**IOC Topic 11b – Advanced Data Science**

Transcript & Notes: PART 1

Author: Dr. Robert Lyon

Contact: robert.lyon@edgehill.ac.uk ([www.scienceguyrob.com](http://www.scienceguyrob.com/))

Institution: Edge Hill University

Version: 1.0

**Topic 11b, Part 1**

**Introduction Slide**

**Slide 1**

Hello and welcome to Part 1 of Topic 11b, Advanced Data Science. During this topic I'll introduce what data science is, the basic principles underpinning data science, and some important data science tools that may be unfamiliar to you. My name is Dr. Robert Lyon, and I’ll be taking you through the learning material.

**Slide 2**

What material will we cover while studying this topic? Well, this topic aims to introduce…

* What data science is all about.
* The Key concepts underpinning good data science – primarily the scientific method.
* Useful terminology that will help you navigate the world of data science.
* Important tools crucial for successful and reproducible data science – these are the tools provided by Statistics.
* Data collection & Experiment Design practices.
* Probability basics – very important for statistical inference.
* Data distributions that describe the characteristics of data.
* Hypothesis testing – a formal method for testing predictions.

The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.

In this part we’ll consider what data science is and describe the scientific method.

**Slide 3**

Data science means different things to different people. To some data science is just business analytics, business intelligence, predictive modelling, or statistics by a different name. I feel there is some truth to this sentiment. Yet today data science is a discipline that combines many of these areas to form something more than the sum of its parts. With that in mind, we can loosely define data science as follows. “Data science is a technical discipline concerned with the extraction of new knowledge from data, via application of the scientific method, in conjunction with the tools of mathematics and statistics.” Please feel free to quote me on that!

This description is very high level and hides much of the complexity of what data science involves in a practical sense. It is a multidisciplinary area. To succeed in data science, you require knowledge from lots of different areas. This includes computer science, mathematics, and of course the domain in which data science is to be applied. It is hard for any one person to be an expert in all three areas. In reality we have to apply a trade-off – either we acquire deep knowledge in one specific area, or shallower knowledge in all three. This trade-off is acceptable so long as we remain aware of our limitations and seek out knowledge to fill any gaps when required.

**Slide 4**

* Let us expand our understanding of what data science is. How do others see data science?
* There are many views depending on an individual's perspective.
* What is your perspective - perhaps you already have a view? Take a few moments to consider what you think a data scientist is or should be. When ready watch the video (3-minute runtime, <https://youtu.be/edZ_JYpOM8U>) shown on the slide. Remember, if your initial thoughts conflict with what you hear during the video, that’s ok – even amongst industry professionals there can be disagreement.

**Slide 5**

* In the real-world, we apply data science in a variety of ways according to the problem domain. For instance, in Medicine we may be asked to determine which patients are most at risk of diabetes, based on patient health records. We collect data much like that shown in the table, explore it, and try to understand its content. Once we've done that, we can construct a statistical model that predicts patient risk using examples of previously seen patients. This type of problem represents a straightforward application of data science. But in other industries data science may involve solving different types of problem, for example:
* Minimize company expenditure by using statistics to determine the area in which savings can be most easily be made.
* Segment customers into groups that accurately characterize their credit risk.
* Determine which products a customer is mostly likely to buy based on past spending habits.
* Determine if a drug is successful at treating a specific condition.

I’m sure you can think of many more examples. What matters at this stage, is that you understand that these problems are tackled using a scientific approach, along with the tools of computing and statistics.

**Slide 6**

The video shown on slide three introduced some interesting ideas.

* Principally that data science is not a new discipline. If we analyze this idea against the definition of data science I provided earlier, then this view certainly holds true.
* For thousands of years humans have collected data about our world. Such data has been used to make predictions, to develop theories, and reveal new knowledge. Astronomy provides perhaps the most famous example of this, as the video explained. For many centuries' humans collected data describing the movement of planets and stars across the night sky. This information was used to form theories that could explain their movement. Ancient astronomers initially explained their observations using a Geocentric or “Earth based” model of the universe. This is shown in the image on the slide. This shows the planets and stars moving around the Earth in loops called Epicycles. People genuinely believed this to be the case for many centuries. However, over time more observational data was collected. Eventually, humans began to realize the Earth was not at the center - the principles of data science helped us get there (eventually!).

