Bootstrapping Privacy Compliance in Big Data Systems (cont'd) + Inferring Data Associations in Black-Box Systems

Giulia Fanti
Based on slides by Anupam Datta
CMU

Fall 2019

Administrative

- HW2 will be released this week
  - Stay tuned
- Recitation on Friday (James)
  - More info about project categories
  - Open office hours
- Project proposals due next Friday, Sept. 20
  - Use Piazza to find partners!

Quiz on Canvas

- Take the quiz on your laptops/tablets/devices
- Please do not look back at your notes
- 10 minutes

Last time (continued)

Bootstrapping Privacy Compliance in Big Data Systems

S. Sen, S. Guha, A. Datta, S. Rajamani, J. Tsai, J. M. Wing

Proceedings of 35th IEEE Symposium on Security and Privacy

May 2014.

Formal Semantics

$$\frac{T^G \sqsubseteq T^C \quad \boxed{\exists_i D_i \ \ \textsf{denies} \ \ T^G}}{\textsf{ALLOW}\ T^C\ \textsf{EXCEPT}\ \boxed{D_1 \cdots D_m}\ \textsf{denies}\ \ T^G}\ (\mathrm{A}_2)$$

- Recursively check exceptions
  - ALLOW clauses have DENY clauses as exceptions
  - Top Level clause determines Blacklist/Whitelist

Designed for Precision

Policy Clause C	::=	$D \mid A$
Deny Clause D	::=	DENY $T_1 \cdots T_n$ EXCEPT $A_1 \cdots A_m$
		$\mid$ DENY $T_1 \cdots T_n$
Allow Clause A	::=	ALLOW $T_1 \cdots T_n$ EXCEPT $D_1 \cdots D_m$
		$\mid$ ALLOW $T_1 \cdots T_n$
Attribute T	::=	$\langle$attribute-name$\rangle$ $v_1 \cdots v_l$
Value v	::=	$\langle$attribute-value$\rangle$

TABLE I
GRAMMAR FOR LEGALEASE

$$\frac{T^G \not\sqsubseteq T^C}{\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m \;\; \text{denies} \;\; T^G} \quad (A_1)$$

$$\frac{T^G \sqsubseteq T^C \quad \exists_i D_i \;\; \text{denies} \;\; T^G}{\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m \;\; \text{denies} \;\; T^G} \quad (A_2)$$

$$\frac{T^G \sqsubseteq T^C \quad \forall_i D_i \;\; \text{allows} \;\; T^G}{\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m \;\; \text{allows} \;\; T^G} \quad (A_3)$$

$$\frac{\bot \in T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \;\; \text{allows} \;\; T^G} \quad (D_1)$$

$$\frac{\bot \notin T^G \sqcap T^C \quad \exists_i A_i \;\; \text{allows} \;\; T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \;\; \text{allows} \;\; T^G} \quad (D_2)$$

$$\frac{\bot \notin T^G \sqcap T^C \quad \forall_i A_i \;\; \text{denies} \;\; T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \;\; \text{denies} \;\; T^G} \quad (D_3)$$

TABLE III
INFERENCE RULES FOR LEGALEASE

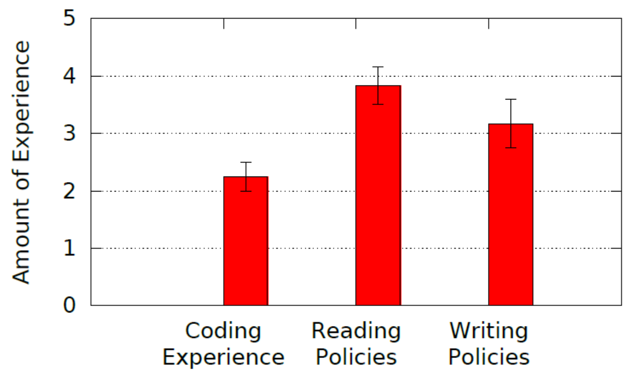
Designed for Expressivity (Bing, October 2013)

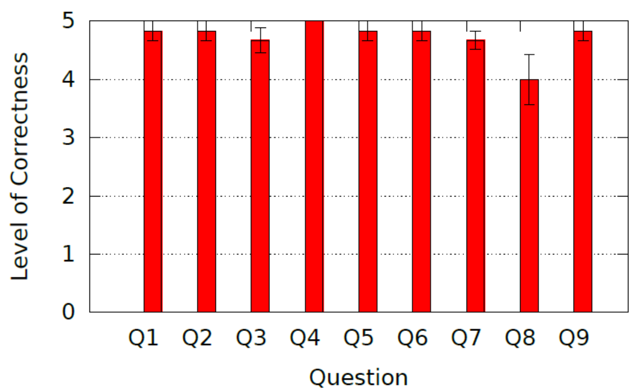
ALLOW	
EXCEPT	
DENY *DataType* IPaddress:Expired	◁ "we remove the entirety of the IP address after 6 months"
DENY *DataType* UniqueIdentifier:Expired	◁ "[we remove] cookies and other cross session identifiers, after 18 months"
DENY *DataType* SearchQuery, PII *InStore* Store DENY *DataType* UniqueIdentifier, PII *InStore* Store	◁ "We store search terms (and the cookie IDs associated with search terms) separately from any account information that directly identifies the user, such as name, e-mail address, or phone numbers."
DENY *DataType* BBEPData *UseForPurpose* Advertising	◁ "we do not use any of the information collected through the Bing Bar Experience Improvement Program to identify, contact or target advertising to you"
DENY *DataType* BBEPData, PII *InStore* Store	◁ "we take steps to store [information collected through the Bing Bar Experience Improvement Program] separately from any account information we may have that directly identifies you, such as name, e-mail address, or phone numbers"
DENY *DataType* BBEPData:Expired	◁ "we delete the information collected through the Bing Bar Experience Program at eighteen months."
DENY *DataType* UserProfile, PII *InStore* Store	◁ "we store page views, clicks and search terms used for ad targeting separately from contact information you may have provided or other data that directly identifies you (such as your name, e-mail address, etc.)."
DENY *DataType* PII *UseForPurpose* Advertising DENY *DataType* PII *InStore* AdStore	◁ "our advertising systems do not contain or use any information that can personally and directly identify you (such as your name, email address and phone number)."
DENY *DataType* SearchQuery *UseForPurpose* Sharing EXCEPT ALLOW *DataType* SearchQuery:Scrubbed	◁ "Before we [share some search query data], we remove all unique identifiers such as IP addresses and cookie IDs from the data."

Designed for Expressivity (Google, October 2013)

ALLOW EXCEPT	
DENY *DataType* PII *UseForPurpose* Sharing	◁ "We do not share personal information with companies, organizations and individuals outside of Google unless one of the following circumstances apply:"
EXCEPT ALLOW *DataType* PII:OptIn	◁ "We require opt-in consent for the sharing of any sensitive personal information."
EXCEPT ALLOW *AccessByRole* Affiliates	◁ "We provide personal information to our affiliates or other trusted businesses or persons to process it for us"
EXCEPT ALLOW *UseForPurpose* Legal	◁ "We will share personal information [if necessary to] meet any applicable law, regulation, legal process or enforceable governmental request."
DENY *DataType* DoubleClickData, PII EXCEPT ALLOW *DataType* DoubleClickData, PII:Optin	◁ "We will not combine DoubleClick cookie information with personally identifiable information unless we have your opt-in consent"

