

Module 1: Introduction to Data Science - Part 2

Introduction

This module consists of two parts.

- **Part 1** - Setup and Installation Instructions
- **Part 2** - Introduction to Data Science

This module part provides a general overview of what constitutes Data Science, and the trends affecting it.

Learning Outcomes

In this module, you will learn the following.

- The history of Data Science and its transition from traditional analytics
- The difference between data analyst and scientist roles
- The major technology shifts driving modern Data Science
- Current Data Science applications and use cases in industry

Readings and Resources

There are no recommended readings and resources for this notebook.

Table of Contents

- [Module 1: Introduction to Data Science - Part 2](#)
- [Introduction](#)
- [Learning Outcomes](#)
- [Readings and Resources](#)
- [Table of Contents](#)
- [Introduction to Data Science](#)

- A Brief History of Data Science
- The Evolution of Analytics
 - Analytics 1.0: Traditional Analytics (up to mid-2000s)
 - Analytics 2.0: Big Data (mid-2000s)
 - Analytics 3.0: Prescriptive Analytics (emerging now)
- From Data Analysts to Data Scientists
 - Traditional Analysts
 - Data Scientists
- Data Science is Multidisciplinary
- Big Data
- Predictive Modeling
 - The Predictive Modeling Process
- Data Mining
- The Analytics Pipeline
- Machine Learning
- Artificial Intelligence
- Applications of Predictive Modeling
 - Applications in Retail
 - Segmentation
 - Recommenders
 - Market Basket Analysis
 - Churn Prevention
 - Applications in Financial Services
 - Fraud Detection
 - Financial Markets
 - Applications in Healthcare
 - Diagnosis
 - Drug Discovery
- In the Coming Weeks
- References

Introduction to Data Science

“Data Science” is a fairly new term for a profession that applies the scientific method to analysis of data, and in particular, Big Data. Collecting, storing, and making sense of Big Data (another fairly new term) is quickly becoming part of every business and everyone’s life.

A Brief History of Data Science

These are a few of the key trends and milestones that led us to modern Data Science:

Year(s)	*Trends*
1960s-1970s	Rapid advances in statistics and computer science
Late 1990s	Google invented a new search engine combining math, statistics, data engineering and computation
2000s	Hard drives and computational power (particularly floating point computation using graphics cards) continued to become less expensive
2001	The term <i>Data Science</i> was coined
2002	CODATA Data Science journal launched
2003	Columbia University began publishing <i>The Journal of Data Science</i> . Google File System paper spawns <i>Hadoop</i>
2006	Moore's Law came to an end resulting in a shift towards parallel processing on multiple CPUs
2010s	Rapid development of Deep Learning (Deep Learning, nd)

The Evolution of Analytics

Analytics 1.0: Traditional Analytics (up to mid-2000s)

Traditional Analytics was primarily reactive:

- Primarily descriptive and focused on reporting
- Internally sourced, relatively small, structured data
- "Backroom" teams of analysts
- Internal Systems of Support

Analytics 2.0: Big Data (mid-2000s)

Modern Analytics involves sophisticated modelling and in some cases requires complex data streaming and processing infrastructure:

- Complex, large, unstructured data sources
- Stored and processed rapidly, with new analytical and computational technologies
- "Data Scientists" emerge
- Online firms create data-based products and services

Analytics 3.0: Prescriptive Analytics (emerging now)

We are entering a new era in which predictive models will guide business leaders in real-time decision-making:

- An environment which combines Analytics 1.0 and 2.0 that yields insights with speed and impact
- Analytics is integral to running a business and becomes part of strategy and operations
- Predictive and Prescriptive Models: models that recommend specific actions
- Artificial Intelligence techniques such as Deep Learning and Reinforcement Learning

From Data Analysts to Data Scientists

The role of the data specialist has been evolving as well:


Traditional Analysts

Traditional data analysts can be found in many corporate departments, and in particular Finance, Marketing and IT (Business Intelligence). They tend to be reactive, responding to specific requests for information or developing and overseeing regular reporting. They typically use proprietary (i.e. not open source) tools such as *Excel*, *SAS*, *SPSS*, *Cognos*, etc. and have strong *SQL* database query skills.

