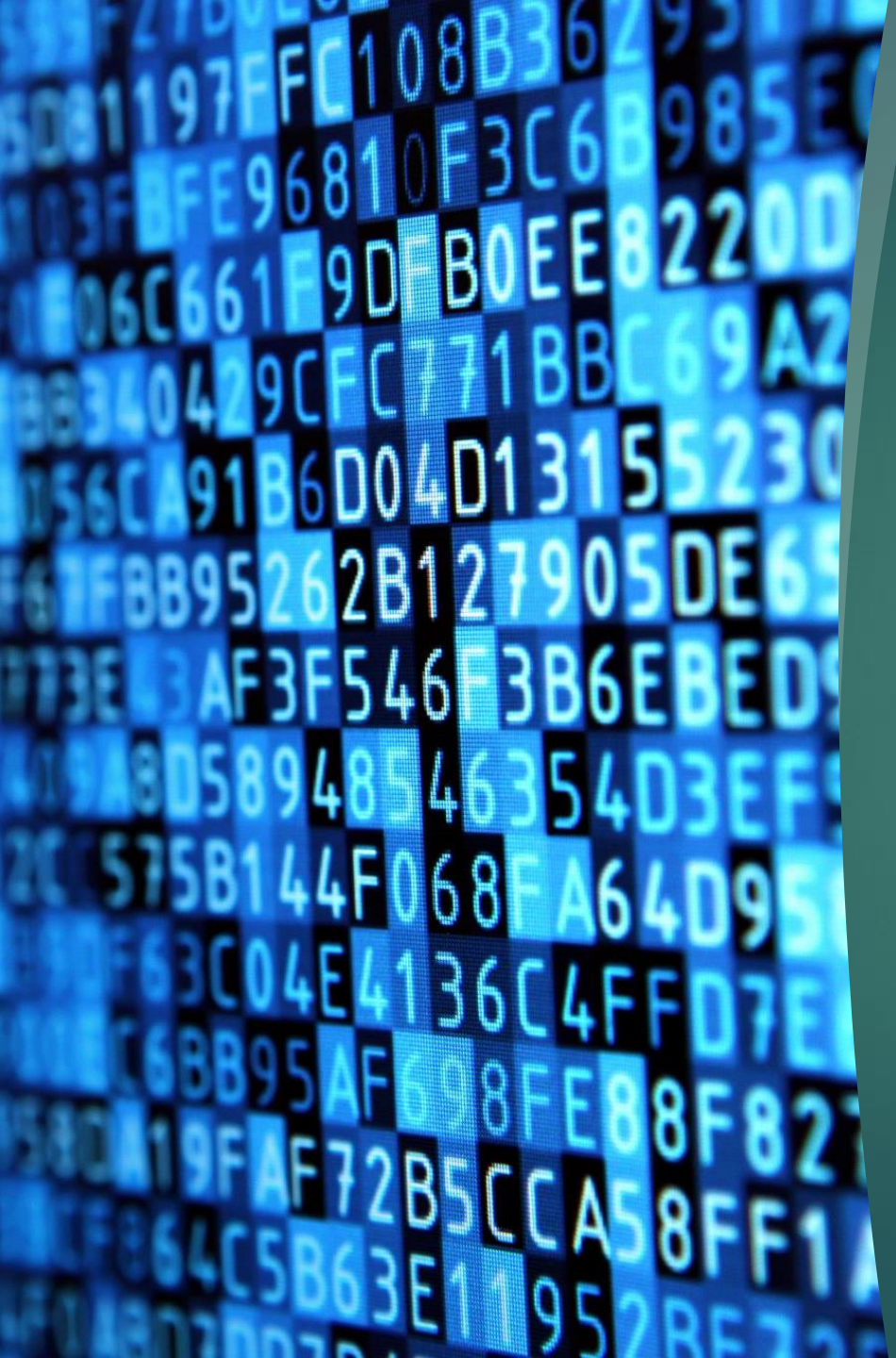

Introduction to Data Science and Machine Learning

COMP-1702

Module 3: Tools & Technologies

Module 3: Tools & Technologies

Learning Outcomes

- By the end of this module, you should be able to:
  - Identify tools and technologies used to perform data science and machine learning operations.
- What are you going to learn in this module?
  - Mainstream tools & technologies used in DSML.
- Why are you going to learn this?
  - To become familiar with some of the tools you may encounter in the industry.

Module 3: Tools & Technologies

Python

- Python is a multi-purpose language with a beginner-friendly and straightforward syntax.
- It's one of the fastest-growing and most popular programming languages used for scientific computing.
- Even people from different backgrounds who are not software engineers, but mathematicians, statisticians, medical students, accountants, and people from other disciplines use Python for various tasks.
- Can be used to build AI applications, perform automation, build web, mobile and desktop applications, perform data analysis, perform software testing, ethical hacking and more.
- https://www.python.org/
- https://en.wikipedia.org/wiki/Python_(programming_language)

https://neuraspike.com/blog/why-is-python-popular-in-data-science/

Module 3: Tools & Technologies

Jupyter Notebook

- A Jupyter Notebook is a powerful tool for interactively developing and presenting Data Science projects.
- Jupyter Notebooks integrate your code and its output into a single document.
- That document will contain the text, mathematical equations, and visualisations that the code produces directly in the same page.
- This step-by-step workflow promotes fast, iterative development since each output of your code will be displayed right away.
- https://jupyter.org/
- https://en.wikipedia.org/wiki/Project_Jupyter#Jupyter_Notebook

https://towardsdatascience.com/a-beginners-tutorial-to-jupyter-notebooks-1b2f8705888a

