Institute of Data

2020

Data Science and AI

Module 0

Introductions, objectives & overview

2

Institute of Data

Logistics

- Zoom and Slack
- Breaks
- Questions

Institute of Data

Agenda of Module 0

- Introductions
- The Data Scientist role
- Objectives
- Overview of the course
- Hands-on labs and homework

Introductions

- Please share with the class:
  - Current role and background
  - Why you are here?
    - Your **objectives and expectations** of attending the course
  - Your current skill levels in:
    - **Programming**
    - **Mathematics**
    - Other related areas (if applicable to you):
      - Information Management
      - Software Engineering
      - Business domain knowledge
  - Your experience completing the prerequisites

What is data scientist's job

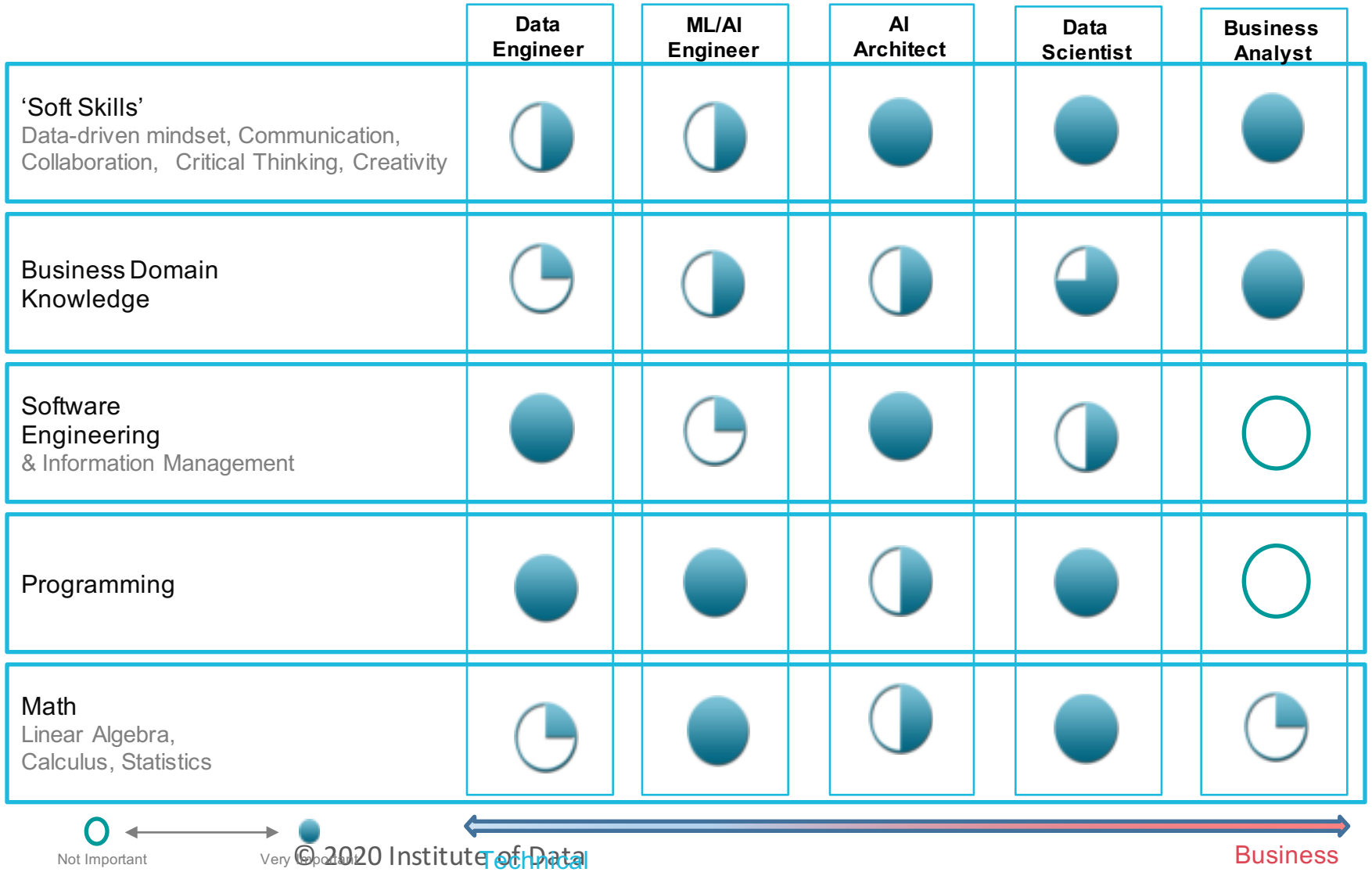
In simple terms, Analysing data for actionable insights.