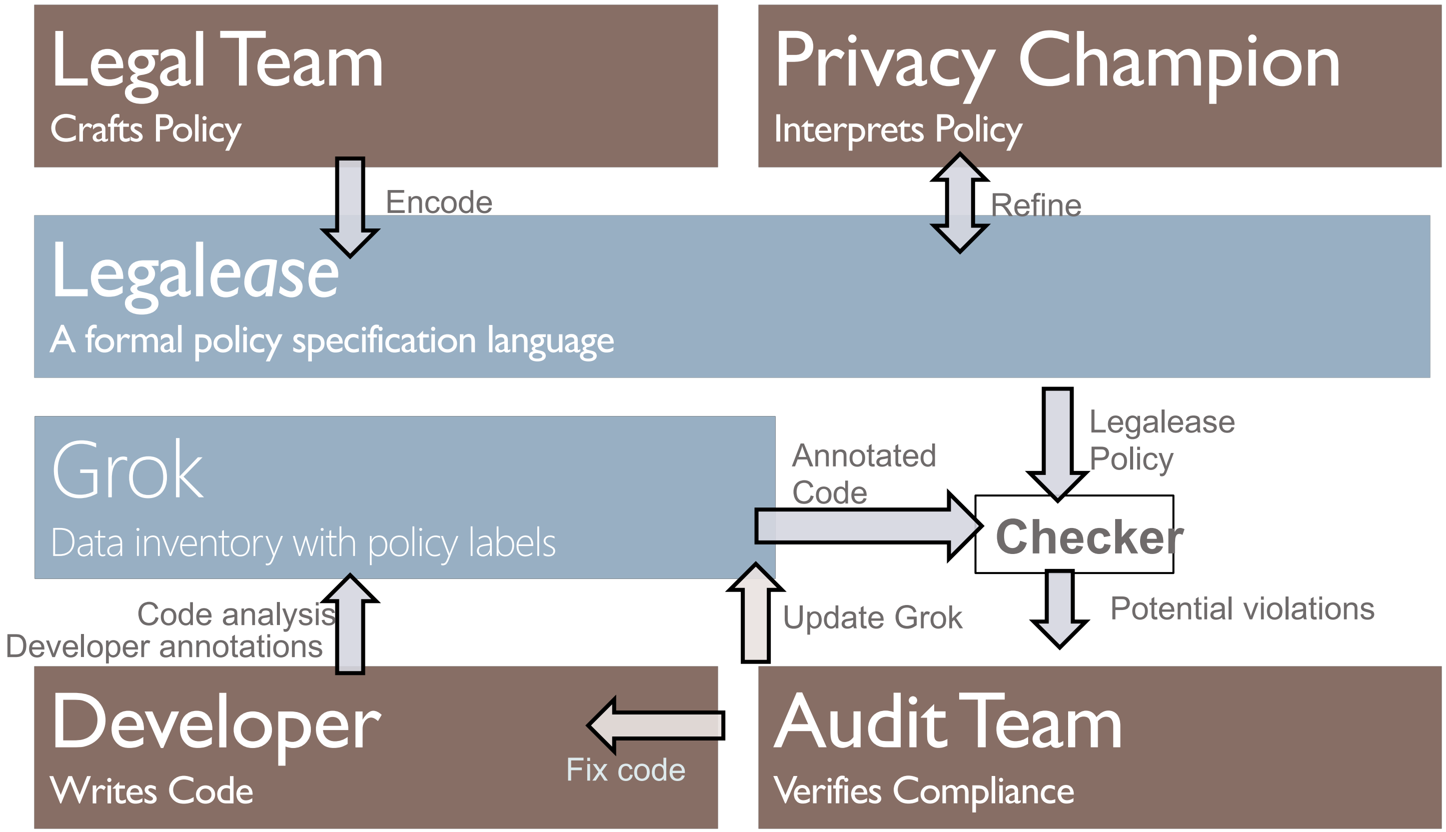
Legalease Usability



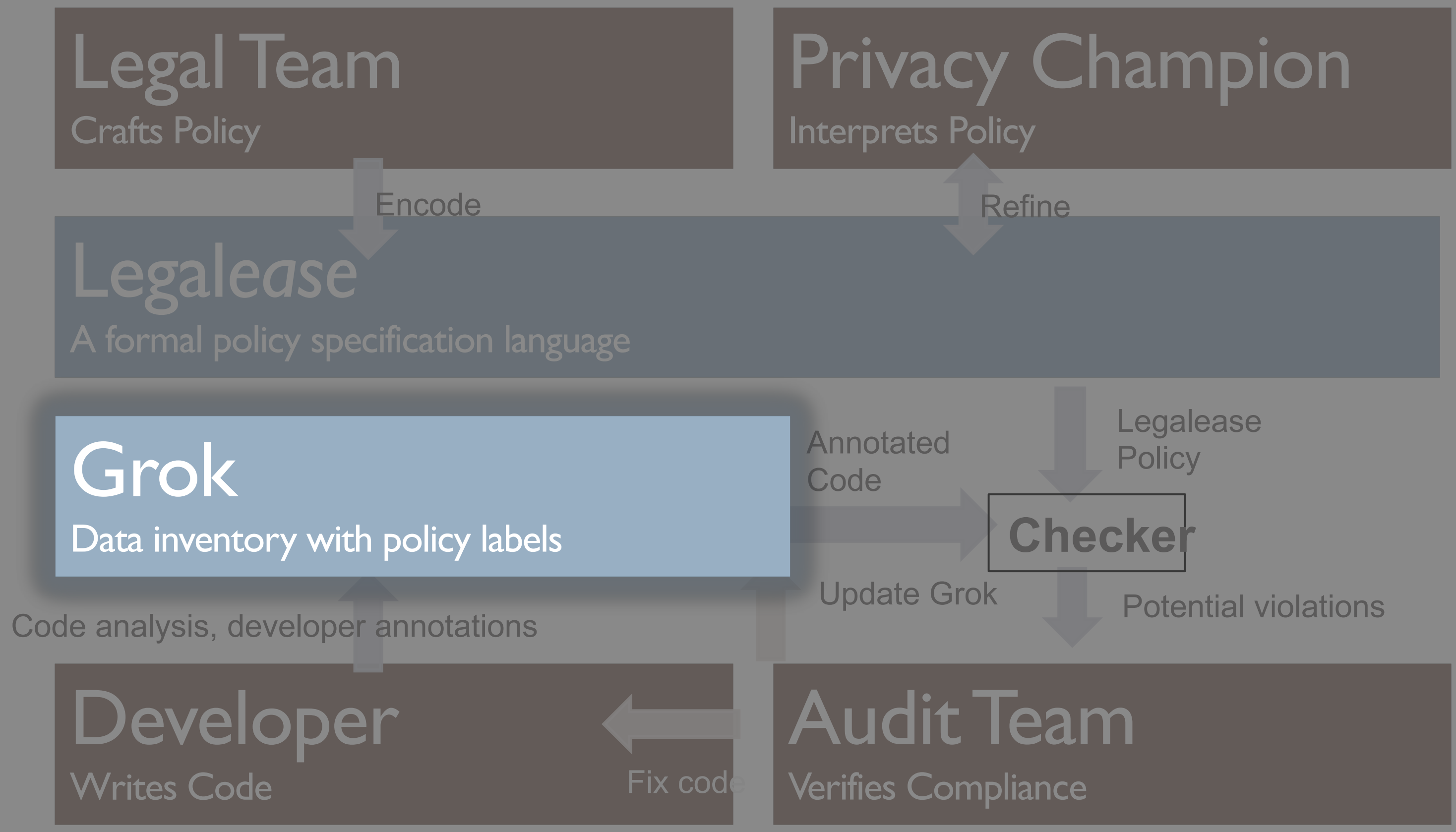


- Survey taken by 12 policy authors within Microsoft
  - Encode Bing data usage policy after a brief tutorial
- Time spent
  - 2.4 mins on the tutorial
  - 14.3 mins on encoding policy
- High overall correctness

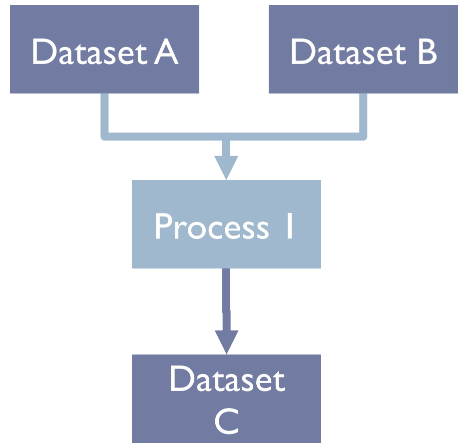
A Streamlined Audit Workflow



A Streamlined Audit Workflow



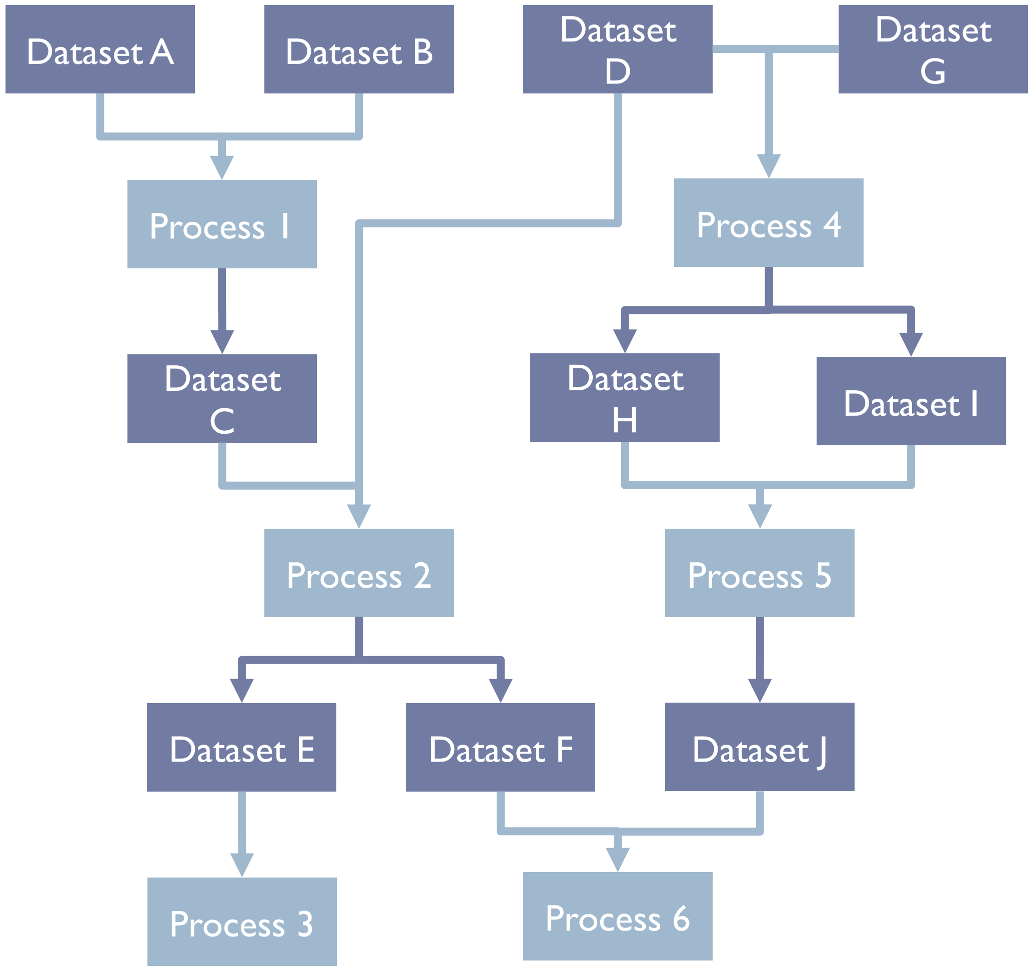
Map-Reduce Programming Systems



- Scope, Hive, Dremel
- Data in the form of Tables
- Code Transforms Columns to Columns
  - No Shared State
  - Limited Hidden Flows

```
users =
   SELECT _name, _age FROM datasetAB
user_tag =
   SELECT GenerateTag(_name, _age)
      FROM users
OUTPUT user_tag TO datasetC
```

