



PREDICT

Protected Repository for Defense of Infrastructure against Cyber Threats

Manish Karir

DHS S&T



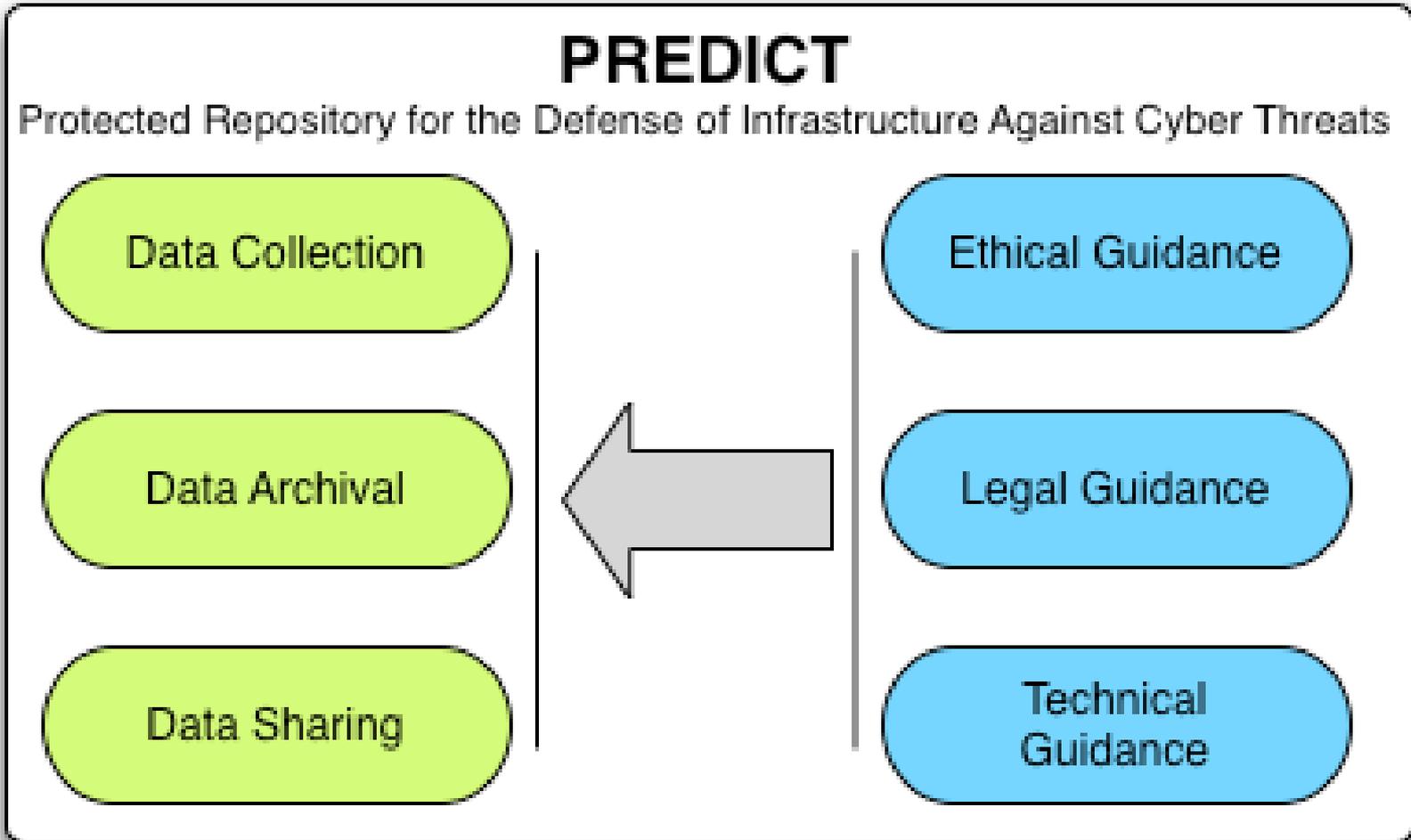
Background



- Rationale / Background / Historical:
 - Researchers with insufficient access to data unable to adequately test their research prototypes
 - Government technology decision-makers and researchers need data to evaluate competing “products”
 - Supports scientific method via repeatability of tests and evaluations
 - Unclear legal and ethical policies for Internet research
- PREDICT project is the only freely-available legally collected repository of large-scale datasets containing real network traffic and system logs.
- Dissemination of data is supported by a streamlined legal framework that controls distribution while protecting researchers, data providers and data hosts.



