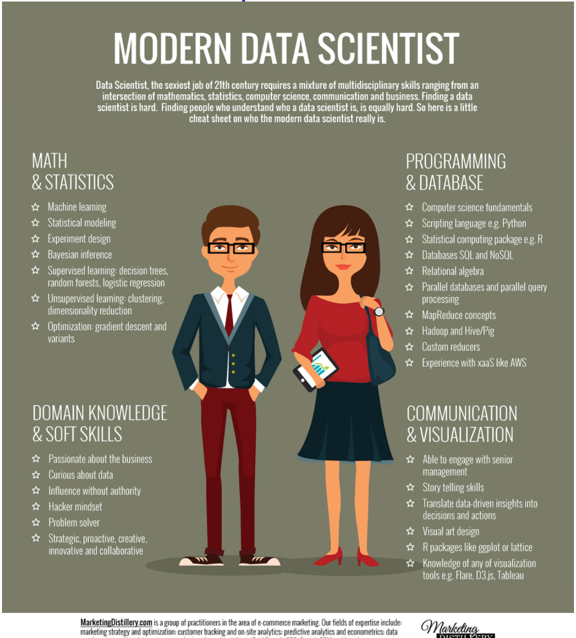
R for Data Science—Introduction

Components of Data Science



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data

What is R

- R is a statistical scripting language.
- You write code (a series of commands) to perform some task.
- R can be used to perform **all** of the tasks of a data analysis.

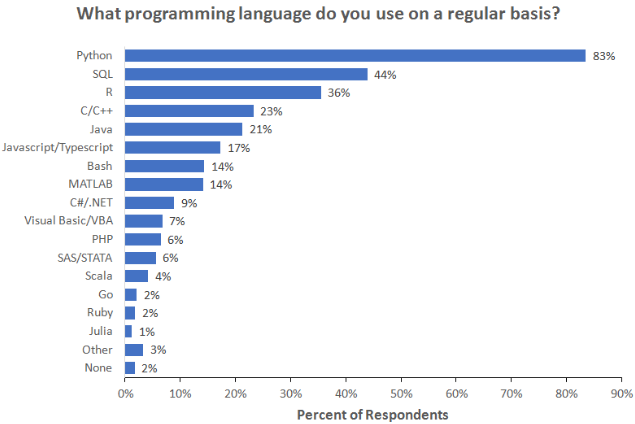
Motivation for R

- It's free.
  - You will always have access to R.
  - Not true for other statistical softwares (Matlab, STATA, SAS).
- It's widely used.
  - If you need to do some special analysis, someone has probably already made an R package for it.
- It's easy (especially graphics and data munging).
- It makes reproducible research easy.
  - When part of the pipeline is copying and pasting excel spreadsheets, people make mistakes.
  - E.g. an excel mistake led countries to adopt austerity measures to increase economic growth.
  - In R, you can automate your analysis, reducing the chance for mistakes and making your analysis transparent to the wider research community.

Could Also Learn Python or Matlab

- Python or Matlab are also very good tools for data science.
- Computer scientists tend to prefer Python because its syntax is more like a standard computer language and is fast.
- MatLab is easy for beginners who are just starting to learn about programming language because the package, when purchased, includes all that you will need.
- Python is excellent for Deep Learning methods.
- Main reason to use either tool is based on what your collaborators use.

Programming languages most used and recommended by data scientists


What programming language do you use on a regular basis?

Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Two main flavors of R

- There are two flavors of R programmers: Base R users and tidyverse users.
- Base R is more general (not fighting against the system when you want to accomplish a unique task that isn't designed to fit within the tidyverse).
- tidyverse is much more convenient for the vast majority of tasks.

Books and Resources:

- All material used in this course is free online.
- R for Data Science: https://r4ds.had.co.nz/
- Tidyverse Style Guide: https://style.tidyverse.org/
- Rstudio Cheat Sheets: https://www.rstudio.com/resources/cheatsheets/
- Hands-on Programming with R: https://rstudio-education.github.io/hopr/