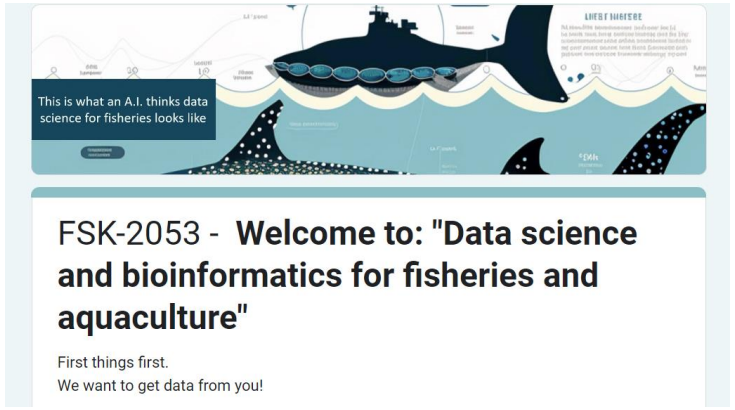
FSK2053
Data Science and Bioinformatics for Fisheries and Aquaculture

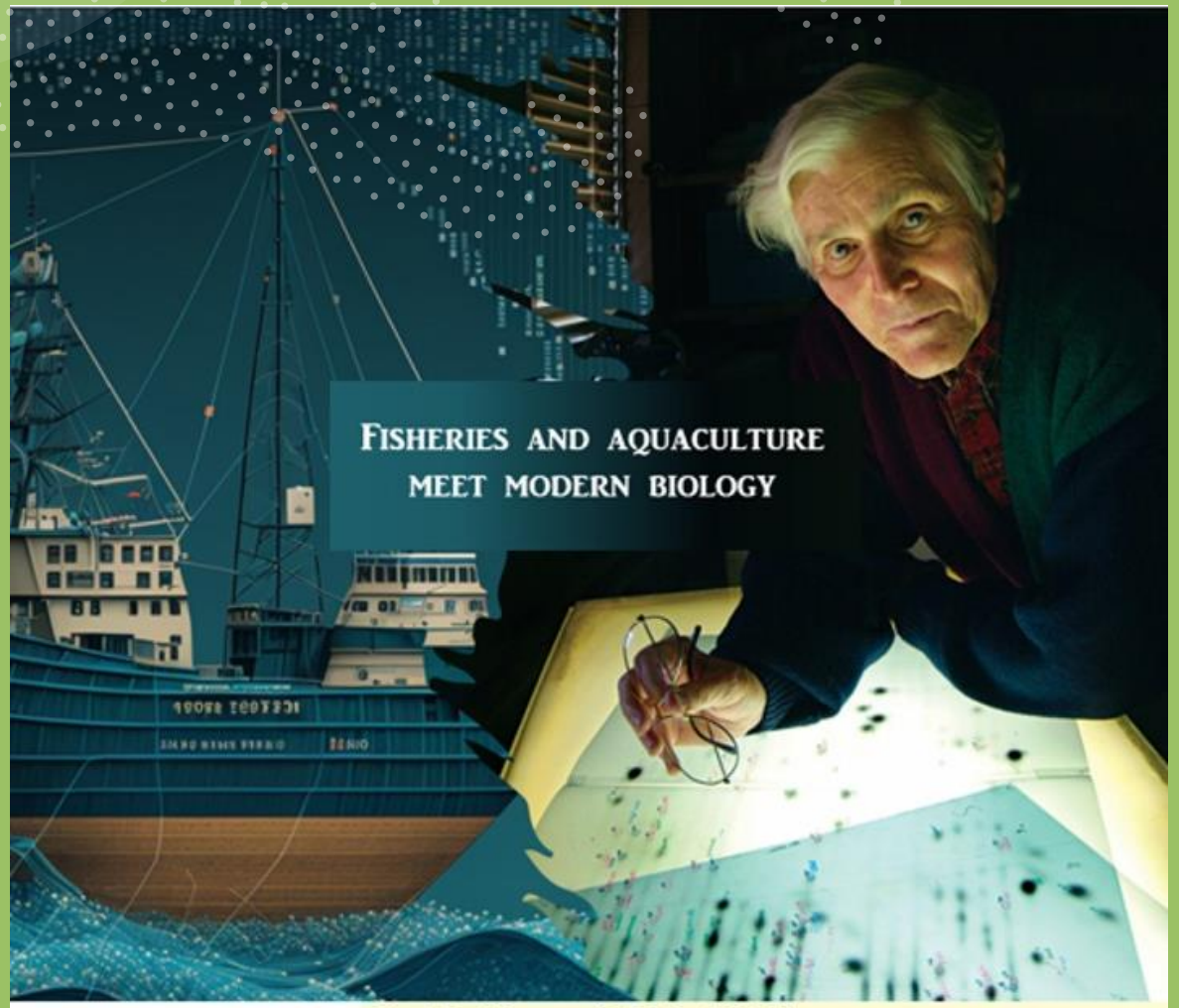
Daniel Kumazawa Morais
daniel.morais@uit.no



Scan here to give us data!

The Start of Data Science in Biology 1977

- Sanger DNA sequencing
- Redefinition of the tree of life using molecular information
- Using molecular data to explore the biological world



Carl R. Woese (1928–2012)

A visionary microbiologist focused on evolution and discovered a third domain of life, archaea.

Goldenfeld, N. and Pace, N.R., 2013. Carl R. Woese (1928–2012). *Science*, 339(6120),

Computational resources we will use for data sciences and bioinformatics

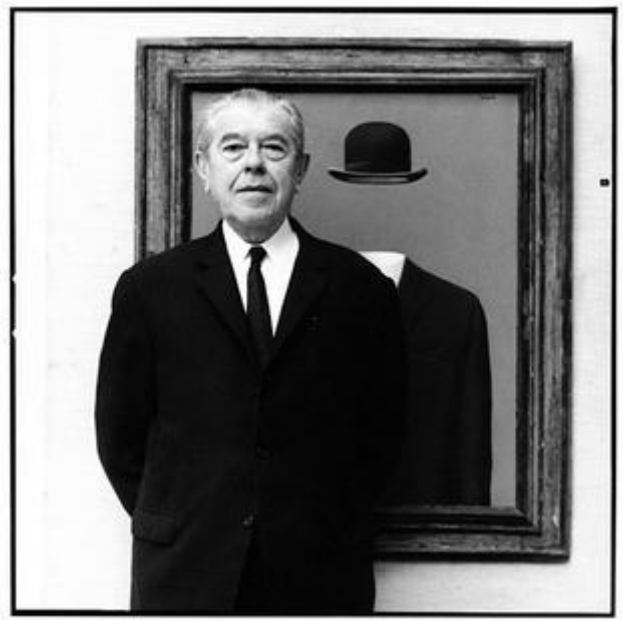
BASH
THE BOURNE-AGAIN SHELL

tidyverse

Nucleotide BLAST
nucleotide ▸ nucleotide

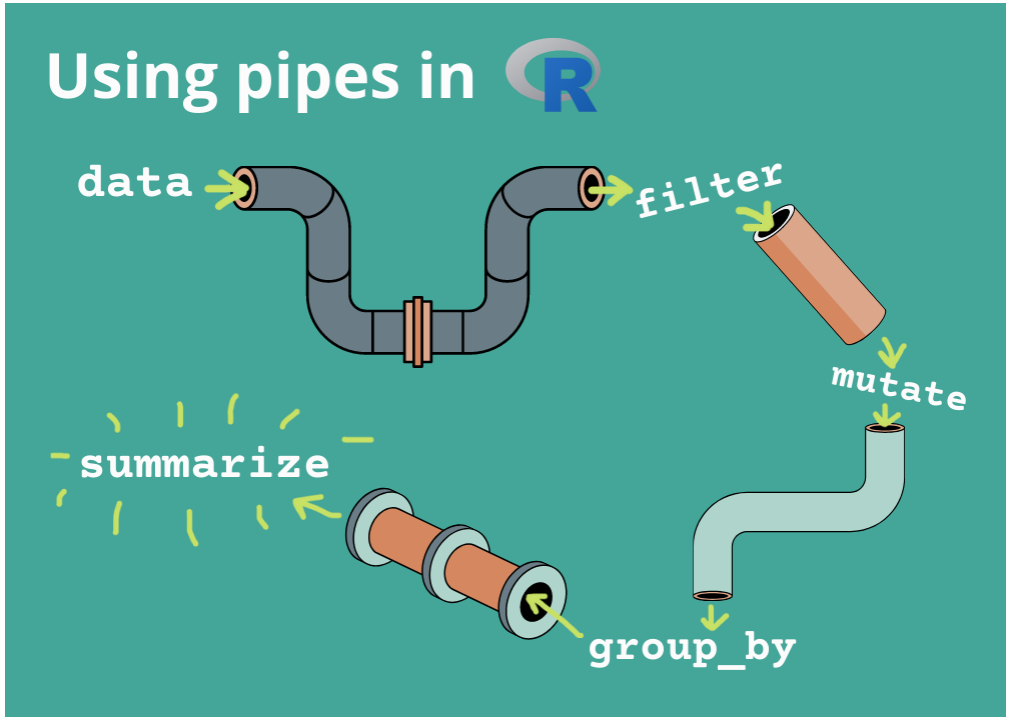
Microsoft

https://en.wikipedia.org/wiki/Ren%C3%A9_Magritte



https://en.wikipedia.org/wiki/Ren%C3%A9_Magritte

%>%




```
data_tuna <- dat %>% select(!starts_with("S_20")) %>%
  pivot_longer(all_of(as.character(2000:2010)), names_to = "Year",values_to = "Catches") %>%
  filter(!near(Catches,0)) %>%
  mutate(Stock = gsub(" tuna Global","",Stock))
data_tuna
```

