Realistic Cyber Ranges & Enabling Machine Learning Within Cybersecurity

Tom Podnar
Dustin Updyke

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Part I (Tom)

- Who we are and what we do
- The state of cyber ranges today
- Data opportunities

SEI > CERT > CWD > Realistic Scenario Simulation Team

- Architect & deliver realistic cyber warfare exercises
- Persistent production spec high-fidelity environment (aka: realism)
- Security hardened and tuned
- Multiple engagements lasting days to 7+ months
- Participants are worldwide DoD operators

Exercise as you Fight

McAfee®

Microsoft

BLUE COAT®

Microsoft Active Directory

splunk>

CISCO™

ENDGAME.

Linux™

ArcSight

Nessus

paloalto NETWORKS®

Windows 10

Exercise as you Fight, continued

dhs.gov/topic/cybersecurity

Official website of the Department of Homeland Security

Contact Us | Quick Links

Homeland Security

Topics | News | In Focus | How Do I? | Get Involved | About DHS

Enter Search Term | On DHS

Topics > Cybersecurity

Topics

Academic Engagement

Border Security

Citizenship and Immigration Services

Civil Rights and Civil Liberties

Critical Infrastructure Security

Cybersecurity

Our daily life, economic vitality, and national security depend on a stable, safe, and resilient cyberspace.

Cyberspace and its underlying infrastructure are vulnerable to a wide range of risk stemming from both physical and cyber threats and hazards. Sophisticated cyber actors and nation-states exploit vulnerabilities to steal information and money and are developing capabilities to disrupt, destroy, or threaten the delivery of essential services.

Machine Learning & Cyber Security – Success Stories

- Breach and lateral movement detection and prevention
- Identification of malicious activity and rogue server behaviors
- Malware classification
- **Supplement and enhance human security analysis techniques**

Machine Learning & Cyber Security – Challenges

- Researchers and engineers face similar challenges to other applications of Machine Learning

- Cyber security is constantly evolving

- **Call to Arms – Data Scientists and Engineers – We need your help!**

Machine Learning & Cyber Security – Data Challenges

- **Acquiring / Data collection:**
  - Harder to get cyber related data – no one wants to share

- **Training Data**
  - How to obtaining the “right” data sets for a specific cyber security issue?
  - Concerns over data quality and whether data sets are relevant
  - Overfitting and under-fitting related problems

- **How to enable efficient development and testing of your cyber security ideas and theories?**

Machine Learning in Cyber-Security - Problems, Challenges and Data Sets

Idan Amit[1] , John Matherly[2] , William Hewlett[1] , Zhi Xu[1] , Yinnon Meshi[1] , Yigal Weinberger[1]
[1]Palo Alto Networks
[2]Shodan
iamit@paloaltonetworks.com, jmath@shodan.io, whewlett@paloaltonetworks.com,
zxu@paloaltonetworks.com, ymeshi@paloaltonetworks.com, yweinberge@paloaltonetworks.com

9

Data Sets

We think that the use of machine learning in cyber-security should change. We also think that the cyber community should help the machine learning community to become more involved in this field.

One of the key obstacles to investigating cyber-security problems is the lack of appropriate data sets. There are many important cyber-security data sets like Microsoft's malware data set (Ronen et al. 2018), Los Alamos's traffic data set (Turcotte, Kent, and Hash 2017) and EndGame's Ember malware properties data set (Anderson and Roth 2018). However, we feel that there are no suitable data sets that will enable academic researchers to cope with the problems and challenges we listed.

22 Apr 2019

Read the article online here.

