2015 Data Science Survey

KARL REXER, PhD

PAUL GEARAN

HEATHER ALLEN, PhD

www.RexerAnalytics.com

Survey Testimonials

The Data Miner Survey is one of the few unbiased resources that provides a comprehensive overview of the data science community. It is a key source of information for analytic professionals regarding vendors, tools, and state-of-the-art algorithms. As a frequently cited publication, it covers essential topics, new trends like "big data", as a well as general market surveillance.
— **Michael Zeller**, PhD; CEO, Zementis

Rexer Analytics has been instrumental in helping advance the field of data mining through applied research, software evaluation, testing, professional conference support and consulting. Their research evaluating the trends and preferences among the data mining community is a great resource for many.
— **Wayne Thompson**, PhD; Chief Data Scientist, SAS

I consider Rexer Analytics' surveys one of the best independent analyses of Data Mining. I stress the word 'independent' as that is the most useful. Karl and his team have been active in this market for many years and bring considerable experience to the topic.
— **John MacGregor**, VP, Predictive Analysis, Products & Innovation, SAP

Rexer Analytics' series of Data Science Surveys is a foundational contribution to this industry's community.
— **Eric Siegel**, PhD; Author of Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die; Founder, Predictive Analytics World; Executive Editor, Predictive Analytics Times

As the longest-running survey of data miners in the industry, the Rexer Analytics survey provides valuable insights and trends into the tools, methods and applications of advanced analytical techniques today.
— **David M. Smith,** R Community Lead, Revolution Analytics (a Microsoft company)

Rexer Analytics' Data Mining Survey is a comprehensive and accurate assessment of the industry's attitudes, performance, and trends. There's no better place to get a firm grip on the direction of this rapidly growing field.
— **Eric A. King**, President, The Modeling Agency, LLC

The Rexer Analytics surveys are extremely useful. They provide industry benchmarks, keep us abreast of new trends, and surprise us with new insights.
— **Julia Minkowski**; Risk Analytics Manager, Fiserv; Co-founder, "Russian Speaking Women in Tech"

As professor in a rapidly changing field, it is of paramount importance for me to keep up with trends in analytics software, techniques and applications. The bi-annual Rexer Analytics Surveys have been a great source of information on global trends in analytics, and have allowed me to teach Babson students skills, tools and techniques that can position them better in the marketplace.
— **Dessislava A. Pachamanova**, PhD; Professor of Analytics and Computational Finance, Babson College; Co-designer of the Babson undergraduate and MBA concentrations in Business Analytics

The Rexer Analytics Data Miner Survey is the best survey of the current state and direction of the data mining and predictive analytics industry. I recommend it to my workshop and course attendees regularly as a way to understand the important trends in software, algorithms, job titles, and vertical markets, as well as issues impacting the analytics industry and which buzz words are gaining traction.
— **Dean Abbott**; Co-Founder & Chief Data Scientist, SmarterHQ; Founder & President, Abbott Analytics

Introduction

The field of data science has evolved in the past decade at a breathtaking pace. The potential applications and sources for rich datasets have expanded as has the attention the field of data science has received in the popular press. In 2015, Rexer Analytics fielded our 7th Data Science Survey, and found that while much has changed in our field over the years, these changes overlay a stable foundation.

Contents

REXER ANALYTICS

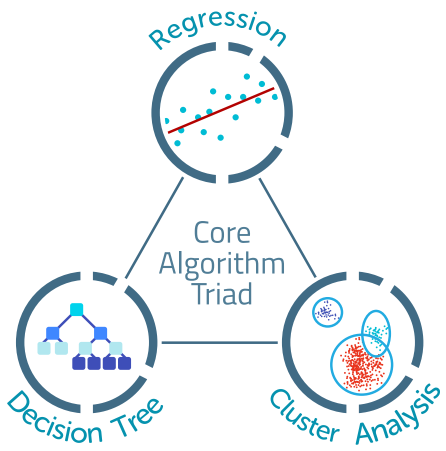
www.RexerAnalytics.com

+1 617-233-8185

DataScienceSurvey@RexerAnalytics.com

Key Findings

Regression, decision trees, and cluster analysis remain the most commonly used algorithms in the field. This has been consistent over the years.



The popularity of R continues to rise.

In 2007, only 23% of respondents reported using R, while this year 76% of respondents did so. Similarly, in 2008 only 5% of respondents indicated that R was their primary tool for data mining and/or statistical analysis. Today more than a third of respondents (36%) identify R as their primary tool — R is selected more than any other tool.

Overall, job satisfaction levels of data scientists are high, but have declined slightly since the 2013 survey. Those with the highest levels of satisfaction have been in the field 10 years or longer.

Deployment continues to be a challenge for organizations.

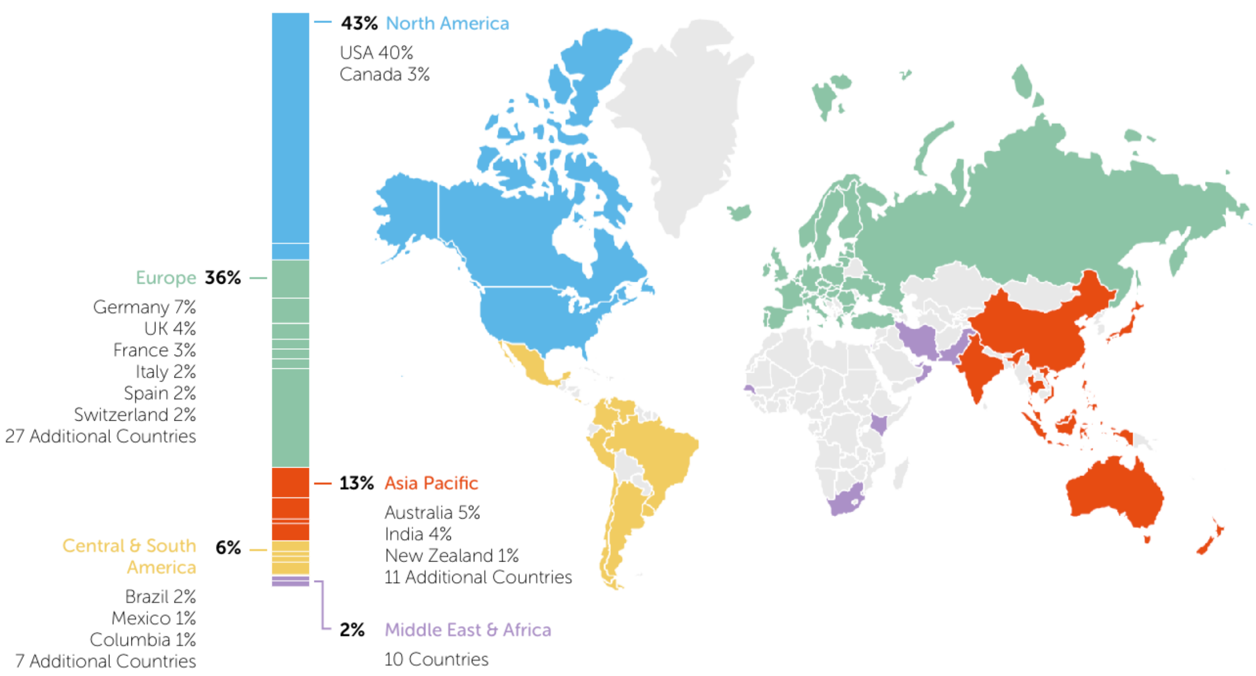
Less than two thirds of respondents indicate that models are deployed "most of the time" or "always." These numbers have not improved over time. Buy-in from customers, managers, and other members of the organization is the largest barrier to deployment, with real-time scoring and other technology issues also playing a significant role.

The term "Data Scientist" has surged in popularity, with over 30% of us describing ourselves as data scientists now, compared to only 17% in 2013

Methodology

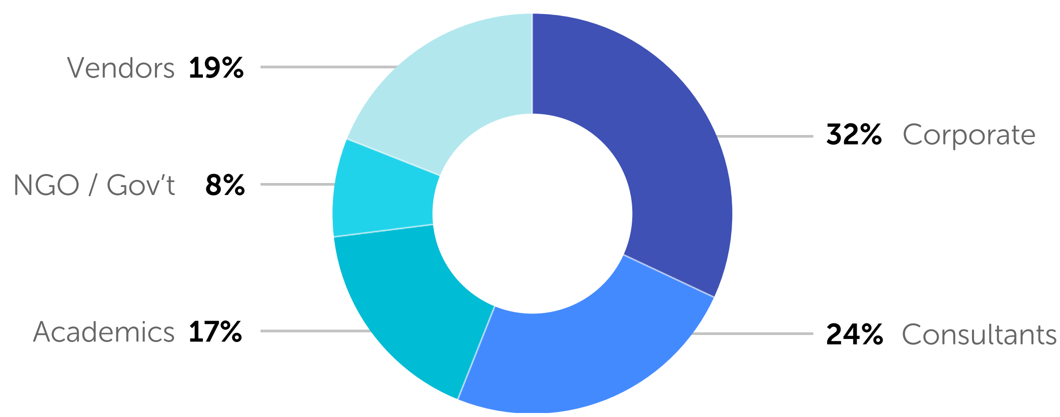
Rexer Analytics' 2015 Data Science Survey was conducted during the first half of 2015. This is Rexer Analytics' 7th survey since 2007 in this series (previously known as the Data Miner Survey). The present survey contained 59 questions that were emailed to over 10,000 data science professionals. The survey was also promoted by several data science newsgroups, tool vendors, and bloggers. There were 1,220 respondents from 72 different countries.

GEOGRAPHIC REPRESENTATION OF SURVEY RESPONDENTS



Respondents came from all sectors of the data science field, including government, corporate employees, consultants, academics, and vendors. Data from vendors are excluded from all analyses of software tools and algorithms. A wide range of age, experience, and work environments are represented.

DATA SCIENTIST JOB SECTORS



The 2015 survey encompassed a variety of questions relevant to data scientists. Respondents were asked about data size, structure, and storage; algorithm and text mining usage; tool preferences, satisfaction, and usage; work environment, challenges and opportunities; and a variety of other issues.

