Introduction to Data Science

Welcome to our introductory workshop. We hope that this will help you learn a bit more about what data scientists do and the process underlying a data science project.

# What Is Data Science?

Data Scientists use data to answer questions and solve problems. You can divide the skills and knowledge that they use into four key domains: Statistics, Programming, Communication and Domain Knowledge.

The tools data scientists use to analyse data come from statistics, for example, statistical models or machine learning methods. However, this domain also helps data scientists design studies so they can work out what data and methods will be needed to answer a certain question. Knowledge of the domain one is studying is also important in deciding how best to answer a question. Data scientists will often work on projects in a range of different areas (for example, geography, business or medicine) so it is useful to be able to learn about these things quickly.

Programming is needed to implement the statistical tools chosen to answer a question – many statistical programs like R are run using code and other languages like Python and SQL are often used for data processing and database management. It is important for data scientists to have good communication skills to they can discuss and explain their results to people who might not be experts in data science.

This might seem like a huge range of skills to learn at first – and it is – but it’s important to note that data scientists will often specialise in specific areas and won’t be experts in every aspect of these four domains. Below are some skill profiles of data scientists (DS 1 – 5) from the Leeds Institute of Data Analytics at the University of Leeds to show this. Blue indicates an area in which they have specialist skills, green an area in which they are confident, yellow an area they are still learning and orange and area they have not yet learned. Note this is not an exhaustive list of skill areas!

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DS 1 | DS 2 | DS 3 | DS 4 | DS 5 |
| STATISTICS | | | | | |
| Study or experiment design |  |  |  |  |  |
| Basic Statistical Modelling |  |  |  |  |  |
| Modelling complex data structures |  |  |  |  |  |
| Spatial statistical modelling |  |  |  |  |  |
| Frequentist Inference |  |  |  |  |  |
| Bayesian Inference |  |  |  |  |  |
| Simulation methods |  |  |  |  |  |
| Optimisation processes |  |  |  |  |  |
| Machine learning |  |  |  |  |  |
| PROGRAMMING | | | | | |
| Basic computer science e.g. shell scripting |  |  |  |  |  |
| Planning and structuring code |  |  |  |  |  |
| Scripting language e.g. Python |  |  |  |  |  |
| Statistical programming language e.g. R |  |  |  |  |  |
| Database management software e.g. SQL |  |  |  |  |  |
| High performance computing/parallel processing |  |  |  |  |  |
| DOMAIN KNOWLEDGE | | | | | |
| Understanding of the application area |  |  |  |  |  |
| Ability to learn about new fields quickly |  |  |  |  |  |
| COMMUNICATION | | | | | |
| Creative data visualisation |  |  |  |  |  |
| Explaining data science to non-data scientists |  |  |  |  |  |
| Translating results into real-world implications |  |  |  |  |  |
| Presenting skills |  |  |  |  |  |
| Writing skills |  |  |  |  |  |

Each of the data scientists have skills in different areas, for example DS 2 is more of a specialist in programming and DS 4 is quite specialised in communication. All the data scientists have a good knowledge of basic statistical modelling and translating their results to real world implications. Most also have a good knowledge of a statistical programming language. All the data scientists have areas of statistics that they are less familiar with, and most have programming skills they are still learning.

On the next page is a skills profile you can fill out for yourself to see what areas of data science you already have skills in and which you have not learned about yet. You could use this as a map to help you find courses and workshops in specific areas and help improve your data science skills.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Skill/Knowledge Area** | **Specialist** | **Fairly confident** | **Still learning** | **Not learned** |
| STATISTICS | | | | |
| Study or experiment design |  |  |  |  |
| Basic Statistical Modelling |  |  |  |  |
| Statistical modelling for complex data structures (e.g. heirarchical) |  |  |  |  |
| Spatial statistical modelling |  |  |  |  |
| Frequentist Inference |  |  |  |  |
| Bayesian Inference |  |  |  |  |
| Simulation methods |  |  |  |  |
| Optimisation processes |  |  |  |  |
| Machine learning |  |  |  |  |
| PROGRAMMING | | | | |
| Basic computer science e.g. shell scripting |  |  |  |  |
| Planning and structuring code |  |  |  |  |
| Scripting language e.g. Python |  |  |  |  |
| Statistical programming language e.g. R |  |  |  |  |
| Database management software e.g. SQL |  |  |  |  |
| High performance computing and parallel processing |  |  |  |  |
| LOCAL DOMAIN KNOWLEDGE | | | | |
| Understanding of the application area for your work |  |  |  |  |
| Ability to learn about new fields quickly |  |  |  |  |
| COMMUNICATION SKILLS | | | | |
| Creative data visualisation |  |  |  |  |
| Explaining data science to non-data scientists |  |  |  |  |
| Translating results into their real-world implications |  |  |  |  |
| Presenting skills |  |  |  |  |
| Writing skills |  |  |  |  |

# The Data Science Process

Most data science projects start with a research question or objective and follow the process below

Data might be given to you alongside the question, or it might be something you find yourself through a process such as web scraping. Most data will have errors, anomalies and missing values or will come in a format you can’t use. Cleaning data is the step in which you deal with this by finding and perhaps correcting for inaccuracies. Some of this may take place in tandem with the next step, exploring the data. In this part you are trying to get an idea of the structure of your dataset and what the relationships between different variables in it are. You might use summary statistics or plots of the data, among other techniques. After this you can then go on to analyse your data in a way that answers your research question. The final step is to communicate these findings. This step is very important as it connects your data science work back to the real-world problem you started with!

The table below shows some examples of things you might do or tools you might use in the different stages of this data science process. See if you can identify which example belongs in which stage (some may fit in more than one). You are encouraged to look things up on the internet to help.

|  |  |
| --- | --- |
| **Example** | **Data Science Process Stage** |
| Neural networks |  |
| Multiple imputation |  |
| Web scraping |  |
| Presenting at a conference |  |
| Computing summary statistics |  |
| Correlation matrix plots |  |
| Agent based modelling |  |
| Multilevel modelling |  |
| Decision trees |  |
| Cartogram |  |
| Simulate data |  |
| Network analysis |  |
| Sankey diagram |  |
| Outlier detection |  |
| Find open data |  |