



# **EBU5601**

# **Data Design**

## **Module Introduction**

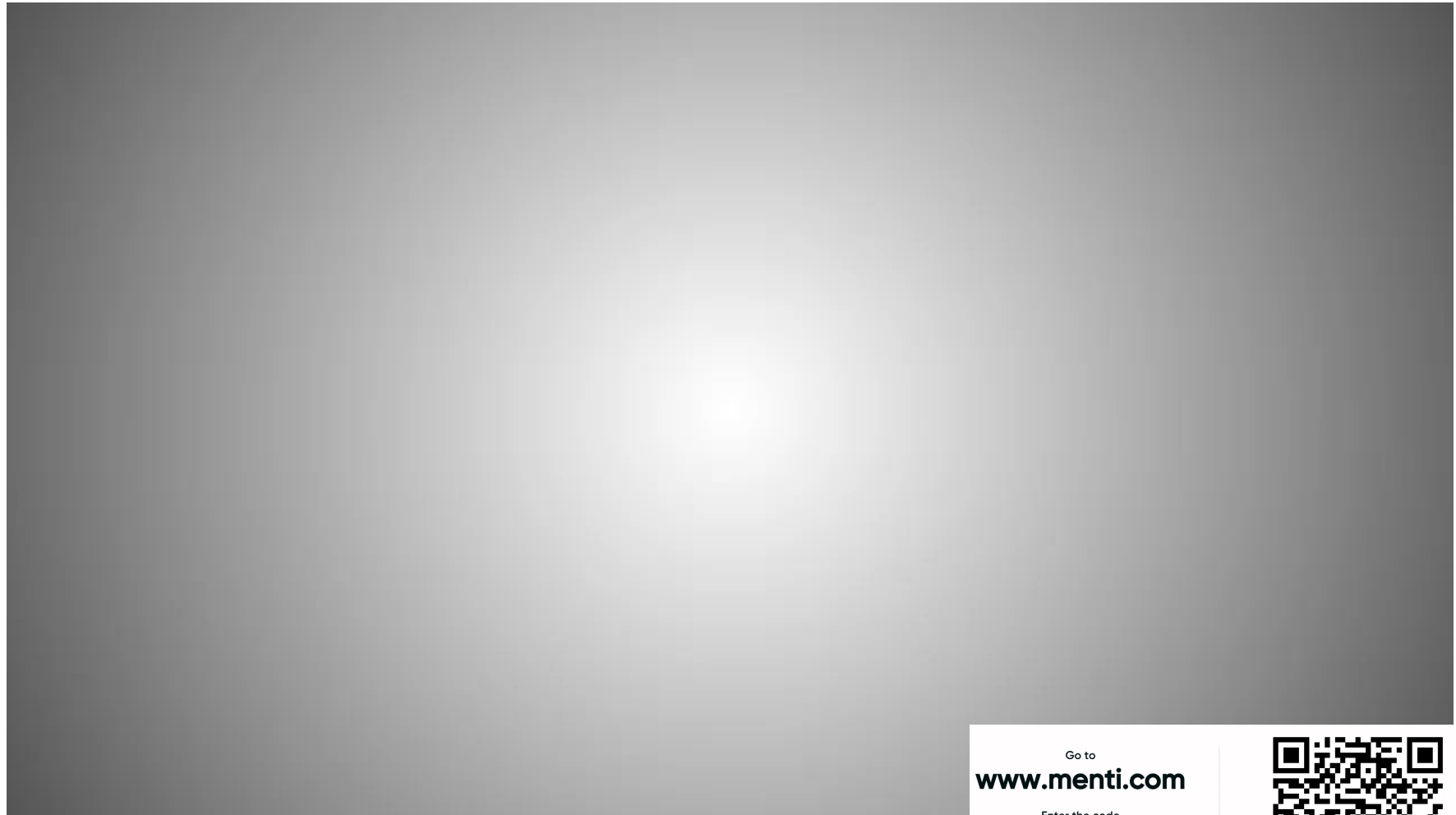
**Dr Chao Shu, Dr Xiaolan Liu**

School of Electronic Engineering and Computer Science  
Queen Mary University of London

Sep. 2024

# What is Data Science

- Watch the [short video](#) and submit keywords you heard in the video that describe what data science is.



