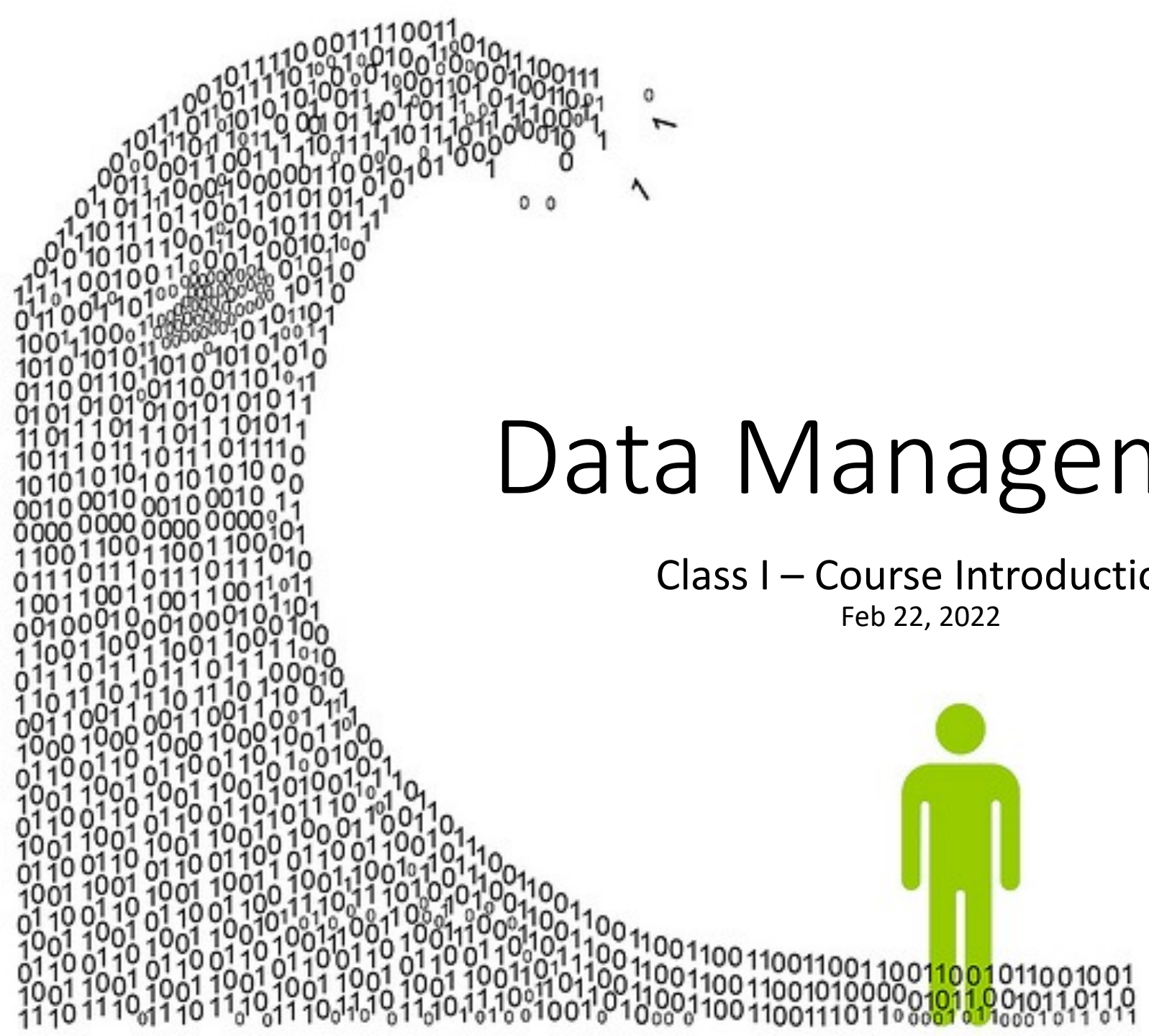
