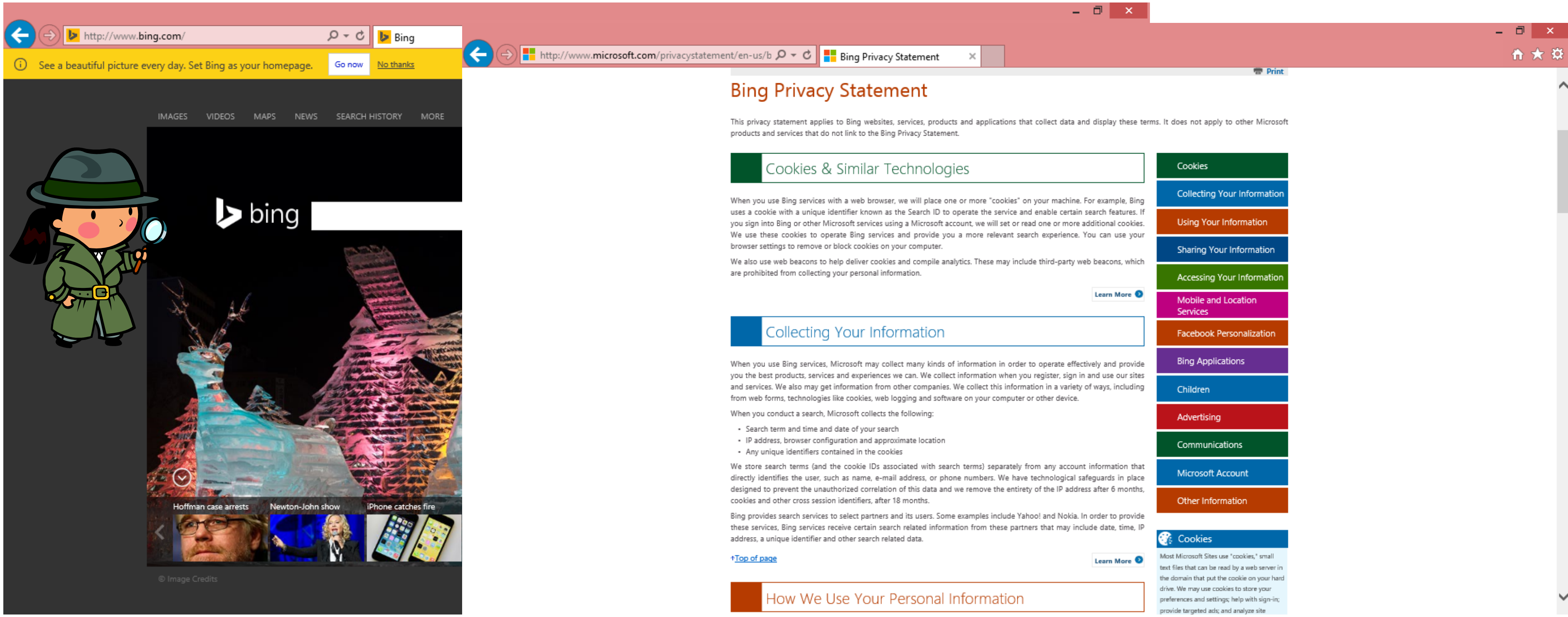
18734: Foundations of Privacy

Bootstrapping Privacy Compliance in Big Data Systems

Anupam Datta

Fall 2015

Privacy Compliance for Bing



Bing Privacy Statement

This privacy statement applies to Bing websites, services, products and applications that collect data and display these terms. It does not apply to other Microsoft products and services that do not link to the Bing Privacy Statement.

Cookies & Similar Technologies

When you use Bing services with a web browser, we will place one or more "cookies" on your machine. For example, Bing uses a cookie with a unique identifier known as the Search ID to operate the service and enable certain search features. If you sign into Bing or other Microsoft services using a Microsoft account, we will set or read one or more additional cookies. We use these cookies to operate Bing services and provide you a more relevant search experience. You can use your browser settings to remove or block cookies on your computer.

We also use web beacons to help deliver cookies and compile analytics. These may include third-party web beacons, which are prohibited from collecting your personal information.

Learn More

Collecting Your Information

When you use Bing services, Microsoft may collect many kinds of information in order to operate effectively and provide you the best products, services and experiences we can. We collect information when you register, sign in and use our sites and services. We also may get information from other companies. We collect this information in a variety of ways, including from web forms, technologies like cookies, web logging and software on your computer or other device.

When you conduct a search, Microsoft collects the following:

- Search term and time and date of your search
- IP address, browser configuration and approximate location
- Any unique identifiers contained in the cookies

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 have technological safeguards in place designed to prevent the unauthorized correlation of this data and we remove the entirety of the IP address after 6 months, cookies and other cross session identifiers, after 18 months.

Bing provides search services to select partners and its users. Some examples include Yahoo! and Nokia. In order to provide these services, Bing services receive certain search related information from these partners that may include date, time, IP address, a unique identifier and other search related data.

↑Top of page

Learn More

How We Use Your Personal Information

Cookies

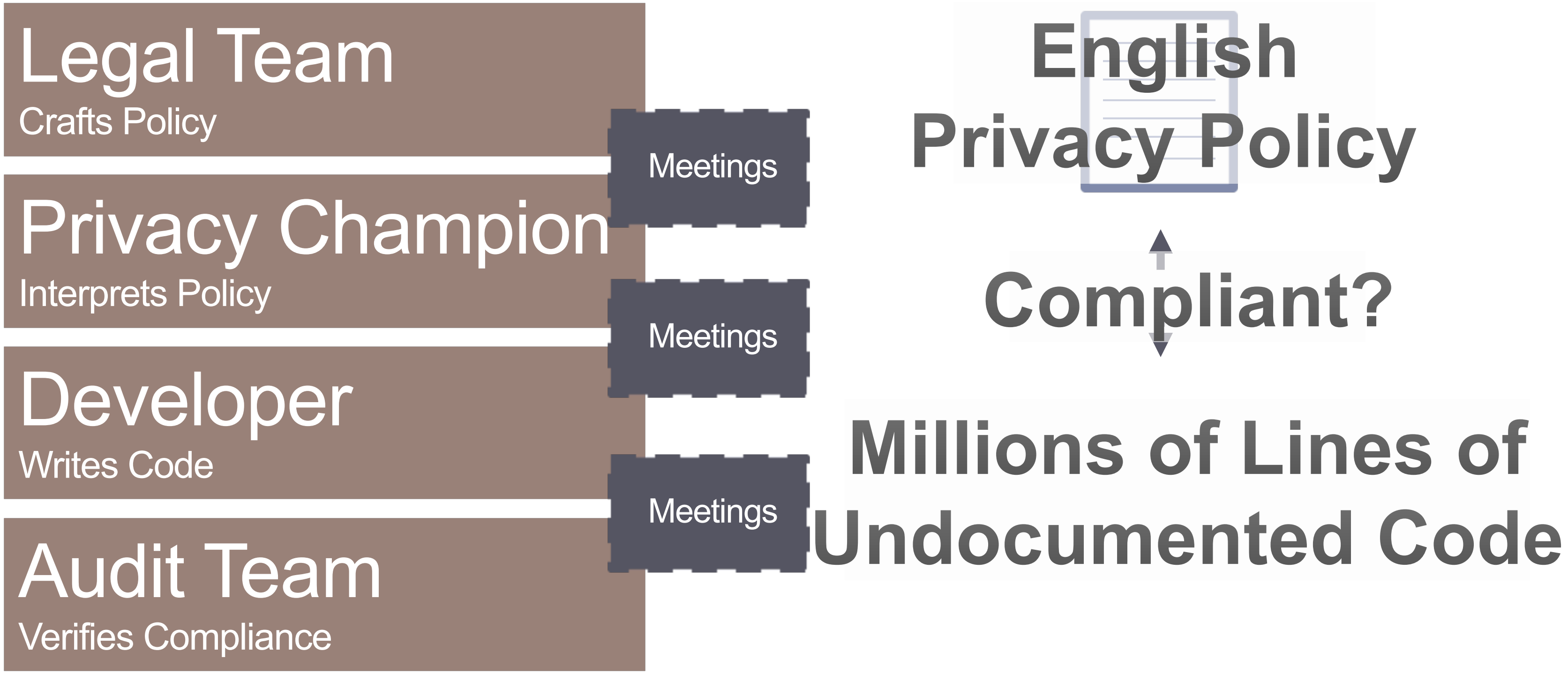
Most Microsoft Sites use "cookies," small text files that can be read by a web server in the domain that put the cookie on your hard drive. We may use cookies to store your preferences and settings; help with sign-in; provide targeted ads; and analyze site

Setting:

- Auditor has access to source code

www.shutterstock.com · 8021692

The Privacy Compliance Challenge



Legal Team
Crafts Policy

Privacy Champion
Interprets Policy

Developer
Writes Code

Audit Team
Verifies Compliance

Meetings

Meetings

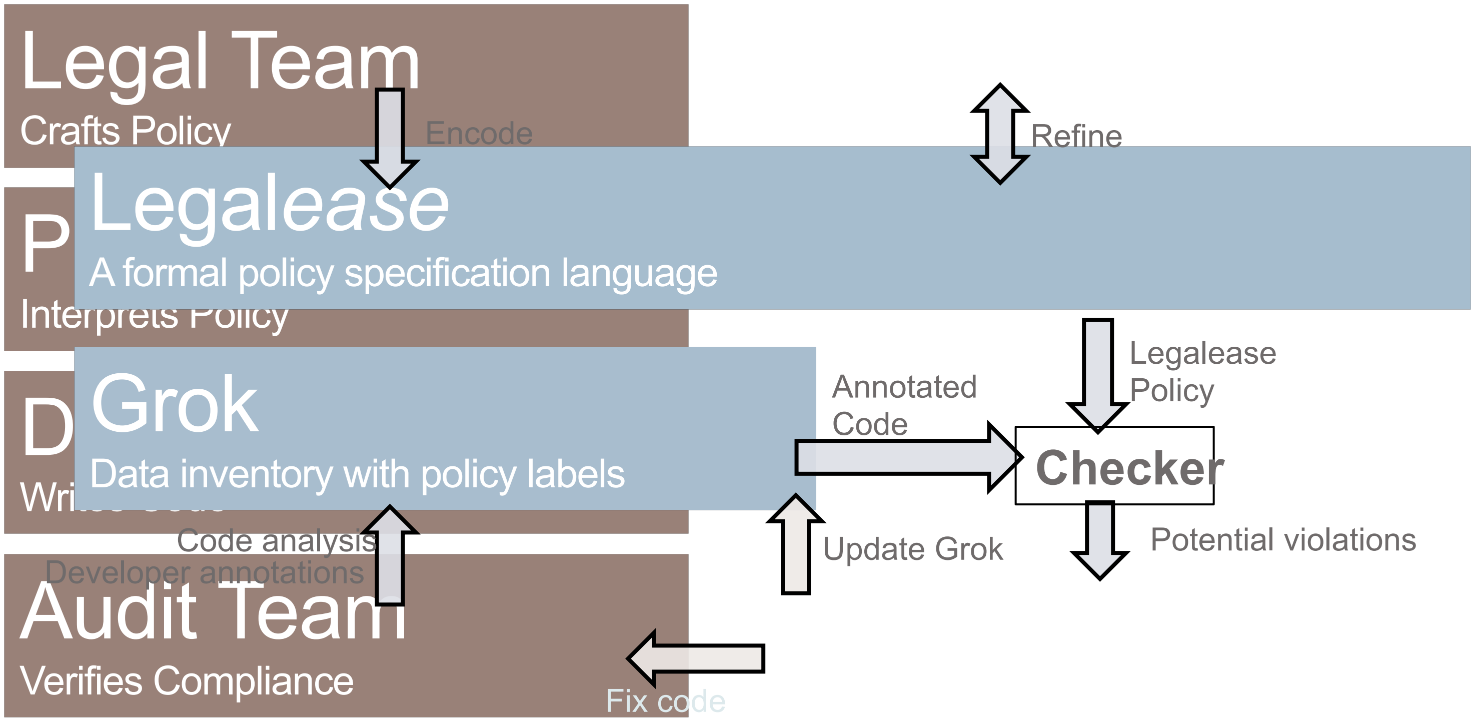
Meetings

English Privacy Policy

Compliant?

