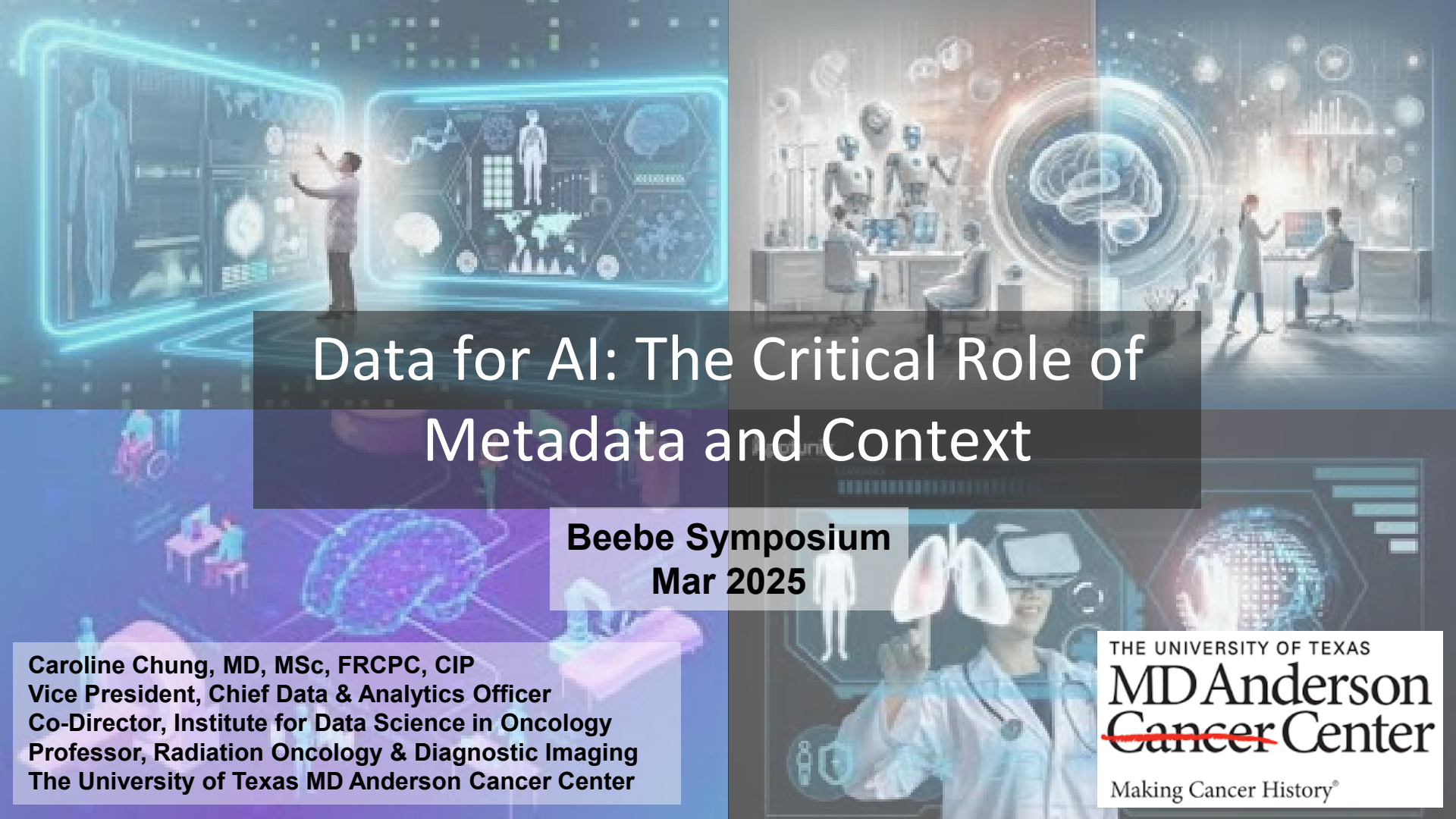




Data for AI: The Critical Role of Metadata and Context

Beebe Symposium
Mar 2025

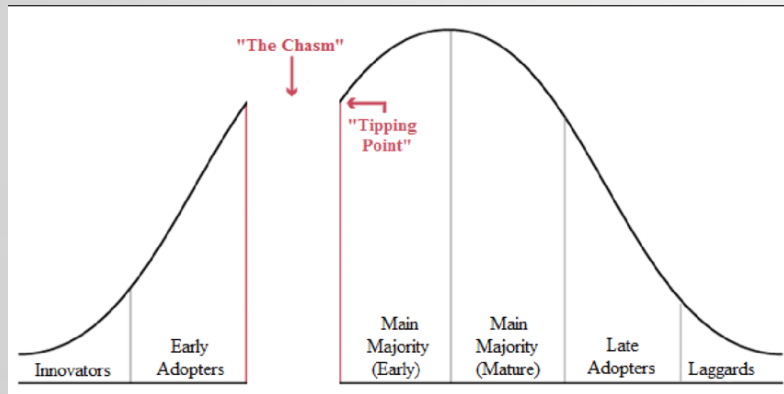
Caroline Chung, MD, MSc, FRCPC, CIP
Vice President, Chief Data & Analytics Officer
Co-Director, Institute for Data Science in Oncology
