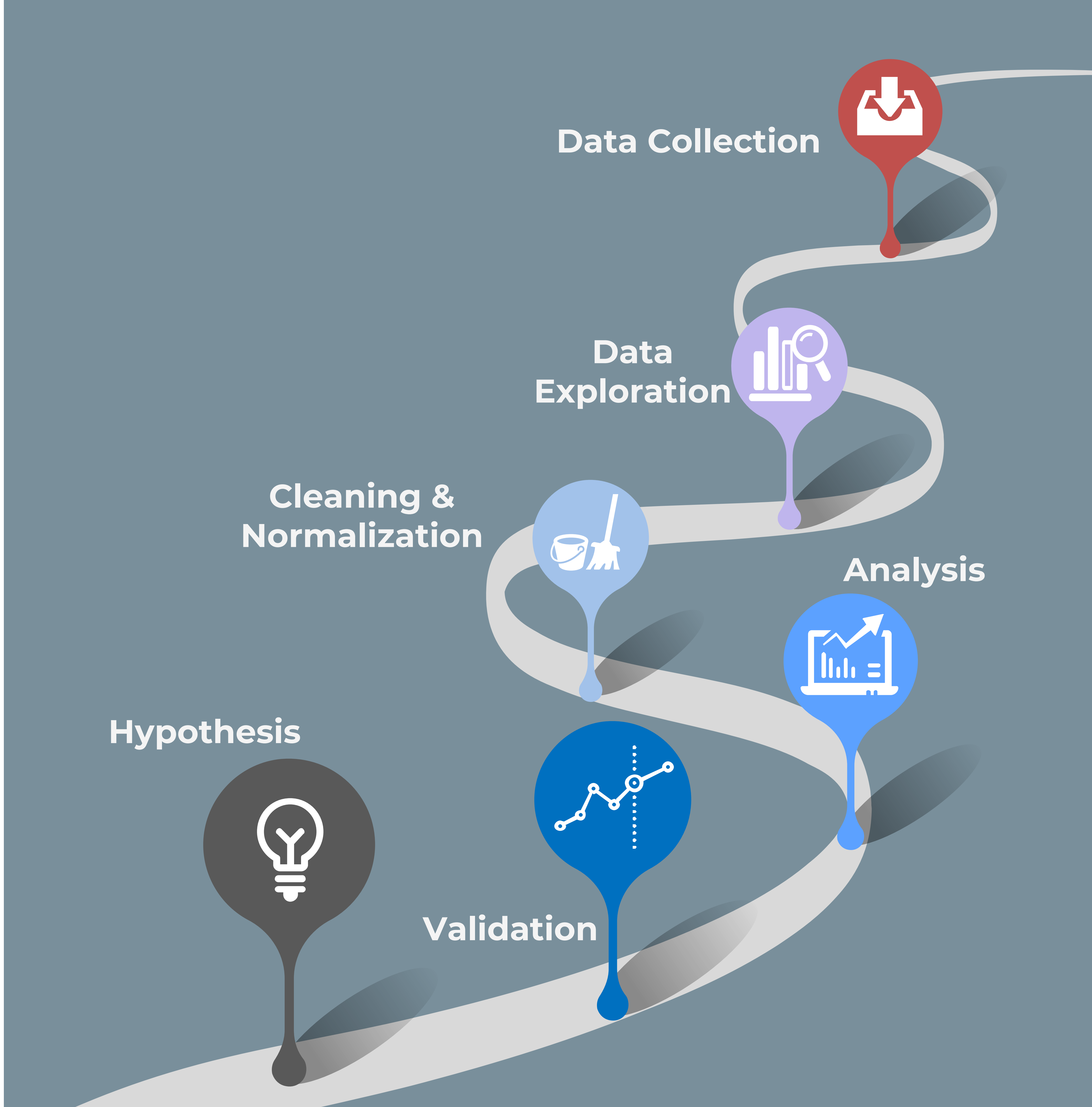
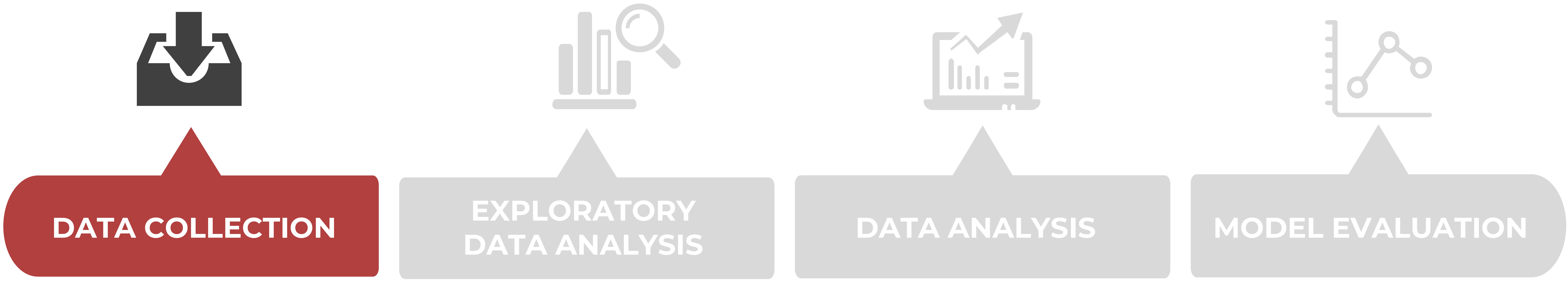
THE DATA'S JOURNEY



