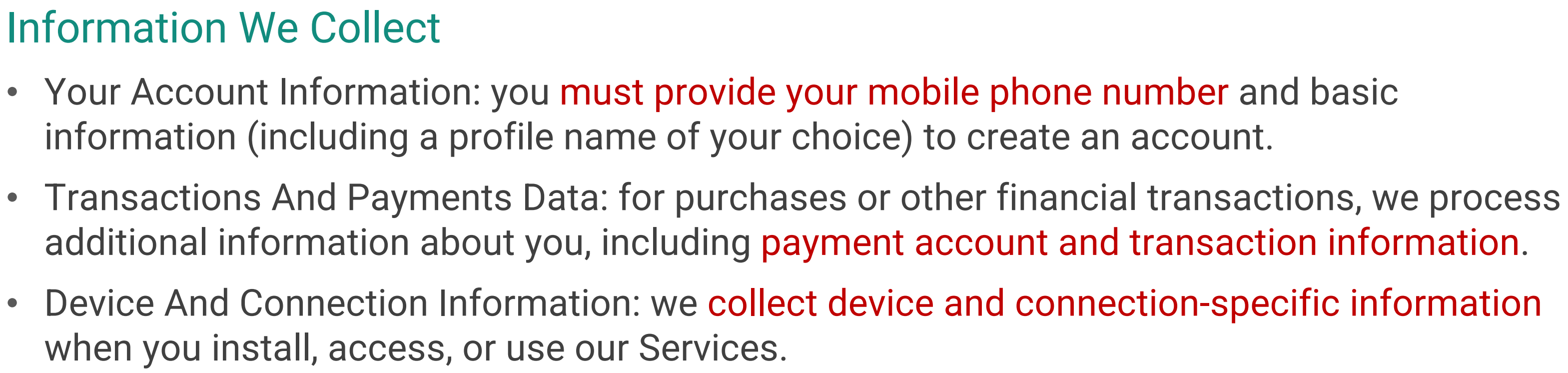
Calpric: Inclusive and Fine-grain Labeling of Privacy Policies with Crowdsourcing and Active Learning

Wenjun Qiu, David Lie and Lisa Austin
Department of Electrical and Computer Engineering, University of Toronto
Faculty of Law, University of Toronto
Schwartz Reisman Institute for Technology and Society

USENIX Security 2023

Privacy Policies

- Legal documents that disclose how a party collects, uses, and shares users’ data
- Required by legislation such as CalOPPA and GDPR
- Long and time-consuming to read & regulate at scale



Information We Collect

- Your Account Information: you must provide your mobile phone number and basic information (including a profile name of your choice) to create an account.
- Transactions And Payments Data: for purchases or other financial transactions, we process additional information about you, including payment account and transaction information.
- Device And Connection Information: we collect device and connection-specific information when you install, access, or use our Services.

Automated Analysis of Privacy Policies

Reading privacy policies is hard and time-consuming!

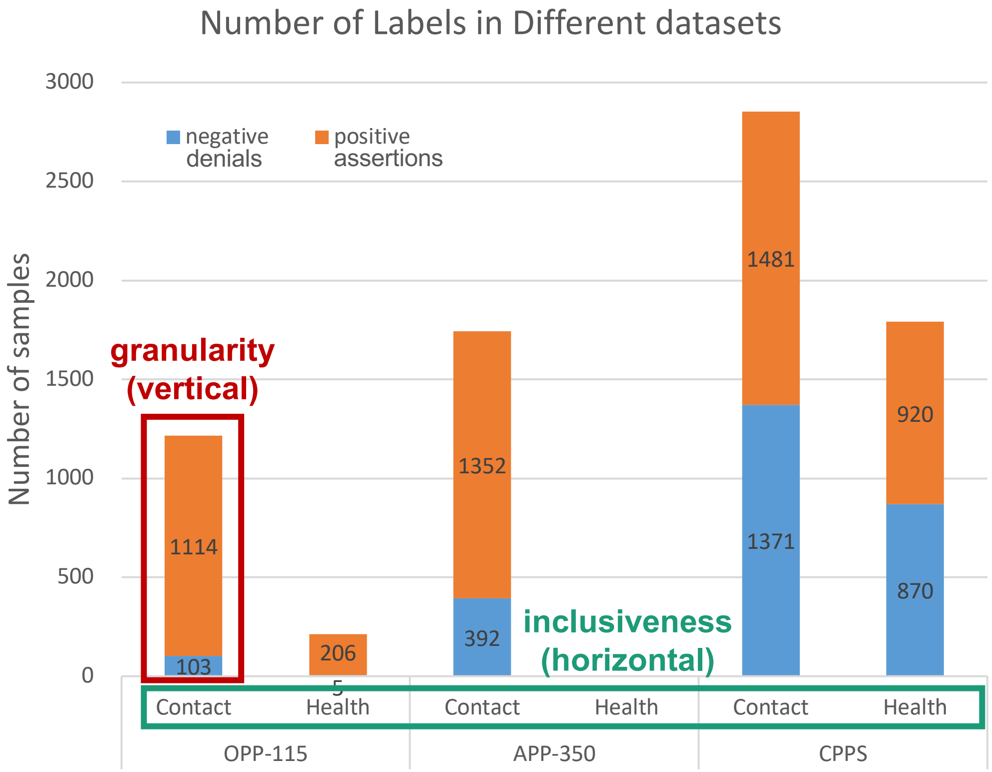
➔ automated privacy policy analysis tool

