PREDICT Overview

PREDICT Workshop
Newport Beach, CA
September 27, 2005

Douglas Maughan, Ph.D.

Program Manager, HSARPA

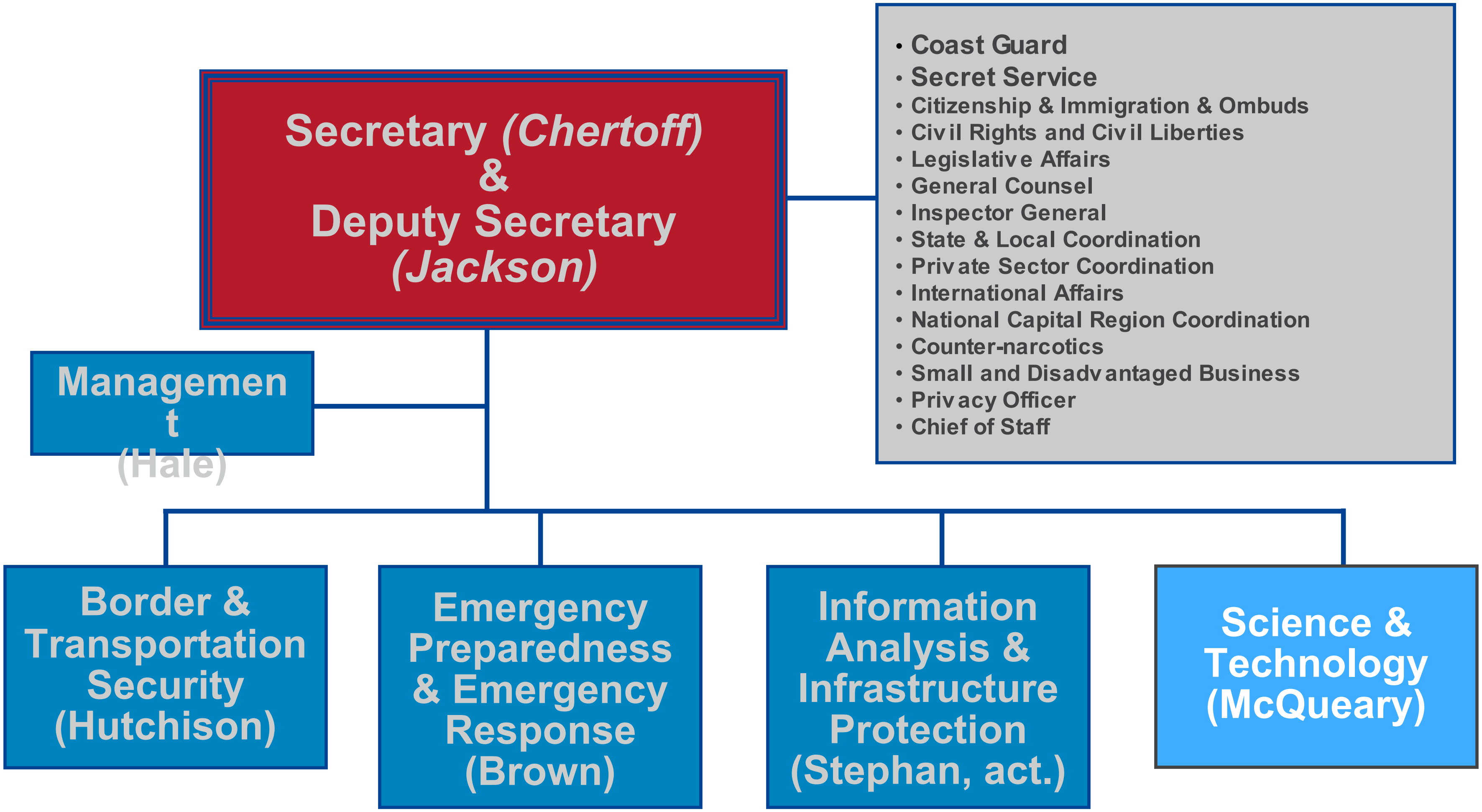
douglas.maughan@dhs.gov

202-254-6145 / 202-360-3170

PREDICT Workshop Objectives

- Describe Cyber Security Research and Development activities within DHS S&T
- PREDICT Overview
- PREDICT Operations and Processes
  - Randy Lucas, RTI
- PREDICT Legal Aspects
- Current PREDICT datasets
  - Several presentations by data providers
- Discussion of future dataset requirements
- Data Anonymization mini-workshop
  - Led by Phil Porras, SRI

General DHS Organization (prior to 7/13/05)



Secretary *(Chertoff)*
&
Deputy Secretary
(Jackson)

- Coast Guard
- Secret Service
- Citizenship & Immigration & Ombuds
- Civil Rights and Civil Liberties
- Legislative Affairs
- General Counsel
- Inspector General
- State & Local Coordination
- Private Sector Coordination
- International Affairs
- National Capital Region Coordination
- Counter-narcotics
- Small and Disadvantaged Business
- Privacy Officer
- Chief of Staff

Managemen
t
(Hale)

Border & Transportation Security (Hutchison)

