

Bootstrapping Privacy Compliance in Big Data Systems

Anupam Datta
CMU

Fall 2016

Privacy Compliance for Bing

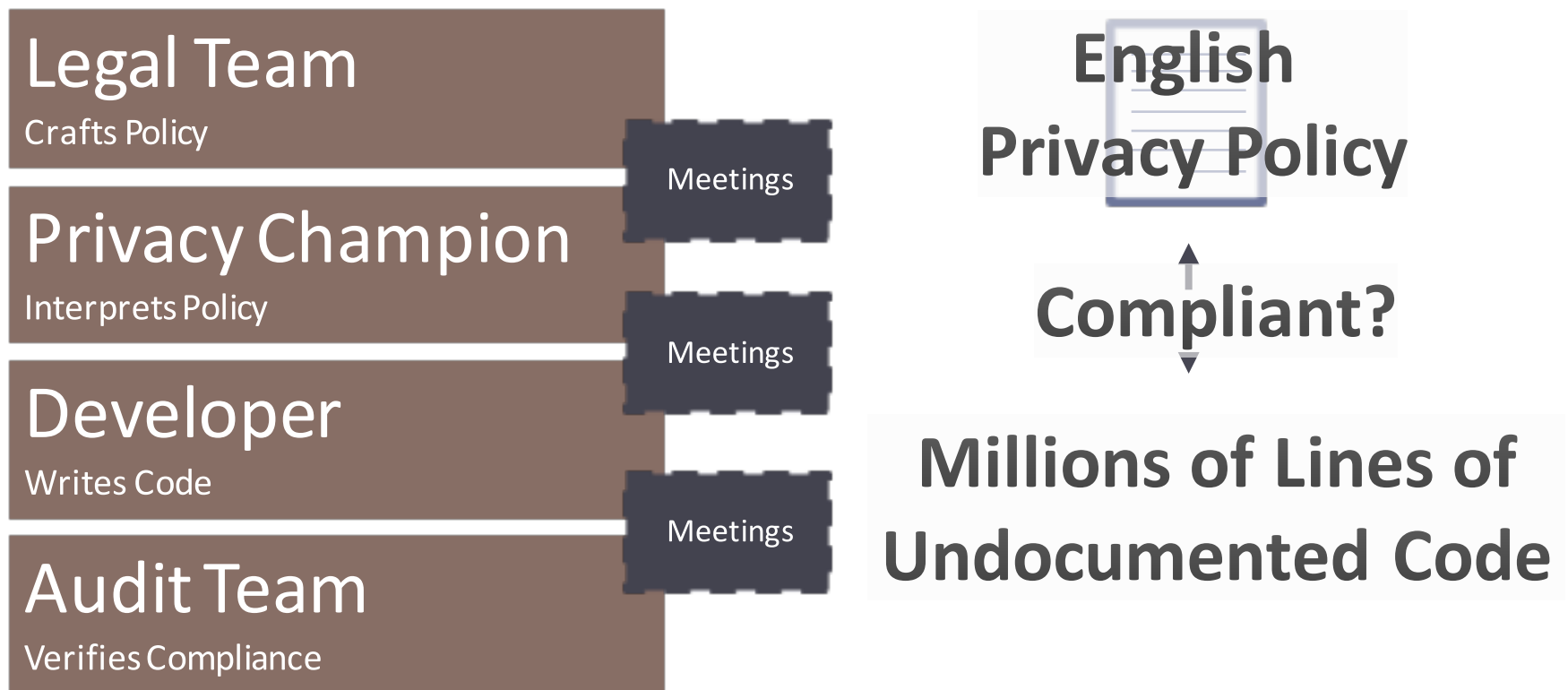
The image shows a side-by-side comparison of two web pages in a browser window. On the left is the Bing homepage (http://www.bing.com/), featuring a search bar, navigation links (IMAGES, VIDEOS, MAPS, NEWS, SEARCH HISTORY, MORE), and a cartoon detective character. On the right is the Bing Privacy Statement page (http://www.microsoft.com/privacystatement/en-us/b...), which contains detailed information about data collection and user rights. The privacy statement is organized into sections: Cookies & Similar Technologies, Collecting Your Information, and How We Use Your Personal Information. A vertical navigation menu on the right side of the privacy statement lists various topics like Cookies, Collecting Your Information, Using Your Information, etc.

Setting:

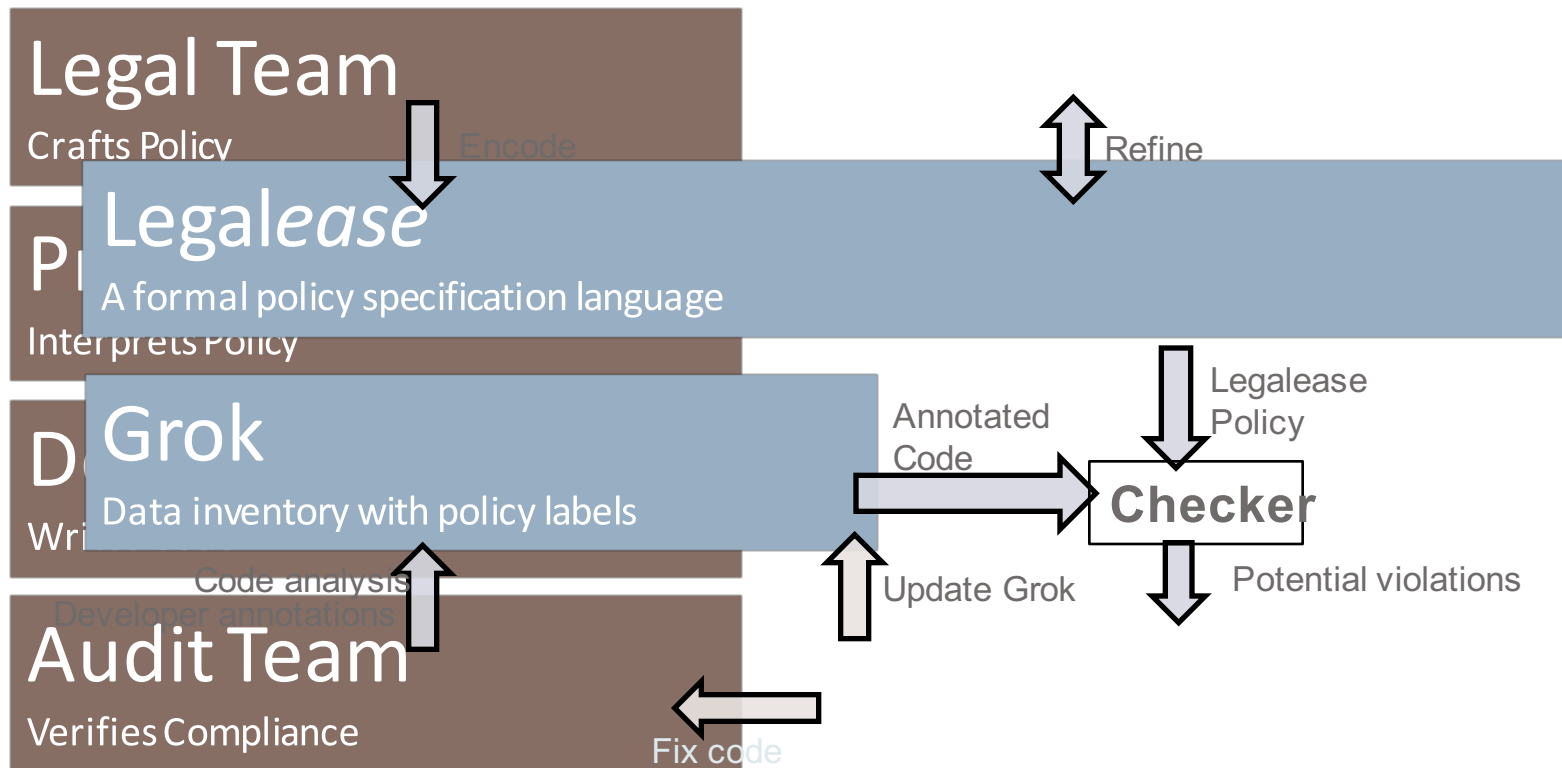
- ▶ Auditor has access to source code



The Privacy Compliance Challenge



A Streamlined Audit Workflow



A Streamlined Audit Workflow

Legal Team

Crafts Policy

Encode

Legalease

A formal policy specification language

Grok

Data inventory with policy datatypes

Code analysis, developer annotations

Developer

Writes Code

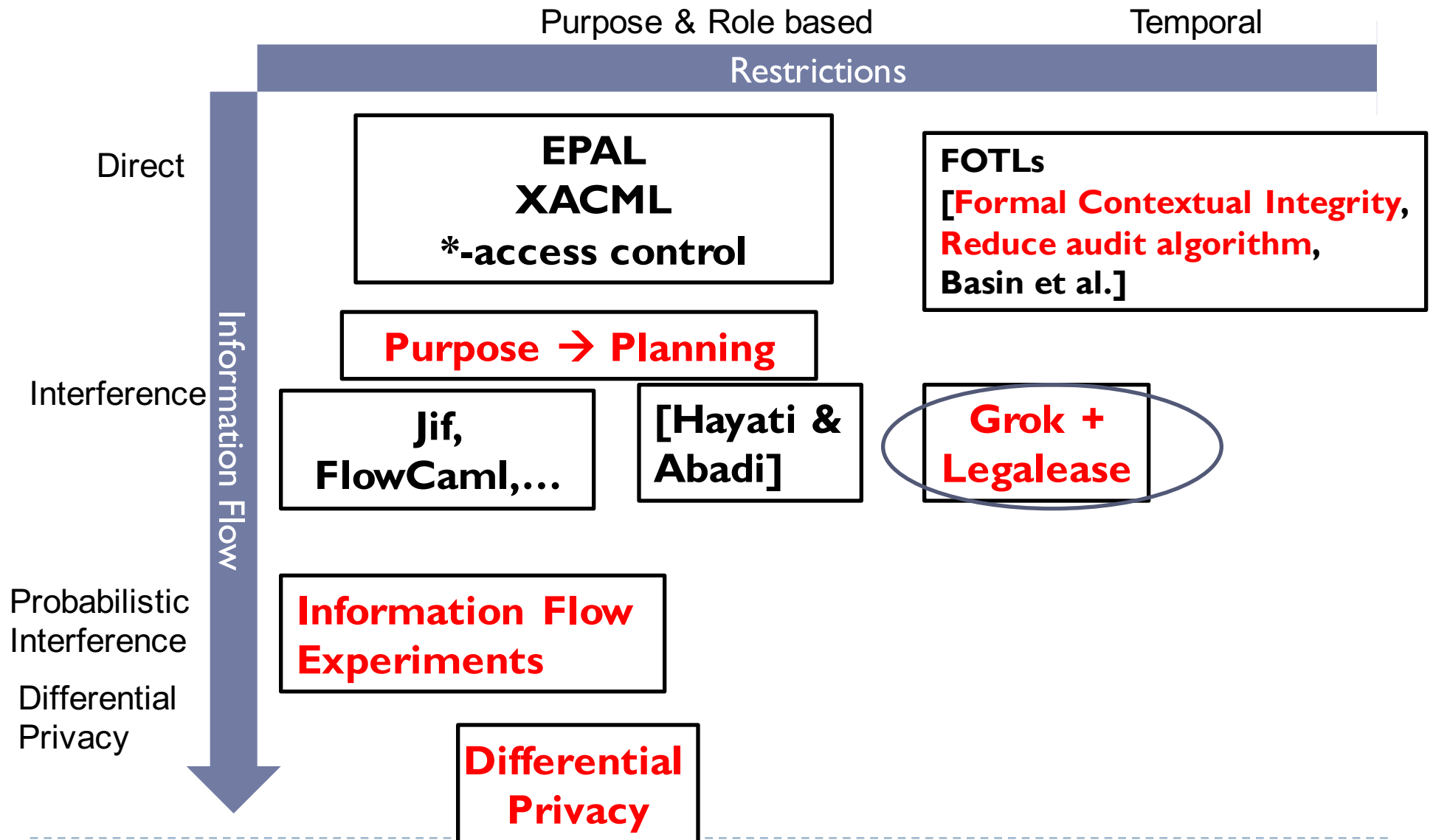
Workflow for privacy compliance

Legalease, usable yet formal policy specification language

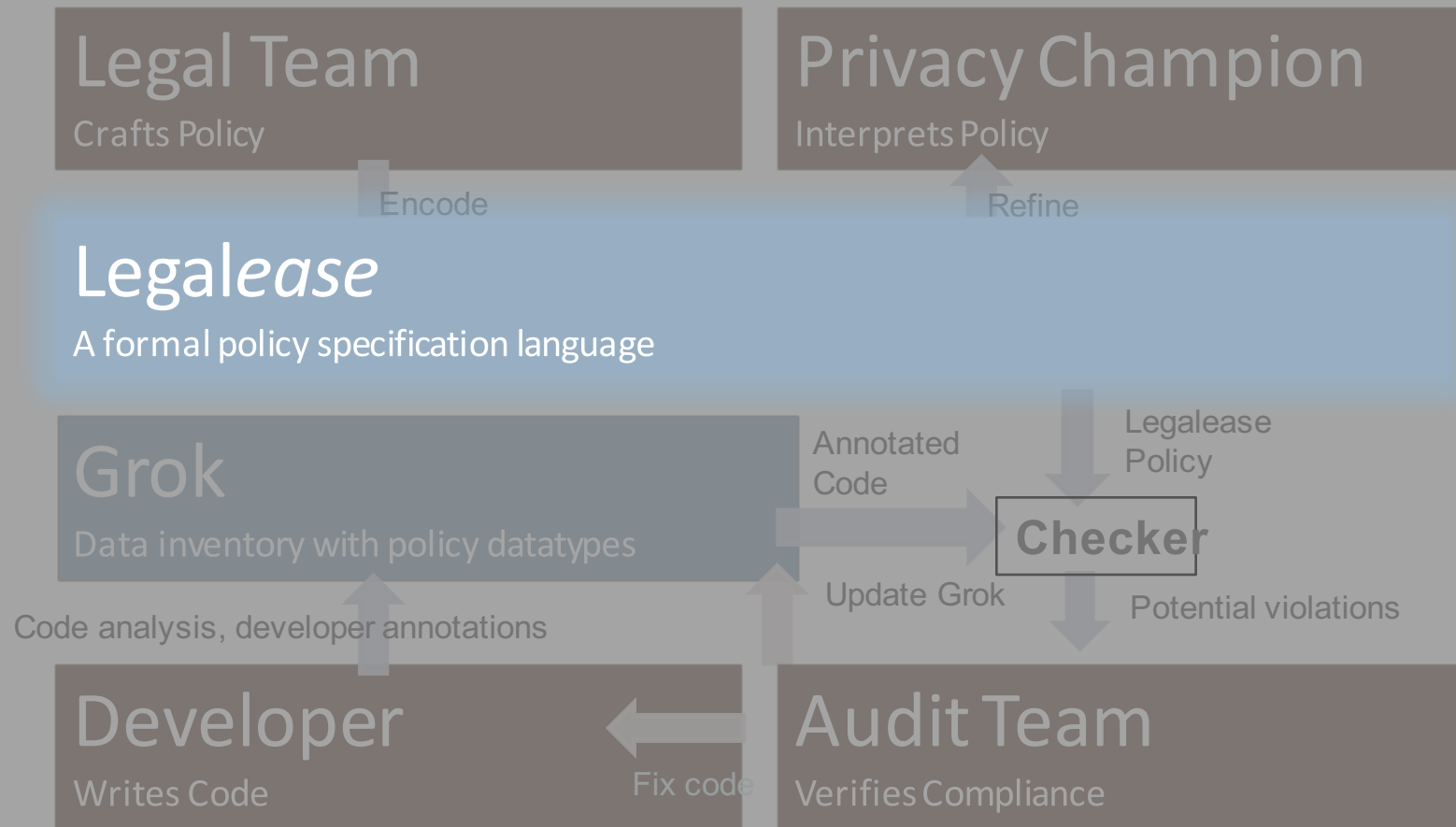
Grok, bootstrapped data inventory for big data systems

Scalable implementation for Bing

Privacy as Restrictions on Personal Information Flow



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 : Syntax

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

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, Account

Exceptions

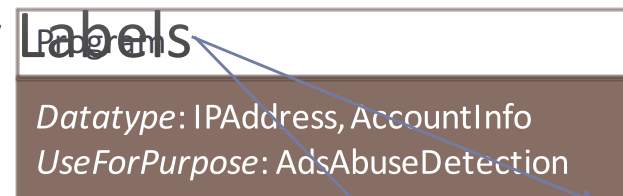
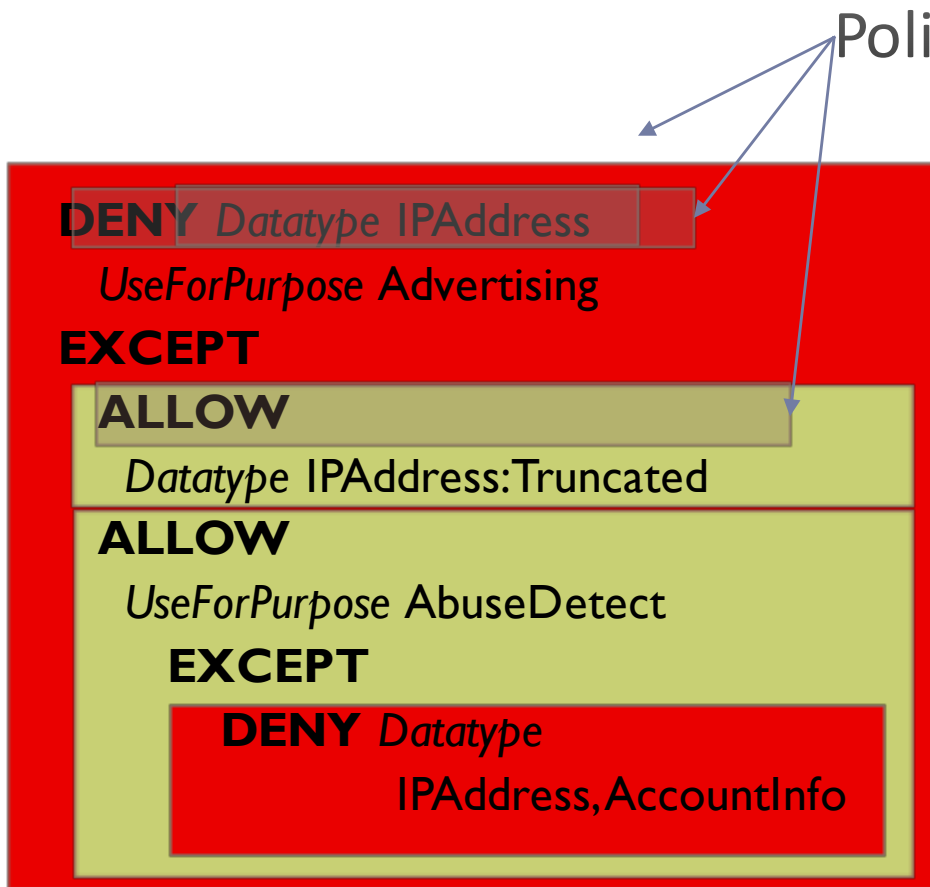
How legal texts are structured
One-to one correspondence