Grok

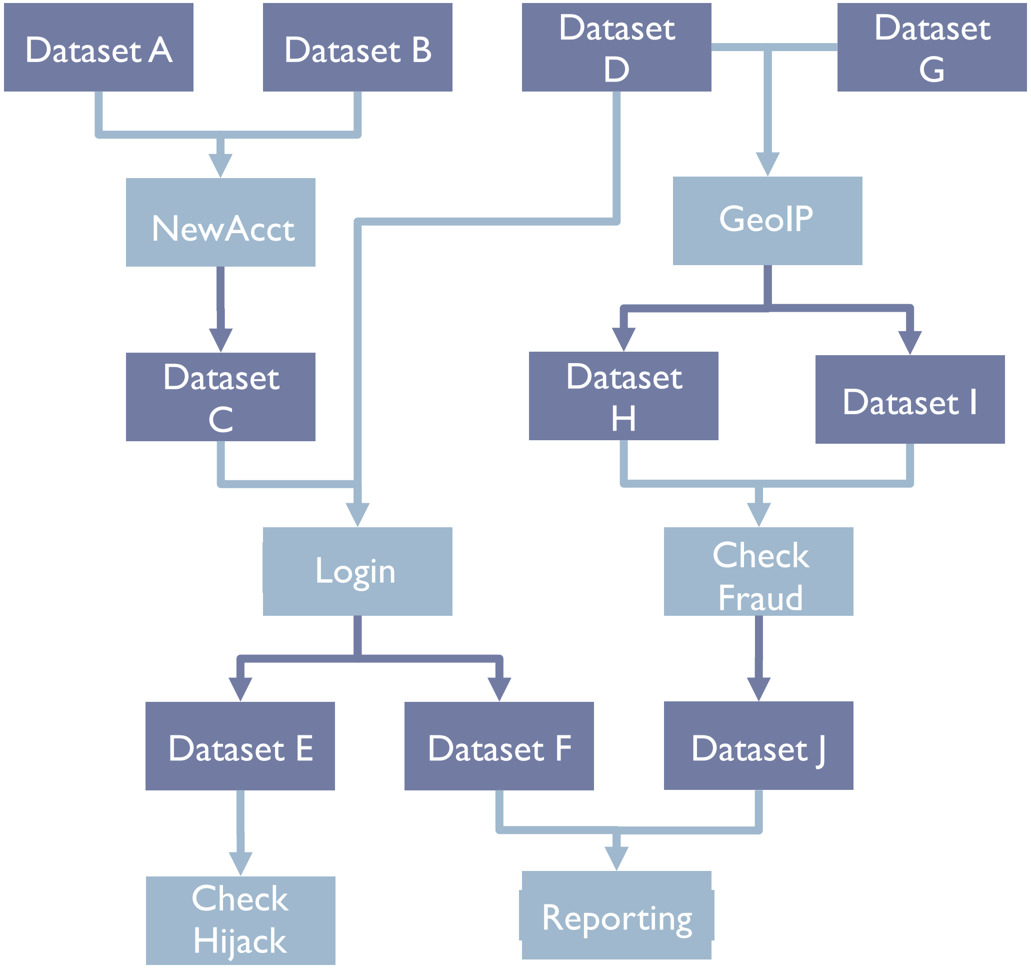


Dataset A
Dataset B
Dataset D
Dataset G
Process 1
Process 4
Dataset C
Dataset H
Dataset I
Process 2
Process 5
Dataset E
Dataset F
Dataset J
Process 3
Process 6

Grok

Purpose Labels

Annotate programs with purpose labels



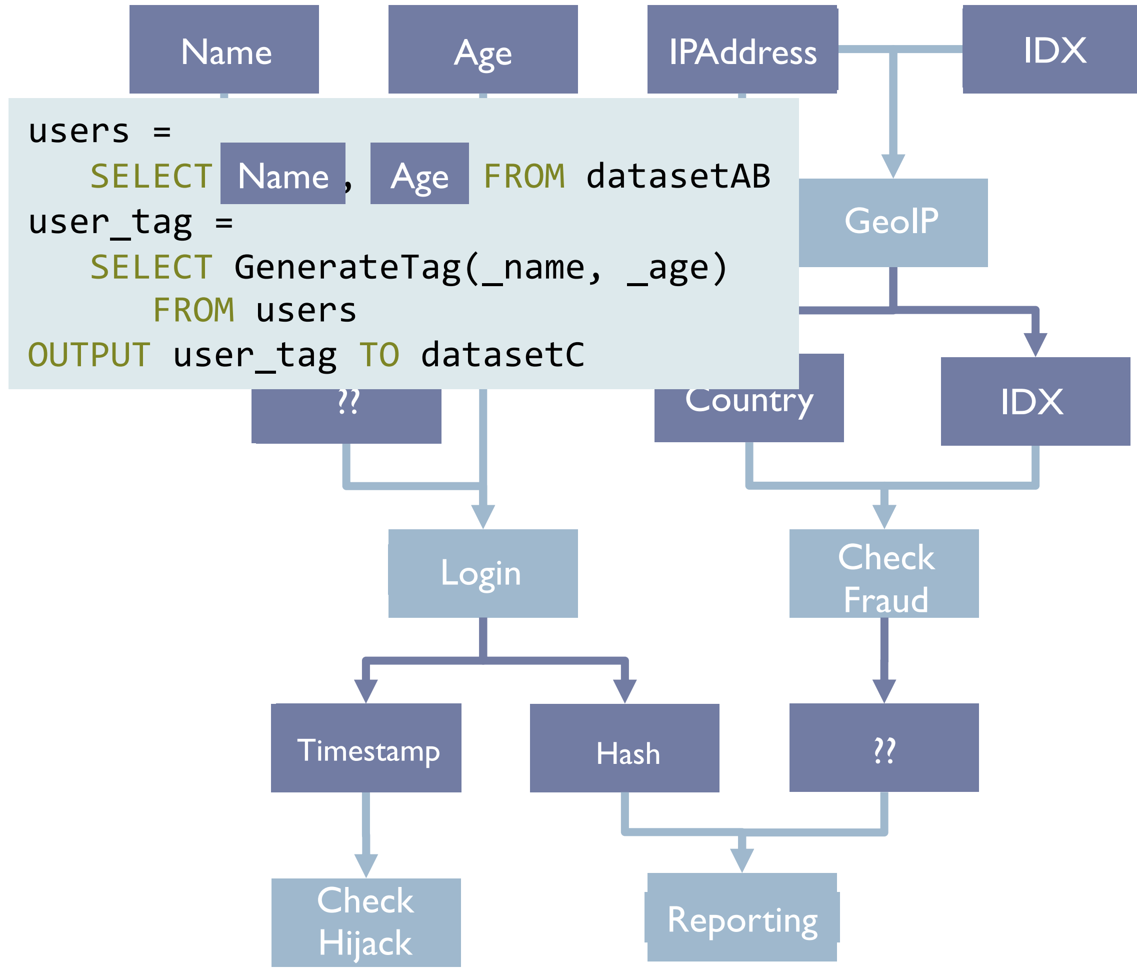
Grok

Purpose Labels

Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations




```
users =
   SELECT Name, Age FROM datasetAB
user_tag =
   SELECT GenerateTag(_name, _age)
      FROM users
OUTPUT user_tag TO datasetC
```

Grok

Purpose Labels

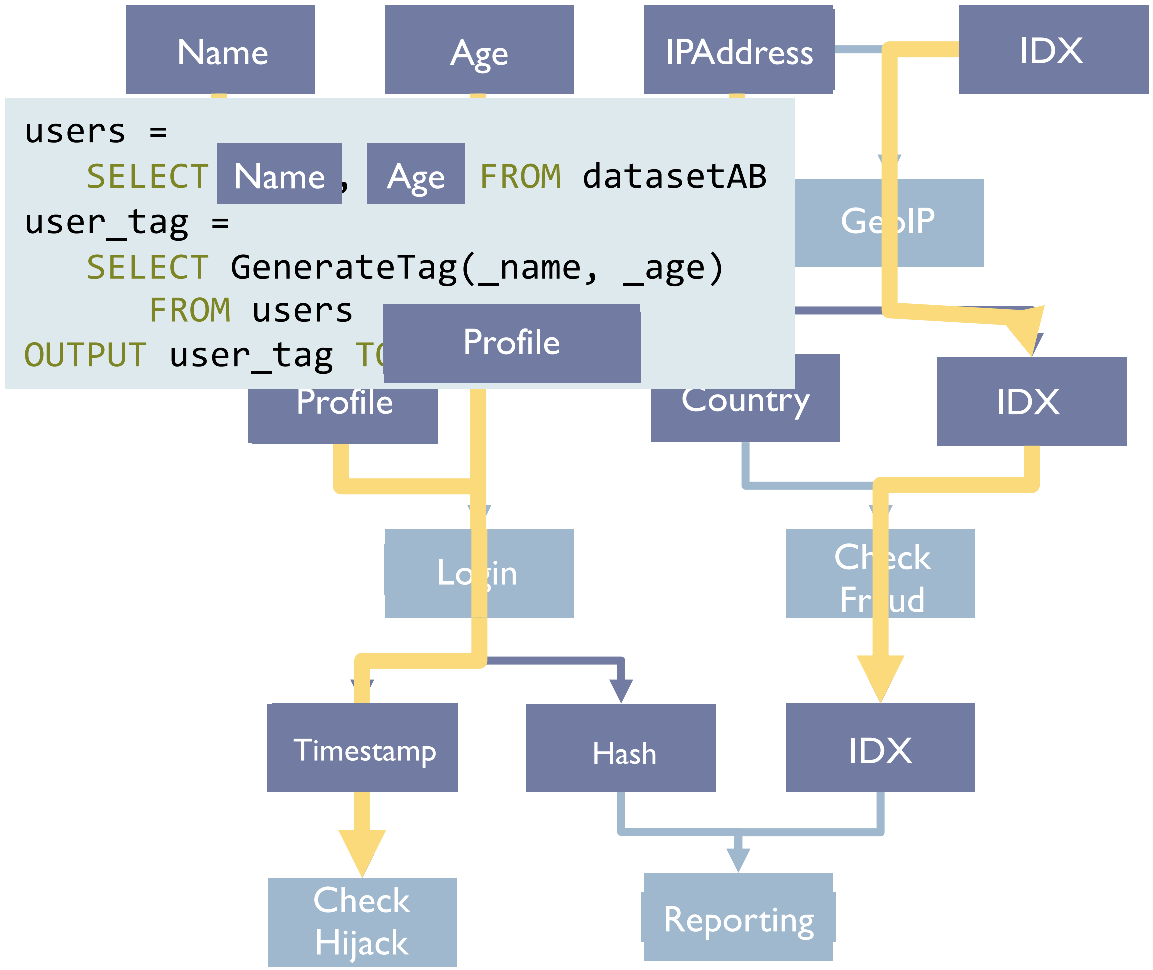
Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations

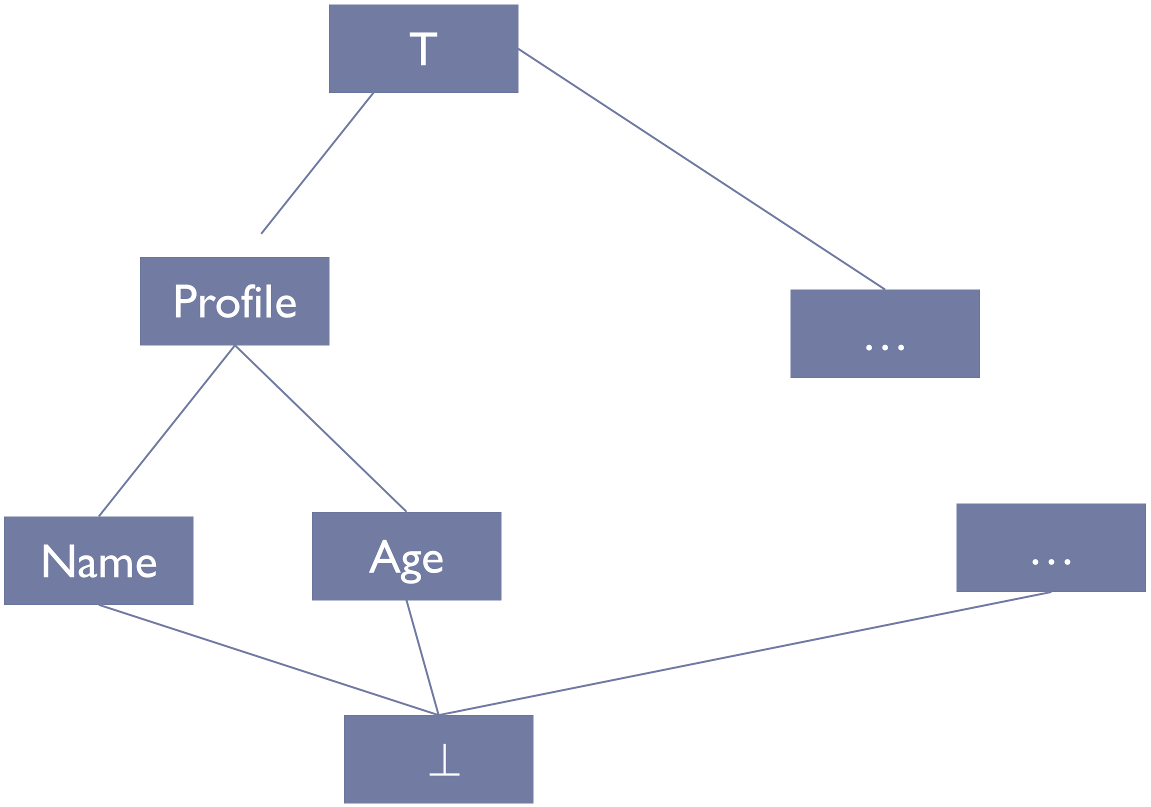
Flow Labels

Source labels propagated via data flow graph



D. E. Denning. "A lattice model of secure information flow"

A Lattice of Policy Labels



- If “Profile” use is allowed then so is everything below it
- If “Name” use is denied then so is everything above it

Implicit flows

```
users =
   SELECT Name , Age FROM datasetAB

users_35 =
   SELECT _name
      FROM users
      WHERE (_age > 35)

OUTPUT users_35 TO Profile
```

Beyond direct flows discussed in healthcare audit examples

Map-Reduce

Map

Operate on rows
in parallel
eg. filtering

Reduce

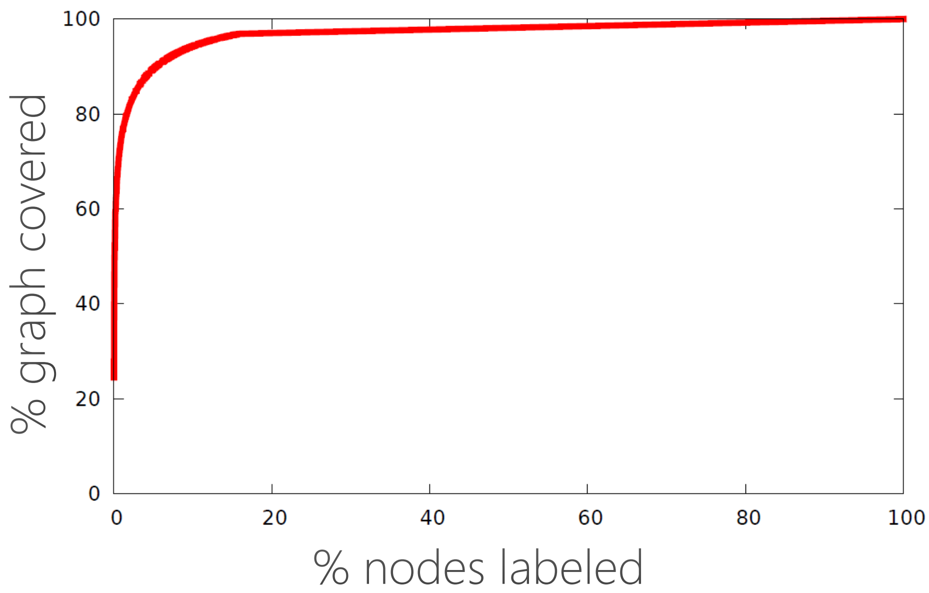
Combine groups of rows
eg. aggregation

```
users =
   SELECT Name, Age FROM datasetAB

users_35 =
   SELECT _name, _age
      FROM users
      WHERE (_age > 35)

ages_35 =
   SELECT _age, COUNT(_name) AS Profile
      FROM users_35
      GROUP BY _age

OUTPUT ages_35 TO datasetC
```

Combine Noisy Sources

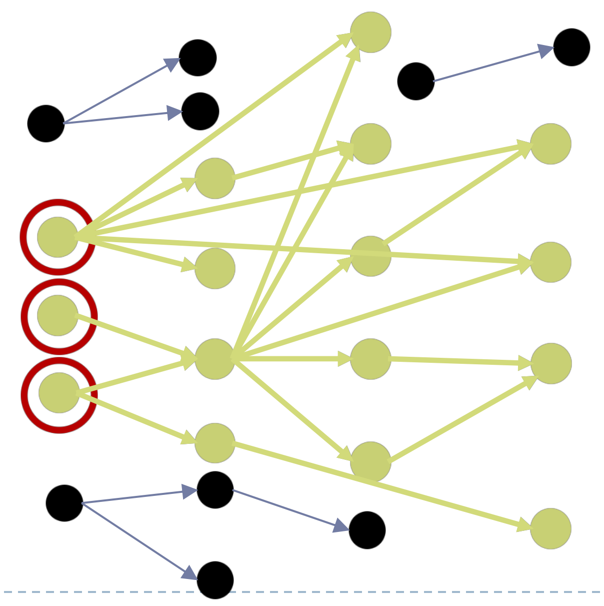
Variable Name Analysis	Developer Annotations	Auditor Verification
Carefully curated regular expressions Leverages developer conventions Significant Noise	Expensive Low Noise	Very Expensive Definitive Need very few of these

Why Bootstrapping Grok Works



Pick the nodes which will label the most of the graph

~200 annotations label 60% of nodes



A small number of annotations is enough to get off the ground.

Scale

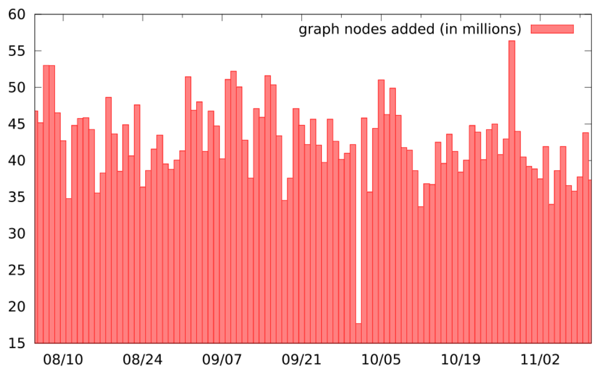
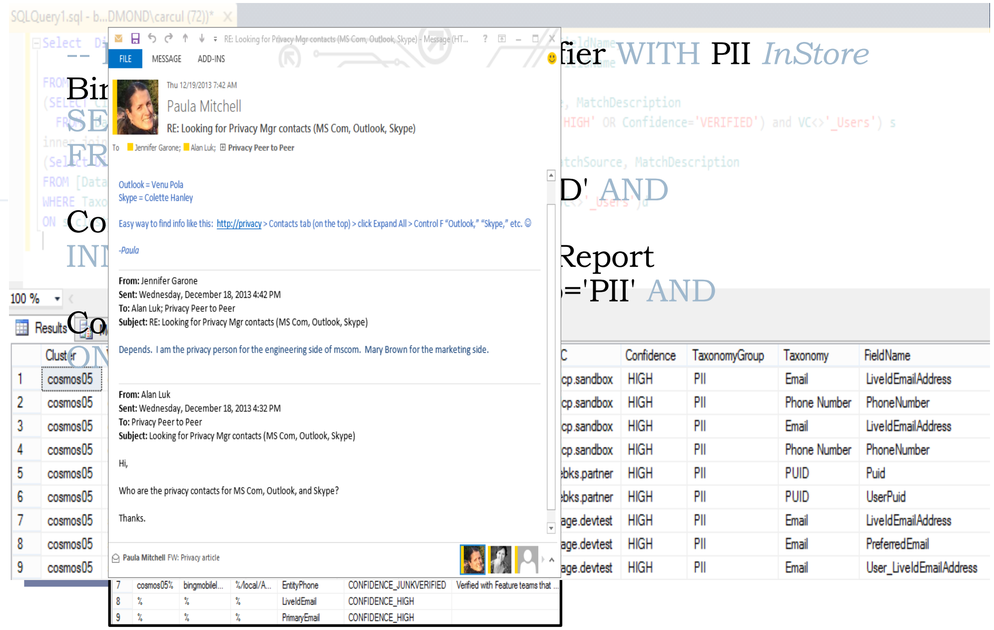


