Automated Text Mining for Requirements Analysis of Policy Documents

Aaron Massey
Postdoctoral Fellow
School of Interactive Computing
akmassey@gatech.edu
@akmassey

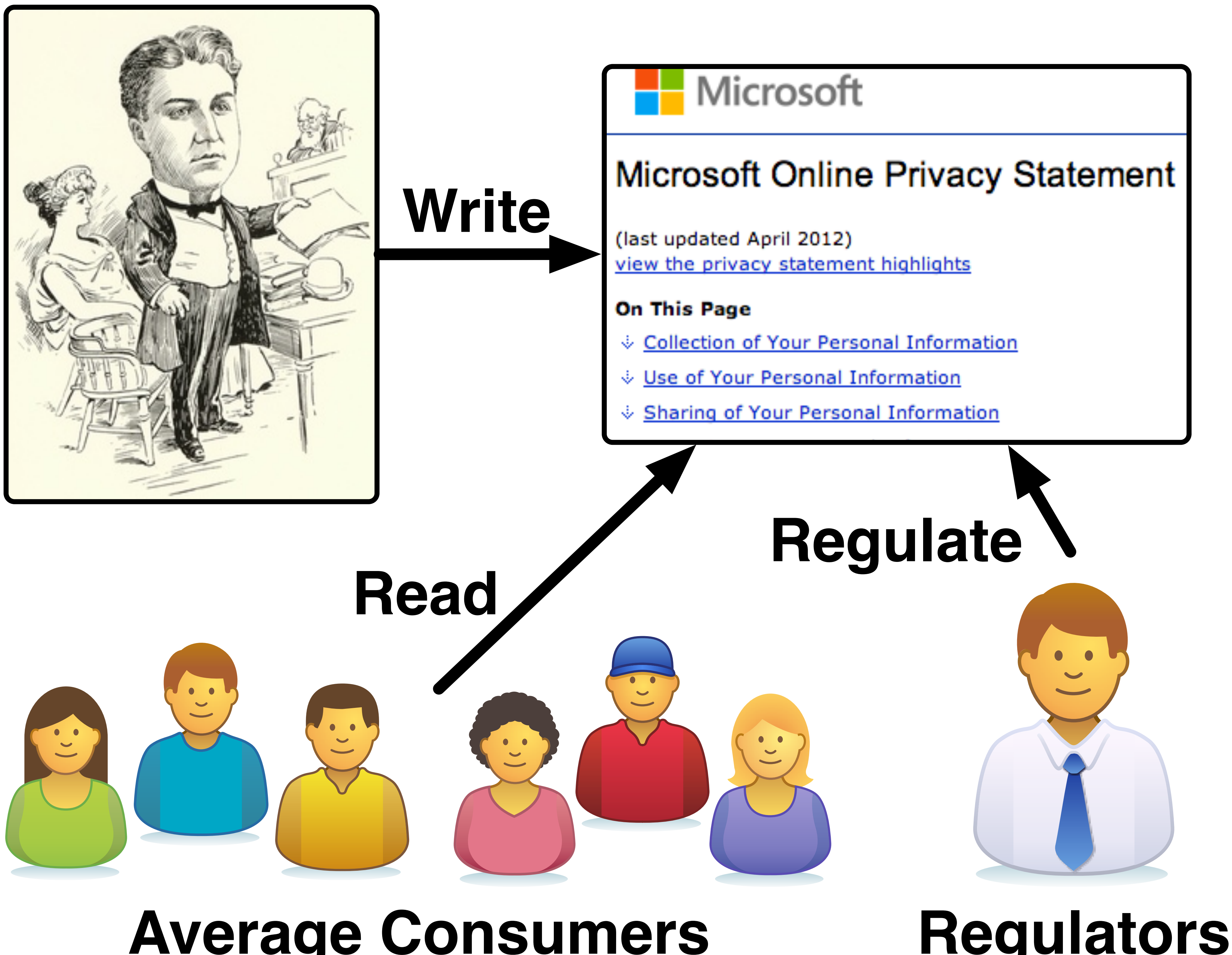
Co-Authors: Jacob Eisenstein, Annie I. Antón, and Peter P. Swire

