Does Defining Privacy Matter?

Ryan C. Getek, Ph.D. CISSP-ISSEP
National Security Agency (NSA)

Session ID: PNG-402

Session Classification: General Interest

RSACONFERENCE2012

Disclaimer

- The views expressed in this talk are those of the presenter only, and do not necessarily reflect the opinions of NSA, RSA, or any other person, entity, or acronym

Background

Definitions

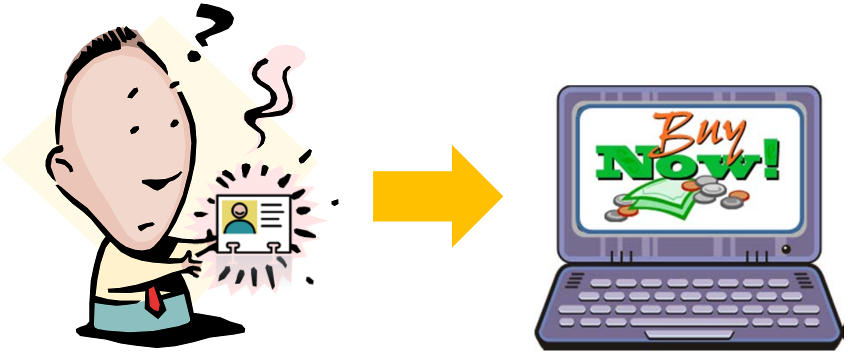
- Personalization (Kim, 2002)
  - Services that "…deliver information that is relevant to an individual or a group of individuals in the format and layout specified and in time intervals specified."
- Privacy (Agre & Rotenberg, 1998)
  - "the capacity to negotiate social relationships by controlling access to information about oneself."
- Security (U.S.Code, 2006)
  - "…unauthorized access, use, disclosure, disruption, modification, or destruction…" of a system or data

Privacy Views From Web Industry Leaders

- Mark Zuckerberg (Guardian, Jan 2010)
  - “People have really gotten comfortable not only sharing more information and different kinds, but more openly and with more people," he said. "That social norm is just something that has evolved over time.”
- Eric Schmidt (cnbc, Dec 2009)
  - “Every time we do anything that would use your information, we think hard about whether you would give us permission, should you give us permission, how could you give us permission, how could you opt out, that sorts of things.”

Data Collection Practices

- Email contents searched for personalized advertising
- Web sites use IP address geolocation
- Personalized news based on data mining
  - Click patterns, time spent on each page, data from other sources such as email, stated preferences, etc.
- Data from various sources for ad purposes



Environmental Assumptions

- U.S. versus European perspectives
- Past studies have shown a difference between privacy preferences and practice
- Even among those who have strong privacy preferences, the actual definition for privacy is either vague or varies
  - And may be defined mostly by academics
- Web service providers may (or may not) have good intent, but how important is intent?
- Web service providers are receptive to feedback

Questions of Interest

- What is privacy, and does the definition inform web hosts on data collection and use practices?
- Are users aware the associated data collection is occurring?
- When disclosed in privacy policies, how clear are the practices to ‘average’ users?
- What types of data are considered private?
- Why is there a divergence between privacy preference and practice?

Ethical Questions

- Three branches of ethics (Mason et. al)
  - Deontological theories
    - Innate right/wrongness, without regard for consequences
  - Teleological theories
    - Take consequences into account
  - Virtue ethics
    - Were motivations virtuous (focuses on intent)
- Is intent enough, or do we have to show benefit?
- Food for thought
- “You can make money without doing evil” (Google philosophy)

