Implementation and Evaluation of Privacy-Preserving Protocols

Thesis Defense

Felipe Saint-Jean

Department of Computer Science

Yale University

July 21, 2010

Acknowledgment: NSF, ONR, IARPA

Overview

To help bridge the gap between theory and practice, we implement and evaluate protocols to enhance privacy in four important tasks:

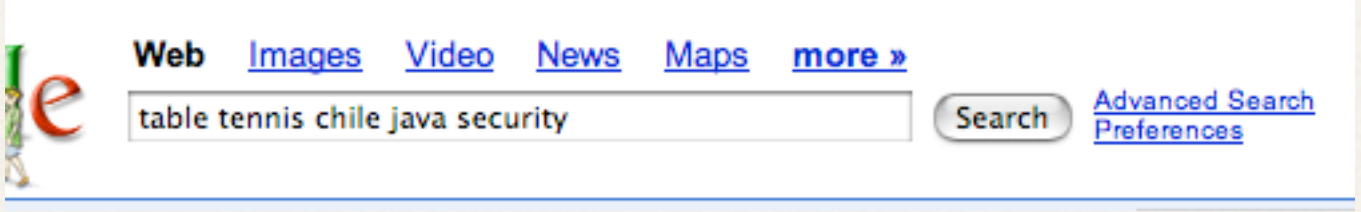
- Web search
- Survey computation
  - Database querying
  - Security-alert sharing

Privacy Issues in Web Search

- Web-search queries are sensitive data
- For example: Search history
  - "Table Tennis Tournament New York"
  - "Java reflection"
  - "Chilean bakery new york"
  - "named buffer overflow"

Privacy Issues in Web Search

- Web-search queries are sensitive data
- For example: Search history
  - "Table Tennis Tournament New York"
  - "Java reflection"
  - "Chilean bakery new york"
  - "named buffer overflow"



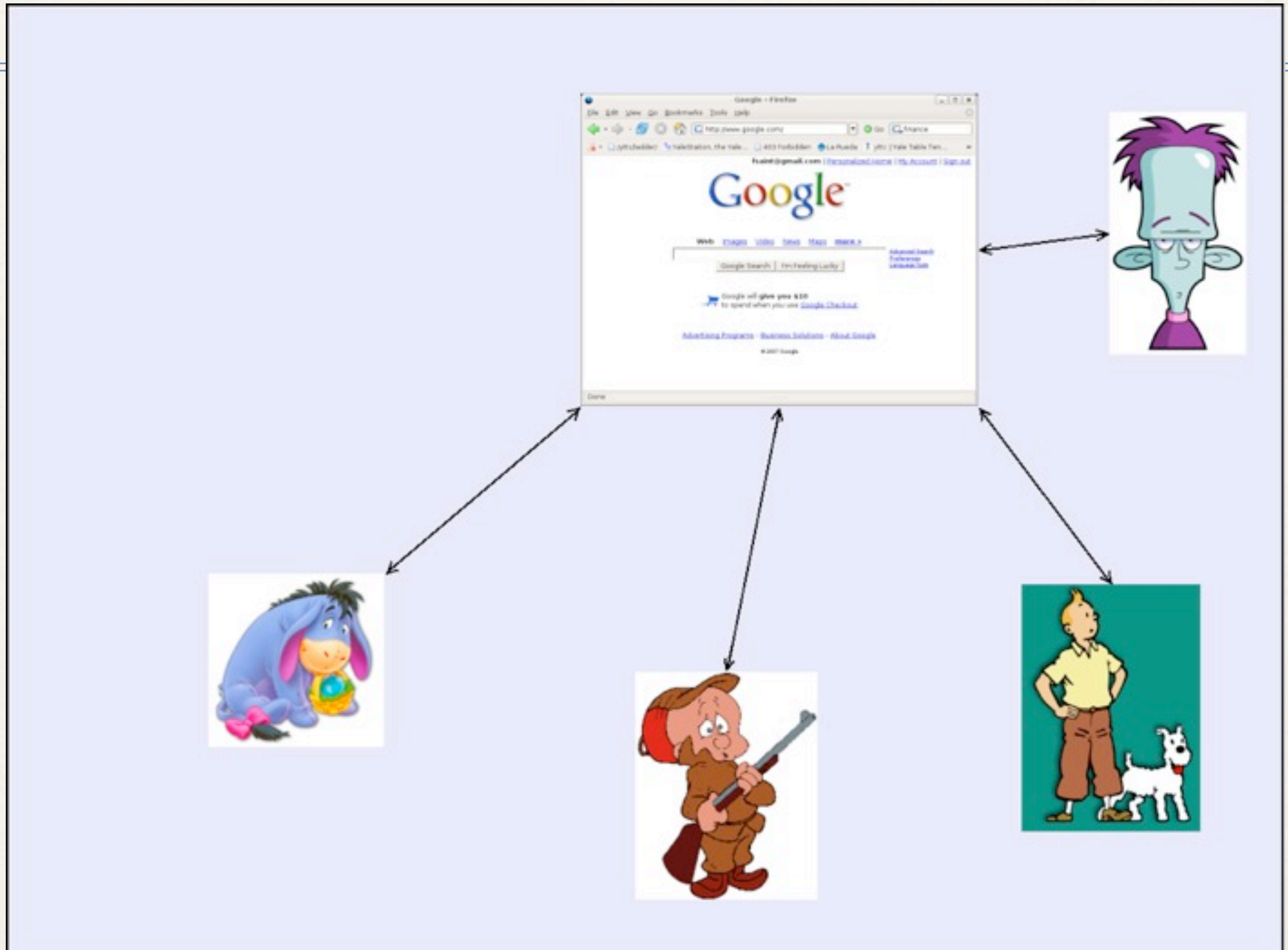
Privacy Issues in Web Search

- Web-search queries are sensitive data
- For example: Search history
  - "Table Tennis Tournament New York"
  - "Java reflection"
  - "Chilean bakery new york"
  - "named buffer overflow"

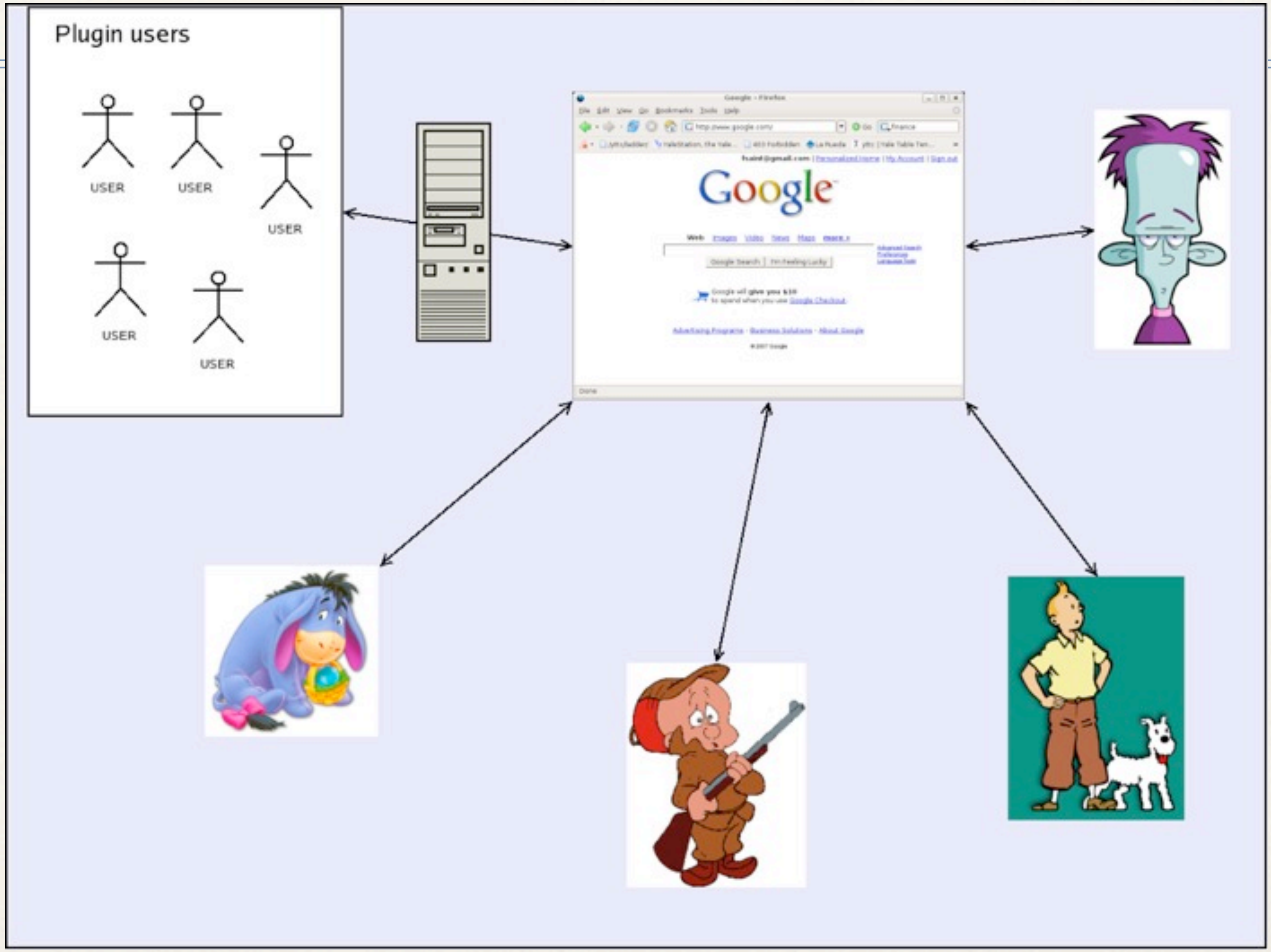
Privacy Issues in Web Search (2)

- What information does the search engine collect?
  - TCP/IP: IP address, Institution or ISP, OS, uptime
  - HTTP Headers: Cookies, Operating system and OS version, Browser make and version, Encodings and language
  - HTML: JavaScript collected information, Timing information, Query terms and time
  - Active components: ...
- ...

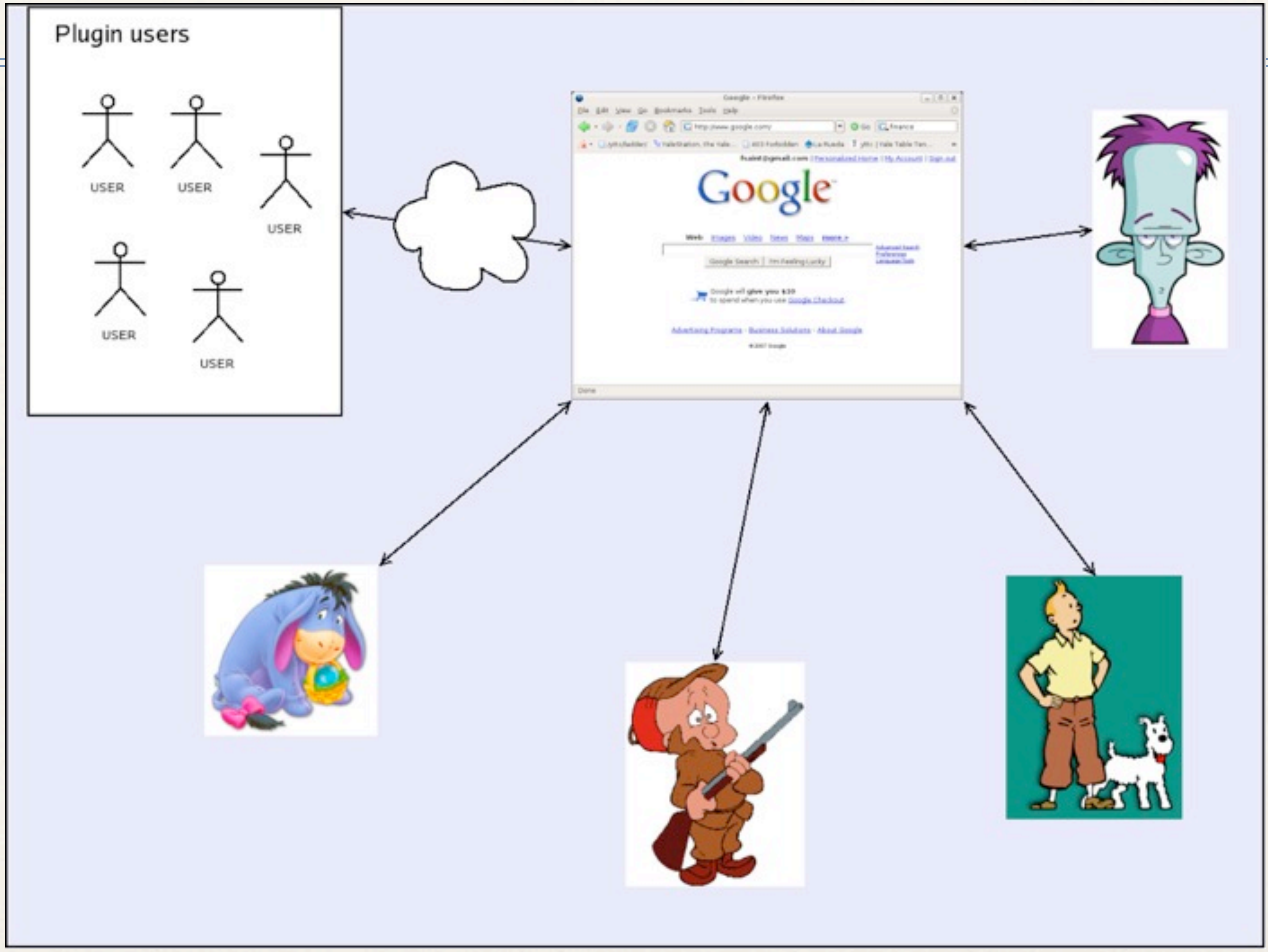
Objective



Objective



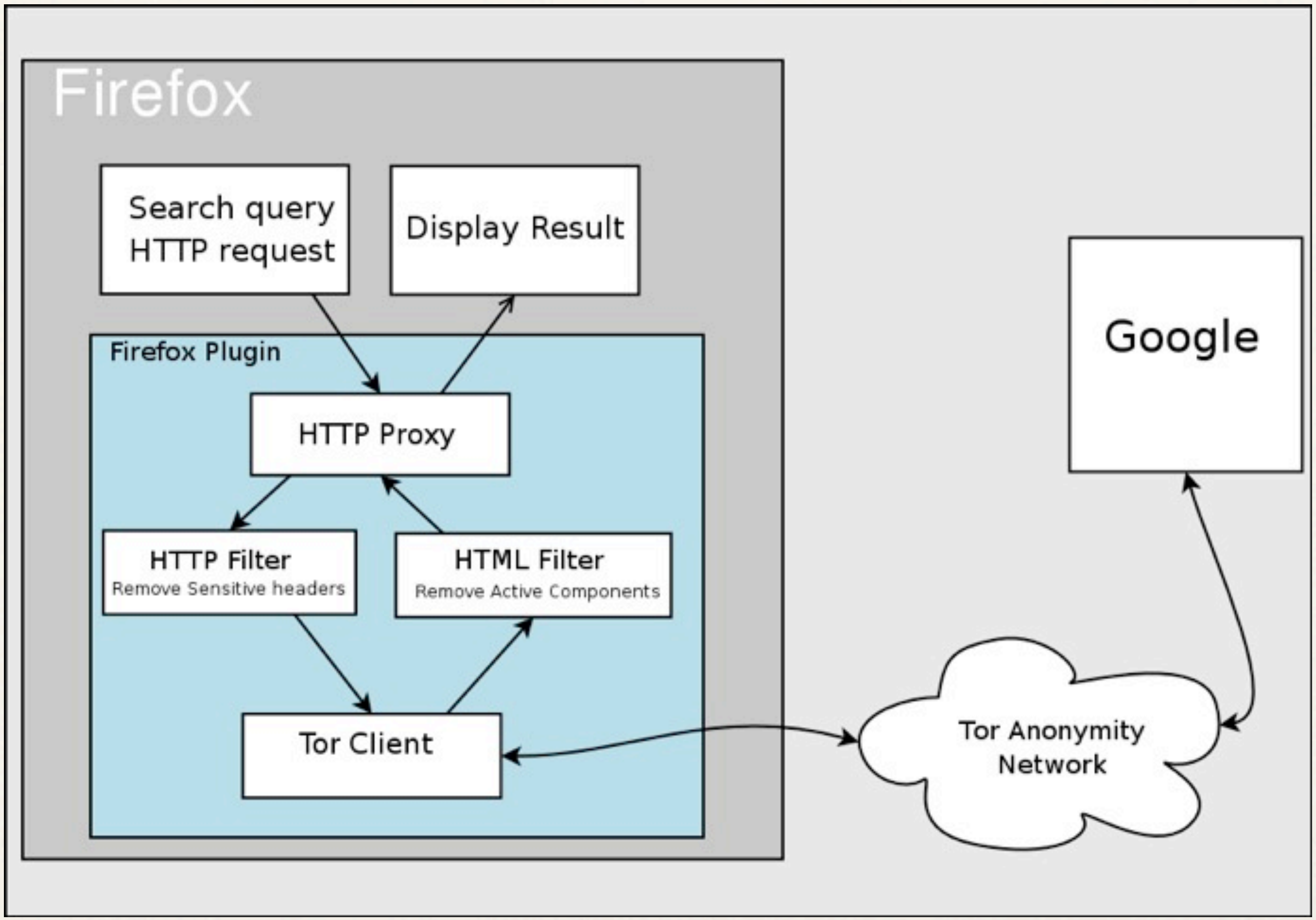
Objective



Objective (2)

- We seek to obfuscate the relationship between queries and the users who issued them.
- We are *not* tackling the harder problem of hiding the content of the query.

