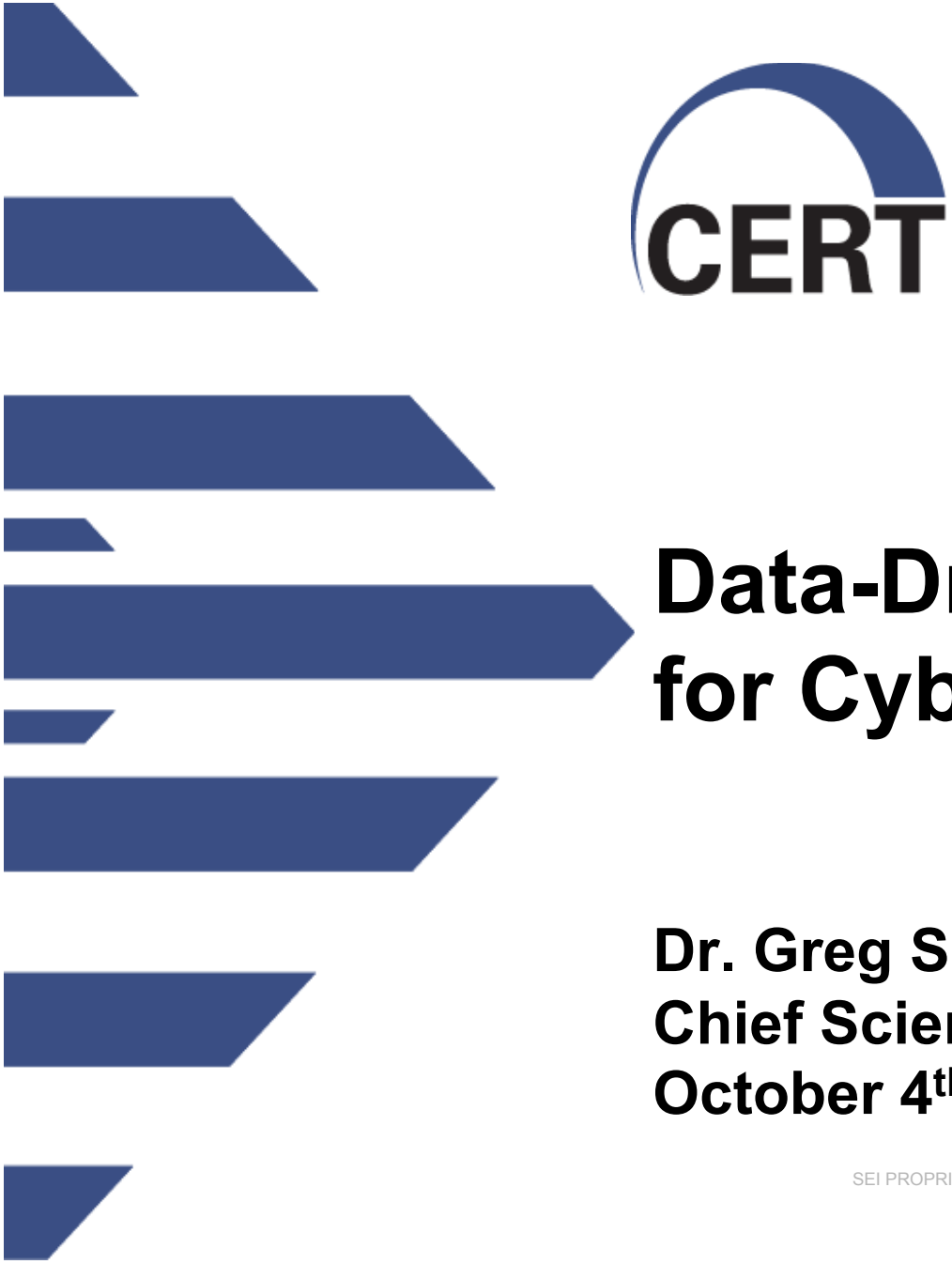


Data-Driven Science for Cyber Security

Dr. Greg Shannon
Chief Scientist
October 4th, 2010

Overview

- Background on SEI and CERT
- Science and Cyber Security
- Malware Analysis for Trending
- Detecting Insider Threat

Objective: Collaboration

Background on SEI and CERT

Software Engineering Institute

- Department of Defense R&D Laboratory; FFRDC
- Created in 1984
- Administered by Carnegie Mellon University
- Headquartered in Pittsburgh;
  offices and support worldwide
- ~400 technical staff

The SEI advances software engineering and related disciplines to ensure systems with predictable and improved quality, cost, and schedule.

CERT® Program

- **Created in response the Morris Worm (1988)**

- **Today**
  ~200 technical staff in directorates, 20 PhDs
  30+ open positions

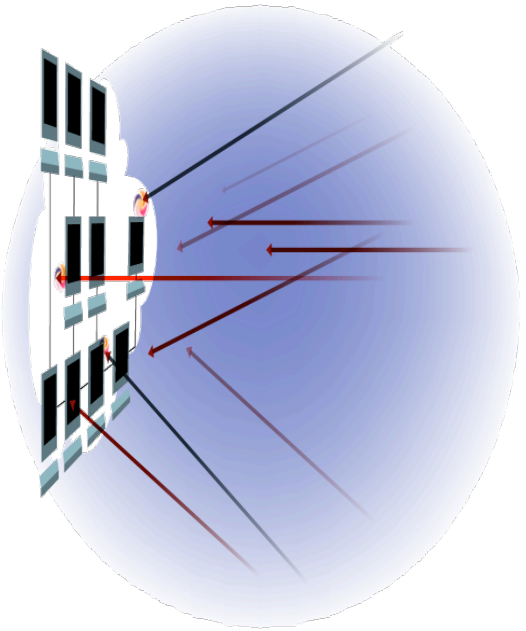
- **Mission**
  Create informed trust and confidence in critical cyber technologies.

- **Vision**
  A securely connected world.

- **Strategy**
  - Reduce the impact of cyber attacks with new:
    - **Software and system development technologies and practices**
    - **System and network monitoring technologies and practices**
    - **Digital investigations and intelligence methods and tools**
  - Anchor R&D efforts in operational challenges and realities.

CERT® Program Directorates

- **Cyber Threat and Vulnerability Analysis (CTVA)**– Discover and resolve vulnerabilities in software products; improve cyber-tradecraft analysis; and quantitatively assess potential threat and subsequent impact of malicious activity.
- **Enterprise and Workforce Development** – Establish the routine use of disciplined approaches to improve survivability and resiliency; and provide security practices and information assurance training and education.
- **Secure Software and Systems Engineering** – Develop technologies and approaches to embed software and system assurance in all aspects of the system development life cycle.
- **Digital Investigations and Intelligence** – Support federal, state, and local investigators through applied research and tool development in large-scale memory extraction and analysis and acquisition and recovery of encrypted data.



Strategically Relevant
Attacks on DoD and Defense Industrial Base networks are common and increasing.