What We Do

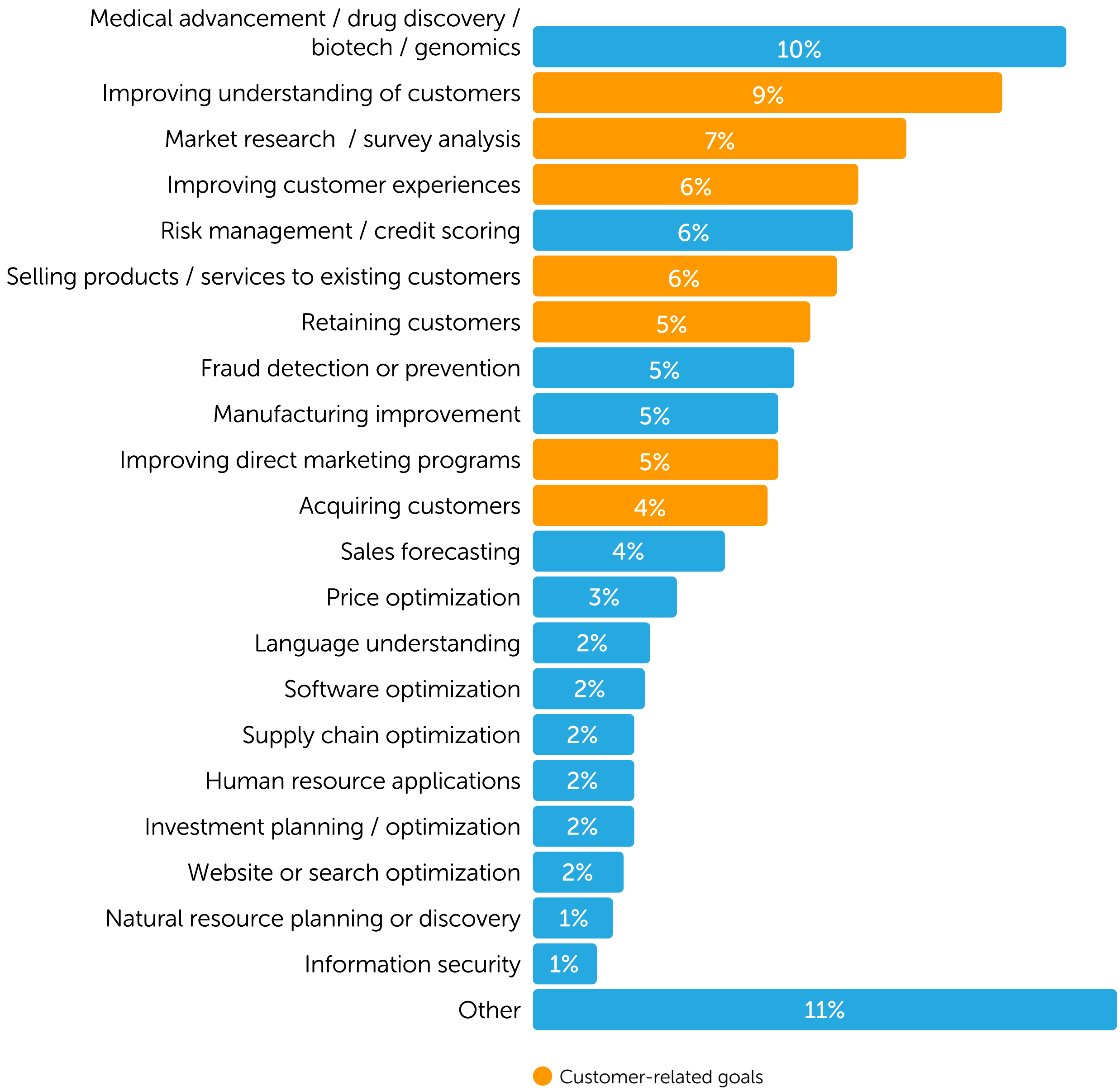
ANALYTIC GOALS (PROPORTION OF PEOPLE)

	2011	2013	2015
Improving understanding of customers	33%	45%	**46%**
Retaining customers	30%	36%	**37%**
Improving customer experiences	22%	36%	**36%**
Selling products / services to existing customers	23%	33%	**35%**
Market research / survey analysis	29%	36%	**34%**
Acquiring customers	23%	32%	**32%**
Improving direct marketing programs	22%	27%	**30%**
Sales forecasting	19%	27%	**27%**
Fraud detection or prevention	21%	23%	**26%**
Risk management / credit scoring	22%	26%	**25%**
Price optimization	14%	22%	**23%**
Manufacturing improvement	10%	15%	**17%**
Medical advancement / drug discovery / biotech / genomics	12%	17%	**17%**
Supply chain optimization	7%	11%	**15%**
Investment planning / optimization	11%	13%	**14%**
Software optimization	7%	9%	**11%**
Website or search optimization	8%	12%	**10%**
Human resource applications	4%	8%	**9%**
Collections	6%	7%	**8%**
Language understanding	4%	7%	**8%**
Criminal or terrorist detection	4%	4%	**7%**
Information security	4%	5%	**6%**
Natural resource planning or discovery	3%	5%	**5%**
Fundraising	3%	3%	**3%**
Reducing email spam	2%	2%	**2%**

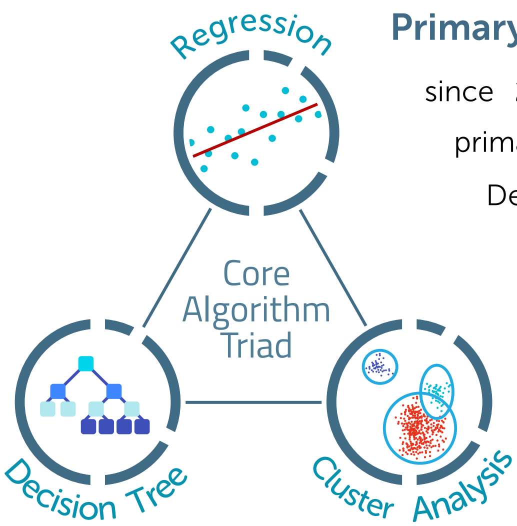
After increases between 2011 and 2013 for several customer-focused analytic goals, there was stability between 2013 and 2015 in the proportion of analytic professionals pursuing the various goals. As in previous years, most people report working towards multiple goals.

The 2015 survey also asked people how they allocate their time across these goals. A collection of customer-related goals occupy 42% of analytic professionals' time. Interestingly, people working toward medical goals were less likely to report splitting their time across diverse goals. This resulted in medical goals rising to the top as the single goal where collectively analytic professionals report spending the most time (10% of overall time).

ANALYTIC GOALS (PROPORTION OF TIME)



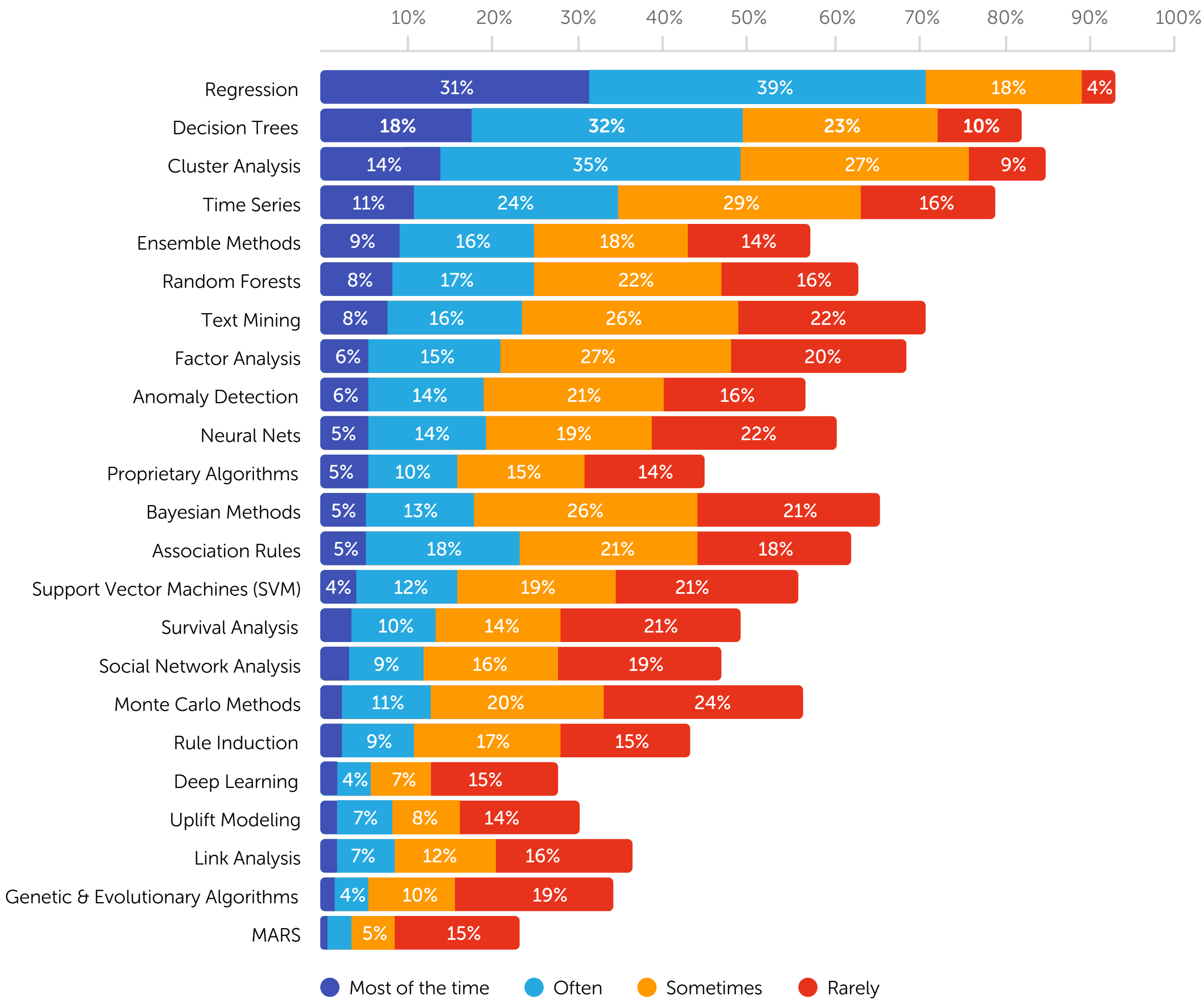
How We Do It



Primary Algorithms We Use. Rexer Analytics' surveys since 2007 have consistently shown that data scientists primarily work with the algorithm triad of Regression, Decision Trees, and Cluster Analysis. Over half of respondents in each year we have conducted our survey indicate that they have used each of these methods in the course of the prior year. Furthermore, in both 2013 and 2015 more than half of respondents reported that they use each of these algorithms either "Often" or "Most of the Time."

Additional Algorithms we use. There are a wide variety of algorithms that data scientists use in addition to these three, from neural nets to time series to rule induction and anomaly detection. While Ran Bi posited in 2014 that Deep Learning (a version of machine learning related to neural networks) may soon make all other learning algorithms obsolete,[1] only 7% of our respondents used these techniques regularly, and 62% report never using Deep Learning. While this is the first year Deep Learning has been included in the survey, future years will illustrate the trends in this algorithm's use. The Rexer Analytics surveys provide a consistent methodology for measuring changes in the use of algorithms. In recent years we have seen increases in the use of Ensemble methods and Uplift Modeling.

2015 ALGORITHM USAGE



The Rise of R

The statistical programming language R was originally developed in the early 1990's by Ross Ihaka and Robert Gentleman at the University of Auckland as a test of how a statistical environment might be built. It was initially based on the computer language "S" which was created by John Chambers in 1976. R quickly outgrew its origins and became a collaborative effort of researchers working via the internet.[2] Early in its history, R development was handed over to the R Development Core Team and became a GNU project, cementing it as free software revised through the process of mass collaboration.

