CST0
COMPUTER SCIENCE TRIPOS Part IA

Friday 12 June 2026 14:00 to 17:00

COMPUTER SCIENCE Paper 3

*Answer **one** question from each of Sections A, B and C, and **two** questions from Section D.*

*Submit the answers in five **separate** bundles, each with its own cover sheet. On each cover sheet, write the numbers of **all** attempted questions, and circle the number of the question attached.*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

STATIONERY REQUIREMENTS
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

(*a*) Dozens of computers around the globe coordinate with each other to implement a distributed database.

(*i*) Name two advantages and two disadvantages resulting from a distributed implementation. [4 marks]

(*ii*) Various policies are possible for deciding at which site(s) a data item is stored. Discuss two extreme possibilities and intermediate approaches. [4 marks]

(*iii*) How might law or security affect such distribution policies and would or should this influence the data model (i.e. the logical schema). [2 marks]

(*b*) Consider a graph database that is subject to updates once a day. It stores data as nodes with attributes connected by arcs that posses further attributes.

(*i*) Discuss whether the distinctive feature of a graph database is the underlying data model (i.e. the logical schema) or its query language, or anything else. [5 marks]

(*ii*) Often there are queries that must compare today's data with that from a month ago and compute metrics on the intermediate changes. Discuss how this might normally be solved and whether a graph database is intrinsically better for this sort of query. Consider using a special form of graph edge. [5 marks]

2 Databases

Tattersalls sells horses, one at a time, by auction. A fresh rDBMS table is created every auction day, with a new record entered each time the hammer falls. The fields are: Date, Time, LotNumber, SellerID, SellerName, HorseName, HorseBreed, HammerPrice, BuyerName, BuyerID.

(*a*) Draw a suitable E/R diagram for this data. Note that sellers and buyers are drawn from the same community of people and both pay a fixed-rate commission to the auction house based on the HammerPrice. Use underlining to denote primary keys and state any needs for synthetic keys or assumptions on recyling of values, such as LotNumber or HorseName. [4 marks]

(*b*) An expert says the resulting table is 'excessively denormalised'. What aspects might they be referring to and why might denormalised data be helpful in this situation? Give a normalised schema which uses several rDBMS tables, but that sticks to their preferred data handling approach of always creating at least one fresh table per auction day. [4 marks]

(*c*) A report is needed with columns HorseBreed and TotalSales that totals the value of sales on a given day for each breed. Give suitable SQL that uses your normalised schema. [3 marks]

(*d*) A second report of TotalSales per breed is needed, formatted as an rDBMS table that has one row for each auction date and one column for each breed of horse. For either of the above schemas, give two reasons why this might be hard to code using simple SQL (*i.e.* SQL that maps directly to the Relational Algebra) and explain what changes or 'fudges' are needed to achieve it within that subset. [4 marks]

(*e*) The rDBMS being used supports many advanced features, such as 'pivot' and meta-programming operations that are, allegedly, elegant and easy-to-use. These could overcome the 'fudges' needed for part (*d*). How might one of these work? Would the most elegant solution be to use a better schema to start with? [5 marks]

SECTION B

3 Introduction to Graphics

(a) Explain the purpose of near- and far-clipping planes in Z-buffer algorithm and how their choice can increase or reduce the possibility of Z-fighting artifacts. [5 marks]

(b) Given the chromaticity coordinates x and y and luminance Y, give the equations for trichromatic coordinates X and Z. [5 marks]

(c) You used the latest Large Language Model to generate shader code using the following prompt:

"Generate both vertex and fragment shaders in GLSL. The shaders should transform vertices and normals using the model matrix M, the view matrix V, and the projection matrix P. Each pixel is shaded using the Phong reflection model, assuming a single point light source at fixed coordinates."

