The University Of Sheffield.

Introduction to Data Science and ML in Life Science

Marta Milo
University of Sheffield
Department of Biomedical Science

September 2019

The University Of Sheffield.

Outline

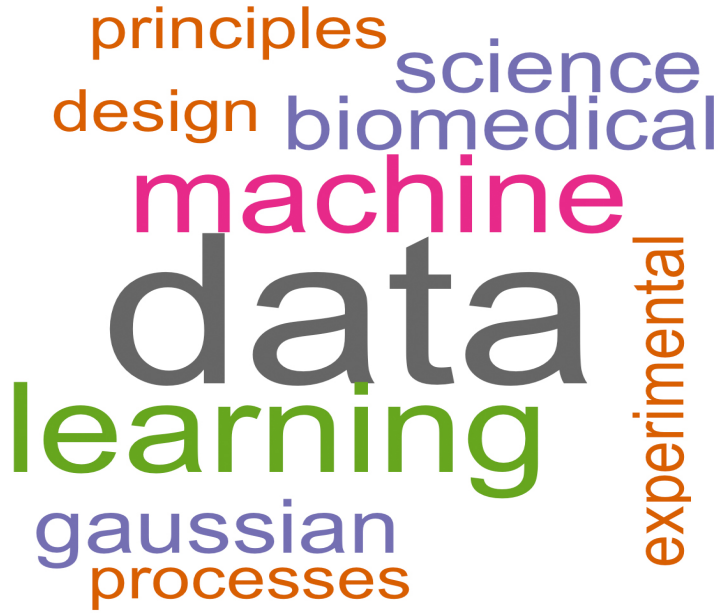
- Introduction: Autumn Data Science School
- Data Science and its principles
- How did the data grow and what are we facing?
- Challenges of ML applications in Data Science
- Making your data ready for the future

Autumn School in Data Science, Cambridge 2019

The University Of Sheffield.

This Autumn School focuses on familiarising life science students/ researchers with principles of **Data Science**.

> ***Data Science*** *is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data..*
>
> *Wikipedia*



Autumn School in Data Science: Roadmap

In this Autumn School we will use Data Science Principles to explore:

how **experimental design** influences choices of models; (through out the programme);

data preparation and characterisation (**Data Readiness**) (Monday- Tuesday: Marta)

machine learning algorithms to handle biomedical data (Tuesday – Wednesday: Adriano);

