PRESS KIT | 2017

ABOUT

The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program supports the global cyber risk research community by coordinating and developing real world data and information sharing capabilities – tools, models, and methodologies. In order to accelerate solutions around cyber risk issues and infrastructure security, the IMPACT program enables empirical data and information sharing between and among the global cyber security research and development (R&D) community in academia, industry and the government.

RESOURCES

URLS:

CSD: https://www.dhs.gov/csd-impact
PORTAL: https://www.impactcybertrust.org

Publications/Articles:

The Cyber Threat Under the Street
https://www.nytimes.com/2015/11/08/sunday-review/the-cyberthreat-under-the-street.html

Hurricane Sandy Internet Outage
http://www.isi.edu/~johnh/PAPERS/Heidemann12d.pdf

BotDigger: Detecting DGA Bots in a Single Network
http://www.cs.colostate.edu/~hanzhang/papers/BotDigger-techReport.pdf

Anycast Latency: How Many Sites Are Enough?
http://www.isi.edu/~johnh/PAPERS/Schmidt16a.pdf

Lost in Space: Improving Inference of IPv4 Address Space Utilization
https://www.caida.org/publications/papers/2014/lost_in_space/lost_in_space.pdf

CONNECT:

Erin Kenneally, Program Manager
erin.kenneally@hq.dhs.gov

Info@ImpactCyberTrust.org

FACT SHEET

DHS Science and Technology Directorate

Trusted Cyber Risk Research Data Sharing

Overview

The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program supports the global cyber risk research community by coordinating and developing real world data and information sharing capabilities – tools, models, and methodologies. In order to accelerate solutions around cyber risk issues and infrastructure security, the IMPACT program enables empirical data and information sharing between and among the global cyber security research and development (R&D) community in academia, industry and the government.

Objective

CSD seeks to coordinate, enhance and develop advanced data and information sharing tools, datasets, technologies, models, methodologies and infrastructure to strengthen the capabilities of national and international cyber risk R&D. These data sharing components are intended to be broadly available as national and international resources to bridge the gap between producers of cyber-risk-relevant ground truth data, academic and industrial researchers, cyber security technology developers, and decision makers in order to inform policy and analysis of cyber-risk and trust.

Background

Cyber security R&D requires real-world data to develop advanced knowledge, test products and technologies, and prove the utility of research in large-scale network environments. Established and funded by the U.S. Department of Homeland Security (DHS) Science and Technology Directorate (S&T) Cyber Security Division (CSD), the predecessor program --Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT)-- was the only freely-available, legally-collected distributed repository of large-scale datasets containing real network and system traffic to advance state-of-the-art cyber security R&D. It provided the research community access to infrastructure and event data to facilitate and develop tools, test theories, and identify workable solutions against cyber threats.

IMPACT Value Proposition

In April 2016, PREDICT transitioned to IMPACT: Information Marketplace for Policy and Analysis of Cyber-risk & Trust, a name meant to reflect an evolved implementation of the program's goals:

- Marketplace - A more open platform to connect and socialize data supply & demand;
- Policy and Analysis - Research infrastructure and analysis driven by and for real world issues; and
- Cyber-risk & Trust - Beyond just "defense" and "threats," information as a critical infrastructure itself; responsible innovation

IMPACT offers a unique, distributed research data repository supported by a streamlined legal framework and coordinated distribution of datasets. This centralized brokering and distributed provisioning between the data providers, data hosts and researchers addresses the operational, trust and administrative costs and challenges that impede sustainable and scalable data sharing.

IMPACT continually adds new data that is responsive to cyber risk management (e.g., threats, vulnerabilities, consequences) to provide the R&D community timely, high-value information to enhance research innovation and quality. The IMPACT model also serves as a laboratory for testing various data sharing models, including batch transfers, newer data-as-a-service (DaaS), and visualization techniques.

IMPACT consists of four components supporting core functional requirements for data sharing: metadata discovery, data and tool matchmaking, trusted brokering, and a social feedback loop.

(1) Metadata Indexing (Find) -- Open, comprehensive, centralized, standardized interface and engine to access metadata from the federation of providers and hosts.

(2) & (3) Data and Tools Matchmaking (Request and Use) -- Standardized policies and procedures to connect researchers with a federation of providers and hosts; Central interface and process to discover and access tools to analyze and/or use data from within and outside of IMPACT.

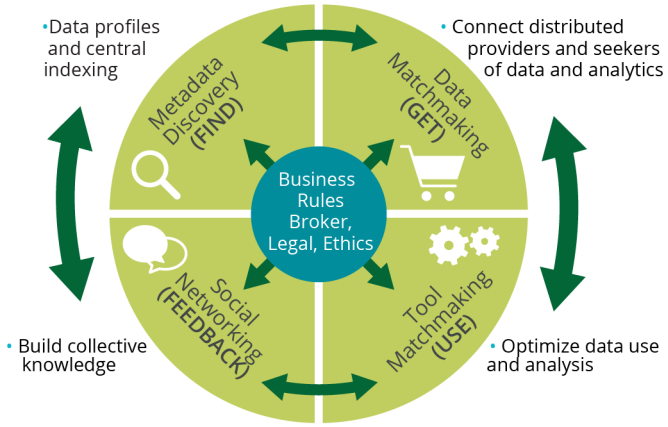
Homeland Security
Science and Technology

To learn more about Cyber Security Division's IMPACT Program, contact Erin Kenneally, Program Manager Erin.Kenneally@HQ.DHS.Gov

DHS Science and Technology Directorate
Trusted Cyber Risk Research Data Sharing

(4) Administrative, Legal, Ethical Brokering – Centralized interface, policies and procedures to request datasets from a federation of providers and hosts; Vetted data source provenance; Mediated access entitlement so sensitive data is shared with legitimate researchers.

(5) Social Networking (Feedback Loopo) – Central platform for exchanging feedback between providers, hosts, Rrsearchers, and domain experts to improve and optimize data, tools, analytics and collective knowledge.



Performers consist of Data Providers (DPs) and Data-Analytics-as-a-Service Providers (DASPs. These entities work independently and/or in concert to provision data and analytic tools using the IMPACT infrastructure, which will continue to mediate between providers and seekers of data and analytics. The DASPs role is to abstract away the low level knowledge- and labor-intensive elements that comprise high dimensional data identification, complex association and fusion, and high-context presentation elements of data for decision analytics. In general, DASPs are responsible for leveraging existing resources to enhance a data and analytical environment or standalone capability for the purpose of repeatedly and reliably providing storage and access to the decision analytics data and/or tools (e.g., visual analysis and awareness front end tools, analytics enabling tools, and/or algorithms).

Ethical Research and Risk-Sensitivity

Research data sharing activities are augmented by community-informed ethics and sensitive data disclosure guidance focused on the principles, controls, and responsible implementation of solutions to issues that impede cyber security research. *The Menlo Report, Ethical Principles Guiding Information and Communication Technology Research* and its *Companion* Report were the initial ethics outputs. They were inspired to pre-empt some of the issues coming down the pike, and to embrace a principals-in-context approach by applying the framework from the Belmont Report to modern information and communication technology research. Current efforts focus on tools and collaboration to apply and enforce research ethics while also supporting research innovation. Foundational work in utility-risk sensitive data disclosure control has similarly helped advance the sociotechnical aspects of research data sharing. These efforts are co-evolutionary with IMPACT, and along with the central broker component, they combine to strengthen trust in sharing data to enhance cyber security capabilities.

International Scope

Presently, more than a petabyte of data is available to the research community, and is being used by academic institutions, commercial entities, government organizations, foreign institutions, and nonprofit organizations. Currently, seven countries participate in IMPACT (Australia, Canada, Israel, Japan, United Kingdom, and Netherlands) with new partners to be on-boarded in the near future.

Join us

For more information visit dhs.gov/cyber-research or contact us at IMPACT@hq.dhs.gov.

Homeland Security
Science and Technology

To learn more about Cyber Security Division's IMPACT Program, contact Erin Kenneally, Program Manager Erin.Kenneally@HQ.DHS.Gov

SLIDE DECK

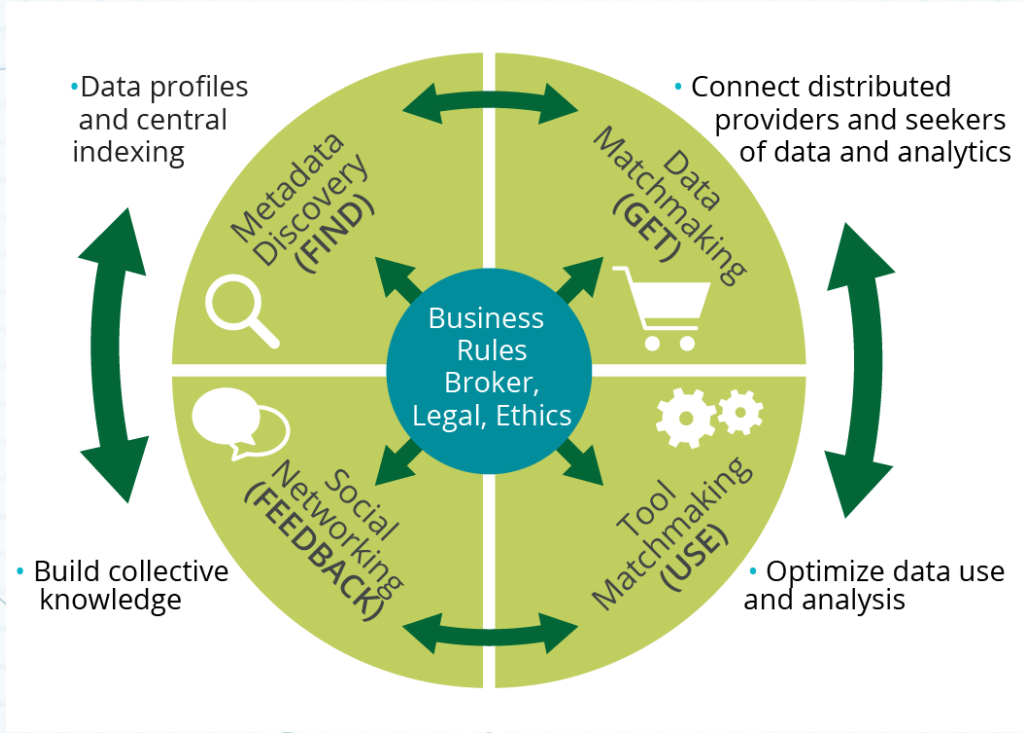
IMPACT:

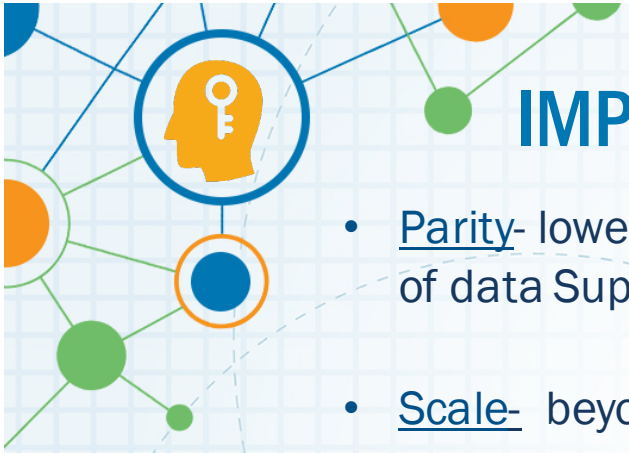
Enabling Information in the Cybersecurity Economy

IMPACT Motivation: The 'Open Secret' of Effective R&D

- Data are critical to R&D capabilities
  - Exactly 0% of R&D possible sans data
  - Cybersecurity needs real-world data to develop, test, evaluate knowledge & tech solutions to counter cyber threats
  - "Big Data" may grow on trees but still has to be picked, sorted, trucked
- Decision analytics are critical to HSE capabilities
  - Cybersecurity needs integrated, holistic understanding of risk environment
  - Gap between Data <-->Decisions: multi-dimensional, complex association and fusion, high-context presentation elements
- Data sharing + Analytics |= Easy
  - High value data = High legal risk + $$
  - Data rich vs. data poor
  - Expensive to abstract away low level knowledge- and labor- intensive tasks
  - Technologists optimize for Efficiency, Lawyers optimize for Certainty

IMPACT Approach





IMPACT ROI

- Parity- lower barrier to entry for data impoverished viz federation of data Supply & Demand (academic, industry, govt)
- Scale- beyond interpersonal relationships, ad hoc acquisitions
- Sustainable- Uniform, repeatable process
- Utility- responsible innovation over risk-aversion
- Trust
  - Vetted data, researchers, providers
  - Balance efficiency and certainty
  - Legal and ethical accountability

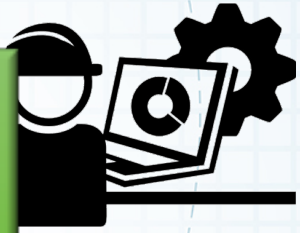
IMPACT Success Stories

* OCIA: Internet Capability Project (Internet Topology)

* NCCIC: Internal Ops (Internet Atlas)

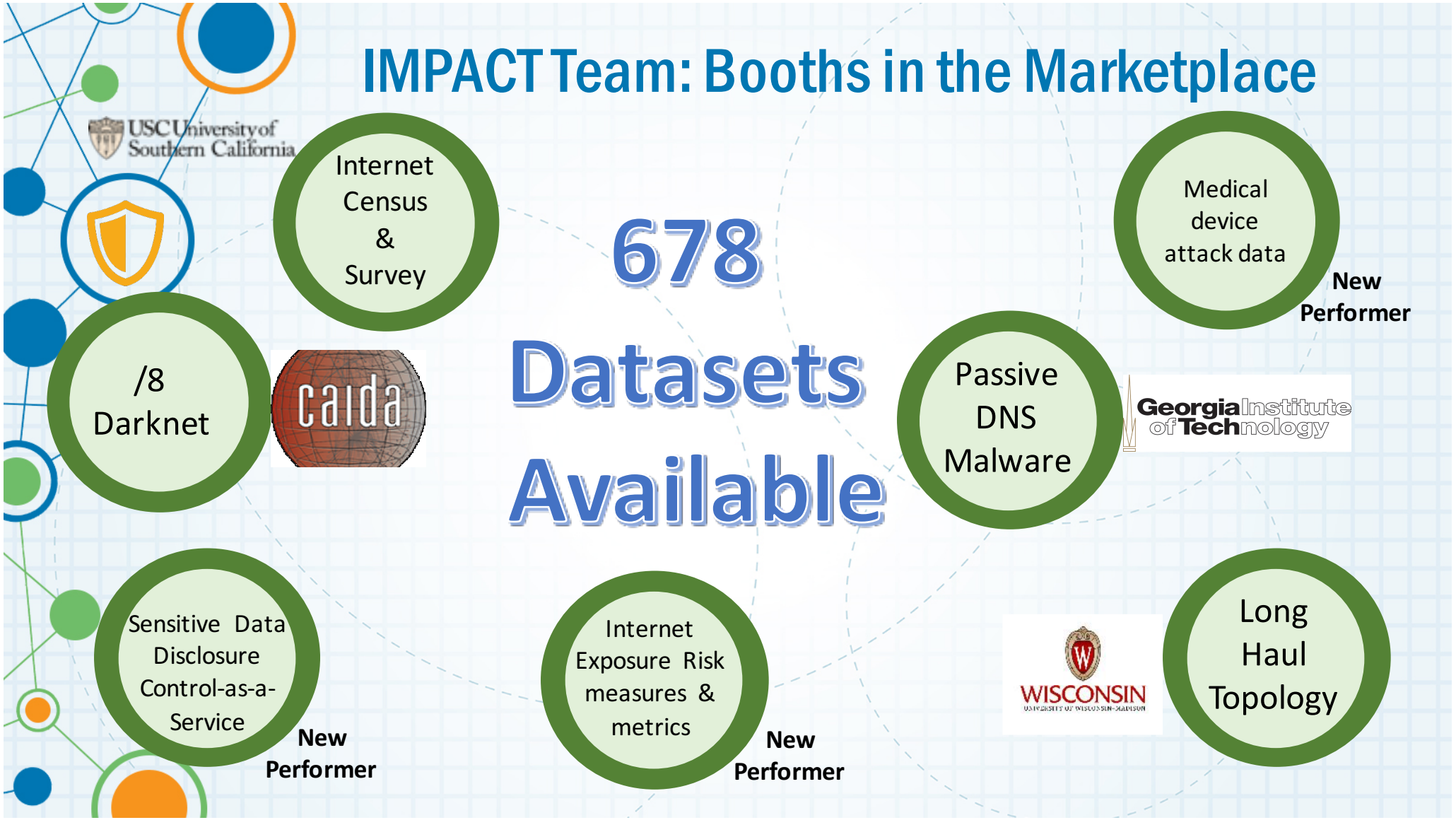
* Galois: 3DCoP ISP DDoS defense

* Comcast: understand scanning for vulnerable IoT devices

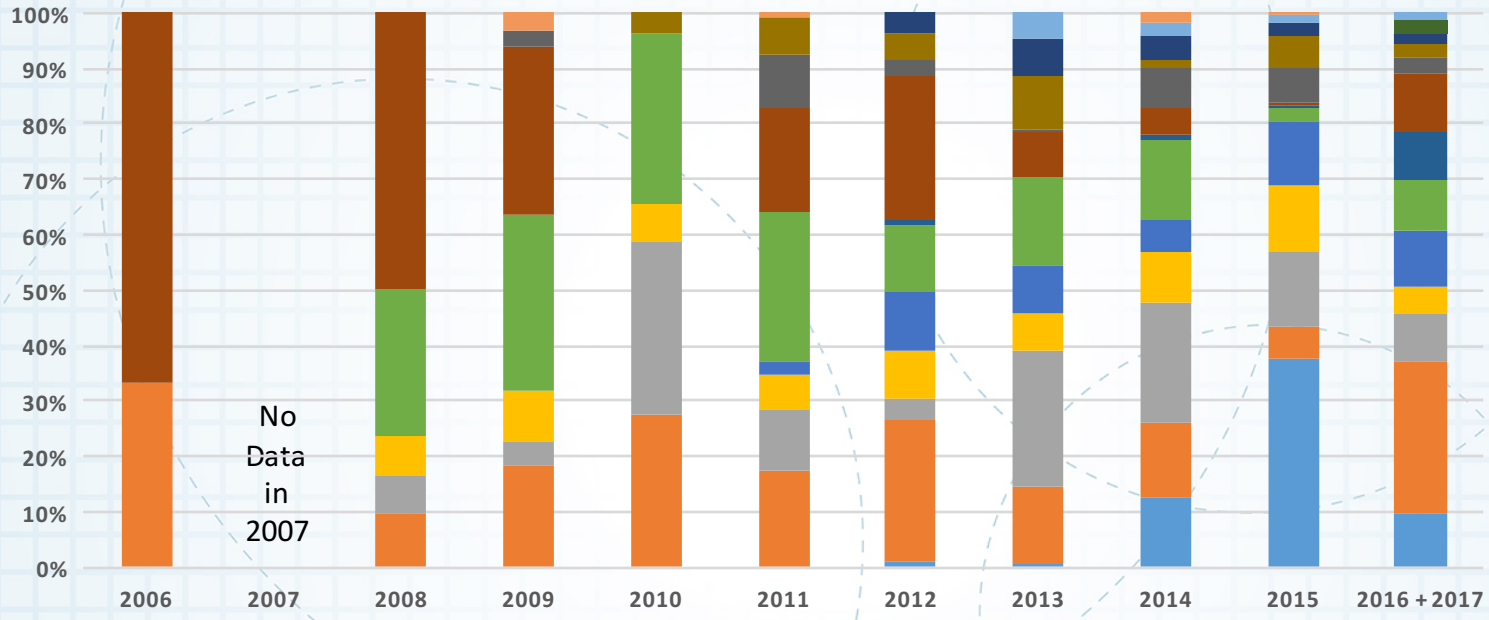


* Ph.D. Thesis, Conference paper, Zhan A characterization of cybersecurity posture from Network Telescope data

IMPACT Team: Booths in the Marketplace

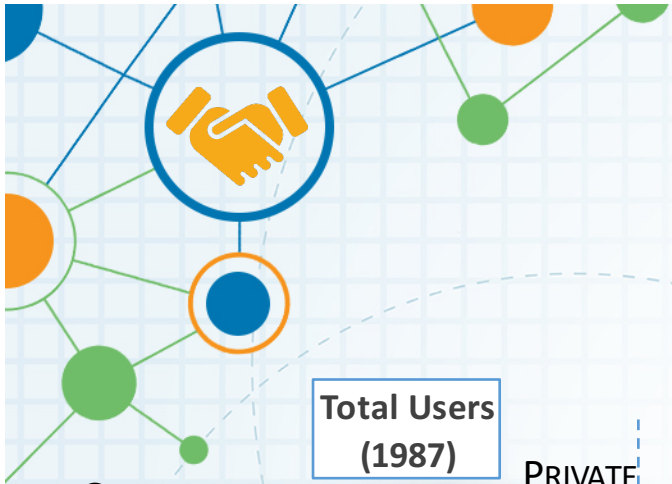


Data Trends

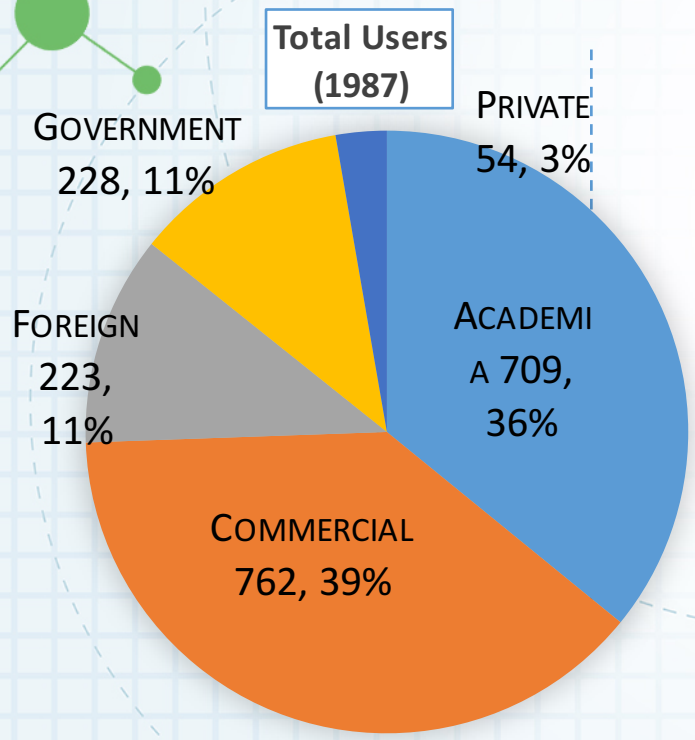


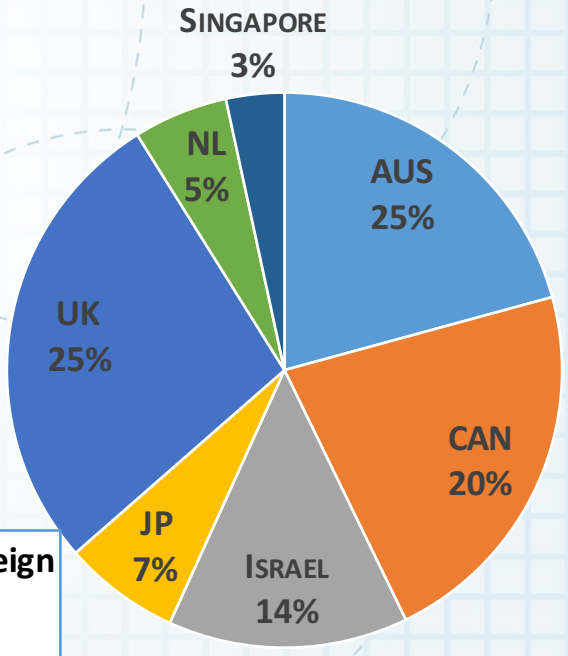
Source: DHS IMPACT program; SRI analysis, Apr '17

IMPACT Stats & Stakeholders



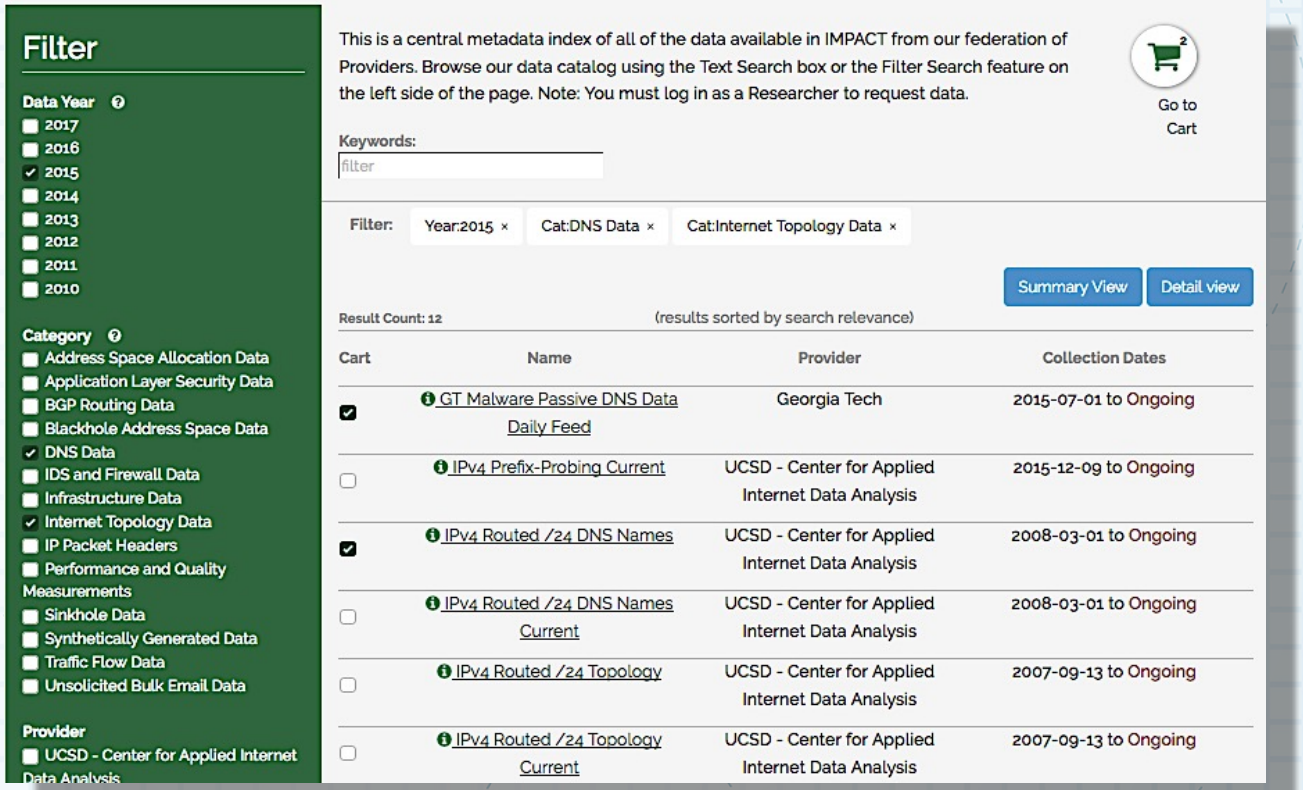
Research papers, journals, tech reports (>275 known)



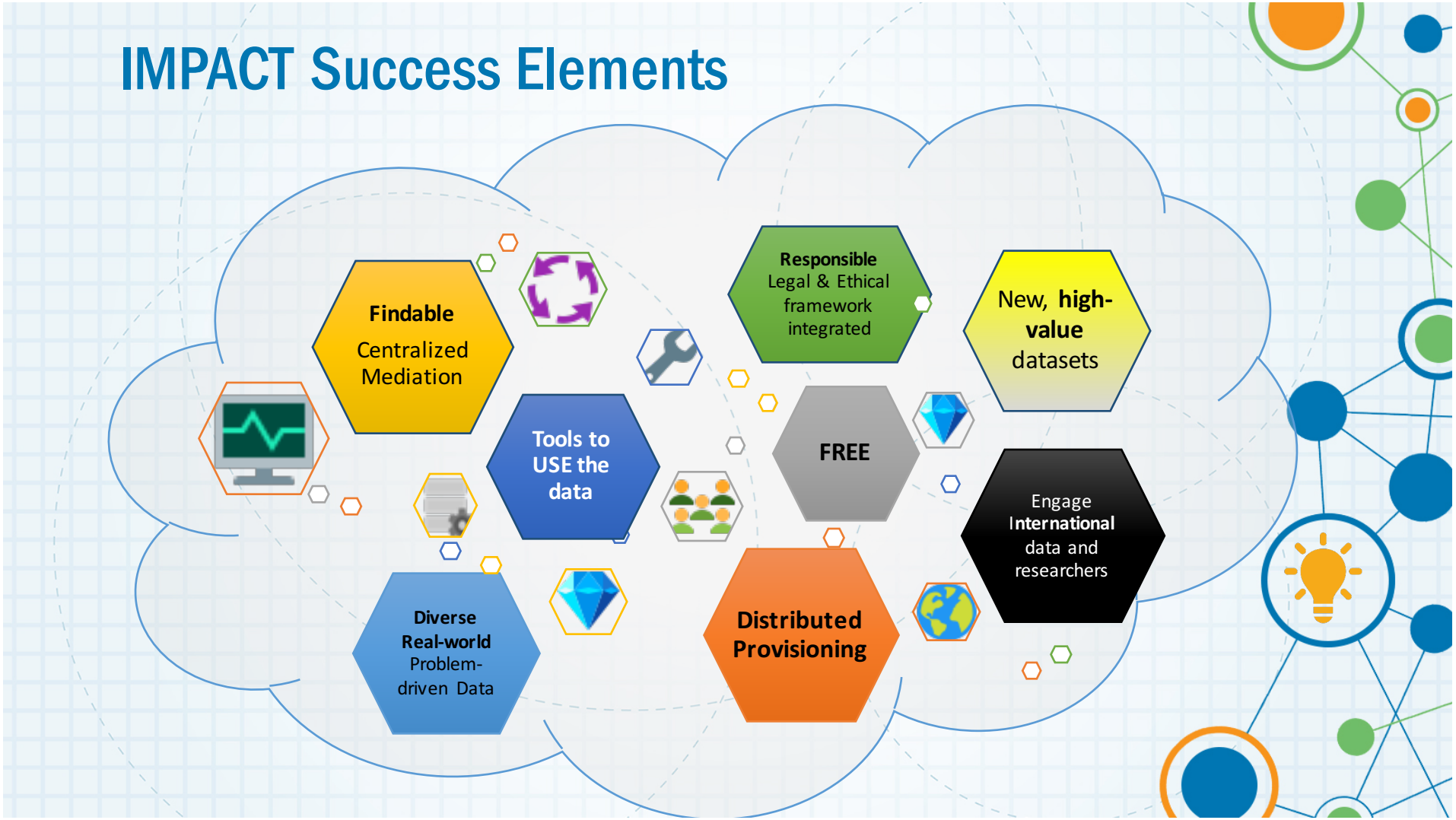


Approved Foreign Users (236 Total)

IMPACT Portal <www.ImpactCyberTrust.org>



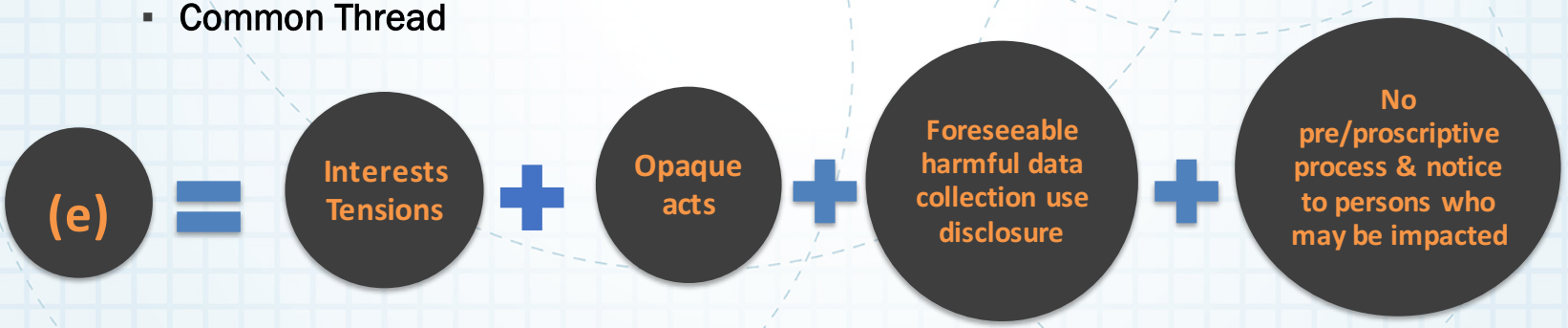
IMPACT Success Elements

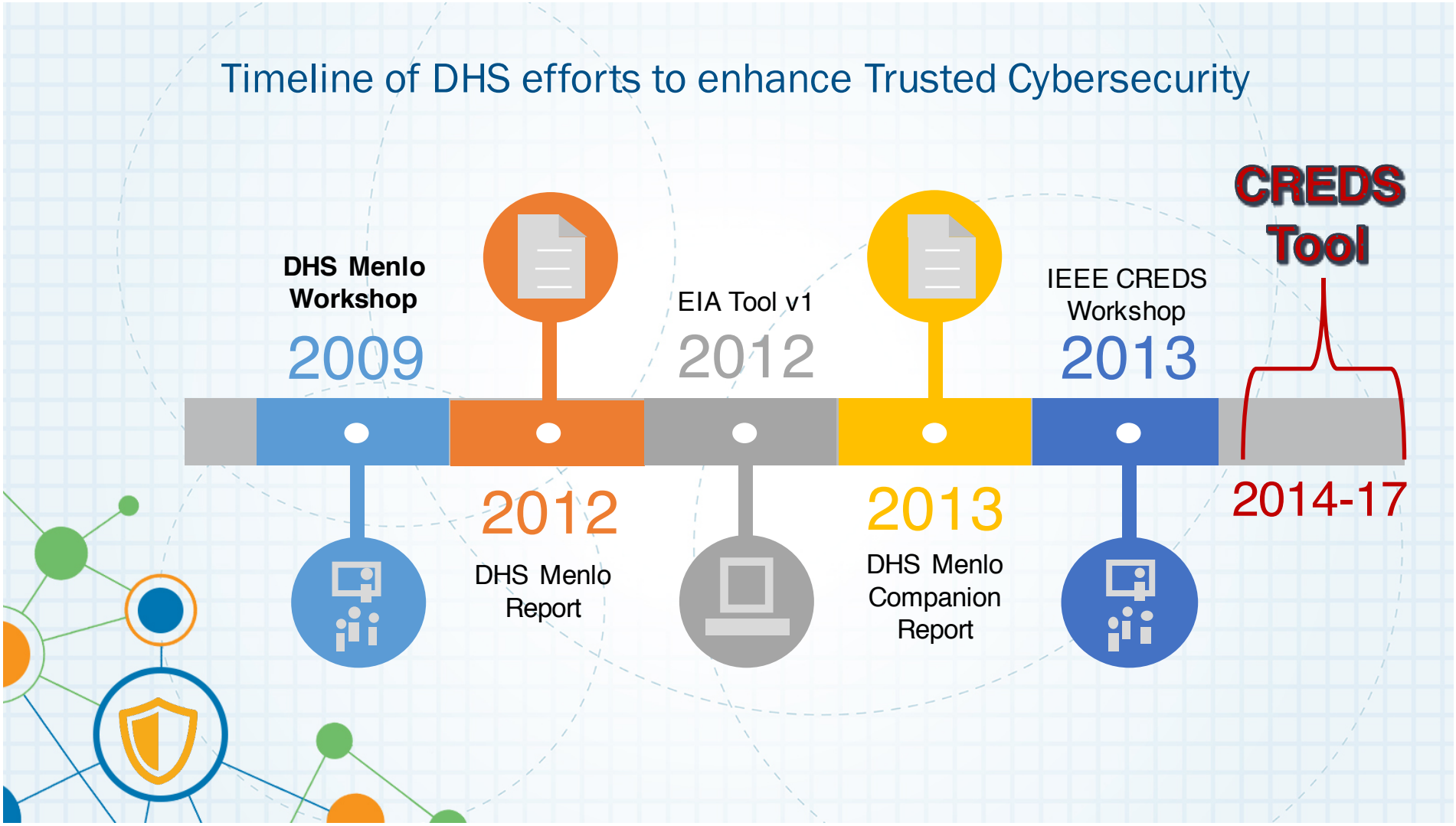


Enabling Trusted Cybersecurity with Ethics

- What characterizes anonymous observation, collection and use of sensitive data online w/o interacting with the data subject?
  - (a) Cyber espionage and surveillance by industry & nation-states
  - (b) Online advertising and data brokering by industry
  - (c) Targeted services and content by industry
  - (d) Security R&D (honeypots, botnet recon, reverse engineering, vuln disclosure)
  - (e) All of the above

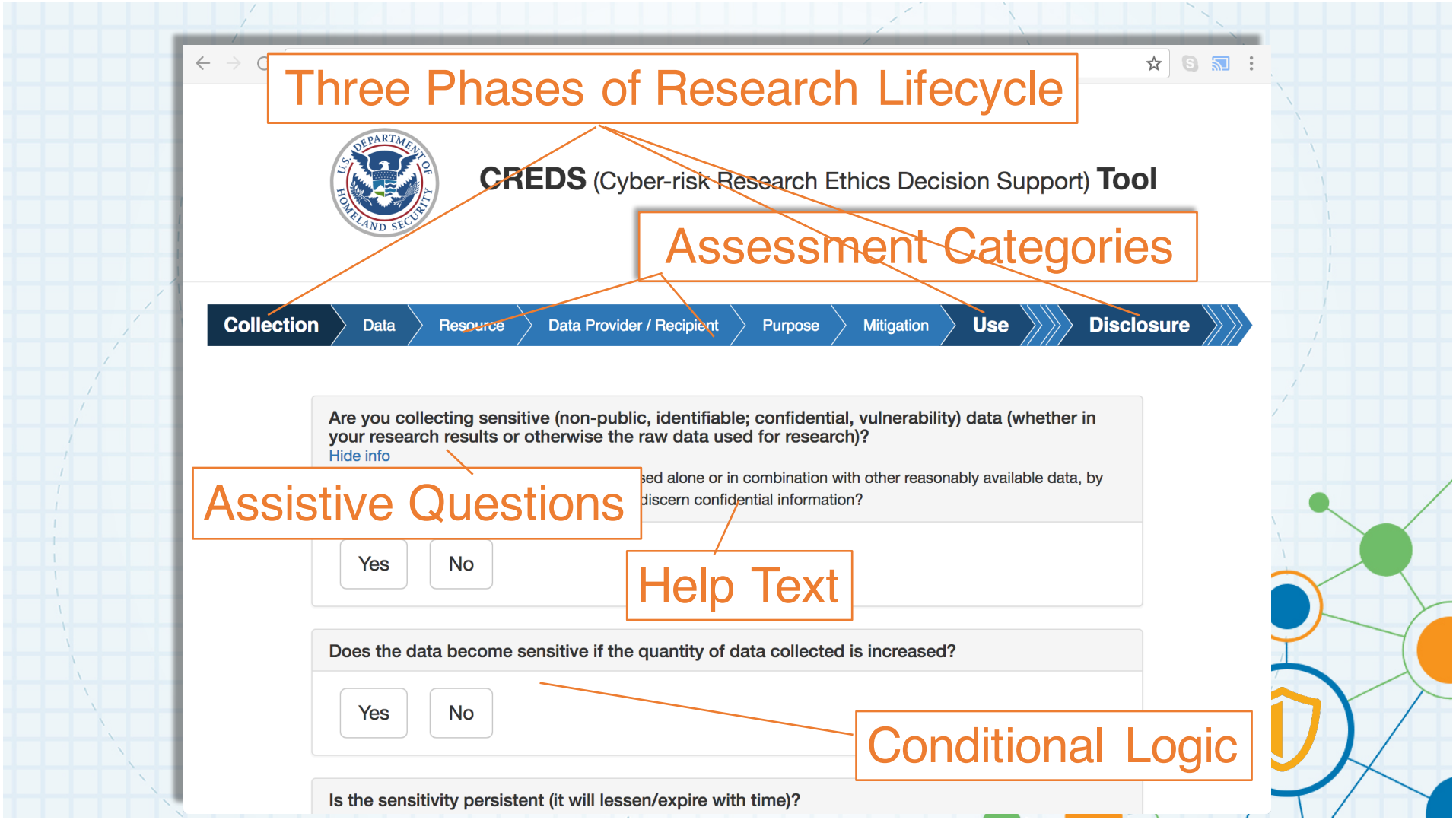
- Common Thread

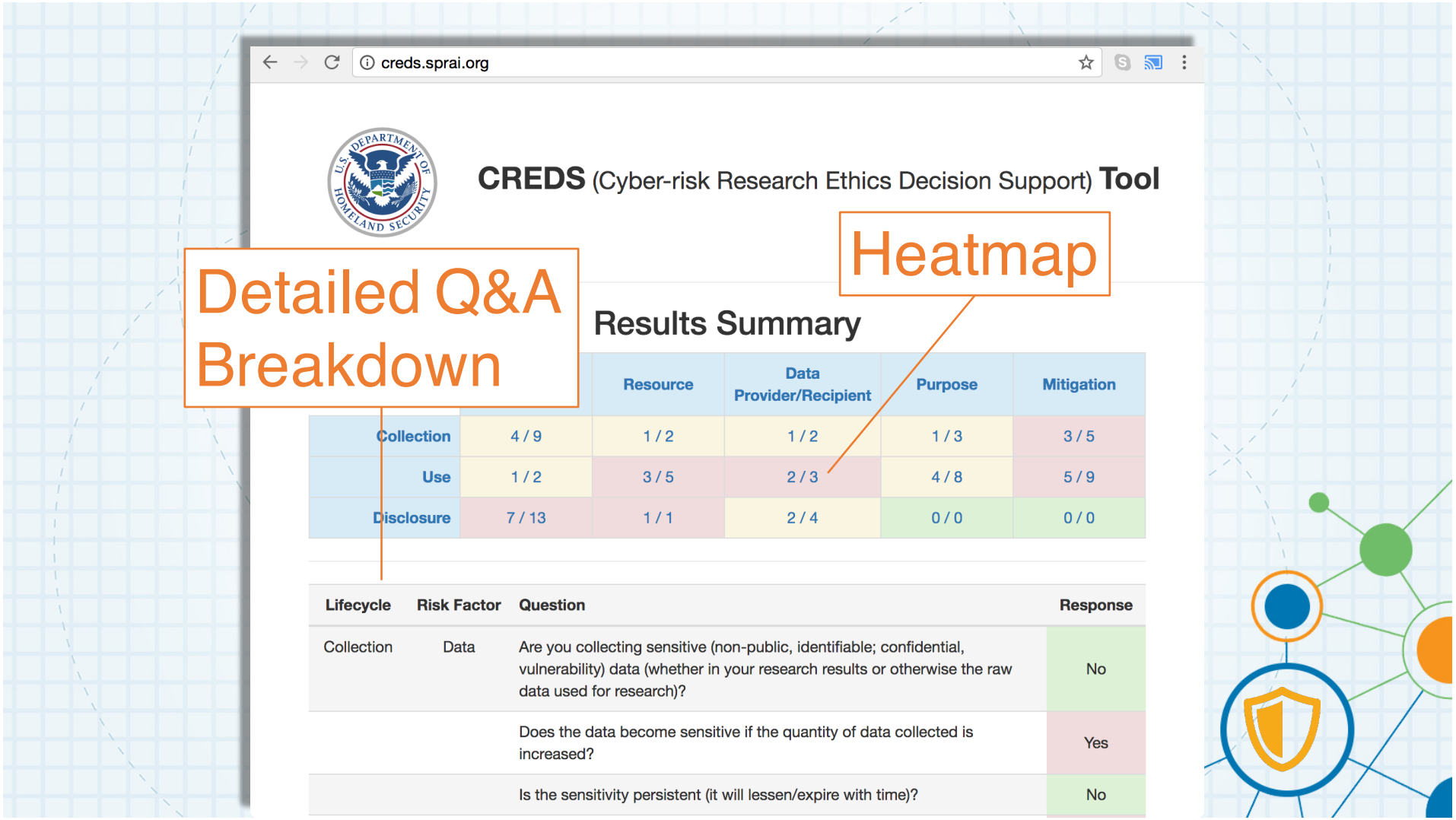


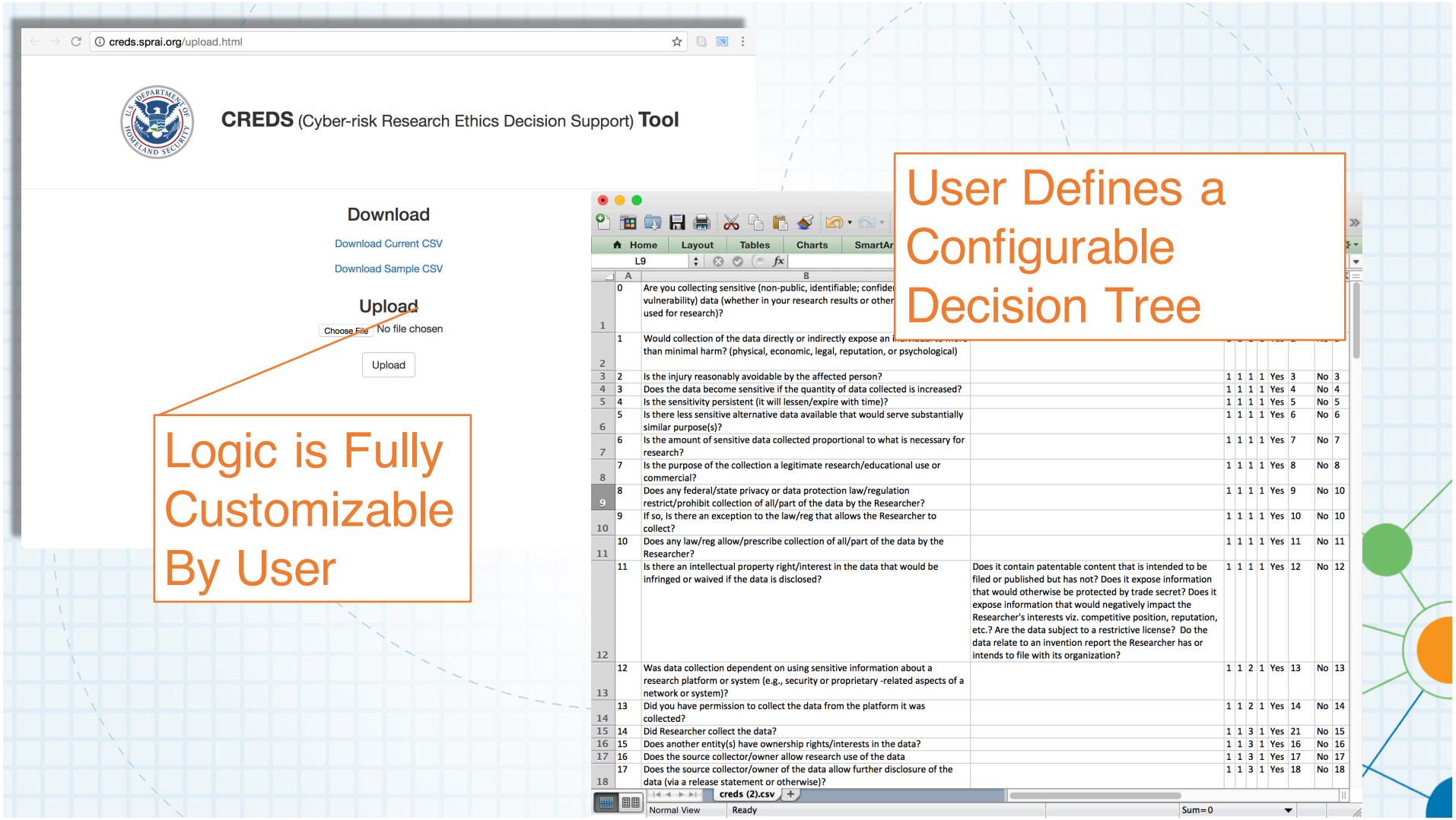

Timeline of DHS efforts to enhance Trusted Cybersecurity
DHS Menlo Workshop
2009
2012
DHS Menlo Report
EIA Tool v1
2012
2013
DHS Menlo Companion Report
IEEE CREDS Workshop
2013
CREDS Tool
2014-17

Cyber-risk Ethics Decision Support (CREDS) Tool

- **What:**
  - Operationalize a decision support framework + methodology into a tool that codifies ethical and legal principles in support of IMPACT
- **Who:** University of Illinois Urbana-Champaign, Michael Bailey
- **How:**
  - Derive responsibilities ethics & laws tenets, best practices
  - Transform EIA logic and methodology into an online decision support tool
  - Test and improve with real world, case-based scenarios and consultation with a range of stakeholders
- **Target:** Researchers, Product Developers, Overseers (ERB, PC, Funders)
- **Goals:**
  - identify and communicate ethical uncertainty and risk;
  - estimate potential ethical impacts of technology;
  - measure and improve human judgment and reasoning.


Three Phases of Research Lifecycle
CREDS (Cyber-risk Research Ethics Decision Support) Tool
Assessment Categories
Collection
Data
Resource
Data Provider / Recipient
Purpose
Mitigation
Use
Disclosure
Assistive Questions
Are you collecting sensitive (non-public, identifiable; confidential, vulnerability) data (whether in your research results or otherwise the raw data used for research)?
Hide info
Yes
No
Help Text
Does the data become sensitive if the quantity of data collected is increased?
Yes
No
Conditional Logic
Is the sensitivity persistent (it will lessen/expire with time)?


creds.sprai.org
CREDS (Cyber-risk Research Ethics Decision Support) Tool
Heatmap
Detailed Q&A Breakdown
Results Summary
Resource
Data Provider/Recipient
Purpose
Mitigation
Collection 4 / 9 1 / 2 1 / 2 1 / 3 3 / 5
Use 1 / 2 3 / 5 2 / 3 4 / 8 5 / 9
Disclosure 7 / 13 1 / 1 2 / 4 0 / 0 0 / 0
Lifecycle
Risk Factor
Question
Response
Collection
Data
Are you collecting sensitive (non-public, identifiable; confidential, vulnerability) data (whether in your research results or otherwise the raw data used for research)?
No
Does the data become sensitive if the quantity of data collected is increased?
Yes
Is the sensitivity persistent (it will lessen/expire with time)?
No



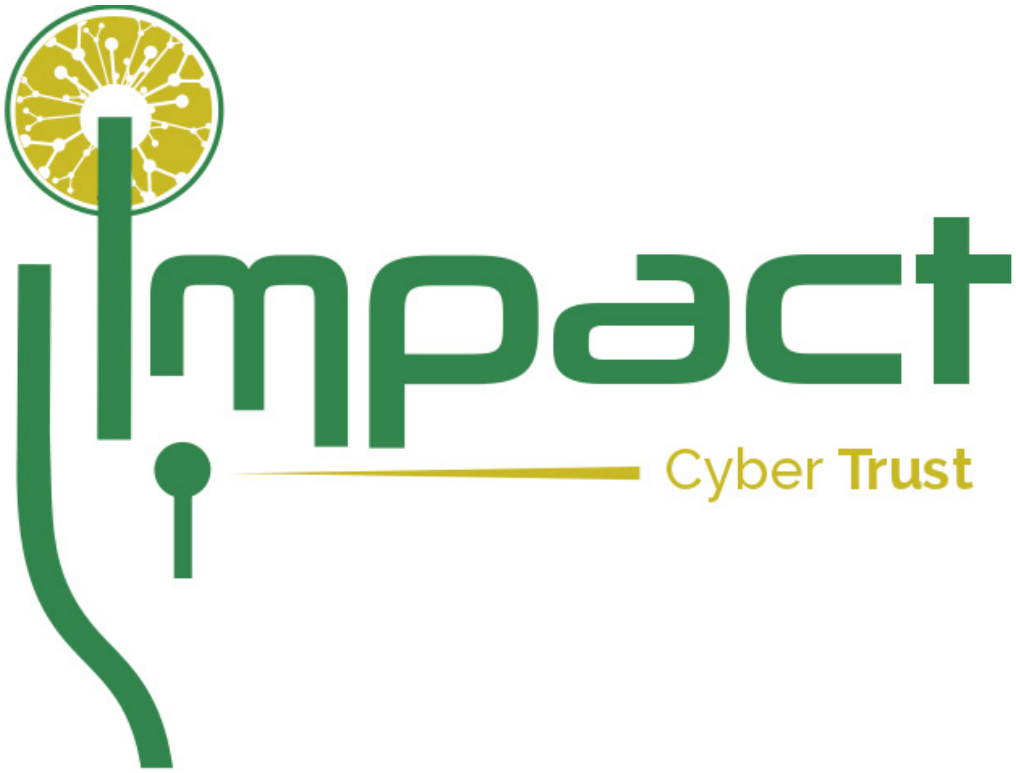
	A	B	C							
1	0	Are you collecting sensitive (non-public, identifiable; confide vulnerability) data (whether in your research results or other used for research)?								
2	1	Would collection of the data directly or indirectly expose an i than minimal harm? (physical, economic, legal, reputation, or psychological)								
3	2	Is the injury reasonably avoidable by the affected person?		1	1	1	1	Yes	3	No 3
4	3	Does the data become sensitive if the quantity of data collected is increased?		1	1	1	1	Yes	4	No 4
5	4	Is the sensitivity persistent (it will lessen/expire with time)?		1	1	1	1	Yes	5	No 5
6	5	Is there less sensitive alternative data available that would serve substantially similar purpose(s)?		1	1	1	1	Yes	6	No 6
7	6	Is the amount of sensitive data collected proportional to what is necessary for research?		1	1	1	1	Yes	7	No 7
8	7	Is the purpose of the collection a legitimate research/educational use or commercial?		1	1	1	1	Yes	8	No 8
9	8	Does any federal/state privacy or data protection law/regulation restrict/prohibit collection of all/part of the data by the Researcher?		1	1	1	1	Yes	9	No 10
10	9	If so, Is there an exception to the law/reg that allows the Researcher to collect?		1	1	1	1	Yes	10	No 10
11	10	Does any law/reg allow/prescribe collection of all/part of the data by the Researcher?		1	1	1	1	Yes	11	No 11
12	11	Is there an intellectual property right/interest in the data that would be infringed or waived if the data is disclosed?	Does it contain patentable content that is intended to be filed or published but has not? Does it expose information that would otherwise be protected by trade secret? Does it expose information that would negatively impact the Researcher's interests viz. competitive position, reputation, etc.? Are the data subject to a restrictive license? Do the data relate to an invention report the Researcher has or intends to file with its organization?	1	1	1	1	Yes	12	No 12
13	12	Was data collection dependent on using sensitive information about a research platform or system (e.g., security or proprietary -related aspects of a network or system)?		1	1	2	1	Yes	13	No 13
14	13	Did you have permission to collect the data from the platform it was collected?		1	1	2	1	Yes	14	No 14
15	14	Did Researcher collect the data?		1	1	3	1	Yes	21	No 15
16	15	Does another entity(s) have ownership rights/interests in the data?		1	1	3	1	Yes	16	No 16
17	16	Does the source collector/owner allow research use of the data		1	1	3	1	Yes	17	No 17
18	17	Does the source collector/owner of the data allow further disclosure of the data (via a release statement or otherwise)?		1	1	3	1	Yes	18	No 18

Help Us Help You to make an IMPACT

- Sign up for an account
- Try out the platform
- Send us feedback about where there's a data::problem gap and we'll help close it.

- Nothing to lose…except better R&D.

www.ImpactCyberTrust.org



Annual Digest

November 2017

The content of this report is CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0. Points of view, conclusions, and opinions contained within this document are those of the respective contributors, and do not represent the official position or policies of any organization.

1. Introduction

Data and analytic tools enable principled science. In a domain like cybersecurity, where threats are dynamic and risks interrelated, effective research and development (R&D) demands a data sharing mechanism that addresses the operational, legal and administrative challenges to responsible innovation. Since an objective notion of truth is defined by the system of one's limitations, what is our collective truth about the security, stability, and resilience of our systems and communications infrastructures where cyber threats are more certain and prolific than their purported defenses? Vigorously stirring Alt-Facts and Post-Truth into a morass of Groundhog-Day discussions of problems and solutions provides few answers about how to move toward effective solutions in our information economy. IMPACT, the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (https://www.dhs.gov/csd-impact; https://www.impactcybertrust.org), is a DHS Science & Technology program aimed at improving our collective truths about cyber security risk to ultimately enhance trust in our solution paths.

The open secret of cyber security is that empirical data and effective analytics are fundamental to high quality R&D; reliable security decisions do not emanate from low-quality or missing data. Consider the cover of the Economist (May 6, 2017), which claims "The World's Most Valuable Resource is No Longer Oil, But Data." If scarcity is an indicator of value, we need look no further than the critical short supply of data for cyber risk R&D; data and analytics foundational elements are essential to develop advanced knowledge and to accelerate design, production, and evaluation of next-generation cybersecurity solutions. However, the value of having research infrastructure that enables real-world, large-scale, timely, and longitudinal data collection, sharing, and analysis is severely underestimated. Too often, such capabilities are assumed to exist without deliberate resource affordances. The result has been a scarcity of data from both industry and the government that is available to the open academic research community for innovation, reuse, and plain old truth-setting. IMPACT is an R&D resource uniquely championed with the support of the DHS Cyber Security Division.

IMPACT enables, sustains, and mediates the provisioning of freely available cybersecurity data and analytics between providers and seekers within the global industrial, academic, and government cybersecurity communities. It lowers the barrier to entry for cybersecurity R&D by addressing the operational, legal, and administrative costs that otherwise impede scalable, sustainable, and responsible data-sharing that underpins valuable and innovative cybersecurity R&D. IMPACT reduces the time and cost to find, acquire, curate, and store data in a manner that is mindful of the associated legal and ethical risks.

This IMPACT Digest is a sampling of recent outputs and outcomes enabled by IMPACT to expand awareness about and engagement with our current and future efforts to inform policy and the analysis of cyber-risk and trust. This Digest is organized along the core dimensions of the IMPACT mission: 2. Data and Analytics, 3. Tools, 4. Trust, and 5. Impact.

2. Data

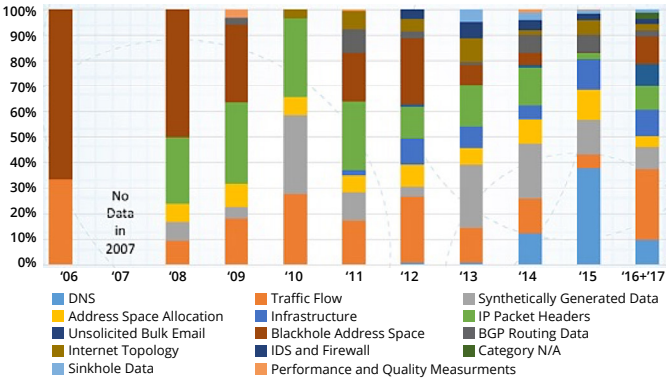


Figure 1: Data Trends; DHS IMPACT Program, SRI Analysis April 2017

2.1 Unsolicited Background Traffic Data

Needs Addressed: Researchers need data to study security and stability-related events including macroscopic connectivity disruptions, trends in malware propagation, and spoofed source-address denial-of-service attacks.

Approach Taken: CAIDA collected Blackhole address-space traffic data by monitoring routed but unused Internet Protocol (IP) address space that does not host any actual networked devices (e.g., hosts or routers). Systems that monitor unoccupied address space have a variety of names including darkspace, darknets, network telescopes, blackhole monitors, sinkholes, and background radiation monitors. Packets observed with such instrumentation can originate from a wide range of security-related events such as scanning for vulnerable targets, backscatter from spoofed denial-of-service attacks, automated spread of Internet worms or viruses, etc. Because unsolicited traffic may incidentally contain information about Internet hosts that are compromised or misconfigured, datasets in this category may be subject to specific disclosure control requirements.

Researchers at MERIT used a different approach; they analyzed traffic captured from MERIT's darknet to obtain

insight into Internet-wide scanning activities in 2016. Their analysis defined a source IP address appearing in the darknet to be a scanner if that address contacted 25 unique destination IPs over a five-minute interval at the same destination port and protocol. (While the definition of scanning can often depend on context, this is the Bro default definition for scanners). Due to the dataset volume, MERIT adopted a sampling approach and analyzed only the first day of each month between Jan. 2016 and Oct. 2016.

Table 1: Top-20 Services Scanned

Port #/Protocol # (%)	Description (Service Name)
23/TCP (60.18%)	Telnet
53413/UDP (9.03%)	Vulnerability scan on Netis routers
No port /ICMP Ping (2.32%)	ICMP Ping
80/TCP (2.29%)	HTTP
3389/TCP (1.09%)	Microsoft RDP
2323/TCP (0.91%)	Mirai (Botnets scanning for IoT devices)
445/TCP (0.80%)	SMB-IP (Microsoft-DS Active Directory)
22/TCP (0.70%)	SSH
2222/6 (0.39%)	Ethernet/IP or DirectAdmin Remote Access
81/TCP (0.33%)	TorPark Onion Routing
8000/TCP (0.29%)	Radio streams such as iTunes Radio, DynamoDB Local
91/TCP (0.27%)	SG Security scan
21/TCP (0.20%)	FTP
443/TCP (0.16)	HTTPS
8123/TCP (0.14%)	Unknown (can be used for Web proxy)
8080/TCP (0.12%)	FilePhile Master/Relay over TCP
53/UDP (0.12%)	DNS
1080/TCP (0.11%)	SOCKS proxy
8888/TCP (0.10%)	HyperVM/ Freenet/ MAMP over TCP
3128/TCP (0.10%)	Squid Caching web proxy

The top twenty services scanned are tabulated in Table 1. Notice the volume of Telnet scanning.