Go to  
**www.menti.com**

Enter the code

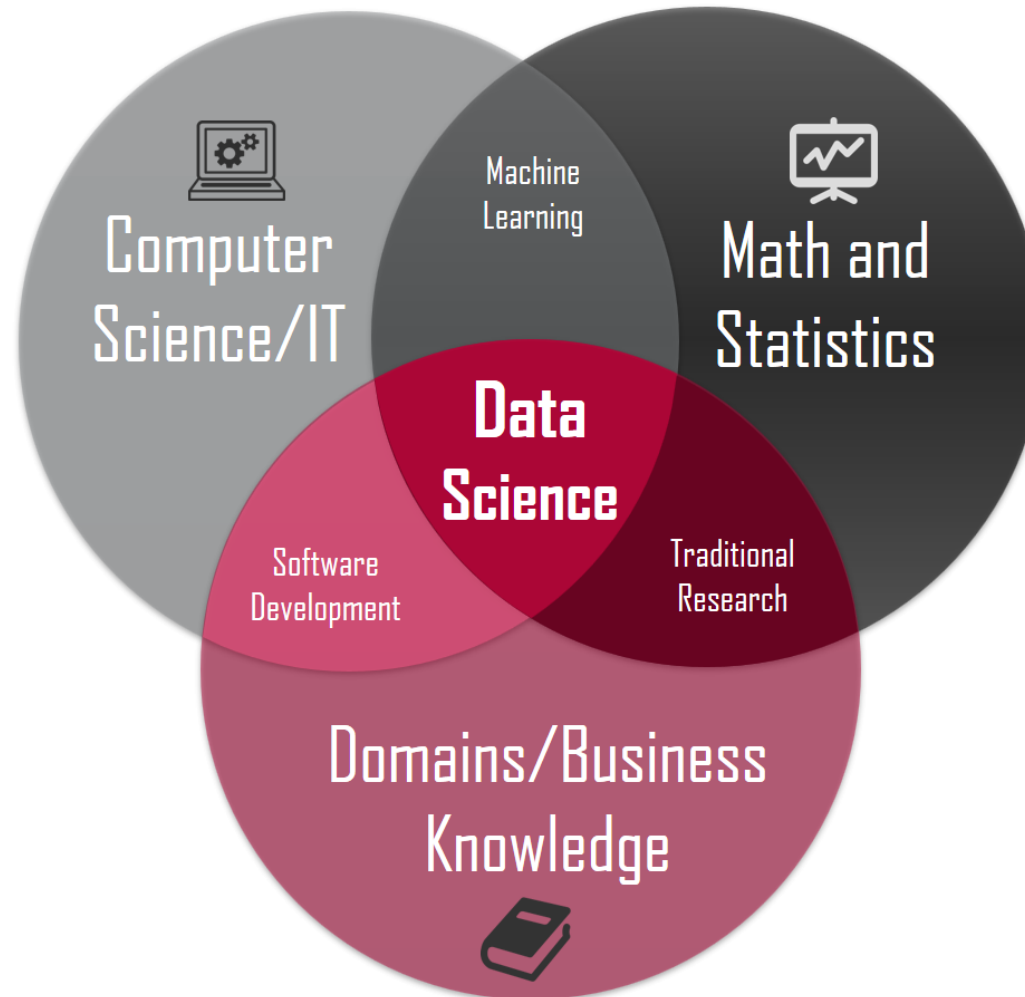
**8915 0742**



Or use QR code

# Data Science

- **Data science** has been hailed as the 'sexiest job of the 21st century'




Source: <https://towardsdatascience.com/introduction-to-statistics-e9d72d818745>

# Data Science

- The US Bureau of Labor Statistics predicts the number of data scientist roles to **grow 36% between 2021 and 2031**.

