

# THE DATA'S JOURNEY

Hypothesis



Validation



Cleaning & Normalization



Data Exploration



Data Collection

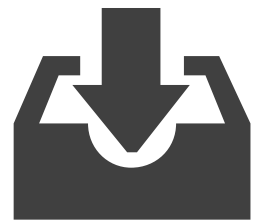


Analysis



# BEGINNING OUR JOURNEY

---



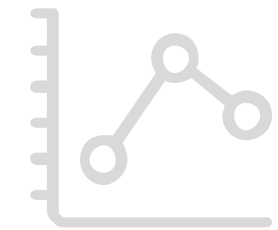
**DATA COLLECTION**



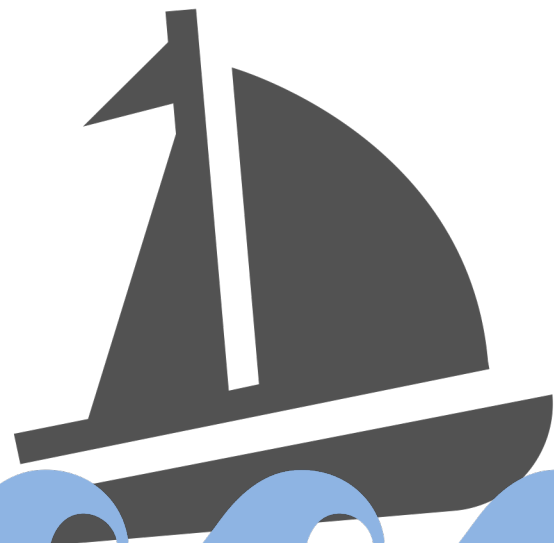
**EXPLORATORY  
DATA ANALYSIS**



**DATA ANALYSIS**

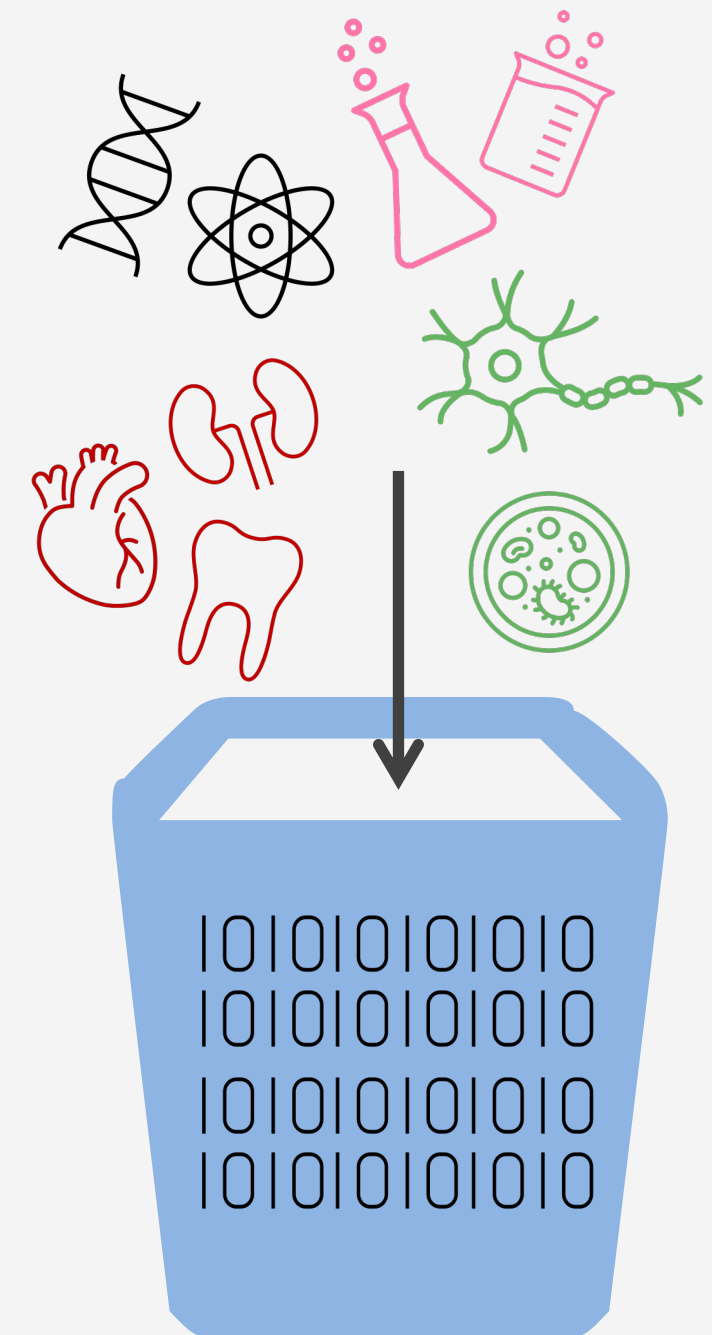


**MODEL EVALUATION**



# DATA COLLECTION & TERMINOLOGY

- Data points are **observations of reality**, made with the help of measuring devices or techniques which means:
  - **Measurement uncertainty**
  - **Introduced biases**
- Type and impact of data uncertainty & biases may be known or unknown
- **Datasets have:**
  - **Observations**, the properties collected, i.e. a cell, a tissue sample
  - **Variables (Features)**, i.e. blood pressure, smoker status, gene expression
    - **Outcome** variable(s) (**dependent**)
    - **Explanatory** variables (**independent**)?



