Intelligent Information Management, **2011, 3, 142-159**
doi:10.4236/iim.2011.34018 Published Online July 2011 (http://www.SciRP.org/journal/iim)

Intelligent Evidence-Based Management for Data Collection and Decision-Making Using Algorithmic Randomness and Active Learning

Harry Wechsler[1], Shen-Shyang Ho[2]
[1]*George Mason University*, *Fairfax*, *USA*
[2]*University of Maryland*, *College Park*, *Maryland*, *USA*
E-mail: *wechsler@gmu.edu*, *hoshensh@umd.edu*
Received April 13, 2011; *revised April* 28, 2011; *accepted May* 10, 2011

Abstract

We describe here a comprehensive framework for intelligent information management (IIM) of data collection and decision-making actions for reliable and robust event processing and recognition. This is driven by algorithmic information theory (AIT), in general, and algorithmic randomness and Kolmogorov complexity (KC), in particular. The processing and recognition tasks addressed include data discrimination and multi-layer open set data categorization, change detection, data aggregation, clustering and data segmentation, data selection and link analysis, data cleaning and data revision, and prediction and identification of critical states. The unifying theme throughout the paper is that of "compression entails comprehension", which is realized using the interrelated concepts of randomness vs. regularity and Kolmogorov complexity. The constructive and all encompassing *active learning* (AL) methodology, which mediates and supports the above theme, is context-driven and takes advantage of statistical learning, in general, and semi-supervised learning and transduction, in particular. Active learning employs *explore* and *exploit* actions characteristic of closed-loop control for evidence accumulation in order to revise its prediction models and to reduce uncertainty. The set-based similarity scores, driven by algorithmic randomness and Kolmogorov complexity, employ strangeness/typicality and p-values. We propose the application of the IIM framework to critical states prediction for complex physical systems; in particular, the prediction of cyclone genesis and intensification.

Keywords: Active Learning, Algorithmic Information Theory, Algorithmic Randomness, Evidence-Based Management, Kolmogorov Complexity, P-Values, Transduction, Critical States Prediction

1. Introduction

Information loaded with meaning and in context is an asset and is referred throughout this paper as evidence. Intelligent evidence-based management (EBM) of data concerns the value-added to raw data in order to transform it into (referential) information and (meaningful) knowledge using purposeful action (DKA). The motivation for EBM comes from Microsoft (MS) Cambridge (UK) Research manifesto Towards 2020 Science [1]. First and foremost the manifesto notes that what will most likely have a profound impact is "the leap from support to 'do' science, *i.e.*, computational science, to the integration of computing into the very fabric of science" leading to science-based innovation. The MS report highlights that an immediate and important challenge is that of end-to-end scientific data management, from data acquisition and data integration, to data treatment, provenance, and persistence" and including "the acquisition of a set of widely applicable complex problem solving capabilities, based on the use of a generic computational environment." This has been also advocated by the Computing Community Consortium http://cra.org/ccc/docs/init/From_Data_to_Knowledge_to_Action.pdf to enable 21st century discovery in science and engineering. The forthcoming revolution will be driven by "computational knowledge extraction. It can be best accomplished when one models and transforms data into evidence and knowledge [suitable to make predictions] and makes then use of existing [domain/meta] knowledge to engage in future actions/behaviors geared for further exploration and exploitation "actions". EBM using DKA can be fur-

ther traced to the "schemata" (kind of meaningful knowledge) proposed by Neisser [2] to account for the perceptual cycle where available information modifies the schema, the schema then directing then exploration, and exploration then sampling information. The outcome of the explorations—the information picked up-modifies the original schema. The Perception-Control-Action-Learning (PCAL) proposed by Wechsler [3] has expanded on the schema framework to include anticipation and learning driven by exploration and exploitation, which are mediated by control and action/manipulation.

This paper advances a comprehensive framework for EBM using DKA, which is concerned with data collection and decision-making related actions for reliable and robust event prediction and recognition. Data collection includes among others data aggregation, data cleaning, data collection, data selection, and data segmentation. Reliability concerns consistency and stability of the predictions made, while robustness is about coping with adversarial information, e.g., incomplete and corrupt information. The motivation for this paper comes from apparent synergies between information theory (IT) [4], algorithmic information theory [5,6], Kolmogorov Complexity [6,7] (see Section 6 and 7), statistical learning theory (SLT) [8], and algorithmic learning [9]. (Algorithmic information theory is a subfield of information theory and computer science that concerns itself with the relationship between computation and information.) The unifying theme throughout is that of "compression entails comprehension", which is realized using the interrelated concepts of randomness opposite regularity, Kolmogorov complexity, and minimum description length (MDL). The constructive active learning interface, which mediates and supports the above theme, is context-driven and takes advantage of semi-supervised learning [10] and transduction [8]. Active learning [11] is first and foremost about the choices made during data collection. It employs explore and exploit actions characteristic of closed-loop control for evidence accumulation in order to reduce uncertainty and to revise the prediction models. The event-recognition tasks addressed include multi-layer data categorization, change detection, data cleaning and data revision, data fusion, data segmentation, data selection and link analysis, and prediction and identification of critical states.

The basic functionalities active learning supports include decisions on where to explore and what (labeled and unlabeled) data to gainfully employ for further adaptation and modeling purposes; on how to process and fuse the information and knowledge acquired; on handling time-varying data streams, detecting change, and choosing when to revise the prediction models. Towards that end, set-based similarity scores are proposed. They are motivated by KC and driven by strangeness/typicality and p-values. The scores take advantage of associations, context and relationships; cohorts and rankings; transformations, e.g., hints and perturbations, censoring and imputation (to account for missing information), and data cleaning and revision (for consistency, error correction, and stability). Note that the terminology used in this paper employs similarity scores to estimate proximity. The use of "metrics" and "distances" is avoided, and the context makes clear when strict definitions are used.

The outline for the paper is as follows. EBM and DKA are discussed in Section 2 Active Learning; Algorithmic Information Theory and Closed-Loop Control; discriminative methods and practical intelligence; Kolmogorov complexity; and algorithmic randomness are discussed in Sections 3 -7, respectively. Strangeness/typicality and p-values; transduction and semi-supervised learning; and multi-layer and multi-set open set categorization are discussed in Sections 8 - 10 Active learning (see Section 3) using the explore and exploit paradigm motivates and supports functionalities related to data collection and evidence accumulation. The specific functionalities, concerning change detection; data aggregation/data fusion; data selection and link analysis; data cleaning and data revision; and prediction and identification of critical states, are described in Sections 11 - 16, respectively. New dimensions and requirements on EBM and DKA, e.g., for life sciences, are discussed in Section 17 The paper concludes in Section 18 with a brief summary and venues for future research.

2. Evidence-Based Management: From Data to Knowledge to Action

Evidence is any data or information so given, whether collected or derived from any source. Management is meant here to denote (organizational) activities pursued to accomplish desired goals and objectives related to data collection and assimilation in an efficient and effective fashion. For all purpose, management comprises planning, organizing, directing, and controlling the activities associated with data collection for the purpose of knowledge discovery and management thereafter. EBM is about the value-added to raw data in order to transform it into information and knowledge using purposeful action. EBM is chartered with the explicit mission to extract, manage, and revise current knowledge. It is data-centric and data-driven, and it is above all about actions directed at decision-making. DKA involves the effective and efficient exploration and exploitation of the data landscape in order to assimilate massive and complex amounts of data and to optimally complete specific tasks. The range of tasks considered throughout this paper, for illustrative

purposes, surrounds recognition, where the primary task is that of multi-layered open set categorization to perform (unlabelled) data annotation and revision. Knowledge amounts to the models learned and relearned ("revised") in order to make consistent and stable predictions mostly on detection, classification, and discrimination.

Knowledge discovery is about modeling the environment for the purpose of reliable and robust prediction. Reliability in terms of consistency (in the limit) and stability of the predictions made, and robustness for dealing with incomplete and corrupt (adversarial) data sources and actions. The knowledge discovered is tasked to facilitate, guide, and support decision-making. EBM and DKA expand first on intelligent information management (IIM) [12], with the latter limited to a life—cycle of data creation, acquisition, organization, storage, retrieval, dissemination and sharing, and use, but devoid of adaptation and knowledge discovery. EBM and DKA expand also on knowledge management (KM) [13,14], with the latter assuming that much of knowledge already exists but devoid of the critical adaptive exploration and exploitation cycles, which are geared to further knowledge discovery and knowledge betterment. KM merely comprises a range of strategies and practices used (in an organization) to identify, create, represent, distribute, and enable adoption of insights and best practices. Such knowledge is already available and needs only to be embedded for practical uses.

EBM and DKA involve for all practical purposes successive hypothetical-deductive cycles of discovery, where an initial predictive model learned from data guides the iterative collection of new data, model revision with new hypotheses/inferences ("labeling") made on unlabeled data, and so on. Examples of such revisions for signal tracking and interpretation include Sequential Monte Carlo (SMC) methods/Sequential Importance Sampling (SIS) also known as particle filtering [15]. The more complex active learning process described later also includes (a) change and anomaly (outlier, surprise, large and extreme values) detection; (b) data perturbations, synthesis, and class membership revisions (caused by possible errors in annotation), with the latter including filling in for missing data using anticipation, censoring and imputation, and/or proactive learning; and (c) mixed modes of adaptation using co-training, transfer learning, and/or multi-task learning for optimal resource-bounded data collection. Towards that end, EBM and DKA have access to multi-set (of instances) similarity scores, which take further advantage of associations, context, and relations.

3. Active Learning

One can approach EBM and DKA in terms of (progressive) evidence accumulation from time-varying data streams for the purpose of data collection, reasoning ("prediction"), and adaptation. This view related to active learning is engaged in data compression while it filters, fills in, and summarizes data contents. Active learning seeks for the patterns most responsible to generate and model the data, while always on look to detect when the data generation models change and to choose the means and ways to best process the data. Active learning is all encompassing, including autonomic computing [16] and W5+. Autonomic computing, also referred to as self-management, is about closed-loop control. It provides basic functionalities, e.g., self-configureuration (for planning and organization), self-optimization (for efficacy), self-protection (for security purposes), and self-healing (to repair malfunctions).

W5+ answers questions related to What data to consider, When to get/capture the data and from Where, and How to best process the data. An additional Who (is) question about identity becomes relevant to biometrics and identity management. Consider now intelligence analysis where directed evidence accumulation should also consider and document the explanation Why dimension. The Why dimension interrelates observations and hypotheses (models) duly ascribed (abducted possibly using analogy reasoning, Bayesian (belief) networks [17], and/or causality [18]) and expectations to be met. The Bayesian networks (for inference and validation purposes) assist with optimal and incremental/progressive intelligence data collection. In a fashion similar to signal processing and transmission, the "progressive" aspect signifies incremental access and/or display of crucial evidence, which at some point is enough to solve the puzzle and/or make recognition apparent.

The dimensions and taxonomy for the resources and processes addressed by active learning are as follows. Data, information, knowledge, and meta-knowledge form one dimension. Searching, categorization, modeling, and prediction define another dimension. Hedging/punting, risk, and decision-making make up yet another dimension. The interplay between active learning and data collection is all encompassing and includes data aggregation, data categorization (detection, discrimination, and classification), data cleaning, data imputation, data revision, data segmentation, and data selection. The same interplay mediates exploration and exploitation, while it addresses specific W5+ questions. Exploration and exploitation entail explanation using domain knowledge encoded using Bayesian Networks an/or Hidden Markov Models (HMM), on one side, and accuracy and confidence in the predictions made, on the other side, for bet-

ter attention and selectivity, focus, anticipation, and improved forecasts.

Everything about active learning is data centric. It is also about on-line learning and prediction with the purpose of data analytics for streaming data. Active learning is further sensitive to change and drift, which suggests that a time-varying D(t)K(t)A(t) model, indexed by time t, needs to supplant DKA in support of EBM. Towards that end, active learning considers associations, context, granularity, relationship, and space *x* time domains; leverages local and global context, on one side, and central and distributed processing, on the other side; and last but not least it involves multi—strategy learning.

4. Algorithmic Information Theory and Closed-Loop Control

The medium that facilitates EBM using DKA takes advantage of Algorithmic Information Theory and Closed-Loop Control (CLC). The starting point for this medium is Information Theory, which is about storage and communication, in general, and capacity, compression, accuracy, rate-distortion, and error correction, in particular. Algorithmic Information Theory provides for information processing, in general, and algorithmic issues related to the interrelated aspects of compression and prediction, in particular. Algorithmic Information Theory, throughout this paper, is about compressibility and generalization for the purpose of learning. The better some model or theory can compress training ("learning") data, the better the generalization achieved is, and the more effective, reliable, and robust the prediction ability becomes. This conforms with the dictum enunciated by Leibniz that "comprehension entails compression," and is conceptually similar to the MDL inductive principle, which gives support to discriminative methods for recognition (see Section 5). AIT, in general, and KC, in particular (see Sections 6 and 7) are related to CLC in terms of "universal prediction" [5] using statistical learning and similarity scores. Solid relations, related to accuracy performance, can be further established between information theory and statistical learning theory [19] to enhance modeling and prediction.

Prediction requires adaptation and learning, with the latter denoting "changes in the system that are adaptive in the sense that they enable the system to do the same tasks drawn from the "same" population more efficiently the next time" [20]. Learning is by default incremental and on-line learning should be preferred to batch learning (see Section 15). Note that learning plays a fundamental role in facilitating "the balance between internal representations and external regularities" [21], with CLC responsible for purposeful and successful action and behavior. One can thus connect IT, AIT, and CLC in terms of technical performance ("efficacy"), meaning and semantics, and action and behavior ("effectiveness").

CLC is also about causation, feedback, and change. Similar to cybernetics, which has been characterized by Louis CoufFigurenal as "the art of ensuring the efficacy of action", CLC can be approached as the interdisciplinary study of autonomous regulatory systems (see Section 3 for the self-management aspect of active learning). It comes to the fore when an action causes some change in the environment and that change becomes manifest to the system via information, or feedback, which then causes the system to adapt to new conditions. CLC participates in circular and causal chains, which transition from action to exploratory sensing ("data collection"); modeling (for exploitation), prediction and evaluation, and again to (sensory) action. This is the very epitome for DKA. Sensory action is much more than merely data collection. It also involves (internal) perturbations of the data available (both labels and representation) and subsequent revision of current models (see Section 15). The challenges CLC faces are twofold, behavior effectiveness subject to efficiency in terms of the resources used, and proper structure and organization for the control system responsible for the displayed behavior. This requires among others context and self-organization using clustering.

Action and adaptation are primordial and most important for CLC. They are interrelated using the observer, structural couplings, and "breaking down" and "thrownness" aspects of behavior-based and grounded intelligence. These are mere reflections of feedback, context, and proper generalization, with practical understanding being more fundamental than detached theoretical understanding (see Section 5). Context includes intricate relationships between understanding and existing ontologies [22].

Along the same lines, Heidegger [22] "insists that it is meaningless to talk about the existence of objects [knowledge] and their properties in the absence of concernful activity, with its potential for breaking down" [wrong predictions]. Closed-loop control supports both meaningful action and adaptation, and involves iterative explore and exploit cycles across the data landscape. Furthermore, the goal for predictive learning is behavior-based imitation (of the unknown data model) rather than its identification. The inductive principle, e.g., transduction (see Section 9), is responsible for the particular model selected and it directly affects the generalization ability in terms of prediction risk, *i.e.*, the performance on unseen/future (unlabeled test) data. This is the main reason to prefer discriminative rather than (normative) generative (prescriptive) methods. Discriminative methods are discussed next.

5. Discriminative Methods and Practical

Intelligence

Discriminative methods support practical intelligence, in general, and inference and prediction, in particular, for the purpose of discrimination. Progressive processing, (transformational and integrative) evidence accumulation, likelihood ratios (LR) and odds, and fast decision-making are the hallmark of practical intelligence. There is no time for expensive density estimation and marginalization, which are characteristic of generative methods (see below). This together with apparent relationships between discriminative methods and Kolmogorov complexity (see Section 6 and 7) are the motivation behind the similarity scores proposed throughout this paper for the purpose of categorization and recognition. Discriminative methods seek for non-accidental coincidences and take advantage of infomax and sparse codes for association [23].

Formally, "the goal of pattern classification can be approached from two points of view: informative [generative]—where one learns the class densities, e.g., HMM, or discriminative—where the focus is on learning the class boundaries without regard to the underlying class densities, e.g., logistic regression and neural networks" [24]. The generative methods, normative and prescriptive in nature, synthesize the classifier by learning what appear to be the specific class distributions, and make their decisions using maximum a-posteriori probability (MAP). The discriminative methods imitate and model only what it takes to make the classification effective. The discriminative methods are in tune with both structural risk minimization [8] and transduction (see Section 9). They are aligned with structural risk minimization (SRM) because they lock on what is essential for discrimination, and they are aligned with transduction because discrimination implements local (cohort) estimation. Jebara [25] has actually suggested that imitation should be added to the generative and discriminative methods as another model for learning and reasoning. Jebara recalls from Edmund Burke that "it is by imitation, far more than by percept, that we learn everything." Discriminative methods make fewer assumptions and fewer assumptions mean less chance to make mistakes.

Discriminative methods avoid estimating how the data has been generated and instead focus on estimating the posteriors (for recognition) in a fashion similar to the use of likelihood ratios (LR) and odds. The informative approach for 0/1 loss assigns some input x to the class $k \varepsilon K$ for which the class posterior probability $P(y = k \mid x)$ yields the maximum. The MAP decision used by generative methods requires instead access to the log-likelihood $P_\theta(x, y)$. The optimal (hyper) parameters θ are learned using maximum likelihood (ML) and a decision boundary is then derived, which corresponds to a minimum distance classifier. The discriminative approach models directly the conditional log-likelihood or posteriors $P_\theta(y \mid x)$. The optimal parameters are estimated using ML leading to the discriminative function,

$$\lambda_k(x) = \log\left[\frac{P(y = k|x)}{P(y = K|x)}\right]$$

which is similar in use to the Universal Background Model (UBM) for similarity score normalization and/or LR definition. The comparison takes place between some specific class membership k and a generic distribution (over K) that describes everything known about the population at large. The discriminative approach has been found [24] to be more flexible and robust compared to informative/generative methods because fewer assumptions need to be made.

6. Kolmogorov Complexity

The Kolmogorov complexity $K(x)$ of a finite string x is the information in x defined by the length of the shortest program for a reference Universal Turing Machine (encoded in binary bits) that outputs the string x [6]. Information distance $d(x, y)$ is the length of the shortest program on the reference Universal Turing Machine computing y from x and x from y. It has been proven that up to an additive logarithmic term, $d(x, y) = \max\{K(x \mid y), K(y \mid x)\}$ [26] where $K(x \mid y)$ is the conditional complexity of x given y. The normalized information distance (*NID*) between two strings x and y is formally defined as

$$NID(x, y) = \frac{\max\{K(x|y), K(y|x,)\}}{\max\{K(x), K(y)\}}$$

with $K(x)$ (and $K(y)$) approximated by a real-life compressor C such that $C(x)$ is the length of the compressed version of x [6].