Module 3: Tools & Technologies

Anaconda Distribution

- Anaconda is the data science platform for data scientists, IT professionals and business leaders of tomorrow.
- It is a distribution of Python & R.
- With more than 300 packages for data science, it becomes one of the best platforms for any project.
- With over 25 million users worldwide, the open-source individual edition is the easiest way to perform Python/R data science and machine learning on a single machine.
- https://www.anaconda.com/products/individual
- https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)

https://www.edureka.co/blog/python-anaconda-tutorial/

Module 3: Tools & Technologies

R

- R is a language and environment for statistical computing and graphics.
- Provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering...) and graphical techniques, and is highly extensible.
- One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.
- https://www.r-project.org/
- https://en.wikipedia.org/wiki/R_(programming_language)

https://www.tutorialspoint.com/r/index.htm

Module 3: Tools & Technologies

NumPy

- NumPy is the fundamental package for scientific computing in Python.
- It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- https://numpy.org/
- https://en.wikipedia.org/wiki/NumPy

https://betterprogramming.pub/numpy-illustrated-the-visual-guide-to-numpy-3b1d4976de1d

Module 3: Tools & Technologies

Pandas

- Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- Aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.
- Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in ANY language.
- https://pandas.pydata.org/
- https://en.wikipedia.org/wiki/Pandas_(software)

https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

Module 3: Tools & Technologies

Matplotlib

- Library for making 2D plots of arrays in Python.
- Has its origins in emulating the MATLAB graphics commands, it is independent of MATLAB, and can be used in a Pythonic, object oriented way.
- Although Matplotlib is written primarily in pure Python, it makes heavy use of NumPy and other extension code to provide good performance even for large arrays.
- https://matplotlib.org/
- https://en.wikipedia.org/wiki/Matplotlib

https://realpython.com/python-matplotlib-guide/

Module 3: Tools & Technologies

Seaborn

- Seaborn is a Python data visualisation library based on Matplotlib.
- It provides a high-level interface for drawing attractive and informative statistical graphics.
- Build on top of Matplotlib and Integrates closely with Pandas data structures.
- Helps you explore and understand your data.
- Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.
- Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.
- https://seaborn.pydata.org/

https://seaborn.pydata.org/introduction.html

Module 3: Tools & Technologies

Scikit-Learn

- Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning.
- It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities,
- Simple and efficient tools for predictive data analysis.
- Accessible to everybody, and reusable in various contexts.
- Built on NumPy, SciPy and Matplotlib.
- Open source, commercially usable BSD license.
- https://scikit-learn.org/stable/
- https://en.wikipedia.org/wiki/Scikit-learn

https://www.dataquest.io/blog/sci-kit-learn-tutorial/

Module 3: Tools & Technologies

SciPy

- SciPy is a Python-based ecosystem of open source software for mathematics, science and engineering.
- Builds on a small core of packages: Python, NumPy, SciPy library, Matplotlib.
- On this base, the SciPy ecosystem includes general and specialised tools for data management and computation, productive experimentation, and high performance computing.
- https://www.scipy.org/
- https://en.wikipedia.org/wiki/SciPy

https://phoenixnap.com/kb/scipy-tutorial

Module 3: Tools & Technologies

TensorFlow

- TensorFlow is a free and open source software library for machine learning.
- It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.
- It's a symbolic math library based on dataflow and differentiable programming.
- Used for both research and production at Google.
- Developed by the Google Brain team for internal Google use.
- Released under the Apache License 2.0 in 2015.
- https://www.tensorflow.org/
- https://en.wikipedia.org/wiki/TensorFlow

https://www.datacamp.com/community/tutorials/tensorflow-tutorial

Module 3: Tools & Technologies

Keras

- Keras is an open source software library that provides a Python interface for artificial neural networks.
- Acts as an interface for the TensorFlow library.
- Contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code.
- Support for convolutional and recurrent neural networks.
- https://keras.io/
- https://en.wikipedia.org/wiki/Keras

https://elitedatascience.com/keras-tutorial-deep-learning-in-python

Module 3: Tools & Technologies

PyTorch

- PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing.
- Primarily developed by Facebook's AI research lab.
- A number of pieces of deep learning software are built on top of PyTorch, including Tesla Autopilot and Uber's Pyro.
- Provides two high-level features:
  - Tensor computing (like NumPy) with strong acceleration via graphics processing units (GPU).
  - Deep neural networks built on a type-based automatic differentiation system.
- https://pytorch.org/
- https://en.wikipedia.org/wiki/PyTorch

https://www.guru99.com/pytorch-tutorial.html

Module 3: Tools & Technologies

Amazon AWS

- Amazon Web Services (AWS) is a subsidiary of Amazon providing on-demand cloud computing platforms and API's to individuals, companies, and governments, on a metered pay-as-you-go basis.
- Provides a variety of basic abstract technical infrastructure and distributed computing building blocks and tools.
- The AWS technology is implemented at server farms throughout the world.
- https://aws.amazon.com/
- https://en.wikipedia.org/wiki/Amazon_Web_Services
- Take 10 minutes to watch the video (right side of slide), then do some quick research on one of the AWS services that interested you, so you can give the class a short description of what it does.

https://youtu.be/a9__D53WsUs

Module 3: Tools & Technologies

Microsoft Azure

- Commonly referred to as Azure, a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers.
- It provides software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) and supports many different programming languages, tools, and frameworks, including both Microsoft-specific and third party software and systems.
- https://azure.microsoft.com/en-us/
- https://en.wikipedia.org/wiki/Microsoft_Azure
- Take 10 minutes to watch the video (right side of slide), then do some quick research on one of the Azure services that interested you, so you can give the class a short description of what it does.

https://youtu.be/eZLcyTxi8ZI

Module 3: Tools & Technologies

Assignment

- For this module, there will be a short quiz to test your understanding of the tools we just covered.
- You can find the quiz in Module 3 in Learn.