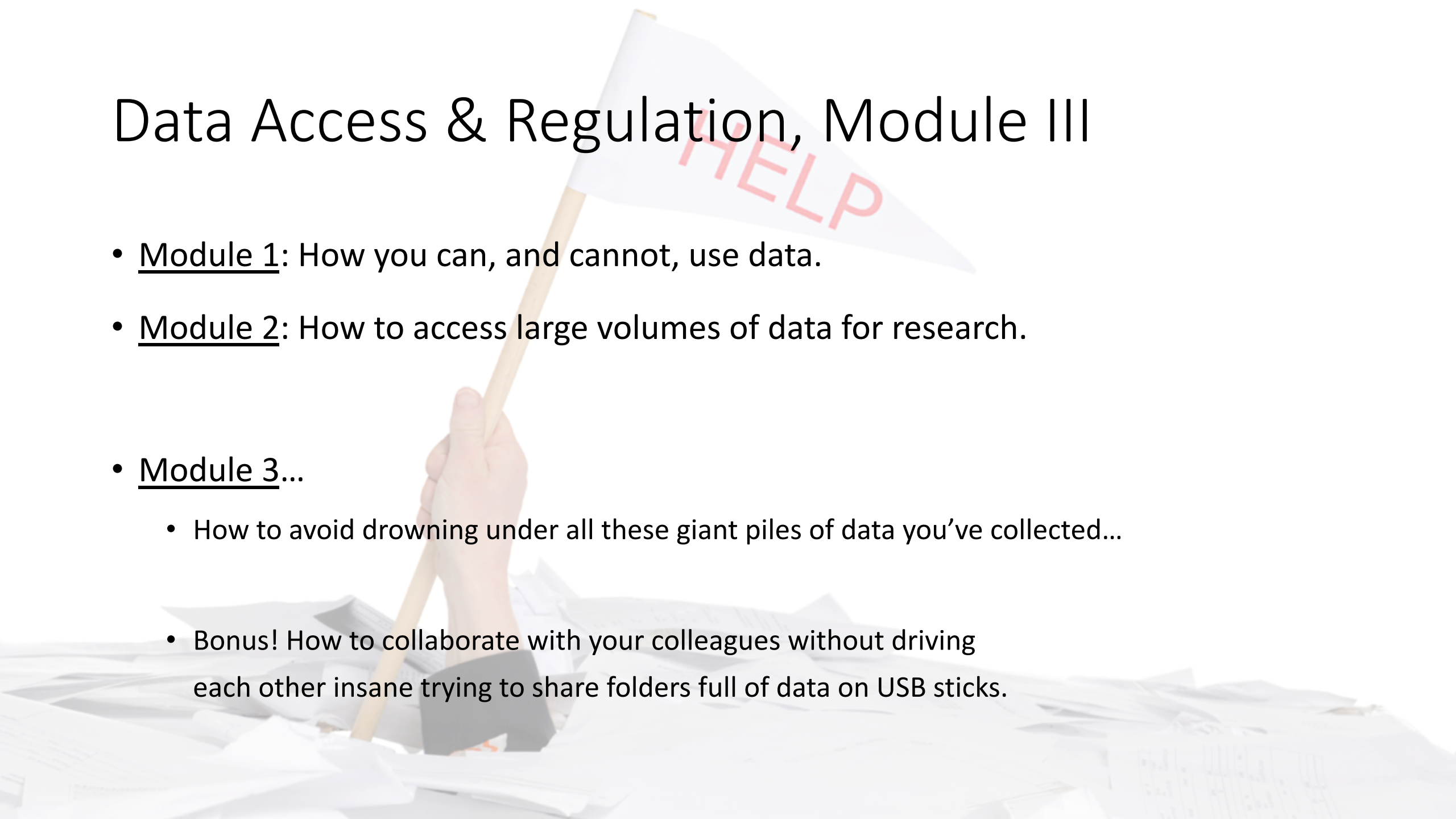