SEI > CERT > CWD > Realistic Scenario Simulation Team

- Architect & deliver realistic cyber warfare exercises
- Persistent production spec high-fidelity environment (aka: realism)
- Security hardened and tuned
- Multiple engagements lasting days to 7+ months
- Participants are worldwide DoD operators

Enabling Machine Learning & Cyber Security Success

- **Persistent cyber ranges and resultant data sets:**
  - On-demand access that can be tuned/adjusted
  - Cleaner, labeled data and more balanced
  - Can produce large amounts of unclassified data
  - Ability to re-run cyber security events and scenarios as required

- **Enables efficient development and testing of your cyber security ideas and theories**

Gasoline

Mid-19th century refinery waste product

Almost worthless — thrown away by the barrel and made Ohio’s Cuyahoga River flammable

The oil industry managed to turn its worthless waste product into the most highly sought fuel of the entire 20th century

Enabling Machine Learning & Cyber Security Success

Types of data sets:

- Blue, red, and white team perspectives and their interactions
- Non player character (NPC) behaviors and actions
- Raw data in any form: JSON, PCAP, Netflow, protocol metadata, etc.
- **All data can be cross-correlated w/ cyber related event occurrences**

Part II (Dustin)

Customer related ML work…

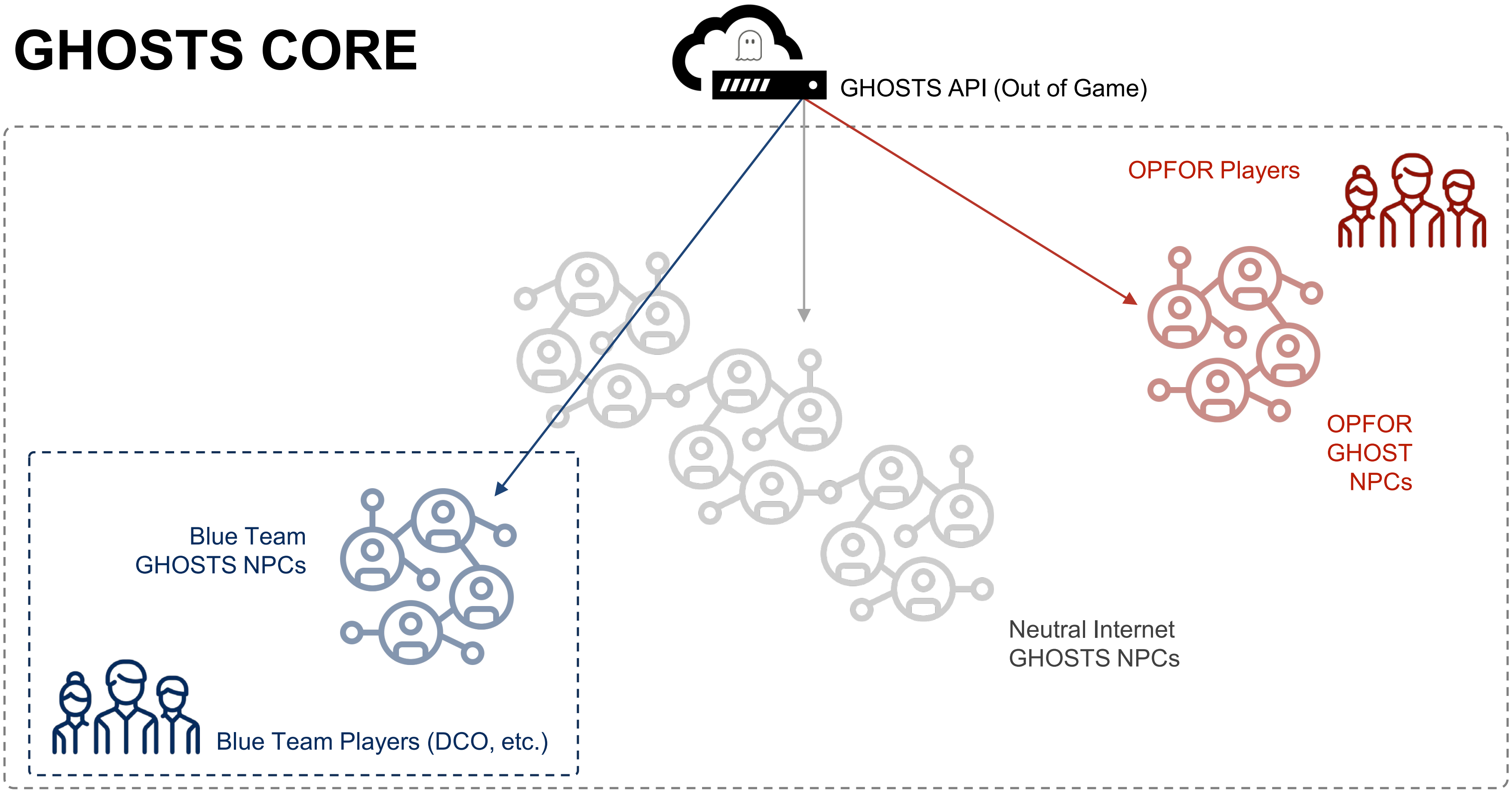
- Large & diverse datasets are a boon for research!
- …but they also make it hard to get started
- Bonus points: How do we enable other teams?

