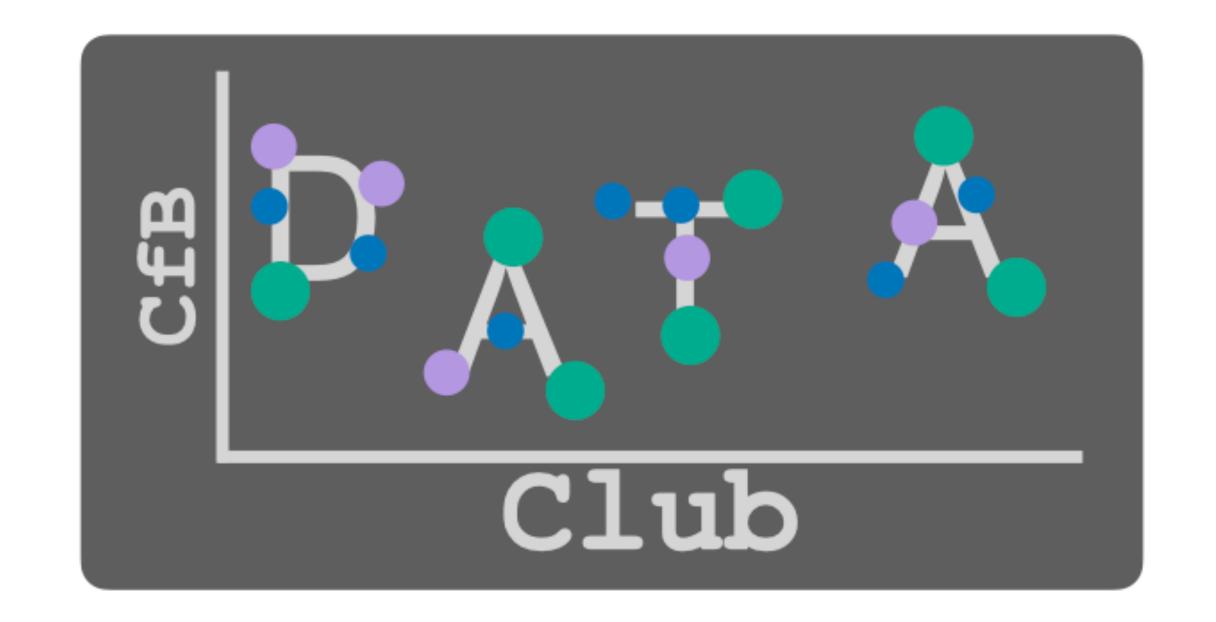
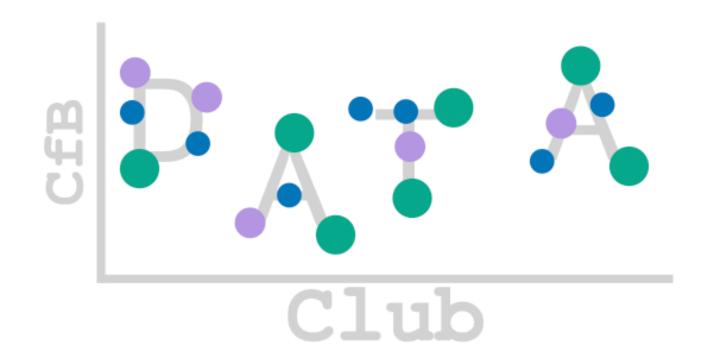
Intro & Data Literacy

Read, understand and communicate data



CfB Data Club

Building a community passionate about data



- A forum for researchers of all levels to discuss, learn, explore and analyse data
- Regular meetings every 2 months for 1.5h
 - A **formal part** with presentations, discussions about data-related topics, case studies, invited speakers, etc.
 - An informal part with sweets and coffee to approach presenters or discuss with other club members

Goals

- —> Foster a community
- -> Create a forum for **discussion** and **solving questions**
- —> Find the right people to help
- -> Learn from each other and share knowledge about data