Fig. 9. Number of GROK data flow graph nodes added each day

- 77,000 jobs run each day
  - By 7000 entities
  - 300 functional groups
- 1.1 million unique lines of code
  - 21% changes on avg, daily
  - 46 million table schemas
  - 32 million files
- Manual audit infeasible
- Information flow analysis takes ~30 mins daily

Nightly Compliance Process



Static code analysis

schemas / files

23M+ / 25M+

Generate report

privacy / audit

candidates / alerts / items*

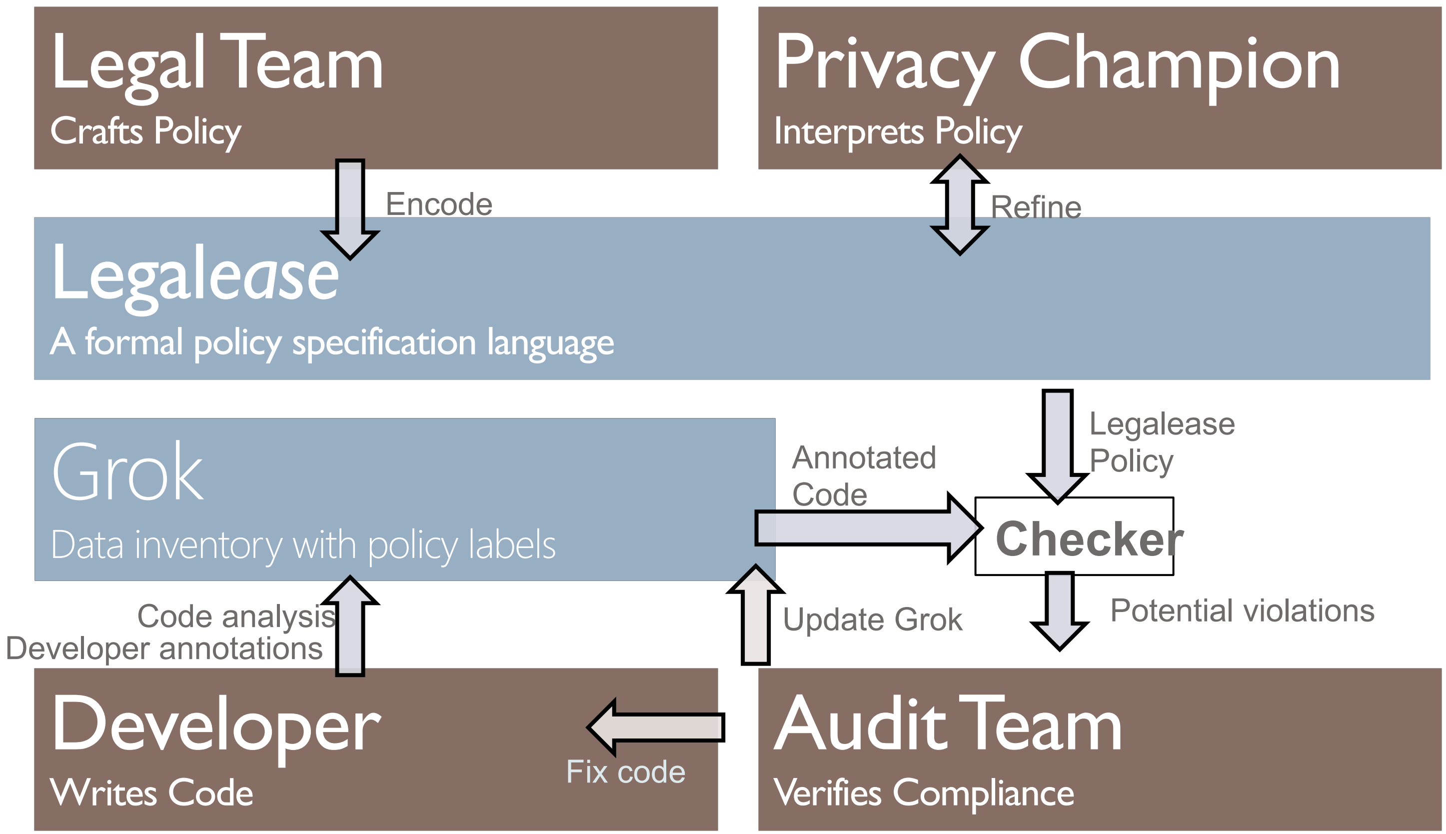
300K+ / 100K+

Manual Audit

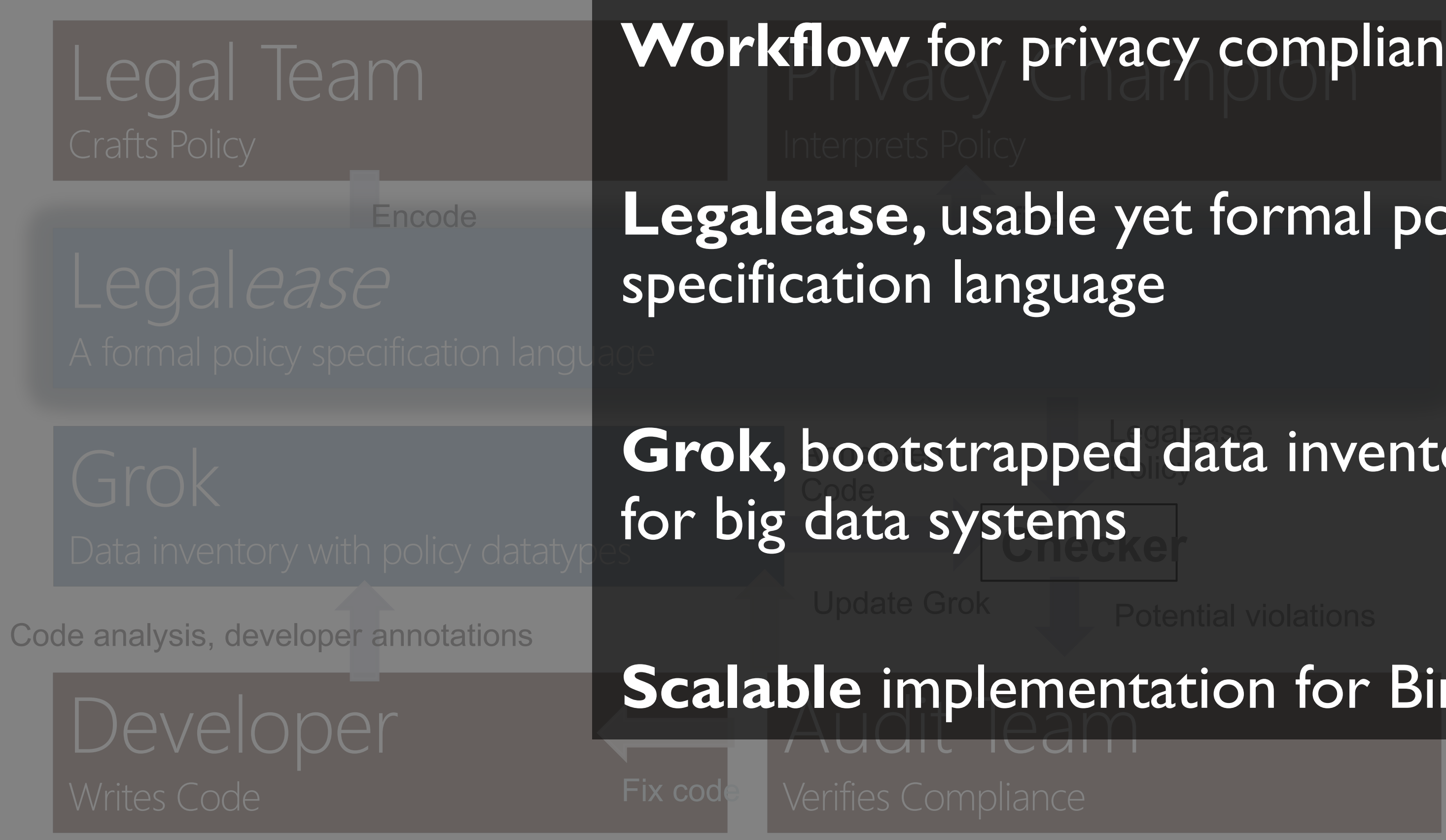
teams

8

A Streamlined Audit Workflow



A Streamlined Audit Workflow



Workflow for privacy compliance

Legalease, usable yet formal policy specification language

Grok, bootstrapped data inventory for big data systems

Scalable implementation for Bing

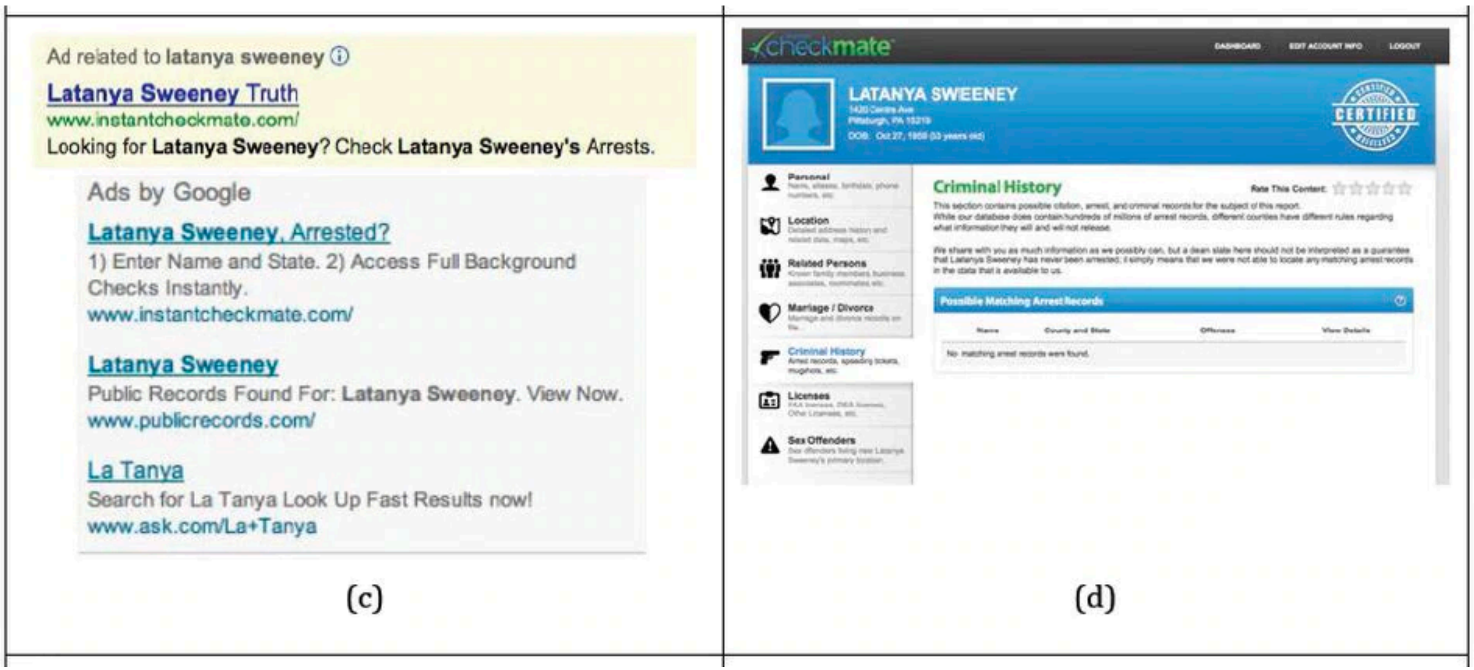
Part II: Inferring Data Usage of Black-Box Systems

So far

- Technique for auditing privacy policies automatically
- Given access to:
  - Developers
  - Code
  - Privacy advocates in the company
- This is really for companies to audit themselves
  - Maybe law enforcement

What if we don’t have access?

- LaTanya Sweeney



What was hard about this study?

- Manual ad checking
  - Limits scale of the study
- She knew what she was looking for
  - Associations between black-sounding names and ads for arrest records
  - Limits scope of the study

Next Up

XRay: Enhancing the Web's Transparency with Differential Correlation

M. Lecuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, R. Geambasu

Proceedings of 2rd USENIX Security Symposium

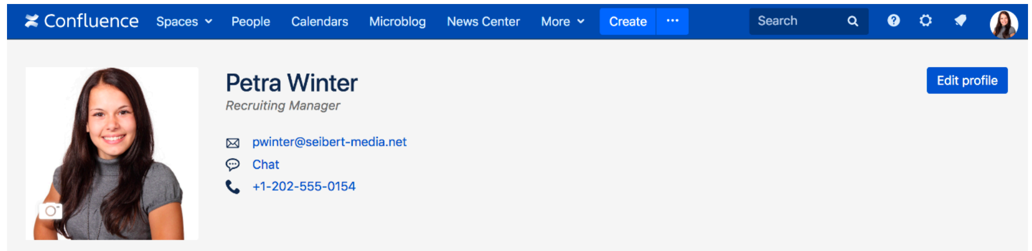
August 2014.

Goals

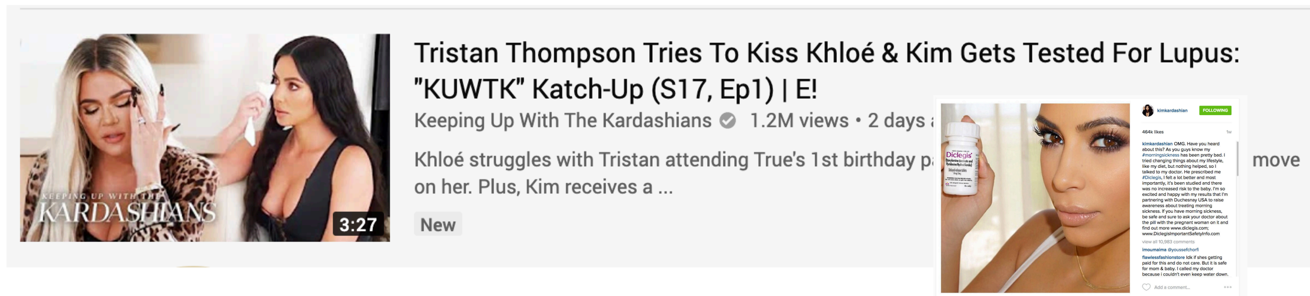
- Fine-grained and accurate data tracking
  - Detect which inputs (e.g., emails) likely triggered which outputs (e.g., ads)
- Scalability
  - E.g., track past month's emails
- Extensibility, generality, self-tuning
  - Limited manual tuning when you switch to general websites

