Unsupervised Learning

Nik Bear Brown

In this module, we study unsupervised learning. Unsupervised learning is about finding patterns and structure in unlabeled data. This is especially true in the analysis of "big-data." A central issue with big-data is that the data is not typically labeled or only a very small subset is, as labeling often requires human annotation. As such some consider unsupervised learning the 'future' of big-data.

Some neural networks may be considered unsupervised learning as well. Neural network models like Self-organizing maps, restricted Boltzmann machines and Hopfield networks are a form of unsupervised learning. However, we'll discuss and implement various neural statistical learning models in a module entirely devoted to neural networks. We'll simply note in this module that some statistical neural learning models are unsupervised and put off discussing the details until the neural network module.

In the first lesson, we define unsupervised learning and introduce essential concepts and jargon.

In the second lesson we cover clustering theory, apply and evaluate various clustering techniques such as hierarchical clustering, k-means clustering and others.

In the third lesson we cover expectation-maximization theory, apply and evaluate expectation-maximization based clustering.

In the fourth lesson we cover linear discriminant analysis (LDA) theory, apply and evaluate linear discriminant analysis (LDA) as a classification technique.

In the fifth lesson we cover association rule learning theory, apply and evaluate association rule learning to generate predictive rules.

Rationale: The power of unsupervised machine learning is that it might spot salient correlations, connections and structure in data absent of human bias (i,e. patterns that no human would have thought to look for), and when the amount of data makes it difficult or impossible to label a significant subset.

Additional packages needed

To run the code you need no additional packages.

Questions

Feel free to tweet questions to [@NikBearBrown](https://twitter.com/NikBearBrown)

Unsupervised Learning

Unsupervised Learning tries to find hidden structure in unlabeled data. Unsupervised Learning has no feedback. This typically means that the data is not labled. Supervised learning has feedback/labels.

For example, if we had some e-mail that labbled some e-mail as spam and not-spam we could use a supervised learning algorthm like Naive Bayes spam filtering to classify any new e-mail as spam or non-spam.

However, if we didn't have labeled data (also called training data) we might still be able to cluster spam and non-spam if we could come up with some measure of similarity or distance between words in the e-mails.

Unsupervised Learning

Unsupervised Learning is computational machinary to extract patterns from unlabled data

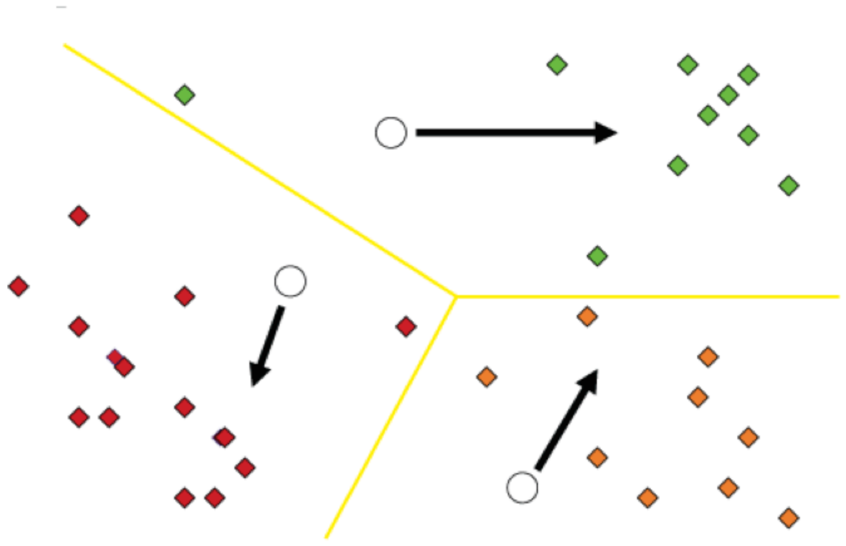
Image courtesy of clipartist.net

Clustering is probably the most common technique whose goal is to group data into similar groups based on a similarity/distance measure. Most of the focus of this module will be on clustering. We'll discuss and implement clustering in Lesson 2 of this module

Types of Clustering

Partitioning-based clustering

Partition-based clustering iterativley assigns data to groups based on a disatnce metric or similarity/dissimilarity measures until the group assignments are stable (i.e. the algorithm converges.)



