CSMDM21 - Data Analytics and Mining

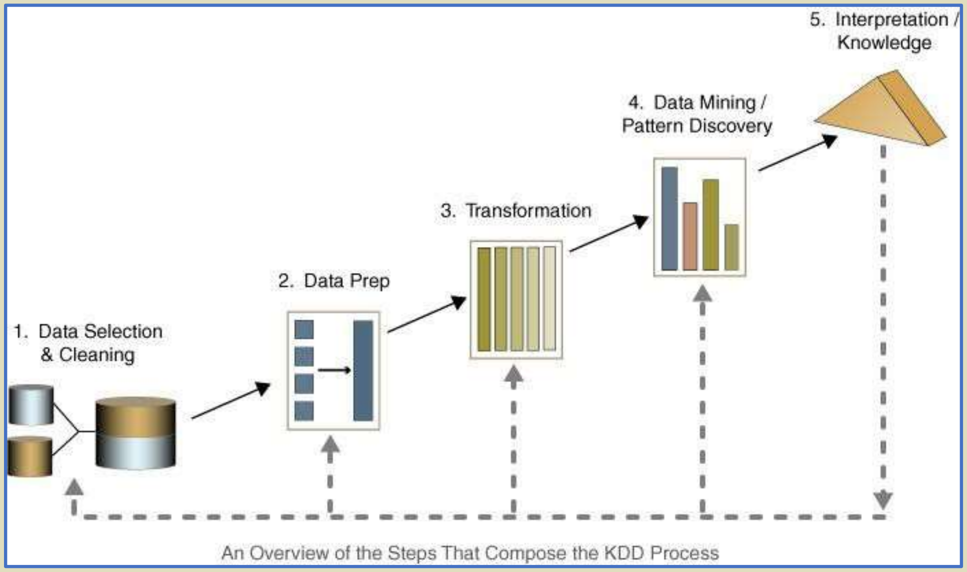
Introduction to Data Science Platforms

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos powered by
Prof. Giuseppe Di Fatta

Integrated Development Environments for Data Science

- Data access and manipulation: Extract, Transform and Load (ETL)
- Data Science/Data Analytics and Mining/Machine Learning: prediction, pattern discovery, etc.
- Information visualization and reporting
- The process of Knowledge Discovery from Databases (KDD)
  - Requirements to facilitate the design and deployment of KDD processes
  - **Data Science workflow platforms** integrate algorithms and tools for data manipulation, analytical, mining and visualization, as well as workflow design, management and deployment.


1. Data Selection & Cleaning
2. Data Prep
3. Transformation
4. Data Mining / Pattern Discovery
5. Interpretation / Knowledge

An Overview of the Steps That Compose the KDD Process

KDD Development Environments – Open Source

- **Weka** 3, Machine Learning algorithms in Java
- **Orange**, a component-based Data Analytics and Mining software (C++)
- **MLC++** is a library of C++ classes for supervised machine learning
- **D2K**, Data to Knowledge (Java)
- **KNIME**, Konstanz Information Miner (Java)

KDD Development Environments - Commercial

RapidMiner (formerly **YALE**, Yet Another Learning Environment) (Java) – free trial version available

- Pentaho – free trial/light version available

  Also free community edition: http://community.pentaho.com/

  (Note: Pentaho Data Analytics and Mining is based on Weka)

- IBM SPSS (Statistical Package for Social Science)
- SAS
- STATISTICA
  - Dell acquired StatSoft in March 2014
  - TIBCO acquired STATISTICA in May 2017

Surveys of Data Science Platforms

- **Gartner’s 2020 Magic Quadrant of Data Science Vendors**
  - https://www.gartner.com/doc/3980855
  - https://www.gartner.com/reviews/market/data-science-machine-learning-platforms
  - **Data Science and Machine-Learning Platform**: "A cohesive software application that offers a mixture of basic building blocks essential both for crating many kinds of data science solution and incorporating such solutions into business processes, surrounding infrastructure and products.“

- **Rexer Analytics 2017 Data Science Survey**
  - http://www.rexeranalytics.com/data-science-survey.html
  - Eight surveys from 2007 to 2017 to examine behaviours, views and preferences of data analytic professionals.
  - Rexer Analytics 2020 Data Science Survey: the results are not yet available