Data Scientists

Modern data scientists have a preference for open source tools such as *Python* and *R*, in addition to traditional toolsets. For large volumes of data they typically use a Big Data analytics system such as *Spark*. Once you learn the foundations of Data Science and gain hands-on experience with some of these tools you can easily transition to others. The underlying skills are the same.

Data Science is Multidisciplinary

 Data Science is oriented at the intersection of a large number of traditional disciplines. Currently, we recognize that Data Scientists encompass expertise from the fields of mathematics, statistics, machine learning and data mining, business, software engineering, data engineering, programming, visualization, storytelling and subject area expertise.

Data Science is oriented at the intersection of a large number of traditional disciplines. Currently, we recognize that Data Scientists encompass expertise from the fields of mathematics, statistics, machine learning and data mining, business, software engineering, data engineering, programming, visualization, storytelling and subject area expertise. (Course Authors, 2018)

Data Science draws on a wide variety of disciplines, skills and bodies of knowledge.

Discipline/Skill/Knowledge	*Reasoning*
Mathematics	Data Science uses mathematical techniques to model phenomena in the world and business to gain useful insights and couple them with a measure of confidence in any conclusions drawn
Statistics	Statistical procedures enable conclusions to be made or insights to be taken from large volumes of data by looking for correlations (relationships) between variables
Machine Learning and Data Mining	Techniques for automated detection of correlations are the "core technology" of data science
Business	In a business setting, all of the usual business-related skills are essential. (<i>i.e., the ability to identify tradeoffs between investment in data science staff and tools vs. likely payoffs, working with other departments to identify potential data science projects and quantifying their value, negotiating for availability of resources, etc.</i>)
Software Engineering	Often predictive models developed by a corporate data science team will be deployed into the data processing infrastructure of the company, which means they must consider tradeoffs such as throughput and latency in addition to predictive power
Data Engineering	When dealing with "Big Data" (defined below), new scalable data streaming and distributed database technologies come into play, along with new issues and strategies for dealing with them
Programming	Although the machine learning part of data science is mostly done using tools that have been developed by specialists, there is a lot of work to get a dataset prepared for analysis, and this often requires writing custom scripts to clean up the data
Visualization	Often relationships in data are easier to see in pictures than numbers and can be much more convincing
Story Telling	A data scientist must be able to tell a story about how their conclusions fit together into the bigger picture, especially if some kind of action must be taken by others as a result
Subject Area Expertise	Data science is a tool in service of business, social goals or other branches of science; all of your existing skills and knowledge will help you identify worthwhile projects to turn the data science lens towards

Data Scientists are rarely deep experts in more than one or two of these areas but must have competence in at least basic statistics and a working knowledge of most other areas. Math and statistics are core skills but a pure degree in statistics isn't usually necessary. Most data scientists come from a *STEM* (*Science, Technology, Engineering and Math*) background but artists with expertise in visualization and storytelling are equally important.

Big Data

Big Data is a relatively new term — however collecting, storing and analyzing data is centuries old. The concept gained momentum in the early 2000s when industry analyst *Doug Laney* articulated the now-mainstream definition of Big Data as the three V's: *Velocity, Variety and Volume* (Downey, 2015).

Data becomes "Big" when the rate at which it arrives, the variety of data formats you need to deal with, and/or the quantities to be processed grow to a level at which a single-server processing solution will no longer be adequate.

Relational databases can't be distributed across more than a few servers, and synchronization between them becomes near impossible when dealing with large or globally distributed data sources. Big Data inevitably involves new technologies designed for distributed, eventual data consistency systems and so a new branch of *Data Science* known as *Data Engineering* has emerged.

Others have extended Laney's three V's in various ways, for example:

 This image shows the extensions of Laney's V's as a diagram.

This diagram shows extensions of Laney's three V. (Course Authors, 2018)

Occasionally you may see references to another V, "Veracity" — which essentially means truthfulness — and can be considered to encompass correctness and accuracy.

In this diagram we see the original three V's enhanced with another three key characteristics:

Additional Characteristics	*Reasoning*
Variability	In the online world, system usage can go from minimal to massive demand in a matter of seconds and should be able scale up and down quickly in response
Complexity	The various kinds of data available and the variety of ways it can be encoded with descriptive tags has exploded
Value	Does acquiring this data represent good value for the costs involved?

