How to Write an Effective Capstone Project Proposal

ST5188 Advanced Data Science Project
2024/2025 Semester 2

ST5188 @ Canvas: https://canvas.nus.edu.sg/courses/72737

1

Reminder: Lecture Attendance & Participation

- ~~**For In-person and Zoom Live Stream Attendees**~~
  - ~~You are expected to complete all activities during lecture time.~~
    - ~~Activities: *Scan the QR codes* located in the top right corner of the respective slides.~~
  - ~~Your attendance will be tracked automatically.~~

- **For Recording Viewers (via Canvas' Panopto)**
  - You must watch the lecture from start to finish (→ tracking is done per second).
    - Activities: *Use the direct activity links* shared via Canvas Announcements and complete all activities by the end of Monday following the original lecture date.
      - Please note that QR codes do not work in lecture recordings at this time.

2

Content Overview

- Reminders & updates
- Project proposal guidance & expectations (revisited)
- **Key terms explained**
- **Good practises versus bad practises**
- Q & A

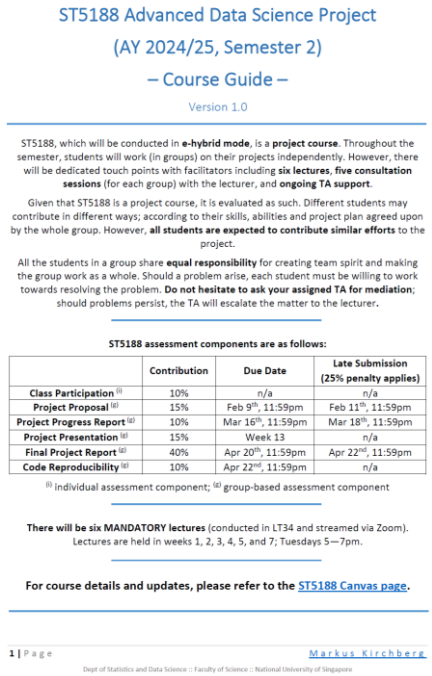
3

Reminders

- ~~Jan 22nd: Consultation bookings available~~
- Jan 27th: Third lecture (available as recording only due to CNY)
- Jan 27th: Consultation sessions commence
- Jan 28th, noon: Office hour (see Canvas Announcements)
- Jan 31st, noon: Book first consultation session (or we will schedule it for you)
- Feb 4th, 5pm: Fourth Lecture (@ LT34 / Zoom)
- Feb 9th: Project Proposal submission deadline
- Feb 11th, 5pm: Fifth Lecture (@ LT34 / Zoom)
- Mar 4th, 5pm: Sixth Lecture (@ LT34 / Zoom)

5

ST5188 Course Guide
(revisited)



ST5188 Advanced Data Science Project
(AY 2024/25, Semester 2)
– Course Guide –
Version 1.0

ST5188, which will be conducted in **e-hybrid mode**, is a **project course**. Throughout the semester, students will work (in groups) on their projects independently. However, there will be dedicated touch points with facilitators including **six lectures**, **five consultation sessions** (for each group) with the lecturer, and **ongoing TA support**.

Given that ST5188 is a project course, it is evaluated as such. Different students may contribute in different ways; according to their skills, abilities and project plan agreed upon by the whole group. However, **all students are expected to contribute similar efforts** to the project.

All the students in a group share **equal responsibility** for creating team spirit and making the group work as a whole. Should a problem arise, each student must be willing to work towards resolving the problem. **Do not hesitate to ask your assigned TA for mediation**; should problems persist, the TA will escalate the matter to the lecturer.

ST5188 assessment components are as follows:

	Contribution	Due Date	Late Submission (25% penalty applies)
Class Participation [i]	10%	n/a	n/a
Project Proposal [g]	15%	Feb 9th, 11:59pm	Feb 11th, 11:59pm
Project Progress Report [g]	10%	Mar 16th, 11:59pm	Mar 18th, 11:59pm
Project Presentation [g]	15%	Week 13	n/a
Final Project Report [g]	40%	Apr 20th, 11:59pm	Apr 22nd, 11:59pm
Code Reproducibility [g]	10%	Apr 22nd, 11:59pm	n/a

[i] individual assessment component; [g] group-based assessment component

There will be six MANDATORY lectures (conducted in LT34 and streamed via Zoom). Lectures are held in weeks 1, 2, 3, 4, 5, and 7; Tuesdays 5—7pm.

For course details and updates, please refer to the ST5188 Canvas page.

1 | Page Markus Kirchberg 6
Dept of Statistics and Data Science :: Faculty of Science :: National University of Singapore

6

ST5188 Lectures

- ~~Jan 14th, 5pm: ***ST5188 Introductory Briefing Session***~~
- ~~Jan 21st, 5pm: **Topic Selection and Foundations of Literature Review**~~
- Jan 27th: ***How to Write an Effective Capstone Project Proposal***
  - **Note, due to CNY, this lecture will be made available as recording only.**
- Feb 4th, 5pm: ***Project Planning, Management and Execution in Data Science***
  - Data Science in Practise
  - Agile Methodologies
  - Project Management Tools for ST5188
- Feb 11th, 5pm: ***Data Science Best Practises Part 1 (Data)***
- Mar 4th, 5pm: ***Data Science Best Practises Part 2 (Model Development and Evaluation)***

All lectures are 90mins each & conducted in LT34 / streamed via Zoom;
recordings will be made available within 2 business days.

