UNIVERSITY OF OXFORD

Introduction to Machine Learning with the Python Data Science Stack

Dan Howarth

Who am I?

UNIVERSITY OF OXFORD

SMARTIA

A new dimension in industrial intelligence.

Dan Howarth
Senior Data Scientist
dan.howarth@smartia.tech
www.linkedin.com/in/drhowarth/

What will we cover today?

UNIVERSITY OF OXFORD

Timing	Session
1100 - 1230	• Overview of Module and Assessment • **Overview of Machine Learning** • **Introduction to ML Workflow** • How python data science libraries are used in ML
1230 - 1315	Lunch
1315 - 1445	• **Introduction to Exploratory Data Analysis** • Machine Learning End to End problem: Regression - Section 1
1445 - 1515	Coffee
1515 - 1700	• Machine Learning End to End problem: Regression - Section 2 • **Introduction to Deep Learning** • Learning How to Learn / Code

- Ask questions, be prepared to discuss and debate
- No-one knows all the answers (especially me!) - this is a wide ranging field

Overview of module

UNIVERSITY OF OXFORD

Aims	Assessment	How Assessed?	Due by?
Module 1: Introduction to python data science stack			
• Introduction to data science libraries	• Introduction to pandas • Introduction to numpy • Introduction to matplotlib • Introduction to scikit-learn	Notebooks with additional questions to be completed (**not assessed**). Support via slack	**End Dec** (suggested)
Module 2: Introduction to Machine Learning			
• Introduction to exploratory data analysis and why is is required • Introduction to machine learning principles and how they are implemented	• Readings and test of comprehension on exploratory data analysis and data science workflow • Readings and test of comprehension on machine learning (different types, examples, limits) • Notebook implementation of machine learning algorithms, including training and assessment • Detailed implementation and explanation of one algorithm	Complete notebook (**to follow**) covering an end to end machine learning problem Support via slack	**End Jan** Submit via slack
Module 3: Introduction to Deep Learning			
• Intro. to Deep Learning • Intro. to Computer Vision • Intro. to transfer learning • Intro. to Keras	• Assessment will be via the successful implementation of the notebooks, which have a series of questions in them that need to be answered • Readings to cover deep learning, including reading a research paper, and test of comprehension	Complete notebooks (**to follow**) covering a deep learning problem (computer vision) Support via slack	**End Feb** Submit via slack

Assessment Criteria

UNIVERSITY OF OXFORD

Exploratory Data Analysis	**Machine Learning**	**Deep Learning**
Understands the **principles** of Exploratory Data Analysis	Understands and can articulate the **principles** of a machine learning workflow	Understands the **principles** of a Deep Learning workflow and can articulate the key aspects
Can generate hypotheses based on data and EDA	Understands the steps to implement the machine learning workflow	Understands how deep learning models are developed, in particular CNNs for Computer Vision
Understands the importance of visualisation	Implements the code required for the key aspects of the ML workflow	Understands what transfer learning is
Can analyse data using univariate and bivariate analysis and interpret the results	Understands the difference between model selection and model evaluation and can suggest appropriate methods	Implements the code required for training a deep learning model
Implements the code required for the key aspects of Exploratory Data Analysis	Understands that there are a range of different algorithms available and that some may be more suitable than others	

Overview of Machine Learning

UNIVERSITY OF OXFORD

- Group Discussion:
  - What is machine learning?
  - What are the different types of machine learning?
  - Where can machine learning be applied?

Overview of ML: Definition

UNIVERSITY OF OXFORD

- Machine learning is often categorized as a subfield of artificial intelligence, but...that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of **machine learning as a means of building models of data**.

- Fundamentally, **machine learning involves building mathematical models to help understand data.** "Learning" enters the fray when we give these models tunable parameters that can be adapted to observed data; in this way the program can be considered to be "learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.