Professor, Radiation Oncology & Diagnostic Imaging
The University of Texas MD Anderson Cancer Center

THE UNIVERSITY OF TEXAS
MD Anderson
Cancer Center
Making Cancer History®

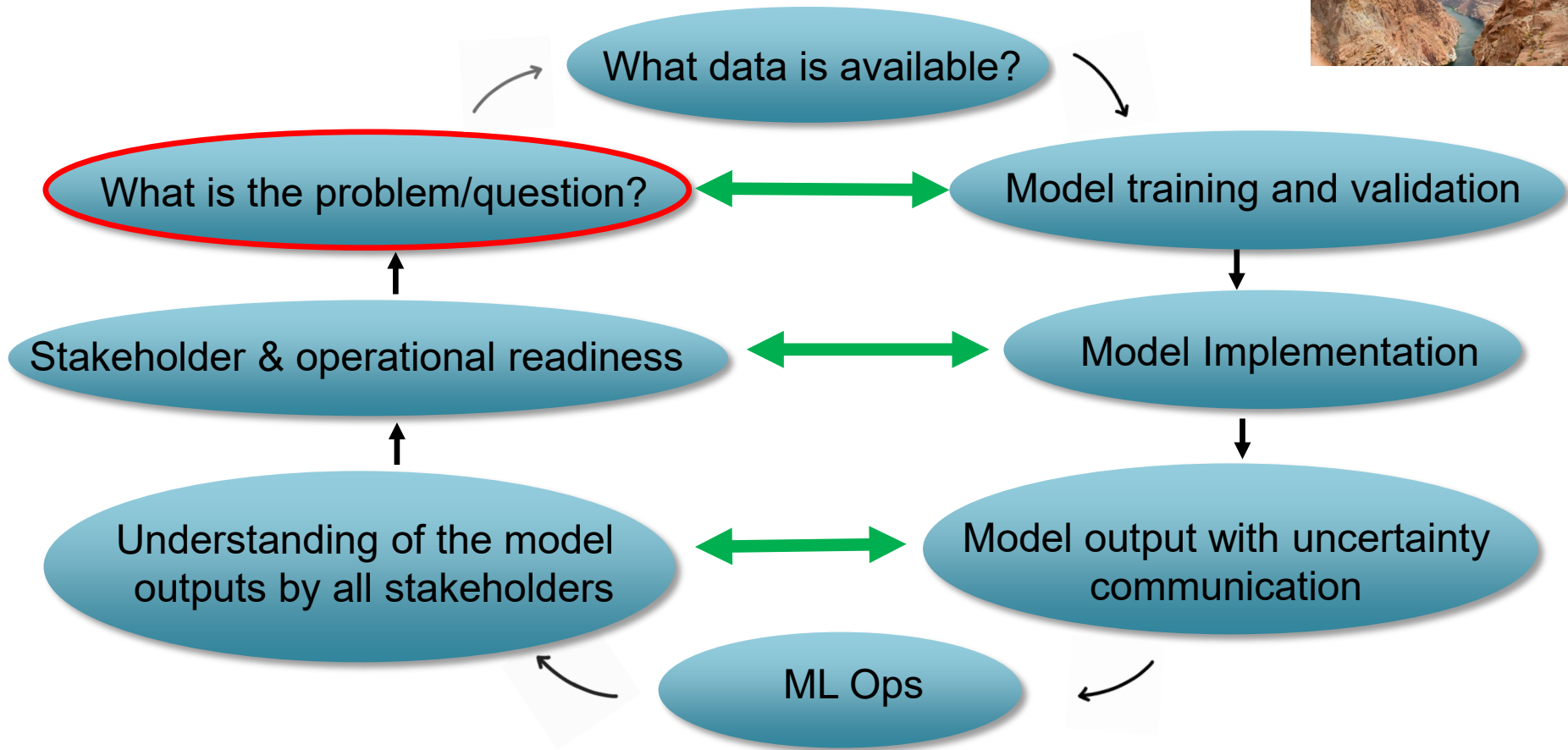
Context is a critical bridge to realize the impact and value of applying AI, models and data science tools into real-world settings.

