Cloud Privacy Labs

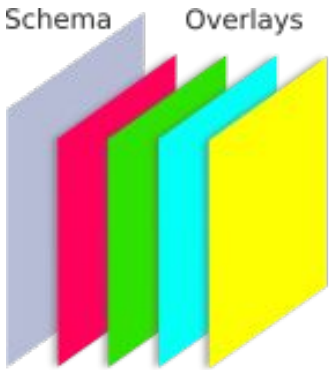
Semantic Pipelines with Layered Schemas

Burak Serdar

bserdar@cloudprivacylabs.com

- Layered Schemas
- Semantic Pipelines
- Use-case: Vaccine Credential Generation
- Use-case: ONC LEAP: Semantic Harmonization of Health Data

Layered Schemas



Schema base defines structure

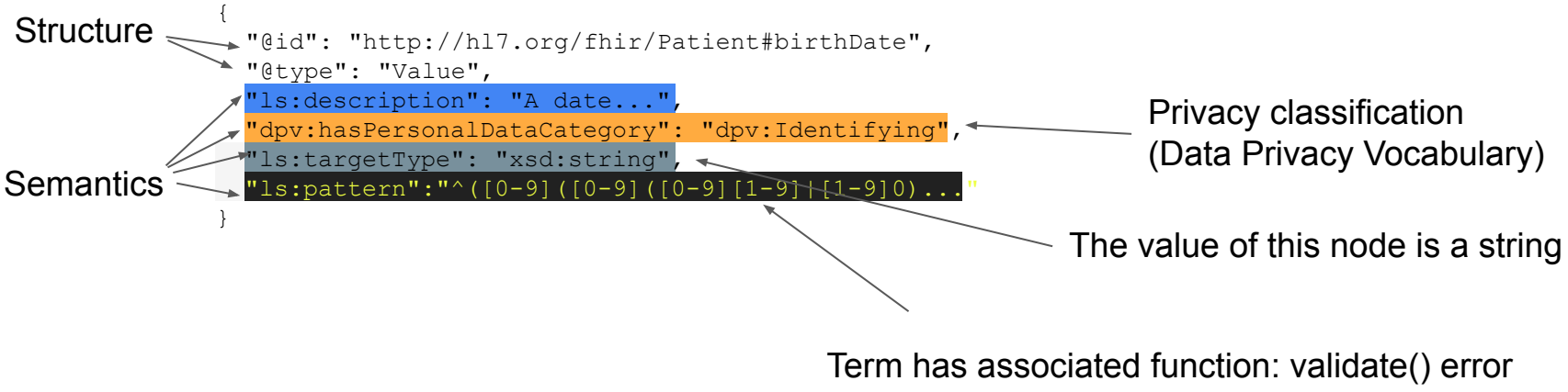
Layers add

- Constraints (min length, max value, required, etc.)
- Format (phone number, date/time, etc.)
- Language (English, Spanish, etc.)
- Dictionary/terminology references
- Privacy/security classifications (PII, Sensitive, etc.)
- Legal basis (HIPAA, consent, etc.)
- Provenance information
- Retention policies
- ...

Schema variant = Schema base + overlays

Layered Schemas: Structure + Semantics

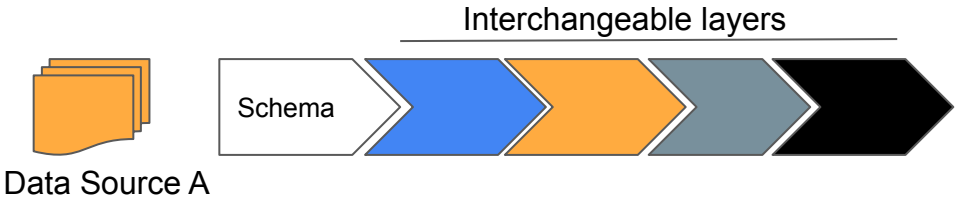
- Metadata as layers: open-ended tags (terms, codes, identifiers)
- Metadata defines meaning **and** functionality
  - Local/jurisdictional variances, representation differences, …
- Data translated to a Labeled Property Graph guided by metadata
- Different layers → different metadata/semantics
- Different data → common graph