However, previous tools lack:

- Inclusiveness: does not cover rare data categories
  - contact category is common; health category is rare
- Granularity: does not include fine-grained labels
  - Does not differentiate between explicit denials ("we do not collect") vs. not mentioned

Challenges

1) Insufficient data:
   a) Privacy policy labels are expensive: annotated by human experts
2) Biased distribution:
   a) Horizontally cross data categories (inclusiveness): e.g. *contact* vs. *health* → few examples on rare categories
   b) Vertically across labels (granularity): e.g. assertion vs. denial → Too few examples of denials



Data imbalance + small dataset = problem!

Solutions

1) Active Learning
  - Selects the samples with the greatest uncertainty for labeling
  - Solves class imbalance

2) Crowdsourcing (Amazon Mechanical Turk)
  - a large group of participants to label privacy policies
  - Enables low-cost labeling

3) Segmentation:
  - Shorter text to label, easier tasks for mTurkers
  - Ensures label reliability: active learning requires 100% reliable oracles but crowdsourced annotators are unreliable

Privacy Policy Annotation

“While using our app, we access your personal information, namely, your ***email address, gender, age*** *and other public information, but we do not keep these data on our server. If you do not wish to provide this data, you can also opt-out this feature in your user setting. However, your email is still required to register your account.”*

Data category	1st party collect/use	3rd party sharing	Data storage
Demographic	Choice	Not mentioned	**Denial**
Contact	Assertion	Not mentioned	**Denial**

granularity

inclusiveness

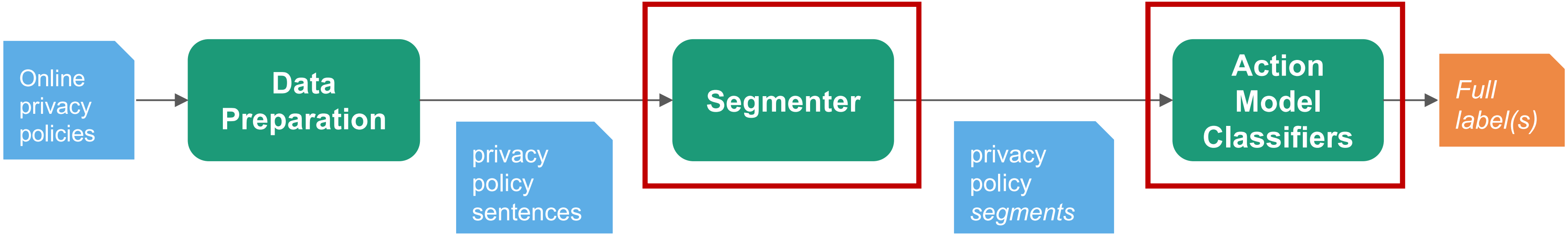
Data categories: contacts, geographic location, device, demographic, financial, health, survey, personal identifier, social media data

Data actions: first party collection/use, third-party sharing, data storage

Action modes: yes, no, opt-out, ambiguous, not mentioned

Calpric Pipeline

Crowdsourcing **A**ctive **L**earning
PRIvacy Policy **C**lassifier



Calpric: Segmenter

Full policy ➔ Sentences

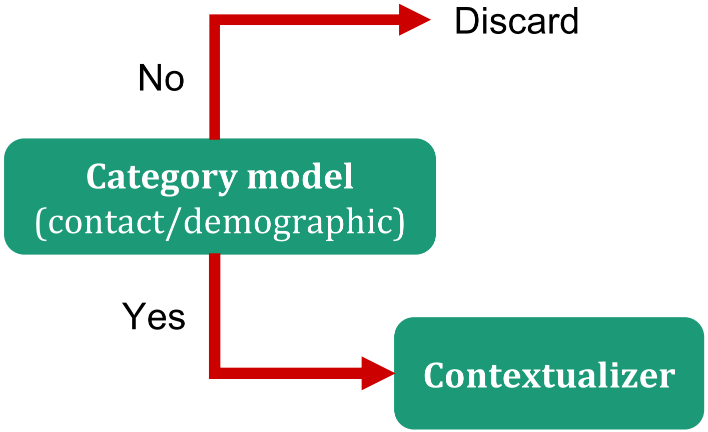
“…Your information allows us to offer you certain products and services, including the use of our website, to fulfill our obligations to you, to customize your interaction with our company and our website, and to allow us to suggest other products and services we think might interest you▌We generally store your data and transmit it to a third party for processing▌However, to the extent we process your data, we do so to serve our legitimate business interests (such as providing you with the opportunity to purchase our goods or services and interact with our website or mobile app).▌While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server.▌You can also opt-out this feature in your user setting. However, your email is still required to register your account▌ INTERNATIONAL DATA: Our website is hosted by servers located in the U.S.▌ Therefore, if you reside in the European Union, some of your data will be transferred internationally to those servers.▌ Transfers will be protected by appropriate safeguards, namely the EU-US Privacy Shield… ”

Sentence tokenizer

Calpric: Segmenter

Sentence

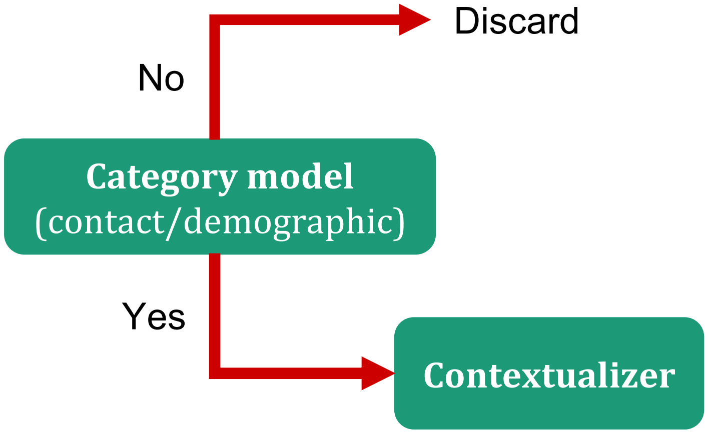
"…Your information allows us to offer you certain products and services, including the use of our website, to fulfill our obligations to you, to customize your interaction with our company and our website, and to allow us to suggest other products and services we think might interest you. We generally store your data and transmit it to a third party for processing. However, to the extent we process your data, we do so to serve our legitimate business interests (such as providing you with the opportunity to purchase our goods or services and interact with our website or mobile app). While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server. You can also opt-out this feature in your user setting. However, your email is still required to register your account. INTERNATIONAL DATA: Our website is hosted by servers located in the U.S. Therefore, if you reside in the European Union, some of your data will be transferred internationally to those servers. Transfers will be protected by appropriate safeguards, namely the EU-US Privacy Shield… "



Calpric: Segmenter

Sentence

While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server.



No

Discard

Category model
(contact/demographic)

Yes

Contextualizer

While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server.

Calpric: Segmenter

Sentence

“…Your information allows us to offer you certain products and services, including the use of our website, to fulfill our obligations to you, to customize your interaction with our company and our website, and to allow us to suggest other products and services we think might interest you. We generally store your data and transmit it to a third party for processing. However, to the extent we process your data, we do so to serve our legitimate business interests (such as providing you with the opportunity to purchase our goods or services and interact with our website or mobile app). While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server. You can also opt-out this feature in your user setting. However, your email is still required to register your account. INTERNATIONAL DATA: Our website is hosted by servers located in the U.S. Therefore, if you reside in the European Union, some of your data will be transferred internationally to those servers. Transfers will be protected by appropriate safeguards, namely the EU-US Privacy Shield… ”