'*Python Data Science Handbook*' by Jake VanderPlass

Overview of ML: Different Types

UNIVERSITY OF OXFORD

- *Supervised learning* is about modeling the relationship between features of data and a label associated with the data. Once this model is determined, it can be used to apply labels to new, unknown data. It can either be a *classification* task (where our labels are in discrete categories) or *regression* task (where our labels are continuous quantities).

- *Unsupervised learning* models the features of a dataset without reference to any label. These models include tasks such as clustering and dimensionality reduction. We can also use unsupervised learning techniques to preprocess data prior to running supervised learning models.

- *Deep learning* can be thought of as a way of implementing *supervised* and *unsupervised* and learning using models that have more layers

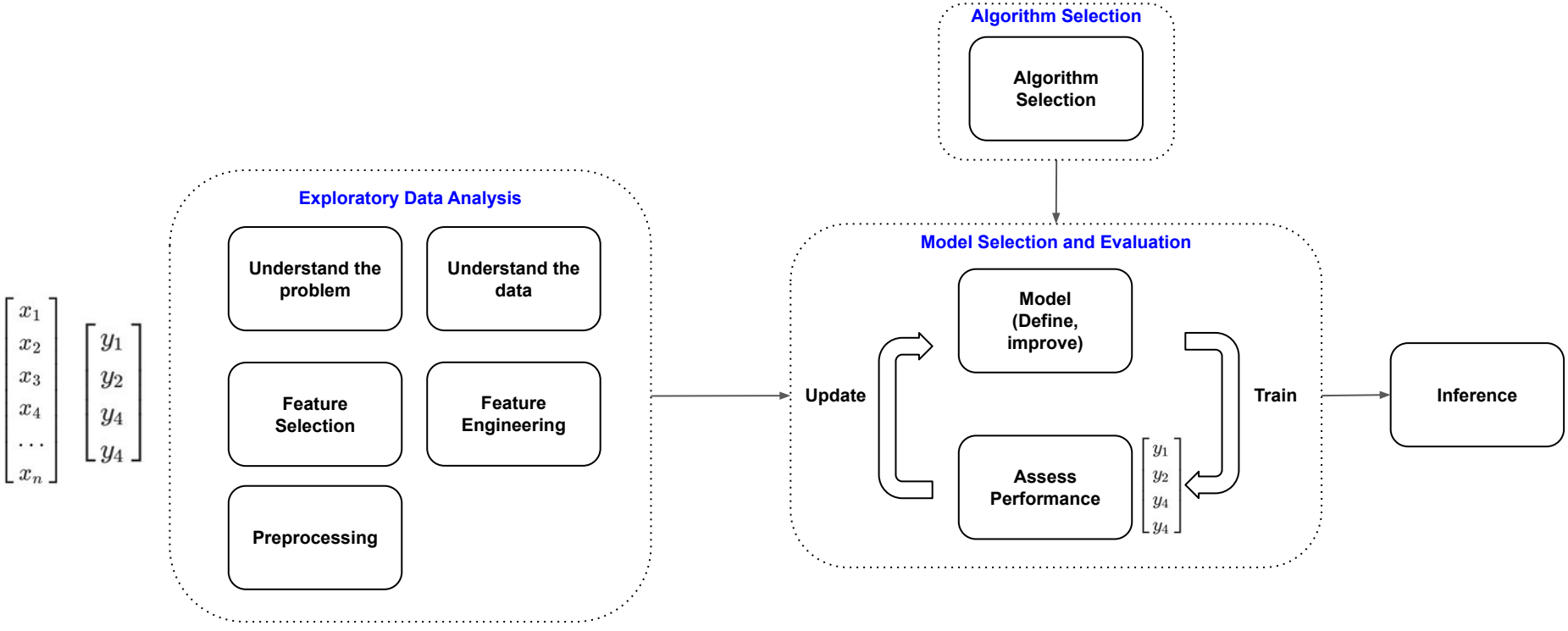
Machine Learning Workflow

UNIVERSITY OF OXFORD

- Group Discussion:
  - What are the key aspects of a machine learning problem?
  - How should we approach solving a machine learning problem?
  - What data science technologies can we use?

