UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Deadline: 4pm Thursday 17 March 2016

Submit to box outside ITO office in Forrest Hill

This paper contains Data & Analysis exam questions from 2013 and 2015. It is being released on Thursday 3 March 2016 as a written coursework assignment. You have **two weeks** to complete this assignment. It will not necessarily take that long, but the time is there to help you schedule against other assignment loads from your different courses. The original exam time was two hours.

Questions 1 and 2 use only material already covered in the course so far. Question 3 requires material that will be covered in Lectures 15 and 16 during Week 8 of semester. The real exam is based on content from throughout the lecture course.

Submit your solutions on paper to the labelled box outside the ITO office in Forrest Hill by **4pm Thursday 17 March 2016**. Please ensure that all sheets you submit are firmly stapled together, and on the first page write your name, matriculation number, tutor name, tutorial group number, and the course code INF1-DA. If these are not clearly stated then your work will not reach your tutor and may not be marked.

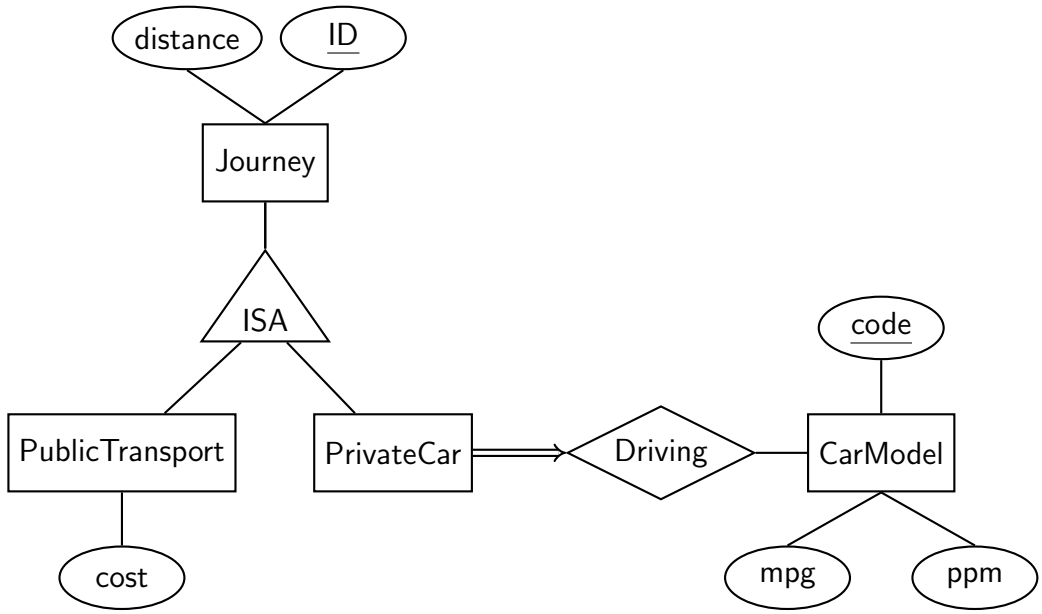
Your tutor will mark your work and return it to you in your Week 11 tutorial, with written and verbal feedback. However, these marks will not affect your final grade for Inf1-DA — this *formative* assessment is entirely for your feedback and learning. Because of this you can freely share help on the questions, discuss on *Piazza*, and talk about your work with other students. Please do.

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **CALCULATORS MAY BE USED IN THIS EXAMINATION.**

1. [*This question is worth a total of 35 marks.*]

The entity-relationship diagram below shows a fragment of a proposed model for a database managing travel claims by university staff. Each journey is labelled with a unique ID, and journeys by car or by public transport have different information recorded.



You also have the following information about the model.

- Attributes distance, cost, mpg (miles per gallon fuel consumption) and ppm (pence per mile travel allowance) are represented as integers.
- All other attributes are alphanumeric values up to 8 characters long.
- Every instance of the CarModel entity must have mpg and ppm values provided.

(a) Draw up an SQL data declaration of appropriate tables to implement this entity-relationship model. (You need not include **on delete** declarations). [*20 marks*]

QUESTION CONTINUES ON NEXT PAGE

Another part of the same database is to contain information about parking provision for cars and bikes at different campuses around the University. This includes the following three SQL tables.

```
create table Parking (
  ID          varchar(10),
  type        varchar(10) not null,
  number      integer not null,
  location    varchar(16) not null,
  primary key (ID),
  foreign key (type) references ParkingType(code),
  foreign key (location) references Campus(name)
)
```

Typical Parking record: {"AHT_FRONT","SHEF_WIDE","3","BioCampus"}

This records that parking area AHT_FRONT on the BioCampus has 3 parking spots of type SHEF_WIDE.

```
create table ParkingType (
  code        varchar(10),
  name        varchar(80) not null,
  capacity    integer not null,
  mode        varchar(6) not null,
  primary key (code)
)
```

Here the transport mode is either bike or car, and a typical ParkingType record would be {"SHEF_WIDE","Sheffield Rack, double width","12","bike"}

This records that a SHEF_WIDE parking spot is a "double-width Sheffield Rack" with a capacity of 12 bikes.

```
create table Campus (
  name    varchar(16),
  primary key (name)
)
```

The Campus table contains rows like "BioCampus", "Central Area" and "Sandwell Site".

