Introduction to Data Science in Life Science

Marta Milo
AstraZeneca
Cambridge, UK

September 2020

Outline

- How did the data grow and what are we facing?
- Challenges of ML applications in Data Science
- Data Science and its principles
- Making your data ready for the future

Needs and Challenges we face

- Growth of data and advances in technologies have highlighted the importance of studying genetic variability to interpret phenotypic observations

- Understanding in the context of biological variability. For example, what type of OMICS variations might lead to the phenotypic changes.

- The need to interpret and exploit all the data/knowledge we have collected.

- Interpretation of modern data in isolation is very hard. Use of integrative models to analyse the data jointly. **What is relevant and what is ambiguous?**

Calling on Machine Learning...

Using the mathematical tools available and integrate the biological knowledge we collected so far, use of statistics in combination to bring the field forward.

Current data is complex and we need models that are able to ***learn*** patterns and associations in the data autonomously. Artificial Intelligence and Machine Learning applications.

Establish the concept of ***trust*** about the data and ensure that it can be ***reused.***

Identify consequences of this advanced of technology and how to control them.

"Technology we create must adapt to our needs and to us, we need support from computers and statistics in responding to this adaptation in order to control it, but not loose control over it."

Prof N. Lawrence

Challenges of ML in Life Science

ML challenges:

- How do we extrapolated the information from the data to train and then test models?
- How do we use domain adaptation in applications to life science?
- How do interpret the data and maintain the ability of the models to learn?

Translated to Life Science…

- **Combine gene–level analyses with pathway–based methods** to generate a comprehensive profile of the functional modules that govern biological processes.
- We want **to use high–throughput data to build models of data integration,** to predict at the systems level.
- Design therapeutic intervention and /or genomic predispositions to disease at individual level.

Summary Part I

Data Science is a multidisciplinary field that has opened new frontiers for Life science

New Challenges are now facing us when implementing Data Science and Machine Learning

These challenges are not just on data acquisition but also how we adapt the models and their “environment” to our questions (needs)

Successful use of Machine Learning methods relies on our knowledge of those challenges and on good practice and data sharing policies.

And on Experimental design (workflow) and on a good characterisation of Data structure

