Privacy Architecture for Data-Driven Innovation

Nishant Bhajaria

Outline

- Introduction
- Privacy in our world
- Privacy architecture for data collection and usage
- Privacy architecture for data sharing
- Lessons
- Questions

Modern companies

- Collect a lot of user data
- Don't always know how to measure risk
- Struggle to protect data preemptively
- .Cannot make informed decisions around data sharing

Customer Trust Sentiment

- 69% believe companies vulnerable to hacks
- 90% feel they lack complete control over their personal information
- 25% believe most companies handle sensitive personal data responsibly
- 15% think companies will use that data to improve their lives.

Citation: PWC

So what does this mean?

- Privacy is “all hands on deck” not just legal
- Security ≠ Privacy
    - Security is necessary but not sufficient for privacy
- Think beyond breaches
    - Data collection and Internal misuse
    - Data sharing and External misuse

Part 1
A privacy architecture for data collection

Privacy by Data and Design

- Classify your data (Planning)
- Set Governance Standards (Planning)
- Inventory your data (Execution)
- Enforce Data Privacy (Execution)

1. Classify Your Data (planning)

- Answers questions
  - “What is this data?”
  - “How sensitive is this data?”

- Tiered ranking of user and business data

Data Classification Examples

Data Classification	Example Category	Example Data Sets
Tier 1: Highly Restricted	Government Identifiers and location data (excludes personal data)	Social Security Card Driver's License
Tier 2: Restricted	Vehicle Data	License Plate Number Proof of Insurance
Tier 3: Confidential	Non-Identifying Vehicle Data	Make and Model Color
Tier 4: Public	Public Information	Press Releases Product Brochures

2. Set Governance Standards (planning)

Data Handling Requirements

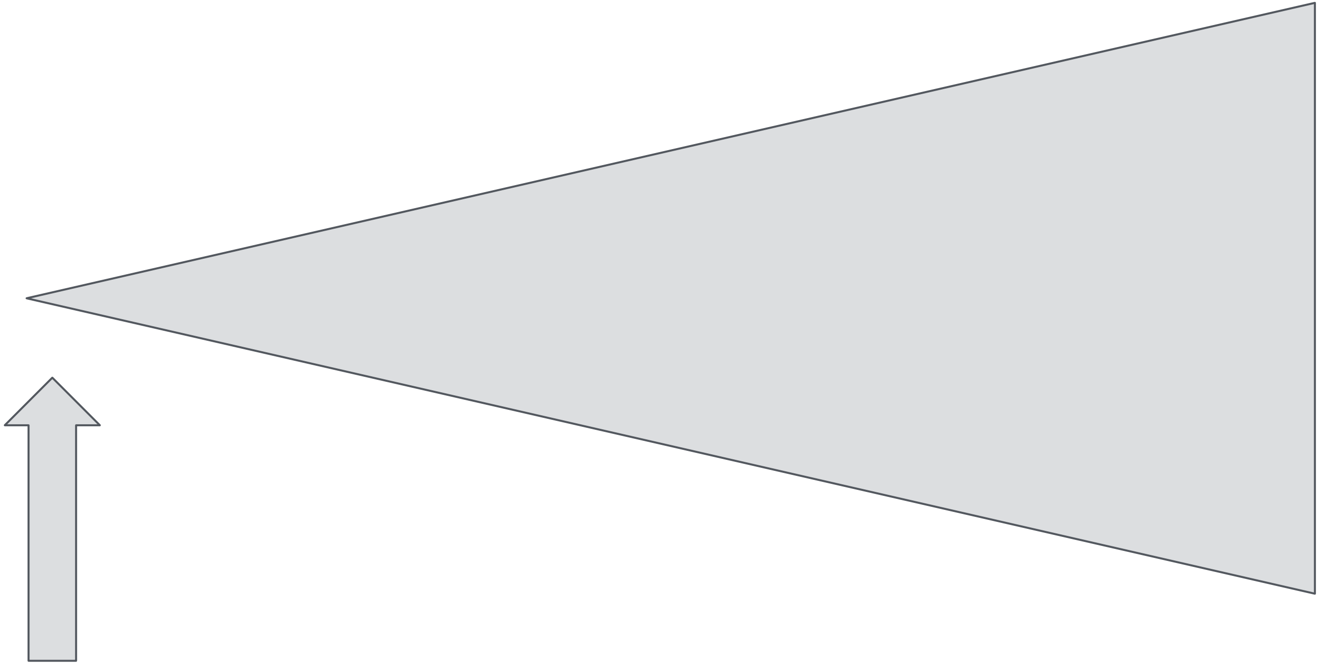
“How can I protect this data?”