PREDICT - Overview





PREDICT Activities



- Support data collection activities to make high quality, timely and relevant dataset available to the research community.
- The development of systems for storage and processing of large volumes of data.
- Support the advancement of tools and techniques for analyzing Internet datasets to extract useful information and the representation of that information.
- Advance the state of the art in data collection techniques, packet formats, new data types, storage techniques, data cataloging/annotation, cross dataset analysis.
- Investigate and highlight legal and ethical issues in Internet data collection and analysis.



PREDICT – Lead Participants



CAIDA/
UCSD

ISI/USC

Colorado
State

University
of
Wisconsin

Packet
Clearing
House

University
of Michigan/
Merit

Georgia
Tech

RTI

Global
Cyber Risk

SRI



Data Categories in PREDICT



- Address Space Allocation Data
- BGP Routing Data
- Blackhole Address Space Data (Darknet)
- DNS Data
- Infrastructure Data
- Internet Topology Data
- IP Packet Headers
- Performance and Quality Measurements
- Synthetically Generated Datasets
- Traffic Flow Data
- Unsolicited Bulk Email Data



Dataset Details – Examples

- Address Space Allocation Data
 - Internet-wide census via probing. Provides information about visibility of hosts on the Internet
 - Used to study Internet outages, understanding of development and growth of the Internet globally
- Internet Topology Data
 - Contains router-level graphs of the Internet, and/or router-level paths through the network. Typically this data is obtained by carrying out traceroute-like probes from monitoring points around the network.
 - Used to support modeling and simulation of malware outbreak, spread, distribution, containment, and countermeasures, as well as macroscopic vulnerability assessments, longitudinal analysis and modeling of the evolution of Internet topology and address usage patterns.



Dataset Details – Examples (2)



- Blackhole Address Space Data (Darknet)
 - Dataset obtained by monitoring large portions of allocated but unused address space. Two kinds – partially spatially synchronized monitoring of last 30 IPv4 /8 address blocks and long term collections of over 5 years each
 - Used for analysis of Internet pollution, censorship studies, worm propagation, DDoS attack detection, Internet and malware propagation and Internet growth and evolution
- Traffic flow Data
 - Datasets obtained by monitoring netflow at both local enterprise as well as regional level. Usually anonymized
 - Used for analysis of actual user traffic, protocol distributions, popularity of various applications, traffic flow distributions, popular destinations, impact of worms, prevalence of botnet infections