“THERE IS A LACK OF AWARENESS OF THE SO-CALLED ‘AI CHASM’, THAT IS THE GULF BETWEEN DEVELOPING A SCIENTIFICALLY SOUND ALGORITHM AND ITS USE IN ANY MEANINGFUL REAL-WORLD APPLICATIONS.”

P. Keane & E. Topol – npj Digital Medicine 2018

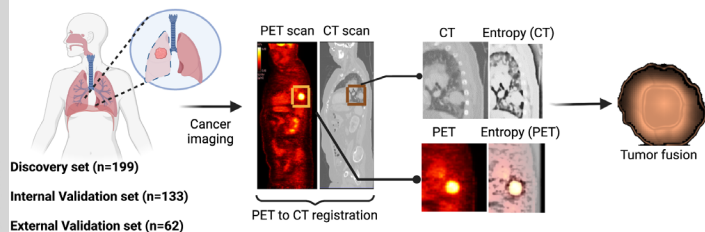


Considering Context Throughout the Model Lifecycle

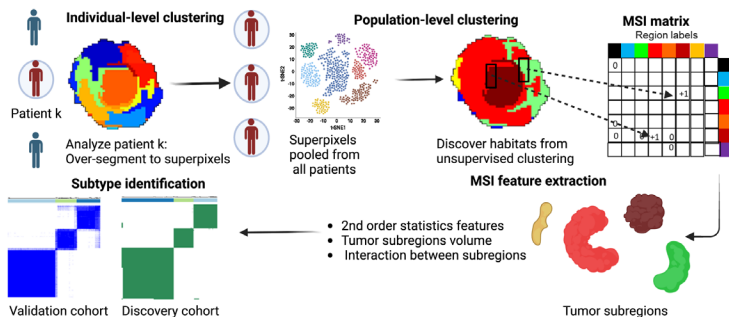


Imaging is a Beacon for Personalizing Cancer Treatment

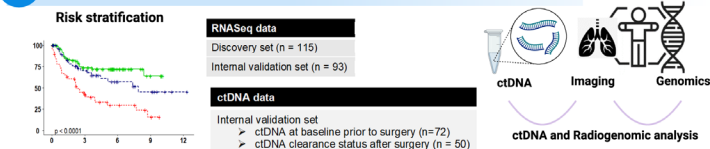
a NSCLC patient data acquisition and preprocessing



