DOE/NCI Aim 1
Natural Language Processing

Spencer Morris, MS

Jessica Boten, MPH

NIH NATIONAL CANCER INSTITUTE

September 21, 2017

Natural Language Processing (NLP)

- Software to perform tasks involving (human) language
- Common examples:
  - Spell checker
  - Translation software
  - Siri
- Medical domain
  - Free-text reports
  - Clinical notes, pathology reports

The Problem

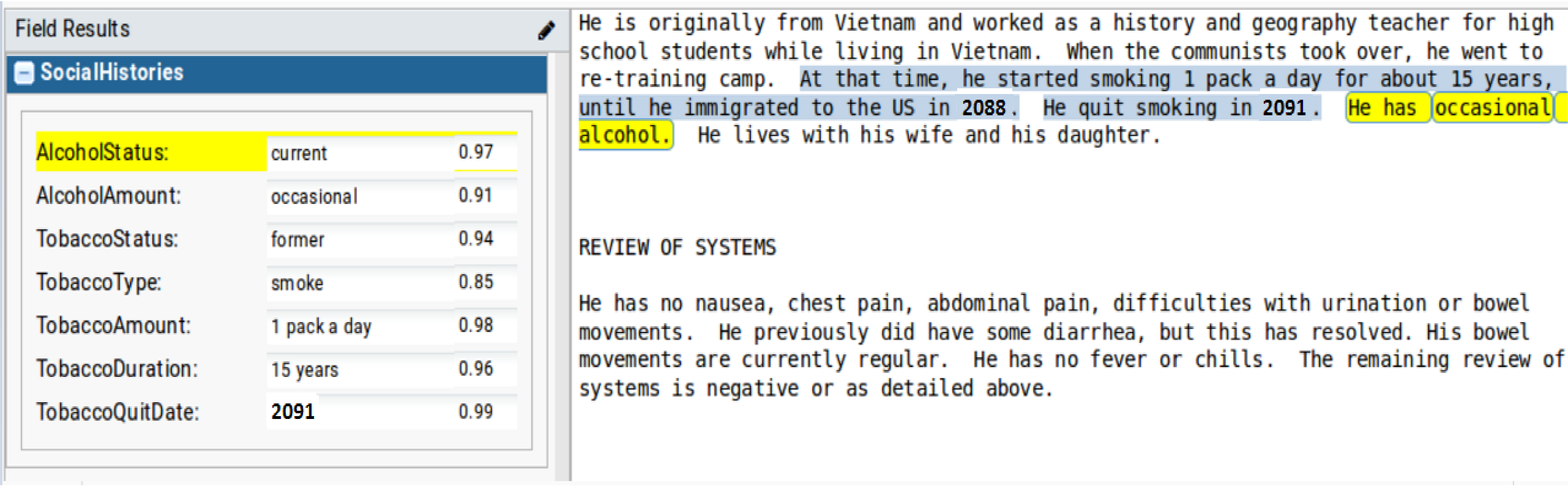
- Cancer researchers want data
  - More data → better research → better patient outcomes
  - Issue: information stored in free text
    - Manual abstraction – time-consuming
    - Results in small amount of data available

Our Solution

- Automatic information extraction with NLP
    - Automate where feasible
        - Greatly increase rate of abstraction
    - Maintain human experts in the loop
        - Ensure quality
        - Handle what current machines can’t

Example System in Action

- Alcohol/Tobacco Use:



Field Results

SocialHistories

AlcoholStatus:	current	0.97
AlcoholAmount:	occasional	0.91
TobaccoStatus:	former	0.94
TobaccoType:	smoke	0.85
TobaccoAmount:	1 pack a day	0.98
TobaccoDuration:	15 years	0.96
TobaccoQuitDate:	2091	0.99

He is originally from Vietnam and worked as a history and geography teacher for high school students while living in Vietnam. When the communists took over, he went to re-training camp. At that time, he started smoking 1 pack a day for about 15 years, until he immigrated to the US in 2088. He quit smoking in 2091. He has occasional alcohol. He lives with his wife and his daughter.

