Queen Mary
University of London

ECS766P Data Mining

Week 12: Revision

Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

December 2021

School of EECS, Queen Mary University of London

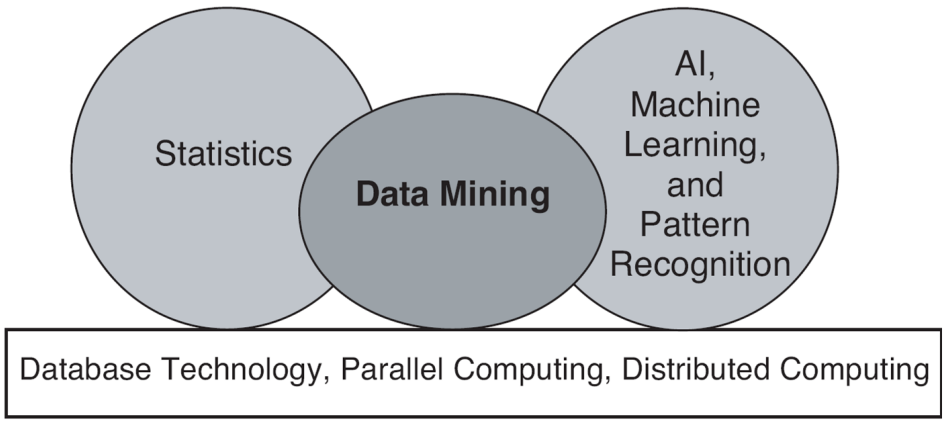
This week’s contents

1. Module revision
2. Sample exam questions

Module Revision

Week 1: Introduction

1. Why data mining?
2. What is data mining?
3. Models in data science
4. Data mining tasks
5. Challenges in data mining

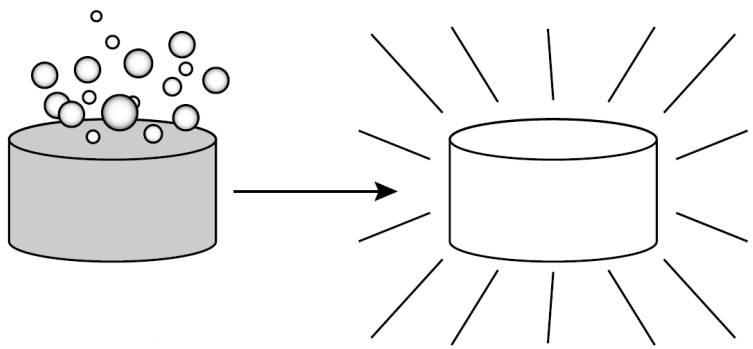


Week 2: Data

1. Attributes and Objects
2. Characteristics of Data
3. Types of Data
4. Data Quality
5. Basic Statistical Descriptions of Data
6. Similarity and Distance

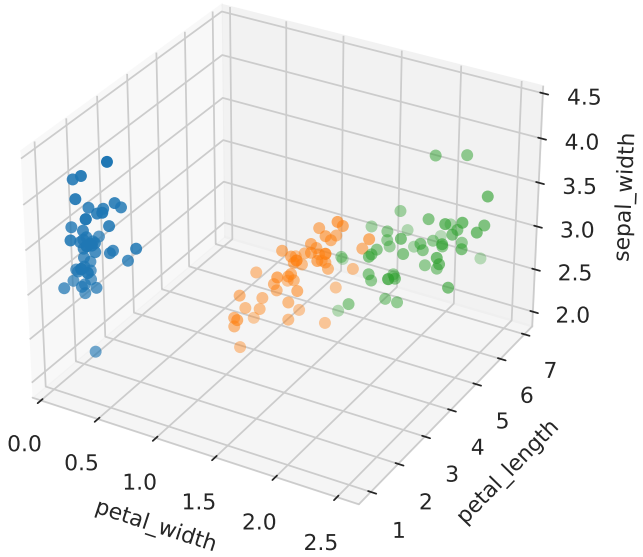
Week 3: Data Preprocessing

1. Data Preprocessing: An Overview
2. Data Cleaning
3. Data Integration
4. Data Reduction
5. Data Transformation and Data Discretisation



