036221

1

# COMP1314

Data Management

2

# Who?

- Danesh Tarapore (Unix & Scripting)
  - d.s.tarapore@soton.ac.uk
- Tom Blount (Structured Data, Databases)
  - t.blount@soton.ac.uk
- Rahman Attar (Relational Algebra, Databases in Practice)
  - aaq1d22@ecs.soton.ac.uk
- Angelos Nikolopoulos (Senior lab demonstrator)
  - arn1r23@soton.ac.uk
- Email the team
  - COMP1314-instructors@all.soton.ac.uk

3

# Our Journey

- Data-based operating systems
- Unix and Unix-like
  - Unix fundamentals
  - Scripting for data processing
- Structured data representation
- Databases
  - How they work
  - What they can do
  - SQL and NoSQL
- Putting it all together

5

# When?

- https://moodle.ecs.soton.ac.uk/course/view.php?id=481
- Always check the Timetable to know what's happening

Timetable

| Lecture | Lecture | Lecture/Tutorial | Lab |
|---|---|---|---|
| Tue. 2pm - 3pm<br>32/1015 | Fri. 10am - 11am<br>54/4011<br>Overflow room: 02A/2065 | Fri. 11am - 12pm<br>46/3001<br>Overflow room: 100/8009 | Wed. 10am - 1pm |
| Capacity: 426 | 213 | 354 | |

6

# Livestreaming and Q&A

- If the lecture theatre is full, you can head to the overflow room or a quiet spot (Library, B60 Labs, B100 etc.) to watch the stream.
- Link to the livestream will be available on the Moodle schedule.
- If you have any questions while watching the livestream,
  - Login to MS Teams using your university account
  - Find your class **COMP1314-45772-25-26**
  - Post questions on "Q and A Board" channel
  - Angelos will be monitoring the channel and communicate your questions to the lecturer.

7

# Resources for you

- Please bookmark us on Moodle:
https://moodle.ecs.soton.ac.uk/course/view.php?id=481

- The module's Moodle page is where you'll find
  - Lecture notes
  - Pre-recorded tutorials
  - Recordings of lectures
  - Links to Lab scrips
  - and more ….

8

# How?

- Lectures
  - To help guide you and start your journey

- Labs (30% from *COMP1300 – the Lab module*)
  - Four assessed labs – D1, D2, D3 and D4 (20%). First lab, D0, is unassessed.
  - Other assessments from COMP1300 (10%)

- Coursework (30%)
  - Putting theory into practice
  - Unix + Databases

- Exam (40%)
  - Making sure you understand the theory

9

UNIX operating system

10

# Data-based Operating Systems

- Some operating systems better tailored to handling data
- UNIX and UNIX-like operating systems
  - Base utilities
    - For processing and manipulating files
  - Pipelines
    - From one application into the next
  - Scripting
    - Getting more complex
  - Developer-oriented
    - Open-source libraries and languages at your fingertips
    - Python, R etc.

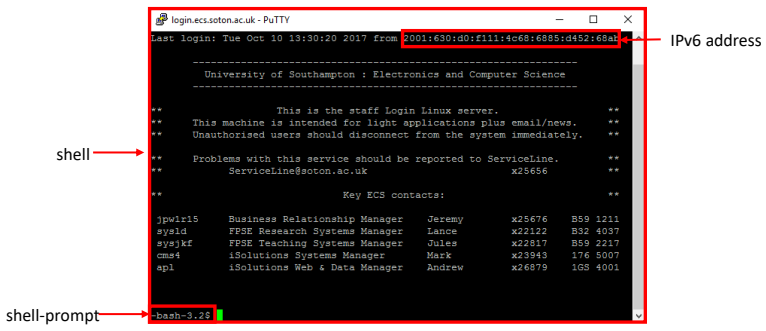
11

# Getting to know UNIX

- The UNIX design philosophy.
- Navigate your way around the UNIX file-system.
- Manipulate UNIX files and data.
- Launch and control jobs (programs) on a UNIX server.
- Write UNIX scripts to automate various tasks:
  - e.g. downloading data from websites and extracting useful information like stock index prices for analysis.

12

# Interactive sessions




13

# “Classic” Databases with Tom

14

# I’m Tom

- I’m Tom!
- I was an undergraduate here at Southampton
- Now, I teach!
- I like games, and creative computing

- *Show, don’t tell. And better yet, **do** don’t show!*

15

# COMP1314

- Teaching approach
  - Learning by doing
  - We'll be mixing practical and theory together
- Mixed sessions
  - We'll have a mixture of lecture and lab sessions
  - But even the lectures will have some interactive bits
- Learn and practice
  - We'll learn some things, try them out, put them into practice, get any help and answer any questions and then move on to the next thing

16

# My Part of the Module

- Structured data
  - Why bother structuring data?
  - How to define, create, invent, validate
  - Common formats and languages
- "Classic" Databases
  - Why and where they're used
  - Some of the maths behind them
  - How to use and manipulate them


17

# Relational Algebra & Databases in Practice with Rahman Attar

18

**Computer Scientist in Artificial Medical Intelligence**

**Dr Rahman Attar**, SMIEEE, MIET, FHEA, PhD, MPhil, BEng
Lecturer @ School of Electronics and Computer Science, University of Southampton | r.attar@southampton.ac.uk
Honorary Research Fellow @ Department of Bioengineering, Imperial College London | r.attar@imperial.ac.uk