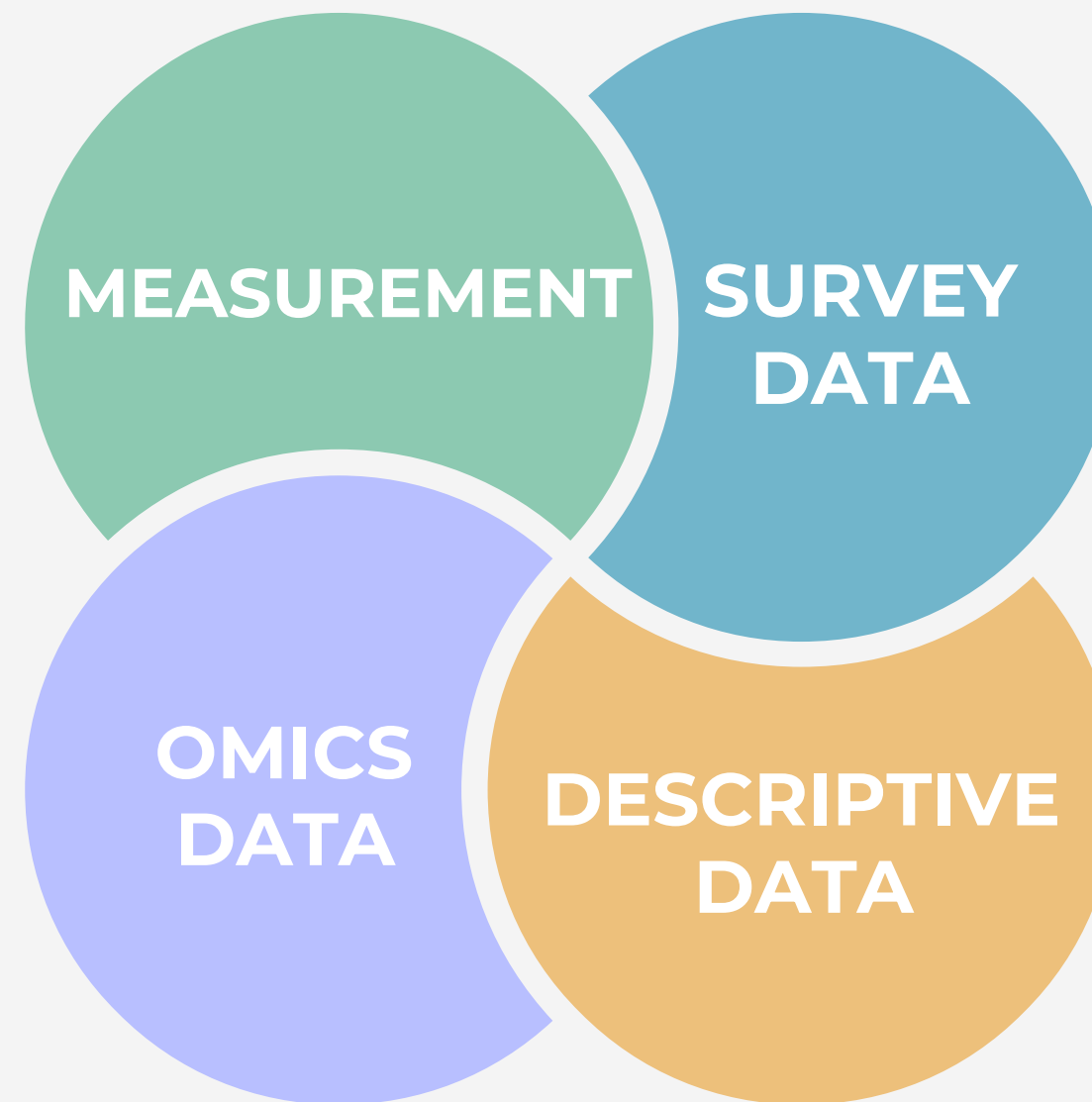
# DATA TYPES

## Measurements

- i.e. height, weight, heart rate
- Can usually be represented as a number or a category
- Typically low dimensional, but *wearables* are changing that

## Omics Data

- i.e. genomics, proteomics
- (Ultra) high-dimensional
- Needs a lot of preprocessing



## Survey Data

- Can be descriptive, numeric or categorical
- Relies on reporting, i.e. very prone to bias

