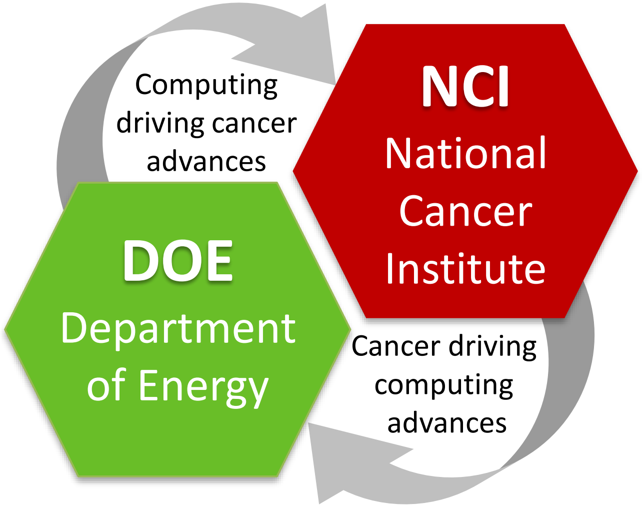
Pilot 3: Population Information Integration, Analysis and Modeling

Paul Fearn
National Cancer Institute

Georgia Tourassi
Oak Ridge National Laboratory

October 18, 2017

Presented at:
Frontiers of Predictive Oncology 2017



Computing driving cancer advances

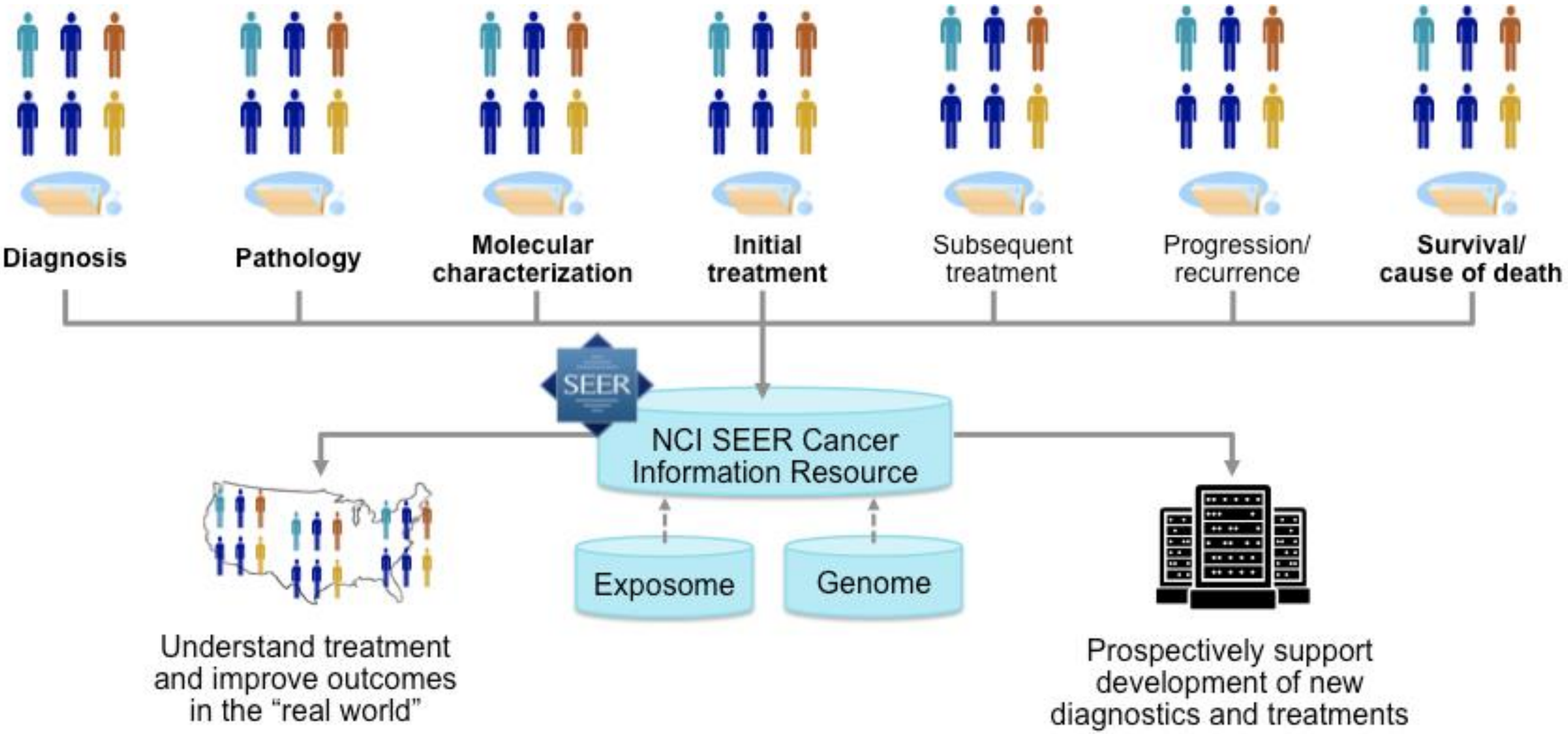
NCI
National Cancer Institute

DOE
Department of Energy

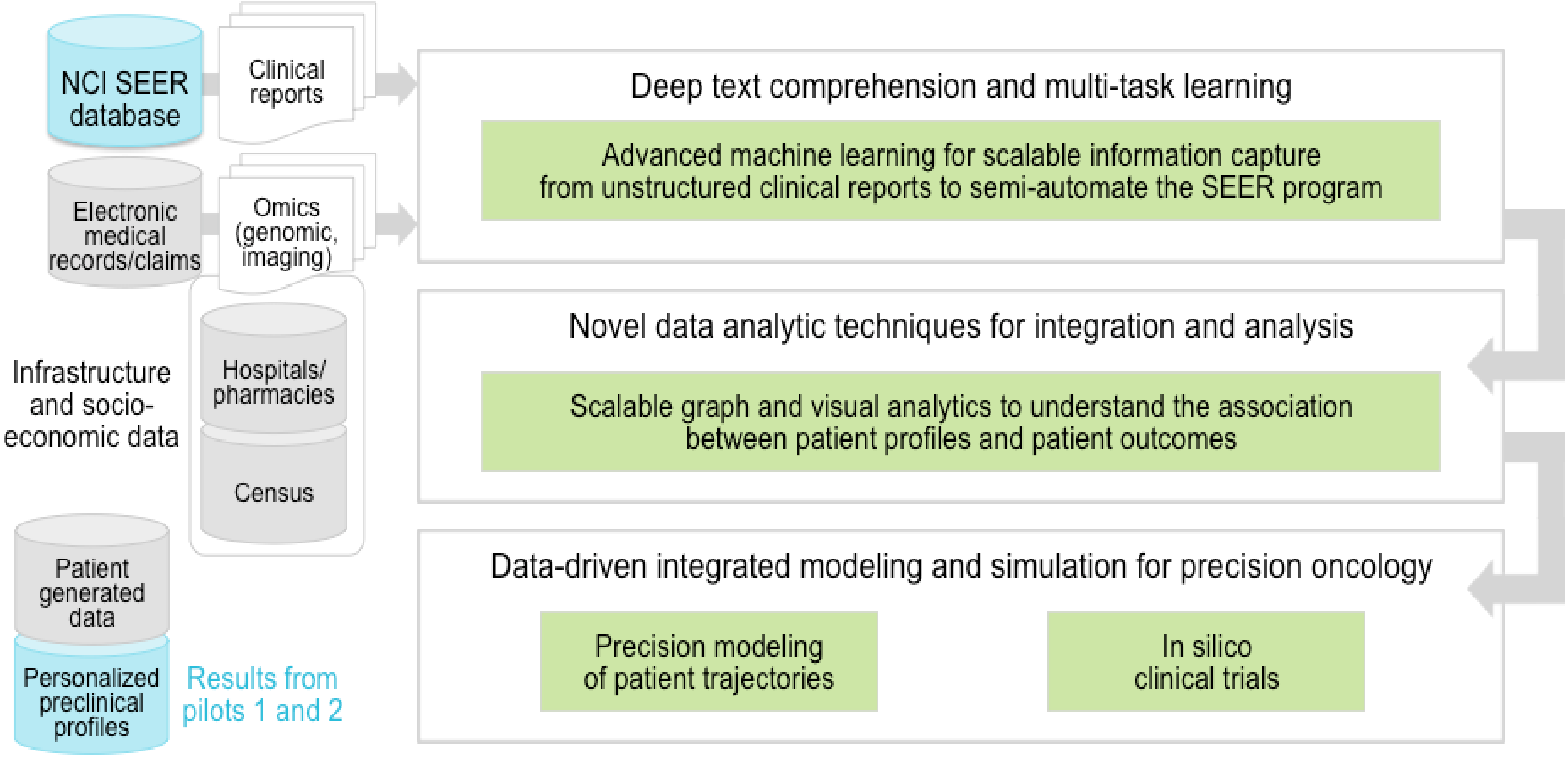
Cancer driving computing advances

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

Improve the effectiveness of cancer treatment in the “real world” through automation: Surveillance Perspective