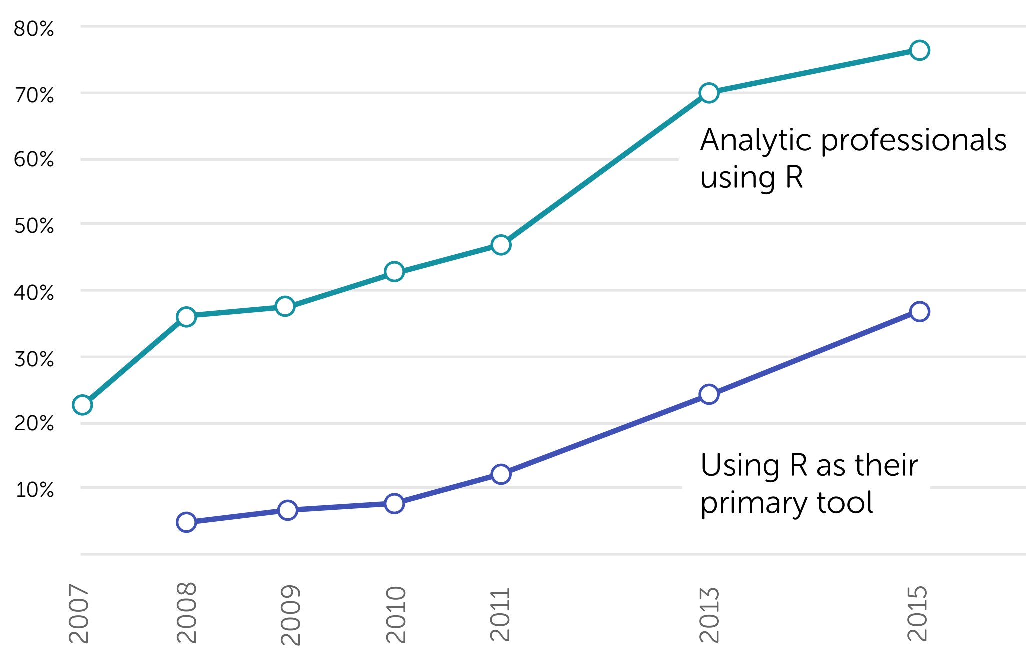
R offers a full array of software capabilities for data preparation, analysis, and graphical display. According to the R project: "R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity."[3]

The popularity of R has risen steadily since its creation. In 2015, as part of its 16th annual Data Mining software poll, KDnuggets asked "What analytics/data mining software you used in the past 12 months for a real project (not just evaluation)." In this year's poll, R was the top-ranked data mining solution, selected by 47% of KDnuggets poll respondents (compared to 39% in 2014).[4] Robert Muenchen uses a variety of methods to explore the popularity of data science software packages at www.r4Stats.com. He reports that job advertisements requiring the use of R have been on a steady rise, eclipsing the demand for SPSS Statistics users and rapidly approaching the demand for those employing SAS.[5] He also reports that in the frequency of use in scholarly publications, R recently passed SAS in popularity (but is still far behind SPSS Statistics). And he notes that R is overwhelmingly the most-discussed package in online discussion forums.

76% of analytic professionals report using R

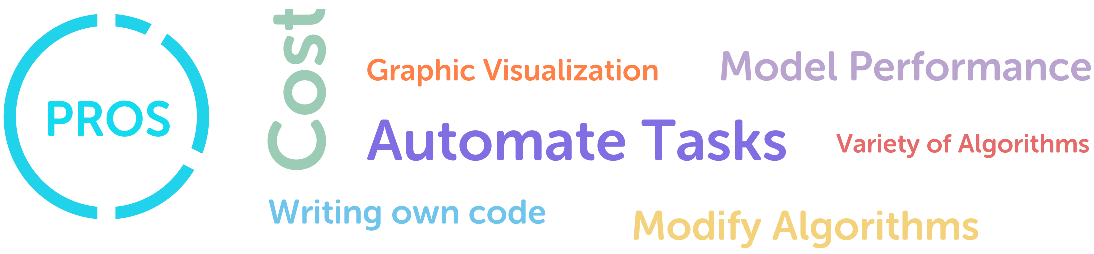
Rexer Analytics has seen a steady rise in the proportion of people who report using R, as well as in the proportion of people who select R as their primary tool. In 2007, only 23% of respondents reported using R, while this year 76% of respondents did so. Similarly, in 2008 (the first year in which we asked the question) only 5% of respondents indicated that R was their primary analytic tool. Today more than a third of respondents (36%) identify R as their primary tool. Since 2010, R has been the most used data mining tool in our survey.

RISE OF R USAGE



In our 2013 Survey[6] we compared those who identified R as their primary tool vs. other data scientists on the importance placed on a variety of factors in selecting software. We determined that while overall, people consider quality and accuracy of model performance, dependability of software, and data manipulation capabilities the most important factors when choosing a data mining tool, those using R as their primary tool identify the ability to write one's own code as their most important priority. Additionally, the quality of user interface was rated as significantly less important by primary R users than by other data scientists. Interestingly, there was no difference in the stated importance of cost between those using R as their primary package and others.

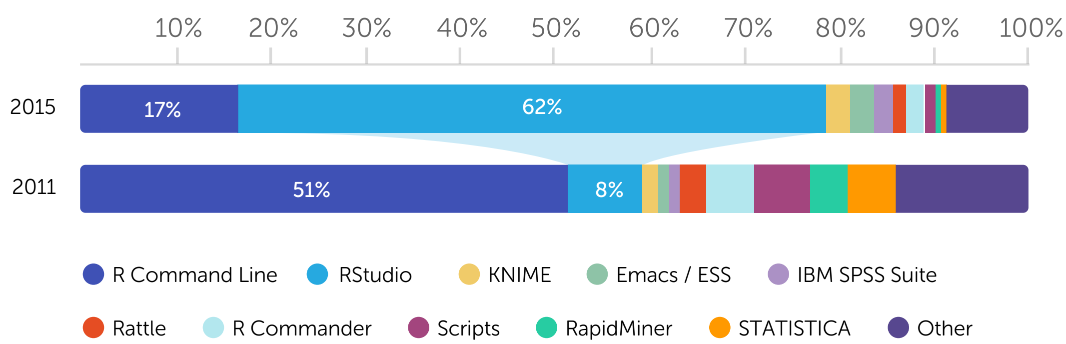
PROS AND CONS OF R USAGE



This year we also looked at the satisfaction of R users compared to others on these same factors. Not surprisingly, R users are significantly more likely to be "very satisfied" with the cost of their (free) software (94% vs. 35%). They are also more likely to be satisfied with the variety of available algorithms, the ability to modify algorithms, the ability to write one's own code, model performance, the ability to automate repetitive tasks, and graphical visualization. Meanwhile, they are less satisfied than others with their software's ease of use and speed. Overall, R Primary users are more likely to be "extremely satisfied" with their primary tool than are those primarily using other tools (57% vs. 40%).

Data scientists interface with R in a variety of ways. However, in recent years RStudio has become the dominant R interface. In the 2015 survey 62% of R users report using RStudio, up from just 8% in 2011. In 2015 only 17% of R users report using the R command line interface, down from 51% in 2011. Small groups of R users report using a variety of other interfaces, such as RKWard, Revolution R GUI, and Rattle. Others report using R from within other tools such as IBM SPSS, STATISTICA and KNIME.

RSTUDIO HAS BECOME THE DOMINANT R INTERFACE



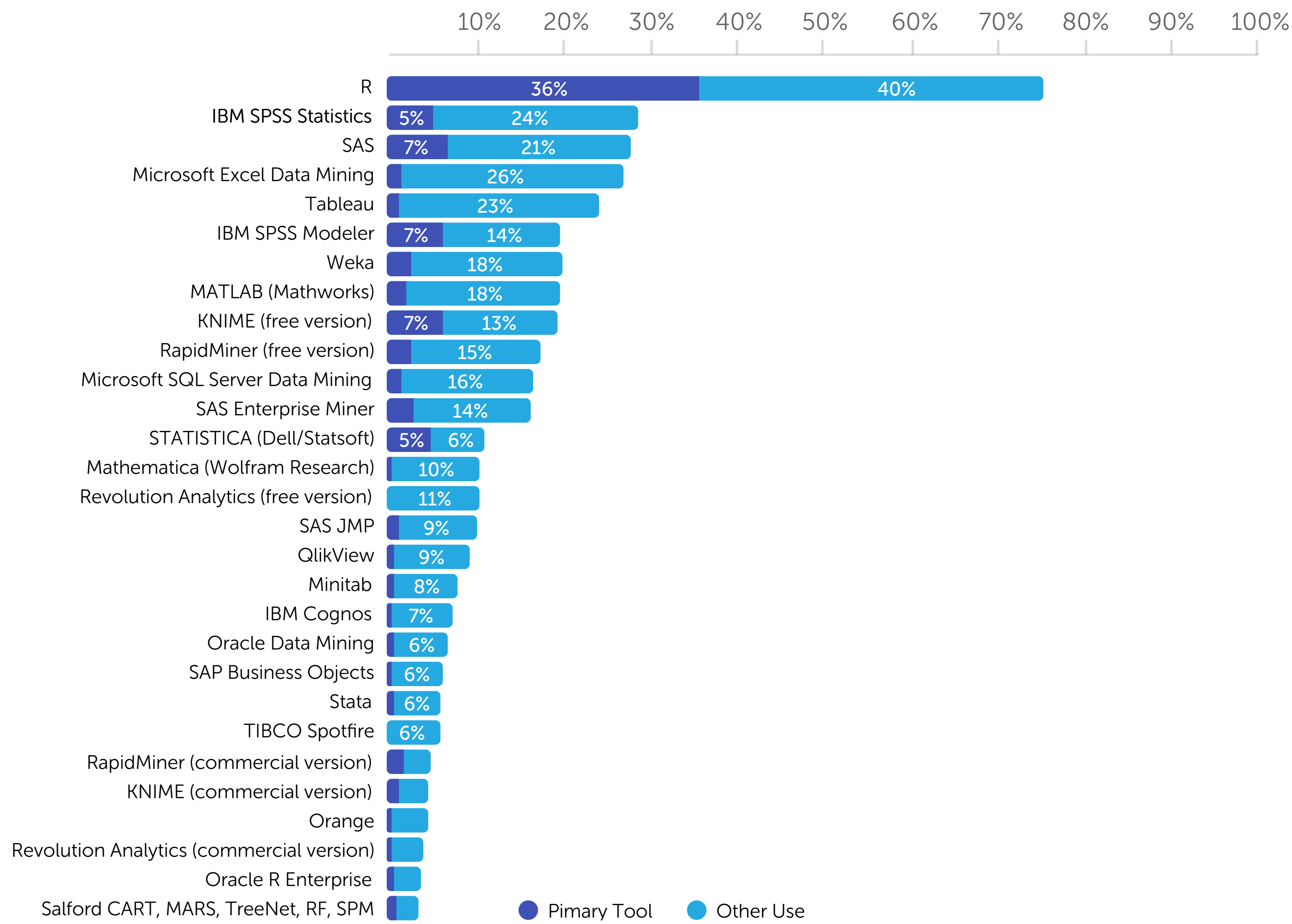
R users come from all settings. While historically those reporting R as their primary package were less likely to work in corporate settings,[6] today this is no longer the case. R users can be found in a variety of work settings and serve a wide variety of industries.

Primary R users are a young and growing population. They are more likely to be under 30 and less likely to be over 50. They are more likely to have worked in predictive analytics for five years or less. They are also more likely to work for organizations that have only recently adopted analytics (used analytics for five years or less). Not surprisingly, given their youth, they are more likely to label themselves using the current term "data scientist" rather than the more traditional "data analyst" or other titles.

