**A “Small” Definition of Big Data**

The term ‘big data’ seems to be popping up everywhere these days. And there seems to be as many uses of this term as there are contexts in which you find it: ‘big data’ is often used to refer to any dataset that is difficult to manage using traditional database systems; it is also used as a catch-all term for any collection of data that is too large to process on a single server; yet others use the term to simply mean “a lot of data”; sometimes it turns out it doesn’t even have to be large. So what exactly is big data?

A precise specification of ‘big’ is elusive. What is considered big for one organization may be small for another. What is large-scale today will likely seem small-scale in the near future; petabyte is the new terabyte. Thus, size alone cannot specify big data. The complexity of the data is an important factor that must also be considered.

Most now agree with the characterization of big data using the 3 V’s coined by Doug Laney of Gartner:

· Volume: This refers to the vast amounts of data that is generated every second/minute/hour/day in our digitized world.

· Velocity: This refers to the speed at which data is being generated and the pace at which data moves from one point to the next.

· Variety: This refers to the ever-increasing different forms that data can come in, e.g., text, images, voice, geospatial.

A fourth V is now also sometimes added:

· Veracity: This refers to the quality of the data, which can vary greatly.

There are many other V's that gets added to these depending on the context. For our specialization, we will add:

· Valence: This refers to how big data can bond with each other, forming connections between otherwise disparate datasets.

The above V’s are the dimensions that characterize big data, and also embody its challenges: We have huge amounts of data, in different formats and varying quality, that must be processed quickly.

It is important to note that the goal of processing big data is to gain insight to support decision-making. It is not sufficient to just be able to capture and store the data. The point of collecting and processing volumes of complex data is to understand trends, uncover hidden patterns, detect anomalies, etc. so that you have a better understanding of the problem being analyzed and can make more informed, data-driven decisions. In fact, many consider value as the sixth V of big data:

· Value: Processing big data must bring about value from insights gained.

To address the challenges of big data, innovative technologies are needed. Parallel, distributed computing paradigms, scalable machine learning algorithms, and real-time querying are key to analysis of big data. Distributed file systems, computing clusters, cloud computing, and data stores supporting data variety and agility are also necessary to provide the infrastructure for processing of big data. [Workflows](http://words.sdsc.edu/words-data-science/workflows) provide an intuitive, reusable, scalable and reproducible way to process big data to gain verifiable value from it in and enable application of same methods to different datasets.

With all the data generated from social media, smart sensors, satellites, surveillance cameras, the Internet, and countless other devices, big data is all around us. The endeavor to make sense out of that data brings about exciting opportunities indeed!

Source: <http://words.sdsc.edu/words-data-science/big-data>

Data Science is about extracting knowledge from data. At the WorDS Center (words.sdsc.edu), we define data science as a multidisciplinary craft that combines people, process, computational and Big Data platforms, application-specific purpose and programmability. Publications and provenance of the data products leading to these publications are also important for data science, but we start by defining 5 P's that take significant part in the data science activities.

* **Purpose:**The purpose refers to the challenge or set of challenges defined by your big data strategy. The purpose can be related to a scientific analysis with a hypothesis or a business metric that needs to be analyzed based often on Big Data.
* **People**: The data scientists are often seen as people who possess skills on a variety of topics including: science or business domain knowledge; analysis using statistics, machine learning and mathematical knowledge; data management, programming and computing. In practice, this is generally a group of researchers comprised of people with complementary skills.
* **Process**: Since there is a predefined team with a purpose, a great place for this team to start with is a process they could iterate on. We can simply say, People with Purpose will define a Process to collaborate and communicate around! The process of data science includes techniques for statistics, machine learning, programming, computing and data management. A process is conceptual in the beginning and defines the course set of steps and how everyone can contribute to it. Note that similar reusable processes can be applicable to many applications with different purposes when employed within different workflows. Data science workflows combine such steps in executable graphs. We believe that process-oriented thinking is a transformative way of conducting data science to connect people and techniques to applications. Execution of such a data science process requires access to many datasets, Big and small, bringing new opportunities and challenges to Data Science. There are many Data Science steps or tasks, such as Data Collection, Data Cleaning, Data Processing/Analysis, Result Visualization, resulting in a Data Science Workflow. Data Science Processes may need user interaction and other manual operations, or be fully automated.Challenges for the data science process include 1) how to easily integrate all needed tasks to build such a process; 2) how to find the best computing resources and efficiently schedule process executions to the resources based on process definition, parameter settings, and user preferences.
* **Platforms**: Based on the needs of an application-driven purpose and the amount of data and computing required to perform this application, different computing and data platforms can be used as a part of the data science process. This scalability should be made part of any data science solution architecture.
* **Programmability**: Capturing a scalable data science process requires aid from programming languages, e.g., R, and patterns, e.g., MapReduce. Tools that provide access to such programming techniques are key to making the data science process programmable on a variety of platforms.

To summarize, data science can be defined as a craft of using the five pieces identified above. Having a process between the more business driven P’s people and purpose and the more technical driven P’s platforms and programmability leads to a streamlined approach that starts and ends with a defined business value, team accountability and collaboration in mind.

Source: <http://words.sdsc.edu/words-data-science/data-science>