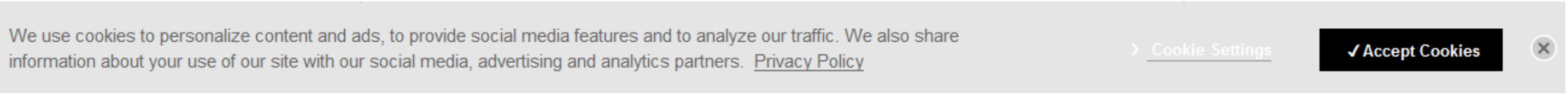
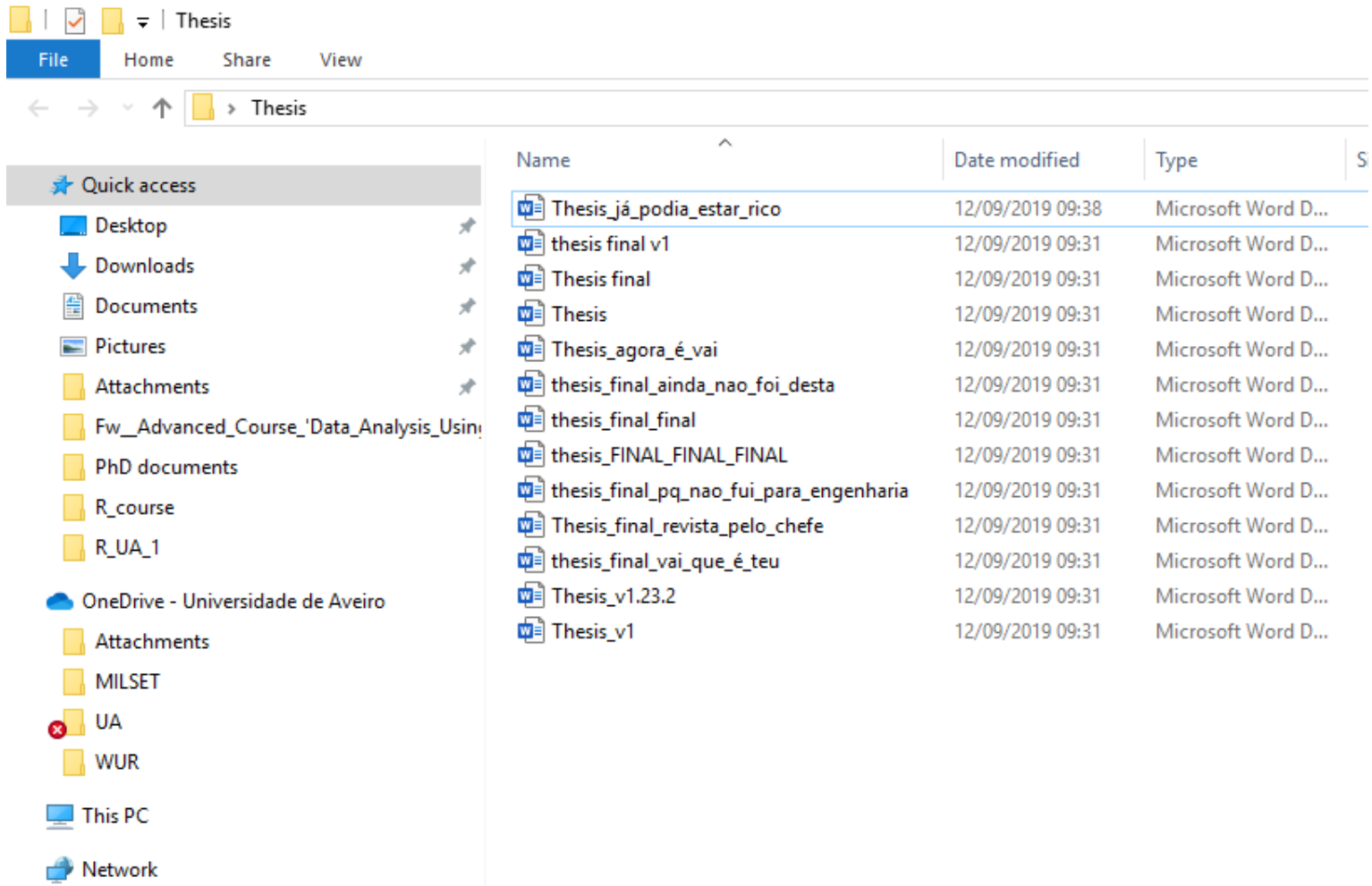
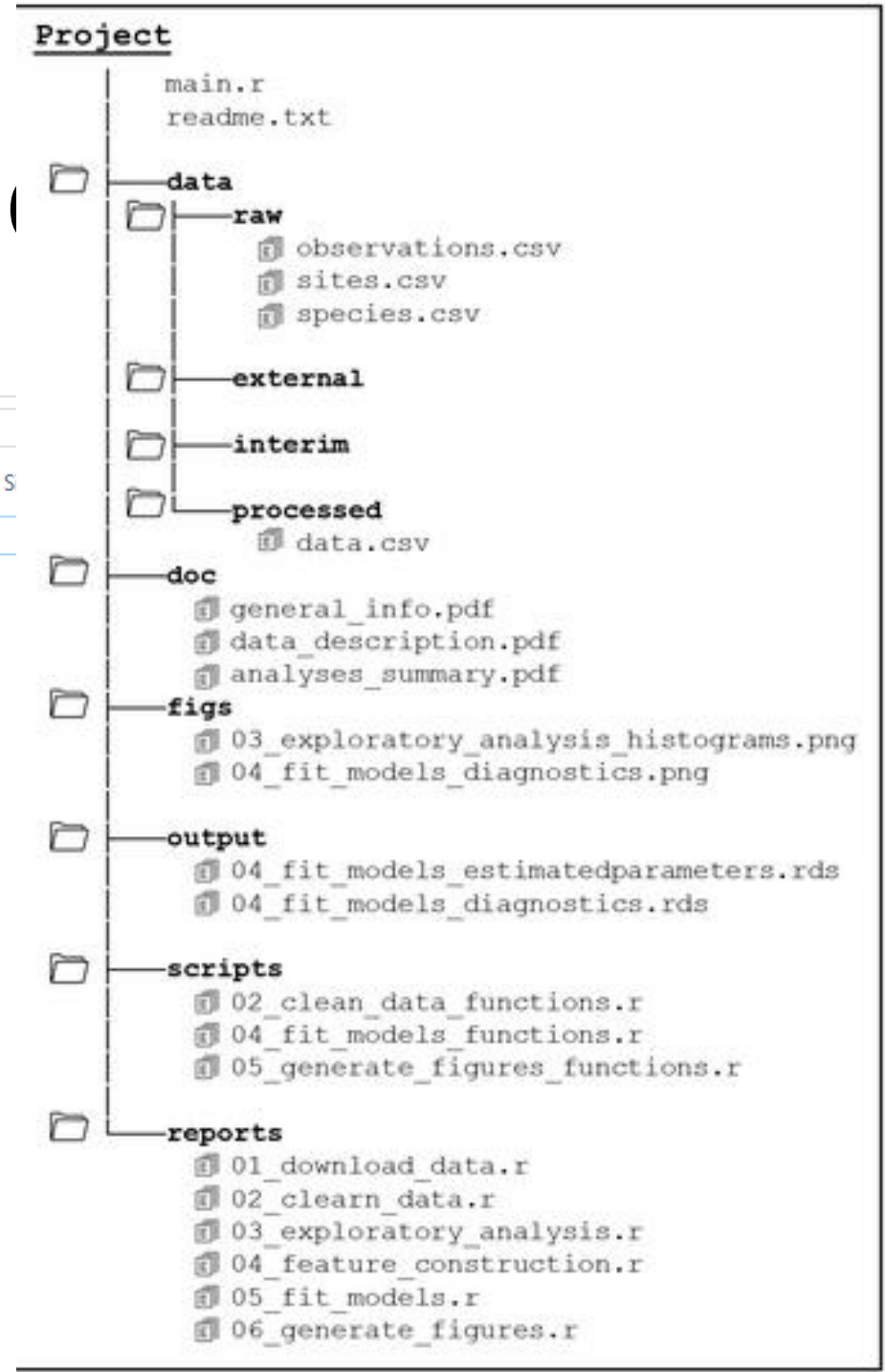
Can big data save the world?