Any exercise contains interactions between:

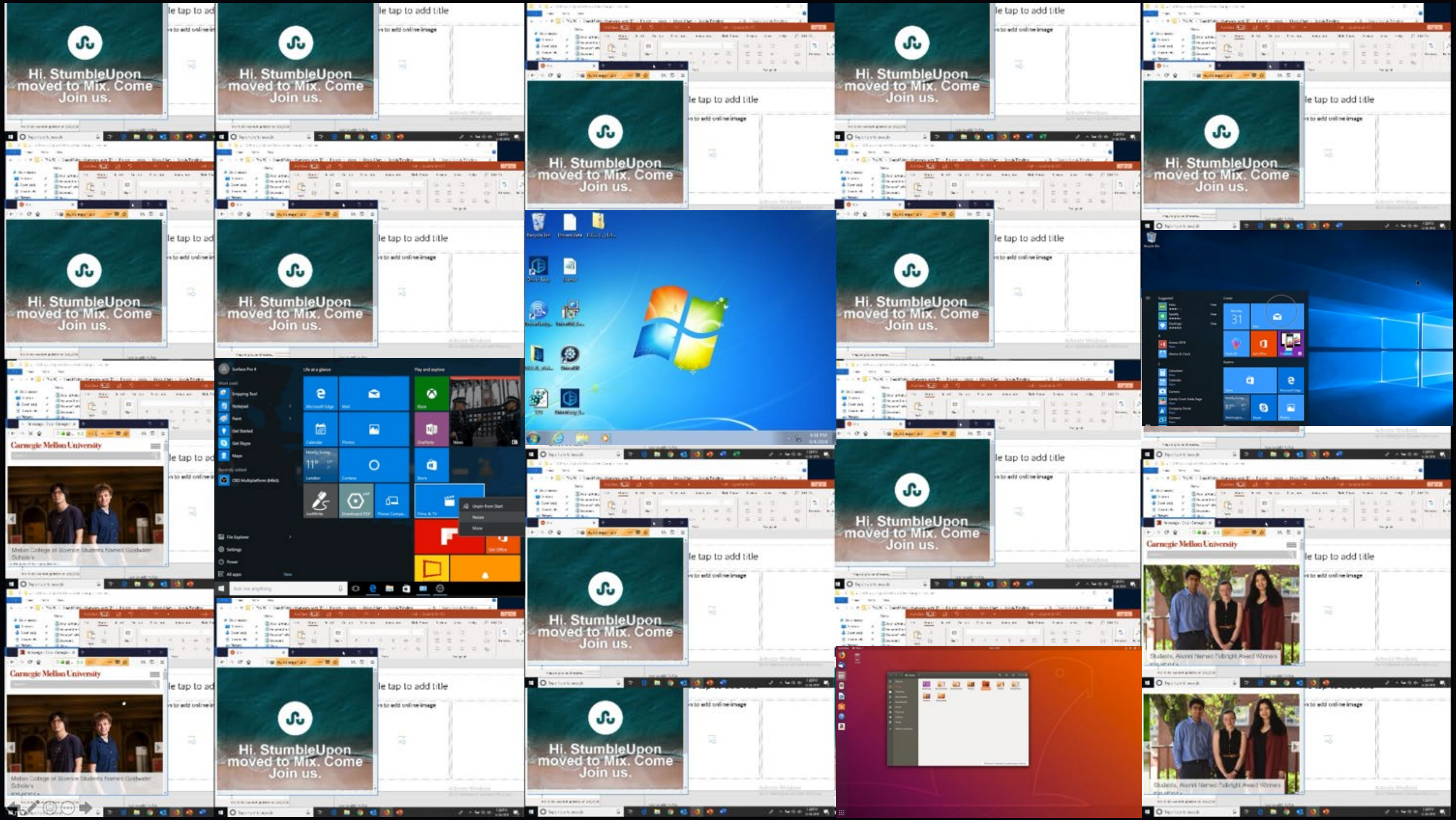
- **Player**
  An active human participant

