Data science and open source

Learn about open source tools for converting data into useful information

M. Tim Jones
Independent author
Consultant

09 August 2013

Data science combines mathematics and computer science for the purpose of extracting value from data. This article introduces data science and surveys prominent open source tools in this rapidly growing field.

The goal of *data science* is the extraction of useful information from a data set. Companies have recognized the value of data as a business asset for a long time. But the huge data volumes that are now available necessitate new ways to make sense of data and manage it efficiently. A growing cadre of engineers and scientists are building systems to apply data science to massive data volumes. This article introduces you to the field of data science and to open source tools that are available for today's data scientist.

Data science and data scientists

Data science begins with the collection of data. Candidates for collection can be open data or data that comes from internal business processes (for example, website statistics). Next comes *refinement*: the inventive process that reduces the data to useful information that answers specific questions. Typically, the questions define the approach to the extraction of the information. Within the collection and refinement steps are other important aspects such as data cleansing (or *preprocessing*) and data visualization.

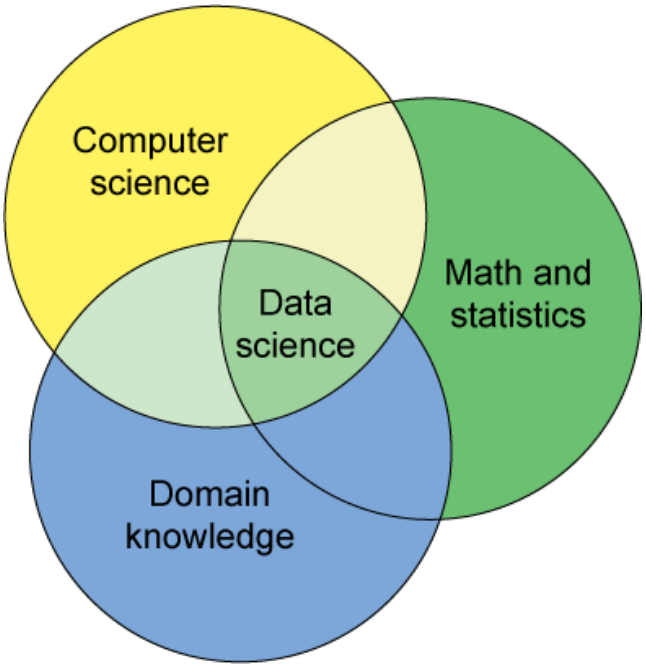
Open data

Open data is the concept of democratizing data by making it freely available to everyone to use as they want. The growing open data movement follows the ideas behind open source. A useful source of open data is Data.gov (see Resources), a US government website that was created to increase public access to data generated by the executive branch of the federal government.

You can also view data science as a business process. Mike Loukides of O'Reilly makes a compelling case that data science is the conversion of data not only into information but also into *products* (see Resources). From that perspective, the field is a modern-day gold rush — a competitive search for the valuable nuggets in mountains of information.

The prospectors in the data gold rush are called *data scientists.* As businesses recognize the value in their data, the need for talented multidisciplinary engineers and scientists is growing. Data scientists must have skills in computer science, math, and statistics. Ideally, they also have *domain* knowledge — an understanding of the source of the data (medical, financial, web, and other domains). Figure 1 illustrates data science as the intersection of computer science, math and statistics, and domain knowledge:

Figure 1. Key disciplines of the data scientist



With this complete skill set, the data scientist can translate domain knowledge and math into an application (from the computer science domain) that mines data and refines it into information. The key is a multidisciplinary focus (which can also include domains such as machine learning and information retrieval).

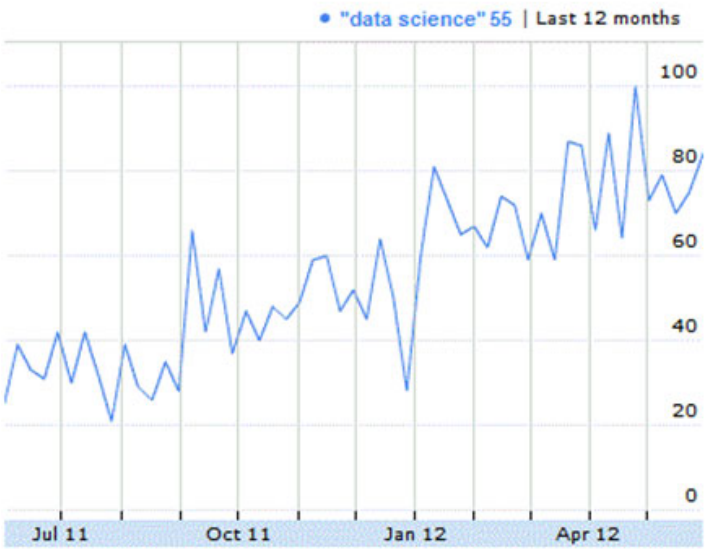
Engineers and scientists with big data analytics experience are in high demand these days. McKinsey & Company predicts that by 2018 a shortage of people who can fit the data scientist role will occur (see Resources). The ideas and approaches in data science are useful in many other disciplines too. Even if you don't aspire to become a data scientist, data science skills can be a great addition to your engineering toolbox.

