



Data management tools and strategy for responsible imaging AI

Beebe Symposium
March 13, 2025

Daniel Marcus, PhD
Professor, Radiology, Washington University
Director, Computational Imaging Research Center
CEO, Embark Labs



Disclosures

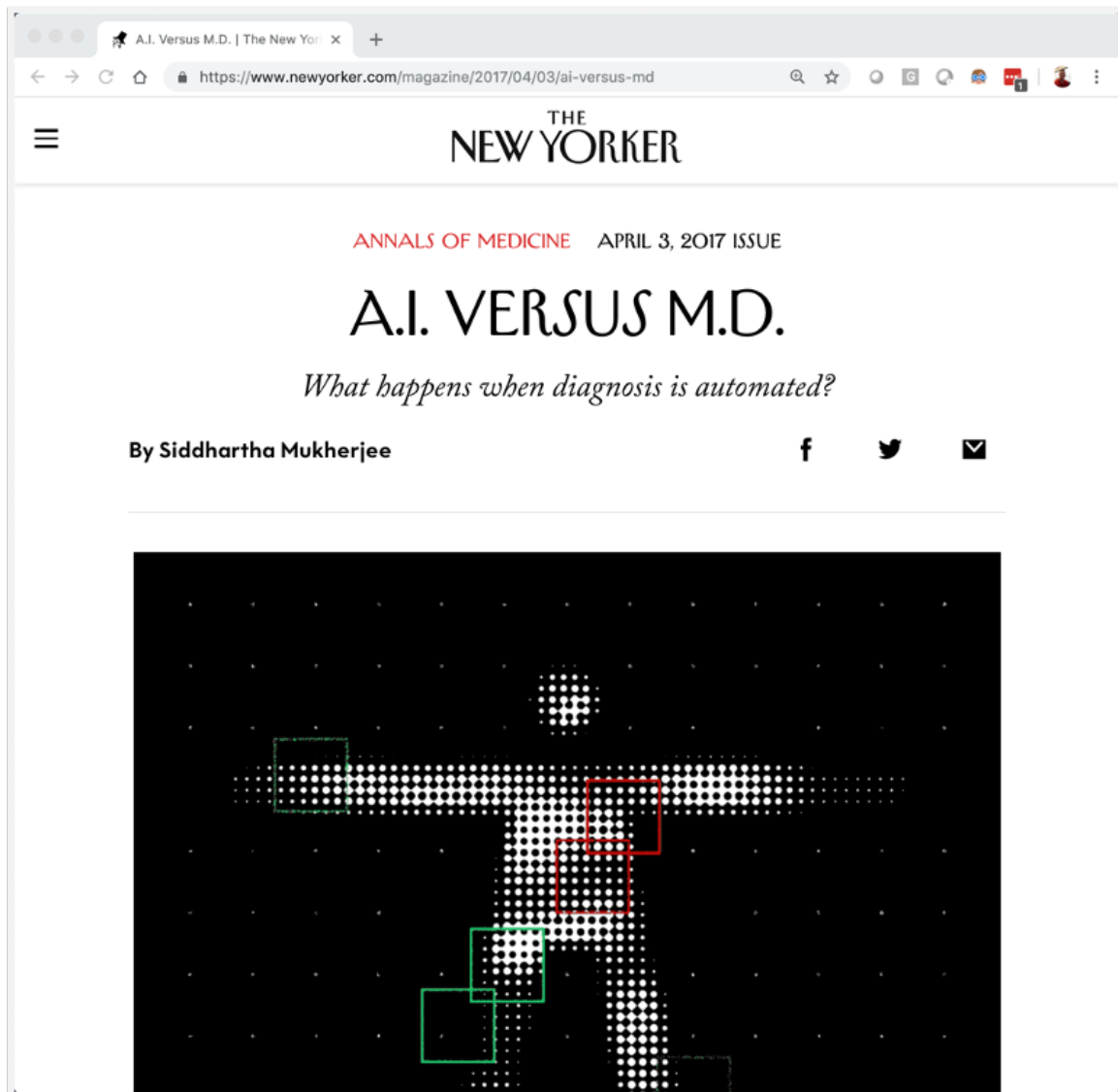
This work was supported in part by funding from the National Institutes of Health:

R01EB009352, U24CA258483, U24CA253531

I have an ownership interest and may financially benefit if the company is successful in marketing its products that are related to this research:

White Rabbit, Embark Labs, and Sora
Neuroscience

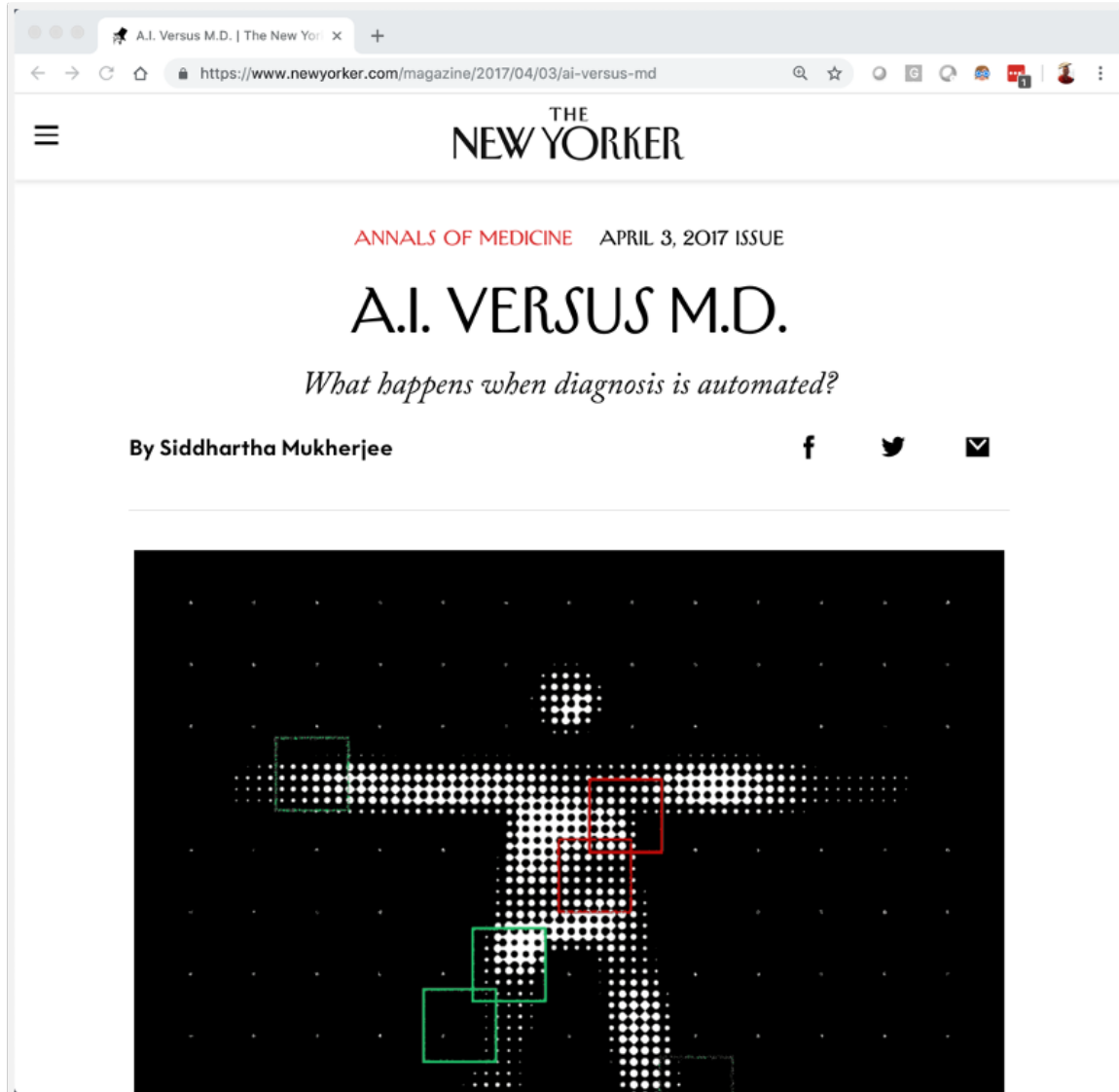
Is imaging AI ready for clinical use?



“I think that if you work as a radiologist you are like Wile E. Coyote in the cartoon. You’re already over the edge of the cliff, but you haven’t yet looked down. There’s no ground underneath. It’s just completely obvious that **in five years** deep learning is going to do better than radiologists. It might be ten years. They should **stop training radiologists now.**”

— Geoffrey Hinton, 2017

Is imaging AI ready for clinical use?



“Deep-learning platform detects malignant nodules on x-ray, **beating radiologists**”

“Chinese AI **Beats Doctors** in Diagnosing Brain Tumors”

“Deep learning **beats radiologists** at fibrotic lung disease classification”

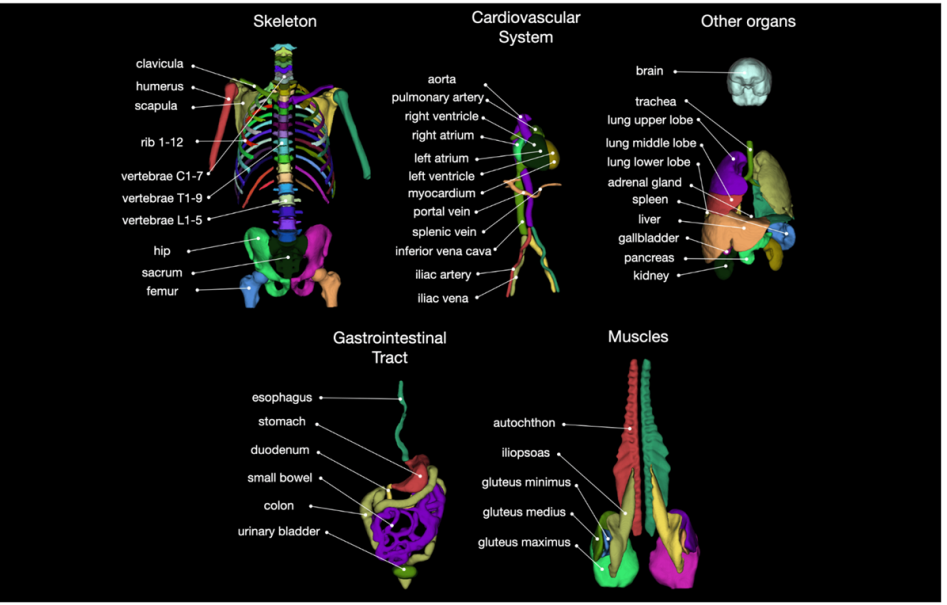
“AI system **beats team of 15 doctors** in competition”

It is getting pretty good

Segment Anything | Meta AI x GitHub - wasserth/TotalSegmentator x TotalSegmentator/overview... | github.com/wasserth/TotalSegmentator

TotalSegmentator

Tool for segmentation of 104 classes in CT images. It was trained on a wide range of different CT images (different scanners, institutions, protocols,...) and therefore should work well on most images. The training dataset with 1204 subjects can be downloaded from [Zenodo](#). You can also try the tool online at totalsegmentator.com.



Skeleton

- clavicula
- humerus
- scapula
- rib 1-12
- vertebrae C1-7
- vertebrae T1-9
- vertebrae L1-5
- hip
- sacrum
- femur

Cardiovascular System

- aorta
- pulmonary artery
- right ventricle
- right atrium
- left atrium
- left ventricle
- myocardium
- portal vein
- splenic vein
- inferior vena cava
- iliac artery
- iliac vena

Other organs

- brain
- trachea
- lung upper lobe
- lung middle lobe
- lung lower lobe
- adrenal gland
- spleen
- liver
- gallbladder
- pancreas
- kidney

Gastrointestinal Tract

- esophagus
- stomach
- duodenum
- small bowel
- colon
- urinary bladder

Muscles

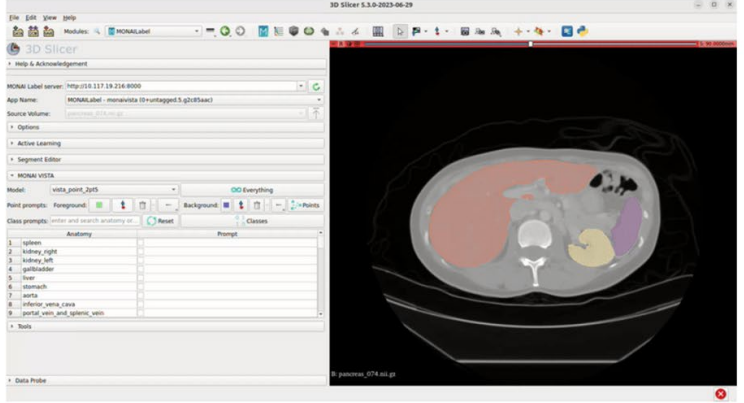
- autochthon
- iliopsoas
- gluteus minimus
- gluteus medius
- gluteus maximus

Created by the department of [Research and Analysis at University Hospital Basel](#).
If you use it please cite our paper: <https://arxiv.org/abs/2208.05868>. Please also cite [nnU-Net](#) since

GitHub - Project-MONAI/VISTA | github.com/Project-MONAI/VISTA

MONAI VISTA

MONAI Versatile Imaging Segmentation and Annotation



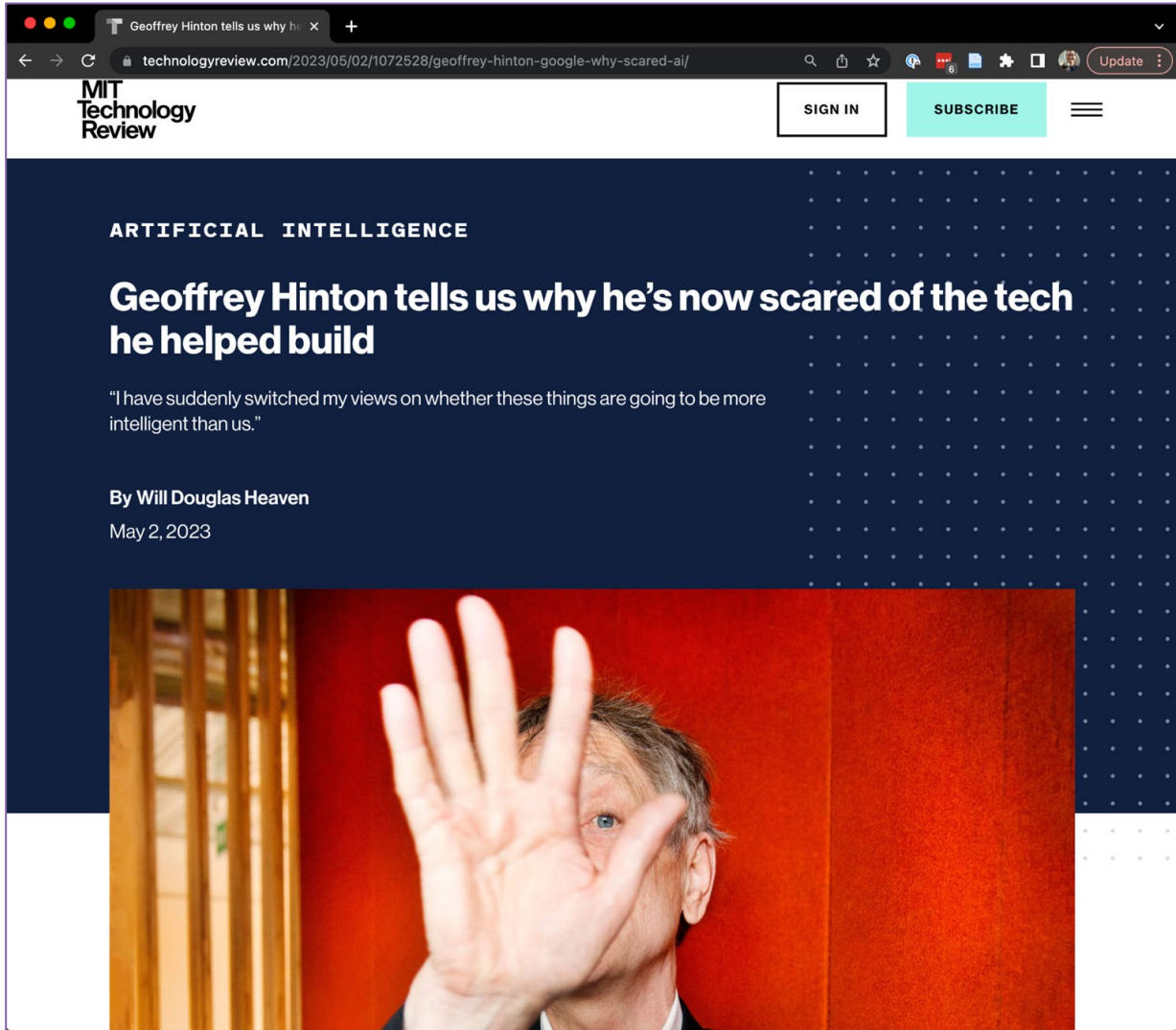
(We're seeking collaborators. If your institution is interested, please fill out the survey: <https://forms.office.com/r/RedPQc9fmw>)

Table of Contents

- [Overview](#)
- [MONAI VISTA Training and FineTuning](#)
- [MONAI VISTA with MONAI Label](#)
 - [Step 1. Installation](#)
 - [Step 2. MONAI Label monaivista app](#)
 - [Step 3. MONAI VISTA - Label Plugins](#)
 - [Step 4. Data Preparation](#)

https://github.com/Project-MONAI/VISTA/blob/main/assets/imgs/demo_gif.gif

Any thoughts Geoff?



