UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Friday 11 th May 2018

0930 to 1130

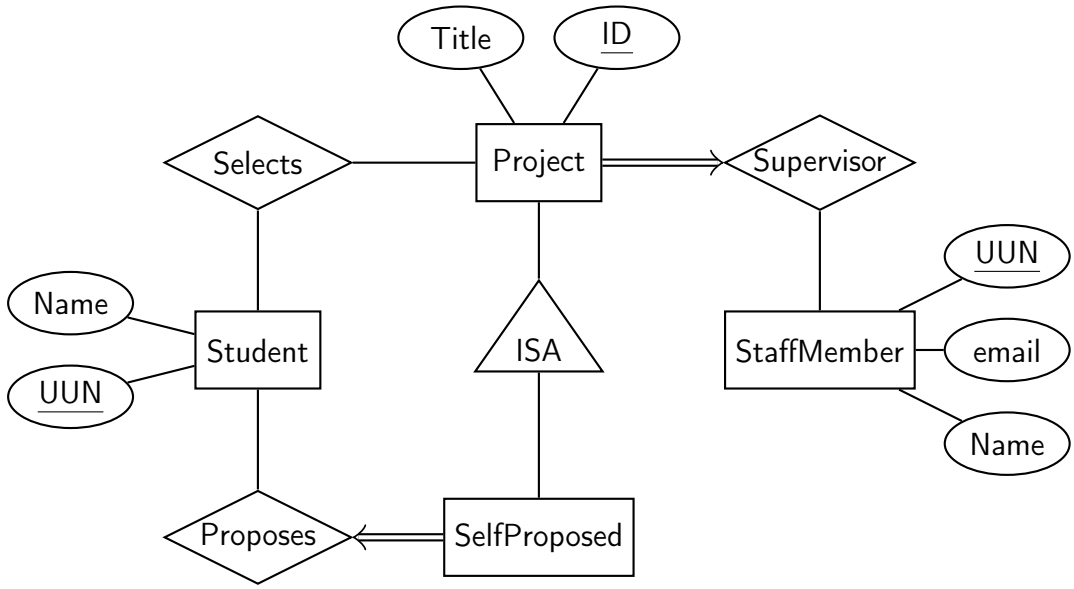
INSTRUCTIONS TO CANDIDATES

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.**

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

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

The following entity-relationship diagram sets out part of the specification for a database to manage allocation of undergraduate projects. All Informatics students undertake an individual project in their final year, supervised by a member of academic staff. This year 143 students had 259 different projects to choose from. Students select several possible projects and there is then a complex matching process to ensure every student has one of the projects they selected and no staff member is supervising too many students.



Some projects are "self-proposed" by individual students: these projects still have a staff supervisor, but cover a topic identified by the student.

(a) The *attributes* of each entity are shown in the linked ovals. Some of them are underlined. What does that indicate? [*1 mark*]

(b) What does the triangle in the middle of the diagram indicate about the relationship between the Project entity above and the SelfProposed entity below? [*2 marks*]

(c) What are the attributes of the SelfProposed entity in this diagram? [*3 marks*]

(d) The link from Project to Supervisor has an arrowhead and a double line. Each of these features indicates a particular kind of constraint. Name these constraints. In combination, what do they say about the Supervisor relationship between staff members and projects? [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

The information described in this diagram is to be held in an SQL database. This requires some additional specification for the attributes.

- Every project must have a title recorded in the database, every student must be named, and each staff member must have both name and email address recorded.
- All of those attributes are to be strings of no more than 260 characters.
- Each project ID is an integer.
- Each UUN (university username *or* universal username) is a string of up to 8 characters.

(e) Write a suitable set of declarations in DDL for the tables of this database. [*25 marks*]

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

The *Scalable Vector Graphics* standard, known as *SVG*, is an XML dialect for describing two-dimensional graphics. Here is a small SVG file for a rectangular image containing a solid red circle overlaid with the letters "SVG".

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <svg height="100" width="200">
3   <circle cx="100" cy="50" r="40" fill="red" />
4   <text x="80" y="55">SVG</text>
5 </svg>
```

(a) Draw the XPath data model tree for this document [*8 marks*]

The following file is a document type definition (DTD) for a highly restricted subset of SVG.