Machine Learning Workflow

UNIVERSITY OF OXFORD



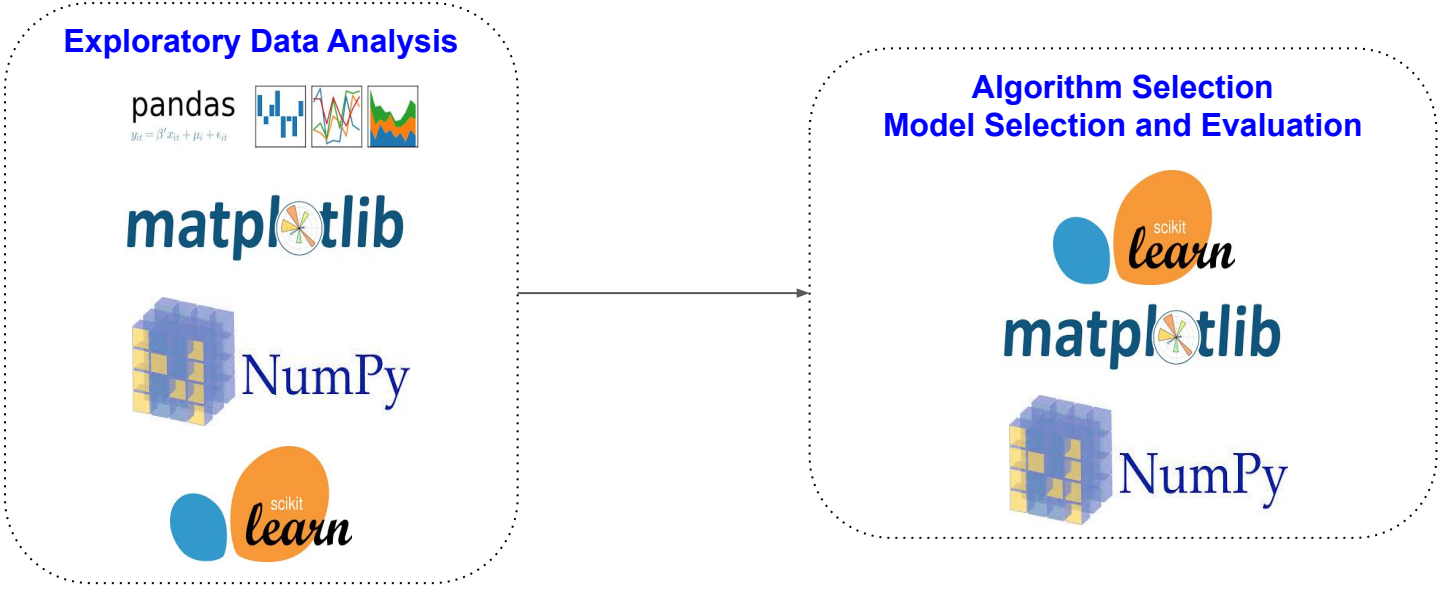
python data science stack

UNIVERSITY OF OXFORD

Library	Description
NumPy	Vectorised arrays, and maths, algebra functionality
pandas	Data structures, data engineering and data analysis tools
matplotlib	Highly customisable plotting functionality
scikit learn	Preprocessing, models, training and inference for ML
TensorFlow	High- and low-level library for deep learning, from Google

python data science stack (2)

UNIVERSITY OF OXFORD



Review / Recap

UNIVERSITY OF OXFORD

- What we will cover in the module
- What Machine Learning Is
- What a Machine Learning Workflow is
- What python technologies we can use for ML

Exploratory Data Analysis

UNIVERSITY OF OXFORD

- Group Discussion:
  - What is Exploratory Data Analysis?
  - What do we need to know about our data before modelling it?
  - What techniques can we use?