eduardobatista@ua.pt
@Batis_Eduardo7

Don’t forget to tell a st





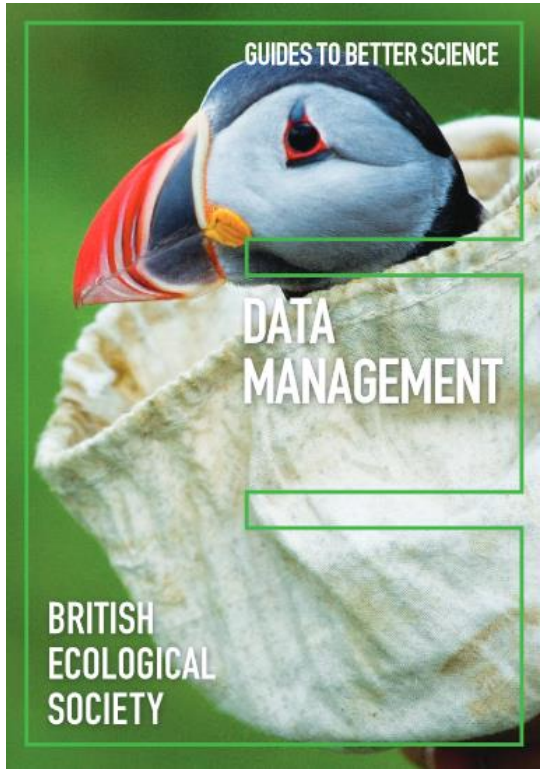
What is big data?

Big data	>5TB
Medium data	10GB – 5 TB
Small data	<10GB

→ R is great at this!

Size of your data > RAM

Can ~~big data~~ save the world!?
small data



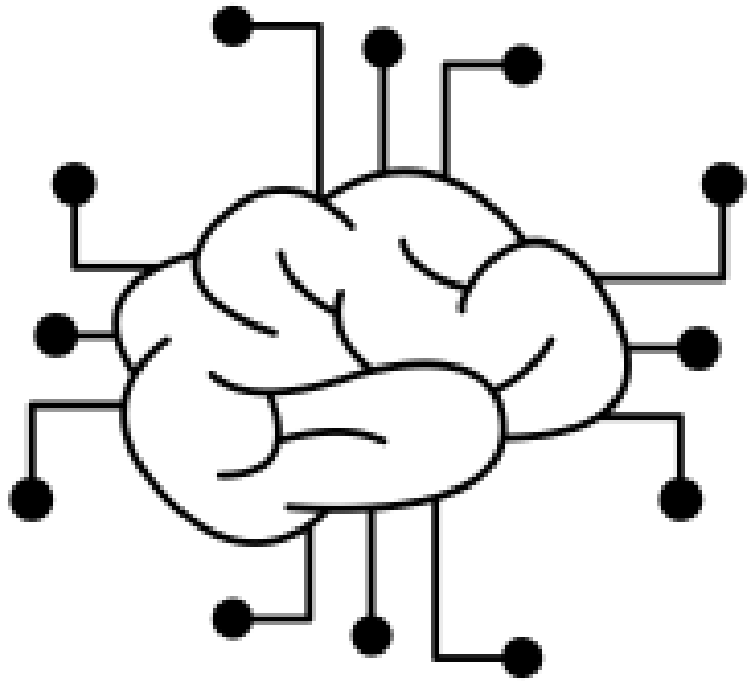
Regulating the internet giants

The world's most valuable resource is no longer oil, but data

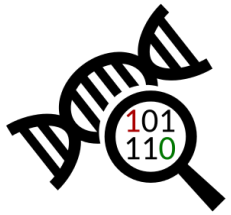
The data economy demands a new approach to antitrust rules

How data can be "valuable" in your research!?

