

Privacy through Accountability

Anupam Datta

Associate Professor

CSD, ECE, CyLab

Carnegie Mellon University

Personal Information is Everywhere



Google

facebook

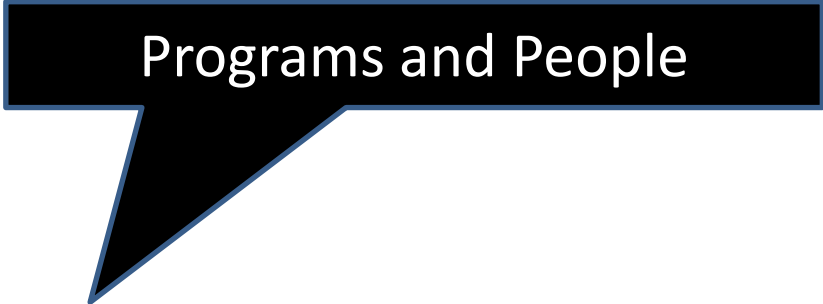


amazon.com



flickr® from YAHOO!

Research Challenge



Programs and People

Ensure organizations respect privacy expectations, regulations, and organizational policies in the collection, use, and disclosure of personal information

Privacy through Accountability: An Emerging Research Area

- Privacy as a right to restrictions on personal information flow
- Computational mechanisms for accountability (internal and external oversight)

<http://www.andrew.cmu.edu/user/danupam/privacy.html>

Today: Two Recent Results

1. Information Flow Experiments

- Methodology for black-box systems
- External oversight tool and application to personal information use in Google's advertising system



2. Bootstrapping Privacy Compliance in Big Data Systems

- Methodology for white-box systems
- Internal oversight tool and application to personal information use Bing's advertising system



Information Flow Experiments

With
Amit Datta (CMU), Michael Tschantz (ICSI) and
Jeannette Wing (MSR)

ADVERTISEMENT



RBC Royal Bank

Enter the DEBIT TO WIN IT™ contest.

Learn More >



THE TIMES OF INDIA China

The Times of India

Search

Advanced Search >

- Home
- World
- US
- Pakistan
- South Asia
- UK
- Europe
- China**
- Middle East
- Rest of World
- Mad, Mad World
- Videos

You are here: Home » World » China

'We'll be back': Hong Kong protesters chant as camp site dismantled

Reuters | Dec 12, 2014, 08:39 AM IST

Time to Hug* by Huggies®

Parenting info, Prizes and Offers! To Meet new Moms like You. : www.facebook.com/TimetoHug

Ads by Google

READ MORE >> [Hong Kong Protesters](#) | ['We'll Be Back'](#) | [Hong Kong](#) | [CY Leung](#)



Police officers stand guard before they move on to remove protesters from a road written 'We Will Be Back' with tarps at an occupied area outside government headquarters in Hong Kong.

HONG KONG: Hong Kong police arrested pro-democracy activists and cleared most of the main protest site on Thursday, marking an end to more than two months of street demonstrations in the Chinese-controlled city, but many chanted: "We will be back".

Most activists chose to leave the Admiralty site, next to the Central business area, peacefully, despite their demands for a free vote not being met. But the overall mood remained defiant.

Hong Kong Federation of Students leader Alex Chow said: "You might have the clearance today but people will come back on to the streets

Connect with us



4

comments

20

Like

Share

77

Tweet

0

+1

1

Share

Share More



AA

RELATED

another day."

LIVE
1ST TEST

cricbuzz

AUS 517/7 dec & 207/3

57.2 Ov

444 IND

Day 4: 3rd Session - Australia lead by 280 runs

Subscribe to our Newsletters

- Top News (Daily)
- Tech News (Daily)

Google's Privacy Policy

When showing you tailored ads, we will not associate a cookie or anonymous identifier with sensitive categories, such as those based on race, religion, sexual orientation or health.

Settings for Google ads

Ads enable free web services and content. These settings help control the types of Google ads you see.

Ads on Google



Search

Google ads across the web ?



Google ads across the web



YouTube

Gender

N/A

Female [Edit](#)

Based on the websites you've visited

Age

N/A

25-34 [Edit](#)

Based on the websites you've visited

Languages

N/A

English [Edit](#)

Based on the websites you've visited

Interests

N/A

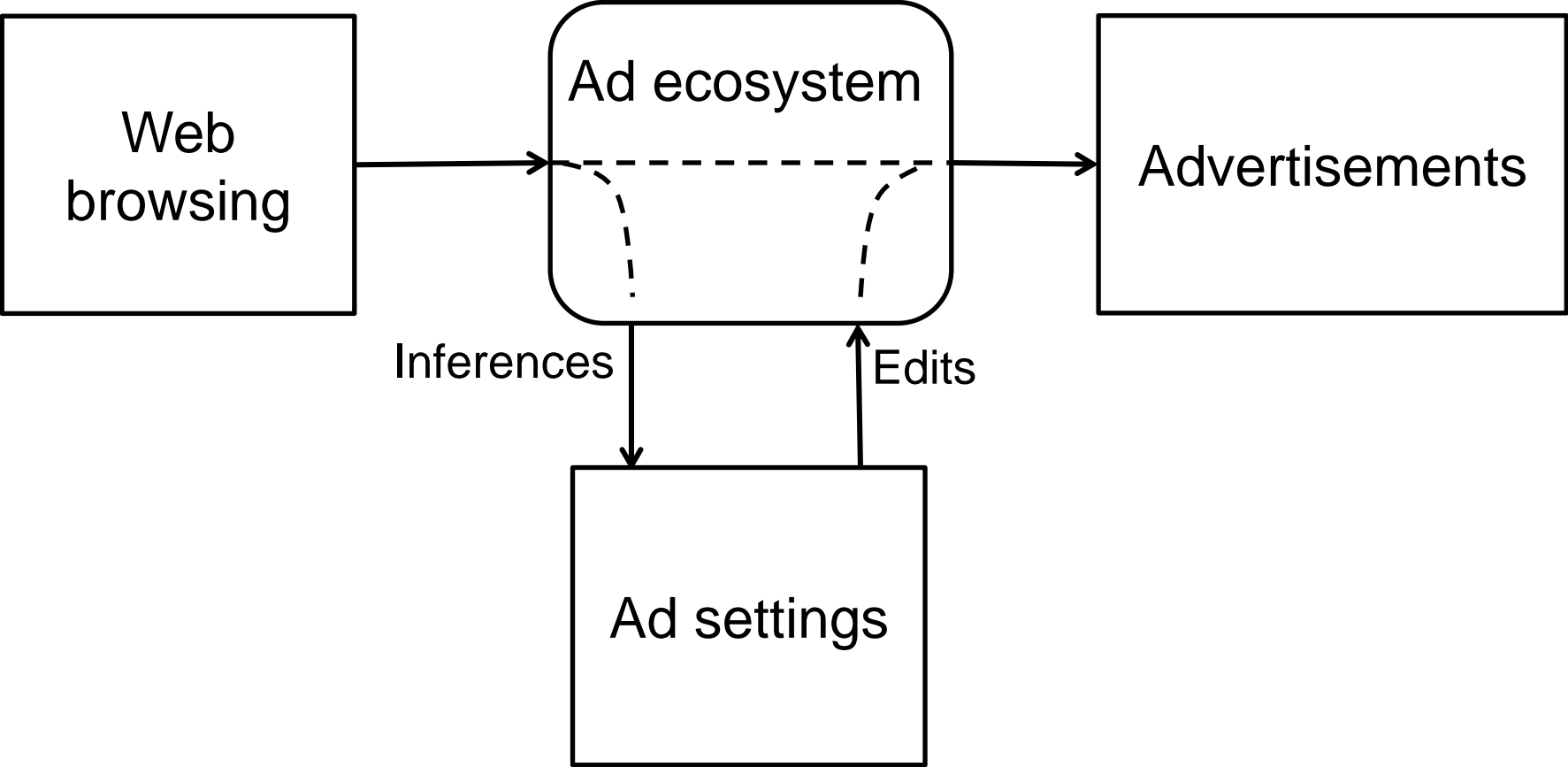
Air Travel, and 30 more [Edit](#)

Based on the websites you've visited

Opt-out settings

You've opted out of *interest-based* ads on Google.
[Opt in](#) to *interest-based* ads on Google