Building a Data Culture

Driven and united by data

- Data and data science as a unifying factor across research disciplines
- Align on the importance of data and the processes to manage and the methods to analyse them
- Provide the support to acquire the skills and knowledge to access, read, understand, and communicate data effectively
- Focusing on collaboration by sharing data across teams and departments
- Use of data as a means to improve and innovate

Organisers











Teddy Groves

PhD in Philosophy of statistics Data scientist, Football Radar Co-PI QMCM kinetics, CfB

Expertise:

Kinetic models of cell metabolism Thermodynamics of biochemical reactions Statistics (especially custom/ Bayesian) Programming (especially Python)

tedgro@biosustain.dtu.dk

Alberto Santos

PhD in Bioinformatics PI Multiomics Analytics, KU Head Data and Infrastructure, Boehringer Ingelheim Pharma PI Multiomics Network Analytics, <u>CfB</u>

Expertise:

Omics Databases Graphs and Knowledge Graphs Programming (especially Python)

albsad@biosustain.dtu.dk

CfB Partnerships & Research Office (PRO)

Sünje Johanna Pamp

sipa@biosustain.dtu.dk



Data Experts

Lea Sommer and RDM team

Shilpa Garg

Kai Blin

. . .











Schedule

Preliminary Schedule

- Data literacy what is it? (Today)
- Data wrangling (May)
- Data visualization (July/August)
- Tools (September)
- Making use of publicly available data and databases (November)
- Analytics for understanding and predicting (December)



Extra

The club is flexible and adapts to your needs

- The schedule provides structure
- But if you want to discuss, learn or work on a specific topic, let us know!
- We can add extra meetings to accommodate requests that are relevant to the members of the club
- For instance:

Do you want to present your project and collect ideas or feedback?

Do you want to **invite someone** to give a talk?

Do you have a specific data problem that you need to discuss about?

Do you have an idea you want to present and gather a team to work on?

Tools

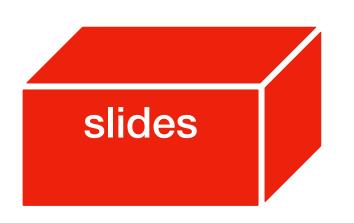
Data Club Website

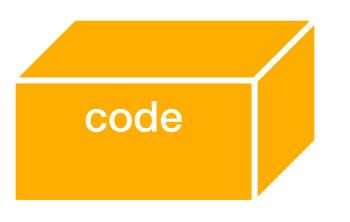
https://github.com/Multiomics-Analytics-Group/cfb_data_club

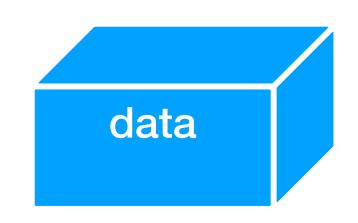
- We use GitHub to maintain the contents of the data club
- GitHub is a framework generally used to maintain and version control software development projects



 Version control is a system that records changes to a file or a set of files over time and who made them so that you can recall specific versions later







Code Python



- Python is a high-level, general-purpose programming language
- It is a good language for both beginners and advanced programmers:
 - Its design philosophy emphasises code readability
 - Many learning resources available
 - Large community behind (i.e. stackoverflow for questions)
 - Excellent Libraries available, especially for data science

Code R



- R is a programming language for statistical computing and graphics
- R is used among data scientists, bioinformaticians and statisticians for data analysis and development of statistical software
- Many learning resources available
- Large community behind
- Excellent Packages available for data and statistics

Jupyter Notebooks

https://jupyter.org/



- Web-based development environment for creating, running and sharing
 Python (and other languages) code
- A notebook is an interactive document that combines live code, equations, text or markdown, and visualisations (output of your code)
- Notebooks are divided into cells that run sequentially! (Need to pay attention)
- It requires having Python installed on your local machine



Colab Notebooks

https://research.google.com/colaboratory/faq.html

- Google Colab is based on Jupyter Notebook open source project hosted on Google's servers
- Advantages:
 - Requires no setup to use (no python installation)
 - Provides free access to computing resources on Google's servers including GPUs
 - Notebooks can be shared just as you would with Google Docs or Sheets.
 - You can import existing Jupyter notebooks
- Own data and notebooks need to be accessed through Google Drive Need Google account or accessible online

Note: Internal Data

If we want to use internal data (confidential) we will work in a secured infrastructure



data

We need you!

