Intro to Machine Learning with scikit-learn

QBS 101.5: Applied Data Science

Simon Stone

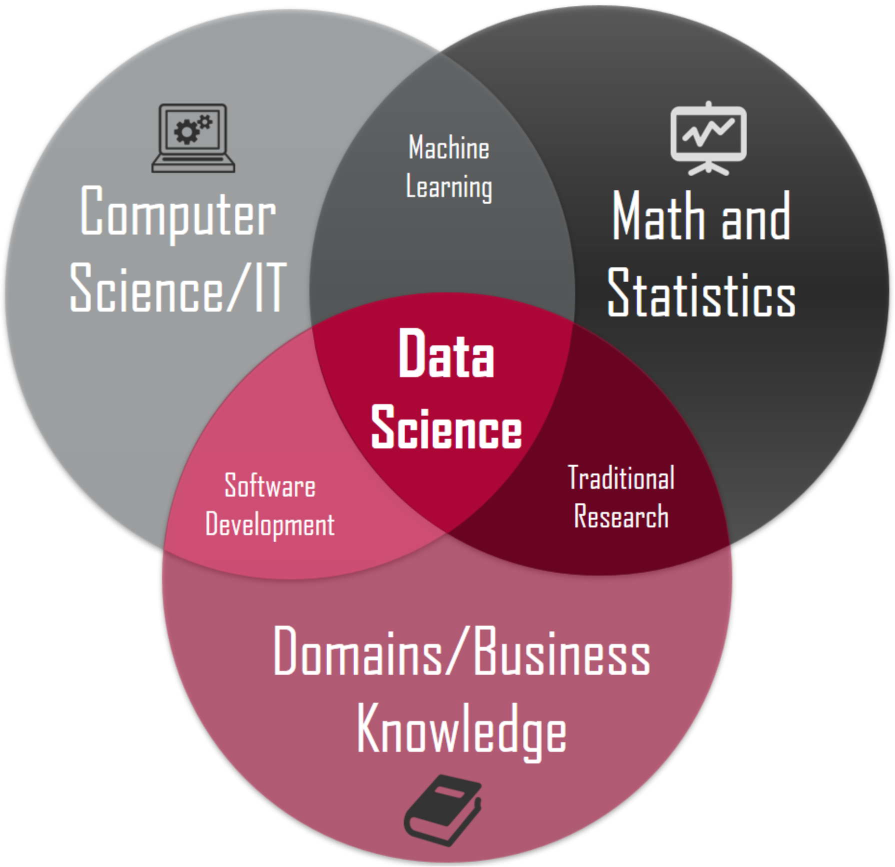
Research Data Services

Dartmouth College Library

Intro

Machine Learning and Data Science

- Data Science is at the intersection of multiple disciplines
- 🛠️ Each discipline brings its own tools and techniques
- 🎲 Statistics lets us describe distributions of observations and make inferences based on the assumed distributions
- 🤖 Machine Learning uses statistical models, but also (optimization) algorithms that can “discover” the best parameters to make inferences based on a given set of observations (learning from data)



Barber, M. (2018). *Venn diagram of disciplines around Data Science.* Data science concepts you need to know! Part 1. Towards Data Science. Retrieved July 17, 2023, from https://towardsdatascience.com/introduction-to-statistics-e9d72d818745.

Machine Learning for Human Teachers

scikit-learn:

💪 Powerful **framework** written with a **Python** frontend

🛠️ Contains a vast number of **popular algorithms** from “classical” machine learning

😕 Only very basic support for **neural networks** (see next session on PyTorch)

📈 Great **evaluation and reporting** functionalities

🧩 Easily **extendable**

Quick poll: “Machine Learning Experience”

