DeepPhe*CR: Natural Language Processing Platform for Cancer Surveillance

Guergana Savova

Boston Childrens' Hospital/Harvard Medical School

guergana.savova@
childrens.Harvard.edu

Harry Hochheiser

University of Pittsburgh

harryh@pitt.edu

Jeremy Warner

Vanderbilt University

jeremy.warner@vumc.org

Eric B. Durbin

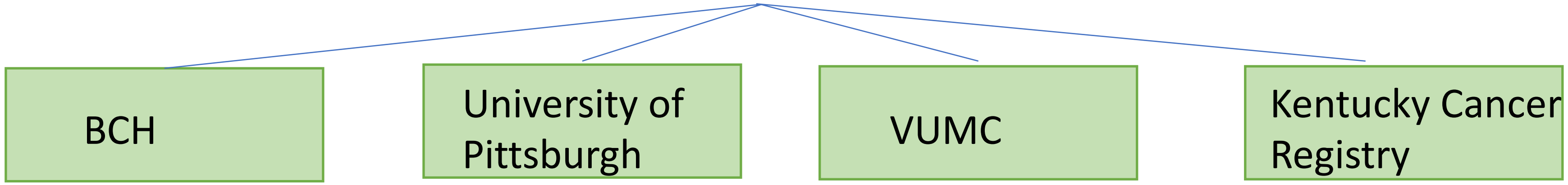
Kentucky Cancer Registry

ericd@kcr.uky.edu

UG3 CA243120

February 8, 2021

Team



BCH

University of Pittsburgh

VUMC

Kentucky Cancer Registry

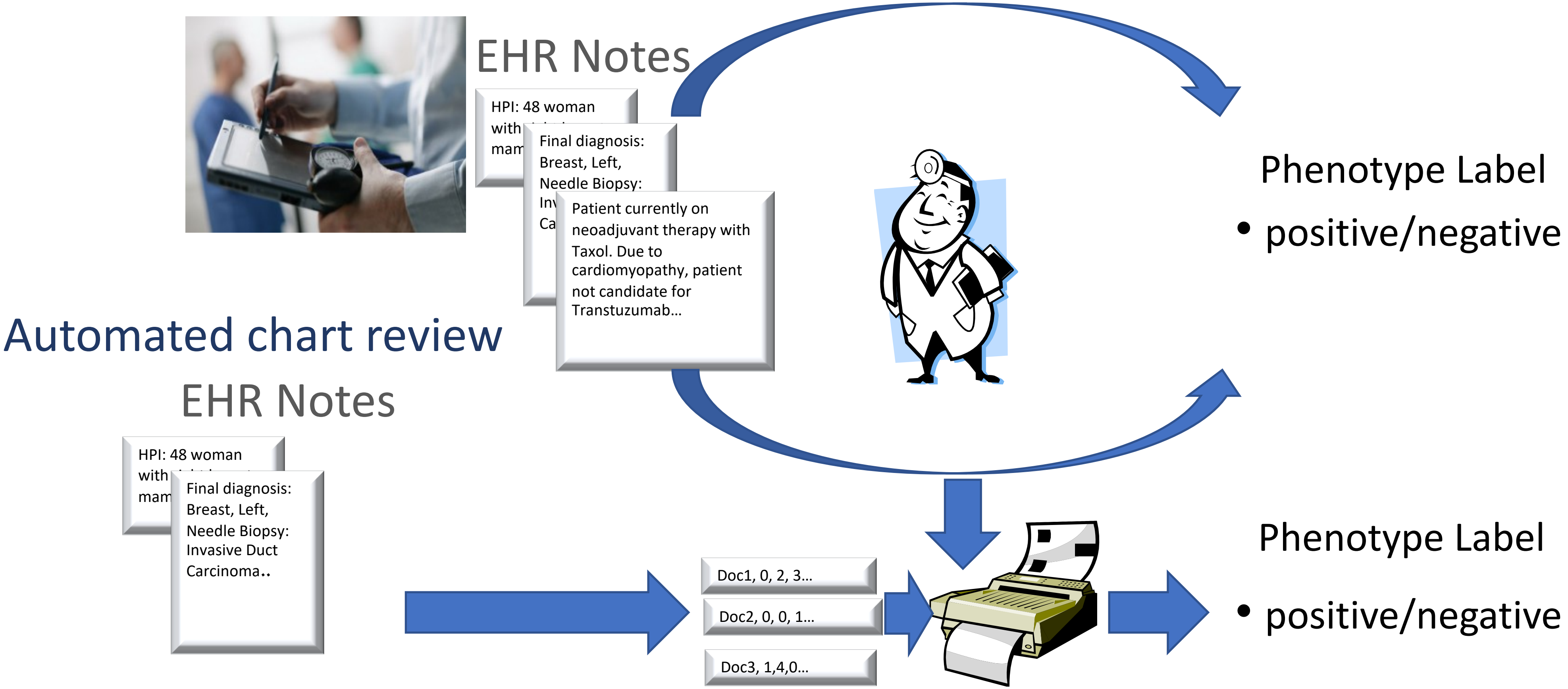
Guergana Savova, MPI
Sean Finan
Timothy Miller
Chen Lin
David Harris
Danielle Bitterman

Harry Hochheiser, MPI
Zhou Yuan

Jeremy Warner, MPI
Li Wen

Eric B. Durbin, MPI
Isaac Hands
JC Jeong
Rama Kavuluru
David Rust
Lisa Witt
Joseph Hurt-Mueller

Phenotype Extraction from Clinical Notes



EHR Notes

HPI: 48 woman
with
mam

Final diagnosis:
Breast, Left,
Needle Biopsy:
Inv
Ca

Patient currently on neoadjuvant therapy with Taxol. Due to cardiomyopathy, patient not candidate for Transtuzumab…

Phenotype Label

- positive/negative

Automated chart review

EHR Notes

HPI: 48 woman
with
mam

Final diagnosis:
Breast, Left,
Needle Biopsy:
Invasive Duct
Carcinoma..

