Towards a Scalable Clinical Data Annotation and Processing Pipeline to Support Cancer Surveillance

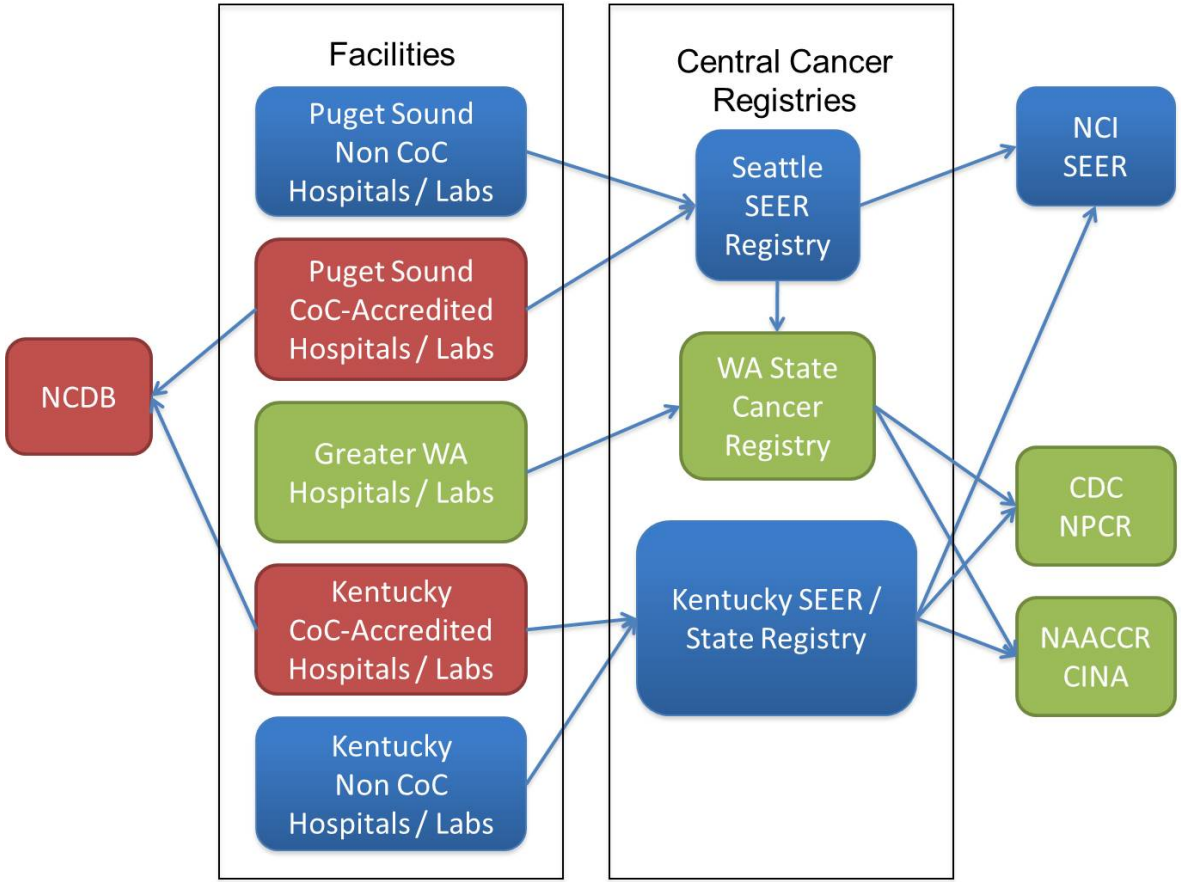
Paul Fearn, PhD, MBA
Surveillance Research Program
Division of Cancer Control and Population Sciences
paul.fearn@nih.gov

NIH NATIONAL CANCER INSTITUTE

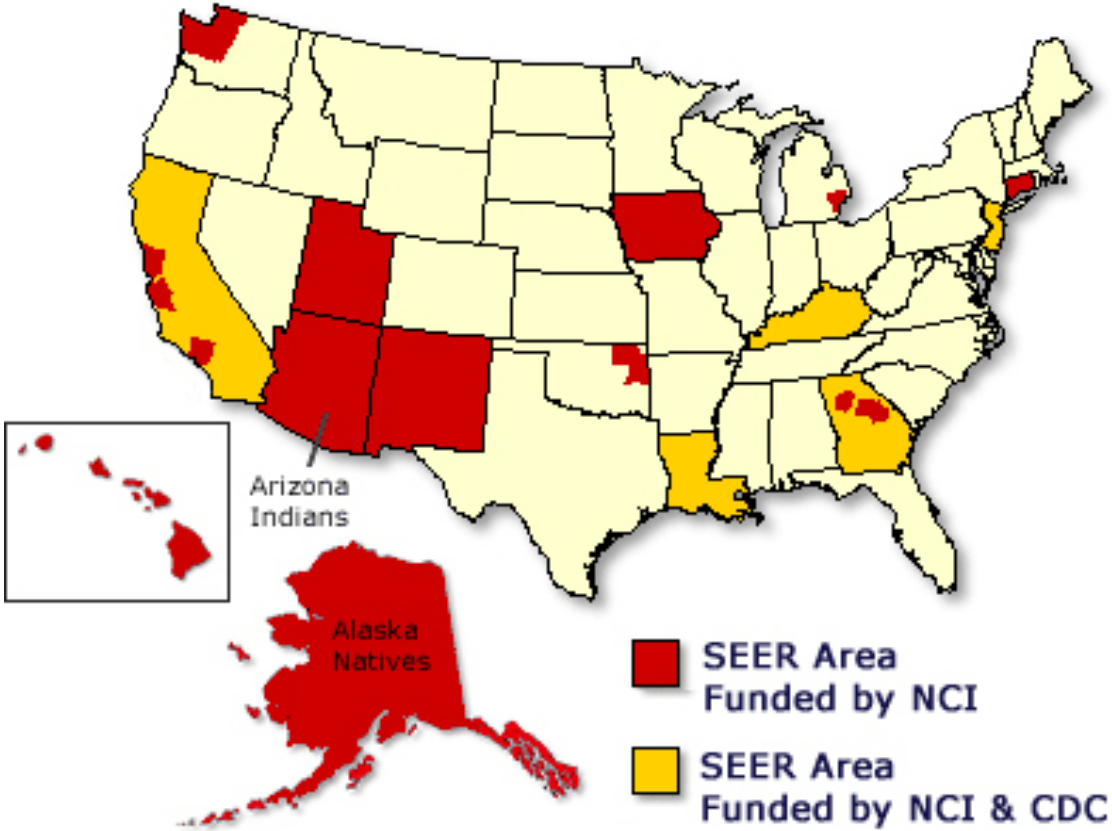
LabKey User Conference
Oct 5th, 2017

Context of Cancer Surveillance

Cancer Surveillance 101



The SEER Cancer Registries



https://seer.cancer.gov/registries/

NCI SEER and Surveillance Research Program (SRP)

