London Office of Technology and Innovation

LOTI and ONS Data Science Bootcamp

Starts 12 July 2021

@LOTI_LDN

www.loti.london

#LOTI

Loti Data Science Campus

Introduction

The Office of National Statistics (ONS) Data Science Campus is at the heart of leading-edge data science capacity building with public sector bodies in the UK and abroad. They equip analysts with the latest tools and techniques, giving them the capability to perform effectively in their roles.

The London Office of Technology and Innovation (LOTI) in collaboration with the ONS Data Science Campus and the Greater London Authority (GLA) have designed a cross-borough Data Science Bootcamp.

The Bootcamp will be available across multiple public sector organisations, working in collaboration in learning clusters. It aims to develop the capability and skills of officers and provide practical demonstrations of the value of data science within London's local government.

The Bootcamp will be offered to members of LOTI's Data Science Network made up of Data Scientists and Analysts learning to code.

Overview

- The Bootcamp will focus on skills, techniques and approaches to data matching
- 12 week programme with 2 ability streams
- 1 day a week commitment
- Participants can join either the Beginner & Intermediate (8 weeks) or Advanced (4 weeks) streams or work through the full 12 week programme
- Mentorship is available for all participants
- Kick off in July running through to September
- Exclusive access to the ONS Learning Hub
- Free to participate
- 15 places

loti Data Science Campus

Details about the Data Science Bootcamp

Format of a typical day

- Morning stand up
- Project work
- ‘Coffee and coding’ participants come together to discuss blockers, share knowledge etc.
- Project work
- Mentors available all day. Individuals have the option to book a 1-2-1 slot with a mentor.

Programme Availability

We have **15 places available**! Places will be distributed fairly across the boroughs, subject to the skills and capabilities of those who apply. We welcome sign-ups from more than one participant per borough, but cannot guarantee everyone a place.

Timetable

Beginner - Intermediate

Monday 12 July – Monday 6 September (This is a 9 week period as the programme won’t take place on Monday 30 August)

Advanced

Monday 13 September – Monday 4 October

Train the Trainer

As there are limited places on the Data Science Bootcamp, we are hoping to develop a process by which learning resources can become learning pack templates for programme participants to share with colleagues in their boroughs.

Loti Data Science Campus

Overview

Data matching is a foundational skill set in data science, enabling data to be brought together for descriptive, diagnostic and predictive analytics.

Taking part in LOTI and ONS's Data Science Bootcamp will give you the ability to use data matching techniques to tackle some of these common challenges:

- Match addresses across multiple data sets
- Get the single view of the resident
- Understand the makeup of households
- Provide a consistent user experience across services and between Boroughs
- Identify resident vulnerabilities
- Predict service demand

Data Matching

Loti Data Science Campus

Overview

Some of the skills and techniques you'll acquire as a result of the programme:

Basic:

- Data / String manipulation + cleaning
- Data Linkage in Python and R
- Good Practice in Data Linkage
- Address Matching

Intermediate:

- Deterministic matching algorithms
- Probabilistic matching algorithms
- Regex

Advanced:

- Machine Learning approaches to address matching

Data Matching

Data Science Campus

Draft Course Overview - Beginner / Intermediate (1/2)

Pre-learning

Required: Into to Data Linkage and either Intro to Python or Intro to R - all available on the ONS Learning Hub.
Optional/recommended: foundations of SQL, RAP pathway, PySpark

Week 1: Consolidate

- Recap of data analysis in Python / R
- Data Quality

Week 2: Exact Matching

- Recapping Joins
- Sections 1-3 (up to exact matching) of Data Linkage course
- Multiple key joins, composite keys and unique identifiers
- Missing data investigation / discussion

Week 3: Further Preprocessing for Data Linkage

- Basic string/ manipulation
- Text Normalisation (NLP course chapter 2)
- Regex (NLP course chapter 3)
- Applying string manipulation to open data set 1

Loti Data Science Campus

Draft Course Overview - Beginner / Intermediate (2/2)

Week 4: Mini Project 1

- Bring your own data set to work on applying methods learned (or have another one provided)
  - Opportunity to apply knowledge in different data set / business area
  - Support from mentor

Week 5: Rule Based + Score Matching (deterministic)

- Sections 4+5 of Data Linkage course
- Application of methods learned to open data sets
  - Compare analysis of data given different matching methods

Week 6: Probabilistic Matching

- Sections 6+7 of Data Linkage course
- Fellegi-Sunter method
- Sampling from probabilistic methods + evaluation
- Application of probabilistic methods for Open data sets
- Comparison with previous methods on analysis

Week 7: Further methods in data linkage

- Exploration of other methods, edit distances, introductory ML methods
- Exploration of other packages / software / language solutions
- Preparation for Mini Project 2

Week 8: Mini Project 2

- Application of intermediate methods to personal / business data set (or other found open data)
- Presentation of analysis
- Retrospective / lessons learned

Loti Data Science Campus

Draft Course Overview - Advanced

Pre-learning

Required: Introduction to Python/R, Data Linkage in Python/R, Machine Learning in Python / R

Participants will ideally have their own business problem they can explore with these techniques to supplement the open data provided.