- **Non-player character (NPC)**
  Any character not controlled by a player within the exercise

GHOSTS CORE



GHOSTS API (Out of Game)
OPFOR Players
OPFOR
GHOST
NPCs
Blue Team
GHOSTS NPCs
Neutral Internet
GHOSTS NPCs
Blue Team Players (DCO, etc.)


Hi. StumbleUpon moved to Mix. Come Join us.
le tap to add title
Carnegie Mellon University

GHOSTS
Last 6 months UTC
NPCs
619
ACTIVITIES
21.6 Mil
POLLING
9.97 Mil
HEALTH
3.89 Mil
ALL ACTIVITIES
min max avg current total
Timeline 2 1.488 K 269 8 14.031076 Mil
Health 0 273 62 7 3.225288 Mil
Polls 0 619 162 33 8.426955 Mil
12/1 12/16 1/1 1/16 2/1 2/15 3/1 3/16 4/1 4/16 5/1 5/16
ALL ACTIVITIES
min max avg current total
Timeline 59 509 237 59 246.774 K
Health 3 127 53 3 55.508 K
Polls 26 270 139 26 145.170 K
2019-05-14 01:40:00
Timeline: 263
Health: 8
Polls: 197
5/13 00:00 5/13 08:00 5/13 16:00 5/14 00:00 5/14 08:00 5/14 16:00 5/15 00:00 5/15 08:00 5/15 16:00 5/16 00:00 5/16 08:00
TIMELINE BREAKDOWN
Data points outside time range
min max avg current total
BrowserFirefox 111 9.19 K 2.02 K 2.75 K 10.61946 Mil
Excel 4 10.98 K 852 2.20 K 4.47662 Mil
Outlook 1 1.41 K 354 1 1.31171 Mil
Word 4 10.98 K 796 2.75 K 4.18565 Mil
PowerPoint 1 10.98 K 177 2.75 K 819.48 K
BrowserChrome 2 15 7 8 88
Command 3 116 87 5 11.66 K
Clicks 3 1.71 K 1.38 K 489 187.85 K

NPCs Make Decisions Based on their Preferences

Alexander.Maxey
+45 Computers
+37 Vim
+55 https://news.ycombinator.com
+20 K8s
+45 Postgresql
-10 EMACS
+10 //files/users/amaxey

username	preference	score
Alexander.Maxey	Computers	44.5
Alfredo.Seaman	Kids	44
Allan.Deal	Recreation	32.5
Ashley.Munson	Arts	37
Carissa.Kelso	Reference	37.5
Cecelia.Nunley	Society	43
Clinton.Belt	Recreation	36.5
Conor.Rouse	Reference	52
Dominick.Ragan	News	56
Donte.Gillette	Home	43.5
Emmanuel.Battle	Games	14
Jaron.Lindstrom	Computers	44
Joey.Crowder	Business	35.5
Joseph.Mosley	Reference	31
Kacy.Kinder	News	39
Krystal.Shepherd	Arts	40.5
Lana.Girard	Society	16
Laurie.Fleming	Shopping	40.5
Leslie.Richmond	Society	42.5
Racheal.Denney	Computers	37
Rashawn.Dow	Science	54
Rodrigo.Rojas	Recreation	32.5
Shayne.Fraley	Science	60
Tarah.Meredith	Shopping	37.5
Tracie.Gamboa	Society	44

These were randomly generated from a sample data set of NPC agents

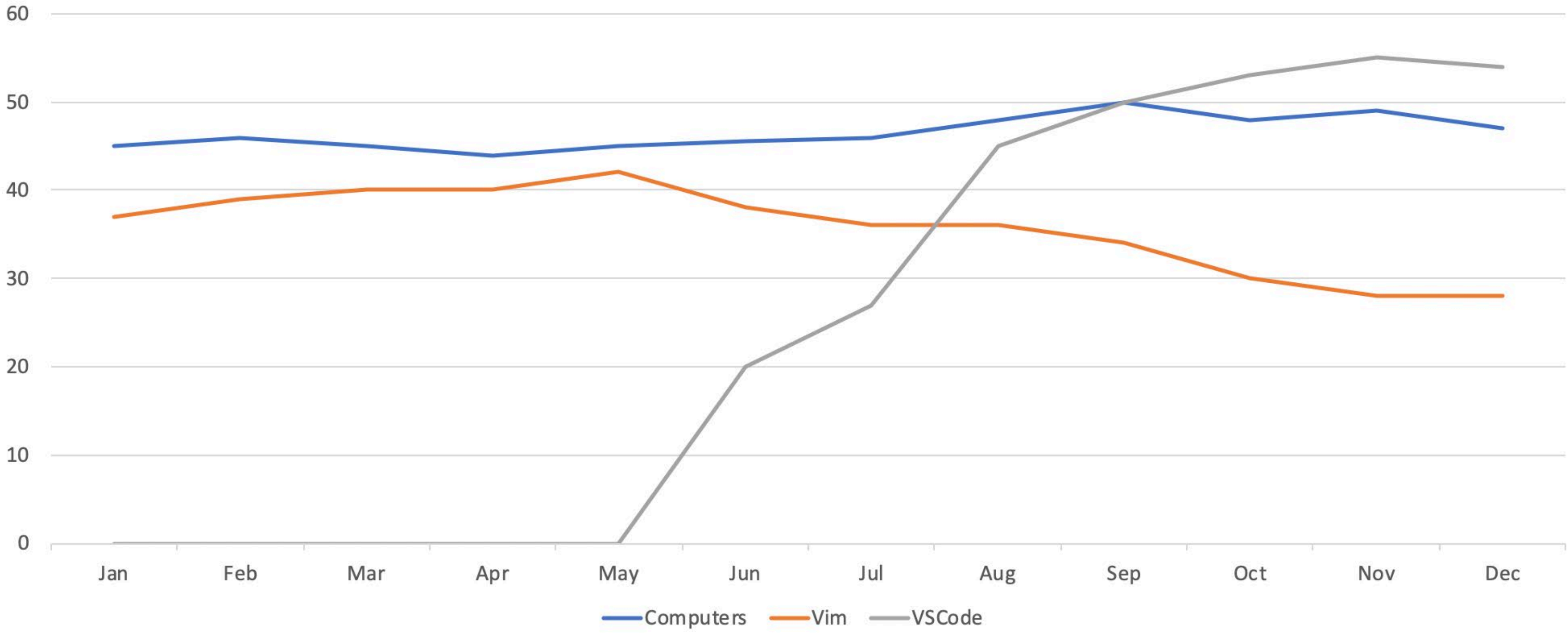
INTRO | USERS AND PREFS | 001_BROWSE_RANDOM | 001_BROWSE_PREFS | 002_BROWSE_PREF | 002_BROWSE_RAND | 003_BROWSE_PREF goldilocks

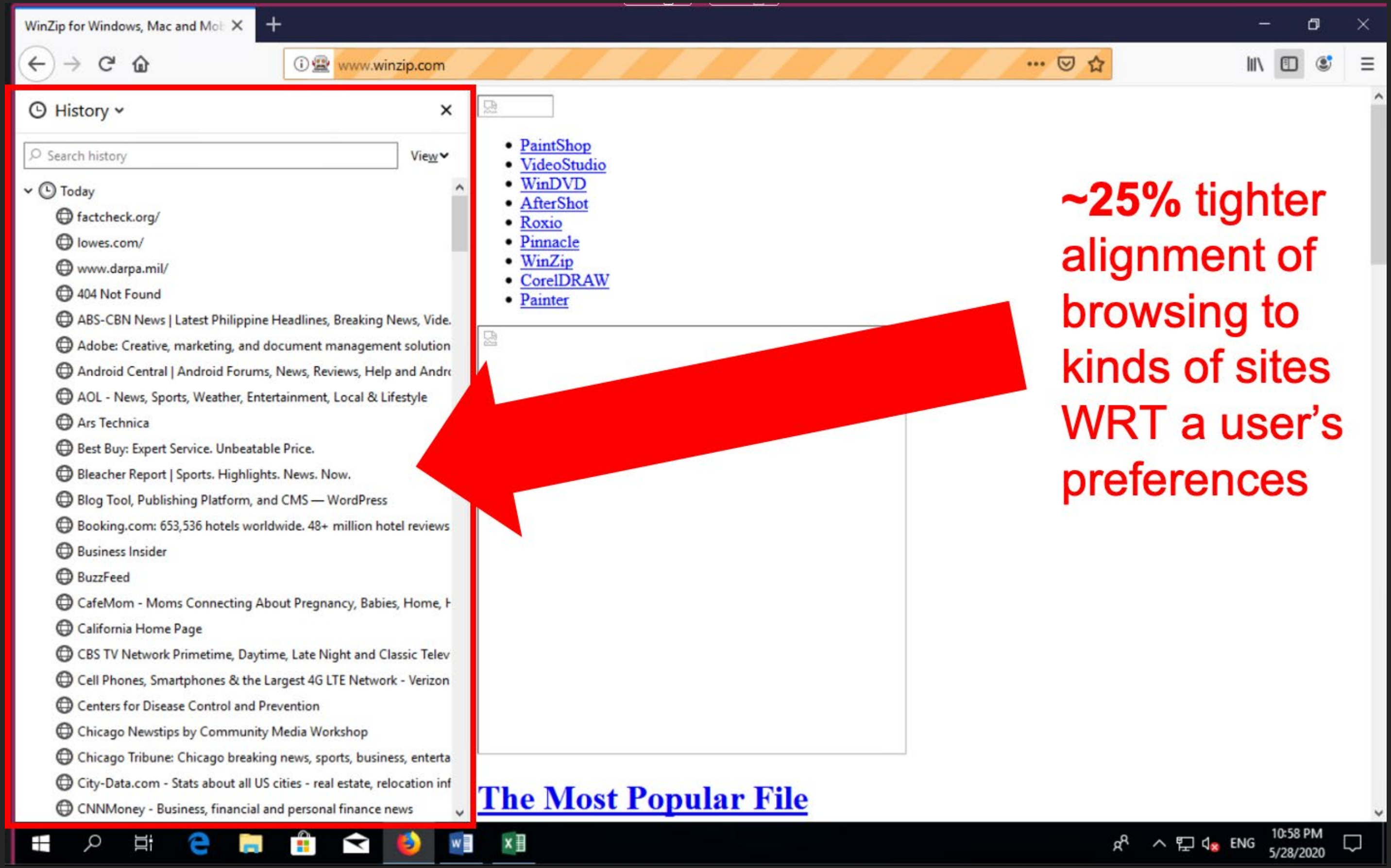
iteration	(All)																
Sum of count	Column Labels																
Row Labels	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Grand Total	Final	Start	Gain
Alexander.Maxey	79	122	34726	15	34	13	17	16	39	34	44	42	44	35225	99%	79%	20%
Alfredo.Seaman	65	113	92	15	32	11	11733	18	37	37	42	44	36	12275	96%	63%	33%
Allan.Deal	84	135	127	18	34	15	26	14	24611	50	49	49	38	25250	97%	68%	30%
Ashley.Munson	21812	123	96	17	39	10	16	27	36	56	42	54	47	22375	97%	70%	28%
Carissa.Kelso	81	106	94	15	38	11	18	31	56	30525	30	45	50	31100	98%	74%	24%
Cecelia.Nunley	94	100	76	15	28	15	9	19	36	45	28	38	34547	35050	99%	79%	20%
Clinton.Belt	89	124	95	16	34	14	22	19	24626	78	49	41	43	25250	98%	68%	29%
Conor.Rouse	66	94	71	8	22	12	14	20	36	30652	41	29	35	31100	99%	80%	19%
Dominick.Ragan	62	79	69	11	30	8	18	20530	27	33	32	33	43	20975	98%	75%	23%
Donte.Gillette	69	102	79	14	36	11991	14	18	41	44	34	46	37	12525	96%	63%	32%
Emmanuel.Battle	102	174	120	3851	59	17	34	25	66	81	55	59	57	4700	82%	26%	56%
Jaron.Lindstrom	73	115	34751	15	31	16	16	19	34	33	35	49	38	35225	99%	80%	19%