Exploratory Data Analysis

UNIVERSITY OF OXFORD

	Initial Data Analysis: *Understand the data quality and properties of data*	**Exploratory Data Analysis:** *Understand what the data is telling us*	**Model Preparation:** *Prepare Data for modelling*
Properties Understand the properties of a dataset	**What are the 'internal' properties of the data?** ● Data types (computational - e.g. strings, integers) ● Data size ● Data shape ● Missing values ● Feature descriptions ● Data provenance	**What are the 'external' properties of the data?** ● Data types (statistical - e.g.continuous, discrete, categorical) ● Unique and Duplicate entries	**What are the properties of the data most relevant to building a model?** ● Class imbalance
Preprocessing Get data in the right shape for the next stage	**How should I treat these 'internal' characteristics?** ● Change data types ● Reshape data **What other changes to the data are required?** ● Drop target variable (if known)	**How should I treat these 'external' characteristics?** ● Normalize or Standardize the data if required for analysis ● Deal with missing values ● Deal with categorical data **What other changes to the data are required?** ● Drop target variable (if known)	***How* to prepare the data for modelling?** ● Normalize and Standardize the data if required for the model ● Deal with class imbalances if necessary (can be done as part of training strategy)
Understanding of data and how it informs the problem we are investigating	**What will help with the next steps?** ● Identify if there is missing data and acquire it ● Develop Data Dictionary and get feature descriptions if required ● Understand provenance and go to the source ● Is this a sample or all the data? ● Data size suggests strategy for processing (e.g. hadoop, etc.)	**Non-computational activities to aid understanding:** ● Generate assumptions about how we are treating the data ● Generate hypotheses for investigation ● Identify new data requirements **Undertake analysis:** ● Univariate ● Bivariate ● Multivariate ● Visualisations	***What* data should be modelled?** ● Feature Selection ● Feature Engineering ● Identify initial ways of modelling data

End to End Regression

UNIVERSITY OF OXFORD

- Implement the steps we have discussed
- Introduction to python and data science libraries
- Introduction to colab environment
- Link here:

  https://github.com/DanRHowarth/Artificial-Intelligence-Cloud-and-Edge-Implementations/blob/master/Oxford_End_to_End_Regression.ipynb

End to End Regression

UNIVERSITY OF OXFORD

- Review
  - What was challenging about the exercise?
  - What did you learn?
  - What do you need to follow-up on?