Forms of Targeting

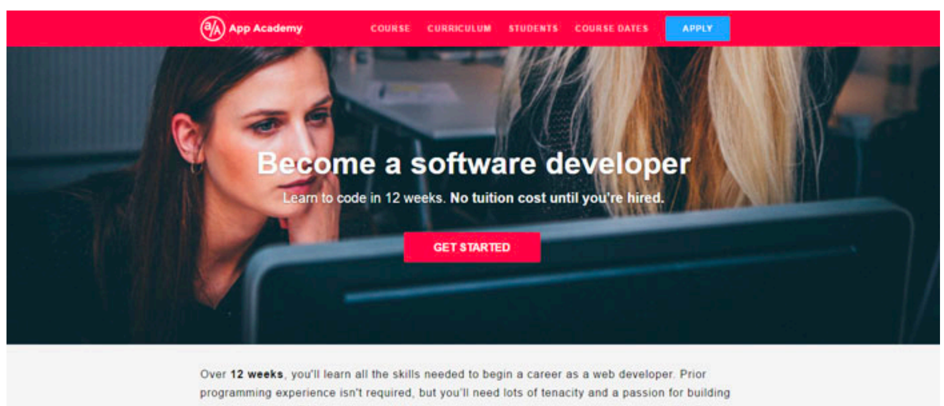
1) Profile Targeting



2) Contextual Targeting



3) Behavioral Targeting



XRay Architecture

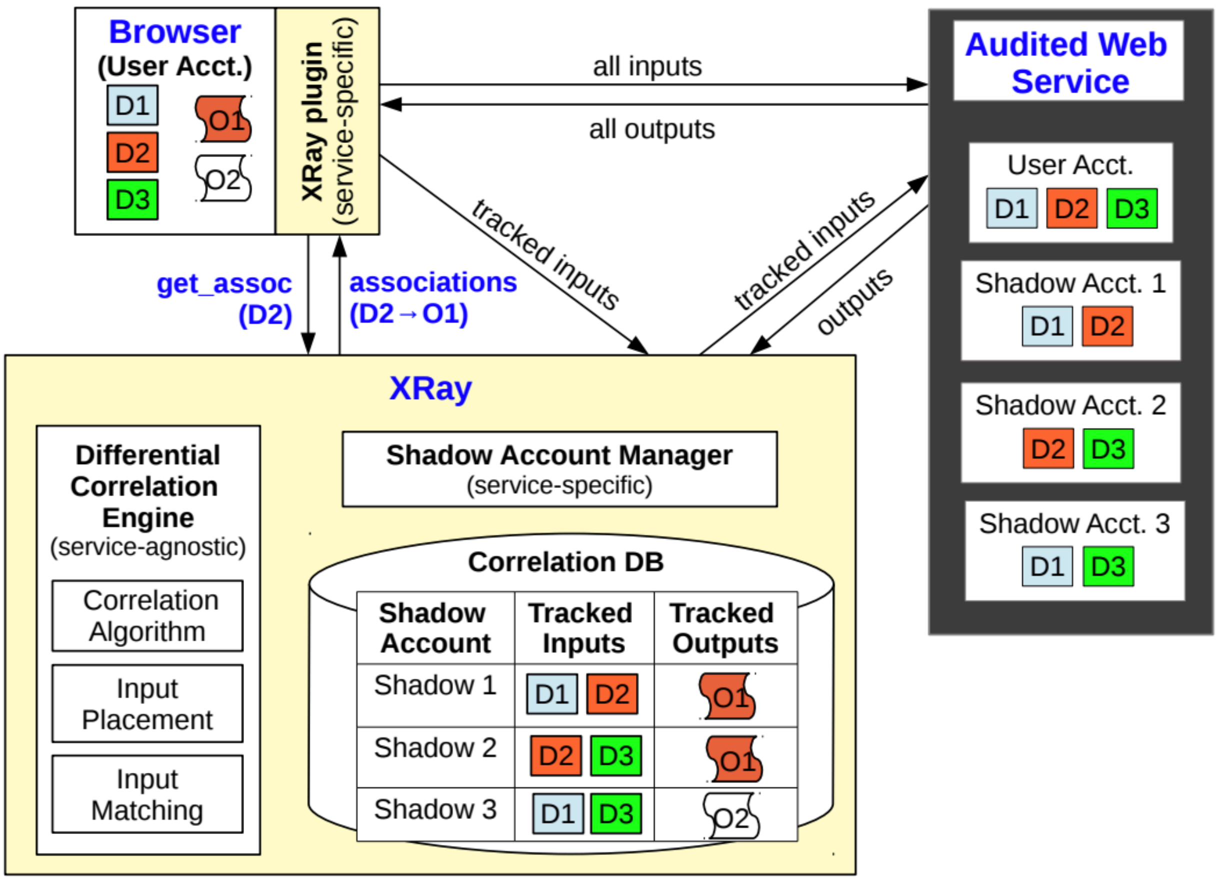


Figure 2: **The XRay Architecture.**

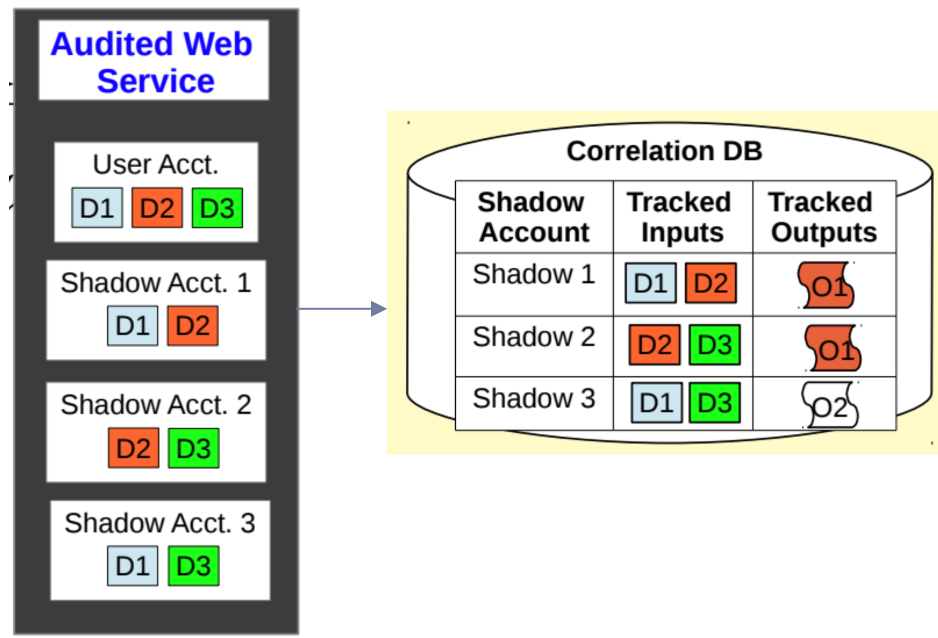
Browser Plugin

- Tracks specific DOM elements in audited services' web pages
- Which elements to track is configuration setting
  - E.g., Gmail
  - Inputs: Emails
  - Outputs: Ads

Shadow Account Manager

- (1) Populate shadow accounts with subsets of user account's tracked inputs
- (2) Periodically retrieves outputs from each audited service for each shadow account

- These are service-specific
- E.g. Gmail
  - Send emails with SMTP
  - Call the ad API



Differential Correlation Engine

- Analyzes correlations in the Correlation DB
- Plugin makes a `get_assoc` request
  - Look up entry in Correlation DB, return pre-computed associations
  - If none found, return `unknown`
- Periodic updates