Where data science is used

Like cloud computing, data science is rapidly gaining interest and adoption. Over the year before this article was written, interest in data science roughly doubled, according to Google Insights for

Search (formerly Google Trends). Google Insights for Search is itself an example of data science in action. Figure 2 shows that the frequency of *data science* as a web search term increased dramatically between the summer of 2011 and the spring of 2012:

Figure 2. Google Insights for Search data on interest in data science



Data science is quickly becoming a staple within organizations that harvest data online (be it crawling-based collection or internal collection that is based on user behaviors such as clicks). Major websites such as Google, Amazon, Facebook, and LinkedIn all have their own data science teams to use their available data (see Resources).

Google's development of the PageRank algorithm is an early example of data science. Google crawls the web and assigns a numerical weight to the hyperlinks on every page to measure the relative importance of those links. (Full details of PageRank are known only within Google.) The algorithm serves as the means of ranking web content as a function of search terms.

Large online retailers such as like Amazon and Walmart use data science to try to increase sales. They generate recommendations to individual users that are based the user's product searches and past purchases.

LinkedIn, a professional networking site, maintains a huge amount of data that is related to people and their careers, interests, and connections. This massive network of data resulted in various recommendation engines (for individuals, groups, and companies) and projects that use the data at a deeper level to produce new products at LinkedIn.

One novel example of data science at a web property is the company bitly. On the surface, bitly is a service that enables users to shorten any URL to a 19-character maximum URL (which is stored permanently in bitly's data center). References to the shortened URL are redirected from bitly to the original URL. bitly can then see which URLs people shorten *and* which URLs other users click. This tactic provides an enormous amount of data that bitly (and its chief scientist, Hilary Mason) can use to generate a wealth of statistics about browsing habits. Users who are registered with

bitly can see when their shortened URLs were clicked, through which referrer (email client, Twitter, or another URL), and from which country. Businesses can also use bitly to track user behavior for a set of content.

Open source tools for data science

Just as computer programming isn't constrained to a single language or development environment, data science isn't associated with a single tool or tool suite. A rich and broad array of tools in the open source domain advance data science. They include tools that process large data sets numerically, and visualization and prototyping tools that aid in the development of complex processing. Table 1 lists prominent open source tools for data scientists and defines their roles:

Table 1. Open source tools for data science

Tool	Description
Apache Hadoop	Framework for processing big data
Apache Mahout	Scalable machine-learning algorithms for Hadoop
Spark	Cluster-computing framework for data analytics
The R Project for Statistical Computing	Accessible data manipulation and graphing
Python, Ruby, Perl	Prototyping and production scripting languages
SciPy	Python package for scientific computing
scikit-learn	Python package for machine learning
Axiis	Interactive data visualization

The list in Table 1 isn't exhaustive but instead represents some of the core elements within the data scientist's toolbox. The open source domain is also filled with highly specialized and domain-specific libraries and tools (for example, utilities for interactive map visualization and for text analysis).

Hadoop, Mahout, and Spark

The Internet creates opportunities to collect masses of data about users' behavior and habits. Apache Hadoop is the premier framework for processing massive data sets. Hadoop is important for data science because it provides a scalable framework for distributed data processing. Not all data science problems require big data processing, but Hadoop is ideal when your problem involves Internet-scale data. The Google MapReduce framework's implementation of the PageRank algorithm is an early example of data science on a big data framework. (Hadoop is an implementation of MapReduce.) Apache Pig can make Hadoop even more accessible, bringing a query language that automatically builds MapReduce applications (see Resources).