However, the resulting code does not work as expected. Fix all 4 bugs in the code, rewritting affected lines of code and explaining the nature of the bug. You may need to modify more than 4 lines. You can use line numbers and the name of the shader (vs/fs) to refer to the lines, e.g., `vs:3` for the vertex shader, line 3. The generated code does not contain any syntax errors. [10 marks]

```
1: #version 330 core

2: in vec3 aPos;
3: in vec3 aNormal;

4: uniform mat4 uModel, uView, uProj; // M, V, P
5: out vec3 vNormal, vFragPos;

6: void main()
7: {
8:    vec4 worldPos = uView * uModel * vec4(aPos, 1.0);

9:    vNormal = normalize((uModel *
                           vec4(aNormal, 1.0)).xyz);
10:   vFragPos = worldPos.xyz;

11:   gl_Position = uProj * uView * worldPos;
12: }
```

```
1: #version 330 core
// Normal and position in world space
2: in vec3 N, vFragPos;
3: out vec4 FragColor;

// light and camera in world space
4: uniform vec3 uLightPos, uCameraPos;

5: uniform vec3 uAmbientColor, uDiffuseColor,
               uSpecularColor;
6: uniform float uShininess;

7: void main()
8: {
9:    vec3 L = normalize(uLightPos - vFragPos);
10:   vec3 V = normalize(uCameraPos - vFragPos);
11:   vec3 R = reflect(-L, N);

12:   float diff = dot(N, L);
13:   float spec = pow(dot(R, V), uShininess);

14:   vec3 ambient  = uAmbientColor;
15:   vec3 diffuse  = uDiffuseColor  * diff;
16:   vec3 specular = uSpecularColor * spec;

17:   FragColor = vec4(ambient + diffuse +
                       specular, 1.0);
18:}
```

4 Introduction to Graphics

An ellipsoid centered at the origin and with semi-axis lengths a, b, and c that are aligned with the axes of the coordinate system has an implicit equation:

$$(P\,E)\cdot(P\,E)=1 \tag{1}$$

where $P=\begin{bmatrix}x & y & z\end{bmatrix}$ is a point in 3D space, E is a diagonal matrix with the elements $\frac{1}{a}$, $\frac{1}{b}$, and $\frac{1}{c}$, and $\cdot$ is the dot product.

(a) Derive an equation for the intersection of the ellipsoid from Eq. 1 with a ray, given by the origin O and direction D. [6 marks]

(b) Use transformation matrices in homogeneous coordinates to find an implicit equation of an ellipsoid centered at point M, and with semi-axes given by line segments $\overline{MA}$, $\overline{MB}$, $\overline{MC}$. All semi-axies are orthogonal to each other, i.e., $\overline{MA}\cdot\overline{MB}=0$, $\overline{MB}\cdot\overline{MC}=0$, $\overline{MA}\cdot\overline{MC}=0$. You can use operator $T(x)$ to denote the translation (translation by vector x), and $R_x(\omega)$, $R_z(\theta)$, $R_y(\phi)$ to denote the rotation by each axis. There is no need to write full matrices. [10 marks]

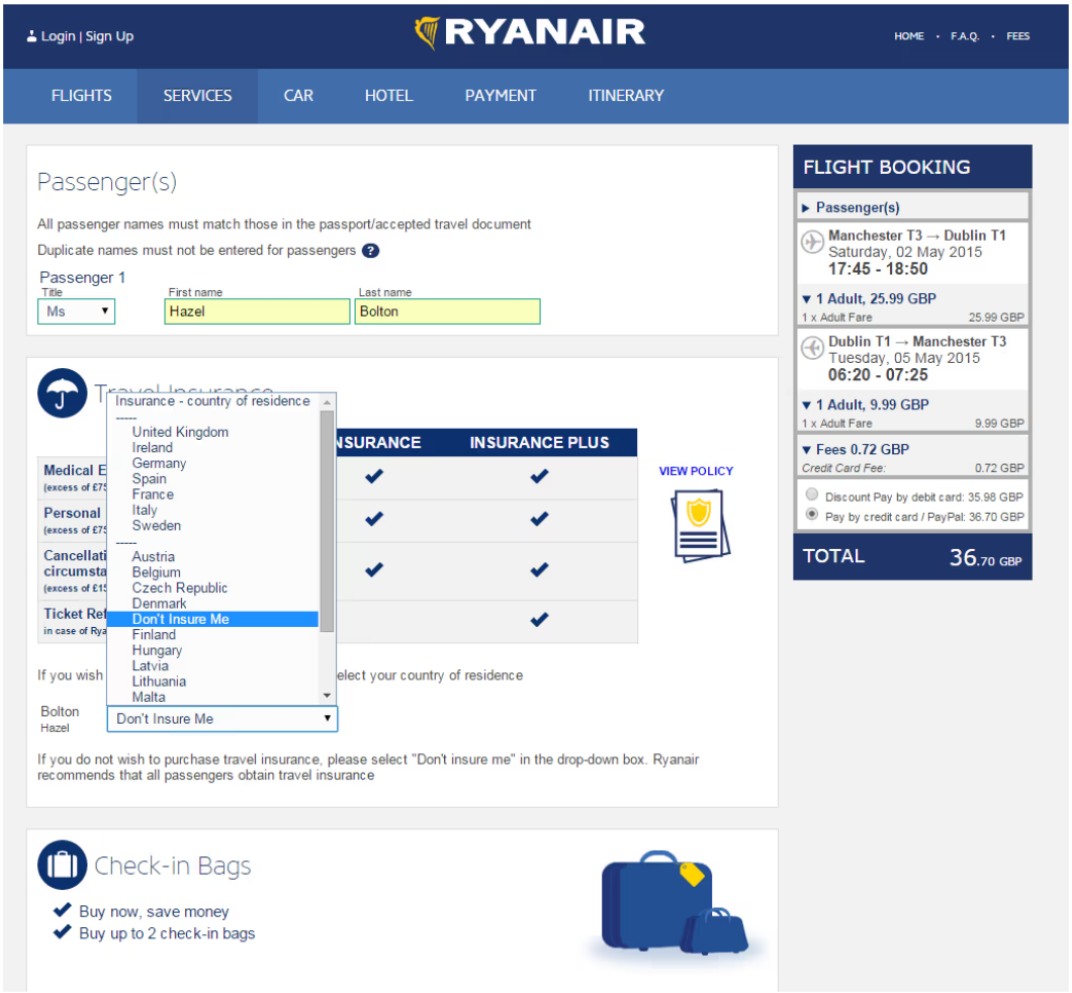
(c) Derive an equation for the intersection of the ellipsoid from Part (b), given by M, $\overline{MA}$, $\overline{MB}$ and $\overline{MA}$, with the ray $\overrightarrow{OD}$. [4 marks]