Collection

Access

Retention, Deletion, Sharing (internal/external)

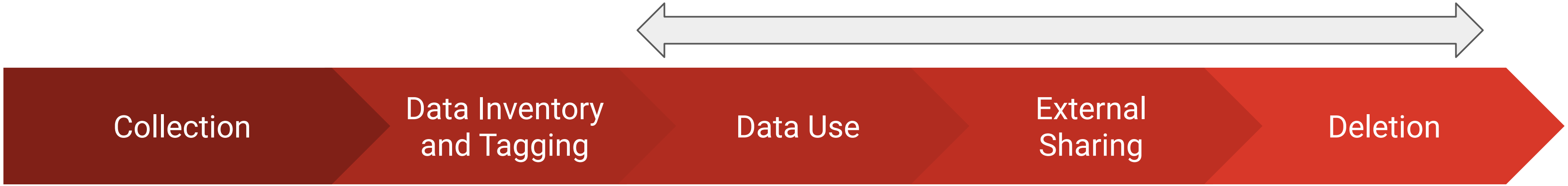
3. Data Inventory (Execution)



Classify and
inventory your data

Why is Data Inventory vital?

Cannot apply data protection post collection without inventory



Collection
- User Apps
- Web Site
- Third-Parties

External Sharing
- User Apps
- Export/DSAR
- Third Party Sharing

Deletion
- Retention Policy

Why data inventory is hard

Data Inventory at Uber

We needed a combined system infrastructure that could

- Crawl various datastores,
- Discover datasets,
- Make those datasets and corresponding metadata available.
- Provide extensibility to add new metadata in self-service fashion.
- Support the categorization of personal data (privacy use case)

How UMS fits into the larger data inventory strategy

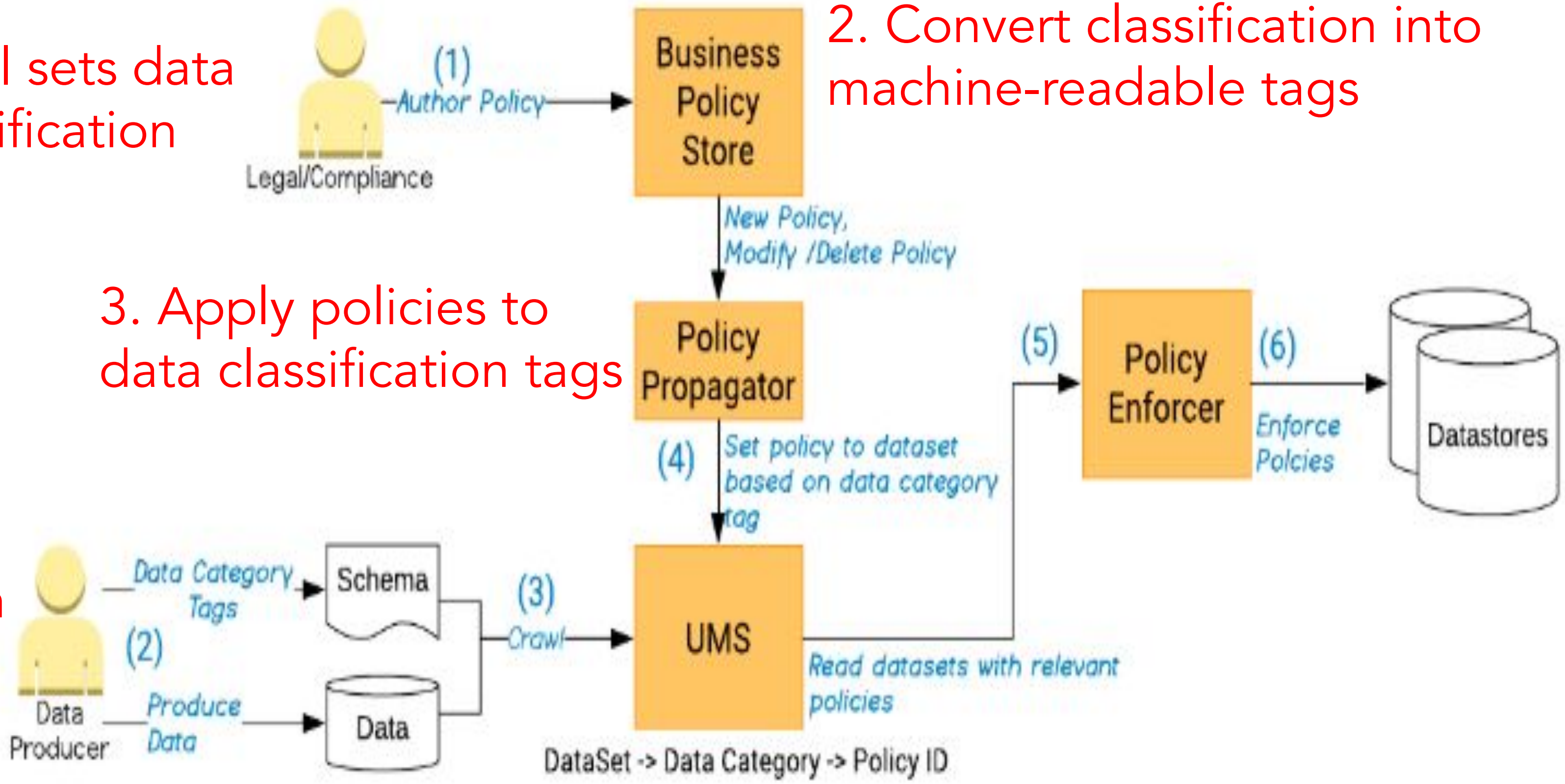
UMS is Uber's Metadata Management Service