The life cycle of a Data Science Project

EXPLORATORY DATA ANALYSIS

DEFINES AND REFINES SELECTION OF FEATURE FOR THE MODEL

DATA MODELLING

TRANSFORMATION

DATA CLEANING

MISSING DATA POINTS
DUPLICATIONS
WRONG ANNOTATIONS

VISUALISATION AND COMMUNICATION

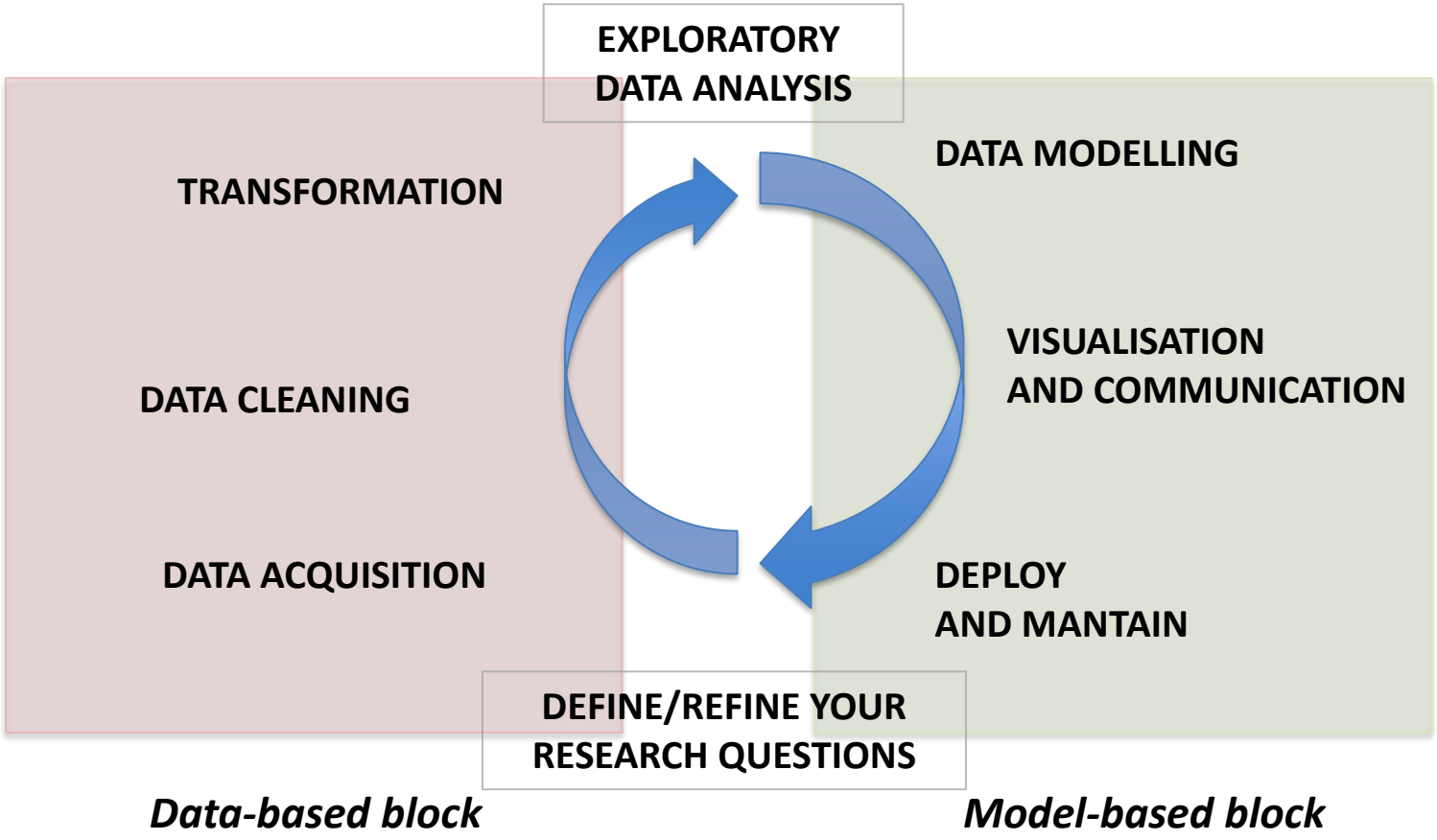
DATA ACQUISITION

- *EXPERIEMENTAL DATA*
- *ON LINE REPOSITORIES*
- *OTHER LABORATORIES*
- *WEB SERVERS*
- *META DATA*
- *IMAGES*

Define your research questions

DEPLOY AND MANTAIN

The life cycle of a Data Science Project (2)