Data Management

Class I – Course Introduction

Feb 22, 2022

Data Access & Regulation, Module III



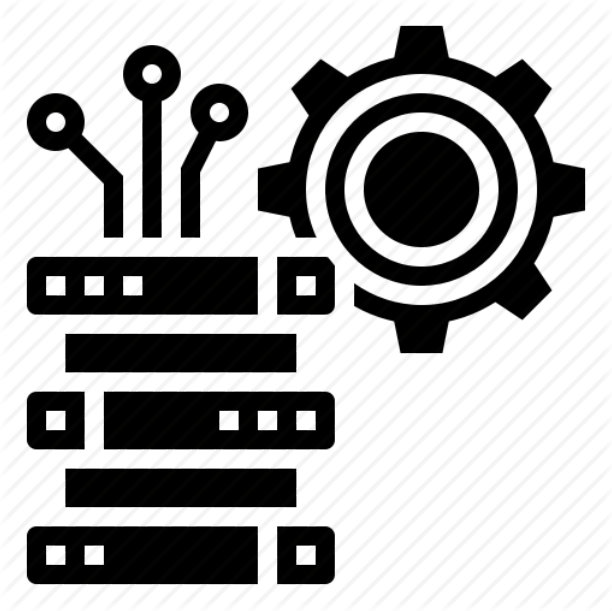
- Module 1: How you can, and cannot, use data.
- Module 2: How to access large volumes of data for research.

- 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 (from next month, I'll be an Assistant Professor at the **Waseda Institute of Advanced Studies**).
- 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 (*a little!*) 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… ~7TB (7,000GB) 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 – who may be using very different software and tools to you.
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.

Collaboration ③

- Another practical problem is that often, your colleagues will not use the same tools (analysis software, programming languages or even computer operating systems) as you.
- Many research projects use a mixture of Python and R, plus other tools (SPSS, Stata, GIS etc.) depending on the details of the project. It's no good if your data only works in <u>one</u> of those tools – it has to work in them <u>all</u>.

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"

- 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

- **Programming Languages:**
  - **R** – you should all be familiar with R to some degree by now. Most of our classes will be conducted in R, and I'll show you how to use it to interface with different kinds of databases.
  - **Python** – currently the most popular programming language in the world. Very widely used in the private sector and in data science. We're going to briefly use the "**Anaconda**" version of Python (because it's easy to download and install) to see how moving data between different programming languages can cause problems, and how to avoid them.
- **SQL** – This is a database language that's been widely used since the 1970s. Many different databases use a version of SQL; the most popular are **PostgreSQL** and **MySQL**, both of which are free software. In this class we'll use **SQLite**, a very small and lightweight database package that can easily be installed in R.
- **MongoDB** – a popular example of the "NoSQL" style of databases which are great for unstructured types of data.

Module Outline

- Week One:
  - **Tuesday**: Introduction class (now!)
  - **Wednesday**: On-Demand Lecture: Key concepts and tools for data storage.
  - **Thursday**: Basics of data access in Python.
  - **Friday**: Common problems with data sharing, and how to fix them in R.
- Week Two
  - **Tuesday**: Introduction to SQL Databases
  - **Wednesday**: More advanced SQL
- Week Three:
  - **Tuesday**: Wrapping up with SQL; introducing NoSQL and MongoDB
  - **Wednesday**: More advanced MongoDB
  - **Thursday**: On-Demand Lecture introducing cloud services and network databases.
  - **Friday**: Preparing a Data Management Plan

Module Assessment

- There will be an assignment every day for this module. It won't necessarily take long, but if you're not sending me an email with an assignment each day, you've missed something.
- 60% of the module's grade will be made up from these assignments.
- I know you are not programmers. You won't lose marks for running into technical challenges or problems you can't solve – I just want to see that you're making a good effort and trying to find solutions. (Hint: Use Google! A lot!)

Today's Objective…

1. Students self-introduce themselves and their projects.

2. Ideally, get the software for the course installed and working on everyone's laptops:
    - Python (Anaconda)
    - R libraries (RSQLite, mongolite)
    - MongoDB
    - MongoDB Compass

Assignment 1

- Full details of each class' assignment can be found (along with these slides) on my site: **www.robfahey.net** (click "Courses/Workshops" in the top menu).

- In brief: I want you to write short descriptions of two datasets; your own dataset from your existing project, and a new dataset you find online (e.g. on Kaggle.com) that you find interesting.
- These descriptions should be both practical (what is this data? What's it about? Where is it from?) and technical (how many observations? Which variables? What format is it in?).
- Write these in an email and send them to me (robfahey@aoni.waseda.jp) by Thursday's class.