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

Transcript & Notes: PART 2

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 2**

**Introduction Slide**

Hello and welcome to Part 2 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 1**

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

* What data science is all about.
* 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 part 2 we're going to get to grips with the terminology most commonly used in data science.

**Slide 2**

The field of data science is filled with terminology that can make it hard for beginners to get started. The terminology is spread across multiple disciplines, which makes things even harder. You almost need a map to understand the data science landscape. To complicate matters that landscape is constantly changing. I hear terms mentioned from time-to-time that I’m unfamiliar with. Often the new terms describe existing concepts/ideas! Yet the new terminology takes hold so we must learn it. To help guide you I’ve provided a map of sorts. Hopefully this will help you learn more quickly, and enable you to participate in data science discussions.

**Slide 3**

During part 1, slide 18, we encountered this diagram describing the scientific process. This is a good starting point for our map, since principled data science involves applying the steps of the scientific method in sequence. As we can see here, we start with a research question that we wish to ask. We then conduct research to equip ourselves to understand it. Once done we’re in a position to develop hypotheses, design and conduct an experiment, analyse and interpret the results, and finally report them. We’ll focus on the enclosed portion of this process to create our map.

Before we do that, let us consider the different umbrella terms often used to refer to this process but related to common application areas.

**Slide 4**

Analytics is a term often used to describe a field focused on the discovery, interpretation, and communication of meaningful patterns in data. We often see modifiers applied to this term. Such as Big Data Analytics, which deals with extremely large dataset. Then there’s Real-time analytics, concerned with decision making and pattern discovery in real-time on high volume streams of data – where a stream is a data set of potentially infinite size that grows over time.

Many consider these terms to be synonymous with Data Science, which I’ve previously described as:

“ 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.”

If you search for definitions of these terms for yourself, you’ll find conflicting descriptions and justifications for why they are the same or different. Though some feel that analytics is a sub-field within data science, as data science has become somewhat all encompassing!

**Slide 5**

Whatever view you take, there are further modifiers you can apply to the terms data science and analytics. There is predictive analytics/data science, which is concerned with making predictions from historical facts and trends in data. Whilst descriptive analytics/data science, focuses on improving our understanding of dataset by interpreting it using summary statistics. Finally, prescriptive analytics/data science combines predictive and descriptive approaches allowing decisions to be made using past data and predictions of the future, to yield improved decision-making outcomes.

Analytics is applied to different fields too. When applied in Business, it is referred to as Business Intelligence (BI) or Business Analytics. In medicine it is usually called Healthcare Analytics. There are many other areas such as Fraud Analytics, not to mention the application of these methods in science. What we’ve covered here provides us with a high-level view of our data science map. Next, we’ll delve a little deeper into specific areas.

**Slide 6**

You’ll find the following video useful when it comes to understanding the terms and buzz words surrounding data science. The info graphic mentioned during the video, can be found here: <https://365datascience.com/wp-content/uploads/2018/05/365-Data-Science-Infographic.jpg> for those that are interested.

Video link: <https://youtu.be/edZ_JYpOM8U>

**Slide 7**

Now we try to break down data science terminology into semantically related areas. This map won’t be perfect - there’s too much terminology out there to cover. So we’ll deal with the common terms as a starting point. I'll also provide details in the notes, so don't worry if you miss a term.

When designing an experiment, we may think about the types of people we made need to undertake it. What sort of roles do we need to fill? We may need,

* A data analyst is a person that may clean, transform and analyse data using statistical tools, with the aim of unlocking insights useful for decision making from existing data.
* A data scientist is a person that applies statistical and machine learning based tools to data, with the aim of unlocking insights useful for decision making. This sounds similar to the data analyst role, though data scientists are usually more technically capable. Data scientists often have extensive coding experience. A data scientist is able to make future predictions from data.
* A data engineer is a person that takes responsibility for the collection, cleaning, and transformation of data. They also develop / maintain the architecture used to undertake data science. This is a crucial supporting role.

