DOE-NCI Collaboration ASCAC Subcommittee

Professor Tony Hey
Chief Data Scientist
Rutherford Appleton Laboratory
STFC/UKRI

ASCAC DOE-NCI Subcommittee

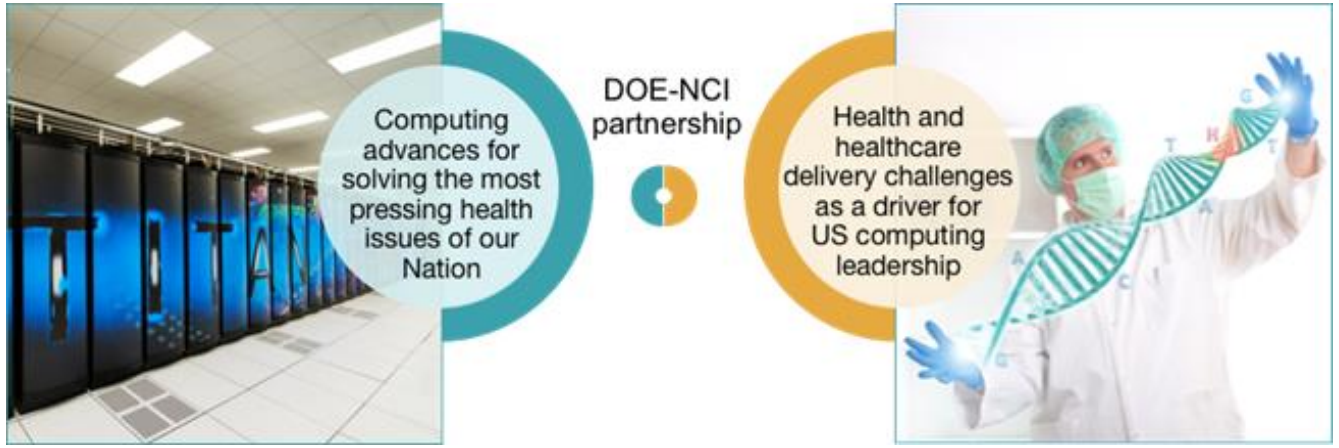
- Tony Hey – STFC, ASCAC (Chair)
- Rick Arthur – GE, ASCAC
- Jay Bardhan – PNNL
- Martin Berzins – Utah, ASCAC
- Bill Gropp – UIUC, NCSA
- Satheesh Maheswaran – DLS, UKRI
- Amanda Randles – Duke
- Vadim Backman – Northwestern
- Caroline Chung – MD Anderson
- Susan Gregurick – NIH, ASCAC
- Amie Hwang – USC
- Gordon Mills - OHSU
- Joel Saltz – Stony Brook

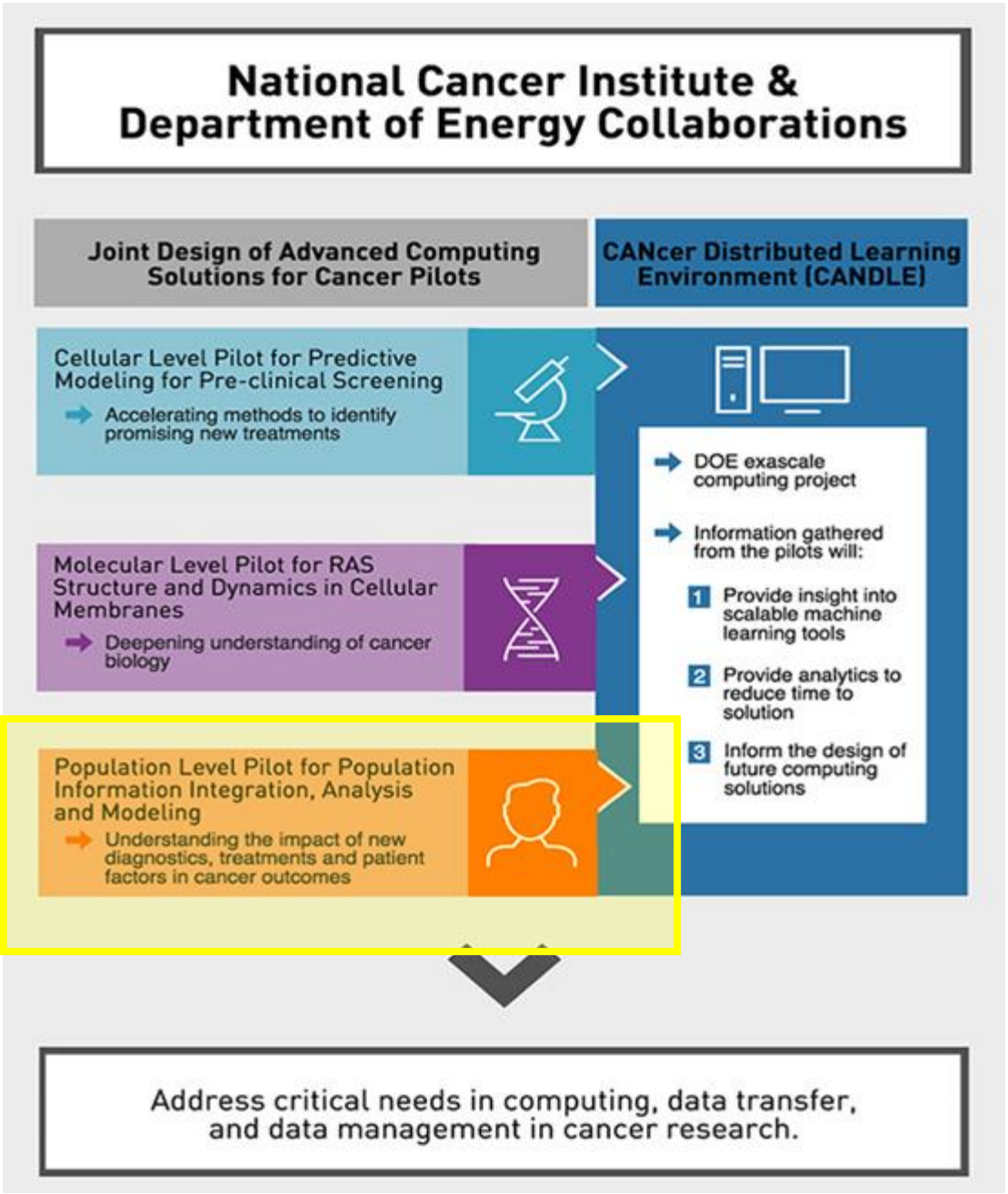
DOE-NCI Subcommittee Meeting
Monday 28th March

1. Introduction and welcome (Tony Hey)
2. Review of MOSSAIC project (Gina Tourassi and Lynne Penberthy)
3. Discussion (led by Amie Eunah Hwang and Joel Saltz)
4. Conclusions (Subcommittee discussion)
5. AOB

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

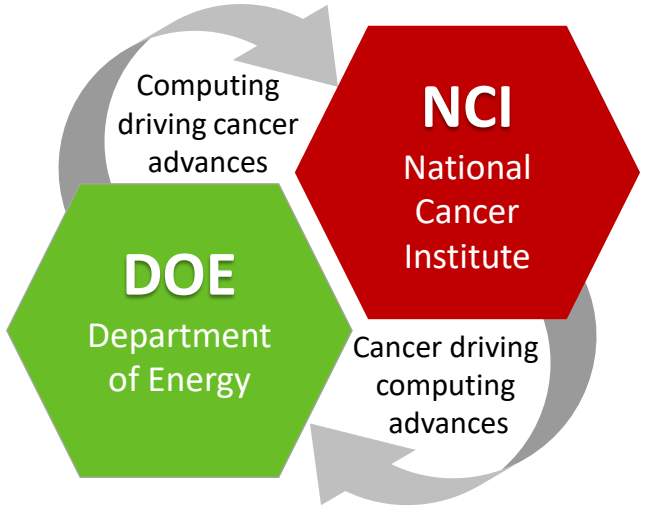
Enable the most challenging deep learning problems in cancer research to run on the most capable supercomputers in the DOE





MOSSAIC: Modeling **O**utcomes using **S**urveillance data and **S**calable **AI** for **C**ancer

