Data Enrichment for data pipelines at scale

Richard Murphy

Sophron Networks LLC

Who/What is Sophron?

- From Ancient Greek *σώφρων* (sṓphrōn, “sane, moderate, prudent”) (from *σῶς* (sôs, “safe, sound, whole”) + *φρήν* (phrḗn, “mind”)) = Of Sound Mind
- Our culture stems from our name, we are a trusted partner.
- What started as a mobile device management startup in 2014 has transformed into a boutique technology consulting firm focused on providing innovative technology solutions to government agencies and private companies.

Who/What is Richard?

- Richard is the founder and CTO of Sophron Networks LLC, which was transformed in 2017 to provide consulting services to government agencies. He is currently consulting with the CDC on their DCMS Cloud and COLO hosting contract.

  Before starting Sophron, Richard spent 8 years with HPE as a Chief Technologist in the Federal Healthcare sector providing DevOps and Cloud consulting to federal agencies. During that time, Richard led modernization and DevOps transformation efforts at multiple agencies.

  Prior to HPE, Richard was a Software Development manager at ADP. Richard led the successful development of divisional data marts, ETL pipelines, and reporting services that delivered business insights to internal teams and reporting to over 10,000 businesses.

  Before that, he joined EDS, and in 1996 started his first 8 year tenure with the CDC as a contractor supporting GRASP and was part of the team that built the first Internet mapping system in CDC in 2000. Before joining EDS Richard worked as a federal employee/civil servant with the United States Probation Office/ US Courts... And his first full time job was as a mainframe and client-server programmer in a county Board of Education.

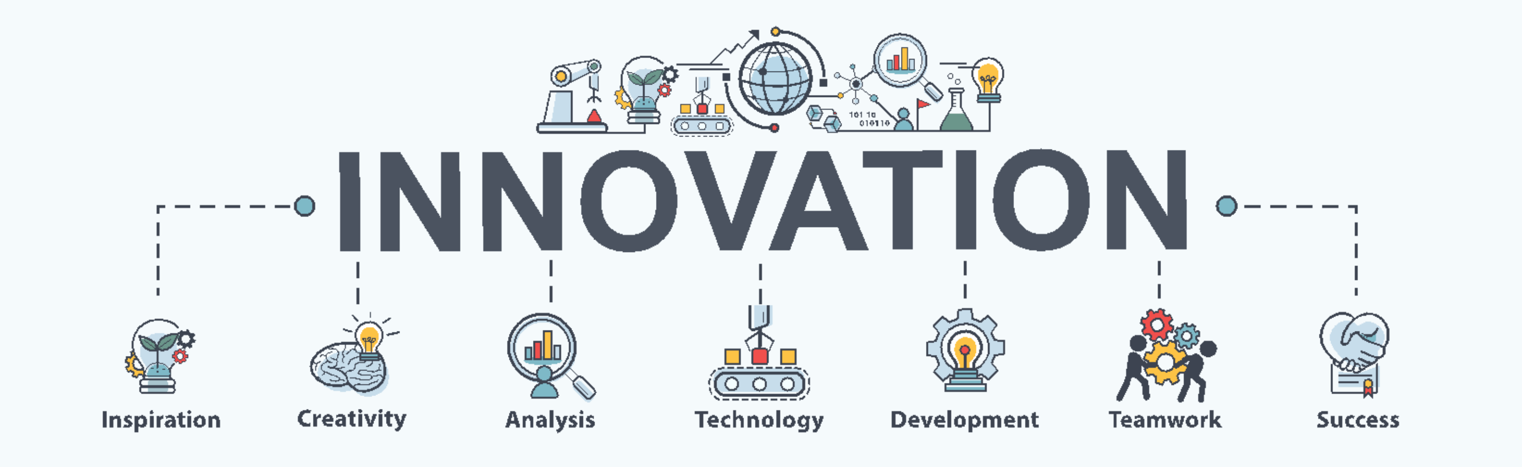
Who am I really?

- The COBOL guy.
- The PERL guy.
- The ETL guy.
- The DevOps guy.
- The Gitlab guy.
- The geocode guy.
- The boat guy.