Local Reasoning

Each exception refines its
immediate parent
Formally proven property

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

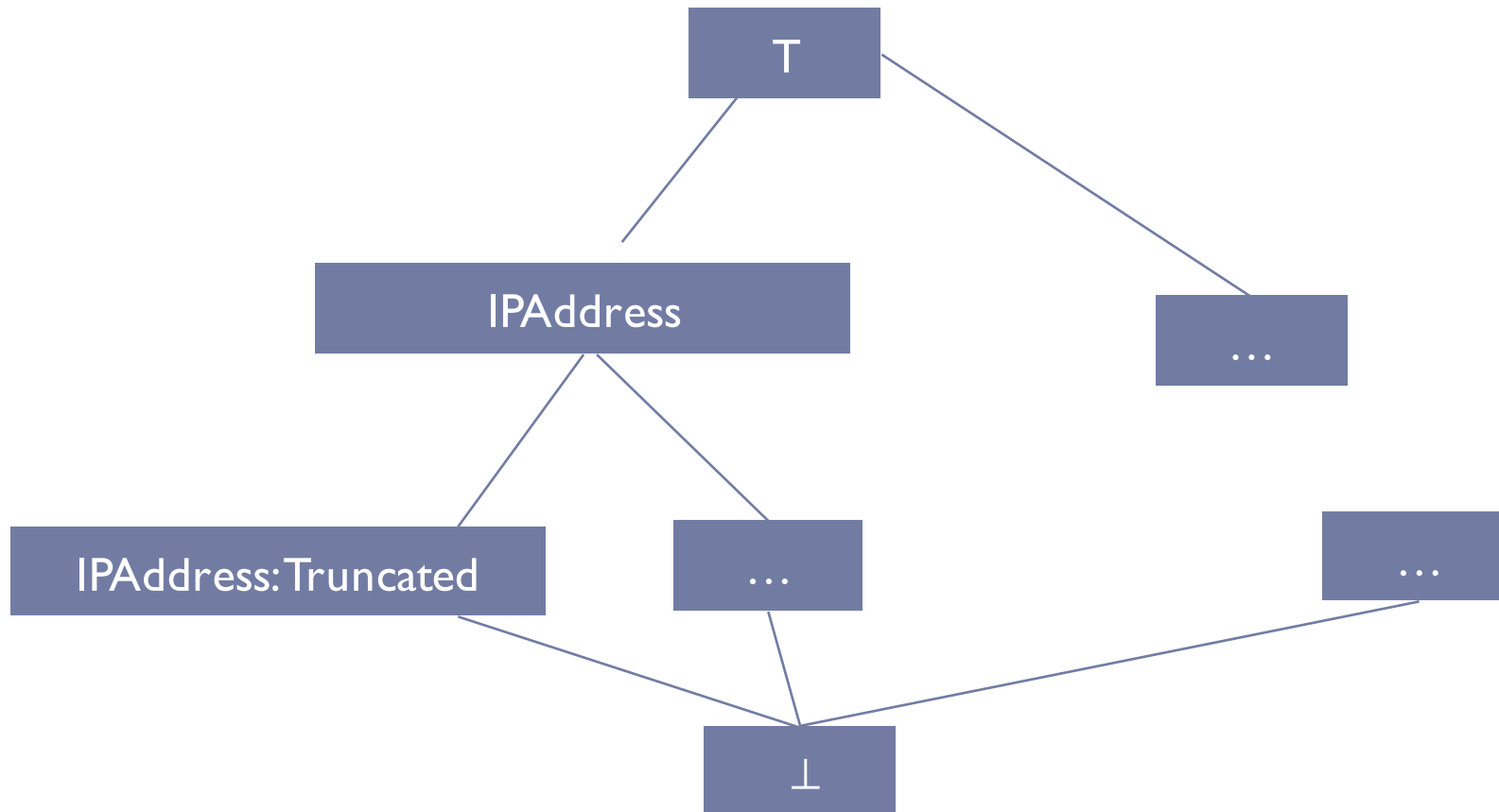
Legalease : In Action



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	::=	$\text{DENY } T_1 \cdots T_n \text{ EXCEPT } A_1 \cdots A_m$ $\text{DENY } T_1 \cdots T_n$
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

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

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

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

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

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

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

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:”

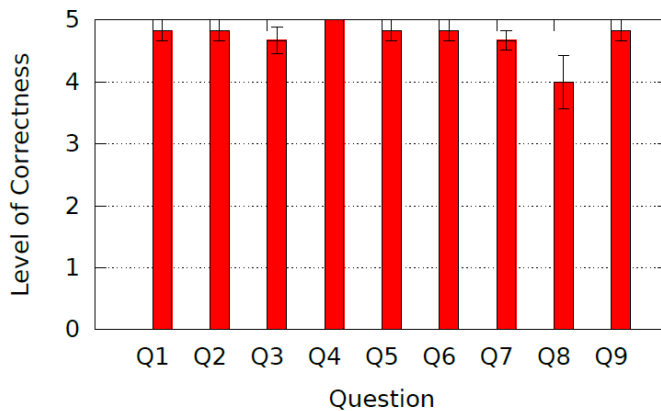
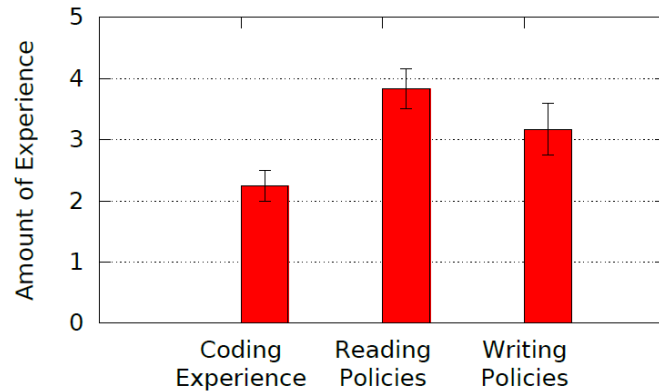
◁ “We require opt-in consent for the sharing of any sensitive personal information.”

◁ “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”

