Curating High Quality, Open Access Data to Enable Machine Learning Research

UAMS
College of Medicine

Fred Prior, PhD
Professor and Chair
Department of Biomedical Informatics
University of Arkansas for Medical Sciences

Resurgence of Machine Learning in Medicine

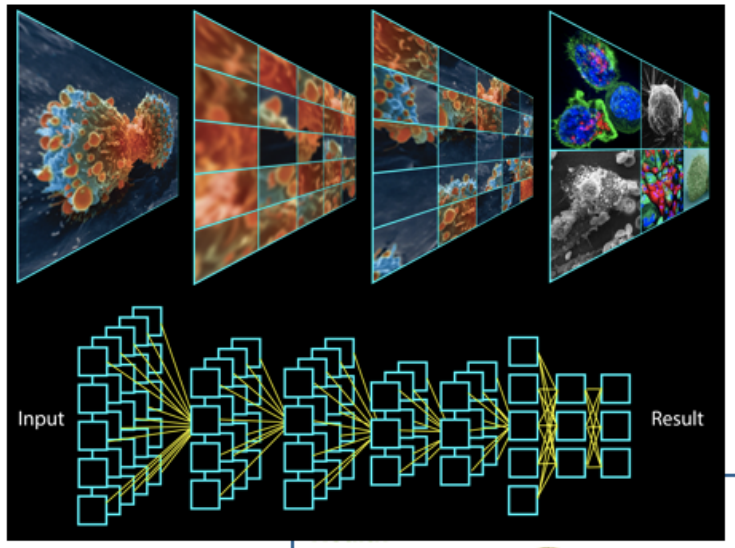
- Computers have been used to detect regions of clinical interest in images since the 1960s.
- The field drew heavily on computer vision and became known as CAD (Computer-aided Diagnosis, Computer-aided Detection)
- CAD researchers generated many of the Machine Learning tools we use today.
- In spite of years of research and development, the number of clinically successful CAD products with FDA approval has been rather limited – **Until Recently**

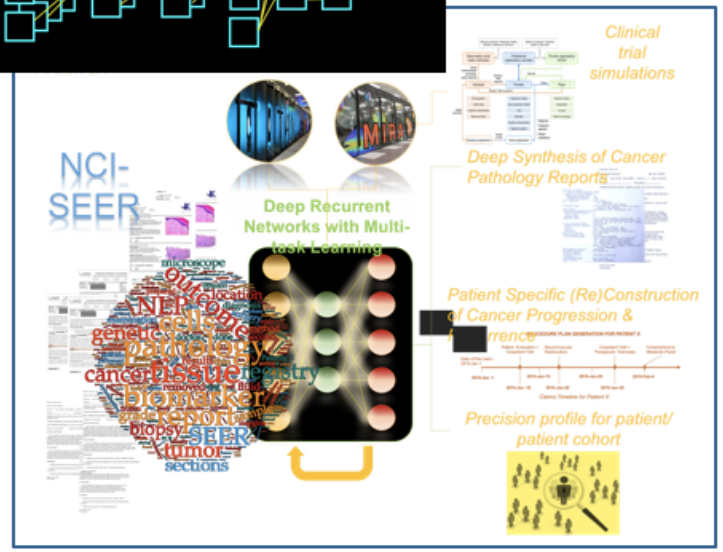
Company	FDA Approval	Indication
Apple	September 2018	Atrial fibrillation detection
Aidoc	August 2018	CT brain bleed diagnosis
iCAD	August 2018	Breast density via mammography
Zebra Medical	July 2018	Coronary calcium scoring
Bay Labs	June 2018	Echocardiogram EF determination
Neural Analytics	May 2018	Device for paramedic stroke diagnosis
IDx	April 2018	Diabetic retinopathy diagnosis
Icometrix	April 2018	MRI brain interpretation
Imagen	March 2018	X-ray wrist fracture diagnosis
Viz.ai	February 2018	CT stroke diagnosis
Arterys	February 2018	Liver and lung cancer (MRI, CT) diagnosis
MaxQ-AI	January 2018	CT brain bleed diagnosis
Alivecor	November 2017	Atrial fibrillation detection via Apple Watch
Arterys	January 2017	MRI heart interpretation

Topol EJ. Nature medicine. 2019 Jan;25(1):44.

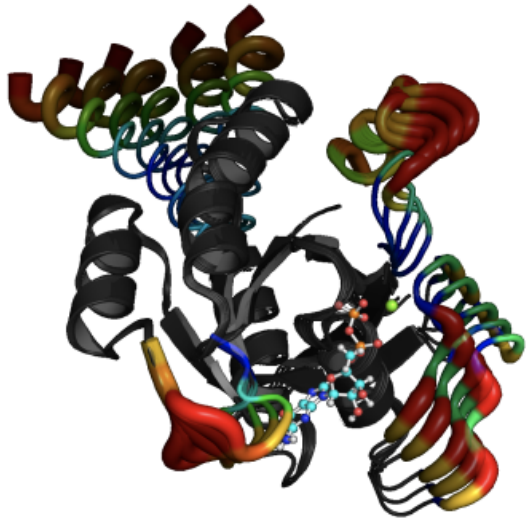
Exascale Deep Learning Enabled Precision Medicine for Cancer

CANDLE accelerates solutions toward three top cancer challenges





- Focus on building a scalable deep neural network code called the CANcer Distributed Learning Environment (CANDLE)
- CANDLE addresses three top challenges of the National Cancer Institute:
  1. Understanding the molecular basis of key protein interactions
  2. Developing predictive models for drug response, and automating the analysis
  3. Extraction of information from millions of cancer patient records to determine optimal cancer treatment strategies



PI: Rick Stevens (ANL)

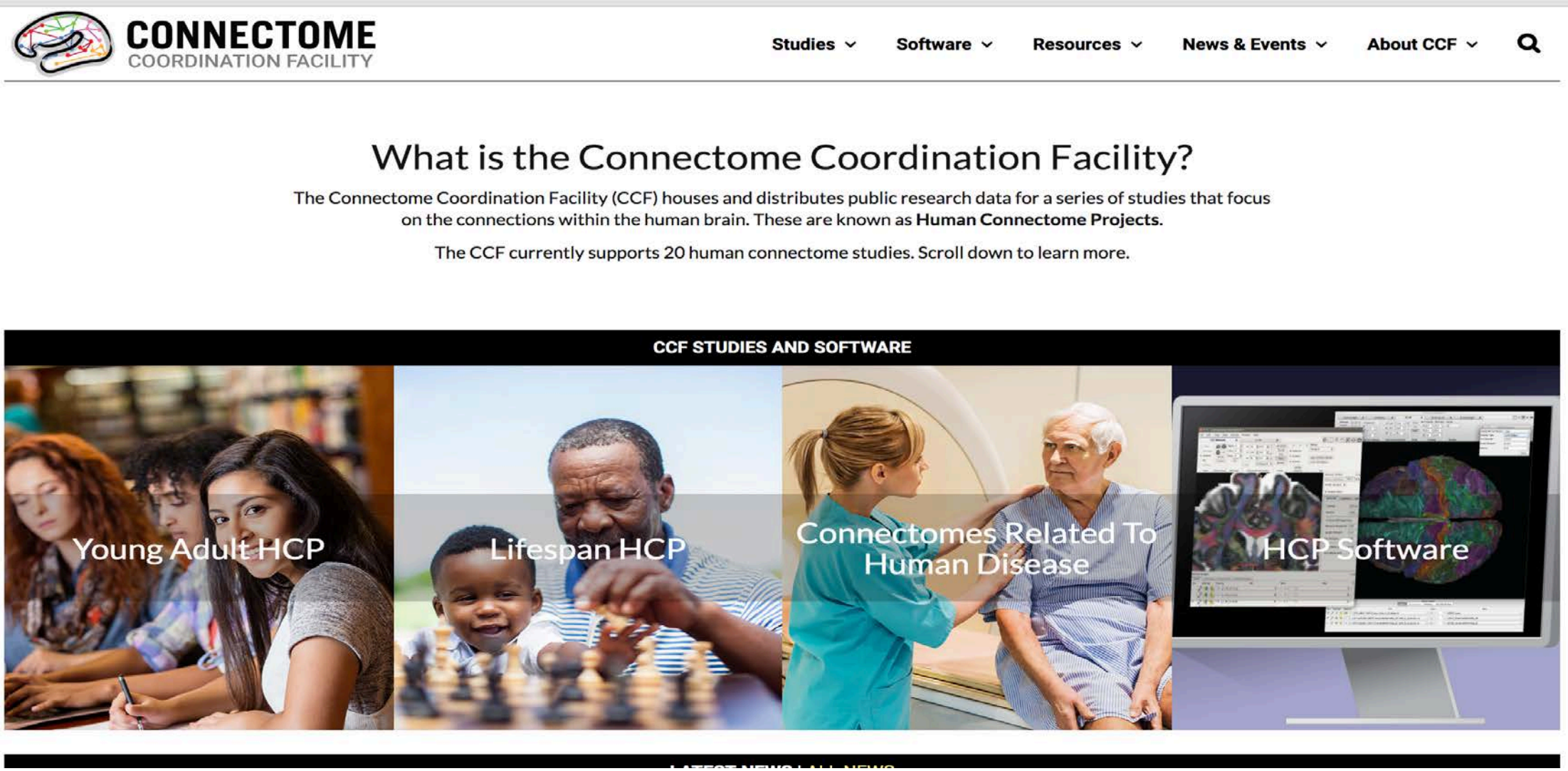
EXASCALE COMPUTING PROJECT

https://www.hpcwire.com/2017/04/26/messina-update-u-s-path-exascale-15-slides/

Machine Learning Algorithms Need Data

- For a ML algorithm to be clinically useful it must be trained on data that appropriately represents the variance in:
  - the human population
  - the presentation of disease
  - the data collection systems
- Data has to be of sufficient quality and acquired with uniform parameters to make certain that conclusions can be validated
- Supervised learning approaches require labeled data for training and validation

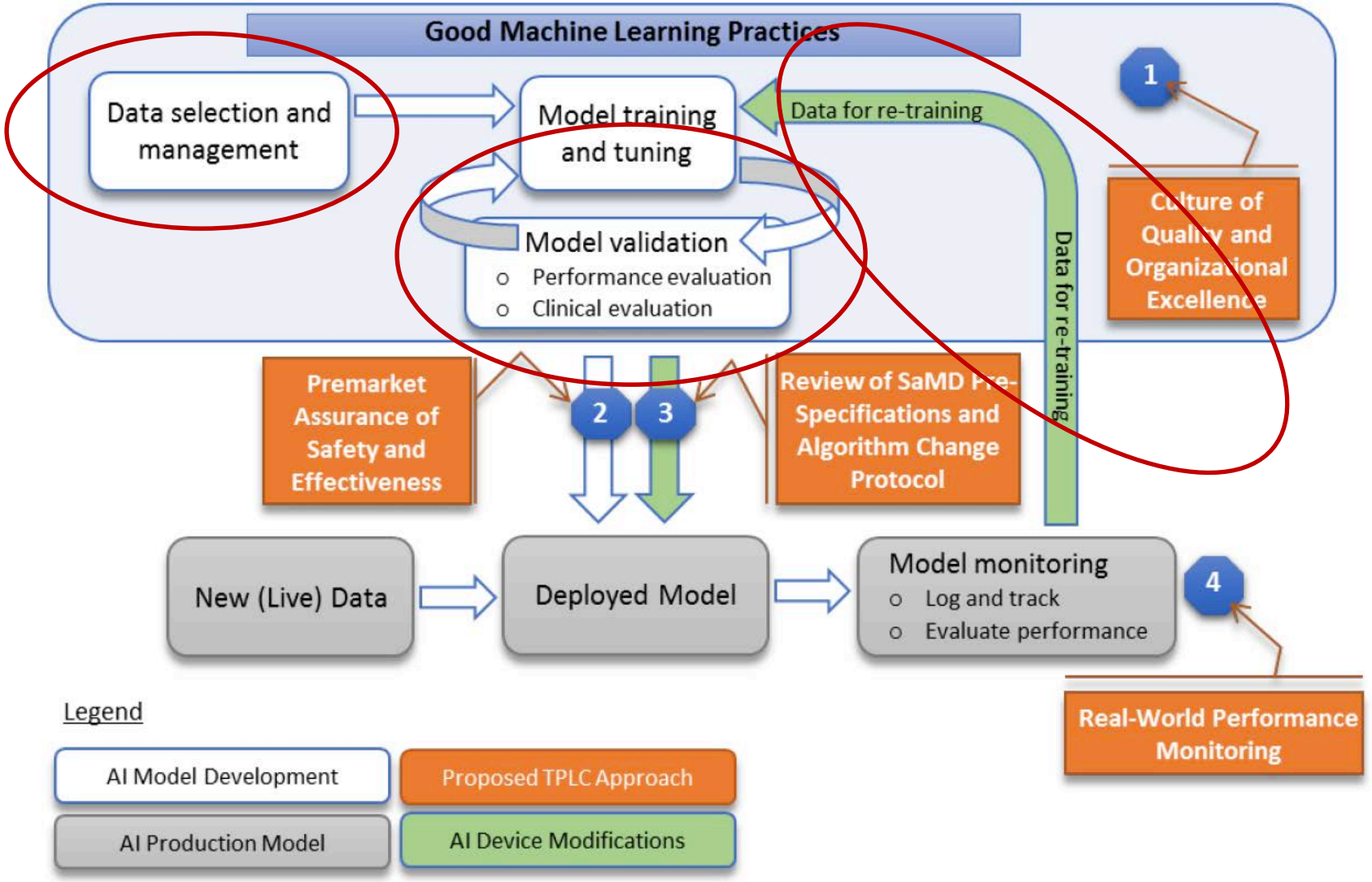
Data Limitations



Intellectual Property

- Intellectual property concerns can limit access to valuable data sets.
- Some researchers and institutions consider data to be their intellectual property and are loath to make it available
- Challenge competitions require sequestered test sets
- FDA has argued for the need for sequestration of data used to validate algorithms approved for commercial use

FDA's Total Product Lifecycle Regulatory Approach for AI/ML-Based SaMD



https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf

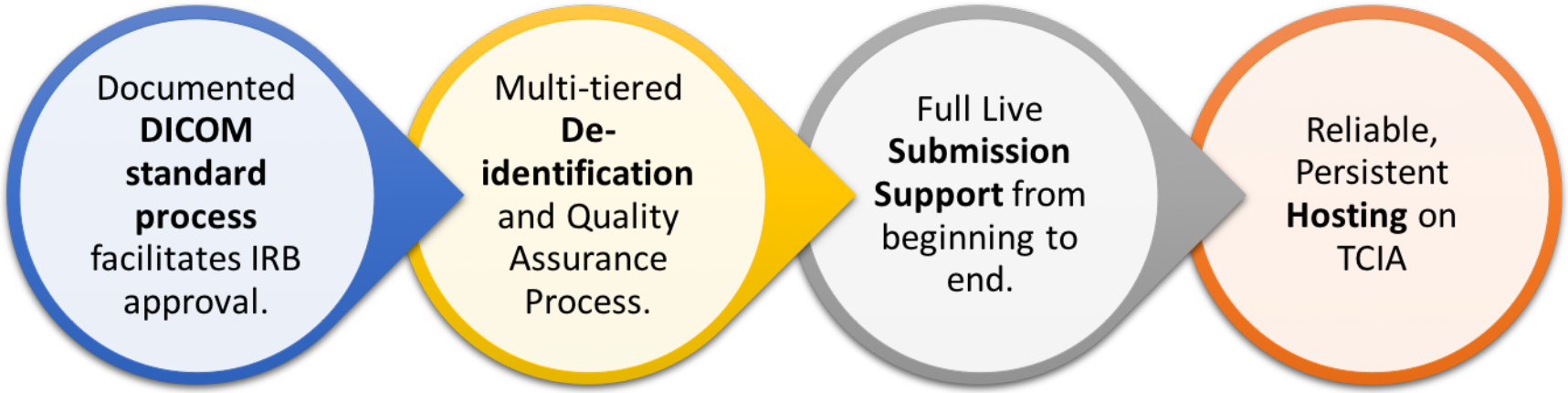
How Much Data?

- Deep learning methods require large, representative and accurately annotated datasets to train robust models and achieve acceptable performance.
- Goodfellow et al. propose the following rule of thumb:
  - "a supervised deep learning algorithm will generally achieve acceptable performance with around **5,000** labeled examples per category and will match or exceed human performance when trained with a dataset containing at least **10 million** labeled examples."
- Accumulating data on this scale poses a significant challenge and requires open access and data sharing

Goodfellow et al. Deep learning: MIT press, 2016.

The Cancer Imaging Archive: Reducing Data-Sharing Barriers

- Increase public availability of high quality cancer imaging data sets for research by complying with the FAIR principle.
- Support NIH data sharing requirements for the cancer imaging community
- Enhance reproducibility in research
- Create a culture of open data sharing and collaboration among cancer imaging researchers



Data Collection Center

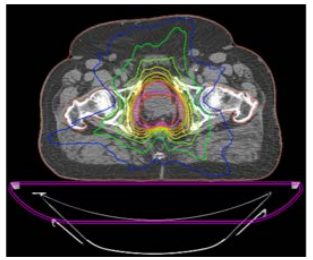
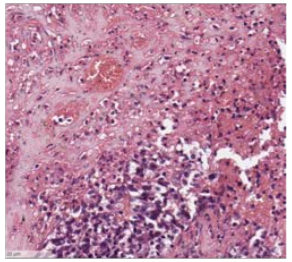
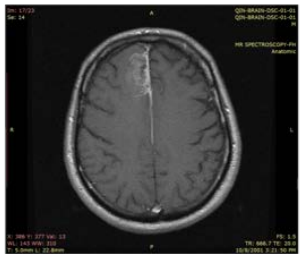
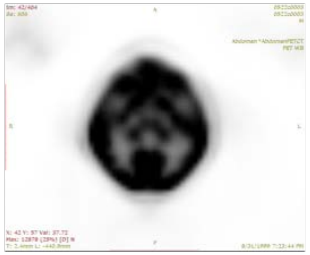
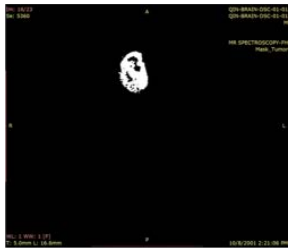
- Tools and staffing to support data collection, curation, and de-identification

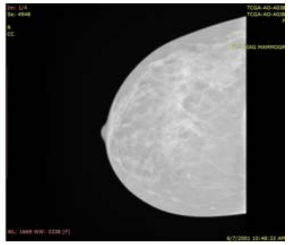
Data Access Portal

- Browse (home page)
- Filter/Search (Data Portal)
- REST API
- Analysis Data

Data Analysis Centers

- 3rd party web sites or tools which connect to TCIA's API or mirror its data

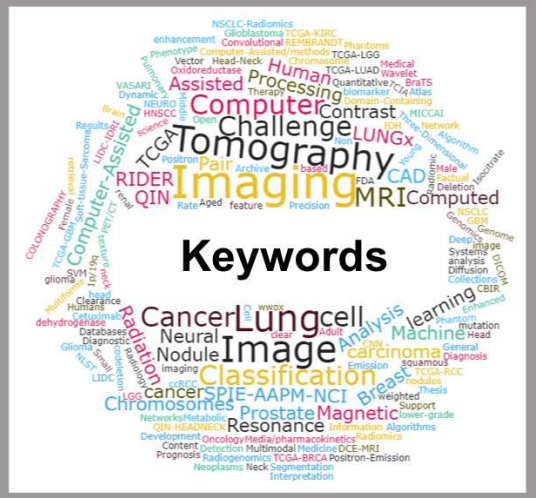







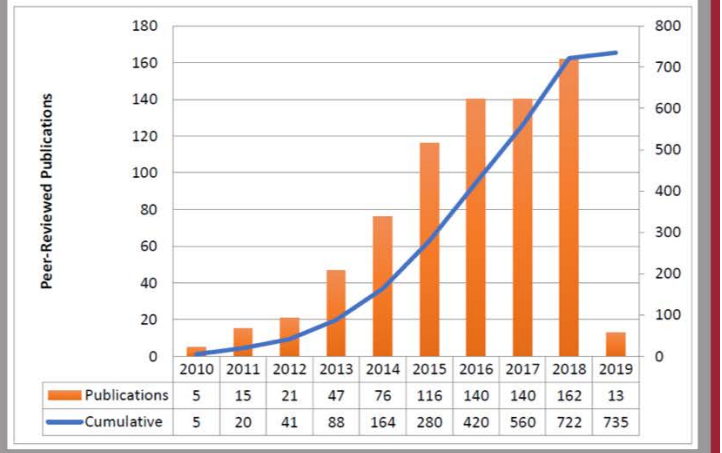
NIH NATIONAL CANCER INSTITUTE
DCTD Division of Cancer Treatment & Diagnosis

