Welcome to Data Science at GA!

Instructor:
Dr Noelia Jiménez Martínez

Teaching Assistants:
Hasan Mahmud & Oliver Laslett

Learning Objectives:

-get to know your instructors

-get to know a little bit about each other

-set up the classroom culture

-get to know about the course

-set up our development environments

-learn about the Data Science pipeline

-practise some Python fundamentals

Dr Noelia Jiménez Martínez

-Head of Data Science Unbound.com

-Data Science Consultant (Barclays, JustGiving, M&S, Royal Mail, assorted SMEs, NGOs and start-ups)

-S2DS Mentor (2016/2017)

Skills:

• Python libraries for ML and data visualization: numpy, pandas, matplotlib, sklearn, ggplot. SQL • C, C++,Fortran, IDL • Version control, Git, GitHub • AWS, GCP, Azure • Linux/Windows • Web scraping, APIs • Data analysis and visualisation • Machine Learning algorithms, parameter optimisation Clustering, Classification, Regression, NLP, Time-series forecasting,Graph Databases, CNN, RNN. • Agile product management. • Statistics, Bayesian modelling, Gaussian Processes.

Academic training:

-Master in Information Theory (MaxEnt)

-PhD in Numerical Astrophysicist

-Postdocs (Buenos Aires, Trieste/Barcelona, St Andrews)

Dr Oliver Laslett

Current job: Freelance Data Scientist
Education: PhD in Theoretical Physics, MEng Engineering
Research Interest: MLOps / infrastructure. Human-in-the-loop and active learning.

Technical skills
Python, javascript, go, C, C++.
Machine learning, data vizualisation, data engineering.
Bayesian statistics, stochastic processes, risk modelling.
Kubernetes, docker, AWS, GCP, Terraform, CICD, serverless, flask, django, RESTful design.

Hasan Mahmud

Current Job:
Data Analytics & Management (KTP Associate)
Research, Innovation, and Enterprise | University of East London
Education: BSc Computer Science, MSc Information Systems, MBA
Research Interest: Learning Analytics, Machine Learning, Human-Computer Interaction
Technical Skills:
Programming: Python, PHP
Frontend: HTML (5), CSS (3)
Database: MySQL
CMS: WordPress, CS-Cart, Open Cart

Turn to your partner:

-introduce yourselves,
-tell each other your motivation to do this course

Report back to the class

To discuss in groups

What is the best classroom experience you ever had?

What made it great?

What are your expectations from us/from each other?

What do you think we expect from you?

Road to Success

- The emotional cycle of change: This course is fast and covers a lot of material.

- There will be times when you may feel discouraged or overwhelmed, but don't give up - this is natural (and part of the design). By the end of the course, you'll feel more confident in your ability to define problems, analyze data, and prototype solutions.

- Student learning responsibility: Our lessons cover topic foundations, but there is always more to learn! You are responsible for your learning experience - but don't get overwhelmed! Instead, just make sure you follow along, practice as much as possible, and ask questions.

- **GA requirements: Show up. Be on time. Participate. Submit your projects. Allow yourself to struggle. Read the docs. Have fun!**

The Pit of Success

The Course

Lessons

Curriculum Structure

General Assembly's Data Science part time materials are organized into four units.

Unit	Title	Topics Covered	Length
Unit 1	Data Foundations	Python Syntax, Development Environment	Lessons 1-4
Unit 2	Working with Data	Stats Review, Visualization, & EDA	Lessons 5-9
Unit 3	Data Science Modeling	Regression, Classification, & KNN	Lessons 10-14
Unit 4	Data Science Applications	Decision Trees, NLP, & Flex Topics	Lessons 15-19

Set up Python

Anaconda is there yet?

GIT


slack

Office Hours

Socials?

DS

Data Science
Units Continued

1 PROGRAMMING BASICS

WHAT IS DATA SCIENCE

- Define the workflow, tools and approaches data scientists use to analyze data
- Apply the data science workflow to solve a task

YOUR DEVELOPMENT ENVIRONMENT

- Navigate through directories using the command line
- Use git and GitHub to share repositories

PYTHON FOUNDATIONS

- Conduct arithmetic and string operations in Python
- Assign variables
- Implement loops and conditional statements
- Use Python to clean and edit datasets

GA

What is Data Science?

Activity: Data Science in the Real World

- List five products or services that you think utilize data science.

Activity: Data Science in the Real World

- List five products or services that you think utilize data science.
- **Examples**
  - Providing movie recommendations on Netflix.
  - Making product suggestions on Amazon.
  - Predicting customer 'churn' in a superannuation fund
  - Developing self-driving cars
  - Trading stocks on the stock-market at high frequency
  - Returning auto-translate and search results on Google.

How to Ask Good Question

Like scientists! Use a scientific method.

- Most practitioners apply a version of the scientific method in order to logically deconstruct and analyze an issue.
- At General Assembly, we call this the **data science workflow.**

The Data Science Pipeline

Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualize, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

- This problem-solving framework will help you produce results that are **reliable** (so that your findings will be more accurate), and **reproducible** (so that others can follow your steps and achieve the same results).