Predictive Modeling

Predictive Modelling is a process of creating a model that, given a set of inputs, will produce a plausible output similar to some process it is meant to mimic. Predictive models are usually statistical in nature. We don't usually know the precise mapping of inputs to outputs or what

all of the input variables are — much less their current values. So we can't define a precise mathematical function but there may be significant value in having a good approximation.

The development of the discipline of Machine Learning gives us tools for discovering such approximations by analyzing the patterns in large datasets. With a large enough dataset of examples of real inputs and outputs these tools can learn a decent input-to-output mapping.


For example:

- Which behaviours (input) are correlated with subsequent fraud (output)
- Which photos (input) do or do not contain a picture of a goat (output)

The rise of the internet has enabled collection of huge datasets that make this kind of analysis feasible and has led to recent advances in areas such as facial recognition.

The Predictive Modeling Process

The process of developing a predictive model involves several steps.

 This image shows the stages of creating a model from start to finish. (Course Authors, 2018)

This image shows the stages of creating a model from start to finish.

Steps	*Description*
Define Problem	Especially in a business context, the best results in terms of value-for-money come from tackling a <i>well-defined</i> problem. The solution to the problem may not be evident at first and some exploration in terms of data collection and analysis may be required, but it is best to have a specific objective in mind.
Prepare Data	This involves finding the data you need (which could be from several sources), confirming its suitability, and transforming it into a format amenable to subsequent analysis.
Create Model	This step involves an initial analysis of the dataset to find useful patterns and selecting one or more modelling approaches. Data scientists will also need to confirm the toolset and which methods to use. If machine learning techniques are used, this stage will also involve <i>training</i> the model on a dataset of previously-observed results for a given set of inputs for the process being modeled.
Test Model	This phase involves confirming that the model is operating as expected.
Validate Model	In this step, the model is presented with test cases of inputs and outputs similar to those used for training the model to confirm that it makes predictions that are similar. As we will see later, we almost always want the model to <i>generalize</i> and not rote-learn an input-to-output mapping. Validating with data that <i>was not used to train the model</i> will help avoid developing models that have not learned to generalize.

Steps	*Description*
Evaluate Model	Data scientists will often develop several models and compare their performance to identify the best. Counterintuitively, sometimes the best model is a weighted average of several models (an <i>ensemble</i> of models) that individually are not the best performers.
Deploy Model	In a business context, this involves integrating the model into existing business processes and utilizing newly received data to update the model's predictions to generate business value.

Data Mining

Sometimes the focus is less on developing a predictive model and more on discovering relationships within a dataset that represent new and valuable insights. The term Data Mining is used to refer to these scenarios.

Some examples include:

- Using a large dataset of network events captured from a credit card processing system and identifying patterns that represent attempts by hackers to break in
- Identifying behaviours of smartphone customers indicating they're leaving for a competitor, so that the customer relationship team can proactively attempt to retain them

These insights could also be used to formulate a predictive model.

The Analytics Pipeline

 This image shows the major stages of producing an analysis. (Course Authors, 2018)

This image shows the major stages of producing an analysis. (Course Authors, 2018)

Many data-intensive companies have now deployed an Analytics Pipeline similar to the one shown in the diagram. The details vary from organization to organization, but generally follow this broad pattern.

Pipeline Stage	*Description*
Data Sources	<p>Data arrives from a variety of systems and providers. Increasingly, this includes the <i>Internet-of-Things (IoT)</i> — devices attached to the Internet (as opposed to people directly). Some examples are:</p> <ul style="list-style-type: none"> • Cellphones generate a huge amount of streaming data about people's location, app usage, phone usage, etc.