**Slide 7**

What's changed to make data science such a hot topic in recent years?

* For most of human history “data science” activities were undertaken by experimental scientists. Scientists would develop an idea, determine how to test that idea, collect experimental data, analyse the data, and produce new knowledge by interpreting the results.
* Contrast this to the modern world we live in. Everything has changed - data is generated about each and every one of us every day. If we own a smart watch, our biometric data is being recorded each second. This data is uploaded to the cloud for us to analyse. Companies can cross reference this information against our medical data during GP or hospital visits. That medical information can be linked to our financial records by insurance companies hoping to improve their pricing strategies. The information may be passed on to advertising companies, who hope to use it to produce targeted ads with the aim of selling products. This could be done in combination with records of your social media activity. I could go on…
* Clearly data is everywhere, and you already know this. But consider this - little of this data was collected with the aim of revealing new knowledge. Much of it is collected for record keeping purposes.

**Slide 8**

Data is everywhere. We are now living in a data-driven world. There are many challenges in this brave new world facing those working with data.

* Data wasn’t necessarily collected to answer any specific question or solve a specific problem. This is an important point worth thinking about. Data can only reveal new knowledge if there is hidden knowledge contained within it - data science isn’t magic. If we aren’t designing the data we collect to maximise its potential to yield new knowledge, we have to accept the possibility that no new insights can be revealed. Consider a simple example - suppose a transit company wants to predict the likelihood of traffic jams on a given day to optimise its operations. They don’t have access to meteorological or accurate traffic data as that would be expensive to obtain. They do have internal records based on driver reports, which describe jams on specific stretches of road, the time this happened and the weather at the time (raining, dry, snow etc). This data wasn’t designed to determine driving conditions – it was recorded to ensure drivers were not blamed for being late. But if they cross reference the data with the current weather, can they optimise their journeys? That’s something for you to think about.

**Slide 9**

* The types of questions we ask of our data are becoming more complex – we require more than just knowledge of statistics to answer them.
* Data volumes are increasing dramatically. New tools are therefore required to process the data collected. New processing paradigms also required, often necessitating the use of specialised data processing hardware.
* Data is increasing in complexity - this complexity arises as the granularity of data collected becomes finer, but also as data sets are combined and cross-referenced.
* The domain knowledge required to stay on top off these issues crosses many areas and is increasing over time.
* Until recently there was no role that covered all these areas. Hence the data scientist role emerged to fill the gap, and with great success.

**Slide 10**

Because of real-world needs, data science has become something of a buzz word in recent years. Companies all over the world are frantically hiring data scientists, and the media is regularly reporting that “Data Scientist” is the hottest (and one of the best paid) jobs around. For these reasons lots of people want to become data scientists. To meet the demand, data science courses have cropped up all-over web, with the aim of upskilling as many people as possible. Universities too are offering full-time bachelors and masters level data science courses – you’re on a similar course right now! But does this hype hold up under scrutiny? Here we see a plot produced by ITJobsWatch (<https://www.itjobswatch.co.uk/jobs/uk/data%20scientist.do>), an organisation that researches the IT job market. This plot has time listed on the -axis and pay listed on the -axis in pounds (£). It shows the salary range for data scientist roles over time. We can see that the median salary, shown via the yellow line, is high. This provides some evidence supporting the view that data science jobs are well paid.

1. <https://www.bloomberg.com/news/articles/2019-05-15/big-data-skills-earn-high-pay-for-today-s-college-graduates>
2. [https://www.forbes.com/sites/louiscolumbus/2019/01/23/data-scientist-leads-50-best-jobs-in-america-for-2019-according-to-glassdoor/#272fb8e07474](https://www.forbes.com/sites/louiscolumbus/2019/01/23/data-scientist-leads-50-best-jobs-in-america-for-2019-according-to-glassdoor/)
3. [https://www.forbes.com/sites/forbestechcouncil/2019/10/14/the-birth-of-the-data-science-generation/#496f3d195845](https://www.forbes.com/sites/forbestechcouncil/2019/10/14/the-birth-of-the-data-science-generation/)

**Slide 11**

Here is a further plot produced by ITJobsWatch (<https://www.itjobswatch.co.uk/jobs/uk/data%20scientist.do>). This plot has time listed on the -axis. While the -axis describes the proportion of IT jobs (in percent) advertised with the title “Data Scientist”. The plot appears to show that an increasing proportion of advertised jobs are “Data Scientist” roles. This provides evidence for my earlier claim that employers are increasingly hiring data scientists in industry. Thus, there appears to be some merit to the hype. Yet you may have some questions:

1. How significant are these trends?
2. What are the data sources, are they trustworthy?
3. Is the data biased or skewed in some way?