Diagnosis
Pathology
Molecular characterization
Initial treatment
Subsequent treatment
Progression/ recurrence
Survival/ cause of death
SEER
NCI SEER Cancer Information Resource
Exposome
Genome
Understand treatment and improve outcomes in the “real world”
Prospectively support development of new diagnostics and treatments

Pilot 3: Aims and Technical Overview



NCI SEER database

Clinical reports

Electronic medical records/claims

Omics (genomic, imaging)

Infrastructure and socio-economic data

Hospitals/ pharmacies

Census

Patient generated data

Personalized preclinical profiles

Results from pilots 1 and 2

Deep text comprehension and multi-task learning

Advanced machine learning for scalable information capture from unstructured clinical reports to semi-automate the SEER program

Novel data analytic techniques for integration and analysis

Scalable graph and visual analytics to understand the association between patient profiles and patient outcomes

Data-driven integrated modeling and simulation for precision oncology

Precision modeling of patient trajectories

In silico clinical trials

Pilot 3

Multi-disciplinary DOE-NCI team w/ clinical & industry partners

NCI

- Jessica Boten
- Paul Fearn
- Rocky Feuer
- Steve Friedman
- Usman Khalid
- Marina Matatova
- Mita Myneni
- **Lynne Penberthy**
- Donna Rivera

SEER

- Eric Durbin (KY)
- Steve Schwartz (Seattle)
- Kevin Ward (GA)
- Xiao-Cheng Wu (LA)

IMS

- Linda Coyle
- Dave Annett
- Nicola Schussler

ORNL

- Folami Alamudun
- Blair Christian
- Shang Gao
- Joe Lake
- Jim Nutaro
- Arvind Ramanathan
- Larry Roberts
- Ozgur Ozmen
- John Qiu
- Johnny Tolliver
- **Georgia Tourassi**
- Hong-Jun Yoon

ANL

- **Tom Brettin**
- Jonathan Ozik

LANL

- **Tanmoy Bhattacharya**
- Kumkum Ganguly
- Nicholas Hengartner
- Benjamin MacMahon
- Sarah Michalak

LLNL

- **Ana Paula De Oliveira Sales**
- Priyadip Ray
- Braden Soper

CLINICAL COLLABORATORS

- Jessica Chubak (Kaiser Permanente)
- Mike Hassett (DF/Harvard)
- Allison Kurian (Stanford)
- Rod Schmidt (U Wash)
- Marty Weiser (MSK)
- Deb Schrag (DF/Harvard)

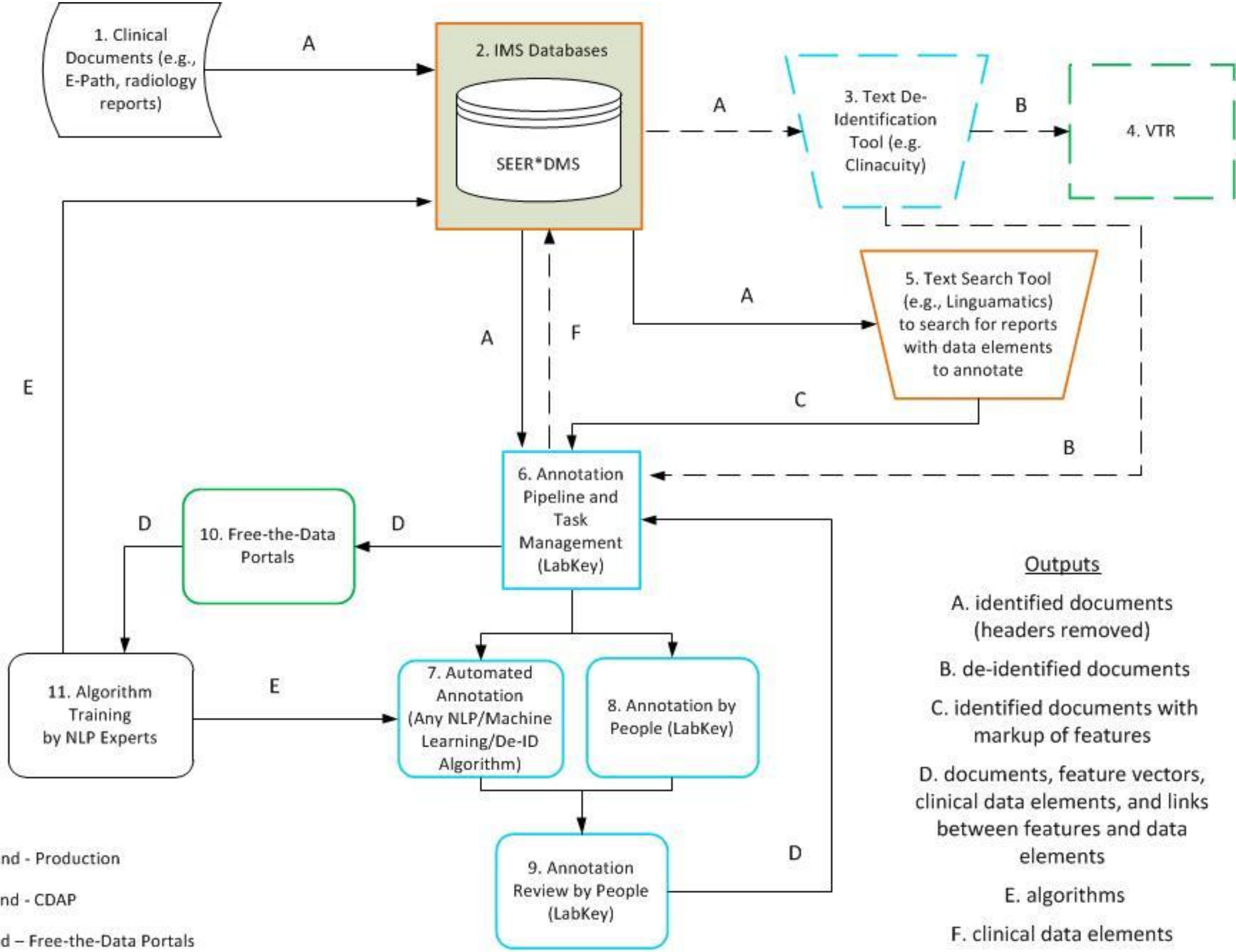
Pilot 3

Update – Aim 1: Data access and Annotation Pipeline

- Access to Louisiana registry data
  - ❖ 105,523 patients
  - ❖ 110,941 cancer diagnoses
  - ❖ 256,816 path reports associated with those diagnoses
- 3 registries have received IRB approval: LA, Seattle, KY; pending: GA
- 1,800 pathology reports annotated for ALK, EGFR by Vasta
- Schema for breast cancer biomarkers and recurrence being finalized (HER2, ER, PR, Neu, distant recurrence)
  - ❖ Use cases for breast recurrence developed and in pipeline
- NCI Investment for annotation pipeline
  - – Enhancements for LabKey
  - – Scaling up of Annotation services (Vasta)

Clinical Document Annotation Pipeline

- Infrastructure to support annotation of unstructured text documents for testing and validation of NLP algorithms
- Represents a critical platform for NLP- large volumes of gold standard annotated data are essential
- Infrastructure will be available to all Federal agencies and their partners for use in annotation for testing of algorithms