1. Legal sets data classification

2. Convert classification into machine-readable tags

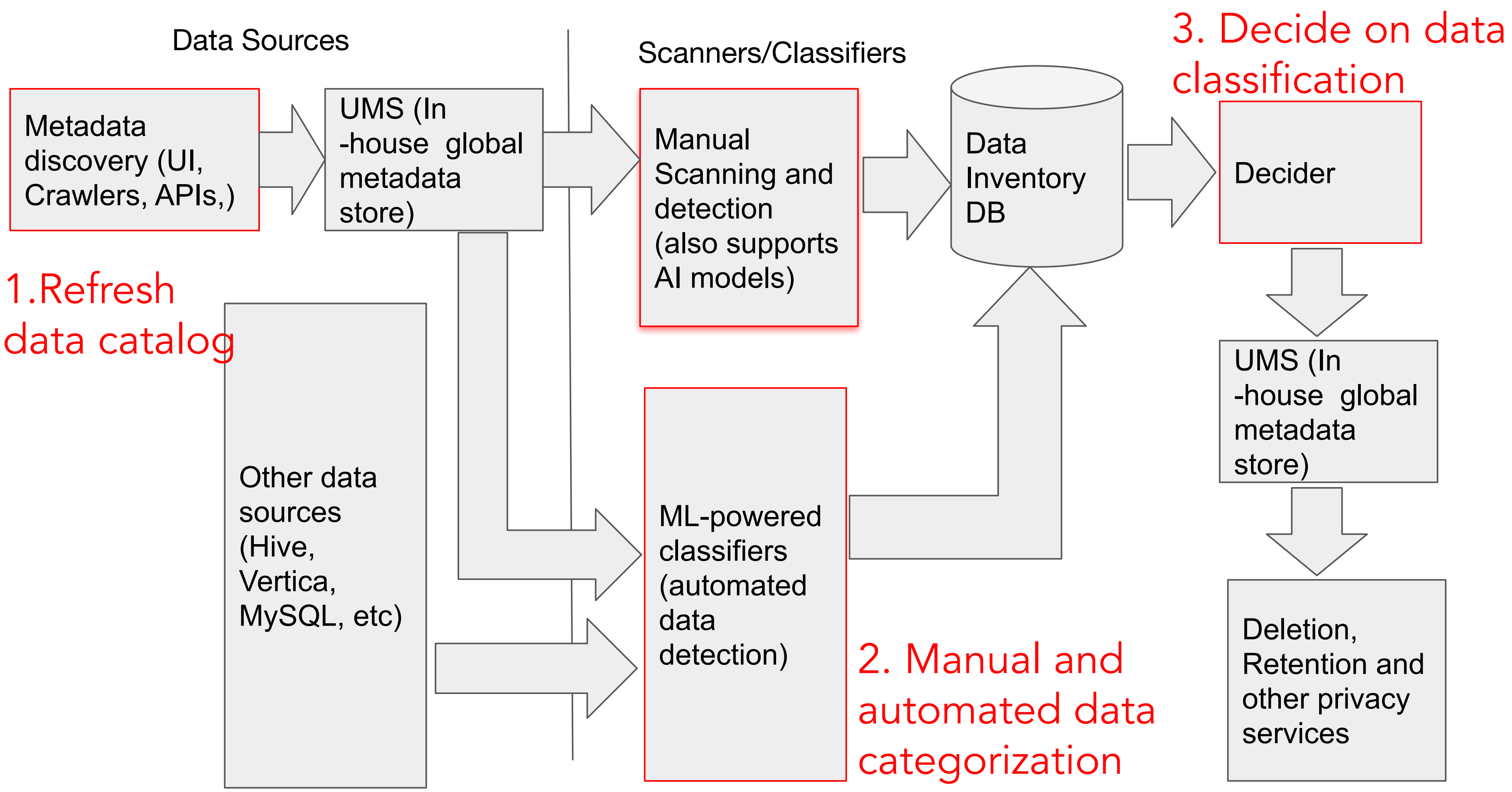
3. Apply policies to data classification tags

