Oak Ridge National Laboratory

March 13th

Computational approaches to Assessing Radiation Exposure Across the Lifecourse: Risks, Insights, and Innovations.

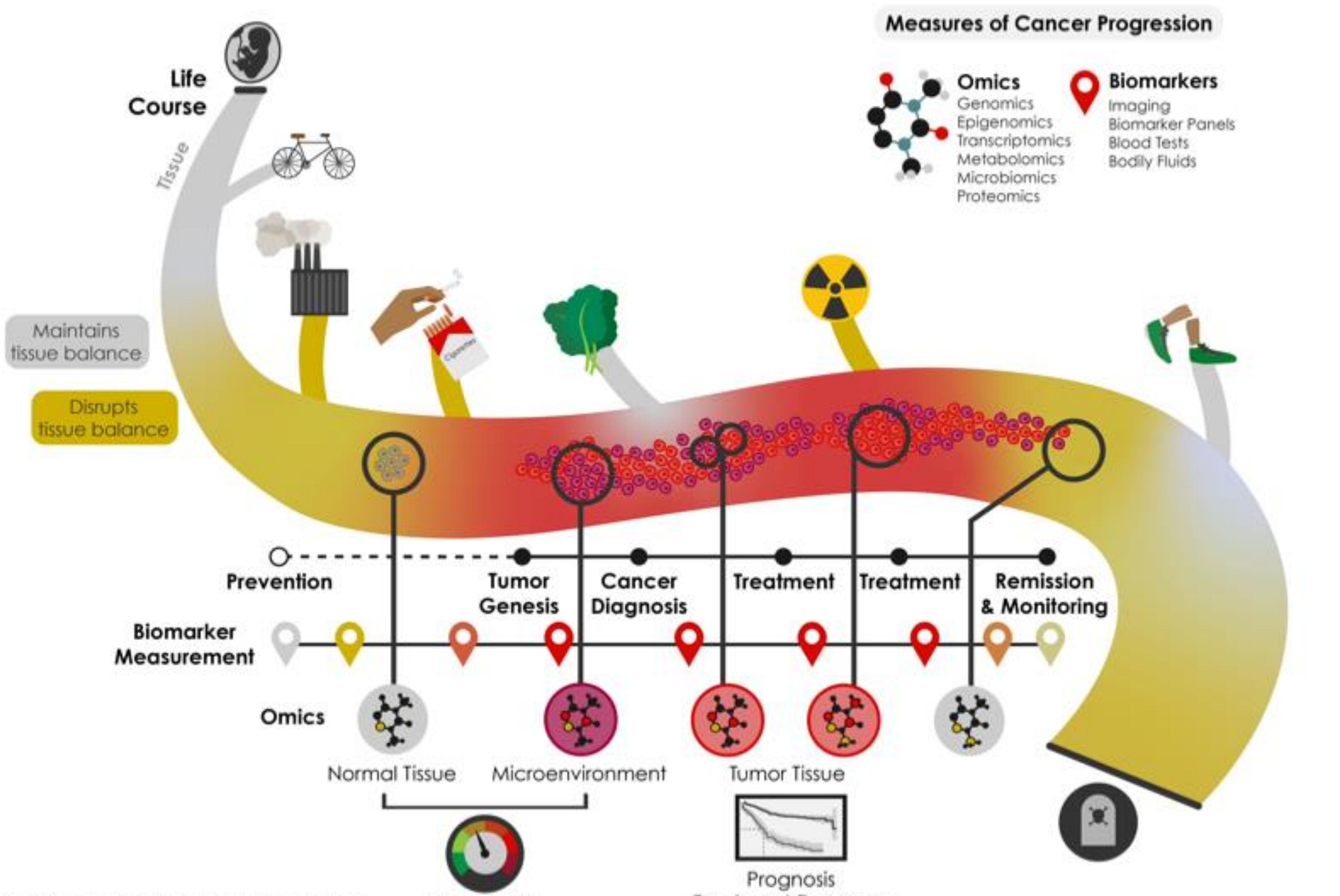
Heidi Hanson

Computational Sciences and Engineering Division

U.S. DEPARTMENT OF ENERGY

ORNL IS MANAGED BY UT-BATTELLE LLC
FOR THE US DEPARTMENT OF ENERGY

Life Course Approach to Radiation Exposure



BC Kelly, University of Utah. HA Hanson, Oak Ridge National Laboratory. NJ Camp, Huntsman Cancer Institute. 2021.

Charting the Life Course: An Interdisciplinary Blueprint

5 Pathways Forward

1 Integrate datasets across the lifespan.

Use transdisciplinary principles and methods, borrowing from:

- Engineering
- Statistics
- Communication
- Mathematics
- Epidemiology

2 Data science cores focus on curation, cleaning, analysis, and modeling of data.

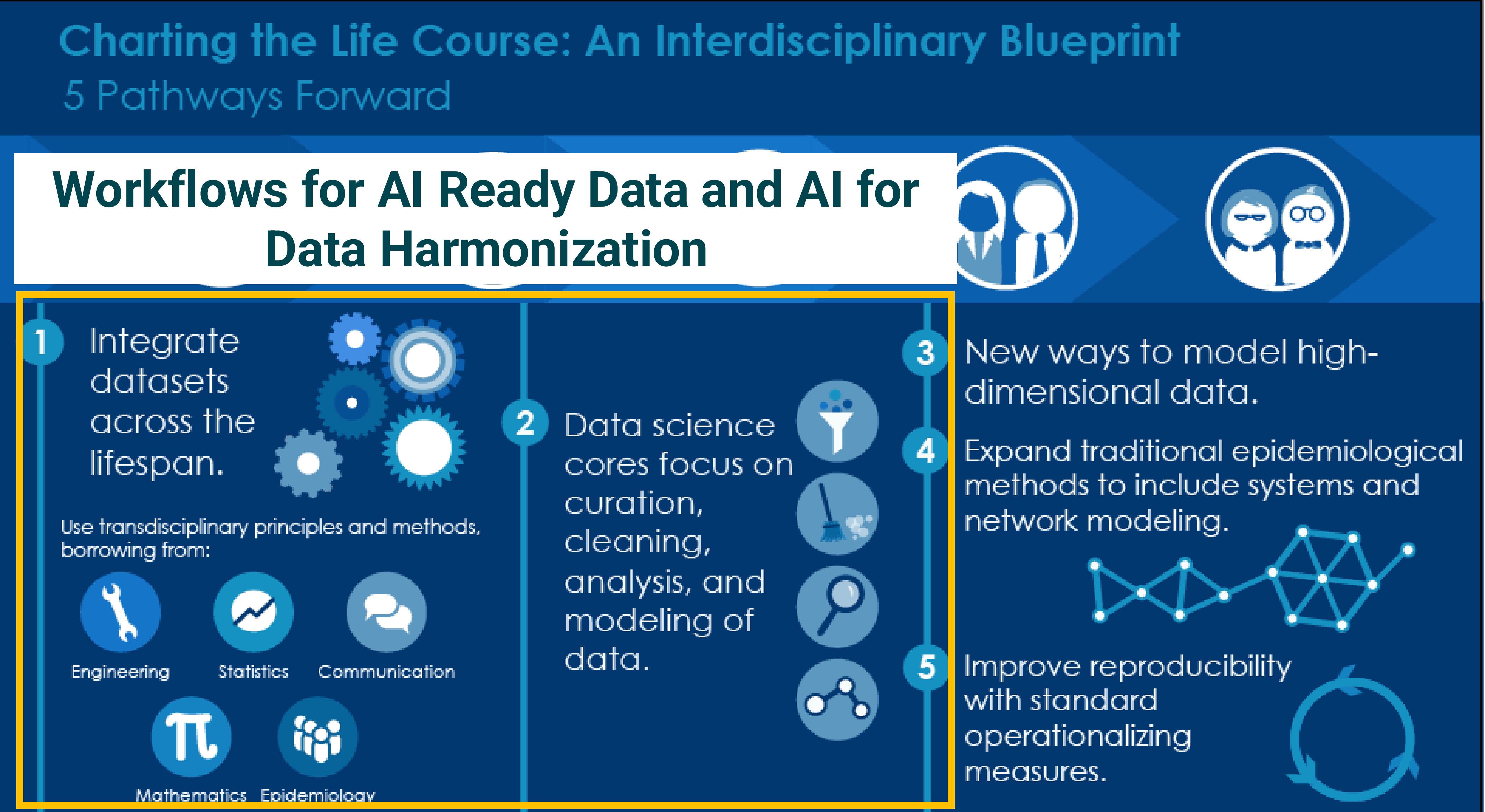
3 New ways to model high-dimensional data.

4 Expand traditional epidemiological methods to include systems and network modeling.

5 Improve reproducibility with standard operationalizing measures.

Heidi Hanson and Shari Barkin

Workflows for AI Ready Data and AI for Data Harmonization



Heidi Hanson and Shari Barkin

Integrated Electronic Health, Administrative, and Population Data

Computational tools enabling decision making for risk assessment, prevention, and treatment

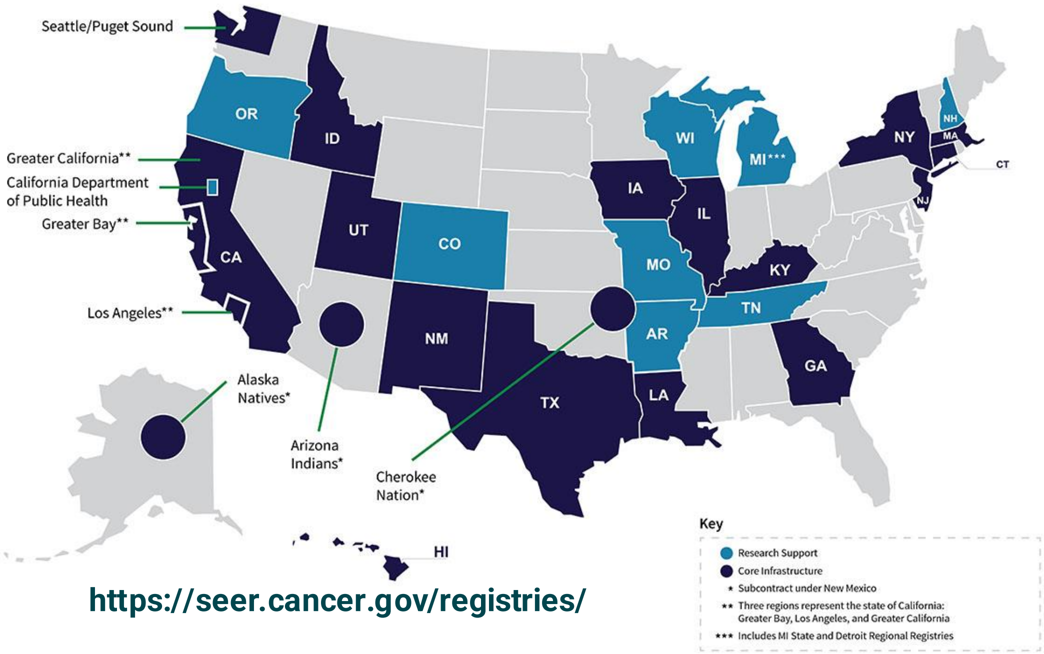
- Easy management of large amounts of heterogenous and siloed sensitive information
- Assessment of longitudinal trajectories of radiation exposure from residential radon, medical treatments, and occupational exposures.
- Identification of geographical areas at risk for high residential radon exposure
- Data-driven tools to guide decision making in the clinic and community

MOSSAIC

Modeling Outcomes Using Surveillance Data & Scalable Artificial Intelligence For Cancer

Real World Evidence: AI for Near Real-Time Health Surveillance Covering 48% of the US Population

Surveillance Epidemiology End-Results (SEER) Registries
> 850,000 Diagnoses Annually



https://seer.cancer.gov/registries/

Decision Making Tool Built into SEER Data Management System (2021)

Auto-Extraction from Pathology Reports:

Accuracy: **Auto-coding of 23-27% of path reports (N ~ 230,000)** with > 98% accuracy across all data elements.

