Principles of Experimental Design

Marta Milo
University of Sheffield
Department of Biomedical Science

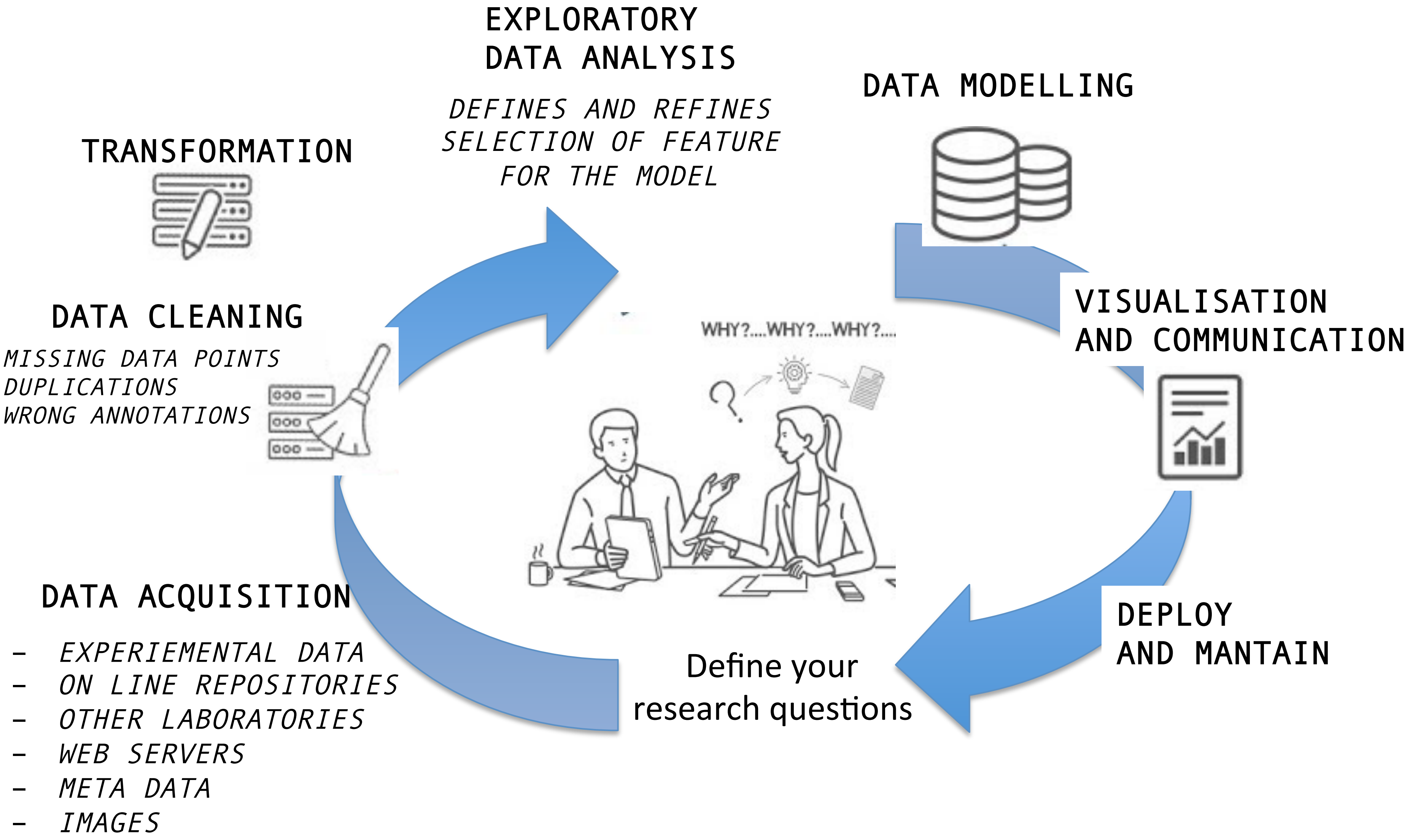
September 2019

Autumn School in Data Science, Cambridge 2019

Outline

- The Life Cycle of a Data Science project
- Data structure to support data modelling
- Data Science principles in the contest of Life Science
- data-based experimental design
- Model-based experimental design
- Points for reflecting