These questions can be answered via application of the tools of data science. As data scientists we must accept the possibility that claims/predictions may hold true. Perhaps “Data Scientist” is the best paid job in IT right now, and perhaps employers are clamouring to hire data scientists. We must remain objective and maintain a healthy scepticism, until we can prove or disprove such claims in a principled and reproducible way. However, with that said, I humbly suggest that the popularity of data science isn’t just down to hype. There is a significant demand for professionals with data science skills capable of answering complex real-world data questions.

**Slide 12**

What key attributes does a typical data scientist have? These can be hard to define as they can vary from business to business, so we keep things high-level. Data scientists are typically,

* Competent programmers in one or more high-level programming languages (Java, Python, C++ etc).
* Knowledgeable of database systems, with some experience of relational and non-relational databases (MySQL, Postgres SQL, MongoDB etc).
* Possess some statistical background and an understanding of data distributions.
* Have prior experience of building/applying machine learning algorithms to data.

**Slide 13**

* Some look at these criteria and worry they can’t meet them – however it is possible to learn these skills with time and persistence. I believe anyone can do this.
* What isn’t mentioned here are the myriad of other skills that many people already possess (including TechUP participants), that great data scientists possess. For instance,
  + the ability to communicate effectively
  + to present methods and results concisely to those without technical backgrounds
  + to problem solve
  + to be rigorous in your approach, I could go on…
* If you’re thinking about heading into data science, remember – do not focus on the attributes you may not have when looking at this list, but on those you already possess! Whilst if you’re someone who may be hiring data scientists in the future, also keep in mind that very few people will be experts in everything. Good people often suffer from imposter syndrome, so you should keep this in mind. Also remember that data science is constantly evolving. Thus, investing in your existing data scientists via providing regular training/upskilling opportunities, is a sensible way to maintain a highly skilled and dedicated analytics team.
* Why not watch this video (2-minute runtime, <https://youtu.be/72JTnRTRGY8>) which shows some leading data scientists talking about what traits they think a data scientist needs.

**Slide 14**

Data scientist roles can be quite varied. But there are some essential steps that all must carry out during the course of their work on a project. This begins with,

* Identifying the business problem at hand. This helps determine the data required, the tools that may be needed, additional resources required and ultimately the objectives & deliverables.
* Then there is data acquisition. This involves securing data from one or more data sources that is needed to tackle the business need or answer a business question.
* Data pre-processing/Preparation follows. Here the data may need to be cleaned or transformed. This might require, for instance, filling in missing values or correcting erroneous entries.
* Once cleaning is complete, we can undertake Data Exploration. During this step we determine the most important variables in the data, the relationships between them, understand the distributions of our data etc. This step is often called Exploratory Data Analysis (EDA).
* After exploration we begin modelling our data. Data modelling may involve building machine learning classifiers to make a predictive model or applying standard statistic tools to answer a business question.
* Finally, the outputs of all the work need to be communicated in some way. Most likely via some form of visualisation. This communication may then initiate some form of internal business response, which brings us full-circle.

**Slide 15**

So far you’ve heard my view on data science, and what a data scientist does day-to-day. The following video (<https://youtu.be/X3paOmcrTjQ>), which lasts around 5 minutes, will provide an alternative view.

**Slide 16**

Next we hear from a data scientist in her own words. She provides an interesting insight into the day-to-day activities of a successful data scientist in industry. I encourage you to watch this video (https://youtu.be/\_Wk9T\_G-u4o) - it lasts around 3 minutes.