“I have suddenly switched my views on whether these things are going to be more intelligent than us. I think they’re very close to it now and they will be much more intelligent than us in the future. How do we survive that?”

— Geoffrey Hinton, May, 2023, upon resigning as head of AI at Google

Even though imaging AI is good... we've got issues

The screenshot shows the NICE website interface. The main heading is "Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment". Below the heading, it says "Health technology evaluation | HTE12 | Published: 28 September 2023". There are navigation tabs for "Guidance", "Tools and resources", "Information for the public", and "History". A "Download Guidance (PDF)" button is visible. The main content area is titled "1 Recommendations" and contains the following text: "1.1 More research is needed on the following artificial intelligence (AI)-derived software to analyse chest X-rays alongside clinician review for suspected lung cancer in adults referred from primary care:" followed by a bulleted list of software products: "AI-Rad Companion Chest X-ray (Siemens Healthineers)", "Annalise CXR (Annalise ai)", and "Auto Lung Nodule Detection (Samsung)".

“More research is needed on the AI-derived software to analyse chest X-rays alongside clinician review.”

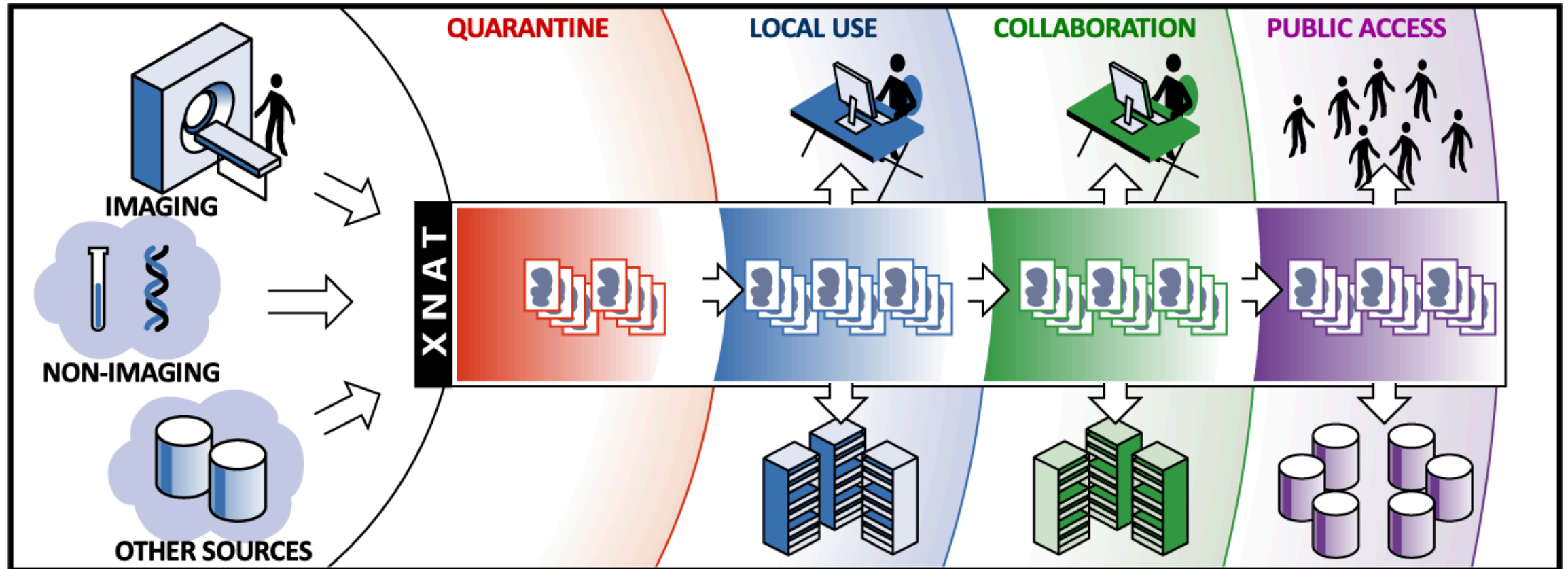
“With the available evidence, it is not possible to assess the clinical and cost benefits or risks of using the technology in the NHS. So, AI-derived software should not be used for clinical decision making in the NHS until more evidence is available.”

“There is no evidence to show how accurate software-assisted clinician review will be at identifying lung abnormalities... Using this software could lead to lung cancer being missed.”

Even though imaging AI is good... we've got issues

- Insufficient evidence of effectiveness and reliability.
- Insidious bias in algorithms.
- Unknown performance in local context (i.e. generalizability).
- Questionable impact on radiologist behavior and workflows.
- How can imaging informatics help address these issues?
 - Plumbing
 - Data

XNAT



XNAT is...

...a feature rich...

Archive, manage, process, view, and share imaging and related data.

...open...

Open source

Open API

Free (though commercial support is available)

Used by organizations around the world

...platform.

Clinical/translational research

Institutional repositories

Multi-center studies

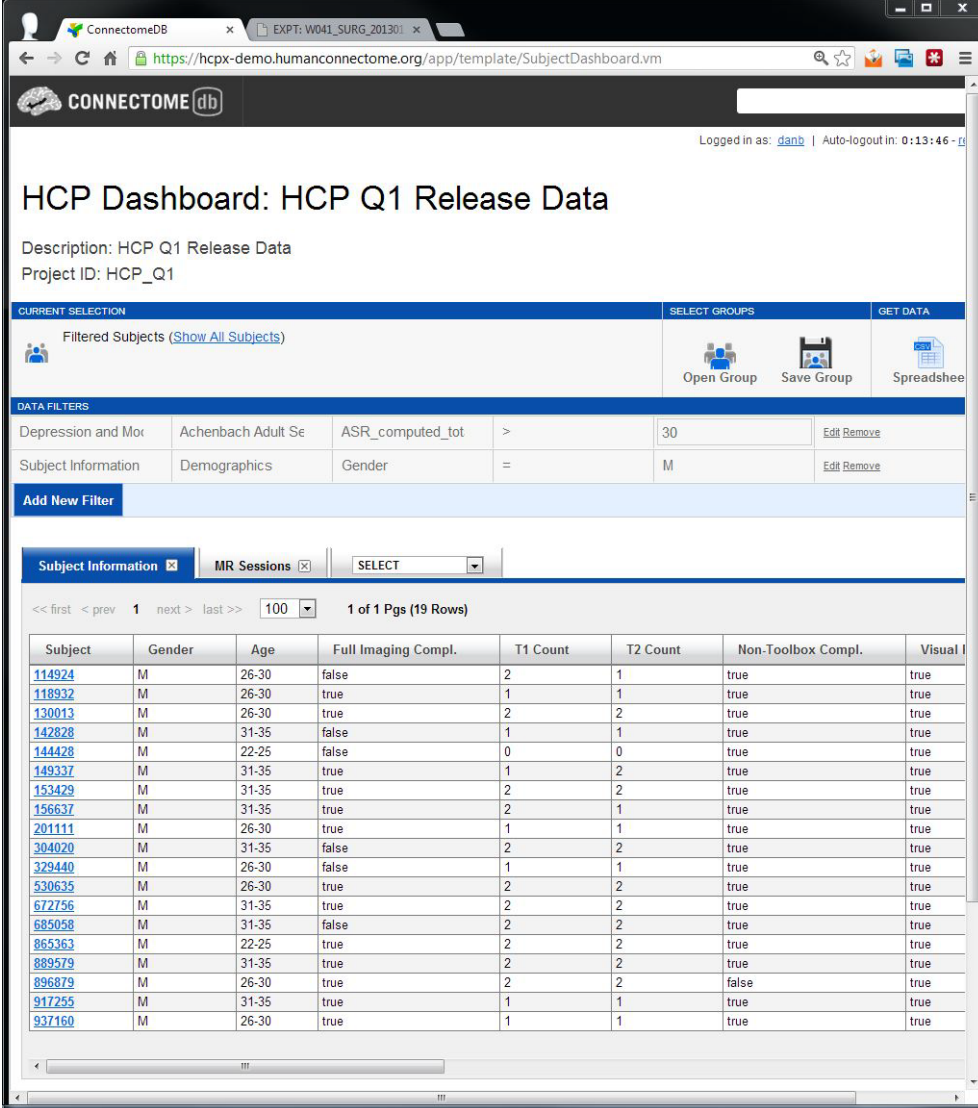
Data sharing

Primary Funding:

NIBIB R01 (Year 16)

NCI ITCR U24 (Year 8)

Commercial open source software



CONNECTOME db

Logged in as: [dant](#) | Auto-logout in: 0:13:46 - t

HCP Dashboard: HCP Q1 Release Data

Description: HCP Q1 Release Data
Project ID: HCP_Q1

CURRENT SELECTION: Filtered Subjects ([Show All Subjects](#))

SELECT GROUPS: Open Group, Save Group, Spreadshee

GET DATA

DATA FILTERS

Depression and Mor	Achenbach Adult Se	ASR_computed_tot	>	30	Edit Remove
Subject Information	Demographics	Gender	=	M	Edit Remove

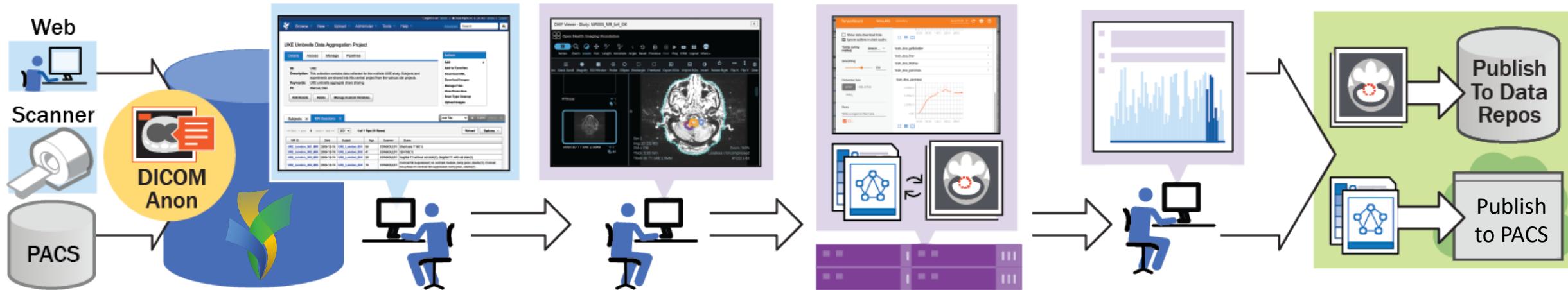
[Add New Filter](#)

Subject Information MR Sessions

<< first < prev 1 next > last >> 100 1 of 1 Pgs (19 Rows)

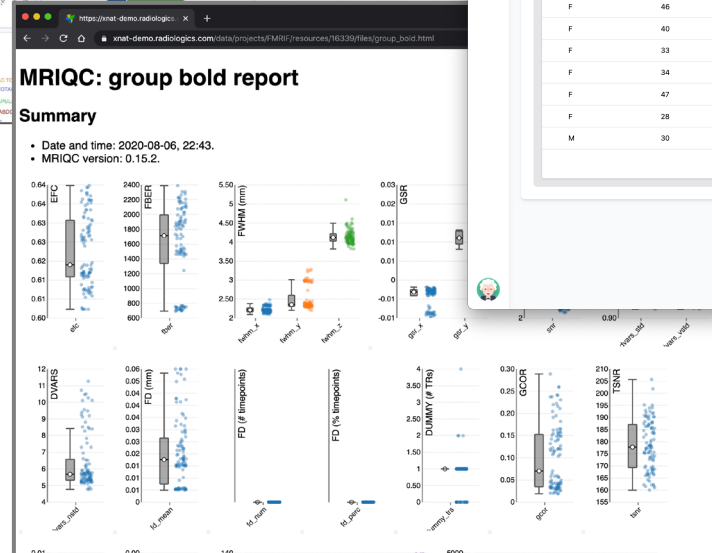
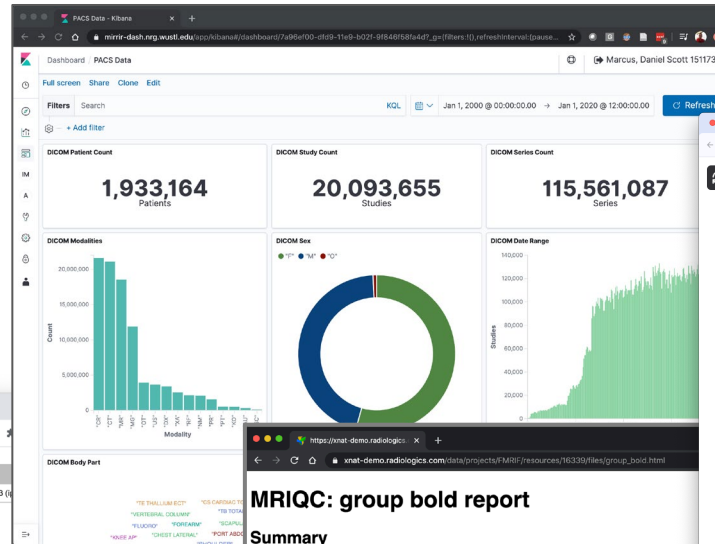
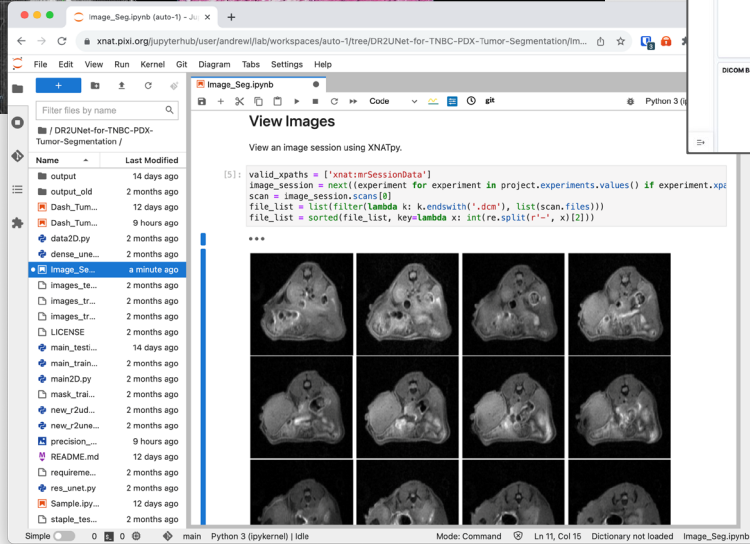
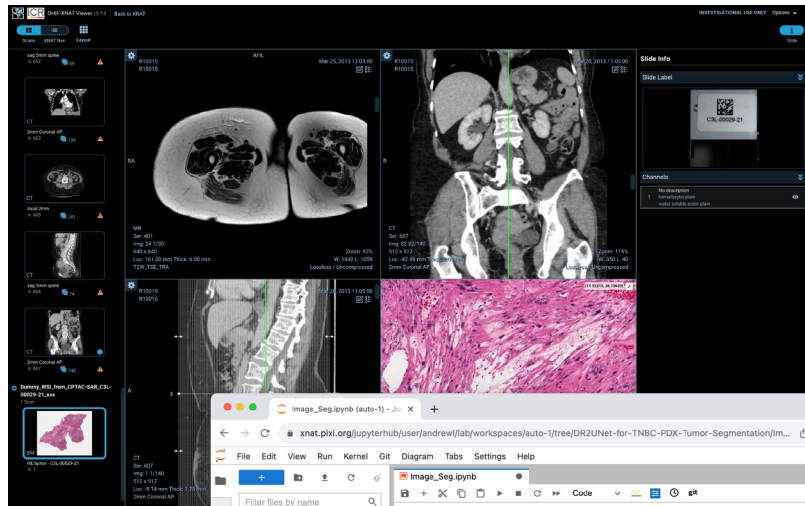
Subject	Gender	Age	Full Imaging Compl.	T1 Count	T2 Count	Non-Toolbox Compl.	Visual
114924	M	26-30	false	2	1	true	true
118932	M	26-30	true	1	1	true	true
130013	M	26-30	true	2	2	true	true
142828	M	31-35	false	1	1	true	true
144428	M	22-25	false	0	0	true	true
149337	M	31-35	true	1	2	true	true
153429	M	31-35	true	2	2	true	true
156637	M	31-35	true	2	1	true	true
201111	M	26-30	true	1	1	true	true
304020	M	31-35	false	2	2	true	true
329440	M	26-30	false	1	1	true	true
530635	M	26-30	true	2	2	true	true
672756	M	31-35	true	2	2	true	true
685058	M	31-35	false	2	2	true	true
865363	M	22-25	true	2	2	true	true
889579	M	31-35	true	2	2	true	true
896879	M	26-30	true	2	2	false	true
917255	M	31-35	true	1	1	true	true
937160	M	26-30	true	1	1	true	true

I3CR: XNAT for cancer research



INGEST	CURATE	ANNOTATE	COMPUTE	ANALYZE	PUBLISH
<ul style="list-style-type: none"> • DICOM send from scanner • DICOM retrieve from PACS • Web-based upload • Metadata anonymization • Pixel anonymization • Face scrubbing • Support for all imaging modalities. • Support for all file formats (DICOM, NIFTI, TIFF, etc.) 	<ul style="list-style-type: none"> • Automated image QC • Manual image review • Pre-archive data staging • Scan type harmonization • Group-level image statistics • Linked clinical, genetic, and other non-imaging data • Automated format conversion 	<ul style="list-style-type: none"> • OHIF integration • Manual annotation tools (ROI, angles, lines, etc.) • AI-assisted annotation • Templated assessment forms • Rapid reader module • Blinded reader mode • Radreport.org integration 	<ul style="list-style-type: none"> • Docker Container Service • HPC integration (Slurm, Torque, PBS, etc.) • Batch processing dashboard • Live Tensorboard and stdout/stderr monitoring • Event-based container automation 	<ul style="list-style-type: none"> • Standalone Jupyter integration • CLARA-based ML service • Scriptable data sets for training, test, validation • Scriptable analysis with Python and Groovy • Integrations with common analysis tools (3D Slicer, MITK, CapTk, etc.) 	<ul style="list-style-type: none"> • Data sharing across projects • Data sharing across nodes • Federated datasets across instances • Automated export to public data repositories (e.g. TCIA) • Automated export to cloud storage • Export models to clinical application hosts (e.g. Sectra Amplifier)

I3CR: XNAT for cancer research



A screenshot of the Scout Data Explorer interface. It shows a cohort of patients with emphysema within a cosine distance of 1. The interface includes a table of patient data and a COPILOT chatbot for generating summary charts.

Scout Data Explorer

Analyze and visualize your data.

Cohort: cohort of patients with emphysema within cosine distance 1 (max results 100000)

File ID: file-0FfjKQ@hzbAU7j2x3MDXW

patientsex	patientage	technique_en	findings_en	impressions_e
F	50	Non-contrast images were	CTO is normal. Calibration	No finding compatib
F	38	Non-contrast images were	Bilateral breast prosthesis	Ground glass denari
M	31	Non-contrast images were	No occlusive pathology wi	Millimetric nodular
F	46	Non-contrast images were	The appearance of a bilate	Several nonspecific
F	40	Non-contrast images were	Trachea, both main bronc	Bilateral silicone bre
F	33	Sections were taken witho	Mediastinal structures cou	Micro and macrocra
F	34	Non-contrast images were	Trachea, both main bronc	Thorax CT examinat
F	47	Non-contrast images were	Port chamber and cathete	Operated endometri
F	28	1.5 mm thick non-contrast	Mediastinal structures wer	Minimal sequelae ch
M	30	Non-contrast images were	Bilateral gynecostasia wa	Bilateral gynecostasi

Rows per page: 10 | 1-10 of 5,319

COPILOT: GPT-4o

Free fee to ask questions about this cohort; for example, requesting summary information or plots and charts.

generate summary charts

Here are the summary charts based on the newly provided cohort data:

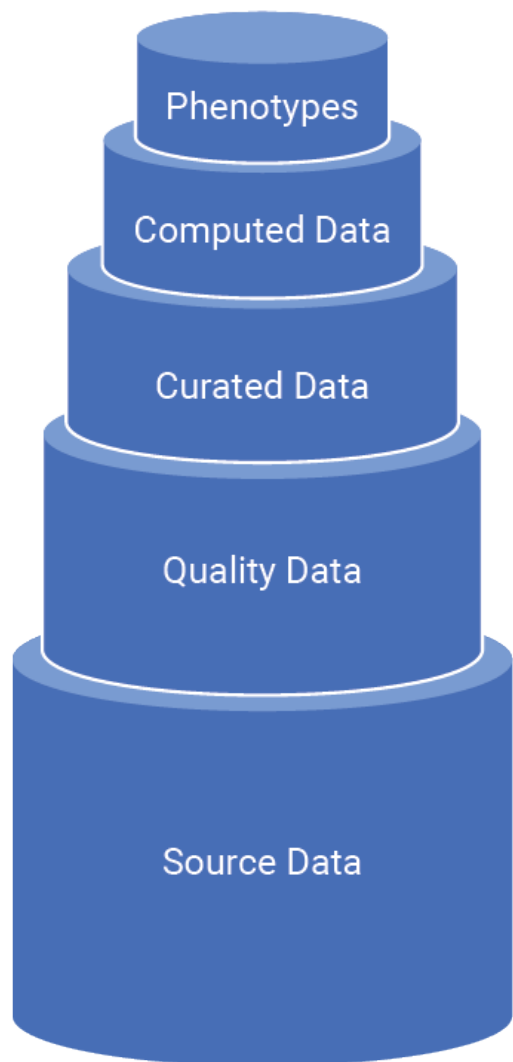
- Distribution of Patient Age:** This histogram shows the frequency distribution of patient ages in the cohort.
- Distribution of Patient Sex:** This bar chart displays the count of patients by sex.
- Distribution of Manufacturers:** This bar chart shows the count of different manufacturers of the equipment used.
- Distribution of Medical Conditions:** This bar chart illustrates the count of various medical conditions present in the cohort.

If you need further analysis or specific details, feel free to ask!

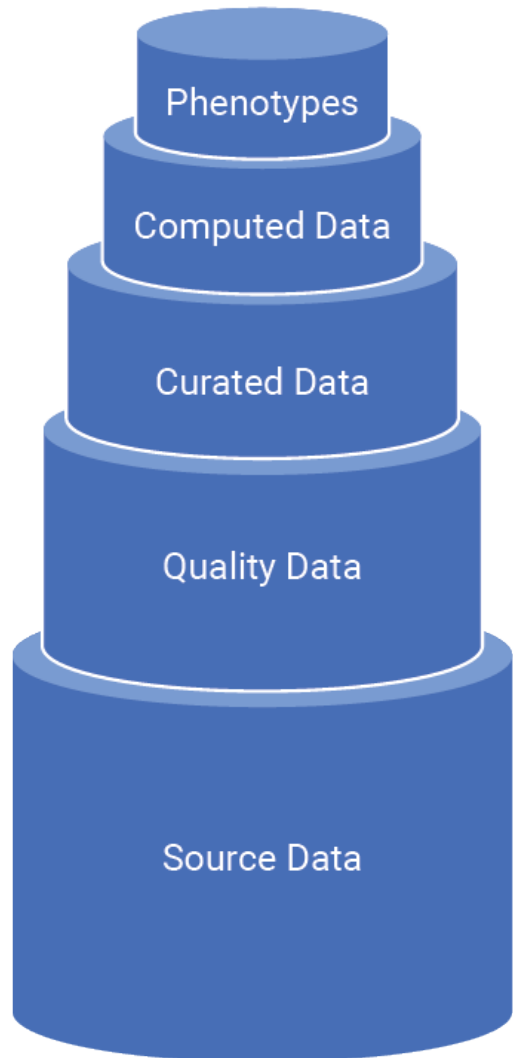
Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack

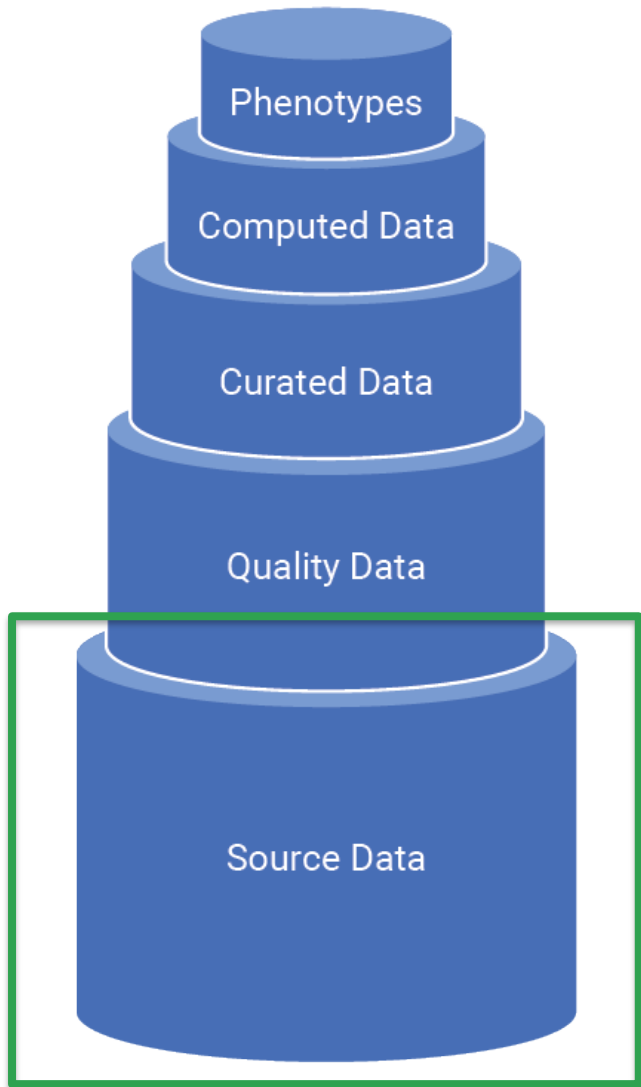


The Feature Stack



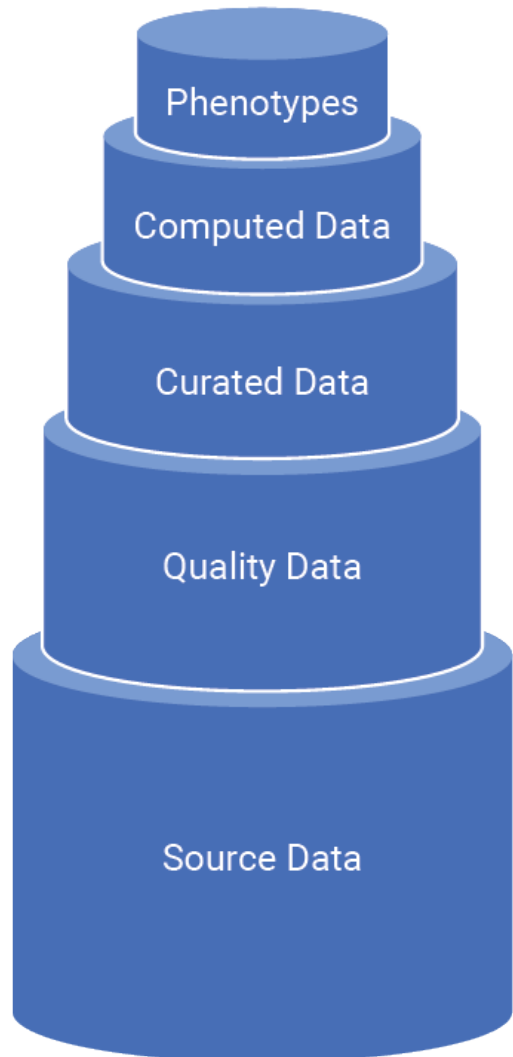
AI enables us to move up the stack to characterize patients in more detail and produce more meaningful & clinically reliable outcomes.

The Feature Stack: Source Data

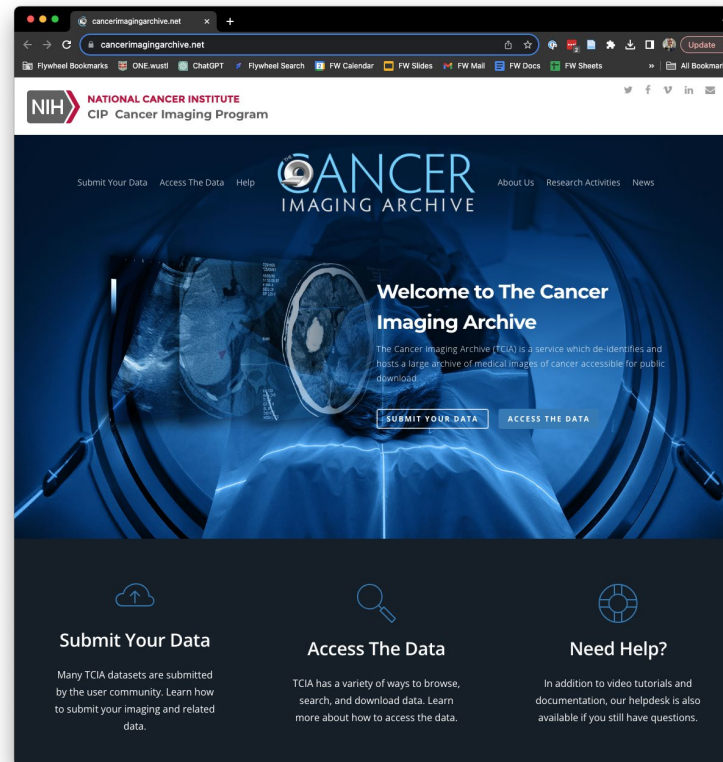


- The raw data on which all subsequent work is done.

The Feature Stack: Source Data



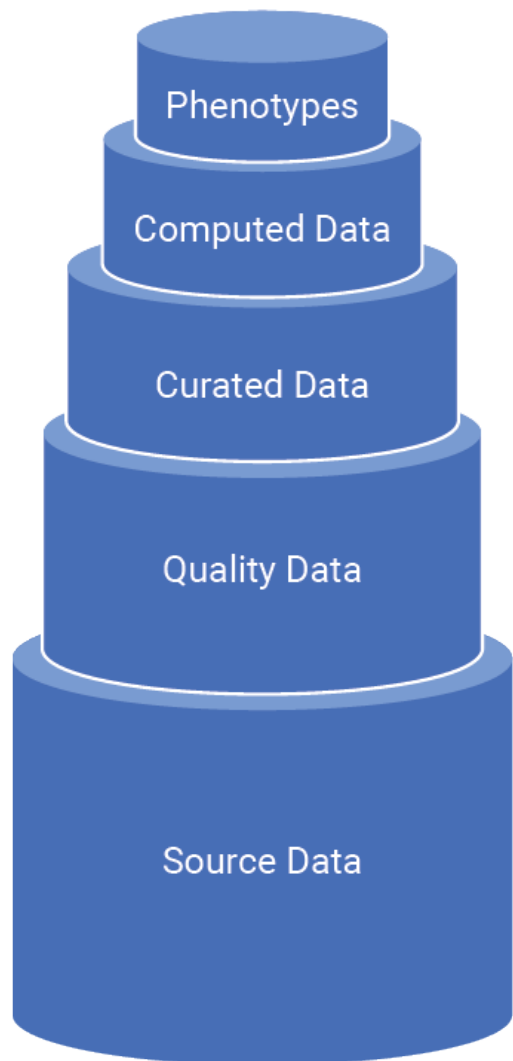
- The raw data on which all subsequent work is done.



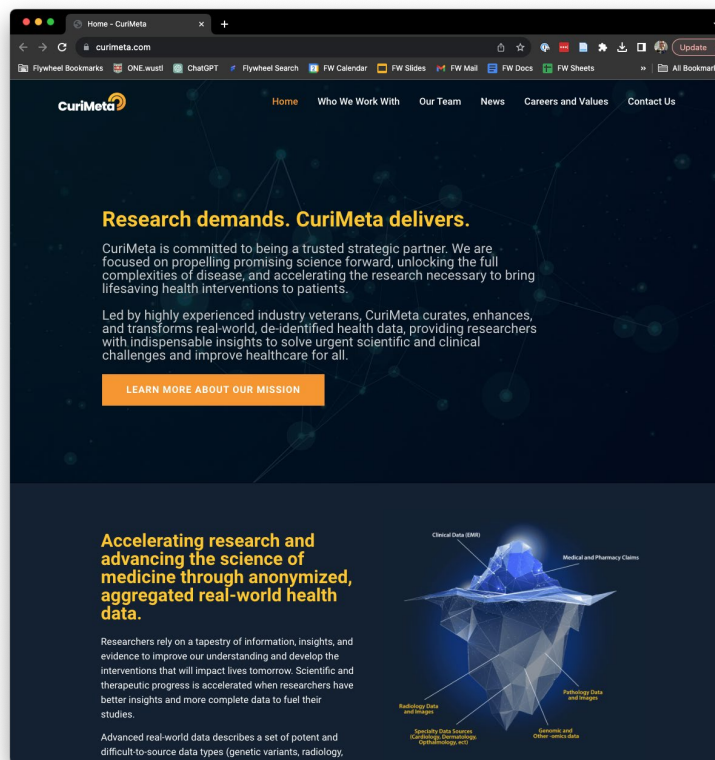
Open Data Resources

- TCIA, MIDRC, Hugging Face, etc.
- Free! (Usually.)
- Regularly used for challenges, can serve as common reference points.
- Limited outcomes and other related metadata.
- Limited novelty and specificity.