Contextualizer

Calpric: Segmenter

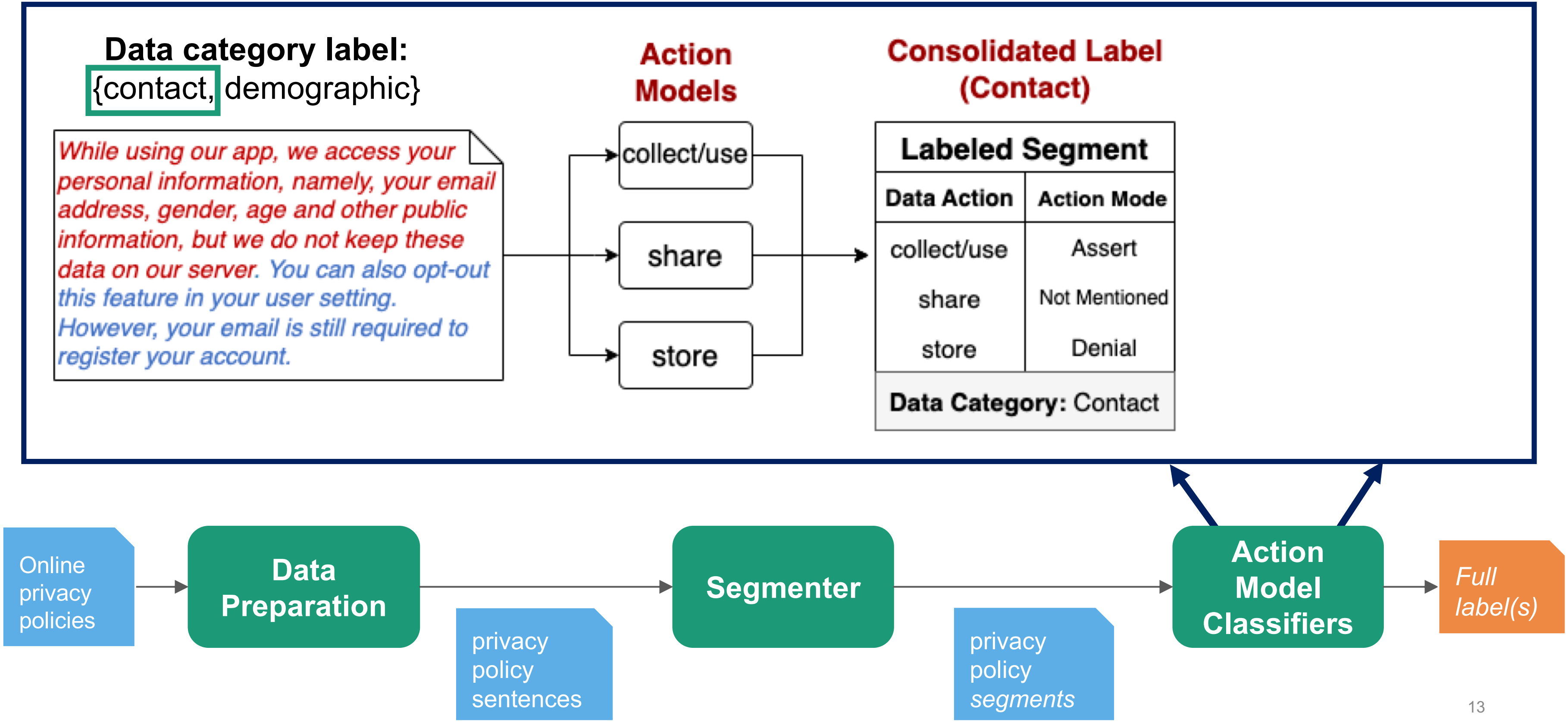
Segment

“…Your information allows us to offer you certain products and services, including the use of our website, to fulfill our obligations to you, to customize your interaction with our company and our website, and to allow us to suggest other products and services we think might interest you. We generally store your data and transmit it to a third party for processing. However, to the extent we process your data, we do so to serve our legitimate business interests (such as providing you with the opportunity to purchase our goods or services and interact with our website or mobile app). While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server. You can also opt-out this feature in your user setting. However, your email is still required to register your account. INTERNATIONAL DATA: Our website is hosted by servers located in the U.S. Therefore, if you reside in the European Union, some of your data will be transferred internationally to those servers. Transfers will be protected by appropriate safeguards, namely the EU-US Privacy Shield… ”