UCPH
HeaDS

BEGINNING OUR JOURNEY

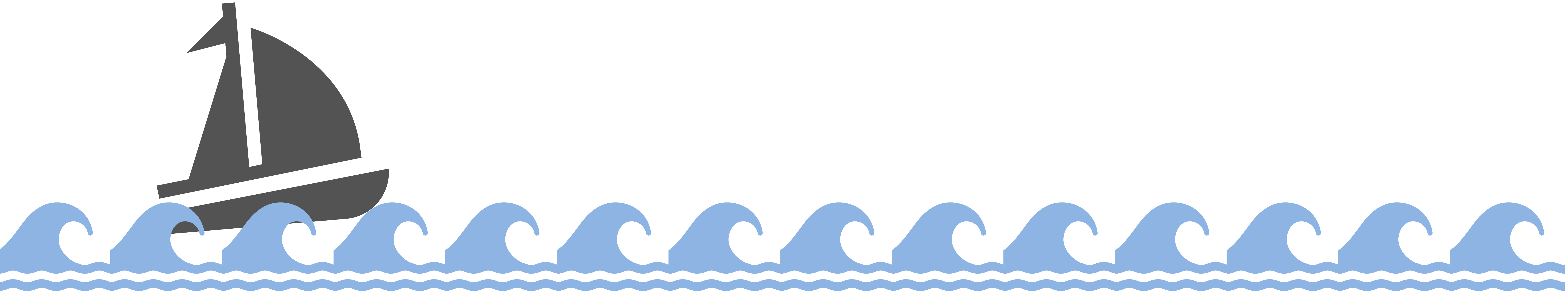


DATA COLLECTION

EXPLORATORY DATA ANALYSIS

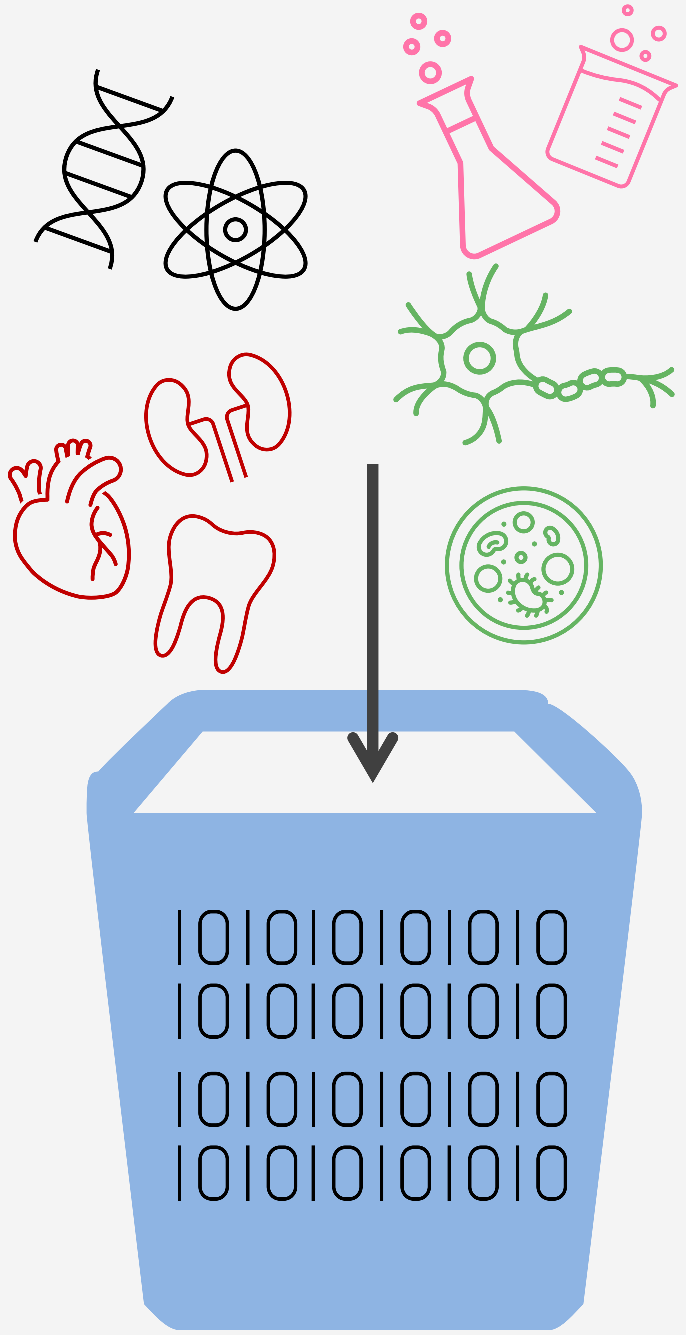
DATA ANALYSIS

MODEL EVALUATION



DATA COLLECTION

- Data points are **observations of reality**, often made with the help of measuring devices or techniques.
- All data has an inherent **measurement uncertainty**, the size of this uncertainty may be unknown.
- Data may have **introduced biases.** If we know what they are we can avoid them or correct for them.