PWS: A *Search-Specific*, Privacy-Preserving Firefox Plugin



How each type of information is handled

- TCP/IP
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML
  - JavaScript collected information
  - Timing information
- Query terms and time
- Active components
  - …

How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML
  - JavaScript collected information
  - Timing information
- Query terms and time
- Active components
  - …

How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML
  - JavaScript collected information
  - Timing information
- Query terms and time
- Active components
  - …

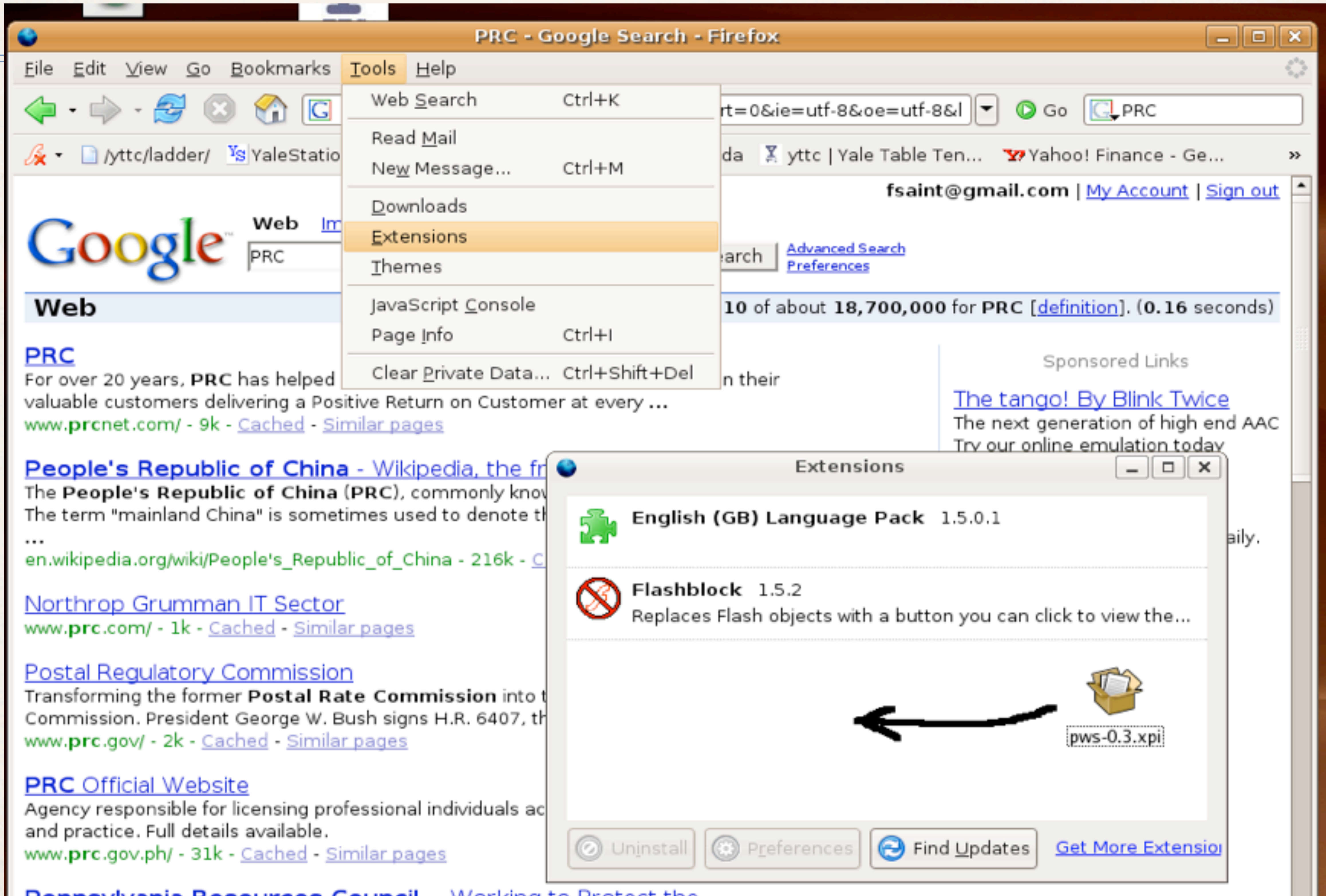
How each type of information is handled

- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time
- Active components ← HTML filter
  - …

How each type of information is handled

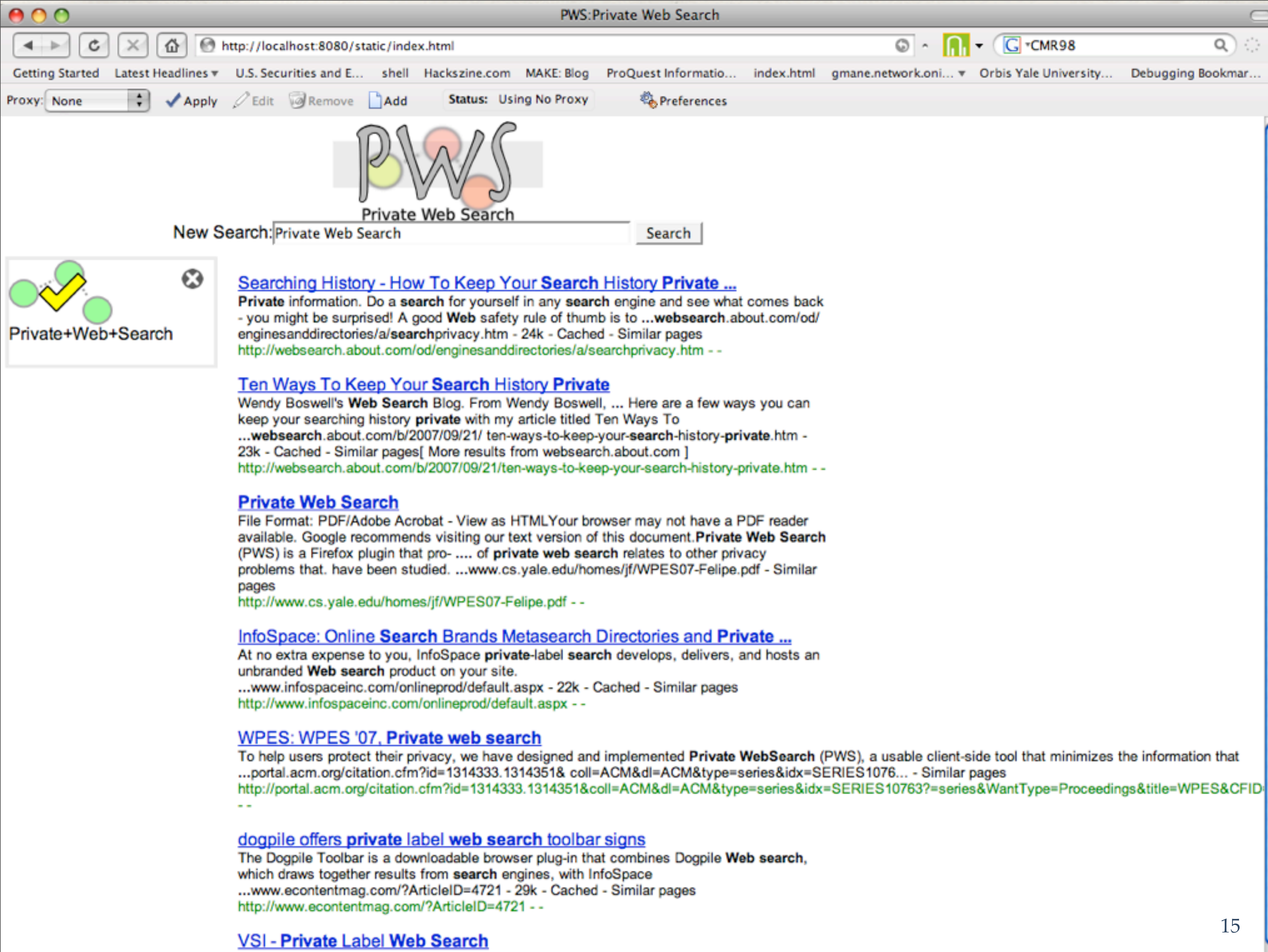
- TCP/IP ← Tor
  - IP address
  - Institution or ISP
  - Operating System
  - uptime
- HTTP Headers ← HTTP filter
  - Cookies
  - Operating system make and version
  - Browser make and version
  - Encodings and language
- HTML ← HTML filter
  - JavaScript collected information
  - Timing information
- Query terms and time ← Open Problem
- Active components ← HTML filter
  - …

Plugin installation



Plugin use

Google - Firefox
File Edit View Go Bookmarks Tools Help
http://www.google.com/
Go
privacy concerns
Google
Yahoo
Amazon.com
Answers.com
Creative Commons
Ubuntu Package Search
eBay
Wikipedia topic
PWS Google search
Wikipedia
Search PWS Google s
Web Images Video News Maps more »
Google Search
I'm Feeling Lucky


PWS:Private Web Search
http://localhost:8080/static/index.html
CMR98
Getting Started
Latest Headlines
U.S. Securities and E...
shell
Hackszine.com
MAKE: Blog
ProQuest Informatio...
index.html
gmane.network.oni...
Orbis Yale University...
Debugging Bookmar...
Proxy: None
Apply
Edit
Remove
Add
Status: Using No Proxy
Preferences
PWS
Private Web Search
New Search: Private Web Search
Search
Private+Web+Search
Searching History - How To Keep Your Search History Private ...
Private information. Do a search for yourself in any search engine and see what comes back - you might be surprised! A good Web safety rule of thumb is to ...websearch.about.com/od/ enginesanddirectories/a/searchprivacy.htm - 24k - Cached - Similar pages
http://websearch.about.com/od/enginesanddirectories/a/searchprivacy.htm - -
Ten Ways To Keep Your Search History Private
Wendy Boswell's Web Search Blog. From Wendy Boswell, ... Here are a few ways you can keep your searching history private with my article titled Ten Ways To ...websearch.about.com/b/2007/09/21/ ten-ways-to-keep-your-search-history-private.htm - 23k - Cached - Similar pages[More results from websearch.about.com]
http://websearch.about.com/b/2007/09/21/ten-ways-to-keep-your-search-history-private.htm - -
Private Web Search
File Format: PDF/Adobe Acrobat - View as HTMLYour browser may not have a PDF reader available. Google recommends visiting our text version of this document.Private Web Search (PWS) is a Firefox plugin that pro- .... of private web search relates to other privacy problems that. have been studied. ...www.cs.yale.edu/homes/jf/WPES07-Felipe.pdf - Similar pages
http://www.cs.yale.edu/homes/jf/WPES07-Felipe.pdf - -
InfoSpace: Online Search Brands Metasearch Directories and Private ...
At no extra expense to you, InfoSpace private-label search develops, delivers, and hosts an unbranded Web search product on your site. ...www.infospaceinc.com/onlineprod/default.aspx - 22k - Cached - Similar pages
http://www.infospaceinc.com/onlineprod/default.aspx - -
WPES: WPES '07, Private web search
To help users protect their privacy, we have designed and implemented Private WebSearch (PWS), a usable client-side tool that minimizes the information that ...portal.acm.org/citation.cfm?id=1314333.1314351& coll=ACM&dl=ACM&type=series&idx=SERIES1076... - Similar pages
http://portal.acm.org/citation.cfm?id=1314333.1314351&coll=ACM&dl=ACM&type=series&idx=SERIES10763?=series&WantType=Proceedings&title=WPES&CFID - -
dogpile offers private label web search toolbar signs
The Dogpile Toolbar is a downloadable browser plug-in that combines Dogpile Web search, which draws together results from search engines, with InfoSpace ...www.econtentmag.com/?ArticleID=4721 - 29k - Cached - Similar pages
http://www.econtentmag.com/?ArticleID=4721 - -
VSI - Private Label Web Search

User study of 39 members of the Yale community

- Compare PWS, TPTV, and Google
  - Ease of installation
  - Speed
  - Accuracy
- Users' attitudes towards privacy in web search and browser-based privacy tools

Study Design

- Required IRB authorization
- Randomly divide subjects into 3 groups:
  - PWS
  - Tor+Privoxy+TorButton+Vidalia (TPTV)
  - Google (no privacy enhancement)
- Each session lasted one hour.