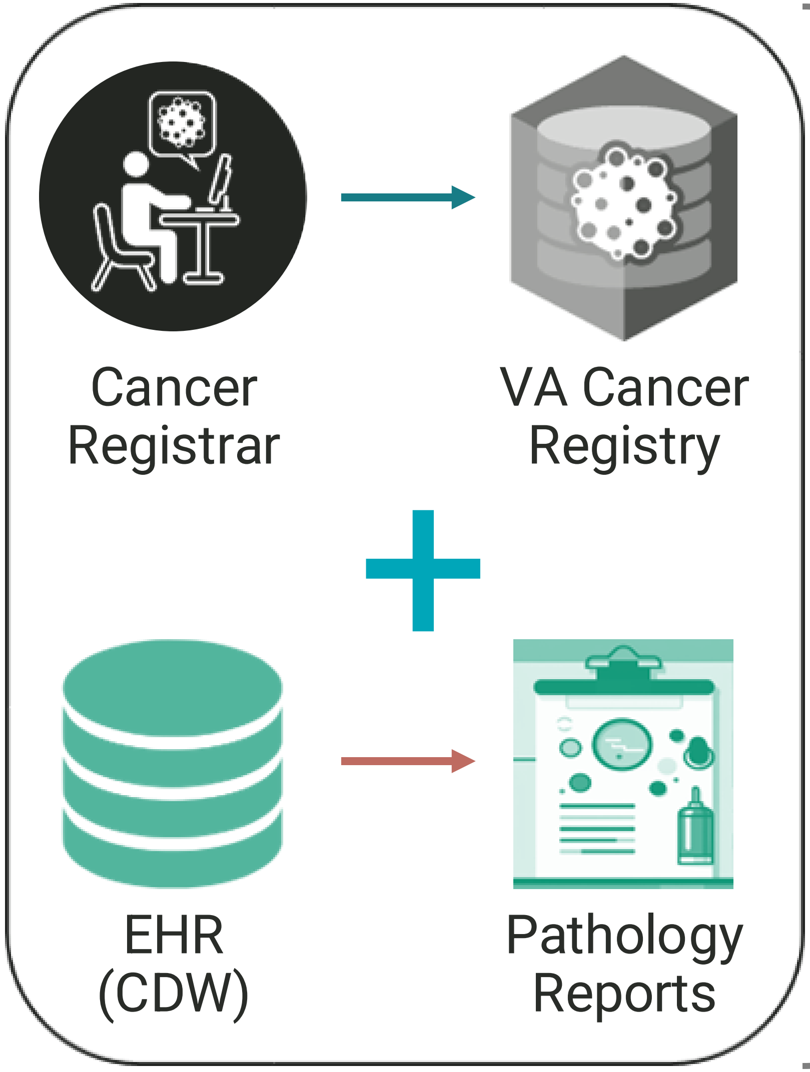
- Phenotype classifications:
  - Site = 70 categories
  - Sub-site = 324 categories
  - Histology = 626 categories
  - Laterality = 7 categories
  - Behavior = 4 categories

Production implementation Hierarchical Self Attention Model (HiSAN) with Deep Abstention:

- **Total 16 registries** (additional 4 since March 2022; ~31% of US population)
- **Default** as part of any new DMS installation, regardless of SEER affiliation
- **5+ new registries** anticipated in 2023/2024
- Testing phase with the Veteran's Health Administration

Oak Ridge National Laboratory

Real World Evidence: Clinical Decision Support with EHR Data



FrESCO

- Classification (Reportability)
- Information Extraction
- Histology
- Behavior
- Site
- Laterality

Near Real-time Case Ascertainment

- National Quality of Care (Systems of Excellence)
- Tele-Oncology
- Cancer Registry

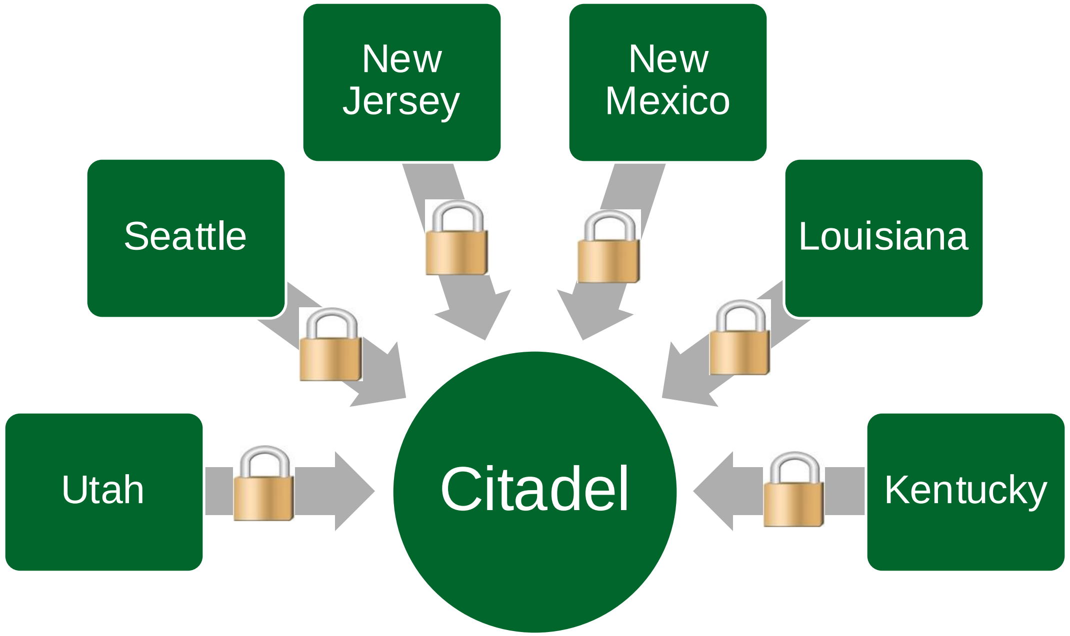
Oak Ridge National Laboratory

J. Stringer, K. Rasmussen, V. Patel, C. Li, A. Halwani

VA U.S. Department of Veterans Affairs

Interoperable AI Systems at Scale:

Bringing Compute Closer to the Source
Rapid Data Integration while Protecting Private Information



The National AI Research Resource (NAAIR) Task Force
National Childhood Cancer Registry
DOE ASCR Biopreparedness (BRaVE) Funding

Our current implementation

- Cross Silo
- Horizontal
- Model Centric – Need for Cooperative Agreement between institutions/ participants

Trusted host

Differential privacy

- “Epsilon Indistinguishability”

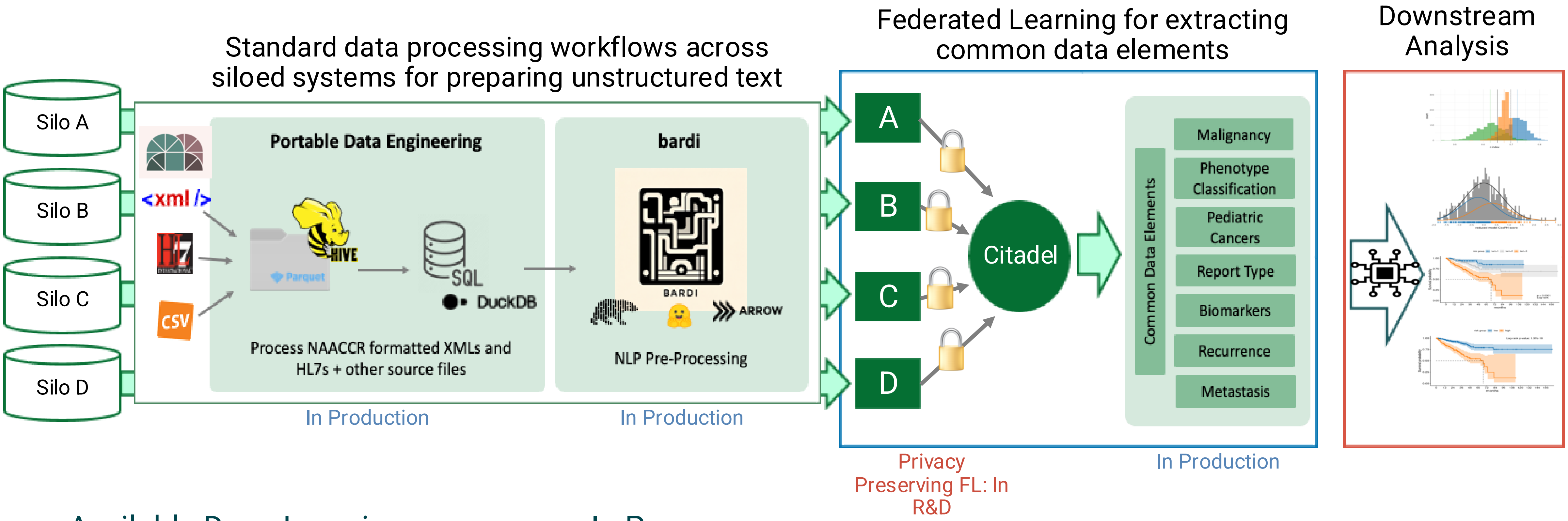
Privacy/accuracy tradeoff

High privacy

High accuracy

Innovation: SEER data make us completely innovative in this space. We can design and test solutions for <u>real-world application</u> at population scale.

MOSSAIC Computational Workflows
Rapid Response for Health Surveillance



<u>Available Deep Learning Tools</u>

MT-CNN, MT-HiSAN, Clinical BigBird, PathBigBird, Deep Abstaining Classifier

<u>In Progress:</u>

Multimodal/longitudinal DNN, Phrase Level Attention DNN, Tools for Bias Detection, Exposomic Data Linkage, Privacy Preserving FL, Synthetic Clinical Reports

FrESCO

https://github.com/DOE-NCI-MOSSAIC

OAK RIDGE National Laboratory

C-HER
CENTRALIZED HEALTH AND ENVIRONMENTAL REPOSITORY

Heidi Hanson, Josh Grant, Maggie Davis,

<u>Metadata and data storage</u>: Metadata is captured with Prefect, which has been integrated with the Centralized Health and Environmental Repository (C-HER) tools. This provides dynamic metadata extraction and ensures robust error handling, making metadata documentation reliable and scalable. Data security is ensured with encryption for data at rest and in transit. We implement strict access controls and authentication measures to safeguard our data.

<u>Raster Data</u>

- Satellite images, aerial photographs, and derived raster products

<u>Vector Data</u>

- Geographic features such as boundaries, roads, and points of interest

<u>Tabular Data</u>

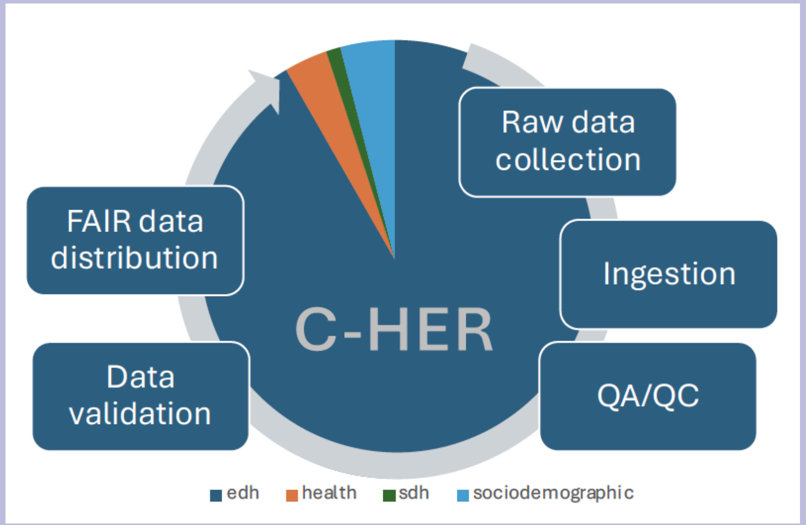
- Meta data, user information, analytical results, modeled data

<u>Large Source Files</u>

- Raw data files sored in MinIO

<u>Data Ingestion Code</u>

- The code used to ingest the data is also stored in the database, allowing us to quickly refresh data and identify issues



The C-HER Ecosystem utilizes Open-Source software such as MinIO, PostgreSQL, Prefect, and Python. The goal of the project is to optimize the reuse of data, thus the C-HER ecosystem adheres to principles of findability, accessibility, interoperability, and reusability (FAIR).

