### 1. Data Modeling and Data Curation (C. Baru)

Developing abstractions of data. Basics of conceptual, logical, and physical data models. Concepts, techniques, and approaches for managing and sharing data. Making data accessible for reuse, re-analysis, long-term preservation. Tracking transformations and changes made to data. Use of these concepts in different application domains, ranging from business applications, Internet-scale Web applications, to scientific computing. Describes data models such as relational, XML, self-describing files; metadata standards; query languages; etc. Describes the curation process, data publication, etc.

### 2. Large-Scale Graph Data Management (A. Gupta)

Learn different application areas that need graph management, data and processing models, storage and indexing techniques for graphs, query processing mechanisms, large graph problems and approaches, hardware-architectures to facilitate graph processing and queries, different tools and engines for graph management, practical projects with these tools in different areas of science including CS.

### 3. Data Semantics, Data Integration, Data Interoperability (A. Gupta, C. Baru, I. Zaslavsky)

Concepts, techniques, and technologies related to data semantics, for extracting meaning from data; representing that information; and computing using that information. Use of semantic information for data access; data integration—the ability to interrelate data from independent data sources, including from different scientific domains; and data interoperability—the ability to use different tools on different types of data.

### 4. Data Processing, Workflow Systems, Provenance (I. Altintas, C. Baru, A. Gupta)

Processing strategies for different genres of data, e.g. graphs, streams, arrays, text, image/video/remote sensing, etc., for different scales of data, from small to large. Concurrent and parallel processing models for data including for shared-memory; shared-nothing, and distributed computing. Introduction to different computing models multi-core, cluster computing, grid computing, and cloud computing. Strategies for chaining together multiple computations using workflow systems. The concept of data provenance and techniques for tracking provenance of data.
Lab: Hands-on programming using large datasets.

### 5. Data Preparation, Preprocessing, and Transformation (N. Balac, C. Baru)

Preparing data for use in analysis and fitting datasets to the application. Data cleaning, data transformation, data reduction models, statistical methods, parsing, descriptive data summarization, discretization and concept hierarchy generation.
Lab: Use of Extract-Transform-Load tools and technologies for preprocessing of real data sets.

### 6. Data Mining (N. Balac)

Data mining concepts, mining of different types of data and knowledge, frequent pattern mining methods, linear and non-linear classification methods, cluster and outlier analysis, ensemble methods, trend analysis, text and web mining, model evaluation techniques.
Lab: Use of statistical and data mining packages for analysis of real data sets.

### 7. Visual Analytics (N. Balac, D. Nadeau, A. Chourasia)

The science of analytical reasoning facilitated by visual interactive interfaces including information visualization and scientific visualization. Visual data mining, visual classifier, class preserving projections, visualizing high-dimensional data, surface and

volume rendering and visual representation of large-scale collections of non-numerical information.
Lab: Use of visual analytics tools with real data.

**8. Practicum: Analysis and Evaluation of Big-Data Infrastructure (A. Gupta, C. Baru)**
Big-data refers to data that is extremely large in scale, arriving into a system at rapid rates, and/or requiring integration of data from a large number of independent, heterogeneous sources. This hands-on course will provide experience with several software stacks for big data processing including open-source systems such as, Hadoop, Cascading, MongoDB, Cassandra, Hyracks, Stratosphere, and Twitter Storm. The course will involve writing code in one or more of these, understanding some module of a chosen system in detail, and developing a comparative evaluation of the same with another system. This course will also provide an introduction to hardware trends for supporting big data applications.

**9. Geographic Information Systems and Spatial Data Analysis (I. Zaslavsky)**
It is estimated that 80-90 percent of today's digital data have some spatial characteristics. Geographic Information Systems (GIS) enable management, visualization, analysis, and modeling of such data. This course will focus on the technical underpinnings of GIS and its use for data analysis. The topics will include key concepts and applications of GIS, spatial data structures and formats, sources of various spatial data, spatial analysis and visualization, Map Algebra, error and uncertainty in spatial data, and foundations of spatial statistics. The hands-on component of the course will include working with widely-used GIS software and understanding its internals, as well as developing a spatial data analysis project. The course will also introduce specific issues and challenges in managing large distributed volumes of spatial data of several common types.

**10. Distributed Spatial and Observational Data Management and Integration (I. Zaslavsky)**
Many spatial and observational data are collected by multiple research groups around the world. Combined together, they would lead to a much more comprehensive portrait of environmental and other phenomena — if only one could consistently discover, interpret, access, and integrate these data in a scientifically rigorous manner. This is a key challenge in the development of cyberinfrastructure for the geosciences and other disciplines that rely on such measurements. This course will present a systematic review of the main issues and accomplishments in spatial data infrastructure development. The topics will include standards for spatial and observational data exchange, spatial databases and GIS servers, spatial and observational data services and middleware, online GIS, and spatial data integration and visualization. Application examples will come from several disciplines, such as environmental science, hydrology, and neuroscience. The hands-on component will consist of a class project to develop an online application that integrates spatial and observational data from several sources.