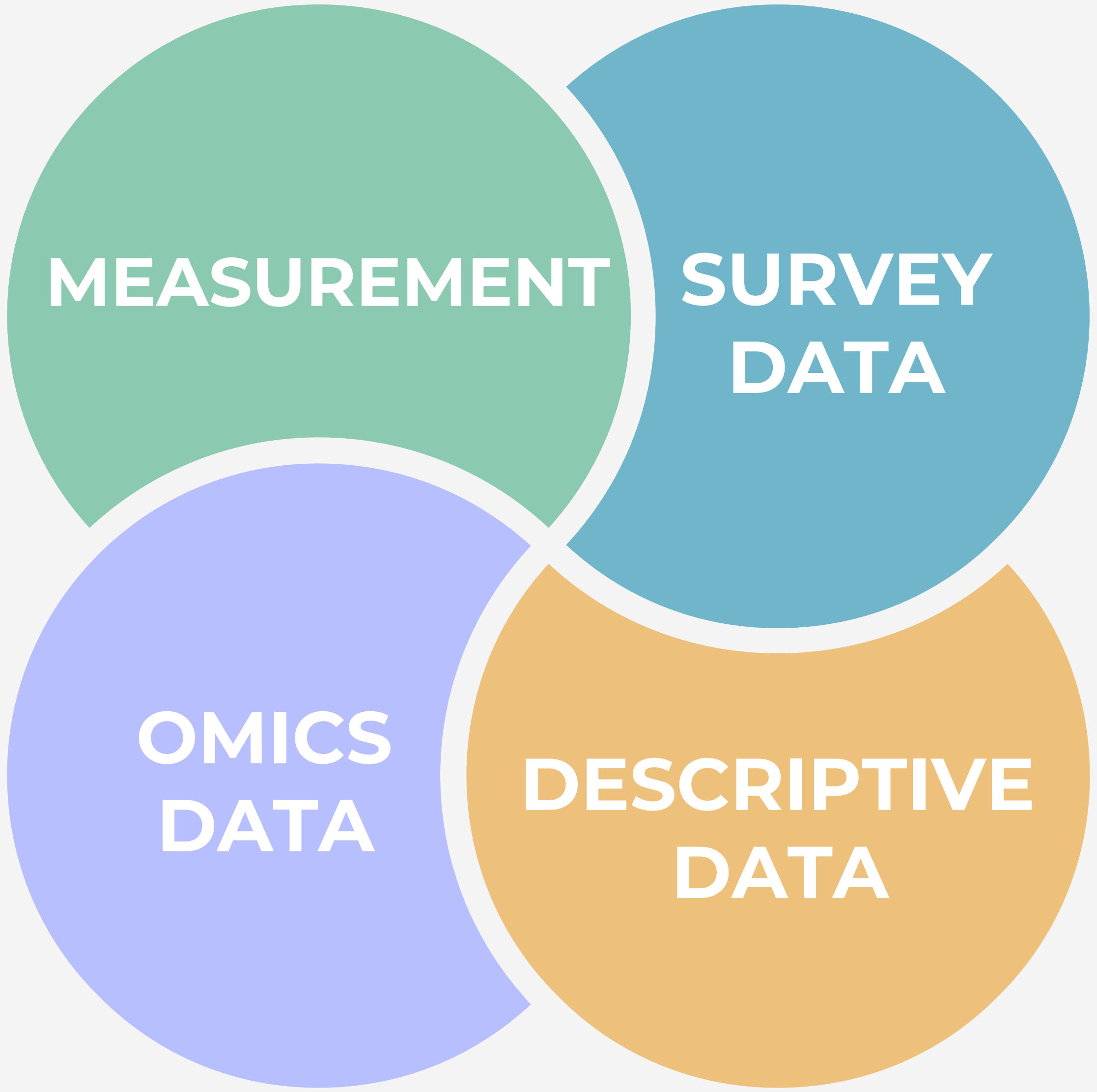


UCPH HeaDS

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



Survey Data

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

Omics Data

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

Descriptive Data

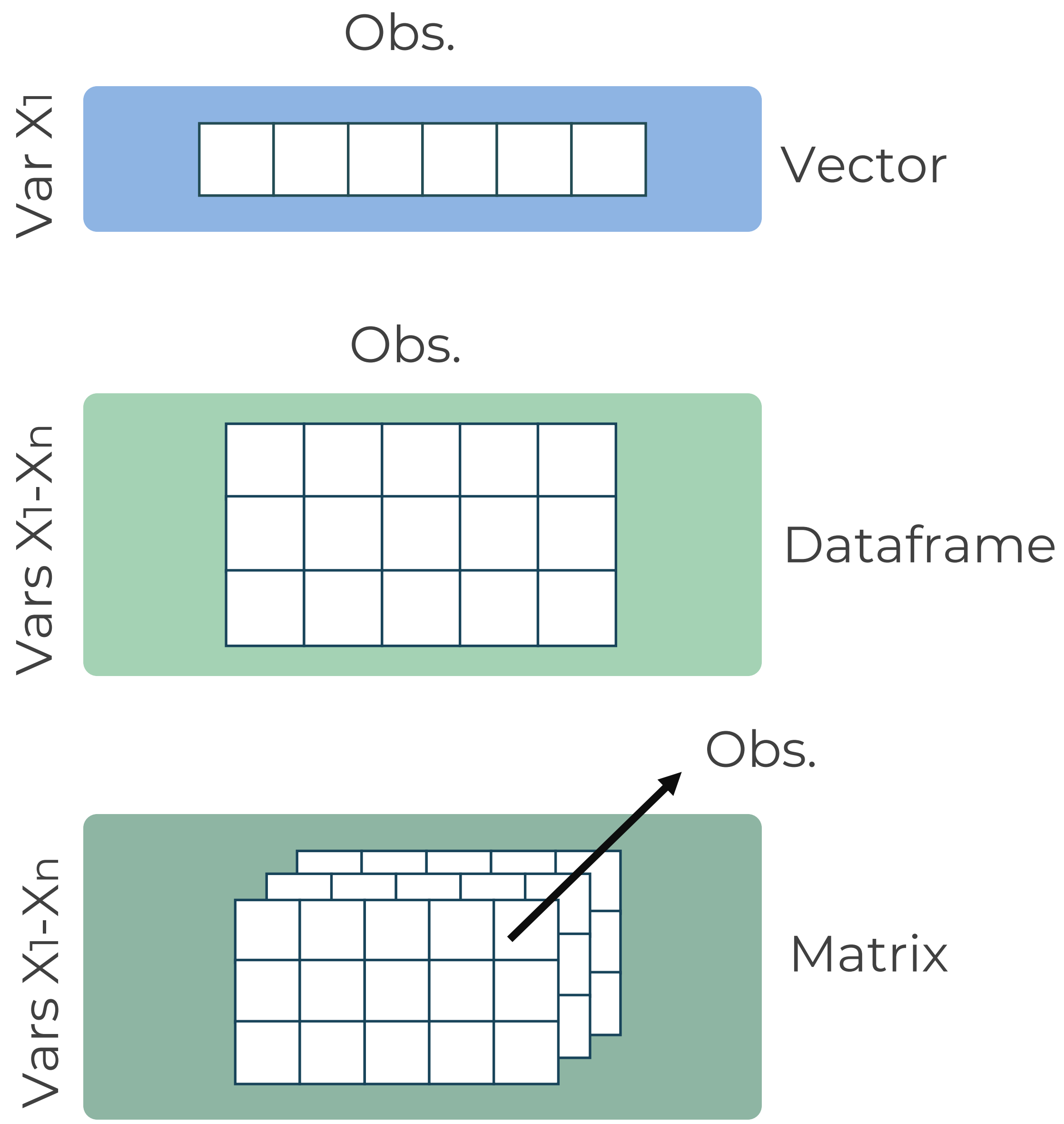
- i.e. patient journals, registry data
- Highly person-sensitive
- Not easily standardized

A LITTLE TERMINOLOGY

- How many observations?
  - **Observation** is the entity whose properties were collected, i.e. a patient, a cell, a tissue sample

