Abstract Title	The Analytics Pipeline: Data Acquisition in the Information Era
Presenting Author	Hao Zhong
Institution/Affiliation	Rensselaer Polytechnic Institute
E-mail	zhongh3@rpi.edu
Additional Authors	Huang F, Prabhu A, Pan F, Eleish A, Ma X, Fox PA, and the Keck Deep-Time Data Infrastructure Team
Abstract Text	Over the past century, enormous amount of data has been produced, archived and published across the geoscience community. Development of new experimental devices, analytical tools, as well as scientific methods have been the driving forces underneath the accelerated increase in quantity and improvement in quality of geoscience-related data. In recent decades, such exponential growth of data has uncovered new, data-intensive approaches towards research questions that were once before unsolvable in the absence of enough data, and even inspired many new discoveries. As data science becomes instrumental in geoscience research, understanding and adoption of suitable data science practices cannot be emphasized enough in order to maximize the utilization of existing data and unleash the full potential of data-driven discoveries in geoscience. Data acquisition represents a starting phase in the data science pipeline where researchers acquire data from its sources. While traditional modes of data acquisition include observation, measurement, and generation, in the information era, data acquisition also includes data rescue and data access. Data rescue generally refers to digitization and curation of “dark data”, i.e. data that exists, often from long time ago and in analogue forms, but are not readily available, easily accessible, and directly usable. Data scientists at Tetherless World Constellation have been actively engaged in and successfully accomplished data rescue tasks, e.g. thermodynamic data rescue and diamond data rescue, in collaboration with geoscientists. Data access, on the other hand, is directly accessing and taking full advantage of data from already well-curated and high-quality geoscience databases. Members of the Keck Deep-Time Data Infrastructure Team have developed a number of applications to facilitate data access from geoscience data portals such as the RRUFF IMA Database of Mineral Properties and Mineral Evolution Database, and the Paleobiology Database, and in turn provided high-quality and comprehensive datasets for several successful deep-time data-driven discoveries.