[Opt out](#) of *interest-based* Google ads across the web

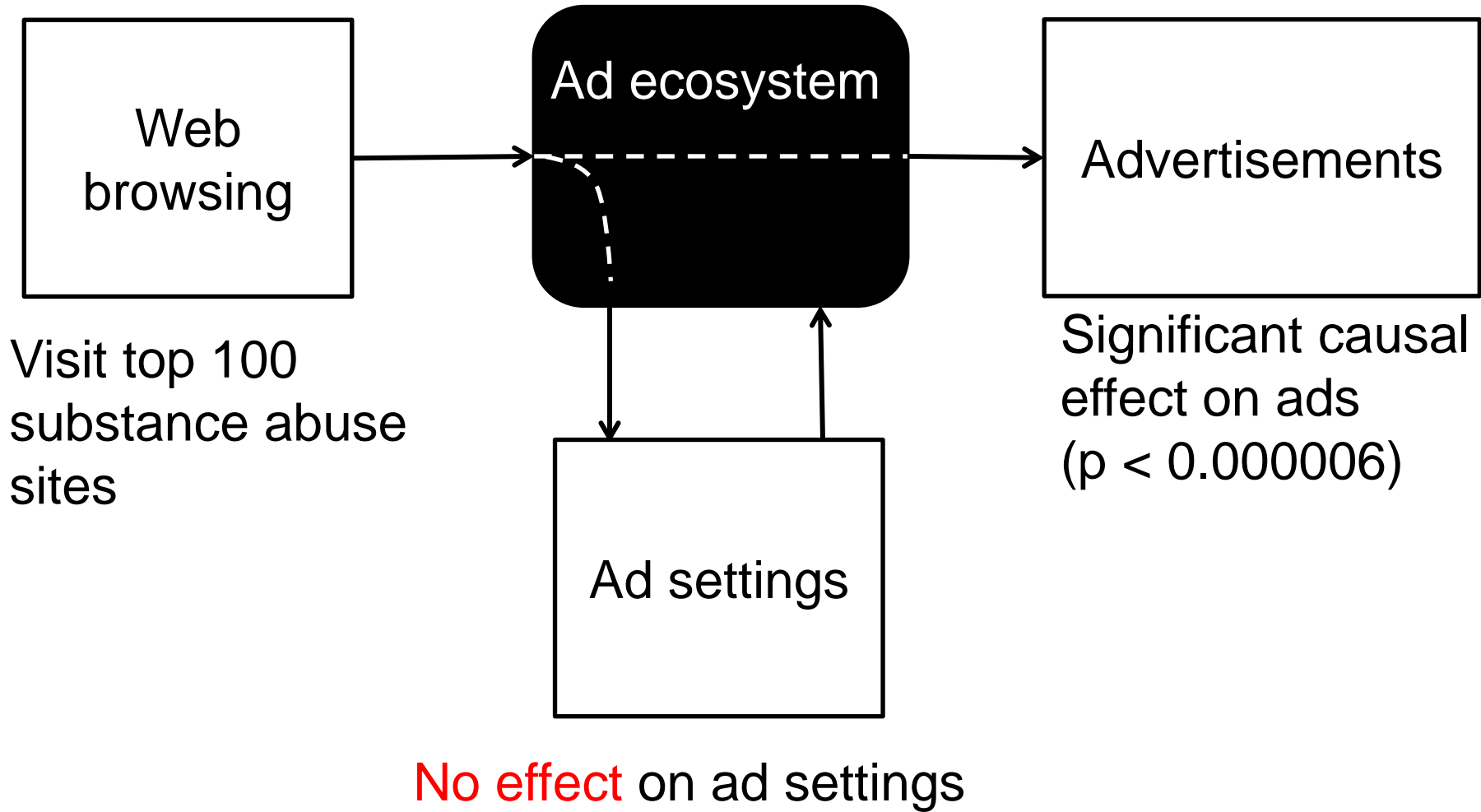


AdFisher



- Emulates users with fresh browser instances
- Randomized assignment
- Statistical analysis to find causal relations
- Open source: github.com/tadatitam/info-flow-experiments

Transparency



Transparency Explanations

Substance Abuse Visitors

The Watershed Rehab
www.thewatershed.com/Help
2276 vs. 0

Watershed Rehab
www.thewatershed.com/Rehab
362 vs. 0

The Watershed Rehab
(none)
771 vs. 0

Control Group

Alluria Alert
www.bestbeautybrand.com
0 vs. 9

Best Dividend Stocks
dividends.wyattresearch.com
24 vs. 54

10 Stocks to Hold Forever
www.streetauthority.com
76 vs. 118

The Watershed Rehab

www.thewatershed.com/Help - Drug & Alcohol Rehabilitation Call Today For Help Now!

Ads by Google

Findings and Non-Findings

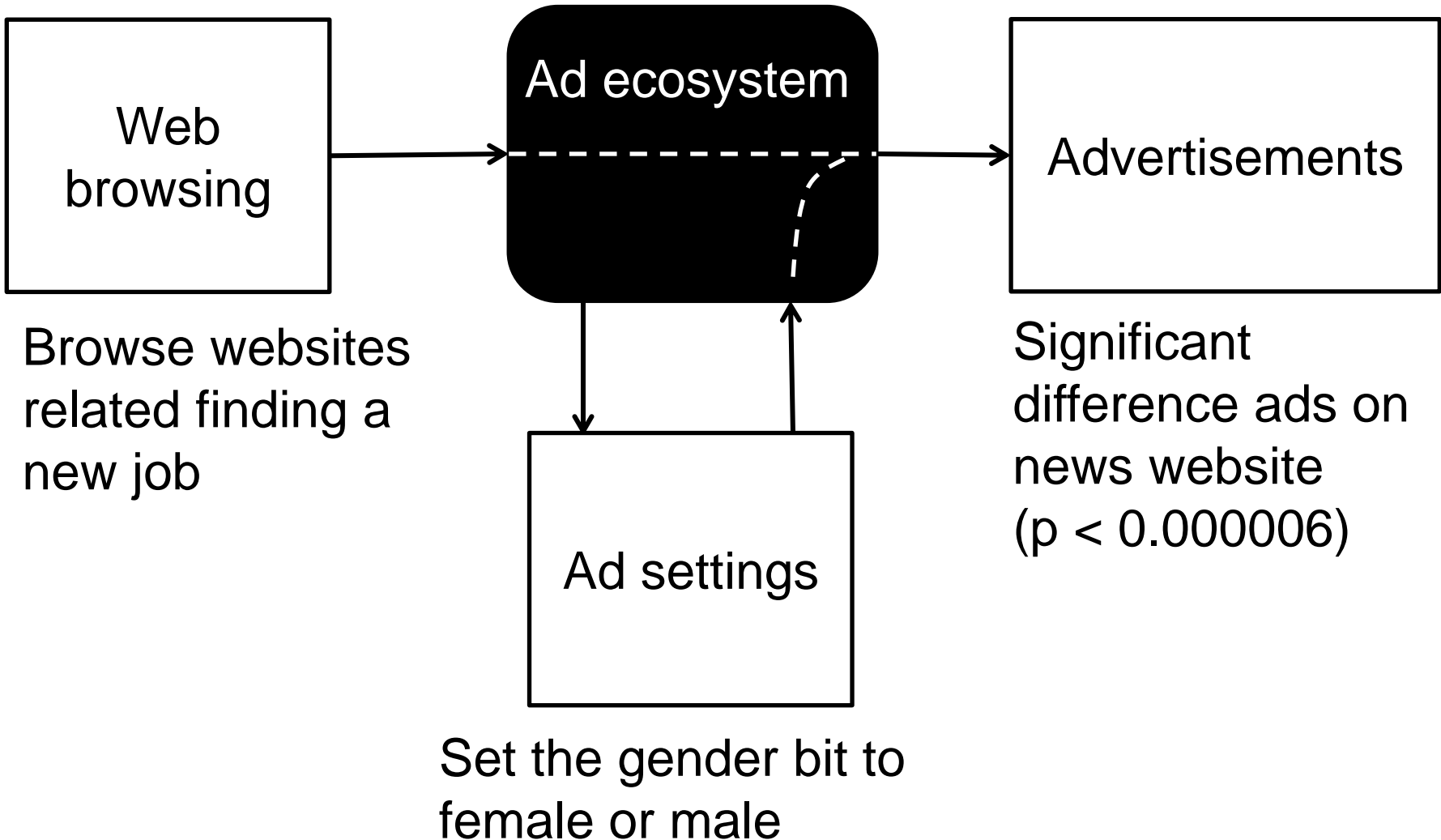
- lack of **transparency**
affected ads without affecting Ad Settings

- might not generalize to other settings

- no blame

- No claims that Google or anyone else violated any policies

Discrimination



Discrimination Explanation

Female Group

Jobs (Hiring Now)

www.jobsinyourarea.co

45 vs. 8

4Runner Parts Service

www.westernpatoyotaservice.com

36 vs. 5

Criminal Justice Program

www3.mc3.edu/Criminal+Justice

29 vs. 1

Male Group

\$200k+ Jobs - Execs Only

careerchange.com

311 vs. 1816

Find Next \$200k+ Job

careerchange.com

7 vs. 36

Become a Youth Counselor

www.youthcounseling.degreeleap.com

0 vs. 310

Information Flow Experiments

Natural Sciences

Natural process

Population of units

...

Causation

Information Flow

System in question

Subset of interactions

...

Information flow

Theorem

Pearl's Causation

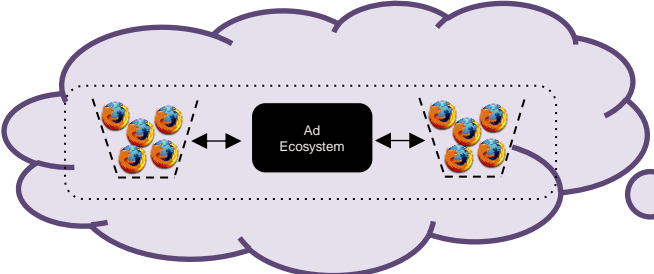
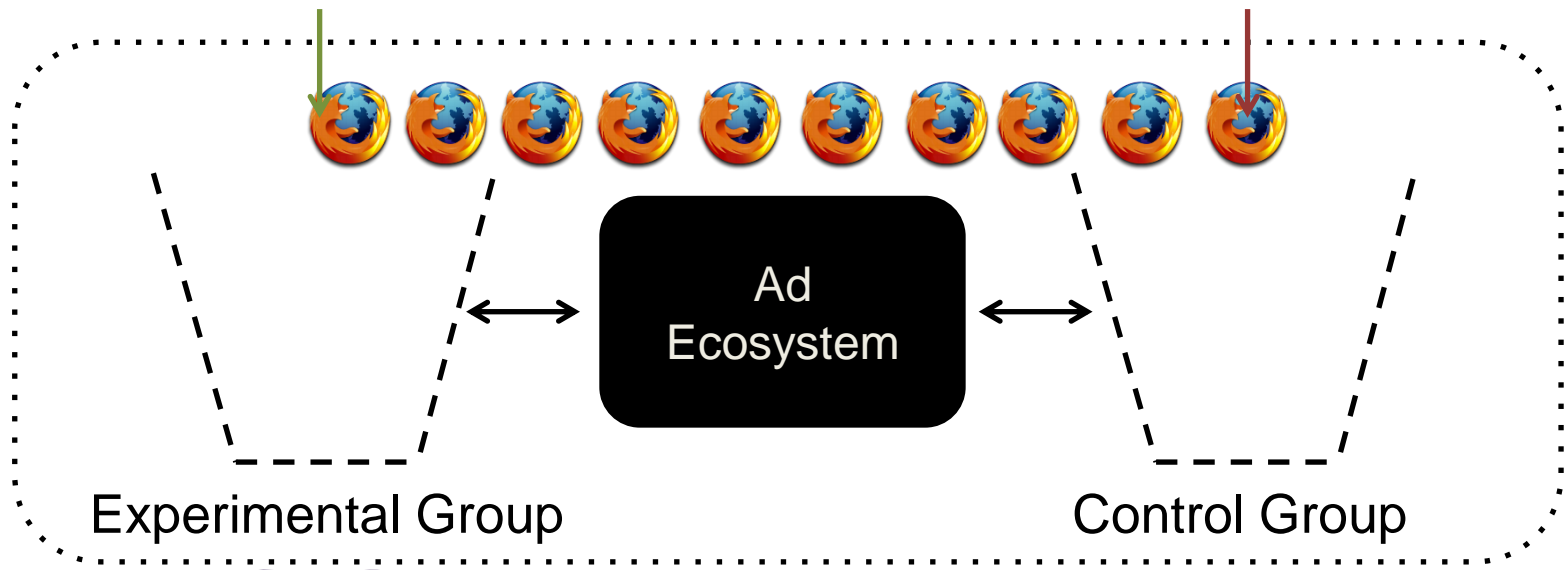
=

Probabilistic Interference

Randomized Controlled Trials

Experimental Treatment

Control Treatment



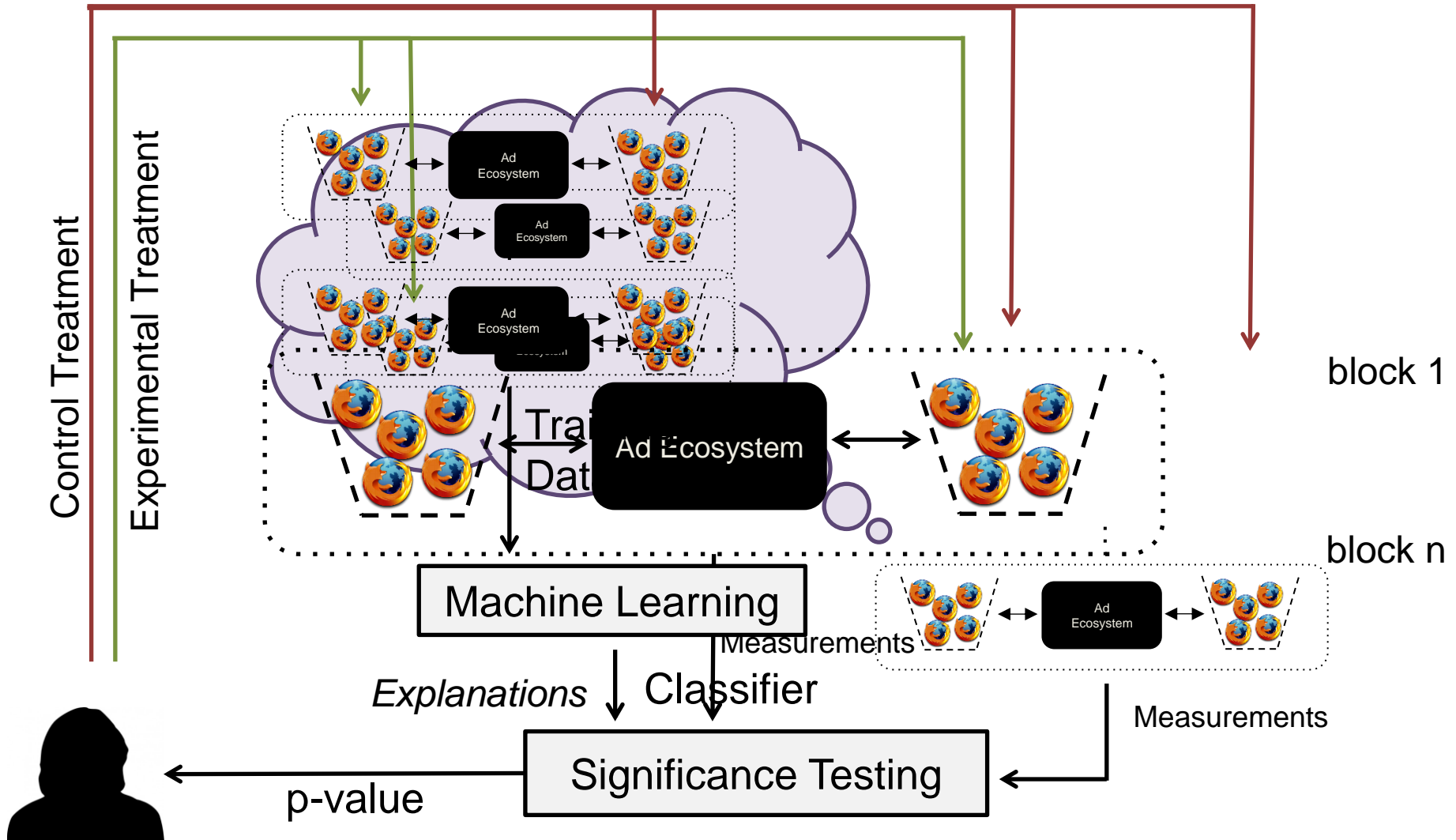
Measurements

Test Statistic

Hypothetical Value

Observed Value

Our Methodology



Prior Work on Behavioral Marketing

Authors	Test	Limitation
Guha et al.	Cosine similarity	No statistical significance
Balebako et al.	Cosine similarity	No statistical significance
Wills and Tatar	Manual examination	No statistical significance
Liu et al.	Process of elimination	No statistical significance
Barford et al.	χ^2 test	Assumes ads identically distributed
Lécuyer et al.	Parametric model	Correlation, not causation; assumes ads are independent
Englehardt et al.	Binomial test	Assumes ads identically distributed

Summary

- Rigorous information flow experiments
 1. Probabilistic interference = Pearl's causation
 2. Experimental design for causal determination
 3. Significance testing with non-parametric statistics
- Experimental study of Google Ads
 1. AdFisher Tool
 2. Findings of opacity, choice, and discrimination

Bootstrapping Privacy Compliance in Big Data Systems

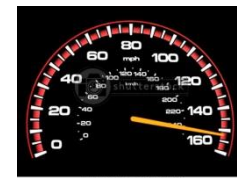
With S. Sen (CMU) and
S. Guha, S. Rajamani, J. Tsai, J. M. Wing (MSR)

Privacy Compliance for Bing

The image shows a side-by-side comparison of two web pages in a browser. The left page is the Bing homepage (http://www.bing.com/), featuring a search bar, navigation links (IMAGES, VIDEOS, MAPS, NEWS, SEARCH HISTORY, MORE), and a large image of a reindeer. The right page is the Bing Privacy Statement (http://www.microsoft.com/privacystatement/en-us/b...), which includes sections for 'Cookies & Similar Technologies', 'Collecting Your Information', and 'How We Use Your Personal Information'. A vertical sidebar on the right of the privacy statement page contains various links like 'Cookies', 'Collecting Your Information', 'Using Your Information', etc.