Other Tools We Use

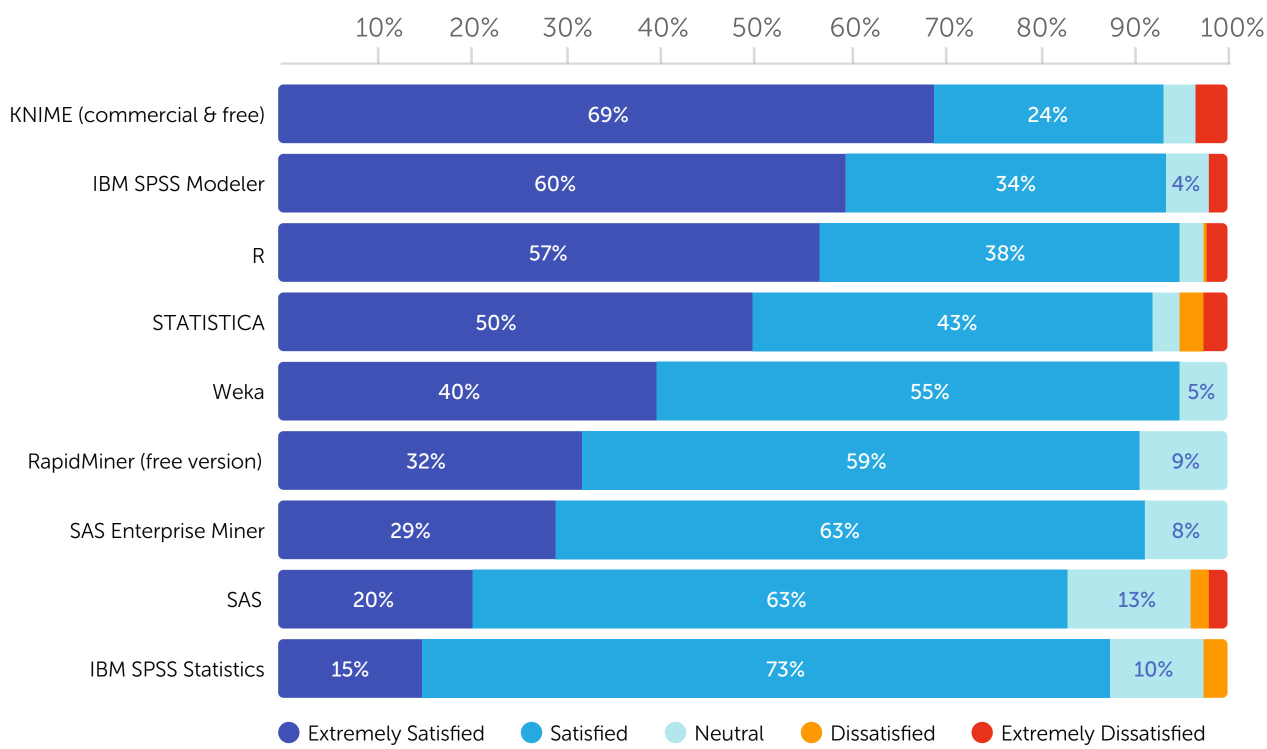
While R has established a dominant position in the world of analytic software, 64% of data scientists select other tools as their primary analytic software, and the average data scientist reports using five tools. Many other platforms continue to show widespread use, including IBM SPSS Statistics, SAS, Microsoft Excel Data Mining, Tableau, and IBM SPSS Modeler. The tool use figure below shows the breadth of analytic software used. R is not the only free tool data scientists report using. Weka is used by 20% of data scientists, and several companies offer free versions of their commercial software that have proven popular: in the 2015 survey 10-20% of respondents reported using the free versions of tools from KNIME, RapidMiner, and Revolution Analytics.

TOOL USE



While there is significant diversity in tool usage, most tools are rated very positively by users. A look at the most commonly used primary tools indicates that at least three quarters of users of all tools rate themselves as "extremely satisfied" or "satisfied" with their primary package. Among the top echelon of highly satisfied users are KNIME, SAS JMP, IBM SPSS Modeler, R, and STATISTICA.

TOOL SATISFACTION



Analysis of concurrent tool use is shown in the figure below, and it reveals some interesting patterns. Consistent with our discussion of R above, users of most other tools are also likely to use R at least occasionally. STATISTICA users are the least likely to use R (67%). STATISTICA users are also less likely to use Tableau and SQL than are users of other tools. Not surprisingly, users of one IBM, SAS, or MS tool are more likely than others to use another tool in that line of products, and users of a free tool (such as WEKA or KNIME free) are more likely than others to use another free tool. Interestingly, however, users of R are more likely to use IBM SPSS Statistics (32%) and SAS (30%) than they are to use other free tools.

CONCURRENT TOOL USE

Antecedent \ Consequent	R	IBM SPSS Statistics	SAS	MS Excel Data Mining	Tableau	IBM SPSS Modeler	Weka	Matlab	KNIME (free)	Rapid Miner (free)	MS SQL Server Data Mining	SAS Enterprise Miner	STATISTICA
R		32%	31%	30%	28%	23%	22%	22%	22%	19%	20%	18%	11%
IBM SPSS Statistics	78%		35%	40%	28%	56%	20%	24%	19%	23%	25%	22%	15%
SAS	79%	37%		31%	35%	30%	15%	23%	18%	20%	24%	55%	14%
MS Excel Data Mining	80%	44%	32%		35%	29%	22%	33%	24%	24%	46%	21%	18%
Tableau	86%	36%	42%	41%		27%	23%	27%	22%	25%	29%	26%	11%
IBM SPSS Modeler	75%	75%	38%	36%	28%		18%	24%	24%	25%	27%	28%	14%
Weka	88%	33%	23%	33%	30%	22%		36%	41%	41%	21%	18%	17%
Matlab	85%	37%	34%	46%	33%	28%	34%		29%	29%	29%	19%	22%
KNIME (free)	88%	30%	27%	35%	28%	29%	41%	30%		40%	26%	19%	14%
Rapid Miner (free)	86%	42%	35%	41%	35%	34%	47%	35%	46%		33%	23%	19%
MS SQL Server Data Mining	84%	44%	39%	73%	40%	35%	23%	33%	28%	31%		28%	14%
SAS Enterprise Miner	79%	39%	92%	33%	36%	37%	20%	21%	20%	22%	28%		14%
STATISTICA	67%	37%	33%	41%	21%	25%	26%	34%	21%	25%	19%	19%	

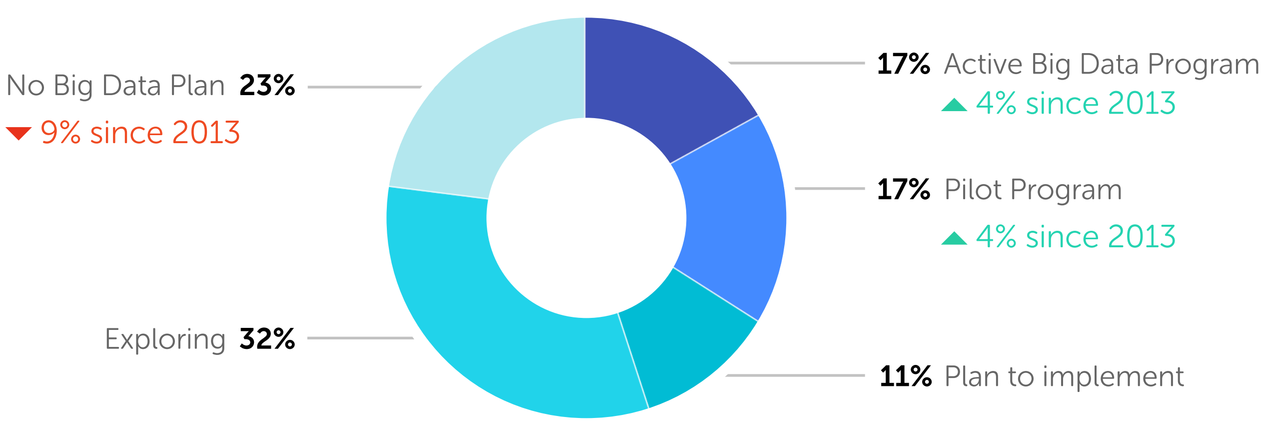
Example: Among Weka users, 88% use R, 41% use the free version of KNIME, 41% use the free verion of RapidMiner, but only 17% use STATISTICA. Among R users, 32% use IBM SPSS Statistics and 31% use SAS.

Big Data

Big data is generally defined as datasets that are so large and complex as to complicate the processes of data capture, storage, and analysis. People have been worrying about the implications of big data for decades. According to Gil Press, writing in Forbes.com, the first mention of the "information explosion" and identification of concerns about the rate of our expanding knowledge base occurred in the 1940s. In the 1960s, 70s, and 80s researchers began in earnest to address the issues that were beginning to arise due to the rapidly expanding information landscape. Concepts such as data compression, assessment of information volume, and information deletion were all explored in these decades.[7] The term "Big Data" itself appears to have been introduced in the late 1990s[8] and efforts to address large datasets and their complications have exploded in the past two decades.

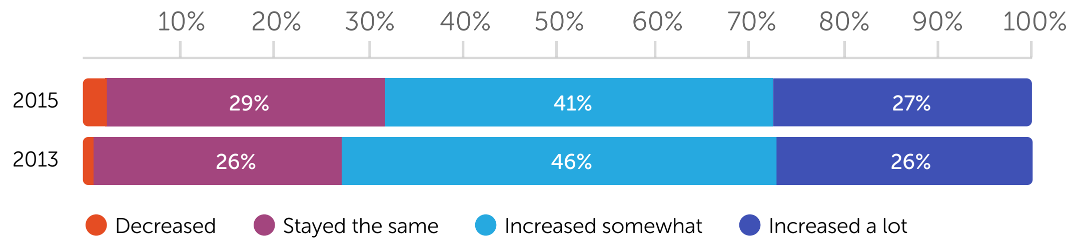
In recent years the literature has abounded with articles focused on the ever-increasing size of datasets, their perils, and the necessity of creating new systems, tools, and algorithms to deal with what is seen as the coming information overload. Data professionals in our 2015 study on one hand confirm the hype about Big Data's growth with the vast majority of organizations at least exploring how to utilize Big Data. These proportions are also substantially higher than in 2013.

STATUS OF BIG DATA IN ORGANIZATIONS



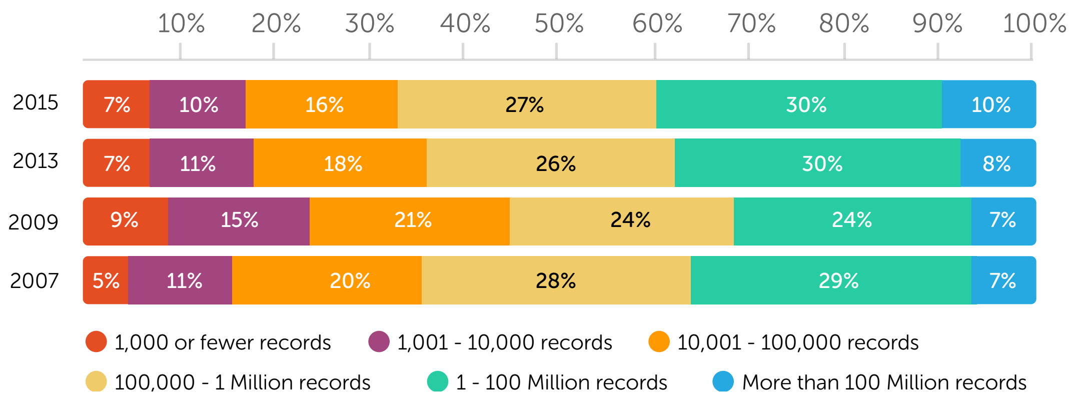
Further, 2013 and 2015 respondents perceive that there have been recent significant increases in the sizes of their datasets: 27% of 2015 respondents report that data volume has "increased a lot".

