Hands-On Data Science with R
Data Science and Analytics

Graham.Williams@togaware.com

23rd August 2014

Visit http://HandsOnDataScience.com/ for more Chapters.

Today we live in a *data* rich, *information* driven, *knowledge* strained, *wisdom* scant world. Data Science is a broad church capturing the endeavour of analysing *data* and *information* by appropriately applying an ever changing collection of tools and technology, to deliver *knowledge* to be synthesized into *wisdom*. The role of a Data Scientist is to perform the transformations that make sense of it all in an evidence based endeavour.

The Data Scientist's skill set augments the humanity and philosophy used to synthesize knowledge into wisdom—they thrive to *resolve the obscure into the known*. It is this that delivers the real benefit of the science—whether that benefit be for business, industry, government, or understanding the human condition, through science.

A Data Scientist, then, is an exceptional person who brings a specific collection of skills together with particularly strong intuitions. They have a desire to continually learn, improve and discover and to improve how things are done. Find all these requisite technical skills and motivation in one person is a pretty tall order. Data Scientists are scarce and the demand for their services continues to grow.

In this brief introductory chapter we present the concepts of Data Science, Analytics, and the role and toolkit of the Data Scientist. We set the foundations for understanding. The remaining chapters of the book—indeed, the bulk of the book—provide a guide to the key tool for doing Data Science, called R.

1 The Art of Data Science

> Science is analytical description, philosophy is synthetic interpretation. Science wishes to resolve the whole into parts, the organism into organs, the obscure into the known. Durant (1926)

In Data Science we ply *the art of excavating data for knowledge discovery* (as the sub-title of the book "Data Mining with Rattle and R" (Williams, 2011) puts it). Whilst we ponder on the art that is Data Science, we might reflect on the observations of the American philosopher, Will Durant, who proposes that *every science begins as philosophy and ends as art* and that *science gives us knowledge, but only philosophy can give us wisdom* (Durant, 1926).

A distinguished scientist is a truly exquisite artist. And so it is with Data Science. One of my themes over the past 30 years of teaching Computer Science, Databases, Artificial Intelligence, Machine Learning, and now Data Mining and Data Science, has been that what we do is an art—it is an expression in a form that communicates to others. Our role is not to simply program computers to do things for us, but to express what it is that we wish the computer to do in an elegant form, to communicate, not (only) for the computer to execute, but for others to read, to marvel, and to enjoy.

So it is with Data Science. As I hope will become evident through the pages of this book, as Data Scientists, we aim to not only turn data and information into actionable knowledge, but we also aim to clearly communicate in such an elegant way so as to *resolve the obscure* and to make it *known* in a form that is a pleasure to read—in a form that makes us proud to share, want to read, and to forever learn.

2 Data Scientist

Specialists who migrate from being Data Technicians (skillful SQL coders) to Data Analysts (who add value to the extracted data), then on to understanding Machine Learning and Statistics (where we begin to understand and model the world based on the data), and able to program with data, will often have the technical know-how and the requisite skills and can practise Data Science admirably. They will ply their tools to increasingly larger volumes, velocity and variety of data to deliver beneficial outcomes.

The next stage of the art form though, and one that is not so easily taught, is the intuition that drives the process of exploring through our data to discover the unknown. It is also the philosophy that we bring to bear on the knowledge we discover, and synthesise in different ways to give us the wisdom to decide how to act. And finally the continual desire to challenge, grow and learn in the art, and to be driving the future, not being pushed along by it. The technical Data Scientist that brings these less tangible skills to the table are those that are most sought after, and indeed are quite rare.

Nonetheless, many prefer to remain within their technical comfort zones and have plenty to offer in delivering basic benefits to organisations with their data management, statistical analysis and model building. It is these technical skills that we cover in this book to build the foundation for the Data Scientist.

So what is a technical Data Scientist? Three core skill categories make up a technical Data Scientist:

1. Programming and Software Engineering;
2. Machine Learning;
3. Mathematics and Statistics; and

Over the last 30 years as researchers and professionals in Computer Science we have observed and, interestingly, contributed to the reduction of the skills in using computers. We seem to be turning out computer users that are driven by computers telling us what to do, rather than users able to control and direct computers to do what we require of them. This has led us today to the shortage of skills that are in demand in areas such as Data Science.

3 Creating an Analytics Capability

Creating an Analytics Capability need not be expensive, despite incredible market pressures to make it so. The expense should be in acquiring expert Data Scientists and providing them with the tools they require and request. Unfortunately, in many traditional organisations we see it instead focused on delivering centrally (ICT) controlled platforms on large and expensive computers running the singularly vetted and extremely expensive statistical software suites. It is odd how funding models prefer the massive expense over the otherwise cost effective reality of the Data Scientists toolkit.