1. Clinical Documents (e.g., E-Path, radiology reports)
A
2. IMS Databases
SEER*DMS
A
3. Text De-Identification Tool (e.g. Clinacuity)
B
4. VTR
A
5. Text Search Tool (e.g., Linguamatics) to search for reports with data elements to annotate
A
F
C
B
E
6. Annotation Pipeline and Task Management (LabKey)
D
10. Free-the-Data Portals
D
11. Algorithm Training by NLP Experts
E
7. Automated Annotation (Any NLP/Machine Learning/De-ID Algorithm)
8. Annotation by People (LabKey)
9. Annotation Review by People (LabKey)
D

Outputs

A. identified documents (headers removed)

B. de-identified documents

C. identified documents with markup of features

D. documents, feature vectors, clinical data elements, and links between features and data elements

E. algorithms

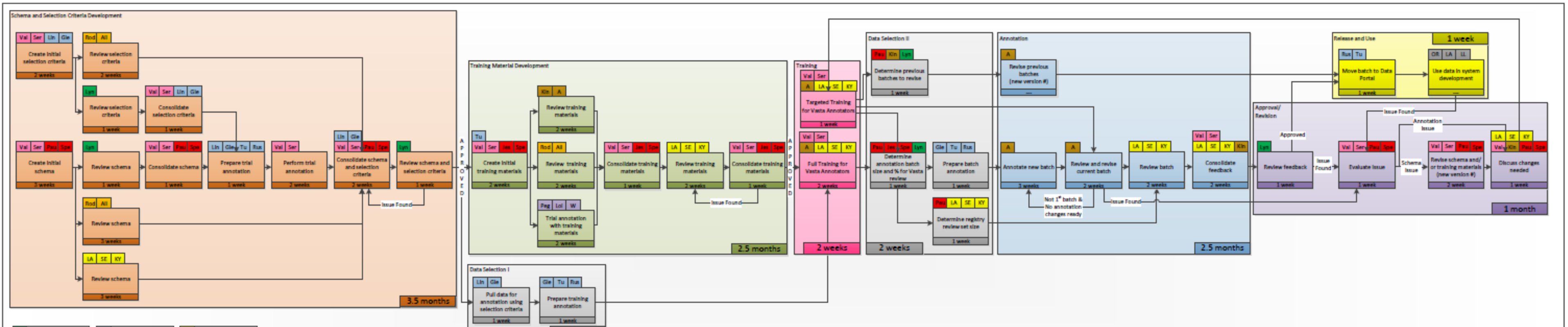
F. clinical data elements

🟧 Registry Island - Production

🟦 Registry Island - CDAP

🟩 Shared Island – Free-the-Data Portals

Complex Annotation Workflow



Schema and Selection Criteria Development

Val Ser Lin Gle
Create initial selection criteria
2 weeks

Rod All
Review selection criteria
2 weeks

Lyn
Review selection criteria
1 week

Val Ser Lin Gle
Consolidate selection criteria
1 week

Val Ser Pau Spe
Create initial schema
3 weeks

Lyn
Review schema
1 week

Val Ser Pau Spe
Consolidate schema
1 week

Lin Gle Tu Rus
Prepare trial annotation
1 week

Val Ser
Perform trial annotation
2 weeks

Lin Gle
Val Ser Pau Spe
Consolidate schema and selection criteria
2 weeks

Lyn
Review schema and selection criteria
1 week

Issue Found

Rod All
Review schema
3 weeks

LA SE KY
Review schema
3 weeks

3.5 months

APPROVED

Training Material Development

Tu
Val Ser Jes Spe
Create initial training materials
2 weeks

Kin A
Review training materials
2 weeks

Rod All
Review training materials
2 weeks

Peg Lol W
Trial annotation with training materials
2 weeks

Val Ser Jes Spe
Consolidate training materials
1 week

LA SE KY
Review training materials
2 weeks

Val Ser Jes Spe
Consolidate training materials
1 week

Issue Found

2.5 months

Data Selection I

Lin Gle
Pull data for annotation using selection criteria
1 week

Gle Tu Rus
Prepare training annotation
1 week

2 weeks

APPROVED

Training

Val Ser
A LA SE KY
Targeted Training for Vasta Annotators
1 week

Val Ser
A LA SE KY
Full Training for Vasta Annotators
2 weeks

2 weeks

Data Selection II

Pau Kin Lyn
Determine previous batches to revise
1 week

Pau Jes Spe Lyn
Determine annotation batch size and % for Vasta review
1 week

Gle Tu Rus
Prepare batch annotation
1 week

Pau LA SE KY
Determine registry review set size
1 week

2 weeks

Annotation

A
Revise previous batches (new version #)

A
Annotate new batch
3 weeks

A
Review and revise current batch
2 weeks

Not 1st batch & No annotation changes ready

Issue Found

LA SE KY
Review batch
2 weeks

Val Ser
LA SE KY Kin
Consolidate feedback
2 weeks

2.5 months

Release and Use

1 week

Rus Tu
Move batch to Data Portal
1 week

OR LA LL
Use data in system development

Approval/ Revision

Lyn
Review feedback
1 week

Approved

Issue Found

Val Ser Pau Spe
Evaluate issue
1 week

Issue Found

Annotation Issue

Schema Issue

Val Ser Pau Spe
Revise schema and/ or training materials (new version #)
2 week

LA SE KY
Val Kin Pau Spe
Discuss changes needed
1 week

1 month

Lyn	Lynne Penberthy	Lin	Linda Coyle	LA	Louisiana Registry Staff
Pau	Paul Fearn	Gle	Glenn Abastillas	SE	Seattle Registry Staff
Jes	Jessica Boten	Tu	Tumen Tumurchudur	KY	Kentucky Registry Staff
Spe	Spencer Morris	Rus	Rusty Shields	OR	Oak Ridge NL
Val	Valentina Petkov	Rod	Rod Schmidt	LA	Los Alamos NL
Ser	Serban Negoita	All	Allison Kurian	LL	Lawrence Livermore NL
Roc	Rocky Feuer	Mar	Marty Wiser		
Peg	Peggy Adamo	Kin	Kinjal Vora		
Lol	Lois	A	Vasta Annotators		
W	Westat CTRs				

Revision loop: ~3.5 months

From finalized schema, no revision loop: ~6 months

Full Process, no revision loop: ~9-10 months

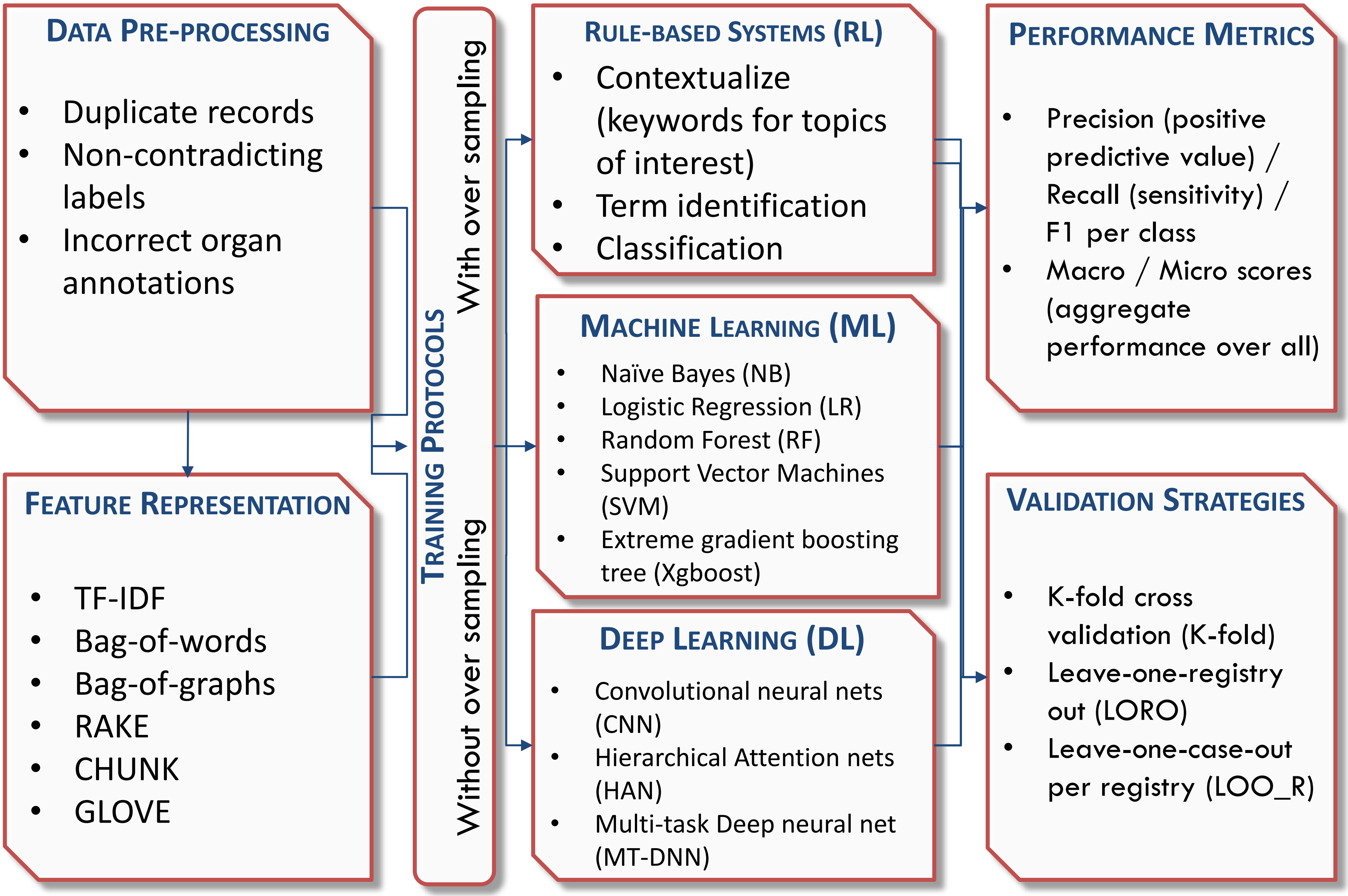
From Spencer Morris

Update – Aim 1: NLP tools

USE CASE 1: Limited dataset of annotated breast and lung cancer pathology reports from 5 different US states

USE CASE 2: Large dataset of pathology reports from Louisiana Cancer Registry

Experimental Pipeline



Data Pre-processing

- Duplicate records
- Non-contradicting labels
- Incorrect organ annotations

Feature Representation

- TF-IDF
- Bag-of-words
- Bag-of-graphs
- RAKE
- CHUNK
- GLOVE

Training Protocols

With over sampling

Without over sampling

Rule-based Systems (RL)

- Contextualize (keywords for topics of interest)
- Term identification
- Classification

Machine Learning (ML)

- Naïve Bayes (NB)
- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machines (SVM)
- Extreme gradient boosting tree (Xgboost)

Deep Learning (DL)

- Convolutional neural nets (CNN)
- Hierarchical Attention nets (HAN)
- Multi-task Deep neural net (MT-DNN)

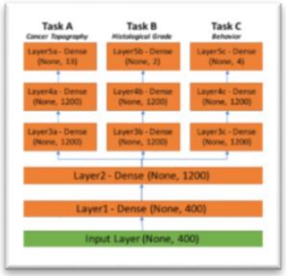
Performance Metrics

- Precision (positive predictive value) / Recall (sensitivity) / F1 per class
- Macro / Micro scores (aggregate performance over all)

Validation Strategies

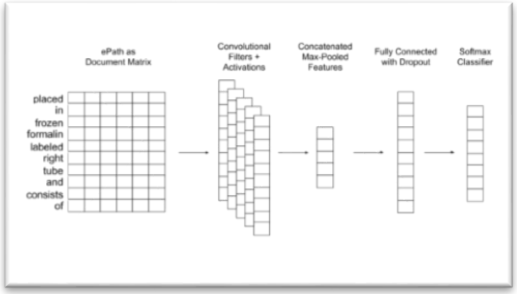
- K-fold cross validation (K-fold)
- Leave-one-registry out (LORO)
- Leave-one-case-out per registry (LOO_R)

Preliminary Investigation on the limited dataset


Task A
Cancer Topography
Task B
Histological Grade
Task C
Behavior
Layer5a - Dense (None, 13)
Layer5b - Dense (None, 2)
Layer5c - Dense (None, 4)
Layer4a - Dense (None, 1200)
Layer4b - Dense (None, 1200)
Layer4c - Dense (None, 1200)
Layer3a - Dense (None, 1200)
Layer3b - Dense (None, 1200)
Layer3c - Dense (None, 1200)
Layer2 - Dense (None, 1200)
Layer1 - Dense (None, 400)
Input Layer (None, 400)

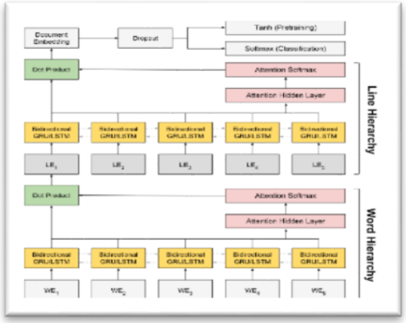
Multi-task Learning Deep Neural Network

"Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports." In INNS Conference on Big Data [2016]



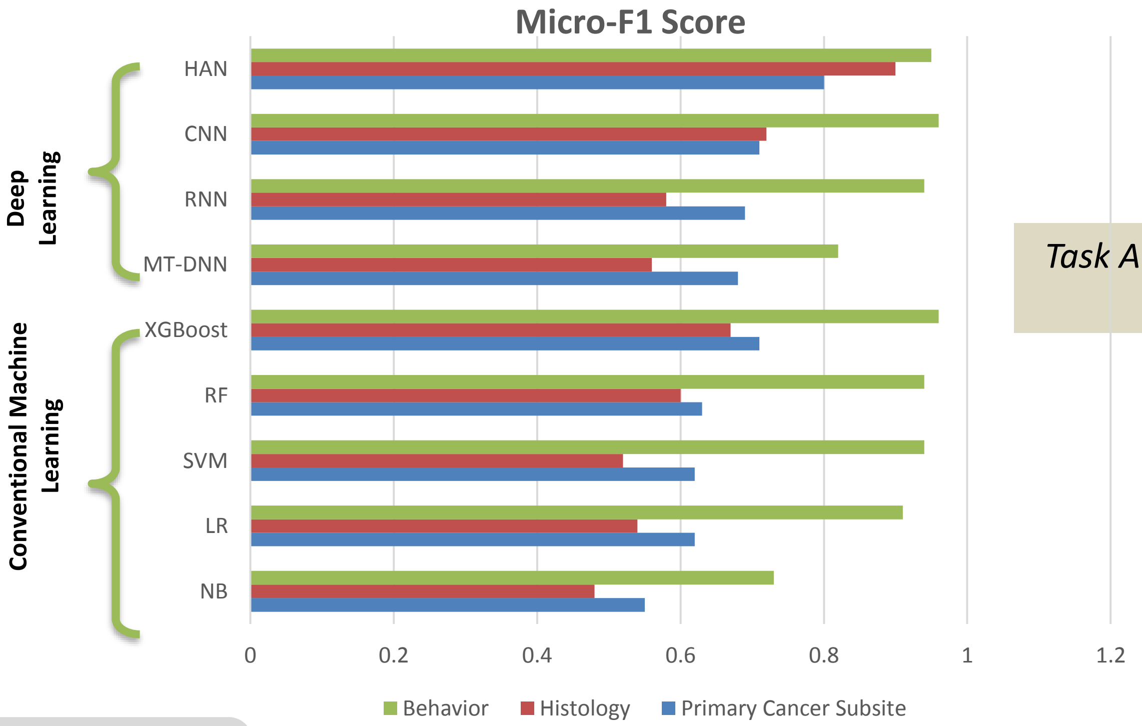
Convolutional Neural Network

"Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports," IEEE Journal of Biomedical and Health Informatics [2017]



Hierarchical Attention Network

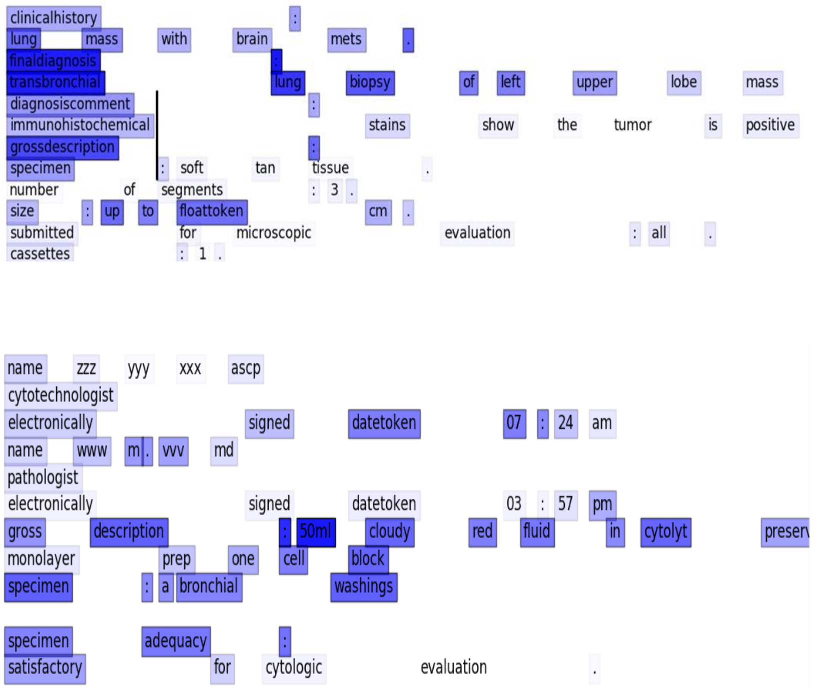
Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports," Journal of American Medical Informatics Association [2017]



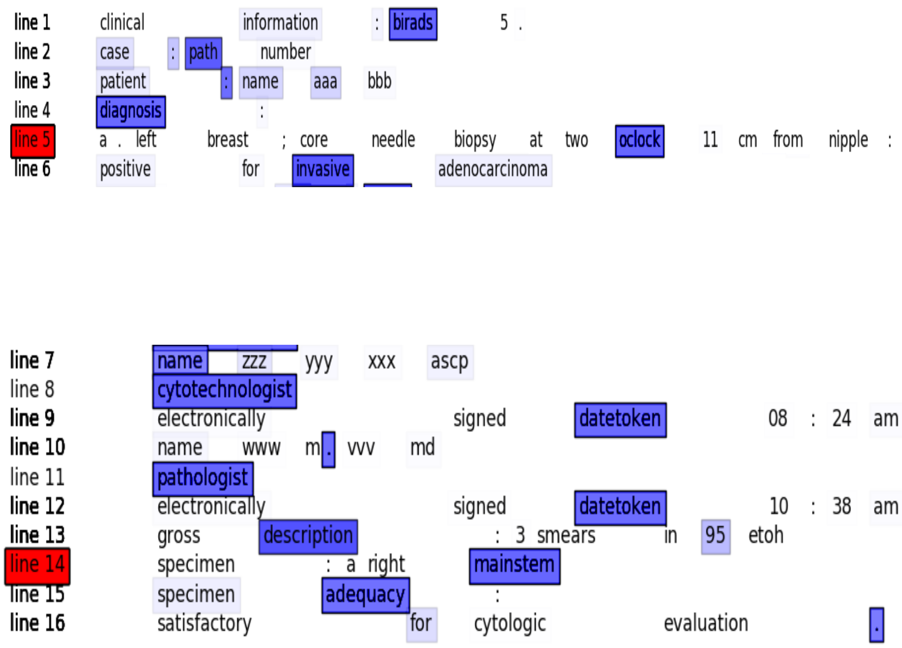
Task Accuracy Performance

Interpretability

CNN



HAN



CNNs associate context with importance based on how often words occur in its neighborhood. Moving along a row, these words may not always capture the required clinical context.

HANs interpret context based on most important words in a sentence → sentences → document. Neighboring words/sentences provide overall importance.

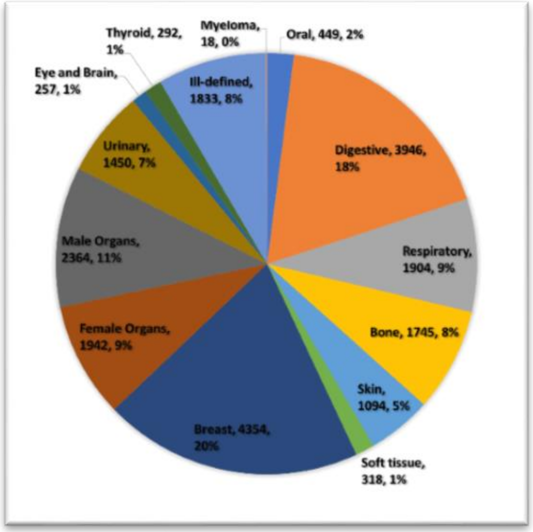
Initial observations with the Louisiana registry data

- 256,816 e-paths total
  - Preliminary experiments with ~5-10% of the path reports

- CNNs for 5 NLP tasks using 10-fold CV and hyper-parameter optimization
  - Primary cancer site
  - Laterality
  - Histology
  - Behavior
  - Grade

- Comparison w/ best performing shallow machine learning

Preliminary results with the LA registry data and Convolutional Neural Network



256,816 e-paths total

26,360 annotated for cancer subsite with >10 cases/subsite

20% reserved for final validation

21,966 cases used for CV

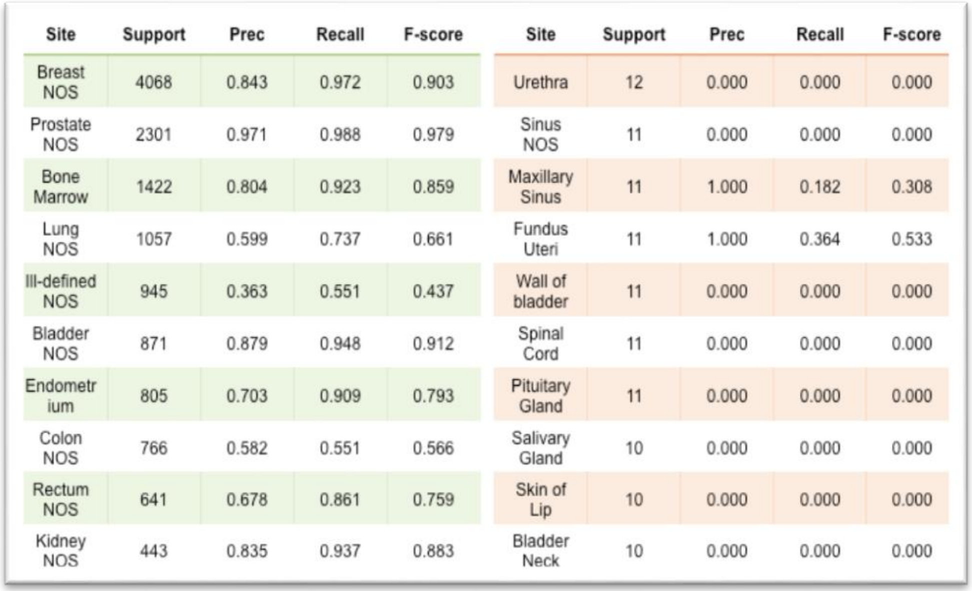
135 classes present

Experiment: 10-fold CV with CNN

# of trainable CNN parameters: 5,483,835

Micro-F1 = 0.71

Subsite Support Size vs. Accuracy



Site	Support	Prec	Recall	F-score	Site	Support	Prec	Recall	F-score
Breast NOS	4068	0.843	0.972	0.903	Urethra	12	0.000	0.000	0.000
Prostate NOS	2301	0.971	0.988	0.979	Sinus NOS	11	0.000	0.000	0.000
Bone Marrow	1422	0.804	0.923	0.859	Maxillary Sinus	11	1.000	0.182	0.308
Lung NOS	1057	0.599	0.737	0.661	Fundus Uteri	11	1.000	0.364	0.533
Ill-defined NOS	945	0.363	0.551	0.437	Wall of bladder	11	0.000	0.000	0.000
Bladder NOS	871	0.879	0.948	0.912	Spinal Cord	11	0.000	0.000	0.000
Endometrium	805	0.703	0.909	0.793	Pituitary Gland	11	0.000	0.000	0.000
Colon NOS	766	0.582	0.551	0.566	Salivary Gland	10	0.000	0.000	0.000
Rectum NOS	641	0.678	0.861	0.759	Skin of Lip	10	0.000	0.000	0.000
Kidney NOS	443	0.835	0.937	0.883	Bladder Neck	10	0.000	0.000	0.000

	$\text{CNN}_{output}>=\theta$			$\text{CNN}_{output}<\theta$		
θ	Support	TP	Accuracy	Support	TP	Accuracy
0	21966	15782	0.718			
0.2	21220	15674	0.739	746	108	0.145
0.4	19557	15232	0.779	2409	550	0.228
0.6	17572	14459	0.823	4394	1323	0.301
0.8	15627	13434	0.860	6339	2348	0.370
0.9	14276	12612	0.883	7690	3170	0.412
0.95	13210	11898	0.901	8756	3884	0.444
0.99	11143	10364	0.930	10823	5418	0.501
0.99999	4378	4299	0.982	17588	11483	0.653

Primary Cancer Site

Name	ICD-O-3 codes	# cases
Bladder	C67	947
Breast	C50	4,414
Colorectal	C18, C19, C20, C21	2,788
Endometrial	C53, C54, C55, C56, C57, C58	1,899
Kidney	C64	458
Leukemia	C42	1,800
Lung	C34	1,569
Lymphoma	C77	741
Melanoma	C44, C51, C60, C63	1,272
Other		4,324
Pancreatic	C25	151
Prostate	C61	2,313
Thyroid	C73	305

Total: 22981

F1 Scores

	CNN	RF
Micro F1	0.9128	0.8583
Macro F1	0.8941	0.8116

Confusion Matrix

908	0	1	4	7	3	0	0	0	16	0	8	0
1	4283	2	15	0	5	9	15	6	76	0	0	2
2	3	2648	22	2	4	12	1	6	88	0	0	0
0	9	23	1795	0	3	4	5	3	57	0	0	0
5	0	1	0	428	1	5	2	0	16	0	0	0
1	3	2	2	2	1704	2	31	9	41	0	2	1
0	9	6	2	2	6	1432	15	1	90	1	2	3
1	48	7	4	4	43	24	468	15	99	3	15	10
0	20	10	2	0	4	4	9	1133	84	0	4	2
36	140	133	82	32	39	122	101	100	3487	30	9	13
0	0	3	0	1	0	0	1	0	29	117	0	0
10	0	3	1	2	5	1	1	1	7	0	2281	1
0	0	0	0	0	0	1	5	1	4	0	0	294

$CNN_{output}>=\theta$

Threshold	Support	PPV
0	22981	0.913
0.2	22978	0.913
0.4	22890	0.915
0.6	22082	0.930
0.8	20745	0.949
0.9	19589	0.961
0.95	18346	0.971
0.96	17941	0.973
0.97	17352	0.976
0.98	16493	0.981
0.99	14808	0.986

Primary Cancer Site

BLADDER	PPV	94
	S	96
	F	95
BREAST	PPV	95
	S	97
	F	96
COLORECTAL	PPV	93
	S	95
	F	94
ENDOMETRIAL	PPV	93
	S	95
	F	94
KIDNEY	PPV	89
	S	93
	F	91
LEUKEMIA	PPV	94
	S	95
	F	94

LUNG	PPV	89
	S	91
	F	90
LYMPHOMA	PPV	72
	S	63
	F	67
MELANOMA	PPV	89
	S	89
	F	89
OTHER	PPV	85
	S	81
	F	83
PANCREATIC	PPV	77
	S	77
	F	77
PROSTATE	PPV	98
	S	99
	F	98
THYROID	PPV	90
	S	96
	F	93

Laterality

Code	Description	# cases
0	Not a paired site	1,432
1	Right: origin of primary	2,036
2	Left: origin of primary	1,926
4	Bilateral	44
5	Paired site: midline tumor	12
9	Paired site, but no information	256

Total: 5706

F1 Scores

	CNN	RF
Micro F1	0.8747	0.7625
Macro F1	0.5166	0.4460

Confusion Matrix

1292	56	55	0	0	29
59	1876	91	0	0	10
55	113	1752	0	0	5
16	12	15	1	0	0
4	3	4	0	0	1
97	46	44	0	0	69

$\mathrm{CNN}_{output}>=\theta$

Threshold	SUPPORT	PPV
0	5705	0.875
0.2	5705	0.875
0.4	5612	0.885
0.6	5052	0.925
0.8	4505	0.953
0.9	4070	0.968
0.95	3676	0.977
0.96	3534	0.979
0.97	3322	0.983
0.98	2973	0.986
0.99	2206	0.991

Laterality

NOT A PAIRED SITE	PPV	85
	S	90
	F	87
RIGHT:ORIGIN OF PRIMARY	PPV	89
	S	92
	F	91
LEFT: ORIGIN OF PRIMARY	PPV	89
	S	91
	F	90
BILATERAL	PPV	100
	S	2
	F	4
PAIRED SITE: MIDLINE TUMOR	PPV	*
	S	0
	F	*
PAIRED SITE, BUT NO INFORMATION	PPV	61
	S	27
	F	37

Behavior

Code	Description	# cases
0	Benign	735
1	Borderline malignancy	158
2	In situ	1,000
3	Malignant	11,751
6	Only Malignant 2010+	112

Total: 13756

F1 Scores

	CNN	RF
Micro F1	0.9264	0.8979
Macro F1	0.6574	0.5010

Confusion Matrix

458	13	22	238	4
49	22	4	83	0
13	0	739	248	0
117	6	161	11450	15
3	1	2	57	49

$\text{CNN}_{\text{output}} >= \theta$

Threshold	SUPPORT	PPV
0	13754	0.925
0.2	13754	0.925
0.4	13712	0.926
0.6	13149	0.942
0.8	12163	0.962
0.9	11177	0.974
0.95	10131	0.983
0.96	9743	0.984
0.97	9149	0.987
0.98	8236	0.989
0.99	6392	0.992

Behavior

BENIGN	PPV	72
	S	62
	F	67
BORDERLINE MALIGNANCY	PPV	52
	S	14
	F	22
IN SITU	PPV	80
	S	74
	F	77
MALIGNANT	PPV	95
	S	97
	F	96
ONLY MALIGNANT 2010+	PPV	72
	S	44
	F	54

Histology

Total: 14173

Code	Description	# cases
8140	Adenocarcinoma	4,469
8500	Ductal Carcinoma	1,484
8070	Squamous Cell Carcinoma	949
8010	Carcinoma in situ	937
8000	Neoplasm, malignant	869
8720	Melanoma in situ	567
8120	Transitonal cell carcinoma	417
8312	Clear cell adenocarcinoma	291
9590	Malignant lymphoma	209
8130	Papillary trans. Cell carcinoma	192
Total 87 classes		

73% of cases distributed among 10 out of 87 classes

F1 Scores

	CNN	RF
Micro F1	0.7922	0.6946
Macro F1	0.4893	0.3113

$\text{CNN}_{\text{output}}>=\theta$

Threshold	SUPPORT	PPV
0	14173	0.792
0.2	13914	0.802
0.4	13023	0.833
0.6	11489	0.874
0.8	9733	0.911
0.9	8226	0.935
0.95	7063	0.951
0.96	6744	0.956
0.97	6320	0.961
0.98	5812	0.967
0.99	4955	0.974

Histologic Grade

Code	Description	# cases
1	Well differentiated	220
2	Moderately differentiated	473
3	Poorly differentiated	367
4	undifferentiated	19
6	t-cell; t-precursor	50
9	Unknown	1,173

Total: 2302

F1 Scores

	CNN	
Micro F1	0.8240	
Macro F1	0.5980	

Confusion Matrix

163	24	12	0	0	21
16	385	22	0	1	49
4	40	272	0	0	51
0	1	6	0	0	12
1	0	0	0	14	35
19	44	38	0	9	1063

$\text{CNN}_{\text{output}} >= \theta$

Threshold	SUPPORT	PPV
0	2302	0.824
0.2	2302	0.824
0.4	2248	0.835
0.6	1986	0.875
0.8	1618	0.916
0.9	1282	0.934
0.95	1008	0.947
0.96	931	0.951
0.97	830	0.955
0.98	682	0.965
0.99	483	0.979