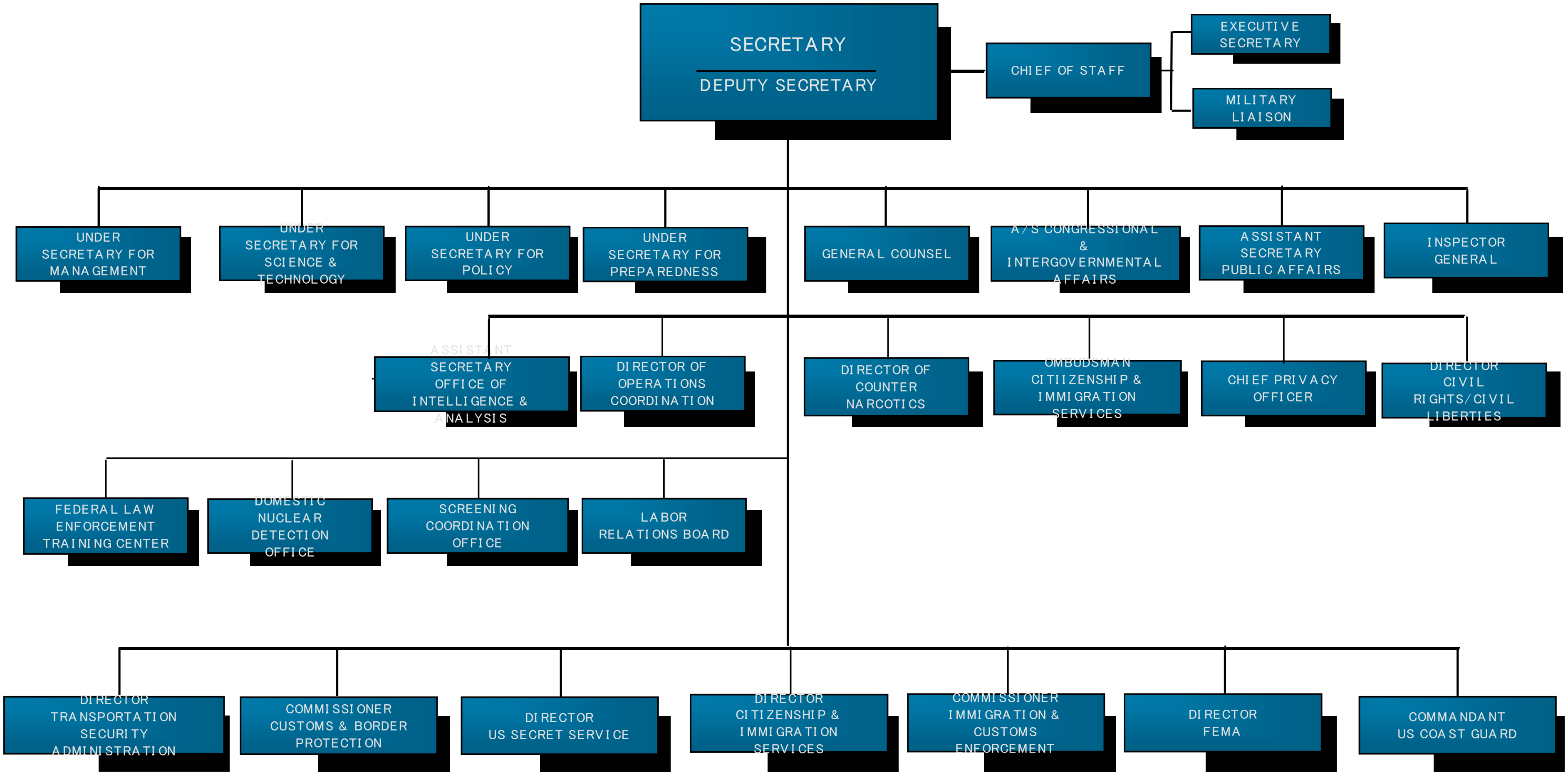
Emergency Preparedness & Emergency Response (Brown)

Information Analysis & Infrastructure Protection (Stephan, act.)

Science & Technology (McQueary)

Organization Chart

(proposed end state)