Doc1, 0, 2, 3…

Doc2, 0, 0, 1…

Doc3, 1,4,0…

Phenotype Label

- positive/negative

DeepPhe*CR

- A generalizable information extraction framework for cancer surveillance
- Component of normalizing/standardizing information
  - Cancer and tumor attributes
  - Treatment and genomic information through widely used ontologies (RxNorm, HemOnc, NAACCR Descriptors)
  - Standards/best-practices for inputs and outputs
    - NAACCR-XML
    - Pathology notes
    - XML sequencing reports
    - ..
- Flexible architecture allows integration with existing tools (SEER*DMS), etc.

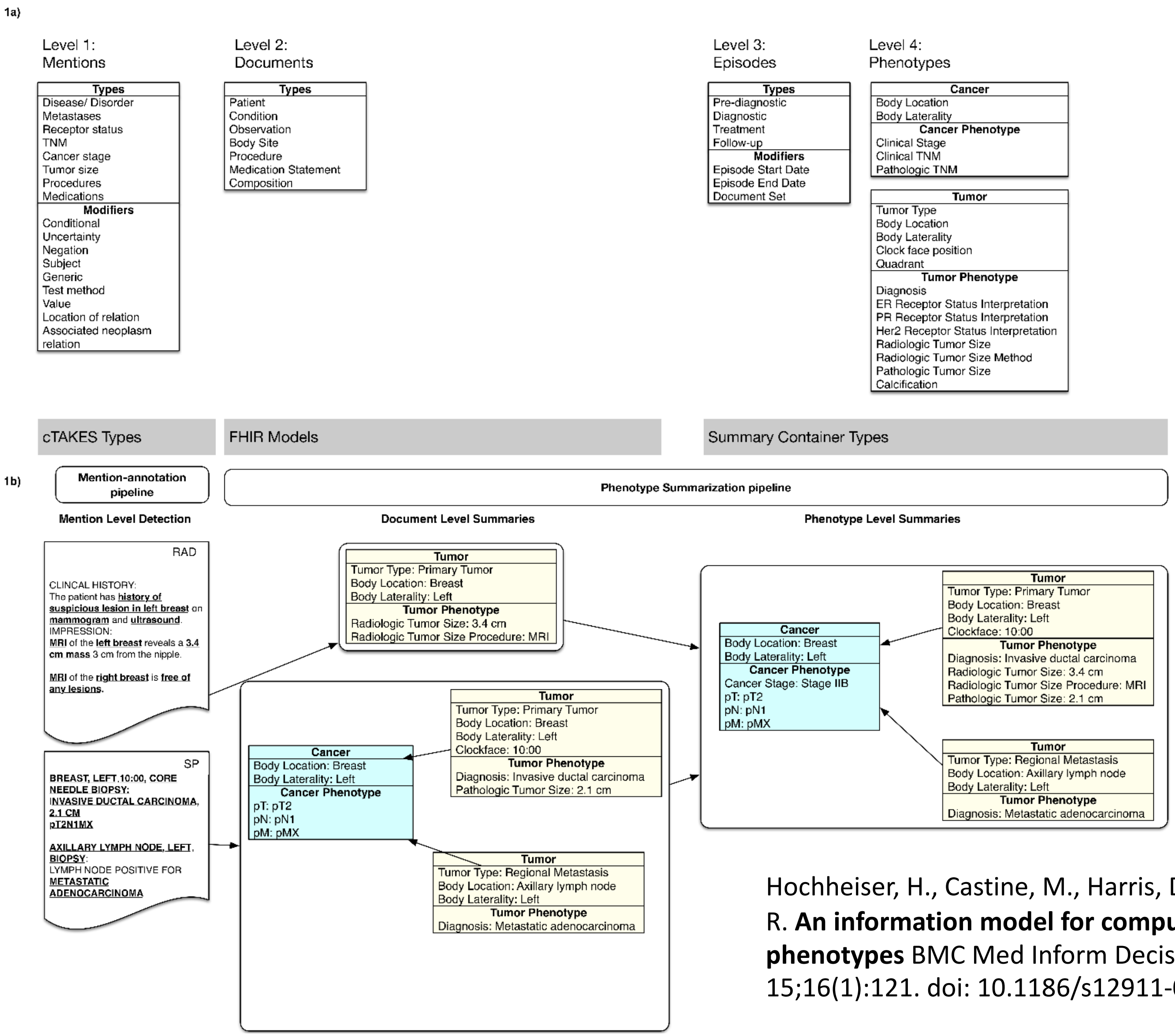
Specific Aims

- **Aim 1**: Develop methods for the automatic extraction of the cancer and tumor characteristics from a variety of data sources.
- **Aim 2**: Extract treatment information via various channels. The extracted treatment information will be mapped to ontologies such as RxNorm and HemOnc.
- **Aim 3**: Develop methods for the extraction of clinical genomics information from (1) XML data feeds from sequencing providers such as Foundation Medicine, (2) pathology notes.
- **Aim 4**: Develop software architectures and tools in support of integrating best-performing DeepPhe*CR methods from SA1-3 into registry abstraction tools.

Core Attributes

primary site, histologic type, behavior code, laterality, grade, TNM

DeepPhe Information Model



1a)

Level 1:
Mentions

Types
Disease/ Disorder
Metastases
Receptor status
TNM
Cancer stage
Tumor size
Procedures
Medications
Modifiers
Conditional
Uncertainty
Negation
Subject
Generic
Test method
Value
Location of relation
Associated neoplasm relation

Level 2:
Documents

Types
Patient
Condition
Observation
Body Site
Procedure
Medication Statement
Composition

Level 3:
Episodes

Types
Pre-diagnostic
Diagnostic
Treatment
Follow-up
Modifiers
Episode Start Date
Episode End Date
Document Set

Level 4:
Phenotypes

Cancer
Body Location
Body Laterality
Cancer Phenotype
Clinical Stage
Clinical TNM
Pathologic TNM

Tumor
Tumor Type
Body Location
Body Laterality
Clock face position
Quadrant
Tumor Phenotype
Diagnosis
ER Receptor Status Interpretation
PR Receptor Status Interpretation
Her2 Receptor Status Interpretation
Radiologic Tumor Size
Radiologic Tumor Size Method
Pathologic Tumor Size
Calcification

cTAKES Types

FHIR Models

Summary Container Types

1b)

Mention-annotation pipeline

Phenotype Summarization pipeline

Mention Level Detection

Document Level Summaries

Phenotype Level Summaries

RAD

CLINCAL HISTORY:
The patient has **history of suspicious lesion in left breast** on **mammogram** and **ultrasound**.
IMPRESSION:
MRI of the **left breast** reveals a **3.4 cm mass** 3 cm from the nipple.

MRI of the **right breast** is **free of any lesions.**

SP

BREAST, LEFT,**10:00**, **CORE NEEDLE BIOPSY:**
INVASIVE DUCTAL CARCINOMA,
2.1 CM
pT2N1MX

AXILLARY LYMPH NODE, LEFT,
BIOPSY:
LYMPH NODE POSITIVE FOR
METASTATIC ADENOCARCINOMA

Tumor
Tumor Type: Primary Tumor
Body Location: Breast
Body Laterality: Left
Tumor Phenotype
Radiologic Tumor Size: 3.4 cm
Radiologic Tumor Size Procedure: MRI

Cancer
Body Location: Breast
Body Laterality: Left
Cancer Phenotype
pT: pT2
pN: pN1
pM: pMX

Tumor
Tumor Type: Primary Tumor
Body Location: Breast
Body Laterality: Left
Clockface: 10:00
Tumor Phenotype
Diagnosis: Invasive ductal carcinoma
Pathologic Tumor Size: 2.1 cm

Tumor
Tumor Type: Regional Metastasis
Body Location: Axillary lymph node
Body Laterality: Left
Tumor Phenotype
Diagnosis: Metastatic adenocarcinoma

Cancer
Body Location: Breast
Body Laterality: Left
Cancer Phenotype
Cancer Stage: Stage IIB
pT: pT2
pN: pN1
pM: pMX

Tumor
Tumor Type: Primary Tumor
Body Location: Breast
Body Laterality: Left
Clockface: 10:00
Tumor Phenotype
Diagnosis: Invasive ductal carcinoma
Radiologic Tumor Size: 3.4 cm
Radiologic Tumor Size Procedure: MRI
Pathologic Tumor Size: 2.1 cm

Tumor
Tumor Type: Regional Metastasis
Body Location: Axillary lymph node
Body Laterality: Left
Tumor Phenotype
Diagnosis: Metastatic adenocarcinoma

Hochheiser, H., Castine, M., Harris, D., Savova G, Jacobson R. **An information model for computable cancer phenotypes** BMC Med Inform Decis Mak. 2016 Sep 15;16(1):121. doi: 10.1186/s12911-016-0358-4

Primary site
Histologic type
Behavior code
Laterality
Grade
cTNM
pTNM

Data: DeepPhe

1. UPMC patients with breast cancer (N=94; 2,836 documents), melanoma (N=24; 674 documents), and ovarian cancer (N=46; 2,797 documents);
2. Vanderbilt University Medical Center (VUMC) breast cancer (N=9,515), ovarian cancer (N=427) and melanoma patients (N=2,460);
3. Dana Farber Cancer Institute (DFCI) melanoma patients (N=2,400);
4. VUMC and Brigham and Women’s Hospital (BWH) prostate cancer patients through a supplement with ITCR grantee Dr. Fedorov (BWH) (N=1,000)
5. SEER patients across the National SEER program, the LTR and the KCR with breast cancer (N=676; 1,310 documents) and melanoma (N=112; 586 documents)

