ISSN:2231-0711

R Progamming in Different Fields

Dr. N. Viswam, M.Sc, Ph.D
HOD & Vice-Principal, Dept. of Statistics
Hindu College, Guntur
Guntur Dist., AP

K. Srinivasa Rao, M. Tech, (Ph.D)
Assoc. Prof.
Bhimavaram Institute of Engg. & Tech.,
Bhimvaram, W.G.Dt., AP

Abstract
R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R programming language is widely used among statisticians and data miners for developing statistical software and data analysis. During the most recent decade, the force originating from both the scholarly community and industry has lifted the R programming language to end up the absolute most significant tool for computational statistics, perception, and data science. For the effective processing and analysis of big data, it allows users to conduct a number of tasks that are essential. R consists of numerous ready-to-use statistical modeling algorithms and machine learning which allow users to create reproducible research and develop data products.

INTRODUCTION:

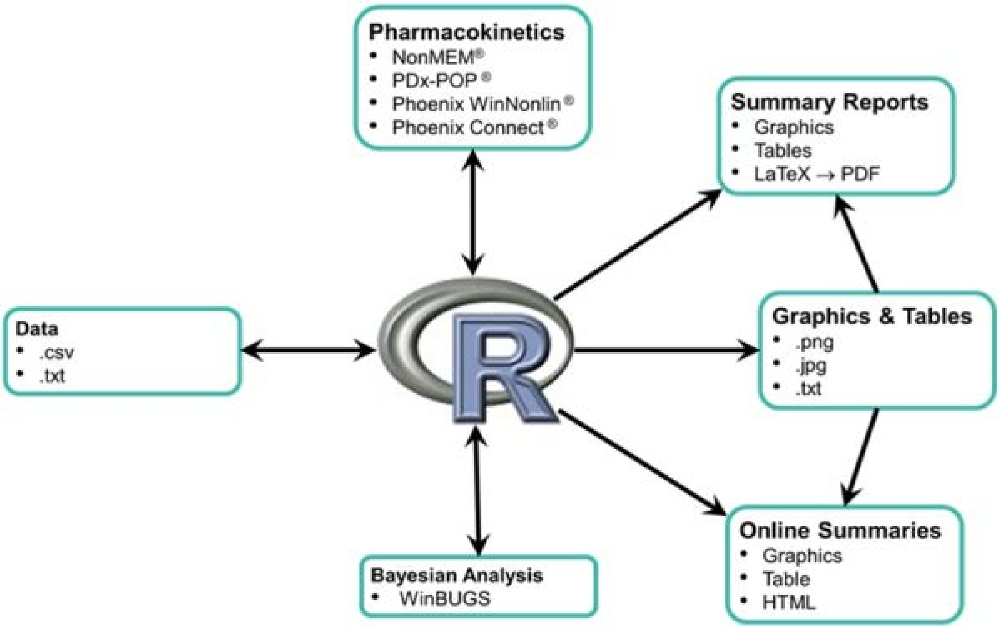
R is a programming language developed by Ross Ihaka and Robert Gentleman in 1993. R possesses an extensive catalog of statistical and graphical methods. It includes machine learning algorithm, linear regression, time series, and statistical inference to name a few. Most of the R libraries are written in R, but for heavy computational task, C, C++ and Fortran codes are preferred. R is not only entrusted by academic, but many large companies also use R programming language, including Uber, Google, Airbnb, Facebook and so on.

Data analysis with R is done in a series of steps; programming, transforming, discovering, modeling and communicate the results

- **Program**: R is a clear and accessible programming tool
- **Transform**: R is made up of a collection of libraries designed specifically for data science
- **Discover**: Investigate the data, refine your hypothesis and analyze them
- **Model**: R provides a wide array of tools to capture the right model for your data
- **Communicate**: Integrate codes, graphs, and outputs to a report with R Markdown or build Shiny apps to share with the world.

