ML: Some Guidelines

University of Victoria - PHYS-555

1 - Problem Formulation

- What am I trying to solve?
    - do I need ML? supervised? regression? would classify those objects really give me a better understanding of the physics?

- Which information do I have access to answer this problem?
    - structured data? noisy? metadata? is there hidden information?
    - how can I inform my model with induction biases that I perceive on the data?

2 - Become one with the data

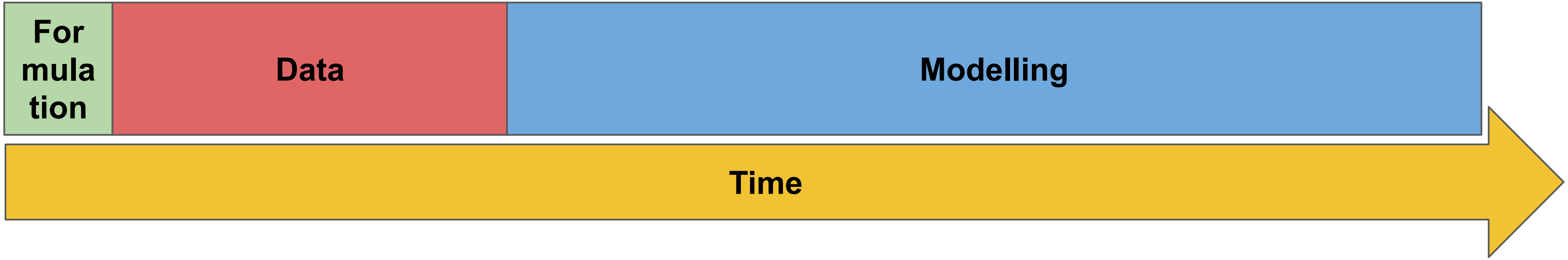
- **Estimate**:
  - How much data do I have access to? How hard is it to get more?
  - What is the data structure: sequences, images, tables. Any known relations?
  - What was the selection process for the data?
  - Can I possibly do data augmentation | simulation?
- **Explore**: plot distributions, correlations, missing, imbalance, get an intuition!
- **Label**: what kind of labels I have access? Are they noisy?
- **Preprocess**: will I need rescaling, do I need feature engineering?

3 - Modelling and interpretation

- **Algorithm**:
  - select algorithms according to the data (tabular, sequential, images, 3D, irregular graphs,...)
  - start with a reliable baseline, i.e. RandomForest, ResNet18, …
  - be ready to pipeline several algorithms
- **Evaluation**: make sure you have metrics
- **Interpretation**: can I trace the results back to the data and act?
- **Robustness:** if I share my pre-trained model, how would it fail?
- **Hyper-parameters**: the gain is often small, reserve for last.

Time Management

Typical perception



Formulation

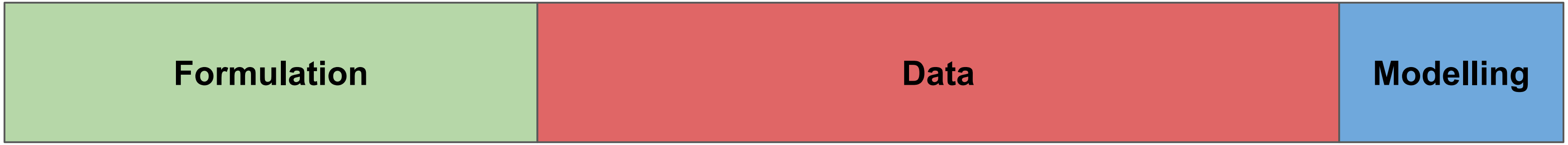
Data

Modelling

Time

Time Management

Perceived impact years after results



Formulation

Data

Modelling

Be ready to seriously revisit your initial time estimates