Internship Program – R&D

- Our Research and Development activities center around our internship program, giving our interns opportunities to learn and grow, while giving our consultants the opportunity to learn new things and mentor young informaticians, scientists, and technologists.

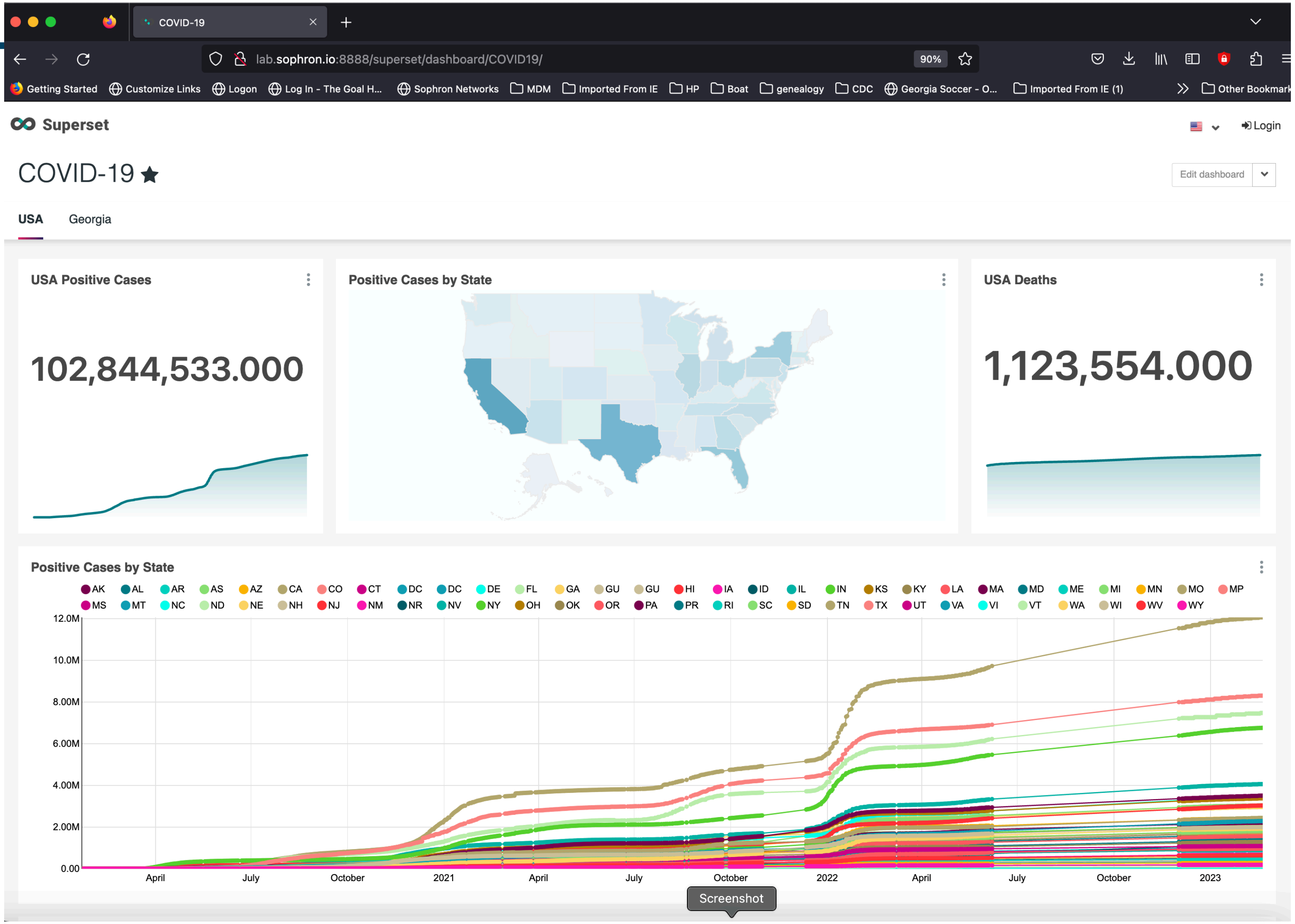
Most recent past project was SLINK, a record linking interface for use in R Studio.