The challenges of big data in biology



Input



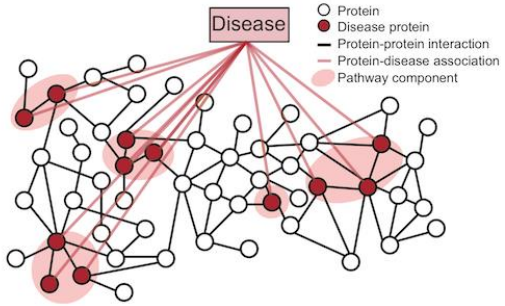
Genome

Financial options

Location

Health status

Protein interactions



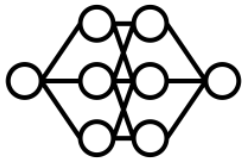
Disease pathway

Social behaviour

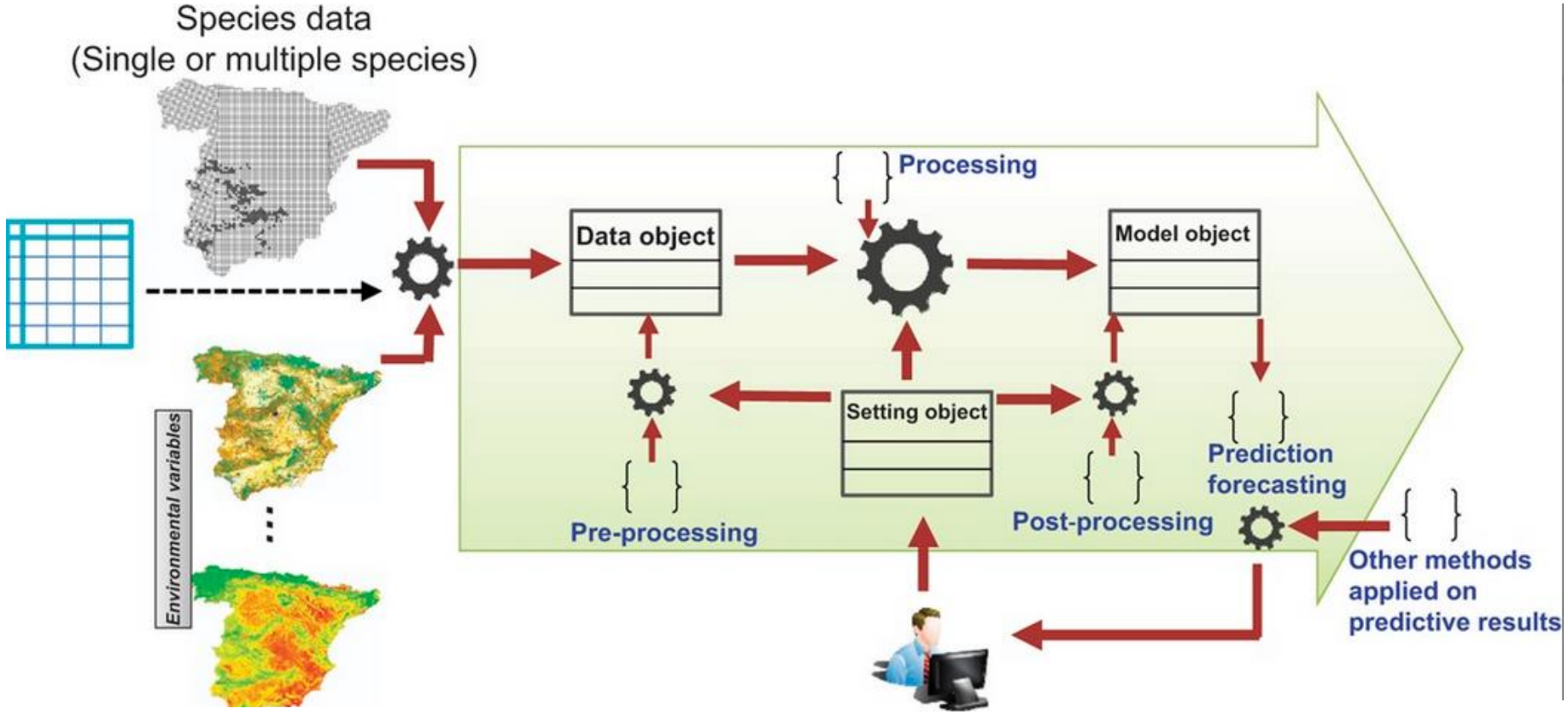
Environmental data

Population studies

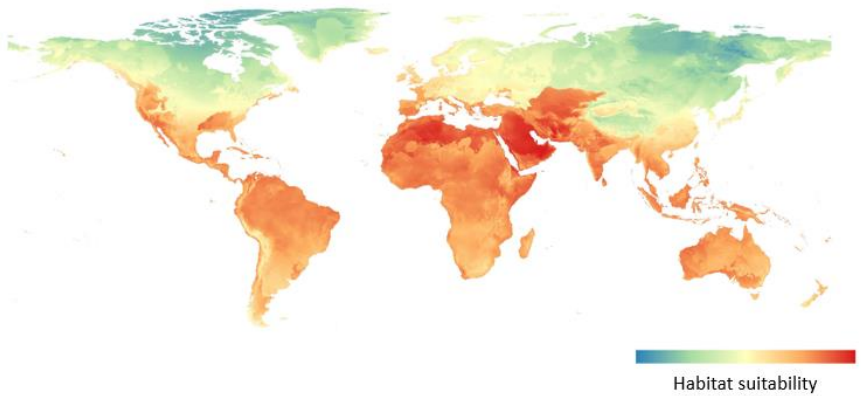
A beer a day keeps the doctor away!