Gartner’s Magic Quadrant for Data Science

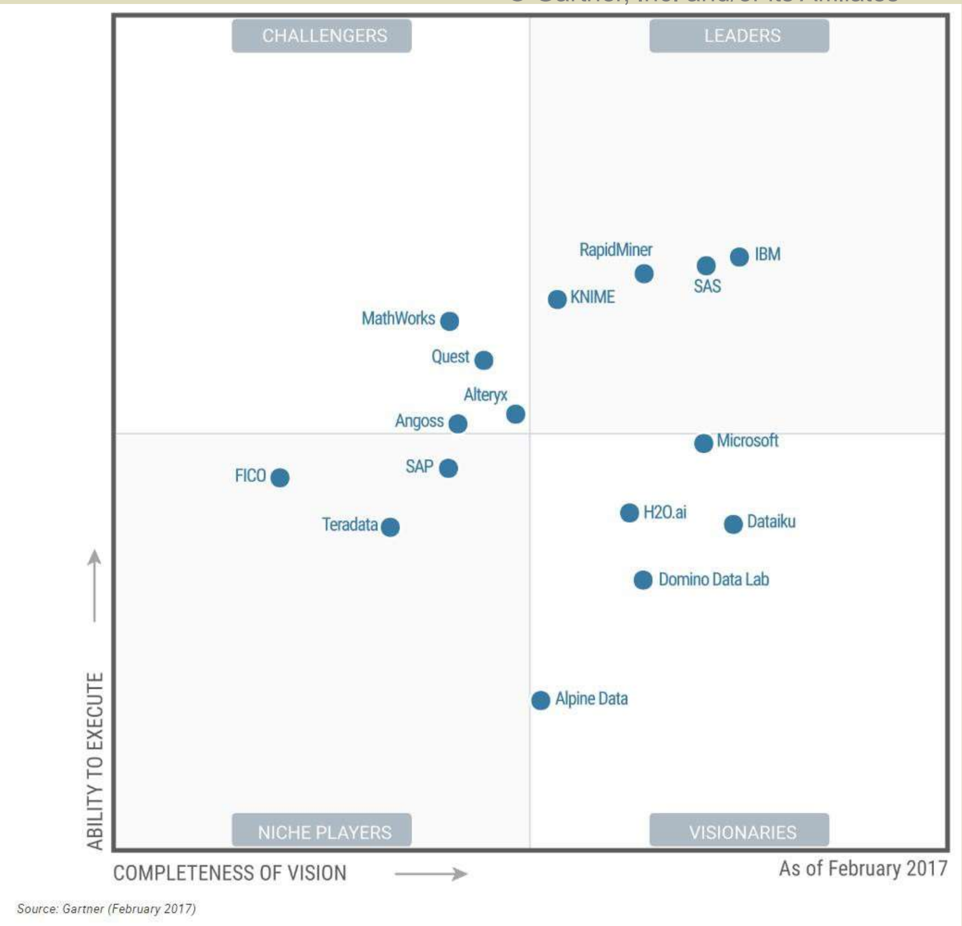
Data Science platforms provide an end-to-end environment for developing and deploying models, including:

- Data access to a variety of data sources.
- Data preparation, exploration and visualization is a key area of functionality
  - analysis is performed by users who may lack familiarity with the data and have increasingly high expectations of tools for automating data discovery, visualization and preparation.
- The ability to develop and build analytic models
  - including clustering, classification and predictive models, forecasting models, simulation models and optimization models.
- Ability to deploy models and integrate them into business processes and applications.
- Capabilities to perform platform, project and model management.
  - The need to be able to validate the performance of models and track them once deployed is necessary.
- High performance and scalability for both development and deployment.
  - The ability to perform at high levels of speed and accuracy with large volumes data and streaming data is still critical for organizations, and with rising data volumes becomes even more of a differentiator.

Gartner's Magic Quadrant for Data Science

2017

For many years SAS, IBM, KNIME and RapidMiner has been considered the Leaders in this market.



Source: Gartner (February 2017)

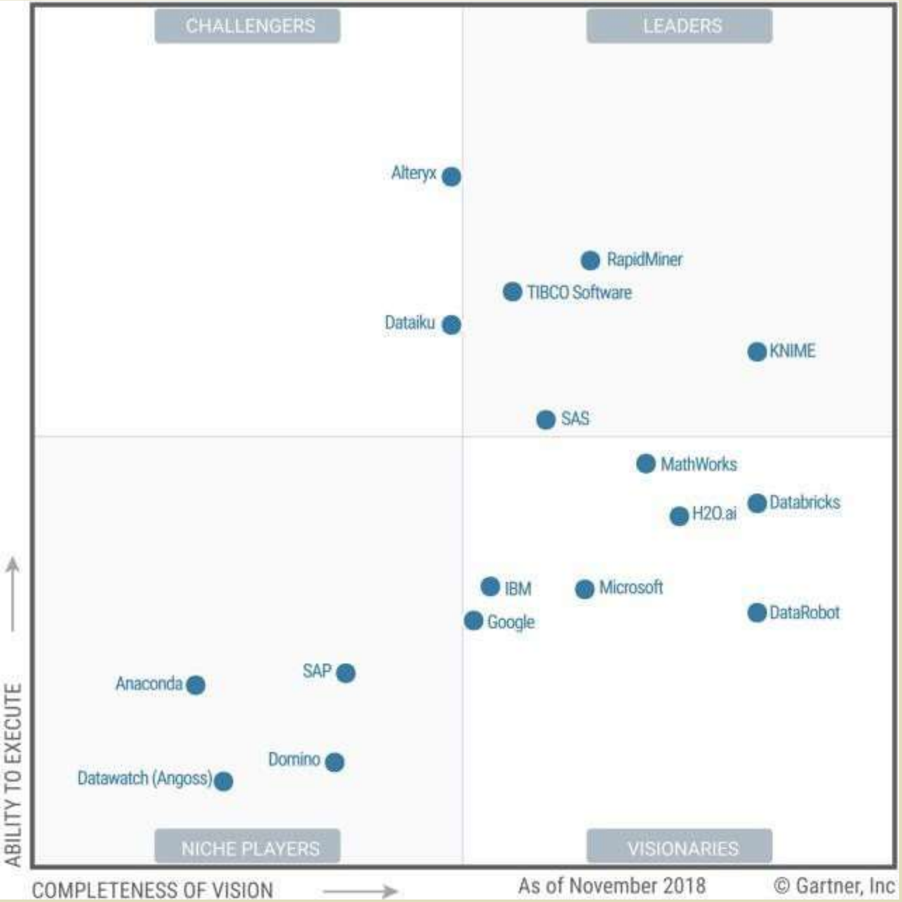
Gartner's Magic Quadrant for Data Science

Gartner's Magic Quadrant for **Data Science and Machine-Learning Platforms**

2019

For the sixth year in a row, Gartner has placed KNIME in the Leaders' quadrant.