The life cycle of a Data Science Project



Define your research questions

DATA ACQUISITION

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

DATA CLEANING

MISSING DATA POINTS
DUPLICATIONS
WRONG ANNOTATIONS

TRANSFORMATION

EXPLORATORY DATA ANALYSIS

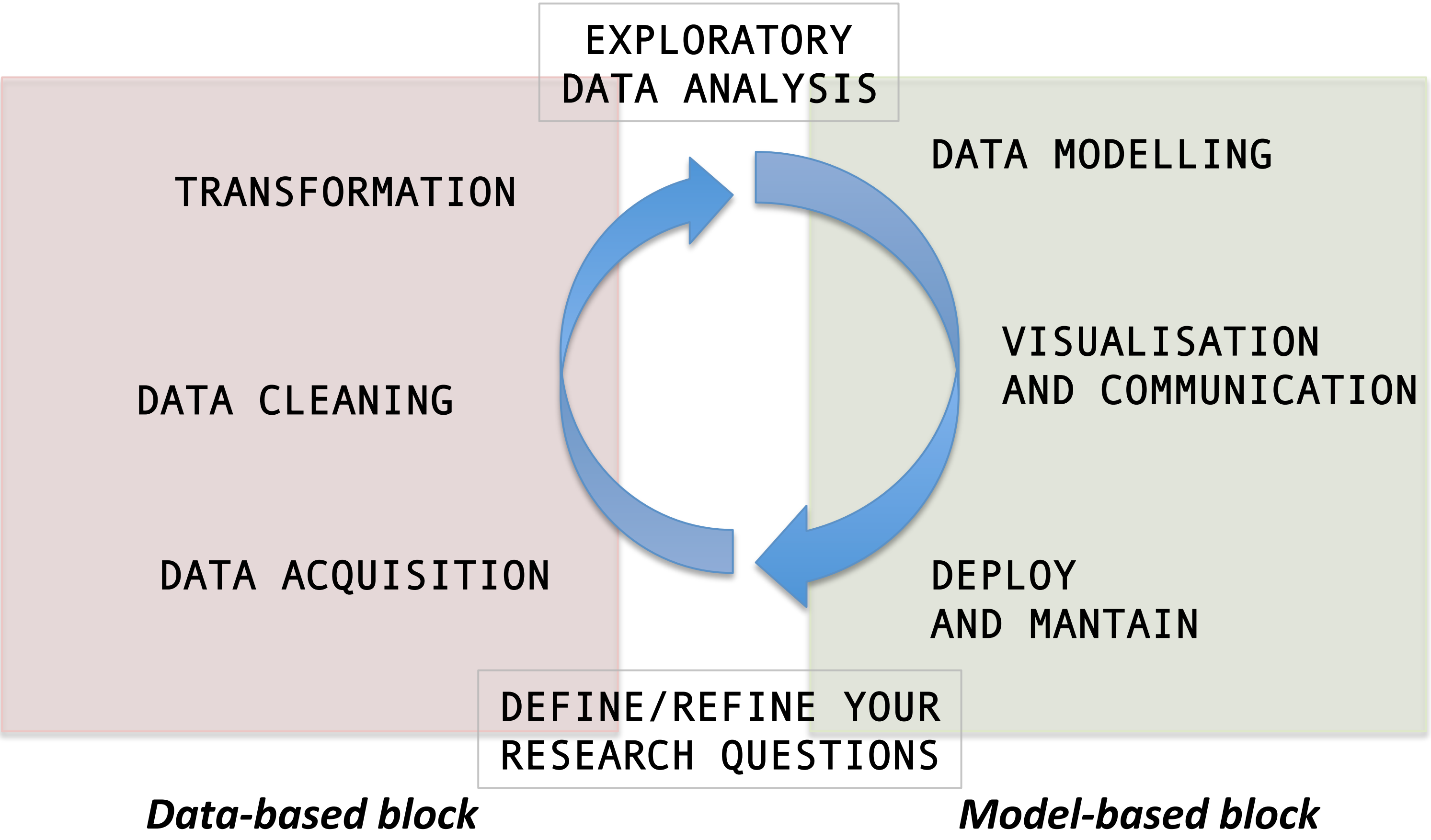
DEFINES AND REFINES SELECTION OF FEATURE FOR THE MODEL

DATA MODELLING

VISUALISATION AND COMMUNICATION

DEPLOY AND MANTAIN

The life cycle of a Data Science Project (2)



