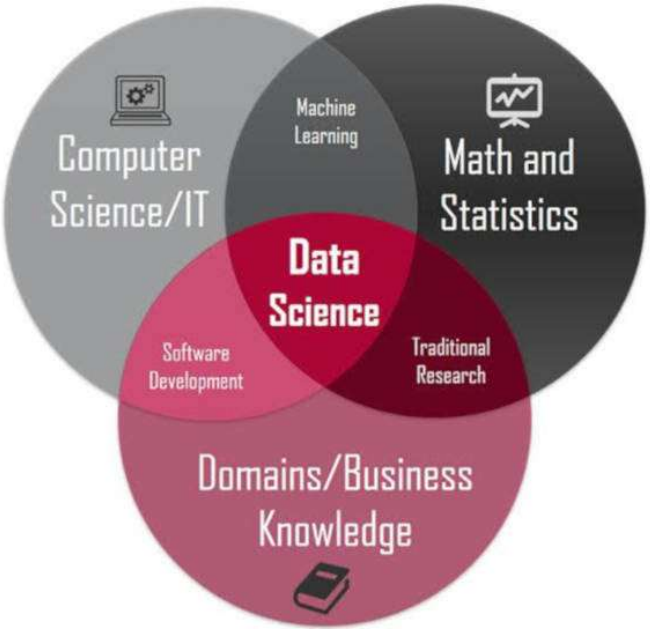
August 26: Introduction to Spatial Data Science

Saturday, August 14, 2021 3:25 PM

Data Science arose in late 1990s and has been growing exponentially since. As a multidisciplinary field, what Data Science is remains debatable. However, most scholars agree the three dimensions that constitute Data Science as shown in the figure below (source: https://www.kdnuggets.com/2020/08/top-10-lists-data-science.html).



To a large degree, the popularity of Data Science is due to its appeals to a wide diversity of government agencies, businesses, and organizations. Some said "data is the new oil." Raw oil remains useless until we refine it to diesel or gas. Likewise, data science aims to extract values out of raw data. On the other hand, all disciplines in science and engineering use data to find information or knowledge useful for questions or tasks on hand, so should all disciplines fall under Data Science? Answer to the question resides in where data come into play in the process of knowledge production; that is, epistemology.

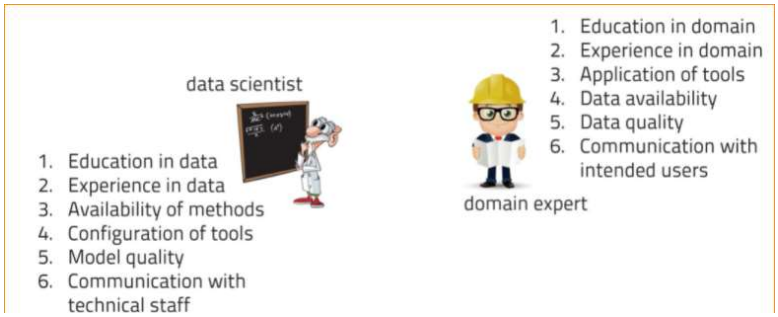
Science aims to get us to know what we don't know. In science, data are facts. Observational, experimental, empirical, and computational sciences apply theoretical frameworks to collect data and then use data to test their propositions or hypotheses. In contrast, data science does not apply a systematic scheme for observations, surveys, or sampling, but use massive amounts of data to find and explain hidden patterns.

Go back to the Venn diagram above. Think about how each dimension contributes to data science, and then think about how spatial or geographic concepts are needed in each dimension for Spatial Data Science (or Geographic Data Science). We will use Spatial Data Science and Geographic Data Science interchangeably in this class. However, spatial differs from geographic, and both differ from geospatial.

So, what competency do we need to do Spatial Data Science (SDS)?

1. Most importantly: ask spatial questions
2. The SDS knowledge production process: objectives and tasks in a SDS workflow
3. Technical skills to implement SDS objectives and tasks
4. Domain knowledge to interpret the SDS findings

Domain knowledge is the most difficult but often neglected in SDS. Many DS and SDS projects are heavy on computing, statistics or mathematics but shallow on domain knowledge with trivial findings. We need to have at least a basic understanding of the domain for our project. If we are working on crime mapping, we need to have basic knowledge in criminology. If we are working on a marketing project, we need to have basic knowledge on marketing and the products we attempt to market. If you are doing a research project for a degree (capstone project, Master's thesis, or doctoral dissertation), you need to master both data science and domain knowledge. If you work for government agencies, organizations, or companies, you need to keep constant communications with domain experts:



4. Configuration of tools
5. Model quality
6. Communication with technical staff

Below are some examples in which data lead to wrong suggestions:

1. Abraham is tasked with reviewing damaged planes coming back from sorties over Germany in the Second World War. He has to review the damage of the planes to see which areas must be protected even more. Abraham finds that the fuel system of returned planes are much more likely to be damaged by bullets than the engines. His data science report suggests the fuel systems receive additional protection. What could be wrong with the suggestion? This is an example of data biases.
2. On July 21, 2021, Iceland reported that 82% of their new COVID-19 cases were fully vaccinated, which someone suggested that the COVID-19 vaccines were ineffective. What could be wrong with the suggestion? This is an example of ignoring the context.

Keep in mind four key elements for good data:

a. Precision: how much uncertainty is in the value
b. Accuracy: how much deviation from reality in the data
c. Representativeness: how much the data reflect all relevant aspects of the domain
d. Significance: how much the data sufficiently reflect every important relationship, behavior or dynamic in the domain