**Tena Koutou Tena Koutou Katoa,**

Thank you all for meeting with me so I can tell you more about myself and my professional experience.

For the next 10 minutes, I will focus on explaining about the process I follow to identify relevant data sources and provide an example where I have used my data wrangling skills to prepare data for analysis and visualization.

First, I would like to show you this diagram which illustrates my process for identifying relevant data sources. In the past, I've have used the Preferred Reporting Items for Systematic Reviews and Meta-Analysis methodology (PRISMA). This methodology is a systematic approach often used to conduct literature reviews. While doing my PhD, I've used this methodology to gather the most relevant papers in my area of research, which I will explain further in the following slides.

A central topic of my doctoral thesis was the application of Species Distribution Models, also known as Ecological Niche Models. For simplicity, I will refer to this as SDMs for short. In a nutshell, SDMs are applications of computer algorithms used to predict the distribution of a species across geographic space and time using species occurrence and environmental data.

Tipically, an SDMs exercise consist of 6 major steps which I'm not going to go over the nitty gritty details in this talk. Here, I will just focus on Step 1 to 3 because this are the steps where data wrangling skills come in handy.

In this diagram, I am showing an actual application of an SDMs which comes from one of the chapters in my doctoral thesis. One of thesis chapters focused on using SDMs to predict the distribution of an invasive species in New Zealand. In the following slides, I will guide you through the steps in the process.

Before writing any code, an Ecologist needs to make a decision about what taxa and for what purpose the predictive models will be used for. In my case, I wanted to use the invasive kelp *Undaria* pinnatifida because of its negative impacts in New Zealand and the data available for modelling its distribution. So, that was easy, we are done with Step 1.

Now let’s see some code in action. To show you the code in interactive mode I will open the file in R studio.

To start with Step 2, I will download some environmental datasets. For modelling the distribution of marine species, a great deal of datasets is readily available for download. These are raster layers of global coverage at a coarse resolution of 9 km and finer grain at 1 km. A very useful R package called *sdmpredictors* contains functions to download and load files into R environment. The code bellow shows an example of data retrieved from the Bio-Oracle dataset. There are close to 50 different variables available, but for the purpose of this demo, I’m only retrieving 5 layers to save time.

Since in the particular case of *Undaria* we are only interested in making predictions around New Zealand, I’ve used the raster package to crop the layers extracted to include only New Zealand coastal waters. Now, we can just plot the raster to make sure the raster extent is correct.

Step 2 continues

For marine species, occurrence data is available in several biodiversity databases. In the case of *Undaria*, data is available for download from three data sources. The Ocean Biodiversity Information System (OBIS), Geographic Bioinformation Facility (GBIF) and The Marine Biosecurity Porthole.

While there is nifty R library called *spocc* that can be used to retrieve occurrence data from GBIF and OBIS, data from the Biosecurity Porthole, can’t be extracted with code. This data has to be imported manually.

Whit this data extraction we can clearly see one of the main drawbacks of working with data from global biodiversity repositories. Name conventions are not standardized, column names differ, species names sometimes include the authority, and coordinate system is either UTM or long/lat.

In this example, I am using basic functions from *dplyr* library such as *mutate, rename* and *select.* But only, because here I am working with a few data frames. For a larger number of items, a custom function will be required to merge the data.

To inspect the data merge was done correctly, I simply plot a map of the data using functions from *ggplpot2* library.

Now let’s move on to Step 3 …

From the previous steps we can conclude that presence data for *Undaria* in New Zealand is abundant, approximately 9,000 presence records. But what about absence? To make predictions we also need locations where *Undaria* is not recorded. Unfortunately, absence data is not easily available. However, pseudo-absence or ‘background’ points can be generated using a surface range envelope that differs from the locations where *Undaria* is present. To do this, the *biomod2* library has a function that can create the absence data needed to make predictions.

The same package has a useful function that can format the environmental and occurrence dataset to be used for ecological modelling. It basically, extracts the values of environmental layers from both presence and absence locations and combines them into a single object that can be used as input for SDMs.

For example, for *Undaria* the input for models consists of 8903 presences, 5000 background points and 6 explanatory variables.

Unfortunately, due to the time constrain I can’t show you the next steps in the process of modelling *Undaria* distributions but if you are interested in seeing the models, I can share with you the code and input data.