EXPLORATORY DATA ANALYSIS

TRANSFORMATION

DATA MODELLING

VISUALISATION AND COMMUNICATION

DATA CLEANING

DATA ACQUISITION

DEPLOY AND MANTAIN

DEFINE/REFINE YOUR RESEARCH QUESTIONS

Data-based block

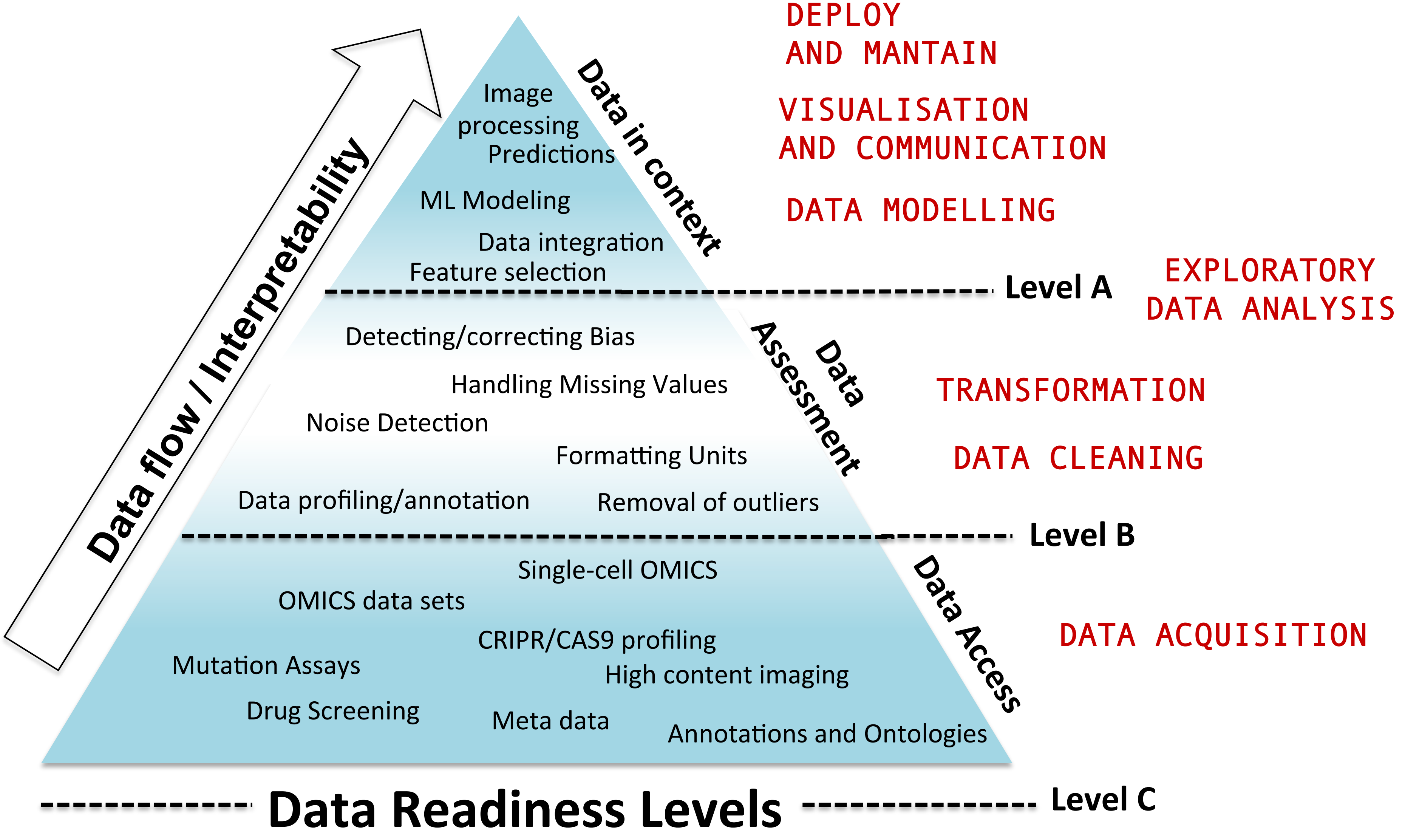
Model-based block

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



Autumn School in Data Science, Cambridge 2019

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.

