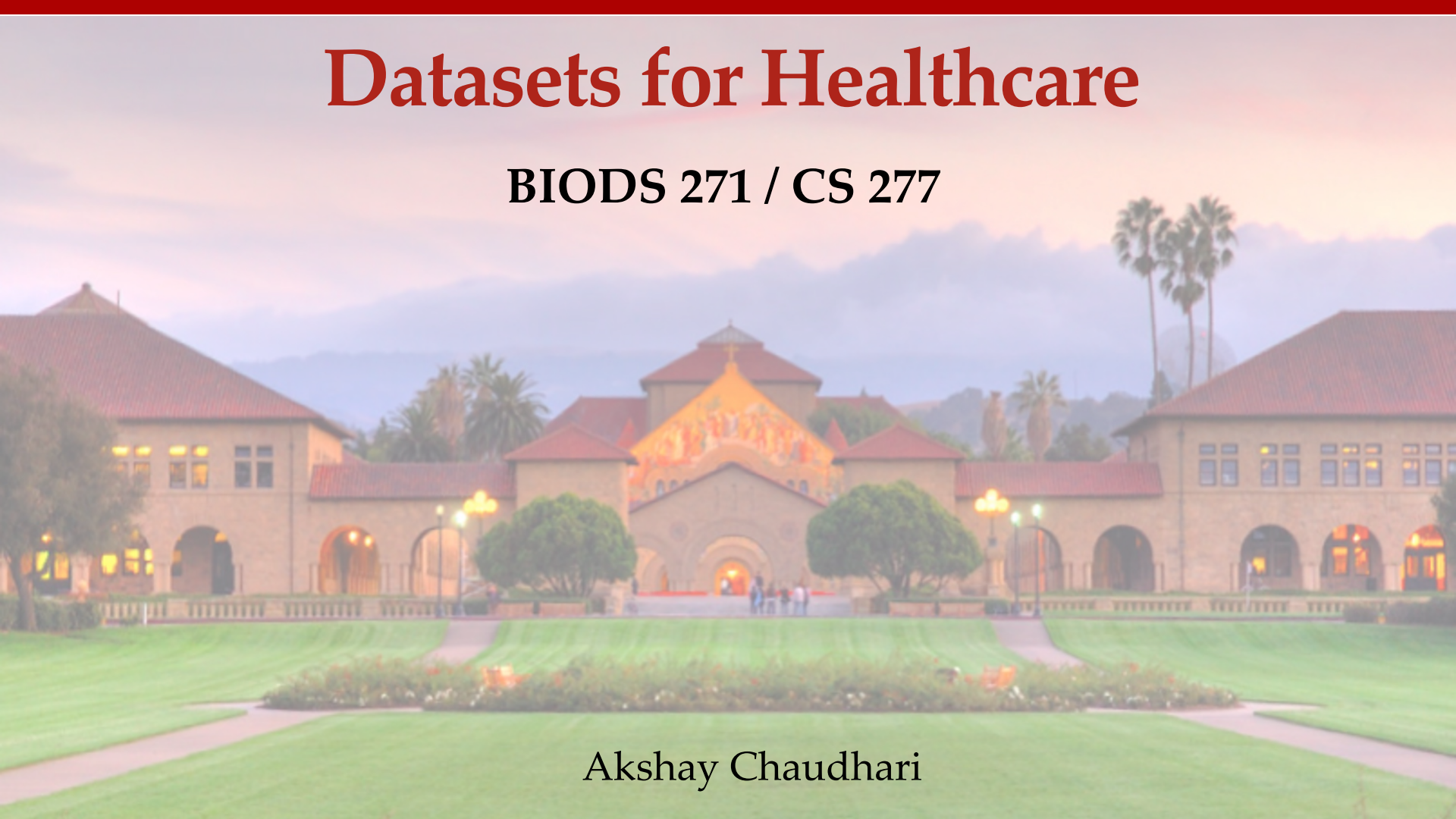
Datasets for Healthcare

BIODS 271 / CS 277

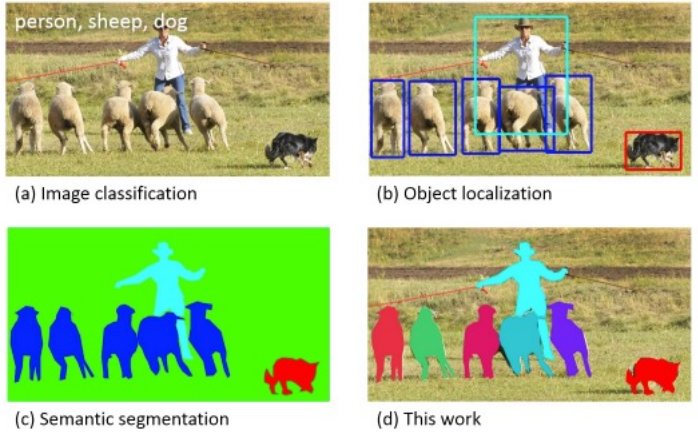


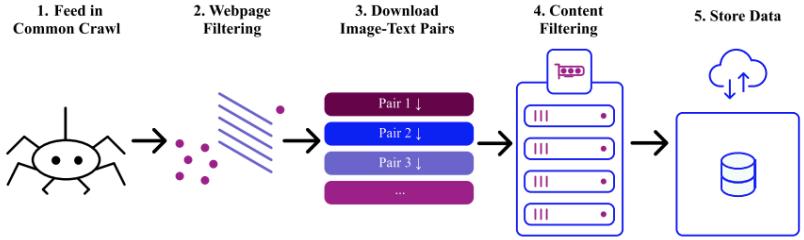
Akshay Chaudhari

Common ML Datasets

Microsoft COCO: Common Objects in Context

Tsung-Yi Lin Michael Maire Serge Belongie Lubomir Bourdev Ross Girshick
James Hays Pietro Perona Deva Ramanan C. Lawrence Zitnick Piotr Dollár