PERCEPTION OF CHANGES IN DATA SETS SIZE



However, Rexer Analytics' surveys have been tracking dataset sizes for close to a decade, and in recent surveys have specifically asked questions about "Big Data," the challenges it poses, and how people are addressing those challenges. We have found, that despite the Big Data hype, and the very real challenges faced by the minority of data scientists, the actual sizes of datasets that data science professionals say they are using have remained remarkably stable since 2007. In 2007, 7% of respondents reported that their datasets typically contained over 100 million records. By 2015 we see that the proportion of people working with datasets of this size has only grown to 10%.

TYPICAL SIZE OF DATASETS



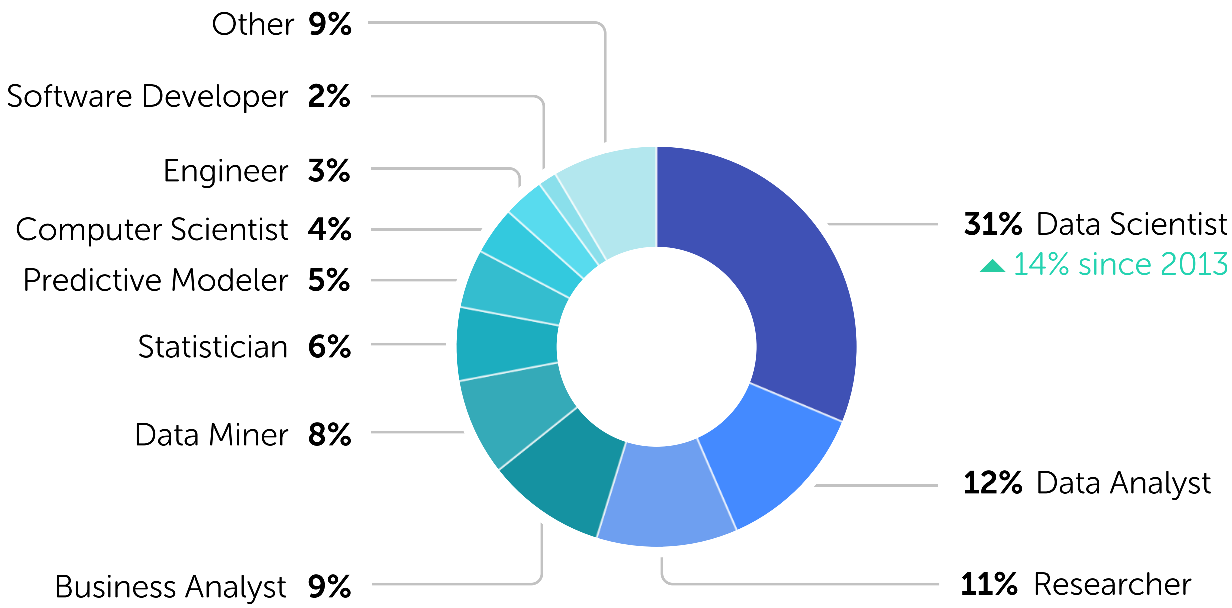
Among analytic professionals working with Big Data, the most common Big Data analyses are predictive modeling, segmentation, and data visualization.

TYPES OF ANALYSIS USING BIG DATA

Who We Are

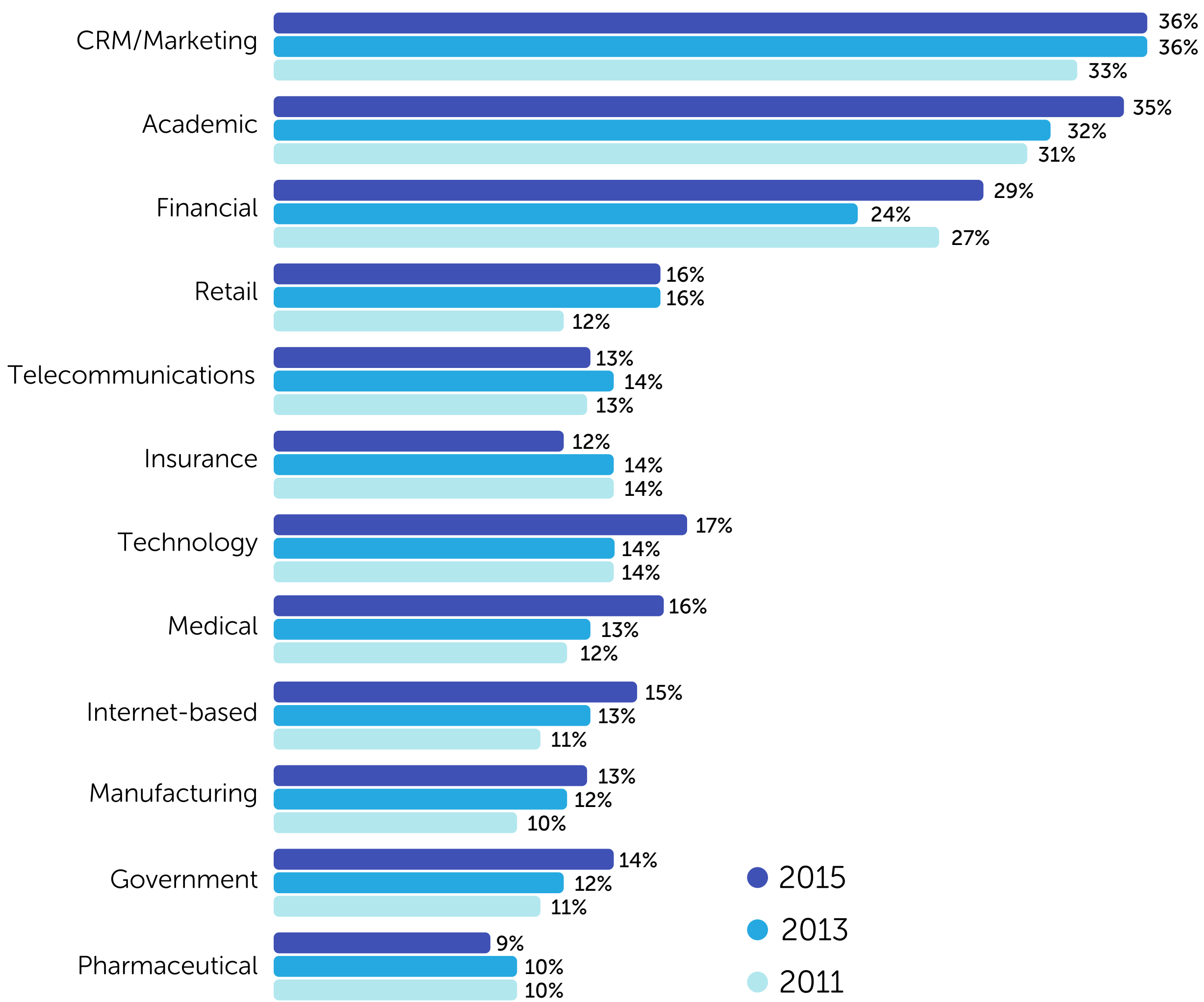
In a field that is ever evolving, how those who work in the field define themselves is also changing rapidly. In just two years, the proportion of us describing ourselves as "data scientists" rose from 17% to 31%, with corresponding declines in the terms "researcher," "business analyst," "statistician," and "predictive modeler."

HOW ANALYTICS PROFESSIONALS DESCRIBE THEMSELVES



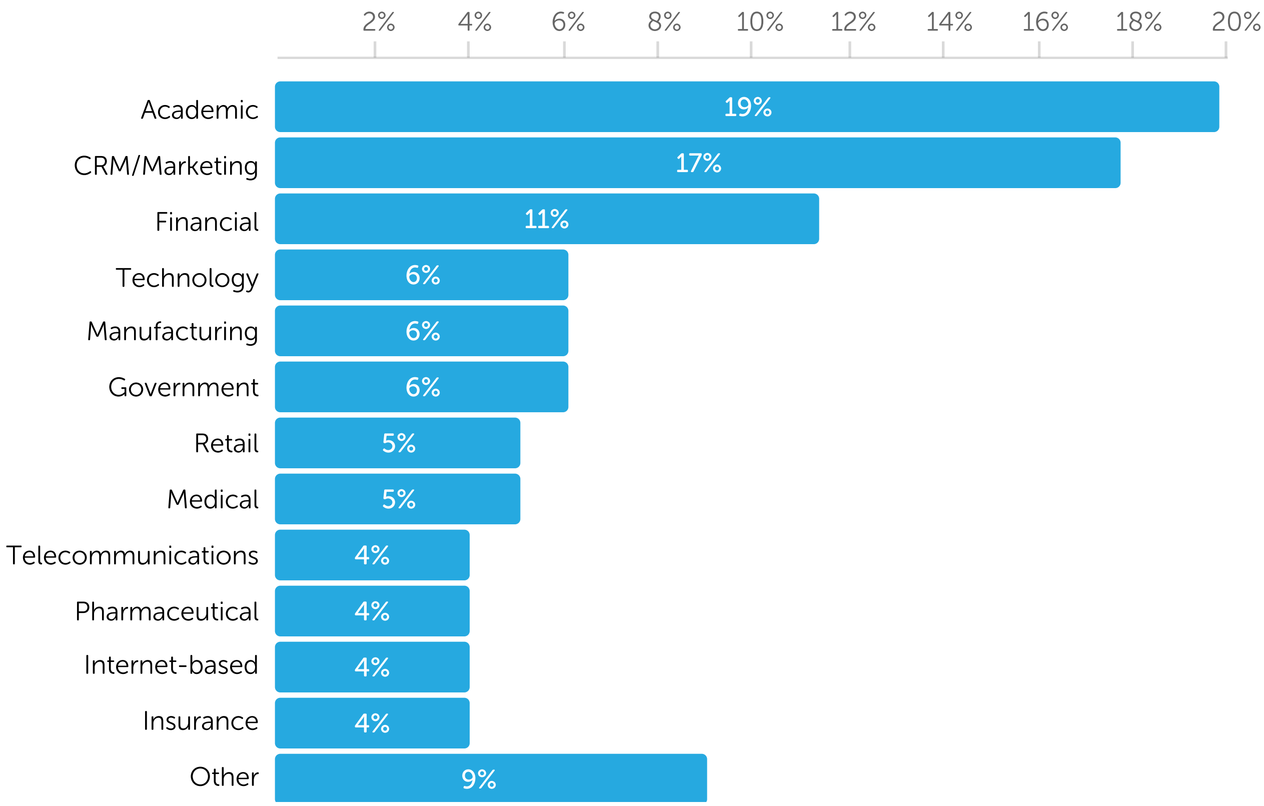
We work in a variety of fields, with over a third of us spending at least some of our time working with CRM/Marketing data. In each of the 7 Rexer Analytics Surveys, more people report applying their analytics in the field of CRM / Marketing than any other field. Respondents reported working in an average of three fields over the course of the past year.

FIELDS WHERE ANALYTIC PROFESSIONALS ARE WORKING



The 2015 survey also asked people how they allocate their time across these fields. An analysis of people's time allocation revealed the same top three fields: across all responders, people spend the most time applying analytics in Academics, CRM/Marketing and Financial Services.