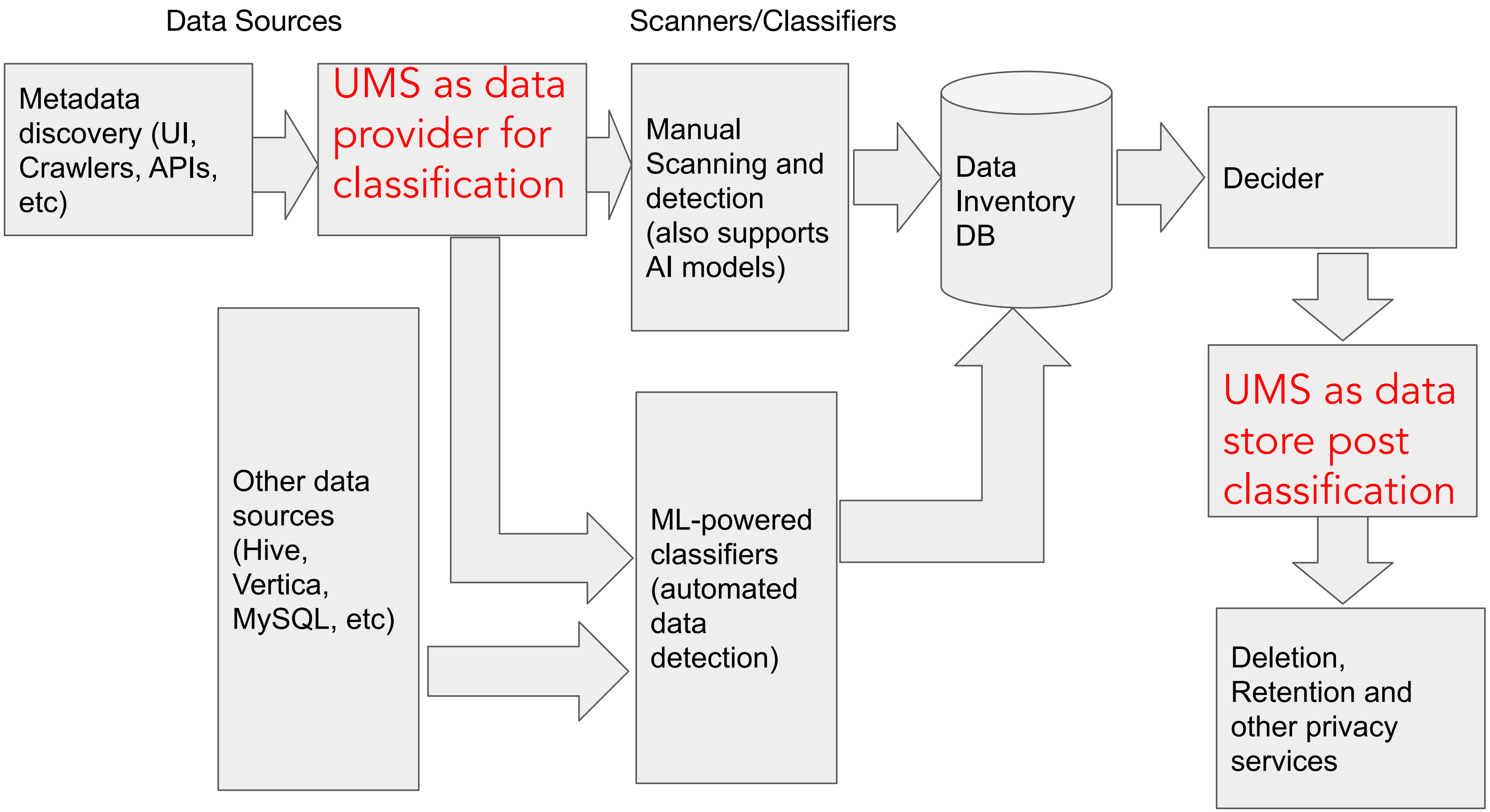
4. Data ingestion stage



The UMS backend

A granular view


Data Sources
Scanners/Classifiers
3. Decide on data classification
Metadata discovery (UI, Crawlers, APIs,)
UMS (In -house global metadata store)
Manual Scanning and detection (also supports AI models)
Data Inventory DB
Decider
1.Refresh data catalog
Other data sources (Hive, Vertica, MySQL, etc)
ML-powered classifiers (automated data detection)
UMS (In -house global metadata store)
Deletion, Retention and other privacy services
2. Manual and automated data categorization