Setting:

- Auditor has access to source code



The Privacy Compliance Challenge

Legal Team

Crafts Policy

Meetings

Privacy Champion

Interprets Policy

Meetings

Developer

Writes Code

Meetings

Audit Team

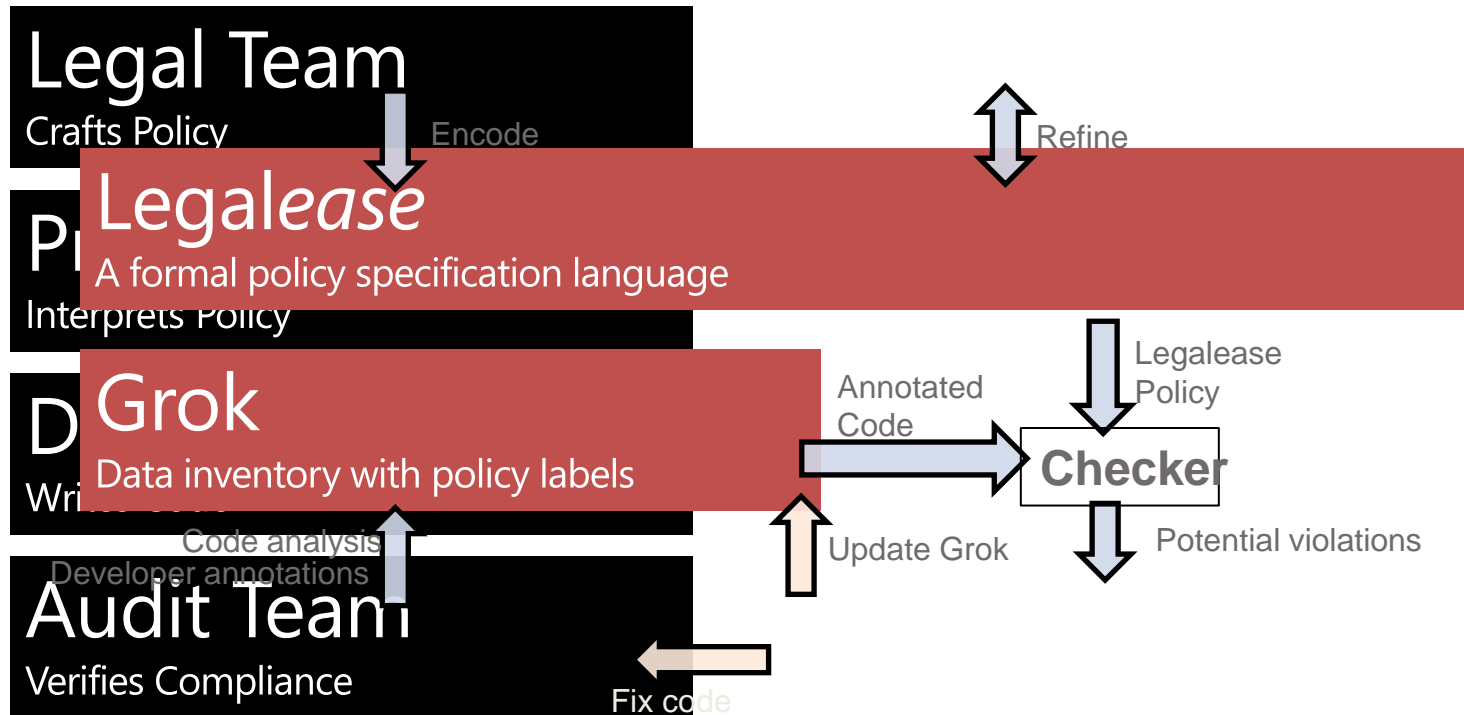
Verifies Compliance

English
Specification
Privacy Policy

Compliance?

Millions of Lines of
Verification
Undocumented Code

A Streamlined Audit Workflow



Specification: *Legalease*

Usable.
Expressive.
Precise.

Usable by
lawyers
and
privacy
champs.

Expressive
enough for
real-world
policies.

Precise
semantics
for local
reasoning.

Legalease: Example Policy

DENY *Datatype* IPAddress

UseForPurpose Advertising

EXCEPT

ALLOW

UseForPurpose AbuseDetect

EXCEPT

DENY *Datatype*

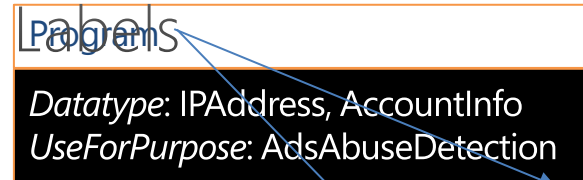
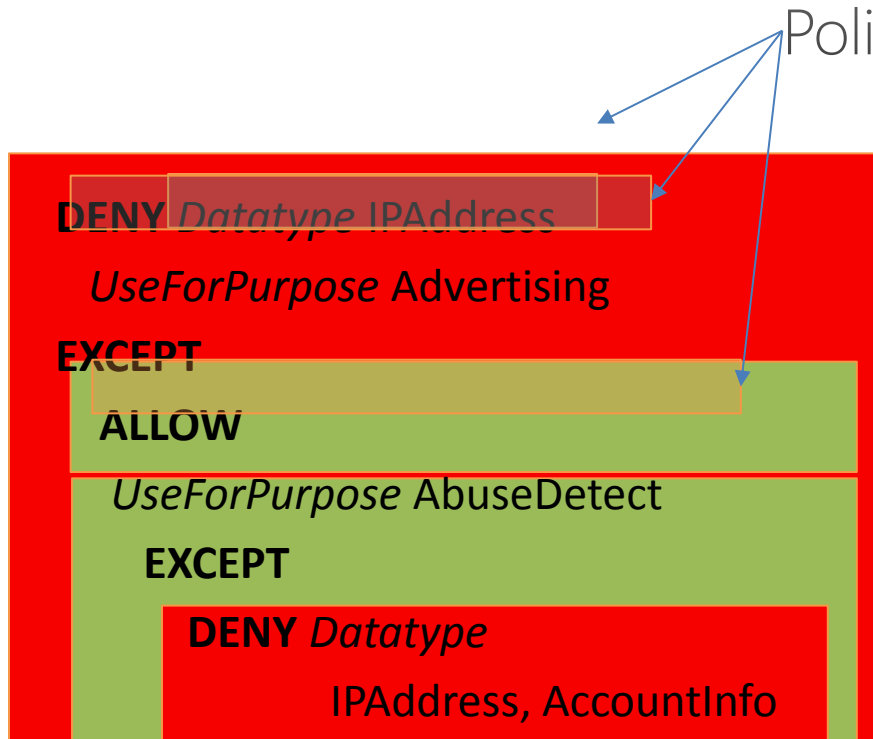
IPAddress, AccountInfo

We will **not** use **full IP Address** for **Advertising**.

IP Address may be used for **detecting abuse**.

In such cases, it will not be combined with **account information**.

Legalease : Policy Checking



We will not use full IP Address for Advertising. IP Address may be used for detecting abuse. In such cases, it will not be combined with account information.



Designed for Precision

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

TABLE I
GRAMMAR FOR LEGALEASE

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

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

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

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

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

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

TABLE III
INFERENCE RULES FOR LEGALEASE

Designed for Expressivity (Bing, October 2013)

ALLOW
EXCEPT

DENY *DataType* IPAddress:Expired

DENY *DataType* UniqueIdentifier:Expired

DENY *DataType* SearchQuery, PII *InStore* Store

DENY *DataType* UniqueIdentifier, PII *InStore* Store

DENY *DataType* BBEPData *UseForPurpose* Advertising

DENY *DataType* BBEPData, PII *InStore* Store

DENY *DataType* BBEPData:Expired

DENY *DataType* UserProfile, PII *InStore* Store

DENY *DataType* PII *UseForPurpose* Advertising

DENY *DataType* PII *InStore* AdStore

DENY *DataType* SearchQuery *UseForPurpose* Sharing

EXCEPT

ALLOW *DataType* SearchQuery:Scrubbed

◁ “we remove the entirety of the IP address after 6 months”

◁ “[we remove] cookies and other cross session identifiers, after 18 months”

◁ “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.”

◁ “we do not use any of the information collected through the Bing Bar Experience Improvement Program to identify, contact or target advertising to you”

◁ “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”

◁ “we delete the information collected through the Bing Bar Experience Program at eighteen months.”

◁ “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.)”

◁ “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).”

