F-SECURE®

White Paper

MLX: Machine Learning to Beat Spam Today and *Tomorrow*

WHITE PAPER
SEPTEMBER 2005

Executive Summary

As spammers employ increasingly sophisticated techniques to avoid detection by the anti-spam rules and heuristics currently in widespread use, these simplistic anti-spam solutions have become less and less effective. Once again, users' email inboxes are flooded with spam and critical emails are being inadvertently discarded or simply lost in the noise. Clearly, a new approach is needed to defend corporate messaging infrastructures and reclaim email's value as a corporate communications medium.

Mounting an effective defense against spam requires detection techniques that can evolve as quickly as the attacks themselves. Without the ability to automatically adapt to detect new types of threats, an anti-spam solution will always be a step behind the spammers. Proofpoint MLX™ technology leverages machine learning techniques to provide a revolutionary spam detection system that analyzes millions of messages to automatically adjust its detection algorithms to identify even the newest spam attacks without manual tuning or administrator intervention.

Proofpoint employs a full range of classification methods, from legacy approaches such as heuristics and Bayesian analysis, to state-of-the-art machine learning algorithms such as those used in genomic sequence analysis.

Unlike other anti-spam solutions, the Proofpoint platform's ability to defend against spam attacks does not degrade over time. Proofpoint's MLX technology enables Proofpoint solutions to evolve to counter emerging threats, ensuring that the corporate infrastructure is secure against tomorrow's spammers as well as today's.

Why Does MLX Matter?

Proofpoint's MLX-based solutions provide the most effective spam detection available today:

- Accurate: Proofpoint's Machine Learning Technology, based on logistic regression and support vector machines, provides the foundation for a powerful, adaptive anti-spam solution capable of analyzing over 20 layers and more than 200,000 attributes to accurately differentiate between spam and valid messages.
- Decisive: Traditional anti-spam solutions evaluate a limited number of attributes and are unable to decisively classify spam, which leads to a low rate of effectiveness and a high rate of false positives.

MLX ensures that Proofpoint's solutions will remain effective against the tactics spammers try to employ tomorrow:

- Predictive: Continuously-evolving spamming techniques can only be countered by a predictive solution capable of learning and self-adjusting. Traditional reactive approaches just can't keep pace.
- Adaptive: Proofpoint's MLX-based solution automatically adapts to counter new threats. As more data from both valid email and spam is added to the machine learning model, the system identifies and weights relevant attributes to automatically tune the classification process. The result is a system that is just as effective at identifying tomorrow's spam as it is at identifying spam today.

Proofpoint is the only vendor that has successfully combined Machine Learning techniques with traditional approaches to achieve near-perfect spam detection. Ongoing efforts by Proofpoint's Anti-Spam Laboratory scientists and Technical Advisory Board secure Proofpoint's position as a technology pioneer and industry-leader in the fight against spam.

The Need for Machine Learning

Defending messaging systems against today's spammers requires an intelligent system that can automatically adapt as the attackers' techniques evolve. Unlike yesterday's anti-spam technologies, Proofpoint's MLX technology enables Proofpoint solutions to counter new spam techniques as they emerge, defending messaging systems against tomorrow's threats as well as today's.

Traditional anti-spam solutions are reactive—they compare new messages to known spam, simply looking for words, phrases, and other attributes previously encountered in spam, and flag messages from "known" spammers. These technologies cannot adapt quickly enough to detect new threats and are losing ground against increasingly sophisticated spam attacks.

Implication for users: they are not very good today, because they will have either low effectiveness or low accuracy, and they will be worse tomorrow. Implication for IT users: some of those techniques require much maintenance time, which will only increase over time.