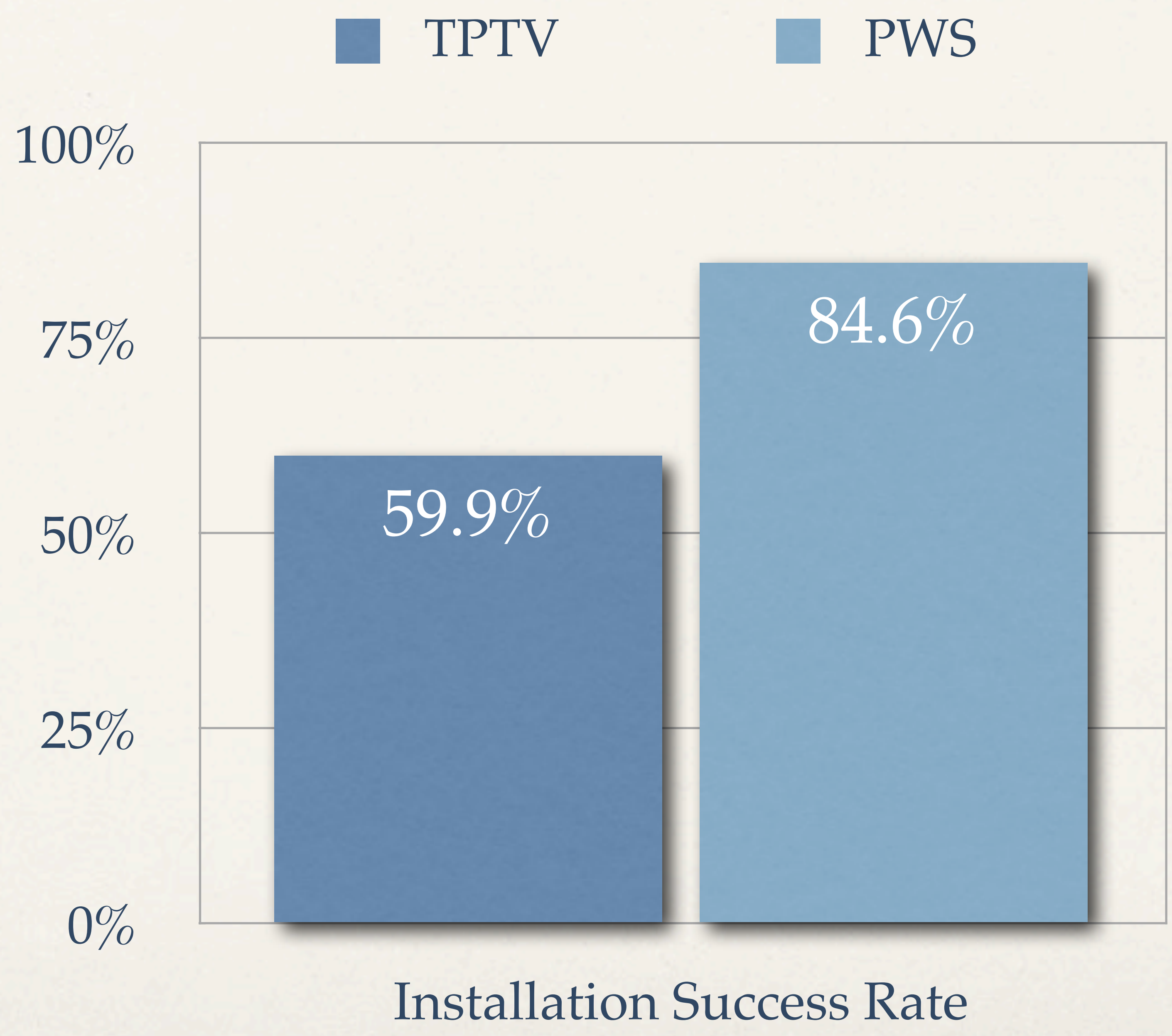
Study Phases

- Installation: Can the user successfully install his assigned privacy tool?
- Switching: Can the user successfully change between privacy-enhanced search and regular search?
- Effectiveness: How fast and how accurately can the user perform ``search tasks''?
- Survey: Poll users' preferences, beliefs, and practices with respect to web search privacy

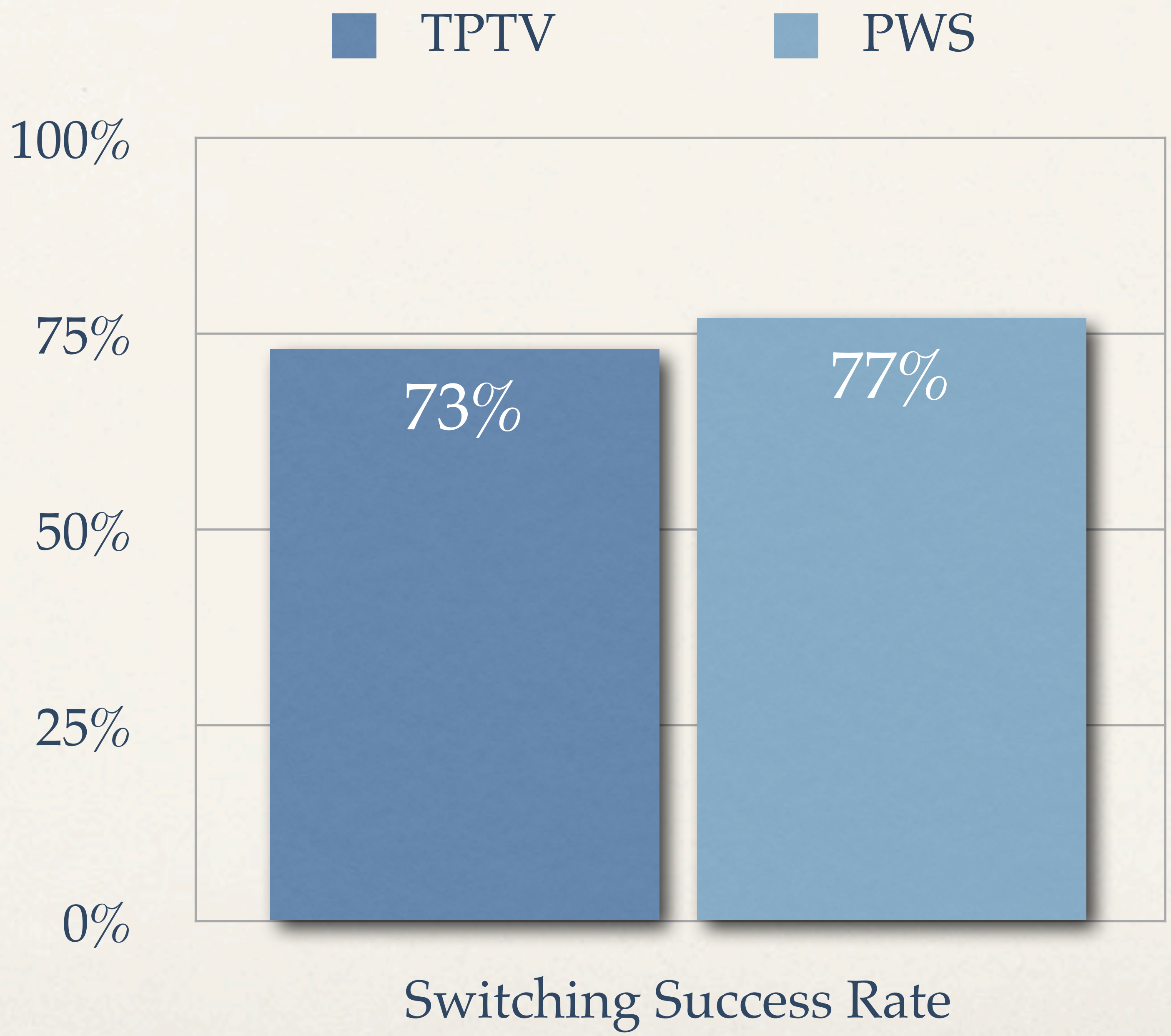
Search Task

- Question
  - A trivia question: e.g., "What is the name of the snowy owl that Hagrid bought for Harry Potter?"
  - A search method: PWS, TPTV, or plain Google
- Answer
  - The answer to the question: "Hedwig"
  - The URL where the answer was found: "http://www.lauraerickson.com/bird/Species/Owls/HarryPotter/HarryPotter.html"

