Informatics 1, 2009

School of Informatics, University of Edinburgh

Data and Analysis

Introductory Lecture:
Overview and Logistics

Alex Simpson

Data — Mirriam-Webster's Dictionary extract:

1: factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation.

2: information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful.

3: information in numerical form that can be digitally transmitted or processed.

In informatics, *"data"* is primarily used to refer to (not necessarily numerical) information which has been put in digital format so it can be stored, transmitted, retrieved or processed.

("Data" was originally the plural of *"datum"*, but is now often used as a singular word as well as a plural word.)

Analysis

All definitions of “data” emphasise the requirement for further processing of data.

Raw data (just the digital information on its own) is meaningless without context

Thus the topic of *data* goes hand in hand with that of the *analysis* required in order to process and interpret data.

Importance of data

- How much data is there on digital storage devices worldwide in total?
- How accurate/reliable is this data?
- How secure is this data?
- Conversely, how accessible is this data?
- How much personal data about you is there?
- How much personal data about you is accessible to you?

There are regular stories in the media that touch upon these questions. E.g., consequences of data inaccuracies; breaches of security with public data; censorship of internet data by governments; etc.

Thus issues about *data* are highly relevant to our everyday lives.

This course . . .

This course is not, however, about the political, sociological and moral issues surrounding data. (Although these issues are both important and interesting.)

This course is about the technologies that underpin the gathering, storage, retrieval, manipulation and analysis of data.

Such technologies are clearly vital given the prevalence of data applications.

However, the focus of this course is not directly on individual technologies themselves. This is not a technology-based course.

. . . is a theory course!

This course is primarily a *theory* course.

You will learn the *principles* underlying a variety of technologies for gathering, storing, retrieving and analysing data.

Learning principles is more important than learning technologies:

- Technologies change and become obsolete relatively quickly.
- The principles underpinning them are much more stable. (They do evolve, but far more slowly than technologies.)

Having said this, the course will also include some discussion of current technologies, mainly as vehicles for explicating general principles.

Structure of course

Lectures are on Mondays and Thursdays, and run from today (12th Jan) until Thursday 19th March (Week 10).

There is no lecture on Thursday 22nd January. This is thus the first of 19 lectures.

The course is divided into 5 parts.

Part I —- Structured Data	(lectures 2–7 approx.)
Part II — Semistructured Data	(lectures 8–10 approx.)
Part III — Corpora	(lectures 11-13 approx.)
Part IV — Data Retrieval	(lectures 14–15 approx.)
Part V — Statistical Analysis of Data	(lectures 16–18 approx.)

Lecture 19 will be a revision lecture.

Textbooks

There is no one book that covers all the topics of the course. There is thus no compulsory course textbook.

Nevertheless, the following book, which is an excellent textbook on databases, covers all the material in Part I in great detail, and most of the material in Part II (more briefly). It is also the recommended textbook for the 3rd-year computer science databases course. So it may make sense to buy now and invest for later:

> Database Management Systems
> R. Ramakrishnan and J. Gehrke
> Third Edition, McGraw-Hill, 2003

However, there is *no obligation* to buy this book for Data & Analysis.

Course notes

The primary reference will be the lecture slides. These will be distributed in lectures. They will also kept (and corrected) on-line on the Data & Analysis webpage, linked of the Informatics 1 course page:

http://www.inf.ed.ac.uk/teaching/courses/inf1/

Occasional clarifications and extensions to the lecture slides will be given in lectures. Thus it will occasionally be helpful to take notes in lectures, possibly by annotating the distributed slides by hand.

In addition to the lecture slides, photocopies of additional reading material will sometimes be distributed in lectures. This material will be set as required reading, which is a *compulsory* part of the course.

Spare copies of all distributed material will be available from the pigeon holes outside AT room 5.03.

Tutorials

Weekly tutorials will be held on Tuesdays and Wednesdays, starting Tuesday 27th January (Week 3) and running until Wednesday 25th March (Week 11), making 9 tutorials in all.

Tutorial allocations will be made by the ITO in Weeks 1 and 2, and you will be contacted by email with your allocation, and given chance to change it if the assigned time is unsuitable.

Attendance at tutorials is *compulsory*. If you are unable to attend a tutorial, you must contact your tutor in advance. If possible, you should attend another tutorial the same week.

Weekly exercise sheets

At tutorials, you will discuss weekly exercise sheets. These will be released (and announced by email) one week before the tutorials in which they are to be covered.

You are expected to attempt to complete the exercise sheets in advance of tutorials, and to take your workings and solutions to tutorials for discussion.

The tutorial exercises will involve paper and pencil exercises, and also, sometimes, on-line exercises which you can complete in the Informatics 1 drop-in labs (held every afternoon in CLW).

It is very important to keep abreast of tutorial work and not get behind with the course. In particular, take care to avoid the *second semester slump*!

Coursework assignment

A coursework assignment will be released Thursday week 8.

Your solutions must be handed in to the ITO (AT room 4.02) by noon on Friday 13th March.

Your solutions will be marked, commented and returned to you in Week 11 tutorials.

The assignment offers practice on exam-like questions and provides invaluable feedback on how you are doing on Data & Analysis.

Assessment

Assessment by means of 2-hour written exam during exam period (April/May).

This exam contributes 100% of your Data & Analysis marks

Prerequisites

The Semester 1 Inf 1 courses, Computation and Logic, and Functional Programming, are prerequisites for Data & Analysis.

The Semester 2 Inf1 course Object-oriented Programming is a corequisite with Data & Analysis.

Course people

Lecturer: Alex Simpson `<Alex.Simpson@ed.ac.uk>`
Teaching Assistant: Srini Janathanam `<s0450680@sms.ed.ac.uk>`

Acknowledgements

Part I of course developed from slides originally written by Stratis Viglas.
Parts III–V developed from slides written by Frank Keller and Helen Pain.