Study Methodology

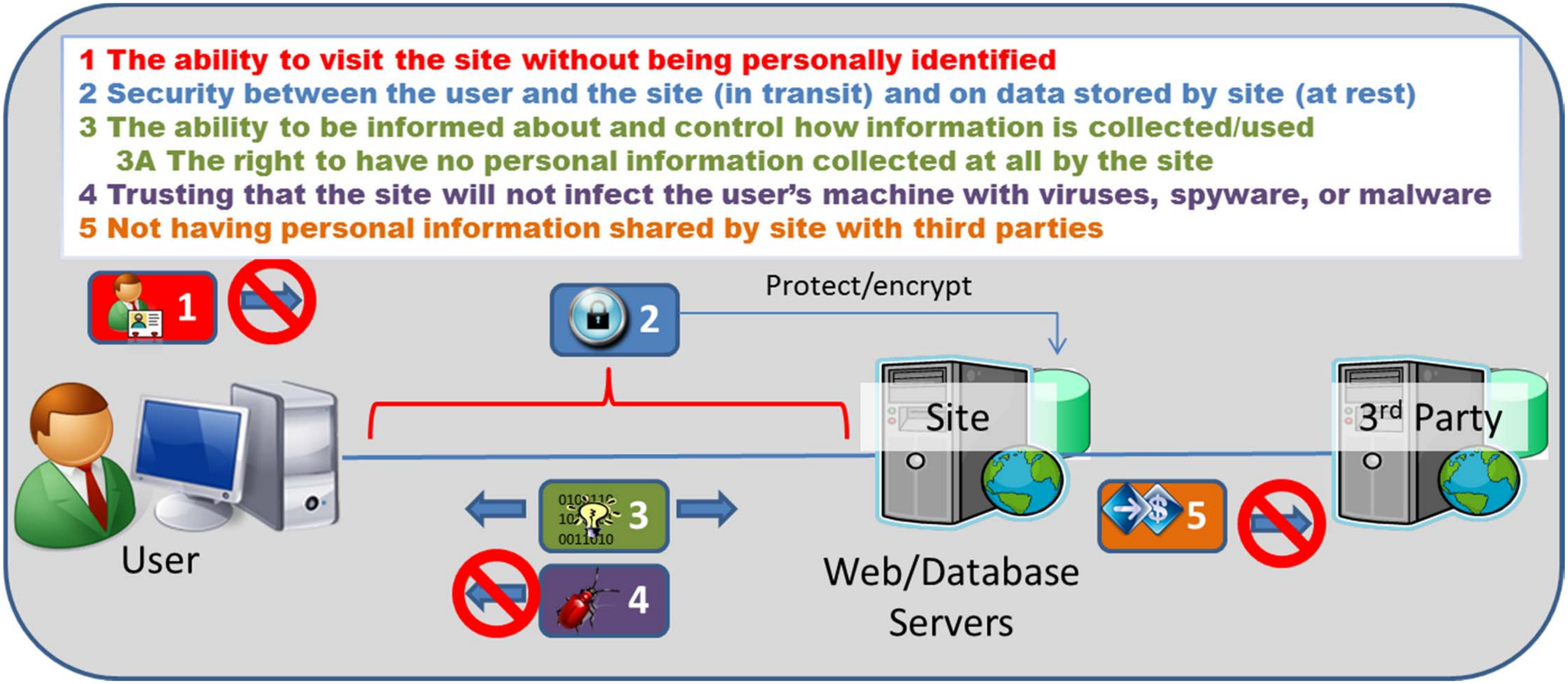
- Multi-phased web-based survey
- Qualitative, then quantitative methodology
  - Open-ended results codified into Likert scale questions
- Rank-based (non-parametric) analysis techniques applied
  - Sign test, rho, categorical PCA
  - Not going to be talking much more about the techniques, just the data!

Analyzing the Results

Updated Privacy Definition

- “the ability to control and remain informed about what information is collected, how it is collected and stored, how it is used and shared by the collector, how it is associated with the user’s identity while using the Internet (including the right to anonymity), and assurance that industry-standard security practices will be applied for data collection and storage.”
  - OK, it’s not exactly succinct
  - There’s an easier way to conceptualize it…

Updated Privacy Definition



1 The ability to visit the site without being personally identified
2 Security between the user and the site (in transit) and on data stored by site (at rest)
3 The ability to be informed about and control how information is collected/used
3A The right to have no personal information collected at all by the site
4 Trusting that the site will not infect the user's machine with viruses, spyware, or malware
5 Not having personal information shared by site with third parties

- Also, that there is 'no such thing' as privacy

Updated Privacy Definition

- Applying the definition…
  - Are users informed of common data collection practices?
  - Do users have knowledge of or choice about how the data is collected or stored?
  - Do users have knowledge of or a choice about how the data is used or shared?
  - Do users have knowledge of or a choice about security mechanisms used to protect the data?
- Even if users can't specify what a site does, knowledge allows them to 'vote with their feet'