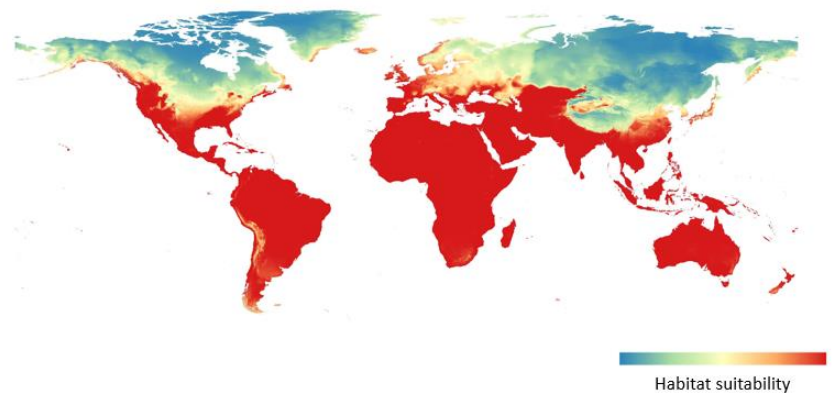
Output



Global/local suitability prediction – Preview



Global/local suitability prediction – Preview



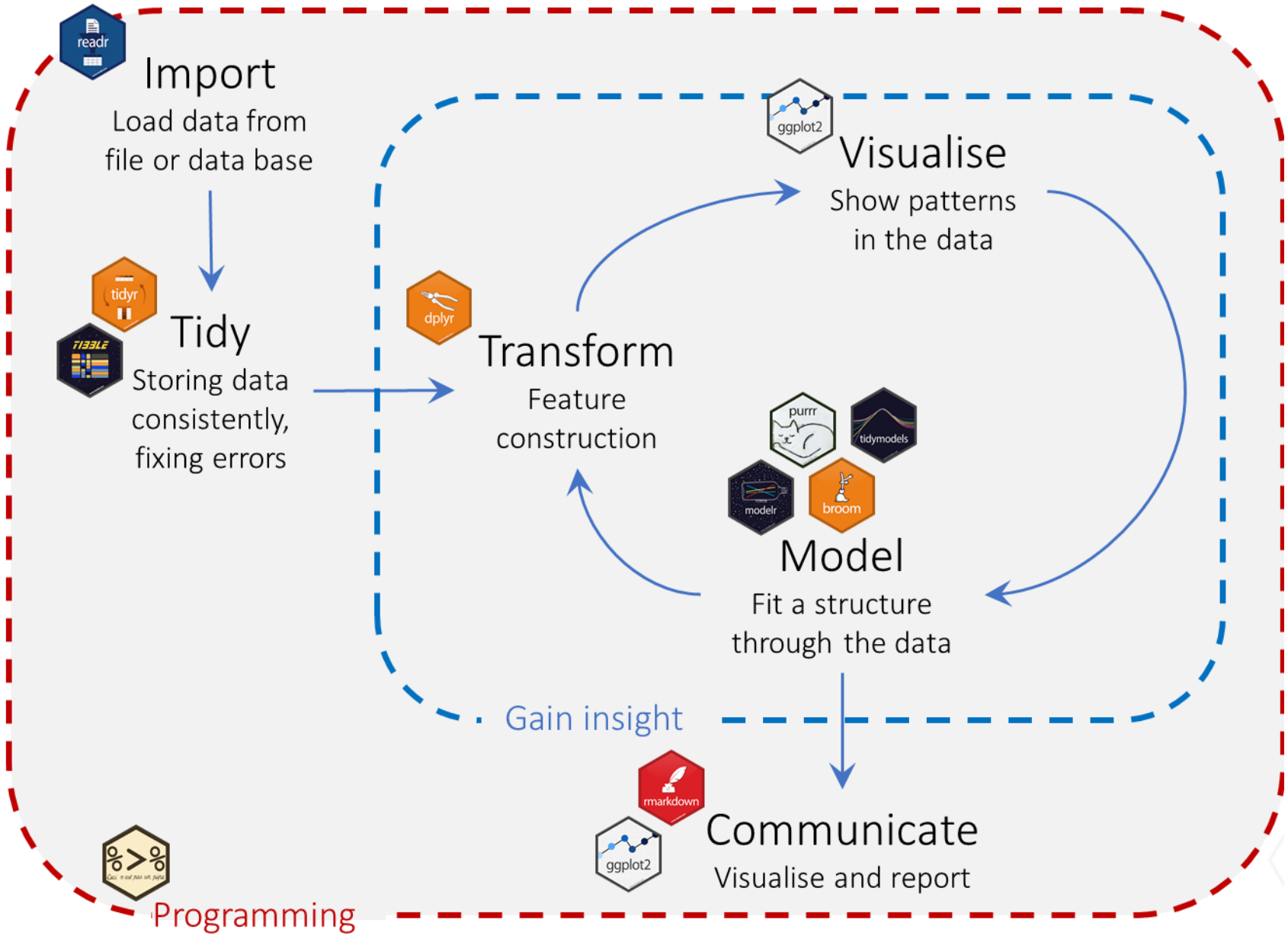
Why should I manage data?

‘Would a colleague be able to take over my project tomorrow if I disappeared, or make sense of the data without talking to me?’

‘Will I be able to find and reuse my own data or recreate this analysis in 10 or 20 years’ time?’

Potential benefits of good data management include:
• ensuring data are accurate, complete, authentic and reliable • increasing research efficiency • saving time and money in the long run – ‘undoing’ mistakes is frustrating • meeting funder requirements • minimizing the risk of data loss • preventing duplication by others • facilitating data sharing • ensuring data discovery and reuse • Reduction of your PhD frustration and saves you time for a beer in the end of the day!