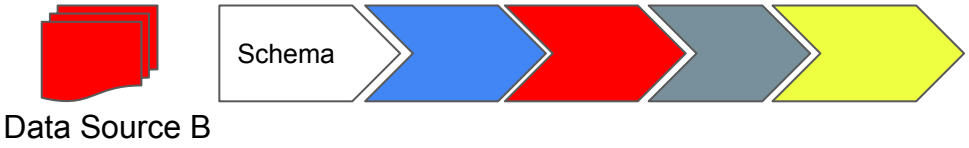
Schema Composition

Cloud Privacy Labs

Schema defines common structure of data
Interchangeable layers account for data-source specific variations during data ingestion
Schema and layers are reusable, shareable




```
{
  "@id": "http://hl7.org/fhir/Patient#birthDate",
  "@type": "Value",
  "ls:description": "A date...",
  "dpv:hasPersonalDataCategory": "dpv:Identifying",
  "ls:targetType": "xsd:string",
  "ls:pattern":"^([0-9]([0-9]([0-9][1-9]|[1-9]0)...
}
```




```
{
  "@id": "http://hl7.org/fhir/Patient#birthDate",
  "@type": "Value",
  "ls:description": "A date...",
  "hl7:confidentiality": "M"
  "ls:targetType": "xsd:string",
  "ls:pattern":"^([0-9]([0-9])/..."
}
```

Data as Labeled Property Graphs

Cloud Privacy Labs

JSON

```
"birthDate": "12/03/1980"
```

XML

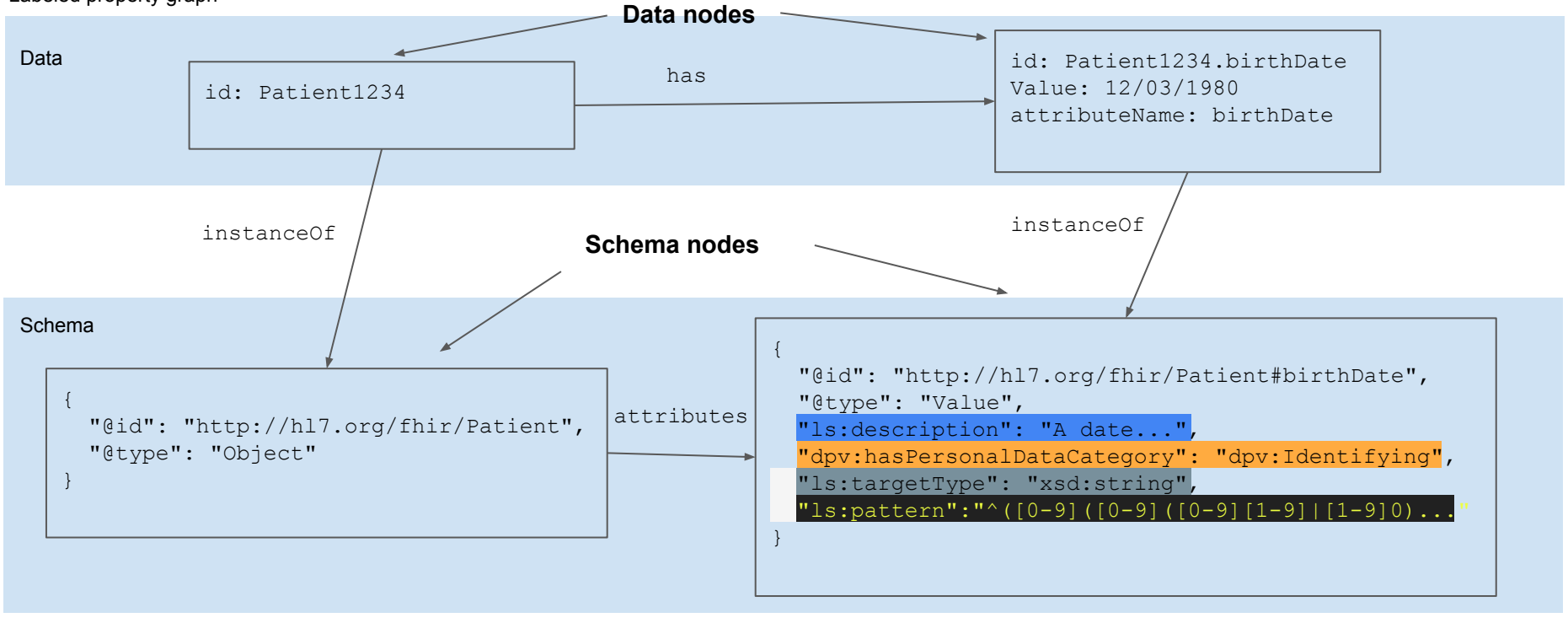
```
<birthDate>12/03/1980</birthDate>
```

CSV

```
birthDate,...
12/03/1980,...
```

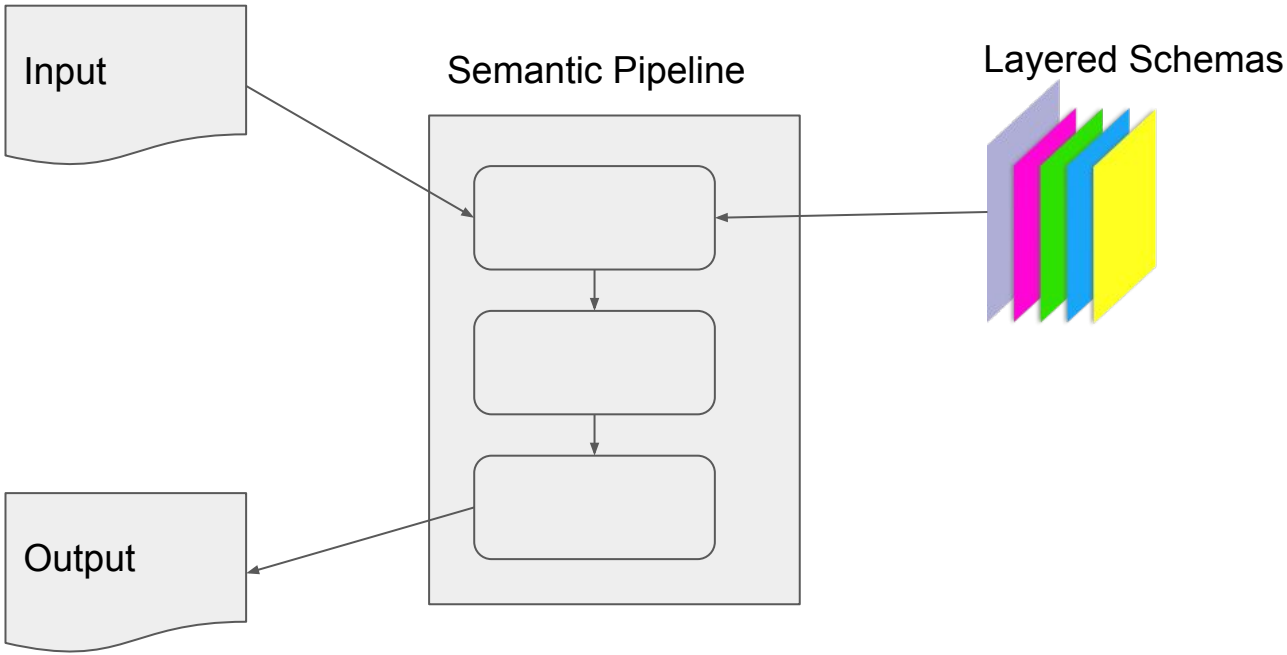
Input data

Labeled property graph



Semantic Pipelines

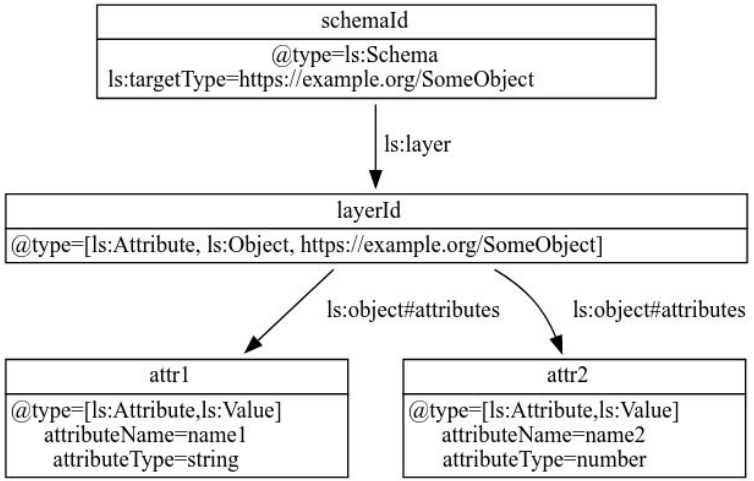
Cloud Privacy Labs



