C3E Workshop

The Role of Experimental Science In Cybersecurity

Ed Giorgio
Ponte Technologies

Keystone, CO
September 25-28, 2011

Cybersecurity research is (in part) an experimental science which benefits from a hypothesis driven analytical approach involving thousands of experiments.

Experimental Science

Model development frequently includes building normal behavioral models from historical data, so that anomalies can later be detected and visualized. The resulting Hypothesize-Test-Revise (HTR) cycle is frequently repeated (with human intervention) many hundreds, if not thousands, of times during the development of a useful model for anomaly detection

Hypothesize – The innovator hypothesizes how observed phenomena might be explained and builds (or modifies) a model capturing that explanation.

Test – The computer scientist develops algorithms & software to test and modify the model to determine how well it predicts the observations. There is frequently a visualization step which involves statistics or data compression (ex. PCA) on the data to help understand the results.

Revise – The mathematician explains the significance of the results (ex. likelihood ratio, Type I/II error separation) and suggests new values of parameters (ex. partial derivatives).

It is easy to draw parallels with other data rich science arising in natural language, life sciences, cryptanalysis, and financial modeling

Data Rich Problems in Science

- Natural Language
  - Expanding Shannon's work to predict natural language
  - Use soft selectors to identify behavior (ex. classified docs, terrorist, etc.)
  - Find instances of a known writer or speaker
- Cryptanalysis
  - Find vowels & consonants in a "substitution with variants" code
  - Separate ciphertext messages into subsets with common keys
- Life Science
  - Find similar amino acid sequences by minimizing "edit distance"
  - Mapping migratory and evolutional trends using PCA generated maps from genome
  - Separating coding from non-coding regions of the genome
- Financial Modeling
  - Hidden Markov Models of macro state of economy (bull, bear, sentiment, hype)
  - Detailed predictive models for multiple financial instruments

Cybersecurity research can focus on either similar data rich problems or on theoretical problems and while both are important, the experimental science seems to get a lot less attention.

Theoretical Problems in Security

- Primality proving
- Formal methods to prove program correctness
- Complexity theory assessment on the membership of a problem to a class of hard problems
- Finding polygraphic repeats (or signatures) in large data sets
- Proof that RSA actually works $m^{ed} \equiv m^{1+k\varphi(n)} \equiv m(m^{\varphi(n)})^k \equiv m \pmod{n}$
- Parsing data generated by a context sensitive grammar
- Zero Knowledge protocol which proves both sides know the same password
- Exhaustion on all cryptovariables
- Factoring using Schor's algorithm on a quantum computer
- Proof of Fermat's "little" or "last" theorem $a^{p-1} \equiv 1 \pmod{p}.$
- Formal language theory to express policy and access control restrictions

Cybersecurity problems frequently do not lend themselves to closed form solutions, and thus there is no unifying theory to discover

Data Driven Problems in Cybersecurity

- Separating malware from normal code
- Finding polymorphic versions of known malware
- Finding network abnormalities from "network flow data"
- Separating normal gateway traffic from exfiltration data from internal host
- Detecting previously unknown steganographic systems
- Crypto attacks based on timing variations due to memory hierarchy
- Separating phishing emails from normal emails
- Building a pfsa to model password creation
- Building Bayesian networks to assign likelihood ratios to attack trees
- Computations on a graph generated from address connection data
- Performing Power Differential Analysis (PDA) on key fobs and USB tokens
- Determining if domain names are real or generated by malware

Experimental and theoretical science overlap, yet is relatively easy to identify an experimental scientist who is virtually always on a computer developing software and running experiments.

Attributes of an Experimental Information Scientist

- Spends >50% of time on computer debugging code and running experiments
- Usually runs at least 10 experiments per hour, each of which requires software modification, testing on data, visualization of results, and generating new ideas for model improvement (usually in a rich environment such as unix/C/emacs)
- Knows Knuth's theory of algorithms and has programmed many of them
- Understands Shannon's information theory, Markov modeling, Chomsky's language theory, statistics, automata theory, and more
- Able to quickly find open source code and incorporate it into research
- Never believes the problem is solved and always seeking improvement
- Face time with other experimental scientists who have clocked 10,000 hours
- Is known by colleagues for solving many real word problems

There are many types of data that will emerge from the sensors, which include routers, firewalls, hosts, IDS/IPS, apps, CDS, VPNs, etc.