◁ “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

EXCEPT

ALLOW *DataType* PII:OptIn

EXCEPT

ALLOW *AccessByRole* Affiliates

EXCEPT

ALLOW *UseForPurpose* Legal

DENY *DataType* DoubleClickData, PII

EXCEPT

ALLOW *DataType* DoubleClickData, PII:OptIn

◁ “We do not share personal information with companies, organizations and individuals outside of Google unless one of the following circumstances apply:”

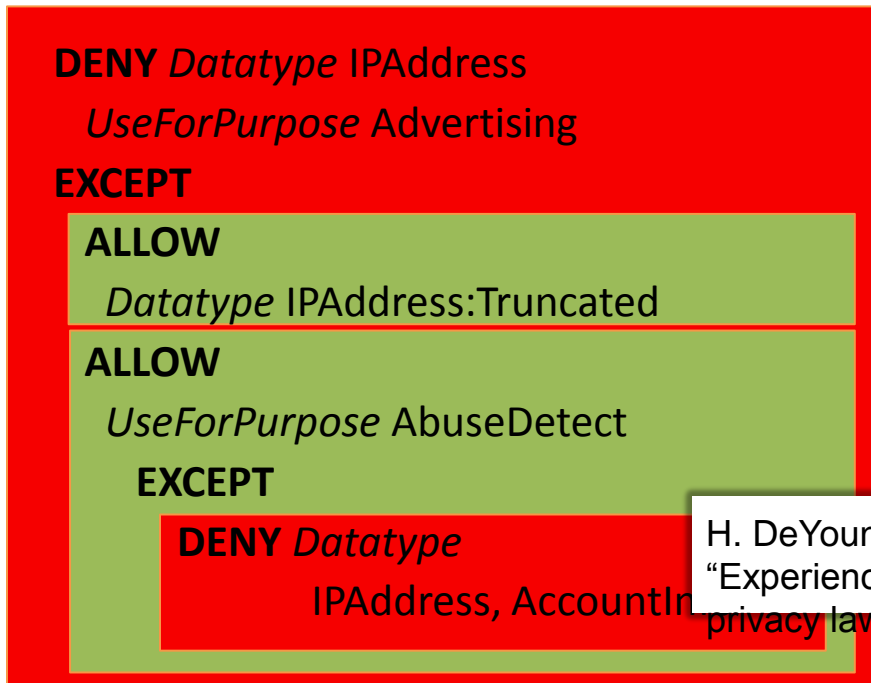
◁ “We require opt-in consent for the sharing of any sensitive personal information.”

◁ “We provide personal information to our affiliates or other trusted businesses or persons to process it for us”

◁ “We will share personal information [if necessary to] meet any applicable law, regulation, legal process or enforceable governmental request.”

◁ “We will not combine DoubleClick cookie information with personally identifiable information unless we have your opt-in consent”

Designed for Usability



Exceptions

How legal texts are structured

One-to one correspondence

Local Reasoning

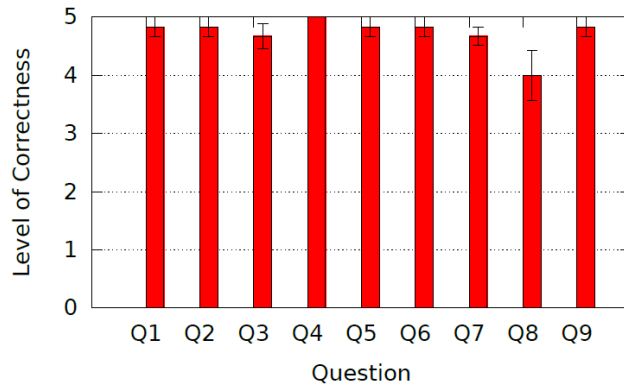
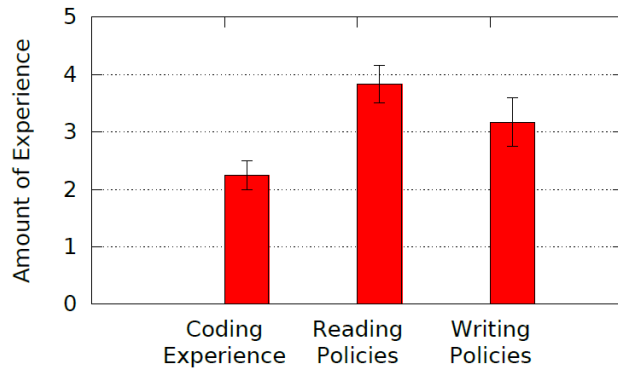
Each exception refines its

H. DeYoung, D. Garg, L. Jia, D. Kaynar, and A. Datta,
“Experiences in the logical specification of the HIPAA and GLBA
privacy laws”

formally proven property

Independent of Code

Legalease Usability



Survey taken by 12 policy authors within Microsoft

Encode Bing data usage policy after a brief tutorial

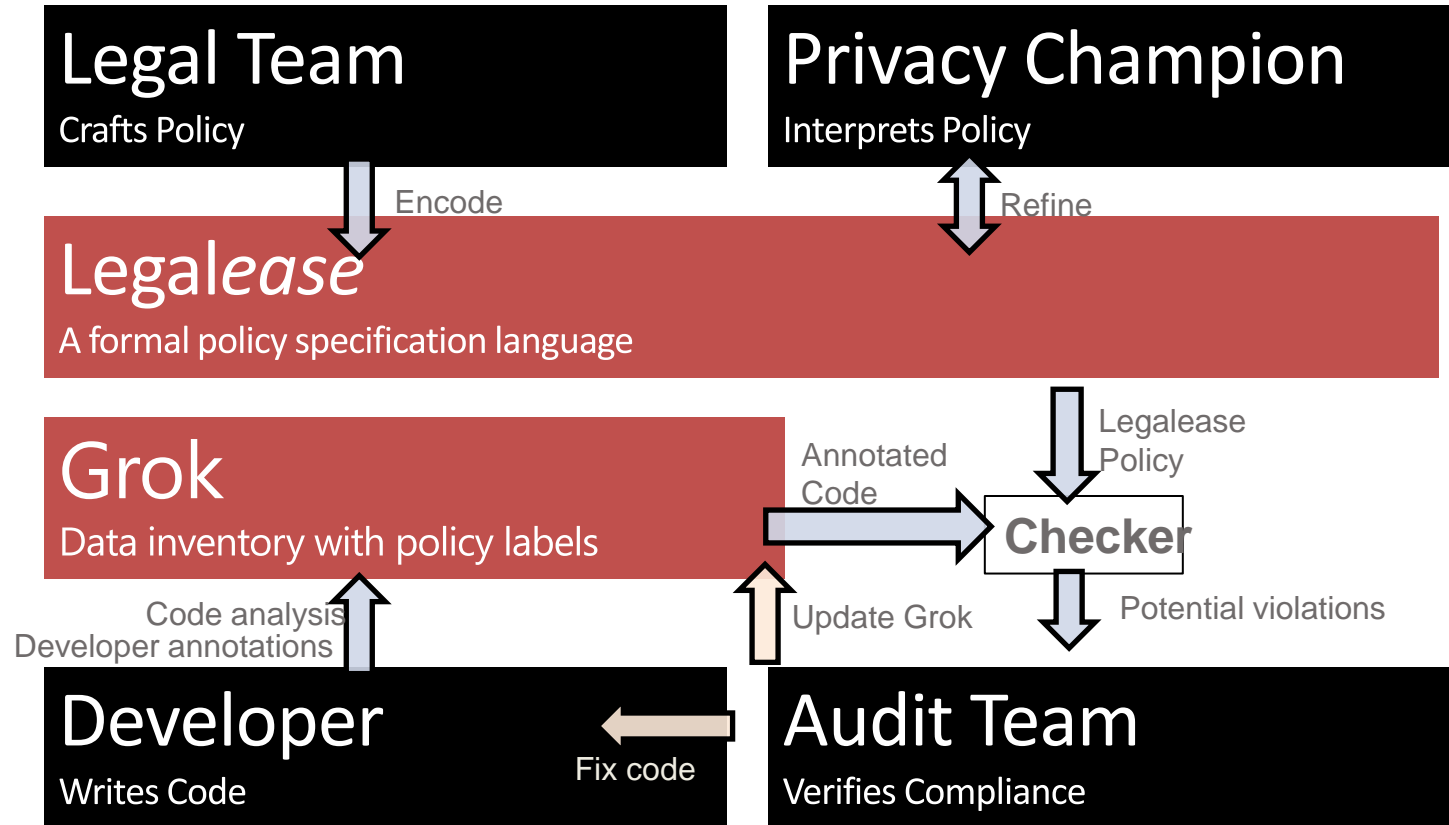
Time spent

2.4 mins on the tutorial

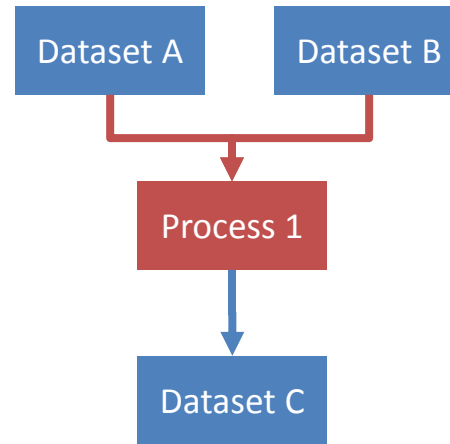
14.3 mins on encoding policy

High overall correctness

A Streamlined Audit Workflow



Map-Reduce Programming Systems



Scope, Hive, Dremel

Data in the form of Tables

Code Transforms Columns to
Columns

No Shared State

Limited Hidden Flows

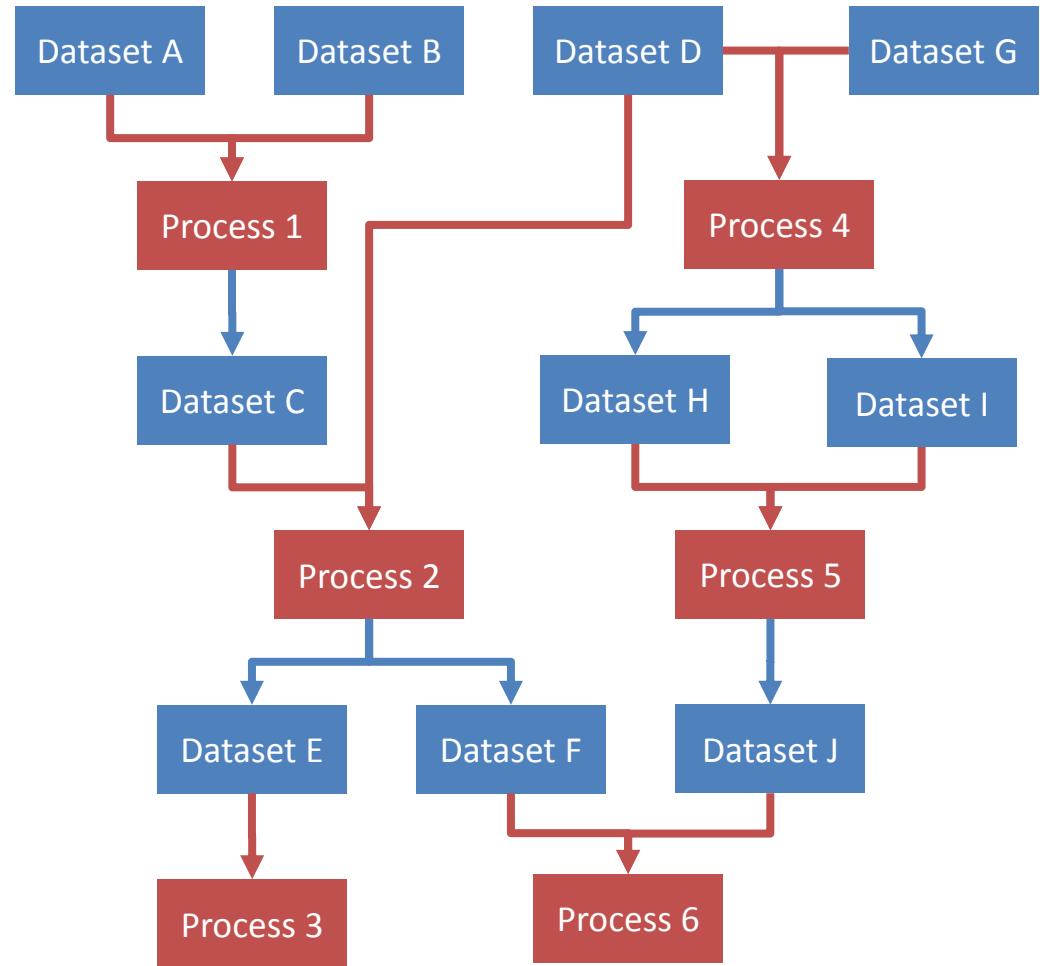
Verification

What data, stored where?
Who used.

Nightly audit of all jobs executed.

Static source code analysis.

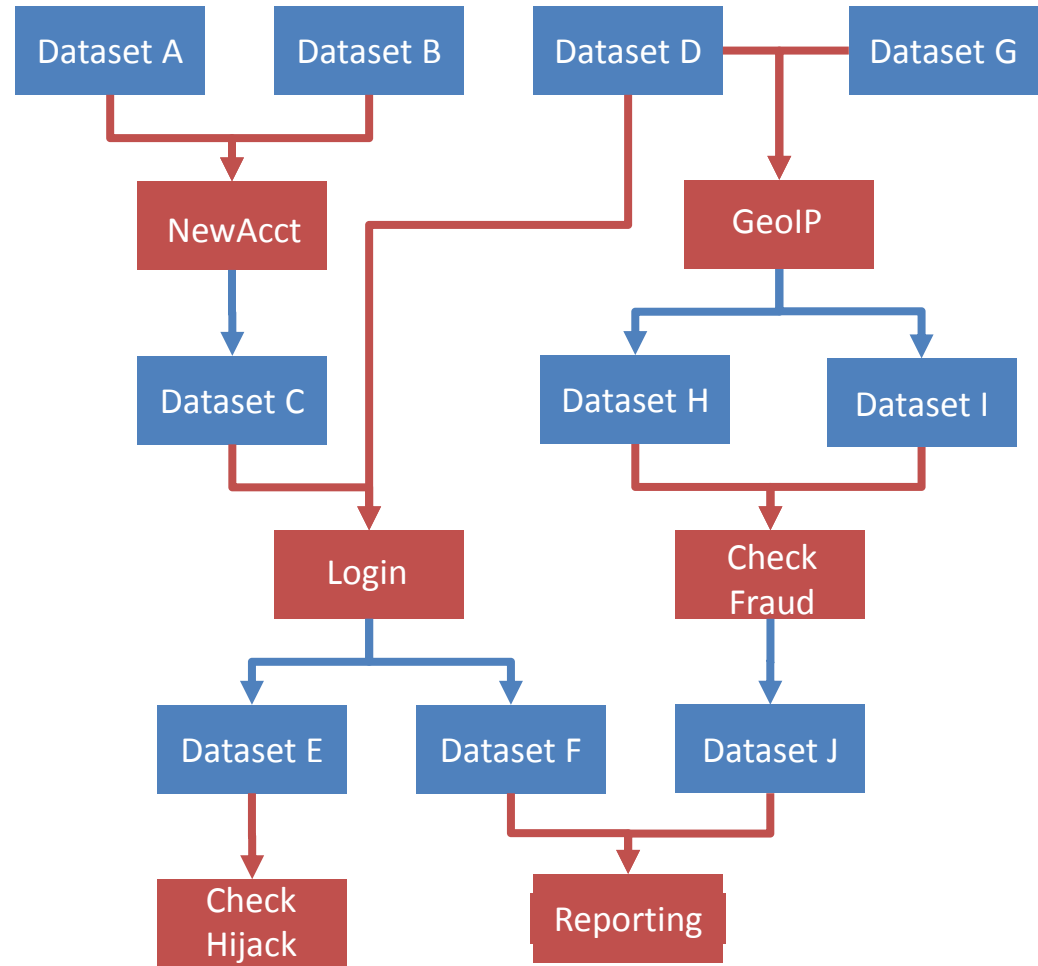
Grok



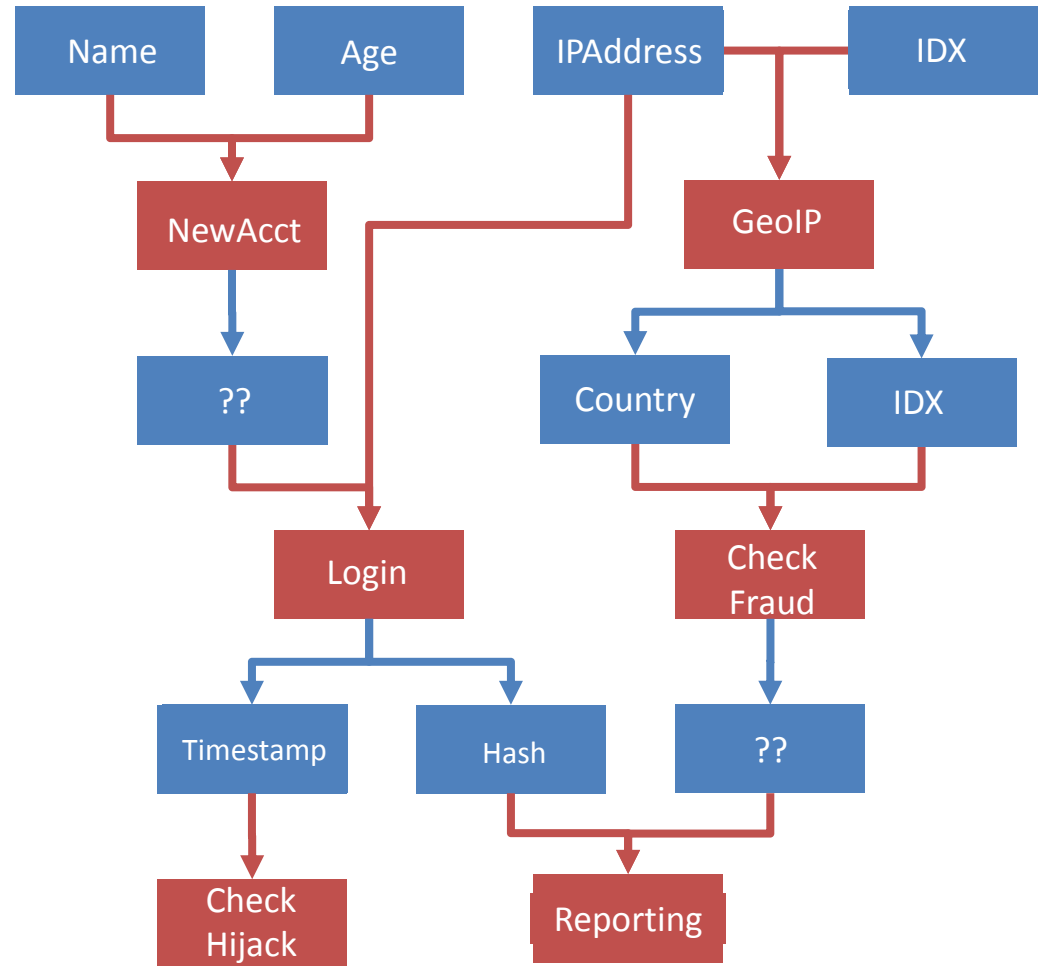
Grok

Purpose Labels

Annotate programs with purpose labels



Grok



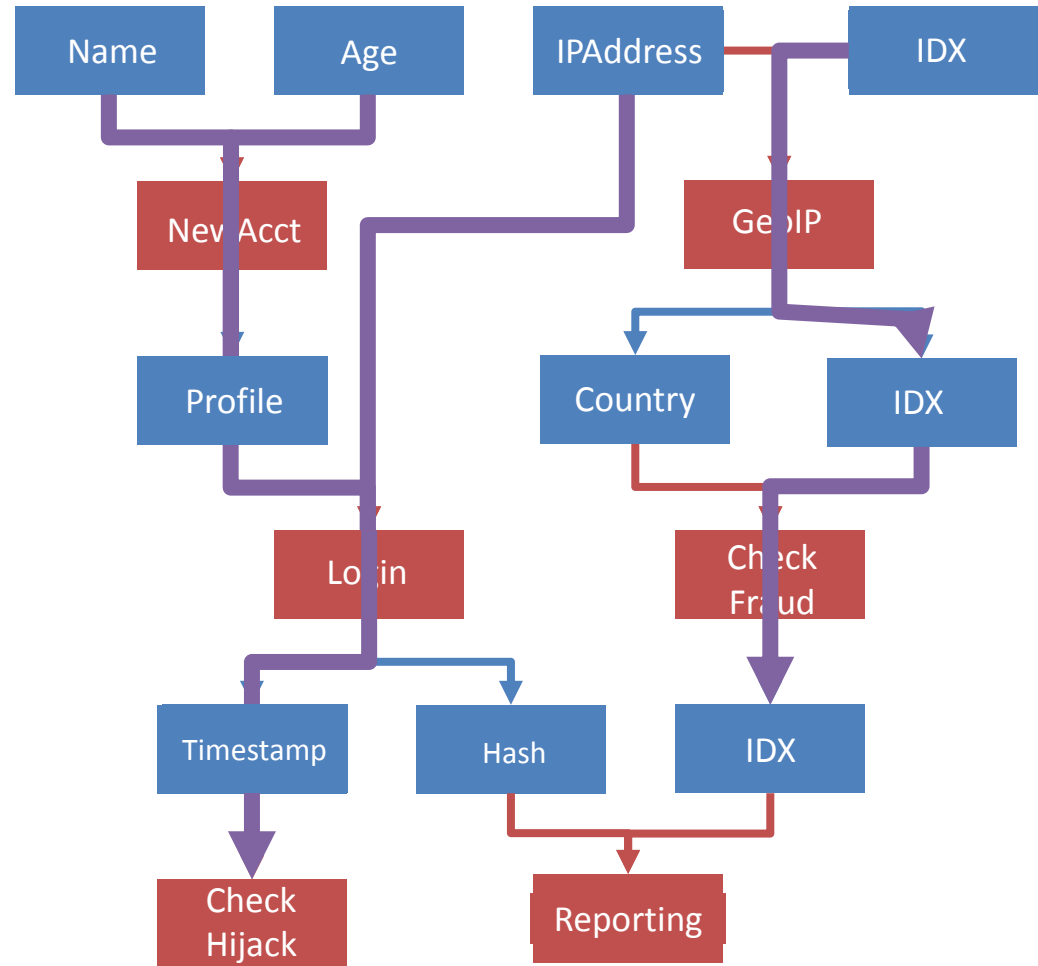
Purpose Labels

Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations

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"

Nightly Compliance Process

The screenshot shows a SQL query window with the following query:

```

-- DE
SELECT
FROM
INNER JOIN
(SELECT Dist
FROM
WHERE Taxonomy
ON s.cluster
ON II

```

Below the query is an email thread with the subject "RE: Looking for Privacy Mgr contacts (MS Com, Outlook, Skype)".

On the right, a table titled "PII InStore BingStore" shows the results of the query:

Confidence	TaxonomyGroup	Taxonomy	FieldName
andbox	HIGH	PII	Email
andbox	HIGH	PII	Phone Number
andbox	HIGH	PII	Email
andbox	HIGH	PII	Phone Number
partner	HIGH	PII	PUID
partner	HIGH	PII	PUID
devtest	HIGH	PII	Email
devtest	HIGH	PII	Email
devtest	HIGH	PII	Email

At the bottom, a small table shows a summary of the results:

7	cosmos05	bingmobile...	%/local/A...	EntityPhone	CONFIDENCE_JUNKVERIFIED	Verified with Feature teams that ...
8	%	%	%	LiveIdEmail	CONFIDENCE_HIGH	
9	%	%	%	PrimaryEmail	CONFIDENCE_HIGH	

