OAK RIDGE National Laboratory

Artificial Intelligence for Near-Real Time Cancer Surveillance: Challenges and Opportunities

Georgia Tourassi, PhD

Director, National Center for Computational Sciences

Oak Ridge National Laboratory

Presented at the NAACCR Annual Meeting

June 25, 2020

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

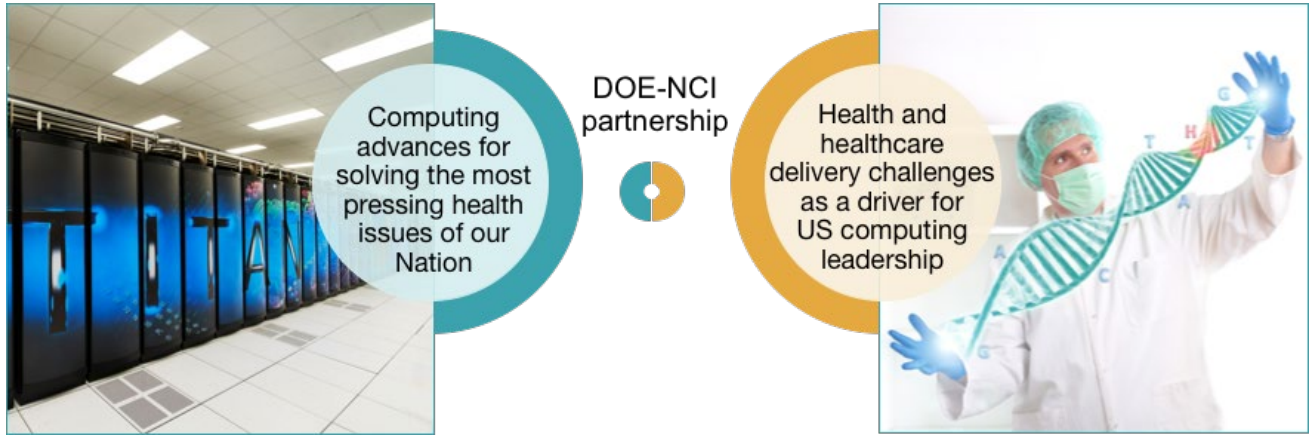
U.S. DEPARTMENT OF ENERGY

NIH NATIONAL CANCER INSTITUTE SEER

No Disclosures

DOE-NCI Partnership:

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



National Cancer Institute & Department of Energy Collaborations

Joint Design of Advanced Computing Solutions for Cancer Pilots

CANcer Distributed Learning Environment (CANDLE)

Cellular Level Pilot for Predictive Modeling for Pre-clinical Screening
- Accelerating methods to identify promising new treatments

Molecular Level Pilot for RAS Structure and Dynamics in Cellular Membranes
- Deepening understanding of cancer biology

Population Level Pilot for Population Information Integration, Analysis and Modeling
- Understanding the impact of new diagnostics, treatments and patient factors in cancer outcomes

