Data Science Syllabus

Summary

The ITC Data Science Fellows course consists of **seven core modules**, each of which contains 1-4 weeks of content:

1. **Fundamentals**
2. **Introduction to Data Science**
3. **Core Data Science**
4. **MLOps**
5. **Deep Learning**
6. **Tracks – CV & NLP**
7. **Workshops and Special Topics**

In addition, there will be afternoon **soft skills sessions**, student-led **ITC TED Talks** and office hours with the tech mentors.

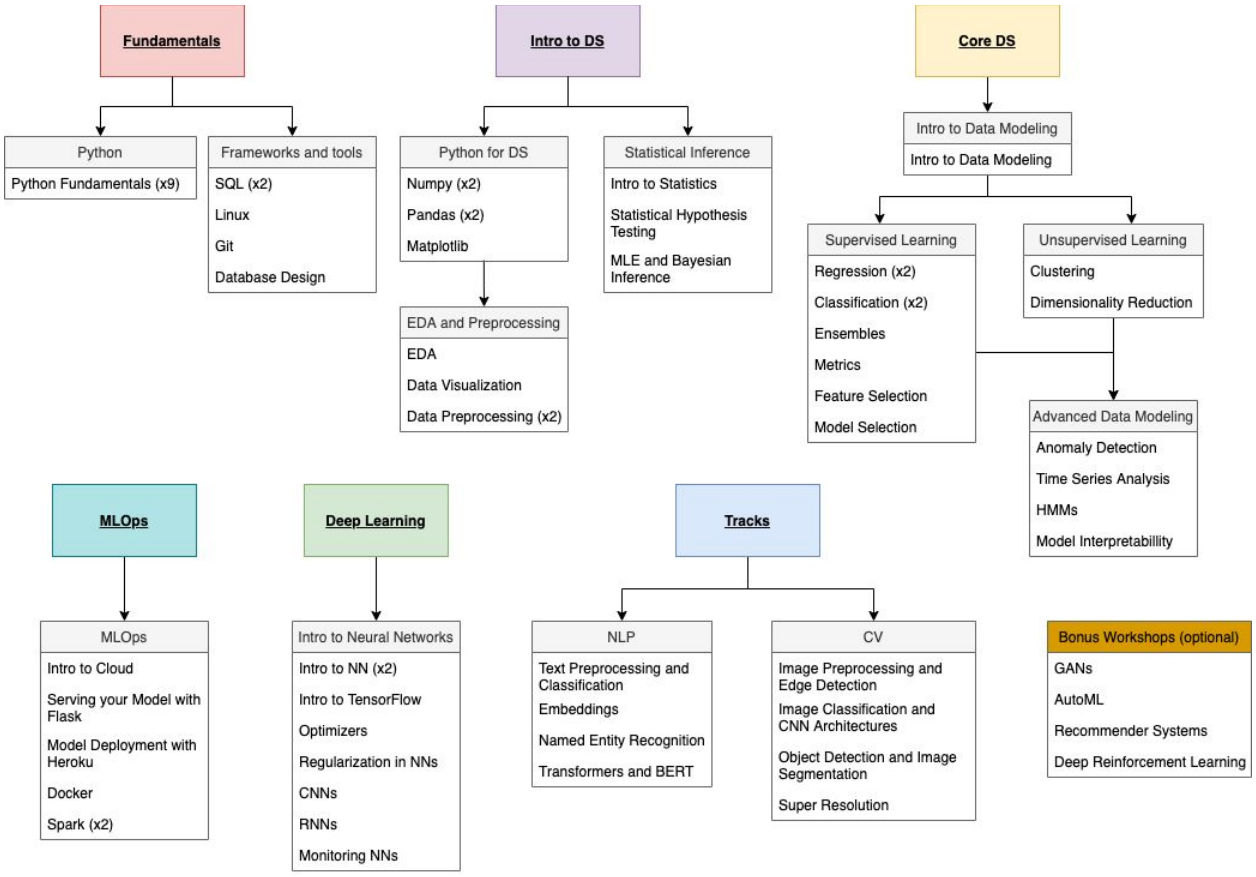
During the course you will complete **two team projects**: a **data mining project** early in the course, and a **final project** which will apply all of the skills you have learned in the course to solve a practical data science problem.