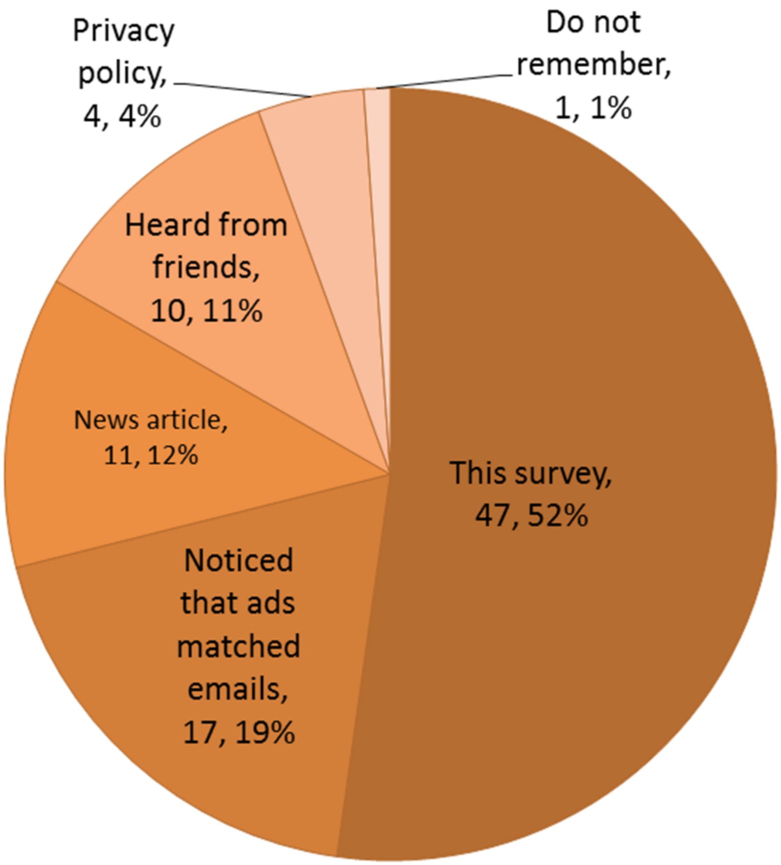
Data Collection Practice Awareness

- Google
  - Google has dozens of privacy policy pages
  - Over 81% of Google users had not visited any of the privacy policies or couldn’t remember having done so
    - 61%+ had not visited, 20% could not remember visiting
- MSN IP address geolocation: 27 (20.5%) aware
- MSN fuller data collection and use practices
  - Links clicked, search history in Bing, friends info through messenger, demographics provided at signup
  - 17 (8.8%) were fully aware
  - 72 (36.4%) were only aware of some of these practices
  - 105 (53.5%) were not aware of any of these practices

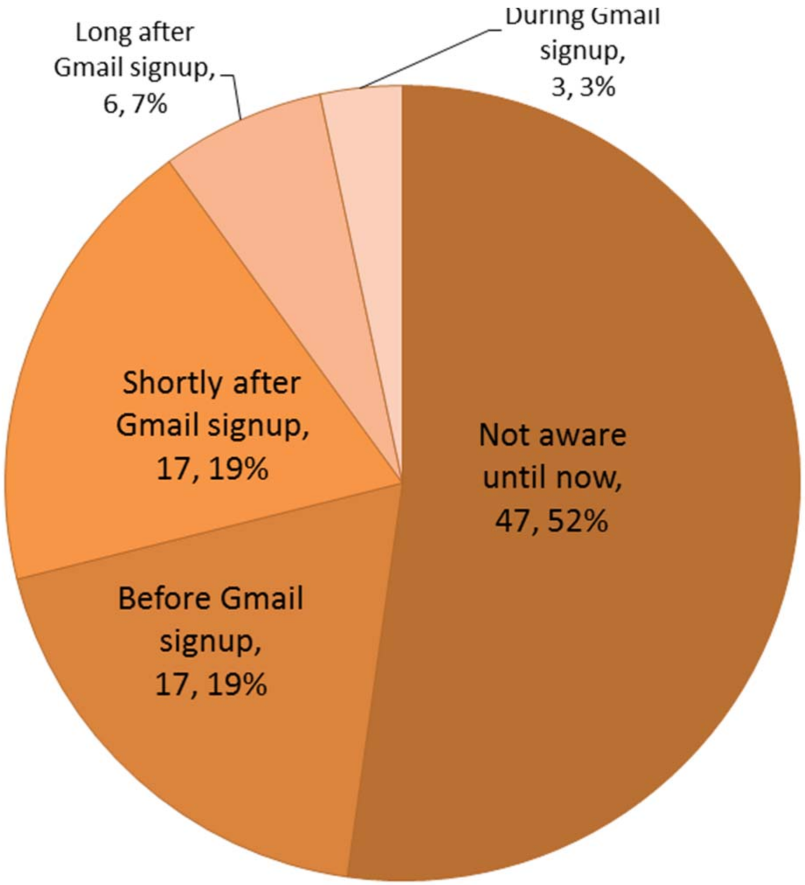
Data Collection Practices in Perspective