Data lifecycle



https://shiny.rstudio.com/gallery/

Big data in biological sciences

GBIF Global Biodiversity Information Facility

Occurrence records	Datasets	Publishing institutions	Peer-reviewed papers using data
1 338 285 019	45 985	1 450	3 859

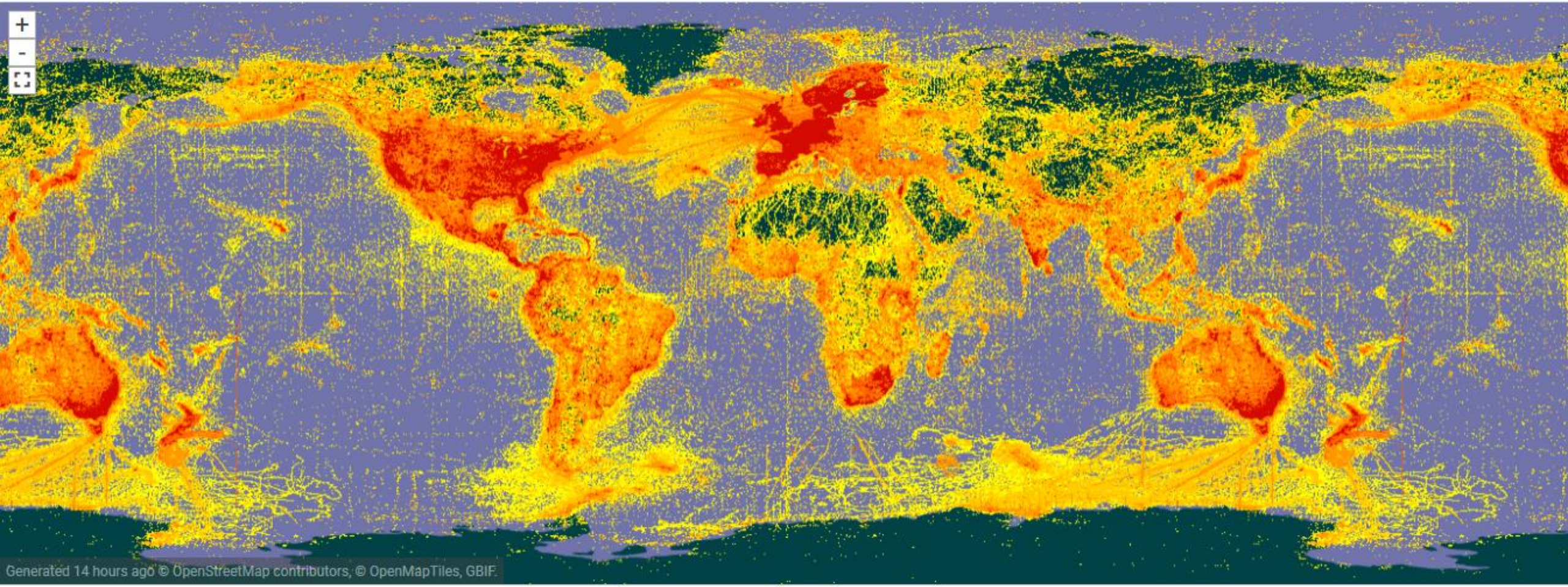
Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

```
GBREL.TXT          Genetic Sequence Data Bank
                        August 15 2019

                NCBI-GenBank Flat File Release 233.0

                    Distribution Release Notes

 213865349 loci, 366733917629 bases, from 213865349 reported sequences
```

Let’s check GBIF first!

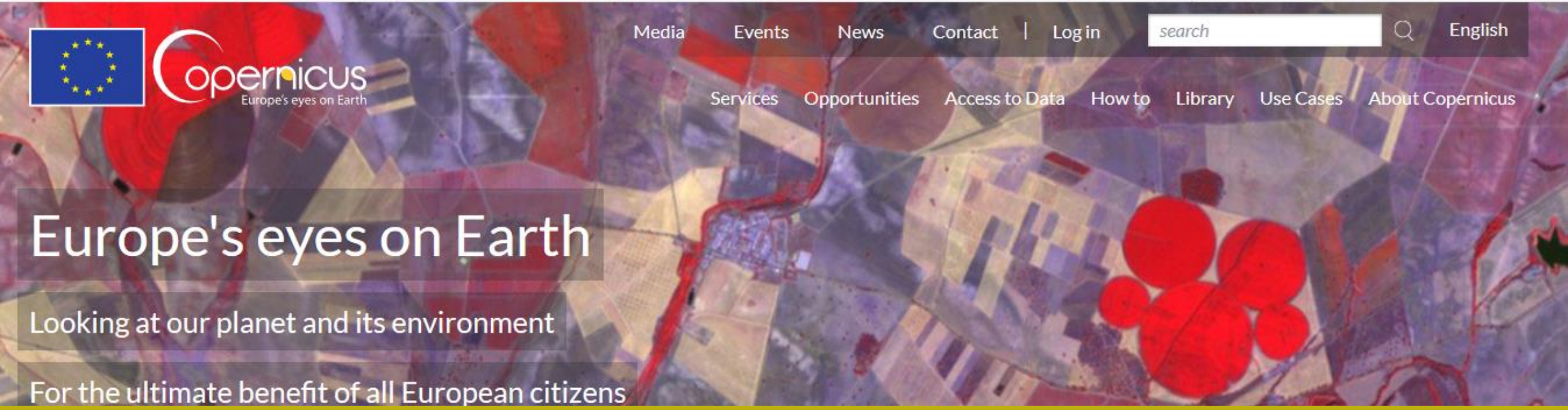
https://www.gbif.org



Task 1

- Download occurrence data for Diplodia spp. from GBIF
- Clean the dataset
- Plot the occurrence