What Does CERT Bring to the Table?

- Government and Industry Experience
  - **Customers** with Pain in Cyber Security
  - **Data** collected
  - **Trusted** 3rd party
  - Operational experience and capabilities

- Full cycle perspective on cyber security
  - Pre-use: Design, Development, Deployment
  - Use: Operations, Continuity of Operations, Training
  - Post-use: Forensics

- Research Focus
  - Science of Security – data and experience driven research
  - Collaboration, publishing, impact

Research Challenge in Cyber Security

- **Threats at Scale in number and time**
  - Small attacks hurt, but aren't what really matter
  - These matter: defense, power grid, financial services, etc.
  - Adversaries can affect millions of connected objects in very compressed time frames with the speed of light as the fundamental limiting factor

- What makes this challenging?
  - Immense attack surfaces: computers, applications, services, networks, routers, users, physical control connections, databases, business operations, etc. etc. Billions of objects.
  - Sub-second timescales for attacks, responses, situational awareness

- We don't know yet how to effectively deter, prevent, detect, respond in a way to mitigate important threats at scale.
  - We're making progress, but the gap is a national security issue
  - **How do we not inhibit innovation, agility, resiliency?**

- CERT's research approach
  - Exploit data collected to mitigate threats and attacks.
  - Exploit data collected to inform development of secure/resilient software, systems, networks, services etc.
  - Develop scalable cyber-security forensics

Research Areas at CERT

- Vulnerability Analysis
- Secure Coding
- Malware Analysis
- Network Situational Awareness
- Incident Response Teams
- Insider Threats
- Cyber-Security Training
- Resiliency and SmartGrid
- Forensics
- Security Measurement
- New Security Mechanisms
- Software Assurance Engineering

CERT's Data Collections

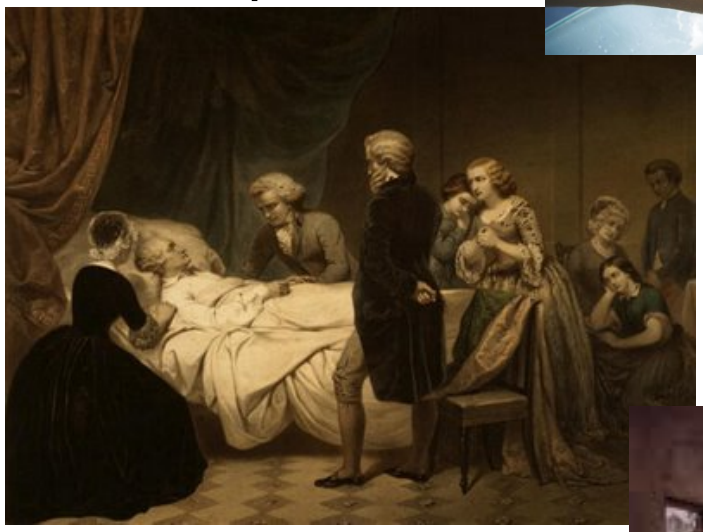
- Malware Catalogue
- Vulnerability Data
- Incident Data
- Assessment Data
  - Insider Threats
  - Trusted Gateways
  - Resiliency
- Network data
  - Netflow
  - DNS
  - BGP
- Forensics Artifacts
- Insider Case Database
- Training Events

Science and Cyber Security

Past Challenges in Research

- Health ~ Religiosity →→→ Health ~ Hygiene
  - Required an understanding of the underlying phenomenologies that degrade health as opposed to the causes of health per se.

- Bloodletting
  - Widely accepted treatment in 1800 for fever, swelling.
  - "Medical statistics" led to better treatments



- Alchemy
  - Broad support; fervently practiced by Newton
  - Eventually overcome by modern chemistry

What is Research?

- Systematic investigation to establish facts and new conclusions (aka new knowledge)
  - Find enduring/useful principles, laws, and models of the essential phenomenology
- More philosophically
  - It's either **math** or → **science** ← **CERT's focus**
  - Math is deductive
  - Science is inductive and driven by observation and data
- Not engineering or development per se
  - These are applied math & science

An Opportunity in Cyber Security

- Getting to game-changing technologies presumes we have discovered sound and pragmatic mathematical and scientific principles for cyber security.

- Cyber Security has important and relevant math principles, but math alone is insufficient…
  - Intractability is a significant negative practicability result
  - Any math used must be empirically validated for utility
    - Paraphrasing Poincaré (IEP): Empirical information is crucial to the choice we make (about which math to use).

- Cyber Security has few scientific (induction-derived) principles to apply in developing of secure systems.
  - We're missing an emperical phenomenology for cyber security
  - Principles must be inferred from the data (follow the data)

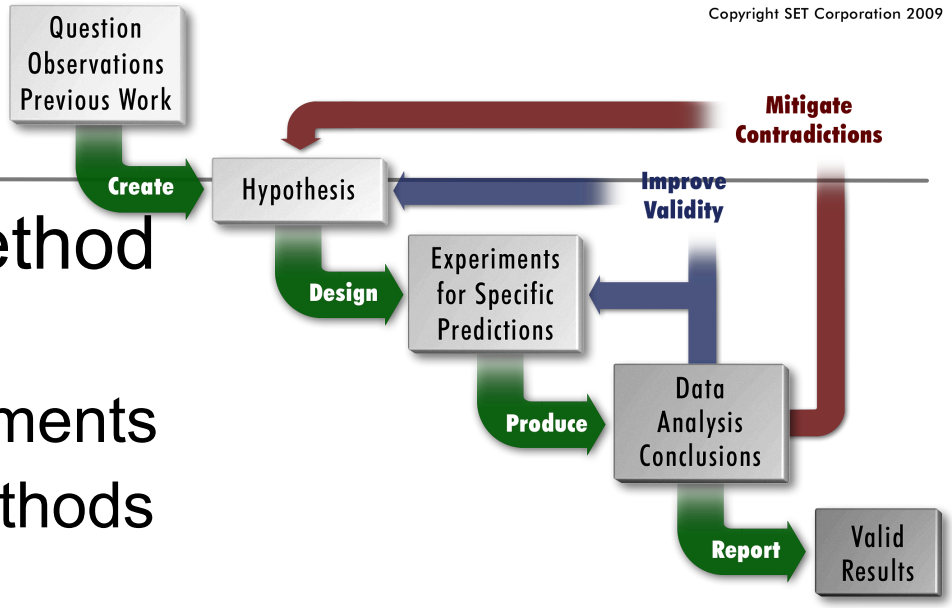
An Opportunity in Cyber Security

The best opportunity for game-changing research in cyber security is to **scientifically** collect and study data about the computing/networked ecology in order to discover the empirically expressed phenomenologies and principles of cyber security.

From these results and existing mathematical principles we can/might develop game-changing technologies.

The alternative is that our adversaries' pragmatism will continue to threaten/dominate us in cyber space.

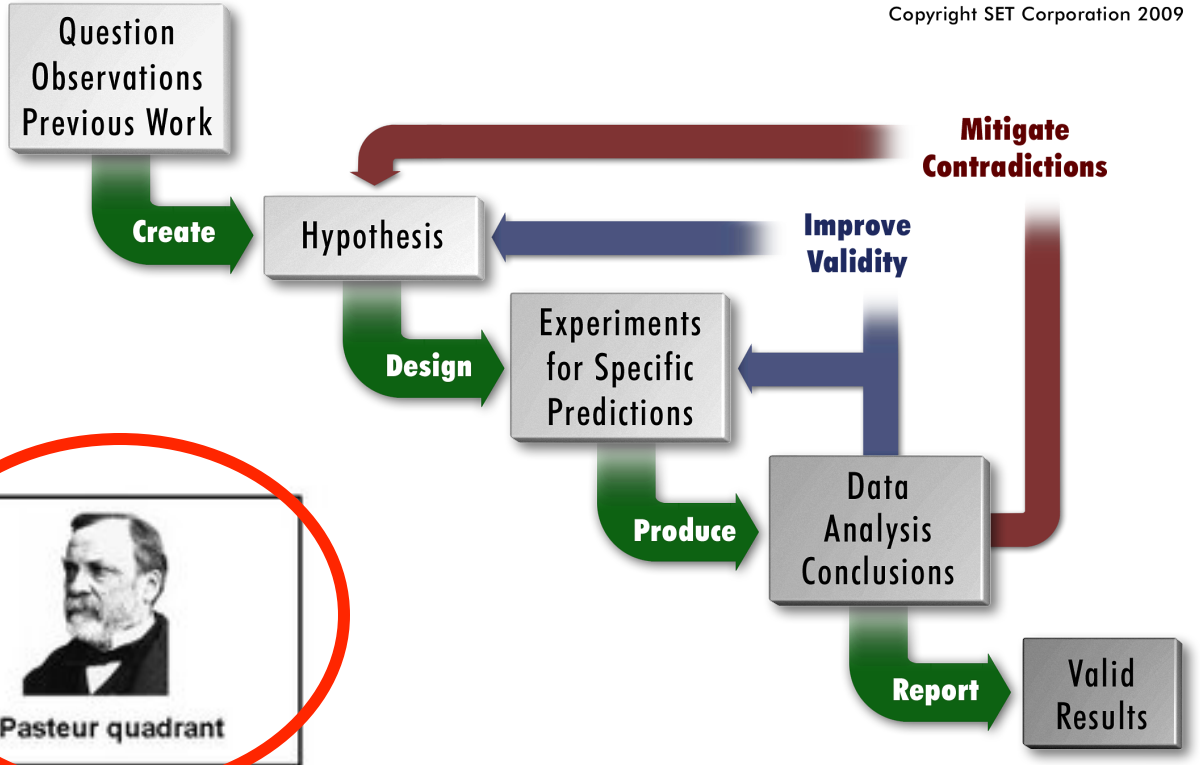
What Can We Do?

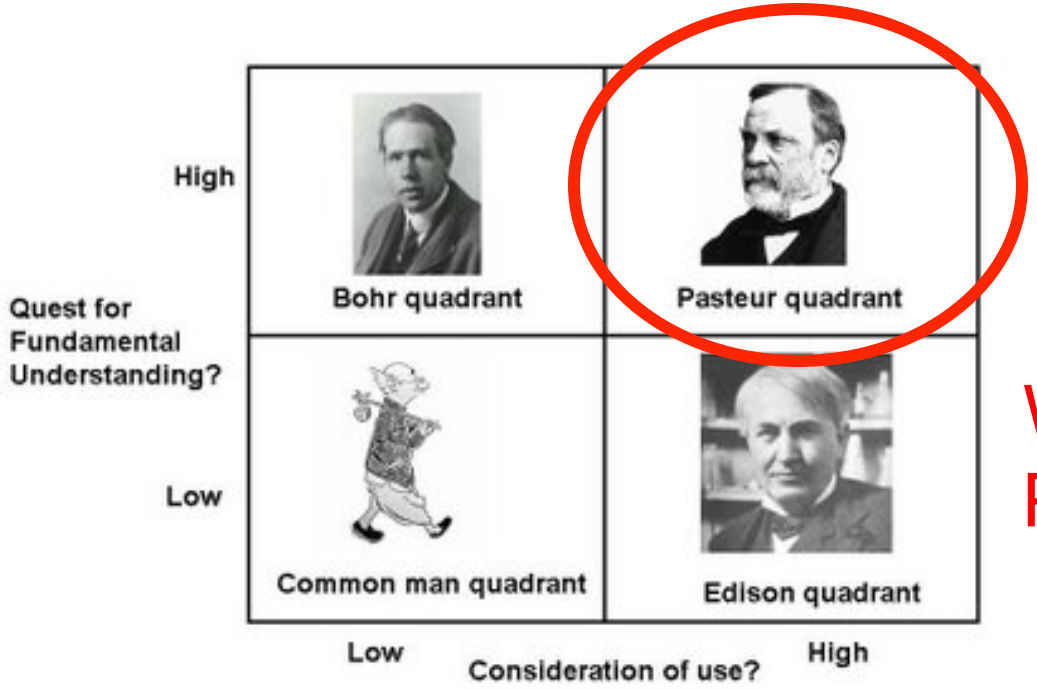


- Espouse the Scientific Method
  - In funding priorities
  - In contract reporting requirements
  - In education – Research Methods
- Support the development of rigorous (aka valid) experimental methods and apparatuses
  - Address validity-changed approaches (like red teams)
  - Develop valid test beds and methods for using them
    - National Cyber Range
- Support broader access to high-fidelity data
  - National-asset data repositories at FFRDCs
  - High-fidelity operational data from large populations

Let's Do Some Good Science!

Use the Scientific Method





Work in Pasteur's Quadrant

http://cataligninnovation.blogspot.com/2008/10/bangalore-innovation-forum-and-pasteurs.html