**Slide 17**

* Data science is becoming an increasingly important discipline.
* As a society we continue to collect more and more data, and the information stored within it has the potential to completely change our world.
* This is because data is now being regularly used to make decisions. Such decisions can affect the lives of people in unforeseen ways, both good and bad.
* It is therefore crucial that we train competent and professional data scientists capable of owning such responsibility.
* If we do this well, data scientists have the potential to help optimise society in many ways – to reduce cost, to reduce harm, to improve supply and to discover unknown issues that can be solved.
* For all the positive possibilities there are an equal number of potentially negative outcomes. There are very few professions which have the capability to impact people in such a way. For instance, if a doctor makes a mistake during surgery, it is possible for a single patient to be harmed. If a data scientist makes a significant mistake, it is possible for many more people to be harmed.
* This is why data science, and principled data science, is crucially important.

**Slide 18**

Unsurprisingly, science lays at the heart of what data science is all about. Some of you may have science backgrounds. Others may not have encountered the topic since school. I want you to put aside any preconceived notations you may have about science as we consider a new perspective. I suggest that science is the most effective tool ever conceived to acquire new knowledge. It provides a systematic method for knowledge acquisition that allows us to both test and develop our understanding of the universe, using evidence to guide us. It is an iterative method that allows for continual improvement. It has allowed the human race to achieve some incredible things. It has enabled humans to land on the moon, unlock the power of the genome, invent medicines to improve our lives, and invent computing technology. Incredibly, these developments can be traced back at least 1000 years to Ḥasan Ibn al-Haytham (<https://en.wikipedia.org/wiki/Ibn_al-Haytham>), who first used these ideas to investigate the world around him.

**Slide 19**

It’s worth taking a moment to reflect on the importance of science and its greatest contribution (in my humble opinion) to our society – the scientific method. Data science relies upon the scientific method to investigate data and reveal actionable insights. When ready, watch the video (<https://youtu.be/Yfkt6WDQNdg>) on the slide. It introduces the ideas behind the scientific method and one of its earliest practitioners.

**Slide 20**

What is the method?

* It consists of a number of a steps which if undertaken in order, help us to establish truth.
* We’ll cover the main steps of the methodology here, noting that other steps can be added in between. This is usually done to describe the process in more detail.

We begin with a question that we aim to answer. The question may be descriptive, where the goal is to uncover the existence of something, or if a process exists. Relational questions seek to ask if a relationship exists between objects, processes or data points. Whilst casual questions seek to determine if a variable or process affects other variables or processes in some way. Always consider what type of question we are asking. This determines the focus for the next step of the method – research.

**Slide 21**

During research we aim to accumulate the domain knowledge required to ask our question, and to understand any potential answer. Research can proceed in many ways, and there is no prescribed approach. All research requires a literature review in order check if the answer is already known. It would be embarrassing to tackle an already answered question, if the answer is not in doubt. Once research is complete, we begin forming hypotheses. Hypotheses are proposed explanations or predictions that we can test. Once hypotheses have been defined, we can design an experiment to test them – do they hold up under scrutiny. This is determined during experimentation, during which we record results for analysis. We report the results, and then make a determination – given the results do we accept our hypotheses as true, or do we reject them as false. If we can accept the results with confidence, we can form an accepted theory that answers the original question. Otherwise we must return to the hypothesis development stage and formulate new hypotheses for testing.

**Slide 22**

Please take a look at the following video (<https://youtu.be/yi0hwFDQTSQ>) that describes the scientific method. It should help you understand the concepts we’ve introduced in the previous slide.

**Slide 23**

* We return to thinking about hypotheses. A hypothesis is a proposed explanation for an observation, phenomena or relationship.
* You no doubt form such explanations each day. Suppose you’re feeling under the weather. You decide some hearty soup might help. You prepare some up and get stuck in. The next day you feel much better. The question is, was the soup responsible for making you feel better? There are many possible explanations for why you recovered so quickly. Each represents a hypothesis. Your best guess may be that the soup helped – this is your working hypothesis.
* The null hypothesis is a proposed explanation that is assumed to be true until proven otherwise. It is normally the default position.
* We must form a null hypothesis as part of the scientific method. If we continue the soup-based example, a null hypothesis maybe that the soup had no impact on recovery. We normally denote the null hypothesis as .
* The alternative hypothesis is a proposed explanation that directly contradicts the null hypothesis. In our example, this could be that the soup did lead to recovery. If we can show this to be true, then this would add some new (and useful) knowledge to the world. We normally denote the alternative hypothesis as or .

