Registry of the Future
Surveillance in the Era of Emerging Technology and Precision Medicine

Tanmoy Bhattacharya
for the
DOE-NCI Pilot 3 team

June 22, 2017

Acknowledgements: team at LANL; Lynne Penberthy, Paul Fearn and team at NCI; Gina Tourassi and team at ORNL
Better technology: individual data

- **Discovery**
  - Proteogenomics
  - Imaging data
  - Clinical trials

- **Patient engaged Research**
  - Clinical Research
  - Observational studies

- **Surveillance Big Data Implementation research**
  - EHR, Lab Data, Imaging,
  - PROs, Smart Devices,
  - Decision Support

- **Well characterized research data sets**
- **Cancer cohorts**
- **Patient data**

**GDC**
Research information donor

**SEER/NPCR**
Real World Data to support Learning from details of every cancer patient
Need for NLP

• It is estimated that at least 65% of desired cancer clinical data elements come from unstructured text
• Similar analyses in many other domains found the number of data elements from unstructured sources to be anywhere from 45% to 80%
• Efforts to advance templated clinical notes are underway, changes in workflow to adopt structured reporting have been slow
• Increasing volume, detail, timeliness, quality, regulation...

Exascale Computing

• Increase in computing power allows many more routine tasks to be automated.
• Computers are reaching the power where they start doing ‘human-like’ tasks
  – Play games like chess and go
  – Recognize images and video
  – Classify texts
• Often beat humans at long tasks
• First progress expected in machine-assisted tasks: not in “complete automation”.
Expected progress in Automation

Grace et al., arXiv:1705.08807

It is very likely that within the next few years, computers can revolutionize routine surveillance and database maintenance tasks.
Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

DOE and NCI have partnered to advance exascale computing through cancer research

Four DOE labs: ANL, LLNL, LANL, ORNL
Integrated Precision Oncology

Crosscut: Integrated Precision and Predictive Oncology

Pilot 1
Pre-clinical Model Development
Aim 1: Predictive Models of Drug Response (signatures)
Aim 2: Uncertainty Quantification and Improved Experimental Design
Aim 3: Develop Hybrid Predictive Models

Pilot 2
RAS Therapeutic Targets
Aim 1: Adaptive time and length scaling in dynamic multi-scale simulations
Aim 2: Validated model for Extended RAS/RAS-complex interactions
Aim 3: Development of machine learning for dynamic model validation

Pilot 3
Precision Oncology Surveillance
Aim 1: Information Capture Using NLP and Deep Learning Algorithms
Aim 2: Information Integration and Analysis for extreme scale heterogeneous data
Aim 3: Modeling for patient health trajectories

Crosscut: Uncertainty Quantification (UQ) and CANDLE exascale technologies
Pilot 3: OVERVIEW

Population Information Integration, Analysis and Modeling

Improve the effectiveness of cancer treatment in the “real world” through computing

- **NCI SEER Database**
- **Electronic Medical Records/Claims**
- **Hospitals/Pharmacies**
- **Census**
- **Patient Generated Data**

**OMICS**
- Genomic
- Imaging

**Free text**
- Clinical Reports
- Temporal & Spatial Trajectory

**Input (n words, k features)**
- Fast Look-up for features
- Convolution network layers
- Unsupervised methods for POS Tagging, NER, SRL, SRW

**Deep Text Comprehension + Multi-Task Learning**
- Traditional NLP + Data Analytic/Machine Learning
  - POS Tagging
  - Named Entity Recognition (NER)
  - Semantic Role Labeling (SRL)
  - Semantically Related Words (SRW)

**Novel Data Analytic Techniques for Integration and Analysis**
- Graph Analytics
- In-memory Analytics
- Visual Analytics
- Uncertainty quantification

**Data-driven Integrated Modeling & Simulation for Precision Oncology**
- New Clinical Bio-markers
- Other patient-relevant data
- Results from Pilots 1 & 2
- Precision profile for patient/patient cohort
- In-silico clinical trials
- Clinical trial simulations on HPC for patient cohort – O(100K) individuals

Operated by Los Alamos National Security, LLC for NNSA
Information Extraction

ePath reports

• First task: extract information automatically from ePath reports, e.g.,
  – Site/Laterality/Grade
  – Histology
  – Behavior
  – Genetic Markers