SECTION C

5 Interaction Design

The term *deceptive patterns* (sometimes referred to as "dark patterns") was coined by Harry Brignull in 2010 to refer to design patterns that prompt users to take an action that benefits the company employing the pattern by deceiving, misdirecting, shaming, or obstructing the user's ability to make another (less profitable) choice. Deceptive patterns harm users by causing financial loss, loss of privacy, and legal control.

(*a*) Identify one deceptive design pattern on the RyanAir web page given in the figure below and explain why it constitutes a deceptive pattern. [2 marks]



(*b*) Describe how you would go about evaluating this RyanAir web page to find out if the identified aspect does indeed constitute a deceptive design pattern or not. Your description should include which research method you expect to use and why, a brief summary of how you would go about using the method in this specific case, what data you expect to collect, and how you plan to analyse it.

[8 marks]

(*c*) Re-design this RyanAir web page so that its design is improved and it no longer contains the deceptive pattern you identified above. Explain what/how the redesign changed the web page. [4 marks]

(*d*) Discuss the statement: 'The widespread use of deceptive patterns is partly driven by the prevalence of Nielsen's Heuristic #4 — Consistency and Standards.' Consider the argument from both supporting and opposing perspectives. [6 marks]

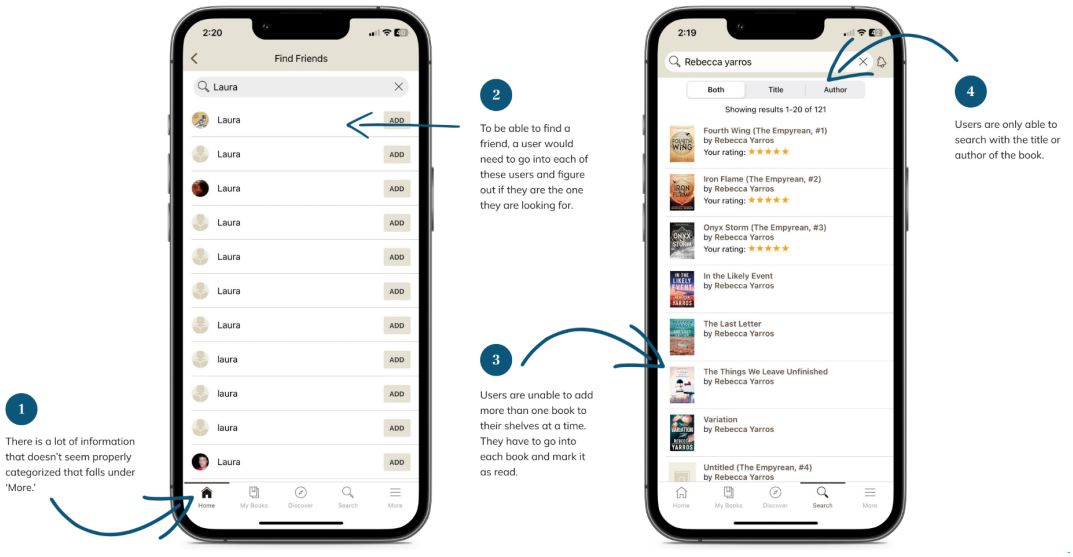
6 Interaction Design

Goodreads is a social app for book lovers. It allows users to track their reading progress, shelve books into different categories, share reviews and interact with other users. Users have been sharing negative feedback of the app in various mediums, including app reviews and social media. An in-lab user study was conducted where the participants were asked to undertake three primary tasks on the Goodreads website:

- **Task 1:** Participants created an account and were asked to find a book they would personally enjoy, aiming to understand how first-time users navigate the site.
- **Task 2:** Participants completed mini-challenges, which involved using different GoodReads functions such as 'Similar Books', 'Favorite Genres', 'Explore', 'Lists', and 'Community', to better understand the site's various sections.
- **Task 3:** Participants used the platform to find a book recommendation for a friend's birthday, helping to analyze the platform's social and recommendation features.

(*a*) Considering the tasks above, identify 3 goals for this user study and the corresponding data that you will be collecting/recording from the users, and which metrics you will use to analyse this data. Provide your answer as a 3x3 table format for clarity and legibility. [6 marks]

(*b*) Identify two distinct methods to analyze the results of this user study on GoodReads, and explain how you will use these methods for data analysis. [4 marks]



(*c*) An analysis of users feedback for GoodReads mobile app identified a variety of

areas for improvement, as shown in the figure above and listed below.