SECRETARY

DEPUTY SECRETARY

CHIEF OF STAFF

EXECUTIVE SECRETARY

MILITARY LIAISON

UNDER SECRETARY FOR MANAGEMENT

UNDER SECRETARY FOR SCIENCE & TECHNOLOGY

UNDER SECRETARY FOR POLICY

UNDER SECRETARY FOR PREPAREDNESS

GENERAL COUNSEL

A/S CONGRESSIONAL & INTERGOVERNMENTAL AFFAIRS

ASSISTANT SECRETARY PUBLIC AFFAIRS

INSPECTOR GENERAL

ASSISTANT SECRETARY OFFICE OF INTELLIGENCE & ANALYSIS

DIRECTOR OF OPERATIONS COORDINATION

DIRECTOR OF COUNTER NARCOTICS

OMBUDSMAN CITIIZENSHIP & IMMIGRATION SERVICES

CHIEF PRIVACY OFFICER

DIRECTOR CIVIL RIGHTS/CIVIL LIBERTIES

FEDERAL LAW ENFORCEMENT TRAINING CENTER

DOMESTIC NUCLEAR DETECTION OFFICE

SCREENING COORDINATION OFFICE

LABOR RELATIONS BOARD

DIRECTOR TRANSPORTATION SECURITY ADMINISTRATION

COMMISSIONER CUSTOMS & BORDER PROTECTION

DIRECTOR US SECRET SERVICE

DIRECTOR CITIZENSHIP & IMMIGRATION SERVICES

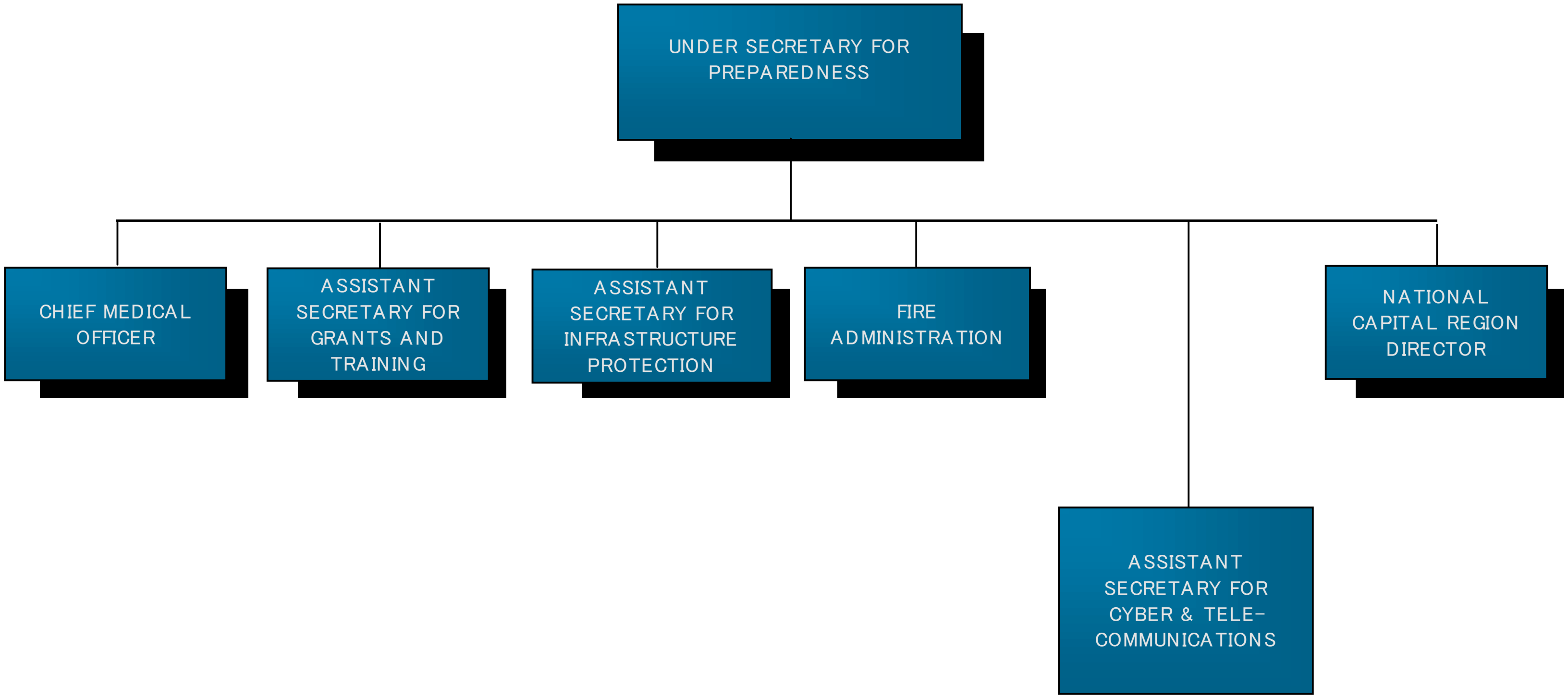
COMMISSIONER IMMIGRATION & CUSTOMS ENFORCEMENT

DIRECTOR FEMA

COMMANDANT US COAST GUARD

Department of Homeland Security
Organization Chart—Preparedness

(proposed end state)



UNDER SECRETARY FOR PREPAREDNESS

CHIEF MEDICAL OFFICER

ASSISTANT SECRETARY FOR GRANTS AND TRAINING

ASSISTANT SECRETARY FOR INFRASTRUCTURE PROTECTION

FIRE ADMINISTRATION

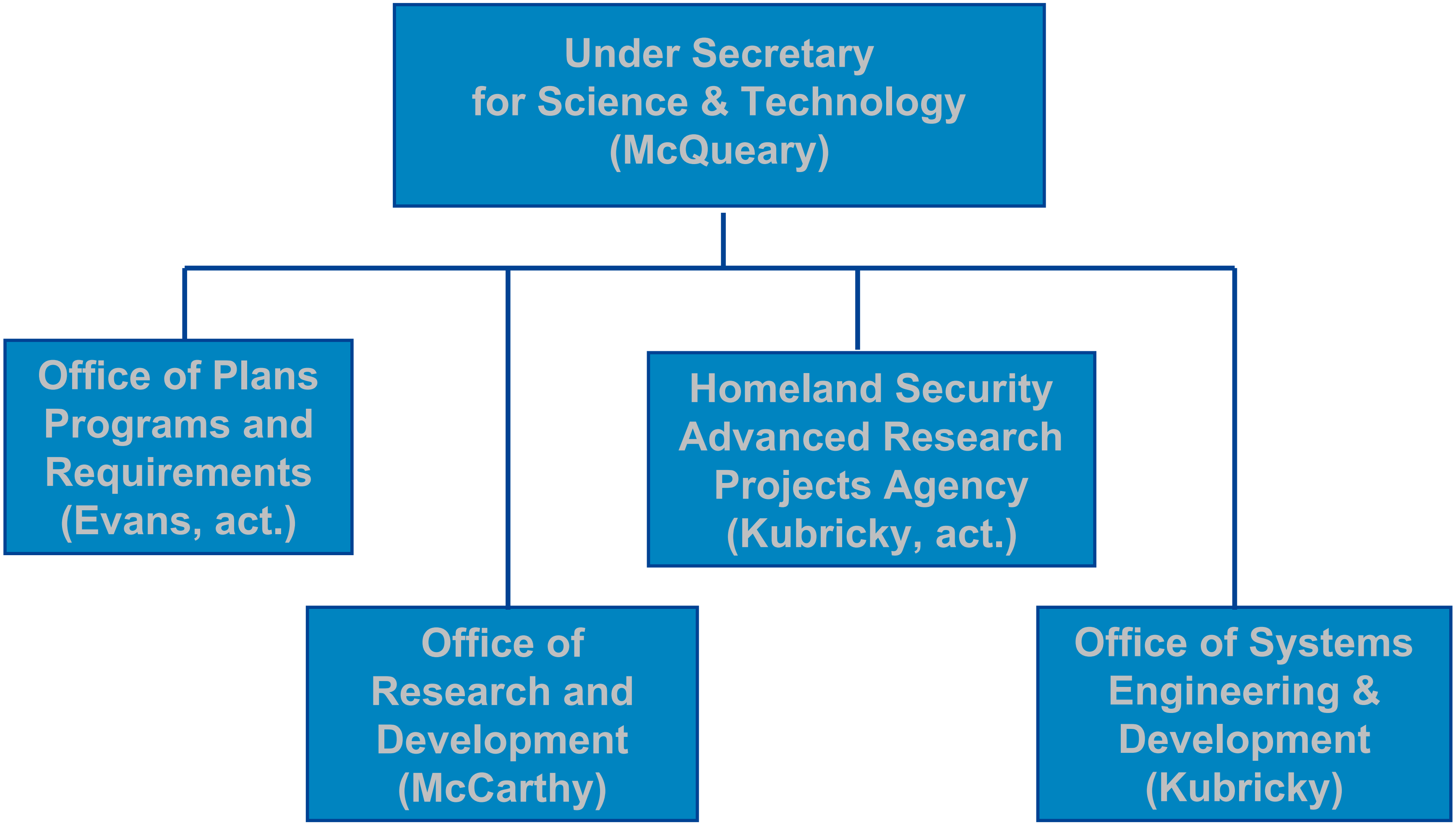
NATIONAL CAPITAL REGION DIRECTOR

ASSISTANT SECRETARY FOR CYBER & TELE-COMMUNICATIONS

Science and Technology (S&T) Mission

Conduct, stimulate, and enable **research, development, test, evaluation and timely transition** of homeland security capabilities to federal, state and local operational end-users.