CIP Cancer Imaging Program

TCIA typically supports 15,000 active users from more than 125 countries that download ~ 75 TB of data per month. Countries March 2019 (right)



Current Listing of Peer-Reviewed Publications Based on TCIA:



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Publications	5	15	21	47	76	116	140	140	162	13
Cumulative	5	20	41	88	164	280	420	560	722	735

Over 700 publications have referenced TCIA and used its data. A help desk provides email and phone support for both data submitters and researchers who download and use TCIA data. As of March 31, 2019 (above right)

Clark, et al. J Digital Imag 26.6 (2013): 1045-1057.

TCIA Collections

Current Collections

The Cancer Imaging Archive

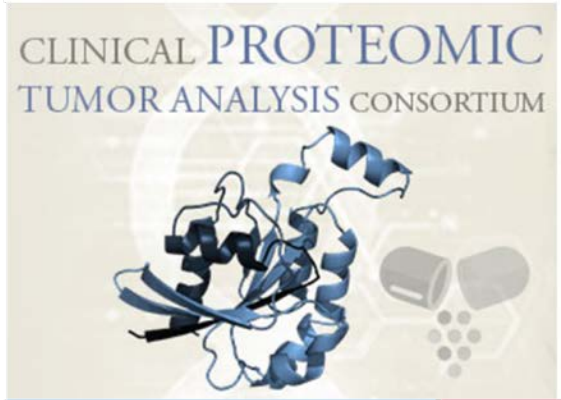
~41,500 subjects

~34 million images

Radiology images
Pathology images
RT data
Clinical data
Image derived features

~100 TB

Collections Planned or in Process



Quantitative Imaging Network

NCI National Clinical Trials Network
a National Cancer Institute program

PDMR NCI Patient-Derived Models Repository
An NCI Precision Oncology Initiative℠ Resource

ECOG-ACRIN cancer research group
Reshaping the future of patient care

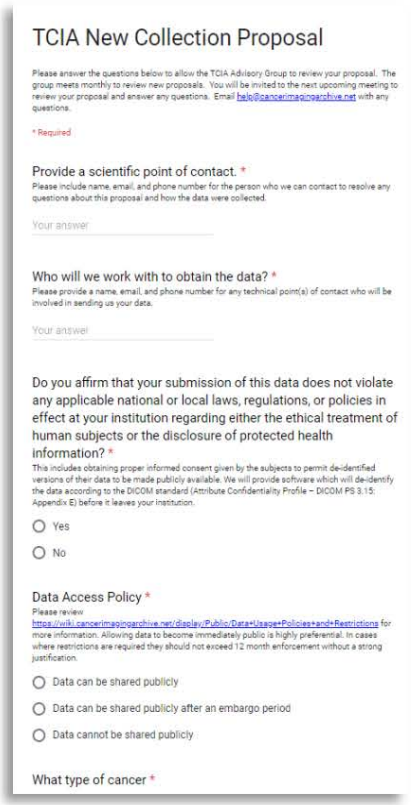
IROC IMAGING AND RADIATION ONCOLOGY CORE
Global Leaders in Clinical Trial Quality Assurance

Community Initiated

~500 TB

Data Governance: Submission Process

Submit Proposal



TCIA New Collection Proposal

Please answer the questions below to allow the TCIA Advisory Group to review your proposal. The group meets monthly to review new proposals. You will be invited to the next upcoming meeting to review your proposal and answer any questions. Email help@cancerimagingarchive.net with any questions.

* Required

Provide a scientific point of contact. *

Please include name, email, and phone number for the person who we can contact to resolve any questions about this proposal and how the data were collected.

Your answer

Who will we work with to obtain the data? *

Please provide a name, email, and phone number for any technical point(s) of contact who will be involved in sending us your data.

Your answer

Do you affirm that your submission of this data does not violate any applicable national or local laws, regulations, or policies in effect at your institution regarding either the ethical treatment of human subjects or the disclosure of protected health information? *

This includes obtaining proper informed consent given by the subjects to permit de-identified versions of their data to be made publicly available. We will provide software which will de-identify the data according to the DICOM standard (Attribute Confidentiality Profile – DICOM PS 3.15: Appendix E) before it leaves your institution.

- Yes
- No

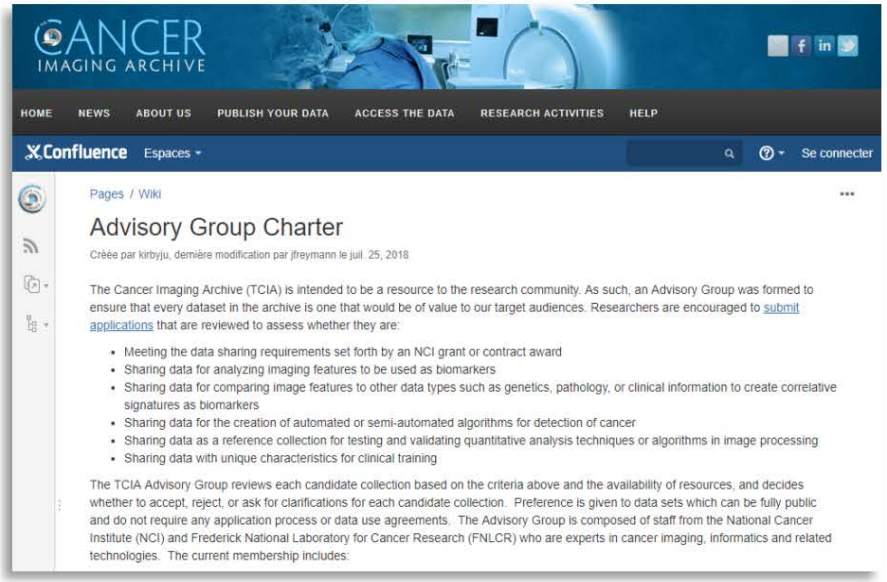
Data Access Policy *

Please review https://wiki.cancerimagingarchive.net/display/Public/Data+Usage+Policies+and+Restrictions for more information. Allowing data to become immediately public is highly preferential. In cases where restrictions are required they should not exceed 12 month enforcement without a strong justification.