Yesterday's Anti-Spam Technologies

Technique	Description	Limitations
Spam Signature Detection	Compare messages to known spam.	• Minor modifications thwart detection, and spammers know this. • Can't detect new threats—always a step behind spammers.
Challenge – Response	Require sender to respond.	• Challenge is offensive in business context. • Misclassifies valid, automatically-generated email.
Text-Pattern Matching	Search for spam keywords such as *Viagra* or *enlargement*.	• Difficult to manage large keyword lists. • Simplicity leads to high false positives. • Effectiveness plummets as new types of spam emerge.
Heuristics or Naïve Bayesian	Apply *rules of thumb* to assign a spam score.	• Based on a small number of independently-assessed attributes • Large administration overhead for manually-tuned systems. • Naïve Bayes models ignore attribute dependencies and systematically under- or over-estimate spam probabilities. This is a key shortcoming of the technology, as will be shown later.
Community Resources	Check messages against RBLs and other public anti-spam resources	• Spammers also use this information to thwart detection. • It takes time to query the network, reducing performance for enterprise-wide usage

Proofpoint's MLX machine learning technology provides the next generation in spam detection today—a highly-effective, intelligent solution that can adapt to detect new types of spam with minimal intervention.

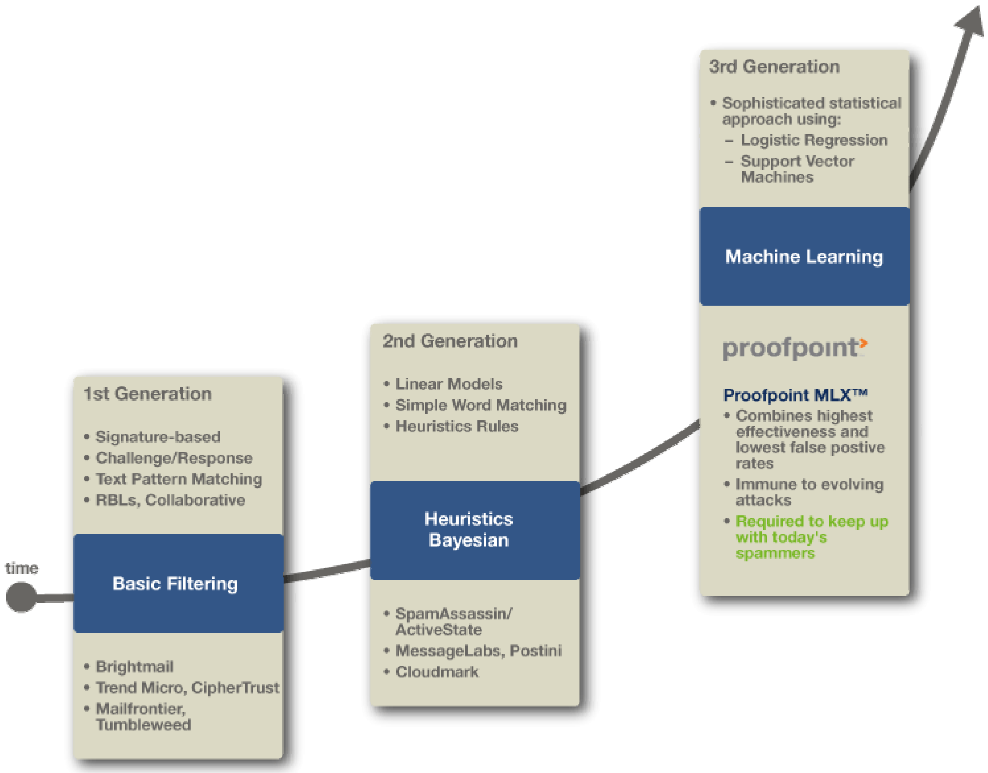

Figure 1 Evolution of Spam Detection

Using Machine Learning to Beat Spam

Proofpoint's MLX technology leverages advanced machine learning techniques to automate the generation of large-scale statistical models for spam and content filtering. Employing a full range of machine learning techniques enables Proofpoint to analyze millions of messages and distill more than 200,000 attributes that reflect the underlying characteristics of spam. The resulting statistical model provides formulas for combining message attributes and weights to estimate the probability that a particular message is spam. With this model in place, the Proofpoint platform can classify messages with a high degree of confidence to maintain high effectiveness rates and a very low occurrence of false positives.

The probability that a message is spam is estimated by applying statistical techniques such as logistic regression. To minimize classification errors, Proofpoint employs Support Vector Machines (SVMs) to train the system to accurately determine whether or not new cases should be classified as spam.