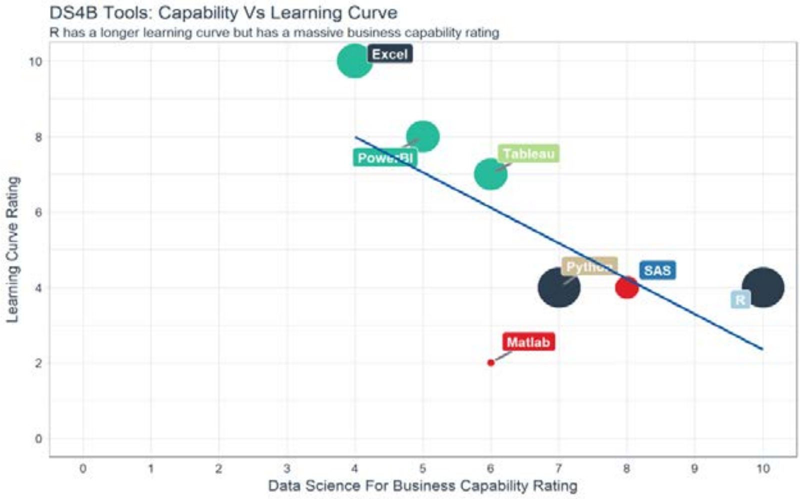
R is an object-oriented language with similarities to C++ and FORTRAN, because of this structure R has the ability to easily manipulate statistics and graphics. R has a set of user developed packages containing groups of functions. There are thousands of well-tested functions available that build on the cutting-edge statistical research; many journals suggest or require available R packages for publication. The packages and the integration with academia means that improvements and experimental procedures are often The statistical programming language R is often underrated within the Pharmaceutical Industry implemented quickly.

The lack of a separate macro language and the code structure when dealing with large datasets and complex data processing can increase the learning curve for R, however R still remains one of the most useable graphics tools. For exploratory graphics there are few software packages that match R with ggplot2 or lattice graphics for the ease of use and customization. As well as this the integration between R and other open source packages such as LaTeX allows for easy and highly customizable reports. These advantages suggest that R should more often be considered when analyzing and summarizing clinical data.



Why use R?

Data science is shaping the way companies run their businesses. Without a doubt, staying away from Artificial Intelligence and Machine will lead the company to fail. The big question is which tool/language should you use? They are plenty of tools available in the market to perform data analysis. Learning a new language requires some time investment. The picture below depicts the learning curve compared to the business capability a language offers. The negative relationship implies that there is no free lunch. If you want to give the best insight from the data, then you need to spend some time learning the appropriate tool, which is R.



On the top left of the graph, you can see Excel and PowerBI. These two tools are simple to learn but don't offer outstanding business capability, especially in term of modeling. In the middle, you can see Python and SAS. SAS is a dedicated tool to run a statistical analysis for business, but it is not free. SAS is a click and run software. Python, however, is a language with a monotonous learning curve. Python is a fantastic tool to deploy Machine Learning and AI but lacks communication features. With an identical learning curve, R is a good trade-off between implementation and data analysis.

When it comes to data visualization (DataViz), you'd probably heard about Tableau. Tableau is, without a doubt, a great tool to discover patterns through graphs and charts. Besides, learning Tableau is not time-consuming. One big problem with data visualization is you might end up never finding a pattern or just create plenty of useless charts. Tableau is a good tool for quick visualization of the data or Business Intelligence. When it comes to statistics and decision-making tool, R is more appropriate.

Stack Overflow is a big community for programming languages. If you have a coding issue or need to understand a model, Stack Overflow is here to help. Over the year, the percentage of question-views has increased sharply for R compared to the other languages. This trend is of course highly correlated with the booming age of data science but, it reflects the demand of R language for data science.

R in Banking

The face of banking is changing and fast. Pick any aspect of banking – risk management, pricing, marketing outreach, customer outreach, product development, cost and revenue allocation – data science is there. The face of the banking industry has been evolving rapidly, particularly post the financial crisis of 2008. Banks were some of the earliest adopters of information technology for processes and security. The growing size of banks and lowered customer loyalty means customers expect quicker and more efficient operational efficiency. Naturally banks are looking at better ways to understand customers and retain them. They are looking at patterns within the data to understand transactional data and engage with customers more meaningfully.

