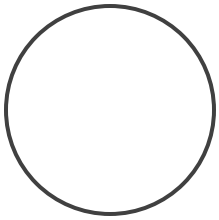
About the module

Programming for Data Analysis

In this **10** credit module, students develop their programming skills towards the effective use of data analysis libraries and software. Students learn how to select efficient data structures for numerical programming, and to use these data structures to transform data into useful and actionable information.

Lecturer: ian.mcloughlin@gmit.ie

Introductory video



https://web.microsoftstream.com/video/90e4ada5-d9c8-4044-a1da-be906a2c3372

Learning outcomes

On completion of this module the learner will/should be able to

1. Perform exploratory analysis on data.
2. Programmatically create plots and other visual outputs from data.
3. Design computer algorithms to solve numerical problems.
4. Create software that incorporates and utilises standard numerical libraries.
5. Employ appropriate data structures when programming for data-intensive applications.
6. Model real-world, data-intensive problems as computing problems.

Indicative syllabus

The following is a list of topics that will likely be covered in this module. Note that the topics might not be presented in this order and might not be explicitly referenced in course materials.

Data
Two-dimensional arrays, matrices, data frames, time series data structures, dictionaries, sets, vectors, slicing, indexing

Programming
Reshaping data structures, unzipping arrays, slicing, calculating descriptive statistics.

Analytics
Exploratory data analysis, scatterplots, histograms, boxplots, principal component analysis.

Assessment

The assessment of this module is broken down as follows.

- Practical assignments (50%)
- Project (50%)

Forums

Announcements

Lecturers will post announcements in this forum but students can't post.

Discussion

You can use this forum to ask questions and start discussions.

Python review

This week we will do a review of the main control structures you will need to know in Python.

Software installation video

https://web.microsoftstream.com/video/12323e90-28a4-4773-88a5-df4a20b9f703

Details of the software you will need for this module. You may already have this software installed.

Setting up a repository video

https://web.microsoftstream.com/video/c0267d8d-77be-47f2-b09d-83923a5f5af0

A review of how to get set up with a repository for starting a new software project.

If statements video

https://web.microsoftstream.com/video/c79619f9-206e-4a97-b4d1-20fd30e40025

A review of `if` statements in Python.

While loops video

https://web.microsoftstream.com/video/61a8b71f-9860-49a6-b657-f274303a3c36

A review of `while` loops in Python.

For loops video

https://web.microsoftstream.com/video/8492c53c-a684-4da9-a2c5-bce1d5c367a9

A review of `for` loops in Python.

Functions video

https://web.microsoftstream.com/video/43ffe3ec-7b08-41dc-b080-2031f45b5c2e

A review of functions in Python.

Plotting basics (matplotlib)

This week we will look at the pyplot plotting library for Python.

Introduction to matplotlib video

https://web.microsoftstream.com/video/07bfc5b8-ab4c-468c-9eb0-b35ca9afeb63

An introduction to matplotlib with pyplot.

Simple plotting video

https://web.microsoftstream.com/video/204cd80b-2054-4587-9a9a-33020e84ba56

Creating a simple plot with pyplot.

Two plots on one set of axes video

https://web.microsoftstream.com/video/7a57ee2c-df42-48d3-882b-cdcabe97c39b

Plotting two plots on one set of axes.

Titles, legends and labels video

https://web.microsoftstream.com/video/c9883d44-2104-44a5-be96-7419e7de993b

Adding extras to plots.

Histograms video

https://web.microsoftstream.com/video/b88a8cbb-2dbd-4cc2-835c-46101baed276

Creating histograms in pyplot.

Side-by-side plots video

https://web.microsoftstream.com/video/ed13c5bf-978f-4d8b-8ee6-c29ac054038f

Creating two plots side by side.

Other plots video

https://web.microsoftstream.com/video/8f09d1fe-9b94-4908-b71e-94ed7b119174

Having fun with pyplot.

---

pyplot examples repository

https://github.com/ianmcloughlin/pyplot-examples

Repository containing example pyplot plots.

Browser workflows (jupyter)

This week we will look at the jupyter package for creating visual workflows to tell a

data analytics story.

Starting jupyter video

https://web.microsoftstream.com/video/28a4f39c-83dc-4e5d-afbf-6e56551a88be

An introduction to getting jupyter running on your computer.

Renaming notebooks video

https://web.microsoftstream.com/video/8752c80e-6175-4799-9434-b8e566d3fa5e

How to rename a jupyter notebook and where they live on your computer.

Cells in jupyter

https://web.microsoftstream.com/video/9c4d56bd-0b5f-4019-b2e6-6185e97154c0

What cells are and how they work in jupyter.

Jupyter keyboard shortcuts video

https://web.microsoftstream.com/video/f542790d-570d-44e3-a6a9-3108652e81d4

How to speed up your jupyter workflow with keyboard shortcuts.

Code and markdown cells in jupyter video

https://web.microsoftstream.com/video/ab0425c8-f4c5-4c91-b972-256e23f47514

Explanation of the difference between code and markdown cells in jupyter.

Jupyter kernel video

https://web.microsoftstream.com/video/31d34f49-725f-45bf-9e93-5f0594e69427

How jupyter runs Python.

Plotting in jupyter video

https://web.microsoftstream.com/video/7bde4555-ca37-4edb-99c6-fcd668393b29

How to plot in jupyter.