Millions of Lines of Undocumented Code

A Streamlined Audit Workflow



Legal Team

Crafts Policy

Encode

Refine

P

Legalease

A formal policy specification language

Interprets Policy

D

Grok

Data inventory with policy labels

Writes Code

Annotated Code

Legalease Policy

Checker

Update Grok

Potential violations

Code analysis

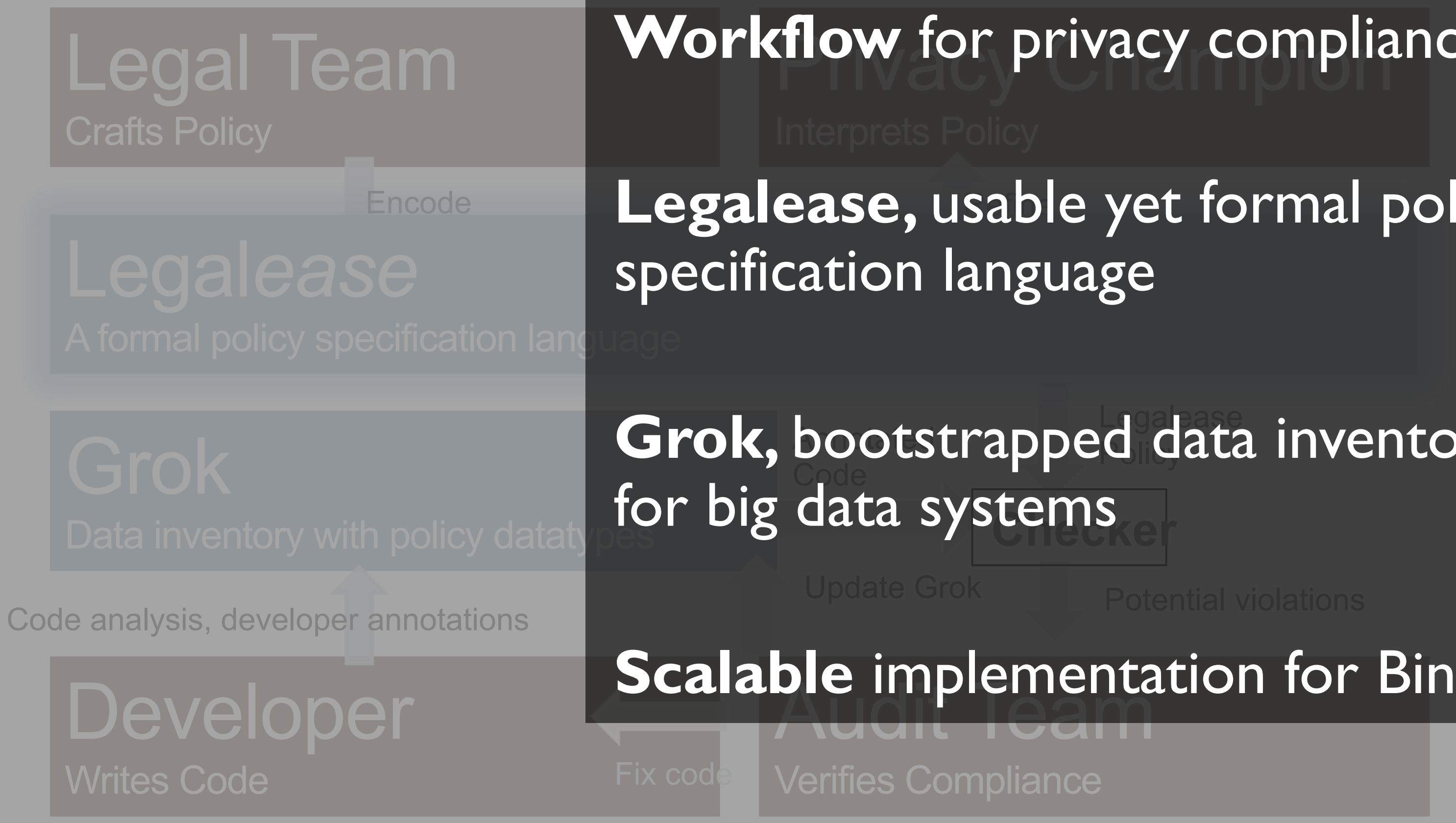
Developer annotations

Audit Team

Verifies Compliance

Fix code

A Streamlined Audit Workflow



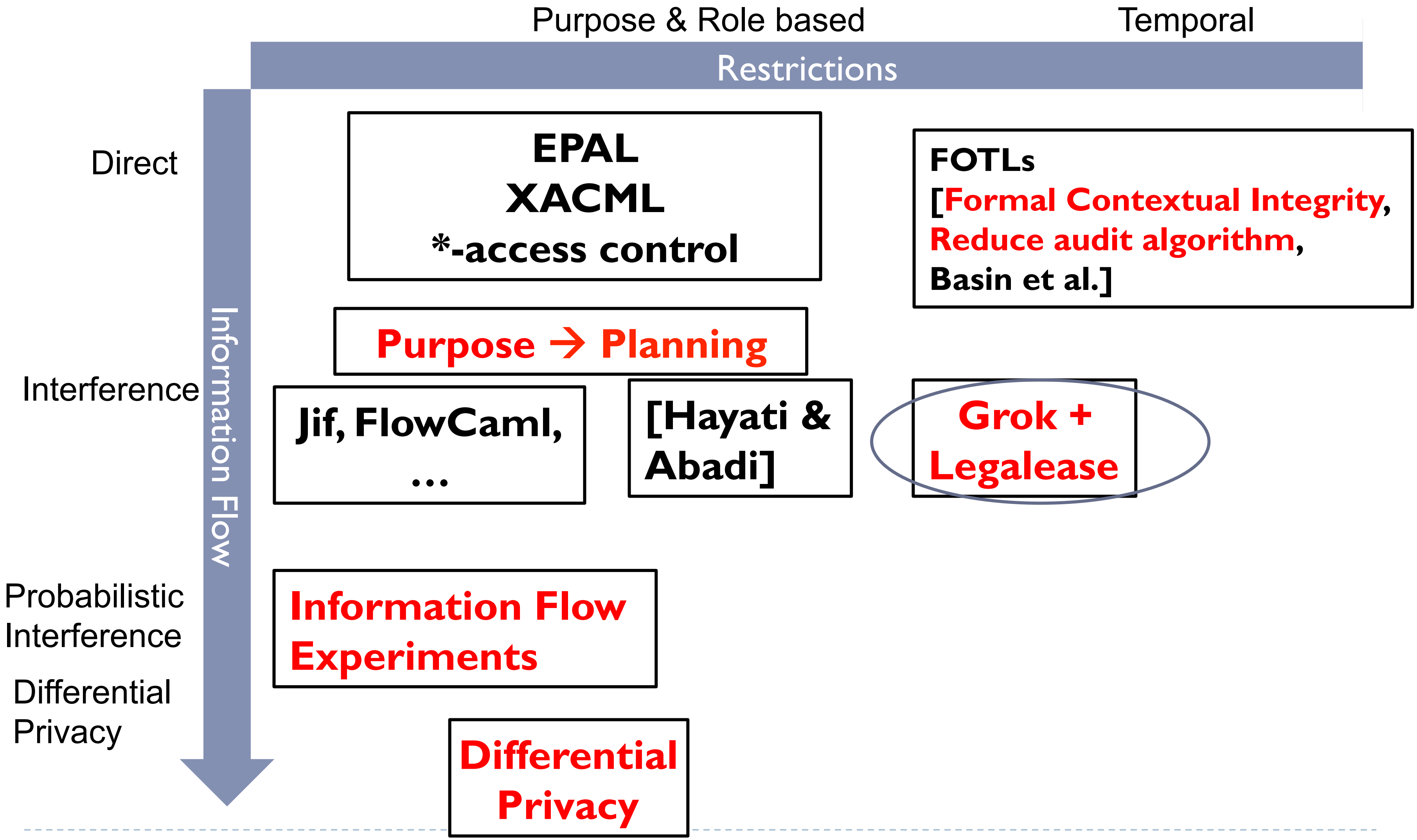
Workflow for privacy compliance

Legalease, usable yet formal policy specification language

Grok, bootstrapped data inventory for big data systems

Scalable implementation for Bing

Privacy as Restrictions on Personal Information Flow



Purpose & Role based

Temporal

Restrictions

Information Flow

Direct

Interference

Probabilistic Interference

Differential Privacy

EPAL
XACML
***-access control**

FOTLs
[Formal Contextual Integrity, Reduce audit algorithm, Basin et al.]

Purpose → Planning

Jif, FlowCaml, …

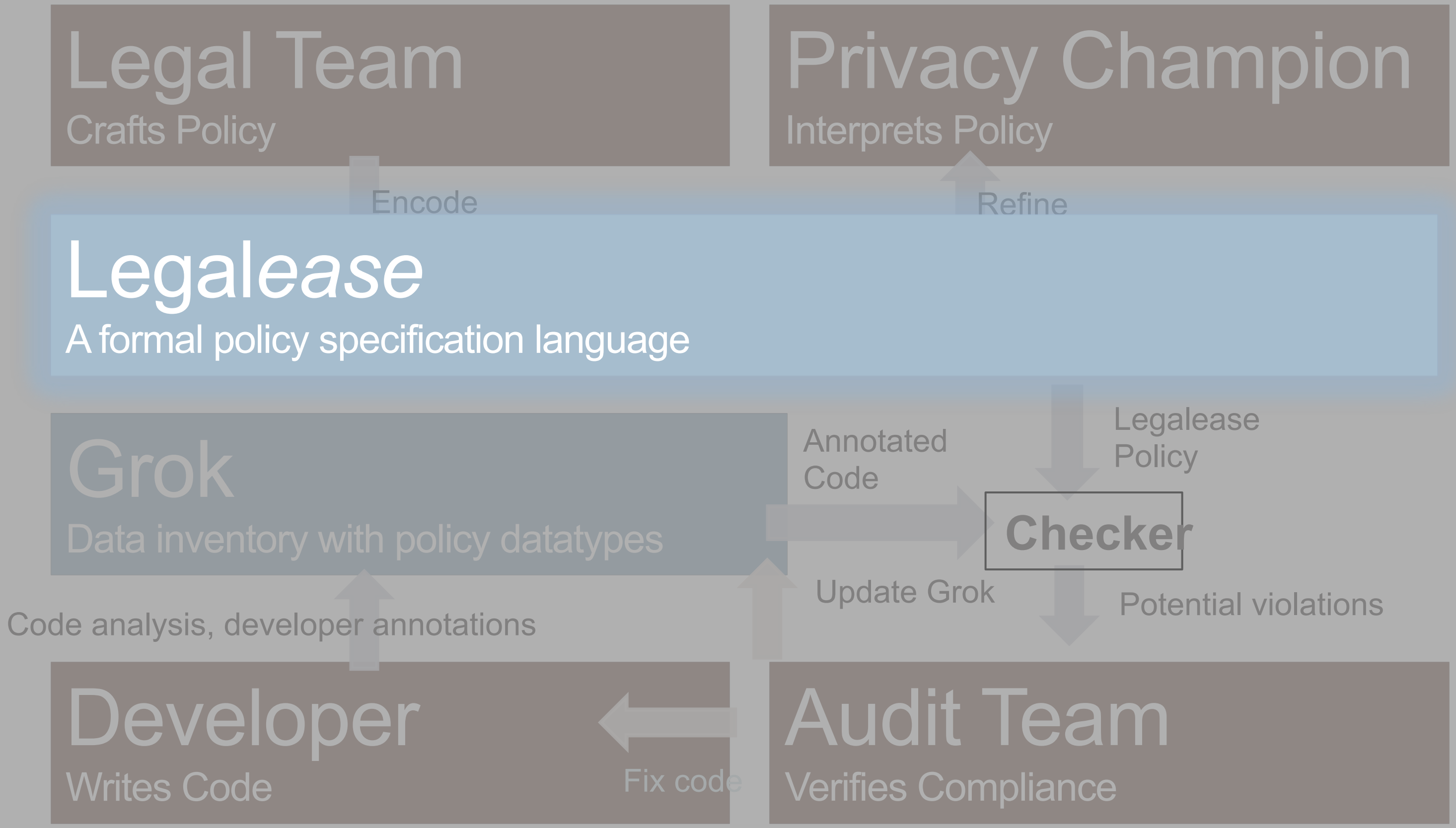
[Hayati & Abadi]

Grok + Legalease

Information Flow Experiments

Differential Privacy

A Streamlined Audit Workflow



Specification: Legal*ease*

Usable. Expressive. Precise.

Usable by lawyers and privacy champs.

Expressive enough for real-world policies.

Precise semantics for local reasoning.

Legalease : Syntax

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

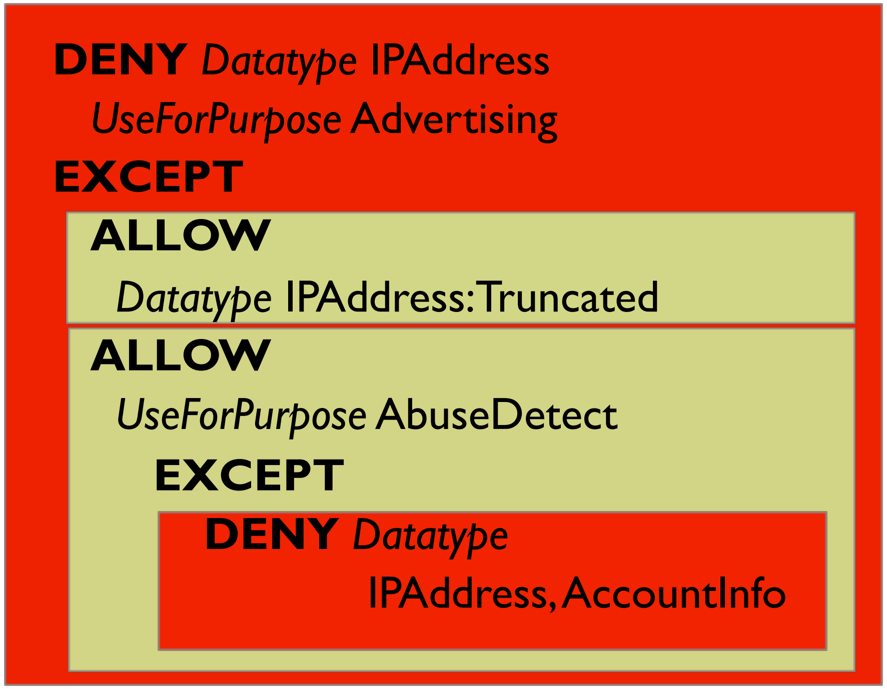
Legalease

DENY *Datatype* IPAddress

UseForPurpose Advertising

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

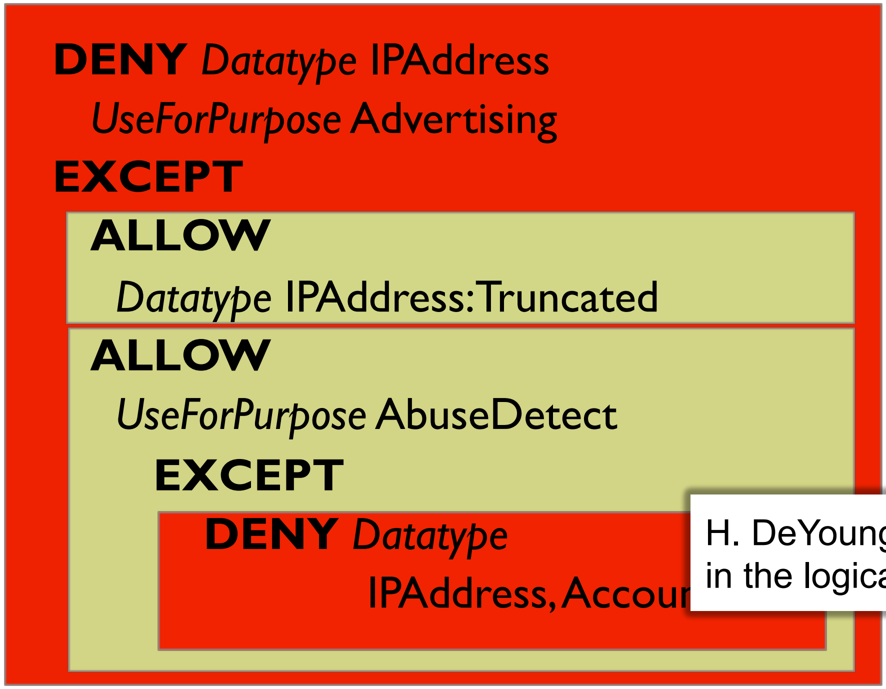
Legalease



```
DENY Datatype IPAddress
  UseForPurpose Advertising
EXCEPT
  ALLOW
    Datatype IPAddress:Truncated
  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.**

Designed for Usability



DENY *Datatype* IPAddress
UseForPurpose Advertising
EXCEPT
ALLOW
Datatype IPAddress:Truncated
ALLOW
UseForPurpose AbuseDetect
EXCEPT
DENY *Datatype*
IPAddress, Accour

Exceptions

- How legal texts are structured
- One-to one correspondence

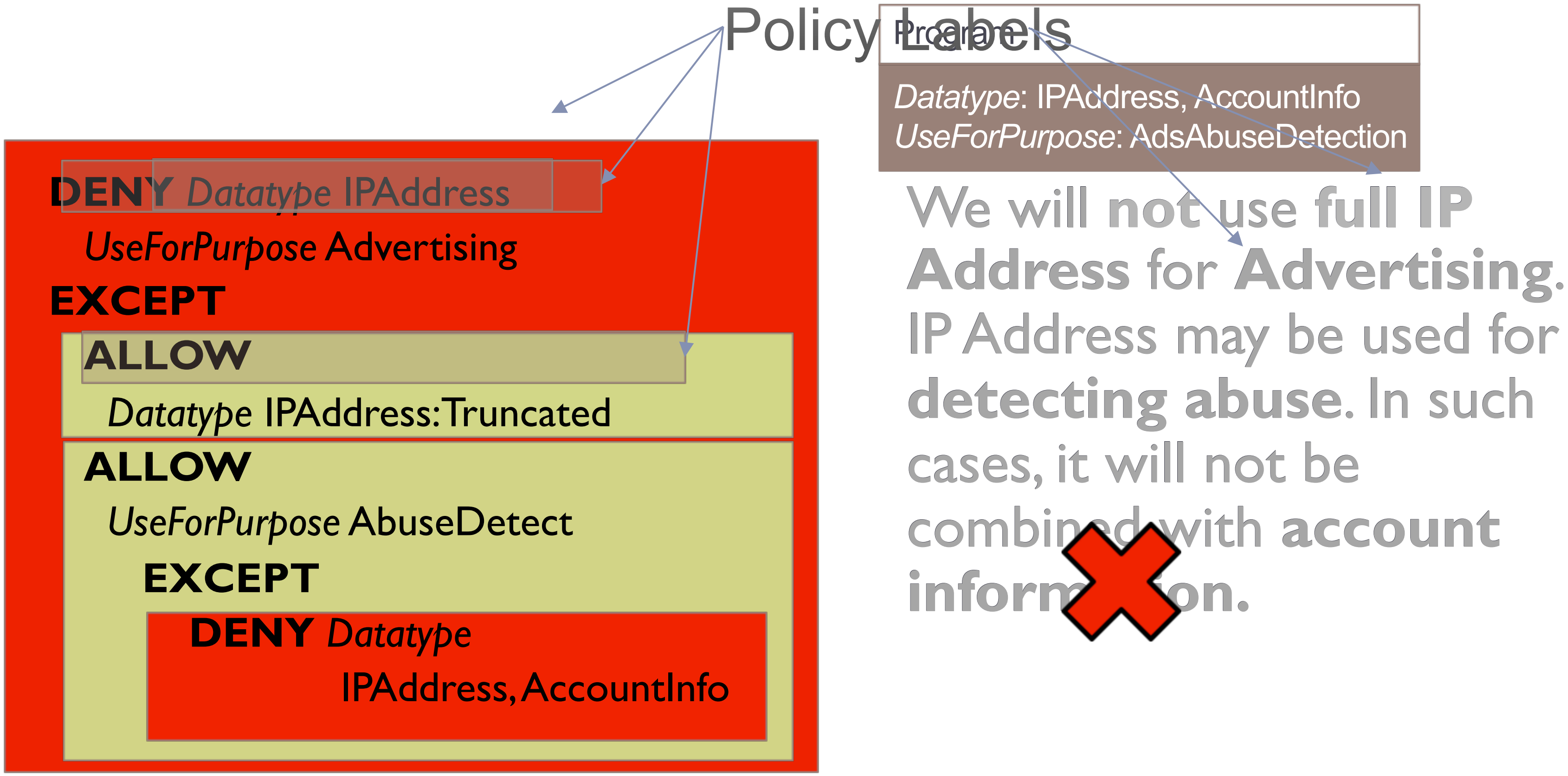
Local Reasoning

- Each exception refines its immediate parent
- Formally proven property

Independent of Code

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

Legalease : In Action



Policy Labels

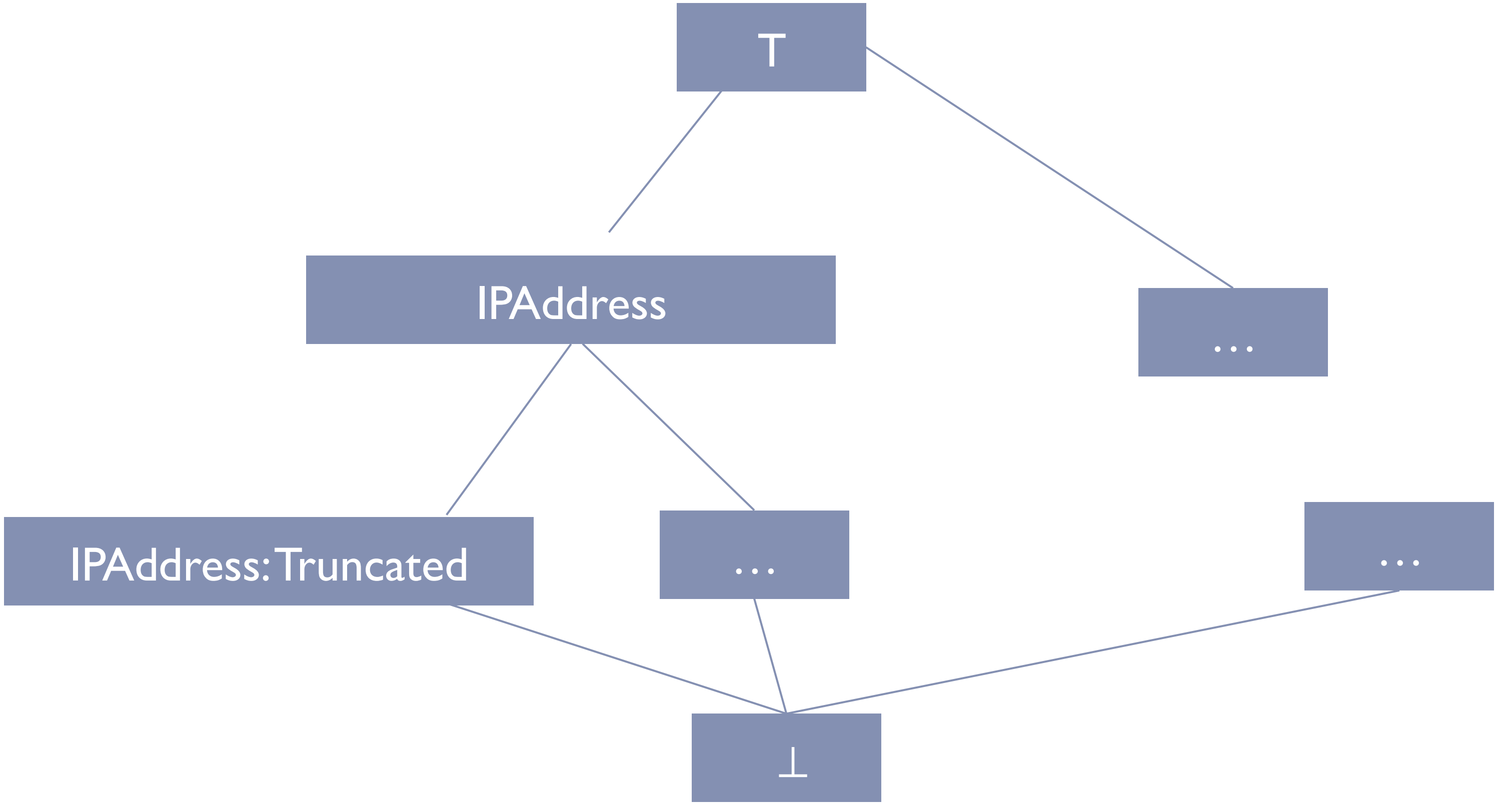
Program

Datatype: IPAddress, AccountInfo
UseForPurpose: AdsAbuseDetection

DENY *Datatype* IPAddress
UseForPurpose Advertising
EXCEPT
ALLOW
Datatype IPAddress:Truncated
ALLOW
UseForPurpose AbuseDetect
EXCEPT
DENY *Datatype*
IPAddress, AccountInfo

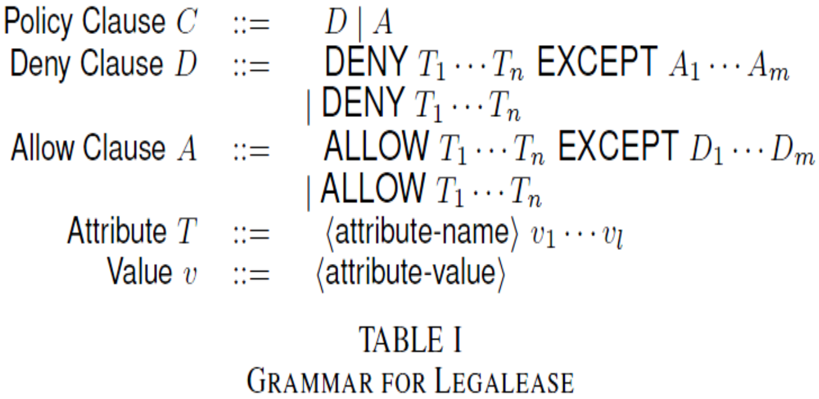
We will **not** use **full IP Address** for **Advertising**. IP Address may be used for **detecting abuse**. In such cases, it will not be combined with **account information.**