The data available to banks thanks to transactions are numerous. These records were used in risk and fraud management. The advent of data science has led to better management of the information that is flowing in from multiple channels in real-time. Banks are beginning to understand the importance of collating and utilizing their internal data such as debit and credit transactions, purchase history and patterns, mode of communication and brand loyalty. Internet banking data, social media and mobile phone usage for banking has further opened the floodgates of data flowing into the banking system. Mining these multiple sources of data is a challenge. The data comes in various formats and sometimes is fairly unstructured. The data has to be therefore 'cleaned' or data munging takes place. Once the data is in comparable form and cleaned of useless data, analysts use various data analytics software, often deeply customized to arrive at insights. Often third party data will be added to the mix to achieve a holistic picture.

Analytical methods used to derive value from the data are hypothesis testing, crowd sourcing, data fusion and integration, machine learning, natural language processing, signal processing, simulation, time series analysis and visualization. Banks use data science in the areas of customer service, fraud detection, forecasting, understanding consumer sentiment, customer profiling and target marketing, among others.

Banks are using unstructured data from social media to assess how customers view the brand and if they are happy with their brand offerings. Data science helps banks get a full segment-wise view of their customers. Sales persons can tweak their pitch by looking at the customer's Code Halo or 'virtual self' that encapsulates all digital transactions. The average cost of each channel can be more accurately arrived at, as can strategies to migrate customers to cheaper channels. Banks use data science for loan appraisals or lending in any form. It also helps in sending up red flags for specs provided to alert banks to fraud.

Banks are using their data to detect fraud even before it happens. How do they do that? Analytics help alert to any change or 'out of pattern' spending. Data science has helped unearth patterns in false insurance claims. A simple detail like the use of 'ed' over the more active 'ing' was noticed in most false claims. The advantages for banks who are relying on data science and analytics are too many. The benefits for banks by using data science are in sales automation, per customer profitability, usage of customized dashboards, regulatory compliance, budgeting and fraud.

R in Communications, Media and Entertainment

Since consumers expect rich media on-demand in different formats and in a variety of devices, some big data challenges in the communications, media and entertainment industry include:

- Collecting, analyzing, and utilizing consumer insights
- Leveraging mobile and social media content
- Understanding patterns of real-time, media content usage

Applications of big data in the Communications, media and entertainment industry

Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:

- Create content for different target audiences
- Recommend content on demand
- Measure content performance

A case in point is the Wimbledon Championships (YouTube Video) that leverages big data to deliver detailed sentiment analysis on the tennis matches to TV, mobile, and web users in real-time.

Spotify, an on-demand music service, uses Hadoop big data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users.

Amazon Prime, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily utilizes big data.

Big Data Providers in this industry include:Infochimps, Splunk, Pervasive Software, and Visible Measures

R in Healthcare Providers

The healthcare sector has access to huge amounts of data but has been plagued by failures in utilizing the data to curb the cost of rising healthcare and by inefficient systems that stifle faster and better healthcare benefits across the board.

This is mainly due to the fact that electronic data is unavailable, inadequate, or unusable. Additionally, the healthcare databases that hold health-related information have made it difficult to link data that can show patterns useful in the medical field.

Source: Big data in the healthcare sector revolutionizing the Management of Laborious Tasks

Other challenges related to big data include: the exclusion of patients from the decision making process and the use of data from different readily available sensors.

Applications of big data in the healthcare sector

Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. A battery of tests can be efficient but they can also be expensive and usually ineffective.

Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

Obamacare has also utilized big data in a variety of ways.

Big Data Providers in this industry include: Recombinant Data, Humedica, Explorys and Cerner

R in Education

From a technical point of view, a major challenge in the education industry is to incorporate big data from different sources and vendors and to utilize it on platforms that were not designed for the varying data. From a practical point of view, staff and institutions have to learn the new data management and analysis tools.qOn the technical side, there are challenges to integrate data from different sources, on different platforms and from different vendors that were not designed to work with one another. Politically, issues of privacy and personal data protection associated with big data used for educational purposes is a challenge.

Applications of big data in Education

Big data is used quite significantly in higher education. For example, The University of Tasmania. An Australian university with over 26000 students, has deployed a Learning and Management System that tracks among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time. In a different use case of the use of big data in education, it is also used to measure teacher's effectiveness to ensure a good experience for both students and teachers. Teacher's performance can be fine-tuned and measured against student numbers, subject matter, student demographics, student aspirations, behavioral classification and several other variables.

