MOSSAIC

Modeling Outcomes Using Surveillance Data &
Scalable Artificial Intelligence For Cancer

Principal Investigators

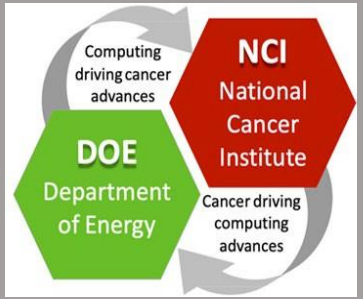
Dr. Lynne Penberthy
Associate Director of the Surveillance Research Program, NCI

Dr. Heidi Hanson
Group Leader, Biostatistics and Biomedical Informatics, ORNL

Technical Leads

Dr. Betsy Hsu
Chief, Surveillance Informatics Branch, NCI

Dr. John Gounley
Group Leader, Scalable Biomedical Modeling, ORNL



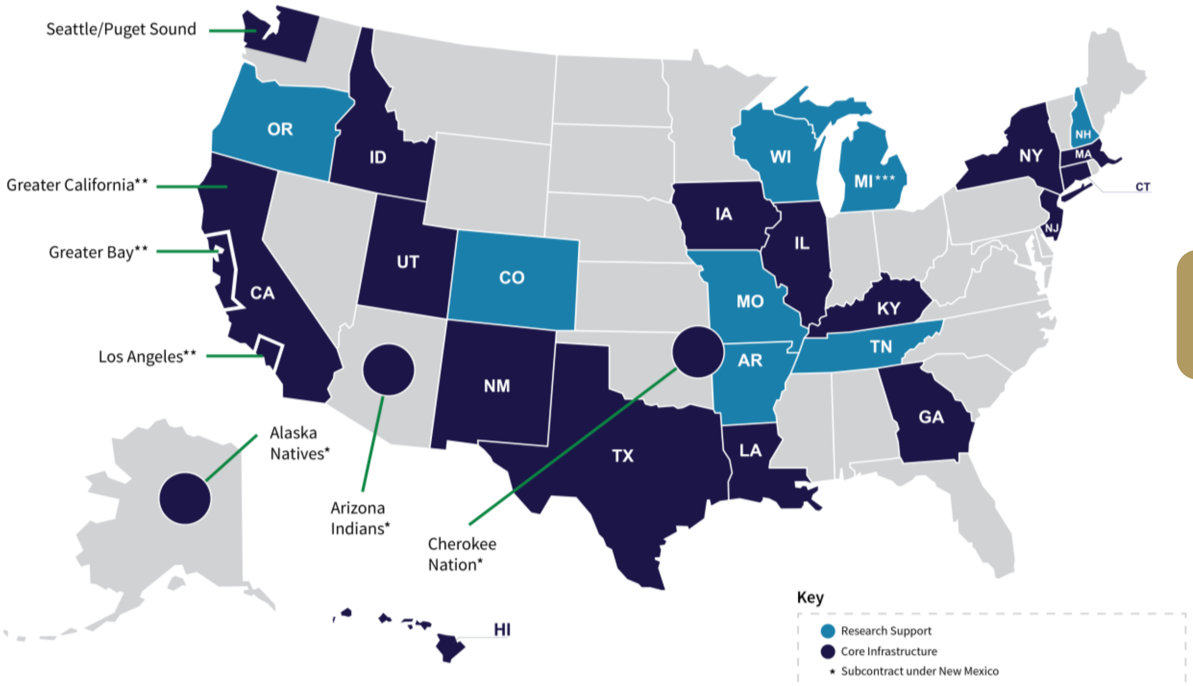
Computing driving cancer advances

NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

Surveillance Epidemiology and End Results (SEER)



Dark Blue represents ***Core Registries*** (Reporting data)

Light Blue represents ***Research Support Registries-*** participate in special projects

SEER authorized by National Cancer Act of 1971 with a mission to support research

- Collects diagnosis, treatment, and outcomes of cancer since 1973
- Provide baseline data on U.S. cancer incidence and survival
- De-identified data submitted to NCI

18 Population-based registries representing ~48% of the U.S. population

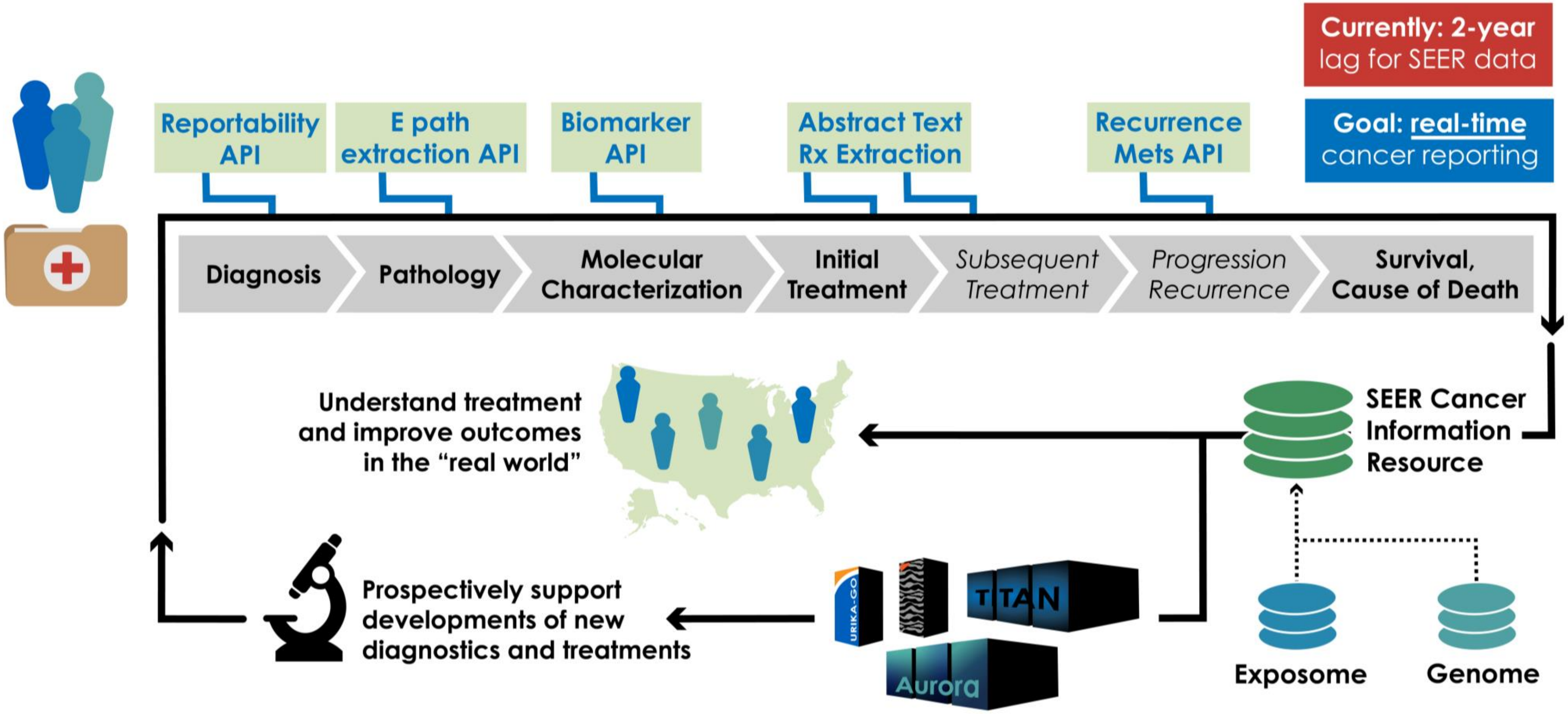
- >850,000 incident cases annually
- Approximately 85% of cases with real time electronic pathology reporting
- A single abstract is consolidated from ~3.6 records/case
- Complex and heterogenous data, including new data sources and real time data feeds
- Largely manual process for screening and abstraction

SEER*Data Management System (SEER*DMS)

- Enables centralized efficient linkages
- Simultaneous implementation of new tools
- Ability develop a common, optimized production workflow

The MOSSAIC Challenge

Translational AI for better cancer surveillance and ultimately better cancer care.

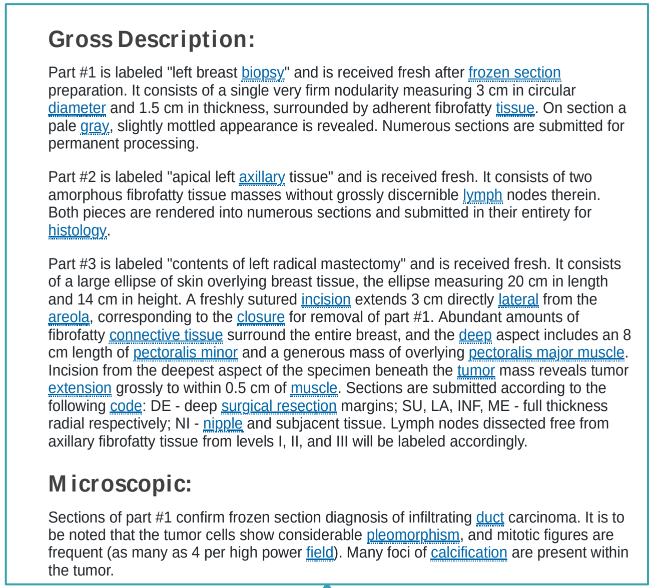


Currently: 2-year lag for SEER data
Goal: real-time cancer reporting
Reportability API
E path extraction API
Biomarker API
Abstract Text Rx Extraction
Recurrence Mets API
Diagnosis
Pathology
Molecular Characterization
Initial Treatment
Subsequent Treatment
Progression Recurrence
Survival, Cause of Death
Understand treatment and improve outcomes in the "real world"
SEER Cancer Information Resource
Prospectively support developments of new diagnostics and treatments
URIKA-GO
TITAN
Aurora
Exposome
Genome

Oak Ridge National Laboratory NIH National Cancer Institute

Real World Evidence

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



Gross Description:

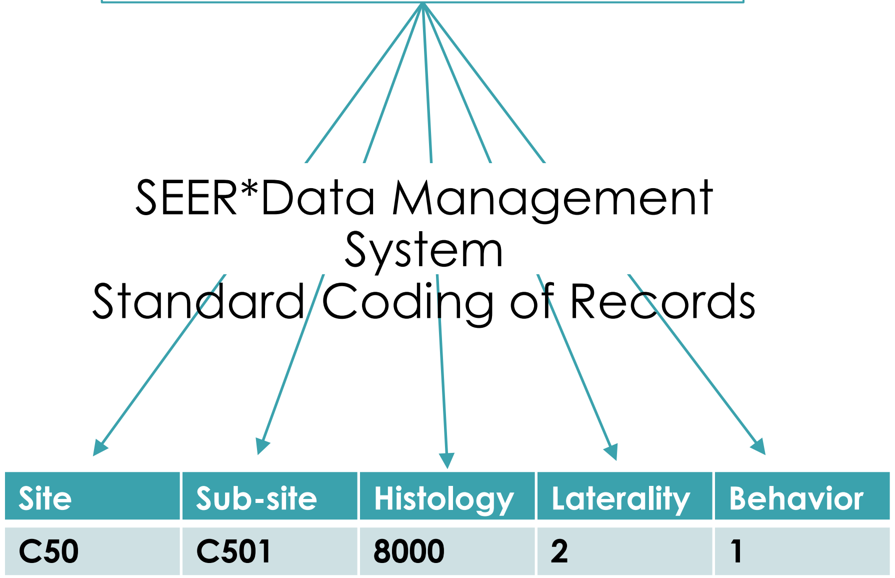
Part #1 is labeled "left breast biopsy" and is received fresh after frozen section preparation. It consists of a single very firm nodularity measuring 3 cm in circular diameter and 1.5 cm in thickness, surrounded by adherent fibrofatty tissue. On section a pale gray, slightly mottled appearance is revealed. Numerous sections are submitted for permanent processing.

Part #2 is labeled "apical left axillary tissue" and is received fresh. It consists of two amorphous fibrofatty tissue masses without grossly discernible lymph nodes therein. Both pieces are rendered into numerous sections and submitted in their entirety for histology.

Part #3 is labeled "contents of left radical mastectomy" and is received fresh. It consists of a large ellipse of skin overlying breast tissue, the ellipse measuring 20 cm in length and 14 cm in height. A freshly sutured incision extends 3 cm directly lateral from the areola, corresponding to the closure for removal of part #1. Abundant amounts of fibrofatty connective tissue surround the entire breast, and the deep aspect includes an 8 cm length of pectoralis minor and a generous mass of overlying pectoralis major muscle. Incision from the deepest aspect of the specimen beneath the tumor mass reveals tumor extension grossly to within 0.5 cm of muscle. Sections are submitted according to the following code: DE - deep surgical resection margins; SU, LA, INF, ME - full thickness radial respectively; NI - nipple and subjacent tissue. Lymph nodes dissected free from axillary fibrofatty tissue from levels I, II, and III will be labeled accordingly.

Microscopic:

Sections of part #1 confirm frozen section diagnosis of infiltrating duct carcinoma. It is to be noted that the tumor cells show considerable pleomorphism, and mitotic figures are frequent (as many as 4 per high power field). Many foci of calcification are present within the tumor.



SEER*Data Management System
Standard Coding of Records

Site	Sub-site	Histology	Laterality	Behavior
C50	C501	8000	2	1

Auto-Extraction from Pathology Reports:

Accuracy: **Auto-coding of 23-27% of path reports** 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.

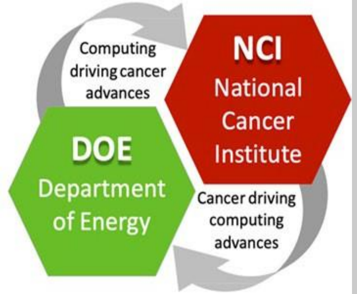
Auto-coding **performance can be easily tuned**

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

- **Total 19 registries** (additional 7 since March 2022; ~39% of US population)
- **Default** as part of any new DMS installation, regardless of SEER affiliation
- **1-3 new registries** anticipated in 2024/2025
- Testing phase with the Veteran’s Health Administration
- Leveraged by registries to increase operational efficiency

Approximately 14,000 person hours/year saved

Democratizing AI for the Oncology Community



Computing driving cancer advances

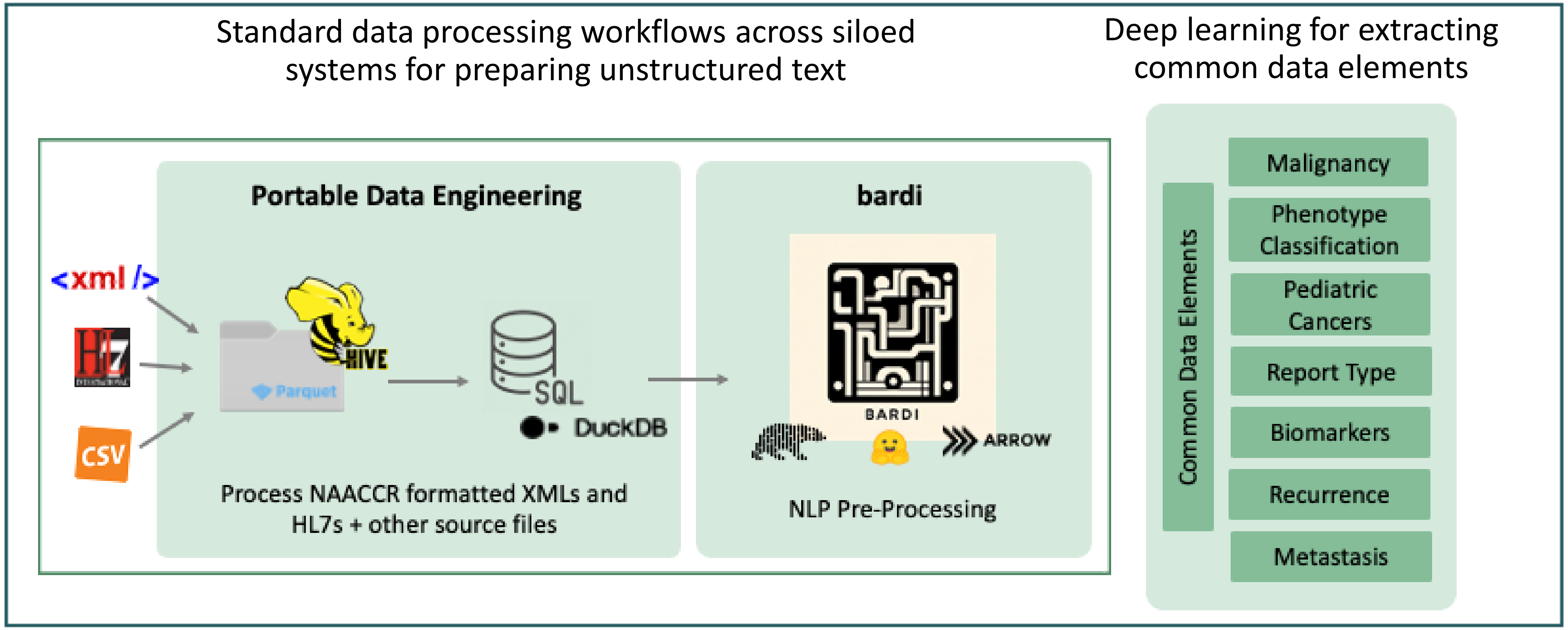
NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

FrESCO
Innovative real-world solutions for automatic case classification

Development of scalable natural language processing and machine learning tools
- Deep text comprehension of unstructured clinical text
- Accurate, automated capture of reportable cancer surveillance data elements



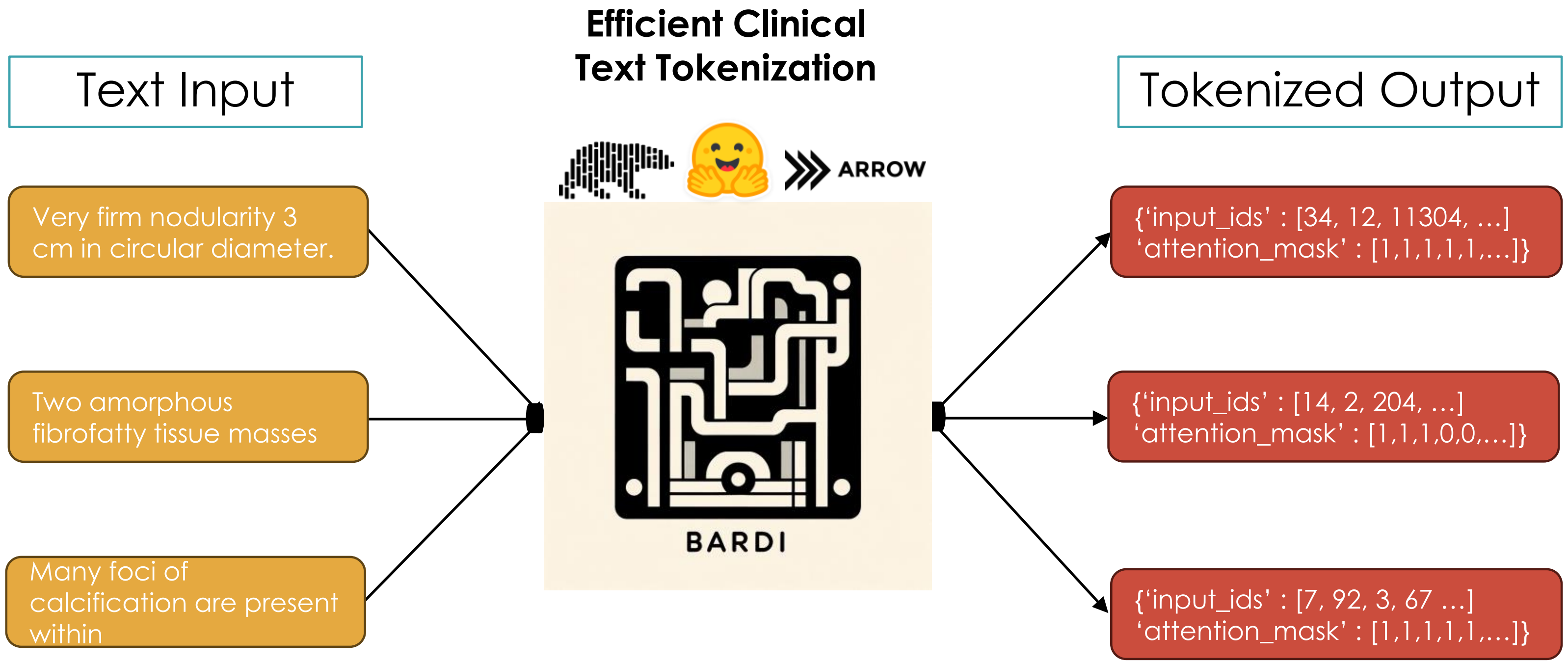
Standard data processing workflows across siloed systems for preparing unstructured text
Deep learning for extracting common data elements
Portable Data Engineering
<xml />
HL7
CSV
Parquet
HIVE
SQL
DuckDB
Process NAACCR formatted XMLs and HL7s + other source files
bardi
BARDI
ARROW
NLP Pre-Processing
Common Data Elements
Malignancy
Phenotype Classification
Pediatric Cancers
Report Type
Biomarkers
Recurrence
Metastasis

OAK RIDGE National Laboratory NIH NATIONAL CANCER INSTITUTE FrESCO https://github.com/DOE-NCI-MOSSAIC

BARDI for AI readiness of clinical data

British Columbia Cancer Registry

- BARDI: our AI-readiness package for clinical data
- Encodes text data into numeric representations ("tokenization") that can be used in downstream deep learning models.
- Supports multiple methods of tokenization for maximum flexibility and interoperability.
- Now ready for research in de-identification and synthetic data generation



Efficient Clinical Text Tokenization

Text Input

Very firm nodularity 3 cm in circular diameter.

Two amorphous fibrofatty tissue masses

Many foci of calcification are present within

Tokenized Output

{'input_ids' : [34, 12, 11304, …] 'attention_mask' : [1,1,1,1,1,…]}

{'input_ids' : [14, 2, 204, …] 'attention_mask' : [1,1,1,0,0,…]}

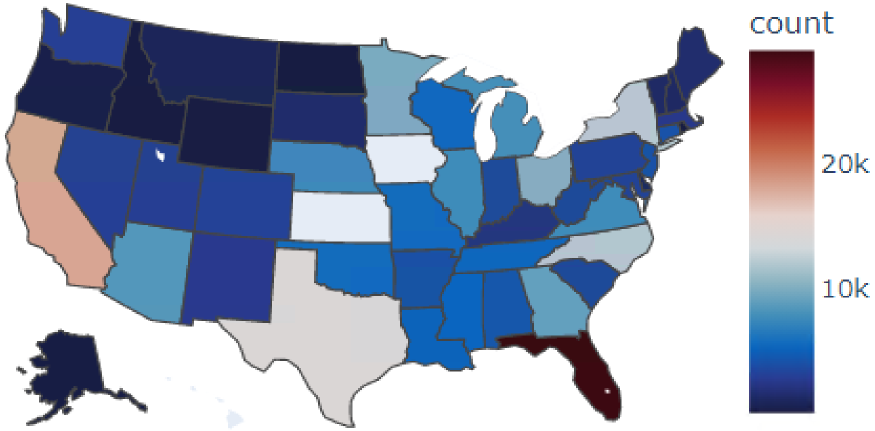
{'input_ids' : [7, 92, 3, 67 …] 'attention_mask' : [1,1,1,1,1,…]}

Framework for Exploring Scalable Computational Oncology (FrESCO)

Near Real-Time Identification of Veteran's with Cancer

- Modular deep-learning natural language processing library for extracting information from clinical documents
- FrESCO has been used to train models with EHR notes and electronic pathology records

Data 2006 - 2018
N=1,081,161



Selected Results: Task	FrESCO		
	F1	Prec.	Rec.
Reportablity	0.95	0.94	0.95
Breast	0.99	0.99	0.99
Cervical	0.88	0.88	0.88
Uterine	0.72	0.73	0.71
Overall Accuracy	Site: 0.92; Histology: 0.79		

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

Improving Autocoding Accuracy

Increase Autocoding rates to 50% by the end of CY24

HiSAN Histology F1Macro = 0.33

- **Path-BigBird:** A large-language model trained from scratch using SEER reports (Histology F1Macro = 0.37)
- **Phrase-Level Attention:** Phrase level attention focuses on phrase level context, such as *non-small cell carcinoma* (Histology F1Macro = 0.41)
- **Mixture of experts:** Develop age/sex-specific models to boost performance on minority classes.
- **Hierarchy:** Create hierarchical model to reflect hierarchical nature of clinical classification tasks
- **Heterogenous ensembling methods:** Assess feasibility of an ensemble approach combining DeepPhe and MOSSAIC model predictions.

Reportability:
Near Real-Time Case Identification

Computing driving cancer advances

NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

Real-time Cancer Case Identification
Reportability API

- Cancer registries only authorized to hold information on cancer cases
- Mix of methods to filter disease-reportable reports from facilities before providing to registries
  - Large number of false positives received by registries
  - Manual review still often required to confirm ”reportable” case
- Essential for goals of:
  - Rapid case ascertainment for research studies
  - Near real-time incidence reporting
- Validating performance across broader set of registries

Real-time Cancer Case Identification

Reportability API

- Model trained with **3,796,593** pathology reports from Seattle
- Reports span 2015 - 2022
- Hierarchical Self-Attention Model (HiSAN)
- A central repository of high-quality SEER data allowed for the development and rapid deployment of an automatic classification algorithm

Accuracy and Macro F1 Score for Seattle Test Set

	Accuracy	Macro
HiSAN: Random	0.9978	0.9969
HiSAN: Year	0.9965	0.9952

Note: Multiple versions of the trained model were tested. The "Random" split model was trained on a 70% train, 15% validation, and 15% test set. We ran a series of tests to test for temporal shifts in performance of the model. We found no evidence of temporal drift. The "Year" results displayed here show the results from a model trained on data from 2015 – 2020 and tested on data from 2021 and 2022.

Near Real Time Recurrence and Metastasis

Computing driving cancer advances

NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

Near Real-Time Recurrence and Metastasis Reporting

Challenge:

- No standard for population level data collection of recurrent metastatic disease – making it difficult to assess risk

Solution:

- MOSSAIC development of algorithm to classify pathology reports by metastasis
- Key to creation of recurrent metastatic disease dataset within SEER
  - Development of methods that incorporate multimodal data from claims data and hospital reported sources

Results: HiSAN model trained with Pathology Report Data

	Prec	Recall	F1
No mets	0.94	0.94	0.94
Mets	0.81	0.87	0.84
Uncertain	0.44	0.19	0.27
Accuracy			**0.90**

Test set of 5389 reports

We can achieve accuracy of 0.94 with a 9% abstention rate.

Real World Use Case: Privacy Preserving Federated Learning

Computing driving cancer advances

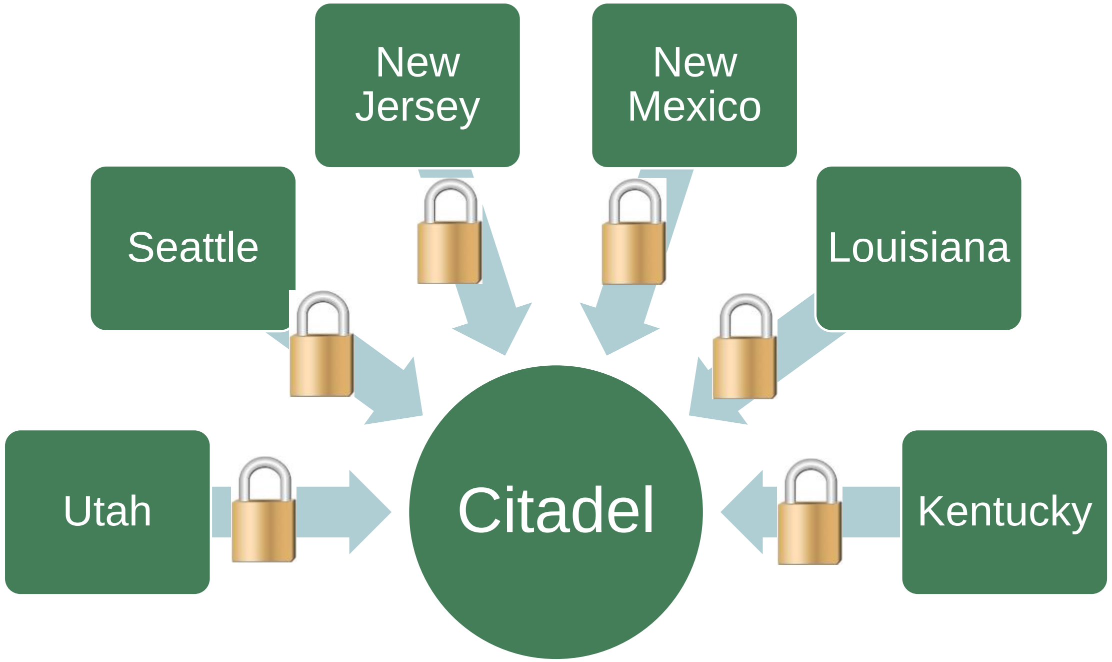
NCI National Cancer Institute

DOE Department of Energy

Cancer driving computing advances

Near Real-Time Health Data Analytics
Privacy Aware Federated Learning

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

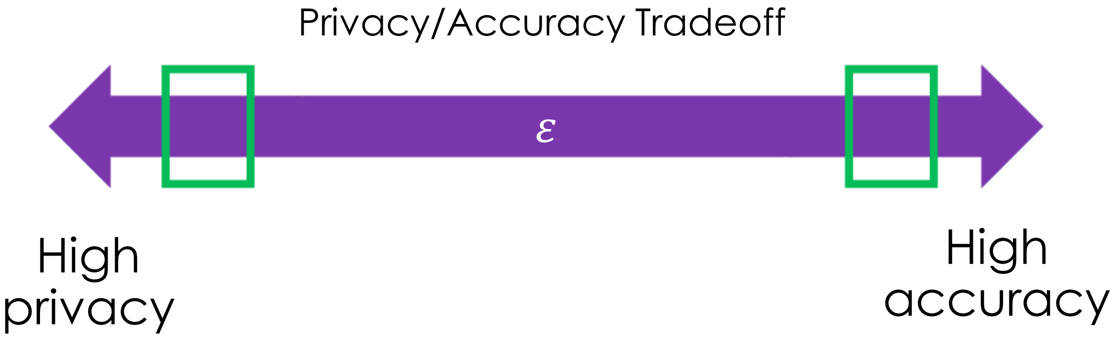


Privacy Preserving Hackathon: June 2024

Our current implementation

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

Differential Privacy



Innovation: SEER data make us completely innovative in this space. We can design and test solutions for real-world application at population scale and potentially leverage for international collaborations

Linking the External Exposome

Computing driving cancer advances

NCI National Cancer Institute

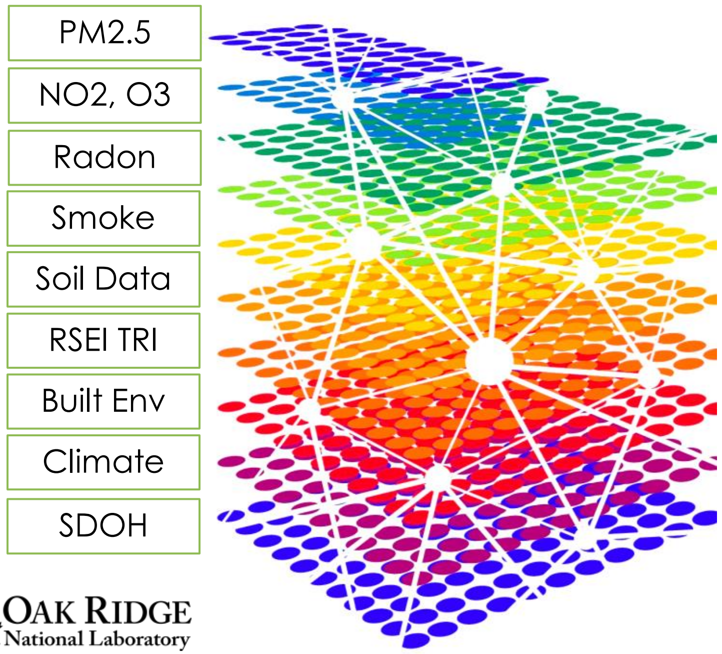
DOE Department of Energy

Cancer driving computing advances

Measuring the External Exposome

Centralized-Health and Environment Repository (C-HER)

- SEER has captured a patient’s residential history from diagnosis year 2005+ via Lexis Nexis linkage
- Residential history from 1995 – Present
- First linkages: Air pollution, toxic releases, and indoor radon
- Over 70 Social and Environmental Measures are in the C-HER database.