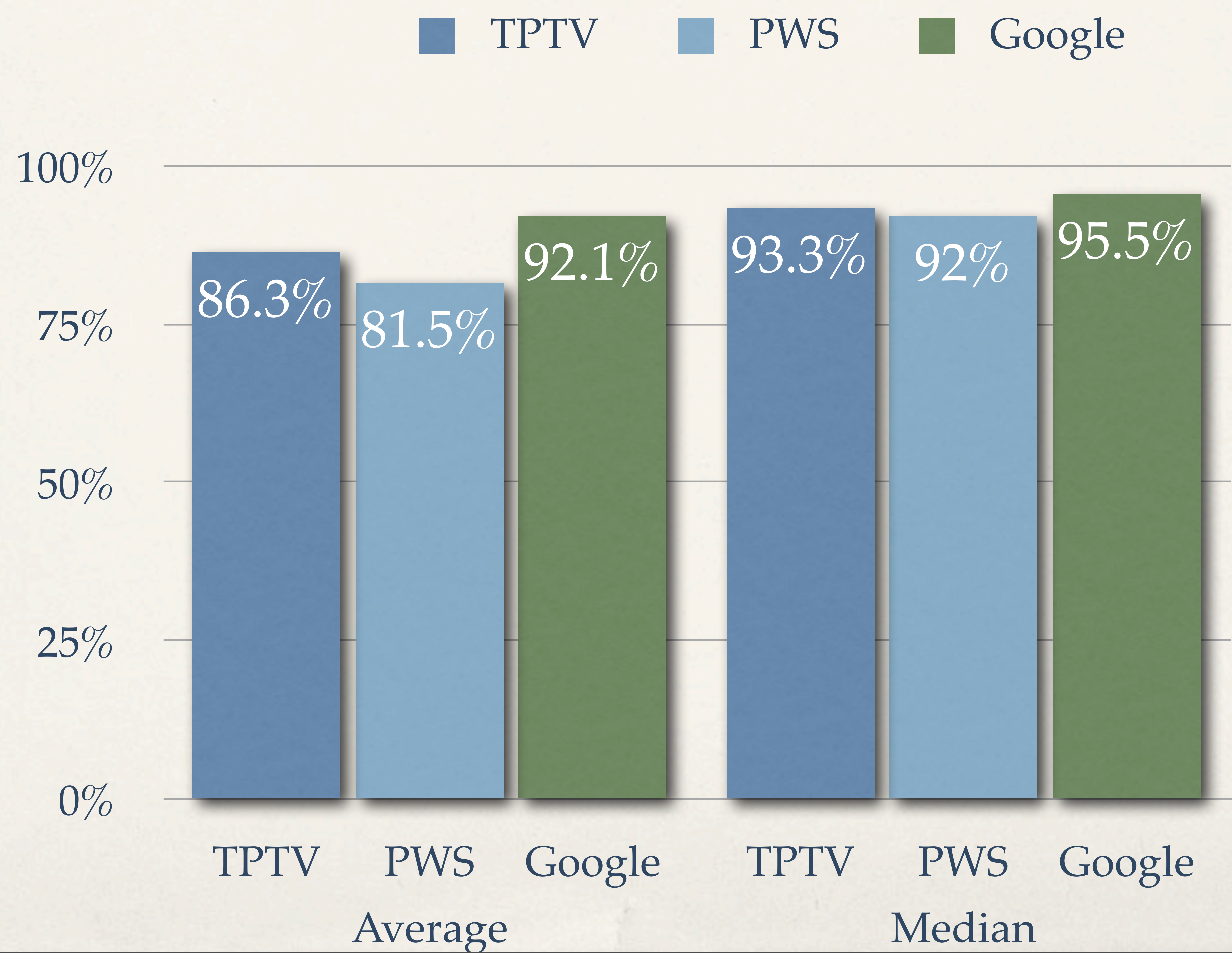
Installation



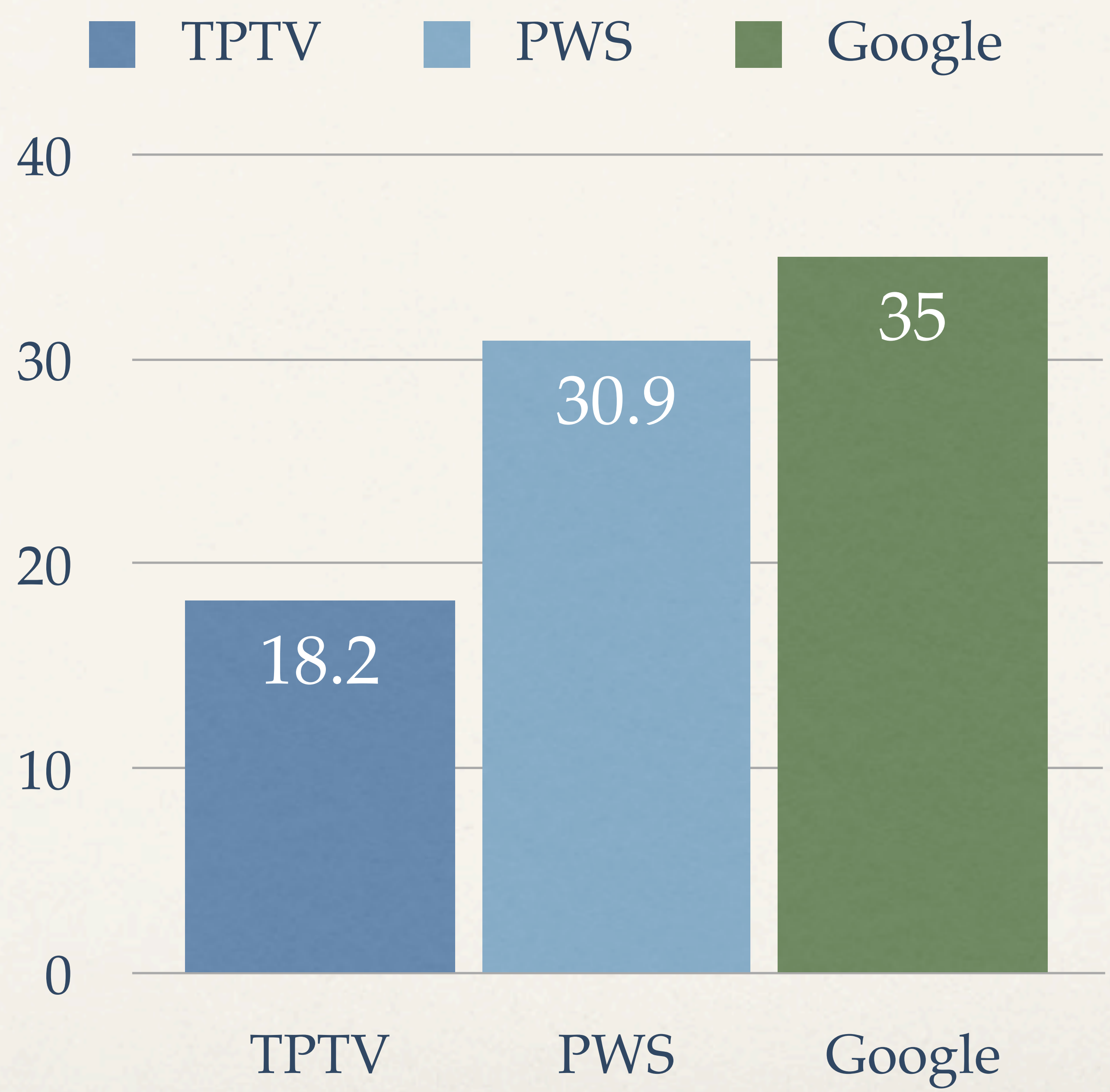
Switching



Accuracy

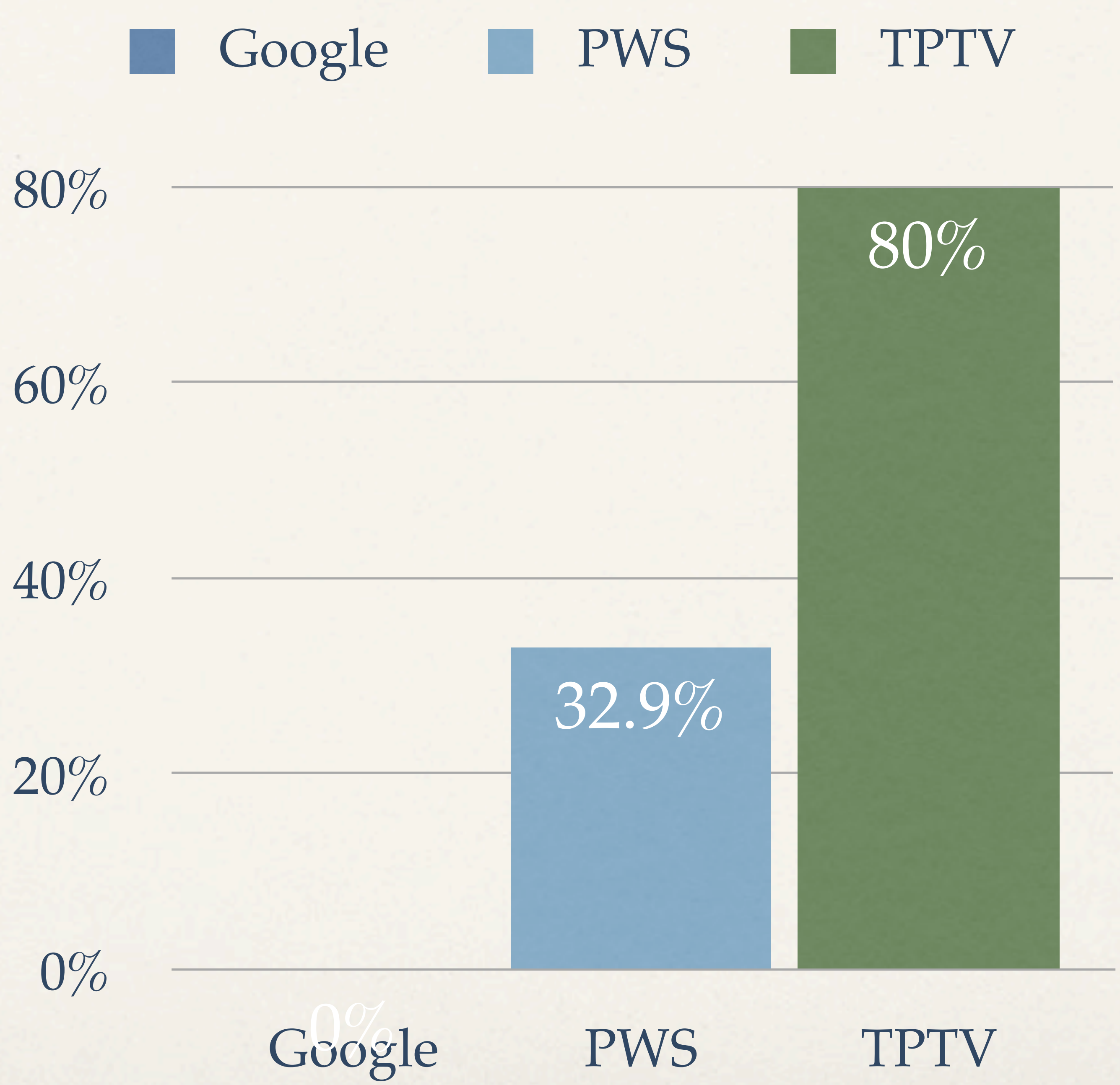

TPTV
PWS
Google
100%
75%
50%
25%
0%
86.3%
81.5%
92.1%
93.3%
92%
95.5%
TPTV
PWS
Google
TPTV
PWS
Google
Average
Median

Speed



Average number of search tasks completed

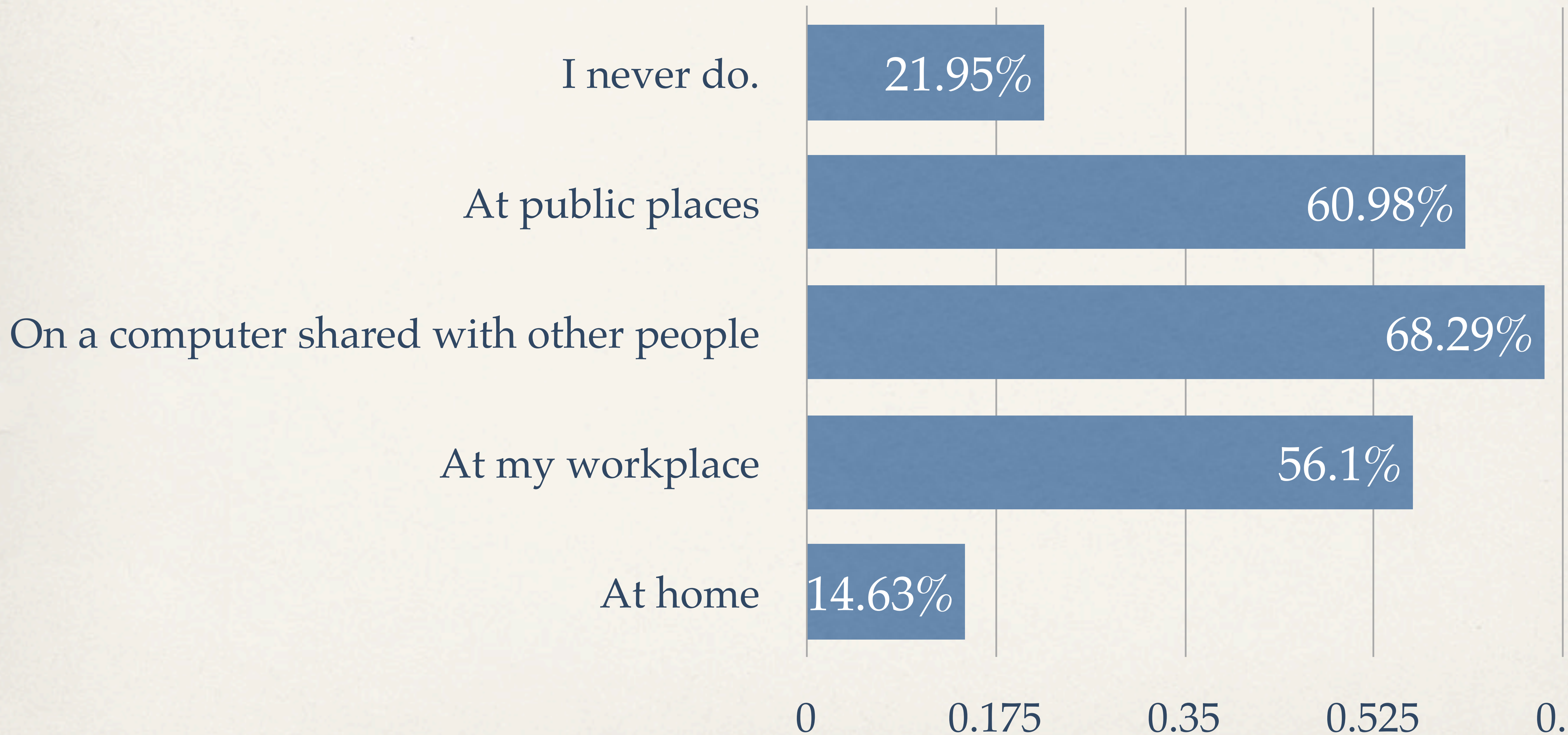
Need for Monitor's Intervention



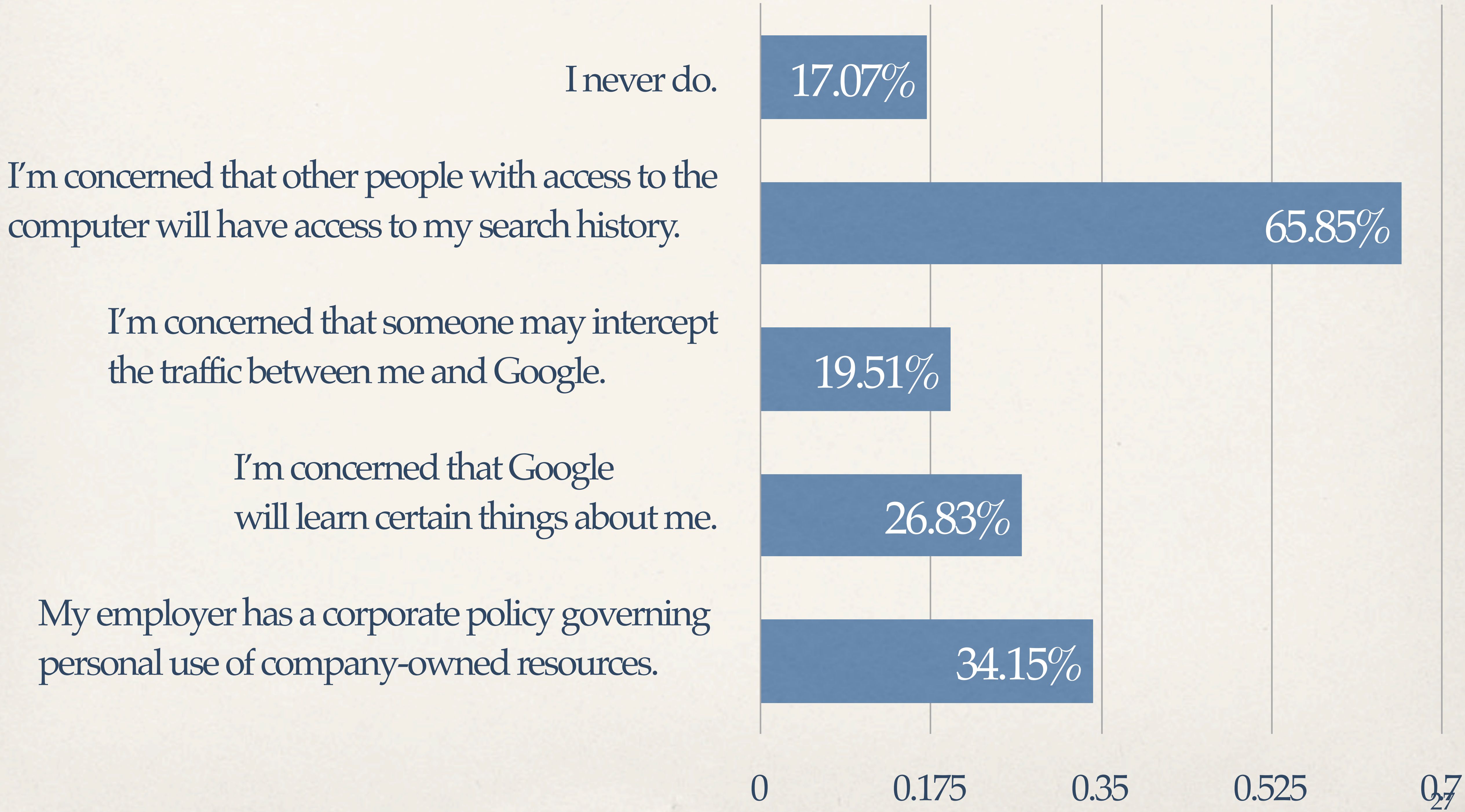
Unrecoverable Errors by TPTV Users

- 33.0% were faced with a Google page in a language they could not understand.
- 33.0% were told by Google that it could not answer queries because the user's machine was infected by spyware.
- 46.6% were not able to figure out how to activate TPTV after installing.

When using Google to search the Web, do you avoid certain topics (select all relevant answers):

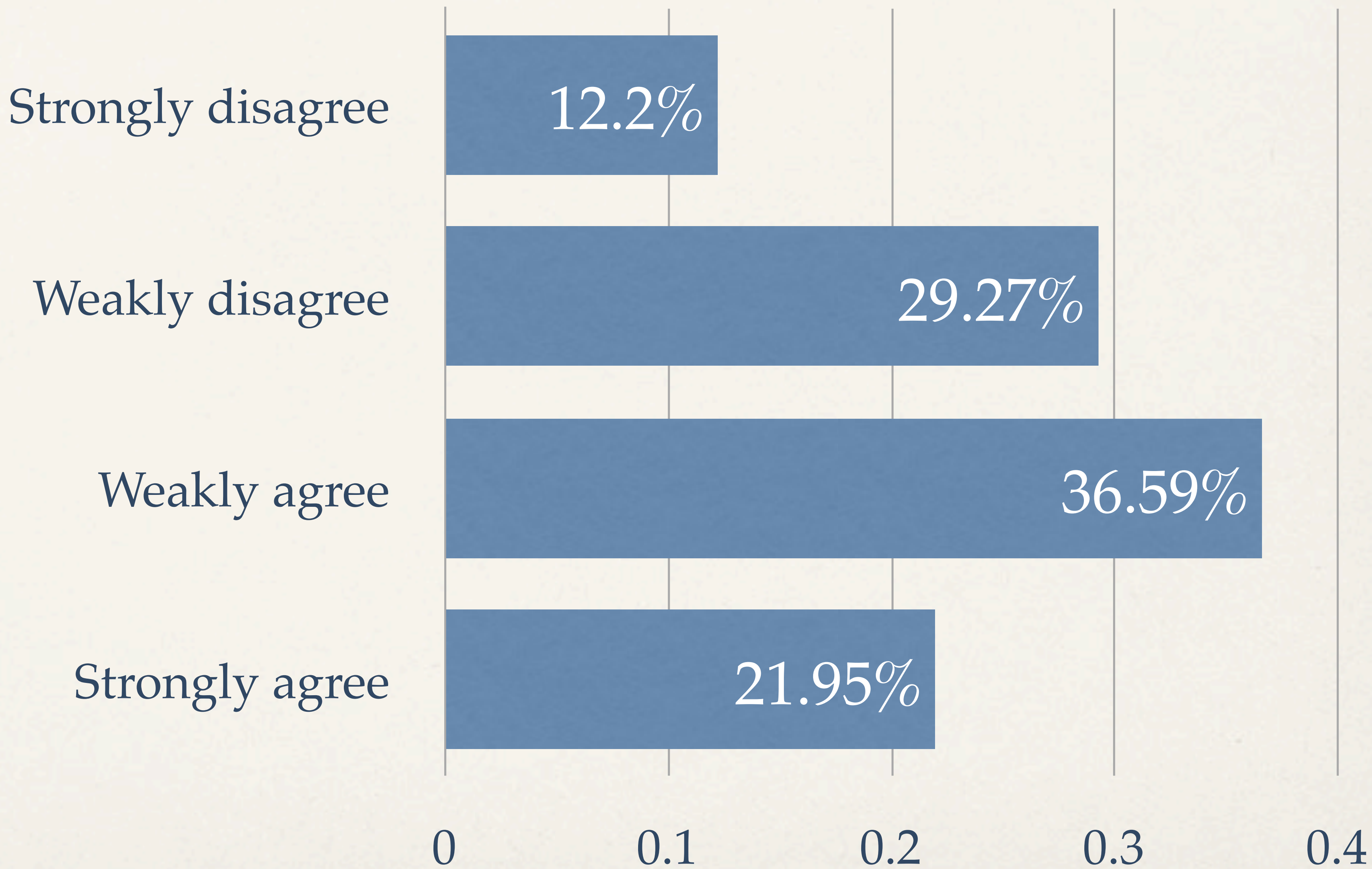


I avoid certain topics when using Google because (select all relevant answers):