7

7

Consultation Sessions

- Available via Canvas → Calendar
  - **Mandatory sessions**: Weeks 3 – 4, weeks 7 – 8, and weeks 11 – 12.
  - **Optional sessions**: Weeks 5 – 6 and weeks 9 – 10.

[booking your 1st consultation session]

- MUST be scheduled and attended BEFORE project proposal submission.
- Book your first consultation slot early → first come, first served.
- **For groups that have not made a booking by Jan 31st noon, we will schedule an online consultation session for your group!**
  - → However, you will then have no choice wrt. the meeting date and time.
- Any group that fails to attend their first consultation session prior to project proposal submission will only receive their project proposal feedback once all other feedback has been returned.

8

8

[prior to your consultation session]

- **Send a summary** (limited to half a page / two paragraphs) **of your project idea, project progress, key discussion points, … PRIOR to the consultation session to the lecturer**.

[during your consultation session]

- Be on time.
- **Have a (written) agenda prepared**.
  - What do you want to ask or seek clarification / guidance / feedback on?
  - Provide sufficient context.
  - Prioritise your agenda items!
- Have supplementary information readily available.
- Respect the meeting time.

8

Project Proposal: Guidance & Expectations (revisited)

10

10

Project Proposal Guidance (revisited)

Guiding Expectations

- Apply current or emerging statistical concepts, methods or techniques to an interesting application or real-world data set(s).
  - Learn **beyond** what was covered in prior coursework.
- Each project should include **some form of analysis** and **some form of experimentation** on real-world or synthetic data sets.
- **Due date: by Feb 9th, 11:59pm**
- Requirements:
  - Use the provided submission template → Canvas.
  - Attend 1st consultation session prior to submission.

Project Proposal Content

Up to 6 pages (using the provided project proposal template):

- Project title (this can be a working title)
- Project introduction / motivation
- **Problem statement** or **hypothesis**
- **Literature review / concepts study** (2-3 pages)
- **Project objective(s)**
- **Requirements** (in terms of data sets, tools, etc.)
- **Success measure(s)**
- **Project plan** including key activities
- References
- Appendix (optional)


Capstone Project Title

Qualities of a Project Proposal

- **Concise, clear**, and direct
- **Structured**
  - Someone should be able to scan through your proposal in minutes and get what you are talking about.
- **Straightforward**
  - A reviewer wants to know: What is your project about? How will you do it? What resources will you need, and how will you get them? Who will be involved? … → Address these questions directly with precise & simple sentences.
- **Compelling**
  - The language you use should be convincing. Be confident about what you want to do, be enthusiastic, and share your enthusiasm.
  - Substantiate why you think the project will work and illustrate the project's relevance.
- **Detailed**
  - Even though the proposal should be brief, include as much detail as is needed to support your points.

Project Selection Considerations (revisited)

- **Scope and Complexity**: The project should be achievable within 9 weeks (→ be ambitious and challenge yourselves but stay realistic).
- **Skills and Expertise**: Align with what you, as a group, already know while also providing opportunities to learn and to apply new skills.
- **Resource Availability**: This includes data, software, hardware, and any other necessary tools or materials.
- Align with the course's **Learning Objectives**.
- **Interest and Engagement**: A project that resonates with your passion or career goals will likely result in higher motivation and better outcomes.
- **Real-World Application**: The project should have a real-world application or relevance (e.g., a problem faced by an industry, a research question, or a social issue) → This not only enhances learning but also adds value to your portfolios.

- **Advanced Statistical Techniques:** Utilize methods beyond basic descriptive statistics.
  - Examples: Inferential statistics, regression analysis, time series analysis, Bayesian methods, or machine learning algorithms.
- **Large or Complex Dataset**
  - **Criteria**: Choose a dataset that has more than **1 million records** or is high-dimensional, with **over 100 features** or use **multi-media data**.
- **Multivariate Analysis:** Analyse and interpret relationships between multiple variables.
- **Validation of Analysis:** Rigorously validate the performance of any analysis.
  - **Methods**: Include data splitting (training and test sets), cross-validation, and appropriate performance metrics.
- **Reproducibility and Documentation:** Maintain a strong emphasis on the ability to reproduce results.
  - **Expectation**: Document code, models, and analysis thoroughly for reproducibility.

→ **Projects should be challenging, relevant, and align with the advanced level of the course / degree!**

13

ST5188 Evaluation Rubric – Project Proposal (15 marks)

Evaluation Criteria	Weak	Satisfactory	Excellent
Introduction (1 mark)			
Has the context been set (→ relevance, clarity, and depth)?	- The context is vague, lacking clear relevance to the broader field. - Little to no connection to current trends or challenges. - The description is shallow, offering minimal background information, making it difficult to understand the importance of the project.	- The context is somewhat relevant, though it may not fully capture current trends or challenges in the field. - The description is clear but lacks depth, providing only a basic understanding of the background. - Some relevance is shown, but it could be stronger or more detailed.	- The context is highly relevant, aligning well with current trends and challenges in the field. - The description is clear, detailed, and comprehensive, providing a strong background that enhances the understanding of the project's importance.
Has a use case been described adequately (→ specificity and detail)?	- The use case is missing or poorly defined, lacking specific examples or scenarios. - Little detail is provided, making it difficult to see how the use case is relevant or important.	- The use case is relevant but lacks some specificity or detail. - While the use case is adequately described, it may not fully illustrate its relevance to the project.	- The use case is clearly defined with specific and detailed examples or scenarios. - The use case is well-chosen and provides a compelling example of the problem or challenge being addressed.
Has a rationale for choosing/tackling the project been provided (→ significance and practical importance)?	- The rationale for the project is unclear or weakly justified. - Little discussion of the significance or practical importance of the project. - The rationale does not convincingly explain why this project is necessary or valuable.	- The rationale is clear but could be stronger in justifying the project's significance and practical importance. - The importance of the project is discussed but may not be fully convincing. - The rationale is adequate but lacks depth or originality.	- The rationale is compelling, clearly explaining the significance and practical importance of the project. - Provides strong justification for why the project is necessary, offering both theoretical and practical reasons. - The rationale highlights innovation and originality in tackling the project.