Histologic Grade

WELL DIFFERENTIATED	PPV		80
	S		74
	F		77
MODERATELY DIFFERENTIATED	PPV		78
	S		81
	F		91
POORLY DIFFERENTIATED	PPV		80
	S		74
	F		76
UNDIFFERENTIATED	PPV	X	
	S		0
	F	X	
T-CELL; T-PRECURSOR	PPV		58
	S		28
	F		38
UNKNOWN	PPV		86
	S		91
	F		88

Training Requirements

- Training a single task CNN with 250,000 path reports requires ~23.9 hours on NVIDIA P100 GPU
- At least 350 trials to obtain optimal hyper-parameter set
- Approximately 1,750 machine days required to complete the 5 NLP tasks

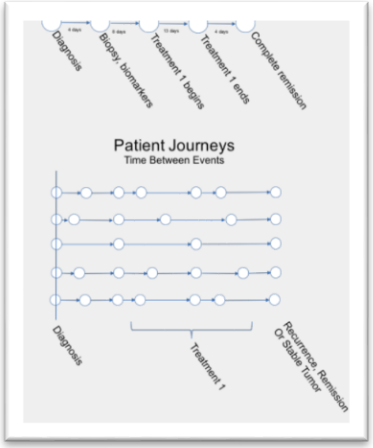
	Baseline	DGX-1	Amazon AWS Cloud	Titan	Summitdev
Platform Specs	1 x P100 GPU	8 x V100 GPU	P2, 16 nodes 8 x K80 GPU	18,688 nodes 1 x K20 GPU	4,600 nodes 6 x V100 GPU
Time	1,750 days	90.8 days	23.24 days	2.7 days	4.15 hours

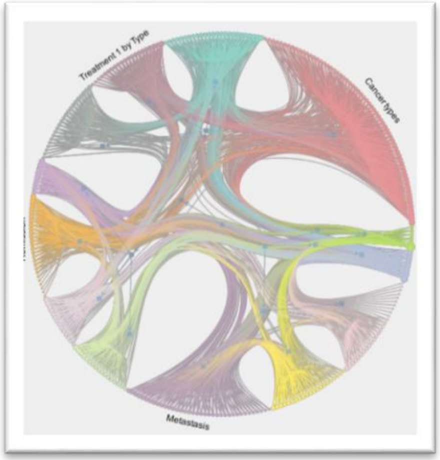
Summary & Conclusions

- **Deep learning for clinical NLP**
  - offers competitive and often state-of-the-art performance
  - CNNs are scalable and effective
  - HANs provide best performance but at the expense of scalability
  - Multi-task learning can exploit task relatedness and provide better results
- **Next steps with DL development**
  - Handling heavily imbalanced datasets
  - Multi-task learning with CNNs and HANs
  - Semi-supervised learning
- **Next steps with clinical translation**
  - Integrate DL NLP tools with prediction-level UQ
  - Address human factor engineering issues

Next Steps for Aims 2-3

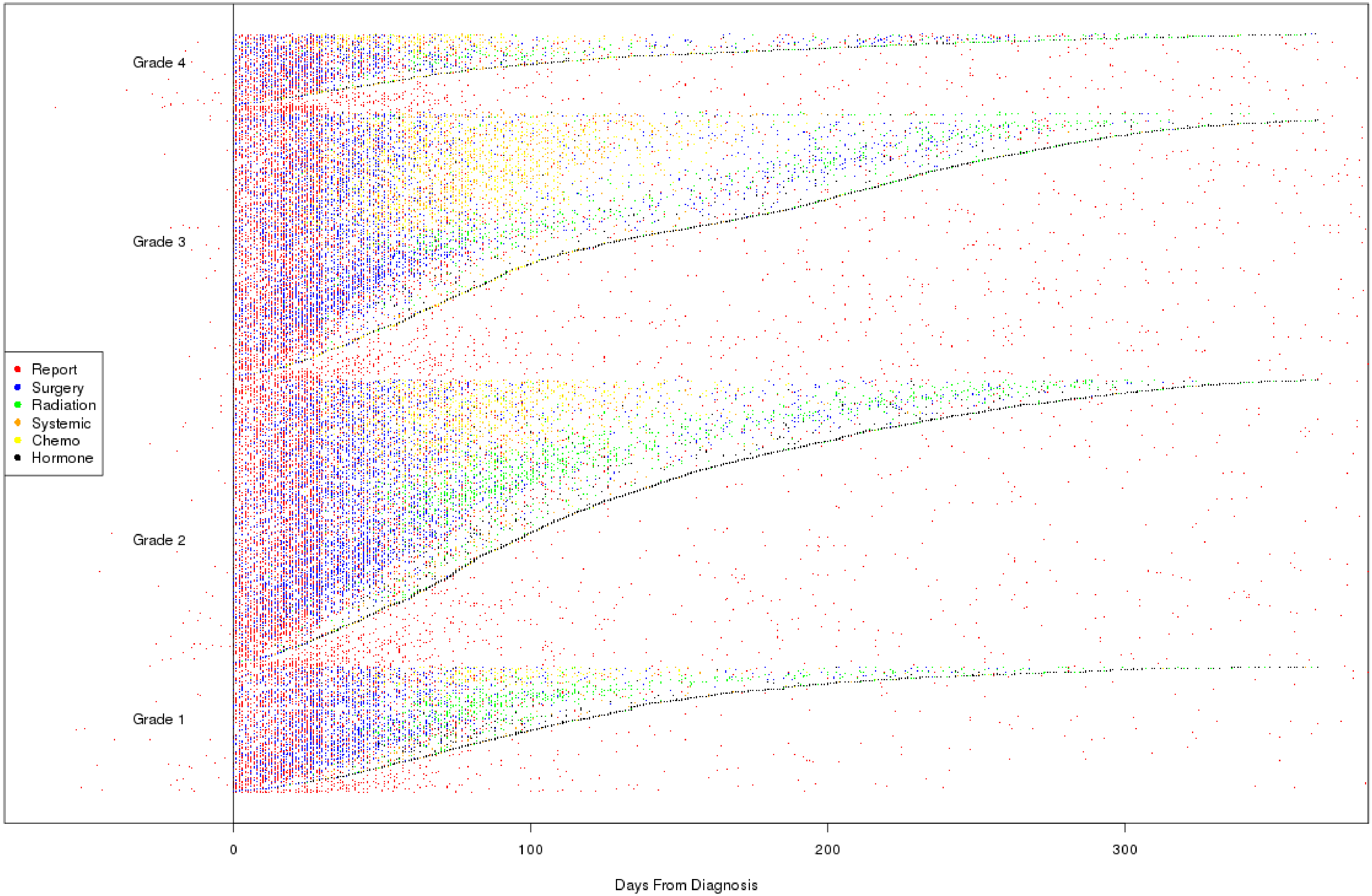
- **STEP 1: Selection of Appropriate Data Sources**
  - Ensure feasibility (Legal, IRB and logistic issues) and relevance to aims
  - Research and methodological questions for each data package
- **STEP 2: Data Linkages and Analytics**
  - Standard operating procedures and infrastructure for data linkages
  - Data analytics and visualization
    - » Parallel coordinates and other multivariate longitudinal visualizations for patient trajectories
    - » Prototyping a scalable, parallel, flexible framework with support for R and python