K-means

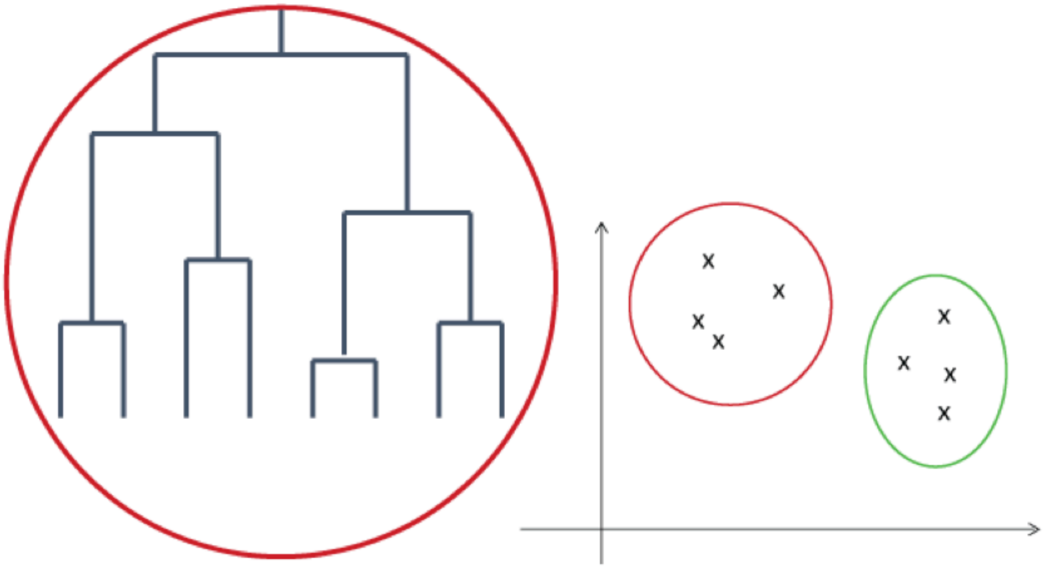
K-means clustering

Image courtesy of clipartist.net

Hierarchical clustering

Hierarchical clustering is a method of cluster analysis which builds a hierarchy of clusters either from the "bottom up" (agglomerative hierarchical clustering) by combiinimg the two "closet" or "most similar" points into a cluster then the next two closet, until all points are merged as one moves up the hierarcy.

Conversely the data is split "top down" (divisive hierarchical clustering) so that there is the most seperation between the split groups and further splits are performed recursively untl there is a single data point in each group.