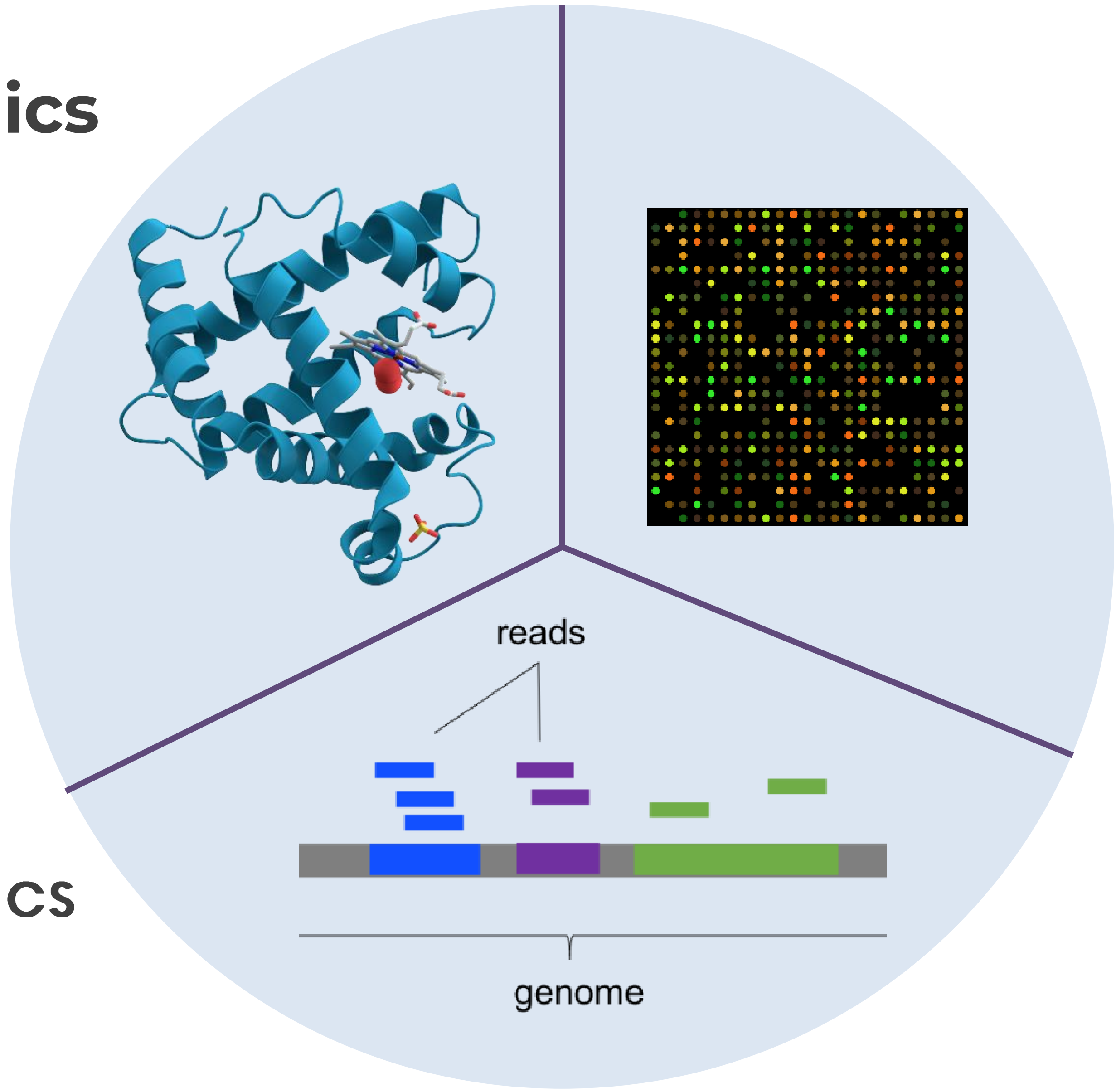
- Which properties were measured?
  - i.e. blood pressure, smoker/non-smoker, gene expression. Also called: **Features or Variables.**

- What is the **outcome** variable(s) (**dependent**) and which are **explanatory** variables (**independent**)?



OMICS DATA

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



Proteomics

Whole genome seq

SNP calling

Transcriptomics

RNAseq

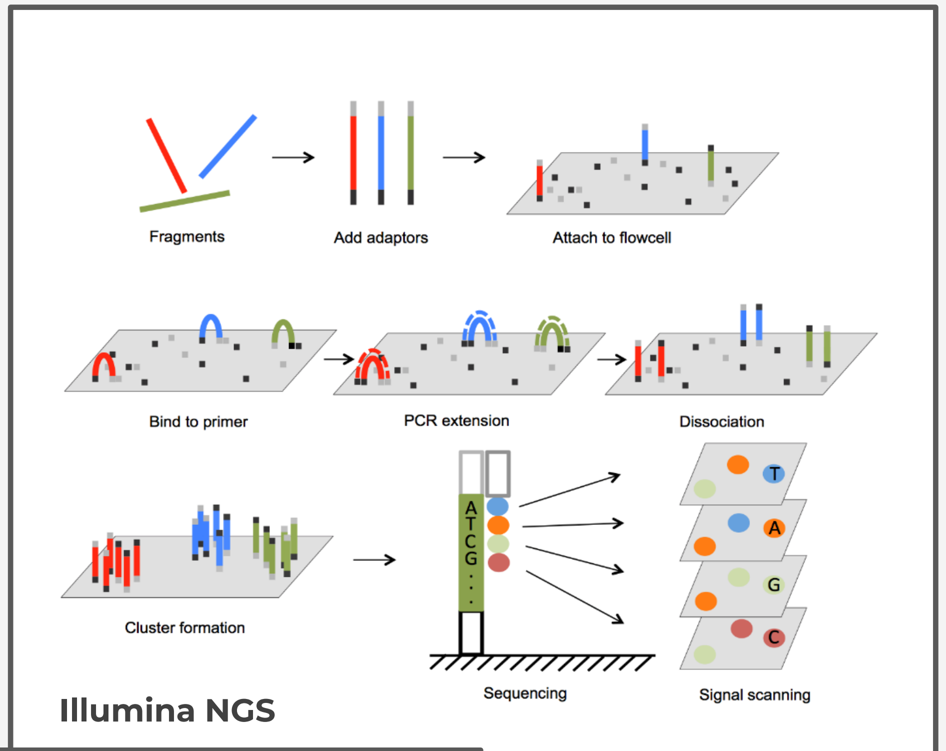
OMICS DATA

Omics data are produced with specialized lab protocols followed by:

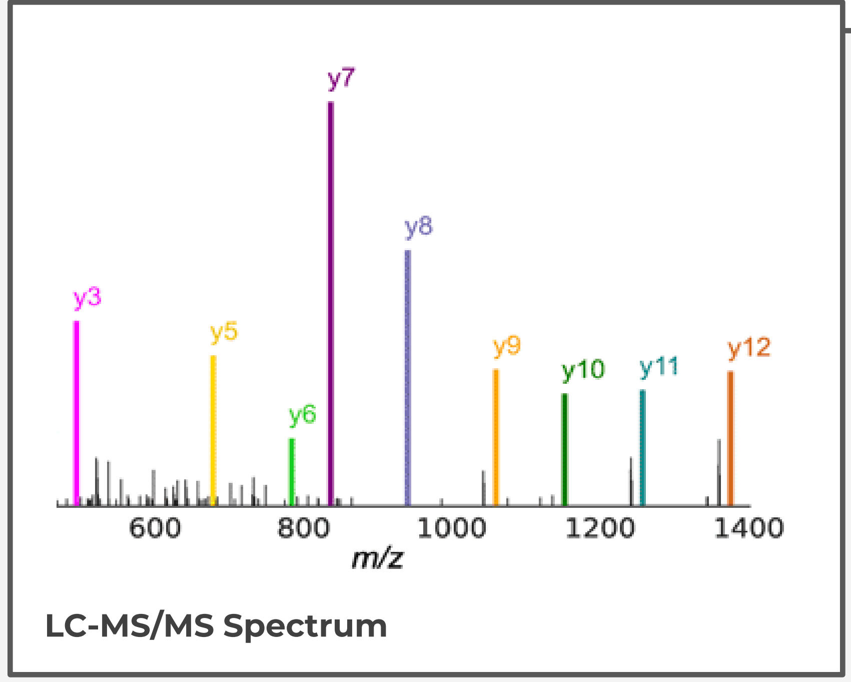
- Next generation sequencing (DNA, RNA)
- Mass spectrometry technologies (proteins, lipids, sugars, metabolites)

Data characteristics:

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



Illumina NGS



LC-MS/MS Spectrum

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

Digitalbevaring.dk

- 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!