Reading

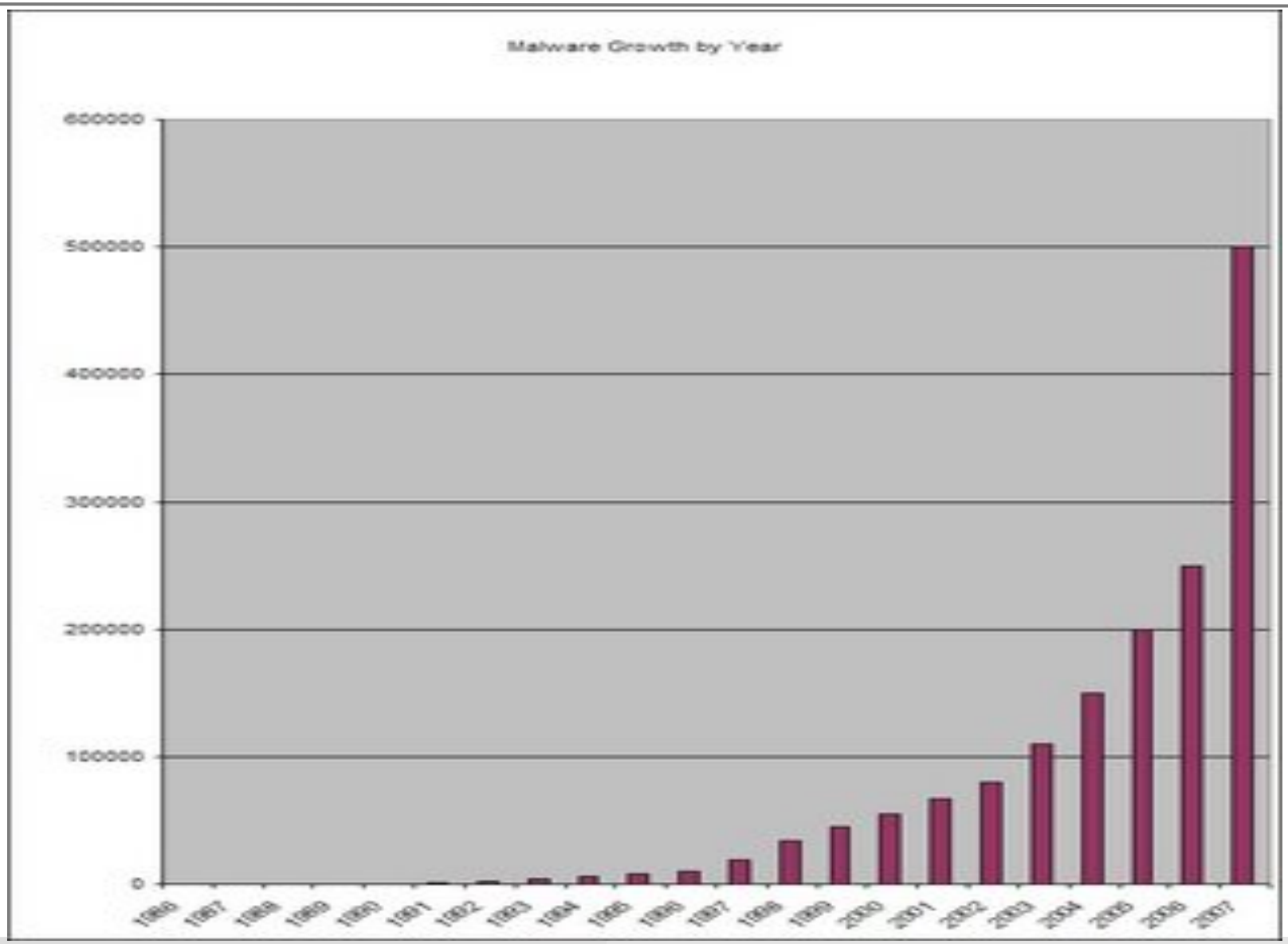
- The Scientific Method in Practice
  - Hugh G. Gauch, Jr., 2002
- The Structure of Scientific Revolutions
  - Thomas S. Kuhn, 1962
  - Father of “paradigm shift”
- Pasteur’s Quadrant
  - Donald E. Stokes, 1997
- Einstein's Clocks, Poincare's Maps
  - By Peter Louis Galison, 2003

Malware Analysis for Trending

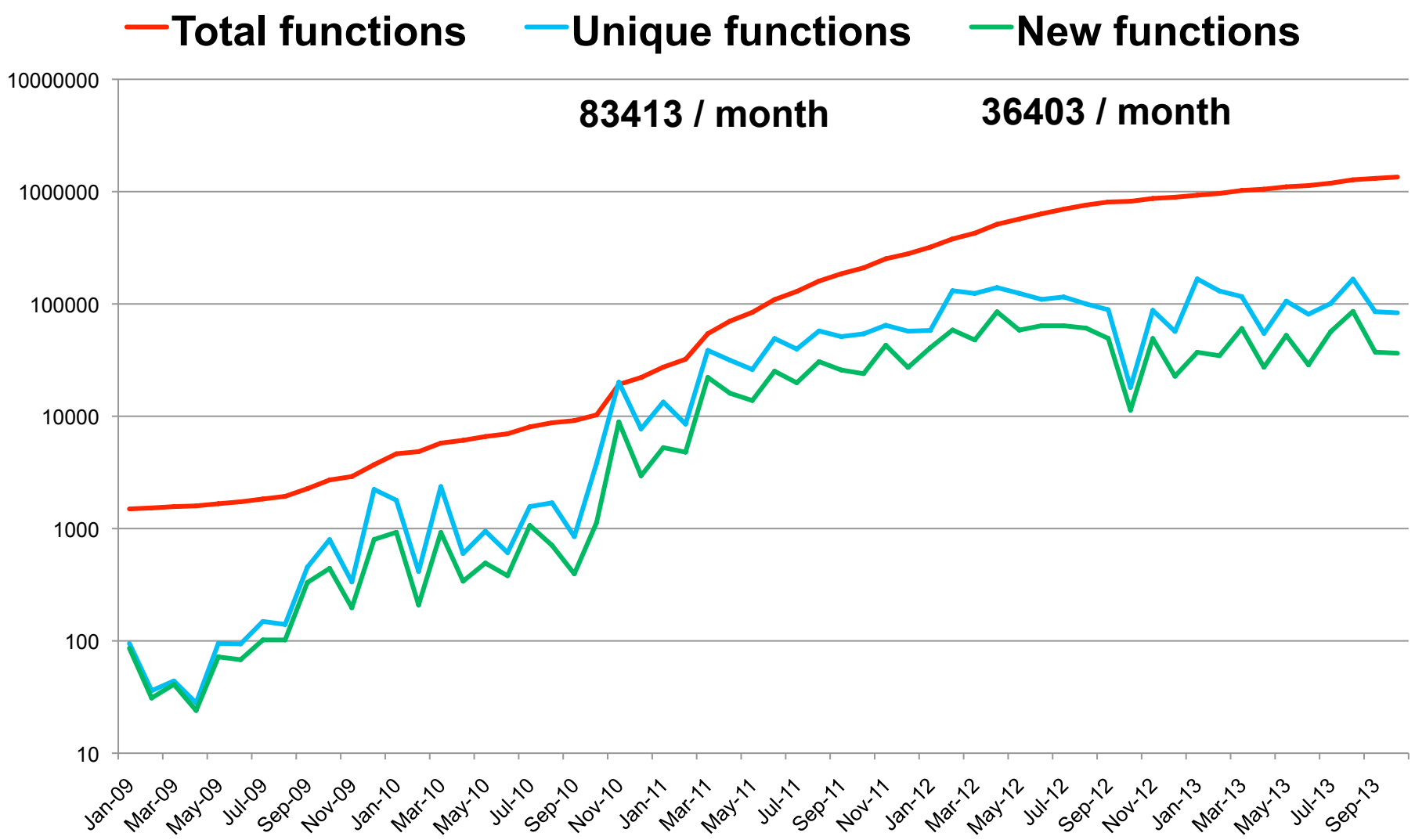
CERT Malware Artifact Catalogue

- CERT maintains and operates the Artifact Catalog for the U.S. Government as directed by congress.
- This unclassified central repository of malicious code enables a broad collection of samples for the purposes of both research and operation analysis.
- It contains ~12 million artifacts collected since 2001 and grows at up to 300,000 artifacts per month.
- There are few (if any) other unclassified malware repositories that are as larger as, or more complete than, this one maintained by CERT.