Legalease Usability



Survey taken by 12 policy authors within Microsoft

Encode Bing data usage policy after a brief tutorial

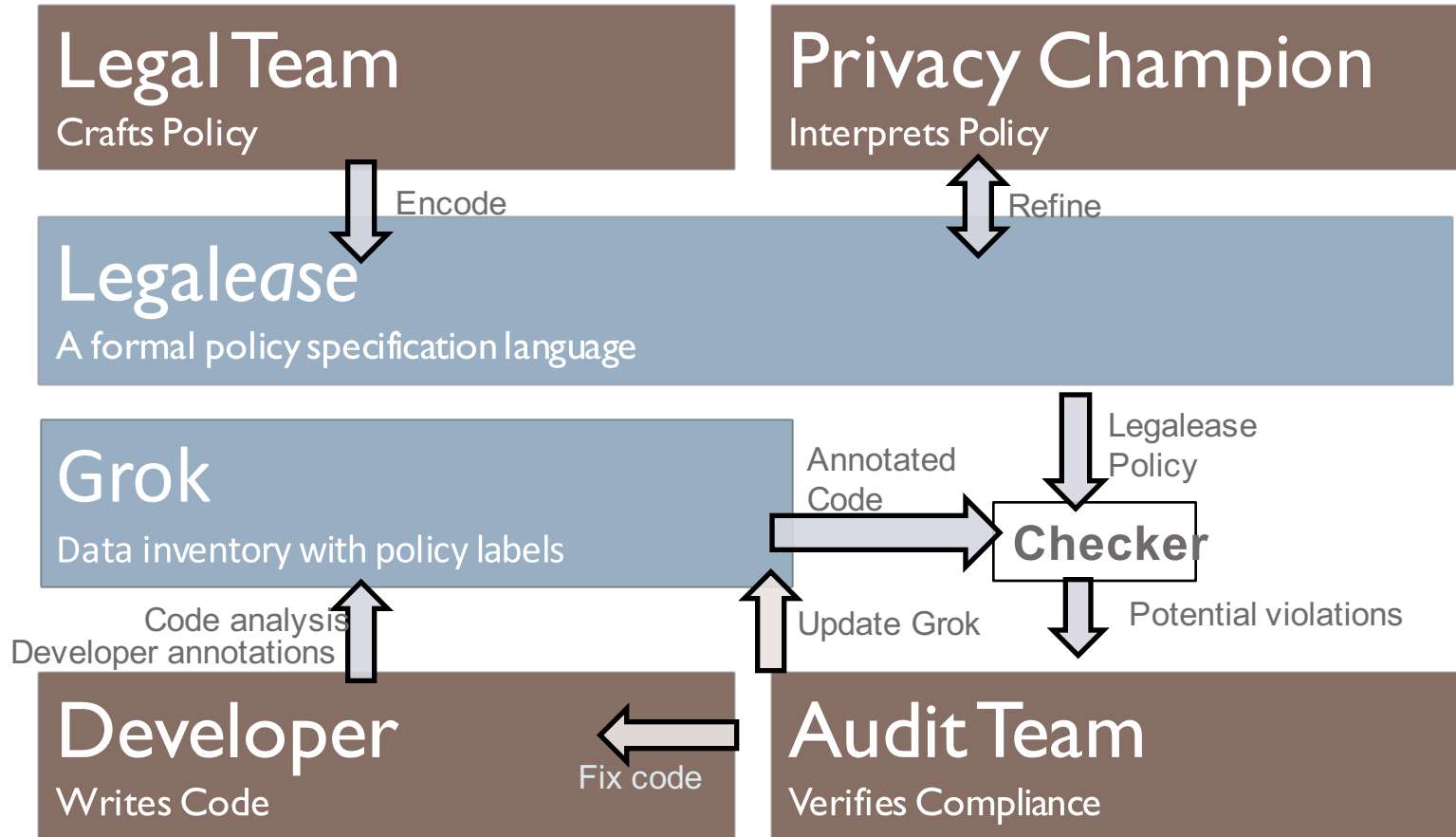
Time spent

2.4 mins on the tutorial

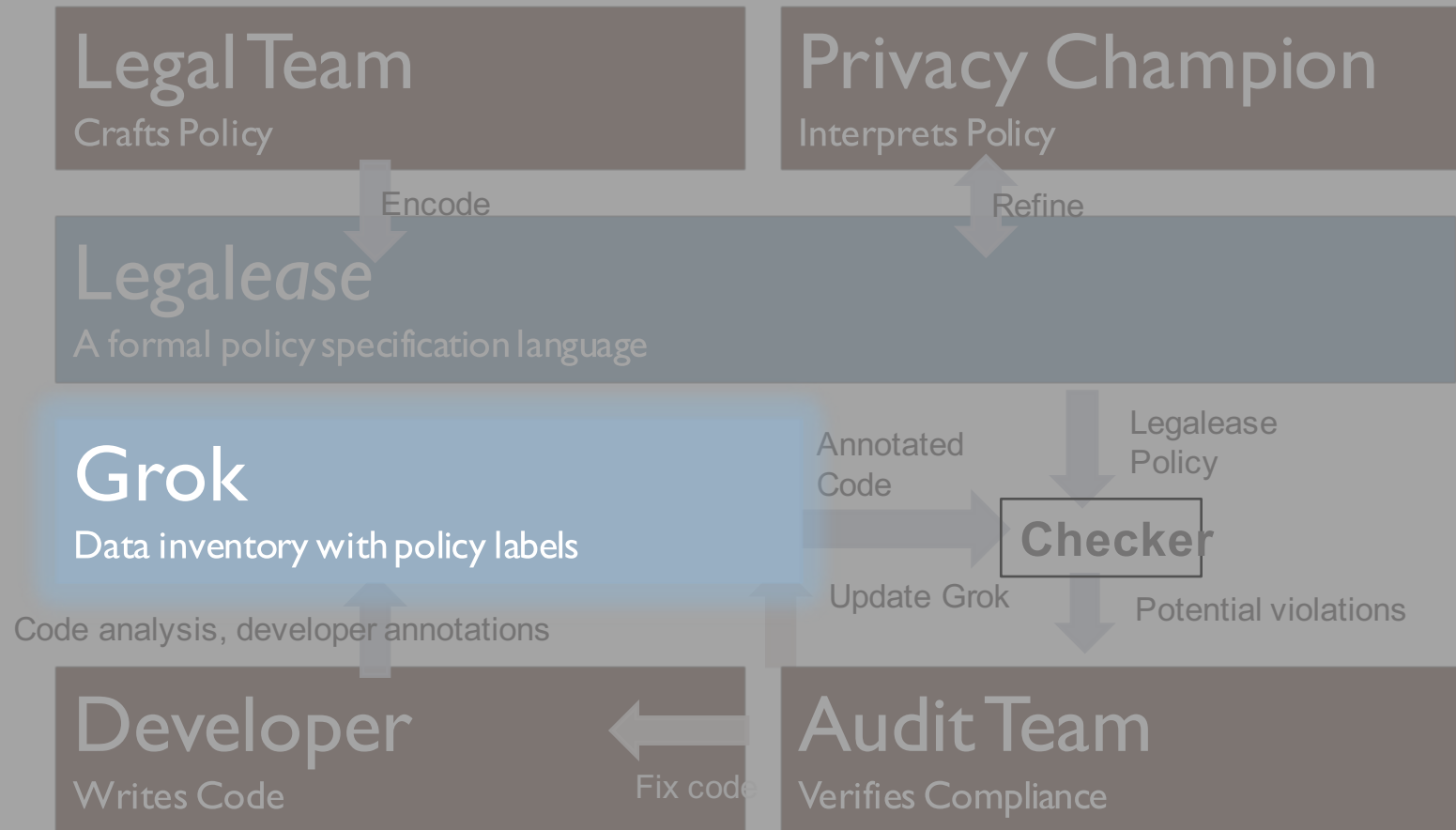
14.3 mins on encoding policy

High overall correctness

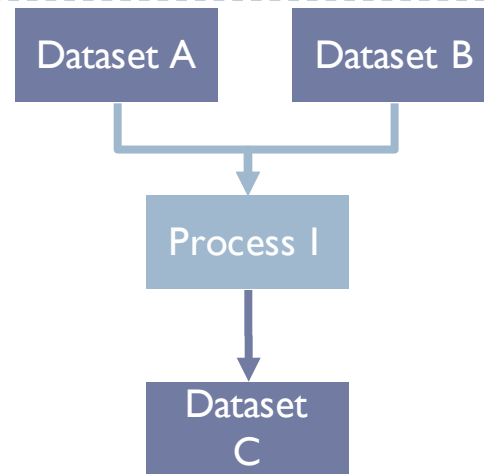
A Streamlined Audit Workflow



A Streamlined Audit Workflow



Map-Reduce Programming Systems



Scope, Hive, Dremel

Data in the form of Tables

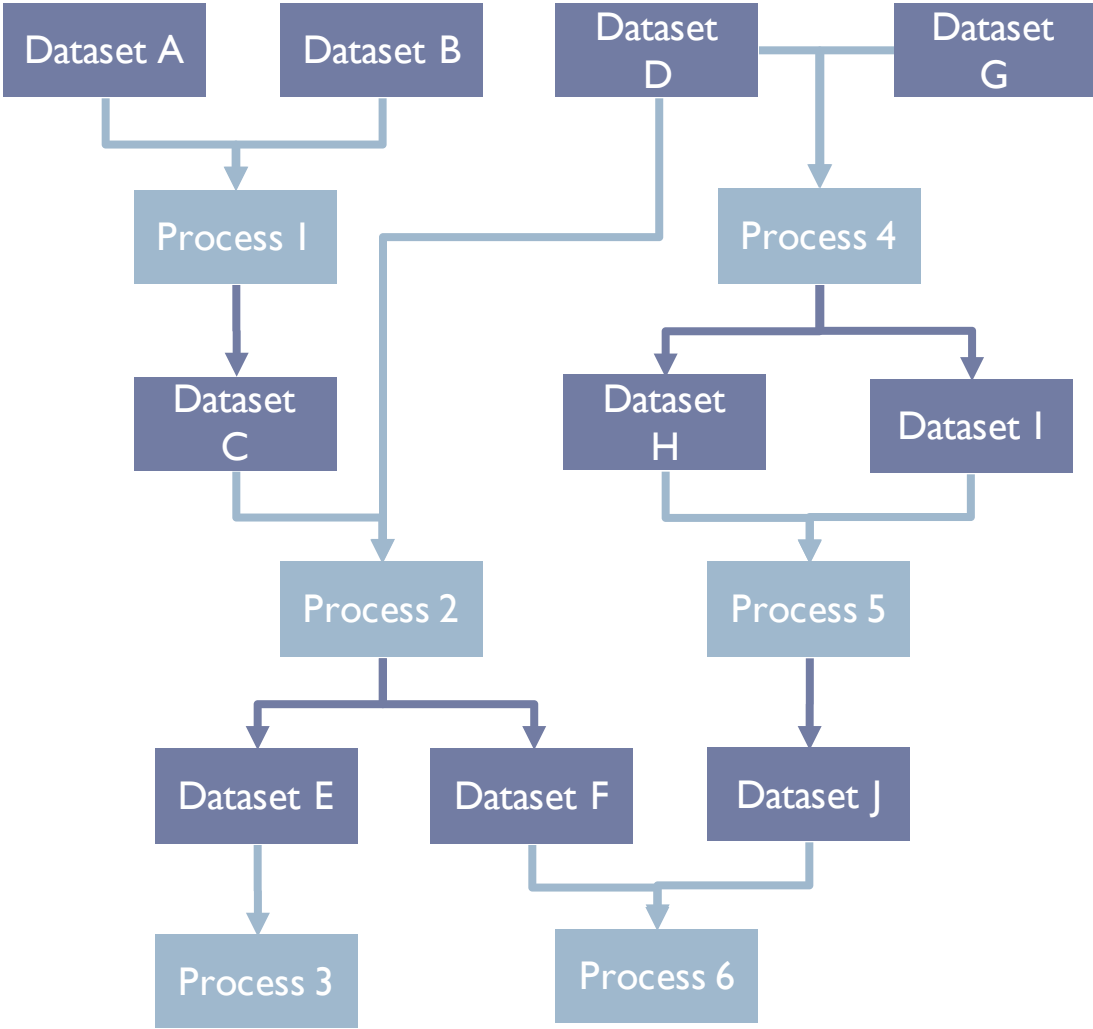
Code Transforms Columns to Columns

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
```

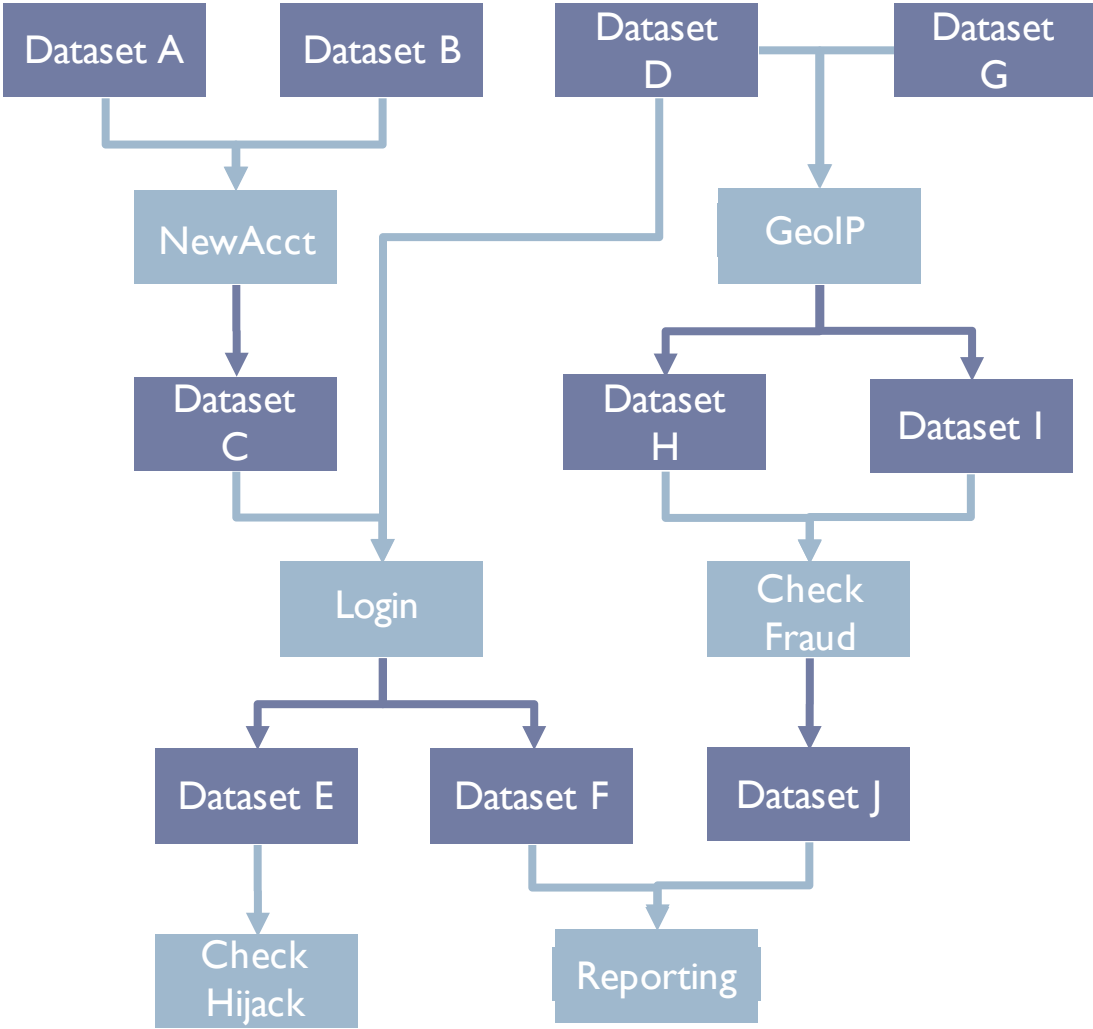
Grok



Grok

Purpose Labels

Annotate programs with purpose labels



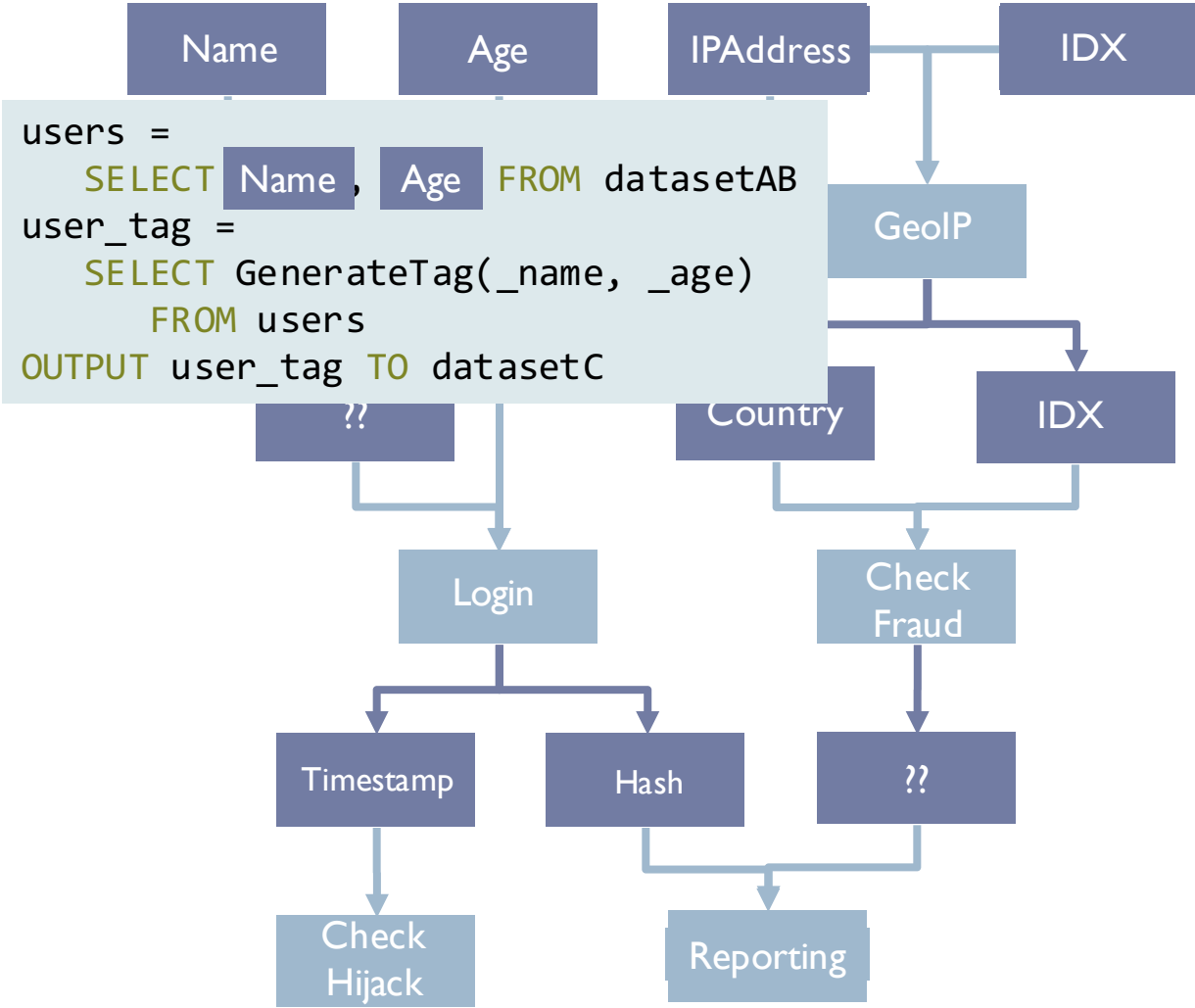
Grok

Purpose Labels

Annotate programs with purpose labels

Initial Data Labels

Heuristics and Annotations



Grok

Purpose Labels

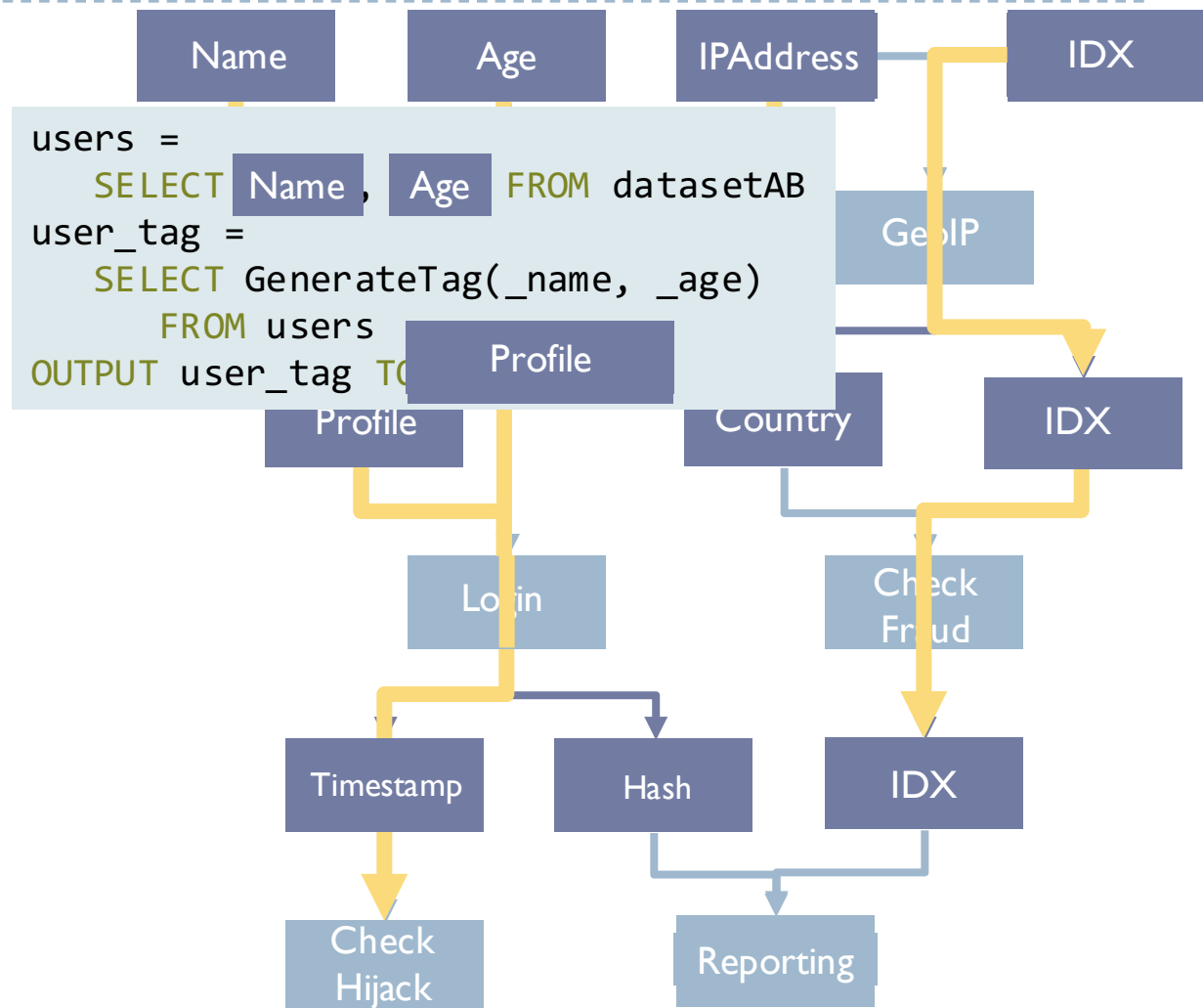
Annotate programs with purpose labels

Initial Data Labels

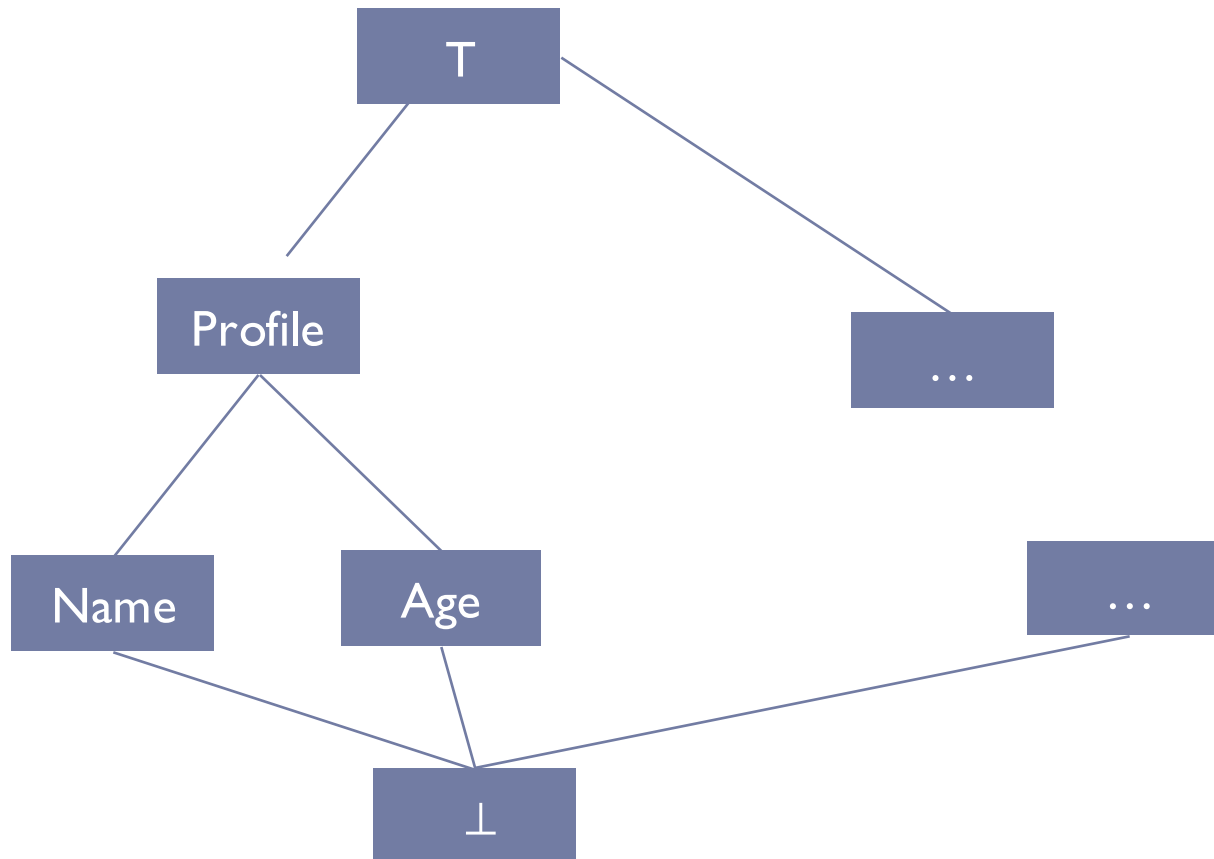
Heuristics and Annotations

Flow Labels

Source labels propagated via data flow graph



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

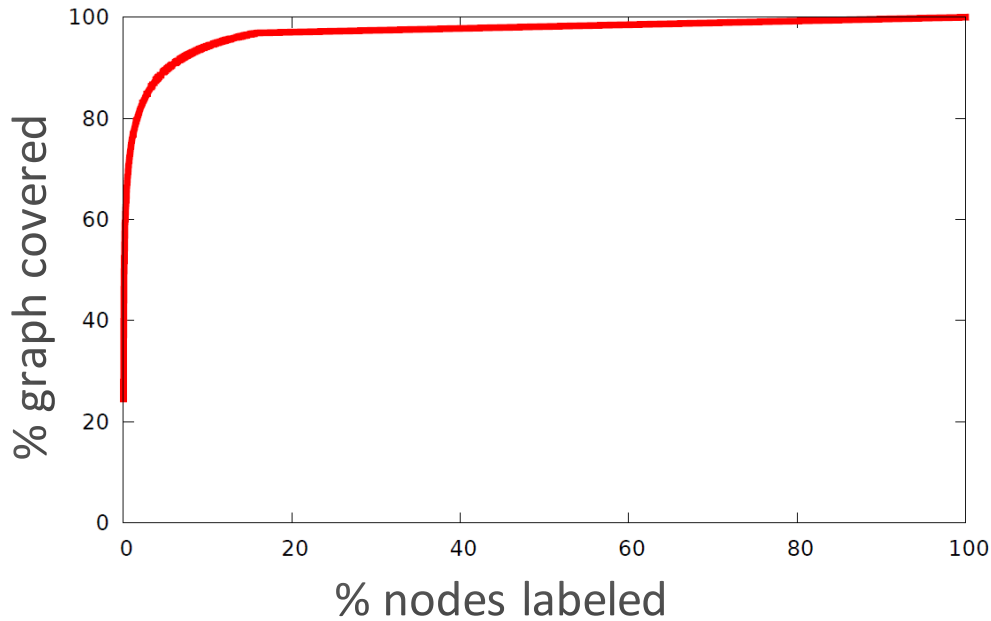