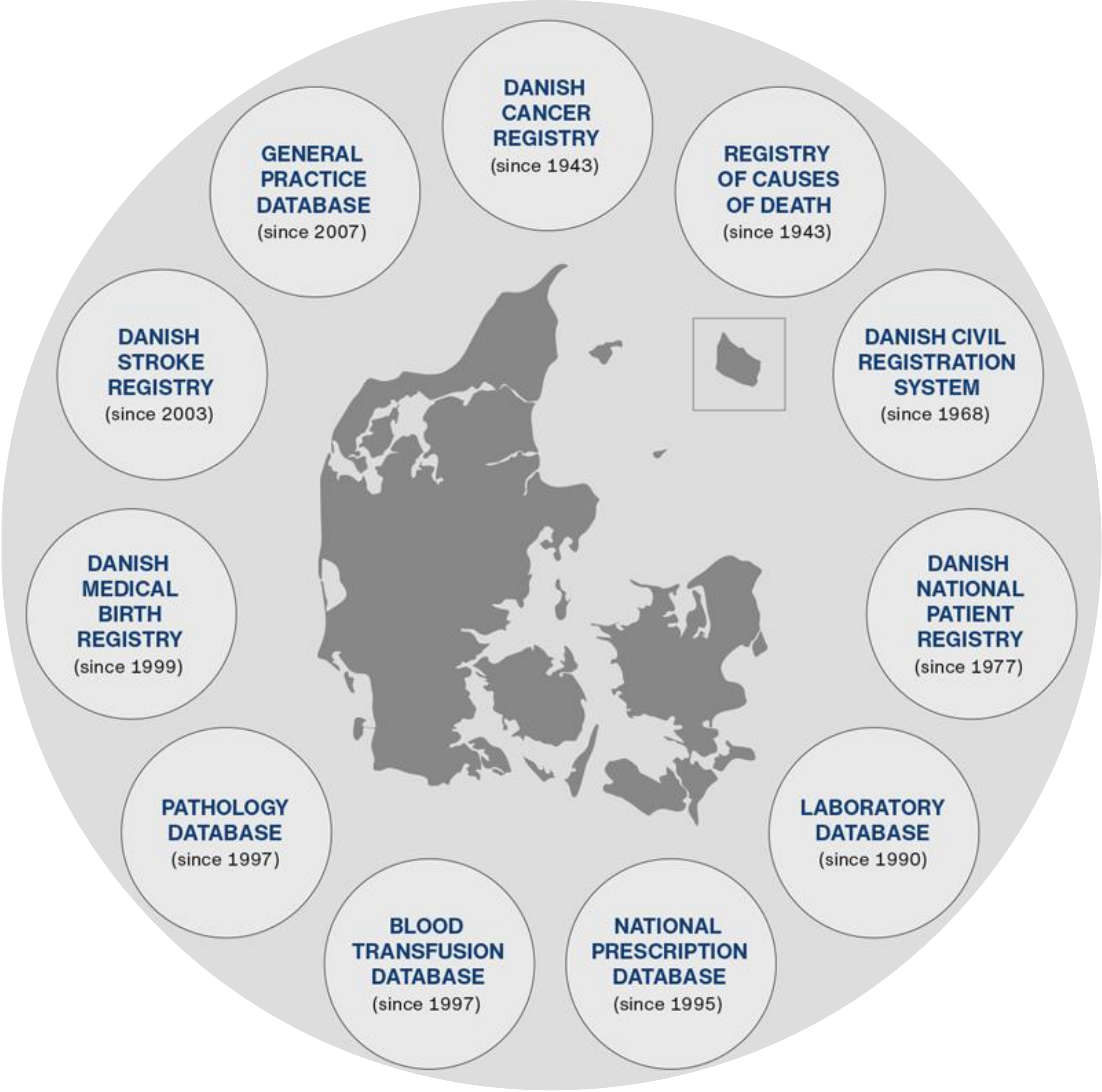
REGISTRY DATA

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).


DANISH CANCER REGISTRY (since 1943)
REGISTRY OF CAUSES OF DEATH (since 1943)
DANISH CIVIL REGISTRATION SYSTEM (since 1968)
DANISH NATIONAL PATIENT REGISTRY (since 1977)
LABORATORY DATABASE (since 1990)
NATIONAL PRESCRIPTION DATABASE (since 1995)
BLOOD TRANSFUSION DATABASE (since 1997)
PATHOLOGY DATABASE (since 1997)
DANISH MEDICAL BIRTH REGISTRY (since 1999)
DANISH STROKE REGISTRY (since 2003)
GENERAL PRACTICE DATABASE (since 2007)

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?