DOE-NCI partnership to advance exascale development through cancer research



Lynne Penberthy
National Cancer Institute

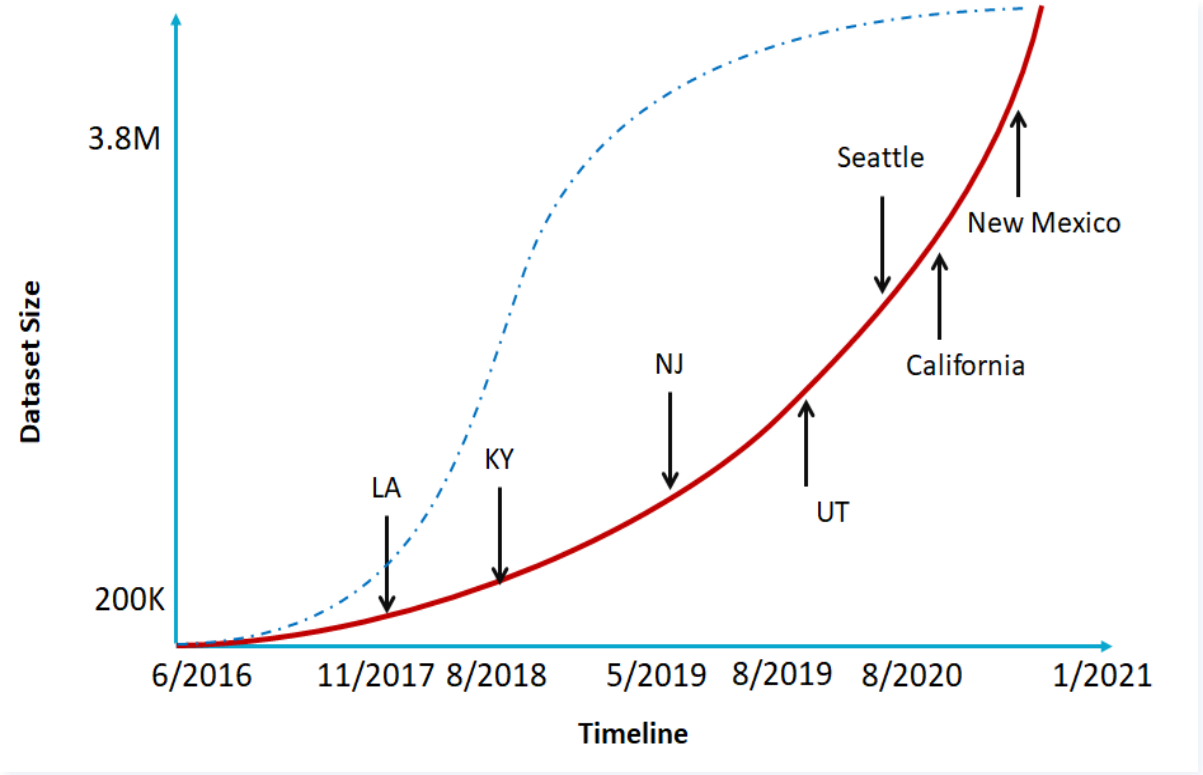
Georgia Tourassi
Oak Ridge National Laboratory

March 28, 2022

Presented to:
ASCAC DOE-NCI Subcommittee

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. [Also adding in the appropriate review and release number]

MOSSAIC built on Partnerships and Community Outreach



NCI

- Division of Cancer Control and Population Sciences
- Coordinating Center for Clinical Trials
- Childhood Cancer Data Initiative (NCCR)

2 DOE labs - ORNL, LANL

10 Registries (9 SEER Registries and 1 non-SEER)

- LA, KY, NJ, UT, CA, Los Angeles, Seattle, NM, ***New York, MN***
- **1.2M patients, 1.4M cancer cases, 3.5M documents**
- **Any registry that uses DMS can request deployment of the API in their production system**

Industry and other Stakeholders

- IMS, CancerLinQ, Health Verity, Exact Sciences (Genomic Health), Castle Biosciences, Decipher, CVS, Walgreens, FMI, Lexis Nexis, OptumLabs, Quest Diagnostics, Intermountain HC

Academia

- Clinical collaborators provide domain expertise
- Fred Hutchison, Emory, UKY, Huntsman CI, Dana Farber, MGH etc.

Other Federal Agencies

- VA, CDC, FDA

The MOSSAIC NLP Framework

AI-Driven NLP algorithmic innovation

NLP Model Search
- Convolutional neural networks
- Recurrent neural networks
- Attention-based neural networks
- Graph convolutional networks
- ELMO and BERT

Supportive Techniques
- Multi-task learning
- Uncertainty Quantification
- Privacy-preserving Learning

Data Representation
- Word/Phrase embeddings: Word2Vec, FastText ....
- Case-Level Multi-Document Representation
- CUI representation

- Supervised
- Semi-supervised
- Unsupervised

Increasing robustness to label quality

Clinical Relevance

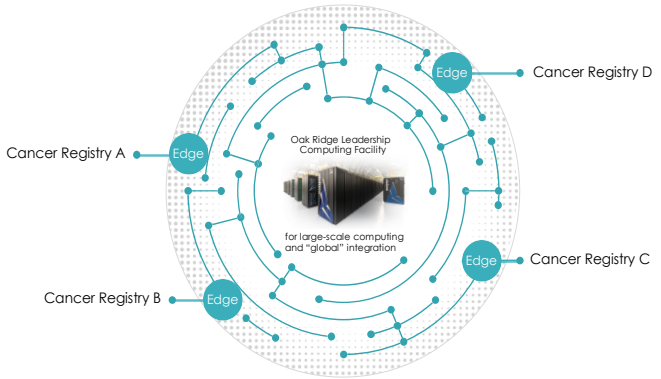
Scalability across pathology labs, SEER registries, and cancer phenotypes

Reportability
Biomarkers
Cancer phenotypes
- Site
- Subsite
- Histology
- Grade
- Behavior
- Laterality
- Metastasis / Recurrence

...

O(11000) cancer phenotypes

Scalable Deployment



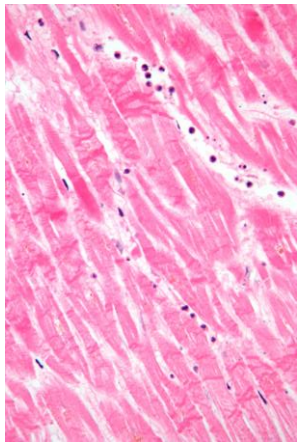
70 cancer sites (330+ subsites); 570+ histologies; 9 grades; 7 lateralities; 4 behaviors;…

→ >11K cancer phenotypes observed based only on 5 attributes

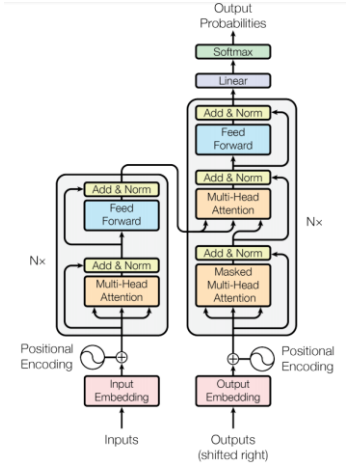
Transformers for large-scale multi-modal data

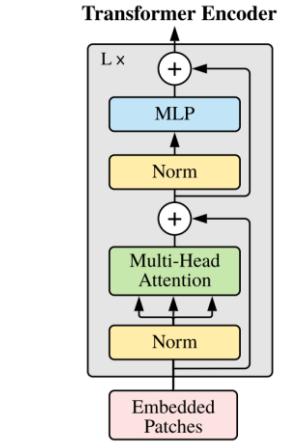
Dataset

Cancer pathology reports



Transformer architecture





HPC resources

Task

Pretraining on dataset: Deployed at scale on exascale machines

Fine-tuning on dataset: Distributed & GPU-accelerated runs

DOE Leadership Computing and MOSSAIC

2021-2022 ASCR Leadership Computing Challenge (ALCC) allocation

- Title: “Next-Generation Scalable Deep Learning for Medical Natural Language Processing”
- 130,000 node hours on OLCF Summit
- Ongoing effort using CITADEL, the OLCF secure computing capability, to train models with MOSSAIC PHI data on Summit

Sustained computing support from DOE over MOSSAIC project lifetime

- Total of 270,000 Summit node hours through the **ALCC** program
- Approximately 300,000 additional Summit node hours provided via the **Exascale Computing Project**, OLCF Director’s Discretionary, and the OLCF Summit Early Science programs
- ORNL secure data enclave resources utilized for data storage and and mid-level computing support at no cost to the project.

Continued development and readiness for the DOE exascale platforms

- Since early 2020, tested on 3 generations of **Frontier development systems (Tulip, Spock, Crusher)**
- Expect Day 1 readiness for MOSSAIC on **Frontier**

MOSSAIC successfully impacting DOE's mission in supercomputing and AI

AI ALGORITHMIC ADVANCES

- Translational solutions for trustworthy, explainable, and secure AI for broad societal impact
- Extensible to emerging NLP applications of critical importance, e.g., cybersecurity
- Key use case for low precision sparse matrix multiply, motivating hardware advances to accelerate peak performance of AI models using sparse tensor algorithms (i.e., Transformers for clinical documents)

PRIVACY-PRESERVING, FEDERATED LEARNING, AND DISTRIBUTED AI

- Driving use case to prototype OLCF's *CITADEL* framework enhanced with federated learning and differential privacy and assess privacy gains vs accuracy loss
- Driving use case to develop a suite of model attack methods and evaluate the security of our methods and models using the CITADEL framework

EDGE-TO-EXASCALE-TO-EDGE

- A prototype of Integrated Research Infrastructure – linking compute and observational facilities, across domains and agencies
- The control and workflows span multiple facilities, storage, and computing: multi-institutional, sub-facility (PHI to moderate); enabling multi-modal exploration

Priorities in 2022

Activity 1: Scalable Transformer language models for clinical information extraction

Activity 2: Data collection and development of new recurrence and biomarker APIs

Activity 3: Uncertainty quantification and interpretability

Activity 4: Clinical integration and translation

Activity 5 (CANDLE): Enabling Transformer training on LCF systems

Preliminary Conclusions and Next Steps

- Subcommittee impressed by achievements of MOSSAIC project
  - Production NLP system now deployed at several SEER Centers
  - Interested in other potential application areas
  - Concerns about sustainability?

- Next meetings will focus on in-depth reviews of
  - ADMIRRAL and IMPROVE pilot projects
  - Cross-cutting CANDLE benchmarking project