Very Expensive

Definitive

Need very few of these

Auditor
Verification

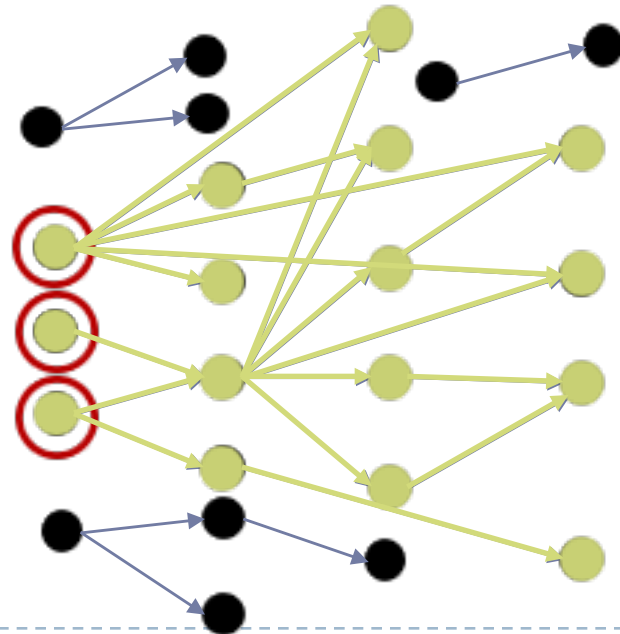
Why Bootstrapping Grok Works



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

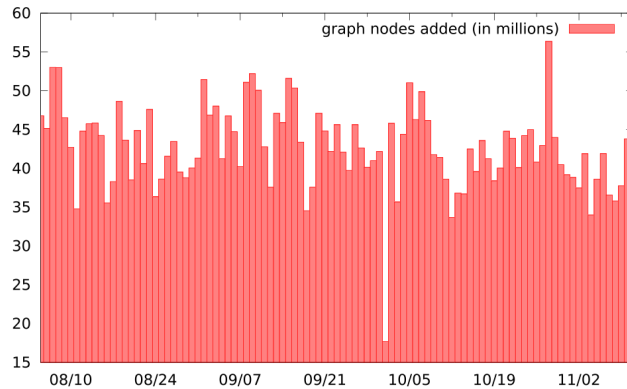


Fig. 9. Number of GROK data flow graph nodes added each day

- ▶ **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**

Nightly Compliance Process

The screenshot displays a SQL query window with a table of PII data and an email thread. The table has columns for Confidence, TaxonomyGroup, Taxonomy, and FieldName. The email thread shows a conversation about privacy management contacts.