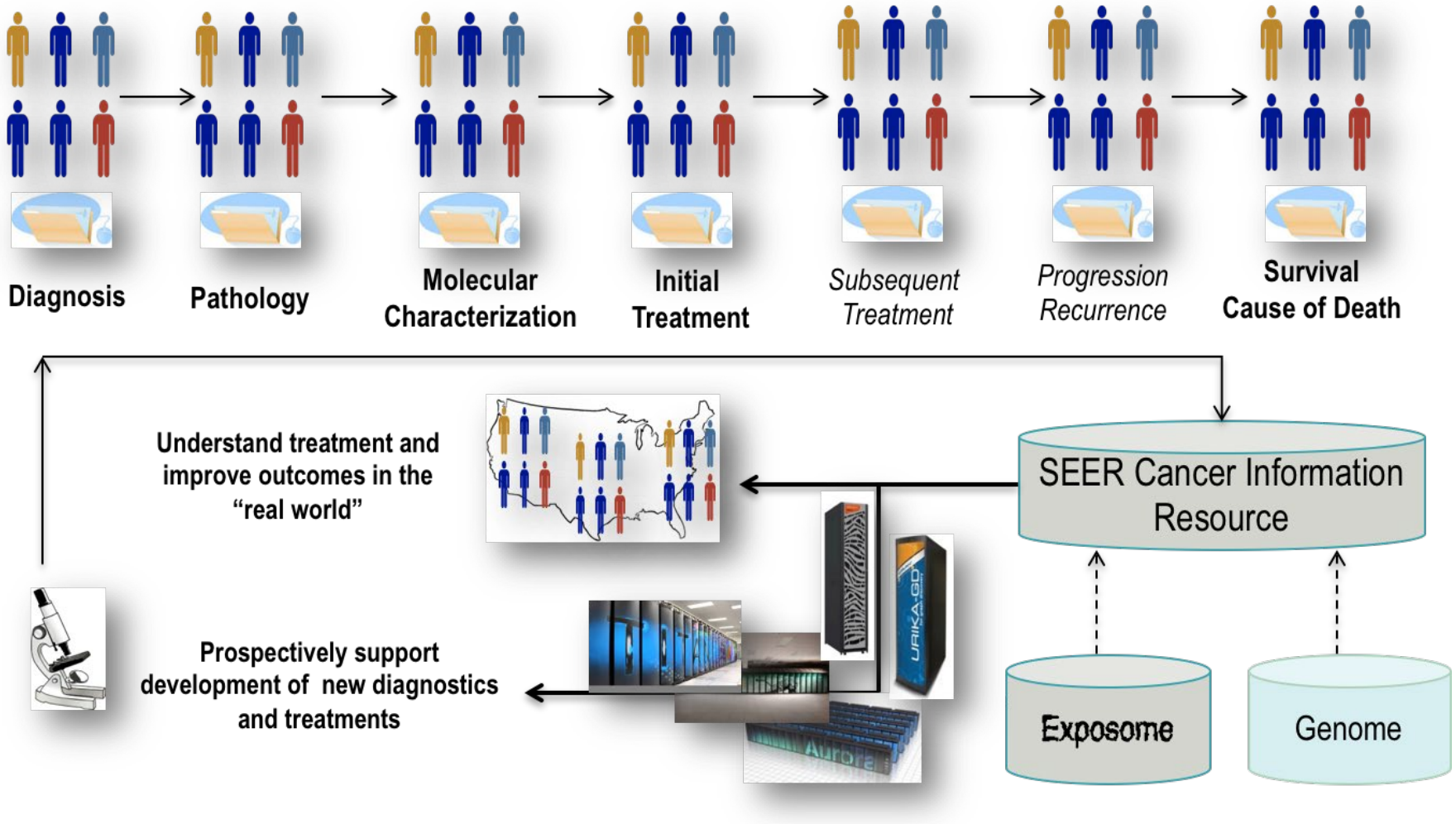
- DOE exascale computing project
- Information gathered from the pilots will:
  1. Provide insight into scalable machine learning tools
  2. Provide analytics to reduce time to solution
  3. Inform the design of future computing solutions

Address critical needs in computing, data transfer, and data management in cancer research.

NIH NATIONAL CANCER INSTITUTE

OAK RIDGE National Laboratory

AI to Modernize the National Cancer Surveillance Program



To develop and deploy robust and scalable AI solutions for automated information extraction from free text pathology reports.

I will present a few vignettes on…

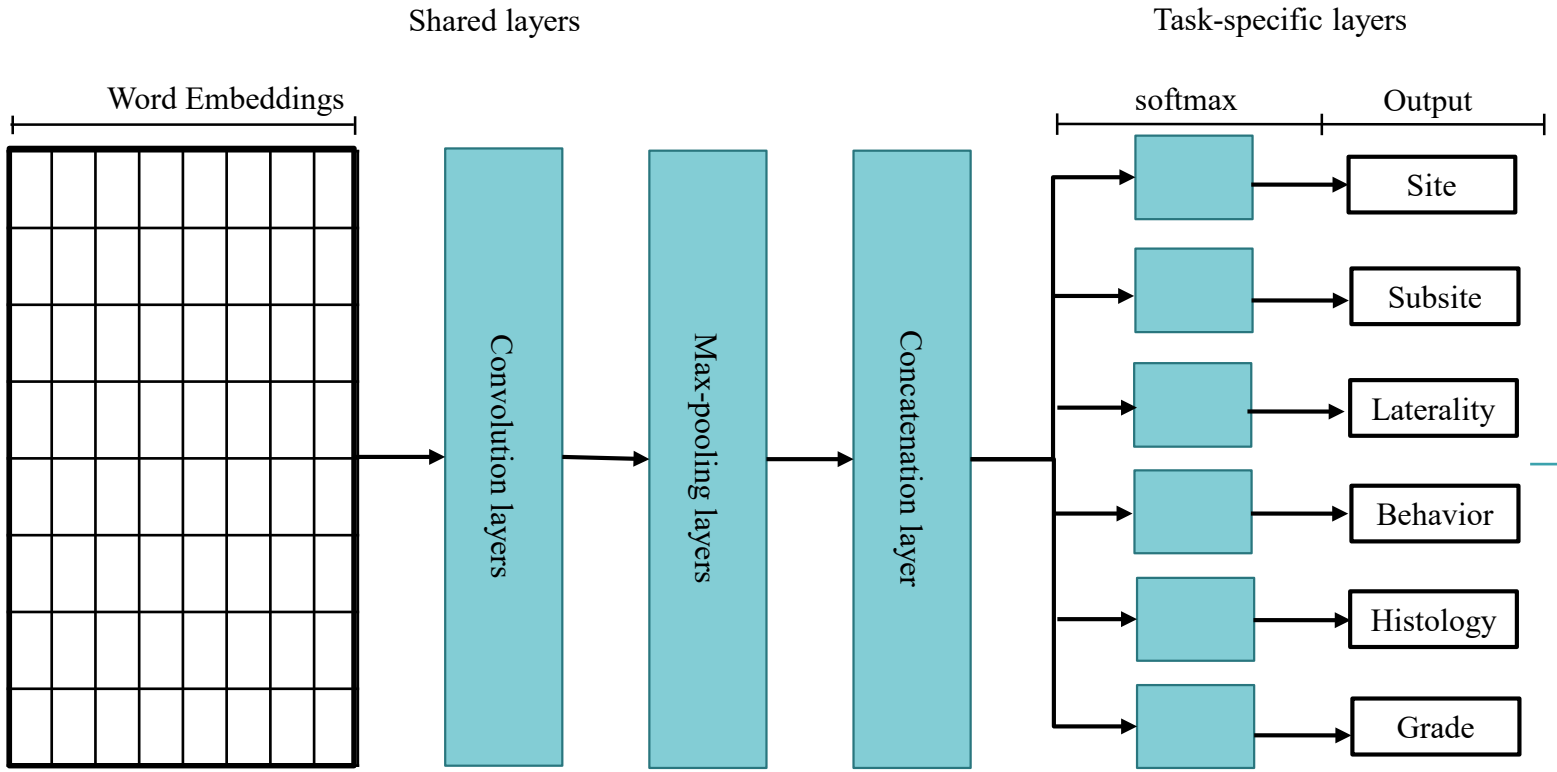
- Information extraction of reportable data elements
- Privacy-preserving AI model sharing
- Reportability
- Recurrence

AI Methodology

- Different deep learning architectures for simultaneous learning of multiple information extraction tasks
  - site, topography, histology, behavior, laterality, grade,…
- Support both report-level and case-level analysis
- Minimal document pre-processing
- Gold standard: Variables coded in the registry abstract
- Benchmarking against traditional machine learning
- Testing within and across SEER registries

Information Extraction of Key Data Elements

NLP Algorithmic innovation and Hyperparameter Optimization



Development & deployment of product

API

70 cancer sites (306 subsites); 515 histologies; 9 grades; 7 lateralities; 4 behaviors

>20,000 cancer phenotypes observed in based only on 6 attributes

Extension to other NLP tasks to extract more data elements (e.g., biomarkers) will increase the number and complexity of cancer phenotypes observed – **combinatorial explosion in computational cancer phenotyping → Exascale computing**

Development and Testing Protocol

- Iterative development and refinement using SEER data
  - Louisiana, Kentucky, Utah, New Jersey, California, Seattle,….
- Broad deployment and testing via IMS
  - Across 13 SEER registries
- Evaluation metrics
  - Overall accuracy
  - Accuracy based on report type
  - Accuracy on over-represented vs. under-represented classes
  - Uncertainty quantification (i.e., confidence)
  - Time efficiency

Iterative Improvement and Testing: 13 SEER Registries, ~4M documents

Trained on 2 registries

V6	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	Average
Site	93.96%	92.04%	93.10%	94.54%	92.84%	92.64%	95.23%	93.13%	93.97%	93.86%	94.38%	93.95%	93.41%	93.62%
Histology	84.52%	79.33%	82.67%	83.03%	84.11%	82.50%	85.63%	82.86%	82.68%	81.03%	80.19%	82.22%	78.82%	82.28%
Laterality	93.38%	92.39%	93.92%	92.95%	93.28%	93.68%	94.76%	93.06%	92.93%	94.24%	92.31%	94.22%	90.39%	93.19%
Behavior	96.61%	96.47%	95.81%	96.51%	95.97%	96.84%	97.52%	95.25%	96.40%	97.23%	95.58%	96.53%	97.45%	96.47%
Grade	79.82%	75.23%	77.06%	81.20%	78.83%	78.66%	82.93%	79.38%	78.16%	78.15%	79.92%	81.55%	79.12%	79.23%
Average	89.66%	87.09%	88.51%	89.65%	89.01%	88.86%	91.21%	88.74%	88.83%	88.90%	88.48%	89.69%	87.84%	

Trained on 4 registries

V7	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	Average
Site	93.88%	91.89%	92.95%	94.48%	93.18%	92.69%	95.37%	93.15%	94.55%	93.91%	94.11%	95.11%	92.67%	93.69%
Histology	87.23%	83.58%	87.12%	86.26%	87.90%	86.33%	88.80%	87.09%	86.91%	84.74%	83.04%	87.56%	83.91%	86.19%
Laterality	94.21%	92.96%	94.11%	93.92%	93.95%	93.99%	95.27%	93.94%	94.02%	94.37%	93.55%	95.27%	92.32%	93.99%
Behavior	96.67%	96.71%	96.22%	96.94%	96.69%	97.10%	97.74%	95.74%	96.96%	97.58%	95.76%	97.17%	97.35%	96.82%
Grade	81.60%	80.21%	78.99%	83.52%	81.13%	81.95%	85.15%	80.47%	81.87%	80.44%	82.77%	85.23%	82.67%	82.00%
Average	90.72%	89.07%	89.88%	91.02%	90.57%	90.41%	92.47%	90.08%	90.86%	90.21%	89.85%	92.07%	89.78%	