The key message is that in setting up an Analytics capability the focus must be on Analytics—strangely enough. Despite the obvious, this is quite a challenge. The Analyst First movement has been promoting this message for some time. It is interesting to read some of its core principles:

- All businesses can create an Analytics function rapidly, cheaply and with a small footprint;
- Analytics, done properly, is scalable;
- The Analyst is the most essential, valuable and rare resource in Analytics;
- The Analyst is the focus of successful Analytics investment;
- Analytics is not IT;
- Analytics requires advanced and flexible software and hardware;
- It not fit into the "standard operating environment";
- It separate, self managed infrastructure.

It is perhaps not surprising that large organisations struggle with deploying Analytics. Traditional ICT departments tend to insist that they drive the infrastructure provision of the organisation. This has been their role traditionally. But unfortunately the world has moved on and they tend to struggle with understanding this and the resultant control that they lose in so doing. Yet, without understanding, they will waste millions for the organisation in purchasing infrastructure, software and hardware, that will not be used effectively for Analytics.

4 Analytics: Build or Buy

Consider the role of the ICT department in acquiring an Accounting capability. Accounting packages mostly provide a common suite of well established capabilities that have changed little over the past two decades supporting techniques that have been with us for centuries. Perhaps the different software tools are packaged in different ways with different levels of ease-of-use. Vendors are invited to tender for the organisation's requirements, and the ICT department will review and choose the appropriate software package. The chosen package is deployed, and accountants will then be employed to do the company accounts. Not a lot changes in how accounts are done from year to year. This approach puts the software first.

This approach simply misses the point for the dynamic and innovative modern technology require for Data Science. The required capabilities are changing, perhaps not quite daily, but sometimes weekly or monthly. Thus instead we need infrastructure (hardware and software) that is agile and can be updated as required, regularly. It thus needs to be cost effective, since we need to be willing to replace it when needed, and not to be tethered to a single vendor. No one vendor has all the answers.

For an Analytics capability, we see examples where traditional organisations can spend up to two years educating themselves about Analytics. They will invite a multitude of vendors, each with their tool suite to sell, to present to the organisation, presneting to up to 20 or 30 ICT staff. In the process, the hope is that the staff of the ICT department will be educated about Analytics. In the meantime the vendors' sales staff will also be attending courses on Analytics to build their capabilities in speaking the language to the clients. Meanwhile, what's missing is the actual hands-on experience of actually doing Analytics. Yet, it is surely that experience that should be driving the infrastructural requirements, rather than the vested interest of the vendors!

5 Closed Source versus Open Source Software

Irrespective of whether software can be obtained freely through a free download or for a fee from a vendor, the important point is that we need our software source codes to be open. That is, we need to have the freedom to be able to review the source code to ensure the software implements the functions correctly and accurately, and to allow us to change and enhance the software to suit our ever changing and increasingly challenging needs, and to allow us to build on and share our developments with others who also have the software.

Today's Internet is built on open source. Most web servers run the open source apache software. Nearly every modem/router is running the open source GNU/Linux operating system. There are more installations of the open source Linux operating system running on devices today than any other operating system ever—Android is an open source operating system running the Linux kernel and today has an installed base far outnumbering all competitors.

Commercial software is typically closed source, presenting challenges to the effective use and reuse of that software. Instead of being able to build on the shoulders of those who have gone before us, we must reinvent the wheel, over and over again—this can't be good for the advancement of humanity though it is undoubtedly financially attractive the shareholders of the vendors.

Gartner has also observed: *this*.

6 Fear, Uncertainty and Doubt

Vested interest in a software product leads salesmen to oversell their interests and undersell or overwhelming dismiss contrary interests. We do need to be careful in evaluating software offerings, and particularly closed source software offerings where the vendors hide their offerings so that we actually can only trust their claims.

An interesting article titled *What Immigration did with just $1m and open source software* was published 6 August 2014, written by John Hilvert for itnews. It quotes the Australian Department of Immigration's Chief Risk Officer Gavin McCairns reporting on their use of R for Analytics. Using open source software (including R) on inexpensive platforms, the organisation's Data Scientists deliver sophisticated risk models to protect Australia's borders. The telling point made by McCairns is that "the department bought $15 million worth of software—but it's gathering dust." Instead, for much less and using freely available open source software, they implemented and deployed a real time passenger risk assessment system.

