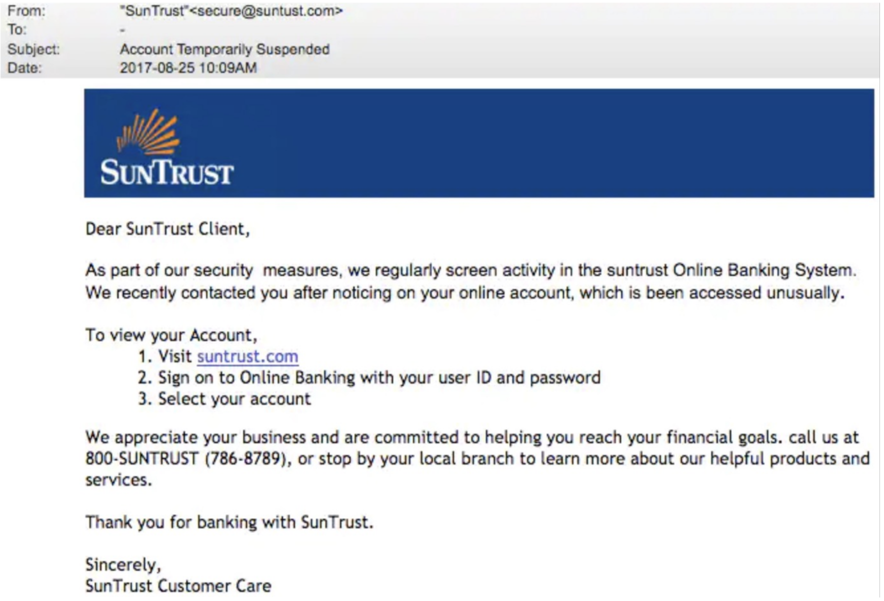
Bright Internet Global Summit 2021 - December 16, 2021

Security and Privacy: *Reconciling the Strengths and Limitations of Human and Artificial Intelligence*

Norman Sadeh
Professor of Computer Science
Carnegie Mellon University

https://normsadeh.org

The Scourge of Phishing



From: "SunTrust"<secure@suntust.com>
To: -
Subject: Account Temporarily Suspended
Date: 2017-08-25 10:09AM

Dear SunTrust Client,

As part of our security measures, we regularly screen activity in the suntrust Online Banking System. We recently contacted you after noticing on your online account, which is been accessed unusually.

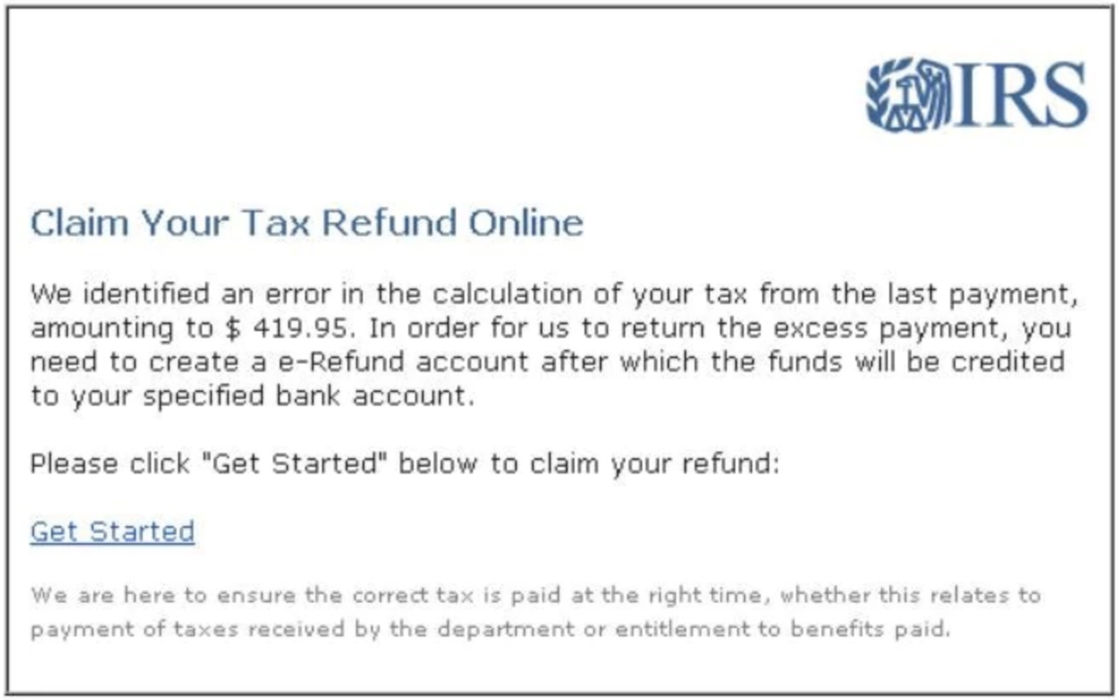
To view your Account,

1. Visit suntrust.com
2. Sign on to Online Banking with your user ID and password
3. Select your account

We appreciate your business and are committed to helping you reach your financial goals. call us at 800-SUNTRUST (786-8789), or stop by your local branch to learn more about our helpful products and services.

Thank you for banking with SunTrust.

Sincerely,
SunTrust Customer Care



Claim Your Tax Refund Online

We identified an error in the calculation of your tax from the last payment, amounting to $ 419.95. In order for us to return the excess payment, you need to create a e-Refund account after which the funds will be credited to your specified bank account.

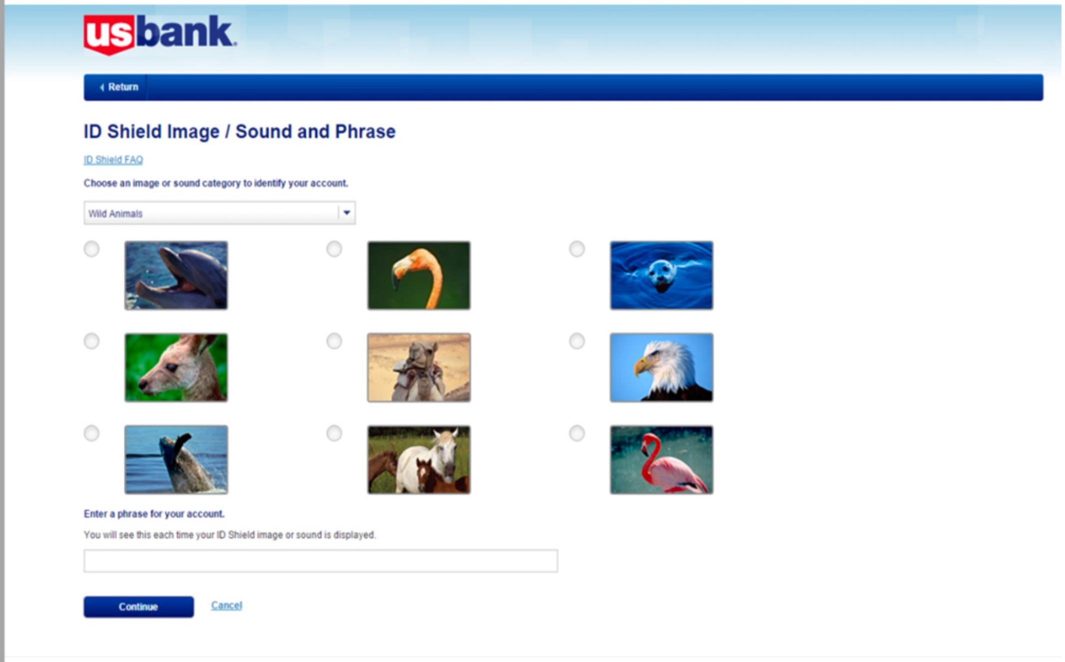
Please click "Get Started" below to claim your refund:

Get Started

We are here to ensure the correct tax is paid at the right time, whether this relates to payment of taxes received by the department or entitlement to benefits paid.

