Informatics 1

School of Informatics, University of Edinburgh

Data and Analysis

Introductory Lecture:

Overview and Logistics

Alex Simpson

17 January 2012

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, legal, sociological and moral issues surrounding data.
(Notwithstanding that 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.

. . . 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

The course is divided into 3 parts.

Part I —- Structured Data (lectures 2–7 approx.)

Principles of relational databases. Entity-relationship diagrams. Relational model. Relational algebra and calculus. SQL.

Part II — Semistructured Data and Text Corpora (lectures 8–13 approx.)

Semistructured data models. XML. DTD's. XPath. Text corpora as semistructured data. Informatics applications of corpora.

Part III — Unstructured Data and its Analysis (lectures 14–18 approx.)

Exploiting and analysing unstructured data. Information retrieval. Statistical analysis of data.

Prerequisites

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

The Semester 2 Informatics 1 course Object-Oriented Programming is a corequisite with Data & Analysis.

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 some 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 be kept (and corrected) on-line on the Data & Analysis course webpage:

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

Occasional clarifications and extensions to the lecture slides will be given in lectures. Thus it will sometimes 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.

Course blog

There is a course blog.

http://blob.inf.ed.ac.uk/inf1da/

(also linked off the DA course webpage)

This carries information about the content of the lectures, tutorial work, references and further discussion of background material.

Students are encouraged to post questions and comments to the blog.

Lectures

Lectures are at 11am Tuesdays (AT LT3) and 2pm Fridays (AT LT1)

The course content will be delivered in lectures 2–18, starting Friday 20th January (Week 1), and ending Monday 27th March (Week 11).

- On Friday 3rd March (Week 3), a lecture by the University Careers Service will be given in the usual DA lecture slot.
- Due to the University *Innovative Learning Week (ILW)*, no lectures will be held in Week 6 (Tuesday 21st and Friday 24th February).

A revision lecture will be held on Monday 3rd April (Week 12).

Tutorials

9 weekly tutorials will be held on Tuesdays and Wednesdays starting Tuesday 31 January (Week 3) and running until Wednesday 4 April (Week 12). Due to ILW, there are no tutorials in Week 6 (21–22 Feb).

Tutorial group allocations can be checked from the course webpage. Please check the list online to find your group. If your assigned group or time is unsuitable, please contact the ITO through their webform to explain the difficulty and request a change.

Attendance at tutorials is *compulsory*. If you are ill or otherwise unable to attend one week then email your tutor, and if possible attend another tutorial group in 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 computer labs in Appleton Tower.

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 on Friday 9 March (Week 8).

The assignment takes the form of a mock exam paper, based on questions from previous years.

Post your written solutions in the box outside ITO (AT room 4.02) by 2pm on Friday 23 March (Week 10).

Your tutor will mark this and return it to you at the final tutorial, with opportunity for discussion and feedback.

Examination

Your grade for Inf1-DA is based on a two-hour written examination at the end of the year. The paper follows a similar format each year, and past papers are available online.

The exam will take place during the April/May diet; the University Registry will publish the precise date later in the Semester.

Course people

Lecturer: Alex Simpson <Alex.Simpson@ed.ac.uk>

Teaching Assistant: Aurora Constantin <A.Constantin-2@sms.ed.ac.uk>

Office hours

Alex Simpson holds a drop-in office hour for Informatics 1 students at 11.30–12.30 on Thursdays in IF (Informatics Forum) room 5.25.
(Office hours will be held in Weeks 1–2, 4–5 and 7–12 of semester only.)

Acknowledgements

This course incorporates material originally contributed by Frank Keller, Helen Pain, Ian Stark and Stratis Viglas.