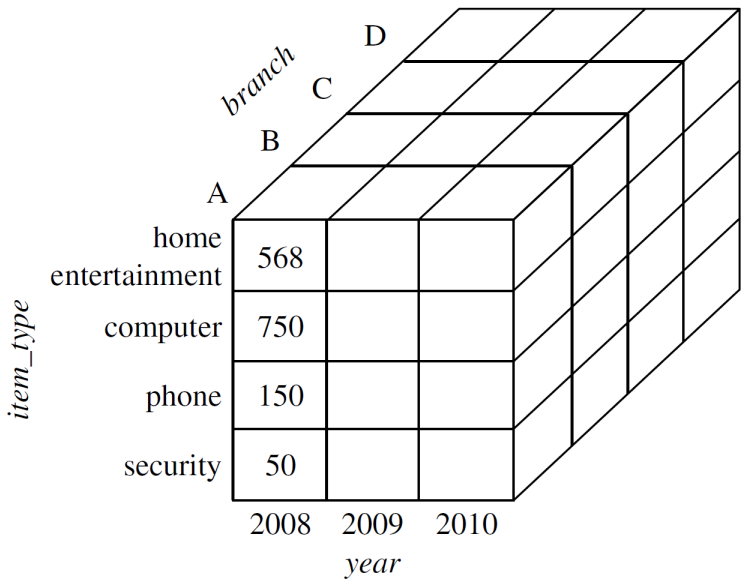
Week 4: Data Exploration and Visualisation

1. Data Exploration
2. Data Summarisation
3. Data Visualisation



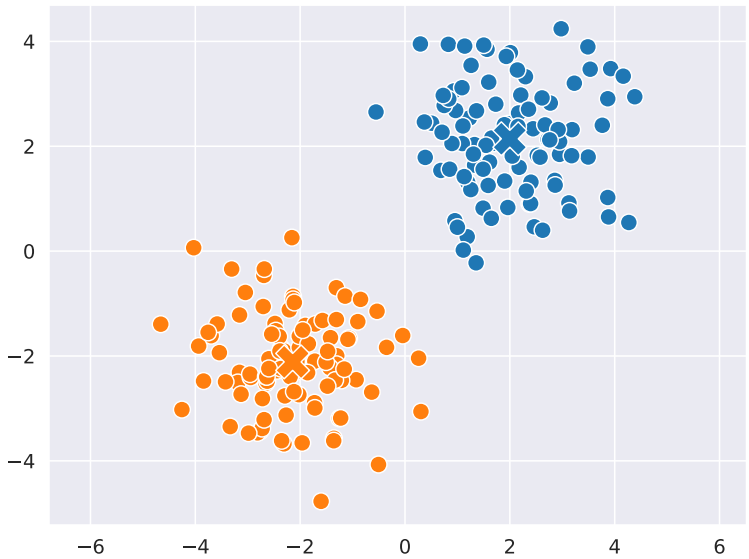
Week 5: Data Warehousing and On-line Analytical Processing

1. Data warehouse - basic concepts
2. Data warehouse modelling
3. Data warehouse design and usage
4. Data warehouse implementation