1

Project Proposal: Key Terms Explained

	Weak	Satisfactory	Excellent
Problem Statement or Hypothesis (2 marks)			
Is there a clearly identifiable problem statement or hypothesis (→ distinct and explicit)?	- The problem statement or hypothesis is vague, poorly defined, or missing. - It lacks distinctness and fails to explicitly state the issue being addressed.	- The problem statement or hypothesis is identifiable but may lack full clarity or explicitness. - It somewhat addresses the issue but could be more distinct.	- The problem statement or hypothesis is clearly identifiable, distinct, and explicit. - It precisely states the issue being addressed, leaving no ambiguity.
Has a gap been identified and explored in depth?	- The gap is either not identified or only superficially mentioned. - Little to no exploration of the gap is provided, leaving it unclear why this issue is significant.	- The gap is identified but not explored in great depth. - There is some discussion, but the exploration lacks depth or thorough analysis.	- The gap is clearly identified and thoroughly explored. - The depth of exploration provides a comprehensive understanding of the issue's significance.
Has the gap been quantitatively assessed (→ extent or impact)?	- There is no quantitative assessment of the gap, or the assessment is unclear or irrelevant. - The impact or extent of the gap is not well understood or documented.	- The gap has been assessed quantitatively but may lack sufficient detail or relevance. - Some understanding of the impact or extent is provided, but it could be stronger.	- The gap is quantitatively assessed in detail, providing clear evidence of its extent or impact. - The assessment is relevant and significantly enhances understanding of the gap.
Has the importance or relevance of addressing the gap been discussed?	- The importance or relevance of addressing the gap is not discussed or is poorly articulated. - There is little to no connection made between the gap and the broader field or objectives.	- The importance or relevance is discussed but lacks depth or strong justification. - Some connection is made, but it may not fully convey the significance of addressing the gap.	- The importance or relevance of addressing the gap is thoroughly discussed and well-justified. - Strong connections are made between the gap and its broader impact, clearly highlighting why it needs to be addressed.

14

2018

Running Example

[used in lectures 3, 5 & 6]

- Aug 2017 → Dec 2018: Helping local e-com fraud startup to modernise & scale their fraud solutions.
- Dec 2017 @ Strata Singapore
  - Presentation on feasibility of auto-encoders for anomaly detection use cases (espc. for new / rare anomalies)
    - → This sounds promising for detecting new / emerging / rare fraud patterns!
- 2018: Embarked on 'side-project' (2 teams):
  - **Explore application of** CNNs and **autoencoders for payments fraud detection**.
  - From exploration to PoC to production (across SEA) in one year.

- Fraud scoring works hand-in-hand with preventive measures.
  - E.g., 3D Secure (payer authentication).
- Historically, fraud systems have relied on **rules hand-curated by fraud experts** to catch fraudulent activity.
- Later, **real-time block-/allow-listing** and **dynamic rules** were added.
- 2000s: ML-based fraud scoring (using sampling methods).
- Early 2010s: **real-time ML-based fraud scoring** became feasible (→ personalised scoring).
  - Due to early successes in payments [Big data].
- Late 2010s: Confirming credentials → **Confirming the person / persona**.

→ 2017/18: **Effective** (advanced) **fraud scoring requires multiple layers**:

- Additional data sources (e.g., mobile location data, device IDs, 3rd-party data, …).
- More complex data representations (→ graph networks).
- Diversify fraud scoring approaches (→ multi-pronged approach).