Efficiency

Time study results:

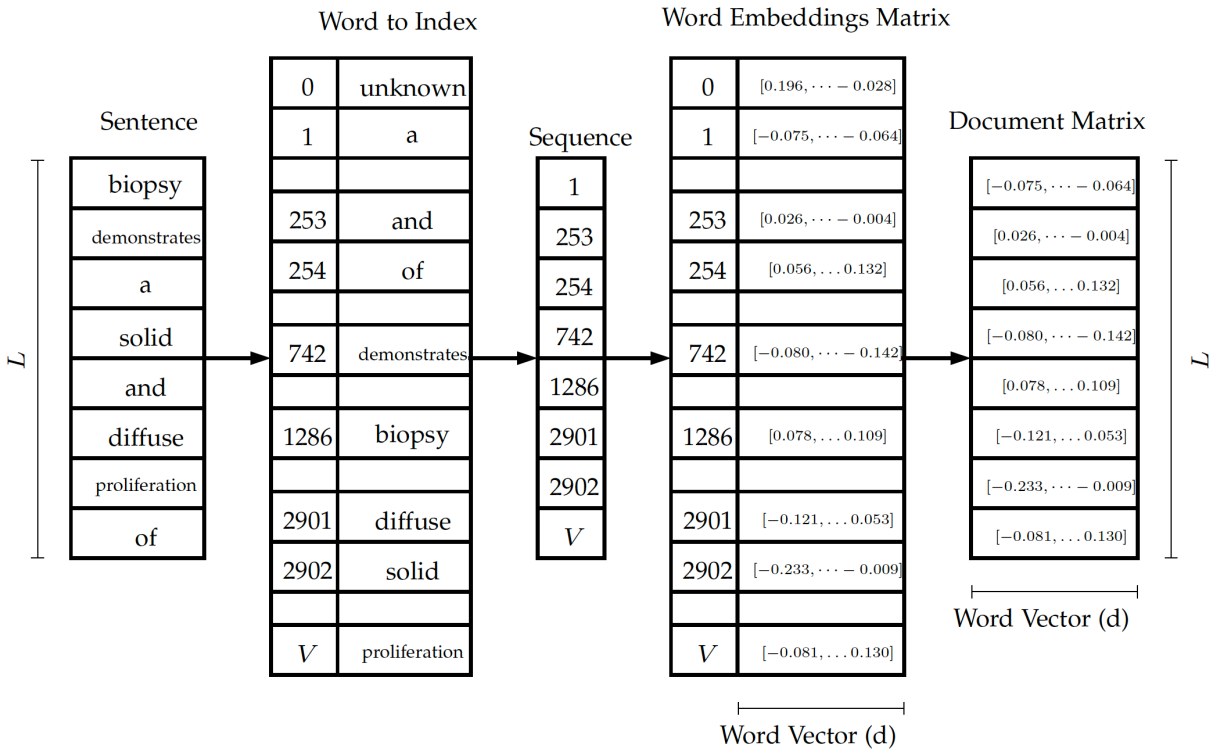
- **Manual** path screening for 5 variables – 1 year,
  - 600,000 path reports: 4,048 hrs (**55sec/report**)
- **AI** path screening on same task: 55 min (**12msec/report**)
- **4500x speed gain to enable near real time cancer surveillance**

Sharing the Model

- If we can't share data, can we share trained models across registries and beyond?
  - Contrary to imaging applications, sharing the trained deep learning NLP model implies sharing also the vocabulary
  - The latter raises privacy concerns, since the vocabulary contains names, addresses, and other PII information.
- Privacy-preserving model

Proposed Solution

- Build a dictionary from all available word in the registry training corpus
- Exclude all words that are not available in a publicly available dictionary
- Train AI model using the reduced dictionary



OAK RIDGE National Laboratory

Iterative Improvement and Testing: 13 SEER Registries, ~4M documents

Model trained on 2 registries

V6	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	Average
Site	93.96%	92.04%	93.10%	94.54%	92.84%	92.64%	95.23%	93.13%	93.97%	93.86%	94.38%	93.95%	93.41%	93.62%
Histology	84.52%	79.33%	82.67%	83.03%	84.11%	82.50%	85.63%	82.86%	82.68%	81.03%	80.19%	82.22%	78.82%	82.28%
Laterality	93.38%	92.39%	93.92%	92.95%	93.28%	93.68%	94.76%	93.06%	92.93%	94.24%	92.31%	94.22%	90.39%	93.19%
Behavior	96.61%	96.47%	95.81%	96.51%	95.97%	96.84%	97.52%	95.25%	96.40%	97.23%	95.58%	96.53%	97.45%	96.47%
Grade	79.82%	75.23%	77.06%	81.20%	78.83%	78.66%	82.93%	79.38%	78.16%	78.15%	79.92%	81.55%	79.12%	79.23%
Average	89.66%	87.09%	88.51%	89.65%	89.01%	88.86%	91.21%	88.74%	88.83%	88.90%	88.48%	89.69%	87.84%	