• Machine learning works by learning a task
• Data hungry: lots of examples of both positive and negative kinds
• Looks only for what is trained on, does not attempt to “understand”
  – Can input relevant “knowledge” to help training

• Provide “Uncertainty Quantification” for registrars
  – Answers come with a measure of certainty (does it need manual review?)
  – Certain answers have error rates far lower than real-world human curated data
  – Goal is to allow machine-assisted manual extraction
• Three modes of learning
  – Unsupervised
  – Supervised
  – Semi-supervised
• Unsupervised gives relevant clusters
• The amount of data required much more for supervised learning
• Semi-supervised uses both clustering and labeling information
Shallow Learning

A traditional way of learning

• Extract relevant “features”
  – Organization, chapters and sections of the staging manual describe different types of cancer.
  – Similarity to such units are likely to be relevant features.

• Use the features to classify
  – If more similar to the “lung” chapter than to “breast” chapter, report probably about lung cancer

• Careful about order in which to use the features
  – Rule out metastasis before concluding that sample is from primary site.
Deep Learning

• Feature selection needed because a lot of data irrelevant to the question
  – English has a lot of articles (e.g., *a, an, the*) that may not help answer the question.
  – Feature selection to weed these out and identify significant groups (e.g., *lower left lobe*) of words.

• Feature selection by hand loses information residing in complex correlations or weak features that are only sometimes relevant.

• Deep learning uses all the data and usually processes in layers
  – Lower layers “learn” feature extraction
  – Higher layers “learn” classification

• Can use expert knowledge as hints into the input layer.
Example model for parsing text using Deep Learning

- Input text e.g., as a sequence of words.
- A few “layers” learn how to parse the text. These learn significant features in the input.
- Separate sets of “layers” extract specific information (e.g., site, laterality, grade) from parsed text. Output tested for mismatch with training data and pushed down to correct the layers.
- Sharing lower layers makes most use of the data since parsing may be better done simultaneously.
Annotation: where training data comes from
Current Status of NLP

- **Text Comprehension:**
  - Developed and benchmarked rule-based, conventional ML-based, and DL-based NLP tools for e-paths for three information extraction tasks: (i) primary cancer site, (ii) histological grade, (iii) behavior
    - DL interpretability
    - Cross-registry robustness validation
  - Established reproducible experimental design pipeline for future studies with new data and different tasks

- **Text Synthesis:**
  - Generative models for e-path text synthesis under development
  - Initial validation in progress

- **Current Status of NLP**
  - J. Qiu, H.-J. Yoon, P. Fearn, G.D. Tourassi. Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports. Submitted to Journal of Biomedical and Health Informatics (revision pending review)
  - J. Boten, P. Fearn, G. Tourassi, et.al. The Development of the Clinical Document Annotation and Processing Pipeline to Facilitate the Integration of Natural Language Processing within Cancer Surveillance. To be presented at the North American Association for Central Cancer Registries Annual Conference in June 2017
  - J. Boten, D. Rivera, M. Myneni, et. al. Leveraging Large-Scale Computing for Population Information Integration, Analysis, and Modeling. Submitted to American Medical Informatics Association (AMIA) 2017 Annual Symposium

- Synthesized Pathology Text
  - Bronchial Mucosa with Moderately Differentiated Squamous Cell Carcinoma.
  - Received in CytoLite solution labeled left mainstem bronchus biopsy are four cylindrical portions of white, soft tissue averaging 0.2 x 0.1 cm.
Progress in Computing

• **Multi-task (MTL) deep neural nets for biomarker extraction** – exploiting task relatedness for simultaneous training and information extraction from unstructured clinical text

![Task-level performance of MTL compared with standard techniques](image1.png)

![Scaling profile of time-to-solution on TITAN with Torch-MPI](image2.png)

![End-to-end visualization of “class-level” importance from documents](image3.png)

• **Recurrent neural nets for generating text with valid clinical context** – Long Short-term memory (LSTM) based approaches for generating synthetic text data to train scalable deep text comprehension algorithms
  
  • Baseline results are tested and scaling is in progress

• **Common deep learning environment for CANDLE testing/development:**
  
  • DD award of 3 million hours (CSC237) with access to DGX-1, summitdev, Titan, Rhea. Container porting on Titan expected to be completed in May 2017
  
