Intelligent Information Management, 2015, 7, 195-222
Published Online July 2015 in SciRes. http://www.scirp.org/journal/iim
http://dx.doi.org/10.4236/iim.2015.74016

Scientific Research Publishing

Cyberspace Security Using Adversarial Learning and Conformal Prediction

Harry Wechsler

Department of Computer Science, George Mason University, Fairfax, VA, USA
Email: wechsler@gmu.edu

Received 4 May 2015; accepted 7 July 2015; published 10 July 2015

Abstract

This paper advances new directions for cyber security using adversarial learning and conformal prediction in order to enhance network and computing services defenses against adaptive, malicious, persistent, and tactical offensive threats. Conformal prediction is the principled and unified adaptive and learning framework used to design, develop, and deploy a multi-faceted self-managing defensive shield to detect, disrupt, and deny intrusive attacks, hostile and malicious behavior, and subterfuge. Conformal prediction leverages apparent relationships between immunity and intrusion detection using non-conformity measures characteristic of affinity, a typicality, and surprise, to recognize patterns and messages as friend or foe and to respond to them accordingly. The solutions proffered throughout are built around active learning, meta-reasoning, randomness, distributed semantics and stratification, and most important and above all around adaptive Oracles. The motivation for using conformal prediction and its immediate off-spring, those of semi-supervised learning and transduction, comes from them first and foremost supporting discriminative and non-parametric methods characteristic of principled demarcation using cohorts and sensitivity analysis to hedge on the prediction outcomes including negative selection, on one side, and providing credibility and confidence indices that assist meta-reasoning and information fusion.

Keywords

Active Learning, Adversarial Learning, Anomaly Detection, Change Detection, Conformal Prediction, Cyber Security, Data Mining, Denial and Deception, Human Factors, Insider Threats, Intrusion Detection, Meta-Reasoning, Moving Target Defense, Performance Evaluation, Randomness, Semi-Supervised Learning, Sequence Analysis, Statistical Learning, Transduction

1. Introduction

Cyber security affects the fabric and infrastructure of modern society. It encompasses the interplay between

science, technology, and engineering practices to protect networks, computers, programs, and data from attacks, damage, insider threat, or unauthorized access (e.g., intrusion) for criminal and nefarious purposes. Cyber security is first and foremost about all-encompassing recognition. It is about intrusion detection and is adversarial in nature. It is crucial for both biological (e.g., immune system) and machine Oracle systems to recognize patterns as friend or foe and to respond to them appropriately. Failure to recognize pathogens or subterfuge such as Trojan horses, characteristic of malware, can be fatal. Recognition is continuous and multi-layered. It includes detection (e.g., intrusion detection system), categorization, and continuous re-authentication. This paper considers the use of adversarial learning as the methodology of choice to enhance cyber security defenses against adaptive, malicious, persistent, and tactical offensive threats. To meet such goals the paper proposes conformal prediction as the principled and unified learning framework to design, develop, and deploy a multi-faceted protection and self-managing defensive shield that supports adversarial learning.

Intrusion detection bears many analogies to biological immunity including the challenges raised by the lack of abnormal patterns (e.g., imposters or pathogens) for modeling and training that would most likely afflict the network and its components, and cause therefore significant harm. Additional challenges come from countering the possibility for denial of service (DoS) and reduced quality of service (QoS) while maintaining high sensitivity (e.g., high detection or true positive rates) for detection purposes and high specificity to avoid high false (positive) alarm rates. Training and learning, which are responsible for model selection and prediction is the core for intrusion detection and depends much on the quality, quantity, and type of data available in order to demarcate normal from abnormal traffic. It is also highly desirable that the decisions made on intrusion carry reliability indices suitable for further processing and protection. Both noisy and faulty data due to poor observations and recordings, annotation mistakes, and deliberate attempts for obfuscation and spoofing, affect training and the safeguards in place for protection purposes.

One can approach intrusion detection using either supervised or unsupervised learning or some mixture of both. Supervised learning, which involves binary or multi-class classification with classes making reference to normal traffic, on one side, and attacking traffic, on the other side, can be therefore addressed using discriminative methods (e.g., decision trees trained over normal and known abnormal data). There is always, however, innovation and novelty about both legitimate and illegitimate traffic so supervised learning cannot be expected to handle zero-shot (unseen) attacks and/or deviations from normal behavior. To address both novelty and lack of enough and representative data for intrusions, one is advised to substitute unsupervised learning for supervised learning vis-à-vis anomaly or outlier detection using amongst others one-class classification (e.g., K-nearest neighbors (KNN) and one-class SVM). Both anomaly detection and the provision of reliability indices for the decisions made are addressed using conformal prediction, in general, and transduction, in particular, as discussed throughout this paper. Here the analogue to outlier detection comes from assessing the extent to which observations are more extreme or strange than normal data. The higher the strangeness or atypicality for a new observation, the more likely it is that the observation of concern can be traced to illegitimate traffic. There is always the possibility that the quality for training data is lacking due to intrusions (e.g., noise) labeled as normal data, and notwithstanding if the noise is deliberate (e.g., insider threat) or not. This is characteristic of adversarial learning and is addressed next.

Annotation for training can be lacking on purpose (e.g., deliberate) or not. Poor annotation affects not only cyber security (e.g., spam, phishing, fake internet accounts, and fraud detection) but basic science as well, as it is the case with coronal mass ejection (CME) events that are fed to design detection and tracking methods for solar physics [1]. It is often difficult to obtain annotated data for pattern recognition tasks; however, public email service providers have the ability to solicit annotation support from their users. A select set of users (e.g., Amazon Mechanical Turk (AMT)), can be asked to occasionally provide a class label for a randomly selected incoming email message. This, of course, allows an adversary (including insider threats) to taint the data used to train the spam detection method. An adversary might mislabel a spam message as not spam in order to allow similar spam messages to be delivered in the future, or alternatively it might mislabel a non-spam message as spam. The same adversary can also overwhelm intrusion detection system (IDS) with highly unbalanced and/or corrupt training data to further compromise the integrity of the spam detection method. While it may be possible to restrict invitations for annotations to well-established accounts, an adversary may create fake accounts with the intent to influence the training of the spam detection method. Even under the best of circumstances annotators may still make unintentional errors.

Adversarial learning (AL) involves effective allocation of finite resources including but not limited to human

and/or machine annotation. The nominal (e.g., computational) and representation (e.g., descriptive) aspect is handled by active learning (ACL) and includes both importance sampling and feature extraction and selection using feature relevance and reputation. To undermine defenses, the adversary places emphasis on denial and deception (D&D) to evade detection and to deceive defenses. The motivation behind active learning comes from the need for promptness and selectivity in separating (e.g., filtering informative contents) from obfuscation using limited resources. This involves what is best to annotate (e.g., queries of interest) and when, the annotation process itself (e.g., the Oracle), and countering what are perceived to be potential vulnerabilities (e.g., to compromise the Oracle) affecting the learning processes involved in intrusion detection systems (IDS) training and/or the reputation of the features involved in social network analysis (SNA).

Miller *et al*. [2] have recently surveyed the field of ACL to promote Security-oriented Active Learning Testbed (SALT) architecture in order to experiment and evaluate diverse strategies surrounding active learning to counter adversarial contexts and deliberate manipulation. SALT evaluation has been so far relatively limited to the continuous 2D feature space where the aim is usually that of learning a binary classifier while evaluating different active learning strategies to prioritize requests (e.g., queries) for annotation. The results reported indicate that for "the system which is not under attack, the maximum uncertainty strategy performs significantly better than random choice. When the system is under attack [the case of interest here], randomization of training samples becomes the best strategy, while the maximum uncertainty choice suffers severe degradations" [2]. Principled randomization using conformal prediction inspired transduction (see Sect. 3 and 4) and randomness (see Sect. 11) are better ways to engage in importance sampling for active learning purposes. Vulnerabilities, both known and unknown are many and each of them affects cyber security differently including functional creep bearing on privacy and interoperability bearing on use across different platforms [3]. Exploratory vulnerabilities, which focus on modifying test samples once a model has been trained, and causative vulnerabilities, which aim to modify both training and test data can be handled using inductive conformal prediction (ICP) driven by conformal prediction and incremental transduction, when unlabeled and training data are complementary to each other in the execution of annotation. The i.i.d. assumption held by SALT and others is relaxed later on using online compression models (see Sect. 10).

Conformal prediction supports a multitude of functional blocks that address the major challenges faced by adversarial learning including denial and deception, pattern representation and classification, and vulnerabilities, deliberate or not, affecting learning, training, and annotation. The solutions proffered are built around meta-reasoning, active learning, and most important and above all adaptive Oracles seeking to be effective and efficient regarding the prediction outcomes. Effective to be resilient to malicious attacks aimed at subverting promptness and selectivity in separating the wheat (e.g., informative patterns) from the chaff (e.g., obfuscation), and efficient to minimize the costs incurred. Additional motivation for using conformal prediction and its immediate offspring, those of semi-supervised learning and transduction, comes from the simple but important realization that such intertwined learning methods are consonant with the deployment of discriminative methods using likelihood ratios; demarcation using cohorts, local estimation, and non-conformity measures; randomness for hypothesis testing and incremental inference using sensitivity analysis; reliability indices, such as credibility and confidence using relative strangeness or a-typicality, to augment the current practice of bare prediction outcomes; high-dimensional change detection using martingale; open set recognition (including the reject option) and negative selection using transductive p-values and skew; and consensus reasoning to upend questionable label annotations, deliberate or not, using aggregation and importance sampling.

Additional motivation for using conformal prediction comes from biometrics and forensics. Many analogies hold between adversarial learning and biometric mass screening and identity management with both vulnerable to impersonation, with ground truth annotation lacking for the non-self, and with uncontrolled settings and change (e.g., covariate shift) the norm rather than the exception. Impersonation and spoofing in biometrics are directly related to obfuscation seeking to confuse training during cyber security engagements.

The structure and outline for the paper address motivation and justification throughout, theoretical foundations, and overall methodology and development tools. The discussion covers both existing methods and promising venues for future R&D. The particular outline for the paper goes as follows. Background for the whole cyber security enterprise continues in Sect. 2 the discussion started in this introductory section. The theoretical foundations on machine learning are discussed in Sect. 3, while conformal prediction, the novel learning core advanced in this paper, is introduced in Sect. 4. Particular methods and development tools including active learning and change detection, semantics and stratification, and the symbiosis of immunity and intrusion detec-

tion, which draw and support conformal prediction, are presented in Sections 5 - 7, respectively. The design and development of an all-encompassing software environment methodology where one can explore and exploit in a coordinated fashion different functional modules that complement each other and address conflicting asymmetries is discussed in Sect. 8 using meta-reasoning and meta-recognition. The remaining sections are dedicated to additional topics that support and provide value-added to a self-management and protective shield vis-à-vis malware attacks, in general, and intrusion detection, in particular. The particulars topics of interest for innovation and their potential impact on cyber security include insider threats and moving target defense, online compression models, randomness, distributed semantic models and vector space representation that substitute context and prediction to traditional bag of words and counting, and human factors (see Sections 9 - 14). Quo Vadis and prescriptive conclusions discuss promising venues for future R & D and bring closure to the paper in Sections 15 and 16, respectively.

2. Background

We review briefly here on current and relevant background and literature regarding adversarial learning [4]. The review takes place at the interface between machine learning, robust statistics, cyber security, and computing. We recall that adversarial learning can be deliberate or not in nature, that information can be missing, corrupt, or superfluous, and that it is most important to assess and respond properly to the very possibility that the Oracle involved with decision-making, including detection and classification, can be compromised. Spam filter (e.g., Spam Bates) can be rendered useless even if the adversary's access is limited to only 1% of the training messages with relatively little system state information and relatively limited control over training data [5]. Focused attacks assisted by insider threat with extra domain specific knowledge can do even more harm. This suggests that the defense should exercise great caution in its use of training data. Towards that end, labeling errors characteristic of lacking proper annotation can be detected and redressed in many ways including constrained regularization driven optimization and label flipping [6] with label flipping best done in a principled way characteristic of importance sampling [7]. The use of soft rather than hard labeled data is more resistant to (adversarial) label noise [3] with soft labels integral to conformal prediction, the learning framework proposed here, and to incremental transduction, in particular [8].

Questions relevant to adversarial learning inquire about how to integrate and process new information, possibly mistaken, which does not fit within the existing mold. Those are also hard questions for medical diagnosis and their resolution will be of great help to clinical practice (e.g., NYU Langone Medical Center announced significant changes in its procedures after the death by septic shock of a 12-year-old boy who was sent home from the center with fever and a rapid heart rate under the erroneous presumption of a run-of-the-mill ailments in children, such as stomach bug). The use of active learning (see Sect. 5) and semantics and stratification (see Sect. 6) would have helped to prevent such fateful misdiagnosis. The mutual relationships between rudimentary biological immunity subliminal to artificial immune system (AIS) and intrusion detection (e.g., IDS) would have further helped with avoiding the above misdiagnosis (see Sect. 7).

The way society conducts its business depends more and more on broadening the cyber space while maintaining proper security and privacy norms of behavior for the very purpose of ensuring the trustworthiness of the cyberspace. Among many concerns deceit and its prevention make the top list. There is offense and there is defense with both attempting to guess and learn from each other. This is the core for adversarial learning. There are many dimensions along which adversarial learning is played out, among them active learning that is tasked to choose what to learn and how to build robust Oracles, moving target defense to counter advanced persistent threats (APT), and moving target defense (MTD) to increase the time it takes to mount and execute an attack while decreasing the time it takes to deploy defensive moves. Much of the adversarial aspect is driven by greed and monetary rewards so the economics of the cyberspace play a major role. Adversarial learning therefore benefits and impacts on the way social and behavioral economics (SBE) are conducted and safeguarded in the cyberspace.

The challenges, which are typically of adversarial nature, are many and they are succinctly tabulated here in context along possible solutions we describe throughout. Adversarial is meant as any attempt to either deceive or defend against deception. Detecting adversarial advertisements in the wild [9] expands on the adversarial aspect. It involves intertwined minority-class and multi-class issues with scale affecting both. The majority of ads are from good-faith advertisers and those of adversarial (e.g., malicious) nature are few and spread between coun-

terfeit goods and inaccurate claims. Local estimation using open set recognition driven by non-conformity measures and ranking putative classification assignments address both such (minority and multi-class) aspects (see Sect. 3 for TCM-DR) with large scale addressed using cascade implementations that trade false positive and false negative rates and can be efficiently run using Map Reduce. One could also consider 1) adversarial behavior aiming to affect personalized recommender systems; and 2) adversarial attempts to tamper the use of mobile devices (e.g., location wise) while seeking for spatial-temporal trajectories when using the sequence of messages as the basic unit of information. This is relevant for both SBE and social network analysis (SNA). Another aspect of interest for adversarial learning is to what extent human experts should be involved and when, something active learning is best qualified to engage in for purposeful annotation with adaptive Oracles (e.g., machine) and/or human expertise competing for the job (see Sect. 13).