- We will use publicly available data to showcase topics and methods
- We will try to find relevant biological datasets
- But we would love to have your help:
 - Propose us to use your own data or type of data
 - Bring your ideas of how to best explore, analyse or visualise them
 - Ask specific questions or issues we should take in the scheduled meetings or extra ones

Data Literacy

Data is like garbage. You'd better know what you are going to do with it before you collect it. Mark Twain

What is it?

Data as a universal language

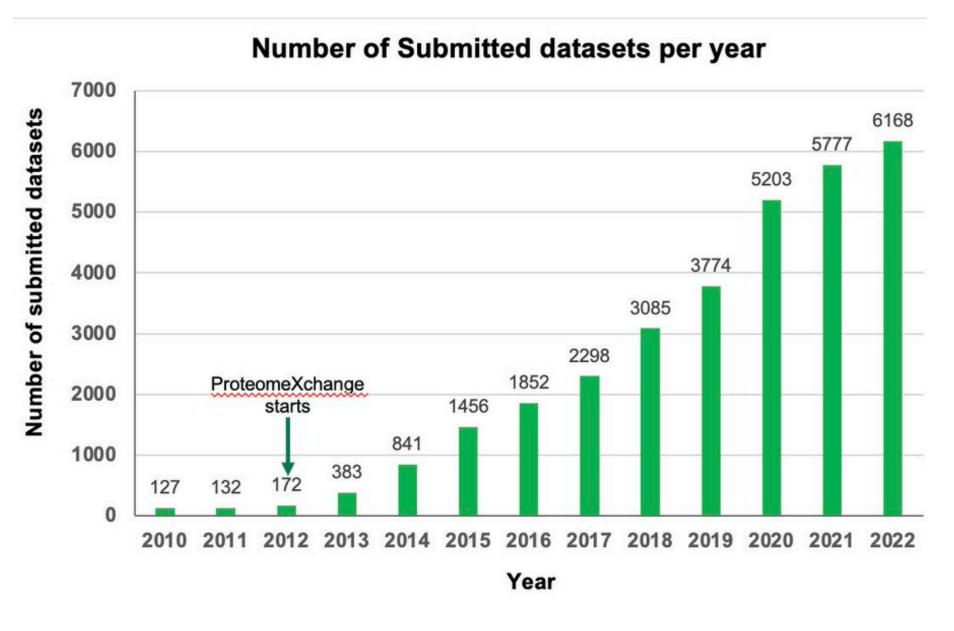
- The ability to read, analyse and communicate data effectively
- Includes also understanding data sources and formats, and the analytical methods to transform them into results
- Requires skills and knowledge in data analysis, statistics, data visualisation, and data management.
- It does not mean that we all need to be Data Scientist ...
- **but** we all need certain skills to extract value from data:
 - Non-technical: interpretation, critical thinking, communication, evaluation
 - Technical: analysis, visualisation, programming, management

Why?

Data! Data! Data! Evidence to support our research

- We generate data everyday: metadata, raw, processed, clean, derived results
- Growing number of resources and datasets available
- The data can have multiple purposes reusability (FAIR)
- Benefits:
 - Better understanding of data
 - Extract data-driven insights
 - Freedom to create, format and analyse
 - Meaningful communication with data
 - Manage your data sustainably