## Fastest Growing Occupations

PRINTER-FRIENDLY 

**Fastest growing occupations:** 20 occupations with the highest projected percent change of employment between 2021-31.

*Click on an occupation name to see the full occupational profile.*

OCCUPATION	GROWTH RATE, 2021-31	2021 MEDIAN PAY
<a href="#">Physician assistants</a>	28%	\$121,530 per year
<a href="#">Nurse practitioners</a>	46%	\$120,680 per year
<a href="#">Information security analysts</a>	35%	\$102,600 per year
<a href="#">Medical and health services managers</a>	28%	\$101,340 per year
<a href="#">Data scientists</a>	36%	\$100,910 per year
<a href="#">Statisticians</a>	33%	\$95,570 per year
<a href="#">Athletes and sports competitors</a>	36%	\$77,300 per year
<a href="#">Web developers</a>	30%	\$77,030 per year
<a href="#">Logisticians</a>	28%	\$77,030 per year
<a href="#">Physical therapist assistants</a>	26%	\$61,180 per year

Source: <https://www.bls.gov/ooh/fastest-growing.htm>

# Module Aims

---

- This module aims to explore the fundamental theory and practice of data analysis.
- Include describing data, uncovering patterns and insights, drawing meaningful conclusions, and clearly communicating critical findings, as well as applying data ethics principles in the data life cycle.
- Minimise the maths and focus on building the intuition and understanding of key statistic concepts for data science.
- Provide a solid foundation in statistical analysis that can be applied to roles such as Data Analyst, Business Analyst, and Data Scientist.

# Main Learning Outcomes

---

*By the end of the module the student will be able to:*

- Understand the typical data **analysis** process
- Evaluate data based on descriptive **statistics**
- Analyse real-world data by referring to the properties of theoretical distributions, such as the binomial distribution and normal distribution
- Apply bootstrapping to analyse sample statistics
- Apply **hypothesis** tests to an A/B testing context by performing and interpreting the results of hypothesis tests
- Analyse relationships between independent variables and a dependent variable using linear or logistic regression
- Apply appropriate plots to **visualise** data in **exploratory** data analysis, present and communicate data findings effectively

# Module Topics

---

- Topics to be Covered
  - Introduction to Data Analysis
  - Descriptive Statistics
  - Data Ethics
  - Data and Sampling Distributions
  - Statistical Experiments and Significance Testing
  - Regression
  - Data Visualisation