A Lattice of Policy Labels



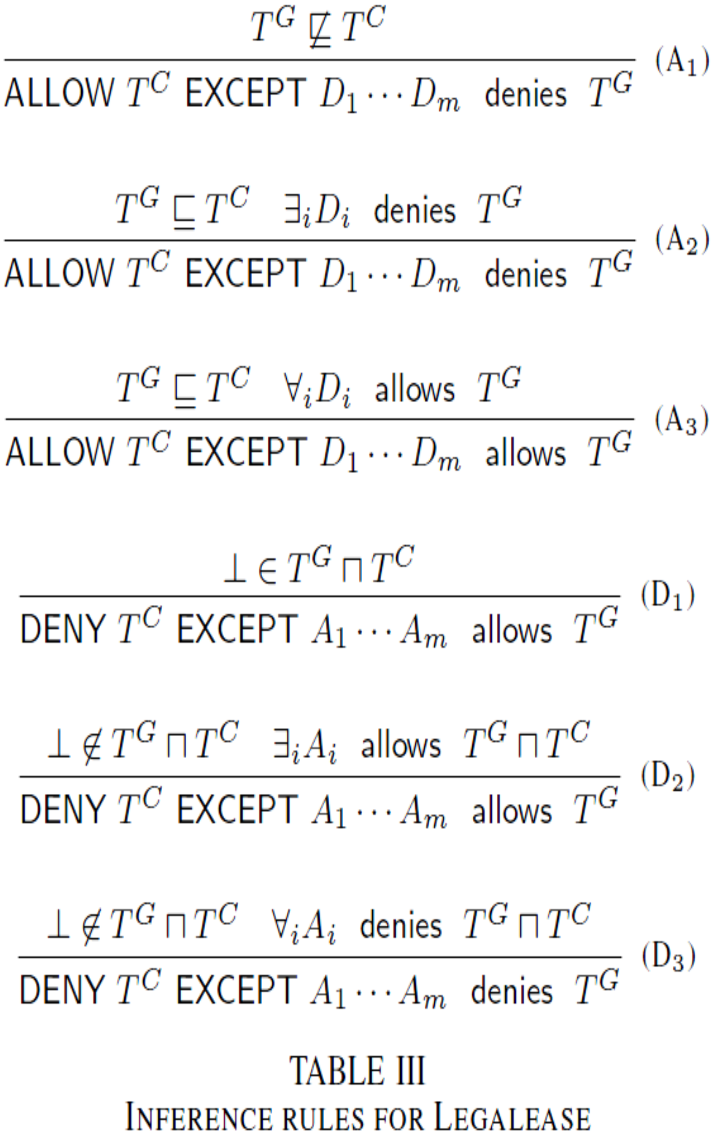
- If “IPAddress” use is allowed then so is everything below it
- If “IPAddress:Truncated” use is denied then so is everything above it

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	::=	⟨attribute-name⟩ $v_1 \cdots v_l$
Value v	::=	⟨attribute-value⟩

TABLE I
GRAMMAR FOR LEGALEASE



$$\frac{T^G \not\sqsubseteq T^C}{\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m \ \text{ denies } \ T^G} \ (\mathrm{A}_1)$$

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

$$\frac{T^G \sqsubseteq T^C \quad \forall_i D_i \ \text{ allows } \ T^G}{\text{ALLOW } T^C \text{ EXCEPT } D_1 \cdots D_m \ \text{ allows } \ T^G} \ (\mathrm{A}_3)$$

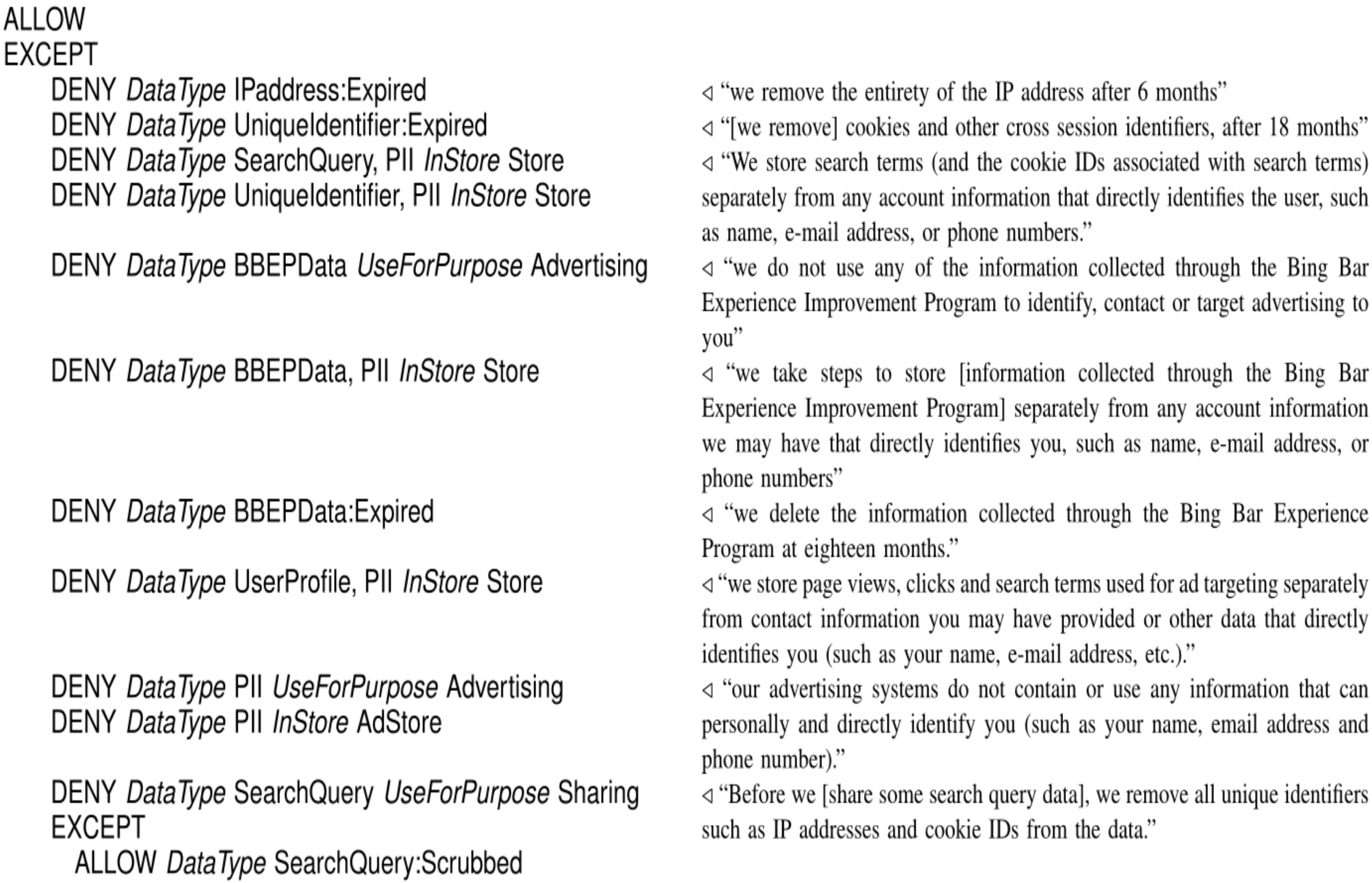
$$\frac{\bot \in T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \ \text{ allows } \ T^G} \ (\mathrm{D}_1)$$

$$\frac{\bot \notin T^G \sqcap T^C \quad \exists_i A_i \ \text{ allows } \ T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \ \text{ allows } \ T^G} \ (\mathrm{D}_2)$$

$$\frac{\bot \notin T^G \sqcap T^C \quad \forall_i A_i \ \text{ denies } \ T^G \sqcap T^C}{\text{DENY } T^C \text{ EXCEPT } A_1 \cdots A_m \ \text{ denies } \ T^G} \ (\mathrm{D}_3)$$

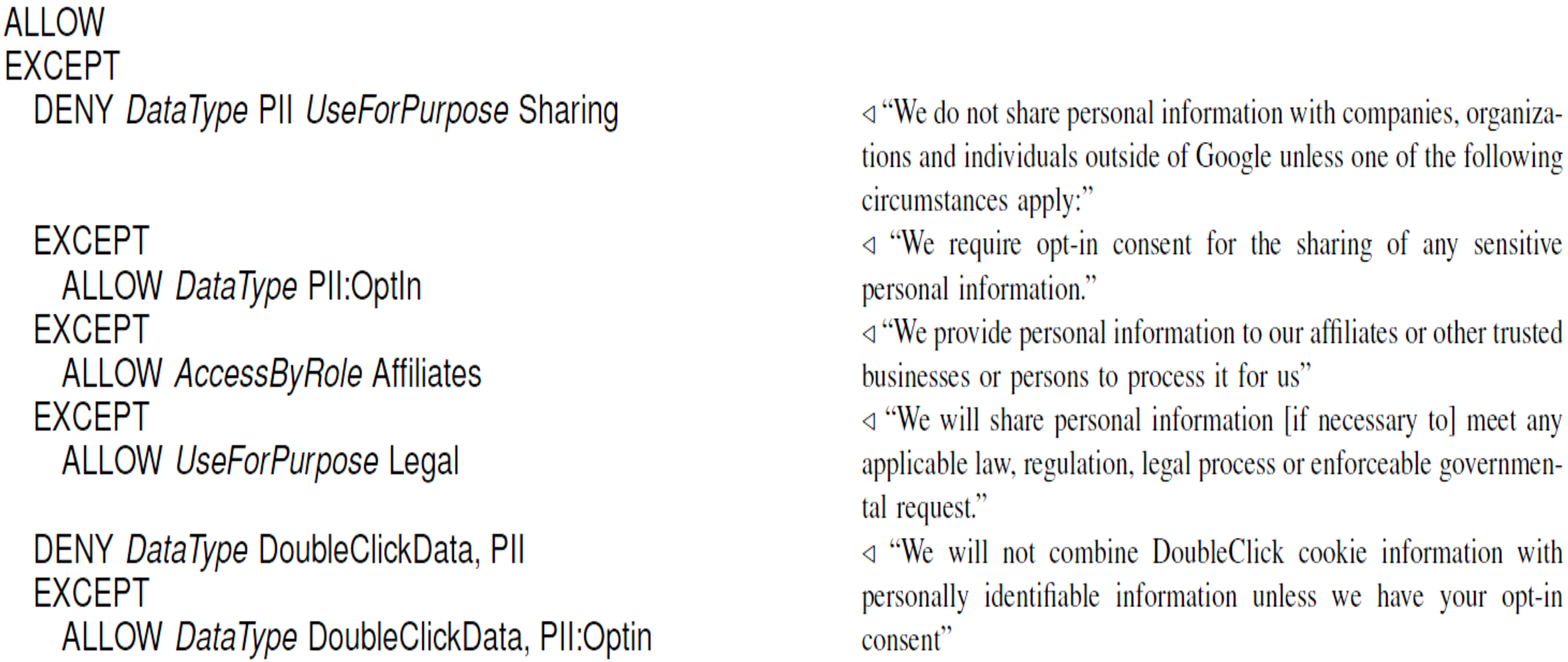
TABLE III
INFERENCE RULES FOR LEGALEASE

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

  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:"

