

AI in Medicine: Near real-time disease surveillance at population scale

Heidi A. Hanson, PhD
Senior Scientist, Advanced Computing
for Health Sciences | Computational Sciences
and Engineering Division

Oak Ridge, TN
September 04, 2024

ORNL is managed by UT-Battelle LLC for the US Department of Energy

Acknowledgements

- This material is based upon work supported by the following:
 - MOSSAIC: UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy.
 - MOSSAIC: National Cancer Institute (NCI) – DOE Collaboration under the Cancer MoonshotSM initiative.
 - The U.S. Department of Energy, Office of Science, Office of Science Advanced Scientific Computing Research (ASCR) as part of Dr. Margaret Lentz's Biopreparedness Research Virtual Environment (BRaVE) initiative.



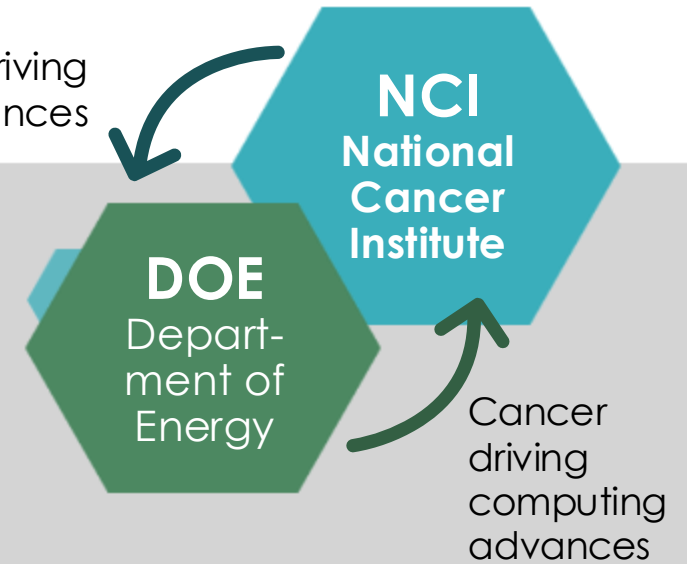
M O S S A I C

Modeling Outcomes Using Surveillance Data &
Scalable Artificial Intelligence For Cancer

Investors and technical leads

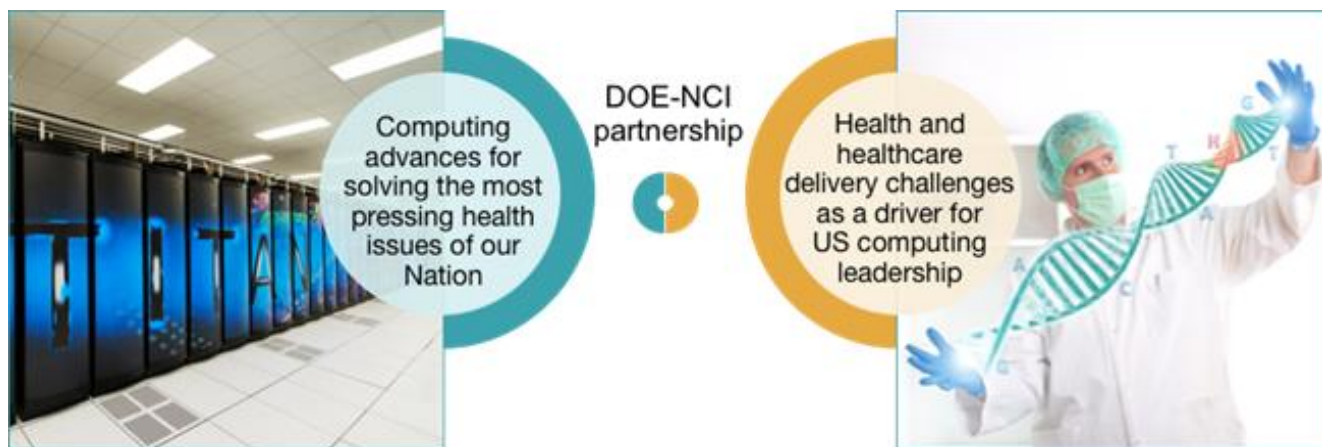
- Principal Investigators
 - Dr. Lynne Penberthy, Associate Director of the Surveillance Research Program, NCI
 - Dr. Heidi Hanson, Group Leader, Biostatistics and Biomedical Informatics, ORNL
- Technical Leads
 - Dr. Elizabeth Hsu, Chief, Surveillance Informatics Branch, NCI
 - Dr. John Gounley, Group Leader, Scalable Biomedical Modeling, ORNL

Computing driving cancer advances



Cancer driving computing advances

Joint Design of Advanced Computing for Cancer (JDACS4C): DOE-NCI Partnership



Enabling the most challenging deep learning problems in cancer research to run on the most capable supercomputers in the DOE

Pilot 3: Dr. Georgia Tourassi and Dr. Lynne Penberthy

NCI-DOE Collaborations

- Foster a growing predictive oncology community
- Provide open access to FAIR AI/ML resources, including datasets and computational models

NCI-DOE Projects

MOSSAIC

Modeling Outcomes Using Surveillance Data and Scalable Artificial Intelligence for Cancer

ADMIRRAL

AI-Driven Multiscale Investigation of the RAS/RAF Activation Lifecycle

IMPROVE

Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

ATOM

Accelerating Therapeutics for Opportunities in Medicine



Infrastructure

CANDLE

CANcer Distributed Learning Environment

MoDaC

NCI Predictive Oncology Model and Data Clearinghouse



MOSSAIC

Innovating AI with Real World Data

A Unique Resource

- Over 9 million pathology reports from 7 SEER registries, along with radiology reports
- Tabular data about cancer diagnoses linked at the patient level
- Residential histories on over 11 million patients
- 8 years experience working with deep learning methods for autocoding RWD
- Two production-level APIs integrated into the IMS SEER*DMS workflows for all central cancer registries
- Scalable HPC workflows for ingesting, preprocessing, and tokenizing electronic health data.

Innovative Edge

- Novel architectural improvements, hierarchical attention mechanisms and deformable phrase level attention mechanisms.
- Multimodal models for identification of cancer recurrence and metastasis
- New methods for bias identification and mitigation for deep learning architectures
- Differentially private federated learning methods for cancer research
- 51 scientific publications exploring machine learning, deep learning, and AI for oncology

Why is there a two-year lag in cancer reporting?

More than 90% of all cancers are histologically confirmed.

Pathology Report

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.

<https://training.seer.cancer.gov/casefinding/sources/pathology.html>

Abstracted information

Answer: Operative Report Example 2

07/20/91. Path. Rpt. #S91-1700 (L) Rad. mast.:

4 cm tumor infiltrates deep fatty tissue; no invasion of muscle, nipple, or lactiferous sinuses. Metas, carcinoma (L) axillary lymph node, level I. Dx: Infiltrating duct carcinoma, (L) breast.

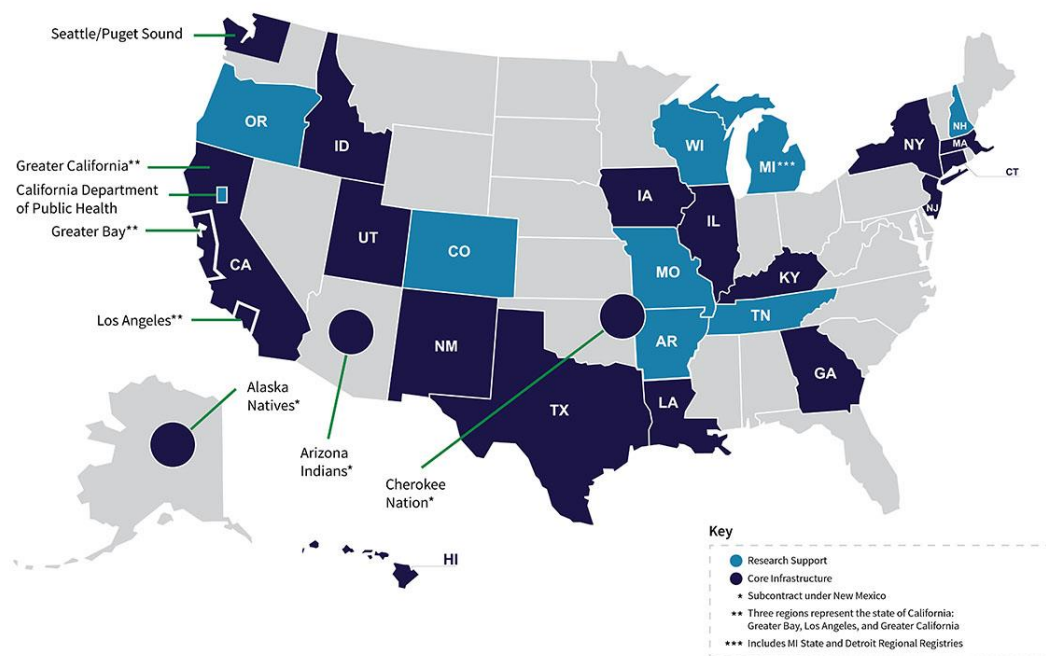
[Close Window](#)

SEER*Data Management System Standard Coding of Records

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

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/>

Auto-Extraction from Pathology Reports:

OncolE Accuracy: Auto-coding of ~32% of path reports (N ~ 270,000) with > 98% accuracy across all data elements

Predictions aid with decision support when records are not autocoded

Abstention rates can be autotuned to fit the need of the research project.

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

- Began testing in Aug 2021
- Total 22 registries (18 SEER + 4 non-SEER)
- Default as part of any new Data Management System installation, regardless of SEER affiliation

OncoIE*

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

OncoID*

- Reportability = 2 categories

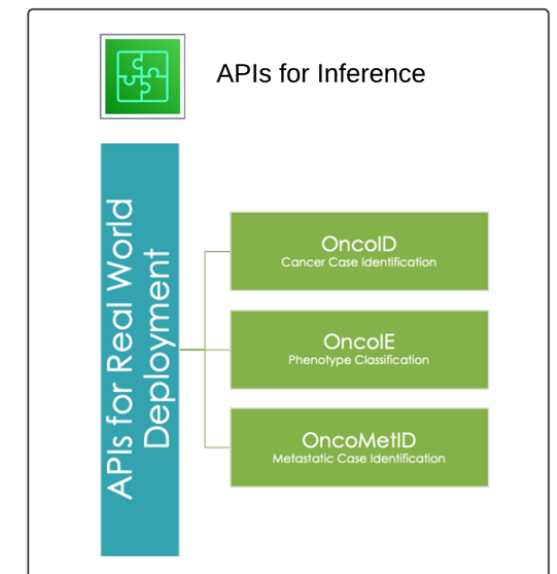
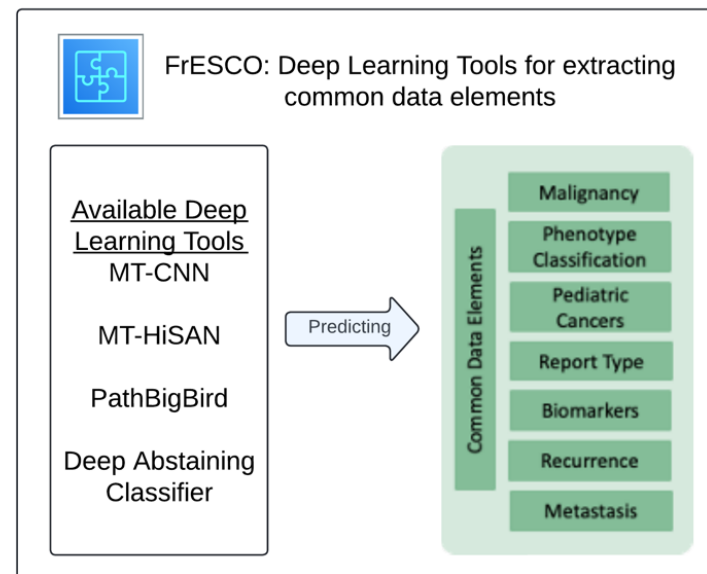
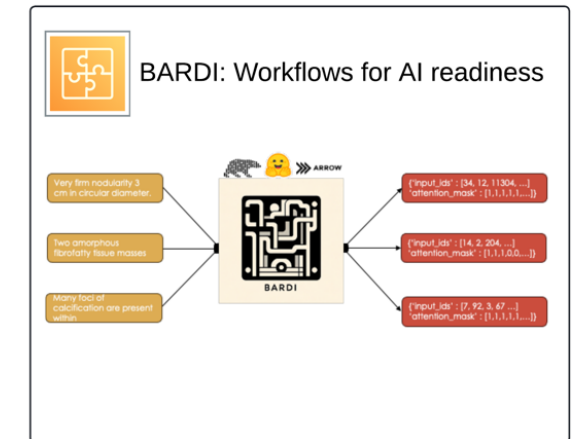
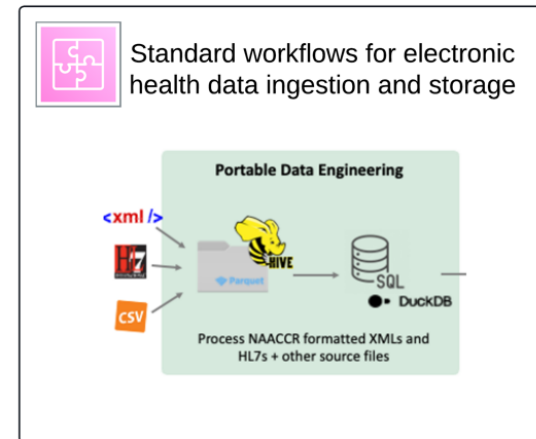
OncoMetsID

- Recurrence = 4 categories
- Metastasis = 2 categories
- Metastatic Site = 4 categories

*Currently in Production in IMS DMS Systems



MOSSAIC Modular tools for AI readiness



Electronic Health Data

Unstructured Clinical Text

- Louisiana, Kentucky, Utah, Seattle, New Jersey, New Mexico, California
- Clinical notes that range from 10 to >10k words. They are 4k words on average.
~9M reports

Tabular sociodemographic, diagnostic, and treatment data

~12M records linked to ~3M diagnoses

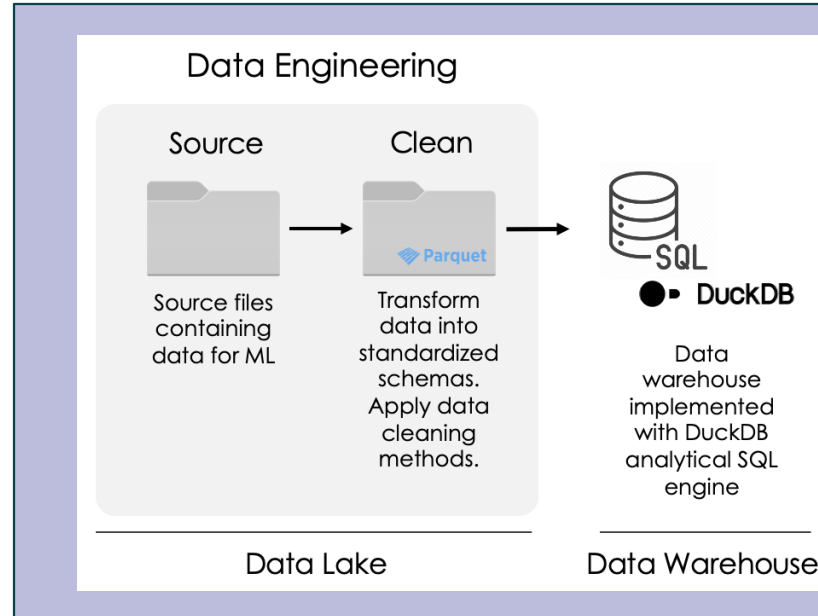
Residential history data

~ 2.5M records with expected growth to 7.5M by the end of CY24

Synthetic Data

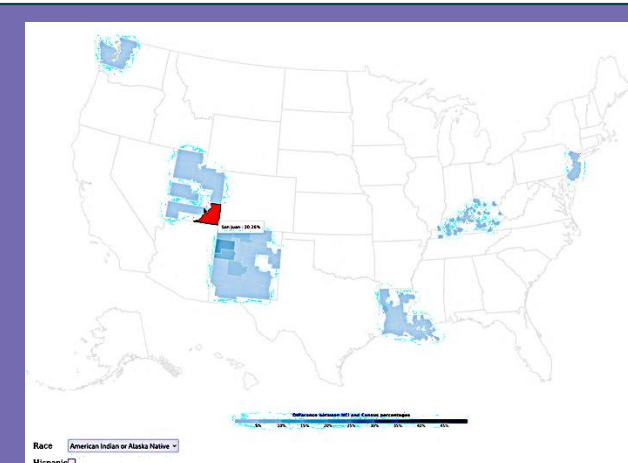
- Generation of privacy preserved synthetic datasets.

Metadata and data storage: Metadata captured during ingestion enabling precise tracking and identification of data origins. Data available in the "clean" layer is accessible through a SQL data warehouse implemented with the DuckDB analytical database management system and query engine.



AI Ready Workflows

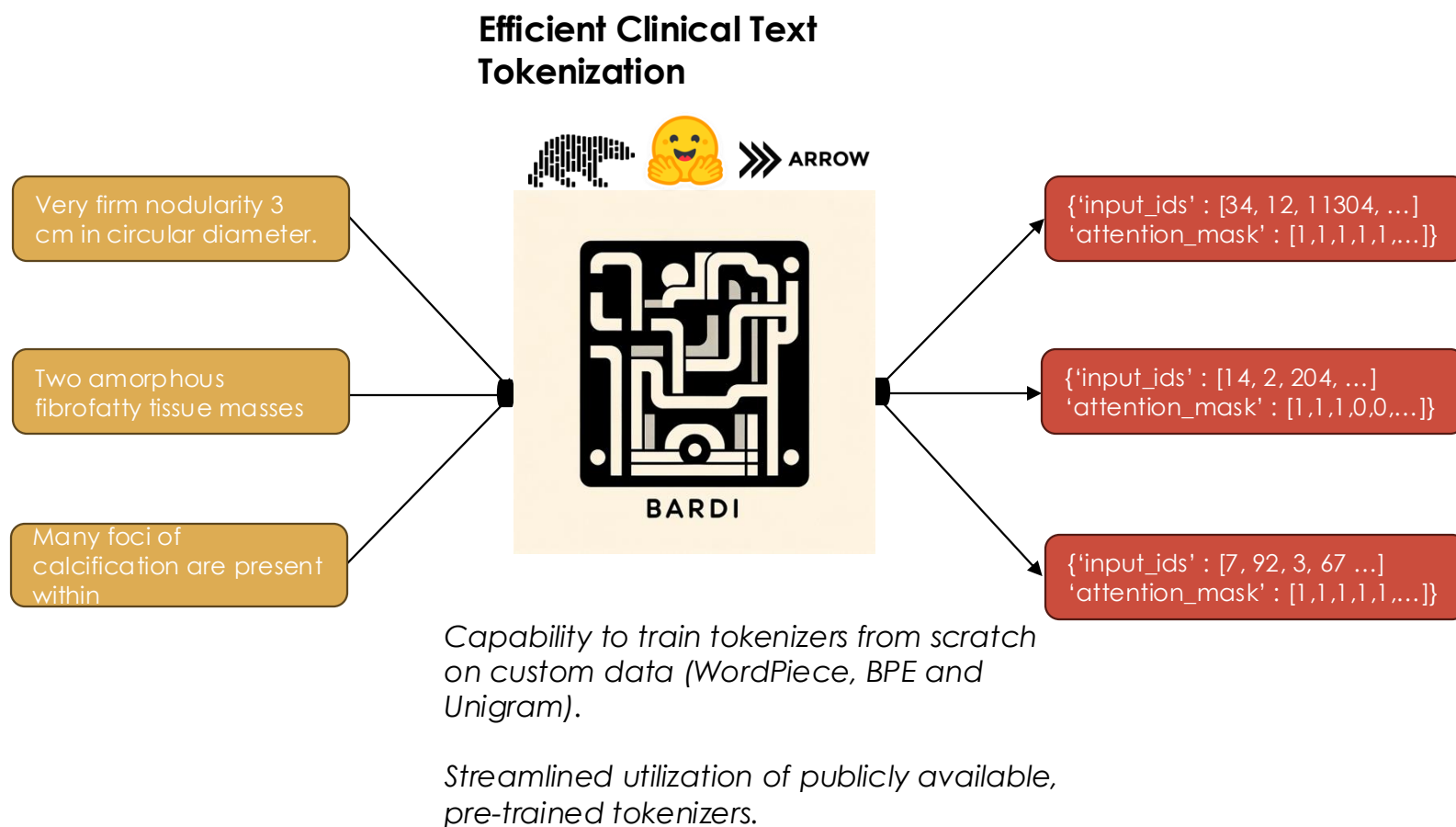
The diagram shows a workflow starting with a data source icon, followed by a smiley face icon, and then an arrow pointing to a 'BARDI' model icon. Below the icon is the text 'BARDI'. The text below the icon reads: 'Capability to train tokenizers from scratch on custom data (WordPiece, BPE and Unigram)'.



Tools for assessing data bias and sources of missingness: Dynamic visualization tools that allow analysts to assess potential sources of data bias and missingness.

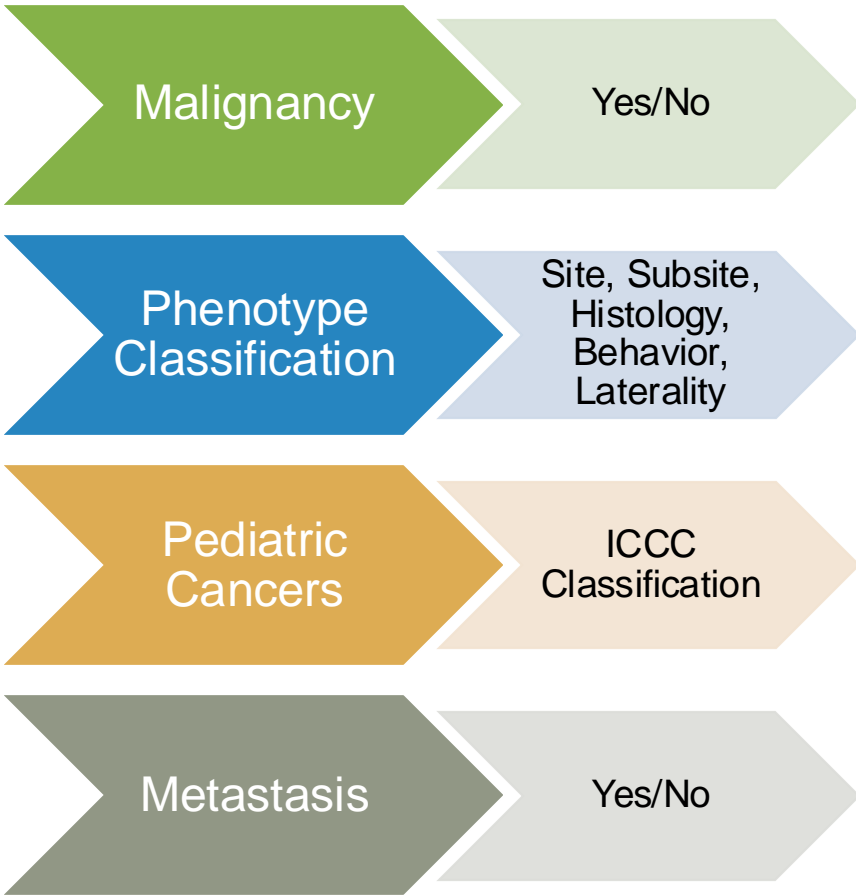
BARDI for AI readiness of clinical data

- BARDI: our AI-readiness package for clinical data
- Increasingly, de-identification and generative AI workflows operate on token-level data
- To better support these workflows, we added built-in tokenization support in BARDI
- Uses standard libraries such as Hugging Face 🤗
- Now ready for NAIRR research in de-identification and synthetic data generation

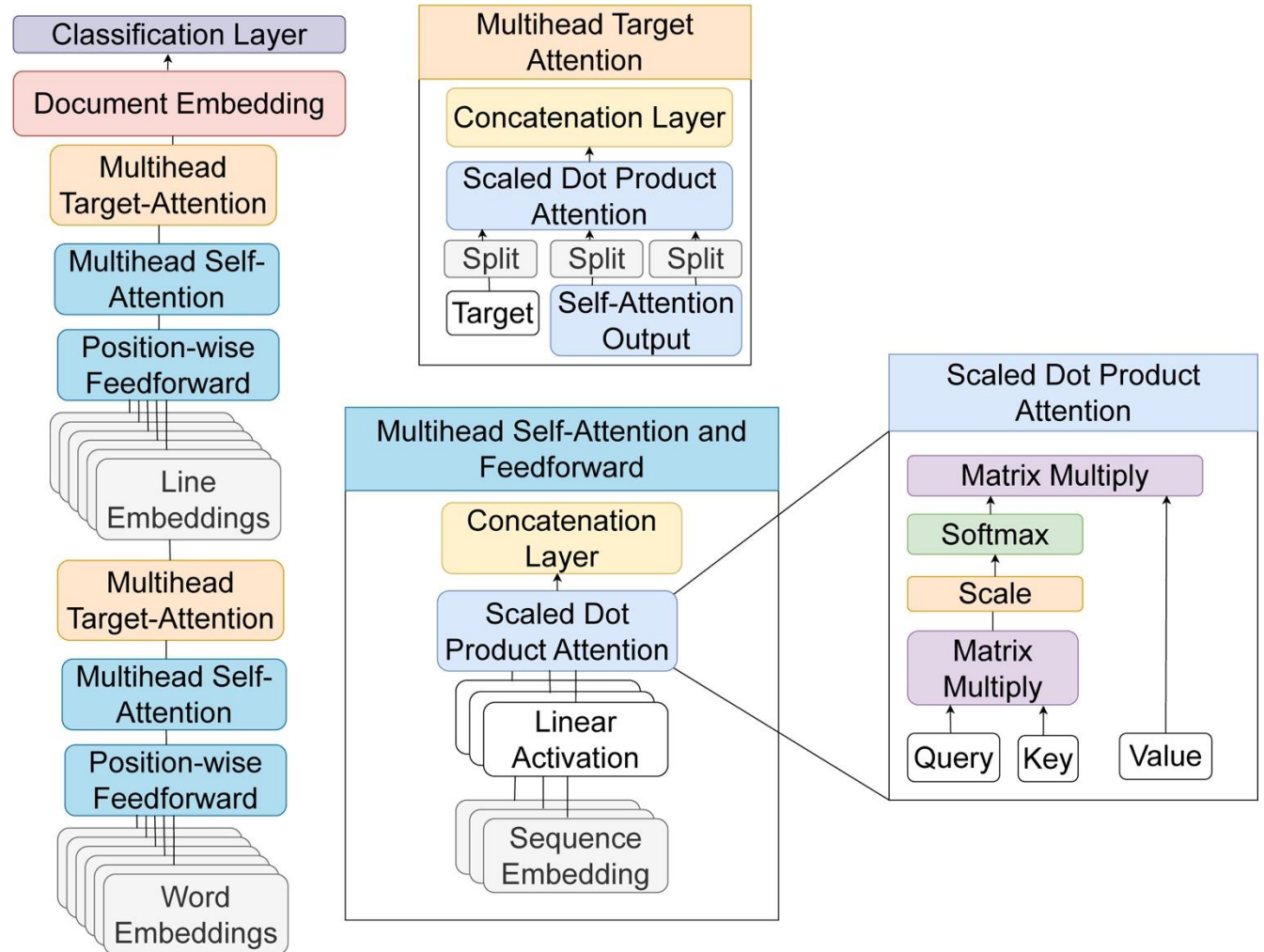




FrESCO

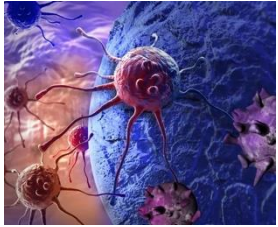


Hierarchical Self-Attention Network (HiSAN)



Case-Level Context Model

Tumor Level



Lung

CTC Label: Manually Annotated Gold Standard for **Overall** Patient/Case



Previous models were at the report level

Lung

Report 1

Lung

Report 2

Lung

Report 3

Breast

Report 4

Lung

Report 5

Case-Level Context

Analyze each pathology report in the **context of the other pathology reports** in the same case (i.e., feeding a sequences of reports into the classifier)

Architecture: Multi-Task Hierarchical Self-Attention (MtHiSAN)

Trusting our Predictions: Deep Abstaining Classifier

Standard cross-entropy training loss

$$\mathcal{L}(x) = - \sum_{i=1}^k t(x)_i \log p(x)_i$$

As probability for true class decreases, $L(x)$ increases

t_i = target for the current sample (1,0)
 P_i = probability that x is classified to the i^{th} class

DAC cross-entropy training loss

Modified cross-entropy loss over the k non-abstaining classes

Penalized Abstention

$$\mathcal{L}(x) = (1 - p(x)_{k+1}) \left(- \sum_{i=1}^k t(x)_i \log \frac{p(x)_i}{1 - p(x)_{k+1}} \right) + \alpha \log \frac{1}{1 - p(x)_{k+1}}$$

p_{k+1} = probability of abstention

α = abstention penalty; when α is closer to 1, there is a high penalty for abstention, when it is close to 0 it may abstain on everything, when in between – the error drive mass into the abstention class

OncoMetsID: Near Real-Time 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

	Total E-Path Reports	No Metastasis	Positive Metastasis	Unknown	Number of Patients
All Registries	60,471	43,286	14,959	2,226	29,632
Seattle	26,507	21,277	4,800	430	12,732
New Jersey	16,770	10,833	5,163	774	8,607
Louisiana	11,886	7,689	3,451	746	5,723
Utah	5,308	3,487	1,545	276	2,570

Results for Metastasis

HiSAN evaluation on 3 classes			
	Precision mean low, high	Recall mean low, high	F1 mean low, high
All data	mean: 0.864 0.856, 0.868	mean: 0.894 0.890, 0.899	mean: 0.878 0.872, 0.883
Breast	0.877 0.869, 0.883	0.903 0.896, 0.908	0.888 0.883, 0.891
Melanoma	0.902 0.871, 0.919	0.919 0.909, 0.932	0.909 0.897, 0.924
Lung	0.788 0.764, 0.810	0.843 0.831, 0.860	0.814 0.799, 0.833
Colorectal	0.877 0.857, 0.885	0.904 0.885, 0.913	0.890 0.870, 0.898
Ovarian/ Fallopian	0.712 0.616, 0.840	0.764 0.679, 0.882	0.735 0.645, 0.860
Other	0.823 0.631, 0.913	0.862 0.744, 0.930	0.839 0.839, 0.917

Results for Metastasis

	HiSAN 3 classes with threshold=0.955			Llama 3 zero-shot learning on 3 classes		
	Precision mean low, high	Recall mean low, high	F1 mean low, high	Precision mean low, high	Recall mean low, high	F1 mean low, high
All data	mean: 0.864 0.856, 0.868	mean: 0.894 0.890, 0.899	mean: 0.878 0.872, 0.883	0.842 0.833, 0.851	0.824 0.817, 0.832	0.825 0.816, 0.834
Breast	0.877 0.869, 0.883	0.903 0.896, 0.908	0.888 0.883, 0.891	0.850 0.842, 0.858	0.833 0.825, 0.842	0.835 0.827, 0.843
Melanoma	0.902 0.871, 0.919	0.919 0.909, 0.932	0.909 0.897, 0.924	0.896 0.884, 0.909	0.850 0.828, 0.871	0.862 0.844, 0.881
Lung	0.788 0.764, 0.810	0.843 0.831, 0.860	0.814 0.799, 0.833	0.772 0.748, 0.796	0.759 0.734, 0.785	0.753 0.726, 0.781
Colorectal	0.877 0.857, 0.885	0.904 0.885, 0.913	0.890 0.870, 0.898	0.853 0.834, 0.871	0.846 0.826, 0.863	0.846 0.828, 0.865
Ovarian/ Fallopian	0.712 0.616, 0.840	0.764 0.679, 0.882	0.735 0.645, 0.860	0.698 0.609, 0.787	0.705 0.626, 0.783	0.689 0.604, 0.773
Other	0.823 0.631, 0.913	0.862 0.744, 0.930	0.839 0.839, 0.917	0.823 0.696, 0.950	0.737 0.624, 0.851	0.758 0.655, 0.861

Uncertainty Quantification/Abstention

- Soft abstention based on softmax scores.
- The confidence threshold is applied to the predicted probabilities for each report (i.e. the softmax values); any report for which the highest predicted probability is lower than the threshold value is omitted from performance calculations. We refer to the proportion of these omitted reports as the abstained percentage.

Threshold	Abstention Rate	Metastasis Abstention Rate	No Metastasis Abstention Rate	Unknown Class Abstention Rate
0.955	mean: 0.271, min: 0.232, max: 0.320	mean: 0.482, min: 0.410, max: 0.549	mean: 0.179, min: 156, max: 0.222	mean: 0.692, min: 0.498, max: 0.716

Metastasis: HiSAN vs Llama 3 8B

	HiSAN 3 classes with threshold=0.955			Llama 3 zero-shot learning on 3 classes		
	Precision mean low, high	Recall mean low, high	F1 mean low, high	Precision mean low, high	Recall mean low, high	F1 mean low, high
All data	0.951 0.941, 0.960	0.969 0.964, 0.974	0.960 0.952, 0.967	0.842 0.833, 0.851	0.824 0.817, 0.832	0.825 0.816, 0.834
Breast	0.955 0.942, 0.963	0.972 0.963, 0.976	0.963 0.952, 0.969	0.850 0.842, 0.858	0.833 0.825, 0.842	0.835 0.827, 0.843
Melanoma	0.961 0.950, 0.975	0.975 0.966, 0.983	0.968 0.958, 0.979	0.896 0.884, 0.909	0.850 0.828, 0.871	0.862 0.844, 0.881
Lung	0.921 0.903, 0.948	0.951 0.941, 0.963	0.935 0.921, 0.955	0.772 0.748, 0.796	0.759 0.734, 0.785	0.753 0.726, 0.781
Colorectal	0.950 0.943, 0.957	0.967 0.965, 0.971	0.958 0.954, 0.964	0.853 0.834, 0.871	0.846 0.826, 0.863	0.846 0.828, 0.865
Ovarian/ Fallopian	0.925 0.875, 0.974	0.949 0.912, 0.979	0.936 0.898, 0.971	0.698 0.609, 0.787	0.705 0.626, 0.783	0.689 0.604, 0.773
Other	0.996 0.979, 1.000	0.996 0.979, 1.000	0.996 0.977, 1.000	0.823 0.696, 0.950	0.737 0.624, 0.851	0.758 0.655, 0.861
Abs. Rate:	mean: 0.271, min: 0.232, max: 0.320					

APIs Installed in SEER*DMS

Diagnostic Information (OncoIE)

- Path Reports in the Training set: 2.5 million reports from 1999 to 2023
- 1,319,992 unique patients
- The model's final predictions are done with case level context.

Reportability (OncoID)

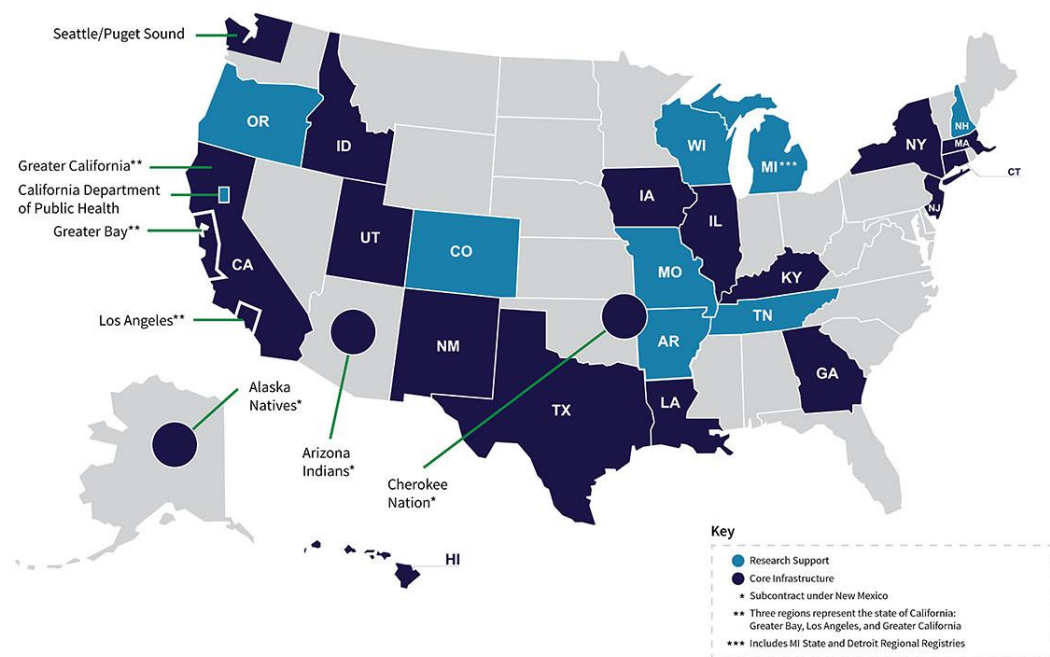
- Path Reports in the Training set: 3.1 million from 2018 to 2022
- Class imbalance is an issue so we offer four weighting schemes so that users can decide the trade-off between False Negative and False Positive rates

Metastasis (OncoMetsID)

- Path Reports in the Training Set: 60k reports annotated by registrars for this task
- API delivered to IMS and in testing phase

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/>

Auto-Extraction from Pathology Reports:

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

Abstention rates can be autotuned to fit the need of the research project.

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

- Total 22 registries (18 SEER + 4 non-SEER)
- Default as part of any new Data Management System installation, regardless of SEER affiliation

Prevailing Challenges: Realizing the Full Potential of AI in Medicine

- Data complexity and regulatory hurdles prevent pooling of data across health care institutions in the US
- Integration of diverse types of social and environmental determinants of health data across space and time requires advanced analytical methods and computational workflows
- Computational limitations prevent scaling algorithms to the population level and have hindered the development and deployment of population health research tools.

Foundation Model for Cancer

Training Foundation Models from Scratch

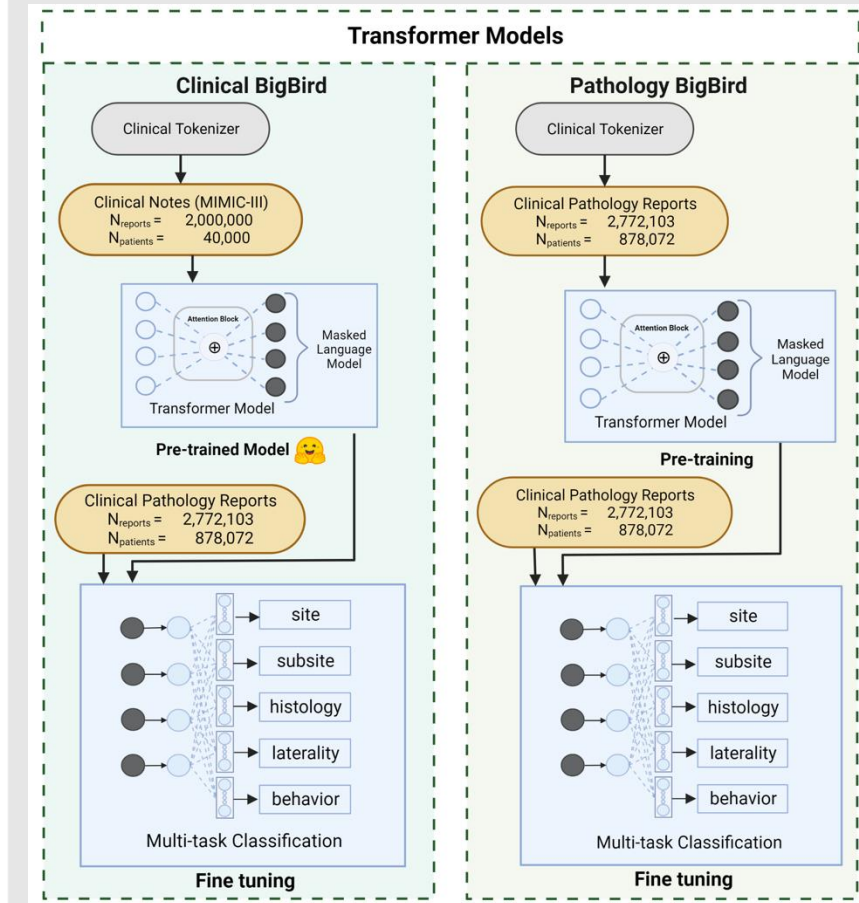
Path-BigBird

Architecture: BERT based BigBird

Data: 2,772,103 pathology reports from six SEER registries

Document Classification: North American Association of Central Cancer Registries

- Histology: Compared to our production level HiSAN model, we see a small increase in micro F1 (80.69 vs 79.25) and moderate increase in macro F1 (37.04 vs 33.22). HiSAN F1 is higher than Clinical BigBird
- This suggests that models trained from scratch have better performance for rare and hard to predict classes
- Less computationally expensive models trained on domain specific data have utility in resource constrained environments



Mayanka Chandrashekar et al., Path-BigBird: An AI-Driven Transformer Approach to Classification of Cancer Pathology Reports. *JCO Clin Cancer Inform* 8, e2300148(2024). DOI:10.1200/CCI.23.00148

Improving Attention Mechanisms

- The first comprehensive comparative study of the initialization strategies for label-wise attention mechanism across multiple text-encoder architectures.
- Initializing label-wise attention with pretrained information improve classification performance of all models.
- We adapted the label-wise attention mechanism to be able to learn the implicit hierarchical structure of common medical encoding systems such as ICD-9.
- Our hierarchical adaptations of label-wise attention can be used to improve model performance for near real-time extraction of patient phenotypes.
- Data: MIMIC-III-Full and MIMIC-III-50

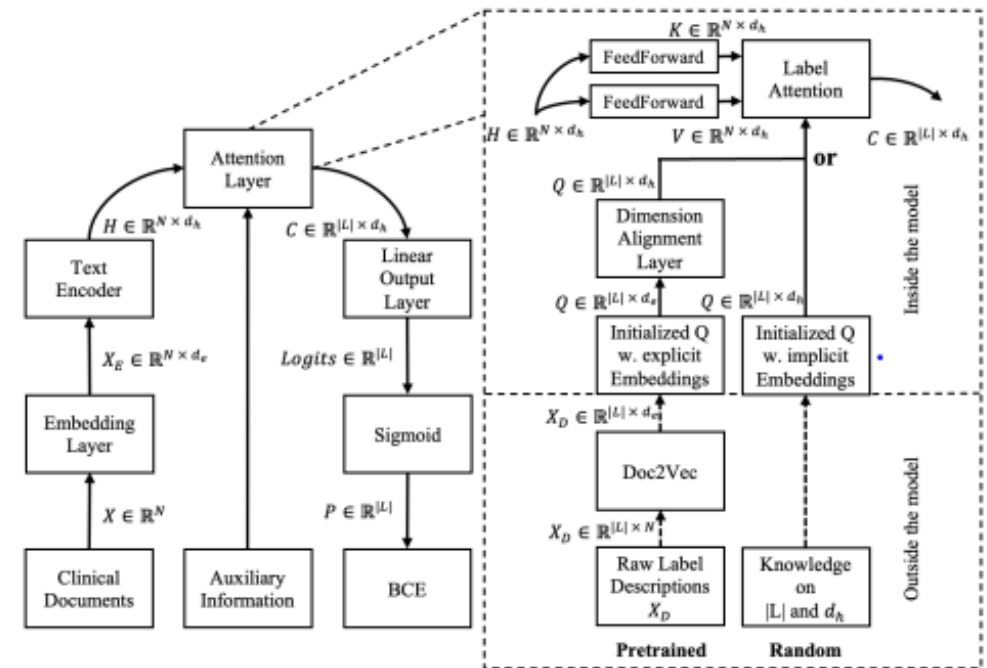


Fig. 1. Structure of the general document classification attention model. The text-encoder segment represents a CNN-, RNN-, or transformer-based architecture. The label-attention layer utilizes implicit (i.e., random embeddings) or explicit (e.g., embedded textual code descriptions) auxiliary information. The multi-label classification models use a sigmoid activation function after the output layer and the binary-cross-entropy objective function.

Charting a path to near real-time data analytics at population scale and data driven clinical decision-making tools: DOE leadership computing

2021-2022 ASCR Leadership Computing Challenge allocation

Title: "Next-Generation Scalable Deep Learning for Medical Natural Language Processing".
130,000 node hours on OLCF Summit



2022-2023 ASCR Leadership Computing Challenge allocation

"Privacy-preserving Transformer models for clinical natural language processing". 150,000 node hours on OLCF Summit and 30,000 node hours on OLCF Frontier

- Use CITADEL, the OLCF secure computing capability
- Secure file system and compute node access: runs normal job on Summit compute nodes in a HIPAA compliant manner

2023-2024 ASCR Leadership Computing Challenge allocation

Title: "Privacy enabled tumor classification for near real time population health analytics".
140,000 node hours on OLCF Frontier



Sustained computing support from DOE over MOSSAIC project lifetime

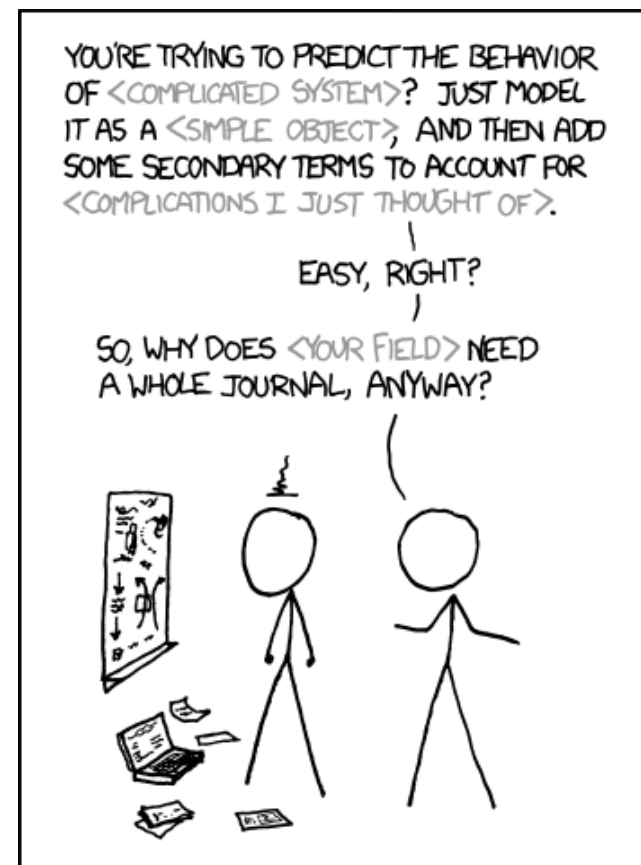
Total of 560,000 Summit node hours through the ALCC program

Approximately 325,000 additional Summit node hours provided via the Exascale Computing Project, OLCF Director's Discretionary, and the OLCF Summit Early Science programs



Team science

Computational Scientists,
Engineers, Biostatisticians,
Biomedical Informaticists,
Geneticists, Subject Matter
Experts



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 (Project 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



Discussion