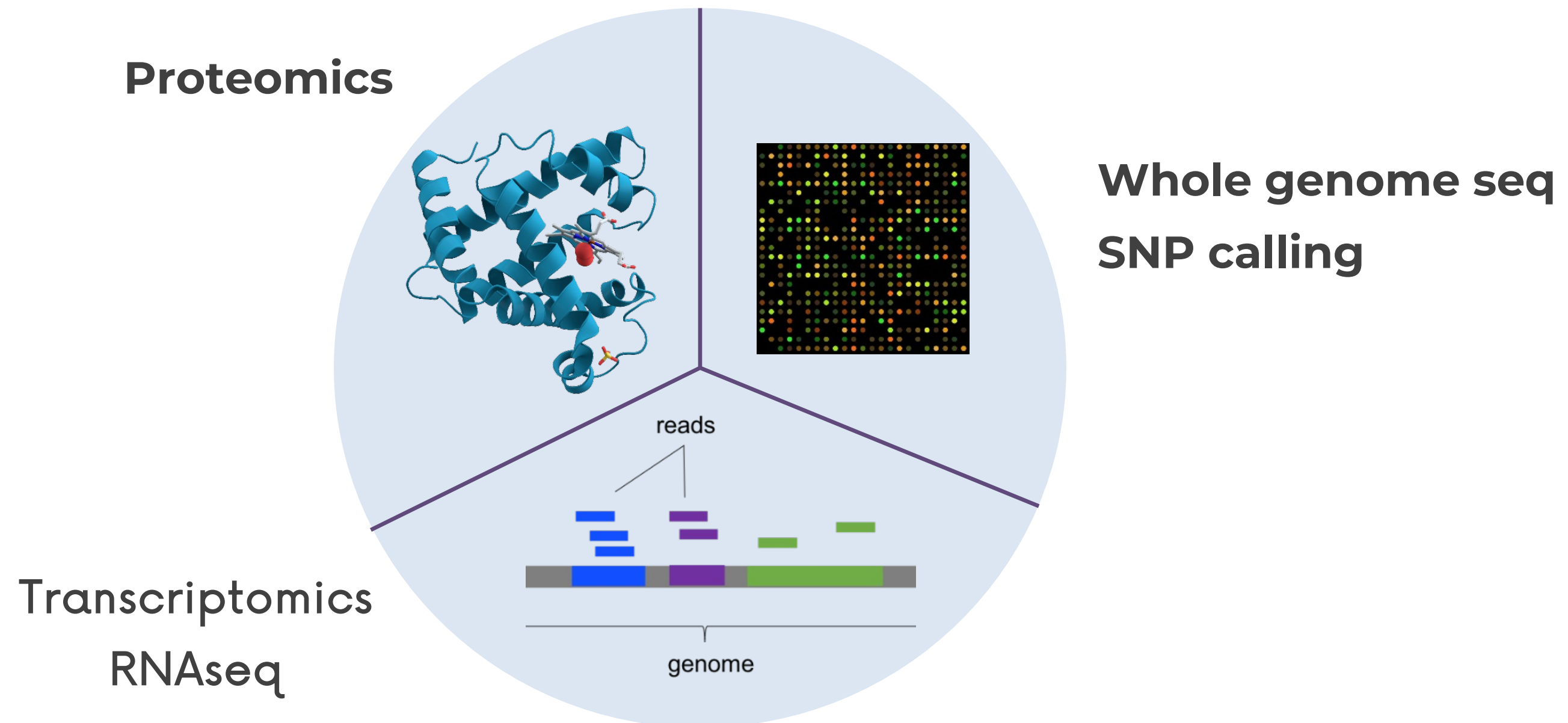
## Descriptive Data

- i.e. patient journals, registry data
- Highly person-sensitive
- Not easily standardized

# OMICS DATA

---

Omics approaches aim to study the entirety of an 'ome' (proteome, transcriptome, genome). Here we name the most commonly used types.