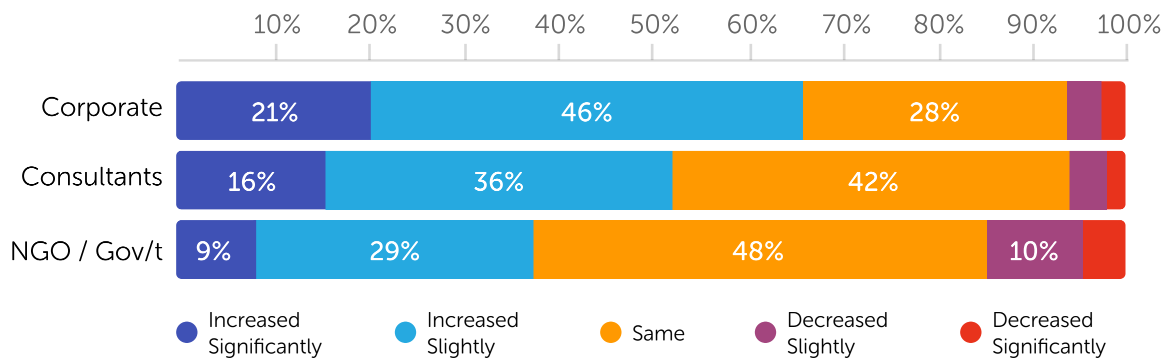
ANALYTIC TIME ALLOCATION ACROSS INDUSTRIES



How We Are

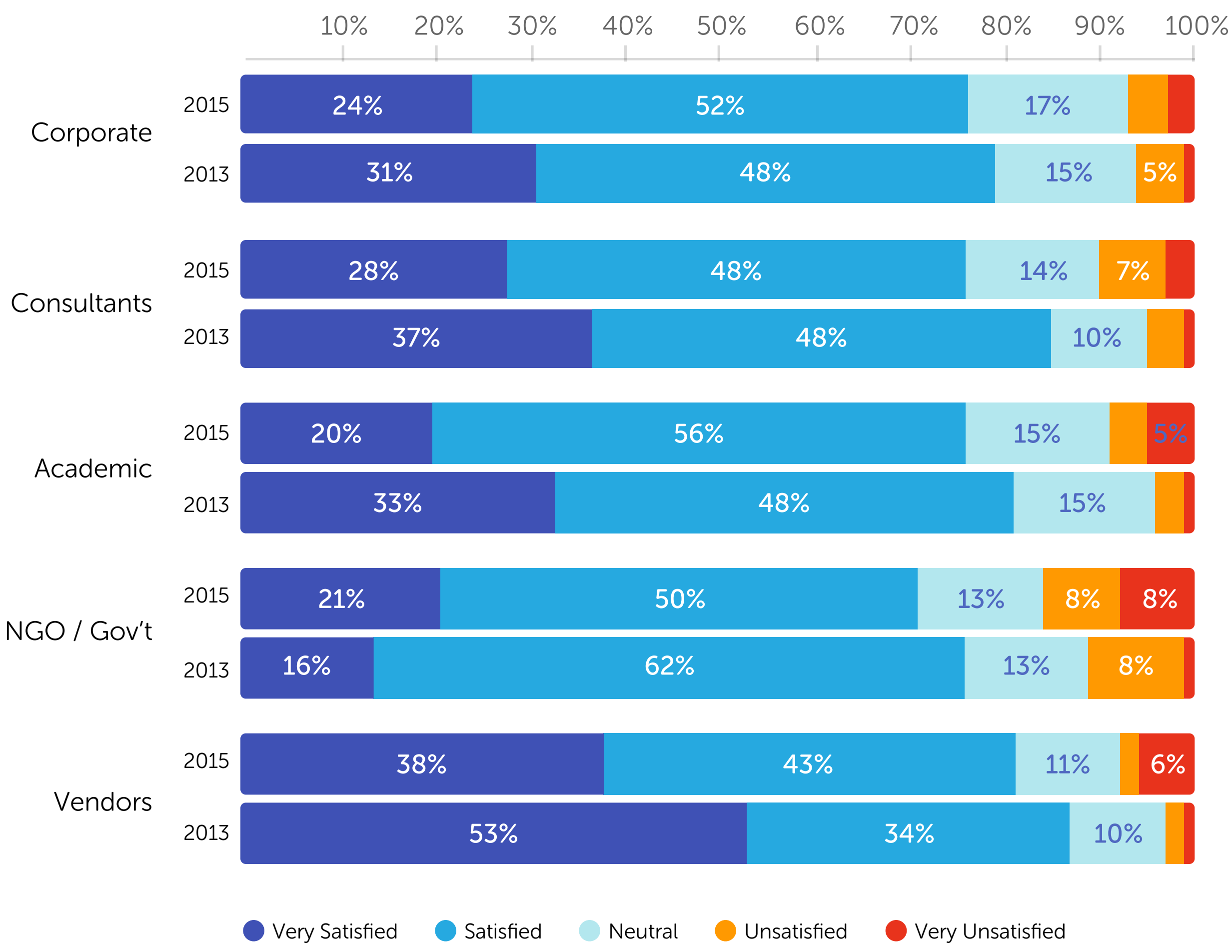
In the past few years, "data science" has also leapt to the top of several "best careers" lists.[9] Fueled by the ever increasing availability of larger transactional data sets (60% of respondents in our study reported working with customer transaction data) and relatively high median incomes, data scientists are increasingly in demand.[10] Approximately two thirds of respondents in corporate environments reported at least some growth in the size of analytic staff at their companies. Growth is slower in government and non-profit sectors, but there are evolving opportunities there as well.

SIZE OF ANALYTIC STAFF



While our study also reveals high levels of job satisfaction, there were small satisfaction declines from 2013 to 2015. Data professionals who had been in the field for 10 or more years had the highest rates of being "very satisfied" (30% versus 20% for those in the field less than 10 years). Perhaps, as more people enter a field being widely touted as desirable, it begins to attract those who discover that the work may be more difficult than they imagine, both in terms of knowledge and skills required to do it well. Further, the challenges to feeling fulfilled as a data analytic professional are wide and varied, as we address in the next section.

JOB SATISFACTION



What Gets in Our Way

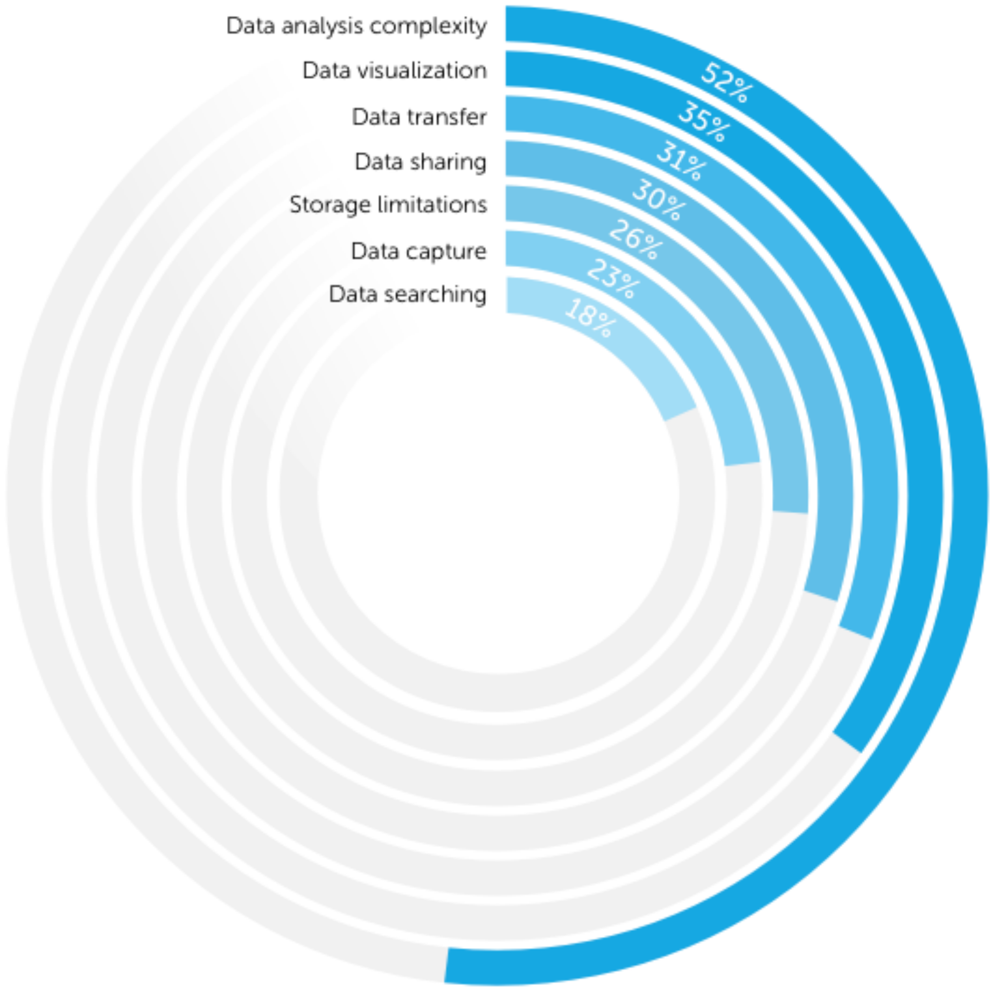
Barriers and Challenges Data Scientists Face

In his article on data science workflow, Phillip Guo describes the challenges that data scientists face during each step of the process.[11] Guo identifies several key data hurdles: data provenance, determining whether data is current, organization and storage, and data combination and cleaning (particularly as it impacts maintenance of data integrity). Analytic issues include runtime and the identification of errors, as well as further issues with data organization and management. Finally, Guo describes the difficulties of data dissemination and validation. We see similar challenges reported by the analytic professionals we've been surveying. Over the years, the Rexer Analytics surveys have asked data scientists about the barriers and challenges they face to the success of their work. Dirty data, inconsistent access to data, explaining their work to others, and deployment issues have consistently ranked among the top challenges.

BIGGEST CHALLENGES

Additionally, data scientists report new challenges due to the rapidly changing nature of their work. Although increases in data size and availability have brought great opportunity, some data scientists face challenges such as higher complexity in data analysis, difficulties with data visualization, and hurdles in sharing, transferring, and storing their larger data sets.

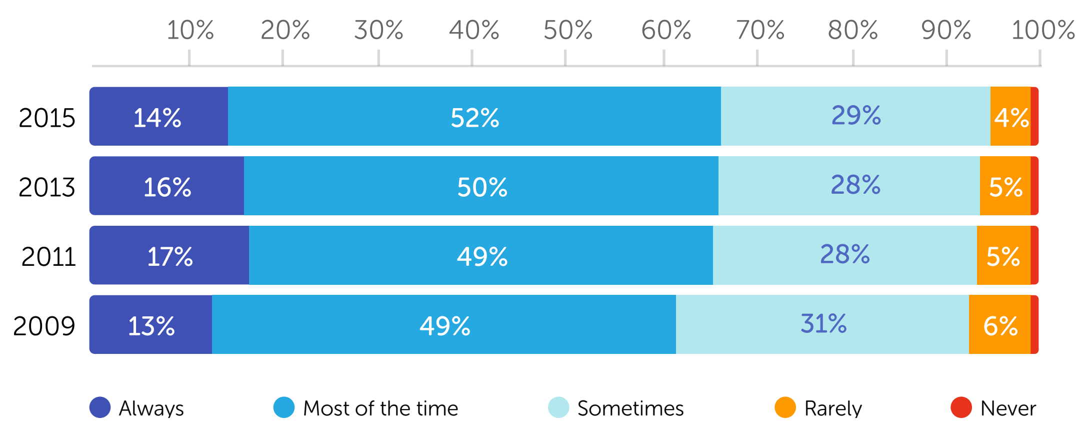
BIG DATA CHALLENGES



Model Deployment

While some models are created for academic, descriptive, or exploratory purposes, many models are designed to be deployed. Many data scientists have identified deployment of analytic projects as an ongoing challenge. Despite advances in many technology areas, deployment challenges have not improved in recent years. Very few people report their results are "always" deployed (13% in 2015). Only half of respondents report that their analytic results are deployed "most of the time." And about a third of respondents report their results are deployed "sometimes" or less often.

