Mphasis
Unleash the Next

# Data science ecosystem

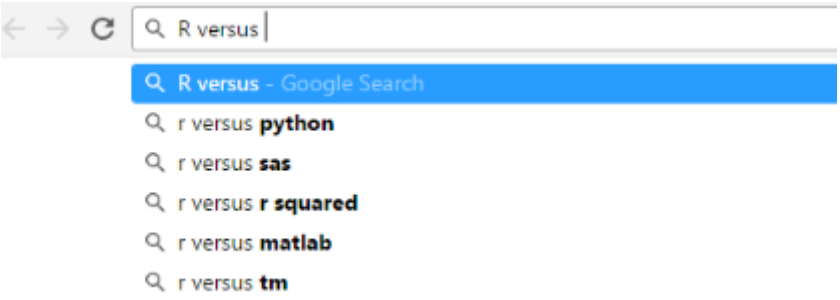
A Whitepaper by

Dr. Archisman Majumdar
Senior Manager, Mphasis NextLabs

While there are a plethora of tools and technologies in the data science ecosystem, users are often confused where to start and which combination of tools to use. In an attempt to understand this, we leverage social network analysis, which is a mathematical technique of investigating relationships among different entities with the help of networks and graph theory. Social network analysis has been gaining popularity over recent years. So, let us use this intelligence to analyze an ecosystem that comprises of the most popular data science programming languages - R and Python, and their competitors. The main aim here is to identify the key tools and languages that are commonly searched in the area of data science related to R and Python. Now the question is – how to identify these?

Well, one of the most convenient ways is to leverage the Google Autocomplete Suggestions – a well-known feature offered by Google. The moment we start typing in a search box, Google offers recommendations before we even finish typing. But do you know how does Google come up with these recommendations? No, it certainly can't read your mind, and there is no magic behind this. So, how does it work? These recommendations come from how people actually search – the freshness and popularity of the searched terms.

## Identifying the entities of an eco-system

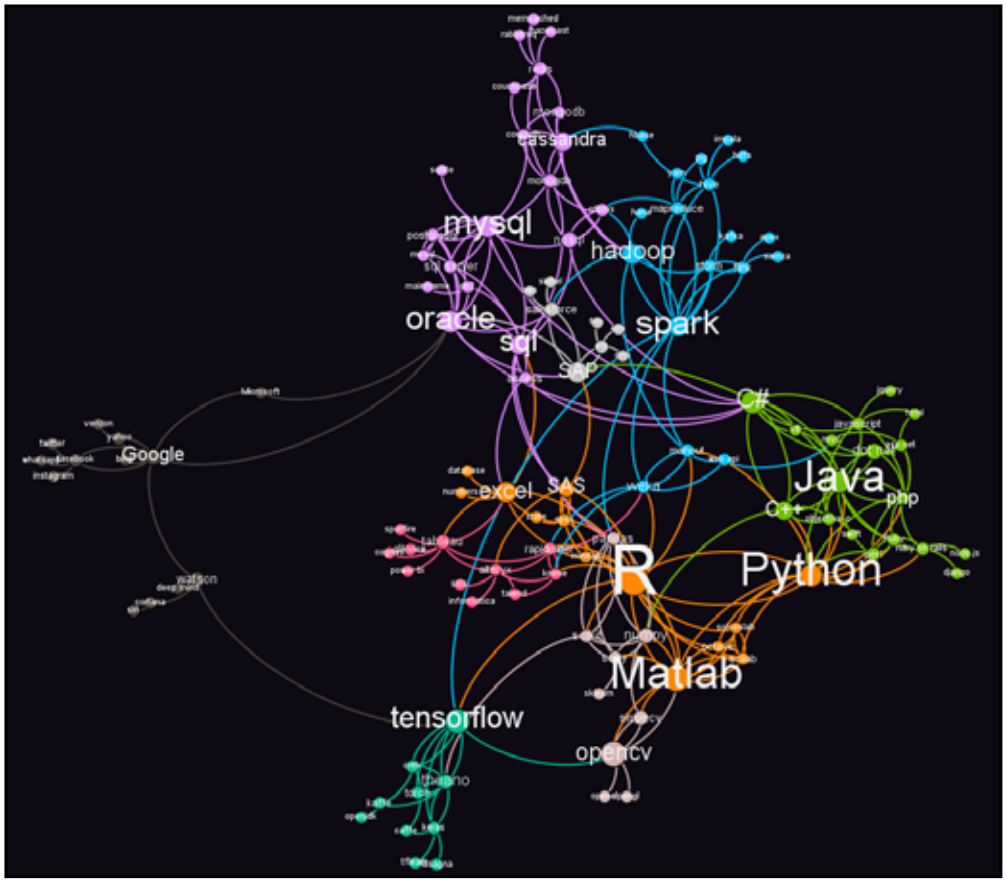



Let us try to understand the concept with the help of an example. When you start typing "R versus.." in the Google search bar, it offers you a list of suggestions – r versus python, r versus sas and so on. All these results are based on real searches done by other people. Popularity is something Google believes in, however there are other factors also such as location, language ,etc., that determine what to show. If we look at our example, it can be seen that "r versus python" is the most frequently searched item, followed by "r versus sas", "r versus r squared". Since, the aim here is to look for products and tools, we can ignore other suggestions and add rest of them to an edge list table to derive a relationship. This table will have one source - R - and multiple targets – python, sas.. This way we can use Google recommendations to identify some of the common substitutes for R, creating an edge.