- Eric Schmidt (cnbc, Dec 2009)
  - "...What they really do is they put it into their email, and their email is, in fact, the number one source because people forget that when they type those emails, they are kept in servers...“
  - “…If you have something you don't want anyone to know, maybe you shouldn't be doing it in the first place…“
- So how does privacy (as defined here) stack up in this modern data paradigm?

Data Collection Practice Awareness



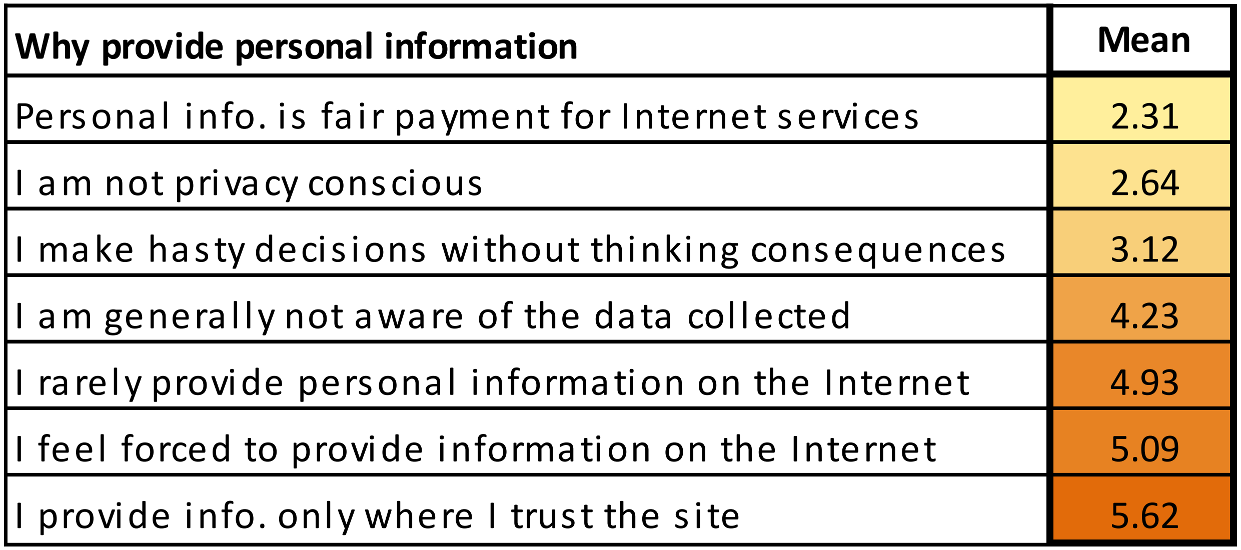
How made aware of Gmail content searching practices



When made aware of Gmail content searching practices

User Knowledge

- When users know about data collection practices, why do they provide the data?
  - “Because there are no, realistic, alternatives short of stopping from using the internet entirely. … I don't bother looking at the privacy policy. No need to sicken myself with the extent of lack of privacy.”



Why provide personal information	Mean
Personal info. is fair payment for Internet services	2.31
I am not privacy conscious	2.64
I make hasty decisions without thinking consequences	3.12
I am generally not aware of the data collected	4.23
I rarely provide personal information on the Internet	4.93
I feel forced to provide information on the Internet	5.09
I provide info. only where I trust the site	5.62

User Knowledge

- How much can the average user really understand about how the information is stored and used?
- Google may employ some of the following (Das et. al)
  - Collaborative filtering using MinHash clustering
  - Probabilistic Latent Semantic Indexing (PLSI)
  - Covisitation counts
- What about competitive advantage?