FREQUENCY OF MODEL DEPLOYMENT



Forty percent of data scientists report that their models are deployed within days of completing their analyses. However, for over 25% of people, model deployment takes months or years. See the appendix for a summary of both the latencies people report between data capture and data availability and the latencies people report between analysis completion and deployment.

DEPLOYMENT CHALLENGE

buy-in

real time scoring

technology

Respondents report a variety of barriers to successful deployment. In the 2015 survey, respondents most consistently identify the difficulty of getting buy-in from customers, managers, and other stakeholders as the primary factor preventing deployment. Data scientists often have to work extensively with stakeholders in order to explain their models and help them to understand the potential value:

> *The Marketing team that hired us didn't see much value in predictive analytics (they were interested more on descriptive analytics). But I asked them to give us a try, and once we showed results with a high accuracy in our predictions, they became convinced and hired us again for a new project the following year.*

> *[Our key challenge was] stakeholder understanding and buy-in. We overcome it by further explanation and subsequent success of the model in real time.*

Others described various technical difficulties with data translation for deployment:

> *The logic/algorithm was developed in Excel and the difficulty was to scale it to the big-data environment. The same logic was developed in R and then deployed in the big-data environment.*

> *Our operational systems are really old and hard to maintain, so we have to simplify our models to use them. There is a lot of binning and rounding of data to get the models to work.*

Data Scientist job satisfaction is higher at companies that deploy their analytic projects

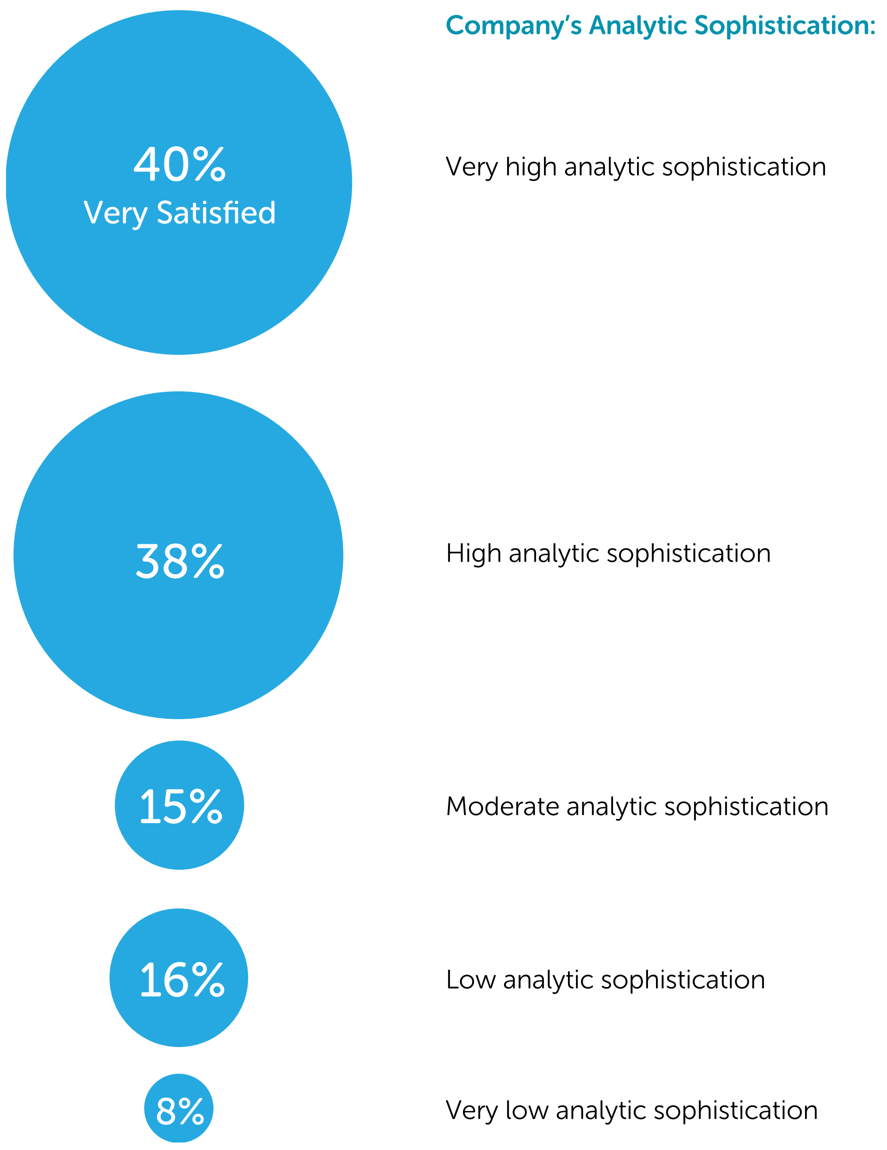
Respondents also reported difficulties interfacing with their IT department in order to translate models into organizational processes. Finally, they described incompatibility across multiple tools, communication problems, and security issues, particularly with cloud-based solutions.

The ability to successfully deploy models has a significant impact on job satisfaction. Those respondents who report that their algorithms are deployed “always” or “most of the time” are three times more likely to be “very satisfied” with their job than are those whose algorithms are deployed “never” or “rarely”.

MODEL DEPLOYMENT AND JOB SATISFACTION

Further, data scientists often work in environments where there is low analytic sophistication. Close to six in ten data scientists working in corporate environments (59%) report that their companies have only moderate to very low levels of analytic sophistication. In government and non-profit settings it is worse, with more than three quarters (77%) reporting moderate to very low levels of sophistication. Workplace analytic sophistication in turn appears to be a driver of job satisfaction: reports of being "very satisfied" are more than twice as common for those in organizations with a high level of analytic sophistication.

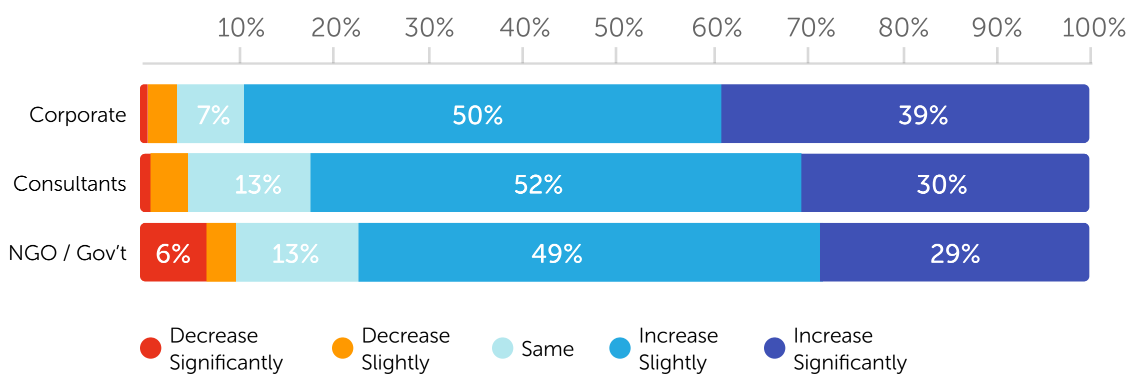
WORKPLACE ANALYTIC SOPHISTICATION AND JOB SATISFACTION



Where We Go from Here

The demand for analytic projects is growing rapidly everywhere. The great majority (89%) of analytic professionals working in corporate settings report the number of analytic projects their companies are planning this year is greater than the number of projects conducted in recent years. Most people working in consulting and NGO / Government settings also report plans to increase the number of analytic projects.

THE DEMAND FOR ANALYTICS PROJECTS IS INCREASING



In a growing number of companies, analytics is no longer in a supportive role. Increasingly, data-driven executives are making analytics core to all company strategies and decisions. With access to databases of a size and breadth which would have been unimaginable 30 years ago, and a corresponding expansion of software tools able to handle such complex datasets, data science has fully embedded itself into the workings of public endeavors and private industry. Analytics have moved from a "nice to know" supplementary role to a core component of organizational structure and strategy.

Operationally & Strategically – Analytics is now central to many companies

There is no end in sight to the avalanche of opportunity data science proffers. As Tom Davenport of The International Institute for Analytics stated, “High-performing companies will embed analytics directly into decision and operational processes, and take advantage of machine-learning and other technologies to generate insights in the millions per second rather than an ‘insight a week or month’.”[12] Data science’s ability to swiftly chart a course for growth and efficiency will be one of the key differentiators between the most highly actualized organizations and those stuck with archaic, inefficient processes.

The rapidly evolving demand for insights and predictive abilities has, of course, pushed the demand for data scientists to fill these roles. In a 2015 Forbes report, 84% of marketing leaders indicated that they intend to increase the role of predictive analytics in marketing over the next 12 months.[13] Our study confirms this prediction is becoming a reality, with 67% of data scientists who work at corporations indicating that the size of their analytic staff has increased in the past year. Data science has become one of the core requirements for a successful business and one of the most desirable new careers for young professionals.

Skilled Data Science professionals are in high demand — And this will continue

As with many professions that experience a spike in demand but require specialized skills to enact, there is expected to be a substantial gap between the demand for and the supply of competent data scientists in the near future. A 2014 McKinsey Global Institute Report cites “big data” as one of the five “game changers” for US economic growth over the next decade.[14] They also believe that there may be as much as a 300,000 person shortfall from the need for 500,000 data scientists to support this revolution of data leveraging. While many colleges and universities are in the process of creating data science and business analytics degree programs, the full actualization of these programs is years away.

Attempting to bridge this personnel gap opens the industry to potential risks. Paramount among these are the need to engage professionals whose background has not prepared them to conduct analytics that are beyond their skill set. While advancements in software tools in the automation of both data processing and analysis has opened the door for wider access to users of various backgrounds, there remain advantages to having professionals well-versed in research methodology and statistical analysis at the helm.

Our study also points to other hurdles data science must traverse to reach full value. The frequency and speed of model deployment remains a significant area of opportunity. While corporate executives recognize the critical need for data science, many organizations have not culturally advanced to an optimal level of analytic sophistication or adoption. Within the data science field itself, there are questions about how open source tools such as R will co-exist with a wide array of other tools: will they be pure competitors, or potentially play complementary roles? Correspondingly, how widely will data scientists need to be trained, or at least familiar with, various analytic platforms? Finally, where will newer, more complex algorithms such as ensemble modeling, uplift modeling and deep learning find the most value, and conversely where will they experience the greatest barriers to adoption?

Despite the significant challenges ahead, the field of data science has never been so exciting and filled with potential value. This formerly dry, back-office, and largely opaque activity has been dragged out into full view and landed in media and public discourse. Data science is everywhere, from the impact on our daily Internet activities, to targeted campaigning in Presidential elections, to the efficiency of our thermostats.