Project Title & **Introduction / Motivation**

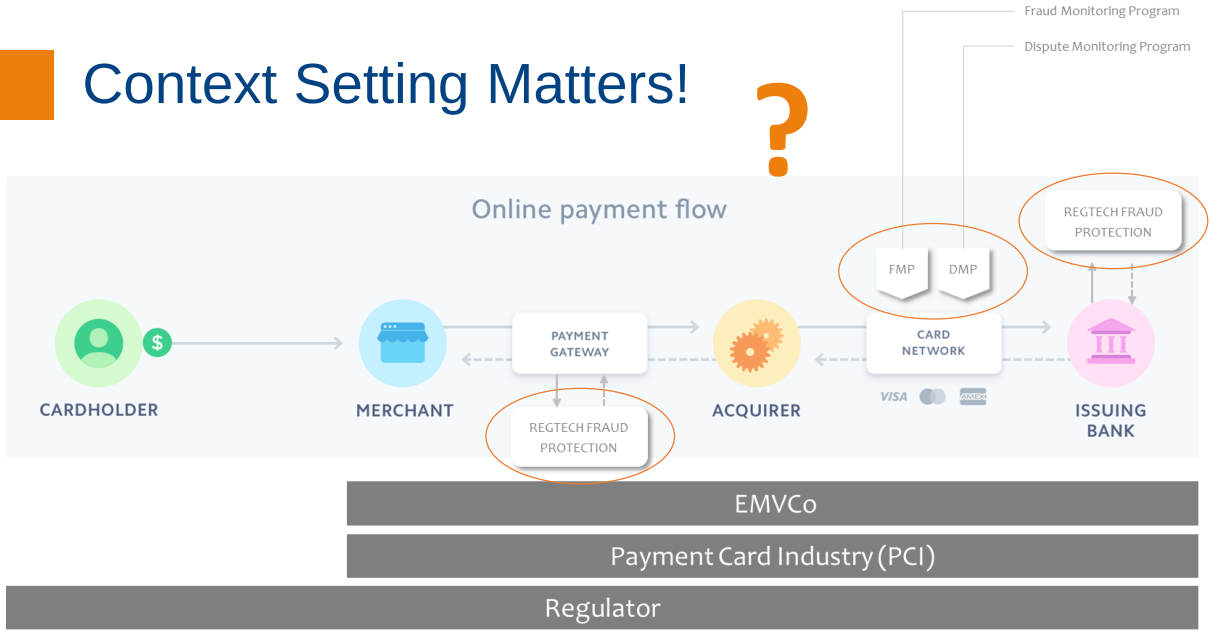
Title

- Craft a **brief, direct, and all-inclusive title** for the proposal.
  - → Ensure it reflects the core idea or problem the project addresses.
  - Start with a working title and refine.
  - E.g., write down the research problem as a question then craft your title in response to that question.
- Try to make the title as intriguing as possible to get your readers interested in what you have to say.

Introduction / Motivation

- Set the scene / **provide context** for the reader (→ broad overview of the field or industry).
- Describe **what the project is about** (→ transition from the broad context to **your specific use case**).
- Give the **rationale** for the project, i.e., why you think it is important and should be carried out.
  - Think of the rationale as to the social implications of the project you are about to undertake.
- Specific ST5188 evaluation criteria:
  - *Has the context been set / use case been described adequately (→ clear and comprehensive)?*
  - *Has a rationale for choosing / tackling the project been provided (→ significance and practical nature)?*

Context Setting Matters!



Sources:
- *"Introduction to online payments"*; published by Stripe; accessed in Jun 2020; https://stripe.com/gb/guides/introduction-to-online-payments
- Markus Kirchberg; *"Module 2: RegTech Use Cases from Different Perspectives: Finance Institutions vs. Regulators vs. FinTechs vs. Consumers"*; Advanced Certificate in RegTech via SMU Academy; 2021.

Problem Statement (or Hypothesis)

- A **problem statement** is a definite or clear expression about an area of concern, a condition to be improved upon, a difficulty to be eliminated, or a troubling question that exists in theory or in practice that points to the need for meaningful understanding and deliberate investigation. [*]
  - **Precisely formulated** (use technical terms).
  - **Specific for your project** (→ what you want to examine and how you want to examine it).
  - May include an explanation, reflection & discussion of the problem.
    - → Precursor to the literature review.
- A research problem does not state how to do something, offer a vague or broad proposition, or present a value question.
  - A problem statement is **open ended** (i.e., it does not predict the outcome).
  - A **hypothesis** states a definite outcome or set of outcomes that you might predict!
    - Must be in testable form!

[*] Adapted from Bryman, Alan. "*The Research Question in Social Research: What is its Role?*" International Journal of Social Research Methodology 10 (2007): 5-20.

- Key elements of an effective problem statement:
  - **Identify the gap** that exists today.
  - Describe **when and where the problem was first observed** and what kind of trend it is following.
    - The latter will serve as a precursor to your literature review / concept study
  - **Quantify the gap** → impact.
  - Comment on the **importance / relevance** of addressing this gap.
- Problem statement in (Project Proposal) context:
  - Problem statement identifies and **precisely formulates the gap**.
  - Your literature review **explores the intricacies of the gap** in detail, outlines relevant literature, and what you can learn from those efforts.
  - Project objectives are clear statements of what you **aim to achieve**.
- Specific ST5188 evaluation criteria:
  - *Is there a clearly identifiable problem statement or hypothesis (→ distinct and explicit)?*
  - *Has a gap been identified and explored in depth?*
  - *Has the gap been quantitatively assessed (→ extend or impact)?*
  - *Has the importance or relevance of addressing the gap been discussed?*

Problem Statement – **Key Elements**

1. **Identification of the Gap:**
   - A clear explanation of **what is missing or lacking** in the current understanding, technology, process, or condition.
   - A description of the current state and **how it differs from the desired state** or potential.
   - How to Approach:
     - **Be specific** about the nature of the gap.
     - Use data, observations, or preliminary research to **substantiate the existence of this gap**.
2. **Historical Context and Trend Analysis:**
   - A brief history of **when and where the problem was first identified**.
   - **Analyse how the problem has evolved** / changed over time.
   - Insight into the trends or trajectories of the problem.
   - How to Approach:
     - Use a chronological narrative or data to illustrate changes over time.
     - Relate this trend to the broader context of your field or area of interest.
3. **Quantification of the Gap and Its Impact:**
   - **Concrete** data or estimations to quantify the gap.
   - An **analysis of the impact** of this gap on relevant fields, communities, industries, etc.
   - How to Approach:
     - Employ statistics, case studies, or models to quantify the gap.
     - Discuss the **direct and indirect impacts** (and potential long-term effects).
4. **Significance of Addressing the Gap:**
   - A justification of **why** it is important or relevant to address this gap.
   - Discussion on the potential benefits or advancements that could arise from addressing the problem.
   - How to Approach:
     - Link the importance of the gap to larger societal, academic, or industry goals.
     - Consider and articulate the consequences of not addressing the gap.

