# Inf1B::Data and Analysis
# 2005 Mock Exam

Frank Keller Stratis Viglas

## 1 Structured Data

In order to manage your large exotic pet collection, you have decided to catalog the relevant information in a relational database system. You have decided to keep track of information concerning types of animals, individual animals, cages in which the animals live (your house is really big) and food that each animal eats. The purpose of the system is to allow you to determine quickly what each animal can eat and which cage it is in. You have determined the following:

- You have many animals, each animal belonging to a specific type.
- Animals have a (unique) name and an age.
- Types have a (unique) name and an original habitat.
- Each animal belongs to a single type.
- Each cage has a size, a (unique) location, and an average temperature.
- Each food has a (unique) name and a price.
- Each type of animal may eat several types of food, and each type of food may be eaten by several types of animals.
- Each animal is assigned to only one cage, although each cage can contain many animals.

1. Design an ER diagram from the information above. Be sure to capture all integrity constraints. What are the types of relationships you have identified (*e.g.*, one-to-one, many-to-many, *etc.*)? if you are making any assumptions, state them *clearly*. *6 marks*

2. Convert the ER diagram to a relational schema. Be sure that any constraints you have identified in the ER diagram are captured by your relational schema. *6 marks*

3. Express the following query in relational algebra: "Find the price for the kind of food that my animal 'Joe' eats." *6 marks*

## 2 Semi-structured Data

1. A linguist would like to find out whether *get a haircut* or *have a haircut* is a valid collocation in English. She conducts a corpus study and obtains the following frequencies from a 1125 word corpus:

| Word | Frequency |
|---|---|
| *get* | 100 |
| *have* | 300 |
| *haircut* | 50 |
| *get haircut* | 25 |
| *have haircut* | 25 |

   (a) Explain how a $\chi^2$ test can be used to decide whether *get a haircut* or *have a haircut* is a valid collocation, based on corpus data such as the one in this example. *4 marks*

   (b) Compute the $\chi^2$ values for both collocations. *8 marks*

2. You want to extract all dates in a corpus of English text. How would you go about doing this? Assume that the corpus is tokenized, but not part-of-speech annotated. Example dates you should be able to recognize include: *7 marks*

   - December 20 2004
   - 20th of December 2004
   - 19/03/70
   - 03/19/70