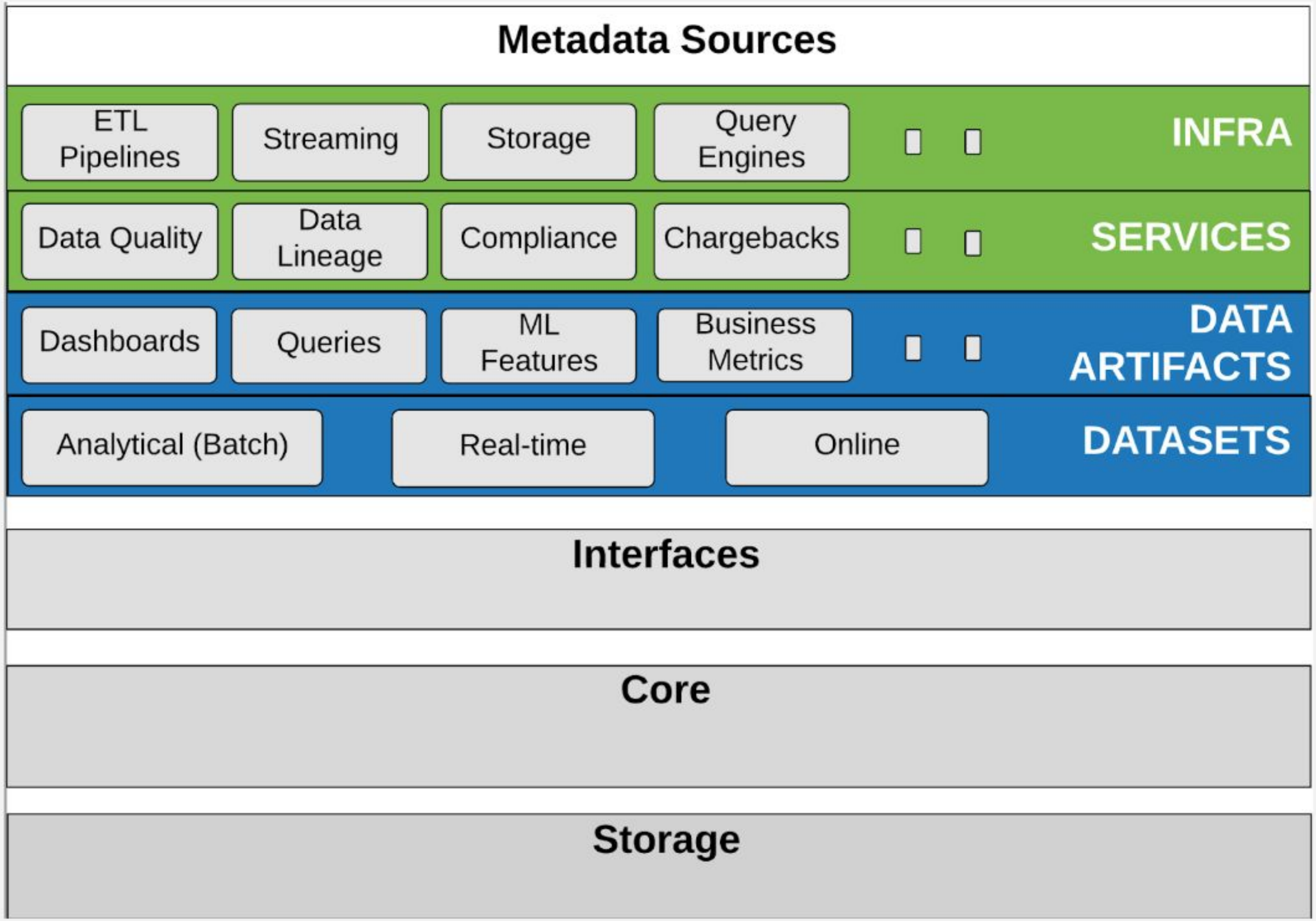

Data Sources
Scanners/Classifiers
Metadata discovery (UI, Crawlers, APIs, etc)
UMS as data provider for classification
Manual Scanning and detection (also supports AI models)
Data Inventory DB
Decider
UMS as data store post classification
Other data sources (Hive, Vertica, MySQL, etc)
ML-powered classifiers (automated data detection)
Deletion, Retention and other privacy services

The UMS is “Privacy Central”

