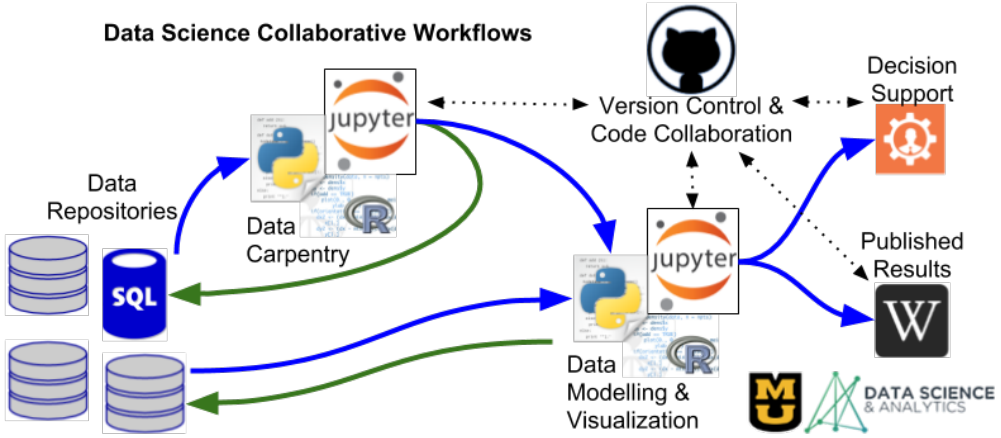
What is Data Science?

Data science is an interdisciplinary field involving key elements from computer science, mathematics/statistics, business analytics, and the sciences. Tools, methods, and techniques from all of these fields are applied within a rigorous scientific approach to document, organize, integrate, manipulate, understand, and exploit data from a wide variety of sources to produce information and analytic products that educate, inform, and support decision making processes across many levels within an organization.

The key word in data science is science and not data! Too often the discussion about data science is erroneously focused on the data (volume, variety, veracity, etc.) and/or tools (Python, R, machine learning, cloud computing, etc.) used to process, analyze, and exploit data. Just because you may have TBytes of data and cloud-computing resources to process that data does not ensure you will be able answer real questions that address key issues and/or knowledge gaps within an organization.

Thus, data science should start with questions (not data!) and then apply the appropriate tools, methods, and techniques from the fields noted above to relevant data sets that can be used to address an organization's key questions. This should be done within a well-documented and repeatable scientific approach that often involves collaboration between a group of individuals with domain-specific expertise from a variety of fields. The scientific process should produce analytic results and information that can be disseminated in a variety of ways to answer an organization's key questions.

The process should, whenever possible, document key assumptions, limitations, and uncertainties in the approach in addition to quantifying the estimated accuracy of the final results. This approach not only informs the decision-making process but also allows for others to assess veracity, repeat the process for validation, and/or to further develop/refine the approach to improve upon it. The benefits of data science to any organization will be maximized through the application of these scientific methods and principles to answer important questions using a wide variety of relevant data sources.



What is a Data Scientist?

A data scientist is an individual that applies scientific methods and practices throughout the entire lifecycle of data from source to decision support to address key questions within an organization. A data scientist must possess an interdisciplinary set of skills to effectively engage, manage, and apply appropriate scientific practices throughout the entire data lifecycle.

A data scientist helps organizational stakeholders formulate and/or refine key questions they seek to answer with relevant data. Then, the first stage of the data lifecycle involves the acquisition, inventory, cleaning, and normalization of relevant data sets using an organized and well-documented methodology to ensure data provenance, repeatability, and extensibility.

Relevant data sets often come from a variety of disparate sources and, as a result, vary widely in size, structure, type, and quality. Thus, a data scientist often must apply a broad spectrum of expertise to deal with these common Big Data challenges during the first stage of the data lifecycle. Also, during this stage, a successful data scientist will often collaborate with data architects and data managers to create standard data integration interfaces for initial exploration, visualization, and modeling of the selected data sets.

Next, a data scientist will engage in exploratory data analysis of the collected, ingested, and well-prepared data along with other existing data repositories when needed. During the exploratory analysis, a data scientist evaluates the suitability of the collected and prepared data to provide decision support for an organization's stakeholders and properly answer the key questions. Throughout this process the data scientist should collaborate with domain experts to ensure data integrity and with downstream analysts to ensure that assumptions and uncertainties regarding the data and processes are properly communicated. Results from the exploratory data analysis may lead to additional data collection/preparation as well as refinement of the scientific methods and approach that will be utilized on a given project.

After this, a data scientist applies the scientific method to analyze the relevant and well-conditioned data sets to answer an organization's key questions. In this stage, the data scientist should be capable of utilizing, as needed for a given project, a variety of tools from various domains, such as statistical analysis/modeling, data mining, computer vision, machine learning, and high-performance computing. Key assumptions, procedures, analysis methods, limitations, and uncertainties should be well documented within a collaborative environment to provide a provenance for the analytic results and products. This provenance will help inform the decision-making process while also allowing for others to assess veracity, repeat the process for validation, and/or to further develop/refine the scientific approach to improve upon it.

Finally, an accomplished data scientist should employ a broad spectrum of visualization and information delivery methods to deliver analytic results and products that educate, inform, and support an organization's decision-making process regarding the key questions.