Adversarial learning that is conducive to cyber security can be conducted using reactive and/or proactive modes of operation [10]. In the reactive mode, the offensive side devises and engages in the attack while the defense is limited to analyzing the attack and developing countermeasures. The proactive case learns from the past and seeks to anticipate and forecast. The defense gets more involved as it models the adversary, simulates attacks, evaluates attacks' impact, and develops countermeasures to prepare for that time when it is under attack and has to raise and deploy adequate defenses. It is a game like environment with learning and adaptation substituting for formal game theory to contend with practical issues of implementation and use of computational resources. The taxonomy for adversarial learning takes place along three axes: influence (exploratory/reconnaissance or causative), specificity (targeted or indiscriminate), and security violation (integrity, availability, or privacy) [11]. The particulars of adversarial learning including task and mode of operation are therefore primed by meta-planning to distinguish between different attacking behaviors, and where are they encountered and for what purpose (e.g., advertising and marketing, sentiment analysis). The persistent arms race between defense and offense brings up continuous tradeoffs between reverse engineering where the attackers want to divine the Oracle used by defense and/or compromise it, and randomization schemes that are used by defense to confuse the attackers. It has been reported that the defender's optimal policy tends to be either to randomize uniformly (ignoring baseline classification accuracy), which is the case for targeted attacks, or not to randomize at all, which is typically optimal when attacks are indiscriminate [12]. This suggests the need for meta-recognition (see Sect. 8) to choose among the arsenal of methods available while considering random matrix theory (see Sect. 11) and random projections for message representation [13].

We have explored both adversarial and active learning along several dimensions and found promising solutions for a number of challenges that involve representational aspects and/or detection methods. Regarding representation we have reported recently a number of innovations: 1) clustering (e.g., stratification) and active learning that yield a 90% reduction in the need for annotation, which is expensive to start with [14]; 2) social network analysis (SNA) reputation features that yield an increase of 70% in detection rate compared to content filters that ignore server reputation [15]; 3) spectral clustering of URL n-grams and transductive semi-supervised learning that yield a 100% increase in detection rate (e.g., doubling the detection rate while maintaining the same false positive rate) of adversarial message modification compared to filters that rely on contents only for classification [16]; 4) reputation and similarity features that yield a 13.5% increase in cost savings when challenged by changes in data distribution [17]; and 5) random boost that yields 75% reduction in computational costs compared to random forest, where both random projections and boosting help with feature randomization to counter adversarial attacks [13].

3. Machine Learning

The learning themes addressed by adversarial learning are those of model selection and prediction and they are dealt with using statistical learning theory (SLT) [18] [19]. SLT involves non-parametric learning and inference from finite samples (rather than asymptotic in nature). It is conceptually modeled using system imitation (e.g., mimicry) rather than accurate identification for both generalization and prediction purposes (e.g., SVM). SLT, which trades margin (for better generalization) against complexity (e.g., MDL), estimates local functional dependencies (e.g., Oracles) and putative classifications using relatively small rather than big collections of data. This helps with the deployment of robust IDS that cannot be easily compromised by adversarial learning using sporadic but dense subversion. There are three learning frameworks that draw from SLT and used here. First comes semi-supervised learning (SSL) [20], with its basic assumptions: the smoothness assumption, characteris-

tic of supervised learning, where similar samples (e.g., messages) share similar labels; the cluster assumption, where samples in the same cluster are likely to be of the same class and their subsequent stratification and prototyping helps with both performance and scale; and the low density separation assumption that seeks for decision boundaries in low-density regions. Transduction, similar in spirit to SSL, comes next. It leverages the complementary use of training and test data, the latter providing insights on its own data distribution. Transduction employs the strangeness or alternatively the typicality of messages to mediate between randomness deficiency (and regularity), Kolmogorov complexity, and minimum description length (MDL). Conformal prediction (CP) hedges and punts on labels to determine putative sets of predictions including their inclusion, order, and ranking in terms of likelihoods. Similar to the immune system, CP engages in open set recognition to distinguish the "self" (e.g., proper training data) from the "non-self" (e.g., adversarial training data) and thus to detect and reject unfamiliar patterns. The proposed learning frameworks go beyond bare predictions and provide reliability indexes (e.g., credibility and consistency) for the classification choices made on missing labels. This facilitates information fusion, meta-reasoning (e.g., control and gating), and meta-recognition, which leads to learning consistency and stability for Oracles and helps with reliable and robust classification including operational resilience to recover from misdiagnoses once changes in the data distribution are observed and tracked.

Randomness deficiency and Kolmogorov complexity are intricately related. The larger the randomness deficiency is the more regular and more probable some string (e.g., message or attack) is. Transduction chooses from all the possible labeling for unlabeled data the one that yields the largest randomness deficiency, *i.e.*, the most probable one. Towards that end, one employs randomness and complexity using similarity and corresponding rankings driven by strangeness and p-values. The strangeness, which stands for non-conformity measures (NCM), measures the lack of typicality with respect to its true or putative (assumed) identity label and the labels for all the other patterns known. The strangeness α is the (likelihood) ratio of the sum of the K nearest neighbor (KNN) distances d from the same class y divided by the sum of the KNN distances from all the other classes $(\neg y)$. The smaller the strangeness, the larger its typicality and the more probable its (putative) label y is, where the use of KNN corresponds to lazy learning. Alternative definitions for the strangeness more suitable for anomaly detection (e.g., unsupervised learning) framework (rather than the supervised learning framework) on intrusion detection involve only the numerator and dispense with the denominator from the earlier definition. This makes the strangeness for samples far away from normal traffic (e.g., illegitimate traffic) larger than the strangeness of samples that belong to normal traffic. The strangeness further facilitates both feature selection (similar to Markov blankets) and variable selection (dimensionality reduction) to decrease the time the defense spends on detection. One finds empirically that the strangeness, classification margin, sample and hypothesis margin, near miss, posteriors, and odds are all related via a monotonically non-decreasing function with a small strangeness corresponding to a large margin. This is similar in spirit to cohort (e.g., context-aware estimation and learning) models using likelihood ratios, universal background model (UBM), and logistic regression [21]. One further notes that the KNN error approaches the Bayes error (with factor 1) if $k = O(\log n)$, that the strangeness α is related to the optimal decision boundary $(\alpha = 1)$ and the posterior $P(c_j | x_j)$; and that the KNN strangeness would smooth boundaries and generalize better than KNN, particularly for overlapping distributions.

The likelihood-like definitions for strangeness are intimately related to discriminative methods. The p-values described next compare ("rank") the strangeness values to determine and rank the credibility and confidence in the putative classifications made. The p-values aggregate information on relative strangeness and inform on the degree of typicality. The p-values bear resemblance to their counterparts from statistics but are not the same [22] [23]. They are derived using the relative rankings of putative classifications against each one of the classes known. The standard p-value construction, where m is the cardinality of the training set T, constitutes a valid randomness (deficiency) test approximation for some putative label y assigned to a new pattern z and is associated with $p_y(z) = \#\left(i : \alpha_i \geq \alpha_{new}^y\right) \Big/ (m+1)$. The p-values are used to assess the extent to which data supports or discredits the null hypothesis $H0$ (for some specific label). The largest p-value defines the credibility of the classification chosen (e.g., all other alternative classifications are stranger and thus more a-typical). The confidence measure is the difference between the top two p-values. It indicates how close to each other the first two classifications are and it thus measures for ambiguity and uncertainty. The larger the confidence the smaller the ambiguity is.

Demarcation, using the transduction confidence machine (TCM) [24], follows from the above discussion. TCM chooses for classification that label that is most consistent with the current and well-localized training set

(e.g., the self) and yields therefore the highest p-value (e.g., most typical and thus least strange label) for credibility of the class (e.g., label) assignment made. The associated confidence measure records the difference between the top two p-values and indicates the degree of ambiguity in the class assignment made. Credibility and confidence are the particular reliability indices that TCM provides for further processing, e.g., meta-reasoning, meta-recognition, and information fusion (see Sect. 8). Note that once one iterates through putative label assignments for the probe of interest, the KNN for training instances can change and their corresponding strangeness and thus p-values may change too. The intuition behind TCM is to model the test sample in a fashion most similar to the training set while minimally changing the original (predictive) model learned. Note that TCM-KNN includes different versions indexed by the number of n nearest neighbors K involved, with the class of demarcation machines referred to as TCM-KNN.

TCM-KNN is most suitable for closed set recognition. It does not, however, address the detection aspect needed for open set recognition, in general, and IDS, in particular. One needs to quantify how strange something needs to be before it can be thought of as non-self and thus be rejected as novel and unfamiliar (see Sect. 7). Towards that end, one expands the basic TCM into TCM-DR (with DR standing for detection and recognition), with detection and rejection suitable for open set recognition and outlier (e.g., intrusion) detection, in general, and impersonation (of legitimate training instances) detection, in particular. The question left for TCM-DR to answer is when to exercise the reject option non-self detection purposes. The challenge here is that one can model only the "normal" class and this leads to one-class classification problems. The approach pursued by TCM-DR leverages the distinction between individual strangeness and ("context aware") p-values, when some novel (e.g., slightly but not much different) pattern characteristic of the "self" is not necessarily a-typical overall compared to alternative interpretations for which the non-self (e.g., alternative putative labels) yields much smaller p-values. One-way to determine the threshold needed for rejection is to re-label each self-pattern as non-self, derive the corresponding p-values distribution under such a false assumption, and then empirically determine a suitable threshold that demarcates self from non-self. Similar to SSL, changing label assignments (characteristic of impersonation) provides the bias needed to determine the rejection threshold required to demarcate the self from non-self. This is characteristic of revision using "ghost" or "virtual" examples subsumed by the symmetrization lemma [25] when one replaces the true risk by an estimate computed on an independent (e.g., auxiliary) set of data, e.g., unlabeled or test data. Towards that end, one re-labels the training samples, one at a time, with all the putative labels except the one originally assigned to it. The PSR (peak-to-side) ratio, $PSR = (p_{\max} - p_{\min}) / p_{stdev}$, records the resulting p-value distribution and determines, using cross validation, the [a priori] threshold used for self-authentication. The PSR values found are low because their relative strangeness is high (and p-value low) and the threshold used is empirically chosen as several standard deviations away from the PSR mean [26]. The PSR distribution derived in such a manner supports negative identification (e.g., negative selection). Edits and revisions, similar to learning from hints and/or side information, are basic mechanisms available throughout for recovery and self-healing from attacks.

4. Conformal Prediction

Conformal prediction (CP) mediates the incremental use of discriminative methods (e.g., likelihood ratios) using varying non-conformity measures (NCM) for strangeness. The purpose for NCM is to support hedging and punting when making predictions and to provide reliability indices such as accuracy (e.g., credibility) and certainty (e.g., confidence). Such indices, radically different from indices such as support and confidence (see Sect. 5) derived during data mining search for associations (e.g., A Priori), are well-calibrated (see below) and are generated assuming only that the data are generated independently by the same (but unknown) (i.i.d.), probability distribution, a standard assumption in machine learning that can be, however, relaxed (see Sect. 10). Transduction, in general, and both the Transduction Confidence Machine (TCM) and the Transduction Confidence Machine for Detection and Recognition (TCM-DR), in particular, which are examples of CP offspring, yield credibility and confidence as reliability indices using different NCM variants for strangeness, p-values, and skew (see Sect. 5). Furthermore, the credibility index is well calibrated (or conservatively valid) to ensure that the frequency of prediction errors does not exceed t (between 0 and 1) at each confidence level $1-t$ (in the long run). Smaller values of t correspond to greater reliability. The confidence measure, which expresses the extent of ambiguity, becomes efficient as the TCM and TCM-DR prediction (nested) sets (regions) Γ shrink (in terms of number of possible outcomes) with the prediction regions as small as possible, e.g., the inductive

conformal (e.g., transductive) predictor (ICP) maps a (labeled) data sequence (training set) T and a new data sample x as $\Gamma^t(T,x)=\{y\in Y: p_y > t\}$ with $\Gamma^{t_1}\subseteq\Gamma^{t_2}$ for $0\le t_1\le t_2\le 1$. Empty predictions correspond to impersonation and spoofing that are therefore rejected. Transduction, similar to CP, is incremental in nature. It leverages the complementarity between training and test data for the purpose of robust and stable predictions (e.g., test data participate in its own disposition and classification). One can use for computational efficiency the following ICP method [27] similar to incremental transduction [8]. Divide the training set T into proper training set $T1$ and calibration set $T2$. Construct prediction rule F for classification using $T1$. Compute NCM (e.g., strangeness) score α for samples (x_i, y_i) in $T2$ using difference $\Delta(y,y')$ with $\alpha_i=\Delta(y_i, F(x_i))$. For each sample x_j and every possible label $y\in Y$, compute $\alpha_j=\Delta(y, F(x_j))$ and the corresponding p-value against calibration set $T2$ using for cardinality $|T1|-|T2|$. The prediction set is $\Gamma^t(T,x)=\{y\in Y: p_y > t\}$ and one can show that ICP is valid.

5. Active Learning and Change Detection

Adversarial learning leverages data and information contents to detect and counter attacks in order to ultimately build robust defenses. Rather than being passive in randomly selecting data instances for learning Oracles, one engages instead in purposeful collection and processing [28]. While collecting large amounts of data lacking annotation for training is straightforward and quite feasible, their annotation requires that significant effort needs to be chosen for inquiry on their annotation are selected dynamically for the purpose of learning and generalization. This follows the resolution and annotation of previous queries made using abstract but intertwined notions of margin, separability, and importance sampling. The choice on queries is made with the implicit expectation for lesser ambiguity and corresponding error reduction. Selection takes place using online or pool-based methods realized as serial or batch, the latter to avoid frequent retraining. Active learning is the functional block that addresses the requirements and objectives listed above. It is incremental and progressive in nature, with data instances once primed for selection subject to online and/or batch annotation expected to augment the training set for further principled selection, annotation, and periodic but timely IDS retraining. Fundamental to active learning is the uncertainty about what is best to query for possible annotation leading to better generalization, with the pseudo-metrics and costs sorting out and ranking the candidates for annotation.