The next step is to pick each of the target entities identified and derive similar results. For instance, type "python versus.." in the google bar and find out the recommendations. Add the relevant entries in the edge list table. This way we can create our edge list table. The question that arises here is – what to do with this table? What is its significance? Well, this table will help us perform Network Analysis of the eco-system to find out the most important entities.

The first step in network analysis is to identify various components in the graph and visualize how the eco-system looks. This can be done using Force Atlas 2, which is one of the force-directed algorithms available in Gephi. The algorithm stimulates a physical system to spatialize a network in which nodes repulse each other (just like magnets) while edges attract their nodes (like springs). The forces between the elements create a movement that congregates to a balanced state, which helps interpret the data. The layout positions the nodes that are often compared, closer as they have edges connecting them.

If we use Force Atlas 2 for our example, we will see that the tools that are often compared to each other appear to be closer in the network. Hence, R, Python, SAS, Matlab are placed close to each other, while PHP and tensorflow are far apart. Once the nodes are placed in the network, we use “indegree” to size them and ‘Page Rank’ to size the labels. The edge colors indicate the cluster to which a tool is compared.




## Analysis

Based on the network analysis, it can be seen that the eco-system comprises of various sets of communities. To be able to identify these communities, eight modularity classes emerged using the Blondel, et. al., 2008, algorithm, which were marked in different colors for easy identification. These eight modules are –

I. Statistical data analysis tools and languages - R, Python, Matlab, SAS, SPSS, Minitab, Excel, etc.

II. Enterprise tools for business intelligence - Weka, Rapidminer, Alteryx, Tableau, etc.

III. Scripting and Traditional Programming languages – Java, php, C#, C++, etc.

IV. Database 1 – sql, Oracle, mysql, mongodb, Cassandra, etc.

V. Database 2 (Hadoop ecosystem) – Hadoop, Spark, Storm, Mapreduce, Hive, etc.

VI. Deep learning tools – Tensorflow, Theano, Cntk, etc.

VII. Modules, libraries and packages for machine learning – pandas, scipy, opencv, tm, etc.

VIII. Machine learning platforms from organizations – Google, IBM (Watson), Microsoft, Alphabet (Deepmind), etc.

Since, the elements of interest here are Python and R, let us try to analyze them using this structure. Python is more useful for teams that are equipped with **programming and scripting** knowledge, which means people who are familiar with Java, php, C# are more interested in Python. It is also often compared to other **statistical analysis languages**, like Matlab, R, and **machine learning languages** on Hadoop, Storm, etc. Hence, we can deduce that Python is used by programmers who want to apply statistical techniques or analyze data.

Talking about R – it is often compared with **statistical data analysis tools** such as SPSS, SAS, Excel, etc. It is also compared to **business intelligence tools and self-service platforms** like Tableau, Rapindminer etc., which implies it is being used by people from research, academics and business analytics background. Also, the frequent comparisons with Python indicates use of R by **machine learning programming** community. Similarly, those who use database technologies such as Oracle, SQL would be more interested in using **machine learning languages available on platforms of Hadoop**, etc. this bunch of users also seem to be comparing existing tools with enterprise **machine learning toolkits and platforms from organizations** like IBM, Google, etc.

Users who are familiar with Matlab seem to compare it with other scientific and machine learning package set from Python and R. They have a good chance to use R and Python. These users, along with those who use R, are also the ones who compare the tools they use against **deep learning technologies** such as Tensorflow.

## Limitations

Like anything else, even Google Autocomplete suggestions feature comes with its own set of limitations. The results depend on region, location as well as history preferences of the user, which in turn affects the suggestions.

# Author:


## Dr. Archisman Majumdar

Senior Manager, Mphasis NextLabs


Archisman is a senior manager at Mphasis NEXTlabs. At Mphasis, he conceptualizes, develops, and leads multiple products in the analytics R&D space. He has extensive experience in the IT industry at various project management, research, and engineering roles.

He completed his PhD from the Indian Institute of Management Bangalore (IIMB) in the Quantitative Methods and Information Systems area and was a visiting researcher at the IT University of Copenhagen during his PhD. His areas of expertise are in business analytics, social media, product management, and information systems research.

## About Mphasis

Mphasis enables customers to reimagine their digital future by applying a unique formula of integrated cloud and cognitive technology. Mphasis X2C$^{2}$™ formula for success, (shift anything to cloud and power everything with cognitive), drives five dimensions of business value with an integrated consumer-centric Front to Back Digital Transformation, enabling Business Operations and Technology Transformation. Mphasis applies advancements in cognitive and cloud to traditional application and infrastructure services to bring much needed efficiency and cost effectiveness. Mphasis' core reference architectures and tools, combined with domain expertise and hyper specialization are the foundation for building strong relationships with marquee customers.


For more information, contact: marketinginfo@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
88 Wood Street
London EC2V 7RS, UK
Tel.: +44 20 8528 1000

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
DoddanakundhiVillage, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com



VAS 07/11/17 US LETTER BASIL 4536