How do we detect a correlation?

- Naïve solution:
  - Create shadow account with every possible combination of inputs
  - Q: If I have N initial inputs and M initial outputs, how many shadow accounts do I need?

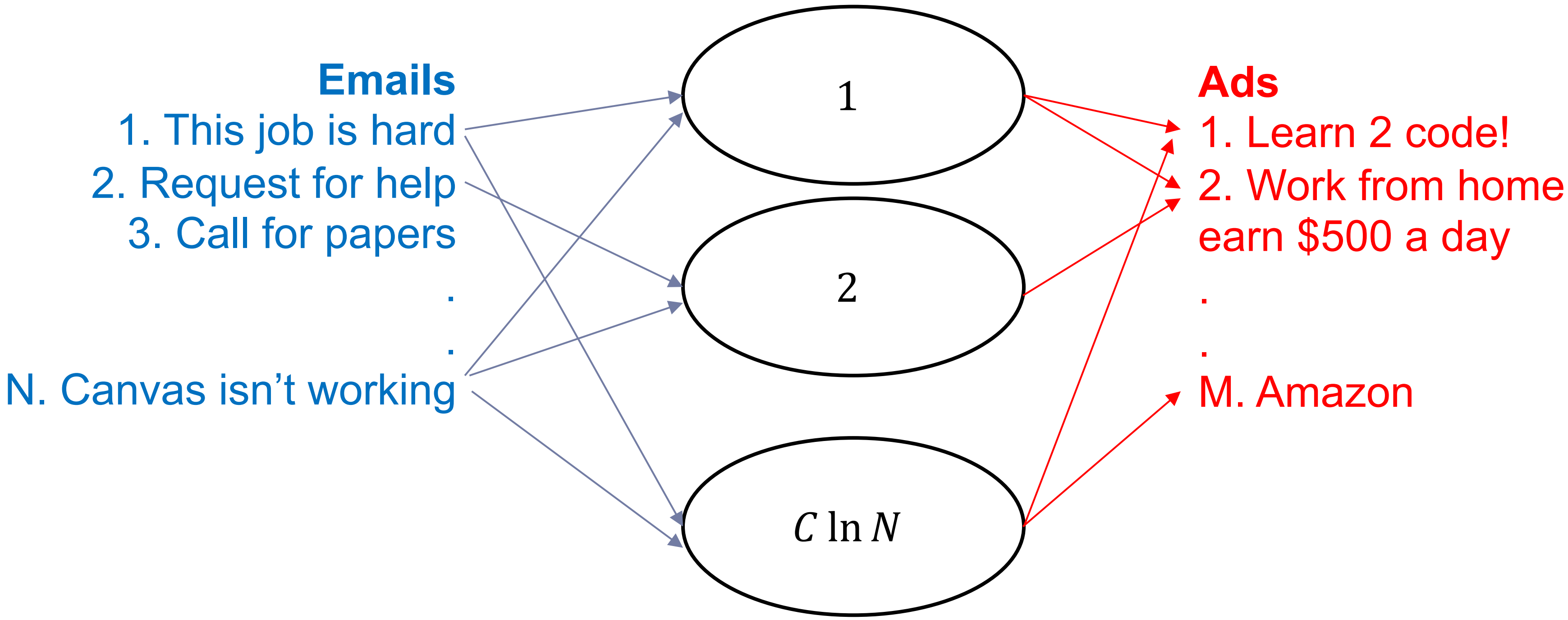
Emails
1. Subject: This job is hard
2. Subject: Request for help
3. Subject: Call for papers
.
.
N. Subject: Canvas isn't working

Ads
1. Learn 2 code!
2. Work from home, earn $500 a day
.
.
M. Amazon

  - A: 2^N. We want every possible subset of inputs

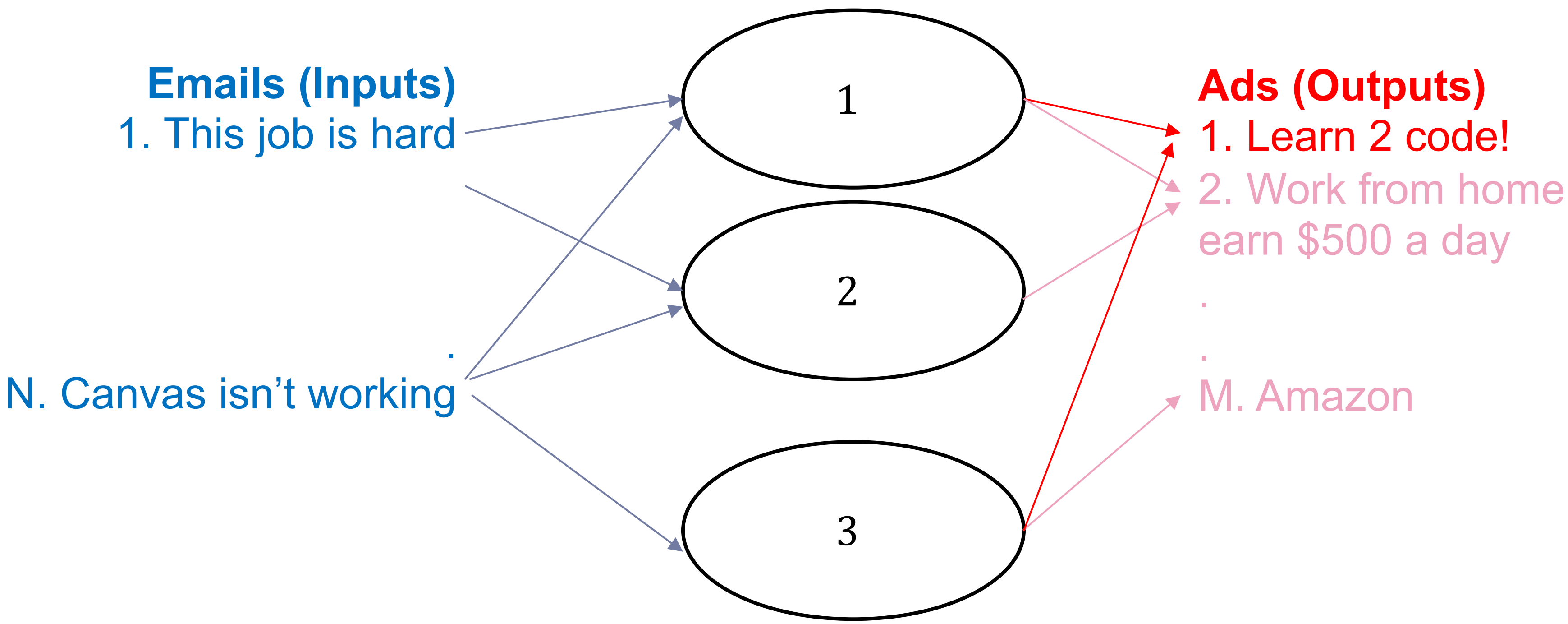
Instead: Set Intersection

- Create $C \ln N$ shadow accounts
- Pick probability $\alpha \in (0,1)$
- Randomly place each input into each shadow account w.p. α



Instead: Set Intersection

- Given output O_k:
  - Compute set A_k of active accounts that saw O_k
  - Compute inputs that appears in fraction β of active accounts
  - Return set of accounts iff $\geq \beta$ contain all remaining inputs



Why should this work?

