# [CRISP\_DM]

[Fundamental concept: From a large mass of data, information technology can be used to find informative descriptive attributes of entities of interest. In our churn example, a customer would be an entity of interest, and each customer might be described by many attributes, such as usage, sometimes this process is referred to roughly as finding variables that “**correlate**” with churn.

A business analyst may be able to hypothesize some and test them, and there are tools to help facilitate this experimentation (see “Other Analytics Techniques and Technologies” on page 35). Alternatively, the analyst could apply information technology to automatically discover informative attributes—essentially doing large-scale automated experimentation. Further, as we will see, this concept can be applied recursively to build models to predict churn based on multiple attributes.

Fundamental concept: If you look too hard at a set of data, you will find something—but it might not generalize beyond the data you’re looking at. This is referred to as **overfitting** a dataset.

Other Analytics Techniques and Technologies:

* Statistics
* Querying: A query is a specific request for a subset of data or for statistics about data, formulated in a technical language and posed to a database system. Many tools are available to answer one-off or repeating queries about data posed by an analyst. These tools are usually frontends to database systems, based on Structured Query Language (SQL) or a tool with a graphical user interface (GUI) to help formulate queries (e.g., query-by example, or QBE). For example, if the analyst can define “profitable” in operational terms computable from items in the database, then a query tool could answer: “Who are the most profitable customers in the Northeast?” The analyst may then run the query to retrieve a list of the most profitable customers, possibly ranked by profitability. This activity differs fundamentally from data mining in that there is no discovery of patterns or models
* Data warehousing: Data warehouses collect and coalesce data from across an enterprise, often from multiple transaction-processing systems, each with its own database. Analytical systems can access data warehouses. Data warehousing may be seen as a facilitating technology of data mining.
* Regression Analysis:
* Machine Learning and Data Mining: Machine Learning as a field of study arose as a subfield of Artificial Intelligence, which was concerned with methods for improving the knowledge or performance of an intelligent agent over time, in response to the agent’s experience in the world. Such improvement often in‐ volves analyzing data from the environment and making predictions about unknown quantities, and over the years this data analysis aspect of machine learning has come to play a key role in the field. The field of Data Mining (or **KDD: Knowledge Discovery and Data Mining**) started as an offshoot of Machine Learning, and they stay intricately linked. Both fields are concerned with the analysis of data to find useful or informative patterns. Techniques and algorithms are shared between the two; indeed, the areas are so closely related that researchers commonly participate in both communities and transition between them seamlessly. Nevertheless, it is worth pointing out some of the differences to give perspective. Speaking generally, Machine Learning is concerned with many types of performance improvement, it includes subfields such as robotics and computer vision that are not part of KDD.
* The topic of explanatory modeling versus predictive modeling can elicit deepfelt debate,