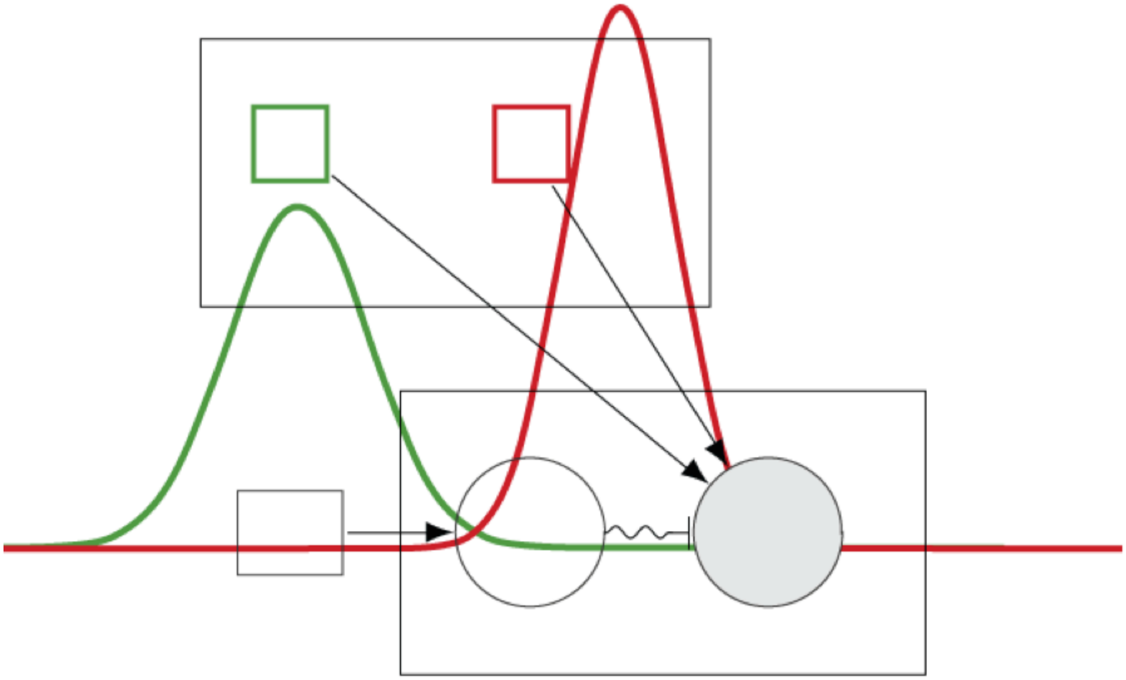
Hierarchical clustering

Hierarchical clustering

Image courtesy of clipartist.net

Density-based clustering

ensity-based clustering is based on probabilty distribution models. Clusters are defined as a set of distibutions then assigning points to the distibutions that they are most likely to belong. The most prominent method is known as Gaussian mixture models that represent the clusters as multivariate normal distributions.



Gaussian mixture model

$$p(\boldsymbol{\theta}) = \sum_{i=1}^{K} \phi_i \, \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$$

Gaussian mixture model

Image courtesy of clipartist.net

Expectation-maximization

The Expectation-maximization algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. We'll discuss and implement association rule learning in Lesson 4 of this module.

Given a statistica which generates a set **X** of observed data, a set of unobserved latent data or missing values **Z**, and a vector of unknown parameters $\boldsymbol\theta$, along with a likelihood function $L(\boldsymbol\theta; \mathbf{X}$, **Z**) $= p($**X**, $\mathbf{Z}|\boldsymbol\theta)$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data.

$$ L(\boldsymbol\theta; \mathbf{X}) = p(\mathbf{X}|\boldsymbol\theta) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol\theta) $$

However, this quantity is often intractable (e.g. if **Z** is a sequence of events, so that the number of values grows exponentially with the sequence length, making the exact calculation of the sum extremely difficult).

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of **Z** given **X** under the current estimate of the parameters

$$ \boldsymbol\theta^{(t)}: Q(\boldsymbol\theta|\boldsymbol\theta^{(t)}) = \operatorname{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol\theta^{(t)}}\left[\log L (\boldsymbol\theta;\mathbf{X},\mathbf{Z}) \right] $$

Maximization step (M step): Find the parameter that maximizes this quantity:

$$ \boldsymbol\theta^{(t+1)} = \underset{\boldsymbol\theta}{\operatorname{arg\,max}} \ Q(\boldsymbol\theta|\boldsymbol\theta^{(t)}) $$

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) which is related to blind signal separation techniques, such as those we've already studied in the dimensionality reduction module like: Principal component analysis, Independent component analysis,Non-negative matrix factorization, and Singular value decomposition. LDA explicitly attempts to model the difference between the classes of data. That is, like regression analysis, LDA attempts to express one response variable as a linear combination of other features or measurements. So while the math is similar to PCA and SVD, the intent is similar to regression (especially logistic regression) in that it creates a model. We'll discuss and implement association rule learning in Lesson 4 of this module.