  • **Container environment** (Theano/Tensorflow/Keras, mxnet, pytorch, Neon, Caffe)
Uncertainty Quantification

• Current data
  • Sparse
  • Opportunistic collection
  • Little “gold standard” ground truth
• Goal of uncertainty reduction can drive design of experiments to collect relevant data.
• Quantify uncertainty in machine learning and statistical conclusions
  • Average performance
  • Estimating confidence in individual predictions
  • Statistical errors and data biases
  • Model transfer uncertainty
• Allows human to focus on cases where machine is uncertain
  • Identify and remove routine tasks
  • Let expert systems and human experts handle the rest
How uncertainty quantification works

- Raw output of the classification process are scores for various answers
- Without UQ, choose the highest scoring answer
- UQ looks for
  - how high the highest score is
  - How low scores the other answers get
- Threshold can be adjusted to meet accuracy criterion.

| Report ID   | Lung | Breast | Colon | Prostate | ...
|-------------|------|--------|-------|----------|------
| CT-REC-XXXX | 0.68 | 0.12   | 0.09  | 0.02     | ...  |
| HI-REC-XXXX | 0.28 | 0.23   | 0.26  | 0.07     | ...  |
Example with shallow learning

- Classify primary site into five categories: breast, lung, colon, prostate, or other.
- Accuracy on the entire classification is about 92%.
- Look for false positives as a function of ‘confidence’
- Setting the thresholds to have undetectable false positives, classifies 48% (607 out of 1253) ePath reports.
- Human workload reduced by half
• 942 ePath reports were used.
  • reports about cancers other than breast and lung were left out.

• As with rule-based methods, false positives (i.e., incorrect confident classification) are rare!

• Setting the thresholds to the vertical red lines give undetectable false positive rates, and correctly identifies 770/951 (81%) of the breast/lung ePath reports.

• These results are optimistic because they were calculated over data used for training and only using breast and lung files.
Future: “better than human” performance

Module → Module → Module/Human

Extract Information → Evaluate confidence → Decide → Pass on

Decision
Future: Individual genomic data

Genomic data now critical to understand each patient’s cancer

The increased availability of targeted agents and proliferation of genomic testing of tumors represents a special challenge to registrars.

Many individual biomarkers - which may or may not be available to registrars in the EMR

Genomic panels can consist of hundreds of mutation tests with varying structure and actionable information.

It is not feasible for registrars to capture all these data so alternative methods must be considered.

Machine learning can help automate extraction of genomic data from various unstructured texts.
Future: Capturing outcomes other than survival, e.g., Recurrence

Identifying patients with distant recurrent disease is critical with >18 million cancer survivors for whom we cannot describe the risk of recurrence.

Identification of recurrence is complex.

It can be diagnosed via multiple methods which vary by:
- cancer site
- time to recurrence
- diagnosing physician type
  - primary care, oncologist, radiation oncologist, radiologist etc.
- diagnostic method (biopsy, imaging, serology)

Accurate measurement of recurrence requires capture of multiple layered, combined data sources and new methods (NLP) to provide comprehensive capture of recurrence(s).
Conclusions

• Exascale computation allows training of complex NLP modules that can extract sophisticated information with known uncertainty.

• Trained models are fast to execute and can be integrated into routine workflow to
  – Reduce human effort in a vast majority of cases
  – Select subset of human work for review

• NLP will allow extraction and integration of ever more complex data-sets.
END OF PRESENTATION
Text Comprehension

Three different DL architectures for text comprehension

Plug-and-Play Experimental Platform with interactive visualization

Task-level performance of DL compared with standard techniques

Text Synthesis

Two different synthetic text generation models (i) inductive modeling, (ii) DL-based

Synthetic pathology reports examples

Initial evaluation metrics
Current performance of site classification – I think this is too complicated for them.

Find probability of assignment to various categories based on density in semantic space.

Uncertainty of prediction given by entropy of this probability distribution.

Hierarchical classifier:

- Invoke expensive layer if uncertain
- Rule based system classifies about half the records

Setting the thresholds to the vertical red lines give no false positive values, and correctly identifies 607 out of 1253 ePath reports = 48%.