One is left to speculate that the "$15 million worth of software" is a particular closed source statistical software suite for Analytics. A senior sales person for the product is sought out to respond. Despite the evidence just presented, their response is that "R in a production system, it can be scary," and that "R has severe limitations when it comes to real time transactions." As one of the article's commentators puts it: *very scary for proponents of proprietary software... that their turf is being eroded by more competive, more secure, more perfoming, cost effective software.*

An earlier example of this kind of fear, uncertainty and doubt being spread about open source was similarly provided by a marketing executive of one of the closed source vendors in the New York Times, 6 January 2009, in the Business Computing section, in an article by Ashlee Vance titled *Data Analysts Captivated by Rs Power*. They claimed "We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet." This only illustrates how little understanding there is of the reliance today we have on open source software. It also of course misses the point that it is not whether the software is "freeware" but whether it is open source.

Commentators readily pointed out that any one of the world's leading statisticians are able to review every line of code available in the open source statistical offerings. By doing so they can confirm or challenge the implementations of critical algorithms. On the other hand, only a very few coders have access to the closed source software, and how is it that we can trust just a small number of eyes reviewing the code? It is a truism that all software contains bugs, irrespective of whether it is closed or open source—just that open source software has more of a chance of the bugs being identified and quickly fixed.

7 The Data Scientist's Toolkit

The open source, and freely available R Statistical Software today provides all of the analytical capabilities required for turning data into information and then turning that information into knowledge. It is one of the tools in our toolkit. We often combine the data processing and statistical tools of R with the powerful command line processing capabilities of the Linux ecosystem. It is interesting to note that the origins of R are the S language which originated in Bell Laboratories in 1976 at the same time that the Unix operating system, on which Linux is based, was developed Chambers (2008).

The complementary nature of the statistical language and the operating system combine to make R on Linux a powerful combination that there is currently no other close contender for the Data Scientist's toolkit. A preferred and modern and easily administered Linux distribution is Ubuntu.

The growing toolkit of big data capable algorithms also include the Hadoop-based systems, and in particular the Spark package that implements algorithms to operate over the Hadoop distributed file system.

R can be used as the front-end to access these systems. It can also be used to access all of the Weka algorithms. And much more.

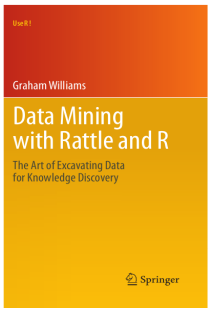
8 Using R in Other Data Science Products

An investment in learning R is an investment in accessing much of today's **and** tomorrow's technology. All major data warehouse vendors, business intelligence software vendors, and statistical software vendors now provide linkages into R or integrate R directly within their products.

The database and data warehouse vendors have each incorporated connectors to R that allow R algorithms to be run "in-database". This includes Oracle, Teradata, Netezza, and GreenPlum.

The business intelligence vendors provide R connections. Information Builders was one of the first, incorporating a version of Rattle, known as R Stat, into their Web Focus product line. Tibco has extensive R capability, including their own efficient implementation of R. Alteryx and SAS can connect to external R processes and allow interaction with then within their native user interfaces.

9 Further Reading



The Rattle Book, published by Springer, provides a comprehensive introduction to data mining and analytics using Rattle and R. It is available from Amazon. Other documentation on a broader selection of R topics of relevance to the data scientist is freely available from http://datamining.togaware.com, including the Datamining Desktop Survival Guide.

This chapter is one of many chapters available from http://HandsOnDataScience.com. In particular follow the links on the website with a * which indicates the generally more developed chapters.

Other resources include:

- http://rattle.togaware.com
- http://datamining.togaware.com
- http://datamining.togaware.com/survivor/index.html

10 References

Chambers JM (2008). *Software for Data Analysis: Programming with R*. Springer, New York. ISBN 978-0-387-75935-7, URL http://stat.stanford.edu/~jmc4/Rbook/.

Durant W (1926). *The Story of Philosophy*. 2012 edition. Simon and Schuster.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Williams GJ (2009). "Rattle: A Data Mining GUI for R." *The R Journal*, **1**(2), 45–55. URL http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf.

Williams GJ (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Use R! Springer, New York. URL http://www.amazon.com/gp/product/1441998896/ref=as_li_qf_sp_asin_tl?ie=UTF8&tag=togaware-20&linkCode=as2&camp=217145&creative=399373&creativeASIN=1441998896.

This document, sourced from DataScienceO.Rnw revision 508, was processed by KnitR version 1.6 of 2014-05-24 and took 1 seconds to process. It was generated by gjw on nyx running Ubuntu 14.04.1 LTS with Intel(R) Xeon(R) CPU W3520 @ 2.67GHz having 4 cores and 12.3GB of RAM. It completed the processing 2014-08-23 15:42:30.