LDA is often based upon s Fisher's linear discriminant. Fisher defined the separation between these two distributions to be the ratio of the variance between the classes to the variance within the classes:

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\vec{w}\cdot\vec{\mu}_1 - \vec{w}\cdot\vec{\mu}_0)^2}{\vec{w}^T\,\sigma_1\,\vec{w} + \vec{w}^T\,\sigma_0\,\vec{w}} = \frac{(\vec{w}\cdot(\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T\,(\sigma_0 + \sigma_1)\,\vec{w}}$$

This measure is, in some sense, a measure of the signal-to-noise ratio for the class labelling. It can be shown that the maximum separation occurs when

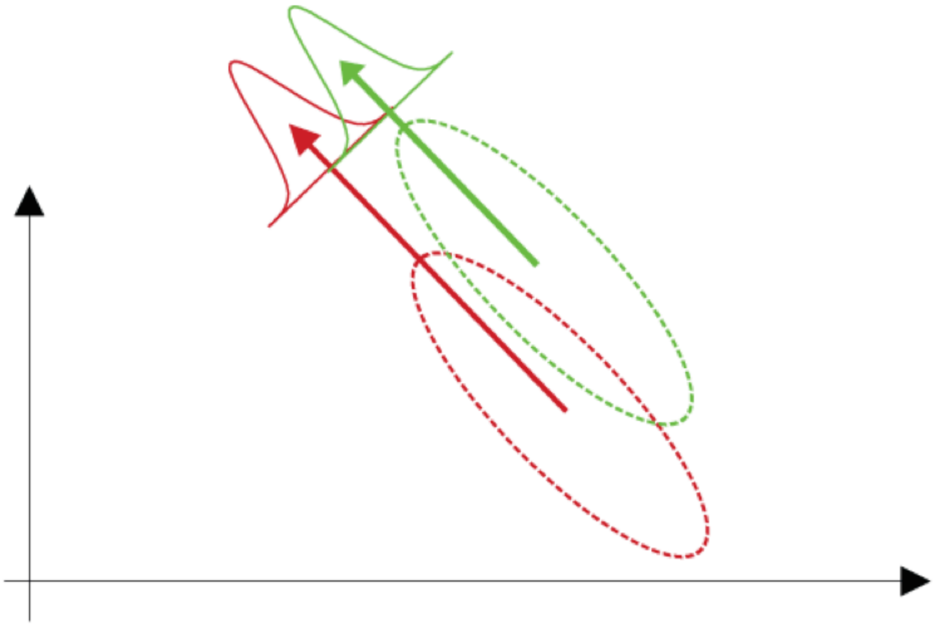
$$\vec{w} \propto (\sigma_0 + \sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$

When the assumptions of LDA are satisfied, the above equation is equivalent to LDA.

In essence, LDA picks a new basis that gives:

To do this, LDA uses eigenvectors based on between-class and within-class covariance matrices.

$$max \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$



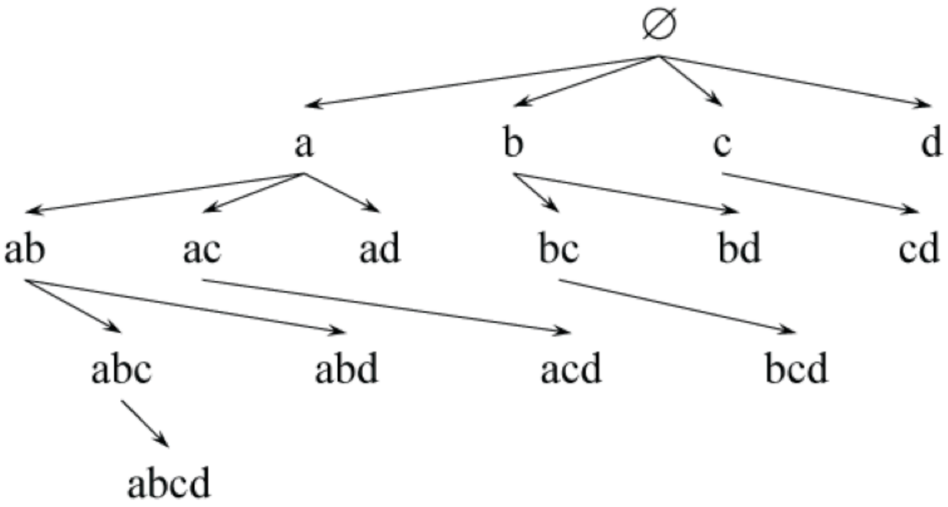
Linear Discriminant Analysis (LDA)