The normalized information distances one may consider can be obtained from naturally observed perturbations, e.g., affine, and (error induced) revisions, e.g., class membership. Additional metrics used to approximate the information distances are derived using statistical learning, in general, and transduction and semi-supervised learning, in particular. Towards that end, strangeness/typicality and p-values (see Section 8), driven by cohort similarity and relative rankings, respectively, quantify either randomness or regularity for the purpose of compression and thus comprehension and recognition. Additional indirect similarity scores suitable for EBM and DKA, which are driven by Kolmogorov complexity, are introduced throughout the paper based on the specific functionalities addressed.

7. Algorithmic Randomness

Let x be a binary string belonging to set S. $K(x \mid S)$ is the Kolmogorov complexity of x given S. The randomness deficiency $D(x \mid S)$ for x given S is $D(x \mid S) = \log |S| - K(x \mid S)$ [6]. The larger the randomness deficiency is the more regular and more probable the string x is. Kolmogorov complexity and randomness are conceptually related through the minimum description length. As an example, transduction (see Section 9), for the purpose of categorization (see Section 10), would choose from all possible labeling ("identities") for (unlabeled) test data the one that yields the largest randomness deficiency, *i.e.*, the most probable labeling. Randomness deficiency is, however, not computable [6]. One has to approximate it instead, using a slightly modified Martin—Löf test for randomness, and the values taken by such randomness tests are called p-values. The p-value construction used (see Section 8) has been proposed by Vovk *et al.* [9] and Predrou *et al*. [27]. This is discussed next.

8. Strangeness and P-Values

Strangeness and typicality [28] [29] are interchangeable. Given a sequence of similarity distances from sample (instance) j to other samples l, the strangeness α_j (or alternatively the typicality) of j with putative label y is defined as:

$$\alpha_j = \frac{\sum_{l=1}^{k} d_{jl}^{y}}{\sum_{l=1}^{k} d_{jl}^{\neg y}}$$

The strangeness measures the lack of typicality with respect to its true or putative (assumed) identity label y and the labels for all the other exemplar patterns. Formally, the strangeness α_j is the (likelihood) ratio (LR) of the sum of the k nearest neighbor (k-*NN*) similarity (Euclidean) distances d for sample j from the same class y divided by the sum of the k nearest neighbor similarity distances for sample j from all the other classes ($\neg y$). Note that other similarity distances, e.g., Mahalanobis, could readily and effectively substitute for the Euclidean distance. The smaller the strangeness, the larger its typicality and the more probable its (putative) label y is. The strangeness facilitates both feature selection (similar to Markov blankets) and variable selection (dimensionality reduction) for recognition purposes. An alternative (cohort) definition for strangeness would tabulate distances only for the most prevalent class surrounding y given the putative assignment $\neg y$. One finds empirically that the strangeness, classification margin, sample and hypothesis margin, posteriors, and odds are all related via a monotonically non-decreasing function with a small strangeness amounting to a large margin. The margin of a hypothesis with respect to an instance is the distance between the hypothesis and the closest hypothesis that assigns an alternative label to that instance. An alternative definition [30] for the hypothesis margin, similar to the strangeness definition proposed above, is

$$\phi(x) = \left(\left\| x - nearmiss(x) \right\| - \left\| x - nearhit(x) \right\| \right)$$

with nearhit (x) and nearmiss (x) being the nearest samples of x that carry the same and a different label, respectively. The strangeness can also be defined using the Lagrange multipliers associated with (kernel) SVM classifiers but this requires a significant increase in computation.

The use of the strangeness gets further support from the Cover-Hart theorem [31], which proves that asymptotically the generalization error for the nearest neighbor classifier exceeds by at most twice the generalization error of the Bayes optimal classification rule. The same theorem also shows that the k-*NN* error approaches the Bayes error (with factor 1) if $k = O(\log n)$. The optimal piecewise linear discrimination boundary includes those samples for whom the strangeness α is constant, *i.e.*, $\alpha = 1$ for the case of two class ("binary") discrimination. The advantage for the strangeness k-*NN* against standard lazy k-*NN* classification is most apparent for overlapping distributions. One can empirically find that the boundaries induced by the strangeness k-*NN* are smoother and closer to the optimal boundary compared to the boundary induced by k-*NN*, while the corresponding errors and their standard deviations are also lower.

Additional relations that link the strangeness and the Bayesian approach using the likelihood ratio can be observed, e.g., the logit of the probability is the logarithm of the odds, $\text{logit}(p) = \log(p/(1-p))$, the difference between the logits of two probabilities is the logarithm of the odds ratio, *i.e.*,

$$\log \frac{p/(1-p)}{q/(1-q)} = \text{logit}(p) - \text{logit}(q)$$

(see also logistic regression and the Kullback-Leibler (KL) divergence). Note that the logit function is the inverse of the "sigmoid" or "logistic" function. Another relevant observation that buttresses the use of the strangeness comes from the fact that unbiased learning of Bayes classifiers is impractical due to the large number of parameters that have to be estimated. The alternative to the unbiased Bayes classifier is logistic regression, which implements the equivalent of a discriminative classifier.

The p-values described next estimate the randomness deficiency (see Section 7), and compare ("rank") the stran-geness values to determine the credibility and confidence in the putative classifications ("labeling") y made. The p-values bear resemblance to their counterparts from

statistics but are not the same [32]. The p-values are determined according to the relative rankings of putative authentications made against each one of the identity classes known. The standard p-value construction shown below, where l is the cardinality of the training set T, constitutes a valid randomness (deficiency) test approximation [33] for some putative label y hypothesis assigned to an unlabeled new sample

$$p_y(e) = \frac{\#\left(j : \alpha_j \geq \alpha_{new}^{y}\right)}{(l+1)}$$

The p-values are used to assess the extent to which data supports or discredits the null hypothesis H0 (for some specific classification attempt). When the null hypothesis is rejected for each of the classes known, one declares that the test ("query") pattern is "unfamiliar" as it fails to "mate" against all the known classes. The query is thus answered with "none of the above." This corresponds in the case of biometrics to forensic exclusion with rejection, and it is characteristic of open set recognition. This is different from closed set recognition where the top choice is the default answer.

9. Transduction

Transduction is different from inductive inference. It is local inference ("estimation") that moves from particular(s) to particular(s). In contrast to inductive inference, where one uses empirical data to approximate a functional dependency (the inductive step [that moves from particular to general] and then uses the dependency learned to evaluate the values of the function at points of interest (the deductive step [that moves from general to particular]), one now infers directly (using transduction) the values of the function only at points of interest [8]. Inference now takes place using both labeled and unlabeled data, which play complementary roles to each other. Transduction incorporates unlabeled data, characteristic of test ("unlabeled") samples, in the decision-making process responsible for their labeling ("prediction"), while seeking for a consistent and stable labeling across both (near-by) training ("labeled data") and test ("unlabeled") data [34]. Transduction "works because the test set provides a nontrivial factorization of the [discrimination] function class" [10].

One key concept behind transduction (and consistency) is the symmetrization lemma [8], which replaces the true (inference) risk by an estimate computed on an independent set of data, e.g., unlabeled/test data, referred to as 'virtual' or 'ghost samples'. We expand on the concept of ghost samples to include guided perturbations and hints in order to broaden the pool of data available for learning and improve on generalization (see Section 15). Revision of the putative class ("label") assignments drives the process responsible with determining ("inferring") the appropriate rejection threshold for multi-layer categorization, in general, and open set recognition, in particular (see Section 10).Transduction seeks consistent labels for both training and test data. It iterates to relabel the test data using local perturbations on the labels already assigned. Poggio *et al*. [35], using similar reasoning, suggest it is the stability of the learning process that leads to good predictions. In particular, the stability property says that "when the training set is perturbed by deleting one example, the learned hypothesis does not change much. This stability property stipulates conditions on the learning map rather than on the hypothesis space." Integral to revision, the concept of stability is expanded to allow for both label changes and training samples deletion (see Section 15). This learning approach can be further analyzed using regularization [36].

The goal for inductive learning is to generalize for any future test set, while the goal for transductive inference is to make predictions for some specific/given working set W. Test data is not merely a passive collection of data waiting for labeling but rather an active player ready to add its own contribution to that provided by the labeled training data. The working or test samples provide additional information about the underlying data distribution and their explicit inclusion in the problem formulation yields better generalization on problems with insufficient labeled points [9]. Transductive inference therefore seeks to find, from all possible labeling L (classifications) L (W) for working (unlabeled) test data set W, the one that yields the largest randomness deficiency, *i.e.*, the most probable labeling. This choice models the working ("test") sample set W in a fashion similar to that used for the training set T. Towards that end, transduction has to minimally change the original model learned for T. "The difference between the classifications that induction and transduction yield for some working sample approximates its randomness deficiency. As one disturbs a classifier (driven by T) using working but putatively labeled samples to augment the training set, the magnitude of the disturbance estimates the classifier's instability (unreliability) in a given region of its problem space. Similar and complementary to transduction is semi-supervised learning (SSL) [10] (see Section 14).

