UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Deadline: 4pm Thursday 20 March 2014

Submit to box outside ITO office on Appleton Tower level 4

This is the Data & Analysis exam from April 2013. It is being released on Thursday 6 March 2014 as a written coursework assignment. You have **two weeks** to complete this assignment. It will not necessarily take that long to complete, but the time is there to help you schedule against other assignment loads from your different courses. The original exam time was two hours.

Submit your solutions on paper to the labelled box outside the ITO office on Level 4 of Appleton Tower by **4pm Thursday 20 March 2014**. Please ensure that all sheets you submit are firmly stapled together, and on the first page write your name, matriculation number, tutor name and tutorial group.

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, ask your tutor for advice, and discuss your work with other students.

INSTRUCTIONS TO CANDIDATES

ALL QUESTIONS ARE COMPULSORY.

DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.

CALCULATORS MAY BE USED IN THIS EXAMINATION.

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

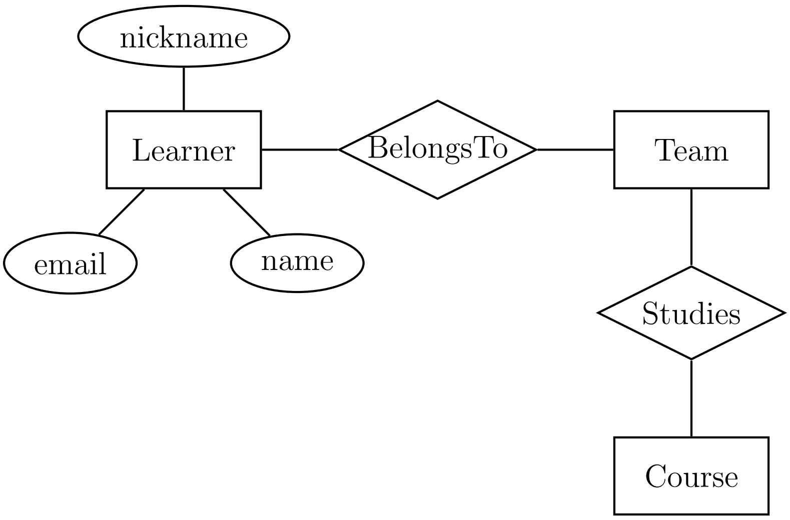
The non-existent startup company *W!L!T!* plans to create a collaborative online community of learners. From their breathless sales pitch:

> Come and join with *We! Learn! Together!* where everyone helps each other! Join teams of other learners just like you, and work together on a host of online study courses! Earn star points as you help others to learn too! You're never alone as We! Learn! Together!

Their plans for a website include a requirement to keep track of the following data.

- For every learner their name, unique identifying email address, and an optional nickname.
- For every study team its title and unique 8-character alphanumeric ID.
- Which learners belong to which teams, where anyone can belong to any number of teams.
- All courses, their titles and the number of star points a team can gain by completing the course.
- Details of which team is studying which courses.

The picture below shows an incomplete entity-relationship (ER) diagram representing some of this information.



(a) Give examples of an *entity set*, a *relationship* set, and an *attribute* from this diagram. [*3 marks*]

(b) Give brief explanations of the following terms: *key*, *candidate key*, *primary key*, *composite key.* [*8 marks*]

(c) Copy and complete the ER diagram, including any missing attributes and choosing appropriate keys. [*7 marks*]

(d) The organisers decide that some learners are more equal than others and so allow study teams to nominate one of their members as leader, if they wish. Extend the ER diagram with a relationship recording this new information. How do you express the constraint that each team can have only one leader?

[*6 marks*]

(e) Elsewhere the *W!L!T!* database has the following SQL for tables describing quizzes in the online courses, where each quiz is made up of several questions.

```
create table Quiz (
    id      varchar(8),
    title   varchar(30) not null,
    primary key (id)
)

create table Question (
    id      varchar(8),
    marks   integer not null,
    quiz    varchar(8) not null,
    primary key (id),
    foreign key (quiz) references Quiz(id)
)
```

What is the effect of the constraint in the last line, beginning **foreign key** ...?

[*2 marks*]

(f) Write an SQL query to count the number of questions held in the database.

[*3 marks*]

(g) Write an SQL query to find every 10-mark question, listing the title of the quiz it appears in and the question id. [*6 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.*]

(a) An information retrieval system is searching a European Parliament archive for documents on the topic of "offshore fishing boundary disputes". The following document matrix indicate three possible matches.

	offshore	fishing	boundary	disputes
Document A	4	2	7	0
Document B	3	3	3	3
Document C	12	6	0	0
Query	1	1	1	1

One way to rank these documents for potential relevance to the topic is the *cosine similarity measure.*

Write out the formula for calculating the cosine of the angle between two four-dimensional vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

Use this to rank the three documents in order of relevance to the query. [*10 marks*]

(b) One way to evaluate the performance of an information retrieval system is to assess its *precision* P and *recall* R. Informally, P can be defined as the proportion of the documents returned by the system which do match the objectives of the original search. Give a similar informal definition of R.

Here is the mathematical formula for calculating precision.

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

Name and define the terms TP and FP here. Give the formula for recall R, explaining any other new terms that appear. [*8 marks*]

(c) You have been given two different information retrieval systems to compare: *Hare* and *Tortoise*. Each one is tested on the same query for a collection of 4000 documents, of which 200 are relevant to the query. *Hare* returns 1200 documents, including 150 that are relevant; while *Tortoise* returns just 160, with 120 of them being relevant.

Tabulate the results for each system and calculate their precision and recall on this test. Show your working.

One way to combine precision and recall scores is to use their *harmonic mean.* Give the formula for this, and calculate its value for each of *Hare* and *Tortoise.* [*12 marks*]