Data Providers

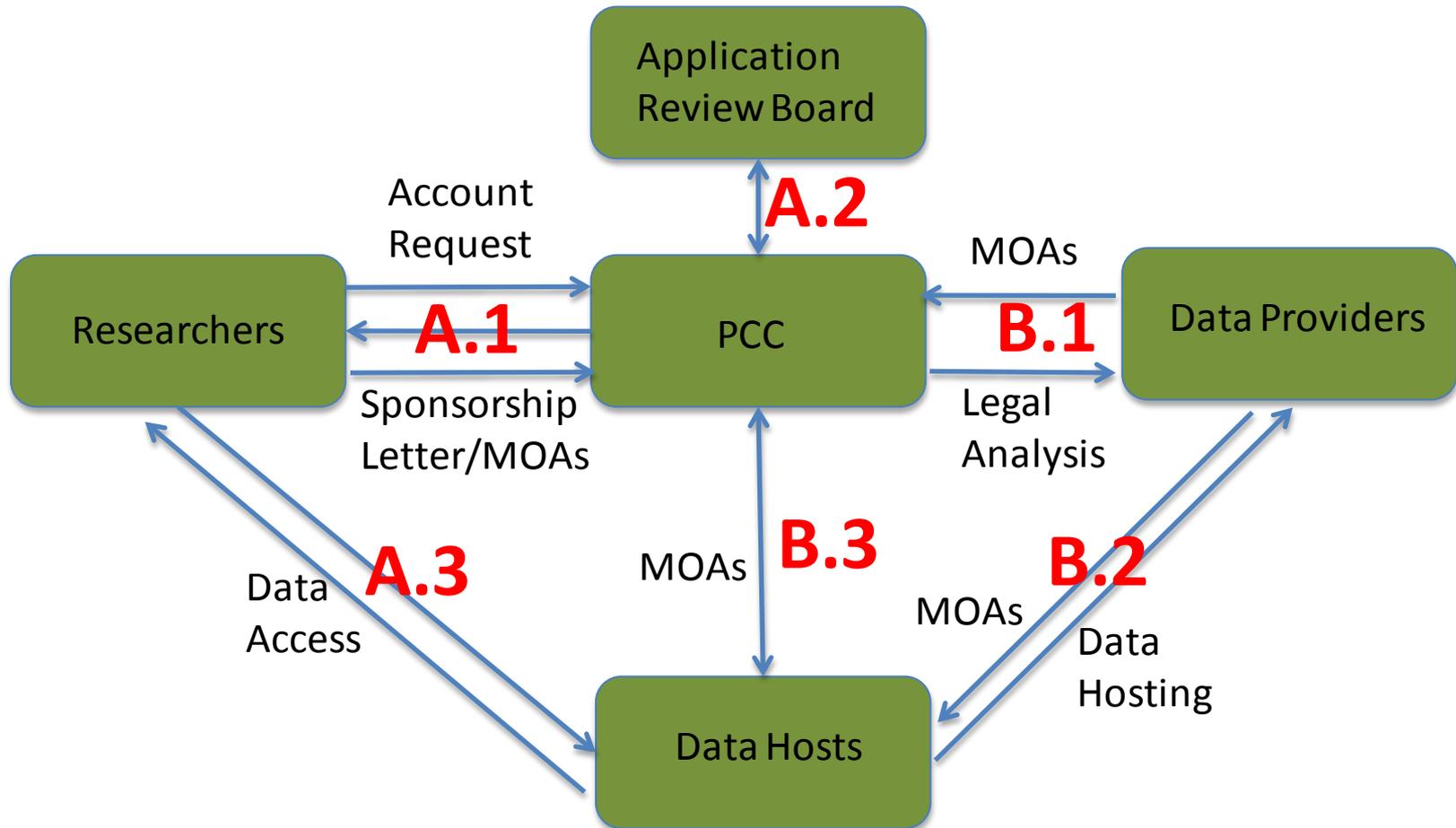


- CAIDA **~80,000 GB**
 - Topology Measurement Data, Network Telescope Data
- USC - LANDER **~50,100 GB**
 - NetFlow Data, Internet Topology Data, Address Allocation Data
- Colorado State University **~200 GB**
 - NetFlow Data, Spam logs
- Merit Networks **~200,000 GB**
 - Netflow Data, BGP Routing Data, DNS, Darknet Data
- University of Michigan **~1,000 GB**
 - Enterprise Netflow Data
- Georgia Tech *** TBD GB**
 - Bulk spam, BGP data, DNS
- University of Wisconsin **~500 GB**
 - Global Intrusion Detection Database
- Packet Clearing House **~10,000 GB**
 - BGP Routing Data, VoIP Measurement Data

TOTAL = ~350+ TB 9



PREDICT Repository Access





PREDICT Impact



- Over 250 research papers/journals/technical reports within the last 3 years, based on research done with data from the PREDICT data repository
- 350+ TB of data is available
- Research groups include:
 - 26 academic institutions
 - 24 commercial entities
 - 11 Government organizations
 - 4 non-profit organizations
- Menlo Report – First attempt at documenting ICTR ethical issues. Still working (expect another 2-3 years) to implement policies on ethical issues with universities, government agencies, professional groups (ACM, IEEE), etc.