Accuracy on the cross-validation set is about 92%.
Moving Forward

Next steps

– AIM 1: Implement and benchmark semi-supervised DL using the new e-path corpus
  • Focus on automated capture of 5 key variables for top 5 cancer sites (Br, Lung, CRC, PrCa)
  • Focus on annotation of variables to support AIM 3 use cases (ER, PR, Her2, path defined recurrence)
– AIM 2: (a) Develop integrated data packages to provide initial resources for more comprehensive modeling of critical concepts (distant recurrence, response to initial and subsequent therapy) working with internal and external partners; (b) Construct patient trajectories using the available data sources and develop visual analytics to study the association between trajectory variations and health outcomes
– AIM 3: Leverage AIMs 1 and 2 targets (NLP captured data and linked data sets) to support development of recurrence modeling

Opportunities

– Integrating key clinical partners to most appropriately target needs from the clinical community (e.g. Allison Kurian)
– Acquisition of new data sources such as claims for under 65 GA, existing recurrence data from radiology reports (LA) to provide pilot data to begin building more complex models more rapidly

Challenges

– IP agreement across all Pilots for sharing of software, APIs, etc.
– Resources
Cross-cutting – Next Steps

• **Challenges**
  • Data generation/collection/curation takes time
  • May appear too late on time-line to be effective in pilot scope but demonstrate feasibility/necessity of experimental design.

• **Next steps**
  • Extend UQ to more realistic scenarios
    • Extend to Deep Learning based classification
    • Rademacher error estimation of classifiers
  • Close coupling between data and analysis teams
    • Pilot 3 Inter-aims discussion
  • Speculative collection efforts guided by preliminary UQ results
    • Pilot 2 experiments of model membranes with K-Ras by Frantz Jean-Francois, FNL containing 1–7 lipids identified unexpected strong interactions
    • Pilot 3 Aim 2 survey of relevant clinical variables based on literature survey of recurrence modeling.
  • Refine collection efforts as problem better quantified.
Pilot 3: AIM 3 - Progress Update

CISNET Pilot 1: 3D discrete event model of esophageal cancer growth and an HPC-enabled framework for automated discovery of optimal cancer screening protocols

Jim Nutaro, Ozgur Ozman (ORNL), Joey Kong (Mass General)

CISNET Pilot 2: HPC calibration of a CRC cancer natural history model with Incremental Mixture Importance Sampling

Jonathan Ozik, Nicholson Collier, Justin Wozniak (ANL), Carolyn Rutter, Jessica Hwang (Rand Corporation)

Aim 3 Short Term Planning

- Framing modeling use cases led by clinical experts
  - First use cases: breast cancer, colorectal cancer
  - Clinical Experts: Allison Kurian (Stanford), Marty Weiser (Memorial Sloan-Kettering)
- Identifying data sources and elements necessary for modeling use cases to coordinate with aims 1 & 2

- Used EMEWS* to execute a (8192 cores * 16 hours) ~ 130k core hours model calibration workflow on Cray XE6 Beagle at the University of Chicago, hosted at ANL
- Incremental mixture importance sampling (IMIS) of CRCSPIN model with 23k initial samples, and 20 iterations of 1k subsequent samplings
Governance Committee Advisement: Joint Working Group

Department of Energy

National Institutes of Health

National Nuclear Security Admin.

Office of Science

Advanced Scientific Computing Research

National Cancer Institute

Defense Programs

Advanced Scientific Computing Advisory Committee (ASCAC)

Defense Programs Advisory Committee (DPAC)

Public Records

Public Records

Public Records

More...

More...

National Cancer Advisory Board (NCAB)

Joint Working Group
MOU Directed Activity Status

The Parties intend to collaborate on the development of advanced computing solutions for cancer and related research activities primarily in the context of the PMI and NSCI.

Through this collaboration, the Parties intend to share best practices and know-how with each other, and to provide access to tools.

<table>
<thead>
<tr>
<th>Item</th>
<th>Action</th>
</tr>
</thead>
</table>
| (Section 1C) Designated contacts meet to review progress and address new opportunities for collaboration; technical and programmatic working groups to make formal recommendations. | - Monthly cross-collaboration pilot co-leads meetings and shared collaboration site established.  
- Established monthly technical progress reports. |
| (Section 1D) Establish a Technical Working Group (TWG) to evaluate progress; meet quarterly, or as needed, and will include membership from the Parties, outside experts as needed, and private sector collaborators that are involved with the activities. | - Governance Review Committee formed with first meeting held on Nov. 7, 2016.  
- Tentative schedule set for quarterly meetings – in person and via WebEx. |
| (Section 1E) Formalize agreements with non-Federal entities in writing. | - Arranging the coordination of collaboration of the State SEER Registries for access to clinical data. (Pilot 3) |
### MOU Directed Activity Status (cont.)