Jupyter lab

https://blog.jupyter.org/jupyterlab-is-ready-for-users-5a6f039b8906

Blog post about the next version of jupyter.

A gallery of interesting Jupyter Notebooks

https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

"This page is a curated collection of Jupyter/IPython notebooks that are notable."

Generating random data (numpy)

This week we will look at the numpy.random package for generating random data in Python.

Introduction to numpy.random video

https://web.microsoftstream.com/video/ea6519e9-f7ed-444c-9b25-d855dfaa363a

An introduction to the numpy.random package.

Introduction to numpy video

https://web.microsoftstream.com/video/b191f6b2-4d80-4ede-8b2b-62a945999585

An introduction to the numpy package.

Setting up the numpy.random repo video

https://web.microsoftstream.com/video/19b0a319-3a98-4cd9-a46b-c1ba13bc25d4

Getting started with a numpy.random notebook.

---

numpy.random docs video

https://web.microsoftstream.com/video/659b15a1-6818-4b71-aa21-5b918351ab20

The numpy.random documentation.

---

rand function video

https://web.microsoftstream.com/video/d44d42fa-5b1a-4152-9d41-8129a3816001

About the rand function in the numpy.random package.

Distributions video

https://web.microsoftstream.com/video/e6f57177-0a17-45a5-8ca8-0e3191e887f8

Figuring out what the numpy.random Distributions functions do.

Seeds video

https://web.microsoftstream.com/video/db596673-6d67-4d13-93c3-074728c693f2

Seeds in the numpy.random package.

Assessments

Practical assignment instructions

https://github.com/ianmcloughlin/progda-assignment-2018/raw/master/assignment.pdf

Click the above link to read the instructions for the practical assignment for this module.

Assignment video

https://web.microsoftstream.com/video/5bb5d2aa-e61e-4612-a7ab-e52e9fb26fa0

Click the above link to view a video describing the assignments for each module.

Submit practical assignment work

Please create a single git repository for the assignment (not including the project) given in this course. You should make the git repository available to the lecturer by URL and enter that URL here. The easiest way to do this is to create a git repository in GitHub. Please only enter your URL in the box provided - if you would like to inform the lecturer of anything regarding your submission you can email ian.mcloughlin@gmit.ie.

Deadline: to be confirmed

Project instructions will be available here soon

Project instructions will be available here soon

Project instructions will be available here soon

Project

Please create a single git repository for project given in this course. You should make the git repository available to the lecturer by URL and enter that URL here. The easiest way to do this is to create a git repository in GitHub. Please only enter your URL in the box provided - if you would like to inform the lecturer of anything regarding your submission you can email ian.mcloughlin@gmit.ie.

Deadline: to be confirmed

Resources

Note: most of the blurbs in this section are taken directly from the websites linked.

Pandas documentation

http://pandas.pydata.org/pandas-docs/stable/

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

Numpy documentation

https://docs.scipy.org/doc/numpy/

NumPy is the fundamental package for scientific computing with Python.

Scipy library documentation

https://docs.scipy.org/doc/scipy/reference/

The SciPy library is one of the core packages that make up the SciPy stack. It provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization.

Matplotlib documentation

https://matplotlib.org/contents.html

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

iPython documentation

https://ipython.readthedocs.io/en/stable/

IPython provides a rich architecture for interactive computing with: a powerful interactive shell, a kernel for Jupyter, support for interactive data visualization and use of GUI toolkits, flexible, embeddable interpreters to load into your own projects, and easy to use, high performance tools for parallel computing.

Jupyter notebook documentation

https://jupyter.readthedocs.io/en/latest/

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Python for Data Analysis, William McKinney

http://shop.oreilly.com/product/0636920050896.do

A book about data wrangling with pandas, numpy, and iPython.

Download Anaconda Distribution

https://www.anaconda.com/download/

A free and open-source Python distribution with lots of pre-installed packages.

Visual Studio Code

https://code.visualstudio.com/

VS Code is a new type of tool that combines the simplicity of a code editor with what developers need for their core edit-build-debug cycle. Code provides comprehensive editing and debugging support, an extensibility model, and lightweight integration with existing tools.

Git homepage

https://git-scm.com/

Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

Cmder | Console Emulator

http://cmder.net/

Windows only: if you are using Windows I recommend you use cmder as your terminal. If you are on Mac or Linux I recommend you use the default Terminal.

Navigation

Home

- Dashboard

Site pages

My courses

52465 -- PROGRAMMING FOR DATA ANALYSIS

Participants

- Badges
- Grades
  - About the module
  - Python review
  - Plotting basics (matplotlib)
  - Browser workflows (jupyter)
  - Generating random data (numpy)
  - Assessments
  - Resources
- User badges

52446 -- FUNDAMENTALS OF DATA ANALYSIS

Administration

Course administration

Unenrol me from 52465ian.mcloughlin@gmit.ie

Student Links

MyApps
Exam Results
Office 365
Email
Library
Student Portal
Students Union
Timetable
Filesender
Staff Directory

Staff Links

MyApps
Learnonline Course Request 2018-19 (Staff Only)
Learnonline Archive
Office 365
Email
Staff Website
Library
CED Portal
Staff Resources
Staff Directory
Programme Catalogue 2018/19
GURU
Web For Faculty
Module Manager

School Links

School of Business
School of Engineering
School of Science
College of Tourism & Arts
GMIT Centre for Creative Arts and Media
GMIT Letterfrack
GMIT Mayo

Library

Get the mobile app