PREDICT – Ethics of Internet Research



- The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research
 - First Published Sept 15 2011 - Comments from public and other stake holders
 - Under revision with final document released July 2012
- Propose a framework of ethical guidelines for Information and Communication Technology research
- Four core principles (3 follow Belmont Principles):
 - Respect for persons – respect for rights of individual protection for those with diminished capacity
 - Beneficence – Systematically assess possible harm and benefits of activity
 - Justice – Fair selection of participants, no specific targets
 - Respect for Law and Public Interest – Engage in legal due diligence

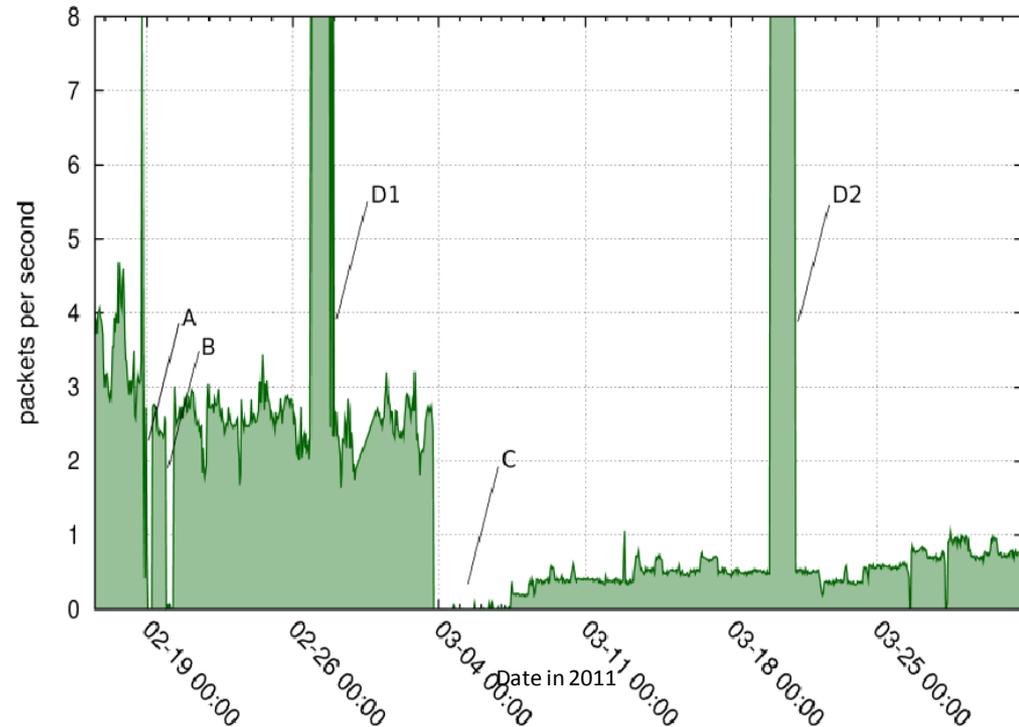


PREDICT Example Research (1)



Darknet Traffic Analysis

- Datasets are useful for studying network security-related events, such as DDoS attacks, the automatic spread of malware and scanning of IP address space by attackers looking for vulnerable targets. Examples:
 - Datasets provide granularity to identify geopolitical events (Internet censorship during recent political unrest in Egypt and Libya) and geophysical events (earthquakes in Japan and New Zealand)
 - Datasets could be used to create “early-warning” system to help detect Internet suppression activities in the future.



