A Decade of Surveying Analytic Professionals:
2017 Survey Highlights

Karl Rexer, PhD
President, Rexer Analytics

www.RexerAnalytics.com
2017 Data Science Survey: Overview

- 8th survey since 2007
- 67 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 1,123 analytic professionals from 91 countries
- Data collected in first half of 2017
2017 Data Science Survey: Overview

- 8th survey since 2007
- 67 questions
- 10,000+ invitations emailed & promoted by newsgroups, vendors and bloggers
- Respondents: 1,123 analytic professionals from 91 countries
- Data collected in first half of 2017

Pie chart showing the distribution of respondents by region:
- North America: 41%
- Europe: 37%
- Asia Pacific: 11%
- Central & South America: 7%
- Middle East & Africa: 4%
Formal Data Science Training is Important

Do you agree or disagree with the following statement? You need to have formal training in data analytics in order to properly model data.

The majority of respondents agree that formal data training is needed to properly model data.

Top problems untrained staff have:
- Poor data preparation
- Mis-interpreting results

Corporate: 74%*
Consultants: 75%
 Academics: 78%
 NGO / Gov’t: 68%
 Vendors: 72%*
Do-it-yourself Tools Can Cause Problems

Have there ever been difficulties when people at your company outside the data science team have utilized these do-it-yourself tools or services?

One third of respondents have seen difficulties when do-it-yourself tools or services are used*

- “Typically, they blow up the ship”
- “People have run analyses that resulted in completely false conclusions, which were then used by the business”
- “A large part of my job is cleaning up those problems”
- “It’s like giving a child a bazooka. Easy to pull the trigger, hard to know what will happen in advance.”

<table>
<thead>
<tr>
<th>Corporate</th>
<th>Consultants</th>
<th>NGO / Gov’t</th>
<th>Vendors</th>
</tr>
</thead>
<tbody>
<tr>
<td>32%</td>
<td>34%</td>
<td>35%</td>
<td>30%</td>
</tr>
</tbody>
</table>

*55% of people report that self-service analytics are being used in their workplaces. Among these people, 59% say that these tools have led to problems.
Most Data Scientists use Multiple Tools

What data science / analytic tools, technologies, and languages did you use in the past year?

Corporate
- SQL: 71%
- R: 70%
- Python: 53%
- Tableau: 38%
- KNIME (free version): 38%
- Hadoop / Hive / Pig: 34%
- Java: 26%
- Microsoft Excel Data Mining: 25%
- SAS Base: 24%
- Spark / MLlib: 21%

Consultants
- R: 73%
- Python: 58%
- SQL: 56%
- IBM SPSS Statistics: 38%
- Tableau: 35%
- KNIME (free version): 35%
- Java: 31%
- Hadoop / Hive / Pig: 28%
- Microsoft Excel Data Mining: 27%
- IBM SPSS Modeler: 27%
- MATLAB: 27%
- RapidMiner (free version): 25%
- Spark / MLlib: 24%
- C / C++: 23%
- Weka: 22%
- STATISTICA: 22%
- SAS Base: 21%

Academics
- R: 65%
- MATLAB: 48%
- STASTICA: 47%
- IBM SPSS Statistics: 40%
- Python: 40%
- Microsoft Excel Data Mining: 32%
- Java: 30%
- SQL: 25%
- C / C++: 24%
- Weka: 21%

NGO / Gov’t
- R: 70%
- SQL: 52%
- Python: 44%
- KNIME (free version): 38%
- Microsoft Excel Data Mining: 35%
- Java: 35%
- STATISTICA: 31%
- Tableau: 30%
- IBM SPSS Statistics: 28%
- Microsoft SQL Server Data Mining: 26%
- Hadoop / Hive / Pig: 25%
- Weka: 22%
- MATLAB: 21%

All tools used by more than 20% of a group are shown.
A Decade of Dramatic Growth in the Use of R

The biggest change in tool adoption we’ve seen over the past decade has been the dramatic growth in the use of R.
RStudio has Become the Dominant R Interface

Most people using R utilize RStudio as their primary interface to R. In the 2017 survey responses, this grew to 69% (we just don’t have the graphic updated yet).

Vendors are excluded from this analysis.
What one data science tool, technology or language do you use most frequently?

**Corporate**
- R: 23%
- Other: 21%
- SQL: 4%
- IBM SPSS Modeler: 5%
- KNIME (commercial version): 6%
- STATISTICA: 9%
- SAS Base: 10%

**Academics**
- R: 26%
- Other: 25%
- STATA: 28%
- Weka: 4%
- Rapidminer (free version): 4%
- MATLAB: 4%
- KNIME (free version): 11%
- IBM SPSS Statistics: 5%

**Consultants**
- R: 28%
- Other: 21%
- Tableau: 4%
- IBM SPSS Statistics: 4%
- SAS Base: 5%
- IBM SPSS Modeler: 6%
- Python: 8%
- STATISTICA: 11%

**NGO / Gov’t**
- Other: 28%
- R: 25%
- STATA: 17%
- IBM SPSS Statistics: 5%
- KNIME (free version): 16%
- Python: 5%
- Weka: 4%
There is a growth in the use of Deep Learning. But it’s still only being used by a small proportion of analytic professionals.

