Data Science Programme

General Information and Further Resources

Sebastian Krantz

18/02/2021

What is Data Science?

From Wikipedia

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.

What is Data Science?

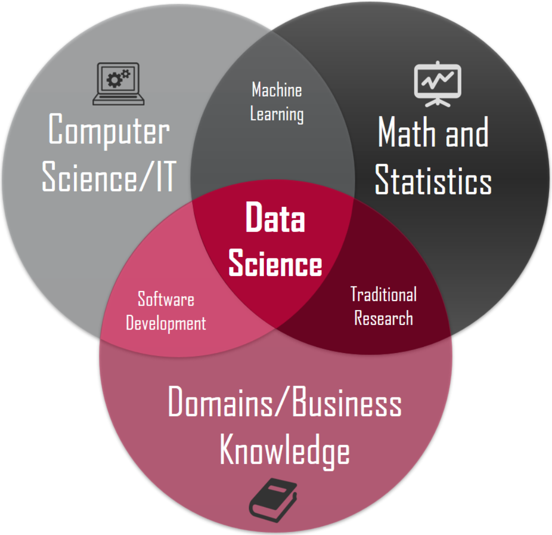


Figure 1: Domain/Business Knowledge = Economics and Finance for us.

Languages for Data Science: R

- **R** is an high-level interpreted functional, and object oriented programming language developed by statisticians for statistical computing and graphics, and supported by the R Foundation for Statistical Computing.
- R is a successor of S, created by John Chambers at Bell Labs in 1976. R was first released in 1993.
- Later many data manipulation, machine learning and graphical libraries were added by users and made easily available on the Comprehensive R Archive Network (CRAN), established in 2000. It currently (December 2020) features 16,801 R packages.
- Other packages for R can be found on Bioconductor and Github. In total there are currently around 21,000 packages.
- RStudio, Inc. is a company providing a like-named integrated development environment (IDE) for R and a popular and consistent set of R packages for data science.
- As of September 2020, R ranks 9th in the TIOBE index, it has more than 2 million active users and is growing at a fast pace.

Languages for Data Science: Python

- **Python** is an interpreted, high-level, multiple-paradigm, and general-purpose programming language. Python's design philosophy emphasizes object orientation and code readability.
- Python was first released in 1991 by Guido van Rossum as a successor to the ABC language, named after Monty Python.
- Python became popular for data science when libraries for scientific computing (SkiPy, 2001) graphics (Matplotlib, 2003), natural language processing (NLTK, 2001), arrays (NumPy, 2005), datasets (Pandas, 2008), machine learning (Scikit-learn, 2007) and deep learning (TensorFlow, 2015), (Keras, 2015), (Apache MXNet, 2015), (PyTorch, 2016) were developed.
- The Python Package Index (PyPI) is the official third-party software repository for Python. Currently $>$ 230,00 Python packages (modules, libraries) can be accessed through PyPI.
- Anaconda is a distribution of Python for data science.
- Spyder is an IDE for scientific programming in Python.
- Python ranks 3rd in the TIOBE index after C and Java, with $>$8 mio. developers and many more users, and growing rapidly.

Languages for Data Science: Julia

- Julia is a high-level, high-performance, dynamic programming language. While it is general-purpose, many of its features are well suited for numerical analysis and computational science.
- Work on Julia was started in 2009, by Jeff Bezanson, Stefan Karpinski, Viral B. Shah, and Alan Edelman, who set out to create a free language that was both high-level and fast.
- Since the 2012 launch, the Julia community has grown, and Julia is used at >10,000 companies with >20,000,000 downloads as of September 2020, up from 9 million a year prior (and is used at more than 1,500 universities).
- Julia has a growing body of currently 4700 packages, among others for econometrics, DSGE and computational models, time series and machine learning. The high performance makes it attractive for future development in many domains of data science, but it is still a very young language.
- See also Julia observer for packages and recent developments.
- Julia ranks 26th in the TIOBE index and is also growing rapidly.

Languages for Data Science: Comparison

Broadly speaking all 3 languages have their place, but the lines are blurring:

- R is best for data manipulation, statistical models / analysis, graphics, and interactive data products.
- Python is best for machine learning and production level programming and deployment (deep learning and data science at scale).
- Julia is best for high-performance and technical programming (numerical optimization / GE models, bootstrapping and other computationally intensive stuff).

Languages for Data Science: Comparison

Further Links

R or Python for Data Analysis

R, Python & Julia in Data Science: A comparison

R Vs Python: What's the Difference?

Overview of the Julia-Python-R Universe

julia for Data Science

Will Julia Replace Python and R for Data Science?