© Gartner, Inc. and/or its Affiliates



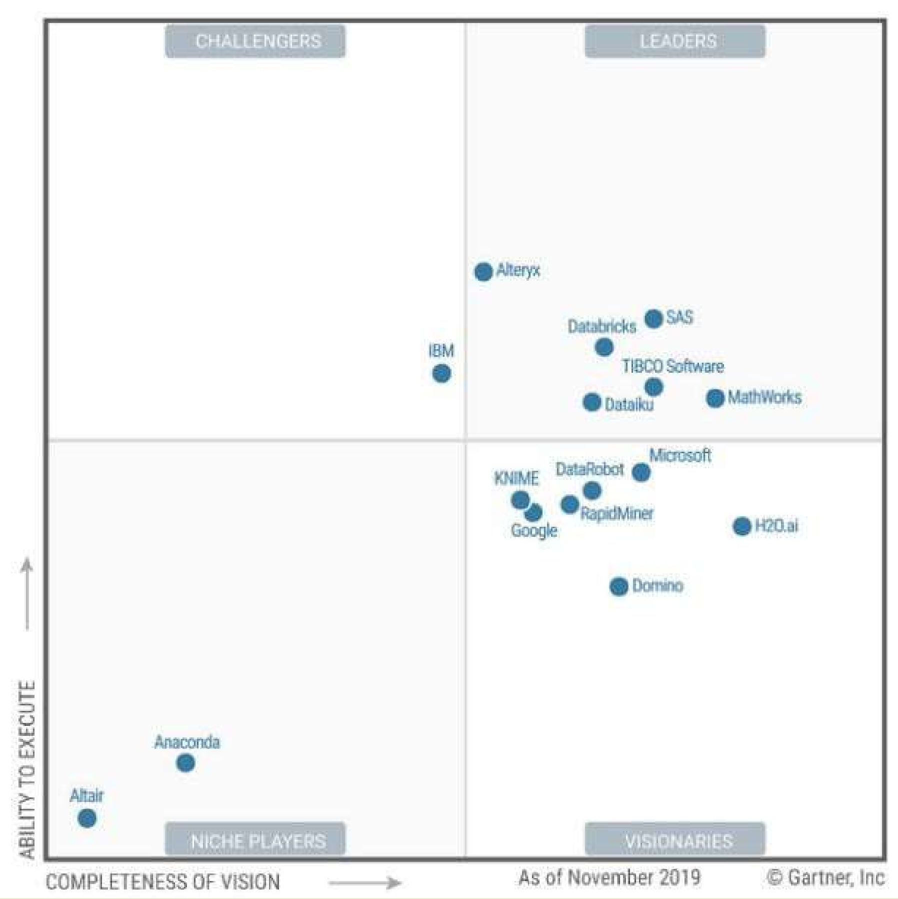
Gartner’s Magic Quadrant for Data Science

Gartner’s Magic Quadrant for **Data Science and Machine-Learning Platforms**

© Gartner, Inc. and/or its Affiliates

2020

“KNIME dropped from Leaders quad to Visionary mainly because of its lower visibility and slow revenue growth relative to other vendors.”



Rexer Analytics: the 2015 Data Science Survey

2015 Data Science Survey (39 pages) released in September 2016:

- **SURVEY & PARTICIPANTS**: 59-item online survey conducted in 2015 on 1,220 participants analytic professionals from 72 countries. The 7th in the series.
- **CORE ALGORITHM TRIAD**: Regression, Decision Trees, and Cluster analysis remain the most commonly used algorithms in the field.
- **THE ASCENDANCE OF R**: 76% of respondents report using R. This is up dramatically from just 23% in 2007. More than a third of respondents (36%) identify R as their primary tool.
- **JOB SATISFACTION**: Job satisfaction in the field remains high, but has slipped since the 2013 survey. A number of factors predict Data Scientist job satisfaction levels.
- **DEPLOYMENT**: Deployment continues to be a challenge for organizations, with less than two thirds of respondents indicating that their models are deployed most or all of the time. Getting organizational buy-in is the largest barrier to deployment, with real-time scoring and other technology issues also causing significant deployment problems.
- **TERMINOLOGY**: 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.