**Slide 8**

* A data wrangler (sometimes called a data "munger") is a person that takes ownership of the transformation of data so it is more easily processed by data scientists (and software programmes). This role is varied and includes cleaning the data, deciding what to do with missing values, and combining possibly very different datasets for use.
* A data architect is someone responsible for designing how data is collected, stored, and processed within an organisation. They focus on designing an infrastructure capable of enabling cutting edge analytics. Generally this is a senior role filled with someone possessing years of experience.
* A unicorn – yes you heard me a "unicorn", is a term you might hear in this space. This describes the (mostly) mythical individual capable of filling all the previously described roles. Such individuals have a wealth of experience which makes them incredibly rare. It is up to you to determine whether or not such individuals really exist. Unicorns bring experience that can greatly accelerate the data science function within in a business, yet if/when they leave, the data science function is greatly impacted.

**Slide 9**

In order provide those in these roles with something to do, we need data – or data sources we aim to tap. There are many different types of data source, online databases, websites, physical documents etc. There is some terminology related to data that you may find useful.

* Velocity describes how quickly data arrives to be processed – in some situations data may arrive in one large lump, for example, when accessed from a database. However data may also arrive incrementally over time, for instance, like the data produced by social media platforms.
* Variety refers to the different characteristics of the data being processed. Data sources may be highly varied, For instance, patient heart rate traces are a form of time-series data. Whilst patient records are in tabular format.
* Veracity refers to how accurate the data is or how far it can be trusted. When monitoring self-reported spending habits we may need to be careful trusting this data. People maybe unwilling to reveal the truth extent of their (perhaps frivolous spending). Whilst banking records should be more accurate, precise, and trustworthy.
* Variability refers to how much the data varies, perhaps over time. Highly variable data can be difficult to model. It becomes somewhat unpredictable.

**Slide 10**

* Data aggregation is the process of collecting data from multiple data sources for reporting or analysis.
* Data streams are sources of information that produce data continually. A sensor control system on a nuclear reactor will continually produce data (at some rate) to describe the health of the reactor. Sometimes data stream processing will be referred to as Online analytical processing (OLAP) or Online transactional processing (OLTP) depending on the data.
* Unstructured data has a non-standard format, for instance, like in JSON files.
* Structured data is data that has a clearly defined format – much like in XML files, CSV files, or database tables.

**Slide 11**

Once we have data, it is important that we store it somewhere safe. There is some terminology related to data storage that you be unfamiliar with.

* A database is a software system optimised for the storage and retrieval of information. Databases maybe relational (which use the SQL database query language), or non-relational.
* A data warehouse is a system designed to store vast quantities of data, collected for some specific purpose. Data warehouses are usually designed to be used optimally for data analytics, making them somewhat different from traditional databases.
* A data lake is a data storage system, or combination of systems, that contain enormous volumes of data not necessarily optimised for analytics. Such data must be pre-processed before it becomes useful. An organisation may deposit large quantities of unstructured highly variable data into a data lake, on the basis that it may prove useful at some point.
* Data governance covers the legal and ethical frameworks that must be taken into account, when responsible for the storage and processing of data. It may also define a set of processes used to ensure data integrity and security.

**Slide 12**

We now provide a visual summary of the first three areas taken together.

**Slide 13**

Once we've designed our data science experiment, we must turn our attention to data acquisition and preparing the data for use:

* Data cleaning refers to the process of modifying data so that it is ready for use by analytic methods. This may involve removing erroneous entries or duplicate records or repairing the structure of a dataset.
* Data labelling is the process of assigning correct labels to data records describing their content, so that the records can be used to teach machine learning algorithms.
* Automatic identification and capture is a process used to identify and collect data from appropriate sources automatically without human intervention.
* Data transformation is used to alter data sets so they become useable for analytics tasks. This may involve, for example, extracting structured information from unstructured data.