Image courtesy of clipartist.net

Association rule learning

Association rule learning is an Unsupervised alogorthm interesting relations between variables in a data set. We'll discuss and implement association rule learning in Lesson 5 of this module. Association rule learning aims to discover interesting correlation or other relationships in large databases. It finds a rule of the form:

$$if \quad A \quad and \quad B \quad then \quad C \quad and \quad D$$

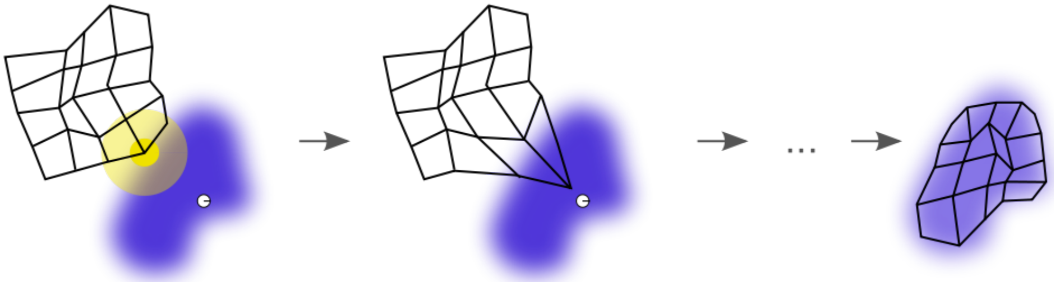


Association rule Set Enumeration Tree

Association rule Set Enumeration Tree

Artificial neural networks (ANN)

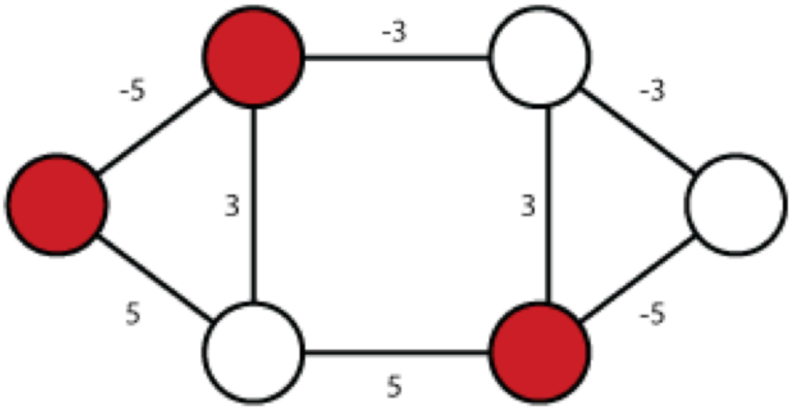
Artificial neural networks are a are a family of statistical learning models inspired by biological neural networks. Some neural network models like Self-organizing map's,adaptive resonance theory (ART), Restricted Boltzmann machines and Hopfield network's. We'll discuss and implement various neural statistical learning models in a module entirely devoted to neural networks. We'll simply note here that some statistical neural learning models are unsupervised and put off discussing the details until the neural network module.