- Data can be shared publicly
- Data can be shared publicly after an embargo period
- Data cannot be shared publicly

What type of cancer *

Review at Monthly TCIA Advisory



CANCER IMAGING ARCHIVE

HOME NEWS ABOUT US PUBLISH YOUR DATA ACCESS THE DATA RESEARCH ACTIVITIES HELP

Confluence Espaces Se connecter

Pages / Wiki

Advisory Group Charter

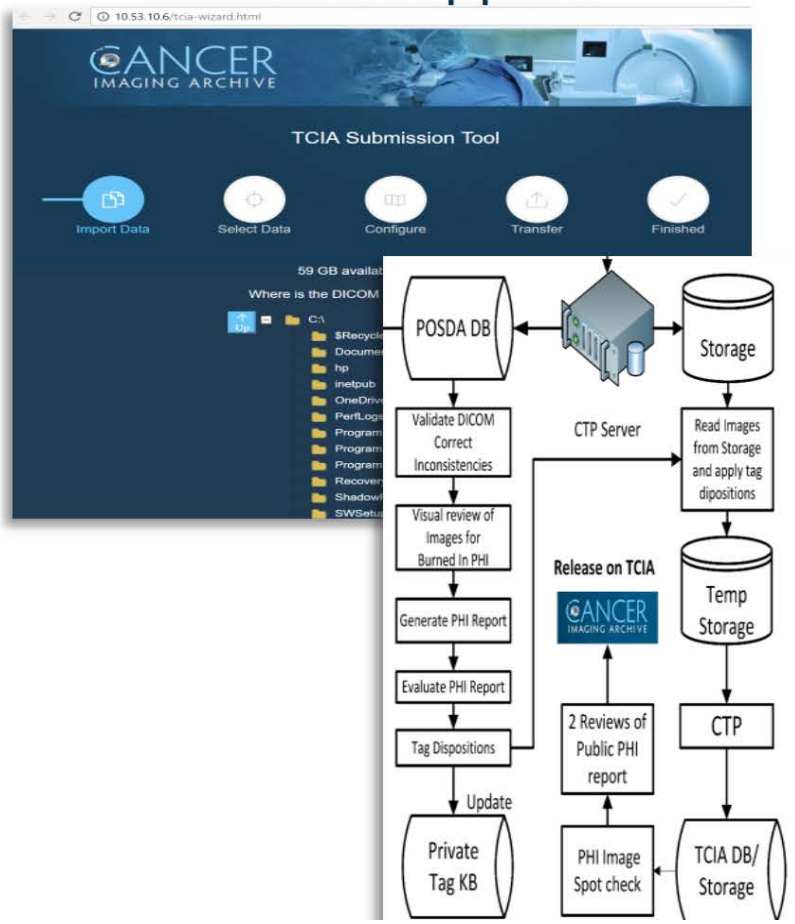
Créée par kirbyju, dernière modification par jfreymann le juil. 25, 2018

The Cancer Imaging Archive (TCIA) is intended to be a resource to the research community. As such, an Advisory Group was formed to ensure that every dataset in the archive is one that would be of value to our target audiences. Researchers are encouraged to submit applications that are reviewed to assess whether they are:

- Meeting the data sharing requirements set forth by an NCI grant or contract award
- Sharing data for analyzing imaging features to be used as biomarkers
- Sharing data for comparing image features to other data types such as genetics, pathology, or clinical information to create correlative signatures as biomarkers
- Sharing data for the creation of automated or semi-automated algorithms for detection of cancer
- Sharing data as a reference collection for testing and validating quantitative analysis techniques or algorithms in image processing
- Sharing data with unique characteristics for clinical training

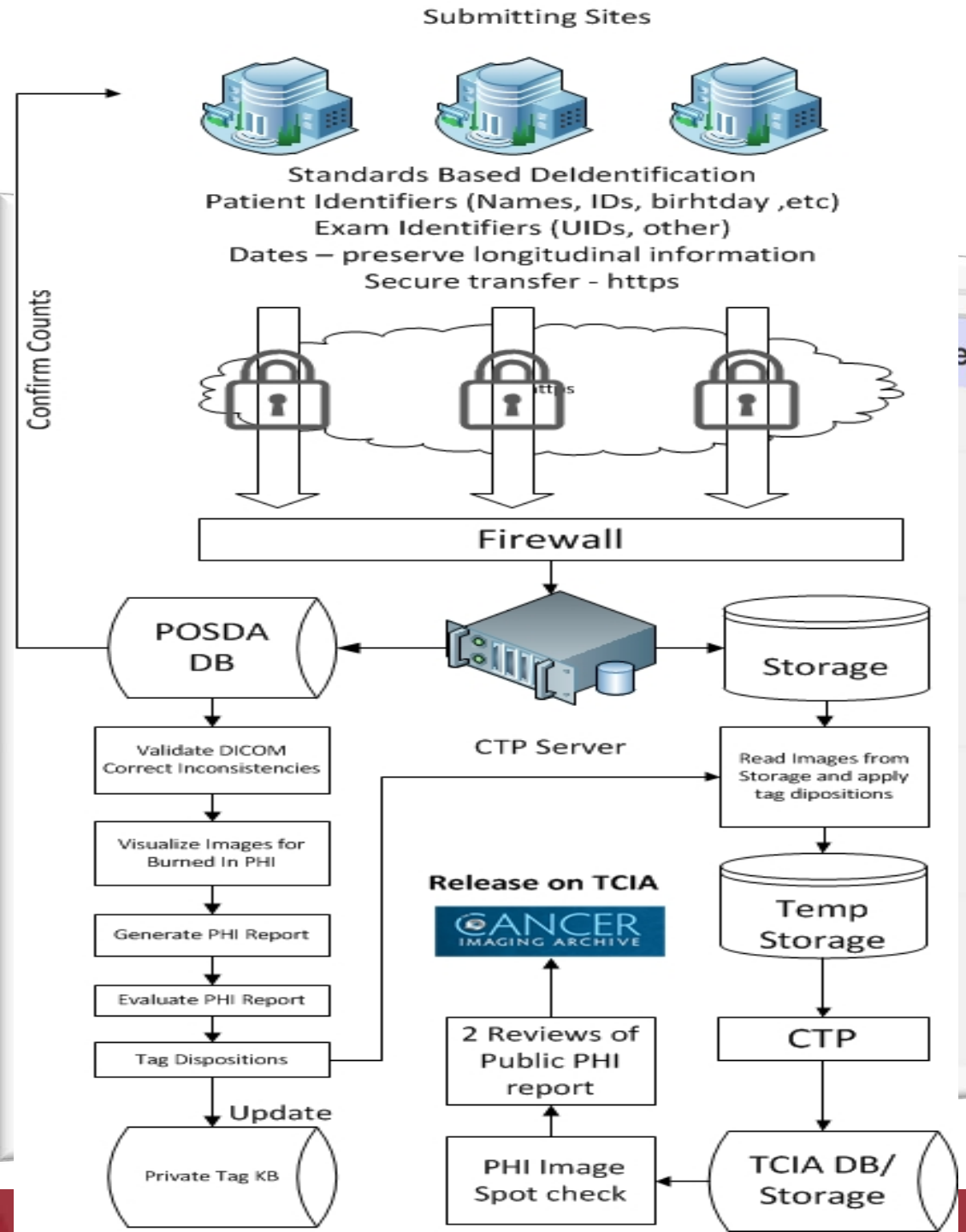
The TCIA Advisory Group reviews each candidate collection based on the criteria above and the availability of resources, and decides whether to accept, reject, or ask for clarifications for each candidate collection. Preference is given to data sets which can be fully public and do not require any application process or data use agreements. The Advisory Group is composed of staff from the National Cancer Institute (NCI) and Frederick National Laboratory for Cancer Research (FNLCR) who are experts in cancer imaging, informatics and related technologies. The current membership includes:

TCIA Submission and Curation Support

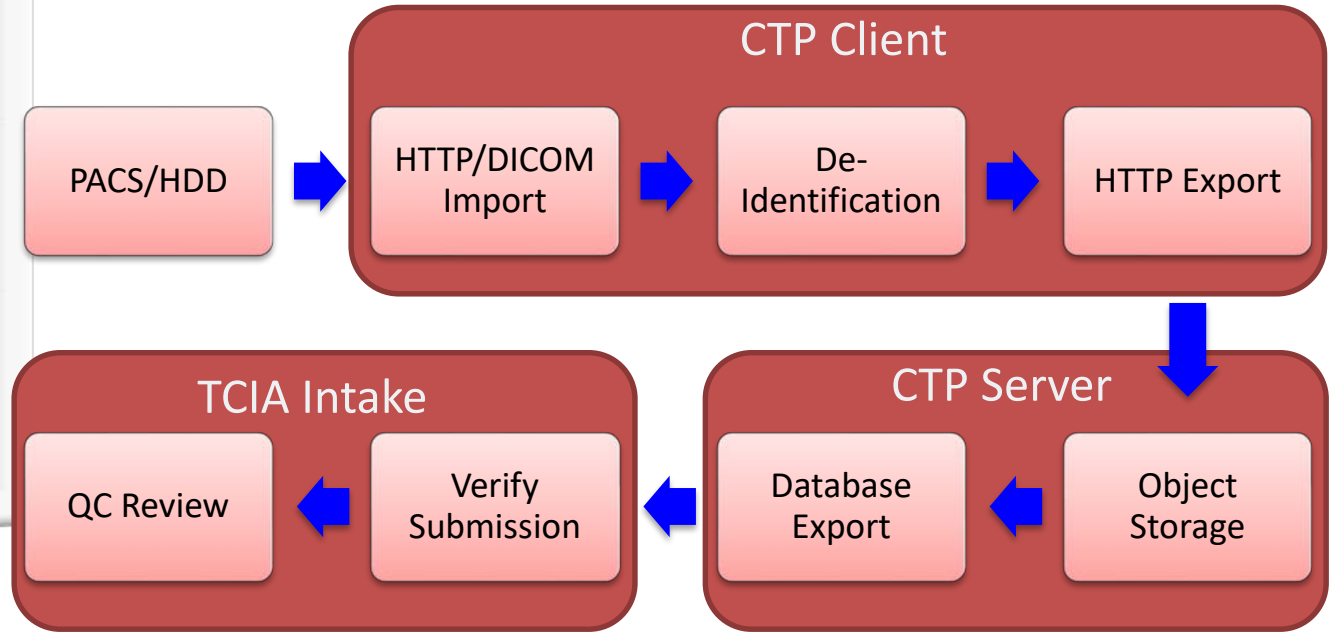


Curation is the Key to Data Quality

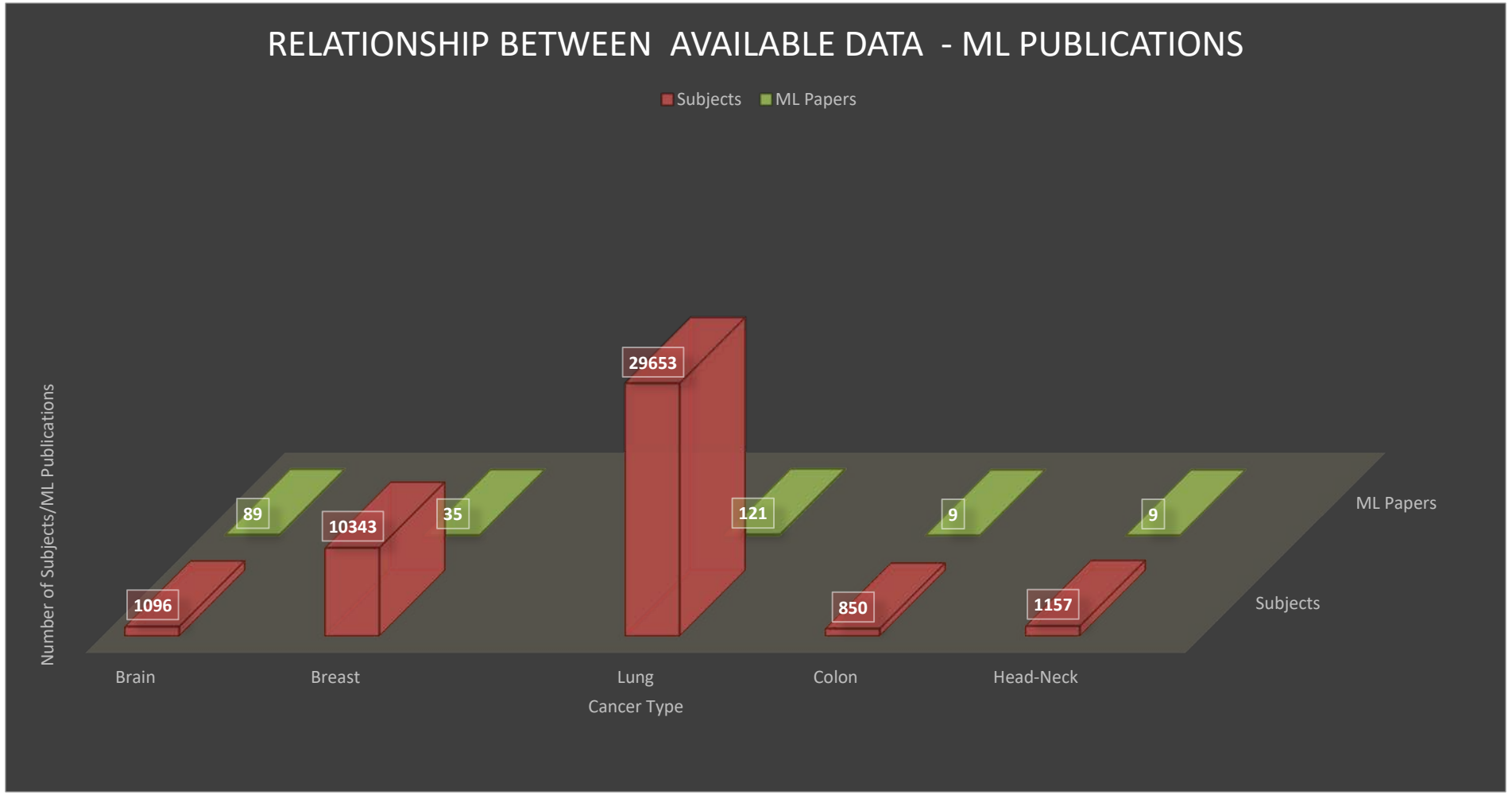
Bennett W, Matthews J, Bosch W. SU-GG-T-262: Open-Source Tool for Assessing Variability in DICOM Data. Medical Physics. 2010;37(6):3245-