Privacy Policies

- What should a privacy policy contain?

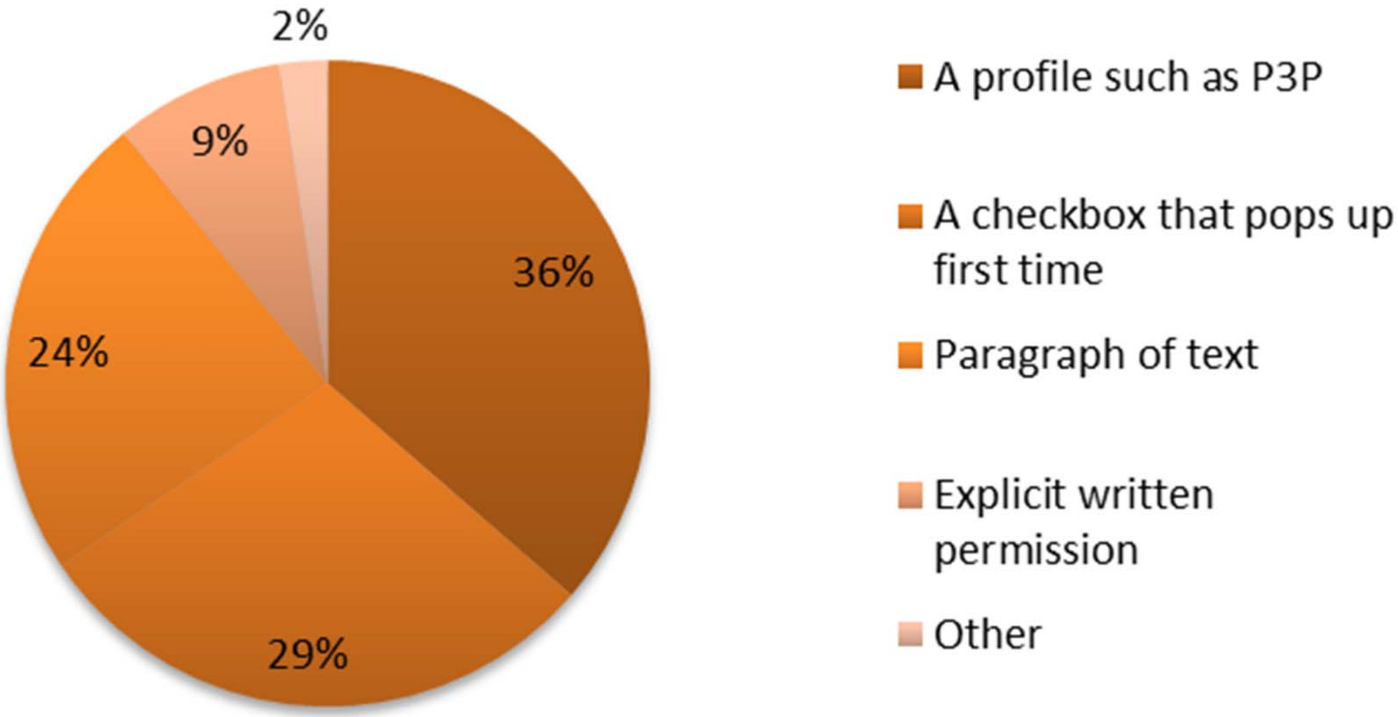
Privacy Policy Components	Mean
Contact information	5.620
How long the data will be kept	5.640
Security features	6.090
Types of data collected	6.150
How data will be used	6.240
How data will be shared	6.430

- Privacy policies as an informational vehicle

Increased data collection disclosure areas	Mean
Privacy policy	6.02
In a popup message	6.14
As part of the web page	6.14
By email	6.39

Privacy Policies

- So, how do users want to be notified of privacy practices?