Examples of Cybersecurity Data

- Results of applying malware signatures
- Report external attempts at port scanning
- Results of "dirty word" searches for content filtering
- Black list and white list IP address violations
- Mail scanning to find malware, phishing, botnets, etc.
- Policy violations for peer-to-peer traffic, ports & connections, CDS, etc.
- Host Based Security System (HBSS) self reporting
- Unpatched systems reported by special network sensors (ex. Trickler)
- Netflow data (Top N talkers & volumes, ports, IP addresses, connection requests TCP flags, ICMP destination/port unreachable, etc.)
- Enterprise wide failed password attempts (or other violations)

This Cybersecurity data can be analyzed by multiple statistical techniques which are valuable in producing displays, plots, graphs, pie charts, 3D displays, etc. , all of which allow humans to quickly understand what is happening

Statistical Techniques

- Multivariate correlation analysis
- Principal Component Analysis (PCA) (eigenvectors of correlation matrix)
- Bayesian Networks
- Probabilistic finite state automata (pfsa)
- Histogram profiling and characterization
- Inter-arrival time analysis
- Chi-squared and other tests for normality
- Subsequence near repeat matching in very long sequences
- Parsing of data using statistical grammars
- Categorizing message types using neural nets and artificial intelligence
- Knowledge representation and expression techniques

Finally, the data available for analysis will depend on what sensors are available to produce the data sets

	Perspective	Description
1	ISP	The inter-domain traffic between the various DoD Points of Presence (POPs) on either DoD WANs or commercial ISPs.
2	Enterprise	The classified networks & hosts which have very different interconnection scenarios than the unclassified networks & hosts.
3	Enterprise Edge	Focuses exclusively on the boundaries between classified and unclassified networks where explicit policy rules exist the gateways (ex. CDS).
4	Network Host	Enterprise hosts and servers that can only see their local OS, application, and network activity

Even ongoing operations, especially situational awareness, can benefit from experimental research which can be broken down into a sequential process which starts with collection and ends with operational use, and includes continuous improvement.

1. **Collecting the data from many sources for research**
2. **Adopting common formats for raw data and event detection**
3. **Data reduction through correlation and other statistical methods**
4. **Normal/anomaly behavior model development**
5. **Software development for visualization**
6. **Insertion into systems for operational use**
7. **Feedback loop to refine model**

Experimental science doesn't end in the laboratory; we will need a large analytic cyber force because the human element cannot be removed from either the research or operations and it takes 10,000 hours to develop the expert.

Now let's look at the use of Principal Component Analysis (PCA) to compress & visualize observed data in several fields.

Organize the data set as a set of *N* data vectors each representing a single grouped observation of the *M* variables.

Calculate the empirical mean $$u[m] = \frac{1}{N}\sum_{n=1}^{N} X[m,n]$$

Calculate the deviations from the mean $$\mathbf{B} = \mathbf{X} - \mathbf{uh}$$ where **h** is a 1 x *N* row vector of all 1's:

Find the covariance matrix $$\mathbf{C} = \mathbb{E}\left[\mathbf{B} \otimes \mathbf{B}\right] = \mathbb{E}\left[\mathbf{B} \cdot \mathbf{B}^*\right] = \frac{1}{N}\sum \mathbf{B} \cdot \mathbf{B}^*$$

Find the eigenvectors and eigenvalues of the covariance matrix $$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$$