Stages: Ingest (data + schema → graph)

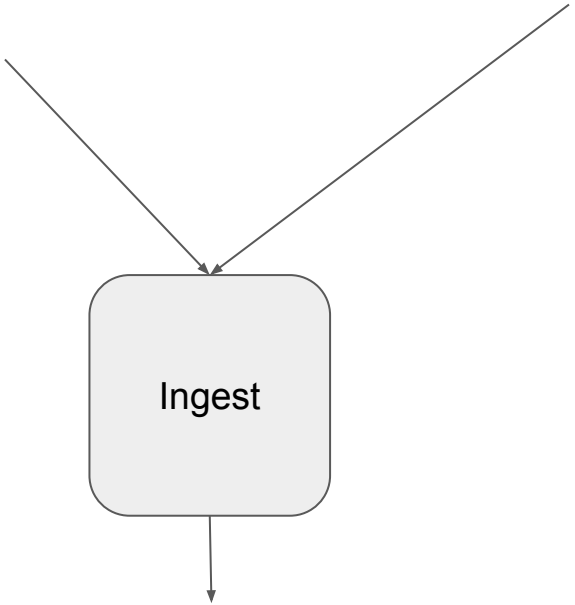
Cloud Privacy Labs

Layered Schema



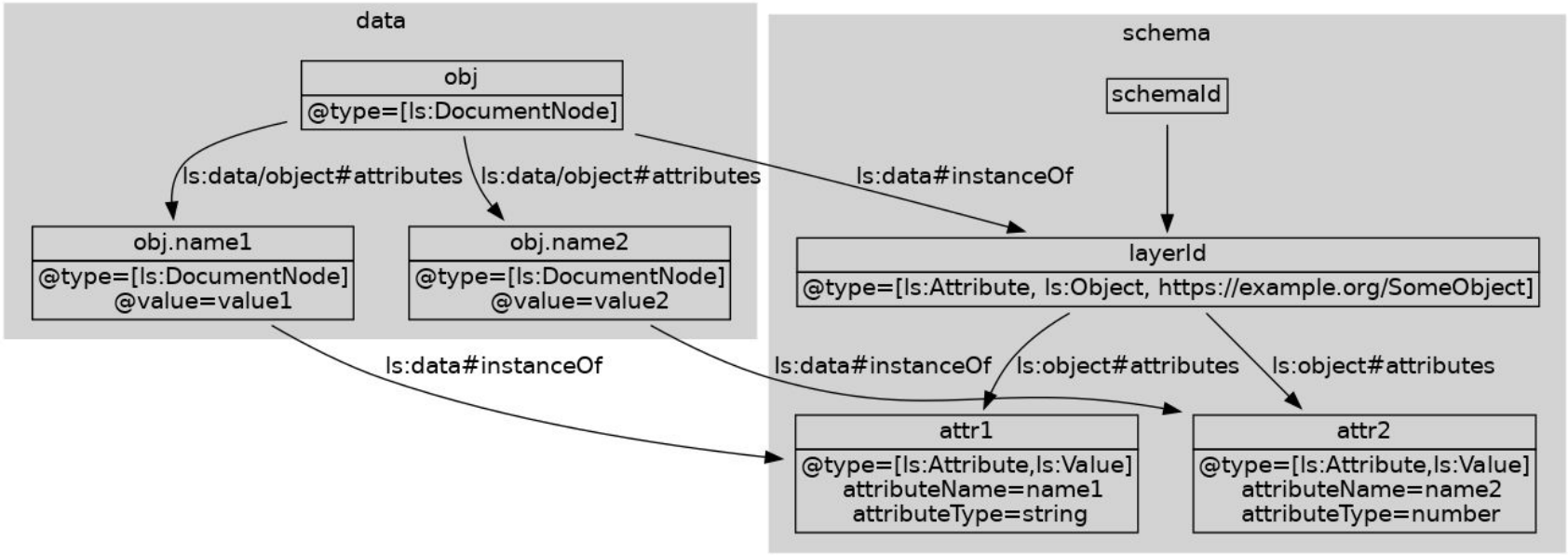
Data

```
{
   "name1": "value1",
   "name2": "value2"
}
```