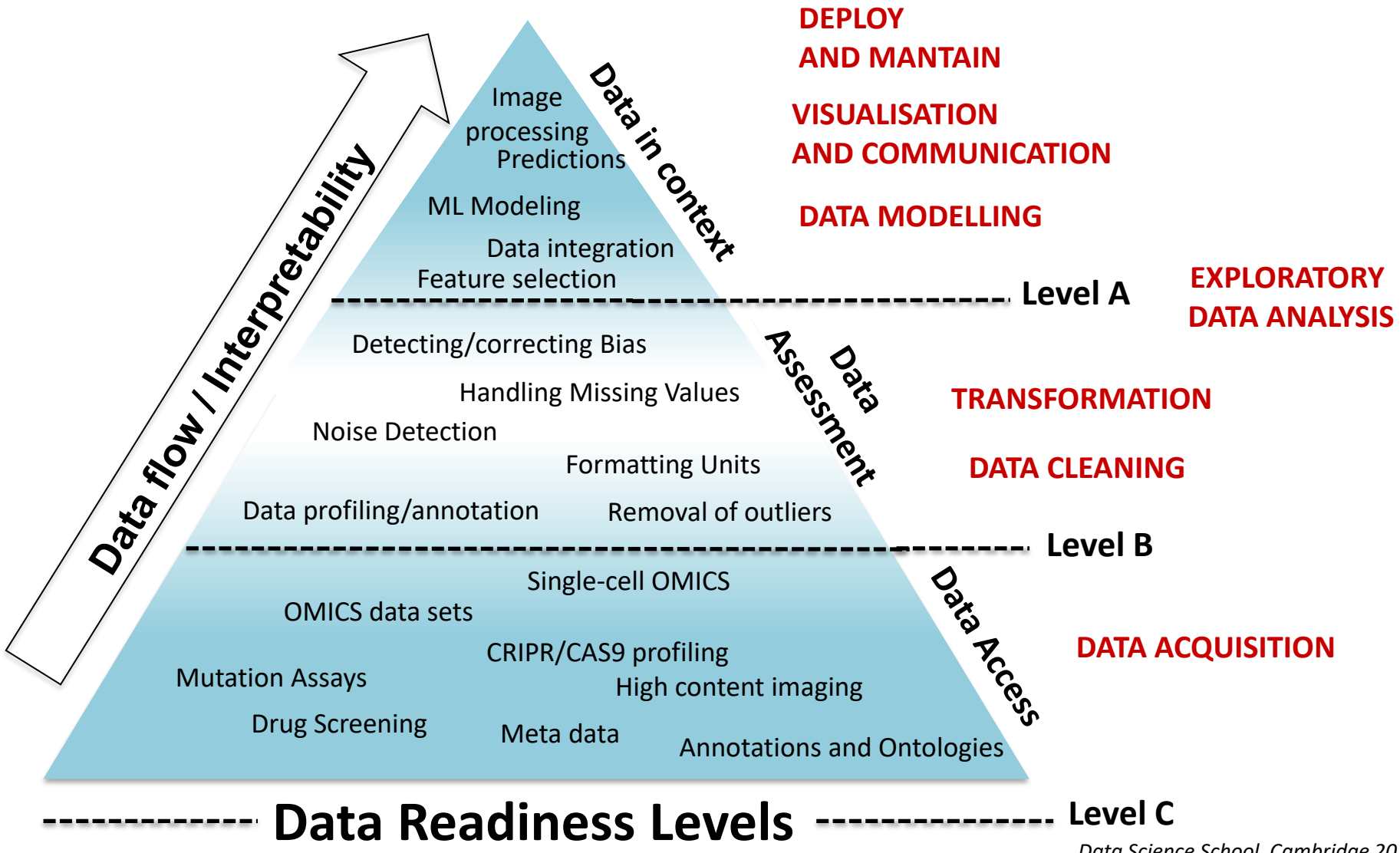


Data Structure to support modelling

- Machine Learning models need data and a lot of it
- In Life Science we have huge amount of data not just quantitative data
- Data is scattered around different labs, repositories and different platforms
- We keep producing data … but how are we going to use all?
- Data Science might be a good approach to make use and interpret current biomedical data data

We need to organise and characterise the data in relation to Data Science

Data organisation and Characterisation



Data Science School, Cambridge 2020

Experimental Design for data analysis

When we set up an experimental design for data analysis we need to keep in mind:

1. Importance of defining your research questions, keeping in mind limitations and effective use of the data

2. Consistency in sample preparation, optimisation of the samples, extensive QC of the data. LOOK at the data generated and QC before processing

3. Clean and prepare that data, what strategy we need to use and we treat missing values

4. Identify noise sources and define possible noise models

5. Choose the correct model to analyse your data, define appropriate parameters to get the maximum information out of your data.