- Dataset 1-3: pathology reports, radiology reports, clinical progress notes, hospital discharge summaries, and ED encounters
- Dataset 4-5: pathology and radiology notes
- Datasets 1 and 5: manually annotated

Additional Data for DeepPhe-CR: lung, breast, prostate

- Lung, breast and prostate cancers
- Data streams
  - E-Path notes
  - E-Rad notes
  - Pilot EMR data from University of Kentucky (this is an exception for cancer registries as SEER registries have only limited direct access to EMRs)
- Gold annotations for 600 patients
- Pre-existing gold annotations for 850 patients

Core Attributes Gold Annotations: Process

- 2 domain experts
- Pilot annotations to stabilize the annotation guidelines (30 patients)
- Disagreements were tracked and discussed.
- Inter-annotator agreement measured as Kappa:
  - Results range from 0.77-1

Methods Overview

- A variety of artificial intelligence methods – pattern matching, rules, machine learning (e.g. SVMs, neural approaches), knowledge engineering, ontologies

Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, Harris D, Hochheiser H, Lin C, Chavan G, Jacobson RS. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer research. Nov 1 2017;77(21):e115-e118. PMID 29092954 PMC5690492

Miller, Timothy; Dligach, Dmitriy; Bethard, Steven; Lin, Chen; Savova, Guergana. 2017. Towards Portable Entity-Centric Clinical Coreference Resolution. Journal of Biomedical Informatics. Vol. 69, May 2017, pp. 251-258. https://doi.org/10.1016/j.jbi.2017.04.015; http://www.sciencedirect.com/science/article/pii/S1532046417300850

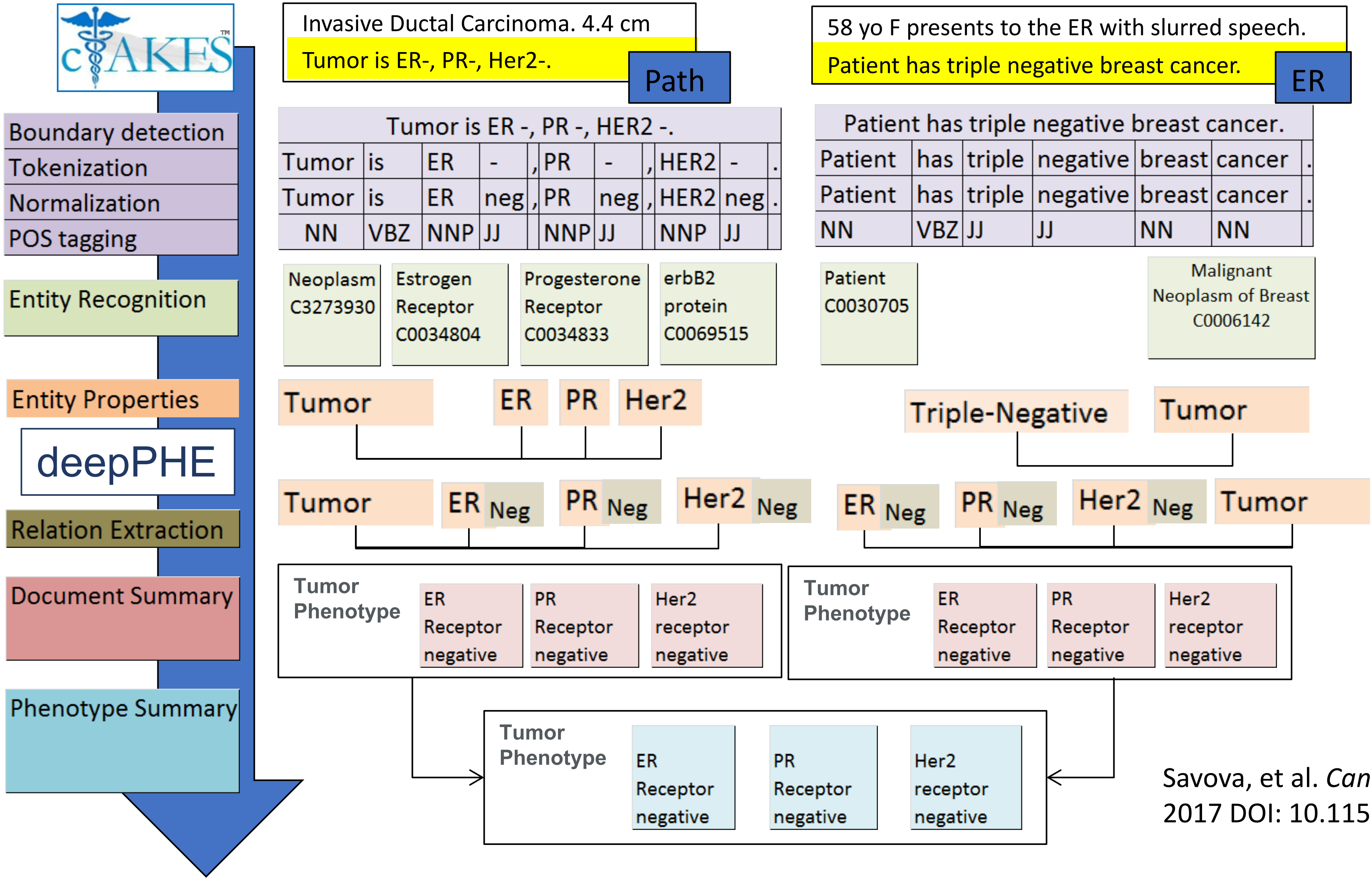
Lin C, Dligach D, Miller TA, Bethard S, Savova GK. Multilayered temporal modeling for the clinical domain. Journal of the American Medical Informatics Association : JAMIA. Mar 2016;23(2):387-395. PMID 26521301 PMC5009920

Warner JL, Cowan AJ, Hall AC, Yang PC. HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals. J Oncol Pract. 2015 May;11(3):e336-50. doi: 10.1200/JOP.2014.001511. Epub 2015 Mar 3.