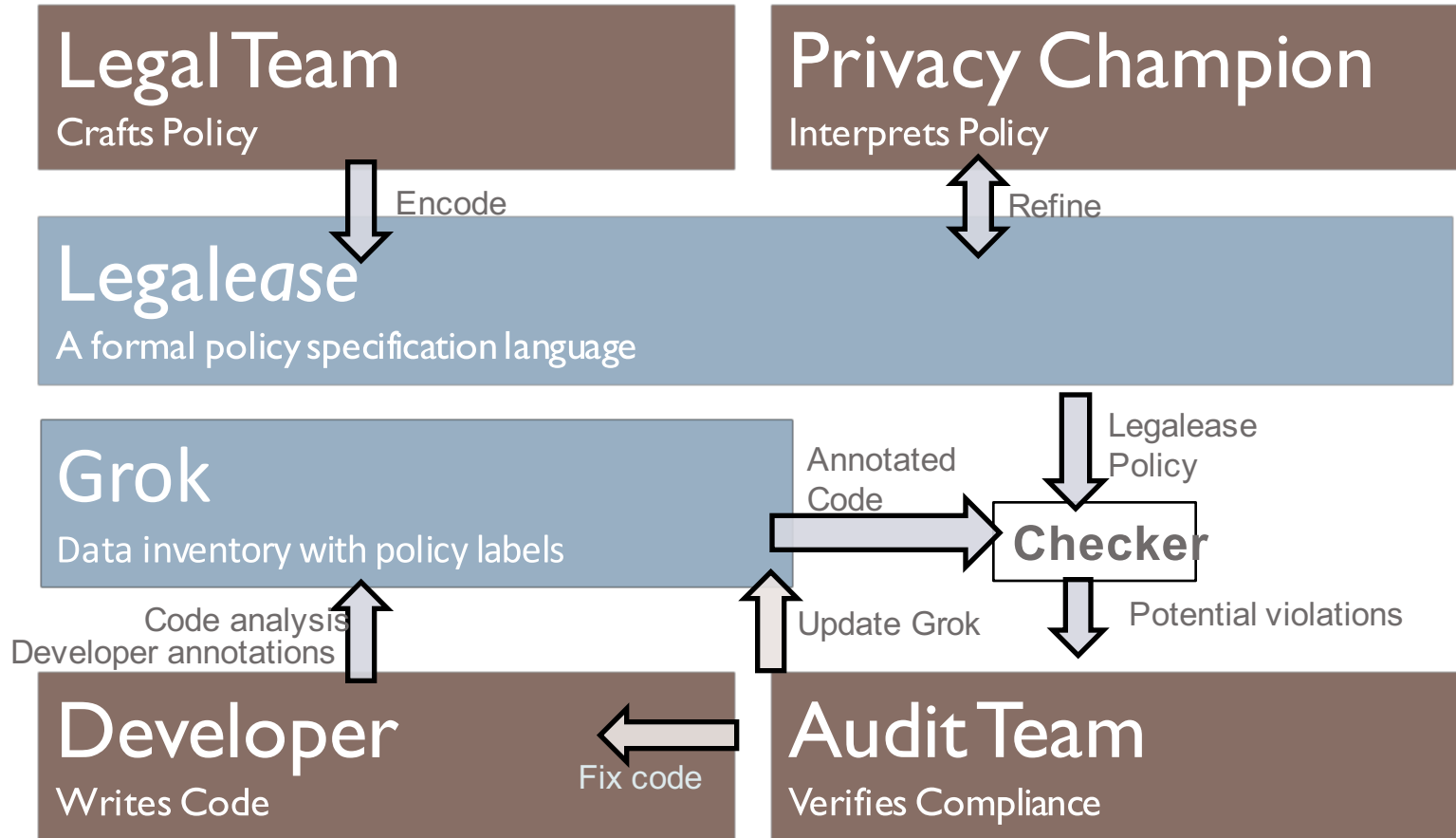
Confidence	TaxonomyGroup	Taxonomy	FieldName
HIGH	PII	Email	LivIdEmailAddress
HIGH	PII	Phone Number	PhoneNumber
HIGH	PII	Email	LivIdEmailAddress
HIGH	PII	Phone Number	PhoneNumber
HIGH	PII	PUID	Puid
HIGH	PII	PUID	UserPuid
HIGH	PII	Email	LivIdEmailAddress
HIGH	PII	Email	PreferredEmail
HIGH	PII	Email	User_LivIdEmailAddress

Static code analysis
 25M+ files

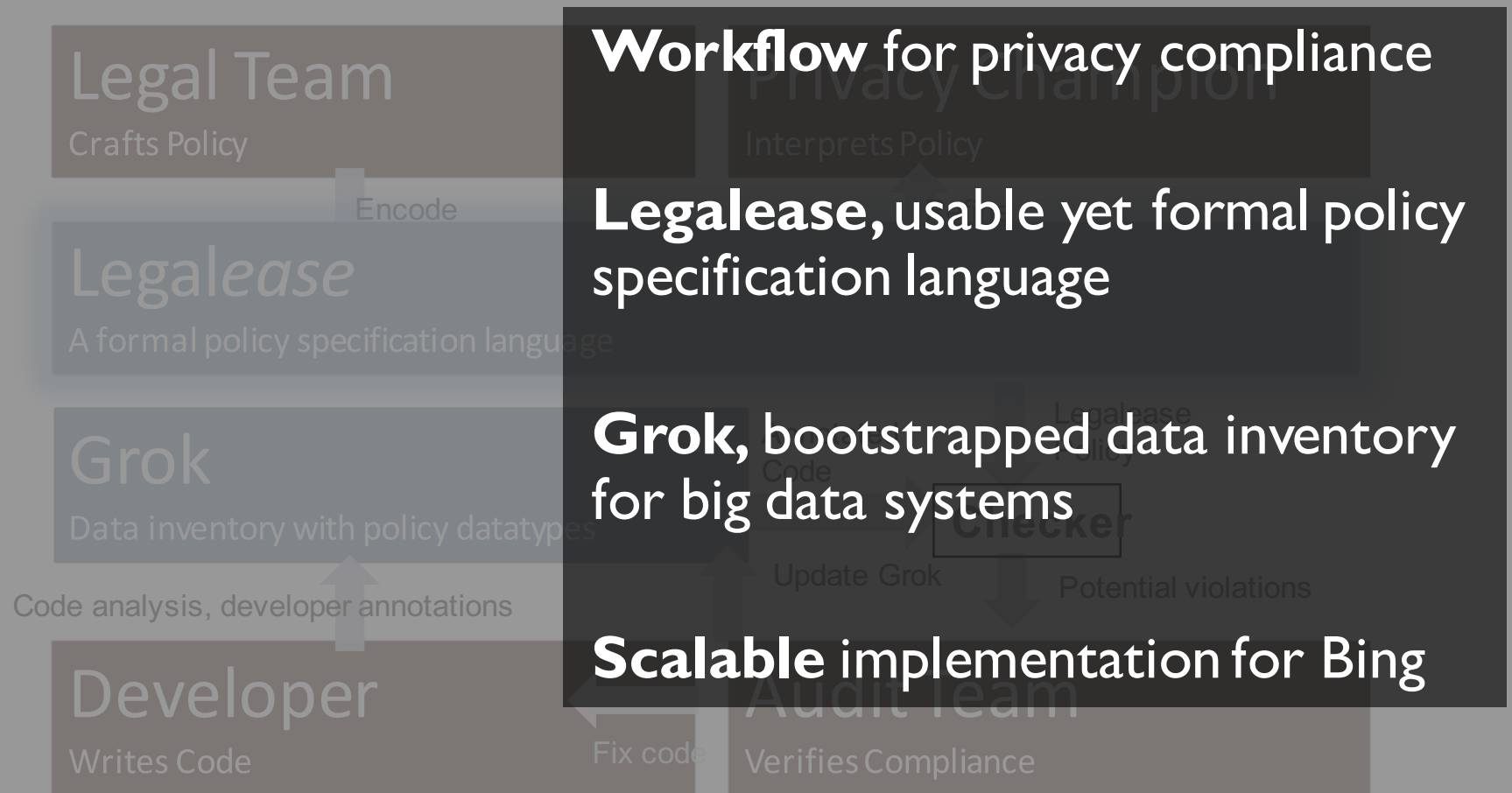
Generate report
 300K+ records

Manual Audit
 8 teams

A Streamlined Audit Workflow



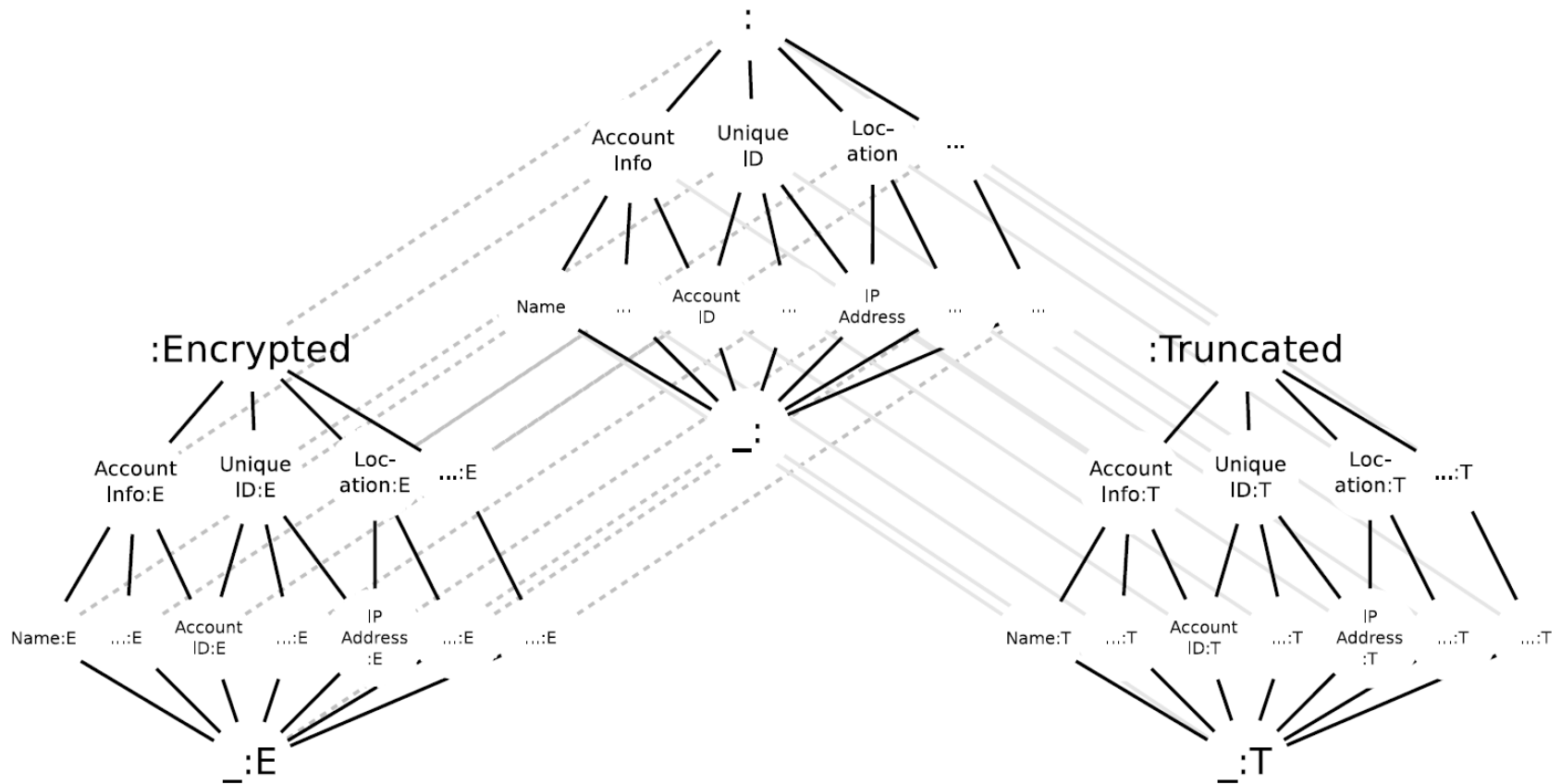
A Streamlined Audit Workflow