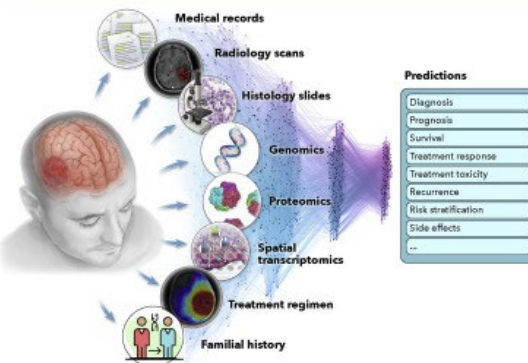
b Habitat image analysis



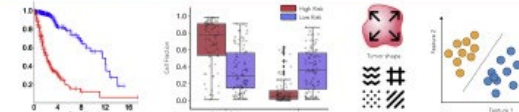
c Clinical and biological value of imaging subtypes



A Multimodal Data Fusion

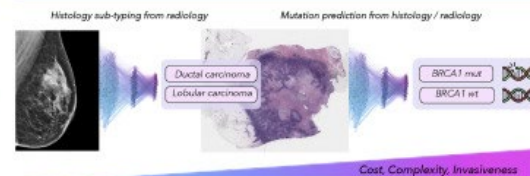


D Quantitative Analysis

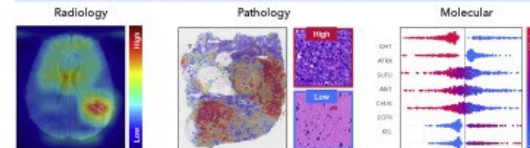


Lipkova – Cancer Cell 2022

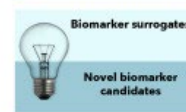
B Multimodal Data Interconnection



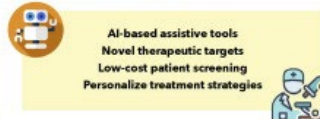
C Multimodal Interpretability & Association Discovery