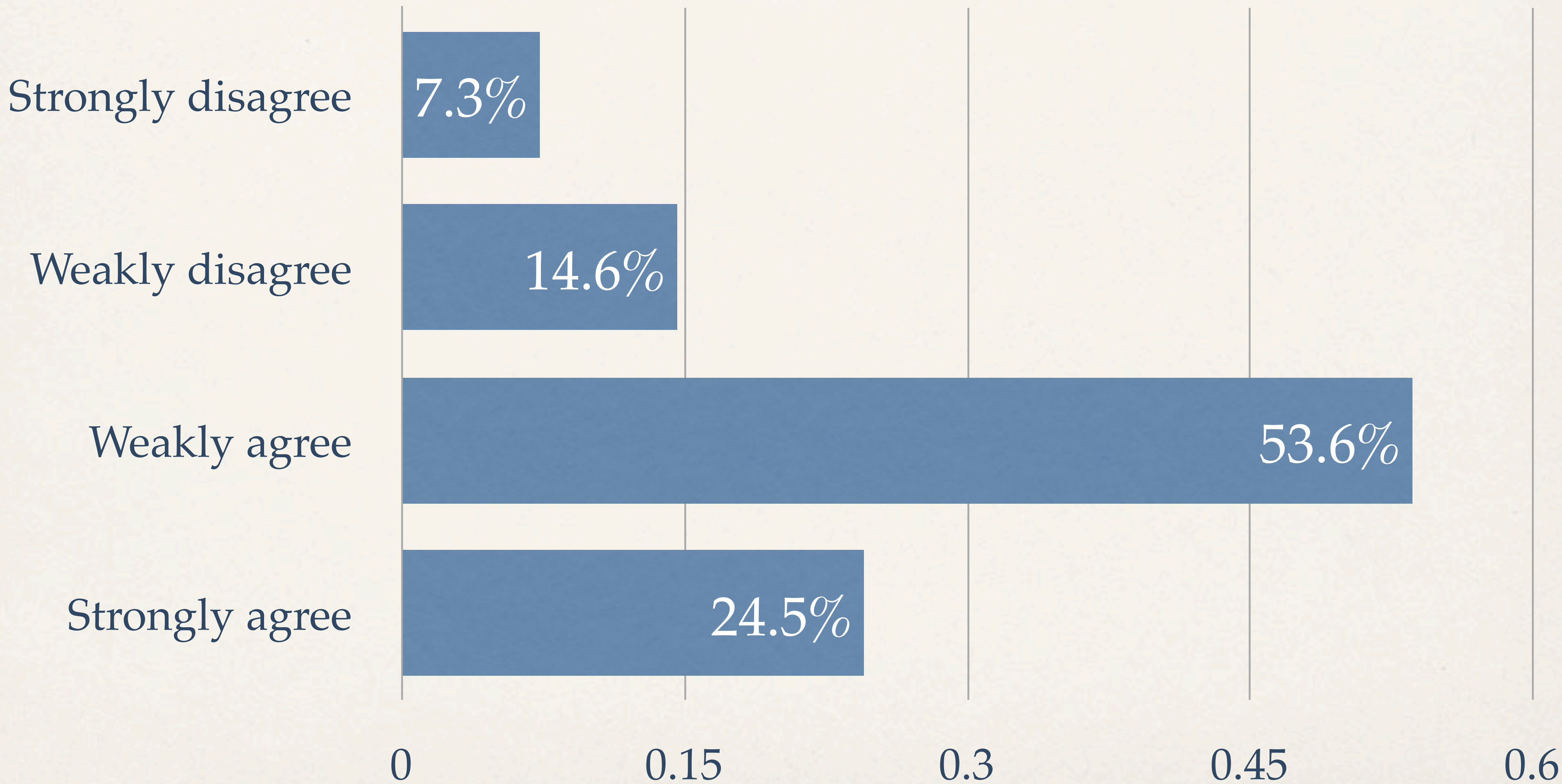


If you never refrain from searching (i.e., if you selected (a) in the previous question), why is this?

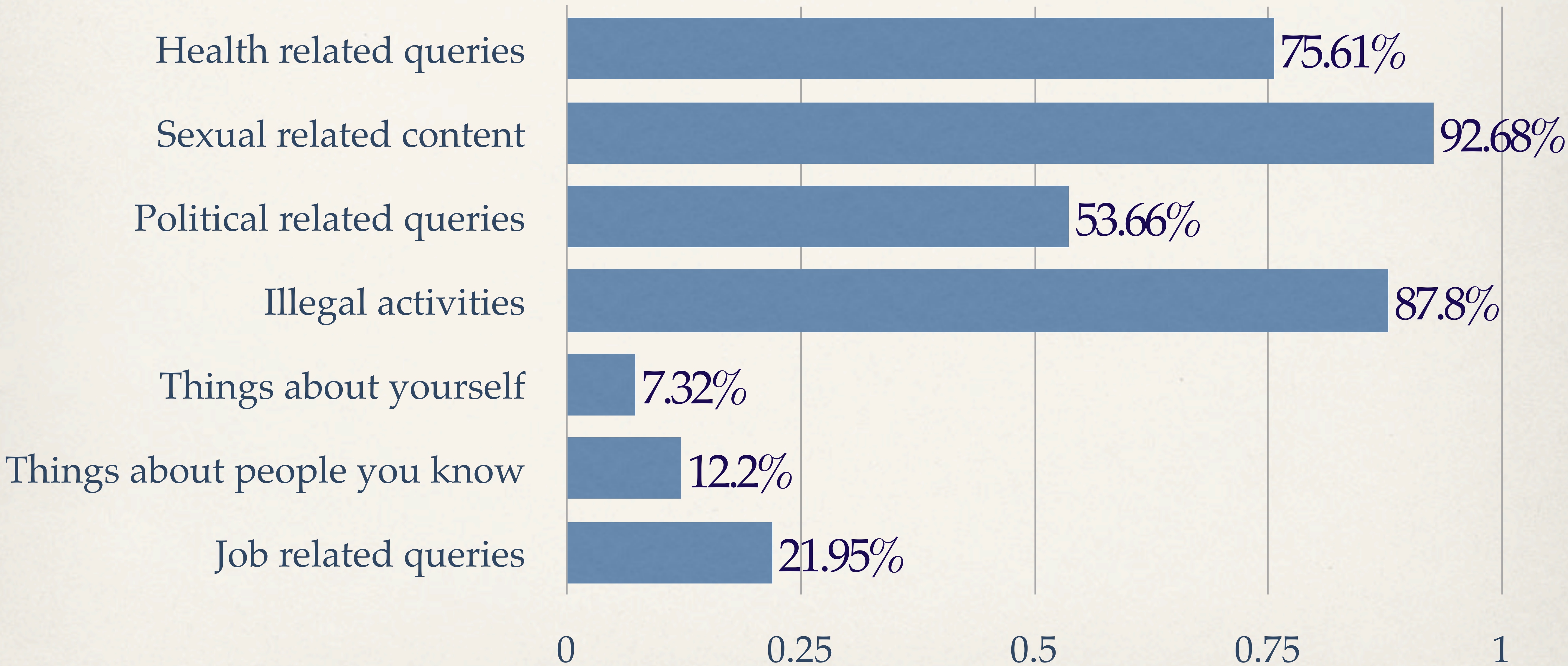
How much do you agree with the following statement: "When I use Google to search the Web, Google has a good chance of associating my identity with each of my queries if it chooses to do so."



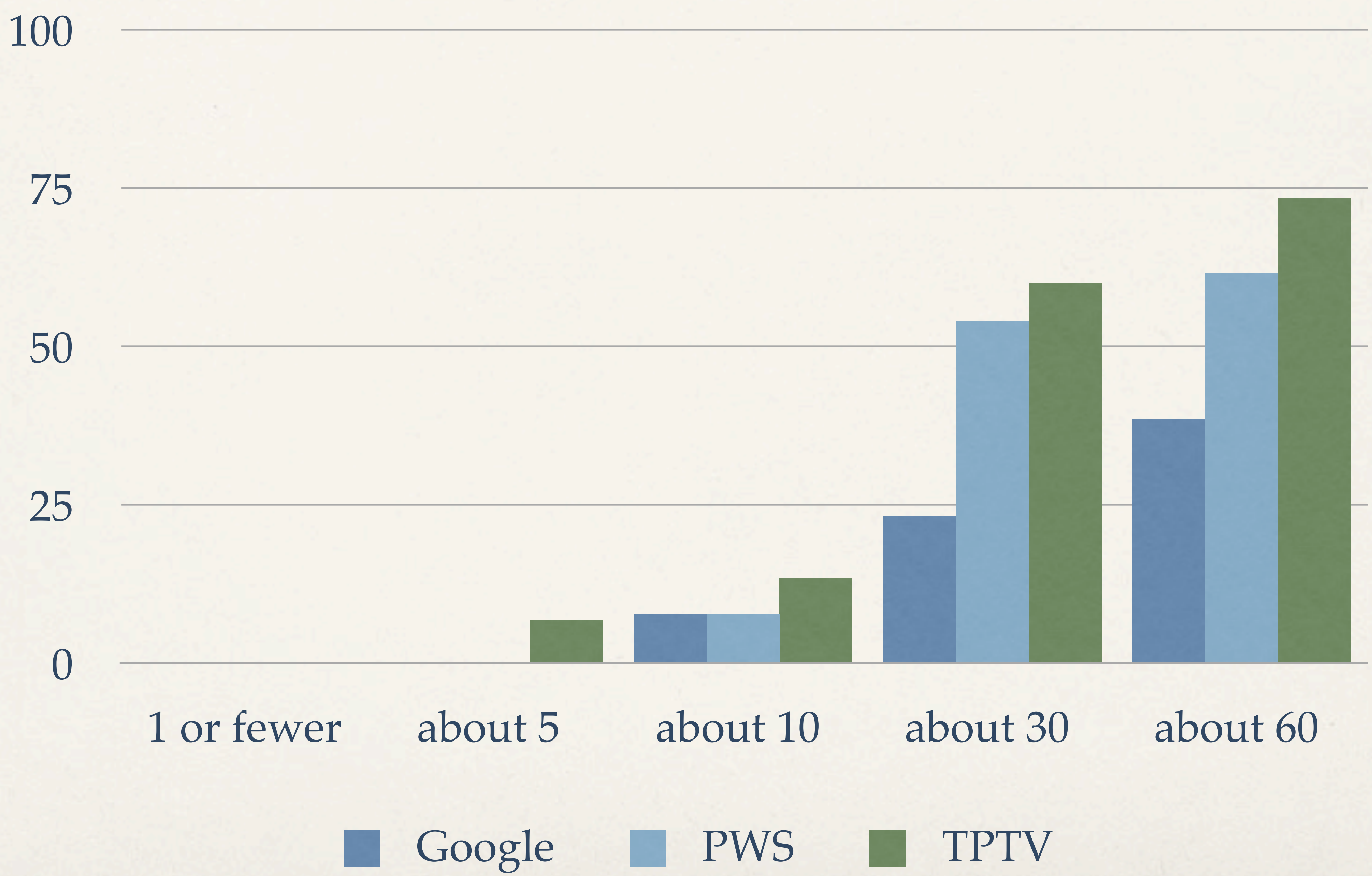
How much do you agree with the following statement: “Google keeps a fairly complete search history associated with my identity.”



If Google were able to associate each query you issue with you, and you had an equally accurate alternative method for searching that protected your identity, you would consider it using it for queries about (select all relevant answers):



Percentage of users in each group who said they would never trade N seconds of delay for identity protection, for $N \in \{1, 5, 10, 30, 60\}$



Conclusions

- PWS shows promise.
    - Provides additional privacy enhancement.
    - Easier to install and use than TPTV.
- Tor-based privacy enhancements are not currently a realistic solution.

Limitations

- Lack of motivation: Why should users do their best?
- Lack of experience: Would PWS users do better the second time they use it?
- Lack of depth: Are trivia questions too easy?
- Lack of time: How unrecoverable are those errors really?

CRA Taulbee Survey of Computer Science Faculty Salaries

- Computer science departments in four tiers:
  12 + 12 + 12 + all the rest
- Academic faculty in four ranks: full, associate, assistant professors, and non-tenure-track teaching faculty
- Intention: Publish aggregate salary statistics per tier-rank without revealing department-specific information or individual salaries.

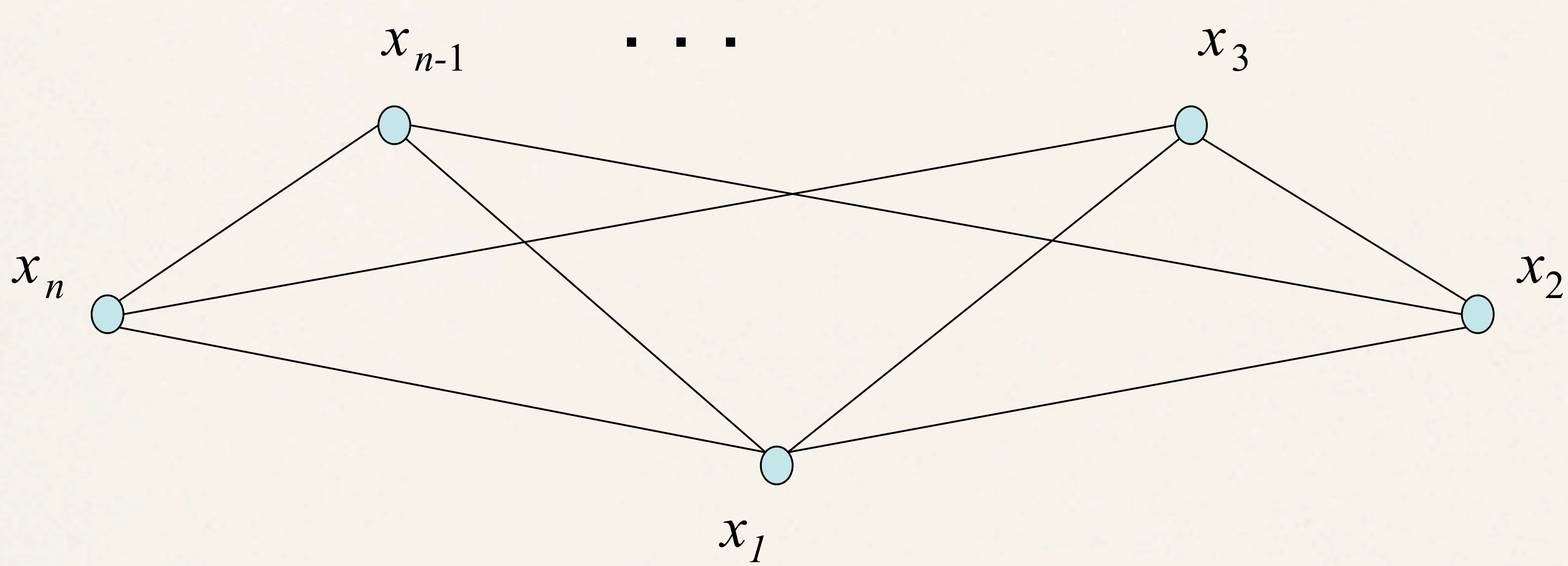
Traditional Taulbee Computation

- Inputs, per department and faculty rank:
  - Minimum
  - Maximum
  - Median
  - Mean
- Outputs, per tier and faculty rank:
  - Minimum, maximum, and mean of:
    - department minima
    - department maxima
    - median of department means (not weighted)
  - Mean (weighted mean of department means)

Our Challenge

- CRA wishes to provide more extensive statistics than the meager data traditionally collected can support.
- Asking departments to provide complete lists of salaries greatly increases the need for trust in CRA's intentions and its security competence.
- Detailed disclosure, even if anonymized, may be explicitly prohibited by the school.
- Hence, there is a danger of significant non-participation in the Taulbee Survey.

Secure Multiparty Computation (SMPC)



$$y = F(x_1, \ldots, x_n)$$

Each i learns y.

