Overview, Key Words, and Subtopic Name

This endeavor aims to bring enhanced two-factor authentication to personal workstations, eventually impacting how the log-in process is handled on smartphones, tablets, and numerous other consumer electronics. Firmly situated within the Information Technology & Cybersecurity topics, this system will use voice recognition and keystroke dynamics analysis to introduce a confidence measure for each login attempt. This confidence measure enables systems to admit users with limited permissions as necessary, or notify users via SMS that their account has been accessed.

Intellectual Merit

This Small Business Innovation Research Phase I project strives to improve the authentication procedure for desktop operating systems, preventing unauthorized access even from actors with the login password. The most significant technical hurdle will be supplementing the existing login process with biometric signatures that are tuned to adequately balance security (unforgeability) with clemency (“forgiveness”). This means that users need not deliberately practice their typing rhythm to perfection, as small deviations will be filtered out in favor of more-significant typing characteristics; worry not about an “off-day” preventing you from accessing your machine.

Beyond improving the login procedure, the project aims to spark a conversation about unobtrusive mechanisms for securing information, and to change the nature of the relationship between enterprises, consumers, and the cyber-liability insurance industry.

Broader/Commercial Impact

This project aspires to become a Software-as-a-Service product for institutions like banks to further secure online portals, and for sensitive information enterprises like Electronic Health Records to quantify their confidence that they are showing information to the right people. This confidence, in turn, provides companies with another tool in the information assurance tool belt, protecting them from legal exposure in the event of a data breach, and ultimately lowering their costs for cyber insurance.

The Customer

Forbes, the NYTimes, and MIT’s Sloan Review all agree that, in this modern digital landscape, every company is a tech company. Industries dealing with sensitive information depend on their information assurance infrastructure to prevent catastrophe from affecting the bottom line. Cyber liability insurance premiums totalled $2 billion in 2015, and, according to InsuranceJournal.com, are slated for a ten-fold increase over the next ten years. Legal costs and liability expenses from a data breach can hamstring businesses focused on rapid growth, like many modern Silicon Valley unicorns. This product exists for enterprises, large and small, with a vested interest in protecting their data and systems.

The Value Proposition

Simple username/password combinations are no longer sufficient to protect a user's data, and organizations that fail to implement modern security measures like two-factor authentication are paying the price. But even these new measures are flawed in some regard; for example, two-factor authentication depends on the user having a mobile phone available whenever they wish to access their account. Our product aims to solve these issues, empowering system administrators with increased transparency and greater control over attempted access to their systems and network.

The Innovation

This project aims to provide the benefits of multi-factor authentication without making too many assumptions about the user's access to a charged mobile phone with an active phone number. Leveraging technologies already available in modern consumer PCs, this project will supplement the traditional password login with additional verification stages as needed.

Keystroke dynamics & typing rhythm are analyzed as a user enters their password, triggering a machine learning algorithm to determine whether or not that user's characteristic rhythm matches the established composite profile of that user. If the machine learning algorithm's confidence in the user's identity does not achieve or surpass an organization's predetermined threshold, the system prompts the user to perform additional voice recognition tasks to complete the login process. Acoustic metrics like pitch, intensity and timbre combine to provide a robust authentication mechanism, guaranteeing only authorized users gain access to a machine, network or web portal.

The Commercial Opportunity

Within the last decade, many industries have witnessed a migration of data to cloud storage. Among numerous benefits, these cloud solutions offer improved scalability and performance to services with users across different continents, and help to manage server maintenance costs. This trend, in confluence with the changing nature of the data being stored, has cast the value of information assurance in a new light. As e-commerce continues to dominate brick-and-mortar retail, consumers gain awareness of just how many ACH and payment processing databases contain their sensitive banking details. The ascension of health-tracking wearables, often subsidized by health insurance companies, has normalized sharing intimate details about our bodies with Fitbit, Misfit, or any other startup de jour. Facebook accounts and email addresses have come to serve as our passports, identifying us to others and behaving as proxies through which we conduct our personal and business affairs.