```
1  <!ELEMENT svg (circle|text)* >
2  <!ELEMENT circle EMPTY>
3  <!ELEMENT text (#PCDATA)>
4
5  <!ATTLIST svg height  CDATA #REQUIRED>
6  <!ATTLIST svg width   CDATA #REQUIRED>
7
8  <!ATTLIST circle cx CDATA #REQUIRED cy CDATA #REQUIRED>
9  <!ATTLIST circle r  CDATA #REQUIRED fill CDATA #IMPLIED>
10
11 <!ATTLIST text x CDATA #REQUIRED y CDATA #REQUIRED>
12 <!ATTLIST text fill CDATA #IMPLIED>
```

(b) The XML document above is well-formed and also valid with respect to this DTD. What do the terms "well-formed" and "valid" mean in this case? [*4 marks*]

(c) For each of the first three and the last two lines of the DTD give a brief explanation of its meaning. (Lines 1, 2, 3, 11 and 12.) [*10 marks*]

Write XPath expressions to find the following information from any XML document that matches this DTD.

(d) The width of the SVG document

(e) The radius of all filled circles

(f) All text coloured green or blue in the document

[*6 marks*]

QUESTION CONTINUES ON NEXT PAGE

The *Openclipart* online community maintain a repository of over 100,000 SVG images released to the public domain. The website http://openclipart.org provides a keyword search facility to find images in the collection: entering words like "bird feeder" should return images relating to feeding birds.

(g) The performance of an information retrieval engine like this can be evaluated in terms of its *precision* P and *recall* R. Give English-language definitions for each of these terms.

(h) Recall is calculated using the following formula.

$$R = \frac{TP}{TP + FN}$$

Name and define the values TP and FN appearing here.

(i) Precision and recall are often combined to give an *F-score*. The "balanced" F-score, $F_{0.5}$, weights each of them equally. Give a formula for $F_{0.5}$ in terms of P and R.

[*7 marks*]

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

Every February 2 in the town of Punxsutawney, Pennsylvania, people gather to watch a groundhog (large rodent) called "Punxsutawney Phil" make a weather prediction. If the sky is clear and Phil sees his shadow then this is a prediction for six more weeks of winter. If Phil sees no shadow, then this is a prediction that spring will come early.

This tradition started in 1887. Jeremy Neiman, a data scientist in New York, has analysed the success of these predictions. Taking 116 years in which there was a clear prediction and comparing data with the average US temperature for April that year, he found the following.

- The temperature was below average for 58 years ("more winter"), and above average for the other 58 ("early spring").
- The groundhog saw his shadow 100 times.
- On 5 occasions, the groundhog did not see his shadow and there was an early spring that year.

This question will explore whether there is a correlation between the groundhog's predictions and the weather.

(a) Draw up a contingency table for this data showing when there was "more winter" or an "early spring" against when the groundhog saw his shadow or not. [*4 marks*]

(b) What is an appropriate null hypothesis when investigating a possible correlation here? [*2 marks*]

(c) Use the *marginals* from this table of observed values to calculate a table of expected values. Show your calculations. [*6 marks*]

(d) Calculate the χ^2 statistic for this data, showing the formula you use and all your working. [*6 marks*]

(e) How many *degrees of freedom* are there in this system? What does this mean? [*3 marks*]

(f) The appropriate critical values for this χ^2 test are as follows.

p	0.10	0.05	0.01	0.001
χ^2	2.71	3.84	6.64	10.83

Based on this information, what can you say about this data as evidence regarding Punxsutawney Phil's ability to predict the weather? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

The χ^2 test here was used to look for correlation in qualitative data: groundhog observations and weather conditions. Where there is quantitative data, it is common to use the *Pearson correlation coefficient*. For two data series $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ this is computed with the following formula.

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n\sigma_x\sigma_y}$$

(g) What do the values n, μ_x, μ_y, σ_x and σ_y represent?

(h) What is the possible range of values that $\rho_{x,y}$ might take? What do the different values within that range indicate about possible correlation?

[*6 marks*]