Apache Mahout is an implementation of scalable machine-learning algorithms on the Hadoop platform (see Resources). Mahout includes scalable implementations of clustering algorithms and batch-based collaborative filtering algorithms (for implementing recommendation systems).

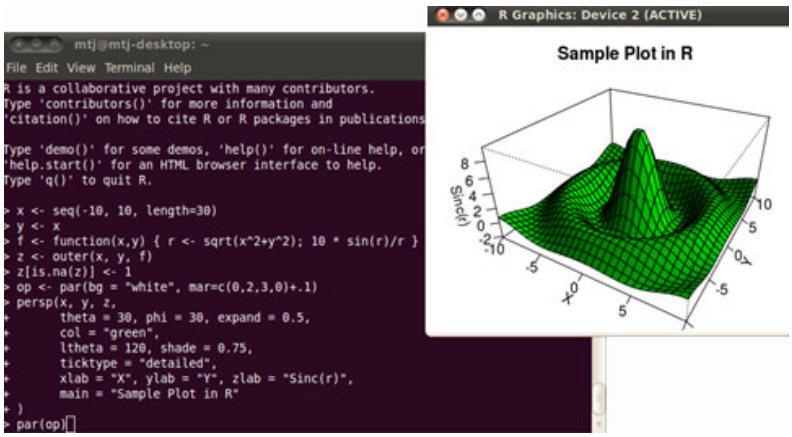
Another noteworthy solution for large data sets is the Spark framework (see Resources). Spark includes optimizations such as in-memory cluster computing with fault-tolerant abstractions.

The R project

A tool that's often found in the data miner's toolkit is a programming language and development environment called *R*. R focuses on statistical computing and graphics. R is relatively simple to learn and is widely used in the domain of data analysis. Being open source and free, R is a popular language with a large user base.

R is a multiparadigm language that supports object-oriented, functional, procedural, and imperative programming styles. The language is interpreted through a command-line interface and also includes extensive production-level graphical capabilities. Static graphics are available out of the box. With additional packages, both dynamic and interactive graphs are possible. Figure 3 shows an example plot that was generated with R:

Figure 3. Sample 3D sinc plot that uses R



The R programming language was developed in C and Fortran. Many of the internal standard functions in R were written in R itself. R supports mixed-language programming, enabling access to R objects from languages such as C and Java™. You can easily extend the capabilities of R by using *packages*, which can be developed in the R, C, Java, and Fortran programming languages.

Scripting languages

Multiparadigm scripting languages such as Python, Ruby, and Perl provide a professional platform for application development and deployment. And they are ideal for prototyping and testing new ideas. These languages also support various data storage and communication formats, such as XML and JavaScript Object Notation (JSON), and a large variety of open source libraries for scientific computing and machine learning. Python is the clear leader in this space, probably because it is the easiest to learn for users who come from backgrounds other than computer science. Knowledge of Python is often a requirement for data scientist jobs.

SciPy and scikit-learn

The SciPy package extends Python into the domain of scientific programming. It supports various functions, including parallel programming tools, integration, ordinary differential equation solvers, and even an extension (called Weave) for including C/C++ code within Python code.

Related to SciPy is scikit-learn, which is a package for Python-based machine learning. Scikit-learn includes many algorithms under the machine-learning umbrella for supervised learning (support for vector machines, naive Bayes), unsupervised learning (clustering algorithms), and other algorithms for data-set manipulation.

Both of these packages extend the capabilities of Python for use as a data science platform.

Axiis interactive data visualization

Many open source solutions focus solely on visualization. One especially interesting example is the Axiis framework, which provides a concise markup language for rich and colorful visualizations. Figure 4 shows an example:

Figure 4. Wedge stack graph visualization using the Axiis framework

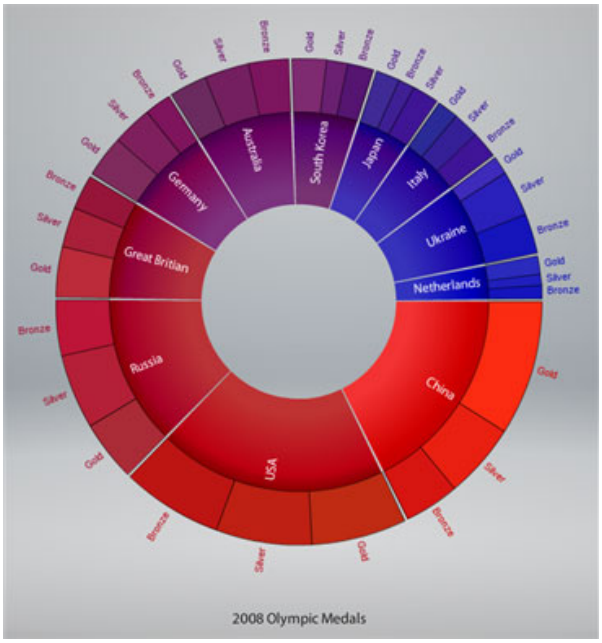


Figure 4 is a static version of an interactive example from Tom Gonzalez, Managing Director at BrightPoint Consulting. See Resources for a link to the interactive version.

Going further

The role of data scientist builds on a solid platform of knowledge and experience. But tools are also an important aspect of the data science field. In emerging disciplines, the open source community is often at the vanguard in establishing software where none existed before. The field of data science is no exception. Data science is relatively new, so more new tools, data protocols, and data formats are almost certainly in the works. But in data science, as in many other disciplines, open source solutions already lead in breadth and depth.

Resources

Learn

- Google Insights for Search: This Google site enables anyone to view search trends for a topic across regions of the world, including comparative trends of two or more topics.
- Open data: Read about open data on Wikipedia.
- "What is data science?" (Mike Loukides, O'Reilly Radar, June 2010): Read a great introduction to data science and the idea behind transforming data into products.
- "Growing Your Own Data Scientists" (Dan Woods, *Forbes*, March 2012): The article series surveys definitions of *data scientist* from leading experts in the field.
- Hadoop on developerWorks: Explore a wealth of articles and other resources on Apache Hadoop and its related technologies.
- "Apache Mahout: Scalable machine learning for everyone" (Grant Ingersoll, developerWorks, November 2011): Mahout committer Ingersoll describes Mahout's features and walks through an example of how to deploy and scale some of Mahout's more popular algorithms.
- "Data visualization tools for Linux" (M. Tim Jones, developerWorks, November 2006): This article presents several useful data visualization tools that bear some similarity to the R Project.
- Big data: The next frontier for competition: Read about research from McKinsey & Co. and on the role of big data and data scientists.
- The developerWorks Big data technical topic: Find extensive how-to information, tools, and products to guide you through the world of big data.
- Data.gov: Browse the Data.gov datasets available through the online catalog and use multiple criteria to filter your search.
- Science.gov: This portal provides access to more than 55 databases and 2,100 websites from 13 federal agencies for US government science information. As on Data.gov, you can restrict your searches by search criteria or by specific agencies.
- "Process your data with Apache Pig" (M. Tim Jones, developerWorks, February 2012): Learn more about Pig and how to put it to work in your applications.
- "Spark, an alternative for fast data analytics" (M. Tim Jones, developerWorks, November 2011): Get to know the Spark approach to cluster computing and its differences from Hadoop.
- The developerWorks Open source technical topic: Find extensive how-to information, tools, and project updates to help you develop with open source technologies and use them with IBM products.

Get products and technologies

- Apache Hadoop: Download Hadoop.
- Apache Mahout: Download Mahout from an Apache mirror.
- Spark: Get the latest Spark release.
- R programming language: Get R, a multiparadigm language and development environment with broad use in statistics and visualization
- Python, Ruby, and Perl: Simplify the development and prototyping of algorithms for data refinement with these multiparadigm scripting languages.

- SciPy and scikit-learn: Use Python's data science capabilities with the SciPy package for scientific computing and the scikit-learn package for machine learning.
- Axiis: The Axiis data visualization framework is a useful solution for both beginners and experts. Check out the examples page to see what's possible with the framework, including the interactive version of Figure 4.
- Evaluate IBM products in the way that suits you best: Download a product trial, try a product online, use a product in a cloud environment.

Discuss

- Get involved in the developerWorks community. Connect with other developerWorks users while you explore the developer-driven blogs, forums, groups, and wikis.

About the author

M. Tim Jones

M. Tim Jones is an embedded firmware architect and the author of Artificial Intelligence: A Systems Approach, GNU/Linux Application Programming (now in its second edition), AI Application Programming (in its second edition), and BSD Sockets Programming from a Multilanguage Perspective. His engineering background ranges from the development of kernels for geosynchronous spacecraft to embedded systems architecture and networking protocols development. Tim is a platform architect with Intel and author in Longmont, Colo.