Informatics 1: Data & Analysis

Lecture 21: Exam Preparation

Ian Stark

School of Informatics
The University of Edinburgh

Friday 7 April 2017
Semester 2 Week 11

THE UNIVERSITY of EDINBURGH

Plan !

In this lecture I shall work through solutions to two past exam questions.

May 2016	Question 1
August 2016	Question 3

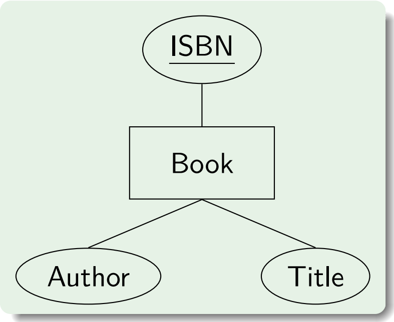
At the end of the lecture you will have an opportunity to fill out the online course feedback survey. Please do stay to do this.

May 2016 Question 1 Section (a)

A book can be identified either by its author and title, or its ISBN (International Standard Book Number). The diagram on the right represents a *Book* entity for a planned relational database.

Here are four concepts from the field of relational databases. For each one, give a one-sentence description and an example using the *Book* entity shown.

1. Key
2. Superkey
3. Composite key
4. Primary key



May 2016 Question 1 Section (a)

1. A *key* is a minimal set of attributes whose values uniquely identify an item in an entity set. For example, the combination of author and book title in the *Book* entity.

   It's essential to mention that a key is a set of attributes, not necessarily just a single attribute, and that it is minimal for identifying an item.

2. A *superkey* is any set of attributes whose values uniquely identify an item in an entity set. For example, the combination of author, book title, and ISBN all together in the *Book* entity.

   The example of a key from the previous part would also serve as a superkey but it's generally clearer to use something that includes more than is necessary for a key

May 2016 Question 1 Section (a)

3. A *composite key* is a key that includes more than one attribute. For example, the combination of author and book title.

   Note that this example isn't the chosen primary key for this entity, it's just one possible candidate key. The primary key, ISBN, is not a composite key.

4. A *primary key* is the key chosen among all candidate keys to be used as the unique identifier of records in a particular database. In this case the ISBN is the chosen primary key.

   It's not essential to mention candidate keys, but it helps to explain the choice of one among several possibilities.

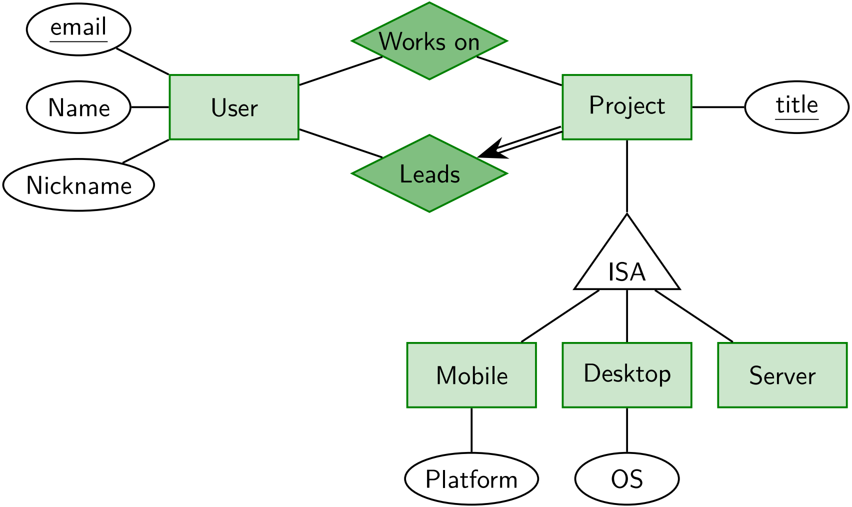
May 2016 Question 1 Section (b)

The fictional online service *BitBarrow* provides a repository for shared software development. It uses a database to track various aspects of the service, including the following.

- Users, who are identified by their personal email address and each have a registered name and nickname.
- Projects, identified by a unique project title.
- Which users work on which projects. Each user can work on several different projects, and each project may have multiple contributors.
- For each project, exactly one user who is the project leader.
- Different kinds of project. Projects may optionally be declared as mobile, desktop, or server. Mobile projects need an identified platform, and desktop projects a named operating system.

Draw an entity-relationship diagram that represents this information.

May 2016 Question 1 Section (b)



May 2016 Question 1 Section (c)

The BitBarrow database also includes information about *endorsements* where one user confirms that another user has expertise in a particular programming language. Users can also endorse themselves, to claim that they have such expertise. This is captured in the following three database tables.

```
create table User (
  email     varchar(254),
  name      varchar(200),
  nickname  varchar(200),
  primary key (email)
)

create table ProgLang (
  name varchar(100),
  primary key (name)
)

create table Endorsement (
  byUser   varchar(254),
  ofUser   varchar(254),
  forLang  varchar(100)
  ...
)
```

May 2016 Question 1 Section (c)

1. The ProgLang table only has one column. Why might it be helpful to have a "relation" like this with just a single field?
2. The Endorsement table is missing primary and foreign key declarations. Write an appropriate set of declarations to complete the definition.
3. Write an expression in the tuple-relational calculus describing the set of all users who have been endorsed as having expertise in Haskell.
4. Write an expression in relational algebra to compute the nicknames of users who have endorsed someone as having expertise in C.
5. Write an SQL command to list without duplication the names and email addresses of all users who have endorsed themselves.

May 2016 Question 1 Section (c)

1. Using a single-column table means that all references to programming languages must pick from this fixed list. This would, for example, avoid multiple alternate spellings or abbreviations of the same language.

2. All fields of Endorsement are required for the primary key, and all three have foreign key constraints to other tables.

```
create table Endorsement (
  ...
  primary key (byUser,ofUser,forLang),
  foreign key (byUser) references User(email),
  foreign key (ofUser) references User(email),
  foreign key (forLang) references ProgLang(name)
)
```

May 2016 Question 1 Section (c)

3. Here are two possible tuple-relational expressions for this set.

$$\{ U \in \text{User} \mid \exists E \in \text{Endorsement} \,.\, E.\text{ofUser} = U.\text{email} \land E.\text{forLang} = \text{"Haskell"} \}$$

$$\{ U \mid U \in \text{User} \land \exists E \in \text{Endorsement}, L \in \text{ProgLang} \,.\, E.\text{ofUser} = U.\text{email} \land E.\text{forLang} = L.\text{lang} \land L.\text{lang} = \text{"Haskell"} \}$$

4. Here are three possible ways to compute the result.

$$\pi_{\text{nickname}}((\sigma_{\text{forLang}='\text{C}'}(\text{Endorsement})) \bowtie_{\text{byUser}=\text{email}} \text{User})$$

$$\pi_{\text{nickname}}(\sigma_{\text{forLang}='\text{C}'}(\text{Endorsement} \bowtie_{\text{byUser}=\text{email}} \text{User}))$$

$$\pi_{\text{nickname}}(\sigma_{\text{forLang}='\text{C}' \land \text{byUser}=\text{email}}(\text{Endorsement} \times \text{User}))$$

May 2016 Question 1 Section (c)

5. Here is a suitable SQL expression.

```
select distinct U.name, U.email
from   User U, Endorsement E
where  E.byUser = E.ofUser and E.byUser = U.email
```

Here is another, slightly different.

```
select distinct U.name, U.email
from   User U, Endorsement E
where  E.byUser = U.email and E.ofUser = U.email
```

There are further legitimate variations on this. However, the **distinct** is essential as a user may endorse themselves multiple times for different languages and should appear only once in the results.

August 2016 Question 3 Sections (a), (b)

1. Explain what it means that data belongs to a *categorical data scale*. Give two examples of categorical data scales.

2. Explain what it means that data belongs to a *ratio data scale*. Give two examples of ratio data scales.

August 2016 Question 3 Sections (a), (b)

1. A *categorical* scale measures data by assigning it into different named categories. There is no ordering or numerical content.

   For example, counting student records by degree programme; or classifying mobile phone sales by operating system.

2. A *ratio* scale uses numeric values which have an absolute notion of zero; this means they can sensibly be added, and multiplied by real numbers.

   For example, the mass of an object measured in kg; or the orbital period of a planet measured in Earth-years.

August 2016 Question 3 Section (c)

Slushtastic! is a new and fictional low-energy soft drink, made of crushed ice and food colouring. It comes in one size and five different varieties. These varieties have no flavour or aroma so are distinguishable only by their colour. Nevertheless, the *Slushtastic!* marketing department are keen to find out whether some varieties are more popular than others.

The company collects data on the first 500 servings to thirsty customers.

Variety	Sales
Blue Blast	95
Red Ripple	124
Purple Power	97
Green Glow	96
Crystal Clear	88
Total	500

Marketing plan to use a χ^2 test to explore whether colour affects sales.

August 2016 Question 3 Section (c)

1. What is the *null hypothesis* for this investigation?
2. Calculate the table of expected frequencies of sales in each variety, under the assumption that the null hypothesis is true.
3. Give the formula for calculating the χ^2 statistic. Compute χ^2 for this sales data, showing your working.
4. In this test the data has 4 *degrees of freedom*. Explain what this means.
5. The critical values for the χ^2 test with four degrees of freedom are as follows.

p	0.1	0.05	0.025	0.01	0.001
χ^2	7.78	9.49	11.14	13.28	18.47

Based on this information, what evidence — if any — does the data provide on whether colour affects *Slushtastic!* sales? Explain how you reach this conclusion.

August 2016 Question 3 Section (c)

1. The null hypothesis is that colour makes no difference to drinks sales.
2. Under the null hypothesis, we expect all frequencies to be equal. This gives the following table.

Variety	Sales
Blue Blast	100
Red Ripple	100
Purple Power	100
Green Glow	100
Crystal Clear	100
Total	500

   The frequency for each variety is the total number sold (500) divided by the number of varieties (5).

August 2016 Question 3 Section (c)

3. The χ^2 statistic is computed as follows:

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$
$$= \frac{5^2}{100} + \frac{24^2}{100} + \frac{3^2}{100} + \frac{4^2}{100} + \frac{12^2}{100} = \frac{770}{100} = 7.70$$

4. The only restriction on the five values in the table is that they must add up to the total of 500. This means that four can take arbitrary values, but the fifth is then determined. These are the four degrees of freedom.

August 2016 Question 3 Section (c)

5. The data provides no evidence at all that colour affects *Slushtastic!* sales, and no justification to reject the null hypothesis.

   The computed χ^2 value of 7.70 lies below the 90% significance level $(p = 0.1)$ from the table of critical values.

> Although there is clearly some variation in this particular run of sales, with Red Ripple outselling everything else, the magnitude is not statistically significant and there is no evidence this is anything other than random noise.
>
> This data doesn't provide, either, any evidence that colour *doesn't* influence sales; although a more sophisticated test might be able to use this to justify or reject proposed upper bounds on the scale of that influence.

Rocket Science

Student Feedback Survey

Please complete the online survey for this course. This is anonymous, and I read every submission.

MyEd $\longrightarrow$ Studies $\longrightarrow$ Course Enhancement Questionnaire

Direct URLs

- For Inf1-DA: **http://is.gd/dasurvey**
- For all courses: **http://is.gd/infsurvey**

You can do this

The Inf1-DA syllabus and exam questions are written to be achievable. Every year large numbers of students pass the exam writing straightforward correct answers about things they understand. You can do this too.

Anything Else?

If you have further questions about the course content, tutorial exercises, the exam, where to buy a disco calculator, or anything else, please:

- Post a question on *Piazza*; *or*
- Ask your course tutor, in person or by email; *or*
- Ask me, in person or by email.

Thank you for your attention

We're done here