International Cartographic Association
Association Cartographique Internationale

Towards Spatial Data Science: Bridging the Gap between GIS, Cartography and Data Science

Jan Wilkening [a]

[a] *Esri Deutschland GmbH, j.wilkening@esri.de*

Keywords: Big Data, Data Science, Spatial Data Science

Abstract:

Data is regarded as the oil of the 21st century, and the concept of data science has received increasing attention in the last years. These trends are mainly caused by the rise of big data – data that is big in terms of volume, variety and velocity. Consequently, data scientists are required to make sense of these large datasets. Companies have problems acquiring talented people to solve data science problems. This is not surprising, as employers often expect skillsets that can hardly be found in one person: Not only does a data scientist need to have a solid background in machine learning, statistics and various programming languages, but often also in IT systems architecture, databases, complex mathematics. Above all, she should have a strong non-technical domain expertise in her field (see Figure 1).

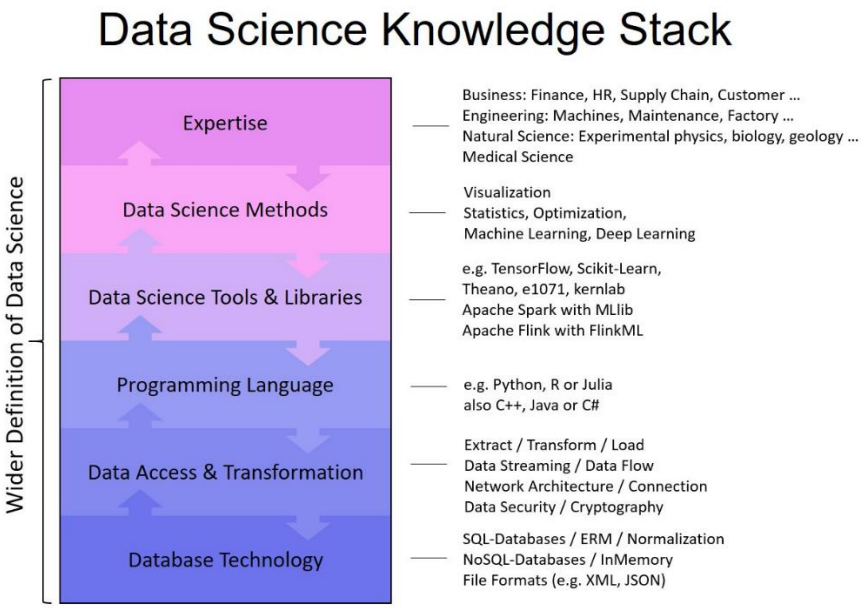

Figure 1. Example of a Data Science knowledge stack[1].

As it is widely accepted that 80% of data has a spatial component, developments in data science could provide exciting new opportunities for GIS and cartography: Cartographers are experts in spatial data visualization, and often also very skilled in statistics, data pre-processing and analysis in general. The cartographers' skill levels often depend on the degree to which cartography programs at universities focus on the "front end" (visualisation) of a spatial data and leave the "back end" (modelling, gathering, processing, analysis) to GIScientists. In many university curricula, these front-end and back-end distinctions between cartographers and GIScientists are not clearly defined, and the boundaries are somewhat blurred.

In order to become good data scientists, cartographers and GIScientists need to acquire certain additional skills that are often beyond their university curricula. These skills include programming, machine learning and data mining. These are important technologies for extracting knowledge big spatial data sets, and thereby the logical advancement to "traditional" geoprocessing, which focuses on "traditional" (small, structured, static) datasets such shapefiles or feature classes.

To bridge the gap between spatial sciences (such as GIS and cartography) and data science, we need an integrated framework of "spatial data science" (Figure 2).

[1] https://data-science-blog.com/blog/2017/09/16/data-science-knowledge-stack-abstraction-of-the-data-scientist-skillset/

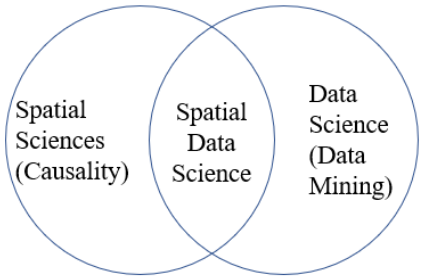

Figure 2. Spatial Data Science at the interface between Spatial Science and Data Science.

Spatial sciences focus on causality, theory-based approaches to explain why things are happening in space. In contrast, the scope of data science is to find similar patterns in big datasets with techniques of machine learning and data mining - often without considering spatial concepts (such as topology, spatial indexing, spatial autocorrelation, modifiable area unit problems, map projections and coordinate systems, uncertainty in measurement etc.).

Spatial data science could become the core competency of GIScientists and cartographers who are willing to integrate methods from the data science knowledge stack. Moreover, data scientists could enhance their work by integrating important spatial concepts and tools from GIS and cartography into data science workflows. A non-exhaustive knowledge stack for spatial data scientists, including typical tasks and tools, is given in Table 1.

Tasks	**Tool**
Preprocessing and automatisation	Java, Python
Big Data Management	Hadoop Ecosystem
Geoprocessing and mapping	(Desktop) GIS
Advanced (spatial) statistics and visualization	R
Database and multi-user management	DMBS / SQL
Distributing spatial data on the intra- and internet	Enterprise GIS / Web GIS

Table 1: Proposed knowledge stack for Spatial Data Scientists

There are many interesting ongoing projects at the interface of spatial and data science. Examples from the ArcGIS platform include:

- Integration of Python GIS APIs with Machine Learning libraries, such as scikit-learn or TensorFlow, in Jupyter Notebooks
- Combination of R (advanced statistics and visualization) and GIS (basic geoprocessing, mapping) in ModelBuilder and other automatization frameworks
- Enterprise GIS solutions for distributed geoprocessing operations on big, real-time vector and raster datasets
- Dashboards for visualizing real-time sensor data and integrating it with other data sources
- Applications for interactive data exploration
- GIS tools for Machine Learning tasks for prediction, clustering and classification of spatial data
- GIS Integration for Hadoop

While the discussion about proprietary (ArcGIS) vs. open-source (QGIS) software is beyond the scope of this article, it has to be stated that a.) many ArcGIS projects are actually open-source and b.) using a complete GIS platform instead of several open-source pieces has several advantages, particularly in efficiency, maintenance and support (see Wilkening et al. (2019)[2] for a more detailed consideration). At any rate, cartography and GIS tools are the essential technology blocks for solving the (80% spatial) data science problems of the future.

[2] Wilkening, J., Kapaj, A. and Cron, J. (2019): Creating a 3D Campus Routing Information System with ArcGIS Indoors. In: Dreiländertagung der DGPF, der OVG und der SGPF in Wien, Österreich – Publikationen der DGPF, 28, 2019. Available at https://www.dgpf.de/src/tagung/jt2019/proceedings/proceedings/papers/11_3LT2019_Wilkening_et_al.pdf (accessed Apr 9, 2019).