On a governmental level, the Office of Educational Technology in the U. S. Department of Education, is using big data to develop analytics to help course correct students who are going astray while using online big data courses. Click patterns are also being used to detect boredom.

Big Data Providers in this industry include: Knewton and Carnegie Learning and MyFit/ Naviance

R in Manufacturing and Natural Resources

Increasing demand for natural resources including oil, agricultural products, minerals, gas, metals, and so on has led to an increase in the volume, complexity, and velocity of data that is a challenge to handle. Similarly, large volumes of data from the manufacturing industry are untapped. The underutilization of this information prevents improved quality of products, energy efficiency, reliability, and better profit margins.

Applications of big data in manufacturing and natural resources

In the natural resources industry, big data allows for predictive modeling to support decision making that has been utilized to ingest and integrate large amounts of data from geospatial data, graphical data, text and temporal data. Areas of interest where this has been used include; seismic interpretation and reservoir characterization. Big data has also been used in solving today's manufacturing challenges and to gain competitive advantage among other benefits. In the graphic below, a study by Deloitte shows the use of supply chain capabilities from big data currently in use and their expected use in the future.

Source: Supply Chain Talent of the Future Findings from the third annual supply chain survey. Deloitte. 2015.
Big Data Providers in this industry include: CSC, Aspen Technology, Invensys and Pentaho

R in Government

In governments the biggest challenges are the integration and interoperability of big data across different government departments and affiliated organizations.

Applications of big data in Government

In public services, big data has a very wide range of applications including: energy exploration, financial market analysis, fraud detection, health related research and environmental protection.

Some more specific examples are as follows:

Big data is being used in the analysis of large amounts of social disability claims, made to the Social Security Administration (SSA), that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims. The Food and Drug Administration (FDA) is using big data to detect and study patterns of food-related illnesses and diseases. This allows for faster response which has led to faster treatment and less death.

The Department of Homeland Security uses big data for several different use cases. Big data is analyzed from different government agencies and is used to protect the country.

Big Data Providers in this industry include: Digital Reasoning, Socrata and HP

R in Insurance

Lack of personalized services, lack of personalized pricing and the lack of targeted services to new segments and to specific market segments are some of the main challenges. In a survey conducted by Marketforce challenges identified by professionals in the insurance industry include underutilization of data gathered by loss adjusters and a hunger for better insight.

Applications of big data in the insurance industry

Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices and CCTV footage. The big data also allows for better customer retention from insurance companies. When it comes to claims management, predictive analytics from big data has been used to offer faster service since massive amounts of data can be analyzed especially in the underwriting stage. Fraud detection has also been enhanced. Through massive data from digital channels and social media, real-time monitoring of claims throughout the claims cycle has been used to provide insights.

Big Data Providers in this industry include: Sprint, Qualcomm, Octo Telematics, The Climate Corp.

R in Retail and Whole sale trade

From traditional brick and mortar retailers and wholesalers to current day e-commerce traders, the industry has gathered a lot of data over time. This data, derived from customer loyalty cards, POS scanners, RFID etc. is not being used enough to improve customer experiences on the whole. Any changes and improvements made have been quite slow.

Applications of big data in the Retail and Wholesale industry

Big data from customer loyalty data, POS, store inventory, local demographics data continues to be gathered by retail and wholesale stores. In New York's Big Show retail trade conference in 2014, companies like Microsoft, Cisco and IBM pitched the need for the retail industry to utilize big data for analytics and for other uses including:

- Optimized staffing through data from shopping patterns, local events, and so on
- Reduced fraud
- Timely analysis of inventory

Social media use also has a lot of potential use and continues to be slowly but surely adopted especially by brick and mortar stores. Social media is used for customer prospecting, customer retention, promotion of products, and more.

Big Data Providers in this industry include: First Retail, First Insight, Fujitsu, Infor, Epicor and Vistex

R in Transportation

In recent times, huge amounts of data from location-based social networks and high speed data from telecoms have affected travel behavior. Regrettably, research to understand travel behavior has not progressed as quickly. In most places, transport demand models are still based on poorly understood new social media structures.