- Note that, depending on the problem, this process is **not always linear.**

- You may require lots of **iteration and repetition** before any conclusions can be drawn!

SMART Framework:

- **Specific**: The data set and key variables are clearly defined.
- **Measurable**: The type of analysis and major assumptions are articulated.
- **Attainable**: The question you are asking is feasible for your data set and not likely to be biased.
- **Reproducible**: Another person (or future you) can read and understand exactly how your analysis is performed.
- **Time-bound**: You clearly state the time period and population to which this analysis pertains.

Machine learning common questions:

- Does X predict Y? (Where X is a set of data and y is an outcome.)
- Are there any distinct groups in our data?
- What are the key components of our data?
- Is one of our observations “weird”?

From a business perspective, we can ask:

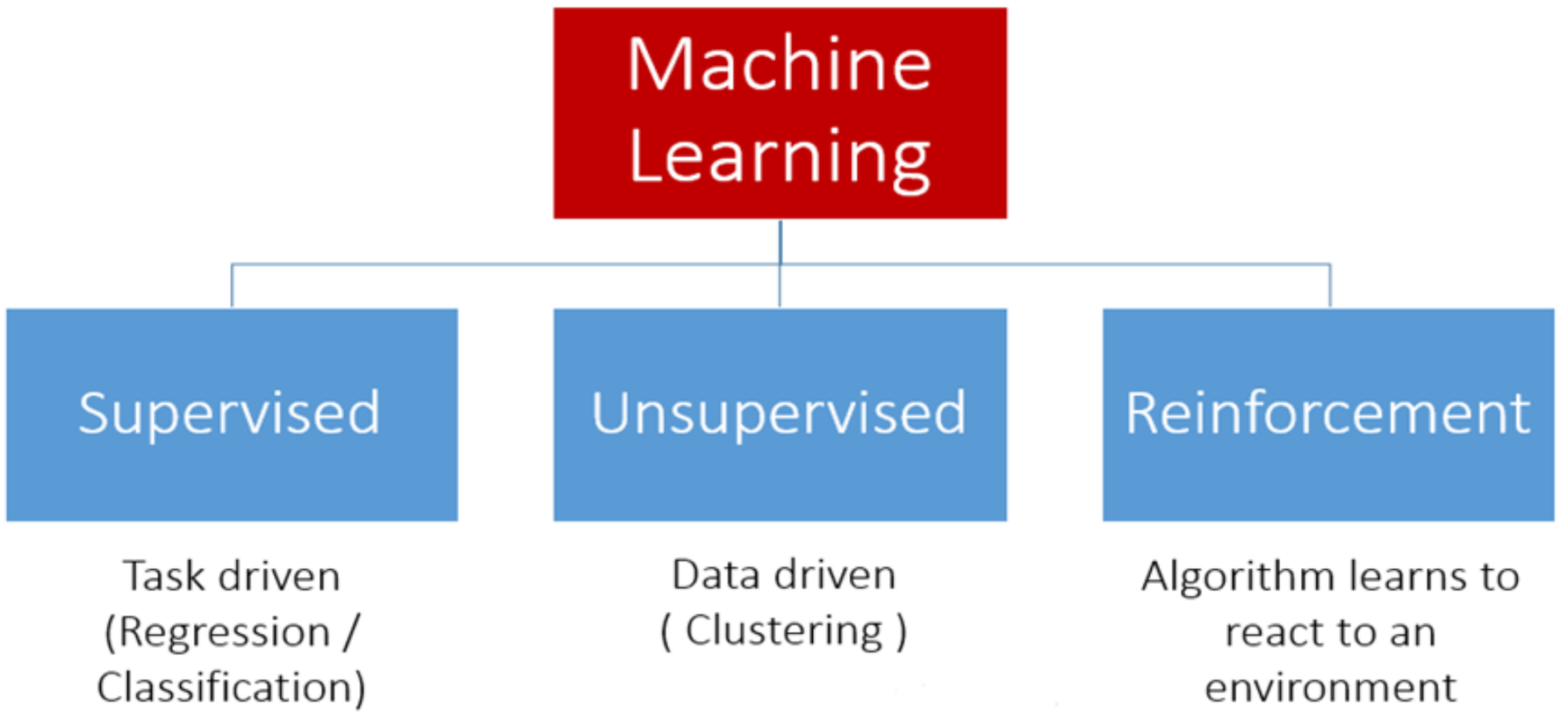
- What is the likelihood that a customer will buy this product?
- Is this a good or bad review?
- How much demand will there be for my service tomorrow?
- Is this the cheapest way to deliver my goods?
- Is there a better way to segment my marketing strategies?
- What groups of products are customers purchasing together?
- Can we automate this simple yes/no decision?