10. Multi-Layer and Multi-Set Open Set Data Categorization

Multi-layer open-set categorization starts with determining the context ("frame problem"), continues with detection ("familiarity"), proceeds with open set recognition ("classification"), and ends up with stratification

("bin-inng"). Open set recognition operates under the assumption that some of the unlabeled patterns asking for recognition are unknown (unfamiliar) and can't be recognized. This is addressed by making available a reject ("unknown") answer. Given an unlabeled (pattern) exemplar *e*, the corresponding p-values, for each putative label (class) drawn from training data, record the likelihood that the new pattern belongs to that putative class. If some p-value p is high enough and it significantly outscores the others, the new pattern can be mated to the corresponding class with credibility p. If the top ranked (highest p-values) choices are very close to each other but outscore the other choices, the top ranked label choice is credible but ambiguous in its classification, and should thus carry a low confidence during further processing. The confidence measures the difference between the first and second largest (or consecutive) p-values. If all p-values are randomly distributed and no p-value significantly outscores the other p-values, any choice of labels is questionable and the new exemplar can't be recognized and should be thus rejected. The credibility and confidence indexes/similarity scores are useful for data fusion purposes (see Section 13).

The implementation details regarding open set decision-making are generic as they apply to each categorization layer. One re-labels the training patterns, one at a time, with all possible putative labels except the ground truth originally assigned to it. As an example, detection involves just two labels. The peak-to-side (PSR) ratio, $PSR = (p_{max} - p_{min})/p_{stdev}$, traces the characteristics of the resulting p-value distribution and determines, using cross validation, the [a priori] threshold needed for open-set recognition. The PSR values found for unknown patterns are low because they do not mate; their relative strangeness is high, and their p-values low. As an example, imagine unknown (unlabeled) exemplars as impostors when recognition concerns biometrics. Impostors should be deemed as outliers and be thus rejected as unknown ("none of the above") vis-à-vis the labels (identities) known (enrolled) [37]. The open set recognition approach just described provides for both reliability and robustness. Reliability in terns of consistency and stability of learning, while robustness vis-à-vis incomplete and corrupt information. The same approach renders itself suitable for both anomaly and/or outlier detection.

Training data consists of (x, y) patterns, where x stands for representation, and y stands for the pattern label (class). The similarity scores discussed so far assumed that each class ("set") consists of only one pattern. This assumption does not hold in the general case when multiple instances can account for variability, increase the signal-to-noise ratio, and provide for better generalization. Towards that end, the Hausdorff distance substitutes for the Euclidean distance d used to define the strangeness (see Section 8) leading to the multi-set distance d_H for sets A and B

$$d_u(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right\}$$

There are other ways to define multi-set similarity scores. As an example, similarity can be defined by the angle between the subspace S_B of unlabeled patterns (set B) and the subspace S_A of reference (labeled) training pattern (set A). The subspaces S_A and S_B are estimated (learned) possibly as eigenspaces. Alternatively, one can learn the subspaces on Grassman manifolds and employ principal angles, Binet-Cauchy and Procrustes metrics for similarity distances [38]. The assumption still holds that both sets A and B are multi-sets consisting of instances coming from only one class. We relax this assumption (see Section 14) when we allow multi-sets A and B whose instances can come from more than one class. The scope for categorization expands using boosting and transduction for data aggregation (see Section 13) and both qualitative and quantitative similarity scores (see Section 18).

Active learning (see Section 3) is context-driven and takes advantage of statistical learning, in general, and transduction and semi-supervised learning, in particular, to provide for additional event-recognition functionalities. Active learning employs explore and exploit actions characteristic of closed-loop control for evidence accumulation in order to revise its prediction models and to reduce uncertainty. Continuous interactions with the environment provide the feedback necessary to iterate on EBM and DKA. Towards that end, algorithmic information theory, algorithmic randomness, and closed-loop control collaborate to advance intelligent data collection, knowledge ("model") discovery ("extraction"), and purposeful behavior ("action"). This is discussed over the next sections (see Sections 11 - 16). It is helpful to note that, according to Maturana, "learning is not [merely] a process of accumulation of representations of the environment; it is a continuous process of transformation of behavior through continuous change in the capacity of the nervous system to synthesize [and incorporate] it' [22].

The basic functionalities active learning supports include decisions on where to explore and what (labeled and unlabeled) data to gainfully employ for further adaptation and modeling purposes (see Section 11); on handling time-varying data streams and detecting change (using martingale) (see Section 12) for choosing when to revise existing recognition models; on how to process and fuse information and knowledge (see Section 13).

The multi-set based similarity scores proposed are all driven by algorithmic randomness and Kolmogorov complexity. Additional similarity scores are described in order to expand on strangeness/typicality and p-values, while taking advantage of associations, context and relationships for the purpose of link analysis and selection (see Section 14). Metrics are also made available for data cleaning and data revision, e.g., label editing (to account for annotation errors) and censoring and imputation (to fill in for missing data), on one side, and appearance perturbations and synthesis (for consistency and stability) (see Section 15). Finally, we address yet another functionality, that of criticality identification and prediction, which belongs to complex system identification and prediction (see Section 16). The latter functionality is an example where the interplay between similarity scores (see Section 8) and change detection (see Section 12) shows how more complex behavior can emerge.

11. Data Collection and Evidence Accumulation

Active learning is concerned with evidence accumulation ("data sampling and collection") towards choosing the most (functionally) relevant examples to improve the classification ("margin") for both effectiveness ("accuracy") and efficiency ("number of examples needed"). The active learning solution proposed here is driven by transduction and it is realized using strangeness and p-values [32]. The p-values provide a measure of diversity and disagreement in opinion regarding the true label of an unlabeled example when it is assigned all possible putative labels. Let p_i be the p-values obtained for a particular example x_{n+1} using all possible labels $i = 1, \cdots, M$. Sort the sequence of p-values in descending order such that the first two p-values, say, p_j and p_k are the two highest p-values found with corresponding labels j and k, respectively. The label assigned to the unknown example is therefore j while its p-value is p_j. This value defines the credibility of the classification. If the credibility for p_j is not high enough (using a priori thresholds found using cross-validation) the prediction is rejected. The difference between the two p-values can be further used as a confidence value, if any, of the prediction. Note that, the smaller the confidence, the larger the ambiguity regarding the proposed label.

One considers now three possible cases of p-values, p_j and p_k, assuming $p_j > p_k$: Case 1: p_j is "high" and p_k is "low." Prediction "j" has high credibility and high-confidence value; Case 2: Both p_j and p_k are "high." Prediction "j" has high credibility but low-confidence value; and Case 3: both p_j and p_k are "low." Prediction "j" has low credibility and low-confidence. High uncertainty in prediction occurs for both Case 2 and Case 3. Note that uncertainty of prediction occurs when $p_j \approx p_k$. Define as information "closeness" the quantity $I(x_{n+1}) = p_j - p_k$ to indicate the quality/relevance of the information contents possessed by the example x_{n+1}. As $I(x_{n+1})$ approaches 0, the more uncertain we are about classifying the example x_{n+1}, and the larger the information gain when one is told about its label. Active learning will add this example, with its (true) label, to the training set because it provides new information about the learning map for classification. Evidence accumulation corresponds to on-line learning. It allows for unlabeled samples to progressively augment the training set T using for their assigned labels the classifications made by some classifier C trained on T. One should be aware that the annotations made on each iteration can be in error and that their propagation could contribute to future annotation errors. The explanation comes from the obvious fact that the classifiers C are updated using the annotations made with some of them being in error. Performance evaluation now includes the temporal dimension; rather than using one-shot cross validation, learning curves need to be graphed over time. Learning consistency/stability implies that after some initialization time the error rate for the learning curve, error(t), is relatively small, its slope is also relatively small, and the overall appearance is that of a smooth curve [9]. The annotations made can be edited as part of some revision process to ensure better consistency and stability for learning (see Section 15).

12. Change Detection Using Martingale

Change detection addresses yet another basic functionality for active learning, that of identifying those time instances when the underlying distribution for time-varying data streams undergoes change or drift. The approach sketched below employs similarity scores driven by strangeness and p-value, and it makes use of martingale [39]. Assume time-varying multi-dimensional data (stream) matrix $R = \{ R(j) = x_j \}$ where $R(j)$ are "columns" and stand for time-varying (data stream) pattern vectors. Assume that seeding provides some initial $R(j)$ with $j = 1, \cdots, 10$. K-means clustering finds (in an iterative fashion) center "prototypes" $Q(k)$ for the data stream (seen so far). Define the strangeness α corresponding to $R(j)$ using the cluster model (with $R = \{x_j\}$ as summary for the data stream and c standing for cluster center) and the Euclidean distance $d(j)$ between $R(j)$ and $Q(k)$ for $j >= 10$ and $k = j - 9$ as

$$s\left(R, x_j\right) = \left\| x_j - c \right\|$$

Define then p-values as

$$p_i\left(\left\{\left(x_1, y_1\right), \cdots \left(x_i, y_i\right), \theta_i\right\}\right) = \frac{\#\left\{j : s_j > s_i\right\} + \theta_i \#\left\{j : s_j = s_i\right\}}{i}$$

where α_j is the strangeness measure for (x_j, y_j), $j = 1, 2, \cdots, I$ and θ_i is randomly chosen from [0, 1] at instance i. Define a family of martingale starting with $M^{\varepsilon}(0) = 1$ and continuing with $M^{\varepsilon}(j)$ indexed by ε in [0, 1]

$$M^{\varepsilon}(j) = \prod_{i=1}^{j} \left\{\varepsilon \left(p_i\right)^{\varepsilon - 1}\right\}$$

Similar to the Neyman-Pearson test, the martingale test

$$0 < M^{\varepsilon}(j) < \lambda$$

rejects the null hypothesis H0 "no change in the data stream" for H1 ("change detected in data stream") when $M^{\varepsilon}(j) > = \lambda$ with the value for λ (empirically chosen to be greater than 2) determined by the false accept rate (FA) one is ready to accept, *i.e.*, $1/\lambda$ = FAR. An alternative (parametric) test, e.g., SPRT, will employ the likelihood ratio (LR) with $B < \text{LR} < A$ and decide for H0 as soon as $\text{LR}(j) < B$, decide for the alternative H1 ("change") when $\text{LR}(j) > A$, with $B \approx \beta\ (1 - \alpha)$ and $A \approx (1 - \beta)/\alpha$ using α for the test significance ("size") and $(1 - \beta)$ for the test power. The changes ("spikes") found correspond to transition states.