Data Science

Session 1. Tuesday 11/04, 12:15-14:00: Introduction to data science. Data science workflows. Databases and public repositories. Data collection. Databases for fisheries & aquaculture. **2h theory**

Session 2. Tuesday 13/04, 12:15-14:00: Data wrangling. Statistics and data science. **2h theory**

Session 3. Monday 17/04 , 12:15-16:00: Data wrangling. Types of data. Filtering and reformatting. R-Studio / Tidyverse. **4h practice -**

***Exercise 1. Data wrangling and exploratory plots. (To be delivered until Sunday - 23.04.2023)**

Session 4. Tuesday 18/04, 12:15-14:00: Data visualization. Plotting tools for exploring big data. The Grammar of Graphics (ggplot2). **2h theory**

Session 5. Wednesday 19/04, 10:15-12:00: Statistics for big data. Descriptive vs inferential statistics. Correlation. Frequentist vs Bayesian inference. Introduction to Machine Learning approaches. **2h theory**

Session 6. Thursday 20/04, 12:15-16:00: Data visualization. Plotting tools for exploring big data. The Grammar of Graphics (ggplot2). 4**h practice**

Session 7. Monday 24/04, 10:15-14:00: Statistics for big data. Descriptive vs inferential statistics. Correlation. Frequentist vs Bayesian inference. Introduction to Machine Learning approaches. **4h practical**

Session 8. Tuesday 25/04, 10:15-15:30: Data modelling and interpretation. Predictions based on data. **2h theory + 4h practice -**

***Exercise 2. Modelling and statistical inference. (To be delivered until Monday - 01.05.2023)**

Session 9. Wednesday 26/04, 12:15-14:00: Turning data into actionable insights. Knowledge base management. **2h seminar** with **Tara Z. Baris from Ocean Data Platform.**

Bioinformatics

Session 10. Thursday 27/04, 12:15-14:00: Introduction to bioinformatics. Linux and command-line tools. Remote servers. **2h theory**

Session 11. Friday 28/04, 08:15-12:00: Introduction to bioinformatics. Linux and command-line tools. Remote servers. **4h practice**

Session 12. Tuesday 02/05, 12:15-14:00: Genetics data and databases. FASTA/FASTQ. NCBI. BLAST. **2h theory + 4h practice** -

***Exercise 3. Use of genetic databases. (To be delivered until Monday - 08.05.2023)**

Session 13. Wednesday 03/05, 12:15-16:00: Genetics data and databases. FASTA/FASTQ. NCBI. BLAST. **4h practice**

Session 14. Thursday 04/05, 10:15-12:00: Phylogenetic inference. Alignment tools. Phylogenetic tree inference. **2h theory**

Session 15. Friday 05/05, 10:15-14:00: Phylogenetic inference. Alignment tools. Phylogenetic tree inference. **4h practice** -

***Exercise 4. Inference of phylogenetic relationships. (To be delivered until Friday - 12.05.2023)**

Session 16. Monday 08/05, 12:15-14:00: Genetics of population differentiation. VCF format. Data Structure. Discriminant Analysis of Principal Components. **2h theory**

Session 16. Monday 09/05, 12:15-16:00: Genetics of population differentiation. VCF format. Data Structure. Discriminant Analysis of Principal Components. **4h practice**

Session 17. Thursday 11/05, 10:15-12:00: The role of genetics data in fisheries, aquaculture and conservation. Defining management units. Decision-making in environmental management. **2h seminar**

Data Science

*Exercise 1. Data wrangling and exploratory plots. (To be delivered until Sunday - 23.04.2023)

*Exercise 2. Modelling and statistical inference. (To be delivered until Monday - 01.05.2023)

*Exercise 3. Use of genetic databases. (To be delivered until Monday - 08.05.2023)

*Exercise 4. Inference of phylogenetic relationships. (To be delivered until Friday - 12.05.2023)

50% of your score will come from these exercises.

Assessment

The assessment consists of 2 parts, each counting 50 % of the final grade.

- **Four exercises** in computer format, either development of short scripts or results of data analyses (counts 50 %)
  - The deadline for wiseflow submission will be on the 13.05.2023 at 14:00.
- Written **home exam** consisting of solving practical problems (counts 50 %)
  - The deadline for wiseflow submission should be on the 26.05.2023 at 14:00.

The students will get feedback on their 4 exercises and written home exam.

The grading scale is A - F, where F is fail.

Both parts must be passed in order to pass the course.

There will not be a **re-sit examination** for students that did not pass the previous ordinary examination.