The Importance of Attribute Dependencies – What Naïve Bayes Models Ignore

Because Machine Learning models take into account the incremental impacts of different spam attributes and dependencies between attributes, the system can very accurately classify messages.

For example, if the phrases "Want to stop Snoring?" and "Get a good night's sleep!!!" appear in an email, the marginal spam effect of the second phrase is lessened so that the likelihood that a message is spam is not overestimated. By employing SVMs to understand subtle differences between valid messages and spam, Proofpoint's MLX technology can accurately classify valid messages that might otherwise be confused with spam.

For example, suppose a user receives an email from a colleague that says, "Bob, did you see the spam message to get a good night's sleep?", the system will recognize that it's valid because other attributes are more important than the fact that the message contains the common spam phrase "good night's sleep".

Standard anti-spam systems are not able to detect these subtleties. For example, suppose a user receives the following email from his doctor:

> Dear Bob,
>
> Hope you did get a good night's sleep after your treatment. Did you sleep well and did you stop snoring? It may take a few days for the medicine to kick in. Let me know if you have any questions. – Dr. Smith

A Naïve-Bayes classifier trained on the following spam will incorrectly classify the doctor's message as spam:

> Did you sleep well last night?? Get a good night's sleep! Stop Snoring Today! Click here to learn more!

Because the phrases "Did you sleep well" and "Get a good night's sleep" have appeared in spam before, and the Naïve-Bayes classifier scores all attributes independently, each attribute gets weighted twice. As a result, it overestimates the probability that the message is spam and mistakenly classifies the doctor's email as spam.

In contrast, Proofpoint's MLX classifiers recognize that certain attributes appear together and combine dependent attributes for smart scoring—as a result, the system is able to correctly conclude that the doctor's message is valid.

Attribute Dependencies

Because attributes associated with spam often have complex relationships and dependencies, taking those dependencies into account is critical for accurate spam detection.

Heuristics and Naïve Bayes classifiers evaluate each spam attribute independently—they cannot take into account dependencies between attributes. Because these systems assume that all attributes are conditionally independent, the benefits of considering a larger number of attributes are overwhelmed by the proportional increases in the missing dependencies. This severely limits the number of attributes that these systems can evaluate—they reach a point where adding attributes can actually degrade their ability to make accurate classifications.

In contract, Proofpoint's MLX classifiers accurately model attribute dependencies, enabling the system to analyze more than 200,000 high-quality attributes selected from a pool of over 8 million. This ability to analyze over 30 times the number of attributes considered by traditional systems results in a highly-effective solution that accurately detect spam while maintaining a low incidence of false positives.

Estimating the Probability that a Message is Spam

To estimate the probability that a message is spam, Proofpoint uses logistic regression to define a statistical model that represents the complicated dependencies observed among spam attributes. Unlike Naïve Bayes classifiers, which evaluate each attribute independently, logistic regression enables smart scoring that leverages the knowledge that certain spam attributes commonly appear together. This not only increases the classifier's effectiveness at identifying spam, it enables it to differentiate between spam and valid messages much more accurately.

Logistic regression calculates the incremental impact each attribute has on a message's spam score. A weight is assigned to each attribute to represent its net effect *after* the effects of other attributes are taken into account. Sets of attributes that are known to be dependent on one another are weighted accordingly and redundant attributes receive less weight, ensuring that the probability that a message is spam is not over or underestimated.

Because each attribute's effect is modeled in relation to other attributes, gaps in the model can be filled by intersecting existing attributes to create new ones. In systems that evaluate attributes independently, continuing to add attributes can actually cause a degradation in the classifier's accuracy and effectiveness. However, adding helper attributes to a logistic regression classifier produces a better model with more predictive power.

Logistic Regression

Proofpoint's statistical models combine attributes and weights to generate an estimate of the probability that a particular message is spam. Logistic regression is one technique used to build these classification models. Logistic regression provides a way to predict a discrete outcome such as group membership from a set of variables that can be continuous, discrete, dichotomous, or a mix.[1]

Logistic regression is a Bayesian technique—the most likely model is inferred from a combination of observed attributes and previous data. Instead of making the Naïve-Bayes assumption that each attribute is conditionally independent, logistic regression provides a mechanism for taking interdependencies into account.