Week 1: Consolidation

- Review of deterministic / probabilistic matching concepts from Data Linkage course
- Discussion of data quality
- Why decide to do ML for linkage?
- Similarity scores, distance metrics

Week 2: Supervised Machine Learning Matching

- Recap of supervised learning methods
- Creating training data for record linkage
- Evaluation metrics for classification
- Case study / practice on open data, comparison with other methods

Week 3: Unsupervised Machine Learning Matching

- Challenges with supervised learning methods
- Recap of clustering / expectation maximisation
- Feature generation of clustering, evaluation methods
- Application of unsupervised classification on open data, comparison with previous methods

Week 4: Applied Project

- Support for applying ML matching techniques to personal data / practice open data
- Presentation from learners on their own work / matching problems

Loti Data Science Campus

Name	Organisation
Chiadi Lionel	Camden
Luke Ballance	Camden
Malgorzata Lachowska	Greater London Authority
Yiran Wei	Greater London Authority
Toby Meller	Kingston upon Thames
Huu Do	Hackney
Lindsey Coulson	Hackney
Lee Latchford	Havering
Anna Trichkine	Hounslow
Ejaz Hussain	Hounslow
Ian Hanson	Kensington and Chelsea
Sean Pedrick-Case	Lambeth
Karen Kemsley	Lewisham
Emmanuel Steadman	Tower Hamlets
David Saxton	Tower Hamlets

Pilot Cohort

Loti Data Science Campus

How many people can sign up from one borough?
We accept multiple applications from the same borough.

How many places are available on the course?
15 places in total, but there are limited spaces on the Advanced stream.

How will ONS & LOTI assess the level of prior knowledge?
Prospective participants will be asked to self-assess their skill levels using our Expression of Interest form [CLOSED].

Must applicants complete the relevant ONS courses on Python / R?
No, equivalent knowledge is acceptable.

How is the course carried out?
Open to flexibility and is adaptable. Participant-led, so please do bring your own datasets. Introduction to content, demos, and then time for participants to put in practice what they've learned, with time aside for 1-2-1 tuition.

What day in the week?
Monday each week

FAQs

Loti Data Science Campus

How do participants access the pre-training materials?
Once the final participants have been decided, then successful applicants will need to approve a privacy statement to access the full raft of content on the ONS Learning Hub. Relevant courses will be highlighted.

What is the recommended language?
R is more user-friendly and Python is more useful for development and deploying machine learning, so it's a matter of preference.

What software and packages need to be installed ahead of the course?
Pre-learning courses will prompt participants to install what will be necessary for the course.
The full list of additional packages and libraries will be circulated to successful applicants as part of the pre-learning starter pack.

Are you expecting people to complete Intro to Python & R as a pre-requisite?
Pre-programme questions will be shared with participants beforehand to gauge what language to focus on. No restrictions on preferred language.

FAQs

Data Science Campus

Is there a benchmark/ standard that is embedded in the prospectus for participants to self assess?

Participants will be asked to complete the recommended ONS courses prior to starting the programme. This will bring them up to the required level.

Utilising libraries to access Python & will it be available after the course?

Some of the content will be taken from pre-existing libraries, but some yet to be determined to allow for flexibility. Majority of packages will have been installed as part of pre-learning content & time will be allowed for participants to download during the course.

Can we use personal laptops?

Yes, but it depends on the data that is being used. If bringing your organisation's data, then participants would be expected to have liaised with their organisation's Information Governance teams beforehand to ensure that will be possible.

How many places are there on the Advanced course?

15 places in total, but there are limited spaces on the Advanced stream.

FAQs

Data Science Campus

Can participants switch between languages?
We would recommend that participants stick to one but won't restrict you from doing so. The theory is the same but participants would be doubling workload. Easier to learn in one language and once comfortable, can switch between them by applying syntax.

How do participants contact their mentors?
Mentors can be contacted by email in-between. Potential for creating a channel on the Government Data Science Slack channel.

Does data matching cover address base as that's what is used in Council?
It depends on the datasets, so please do bring the datasets that would be of most use and we will remain flexible.

Is access to the Learning Hub time-limited?
After completing the programme, participants will have access to the learning hub for an indefinite period.

FAQs

Loti Data Science Campus

London Office of Technology and Innovation
ONS Data Science Campus

For further information

Jay Saggar - Programme Manager (Data, Smart City & Cyber Security), LOTI

Onyeka Onyekwelu - Strategic Engagement Manager, LOTI

Sophie Nelson - Programme Officer, Data Science Campus

www.loti.london

Loti Data Science Campus