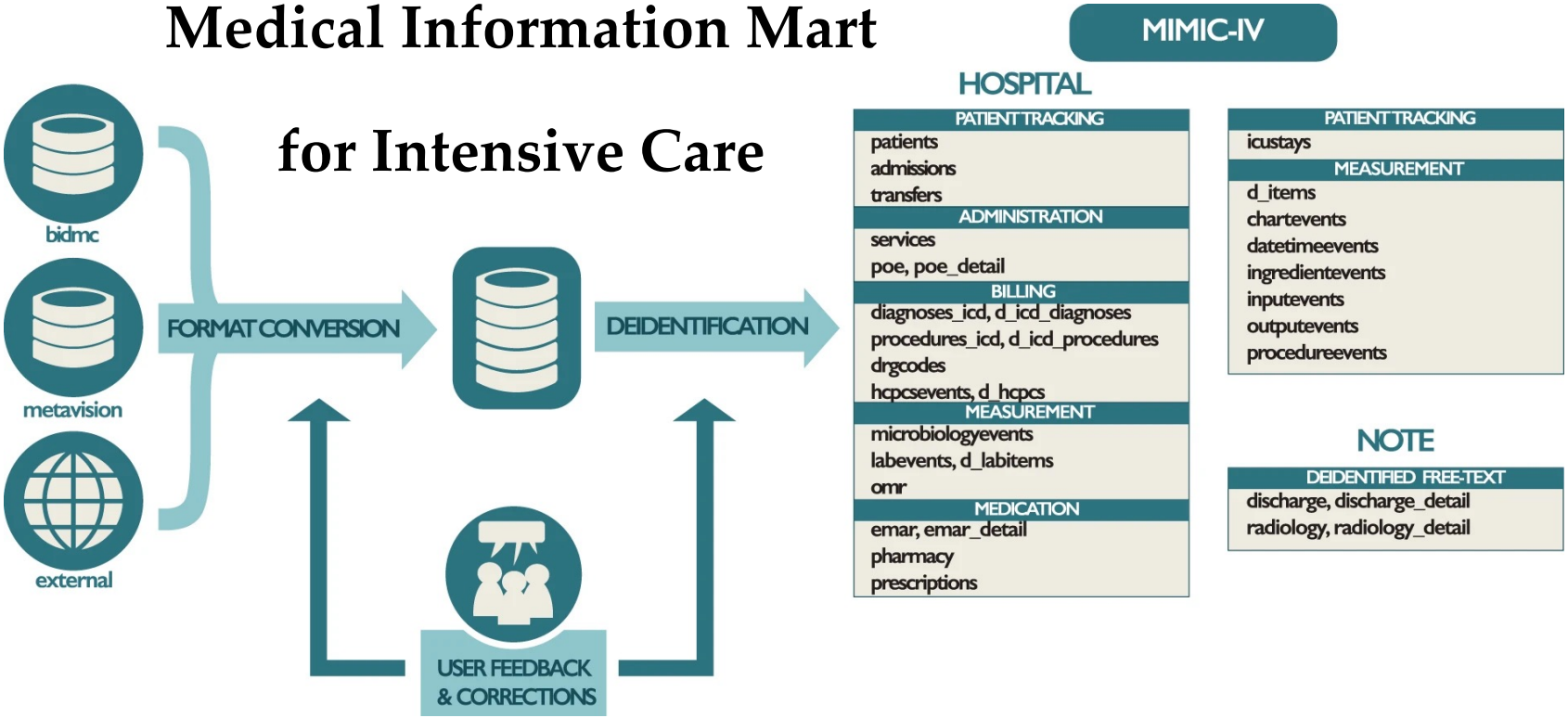




LAION-5B: An open large-scale dataset for training next generation image-text models

COMMON CRAWL

MIMIC Dataset



[1] Johnson et al. MIMIC-IV, a freely accessible electronic health record dataset. Nature Scientific Data 2023.

Types of Studies Hailing from MIMIC

- Automatic sepsis detection using vital signs
- Chest x-ray classification + radiology report analysis
- Automatic documentation generation
- Coding assignment
- …

UK Biobank

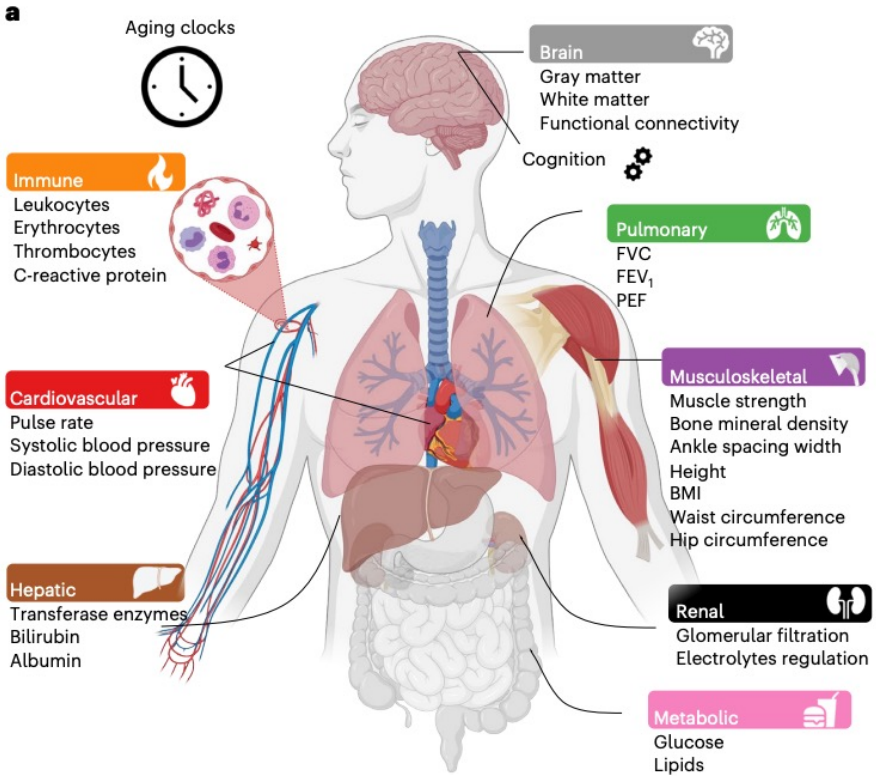
uk biobank
Enabling scientific discoveries that improve human health

- Imaging: Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- Genetics: Whole genome sequencing for all 500,000 participants, whole exome sequencing for 470,000 participants, genotyping (800,000 genome-wide variants and imputation to 90 million variants).
- Health linkages: Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.
- Biomarkers: Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.
- Activity monitor: Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.