**What algorithms/analytic methods do you use?:**

- Never: 68%
- Most of the time: 2%
- Often: 5%
- Sometimes: 12%
- Rarely: 13%

**Which Deep Learning algorithms have you used?**

- Convolutional Neural Networks: 65%
- Recurrent Neural Networks (RNN): 49%
- Deep belief networks: 23%
- Long Short-term Memory Recurrent Neural Networks: 20%
- Deep Boltzmann Machine: 17%
- Stacked Autoencoders: 14%
Selected Success Stories

- Medical imaging segmentation
- Analysis of histopathological images
- Colorization of black and white images
- Classification of 20,000+ species using crowd-sourced images
- Handwriting interpretation and voice recognition
- Closed vocabulary verbal intent determination
- Impact of weather on airport arrival and departure rates
- Chip defect identification in semiconductor industry
- Prediction of negative prices in energy market
- Estimation of the daily demand on ATM knowing historical demands, stock and recharge times

If you have had a great experience using Deep Learning on a project, please provide the details of that project
## Employment: It’s a Good Time to be a Data Scientist

### If you left your current role now, how easy would it be for you to find a job fairly equivalent to your current one?

<table>
<thead>
<tr>
<th></th>
<th>Very difficult</th>
<th>Somewhat difficult</th>
<th>Somewhat easy</th>
<th>Very easy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate</td>
<td>7%</td>
<td>25%</td>
<td>50%</td>
<td>18%</td>
</tr>
<tr>
<td>Consultant</td>
<td>8%</td>
<td>26%</td>
<td>40%</td>
<td>26%</td>
</tr>
<tr>
<td>NGO/Gov't</td>
<td>18%</td>
<td>37%</td>
<td>34%</td>
<td>12%</td>
</tr>
<tr>
<td>Academic</td>
<td>20%</td>
<td>31%</td>
<td>34%</td>
<td>15%</td>
</tr>
</tbody>
</table>

### What is your current level of job satisfaction?

<table>
<thead>
<tr>
<th></th>
<th>Very Unsatisfied</th>
<th>Unsatisfied</th>
<th>Neutral</th>
<th>Satisfied</th>
<th>Very Satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corporate</td>
<td>8%</td>
<td>22%</td>
<td>43%</td>
<td>23%</td>
<td></td>
</tr>
<tr>
<td>Consultants</td>
<td>7%</td>
<td>13%</td>
<td>47%</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>NGO/Gov't</td>
<td>9%</td>
<td>17%</td>
<td>47%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>Academics</td>
<td>16%</td>
<td>47%</td>
<td>29%</td>
<td>23%</td>
<td></td>
</tr>
</tbody>
</table>
Analytic Sophistication: Opportunity for Improvement

In general, with what degree of sophistication does your company/organization approach analytic problems?

Only a minority of data scientists feel that their firms have high analytic sophistication.

Interpretation: When a company is judged to have very high analytic sophistication, 59% of analytic staff report being very satisfied with their jobs.
Data Science Skills Needed: Varied and Growing

Please rank the following in terms of which are the most important attributes/background/skills for a data miner/data science professional to possess?

- Data preparation and management skills: 37%
- Domain knowledge: 35%
- General business experience or knowledge: 28%
- Math background: 22%
- Classical statistics: 22%
- Knowledge of a breadth of algorithms: 19%
- Programming knowledge: 17%
- Database skills: 14%
- Data visualization skills: 12%
- Experience with a breadth of tools: 10%
- Deep experience with a single tool: 5%

What is the most important step you would like to take to advance your career?

- Learn new technologies/skills/algorithms: 46%
- Work on more interesting/important projects: 22%
- Move into a manager/leadership role: 12%
- Switch companies: 6%
- Learn more business skills: 6%
- Learn more about the industry: 5%

Top 2 rank, 2105 Survey
Data Science Survey

Rexer Analytics designed and began this research in 2007. The eight surveys from 2007 through 2017 examine the analytic behaviors, views and preferences of data analytic professionals. We use a small set of consistent questions to enable the tracking of trends across the years. We also incorporate new questions into each survey to explore emerging topics. We appreciate the many people who have submitted survey questions – many have been included in the surveys. Please send us your ideas and feedback about this research by using the form on our CONTACT page or emailing us at DataScienceSurvey@RexerAnalytics.com. The survey summary reports can be downloaded using the links below.

2017 Survey
2015 Survey
2013 Survey
2011 Survey + “Best Practices” Verbatim Responses
2010 Survey + “Best Practices” Verbatim Responses
2009 Survey
2008 Survey
2007 Survey

www.RrexerAnalytics.com
For additional information, please contact
Karl Rexer, PhD
President
krexer@RexerAnalytics
617-233-8185