<u>Data Quality Assurance</u>: Regularly validate data using automated tools and manual checks to ensure accuracy and reliability. Document all QA-QC procedures and maintain logs of all data checks.

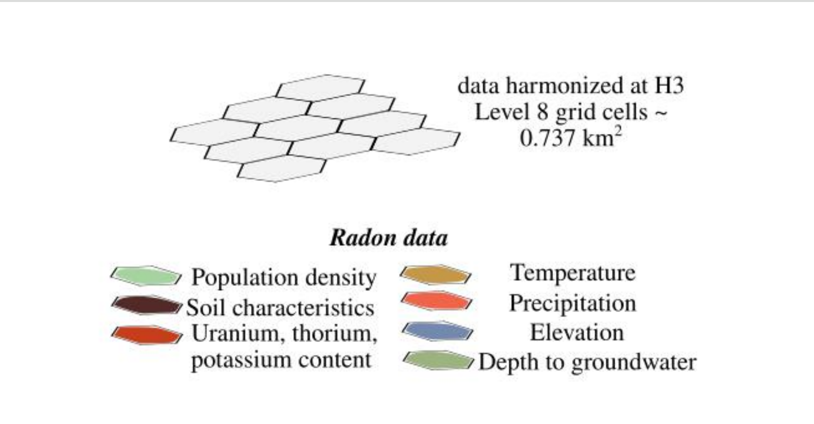
<u>Ethical Concerns</u>: We consider and address privacy concerns for data that is potentially identifiable. We ensure that data usage complies with all applicable laws and ethical guidelines.

OAK RIDGE National Laboratory

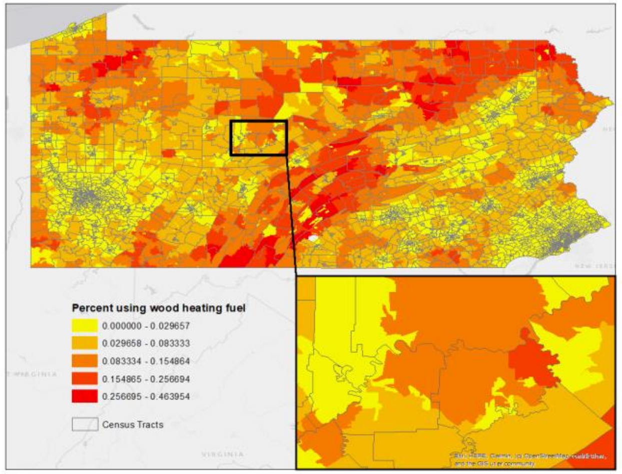
C-HER Serves as the foundation for modeling exposures

Example: Indoor Radon

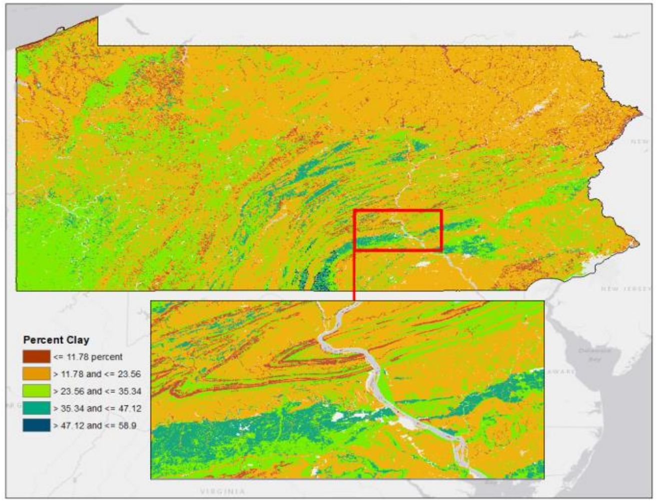
C-HER Protocol for AI Ready Data



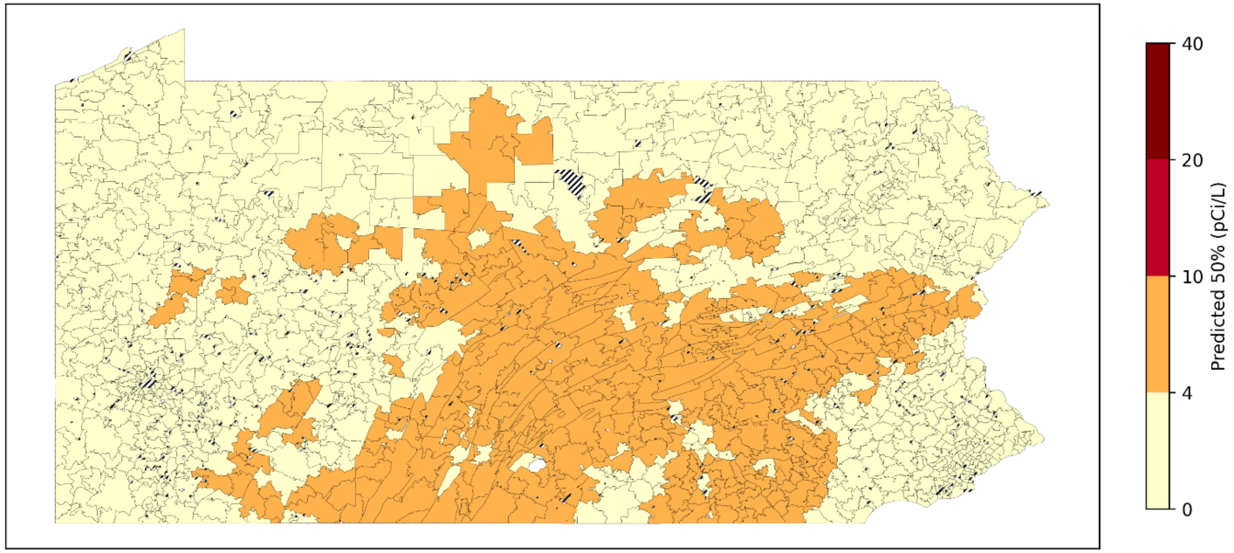
ZCTA level housing Characteristics
Individual test results with ZCTA information