- SRP organizationally
- SRP goals for SEER registries
  - **Natural language processing (NLP)** and other automation to improve efficiency and reliability and accuracy of data collection
  - **Linkages** of new detailed diagnostic and treatment data sources to improve
  - Data **quality** initiatives
  - Grow and sustain national population-based research resource

https://surveillance.cancer.gov/

Surveillance Informatics Branch (SIB) Initiatives

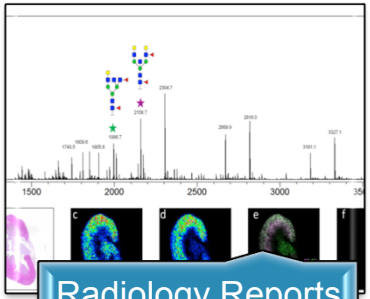
- Data Acquisition and Linkages
- SEER*DMS enhancements
- NLP: DOE collaboration, inter-agency collaboration with CDC, FDA, de-identification, data quality assessment and improvement

https://surveillance.cancer.gov/branches/sib/

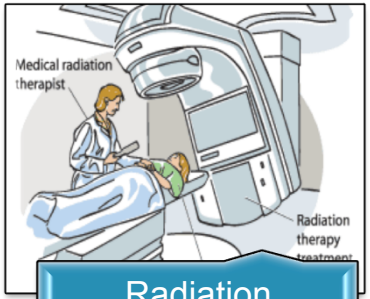
Data Acquisition and Linkage


Pathology

Pharmacy

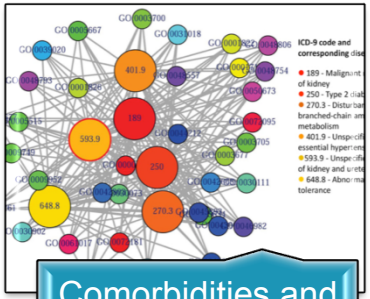


Radiology Reports and Imaging

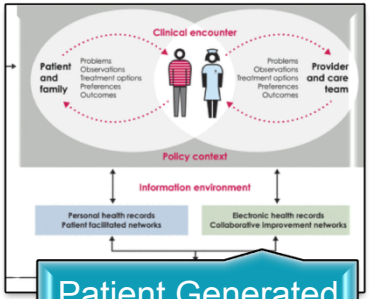


Radiation Oncology


Genomic/Lab

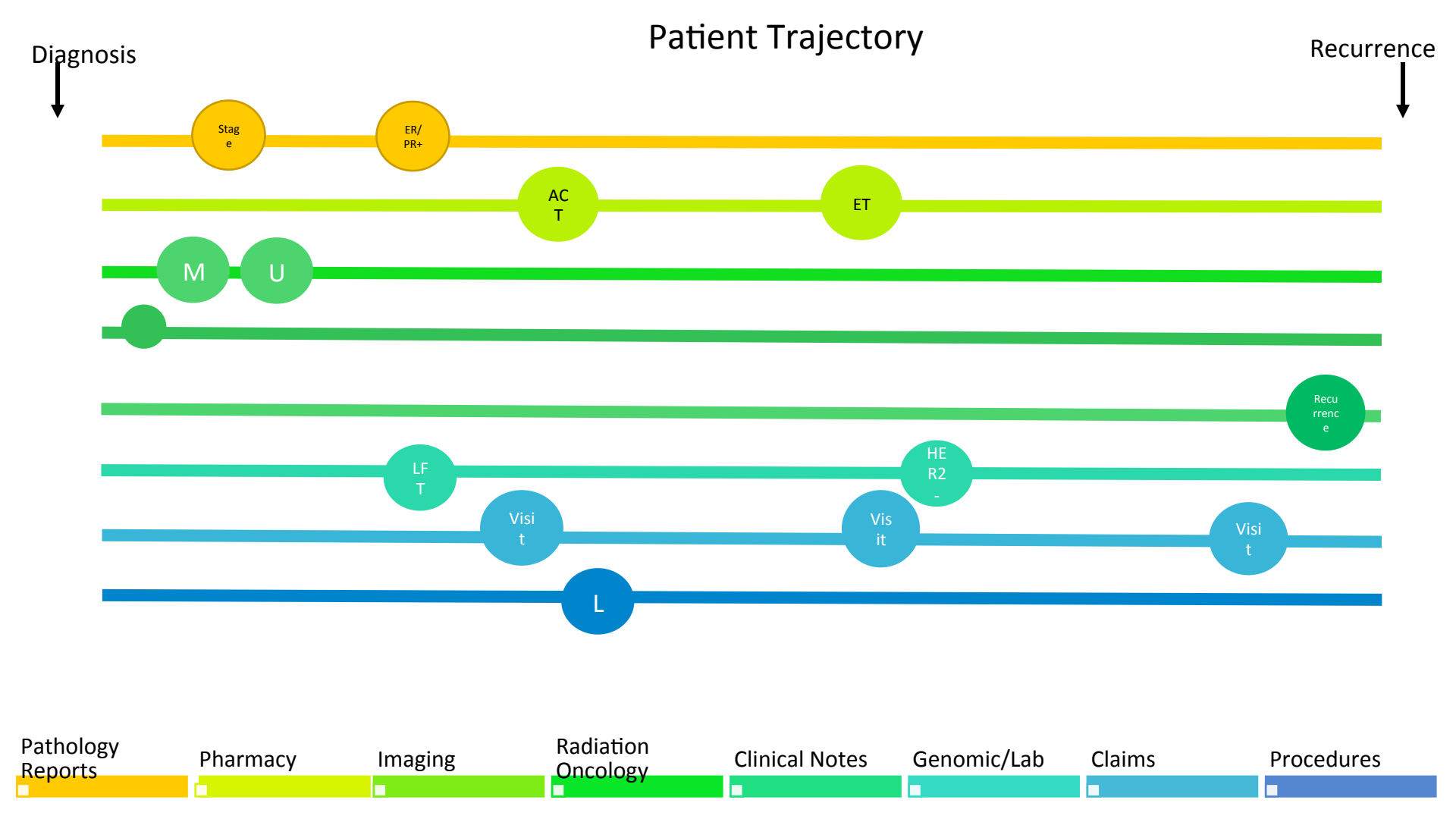

Comorbidities and Diagnoses


Procedures



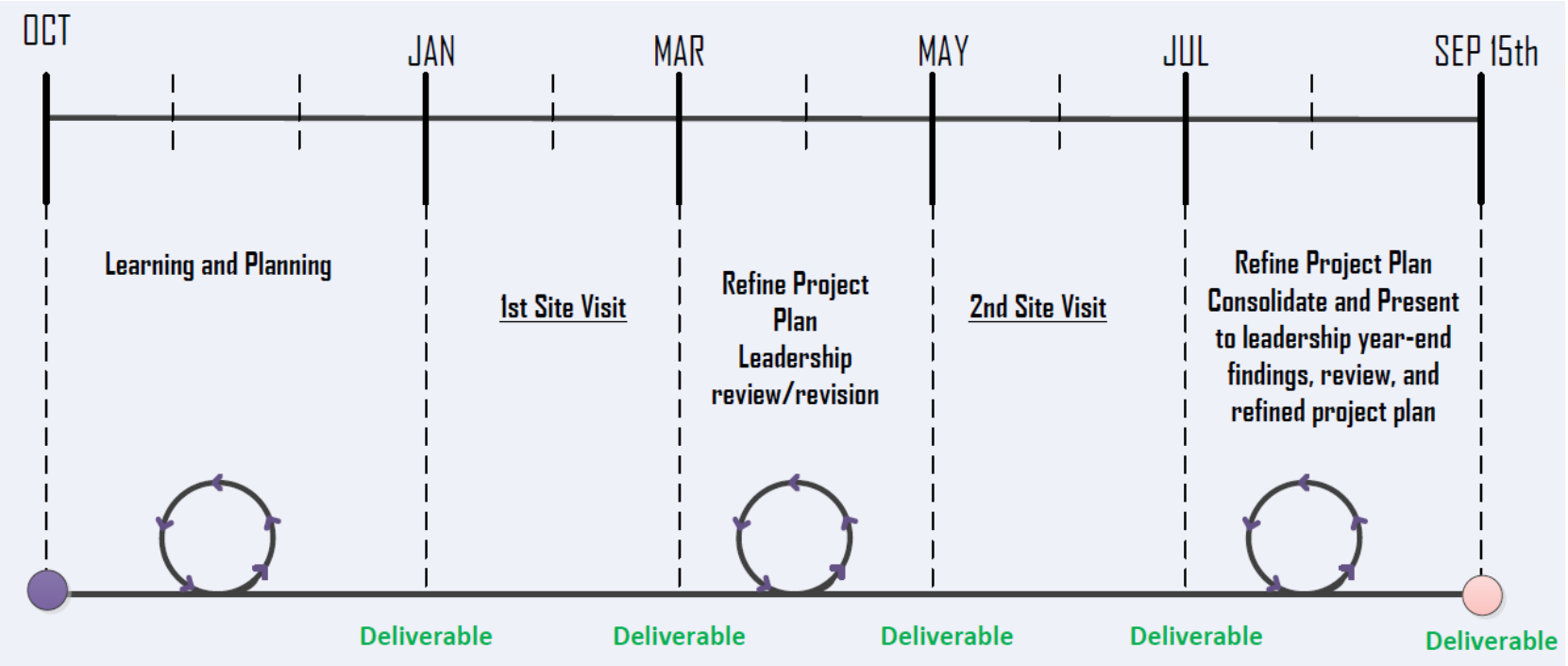
Patient Generated Information

From Donna Rivera


Patient Trajectory
Diagnosis
Recurrence
Stag e
ER/ PR+
AC T
ET
M
U
Recu rrenc e
LF T
HE R2 -
Visi t
Vis it
Visi t
L
Pathology Reports
Pharmacy
Imaging
Radiation Oncology
Clinical Notes
Genomic/Lab
Claims
Procedures

SEER*DMS Enhancements

Usability assessment and human computer interaction design framework

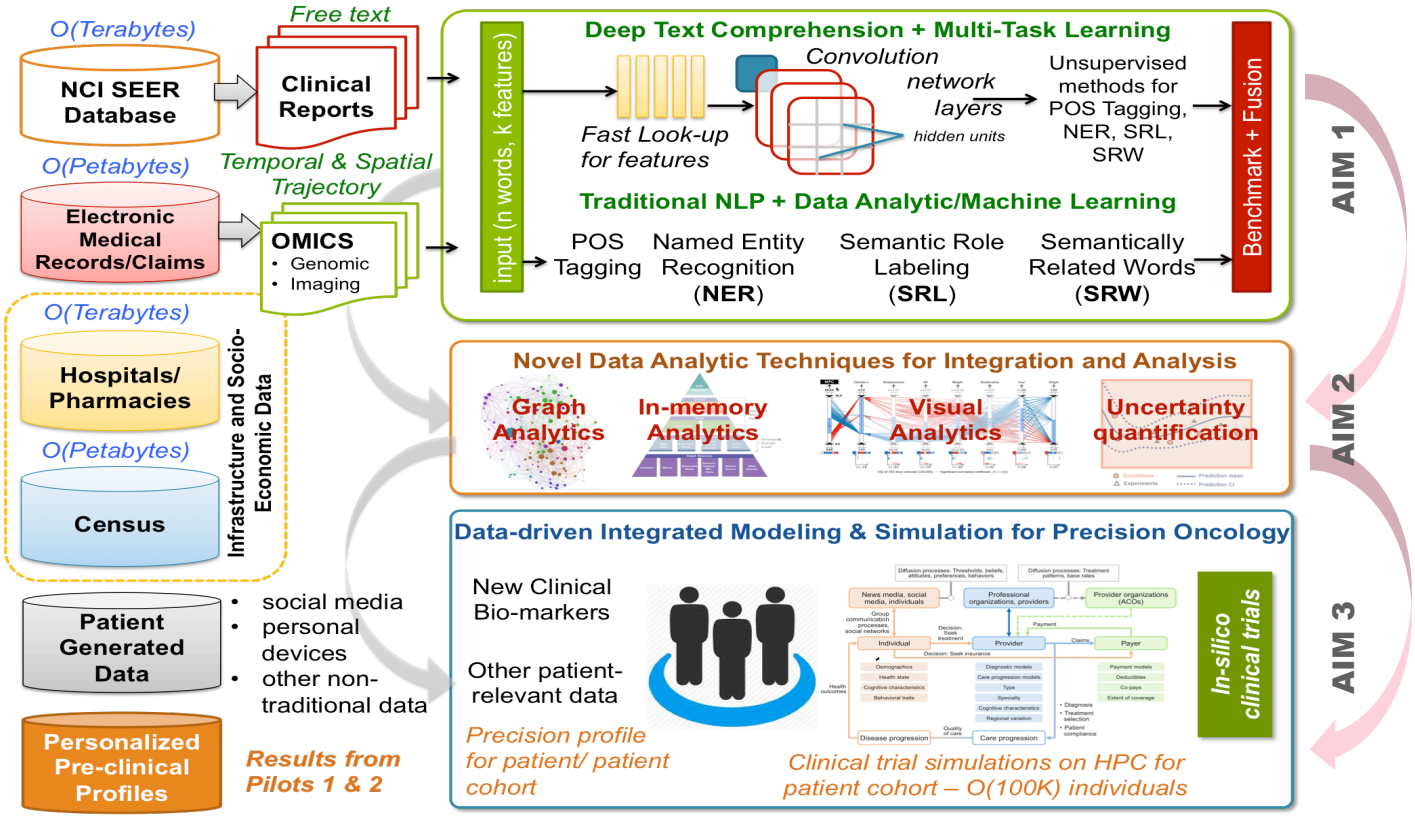


From Marina Matatova

NCI DOE Pilot 3: Overview

Population Information Integration, Analysis and Modeling