**Slide 14**

* Extract, transform, load is a process used to extract data from large data lakes. Here extraction involves retrieving raw data from an unstructured data pool and migrating it into a temporary staging area for transformation. During transformation the data is converted into a format amenable to further processing. Once the data is transformed in to some structured format, it is loaded into a data warehouse. This approach is often compared to Extract, load and transform, which transforms the data once loaded in to some database or analysis tool. This approach is not practical for large datasets, while ETL is.
* Duplicate elimination involves removing duplicate entries in datasets.
* Data pre-processing is a catch all term often used to describe multiple processes applied to data before it’s ready for use. This may involve cleaning, transformation, and duplicate elimination.
* Data munging is the process of transforming and mapping data from one "raw" data form into another format so it is ready for use. This process is similar to pre-processing.
* Data wrangling is a synonym for data munging.

**Slide 15**

To conduct our data science experiments, we need tools. There are some obvious tools we may need, a computer for instance. But there are some specific terms used to describes tools in this area.

* Data mining tools are an abstract set of software systems and statistical methods used to extract knowledge from data.
* A data science Stack, or data science software stack, is a collection of software tools used to carry out the various stages of the data science process, usually from start to finish. This includes tools capable of storing, transforming, modelling, analysing and visualising data. We can see some examples here in the image on the left of the slide. We can see database tools such as MySQL, Python for coding tasks, and machine learning libraries for modelling. Companies usually develop their own stack. If you hear to the term-full stack, then know that this refers to a collection of software supporting all aspects of data science - not just an individual task like visualisation.

**Slide 16**

There are some terms arising from the field of statistics related to experimentation.

* Linear regression – a statistical method used to model the relationship between variables in an experiment.
* Sampling – a process used to select data for analysis.
* Cross validation – a method for evaluating machine learning algorithms or for evaluating statistical results, to ensure the results are robust and not specific to a sampled data set only.
* Hypothesis testing – a statistical framework for checking and interpreting available evidence related to a hypothesis.
* A/B Testing – a form of randomised experiment that splits test subjects into control and experimental groups. The results of the experiment are typically evaluated using hypothesis testing.

**Slide 17**

There are some terms arising from the field of machine learning related to experimentation.

* Supervised Learning – a form of machine learning capable of making predictions or grouping data. These methods are given information to learn from by some form of teacher which can be as simple as labelled data.
* Unsupervised Learning – similar to supervised learning except there is no teacher. Instead unsupervised methods learn from a collection of unlabelled examples. These methods learn to group or predict using just the inherent structure of data.
* Semi-supervised Learning - a form of machine learning that combines supervised and unsupervised learning.
* Classification - the process of assigning predicted labels to data points. Usually done using automated algorithms.

**Slide 18**

There are a couple of types of analysis worth knowing about.

* Exploratory analysis is as approach to studying data that uses summary statistics – such as the mean or standard deviation, to describe the main characteristics of the data. This allows general patterns or trends to be identified quickly.
* Correlation analysis is an approach used to measure the significance or strength of a relationship between one or more variables in a dataset.
* Comparative analysis compares data sets or variables side-by-side, allowing differences between them to be quickly found.

**Slide 19**

We’ve now introduced many of the most common terms you’ll come across in data science. I’m sure you are now experiencing information overload, and perhaps can’t connect these terms to any intuitive understanding yet. We may come across these terms again as we move forward, helping us to build a better understanding.

**Slide 20**

There are lots of terms floating around when it comes to data science. We simply can’t cover them all here. So here are some resources that you may find useful if you come across an unfamiliar term.

The data science glossary is a fairly complete resource that you can turn to: <http://www.datascienceglossary.org/>

While companies like Microsoft have provided glossaries describing machine learning and data science terminology: <https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/glossary>

Then there are some other useful online discussions, such as this: <https://towardsdatascience.com/the-a-z-of-ai-and-machine-learning-comprehensive-glossary-fb6f0dd8230>