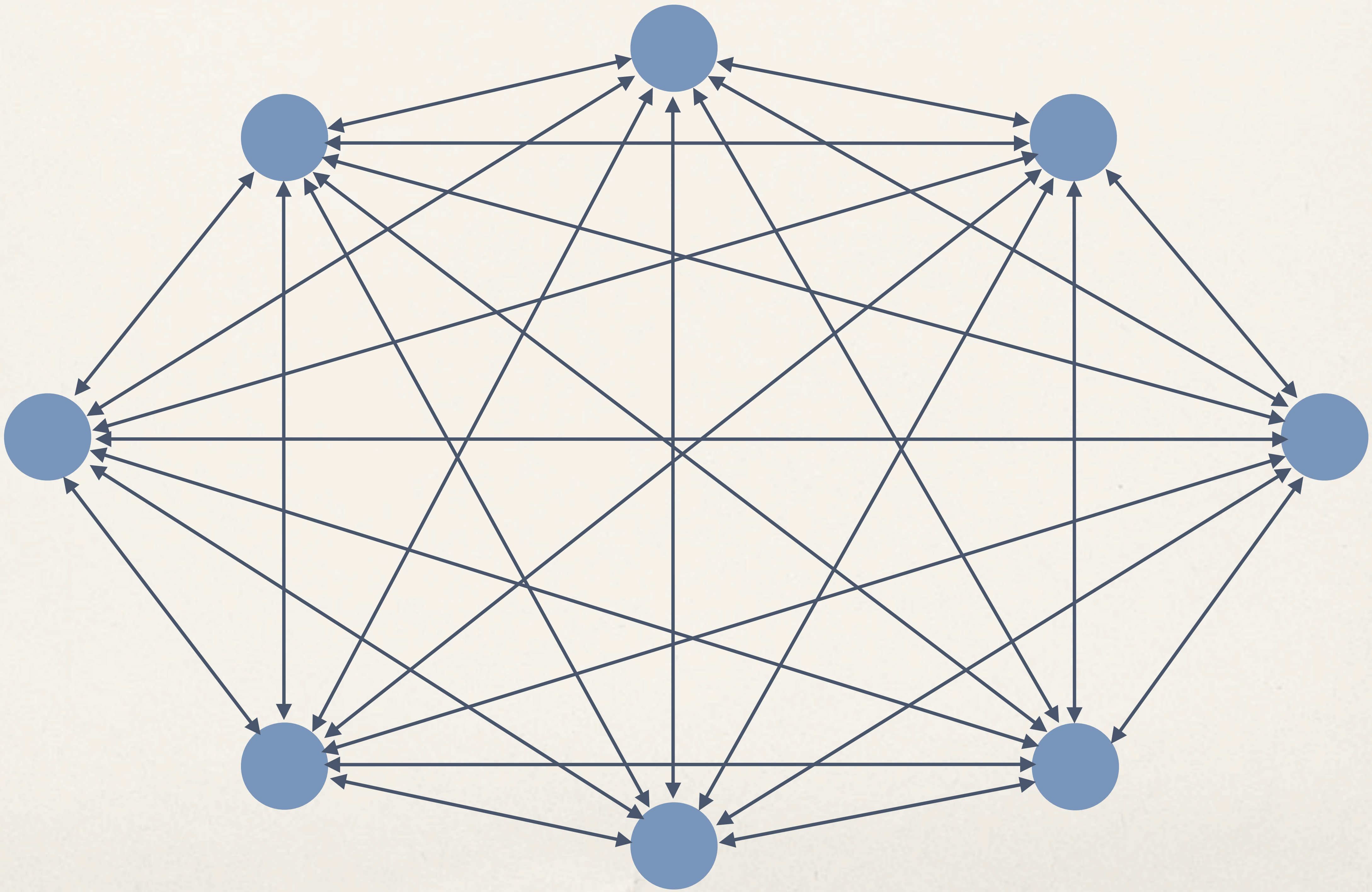
No i can learn anything about x_j (except what he can infer from x_j and y).

Very general positive results. Not very efficient.

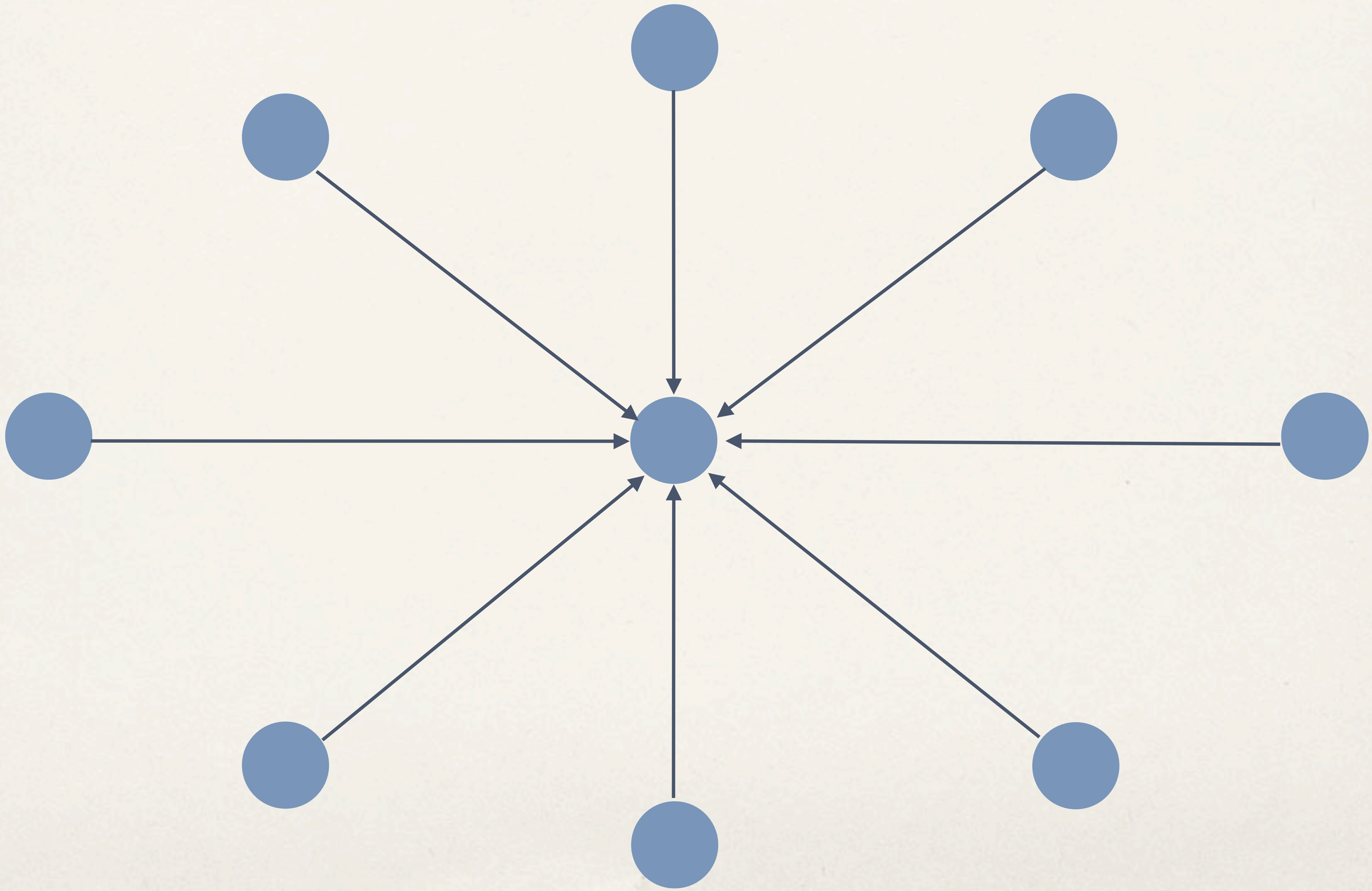
Applying SMPC to the Taulbee Survey

- We cannot simply run an SMPC protocol "off the shelf" but rather must arrive at a reasonable *coordination architecture*. [Ryger]
- We can use the *Fairplay* S2PC system of Malkhi *et al.* [USENIX Sec. 2004], but we must supply:
  - a user interface aimed at the Taulbee survey
  - a specialized "circuit constructor"

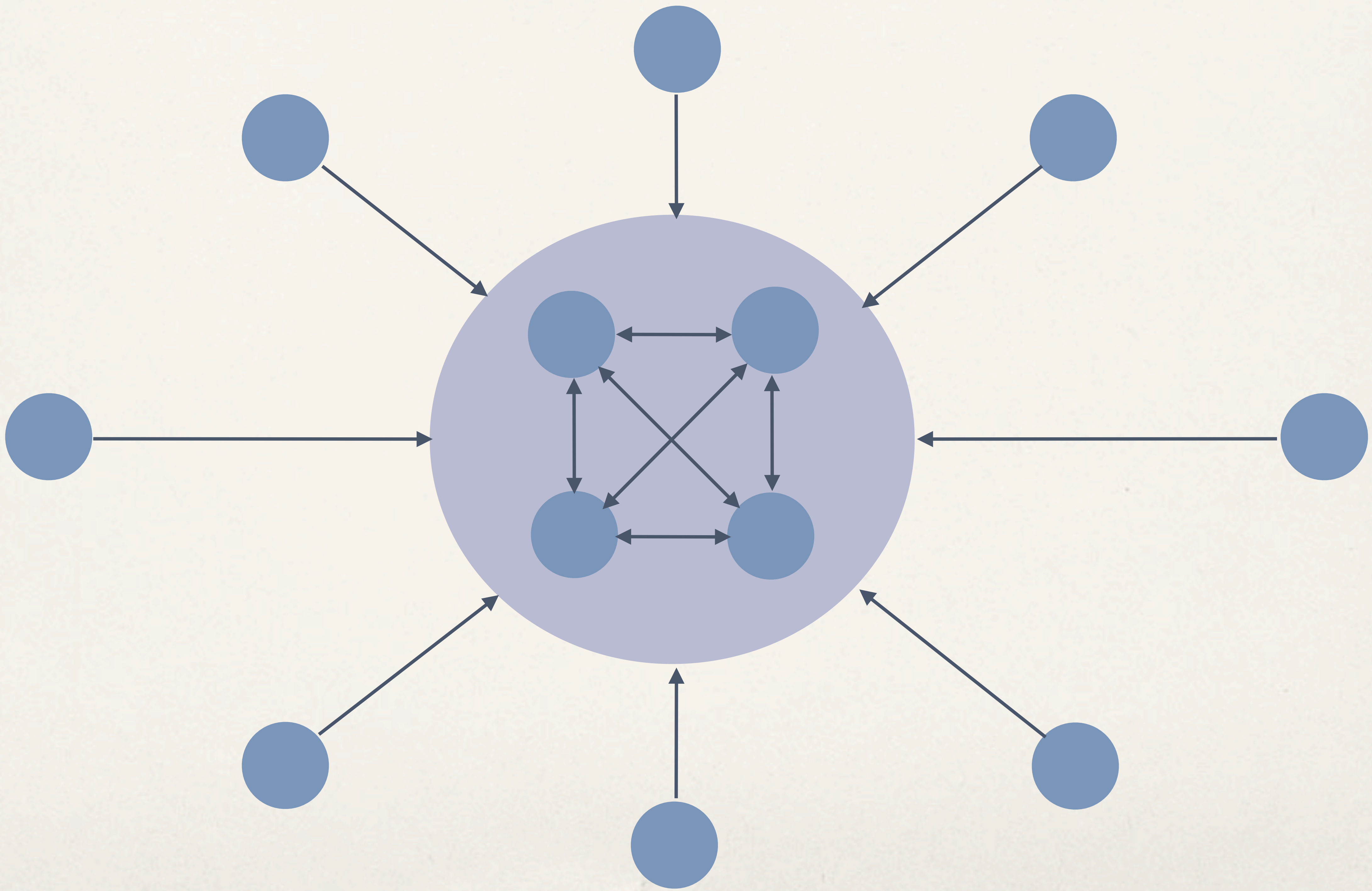
Communication Pattern: General SMPC Protocols



Communication Pattern: Surveys and Other Trusted-Party Computations



Communication Pattern: M-for-N-Party SMPC



Privacy-Preserving Data Entry

Browser Context

Full	120K
Full	110K
Assoc	75K
Assoc	100K
Assoc	100K

Input x is a set of salaries.

S1

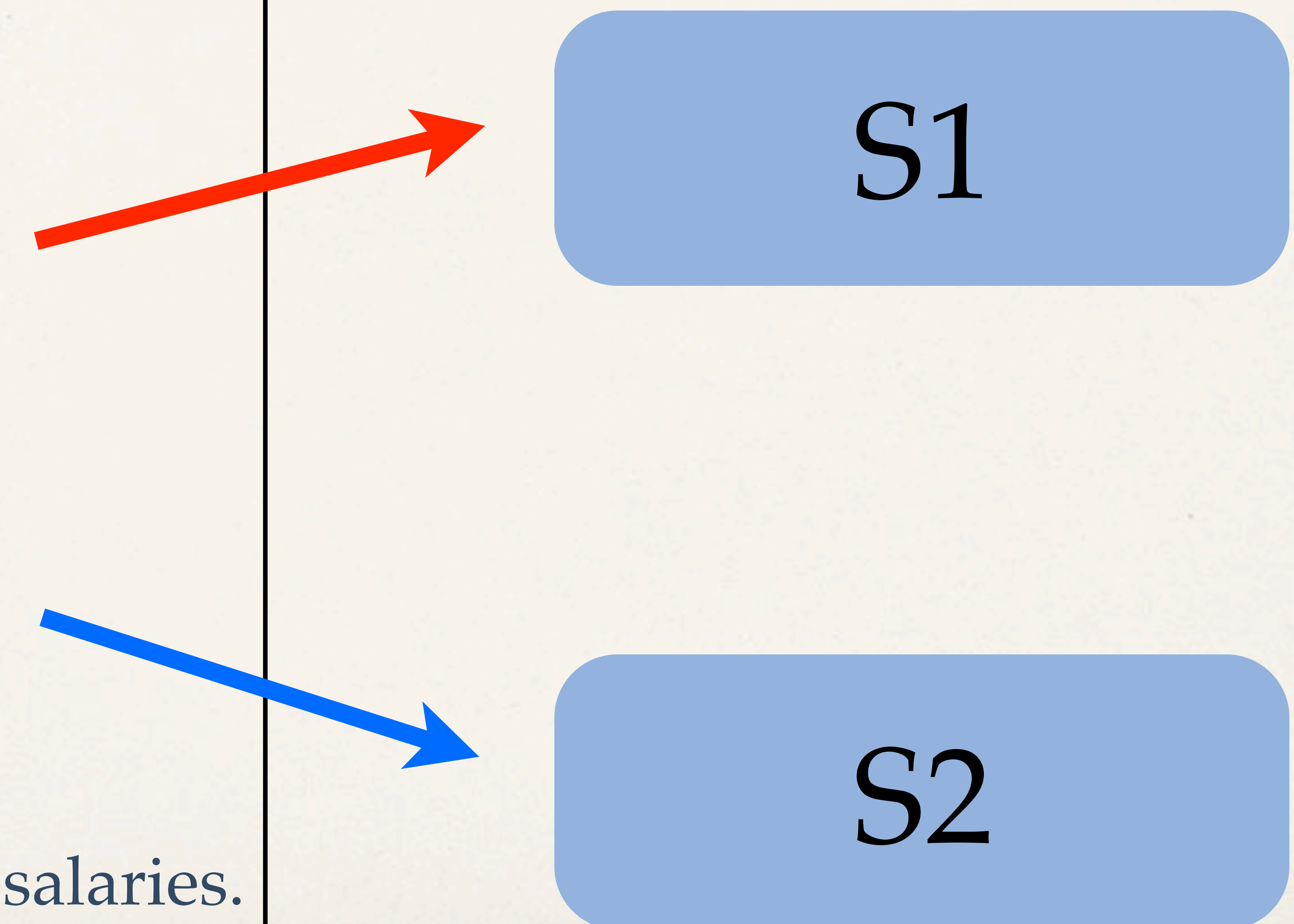
S2

Privacy-Preserving Data Entry

x(b) = RED(b) ⊕ BLUE(b)