REVIEW OF SYSTEMS

He has no nausea, chest pain, abdominal pain, difficulties with urination or bowel movements. He previously did have some diarrhea, but this has resolved. His bowel movements are currently regular. He has no fever or chills. The remaining review of systems is negative or as detailed above.

Current Stage

- Create NLP systems
- Verify NLP system performance

Annotated Data

- NLP systems need annotated data
  - Data with the desired info recorded by humans
  - Example annotated data – Sentence Has Tobacco Information (Yes/No)

Original Data	Annotation – Has Tobacco Info
Patient drinks heavily.	No
Smokes heavily.	Yes
Smokes 3 packs/day	Yes
Enjoys an occasional cigar	Yes

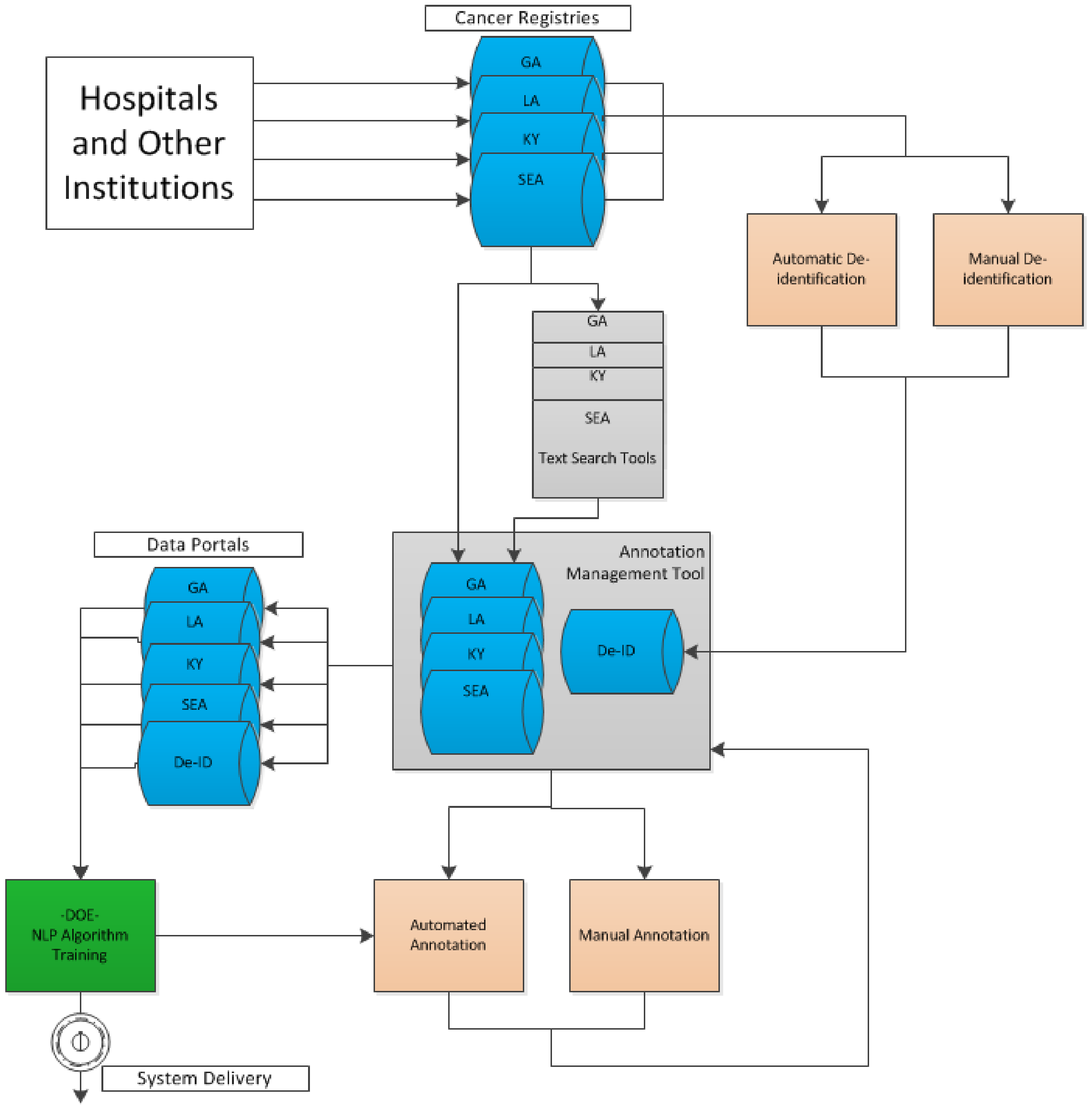
  - System sees many of these examples, learns patterns
    - More examples → learns better