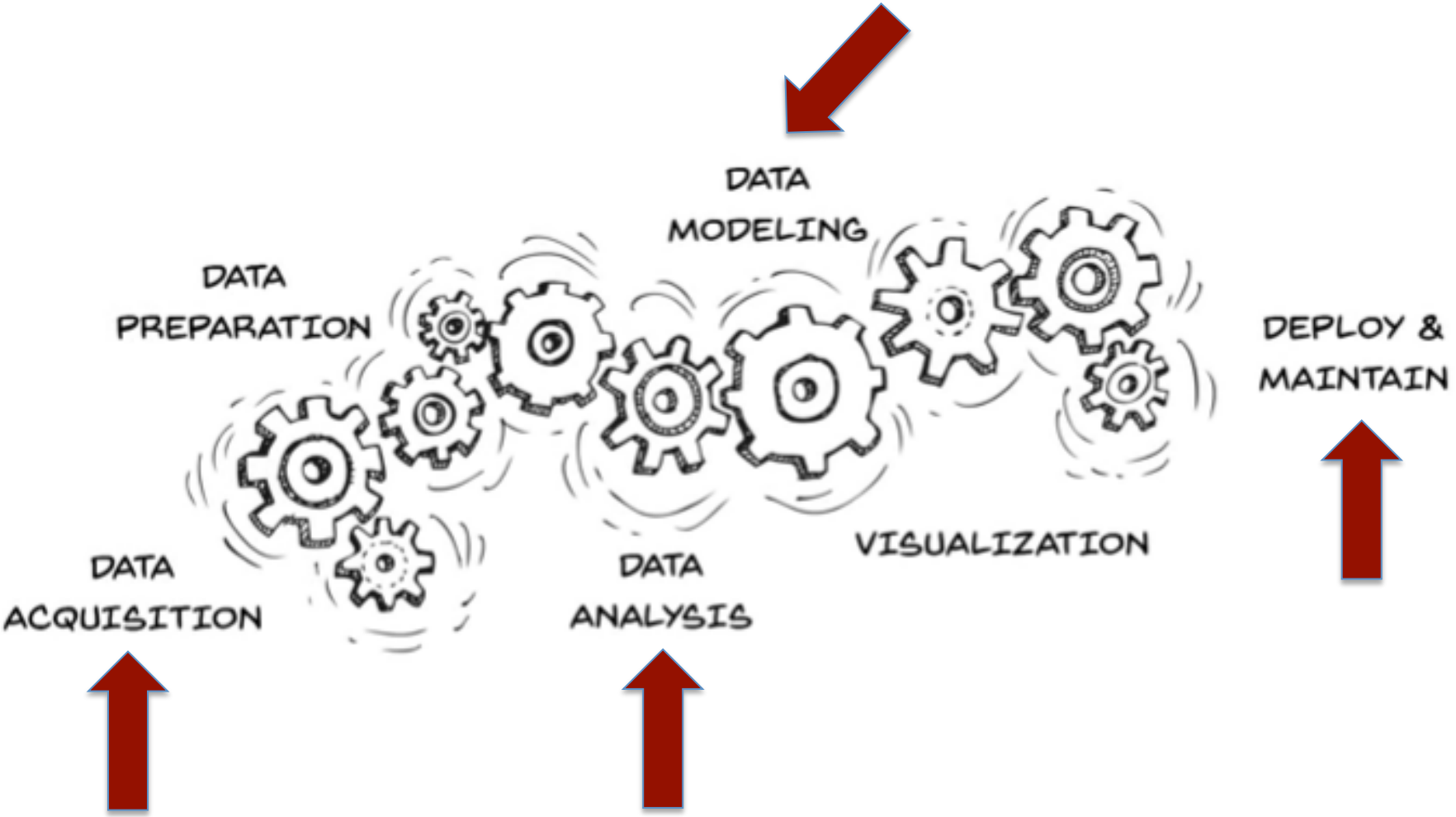
Pipeline implementations in Python and related statistics; (Tuesday-Thursday: Adriano, Alexis, Javier, Catherine)

Application of **Gaussian Process** models (Wednesday – Thursday: Alexix, Javier)

Prof Neil Lawrence on Future of AI in biomedical research (Thursday)

Principles of Data Science

The University Of Sheffield.



In applying these principles in Life Science we encounter a number of bottlenecks at different stages

Once there was the darkness of the microscope room….

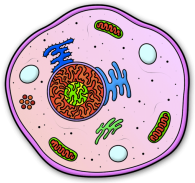
The University Of Sheffield.

Past – HGP
Biological systems were investigated by examining their parts. System-level analysis only theoretically possible

Human Genome Project

HGP – present
High–throughput profiling studies to quantify the activity of every gene**, System-level analysis and predictions**.

\+



\+

Present –Future
"omics" analysis at single-cell resolution.
Data integration with use of AI
Meta genomics and tissue composition at cellular level.

Needs and Challenges we face

The University Of Sheffield.

- 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...

The University Of Sheffield.

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

The University Of Sheffield.

How do we extrapolated the information from the data to train and then test models? (**DATA ACQUISITIONS/ DATA ANALYSIS**)

How do we use domain adaptation in applications to life science? (**DATA MODELLING**)

How do interpret the data and maintain the ability of the models to learn? (**DEPLOY AND MAINTAIN**)

Some examples:

- **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

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)

DATA TRUST becomes a key issue for successful use of advanced ML in Life Science

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 and on a good characterisation of Data structure