Low spatial resolution geologic, soil, and other data



Predicted 50% by ZCTA



Heechan Lee, Dakotah Maguire, Jeremy Logan, Shaheen Dewji, Greeshma Agastha, Heidi Hanson

25+ years of multimodal exposure data linked to cancer records

MOSSAIC

Modeling Outcomes Using Surveillance Data & Scalable Artificial Intelligence For Cancer

C-HER

CENTRALIZED HEALTH AND ENVIRONMENTAL REPOSITORY

LexisNexis Residential History data

- 11 SEER registries have been linked (3.2 million individuals diagnosed from 2005 - 2022)
- 15 SEER registries should be linked by the end of 2023
- Residential history constructed for Louisiana and expected to be expanded to all linked registries this year
- High quality data from 1995 – 2020
- 83% are geocoded to the point location

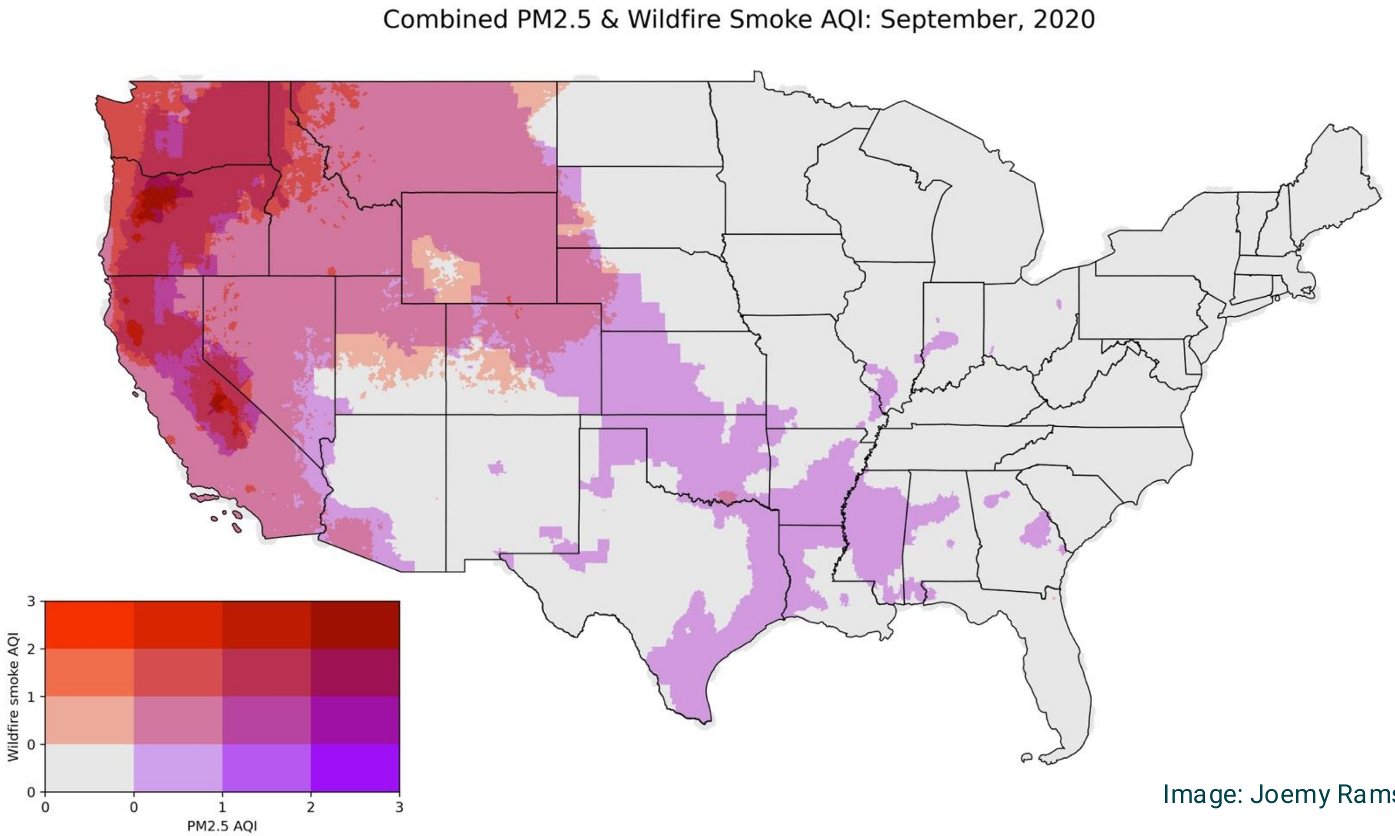
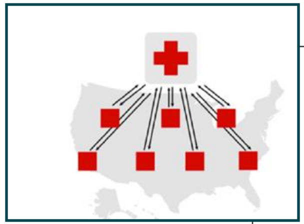


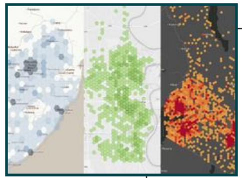
Image: Joemy Ramsay

"Electronic Health Record informed LagranglIan method for precision public Health (EHRLICH)" will be a set of new computational capabilities developed to enable the rapid assimilation of real-world data for digital twins of population level biological threats.