Methods: Pipeline Approach

- Modules responsible for different tasks, e.g.
  - Sentence boundary
  - Token boundary
  - Entity mentions
  - Attributes of the entity mentions
  - Relations between the entities
  - Summarization
- Data usage from different points of view
- Modules implement different methods
- Usage of pre-existing modules, e.g. sentence boundary detection
- Reasonable computational demands

The DeepPhe System



cTAKES

Boundary detection

Tokenization

Normalization

POS tagging

Entity Recognition

Entity Properties

deepPHE

Relation Extraction

Document Summary

Phenotype Summary

Invasive Ductal Carcinoma. 4.4 cm
Tumor is ER-, PR-, Her2-.

Path

58 yo F presents to the ER with slurred speech.
Patient has triple negative breast cancer.

ER

Tumor is ER -, PR -, HER2 -.

Tumor	is	ER	-	,	PR	-	,	HER2	-	.
Tumor	is	ER	neg	,	PR	neg	,	HER2	neg	.
NN	VBZ	NNP	JJ		NNP	JJ		NNP	JJ	

Patient has triple negative breast cancer.

Patient	has	triple	negative	breast	cancer	.
Patient	has	triple	negative	breast	cancer	.
NN	VBZ	JJ	JJ	NN	NN	

Neoplasm C3273930

Estrogen Receptor C0034804

Progesterone Receptor C0034833

erbB2 protein C0069515

Patient C0030705

Malignant Neoplasm of Breast C0006142

Tumor Phenotype: ER Receptor negative; PR Receptor negative; Her2 receptor negative

Tumor Phenotype: ER Receptor negative; PR Receptor negative; Her2 receptor negative

Tumor Phenotype: ER Receptor negative; PR Receptor negative; Her2 receptor negative

Savova, et al. *Cancer Research* 2017
2017 DOI: 10.1158/0008-5472.CAN-17-0615

Treatment Information Extraction

Treatment Information

- Data streams
  - NAACCR abstracts (narrative text components)
  - Pilot EMR data from University of Kentucky (this is an exception for cancer registries as SEER registries have only limited direct access to EMRs)
- Modules for medication extraction, radiotherapy treatments, and temporality

Named Entity Extraction

Identifying and labeling pertinent treatment entities
Example: Dose and Treatment Site

She presented after a screening mammogram showed a nodule in the left breast upper outer quadrant. After lumpectomy, she was treated with radiation to a dose of 50 Gy in 25 fractions to the left breast, followed by a boost of 10 Gy in 5 fractions to the tumor bed.

ASTRO 2020
Clinical NLP Workshop at EMNLP 2020

Courtesy of Dr. Danielle Bitterman

Results: NER

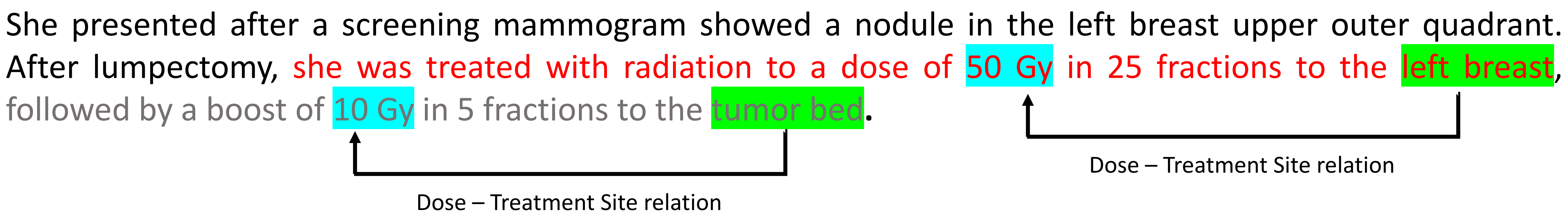
Entity	IAA F1	Instances			Precision - P (PPV)	Recall – R (sensitivity)	F1 (harmonic mean of P and R)
		Train Set	Development Set	Test Set			
Radiotherapy Dose	0.99	397	129	178	0.96	0.95	0.95
Fraction Number	0.83	163	55	90	0.86	0.74	0.8
Fraction Frequency	1	52	15	15	0.93	0.93	0.93
Boost	1	23	16	10	0.7	0.7	0.7
Treatment Site	0.8	153	59	120	0.97	0.94	0.95
Treatment Dates	1.00	55	36	37	0.73	0.53	0.61

Courtesy of Dr. Danielle Bitterman

Relation Extraction

Labeling pertinent treatment entities that refer to the same radiotherapy phase

Example: Dose - Treatment Site



She presented after a screening mammogram showed a nodule in the left breast upper outer quadrant. After lumpectomy, she was treated with radiation to a dose of 50 Gy in 25 fractions to the left breast, followed by a boost of 10 Gy in 5 fractions to the tumor bed.

Dose – Treatment Site relation
Dose – Treatment Site relation

Red = First course

Blue = Boost course

ASTRO 2020
Clinical NLP Workshop at EMNLP 2020

Courtesy of Dr. Danielle Bitterman

Results: Relation Extraction

Flair Models	IAA	Precision		Recall		F1	
		92 Word Windows (n=1444)	180 Word Windows (n=2520)	92 Word Windows (n=1444)	180 Word Windows (n=2520)	92 Word Windows (n=1444)	180 Word Windows (n=2520)
Dose-Dose	0.94	0.77	0.79	0.75	0.83	0.76	0.81
Dose-Treatment Site	0.90	0.84	0.61	0.86	0.92	0.85	0.73
Dose-Fraction Frequency	1.00	0.79	0.84	0.95	1.00	0.86	0.91
Dose-Fraction Number	0.98	0.95	0.90	0.93	0.92	0.94	0.91
Dose-Boost	0.67	1.00	0.56	0.69	0.69	0.82	0.62
None	0.74	0.95	0.98	0.95	0.94	0.95	0.96
Average	0.87	0.88	0.78	0.86	0.88	0.86	0.82

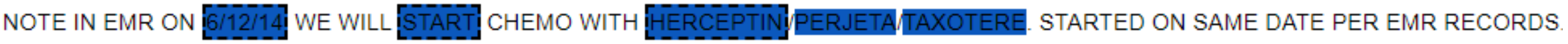
BERT Models	IAA	Precision		Recall		F1	
		92 Word Windows (n=1444)	180 Word Windows (n=2520)	92 Word Windows (n=1444)	180 Word Windows (n=2520)	92 Word Windows (n=1444)	180 Word Windows (n=2520)
Dose-Dose	0.94	0.88	0.68	0.82	0.81	0.85	0.74
Dose-Treatment Site	0.90	0.79	0.76	0.88	0.89	0.83	0.82
Dose-Fraction Frequency	1.00	0.78	0.88	0.90	1.00	0.84	0.93
Dose-Fraction Number	0.98	0.94	0.87	0.96	0.95	0.95	0.91
Dose-Boost	0.67	0.85	0.60	0.85	0.92	0.85	0.73
None	0.74	0.96	0.98	0.94	0.95	0.95	0.96
Average	0.87	0.87	0.80	0.89	0.92	0.88	0.85

ASTRO 2020
Clinical NLP Workshop at EMNLP 2020

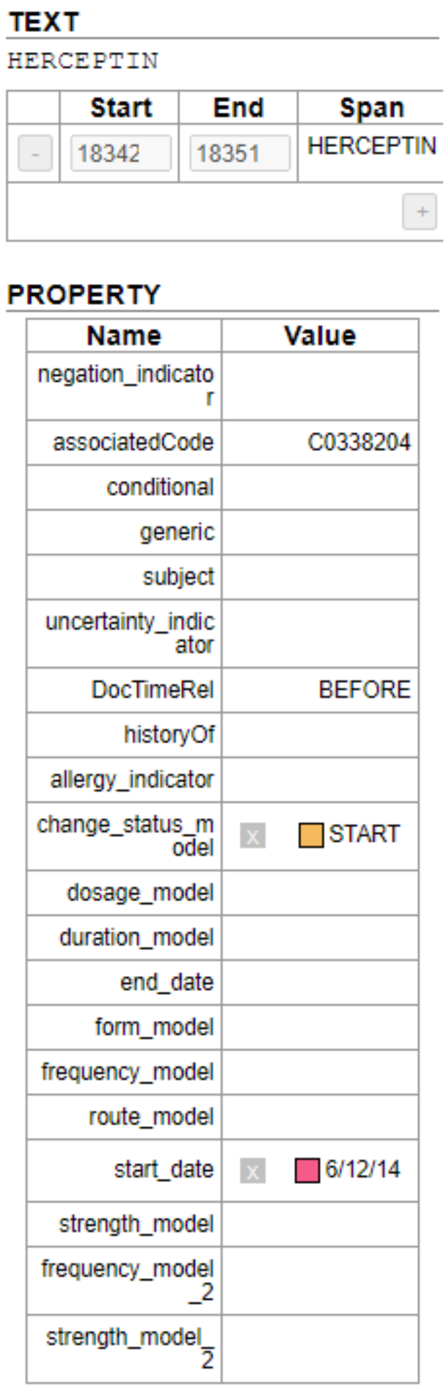
Courtesy of Dr. Danielle Bitterman

Medication Signature Extraction

- Gold annotations in progress
- Medication attributes



NOTE IN EMR ON 6/12/14 WE WILL START CHEMO WITH HERCEPTIN/PERJETA/TAXOTERE. STARTED ON SAME DATE PER EMR RECORDS.



TEXT

HERCEPTIN

	Start	End	Span
-	18342	18351	HERCEPTIN
			+

PROPERTY

Name	Value
negation_indicator	
associatedCode	C0338204
conditional	
generic	
subject	
uncertainty_indicator	
DocTimeRel	BEFORE
historyOf	
allergy_indicator	
change_status_model	☒ 🟧 START
dosage_model	
duration_model	
end_date	
form_model	
frequency_model	
route_model	
start_date	☒ 🟥 6/12/14
strength_model	
frequency_model_2	
strength_model_2	

Methods: Medication Extraction

- Medication extraction and code mapping
  - Ontology-driven method
- Medication signature extraction
  - Neural model (in progress)

Temporality

- Relations of particular interest:
  - DocTimeRel – relation of the event to the document creation time
  - CONTAINS – date or date range containing the event
- Trained on colorectal cancer notes, method is BERT-style multi-task learning

Clinical Genomics Extraction

Clinical Genomics Extraction

- Data feeds:
  - Foundation Medicine XML
  - Pathology notes
- Focus on biomarker priorities for cancer registry reporting
- Ontology
  - Mostly NCIt with some protein classes from HPO
  - Some 30 custom relations between appropriate cancer branches and biomarkers
- XML parser for Foundation Medicine documents
- NLP module for biomarker extraction

HPO: Human Phenotype Ontology

Integration of Clinical Genomics

- Foundation Medicine, Inc. (FMI) XML data feeds in Kentucky
    - University of Kentucky
    - University of Louisville
    - Norton Healthcare (pending)
    - Statewide (pending)
  - FMI XML includes clinically reported mutations, mutations of unknown significance and other biomarkers such as Tumor Mutation Burden
- KCR XML parser translates XML into discrete data elements for database storage and retrieval

Software architectures and tools for integrating best-performing DeepPhe*CR methods into registry abstraction tools.

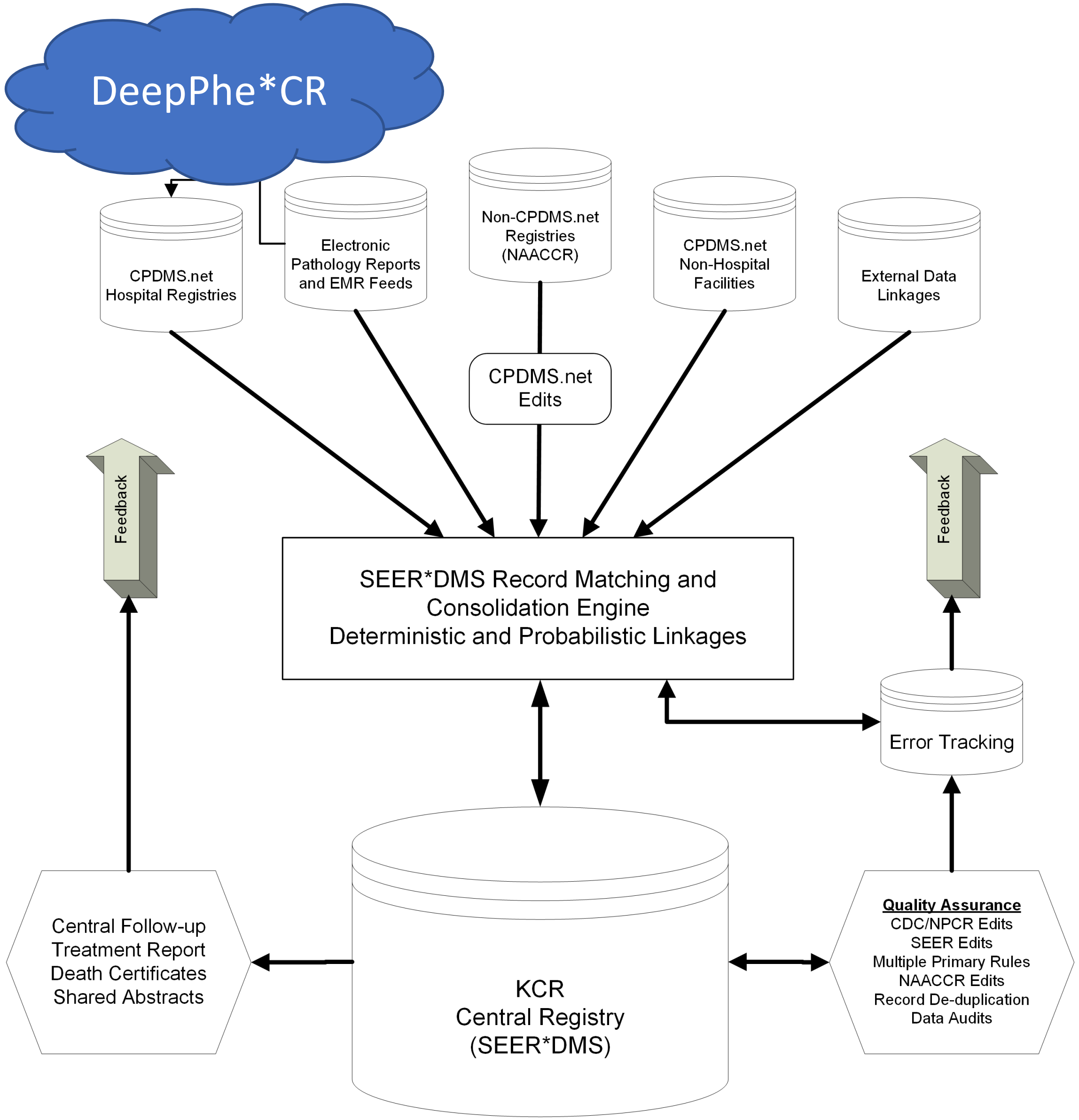
Software Development

- Understand registrar workflows and user needs
- Develop and document Application Programming Interfaces (APIs)
- Provide containers, documentation, and support to encourage deployment
- Refine, revise, and harden APIs, containers, and documentation

KCR's Cancer Patient Data Management System (CPDMS)

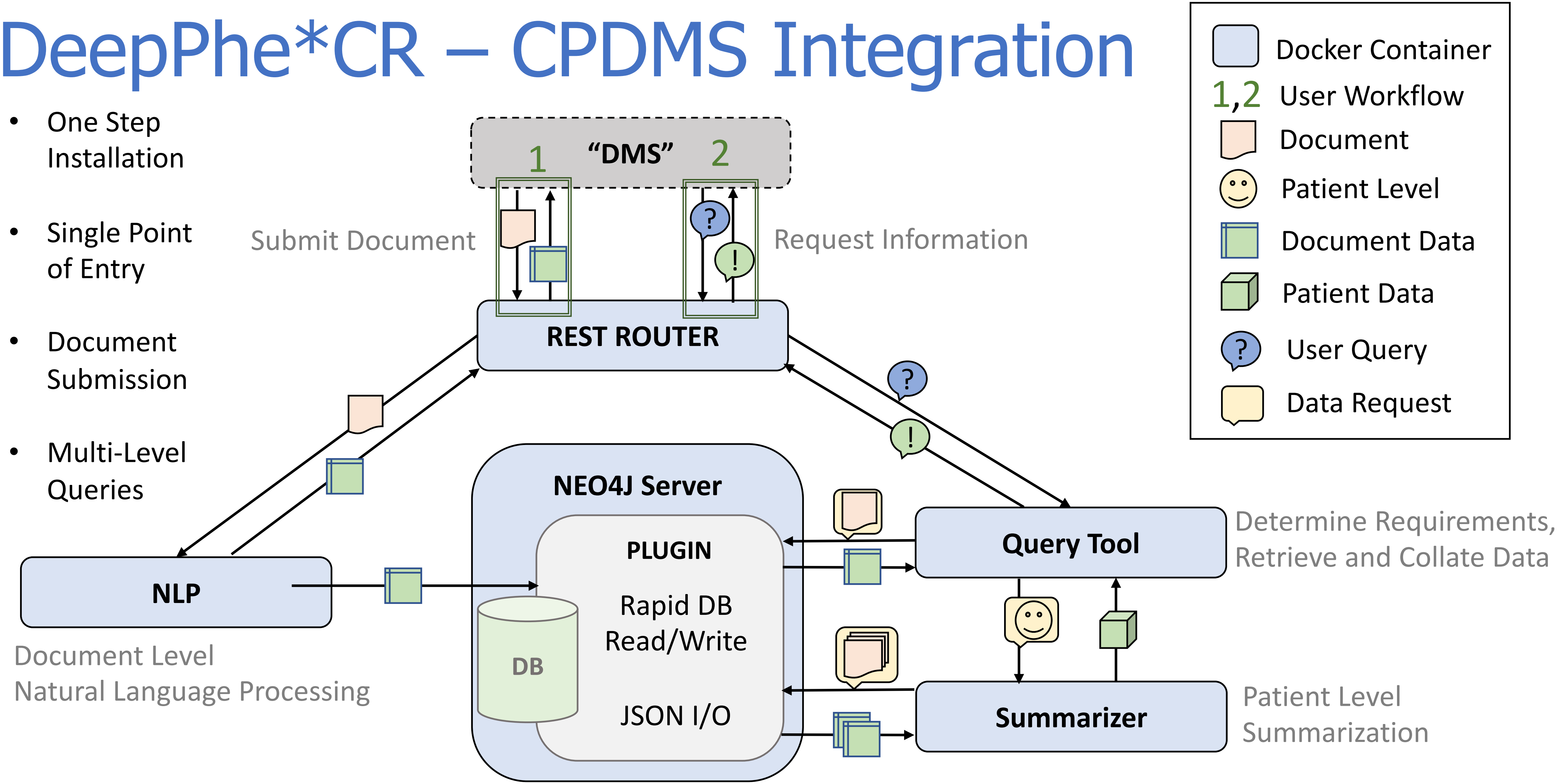
- CPDMS is the cancer registry abstracting and data management system used by all non-federal hospitals and facilities in Kentucky
- CPDMS is primarily used to create and maintain high quality longitudinal cancer abstracts
  - Similar to SEER*Abs, but many more features
- CPDMS has been fully integrated into SEER*DMS for KCR

CPDMS Integration with SEER*DMS



DeepPhe*CR – CPDMS Integration

- One Step Installation
- Single Point of Entry
- Document Submission
- Multi-Level Queries



Next steps: SEER*DMS

- Discussions with NCI and IMS underway
- Containerized architecture provides pathway for adoption
- Potential for co-existence with similar systems – DOE, etc.
- Collaboration with
  - Louisiana Tumor Registry
  - Massachusetts Cancer Registry