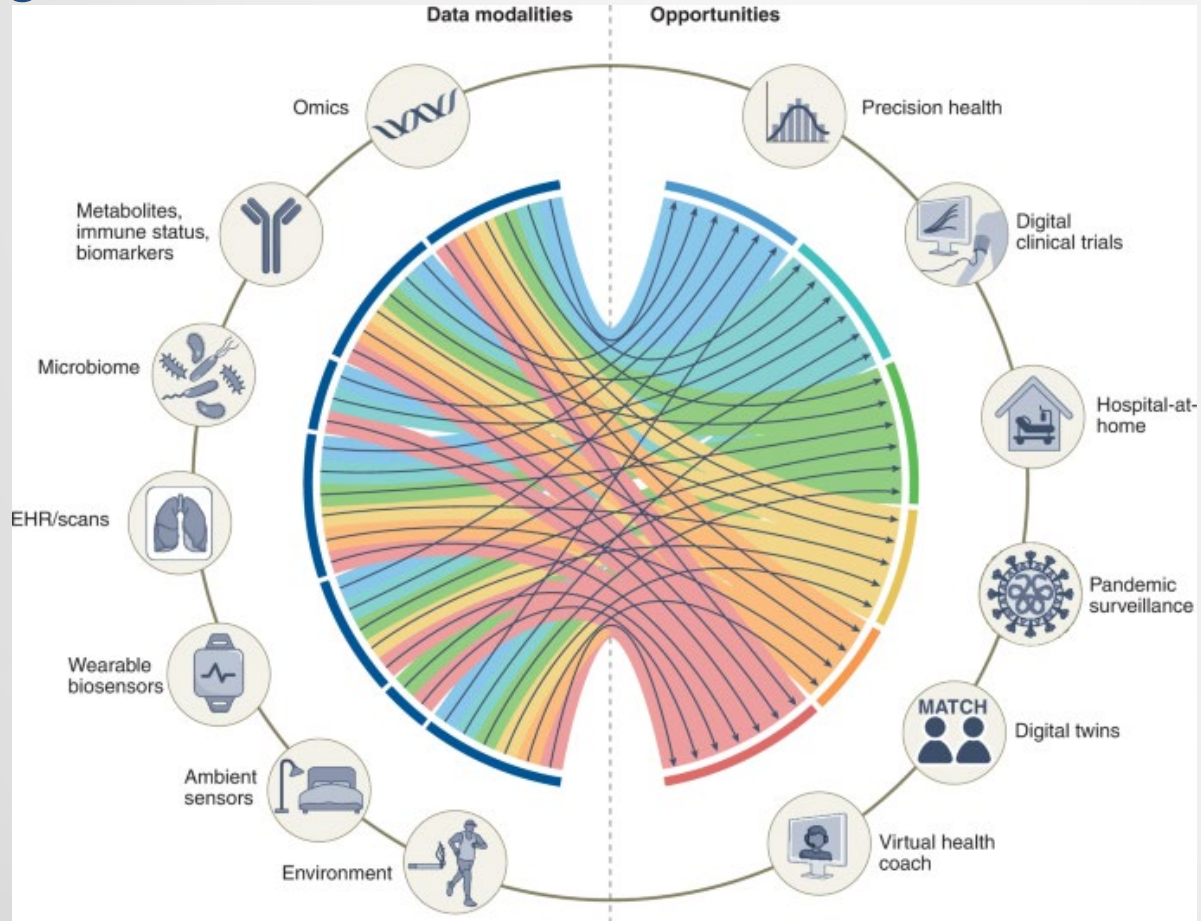
E Biomarker Exploration



F Translation to Practice



Broad Range of Data Domains for Multimodal Biomedical AI



Team Data Science Principles: Grounding Our Approach with **'Data in Context'**



Observations in context



Dynamic assessment of quality based on context



Provenance linking insights to observations to build confidence/trust and attribution for data and insight contributions



Integrated data governance - ensure appropriate data access and attribution based on role & intended use

The underlying glue to bring together these principles into action is METADATA

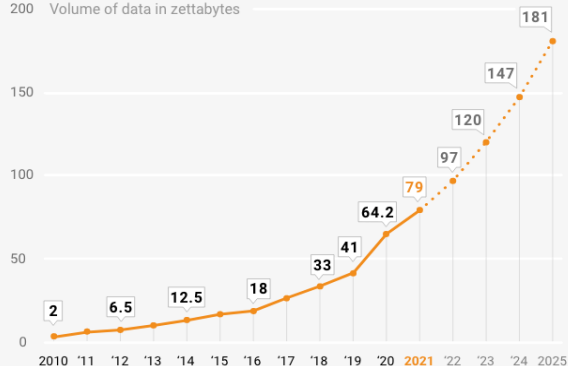
Finding signal from the noise of exponential growth of health data