Automated Standards-based Anonymization Profile for Image Sharing Using RSNA's Clinical Trial Processor . Justin Kirby, John Perry, Carl Jaffe, John Freymann



We Still Do NOT Have Enough Data



The Truth Problem

- Supervised learning techniques, commonly used in radiomic/pathomic studies, are hampered by the lack of labeled data for training and testing
- Labeled data is created manually by human experts resulting in high cost and limited volume of high-quality training (and testing) datasets
- Problem of ground truth - Approaches to modeling truth by combining observations from multiple observers (machine and human) attempt to create standards that do not penalize algorithms that outperform the human observer

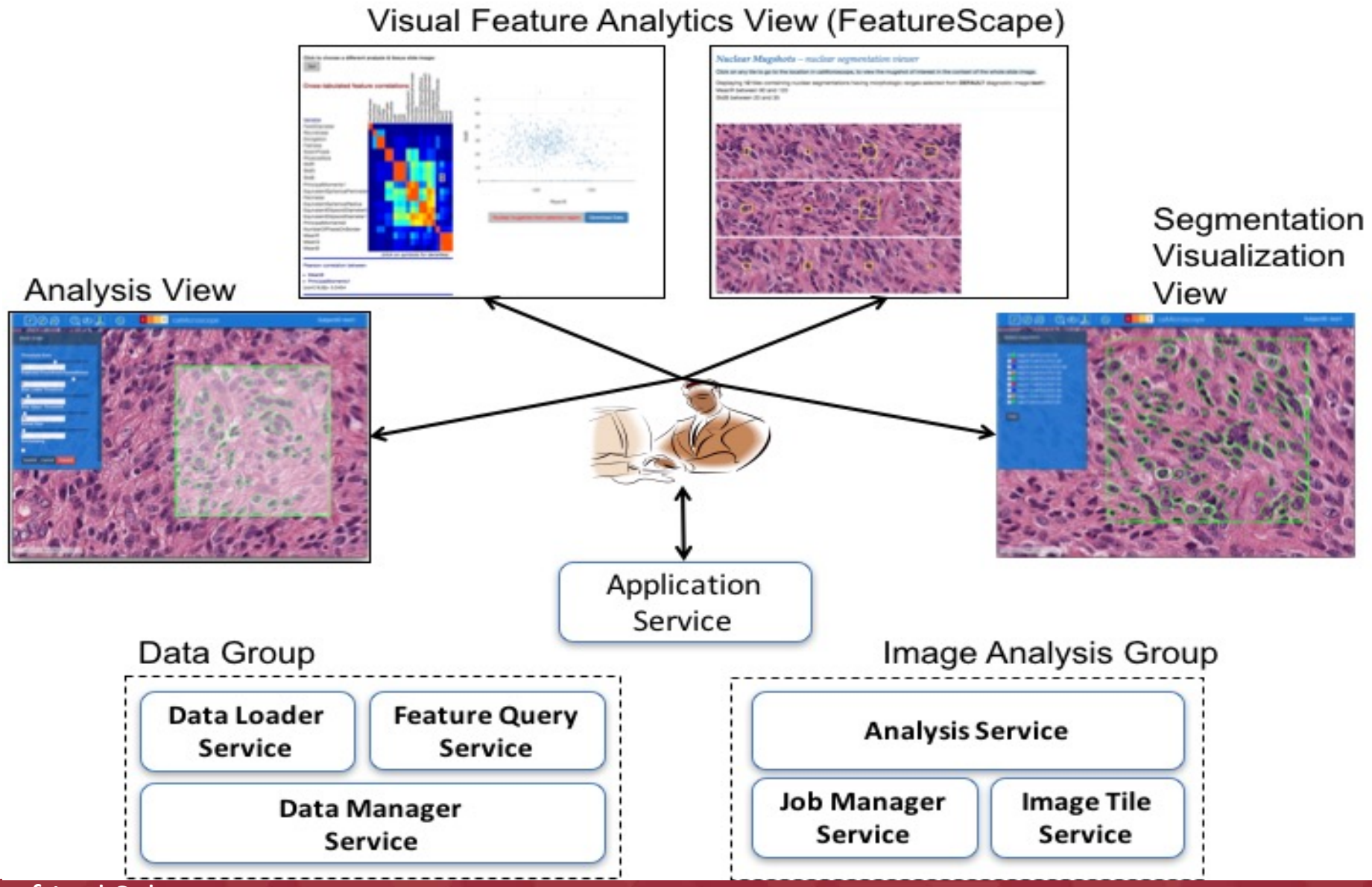
Capturing Labeled Data

- The generation of labeled data remains a roadblock.
  - crowdsourcing and the use of augmentation and synthetic data generation
- Management of labeled training and test sets and the resulting image derived features raise special curation and validation issues.
- Currently, standards and standard operating processes for data representation, curation, evaluation and sharing of labeled datasets are in early stages of development
- Reproducibility of radiomics/pathomics analyses are difficult to assess due to a lack of standards for validating results.

Unique Challenges in Developing High-quality Training Datasets in Digital Pathology

- Tissue images capture much denser information than many other imaging modalities 100,000 x 100,000 pixels
- Whole slide tissue image can contain more than a million cells, nuclei and other cellular-level structures making it impossible to manually segment and annotate
- Heterogeneity across and within tissue specimens and variations in tissue preparation lead to high variability in quantitative analyses
- Consistent and unified protocols are necessary to reduce artifacts in image datasets used to generate training data.
- While there is a DICOM standard image file format for digital pathology it is not universally accepted, despite on-going efforts.