With a partner, rephrase these

data science questions to fit the SMART framework:

- How's Google's share price doing?
- Is the weather getting hotter?
- Is the Prime Minister popular?

Types of Machine Learning



Supervised Learning

Predictive modelling.

Learn the relationship between some inputs (x) and an output (y) based on a 'training' set of data points.

For new, 'unseen' data points, the aim is to accurately predict y based on x.

We want our model to generalise.

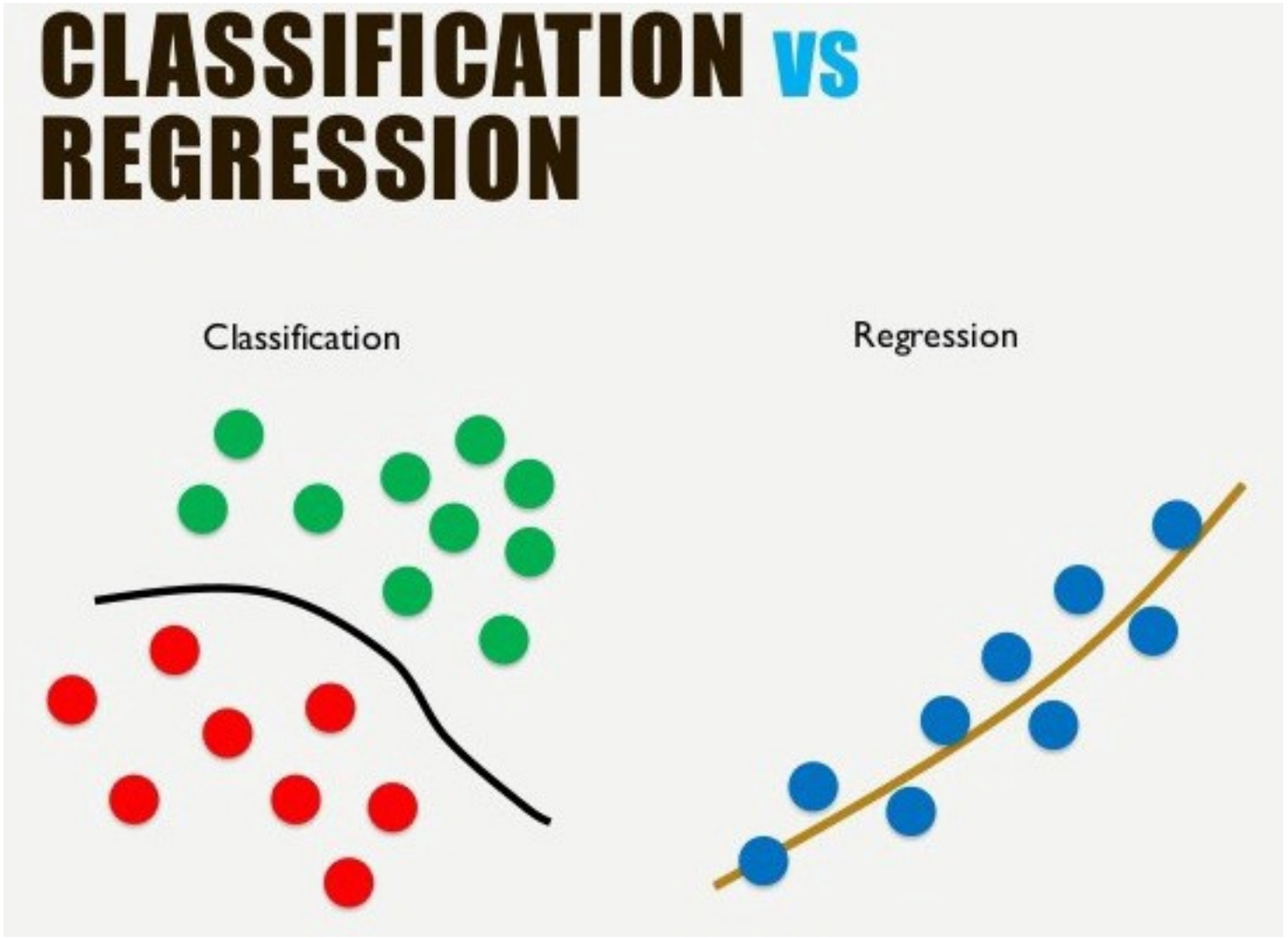
Supervised learning Training set consists of features (x) and labels (y)

$$\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$$

The learning algorithm finds a function (g) that maps features to labels

$$g: X \rightarrow Y$$

Classification tasks involve categorical outputs.
Regression tasks involve continuous outputs.



Unsupervised Learning

Clustering and dimensionality reduction.

Extract patterns and groupings from data.

We don’t necessarily know what we’re looking for…

Machine Learning in the real world

Supervised Learning: fraud detection, lifetime value model, price optimisation

Unsupervised Learning: Anomaly detection (e.g. unusual objects in astronomy)

Deep Learning: autonomous cars, passport gates in the airport

RECAP

Python fundamentals