Parallel Training of Large Knowledge Graph Convolutional Networks

Hong-Jun Yoon, PhD

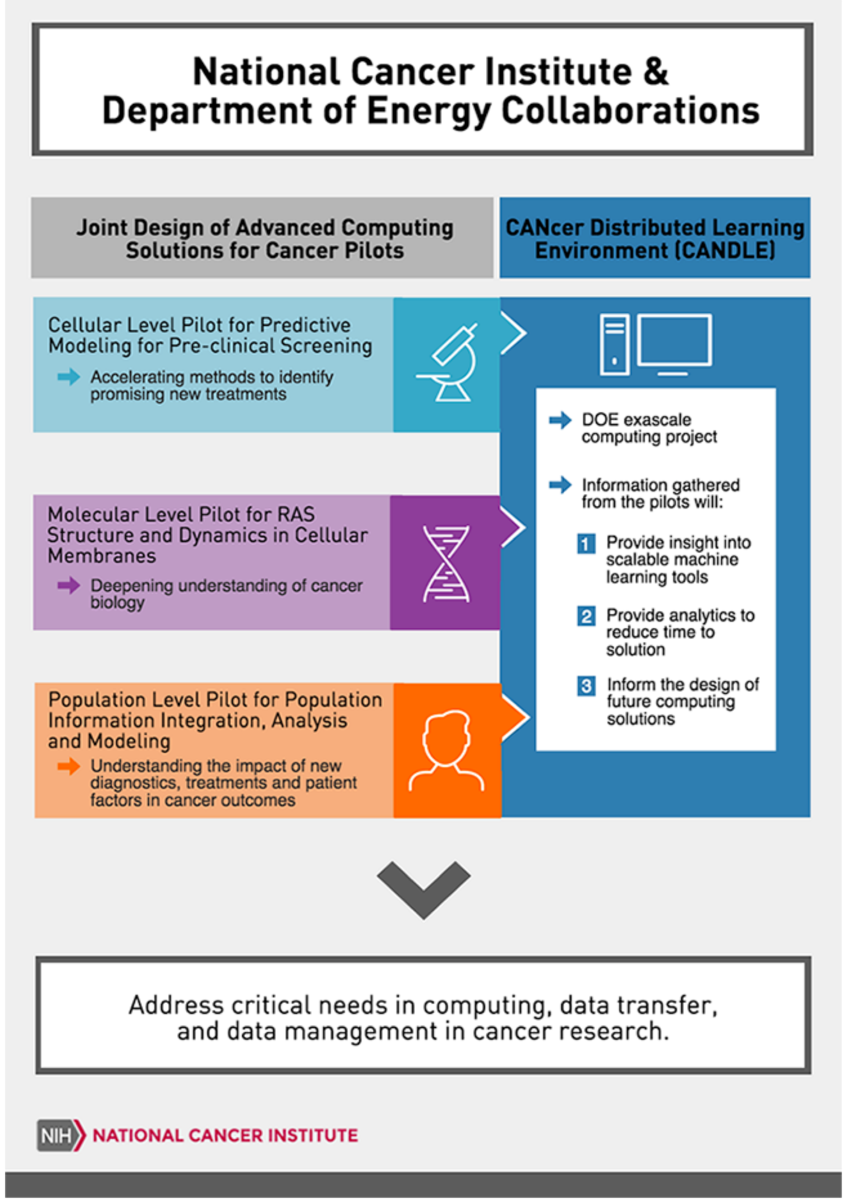
2020 Performance, Portability, and Productivity in HPC Forum

September 2020

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

- US Department of Energy/National Cancer Institute Collaboration
- Three Pilots
  - Pilot 1: Cellular
  - Pilot 2: Molecular
  - Pilot 3: Population-level (ORNL leading)
- Our Team
  - NLP
  - Information Extraction
  - Knowledge Discovery
  - Hypothesis Testing



National Cancer Institute & Department of Energy Collaborations

Joint Design of Advanced Computing Solutions for Cancer Pilots

CANcer Distributed Learning Environment (CANDLE)