Improve the effectiveness of cancer treatment in the "real world" through computing



NCI-DOE Pilot 3 - Aims and Mission

Aims	NCI	DOE
Aim 1: Information Capture Deep text comprehension of unstructured clinical text to improve the capacity of the cancer surveillance program	Enhance breadth, quality and timeliness of data collected in SEER with confidence-rated (UQ) automated data element identification from unstructured data sources	Advance *descriptive* analytics with scalable deep learning based NLP tools for CORAL and exascale architectures
Aim 2: Information Integration Scalable graph, visual, and in-memory heterogeneous data analytics and inference methods to understand drivers in patterns of cancer outcomes and predict clinical endpoints	Integrate a variety of new data sources containing relevant clinical information feeding cancer patient surveillance databases to understand patient outcomes and cancer drivers in the real world population – beyond clinical trials	Advance *predictive* analytics with scalable statistical inference tools, graph and visual analytics for CORAL and exascale architectures
Aim 3: Population-level Modeling Data-driven modeling of patient-specific and population level cancer health trajectories to guide precision cancer care	Develop capability for data-driven modeling of individual patient outcomes using an integrative cancer patient model based on comprehensive data from Aims 1 and 2	Advance *prescriptive* analytics that involve large-scale data-driven modeling and simulation for CORAL and exascale architectures