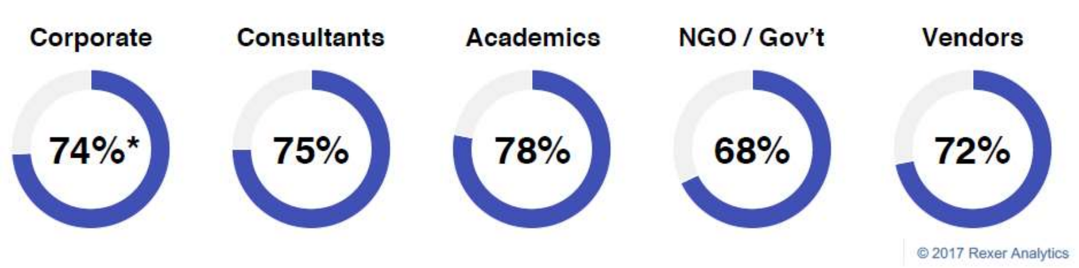
Rexer Analytics

Rexer Analytics: the 2017 Data Science Survey

- 8th survey since 2007, 67 questions, 1123 respondents (professionals from 91 countries)
- *(data collected in first half of 2017; highlights of the results in Oct. 2017)*

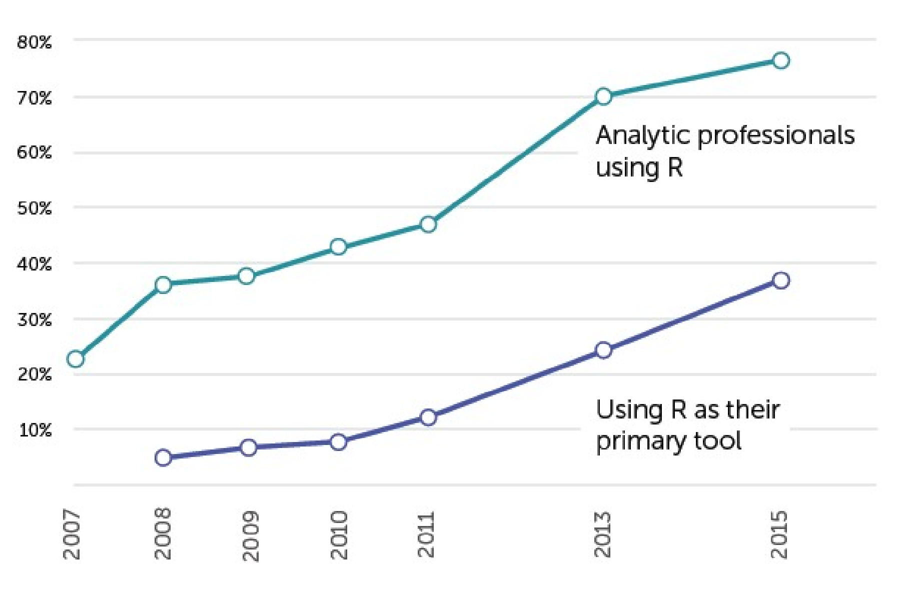
- The majority of respondents agree that formal data training is needed to properly model data
  - Do you agree or disagree with the following statement?

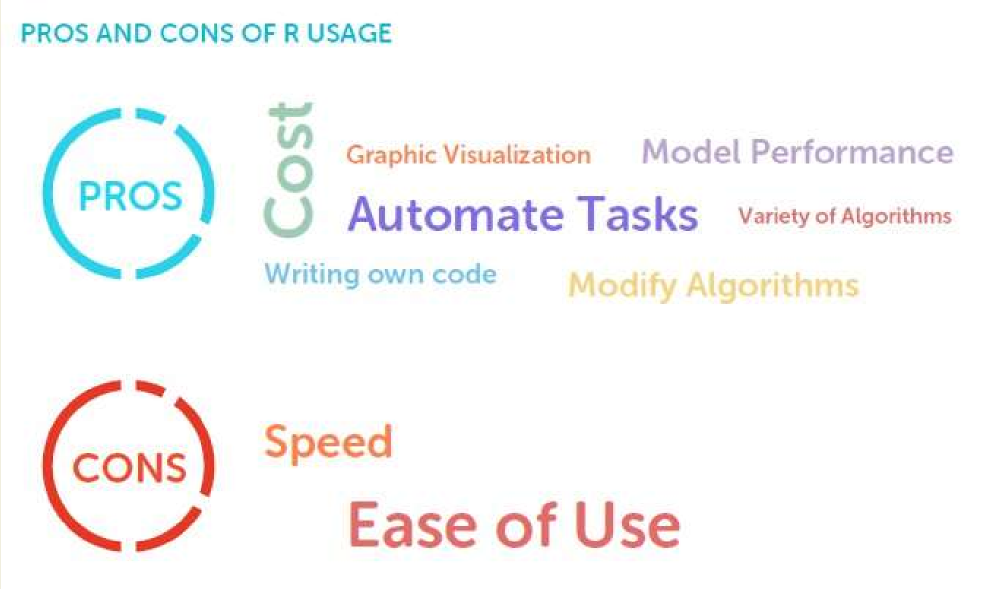
    You need to have formal training in data analytics in order to properly model data.