17 July 2013

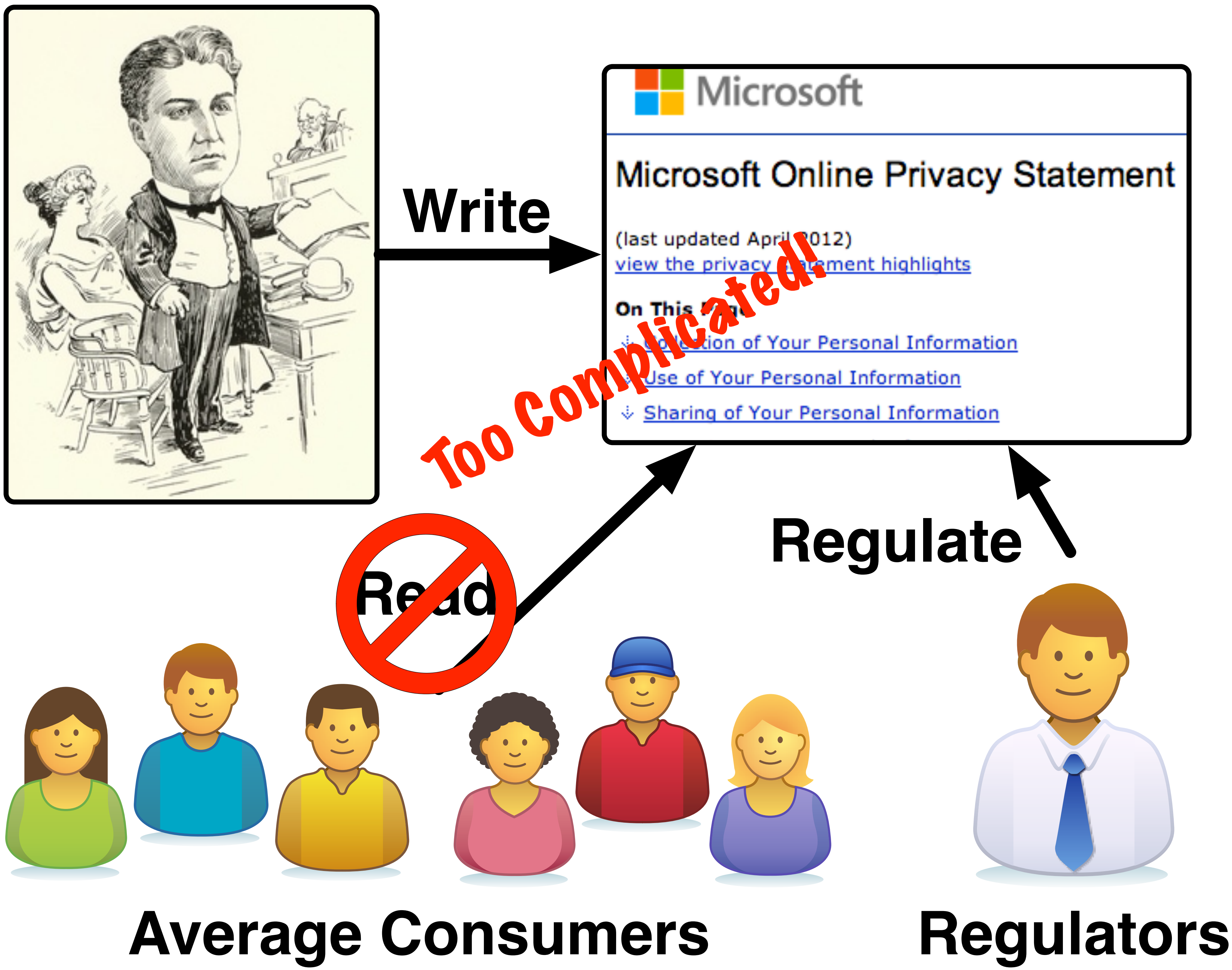
the privacyplace .org

Georgia Institute of Technology

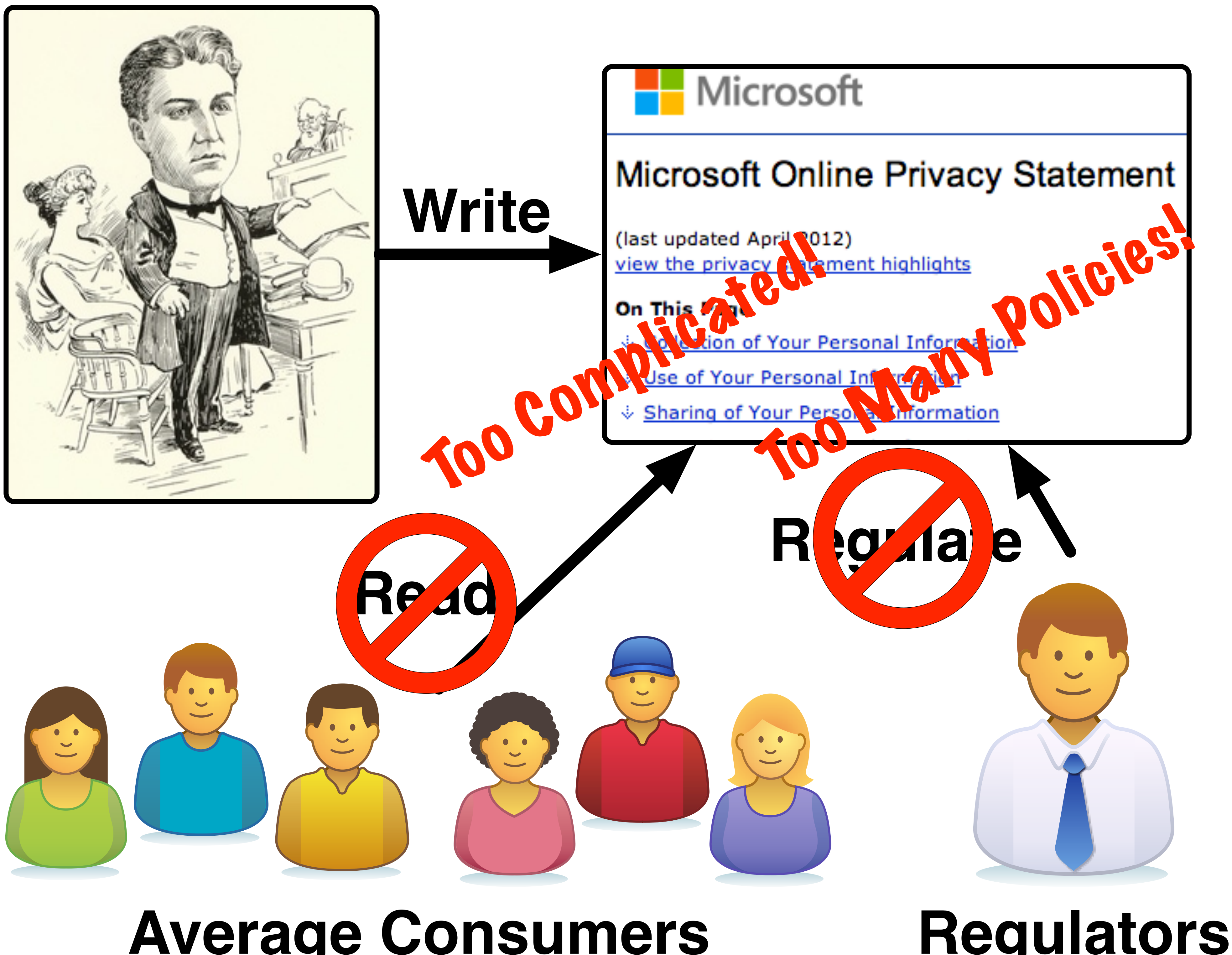
Idealized Policy Documents



Real Policy Documents



Real Policy Documents



Policy Document Readability

- Most research focuses on relatively small sets of privacy policies
  - 40 financial privacy policies [AE04]
  - 24 healthcare privacy policies [AEV07]
  - 75 privacy policies from popular websites [MC08]
- About half of the U.S. population doesn't have the level of education required to understand most privacy policies! [AE07]

Privacy Policy Taxonomy

[AEH04, AE04, AEV07]

- Privacy Policies consist of both privacy protection goals and possible privacy vulnerabilities.
- Goals and Vulnerabilities can be expressed in a semi-formal structure using keywords.
- Some Examples:
  - `COLLECT` date and times at which site was accessed
  - `STORE` credit card information until dispute is resolved
  - `ALLOW` affiliates to use information for marketing purposes

Engineers Must Participate!

Engineers are the Internal Audience

- **Engineers:** Must **ensure that software systems comply** with stated policies.
- Policy documents contain software requirements. [AE04, AEV07]
  - Some software requirements represent **privacy protection goals**
  - Other software requirements represent **vulnerabilities**
- Regulators need to understand these requirements because they represent possible areas of non-compliance.

Problem Statement

Can automated text mining help identify requirements found in prior research in at scale?

Research Questions

- **RQ1:** How similar, with respect to readability, are policy documents of different types, organizations, and industries?
- **RQ2:** Can automated text mining help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections and vulnerabilities?
- **RQ3:** Can topic modeling be used to confirm the generalizability of the Antón-Earp privacy protections and vulnerabilities taxonomy? [AE04]

Data Sets and Collection

- Corpus includes 2,061 policy documents

1. Two requirements engineering studies [AE04, AEV07]
   - 64 documents (all privacy policies)
2. The Google Top 1000 most visited websites
3. The Fortune 500 companies

- Data collection process:
  - Visit main organizational website manually
  - Manually identify any policy documents: privacy notice, privacy policy, terms of use, terms of service, etc...