PM2.5

NO2, O3

Radon

Smoke

Soil Data

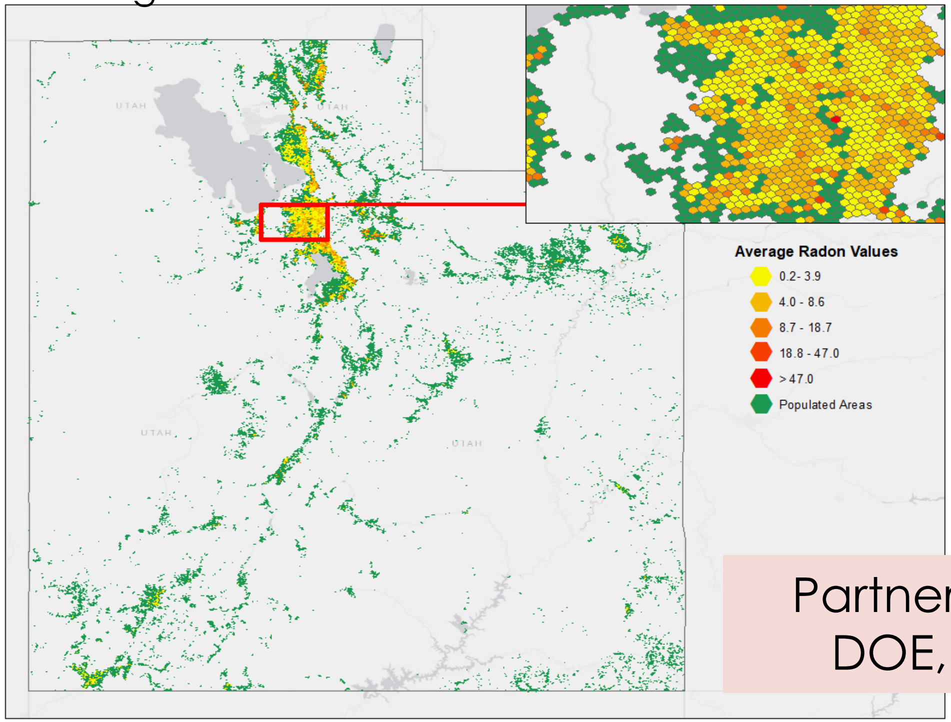
RSEI TRI

Built Env

Climate

SDOH

Average Radon Value: 35k home radon test kits



Partners: VA, DOE, NCI

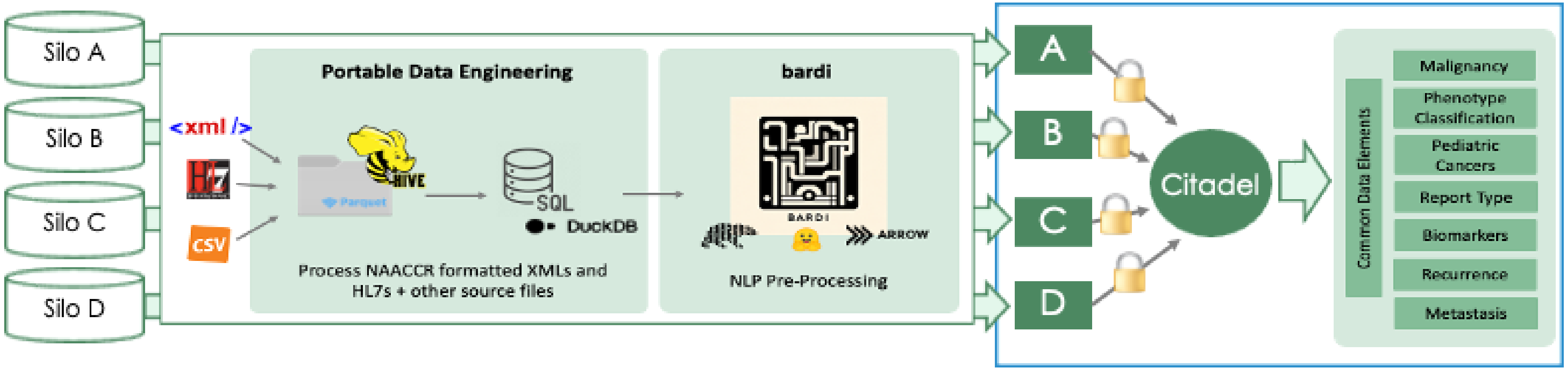
MOSSAIC: Translational AI for better cancer surveillance and ultimately better cancer care.

Real World Solution

FrESCO

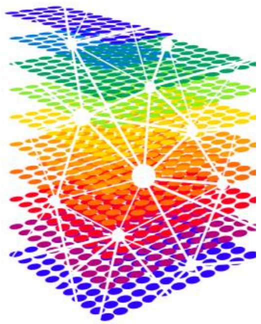
https://github.com/DOE-NCI-MOSSAIC

Real World AI Solutions for Population Surveillance

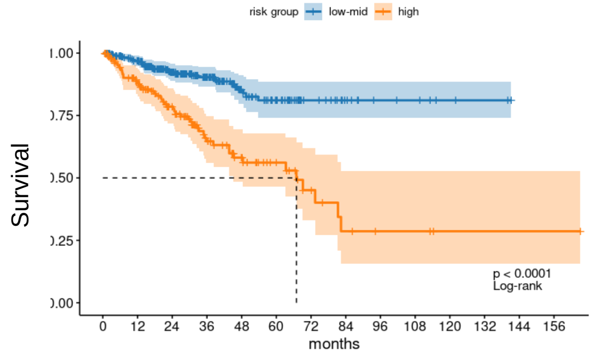


Yield Real World Solutions for Precision Health

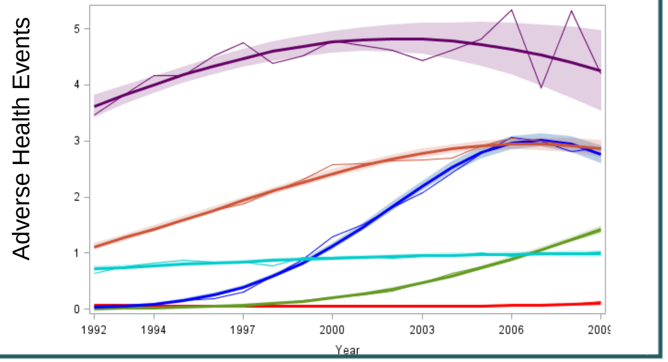
Risk Factors



Prognosis



Patient Trajectories


Adverse Health Events
Year