Data Inventory back-end infrastructure

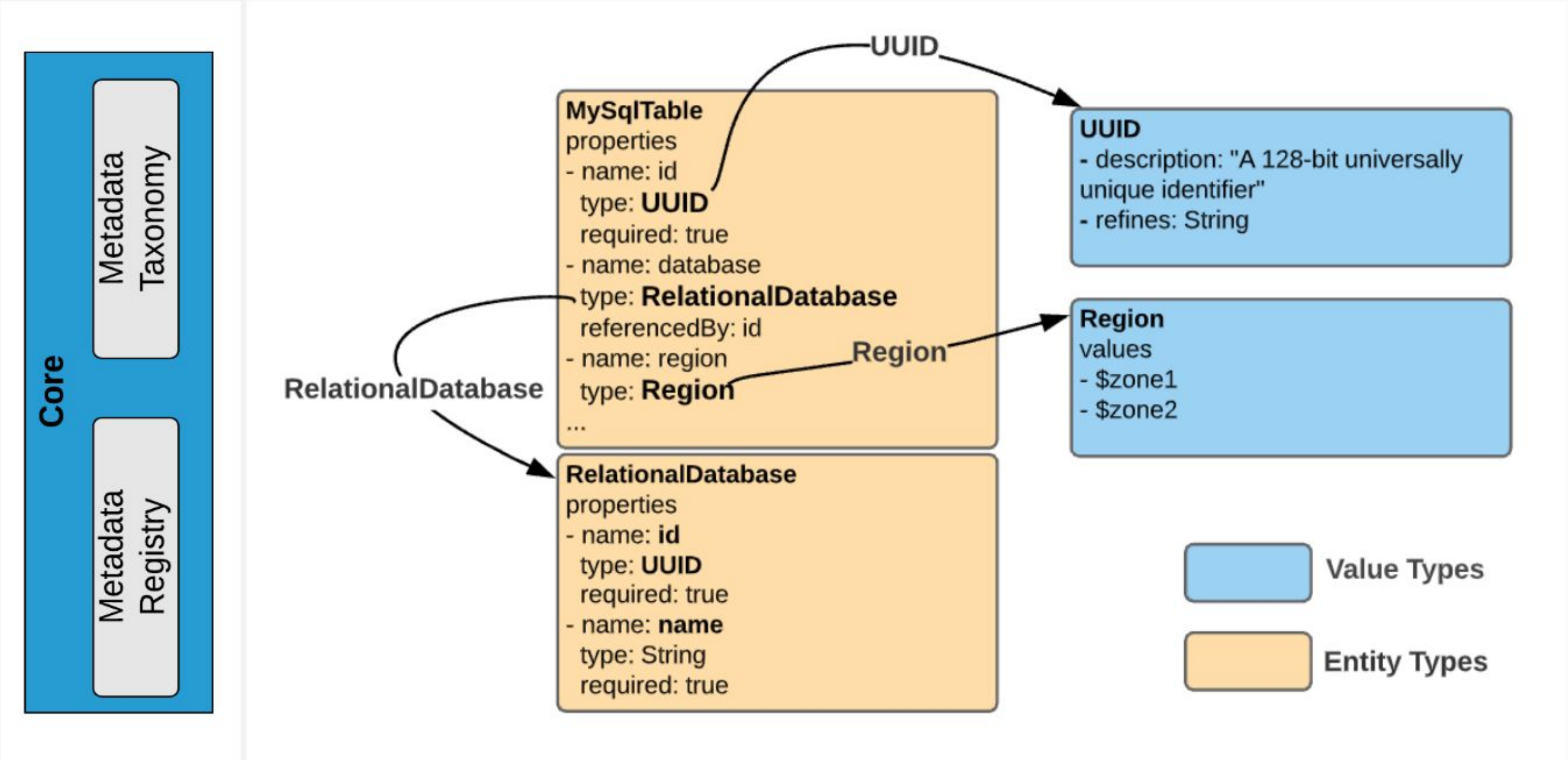
Metadata Sources

UMS

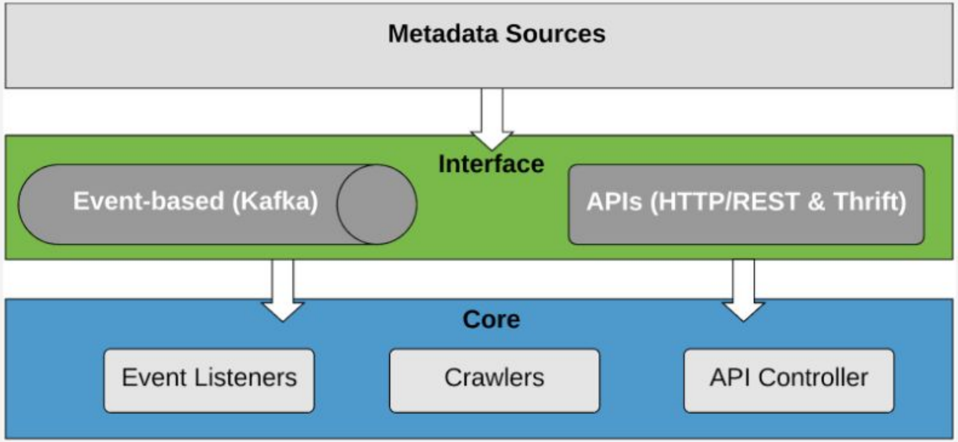


A consistent Metadata definition for Data Inventory

Metadata Registry/Definition



Metadata Collection



Pull model	Push model
○ **Crawler** (periodic) ○ **Event-based** (Event Listeners)	○ **Automated** ○ **Manual entry**

Classification techniques

Categorization method	Coverage	Accuracy	Performance
ML method #1	Very High	Medium	Very High
ML method #2	High	High	Very High
ML method #3	Medium	High	Very High

Data Inventory high level milestone

Data Source	Results	Granularity
Databases (Structured data)	Data volume (TB/PB), % of columns (by risk level)	Storage instance (Eg: Hive instance)
AWS S3 bucket	Data volume (TB/PB), % of objects (by risk level)	Bucket
3rd party SaaS Apps	Data volume (TB/PB), % of objects (by risk level)	Application instance (Eg: Drive instance)

The Privacy Challenge

Could your security infrastructure keep up with data growth?

Concerns/Learnings

- Data quantity
- Inflection point
- Rate of collection <= Rate of deletion?
- Data Quality

Part 2
A privacy architecture for data sharing

3rd party data sharing checklist

- Will the data be secure (at rest and in transit)?
- How granular must shared data be?
    - Location precision
    - Aggregation and anonymization
- Will 3rd party monetize the data?