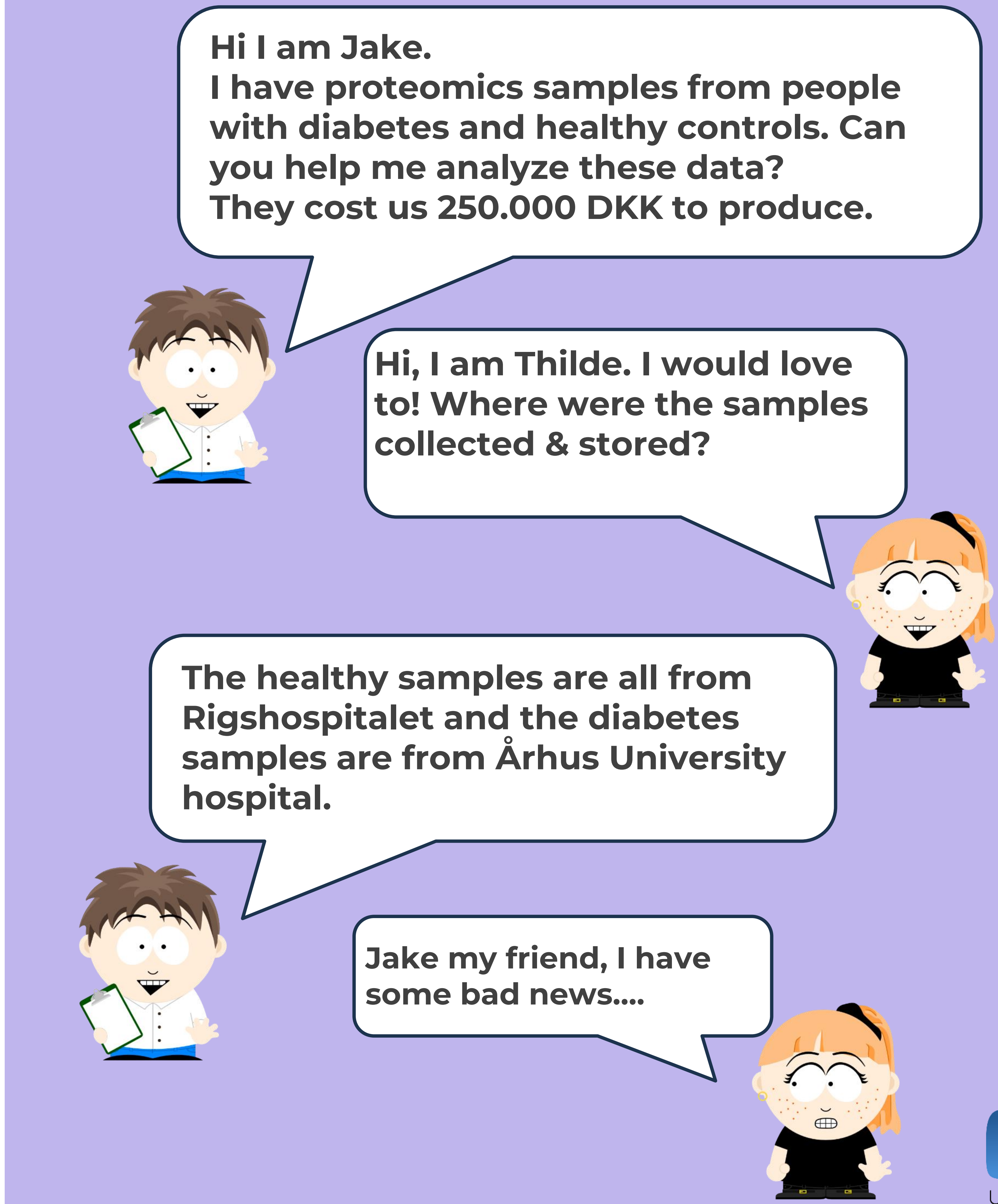
- You will be able to correctly answer your scientific question
- Saves you time/work
- Saves you money

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).


Hi I am Jake.
I have proteomics samples from people with diabetes and healthy controls. Can you help me analyze these data?
They cost us 250.000 DKK to produce.
Hi, I am Thilde. I would love to! Where were the samples collected & stored?
The healthy samples are all from Rigshospitalet and the diabetes samples are from Århus University hospital.
Jake my friend, I have some bad news....

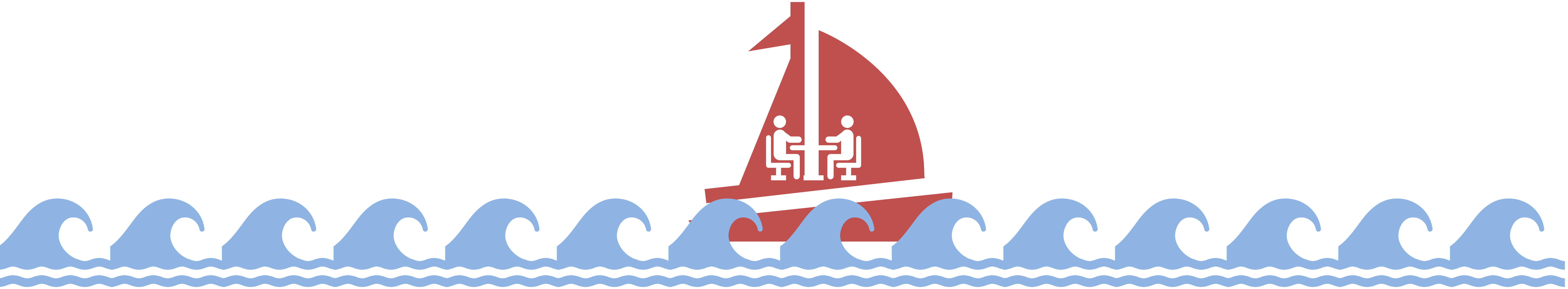
UCPH
HeaDS

GROUP DISCUSSION

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?