Scalable, Trustworthy and Responsible AI

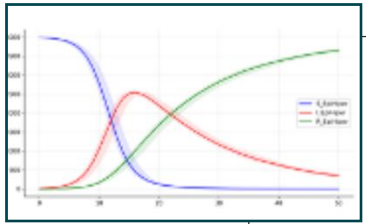


Privacy-Aware Simulation for Clinical AI and Learning (PASCAL)



AI Ready Social and Environmental Determinants of Health Data

High-spatial resolution and high-fidelity synthetic populations that include demographic and biologic characteristics



ENABLE: An agent-based simulation framework

OAK RIDGE National Laboratory · Argonne NATIONAL LABORATORY · Los Alamos NATIONAL LABORATORY · MOREHOUSE SCHOOL OF MEDICINE · USC UNIVERSITY OF SOUTHERN CALIFORNIA · THE UNIVERSITY OF UTAH · THE UNIVERSITY OF ARIZONA · Duke UNIVERSITY · THE UNIVERSITY OF CHICAGO

Interoperable AI Systems at Scale:
Bringing Compute Closer to the Source
Rapid Data Integration while Protecting Private Information

Alternative approach:
synthetic data generation

Our Solution:
- Leverage generative AI
- Minimize leakage of PHI & other identifiable info
- Evaluate for **utility**, **similarity**, and **privacy**

Initial demonstration on SEER data

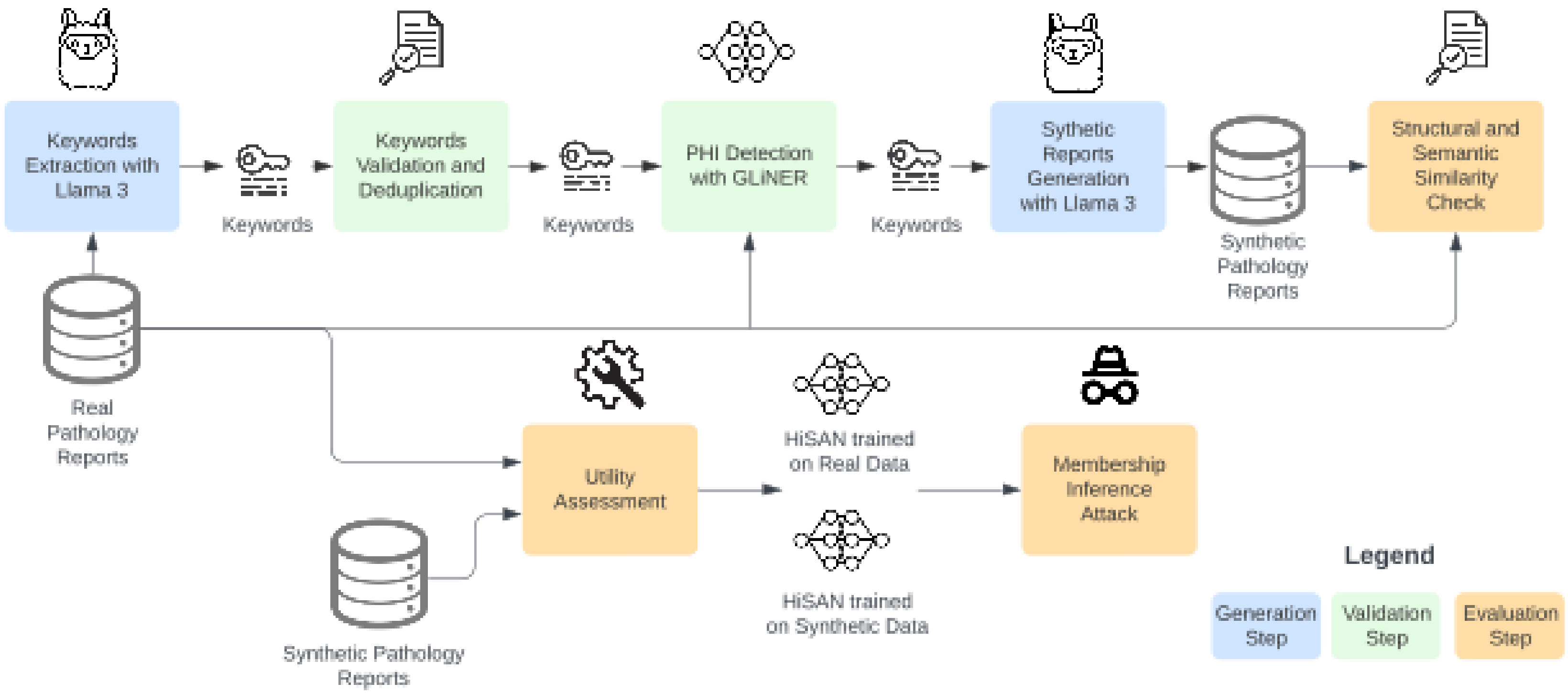
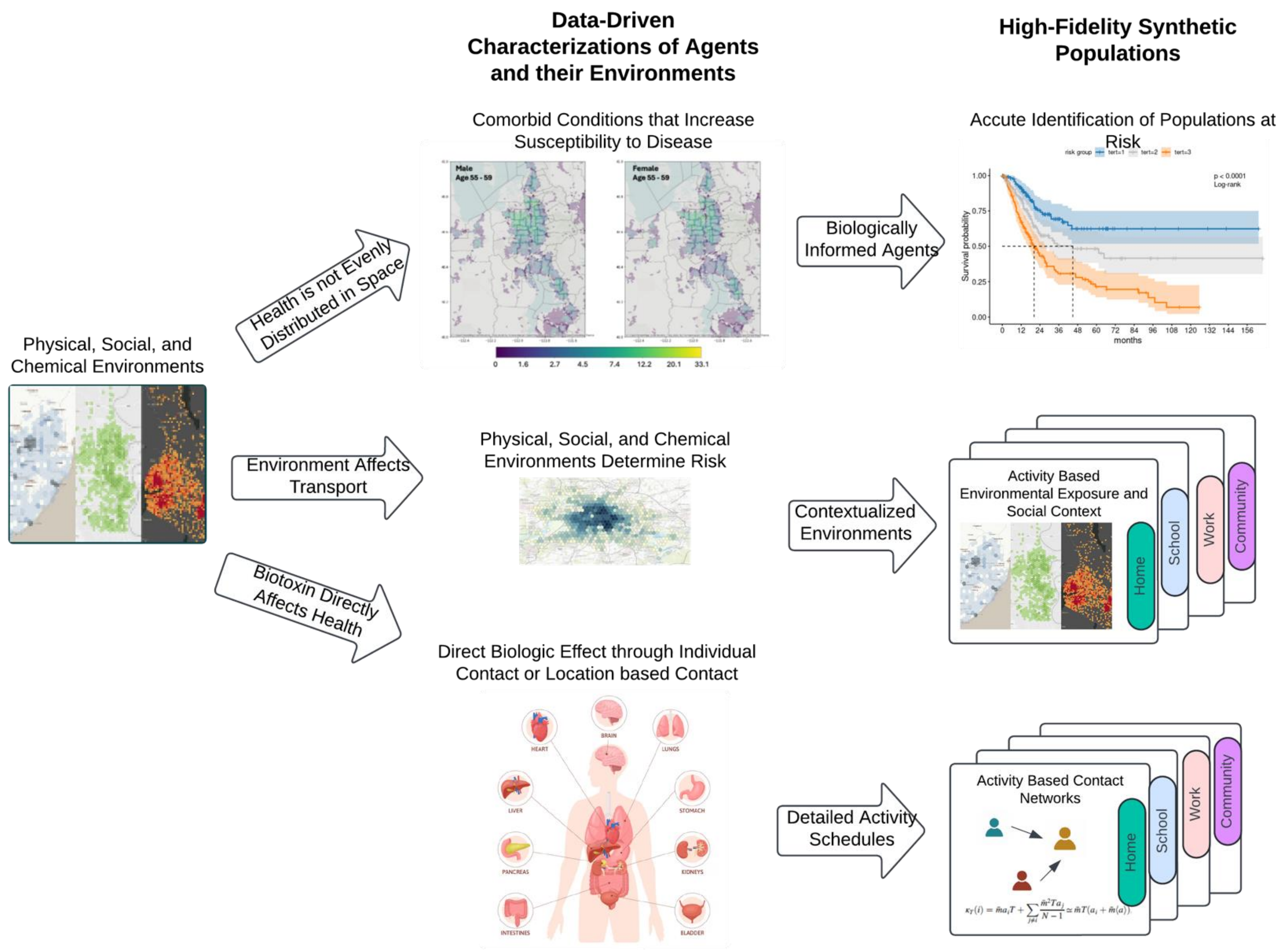