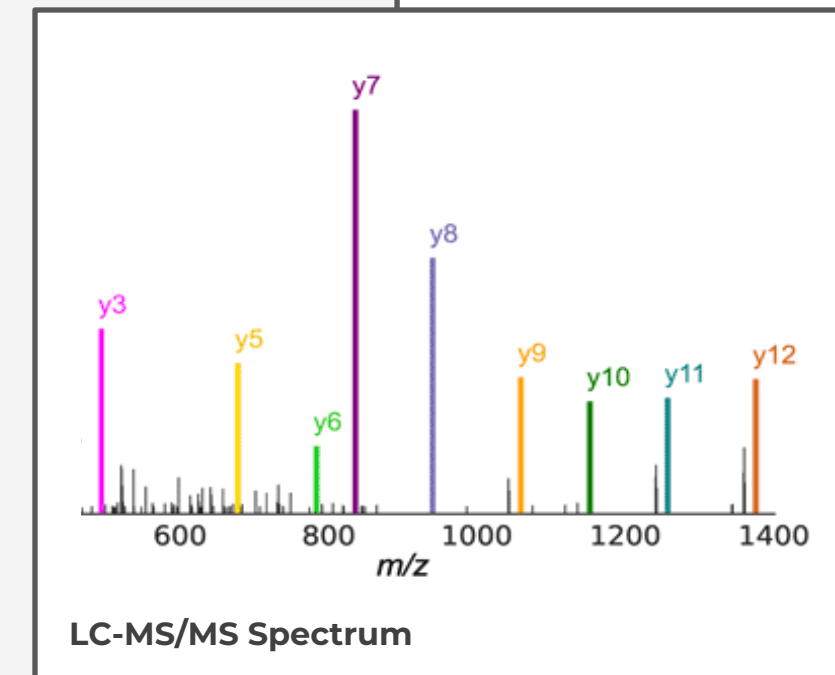
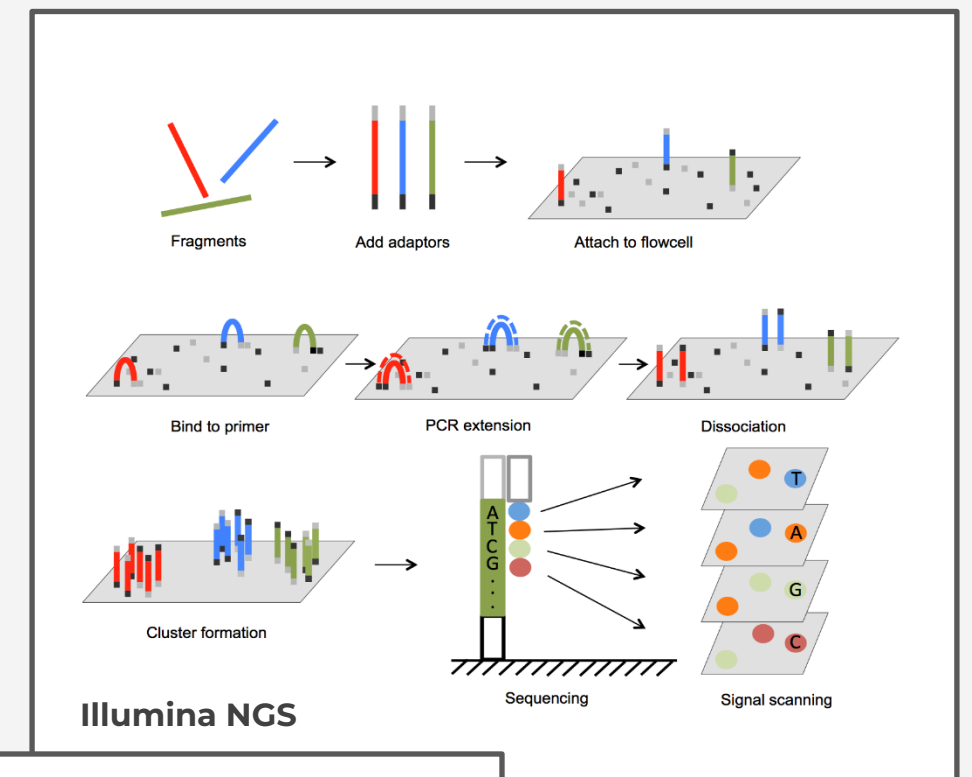
# OMICS DATA

Omics data are produced with specialized lab protocols followed by:

- Next generation sequencing (DNA, RNA)
- Mass spectrometry technologies (proteins, lipids, sugars, metabolites)

Data characteristics:

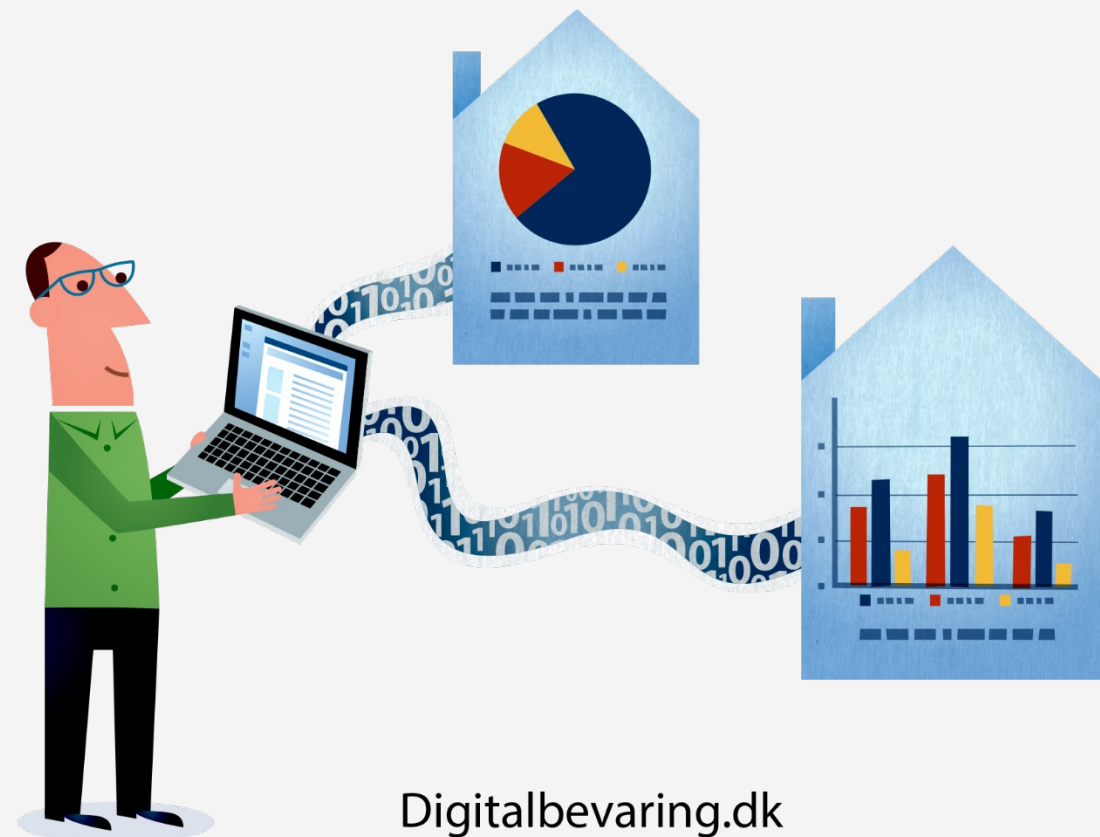
- Extremely high-dimensional
- Expensive to generate
- Measurement is indirect
- Need for pre-processing
- Can be prone to high variability (therefore replicates)



