For more than two thousand years of human history, science was empirical. Several hundred years ago

theoretical methods based on models and generalization were introduced, allowing substantial progress

in human knowledge. In the last few decades, we have witnessed the explosion of computational science

based on the simulation of complex phenomena.

In a talk delivered in 2007 and posted on his Web site just before he went missing in January

2007, computer scientist Jim Gray discussed eScience as a transformative scientific method [163].

Today, eScience unifies experiment, theory, and simulation; data captured from measuring instruments

or generated by simulations are processed by software systems, and data and knowledge are stored by

computer systems and analyzed using statistical packages.

The generic problems in virtually all areas of science are:

• Collecting experimental data.

• Managing very large volumes of data.

• Building and executing models.

• Integrating data and literature.

• Documenting experiments.

• Sharing the data with others; data preservation for long periods of time.

All these activities require powerful computing systems.

A typical example of a problem faced by agencies and research groups is data discovery in large

scientific data sets. Examples of such large collections are the biomedical and genomic data at NCBI,8

the astrophysics data at NASA,9 or the atmospheric data at NOAA10 and NCAR.11

The process of online data discovery can be viewed as an ensemble of several phases [282]: (i)

recognition of the information problem; (ii) generation of search queries using one or more search

engines; (iii) evaluation of the search results; (iv) evaluation of theWeb documents; and (v) comparison

of information from different sources. The Web search technology allows scientists to discover text

documents related to such data, but the binary encoding of many of the documents poses serious

challenges.

Metadata is used to describe digital data and provides an invaluable aid for discovering useful

information in a scientific data set. A recent paper [282] describes a system for data discovery that

supports automated fine-grained metadata extraction and summarization schemes for browsing large

data sets and is extensible to different scientific domains. The system, called Glean, is designed to run

on a computer cluster or on a cloud; its run-time system supports two computational models, one based

on MapReduce and the other on graph-based orchestration.