Minimizing Classification Errors

An anti-spam solution must be able to accurately classify messages—it must effectively block spam while avoiding false positives. To achieve this, Proofpoint employs SVMs to use a set of training examples to determine whether or not new cases should be classified as spam. Support Vector Machines identify the most important attributes (support vectors) of a message and then differentiate between spam and valid messages based on those attributes.

Support vector machines minimize classification errors by focusing on messages that are close to the boundary, rather than treating all messages equally. By defining the boundary between "spam" and "non spam" based on the training examples, SVMs enable messages to be classified according to which side of the boundary they fall.

Support Vector Machines

Support Vector Machines (SVMs) provide a method for creating classification functions from a set of labeled training data. SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface attempts to split the positive examples from the negative examples. The split is chosen to maintain the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data. To avoid overfitting the model to the training data and maximize its accuracy for new data, Proofpoint performs large-scale cross validation testing utilizing data from a wide variety of sources.

[1] Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics.* NY: HarperCollins.

Machine Learning in Action: Proofpoint MLX

Proofpoint's pioneering research into ways to apply tried and true statistical techniques to the spam problem and focus on the needs of large enterprises has resulted in a complete message-processing platform that provides a comprehensive defense against spam, viruses, and other messaging threats.

Proofpoint's advanced machine-learning classifiers and enterprise-strength platform enables the Proofpoint Protection Server software and Proofpoint Messaging Security Gateway appliance to synthesize large amounts of data and analyze more message characteristics to classify messages with a very high degree of confidence, resulting in a high rate of effectiveness and a very low rate of false positives.

Powered by Machine Learning MLX technology, the Proofpoint platform provides the most effective anti-spam solution available. By leveraging the best of first and second generation spam-detection techniques, applying state-of-the-art MLX classifiers, and adapting to enterprise-specific message characteristics and policies, Proofpoint solutions keep pace with emerging message threats and changing corporate needs.

The Proofpoint system is an enterprise-grade platform designed from the ground up to ensure high availability and performance, minimize management overhead, and integrate seamlessly with existing enterprise management tools.

Maximum Protection Today: High Confidence

The large number of attributes that the Proofpoint platform is able to analyze ensures that messages can be classified with a high degree of confidence.

Proofpoint's advanced classifiers enable the system to classify messages decisively—most messages score very high or very low, with only 1.5% falling between 20 and 80 on a scale of 1-100. Competitors' products often aren't really sure how to classify messages—upwards of 40% of messages typically receive scores between 20 and 80.

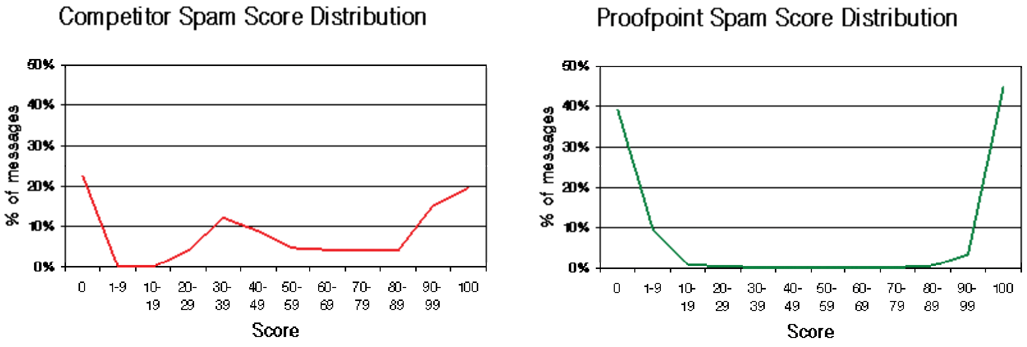

Figure 2 Was it really spam? Decisive classification eliminates the uncertainty.

Systems that are unable to decisively classify spam are left with a difficult dilemma—should the messages that fall in the in the middle of the scoring range be sent to the user as valid email, or blocked as spam? Sending the messages to the user will lower the overall spam detection rate, greatly reducing the solution's effectiveness. On the other hand, blocking the messages will cause a spike in false positives, which can be very detrimental to users.