Citations

1 Bi, Ran. (2014), *Will Deep Learning take over Machine Learning, make other algorithms obsolete?*, KDnuggets, October, 2014. Retrieved from www.KDnuggets.com

2 Ihaka, R (1998), *R: A Past and Future History*. Draft of a paper for Interface '98. Retrieved from www.stat.auckland.ac.nz

3 *What is R? The R Project*. Retrieved from https://www.r-project.org/about.html

4 Piatetsky, G. (2015), *R leads RapidMiner, Python catches up, Big Data tools grow, Spark ignites*. KDnuggets. May 2015. Retrieved from www.KDnuggets.com

5 Muenchen, R. *The Popularity of Data Analysis Software*. Retrieved from http://r4stats.com/articles/popularity/

6 Rexer, K., Allen, H., & Gearan, P. (2013), *Rexer Analytics 2013 Data Miner Survey*. Retrieved from www.RexerAnalytics.com

7 Press, G. (2013), *A Very Short History of Big Data*. Forbes, May 9, 2013. Retrieved from www.Forbes.com

8 Lohr, S. (2013), *The Origins of 'Big Data': An Etymological Detective Story*. The New York Times, February 1, 2013. Retrieved from http://bits.blogs.nytimes.com

9 Glassdoor (2016), *25 Best Jobs in America for 2016*. Glassdoor, January 19, 2016. Retrieved from www.Glassdoor.com

10 DeZyre (2016), *Data Scientist Salary Report of 100 Top Tech Companies*. DeZyre, February 24, 2016. Retrieved from www.dezyre.com

11 Guo, P., (2013), *Data Science Workflow: Overview and Challenges*. Communications of the ACM, October 30, 2013. Retrieved from http://cacm.acm.org/blogs

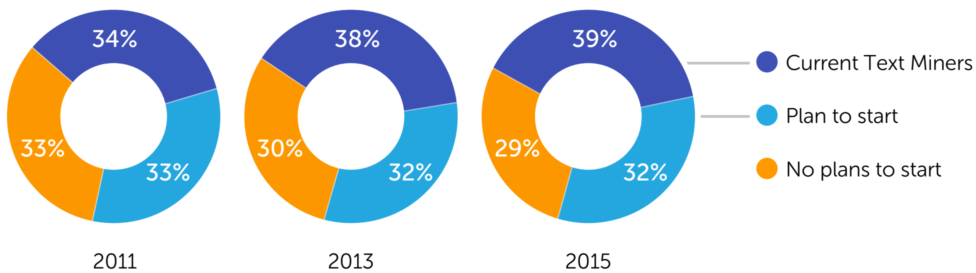
12 Davenport, T. (2013), *What is Analytics 3.0?*, Retrieved from http://iianalytics.com/analytics-resources/analytics-3.0

13 Forbes Insights (2015), *The Predictive Journey: 2015 Survey on Predictive Marketing Strategies*. Forbes Insights, October 2015. Retrieved from http://www.forbes.com/forbesinsights/lattice_engines/

14 Lund, S., Manyika, J., Nyquist, S., Mendonca, L., & Ramaswamy, S. (2013), *Game Changers: Five Opportunities for US Growth and Renewal*. McKinsey Global Institute, July 2013. Retrieved from www.McKinsey.com

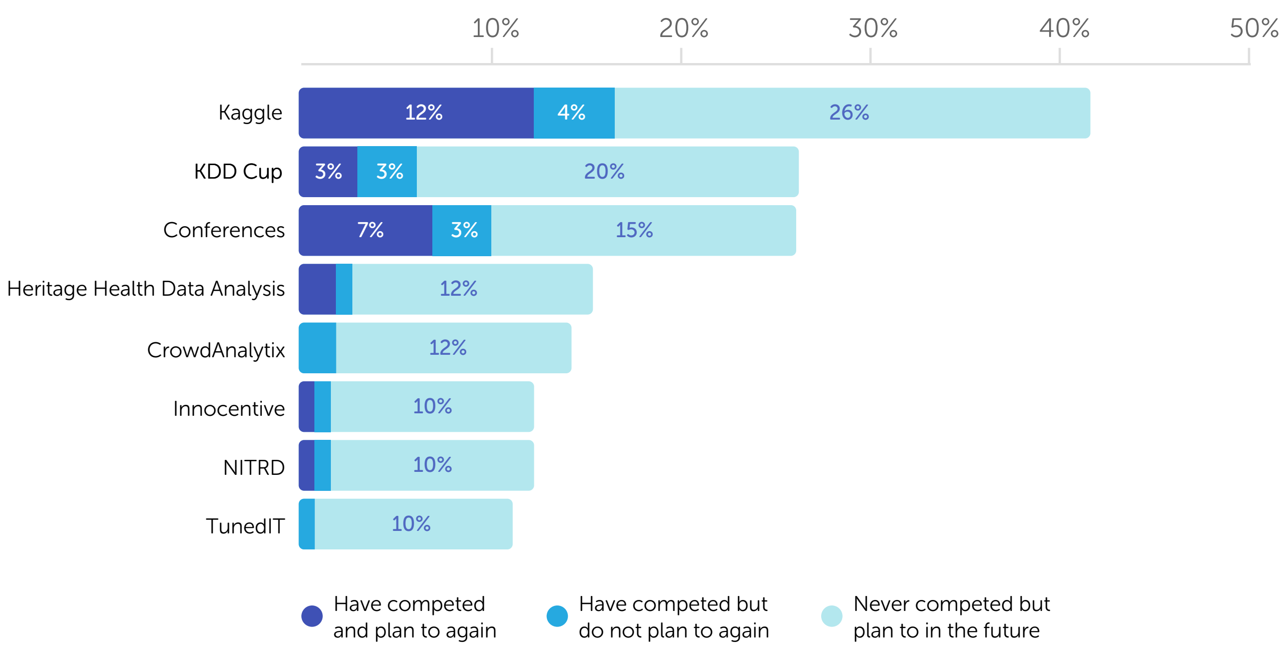
Appendix

TEXT MINING



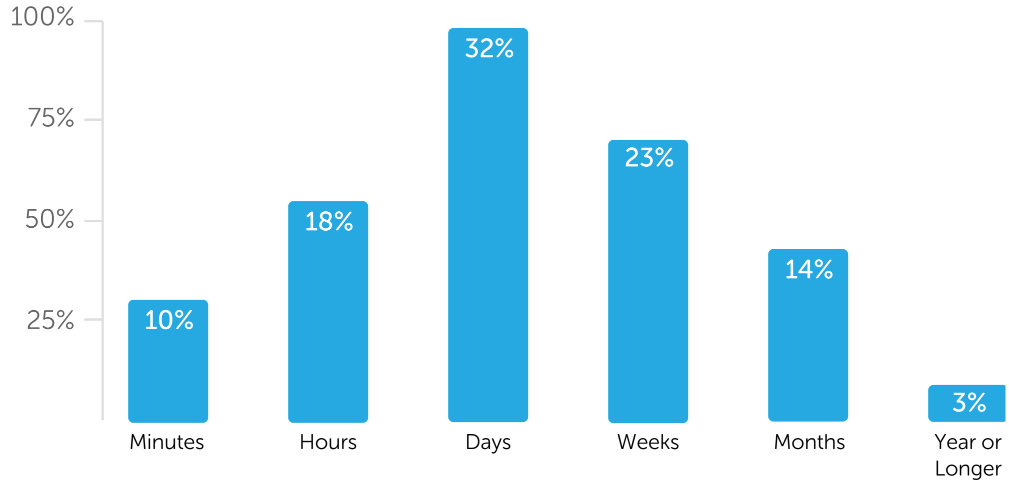
Text mining adoption has slowly but steadily increased since 2011, to its present state, where 39% of analytic professionals incorporate text mining into their analyses.

PARTICIPATION IN ANALYTIC COMPETITIONS



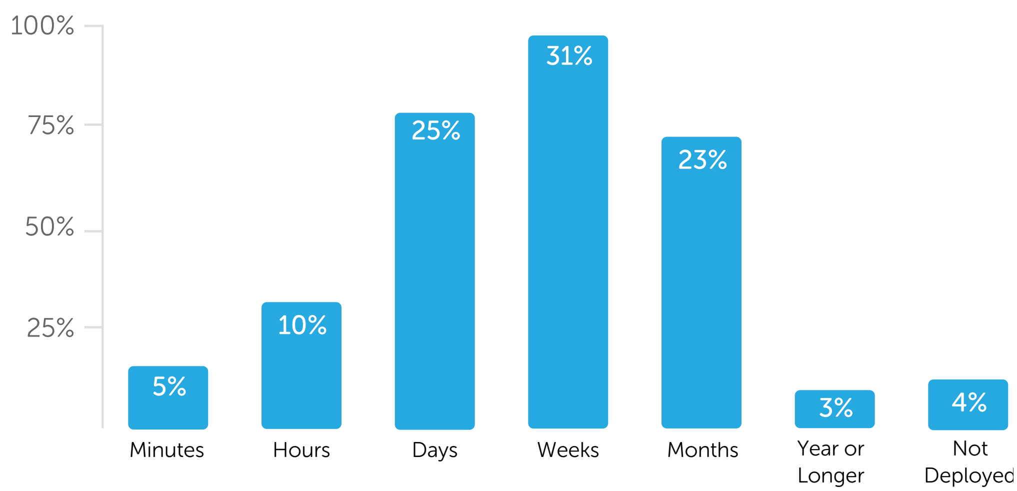
Many analytic professionals have participated in analytic competitions or have plans to.

TIME BETWEEN DATA CAPTURE AND AVAILABILITY FOR ANALYSIS



For 60% of data scientists, data is available for analysis within a few days of data capture.

TIME BETWEEN ANALYSIS COMPLETION AND DEPLOYMENT



Forty percent of data scientists report that their models are deployed within days of completing their analyses. However, for over 25% of people, model deployment takes months or years.

REXER ANALYTICS

I highly recommend Rexer Analytics. Very few quant people possess the total package: excellent credentials, strong statistical and data mining skills, the initiative to anticipate client needs, compelling and concise communications, strong client management skills, and a results-driven mindset. They've led numerous projects that have helped our company maximize marketing efforts and understand our customers better. In all, the quality has been superior and their client-service attitudes are tremendous.

— **A. Charles Thomas**, PhD; Senior Manager, Advanced Analytics, Hewlett-Packard

Rexer Analytics hit a grand slam for us! Their expertise and mastery of data mining and predictive analytics found important patterns in our data where others had failed. They built multiple predictive models that predicted things like student success, the likelihood of a student dropping out of college, and the likelihood of a student defaulting on their student loans. And Karl was outstanding when presenting the information to non-technical audiences. His slides and explanations were clear and resonated with the audience. Karl and his team were amazing to work with and I would hire them again in a heartbeat.

— **Robert Reeder**, CIO, Cegment, Inc.

Rexer Analytics assisted our international PwC benchmarking team in the production of our 2013 and 2015 Corporate Finance white papers. They did an awesome job, as always. As good as Karl and his team of Paul and Heather were two years ago, they upped the result this time around. Their data analysis, ad hoc data fulfillment, writing, editing, proofing, project management, dozens and dozens of interviews and, not the least, positive attitude and continuous assistance were tremendous.

— **Ed Shapiro**, Director, Finance Effectiveness, PricewaterhouseCoopers

The models created by the Rexer Analytics team enabled us to target our retention efforts and take immediate action to retain at-risk customers. The insights from their analyses also enabled us to make strategic decisions which boosted customer lifetime value.

— **Betsi Harris**; Director, Operational Excellence, Training & Quality Assurance/ Compliance, Tyco Integrated Security; Author of "Transactional Six Sigma and Lean Servicing: Leveraging Manufacturing Concepts to Achieve World Class Service"

www.RexerAnalytics.com
+1 617-233-8185
DataScienceSurvey@RexerAnalytics.com

Graphic design assistance provided by Brainspin