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

R Usage in the Rexer's surveys

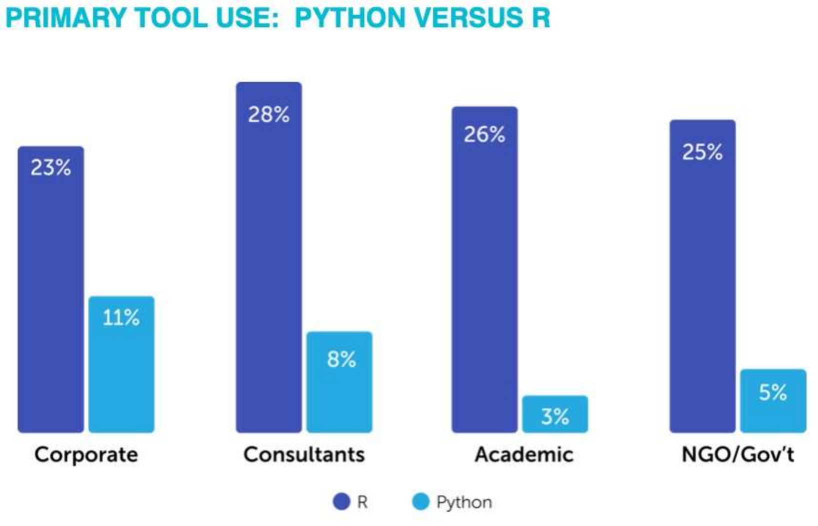




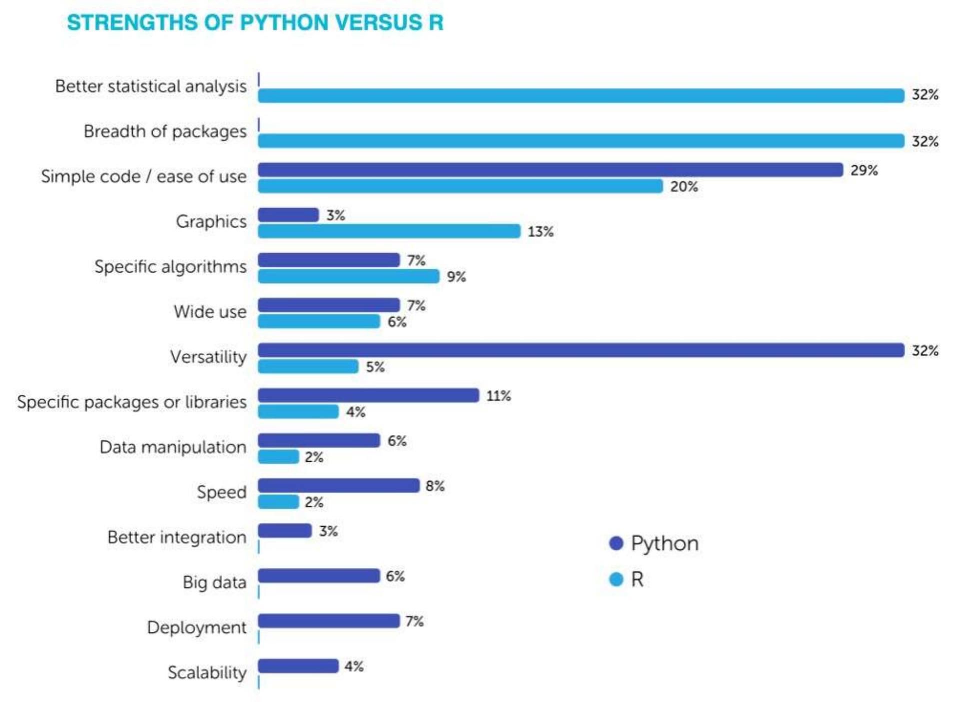
Rexer Analytics

Rexer Analytics: the 2017 Data Science Survey

Python vs R

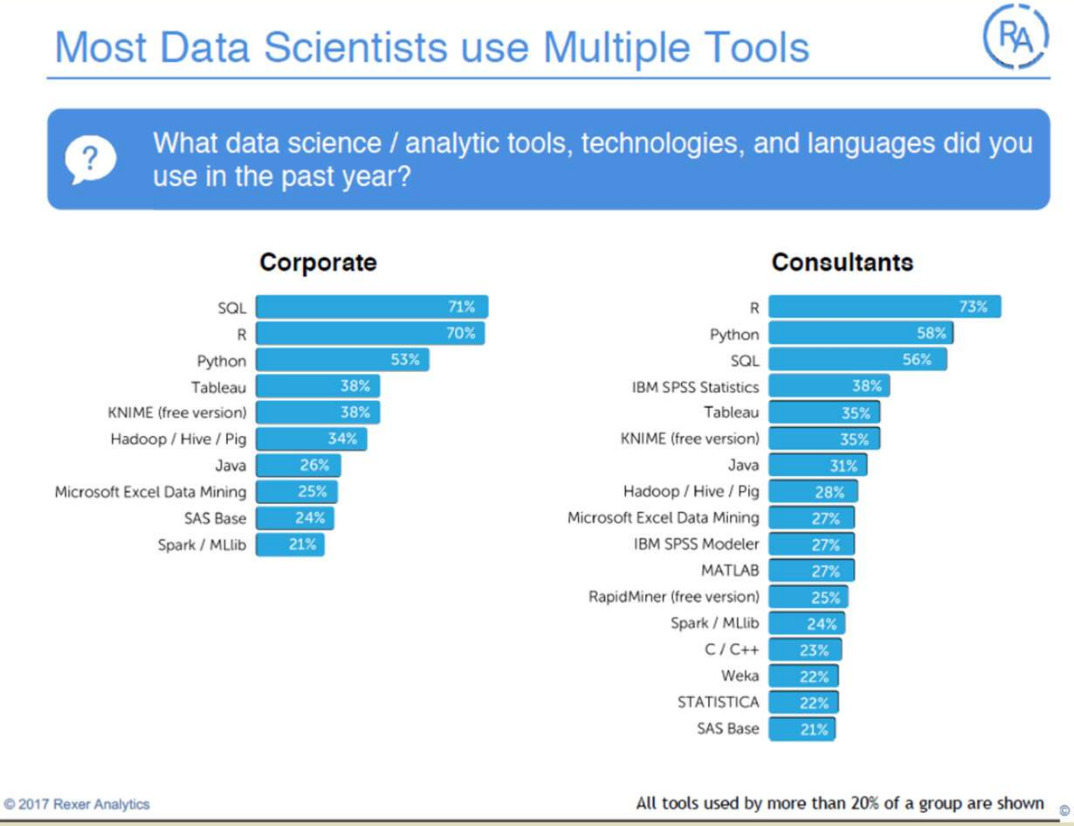


R is considered the primary tool by more respondents overall than Python.