Applications of big data in the transportation industry

Some applications of big data by governments, private organizations and individuals include:

- Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)
- Private sector use of big data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)
- Individual use of big data includes: route planning to save on fuel and time, for travel arrangements in tourism etc.

Source: Using big data in the transport sector
Big Data Providers in this industry include: Qualcomm and Manhattan Associates

R in Energy and Utilities

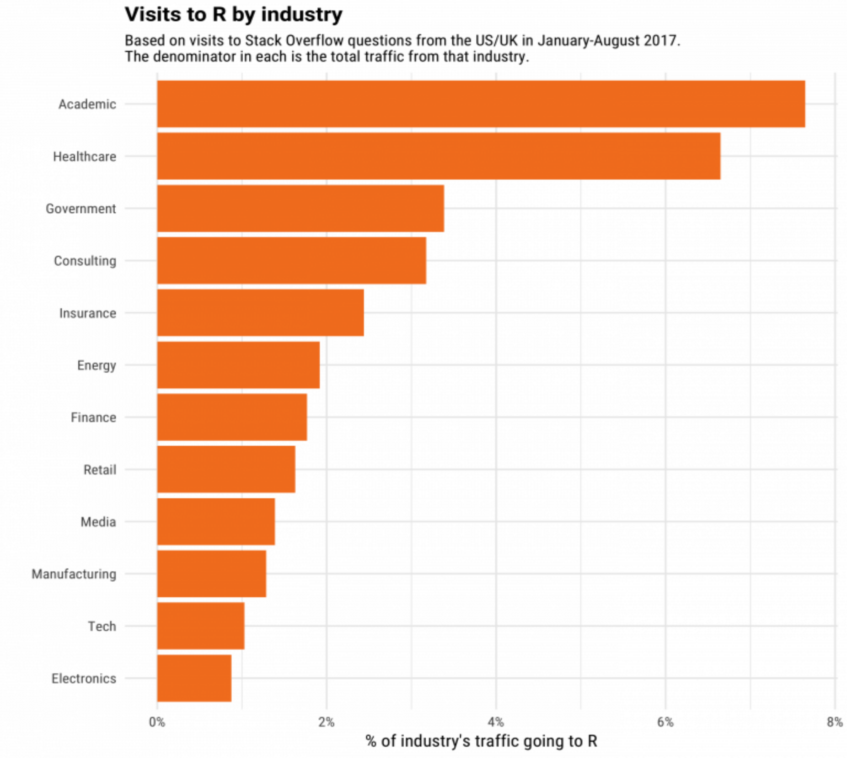
The image below shows some of the main challenges in the energy and utilities industry.

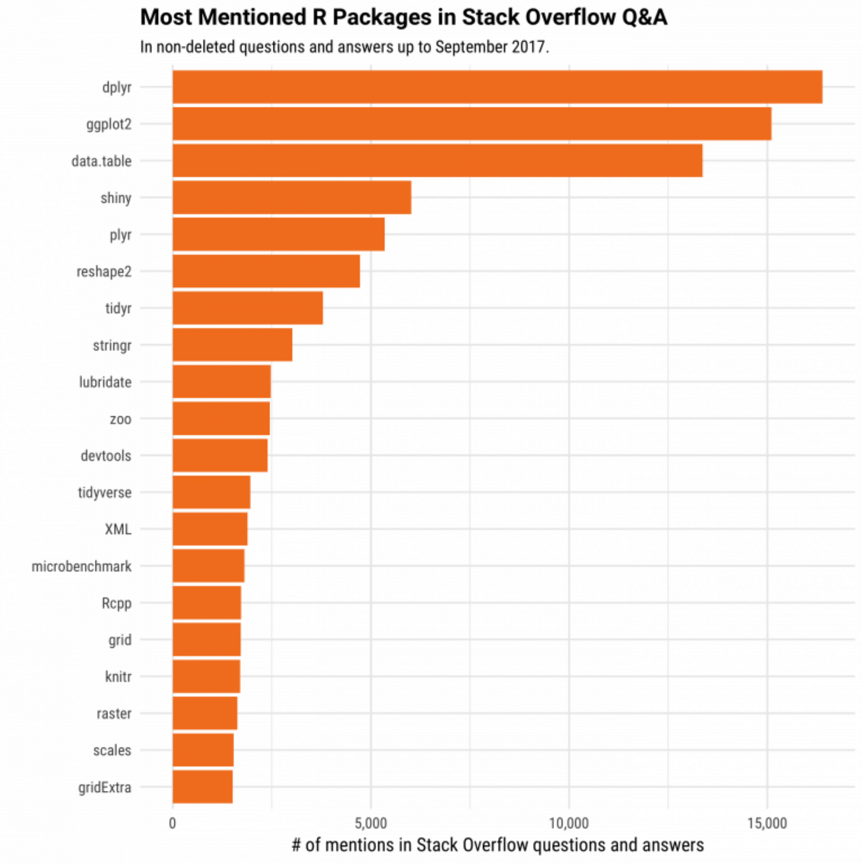
Applications of big data in the energy and utilities industry

Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use.

In utility companies the use of big data also allows for better asset and workforce management which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

Big Data Providers in this industry include: Alstom Siemens ABB and Cloudera





R in Airbnb

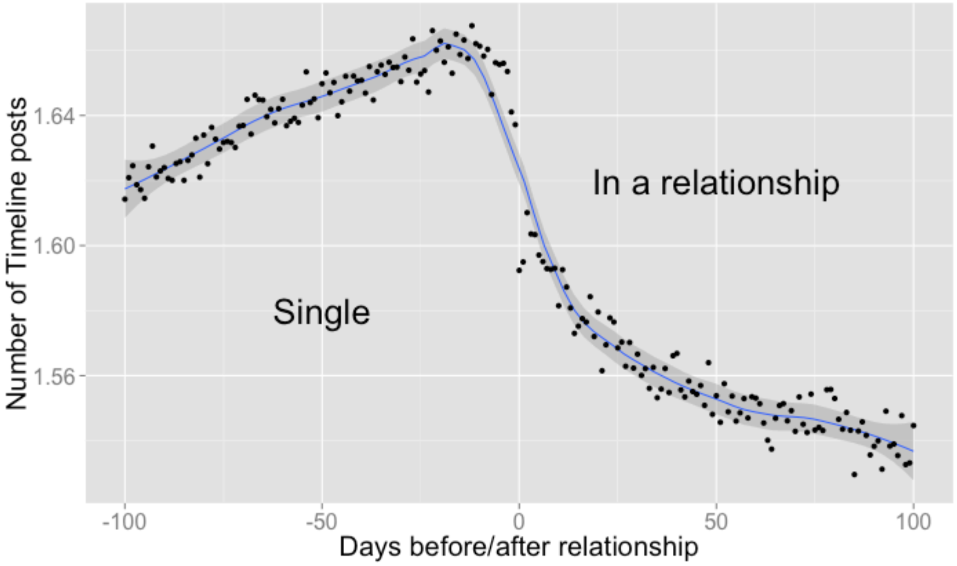
Airbnb uses R to drive a number of company initiatives, using an internal R package called Rbnb. 80% of Airbnb data scientists use R, and 64% of them use it as their primary analysis tool. Analysts use R to predict re-booking rates using past guest ratings and to automate guest/host matching. The tool is also used for internal reporting and data visualizations.