Throughout the course there will be **three exams (two midterms, one final)**, **checkpoint interviews** to monitor your personal progress in the course, and **mock interviews** to prepare you for your job search.

After completing your course studies, you will be hosted by one of our partner companies for a **five-week project** that will use your data science skills to tackle real challenges from the industry. Examples of projects from previous cohorts include the following videos - Aidoc, BMW and can also be found in the ITC project mini-site.

Syllabus Flowchart

The following flowchart gives an overview of the lessons in the course in each module and their order. Note: **These may be subject to minor change per cohort**.



The following sections describe each module in more detail.

Module 1: Fundamentals

The gist: core skills that are fundamental in most DS workflows.

Curriculum:

In the **Python** section, students will review and deepen their **Python coding skills**, from basic building blocks of Python code to advanced concepts like **decorators, complexity and big-O notation, concurrent programming, OOP** and **unit testing**. Students will use both **PyCharm** and its debugging tools as well as **Jupyter notebooks**, and will be introduced to **Anaconda** and **Python virtual environments.**

In the **Frameworks and Tools** section, students will learn about the core tools that are essential to practicing Data Science in the industry, including version control with **Git** and becoming comfortable working in **Linux**. Students will learn how **databases** are structured, and will use **SQL** to query various types of databases.

Module 2: Introduction to Data Science

The gist: In this section, students will learn the fundamental concepts that are required for understanding most work in the field of Data Science and modules and syntax for Data Science coding.

Curriculum:

The **Python for Data Science** classes will introduce students to the use of various Python libraries that are crucial for data science work – **Numpy**, which is used for linear algebra calculations; **Pandas**, used for tabular data and data analysis; and **Matplotlib** which is the most commonly-used Python data visualization library.

The **Statistics** section will teach the core concepts in statistics that are required to understand the workings of machine learning models that will be later in the course. This includes **basic statistical concepts**, **statistical hypothesis testing, maximum likelihood estimation, Bayesian inference** and **MAP estimation.**

The **EDA and Preprocessing** section will cover the fundamentals of **exploratory data analysis (EDA)**, various **data visualization** techniques and **data preprocessing.**

Module 3: Core Data Science

The gist: common models used in machine learning, and how to evaluate and interpret their output

Curriculum:

In the introduction to data modeling lecture and following classes, students will be introduced to important concepts in Machine Learning such as **underfitting** and **overfitting, loss functions,** the **bias-variance tradeoff**, **cross-validation**, and **regularization**.

In the **supervised learning** classes, students will learn about common **classification** and **regression** algorithms, including **Naive Bayes, Logistic Regression, Decision Trees, SVMs**, **linear regression**, and more. Students will learn the inner workings of these methods as well as how to apply them in practice with **scikit-learn**. They will also learn about **ensemble methods** such as **Random Forests** and **XGBoost**, and they will learn which **metrics** are appropriate for evaluating each type of model in given use cases.

In the **unsupervised learning** classes, students will learn algorithms for finding patterns in unlabelled data, including **clustering algorithms** such as **K-Means** and **DBSCAN**, and **dimensionality reduction** algorithms including **PCA**.

In the **advanced data modeling** sessions, students will learn various advanced topics that give more powerful tools for creating machine learning models. They will learn various methods for performing **feature selection** and **model selection**. The section will also cover various types of modeling including **autoregressive time series modeling**, unsupervised **anomaly detection**, and **Hidden Markov Models** and their application to problems such as part-of-speech tagging of text.

Module 4: MLOps

The gist: technologies for data science using big data, deploying and maintaining models in production systems

Curriculum:

In the **MLOps** sessions, students will be exposed to various technologies used for training machine learning models on big data and deploying them for use in production systems.