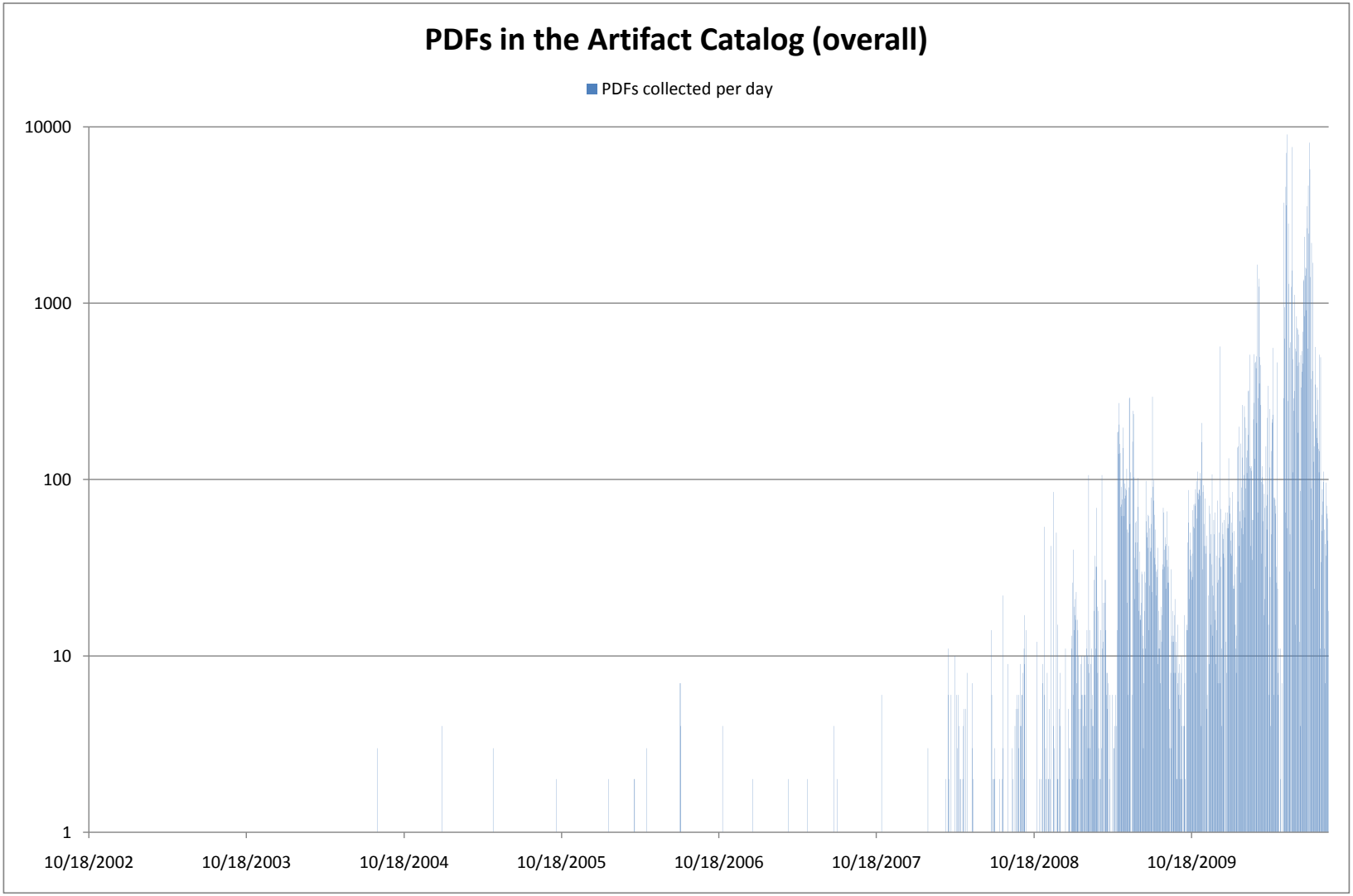
Conventional Wisdom on Malware



New Functions in Binaries



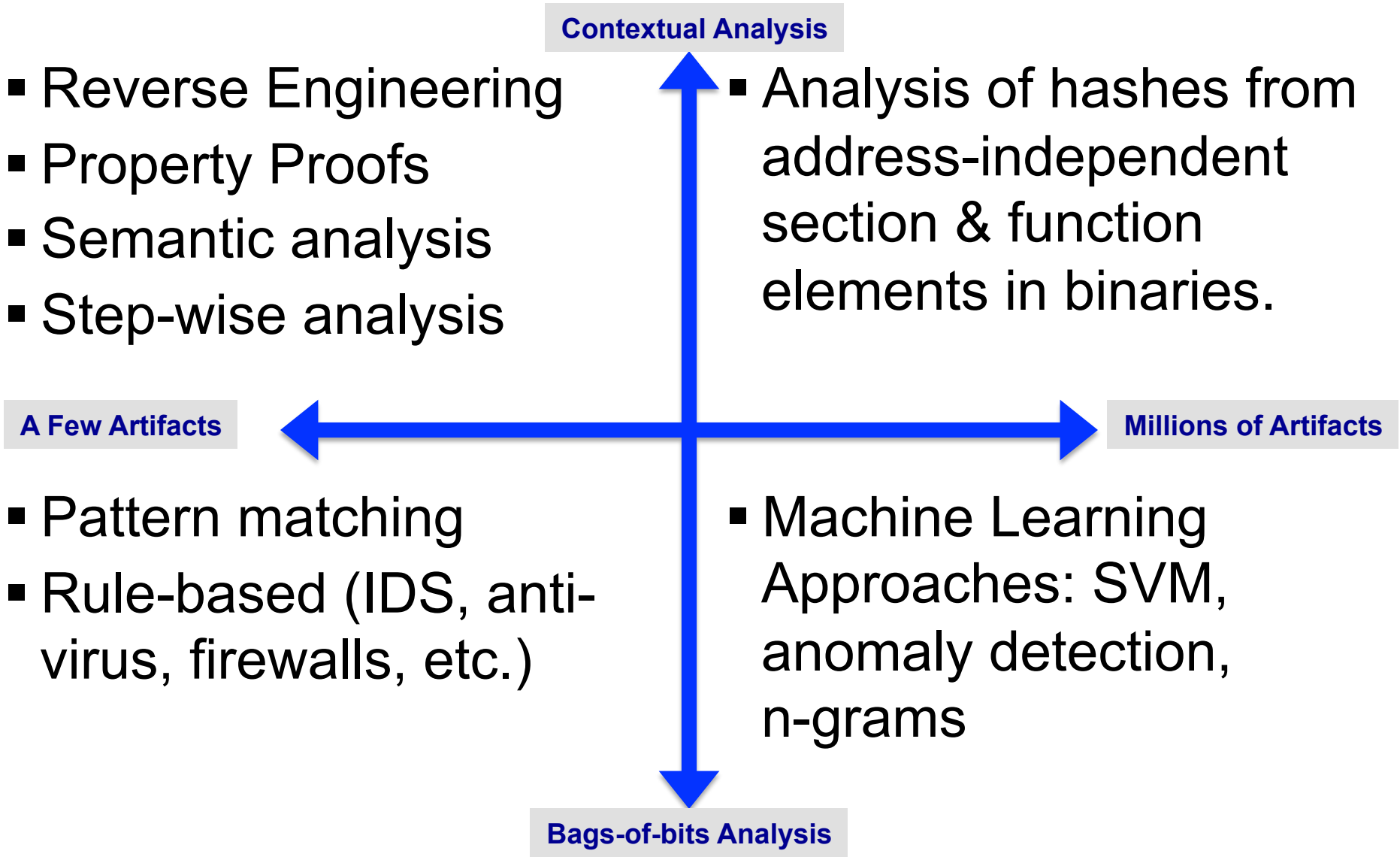
Trend in PDFs



Malware Analysis: Classification and Pedigree

- **Objective**
  - Leverage large malware collections in order to quickly understand and categorize the properties and pedigree of malware instances among the malware reported daily; identify and track trends in malware.
- **Challenges**
  - Scale of artifact catalogue (10M), daily volume of new artifacts (10K)
  - Packed, obfuscated, broken, near duplicate binaries and sub-elements
- **Research Approach & Innovations**
  - Continue improving unpacking knowledge and tools
  - Continue improving fast, reliable decomposition tools
  - Create advanced hashing techniques to better identify duplicates
  - Create advanced data structures for fast queries at scale
  - **Create machine learning techniques for accurate automatic classification**
- **Impact to DoD**