Use cases for data sharing with cities

- Impact on traffic, parking, emissions, etc.
- Collecting per-vehicle fees
- Enforcing parking rules for bikes/scooters
- Responding to service and safety issues

Other Data sharing use-cases

- Geolocations
- Trip telemetry
- Vehicle and driver license numbers

Data retention guidelines

- Delete unique IDs, precise times and locations after 90 days
- Delete coarsened times and locations after 2 years
- Internal, infinitely retained data should be at least 5-anonymous
- Bulk shared data should be at least 100-anonymous

The more precise the data, the lower the retention period

Privacy preservation techniques 1 (Uber)

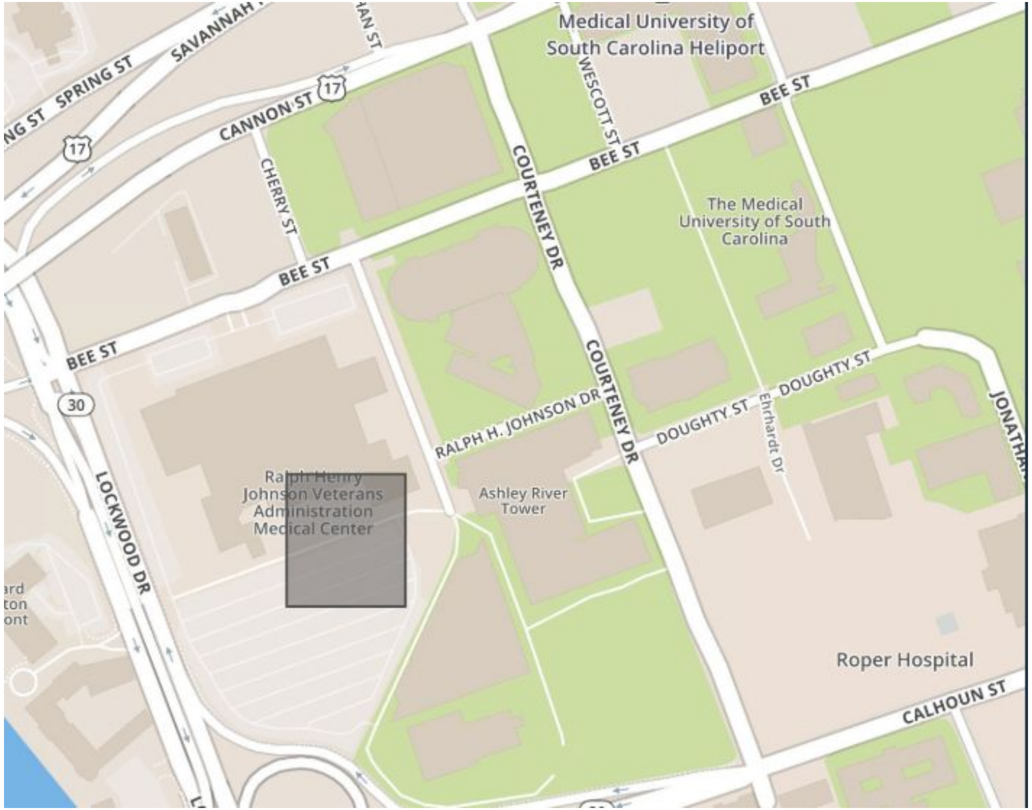
- Remove or replace unique identifiers
- Recommendations:
  - Replace IDs from providers with internal IDs
  - Remove PII or replace w/ keyed pseudorandom function

Privacy preservation techniques 2 (Uber)

Coarsen precision of stored data

- Round times to nearest 30-minute increment
- Convert GPS coordinates to street segment start/center/end
- Truncate GPS coordinates to 3 decimal degrees

Time/Location coarsening has its limits



[Citation]

Privacy preservation techniques 3 (Uber)

- Suppress data that does not meet a minimum k-anonymity

Uber Movement Portal

K-Anonymity

A case study: 40,000 Boston trips

K-Anonymity with 0 decimal points

K-anonymity

GPS rounding

	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

K-Anonymity with 4/5 decimal points

K-anonymity

GPS rounding

	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

5-Anonymity for 0-5 GPS decimal points

K-anonymity

GPS rounding

	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

Privacy preservation techniques 4 (Uber)

- Allow noise infusion as use cases allow
- Recommendations:
  - Publish expected statistical and aggregate queries
  - Publish acceptable error tolerances

Data Sharing

Case study: Minneapolis

Privacy in collection (Minneapolis)

- Trip IDs from MDS
  - Already hashed, still discarded
  - Generated a new unique city trip ID to make identification harder
- Discard trips that did not have start and end points
- Round off start and end times for trips (12:21 == 12:30 == 12:24)

Privacy in processing (Minneapolis)

- Access control for data stores and APIs
- Anonymize data in memory
  - Do not persist data used solely for aggregation
  - Keep individual-level data in memory; only processed data to disk

Location Binning for Anonymization (Minneapolis)