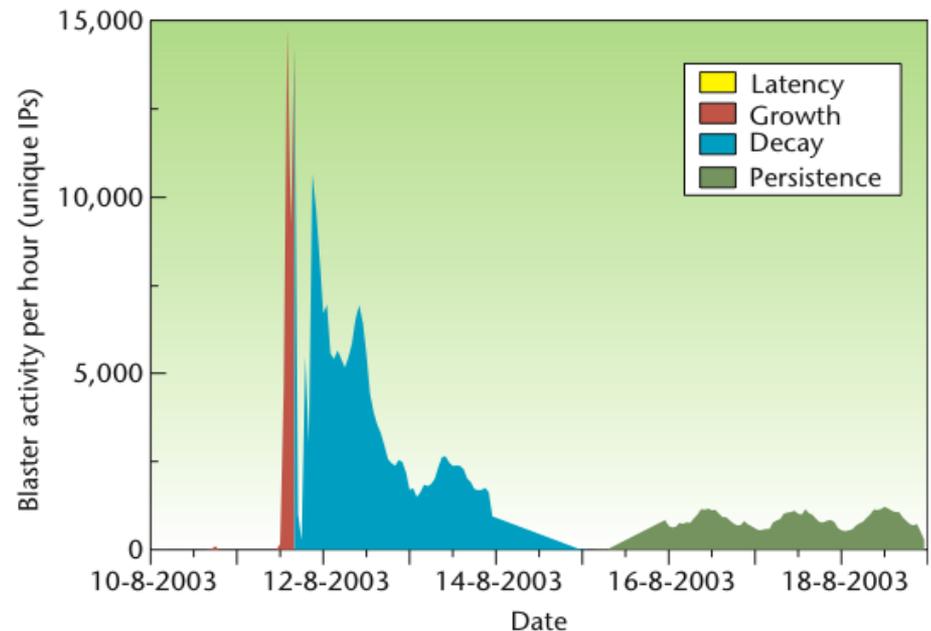
UCSD Darknet Traffic coming from Libya. Labels A, B, C indicate the three outages. Spikes labeled D1 and D2 are due to backscatter from two denial-of-service attacks.



PREDICT Example Research (2)



- Botnets, Worms, and DDoS attacks rank among the most significant operational threats to today's Internet
- The cost of building and operating such attacks are dropping while the size, sophistication, and targeting of these attacks is increasing
- Traffic Flow datasets help understand how these attacks are launched, and how effective new defenses are
- Traffic Flow datasets can help determine impact of a security event, detect impacted hosts, as well as how critical it is



The Blaster worm lifecycle



PREDICT Example Research (3)



- DNSFlow is a way to enable netflow like telemetry for DNS data on a network
- PREDICT funded effort has create this new visibility into local data which can allow network operators to uncover a wide range of impacting Internet events such as censorship as well as the impact of malware such as DNSChanger
- DNS datasets can also help determine structure of the Internet which is increasingly becoming hidden behind layers and layers of interconnected clouds
- Locally DNS data can reveal which hosts have been victims of successful phishing and malware infections