The martingale method is incremental and single - pass, does not require a sliding window on the data stream, does not require monitoring the explicit performance of the classification or clustering model as data points are streaming, and it works well for high-dimensional data streams. Furthermore, the change detection method is non-parametric and it works on both labeled and unlabeled data. The proposed method has a theoretical false positive error bound given a specific threshold, and the delay time between the true change point and the detected change point can be approximated.

13. Data Aggregation

DKA goes beyond data collection to include data aggregation. The knowledge component and the actions pursued come from building up the knowledge blocks using data aggregation. The Gestalt whole, *i.e.*, the knowledge, is more than the sum of its parts, which aggregates as a result of data collection. Data aggregation is widely referred to as data fusion. It covers for multi-level and multi-layer fusion (in terms of functionality and granularity), voting methods, mixture of experts, ensemble methods, and/or gating networks. As an example, the visual cortex in charge of human vision, is nothing more than a feed-forward architecture (supplemented by appropriate feedback) where data aggregation takes for its raw input fine retinal representations and sequentially fuses them into more elaborate but coarse representations suitable for recognition and action. Successive field of views (FOV) get larger with categorical Gestalts emerging to trigger recognition and action. Data aggregation employs the same similarity scores we have used so far, *i.e.*, the strangeness and p-values.

One method for data aggregation is based on (spatial-temporal) data segmentation using similarity and clustering. Strangeness-based (unsupervised K-Means) clustering has two intuitive assumptions. First, well-separated groups of data should aggregate as different clusters. The "well-separated" or "purity" assumption says that the samples (from those groups) have strangeness value less than some threshold γ with $\gamma < 1$. This defines the minimal margin for groups that can be characterized by different labels. The second assumption says that when different groups of samples are not (well) separated but instead represent different clusters, there should be a minimal number of samples whose strangeness is greater than γ. Those samples are close to the separating boundary.

Data aggregation can also employ boosting and transduction for decision-making, in a fashion similar to cascade recognition [40]. Boosting amounts to multi-level fusion when it involves feature/parts, score ("match"), and detection ("decision"), or alternatively amounts to multi-layer fusion when it involves modality, information quality, and method (algorithm) used. The components are iteratively realized as weak learners (see below) whose relative performance are driven by strangeness and p-value (see Section 8), transduction (see Section 9) and open set recognition using transduction (see Section 10). Strong inference, characteristic of boosting, takes advantage of both localization and specialization to combine expertise. The substance of boosting is sketched next.

Logistic regression amounts to a sigmoid function that directly estimates the parameters of $P(y \mid x)$, e.g., $P\{y = 1 \mid x\}$ for the case when y is Boolean. Logistic regression supports discriminative methods and likelihood ratios, e.g., find y as $y = 1$ if $P\{y = 1 \mid x\}/P\{y = 0 \mid \text{x}\} > 1$ (see Section 5). Logistic regression is approximated by Support Vector Machines (SVM). Note that AdaBoost [41] minimizes (using greedy optimization) some functional whose minimum yields the equivalent of logistic regression [42], while an ensemble of SVM is functionally similar to AdaBoost [43]. The basic assumption behind boosting is that "weak" learners can be combined to learn any target concept with probability close to 1. Weak learners, usually built around features easy to compute, can classify at better than chance (with probability $1/2 + \eta$ for $\eta > 0$). AdaBoost works by adaptively

and iteratively re-sampling the data to focus its learning on samples that the previous weak (learner) classifier could not master, with the relative weights of misclassified samples increased ("refocused") after each iteration. AdaBoost involves choosing v effective components h_v to serve as weak (learners) classifiers and using them to construct separating hyper-planes. The mixture of experts or final boosted (stump) strong classifier H is

$$H(x) = \sum_{v=1}^{v} \gamma_v h_v(x) > \frac{1}{2} \sum_{v=1}^{v} \alpha_v$$

with r denoting the reliability or strength of the weak learner. The constant 1/2 comes in because the boundary is located mid-point between labels 0 and 1. If the negative and positive examples are labeled as –1 and +1 the constant used is 0 rather than 1/2. The goal for AdaBoost is margin optimization with the margin viewed as a measure of confidence or predictive ability. The weights associated with data samples are related to their location relative to the margin and affect AdaBoost's generalization ability. AdaBoost minimizes (using greedy optimization) a risk functional whose minimum defines logistic regression. AdaBoost converges to the posterior distribution of y conditioned on x, and the strong but greedy classifier H in the limit becomes the log-likelihood ratio test.

The multi-class extensions for AdaBoost are AdaBoost. M1 and .M2, the latter one used to learn strong classifiers with the focus now on both difficult samples to recognize and labels hard to discriminate. The possible use of appearance features for weak learners is justified by their apparent simplicity. One drawback for AdaBoost. M1 comes from the expectation that the performance for the weak learners selected is better than chance. When the number of classes is $k > 2$, the condition on error is, however, hard to be met in practice. The expected error for random guessing is $1 - 1/k$; for $k = 2$ the weak learners need to be just slightly better than chance. AdaBoost.M2 addresses this problem by allowing the weak learner to generate instead a set of labels together with their plausibility (not probability), *i.e.*, $[0, 1]^k$. AdaBoost.M2 focuses on the incorrect labels that are hard to discriminate. Towards that end, AdaBoost.M2 introduces a pseudo-loss e_v for hypotheses h_v such that for a given distribution D_v one seeks $h_v : x \times y\ [0, 1]$ that is better than chance. "The pseudo-loss is computed with respect to a distribution over the set of all pairs of examples and incorrect labels. By manipulating this distribution, the boosting algorithm can focus the weak learner not only on hard-to-classify examples, but more specifically, on the incorrect labels y that are hardest to discriminate" [41].

The strangeness is the thread to implement both (1) representation and (2) decision-making, the latter using boosting (learning, inference, and prediction for the purpose of classification). The strangeness, which implements the interface between the components describing the representation, e.g., attributes and/or components/parts, and boosting, combines the merits of filter and wrapper classification methods. The coefficients and thresholds for the weak learners, including the thresholds needed for open set recognition and rejection are learned using validation patterns [44]. The best feature correspondence for each component is sought between validation and training patterns over existing components. The strangeness of the best component found during training is computed for each validation pattern under all its putative class labels c ($c = 1, \cdots, C$). Assuming M validation pattern from each class, one derives M positive strangeness values for each class c, and M (C - 1) negative strangeness values. The positive and negative strangeness values correspond to the case when the putative label of the validation and training pattern are the same or not, respectively. The strangeness values are ranked for all the components available, and the best weak learner h_v is the one that maximizes the recognition rate over the whole set of validation patterns V for some component i and threshold θ_i. Boosting is similar to cascade classification as on each iteration a weak learner component is chosen.

The level of significance α (not to be confused with strangeness notation α) determines the scope for the null hypothesis H0. Different but specific alternatives can be used to minimize Type II error or equivalently to maximize the power $(1 - \beta)$ of the weak learner. During cascade learning each weak learner ("classifier") is trained to achieve some (minimum acceptable) hit rate $h = (1 - \beta)$ and (maximum acceptable) false alarm rate α. Upon completion, boosting yields the strong classifier $H(x)$, which is a collection of discriminative components playing the role of weak learners. The hit rate after V iterations is h^V, while the false alarm is α^V.

14. Data Selection and Link Analysis

Link analysis concerns data fragmentation, e.g., time-varying surveillance of data streams. The goal is to aggregate separate pieces of information back into one coherent and possibly identifiable pattern. For illustration purposes consider screening biometrics for security purposes where the goal is to select and track faces, and possibly identify them. Assume first that mug shots and/or multiple still image sets and/or video sequences for known subjects become available during biometric enrollment for mass screening. Data streams shaped as

time-varying video sequence(s) of crowds and consisting of both foreground faces and background structured noise are then captured during surveillance. The goal is to identify the subset of CCTV frames, if any, where wanted subjects show up or alternatively to locate subjects (of unknown identity) across the video whose behavior is suspicious. Subjects can appear and disappear as time progresses and the presence of any face is not necessarily continuous across (video) frames. Faces belonging to different subjects appear in a sporadic fashion across the video sequence. Some of the CCTV frames could actually be void of any face, while other frames could include occluded or disguised faces from different subjects. Kernel K-means and/or spectral clustering [45] using biometric image patches, parts, and multi-sets similarity scores driven by strangeness and p-values for typicality and ranking, are suitable for link analysis, face selection, and tracking.

Spectral clustering [46] is a recent methodology for data segmentation and data aggregation/clustering. The inspiration for spectral clustering comes from graph theory (minimum spanning trees (MST) and normalized cuts) and the spectral (eigen decomposition) of the adjacency/proximity ("similarity") matrix and its subsequent projection to a lower dimensional space, which describes in a succinct fashion the graph induced by the set of data samples ("patterns"). Minimizing the "cut" (over the set of edges connecting K clusters) yields "pure" (homogeneous) clusters. Similarity is again defined using strangeness/typicality, p-values, and rankings.