Joey.Crowder	91	12146	93	16	33	17	25	23	39	51	43	39	59	12675	96%	64%	32%
Joseph.Mosley	90	123	104	7	39	7	37	21	57	30478	36	44	57	31100	98%	72%	26%
Kacy.Kinder	74	124	96	14	34	19	19	20402	37	44	36	38	38	20975	97%	68%	29%
Krystal.Shepherd	21832	138	94	14	31	10	13	16	34	51	48	55	39	22375	98%	71%	27%
Lana.Girard	103	197	129	20	34	16	33	24	61	70	49	48	33866	34650	98%	67%	31%
Laurie.Fleming	83	113	85	15	35	14	18	19	44	42	34	20548	50	21100	97%	69%	28%
Leslie.Richmond	77	111	94	14	30	11	23	21	37	41	32	52	34507	35050	98%	77%	21%
Racheal.Denney	78	124	34696	14	32	12	27	18	40	51	50	29	54	35225	98%	78%	21%
Rashawn.Dow	66	97	57	13	20	11	12	18	30	30	27227	31	38	27650	98%	80%	19%
Rodrigo.Rojas	80	118	113	21	38	15	27	26	24619	56	46	37	54	25250	98%	68%	30%
Shayne.Fraley	48	81	58	11	17	12	13	7	23	49	27267	36	28	27650	99%	81%	17%
Tarah.Meredith	77	128	83	21	35	12	29	20	46	53	38	20516	42	21100	97%	68%	30%
Tracie.Gamboa	91	111	62	13	26	5	27	19	35	65	39	35	34522	35050	98%	78%	21%
Grand Total	45,466	14,998	106,160	4,203	821	12,294	12,240	41,390	74,747	92,749	55,426	42,037	138,369	640,900			

INTRO | USERS AND PREFS | 001_BROWSE_RANDOM | 001_BROWSE_PREFS | 002_BROWSE_PREF | 002_BROWSE_RAND | 003_BROWSE_PREF

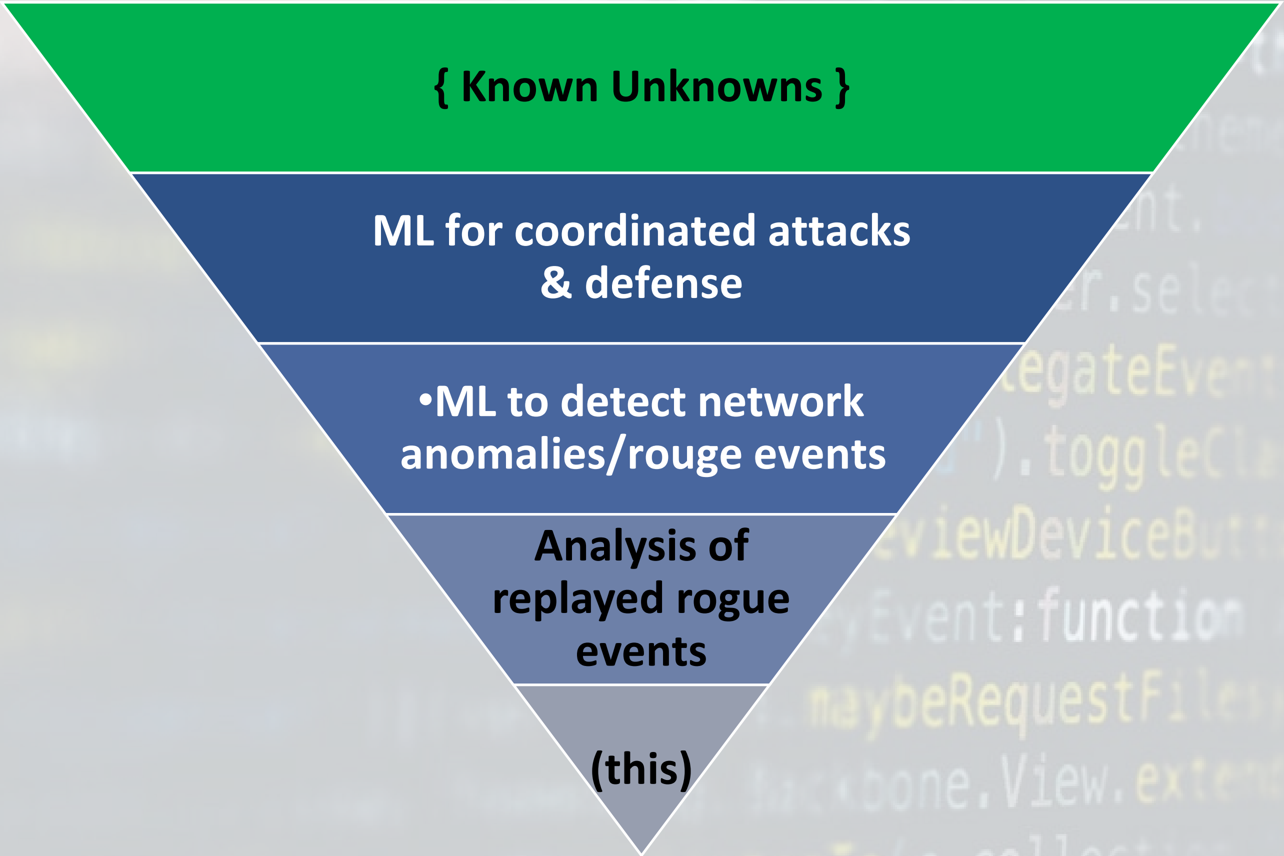
Average: 39.78 Count: 50 Sum: 994.5

An Agent’s Preference Over Time





~25% tighter alignment of browsing to kinds of sites WRT a user's preferences


{ Known Unknowns }
ML for coordinated attacks & defense
•ML to detect network anomalies/rouge events
Analysis of replayed rogue events
(this)

Takeaways

1. All data has ML potential
2. ML projects beget larger & more complex projects
3. Start simple (even for the most daunting datasets)
4. Output of a project feeds the next