Pipeline Stage	*Description*
	<ul style="list-style-type: none"> • In auto insurance and fleet management, real-time information about car and truck drivers' behavior is increasingly being monitored • Sensors in modern tractors continually monitor the soil conditions for farmers and are sent to agricultural product manufacturers who use the information to recommend improvements to drainage, pesticide and fertilizer use <p>A second source of data is a company's internal systems:</p> <ul style="list-style-type: none"> • Banks use internally-sourced data aggregated from across their many lines of business to create a consolidated risk management view • Retailers scour sales data from their <i>point-of-sale</i> terminals to spot trends and behaviors they can use to drive profitability • For large tech companies (<i>Google, LinkedIn, etc.</i>) their users are a massive source of real-time data about opinions, trends and behaviors at the individual level • And, in our connected society, companies increasingly share data from outside their walls: with trading and supply chain partners, and by purchasing data from 3rd party data providers
Data Capture and Processing	<p>The arriving data needs to be captured and saved for subsequent analysis. Organizations are increasingly capturing data in its raw form in a database optimized for fast data input. This is a departure from the past where data was often carefully validated and preprocessed prior to storage in a near-normalized form in a relational database. In some applications today (especially Web or IoT-based) the arriving data rates are so high that data capture using relational technology is infeasible. Relatively new database technologies such as <i>Hadoop</i> were designed for this purpose: very fast writing of large volumes of data. A database used this way, as a capture area for raw data is often called a <i>data landing area</i> or <i>data lake</i>.</p> <p>Another reason for storing raw data is that preprocessing assumes a particular use for the data and as a result throws away some of its content. If the data is stored in its raw form, all of the information content is preserved in case additional forms of analysis might require it. Once the data has landed, data scientists can use it to build models or conduct enquiries, often transforming the data to improve its quality or make it more amenable to analysis. Transformations and models that prove to be useful over the long term can be automated and deployed as on-going processes that produce regular outputs.</p>
Analysis Results	<p>The outputs of this overall process include ad-hoc and on-going analysis, predictions, standing reports, and insights into the data that may lead to business process improvements or better margins.</p>

Machine Learning

Throughout most of the twentieth century, the standard approach for building a predictive model started with identifying the features of the business, social or economic process of interest that were likely to have some predictive power. In other words, we identified the independent variables (inputs) to the model in advance of building the model. This approach

is known as *Feature Engineering* and is still important. But we now have many tools to assist us in finding predictive variables empirically.

Machine Learning is a term used to describe the automatic extraction of knowledge from data. By knowledge, we mean a set of generalizations that can be made about the data. These generalizations can be hard rules, statements that appear to always be true (at least within a given dataset), or soft rules that are approximately or probabilistically true. Suppose, for example, you knew nothing about the rules of the road. Analyzing a dataset of nothing but traffic patterns might allow you to infer that driving only one way on some streets is a hard rule (unless you have a dataset that's big enough to contain an example of someone doing it). Speed limits would become apparent as a very soft rule (in fact, a distribution around a speed is no doubt a little above the posted limit).

There are a variety of Machine Learning algorithms that have been developed for this purpose, each with their own strategy for extracting and representing knowledge: some as actual rules, some as tree structures, some as statistical distributions, often as combinations of these. The process of extracting the knowledge from a dataset is called *training* or learning a model.

Each algorithm has its own performance characteristics. Some are fast to train but slow to compute a prediction whereas the opposite is true of others. Some are more resilient to noise in the data than others. Paradoxically, simple models often perform as well as highly sophisticated ones for many business applications. We will encounter many of these algorithms in the courses in this series.

Artificial Intelligence

You may be wondering about the difference between Machine Learning and *Artificial Intelligence (AI)*. Machine Learning is a particular set of knowledge and techniques within the much broader topic of AI.

- AI comprises the entire question of whether and how a machine can show behaviors that are sufficiently human-like to be indistinguishable from the real thing. It includes general theories of knowledge representation and strategies for searching for solutions to complex problems in a large space of alternatives.
- Machine Learning is a subset of AI that is focused on extracting knowledge from observed data with little human guidance whereas AI more generally allows for incorporating generic or hand-crafted problem-solving strategies.

Applications of Predictive Modeling

Let's take a look at the myriad applications of Data Science and Predictive Modeling. Here are a few industries we will focus on:

- Retail
- Financial Services
- Healthcare

Applications in Retail

Segmentation

Customer segmentation is a type of predictive model. For example, we would like to predict which segment a customer is likely to fall into, so that we can tailor our marketing strategy to meet their needs and increase their satisfaction.