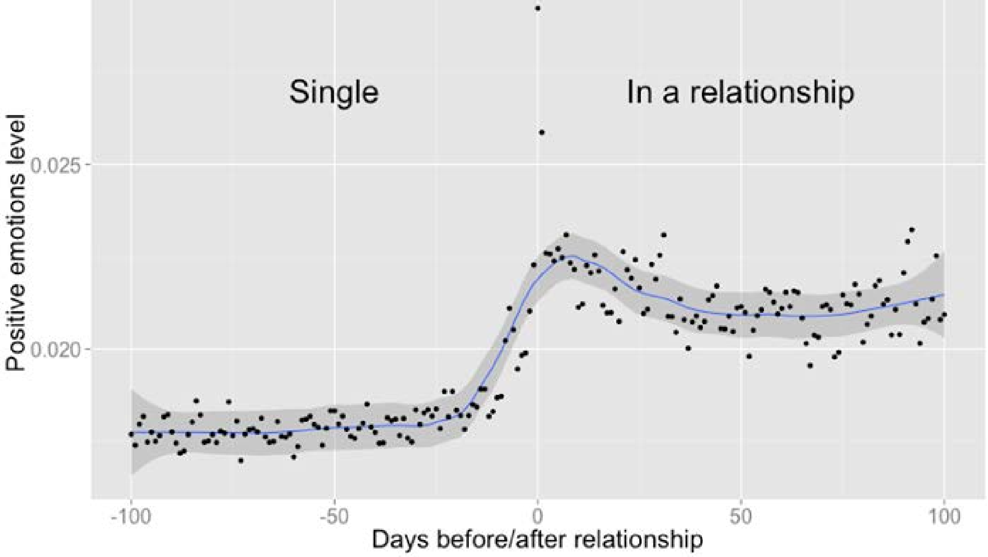
R in Facebook

Facebook used R to perform sizable behavior analysis based on status and profile picture updates. Carlos Diuk, a Facebook data scientist, created an analysis on the formation of love, based on Facebook relationship status updates. What they found is that as a relationship is formed, the number of timeline posts go down. However, the amount of positive emotions in posts rise.

R in IBM

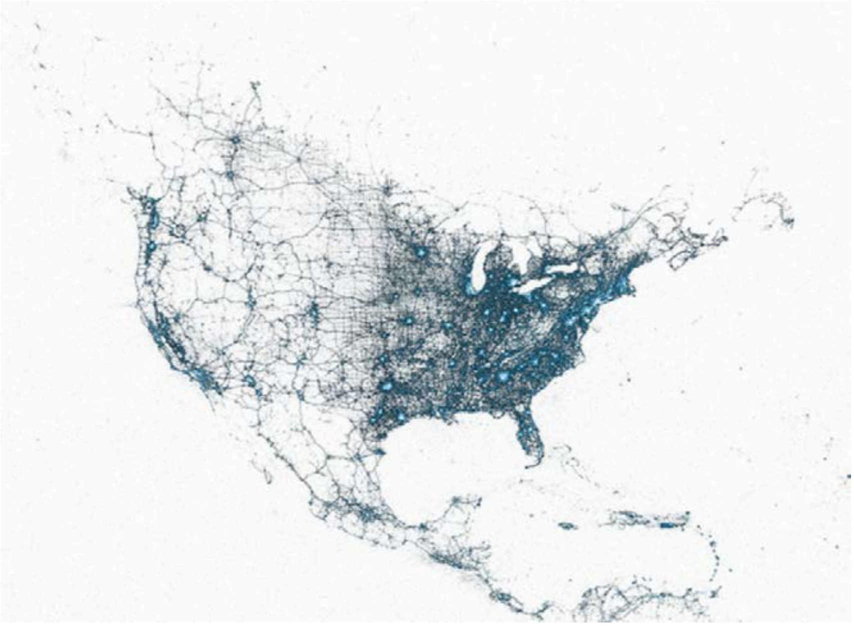
IBM Invests in R Programming Language for Data Science; joins R Consortium. The company has relied on R, among other data languages, to help create innovative solutions like IBM Watson, an open, cognitive computing technology platform that represents a new era in computing where systems understand the world like humans do: through senses, learning, and experience. As a member of the R Consortium, IBM will collaborate with the R user community and support the project’s mission to identify, develop and implement infrastructure projects that drive standards and best practices for R code.





R in Twitter

Using R, Twitter has created some pretty impressive projects. Below is geo-tagged data that represent every tweet in the United Stated since 2009. Twitter also has created open-source packages for anomaly and breakout detection, which has helped improve their customer experience.



R in Microsoft

For those Xbox gamers out there, you can thank Microsoft for choosing R for visualization in their matchmaking system. Matchmaking is how Xbox pairs gamers up with someone of their equal skill level, because no one with an advanced skillset wants me on their team! Analysis of both the gaming community and the games themselves (for example, where players are getting stuck) is achieved using the R language and statistical modeling.

R in John Deere

Farming equipment manufacturer, John Deere, saw huge company savings when they dropped their “flop forecasting software” and adopted R. John Deere uses R for short and long-term forecasting, forecasting crop yields, data coordination, and optimizing the build order on the production line.