Types of Studies Hailing from UK Biobank

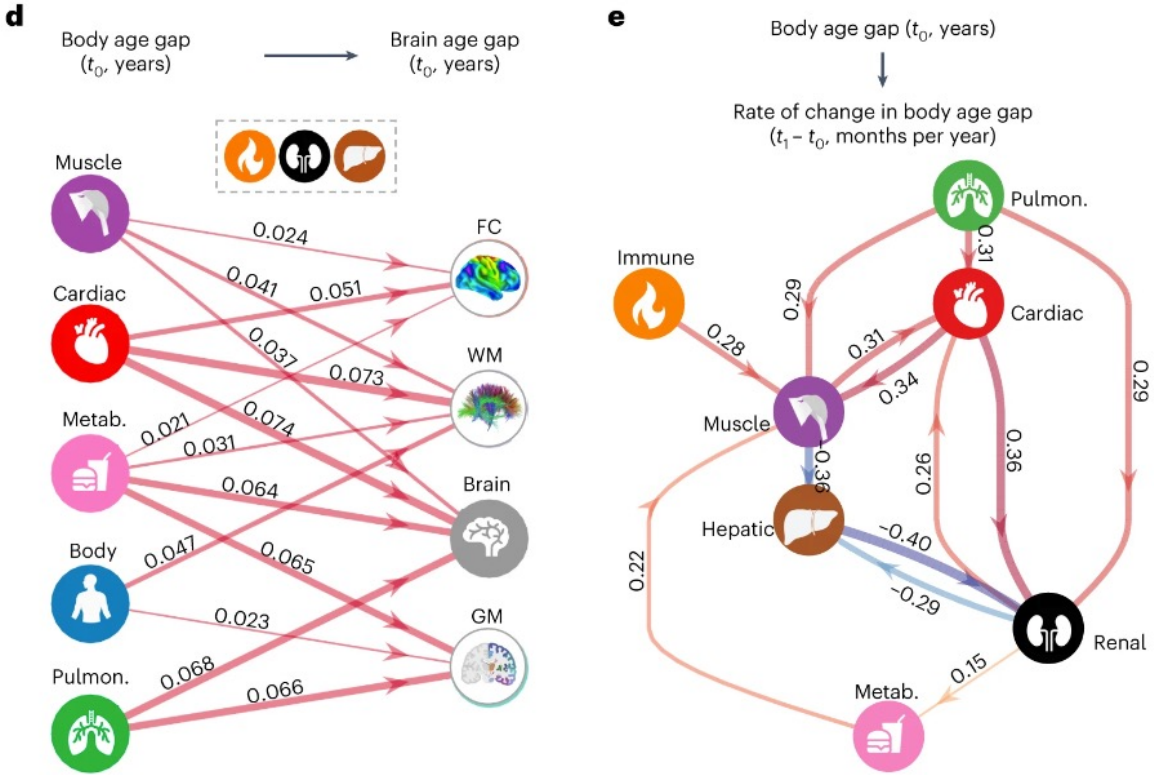
- Normative population assessment
- Genome/phenome-wide association studies
- Medical imaging segmentation
- New biomarker discovery

Biological vs Chronological Aging



[1] Tian et al. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. Nature Medicine 2023.

Biological vs Chronological Aging



[1] Tian et al. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. Nature Medicine 2023.