# BIOBANKS & REPOSITORIES

---

- Biobanks and repositories are great sources of both bio data and patient metadata
- We have many such resources in Denmark but getting access can be a cumbersome process, since this is (highly!) person-sensitive data



- Access has to be applied through the proper channels and compliance has to be ensured while working with them.

You can hear more about this in our **GDPR course** for biomedical researchers!



# REGISTRY DATA

---

Most doctors, epidemiologists, ..., and statisticians work with registry data (birth, death, diseases, medications, biometrics)

Data characteristics:

- Data are sensitive & hard to access
- There are many inconsistencies and errors
- Missing data are very common
- Needs a lot of clean-up and restructuring

The registries are governed by **Sundhedsdatastyrelsen**  
(The Danish Health Data Authority).



Dan Med J 2023;70(4):A12220796



# EXPERIMENTAL DESIGN

---

You have heard it before and you will hear it again, **experimental design is important.**

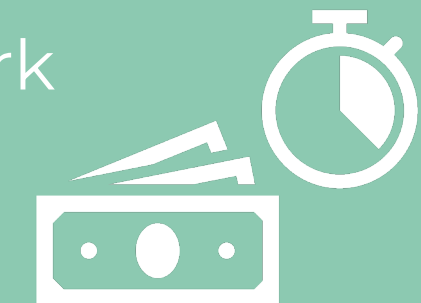
## WHAT IS GOOD EXPERIMENTAL DESIGN?

- True normal controls
- Power calculations
- Randomization (data collection & laboratory)
- Bias prediction
- Documentation



## WHY IS IT SO IMPORTANT?

- You will be able to correctly answer your scientific question
- Saves you time/work
- Saves you money



# CONFOUNDING

---

We often see non-randomized design, and sometimes total **confounding**.

**Confounding:** We cannot distinguish the separate effects of two different sources of variation.

There is **no way to correct confounding** when it is complete (this means the two variables 100% co-occur).



# GROUP DISCUSSION 1.1

Consider the following **scenario (1)**:

You would like to study whether smokers have an increased risk of heart disease. For this, you have collected some data from both men and women on whether they smoke and have heart disease. When you examine the data you see that all of your smokers are men.

**Can you answer your research question with this dataset?**



## GROUP DISCUSSION 1.2

Consider the following **scenario (2)**:

You would like to study differences in gene expression between tumor and healthy tissue. Since you have a lot of samples you ask two lab techs to each process half.

**What are potential biases and confounders in this set-up?**



# BEFORE THE ANALYSIS COMES...

---

- **File formats in data science usually need to be a table** (csv, xlsx, tab delimited, ...)
- **Omics platforms return:**
  - Fasta, fastq, SAM, GFF, MzML, ... huge files do not fit in DS analysis
- **Registries are:**
  - In formats more easily legible by humans, but still poorly structured for DS
- At the point where we do what is typically regarded as DS, the raw data has been translated into the quantities we want to measure and/or restructure.
- The steps we take before Data Science may be called: ***data management, set-up or wrangling***

# DATA MANAGEMENT

---

## DOS

Write down steps (protocol & comments)

Record date of downloads

Consistency of file naming and folder structure

The raw data is untouched!

Consistency of data management

Don't store BIG DATA in your computer

Avoid XLSX spread sheets

No color coding

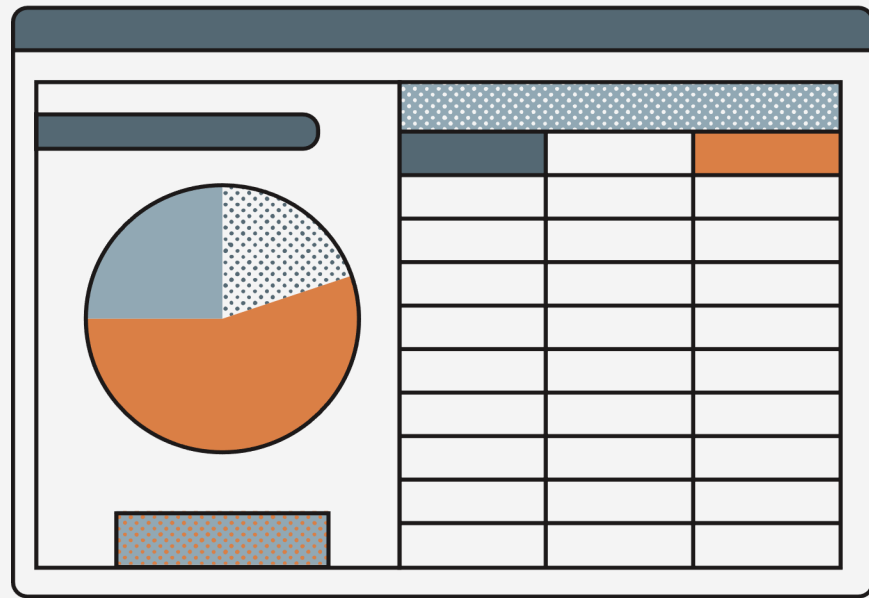
No spaces nor special characters in naming

