**SDC #2**

**Big Data Terminology**

**Data curation** is the management of **data** throughout its lifecycle, from creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted. The main purpose of **data curation** is to ensure that **data** is reliably retrievable for future research purposes or reuse.

**Knowledge Discovery and Data Mining** (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

**Data** **scrubbing**, also called **data cleansing**, is the process of amending or removing **data** in a database that is incorrect, incomplete, improperly formatted, or duplicated.

**Data pre-processing** includes cleaning, normalization, transformation, feature extraction and selection.

**Data Models** define how **data** is connected to each other and how they are processed and stored inside the system. The very first **data models** could be flat **data**-**models**, where all the **data** used are to be kept in the same plane.