Supplemental Findings

Data Type Privacy

- Categorizing data types
  - Principal Components Analysis (PCA)

Data Type	Mean
Screen Size	2.01
Connection Speed	2.35
Browser Type	2.70
Gender	3.19
Age	3.68
Hobbies/Interests	3.99
Location: ZIP Code	4.10
Shopping Habits	4.51
Web History	4.95
IP Address	5.16
Email Address	5.45
Income	5.72
Location: Address	6.26
Email Contents	6.44
Hard Drive Contents	6.64
SSN	6.75

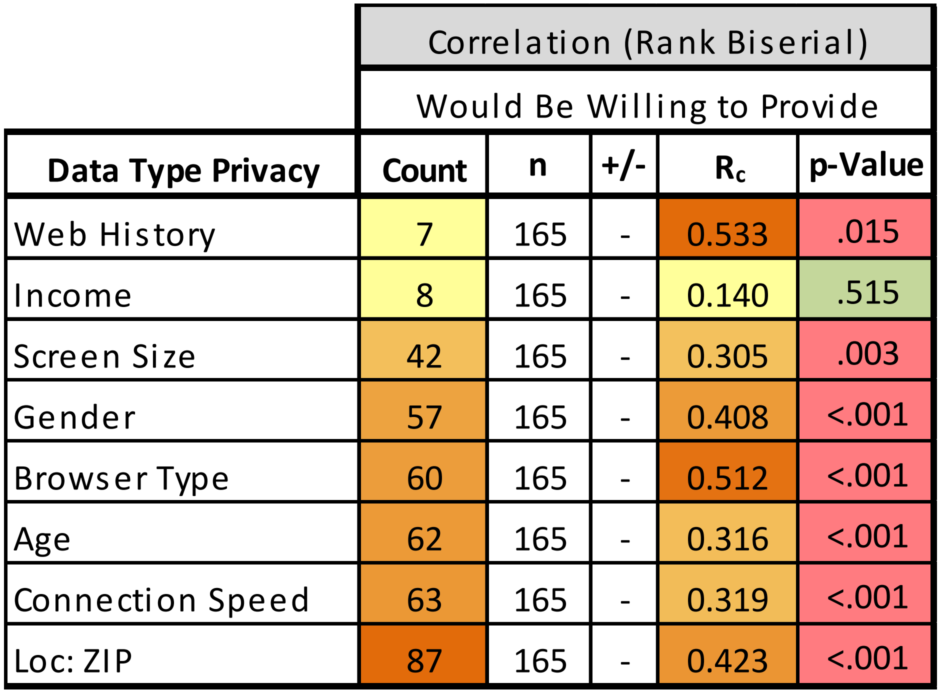
	Component Loadings	
	Dimension	
	1	2
ScreenSize	-0.180	**0.740**
InternetConn	0.069	**0.693**
Browser	-0.080	**0.795**
Gender	0.286	**0.659**
Age	0.443	**0.612**
HobbiesInterests	0.440	**0.543**
LocZIP	0.378	**0.429**
ShoppingHabits	**0.571**	0.257
WebHistory	**0.675**	0.215
IpAddr	**0.603**	0.087
EmailAddr	**0.577**	-0.020
Income	**0.671**	0.081
LocStreetAddr	**0.650**	-0.108
EmailContents	**0.719**	-0.461
HardDrive	**0.687**	-0.434
SSN	**0.642**	-0.622

Semi-Personal

Personal

Data Type Privacy

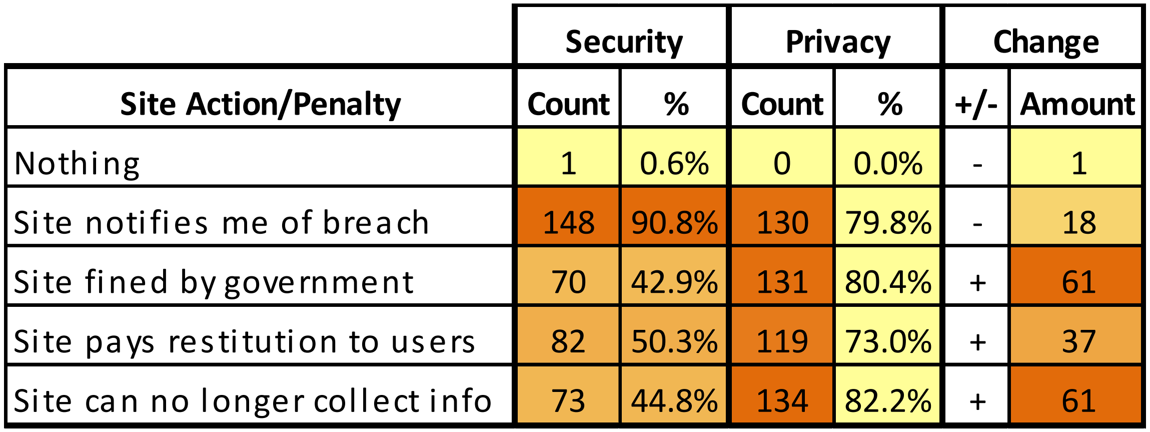
- Data type privacy relationship with willingness to provide to a site