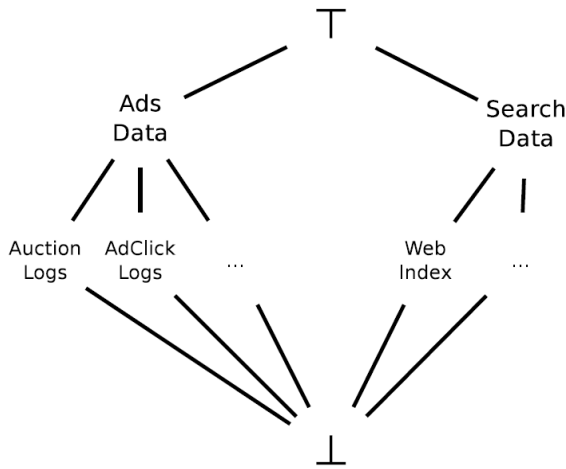
Reference

- ▶ 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.

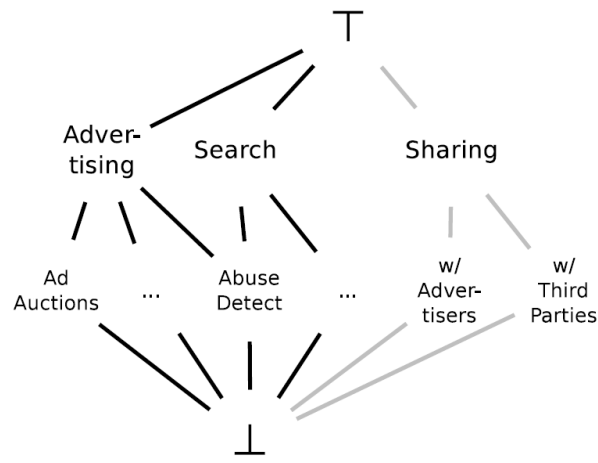
Policy Labels : Datatypes



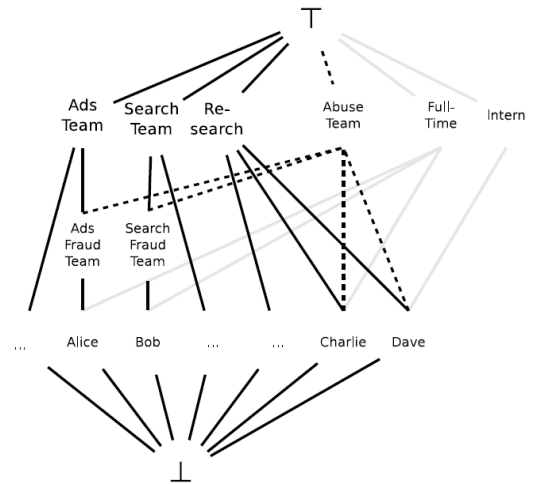
Policy Types : Concept Lattices



InStore Lattice



UseForPurpose Lattice



AccessByRole Lattice

Formal Semantics

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

Based on Lattice Orderings on Policy Types

Formal Semantics

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

Recursively check exceptions

ALLOW clauses have DENY clauses as exceptions

Top Level clause determines Blacklist/Whitelist