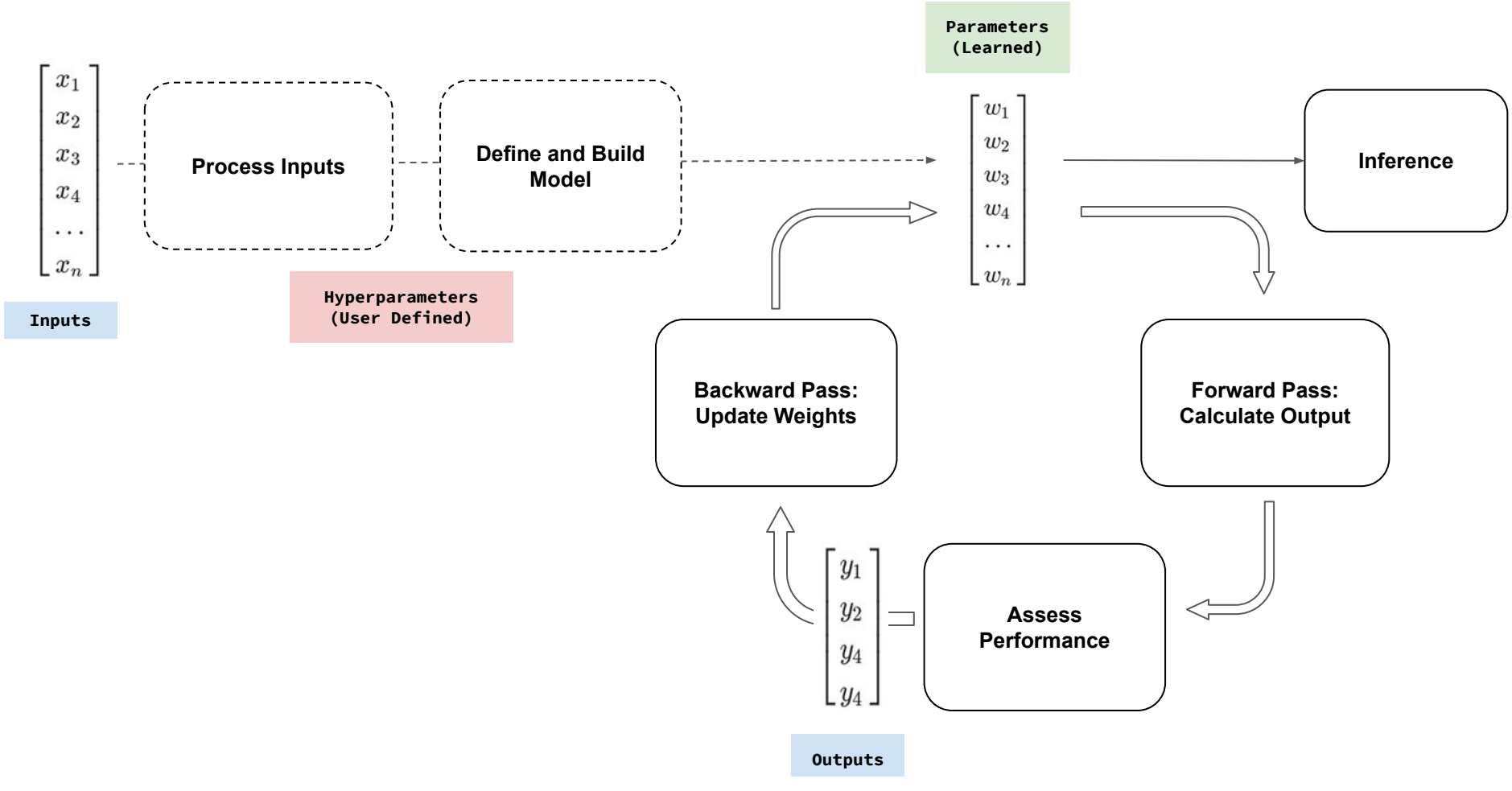
Deep Learning Workflow

UNIVERSITY OF OXFORD

- Group Discussion:
  - What is Deep Learning?
  - What is Deep Learning workflow?
  - How does it differ from machine learning?

Deep Learning Workflow

UNIVERSITY OF OXFORD



Learning How to Learn

UNIVERSITY OF OXFORD

- Understand the ‘Big Picture’ and the Detail -> relate one to the other:
  - For Detailed understanding”
    - ‘Where does this fit with the overall ML framework?’
    - Go ‘one level down’
  - For big picture understanding:
    - Understand the principles that underpin what you are doing
    - Go ‘one level up’ and ‘across’
    - Think about why something exists, and what would happen if ir didn’t
    - Develop a model - it won’t be perfect but it helps join the dots
- Think about learning techniques:
  - Deliberate Practice - be conscious of what you are doing
  - Feynman Technique - explain something to a non-technical person
  - The number of iterations are important - Naval
  - ‘With trial and error, your IQ is a 1000’ - Nassim Taleb

Learning How to Code

UNIVERSITY OF OXFORD

- Type out every line of code - don’t copy and paste
- Use comments to explain what the code is doing
- Think about the requirements first, then break this down into smaller requirements, write the pseudocode - and only then start to code it
- Follow this ‘learning hierarchy’ for a given problem
    - Develop code that works and compare answers;
    - Develop code that doesn’t work and compare answers;
    - Work out the code structure/pseudocode and compare;
    - Look at the solution then implement it ‘by hand’, with comments

Notes of Discussions

UNIVERSITY OF OXFORD

-