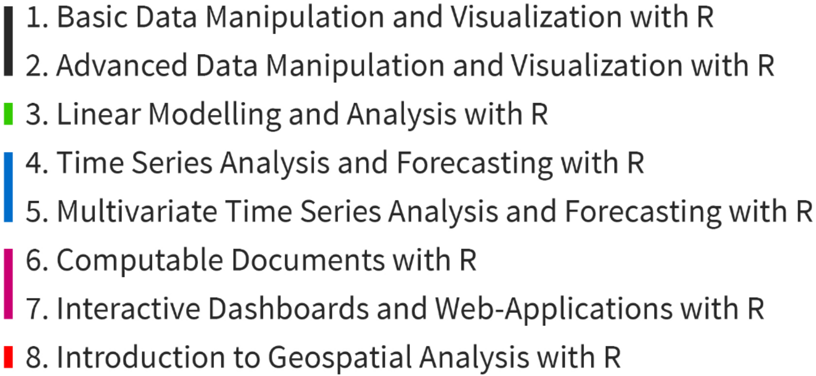
This Programme: Data Science in R

Focus on data manipulation, graphics, statistical models, time series, reproducible research, interactive outputs, and geospatial computing.

Core Aims: Participants should be able to

- Import, manage and maipulate data in R
- Produce publication quality graphics
- Write good code and clean analysis scripts allowing reproduction of results
- Estimate linear regression models and report results
- Analyze time series data and report results
- Generate interactive documents, presentations, dashboards and web applications
- Basic geospatial analysis and computing

The Programme Concrete



1. Basic Data Manipulation and Visualization with R
2. Advanced Data Manipulation and Visualization with R
3. Linear Modelling and Analysis with R
4. Time Series Analysis and Forecasting with R
5. Multivariate Time Series Analysis and Forecasting with R
6. Computable Documents with R
7. Interactive Dashboards and Web-Applications with R
8. Introduction to Geospatial Analysis with R

Figure 2: Courses of the same colour are consecutive. Course 1 is mandatory for all further courses.

Programme Organisation

- Courses are 2-days requiring full attendance.
- We start at 9am.
- To get most out of the course you need to participate in in-class exercises and homework assignments.
- Dates (and possibly alternative locations) for further courses will be communicated in advance. The plan is to have a course every 2-3 weeks so that we are done in the summer.

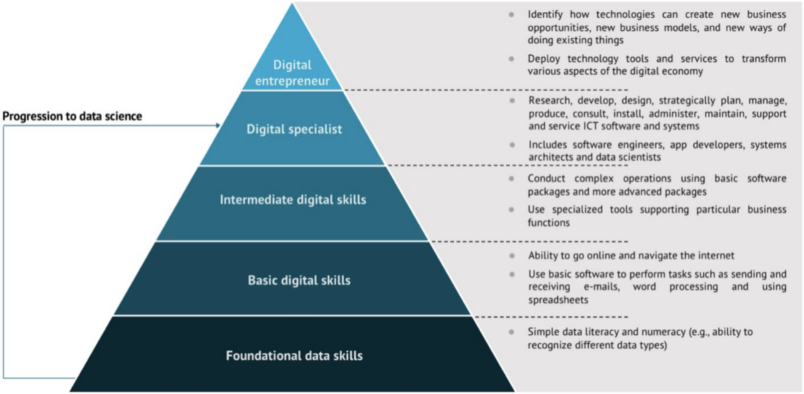
The course material can be accessed via:

- Google Drive: https://drive.google.com/drive/folders/1qz5hA-wkXVspta63O2___ZX5E6pkxrI4d?usp=sharing
- Github: https://github.com/SebKrantz/Data-Science-Programme

Further Resources for Data Science with R

- Swirl lets you learn R in R
- Cheat sheets for important packages
- Books for programming, statistics, applied statistics, econometrics, time series, statistical learning and geocomputation
- Websites for quick help and full university courses
- Online MOO Courses. The John's Hopkins University Data Science Specialization is a great general reference. More specialized courses are available on Coursera and Edx and several smaller providers (e.g. see here for a geocomputation course given by a friend on Blossom Academy in Ghana).
- See here for a good overview of the available online learning opportunities.

The Benefit of Learning Data Science



Progression to data science

Digital entrepreneur

- Identify how technologies can create new business opportunities, new business models, and new ways of doing existing things
- Deploy technology tools and services to transform various aspects of the digital economy

Digital specialist

- Research, develop, design, strategically plan, manage, produce, consult, install, administer, maintain, support and service ICT software and systems
- Includes software engineers, app developers, systems architects and data scientists

Intermediate digital skills

- Conduct complex operations using basic software packages and more advanced packages
- Use specialized tools supporting particular business functions

Basic digital skills

- Ability to go online and navigate the internet
- Use basic software to perform tasks such as sending and receiving e-mails, word processing and using spreadsheets

Foundational data skills

- Simple data literacy and numeracy (e.g., ability to recognize different data types)

Figure 3: The Pyramid of Digital Proficiency.