The Feature Stack: Source Data



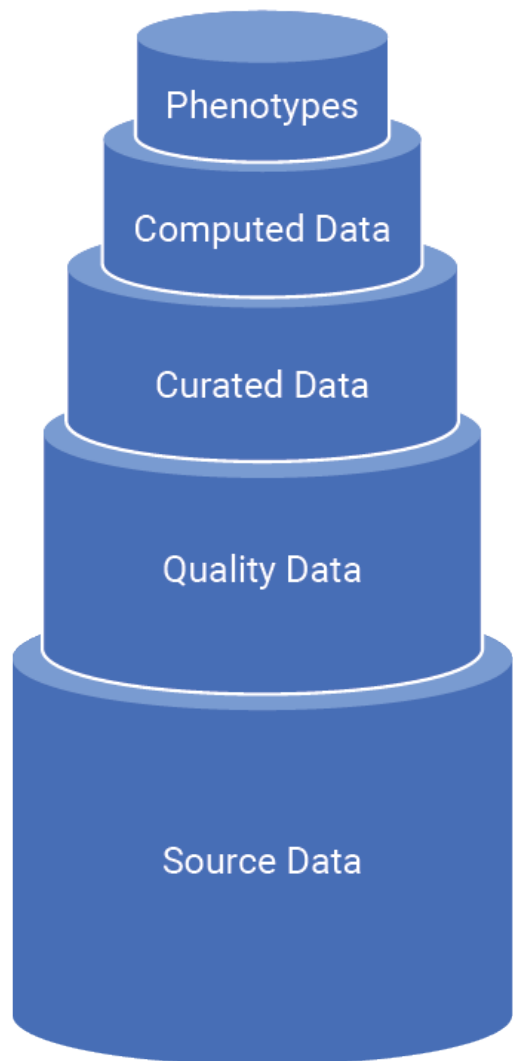
- The raw data on which all subsequent work is done.



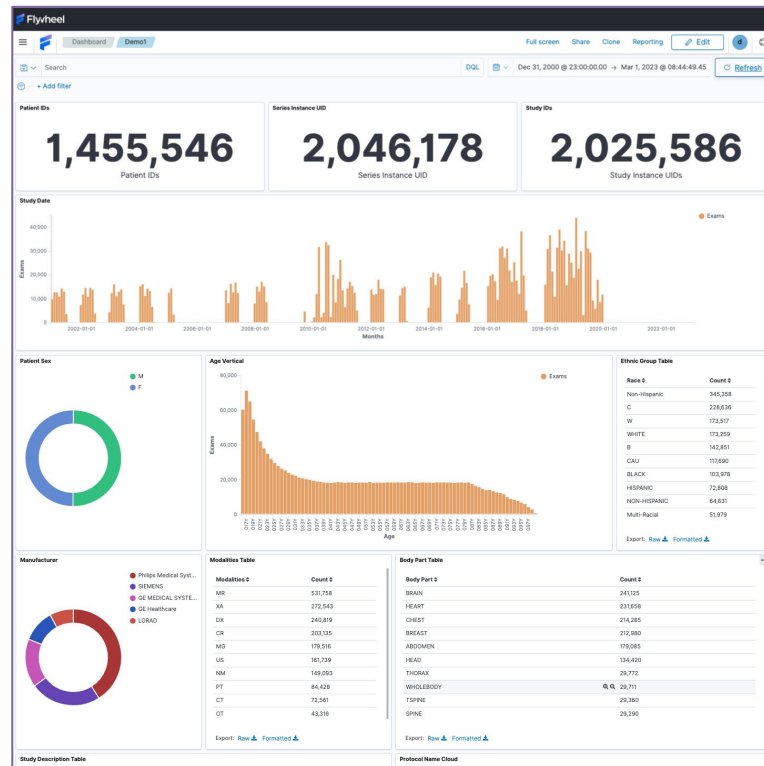
Commercial Data

- Gradient Health, CuriMeta, etc.
- Costly! (Always.)
- Tuned to needs of project
- Typically well validated and well characterized.
- Time to access ranges from instantaneous to protracted.

The Feature Stack: Source Data



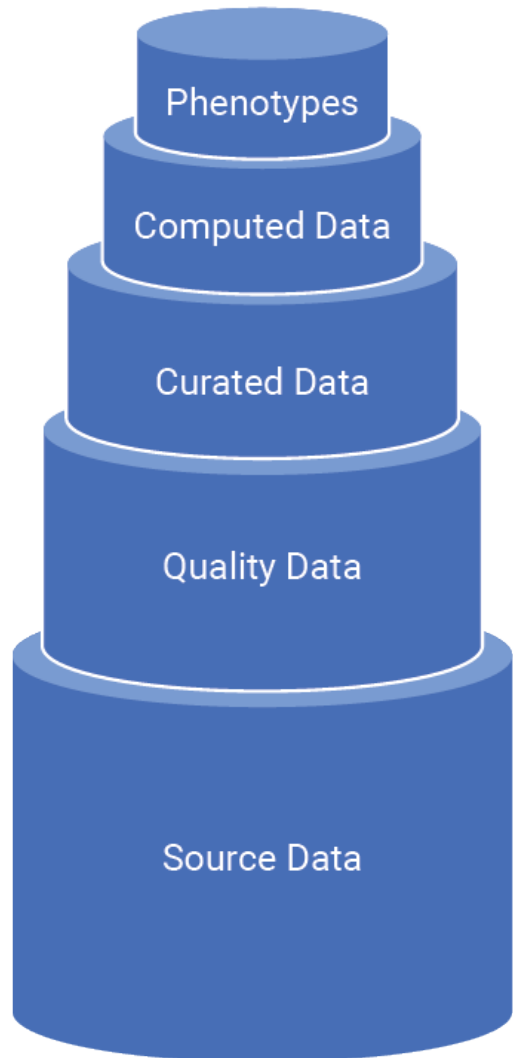
- The raw data on which all subsequent work is done.



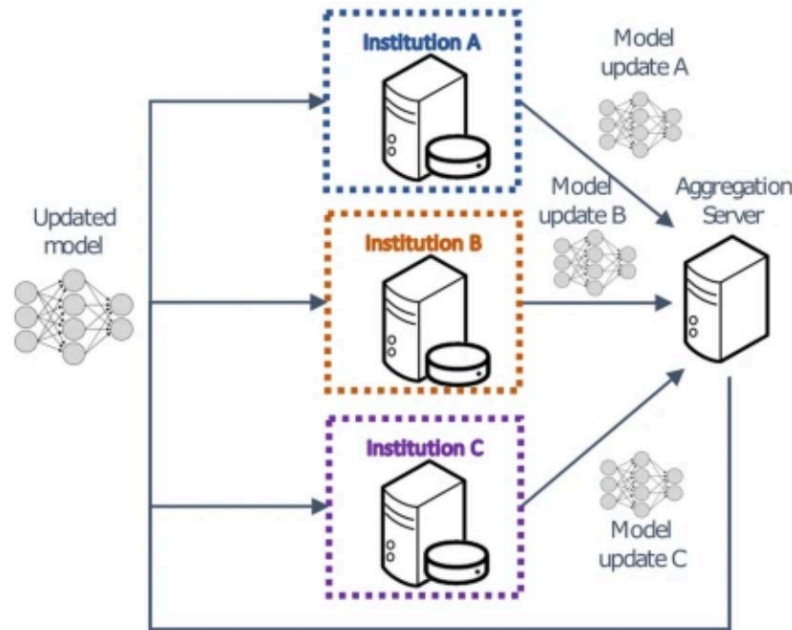
Institutional Data

- Clinical PACS, research data sets
- Low cost! (If you're lucky enough to be in clinical research setting.)
- Scale to fit project
- Cohort identification is up to researcher
- Time to access depends on institution

The Feature Stack: Source Data



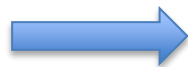
- The raw data on which all subsequent work is done.



Federated Data

- Localized data aggregated at the time of training/analysis.
- Supports more diverse data sets.
- Manages data privacy concerns.
- Introduces new challenges in model development, validation, maintenance.

The Feature Stack: Source Data



Scout Data Explorer
Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer within cosine distance 1 (max results 10,000)
File ID: assistant-gaWG134KoJ60JsHcSJfse8OU

[Download CSV](#) [Export cohort](#)

ID	Sex	Age	Technique	Findings	Impressions	Distance
train_8002_a	M	32	Transverse sections of 1.5	Trachea and lumens of bot	Malignant neoplasm of tes	0.26646137
train_8137_a	M	72	1.5 mm thick non-contrast	Trachea and lumens of both	Lung ca in follow-up; Multi	0.28574675
train_7238_a	F	44	Non-contrast images were	Trachea, both main bronch	Dimensional increases in le	0.31818104
train_6732_a	M	68	Non-contrast images were	Trachea, both main bronch	In a patient with a prediagr	0.33376628
train_12443_a	F	63	Sections were taken witho	Mediastinal structures cou	Lymphadenopathies in the	0.3435843
train_11184_a	M	57	Non-contrast images were	No pneumonic infiltration c	Lung Ca, primary lesion in	0.3610168
train_11961_a	M	59	Sections were taken in the	Mediastinal structures can	Lymphadenopathies in the	0.36426234
train_8809_a	M	56	Non-contrast images were	Trachea, both main bronch	Diffuse metastatic lesions in	0.38882113
train_4331_a	M	40	Non-contrast images were	Trachea and both main brc	Lung Ca in the follow-up, s	0.3913501
train_4279_a	F	52	1.5 mm thick non-contrast	Trachea and lumen of both	Operated RCC at follow-up	0.39915276
train_4080_a	M	57	1.5 mm thick sections were	In the patient with multiple	Multiple myeloma at follow	0.40074855
train_11885_a	F	50	Non-contrast images were	Trachea and both main brc	Breast Ca. In the upper lob	0.40704226
train_3036_a	M	58	1.5 mm thick non-contrast	No occlusive pathology wa	Not given.	0.41336262
train_10441_a	F	32	Before IVCM was given, ax	Mediastinal structures can	Metastases in both lungs,	0.42043048
train_8987_a	F	50	1.5 mm thick non-contrast	An image of a catheter ext	Hodgkin lymphoma at follc	0.42082763

Rows per page: 15 | 1-15 of 3,685

COPILOT Azure GPT-4o

patients under 85 with lung cancer

I found 3,685 hits for query "patients under 85 with lung cancer" within cosine distance ≤ 1 . Your data explorer has been updated with this cohort.

Let's generate some charts for this data!

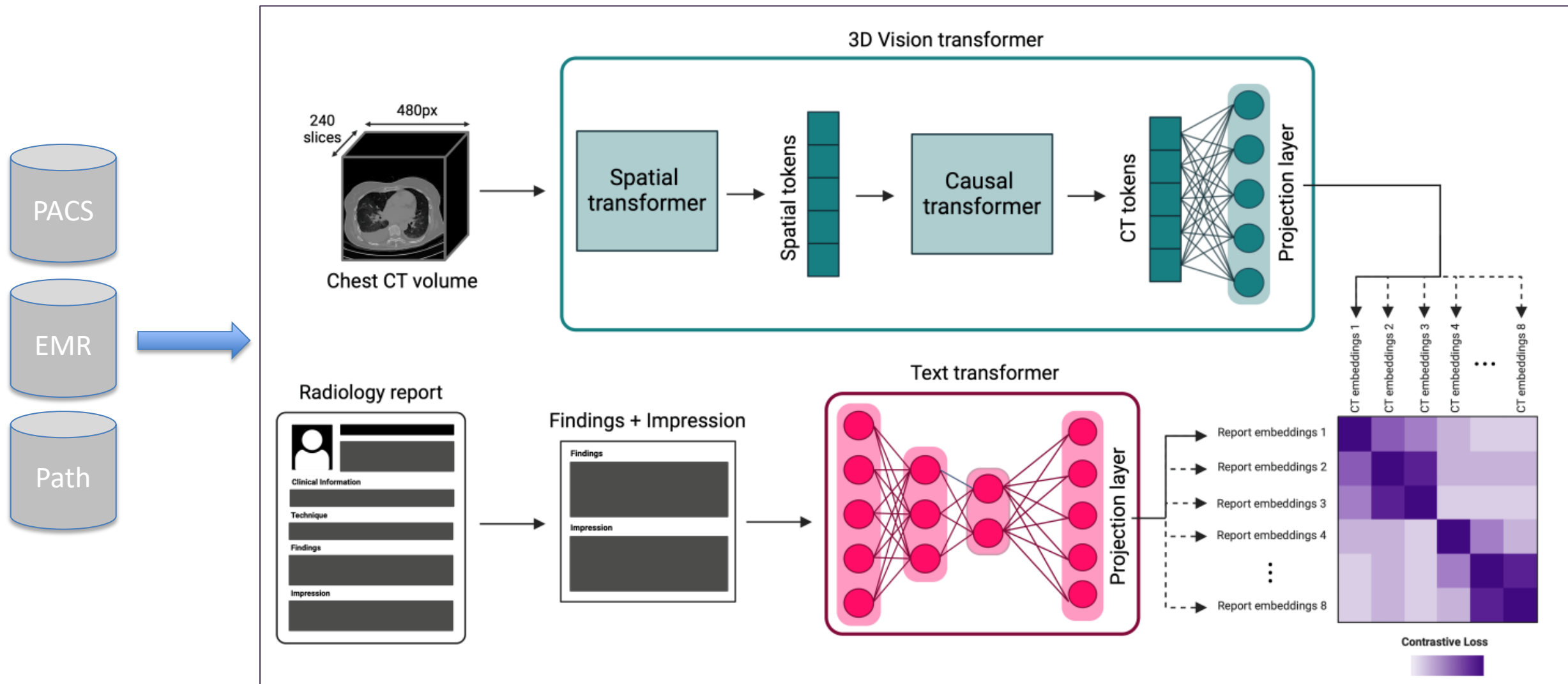
I am generating some code to fulfill the request.

[Review code](#)

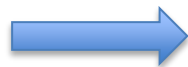
Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack: Source Data



The Feature Stack: Source Data



Scout Data Explorer

Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer within cosine distance 1 (max results 10,000)
File ID: assistant-gaWG134KoJ60JsHcSJfse8OU

[Download CSV](#) [Export cohort](#)

ID	Sex	Age	Technique
train_8002_a	M	32	Transverse sections of 1.5
train_8137_a	M	72	1.5 mm thick non-contrast
train_7238_a	F	44	Non-contrast images were
train_6732_a	M	68	Non-contrast images were Trachea, both main bronch In a patient with a prediagr 0.33376628
train_12443_a	F	63	Sections were taken witho Mediastinal structures cou Lymphadenopathies in the 0.3435843
train_11184_a	M	57	Non-contrast images were No pneumonic infiltration Lung Ca, primary lesion in 0.3610168
train_11961_a	M	59	Sections were taken in the Mediastinal structures can Lymphadenopathies in the 0.36426234
train_8809_a	M	56	Non-contrast images were Trachea, both main bronch Diffuse metastatic lesions in 0.38882113
train_4331_a	M	40	Non-contrast images were Trachea and both main brc Lung Ca in the follow-up, s 0.3913501
train_4279_a	F	52	1.5 mm thick non-contrast Trachea and lumen of both Operated RCC at follow-up 0.39915276
train_4080_a	M	57	1.5 mm thick sections were In the patient with multiple Multiple myeloma at follow 0.40074855
train_11885_a	F	50	Non-contrast images were Trachea and both main brc Breast Ca. In the upper lob 0.40704226
train_3036_a	M	58	1.5 mm thick non-contrast No occlusive pathology wa Not given. 0.41336262
train_10441_a	F	32	Before IVCM was given, ax Mediastinal structures can Metastases in both lungs, 0.42043048
train_8987_a	F	50	1.5 mm thick non-contrast An image of a catheter ext Hodgkin lymphoma at follc 0.42082763

Rows per page: 15 | 1-15 of 3,685

COPILLOT Azure GPT-4o

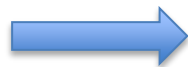
patients under 85 with lung cancer

I found 3,685 hits for query "patients under 85 with lung cancer" within cosine distance ≤ 1 . Your data explorer has been updated with this cohort.

Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack: Source Data



Scout Data Explorer

Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer and emphysema within cosine distance 1 (max results 10,000)
File ID: assistant-upA7FsZZvPTFBXqKDL6tnlz3

[Download CSV](#) [Export cohort](#)

ID	Sex	Age	Technique	Findings	Score
train_3036_a	M	58	1.5 mm thick non-contrast		
train_10850_a	M	72	1.5 mm thick non-contrast		
train_5544_a	F	45	Sections were taken in the		
train_7910_a	M	54	Non-contrast images were		
train_6240_a	M	46	Axial sections with a thick	The mediastinal vascular s Other findings are stable.	0.48818338
train_7588_a	F	61	Sections were taken witho	Trachea and both main brc Pancreatic Ca in the follow	0.49542022
train_3098_a	M	59	Sections were taken witho	Mediastinal structures cou Nodules evaluated in favor	0.4964621
train_7128_a	M	76	Before IVKM was given, se	Trachea and both main brc Metastatic colon ca, metas	0.5044378
train_8453_a	M	77	1.5 mm thick non-contrast	Trachea, lumen of both ma Operated lung Ca on follow	0.50992274
train_11744_a	M	68	1.5 mm thick non-contrast	Mediastinal structures wer Soft tissue lesion in the rig	0.51083267
train_9113_a	F	37	Non-contrast images were	Trachea, both main bronch Operated hemangioendotr	0.511675
train_11025_a	F	74	Sections were taken witho	Trachea and both main brc Findings evaluated in favor	0.51344216
train_3414_a	M	46	Non-contrast images were	Tracheal cannula is observ Infiltrative mass in right lur	0.5159209
train_7244_a	M	61	Sections were taken witho	Trachea and both main brc Malignant mass in the upp	0.5244207
train_8037_a	F	66	Non-contrast images were	Trachea and both main brc In the follow-up, a few milli	0.52541083

Rows per page: 15 | 1-15 of 768

COPILOT Azure GPT-4o

narrow to cases with emphysema

I found 768 hits for query "patients under 85 with lung cancer and emphysema" within cosine distance ≤ 1 . Your data explorer has been updated with this cohort.

Medical Condition

Medical Condition	Count
atelectasis	~300
peribronchovascular thickening	~250
bronchiectasis	~200
consolidation	~150
pleural_effusion	~120
hilar_hemia	~100
interlobular_septal_thickening	~80
pericardial_effusion	~60
cardiomegaly	~40
mosaic_attenuation_pattern	~20

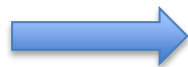
Age Distribution

Frequency vs Age

Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack: Source Data



Scout Data Explorer

Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer and emphysema within cosine distance 1 (max results 1000)
File ID: assistant-upA7FsZZVpTFBXqKDL6tnlz3

[Download CSV](#) [Export cohort](#)

ID	Sex	Age	Technique
train_3036_a	M	58	1.5 mm thick non-contrast
train_10850_a	M	72	1.5 mm thick non-contrast
train_5544_a	F	45	Sections were taken in t
train_7910_a	M	54	Non-contrast images were Trachea, both main bronch Consolidations (pneumoni 0.47161317
train_6240_a	M	46	Axial sections with a thick The mediastinal vascular s Other findings are stable. 0.48818338
train_7588_a	F	61	Sections were taken witho Trachea and both main brc Pancreatic Ca in the follow 0.49542022
train_3098_a	M	59	Sections were taken witho Mediastinal structures cou Nodules evaluated in favor 0.4964621
train_7128_a	M	76	Before IVKM was given, se Trachea and both main brc Metastatic colon ca, metas 0.5044378
train_8453_a	M	77	1.5 mm thick non-contrast Trachea, lumen of both ma Operated lung Ca on follow 0.50992274
train_11744_a	M	68	1.5 mm thick non-contrast Mediastinal structures wer Soft tissue lesion in the rig 0.51083267
train_9113_a	F	37	Non-contrast images were Trachea, both main bronch Operated hemangioendotr 0.511675
train_11025_a	F	74	Sections were taken witho Trachea and both main brc Findings evaluated in favor 0.51344216
train_3414_a	M	46	Non-contrast images were Tracheal cannula is observ Infiltrative mass in right lur 0.5159209
train_7244_a	M	61	Sections were taken witho Trachea and both main brc Malignant mass in the upp 0.5244207
train_8037_a	F	66	Non-contrast images were Trachea and both main brc In the follow-up, a few milli 0.52541083

Rows per page: 15 1-15 of 768

COPILOT Azure GPT-4o

Exclude cases with pneumonia

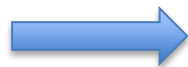
I found 587 hits for query "patients under 85 with lung cancer and emphysema without pneumonia" within cosine distance ≤ 0.9 . Your data explorer has been updated with this cohort.

Medical Condition	Count
lung_opacity	300
atelectasis	250
peribronchovascular_thickening	200
bronchiectasis	150
consolidation	100
pleural_effusion	100
hilar_hernia	100
interlobular_septal_thickening	100
pericardial_effusion	100
cardiomegaly	100
mosaic_attenuation_pattern	100

Age	Frequency
20	5
30	10
40	20
50	35
60	55
70	75
80	50

Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

The Feature Stack: Source Data



Scout Data Explorer
Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer and emphysema within cosine distance 1 (max results 10,000)
File ID: assistant-upA7FsZZVpTFBXqKDL6tnlz3

Download CSV | Export cohort

ID	Sex	Age	Technique	Findings	Impressions	Distance
train_3036_a	M	58	1.5 mm thick non-contrast	No occlusive pathology wa	Not given.	0.41336262
train_10850_a	M	72	1.5 mm thick non-contrast	Trachea and lumen of both	The differential diagnosis c	0.44062293
train_5544_a	F	45	Sections were taken in the	Trachea and both main brc	Metastatic rectum Ca, lung	0.4666534
train_7910_a	M	54	Non-contrast images were	Trachea, both main bronch	Consolidations (pneumoni	0.47161317
train_6240_a	M	46	Axial sections with a thick	The mediastinal vascular s	Other findings are stable.	0.48818338
train_7588_a	F	61	Sections were taken witho	Trachea and both main brc	Pancreatic Ca in the follow	0.49542022
train_3098_a	M	59	Sections were taken witho	Mediastinal structures cou	Nodules evaluated in favor	0.4964621
train_7128_a	M	76	Before IVKM was given, se	Trachea and both main brc	Metastatic colon ca, metas	0.5044378
train_8453_a	M	77	1.5 mm thick non-contrast	Trachea, lumen of both ma	Operated lung Ca on follow	0.50992274
train_11744_a	M	68	1.5 mm thick non-contrast	Mediastinal structures wer	Soft tissue lesion in the rig	0.51083267
train_9113_a	F	37	Non-contrast images were	Trachea, both main bronch	Operated hemangioendotr	0.511675
train_11025_a	F	74	Sections were taken witho	Trachea and both main brc	Findings evaluated in favor	0.51344216
train_3414_a	M	46	Non-contrast images were	Tracheal cannula is observ	Infiltrative mass in right lur	0.5159209
train_7244_a	M	61	Sections were taken witho	Trachea and both main brc	Malignant mass in the upp	0.5244207
train_8037_a	F	66	Non-contrast images were	Trachea and both main brc	In the follow-up, a few milli	0.52541083

Rows per page: 15 | 1-15 of 768

COPILLOT Azure GPT-4o

narrow to cases with emphysema

I found 768 hits for query "patients under 85 with lung cancer and emphysema" within cosine distance ≤ 1 . Your data explorer has been updated with this cohort.

Let's generate some charts for this data!
I am generating some code to fulfill the request.

[Review code](#)

Medical Condition	Count
emphysema	750
lung_nodule	550
arterial_wall_calcification	450
lymphadenopathy	400
pulmonary_fibrosis sequelae	350
lung_specity	300
atelectasis	250
peribronchovascular thickening	200
bronchiectasis	180
consolidation	150
pleural_effusion	120
hilar_hernia	100
interlobular_septal_thickening	80
pericardial_effusion	60
cardiomegaly	40
mosaic_attenuation_pattern	20

Age	Frequency
20	5
25	10
30	15
35	25
40	35
45	45
50	55
55	65
60	75
65	80
70	70
75	55
80	35
85	15

Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack: Source Data



Scout Data Explorer

Analyze and visualize your data.

Cohort
Query: patients under 85 with lung cancer and emphysema within cosine distance 1 (max results 10,000)
File ID: assistant-upA7FsZZVpTFBXqKDL6tnlz3

Download CSV | Export cohort

COPILOT Azure GPT-4o

narrow to cases with emphysema

I found 768 hits for query "patients under 85 with lung cancer and emphysema" within cosine distance ≤ 1 . Your data explorer has been updated with this cohort.

Let's generate some charts for this data!

Findings

Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are findings in favor of pleuroparenchymal sequela fibrotic changes in both lung apex and left lung upper lobe anterior segment anterior segment. There are also occasional linear atelectasis and minimal emphysematous changes in both lungs. There are nodules in both lungs. The nodules were considered to be metastatic. The largest of these nodules are observed in the right lung lower lobe superior segment and right lung lower lobe posterobasal segment, measuring approximately 11x8 mm and 8x9 mm, respectively. However, it is understood that all of them have increased in size. There was no finding in favor of a mass or pneumonic infiltrative in both lungs. A port chamber is observed in the subcutaneous adipose tissue in the right hemithorax. The port catheter terminates in the right atrium. Heart contour and size are normal. No pleural or pericardial effusion was detected. The widths of the mediastinal main vascular structures are normal. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. No pathological wall thickness increase was detected in the esophagus within the sections. No upper abdominal free fluid-collection was detected in the sections. Vertebral corpus heights, alignments and densities within the sections are normal. There are osteophytes in the vertebral corpus corners. The neural foramina are open.

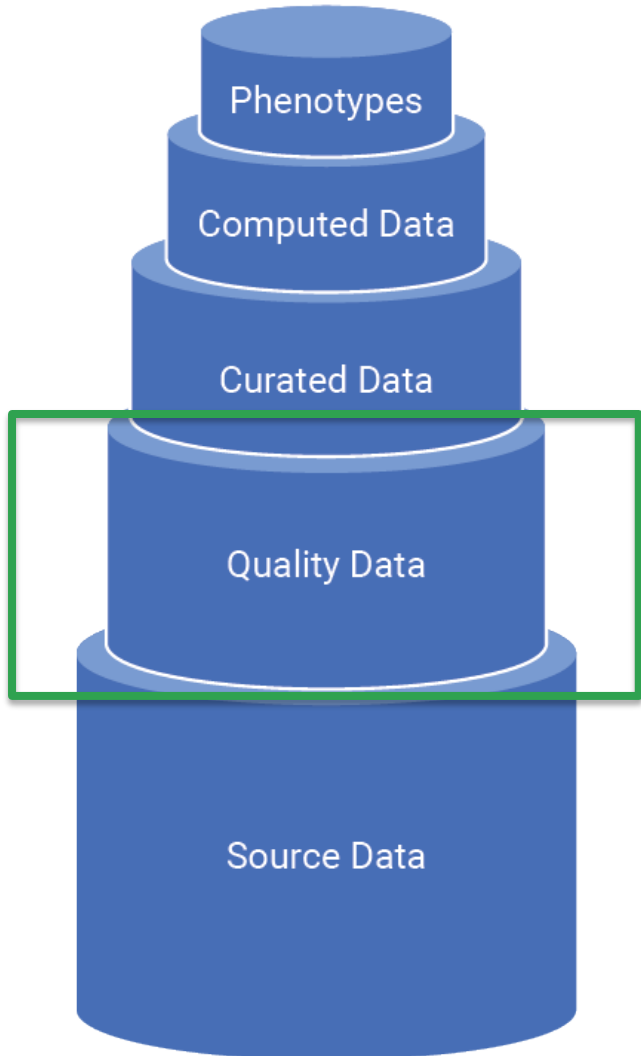
train_3414_a	M	46	Non-contrast images were Tracheal cannula is observ Infiltrative mass in right lur	0.5159209
train_7244_a	M	61	Sections were taken witho Trachea and both main brc Malignant mass in the upp	0.5244207
train_8037_a	F	66	Non-contrast images were Trachea and both main brc In the follow-up, a few milli	0.52541083

Rows per page: 15 | 1-15 of 768

Find data with the requested characteristics. If you include an image, please explain how you wish to use it to inform the query.

Enter your search...

The Feature Stack: Quality Data



- Data on which quality metrics have been obtained systematically to drive subsequent decisions.
- Quality is always an issue, but it's potentially a bigger issue with quantification.
- “Low” quality isn't necessarily bad, but it needs to be understood.
- E.g. if quality is different by site, geography, demographics, etc, you need to ensure that measurements aren't biased.

The Feature Stack: Quality Data

Summary Visual reports Other Toggle rating widget

aor		0.0018386956521739134
aqi		0.0005369426923076923
dummy_trs		1
dvars	nstd	8.399083999865779
dvars	std	1.1955031351342273
dvars	vstd	0.9671125125838926
efc		0.6317237289493478
fber		1802.6905517578125
fd	mean	0.03666213530525616
fd	num	0
fd	perc	0.0
fwhm	avg	2.964602222222222
fwhm	x	2.2335033333333336
fwhm	y	2.391103333333333
fwhm	z	4.2692000000000005
gcor		0.0338436
gsr	x	-0.0023884086403995752
gsr	y	0.009094901382923126
size	t	299

Rate Image ×

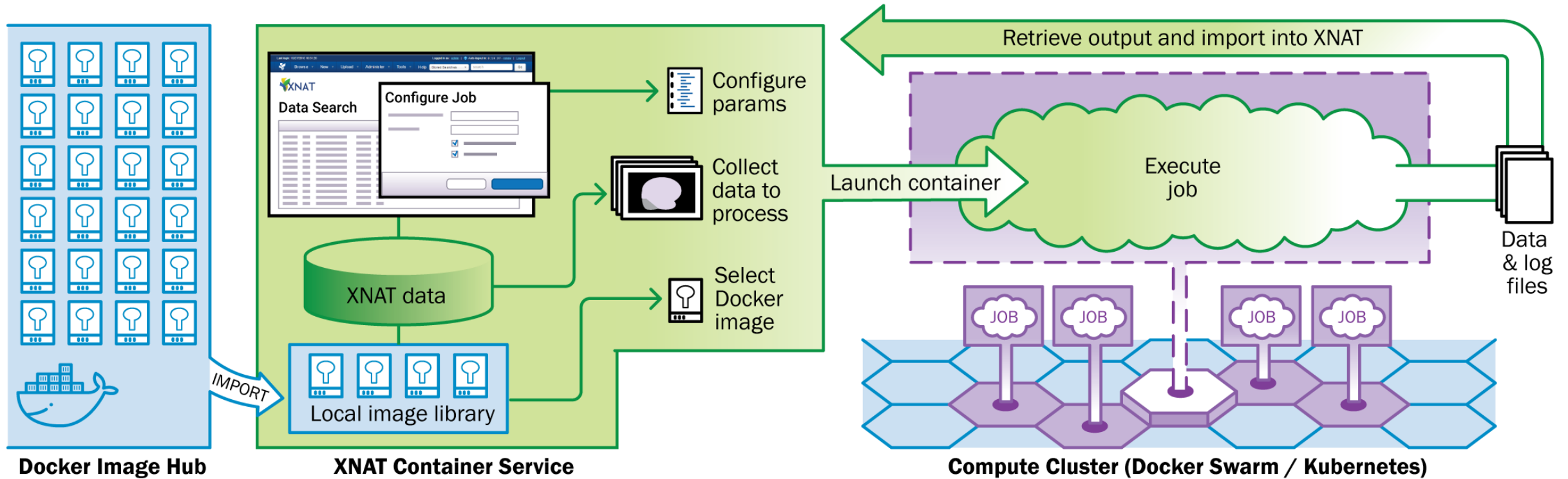
Exclude Poor Acceptable Excellent

- Head motion artifacts
- Eye spillover through PE axis
- Non-eye spillover through PE axis
- Coil failure
- Global noise
- Local noise
- EM interference / perturbation
- Problematic FoV prescription / Wrap-around
- Aliasing ghosts
- Other ghosts
- Intensity non-uniformity
- Temporal field variation
- Reconstruction and postprocessing (e.g. denoising, defacing, resamplings)
- Uncategorized artifact

Download Post to WebAPI

MRQy (Sadri et al, 2020)