Clinical Document Annotation and Processing Pipeline

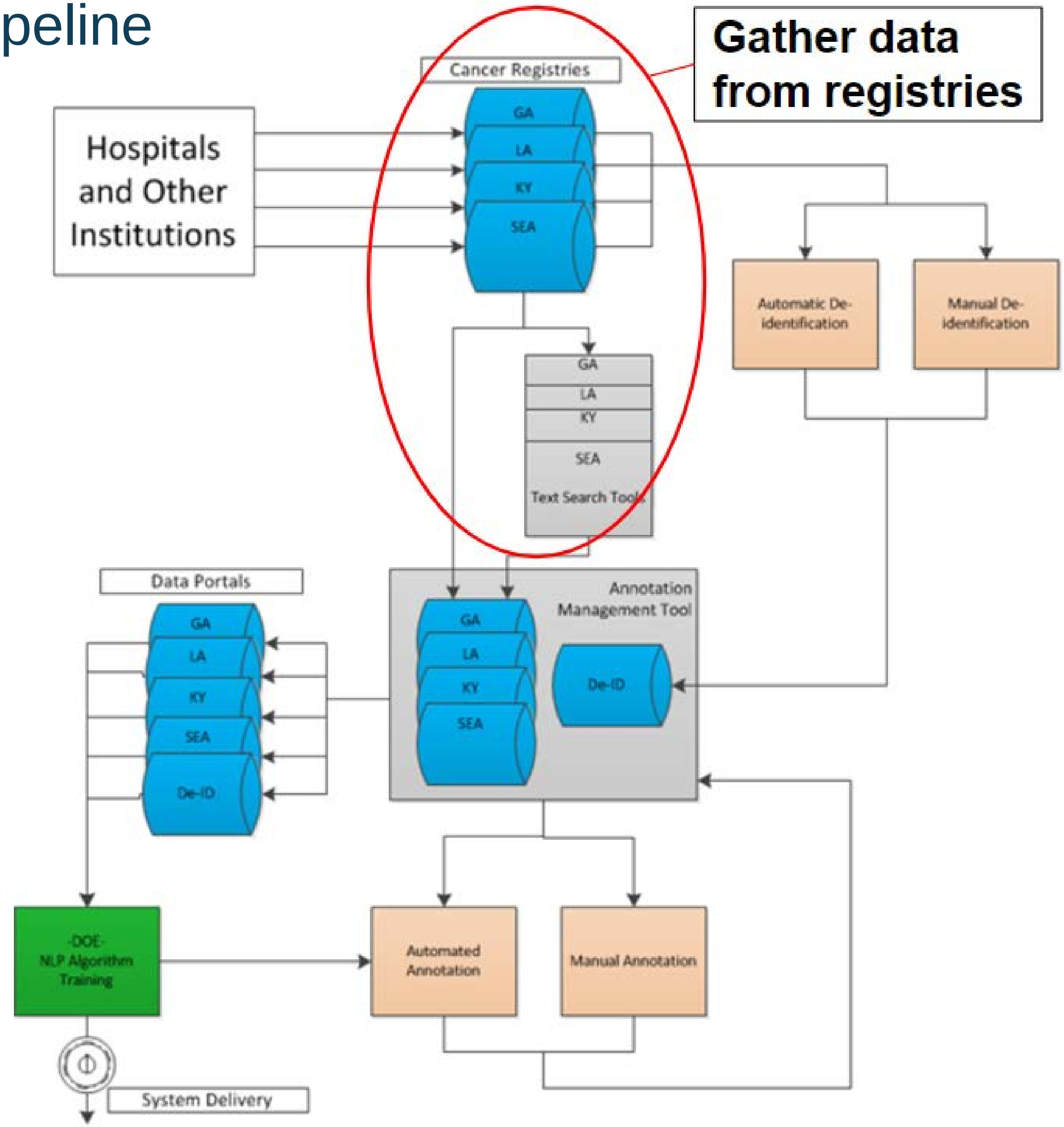
- Issue: NLP systems need annotated data
- Issue: NLP developers have limited access to medical data
  - Restrictions on sharing health data
  - Small amount of existing annotations