Stages: Ingest

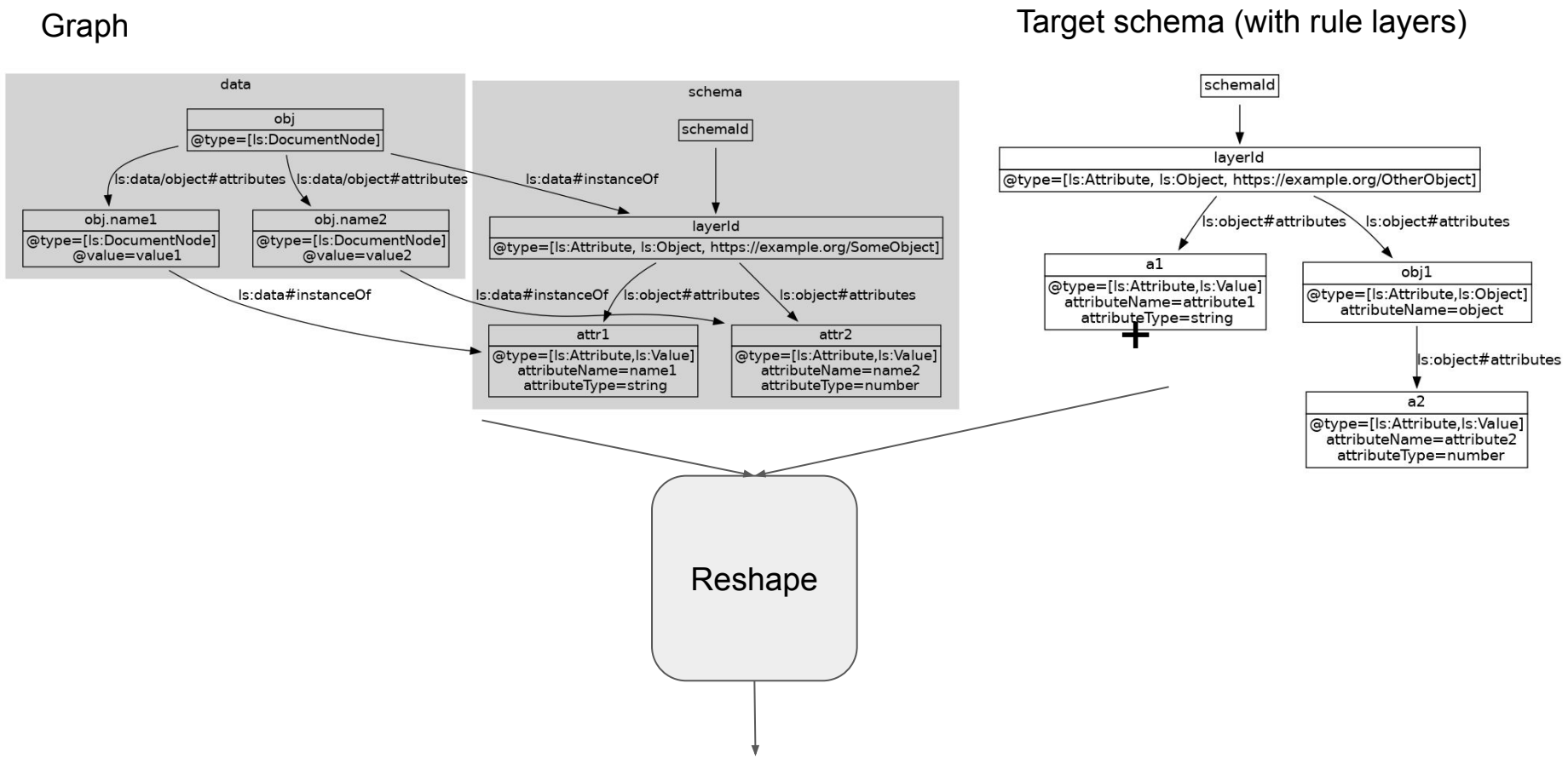
Cloud Privacy Labs



Labeled property graph with embedded schema