Note: since 2017, Python is believed to have overtaken R in popularity. In 2020, a new survey has been carried out and the results are expected in the new few months.

Rexer Analytics: the 2017 Data Science Survey

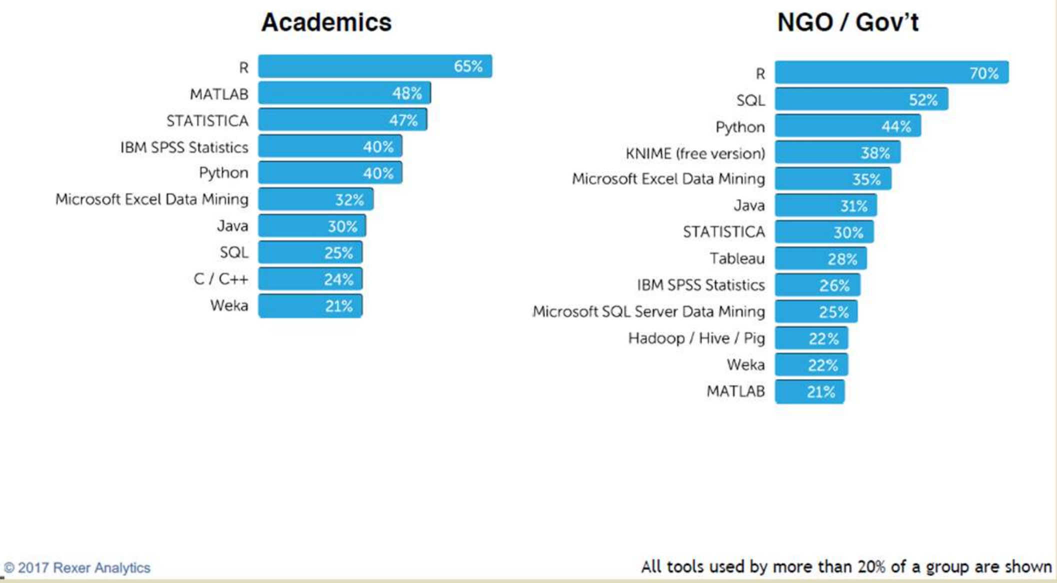


Most Data Scientists use Multiple Tools

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

© 2017 Rexer Analytics

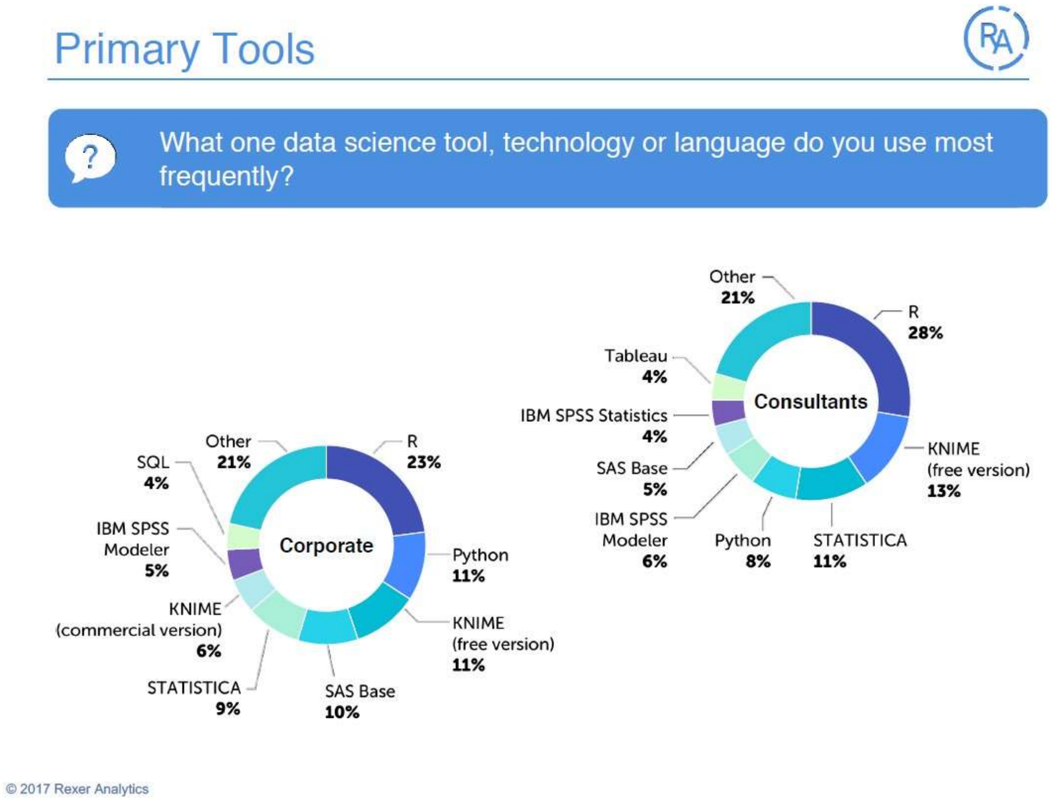
All tools used by more than 20% of a group are shown