- https://github.com/Batis007/R_course_2021
- GBIF data extraction



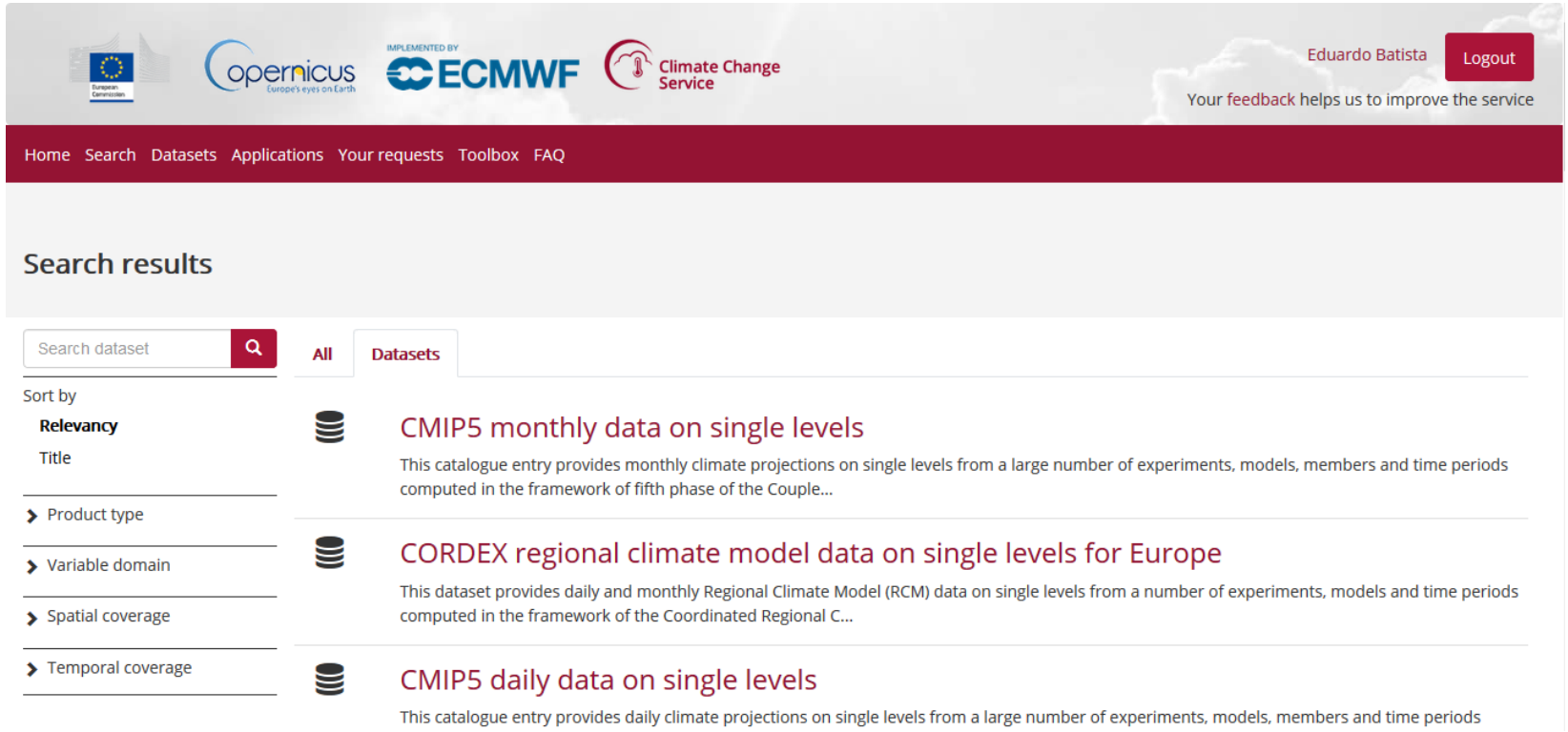
WorldClim - Global Climate Data

Free climate data for ecological modeling and GIS

Bioclimatic variables

Bioclimatic variables are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables. These are often used in species distribution modeling and related ecological modeling techniques. The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters). A quarter is a period of three months (1/4 of the year).

Environmental data extraction using the climate data store of the Copernicus project



https://github.com/Batis007/C3S_Data_use

5

NOT STARTED 0%

Introduction to the Portuguese Blended training

90 min • Maria del Pozo 23/05/2019

Introduction to the blended training, learning path and assigments for Portugal event on the 19th of June. Please note this is part of an event. The lesson includes the record of the Webinar

5

NOT STARTED 0%

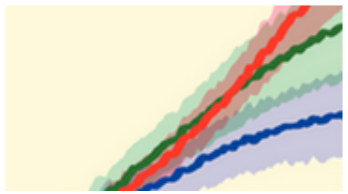
C3S ULS: Introduction to Copernicus program

12 min • Maria del Pozo 09/05/2019

Why are we here? introduction to the general overview of the Copernicus program and C3S

5

FINISHED 100%



C3S ULS: Using climate models for climate scenarios

60 min • Emma Daniels 28/03/2019

Climate data • Projections

This lessons will teach you about the methods for using climate models to develop national climate scenarios.

Environmental data using Bioclim http://worldclim.org

BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter

This scheme follows that of ANUCLIM, except that for temperature seasonality the standard deviation was used because a coefficient of variation does not make sense with temperatures between -1 and 1).

To create these values yourself, you can use the 'biovars' function in the R package dismo

