Informatics 1: Data & Analysis

Lecture 1: Introduction

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 15 January 2019
Semester 2 Week 1

What’s in This Lecture

- Introduction
- Data, analysis, and what this course is about
- Survey of operating system use
- Three areas of study:
  - Structured Data
  - Semi-Structured Data
  - Unstructured Data
- Review and homework

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:	Informatics Forum IF 5.04
Drop-in:	1030–1130 Wednesdays

Course Secretary:	Rob Armitage
Contact:	ITO Appleton Tower 6.05
Office open:	0930–1230, 1330–1630 Monday to Friday

Supporting Staff !

Course Tutors

Nikola Pavlov, Rik Sarkar, Jane Hillston, Santi Guillen Garcia, Piotr Jander, Heather Yorston, Jázon Szabó, Stefani Genkova, *and many others*

Tutorials start in week 3

ITO — Informatics Teaching Organisation

Opening hours: Monday–Friday 0900–1230 / 1330–1630
Online: http://www.inf.ed.ac.uk/teaching/contact
Go to the *internal enquiries* form and select category "ITO: Informatics 1".

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

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

Data

Data *(noun)* Oxford English Dictionary

2a. Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation.

Edinburgh New Philosophical Journal, 1826:
"Inconsistent data sometimes produces a correct result."

2b. *Computing.* Quantities, characters, or symbols on which operations are performed by a computer, considered collectively. Also (in non-technical contexts): information in digital form.

Moore School Lectures, U. Pennsylvania, 1946:
"The data is stored in the memory in a systematic fashion"

Originally *"data"* was the plural, with singular *"datum"*. For some time now, though, it has been very widely used as a singular mass noun: "your personal data is at risk".

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)

This Course is About…

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 will 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

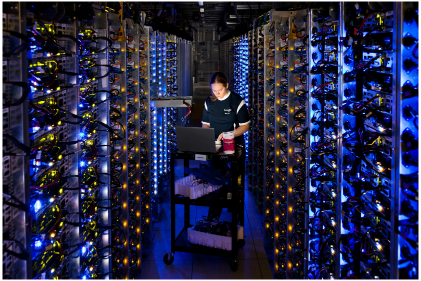
What's so special about using electronic computers to handle data?

Scale

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

Speed

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



Operations Engineer Denise Harwood
Google Data Centre, Oregon

Flexibility

Computers are *programmable* — they will do with data whatever we ask. We can even devise new languages to describe the new ways we create, manipulate and analyse data

Intermission

Anonymous Survey

Q1: How many hours do you estimate you spent asleep last night?

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

Q3: Pick your operating system.

Operating System

Please fill in one of W, X, L, A, C, Z or N — explanation on next slide.

Pick Your Operating System

This question is about how you expect 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 use for work away from the computer labs?

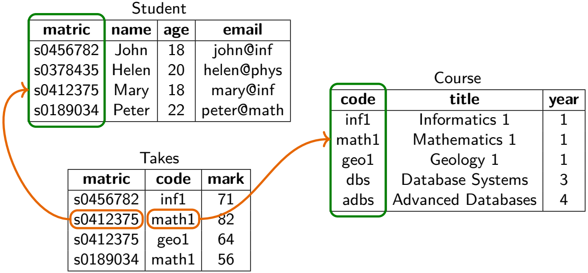
W	Microsoft Windows	
X	macOS, OS X, iOS	
L	Linux	(Mint, Ubuntu, Fedora, Debian, SUSE, . . .)
A	Android	(Also Linux, but it's helpful to distinguish)
C	ChromeOS	(Chrome, ChromeBox, ChromeBase, . . .)
Z	Something else	(FreeBSD, AmigaOS, z/OS, Analytical Engine,. . .)
N	None — I expect to do all my work on lab DICE machines.	

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

Structured Data

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

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



Student

matric	name	age	email
s0456782	John	18	john@inf
s0378435	Helen	20	helen@phys
s0412375	Mary	18	mary@inf
s0189034	Peter	22	peter@math

Course

code	title	year
inf1	Informatics 1	1
math1	Mathematics 1	1
geo1	Geology 1	1
dbs	Database Systems	3
adbs	Advanced Databases	4

Takes

matric	code	mark
s0456782	inf1	71
s0412375	math1	82
s0412375	geo1	64
s0189034	math1	56

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

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

Structured Data

Example Data

The University of Edinburgh keeps records of students, degree programmes, lecturers, courses, lecture theatres, *etc.*

What is a good way to organise this data? How much is there? How does it change over time?

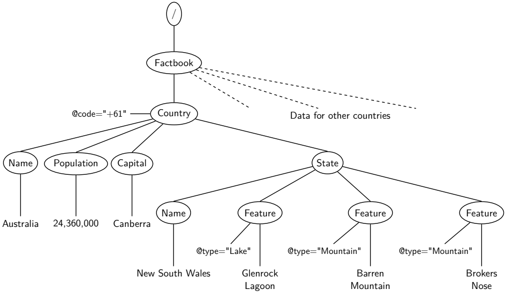
Example Analysis

- What proportion of students this Semester have to travel between two sites for consecutive lectures or tutorials?
- How can we timetable the exams?

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 many kinds of information.



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

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

Semistructured Data

Example Data

The BioModels Database is a repository of computational models of biological processes. Some are taken from original scientific papers, others are automatically generated from five other different databases. As of 15 January 2019 the database holds 151,472 models of biological systems.

What is a good way to organise this data? How can we explore it? What if it contains contradictory information?

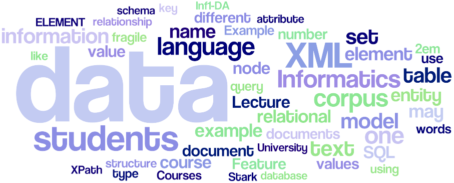
Example Analysis

- What models contain a reversible reaction involving calcium ions?
- What drugs might help manage blood cholesterol levels?

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.

Unstructured Data

Example Data

The British National Corpus contains 100 million words of written and spoken English from a range of sources covering the late twentieth century. It is extensively annotated with linguistic and grammatical information.

What is a good way to organise this data? How can we incorporate useful extra annotations?

Example Analysis

- Which documents mention pineapples?
- Is written English getting simpler over time?

Speed isn't Everything

Managing and analysing large amounts of data is hard.

At first sight, the fact that computers are big, fast and programmable (so 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.

Lectures !

When and Where

Tuesday	1110 – 1200	Gordon Aikman Lecture Theatre
Thursday	0900 – 0950	Gordon Aikman Lecture Theatre

Lectures are usually recorded and will appear online for later reference.

I shall often in one lecture set reading or other preparation for the next lecture: this homework is part of the examinable content of the course.

If you wish to participate in Inf1-DA then you should attend all lectures and carry out the homework.

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 PDF and photocopies of an individual textbook chapter when it is especially relevant to the course.

Coming Soon !

That's enough administration for now. In future lectures, I shall cover the following:

- Tutorials (these start in Week 3)
- Coursework (take-home exam practice in Week 7)
- Inf1-DA online: Piazza discussion group, blog, Facebook, Twitter,. . .
- Working hours
- Assessment and feedback
- Exams and exam preparation
- The colour-coded lecture slides
- Places to go for help

Summary

Structured Data

- e.g. University database of students, staff, courses, rooms, *etc.*
- Which students take Inf1-DA? How do we timetable exams?

Semistructured Data

- e.g. Tourist factbook about countries, regions, cities, . . .
- e.g. BioModels repository of computational models for biochemical reactions and metabolic pathways.

Unstructured Data

- e.g. British National Corpus of spoken and written English.

How to Succeed at Inf1-DA

- Be there at every lecture
- Take notes, paper or electronic (slides will be available online and as paper handouts)
- Write up your notes after the lecture: rearrange, reflect, summarise
- Do the homework: read this, do that, watch these
- Work through the exercises before each tutorial
- Be there at every tutorial: take part, work together with others
- After each tutorial, do some of the additional example questions
- If you have questions or problems then ask for help: on Piazza or Facebook; other students, tutors, me; in person, by email, after lectures.

Homework

Before the next lecture, on Friday:

Read This

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

Do This:

If you use Piazza or Facebook, then enrol in the Inf1-DA 2019 class/group.

Acknowledgements

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