Similar to decision trees, where information gain is replaced by gain ratio to prevent spurious fragmentation, one substitutes the "normalized cut" (that minimizes the cut while keeping the size of the clusters large) for "cut." To minimize the normal cut (for $K = 2$) is equivalent to minimize the Raleigh quotient of the normalized graph Laplace matrix L^* where $L^* = D^{-1/2}LD^{-1/2}$ with $L = D - W$; W is the proximity ("similarity") matrix and the (diagonal) degree matrix D is the "index" matrix that measures the "significance" for each node. The Raleigh quotient (for $K = 2$) is minimized for the eigenvector z corresponding to the second smallest eigenvalue of L^*. Given n data samples and the number of clusters expected K, spectral clustering (for $K > 2$) employs the Raleigh-Ritz theorem and describes among others algorithms such as Ng, Jordan, and Weiss [47] where one (i) computes W, D, L, and L^*; (ii) derives the largest K eigenvectors z_i of L^*; (iii) forms the matrix $U \varepsilon R^{n x k}$ by normalizing the row sums of z_i to have norm 1; and (iv) cluster the samples x_i corresponding to z_i using K-means.

An expanded framework that integrates graph-based semi-supervised learning [48] and spectral clustering for the purpose of iterative grouping and classification, *i.e.*, label propagation, can be developed. One takes advantage of both labeled and mostly unlabeled (biometric) patterns. The graphs reflect domain knowledge characteristics over nodes (and sets of nodes) to define their proximity ("similarity") across links ("edges"). The solution proposed is built around label propagation and relaxation. The graph and the corresponding Laplacian, weight, and diagonal matrices L, W, and D are defined over both labeled and unlabeled (biometric) patterns. The harmonic function solution [48] finds (and iterates) on the (cluster) assignment for the unlabeled biometric patterns Y_u as $Y = -(L_{uu})^{-1}L_{ul}Y_l$ with L_{uu} the sub-matrix of L on unlabeled nodes and Y_l the group indicator over the labeled nodes. Each row of Y_u reports on the posteriors for the Cartesian product between K clusters and n biometric samples. Class proportions for the labeled patterns can be estimated and used to scale the posteriors for the unlabeled biometric patterns. The harmonic solution is in sync with a random (gradient) walk on the graph that makes predictions on the unlabeled (biometric) patterns according to the weighted average of their labeled neighbors.

15. Data Cleaning and Data Revision

DKA is all encompassing. The derivation of knowledge from data amounts to searching for regularities, on one side, and model construction for prediction purposes, on the other side. This is usually referred to as data mining. Data is not only subject to collection (using active learning) but it is also subject to aggregation, on one side, and cleaning and revision, on the other side. Data cleaning also referred to as data cleansing and/or preparation, concerns itself with data quality. It plays a major role in supporting data mining, in general, and exploration and exploitation bound closed-loop control, in particular. Traditionally, "data cleaning has taken a backseat to the more alluring question of how best to extract meaningful knowledge." Dorian Pyle is very convincing in dispelling such misconceptions and strongly supports data cleaning [49].

Data cleaning addresses issues related to measurement and data collection. This includes errors, e.g., noise and artifacts; outliers; missing values; duplicate data; and precision, bias, and accuracy. Data cleaning can be addressed using statistical methods, e.g., robust statistics [50]. One can expand on data cleaning and entertain data revision, which is driven by considerations regarding learning consistency and stability. One can further expand on the revision process itself using the lawful synthesis of additional patterns to enhance modeling, while alternating between data analysis and synthesis. This is discussed next.

Learning is first and foremost about generalization. Towards that end, one is interested to control to what extent learning behavior is consistent and stable. Consistency is about the asymptotic analysis regarding the rate of convergence, while stability is about sensitivity to changes, e.g., the deviations in learning performance experienced when data collection is subject to perturbations. For the specific case of transduction this takes place using the apparent complementary between training and test patterns. Both consistency and stability are measured over time, as it is the case for on-line learning. This allows for annotation errors to accrue and for robustness (vis-à-vis incomplete and corrupt information) to play out. Errors and/or perturbations can naturally occur or be artificially induced. Data revision is risk driven and includes editing, e.g., insertion, deletion, substitution, and transposition; perturbations and synthesis; and censoring and imputation to fill in for missing information.

Perturbation and synthesis are data driven, take advantage of existing models, and serve as constraints and hints for better discrimination and generalization. The constraints are usually about distances, while the hints involve both appearance and class membership ("label"). This is characteristic of semi-supervised learning and is sometimes referred to as "hallucinations." Hints and/or assumptions that underlie semi-supervised learning include (a) smoothness, *i.e.*, if two sample patterns are close so should be their outputs; (b1) cluster, *i.e.*, samples in same cluster are likely to share the same label; or alternatively (b2) low-density, *i.e.*, the decision boundary should lie in a low-density region; and (c) manifold, *i.e.*, the high dimensional data lie (roughly) on a low-dimensional manifold [10]. Yet another example for hints, *i.e.*, prior semi-supervised learning knowledge, comes from binary inference using the Universum [51]. "Unlabeled examples known not to belong to either class but belonging to the same pattern domain implicitly specify a prior distribution. Supplying such examples, rather than defining explicitly the underlying distribution, can be a far easier task." The disturbances created regarding the classifications made when adding new data, which takes the form of hints, virtual examples, noise injection, expands on data collection and facilitates sensitivity analysis (see below). It leverages priors to provide an alternative capacity concept (using the VC dimension) to the large margin approach embedded in structural risk minimization [52].

Censoring often consists of conducting a test on an item (under specified conditions) to determine the time it takes for a "failure" to occur, *i.e.*, time-to-event outcomes. Censored data occurs (a) when the value of an observation is only partially known; and/or (b) when a value of interest occurs outside the range of a measuring instrument. Examples for censoring include clinical trials related to survival rates, disease progression, and times to recovery. "Systematic reviews of published time-to-event outcomes commonly relay on calculating odds-ratios (OR) at fixed points in time and where actual numbers at risk are not present" [53]. Meta-analysis using strangeness driven likelihood ratios can estimate OR and, similar to risk analysis, can estimate proposed hazard ratio (HR) as well. The problem of censored data, in which the observed value of some variable is partially known, is closely related to the problem of missing data, where the value of some variable is unknown (see below the case for imputation).

Imputation is the substitution of some value for missing pattern and/or missing component ("feature"). This becomes relevant for clinical / longitudinal studies due to random drops-out and withdrawals caused by lack of efficacy, with the latter responsible for bias. The SSL smoothness and cluster assumptions suggest using cluster membership and cluster prototypes for making suitable substitutions. Preeminent among clustering methods are self-organizing map (SOM [54]) and hybrid (unsupervised-supervised) label-vector-quantization (LVQ) [54]. Alternatively, one can use multiple imputations (possibly augmented by jackknife stratification including variance estimation) and combine their outcomes to produce unbiased estimates. Throughout the strangeness and p-values estimate similarity and ranking among alternative substitutions.

16. Criticality Identification and Prediction

We consider here criticality identification and prediction, which belongs to complex system identification and prediction. This functionality is an example where the interplay between similarity scores (see Section 8) and change detection (see Section 12) shows how complex behavior driven by algorithmic randomness can emerge. Towards that end, we detail a novel application for identifying critical states characteristic of a complex physical system that is nonlinear and heterogeneous and involves a high degree of freedom [55]. This is an important subject of study in many areas of natural and social sciences. The prediction of critical states of a complex system is a useful but challenging issue, in particular the occurrence of a catastrophic event such as tropical cyclones in the Earth atmosphere. In principle, a physical system should strictly follow physical laws that have been proven to be correct and universal. Indeed, many physical systems can be modeled by these laws, e.g., planetary motions around the Sun, the precise working of mechanic engines, etc. However, many other systems are not. Examples of complex physical systems and their critical states include

solar flares in the solar atmosphere, cyclones in the terrestrial atmosphere, and earthquakes in the geological system.

A novel architecture and methodology for predicting critical states for complex physical systems is proposed here (see **Figure 1**). The conceptual architecture, main

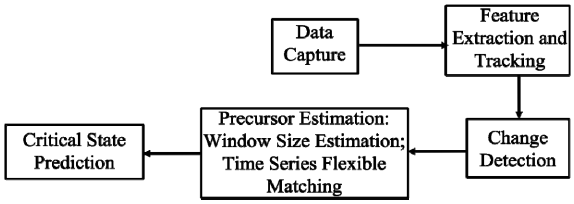

Figure 1. An IIM Framework for Predicting Critical States in Complex Physical Systems.

processing stages, and flow of control are built around novel methods for modeling and prediction using statistical learning, in general, and semi-supervised learning and transduction, in particular. The architecture proposed approaches complex physical systems as a time evolving system; it parses and translates data streams into knowledge regarding the emergence of critical states. Towards that end, the specific methodology proposed here takes advantage of better sensor technology to employ advanced computational methods for both data representation and prediction of critical states. While the methodology proposed to map data into knowledge is generic, we apply the methods for predicting the genesis and intensification of cyclones.

The main processing stages are as follows:

1) Data capture: Archived Data and Near Real-Time (NRT) data are available for weather forecasting and modeling. Archived data are used for system training and validation, while NRT data are used for testing and operational purpose. These data are the inputs to the processing pipeline outlined in **Figure. 1**. To ensure rapid data availability, the technique for NRT satellite data processing is different from science data product processing. While NRT satellite data and related science products are not identical, the NRT data should be accurate enough for operational use. The data needed for predicting cyclones are available from various NASA data archives.