→ **Connection to Literature Review:**
- Provides the current knowledge landscape related to the gap.
- Indicates how your project will build upon, challenge, or diverge from existing literature.

2018

Problem Statement – Key Elements:
Detecting Emerging Fraudulent Transaction Patterns with Auto-Encoders

- **Identification of the Gap:**
  - **Current State:** Traditional fraud detection systems in financial institutions primarily rely on predefined rules and patterns. These systems are efficient at detecting known types of fraud but struggle with new, evolving, or previously unseen patterns.
  - **Desired State:** A dynamic, adaptable system capable of identifying emerging fraudulent patterns in real-time, even when these patterns deviate significantly from known fraud signatures.
  - **A closely related gap**: Lack of representation of underrepresented groups in traditional fraud detection models.
    - Example: Small businesses and low-income individuals are disproportionately affected by fraud as their transaction data is not be well-represented in current models. Auto-encoders will also help to better uncover such patterns in the transaction data that would otherwise be overlooked, thus providing a more comprehensive approach to fraud detection.
- **Historical Context and Trend Analysis:**
  - **Origin of the Problem:** The limitation of rule-based systems became apparent with the rise in sophisticated, ever-changing fraudulent schemes, especially in online transactions.
  - **Trend Analysis:** There has been a significant increase in the complexity and variability of fraudulent transactions over the past decade, with fraudsters continually adapting their strategies to circumvent existing detection mechanisms.
    - 45% increase in account takeovers in Q2 2017; $3.3B lost by merchants to account takeovers in Q2 2017 [source: https://www.pymnts.com/global-fraud-index/]
- **Quantification of the Gap and Its Impact:**
  - **Quantification:** Studies have shown that new fraud patterns can take weeks to months to identify using traditional methods, during which millions of dollars could be lost.
  - **Impact:** The financial industry faces substantial losses due to undetected or belatedly detected fraud. Additionally, there is a loss of customer trust and potential regulatory penalties.
    - Estimated fraud cost globally: $31.67B by 2020 [https://nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf]
- **Significance of Addressing the Gap:**
  - **Benefits of Addressing the Gap:** Implementing a system capable of detecting emerging fraud patterns in real-time can significantly reduce financial losses. It enhances customer trust and meets evolving regulatory requirements for fraud prevention.
  - **Consequences of Inaction:** Failure to address this gap will likely result in continued financial losses, damage to reputation, and potential legal repercussions for financial institutions.

Hypothesis Testing

- What is a Hypothesis?
  - A hypothesis is a proposed relationship between two or more variables.
  - It typically follows an "*If... then...*" structure.
- Key Characteristics of a Hypothesis:
  - **Specific:** Clearly define the variables you will test.
  - **Testable:** The hypothesis can be supported or refuted through experimentation or research.
  - **Logical:** There should be a rationale for the hypothesis based on existing theories or observations.
  - **Predictive:** It makes a statement about the expected outcome.
- Example:
  - "If auto-encoders, trained on *normal network traffic*, can successfully detect *anomalous patterns* that deviate from expected traffic behaviour, then they should also be capable of identifying emerging *fraudulent transaction patterns* in payment data by flagging anomalies that deviate from *normal transaction behaviour*."

Literature Review (see Lecture 2)

- A literature review is a **descriptive summary of prior research on your chosen topic** (with particular focus on the problem statement).
- Purpose 1: Inform readers of the significant **knowledge and ideas that have been established**.
  - *We evaluate breadth, depth, and inclusion of recent (i.e., past 5 years) literature!*
- Purpose 2: **Compare, contrast and/or connect** findings that were identified when reviewing researchers' work.
  - *We evaluate whether the review demonstrates a deep understanding of nuances and contrasts in the literature, rather than just summaries!*
- Purpose 3: Establish an **educational research context for your own research project**.
  - → Identify what is known about the topic and the problem statement that you are studying.
  - → Are there any gaps, shortcomings and/or failures?
  - → Provide a broad and critical analysis of relevant literature.
  - → Cite all sources!
  - *We evaluate whether there are connections between the reviewed literature and your project's intentions!*

What to include?

- **Identify relevant research in the field (→ breadth).**
  - Review previous studies (espc. recent efforts) that have been conducted on the topic of interest and identifying gaps in the existing literature that the proposed project aims to fill.
  - Highlight achievements and limitations of these studies to depict the field's evolution.
- **Analysis and synthesis existing literature (→ depth).**
  - Contrast different methodologies, findings, and perspectives to demonstrate a nuanced understanding of the field and to highlight where opinions or findings diverge.
- **Define key terms and concepts.**
  - Clearly define and explain any key terms or concepts that are central to the project.
- **Identify potential data sources and research gaps.**
  - What data did other researchers use? Is it applicable to your project?.
  - Identify gaps or under-researched areas in the existing literature that your project could address.
- Identify any potential challenges or limitations that may arise during the project and propose strategies for addressing them.
- **Integrate with your project.**
  - Make clear and logical connections between the literature reviewed and your own project intentions (e.g., how will your project build upon, diverge from, or fill the gaps in existing literature?).