Do not do things 'manually' if avoidable

## DON'Ts

# WHAT HAS CHANGED?

---



- Absolutely massive amount of data! Both in terms of number of observations as well as number of measured variables
- New types of data (i.e. omics, geolocation, wearables, ect)



change of tools

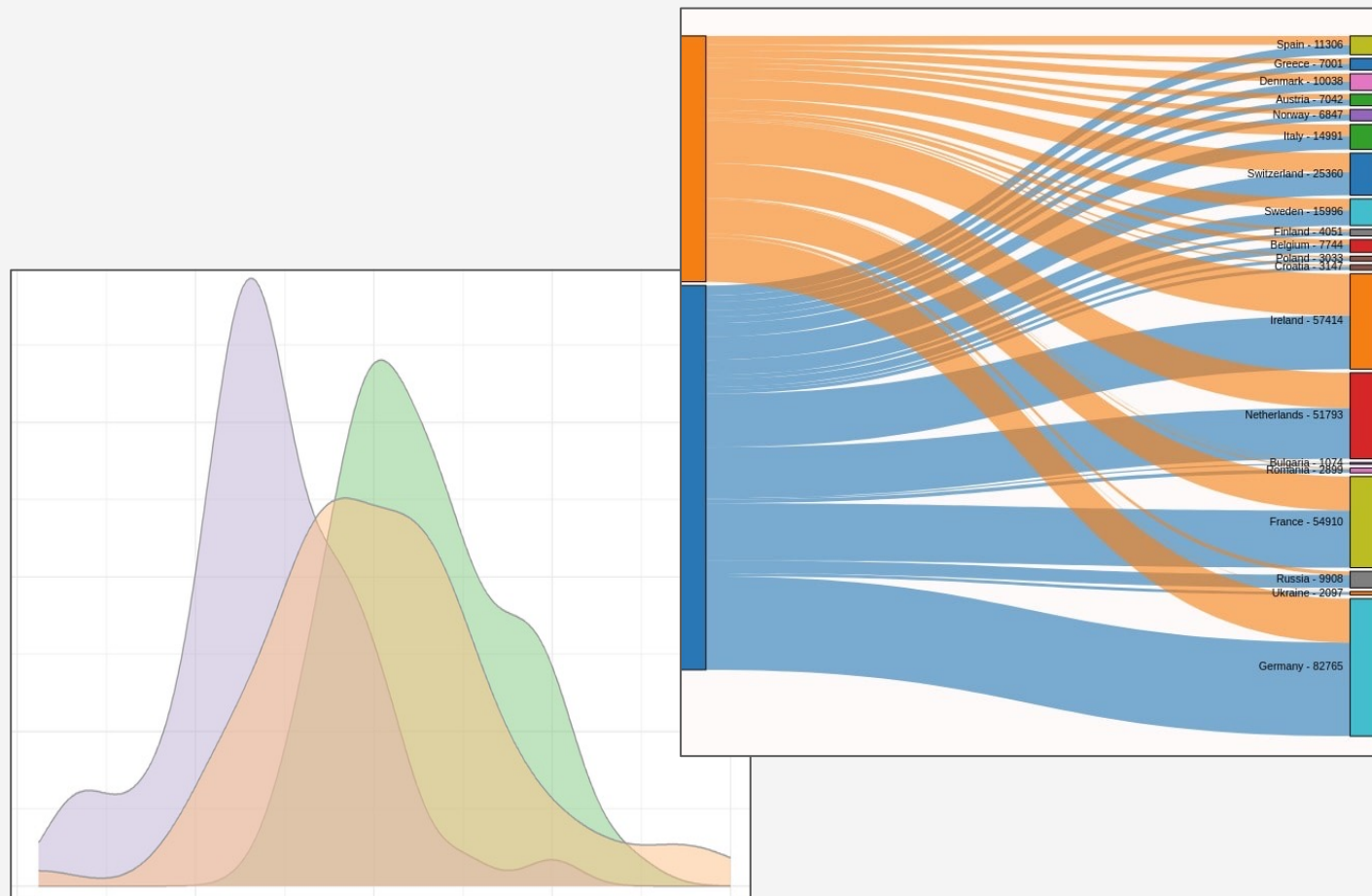


new analysis techniques



# PROGRAMMING LANGUAGES

- Part of the new tools is the use of **formal programming languages** to analyze data instead of tools like Excel.