  - Operational situational awareness (SA) for cyber defense, operations
  - Threat trending, TTP discovery and tracking

Malware Analysis: Classification and Pedigree



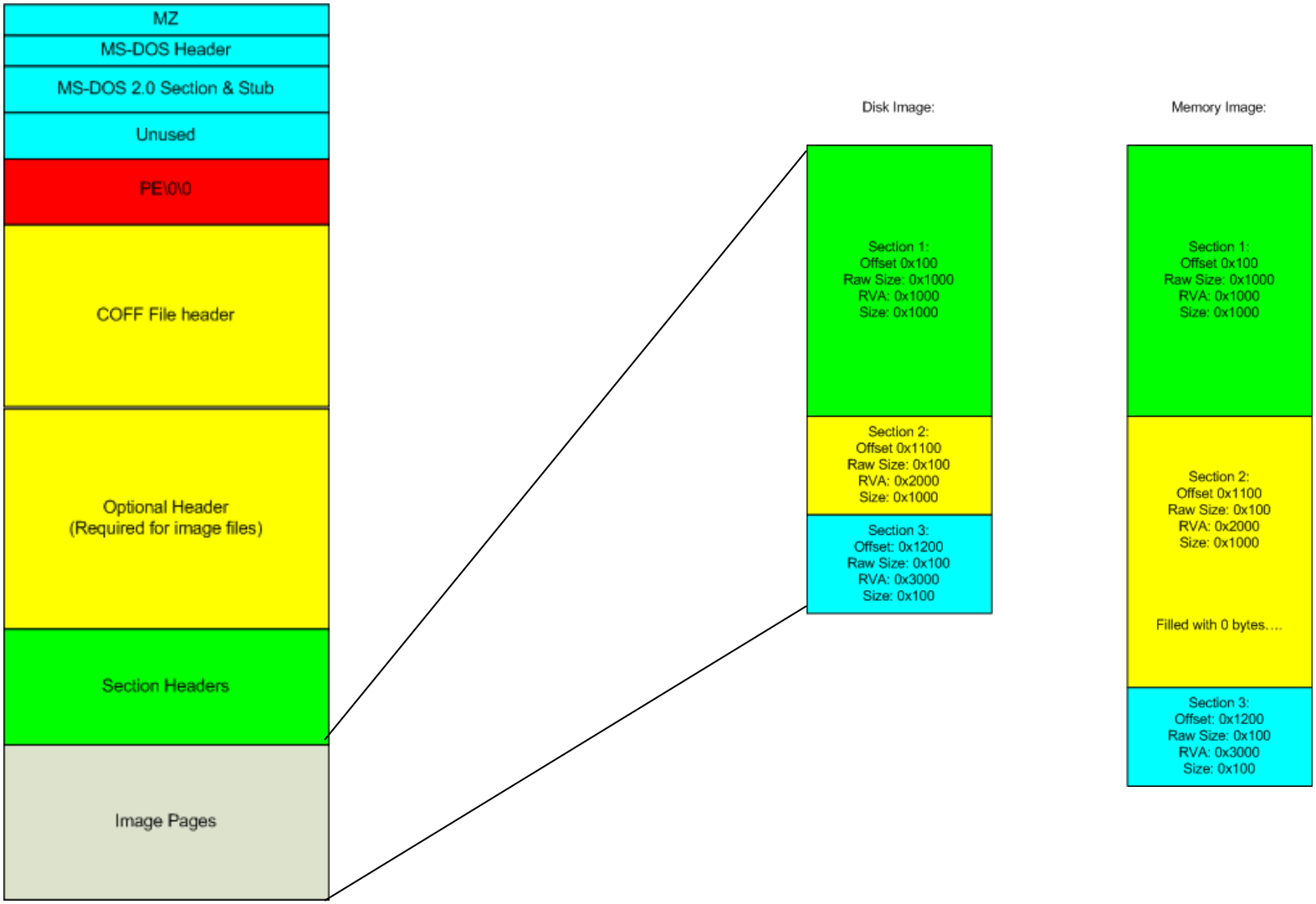
Some Details (Presented at GFIRST, MTEM)

- CERT has access to ~10M malicious code samples
- We have developed analytic techniques
  - Data transformation
  - Data comparison
- We have observed malicious code trends
  - Packer proliferation
  - Malware family proliferation
  - Malware duplication
- We would like to present some guidelines for observing these trends for your own malware

Objective vs. subjective

- Trends need to be informed by data
- Useful trends need useful data
  - Reproducible
  - Consistent
- Suspect data yields suspect trends
  - "Black box" tools ("how does this thing produce its answers?")
  - "Fuzzy" data ("this looks sorta kinda like that other thing")
- Bottom line: Prefer objective measurements to subjective ones
  - Corollary: when you must cheat, know why and what cost