Clearly, when messages can't be classified decisively, there are no good options. The ability of Proofpoint MLX to classify messages with a high degree of confidence eliminates this dilemma and greatly improves the system's overall effectiveness while maintaining a low rate of false positives.

Maximum Protection Tomorrow: The Learning Cycle

Unlike traditional anti-spam tools that quickly degrade as spammers change their tactics to thwart detection, Proofpoint's MLX-based solutions are capable of learning and automatically adjusting to detect new threats. As more data from both valid email and spam is added to the statistical model, the system identifies and weights relevant attributes to tune the classification process. The result is a system that is just as effective at identifying tomorrow's spam as it is at identifying spam today.

Spam Detection Process

Every email message can be broken down into three main components:

- The Message Envelope—contains information used by the Mail Transfer Agent to route the message. Spammers can change the message envelope by capturing open relays or planting zombies at unsuspecting computers and using them to send spam with a 'valid' email address. MLX catches envelope-based spammer tricks.

- The Message Headers—key value pairs that provide source and routing information for the message, as well as other meta information such as the message sender, subject, and recipients. Headers are often spoofed by spammers. MLX catches header-based spammer tricks.

- The Message Body—the actual content of the message. Spammers often obfuscate the text of the message body using HTML and other encoding tricks, in an attempt to exploit first and second generation spam filters. MLX catches message-body-based spammer tricks.

The MLX detection process begins at the Proofpoint Anti-Spam Laboratory, where scientists and engineers build and refine mathematical models that represent Internet spam. These models are delivered to customers on a frequent basis and are constantly updated to ensure customers stay ahead of the latest spam attacks.

Proofpoint examines every aspect of incoming messages, from the sender's IP address, to the message envelope, headers, and structure, and finally the content and formatting of the message itself. In all, 20 layers of analysis and more than 200,000 attributes/ structural components are analyzed. A typical message may trigger more than 300 MLX structural components.

The first level of screening examines the network stream to identify the source of each incoming email. The system then performs an in-depth contextual analysis of the email, from distilling the e-mail's linguistic structure to normalizing permutations of words (permutations are a common exploit whereby spammers may replace 'Viagra' with a term like 'v1a<b>g</b>gra'). Once the contextual analysis is complete, the system evaluates the message according to the preferences set by both the end user and administrator.

The results of this in-depth analysis are fed into Proofpoint's advanced classifiers to determine the appropriate disposition for the message. On its own, no single test classifies a message as spam—by taking all attributes of a message into account, Proofpoint's advanced classifiers can classify each message with a high-degree of certainty to accurately identify spam and minimize false positives.