The increase in Telnet scanning can be attributed to malicious activities at the time scanning for vulnerable IoT (Internet-of-Things) devices. Careful examination of the Mirai botnet source code[1] revealed that it was scanning for TCP port 23 (Telnet) and TCP 2323. The increase in darknet activities regarding these two ports is evident in Figure 2 and in the temporal summary contained in Table 2.

Table 2: Temporal Comparison of all Darknet Activities (not just scanning)

2004 [2]	2010 [3]	2014 [4]	2016 (this study)
HTTP (80)	SMB-IP (445)	SMB-IP (445)	Telnet (23)
NetBIOS (135)	NetBIOS (139)	ICMP Ping	UDP 53413 (Netis)
NetBIOS (139)	eMule (4662)	SSH (20)	HTTP (80)
DameWare (6129)	HTTP (80)	HTTP (80)	ICMP Ping
MyDoom (3127)	NetBIOS (135)	RDP (3389)	SSH (20)

Table 3: Top Countries Associated with Scanning

United States (21.68%)	Taiwan (3.63%)
China (10.16%)	Netherlands (1.91%)
Brazil (6.98%)	Turkey (1.77%)
Vietnam (5.25%)	Romania (0.57%)
South Korea (4.47%)	Russia (0.43%)

Table 4: The Top-20 Scanners Originate from These Providers (ASes)

Akamai	QUASINetworks	SingleHop, Inc
CariNet, Inc.	Steadfast	WELLPOWER-TW

Table 3 identifies countries associated with scanning activities (information obtained from the geolite2 MaxMind database). In 2016, Telnet ranks at the top, and activity for TCP 2323 made the top 10 (Table 2). Table 4 lists the providers associated with autonomous scanning systems (information obtained from "Whois" data).

Resulting Benefits: Darknet analysis provides valuable insights to network operators regarding trends in global scanning activities (associated with active cyber-threats), denial of service attacks, network outages, and other threats. The results of the analyses enable further activities for situational awareness and attribution, which enhance the security posture of both public and private organizations.

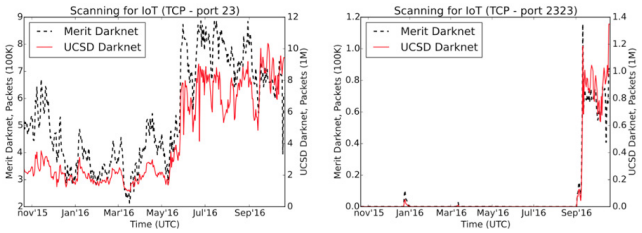


Figure 2. Darknet traffic to TCP ports 23 and 2323, as recorded by two large darknets, namely the Merit and the UCSD darknets. Data source: Merit and UCSD darknets. Graph data obtained from CAIDA's Charthouse.

2.2 Internet Addresses and Topology Data

Exchanging information over the Internet requires computers on both ends with IP addresses. Almost all servers on the Internet have their own addresses, although client computers sometimes share an address using network address translation (NAT). All commercial services today require addresses using IP version 4 (IPv4). IPv4 addresses are a limited resource; there are only 4 billion of them. In 2011, IANA (the organization managing global addresses) allocated the last open IPv4 addresses, and by 2015, all regional registries except one have exhausted their pools.

[1] Mirai was responsible for some of the largest DDoS attacks ever recorded, including the attacks against KrebsonSecurity and Dyn.

Marketplaces that trade IPv4 addresses are raising questions about imposing technical restrictions on what can be traded. Completely unfettered trading could fragment addressing and increase costs. IPv6 exists, and its use is increasing, but concerns include unanswered questions regarding how the cost of supporting and deploying IPv6 compares to more careful management of IPv4. IPv4 addresses affect factors besides cost, for example, address usage and density affect worms and port scanning. These issues open other unanswered questions:

- *How effective is scanning by botnets like Mirai?*
- *Can we estimate botnet size and capability by the probes we see?*
- *What about millions of new mobile phones today?*
- *What about millions more IoT devices tomorrow?*

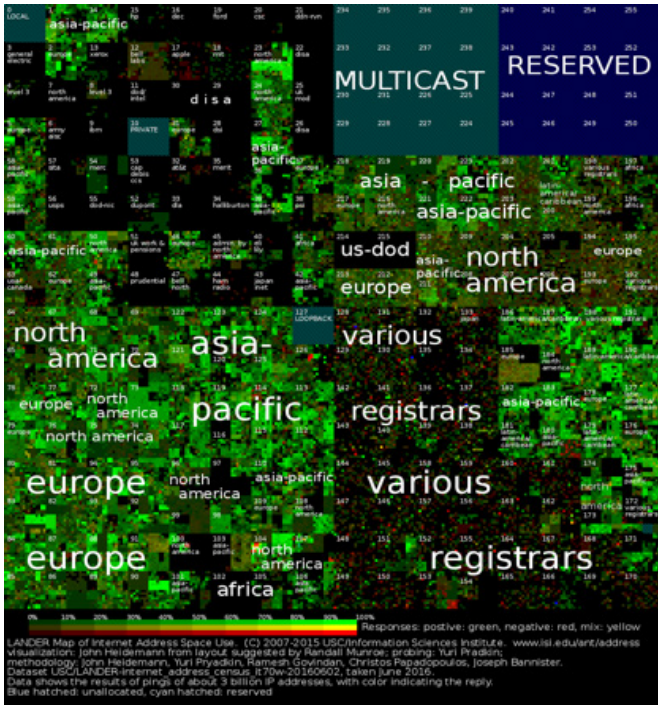


Figure 3.A map of the IPv4 address space. Dataset: https://www.impactcybertrust.org/dataset_view?idDataset=621.

Clearly, effective planning of this important global resource requires high-quality, objective, and public insight into current allocation and activity trends.

Approach Taken: In 2003, researchers at ISI started collecting data about the Internet (IPv4) address space, and regularly probe all addresses in the allocated Internet address space, more than 3 billion. Figure 3 shows their most recent map of Internet addresses. Internet addresses are allocated in blocks of adjacent addresses, and these blocks turn into squares of different size in this map. Addresses usually are numbers like 192.0.2.0; there are about 4 billion addresses, from 0.0.0.0 to 255.255.255.255. Addresses follow the Hilbert fractal, so numerically close addresses are physically close on the map.

In the map, brighter areas indicate more replies, darker, fewer, with color indicating positive (green) or negative replies (red), and black showing areas that do not reply.

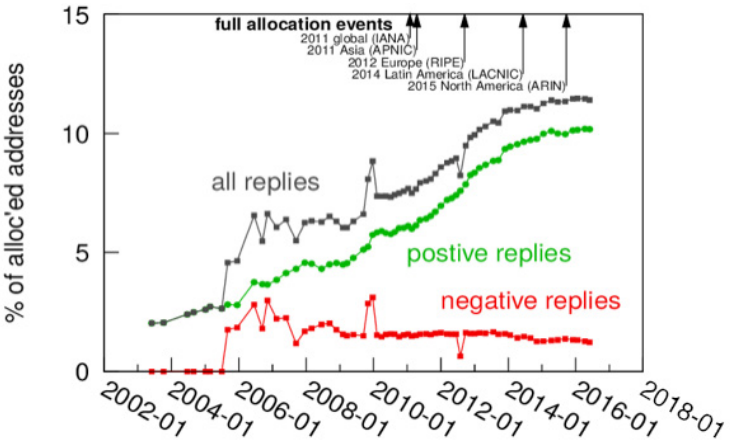


Figure 4. Address usage over time as seen by ISI's probing.

The cyan hatched areas are not studied, typically because they are reserved or private. With a consistent, careful methodology, our data provides a long-term longitudinal view of changes in IPv4 address use.

Our data (Figure 4) gives an estimate of the instantaneous size of the Internet. Scaling replies by our correction factor of 1.6 to 1.9. Our best estimate is that about 676 to 803 million addresses are active at any instant. We used the correction factor from "Census and Survey of the Visible Internet," by Heidemann et al., ACM IMC 2008.

No census of billions of addresses will be perfect, but we correct for addresses we cannot directly measure with statistical estimation. Our peer-reviewed study describes sources of under-counting: A few percent of probes and replies are lost due to congestion. Some addresses, such as those behind firewalls, chose not to receive or reply to our requests. Other computers use private addresses.

Table 5: Instantaneous size of the Internet (est.) Basic data from https://www.impactcybertrust.org/dataset_view?idDataset=621 in June 2016.

What	Addresses
IPv4 addresses:	4,294,967,296 (100% all)
..unprobed:	588,972,032 (13% all)
....special (multicast, etc.):	587,202,560 (13% all)
....unallocated:	1,769,472 (0% all)
..probed:	3,705,995,264 (86% all)
....replies:	422,662,606 (9% all)
......positive:	377,166,966 (8% all)
......negative:	45,488,476 (1% all)
....non-replies:	3,283,332,658 (76% all)

After we correct for under-counting, our measurements represent the most accurate current snapshot estimating IPv4 address usage.

Dynamic addresses are another source of change in the number of addresses that respond. Our census gives a virtual snapshot of how many addresses are active at any instant, on average. In some countries and regions, however, this number varies up or down by 20% over the course of a day. (For more information, see "When the Internet Sleeps: Correlating Diurnal Networks with External Factors," by Quan et al., ACM IMC 2014).

While the previous numbers capture instantaneous snapshots, far more addresses are used when considering an entire day or multiple days. CAIDA added data to ISI's IPv4 2013 Census data set that elucidates the IPv4 address-space usage. CAIDA cross-correlated results of analysis of passive and active measurements to taxonomize 24 address blocks as IETF-reserved, used, routed-unused, unrouted-assigned, and available. These datasets include both raw and curated forms of topology data gathered from across the global Internet. Typically, this data is obtained by deploying traceroute-like probes from monitoring points around the network. Raw IP topology data can include IP addresses on machines that a packet traverses along the forward path to a target destination, allowing heuristic-based inference of router-level and AS-level topologies. Some topology datasets are already curated into router-level or autonomous system-level topologies for ease of researcher use. Datasets in this category support modeling and simulation of malware outbreaks, spread, distribution, containment, and countermeasures; macroscopic vulnerability assessments; longitudinal analysis; and modeling of Internet evolution.

Resulting Benefits: Our data provides a unique, long-term view of Internet use with several applications:

- *Work with census and aggregated datasets helps inform address usage. Our long-term study shows that address usage is growing, but overall use is still low (less than 10%). We could improve the management of the IPv4 address space, but better management comes with a cost.*
- *Censuses help improve studies of Internet topology. We use censuses to build hitlists, a comprehensive list of addresses to probe for good coverage of the Internet. Multiple research groups (71 different researchers) use these hitlists to conduct new research studies about Internet topology.*
- *Internet census-taking has inspired new approaches. Based on census findings, USC/ISS have developed techniques to estimate network outages around the globe, and have shown how diurnal shifts in Internet address usage vary from country to country, reflecting the maturity and policies in each country's part of the Internet. CAIDA has aggregated additional datasets, showing that passive and active measurement approaches can produce a fuller picture of long term Internet address use.*
- *The USC/ISI censuses have inspired faster versions. Multiple research groups (Michigan, MassScan by Graham et al.) have focused on conducting censuses as fast as possible. While our censuses are intentionally paced at a slow rate to minimize concerns, fast probing has a place in security studies.*
- *Data from the USC/ISI censuses have been used by others. Since 2006, 71 different researchers have acquired copies of some or all our censuses and related datasets, over 53TB of data. This data has been used in dozens of papers and follow-on studies.*

2.3 Malware Data

Needs Addressed: Malicious software is a centerpiece among current threats on the Internet. It is used to create botnets that, for example, generate unsolicited email, conduct DDoS attacks, and steal sensitive financial information and intellectual property, fueling the intentions of criminals and nation-states alike. Efforts to understand and defend against malicious software are therefore critical to cyber-security research and applied defense. These efforts, which include cyber-threat discovery, compromise detection, and asset remediation, require the ability to observe and study current malware behavior.

Major information security organizations collect over 100,000 new malware samples each day from spam traps, web crawlers, user submissions, and malware exchanges. Each sample holds actionable network information that, once derived via a dynamic analysis sandbox, has both research and operational utility. However, the sensitive nature of malware, the resources required to process the substantial volume of new samples that appear each day, and the commoditization of anti-analysis techniques leave many researchers and practitioners with an ongoing, unmet need for this data.

Figure 5. GTISC 2016 Malware DNS Wordle.

Approach Taken: Through DHS IMPACT, the Georgia Tech Information Security Center (GTISC) leverages its extensive malware collection and analysis experience to facilitate the availability of real-world, large-scale malware network datasets. In providing such data to approved requestors vetted by the IMPACT Coordinating

logs from NTP servers. Using NTP-derived data for studying latency is compelling due to NTP's pervasive use in the Internet and its inherent focus on accurate

Approach Taken: Fortunately, one approach appeases privacy advocates while foiling malicious plans to capture Internet traffic: we can monitor the entire Internet at the routing level. We can lay down a coarse map of the major streets and highways on the Internet without actually looking at the traffic on them, every few minutes. We can quickly detect changes in the because the Internet uses a routing protocol called Border Gateway Protocol (BGP) that tells every router how to direct traffic anywhere on the Internet. BGP is a verbose protocol, telling each router the next hop to send traffic and the series of networks that traffic will follow to get to that destination. This allows routers to implement policies and make local decisions about which paths they prefer.

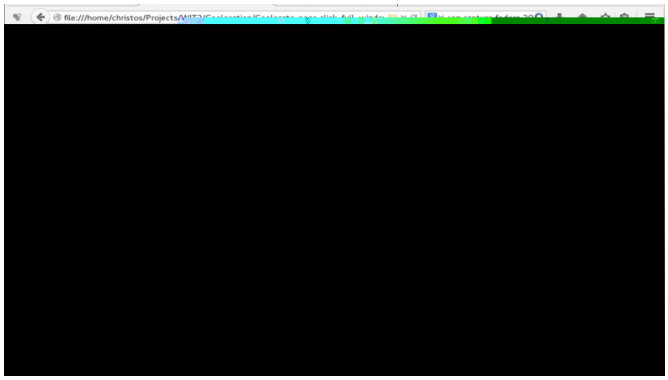


Figure 7: The BGPmon eyes: IPv4 (orange) and IPv6 (blue)

Resulting Benefits: The BGPmon project (http://www.bgpmon.io) is one of the many projects contributing data to IMPACT. By leveraging data collection started by RouteViews about 15 years ago, BGPmon offers a continuous live feed of BGP information from about 400 points around the world. BGPmon is a quiet listener, it never talks back, so it is safe to deploy. The information it collects is time-stamped, geolocated, and collected in a fast database at Colorado State University, where it is made available in many convenient formats to the entire research community.

An important operational benefit of BGPmon, is its ability to detect route hijacks. A route hijack occurs when a router, accidentally or maliciously, announces a path to an unauthorized destination network. BGP cannot detect lying routers; the only defenses are the instincts, skills, and expertise of network operators. Sometimes these qualities fail, especially in networks with less experienced operators or when a trusted network makes a mistake. Such a mistake results in the diversion of traffic, which can cause a routing outage or a man-in-the-middle attack. Only the operators can remedy this situation; however, BGPmon can detect such incidents and notify the operators immediately so they can respond.

BGPmon offers more than just a stream of BGP messages to help detect hijacks. The BGPmon team has geolocated all the monitors, so when a hijack is reported, operators will know the prefix, time, and duration of the hijack and the geographical regions that were affected. The BGPmon team provides an AS geolocation service, so users can issue queries about the prefixes an AS advertises at a particular time, and where those prefixes geolocate. While geolocating Internet hosts is hardly an exact science, BGPmon provides a good approximation.

3.2 Mapping Internet Physical Infrastructure—Internet Atlas and BigFoot

Needs Addressed: Mapping and understanding the physical underpinnings of our Internet infrastructure helps identify shared risk and defend against adversarial and natural threats to our communications infrastructure. A large body of economic research has shown a strong correlation between broadband connectivity and economic productivity. These findings motivate government agencies, such as the FCC in the US, to provide incentives to Internet Services Providers to deploy broadband infrastructure in unserved or underserved areas.

Approach Taken: We developed a framework for automating the identification of target areas for network infrastructure deployment. Our approach considers infrastructure availability, user demographics, and deployment costs. We used multi-objective optimization to identify geographic areas that have the highest concentrations of unserved and underserved users and that can be upgraded at the lowest cost. To demonstrate the efficacy of our framework, we considered physical infrastructure and demographic data from the US and two deployment cost models. Our results identified a list of counties that would be attractive targets for broadband deployment from cost and impact perspectives. We validated our findings by comparing the results with the FCC's Connect America report. We believe that there are a variety of broader applications of our framework, which is now incorporated into the Internet Atlas portal.

The Internet Atlas project is focused on building and maintaining a repository of geocoded maps of Internet physical infrastructure. We define physical infrastructure as nodes (PoPs, co-location centers, IXPs, etc.) and links (fiber optic cables) that carry Internet traffic. The repository currently includes maps of over 1,100 networks from around the world, which have been carefully audited over the past year. The process of adding new maps to the repository is on-going. The repository is complemented by a GIS-based web portal, which enables data to be visualized and analyzed. The portal also enables a wide variety of Internet and related data to be imported and visualized.

Regarding ethics, a satellite aspect of IMPACT has focused on responsible innovation by setting ICT research ethics standards. IMPACT supports cyber-risk R&D by enabling data and tool sharing, and addresses the socio-technical layer that can impede R&D. IMPACT has been a leader in community-informed ethics and sensitive data-disclosure guidance that address the principles, controls, and responsible implementation of solutions to issues that impede cyber security research. The Menlo Report, Ethical Principles Guiding Information and Communication Technology Research and its Companion Report were the flagship ethics outputs. They were inspired to pre-empt some of the socio-technical issues coming down the pike, and to embrace a principals-in-context approach by applying the framework from the Belmont Report to modern information and communication technology research.

Current efforts are focused on Cyber-risk Ethics Decision Support (CREDS). The CREDS Tool is an applied research and development project intended to operationalize a decision support methodology, conceptual framework, and an interactive online tool to identify, reason, and

Figure 13. IMPACT Footprint

manage ethical and legal issues related to cyber-based research (e.g., network and system security). It operationalizes the Ethical Impact Assessment (EIA) framework that was coined and commenced under the IMPACT ethics project. The objectives of the tool are to facilitate research that minimizes potential harm while enabling innovation, and to advance the collective dialogue between and among researchers, oversight entities and policymakers about research ethics principles and practices.

The functional goals include estimating and communicating ethical risk; identifying potential impacts of technology; and measuring and improving judgment and reasoning. The methodology involves deriving principles and practices from established law, ethics and best practices, and then using that output to drive the underlying logic of the tool. The CREDS tool is intended to be a resource for the entire community to engage in repeatable and transparent decision-making to prevent diminished public trust and reputational blowback caused by association with undifferentiated comparisons to public or private surveillance and cyber opportunism. CREDS is currently in prototype development and is scheduled for release to the community in Fall 2017.

Stakeholder Engagement. In February 2016, CAIDA organized and hosted a two-day BGP Hackathon event for teams of researchers and students with interests in the development of tools to measure, monitor, and model the routing infrastructure of the Internet[7]. Sponsored by industry, professional organizations, and government agencies, the teams competed to develop practical solutions to live BGP measurements and monitoring challenges. All fifteen teams made significant progress in critical BGP areas including investigating anycast routing, automating detection of BGP anomalies (e.g., hijacking events), improving the BGPStream framework by adding customized filters, and developing informative visualizations of live BGP data.

5. Conclusion

Data are critical to R&D capabilities, and despite interpretation challenges with statistics, we can confidently conclude that impactful R&D is impossible without empirical data. Cybersecurity needs real-world data to develop, test, and evaluate knowledge and technology solutions to counter cyber threats. Big data may grow on trees, but it must be picked, sorted, and trucked. IMPACT addresses these needs and moves forward with the IMPACT vision. Decision analytics are critical to HSE capabilities: cybersecurity needs an integrated, holistic understanding of the risk environment for strategic interventions. IMPACT is addressing the noticeable gap between data and decisions by providing the multi-dimensional data, complex associations and fusion, and high-context presentation elements that will close that gap. Finally, data sharing and analytics are not easy; high value data often results in high legal risk and/or cost. IMPACT addresses the technologists' need to optimize for efficiency and the lawyers' need to optimize for certainty. Help us help you make an IMPACT. https://ImpactCyberTrust.org

[7] A report is available at http://caida.org/publications/papers/2016/bgp_hackathon_2016_report.