

Introduction to Data Science and Engineering

- About the project



Some materials courtesy of
Rafael A. Irizarry, and are modified
from the original version.

Zhenqin (Michael) Wu / 吳楨欽

School of Computing and Data Science
University of Hong Kong

Slide deck originally created by RB Luo




Why are we doing this?

- Learn data science skills from a real-world project
- Train your capabilities in communicating your findings
- Make use of your own **domain expertise**

Project Proposal

	Week	Day	Time	Hour	Content	Notes
Sept. 1st	1	Mon	9:00-9:50, 10:00-10:50 am	2	Introduction and R basics	
Sept. 4th	1	Thu	3:00-3:50pm	1	R basics	
Sept. 8th	2	Mon	9:00-9:50, 10:00-10:50 am	2	N/A	
Sept. 11th	2	Thu	3:00-3:50pm	1	R markdown, Introduction to tidyverse	
Sept. 15th	3	Mon	9:00-9:50, 10:00-10:50 am	2	tidyverse, working with external data	Assignment 1
Sept. 18th	3	Thu	3:00-3:50pm	1	Data visualization	
Sept. 22nd	4	Mon	9:00-9:50, 10:00-10:50 am	2	Data visualization principles	
Sept. 25th	4	Thu	3:00-3:50pm	1	Lab session 1	
Sept. 29th	5	Mon	9:00-9:50, 10:00-10:50 am	2	Data wrangling: reshaping, joining	
Oct. 2nd	5	Thu	3:00-3:50pm	1	Data wrangling: web scraping, regex & string processing	Assignment 2
Oct. 6th	6	Mon	9:00-9:50, 10:00-10:50 am	2	Text mining	
Oct. 9th	6	Thu	3:00-3:50pm	1	Overview: AI for science	Assignment 1 deadline
Oct. 13th	7	Mon		2	Reading Week	
Oct. 16th	7	Thu		1	Reading Week	
Oct. 20th	8	Mon	9:00-9:50, 10:00-10:50 am	2	Mid-term	
Oct. 23rd	8	Thu	3:00-3:50pm	1	Project proposal write-up & Consultation	Project proposal submission open
Oct. 27th	9	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: probability	
Oct. 30th	9	Thu	3:00-3:50pm	1	Statistics: probability & statistical methods	Assignment 2 deadline
Nov. 3rd	10	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: statistical methods	Assignment 3
Nov. 6th	10	Thu	3:00-3:50pm	1	Statistics: Bayesian statistics	
Nov. 10th	11	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: Inference & Regression	
Nov. 13th	11	Thu	3:00-3:50pm	1	Machine learning	
Nov. 17th	12	Mon	9:00-9:50, 10:00-10:50 am	2	Deep learning/artificial intelligence/LLM in practice	Project proposal submission deadline
Nov. 20th	12	Thu	3:00-3:50pm	1	Lab session 2	
Nov. 24th	13	Mon	9:00-9:50, 10:00-10:50 am	2	Student presentations	7 slots
Nov. 27th	13	Thu	3:00-3:50pm	1	Student presentations	3 slots; Assignment 3 deadline
TBD				2	Final Exam	Project report deadline

**Submission window:
from Oct 23rd
to Nov 17th (@23:59)**



Project Proposal (5% of assessment)

The project proposal submission will be in the form of Moodle quiz, answer the following:

1. A tentative **topic** of your project;
2. The tentative **source of data** you plan to use for the project;
3. One to two data science **questions** you want to answer with the data;
4. A short proposal (within 300 words) stating:
 - **Importance** of the question(s);
 - **Challenges/difficulties** envisioned in answering the question(s)?
 - If there are notable **existing works** related to the question(s), describe them briefly.
 - A brief overview of your planned **approach or methods**.
5. *Do you want to give a short 10 min presentation (+5 min Q&A) in-class?
- 5-point bonus to your raw midterm (60) and final (~90) scores,
subject to the maximum score limit*