- Solution: Clinical Document Annotation and Processing (CDAP) Pipeline
  - Goal: get data to NLP developers
  - Pilot project with Louisiana, Kentucky, Seattle, and Georgia registries

CDAP Pipeline

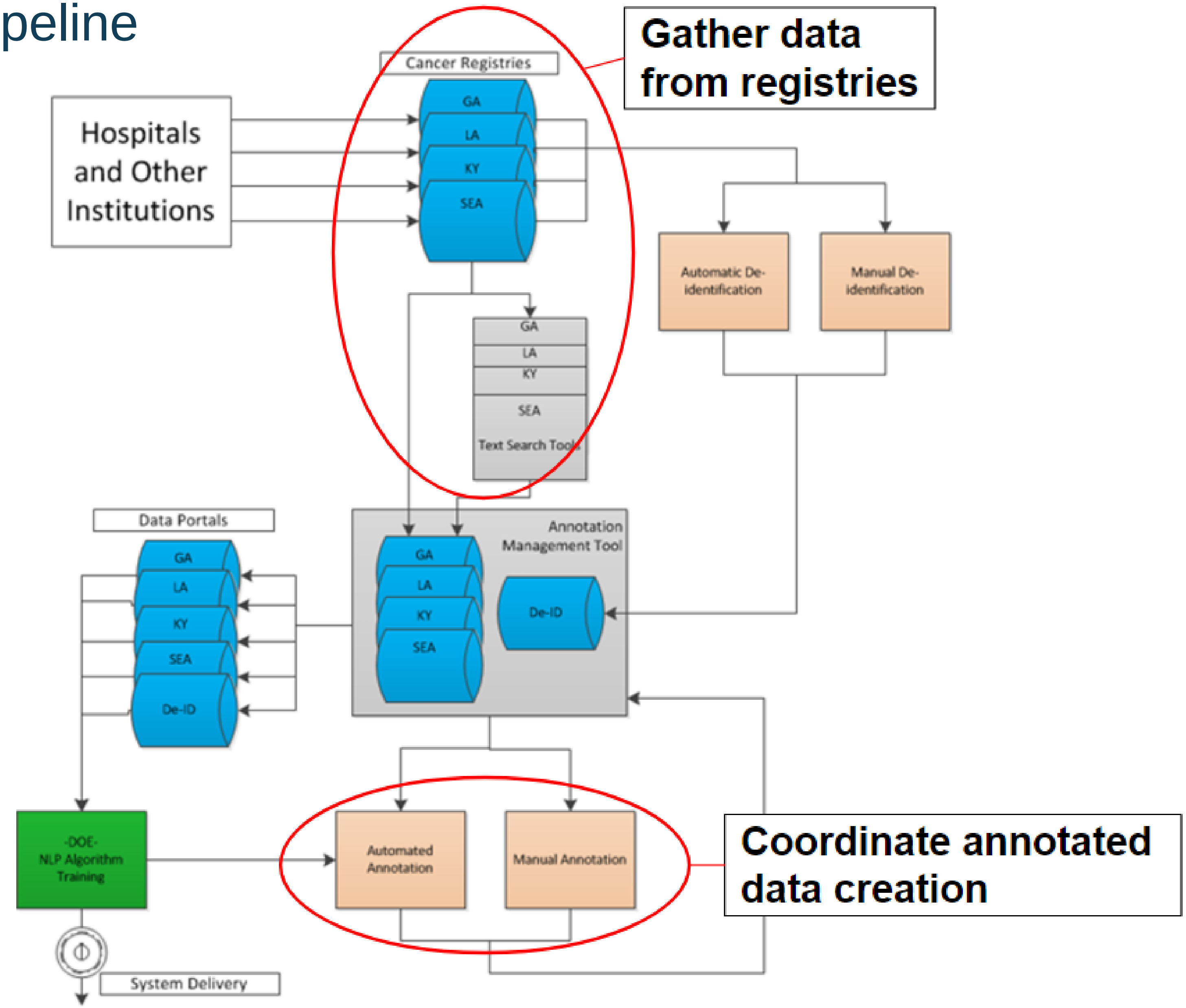


CDAP Pipeline



Gather data from registries

CDAP Pipeline



Gather data from registries

Coordinate annotated data creation

Hospitals and Other Institutions

Cancer Registries

GA

LA

KY

SEA

Automatic De-identification

Manual De-identification

GA

LA

KY

SEA

Text Search Tools

Data Portals

GA

LA

KY

SEA

De-ID

Annotation Management Tool

GA

LA

KY

SEA

De-ID

-DOE-

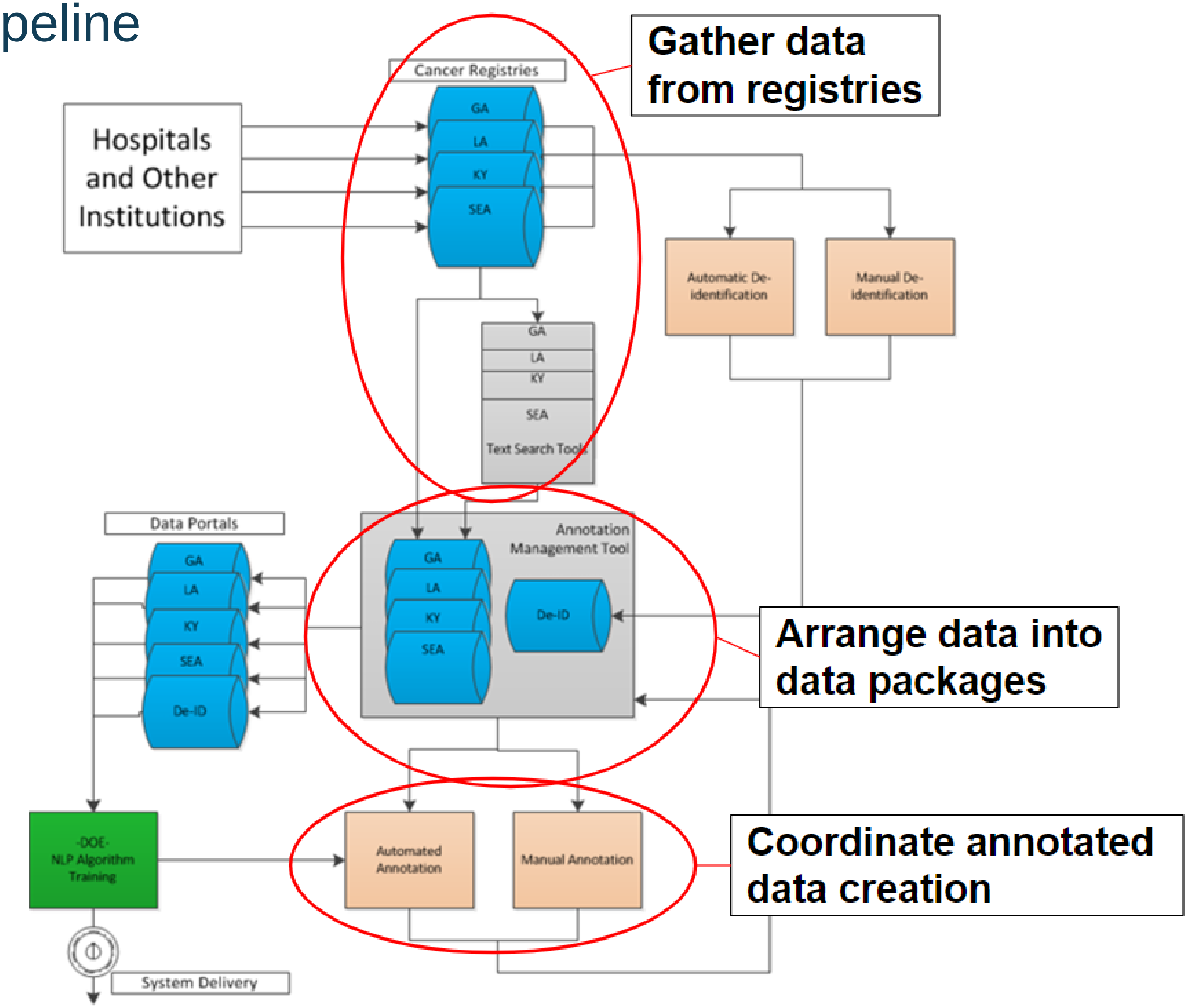
NLP Algorithm Training

Automated Annotation

Manual Annotation

System Delivery

CDAP Pipeline

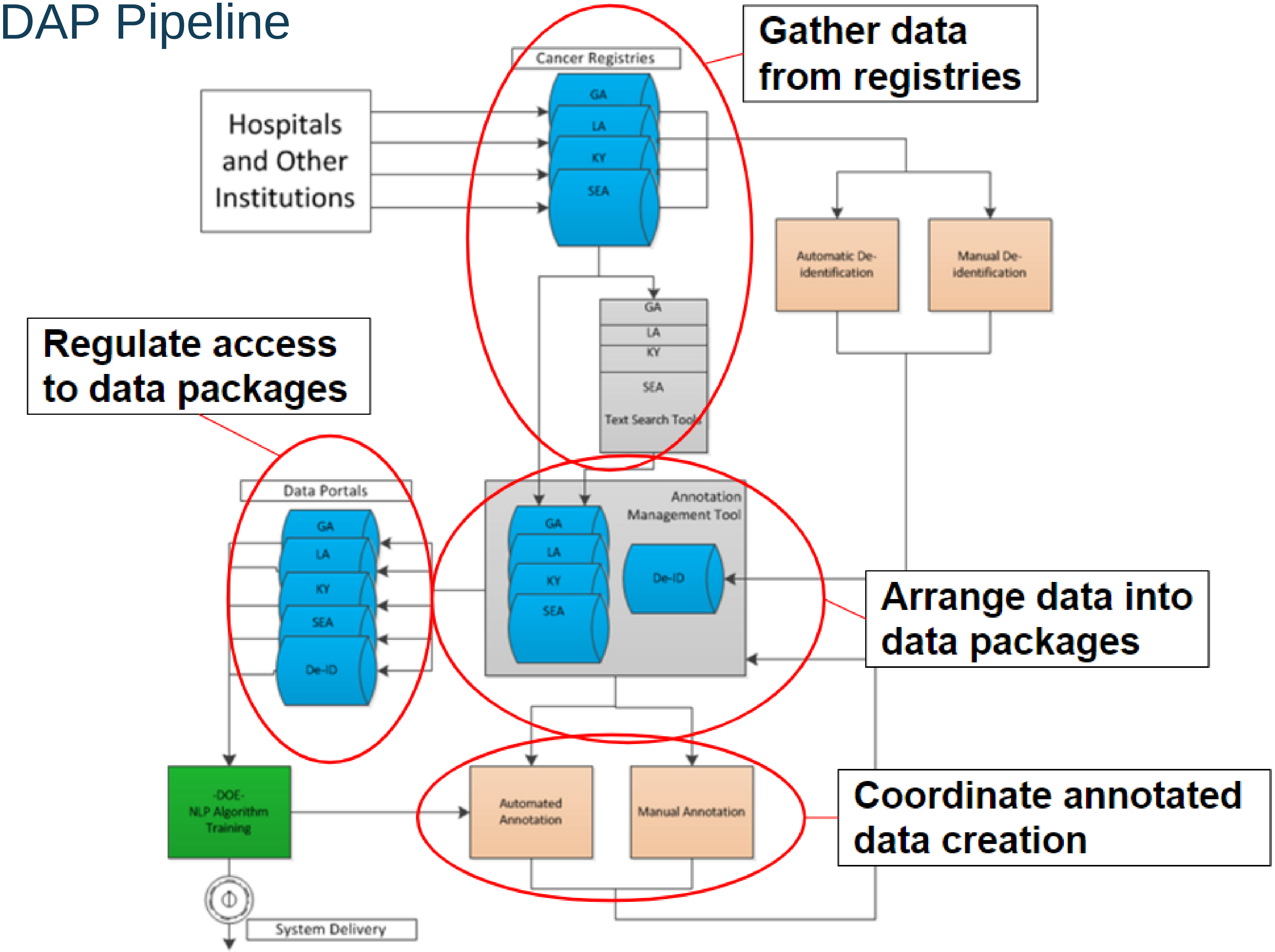


Gather data from registries

Arrange data into data packages

Coordinate annotated data creation

CDAP Pipeline