1. There is a lot of information that doesn't seem properly categorized that falls under 'More'.

2. To be able to find a friend, a user would need to go into each of these users and figure out if they are the one they are looking for.

3. Users are unable to add more than one book to their shelves at a time. They have to go into each book and mark it as read.

4. Users are only able to search with the title or author of the book.

Considering the four issues identified above, create the illustration of a user journey map, and map the user journey from discovery to retention and highlight key pain points experienced by users at different stages. [2 marks]

(*d*) Considering the two screens provided and the four issues identified above in (c), re-design two screens by utilising the various design principles and guidelines you learned in the lectures to address at least two of the four issues. For each screen redesigned, indicate clearly which principle or guideline you used and how. [8 marks]

SECTION D

7 Machine Learning and Real-world Data

You are interested in developing a system to recommend music genres to users according to their moods. As different users have different tastes, each of them provides you with a playlist containing the genre ((C)lassical, (R)ock, (J)azz, (E)lectronica, (H)ip-hop) listened to and their mood (U)pbeat, (D)ownbeat, (N)eutral). Here is an example from a user:

```
Mood   U  U  N  D  D  N  N  U  U  ...
Genre  R  R  J  C  J  J  E  E  H  ...
```

Your manager tells you that you should model this data using a first order hidden Markov model (HMM), with mood as the hidden states and genre as the observed states.

(*a*) Define and estimate the parameters of the HMM using the playlist above. Provide equations as needed. Include the start state in your calculations, but not the end state as the example continues in perpetuity. No smoothing should be applied. [4 marks]

(*b*) What are the two assumptions made by the HMM? Are they appropriate in this application? [4 marks]

(*c*) Using the model you estimated above, predict the most likely music genre to continue this playlist. [4 marks]

(*d*) Using the model you estimated above, what is the probability of following Rock immediately with Classical? [2 marks]

(*e*) What would happen if you used the HMM you have trained to generate a playlist? Could you generate different playlists? Give reasons for your answer. [2 marks]

(*f*) Imagine you now define finer genres, e.g. by splitting rock into classic rock vs heavy metal. Which advantages and disadvantages do you foresee? [2 marks]

(*g*) What would happen if you replaced moods with the user's own reactions (e.g. likes, faves, etc.) and learned the parameters of the HMM ? Would the modelling assumptions of the HMM still be valid? [2 marks]

8 Machine Learning and Real-world Data

Two systems are classifying documents into relevant to a particular topic (e.g., "remedies against hair loss"), or not. System A is given 5 documents. It classified all as relevant, but only 4 of these are relevant according to a human. System B is given 20 documents. It classified all as relevant, but only 5 are indeed relevant.