Contextualizer

Data category label:
{contact, demographic}

Calpric: Classifier

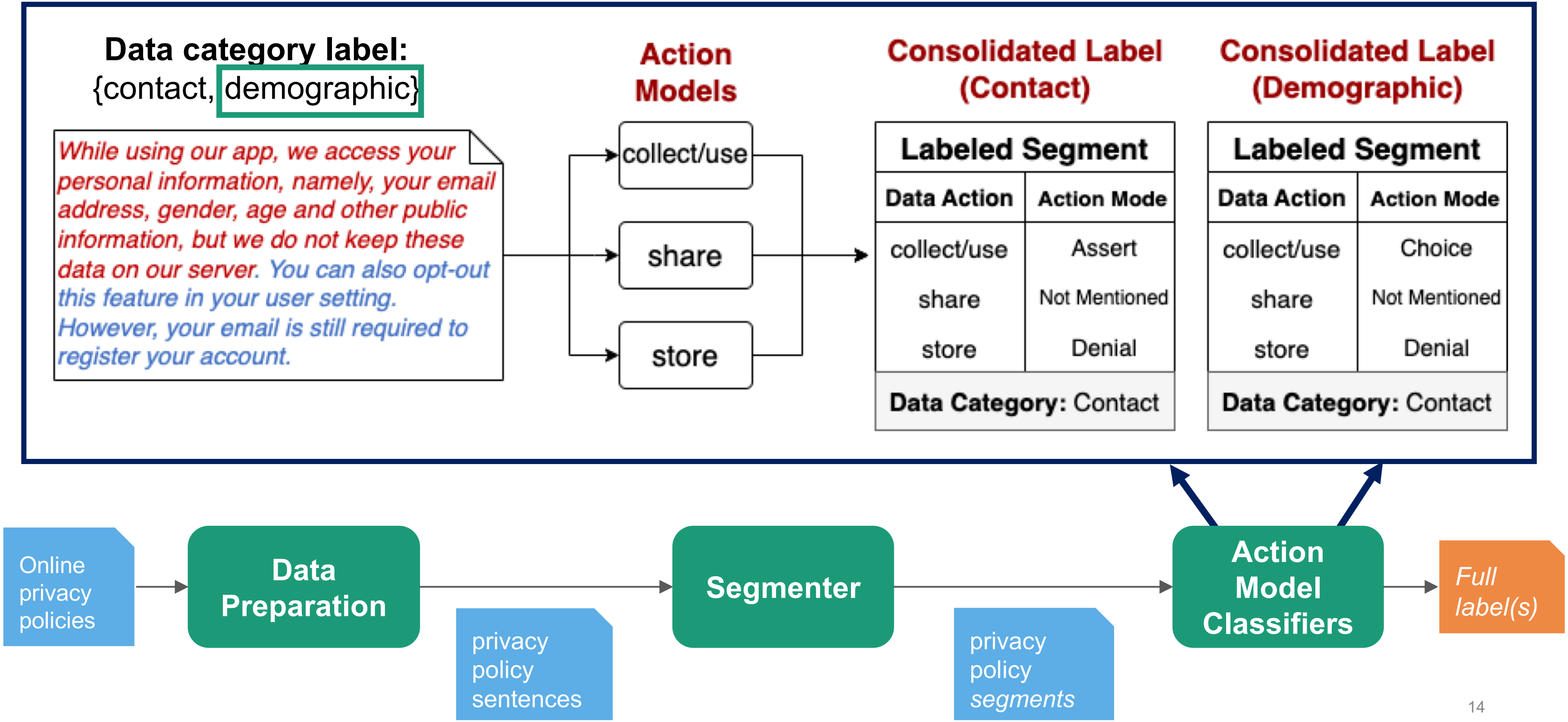


Data category label:
{contact, demographic}

While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server. You can also opt-out this feature in your user setting. However, your email is still required to register your account.

Labeled Segment	
Data Action	**Action Mode**
collect/use	Assert
share	Not Mentioned
store	Denial
Data Category: Contact	

Calpric: Classifier



Data category label:
{contact, demographic}

While using our app, we access your personal information, namely, your email address, gender, age and other public information, but we do not keep these data on our server. You can also opt-out this feature in your user setting. However, your email is still required to register your account.