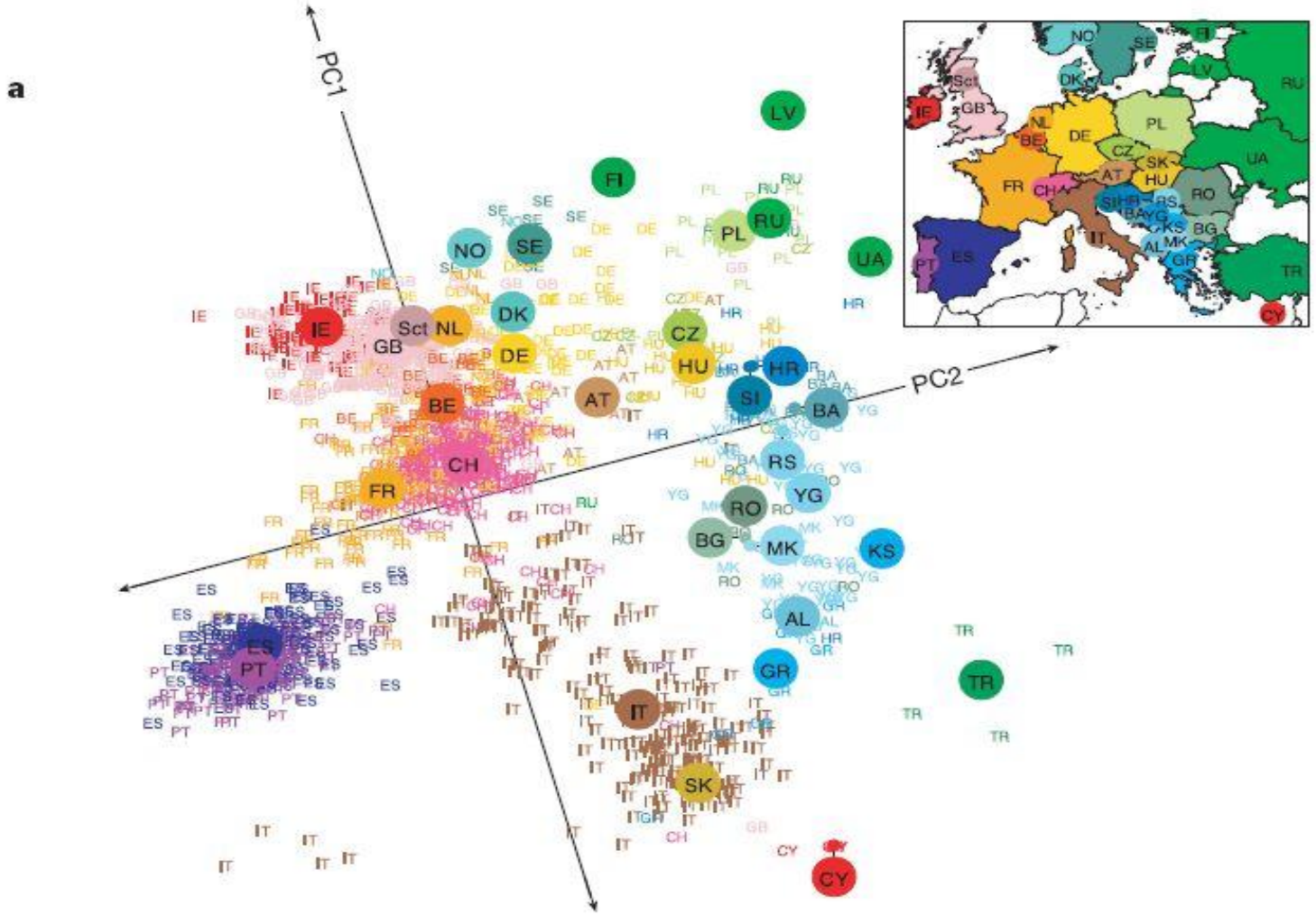
Rearrange and Compute the cumulative energy content for each eigenvector $$g[m] = \sum_{q=1}^{m} D[p,q] \quad \text{for} \quad p = q \quad \text{and} \quad m = 1, \ldots, M$$

Select a subset of the eigenvectors as basis vectors $$W[p,q] = V[p,q] \quad \text{for} \quad p = 1, \ldots, M \quad q = 1, \ldots, L$$

Convert the source data to z-scores $$\mathbf{Z} = \frac{\mathbf{B}}{\mathbf{s} \cdot \mathbf{h}} \qquad \mathbf{s} = \{s[m]\} = \sqrt{C[p,q]} \qquad \text{for } p = q = m = 1 \ldots M$$