**Slide 24**

* Hypotheses must be falsifiable, otherwise they can never be properly tested. For example - if I hypothesise that Loch Ness contains a large green reptile that answers to the name “Nessie”, this is impossible to test. This is because with each test I will only encounter an absence of evidence, for which people will propose an explanation. If I didn’t find Nessie today when searching, this is because it was hiding. When I search for tracks I find none - this is because Nessie never leaves the water, and so on. In general, a falsifiable statement only needs one observation to disprove it.

**Slide 25**

The video on the slide will provide another perspective on the null and alternative hypotheses. Please watch the video when ready (it lasts around 4 minutes, <https://youtu.be/ZzeXCKd5a18>).

**Slide 26**

When faced with a null hypothesis and an alternative hypothesis, we must design an experiment to test them.

* The design will vary according to the hypotheses under consideration, but there are some general similarities between all experiments. This is best illustrated via an example.
* Suppose we aim to determine if users of a shopping website spend more if targeted advertisements are removed.
* We form the null hypothesis which says removing ads has no effect on sales, whilst the alternative hypothesis indicates there is an effect (positive or negative) on sales. To test this, we design an experiment which measures sales for two groups of individuals. Those targeted by ads, and those not targeted.
* In this case, the group targeted by ads is known as the control group. Whilst the group not targeted is called the experimental group. Next we define the variables within our experiment.
* There are independent variables. In this case the independent variable describes the quantity of ads a user is exposed to. The independent variable is what we change between the control and experimental groups.
* Then there is the dependent variable, sometimes called the response variable. In this case this would be the value of sales per customer – this is the value we measure during our experiment.

**Slide 27**

* Finally, there are one or more control variables, which are constant, and don’t change between the groups. By designing our experiment in this way, having only one variable difference between the control and experimental groups, we can measure any real difference that may be present. If we altered more than 1 variable between the groups, we wouldn’t know which was responsible for any change we may observe. We can summarise experimental design as a process whereby you identify groups and determine the experimental variables. When designing experiments we must also keep in mind replicability. A good experiment is a reproducible experiment. The variables able we’ve used could present some replicability issues. What happens if the experiment is conducted in November / December, which in western countries, may see an increase in sales due to the holiday season. This sort of influence would have to be controlled for.

**Slide 28**

The following video (<https://youtu.be/DaBq0naj0YY>) will help recap some of the issues we’ve just covered with regards to experimental design. Please take a look and think about how you could apply this approach to data science challenges.

**Slide 29**

* Once you’ve designed your experiment, you ultimately prepare it and then carry it out.
* In data science this usually involves first collecting and cleaning data. Then we can write code/use tools to run the experiment and ensure results are collected in a standard intuitive format amenable to analysis.
* Perhaps the most important part of the experiment involves analysis. Here the null and alternative hypotheses are considered and evaluated against the results collected. Usually the results are analysed using the tools of statistics, and acceptance/rejection of any hypotheses decided using a statistical test – it is not a matter of personal opinion.
* Eventually results are reported to various stakeholders. Reports usually contain an executive summary the outcome for non-data scientists to digest, but with additional detail provided in the main body of the report – including a description of the experimental methodology. Software tools are often used to automatically produce sophisticated reports, which may include various forms of data visualisation.
* Traditionally negative outcomes have been viewed as failures – if we spend time developing an experiment and it fails, it can feel like a waste of time. However, negative results should not be viewed this way. Negative results are just as useful as positive results. They lead us to the objective truth and are crucial signposts on that path.

**Slide 30**

Now that we’ve become familiar with data science and the scientific process, let’s hear about some case studies. When ready watch the video (<https://youtu.be/ypbSMS8XrAE>) on the slide. Think about how you might solve these problems, what steps might be involved to solve them – what hypotheses might you need to form? I don’t expect answers, but thinking about these things will give you an insight into how data science is done in the real-world.

**Slide 31**

So far, we’ve introduced:

* What data science is, the science of analysing data to reveal new knowledge and insights.
* The nature of the data science role, what the discipline involves day-to-day and the skills required to be a data scientist.
* The data science process – from the formation of a business need, the collection of data, and the analysis.
* The scientific method in relation to data science. The process of developing a science question, experiment design, and data analysis.

Next, we’ll cover some of the common terminology used in the field of data science.