

## Machine Learning – a unifying perspective & new paths

PG Madhavan, Ph.D.  
Chairman, Syzen Analytics, Inc., Seattle, WA, USA  
pgmad@syzenanalytics.com

*Dr. PG Madhavan is the Founder of Syzen Analytics, Inc. He developed his expertise in Analytics as an EECS Professor, Computational Neuroscience researcher, Bell Labs MTS, Microsoft Architect and startup CEO. PG has been involved in four startups with two as Founder.*

### **Major Original Contributions:**

- Computational Neuroscience of Hippocampal Place Cell phenomenon related to the subject matter of 2014 Nobel Prize in Medicine.
- Random Field Theory estimation methods, relationship to systems theory and industry applications.
- Early Bluetooth, Wi-Fi, 2.5G/EDGE and Ultra-wideband wireless technology standards and products.
- Currently developing Systems Analytics bringing model-based methods into current Analytics practice.

*PG has 12 issued US patents and over 100 publications & platform presentations to Sales, Marketing, Product, Industry Standards and Research groups. More at [www.linkedin.com/in/pgmad](http://www.linkedin.com/in/pgmad)*

**Pedro Domingos in his new book**, “The Master Algorithm”, has done us a huge favor. As is true of any emerging technology field, Machine Learning (ML) is a “bag of tricks” today; it takes a while for a unifying framework to emerge. Then, one can see various aspects of ML as special cases of a general theory rather than a grab-bag of tools and techniques.

Pedro has taken a great early step to such unification. He has collected all major ML initiatives into a taxonomy that makes sense; five schools of thought: the evolutionaries, connectionists, symbolists, Bayesians, and analogizers. I believe this does not go far enough in the unification of ML thought however . . .

From the early days of “ML”, I see Pattern Recognition and Classification as a better unifying perspective. In particular, the classic textbook of Duda & Hart, “Pattern Classification & Scene Analysis”, published in 1973 is my starting point!

Duda & Hart’s approach in simple terms is as follows. Given labelled samples, obtain a class description consisting of either a distance metric (Euclidean, intra-class, etc.) or a probability density function and then derive a decision rule (Maximum A-posteriori Probability, Bayes, etc.) from the description. The decision rule specifies a decision boundary in feature space among classes.

***Alternatively, decision surface can be derived directly from labelled samples which is then called a “Discriminant Function”, perceptron being an example. Then, most if not all current ML techniques can be seen as dueling methods to derive Discriminant Functions!***

Discriminant Functions can be linear or nonlinear (neural network with back-propagation, deep learning, support vector machines, kernel PCA, etc.) and outputs can be binary, integer or real valued. Various learning algorithms can be seen as belonging to the family of iterative/ recursive/ adaptive learning algorithms (Least Mean Square being a great old standby!) that update the parameters of the Discriminant Function as new data arrive.

In the discussion above, features were considered as “static” and not context-sensitive (for identifying a word within a sentence as an example). Context-sensitivity or Dynamics can be added to improve classification by incorporating Markov models (or Hidden Markov Models for tractable computations). Markov model is a special case of State Space Models which are well-studied in Systems Theory.

Setting aside Supervised Classification when labelled samples (or “desired signals”) are available, what can we do when there is no supervision? This is the realm of much harder Unsupervised Learning, which is very useful in transforming basic features into more and more meaningful ones. One usually brings in some overall desirable property to guide unsupervised learning. From the domain of “blind processing” (Radar signal processing, for example), Mutual Information among classes can be minimized as a learning process in the belief that the “best” classification happens when the classes have least overlapping information (better “efficiency” in representation).

Instead of entropy-related quantities that are hard to estimate, it is likely that Scale of Fluctuation which is related to “order” and “state space volume” may be a quantity to optimize for a new unsupervised learning process. (For more information on Scale of Fluctuation, refer to my papers, “[Instantaneous Scale of Fluctuation Using Kalman-TFD and Applications in Machine Tool Monitoring](#)”, 1997 & “[Kalman Filtering and time-frequency distribution of random signals](#)”, 1996).

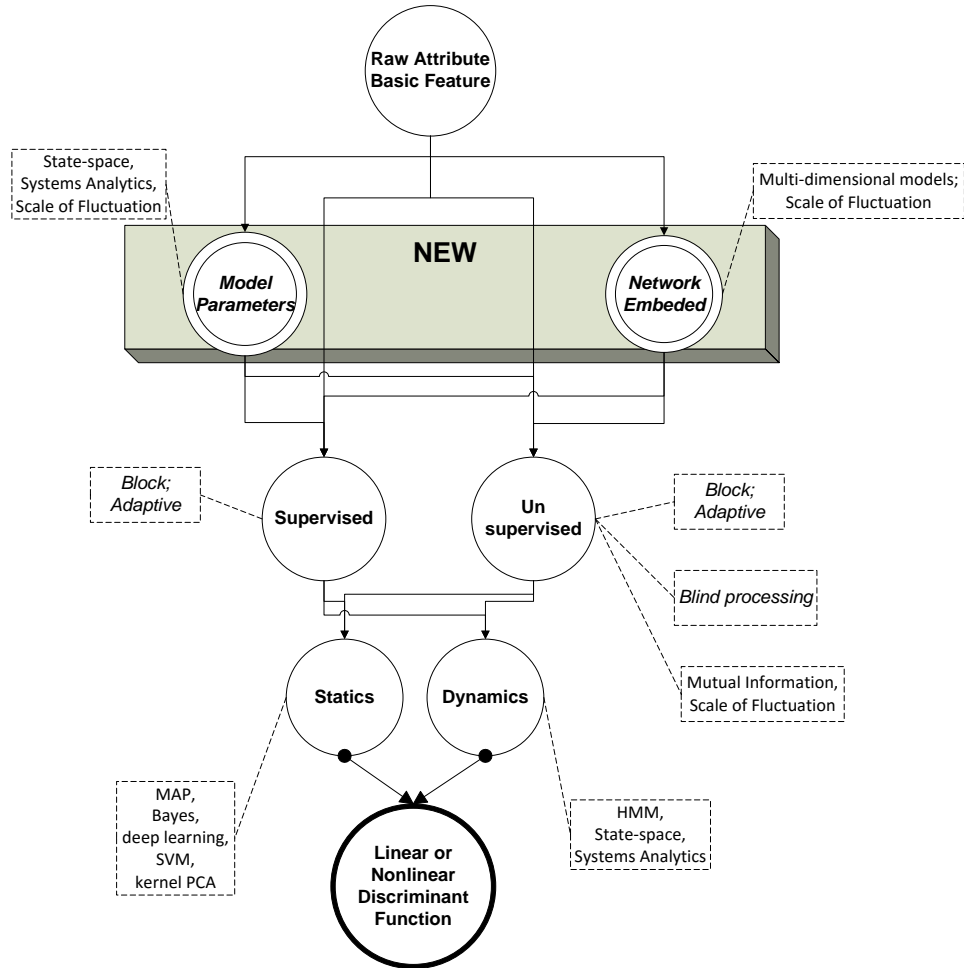
In all of the existing ML bags of tricks, we are still staying at the surface level! ***We are modeling the attributes or data DIRECTLY. What if we went one level deeper? Model the SYSTEM that generates the data!*** Syzen Analytics, Inc., takes such an explicit approach in what we call “SYSTEMS” Analytics” which has already demonstrated significant value in business applications.

In Syzen’s retail commerce application, our ***Systems Analytics approach hypothesizes that there is a system, either explicit or implicit, behind the scenes*** generating customer purchase behaviors and purchase propensities. This ‘one-level-deeper system model parameters’ can be more effective for pattern recognition and classification purposes instead of the data that the model generates! There is a long history of model parameters providing better estimates (in power spectrum analysis, for example). Scale of Fluctuation mentioned earlier seems to have another desirable property of quantifying “coupling” among deeper-level model parameters.

Context-sensitivity dynamics is a very good avenue to exploit. The dynamics could be over any independent variable (time always comes to mind first but it is only one of the possibilities). As I

noted in my recent blog ([“SYSTEMS Analytics – the next big thing in Big Data & Analytics”](#)), “Extensions to Systems Analytics in the future will be inspired by the insight that in reality, data exist in *\*embedded\** forms in preference and influence networks which are distributed in time and space” AND other independent dimensions (shopper preference, for example).

Let me pull all of the notions discussed so far into a diagram.



Once the patterns have been recognized and classes identified, the resulting classes can be used for all sorts of applications such as Recommendation Engine, Language Translation, Fraud Detection and many others. The approach I outline above allows you to take a unified approach till the application development stage. In doing so, the unified approach also points out new paths ahead for ML!

Some readers would have noticed an undertow of dichotomies while reading this “opinion piece”: Theoretic vs Heuristic; Formal vs Ad hoc; Mathematics vs AI; Electrical Engineering vs Computer Science academic departmental affiliations! I am firmly in the former camps. However, as an engineer, I am personally happy to start with heuristic solutions but quickly put them on firm mathematical foundations before “gotchas” and unintended consequences of ad hoc methods catch up with me. ☺

**The unification of ML proposed here opens up a multilane highway – join the journey and create more breakthroughs with us or on your own!**

*In this blog, I have not provided many references – web search will get you most; Pedro Domingos’ “The Master Algorithm” book is an excellent source of ML-related literature. For the newer and less familiar work, please contact me directly.*