Feature Management System
Quantitative Imaging Pathology - QuIP
(Saltz et al. ITCR U24)

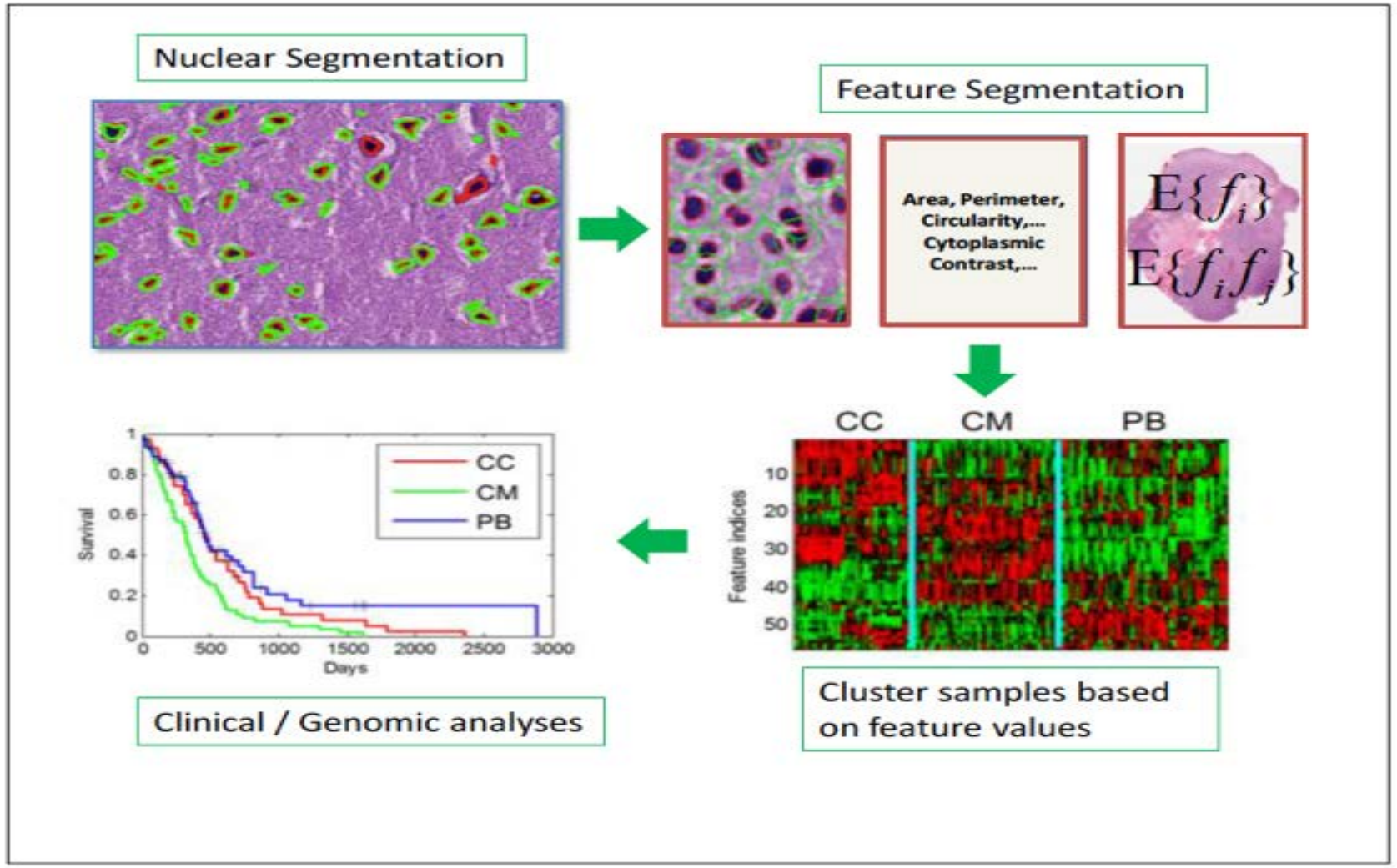


Slide Courtesy of Joel Saltz

Integrative Morphology/"omics"

J Am Med Inform Assoc. 2012
Integrated morphologic analysis for the identification and characterization of disease subtypes.

Lee Cooper, Jun Kong



Slide Courtesy of Joel Saltz

Rad/Path Image Feature Base

- Manage all feature classes
- Semantic integration of feature information
- General format for feature representation: 4D Tensor
- Linkage of features to images from which they were generated, feature extraction and classification algorithms, quality metrics

Summary

- Using TCIA as an example and drawing on our own research I have attempted to address one of the objectives of the conference:
  - What and how should the curated data sets be developed to optimize specific tasks of today and tomorrow?
- Data has to be of sufficient quality to make certain that conclusions can be validated
- Large quantities of labeled data are required as are methods to generate, curate and manage such data.
- Open access data repositories hold the key to providing sufficient data

Acknowledgements

University of Arkansas

Fred Prior
Lawrence Tarbox
Mathias Brochhausen
Yasir Rahmatallah
Xiuzhen Huang
Jonathan Bona
Kirk Smith
Bill Bennett
Tracy Nolan
Dwayne Dobbins
Jeremy Jarosz
Jeff Tobler
Joseph Utecht
Julie Frund
Sonya Utecht
Betty Levine
Diana Stockton
Erica Bilello
Geri Blake
Robert Brown
Pam Angelus
Brittney Camp
Claren Freeman
Natasha Honomichl

Frederick National Laboratory for Cancer Research

John Freymann
Justin Kirby
Brenda Fevrier-Sullivan
Carl Jaffe
Luis Cordeiro
Craig Hill

Emory University

Ashish Sharma
Ryan Birmingham

Funding

1. **NCI 1U01CA187013-01** *Resources for development and validation of Radiomic analyses & Adaptive Therapy* (Prior, Sharma)
2. **Leidos Biomedical Research, Contract 16X011** for NCI, *Maintenance and Extension of The Cancer Imaging Archive (TCIA*) (Prior)
3. **NCI 1U24CA215109**, *TCIA Sustainment and Scalability - Platforms for Quantitative Imaging Informatics in Precision Medicine* (Prior, Sharma, Saltz)
4. **NCI 3U24CA215109-02S1,** *Informatics Platform for Cancer-Related Cognitive Impairment and Dementia Research (Prior)*

Stony Brook University

Joel Saltz
Tahsin Kurc
Erich Bremer
Feiqiao Wang
Joseph Balsamo

Washington University St. Louis

Malcolm Tobias
Walter Bosch

QARC/University of Massachusetts

TJ Fitzgerald
Richard Hanusik

NBIA Team (NCI/FNLCR/Ellumen)

Ulli Wagner
Scott Gustafson
Qinyan Pan
Russ Rieling
Carolyn Klinger
Martin Lerner
Tin Tran