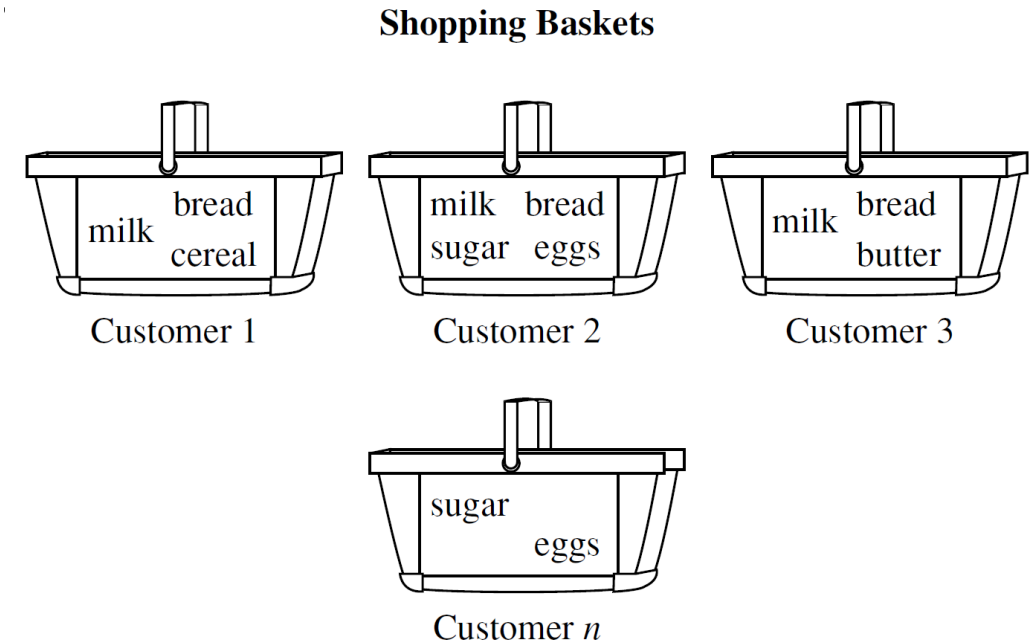
Week 6: Classification and Clustering

1. Classification
2. Clustering



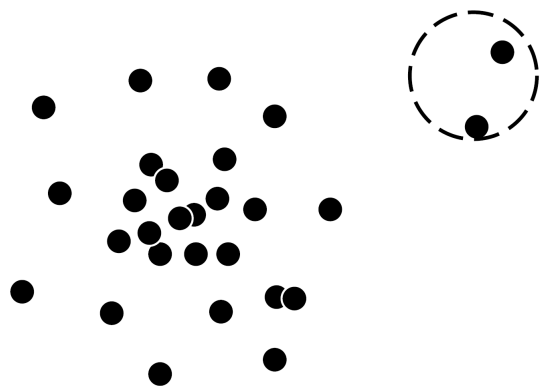
Week 8: Association Analysis

1. Frequent itemsets - basic concepts
2. Frequent itemset mining methods
3. Association rule mining
4. Pattern evaluation methods



Week 9: Outlier Detection

1. Outliers and Outlier Analysis
2. Outlier Detection Methods
3. Statistical Approaches
4. Proximity-Based Approaches
5. Clustering-Based Approaches
6. Classification Approaches
7. Mining Contextual and Collective Outliers



Week 10: Web Mining

1. Six Paradigms for Today's Internet
2. Technology Review
3. Internet Mining Applications
4. Ingesting Internet data
5. Search Engine Indexing & Ranking

Week 11: Data Mining Applications & Data Ethics

1. Mining Text Data
2. Mining Timeseries Data
3. Data Ethics

Sample Exam Questions

Question 1

Consider the below dataset:

Animal	Body mass [g]	Heart rate [bpm]	Order
Wild mouse	22	480	Rodentia
Rabbit	2.5×10^3	250	Lagomorpha
Humpback whale	30×10^6	30	Artiodactyla

For each attribute in the below dataset, describe its attribute type and justify your response.

Question 2

Assume an online store, where we need to capture the following information:

1. Information for all products currently in stock in the store in terms of product name, brand, quantity, and retail price.
2. Information on customer purchases, in terms of purchase number, items purchased, and total cost.
3. Information on total sales per week over the timespan of the last year.

Which data type and data sub-type (when applicable) should we use for each of the above cases? Justify your response.