S&T Organization Chart



Under Secretary for Science & Technology (McQueary)

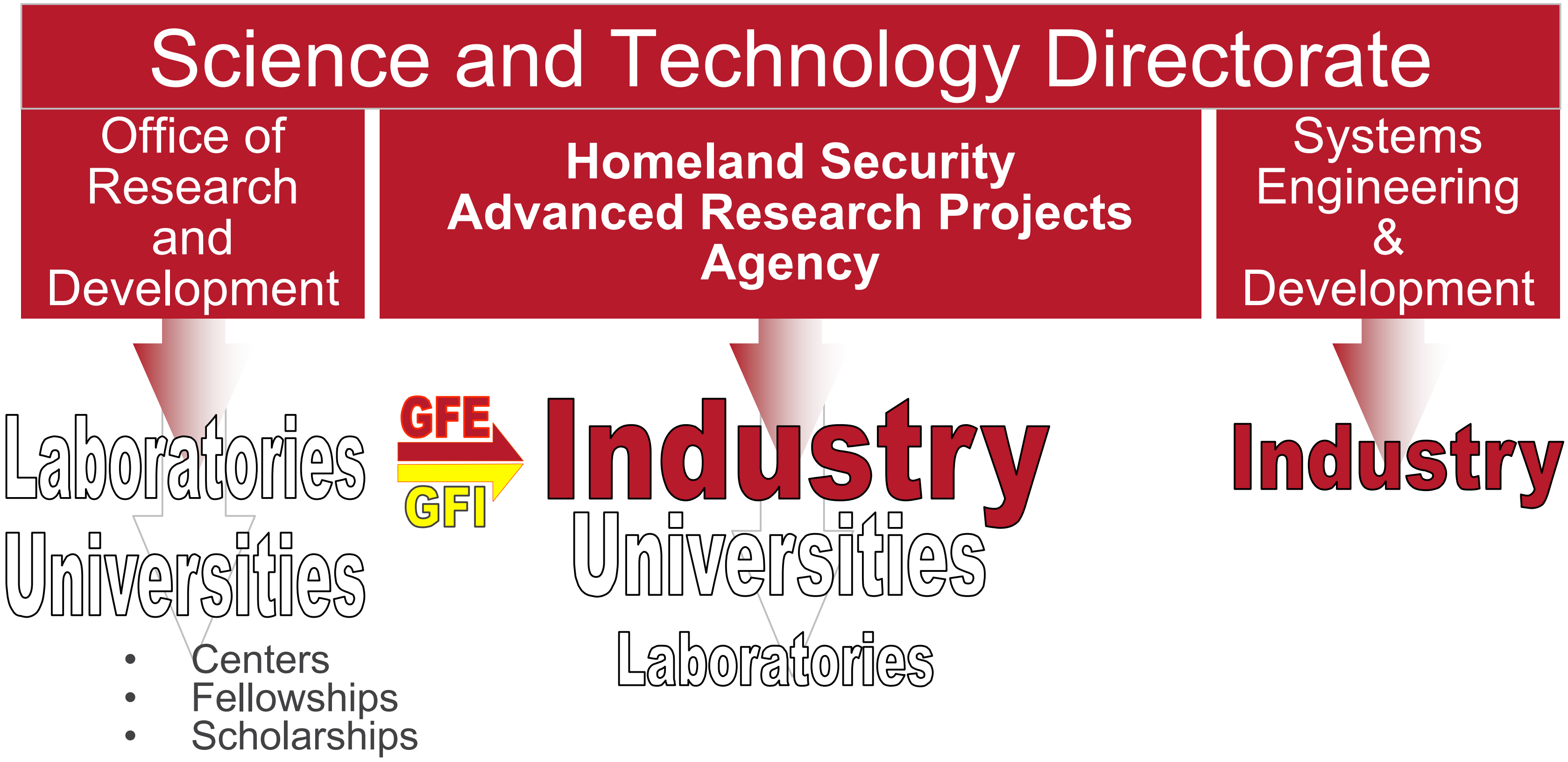
Office of Plans Programs and Requirements (Evans, act.)

Office of Research and Development (McCarthy)

Homeland Security Advanced Research Projects Agency (Kubricky, act.)

Office of Systems Engineering & Development (Kubricky)

Execution



Science and Technology Directorate

Office of Research and Development

Homeland Security Advanced Research Projects Agency

Systems Engineering & Development

Laboratories

Universities

- Centers
- Fellowships
- Scholarships

GFE

GFI

Industry

Universities

Laboratories

Industry

Stewardship of an enduring capability

Innovation, Adaptation, & Revolution

Development Engineering, Production, & Deployment

PREDICT

Protected Repository for the Defense of Infrastructure Against Cyber Threats

Crosscutting Portfolio Areas

- Chemical
- Biological
- Radiological
- Nuclear
- High Explosives
- Cyber Security
- Critical Infrastructure Protection (CIP)
- USSS

Legacy of HSARPA Name

How is it different from DARPA?