# Tools

- Packages/Libraries used in this module

Arrays



Data Frames  
Manipulate and  
analyse tabular data



Data Visualisations



IP[y]: IPython  
Interactive Computing  
Interactive computing



Scientific Computing



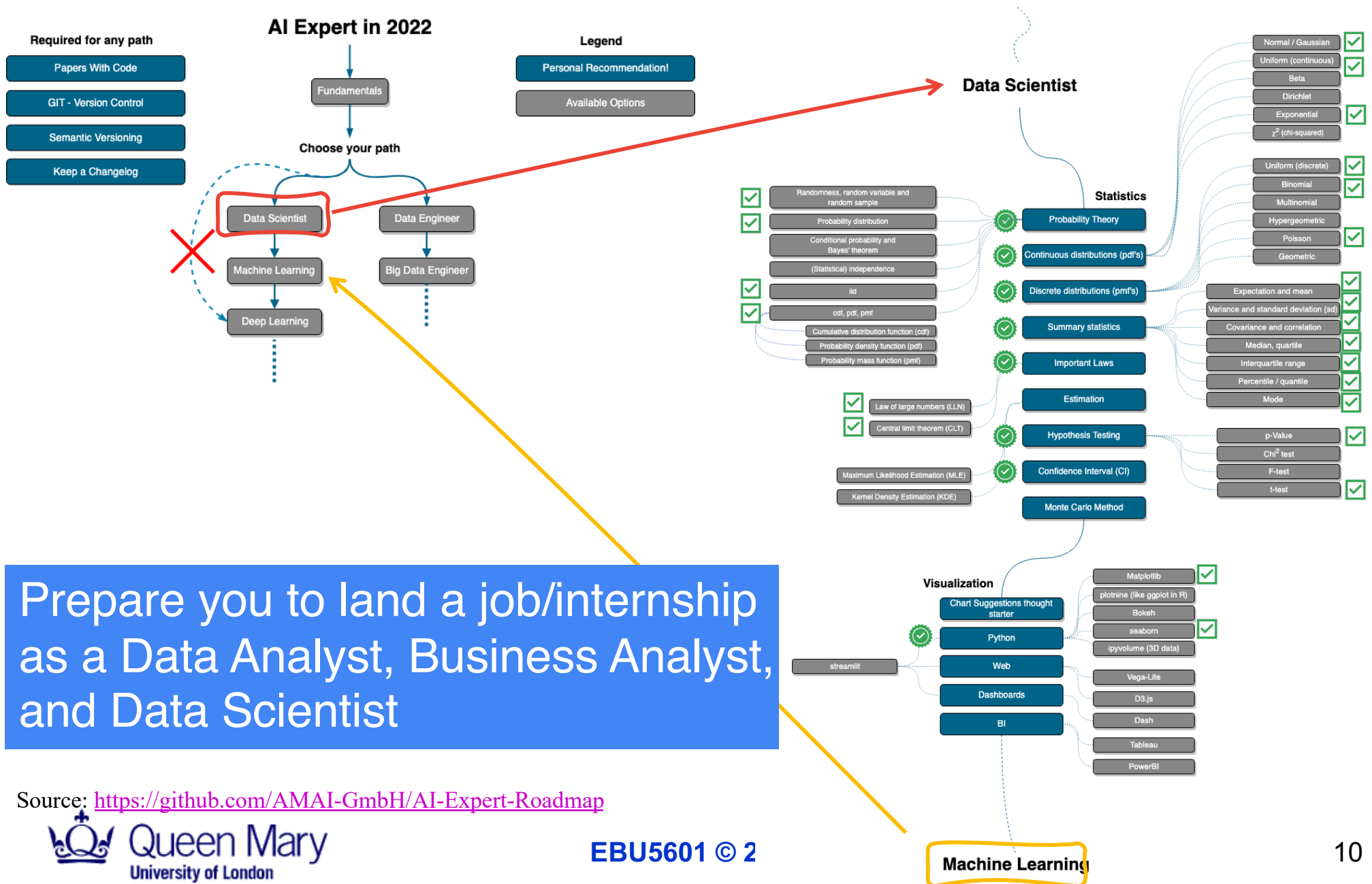
Statistic Models and  
Tests



Data Visualisations



# Data Science Roadmap



# Teaching Format

- Lectures
  - 45 mins x 10 per teaching block
  - Walking through lecture notes
- Tutorials
  - 45 mins per teaching block
  - Guided mini-projects, code templates
- Office Hours
  - 45 mins per teaching block
  - Ask your lecturer questions at International School (IS) Building 102
- In lecture notes you will see:

**Example:** *Solution provided in lecture notes and discussed in lectures*

**Exercise:** *Solution ONLY discussed in lectures*

**Demo:** *Codes demonstration using Jupyter notebooks in lectures*

# Teaching Schedule

- Block Teaching
  - 4 Blocks over 4 separate weeks

BUPT Week:	6	7	8	9	10	11	12	13	14	15
W/C:	00-Sep	07-Oct	14-Oct	21-Oct	28-Oct	04-Nov	11-Nov	18-Nov	25-Nov	02-Dec
8, 9, 12, 15			XL	XL			CS			CS

week8, 9, 12, 15		Mon	Tue	Wed	Thu	Fri
	08:00-09:35					
	09:50-11:25			3-417		
	11:30-12:15			IS 102 (OH)		
	13:00-14:35	3-417				3-417
	14:45-16:25					
	16:35-18:10		3-417		3-417	
	18:30-19:15				3-417(Tut)	
	19:20-20:55					