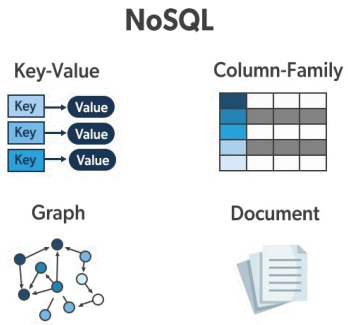
ATTAR Lab
Advanced Technologies for Translational Artificial-Intelligence Research

attarlab.com

19

## What topics will I be covering?

- Overview of traditional relational databases
- Limitations of relational databases
- Introduction to NoSQL databases
- Types of NoSQL databases
- Basic operations and queries in NoSQL databases
- Schema design in NoSQL databases
- Challenges and considerations in choosing NoSQL databases




20

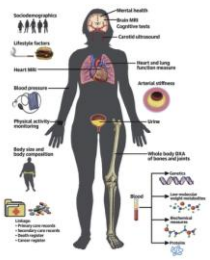
## Why are these topics important?

- The UK Biobank contains a vast amount of health-related data, imaging data, genetic information, and more from a large number of participants. A combination of technologies is used to efficiently manage and query such a diverse and extensive dataset.

- UK Biobank utilises a complex and comprehensive data storage infrastructure to manage its vast dataset. The UK Biobank doesn't rely on a single type of database but rather employs a combination of databases.

biobank uk

one of the world's largest population studies



21

## COMP1314 Coursework


Counts for 30% of this module

Write UNIX scripts to extract useful data from a very large dataset.

Then structure this data as a database and write queries to efficiently access it.

Write a report on the project in LaTeX

Coursework will be set in the end of Week 2 of the module.

22

# The COMP1314 exam

- 40% of the module
- Testing the underlying theory to the practice

- An entirely computer-based multiple-choice exam
  - However, testing not just memorisation, but **understanding**
  - Type of MCQ you can expect will be covered in revision lectures.

23

# What you need to know for your labs with Angelos Nikolopoulos

24

## A little bit about me…

- Currently a 3rd Year PhD Student within ECS
  - **Research Group:** Electrical Power Engineering (EPE)
  - **Research interests**: Solar Forecasting and Statistical Analysis.
- Have been a lab demonstrator for:
  - Student ECS Mentoring Program
  - Machine Learning Technologies (COMP3222)
  - Data Management (COMP1314)
- Senior demonstrator for COMP1314
  - Feel free to approach me and **ask questions**!

arn1r23@soton.ac.uk

25

# Lab Sessions

- Wednesday - B59/3229 ECS Computing Lab
- Please check the schedule for your cohort slot

**Schedule**

This is a provisional schedule and is subject to change - always check this page to see what's coming up!

| Week | Lecture Topics | Lab session - Cohort B | Lab session - Cohort A |
|---|---|---|---|
| 1 | **Tue:** Introduction to module. Webcast stream.<br><br>**Fri-1:** Unix Fundamentals.<br>- UNIX philosophy.<br>- Introduction to file system<br><br>**Fri-2:** Typesetting with LaTeX. | | |
| 2 | **Tue:** Processes, pipes and filters.<br><br>**Fri-1:** Bash scripts.<br><br>**Fri-2:** Grep, sed and awk - Part-1. | Lab D0: Getting familiar with the shell | |
| | *Coursework released* | | |
| 3 | **Tue:** Grep, sed and awk - Part-2.<br><br>**Fri-1:** Grep, sed and awk - Part-3.<br><br>**Fri-2:** Wildcards and file permissions. | | Lab D0: Getting familiar with the shell |
| 4 | **Tue:** Process management<br><br>**Fri-1&2:** Structured Data | Lab D1: Pipes and filters | |
| 5 | Structured Data<br>Relational Models | | Lab D1: Pipes and filters |
| 6 | Relational Algebra | Lab D2: Structured data and XML | |
| 7 | Normalization | | Lab D2: Structured data and XML |

https://moodle.ecs.soton.ac.uk/course/view.php?id=481

26

# Lab Structure

- D0 – Introduction, setting up your UNIX environment.
- D1 – UNIX bash scripts, pipes and filters.
- D2 – Structured data
- D3 – SQL databases
- D4 – NoSQL databases

**Assessed labs**

27

# Assessed labs

- Each lab D1-D4 assessed based on four equally weighted criteria:
  - **Preparation** – How well are you prepared for the lab. Questions released a week before the scheduled lab.
  - **Progress** and **Understanding** – Assessed during the lab.
  - **Logbook** – Note down answers to lab exercise in your logbook.

28

# A word of caution regarding lab environment and CW

- Some of you are already familiar with Linux/other UNIX systems
  - That's great!
- Some of you probably have your own Linux boxes up and running
  - That's also fantastic!
- In this module, we expect you to use a specified Linux distribution, and not any other environments!
- There are occasional subtle differences between implementations of core tools, so make sure your code works on the environment provided!

29

# What's next!

- Today: Introduction
- Friday lecture: UNIX philosophy + file-system introduction.
- Friday tutorial: LaTeX typesetting system.

- Next week, *D0* lab session on Wednesday (for Cohort B): Getting Started
  - Setting up your environment
  - Becoming familiar with the tools
  - Getting ready for the coursework

30