Specific tasks include:

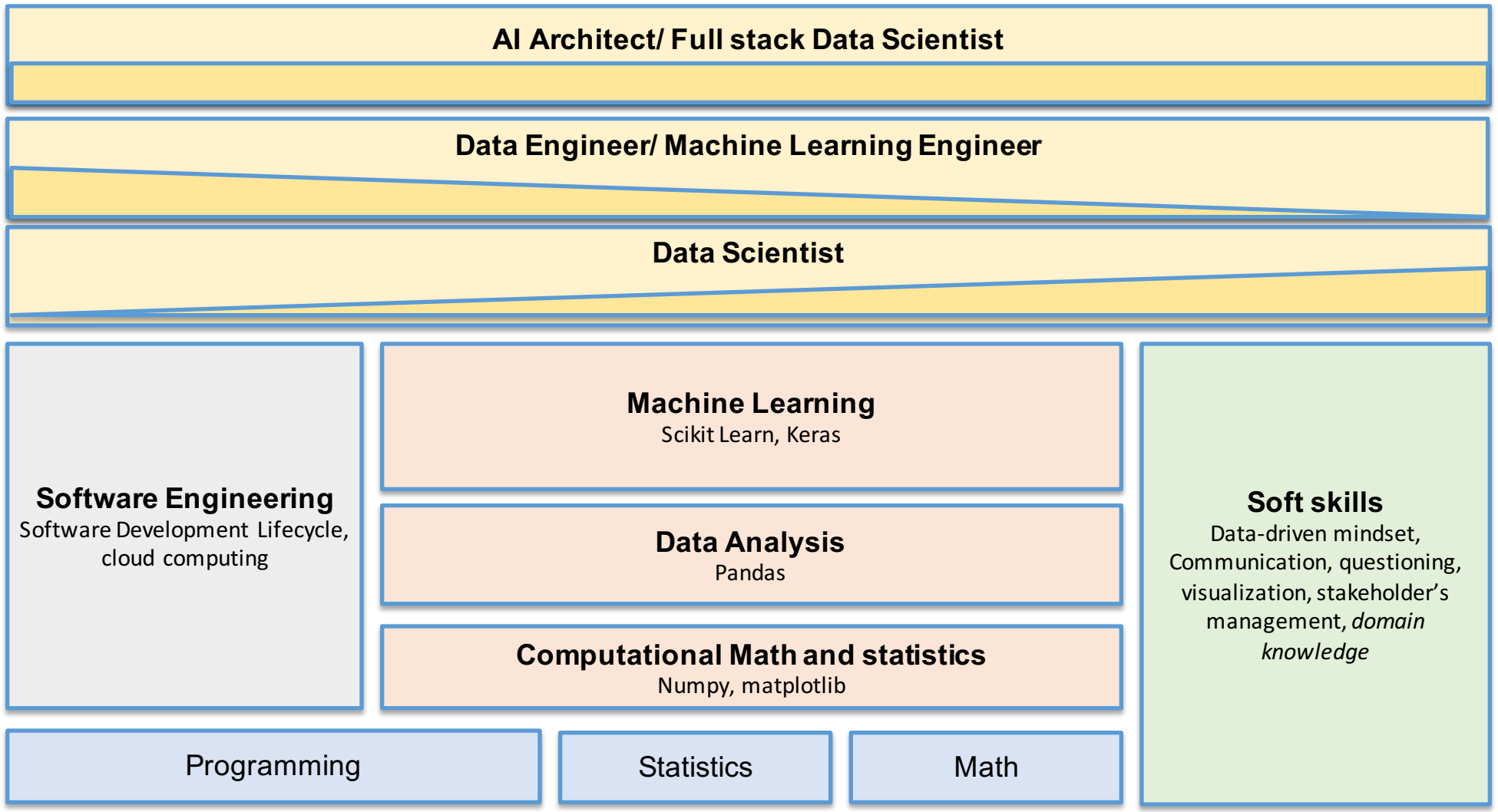
- **Identifying** the **data-analytics problems** that offer the greatest opportunities to the organization
- Determining the **correct data sets** and **variables**
- **Collecting** large sets of structured and unstructured data from disparate sources
- **Cleaning** and validating the data to ensure accuracy, completeness, and uniformity
- **Devising and applying models and algorithms** to mine the stores of big data
- Analyzing the data to **identify patterns and trends**
- **Interpreting the data** to discover solutions and opportunities
- **Communicating** findings to **stakeholders using visualization** and other means

Skills of various roles in Data Science and AI

- There are a number of variations of roles that are required to deliver Data Science/AI projects.
- Some can be considered closer to business while others being more technical.
- There is a growing demand for Data Scientists to be able to contribute directly to systems in 'production'.