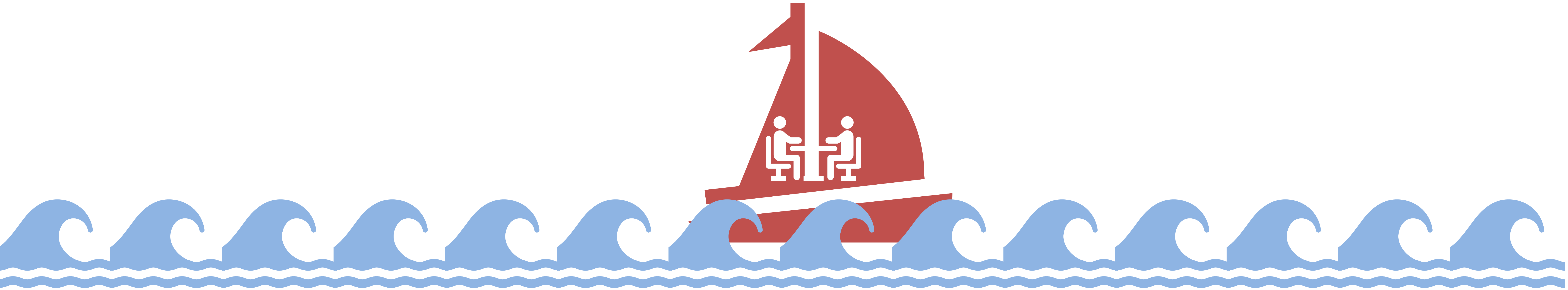


GROUP DISCUSSION

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'Ts

Don't store BIG DATA in your computer

No embedded XLSX sheets

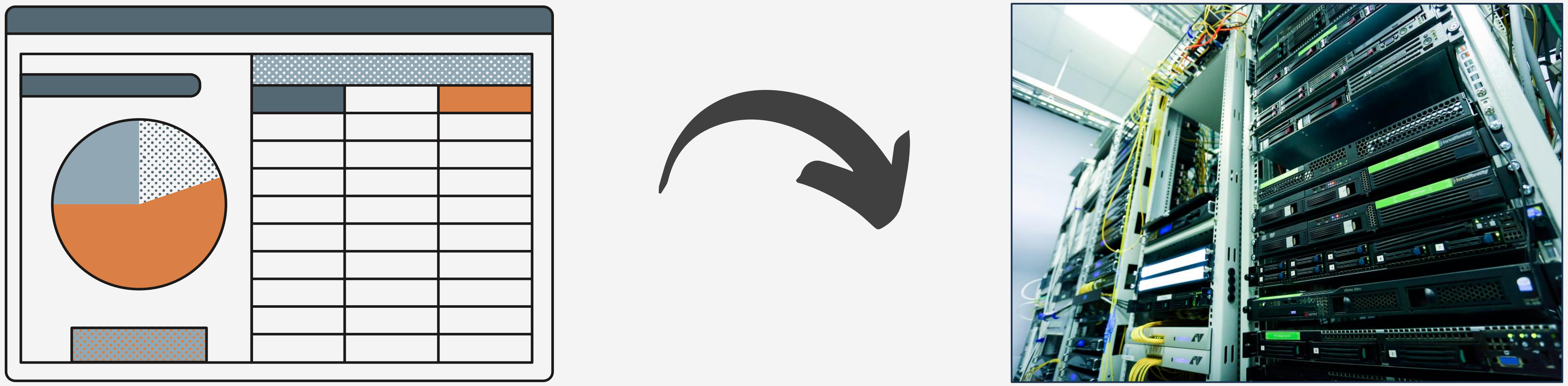
No color coding

No spaces nor special characters in naming

Do not do things 'manually' if avoidable

UCPH HeaDS

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. This has a couple of advantages:

- **Automation** of tasks
- **Workflow** is explicit and visible
- **Separate raw data and data processing**. The raw data is never touched
- **Working code** can be reused (perhaps with some tweaks) to analyze other similar datasets