WorldClim Version2

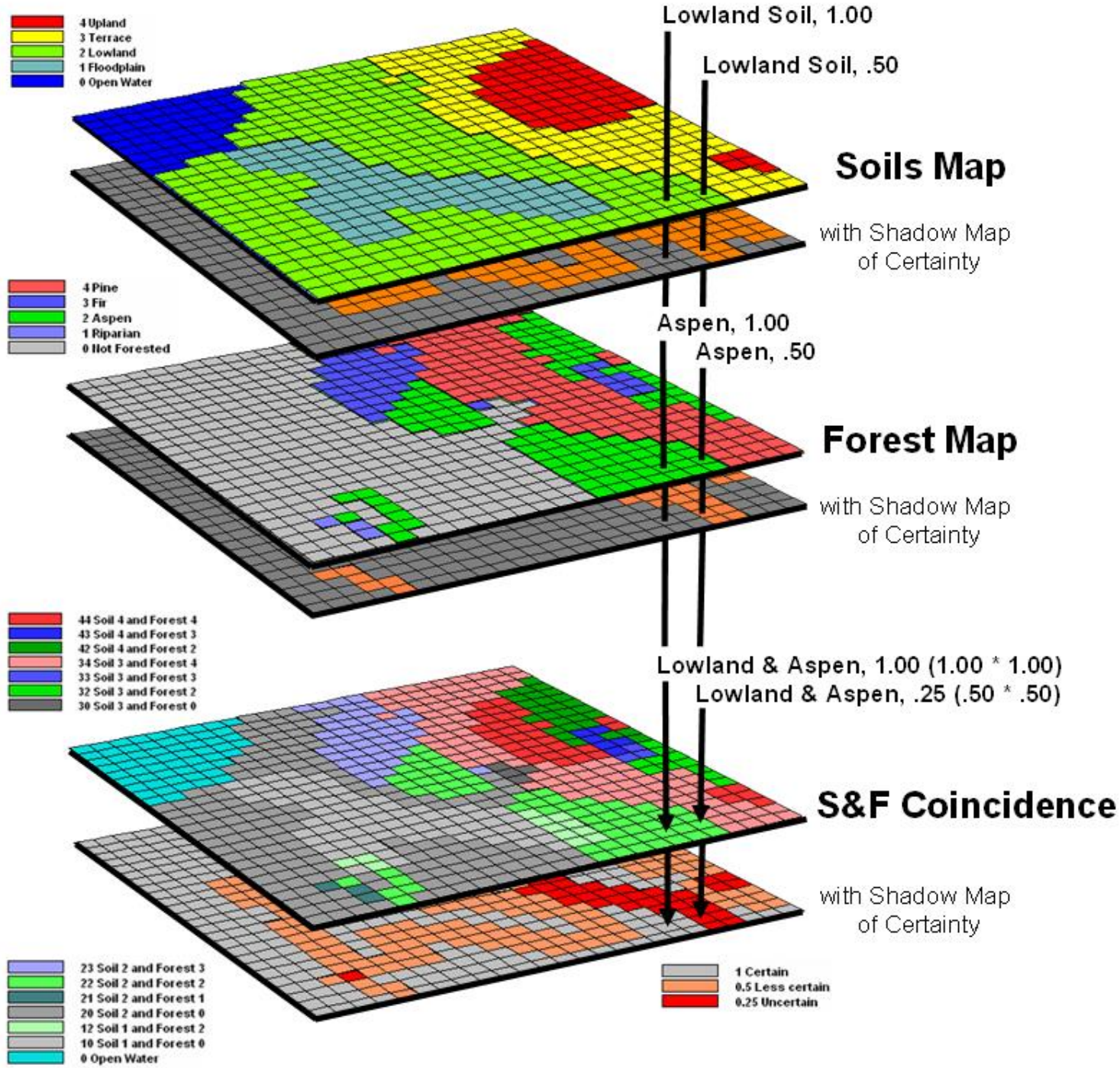
WorldClim version 2 has average monthly climate data for minimum, mean, and maximum temperature and for precipitation for 1970-2000.

You can download the variables for different spatial resolutions, from 30 seconds (~1 km^2) to 10 minutes (~340 km^2). Each download is a "zip" file containing 12 GeoTiff (.tif) files, one for each month of the year (January is 1; December is 12).

variable	10 minutes	5 minutes	2.5 minutes	30 seconds
minimum temperature (°C)	tmin 10m	tmin 5m	tmin 2.5m	tmin 30s
maximum temperature (°C)	tmax 10m	tmax 5m	tmax 2.5m	tmax 30s
average temperature (°C)	tavg 10m	tavg 5m	tavg 2.5m	tavg 30s
precipitation (mm)	prec 10m	prec 5m	prec 2.5m	prec 30s
solar radiation (kJ m^{-2} day^{-1})	srad 10m	srad 5m	srad 2.5m	srad 30s
wind speed (m s^{-1})	wind 10m	wind 5m	wind 2.5m	wind 30s
water vapor pressure (kPa)	vapr 10m	vapr 5m	vapr 2.5m	vapr 30s

Below you can download the standard (19) WorldClim Bioclimatic variables for WorldClim version 2. They are the average for the years 1970-2000. Each download is a "zip" file containing 19 GeoTiff (.tif) files, one for each month of the variables.