Cellular Level Pilot for Predictive Modeling for Pre-clinical Screening

→ Accelerating methods to identify promising new treatments

Molecular Level Pilot for RAS Structure and Dynamics in Cellular Membranes

→ Deepening understanding of cancer biology

Population Level Pilot for Population Information Integration, Analysis and Modeling

→ Understanding the impact of new diagnostics, treatments and patient factors in cancer outcomes

→ DOE exascale computing project

→ Information gathered from the pilots will:

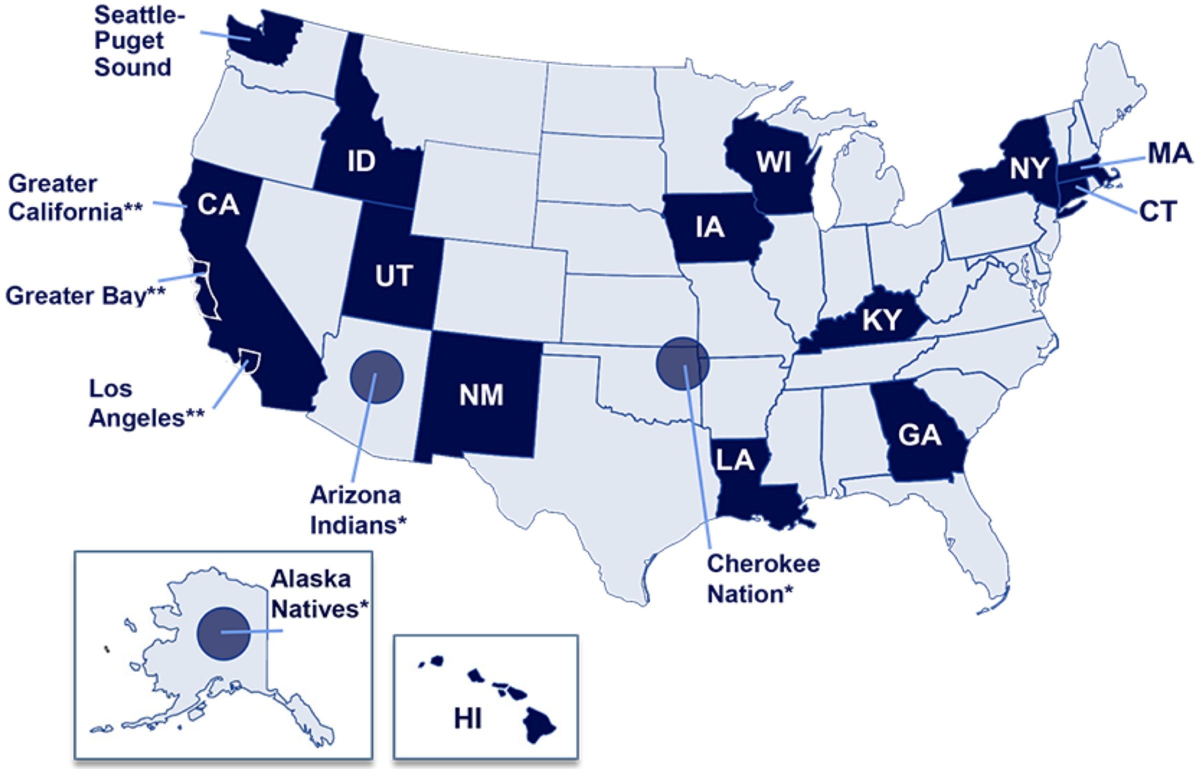
1. Provide insight into scalable machine learning tools
2. Provide analytics to reduce time to solution
3. Inform the design of future computing solutions

Address critical needs in computing, data transfer, and data management in cancer research.