Consumers are willing to relinquish this information in pursuit of convenience; users can access their documents, correspondence, photos, music, and health information from any internet-connected device in the world. In exchange, they are transferring ownership of their data to the companies they entrust this information to. Companies eagerly accept this information for at least one of two reasons; they can charge users for the privilege of using a company's storage (cloud backup services), or they use the information to develop composite profiles of potential customers, in turn selling these profiles to advertisers & ad networks (e.g. Google's DoubleClick). By that same token, these companies inherit liability; they become stewards of this information, and it is in their best interest to protect it to the greatest extent possible.

One need not look further than recent news headlines to see the financial impact this liability can have. Yahoo, eagerly attempting to be sold for a mere 10% of its valuation 8 years ago, has suffered an additional 20% decrease in Verizon's buyout offer in light of a breach resulting in the release 500 million user credentials. Bitdefender, in a 2015 article, cites a Forrester analysis asserting double-digit growth in the security expenditure of certain industries. Whereas incident response is a prudent investment for all enterprises, it's importance is eclipsed by the value of investments in *preventing data breaches* in the first place.

This project aims to provide enterprises of all sizes with increased confidence in their ability to prevent data breaches. Licensed on a per-user basis, the software will introduce additional authentication factors as needed to prevent bad actors from gaining access to a system.

Biometric authentication is in no way a novel concept; companies like KeyTrac provide authentication systems based on a user's keystrokes, and multinationals like Nuance offer voice analysis to customer service organizations as an authentication measure. What differentiates this project from competitors is the fact that additional authentication factors are introduced as needed, significantly reducing hassle and opportunities for social engineering exploits.

The primary risk to bringing this innovation to market is the fact that, presently, nothing prevents competitors from emulating the technology. With the requisite capital, patents can be filed to protect this intellectual property, but none are currently in place.

Initially, commercialization will target small enterprises across different industries and sectors. Implementation engineers will be assigned to each of these customers, physically visiting their headquarters and aiding in the setup of the software system. Beyond simply installing the software and training the on-site IT admin, implementation engineers will be able to gain insight into pain-points and feature requests, informing subsequent iterations on the product. After this preliminary information-gathering phase, a second release of the software will include a simple bootstrapping procedure to aid IT professionals with installation, as well as a feedback system to automate further data-collection, and improve later releases. Using a continuous delivery release schedule, this project will be able to quickly respond to innovation in cybersecurity, as well as broader changes in consumer demand.

Societal Impact

Beyond the commercial aspects, broader societal impact is paramount in determining the direction and approach this project takes. Considering the changing nature of information stewardship, with consumers handing over sensitive data for companies to store in cloud databases, the time is nigh to improve information security standards. We aim to inspire innovation in how technology services approach multi-factor authentication.

Assuming modest adoption rates, the software would encourage digital service providers to prioritize inherent traits of a user, like their vocal timbre or keystroke rhythm, over coincidental knowledge factors, such as knowing a password or having access to a registered mobile device to which an verification code is sent. The user experiences greater convenience in tandem with increased protection, and enterprises succeed in democratizing multi-factor authentication without worrying about whether users have access to a working phone or inbox for SMS or email verification.

First and foremost, consumers would be impacted directly, benefitting from tighter security standards guarding their sensitive information. Beyond the tangible benefits of more highly-secured banking/health data, the consumer gains peace of mind. This, in turn, benefits digital enterprises, personifying them as noble and worthy of entrusting one's most-personal information to.

However, we do acknowledge that this is a significant undertaking. Any exploitation of this project will have trickle-down consequences for the organizations implementing this software, and care must be taken to transparently address and resolve vulnerabilities in any aspect of the system. The last thing we want is to damage the brand integrity of our enterprise clients, so we must be forthright in disclosing suspected vulnerabilities as we learn of them. This underlies the decision to use a continuous delivery schedule; flaws can be patched quickly with minimal intervention by a given client’s IT personnel.

Ultimately, it is up to enterprises to determine whether this system is appropriate for achieving higher levels of security in their particular application and use cases. Products and services intended to be used by multiple individuals on a single account might not benefit from such strict authentication, and the value of password manager applications might be negatively impacted. As such, potential customers must consider the Latin principle “caveat emptor”, roughly translating to “let the buyer beware.”