2) Feature extraction and tracking: It produces the necessary high-dimensional feature vector and its time variation in order to find the change in the state of the physical system of interest. The feature extraction approach is either an image data processing approach or a dimensionality reduction approach, or a combination of both. Based on domain knowledge and predictive power, one decides on the features to use. Tracking and monitoring of physical systems can be achieved using a Kalman Filter approach or a particle filter approach. These approaches have been shown to work robustly for tropical cyclone tracking using multiple satellite observations [56,60].

3) Change detection using martingale: See Section 12.

4) Precursor estimation: The input from the change detection stage provides the temporal extent of significant events with only some of them preceding critical states. The events referred to as precursors are subject to classification using time-series flexible matching. The two step process is discussed next.

a) Window size estimation: The online martingale change detection approach is used to identify a data subsequence (time series) that behaves anomalously. First, two threshold values, θ_L and θ_R, are selected such that $\theta_L < \theta_R$. As the system is monitoring the data (vector) time-series sequence $t(k)$, $k = 1, 2, \cdots$, martingale values M(i) at time instance $i = 1, 2, \cdots$, are computed. When the first $M(i) > \theta_L$ is detected, one seeks for the next time instance j when $M(j) > \theta_R$. The window size for this time series is $j - i + 1$, and the time series "window" $t = [t(i), t(i+1), \cdots, t(j)]$ is extracted and recorded as $[T_1, T_2]$. Note that the window is merely a candidate to serve as a PRECURSOR for a future critical state. Hence, the θ_L and θ_R values are empirically selected using Doob's inequality [39] at a lower value, maybe less than 10, e.g. $\theta_L = 3$ and $\theta_R = 4$, which yields a FAR of 25% to 33 1/3% (considered high). $\theta_L < \theta_R$ ensures that a change is more likely to take place at the more recent time instance T_2. This leads to a collection of training sets where the sets differ by how far in time T_2 is away from the true critical state. As long as the martingale at $(T_2 + n)$ is greater than θ_L, the window $[T_1, T_2]$ is widen to $[T_1, T_2 + n]$ to augment the set of potential precursors. The same process iterates until the martingale $M(T_2 + n)$ becomes less than θ_L. The next candidate window starts when the martingale M is again greater than θ_L. The candidate precursors are subject to time-series flexible matching to yield the similarity scores required for predicting critical states.

b) Time-series flexible matching and classification: In the time series classification problem, one assigns a label to an unlabeled time series based on training examples. The main research task for this problem involves similarity measures. Many similarity measures for data sequences have been proposed [57]. The two main categories of similarity measures are the L_p-norm based similarity measures and the elastic similarity measures. The L_p similarity measures are metric, but they assume fixed length data sequences and do not support local time shifting; the elastic similarity can be used to compare arbitrary length data sequences and support local time shifting but they are not metric. The classical elastic measure to overcome the weakness of L_p norms is the Dynamic Time Warping (DTW) [58]. The Longest

Common Subsequence (LCSS) elastic measure was proposed to handle two- and three-dimensional arbitrary length data sequences. The LCSS is robust to noise and gives more weight to the similar portion of the sequences [59]. **Figure. 2** shows a comparison of two hurricane intensity time series using LCSS similarity measure. The top graph shows the minimum bounding envelope for the two time series and the bottom graph shows the corresponding points between the two time series. We use LCSS as in [60].

5) Online confident prediction and point estimation of critical states: The inputs are the time-series window $[T_1, T_2]$ and its potential successors $[T_1, T_2 + n]$ extracted in the preceding stage. We now classify/predict whether the precursor window predicts a critical state, if at all, and at what confidence, if it does. Prediction employs boosting and label propagation using spectral clustering [47], with spectral clustering revising the results of boosting, if needed, using the cluster assumption. The confidence for predictions is computed using p-value estimation driven by transductive inference. We then estimate the temporal lag, relative to the precursor state, and its characteristics. The lag indicates after how many time instances the critical state, if any, will occur, its intensity, and the level of confidence.

17. Discussions

Life sciences provide further insights on specific ways and means for expanding and/or revising the current EBM and DKA framework. The goals here are to understand among others biological function and evolution. The metrics are set-based, have to discount both random and redundant information (see related "paradoxes" below), while context, storage, and flow of information become relevant [61]. Gibberish doesn't help with any biological process, to paraphrase Gell-Mann [62], and indeed the 1st paradox for life sciences states that a random string adds zero information to the set. The 2nd paradox states that an exactly duplicated ("pre-existing") string adds little or nothing to the overall information in the set. Note that any measure of information proposed should "include the information content of the strings individually as well as the information contained in the relationship with other members of the set." Relationships are about sharing information and possible interactions, e.g., catalysts and enzymes, and determine "function." Complexity, however it is encoded, and emergent behavior and functionality, are intertwined.

The scope of image representations and their association cods relevant to sensory data collection can be also expanded. The weak learners introduced earlier to build strong classifiers (see Section 13) are not limited to simple features. They can also stand for "parts" in the context of object recognition, e.g., face recognition [44], with parts represented as clusters of image patch instances. Such recognition-by-parts architectures employ boosting and transduction and are driven by algorithmic randomness. The architectures can further employ region-based strategies that compare noncontiguous image regions [63]. Towards that end "under certain circumstances, comparisons [using dissociated dipole operators] between spatially disjoint image regions are, on average, more valuable for recognition than features that measure local contrast." This leads to the obvious observation that the recognition-by-parts architecture should learn to sample "optimal" sets of regions' comparisons for recognizing objects, e.g., faces, across varying pose and illumination. The choices made on such combinations (during the boosting feature selection stage) amount to "rewiring" operators. Rewiring corresponds to an additional processing and competitive stage for the feed-forward recognition-by-parts architecture. As a result the repertoire of feature ranges over local, global, and non-local (disjoint) operators ("filters"). Ordinal rather than absolute codes become available in order to gain invariance to small changes in inter-region contrast. Similarity scores using strangeness and p-values can accommodate both ordinal and absolute codes.

18. Conclusions

This paper describes a general framework for evidence-based management, which is geared for data collection and decision making, on one side, and discovery, inference, and prediction, on the other side. It is driven by information theory and statistical learning, algorithmic probability and inference, algorithmic randomness and Kolmogorov complexity, and conformal prediction [9,64]. Reliable confidence is vital for many real world applications that involve anomaly and change detection, categorization, diagnosis and discrimination, link analysis, and prediction and identification of critical states. This paper expands on basic concepts, proposes new theoretical development, and sketches applications that show the feasibility and utility of evidence-based management. The unifying theme throughout the paper is that of "compression entails comprehension", which is realized using the interrelated concepts of randomness as opposed to regularity, Kolmogorov complexity, minimum description length, and ranking scores using strangeness and p-values. Venues for future research are discussed below (see Section 17 for a brief discussion on life sciences).

Drug design and synthesis for successful therapy are challenging and very important problems [65]. They expand on data selection (see Section 14). An example for such a problem goes as follows. Assuming that one is

given triplets (x_i, y_i, g_i) in terms of description x, action y, and score g, respectively, find for a new situation x^* the action y^* that "guarantees" that the corresponding evaluation score g^* falls within some bounded confidence interval. Towards that end, selection (as action) seeks for some cocktail of drugs whose score ("prognosis") would improve on current therapy [52]. Another application would involve exploration and exploitation driven by the interplay between mutagenesis and on-line transduction for the purpose of protein function prediction [66].

Another area ripe for significant development is that of social media analytics. Information from sources such as RSS, Facebook, and Twitter, gets disseminated and its fast growing reach is in the hundred of millions of people. The potential is for both enrichment and subversion. The ever changing information helps to alert to impending and developing critical states, e.g., natural disasters, emergencies, and pandemics. It can be, however, also subject to manipulation in order to bias sentiment and subject preference. It is up to evidence-based management to use spatial context, temporal clustering, and source reputation, to sort out truth from fiction and to do that in real-time.

19. References

[1] S. Emmott, "Towards 2020 Science," MS Research, Cambridge, 2006.

[2] U. Neisser, "Cognition and Reality," *The American Journal Psychology*, Vol. 90, No. 3, 1977, pp. 541-543.

[3] H. Wechsler, "Computational Vision," Introduction, Academic Press, Cambridge, 1990.

[4] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, Urbana-Champaign, 1949.

[5] R. J. Solomonoff, "A Formal Theory of Inductive Inference," *Information and Control*, Vol. 7, 1964, pp. 1-22, 224-254.

[6] M. Li and P. Vitanyi, "An Introduction to Kolmogorov Complexity and Its Applications," 3rd Edition, Springer Verlag, Berlin, 2008. doi:10.1007/978-0-387-49820-1

[7] T. M. Cover and J. A. Thomas, "Elements of Information Theory," 2nd Edition, Wiley, New York, 2006.

[8] V. Vapnik, "Statistical Learning Theory," Springer, Dordrecht, 1998.

[9] V. Vovk, A. Gammerman and G. Shafer, "Algorithmic Learning in a Random World," Springer, Dordrecht, 2005.

[10] O. Chapelle, B. Scholkopf and A. Zien (Eds.), "Semi-Supervised Learning," Massachusetts Institute of Technology Press, Cambridge, 2006.

[11] B. Settles, "Active Learning Literature Survey," University of Wisconsin, Madison, 2010.

[12] R. Galliers and D. Leidner (Eds.), "Strategic Information Management," 4th Edition, Routledge, Cornwall, 2009, pp. 1-2.