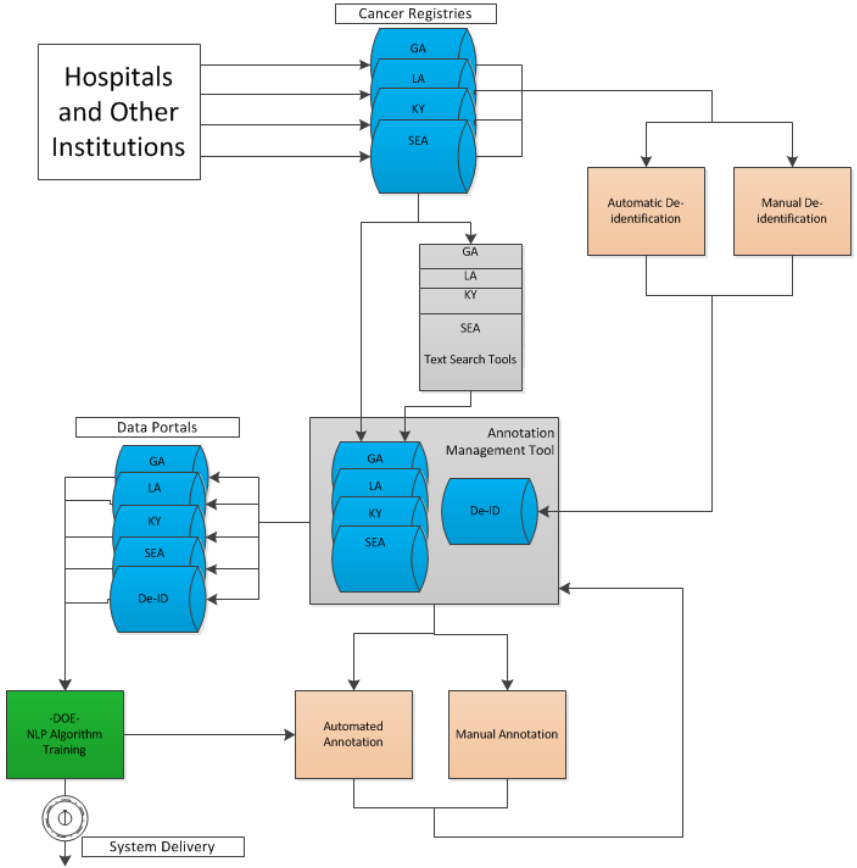
Challenges

NCI-DOE Pilot 3 Challenges

- Develop scalable NLP tools and processes
- Need for robust and scalable tool for annotation and review
  - Developing training and validation datasets through human annotation and review
  - External review and validation of human annotation
  - Review and validation of NLP algorithms with UQ
- Scale to multiple batches, schemas
- Tracking and sharing data through a secure data sharing portal