Key Technical Challenges and Risks

The primary technical challenge will be collecting enough data from a user to train a significantly robust classifier. In order to precisely gauge a particular user’s keystroke dynamics, we would need to record the user typing a variety of sentences with every letter in the alphabet appearing multiple times. While there are a few different possible solutions, their costs and benefits must be considered thoughtfully.

The first solution is to run a keylogger on a user's system, collecting plenty of data for use in training the classifier. While this is ideal from a machine learning perspective, the end user might find it counter-intuitive to install a keylogger in pursuit of better information security. Especially in the initial stages of the product offering, the inclusion of a keylogger would necessitate third-party security certifications, giving the customer confidence that the product is not an eloquent scheme to steal user data. Beyond potential damage to the brand reputation, there are engineering challenges involved in running a keylogger. This collected data would need to be fed back into the training set, which (depending on the architecture chosen by a sysadmin customer) may introduce encryption and networking requirements.

The second solution is to engineer better features from the existing training data provided by users, and lower the max confidence interval values that a sysadmin can choose for authenticating their users by keystrokes alone. Currently, aggregating dwell times for every key into a single set of descriptive statistics (average, minimum, max, quartiles, standard deviation) achieves ~65% accuracy. Considering the unique value proposition of this project is to dynamically introduce authentication factors as needed, this inadequate accuracy means that users would most likely have to complete additional levels of authentication.

Description of Innovation

Traditional multi-factor authentication systems require users to complete all the authentication factors before gaining access. This process can be time consuming, inefficient, and have unnecessary requirements that make access impossible at times. One common example is SMS or email confirmation, where users are expected to be able to receive an access code in their inbox, but this fails to consider situations where customers are unable to access these messages due to connectivity issues or a dead phone battery.

The aim of this project is to allow system administrators to choose, and even implement, the authentication factors they deem appropriate for their users, and for these factors to be dynamically introduced as needed. This is where the novel innovation lies: administrator-mandated confidence intervals create a threshold for the user to meet or exceed, and failure to do so introduces an additional authentication factor & confidence interval. The outcome is an efficient login procedure that intrudes as little as possible on the user, who very well may not even know heightened security protocol is in place.

As a user enters their login credentials, the system logs timestamped keypresses, including **keydown** and **keyup** events. These timestamped logs are parsed and fed into a pre-trained machine learning classifier, which outputs a predicted label (user or not-user) and a degree of confidence in that prediction. These logs are then cached until the end of the login process, where they will either be labelled "user" or "not user" and fed into the classifier as additional training data. This machine learning algorithm is considered to be ‘online’, since it is continually training on new data and becoming more accurate as time goes on.

Assuming the user successfully enters their username and password, and their keystroke dynamics match the user profile generated by the machine learning classifier, the user is then granted access to the system. If the user’s keystroke dynamics do not sufficiently indicate that they are the correct user past a certain confidence interval threshold, then the next authentication factor is introduced.

For Phase I, this second authentication factor will be a digital signature that the user inputs via the computer’s webcam. The user will sign their name with their finger in front of the webcam, generating an image of their signature on-screen. These signatures will be cached as a list of pixel values and, similar to the keystroke dynamics classifier, will be fed into the signature evaluation system as additional training data in the event that the user successfully gains access to the system. The signature evaluation system will operate in a slightly different manner than the keystroke classifier; rather than labeling an input as **user** or **not user**, the signature evaluator will output a percentage match, indicating how closely the input resembles the average of known-correct inputs for a given user. Since there are theoretically infinite signature possibilities, generating training data signatures for **not-user** would be an intractable problem. Therefore, we will focus on using distance algorithms to measure the similarity between unlabeled inputs and labeled, correct inputs in order to judge whether or not a signature matches the expected form and style.

Key Phase I Objectives

The key objectives for Phase I are determining the proper algorithms, feature engineering techniques, and evaluation criteria to successfully distinguish between users by their keystroke dynamics and virtual signatures. This experimentation phase will need to be automated, allowing us to run batch processes to grid search for the ideal parameters. Fortunately, we’ve performed our experimentation using the Scikit-Learn library for Python, which means we have access to the GridSearchCV submodule, and its methods, for determining the most performant model selection.

Selecting the proper algorithm for each task is important, as each machine learning algorithm has a variety of advantages and disadvantages that make it ideal for specific tasks. For example, Decision Trees can be useful for exploratory analysis in their easy interpretability, but are prone to overfitting on data, meaning they generalize poorly to data outside the training set.