Overview of PE structure



Observables from PE structure

- Entry point
  - Relative virtual address (RVA), specified by IMAGE_OPTIONAL_HEADER
  - Bytes represent the instructions executed first
- Sections
  - Natural boundaries for “types” of data in PE
  - Specified by IMAGE_SECTION_HEADERs
  - MS Windows Loader behaves differently from PECOFF specification!
    - More on this later

Observables from PE structure

- Functions
  - Functions consist of bytes passed to instruction cycle
    - Opcodes, operands, addresses, data, etc
  - Not directly available by parsing header
    - Header tells you which bytes to start at
    - Interpreting them is the processor's job
  - Must recreate the context/interpretation of bytes to observe a "function" = disassembly
  - Third-party tools to the rescue!
    - IDA-Pro is de-facto standard for disassemblers
    - Others currently under evaluation (ROSE, etc)

Tools to exploit observables

- MD5
  - Standard cryptographic hash algorithm
  - Used to reasonably* assert “uniqueness” of data
  - Allows significance to be asserted by collision

- Composite hashing
  - Hash of hashes
  - Separates data from its underlying structure
  - Removes duplicate data from consideration

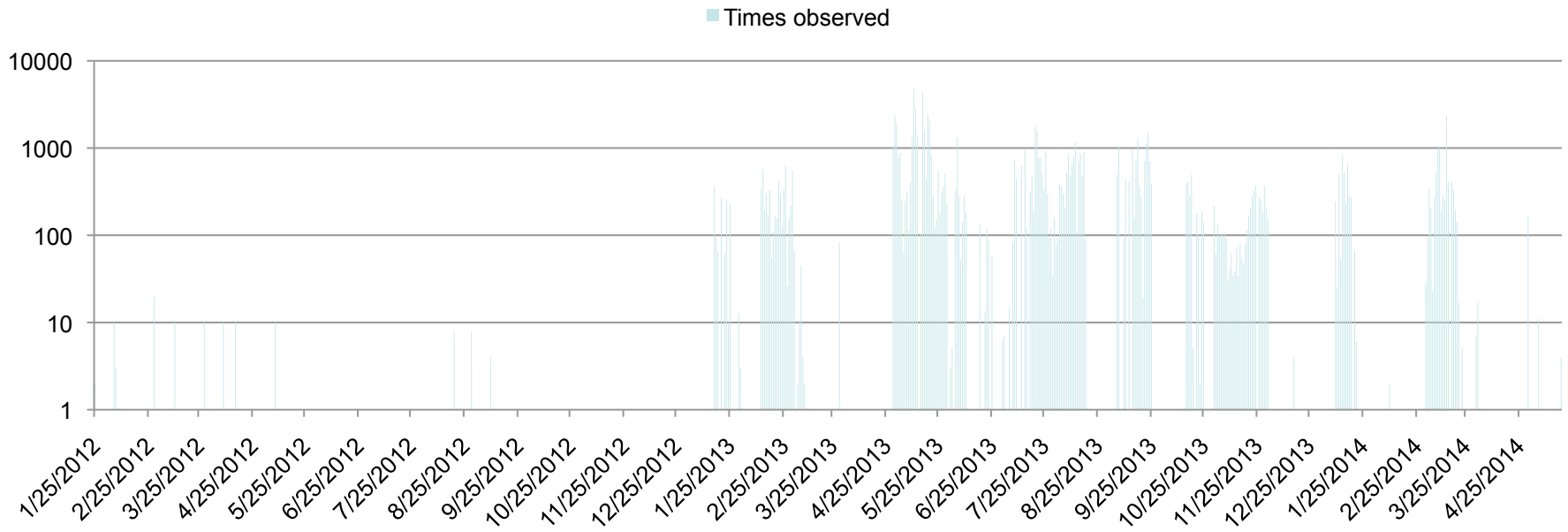
- * **Caveat**: *On Collisions for MD5*, M.M.J. Stevens, 2007

Tools to exploit observables

- PEClass
  - Parsing library for Portable Executables
  - Provides access to header values, sections
    - Also provides access to unaddressed “slack” space
  - Parses several types of sections
    - Resources
    - Imports
    - MS Rich Header
  - Created to overcome differences between Windows Loader behavior and PECOFF v8 specification

Composite section hit list #4

- b8f8e51eaf8e1a935303ade6d8082622
  - Name: Yuner
  - Quantity seen: 92907
  - Earliest date: 24-Jan-2008
  - Peak: 10-May-2009
  - Most recent date: 18-May-2010