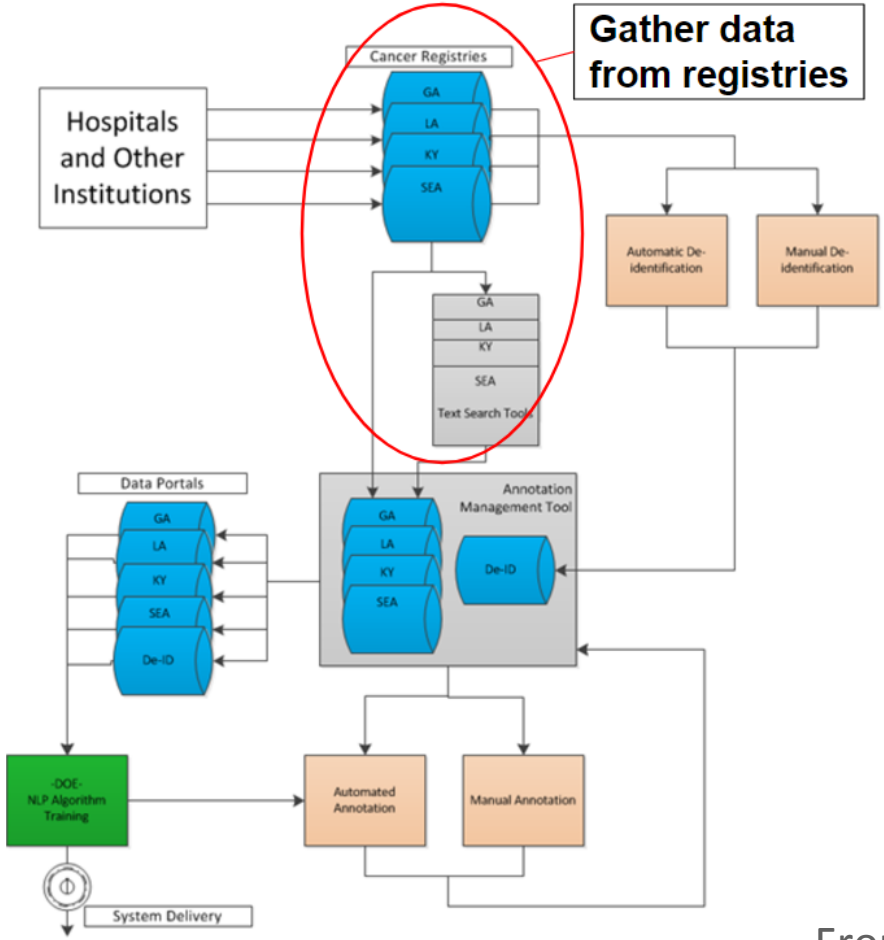
Role of LabKey

Clinical Document Annotation and Processing (CDAP) Pipeline



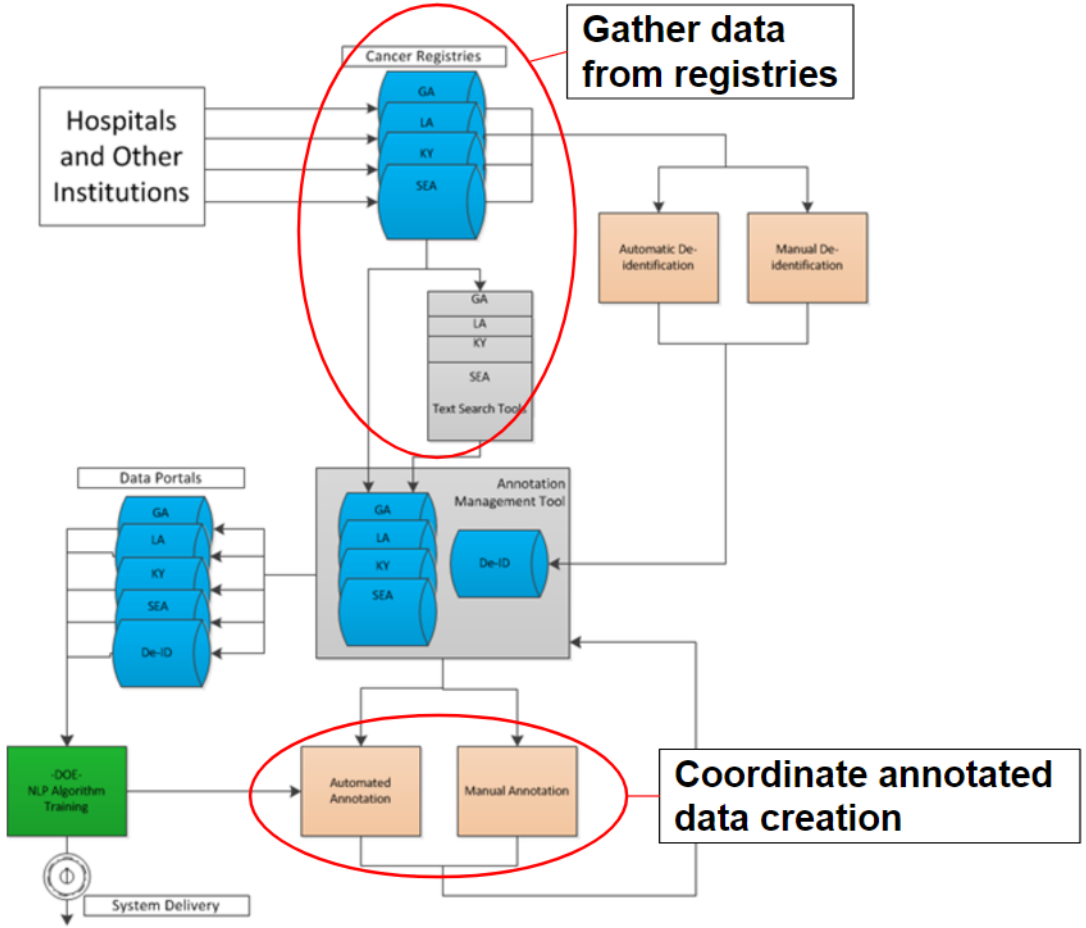
From Spencer Morris

CDAP Pipeline



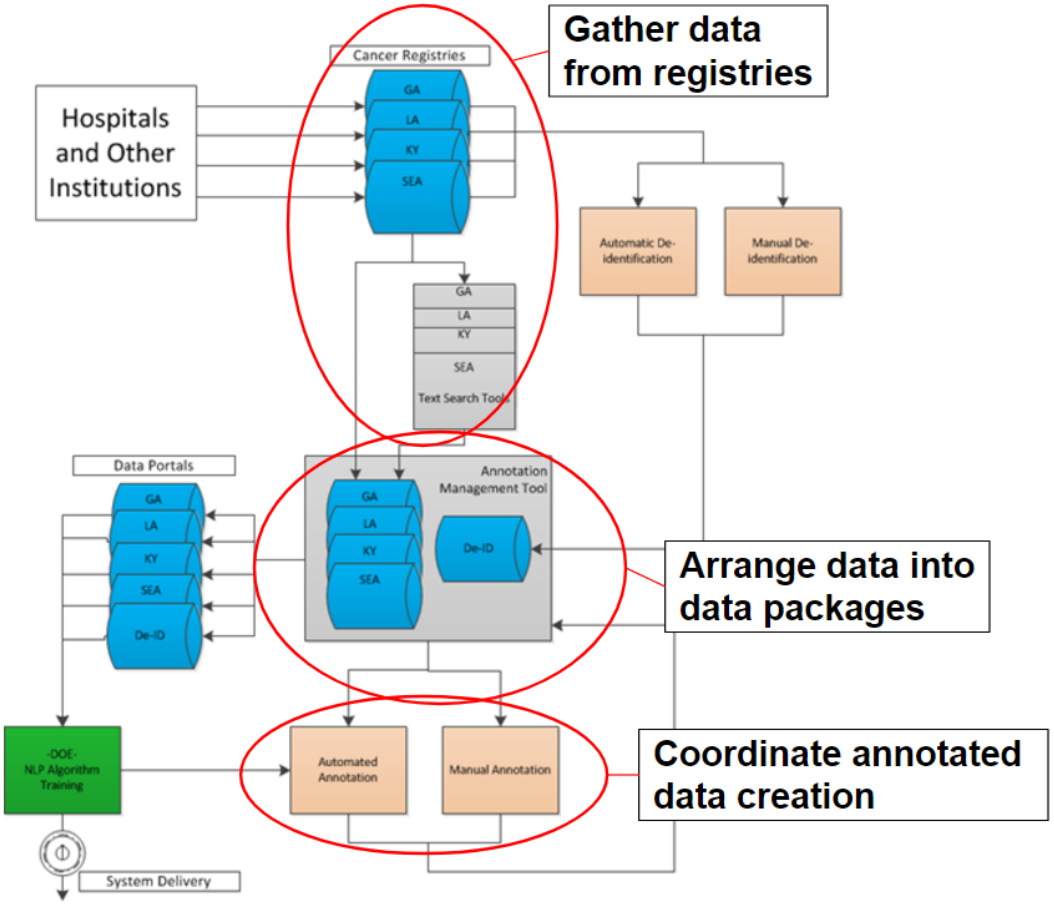
From Spencer Morris

CDAP Pipeline



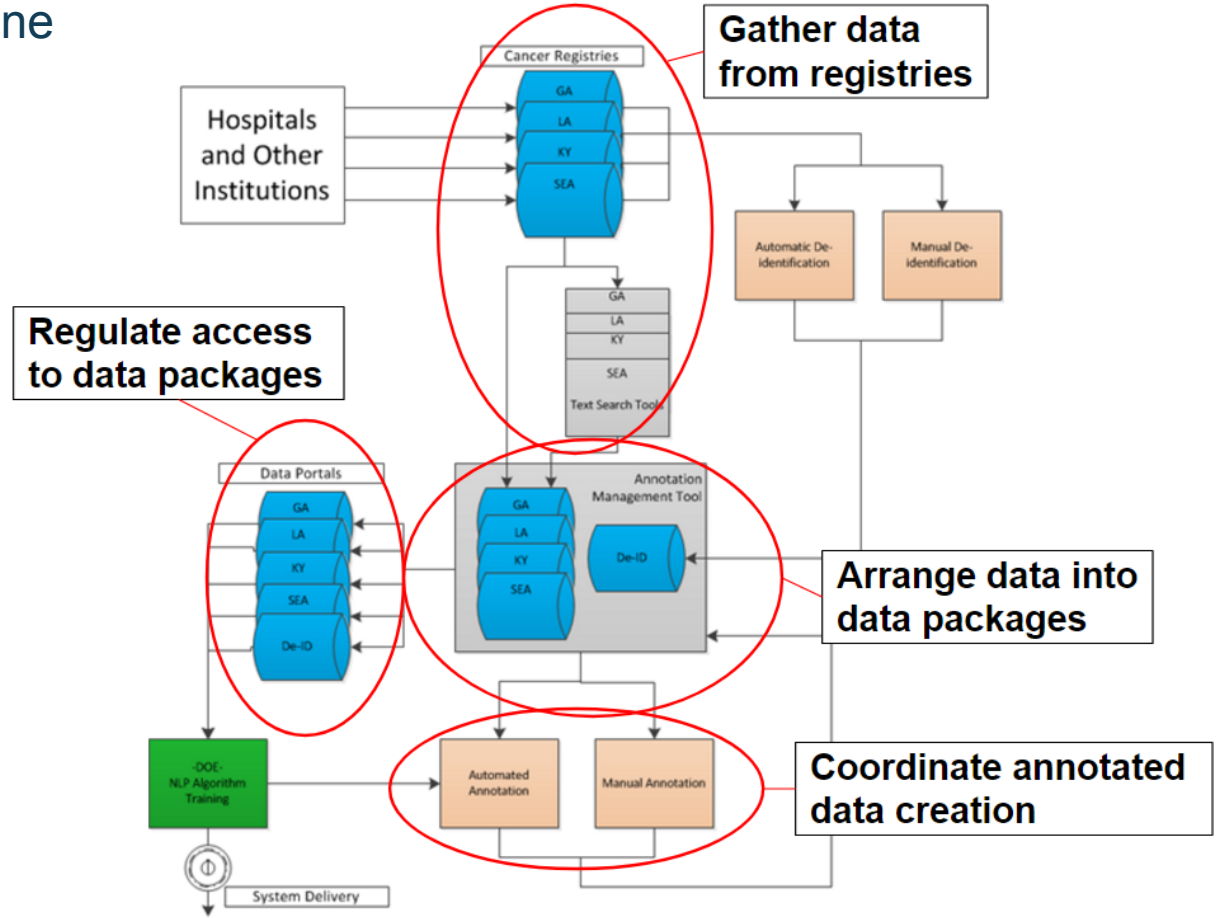
From Spencer Morris

CDAP Pipeline



From Spencer Morris

CDAP Pipeline



From Spencer Morris

Working List for Managers, Annotators, and Reviewers

Abstraction Task List

3 cases for me to assign

GRID VIEWS | REPORTS | CHARTS | EXPORT | PRINT | PAGING

		Tasks	Report Number	MRN	Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
☐	DETAILS	ASSIGN	01-REC-0000000441	484932	2016-10-08 21:07		Ready for assignment	Pathology Reports		