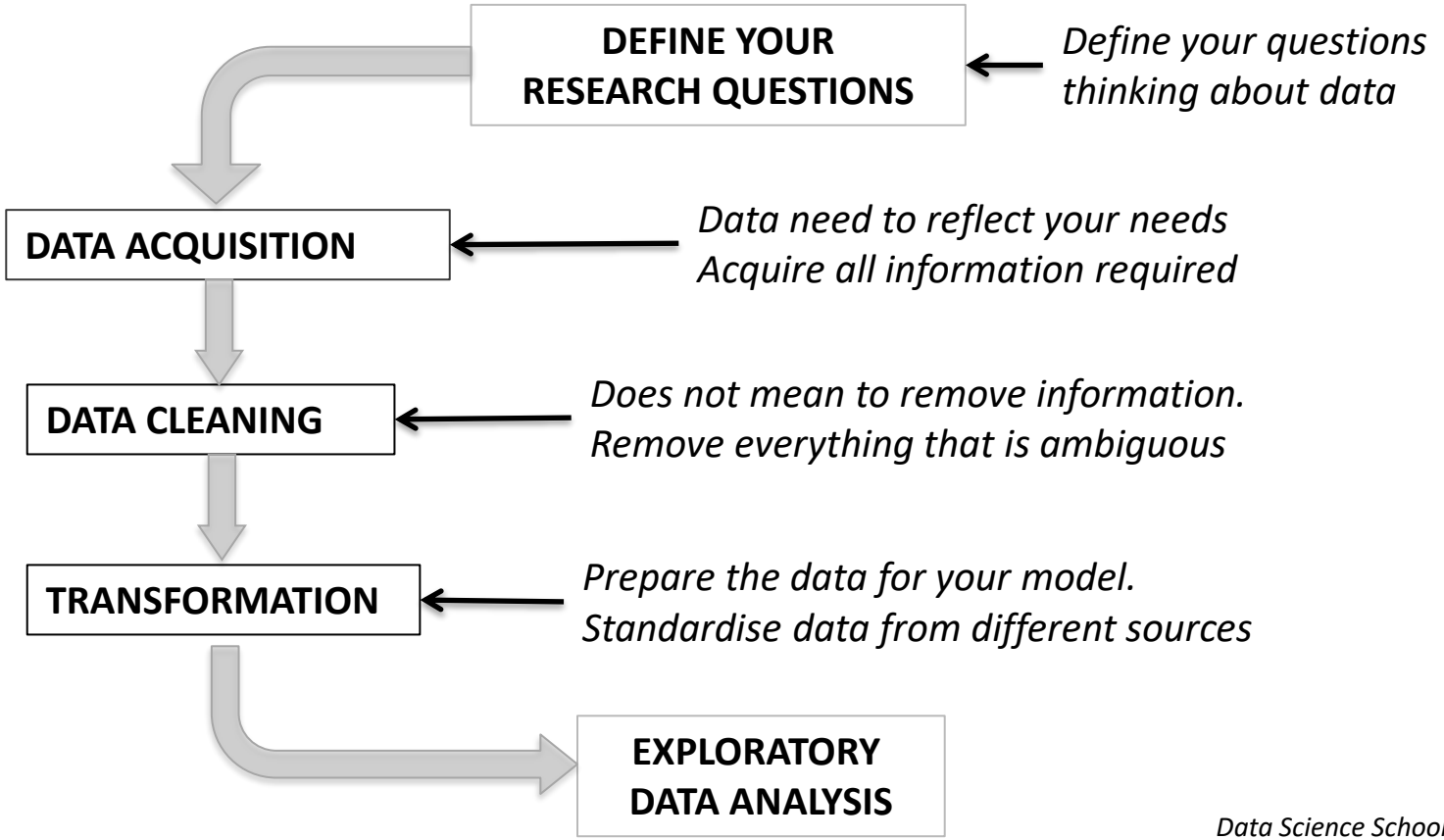
6. Use the best tool to visualise your data, to discriminate, cluster and rank your significant targets

Data-based experimental design

1. Importance of defining your research questions, keeping in mind limitations and effective use of the data

2. Consistency in sample preparation, optimisation of the samples, extensive QC of the data. LOOK at the data generated and QC before processing