Browser Context

Full	120K
Full	110K
Assoc	75K
Assoc	100K
Assoc	100K



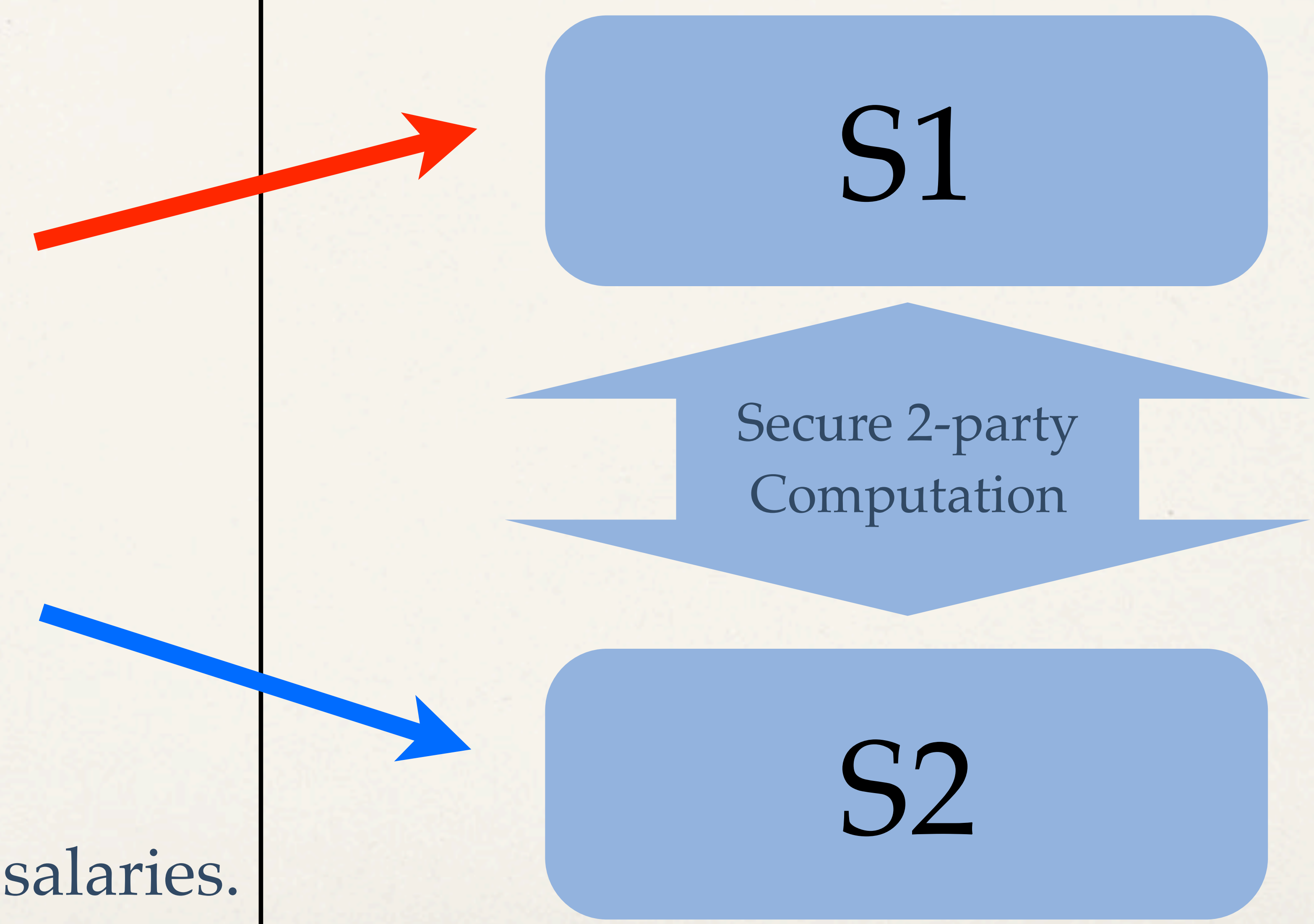
Input x is a set of salaries.

Privacy-Preserving Data Entry

$x(b) = \text{RED}(b) \oplus \text{BLUE}(b)$

Browser Context

Full	120K
Full	110K
Assoc	75K
Assoc	100K
Assoc	100K



Input x is a set of salaries.

Summary: Input-Collection Phase

- Department representative enters salary list and ranks.
- Per rank, in JavaScript, computation of XOR shares of the individual salaries for the two (M = 2) computation servers
- Per rank, HTTPS transmission of XOR shares to their respective computation servers
- Note that cleartext data never leave the client machine.

Computation Phase
(for each tier-rank)

- Construction of a Boolean circuit to
  - reconstruct inputs by XOR-ing their shares
  - sort the inputs in an odd-even sorting network
- Secure computation
  - Fairplay [Malkhi et al., 2004] implementation of the Yao S2PC protocol for the constructed circuit and the collected input shares
  - Output is a sorted list of all salaries in this tier-rank.
- Postprocessing
  - arbitrary, statistical computation on the sorted, cross-departmental salary list

Fairplay

- Fairplay = Compiler + S2PC Runtime
- Compiles a Pascal-like specification of F( ) into boolean circuit
- The Fairplay compiler was not efficient enough! It produced circuits that were too large for S2PC execution. We created a custom circuit generator.

Summary: Circuit Construction

- Implement Compare-and-Swap as a truth table
- Link Compare-and-Swap operations into a sorting circuit of the appropriate size
- We use OddEven sorting networks (Batcher). $O(k \log^2 (k))$ Compare-and-Swap operations to sort k integers -- adequate for S2PC protocol execution.

The Heartbreak of Cryptography

- User-friendly, open-source, free implementation
- NO ADOPTION !@%$#
- CRA's reasons
  - Need for data cleaning and multiyear comparisons
  - Perhaps most member departments will trust us.
- Yale Provost's Office's reasons
  - No legal basis for using SMPC on data that we otherwise don't disclose
  - Correctness and security claims are hard and expensive to assess, despite open-source implementation.
  - All-or-none adoption by Ivy+ peer group.

Conclusions

- From a technical point of view, implementation and deployment of SMPC theory is more tractable than many in the security-research community believe. Fairplay is a useful platform, and there is now a multiparty (*i.e.*, M > 2) version ``FairplayMP'' (Ben-David *et al.*, CCS 2008).
- Widespread adoption of SMPC protocols will require overcoming substantial economic, social, and legal barriers.

Private Information Retrieval

Q(uerier)

4

Lettuce

	D(atabase)
1	Apple
2	Pear
3	Tomato
4	Lettuce
5	Avocado
6	Cherry

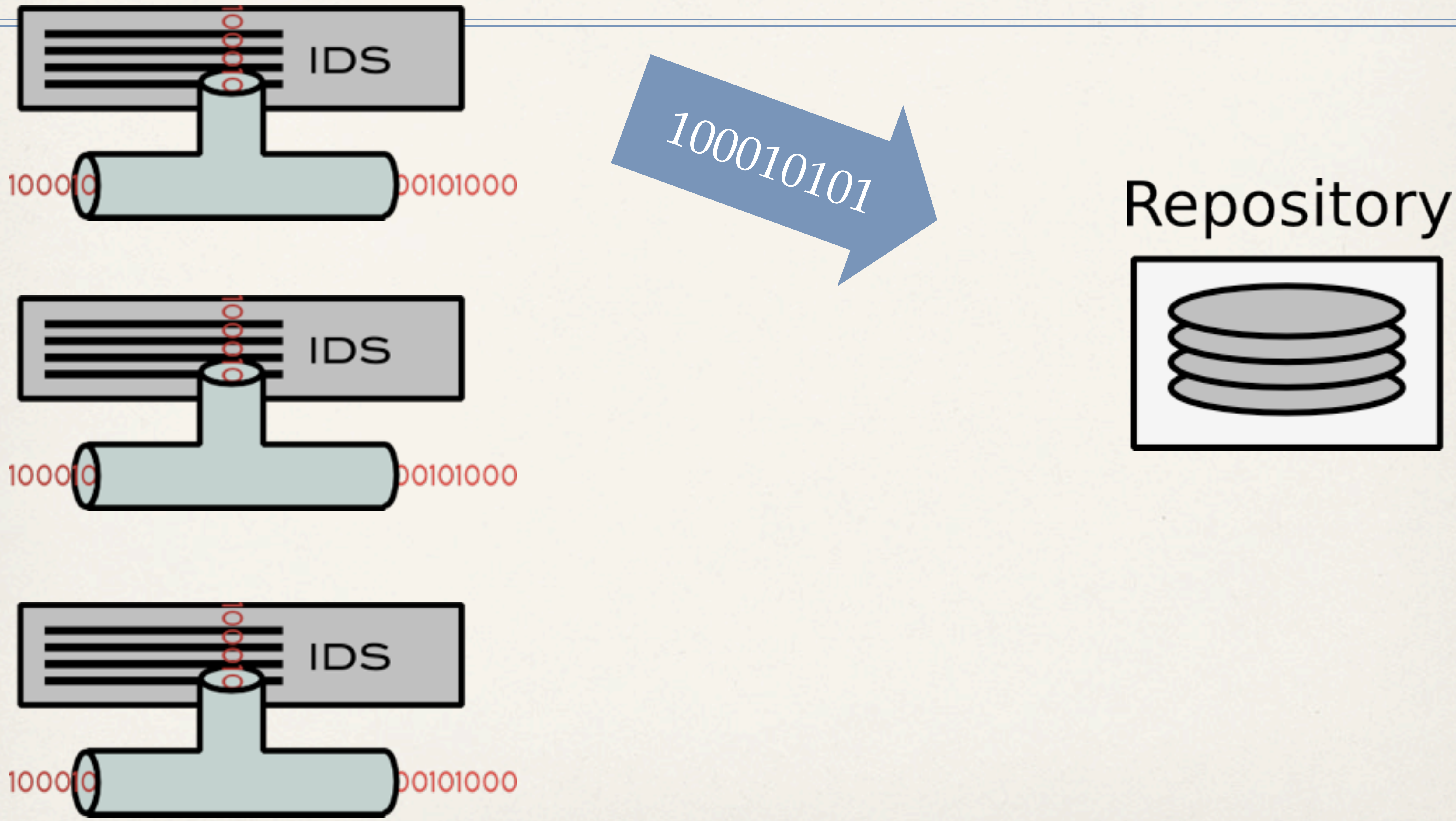
- PIR: Q learns D[4] and D learns nothing about the query (4)
- SPIR: PIR + Q learns D[4] and nothing else about D

Our Contributions

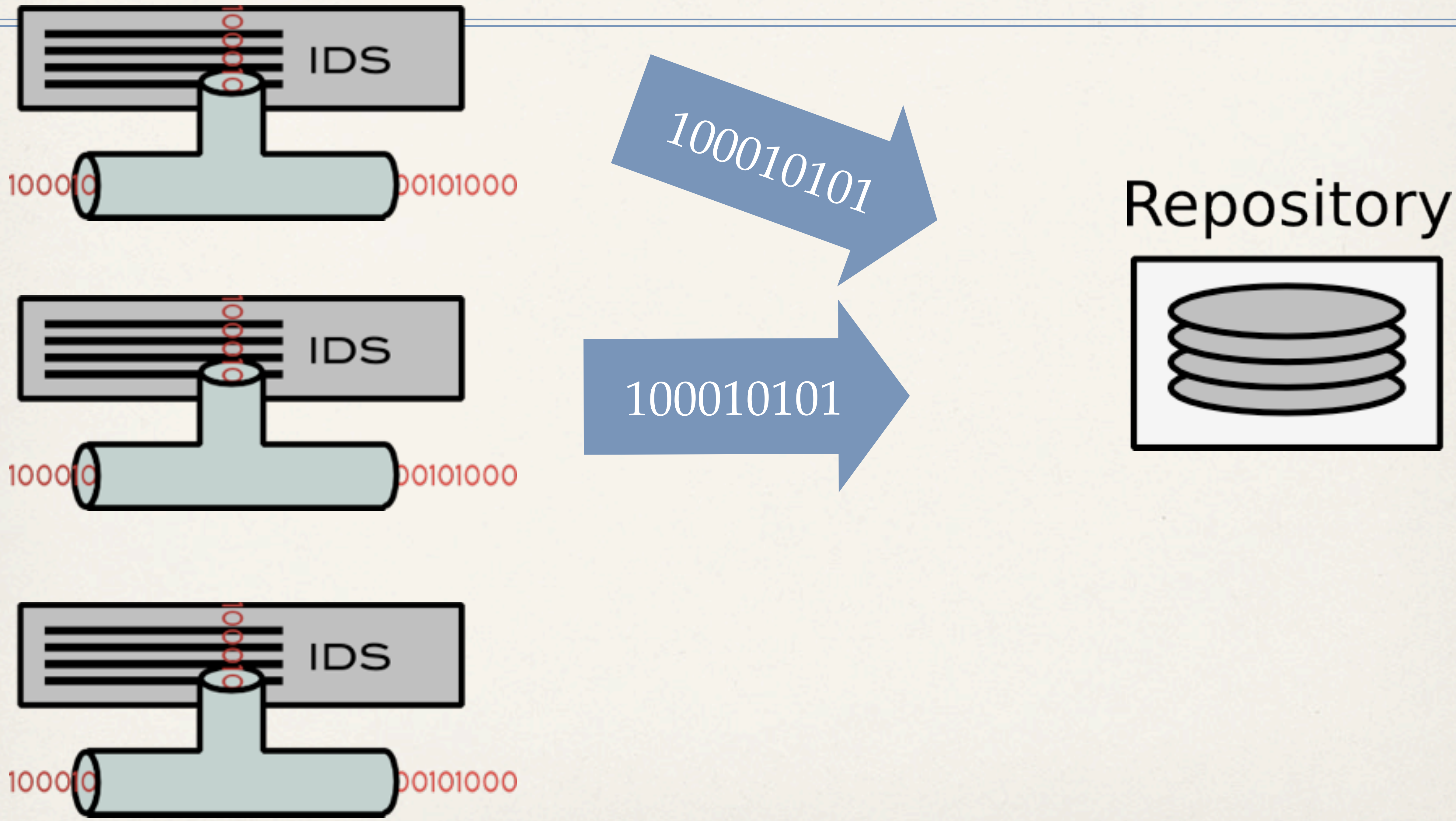
- Java implementation of PIR and SPIR protocols of Naor and Pinkas (Crypto 1999)
- Efficiency enhancement to PIR that replaces an O(n)-communication initialization step with O(n) local computation
- Resulting implementation is fast enough for modest-sized databases but not for large databases.