☐	DETAILS	ASSIGN	01-REC-0000000442	485263	2016-10-08 21:07		Ready for assignment	Pathology Reports		
☐	DETAILS	ASSIGN	01-REC-3000698470	10049074	2016-10-08 21:07		Ready for assignment	Pathology Reports		

2 cases for me to abstract

GRID VIEWS | REPORTS | CHARTS | EXPORT | PRINT | PAGING

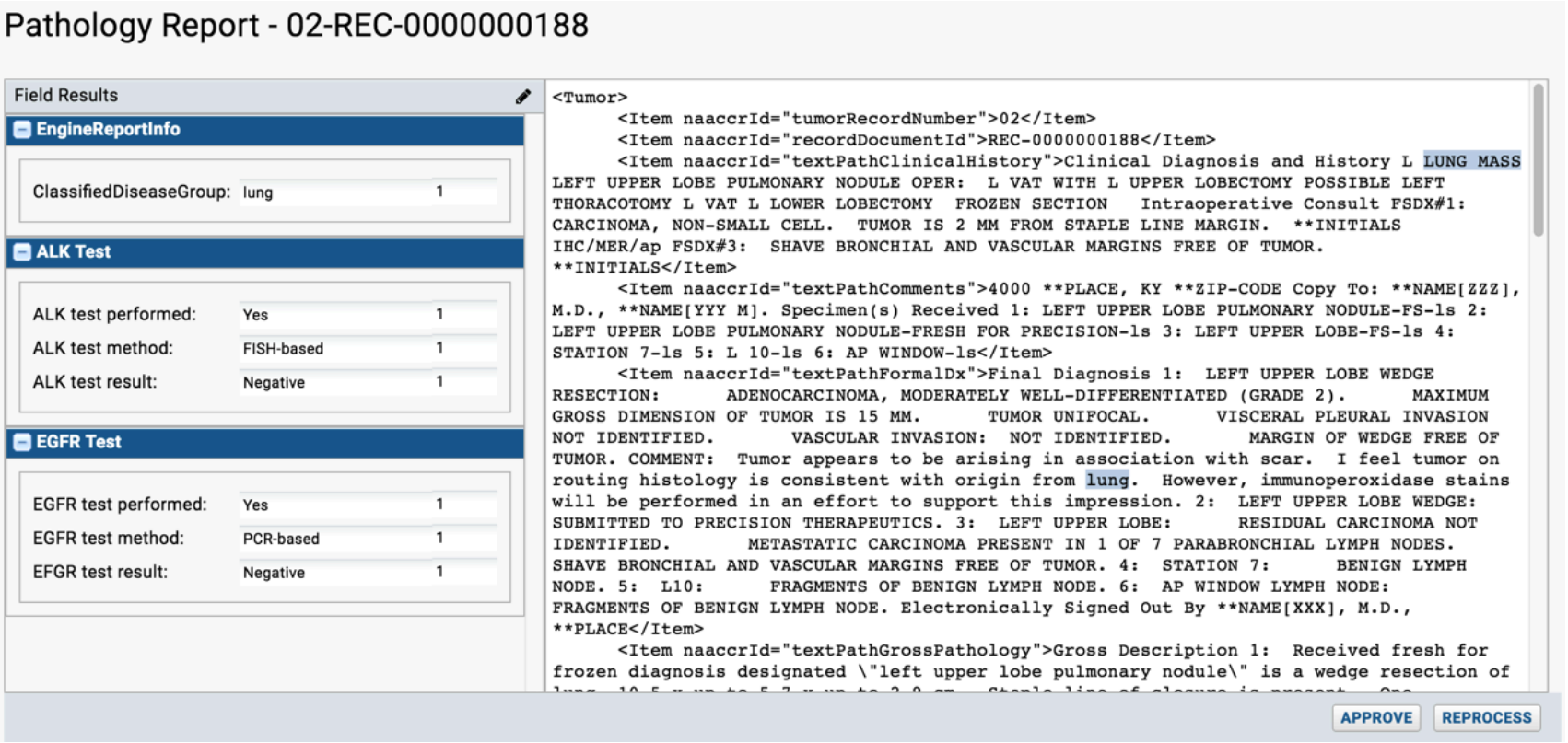
		Tasks	Report Number	MRN	Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
☐	DETAILS	ABSTRACT	02-REC-3000679345	01036995	2016-10-08 21:07		Ready for initial abstraction	Pathology Reports	kristinf	reviewer1
☐	DETAILS	ABSTRACT	02-REC-3000697853	00833487	2016-10-08 21:07		Ready for initial abstraction	Pathology Reports	kristinf	