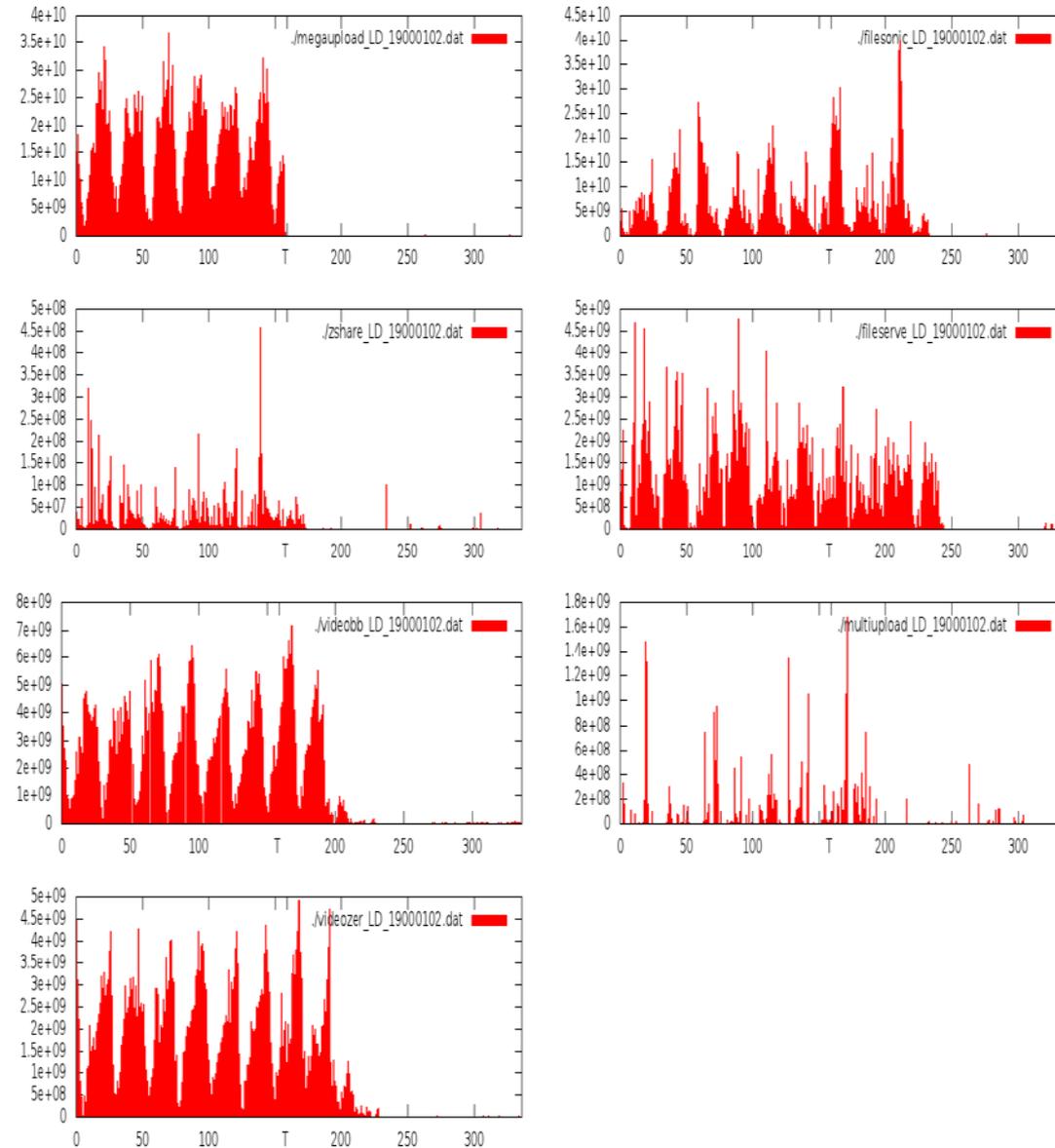
alert_id	set_id	IP f/ response	client IP	timestamp	BL source	name from DNS
7907823	4819173	xxx.235.133.xxx	10.1.1.2	2012-04-09 00:57:57.379013	malwaredomains	www.sexy-screen-savers.com.
7924562	4829607	xxx.233.142.xxx	10.1.1.1	2012-04-09 05:45:08.815553	malwaredomains	www.meb.gov.tr.



PREDICT Example Research (4)



Filesharing Traffic (Bytes)



- Netflow data can help to understand client population behaviors
- What would be the impact of service shutdowns and disruptions
- MegaUpload service shutdown but what was the broader impact
- Data analysis shows a chilling effect on other similar services as they scramble to modify their policies



Conclusions



- PREDICT's role encompasses a wide range of activities which includes support for enabling new methods of enhancing network data visibility, data storage, and data sharing activities
- Legal and Ethical issues are a primary concern and great effort is taken to ensure we do the *right thing*
- PREDICT datasets are the best datasets that help understand the Internet structure and performance
- PREDICT also supports a wide range of research activities which enhance our understanding of Internet data
- PREDICT processes are well defined and structured to ensure researcher requests are handled systematically and consistently



Conclusions



- Future Work includes:
 - Streamlining access to some data categories
 - Additional datasets and categories – Public/restricted and live streaming/archival data
 - Additional Data Access Methods such as VMs/Virtual Enclaves
 - Expansion of Ethics of Cybersecurity research to include the development of practical guidelines for research community
 - Inclusion of current/active datasets from the community

<https://www.predict.org>