NIH NATIONAL CANCER INSTITUTE

Data Sources

- NCI Surveillance, Epidemiology, and End Results (SEER) Program
  - Since 1973
  - 450,000+ cases / year
  - 1/3 US population

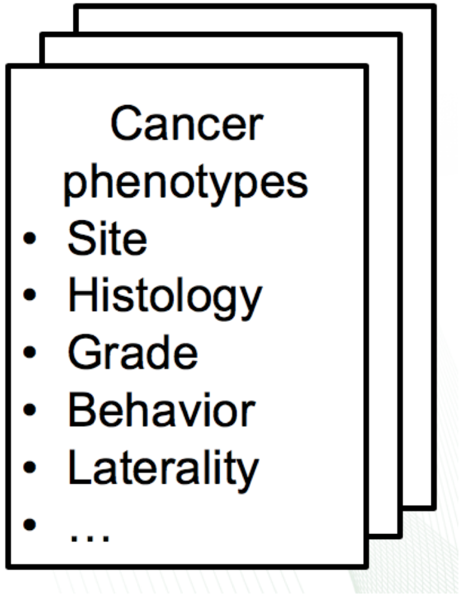


*Subcontract under New Mexico

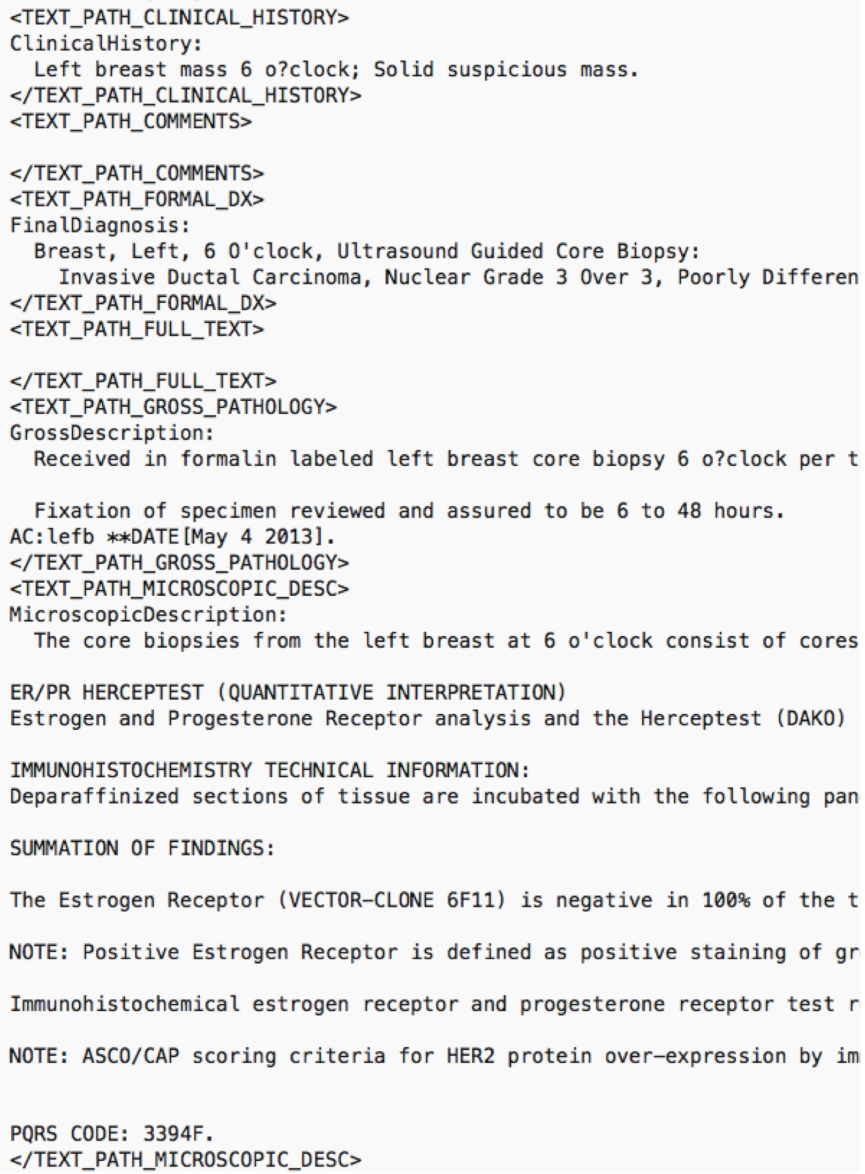
**Three regions represent the state of California: Greater Bay, Los Angeles, and Greater California

Cancer Pathology Reports

- Automated information extraction
  - Replace manual or rule-based approaches
  - Scalable training of solutions
  - Deploy API to SEER registries