Quick overview

Processing power

Easy starting

Data Wrangling

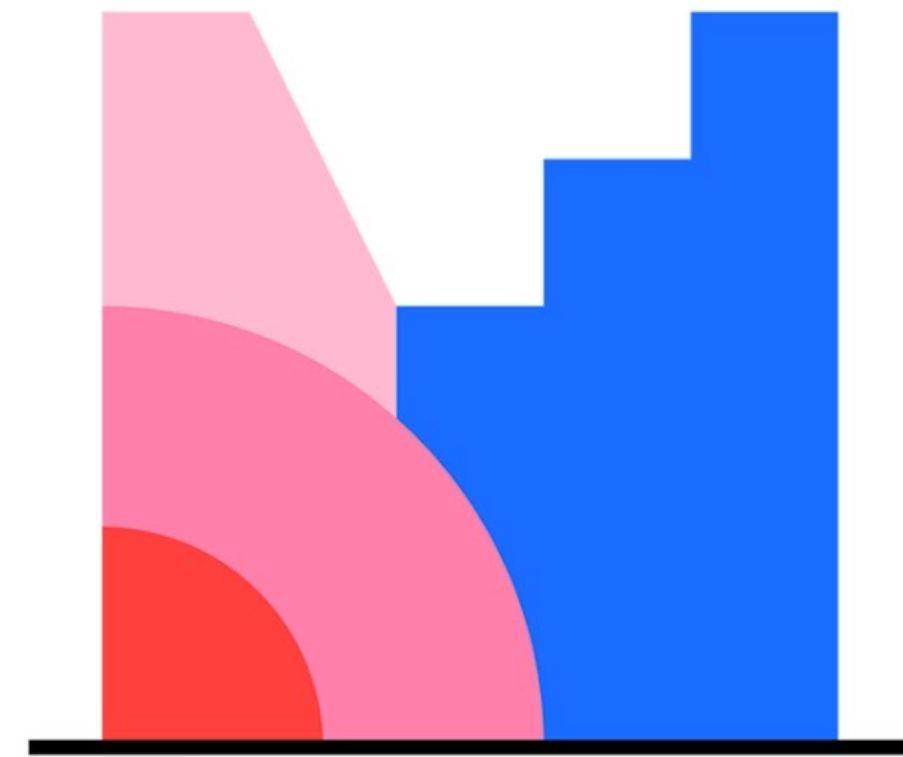
Reproducibility

Raw data exposed

Complex analyses

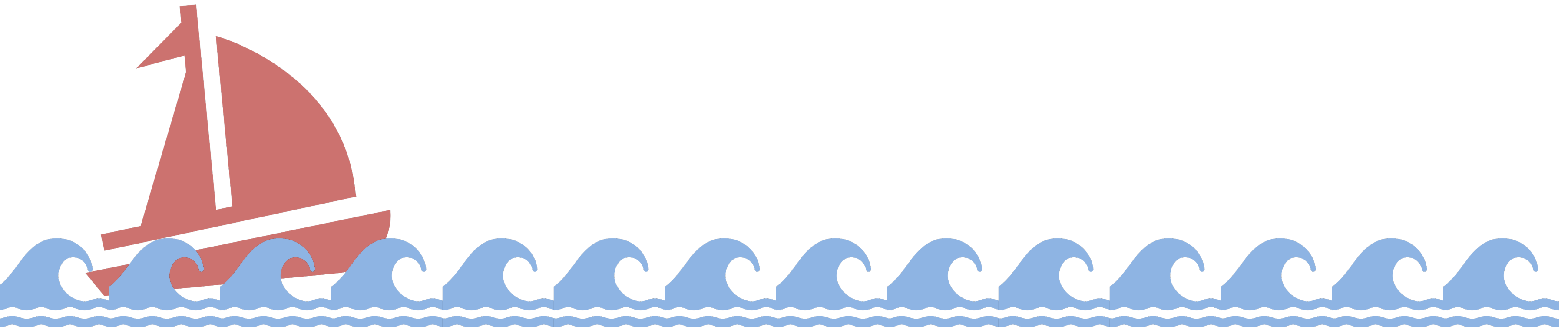
Manual analysis

Programming language



**Mentimeter**

**Which programming languages  
have you had contact with?**



# THE DATA-HYPOTHESIS CYCLE

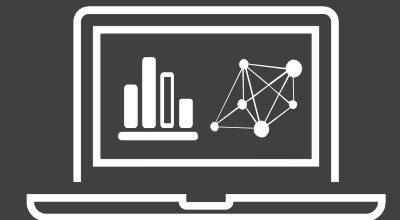
## Hypothesis-driven:

- 'traditional' way of research
- formulate hypothesis
- design experiment to challenge
- hypothesis supported or rejected