© 2017 Rexer Analytics

All tools used by more than 20% of a group are shown

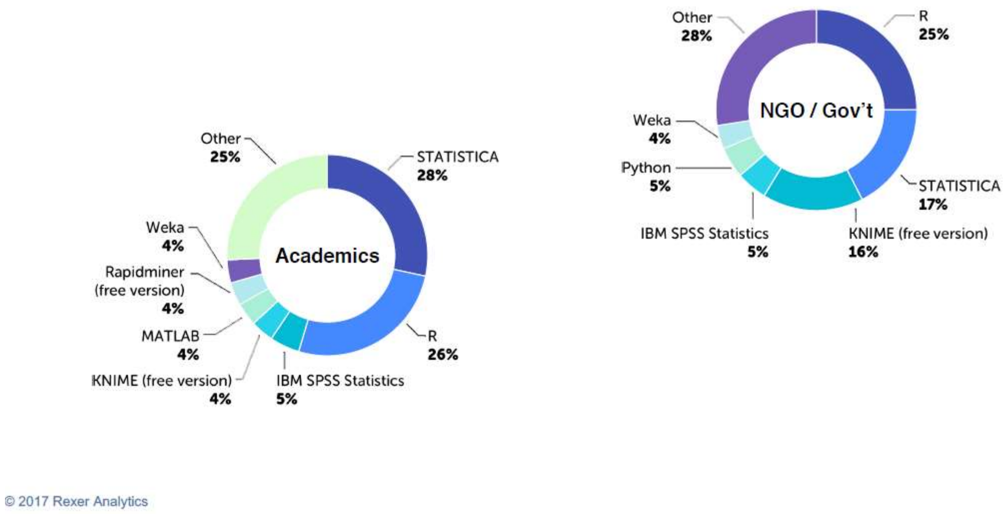
Rexer Analytics: the 2017 Data Science Survey



Primary Tools

What one data science tool, technology or language do you use most frequently?