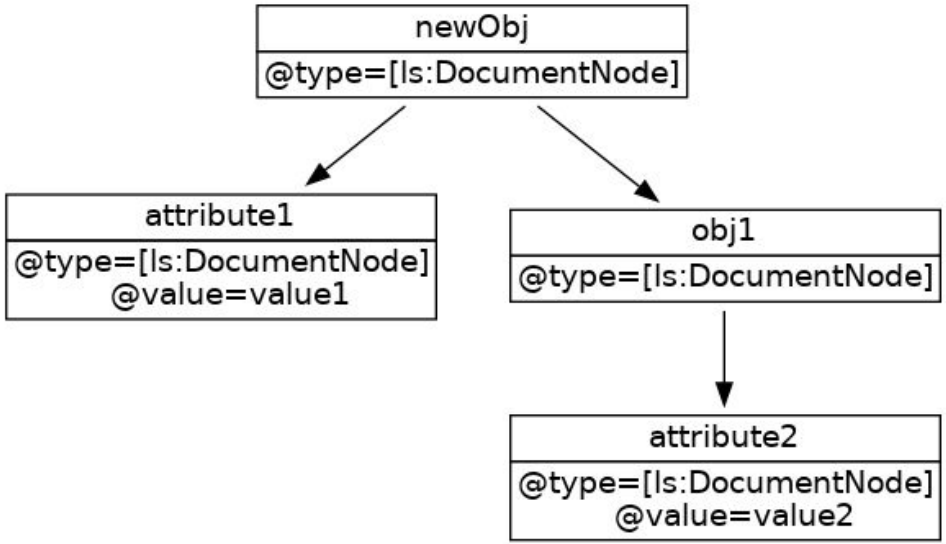
Stages: Reshape (graph + schema → new graph)

Cloud Privacy Labs



Stages: Reshape (graph + schema → new graph)