Cancer phenotypes
- Site
- Histology
- Grade
- Behavior
- Laterality
- …



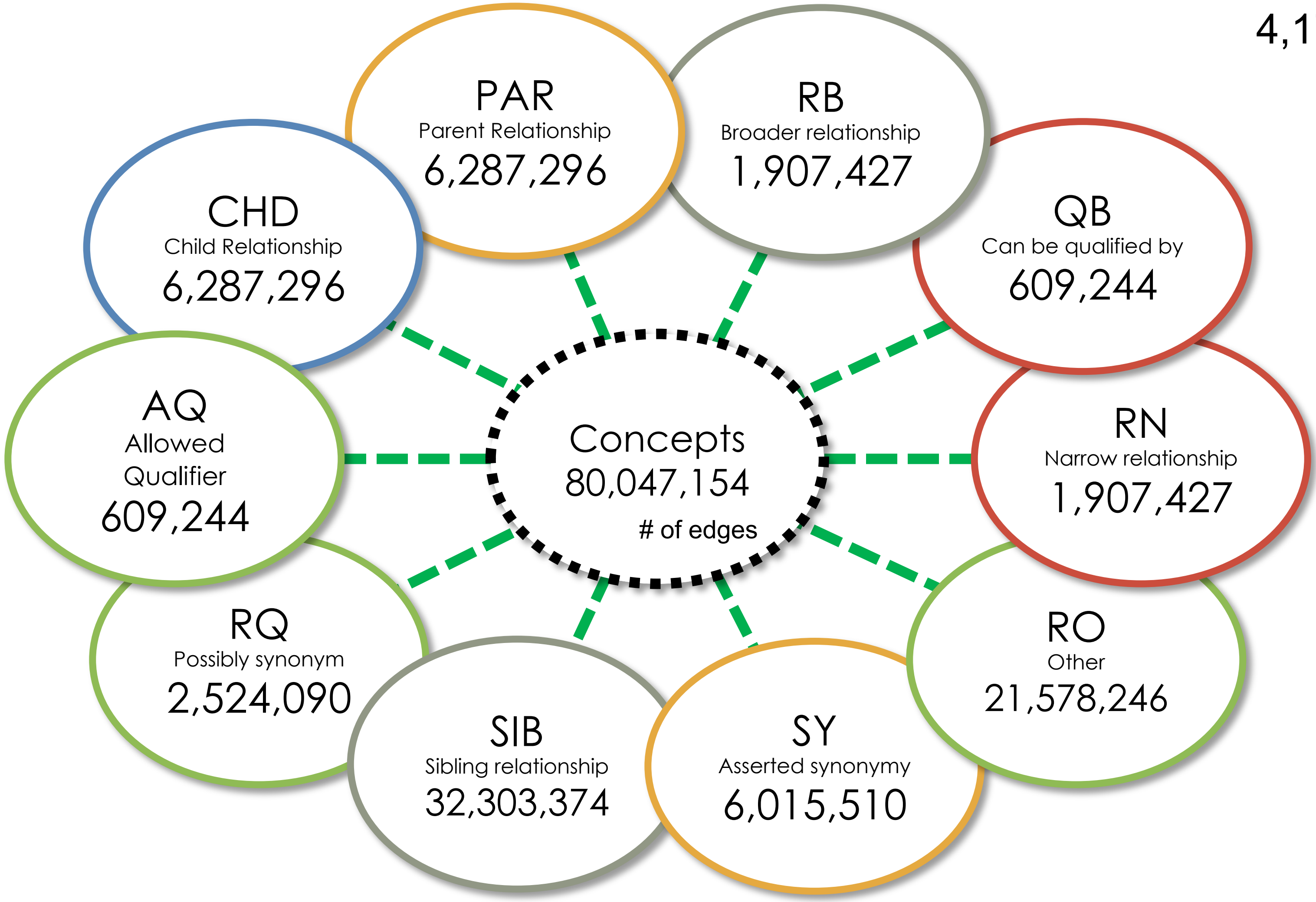
```
<TEXT_PATH_CLINICAL_HISTORY>
ClinicalHistory:
  Left breast mass 6 o?clock; Solid suspicious mass.
</TEXT_PATH_CLINICAL_HISTORY>
<TEXT_PATH_COMMENTS>

</TEXT_PATH_COMMENTS>
<TEXT_PATH_FORMAL_DX>
FinalDiagnosis:
  Breast, Left, 6 O'clock, Ultrasound Guided Core Biopsy:
    Invasive Ductal Carcinoma, Nuclear Grade 3 Over 3, Poorly Differen
</TEXT_PATH_FORMAL_DX>
<TEXT_PATH_FULL_TEXT>

</TEXT_PATH_FULL_TEXT>
<TEXT_PATH_GROSS_PATHOLOGY>
GrossDescription:
  Received in formalin labeled left breast core biopsy 6 o?clock per t

  Fixation of specimen reviewed and assured to be 6 to 48 hours.
AC:lefb **DATE[May 4 2013].
</TEXT_PATH_GROSS_PATHOLOGY>
<TEXT_PATH_MICROSCOPIC_DESC>
MicroscopicDescription:
  The core biopsies from the left breast at 6 o'clock consist of cores

ER/PR HERCEPTEST (QUANTITATIVE INTERPRETATION)
Estrogen and Progesterone Receptor analysis and the Herceptest (DAKO)

IMMUNOHISTOCHEMISTRY TECHNICAL INFORMATION:
Deparaffinized sections of tissue are incubated with the following pan

SUMMATION OF FINDINGS:

The Estrogen Receptor (VECTOR-CLONE 6F11) is negative in 100% of the t

NOTE: Positive Estrogen Receptor is defined as positive staining of gr

Immunohistochemical estrogen receptor and progesterone receptor test r

NOTE: ASCO/CAP scoring criteria for HER2 protein over-expression by im


PQRS CODE: 3394F.
</TEXT_PATH_MICROSCOPIC_DESC>
```