	Correlation (Rank Biserial)				
	Would Be Willing to Provide				
Data Type Privacy	**Count**	**n**	**+/-**	**R_c**	**p-Value**
Web History	7	165	-	0.533	.015
Income	8	165	-	0.140	.515
Screen Size	42	165	-	0.305	.003
Gender	57	165	-	0.408	<.001
Browser Type	60	165	-	0.512	<.001
Age	62	165	-	0.316	<.001
Connection Speed	63	165	-	0.319	<.001
Loc: ZIP	87	165	-	0.423	<.001

- Trust and privacy concern inversely correlated

Privacy Vs. Security



	Security		Privacy		Change	
Site Action/Penalty	Count	%	Count	%	+/-	Amount
Nothing	1	0.6%	0	0.0%	-	1
Site notifies me of breach	148	90.8%	130	79.8%	-	18
Site fined by government	70	42.9%	131	80.4%	+	61
Site pays restitution to users	82	50.3%	119	73.0%	+	37
Site can no longer collect info	73	44.8%	134	82.2%	+	61

- Intent matters to users!
- Unauthorized disclosure is not the ‘bottom line’
- Users want sites to be held responsible
- Failure to honor privacy policy is likely to cost a site some business (if disclosed)

Applying the Conclusions

Does Defining Privacy Matter?

- Short answer is… Absolutely!
  - Users often define privacy differently than academia or big business
  - Informs behavior of data collector
  - Allows creation of metrics
  - Helps explain gap between stated behavior and practice
  - Identifies gaps in content of existing privacy policies
  - Shows areas where concerns exist but data cannot be shared
    - Privacy policies have traditionally focused on data that is commonly shared…

Apply: Web Service Providers

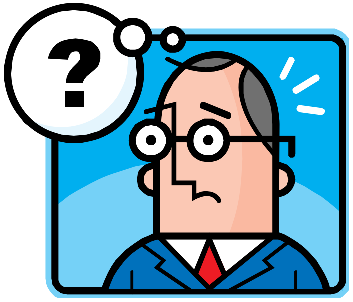
- Key points of privacy policy should be stated clearly and succinctly (such as in bullets)
  - Especially important for ‘trusted’ sites
- Describe how data is:
  - collected
  - used (tricky to describe)
  - shared
  - protected
- Not just in privacy policies
  - Plain language when user enrolls in service

Apply: Consumers

- Carefully consider your ‘line in the sand’
- Vote with your feet!
- Read the privacy policies
- Use tools such as P3P
- Communicate with the service provider
- Don’t give up!

Summary

- Ethical considerations are not the whole story…
- Ethics of privacy
  - Mean well, do good, consequences ‘good’
- Practical considerations of privacy
  - Many users uninformed
    - If not informed, how can they choose?
  - The details are left vague
    - If users were told, would they understand?
  - The definition and research show serious gaps



References

- Kim, W. (2002). Personalization: Definition, Status and Challenges Ahead. Journal of Object Technology, 1(1), 29-40 http://www.jot.fm/issues/issue_2002_2005/column2003.
- Agre, P. E., & Rotenberg, M. (1998). Technology and Privacy: The New Landscape. Cambridge, MA: MIT Press.
- Information Security: 44 USC Sec. 3542 (2006).
- http://www.guardian.co.uk/technology/2010/jan/11/facebook-privacy
- http://video.cnbc.com/gallery/?video=1372176413
- Mason, R. O., Mason, F. M., & Culnan, M. J. (1995). Ethics of Information Management: Thousand Oaks: Sage.
- http://www.google.com/about/corporate/company/tenthings.html
- Das, A.S., Datar, M., Garg, A., Rajaram S. Google news personalization: scalable online collaborative filtering WWW '07: Proceedings of the 16th international conference on World Wide Web