The most popular languages are python and R

Programming languages

Quick overview

Easy starting

Raw data exposed

Manual analysis

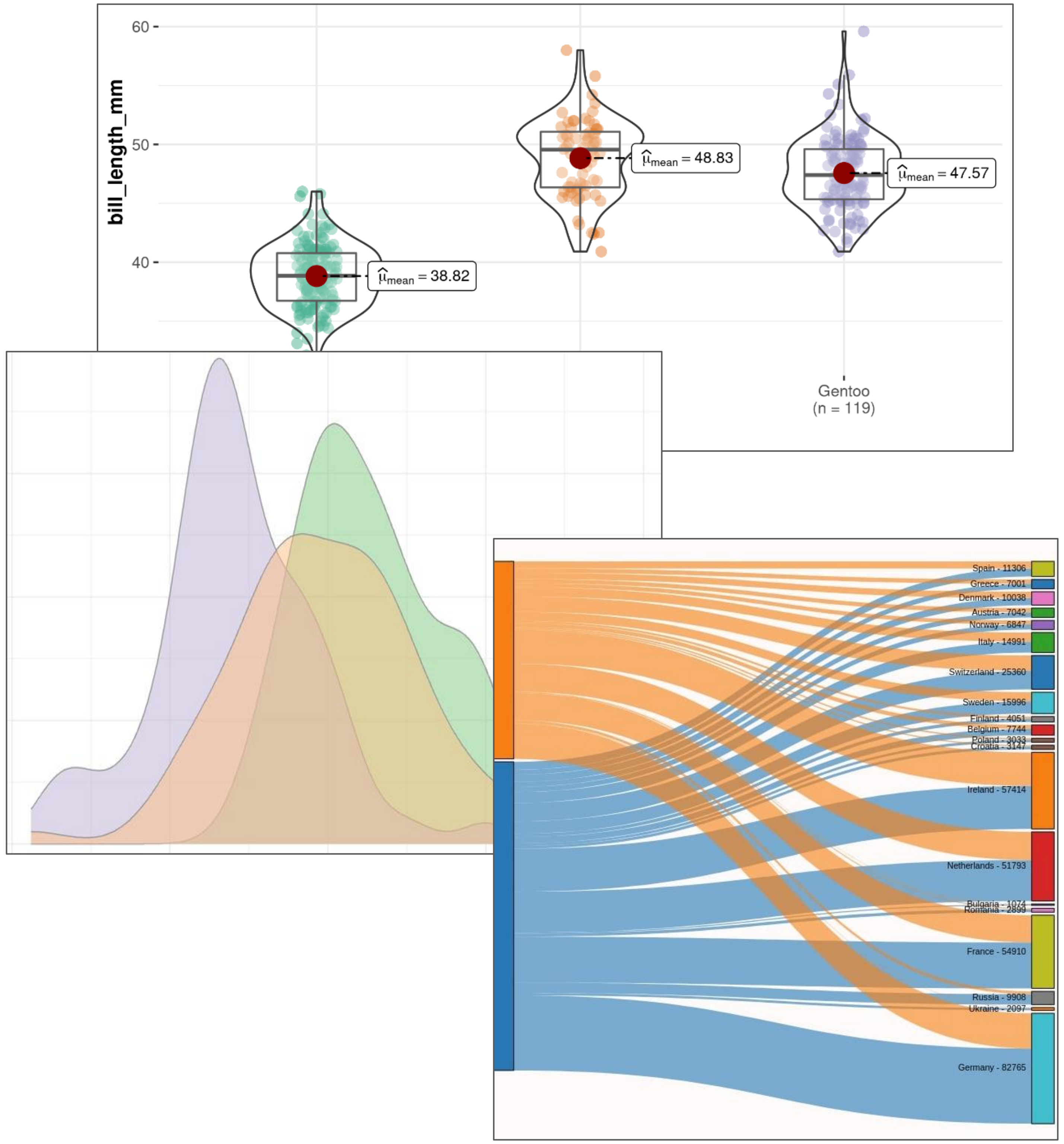
Processing power

Data Wrangling

Reproducibility

Complex analyses

Programming language

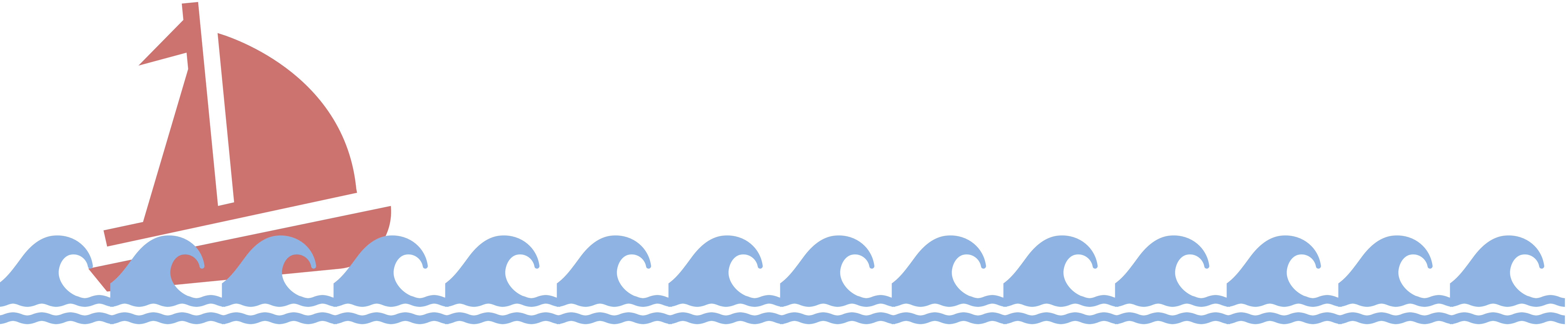


Beginner's Courses

UCPH HeaDS

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

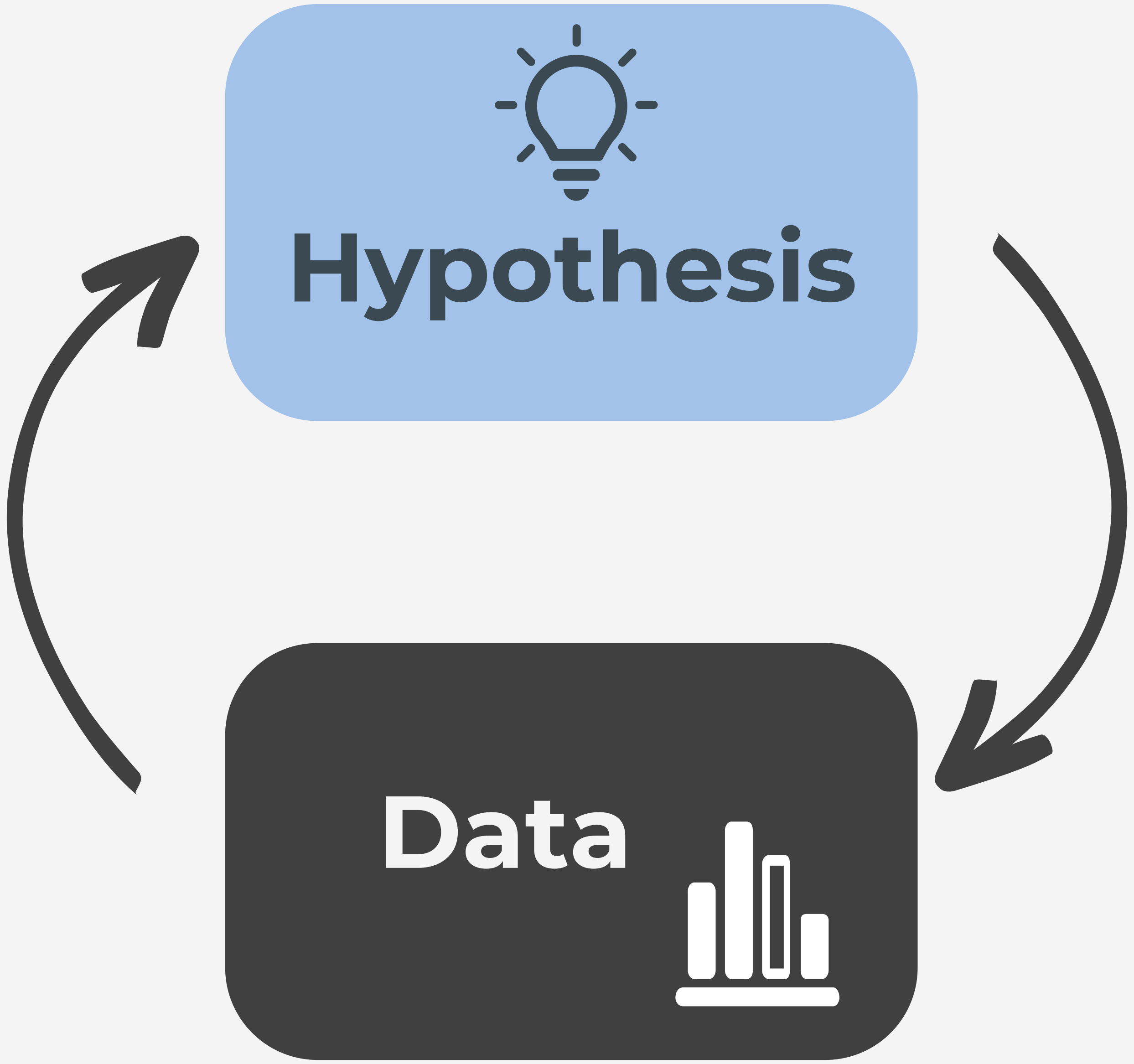
Data-driven:

- Discover properties of data set
- Identify patterns and relationships
- Mostly possible with BIG datasets
- data-driven ≠ 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



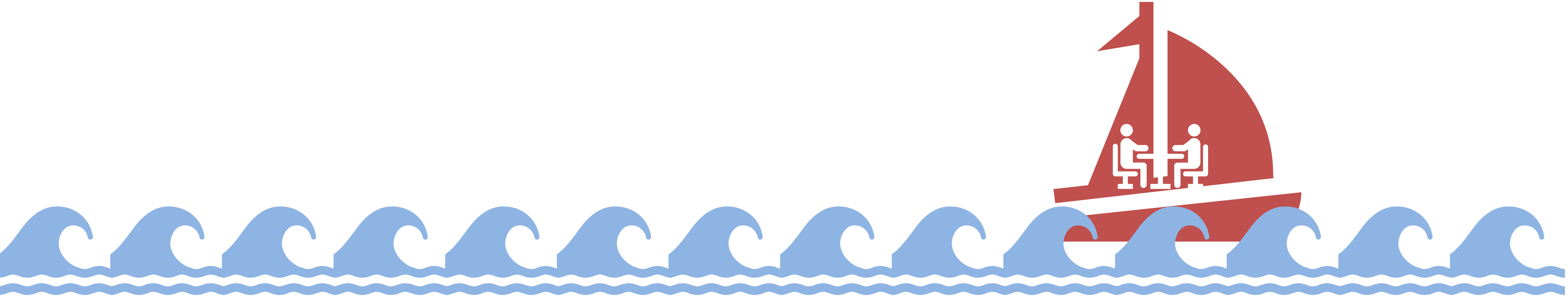
Gather new data to confirm or deny new hypothesis

We do both alternatingly

GROUP DISCUSSION

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