Readability Results

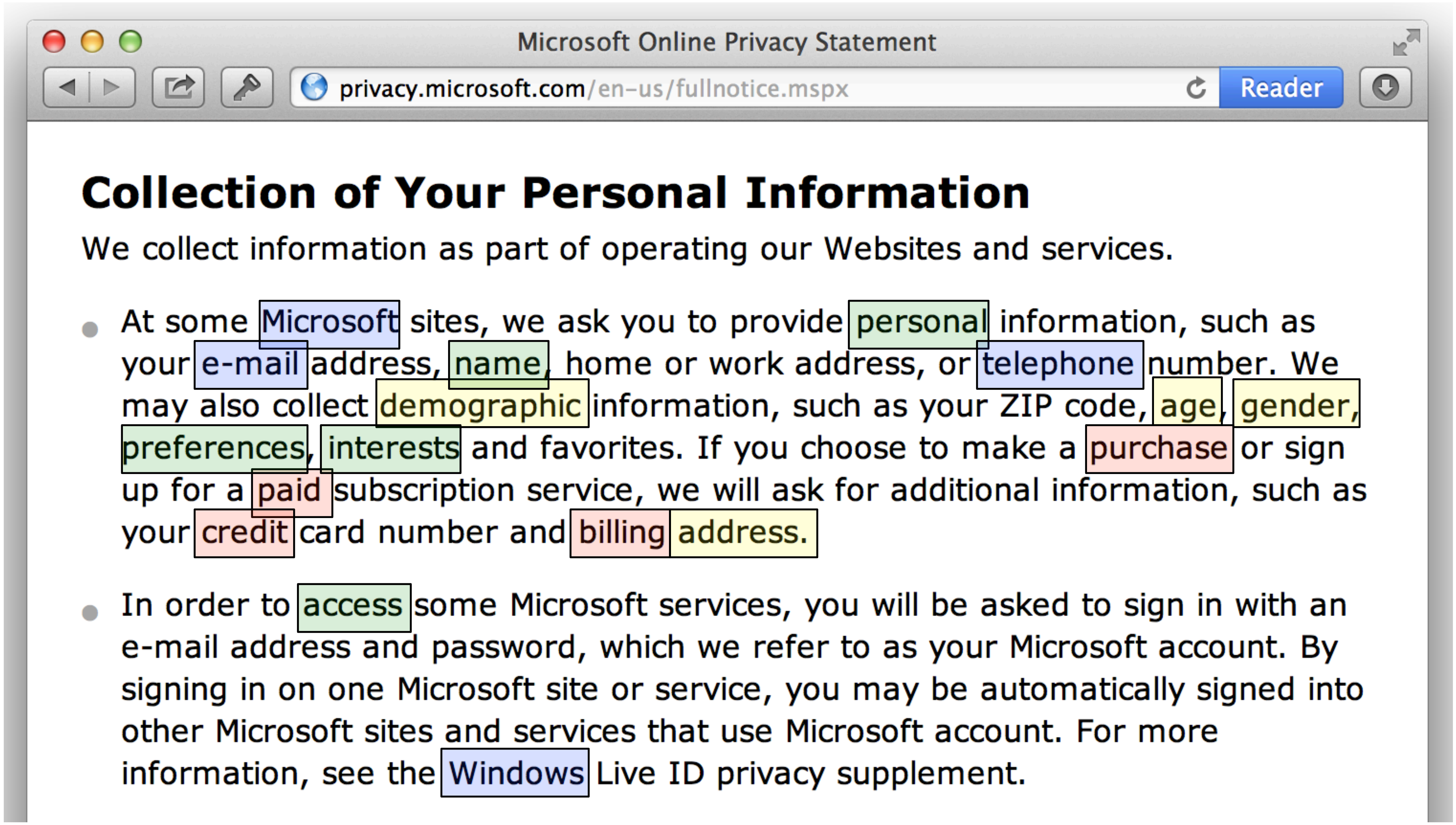
RQ1: Yes, other domains are similarly hard to read.

Document Set	FGL	FOG	SMOG	ARI
AE04 (40 policies)	13.5 (2.34)	14.9 (2.23)	15.2 (1.72)	13.7 (2.87)
AEV07 (24 policies)	13.9 (2.81)	15.5 (2.08)	15.6 (2.10)	13.6 (2.96)
Google Top 1000 Sites	15.4 (3.27)	16.0 (2.9)	16.6 (2.15)	15.3 (4.00)
Fortune 500	14.8 (3.67)	15.7 (3.28)	15.9 (2.09)	14.7 (4.47)

Topic Modeling: Latent Dirichlet Allocation (LDA)

- LDA is an approach to Probabilistic Topic Modeling that makes the following assumptions:
  1. Documents are made of topics, topics are made of words
  2. Topics are identified by the algorithm, not manually
  3. Topics are shared across documents in a corpus
- Caveat: the number of topics must be decided in advance
- Used successfully in bioinformatics, political science, and information retrieval

- **Our Goal: Can we identify documents likely to contain system requirements?**


Microsoft Online Privacy Statement

privacy.microsoft.com/en-us/fullnotice.mspx

Reader

Collection of Your Personal Information

We collect information as part of operating our Websites and services.