Early project was COVID data tracker.
Recent projects have included data pipelines with Azure Logic Apps, Functions, and Liquid templates.

COVID Tracker

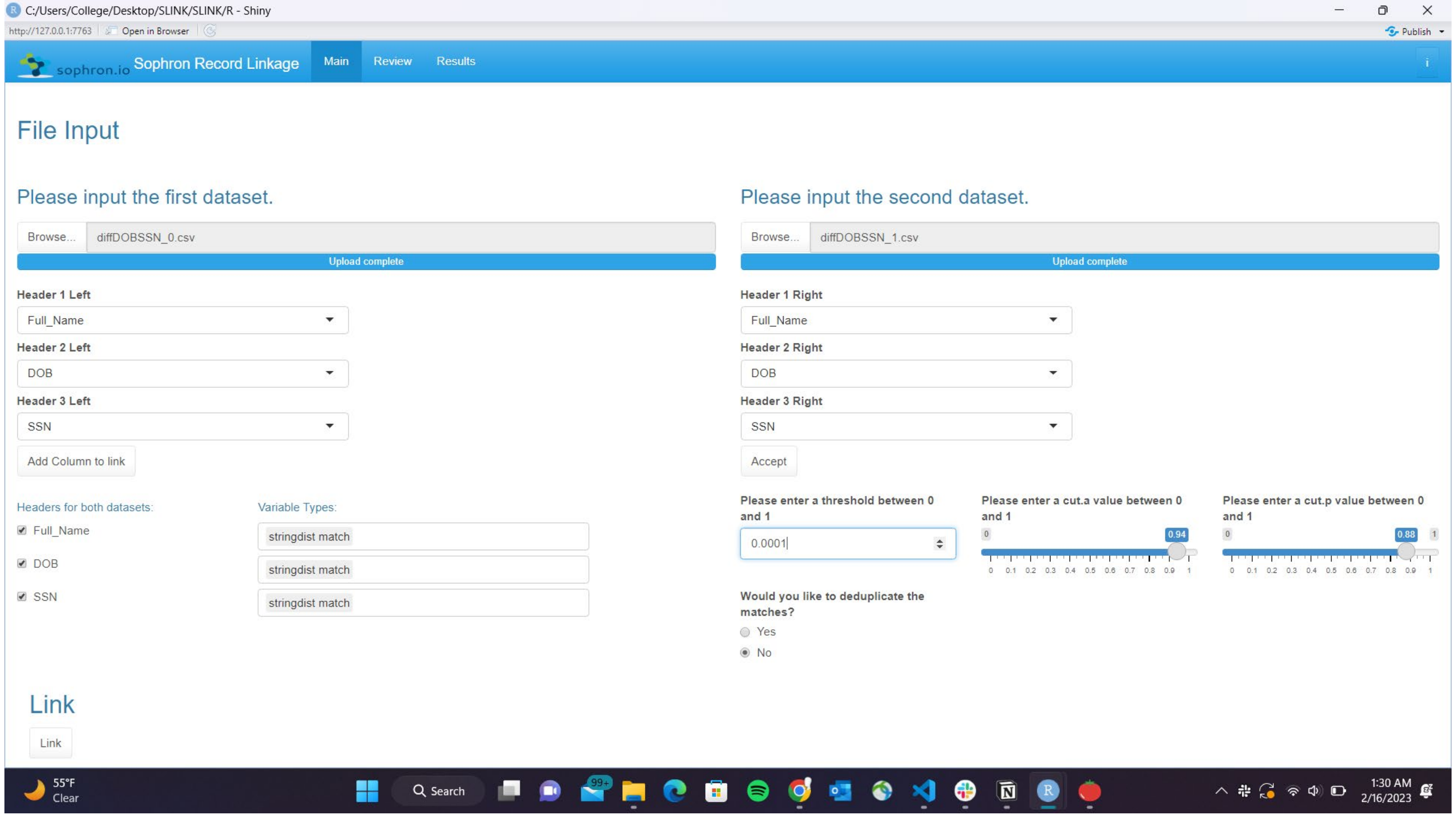
COVID tracker was the first intern group project. Utilizing ETL jobs to gather data from different sources and visualize using open source software. (retired last year, data no longer monitored)



SLINK

SLINK is an R Shiny program which provides a GUI interface to the fastLink package.

https://github.com/sophronio/SLINK



File Input

Please input the first dataset.