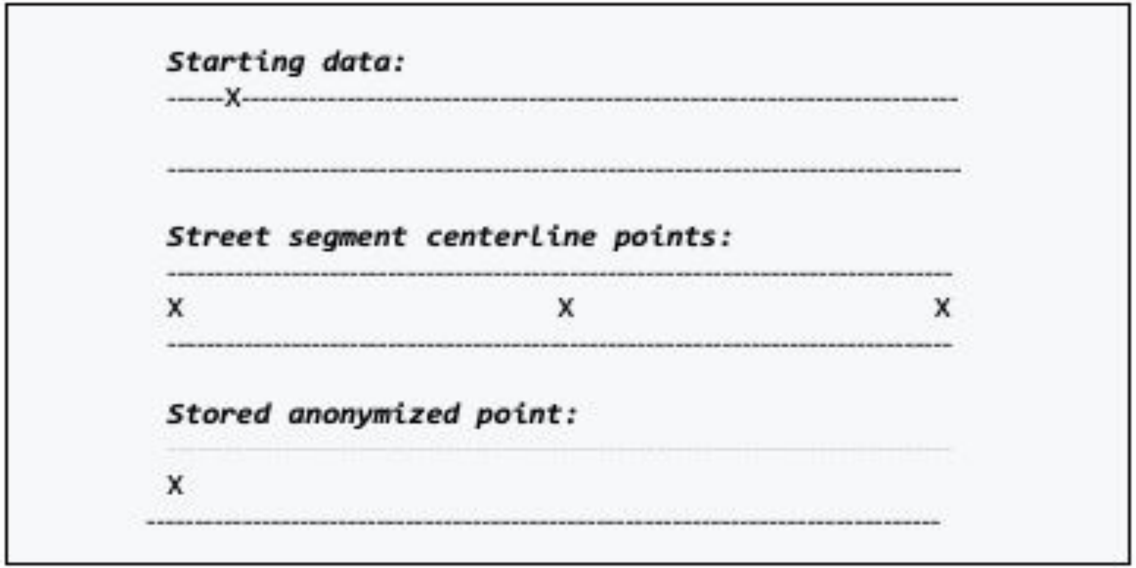


Figure 1: Centerline Anonymization Binning Methodology

Discard trip start and end points for all trips

APPENDIX

Privacy and Precision

White paper: “Unique in the Crowd: The privacy bounds of human mobility”

Citation

The Privacy Challenge

Could your digital fingerprint identify you more than your real fingerprint?

Research TL;DR

- 12 points needed to uniquely identify a fingerprint
- How many points can identify a human on the move?
  - Fewer points required to identify means less privacy
  - Don't forget the power of outside info

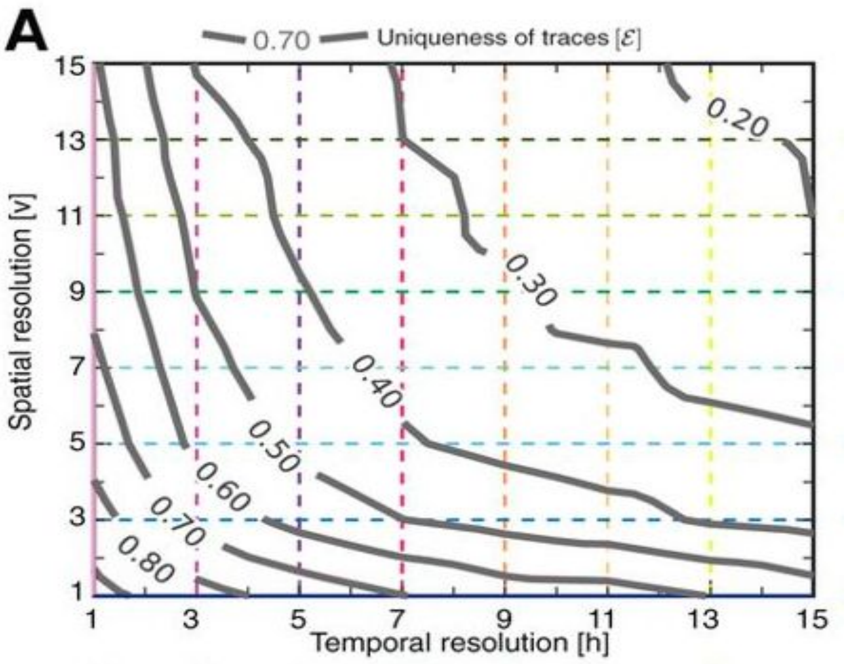
Research TL;DR

- 15 months of human mobility data for 1.5 million users
- Findings
  - 4 spatio-temporal points ID 95% of individuals
  - Coarsening costs more in quality than rewards in privacy

Research TL;DR

- Uniqueness of trace decays at 1/10th power of resolution
- Challenge
  - Even coarse datasets may not provide sufficient anonymity
  - At some point, data may start losing value due to coarseness

Sacrificing time and location for privacy



Lose a pound of precision for a penny of privacy

The challenge of outside info

- Even fully anonymized datasets can pose privacy risks
- Privacy challenge is not just data, but patterns
- Example:
  - Medical DB + Voter list $\Rightarrow$ MA Gov. health record

So, how do we solve this?

Data Minimization

Takeaways

- Privacy is not just for lawyers, but a cross-functional discipline
- Know what you collect, classify it and do it early
- In using and sharing data, make it coarser to protect privacy
- Minimize your data