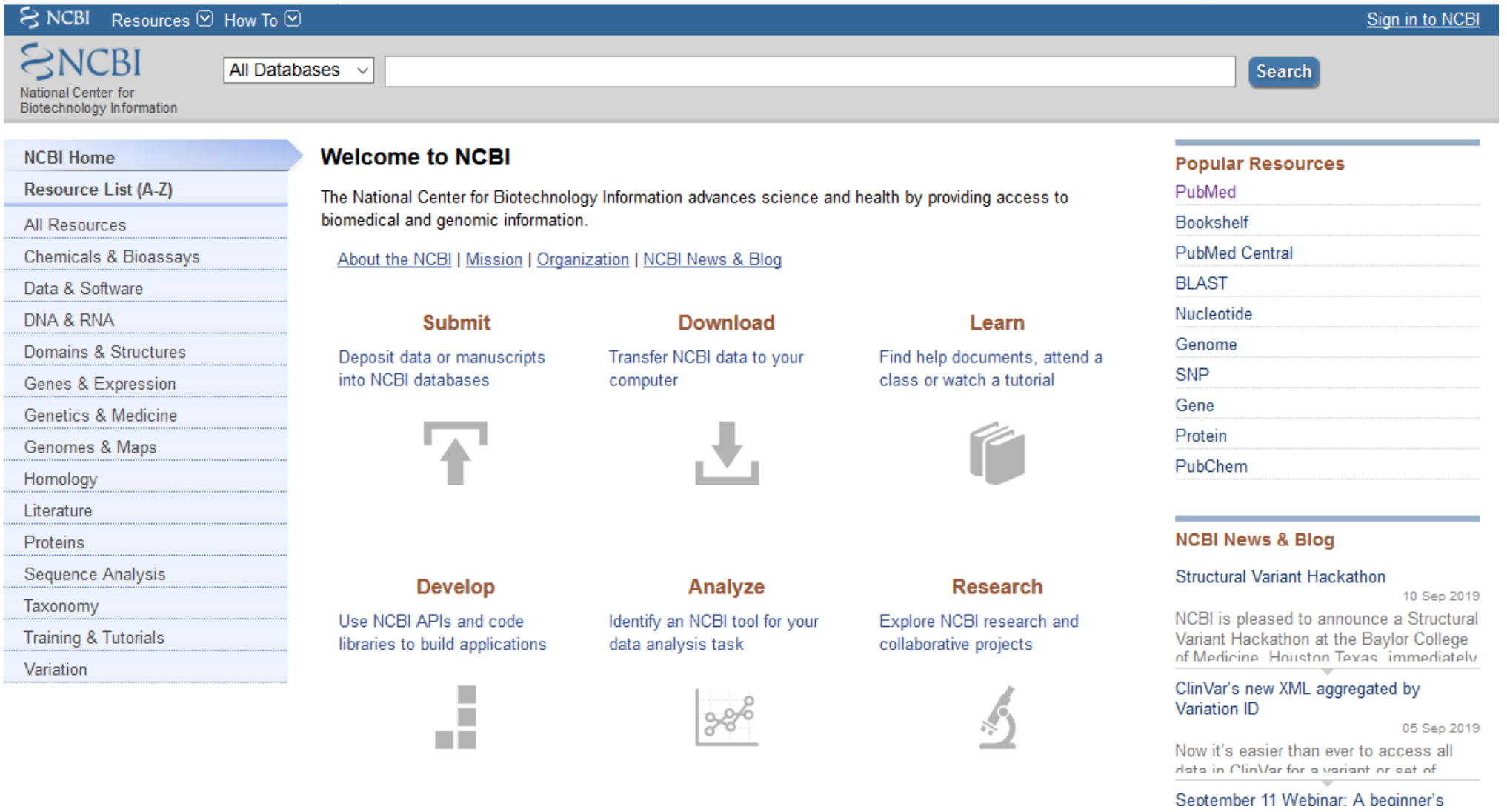
variable	10 minutes	5 minutes	2.5 minutes	30 seconds
Bioclimatic variables	bio 10m	bio 5m	bio 2.5m	bio 30s


4 Upland
3 Terrace
2 Lowland
1 Floodplain
0 Open Water
Lowland Soil, 1.00
Lowland Soil, .50
Soils Map
with Shadow Map of Certainty
4 Pine
3 Fir
2 Aspen
1 Riparian
0 Not Forested
Aspen, 1.00
Aspen, .50
Forest Map
with Shadow Map of Certainty
44 Soil 4 and Forest 4
43 Soil 4 and Forest 3
42 Soil 4 and Forest 2
34 Soil 3 and Forest 4
33 Soil 3 and Forest 3
32 Soil 3 and Forest 2
30 Soil 3 and Forest 0
Lowland & Aspen, 1.00 (1.00 * 1.00)
Lowland & Aspen, .25 (.50 * .50)
S&F Coincidence
with Shadow Map of Certainty
23 Soil 2 and Forest 3
22 Soil 2 and Forest 2
21 Soil 2 and Forest 1
20 Soil 2 and Forest 0
12 Soil 1 and Forest 2
10 Soil 1 and Forest 0
0 Open Water
1 Certain
0.5 Less certain
0.25 Uncertain

Task2

- Extract the bioclim variables for Diplodia spp.
- Analyse the Annual mean temperature for the occurrence data of Diplodia spp.

THE AMAZING WORLD OF NCBI



Task3

- Download from pubmed the number of papers about Diplodia from 2008 to 2014

The underworld of NUCLEOTIDE!!!!


```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <INSDSet>
  - <INSDSeq>
      <INSDSeq_locus>NR_111152</INSDSeq_locus>
      <INSDSeq_length>590</INSDSeq_length>
      <INSDSeq_strandedness>double</INSDSeq_strandedness>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_topology>linear</INSDSeq_topology>
      <INSDSeq_division>PLN</INSDSeq_division>
      <INSDSeq_update-date>08-NOV-2018</INSDSeq_update-date>
      <INSDSeq_create-date>25-MAR-2014</INSDSeq_create-date>
      <INSDSeq_definition>Diplodia corticola CBS 112549 ITS region; from TYPE material</INSDSeq_definition>
      <INSDSeq_primary-accession>NR_111152</INSDSeq_primary-accession>
      <INSDSeq_accession-version>NR_111152.1</INSDSeq_accession-version>
    - <INSDSeq_other-seqids>
        <INSDSeqid>ref|NR_111152.1|</INSDSeqid>
        <INSDSeqid>gi|597900487</INSDSeqid>
      </INSDSeq_other-seqids>
      <INSDSeq_project>PRJNA177353</INSDSeq_project>
    - <INSDSeq_keywords>
        <INSDKeyword>RefSeq</INSDKeyword>
      </INSDSeq_keywords>
      <INSDSeq_source>Diplodia corticola</INSDSeq_source>
      <INSDSeq_organism>Diplodia corticola</INSDSeq_organism>
      <INSDSeq_taxonomy>Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina; Dothideomycetes; Dothideomycetes incertae sedis; Botryosphaeriales;
        Botryosphaeriaceae; Diplodia</INSDSeq_taxonomy>
    - <INSDSeq_references>
      - <INSDReference>
          <INSDReference_reference>1</INSDReference_reference>
          <INSDReference_position>1..590</INSDReference_position>
        - <INSDReference_authors>
            <INSDAuthor>Schoch,C.L.</INSDAuthor>
            <INSDAuthor>Robbertse,B.</INSDAuthor>
            <INSDAuthor>Robert,V.</INSDAuthor>
            <INSDAuthor>Vu,D.</INSDAuthor>
```

Task 4

- Download ITS sequences and feature data for Diplodia isolates.

An example of a cured dataset

- Shiny app
- https://mdr-bot-cesam-ua.shinyapps.io/bot_database

- Find more
- https://www.mdpi.com/1999-4907/12/3/313