Under-represented Classes

- Rare cancers
  - Low incidence
  - Low number of training samples
  - Not enough to train our DL models
  - Low classification accuracy
- Solution
  - Import external knowledge sources
  - Knowledge graph, graph convolutional networks

UMLS Concept Relation Graph

4,177,638 Nodes (CUI's)



Concepts
80,047,154
# of edges

PAR
Parent Relationship
6,287,296

RB
Broader relationship
1,907,427

QB
Can be qualified by
609,244

RN
Narrow relationship
1,907,427

RO
Other
21,578,246

SY
Asserted synonymy
6,015,510

SIB
Sibling relationship
32,303,374

RQ
Possibly synonym
2,524,090

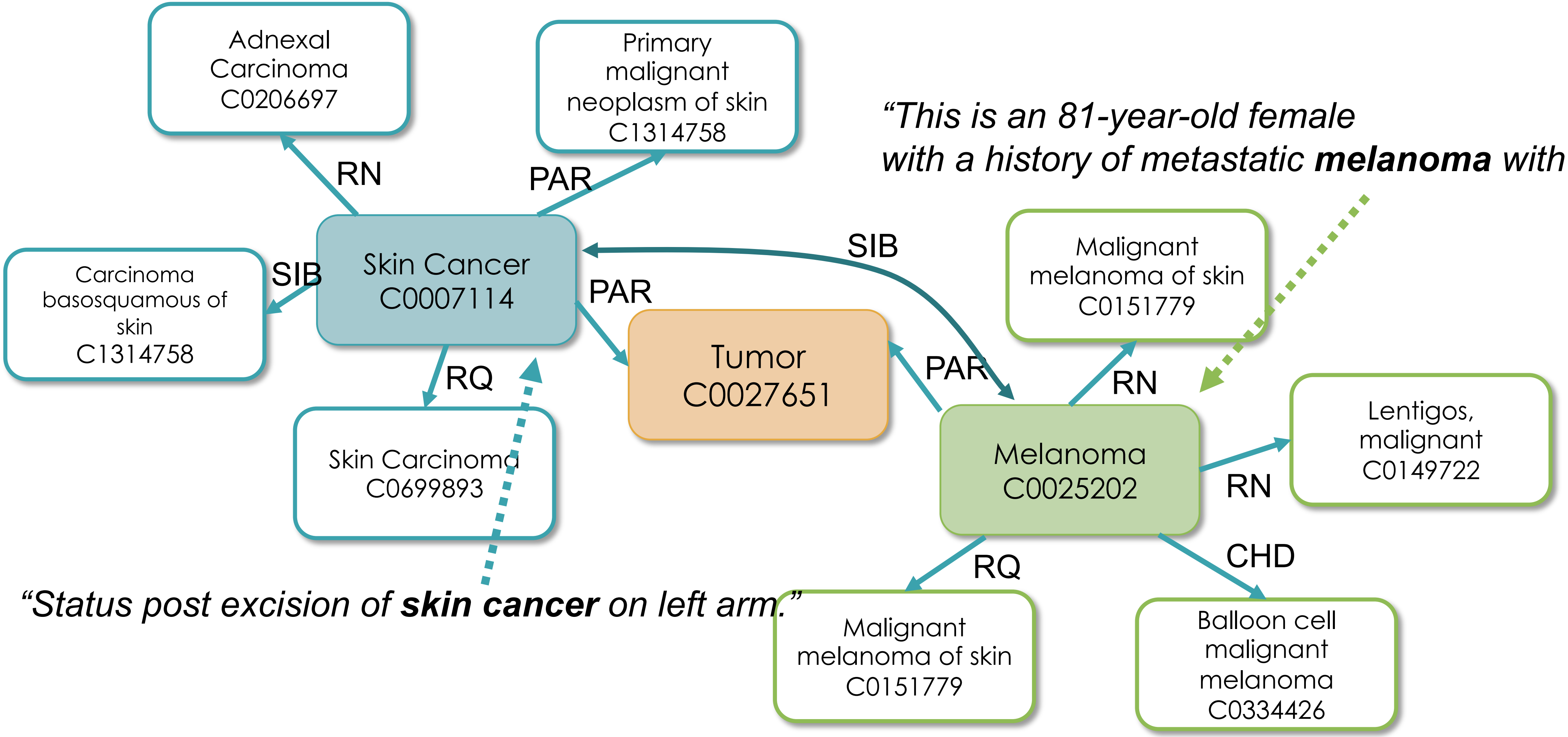
AQ
Allowed
Qualifier
609,244

CHD
Child Relationship
6,287,296

96 Data Sources

MSH - 3,082,856
ICD9CM - 219,834
ICD10PCS – 1,263,764
SNOMEDCT – 6,599,872
CCS – 57,718
CSP – 179,448
GO – 2,498,396
OMIM – 618,308
LNC – 3,717,352
MDR – 2,022,824
…

Graph-based Disambiguation of Terms


Adnexal Carcinoma C0206697
Primary malignant neoplasm of skin C1314758
RN
PAR
Carcinoma basosquamous of skin C1314758
SIB
Skin Cancer C0007114
PAR
SIB
RQ
Tumor C0027651
Skin Carcinoma C0699893
PAR
Malignant melanoma of skin C0151779
RN
Melanoma C0025202
RN
Lentigos, malignant C0149722
RQ
CHD
Malignant melanoma of skin C0151779
Balloon cell malignant melanoma C0334426

*“This is an 81-year-old female with a history of metastatic **melanoma** with…”*

*“Status post excision of **skin cancer** on left arm.”*

Oak Ridge National Laboratory

Graph Convolution with Large Knowledge Graph(s)

- Loosely-coupled: Cluster GCN
  - Wei-Lin Chiang et al., “Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks,” arXiv:1905.07953
  - Divide big one into multiple small dense graphs
  - Concatenate decisions from the multiple GCNs
- Tightly-coupled: Model-parallel GCN
  - Alok Tripathy et al., “Reducing Communication in Graph Neural Network Training,” arXiv:2005.03300
  - Divide one big adjacency matrix
  - Communication overhead

Work In-Progress

- Medical document classification using CUIs
  - Disambiguation of terms
  - Abstraction of various expressions
- GCN
  - Matrix multiplication – GPU-friendly
  - Competitive/higher task performance
- Knowledge Graph
  - Big adjacency matrix – too big to fit one GPU
  - Two approaches
    - Cluster GCN
    - Model-parallel, distributed GCN

Thank you!

- Questions?