Volume of data created, captured, copied, and consumed worldwide



The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025

Volume of data in zettabytes



Source: statista.com



firstsiteguide.com

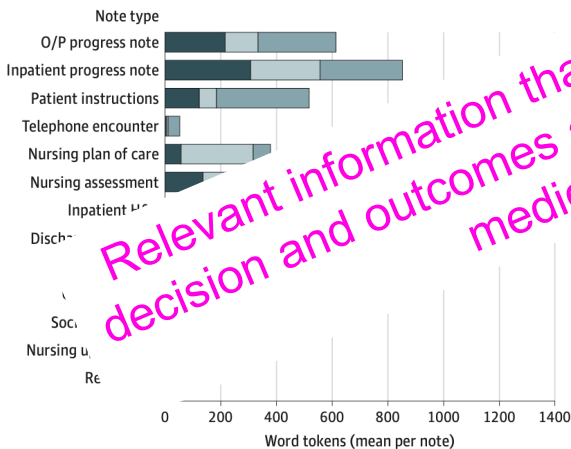


Original Investigation | Health Informatics

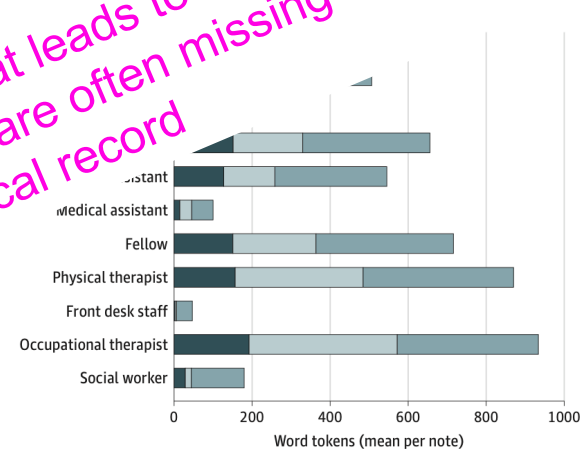
Prevalence and Sources of Duplicate Information in the Electronic Medical Record

Jackson Steinkamp, MD; Jacob J. Kantrowitz, MD, PhD; Subha Airan-Javia, MD

A Mean duplicate text per note by note type



B Mean duplicate text per note by provider type



Relevant information that leads to the treatment decision and outcomes are often missing from the medical record

Impact of Wide Variability in Data

e.g. Standard of Care Imaging

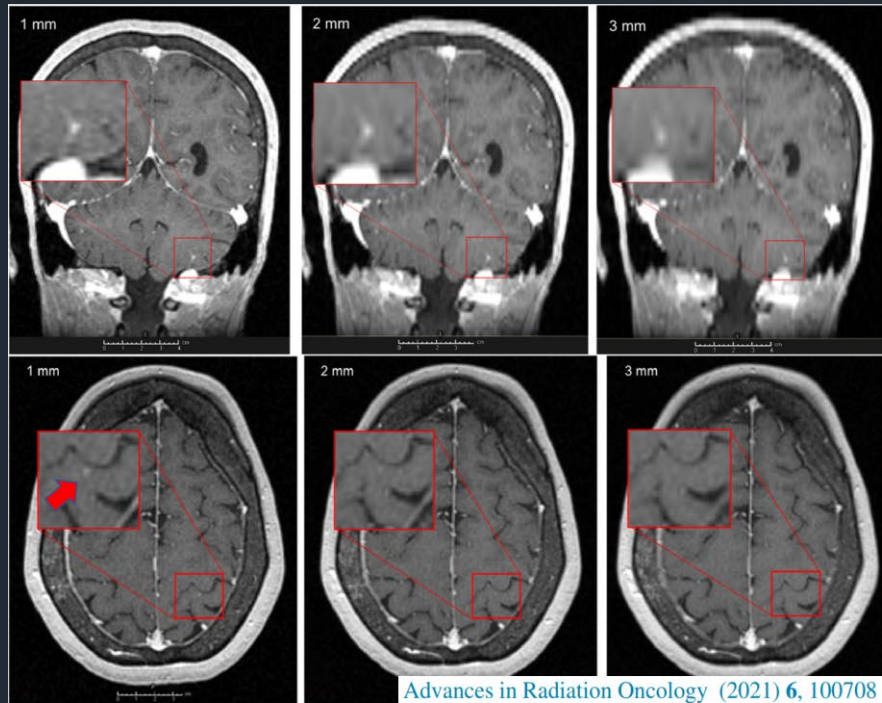
EMERGING TECHNOLOGY DAY ONE PROJECT

A Quantitative Imaging Infrastructure To Revolutionize AI-Enabled Precision Medicine

12.12.24 | 9 MIN READ | TEXT BY WILLIAM HOLLANDER & CALLIE WEIANT & CAROLINE CHUNG & NOLA HYLTON & MATTHEW ROSEN

<https://fas.org/publication/ai-enabled-precision-medicine/>

Impact of Focusing on Data without Context



Advances in Radiation Oncology (2021) 6, 100708

Thrower et al. – Advances in Radiation Oncology (2021)



Hundreds of AI tools have been built to catch covid. None of them helped. | MIT Technology Review

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better. When covid-19 struck Europe in March 2020, hospitals were plunged into a ...