- **Differences**
  - **85-90% of funds for identified DHS requirements**
  - **10-15% of funds for revolutionary research**
    - **Breakthroughs,**
    - **New technologies and systems**
  - **These percentages likely to change over time, but we need to meet today’s requirements**

HSARPA Funding

HSARPA funding is **allocated** from Appropriated line items

SCIENCE AND TECHNOLOGY DIRECTORATE
FY 2005 Budget Execution Distribution
Dollars $M
FY05 Allocations 1 NOV04

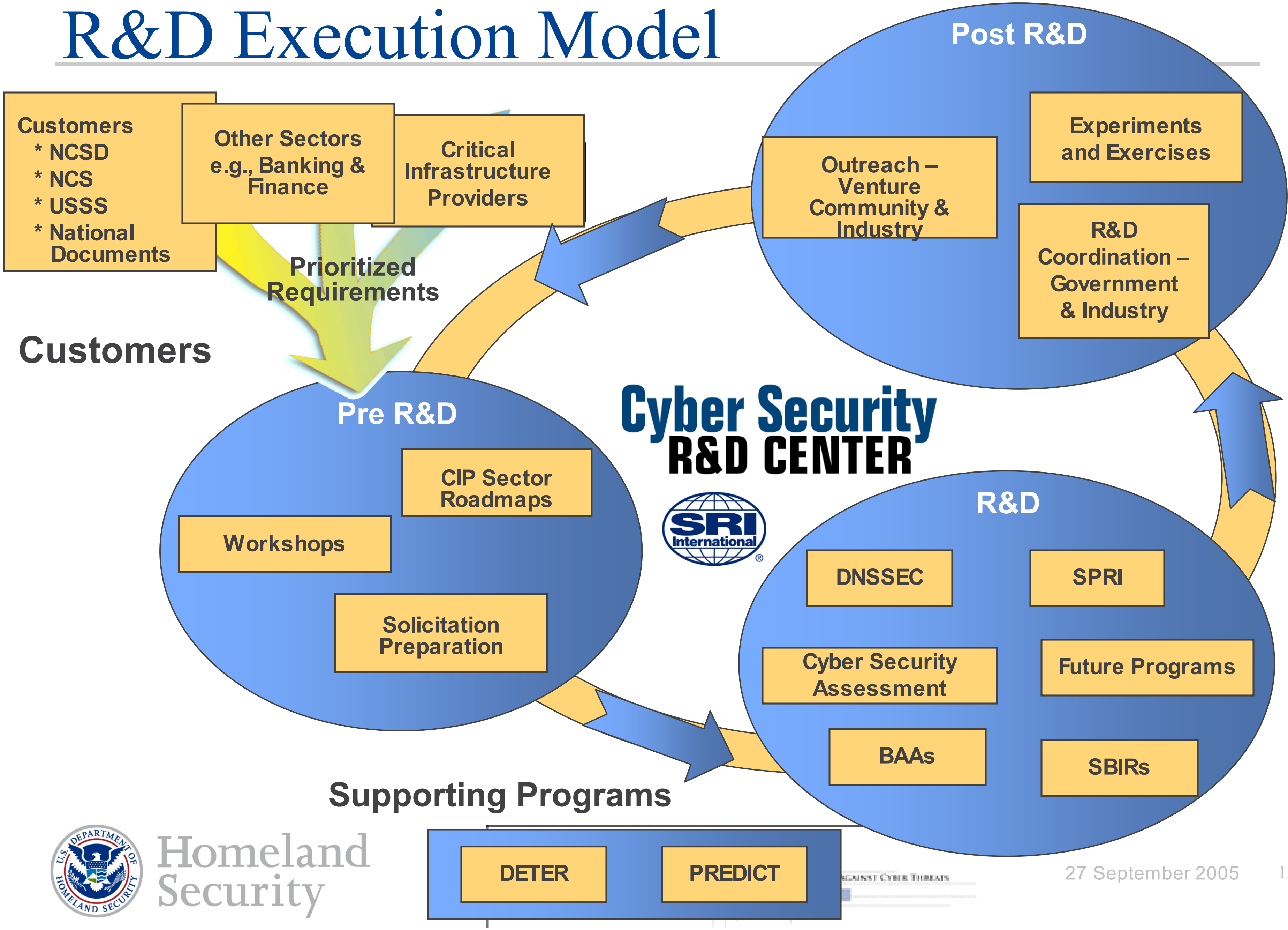
PROGRAM ELEMENT	FY 2005	**HSARPA**
PORTFOLIO/PAD	Appropriation	
Biodefense	362.7	**77.8**
Rapid Prototyping	76.0	**56.9**
Rad/Nuc	122.6	**39.0**
Chemical Countermeasures	53.0	**33.0**
Threat and Vulnerability Testing and Assessment	65.8	**5.0**
High Explosives	19.7	**3.9**
Standards	39.7	**0.0**
University Programs/Fellowships	70.0	**0.0**
Critical Infrastructure Protection	27.0	**4.0**
Conventional Missions	54.7	**26.8**
Emerging Threats	10.8	**4.0**
National Biodefense Analysis and Countermeasures Center (NBACC)	35.0	**0.0**
Cyber Security	18.0	**15.8**
Counter-MANPADS	61.0	**0.0**
Safety Act	10.0	**0.0**
Office of Interoperability and Compatibility	21.0	**0.0**
SBIR		**23.0**
Grand Total	1,046.9	**289.1**

Cyber Security R&D Portfolio: Scope

- **DHS S&T focus is on those research and operational threats and issues that warrant national-level concerns**

- The Internet serves a significant underlying role in many of the Nation’s critical infrastructures
  - Communications, monitoring, operations and business systems
- Adversaries face asymmetric offensive / defensive capabilities with respect to traditional warfare
  - Makes cyberspace an appealing battleground
- The most significant cyber threats to the nation are very different from “script-kiddies” or virus writers

R&D Execution Model



Customers
- * NCSD
- * NCS
- * USSS
- * National Documents

Other Sectors e.g., Banking & Finance

Critical Infrastructure Providers

Prioritized Requirements

Customers

Pre R&D

CIP Sector Roadmaps

Workshops

Solicitation Preparation

Post R&D

Experiments and Exercises

Outreach – Venture Community & Industry

R&D Coordination – Government & Industry

Cyber Security R&D CENTER

SRI International®

R&D

DNSSEC

SPRI

Cyber Security Assessment

Future Programs

BAAs

SBIRs

Supporting Programs

DETER

PREDICT

AGAINST CYBER THREATS

Homeland Security

A Protected REpository for Defense of Infrastructure against Cyber Threats

- PREDICT Program Objective

  “To advance the state of the research and commercial development (of network security ‘products’) we need to produce datasets for information security testing and evaluation of maturing networking technologies.”

- Rationale / Background / Historical:
  - Researchers with insufficient access to data unable to adequately test their research prototypes
  - Government technology decision-makers with no data to evaluate competing “products”

End Goal: Improve the quality of defensive cyber security technologies

Homeland Security

PREDICT

PROTECTED REPOSITORY FOR THE DEFENSE OF INFRASTRUCTURE AGAINST CYBER THREATS

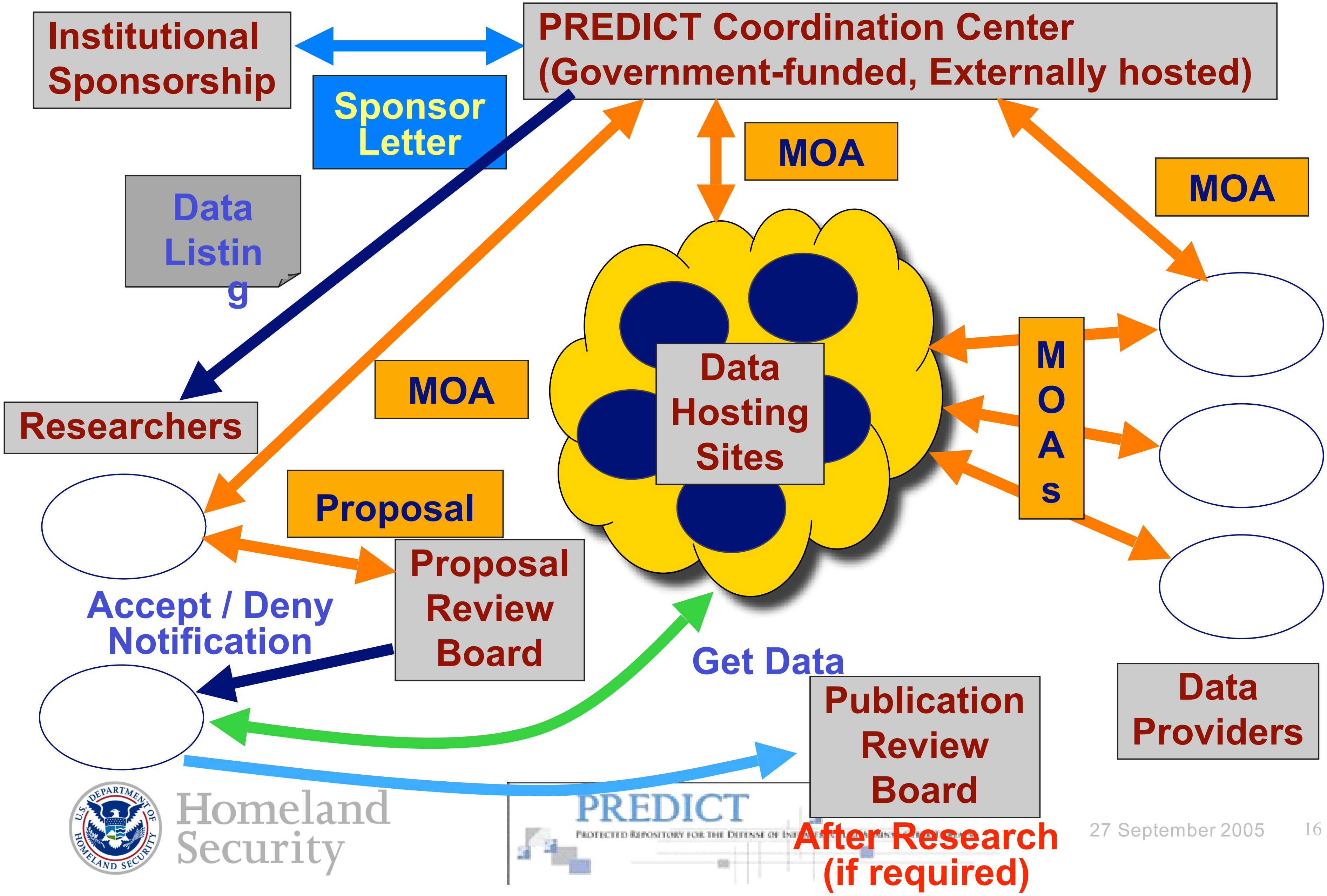
Industry Workshop (2/11-12/2004)

- Begin the dialogue between HSARPA and industry as it pertains to the cyber security research agenda
- Discuss existing data collection activities and how they could be leveraged to accomplish the goals of this program
- Discuss data sharing issues (e.g., technical, legal, policy, privacy) that limit opportunities today and develop a plan for navigating forward
- Develop a process by which "data" can be "regularly" collected and shared with the network security research community

ATTENDEES

AOL	
UUNET	
Verio	PREDICT participant
XO Comms	
Akamai	
Arbor Networks	
System Detection	
Cisco	
PCH	PREDICT participant
Symantec	
USC-ISI	PREDICT participant
Univ. of WA	PREDICT participant
CERT/CC	
LBNL	PREDICT participant
Internet2	PREDICT participant
CAIDA	PREDICT participant
Merit Networks	PREDICT participant
Citigroup	

PREDICT Repository Access Process



Institutional Sponsorship

Sponsor Letter

PREDICT Coordination Center (Government-funded, Externally hosted)

Data Listing

MOA

Researchers

MOA

Proposal

Proposal Review Board

Accept / Deny Notification

Data Hosting Sites

MOA

MOAs

Data Providers

Get Data

Publication Review Board

After Research (if required)

Data Collection Activities

- Classes of data that are interesting, people want collected, and seem reasonable to collect
  - Netflow
  - Packet traces – headers and full packet (context dependent)
  - Critical infrastructure – BGP and DNS data
  - Topology data
  - IDS / firewall logs
  - Performance data
  - Network management data (i.e., SNMP)
  - VoIP (1400 IP-phone network)
  - Blackhole Monitor traffic

Netflow Traffic Measurement

- This data consists of statistics regarding data collected from routers. It identifies two end points. Raw NetFlow data will indicate not only traffic totals, but the application breakdowns at each peering point. The individual flows will be stored in a method compatible with free analysis tools. All IP addresses will be anonymized.
- Research Use: Netflow data can be used to develop tools and techniques that will lead to a better understanding of the tradeoffs in traffic management. It can also be used in the development of anomaly detection and intrusion detection applications and for traffic characterization applications, such as self similarity and bottleneck bandwidth estimation applications.
- Data Provider(s): U Michigan/Merit Networks, Internet 2 (Host: U Michigan/Merit Networks), Los Nettos

Enterprise Data

- This data is internal traffic data from a large enterprise. It will consist of only headers with anonymized IP addresses. No content will be included. Prior to this effort, there have been no such enterprise background traces available, to the significant detriment of researchers attempting to devise enterprise-level network security mechanisms that will actually work soundly in practice.
- Research Use: The principle security-oriented use of the enterprise datasets will be as background traffic. By providing a large amount of real network traffic, the goal is to provide a resource for researchers to use in assessing the false positive rates and/or collateral damage of deploying proposed detection algorithms.
- Data Provider: LBNL

BGP Routing Data

- This dataset captures “snapshots” of the topological state of the Internet by archiving Border Gateway Protocol (BGP) routing tables from Internet routers in many locations around the world (these are called Internet Exchange Points). Each routing table expresses the “view” of the Internet from that router’s point in the overall topology and, taken together, all of these views provide a relatively complete roadmap of the connectivity within the Internet Service Provider core of the Internet. This dataset contains only backbone topology information; it does not contain any packet header information or information which relates to individuals.
- Research Use: BGP Routing Table Data is used by researchers who study the overall growth patterns of the Internet over time, as well as those who are looking specifically at individual carriers, regions, or resources. It shows historical trends in the utilization of the two principal Internet resources, IP addresses and Autonomous System Numbers (ANS), and this presents the basic backdrop against which many other trends are tracked.
- Data Provider(s): U Michigan/Merit Networks, Packet Clearing House (PCH), Internet 2 (Host: U Michigan/Merit Networks)

DNS Root Server Data

- This is root server data from the hosts of major DNS root servers. The data will identify the user by IP address. It will show what site is asked for, but it will not indicate whether the person associated with the IP address actually connected to that site. Generally, requests are aggregated by multiple users, but some are not. All IP addresses will be anonymized.
- Research Use: This data will be used for DNS root server traffic analysis and characterization and DNS root server attack analysis and characterization.
- Data Provider(s): Internet Software Consortium (Host: CAIDA), Los Nettos

Topology Measurement Data

- This data is obtained from computers that the data provider puts on the network in order to map the network connections of the Internet connecting out from that point. The computers send out probe packets with Time to Live (TTL) (the number of machines that can touch a packet before it gets sent back). The packet is owned by the data provider. It is sent out and comes back with information about routing, but no data is transmitted in the process. The data provider makes an Anonymous System (AS) core and ISP level map of Internet connectivity. The data provider requests that researchers not probe certain Internet Protocol (IP) addresses or disclose IP addresses to anyone else.
- Research Use: Better understanding of Internet traffic, latency, connectivity, and stability
- Data Provider(s): CAIDA, Los Nettos

Intrusion Detection Logs

- An intrusion detection system scans traffic to detect unauthorized or malicious activity. When it detects an attack, it can trigger protective actions. It is essentially a sensor that is watching for malicious activity.
- Research Use: Researchers can study IDS traffic in order to understand the evolution, rise, and decay of malicious traffic. It is possible to identify the end point responsible for originating the suspicious activity.
- Data Provider: University of Wisconsin (Host: Univ. of Michigan/Merit Networks), University of Washington (Host: U Michigan/Merit Networks)

Firewall Logs

- Firewalls detect distributed denial of service (DDOS) attacks and other malicious activity. Firewall logs contain detailed information regarding the end point that is directing harmful activity towards the network they are protecting. They contain the number of packets, origin, and where it went. All IP addresses will be anonymized.
- Research Use: Firewall logs are used by researchers that study attempted attacks on systems, such as port scanning, DDOS traffic, worm traffic, and can also be used for detection of insider threat activity.
- Data Provider: University of Wisconsin (Host: Univ. of Michigan/Merit Networks)

VoIP End-to-End Quality Data

- This dataset contains data which characterizes the quality of the paths which Voice Over Internet Protocol (VOIP) telephone calls take across the global Internet. It consists of anonymized Session Initiation Protocol teardown messages collected from both ends of the conversations on the INOC-DBA hotline phone system, and includes call duration; volume of data sent; number of packets sent, received and lost; and the number delivered out of order. The endpoints of the call are identified by country, Autonomous System Numbers (ASN), and subnet, but anonymity is preserved by not including either IP address or the caller or called phone numbers.
- Research Use: It is anticipated that the VOIP End-to-End Quality data will be used by researchers who wish to compare differential quality of service in similar and dissimilar regions of the Internet, such as across different backbone carriers which utilize different technology or capacity-planning methodologies. The data could also be used by researchers who are interested in correlations between the quality of service underlying voice communications and users patterns of utilization.
- Data Provider: Packet Clearing House (PCH)

Blackhole Address Space Data

- The data provider owns a large number of IP addresses. Traffic to legitimate addresses owned by the provider is delivered, and the remainder goes back to the data provider because the traffic is targeting unassigned IP addresses. Since this traffic was targeting illegitimate IP addresses, it is usually malicious traffic such as scanners and worms. In addition, all IP addresses will be anonymized.
- Research Use: This data can be useful for studying backscatter from distributed denial of service (DDOS) attacks, worm spread (growth rates, population size, and affected population), scanning and backdoor activity, and evaluating various honeypot responders.
- Data Provider(s): CAIDA, U Michigan/Merit Networks

What other things have we done?

- Internal Pilot
  - 6 weeks – Mid-January to end of February
  - 18 participants – academia, industry, government
  - Tested all parts of the system (except Pub Review Board)
- Sandia Red Team Evaluation
  - PREDICT portal penetration testing (outsider)
  - Data exfiltration testing (insider)

What’s Next??

- Final PIA and MOAs submitted to DHS Privacy Office
  - Comments received; now addressing
- DHS Privacy Office to post on their website
- Inform research community that PREDICT system is available for use
  - Purpose of this workshop
- Monitor usage, performance, issues, etc.
- Data Anonymization
  - Work out remaining issues
- Public Relations – several articles, etc. to be published over the next few months

Outstanding Issues

- Legal Issues
  - Current laws are written to ensure law enforcement doesn’t have unnecessary access to data
  - Because of these laws, Government researchers cannot have access to PREDICT data
  - However, …. We are working the issue with DHS lawyers to get things changed.

Summary

- DHS S&T is moving forward with an aggressive cyber security research agenda

- PREDICT is a national-level research resource that the cyber security community has really needed
- We hope you'll "pitch in" and help – as a provider and/or researcher

End Goal: Improve the quality of defensive cyber security technologies

Douglas Maughan, Ph.D.

Program Manager, HSARPA

douglas.maughan@dhs.gov

202-254-6145 / 202-360-3170

Homeland
Security

Back Up Slides

PCC – Provider MOA

- They will make the data available to data hosts, for release to approved researchers and no others, under the terms and conditions for access and use as specified by them and the PCC.
- They will provide the PCC with metadata on the data they agree to make available and they will not provide any data or metadata to anyone other than those researchers approved by the PCC.
- They will provide **terms and conditions for access to and use of the data**, including identification requirements for the data custodian; permitted uses and specific restrictions; minimum safeguards to protect the data; procedures for receipt, handling, control, dissemination, and return of data; and restrictions on publishing or releasing information about the data (which is addressed below under Publication Review Board).
- **They are responsible for ensuring that any data they release complies with all applicable statutes and regulations of applicable governing or regulating bodies and contractual agreements and is consistent with the provider's privacy, security, or other policies and procedures.**
- They certify that the data provided for use in the PREDICT program has been sanitized, de-identified, or cleaned of any and all information that would not be in compliance or consistent with the privacy requirements as determined by PCC and DHS.
- Non compliance with these requirements may result in the data provider's expulsion from the PREDICT project.

Back

PCC – Data Hosting Site MOA

- They will accept data from approved data providers, for release to approved researchers, subject to the terms and conditions set forth by the providers and hosts.
- They will provide terms and conditions for access to, transfer, storage, and use of the data as required by the provider and PCC, as well as any other restrictions the host deems necessary to accomplish efficient and secure access to the data.
- They acknowledge that the data access approval given to a researcher in any application will permit access to the requested data by that researcher, regardless of approval or denial of access to that researcher in any other application.
- **They are solely responsible for ensuring that any data they release complies with the host's separate agreement with the data provider, all applicable statutes and regulations applicable to the data, and all contractual agreements it has with any other third parties. The host must also ensure that the data they release is consistent with their own privacy, security, or other policies and procedures**.

PCC – Researcher MOA

- They will not use the data for purposes other than described in their application.
- They **will not disclose the data to any persons other than those identified in their application**.
- They will establish and maintain the appropriate administrative, technical, and physical safeguards to protect the confidentiality of the data and to prevent unauthorized use of or access to the data.
- They will permit others to use the data only in accordance with the terms of the MOA and the procedures in the researcher’s application.
- If the researcher moves to a different institution, they will notify PCC and the sponsoring institution in writing regarding the disposition of all copies of the data and follow PCC’s directions and the sponsoring institution’s guidelines.
- No findings, analysis, or information derived from the data may be released if such findings contain any combination of data elements that might allow for identification or the deduction of a person’s or institution’s identity.

PCC – Researcher MOA (continued)

- Any findings, results of analysis, or manuscripts proposed for public release, publication, or any other type of disclosure to persons not listed and approved in this application (e.g., abstracts, presentations (oral or written), publications) must be submitted for a stringent review by the researcher's sponsoring institution and by a Publications Review Board managed by PCC prior to release to assure that data confidentiality is maintained, entities or individuals cannot be identified, and the terms and conditions attached to the use of the data have been followed.
- They **will report immediately to PCC any use or disclosure of the Data other than as permitted** and will take all reasonable steps to mitigate the effects of such improper use or disclosure, cooperating with all reasonable requests of PCC towards that end.
- In the event PCC determines or has a reasonable belief that researcher has violated any terms of the MOA, PCC may terminate the MOA and require the researcher to return the data and all derivative files. PCC may also seek injunctive relief against the researcher or the sponsoring institution to prevent any unauthorized disclosure of data. In addition, PCC will report any misuse or improper disclosure of the data to the data provider and host and to appropriate authorities as required by applicable Federal or state law.
- They will destroy all copies of the data when the MOA expires or as specified in the MOA and will certify such destruction or return by signing and providing to PCC a Certification of Data Return or Destruction.

Back

New Datasets

- Application oriented datasets
  - Phishing e-mails
  - E-commerce content (e.g., electronic trading)
- Tracking bio-chem
- Keystroke data with context
  - OS, processes
- Ground truth – attacks
  - Tools for extraction of attacks
  - Threat characterization – post-analysis
- First responder communications
- Malware data
  - Authors / Source, evolution, etc.

New datasets

- High volume/interaction honeypots
- IPv6, VOIP, IM, etc., etc. – usage specific
- Telecom
- P2P
- Anti-virus logs, zone alarm – Host IPS
- Steganographic traffic
- Systems forensics – File system, configurations, etc.
  - Stuart - MIT CSAIL
- Sensor network traffic

- Slides will be available at:
  http://www.hsarpacyber.com