COMP0235: Engineering For Data Analysis

Welcome

- Many forms of science, maths and statistics require high throughput data analysis
- Often these skill as learnt in an adhoc manner
- This module (and COMP0239) want to teach the Dev-Ops, and ML-Ops that will let you process large volumes of data

Aims

- Introduction to ML-Ops
- Introduction to scaled data storage
- Introduction to deploying applications at scale
- Introduction to running large scale data analysis programs
- Sufficient knowledge to install a data analysis system that can process a lot of data

General Information

Module Lead(s)	Dr Daniel Buchan Prof James Hetherington
Term	Term 1
Teaching	F2F interactive coding sessions
Format	F2F sessions – concepts, theory and practical Home exercises Reading - theory
Assessment	50% Coursework, code and short report 50% Written Exam
Moodle	All information should be here.

Who we are

Dr Daniel Buchan

Who we are

Prof James Hetherington

Who we are

Dr Owain Kenway

Class Information

Size	35
Teaching	2 lecturers
Resources	You must bring your own laptop You will be issued some cloud computers
Courses	Data Science and Machine Learning Software Systems Engineering Scientific and Data Intensive Computing
Depts	2

AI Usage

- **No** – Oral or written exam, test, discussion-based assessments, lab book or results, discussion, drafting
- **YES** – Learning (e.g. tutor), developing code, grammar checking
- https://www.ucl.ac.uk/students/exams-and-assessments/assessment-success-guide/engaging-ai-your-education-and-assessment

Studying

- Attend the classes!
- Classes are F2F
- Catch up with recordings if you miss something
- Try and understand why each topic is linked to the others
- Ask questions during classes
- Post questions on the group notes or forum so the whole class can learn
- Reading the readings!

Missing a Class

- Catch up with recordings if you miss something
- Reading the readings
- Use the forum to ask questions you didn’t get a chance to in the class

Assessment

- Coursework

    Short exercise in installing a dataset, some data analysis tools, and building a system to distribute some data analysis

    The code you used to achieve this

    A short viva demonstrating your application

- Exam

    Short form answer questions

    2 long form answer questions

Feedback

- Please give feedback so we can work out what is working
- One good thing, one bad thing post-its
- Feedback on Moodle ASAP

  Anonymous so be honest

  Simple rating system

Overview

Take you through the whole process of automated **commissioning of machines in consistent configurations**

带你了解整个自动化部署机器并保持配置一致性的过程

...And **file store scaling and data analysis pipeline scaling**

...以及文件存储扩展和数据分析管道扩展

10 Topics

1. Intro to ML-Ops, Intro to Systems Administration, Basic Software Eng Practice
2. Idempotent deploying
3. Intro To Filestores
4. Intro to File & Compute Paralellism
5. Message Queues

10 Topics

6. Pipelines & Scaling
7. Containers
8. Container orchestration
9. Security
10. Integrating everything

Goal

1. Create a set of machines
2. Install the software you need to distribute some analysis
3. Install the data you need to analyse
4. Run the data analysis in a distributed fashion
5. Monitor and secure the “health” of you machines and analysis

What we’re going to teach

A weird combo of software engineering, systems engineering, systems administration and some ML

What *IS* ML-Ops?

Machine Learning Operations

Why *IS* ML-Ops?

Why would we need this?

The scale and complexity of ML-applications

Etherpad

https://etherpad.wikimedia.org/p/COMP0235

Remedial software engineering

1. Basic unix commandline
2. Good python programming practice
3. Introductory use of source control git

Your experience of sys admin or distributed systems

1. What are the challenges/benefits

2. What have you liked or not?

3. What are you looking forward to?