3 cases for me to review

GRID VIEWS | REPORTS | CHARTS | EXPORT | PRINT | PAGING

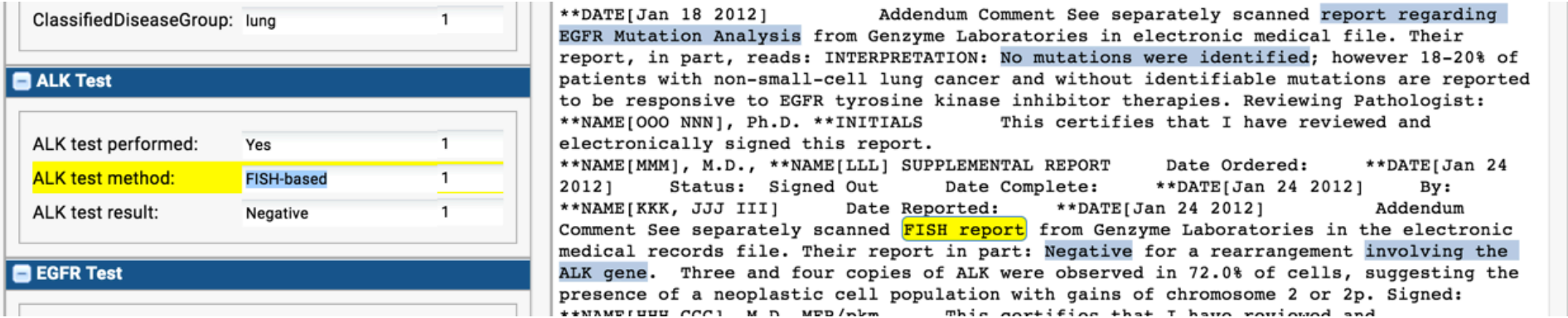
		Tasks	Report Number	MRN	Last Modified	Pulled Date	Status	Document Type	Abstractor	Reviewer
☐	DETAILS	REVIEW	01-REC-0000000443	485401	2016-10-08 21:08		Ready for review	Pathology Reports		kristinf
☐	DETAILS	REVIEW	01-REC-0000000445	486414	2016-10-08 21:09		Ready for review	Pathology Reports		kristinf
☐	DETAILS	REVIEW	01-REC-3000698369	10046190	2016-10-08 21:09		Ready for review	Pathology Reports		kristinf

Main Annotation UI



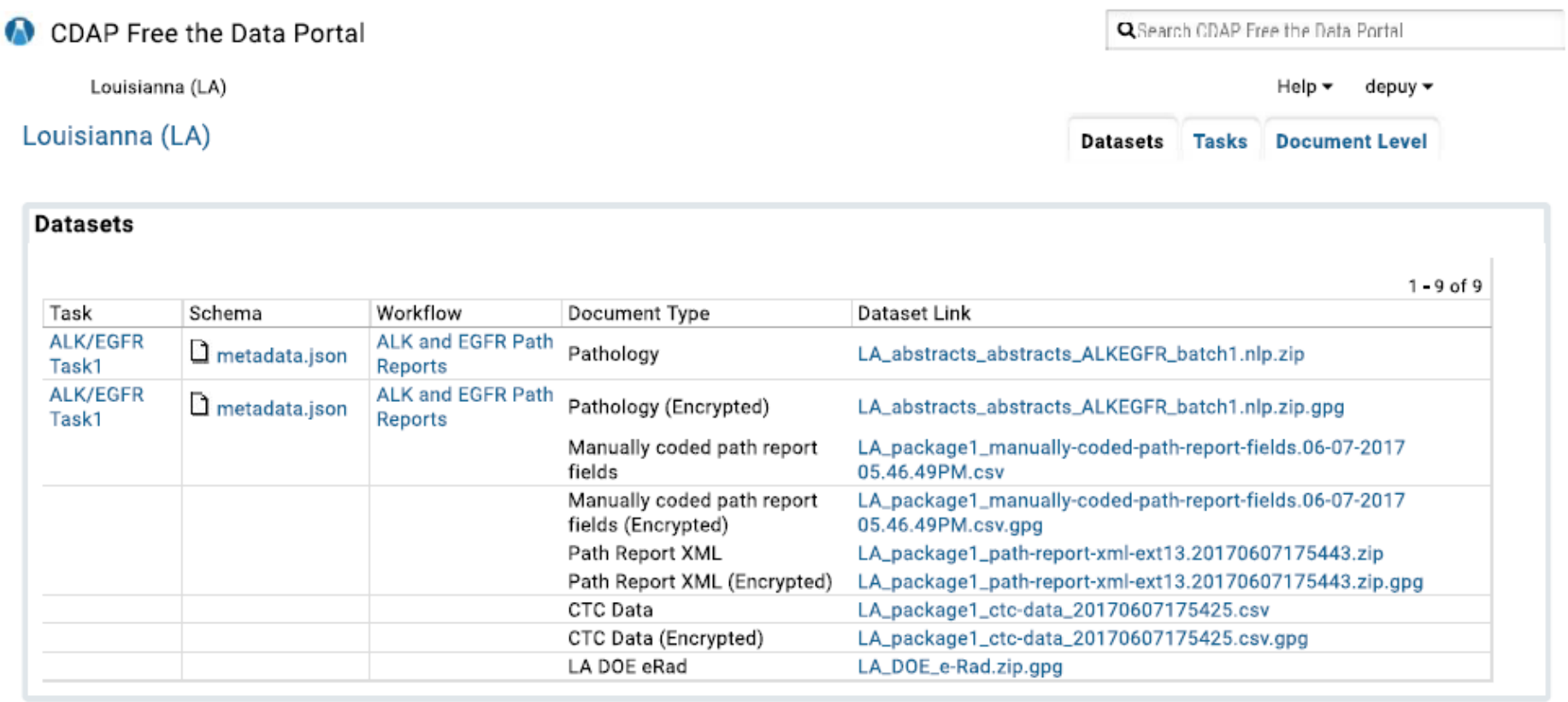
In view mode, an annotator can see the abstracted values on the left compared with the original document text on the right.

Highlight supporting text in source document

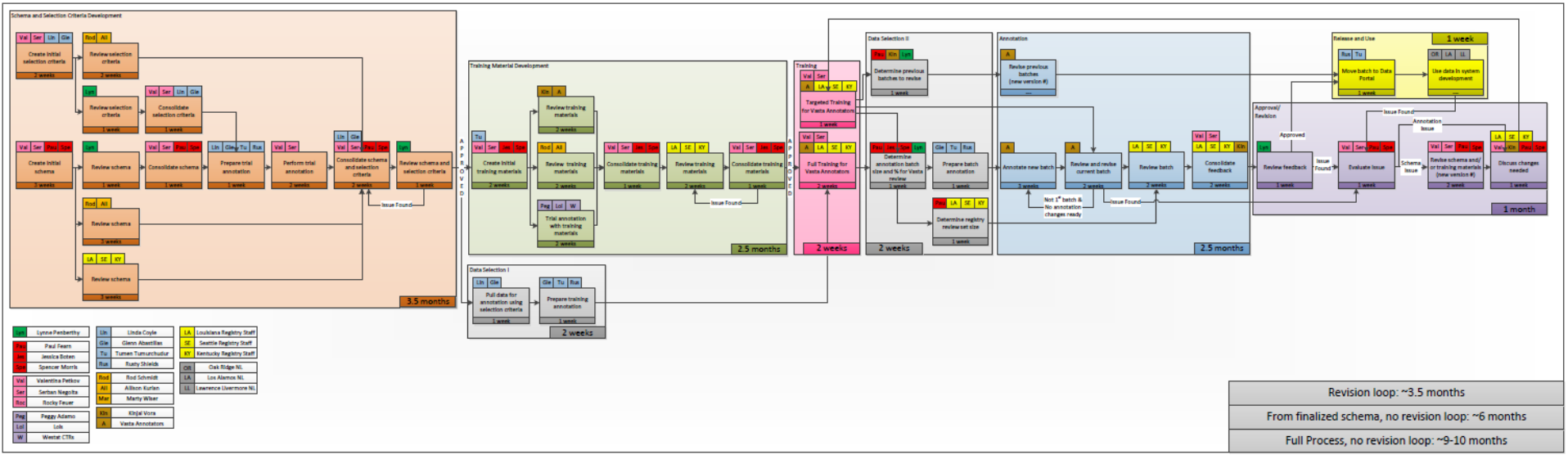


Clicking on a field result highlights all of the supporting strings of text in the document that provide evidence for the value (label) selected