Literature Review vs. Related Work

Lit. Review for Project Proposal

- **Purpose**:
  - To justify the need for the proposed research.
  - To demonstrate familiarity with the field and to identify gaps that the proposed research is intending to fill.
- **Content**:
  - A broad overview of the field to situate the proposed research within the larger academic conversation.
  - Identification of gaps in the existing literature that the research will address.
  - Review of methodologies and findings of previous research to justify the chosen approach.
- **Depth and Breadth**:
  - Typically, broader in scope to show the research space and to argue the proposal's relevance.
  - May not delve into excessive detail about each source but should provide enough depth to substantiate the research need.
- **Style**:
  - Persuasive, aiming to convince the reader of the importance and viability of the proposed research.
  - Forward-looking, discussing how the proposed research will contribute to the field.

Related Work for Final Report

- **Purpose**:
  - To contextualize the results and discussion of the research findings within the existing body of knowledge.
  - To acknowledge previous work and to delineate how the current research adds to or differs from existing studies.
- **Content**:
  - A focused examination of studies directly related to the research question and methodology.
  - Critical analysis of the most relevant research, discussing similarities and differences in findings and methodologies.
- **Depth and Breadth**:
  - More depth into specific studies that are directly related to the research question and outcomes.
  - Less breadth than in a proposal literature review, as the focus is narrower and more aligned with the research results.
- **Style**:
  - Analytical, providing a critique of previous work and discussing how the current research engages with it.
  - Retrospective, reflecting on how the research fits into and expands the existing literature.

- **Comparison**:
  - The literature review in a project proposal is more about **setting the stage for the research**, identifying why it is necessary, and what it intends to achieve.
  - The related work in the final report is about **positioning the research** within the existing academic work, showing how it adds new insights or data to the field.
- **Contrast**:
  - The proposal's literature review is more generalized and preparatory, while the final report's related work is specific and reflective.
  - The literature review in **the proposal is used to secure approval** or funding for the project; the related work section in **the final report is used to demonstrate the research's contribution** to the field.
  - Audience: Proposal (funders or a thesis committee who need to be persuaded of the work's value) vs. final report (peers and scholars who are interested in the detailed contributions and implications of the findings).

Project Objectives

- […] are **clear statements of what you aim to achieve** through your research.
  - Should relate to both, your problem statement and your literature review.
  - Objectives are highly dependent on the problem, but they are different in that they state what you aim to achieve.
- Objectives should always be SMART (**specific, measurable, achievable, realistic** & timebound).
  - Specific → Requirements, success measures
  - Measurable → Success measures
  - Achievable → Project plan
  - Realistic → Project plan
  - Timebound → Out-of-scope
- Specific ST5188 evaluation criteria:
  - *Are the objectives (goals or aims, not activities!) clearly identified?*
  - *Do the objectives align with the problem statement?*
  - *Are the objectives SMAR(T)?*

Examples:

- Not SMART
  - "*Develop a model that detects emerging fraud patterns in financial transactions.*"
- Somewhat SMART
  - "*Develop an unsupervised learning model using auto-encoders to identify emerging fraud patterns in transaction data from a financial institution.*"
- SMART
  - "*Develop and optimize an auto-encoder model to identify anomalies indicative of potential emerging fraudulent transaction patterns in real-time from a dataset of at least 1 million financial transactions. The model aims to achieve an AUC-ROC of greater than 0.XX [1], maintain a false positive rate below X% [2], and ensure a high recall rate.*"

Project Requirements

- **Anything that is needed throughout the project life cycle** for you to complete the project successfully.
  - Resources (e.g., data, compute capabilities, …), tools, access to domain knowledge, …
- What data set(s) do you plan to use?
  - Describe the total amount of data that will be available and justify/articulate that it is sufficient for the project.
  - Must you do significant work to get, merge, and/or convert the data?
    - → Describe the process and approximate effort required.
- Specific ST5188 evaluation criteria:
  - *Are data and technical (or skills / domain) requirements clearly identified?*
  - *Do they seem reasonable wrt. the identified problem statement and objectives?*

Examples of what to cover / questions to ask yourselves:

- Data Requirements:
  - Sources: From where will you obtain the data?
  - Access: Do you need special permissions or licenses to access or use the data?
  - Volume: How much data is needed for meaningful results?
  - Formats: In what formats is the data available, and do you need to convert it?
  - Quality: Is the data clean, or does it require preprocessing?
  - Update Frequency: How often is the data updated? Do you need real-time, daily, monthly data, etc.?
  - Privacy and Ethical Considerations: Does the data contain personal information? If so, how will you handle and protect it?
- Technical Requirements:
  - Software & Tools: What software or tools will you use? Think about programming languages, database systems, libraries, frameworks, etc.
  - Hardware: Do you need specialized hardware? Consider storage, compute power, GPUs for certain tasks, etc.
  - Infrastructure: Cloud services, servers, APIs, etc.
- Skill and Expertise Requirements:
  - What domain knowledge is necessary?
  - What technical expertise is essential?