Static code analysis

schemas

25M+

Generate report

privacy audit
calculated

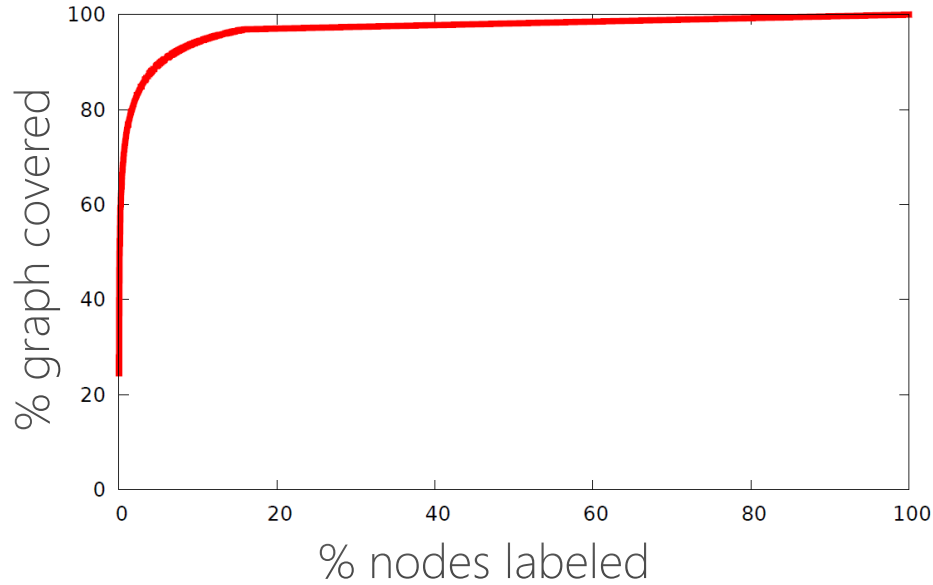
300K+

Manual Audit

teams

8

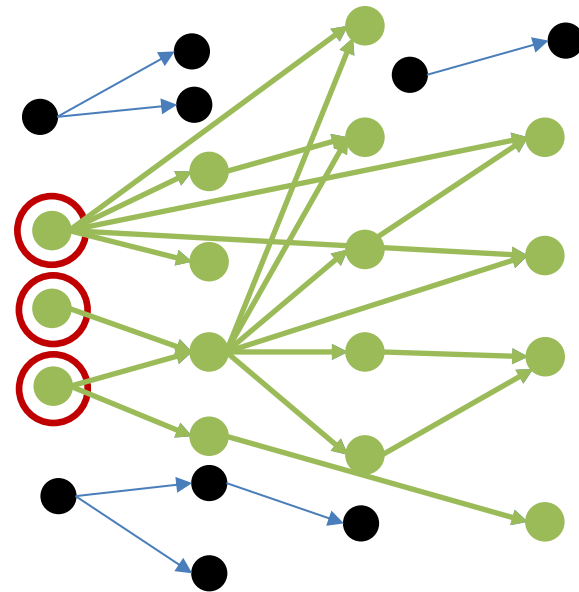
Why Bootstrapping Grok Works



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

Pick the nodes which will label the most of the graph

~200 annotations label 60% of nodes



Scale

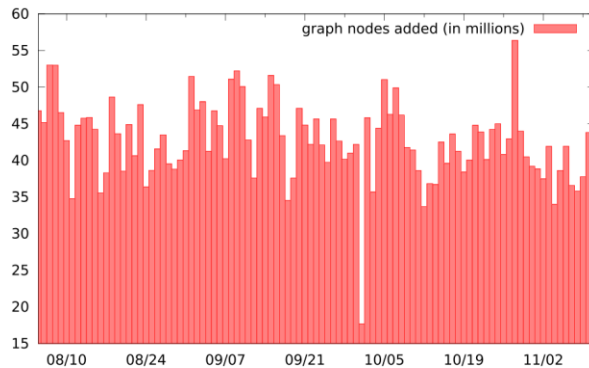
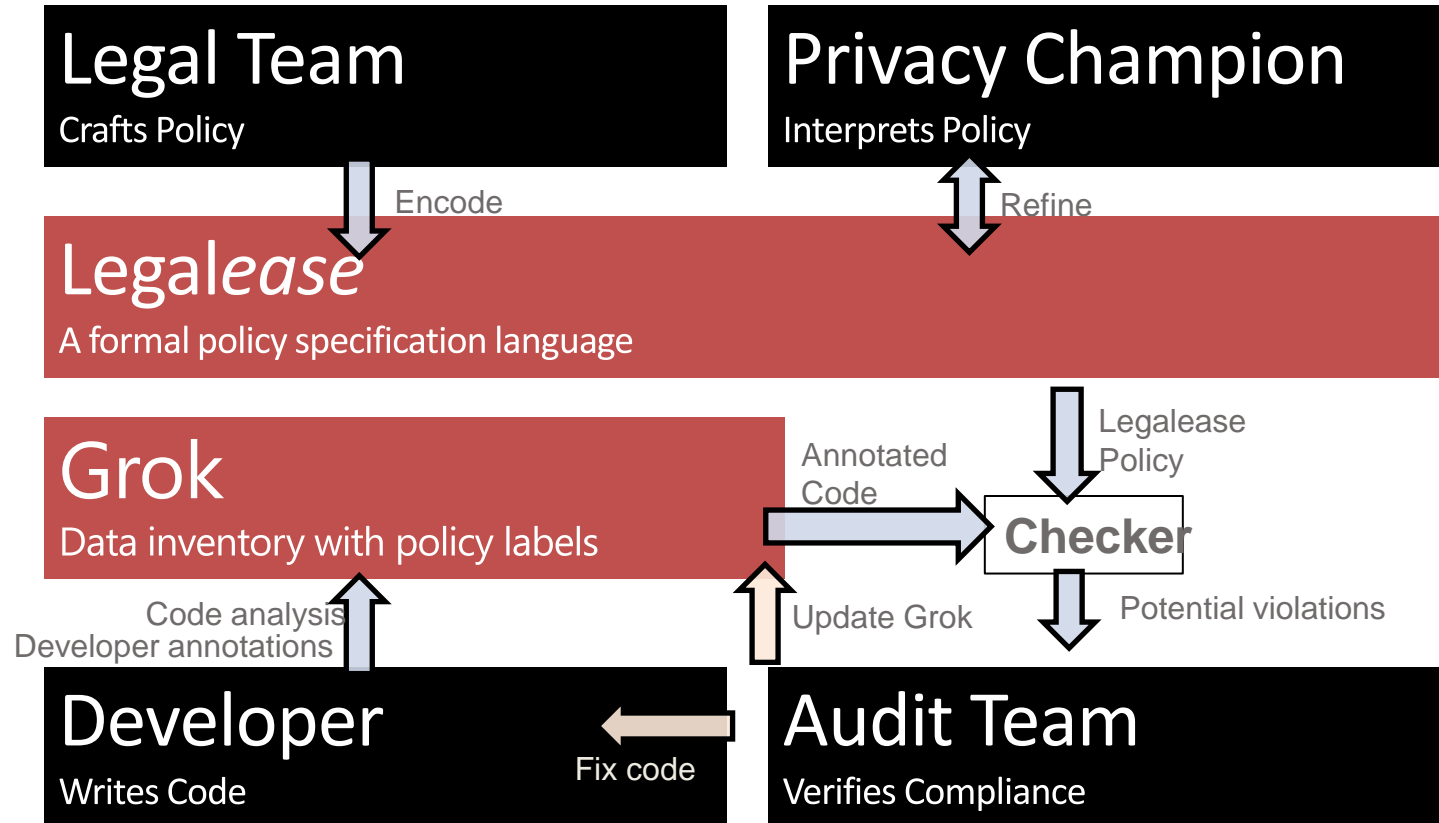


Fig. 9. Number of GROK data flow graph nodes added each day

- 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

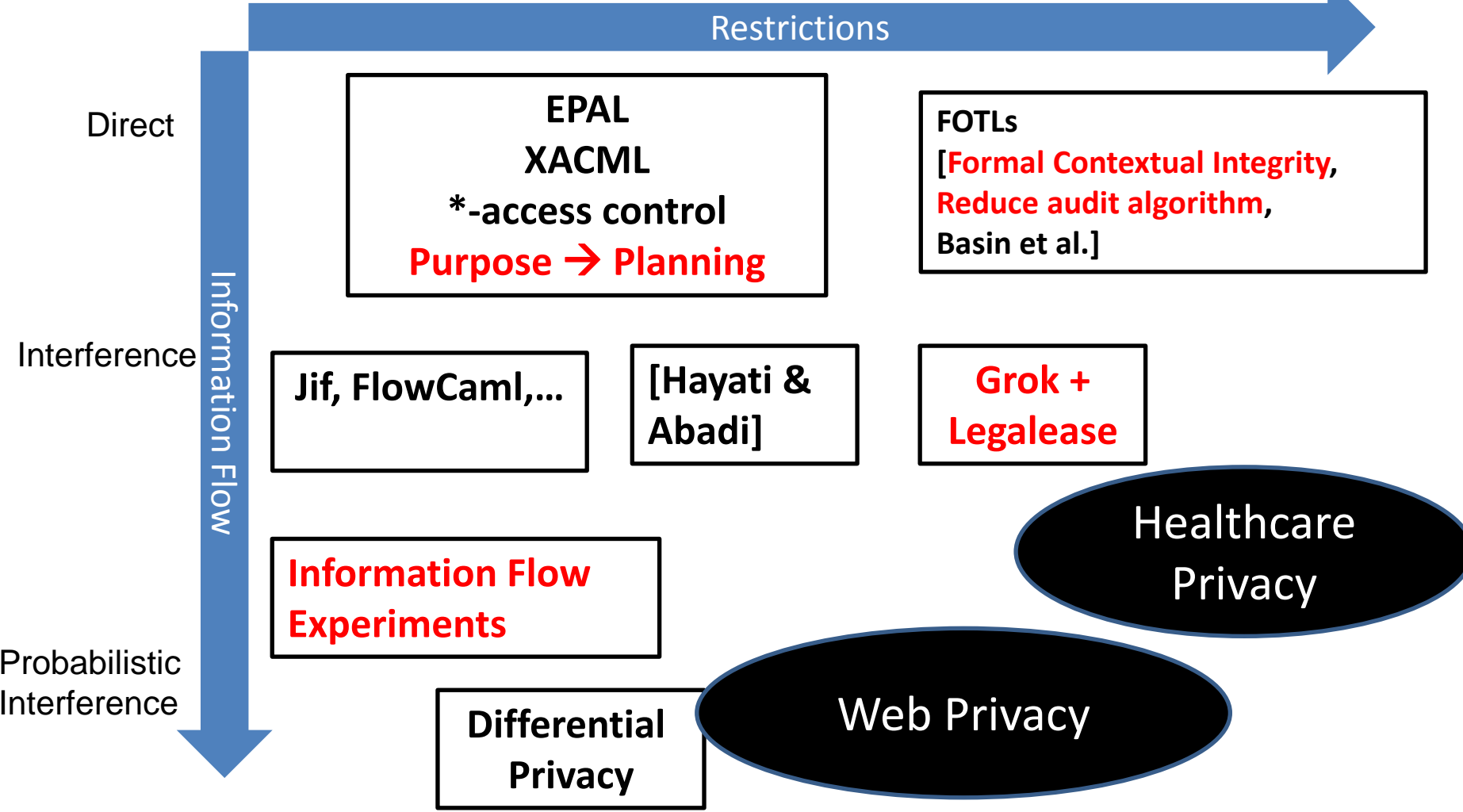
A Streamlined Audit Workflow



Privacy as Restrictions on Personal Information Flow

Purpose & Role based

Temporal



Today: Two Recent Results

1. Information Flow Experiments

- Methodology for black-box systems
- External oversight tool and application to Google's advertising system



2. Bootstrapping Privacy Compliance in Big Data Systems

- Methodology for white-box systems
- Internal oversight tool and application to Bing's advertising system



Privacy through Accountability: An Emerging Research Area

- Privacy as a right to restrictions on personal information flow
- Computational mechanisms for accountability (internal and external oversight)

<http://www.andrew.cmu.edu/user/danupam/privacy.html>