- Key idea: argue that every non-targeting input would have a vanishingly small probability of being in a significant fraction of active accounts

- Try to prove this yourself before next class

- Connections to the idea of group testing
  - Technique from WWII for blood testing

Extension

- To get rid of parameter tuning (C, α, β), they introduce Bayesian inference-based detection mechanism

- Behavioral Targeting
  - Defines a generative model for observations, computes likelihood
  - Uses same method of data collection as before
- Contextual targeting
  - Compute likelihood based on assumptions about
    - p_{in} = P(see ad | targeted input is present)
    - p_{out} = P(see ad | targeted input is not present)
    - p_0 = P(see ad | no targeting)
  - Iteratively train parameters, then likelihoods
- Composite model
  - Arithmetic mean of scores

Experimental Methods

- Implemented in 3,000 lines of Ruby
  - Google, YouTube, and Amazon
  - Service-specific shadow account manager
    - ~500 lines of code each
- Ground truth exists for ads on Amazon and YouTube
  - "Why recommended"
- Google labelled manually

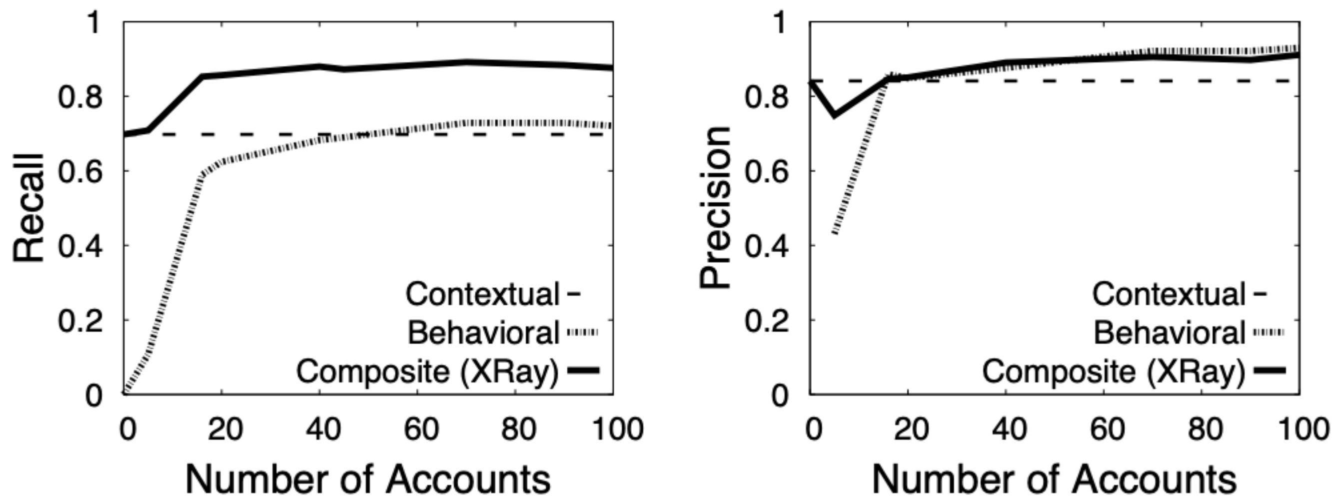
Results: Self-Targeted Ads (Sanity Check)

- Check for Gmail targeting via AdWords

Ad Keyword	Targeted Email	Detected by XRay?	XRay Scores	# Accounts & Displays
Chaldean Poetry	Like Chaldean Poetry?	Yes	0.99, 1.0	13/13, 1588/1622
Steampunk	Fan of Steampunk?	Yes	0.99, 1.0	13/13, 888/912
Cosplay	Discover Cosplay.	Yes	0.99, 1.0	13/13, 440/442
Falconry	Learn about Falconry.	Yes	0.99, 1.0	13/13, 1569/1608

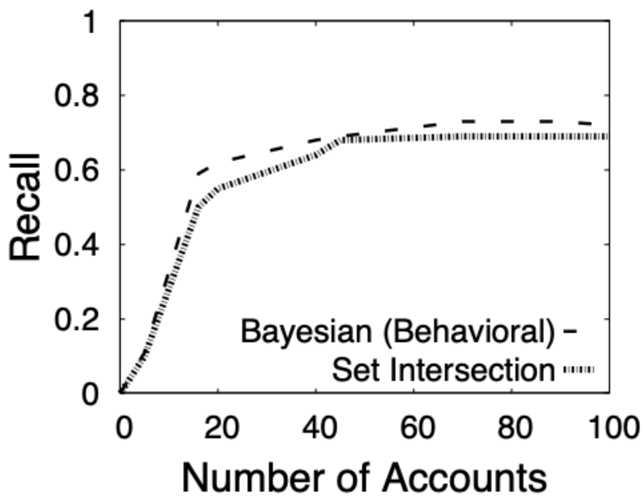
Bayesian Model Accuracy

- Experiment on Gmail

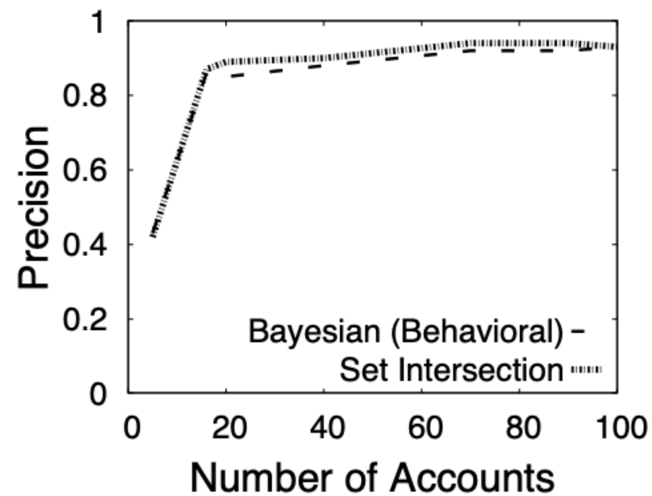


(a) **Recall** (b) **Precision**

Bayesian vs. Set Intersection Comparison



(a) **Recall**



(b) **Precision**

Results: Examples of Targeted Ads

Topic	Targeted Ads	XRay Scores	# Accounts & Displays
Alzheimer	Black Mold Allergy Symptoms? Expert to remove Black Mold.	0.99, 0.05	9/9, 61/198
	Adult Assisted Living. Affordable Assisted Living.	0.99, 0.99	8/8, 12/14
Cancer	Ford Warriors in Pink. Join The Fight.	0.96, 0.98	9/9, 1022/1106
	Rosen Method Bodywork for physical or emotional pain.	0.98, 0.05	7/7, 24/598
Depression	Shamanic healing over the phone.	0.99, 0.99	16/16, 117/117
	Text Coach - Get the girl you want and Desire.	0.93, 0.04	7/7, 31/276
African American	Racial Harassment? Learn your rights now.	0.99, 0.2	10/10, 851/5808
	Racial Harassment, Hearing racial slurs?	0.99, 0.2	10/10, 627/7172
Homosexuality	SF Gay Pride Hotel. Luxury Waterfront.	0.99, 0.1	9/9, 50/99
	Cedars Hotel Loughborough, 36 Bedrooms, Restaurant, Bar.	0.96, 1.0	8/8, 36/43
Pregnancy	Find Baby Shower Invitations. Get Up To (60% Off) Here!	0.99, 1.0	9/9, 22/22
	Ralph Lauren Apparel. Official Online Store.	0.99, 0.6	10/10, 85/181
	Clothing Label-USA. Best Custom Woven Labels.	0.99, 1.0	9/9, 14/14
	Bonobos Official Site.	0.99,	9/9,

Results: Scalability

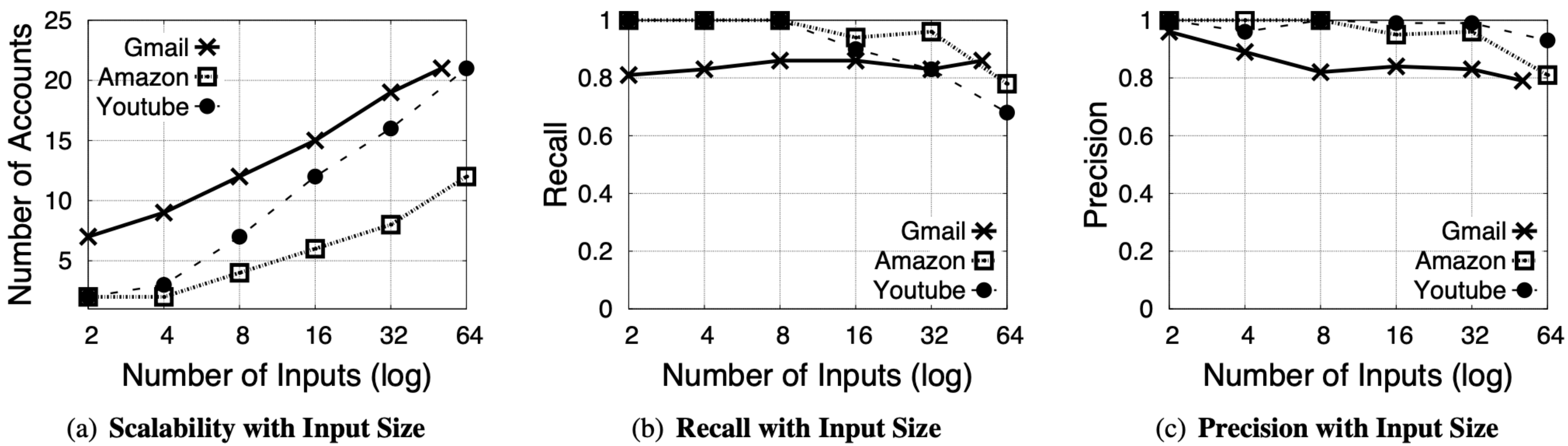


Figure 8: **Scalability.** (a) Number of accounts required to achieve the knee accuracy for varied numbers of inputs. (b), (c) Recall/precision achievable with the number of accounts in (a). Behavioral uses the Bayesian algorithm.

What are some of the challenges?

- Only detect correlation, not causation
- Required manual tuning for each service