Question 3

Consider the below datasets where we need to calculate the similarity or distance between objects:

1. A collection of documents, where we need to calculate document similarity with respect to the terms included in those documents.
2. A list of locations in a city specified by their coordinates, where we need to calculate the distance that a vehicle would need to travel between two locations.
3. Results of a customer survey, where participants are asked to agree or disagree with a series of statements; we would need to calculate the similarity between survey participants.

Which similarity or distance metric should we use in each case and why?

Question 4

Consider a dataset of sales for a shop, where we have information on customer IDs, customer postcode, payment method (e.g. credit card or debit card), and total amount of purchases per customer. Assume that the dataset is too large to analyse, and we would like to represent the dataset by a smaller sample. Suggest an appropriate method to obtain a smaller but representative dataset.

Question 5

Consider the below dataset which includes average temperature and pressure measurements for different altitudes:

Altitude (m)	Temperature (°C)	Absolute pressure (Hg)
0	15	29.92
610	11.1	27.82
1219	7.06	25.84
1829	3.11	23.98

What correlation relationship can be inferred between altitude and temperature? Justify your response.

Question 6

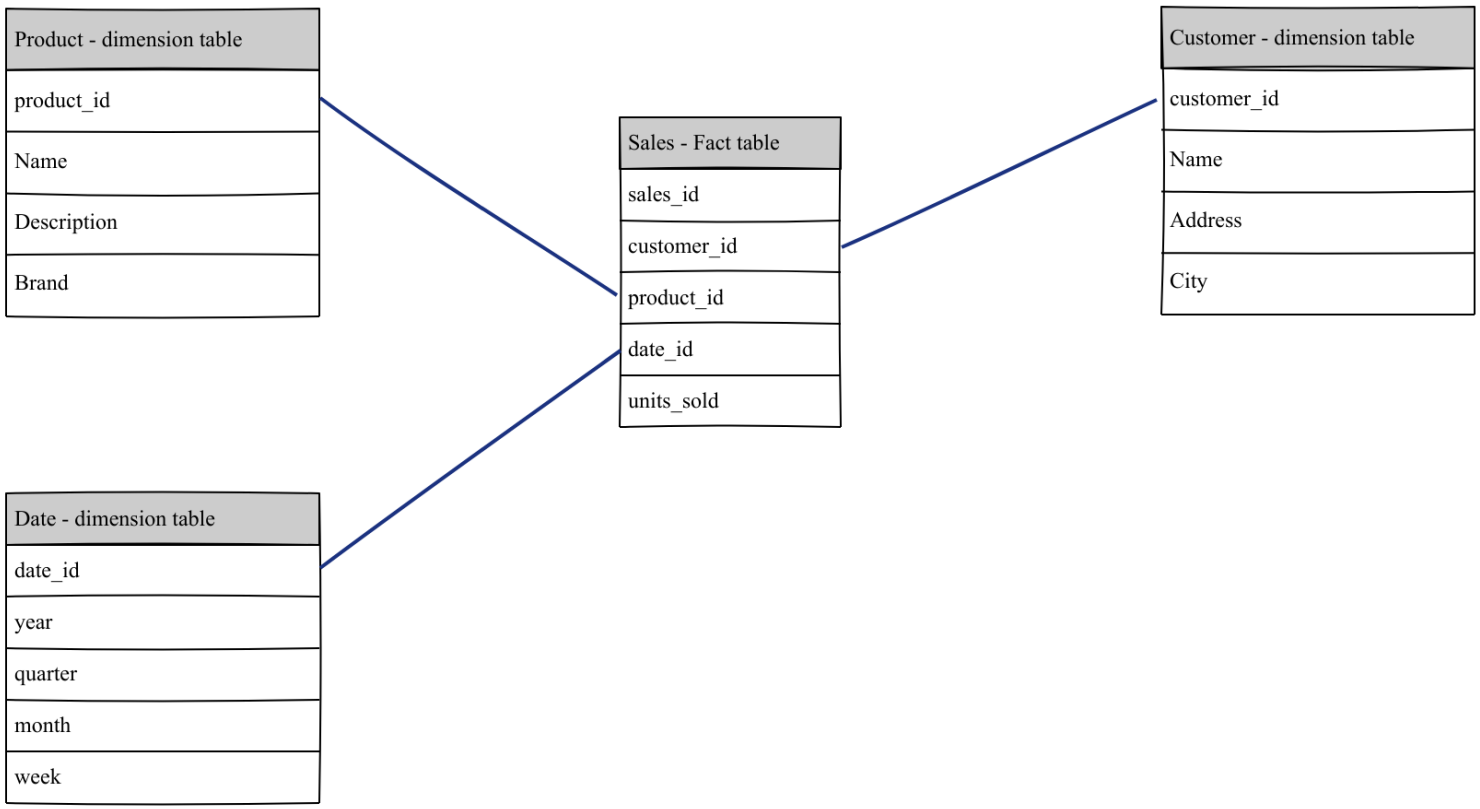
Consider the following dataset represented by a table:

ID	Feature 1	Feature 2	Feature 3	Feature 4
0	2	12	8	8
1	12	4	16	3
2	5	12	4	8
3	2	13	14	8
4	7	8	10	4
5	12	1	9	0
6	11	3	13	3
7	16	3	5	3

1. What are the modes of feature 4?
2. What is the 50th percentile of feature 3?
3. Draw an equal-width histogram with four bins for feature 2. Do not use code to generate the plot. Assume that the feature typically ranges from 0 to 20.
4. Create a scatterplot to visualise features 1 and 3. Do not use code to generate the plot. What can you say about the correlation coefficient between these features based on this visualisation? You do not need to compute the correlation coefficient.

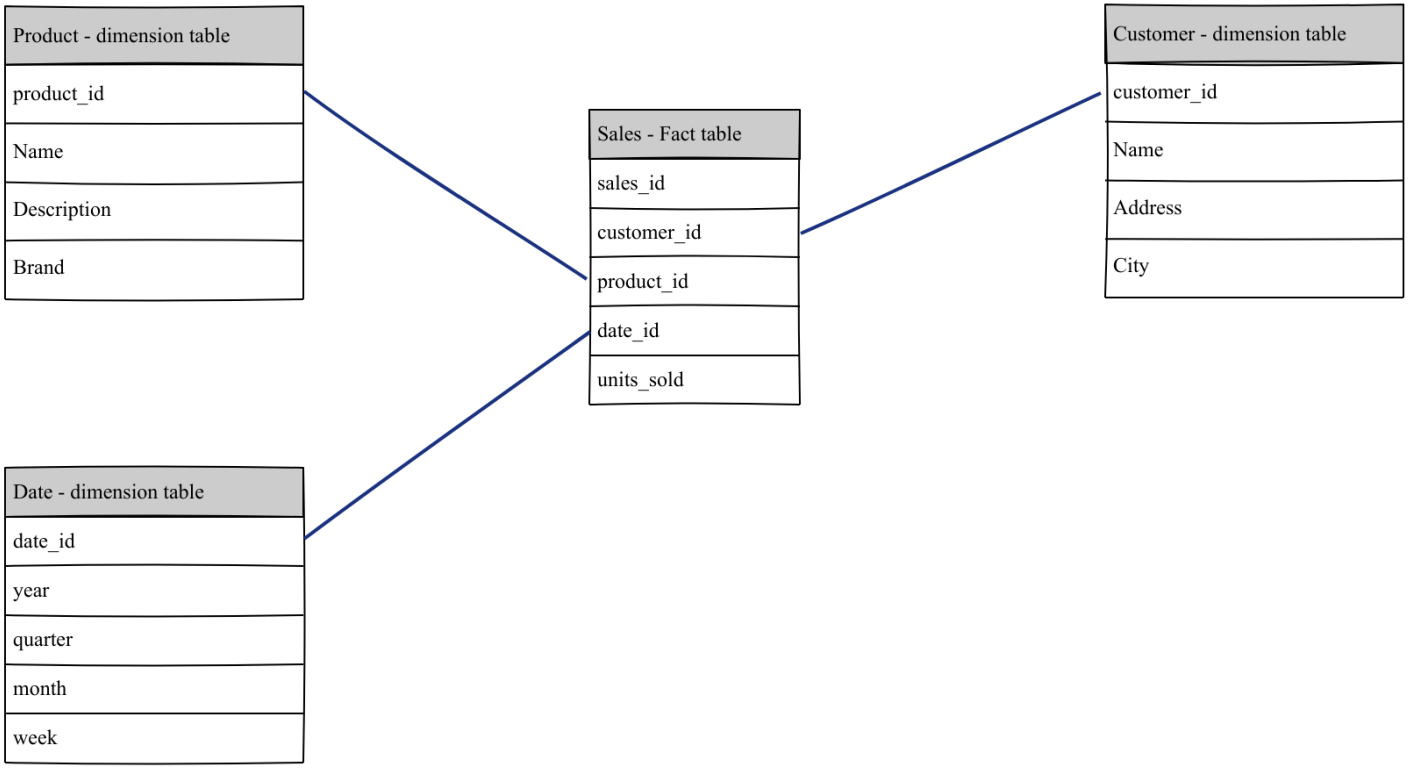
Question 7

Consider the below star schema representing the data warehouse model for a company:



The schema supports the following concept hierarchies: city > address > Name (for Customer); and year > quarter > month > week (for Date).

Question 7 (cont'd)



A query to be processed is on calculating the units sold for a given product for a given quarter at a given city. Assume that we have access to the following materialised cuboid: {Date = year, Customer = Name, product}. What specific OLAP operations should be performed?

Question 8

Answer the following questions regarding classification tasks:

1. How can a classification dataset be used to evaluate the capacity of a learning algorithm to produce classifiers that generalise to unseen data?
2. Consider the two types of errors (false positives and false negatives) in a binary classification task (a classification problem with two classes). Explain why one of these errors may be worse than the other.