The Feature Stack: Quality Data



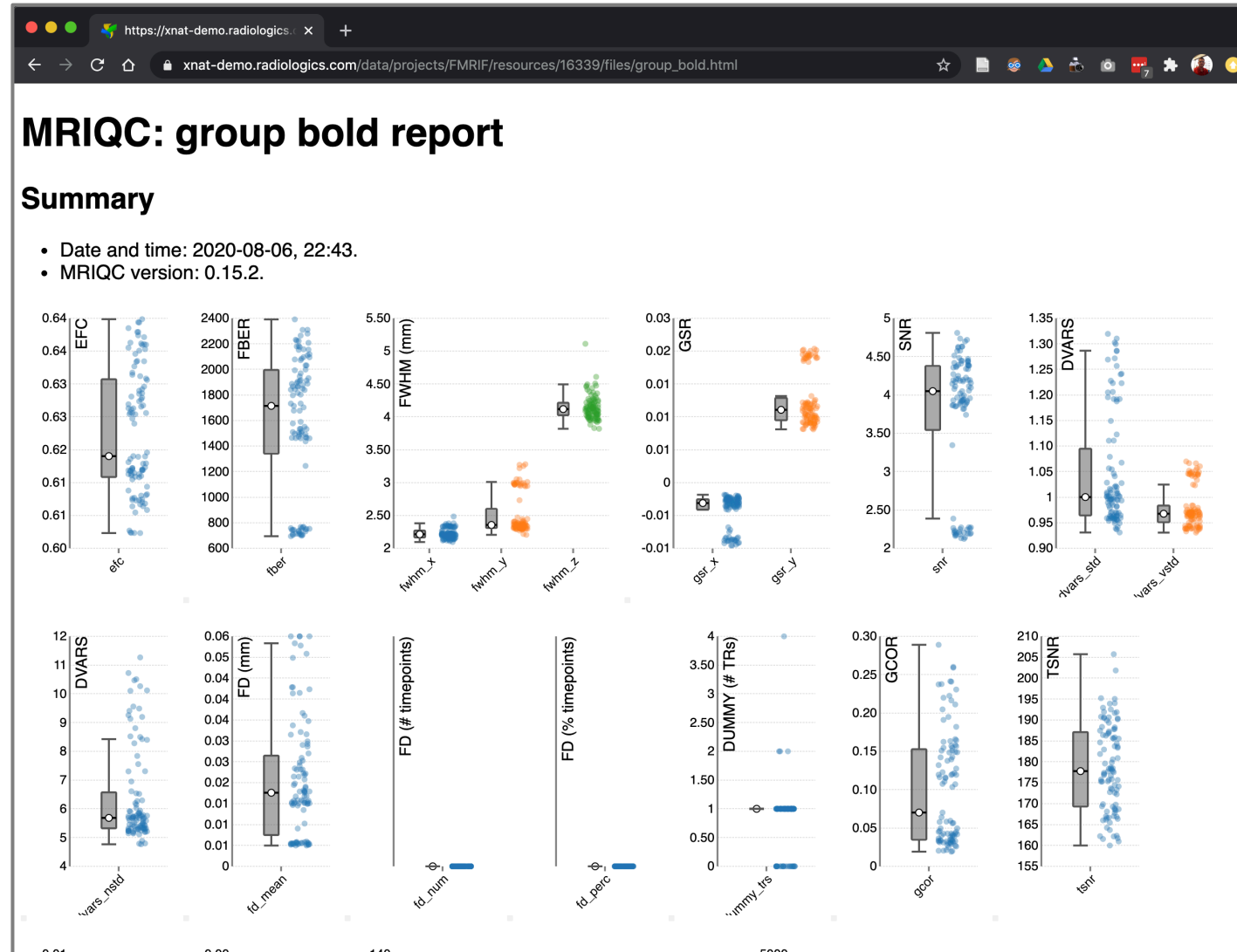
The Feature Stack: Quality Data

PROJECT: CCF_HCD_ITK > Pipeline Control Panel > DCM2NII

Refresh Update Cache Remove Selected Reset Status Info Report Launch Pipeline

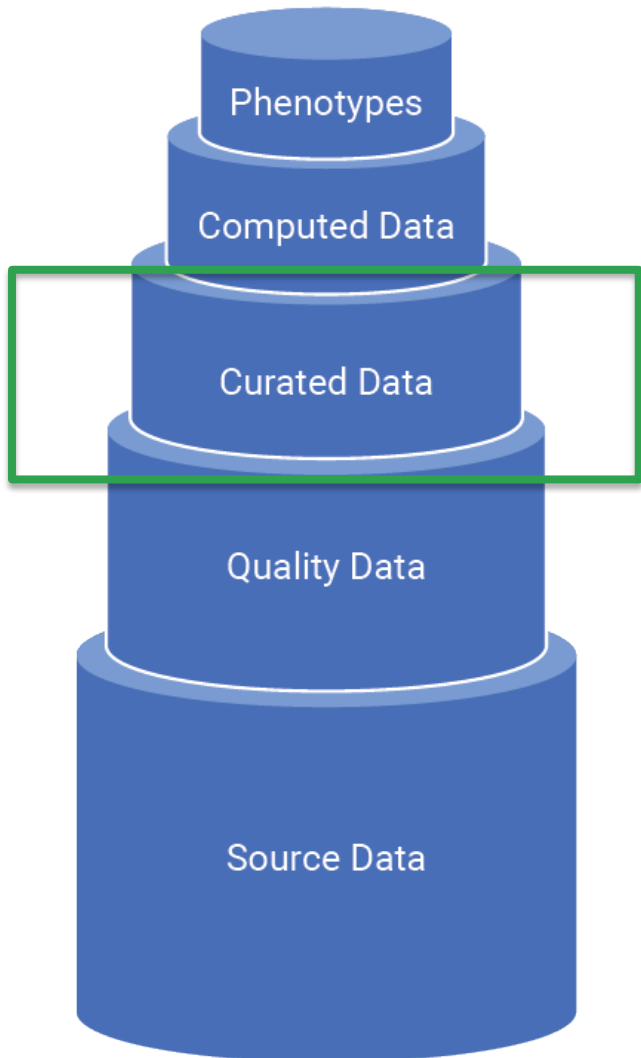
<input type="checkbox"/>	Entity	Subgroup	Status	StatusTime	Prereqs Met	Validated	Issues	Runnable
	Filter by Entity	Filter by	Filter by Status	Filter by StatusTime	All	All	All	All
<input type="checkbox"/>	HCD0001305_V1_A	ALL	ERROR ✖	2018-01-26 17:08:45	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0001305_V1_B	ALL	COMPLETE ✔	2018-01-10 10:47:39	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0015417_V1_A	ALL	COMPLETE ✔	2018-03-13 17:27:55	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0015417_V1_B	ALL	COMPLETE ✔	2018-04-24 19:44:29	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0021614_V1_A	ALL	EXT_COMPLETE ✔	2018-08-05 16:11:44	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0021614_V1_B	ALL	EXT_COMPLETE ✔	2018-08-29 09:16:24	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0022919_V1_A	ALL	EXT_COMPLETE ✔	2018-08-01 23:03:49	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0022919_V1_B	ALL	EXT_COMPLETE ✔	2018-08-07 14:55:03	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0022919_V1_X1	ALL	COMPLETE ✔	2018-08-16 16:53:53	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0024418_V1_A	ALL	EXT_COMPLETE ✔	2018-07-13 03:30:59	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0024418_V1_B	ALL	COMPLETE ✔	2018-08-17 09:12:52	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	HCD0031617_V1_A	ALL	COMPLETE ✔	2018-01-26 17:13:38	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

The Feature Stack: Quality Data



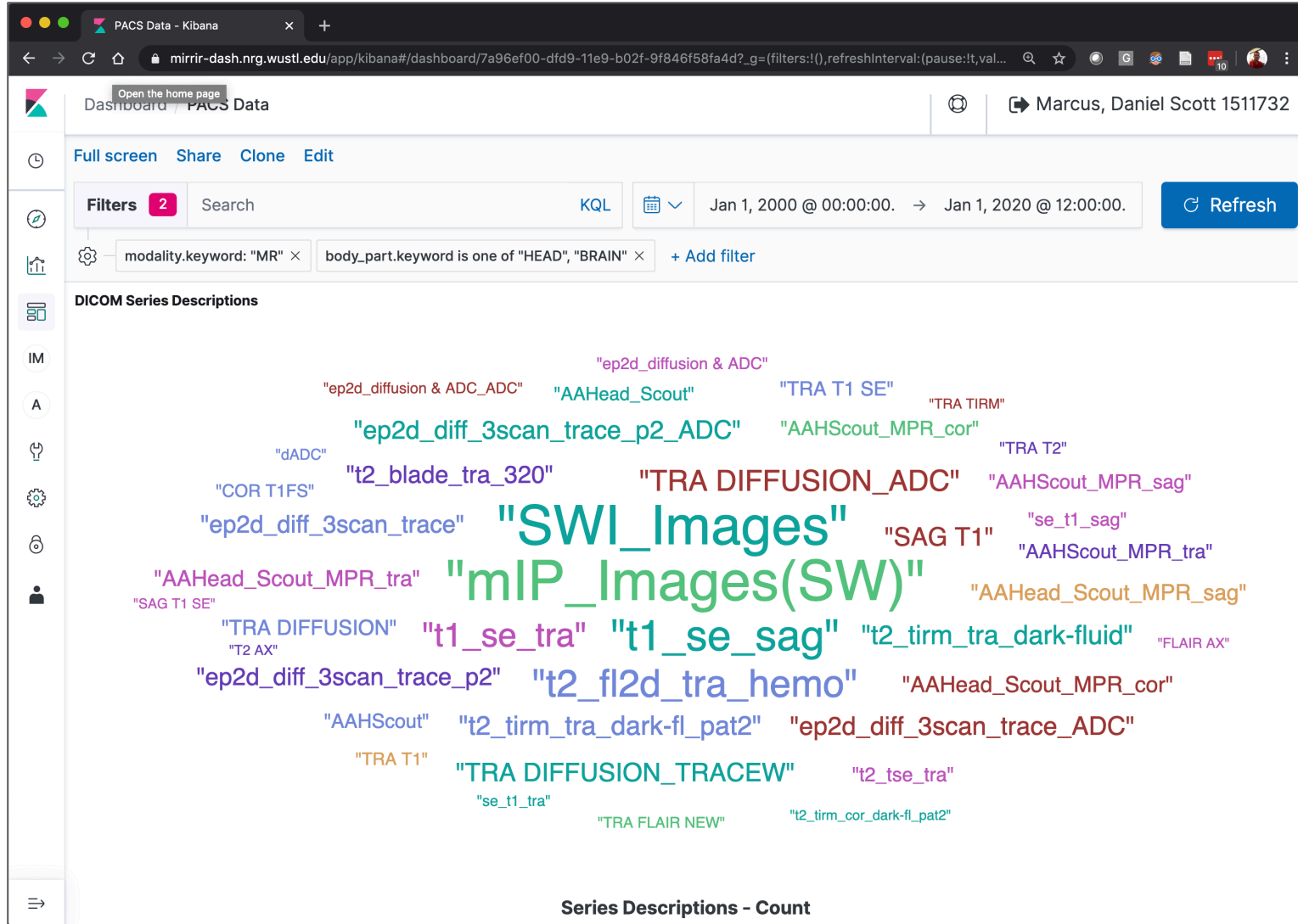
MRQy (Sadri et al, 2020)

The Feature Stack: Curated Data



- Data curation tasks include harmonization, labelling, annotating, sorting, and organizing.
- Curation is traditionally manual, time consuming and error prone.
- AI is really good at handling disparate data, so it's particularly suitable for automating curation tasks.

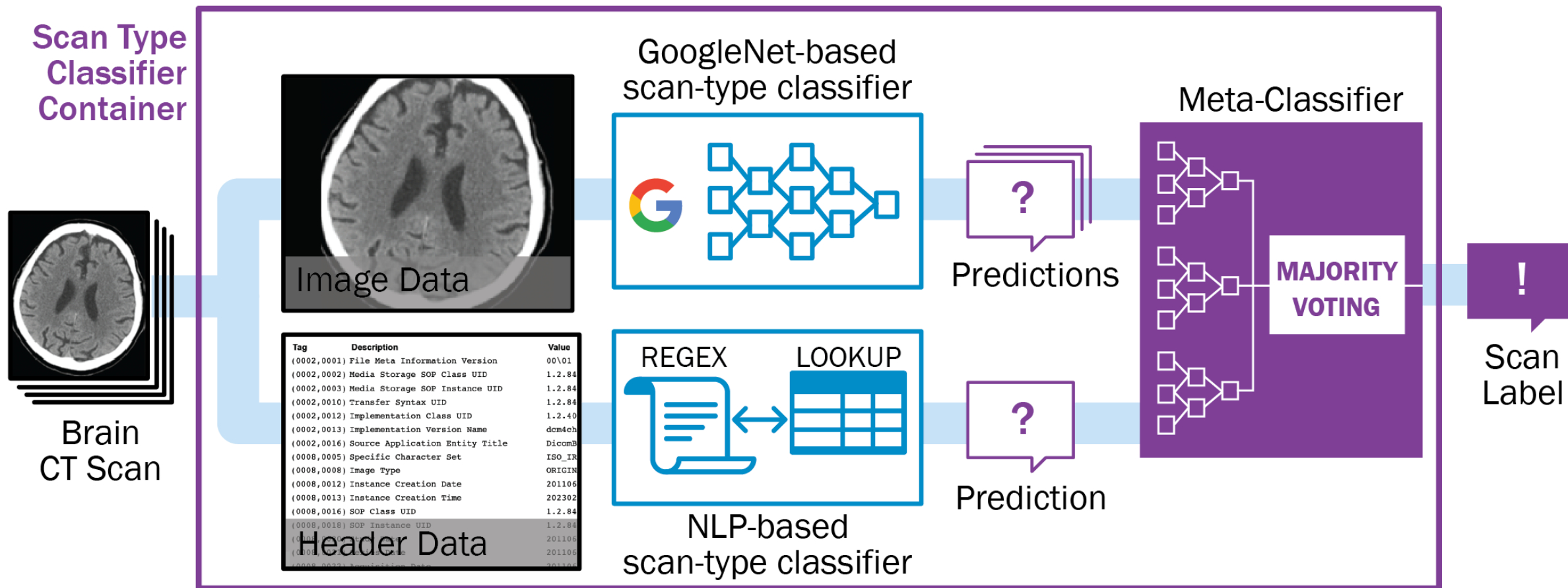
The Feature Stack: Curated Data



DICOM Series Descriptions

Series Descriptions	Count
"mIP_Images(SW)"	67,135
"SWI_Images"	67,053
"t1_se_sag"	50,716
"t2_fl2d_tra_hemo"	43,570
"t1_se_tra"	37,099
"TRA DIFFUSION_ADC"	31,353
"t2_tirm_tra_dark-fl_pat2"	28,995
"t2_tirm_tra_dark-fluid"	28,657
"ep2d_diff_3scan_trace_p2_ADC"	27,426
"TRA DIFFUSION_TRACEW"	27,009
"SAG T1"	26,830
"ep2d_diff_3scan_trace_ADC" ⊕ ⊖	26,259
"ep2d_diff_3scan_trace"	26,011

The Feature Stack: Curated Data



The Feature Stack: Curated Data

PROJECT: test > SUBJECT: S001 > S001_MR_091213_0347_Preop

MR Session: S001_MR_091213_0347_Preop

Details | Projects

Accession #: XNAT_E00001 Subject: S001
 Date Added: 05/03/2019 15:44:06 (admin) Gender:
 Date: 12/12/2009 Handedness:
 Time: 19:47:47 Age: --
 Operator: KEL
 Scanner Name: P-FILESCDBURN,GEHCGEHC
 Scanner Type:
 Acquisition Site: SWEDISH 3T MR750

Actions

- Edit
- View ▶
- Download ▶
- Email
- Manage Files
- View Images
- Delete
- Run Containers ▶