Why segment markets? Customers may differ in:

- What they want to buy
- Amount willing to pay
- Quantity they buy
- Time, place, and frequency of purchase
- Personal taste (likes and dislikes), for example in:
 - Media
 - Phone plan
 - Newspapers
 - Magazines
 - Movies
 - Social media platform

If a retailer knows that you belong to a segment that shops often, and on average spends \$100 per transaction, their strategy may be to try to "stretch you" to spend \$125 on your next purchase in the hope that it will become a new pattern. They could do this by offering you a coupon, or a certain promotion (spend \$150 and get \$25 back).

Some customers may even represent negative value: ones that come in and scoop up advertised loss leaders but don't buy anything else while in the store. A retailer may devise strategies to minimize traffic from this segment.

Recommenders

Online retailers in particular have developed sophisticated methods for recommending products to customers based on their online history, interests or similarity to other customers. Online recommendation engines seek to match consumers behavior profiles with those of others and to recommend products that similar customers have bought in the past.

Market Basket Analysis

Market Basket Analysis is a technique where retailers look for correlations in items that consumers tend to buy at the same time. This kind of analysis can be used for recommendations, but also as a way of maximizing profit. Some items naturally go together, like burgers and condiments, so a grocery store could advertise a special on ground beef yet ensure all condiments are at full price that week. Retailers continually mine stored data for more subtle associations between items and use the insights to inform their pricing strategy.

Churn Prevention

It's an old saying that it takes about ten times as much money to acquire a new customer as it does to retain an existing one. Retailers and services providers such as Telcos are deeply interested in ways to boost brand loyalty or intervene meaningfully when a customer is considering switching to another brand. As part of churn prevention, some companies mine the data they have on the behavior of their customers or their comments on social media to determine if customers are considering competitors.

Applications in Financial Services

Fraud Detection

Fraud detection is serious business in the online world. Online marketplaces perform real-time fraud detection to decide whether to allow transactions to take place between parties. Modern fraud detection systems consider a wide variety of factors such as:

- The location of the parties to the transaction
- Whether a user has a new account but is using the app's features like an experienced user
- Whether an account is linked with other accounts in a fashion that would facilitate money laundering
- With whom the parties have other relationships
- Whether the names and addresses of the parties are credible

Financial Markets

Attempting to "beat the market" is as old as exchanges and people continue to build increasingly more sophisticated predictive models to attempt to do so. Some also incorporate other AI techniques to try to discover positive or negative impacts from electronic news releases and execute trades faster than people can react.

Applications in Healthcare

Diagnosis

Startups are developing diagnoses for various kinds of cancer by analyzing skin photos or breath samples, using predictive models trained to recognize telltale signs. *Deep Learning*, a branch of Machine Learning that has revolutionized face and object recognition in photos, is now being used to analyze X-ray photos for signs of disease.

Drug Discovery

Startups are using Machine Learning to identify potential drug candidates by analyzing patterns in the chemical properties of previously successful and unsuccessful drugs. Lab testing of compounds is a slow and costly process which can be dramatically accelerated by focusing the search on compounds most likely to have desirable properties.

In the Coming Weeks

We will be covering a broad introduction to Data Science over the next few weeks. Throughout the course (and overall program) we will be working in the *Python* programming language, which has become the number one choice for Data Science, slightly ahead of *R*. We will begin by learning the basics of Python, then focus on the **Numpy** (Numpy developers, 2018), **Pandas** (Pandas community, 2018), and **Scikit-Learn** (Scikit-learn developers, 2018) libraries for Statistics and Machine Learning. As you build your skills, you will use them to characterize datasets and begin building simple predictive models using these tools.

End of Module

You have reached the end of this module.

If you have any questions, please reach out to your peers using the discussion boards. If you and your peers are unable to come to a suitable conclusion, do not hesitate to reach out to your instructor on the designated discussion board.

When you are comfortable with the content, and have practiced to your satisfaction, you may proceed to any related assignments, and to the next module.

References

Deep Learning, (nd). wikipedia, the free encyclopedia. Accessed Sept 4, 2018. [online](#)

Downey, A. (2015). Think python--how to think like a computer scientist. [online](#)

NumPy developers, 2018. Numpy. [online](#)

Pandas community, 2018. Pandas. [online](#)

Scikit-learn developers, 2018. Scikit-learn: machine learning in python — scikit-learn 0.19.2 documentation. [online](#)

In []: