MATERIALS TECHNOLOGY @TMS

MaterialsTechnology.org

Materials Informatics Glossary : Terms and Acronyms – v. 1.0

Kim Ferris
Pacific Northwest National Laboratory

February 2007

The following is a compilation of terms utilized in materials informatics, with hyperlinks provided for additional information.

Anomaly Detection Identify data or observations that are significantly different from the norm. The ability of identify such anomalies can identify errors in reported properties or new classes of materials with unusual properties

Association Analysis Used to discover patterns that describe strongly associated features in data (e.g. the frequency of association of a specific materials property to materials chemistry)

CALPHAD Polynomial-based model for the phase diagram behavior of a material.

Classification Organization of library materials by a hierarchy of subject categories.

Correlation In probability theory and statistics, correlation, also called correlation coefficient, is a numeric measure of the strength of linear relationship between two random variables. In general statistical usage, *correlation* or co-relation refers to the departure of two variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

Covariance The expected value of the product of the deviations of two random variables from their respective means. It is the arithmetic mean of the products of the deviations of corresponding values of two quantitative variables from their respective means. Intuitively, the covariance is the measure of how much two variables vary together.

Data Mining/Knowledge Extraction An interdisciplinary field merging ideas from statistics, machine learning, databases, and parallel and distributed computing. The goal of data mining is the extraction of knowledge and insight from massive databases.

Database An organized collection of data.
First Principles Electronic structure calculation methodology using the density functional theory (DFT) formalism. First principles calculations provide a broad spectrum of materials descriptors; thermodynamic, electronic, spectral, mobilities.

Information Architecture Organized structure of data evolution from raw data to developed model to predictive behavior.

Linear Search A search algorithm, also known as sequential search, that is suitable for searching a set of data for a particular value.

Linear Discriminant Analysis (LDA) LDA is a common statistical classification technique that guarantees maximal linear separability between two classes of data. This is performed by maximizing the ratio of between-class variance to within-class variance. Unlike PCA or SVMs, LDA does not change the shape and location of the original data but focuses on discovering a decision boundary between the given classes.

Machine Learning An area of artificial intelligence concerned with the development of techniques which allow computers to "learn". More specifically, machine learning is a method for creating computer programs by the analysis of data sets.

Material Signature The signature of a material is defined using both its crystal structure and composition.

MatML Materials markup language.

Model Development Identify and/or evolve a physical model for materials properties; perform knowledge extraction.

Pettifor Map Structure map of crystal systems plotted against Mendeleev A and B numbers.

Partial Least Squares (PLS) Partial least squares bears some relation to principal component analysis; instead of finding the hyperplanes of maximum variance, it finds a linear model describing some predicted variables in terms of other observable variables. It is used to find the fundamental relations between two matrices (*X* and *Y*), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space. Intuitively it is a data mining technique used in situations where there are fewer observations than predictor variables.

Principal Component Analysis (PCA) Mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. PCA is a statistical approach that yields patterns and relationships in multivariate datasets. PCA generates a new orthogonal coordinate system where the new variables are independent linear combinations of the original variables and capture features of the original data.

Regression Analysis Statistical method where the mean of one or more random variables is predicted conditioned on other (measured) random variables. In particular, there is linear regression, logistic regression, Poisson regression and supervised learning. Regression analysis is more than curve fitting (choosing a curve that best fits given data points); it involves fitting a model with both deterministic and stochastic components. The deterministic component is called the **predictor** and the stochastic component is called the **error term**. The simplest form of a regression model contains a dependent variable (also called "outcome variable," "endogenous variable," or "Y-variable") and a single independent variable (also called "factor," "exogenous variable," or "X-variable").

Relational Database A database based on the relational model. Strictly speaking the term refers to a specific collection of data but it is invariably employed together with the software used to manage that collection of data. That software is more correctly called a relational database management system, or RDBMS.

Structure Characteristic identifier for a data entry; in simple form the molecular or crystal structure

Structure Map Classically, a two dimensional representation of a material property against defining structural variables. Structure maps such as the Pettifor map are analogous to the scores plots from PCA with the Mendeleev A and B numbers representing the first two latent variables. More recent developments in visual analytics have allowed for higher dimensional presentations.

Supervised Learning Machine learning technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs. The output of the function can be a continuous value (called regression), or can predict a class label of the input object (called classification). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output). To achieve this, the learner has to generalize from the presented data to unseen situations in a "reasonable" way (inductive bias).

Support Vector Machines (SVMs) A set of related supervised learning methods used for classification and regression. SVMs have been noted as an excellent generalized supervised learning approach to classification based on statistical optimization theory.