3. Clean and prepare that data, what strategy we need to use and we treat missing values

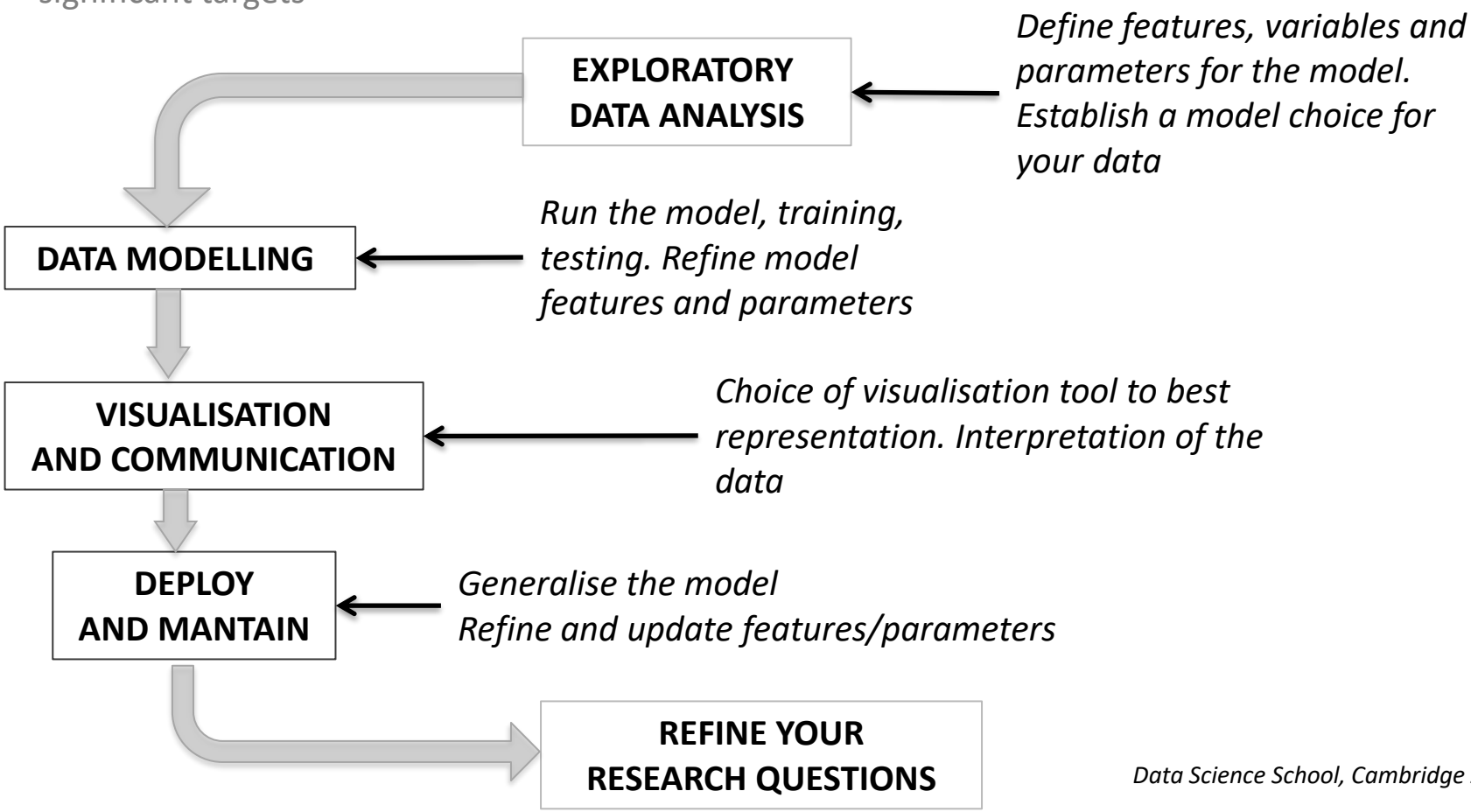


Model-based experimental design

4. Identify noise sources and define possible noise models

5. Choose the correct model to analyse your data, define appropriate parameters to get the maximum information out of your data.

6. Use the best tool to visualise your data, to discriminate, cluster and rank your significant targets



Summary Part II

Data Science might be the best approach to discover more knowledge from current and future Life Science data

The experimental designs need to reflect the principles of data science and all phases of the cycle of a Data Science project

When applying ML to Life Science there are types of experimental designs: data-based and model-based. They intercommunicate

A Data Science project aims to optimally deploy its results but for a life science application this means that we need refine our original research questions in order to advance.