Detecting Insider Threats

Insider Threat Modeling from Case Data

Appendix B: Insider IT Sabotage Model

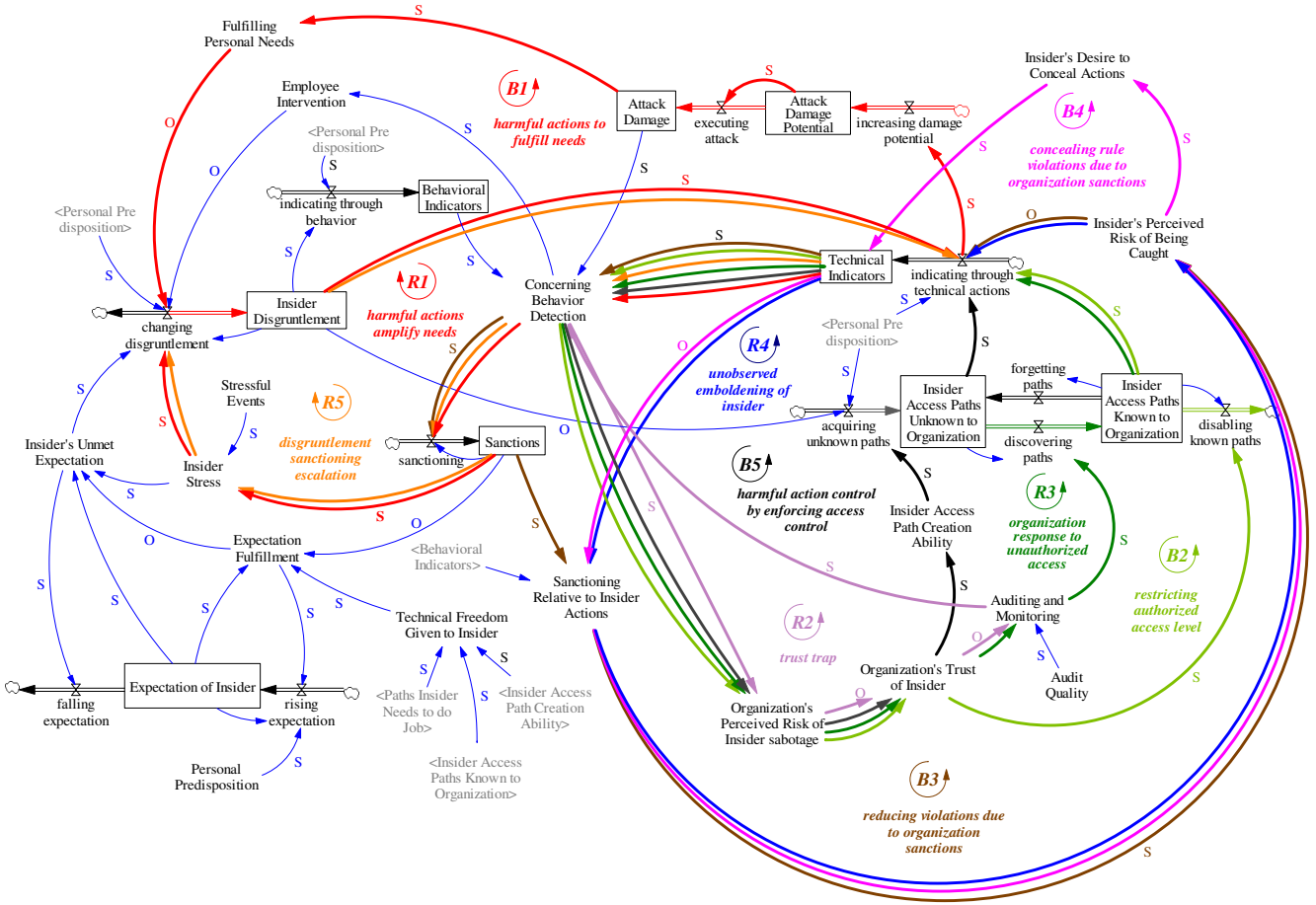
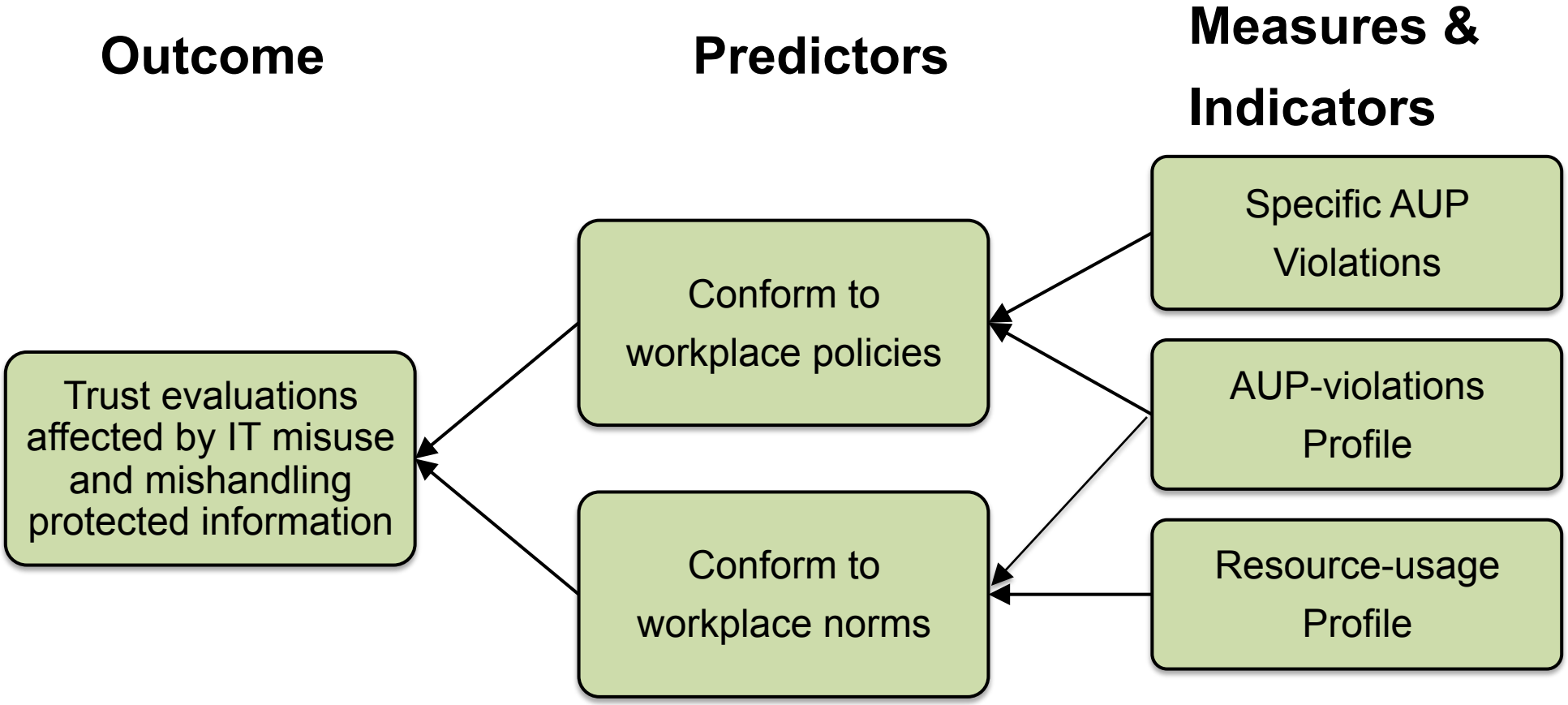


Figure 10: Insider IT Sabotage Model

Exploiting High-Fidelity Monitoring



Phase 1: Show that the measures are indicators for the predictors.

Phase 2: Show that the predictors predict the outcome.

Predictors face-validated with manager-provided staff assessments.

CERT Contact Information

Rich Pethia, Program Director
(412) 268-7739
rdp@sei.cmu.edu

Archie Andrews, Technical Director
Survivable Systems Engineering
(412) 268-9217
ada@sei.cmu.edu

Roman Danyliw, Technical Director
Cyber Threat and Vulnerability Analysis
(412) 268-5466
rdd@sei.cmu.edu

Barbara Laswell, Technical Director
Enterprise Workforce Development
(412) 268-5466
blaswell@sei.cmu.edu

Rich Nolan, Technical Director
Digital Intelligence and Investigations
(412) 268-3619
ran@sei.cmu.edu

Bill Wilson, Deputy Director
(412) 268-8003
wrw@sei.cmu.edu

Greg Shannon, Chief Scientist
(412) 268-8545
shannon@sei.cmu.edu

www.cert.org
www.sei.cmu.edu/security

Lawyers on Parade

▪NO WARRANTY

▪THIS MATERIAL OF CARNEGIE MELLON UNIVERSITY AND ITS SOFTWARE ENGINEERING INSTITUTE IS FURNISHED ON AN “AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

▪Use of any trademarks in this presentation is not intended in any way to infringe on the rights of the trademark holder.

▪This Presentation may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

▪This work was created in the performance of Federal Government Contract Number FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. The Government of the United States has a royalty-free government-purpose license to use, duplicate, or disclose the work, in whole or in part and in any manner, and to have or permit others to do so, for government purposes pursuant to the copyright license under the clause at 252.227-7013.