Success Measures

- How will the results of the project be evaluated?
- Benefits:
  - → Enables tracking of progress.
  - → Facilitates decision making.
  - → Helps to avoid project failure.
- Specific ST5188 evaluation criteria:
  - *Are the success measures consistent with the objective?*
  - *Have the success measures been defined formally or given in pseudocode?*

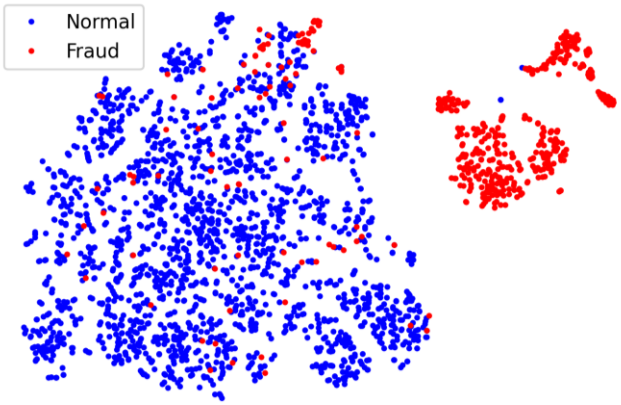
→ Often **challenging to define**!

- You need to take the business context in account.

- How do you plan to monitor & evaluate your ongoing work, results, and final contributions?
  - Includes data selection, feature generation, model training, performance measures, …
- **How will you** know if you are on track / doing the 'right' thing or **determine the utility of your results**?
  - E.g., how do you plan to determine if / when a model performs well? When is it good enough to satisfy your objectives?
    - Clearly define what '*well*', '*good enough*', … mean (**in technical / mathematical terms**) wrt. the project context.
  - Are there other papers (→ literature review) that you can compare your results against, or is this a novel problem?
    - → Can often be used as **baseline** or to set a **target value**!

Success Measures Example
(E-com Fraud: Autoencoder Representation)

2D visualisation of the learnt data representation using normal and fraudulent transactions.



How to define success?

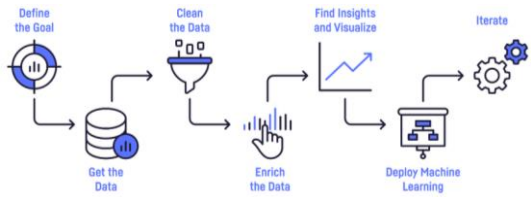
- False positives have a huge impact → accuracy alone is not a good measure of success.

Source: "*False Positives: The Biggest Loss of Value in e-Commerce*"; Fraugster blog post; https://www.fraugster.com/resources/post/false-positives-biggest-ecommerce-loss

Project Plan
with Key Activities

- A project plan should give you (and us) an indication of **what you want to achieve by when and who is assigned to do which tasks**.
  - What are the main data science tasks?
    - Descriptions can be brief but must be clear and well-defined.
  - How to break down the work (into major and minor items)?
    - Every activity that required more than 5 days of effort needs to be broken down.
  - How to allot the work to the different members in your group?
    - → Assemble a tentative timeline (9 weeks!).
  - What difficulties could you encounter?
    - If you cannot achieve the expected result, what could you try next? …

→ **Serves as a 'contract' for your group work** (seek agreement and track progress).



Source: Alvia Smith; "*7 Fundamental Steps to Complete a Data Analytics Project*"; dataiku blog; Jul 2019; https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success

- Specific ST5188 evaluation criteria:
  - *Primary: Have the key data science tasks been clearly identified?*
  - *Secondary:*
    - *Major and minor tasks have been identified and logically sequenced;*
    - *Timeline (weekly) or time estimates (for minor tasks) are reasonable;*
    - *Tasks are assigned to group member(s);*
    - *Flexibility (e.g., time set aside) for adjustments based on challenges or new findings; and*
    - *The cyclic nature of data science activities is evident.*

Project Plan Example (weeks 01 and 02 only)

Week	Main Objective(s)	Key Activities / Tasks	Responsible Member(s)	Estimated Time Commitment (by Member)	Milestones	Potential Challenges & Contingency Plans
1 [week 5]	Project Initiation & Data Collection	• Team alignment meeting	All members	3hrs each	• Scope finalized • Data collection initiated	• Delays in data access (→ adjust data source)
		• Finalize & sign off project plan	All members	1hr each		
		• Data collection and validation	DC: K, L & M DV: X, Y & Z	DC: 4hrs each DV: 6hrs each		
2 [week 6]	Data Integration & Comprehensive Analysis	• In-depth exploratory data analysis (EDA)	K & M	10hrs each	• Data cleaned and integrated	• Data quality issues (→ explore alternative data preparation methods)
		• Rigorous data cleaning and harmonization	L, X, Y & Z	L: 12hrs X, Y & Z: 10 hrs each		

Project Proposal: Good vs. Bad Practises

Introduction and Motivation

Bad practises

- Providing too much detail or **too little context**.
- Being **overly technical** in the introduction.
- Failing to clearly establish the significance of the project.

Good practises

- Clearly articulate why the project is important and relevant.
- Provide context and background information.
- **Engage the reader** with compelling reasons for the project's necessity.

Problem Statement and/or Hypothesis

Bad practises

1. **Focus on symptoms** associated with the problem.
   - Leads to distractions.
2. Offering solutions early (→ natural tendency to eliminating pain).
   - Tends to 'fix' symptoms.
3. Assign blame to those who 'caused' the problem.

Good practises

1. Identify and focus on THE core problem / gap.
   - Be precise and clear.
2. First **gain a "profound" understanding of the current state**.
   - Tackle the core problem.
3. Failures / mistakes are opportunities to learn!