Active learning is about making optimal choices to improve accuracy (including tradeoffs between false positives and false negatives), trading latencies between defense and offense, and generating cost savings on the resources used. Active learning, however, is not only about data instance selection. It is also about data reduction for better use of computational resources and deployment of more powerful intrusion detection methods. Active selection addresses both anomaly-based intrusion detection, which is most suitable to unknown attacks, and signature-based intrusion detection, which is most suitable for detection of known attacks. Anomaly detection, similar to immunity (see Sect. 7) is usually afflicted by high rates of false positives due to lack of knowledge about the non-self (e.g., unknown types of attacks) relative to possible deviations from legitimate behavior (e.g., the self) that are characteristic of signature-based intrusion detection. One can observe that anomaly detection or alternatively intrusion detection can benefit from the use of TCM-KNN [29] [30] to achieve higher true positive rates and lower false positive rates. In addition, data reduction in terms of less instance samples for training using active learning and lower dimensionality for those samples using feature selection becomes feasible at reduced computational costs. Note that one can redefine to advantage strangeness using only the KNN distances to normal instances. This is similar and consonant with the concept of one-class classification when anomalies are not known ahead of time.

Conformal prediction driven by algorithmic randomness, hypothesis testing, and transductive inference, provides rigorous theoretical guarantees on the error frequencies of the predictions made on unseen data samples. Transductive active learning is much cheaper in a stream-based setting (especially with the use of incremental classifiers such as ICP) where data instances are observed sequentially. It is therefore intuitive to use p-values for active learning in the stream-based (online) setting (e.g., query by transduction (QBT) [31]. The samples are queried based on the difference between the top two p-values computed using the likes of TCM-KNN. Using existing relations between transduction, Bayesian statistical testing, Kullback-Leibler divergence, and Shannon information, QBT was found related to the Query-by-Committee (QBC) paradigm for active learning. The specifics for QBT are straightforward. Let p_i be the p-values obtained for a particular instance x_{n+1} using all the possible class labels $i=1,\cdots,M$ and let p_j and p_k be the two highest p-values when sorted in descending order. The absolute difference between p_j and p_k provides a degree of information contents for the unla-

beled instance, with a smaller value of the difference denoting a larger ambiguity regarding the proposed label. To quantify the uncertainty of the information contents possessed by each instance or sample, one defines its ambiguity as $I(x_{n+1}) = |p_j - p_k|$. As $I(x_{n+1})$ approaches 0, the uncertainty in classifying the unlabeled instance increases. Thus, the addition of this unlabeled sample with its actual label (obtained using a human or machine Oracle and with incremental machine transduction the Oracle of choice here) to the training set provides substantial information regarding the structure of the data model. Such an unlabeled instance represents a promising data sample from an active learning perspective.

One can expand on QBT to advance a generalized version based on eigen-decomposition of matrices, where the p-values corresponding to all the putative class labels of a given sample point are integrated to decide whether or not to query any particular data sample [32]. As an example, lightweight TCM-KNN [33] subjects the input space (e.g., web transactions) with attributes such as one-way delay, request/response delay, packet loss, overall transaction duration, and delay variation (jitter) (similar to Covert Time Channels (CTC) [34]) (see Sect. 9) toward instance selection. The result for lightweight TCM-KNN is that 5600 rather than the original 98,000 training points yield similar (TP, FP) indices but at a much-reduced computational cost, e.g., original TCM-KNN (100%, 1.28%) vs. light TCM-KNN (98.38%, 1.87%). Savings of 98.65% building time for the training set and 66.45% detection time for TCM-KNN are further reported using active learning and feature selection, respectively [35].

Yet another dimension for active leaning is that of data reduction "to avoid the curse of dimensionality and alleviate the annotation workload" [30]. First and foremost and characteristic of data reduction is feature selection. There are many methods for feature selection including (Pearson) chi-square and/or Fisher discriminant tests, to determine whether there is a significant difference between the expected ("feature") and observed ("class") frequencies in one or more categories, entropy and mutual information, and/or (filter and/or wrapper) methods using genetic algorithms (GA), characteristic of evolution and including generational GA (GGA), steady-state GA (SGA), heterogeneous recombination and cataclysmic mutation adaptive search algorithm (CHC), and population-based incremental learning (PBIL) [33].

As information contents and ambiguity are further primed by changes in the data distribution, another dimension for active learning is to first seek for and detect such changes and then to leverage them. This takes place in particular for the purpose of online change (and drift) detection for multidimensional data sequences based on testing exchangeability using martingale [23] [31]. We recall first that some patterns found strange are not counted as a-typical or novel when compared to alternative non-self samples. We also recall that both the strangeness and p-values provide the information needed for open set recognition. This holds for both anomaly (e.g., outlier) and novelty detection using the relative order of p-values for ranking. Change detection seeks to identify those time instances when the underlying distribution for time-varying data streams (e.g., attacks) undergoes a significant change, which shows as a break down in exchangeability. Given unlabeled training set $T = \{x_1, \cdots, x_n\}$, the strangeness of a sample x_i with respect to a cluster model induced by T is $s(T, x) = \|x_i - c\|$ with c standing for cluster centers. Using the strangeness, a family of martingales, indexed by ε drawn from $[0,1]$ and referred to as the randomized power martingale, is defined as $M_n^{(\varepsilon)} = \prod(\varepsilon p_i^{\varepsilon-1})$ with the p-values p_i approximately distributed uniformly over $[0,1]$. One notes that the skewness, a measure of the degree of asymmetry of a pdf distribution, deviates from close to zero (for uniformly distributed p-values) when the underlying model changes, due to adversarial learning, in general, and impersonation, in particular. The skewness is small and stable when there is no change. The skewed p-value distribution plays an important role in the martingale test for change detection as small p-values inflate martingale values. As a result, the martingale $M_n^{(\varepsilon)}$ increases, and evidence starts to accrue against the null hypothesis $H0$ of exchangeability in a data sequence. The increase in martingale value is used to test for change detection. The martingale, non-parametric, incremental, single-pass and working on both labeled and unlabeled data instances, does not require a sliding window on the data stream or the need to monitor the explicit performance of the classification or clustering model as data samples are streaming. The martingale is most suitable for high-dimensional data streams, compared to the Sequential Probability Ratio Test (SPRT) and Cumulative Sum (CUSUM) [36] that are suitable for time-series confined to 1D. The martingale 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.

6. Semantics and Stratification

The motivation here is to confront attack diversity and scale using semantics and stratification using clustering

(e.g., Chinese Restaurant Process and spectral clustering), topic discovery using probabilistic latent semantic analysis (PLSA) and latent Dirichl *et al.* location (LDA)) [37], and aggregation and consensus reasoning using RANSAC and Random-Hough Transform) [7]. Defense can be further enhanced using feature reputation [17].

We start by reviewing here our own ongoing efforts on semantics and stratification. Towards that end we have developed phish GILLNET, a multi-layer methodology characteristic of meta-reasoning, for phishing detection [38]. The first layer (phishGILLNET1) employs PLSA to build a topic model that captures diversity using semantics for stratification. The topic model handles synonymy and polysemy (words with multiple meanings) and other linguistic variations found in phishing to counter adversarial attacks that involve deception. Deliberate misspellings are handled using Levenshtein editing and Google APIs for correction. The term document (TD) frequency matrix is fed to PLSA to find phishing and non-phishing topics using tempered expectation maximization. The performance of phish GILLNET1 is evaluated using PLSA fold-in technique with classification driven by Fisher similarity. The second layer, phishGILLNET2 employs AdaBoost [39] to build a robust ensemble classifier using the topics found earlier for stump classifiers. The third layer, phish GILLNET3, expands on phish GILLNET2 while learning a classifier from labeled and unlabeled examples using co-training [40]. Experiments conducted using one of the largest public corpora of email data containing 400,000 emails show phish GILLNET3 outperforms state of the art phishing detection methods to achieve an F1-score of 100%. Moreover, phish GILLNET3 requires only a small percentage (10%) of data to be annotated thus saving significant time for defense and avoiding errors incurred in annotation.

We expanded on phish GILLNET for the dual purposes of phishing detection and impersonated entity discovery using conditional random fields (CRF) and LDA, the first leveraging and the latter adding a Dirichlet prior for per-document topic distribution that serves as a proper generative model for new messages [37]. Impersonated entity discovery helps with black hat characterization and locks on potential attackers, which pretend to be from a legitimate organization and direct users to fake websites, which resemble legitimate sites aiming to collect users' personal information. Towards that end, phishing web side detection (PWD) involves LDA, intelligent web crawler, image to text conversion, and is device and language neutral. Our approach engages name entity recognition (NER) and discovers the impersonated entity from messages that are classified as phishing at a rate of 88.1% [41].

The Chinese Restaurant Process (CRP) is a recent application of non-parametric clustering that does not fix in advance the number of clusters. This is relevant to situations when attacks are varying and characteristic of multi-class environments, as it is the case when detecting adversarial advertisements in the wild [9]. Similar to the infinite Gaussian mixture model (IGMM) but different from K-means, CRP does not need to manually set in advance the number of clusters to be found. The arriving "customers" (e.g., messages or attacks) can choose to sit alone at the first free/unoccupied table or join other "customers" at any of the already occupied tables. The corresponding probabilities for the nth arriving customer are $a/(n-1+a)$ and $n_k/(n-1+a)$, respectively, where a determines how likely it is that a customer would choose to sit by herself at the first unoccupied table and n_k is the number of customers already seated at one of the k tables with k ranging from 1 to the number K of tables currently seating customers. The CRP distribution of table assignments for any arriving customer, proportional to either a or n_k, thus favors crowded tables.

The goal for most customers is to get seated with similar or familiar customers with the possibility to "open" new tables for customers who don't have much to share with those already seated. This is the motivation behind the recently introduced distance-dependent CRP (dd-CRP) [42] where table assignments are dependent on direct familiarity or through customers' connections. The distribution of table assignments for customer c_i is now a (for an unoccupied table) and $f(d_{ij})$ to join customer j at her table with similarity distance d_{ij} between customers i and j, and with f standing for the (exponential or logistic) decay function, e.g., $f(d)=\exp(-d/a)$ or $f(d)=\exp(-d+a)/(1+\exp(-d+a))$ [43]. One can expand on dd-CRP using NCM that leverage Levenshtein editing to choose among alternative readings of the attacks (e.g., hallucinate and revise). Another possibility is to have dd-CRP operate in a lower dimensionality manifold generated using spectral clustering. Table assignments would use NCM and TCM-DR, including the reject option characteristic of open set recognition, for impersonation detection (e.g., new customers unfamiliar with those already seated) and have them seated at a new table. Imposters, characteristic of unfamiliar customers (e.g., obfuscation attacks to confuse training), can be reassigned to earlier tables (e.g., known attacks) as more evidence streams in and links are established with customers (e.g., attacks) already seated at tables.

One goes beyond active learning to leverage the relative importance of data instances and build better de-

fenses using importance sampling, characteristic of RANSAC and Random-Hough Transform (RHT) for consensus reasoning methods. The motivation is to pursue evidence accumulation and accrue cues for aggregation and ultimate detection of intrusions. The whole is more than the sum of its parts. RANSAC, characteristic of robust estimation methods, starts with a small but randomly chosen data set of instances and estimates some parametric model that fits data best. It then iteratively enlarges the set with consistent data when possible and re-estimates the model and its error. Starting from different subsets, The Randomized Hough Transform-Support Vector Machine (RHT-SVM) [7] leverages multiple versions of the decision boundary to identify messages that have been mislabeled deliberately, as result of persistent attacks, or not deliberately but due to poor Oracle annotation. The RHT-SVM uses the product of the actual classification label and the average signed distance of an observation from the decision boundary to determine if a training message has been mislabeled. The labels for messages, which on the average appear on the wrong side of the boundary, are flipped and a final SVM model is trained using the modified data. We note that RHT expands on agnostic active (A2) learning [44] [45] that maintains both a current version space and a region of uncertainty. Two data sets, TREC 2007 and CEAS 2008 were used for comparing the performance of RHT-SVM to the performance of Reject On Negative Impact (RONI) [5] as well as the performance of an SVM trained on the tainted training data. To preserve the time ordered nature of the data stream, for each of the data sets the first 10% of the messages are used for training and the remaining 90% of the messages are used for evaluation. Separate adversarial experiments are conducted for flipping spam labels and non-spam labels. For 10 iterations, labels are flipped for a randomly selected subset of 5% of the training data and the final RHT-SVM is evaluated on the test set. RHT-SVM shows an average 9.3% increase in the F1- (harmonic mean of precision and recall) score compared to RONI (99.0% versus 90.6%). The flip sensitivity for RHT-SVM is 95.9% and the flip specificity is 99.0%. It also takes over 90% less time to complete the RHT-SVM experiments compared to the RONI experiments (20 minutes per experiment instead of 360 minutes).

7. Immunity and Detection

The observation that biological immune system (BIS) and information detection systems (IDS) aims are functionally similar is not new. This observation has led to the design of artificial immune systems (AIS) that interface between BIS and IDS using concepts borrowed from evolutionary computation (EC) and genetic algorithms (GA). AIS analogies for IDS are appealing for two reasons in terms of reach and scope. First, BIS provides a high level of protection from invading pathogens but can still fail as one knows very well. Second, computer security techniques "are not able to cope with the dynamic and increasingly complex nature of computer systems and their security" [46]. The immunity characteristics of AIS are not a literal translation of BIS, which is still shrouded in mystery, but rather a conceptual and functional transliteration of BIS. Note that AIS help also with misbehavior detection (e.g., routing misbehavior) in mobile ad-hoc networks [47], recommender systems, and detecting security attacks in software-defined networks (SDN).

Conformal prediction and open set recognition (e.g., TCM-DR) provide a suitable framework to emulate the immunity paradigm. Basically, AIS have to demarcate adversarial attacks that have never been seen before (e.g., non-self) from current self (e.g., normal patterns of behavior and expression). This is accomplished using self-generated antibodies and evaluating them for fitness and avidity to match (e.g., negative selection) continuously morphing antigens (non-self) (e.g., network access patterns) characteristic of pathogens that use clonal selection. Both affinity and avidity are readily available using transduction. Affinity, between detector and specific antigen, is similar to strangeness; avidity, which reflects on the interactions between one detector and all the antigens, is similar to p-values including credibility and confidence for fitness. Together, affinity and avidity support attribute weighting and priorities needed to establish the degree of matching and to facilitate clonal selection including positive selection (PS), when new and advantageous genetic variants sweep a population [48]. A good intrusion detector should have a high non-self avidity and low self-avidity. Additional AIS functionalities that cope with dynamics include apoptosis (e.g., programmed cells death), and the provision of danger SOS signals that indicate damage to self-cells during positive selection. Negative selection and recognition takes place using open set recognition, with innovations on re-identification that seek among others for similar antigens that can be traced to the same source of attack. This leverages consensus reasoning (e.g., RANSAC) and helps with both adversarial learning and clonal selection. The motivation behind re-identification comes from the fact that the multitude of antigens share common characteristics and can be traced to some common source(s)

using amongst others longest common sequence (LCS) (e.g., positive selection) for similarity using dynamic programming and/or RANSAC/Random Hough Transform for realization [7]. Positive selection further helps with better designs for adversarial attacks as they anticipate weak points in defense susceptible to be overwhelmed by persistent attacks.