“Free the Data” Portal



Complex Annotation Workflow



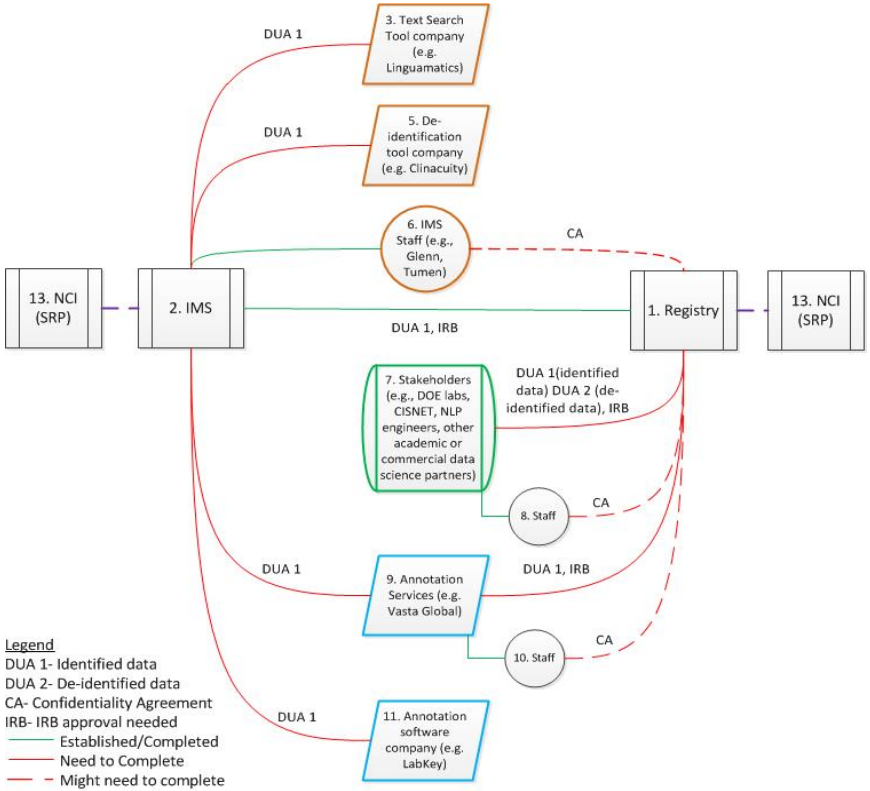
From Spencer Morris

Lessons Learned

Lessons Learned

- Have not scaled up volume yet; still scaling for complexity (variety), data quality (veracity), and velocity
- Looked at many annotation tools; none build for this scale
- Engaging LabKey more as a partner than a vendor

Results


DUA 1
3. Text Search Tool company (e.g. Linguamatics)
DUA 1
5. De-identification tool company (e.g. Clinacuity)
6. IMS Staff (e.g., Glenn, Tumen)
CA
13. NCI (SRP)
2. IMS
DUA 1, IRB
1. Registry
13. NCI (SRP)
7. Stakeholders (e.g., DOE labs, CISNET, NLP engineers, other academic or commercial data science partners)
DUA 1(identified data) DUA 2 (de-identified data), IRB
8. Staff
CA
DUA 1
9. Annotation Services (e.g. Vasta Global)
DUA 1, IRB
10. Staff
CA
DUA 1
11. Annotation software company (e.g. LabKey)
Legend
DUA 1- Identified data
DUA 2- De-identified data
CA- Confidentiality Agreement
IRB- IRB approval needed
Established/Completed
Need to Complete
Might need to complete
Contract Relationship
1 Master IRB File
- Appendix with each company, description of their security environment, and DUA
- Registry-specific CA or DUA for company staff that may access registry data
Registry Island - Production
Registry Island - CDAP
Shared Island – Free-the-Data Portals

Where we are now

- Multiple instances of LabKey up and running in secure enclaves for annotating documents from three SEER cancer registries
- 3 batches of ALK/EGFR annotated
- Enhancements in-progress

Enhancements and Support Completed for this project

1. Development of Batch Assignment of Cases
2. Development of Alternate Metadata Location Configuration
3. Development of NCI Folder Template
4. Development to Improve XML Document Display
5. Development to Specify Metadata Location, Document Type, Disease Group via Protocol
6. Development of Ability to Have Second and Third Review Step of Approved Results
7. Provision of Help Desk Support

Planned Phase 2 Enhancements In-Progress

Structured Abstract Result Export

Expert Services Provider shall implement a feature for authorized researchers to download the results associated with a particular annotation task in a structured format (LabKey's JSON format) so that is easy to import the data into external tools.

Process to Remove Non-printable Characters when Importing Text Documents

Expert Services Provider shall implement a process that removes any non-printable characters (e.g., anything less than ASCII 31) from imported text files.

Custom View Providing Task & Dataset Details

Expert Services Provider shall implement a feature in the annotation interface that gives authorized users a view of the Tasks and Datasets details in the Natural Language Processing (NLP) pipeline

Add Document Counts to Task & Dataset View

Expert Services Provider shall implement a feature in the annotation interface that allows authorized users to view document counts as a part of the Tasks and Datasets custom view in the Natural Language Processing (NLP) pipeline.

Select Multiple Values from a Dropdown

Expert Services Provider shall modify fields in the annotation interface to allow abstractors and reviewers to select one or more values from the list provided in the dropdown menu. The selected values must be delimited or maintained as a list of separate values.

Future plan

Next steps

- Annotation of path reports for breast cancer recurrence
- Review and validation of NLP and UQ from DOE labs
  - Site, laterality, histology, grade, behavior
- List of anticipated enhancements to CDAP pipeline in 2018

Potential Phase 3 Enhancements for 2018

- Integration with NLP algorithms from DOE labs
- Multiple independent annotators with adjudication and inter-annotator agreement reports
- Branching logic for annotation schemas
- Case-level annotation (multiple source documents)
- Refinements to annotation workflow and UI
- Refinements to the administration workflow and UI
- Annotation with hierarchical vocabulary in addition to by defined fields
- Annotation of relations
- Document lifecycle reporting
- Making workflows more automated and robust

Acknowledgements

- NCI Staff
  - Jessica Boten
  - Donna Rivera
  - Marina Matatova
  - Spencer Morris
- IMS
  - Linda Coyle
  - Glenn Abastillas
- DOE labs
- LabKey
  - Kristin Dubrule
  - Adam Rauch
  - Michael Gersch
- SEER Registries
  - Seattle, LA, KY, GA
- Other contributors
  - Emily Silgard

Questions?

Paul Fearn
paul.fearn@nih.gov

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol