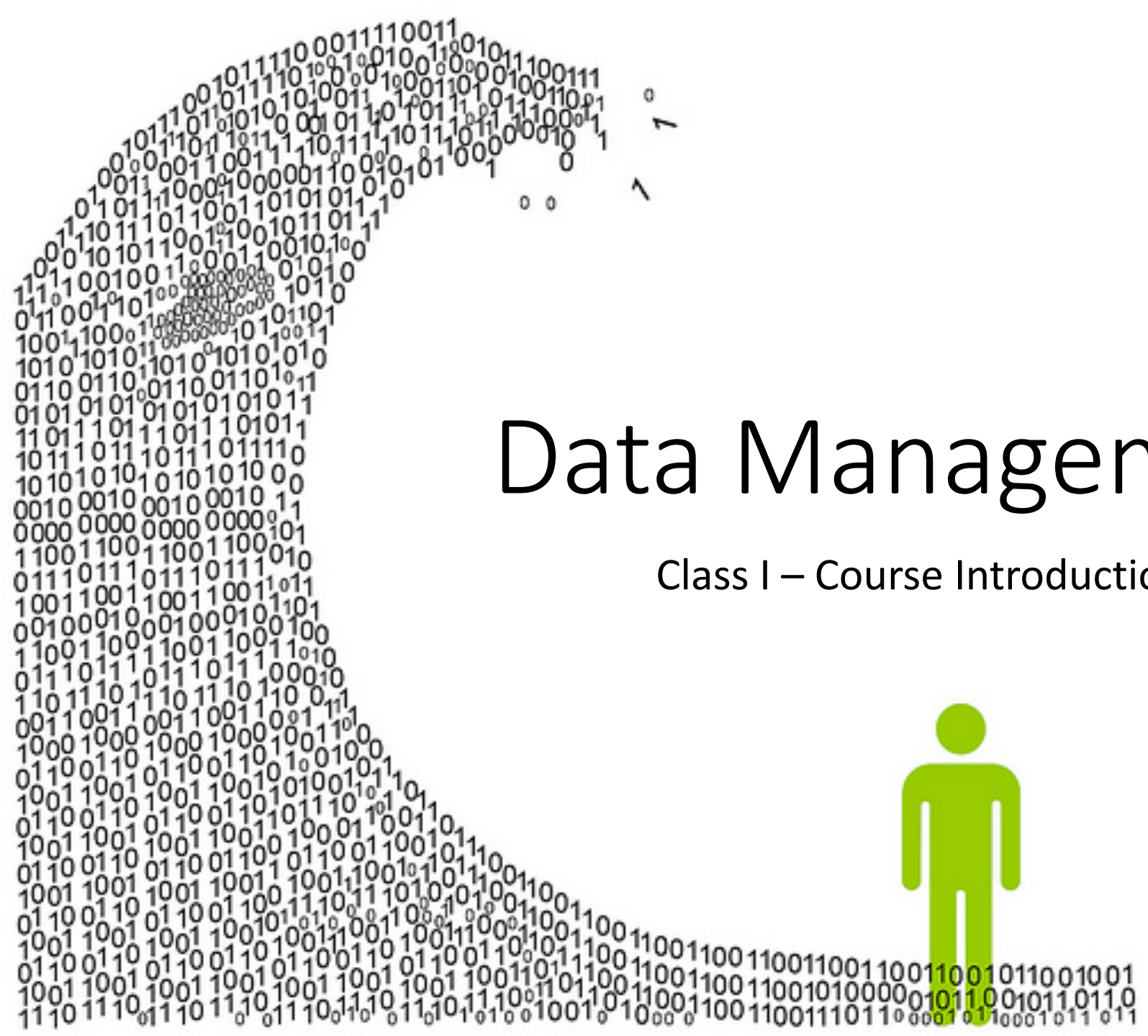
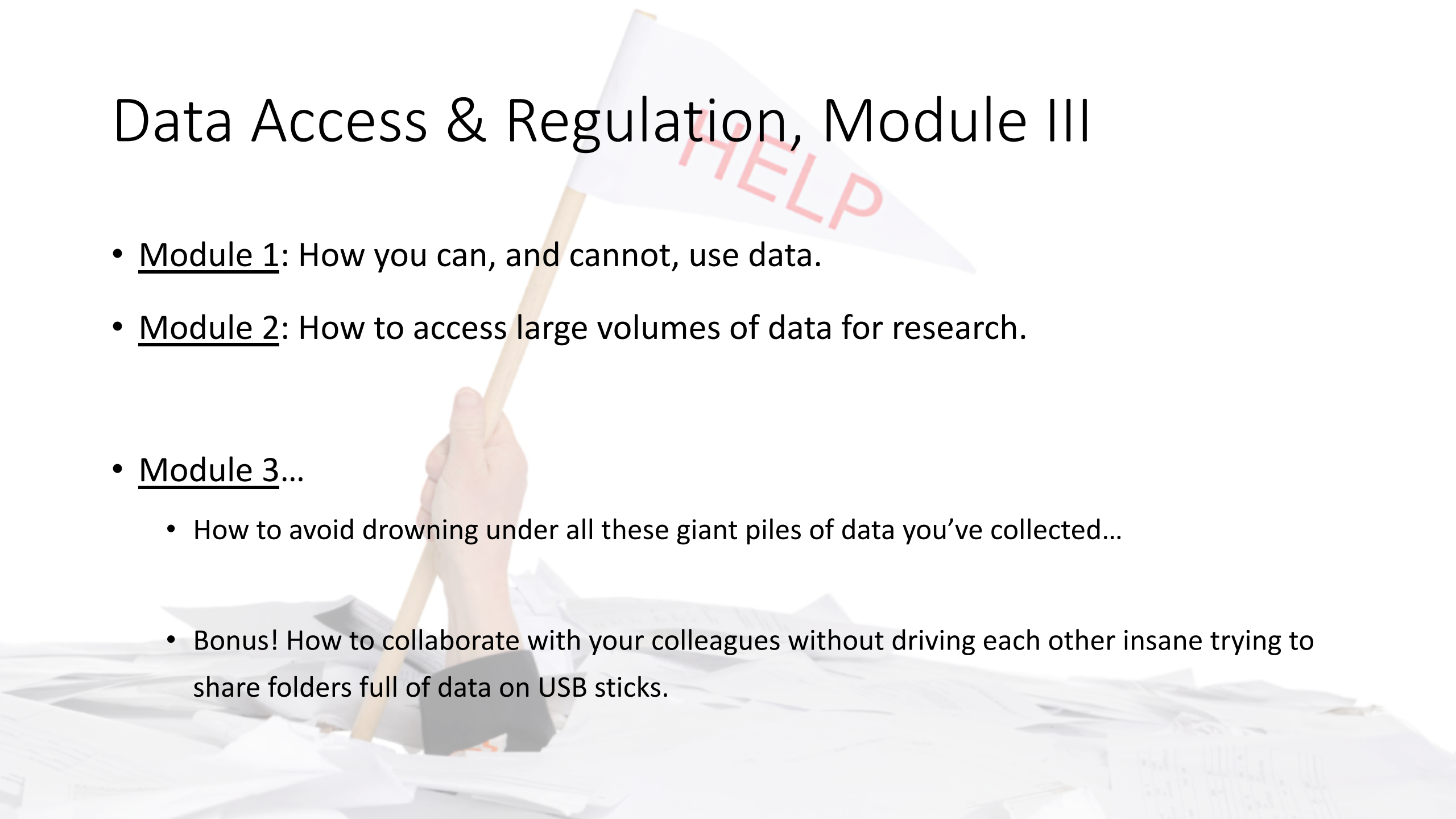
Data Management

Class I – Course Introduction



Data Access & Regulation, Module III



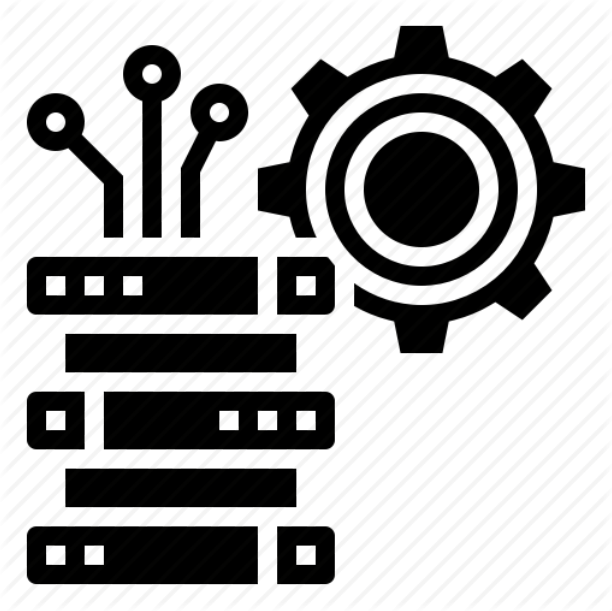
- Module 1: How you can, and cannot, use data.
- Module 2: How to access large volumes of data for research.

- Module 3…
  - How to avoid drowning under all these giant piles of data you've collected…
  - Bonus! How to collaborate with your colleagues without driving each other insane trying to share folders full of data on USB sticks.

Who am I?

- Researcher at the **Waseda Institute of Political Economy** in Tokyo
- We work with a lot of large data sets: social media data (Twitter posts, social network connection data), newspaper data, large-scale public opinion surveys…
- I have some (*very minor!*) background in programming, so ended up handling data storage, access and analysis tasks for many research projects.
- Largest to date: an archive of Twitter data for academic research… ~5TB (5,000GB) of social media posts.

Data Management: Challenges



- Three major trends have made it increasingly important to have a good strategy for managing research data…

1. We increasingly use *complex data* – text, images, audio etc. – for social science research.
2. It's increasingly common to *collaborate* with colleagues all over the world on research projects.
3. Journals are increasingly aware of the need for *data sharing* – many journals won't accept articles if you don't make your research data available!

“Complex Data”

- Traditionally, the data used for empirical / quantitative analysis in the social sciences was *structured* data – tables made up of variables and observations, like an Excel spreadsheet.
    - This data could be messy – missing values etc. – but it obeyed a clear structure.
- Today, we can analyse many other kinds of data… Any kind of text (from legislative speeches to social media posts), images, audio recordings, network connections between individuals, etc.

“Complex Data” ②

- What do these new kinds of data have in common?
  - They’re usually *unstructured* – i.e. the information they contain isn’t numeric or tabular, so it requires a lot of pre-processing before we can perform statistical analysis.
  - Sometimes, they’re structured, but not in a conventional, tabular way – network data, for example, has a structure you can’t easily represent in a table.
  - More importantly… They’re *big*. Projects using this kind of data end up storing far more data than you’d ever imagine handling if you were looking at survey results etc.

Collaboration

- When you work alone, your data management strategy is still important (so you don't waste your own time and effort!) but in collaborative research it becomes vital.
- Your colleagues need to *access* research data; to *modify* it in a way that's tracked (and reversible!); and to ensure everyone is always using the latest version of the data.

Collaboration ②

- Even if you're working with people in the same office or campus, that can be tricky – but working remotely with people in different universities or countries makes it even more challenging.
- I can't just email / Slack you to ask for a certain data file if I'm 9 time zones away and you're asleep… We need a persistent data store that's accessible to us all.

Data Sharing

- In the past, it was often difficult or impossible to access the data other researchers had used for their work – especially if it was published a long time ago.
- Now, many journals demand that you make your research data available in a permanent, easy-to-access archive as a condition of publishing your article.

Data Sharing ②

- This means you need to keep your data in a “clean”, easy to understand format; carefully record how you’ve changed or filtered it; and be able to output it in files other researchers can use.
- This doesn’t stop with journals; many public- and private-sector bodies demand total transparency with research data, to ensure high-quality analysis is being conducted.

Today's Buzzword: "Big Data"

- Looming behind all of these issues and challenges is the idea of "Big Data" – which is a very popular buzzword in tech circles, and increasingly in political circles too.
- There are various definitions of "Big Data", and some of what we'll cover in this module certainly qualifies as "Big Data" handling and management.
- "Big Data" refers to the size of the data files (usually data sets so large an average PC can't process them) – but also to the broad idea that our society is producing huge amounts of data every minute of every day.

Module Objectives

- I'll introduce you to a set of technologies and tools that can help you solve these problems and challenges in your research projects.
- I can't make you a data management expert in two weeks – but I can show you the kinds of solutions that are available to you and the basics of how you work, so when you encounter a real challenge you'll know where to start looking for solutions.
- I do want you to gain some technical skills – but it's much more important to gain a good understanding of the concepts behind data management, and why certain solutions are a good fit for certain problems.

Software & Tools

- **Python** – free programming language, currently the most popular in the world. Very widely used in the private sector and in some research areas, and excellent for data management tasks.
  - We're going to use the "**Anaconda**" version of Python, because it's easy to download and install.
- **MySQL** – widely used free database software, based on the SQL database language that's been widely used since the 1970s.
  - We'll also use the **MySQL Workbench** free software to examine our databases.
- **MongoDB** – a popular example of the "NoSQL" style of databases which are great for unstructured types of data.
  - We'll also use **Robo 3T**, a tool for looking into MongoDB databases.

Module Outline

- Week One:
  - Working with Python (programming language) and SQL (database language).
  - How to get your data into Python, and then into a database;
  - How to sort, filter and export it from the database;
  - How to control the database directly from your program.
- Week Two:
  - Other kinds of databases: NoSQL (for unstructured data) and Network Databases (for network data).
  - Cloud Services – where to turn when either your data or your research team (or both) gets really, *really* big.

Today's Objective…

1. Students self-introduce themselves and their projects.

2. Ensure that everyone has the software for the course installed and working on their laptops:
    - Python (Anaconda)
    - MySQL
    - MySQL Workbench
    - MongoDB
    - Robo 3T