Browse... diffDOBSSN_0.csv

Upload complete

Please input the second dataset.

Browse... diffDOBSSN_1.csv

Upload complete

Header 1 Left: Full_Name

Header 2 Left: DOB

Header 3 Left: SSN

Header 1 Right: Full_Name

Header 2 Right: DOB

Header 3 Right: SSN

Headers for both datasets:	Variable Types:
☑ Full_Name	stringdist match
☑ DOB	stringdist match
☑ SSN	stringdist match

Please enter a threshold between 0 and 1: 0.0001

Please enter a cut.a value between 0 and 1: 0.94

Please enter a cut.p value between 0 and 1: 0.88

Would you like to deduplicate the matches?

○ Yes

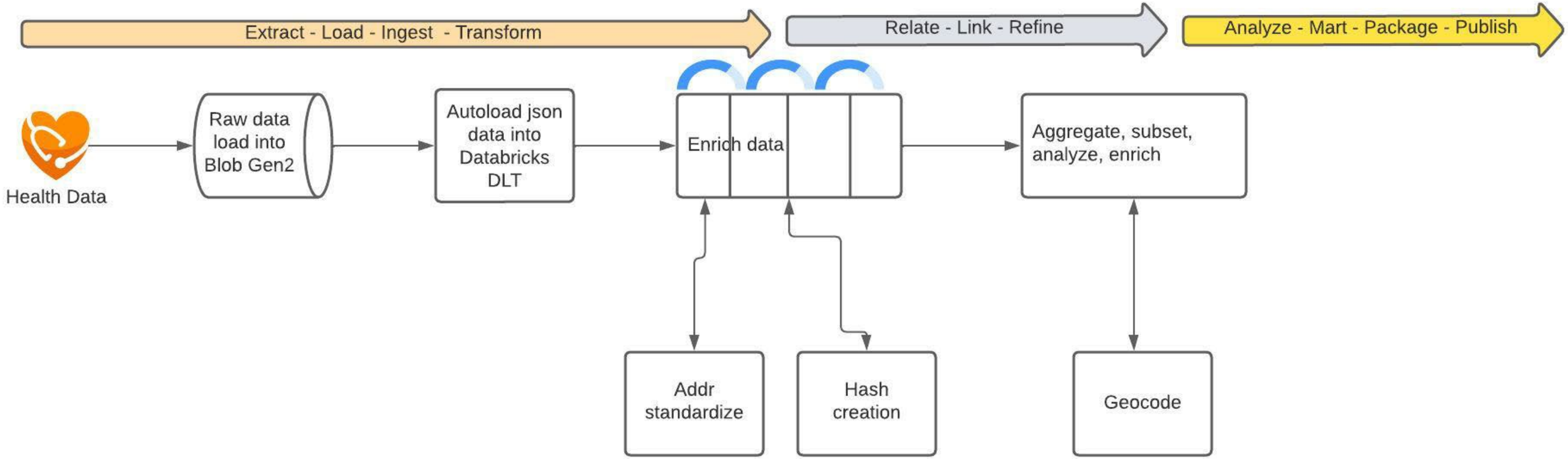
● No

Link

Current Project -Data Enrichment(services)

- Create Address and record linking services that can be of use in a data pipeline
  - Address standardization- convert addresses to USPS standard
  - Hashing service- create identifying hash based on name,DOB,address to enable fast deterministic matching
  - Geocoding- geocoding addresses to lat/lon to be used in analysis; we are not storing lat/lon as part of record because results from geocoding are not absolute
  - Record Linking- de-duplicating records and associating like records with one another; primarily *patient*

Data Flow


Extract - Load - Ingest - Transform
Relate - Link - Refine
Analyze - Mart - Package - Publish
Health Data
Raw data load into Blob Gen2
Autoload json data into Databricks DLT
Enrich data
Aggregate, subset, analyze, enrich
Addr standardize
Hash creation
Geocode