R in Google

Search giant, Google, uses R in many ways. One specific way Google uses R is to determine the effectiveness of display ads. For example, a brand can set an ad that would show on a site such Animal Planet or HGTV. Just because a visitor doesn’t click on the ad does not mean they did not convert. Google uses R to gather and visualize search behaviors of those who saw the ad and didn’t visit the site, versus viewers who saw the ad and visited the advertiser’s site without clicking the ad itself. This ensures advertisers are getting the most for their dollar.

R in Trulia and Zillow

If you’ve been on the hunt for a house or even a rental, you probably have turned to Trulia or Zillow. Both companies use R in a way that has a huge impact on your purchasing decision. Zillow’s “Zestimate” gives users an estimate of a property’s value. The “Zestimate” takes data and runs it through proprietary software powered by R to give the user the estimate they see on the screen. In fact, Zillow is currently offering a $1.2M prize in a Kaggle competition for whoever can develop the most accurate home pricing algorithm.

Have you ever found that house or apartment that looked amazing, then you scrolled down to the bottom and recoiled in horror over the amounts of stabbing and domestic disputes are in the crime reports? That’s Trulia using R and statistical modeling to scare you into looking somewhere else.

CONCLUSION:

In a nutshell, R is a great tool to explore and investigate the data. Elaborate analysis like clustering, correlation, and data reduction are done with R. This is the most crucial part, without a good feature engineering and model, the deployment of the machine learning will not give meaningful results.

REFERENCES

1. Peter Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". *Proceedings of the Association for Computational Linguistics*. pp. 417–424. arXiv:cs.LG/0212032.
2. Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002)."Thumbs up? Sentiment Classification using Machine Learning Techniques". *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79–86.
3. Chang J (2010). lda: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.2.3, URL http://CRAN.R-project.org/package=lda.
4. Daum´e III H (2008). HBC: Hierarchical Bayes Compiler. Pre-release version 0.7, URL http://www.cs.utah.edu/~hal/HBC/.
5. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990). "Indexing by Latent Semantic Analysis." Journal of the American Society for Information Science, 41(6), 391–407.
6. Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data Via the EM-Algorithm." Journal of the Royal Statistical Society B, 39, 1–38. Feinerer I (2011). tm: Text Mining Package. R package version 0.5-5., URL http://CRAN. R-project.org/package=tm.
7. Feinerer I, Hornik K, Meyer D (2008). "Text Mining Infrastructure in R." Journal of Statistical Software, 25(5), 1–54. URL http://www.jstatsoft.org/v25/i05/.
8. Griffiths TL, Steyvers M (2004). "Finding Scientific Topics." Proceedings of the National Academy of Sciences of the United States of America, 101, 5228–5235.
9. Tutorial on Discovering Multiple Clustering Solutions http://dme.rwth-aachen.de/en/DMCS
10. Time-Critical Decision Making for Business Administration http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm.
11. A paper on Open-Source Tools for Data Mining, published in 2008 http://eprints.fri.uni-lj.si/893/1/2008-OpenSourceDataMining.pdf
12. An overview of data mining tools http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf
13. Textbook on Introduction to social network method http://www.faculty.ucr.edu/~hanneman/nettext/
14. Information Di_usion In Social Networks: Observing and Inuencing Societal Interests, a tutorial at VLDB'11 http://www.cs.ucsb.edu/~cbudak/vldb_tutorial.pdf
15. Tools for large graph mining: structure and di_usion, a tutorial at WWW2008 http://cs.stanford.edu/people/jure/talks/www08tutorial/
16. Graph Mining: Laws, Generators and Tools http://www.stanford.edu/group/mmds/slides2008/faloutsos.pdf

View Publication

http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-3.html
http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data
http://www.datasciencecentral.com/profiles/blogs/the-free-big-data-sources-everyone-should-know
http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609?image_number=1
http://en.wikipedia.org/wiki/Cluster_analysis
http://en.wikipedia.org/wiki/Tag_cloud
http://thinktostart.com/cluster-twitter-data-with-r-and-k-means/
http://www.rdatamining.com/