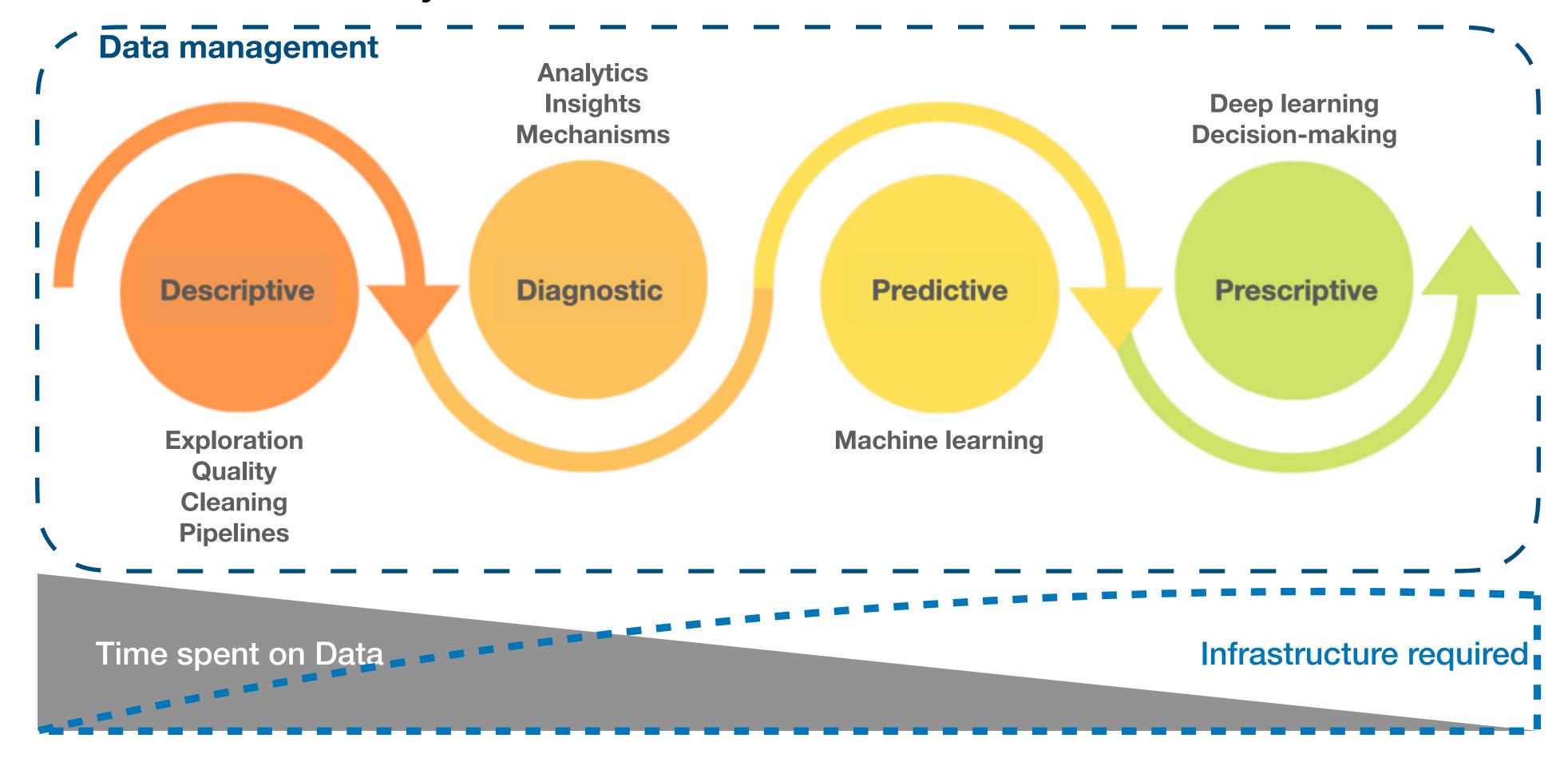




Our Data Journey

Be Data Literate: The Data Literacy Skills Everyone Needs To Succeed. Jordan Morrow