<table>
<thead>
<tr>
<th>Item</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Section 2B)</strong> There may be opportunities for independent collaborations and activities which are outside the scope.</td>
<td>- Continued engagement with CISNET introduction programs; exploring feasibility of future collaborations (Pilot 3).</td>
</tr>
</tbody>
</table>
| **(Section 2C)** Materials, data, or technology being analyzed/studied under the terms and conditions of this MOU may be shared between the Parties; such transfers will be carried out under separate, written agreements. | - The JDACS4C Multi-party Bi-lateral CDA/DUA was completed and signed by all parties March 8, 2017. The agreement authorizes all anticipated data sharing and collaboration between the DOE labs and Pilots in support of JDACS4C.  
- Oakridge established a centralized DOE IRB to cover all labs in Pilot 3 (March 2017).  
- Working through Data Use Agreements for transmittal and sharing of registry data and SEER*Medicare data (Pilot 3).  
- Lab legal and cyber teams are working through HIPAA compliance and connectivity. |
| **(Section 2D)** Software and related artifacts are expected to the extent possible to be made open source and made available. | - Created a shared JDACS4C repository (GitHub) for cross-collaboration development and release.  
- The default open source license for the JDACS4C repository will be the MIT license, which is the same license being used by the CANDLE effort. All non-compatible licenses will need to be presented to the IP Agreements team for review. |
| **(Section 2E)** Rights to intellectual property developed during the course of research under this MOU will be addressed. | - Established an integrated IP Management Team across labs and agencies.  
- JDACS4C IP Management Plan is currently in development by the IP Management Team. |
Pilot 3: Data Access and Hosting Update

- **IRBs**
  - IRBs for sharing of PHI data with DOE labs is approved
    - Includes: Path reports, claims, abstracts, radiology reports, etc.
  - 3 State Registries have received IRB approval for the sharing of clinical documents; 1 additional in progress

- **DUAs**
  - DUAs for acquisition and annotation of registry data in place with IMS and registries
  - DUAs for potentially identifiable datasets between registries, IMS, and DOE labs in progress
    - DUA from Louisiana and Kentucky currently being reviewed
    - DUA from SEER*Medicare being reviewed
    - Data to be shared:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Louisiana</th>
<th>Seattle</th>
<th>Georgia</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Path Reports</td>
<td>328,176</td>
<td>1,107,203</td>
<td>87,393</td>
<td>1,522,772</td>
</tr>
<tr>
<td>Abstracts</td>
<td>741,546</td>
<td>624,114</td>
<td>1,310,657</td>
<td>2,676,317</td>
</tr>
<tr>
<td># SEER Reportable Cases</td>
<td>511,892</td>
<td>838,978</td>
<td>940,261</td>
<td>2,291,131</td>
</tr>
</tbody>
</table>

Pilot 3: Community Engagement

- Clinical experts collaborating with pilot to identify relevant data elements and provide content expertise for modeling use cases
- Continuing stakeholder engagement for the following data types: Pharmacy, Claims, Radiology, Pathology, Genomic, Environmental
Pilot 3: Dissemination

- **Tools / Software:** Established an agile process of iterating deliverables to IMS, which are machine learning backed API, distributed as a docker image, an agnostic, flexible solution suitable for a production software development environment

- **Publications**
    - *J. Qiu, H.-J. Yoon, P. Fearn, G.D. Tourassi. Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports. Submitted to Journal of Biomedical and Health Informatics (revision pending review)*
  - 1 conference paper published (2/2017)
  - 2 abstracts submitted
    - *J. Boten, P. Fearn, G. Tourassi, et.al. The Development of the Clinical Document Annotation and Processing Pipeline to Facilitate the Integration of Natural Language Processing within Cancer Surveillance. To be presented at the North American Association for Central Cancer Registries Annual Conference in June 2017*
    - *J. Boten, D. Rivera, M. Myneni, et. al. Leveraging Large-Scale Computing for Population Information Integration, Analysis, and Modeling. Submitted to American Medical Informatics Association (AMIA) 2017 Annual Symposium*