- At some Microsoft sites, we ask you to provide personal information, such as your e-mail address, name, home or work address, or telephone number. We may also collect demographic information, such as your ZIP code, age, gender, preferences, interests and favorites. If you choose to make a purchase or sign up for a paid subscription service, we will ask for additional information, such as your credit card number and billing address.
- In order to access some Microsoft services, you will be asked to sign in with an e-mail address and password, which we refer to as your Microsoft account. By signing in on one Microsoft site or service, you may be automatically signed into other Microsoft sites and services that use Microsoft account. For more information, see the Windows Live ID privacy supplement.

LDA Example Topics

Topics are Lists of Words

- **Blue Words:** Microsoft, e-mail, telephone, Windows
- **Yellow Words:** demographic, age, gender, address
- **Red Words:** purchase, paid, credit, billing
- **Green Words:** personal, name, preferences, interests, access

- **Caveat: It is dangerous to label these topics with semantic meaning.**

- Some words appear more often than others, and we can build **a distribution of how often these words appear** in a given topic.

The LDA Model

- Intuitions:
  - Documents consist of multiple topics, some of which appear more than others.
  - Topics consist of multiple words, some of which appear more than others.
- If we assume that all documents in the corpus share a common set of **possible** topics, then we can build a statistical model!
- Once we have this model, we can use it to determine **what topics appear most often** in the corpus or in a particular document.

Basic Methodology

- Normalize and preprocess the documents (downcase, stemming, drop stopwords, etc...)
- Select a subset of the policy documents to hold out for validation
- Build a series of topic models using LDA
- Identify the least perplexed model using the held out data
- Determine the extent to which the model helps identify requirements

Selecting a Topic Model

1. Started with 20 models with a pre-selected number of topics chosen evenly from K=10 to K=160
2. Selected the value for K that created the least perplexed model
3. Built an additional 15 models centered around that K
4. Selected the least perplexed model a second time

- Other approaches could be used to select the model:
  - Additional rounds to build and select models
  - Could have used something other than perplexity to accept the model, but perplexity is commonly used for this.

Using the Topic Model

Goal Keywords	Topics with Term	Documents with Topic
ALLOW	20	YouTube Terms of Service,
COLLECT	68	***Microsoft Privacy Statement,***
CUSTOMIZE	***150***	ConocoPhillips Legal and Privacy Statement
DISCLOSE	9	
INFORM	125	

- Select a Goal Keyword
- Select the topic in which the keyword is most likely present
- Select documents in which that topic is most likely present

Finding Requirements in Policy Documents

TABLE II
NUMBER OF POLICY DOCUMENTS (OUT OF 2,061) IDENTIFIED AS POTENTIALLY CONTAINING GOAL STATEMENTS

Key-word	Docu-ments	Key-word	Docu-ments	Key-word	Docu-ments
access	904	apply	331	change	31
collect	202	comply	339	connect	121
display	308	help	61	honor	19
inform	23	limit	52	notify	347
opt-in	32	opt-out	76	post	76
request	31	reserve	51	share	300
specify	38	store	38	use	525

Research Question Summary

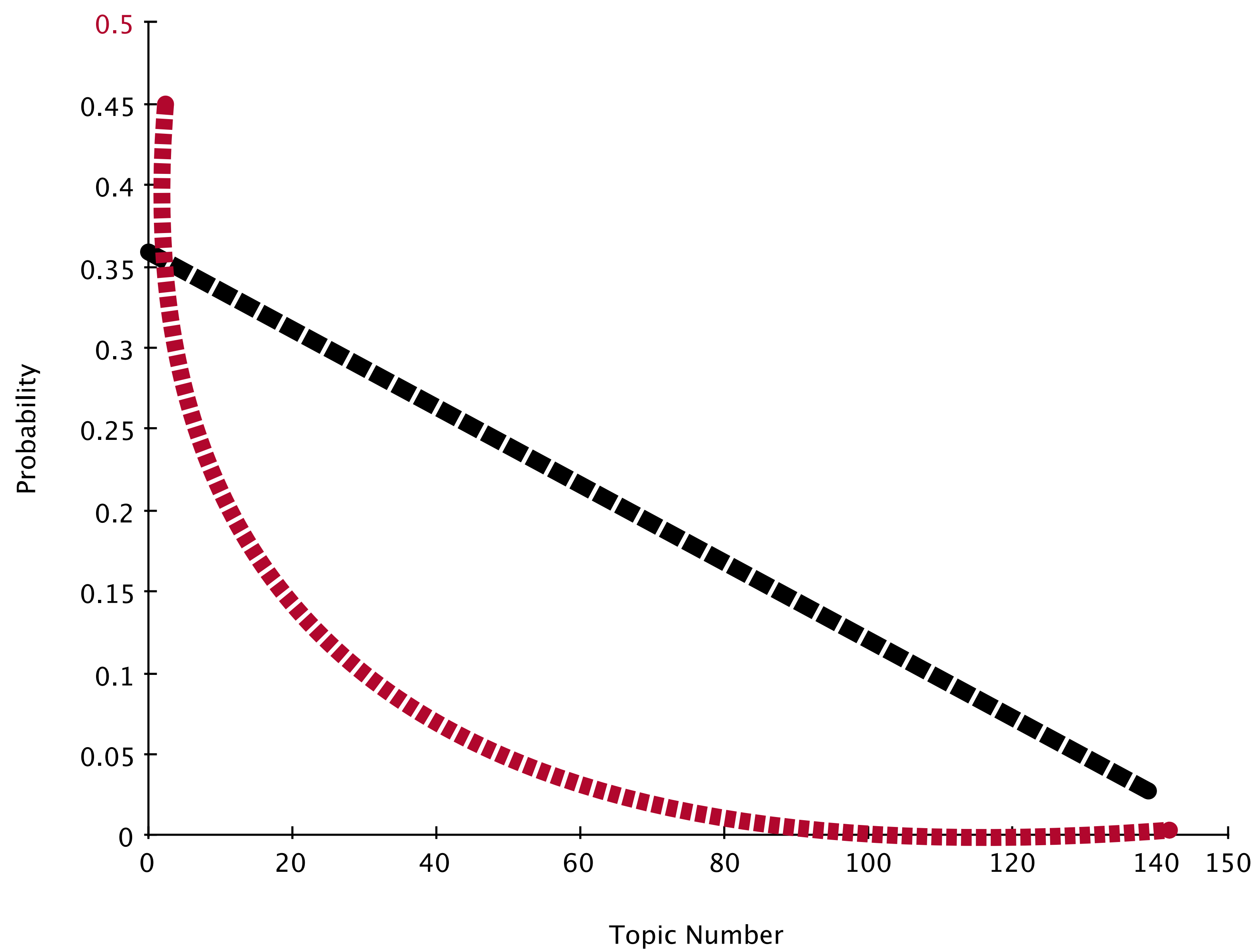
- **RQ1:** Are the documents similarly hard to read? **Yes.**
- **RQ2:** Can topic modeling help requirements analysts? **Found Supporting Evidence**
- **RQ3:** Can topic modeling confirm broader use of the Antón-Earp taxonomy [AE04]? **Found Supporting Evidence**

Areas of Future Work

- How can we validate these models are useful?
- Can we improve our ability to find requirements by including additional parts of the goal-based requirements analysis? (i.e. Can we relax LDA's assumptions to improve performance?)
- What approaches to visualizing the model would improve their usefulness for engineers, consumers, and regulators?

Additional Future Work

- We only explored the most probable topic for a keyword and the most probable document for a topic.
- **We could look at the actual distributions!**



Thank You!
Questions?

Aaron Massey
Postdoctoral Fellow
School of Interactive Computing
akmassey@gatech.edu
@akmassey

Co-Authors: Jacob Eisenstein, Annie I. Antón, and Peter P. Swire

17 July 2013

the privacyplace .org

Georgia Institute of Technology