Literature Review

Bad practises

1. Over-reliance on **low-quality sources**.
   - Non-academic resources: blog posts, ML notebooks, news articles, …
2. **Lack of current literature**.
   - Danger of tackling a solved problem.
3. Simply summarise relevant literature.
   - Fails to demonstrate your learnings.
4. Irrelevant or **unfocused content**.
5. **Plagiarism** and poor referencing.
   - Plagiarism is a serious form of academic misconduct.

Good practises

1. Use credible academic literature.
   - Validate non-academic resources; use them sparingly.
2. Have a mix of classic / seminal and current literature.
3. **Integrate existing research to show how it fits together**.
   - Relate it directly to your problem statement.
4. The purpose of the literature review is to build the foundation that will help you tackle the problem statement in order to achieve your objectives.
5. Reference your source materials and **explain concepts in your own words** (→ integrate).

Success Measures

Bad practises

1. Be “better” than xyz.
   - Avoid vague terms; quantify what “better” means.
2. Solely rely on AI / ML metrics.
   - Typically, data science capabilities are only a part of the solution.
3. Choose the AI / ML technical metrics without considering business context.
   - Involvement of domain experts from the onset of a project is essential (but often neglected).

Good practises

1. Quantify success measures in technical / mathematical terms.
2. Use AI / ML technical metrics in conjunction with business value metrics / operational KPIs to measure the effectiveness of an AI / ML solution.
3. Let the context guide the definition of the success measure.
   - E.g., fraud use cases.

Project Plan

Bad practises

Week	Core Activities (Example)	Deliverables (Example)
1	• Read up on ST5188 • Attend lecture 1 • Form and register project group • Brainstorm / review project ideas	
2	• Attend lecture 2 • Explore project topics • Commence literature review	
3	• Attend lecture 3 • Formulate project scope, problem statement, objectives, approach, success measure, ...	
4	• Attend lecture 4 • Schedule & attend first consultation session with lecturer • Formulate project plan • Finalise project proposal	Project Proposal
5	• Commence project work • Attend lecture 5	
6	• Continue project work • Attend lecture 6	
	Recess Week	
7	• Continue project work • Schedule and attend second consultation session with lecturer	Peer Group Evaluation 1
8	• Continue project work • Write project progress report	Project Progress Report
9	• Continue project work	
10	• Continue project work	
11	• Continue project work • Schedule and attend third consultation session with lecturer	
12	• Continue project work • Plan for project presentation	
13	• Present your project to lecturer, TAs, and peers • Complete project work • Attend two peer project presentations and write peer project evaluations • Finalise final project report • Complete second peer group evaluation form	Project Presentation Peer Project Evaluations Project Report Submission Peer Group Evaluation 2
14+	Reading Week & Examination Weeks 1 & 2	

Good practises

- Provide a detailed timeline with milestones.
- Assign roles and responsibilities within the group.
- Include risk assessment and mitigation strategies.

Week	Main Objective(s)	Key Activities / Tasks	Responsible Member(s)	Estimated Time Commitment (by Member)	Milestones	Potential Challenges & Contingency Plans
1 (week 5)	Project Initiation & Data Collection	• Team alignment meeting	All members	3hrs each	• Scope finalized • Data collection initiated	• Delays in data access (→ adjust data source)
		• Finalize & sign off project plan	All members	1hr each		
		• Data collection and validation	DC: K, L & M DV: X, Y & Z	DC: 4hrs each DV: 6hrs each		
2 (week 6)	Data Integration & Comprehensive Analysis	• In-depth exploratory data analysis (EDA)	K & M	10hrs each	• Data cleaned and integrated	• Data quality issues (→ explore alternative data preparation methods)
		• Rigorous data cleaning and harmonization	L, X, Y & Z	L: 12hrs X, Y & Z: 10 hrs each		

36

36

Questions?

Markus Kirchberg AUTHOR | TEACHER

Created 24 Jan 22:00 | Posted 24 Jan 22:00

Office Hours for Lectures 1--3 and General Q&A

Dear students,

I hope you are all doing well!

To support your learning and provide an opportunity to clarify concepts from Lecture 3, I will be hosting an **online office hour session** on **Tuesday (28/Jan) from 12:00 PM to 2:00 PM** via **Zoom**. The session will remain open until all questions are answered or until 2:00 PM, whichever comes first.

Details

- **Purpose**:
  - Answer questions related to **Lectures 1–3**.
  - Address general project-related questions.

Please note: We will not be addressing questions specific to individual project groups. Consultation sessions are designed for that purpose.

- **Format**:
  - You can submit questions in advance via the provided PollEverywhere link or during the live session.
  - Priority will be given to the most upvoted questions submitted during the session.
- **Recording**:
  - The session will be recorded, and the recording will be shared later for those who cannot attend live.

How to Submit Questions?

- Use this link to submit your questions in advance or during the session: PollEverywhere link
- You can also upvote questions you find most relevant, helping prioritize the discussion.

How to Attend?

- Time: Jan 28, 2025 12:00 PM Singapore
- Join Zoom Meeting via https://nus-sg.zoom.us/j/81289288942?pwd=4Thz9c0FUBY3EaOfbCSWBaQYjma2aH.1
- Meeting ID: 812 8928 8942
- Passcode: 470218

I look forward to your participation!

Cheers,

Markus

37

37

Thank You

Markus Kirchberg
Markus.Kirchberg@nus.edu.sg
www.markuskirchberg.net

40