Informatics 1: Data & Analysis

Lecture 1: Introduction

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 14 January 2014
Semester 2 Week 1

Welcome

Informatics 1: Data & Analysis

This course provides an introduction to representing and interpreting data from areas across Informatics; treating in particular structured, semi-structured, and unstructured data models.

Lecturer: Dr Ian Stark
Email: Ian.Stark@ed.ac.uk
Office: IF 5.04
Drop-in: 1030–1130 Wednesdays

Assistant: Dr Areti Manataki
Email: A.Manataki@ed.ac.uk
Drop-in: 1630–1730 Tuesdays at InfBASE in AT 5.02

Supporting Staff

Course Tutors

Jane Hillston; Eleni Zarogianni; Mihaela Dragomir; Tom Spink; Çigdem Beyan; Areti Manataki; Ian Stark

Tutorials start in week 3

ITO (AT Level 4)

Course secretary: Sue Cade
Contact: http://www.inf.ed.ac.uk/teaching/contact

Year Organiser

Paul Anderson

Degree Programmes and Related Courses

Degrees

- Computer Science
- Software Engineering
- Artificial Intelligence
- Cognitive Science
- Cognitive Science (Humanities)
- Informatics

...perhaps with Mathematics, Electronics, Physics, or Management.

Other Courses

- Informatics 1: Functional Programming
- Informatics 1: Computation & Logic
- Informatics 1: Object-Oriented Programming

- Informatics 1: Cognitive Science
- Introduction to Linear Algebra
- Calculus and its Applications

Data

Data *(noun)*

Facts and statistics collected together for reference or analysis: *"there is very little data available".*

- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- *Philosophy* Things known or assumed as facts, making the basis of reasoning or calculation.

Oxford Dictionaries Online

In Latin *"data"* is the plural, with singular *"datum"*, but it is now widely used as a singular mass noun: "data was collected".

Analysis

Definitions of "data" generally emphasise the requirement for further processing of data: invariably, the purpose of collecting data is to make some further use of it.

We shall be looking at several different kinds of data, but for all of them the topic of *data* goes hand in hand with that of the *analysis* necessary to process and interpret it.

Indeed, before even starting to collect data it's usually important to know what kind of analysis will be done with it, in order to gather, organise and manage the data appropriately.

Bits → Data → Information → Knowledge → Understanding → Wisdom

(We'll be spending most of our time towards the left)

Significance of Data and Analysis

- How much data is held on digital storage devices, worldwide?
- How accurate is this data?
- How secure is this data?
- What do people do with this data?
- How much of it is personal data about you?

There are regular stories in the media that touch upon these questions — consequences of data inaccuracies; breaches of security with personal data; covert interception of data; censorship of data; etc.

Issues about data, how it is organised and analysed, have real relevance to our everyday lives.

This Course is About. . .

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

This course covers the methods and technologies used for large-scale collection, storage, retrieval, manipulation and analysis of data.

However, the technologies are a vehicle: really, this is about the *principles* which guide these technologies and what *challenges* they aim to address.

When studying this course, try to take a longer-term view:

- Notice the general principles which have been developed and applied with success in these specific cases.
- Use these particular models and languages as practice in the general skill of learning new things.

Challenges and Solutions

Example

Challenge — How can we use computers to help us extract information more efficiently from large quantities of data?

Technology — SQL and query optimization engines.

Principle — Use a custom language to describe the analysis required at a high level of abstraction, and have the computer identify the most efficient algorithm to carry it out.

Although we shall discuss specific technologies like SQL, and you should acquire skill in using them, the long-term goal is to understand the challenge and what it is that makes for a good solution.

Computers and Data

What's so special about using electronic computers to handle data?

Scale

Terabytes, Petabytes, Exabytes; the internet, genomes, lifebits, data smelters.

Speed

Gigahertz, Teraflops, Megabits/second; multicore, data pipes, fibre.

Flexibility

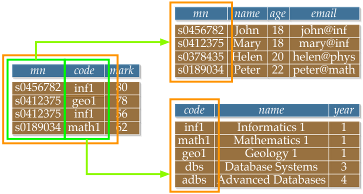
Computers are *programmable* — they will do with data whatever we ask them to.

We can even devise new languages to describe the new ways we create, manipulate and analyse data

Structured Data

In this course Structured Data refers to the classic model of databases keeping highly-structured records and files of information.

The currently-dominant approach is relational databases: rectangular tables with fixed structure and links between them.



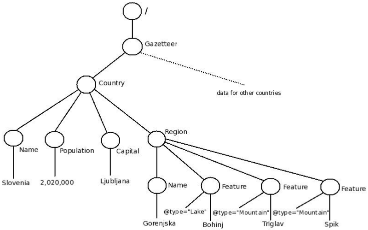
We analyse the data using the high-level declarative language SQL.

A key principle is that an SQL declaration states the desired solution; but a query optimizer works out how best to carry out the computation.

Semistructured Data

What we now call Semistructured Data originated with languages like SGML and HTML for annotating text documents.

Their more general descendant XML is now widely used for repositories information of many kinds.



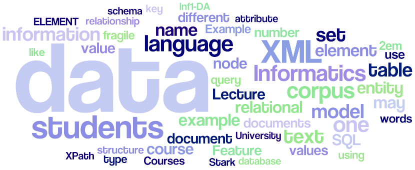
Compared to classic relational database tables, XML trees offer more flexibility in arrangement, but still with some structure and the possibility of validition and type-checking.

There are several specialised languages for describing and analysing XML files: we look at DTD and XPath.

Unstructured Data

Almost all data in machine-readable form has a least *some* structure: bits, bytes, characters, files.

By Unstructured Data we generally mean there is no additional large-scale or data-specific structure.



We shall look at unstructured data from written texts — documents, books, or whole libraries of them — as well as numeric data from surveys or experiments.

This data may be locally annotated and tagged, but without global structure.

Methods for the analysis and retrieval of information from unstructured data are less standardized, and in some ways much more challenging.

Speed isn't Everything

Managing and analysing large amounts of data is hard.

At first sight, the fact that computers are (unimaginably) fast and programmable (can do anything) may suggest that our problems are over.

However, it turns out — perhaps surprisingly often — that:

- Sheer speed and capacity are not nearly enough.
- We may not know how to do the analysis we want.

(or even exactly what it is that we do want)

The real advances are

> Solutions that are better than the thing you first thought of

Many things in this course are firm fixtures in computing not because they are "the" solution, but because they turned out to be *better than what was done before*. And there may be better ones yet to be invented.

Anonymous Survey

Q1: How many hours do you estimate you slept last night?

Q2: About how many hours of physical exercise do you usually do in a week?

Q3: Elaborate question about operating systems.

My Z is:

Elaborate Question About Operating Systems

This question is about what students use to carry out coursework and connect remotely to Informatics accounts. If you routinely use multiple devices for this, running multiple different operating systems, then please choose the one which you use most often.

Which operating system do you expect to be using while working on this course when away from the computer labs?

- L Linux (Mint, Ubuntu, Fedora, Debian, SUSE, ...)
- C ChromeOS (Chrome, ChromeBox, ChromeBase, ...)
- W Microsoft Windows
- X OS X
- Z Something else (BSD, AmigaOS, z/OS, Difference Engine,...)
- N None — I expect to do all my work on lab DICE machines.

If Z, then please write a more specific answer in the space provided.

Yes, I know that ChromeOS is a Linux variant; in this case I want to distinguish it.

Lectures !

When and Where

Tuesday	11.10 – 12.00	Lecture Theatre 5, Appleton Tower
Friday	14.10 – 15.00	Lecture Theatre B, David Hume Tower

These are usually video-recorded and put online for later reference.

I shall often in one lecture set reading or other preparation for the next lecture.

You are strongly advised to attend all lectures and to carry out the reading or preparation.

Textbooks

This is not a textbook course, and there is no single compulsory book.

For certain parts of the course, however, I shall indicate one or more books which cover the current material — usually in much more depth and generality than required for this introductory course.

You can consult these books in the library, or borrow them, and you may find one or other helpful to you. Although the content is often similar, styles and tastes can differ significantly.

Occasionally I shall distribute photocopies of an individual textbook chapter when it is especially relevant to the course.

Coming Soon

That's quite enough information for now. In future lectures, I shall cover what I can of the following:

- Tutorials (these start in week 3)
- Coursework
- Assessment and feedback
- Exams
- The colour-coded lecture slides
- Collaborative annotation with Nota Bene
- Informatics 1 discussion forum
- Mailing list inf1-da-students@inf.ed.ac.uk
- Course blog
- Informatics 1 IRC channel
- Places to go for help

Fin.

Homework

Before the next lecture, on Friday, read these two things:

- The Inf1-DA course web page.
  http://www.inf.ed.ac.uk/teaching/courses/inf1/da
- Jeannette M. Wing.
  Computational Thinking. *Communications of the ACM* 49(3):33–35.
  DOI: 10.1145/1118178.1118215
  PDF: http://www.cs.cmu.edu/~CompThink/papers/Wing06.pdf

Acknowledgements

This course includes material originally contributed by Frank Keller, Helen Pain, Alex Simpson and Stratis Viglas.