Image credit: Patrycja Krawczuk

EHRLICH tools are designed to be agile and adaptable



OAK RIDGE National Laboratory
USC UNIVERSITY OF SOUTHERN CALIFORNIA
THE UNIVERSITY OF UTAH®

PLACING A PULSE ON POPULATION HEALTH

- Federated Learning
- Synthetic Data Generation
- Privacy Preserving AI

- Identify existing patterns and anomoly detection
- Predict patterns relating to environmental exposures
- High-performance computing for modeling and simulating health outcomes

C-HER
CENTRALIZED HEALTH AND ENVIRONMENTAL REPOSITORY

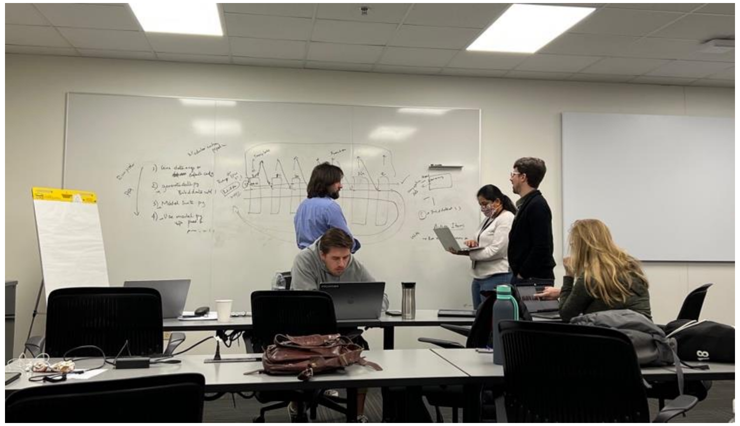
MOSSAIC

National Cancer Institute
Lynne Penberthy (PI)
Elizabeth Hsu (Technical Lead)
Valentina Petkov
Serban Negoita
Ola Adeyemi
Sylkk Ansah
Sarah Bonds

IMS
Linda Coyle
Jennifer Stevens
Scott Depuy
Rusty Sheilds
Gary Beverungen

Los Alamos National Laboratory
Jamaludin Mohd Yusof
Sayera Dhaubhadel
Tanmoy Bhattacharya

Oak Ridge National Laboratory
Heidi Hanson (PI)
John Gounley (Technical Lead)
Shelaine Curd (Program Manager)
Georgia Tourassi
Joe Lake
Adam Spannaus
Dakota Murdock
Zachary Fox
Patrycja Krawczuck
Dakotah Maguire
Jordan Miller
Mayanka Chandra Shekar
Noah Schaefferkoetter
Sajal Dash
Isaac Lyngaas
Abhishek Shivanna
Robert Bridges
Christopher Stanley
Vandy Tombs
Christoph Metzner



EHRLICH

Oak Ridge National Laboratory
Heidi Hanson (PI)
John Gounley (Technical Lead)
Shelaine Curd (Program Manager)
Adam Spannaus
Zachary Fox
Patrycja Krawczuck
Dakotah Maguire
Mayanka Chandra Shekar
Robert Bridges
Christopher Stanley
Vandy Tombs
Sudip Seal
James Nutaro
Sifat Moon
Christoph Metzner

Los Alamos National Laboratory
Jamal Mohd-Yusof
Cristina Garcia Cardona

Argonne National Laboratory
Rick Stevens
Thomas Brettin

University of Utah
James VanDerslice
Joemy Ramsay

University of Arizona
Nirav Merchant
Ravi Tandon

University of Southern California
Rima Habre

Duke
Amanda Randles

University of Chicago/Morehouse School of Medicine
Lilly Immergluck