Security-Alert Sharing

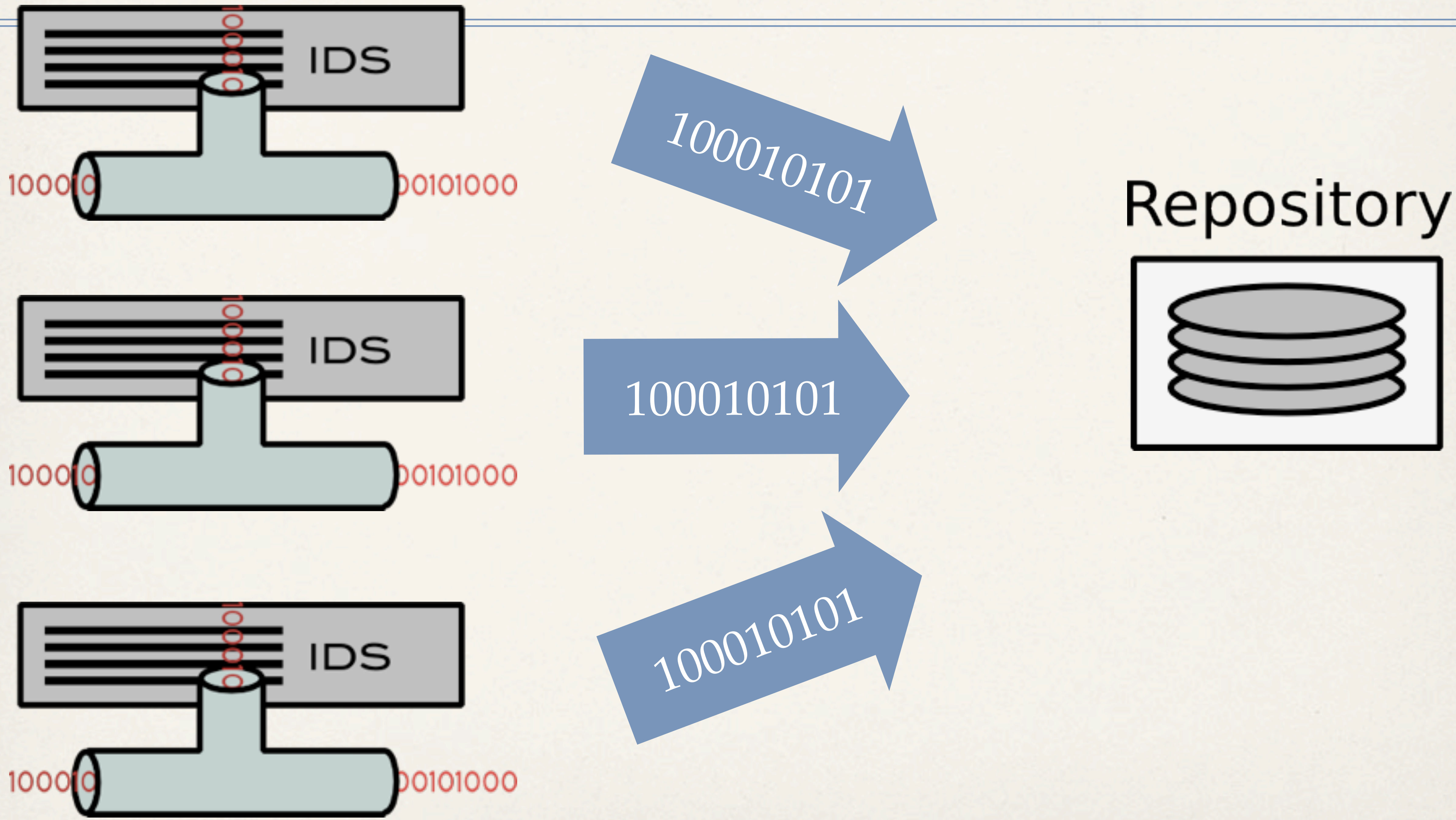
Security-Alert Sharing



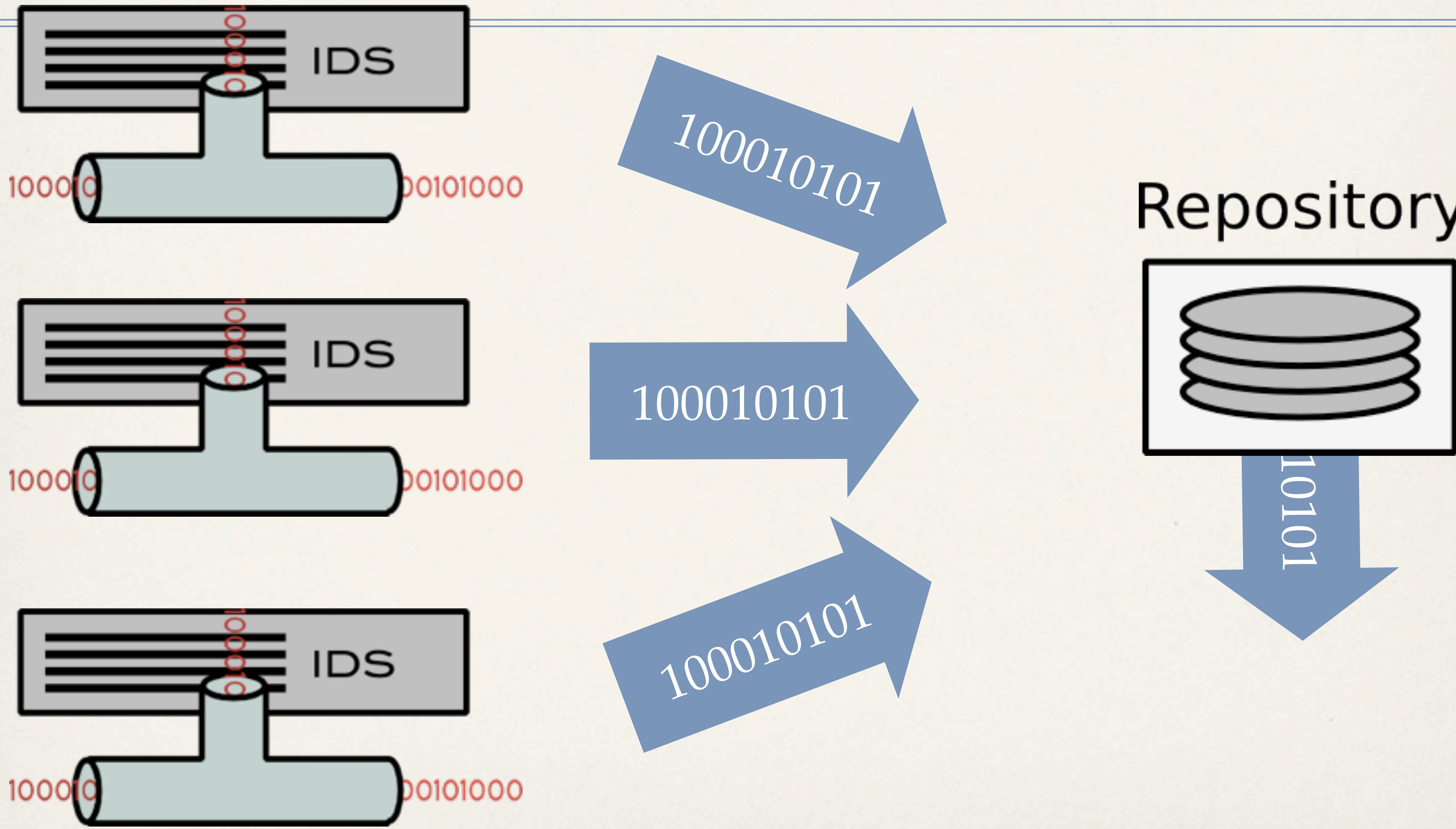
Security-Alert Sharing



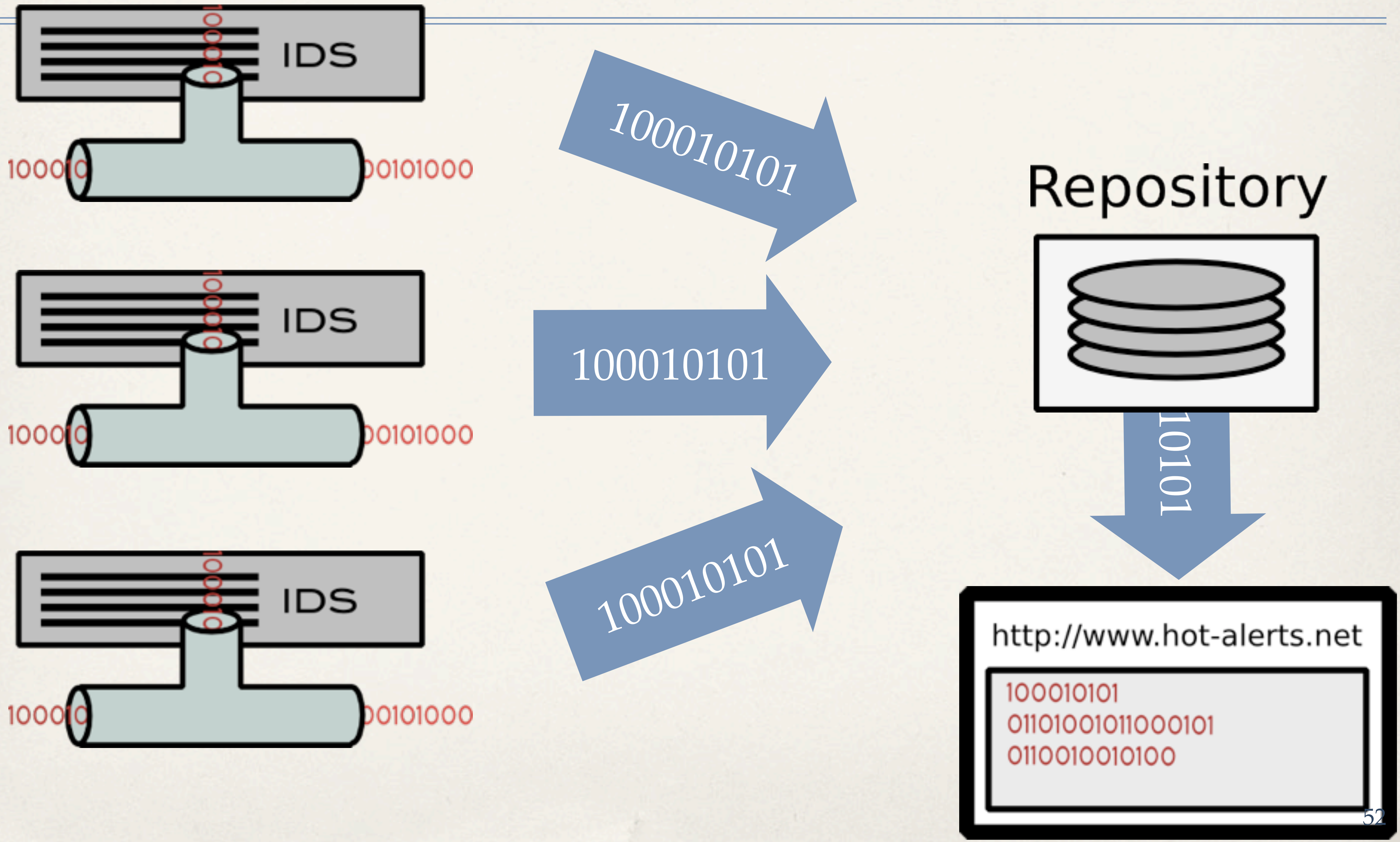
Security-Alert Sharing



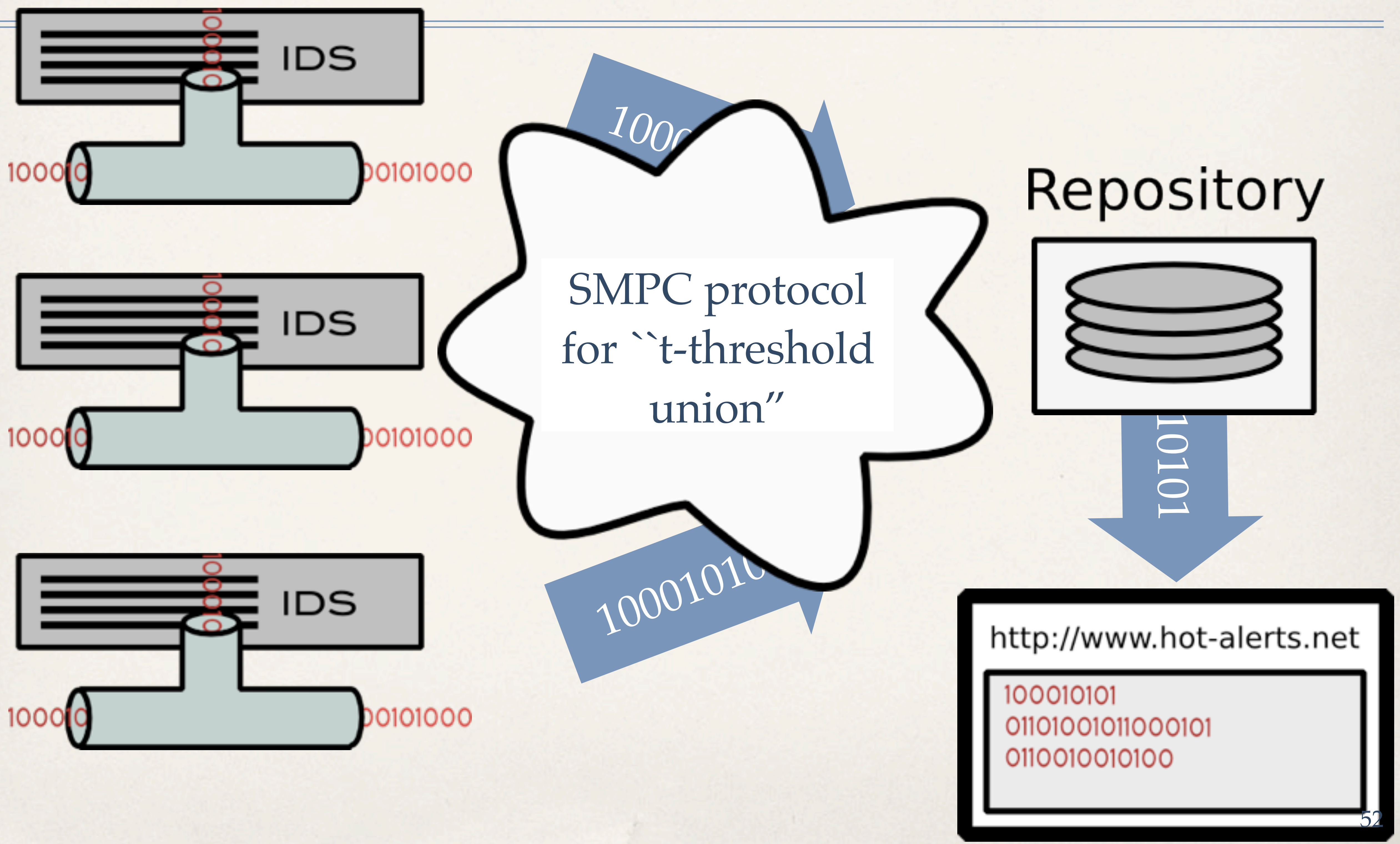
Security-Alert Sharing



Security-Alert Sharing



Security-Alert Sharing



Our Contributions

- Solution based on Threshold Identity Based Encryption (TIBE)
    - Each contributor sends to the repository one *share* of each alert he receives.
    - Achieves *entropic security* (assuming a high-entropy alert space)
    - Significantly more scalable than t-threshold union protocol of Kissner and Song (Crypto 2005)
- Open problem: What is the real alert-space distribution?

Conclusions

- Less and less sensitive information is truly inaccessible. The question is the cost of access, and that cost is decreasing.
- Foundational legal theories to support obligations and rights in cyberspace are lacking.
- Technological progress is still going strong, 34 years after the publication of Diffie and Hellman's seminal paper, but adoption is slow.
- Client-side defenses can only go so far.

What's Next?

- More technological progress, but ...
- We need a paradigm shift on sensitive data: Strive for *accountability instead of secrecy*.
- Traditional data security is based on *preventing unauthorized access* to sensitive information.
- Internet-age data security should be based on *ensuring appropriate use* of sensitive information.

Support for an Accountability Agenda

Lampson, CACM 2009:

Misplaced emphasis on prevention (``security based on locks”) rather than accountability (``security based on deterrence”) has resulted in unusable security technology that people do not understand and often work around.

Support for an Accountability Agenda (2)

Weitzner et al., CACM 2008:

``For too long, our approach to information protection policy has been to seek ways to prevent information from `escaping' beyond appropriate boundaries, then wring our hands when it inevitably does. This hide-it-or-lose-it perspective … on privacy, copyright, and surveillance is increasingly inadequate. … As an alternative, accountability must become a primary means through which society addresses appropriate use."

Questions?