3. Let f_1, f_2, and f_3 be three different classifiers trained on the same training set. Suppose that someone decides to use f_2 for a specific application because it obtains the highest accuracy on the training set. Explain why this methodology is flawed, and suggest a better alternative.

Question 9

Assume that the K-means algorithm is employed in a clustering task using a Euclidean distance between observations:

1. Suppose the cluster centers for a given iteration of the algorithm are $\boldsymbol{\mu}_1 = (4, 2)$ and $\boldsymbol{\mu}_2 = (-2, 4)$. To which of the clusters would the observation $\mathbf{x} = (2, -2)$ be assigned? Show your calculations.
2. Suppose a cluster center is associated to the observations $\mathbf{x}_1 = (5, 3)$, $\mathbf{x}_2 = (-5, 2)$, and $\mathbf{x}_3 = (3, 1)$ after the assignment step. What would be the new position of this cluster center after the movement step?

Question 10

Consider the following transaction dataset represented by a table:

TID	Transaction
1	$\{I2, I3, I4, I5\}$
2	$\{I1, I3\}$
3	$\{I2, I3, I5\}$
4	$\{I1, I4, I5\}$
5	$\{I1, I2, I4\}$
6	$\{I3, I5\}$

1. What is the support count of the itemset $\{I2, I4\}$?
2. Compute the Kulczynski measure of the itemsets $\{I1, I2\}$ and $\{I4\}$.
3. Suppose that an itemset $\mathcal{A}$ has n items. How many subsets does $\mathcal{A}$ have? Why can this be a problem for frequent itemset mining?
4. Explain why frequent itemset mining is often a preliminary step in association rule mining.