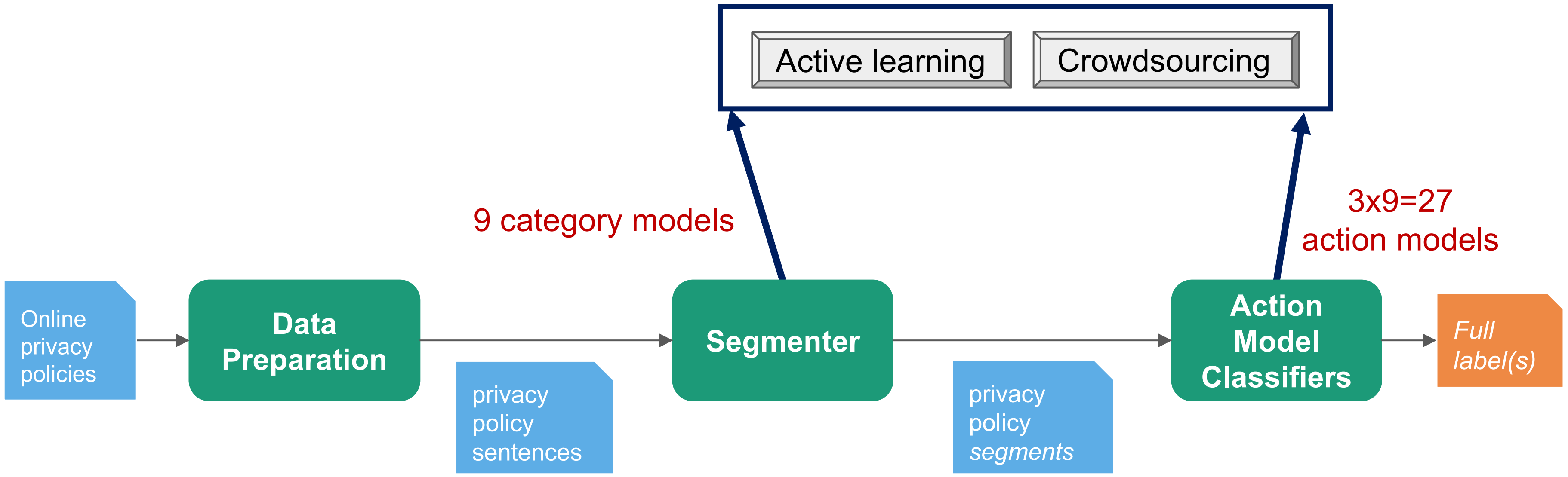
Consolidated Label (Contact)

Labeled Segment	
Data Action	**Action Mode**
collect/use	Assert
share	Not Mentioned
store	Denial
Data Category: Contact	

Consolidated Label (Demographic)

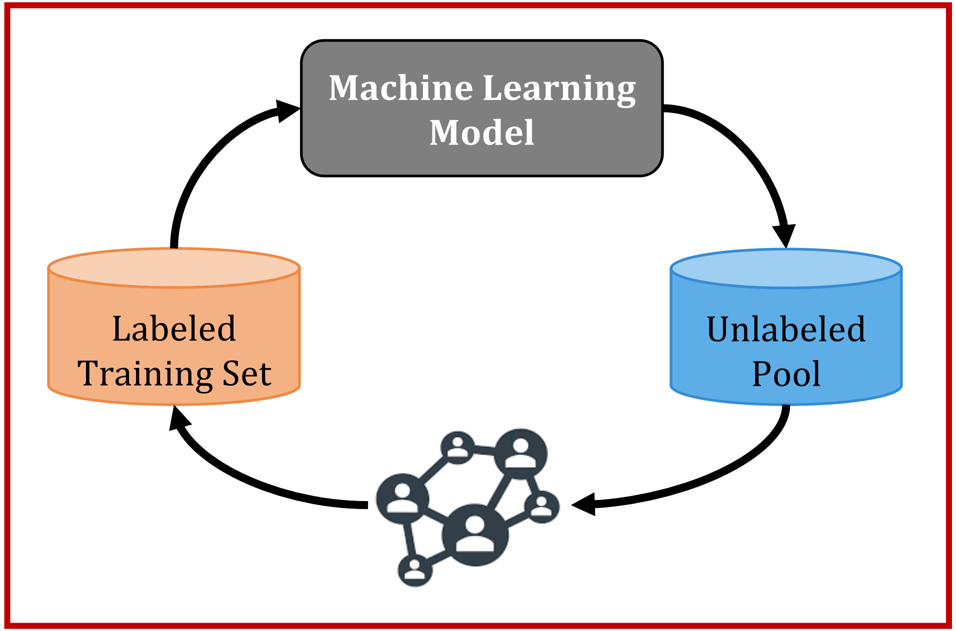
Labeled Segment	
Data Action	**Action Mode**
collect/use	Choice
share	Not Mentioned
store	Denial
Data Category: Contact	