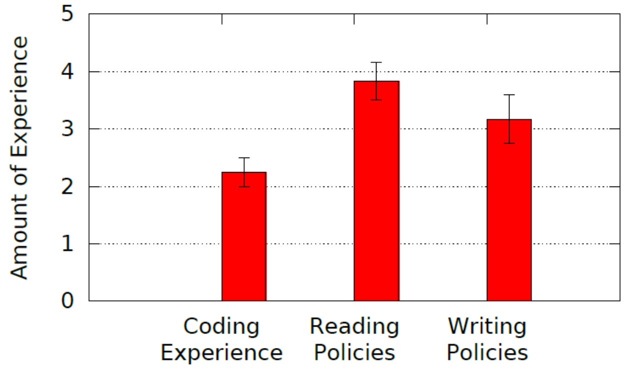
◁ "We require opt-in consent for the sharing of any sensitive personal information."

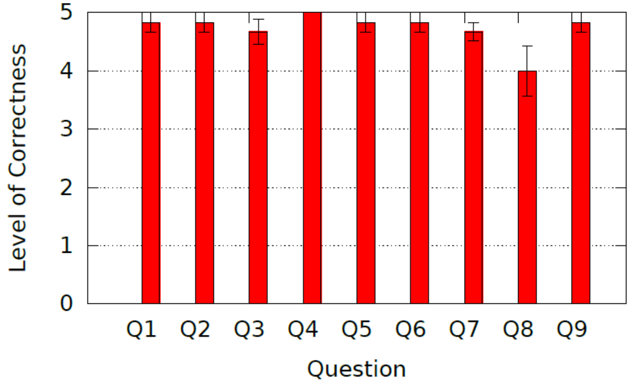
◁ "We provide personal information to our affiliates or other trusted businesses or persons to process it for us"

◁ "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"

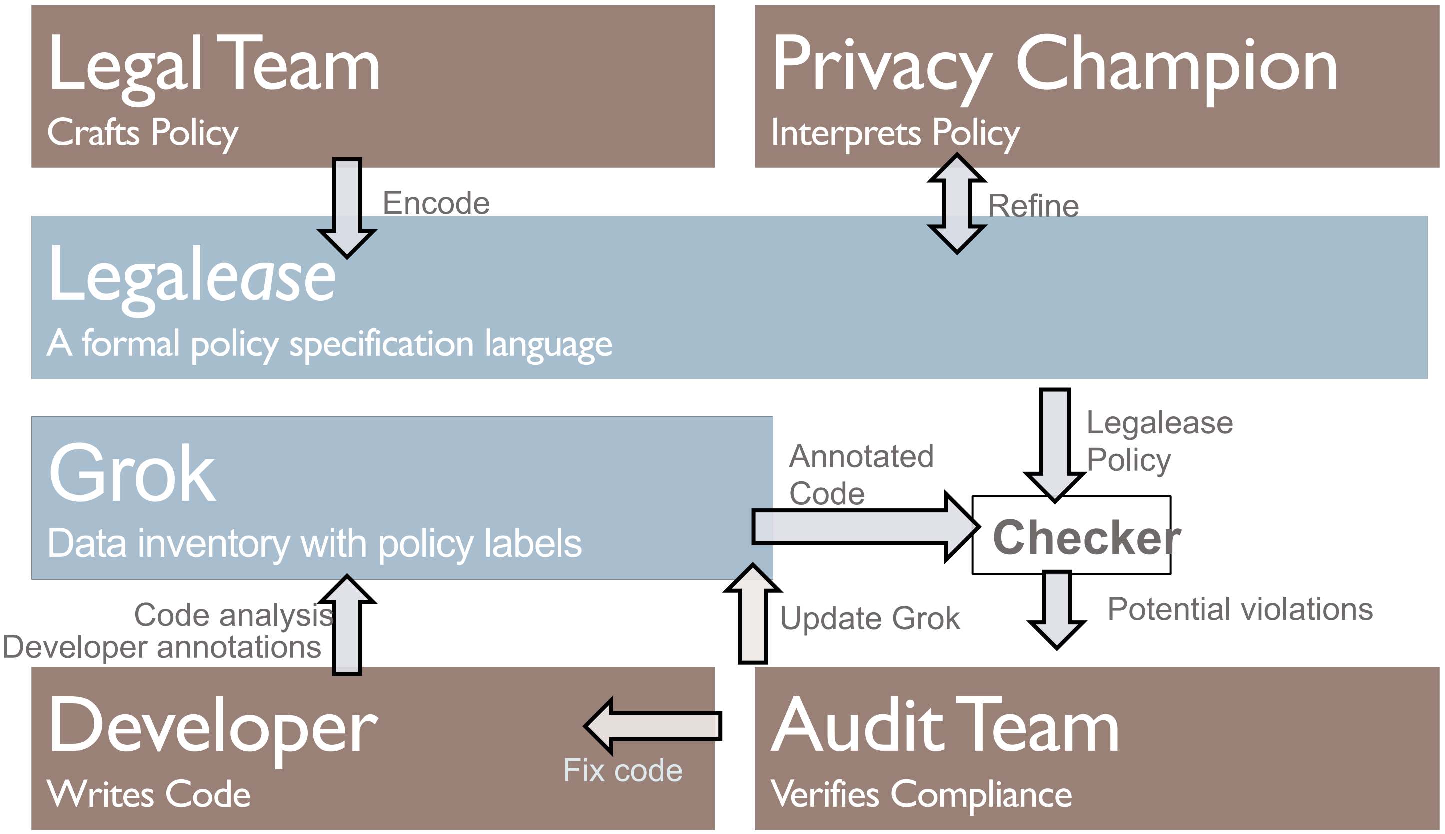
Legalease Usability





- 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



Legal Team

Crafts Policy

Privacy Champion

Interprets Policy

Encode

Refine

Legalease

A formal policy specification language

Grok

Data inventory with policy labels

Legalease Policy

Annotated Code

Checker

Code analysis
Developer annotations

Update Grok

Potential violations

Developer

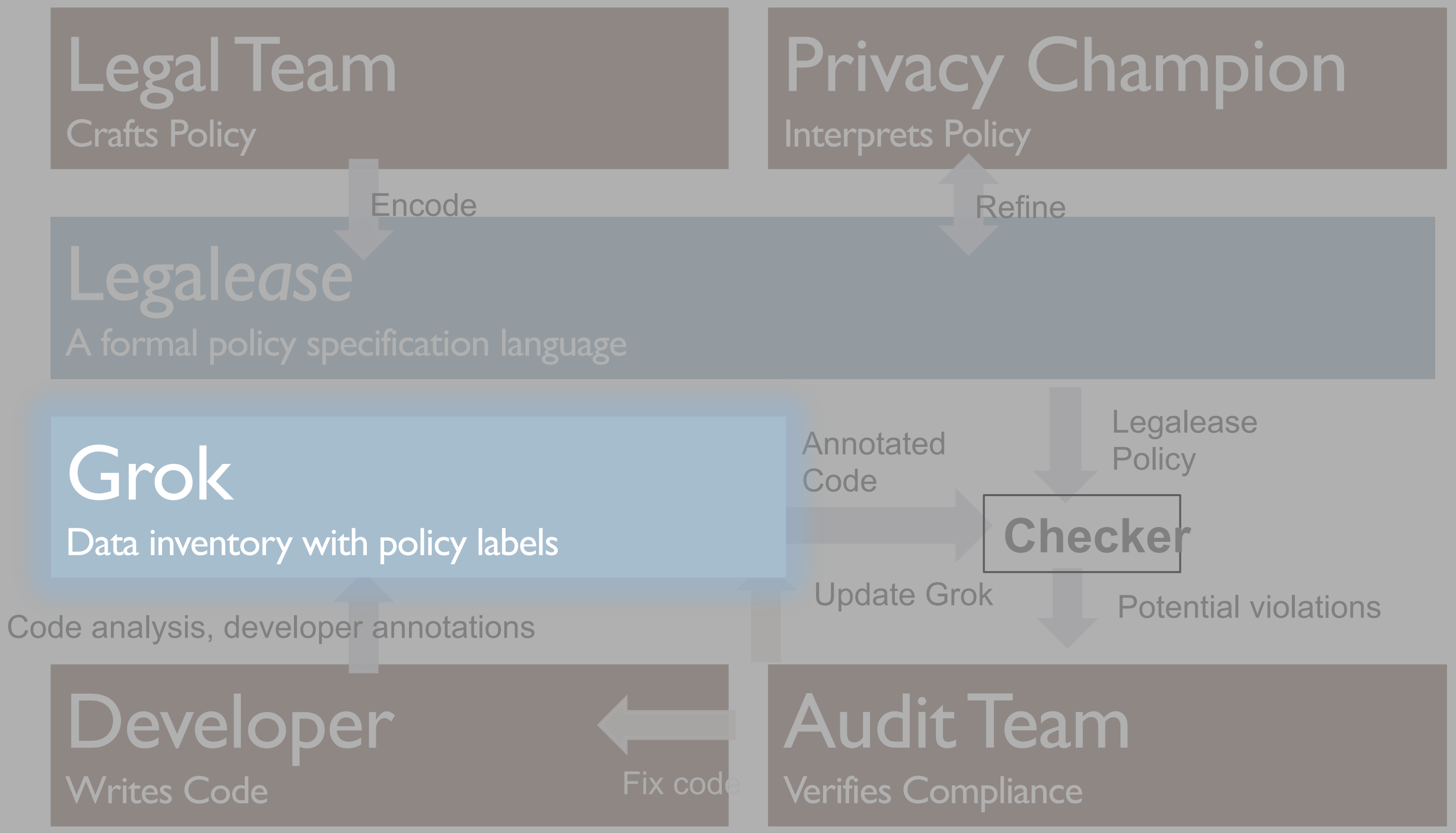
Writes Code

Fix code

Audit Team

Verifies Compliance

A Streamlined Audit Workflow



Legal Team

Crafts Policy

Privacy Champion

Interprets Policy

Encode

Refine

Legalease

A formal policy specification language

Grok

Data inventory with policy labels

Annotated Code

Legalease Policy

Checker

Update Grok

Potential violations

Code analysis, developer annotations

Developer

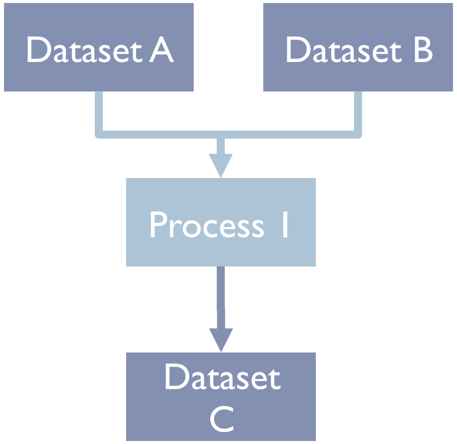
Writes Code

Fix code

Audit Team

Verifies Compliance

Map-Reduce Programming Systems



Scope, Hive, Dremel

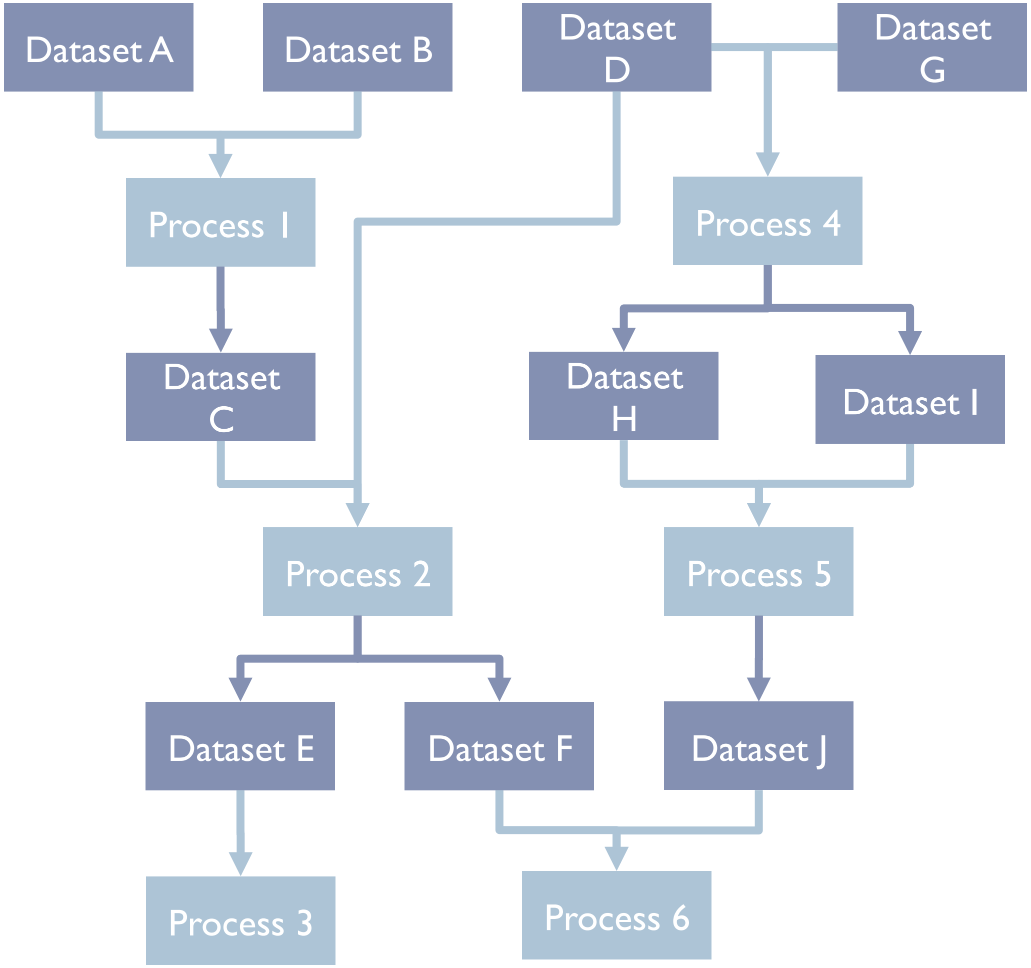
Data in the form of Tables

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
```