(b) The Campus table is referenced from Parking, but only has one column. Why might it be helpful to have a "relation" like this with just a single field? [*2 marks*]

(c) Give SQL queries, using the tables above, that carry out the following.

(i) List the names of every different type of bicycle parking in the database.

(ii) List, without duplication, all university campuses where there is parking for cars.

(iii) Compute how many space for bike parking there are on the BioCampus. [*13 marks*]

2. [*This question is worth a total of 35 marks.*]

The following small XML document is a marked-up version of a speech from one of Shakespeare's plays.

```
<speech speaker="First Witch">
    <line>
        <w>When</w>
        <w>shall</w>
        <w>we</w>
        <w>three</w>
        <w>meet</w>
        <w>again</w>
    </line>
    <line>
        <w>In</w>
        <w>thunder</w>
        <punct>,</punct>
        <w>lightning</w>
        <punct>,</punct>
        <w>or</w>
        <w>in</w>
        <w>rain</w>
        <punct>?</punct>
        </line>
</speech>
```

(a) Draw this XML document as a tree, following the XPath data model. [*9 marks*]

(b) Write an XML DTD for a Speech document type to validate such speeches. Assume that every speech must have an identified speaker. [*12 marks*]

(c) Suppose a large XML document contains many such speeches, nested at various levels inside Plays, Acts, Scenes and so forth. Write XPath expressions to identify:

(i) All lines spoken by Macbeth

(ii) All speakers using the word "blood" in a speech. [*8 marks*]

(d) The lines above come from the works of a single author. Standard corpora for linguistic research like the *British National Corpus* or the *Penn Treebank* bring together work from many sources. Building them requires balancing and sampling in order to ensure that they are representative.

Explain the meaning of *balancing*, *sampling* and *representative* here. [*6 marks*]

3. [*This question is worth a total of 30 marks.*]

The standard "information retrieval (IR) task" is, given a *query* and a collection of *documents*, to find those documents that are relevant to the query.

(a) One measure of performance of an information retrieval algorithm is its *precision* P, usually computed using the following formula.

$$P = \frac{TP}{TP + FP}$$

State the matching formula for the other common performance measure, *recall* R. [*2 marks*]

(b) Explain the following abbreviations used in these formulae — for each one, state what it stands for, and describe in a single sentence what it means.

$$TP, FP, TN, FN$$

[*8 marks*]

(c) You are evaluating information retrieval systems to use with a collection of 2000 documents from an archive of 20th-century material. There are two candidate systems: *Happy* and *Sleepy*. Each is tested on the query "Enigma machine", for which there are 150 relevant documents in the archive.

Happy returns 600 documents, with 100 of those being relevant; *Sleepy* returns only 25, but all of them are relevant.

For each system, draw up a table of retrieval against relevance from this data, and use it to calculate precision and recall on this test. Show your working. [*8 marks*]

(d) The F_α performance measure combines precision and recall using the formula

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}\,.$$

Compute $F_{0.8}$ for *Happy* and *Sleepy*. Which performs more strongly?

The *balanced* F-score uses $alpha = 0.5$. Does the choice of $\alpha = 0.8$ favour systems that are good at precision, or recall? [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

A simple document retrieval algorithm might look at whether certain words appear in a document. A more sophisticated algorithm might weigh those words by their significance, using a a measure such as *term frequency - inverse document frequency* (tf-idf).

One of the items in the 2000-document collection described earlier is identified as *Report Q*. You have the following information about word occurrences in this document, and across the collection.

- The word "Turing" appears 590 times in the collection, across 85 different documents; it appears 8 times in Report Q.
- The word "computer" appears 22 times in Report Q, and a total of 1700 times in the whole collection — it is in 400 of the 2000 documents.
- Of all the documents, 1600 contain no mention at all of either "Turing" or "computer".

(e) Use these figures to compute tf-idf for both "Turing" and "computer" in Report Q. (Note: you don't need all of the numbers above for this calculation, just some of them.) [*6 marks*]

(f) Which word is most characteristic of Report Q, "Turing" or "computer"? Explain your answer. [*2 marks*]