Cloud Privacy Labs




```
{
   "name1": "value1",
   "name2": "value2"
}
```

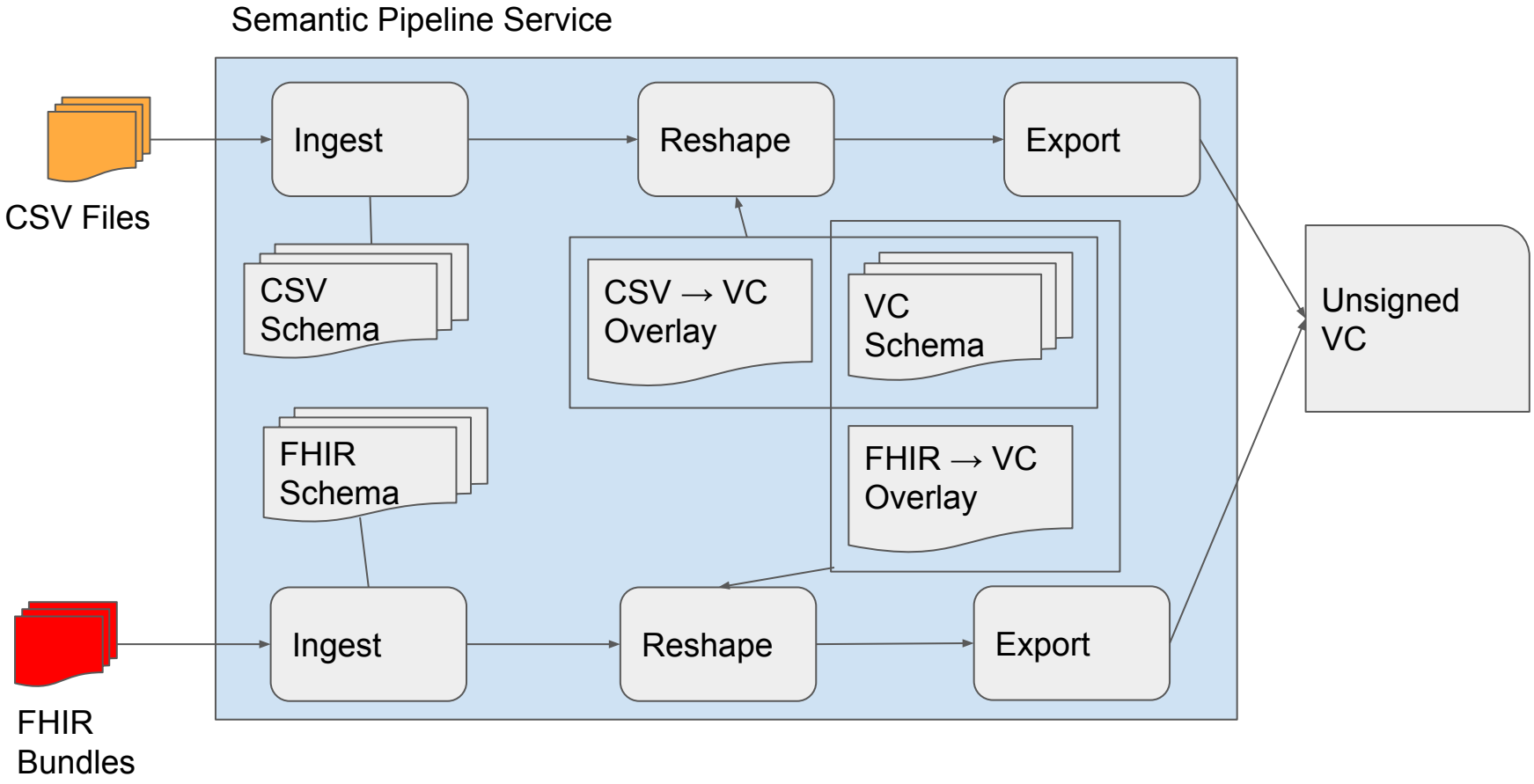
Reshape

```
{
   "attribute1": "value1",
   "obj1": {
       "attribute2": "value2"
   }
}
```