Geocoding

- Custom REST service built with Python/FastAPI which fronts a PostGIS server loaded with TIGERLine 2020 data.
- ArcGIS geocoding API
- Azure Maps gocoding API
- Precisely geocoding service (formerly Pitney Bowes Geostan)

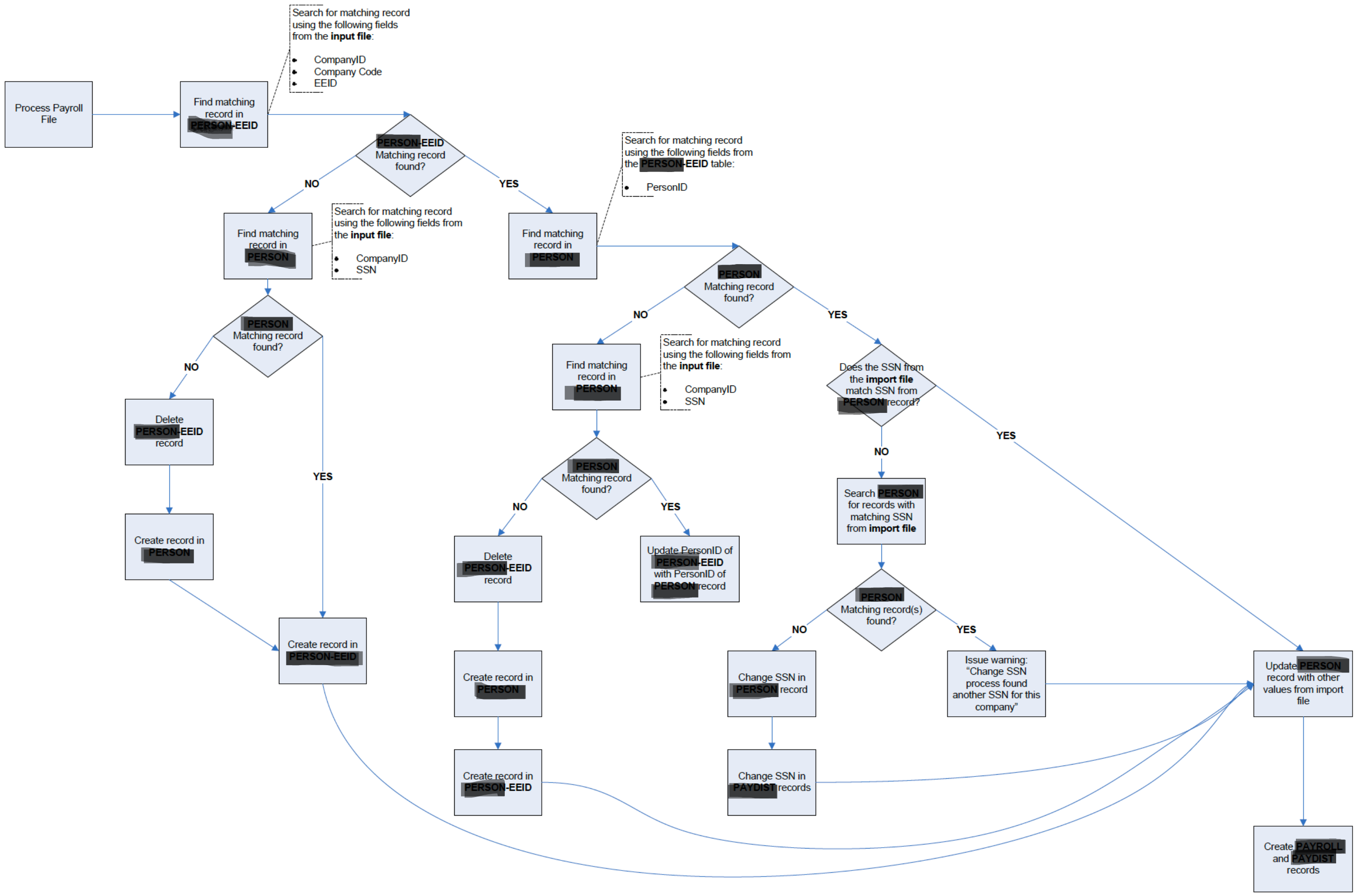
Geocoding Considerations

- Accuracy – many data sets are updated with new streets, street ranges but do not contain point data
    - Some data sets are more accurate- TeleAtlas, Navteq are more accurate and updated more frequently than TIGER
        - Engines that use these rely on interpolation algorithms to approximate lat/lon
        - Most geocoding is done with these data sets
    - Some data sets also include their own points database
        - Precisely offers their own points database
- How accurate does it need to be? How fast does it need to be?
- Lat/lon calculated from geocoding is not absolute, can change based on updates to data sets, updates to interpolation algorithms, results should be annotated with metadata on the date, software, data set used

Record Linking

Record linkage (also known as **data matching**, **data linkage**, **entity resolution**, and many other terms) is the task of finding records in a data set that refer to the same entity across different data sources (e.g., data files, books, websites, and databases). Record linkage is necessary when joining different data sets based on entities that may or may not share a common identifier (e.g., database key, URI, National identification number), which may be due to differences in record shape, storage location, or curator style or preference. A data set that has undergone RL-oriented reconciliation may be referred to as being *cross-linked*. - Wikipedia

Record Linking – Financial Services
Person Use Case



Process Payroll File

Find matching record in [redacted]-EEID

Search for matching record using the following fields from the **input file**:

- CompanyID
- Company Code
- EEID

[redacted]-EEID Matching record found?

NO

Find matching record in [redacted]

Search for matching record using the following fields from the **input file**:

- CompanyID
- SSN

[redacted] Matching record found?

NO

Delete [redacted]-EEID record

Create record in [redacted]

YES

Create record in [redacted]

YES

Find matching record in [redacted]

Search for matching record using the following fields from the [redacted]-**EEID** table:

- PersonID

[redacted] Matching record found?

NO

Find matching record in [redacted]

Search for matching record using the following fields from the **input file**:

- CompanyID
- SSN

[redacted] Matching record found?

NO

Delete [redacted]-EEID record

Create record in [redacted]

Create record in [redacted]-EEID

YES

Update PersonID of [redacted]-EEID with PersonID of [redacted] record

YES

Does the SSN from the **import file** match SSN from [redacted] record?

NO

Search [redacted] for records with matching SSN from **import file**

[redacted] Matching record(s) found?

NO

Change SSN in [redacted] record

Change SSN in [redacted] records

YES

Issue warning: "Change SSN process found another SSN for this company"

YES

Update [redacted] record with other values from import file

Create [redacted] and [redacted] records

sophron.io

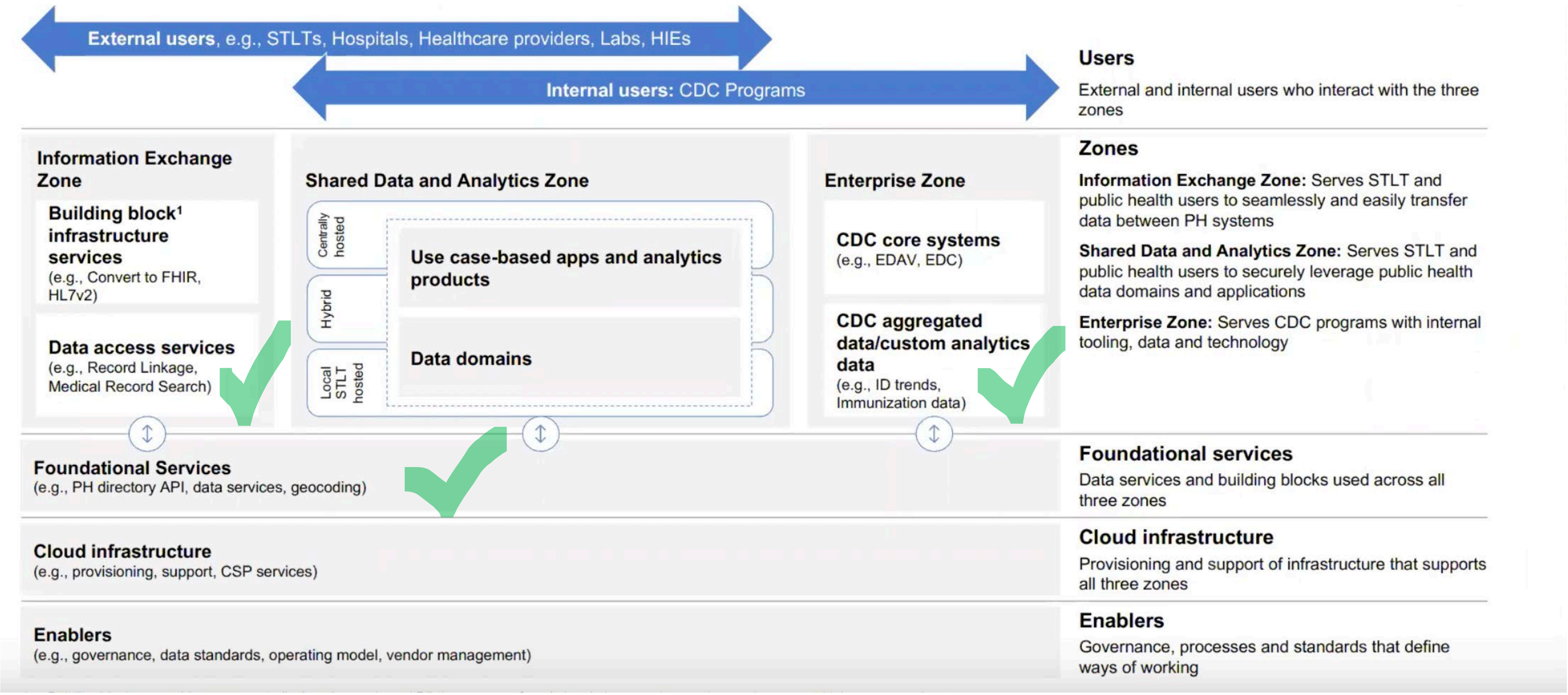
Two Phased Linking

- Before ‘linking’ takes place, but after after raw data is ingested into ‘rare’ tables, a job will run that utilizes the hashing service to create a hash based on name, address, DOB. This hash is then stored along with the record in the patient table.
- The ‘linking’ job will then run as one of the jobs in the next part of the pipeline. This job will first perform a deterministic linking of the new record against the patient table to see if there is a match utilizing the value of the new ‘hash’ field previously calculated.
- If there is no match, then the second phase of linking will occur and a probabilistic algorithm will be executed and any possible matches will be placed into a ‘possible_match’ table, along with the score, for further review.

Project Status

- Custom TIGER geocoding service is built and undergoing testing
- Geocoding testing is underway with Azure Maps, Google Maps, Precisely, ESRI
  - Comparing accuracy, standardization, speed
- Hashing service is built and undergoing testing
- Linking system is in development
- Noodling on Machine Learning for Record Linking

Relation to Northstar Architecture



External users, e.g., STLTs, Hospitals, Healthcare providers, Labs, HIEs

Internal users: CDC Programs

Information Exchange Zone

Building block[1] infrastructure services (e.g., Convert to FHIR, HL7v2)

Data access services (e.g., Record Linkage, Medical Record Search)

Shared Data and Analytics Zone

Use case-based apps and analytics products

Data domains

Enterprise Zone

CDC core systems (e.g., EDAV, EDC)

CDC aggregated data/custom analytics data (e.g., ID trends, Immunization data)

Foundational Services (e.g., PH directory API, data services, geocoding)

Cloud infrastructure (e.g., provisioning, support, CSP services)

Enablers (e.g., governance, data standards, operating model, vendor management)

Users

External and internal users who interact with the three zones

Zones

Information Exchange Zone: Serves STLT and public health users to seamlessly and easily transfer data between PH systems

Shared Data and Analytics Zone: Serves STLT and public health users to securely leverage public health data domains and applications

Enterprise Zone: Serves CDC programs with internal tooling, data and technology

Foundational services

Data services and building blocks used across all three zones

Cloud infrastructure

Provisioning and support of infrastructure that supports all three zones

Enablers

Governance, processes and standards that define ways of working

Closing & Questions

- Themes: Flexibility, modularity, KISS, Enterprise, Prototype
- Results? http://sophron.io/

- Questions?
- Feel free to reach out: Richard.murphy@sophronnet.com