Introduction
Data Science

Data Science is OSEMN!*

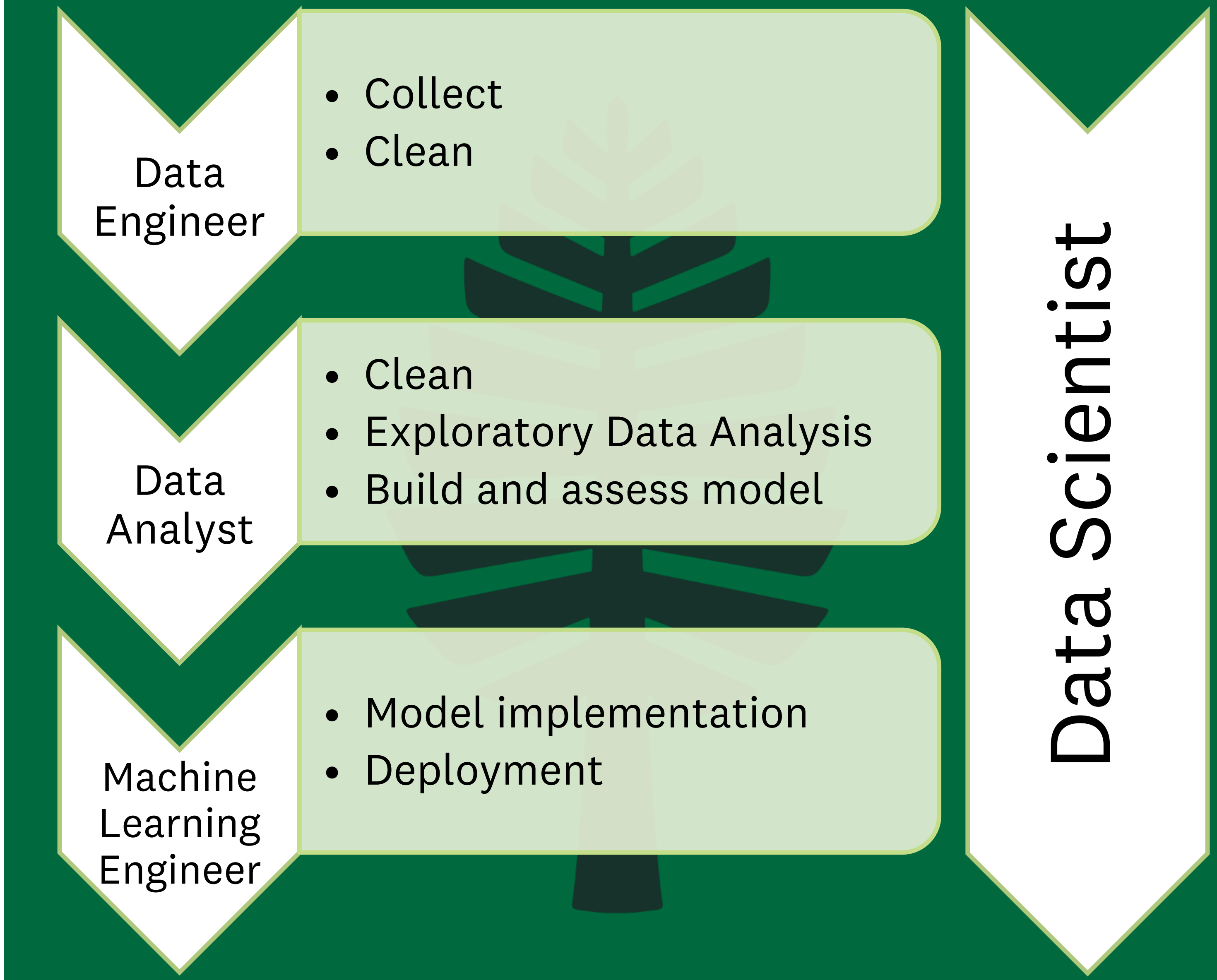
📥 Obtain

🧼 Scrub

🔎 Explore

🤖 Model

⁉️ iNterpret



Data Engineer
- Collect
- Clean

Data Analyst
- Clean
- Exploratory Data Analysis
- Build and assess model

Machine Learning Engineer
- Model implementation
- Deployment

Data Scientist

*pronounced “awesome” - /ˈɔ.səm/
https://www.datascience-pm.com/osemn

Why use a framework for machine learning?

- 😫 Machine Learning is full of **trial & error**
- 🤓 Algorithms can be very computationally and/or conceptually **complex**
- 🔁 **Recurring programming patterns** across projects

- 🧩 Use a framework to harness **efficient implementations** and **modular code design**

What you will learn in this session

- Basic **structure** of scikit-learn
- **Preprocessing**
- **Dimensionality reduction**
- **Training** and **testing** a classifier
- Hyperparameter **tuning**
- **Reporting** results

Creating a complete **pipeline**

What we will work with in this session

- We will use **Python**, a little bit of **Pandas**, and plenty of **scikit-learn**
- As a **programming environment**, we will use a **Jupyter notebook**

scikit-learn at a glance

Project website:

https://scikit-learn.org/

- Installation guide
- Excellent user guide
- API reference
- Examples
- Community
- …

Let’s get started…