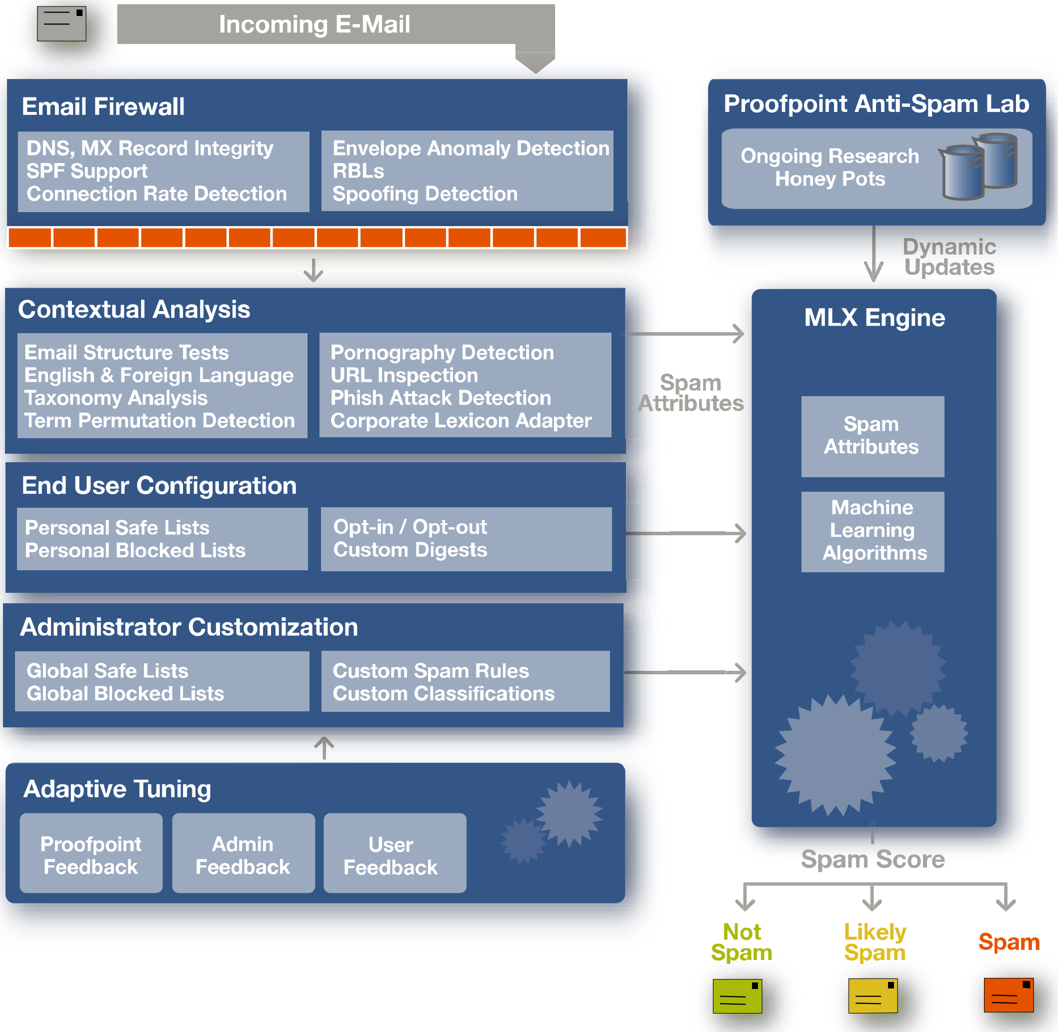

Figure 3 Spam Detection Process

Phish Attacks and Pornographic Spam

Because of the mathematical foundation of MLX, its models can be easily adapted to subcategories of spam. Two common email attacks are phish or scam attacks and pornographic spam.

A phish attack is a type of fraud. Phishing email looks like a legitimate message from a business familiar to the recipient—typically a bank or a well-known online brand such as Amazon or Paypal—but is actually a fraudulent attempt to extract personal identity or financial information. Thinking the message is valid, the recipient posts personal account information. This information is then collected by the sender and used illegally. Proofpoint has applied its machine learning algorithms to detecting phish attacks, thereby ensuring they are blocked from end users' inboxes.

In addition to phish attacks, organizations are being bombarded with pornographic spam, exposing the organization to risk, liability, and embarrassment. Administrators need tools to enforce a policy of zero-tolerance for pornographic spam. Proofpoint not only detects pornographic spam with a high degree of accuracy using MLX technology, but also allows administrators to define separate, more aggressive policies for pornographic spam. Each message analyzed by Proofpoint MLX is assigned a general spam score as well as an adult spam score, enabling each type of message to be handled differently. For example, an organization might delete all pornographic spam, while quarantining non-pornographic spam.

Proofpoint Anti-Spam Laboratory

Talking about machine learning is relatively easy. Developing an enterprise-class anti-spam solution that effectively leverages machine learning techniques and the best traditional techniques requires a major R&D investment and a world-class team. Proofpoint's Anti-Spam Laboratory brings together the expertise and resources necessary to ensure that Proofpoint's solutions continue to set the standards that the rest of the industry strives to match.

Staff Scientists

Proofpoint's success at defending corporate infrastructures against spam is the result of the ongoing work at the Proofpoint Anti-Spam Laboratory. In the laboratory, Proofpoint's scientists continually monitor Internet spam to identify and analyze new spam attributes, update their statistical models, and develop even more effective classifiers. Proofpoint releases multiple updates per week based on the laboratory's extensive testing. These updates are automatically delivered to Proofpoint customer sites via the Proofpoint Dynamic Update service, ensuring that the most accurate statistical models and machine learning classifiers are always used.