In the **Intro to Cloud** lecture, students will learn about cloud computing and infrastructure, and how it relates to data science work.

Students will learn about containerization with **Docker** and distributed computing with **Spark**. In the **Serving Your Model** and **Model Deployment** sessions, students will learn how to use **Flask** and **Heroku** to create **REST APIs** for their data science models and serve them as a web application.

Module 5: Deep Learning

The gist: the nuts and bolts of Neural Networks and Deep Learning

Curriculum:

We start with the fundamentals of Deep Learning and Neural Networks, learning how to implement **feed-forward networks** from scratch. We cover basic concepts such as **backpropagation** and **stochastic gradient descent**, **activation functions**, various types of **loss functions**, the main **types of layers** used in neural networks, and **regularization techniques**.

Students will learn how to use **TensorFlow 2** and **Keras** to implement deep neural networks, along with advanced techniques for **optimizing** and **monitoring** their performance, including using tools such as **Tensorboard** to understand how neural networks learn.

In this module, students will also learn the theory behind **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**, including **Long Short-Term Memory Networks (LSTMs)** and **Gated Recurrent Networks (GRUs).** They will also learn how to use these in practice with TensorFlow to solve various problems in computer vision, natural language processing and more. These will serve as the basis for the state-of-the-art Deep Learning algorithms that students will be exposed to in later classes and workshops.

Module 6: Tracks – CV and NLP

The gist: Advanced topics in the fields of Computer Vision and Natural Language Processing

Curriculum:

In the **Computer Vision (CV) and Natural Language Processing (NLP) Tracks**, students will learn classic methods all the way to Deep Learning applications specifically focused on vision and language tasks. This brings students close to the state-of-the-art in these fields, and enables them to understand the main paradigms behind the famous recent papers while also understanding how to apply them in practice.

In the **Computer Vision track**, students will begin by learning **classic CV techniques** for preprocessing and analyzing images, such as, and will learn how to perform **edge detection** using **openCV.** We will then continue to learn deep learning methods for **image classification** using **modern CNN architectures,** and proceed to **object detection with the YOLO algorithm** and **image segmentation with U-Nets.** We will then learn algorithms for **super-resolution,** including **example-based super resolution and super-resolution with CNNs.**

In the **Natural Language Processing track,** students will start by learning how to preprocess textual data, including **tokenization, stemming, lemmatization and POS tagging**. We will learn classic NLP algorithms for **document classification** with **bag-of-words models** and **Latent Dirichlet Allocation**-based topic modeling. This will include using NLP libraries such as **SpaCY, NLTK,** and **Gensim.** We will then move to **neural word embeddings** and explain the theory behind deep learning word embedding architectures such as **word2vec** and contrast them with modern **contextual embedding** methods. We will proceed to apply deep learning to specific NLP tasks such as **Named Entity Recognition** with **BiLSTMs**. Finally we will learn about **Transformer architectures and BERT**, which have become the state-of-the-art for many NLP tasks in recent years, learning how they work and how to use them in practice.

Module 7: Workshops and Special Topics

The gist: workshops with companies are meant to expose the students to additional techniques and get a better sense of real-world challenges. They are a chance to learn more about the Data Science industry in general and about different companies in particular.

In the workshops, companies present an important problem that they face, as well as some techniques that they use to solve the problem. This includes hands-on exercises where the students will work with real data, apply the techniques that they have learned and implement their own ideas.

Examples of past workshops:

- Recommender Systems: Outbrain, Taboola
- Autonomous Vehicles: Mobileye
- Cyber: Deep Instinct, Check Point
- Feature selection and engineering: SparkBeyond
- Computer Vision: DataDen, iCarbonX
- Information Retrieval: Yahoo, Sefaria
- Anomaly Detection: Anodot
- GPUs and DL Solutions: NVIDIA
- Medical Imaging: Zebra medical
- Fin-tech: Istra
- Pre-processing and data modeling: Intel
- Text Classification and Topic Modelling: Chorus