4 Levels of Data Literacy

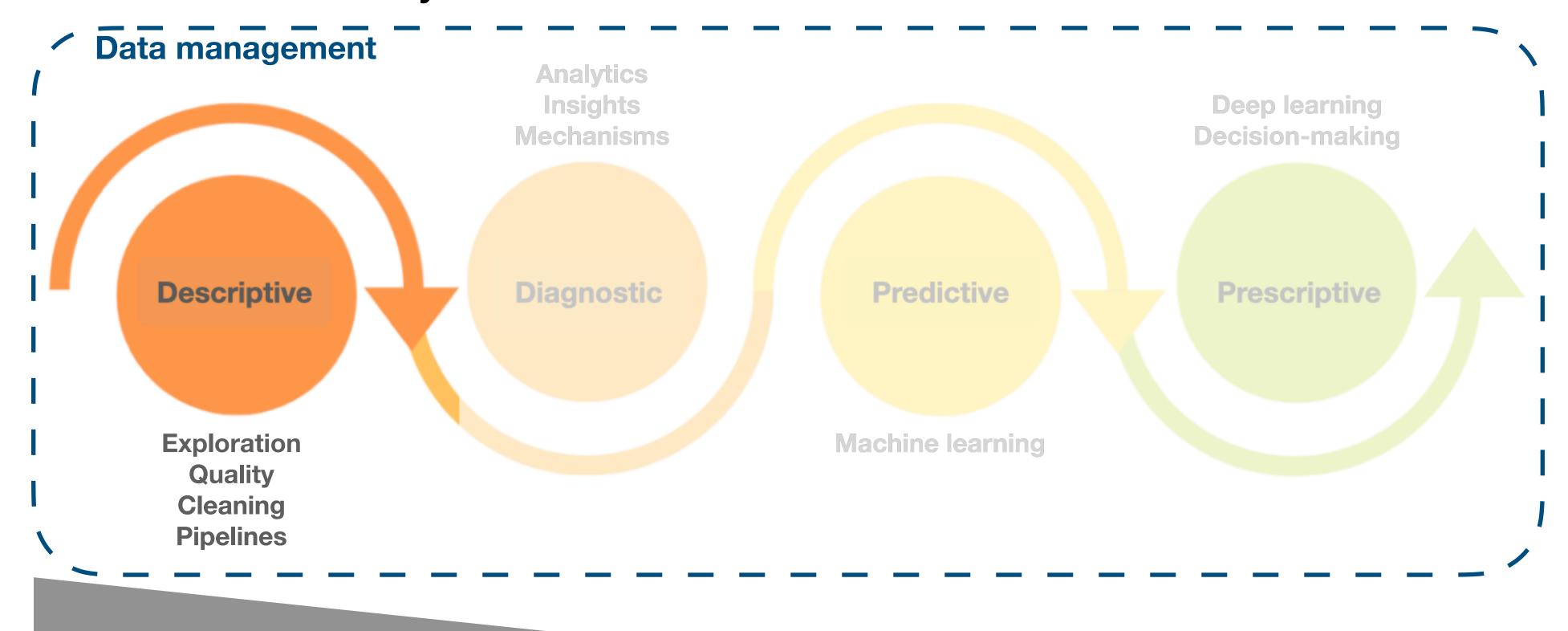




Our Data Journey

Be Data Literate: The Data Literacy Skills Everyone Needs To Succeed. Jordan Morrow

4 Levels of Data Literacy



Time spent on FAIR Data

You are not alone

Support:

- RDM team (Lea Sommer)
- Data Science Platform (Alberto Santos)

Data and Data Science collaborations and discussions:

- Quantitative Modeling of Cell Metabolism (Teddy Groves)
- Genomics Sustainable Solutions (Shilpa Garg)
- Natural Products Genome Mining (Kai Blin)
- Multiomics Network Analytics (Alberto Santos)

Data Club!

Descriptive level

- Focus on reading, and visualising data
- Output is generally a graph, a dashboard or a report
- Requires:
 - Exploring the data
 - Shaping and standardising the data
 - Evaluating the quality of the data
 - Extracting some metrics
 - Visualising these metrics



Data wrangling (May)

Data exploration

Data Quality

Data cleaning

Data pipelines: nextflow, snakemake

Data visualisation (July/August)

Data simplification and communication

Plots and Colours

Dashboarding tools (Jupyter notebooks, markdown, streamlit)

Any Questions, Ideas, Suggestions?

Time for cake!