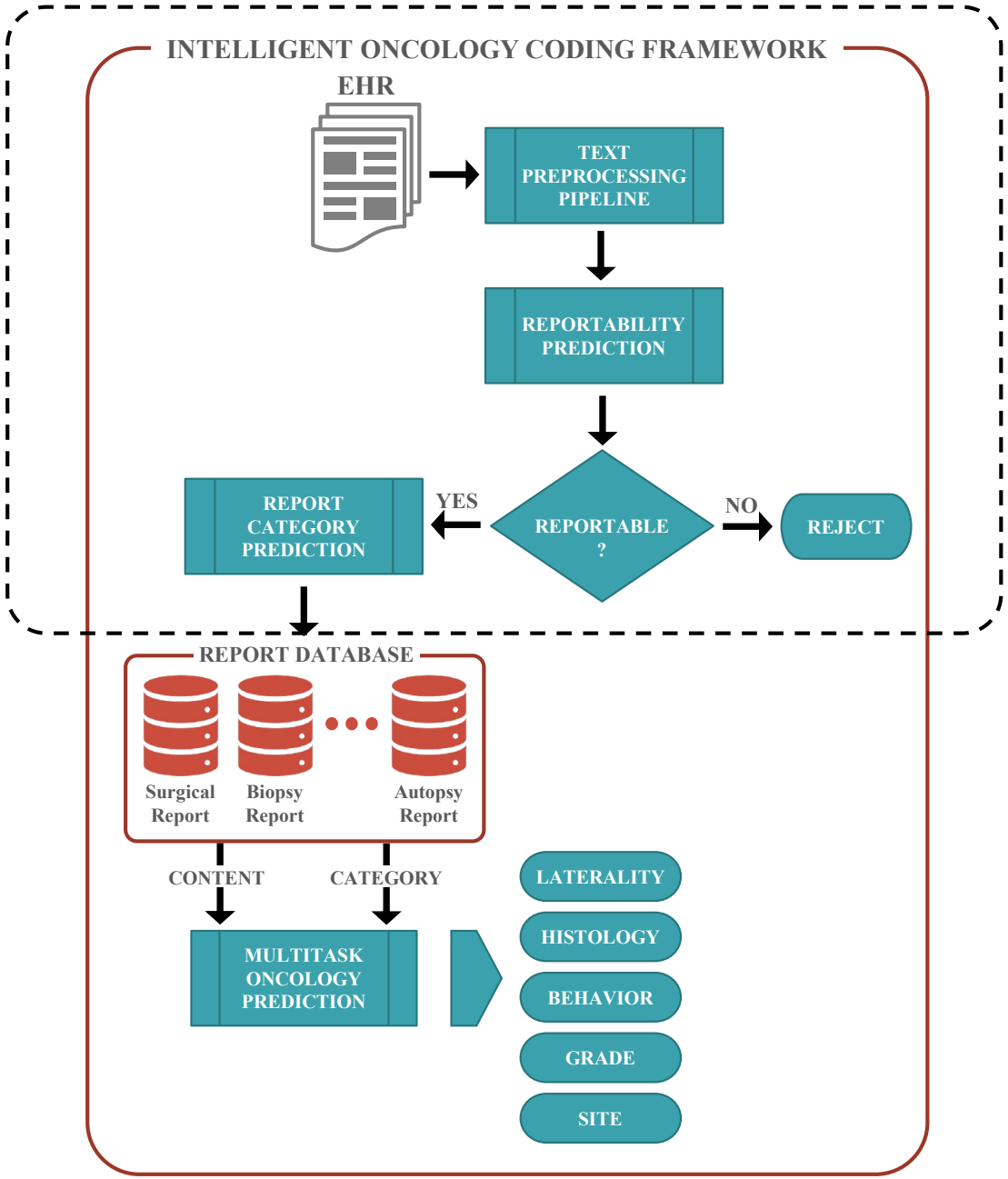
Model trained on 4 registries

V7	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	Average
Site	93.88%	91.89%	92.95%	94.48%	93.18%	92.69%	95.37%	93.15%	94.55%	93.91%	94.11%	95.11%	92.67%	93.69%
Histology	87.23%	83.58%	87.12%	86.26%	87.90%	86.33%	88.80%	87.09%	86.91%	84.74%	83.04%	87.56%	83.91%	86.19%
Laterality	94.21%	92.96%	94.11%	93.92%	93.95%	93.99%	95.27%	93.94%	94.02%	94.37%	93.55%	95.27%	92.32%	93.99%
Behavior	96.67%	96.71%	96.22%	96.94%	96.69%	97.10%	97.74%	95.74%	96.96%	97.58%	95.76%	97.17%	97.35%	96.82%
Grade	81.60%	80.21%	78.99%	83.52%	81.13%	81.95%	85.15%	80.47%	81.87%	80.44%	82.77%	85.23%	82.67%	82.00%
Average	90.72%	89.07%	89.88%	91.02%	90.57%	90.41%	92.47%	90.08%	90.86%	90.21%	89.85%	92.07%	89.78%	

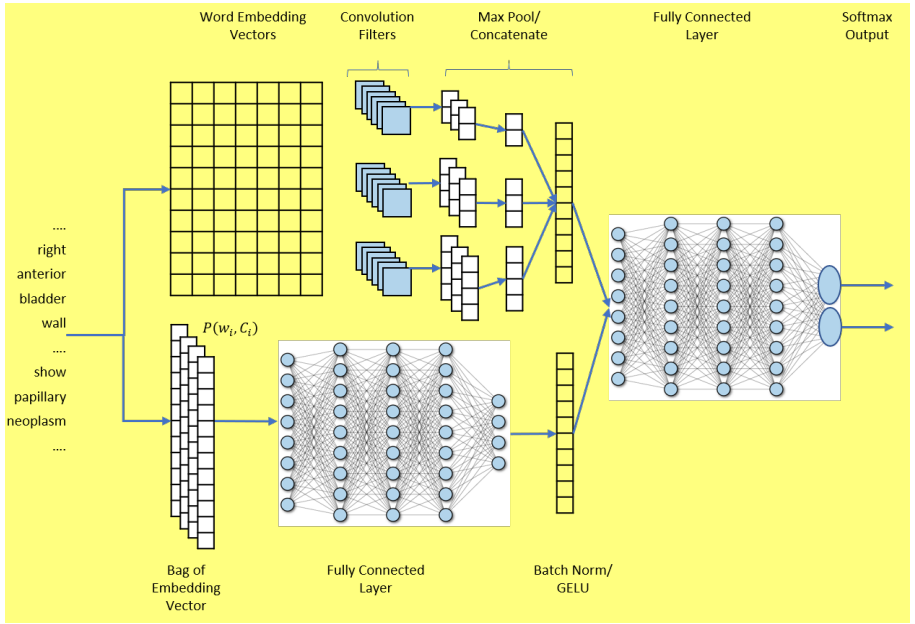
Privacy-Preserving Model

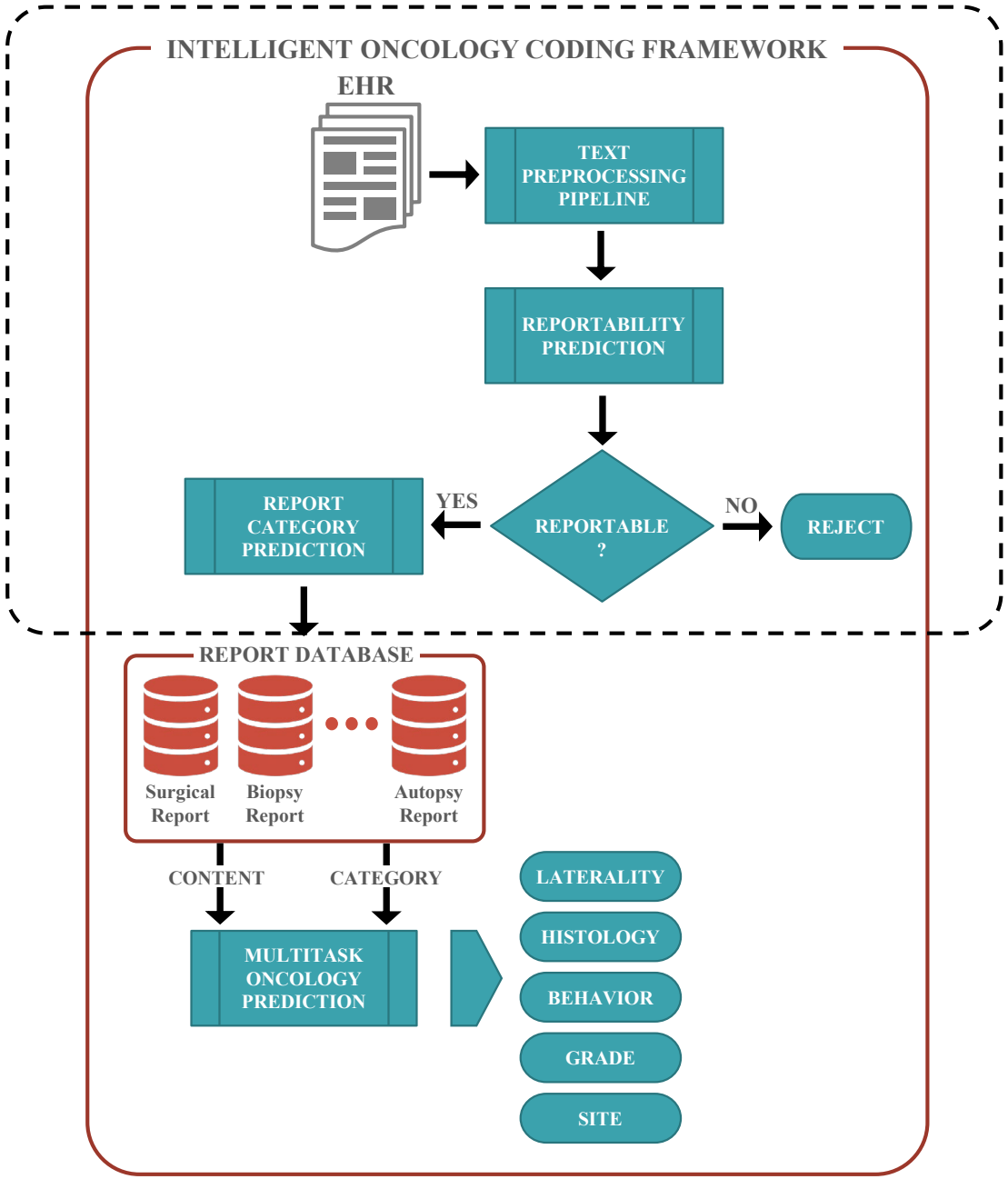
V7PP	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	Average
Site	93.88%	92.35%	93.11%	94.70%	93.12%	92.64%	95.50%	93.19%	94.69%	94.12%	94.34%	95.26%	92.38%	93.79%
Histology	87.62%	83.76%	87.01%	85.88%	87.66%	86.77%	89.02%	87.30%	87.15%	84.84%	83.88%	87.68%	83.42%	86.31%
Laterality	94.17%	92.92%	94.08%	93.76%	93.93%	93.96%	95.18%	93.27%	93.95%	94.22%	93.34%	95.20%	91.92%	93.84%
Behavior	96.62%	96.75%	96.22%	97.06%	96.72%	97.08%	97.81%	95.81%	96.96%	97.62%	95.92%	97.19%	97.68%	96.88%
Grade	81.28%	79.34%	78.91%	83.44%	80.76%	80.35%	84.78%	80.45%	81.51%	79.83%	82.32%	84.95%	82.80%	81.59%
Average	90.71%	89.02%	89.87%	90.97%	90.44%	90.16%	92.46%	90.00%	90.85%	90.13%	89.96%	92.06%	89.64%	

Reportability



Reportability





Preliminary Results

- Datasets:
  - Kentucky cancer registry (n=1509)
  - Louisiana cancer registry (n=61610)
  - New Jersey cancer registry (n=14463)
  - Utah cancer registry (n=27677)
- Inference Task:
  - Reportability (reportable; non-reportable and unclear)
- Model training and evaluation (k-fold cross validation):