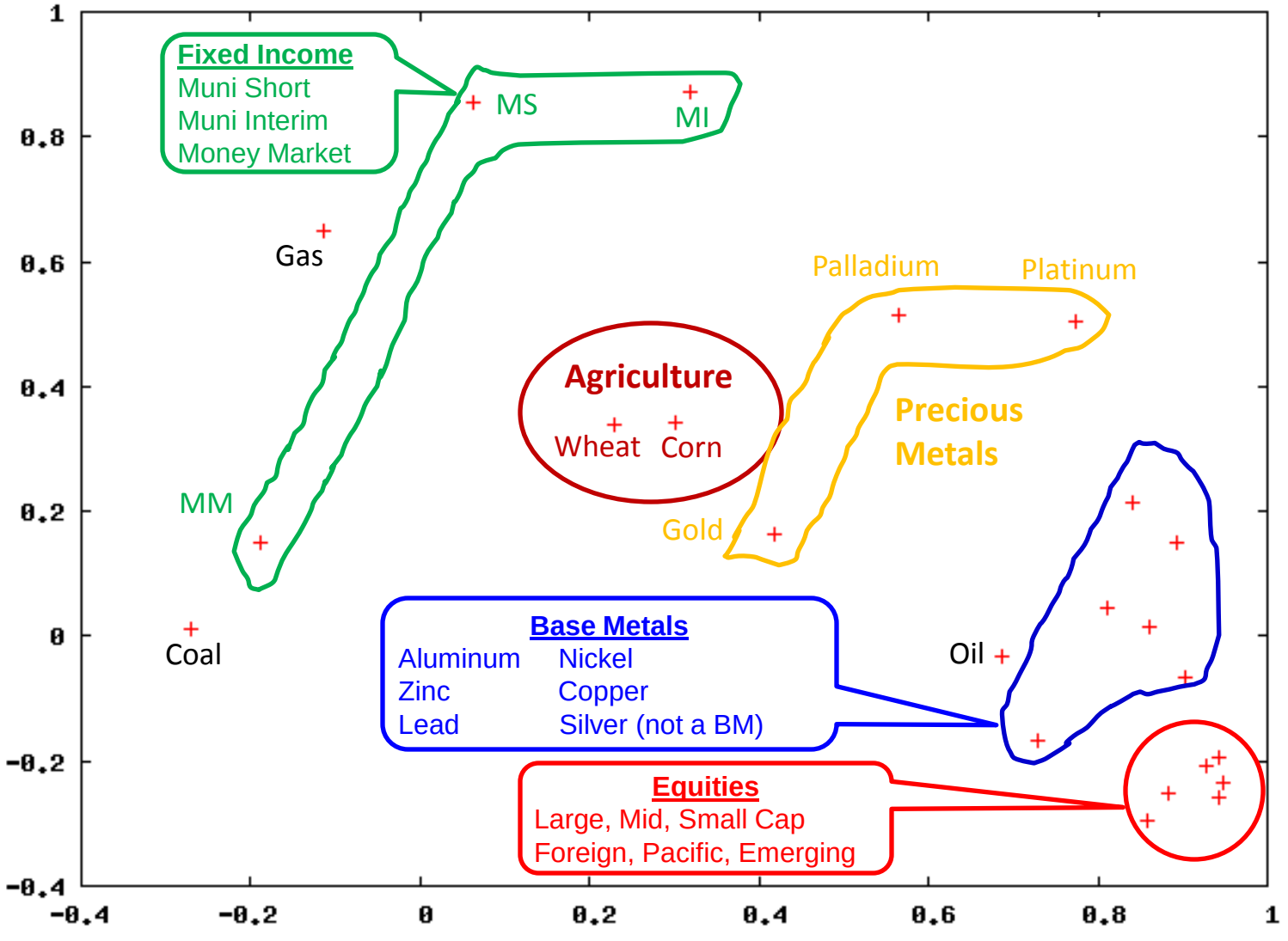
Project the z-scores of the data onto the new basis $$\mathbf{Y} = \mathbf{W}^* \cdot \mathbf{Z} = \mathbb{KLT}\{\mathbf{X}\}.$$

From PCA analysis of human genome samples it is evident that the first two principal components bear a strong relationship to the geographical map of Europe.



Nature, November 2008 , Novembre et al: Genes mirror geography within Europe

Looking at financial data we found very little performance diversity within similar asset classes; only fossil fuels acted unpredictably.