It is important to know not only what works and to what extent it works, but also to know what does not work and why. This affects AIS (e.g., detection and re-identification). Anecdotal evidence suggests that 90 percent of mass screening errors is due to only 10 percent of the biometric (face) patterns and that even 1% obfuscation and spoofing are enough for cyber security attacks to win. The contribution made by varying attacks patterns on the overall system error is not even. Pattern Specific Error Inhomogeneity (PSEI) analysis [49] shows that the error rates vary across the population being screened according to its diversity. This has led to the jocular characterization of the target population as being composed of "sheep" and "goats." In this characterization, the sheep, for whom classification systems perform reasonably well, dominate the population, whereas the goats, though in a minority, tend to determine the performance of the system through their disproportionate contribution of false reject errors. Impersonation has additional barnyard appellations, which follow from the observed inhomogeneity in performance observed across the population. Specifically, there are some malicious attacks, which have unusually good success at impersonating many different targets. These are called "wolves." There are also some targets that are easy to imitate and thus seem unusually susceptible to many different impersonations. These are called "lambs." PSEI can be addressed using meta-reasoning and transduction [26] to recognize attack patterns that are difficult to defend against, and thus to gate and process them accordingly (see Sect. 8). PSEI spans the analogue of a biometric/forensic menagerie [50]. It expands on the type of attacks and also addresses covariate shift (e.g., changes in the data distributions characteristic of the arm race encountered during evolution) on one side, and change and drift detection, on the other side. PSEI and fraud detection can further leverage reputation and implicit NCM toward deploying one-class SVM [17] in order to mediate between different types of attacks and their constructive resolution.

One can observe and investigate (rather than be constructive and proffer solutions) different automated evasion techniques in the "wild" that "enable malware writers to generate highly variable polymorphic versions of malware that all exploit the same software vulnerability" [51]. Two quantitative measures, similar to evolution and immunity, were proposed for the evaluation of the strength of polymorphic engines: the variation (e.g., diversity) strength and the propagation (e.g., survival) strength. Using these measures, "the authors [51] analyze variability of real shell code samples and claim that the degree of variability attainable by polymorphic engines raises a strong doubt that attacks can ever be modeled by the simple generative approach (*i.e.*, attack signatures) used in many common intrusion detection and antivirus tools" [52]. This could be handled, however, using anomaly detection methods, on one side, and techniques similar to those deployed using PSEI, randomness, and distributed semantic models using vector space representations (see Sect. 12), on the other side. Such observations suggest that negative and clonal selection is ultimately better off using cohort (e.g., context aware) learning and open set recognition rather than the intricacies of polymorphic engines and signature-based intrusion detection.

For completeness we mention that [53] have reported that they built and deployed a coherent, scalable, and extensible real time system, the Facebook Immune System (FIS), to protect users and the social graph (SG) they span. The use of "Immune" refers to overall defense against attackers (e.g., intruders similar to pathogens) without any specific adherence to AIS except a reference to generic mutation. The FIS has two advantages over the attacker: user feedback and global knowledge, something that is not usually pervasive and therefore available to neither AIS nor much of IDS. User feedback for FIS is both explicit and implicit. Explicit feedback includes mark as spam or reporting. Implicit feedback includes deleting a post or rejecting a friend request. Both implicit and explicit feedback is valuable and central to FIS defense. In addition to user feedback, the system has knowledge of aggregate patterns and what is normal and unusual, again something that is not usually available to standard and continuous IDS operation.

Some of the findings and solutions reported by FIS are consonant with the statistics and methods reported throughout this paper about the mode and effectiveness of current attacks. A 2% false-positive rate today on an attack affecting 1000 users is better than a 1% false-positive rate tomorrow on the same attack affecting 100,000 users. As time progresses, attacks mutate and training data becomes less relevant. Similar to learning good is often better than perfect when the complexity involved is lower because it yields better generalization. Optimizing the classification methods or reducing their feature space further improves the classification latency. FIS per-

forms real time checks and classifications on every read and write action. As of March 2011, this was 25B checks per day, reaching 650 K per second at peak.

The effort to design, develop, and deploy FIS has been major, with many Facebook (FB) engineers involved in addition to the authors of the reference paper. FIS findings, which are specific to social media, are helpful overall with both meta-reasoning and meta-recognition (Sect. 8), on one side, and with the immunity aspect discussed in this section, on the other side. Attackers target the social graph in two ways: either by compromising existing graph nodes or by injecting new but fake nodes and relationships. The defense is tasked to protect the graph from attackers who aim to hide patterns and subvert detection. "To be effective, the defense must respond fast and target the features that are most expensive for the attacker to change, being careful to avoid over-fitting on the superficial features that are easy for the attacker to change. The defender seeks to shorten Attack and Detection phases while lengthening the Defense and Mutate stages. The attacker seeks the opposite: to shorten Defense (by obscuring responses and subverting attack canaries) and Mutate while lengthening Attack and Detect" [53]. This illustrates why detection and response latencies are so important for effective defense with any AIS designed to shorten the phases controlled by attackers and to lengthen the phases under defense control. FIS ultimately deploys an integrated IDS that is scalable and responsive to attacks coming from multiple and heterogeneous channels. This affects "the metric interplay between fitness and immunity as response and detection latencies become more important than precision and recall". Damage accumulates quickly [53]. The above interplay and further analogies between AIS and IDS support evolution and co-evolution for both defense and offense whose role are interchangeable, and ultimately affect the effectiveness of the self-protection shield to buttress the defense. Topics of further interest for investigation in the context of adversarial learning include the digital analog of immune disease and immunosuppression using sensitivity analysis driven by cohorts and NCM related concepts.

We note here for completeness that data mining methods have also been used for intrusion detection [54]. Data mining, however, lacks the local estimation and training aspects characteristics of conformal prediction, in general, and transduction, in particular, which provide for locality that reveals specific context, location, and time stamps. One early example for data mining use is audit data analysis and mining (ADAM) system [55] to discover attacks in a TCP dump audit trail using KDD 1999 for test bed and seeking DOS and PROBE attacks. ADAM leverages A Priori association mining to derive (antecedent to consequent) rules of legitimate behavior (e.g., profiles free of attacks) in terms of "normal" frequent item sets. The data mining output augments the rules found with support and confidence indices, which are characteristic of the whole transaction data set T. This is different from the reliability indices for putative class assignments found using TCM-KNN, which correspond to different types of localized attacks. The rules $X \to Y$ found using association mining have support s in the (big data) transaction set T if s% of transactions contains $X \bigcup Y$, and confidence c if c% of transactions that contains X also contains Y.

8. META-Reasoning and META-Recognition

There are two complementary dimensions discussed in this paper. One dimension is about advancing and developing a unified learning framework built around conformal prediction for adversarial learning purposes. The other dimension is about designing and developing a software environment where one can explore and exploit in a coordinated fashion different functional modules that complement each other and address conflicting asymmetries. The proposed integrated environment should expand on the likes of SALT [2] and optimally engage and gate modules that challenge both offense and defense while at the same time enhances and evaluates both. Such an enterprise is supported by meta-reasoning and meta-recognition, whose workings are intertwined. Meta-reasoning plans the best layout and disposition of functional modules, and meta-recognition navigates the maze of data and detection options in order to sort out and rank alternative hypotheses and feasible solutions according to reliability indices and sensitivity analysis. Best defense is multi-prong as contents are multi-varied in reach and scope. This is where re-identification comes in. It is broader than both stand-alone static and/or dynamic recognition and it is incremental in nature. The patterns (e.g., messages) characteristic of adversarial attacks are sporadic rather than continuous in terms of location and time stamps, and can be at times only partial in appearance and disposition. Re-identification [56] is about threading, on one side, and countering disparate and sporadic events, on the other side. This takes place as prior information is lacking and with nothing yet available to track when defenses start looking around for possible attacks.

Isaiah Berlin recalls in his landmark work The Hedgehog and the Fox that "There is a line among the fragments of the Greek poet Archilochus which says: 'the fox knows many things, but the hedgehog knows one big thing. For there exists a great chasm between those, on one side, who relate everything to a single central vision, one system less or more coherent or articulate, in terms of which they understand, think and feel—a single, universal, organizing principle in terms of which alone all that they are and say has significance—and, on the other side, those who pursue many ends, often unrelated and even contradictory, connected, if at all, only in some de facto way, for some psychological or physiological cause, related by no moral or aesthetic principle; these last lead lives, perform acts, and entertain ideas that are centrifugal rather than centripetal, their thought is scattered or diffused, moving on many levels, seizing upon the essence of a vast variety of experiences and objects for what they are in themselves, without consciously or unconsciously, seeking to fit them into, or exclude them from, any one unchanging, all-embracing, sometimes self-contradictory and incomplete, at times fanatical, unitary inner vision. The first kind of intellectual and artistic personality belongs to the hedgehogs, the second to the foxes [57]. The clash between "monist and pluralist", with the latter aware of the permanence of ambiguity and uncertainty, parallels the competition between generative (hedgehog) and discriminative (fox) methods. For modern audience the fox is "divergent" as it displays many not necessarily complementary traits. What is unique are only the philosophical underpinnings for using conformal prediction to learn, on one side, and hedging and punting for demarcation purposes, on the other side, while all along training and querying for annotation and testing are complementary to each other.

Best defense needs to be multi-prong with adversarial learning expected to leverage both contents and context. Contents are multi-varied in reach and scope. Their description starts from raw messages and/or events and moves up the information ladder to include pragmatics and semantics, expected vulnerabilities, and linkages that seek to tie everything together. This is where re-identification comes in. It is about threading and explaining disparate and sporadic events lacking prior information and without anything yet to track when defenses get started to look around for possible attacks. Some functional modules have been motivated and described in the preceding sections. Additional functionalities are introduces in subsequent sections including moving target defense, on-line compression, randomness, and distributed semantic models and vector space representation.

The motivation for meta-reasoning including meta-planning and gating networks draws from anticipation and control, on one side, and context and goals, on the other side. It has been apparent to all that there is no single method for all pattern recognition problems but rather a bag of tools and a bag of problems. Pragmatic and constructive context-aware information fusion supports reliable adversarial learning for intrusion detection and re-identification using principled conformal prediction, in general, and incremental transduction and consensus reasoning for aggregation and stratification, in particular. Inference leverages localization and specialization to combine and deploy expertise. This bears analogies to ensemble of methods, mixtures of experts, and voting machines. Re-identification, which is integral to meta-recognition, accrues evidence for recognition of sporadic (site and time wise) and partial but potentially adversarial events, and supports incremental learning about the adversary. Re-identification helps to interface and mediate between AIS and IDS deployment for the dual purposes of negative and positive selection, on one side, and clonal selection, on the other side.

Meta-reasoning mediates among functional modules while making strategic choices among methods, strategies, and tactics. Adversarial architectures, in general, and IDS architectures, in particular, should be modular and integrated, on one side, and discriminative and incremental in nature, on the other side. The ultimate goal for cyber security is that of deploying a protective defense shield that at its core implements the analogue of autonomic computing [58] and W5+. Autonomic computing, referred to as self-management, provides basic functionalities, e.g., self-configuration (for planning and organization), self-optimization (for efficacy), self-protection (for security purposes), and self-healing (to repair malfunctions and display resilience). 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. The WHO question, about adversarial identity (e.g., source and reputation), is about identity management. Directed evidence accumulation seeks also to explain intent or mal-intent using the WHY question. This question is tasked with linking observations and hypotheses (models) (abducted using analogy reasoning or inferred using Bayesian (belief) networks). The Bayesian networks (for inference and validation purposes) can assist with optimal and incremental/progressive smart data collection, e.g., multi-view integration. In a fashion similar to signal processing and transmission, the incremental aspect signifies continuous access and/or use of crucial evidence, which at some point is enough to solve the IDS "puzzle" and/or make re-identification apparent. Exploration and exploitation, training and detection, active learning, adaptation and co-evolution, can be

threaded according to the confidence placed in prediction outcomes and the calibration of the confidence obtained when using ICP, TCM, and TCM-DR. As an example, Integrated Adaptive Cyber Defense (IACD) where autonomic computing is coupled to human-centric automation (see Sect. 14), promises "to create a healthy cyber ecosystem by automating many risk decisions and optimizing human oversight of security processes too complex or important for machines alone to solve" [59].

Meta-recognition [60], complementary to meta-reasoning, is a post-recognition and score normalization analysis that considers the underlying nature of the prediction sets and their score distribution, and evaluates the extent to which a recognition algorithm succeeds or fails. It can adjust the recognition outcomes if necessary, and it feeds control information to signal to meta-reasoning that a specific response action, e.g., operator intervention or further acquisition of data, is needed. Three basic but different techniques can address the interplay between non-match and matching distributions for putative intrusions: score normalization using cohort analysis and non-conformity measures, statistical extreme value theory (EVT) (without requiring training data), and machine learning. Score normalization essentially leverage varying cohorts, which are not available during real-world operation [61] and the corresponding prediction sets that ICP and its variants yield. The key insight for EVT is that "if the best score is a match, then it should be an outlier with respect to the non-match (tail distribution) model" [60]. P-values and skew discussed earlier are the extreme value solutions that are advanced here using conformal prediction, in general, and transduction, in particular. While machine learning using SVM is reported to perform undertook on cyber security. The motivation for such findings is aligned with the sensible observation already made that a multitude of methods and algorithm instantiations including randomness is needed to face a diversity of attacks. As an example, anomaly detection renders itself to meta-recognition using the choice or (stage-wise cascade or weighted) combination of one-class SVM to account for access to the normal class only, TCM-DR introduced earlier that avails itself of the reject option of open set recognition for anomaly detection of the novel and unfamiliar, the sequential Hausdorff NN conformal anomaly detector (SHNN-CAD) [62] for online learning and sequential anomaly detection, and the discords algorithm characteristic of SAX for time-series [63].