  - Comparison with alternative predictive models
    - Logistic Regression (LR)
    - Naïve Bayes (NB)
    - Support Vector (SVM)
    - Convolutional neural network (single layer CNN)
    - CNN + bag-of-embeddings (CNN*)

Model	Sensitivity	Specificity	AUC
LR	90.13 (90.03, 90.22)	70.28 (70.04, 70.51)	89.82 (89.73, 89.92)
NB	89.12 (89.02, 89.22)	59.14 (58.88, 59.39)	84.82 (84.71, 84.94)
SVM	93.46 (93.38, 93.54)	49.73 (49.47, 49.98)	85.28 (85.16, 85.40)
CNN	90.12 (90.02, 90.23)	69.90 (69.68, 70.12)	90.25 (90.15, 90.35)
CNN*	90.91 (90.82, 91.01)	71.56 (71.33, 71.79)	91.80 (91.71, 91.88)

Recurrence

- Recurrence is essential to report given that 5% of the US population are cancer survivors and at risk
- Real time identification of recurrence is also critical to the clinical trials infrastructure as many trials are focused on recurrence
- Initial focus on unstructured pathology reports
- Objective: Identify pathology reports indicating "de novo" mets
- Hypothesis: Model trained to detect metastasis at the time of diagnosis (for which we have CTC gold standard) can be used to detect metastasis indicative of disease progression.