Project Report & Presentation

	Week	Day	Time	Hour	Content	Notes
Sept. 1st	1	Mon	9:00-9:50, 10:00-10:50 am	2	Introduction and R basics	
Sept. 4th	1	Thu	3:00-3:50pm	1	R basics & R markdown	
Sept. 8th	2	Mon	9:00-9:50, 10:00-10:50 am	2	Tidyverse	
Sept. 11th	2	Thu	3:00-3:50pm	1	Tidyverse, work with external datasets	Assignment 1
Sept. 15th	3	Mon	9:00-9:50, 10:00-10:50 am	2	Data visualization	
Sept. 18th	3	Thu	3:00-3:50pm	1	Data visualization in practice	
Sept. 22nd	4	Mon	9:00-9:50, 10:00-10:50 am	2	Data visualization principles	
Sept. 25th	4	Thu	3:00-3:50pm	1	Lab session 1	
Sept. 29th	5	Mon	9:00-9:50, 10:00-10:50 am	2	Data wrangling: reshaping, joining, web scraping	Assignment 2
Oct. 2nd	5	Thu	3:00-3:50pm	1	Data wrangling: regex & string processing	Assignment 1 deadline
Oct. 6th	6	Mon	9:00-9:50, 10:00-10:50 am	2	Text mining	
Oct. 9th	6	Thu	3:00-3:50pm	1	TBD: Recent topics in data science	
Oct. 13th	7	Mon		2	Reading Week	
Oct. 16th	7	Thu		1	Reading Week	
Oct. 20th	8	Mon	9:00-9:50, 10:00-10:50 am	2	Mid-term	
Oct. 23rd	8	Thu	3:00-3:50pm	1	Project proposal write-up & Consultation	Project proposal submission open
Oct. 27th	9	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: probability	
Oct. 30th	9	Thu	3:00-3:50pm	1	Statistics: probability & statistical methods	Assignment 2 deadline
Nov. 3rd	10	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: statistical methods	Assignment 3
Nov. 6th	10	Thu	3:00-3:50pm	1	Statistics: Bayesian statistics	
Nov. 10th	11	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: Inference & Regression	
Nov. 13th	11	Thu	3:00-3:50pm	1	Machine learning	
Nov. 17th	12	Mon	9:00-9:50, 10:00-10:50 am	2	Deep learning/artificial intelligence/LLM in practice	Project proposal submission deadline
Nov. 20th	12	Thu	3:00-3:50pm	1	Lab session 2	
Nov. 24th	13	Mon	9:00-9:50, 10:00-10:50 am	2	Student presentations	7 slots
Nov. 27th	13	Thu	3:00-3:50pm	1	Student presentations	3 slots; Assignment 3 deadline
TBD				2	Final Exam	Project report deadline

Final report & presentation video due on Nov 30th (@23:59)

10 in-class presentation slots. No need to submit the video if you presented in class.



Project Report & Presentation (15% of assessment)

- Project report:
 - Compose your report in **R Markdown** and knit it into an **html** page;
 - Include the **data files** used in your analysis in the submission, i.e., zip them together;
 - Or add a note if the data is private or too large to be uploaded (>100M, in which case attach a link to the data);
 - Ensure your R Markdown file is fully **runnable and reproducible**.
 - Include **graphics (preferably), data visualization, description, and analysis** to explain what you are doing and what data insights you have gained from the results/plots. **The report should NOT contain only codes.**
 - Add **in-line comments** in code blocks to explain your codes wherever necessary.
 - Note that your project report may be shared with future students of this course.



Project Report & Presentation

- Presentation video (5-10 min), share the following:
 - Problem definition & background
 - Data science questions in the project, their importance, challenges, related works
 - Data you have used
 - Results:
 - methods and approaches
 - major findings
 - Conclusion:
 - Insights from the data
 - Possible future extensions
 - Acknowledgements, References



Example topics

- Global temperature trends
- Tropical storms in the west pacific
- COVID-19 data analysis
- Music genres and trends
- Video game genres and trends
- Sentiment analysis of text corpus
- Usage of large language models

It is most encouraged to leverage your own **domain expertise** to propose unique ideas

Example topics

- From 2025 spring:

Date	Time	Last name	First name	Tentative title
Apr 24th	15:30-15:40	Salam	Sadiq	Does Money Buy Quality? Analyzing the Relationship Between Movie Budgets and Critical Success
Apr 24th	15:40-15:50	Seong	Hyun Soo	Werther's Effect on South Korean Teenagers
Apr 24th	15:50-16:00	Wu	Sihan	What Affects College Students' Sleep?
Apr 24th	16:00-16:10	Tanto	Caroline Avery	Which Kid My Parents Love The Most?
Apr 24th	16:10-16:20	Ng	Ching Lap	Decoding Decades of Rhythms - A Sentiment Analysis on the Lyrics of Trending Cantonese Songs from 1978 to 2023
Apr 28th	15:30-15:40	Lam	Kwong Chiu	Elon Musk's X Posts and TSLA Stock Volatility: Comparing Tesla-Related and Non-Tesla-Related Impacts in 2025
Apr 28th	15:40-15:50	Guo	Yixuan	Decoding Hong Kong Residents' Cross-border Consumption Trends: A data-driven analysis of northbound consumption behaviors
Apr 28th	15:50-16:00	Lau	Wai Man	E-sports @ Hong Kong
Apr 28th	16:00-16:10	Macleod	Luke Campbell	Gaelic Medium Education: Analysis of supply and demand to inform policy and practice
Apr 28th	16:10-16:20	Mukayev	Asset	Predicting Student Dropout Risk in Early University Stages
Apr 28th	16:20-16:30	Chap	Kin Cheung	How can data science unlock the secrets of China's football development?
Apr 28th	16:30-16:40	Cheng	Lefan	Hiking Downhill Blindfolded: How Gradient Descent Solve Real-World Puzzles
Apr 28th	16:40-16:50	Qiu	Wai Kit	What stops love in Hong Kong

- More candidate topics and example reports are shared on Moodle.