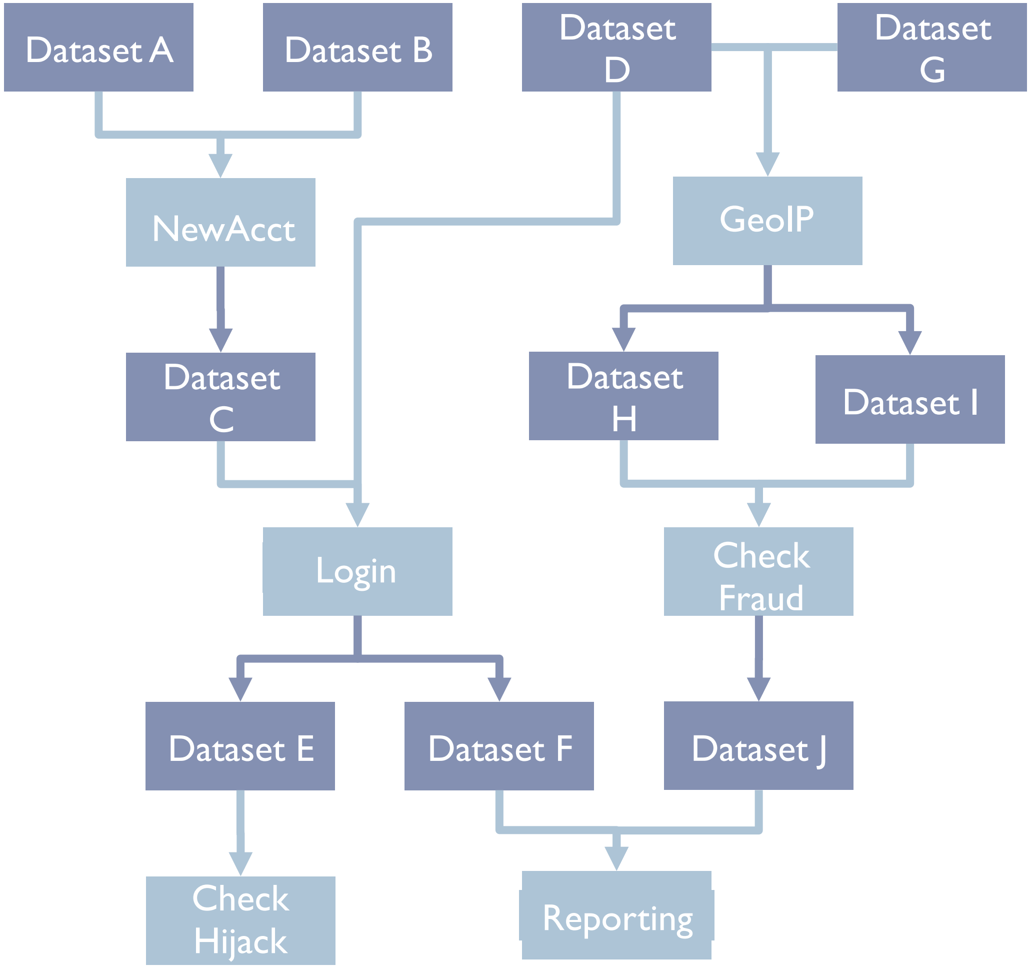
Grok


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



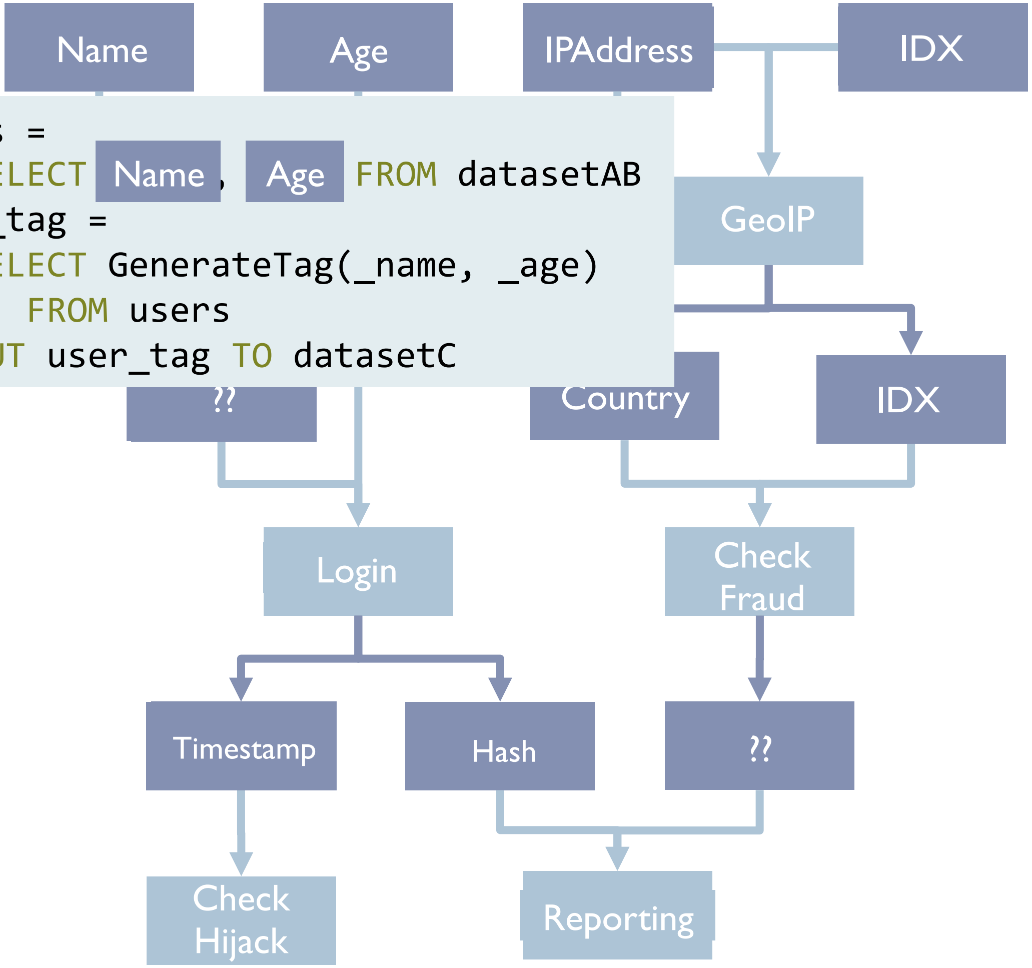
Grok

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

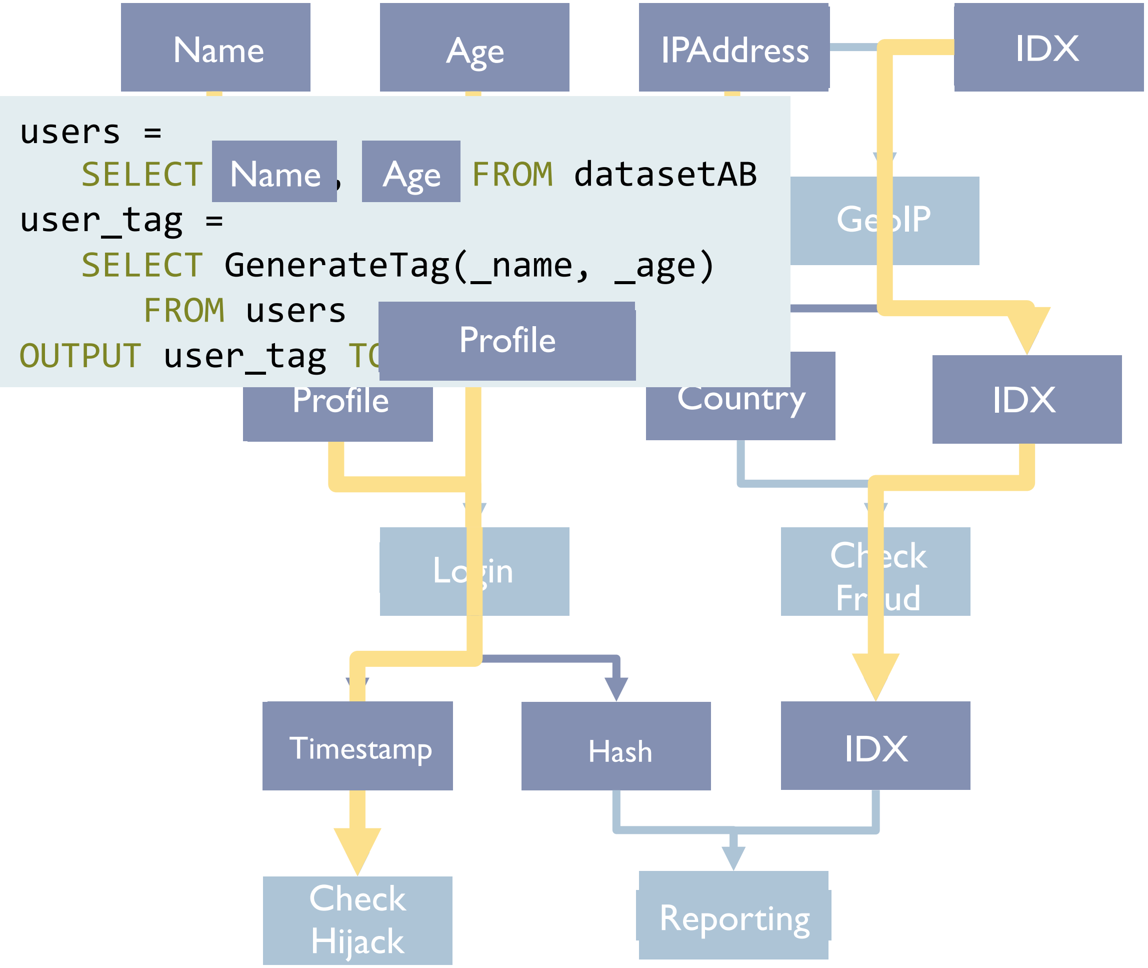
Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations

Flow Labels

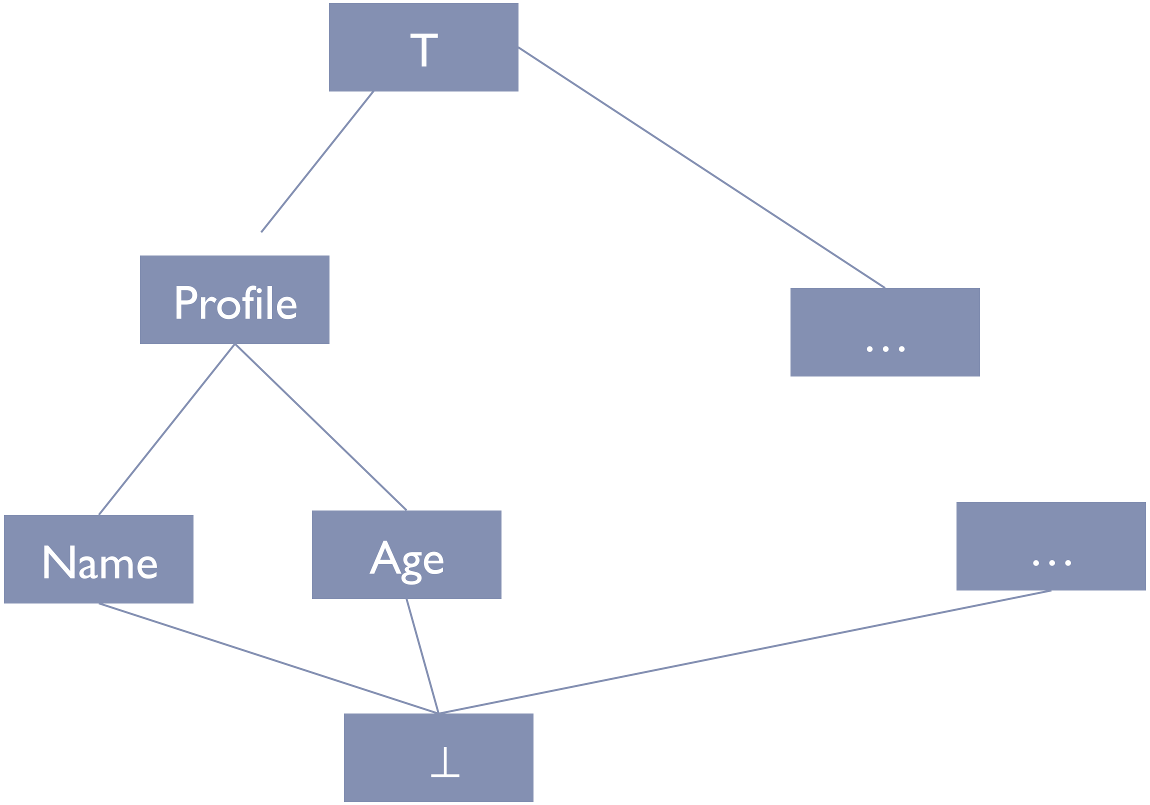
Source labels propagated via data flow graph



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

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

Carefully curated regular expressions

Leverages developer conventions

Significant Noise

Variable Name Analysis

Expensive

Low Noise

Developer Annotations

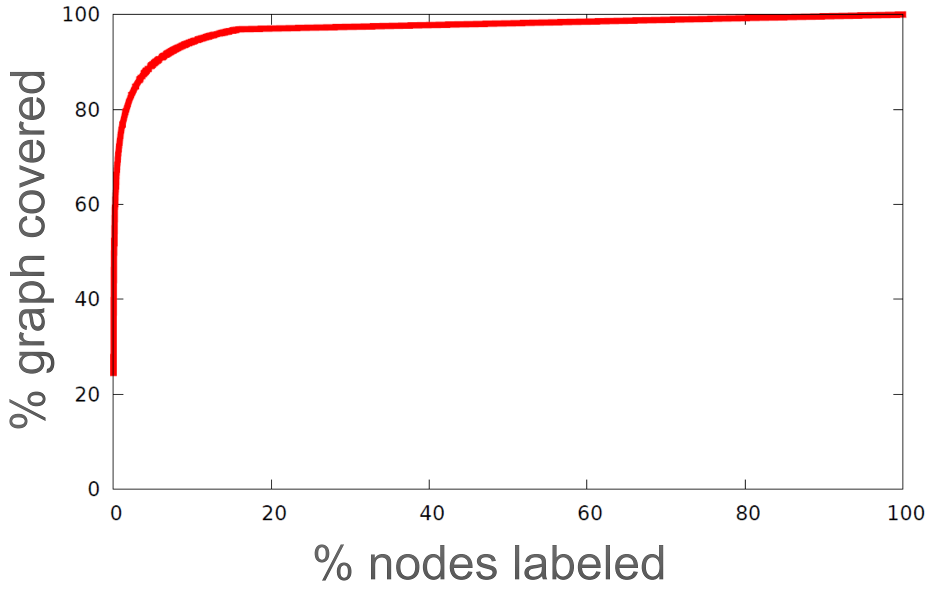
Very Expensive

Definitive

Need very few of these

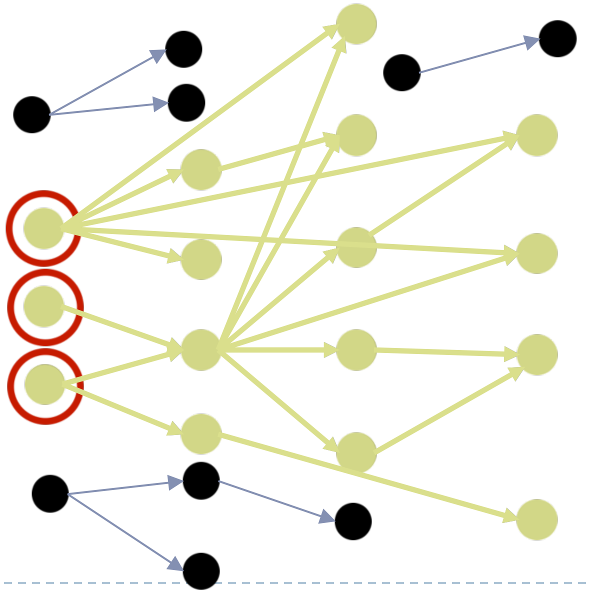
Auditor Verification

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

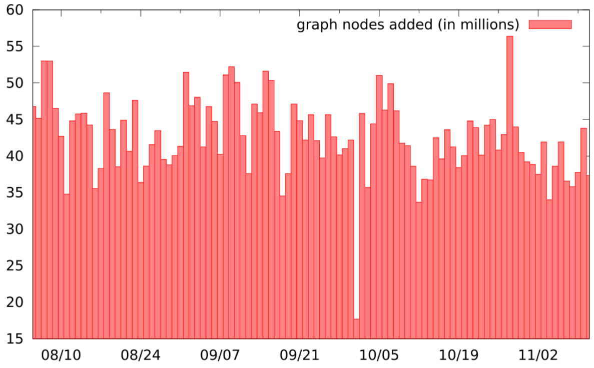
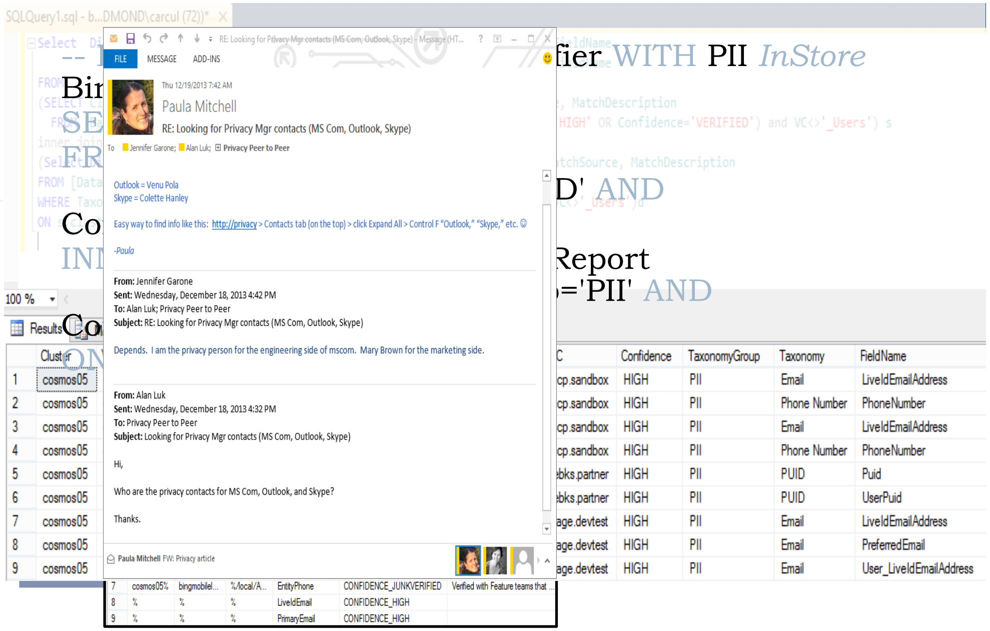


Fig. 9. Number of GROK data flow graph nodes added each day

- 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


SQLQuery1.sql - b...DMOND\carcul (72))*
FILE
MESSAGE
ADD-INS
RE: Looking for Privacy Mgr contacts (MS Com, Outlook, Skype) - Message (HT...
100 %
Results
Bin
SE
FR
Co
INN
Co
ON
fier WITH PII InStore
D' AND
Report
='PII' AND

Thu 12/19/2013 7:42 AM

Paula Mitchell

RE: Looking for Privacy Mgr contacts (MS Com, Outlook, Skype)

To Jennifer Garone; Alan Luk; Privacy Peer to Peer

Outlook = Venu Pola
Skype = Colette Hanley

Easy way to find info like this: http://privacy > Contacts tab (on the top) > click Expand All > Control F "Outlook," "Skype," etc. ☺

-Paula

From: Jennifer Garone
Sent: Wednesday, December 18, 2013 4:42 PM
To: Alan Luk; Privacy Peer to Peer
Subject: RE: Looking for Privacy Mgr contacts (MS Com, Outlook, Skype)

Depends. I am the privacy person for the engineering side of mscom. Mary Brown for the marketing side.

From: Alan Luk
Sent: Wednesday, December 18, 2013 4:32 PM
To: Privacy Peer to Peer
Subject: Looking for Privacy Mgr contacts (MS Com, Outlook, Skype)

Hi,

Who are the privacy contacts for MS Com, Outlook, and Skype?

Thanks.

Paula Mitchell FW: Privacy article

	Cluster	C	Confidence	TaxonomyGroup	Taxonomy	FieldName
1	cosmos05	cp.sandbox	HIGH	PII	Email	LiveIdEmailAddress
2	cosmos05	cp.sandbox	HIGH	PII	Phone Number	PhoneNumber
3	cosmos05	cp.sandbox	HIGH	PII	Email	LiveIdEmailAddress
4	cosmos05	cp.sandbox	HIGH	PII	Phone Number	PhoneNumber
5	cosmos05	ebks.partner	HIGH	PII	PUID	Puid
6	cosmos05	ebks.partner	HIGH	PII	PUID	UserPuid
7	cosmos05	age.devtest	HIGH	PII	Email	LiveIdEmailAddress
8	cosmos05	age.devtest	HIGH	PII	Email	PreferredEmail
9	cosmos05	age.devtest	HIGH	PII	Email	User_LiveIdEmailAddress