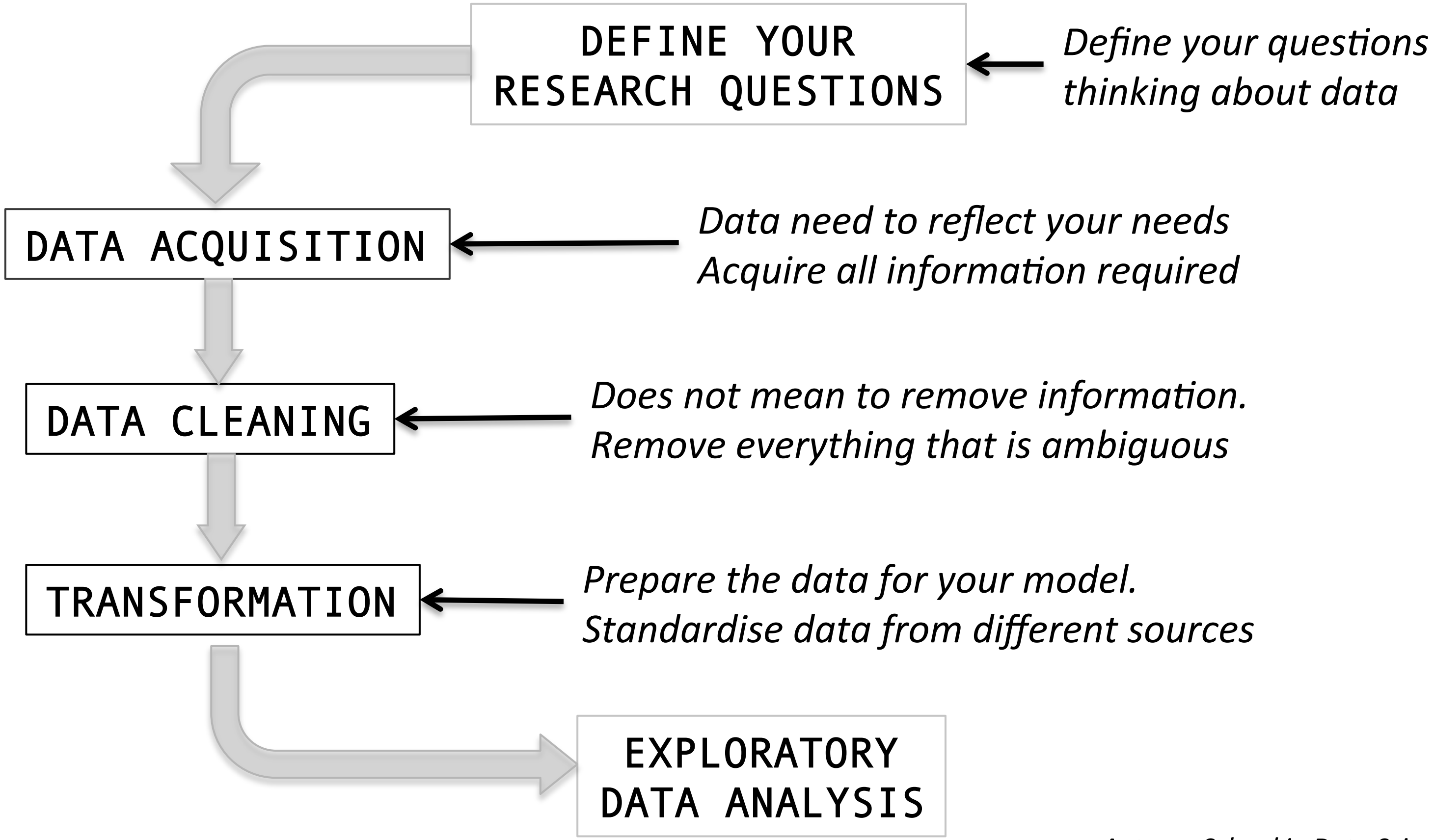
6. Use the best tool to visualise your data, to discriminate, cluster and rank your significant targets