[13] G. Schreiber *et al*., "Knowledge Engineering and Management," Massachusetts Institute of Technology Press, Cambridge, 2000.

[14] E. M. Awad and H. M. Ghaziri, "Knowledge Management," Upper Saddle Rive, 2004.

[15] A. Doucet, and A. Johansen, "Technical Report, Department of Statistics, University of British Columbia," 2008. http://www.cs.ubc.ca/%7Earnaud/doucet_johansen_tutorialPF.pdf,

[16] A. Ganek and T. Corbi, "The Dawning of the Autonomic Computing Era," *IBM Systems Journal*, Vol. 42, No. 1, 2003, pp. 5-18. doi:10.1147/sj.421.0005

[17] A. Darwiche, "Modeling and Reasoning with Bayesian Networks," Cambridge University Press, Cambridge, 2009.

[18] J. Pearl, "Causality," 2nd Editon, Cambridge University Press, Cambridge, 2009.

[19] N. Schmid and H. Wechsler, "Information Theoretical and Statistical Learning Theory Characterizations of Biometric Recognition Systems," *SPIE Electronic Imaging: Media Forensics and Security*, San Jose, 2010.

[20] H. Simon, "The Science of the Artificial," Massachusetts Institute of Technology Press, Cambridge, 1982.

[21] S. Nayar and T. Poggio (Eds.), "Early Visual Learning," Oxford University Press, Oxford, 1995.

[22] T. Winograd and F. Flores, "Understanding Computer and Cognition," Addison Wesley, Boston, 1988.

[23] H. B. Barlow, "Unsupervised Learning," *Neural Computation*, Vol. 1, No.3, 1989, pp. 295-311. doi:10.1162/neco.1989.1.3.295

[24] Y. D. Rubinstein and T. Hastie, "Discriminative vs. Informative Learning," Knowledge and Data Discovery, 1997, pp. 49-53.

[25] T. Jebara, "Discriminative, Generative and Imitative Learning," MIT Press, Cambridge, 2002.

[26] C. H. Bennett, P. Gacs, M. Li, P. M. B. Vitanyi and W H. Zurek, "Information Distance," *IEEE Transactions on Information Theory*, Vol. 44, No. 4, 1998, pp. 1407-1423. doi:10.1109/18.681318

[27] K. Proedrou, I. Nouretdinov, V. Vovk and A. Gammerman, "Transductive Confidence Machines for Pattern Recognition," Royal Holloway, University of London, 2001.

[28] M. Kukar, "Quality Assessment of Individual Classifications in Machine Learning and Data Mining," *Knowledge and Information Systems*, Vol. 9, No. 3, 2006, 364-384. doi:10.1007/s10115-005-0203-z

[29] M. Kukar and I. Kononenko, "Reliable Classifications with Machine Learning," *Proceedings European Conference on Mashine Learning*, Banff, Canada, 2002, pp. 219-231.

[30] R. G. Bachrach, A, Navot and N. Tishby, "Margin Based Feature Selection-Theory and Algorithms," *Conference on Machine Learning*, Banff, Canada, 2004.

[31] T. M. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE on Information Theory*, Vol. IT-13, 1967, pp. 21-27.

[32] S. S. Ho and H. Wechsler, "Query by Transduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 9, 2008, pp. 1557-1571. doi:10.1109/TPAMI.2007.70811

[33] T. Melluish, C. Suanders, I. Nouretdinov and V. Vovk, "The Typicalness Framework: A Comparison with the Bayesian Approach," Royal Holloway, University of London, 2001.

[34] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by Transduction," Elsevier Publish, New York, 1998, pp. 148-155.

[35] T. Poggio, R. Rifkin, S. Mukherjee and P. Niyogi, "General Conditions for Predictivity of Learning Theory," *Nature*, Vol. 428, 2004, pp. 419-422. doi:10.1038/nature02341

[36] T. Poggio and S. Smale, "The Mathematics of Learning: Dealing with Data," *Notices of the American Mathematica Society*, 2003, pp. 537-544.

[37] F. Li and H. Wechsler, "Open Set Face Recognition Using Transduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 11, 2005, pp. 1686-1698,. doi:10.1109/TPAMI.2005.224

[38] J. Hamm and D. L. Lee, "Grassman Discriminant Analysis: A Unifying View of Subspace-Based Learning," 25*th International Conference on Machine Learning*, Helsinki, 2008.

[39] S. S. Ho and H. Wechsler, "A Martingale Framework for Detecting Changes in the Data Generating Model in Data Streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 12, 2010, pp. 2113-
2127, doi:10.1109/TPAMI.2010.48

[40] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Conference on Computer Vision and Pattern Recognition*, Kauai, 2001.

[41] Y. Freund and R. E. Shapire, "Experiments with a New Boosting Algorithm," 13*th International Conference on Machine Learning*, Bari, 1996, pp. 148-156.

[42] F. H. Friedman, T. Hastie and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting," *Annals of Statistics*, Vol. 28, 2000, pp. 337-407. doi:10.1214/aos/1016218223

[43] V. Vapnik, "The Nature of Statistical Learning Theory," 2nd Edition, Springer, Berlin, 2000.

[44] F. Li and H. Wechsler, "Face Authentication Using Recognition-by-Parts, Boosting and Transduction," *International Journal of Artificial Intelligence and Pattern Recognition*, Vol. 23, No. 3, 2009, pp. 545-573. doi:10.1142/S0218001409007193

[45] I. S. Dhillon, Y. Guan and B. Kulis, "Kernel K-Means, Spectral Clustering and Normalized Cuts," *Proceedings of the Conference on Knowledge and Data Discovery*, Seattle, Western Australian, 2004.

[46] M. Filippone, F. Camastra, F. Masulli and S. Rovetta, "A Survey of Kernel and Spectral Methods for Clustering." *Pattern Recognition*, Vol. 41, No. 1, 2008, pp. 176-190. doi:10.1016/j.patcog.2007.05.018

[47] A. Y. Ng, M. I. Jordan and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm, NIPS 14," Massachusetts Institute of Technology Press, Cambridge, 2002.

[48] X. Zhu, Z. Ghahramani and L. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," *Proceeding* 20*th International Conference on Machine Learning*, Washington DC, 2003.

[49] D. Pyle, "Data Preparation for Data Mining," Morgan Kaufmann, Waltham, Massachusetts, 1999.

[50] P. J. Huber, "Robust Statistics," Wiley, NewYork, 2004.

[51] J. Weston, R. Collobert, F. Sinz, L. Bottou and V. Vapnik, "Inference with the Universum," *Proceeding of the* 23*rd International Conference on Machine Learning*, 2006, NewYork, pp. 1009-1016.

[52] V. Vapnik, "Estimation of Dependence Based on Empirical Data (2nd. ed.)," Springer Verlag, Berlin, 2006.

[53] C. L. Vale, J. F. Tierney and L. A. Stewart, "Effects of Adjusting for Censoring on Meta-Analysis of Time-to-Event Outcomes," *International Journals of Epidemiology*, Vol. 31, 2002, pp. 107-111. doi:10.1093/ije/31.1.107

[54] T. Kohonen, "Self-Organizing Maps," 2nd Edition, Springer Verlag, Berlin, 1996.

[55] S.-S. Ho and A. Talukder, "Automated Cyclone Discovery and Tracking Using Knowledge Sharing in Multiple Heterogeneous Satellite Data," Proceeding KDD, 2008, pp. 928-936.

[56] A. Panangadan, S.-S. Ho, and A. Talukder, "Cyclone Tracking Using Multiple Satellite Image Sources," Proceeding GIS, 2009, pp. 428-431. doi:10.1145/1653771.1653836

[57] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang and E. J. Keogh, "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," *Proceeding of Large Data Base*, Vol. 1, No. 2, 2008, pp. 1542-1552.

[58] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," KDD Workshop, 1994, pp. 359-370.

[59] M. Vlachos, D. Gunopulos and G. Kollios, "Discovering Similar Multidimensional Trajectories," Procedeeing ICDE, 2002, pp. 673-684.

[60] S.-S. Ho, W. Tang and W. T. Liu, "Tropical Cyclone Event Sequence Similarity Search Via Dimensionality Reduction and Metric Learning," Procedeeing KDD, 2010, pp. 135-144.

[61] D. J. Galas, M. Nykter, G. W. Carter, N. D. Price and I. Shmulevich, "Biological Information as Set-Based Complexity," *IEEE Transaction on Information Theory*, Vol.

56, No. 2, 2010, pp. 667-667.
doi:10.1109/TIT.2009.2037046

[62] M. Gell-Mann, "The Quark and the Jaguar: Adventures in the Simple and the Complex," Freeman, New York, 1994, p. 392.

[63] B. J. Balas and P. Sinha, "Region-Based Representations for Face Recognition," *ACM Transactions on Applied Perception*, Vol. 3, No. 4, 2006, pp. 354-375. doi:10.1145/1190036.1190038

[64] F. Emmert-Streib and M. Dehmer (Eds.), "Information Theory and Statistical Learning," Springer, Berlin, 2009, pp. 1-3.

[65] J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff and B. Scholkopf, "Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design," *Bioinformatics*, Vol. 19, No. 6, 2003, pp. 764-771.

[66] N. Basit and H. Wechsler, "Computational Mutagenesis and Protein Function Prediction Using Computational Geometry," *Journal of Biomedical Science and Engineering*, 2011.

[67] G. Kotonya and I. Sommerville, "Requirements Engineering," Wiley, NewYork, 1998.

