to the work, including identification of information gaps - basis for subsequent discussion of the results. Usually more than 1 paragraph	Compulsory; publications most directly related to the topic being studied.
3. Define the relevance of the issue to be analyzed, taking into account the information gaps. Usually 1 paragraph.	Not compulsory.
4. Clarify and define the focus of the work, questions to be addressed, problems and/or hypotheses. 1-2 paragraphs.	Not compulsory.
5. Justify the relevance or importance of the problem you have chosen to focus on - e.g. applications. 1 paragraph.	Not compulsory.

Recommended writing calendar



Guidelines for Master's Dissertations at ISA

- *Update of MSc themes*
- *Guidelines for Master's Dissertations at ISA*
- *Structure of a Data Science dissertation/report*
- ***Best practices of formal and informal planning***

Best practices of formal and informal planning

Formal planning – A plan to be submitted in a call or requested by the supervisor or institution/company.

Informal planning – A more realistic plan that is intended to be followed more closely.

Best practices of formal and informal planning

Formal planning

ChatGPT

"Write me the sections for a formal planning of a data science project"

1. Title Page
2. Executive Summary
3. Project Introduction
4. Problem Statement
5. Objectives and Deliverables
6. Stakeholder Analysis
7. Project Scope
8. Methodology
9. Project Timeline
10. Resource Allocation
11. Data Management
12. Quality Assurance
13. Communication Plan
14. Evaluation and Success Criteria
15. Conclusion
16. Appendices

Best practices of formal and informal planning

1. Title Page
2. Executive Summary
3. Project Introduction
4. Problem Statement
5. Objectives and Deliverables
6. Stakeholder Analysis
7. Project Scope
8. Methodology
9. Project Timeline
10. Resource Allocation
11. Data Management
12. Quality Assurance
13. Communication Plan
14. Evaluation and Success Criteria
15. Conclusion
16. Appendices

Recommended:

1. Title Page
2. Summary
3. Introduction
5. Objectives
8. General methodological approach
9. Tasks
10. Project Timeline
12. Data Management
13. Quality Assurance
14. Communication Plan

Best practices of formal and informal planning

Informal planning

It is often advantageous to start the planning by defining the end goals and working backward to develop a roadmap for the project: **Reverse or Backward Planning!**

Best practices of formal and informal planning

Reverse planning

Main steps:

1. Define your end goal.
2. What are the results needed?
3. Which data is needed to accomplish these results?
4. What protocol is needed to obtain this data?
5. Schedule the tasks taking into account the time available
6. Start work and deal with reality!
7. Update your planning whenever needed

Best practices of formal and informal planning

Watch these videos

https://www.youtube.com/watch?v=7vQ9zxT6uhs

https://www.youtube.com/watch?v=wHRqO61-myY

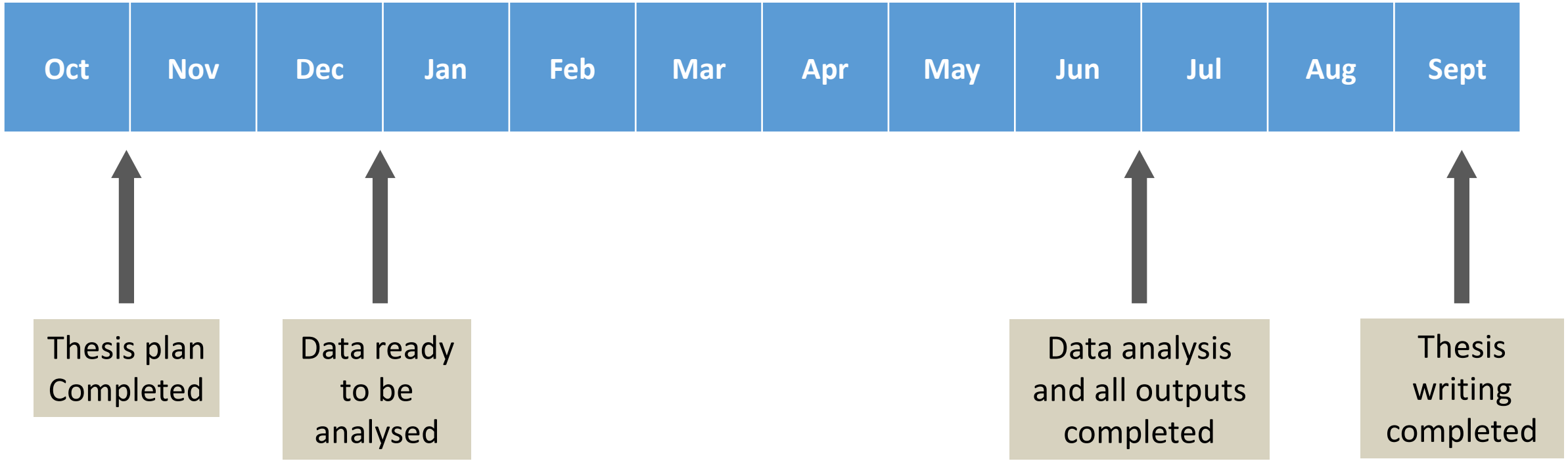
Best practices of formal and informal planning

Home exercises: try asking these questions to ChatGPT (or other)

- Write me the sections for a formal planning of a data science project
- Write me the structure of an informal plan for a data science project
- What is the best strategy to plan a data science project?
- Are there benefits in a reverse planning for a data science project?

Best practices of formal and informal planning

Recommended general timeline of your MSc Thesis work



Best practices of formal and informal planning

Home work: write a summary of your thesis plan

- 300 words + up to 5 key words
- To be delivered on 15 nov.

This implies:

- First literature search
- Define a first working title
- Include a short sentence with the expected results
- Write keywords

Best practices of formal and informal planning

Example

Disentangle the effects of environment and disturbance on landscape dynamics using LANDIS forest landscape model (https://www.sciencedirect.com/science/article/abs/pii/S1364815222002134)

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model. We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery. Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010). This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.

Keywords

Changbai mountain; Environment; Historical landscape reconstruction; LANDIS PRO; Typhoon disturbance; Post-volcanic-eruption forest landscape recovery

Best practices of formal and informal planning

Abstract example 1

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model. We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery. Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010). This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.

Introduction/State-of-art:

- What is the relevance of the topic?
- What is the current knowledge, gaps or limitations?

Best practices of formal and informal planning

Abstract example 1

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. **Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model.** We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery. Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010). This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.

Aim of the study

Best practices of formal and informal planning

Abstract example 1

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model. **We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery.** Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010). This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.

Methods

Best practices of formal and informal planning

Abstract example 1

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model. We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery. **Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010).** This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.

Results and discussion

Best practices of formal and informal planning

Abstract example 1

Abstract

Forest landscapes pattern and development are affected by environment and disturbance. Disentangling their effects is important to understanding current landscape and predicting future changes. Such studies are limited by short-term observation and sparse disturbance-history data. Spatially-explicit forest landscape modeling represents a solution to these limitations. Here, we reconstructed the 300-year-time-series (1710–2010) of post-volcanic-eruption forest landscapes experiencing periodic-typhoons in Changbai Mountain, China, using LANDIS forest landscape model. We used a factorial simulation design to quantify the main and interactive effects of environment and typhoon on forest landscape recovery. Results showed environment had dominant effects (>80%) on early recovery (1710–1760), suggesting early forest development follows deterministic community-assembly processes governed by environment. However, as forest matured, disturbance became dominant (>50%) at later-recovery stages (1860–2010). **This study showed that historical landscape reconstruction reveals the full spectrum of interplays of environment, disturbance, and succession in forest ecosystems, which may not be captured by short-term studies.**

Conclusions