An illustration of the training of a self-organizing map

An illustration of the training of a self-organizing map

Art "Somtraining" by Mcld. Licensed under CC BY-SA 3.0



Hopfield network

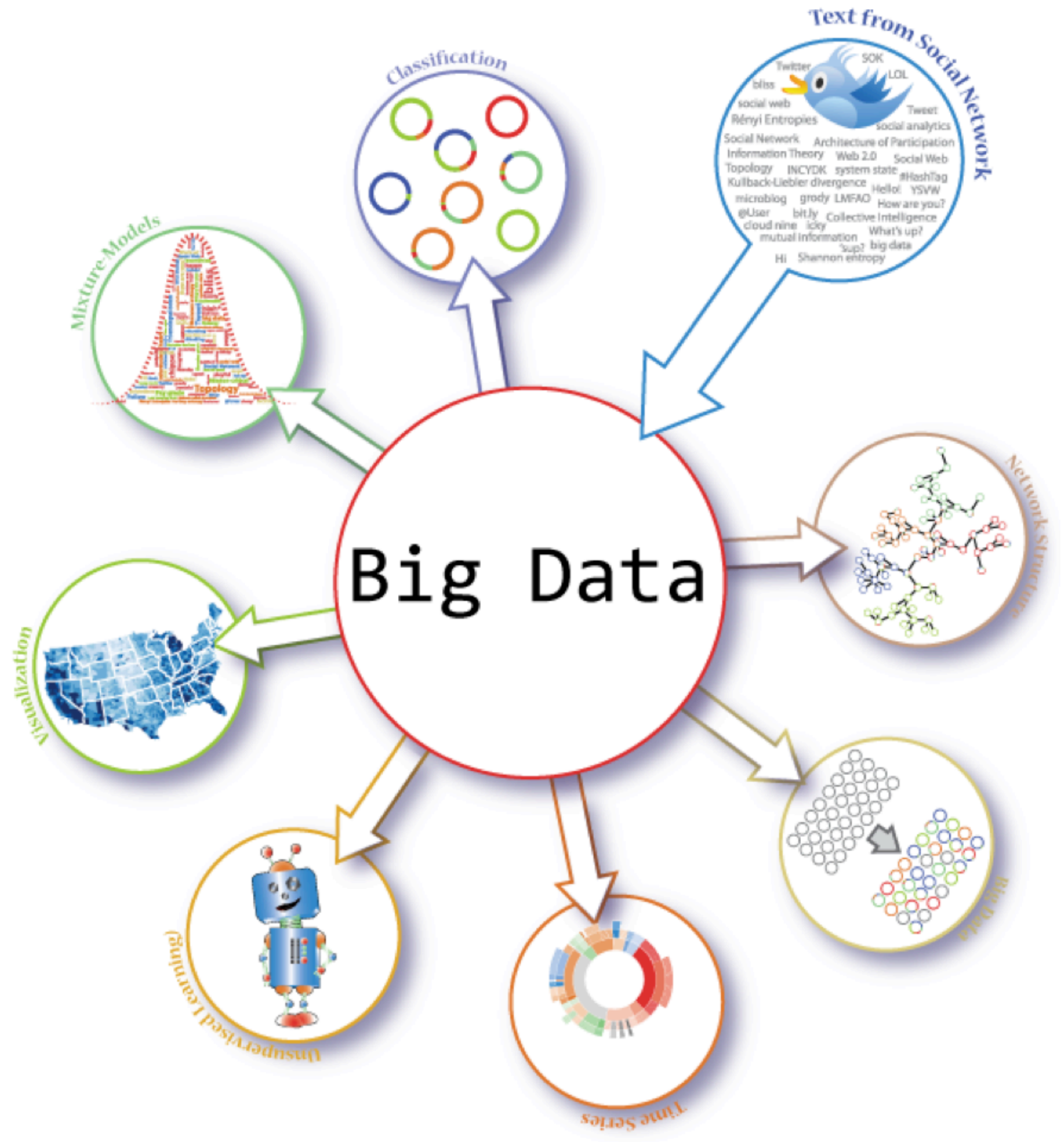
An illustration of a stable Hopfield network

Image courtesy of clipartist.net

Big Data and Unsupervised Learning

“Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone.” [1] Due to the growing complexity of digital social networks and the huge quantity of data they produce daily, it’s important that we deal with “big-data” efficiently. A number of tools and technologies (Map-Reduce, NoSQL, Hadoop, Hive, cloud computing, parallel processing, clustering, MPP, virtualization, large grid environments, and so on) have been developed to store and process big data. While big-data technologies have established the ability to collect and process large amounts of data, most organizations struggle with understanding the data and taking advantage of its value. According to an Economist report: “Extracting value from big data remains elusive for many organizations. For most companies today, data are abundant and readily available, but not well used.”[2,3]

A central issue with “big data” is that not all of the data will labled. As such, there is an expectation that unsupervised learning and research in unsupervised learning will take on an even greater importance in machine learning.



big data

Image courtesy of clipartist.net

References

(1) IBM. 2012. What is big data?

(2) Economist. 2012. Economist Intelligence Unit

(3) Quiñonero-Candela, Joaquin. 2008. Data Set Shift The MIT Press

(4) “Unsupervised learning” and the future of analytics

(5) Distance and Similarity Measures