(a) Define precision, recall and accuracy and calculate these for Systems A and B. [4 marks]

(b) You now find out that the situation was actually quite different: The systems were asked to make relevance decisions on documents from a large pool (1 million documents), and System A classified the 5 documents mentioned above from the document set as relevant, and the rest as irrelevant; in a similar way, System B classified the 20 documents mentioned above as relevant. You also learn that this topic had 6 relevant documents in total. With this information, how do your measurements of precision, recall and accuracy change? [2 marks]

(c) In a situation such as the one in Part (b), which of the three evaluation metrics accuracy, precision and recall are appropriate, and why? [2 marks]

(d) Consider now the four systems A, B, C, D in the table below, which ranked documents according to relevance, for the topic introduced in Part (b). The table shows the top 20 ranks. X stands for "relevant document".

	Rank																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	–	–	X	–	–	X	–	–	–	X	X	–	–	–	–	–	–	–	X	–
B	X	X	–	X	–	–	–	X	–	–	–	–	–	–	–	–	–	X	–	X
C	–	–	–	–	X	–	–	–	–	–	–	–	X	X	–	X	–	–	X	X
D	X	X	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	X	X	X

Precision at rank 10 ($P_{Rank10}(A)$) and recall at rank 10 ($R_{Rank10}(A)$) are two metrics which are calculated at rank 10. Calculate these for systems A–D. [3 marks]

(e) Two new metrics are proposed: $Precision@Recall = 0.5$, abbreviated $P@R = 0.5$, is defined to be the precision at the rank where recall first reaches 50%; $Recall@Precision = 0.5$, abbreviated $R@P = 0.5$, is defined to be the rank where precision first reaches 50%. Calculate these metrics for the systems A–D. (Tip: it is not guaranteed that these measurements can be taken in all cases.) [4 marks]

(f) What is problematic about the metrics from Part (d), and what is problematic about the metrics from Part (e)? Which properties would an ideal metric for measuring classification performance in the context of extremely imbalanced tasks for ranked systems have? [5 marks]

9 Machine Learning and Real-world Data

You want to classify sentence pairs of sentences $S1$ and $S2$ as to whether their meaning is identical or not. You have collected some sentence pairs and labels assigned by humans, as below.

Sentence pair $\langle S1, S2 \rangle$		**Class**
1) Mini-poodle feeds man delicious acorns.	Man feeds dog tasty acorns.	DIFFERENT
2) Elsa didn't serve chicken today.	Chicken wasn't cooked by Elsa.	SAME
3) Dog was bitten by squirrel.	Squirrel attacked tiny dog violently.	SAME
4) This school happily admits 6-year-olds.	This school admits no 6-year-olds.	DIFFERENT

(a) As features, you want to use the tokens contained in the sentence pair. Give the formula for the Naive Bayes Classifier, defining all parameters in the formula, and explain how this formula can be applied to the situation above. [3 marks]

(b) What is the main disadvantage of using tokens as features here? [3 marks]

(c) In an alternative approach to feature extraction, we derive features for each sentence $S1$ and $S2$, expressing how long it is (F1, F2), how many negation tokens such as *not, none, nobody...* it contains (F3, F4), and whether it contains passive voice (F6, F7). For each sentence pair $\langle S1, S2 \rangle$ we report the number of shared tokens between $S1$ and $S2$ (F7), and whether the sentences have the same main verb (F8) and the same subject (F9), as shown in the table below.

	F1 S1 length	F2 S2 length	F3 S1 neg.	F4 S2 neg.	F5 S1 passive	F6 S2 passive	F7 overlap	F8 same verb	F9 same subj
a	5	5	1	1	N	Y	2	N	N
b	5	5	0	1	N	N	4	Y	Y
c	5	5	0	0	N	N	3	Y	N
X	?	?	1	2	N	N	5	N	Y

(i) Feature sets a,b,c were derived from 3 of the sentence pairs 1)–4) above, but the correct assignment got lost. Assign each feature set to its sentence pair. [2 marks]

(ii) For the sentence pair from the table above that remains non-aligned, provide the values for features F1–F9. [2 marks]

(iii) Write a sentence pair that corresponds to feature set X. You can freely decide how long the sentences are. The class should be SAME. [3 marks]

(d) State the central assumption behind the NB classifier, and whether you think it holds in the scenario in Part (c). If so, explain why. If not, provide a counter example. [3 marks]

(e) Give two ways how the classifier in Part (c) might be improved. [4 marks]

END OF PAPER