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

Static code analysis

files
25M+

schemas
2M+

Generate report

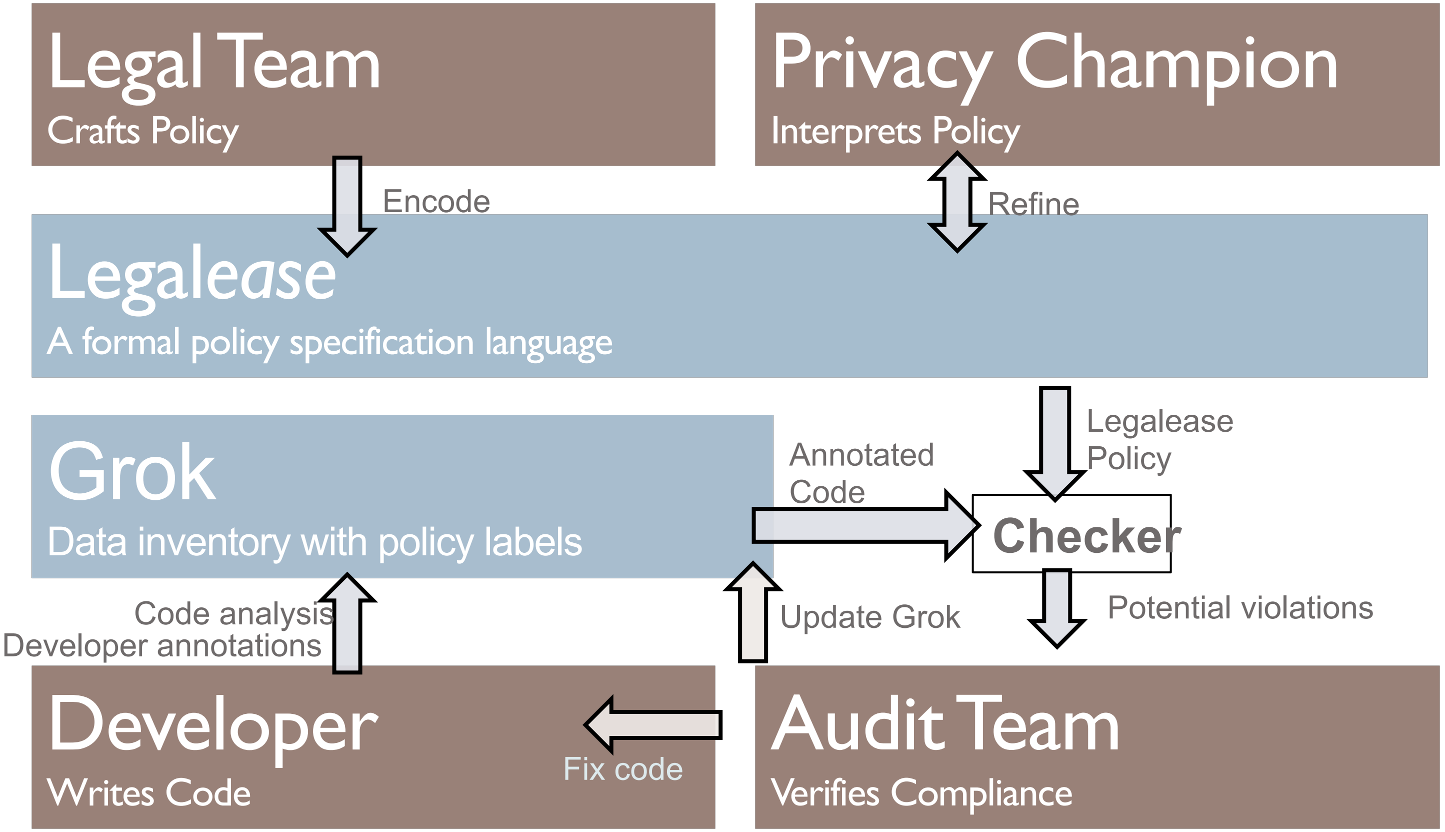
privacy
candidates
300K+

audit
alerts*
100K+

Manual Audit

teams
8

A Streamlined Audit Workflow



Legal Team
Crafts Policy

Privacy Champion
Interprets Policy

Encode

Refine

Legalease
A formal policy specification language

Grok
Data inventory with policy labels

Annotated Code

Legalease Policy

Checker

Code analysis
Developer annotations

Update Grok

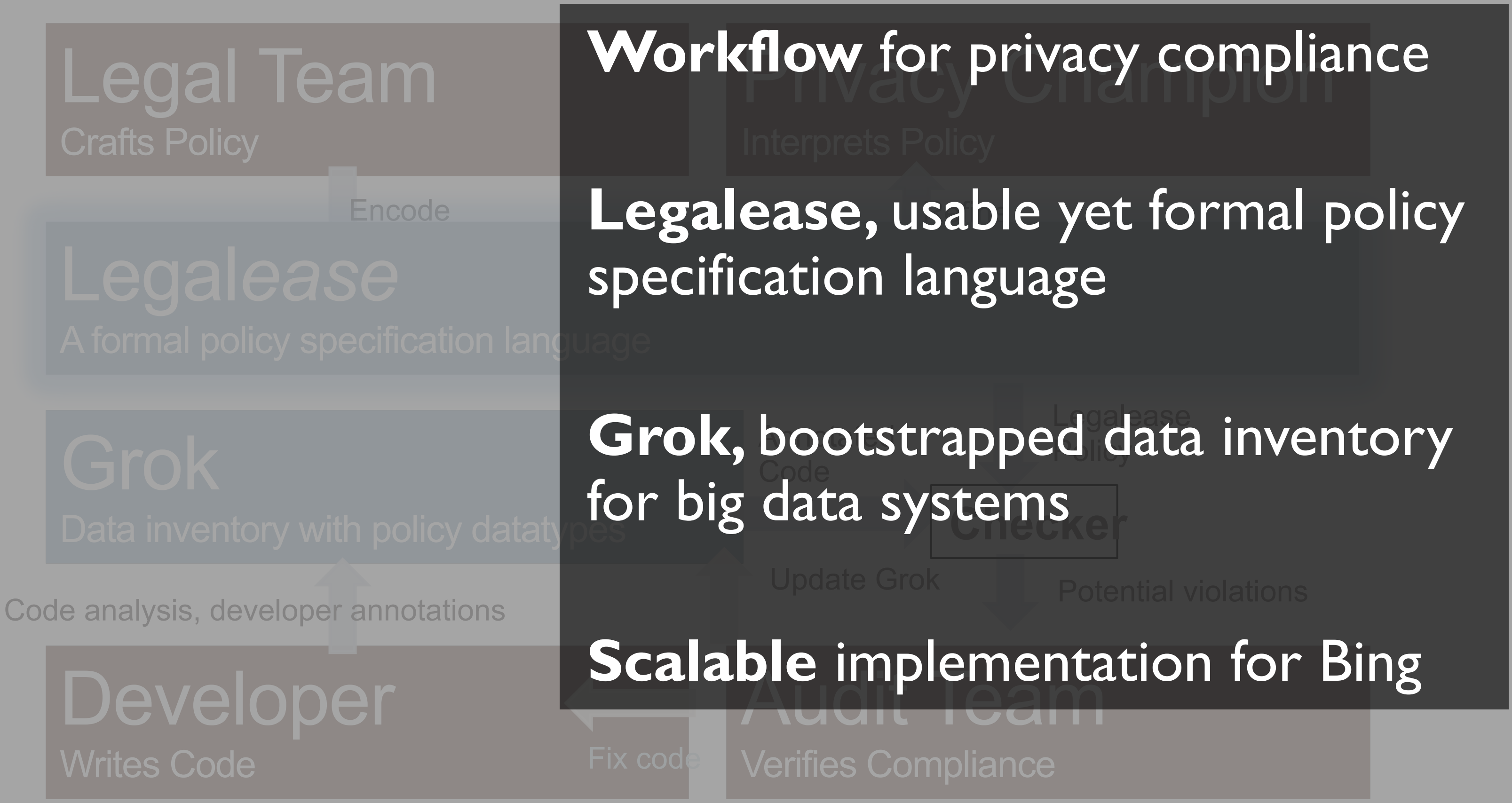
Potential violations

Developer
Writes Code

Fix code

Audit Team
Verifies Compliance

A Streamlined Audit Workflow



Workflow for privacy compliance

Legalease, usable yet formal policy specification language

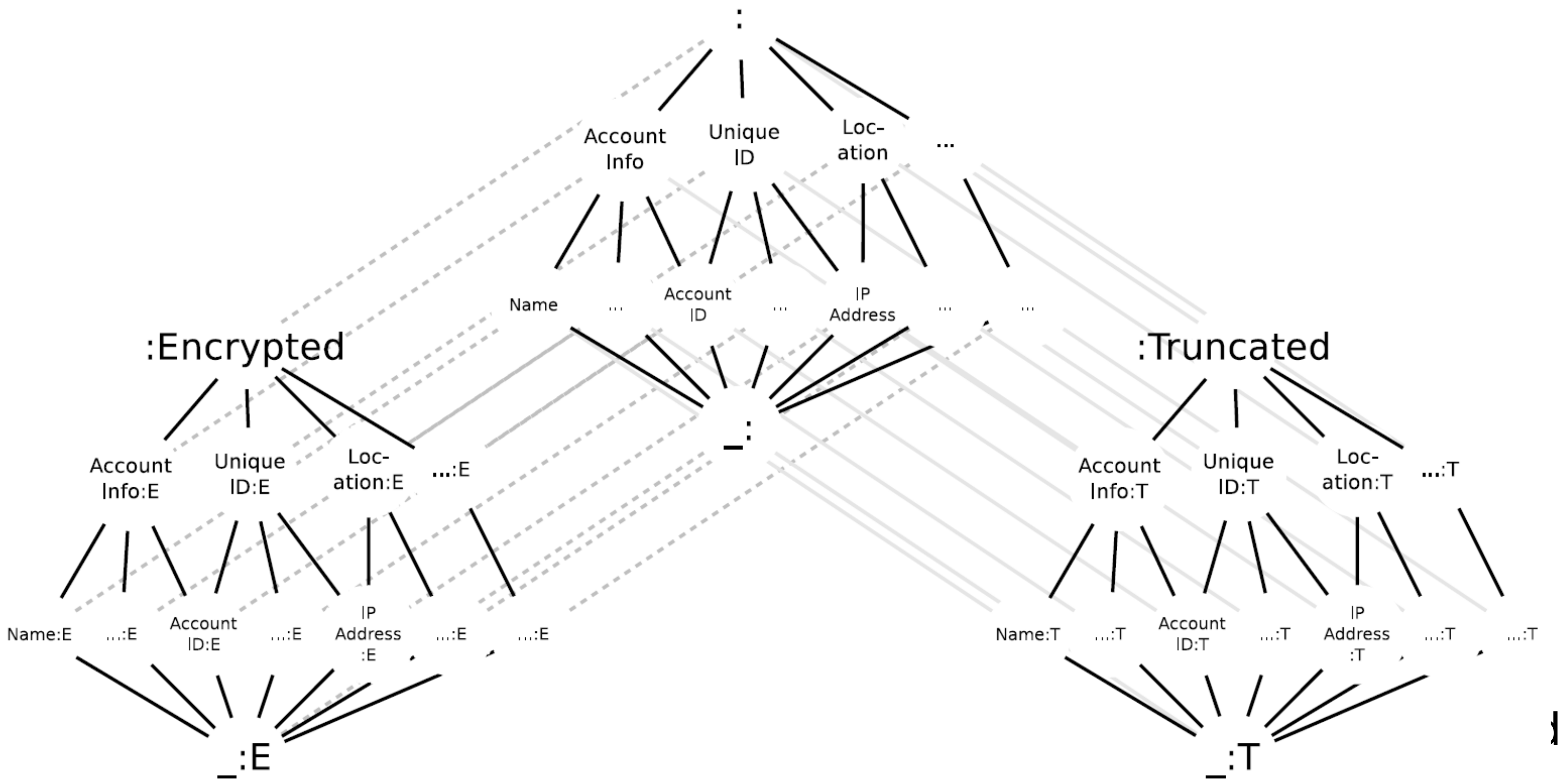
Grok, bootstrapped data inventory for big data systems

Scalable implementation for Bing

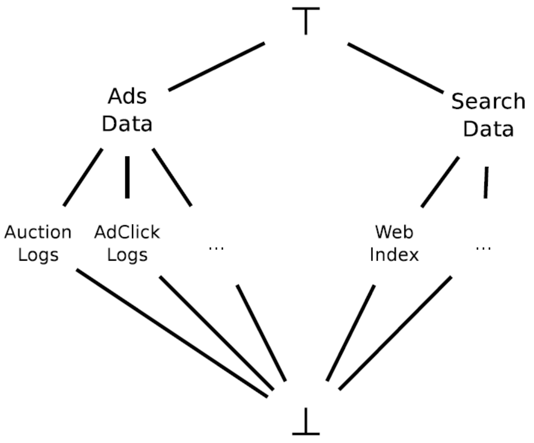
Reference

- S. Sen, S. Guha, A. Datta, S. Rajamani, J. Tsai, J. M. Wing, Bootstrapping Privacy Compliance in Big Data Systems, in *Proceedings of 35th IEEE Symposium on Security and Privacy*, May 2014.

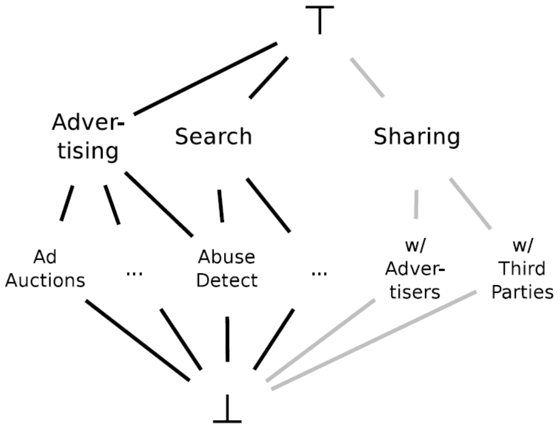
Policy Labels : Datatypes



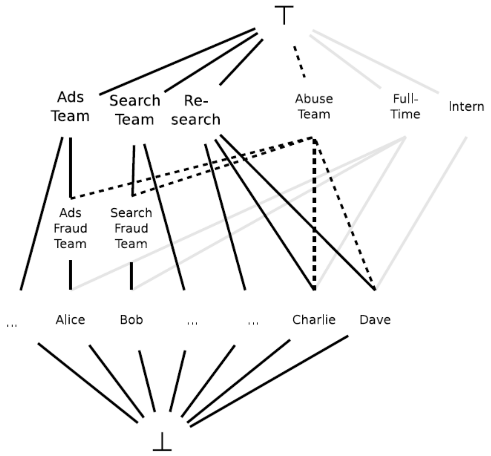
Policy Types : Concept Lattices



InStore Lattice



UseForPurpose Lattice



AccessByRole Lattice

Formal Semantics

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

Based on Lattice Orderings on Policy Types

Formal Semantics

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

Recursively check exceptions

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