0.1 Data Science

Data science incorporates varying elements and builds on techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.

Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common. Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners. Some areas of research are:

- Cloud computing
- Databases and information integration
- Learning, natural language processing and information extraction
- Computer vision
- Information retrieval and web information access
- Knowledge discovery in social and information networks

Data scientists use an extensive understanding of business, combined with technical skills and statistical knowledge, to create methods for organizations to collect and interpret data. A data scientist helps an organization determine the questions that need answers, develops the methodology and technological tools needed to collect pertinent data, and builds the statistical models needed to derive answers from the data collected. In short, these professionals have the technical skills and business understanding needed to help an organization make decisions or develop useful products and services for customers, based on the analysis of data.

The exact particulars of a data scientist's job vary based on the industry in which the professional works. One data scientist might focus on the programming needed to gather specific data, while another uses existing tools in unique ways to further enhance the accuracy or effectiveness of data. Still another data scientist might combine existing tools and specially made tools to collect and analyze data in a way that helps the organization offer a new service or product.

For example, data scientists help develop many of the convenience applications used on social networking websites. Data is collected about each individuals employment and educational history, then this information is compared with current affiliations. Based on each individuals history and current connections, an application makes recommendations for additional connections, possible job leads, or products and services of interest to individual members. Results typically display on the users home page or main profile screen. Using technical skills and creativity, the data scientists who developed these applications helped each website create a more useful user experience.

Similar applications allow a data scientist to collect, analyze, and report information about website visitors, in-store shoppers, and other customer information. Depending on the goals of the organization, such information may be used to create custom shopping experiences or test various marketing strategies. Many websites, for example, have applications that display tailored advertisements based on customer behavior. Before launching these applications, a professional data scientist had to program a means to collect customer information, analyze it, and produce an appropriate result to display.

Different industries and, in fact, different companies within the same industry, have different needs when it comes to daily tasks completed by a data scientist. While the tasks may differ, the skills needed remain

the same. Professionals in this line of work need programming and other technical skills in order to develop appropriate tools to collect and manipulate data. Additionally, such professionals need creativity and critical thinking skills, as well as the ability to understand business needs, in order to know what data to collect and the different ways to interpret information.

Data Sciences

Data analytics (DA) is the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories. Data analytics is distinguished from data mining by the scope, purpose and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false. Qualitative data analysis (QDA) is used in the social sciences to draw conclusions from non-numerical data like words, photographs or video. In information technology, the term has a special meaning in the context of IT audits, when the controls for an organization's information systems, operations and processes are examined. Data analysis is used to determine whether the systems in place effectively protect data, operate efficiently and succeed in accomplishing an organization's overall goals.

Analytics

The term "analytics" has been used by many business intelligence (BI) software vendors as a buzzword to describe quite different functions. Data analytics is used to describe everything from online analytical processing (OLAP) to CRM analytics in call centers. Banks and credit cards companies, for instance, analyze withdrawal and spending patterns to prevent fraud or identity theft. Ecommerce companies examine Web site traffic or navigation patterns to determine which customers are more or less likely to buy a product or service based upon prior purchases or viewing trends. Modern data analytics often use information dashboards supported by real-time data streams. So-called real-time analytics involves dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use.

Data modeling

Data modeling is the formalization and documentation of existing processes and events that occur during application software design and development. Data modeling techniques and tools capture and translate complex system designs into easily understood representations of the data flows and processes, creating a blueprint for construction and/or re-engineering.

A data model can be thought of as a diagram or flowchart that illustrates the relationships between data. Although capturing all the possible relationships in a data model can be very time-intensive, it's an important step and shouldn't be rushed. Well-documented models allow stake-holders to identify errors and make changes before any programming code has been written.

Data modelers often use multiple models to view the same data and ensure that all processes, entities, relationships and data flows have been identified. There are several different approaches to data modeling, including:

- Conceptual Data Modeling - identifies the highest-level relationships between different entities.
- Enterprise Data Modeling - similar to conceptual data modeling, but addresses the unique requirements of a specific business.

- Logical Data Modeling - illustrates the specific entities, attributes and relationships involved in a business function. Serves as the basis for the creation of the physical data model.
- Physical Data Modeling - represents an application and database-specific implementation of a logical data model.

Predictive Modeling

Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends.

A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. In marketing, for example, a customer's gender, age, and purchase history might predict the likelihood of a future sale.

In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.

Predictive modeling is used widely in information technology (IT). In spam filtering systems, for example, predictive modeling is sometimes used to identify the probability that a given message is spam. Other applications of predictive modeling include customer relationship management (CRM), capacity planning, change management, disaster recovery, security management, engineering, meteorology and city planning.