Incremental transduction similar to ICP augments the training set T with sample (*q*, *y*(*q*)) for query q that has been annotated as $y(q)$. One endows the new instance with some reliability index to measure the difference between the T distributions before ($T1$) and after instance augmentation ($T2$) using Kullback-Leibler (KL) divergence (e.g., sensitivity analysis), *i.e.*, $r(q, y(q)) = 1 - KL(T1, T2)$. The strangeness (e.g., NCM) for the instance is $\alpha(q) = 1 - r(q, y(q))$. The reliability index r can be revised using information on class (label), e.g., priors and the sensitivity and specificity of the classification method C used by ICP. Confidence intervals for both prediction and sequential importance sampling are established accordingly. Assume now a regression model where the strangeness for putative label y is defined as $\alpha(y) = p(x)|y - f(x)|/\exp(g(x))$ with regression function f built upon training set T, error of regression function $|y - f(x)|$, estimate of accuracy for regression $g(x)$ built using Support Vector Regression (SVR) from data set $G = \{(x_i, \ln(|y - f(x_i)|)), i = 1, \cdots, N\}$, and with $p(x)$ characteristic of the input data density. Assuming confidence level t and predictive region $\Gamma^t(T, x) = \{y \in Y : p_y > t\}$ or equivalently the set $\{y : \alpha(y) < \alpha a\}$, which is a valid $(1 - a/(N+1)) \times 100\%$ confidence region with $t = \#\{\alpha_i : \alpha_i \geq \alpha_a\}/(N+1)$, the predictive region is then $|y - f(x_{new})| < \alpha_a * \exp(g(x_{new})/p(x_{new}))$. The obtained predictive region will then be smaller for points at which SVR prediction is good and large for points where the prediction is bad, with more confidence in the regions of high input density [64]. Similar sensitivity analysis leverages the observed changes in class posteriors due to guided perturbations and informed edits.

The decision about how to proceed in a step-wise fashion depends on both the current prediction sets (see above) and on management and/or policy considerations. For example, an action in one region (e.g., prediction set) might be more undesirable than in another region. Another example would be applying a more aggressive spam classifier to pages depending on admin preferences. Towards such ends, meta-reasoning includes mechanisms to evaluate classifier performance and leverage in an incremental fashion the prediction sets that are advanced by ICP-like methods (e.g., transduction). Subsequent gating is realized using a multi-layer architecture that includes 1) PSEI (see Sect. 7); 2) Psychology and (behavioral) economics using biases (see Sect. 14); 3) AIS and IDS (see Sect. 7); 4) incremental ICP and sensitivity analysis (see above); 5) stratification using topics (see Sect. 6); and 6) targeted strategies and tactics. Other anti-abuse and adversarial learning problems are likely to benefit from focusing on fast detection and response, sharing data across information channels, and integrated

feedback loops.

Adversarial learning is challenging because attackers can detect defenses and mutate their exploits relatively quickly. Towards that end, an important functionality is that of sequential anomaly detection and mining of trajectories, which is relevant amongst other to moving targets defense (MTD) that modify network environments in response to adversarial activity and persistent threats following reconnaissance undertaken by adversary (e.g., data collection on targets of interest). "Existing methods are not designed for sequential analysis of incomplete trajectories or online learning based on an incrementally updated training set [as ICP and incremental transduction do] and involve ad-hoc thresholds, and may suffer from over fitting and poorly calibrated alarm rates"; the sequential Hausdorff NN conformal anomaly detector (SHNN-CAD) is a sequential "parameter-light anomaly detection that offers a well-founded approach to the calibration of the anomaly threshold" [62] and is consonant with re-identification that threads to fill in for missing or corrupt information. Traffic can also be approached using time series encoded using discrete representations including symbolic aggregate approximation (SAX) like methods that replace standard representations of time series (e.g., DWT, DFT) in order to find discords and most unusual time series subsequences [63]. Complementary to SAX is the search for indexing and mining very large collections of time series (e.g., iSAX2.0) [65] to demarcate self from non-self (e.g., alien DNA). Search and storage can be accommodated to advantage using locality sensitive hashing (LSH) [66].

9. Moving Target Defense

Another desirable functionality for adversarial learning is to leverage the cognitive footprints including their timing left [34] by both legitimate and adversarial users, while at the same time covertly deploying decoys (e.g., hotspots) that do not interfere with the normal operation of the network and/or mobile devices. One now seeks to ensure that the current user is still one of the legitimate ones who were initially authenticated in order to prevent intrusions characteristic of advanced persistent threats (APT). This is similar in concept to moving target defenses (MTD), in general, and to continuous and covert re-authentication (CCA), in particular, it searches among others for "active indicators and corresponding automatic detection tools to ferret out individual with privileged access who are engaged in malicious behaviors such as espionage, sabotage, or violence" and the development of inference enterprise models (IEMs) designed to forecast an enterprise accuracy in detection potential threats [67].

Expert ("voting") methods (e.g., random forests) support the deployment of active indicators for user engagements using command streams [68] or scrolling behaviors [69]. As both MTD and CCA have to cope with change and novelty detection, solutions built around zero-shot learning [70] using one-class SVM, are yet another possibility [71] [72], while the control strategy is to explore and exploit. Similar to decoys, MTD challenges any potential adversary to divine and fathom friend from foe, on one side, and exploit the outcomes the decoys trigger in order to make future engagements (e.g., exploration) better focused and effective. MTD is adaptive by nature. It is complementary to gating in supporting meta-reasoning and meta-recognition, can leverage the conformal prediction sets competing to become effective (e.g., shrink), while at the same time striving for better intrusion detection.

The exploitation dimension seeks for enhanced re-authentication using the analog of recommender system like strategies, e.g., user profiles (e.g., contents filtering) and crowd out sourcing (e.g., collaborative filtering). Similar strategies also support system agility. This "[to engage] in any reasoned modification to a system or environment in response to a functional, performance, or security need" using MTD [73]. System agility is expected to be responsive to W5+ like questions (see Sect. 8), in particular to those of when, what, and how to employ autonomic computing to improve the security of an environment, as well to consider how to measure and weigh the effectiveness of different approaches to agility. Challenges of interest to moving target defenses include concealing the strategy from the adversary (e.g., randomness and subliminal covert challenges), sustaining security across layers, and managing costs including utility and usability (see Sect. 14).

Covert Time Channels (CTC), both active and passive in nature, manipulate the timing of existing traffic (e.g., ordering of network events) for exfiltration purposes. Towards that end, obfuscation takes place when one rather than transmitting at a high rate of speed in order to achieve large bandwidth would deliberately extend the duration of the CTC transmission process. IDS are tasked currently to handle CTC by analyzing deviations from legitimate network traffic statistics. Such approaches, however, are not applicable to highly dynamic and noisy environments because they rely on historical traffic and tedious model training. One could instead use traffic-

mining [62] and/or transduction (with/without i.i.d. assumptions) (see Sect. 10) and its variants. An NCM-CTC combination can model slow CTC and use wavelet-based detection (WBD) [34] that does not require historical traffic data and still yields a high detection rate and a low false positive rate. WBD measures the distance between outbound traffic generated by two virtual machines (VM), one the potentially infected VM and the other a benign VM. The detection metric employs observable variables derived from the discrete wavelet-based multi-resolution transform (DWMT) to measure the variability of the timing difference at all decomposition scale. Yet another possibility is to expand on NCM-CTC using random matrix spectral theory (see Sect. 11).

10. On-Line Compression Models

We expand here the scope of conformal prediction using online compression models (OCM) that start from scratch (e.g., zero-shot learning) and without assuming that the samples are i.i.d. and exchangeable. This supports the adversarial and incremental aspect of the arm race waged between defense and offense. Conformal inductive predictors (CIP) including their validity and efficiency or alternatively their credibility and confidence regarding their prediction outcomes become relevant here. The OCM are lossy in nature and one "can argue that the only information lost is noise, since the summary is required to be a 'sufficient' statistics" [27]. OCM-like summaries include the Gauss and Markov models, which are stronger models than exchangeability. The summary statistics for the Gaussian model are, as one would expect, the mean and variance (or alternatively the first two moments) of the data sequence T seen so far. The non-conformity measure (NCM) for some data sequence $T\left(x_1,\cdots,x_{n+1}\right)$ is $\left|x_{n+1}-mean\left(x_1\cdots x_n\right)\right|$ and the prediction regions Γ^t are built around the t-distribution with $n-1$ degrees of freedom using the t-test with the bonus feature of errors being independent of each other in the online setting [27].

11. Randomness

Another dimension that needs to get embedded in the self-protective shield is that of random matrix theory (RMT). This test the degree and distribution of randomness observed using novel non-conformity measures (NCM) that support demarcation for data lacking proper annotation. A random matrix (RM) is a matrix of given type and size whose entries consist of random numbers from some specified distribution. Random matrix theory, similar to statistical mechanics, can describe the dynamics of socio-economic systems and is a basic tool for financial markets that consist of many components (e.g., stocks) and complex interactions between stock prices (e.g., macroscopic variables $g_i\left(t\right)$). The cross-correlation matrix C (e.g., between stock prices) is time-dependent and varies, with meaningful correlations described by large any random matrices. One can then compare the eigenvalues of C against a "null hypothesis"-a random correlation matrix R constructed from mutually uncorrelated time series [74] [75] and derive NCM of interest. The eigenvectors of C corresponding to the eigenvalues outside their RMT bound display systematic deviations from RMT predictions, with the deviating eigenvectors showing distinct groups (e.g., explaining factors). Such randomness tests support conformal prediction, in general, and adversarial learning, in particular, in their search for demarcation and grouping. Similarity (e.g., agreement and regularity) characteristic of randomness and deviations characteristic of contents can therefore inform on advanced persistent threats (APT) and/or purposeful (spear) attacks. RMT is particularly useful when there is no informative historical data (e.g., traffic) available (e.g., zero-day attacks) and one still seeks to connect the dots for seemingly disparate and sporadic events and patterns. RMT provides also additional insights in problems related to detection (e.g., number of sources embedded in noise) [76]. Yet another use of randomness is to hash and salt, with the latter about random data that augments the hash. There are also links to be leveraged between RMT and Covert Time Channels (CTC). In particular, one would use RMT and/or transduction to detect CTC, which for obfuscation purposes rather than transmitting at a high rate of speed in order to achieve large bandwidth would deliberately extend the duration of the CTC transmission process for information leak purposes. The expanded RMT-NCM-ICP framework can further leverage the benefits accrued from using random matrix spectral theory and conformal prediction (e.g., NCM-ICP).

12. Distributed Semantic Models and Vector Space Representations

Chinese restaurant process (CRP), clustering, and stratification models can use any of TD/IDF, bag of words (BOW), or latent Dirichl *et al.* location (LDA) for message representation. An alternative spectrum of represen-

tations, which has been recently proposed, makes reference to distributional semantic models (DSM) including semantic vector space representations (VSR) that span the universe of classic context-counting (e.g., co-occurrence) and context-predicting (e.g., embedding or neural language models) semantic vectors, with the latter claimed to "obtaining a thorough and resounding victory against their count-based [only] counterparts" [77]. VSR using continuous bag-of-words (CBOW) (e.g., predict a word given the context) and continuous skip-gram (CSG) (e.g., predict context given a word) models [78]. Both CBOW and CSG, characteristic of context-predicting VSR models, have recently become available (e.g., word2vec) from Google using Python. There is also the possibility to combine global matrix factorization (e.g., latent semantic analysis (LSA)) and local context window (e.g., skip-gram) methods (e.g., Global Vector (GloVe) [79]) to enjoy both worlds (e.g., BOW and CBOW). VSR, one among several directions for developing new NCM in support of adversarial learning, facilitates message detection and stratification, on one side, and can overcome message obfuscation using quasi vector operations, on the other side. The justification and motivation come from one moving away from simple scalar or angle distances between word vectors toward measuring instead for varying word analogies and relational similarities using vector arithmetic in the embedded space. The NCM can further encode the difference between predictions and observations, use likelihoods (similar to GloVe) and strangeness to filter out noise, and leverage the complementarity between CBOW and skip-grams. DSM and VSR are similar in their use of analogies for learning with the use of auxiliary side information including higher-order similarities and chorus of prototypes.

13. Experimental Methods and Performance Evaluation

Experimental design and performance evaluation include nested and multiple k-fold cross-validation ((train, tune/validate), test)) for model selection and generalization for ultimate prediction on sequestered data. This includes parameter setting, pruning (e.g., feature selection and dimensionality reduction) for complexity control, and stopping criteria (e.g., using extreme values and/or lowest perplexity). Note that first and foremost one is most interested in model selection and prediction using the expected prediction risk for any method M contemplated. The prediction risk for method M is found as that $\text{prediction risk}(\text{M}) = \text{empirical risk}(\text{M}) * f(d/n)$, with "$d$" the number of degrees of freedom (e.g., parameters) for the prediction method M trained on "n" instances/samples/examples/and "f" a monotonically increasing penalty function. As the number of degrees of freedom increases and the data sample size decreases the prediction risk on sequestered or new test data increases. One therefore seeks to decrease the number of parameters for any given training data set. Additional dimensions of interest include the use of randomized control trials (RCT) to avoid correlation to become causation and substitute foresight to hindsight (e.g., back stacking), and sampling and resampling strategies including balanced training populations using over and under sampling and/or mixed strategies (e.g., 0.632 bootstrap). In addition to the reliability indices for confidence on the decisions associated with conformal prediction (e.g., risk for misdetection), one also employs standard performance metrics including overall accuracy, ROC and AUC, and sensitivity (e.g., recall), precision, specificity, and F1-score. Similarity, which is intrinsic to both metrics and non-conformity measures, includes cosine, Jaccard, Hamming, Hausdorff (e.g., maximum distance from a point in set A to the closest point in set B, quite useful for anomaly detection), and weighted distances. Scale is addressed using vector space representations, (distributed) semantics and topics, and stratification and importance sampling (see Sects. 6 and 12), the latter expanding on active learning using consensus reasoning and guided perturbations that are characteristic of sensitivity analysis (e.g., mutations using flipping annotation and/or revising contents to assess the extent to which the Oracles become affected). Last but not least we recall that what to evaluate is as important as how to evaluate.

Yet another factor affecting overall cyber security and malware detection, which is traced to epidemiology, is the intrinsic relationship between prevalence (e.g., proportion of a population to show a condition of interest) and fallacy (e.g., invalid reasoning) [80]. The corresponding base-rate fallacy and its implications for the difficulty of intrusion detection [81] are examined next. The (Sensitivity, Specificity) and (Recall, Precision) pairs, which are set according to the entries of the confusion matrix (e.g., TP for true positives and FN for false negatives) and serve as useful figures of merit (FOM) to characterize overall intrusion detection performance, An intrusion detection engine with constant sensitivity of 99.9% and specificity of 99.9% would appear to provide excellent performance when just 1% or 10,000 out of 1,000,000 of messages are adversarial attacks. Since the engine is 99.9% sensitive, it will detect 9990 <TP> attacks and miss 10 <FN> attacks. To continue this analysis,

recall that out of one million messages, 990,000 are adversarial attacks. If the specificity is also 99.9%, one can see that 989,010 legitimate messages (e.g., true negatives (TN)) have to be allowed to pass through, while 990 legitimate messages, or approximately 0.1% of the original population, are labeled as false positives (FP) and need to be denied access. What is still needed to complete a meaningful analysis is the prevalence of adversarial attacks in the general population, which is referred to as the prior odds. Assume now that the prevalence for adversarial attacks is 0.1% rather than 1%, *i.e*., there are 1000 rather than 10,000 adversarial attacks. At 99.9% sensitivity, the intrusion detection engine will pick up 999 of them, leaving only one adversarial attack to slip through. Of the 999,000 genuine messages, the detection engine lets through 998,001 of them, and falsely labels 999 of them as attacks. The evaluation yields now the same number of false positive as true positive, and the PPV (+ predictive value) for attacks is now only 50%. Each other message labeled as an attack is a mistake! When the prevalence is 1%, the decision on detection is worth much more because the PPV changes drastically and goes up to 90%, *i.e*., only one tenth rather than half of genuine messages are quarantined. PPV is further affected by sensitivity and specificity changes that are related to the differences in the populations trained on and then screened for detection purposes.

14. Human Factors

Human factors are about tactics, in general, and insights, ingenuity and intuition in particular. Such characteristics are hard to replicate by machines but are most important for human learning, critical thinking, and problem solving [19], with automation expected to amplify human skill and therefore enhance cyber security infrastructure. Human factors should ease, facilitate and induce human "serendipity" in exploring and exploiting alternative scenarios leading to better and original security solutions. The ultimate challenge is to move away from "technology-centered automation toward human-centered automation where the human operator is kept in the decision loop, which is a continuous process of action, feedback and judgment making" [82]. Furthermore, the same author [82] argues that the best decision-support systems provide professionals with "alternative interpretations, hypotheses, or choices." This entails adaptive automation, which employs "cutting-edge sensors and interpretive algorithms to monitor people's physical and mental states, then uses that information to shift tasks and responsibilities between human and computer. When the system senses that an operator is struggling with a difficult procedure, it allocates more tasks to the computer to free the operator of distractions. But when it senses that the operator's interest is waning, it ratchets up the person's workload to capture their attention and build their skills." Conformal prediction supports such adaptive automation using sensitivity analysis driven by reliability indices, randomness, and self-adaptation surrounding active learning.

Another dimension of interest for future research is that usability and security go hand in hand [83]. Most recently (in February 2015), and following recent attacks on Target and Sony, medical insurance giant Anthem was hit by a massive data breach affecting personal information for about 80,000,000 of its customers and employees in one of the largest attacks in corporate history with tens of millions of records stolen. The security breach exposed names, birthdays, and social security numbers among others. It appears that Anthem didn't encrypt the data, "the result of what a person familiar with the matter described as a difficult balancing act between protecting the information and making it useful" [84]. This balancing act reflects on what the National Academy of Engineering (NAE) has recently noted "if security systems are burdensome, people may avoid using them, preferring convenience and functionality to security." Usability and security involves first and foremost human factors and purposeful adaptation to ensure that "usable privacy and security researchers adhere to human activity-centered design to improve system security" and that "usable security research frequently benefits from studying user behavior in the presence of an adversary including deliberate deception" [83]. The self-protective shield, promoted in this paper and characteristic of autonomic computing, would go a long way to lessen the burden of security and increase the usability of the defenses deployed, and ultimately reach "the best security interface that would generally be no security interface." To achieve usability gains, researchers must go beyond adopting human-centered design principles and embrace user decision-making, moving from "command by direction, to command by plan, and ultimately to command by intention" [83]. Additional factors that bear on usability or alternatively utility come from differential privacy (see Sect. 15).

Adversarial learning is complex in nature with many strategies (e.g., overall plans) and tactics (e.g., maneuvers that achieve specific missions) in play. This aspect, related to cognitive biases arising from limits on optimization and/or characteristic of psychology and (behavioral) economics (e.g., systematic deviations from rational-

ity) rather than formal game theory, concerns adversaries (e.g., user) habits and profiles. Examples of biases include anchoring (e.g., focusing on one trait) and attention bias (e.g., focus on preponderance of information), which are triggered by unbalanced populations, base rate fallacy (e.g., ignore base rate information), characteristic of intrusion detection (e.g., seeking for a needle in a haystack), confirmation bias (e.g., tendency to search for information that confirms prior beliefs), hyperbolic discounting (e.g., display stronger preference for payoffs sooner rather than later). First interpretations tend to be decisive including stubbornness in abandoning prior but still cherished beliefs.

Most recently Stahl and Feigenson [85] have shown that babies are better scientists than adults often are. The adults are usually afflicted by confirmation bias and ignore current information most likely to enhance their behavior and better their performance. Human factors and cyber security reach and scope, however, are tightly intertwined and should mimic the babies in their hunger for the unexpected. Similar to Popper the babies seek for those facts that falsify their current theories and understanding about how the world surrounding them works. The unexpected after all enhances both adaptation and exploration geared for choosing what to learn about and what to ignore. Towards such ends the authors show that 11-month-old infants used violations of prior expectations as special opportunities for further exploration and hypothesis testing for learning purposes. “Much as scientists faced with unexpected patterns of data are propelled to think harder, run further experiments, or change their methods of inquiry, untutored preverbal minds are sensitive to conflict between the predicted and the observed, and use this conflict as a scaffold for new learning” [85]. Active learning using the NCM introduced earlier can seek and lock on the unexpected.

Defense and immunity from such cognitive biases further comes from using randomness and/or implementing strategic aims reminiscent of Sun Tzu’s Art of War that advises on competition and conflict. In particular, the conformal prediction framework must be able to anticipate and repair, and to quote from Sun Tzu “deep knowledge is to be aware of disturbance before disturbance, to be aware of danger before danger, to be aware of destruction before destruction, to be aware of calamity before calamity”; and that “by deep knowledge of principle, one can change disturbance into order, change danger into safety, change destruction into survival, change calamity into fortune” [86]. Such lofty aims can be achieved by deploying a self-defensive and protective shield using sensitivity and stability analysis for exploration and exploitation (e.g., control, gating, and navigation), which predates events, determines weaknesses and anticipates vulnerabilities, and self-heals as much as it can before harm is caused. This is similar to what vaccination does in administering antigenic material (e.g., non-self) to stimulate the IDS immune system to develop adaptive immunity to adversarial attacks (e.g., pathogens). The challenge is to heal, repair, and transform the battlefield before calamity strikes.

15. QUO VADIS

Adversarial learning using conformal prediction is a major direction on the road leading to better cyber security but not the end of the road. Worth to note that learning requires both better classification and detection for IDS purposes and better human-centered problem solving skills. Adversarial learning goes beyond cyber security and can serve intelligence analysis on intruders and their threats. This will entail linking between the dots in analogy to solving gumshoe stories that make authentication and detection hard. A majority of detective stories follow the “whodunit” format. The events of the crime and the subsequent events of the investigation are presented such that the reader is provided with enough clues from which the identity of the perpetrator of the crime and her intent may be deduced. The solution is usually not revealed until the final pages of the book. This is similar to expanding re-identification into multiple event re-identification (MERI) to account and detect sporadic and partial events, thread the clues found and trace them to multiple and different causative events and entities while accounting for intrinsic sources of vulnerabilities. An additional outcome that should be sought by adversarial learning is to preclude functional creep and to enable functional interoperability for similar methods across different infrastructure and platforms.

The lessons learned from the 1998 DARPA off-line intrusion detection evaluation still hold today as they did back then. One major challenge, which still waits for feasible but satisfactory solutions, is that further research should focus on developing techniques to find new attacks instead of extending rule-based approaches [54] (see Sect. 7 for ADAM). The conformal prediction framework, open set recognition using TCM-DR for KNN, and the synergy between natural immunity and intrusion detection, are the solutions advanced here to cope with new attacks. Such zero-day attacks can be detected even if training data for them is lacking using zero-shot learning

[87] and self-taught learning [88]. Such an approach draws from cross-modal transfer learning, which leverages large corpora of known attacks and unsupervised learning, side information using domain knowledge and shared distributed semantics (see Sect. 12), and adaptive and smart mutations (see Sect. 7) geared for both better immunity and continuous and targeted testing.

The ultimate challenge is to have cyber security avail itself of a self-protective shield that can engage in precision intrusion detection. Similar to challenges and goals recently set for optimal delivery of precision medical care what cyber security is expected and has to do is to be context-aware and cure the particular service or user engagements from specific intrusions while taking into account the specific infrastructure and platform where the intrusion momentarily takes place rather than handling the open-ended generic "disease" of intrusion detection. It is here and now, it is local, and it is consonant with conformal prediction and the advice proffered by Vapnik [85] that everything is local in nature and one should not solve a problem more general than needed to be. An interesting line of future research concerning lack of information for attacks and their attribution would consider differential privacy, anonymity, and meta-data, with a constant "tension between minimizing privacy loss and maximizing utility" [89]. Loss of privacy concerns amongst others identity theft using spear phishing.

One can pursue important leads for future research by expanding the earlier discussion on immunity and intrusion detection (see Sect. 7) using social network analysis and herd immunity. Towards that end and using the obvious relationships between vaccination and overall network protection using the likes of antivirus protection and firewall policies, one can follow and benefit from the current discussion on how anti-vaccine views hurt herd immunity. In particular, when large segments of a population are immunized against measles, it reduces the risk of exposure in the community, including families who refuse vaccines. The concept is called herd immunity. But when too many healthy people forgo vaccinations, the whole herd becomes more vulnerable, not just those who skipped the shots. "Without vaccines measles, other infectious diseases can proliferate, and people who were previously protected may become imperiled" [90]. Future studies on quantitative and qualitative aspects of cyber security should therefore leverage links between social networks and epidemics using the importance of contact patterns for outcomes including but not limited to the "informatics of complex interacting systems scaling the microbiome to the entire globe" [91] [92].

Epidemics and their spread depend first and foremost of the vaccination rate that confers protection to the herd (e.g., about 93% for measles). Further relevant to this discussion is to recall that immunity is never 100%, that "even people who have gotten the vaccine need the protection provided by the herd to minimize their odds of contracting the disease" (e.g., being afflicted by cyber-attacks), and that the economics of malware attacks require the equivalent of epidemics to justify investment and ensure that significant profits are made. Proliferation is such that "highly infectious diseases like measles ricochet through a community, leaping from person to person until the chain of transmission is interrupted" [90]. Furthermore, "it's not that no one will be infected, but the chain of infections will burn out before it spreads to a large population" and "social mixing fundamentally changes the epidemiological landscape and, consequently, that static network approximations of dynamic networks can be inadequate" [93]. Vaccines "help by breaking the chain of transmission and lowering a disease reproduction rate" [92]. To avoid spread and pandemics the reproduction rate has to stay below 1 and this determines the fraction of the population that needs to be inoculated.

Additional challenges that need to be addressed include better carving of space and time and not necessarily all reductionist in nature. Some things don't bear dissection [94] [95]. Proper layers of representation and execution including mutual dependencies and recurrent feedback are important. The ecosystems of interest for cyber security are intricately complicated with the possibility of indeterminacy always present despite the very ability to improve the odds. Resolution of perennial philosophical questions on forecast ability between description, explanation and prediction, nature vs. nurture, and "holistic conundrums," on one side, and correlation and causality, on the other side, come immediately to mind. Is the future predestined or emergent? It is really the case that to understand the whole all what one needs is to understand its parts or to conflate simpler with smaller when lacking proper layers of description and aggregation? Possible solutions include consensus (e.g., aggregation) and context-aware reasoning that makes problem solving circumscribed and consonant with basic statistical learning theory tenets as expressed in conformal prediction, in general, and transduction, in particular. Additional goals would include investigating integration issues ranging from fusing multiple information channels (e.g., appearance, behavior, location, and profiles) using conformal prediction, best ways and means to combine conformal prediction with deep learning using vector space representations, models that incorporate in a competitive fashion particular strategies and tactics and avoid human biases, and addressing W5+ in an integrated

fashion. One notes in particular that "it is now widely recognized that traditional approaches to cyber defense [including boundary controllers, firewalls, and virus scanners] have been inadequate. Nevertheless, sophisticated adversaries using zero-day exploits are still able to get in and, in many cases, establish a persistent presence, so we need to study and engage the adversary on the defender's turf in order to influence any future moves. The key component in this new paradigm is cyber denial and deception (D & D) [96]. Denial prevents the adversary from capturing information, while deception provides misleading information. Together D & D lessen the amount of reliable information available to an adversary to pursue its malfeasance purposes, whatever they might be. The net result of D & D is increased uncertainty that requires more (precious time) deliberation to move on with attacks, which are less focused and allow the defense to have better and more opportunity to counter the perceived threats. The tradeoffs, which are dynamic in nature and affect D & D effectiveness, are about what to reveal and what to conceal, on one side, and facts or fiction for make believe information, on the other side. The threat-based active defense using D & D can leverage the multi-faceted virtual protective shield described throughout with adversarial learning and conformal prediction engaging both immunity and immunosuppression, on one side, and human factors, on the other side.

16. Conclusions

This paper promotes cyber security using adversarial learning and non-parametric conformal prediction. It advances reliable, robust, and resilient intrusion detection or alternatively consistency in performance and the ability to withstand and recover from adversarial, hostile, and malicious attacks and subterfuge to improve system resilience across networks and computing platforms. This is achieved using strong fundamentals driven by intertwined learning themes that combine statistical learning, semi-supervised learning, and transduction, in general, prediction validity and calibration and reliability indices on outcomes, non-conformity measures and open set recognition, change detection, and active learning. The methodology proposes links and leverages natural immunity and intrusion detection. Venues for future research including economics, epidemics and pandemics, denial and deception, interoperability, and zero-day attacks are motivated and addressed as well. The envisioned software environment would challenge both offense and defense while at the same time enhance defenses for continuous and better cyber security performance.

There is awareness at all times of the permanence of ambiguity, risk, and uncertainty. The infrastructure proposed displays many complementary traits built using conformal prediction to learn, on one side, and to hedge and punt for demarcation purposes, on the other side. The traits include training, querying, and re-training for annotation, detection, disruption, and protection, which are complementary to each other with helping defenses. This comes courtesy of conformal prediction (using non-conformity measures that quantify the resemblance of a data point or event to a particular class) that draws from validation, calibration, and efficiency (e.g., the frequency of prediction errors does not exceed some a priori chosen confidence level in the long run and the prediction outcomes sets are as small as possible), incremental transduction and sensitivity analysis that guarantee the confidence values derived match actual errors, anomaly detection using active learning for self from non-self demarcation, and self-adaptation using randomness.

The objectives here, to enhance and protect cyber security, are multi-pronged. The habitat for both defense and offense is that of an ecosystem where mutual adversarial learning takes place. The functional goals, similar to those espoused by autonomic computing, aim to deploy a self-management protective shield, which coordinates and parcels recurrent activities geared to detect, disrupt, and deny intrusive attacks and subterfuge. Diversity, scale, and utility, which come from using annotation, co-evolution, randomness, and semantics and stratification, raise attack costs to enable and facilitate better defenses. This gains expression in terms of layout and purposeful disposition of functional modules using meta-reasoning and meta-recognition that leverage aggregation, context, traffic, and trajectories, to carve space and time better and to undermine hostile and malicious behavior. In particular, purposeful coordination involves control and gating to navigate the maze of data and detection options, and to sort out and rank alternative hypotheses according to reliability indices, risk, and sensitivity analysis, which are intrinsic to active learning. Significance comes from advancing a science of cyber security built around exploration and exploitation that has learning, anticipation and imputation, control, self-adaptation and self-healing, and prevention work in tandem towards detecting the alien and malicious DNA and making it less harmful as much as possible.

The impact for cyber security is wide open and versatile. Beyond the obvious need to maintain quality of ser-

vice, safeguard proprietary information and privacy, there is also much to gain from straightforward economic well-being. This comes from the simple observation that cyber-crime can be pervasive and relatively facile to undertake, that lack of violence makes cyber-crime less susceptible to further investigation and leads to under-reporting due in part to lack of full awareness, on one side, apprehension not to have reputation compromised, on the other side, the economic harm can be huge when cyber security is not properly safeguarded. Additional applications bear on social and behavioral economics and social network analysis include cyber economics and mobility. Cyber security is multi-faceted as it is expected to protect financial, power, and water grids, and smart identity management [97] interests. It provides in particular immunity to exfiltration and identity theft, protection of intellectual property and e-commerce, and coming and emergent applications geared for mobile wearable devices and personal health surveillance, machine-to-machine (M2M) communication including smart transportation, and Internet of Things (IoT). Cyber security also contributes to big data as it promotes effective and efficient access and leverage of rich lodes of data that might lack proper annotation and safeguard them against disruptive and malicious attacks.

A recent cyber security outlook for the future reports on six potential cyber game changers and priorities for future research [98]. The game changers include new computing paradigms and new territories for network complexity for cyber environment changes, big data analytics and resilient self-adaptation for technology trends, and mixed-trust systems and active defenses for cyber technology breakthroughs. This paper involves many of the changers contemplated. This includes adversarial learning using conformal prediction and immunity for new computing paradigms, meta-reasoning and meta-recognition to support the emergence of a self-protective management shield for resilient self-adaptation, and human factors for mixed-trust systems and usability. Additional directions and factors for promising and rewarding future R & D are discussed throughout and in particular in the preceding Sect. 15 on Quo Vadis.

Last but not least any advances in cyber security are predicated on some astute remarks expressed recently about the computer security industry itself. Among them Amit Yoran from RSA makes users aware that "first, much commercial software is riddled with flows, making it tricky for any technology to make it secure after the fact" and "second, security companies are competitive, and pitch their own solutions as a cure-all. Executives are loath to suggest that customers would be more secure if they used other technologies as well [99]". This calls for the deployment of secure and sustainable cyber ecosystems, in general, and similar to autonomic computing seeking for "computer architectures that are engineered from the foundation to promote hardware-enhanced security-for example, creating combined hardware-software architecture to support self-protecting data, secure enclaves for executing trusted software components, and new hypervisor models for more security in cloud environments. A more fundamental goal is to engineer secure hardware than can itself limit security breaches, such as the cache side-channel attacks that today's cache architectures allow" [100].

References

[1] Olmedo, O., Zhang, J., Wechsler, H., Poland, A. and Borne, K. (2008) Automatic Detection and Tracking of Coronal Mass Ejections (CMEs) in Coronagraph Time Series. *Solar Physics*, **248**, 485-499. http://dx.doi.org/10.1007/s11207-007-9104-5

[2] Miller, B., Kantchelian, A., Afroz, S., Bachwani, R., Dauber, E., Huang, L., Tschantz, M.C., Joseph, A.D. and Tygar J.D. (2014) Adversarial Active Learning. *Proceeding of the* 2014 *Workshop on Artificial Intelligent and Security Workshop AI*, Scottsdale, 3-7 November 2014, 3-14. http://dx.doi.org/10.1145/2666652.2666656

[3] Thiel, C. (2008) Classification on Soft Labels Is Robust Against Label Noise. *Proceeding of the* 12*th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, Zagreb, 3-5 September 2008, 65-73. http://dx.doi.org/10.1007/978-3-540-85563-7_14

[4] Tygar, J.D. (2011) Adversarial Machine Learning. *IEEE Internet Computing*, **15**, 4-6. http://dx.doi.org/10.1109/MIC.2011.112

[5] Nelson, B., Barreno, M., Chi, F.J., Joseph, A.D., Rubinstein, B.I.P., Saini, U., Sutton, C., Tygar, J.D. and Xia, K. (2008) Exploiting Machine Learning to Subvert Your Spam Filter. *Proceeding of* 1*st Usenix Workshop on Large Scale Exploits and Emergent Threats*, San Francisco, 15 April 2008, 1-9.

[6] Bootkrajang, J. and Kaban, A. (2012) Label-Noise Robust Logistic Regression and Its Applications. *Proceedings of the* 2012 *European Conference on Machine Learning and Knowledge Discovery in Database*, Bristol, 24-28 September 2012, 143-158. http://dx.doi.org/10.1007/978-3-642-33460-3_15

[7] DeBarr, D., Sun, H. and Wechsler, H. (2013) Adversarial Spam Detection Using the Randomized Hough Transform-

Support Vector Machine. *Proceedings of the IEEE International Conference on Machine Learning and Applications* (*ICMLA*), Miami, 4-7 December 2013, 299-304. http://dx.doi.org/10.1109/icmla.2013.61

[8] Basit, N. and Wechsler, H. (2011) Function Prediction for in Silico Protein Mutagenesis Using Transduction and Active Learning, *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, Atlanta, 12-15 November 2011, 939-940. http://dx.doi.org/10.1109/bibmw.2011.6112511

[9] Sculley, D., Otey, M.E., Pohl, M., Spitznagel, B., Hainsworth, J. and Zhou, Y. (2011) Detecting Adversarial Advertisements in the Wild. *Proceeding of the* 17*th ACM SKIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, 21-24 August 2011, 274-282.

[10] Biggio, B., Fumera, G. and Roli, F. (2014) Security Evaluation of Pattern Classifiers under Attack. *IEEE Transaction on Knowledge and Data Engineering*, **26**, 984-996. http://dx.doi.org/10.1109/TKDE.2013.57

[11] Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B. and Tygar, J.D. (2011) Adversarial Machine Learning. *Proceedings of the* 4*th Workshop on Artificial Intelligence and Security and Artificial Intelligence*, Chicago, 17-21 October 2011, 43-57. http://dx.doi.org/10.1145/2046684.2046692

[12] Vorobeychik, Y. and Li, B. (2014) Optimal Randomized Classification in Adversarial Settings. *International Conference on Autonomous Agents and Multi-Agents Systems* (*AAMAS*), Paris, 5-9 May 2014, 485-492.

[13] DeBarr, D. and Wechsler, H. (2012) Spam Detection using Random Boost. *Pattern Recognition Letters*, **33**, 1237-1244. http://dx.doi.org/10.1016/j.patrec.2012.03.012

[14] DeBarr, D. and Wechsler, H. (2009) Spam Detection Using Clustering, Random Forests, and Active Learning. *Proceedings of the* 6*th Conference on E-Mail and Anti-Spam* (*CEAS*), Mountain View, 16-17 July 2009, 16-17.

[15] DeBarr, D. and Wechsler, H. (2010) Using Social Network Analysis for Spam Detection. *Proceedings of the* 3*rd Conference on Social Computing, Behavioral Modeling and Prediction* (*SBP*), Bethesda, 30-31 March 2010, 62-69. http://dx.doi.org/10.1007/978-3-642-12079-4_10

[16] DeBarr, D., Ramanathan, V. and Wechsler, H. (2013) Phishing Detection Using Traffic Behavior, Spectral Clustering, and Random Forests. *Proceedings of the IEEE International Conference on Intelligence and Security Informatics* (*ISI*), Seattle, 4-7 June 2013, 67-72. http://dx.doi.org/10.1109/isi.2013.6578788

[17] DeBarr, D. and Wechsler, H. (2013) Fraud Detection Using Reputation Features, SVMs, and Random Forests. *Proceedings of the* 9*th International Conference on Data Mining*, Las Vegas, 22-25 July 2013, 238-244.

[18] Cherkassky, V. and Mulier, F. (2007) Learning from Data. 2nd Edition, Wiley, Hoboken. http://dx.doi.org/10.1002/9780470140529

[19] Vapnik, V. (1998) Statistical Learning Theory. Springer, Berlin.

[20] Chapelle, O., Scholkopf, B. and Zien, A. (Eds.) (2006) Semi-Supervised Learning. MIT Press, Cambridge. http://dx.doi.org/10.7551/mitpress/9780262033589.001.0001

[21] Wechsler, H. and Li, F. (2014) Biometrics and Face Recognition. In: Balasubramanian, V., Ho, S.S. and Vovk, V., Eds., *Conformal Predictions for Reliable Machine Learning*: *Theory*, *Adaptations*, *and Applications*, Elsevier, Amsterdam, 189-215. http://dx.doi.org/10.1016/B978-0-12-398537-8.00010-9

[22] Ho, S.S., and Wechsler, H. (2010) A Martingale Framework for Detecting Changes in The Data Generating Model in Data Streams. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **32**, 2113-2127. http://dx.doi.org/10.1109/TPAMI.2010.48

[23] Ho, S.S., and Wechsler, H. (2014) On Line Change Detection Using Exchangeability. In: Balasubramanian, V., Ho, S.S. and Vovk, V., Eds., *Conformal Predictions for Reliable Machine Learning*: *Theory*, *Adaptations*, *and Applications*, Elsevier, Amsterdam, 99-114. http://dx.doi.org/10.1016/B978-0-12-398537-8.00005-5

[24] Proedrou, K., Nouretdinov, I., Vovk, V. and Gammerman, A. (2002) Transductive Confidence Machine for Pattern Recognition. *Proceeding of the* 13*th European Conference on Machine Learning*, Royal Holloway, 19-23 August 2002, 81-390. http://dx.doi.org/10.1007/3-540-36755-1_32

[25] Vapnik,V. (2000) The Nature of Statistical Learning Theory. 2nd Edition, Springer, New York. http://dx.doi.org/10.1007/978-1-4757-3264-1

[26] Li, F. and Wechsler, H**.** (2005) Open Set Face Recognition Using Transduction. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **27**, 1686-1698. http://dx.doi.org/10.1109/TPAMI.2005.224

[27] Vovk, V., Gammerman, A. and Shafer, G. (2005) Algorithmic Learning in a Random World. Springer, Berlin.

[28] Wechsler, H. and Ho, S.S. (2011) Evidence-Based Management of Data Collection and Decision-Making Using Algorithmic Randomness and Active Learning. *Journal of Intelligent Information Management*, **3**, 142-159. http://dx.doi.org/10.4236/iim.2011.34018

[29] Li, Y., Fang, B., Guo, L. and Chen, Y. (2007) Network Anomaly Detection based on TCM-KNN Algorithm. *Proceeding of the* 2*nd ACM Symposium on Information, Computer and Communications Security*, Singapore, 20-22 March

2007, 13-19. http://dx.doi.org/10.1145/1229285.1229292

[30] Li, Y. and Guo, L. (2007) An Efficient Network Anomaly Detection Scheme Based on TCM-KNN Algorithm and Data Reduction Mechanism. *Proceeding of the IEEE Workshop on Information Assurance*, West Point, 20-22 June 2007, 221-227. http://dx.doi.org/10.1109/iaw.2007.381936

[31] Ho, S.S. and Wechsler, H. (2008) Query by Transduction. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **30**, 1557-1571. http://dx.doi.org/10.1109/TPAMI.2007.70811

[32] Balasubramanian, V., Chakraborty, S., Ho, S.S., Wechsler, H. and Panchanathan, S. (2014) Active Learning. In: Balasubramanian, V., Ho, S.S. and Vovk, V., Eds., *Conformal Predictions for Reliable Machine Learning: Theory, Adaptations, and Applications*, Elsevier, Amsterdam, 49-70. http://dx.doi.org/10.1016/B978-0-12-398537-8.00003-1

[33] Li, Y., Guo, L., Fang, B.X., Tian, Z.H. and Zhang, Y.Z. (2008) Detecting DoS Attacks Against Web Server via Lightweight TCM-KNN Algorithm. *Proceeding of the ACM SIGCOMM* 2008 *Conference on Data Communication*, Seattle, 17-22 August 2008, 497-498.

[34] Liu, A., Chen, j.X. and Wechsler, H. (2013) Real-Time Covert Timing Channels Detection in a Networked Virtual Environment. *Proceeding of the* 9*th Annual International Federation for Information Processing*, Orlando, 28-30 January 2013, 273-288. http://dx.doi.org/10.1007/978-3-642-41148-9_19

[35] Li, Y. and Guo, L. (2007) An Active Learning based TCM-KNN Algorithm for Supervised Network Intrusion Detection. *Computers and Security*, **26**, 459-467. http://dx.doi.org/10.1016/j.cose.2007.10.002

[36] Basseville, M. and Nikiforov, I.V. (1993) Detection of Abrupt Changes: Theory and Application, 104. Prentice Hall, Englewood Cliffs.

[37] Ramanathan, V. and Wechsler, H. (2013) Phishing Detection and Impersonated Entity Discovery Using Conditional Random Field and Latent Dirichlet Allocation. *Computer and Security*, **34**, 123-139. http://dx.doi.org/10.1016/j.cose.2012.12.002

[38] Ramanathan, V. and Wechsler, H. (2012) PhishGILLNET—Phishing Detection Methodology Using Probabilistic Latent Semantic Analysis, AdaBoost, and Co-Training. *EURASIP Journal of Information Security*, **2012**, 1. http://dx.doi.org/10.1186/1687-417X-2012-1

[39] Freund, Y. and Shapire, R.E. (1996) Experiments with a New Boosting Algorithm. *Proceeding of* 13*th International Conference on Machine Learning* (*ICML*), Bari, 3-6 July 1996, 148-156.

[40] Blum, A. and Mitchell, T. (1998) Combining Labeled and Unlabeled Data with Co-Training. *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, 24-26 July 1998, 92-100. http://dx.doi.org/10.1145/279943.279962

[41] Ramanathan, V. and Wechsler, H. (2012) Phishing Website Detection using Latent Dirichlet Allocation and AdaBoost. *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, Washington, 11-14 June 2012, 102-107. http://dx.doi.org/10.1109/isi.2012.6284100

[42] Blei, D.M. and Frazier, P. (2010) Distance Dependent Chinese Restaurant Process. *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*), Haifa, 21-24 June 2010, 87-94.

[43] Sun, H., Chen, J.X. and Wechsler, H. (2014) A New Segmentation Method for Broadcast Sports Video. *Proceedings of the* 8*th International Conference on Frontier of Computer Science and Technology* (*FCST*), Chengdu, 19-21 December 2014, 1789-1793. http://dx.doi.org/10.1109/cse.2014.328

[44] Balcan, M.F., Beygelzimer, A. and Langford, J. (2009) Agnostic Active Learning. *Journal of Computer and System Sciences*, **75**, 78-89. http://dx.doi.org/10.1016/j.jcss.2008.07.003

[45] Balcan, M.F., Beygelzimer, A. and Langford, J. (2006) Agnostic Active Learning. *Proceedings of the International Conference on Machine Learning* (*ICML*), Pittsburgh, 25-29 June 2006, 65-72. http://dx.doi.org/10.1145/1143844.1143853

[46] Kim, J., Bentley, P., Aiklelin, U., Greesmith, J., Tedesco, G. and Twycross, J. (2007) Immune System Approaches to Intrusion Detection—A Review. *Natural Computing*, **6**, 413-466. http://dx.doi.org/10.1007/s11047-006-9026-4

[47] Boudec, J.Y. and Sarafijanovic, S. (2004) An Artificial Immune System Approach to Misbehavior Detection on Mobile Ad-Hoc Networks, *Proceeding of Biologically Inspired Approaches to Advanced Information Technology*, Lausanne, 29-30 January 2004, 96-111. http://dx.doi.org/10.1007/978-3-540-27835-1_29

[48] Tang, W., Yang, X.M., Xie, X., Peng, L.M., Youn, C.H. and Cao, Y. (2010) Avidity-Model based Clonal Selection Algorithm for Network Intrusion Detection. *Proceedings of the* 18*th International Workshop on Quality of Service* (*IWQoS*), Beijing, 16-18 June 2010, 1-5. http://dx.doi.org/10.1109/iwqos.2010.5542731

[49] Doddington, G.R., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D. (1998) Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance. *Proceedings of* 5*th International Conference Spoken Language Processing*, Sydney, 30 November-4 December 1998, 1351-1354.

[50] Yager, N. and Dunstone, T. (2010) The Biometric Menagerie. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **32**, 220-230. http://dx.doi.org/10.1109/TPAMI.2008.291

[51] Song, Y., Locasto, M.E., Stavrou, A., Keromytis, A.D. and Stolfo, S.J. (2010) On the Infeasibility of Modeling Polymorphic Shell Code. *Machine Learning Journal*, **81**, 179-205. http://dx.doi.org/10.1007/s10994-009-5143-5

[52] Laskov, P. and Lippmann, R. (2010) Machine Learning in Adversarial Environments. *Machine Learning Journal*, **81**, 115-119. http://dx.doi.org/10.1007/s10994-010-5207-6

[53] Stein, T., Chen, E. and Mangla, K. (2011) Facebook Immune System. *Proceeding of the* 4th *Workshop on Social Network Systems* (*SNS*), Salzburg, 10-13 April 2011, 1-8. http://dx.doi.org/10.1145/1989656.1989664

[54] Lippmann, R. *et al*. (2000) Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation. *Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX)*, Los Alamitos, 25-27 January 2000, 12-26.

[55] Barbara, D., Couto, J., Jajodia, S., Poypack, L. and Wu, N. (2001) ADAM: Detection Intrusions by Data Mining. *Proceedings of the IEEE Workshop on Information Assurance and Security*, West Point, 5-6 June 2001, 11-16. http://dx.doi.org/10.1145/604264.604268

[56] Nappi, M. and Wechsler, H. (2012) Robust Re-Identification Using Randomness and Statistical Learning: Quo Vadis. *Pattern Recognition Letters*, **33**, 1820-1827. http://dx.doi.org/10.1016/j.patrec.2012.02.005

[57] Berlin, I. (1953) The Hedgehog and the Fox. Weidenfeld & Nicolson, London.

[58] Ganek, A. and Corbi, T. (2003) The Dawning of The Autonomic Computing Era. *IBM Systems Journal*, **42**, 5-18. http://dx.doi.org/10.1147/sj.421.0005

[59] Fonash, P. and Schneck, P. (2015) Cybersecurity: From Months to Milliseconds. *Computer*, **48**, 42-49. http://dx.doi.org/10.1109/MC.2015.11

[60] Scheirer, W.J., Rocha, A., Parris, J. and Boult, T.E. (2012) Learning for Meta-Recognition. *IEEE Transactions on Information Forensics and Security*, **7**, 1214-1224. http://dx.doi.org/10.1109/TIFS.2012.2192430

[61] Wechsler, H. (2007) Reliable Face Recognition Methods. Springer, New York. http://dx.doi.org/10.1007/978-0-387-38464-1

[62] Laxhammar, R. and Falkman, G. (2014) On-Line Learning and Sequential Anomaly Detection in Trajectories. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **36**, 1158-1173. http://dx.doi.org/10.1109/TPAMI.2013.172

[63] Keogh, E., Lin, J. and Fu, A. (2005) HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. *Proceeding of the* 5*th IEEE International Conference on Data Mining* (*ICDM*), Houston, 27-30 November 2005, 226-233. http://dx.doi.org/10.1109/ICDM.2005.79

[64] Nischenko, I. and Jordaan, E.M. (2006) Confidence of SVM Predictions Using a Strangeness Measure. *Proceedings of the International Joint Conference on Neural Networks* (*IJCNN*), Vancouver, 16-21 July 2006, 1239-1246.

[65] Camerra, A., Palpanas, T., Shieh, J., and Keogh, E. (2010) *i*SAX 2.0: Indexing and Mining One Billion Time Series, *Proceeding of the* 10*th IEEE International Conference on Data Mining* (*ICDM*), Sydney, 13-17 December 2010, 58-67. http://dx.doi.org/10.1109/icdm.2010.124

[66] Leskovec, J., Rajaraman, A. and Ullman, J.D. (2015) Mining of Massive Data Sets. 2nd Edition, Cambridge University Press, Cambridge.

[67] Rockwell, M. (2015) IARPA Eyes Insider-Threat Tech. http://fcw.com/articles/2015/03/30/iarpa-insider-tech.aspx

[68] El Masri, A., Likarish, P., Wechsler, H. and Kang, B.B. (2014) Identifying Users with Application-Specific Command Streams. *Proceedings of the* 12*th International Conference on Privacy, Security and Trust* (*PST* 2014), Toronto, 23-24 July 2014, 232-238. http://dx.doi.org/10.1109/pst.2014.6890944

[69] El Masri, A., Likarish, P., Wechsler, H. and Kang, B.B. (2015) Active Authentication Using Scrolling Behaviors. *Proceedings of the* 6*th IEEE International Conference on Information and Communication Systems* (*ICICS* 2015), Amman, 7-9 April 2015, 257-262. http://dx.doi.org/10.1109/IACS.2015.7103185

[70] Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C.D. and Ng, A.Y. (2013) Zero-Shot Learning through Cross-Modal Transfer. *Proceedings of the Advances in Neural Information Processing Systems* (*NIPS*), **26**, Lake Tahoe, 5-10 December 2013, 935-943.

[71] Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J. and Platt, J.C. (2000) Support Vector Machine for Novelty Detection. MIT Press, Cambridge, 582-588.

[72] Tax, D.M.J. and W Duin, R.P. (2004) Support Vector Data Description. *Machine Learning*, **54**, 45-66. http://dx.doi.org/10.1023/B:MACH.0000008084.60811.49

[73] McDaniel, P., Jaeger, T., La Porta, T.F., *et al*. (2014) Security and Science of Agility. *Proceedings of the* 1*st ACM Workshop on Moving Target Defense*, Scottsdale, 3-7 November 2014, 13-19.

http://dx.doi.org/10.1145/2663474.2663476

[74] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L., Guhr, T. and Stanley, H.E. (2002) A Random Matrix Theory Approach to Quantifying Collective Behavior of Stock Price Fluctuations. *Empirical Science of Financial Fluctuations*, **88**, 35-40. http://dx.doi.org/10.1007/978-4-431-66993-7_5

[75] Rosenow, B. (2005) DPG-School on Dynamics of Socio-Economic Systems. Bad Honnef, Germany.

[76] Kritchman, S. and Nadler, B. (2009) Non-Parametric Detection of the Number of Signal: Hypothesis Testing and Random Matrix Theory. *IEEE Transactions on Signal Processing*, **57**, 3930-3941. http://dx.doi.org/10.1109/TSP.2009.2022897

[77] Baroni, M., Dinu, G. and Kruszewski, G. (2014) Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors. *Proceeding of the* 25*nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 23-25 June 2014, 238-247. http://dx.doi.org/10.3115/v1/p14-1023

[78] Mikolov, T., Chen, K., Conrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. *Proceedings of the Workshop at ICLR*, Scottsdale, 2-4 May 2013, 1-12.

[79] Pennington, J., Socher, R. and Manning, C.D. (2014) GloVe: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing* (*EMNLP* 2014), Doha, 26-28 October 2014, 1532-1543. http://dx.doi.org/10.3115/v1/d14-1162

[80] Kraemer, H.C. (1992) Evaluating Medical Tests: Objectives and Quantitative Guidelines. Sage Publication, Thousand Oaks.

[81] Axelsson, S. (1999) The Base-Rate Fallacy and Its Implications for The Difficulty of Intrusion Detection. *Proceedings of the* 6*th ACM Conference on Computer and Communications Security*, Singapore, 1-4 November 1999, 1-7. http://dx.doi.org/10.1145/319709.319710

[82] Carr, N. (2014) http://online.wsj.com/articles/automation-makes-us-dumb-1416589342

[83] Cranor, L.F. and Buchler, N. (2015) Better Together: Usability and Security Go Hand in Hand. *IEEE Security and Privacy*, **12**, 89-93. http://dx.doi.org/10.1109/MSP.2014.109

[84] Yadron, D. and Beck, M. (2015) Investigators Eye China in Anthem Hack. http://www.wsj.com/articles/investigators-eye-china-in-anthem-hack-1423167560?mod=WSJ_hp_LEFTWhatsNewsCollection

[85] Stahl, A.E. and Feigenson, L. (2015) Observing the Unexpected Enhances Infants' Learning and Exploration. *Science*, **348**, 91-94. http://dx.doi.org/10.1126/science.aaa3799

[86] Sun, T. (1988) The Art of War. Thomas, C. (Translator), Shambhala Publications, Boston & London.

[87] Rohrbach, M., Stark, M. and Schiele, B. (2011) Evaluating Knowledge Transfer and Zero-Shot Learning in A Large Scale Setting. *Proceedings of the Computer Vision and Pattern Recognition* (*CVPR*), Colorado Springs, 20-25 June 2011, 1641-1648. http://dx.doi.org/10.1109/cvpr.2011.5995627

[88] Raina, R., Battle, A., Lee, H., Packer, B. and Ng, A.Y. (2007) Self-Taught Learning: Transfer Learning from Unlabeled Data. *Proceedings of the* 24*th International Conference on Machine Learning*, Corvalis, 20-24 June 2007, 759-766. http://dx.doi.org/10.1145/1273496.1273592

[89] Dwork, C. (2009) The Differential Privacy Frontier. *Proceedings of the* 6*th Theory of Cryptography Conference* (*TCC*), San Francisco, 15-17 March 2009, 496-502.

[90] McGinty, J.C. (2015) How Anti-vaccine Views Hurt Herd Immunity. *Wall Street Journal*. http://en.wikipedia.org/wiki/Self-information

[91] Eubanks, S. (2003) Social Networks and Epidemics. http://silver.ima.umn.edu/talks/workshops/11-3-6.2003/eubank/eubank.html

[92] Eubank, S., Kumar, V.S., Marathe, M., Srinivasan, A. and Wang, N. (2006) Structure of Social Contact Networks and Their Impact on Epidemics. *AMS-DIMACS Special Volume on Epidemiology*, **70**, 181-213.

[93] Volz, E. and Meyers, L.A. (2009) Epidemic Thresholds in Dynamic Contact Networks. *Journal of The Royal Society Interface*, **6**, 233-241. http://dx.doi.org/10.1098/rsif.2008.0218

[94] Lewontin, R. (2000) The Triple Helix. Harvard University Press, Cambridge.

[95] Rothman, S. (2002) Lessons from the Living Cell: The Limits of Reductionism. McGraw-Hill, New York.

[96] Heckman, K.E., Stech, F.J., Schmocker, B.S. and Thomas, R.K. (2015) Denial and Deception in Cyber Defense. *Computer*, **48**, 36-44. http://dx.doi.org/10.1109/mc.2015.104

[97] Wechsler, H. (2012) Biometrics, Forensics, Security, and Privacy using Smart Identity Management and Interoperability: Validation and Vulnerabilities of Various Techniques. *Review of Policy Research*, **29**, 63-89. http://dx.doi.org/10.1111/j.1541-1338.2011.00538.x

[98] Kott, A., Swami, A. and McDaniel, P. (2014) Security Outlook: Six Cyber Game Changers for the Next 15 Years. *Computer*, **47**, 104-106. http://dx.doi.org/10.1109/MC.2014.366

[99] Yoran, A. (2015) Computer-Security Industry Critiques Itself Following High-Profile Breaches. http://www.wsj.com/articles/computer-security-industry-critiques-itself-following-high-profile-breaches-1429573277

[100] Lee, R.B. (2015) Rethinking Computers for Cyber Security. *Computer*, **48**, 16-25. http://dx.doi.org/10.1109/MC.2015.118