Traditional Solutions Have their Limits

- Spam email filters
- Personalized Security Images
- DKIM
- Emails from system administrators
- Lecturing employees once a year about security
- etc.



ID Shield Image / Sound and Phrase

ID Shield FAQ

Choose an image or sound category to identify your account.

Wild Animals

Enter a phrase for your account.

You will see this each time your ID Shield image or sound is displayed.

Source: US Bank

Making Humans Part of the Solution

Phishing takes advantage of technical limitations when it comes to authenticating different entities (e.g., bank website)...

…but fundamentally it is a **social engineering attack**…

So, why not make humans part of the solution?

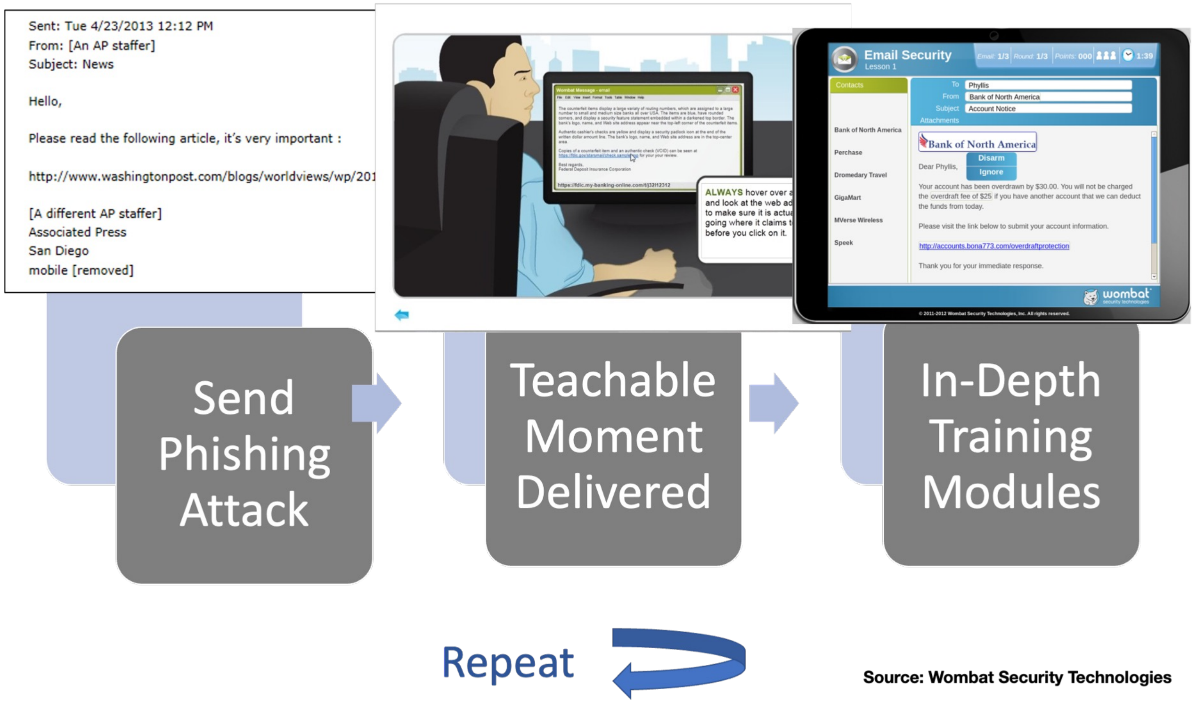
Prevailing View of Cybersecurity Training Circa 2005

Yet, humans are often the greatest source of vulnerability

Mock Phishing Attacks To Educate Users

Phishing Education Example

- Started as research project at CMU
- Built on Learning Science



Sent: Tue 4/23/2013 12:12 PM
From: [An AP staffer]
Subject: News

Hello,

Please read the following article, it's very important :

http://www.washingtonpost.com/blogs/worldviews/wp/201

[A different AP staffer]
Associated Press
San Diego
mobile [removed]

The counterfeit items display a large variaty of routing numbers, which are assigned to a large number to small and medium size banks all over USA. The items are blue, have rounded corners, and display a security feature statement embedded within a darkened top border. The bank's logo, name, and Web site address appear near the top-left corner of the counterfeit items.

Authentic cashier's checks are yellow and display a security padlock icon at the end of the written dollar amount line. The bank's logo, name, and Web site address are in the top-center area.

Copies of a counterfeit item and an authentic check (VOID) can be seen at https://fdic.gov/starsmail/check.sample[illegible] for your your review.

Best regards,
Federal Deposit Insurance Corporation

https://fdic.my-banking-online.com/t/j32l12312

ALWAYS hover over a
and look at the web ad
to make sure it is actua
going where it claims t
before you click on it.

To: Phyllis
From: Bank of North America
Subject: Account Notice

Dear Phyllis,

Your account has been overdrawn by $30.00. You will not be charged the overdraft fee of $25 if you have another account that we can deduct the funds from today.

Please visit the link below to submit your account information.

http://accounts.bona773.com/overdraftprotection

Thank you for your immediate response.

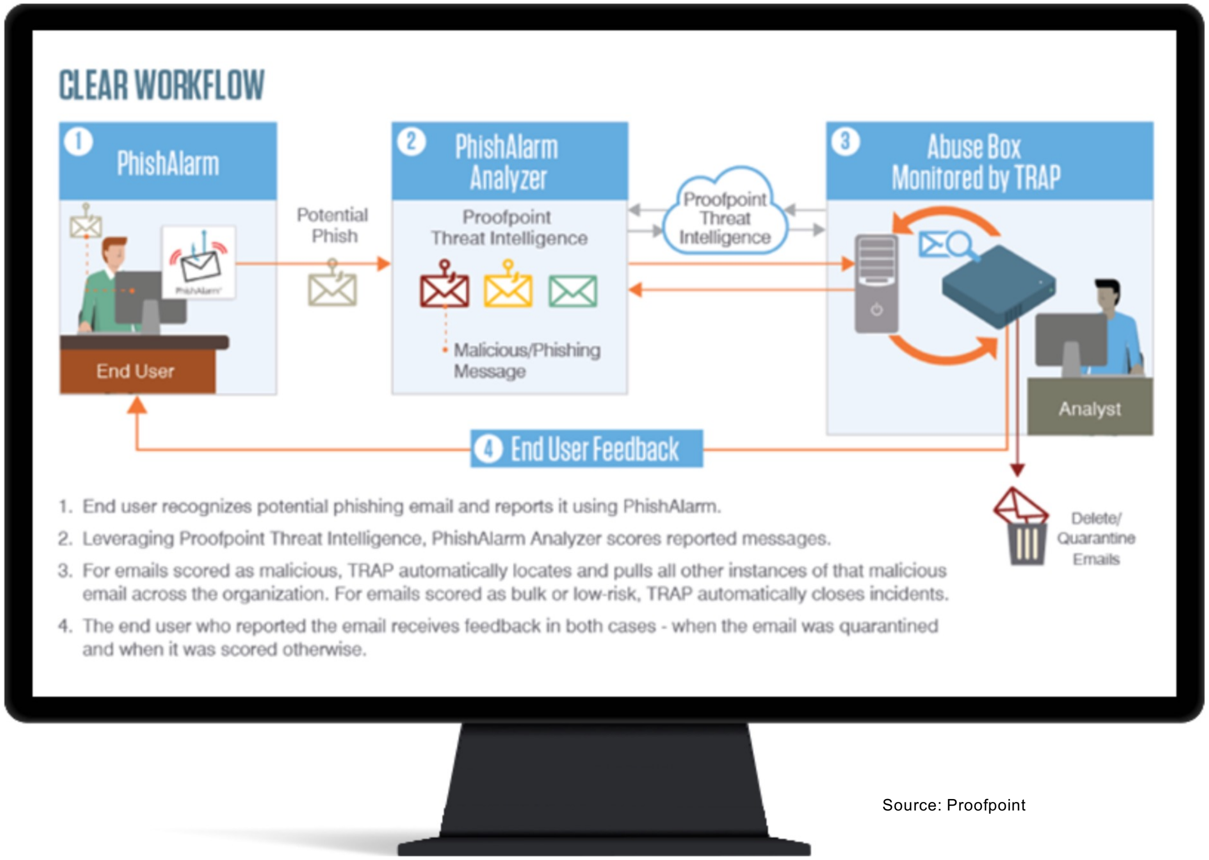
Wombat Message - email
File Edit View Insert Format Tools Table Window Help
Email Security
Lesson 1
Email: 1/3 Round: 1/3 Points: 000 1:39
Contacts
Bank of North America
Perchase
Dromedary Travel
GigaMart
MVerse Wireless
Speek
Attachments
Bank of North America
Disarm
Ignore
wombat security technologies
© 2011-2012 Wombat Security Technologies, Inc. All rights reserved.
Send Phishing Attack
Teachable Moment Delivered
In-Depth Training Modules
Repeat

Source: Wombat Security Technologies

- **Show people they are susceptible to attacks to get their attention**
- **Use teachable moment to teach them practical tips**

- Incorporated as Wombat Security Technologies in 2008
- Became de facto standard for training users & developed a suite of products for other threats
- Acquired by Proofpoint (NASDAQ: PFPT) in March 2018
- Among 500 fastest growing businesses in the US for 3 years in a row
- Thousands of corporate customers, tens of millions of users; about half of Fortune 500 companies as customers

PhishAlarm Analyzer: AI and Humans Working Together


CLEAR WORKFLOW
1 PhishAlarm
PhishAlarm
End User
Potential Phish
2 PhishAlarm Analyzer
Proofpoint Threat Intelligence
Malicious/Phishing Message
Proofpoint Threat Intelligence
3 Abuse Box Monitored by TRAP
Analyst
4 End User Feedback
Delete/ Quarantine Emails

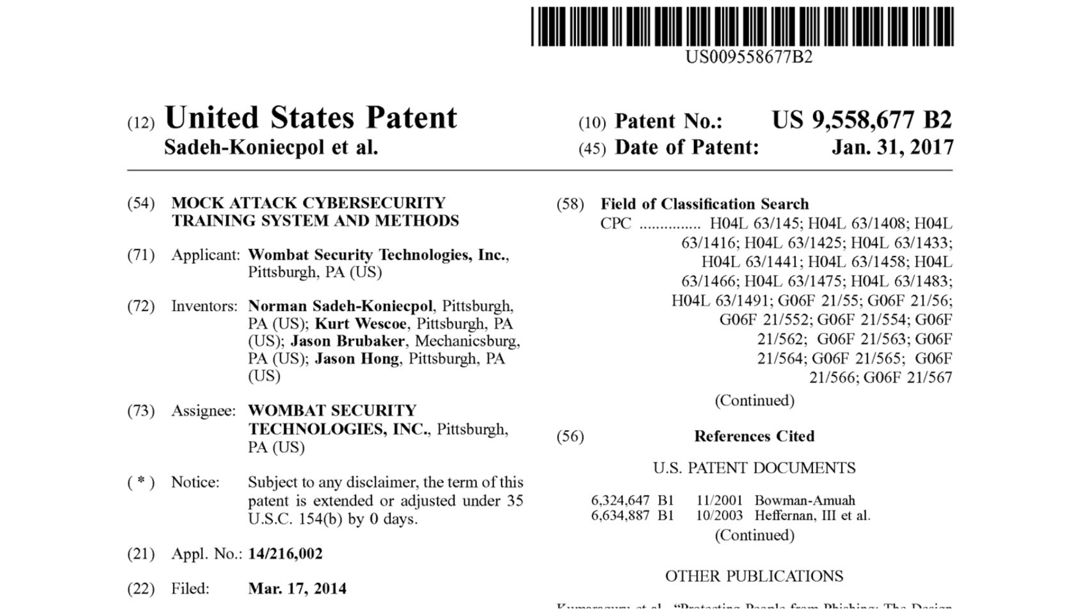
1. End user recognizes potential phishing email and reports it using PhishAlarm.
2. Leveraging Proofpoint Threat Intelligence, PhishAlarm Analyzer scores reported messages.
3. For emails scored as malicious, TRAP automatically locates and pulls all other instances of that malicious email across the organization. For emails scored as bulk or low-risk, TRAP automatically closes incidents.
4. The end user who reported the email receives feedback in both cases - when the email was quarantined and when it was scored otherwise.

Source: Proofpoint

- **User reports are triaged using AI/ML & security analysts make the final decision**
- **Adapted filtering technology - rather than let AI/ML have the final say, reintroduce people into the process**

Phishing Can Take Many Forms

- SMS
- Facebook
- QR codes
- Phone calls - incl. Deepfakes…
- Malicious WiFi access points
- etc.



US009558677B2

(12) **United States Patent**
Sadeh-Koniecpol et al.

(10) **Patent No.:** **US 9,558,677 B2**
(45) **Date of Patent:** **Jan. 31, 2017**

(54) **MOCK ATTACK CYBERSECURITY TRAINING SYSTEM AND METHODS**

(71) Applicant: **Wombat Security Technologies, Inc.**, Pittsburgh, PA (US)

(72) Inventors: **Norman Sadeh-Koniecpol**, Pittsburgh, PA (US); **Kurt Wescoe**, Pittsburgh, PA (US); **Jason Brubaker**, Mechanicsburg, PA (US); **Jason Hong**, Pittsburgh, PA (US)

(73) Assignee: **WOMBAT SECURITY TECHNOLOGIES, INC.**, Pittsburgh, PA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **14/216,002**

(22) Filed: **Mar. 17, 2014**

(58) **Field of Classification Search**
CPC H04L 63/145; H04L 63/1408; H04L 63/1416; H04L 63/1425; H04L 63/1433; H04L 63/1441; H04L 63/1458; H04L 63/1466; H04L 63/1475; H04L 63/1483; H04L 63/1491; G06F 21/55; G06F 21/56; G06F 21/552; G06F 21/554; G06F 21/562; G06F 21/563; G06F 21/564; G06F 21/565; G06F 21/566; G06F 21/567

(Continued)

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,324,647	B1	11/2001	Bowman-Amuah
6,634,887	B1	10/2003	Heffernan, III et al.

(Continued)

OTHER PUBLICATIONS

Kumaraguru et al. "Protecting People from Phishing: The Design

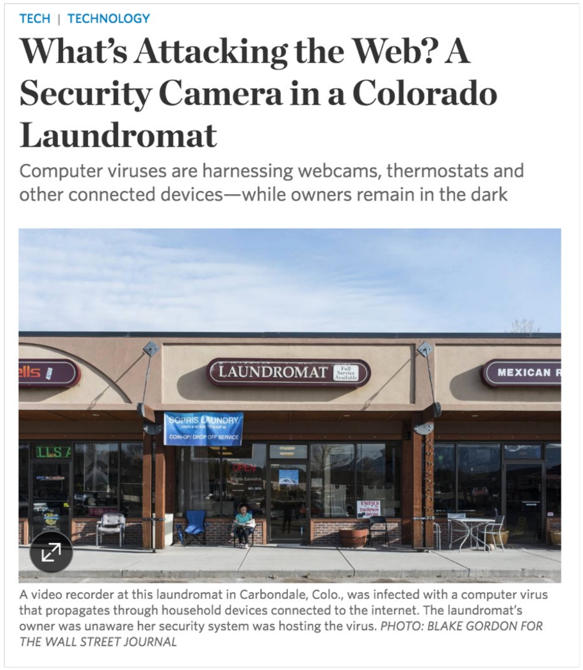
Phishing is Just An Example of…

…the important **role played by people in security**…and the challenges people are confronted to when dealing with an **increasingly more complex and diverse set of attack surfaces**

- **People as users**
- **People as developers**
- **People as platform providers**
- **People as regulators**

Everyone is now a Sysadmin

Mirai Botnet Attack of Oct. 2016: 600,000 compromised devices creating traffic in excess of 1.2Tbps - DDoS Attacks



TECH | TECHNOLOGY

What's Attacking the Web? A Security Camera in a Colorado Laundromat

Computer viruses are harnessing webcams, thermostats and other connected devices—while owners remain in the dark

A video recorder at this laundromat in Carbondale, Colo., was infected with a computer virus that propagates through household devices connected to the internet. The laundromat's owner was unaware her security system was hosting the virus. *PHOTO: BLAKE GORDON FOR THE WALL STREET JOURNAL*

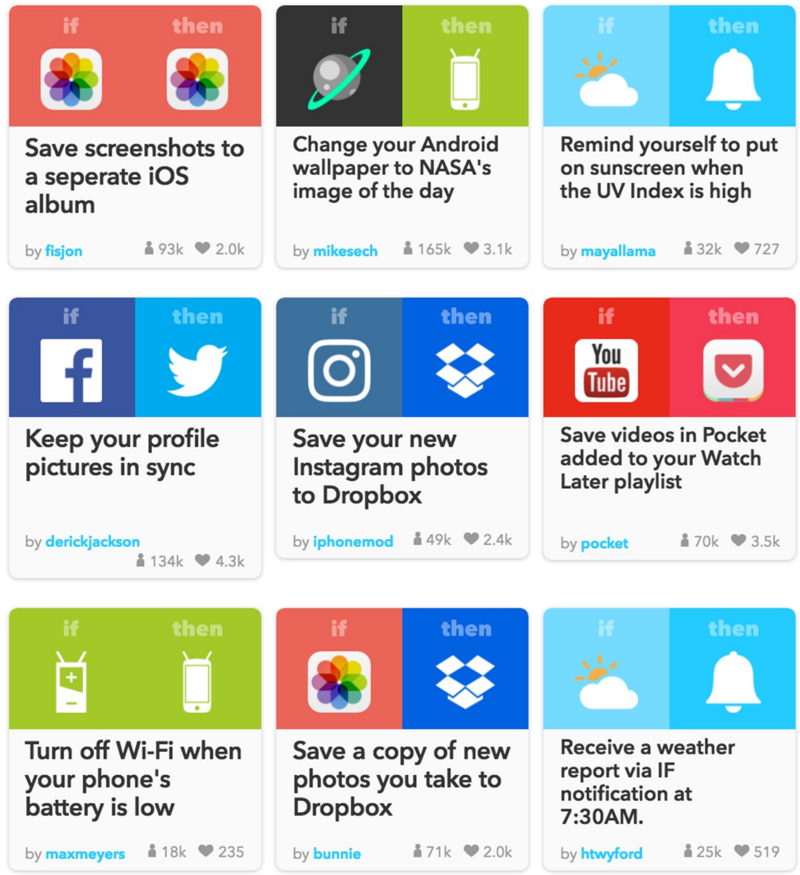
- Owner didn't notice traffic generated by her camera
- Camera would regularly crash but she learned to just restart it
- She lost her password but the manufacturer just resets the password to its default (123456) when this happens
- The security person who installed the camera learned about the virus after being contacted by the press
- Camera manufacturer denies any responsibility

Everyone is a Developer - I

App Stores, IoT Platforms, etc.

Sample Shared Recipes

https://ifttt.com/recipes



Save screenshots to a seperate iOS album

by fisjon 93k ♥ 2.0k

Change your Android wallpaper to NASA's image of the day

by mikesech 165k ♥ 3.1k

Remind yourself to put on sunscreen when the UV Index is high

by mayallama 32k ♥ 727

Keep your profile pictures in sync

by derickjackson 134k ♥ 4.3k

Save your new Instagram photos to Dropbox

by iphonemod 49k ♥ 2.4k

Save videos in Pocket added to your Watch Later playlist

by pocket 70k ♥ 3.5k

Turn off Wi-Fi when your phone's battery is low

by maxmeyers 18k ♥ 235

Save a copy of new photos you take to Dropbox

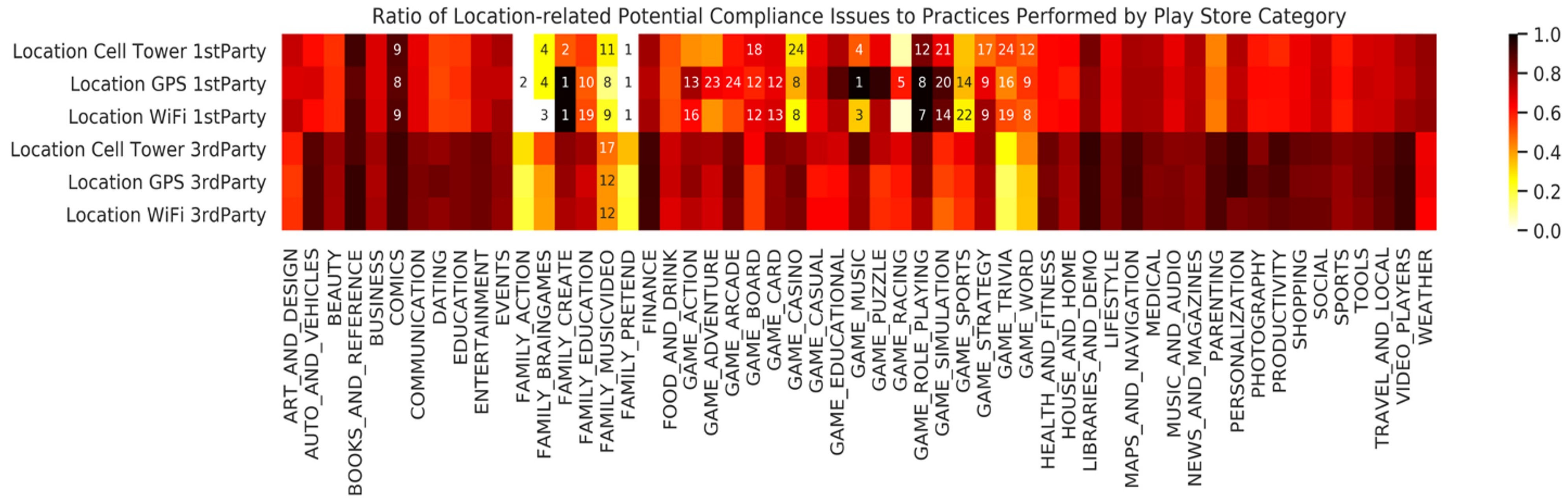
by bunnie 71k ♥ 2.0k

Receive a weather report via IF notification at 7:30AM.

by htwyford 25k ♥ 519

Everyone is a Developer - II

Potential Privacy Compliance Issues - Automated Analysis of over 1 million Android Apps - Darker color indicates a data practice appears to be performed but not disclosed



The Human Bottleneck

Lack of:

- **Expertise**
- **Time**
- **Attention**
- **Motivation**

Source: https://www.datanami.com/2016/09/13/sas-goes-back-future-cognitive-computing-viya/

Privacy as a Usability Challenge - I

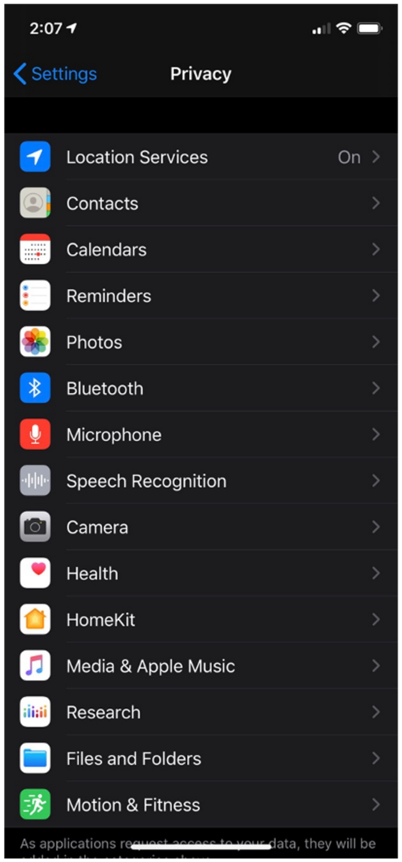
“**Notice and Choice**” is at the core of privacy regimes around the world.

Yet, as someone once said: “***Only in Fantasy Land, do people read the text of privacy policies***”

Privacy as a Usability Challenge - II

..and who has the time to configure all their privacy settings - let alone understand what they really mean?

“*Notice and Choice is broken*”
Fred Cate

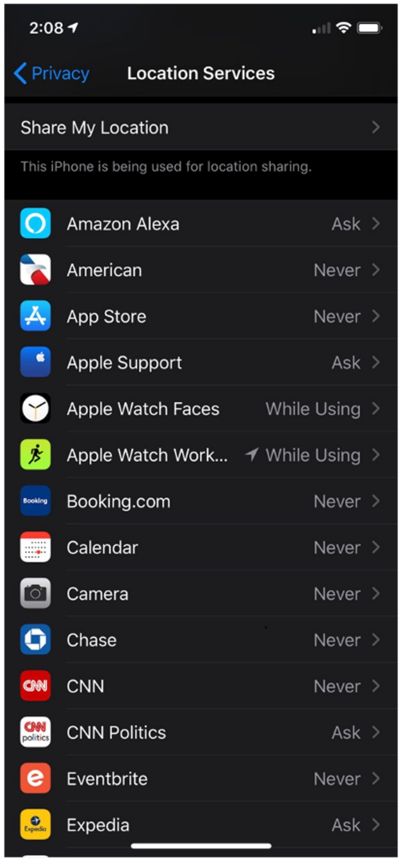


2:07

< Settings | Privacy

Location Services	On >
Contacts	>
Calendars	>
Reminders	>
Photos	>
Bluetooth	>
Microphone	>
Speech Recognition	>
Camera	>
Health	>
HomeKit	>
Media & Apple Music	>
Research	>
Files and Folders	>
Motion & Fitness	>

As applications request access to your data, they will be



2:08

< Privacy | Location Services

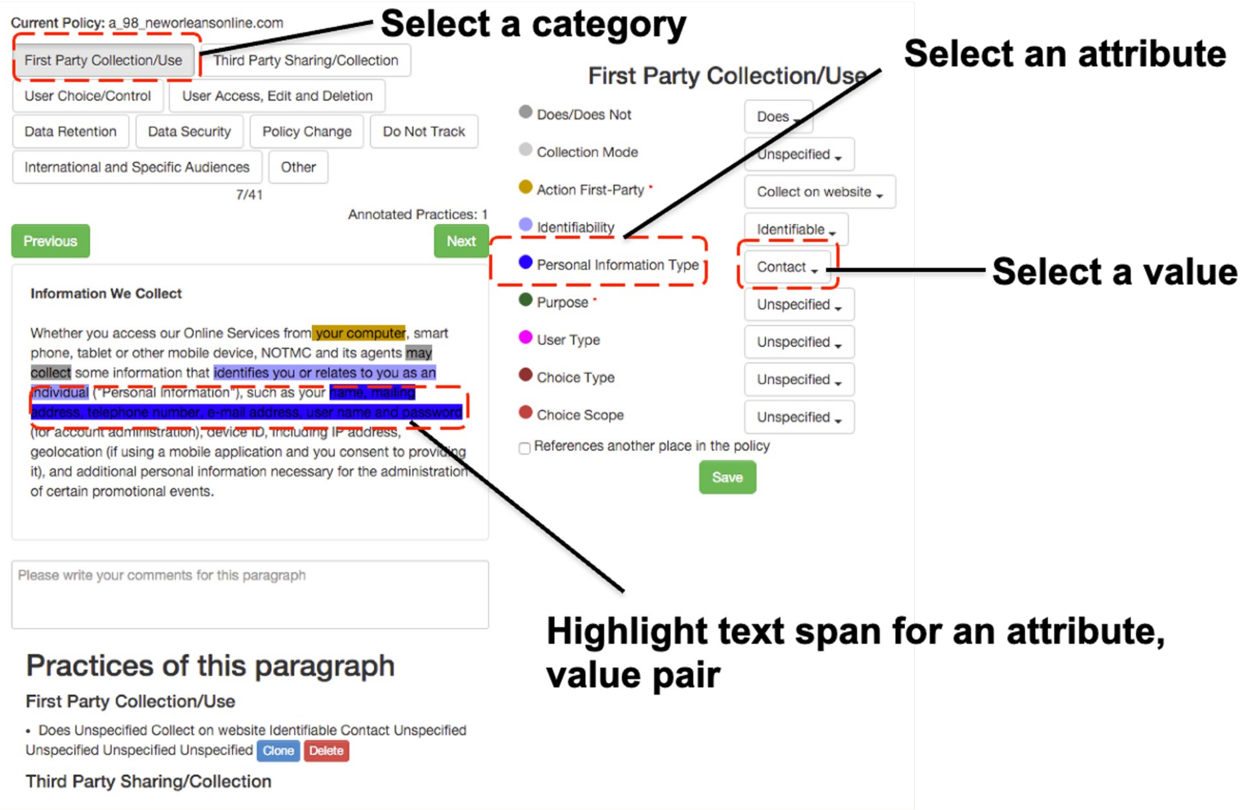
Share My Location >

This iPhone is being used for location sharing.

Amazon Alexa	Ask >
American	Never >
App Store	Never >
Apple Support	Ask >
Apple Watch Faces	While Using >
Apple Watch Work...	While Using >
Booking.com	Never >
Calendar	Never >
Camera	Never >
Chase	Never >
CNN	Never >
CNN Politics	Ask >
Eventbrite	Never >
Expedia	Ask >

What If Computers Understood the Text of Privacy Policies?

Annotation Tool



Current Policy: a_98_neworleansonline.com

First Party Collection/Use
Third Party Sharing/Collection
User Choice/Control
User Access, Edit and Deletion
Data Retention
Data Security
Policy Change
Do Not Track
International and Specific Audiences
Other
7/41
Annotated Practices: 1
Previous
Next

Information We Collect

Whether you access our Online Services from your computer, smart phone, tablet or other mobile device, NOTMC and its agents may collect some information that identifies you or relates to you as an individual ("Personal Information"), such as your name, mailing address, telephone number, e-mail address, user name and password (for account administration), device ID, including IP address, geolocation (if using a mobile application and you consent to providing it), and additional personal information necessary for the administration of certain promotional events.

Please write your comments for this paragraph

Practices of this paragraph

First Party Collection/Use

- Does Unspecified Collect on website Identifiable Contact Unspecified Unspecified Unspecified Unspecified

Third Party Sharing/Collection

First Party Collection/Use

● Does/Does Not	Does
● Collection Mode	Unspecified
● Action First-Party *	Collect on website
● Identifiability	Identifiable
● Personal Information Type	Contact
● Purpose *	Unspecified
● User Type	Unspecified
● Choice Type	Unspecified
● Choice Scope	Unspecified

□ References another place in the policy

Select a category

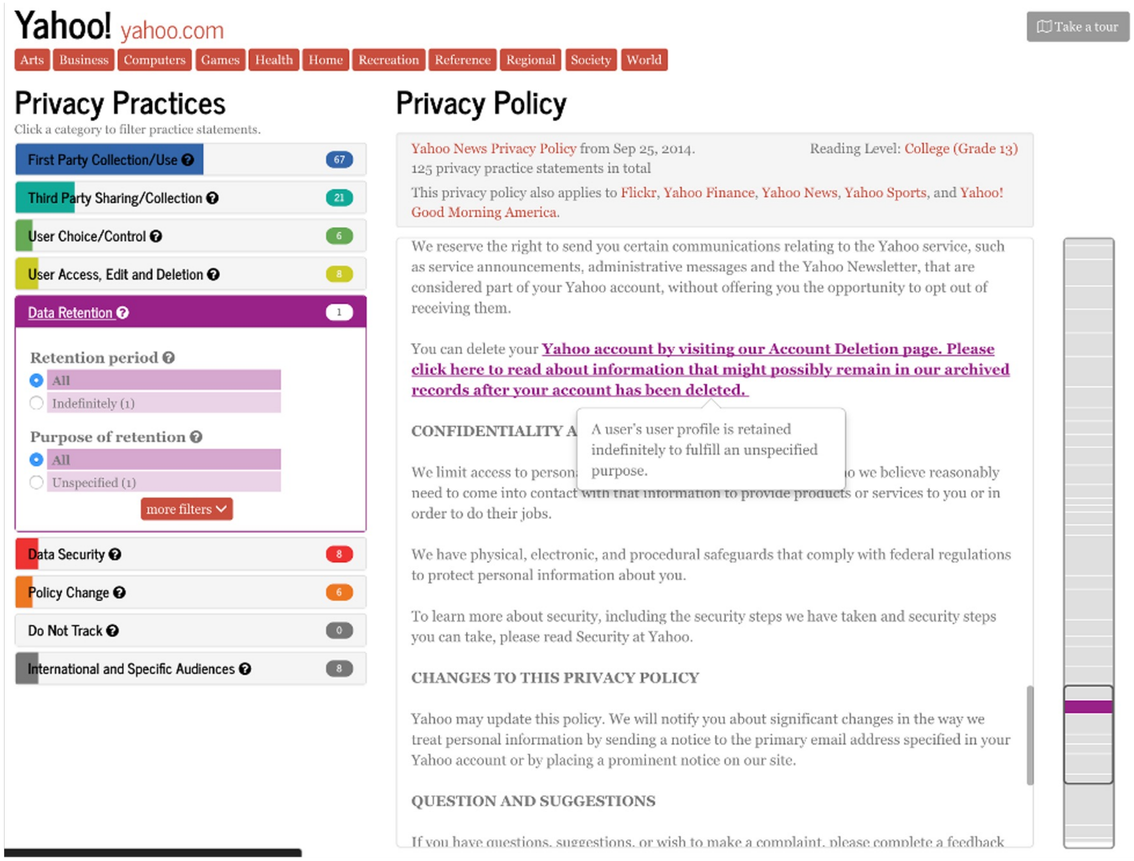
Select an attribute

Select a value

Highlight text span for an attribute, value pair

S. Wilson, F. Schaub, A. Dara, F. Liu, S. Cherivirala, P.G. Leon, M.S. Andersen, S. Zimmeck, K. Sathyendra, N.C. Russell, T.B. Norton, E. Hovy, J.R. Reidenberg, N. Sadeh, "The Creation and Analysis of a Website Privacy Policy Corpus", ACL '16: Annual Meeting of the Association for Computational Linguistics, Aug 2016

Automatic Identification of Data Practice Disclosures



Yahoo! yahoo.com

Privacy Practices

Click a category to filter practice statements.

First Party Collection/Use	67
Third Party Sharing/Collection	21
User Choice/Control	6
User Access, Edit and Deletion	8
Data Retention	1

Retention period

● All
○ Indefinitely (1)

Purpose of retention

● All
○ Unspecified (1)

Data Security	8
Policy Change	6
Do Not Track	0
International and Specific Audiences	8

Privacy Policy

Yahoo News Privacy Policy from Sep 25, 2014. Reading Level: College (Grade 13)
125 privacy practice statements in total

This privacy policy also applies to Flickr, Yahoo Finance, Yahoo News, Yahoo Sports, and Yahoo! Good Morning America.

We reserve the right to send you certain communications relating to the Yahoo service, such as service announcements, administrative messages and the Yahoo Newsletter, that are considered part of your Yahoo account, without offering you the opportunity to opt out of receiving them.

You can delete your **Yahoo account by visiting our Account Deletion page. Please click here to read about information that might possibly remain in our archived records after your account has been deleted.**

A user's user profile is retained indefinitely to fulfill an unspecified purpose.

CONFIDENTIALITY A

We limit access to person o we believe reasonably need to come into contact with that information to provide products or services to you or in order to do their jobs.

We have physical, electronic, and procedural safeguards that comply with federal regulations to protect personal information about you.

To learn more about security, including the security steps we have taken and security steps you can take, please read Security at Yahoo.

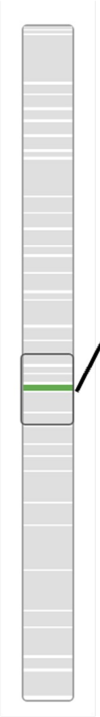
CHANGES TO THIS PRIVACY POLICY

Yahoo may update this policy. We will notify you about significant changes in the way we treat personal information by sending a notice to the primary email address specified in your Yahoo account or by placing a prominent notice on our site.

QUESTION AND SUGGESTIONS

If you have questions, suggestions, or wish to make a complaint, please complete a feedback

User Choice Instance Extraction



Choice Instance !!!
If you do not want us to use personal information that we gather to allow third parties to personalize advertisements we display to you, please adjust your Advertising Preferences .

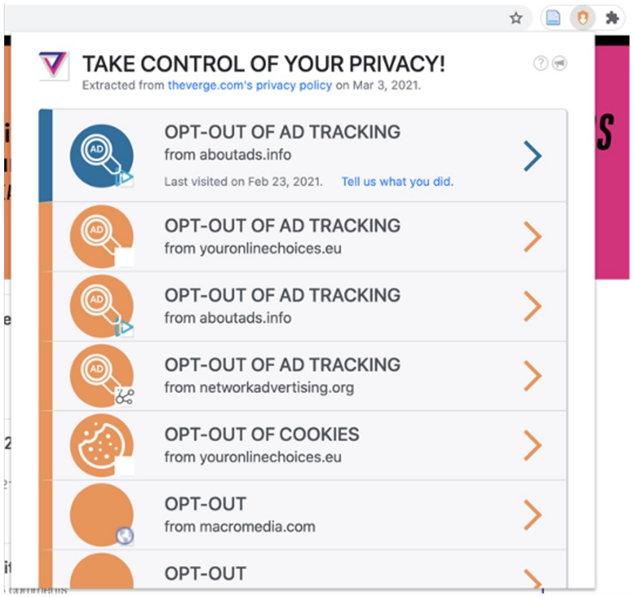
- User choices often buried deep in the text of long policies
- Is it possible to **automatically extract information** about such “choice instances” from privacy policies?
- Use Natural Language Toolkit tokenizer to subdivide segments into sentences & build classifiers

Results: Recall & Accuracy > 90%

Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Faith Cranor, Shomir Wilson, Florian Schaub, Norman Sadeh, **"Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text"**, WWW ’20, Apr 2020 [pdf]

Applications

- **Helping end users**
  - Opt-Out Easy Browser Extension (available in Chrome store)
  - Privacy Q&A
- **Helping Product Managers, Privacy Engineers and Developers**
  - MAPS Privacy Compliance tool - used to help with GDPR compliance
- **Helping Platforms and Regulators**
  - Automated compliance analysis of mobile apps - results shared with FTC, Cal AG



TAKE CONTROL OF YOUR PRIVACY!

Extracted from theverge.com's privacy policy on Mar 3, 2021.

OPT-OUT OF AD TRACKING
from aboutads.info
Last visited on Feb 23, 2021. Tell us what you did.

OPT-OUT OF AD TRACKING
from youronlinechoices.eu

OPT-OUT OF AD TRACKING
from aboutads.info

OPT-OUT OF AD TRACKING
from networkadvertising.org

OPT-OUT OF COOKIES
from youronlinechoices.eu

OPT-OUT
from macromedia.com

OPT-OUT

…But Automation Only Goes So Far…

Example: Privacy Question Answering:

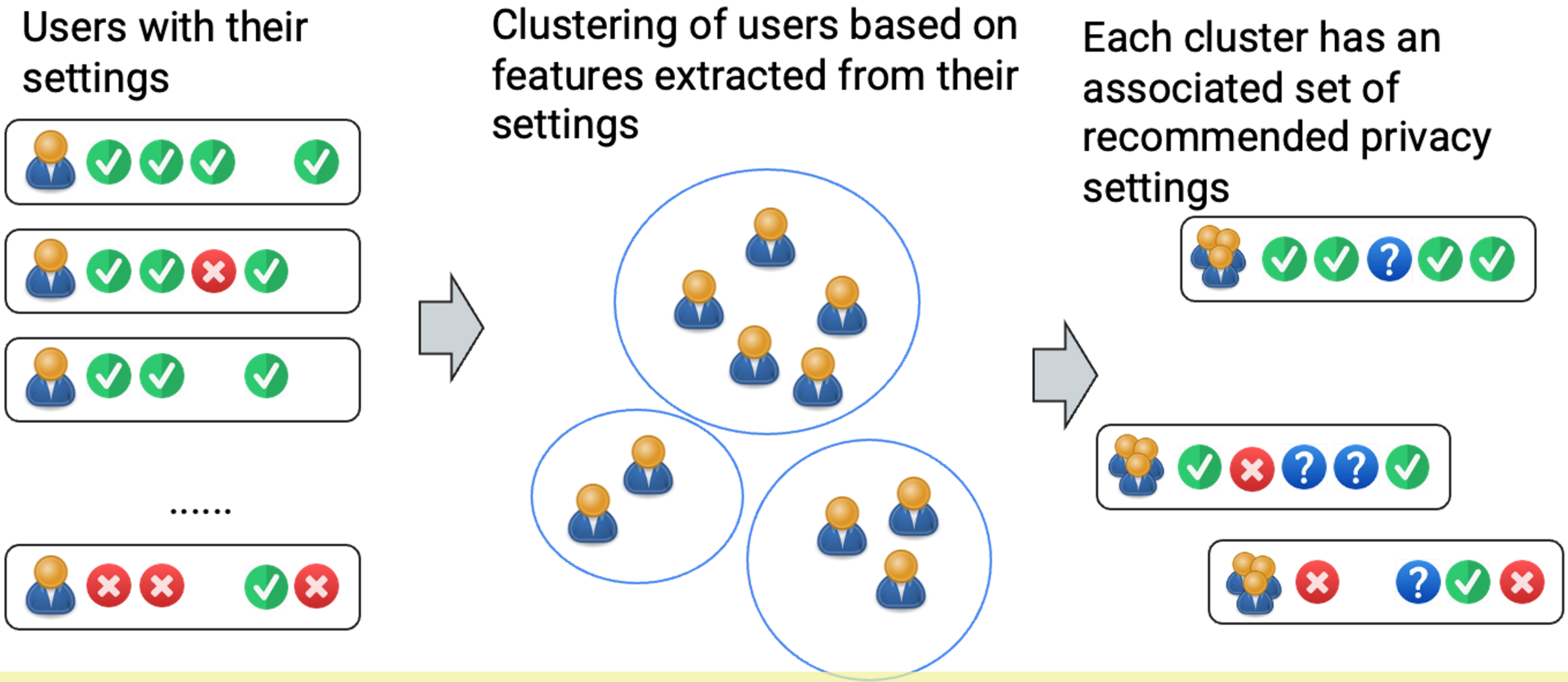
- Difficulty of user to articulate their questions
- Privacy Policies are vague and ambiguous
- Need to come up with answers that are useful and legally sound at the same time

Example: Automated Compliance Analysis

- Need human verification (e.g., interpretation of policy statements or of what the code does)

What If Computers Understood People's Privacy Concerns and Expectations?

Privacy Assistants - I



Users with their settings

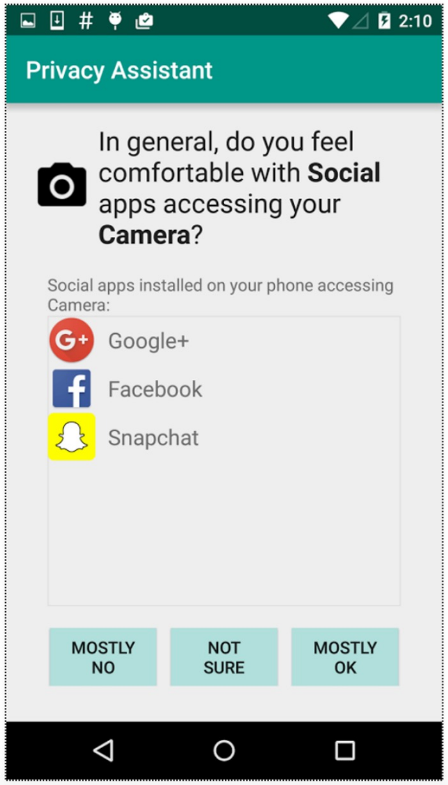
Clustering of users based on features extracted from their settings

Each cluster has an associated set of recommended privacy settings

B. Liu, M.S. Andersen, F. Schaub, H. Almuhimedi, S. Zhang, N. Sadeh, A. Acquisti, and Y. Agarwal, **"Follow My Recommendations: A Personalized Assistant for Mobile App Permissions"**, SOUPS 2016- US patents 10,956,586

Privacy Assistants II

Generating recommendations rather than automating privacy decisions

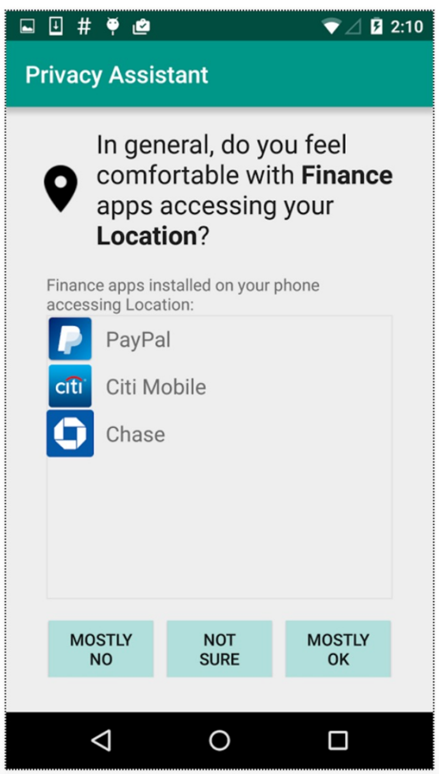


Privacy Assistant

In general, do you feel comfortable with **Social** apps accessing your **Camera**?

Social apps installed on your phone accessing Camera:

- Google+
- Facebook
- Snapchat

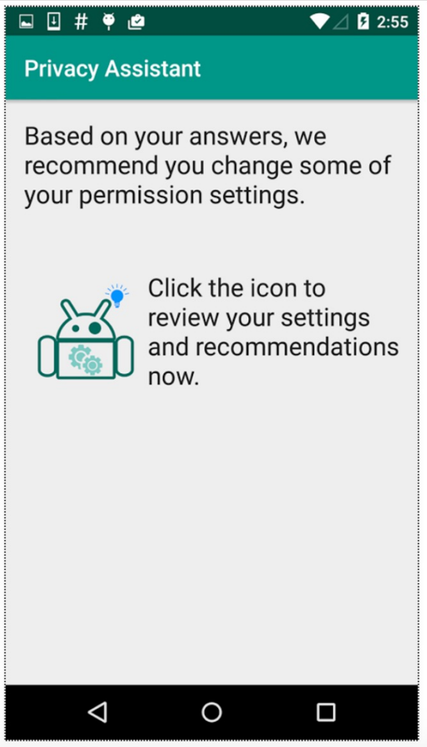


Privacy Assistant

In general, do you feel comfortable with **Finance** apps accessing your **Location**?

Finance apps installed on your phone accessing Location:

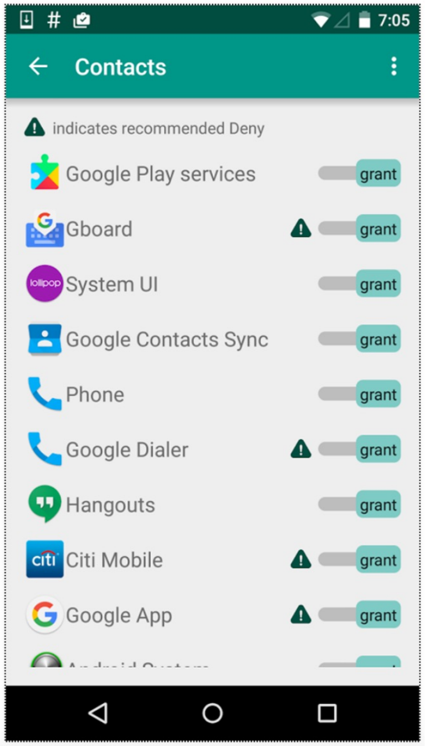
- PayPal
- Citi Mobile
- Chase



Privacy Assistant

Based on your answers, we recommend you change some of your permission settings.

Click the icon to review your settings and recommendations now.



Contacts

⚠ indicates recommended Deny

Google Play services		grant
Gboard	⚠	grant
System UI		grant
Google Contacts Sync		grant
Phone		grant
Google Dialer	⚠	grant
Hangouts		grant
Citi Mobile	⚠	grant
Google App	⚠	grant

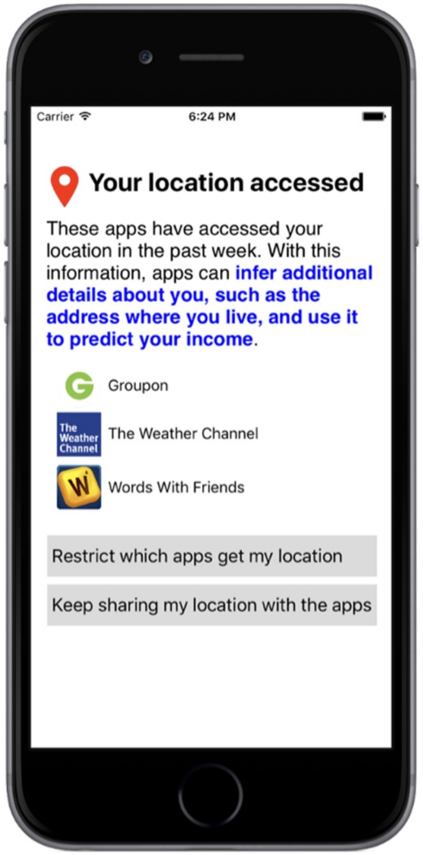
Successfully deployed in Google Play store for rooted phones for several years

Why Recommendations?

Agency is a major part of privacy: users should remain in charge of their decisions…but **AI can help** them make these decisions and can help overcome fundamental **usability limitations**

- **Major requirement**: the recommendations have to be ***understandable*** and ***auditable***

Explanation is not easy - Example: Framing



Your location accessed

These apps have accessed your location in the past week. With this information, apps can **infer additional details about you, such as the address where you live, and use it to predict your income**.

Groupon
The Weather Channel
Words With Friends
Restrict which apps get my location
Keep sharing my location with the apps

The Impact of Framing Measured by Change in Privacy Settings

Condition	Wording
(1) Baseline	“These apps have **accessed** your location in the past week.”
(2) Frequency	“These apps have accessed your location **1,865 times** in the past week.”
(3) Background	“These apps have accessed your location in the past week although **you did not use them**.”
(4) Purposes	“These apps have accessed your location in the past week for purposes **not related to the apps’ main function**.”
(5) Purposes + Example	“These apps have accessed your location in the past week for purposes **not related to the apps’ main function, such as location-based advertising**.”
(6) Inferences	“These apps have accessed your location in the past week. With this information, apps can **infer additional details about you.”**
(7) Inferences + Example	“These apps have accessed your location in the past week. With this information, apps can **infer additional details about you, such as the address where you live**.”
(8) Predictions	“These apps have accessed your location in the past week. With this information, apps can **infer additional details about you, such as the address where you live, and use it to predict your income**.”
(9) Predictions + Implications	“These apps have accessed your location in the past week. With this information, apps can **infer additional details about you, such as the address where you live, and use it to predict your income. Knowing your income can affect prices and discounts you see in ads**.”

Purple color indicates statistically significant difference compared to baseline (Halmuhimedi, 2018)

The Big Picture - I

Security and privacy are increasingly challenging

- **Software-centric** and **data-centric** economy
- Layers upon layers of functionality/system of systems
- Complex **dataflows** and a world of **poorly documented APIs**
- **Everyone is a user, sys admin, developer**
- 90% of all security breaches can be traced to some kind of **human failure** - lack knowledge, time, motivation

The Big Picture - II

- **AI is compounding the complexity** of these challenges …**but just like humans it can also be part of the solution**…
- AI can help speed up the **detection of attacks**; it can help make cars more **secure**; it can help **authenticate** people; it can help us **manage our privacy** and much more
- …but it can also introduce its own **vulnerabilities** (e.g. vulnerable ML models)
- …and can also further add to **people's confusion** (e.g. lack of transparency leads to lack of trust)
- …and it can also be used for **malicious purposes** (e.g., deep fakes, social media manipulation, automated attacks)

Combining Human and Artificial Intelligence

-Developing solutions that effectively combine the strengths of both human and artificial intelligence requires:

-Developing a **deeper understanding and better modeling of the strengths and weaknesses of both humans and AI**

- What are people realistically capable of doing and how we can best help them
- What are AI systems capable of doing, what are their limitations, how can we configure them to benefit from their capabilities without paying a price for their limitations? **What guarantees should we require in different contexts?**

-Training People to Understand the **evolving capabilities and limitations of AI** (e.g. to avoid falling for deep fake attacks)

-Increasing critical in security and privacy where the **stakes are increasingly higher - not just money but human lives and democracy**

Q&A

The **Usable Privacy Policy Project** and the **Personalized Privacy Assistant Project** involve collaborations with a number of individuals

More details at:

Usableprivacy.org

Privacyassistant.org

Explore.usableprivacy.org

iotprivacy.io

Selection of References

- Janice. Tsai, S. Egelman, L. Cranor, A. Acquisti, "**The effect of online privacy information on purchasing behavior: An experimental study**," Information Systems Research, 22 (2), 2010
- Michael Benisch, Patrick Gage Kelley, Norman Sadeh, Lorrie Faith Cranor. Capturing Location Privacy Preferences: Quantifying Accuracy and User Burden Tradeoffs. *Journal of Personal and Ubiquitous Computing,* 2011.
- Zhang, Y Feng, L Bauer, LF Cranor, A Das, and N Sadeh, "**Did you know this camera tracks your mood?": Understanding Privacy Expectations and Preferences in the Age of Video Analytics**", Proceedings on Privacy Enhancing Technologies, 2, 1, Apr 2021
- Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Faith Cranor, Shomir Wilson, Florian Schaub, Norman Sadeh, "**Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text**", WWW '20, Apr 2020
- Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh, "**MAPS: Scaling Privacy Compliance Analysis to a Million Apps**", Privacy Enhancing Technologies Symposium (PETS 2019), 3, Jul 2019
- H. Almuhimedi, F. Schaub, N. Sadeh, Y. Agarwal, A. Acquisti, I. Adjerid, J. Gluck, L. Cranor, "**Your Location Has Been Shared 5398 Times! A Field Study on Mobile Privacy Nudges**", in Proc. CHI 2015, Jul 2015

Selection of References - III

- Examples of ongoing research projects at CMU:
  - The Usable Privacy Policy Project: https://usableprivacy.org
  - The Privacy Assistant Project: https://privacyassistant.org