## Data-driven:

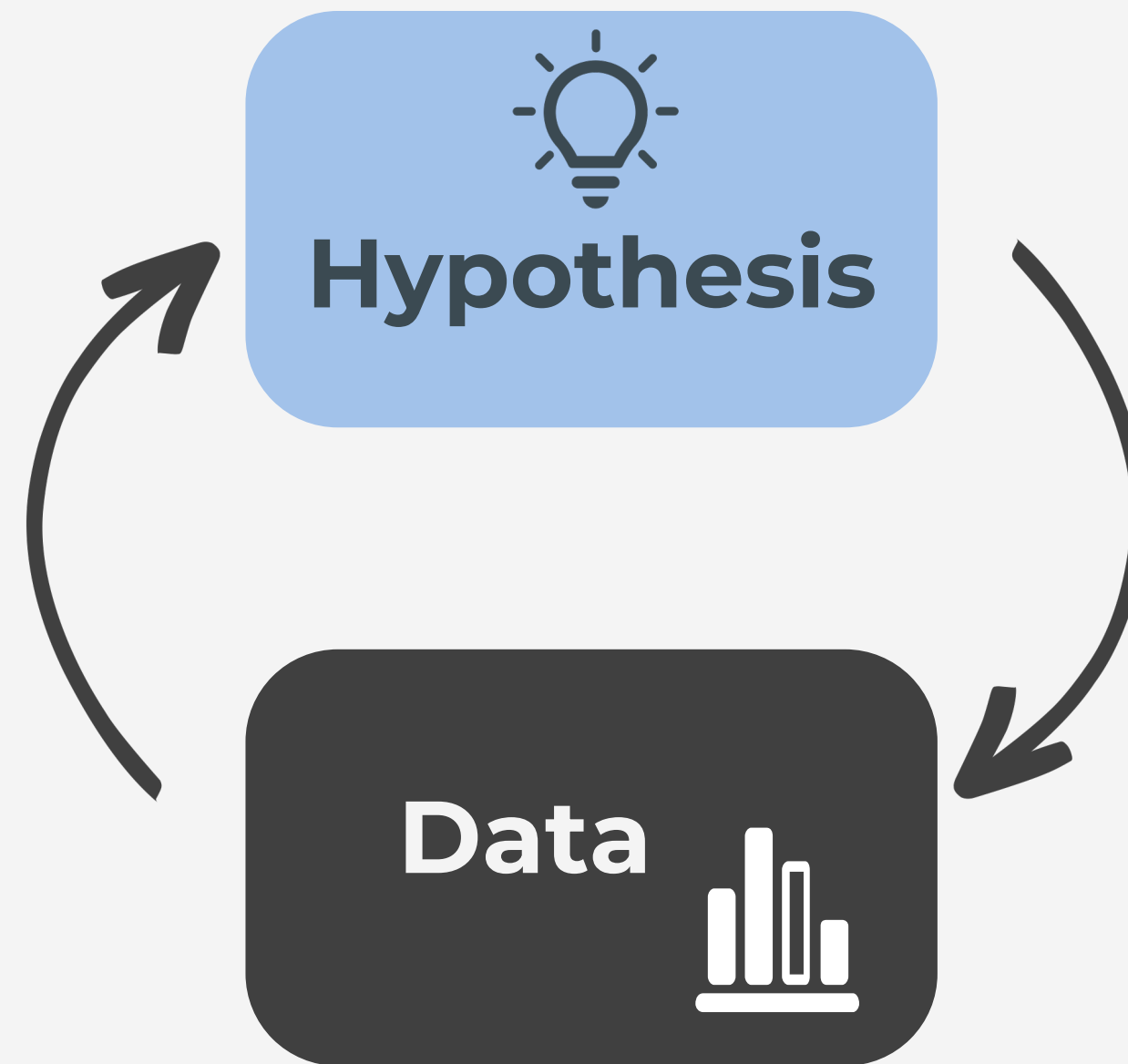
- Discover properties of data set
- Identify patterns and relationships
- Mostly possible with BIG datasets
- data-driven  $\neq$  fishing!



# THE DATA-HYPOTHESIS CYCLE

High-dimensional data with many observations

- Discover complex patterns that humans cannot identify without algorithms & computer power
- Find new patterns from re-analysis with new (multiple) methods or combination of datasets



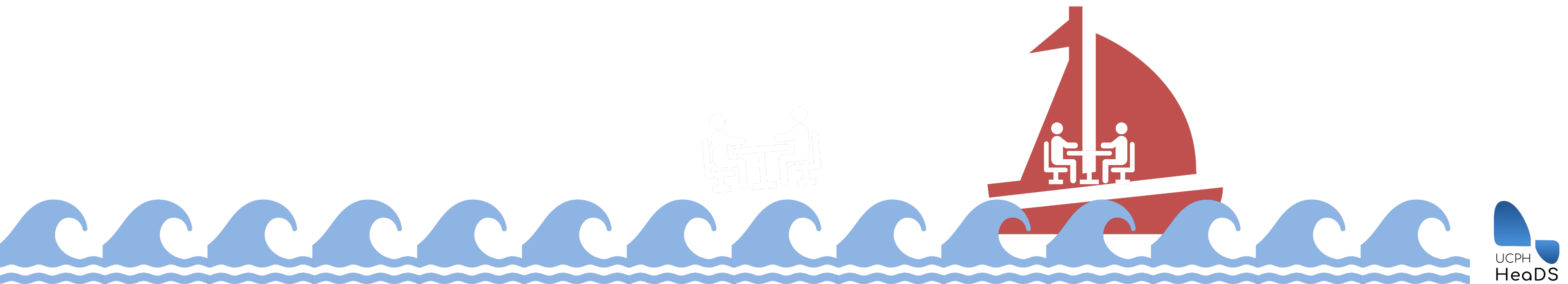
Gather new data to confirm or deny new hypothesis

We do both alternately

# GROUP DISCUSSION 1.3

In your groups discuss:

- What **data types** do you and your collaborators currently work with and/or what are you interested in working with in the future?
- What considerations are there in terms of **experimental design, data collection** and/or **data management/set-up**?



# BREAK

