idss_lecture_01_introduction

September 20, 2022

[1]:

```
%%javascript
$('#menubar').toggle();
```

```
<IPython.core.display.Javascript object>
```

1 Introduction to Data Science and Systems 2022-2023

1.1 Lecture Week 1: Introduction

University of Glasgow v20222023a

1.2 Outline:

- Part 1: Introduction and motivation
- Part 2: Structure and logistic of the course.
- Part 3: The basics of data science - a demo

By the end of this week you should:

- know your lecturer(s) and lab assistant
- know the structure of the course (including assessment and schedule)
- become familiar with common definition(s) of data science
- know the basic steps in data science: load data, identify the data type, perform basic quality control, data cleaning/curation, simple visualizations.
- become familiar with Numpy through the self-study and (supervised) labs

[2]:

```
import IPython.display
IPython.display.HTML("""
<script>
  function code_toggle() {
    if (code_shown){
      $('div.input').hide('500');
      $('#toggleButton').val('Show Code')
    } else {
      $('div.input').show('500');
      $('#toggleButton').val('Hide Code')
    }
    code_shown = !code_shown
  }

  $( document ).ready(function(){
    code_shown=false;
    $('div.input').hide()
  });
</script>
<form action="javascript:code_toggle()"><input type="submit" id="toggleButton"␣
 ↪value="Show Code"></form>""")
```

[2]: `<IPython.core.display.HTML object>`

[3]:
```
# standard imports
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
plt.rc('figure', figsize=(10.0, 8.0), dpi=140)
from jhwutils.matrices import show_matrix_effect, print_matrix
from jhwutils import ellipse as ellipse
```

---

2 Part 1: What is data science and systems ?

2.0.1 Data is transforming the world!

Technology trends: (adapted from J. Gonzales)

- 1940-1980s: Hardware - building digital computers
- 1990s: Software industry (developing increasingly more complex software)
- 2000s: Internet industry (online retailing and services)
  - At the same time: Mobile devices
- 2010s: Data industry (sell, buy and share information about people and objects)
  - At the same time: AI revolution
- 2020s: ?

Examples: > - Universities: UofG collects data about students such as attendance and grades to support students and ensure quality of our degree programmes and individual courses. Some of this data is subjective and collected by asking you - the students - questions about your perceived quality of a course or perceived difficulty. Other types of data is objective i.e. attendance, activity level on Moodle etc. >

[Source: Moodle]

- Amazon: When you buy a book on Amazon they record not only your purchase but also you interest (e.g. through searches) or disinterest in other produces.

  [Source: Amazon]

- Politics: Politicians use data collected on e.g. Facebook for target political advertisement.

  [Source: BBC]

- Dating sites: Dating sites ask you 100s of personal questions with the promise that they can find the perfect partner for you (others just show a picture…).

- Covid tracking: Many countries have launched Covid tracking apps which informs you if you have been in the vicinity of an infected person.

2.0.2 Is data all you need?

Data science an highly interdisciplinary activity concerned with gaining insights from often large and messy data which you sometimes have to collect, store and curate yourself. It is the application of data centric, computer science, and inferential thinking to understand the world (science) and solve problems (engineering).

[Source: Drew Conway 2010]. Warning: There are many variations of this figure (not all sensible).

Example of questions we might try to answer with data science: - How much exercise did you get in 2020? - Is the use of the COMPAS algorithm for prison sentencing in the US fair? - How will Brexit influence the UK economy? - What should we eat to maximize our live expectancy? - Who should get a COVID-19 vaccine first? - Has the world gotten better or worse over the last 10 years (on average)? - Will there be a forth wave of COVID19 in the UK?