Feature engineering is particularly important for the keystroke detection classifier, and the bulk of the requisite research. Considering many users may have similar typing styles and rhythms, it’s important to build features that emphasize slight, nuanced differences in typing styles between otherwise similar typists. Standard features like dwell time, the time between a key press and key release, and flight time, the time between releasing a key and pressing another, are not enough to reliably distinguish between two similar typists. This means we will need to expand and combine these features to gain a more distinct, unique profile of each typist. With the automation of the model training and evaluation process, we will be able to more-rapidly iterate on successful feature engineering techniques, and quickly abandon approaches that show little to no promise.

Selecting useful evaluation metrics guarantees the machine learning models don't train specifically for the test, helping us construct models that generalize well for previously-unseen data. Whereas traditional accuracy scores can help us determine how well the model performed on the test set, novel metrics like the area under the curve for the receiver operating characteristic will give us a better sense of the distribution of false positives, false negatives, true positives and true negatives. When we get a grasp on the particular ways a model underperforms, we stand a better chance of being able to optimize that model to supplement those shortcomings.

Additional thought must be put into allowing system administrators to implement their own authentication factors as needed. This is as much an engineering problem as it is a product design problem; for example, should administrators be expected to write their own authentication factor code and integrate with the product via an API or SDK provided by the product? These decisions constitute and impact the R&D process, demonstrating the importance of proper market research and user experience design in the ultimate marketability of the product.

Critical Technical Milestones

The primary technical milestones standing between the current project and the final, ready-for-market product are research-based. Rapid experimentation must take place in the feature engineering department to determine the best ways to interpret the user-provided data. Without groundbreaking innovation in feature selection, the machine learning models that the entire project operates on will perform inadequately, meaning users will likely have to complete more authentication steps to gain access to a system.

The entire value proposition of the project is a system that intrudes on the user as little as possible, maximizing the accuracy of each authentication factor is an imperative. Beyond customer dissatisfaction, a subpar security system has implications for information security, and ultimately affects the product's reputation in the market.

R&D Plan

I. Keystroke Detection
    A. Data Collection - **complete**
        1. Set up a web page to collect timestamped keystroke data
        2. Save results to cloud database

B. Feature Engineering - **in progress; predicted completion 12/20/16**

1. Research best-practices for metrics depicting keystroke rhythm
2. Automate grid-search for ideal parameters
3. Analyze different features and their predictive value, revising those that are minimally-useful

C. Trained Model saved to Pickle file - **predicted completion 12/31/16**

1. Tune parameters and export to Pickle file for rapid use in production's authentication workflow

II. Virtual Signature Comparison

A. Initial library integration - **in progress; predicted completion 12/31/16**

1. Live Kinect stream on screen - **complete**
2. Draw bounding box on screen - **in progress**
3. Follow joints

B. Data collection - **predicted completion 1/15/17**

1. Log finger movement as coordinate pairs
2. Preliminary data analysis

C. Comparison algorithm - **predicted completion 2/15/17**

1. Comparative analysis of distance metrics for signature analysis
2. Implementation, testing and parameter tuning
3. Export for production workflow

III. Administrator Control Panel

A. Preliminary UI mockups - **predicted completion 1/15/17**

1. Requirements gathering
2. Determine ideal configuration of UI elements to collect information from administrator

B. Implementation - **predicted completion 2/15/17**

1. Build & style control panel frontend
2. Configure log-files to describe administrator's choices for the authentication factors
3. Integrate control panel frontend with log-file generator

IV. Integrate Control Panel with network of Systems

A. Network Architecture Design - **predicted completion 3/1/17**

1. Determine how to store and distribute log-files to client systems

B. Implementation - **predicted completion 3/30/17**

1. Build interacting modules to facilitate configuration sharing, config installation (into client system), and reporting successful installation back to administrator's control panel

Joseph Haaga

Assignment 4 Revision Plan

- Pg1: Change semi-colon to colon on page 1
- Pg2: Move sentence down to next paragraph, serving as introduction sentence
- Pg3: Remove ‘to’ and change “emphasis” to “emphasize”