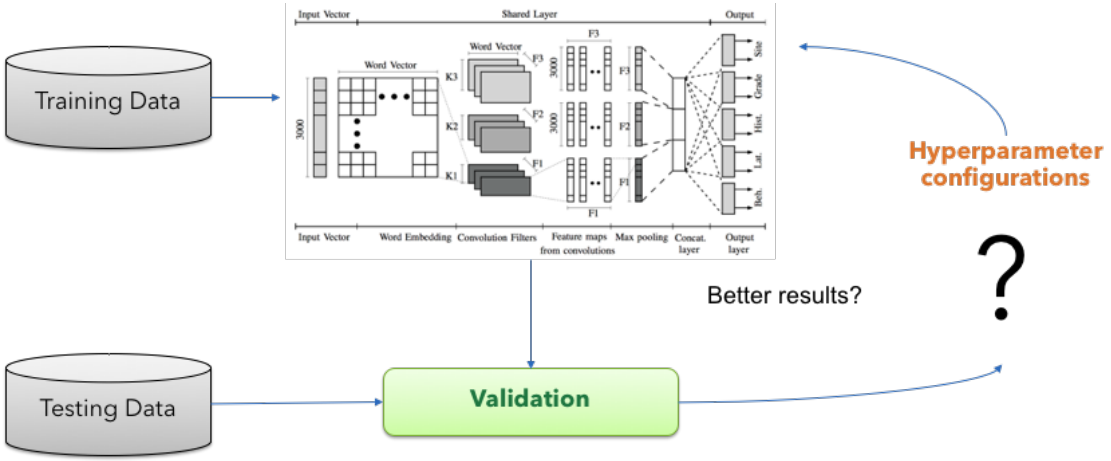
Preliminary Results

- Denovo metastasis classification performance:
  - 98% accuracy
  - 90% sensitivity and 2.9% false positive rate

- Performance on additional, manually annotated data including all metastases
  - 85% accuracy
  - 82.6% sensitivity and 14.5% false positive rate

Building & Scaling AI NLP Tools on Summit

Peak of 3.3 ExaOps for data analytics and AI



Multi-task learning convolutional neural network trained to extract primary cancer site, histology, behavior, laterality , and grade

Computational Scalability

- Implemented an inherently parallel hyperparameter optimization approach called HyperSpace that simultaneously explores 20,480 deep learning models on Summit while optimizing model parameters that affect clinical performance.

	Titan	Summitdev	Summit
Platform	Cray XK 7	IBM Power8 S822LC	IBM Power9
# nodes	18,688 nodes	54 nodes	4,600 nodes
Specs	1 x K20 GPU	4 x P100 GPU	6 x V100 GPU
Time (single thread)	16.67 hours	1.67 hours	1.39 hours
Time (single, multi-GPU)	16.67 hours	0.73 hours	0.41 hours (FP) 0.25 hours (HP)
Hyperparameter optimization	266.67 hours	221.49 hours	6.56 hours (FP) 4.0 hours (HP)

Summary & Ongoing Challenges

- AI shows promise for automated information extraction from unstructured pathology reports to increase efficiency, data quality, and timeliness of cancer surveillance.
- Understanding the sources of AI errors is important for continuing improvement
  - More errors with low prevalence classes
  - Human ground truth presents limitations
- Collaboration across registries is essential to fully exploit the promise of AI
- Data sharing OR Model sharing?
  - Reliable de-identification of unstructured text reports is difficult
  - Need for privacy preserving AI solutions to handle data privacy and confidentiality restrictions
- Human-AI integration is an open-ended question
  - What is the most effective way to integrate AI in national cancer surveillance?
  - Is interpretability possible and/or important?
  - Model-level and case-level uncertainty quantification maybe helpful

Final Thoughts on AI for Cancer Registries

- Hope
  - The convergence of big data and AI will enable near real time cancer surveillance
- Hype
  - AI solutions are superior to collective intelligence of the experts
  - Practical translation of AI tools is straightforward
- Hard Truth
  - AI solutions have a single point of failure: data quality
  - Human-AI integration approach will impact real-world value
    - AI interpretability and (real-time) uncertainty quantification are important future directions
  - Vulnerability issues for AI models and AI users (cognitive hacking) are critical

The Team

- NCI SRP
  - Lynne Penberthy, Betsy Hsu, Serban Negoita
- SEER Registries
  - LA, KY, NJ, UT, Seattle, CA
- IMS
  - Linda Coyle, Jennifer Stevens, Scott Depoy
- Oak Ridge National Lab
  - Folami Alamudun. Devanshu Agarwal, Mohammed Alawad, Blair Christian, Ioana Dansiu, Shamimul Hasan, Shang Gao, Jacob Hinkle, Alina Peluso, Noah Schaefferkoetter, Hong-Jun Yoon, Todd Young,
- Los Alamos National Lab
  - Tanmoy Bhattacharya, Kumkum Ganguly, Nick Hengartner, Ben MacMahon, Jamal Mohd-Yusof

OAK RIDGE National Laboratory

ACKOWLEDGEMENTS

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.

The authors gratefully acknowledge the contributions of the state and regional cancer registry staffs for their work in collecting the data used in this study.


THANK YOU!!!
tourassig@ornl.gov