2.1 Example: Covid-19 cases in the UK

[Source: Daily Covid cases from OutWorldInData, CC BY]

Interactive version

Question: Why does the number of cases drop on certain days? (consider what kind of information and knowledge you would need to answer this question)

2.2 Data science and systems overview

2.2.1 Question

-***why do you want to (or why did you) collect/analyse the data?*** > > a) Problem solving (engineering): E.g. recommender systems, student support, Covid tracking. > > b) Understand the world (science). E.g. biological image analysis, human behavior. > > c) Don't know (yet): Collecting data without a specific goal is not uncommon expecting a later need or question to present itself. >

2.2.2 Data acquisition

*-**how do we (or did we) obtain the data - and in a robust and fair manner?*** > > Given a specific question of interest, a key challenge is how to obtain the required data which would allow you to answer the question. This involves asking the right question to the right people - or measuring the correct physical signal using the correct sensor. > > Many premature or even wrong conclusions have made due to biased dataset, faulty data, wrong sampling or other issues with basic data acquisition. > > > **Experimental design**: There is an entire discipline in statistics dedicated to the design of experiments to make sure the data is obtained and collected in a fair and robust manner. For example: how do you designs polls for election polls to query a representative sample of the population (hint: there is room for improvement). > > **Provenance**: As a data scientists you should ideally know where the data came from and if/how it has been pre-processed before you got your hands on it! >

2.2.3 Data storage

- ***how do you store the data so it is secure and easily accessible?*** > > - **Security** > > - **Accessibility** > > - **Efficiency** >

2.2.4 Data curation, quality control and exploration ("understanding the data")

- ***what's the data type, are there mistakes, are there missing data or other issues that needs to be addressed now (or perhaps in the data acquisition process)*** > > **Exploration:** > - What does the data represent represent and how is it represented? > > - What data types are in the dataset (perhaps you already know this if you collected the dataset)? This includes the structure structure, data types of columns and the granularity of rows in the dataset. > - Numerical data, which represents amounts or quantities. For example: temperature, price, and height. > - Nominal data, which represents categories that do not have a natural ordering. For example: political party affiliation (Labor, Conservative, SNP), mood (happy, sad), and country (UK, US, China). > - Ordinal data, which represents ordered categories. For example: Shirt sizes (small, medium, large), Likert-scale responses (disagree, neutral, agree), and level of education (high school, university, graduate school). > - Temporal: E.g. dates, timestamps, time series. > - Freetext which represents comments, tags etc > - … > > - Summary statistics: The distributions of qualitative and quantitative data (e.g. measures of center and spread). > > - Relationships between quantities in the dataset (basic correlation and visualisation). > > > **Quality and curation** > > Many real-world dataset often contain "messy", missing or faulty data which must be accounted for before processing to the next steps. A few common examples include: > - Incorrect or unrealistic values. For example, dates in the future, colours that don't exist, negative counts, or large outliers. > > - Violations of obvious dependencies: For example, age and birthday don't match. > > - Data provided by humans: These are typically filled with spelling errors and inconsistencies. > > - Data falsification or adversarial attacks > > - Missing data > > - **Key rule**: All steps involved in modifying or augmenting the dataset must be clearly documented and justified. > > > **Preparation** > - Feature extraction: Many domains rely on expert knowledge to extract. E.g. in analysing speech we often extract socalled cepstral coefficients which represents what is being said without capturing the charateristics of the speaker (e.g. female vs male). > > - Transformations: Transform the data to be. For example: Reduce the dimensionality of numerical data, transform a sentence in to a bag-of-words. > > > - **Key rule**: All steps involved in pre-processing/preparing the data must be documented. > > **Visualization** > - Visualisation is a essential components of modern data science from undersatnding the data, cleaning it, displaying the results and prediction to finally communication the result of the whole analysis. >

2.2.5 Modelling & Inference

- ***how do we find patterns and trends in the data which allows us to make inferences and predictions about the world?*** > > *Essentially, all models are wrong, but some are useful.* George Box (1919-2013) > > A model is an *idealised* representation of for example the world, a system or a person which can be used to do prediction, estimation and description. For example a model of the world is the differentiation equations governing the trajectory of particle, e.g. the path that an call with mass in motion follows through space as a function of time.
> > > In a data science context a model is learned (at least to some degree) from the available data. This process results in an approximation which might be useful if the the approximation is reasonable precise. For example: predicting the popularity of a a baby name based on previous years popularity. > > > A typically modelling process often looks like (many of you will explore this further in the ML&AI course): » 1. Select a model (e.g. a linear model mapping from year to the frequency of a baby name). » 2. Select a loss function. » 3. Fit the model by minimizing the loss using numerical optimization. » 4. Evaluate the results in terms of generalization error (visualization, etc) » - … repeat if needed >

2.2.6 Evaluation

-***where we produce the visualisaiton and reports, and provide objective observations on the findings in light of the data, data pre-processing and models*** > > - Visualisation > - Descriptions > - Critical reflection >

2.2.7 Results & Interpretation

-***where we present the factual results and interpretations to decision-makers*** > > Traditionally, data science is often about presenting clear and objective results and findings to decision-makes, who - based on your findings - will make a decision. This distinction is blurred and becoming increasingly vauge as data science as a whole starts considering and accounting for the consequences of the recommendations (e.g. decision-theory).

---

3 Part 2: Structure and logistics of IDSS

3.0.1 Focus of this course

Mathematical foundations using Python:

- Vector spaces (data representation)
- Computational linear algebra (e.g. PCA/SVD)
- Optimisation
- Probability (the basics and inference)

Data storage:

Aside from built-in Python packages for handling tabular data, we will be presenting core material on: - (Databases) - Indexing - Querying

Other skills

- Numpy (in detail) and Pandas (self-study / recommended material)
- Visualization
- Critical analysis and reflection
- Practical with a real-world data science case

The course is not: Machine learning, artificial intelligence, a "tools course" teaching you specific data science systems .

- **Credits and workload**:
- The course is a 15 credit course which amounts to about 150 hours of work.
- **Lectures** (Tuesdays)
- In person Thomson 238. Recordings of the live lecture will typically be provided on Moodle as Zoom/MS Streams
- **Lab sessions** (Fridays)
- In Boyd Orr 1028
- You will be allocated to a 1h group on MS Teams where you can interact with and the allocated lab assistant and your teams mates. You can find the Lab instructions (and deadlines) on Moodle.
- Peer support: Please be active in the MS Team groups and answer each other questions etc.
- **Assessment**
- Labs (4 labs, each worth 6% of the final grade, recommend effort is 9h each) - deadlines posted on Moodle.
  - Labs are typically due two weeks after hand-out (see Moodle / LTC system for details).
- Weekly quizzes (10 quizzes, 6% in total, top 80% counts, i.e. you can drop two quizzes and still get top mark)
- Exam, written 2 hour in April/May.
- **Material**.
  - A lecture notes (e.g. interactive notebooks, slides) with references to additional/supporting material.
  - Explore on your own! There are many excellent source of information on the internet and we strongly encourage you to read outside the predefined curriculum.

A note on marking and effort: The suggested 9 hours for each assessed lab is the time we expect you to spend on the lab (effective work hours) to get the grade you would normally get in a similar course. An A1 (top grade) student should spend 9h getting an A1, a B2 (average) student should spend 9h getting a B2, etc.

A note on plagiarism: Don't do it! We automatically check all submissions against each other - and we are good at finding patterns. If you find yourself in a situation where copying other peoples work is the only option then reach out to the teaching staff on the course and we can help/guide you.

A heads up: We will be monitoring and analysing the progress throughout the 11 weeks - we are data scientists after all! We might make changes to the schedule as we go along based on your feedback and our observations.

4 Part 3: Data Science basics - *by example*

We will go through a simple data science example in Python to demonstrate the following aspects of data science: - Question formulation - Data acquisition - Data storage and access - Data curation and preparation - Quality control - Curating (/cleaning) - Visualization - Basic modelling - Evaluation - Results & Conclusion

[A Jupyter notebook will be presented during the lecture]