Question 11

Answer the following questions regarding outlier detection:

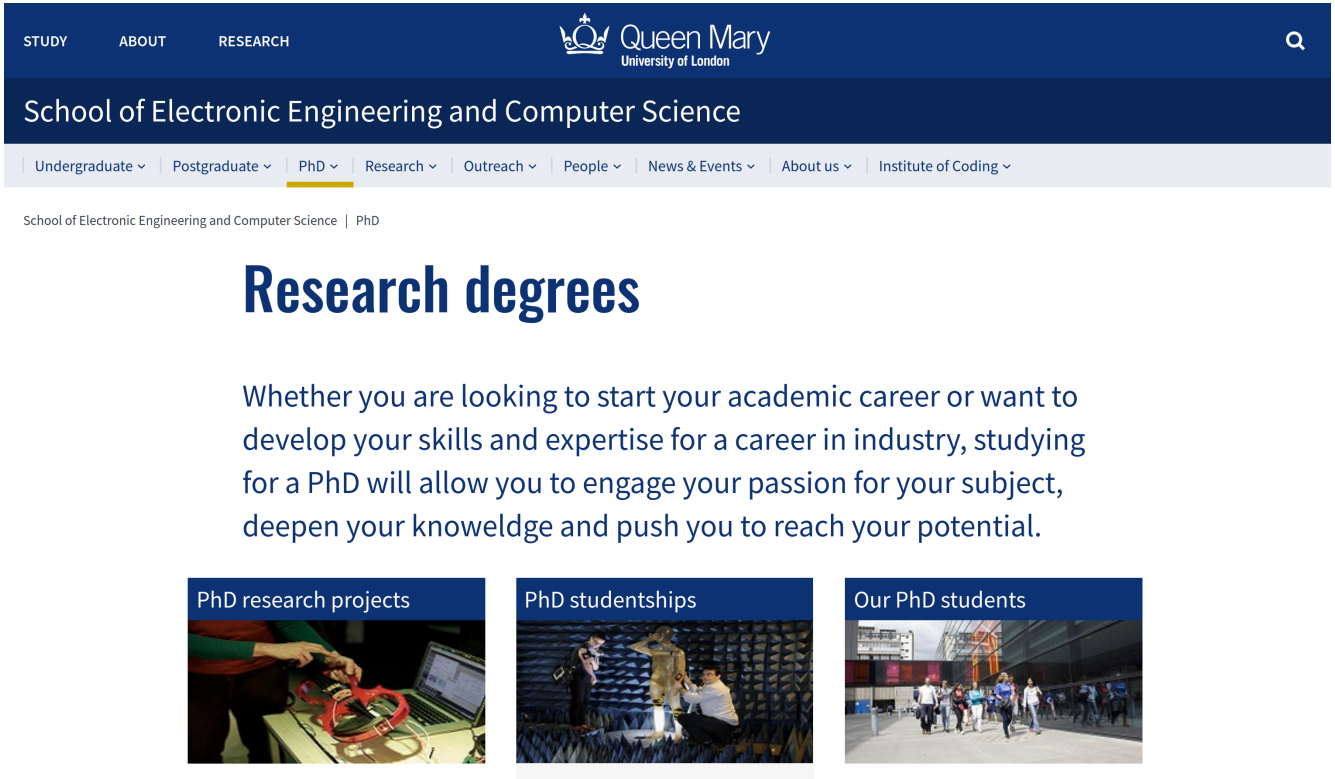
1. Explain the distinction between global outliers and collective outliers.
2. Consider the dataset $\mathcal{D} = \{5.63, 5.285, 2.785, 3.285, 3.535, 5.395, 6.605, 3.49\}$. Suppose that each observation was drawn independently from the same Gaussian distribution. Explain why a new observation $x = 11$ should be considered as an outlier.

Question 12

Answer the following questions regarding web mining:

1. During a run of a crawler algorithm, the process terminates when the frontier list is empty. Does this imply that the whole Web has been crawled?
2. Explain why frequency-based crawlers are often used in practice compared to other types of web crawlers.

Off-topic: PhD at EECS



http://eecs.qmul.ac.uk/phd/

http://eecs.qmul.ac.uk/phd/phd-studentships/

Thank you for attending Data Mining!