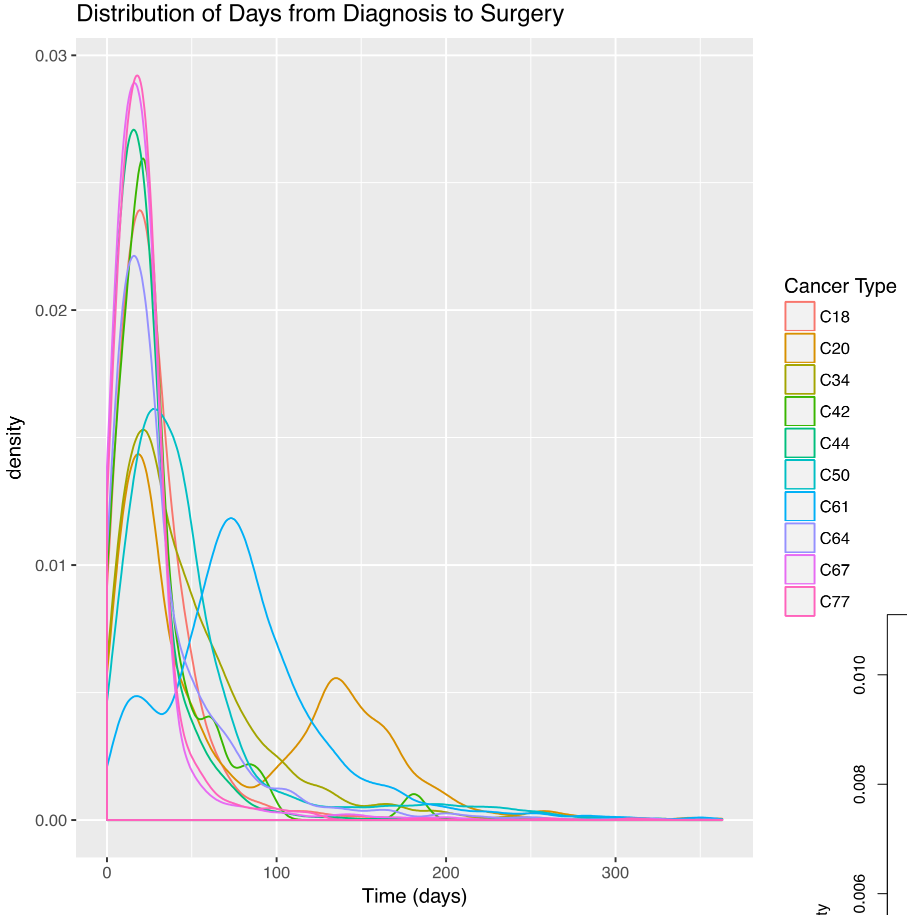


Building Patient Trajectories

- **What events happened to individual patients?**
- **What events happened across a population of patients?**
- **What do the (statistical) distributions look like across patients?**
- **What covariates (eg location, payor, sex, age, biomarkers, cancer characteristics) associated with the *most commonly used treatment regimes in a real world population?***
- **Set the stage for analysis of individual and population outcomes**

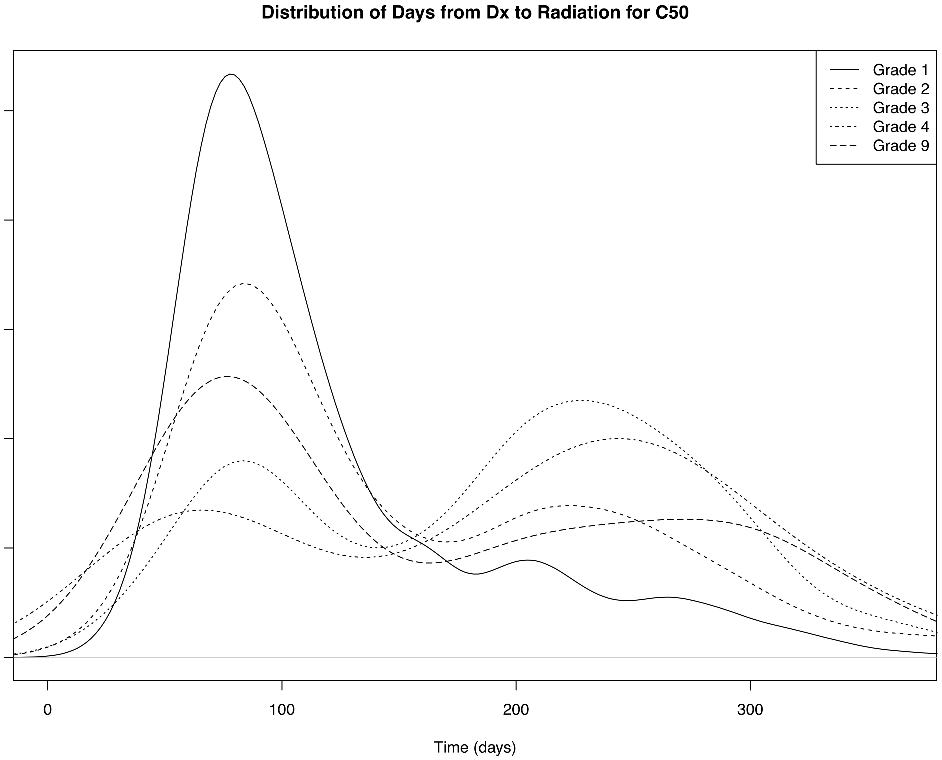


~15,000 primary tumor trajectories for breast cancer, demonstrates variation meriting further analysis



Time to surgery varies by cancer type

Time to radiation varies by grade within cancer type



Pilot 3

Scientific Outcomes since 10/2016

- **Peer-Reviewed Journal Publications:**
  - J.X. Qiu, H.-Y. Yoon, P.A. Fearn, G.D. Tourassi, “Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports,” IEEE Journal of Biomedical and Health Informatics [05/2017]
  - S. Gao, M.T. Young, J.X. Qiu, J.B. Christian, P.A. Fearn, G.D. Tourassi, A. Ramanathan, “Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports,” Journal of American Medical Informatics Association [*accepted 10/2017*].
- **Peer-Reviewed Conference Articles & Posters:**
  - H.-J. Yoon, A. Ramanathan, G.D. Tourassi. "Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports." In INNS Conference on Big Data, pp. 195-204. Springer International Publishing, 2016.
  - H.-J. Yoon, L.W. Roberts, G.D. Tourassi, Automated histologic grading from free-text pathology reports using graph-of-words features and machine learning. 2017 IEEE International Conference on Biomedical and Health Informatics, Orlando, Florida, February 16-19, 2017 [Available in IEEE Xplore 04/2017] .
  - J. Boten, D. Rivera, M. Myneni, G.D. Tourassi, T. Bhattacharya, A.P. de Oliveira Sales, T. Brettin, P. Fearn, L. Penberthy, “Leveraging Large-Scale Computing for Population Information Integration,” AMIA 2017 Annual Symposium, November 4-8, 2017, Washington, DC [*Accepted*].
  - G. Abastillas, S. Morris, J. Boten, T. Tumurchudur, K. Vora, P. Fearn, “Characterizing a Learning Curve for Annotating Data for Training and Validation of Natural Language Processing Systems,’ AMIA 2017 Annual Symposium, November 4-8, 2017, Washington, DC [*Accepted*].
- **Invited Presentations:**
  - L. Penberthy, G.D. Tourassi, ““Population Information Integration, Analysis and Modeling”, Computational Approaches for Cancer Workshop, Supercomputing 2016, Salt Lake City, UT, November 13, 2016.
  - A. Ramanathan, “Exascale deep text comprehension tools for cancer surveillance”, GPU Tech Conference (GTC), San Jose, May 2017.
  - G.D. Tourassi, “Deep Learning Enabled National Cancer Surveillance to Support Precision Oncology”, 21st Century Cures: Southeast Conference, Knoxville, TN, June 1, 2017.
  - T. Bhattacharya, “Surveillance in an Era of Emerging Technology and Precision Medicine,” NAACCR 2017 Annual Symposium, June 16-23, 2017, Albuquerque, NM.
  - J. Boten, “The Development of the Clinical Document Annotation and Processing Pipeline to Facilitate the Integration of Natural Language Processing to Enhance Cancer Surveillance,” NAACCR 2017 Annual Symposium, June 22, 2017, Albuquerque, NM.
- **Educational Outreach:**
  - G.D. Tourassi, “Advanced Deep Learning for NLP”, NCI NLP Workshop, Rockville, MD, December 8, 2016
  - A. Ramanathan, “Building deep text comprehension tools for cancer surveillance”, NCI-DOE Workshop on Cancer Deep Learning Environment (CANDLE), National Cancer Institute, Bethesda, MD, April 2017.

Future Directions

Pilot 3

- **DUAs:** gain access to additional registries data and regular updates as new data arrives
- **Aim 1:**
  - Annotate pathology reports for breast cancer biomarkers & recurrence
  - Scale up annotation pipeline to up to 10,000 documents per month
  - Identify and prioritize other key biomarkers for inclusion in the annotation pipeline
  - Test and scale supervised and semi-supervised DL algorithms for automated extraction of 5 key variables (histology, laterality, behavior, grade and organ site) with uncertainty information for use by registries
- **Aim 2:**
  - 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;
  - Incorporate detailed treatment data on a subset of the population for use in algorithms and modeling (e.g. recurrence and response to therapy)
  - Develop scalable visual and graph 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 and modeling response to initial and subsequent therapies for selected cancer sites

THANK YOU!!!