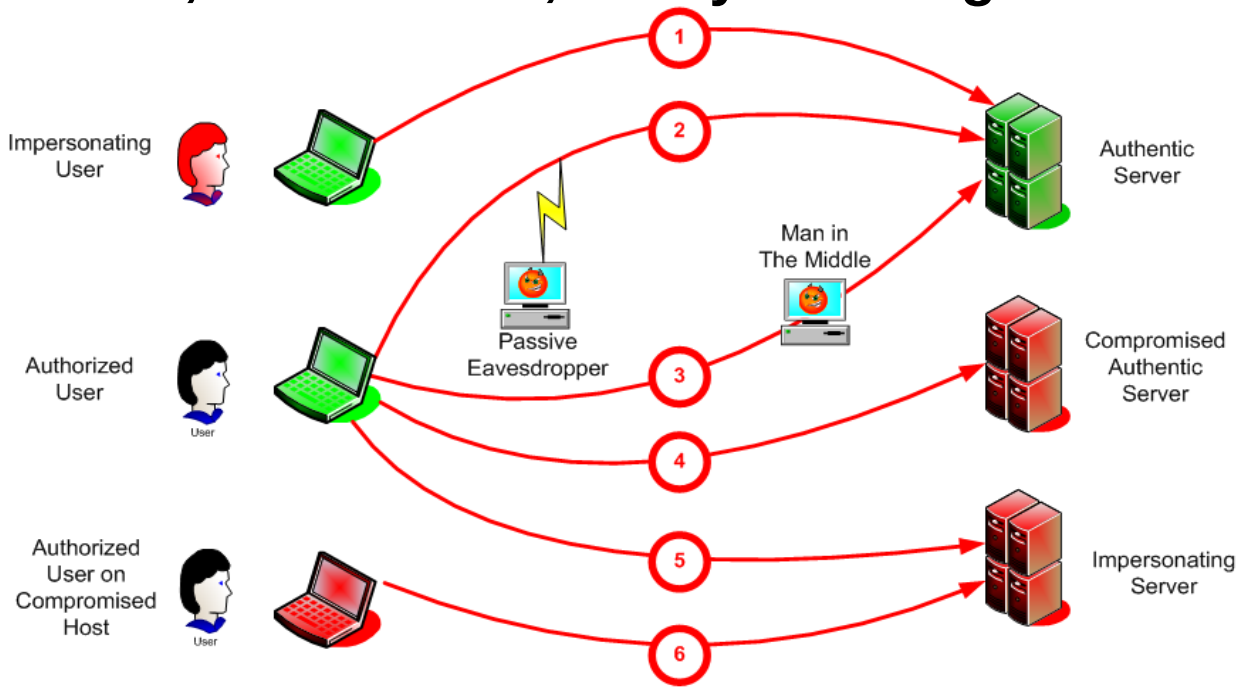
The use of PCA on “network flow data” has demonstrated the ability to detect anomalous activity.

“Network traffic arises from the superposition of Origin-Destination (OD) flows. Hence, a thorough understanding of OD flows is essential for modeling network traffic, and for addressing a wide variety of problems including traffic engineering, traffic matrix estimation, capacity planning, forecasting and anomaly detection. ..

Using Principal Component Analysis (PCA), we find that the set of OD flows has small intrinsic dimension. In fact, even in a network with over a hundred OD flows, these flows can be accurately modeled in time using a small number (10 or less) of independent components or dimensions. We also show how to use PCA to systematically decompose the structure of OD flow time-series into three main constituents: common periodic trends, short-lived bursts, and noise. We provide insight into how the various constituents contribute to the overall structure of OD flows and explore the extent to which this decomposition varies over time. “

Anukool Lakhina et al: Structural Analysis of Network Traffic Flows

Finally, shouldn't we be using more decision logic involving attack surfaces, attack trees, & Bayesian logic?



	Attack Agent	Actor Vector	Prob.*
1	Impersonating User	Physical Compromise of Password	.40
2	Passive Eavesdropper	Passive Access to Channel	.25
3	Man-in-the-Middle	In-Line Access to Channel	.15
4	Impersonating Server	Induces trust from User	.10
5	Man-in-the Computer	Compromised Client	.10
6	Man-in-the Server	Compromised Server	.05

* Probabilities from Mudge (Peiter Zatko, DARPA PM)

Ponte Technologies

Bridging technology with business innovation, security with emerging applications, and theory with practice.

Ed Giorgio
Ponte Technologies
443-226-0944
edgiorgio@pontetec.com