Gather data from registries

Regulate access to data packages

Arrange data into data packages

Coordinate annotated data creation

DOE Collaboration

- NCI will generate large amounts of annotated data

- Partnership with Department of Energy (DOE) labs
  - First NLP developers to use the CDAP pipeline
  - Can leverage huge amounts of data
    - Supercomputing infrastructure
    - Advanced techniques in artificial intelligence (AI)

DOE Collaboration

- Pilot projects:
    - Extract information about biomarkers of cancer
        - ALK and EGFR
        - HER2, ER, PR, Ki-67
    - Extract information about recurrence/progression of cancer
        - Breast and colorectal cancer
    - Model recurrence/progression of cancer with AI

Registry Review

- Participating registries conducting review of:
  - Annotations used to train NLP systems
  - NLP system output
- Current workload for review:
  - ~1 hour per 1.5 months
  - Registries decide how much to review
- Development and review take place in CDAP Pipeline environment
- Until registries are confident in system performance

Future Stage

- Integrate NLP systems to facilitate registry work

Incorporating NLP Systems

- Goal: Incorporate NLP systems in SEER*DMS for registries in the pilot project
  - Automate elements that machines can handle consistently
  - Registrars can focus on variables that need human abstraction

NLP Algorithms with Uncertainty Quantifications being developed by DOE Labs

Data Elements:

- Primary Site
- Histology
- Laterality
- Behavior
- Grade

Cancer Sites:

- Breast
- Colorectal
- Lung
- Prostate

Incorporating NLP Systems

- Uncertainty Quantification
  - Different performance for different tasks
  - Machines provide confidence score for each element
- Maintain human experts in the loop
  - Low confidence score → human expert takes over
  - Reliability of confidence score verified in review
  - Humans tackle elements that current machines can’t
  - Machine-assisted manual extraction

The Future: Integrate NLP system to facilitate registry work

- Need: to transfer algorithms and tools that the DOE develops and integrate them into the registry workflow

- Issue: Different computing environments → way to transfer the algorithms in a way that it works in both environments

The Future: Integrate NLP System to facilitate registry work

- Solution: iterative process between the IMS, DOE labs, and registries
  - API & Docker Container
    - Transfer from the DOE
    - IMS can test it in the new computing environment
    - Registries can test the algorithms and help refine them
      - CDAP Pipeline

- Long term goal: incorporate refined tools into SEER*DMS workflow

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH NATIONAL CANCER INSTITUTE

www.cancer.gov www.cancer.gov/espanol