PDMP Patient Matching Challenges and Opportunities – Utah's Perspective

Naresh Sundar Rajan, PhD, MS

Senior Health Informaticist,

Controlled Substance Database,

Division of Occupation and Professional Licensing,

Utah Department of Commerce

Patient Matching Challenges/Opportunities

- Universal problem
  - *"No single solution to patient matching" – GAO, Jan. 2019 Report to Congress*
- Why the movement toward a patient identifier is only a start
  "Instead of looking at a national patient identifier as the definitive answer for solving our patient identification issues, we should not lose sight of a pragmatic, multi-faceted approach to improving patient matching—one that relies on a combination of ***probabilistic algorithms for connecting identifiers, processes and standards for data capture, biometric and identity verification tools*** *and* ***consumer involvement for managing their demographic information****." – Dan Cidon*

Patient Matching Efforts in Utah – Beyond PDMP

- Department of Health MPI (DOHMPI)
  - Create a gold-standard MPI by linking different data sources across Utah, Vital Records, Cancer Registry, Controlled Substance Database, All-Payer Claims Database and etc.
- Utah Health Information Network (UHIN) - MPI
  - A REST-ful MPI Service to search patients across Utah's population.
  - Authorized organizations can search using: Name, Gender, DOB, Address, Phone (Home, Work and Mobile) and SSN

Utah’s Controlled Substance Database

- Legislatively created 1995
- Collects data on dispensing of Schedule II– V drugs from retail, institutional, and outpatient hospital pharmacies, and in-state/out-of-state mail order pharmacies
- Housed with Division of Occupational and Professional Licensing (DOPL), Department of Commerce
- Current reporting standard: ASAP Version 4.2

Utah's CSD – Data Elements for Patient Matching

- Demographics (PATIENT TABLES)
  - First Name **(required)**
  - Last Name **(required)**
  - DOB **(required)**
  - Address **(required)**
  - City **(required)**
  - Zip-code **(required)**
  - Gender (optional)
  - Middle Name (optional)

Common Demographics Data Elements Issues

- Quality of data elements
  - Missing values
  - Inaccurate values
  - Validity of address fields (address, zip, city)
- Other issues
  - Nicknames
  - Swapped names
  - Abbreviated names, addresses
  - Misspelled names

Existing Methods

- Traditional approaches – Fuzzy Algorithms
- Edit distance methods (Levenshtein, Affine Gap Distance)
  - character-by-character distance between two names
  - Cindy vs Cyndi
- Computationally expensive due to large pair-wise comparisons, for example: 10000 patients, close to 5 million comparisons.
- Other solutions: Proprietary and expensive

Existing Methods

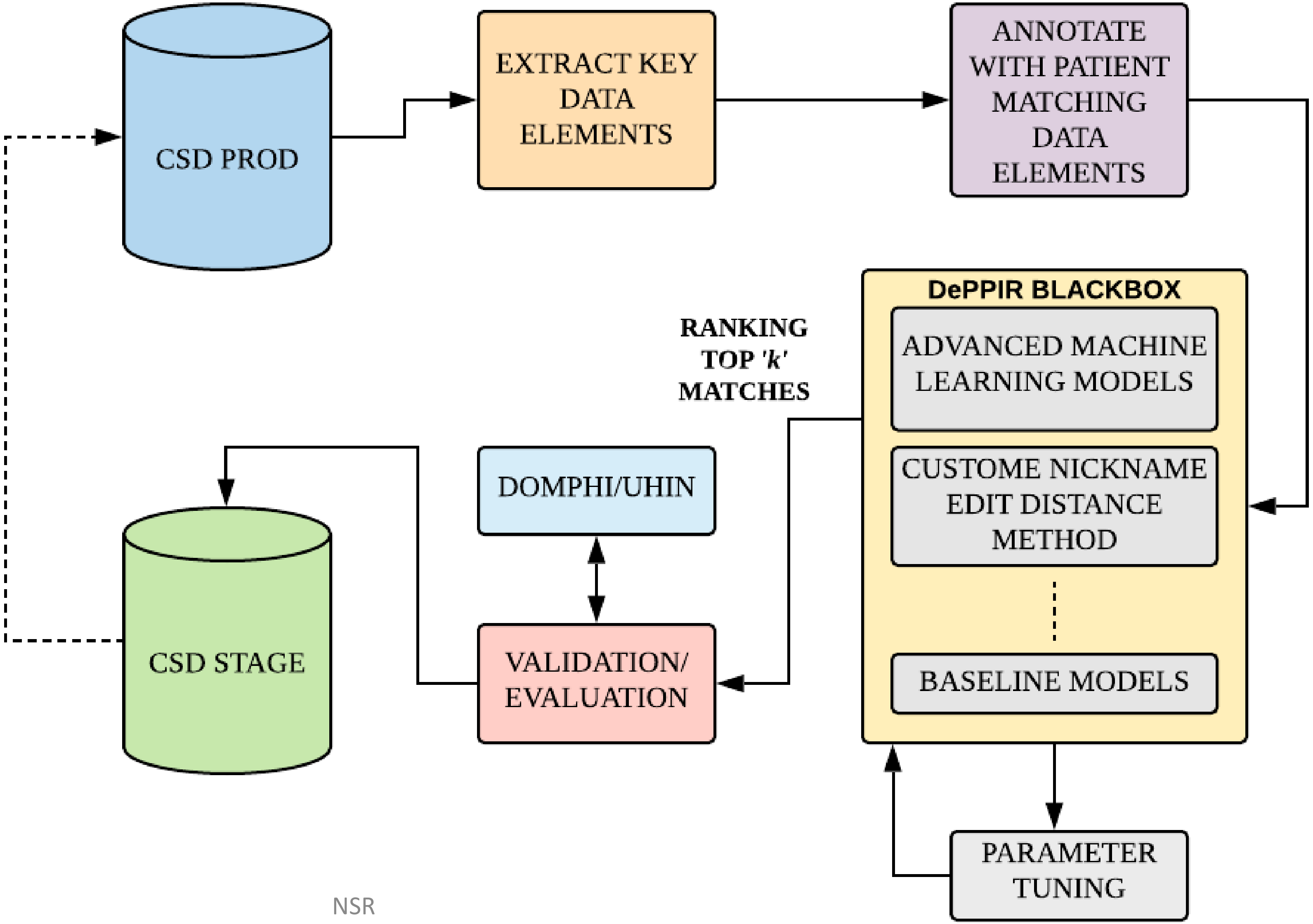
- Dedupe Library
  - Open-source/paid version for advanced user (millions of records)
  - Python based library
  - Segmenting the data using first few characters of firstname/lastname
  - Scalability is still a bottleneck when applying on big-data (or even few million records)

Deep Probabilistic Patient Identity Resolution (DePPIR)

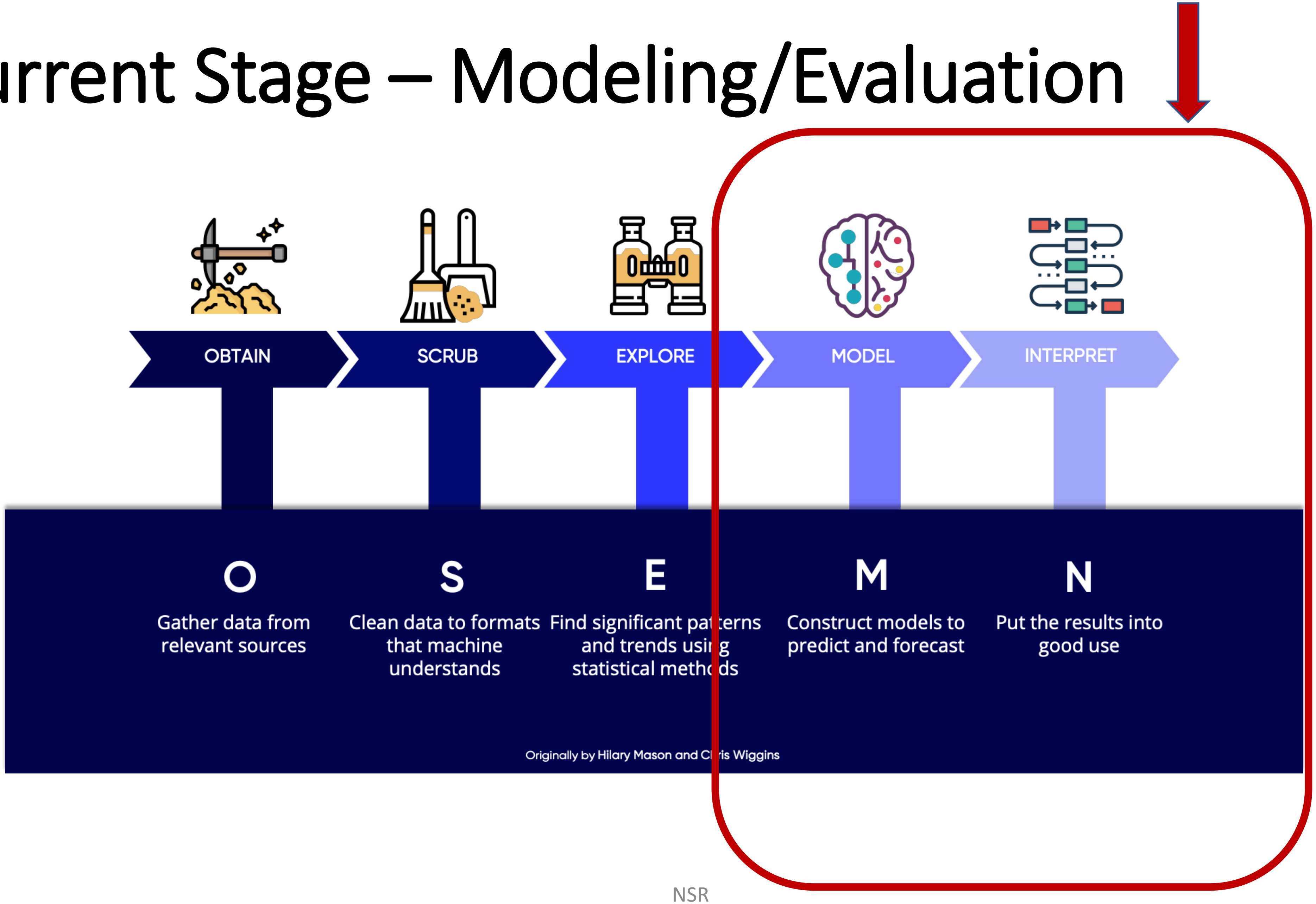
- Patient matching problem as a data science challenge
- Open-source (Python, PySpark/Apache Spark, TensorFlow)
- Supervised Machine Learning based methods and annotated ASAP 4.2 version data model
- Hybrid approaches for blocking data to reduced pair-wise comparison by a significant number

DePPIR – Architecture

- Architecturally Open-sourced supervised learning methods using annotated ASAP 4.2 data elements
- Custom built edit-based method for identifying nicknames
- Outputs probability of match, ranked *k* best matches



Current Stage – Modeling/Evaluation



Next Steps

- Result Validation/Evaluation
  - UHIN's MPI
  - DOH-MPI
- Explore feasibility of exposing DePPIR as a service (using FHIR standards) for enhancing interstate PDMP Patient Matching
- Standardizing validity checks at the point of data ingestion to increase quality of data thereby increasing quality of matching downstream
  - Specifically for Address/zip/city fields

Summary

- Leverage use of sophisticated technologies (probabilistic theory/statistical methods)
- Reduce human errors, create standardized data capture methods, and validity checks at the point of data ingestion
- Improve matching by including external sources such as biometrics, and Internet of Things (IoT)

Acknowledgements

- Jeya Balaji Balasubmramanian, PhD,
  School of Computing and Information,
  University of Pittsburgh
- Kailah Davis, PhD,
  Director, Center for Health Data Informatics,
  Utah Department of Health

Thank you!

- Naresh Sundar Rajan
- Email: nsr@utah.gov or nsr.informatics@gmail.com