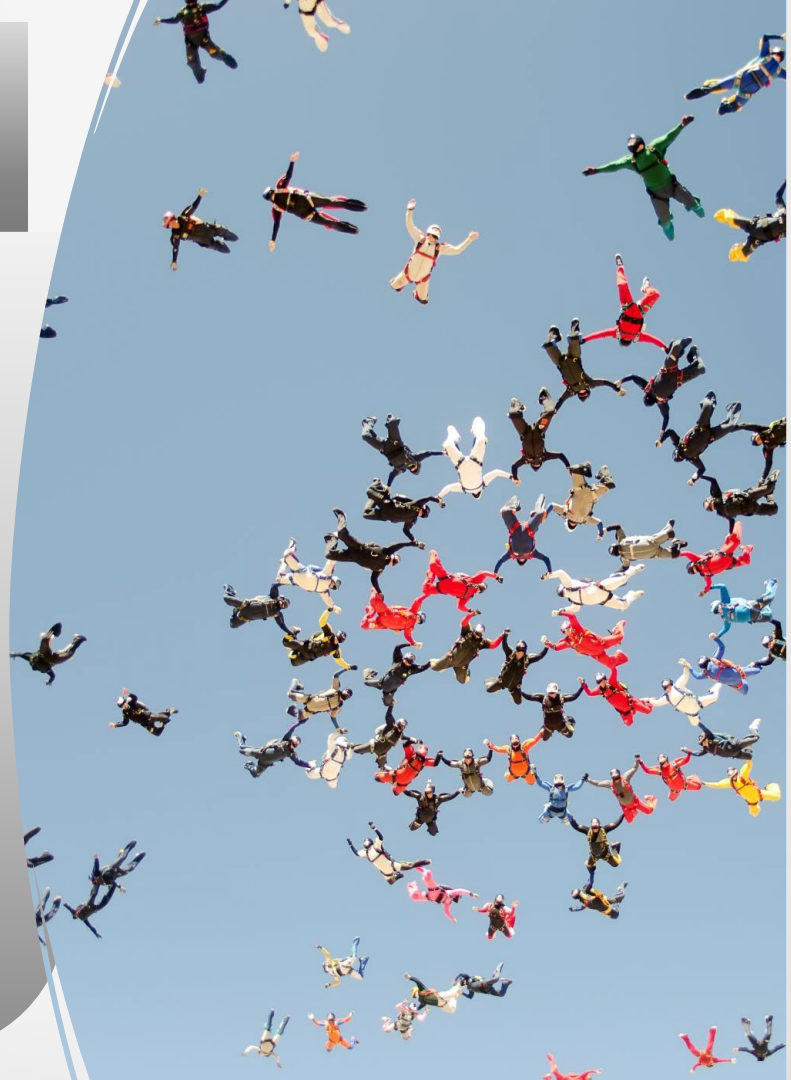
www.technologyreview.com

Building Meaningful Training, Validation, Test Data Sets

Integrate & consider metadata because 'CONTEXT MATTERS' during model development, validation and deployment.

Some CONTEXT that may introduce bias:

- **Cohort selection**
 - Age
 - Race/ethnicity
 - SES
 - Geographic location – environmental, diet
 - Comorbidities – impact of other conditions or related medications on the observations/outcomes
 - Classification/subtype/stage of disease
- **Availability of data** — frequency, timing & types of data could be connected to geography, insurance status, age, etc.
- **Measurement error bias** — scanner, protocol, patient cooperation (e.g. motion, sedation)



Summary

- When tackling complex questions like cancer, we can benefit from ***empowering the complexity to deepen our understanding.***
- We need to consider **Data (content) + Metadata (context)** as we generate, curate, and utilize data to ensure meaningful insights are generated.
- Starting with a clearly defined purpose (or problem that needs to be solved) is critical to building the appropriate team, bringing together the best fit data, and considering the context of development and implementation to drive to impact.
- Verification, validation and uncertainty quantification (considering the data and the model) are critical for building trust/confidence and informing decision-making.