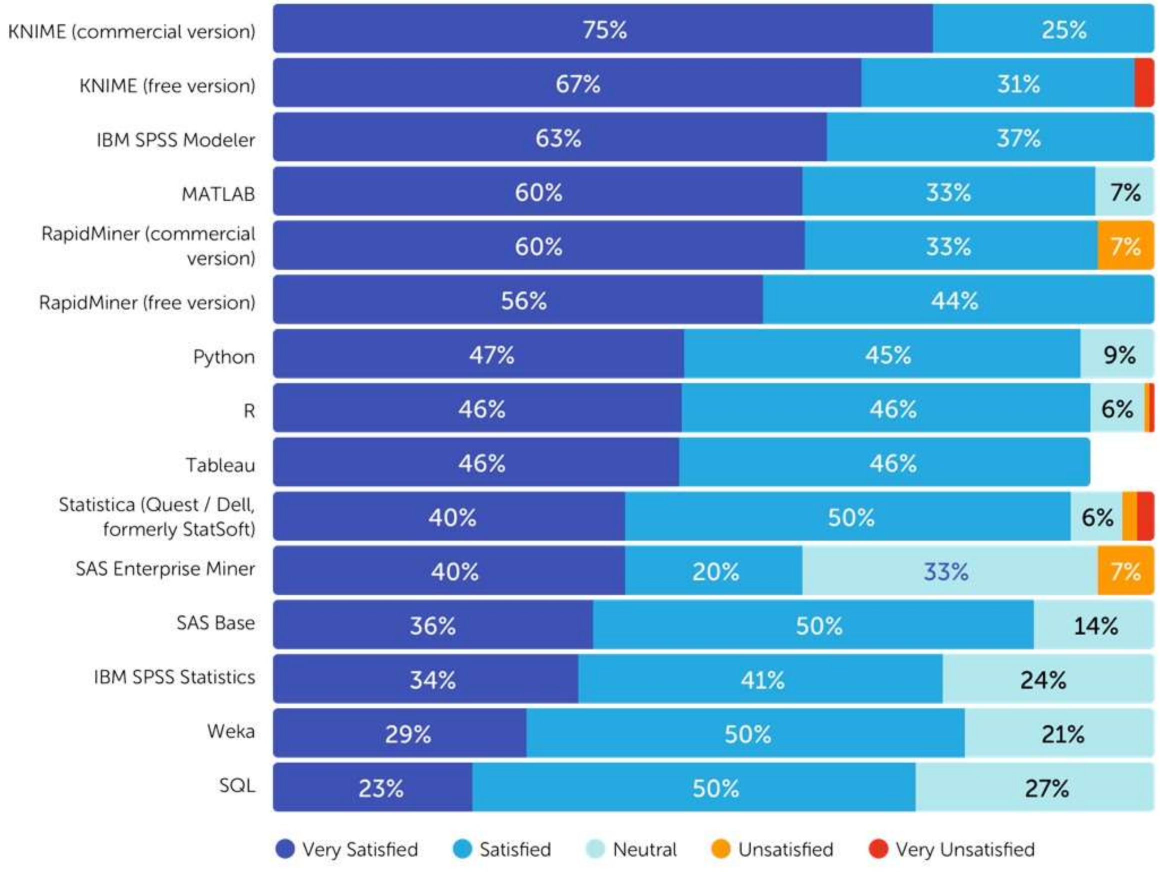
© 2017 Rexer Analytics



© 2017 Rexer Analytics

Rexer Analytics: the 2017 Data Science Survey

Overall satisfaction in the primary tool

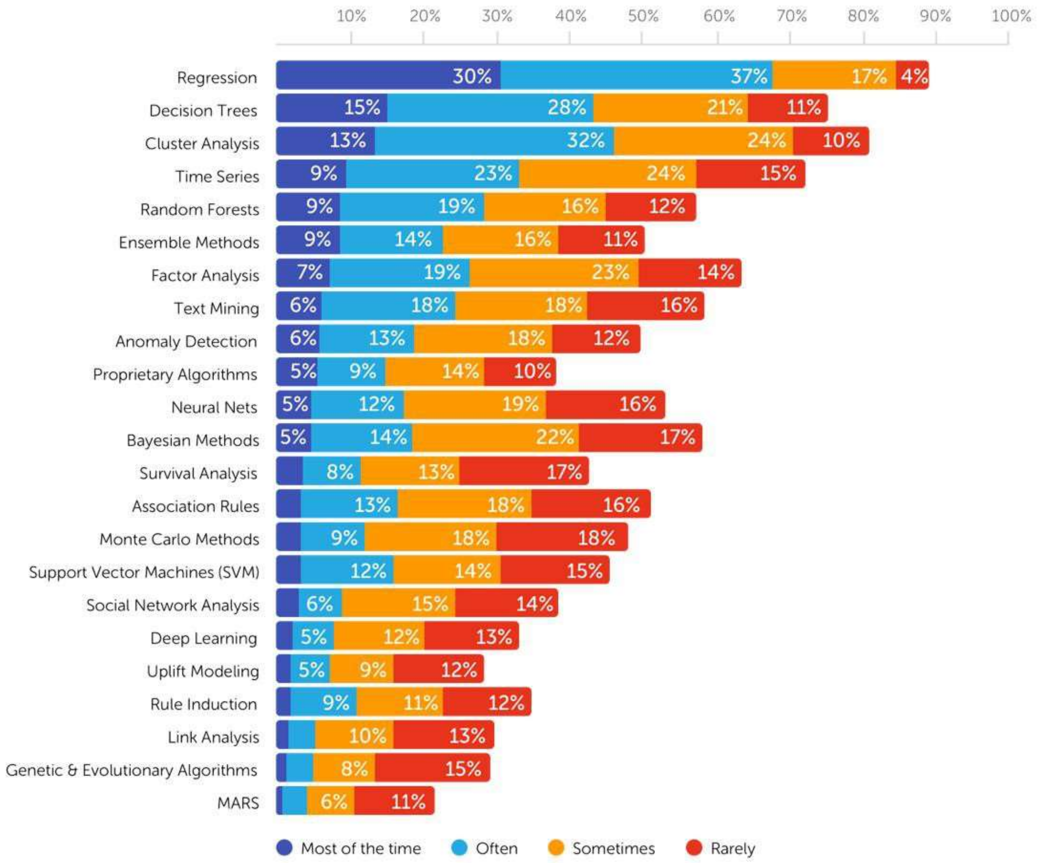


Rexer Analytics: the 2017 Data Science Survey

<u>CSMDM16:</u>

- ✓ Regression
- ✓ *Decision Trees*
- ✓ *Cluster Analysis*
- ▪ *Time Series*
- ✓ *Random Forests*
- ✓ *Ensemble Methods*
- ▪ *Factor Analysis*
- ▪ *Text Mining*
- ▪ *Anomaly Detection*
- ▪ *Neural Nets*
- ▪ *Bayesian Methods*
- ▪ *Survival Analysis*
- ✓ *Association Rules*
- ▪ *Monte Carlo Methods*
- ▪ *Support Vector Machines*
- ▪ *Social Network Analysis*
- ▪ *Deep Learning*
- ▪ *Uplift Modeling*
- ▪ *Rule Induction*
- ▪ *Link Analysis*
- ▪ *Genetic&Evolutionary Algorithms*
- ▪ *MARS*

What algorithms are most popular?



"Despite extensive media hype about AI, Cognitive Computing, Deep Learning, and the rise of machine learning and its related algorithms, no algorithms showed substantial increased usage since the 2015 survey."
© 2018 Rexer Analytics

Conclusions

❑Modern Data Science platforms

<u>must have:</u>

- data I/O, manipulation, visualization tools
- a workflow management system
- a comprehensive repository of algorithms (data mining, machine learning, statistics)
- fast prototyping and deployment capability

<u>should have:</u>

- computationally advanced and efficient execution environment (e.g., multi-threading, support for Cloud Computing)
- modularisation (nested workflows, reuse of general solutions to subtasks)
- user-friendly interface
- open environment (integration of external tools)
- scalability for Big Data

- <u>can be:</u>
  - free and open source

Next video lecture:

➢An Overview of Input Data