**Subject to further updates.**

# Assessment

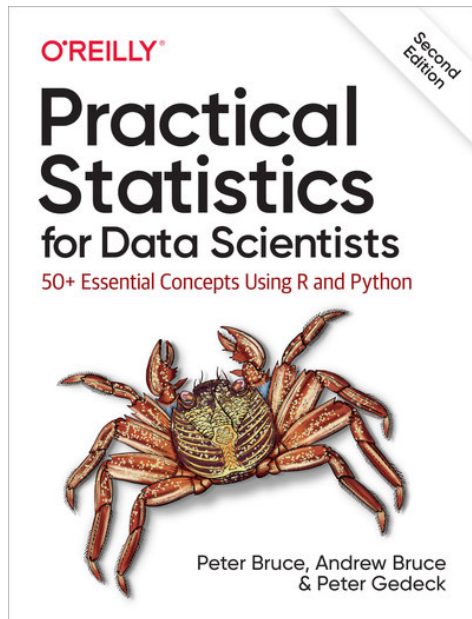
Summative Coursework	Items	Weights	Remarks
	Assignment	5%	Topic: Data Ethics Format: Group report Weeks: Week 10 –11 (2 weeks)
	Lab 1	15%	Topic: Analyse A/B test result Format: Individual project using Jupyter notebook Weeks: Week 13 – 14 (2 weeks)
	Lab 2	20%	Topics: Communicate data findings Format: Individual project using Jupyter notebook + Video presentation Weeks: Week 16 – 17 (2 weeks)
	Exam	60%	Covers everything - lectures, labs and other coursework except Data Ethics.

All other exercises, online quizzes, etc. are **formative** and they are useful learning resources even though no marks are given.

# Recommended Books

- Recommended (NOT Required):
- Available online via [QMUL library](#) website and [O'Reilly](#)

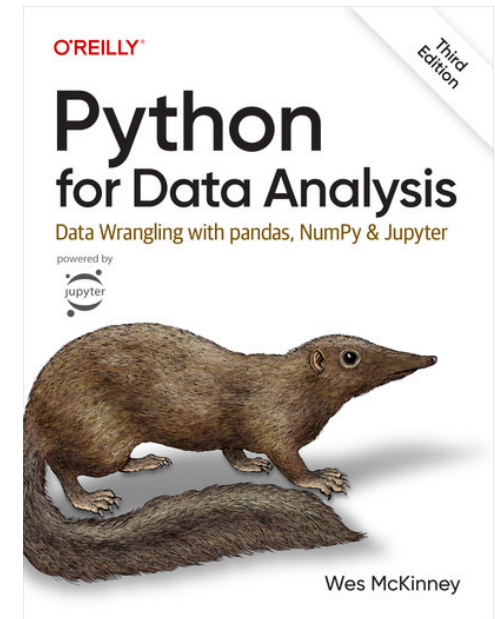
Peter Bruce, Andrew Bruce, Peter Gedeck, Practical Statistics for Data Scientists, 2nd ed. O'Reilly Media, Inc., May 2020.



Joel Grus, Data Science from Scratch: First Principles with Python, 2nd ed. O'Reilly Media, Inc., May 2019



Wes McKinney, Python for Data Analysis. O'Reilly Media, Inc., August 2022



# Module Representatives

- Call for Module Representatives (1):
  - Collect feedback from classmates about the lectures
  - Assist lecturers in disseminating important information
  - Attend ONE Student Staff Liaison Committees (SSLC) meeting in the semester
  - Volunteers (**one**) will be welcomed or chosen during the first teaching week of the module
  - Email to [chao.shu@qmul.ac.uk](mailto:chao.shu@qmul.ac.uk)
  - Or scan the QR code



Module Reps 2024-25 Sem1





# Quiz

## Case Study


- Read the job descriptor of a data scientist job from Monzo Bank in London, identify skills/techniques covered in this module (check the module LOs)



### Data Scientist

Monzo Bank · London, England, United Kingdom (Remote)

 Full-time · Entry level

 1,001-5,000 employees · Banking

### About the job

 Location | UK Remote | London |  £65,000 - £75,000 + Benefits

Go to

**www.menti.com**

Enter the code

**5993 6343**



Or use QR code

### Job Descriptor

- Applying your skills in quantitative analysis, data mining, and the presentation of data to see beyond the numbers and understand how our users interact with our products and how those insights can inform our product strategy
- Guide and enable product teams to measure things that matter; initiate or help run A/B experiments to keep improving everything we do
- Drive together with the finance team a unified company-wide understanding of the lifetime value of our users and how different product features are impacting user profitability
- Liaise with engineers to keep making sure we collect the right data to produce relevant business insights

Source: LinkedIn

# Questions

---

Use student forum on QM+

[chao.shu@qmul.ac.uk](mailto:chao.shu@qmul.ac.uk)

[xiaolanliu@qmul.ac.uk](mailto:xiaolanliu@qmul.ac.uk)