Data reshaped to conform to another schema

Cloud Privacy Labs

Pulse Connect Vaccine Credentials



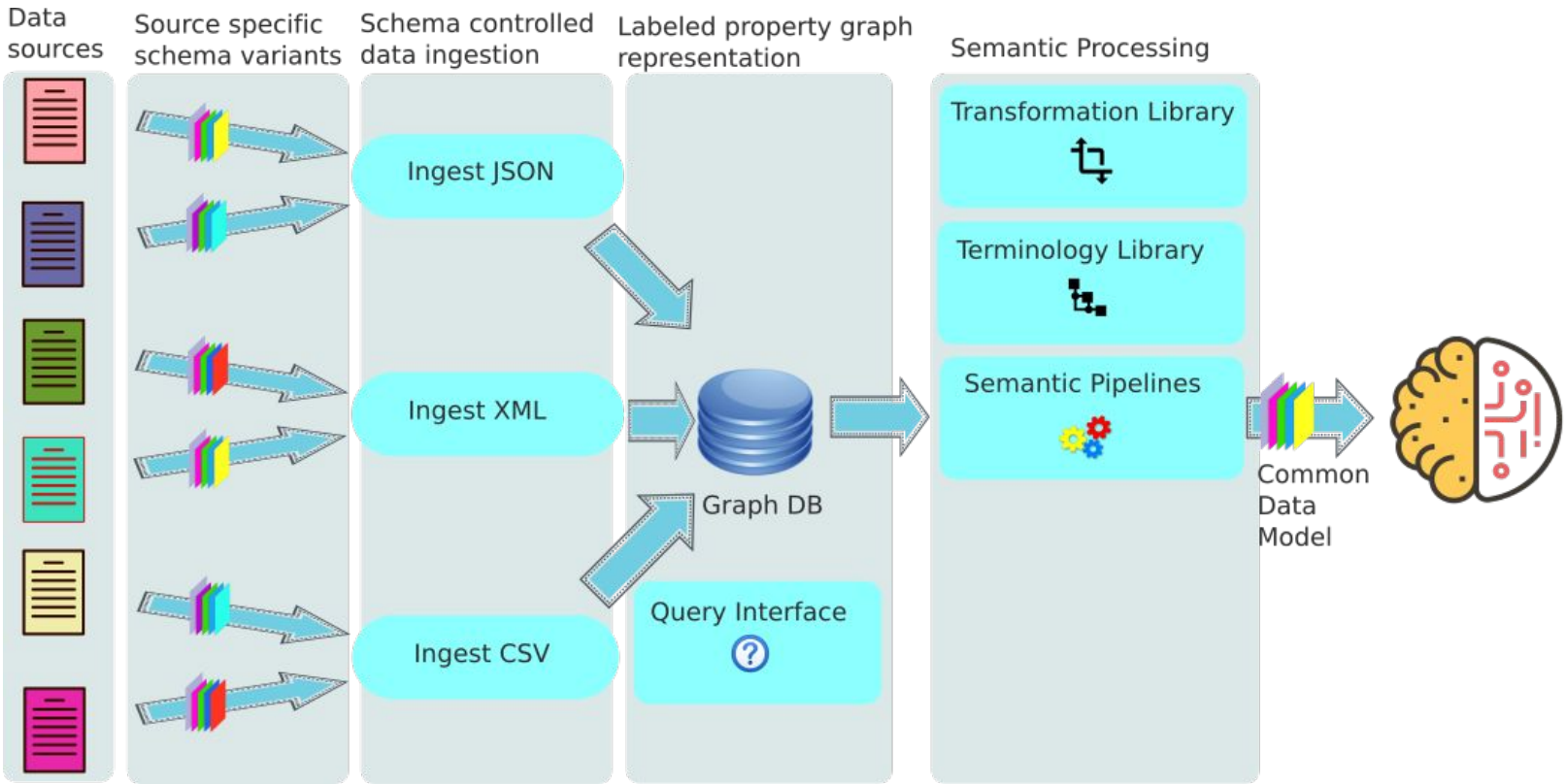
Cloud Privacy Labs

Use Case: A Semantic Data Warehouse for Health Data

- Collaboration with the DARTNet Institute and Cloud Privacy Labs
- Over 5000 clinical organizations contribute EHR data
- EHR data sets that cover over 30 million people
- Project Goals:
  - Replace source-specific ETL practices with a semantic pipeline based on the Layered Schema Architecture
  - Build reusable semantic mappings and components
  - Enable providers of all sizes to contribute data to a research data commons

This project is supported by the Office of the National Coordinator for Health Information Technology (ONC) of the U.S. Department of Health and Human Services (HHS) under grant number 90AX0034, **Semantic Interoperability for Electronic Health Data Using the Layered Schemas Architecture**, total award $999,990 with 100% financed with federal dollars and 0% financed with non-governmental sources. This information or content and conclusions are those of the author and should not be construed as the official position or policy of, not should any endorsements be inferred by ONC, HHS, of the U.S. Government.

Use Case: A Semantic Data Warehouse for Health Data



Use Case: A Semantic Data Warehouse for Health Data

Challenges:

- Values as code (2089-1) vs. text (Cholesterol in LDL [Mass/volume] in Serum)
- Same event represented in multiple tables (vaccination event as procedure and drug)
- Value sets, dictionaries: gender ("M" vs. "Male"), race ("9178" vs. "Non-white")
- Input variations (birth date as yyyy/mm vs. yyyy)
- Terminology crosswalks (LOINC, SNOMED, provider-specific codes)