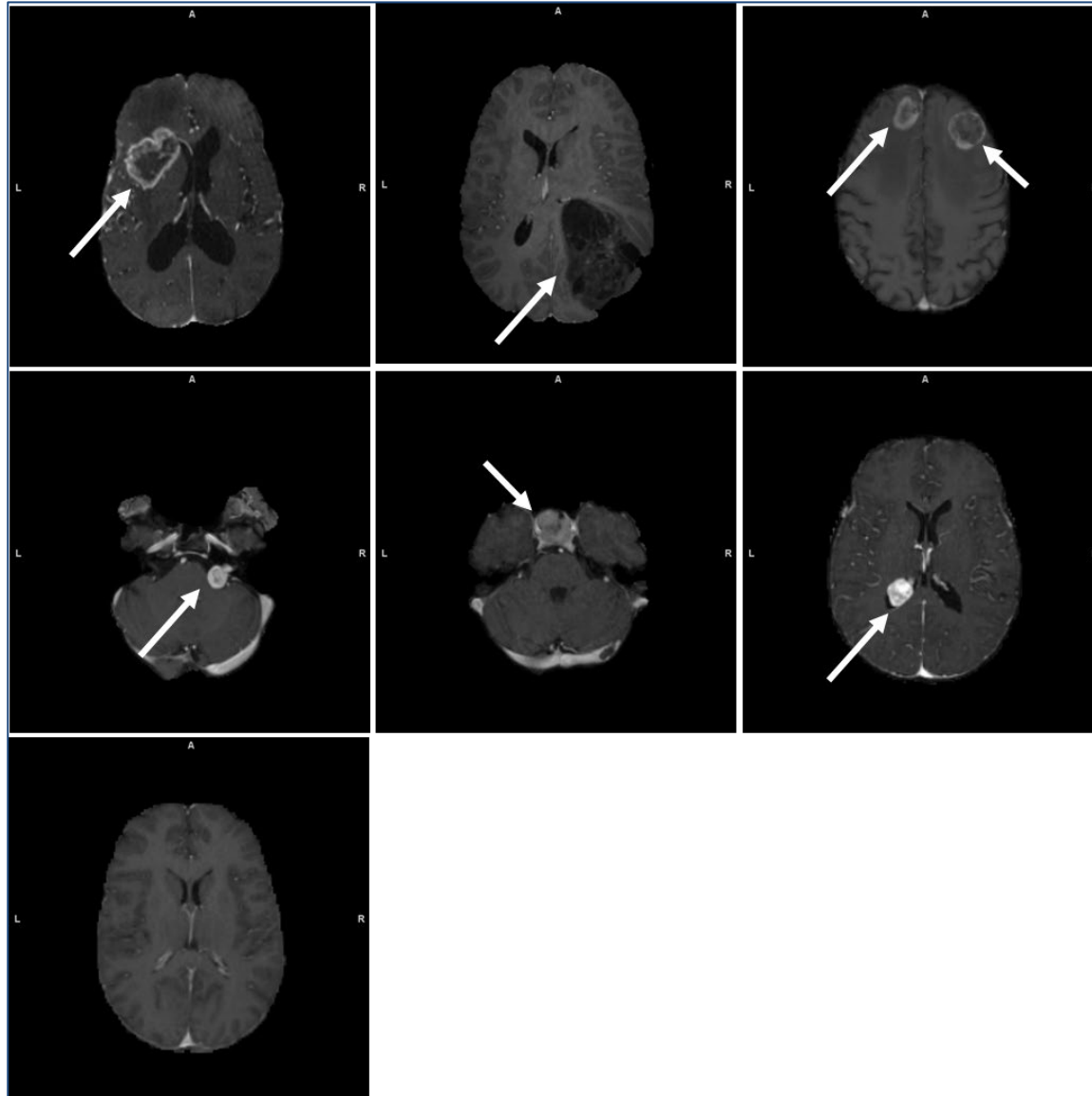
Scans

Bulk Actions: Download Run Container

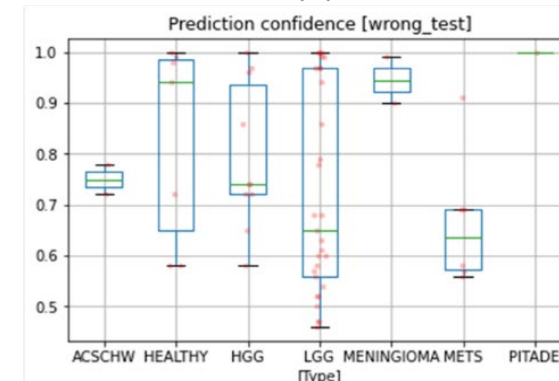
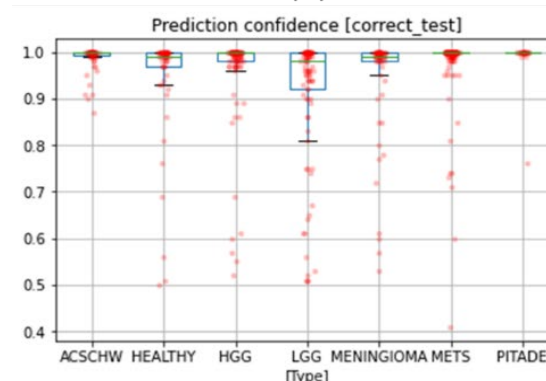
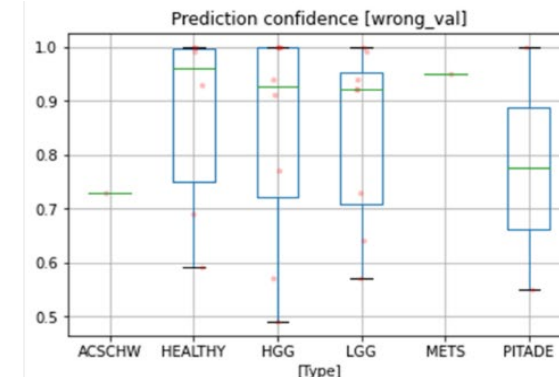
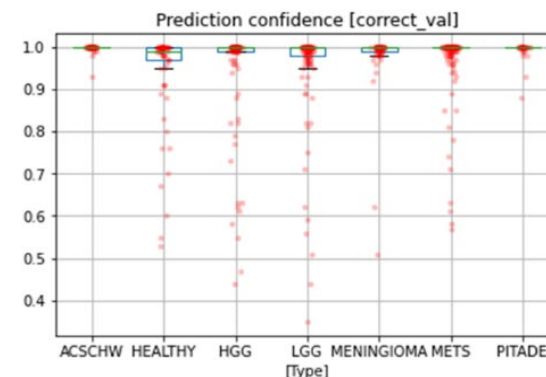
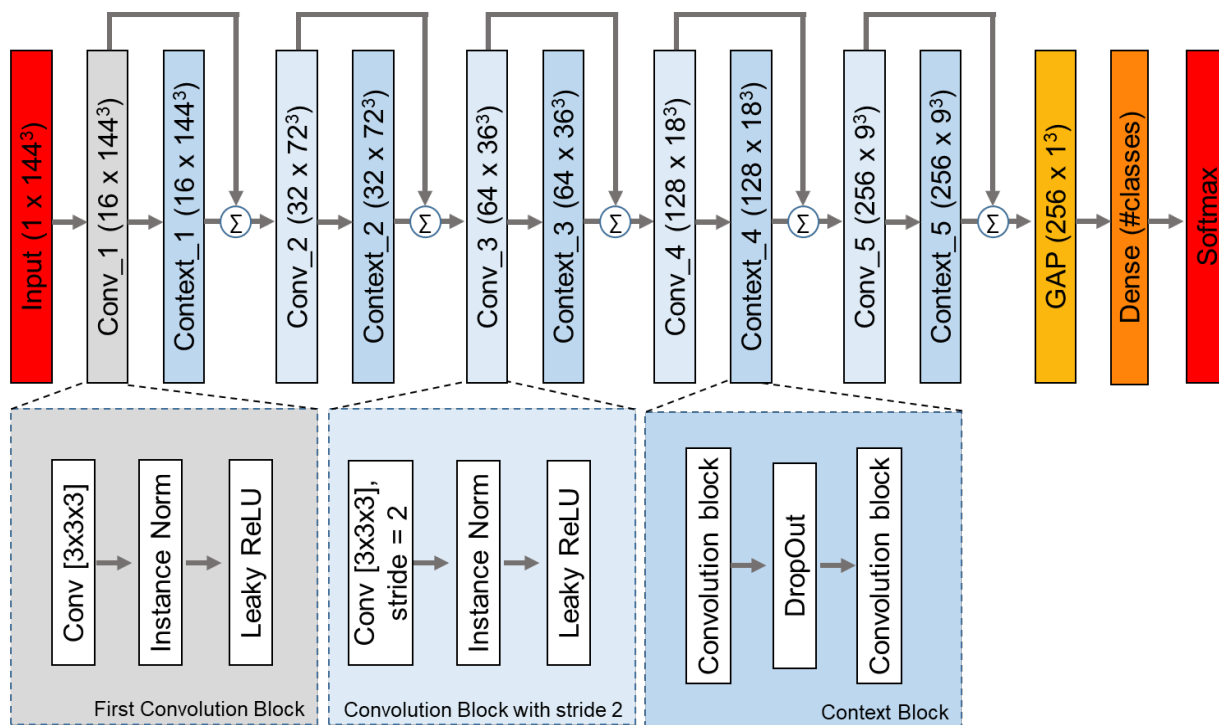
Scan	Series Class	Type	Series Desc	Usability	Files	Note	Run
<input type="checkbox"/>	3	Flair	SAG T1 FLAIR				
<input type="checkbox"/>	4	DWI	AX EPI DTI				
<input type="checkbox"/>	5	TSE	AX T2 FRFSE-XL 2MM	usable	11.8 MB in 85 files		▶
<input type="checkbox"/>	6	Flair	AX T2 FLAIR 3MM	usable	7.9 MB in 57 files		▶
<input type="checkbox"/>	7	ce-Gad_T1w	AX T1 FSPGR 3D	usable	18.9 MB in 136 files		▶
<input type="checkbox"/>	8	DWI	AX EPI DWI	usable	7.2 MB in 52 files		▶
<input type="checkbox"/>	9	ce-Gad_Perfusion	AX EPI PERFUSION	usable	56.2 MB in 1248 files		▶
<input type="checkbox"/>	10	ce-Gad_TSE	COR T2 FSE-XL FS POST	usable	24.2 MB in 47 files		▶
<input type="checkbox"/>	11	ce-Gad_Perfusion	AX EPI PERFUSION #2	usable	56.2 MB in 1248 files		▶
<input type="checkbox"/>	12	ce-Gad_Flair	AX T1 FLAIR POST	usable	13.4 MB in 26 files		▶

Series Class	Type
Flair	SAG T1 FLAIR
DWI	AX EPI DTI
TSE	AX T2 FRFSE-XL 2MM
Flair	AX T2 FLAIR 3MM
ce-Gad_T1w	AX T1 FSPGR 3D
DWI	AX EPI DWI
ce-Gad_Perfusion	AX EPI PERFUSION
ce-Gad_TSE	COR T2 FSE-XL FS POST
ce-Gad_Perfusion	AX EPI PERFUSION #2
ce-Gad_Flair	AX T1 FLAIR POST

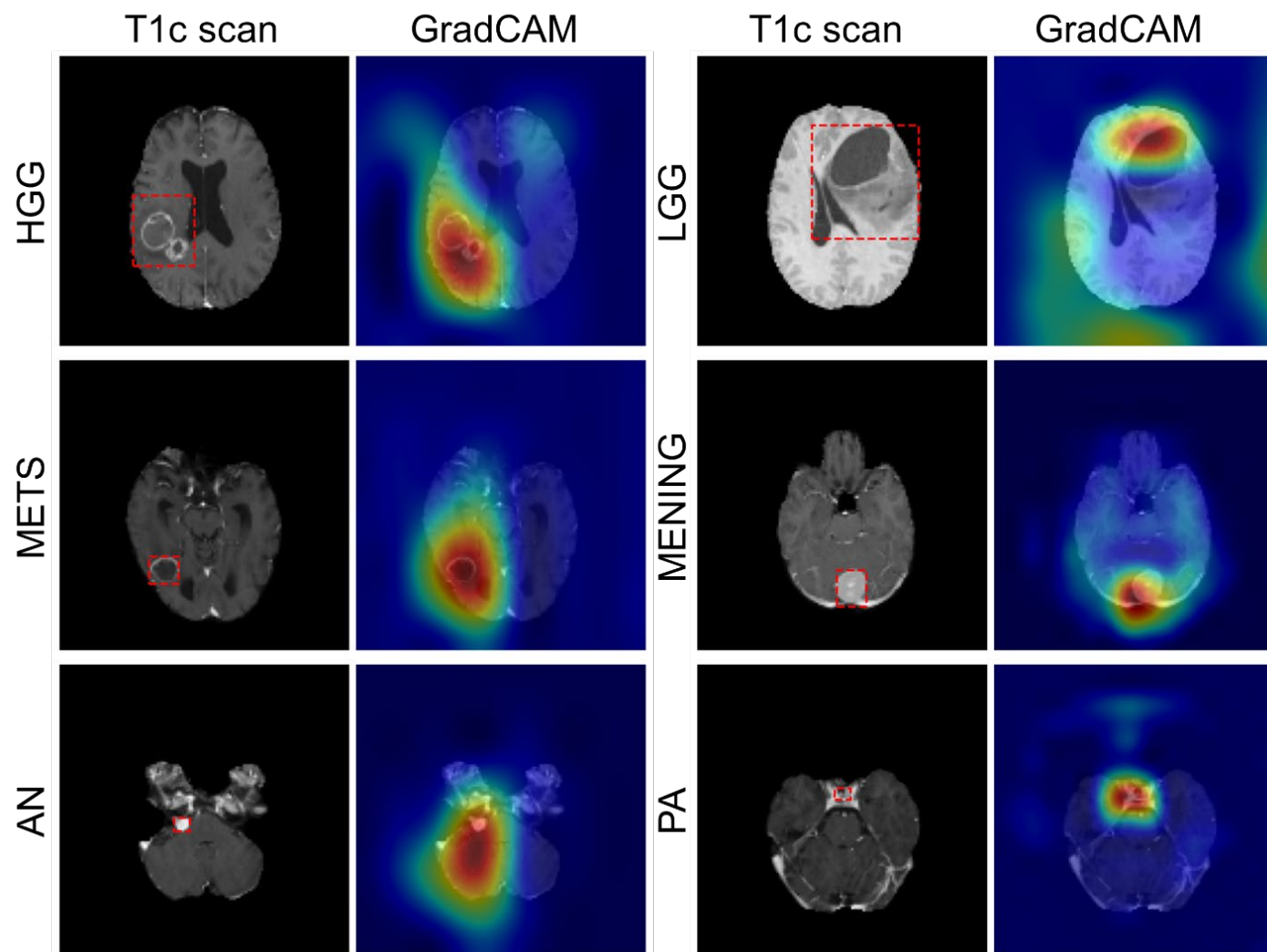
The Feature Stack: Curated Data



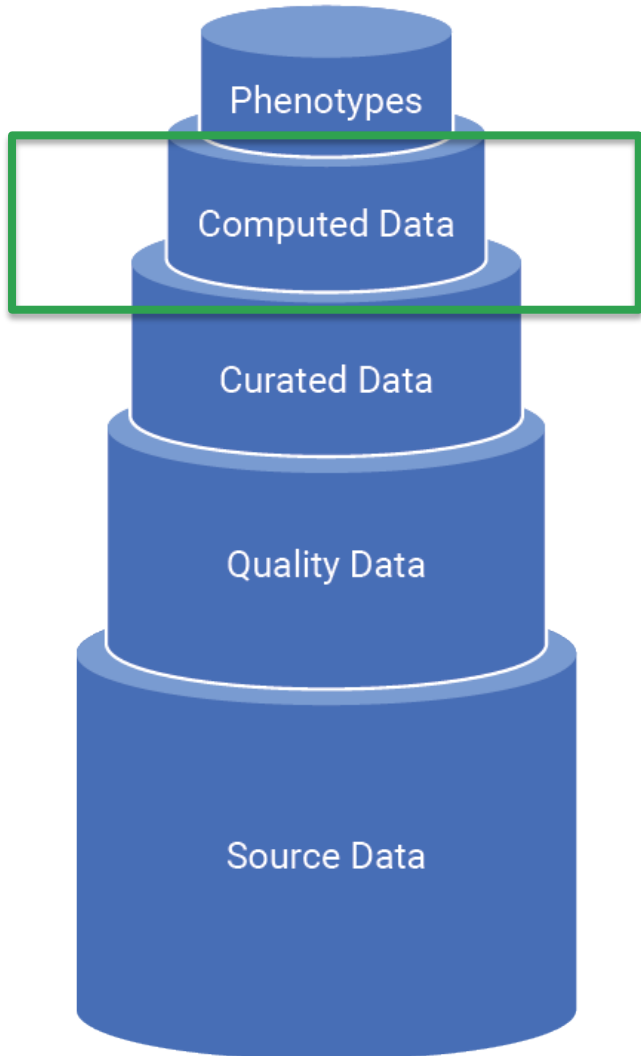
The Feature Stack: Curated Data



The Feature Stack: Curated Data



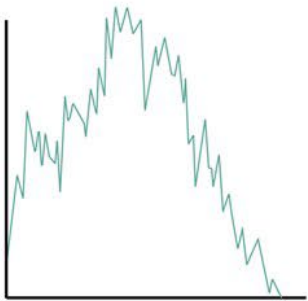
The Feature Stack: Computed Data



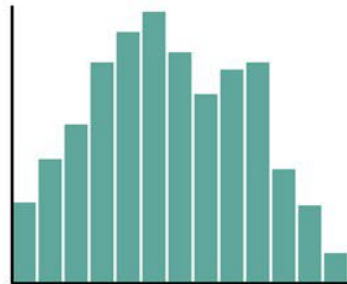
- Features extracted using computational methods including (but not limited to) AI.
- Computed features can be highly sensitive to underlying data distribution, so the preceding layers of the feature stack are critical.
- Computed features are analysis ready – leaving image space and into numeric space.