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



DEFINE YOUR RESEARCH QUESTIONS

Define your questions thinking about data

DATA ACQUISITION

Data need to reflect your needs
Acquire all information required

DATA CLEANING

Does not mean to remove information.
Remove everything that is ambiguous

TRANSFORMATION

Prepare the data for your model.
Standardise data from different sources

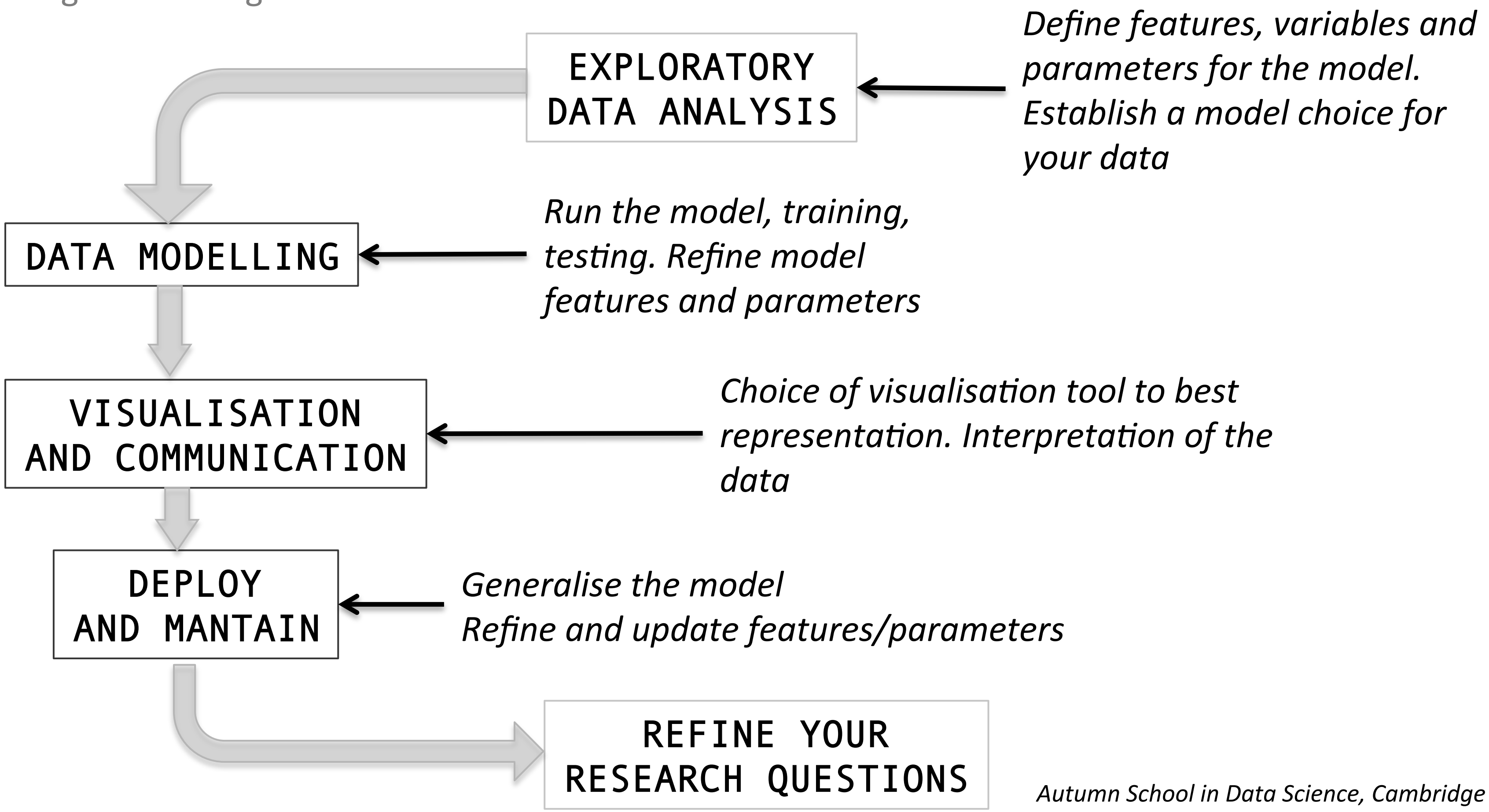
EXPLORATORY DATA ANALYSIS

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



EXPLORATORY DATA ANALYSIS

Define features, variables and parameters for the model. Establish a model choice for your data

DATA MODELLING

Run the model, training, testing. Refine model features and parameters

VISUALISATION AND COMMUNICATION

Choice of visualisation tool to best representation. Interpretation of the data

DEPLOY AND MANTAIN

Generalise the model
Refine and update features/parameters

REFINE YOUR RESEARCH QUESTIONS

Summary

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.

Challenges of ML in Life Science (Q&A session)

The University Of Sheffield.

- Working with a multidisciplinary approach, collaboratively in order to acquire a suitable level of domain expertise
- Often we have limited/"frozen"/ retrospective data which can be still considered "big data". High complexity
- Correlations vs causation and the role that confounding factors play in the model
- Interpretability of the confounders and difficulties in getting outputs deployed in the "real world".
- Different scales of measurements, problem of transforming the data. Use of latent variable models to overcome this.
- In some cases we have "ill-defined" phenotypes. What we observe is not directly connected to the data. Behavioral phonotypes and omics data.
- Privacy of data which clashes with ML needs.
- Ethics and Fairness to reuse and recycle the data.
- Mitigation of risks when ML is applied and define what population will benefit