Proofpoint has assembled a world-class team of scientists for the Anti-Spam Laboratory, unparalleled in its cross-disciplinary depth and breadth. The team consists of researchers and engineers with deep roots across several relevant disciplines including Machine Learning, statistics, natural language processing, information classification, messaging and security.

Technical Advisory Council

To ensure that the Proofpoint Anti-Spam Laboratory remains at the forefront of the industry, Proofpoint has established a technical advisory council that provides tactical and strategic guidance on specific areas related to MLX technology, content classification, messaging, and security. Members are respected experts in their fields who bring different areas of expertise to the council and provide Proofpoint with a well-rounded, detailed understanding of the challenges and opportunities facing the industry. Members include:

- David Crocker, inventor of internet mail format (RFC822) & 45 other key internet standards

- Philip Hallam-Baker, Principal Scientist, VeriSign, Inc., leader in Internet security initiatives

- Tim Howes, CTO Opsware, Inc., co-inventor of LDAP (Internet standard for directories);

- Dan Jurafsky, Asst. Professor, Stanford University, recognized authority in natural language processing, computational linguistics and probabilistic models for human syntactic parsing;

- Hongyuan Zha, Assoc. Professor, Penn. State – leading researcher in scientific computing, Machine Learning and pattern recognition;

ChengXiang Zhai, Asst. Professor, U. of I. – Urbana-Champaign – expert in natural language processing, personalization of information retrieval and biosequence pattern analysis

Conclusion

Proofpoint's MLX system is continually training to detect the latest forms of spam. Information is fed back into the system to enable it to automatically tune its spam attributes, statistical processes, and classifications. Rather than relying on any one technology, the MLX engine dynamically chooses the most effective set of attributes and models to process each message.

Proofpoint's MLX Technology

- Continuously adapts to detect new types of spam without manual intervention—the system's ability to identify spam does not degrade as spammers change their tactics.
- Employs next generation machine leaning techniques like Logistic Regression and Support Vector Machines to build large-scale statistical models that accurately represent dependencies among spam attributes and delineate the boundary between spam and valid messages.
- Analyzes more than 200,000 spam attributes, including message envelope and header characteristics as well as the actual message content, to accurately classify messages and ensure a low rate of false positives.
- Ensures the maximum protection today and improves in performance, even as spam evolves.

About F-Secure Corporation

F-Secure Corporation protects individuals and businesses against computer viruses and other threats coming through the Internet or mobile networks. Our award-winning solutions include antivirus, desktop firewall with intrusion prevention and network encryption. Our key strength is the speed of response to new threats. For businesses our solutions feature centralized management. Founded in 1988, F-Secure has been listed on the Helsinki Exchanges since 1999. We have our headquarters in Helsinki, Finland, and offices in USA, France, Germany, Sweden, the United Kingdom and Japan. F-Secure is supported by a global ecosystem of value added resellers and distributors in over 50 countries. F-Secure protection is also available through major Internet Service Providers, such as Deutsche Telekom and France Telecom.

Europe	***USA***
F-Secure Corporation PL 24 FIN-00181 Helsinki, Finland Tel +358 9 2520 0700 Fax +358 9 2520 5001 http://www.f-secure.com/	***F-Secure Inc.*** F-Secure Inc. 100 Century Center Court, Suite 700 San Jose, CA 95112, USA Tel. (408) 938 6700 Fax (408) 938 6701

"F-Secure" and the triangle symbol are registered trademarks of F-Secure Corporation and F-Secure product names and symbols/logos are either trademarks or registered trademarks of F-Secure Corporation. All other product and company names referenced herein are trademarks or registered trademarks of their respective companies. F-Secure Corporation disclaims proprietary interest in the marks and names of others. Although F-Secure Corporation makes every effort to ensure that this information is accurate, F-Secure Corporation will not be liable for any errors or omission of facts contained herein. F-Secure Corporation reserves the right to modify specifications cited in this document without prior notice.

We are continually evaluating and developing our products. Please visit www.f-secure.com for the most recent information.