Data Science skills for industry



Foundational skills

- Programming for Data Science (Python)
- Maths and Statistics for Data Science

Core data science and AI skills

- Exploratory Data Analysis (EDA) and data wrangling
- Data Visualisation
- Database access
- Application Programming Interfaces (APIs)
- Supervised learning (Regression and Classification)
- Unsupervised learning (Clustering and Dimensionality reduction)
- Deep Learning
- Natural Language Processing (NLP)
- Artificial Intelligence
- Data science industry practices

Applying data science in industry

- Applying data science on different data structures and domains
- Defining a data science project
- Designing a data science project
- Delivering data science project
- Optimising Machine Learning model algorithms
- Overall end-to-end solution
- Presenting to stakeholders and obtaining buy-in
- Capstone project

Soft skills

Consulting, Questioning, Critical Thinking, Problem Solving, Documenting, Presenting

Learning how to learn effectively framework

Minimal Viable Learning (MVL), Multimodal learning, Learn-Create cycle

Approach and principles of the course

This course aims particularly to prepare students to **get a role as a data scientist and perform well in this role in the industry**. This aim shapes the curriculum and the delivery of the course through the following principles:

- Emphasis on **practical skills** for succeeding as a data scientist in the industry
- **Workshop-style, highly interactive** and **collaborative** teaching techniques
- Use of **computational math and statistics** rather theoretical aspects
- Priority on **doing than remembering**
- **Minimal Viable Learning** (MVL) approach

10

Objective of Data Science and AI course

By the end of the Data Science and AI program you will be able to:

Help business to make effective data-driven decisions and track their effectiveness using the appropriate combination of the following tasks:

- Collect, extract, query, clean, and aggregate ***data*** for advanced analytics purposes
- Perform ***statistical and visual analysis*** on data using Python and its libraries and tools
- Build, implement, and evaluate advanced analytics problems using appropriate ***machine learning models*** and algorithms
- Use data visualisation tools *to* ***communicate*** findings
- Create clear **and *reproducible*** reports for stakeholders
- Use *business* ***consulting*** skills and frameworks in data science to assist managers and stakeholders understand the application of AI technology
- Identify ***big data*** problems in businesses and understand how computing technologies are solving these challenges
- Apply ***hypotheses testing, modelling, and validation problem-solving*** processes to datasets from different industries in order to provide insight into real-world problems and solutions

Data Science skills for industry

- **Foundational skills** that are required to learn Data Science:
  - Programming
  - Math, Statistic
  - Basic software engineering
  - Soft skills

Data Science skills for industry

- **Core** Data Science skills
  - Computational math and statistics
  - Data Analysis
  - Machine Learning
- **Complementary** Data Science skills
  - Business domain knowledge
  - Software Engineering
  - Soft skills
    - Questioning
    - Critical Thinking
    - Communication and presentation
    - Problem solving

Programming Data Science in Python

Programming is:

the **process of creating a set of instructions** that tell a computer how to perform a task.
thinking **systematically and critically**
breaking a task into steps. Examples include: a recipe, directions to a destination and mathematical problem solving.

Python has a very **active community** with a vast selection of **libraries**, especially in scientific computing, data analysis and visualisation which makes it **very suitable for Data Science**.

There are a number of tools available to support the development of Python.

Jupyter notebook has emerged as an effective way to develop and share Data Science projects.
Visual Studio Code (VSC) is an alternative for developing reusable software modules.

Programming (**computational mathematics and statistics**) can be crucial for developing deep mathematical and statistical knowledge and skills.

Why statistics is important for a Data Scientist?

- **Statistical Thinking** is an an essential component of a data-driven mindset which is crucial for a Data Scientist
  - Statistical analysis must start with the appropriate **data** (sample)
  - Statistical Inference (reasoning) should start with measurement, ideally, via **controlled experiments**
  - Statistics uses samples (a small subset of the population) and therefore always has a degree of **uncertainty.**
  - Sampling must be **random, and preferably, independent.**

- The best way to learn statistics is by **experimenting with data using Python code and visualisation.**

Hands-on labs and homework

- The course focus on the practical aspects of Data Science to prepare for real-life role.
- You will need around 6 hours/ week for homework
- Programming environment
  - **We will use Google Colaboratory (Colab) for *coding and sharing* Notebooks**
    - Colab is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.
    - With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.
  - **We will use Jupyter Notebook with Anaconda for coding on your own machine**

Questions?

End of Presentation!