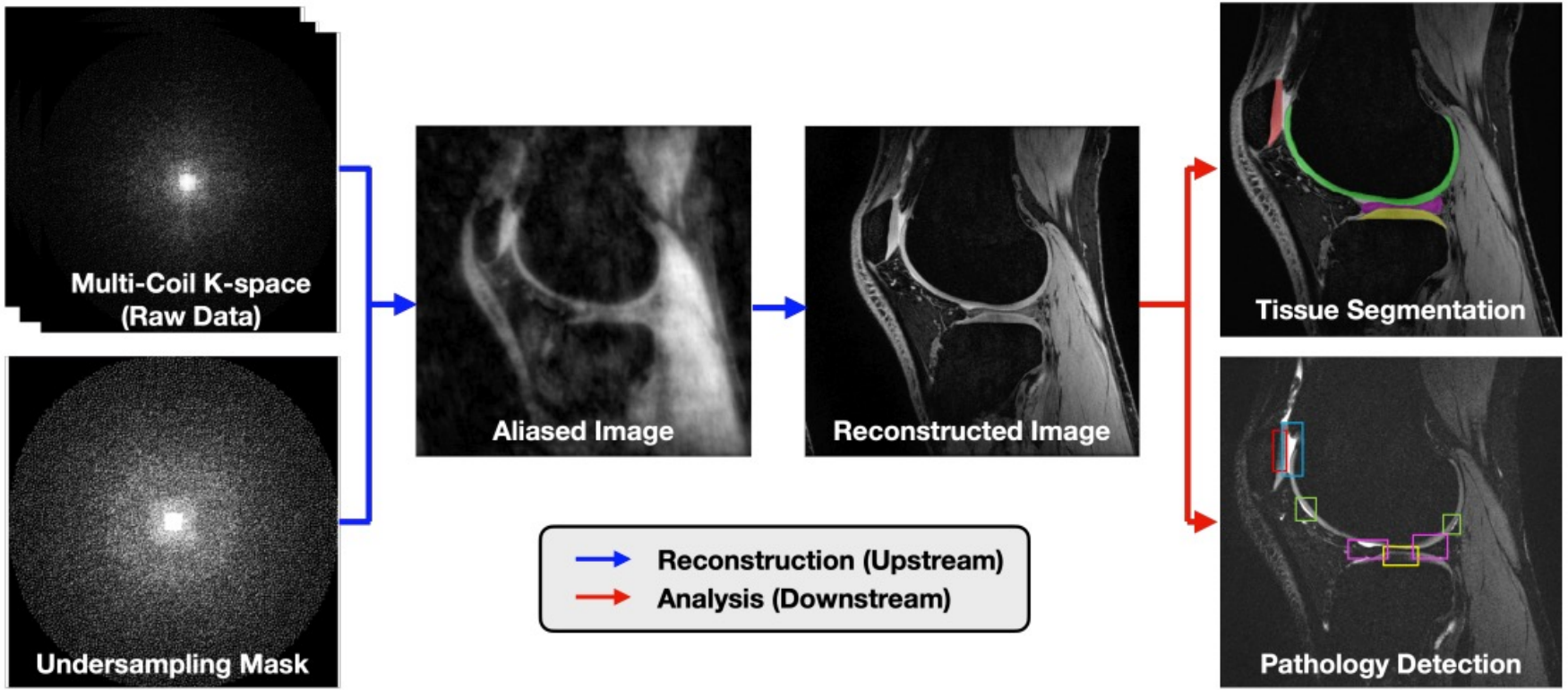
General Guidelines for Clinical Needs

- Why does your innovation **matter**?
- **Who benefits** from your benefit? Can you quantify it?
- How does your innovation get **implemented**?
- How does your innovation get used in a **care pathway**?
- Do all patients **equally benefit** from your innovation?
- How can we **scale/deploy** your innovation *broadly*?

What Makes a Good Dataset?

- Availability/Access
- Quality of data – inputs + outputs
- Completeness of data
- Fit within ecosystem + societal needs
- Usability of data
- Addressing clinical needs

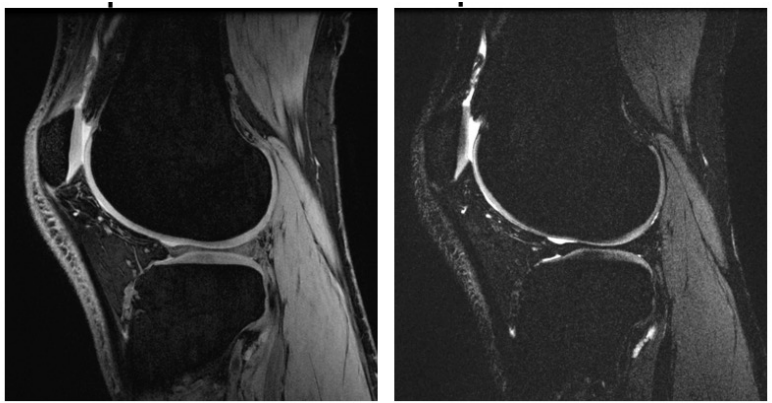
Example Dataset: SKM-TEA



[1] Desai et al. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. NeurIPS D&B 2022.

Example Metrics for Reconstruction

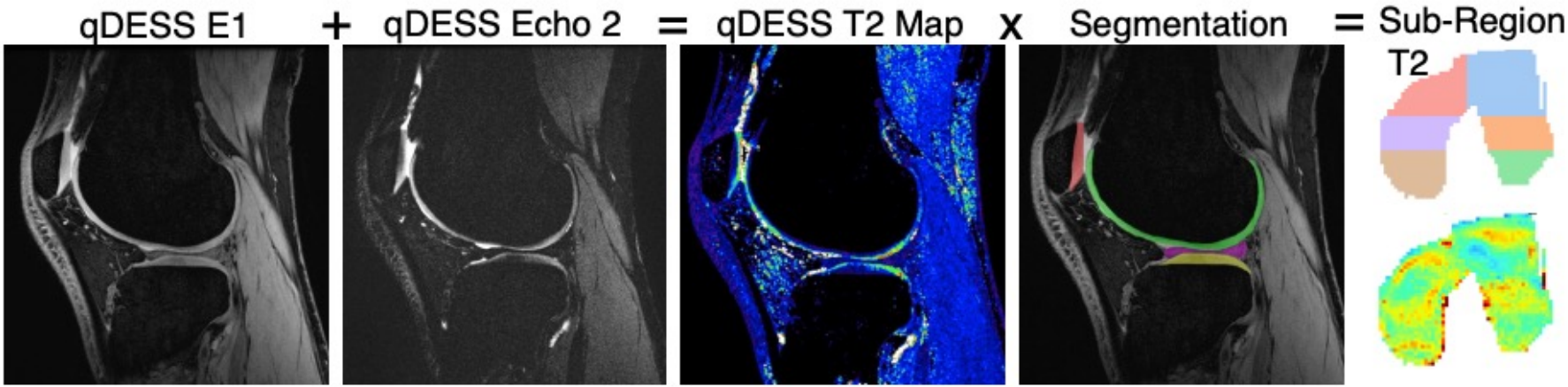
Acc.	Metric / Model	pSNR (dB) E1	pSNR (dB) E2	SSIM E1	SSIM E2
6x	U-Net (E1/E2)	31.5 (1.38)	33.7 (1.02)	0.77 (0.027)	0.73 (0.032)
	U-Net (E1+E2)	31.1 (1.38)	33.2 (1.05)	0.77 (0.024)	0.74 (0.030)
	U-Net (E1⊕E2)	31.1 (1.63)	33.5 (1.02)	0.76 (0.026)	0.73 (0.034)
	Unrolled (E1/E2)	**35.0 (1.08)**	34.5 (1.09)	0.83 (0.024)	0.76 (0.031)
	Unrolled (E1+E2)	35.0 (1.07)	**34.5 (1.09)**	**0.84 (0.022)**	**0.76 (0.030)**
	Unrolled (E1⊕E2)	35.0 (1.08)	34.2 (1.08)	0.83 (0.023)	0.76 (0.030)



[1] Desai et al. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. NeurIPS D&B 2022.

Downstream Clinical Metrics

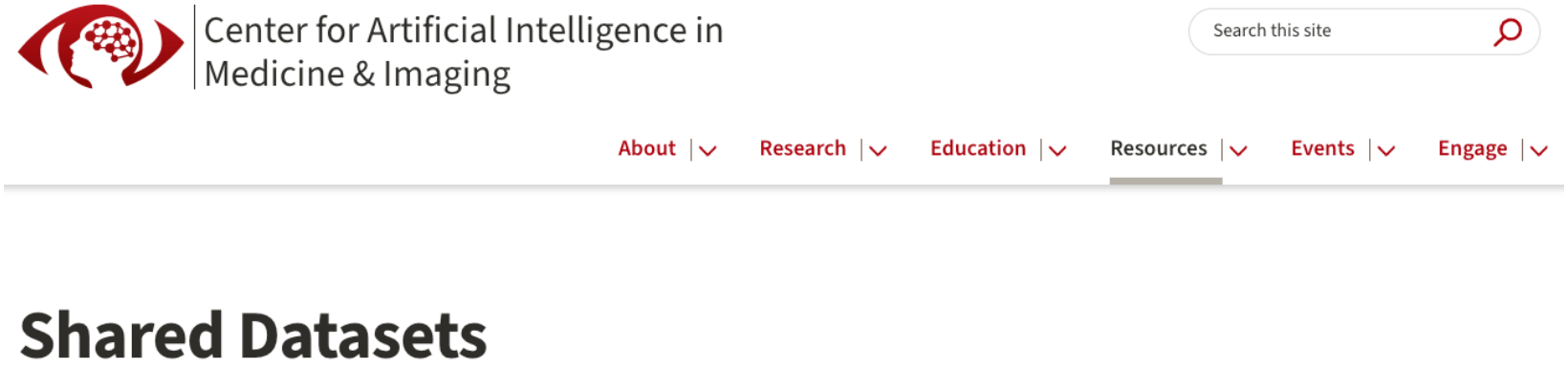
- Assessing what the information is used for



[1] Desai et al. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. NeurIPS D&B 2022.

Sharing Data at Stanford

- Data sharing as a secondary use of data while ensuring adequate privacy safeguards



Questions?

akshaysc@stanford.edu