The Feature Stack: Computed Data

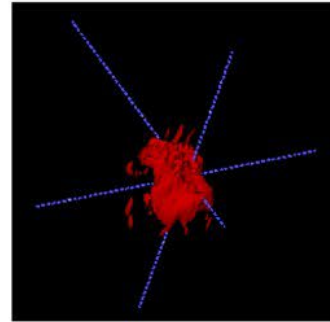
Intensity-Based
Features



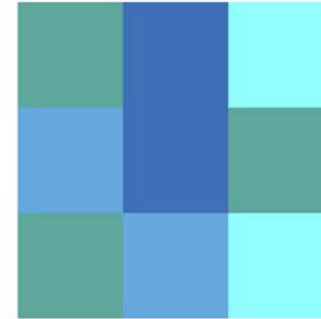
Histogram-Based
Features



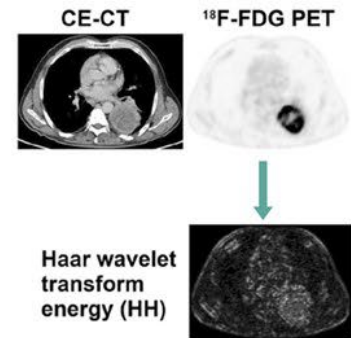
Volumetric &
Morphologic
Features



Textural
Features



Higher Order
Features

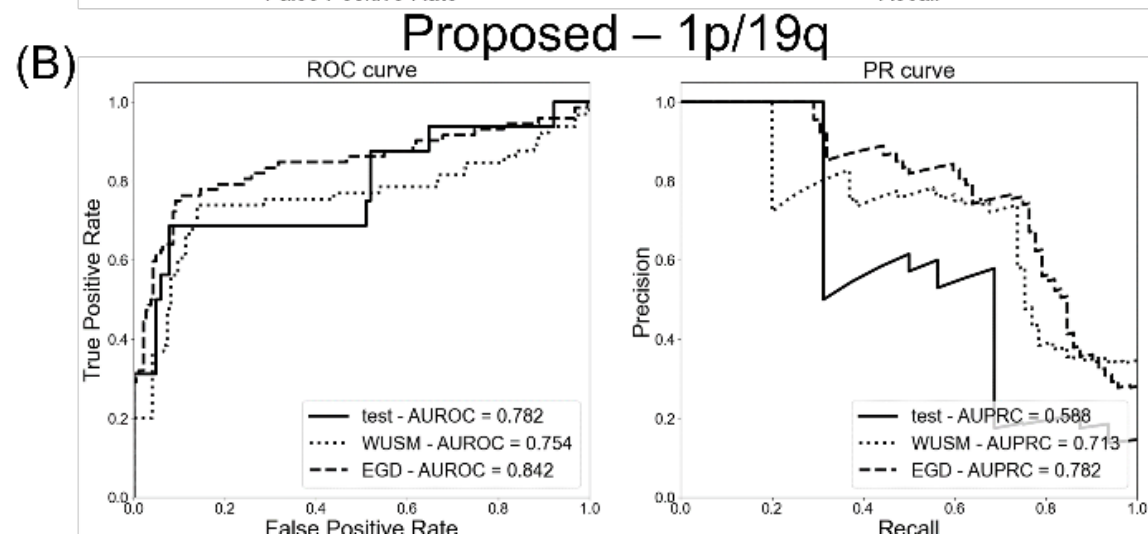
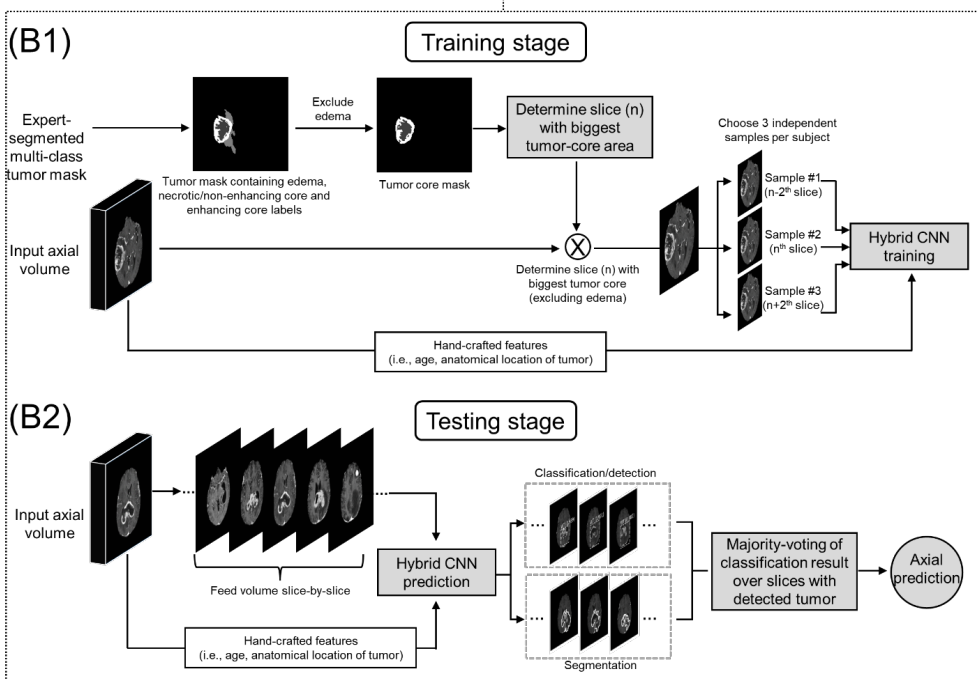
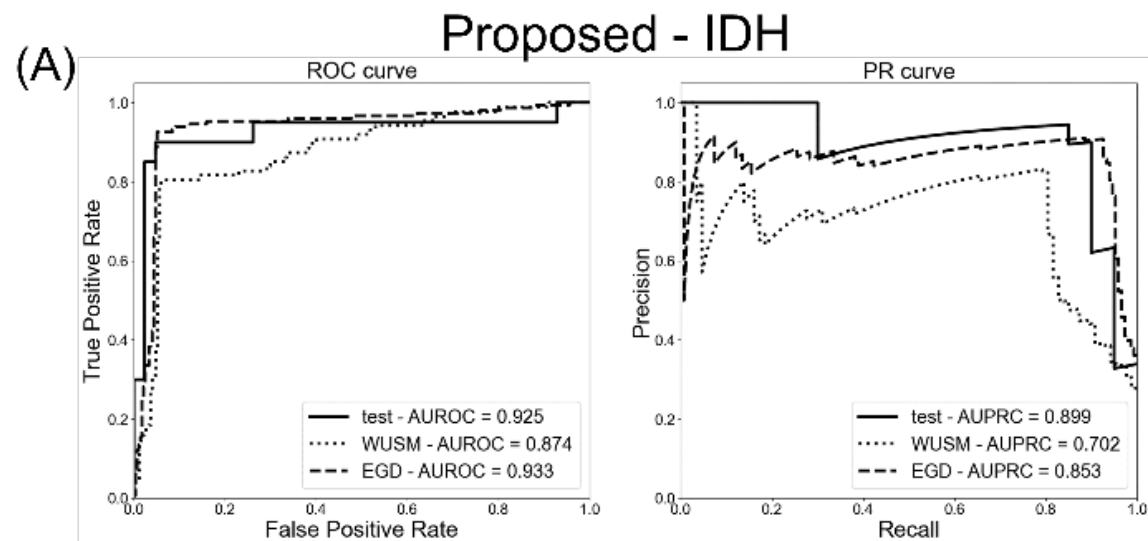
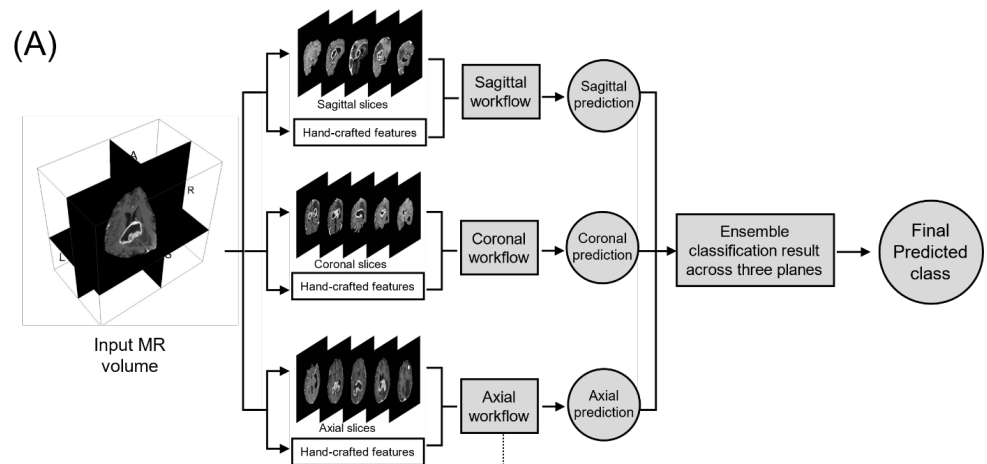


The Feature Stack: Phenotypes

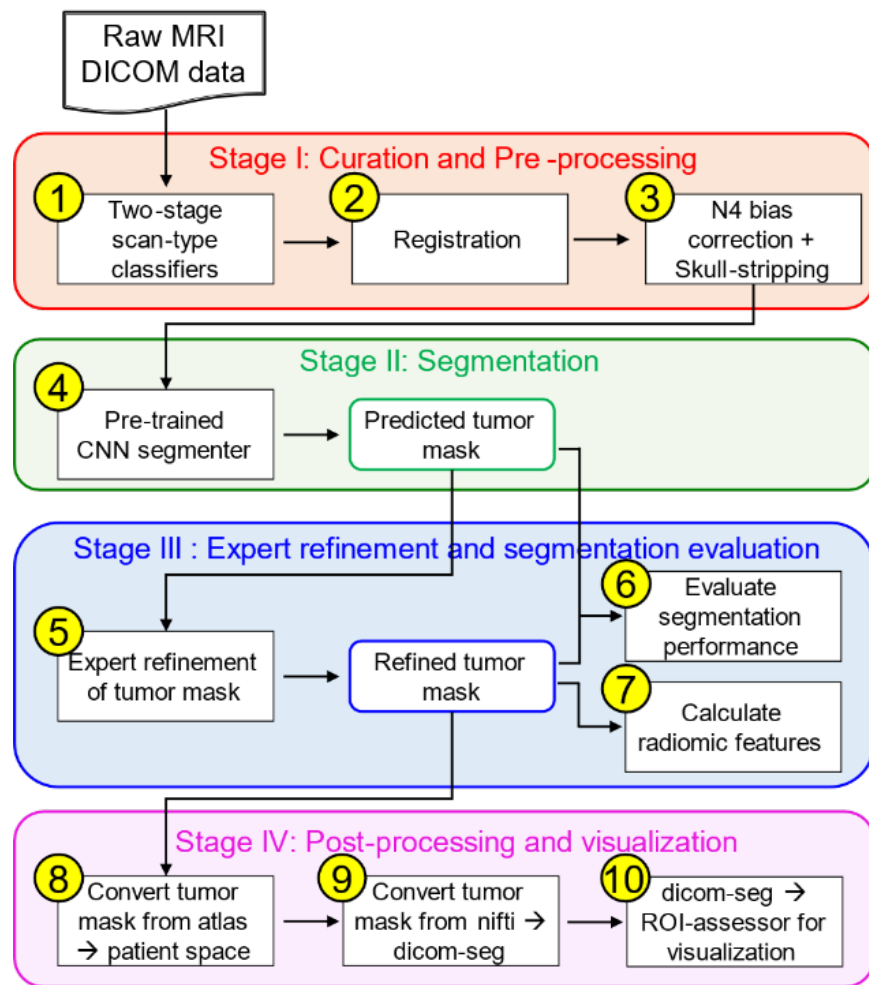


- Computed phenotypes align to known attributes of disease.
- Phenotypes may be alternatively obtainable through more invasive methods (e.g. biopsy).
- Useful phenotypes are predictive of outcomes.
- You might also call them biomarkers.

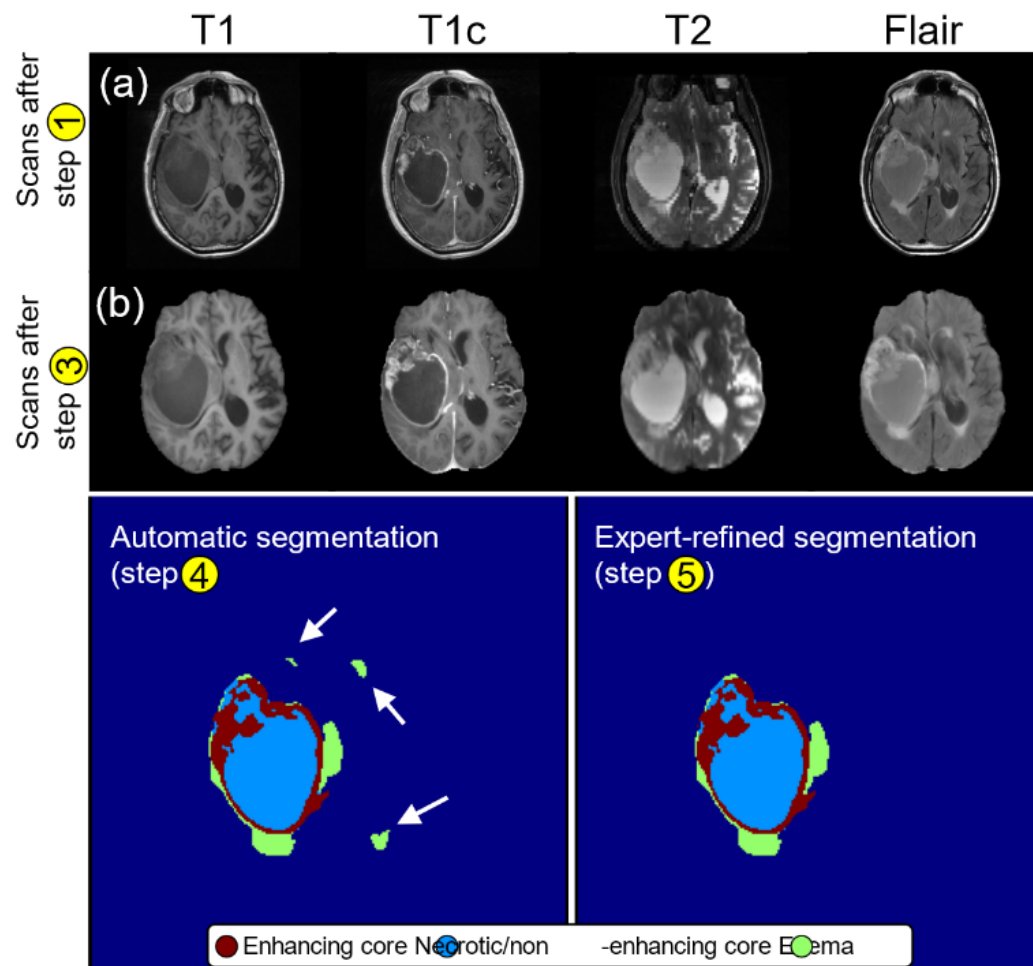
The Feature Stack: Phenotypes



The Feature Stack: Glioma – a full proof of concept

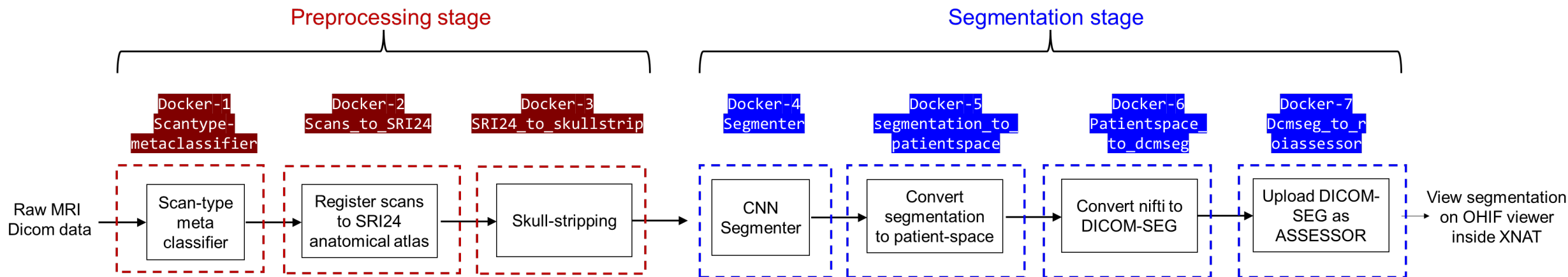


(A)



(B)

The Feature Stack: Glioma – a full proof of concept



Set up orchestration

Name:

Enter a name for the orchestration, this will display when a batch-launch initiates the orchestration.

Command: ✖

This command will determine the context (project, subject, session, etc) for the subsequent commands. Changing it may disable orchestration if previously-selected commands are no longer in context.

Command: ✖

Command: ✖

Set up orchestration

Name:

Enter a name for the orchestration, this will display when a batch-launch initiates the orchestration.

Command: ✖

This command will determine the context (project, subject, session, etc) for the subsequent commands. Changing it may disable orchestration if previously-selected commands are no longer in context.

Command: ✖

Command: ✖

Command: ✖

The Feature Stack: Glioma – a full proof of concept

PROJECT: HGG_seg > SUBJECT: TCGA-06-0175 > TCGA-06-0175

MR Session: TCGA-06-0175

Details Projects

Accession #: XNAT_E00010 Subject: TCGA-06-0175
Date Added: 2021-09-10 04:20:58 (admin) Gender:
Date: 2001-11-20 Handedness:
Scanner Name: MR01MROW Age: --
Scanner Type:

Actions

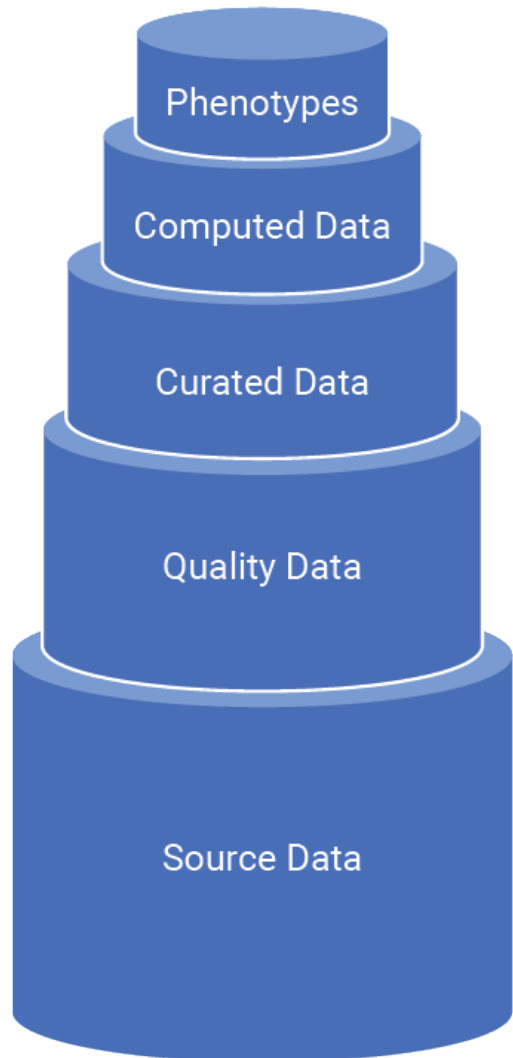
- Edit
- View ▶
- Download ▶
- Email
- Manage Files
- View Images
- Delete
- Run Containers ▶
 - Preprocessing_orchestrator
 - Segmentation_orchestrator

Scans

Bulk Actions: Download

<input type="checkbox"/>	Scan	Type	Series Desc	Usability	Files	Note
<input type="checkbox"/>	1	3 PLANE LOC	3 PLANE LOC	usable	1.1 MB in 9 files	
<input type="checkbox"/>	2	SAG T1	SAG T1	usable	2.3 MB in 18 files	
<input type="checkbox"/>	3	SAG T1	SAG T1	usable	2.3 MB in 18 files	
<input type="checkbox"/>	4	AXIAL FSE	AXIAL FSE	usable	15.7 MB in 124 files	
<input type="checkbox"/>	5	AXIAL FLAIR	AXIAL FLAIR	usable	7.9 MB in 62 files	
<input type="checkbox"/>	6	AXIAL DIFFUSION	AXIAL DIFFUSION	usable	15.0 MB in 118 files	

Conclusions



- Many relevant AI methods... transformers, classifiers, segmentation, clustering, etc.
- Attending to all levels of the feature stack supports reliable, trustworthy AI.
- Scalable informatics tools like XNAT enable development, validation, and deployment of AI-powered data pipelines.

Thank you