Calpric Pipeline



Calpric Pipeline

Active learning



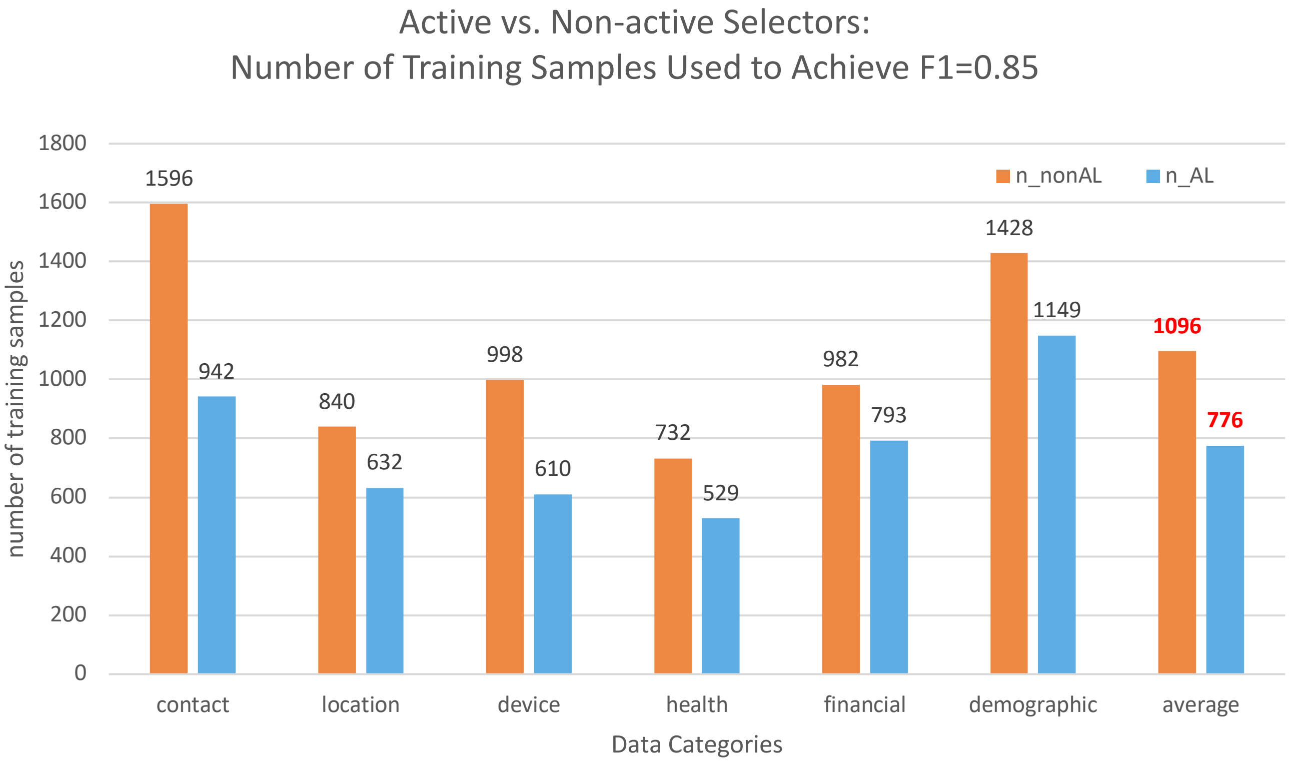
Machine Learning Model

Labeled Training Set

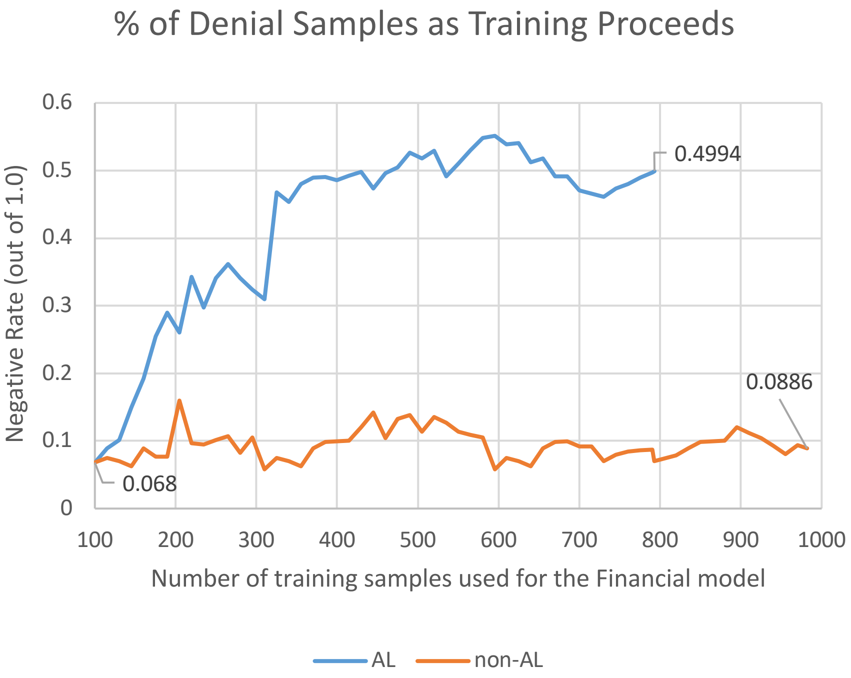
Unlabeled Pool

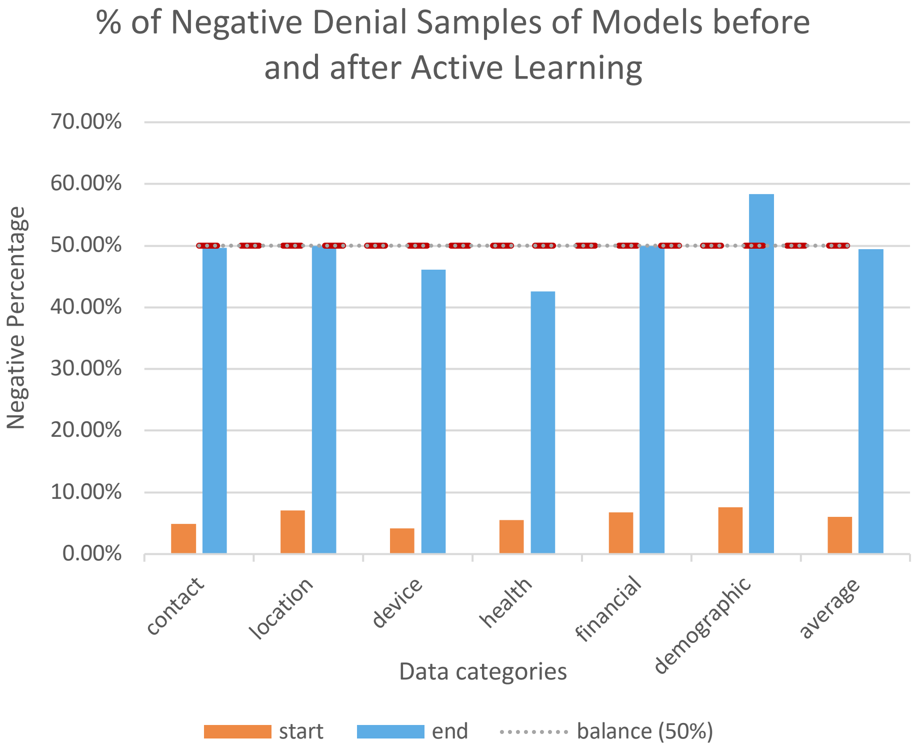
Crowdsourcing

Label Savings



Class Balance





Summary

Automated analysis of privacy policies:

- Annotation goal: inclusive and fine-grained
- Challenges: insufficient data, biased distribution
- Design considerations: class imbalance, labeling cost, label reliability
- Calpric: crowdsourcing active learning privacy policy classifier
  - Active learning: automatic text selection --> high accuracy with fewer training labels
  - Crowdsourcing: enable low-cost labeling
  - Segmentation: ensure label reliability
- Calpric and CPPS are available at: https://github.com/dlgroupuoft/Calpric
- Author web page: http://individual.utoronto.ca/wenjunqiu/