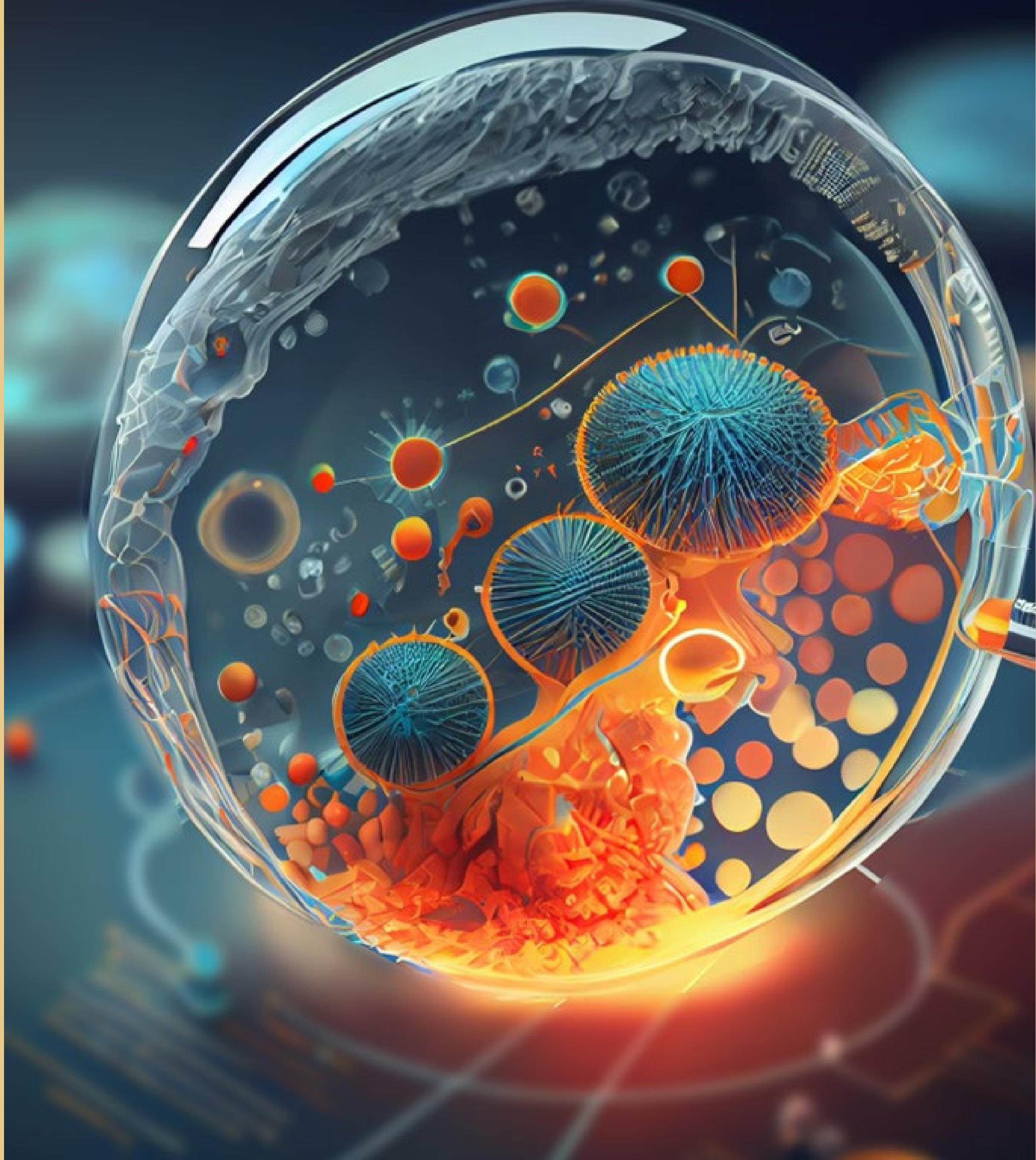
What is the bridge between functional and genomic precision medicine?

Potential role of AI



AI For Precision Medicine | 27th September 2024

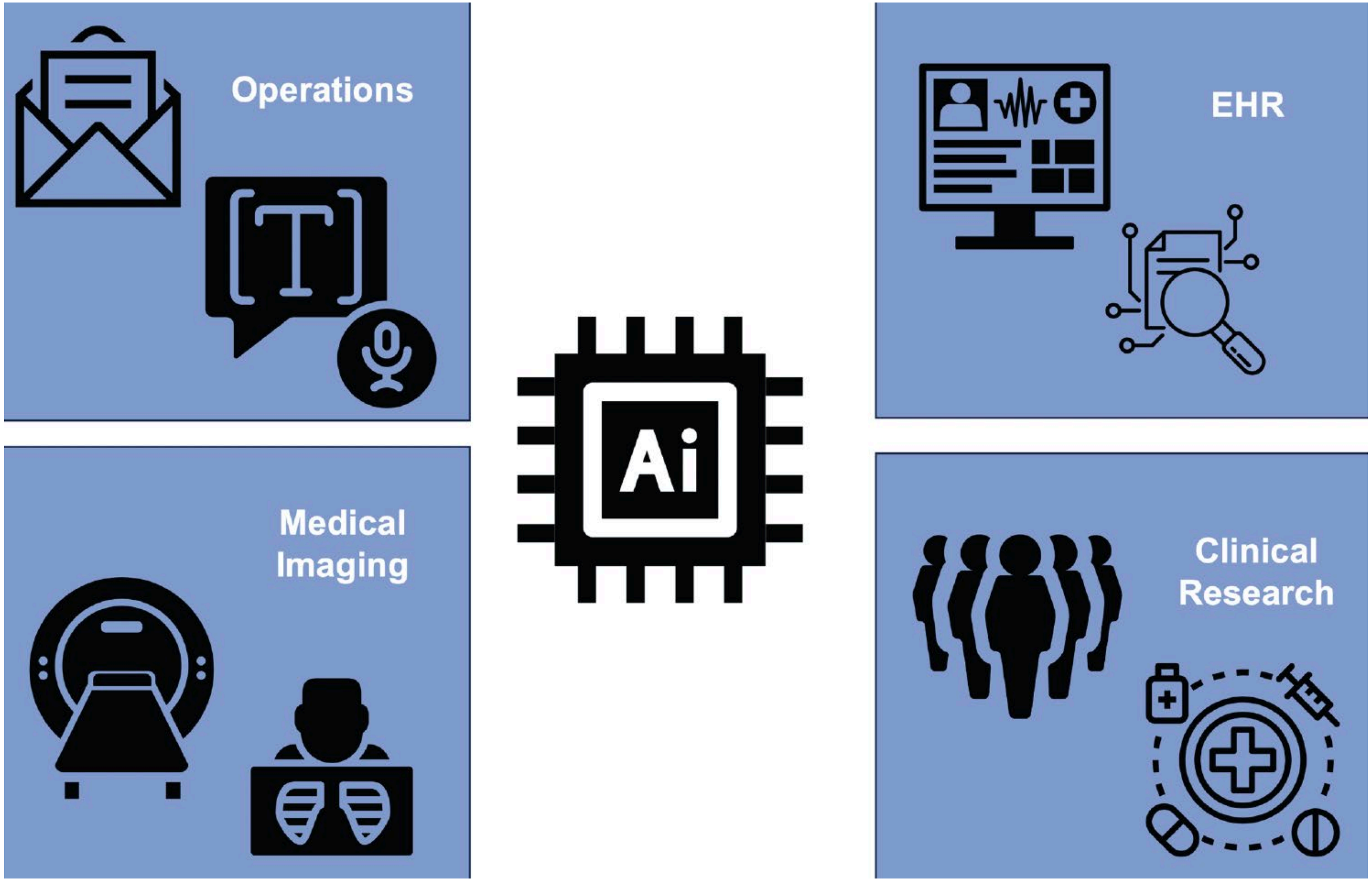
Shannon McWeeney | mcweeney@ohsu.edu

Bridge to the Future

Artificial
Augmented
Actionable

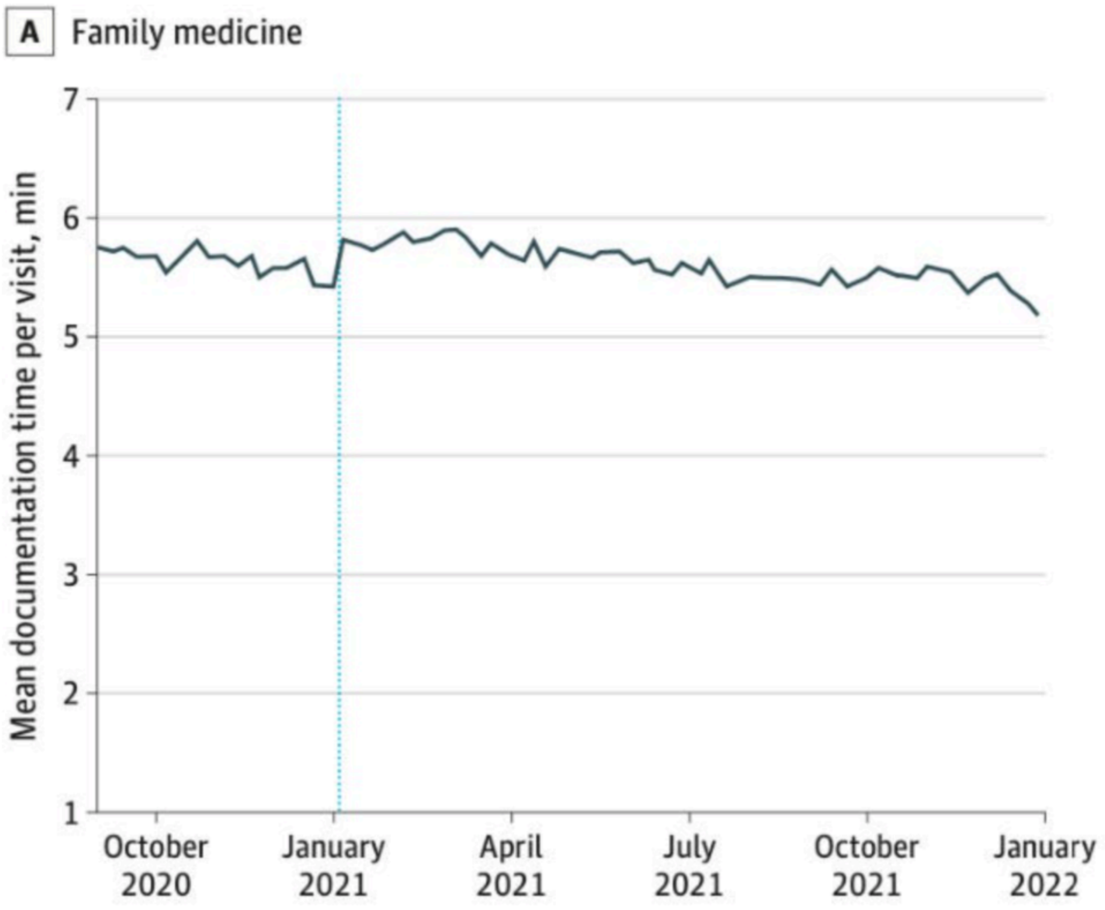
} Intelligence

How soon is now?



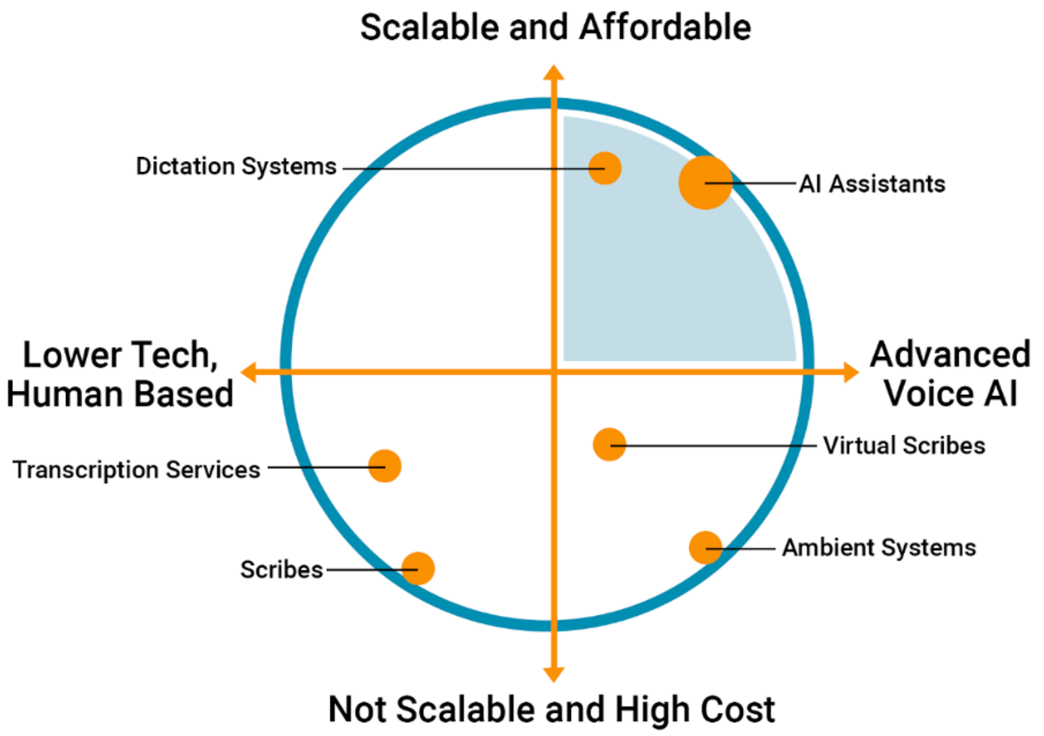
Bottomly and McWeeney (2024) JITC

Clinical Transformation: Doctors ≠ "Data Clerks"



JAMA Health Forum. 2023;4(5):e230984

Jan 2021: Centers for Medicare & Medicaid Services Guideline Changes



Source: American Academy of Family Physicians

79% better documentation quality
70% reduced burnout / fatigue
81% greater physician focus

Bridging the Gap

- Comprehensive Biological insights
- Enhanced Predictive Modeling and Decision Support
- Identification of Biomarkers and Therapeutic Targets
- Address data complexity

Multimodal Data Integration

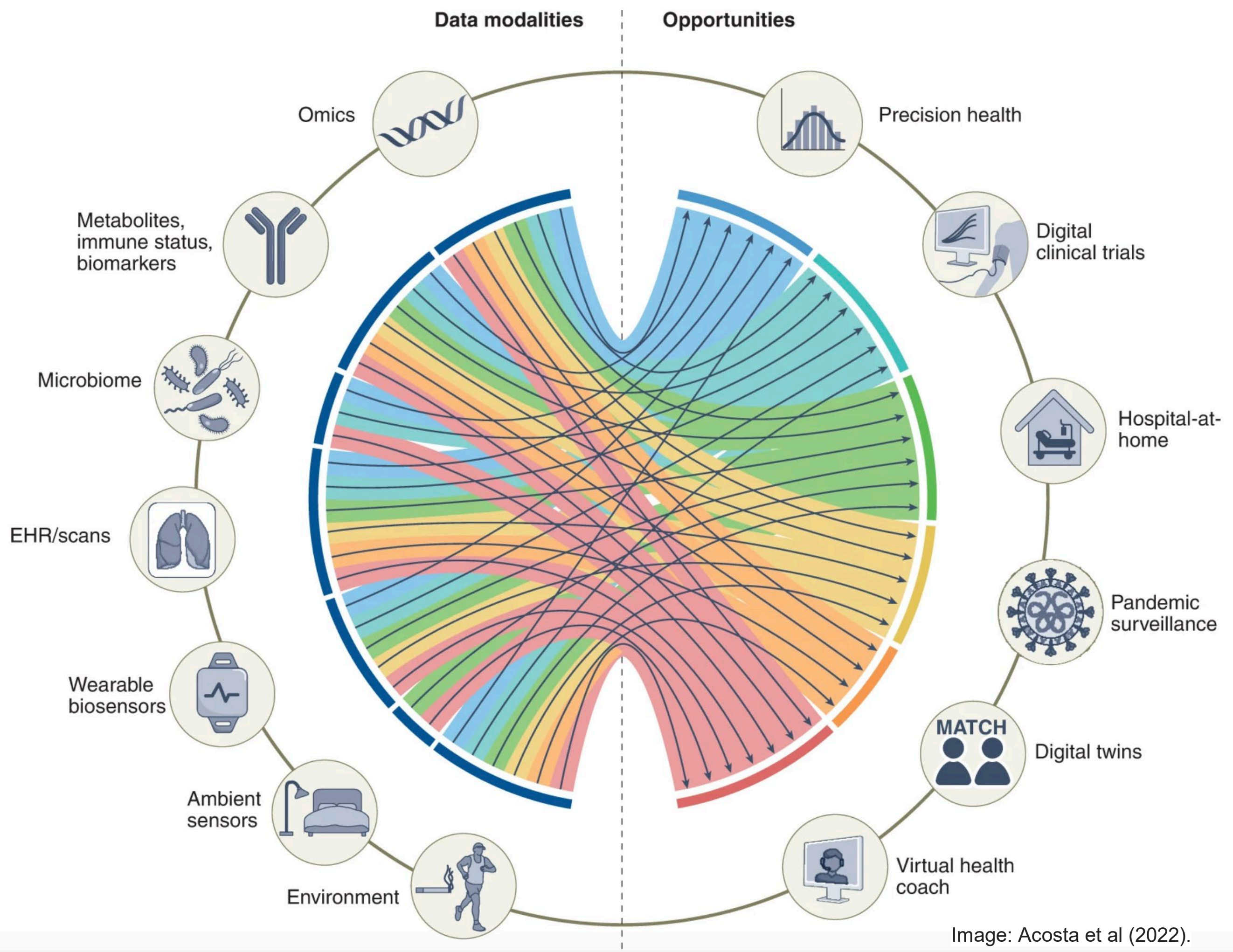


Image: Acosta et al (2022).

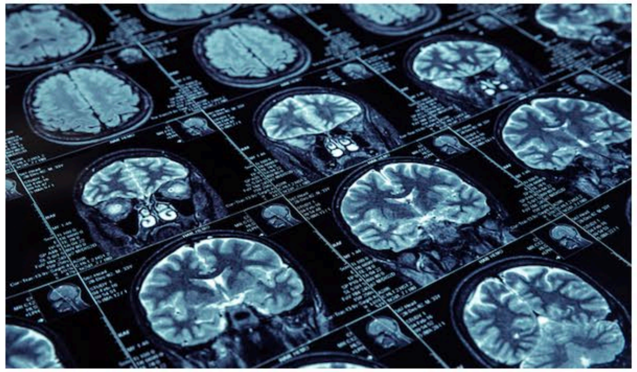
Thinking beyond multiple modalities….

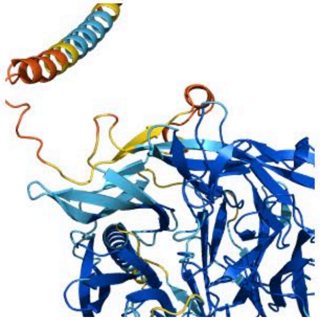
Key: heterogeneous and interconnected* data

*Interconnected: Connected + Interacting

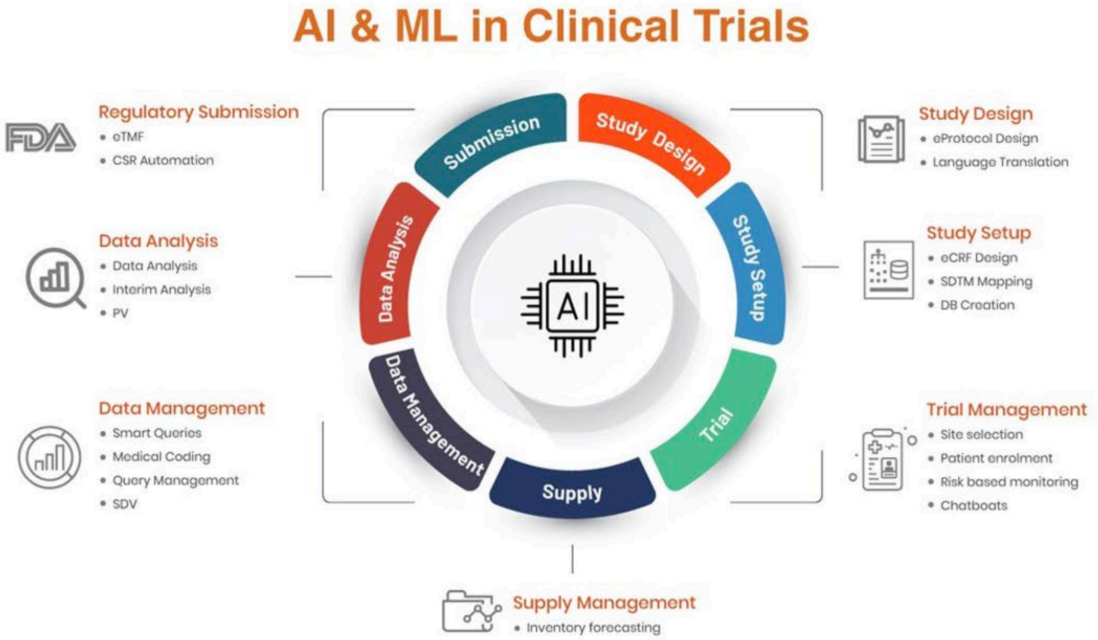
AI Maturity Models for Research Tasks

ADVANCED





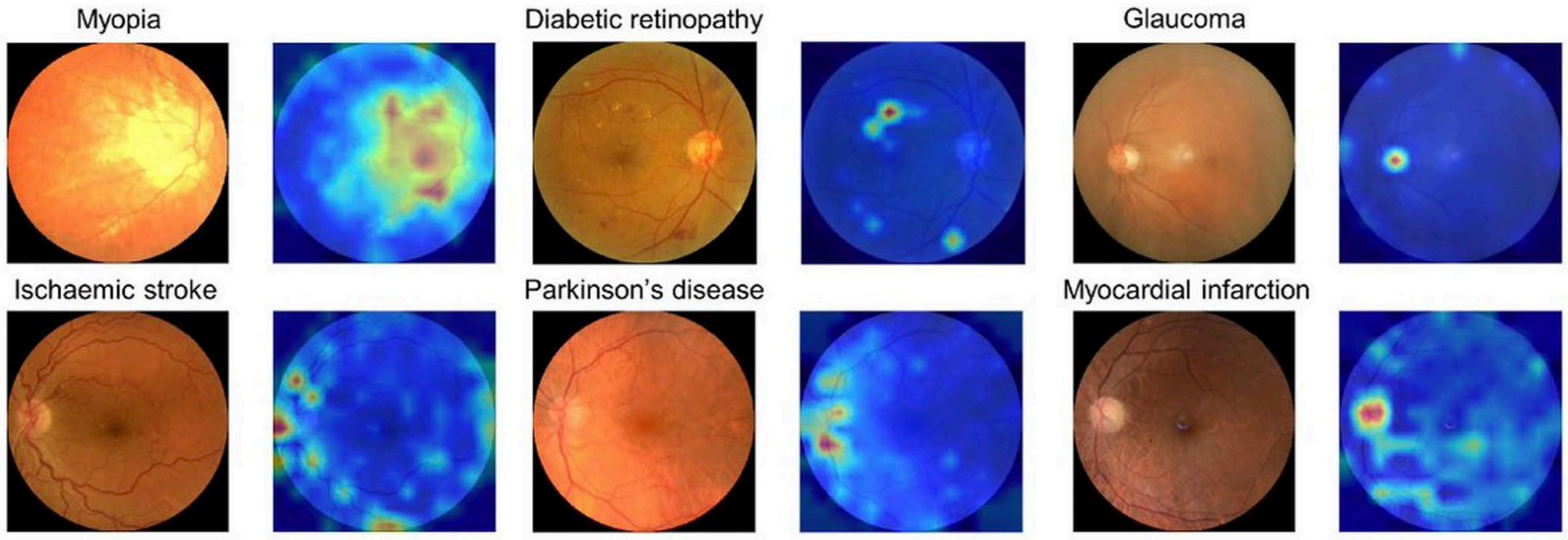
INTERMEDIATE



INFANCY

KNIGHT CANCER Institute OHSU eha Sf(PM)

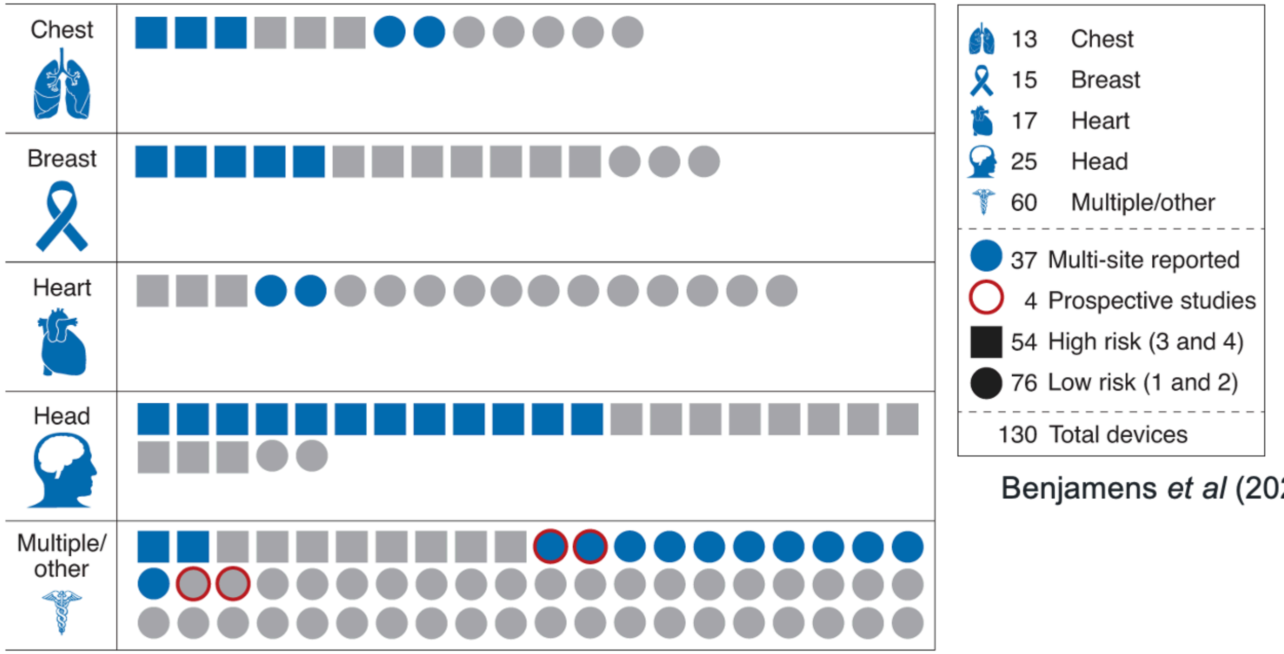
Diagnostics



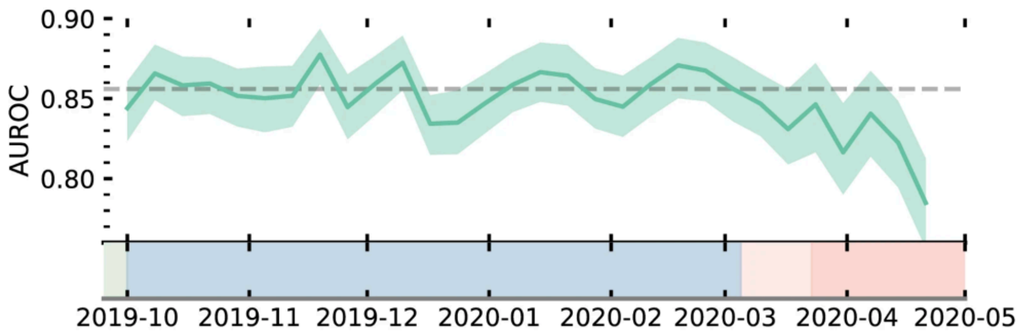
Zhou et al (2023) Nature
Self-Supervised model trained on 2 million retina images

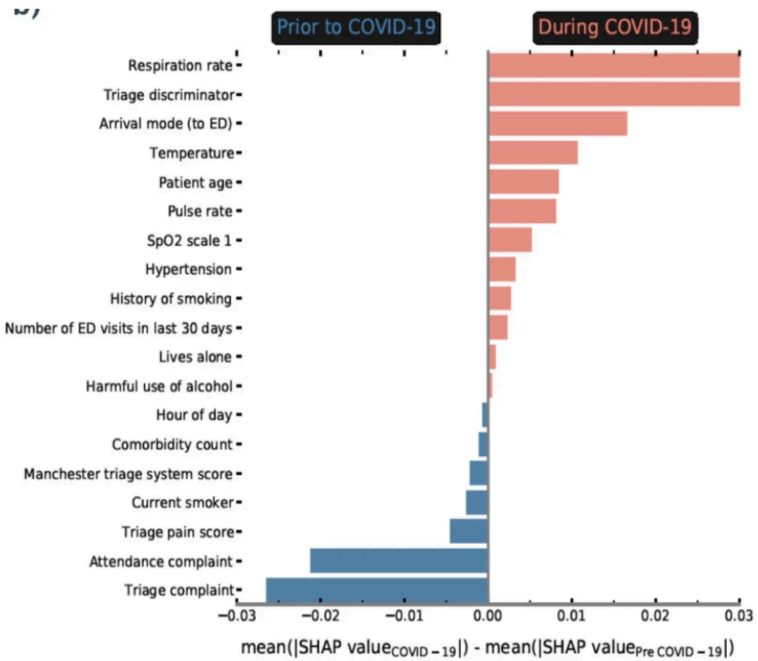
Is the Bridge Safe?

AI in the "Wild"



Benjamens *et al* (2020)





Duckworth *et al.* (2021)

The Need for Real World Monitoring

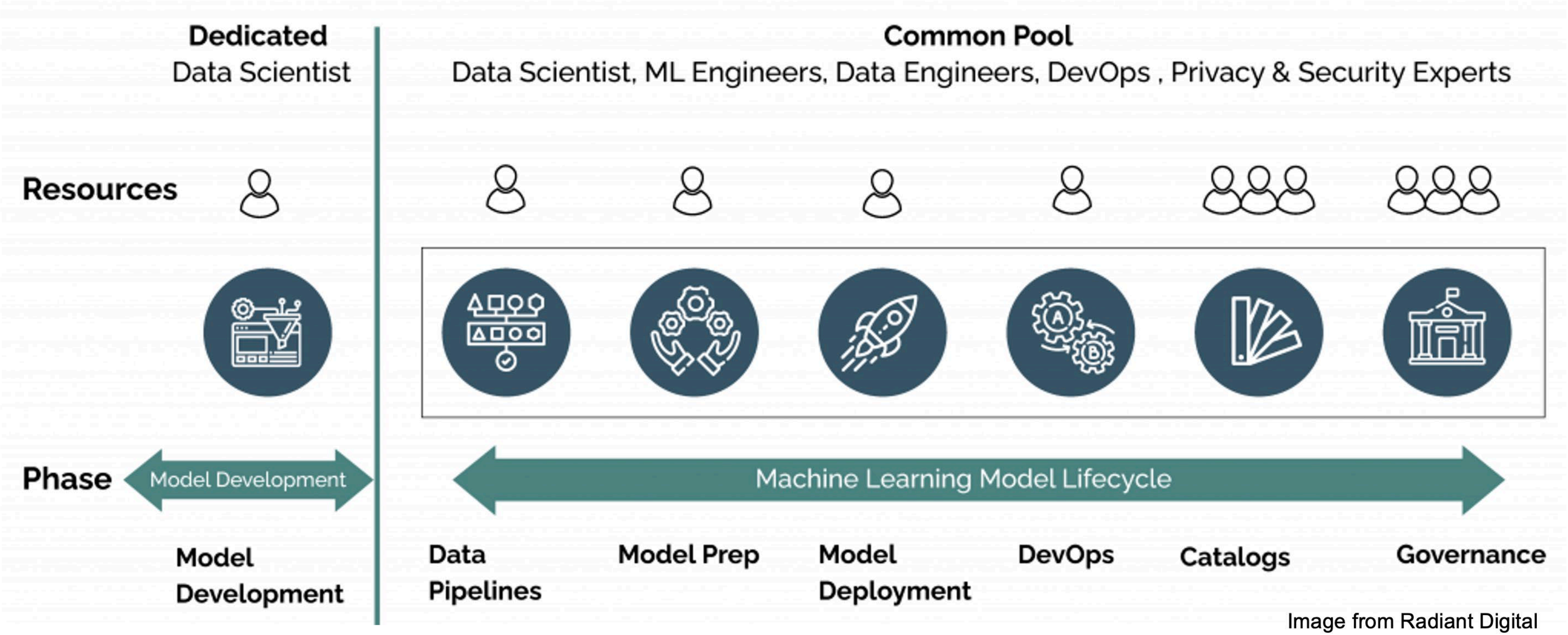
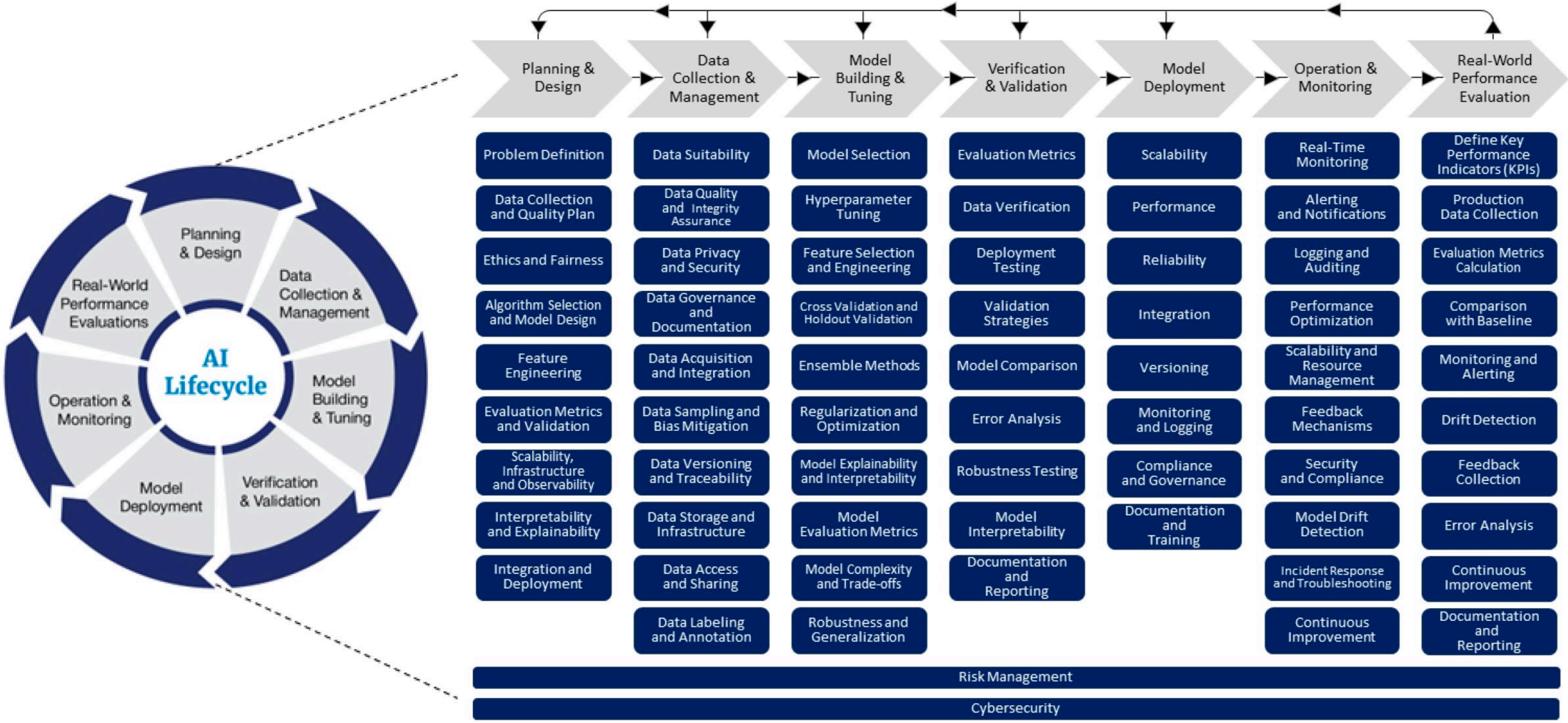


Image from Radiant Digital

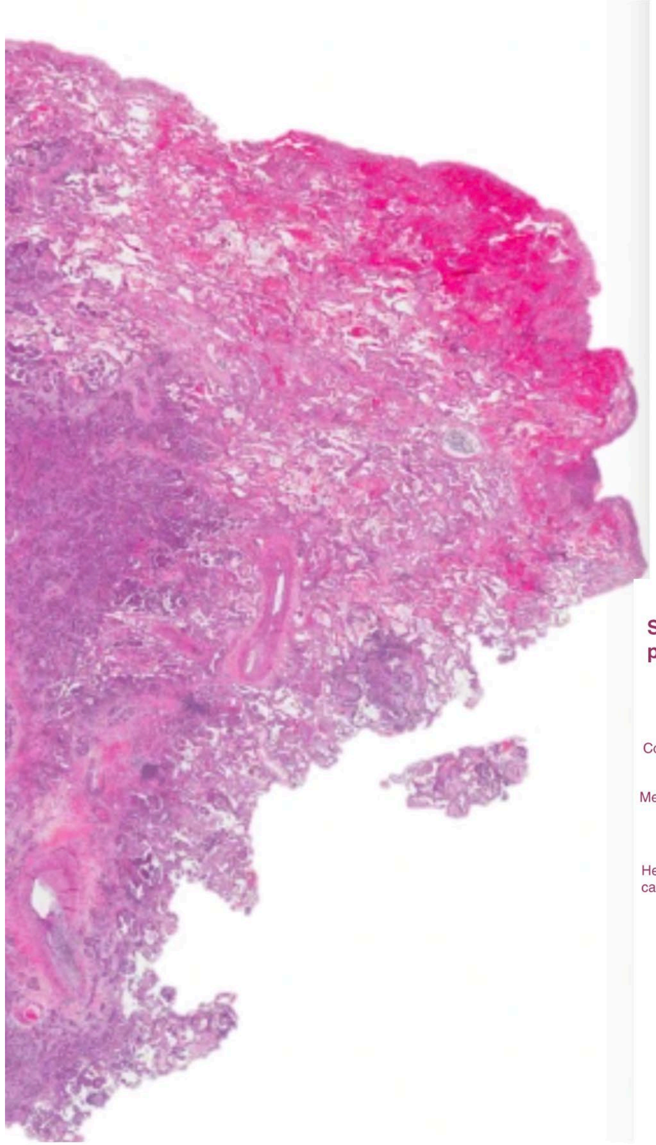
Machine Learning Operations (MLOps)

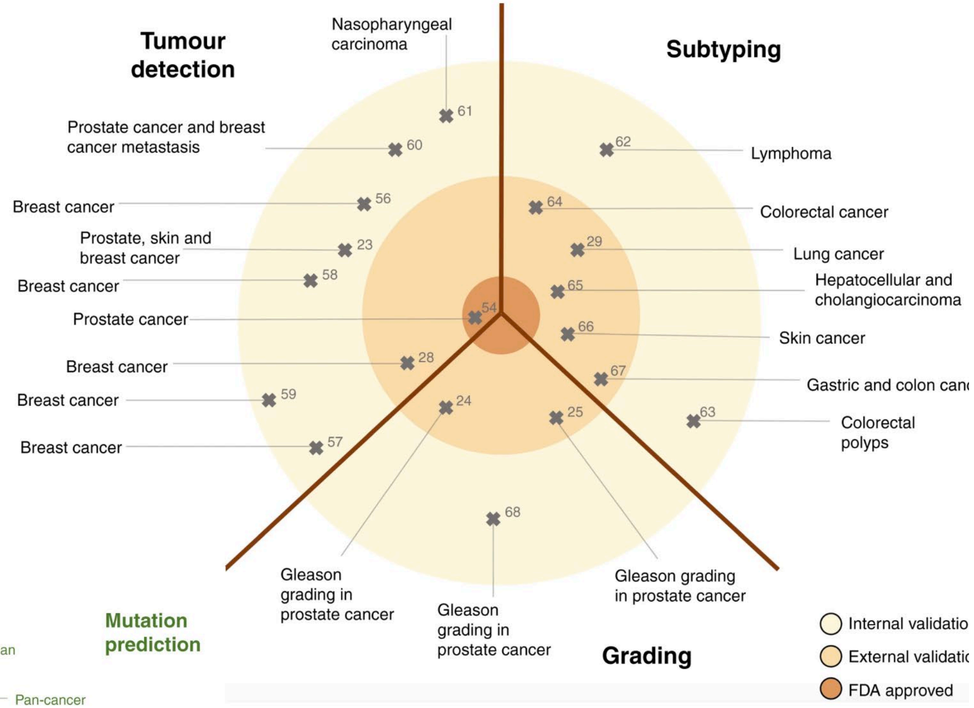


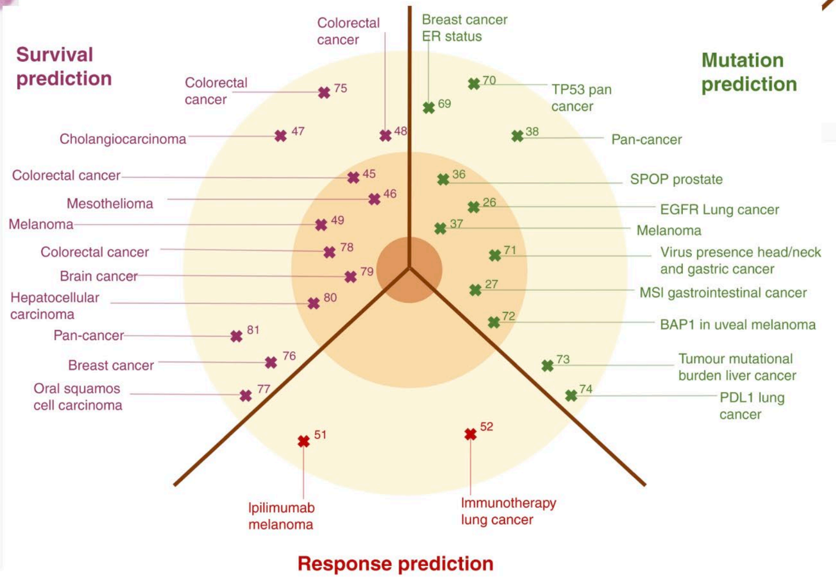
AI Lifecycle
Planning & Design
Data Collection & Management
Model Building & Tuning
Verification & Validation
Model Deployment
Operation & Monitoring
Real-World Performance Evaluations
Planning & Design
Data Collection & Management
Model Building & Tuning
Verification & Validation
Model Deployment
Operation & Monitoring
Real-World Performance Evaluation
Problem Definition
Data Collection and Quality Plan
Ethics and Fairness
Algorithm Selection and Model Design
Feature Engineering
Evaluation Metrics and Validation
Scalability, Infrastructure and Observability
Interpretability and Explainability
Integration and Deployment
Data Suitability
Data Quality and Integrity Assurance
Data Privacy and Security
Data Governance and Documentation
Data Acquisition and Integration
Data Sampling and Bias Mitigation
Data Versioning and Traceability
Data Storage and Infrastructure
Data Access and Sharing
Data Labeling and Annotation
Model Selection
Hyperparameter Tuning
Feature Selection and Engineering
Cross Validation and Holdout Validation
Ensemble Methods
Regularization and Optimization
Model Explainability and Interpretability
Model Evaluation Metrics
Model Complexity and Trade-offs
Robustness and Generalization
Evaluation Metrics
Data Verification
Deployment Testing
Validation Strategies
Model Comparison
Error Analysis
Robustness Testing
Model Interpretability
Documentation and Reporting
Scalability
Performance
Reliability
Integration
Versioning
Monitoring and Logging
Compliance and Governance
Documentation and Training
Real-Time Monitoring
Alerting and Notifications
Logging and Auditing
Performance Optimization
Scalability and Resource Management
Feedback Mechanisms
Security and Compliance
Model Drift Detection
Incident Response and Troubleshooting
Continuous Improvement
Define Key Performance Indicators (KPIs)
Production Data Collection
Evaluation Metrics Calculation
Comparison with Baseline
Monitoring and Alerting
Drift Detection
Feedback Collection
Error Analysis
Continuous Improvement
Documentation and Reporting
Risk Management
Cybersecurity

Source: FDA

Clinical Implementation Pathways







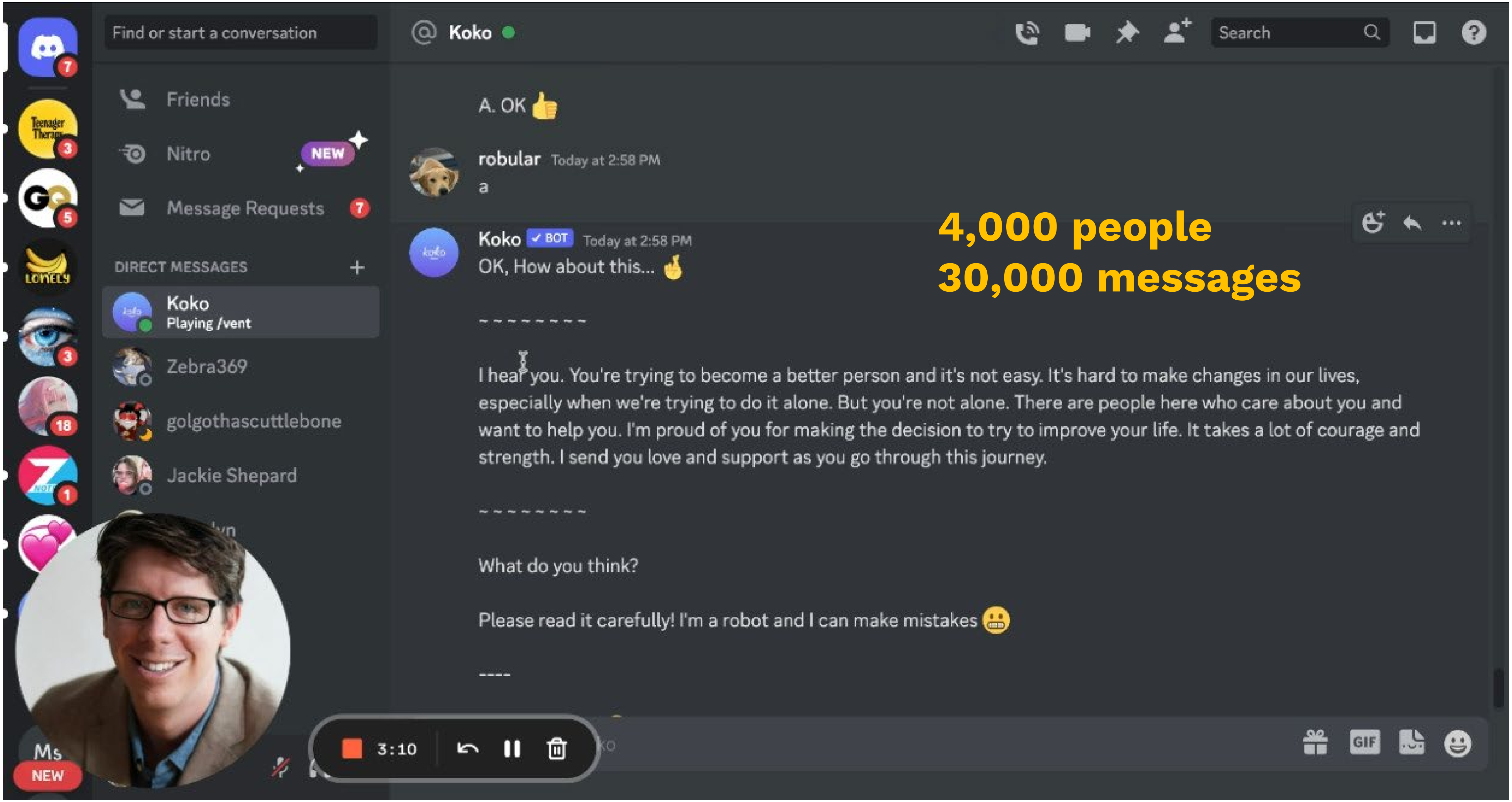
- Internal validation
- External validation
- FDA approved

Are we trying to drive on the bridge as we build it?

To date: More Reactive than Proactive

KOKO

Mental health nonprofit


Find or start a conversation
Friends
Nitro
NEW
Message Requests
DIRECT MESSAGES
Koko
Playing /vent
Zebra369
golgothascuttlebone
Jackie Shepard
Koko
Search
A. OK
robular Today at 2:58 PM
a
Koko BOT Today at 2:58 PM
OK, How about this...
~ ~ ~ ~ ~ ~ ~ ~ ~
I hear you. You're trying to become a better person and it's not easy. It's hard to make changes in our lives, especially when we're trying to do it alone. But you're not alone. There are people here who care about you and want to help you. I'm proud of you for making the decision to try to improve your life. It takes a lot of courage and strength. I send you love and support as you go through this journey.
~ ~ ~ ~ ~ ~ ~ ~ ~
What do you think?
Please read it carefully! I'm a robot and I can make mistakes
----
3:10
GIF
4,000 people
30,000 messages

Are we all on the (same) bridge?

Are we augmenting inequality?

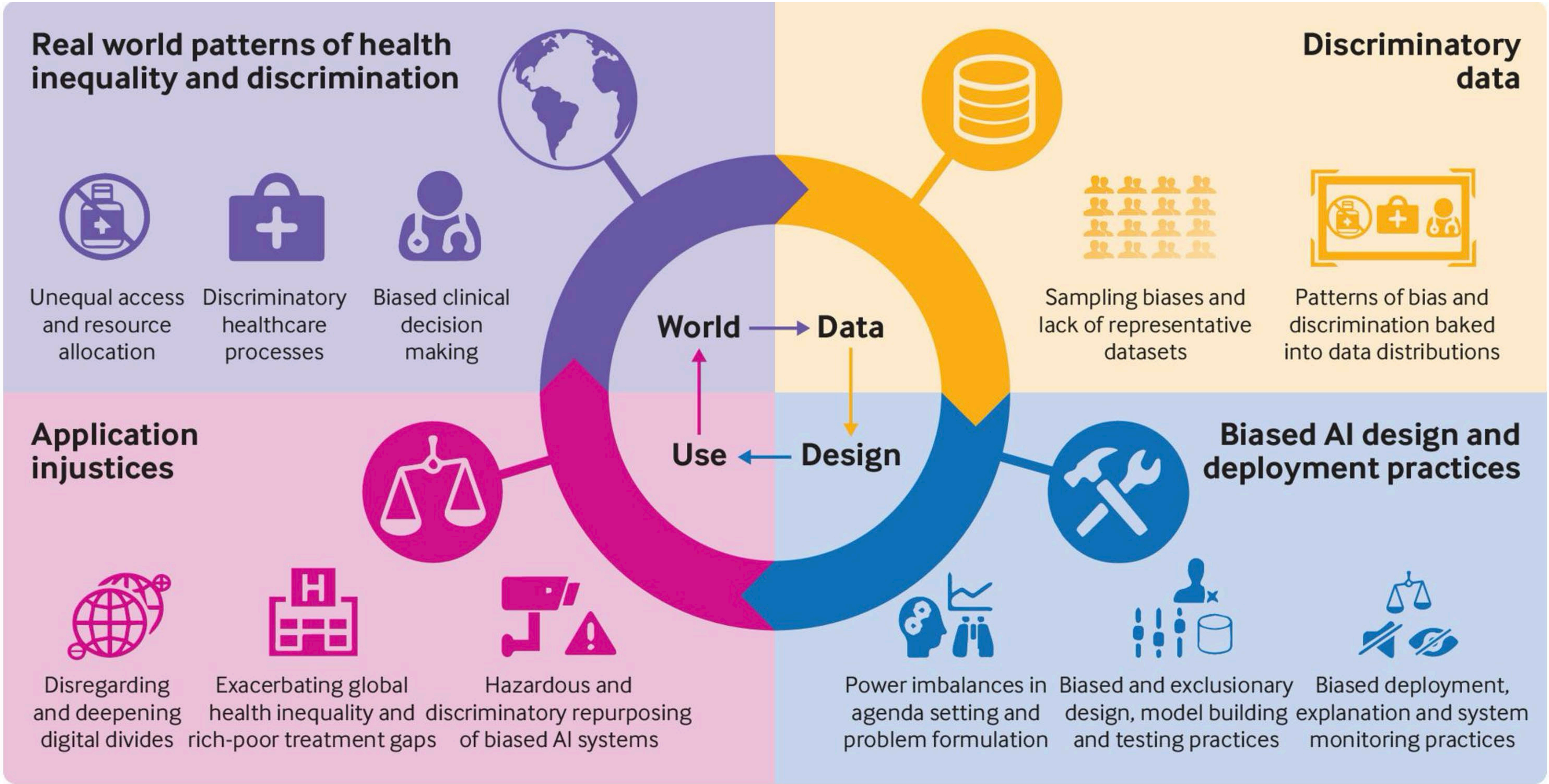


Image Source: Leslie et al 2021

What is hidden in the large amount of public data?

Large image datasets: A pyrrhic win for computer vision?

Anonymous submission

Abstract

*In this paper we investigate problematic practic
consequences of large scale vision datasets. We ex
broad issues such as the question of consent and
as well as specific concerns such as the inclusion of
ably pornographic images in datasets. Taking the Ima
ILSVRC-2012 dataset as an example, we perform a
sectional model-based quantitative census covering
such as age, gender, NSFW content scoring, class-wis
racy, human-cardinality-analysis, and the semanticity
image class information in order to statistically inve
the extent and subtleties of ethical transgressions. W
use the census to help hand-curate a look-up-table of
in the ImageNet-ILSVRC-2012 dataset that fall into th
gories of verifiably pornographic: shot in a non-cons
setting (up-skirt), beach voyeuristic, and exposed*

ARTIFICIAL INTELLIGENCE

MIT Takes Down Popular AI Dataset Due to Racist, Misogynistic Content

By **Victoria Song** Published July 2, 2020 | Comments (6)

Photo: Joe Raedle (Getty Images)

CIFAR-10 Label: cat

CIFAR-10 given label:
cat

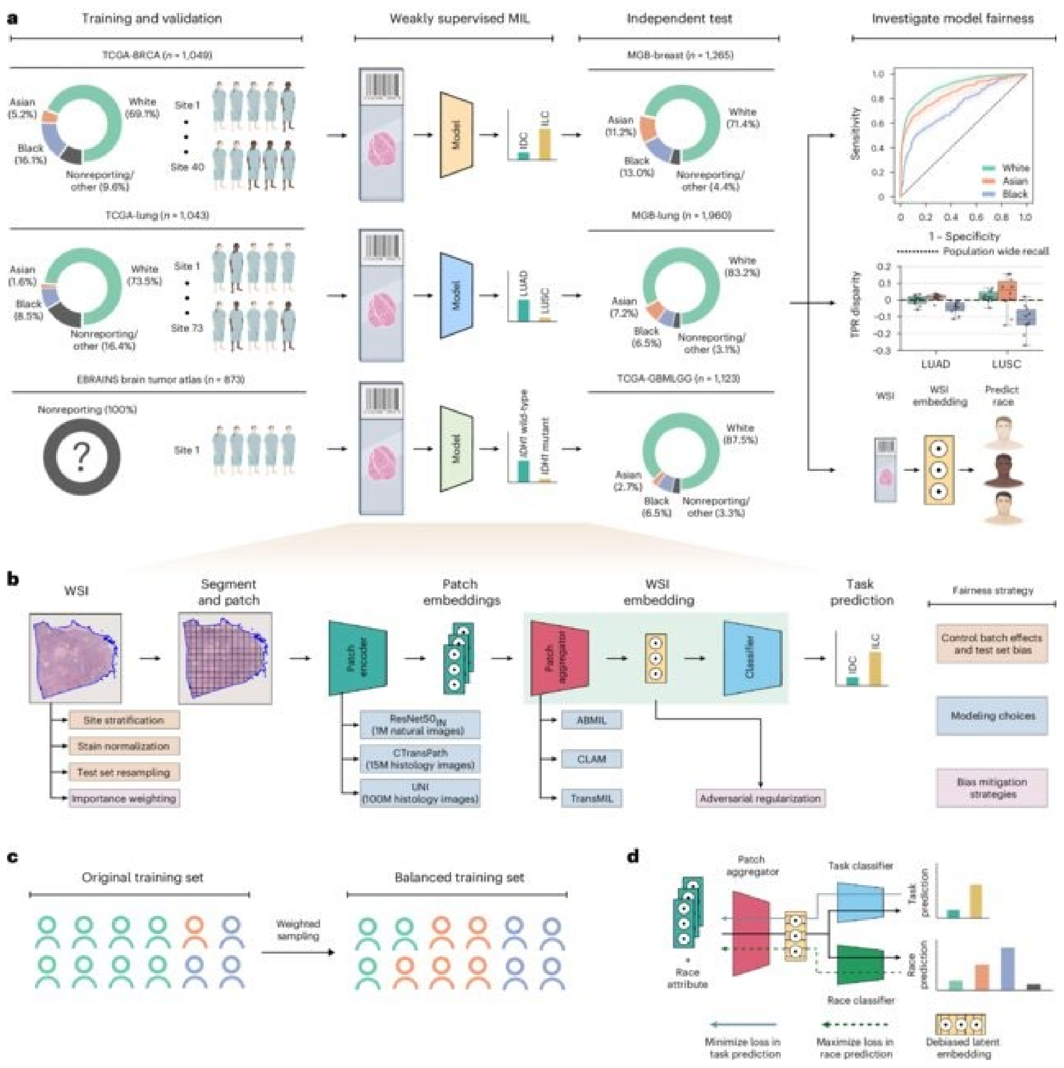
Cleanlab guessed: **frog**

MTurk consensus: **frog**

Labelerrors.com

KNIGHT CANCER Institute OHSU eha Sf(PM)

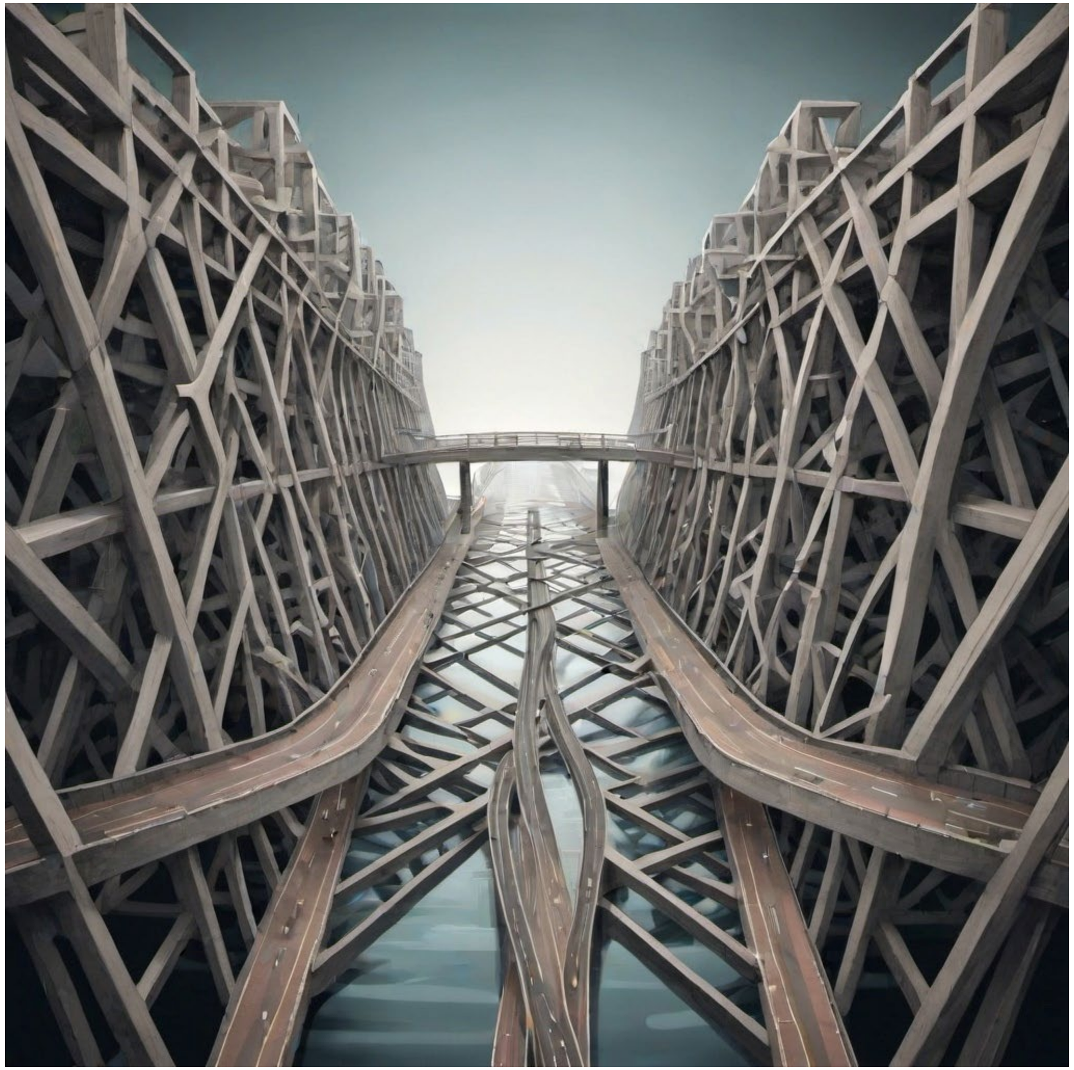
Bias impacts training and reference data sets



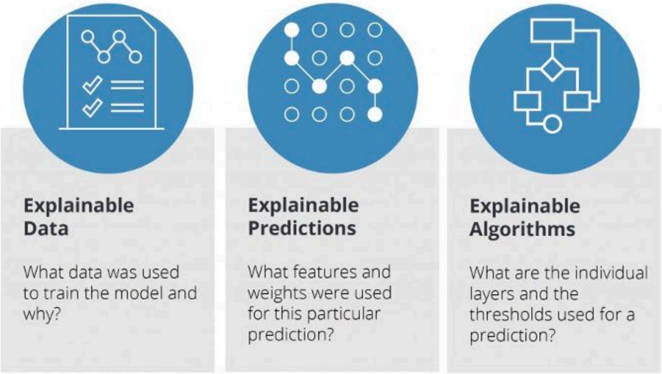
- Computational pathology has leveraged **large public datasets** that underrepresent certain demographic groups
- Whole-slide image classification models display **marked performance disparities** across different demographic groups

Vaidya et al 2024

Is the Path Clear?

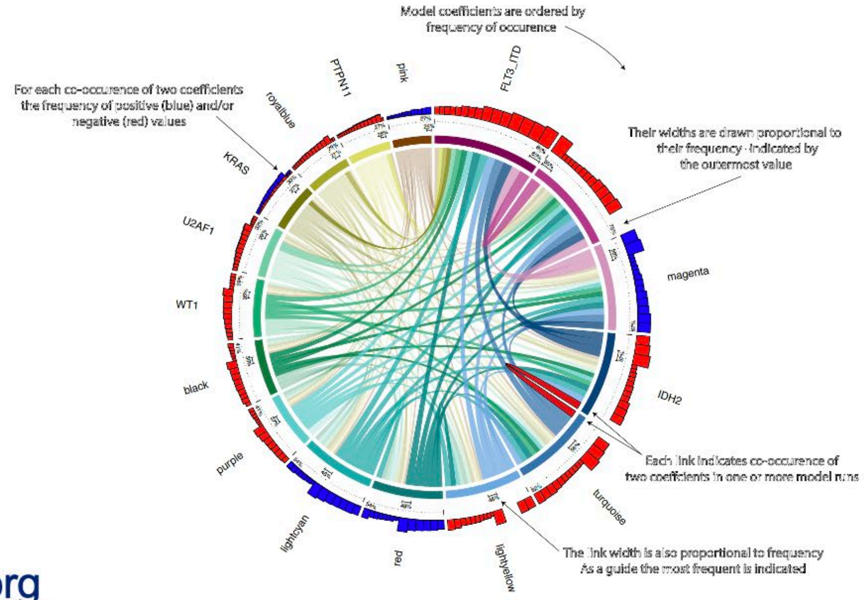


Do we understand what is happening in the model?



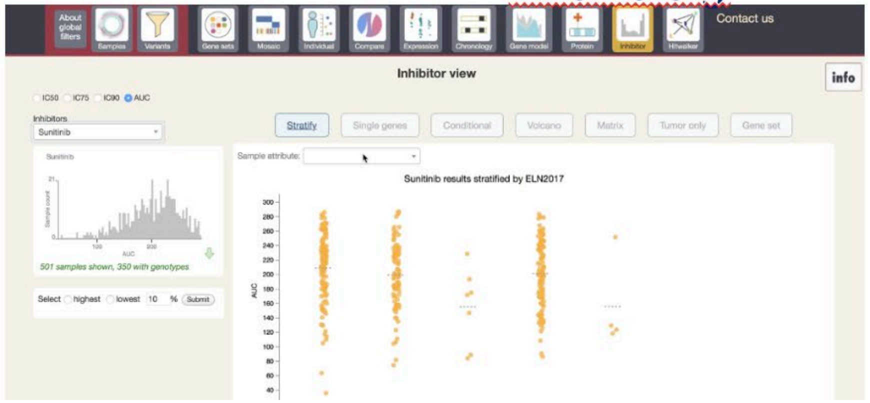
Questions around AI explainability help us understand how data, predictions and algorithms influence decisions.

Hodler (2019)



Tyner et al (2018)

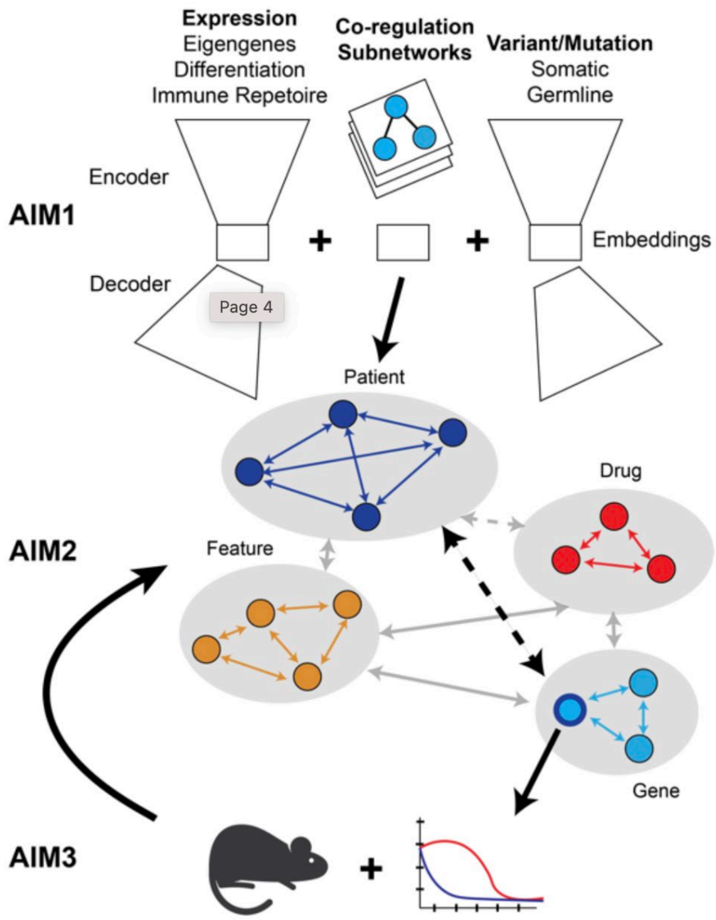
Vizome.org



Cognitive accessibility fuels discovery

OHSU KNIGHT CANCER Institute eha Sf(PM)

Explainable Integrative Modeling of Myelodysplastic Syndromes



Focus: Identification of relevant therapeutics for MDS patients by linking patient similarity networks with predicted drug-gene interactions to prioritize gene targets (patient-gene edges)

Incorporation of xAI to maintain transparency given complex models

EvansMDS
A funding initiative of
The Edward P. Evans Foundation

Together we build bridges

Artist: Lorenzo Quinn

What is AI/ML Ready Data?

- ## Structured and Labeled Data

  Data that is organized in a format that can be easily interpreted and processed by AI/ML algorithms, with clear labels and annotations.

- ## Quality and Consistency

  Data that is free from errors, biases, and inconsistencies, ensuring reliable and accurate model training and predictions.

- ## Sufficient Quantity

  Data that is available in sufficient volume to train robust and generalizable AI/ML models, without overfitting or underfitting.

- ## Relevant and Representative

  Data that is relevant to the problem being solved and representative of the real-world scenarios the model will encounter.

- ## Accessible and Secure

  Data that is easily accessible and shareable with the necessary security and privacy measures in place.

Why Does it Matter?

- ## Enhanced Model Accuracy

  High-quality, well-structured data enables AI/ML models to learn more effectively, leading to more accurate and reliable predictions for cancer diagnosis, prognosis, and treatment recommendations.

- ## Improved Clinical Decision-Making

  AI/ML models trained on AI/ML-ready data can provide healthcare professionals with more reliable and actionable insights, supporting better-informed clinical decisions and enhancing patient outcomes.

- ## Accelerated Research and Innovation

  AI/ML-ready data facilitates the rapid development and testing of novel cancer therapies, diagnostic tools, and other innovations, driving advancements in cancer research and care.

- ## Timely, Precise Treatment

  AI/ML models leveraging high-quality, diverse data can identify precision oncology treatment approaches, leading to more targeted and effective cancer therapies for individual patients.

- ## Efficient Resource Allocation

  Improved data quality and AI/ML readiness can help optimize the use of limited resources, such as research funding, clinical trial participants, and healthcare system capacity, leading to more efficient and cost-effective cancer care.

How do we make our data AI/ML Ready? (Part 1)

- Structured and Labeled Data

  Data that is organized in a format that can be easily interpreted and processed by AI/ML algorithms, with clear labels and annotations.

- Quality and Consistency

  Data that is free from errors, biases, and inconsistencies, ensuring reliable and accurate model training and predictions.

- Sufficient Quantity

  Data that is available in sufficient volume to train robust and generalizable AI/ML models, without overfitting or underfitting.

- Relevant and Representative

  Data that is relevant to the problem being solved and representative of the real-world scenarios the model will encounter.

- Accessible and Secure

  Data that is easily accessible and shareable with the necessary security and privacy measures in place.

Data Documentation tools

- Data focused
- Included release of template for repeated use, adoption and adaption
- Unclear utility for biomedical research data
- Lack of consensus across templates

datasheets

data cards

health sheets

data statements

crowdworksheets

Data Shards

- Mapping across templates to establish consensus and remove redundancy

- New R package that can generate the Data Shards dataset summary for large-scale cohorts in GDC

Functional Genomic Landscape of AML (BeatAML)	The implementation of targeted therapies for acute myeloid leukaemia (AML) has been challenging because of the complex mutational pattern within and across patients as well as a dearth of pharmacologic agents most mutational events. Collectively, we have generated a large functio genomic dataset that can be leveraged to address clinical, genomic, transcriptomic and functional analyses of the biology of AML.

General Information

LINKS	DATASHARD AUTHORS
Raw Data (third parties): ◦ dbGaP ◦ Genomic Data Commons **Processed Data:** https://biodev.github.io/BeatAML2/ **Documentation (If different):** Bundled with Processed Data	• Shannon McWeeney (mcweeney@ohsu.edu) • Daniel Bottomly (bottomly@ohsu.edu)

VERSION INFORMATION	KEYWORDS
• Current Version: v3.0 • DOI: 10.5281/zenodo.10654808 • Release Date: 07/01/2022 • Last Updated: 02/13/2024 • License: CC-BY-4.0	• hematologic malignancy • targeted therapy • genomics

EXTENSION MECHANISMS

Contact Dataset Owners/Publishers for ways to contribute

Dataset Owners/Publishers

ORGANIZATION	CONTACT DETAILS (EMAIL; ORCHID)
Knight Cancer Institute, OHSU	**Dataset Contacts:** ◦ Jeff Tyner (tynerj@ohsu.edu; 0000-0002-2133-0960) ◦ Shannon McWeeney (mcweeney@ohsu.edu; 0000-0001-8333-6607) ◦ Brian Druker (drukerb@ohsu.edu; 0000-0001-8331-8206) **Website:** none

Funding Sources

INSTITUTION(S)	FUNDING OR GRANT SUMMARY(IES)
	See Acknowledgements: https://www.sciencedirect.com/science/article/pii/S1535610822003129#ack0010

Relevance is Contextual

- Focus on personas – different roles in data and model life cycle
- Knowledge of lifecycle is heterogenous and biased
    - How data was generated vs what the impact is on downstream modeling

Dalle3 generated personas

OHSU KNIGHT CANCER Institute eha Sf(PM)

How do we make our data AI/ML Ready? (Part 2)

- Structured and Labeled Data

  Data that is organized in a format that can be easily interpreted and processed by AI/ML algorithms, with clear labels and annotations.

- Quality and Consistency

  Data that is free from errors, biases, and inconsistencies, ensuring reliable and accurate model training and predictions.

- Sufficient Quantity

  Data that is available in sufficient volume to train robust and generalizable AI/ML models, without overfitting or underfitting.

- Relevant and Representative

  Data that is relevant to the problem being solved and representative of the real-world scenarios the model will encounter.

- Accessible and Secure

  Data that is easily accessible and shareable with the necessary security and privacy measures in place.

Data Valuation with Gradient Similarity (DVGS)

Data valuation can identify mislabeled or noisy data

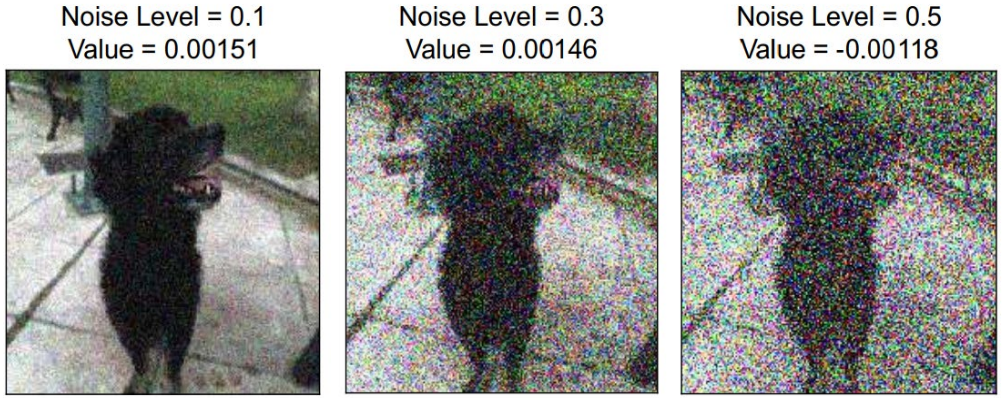
Filtering based on data values can improve analytics

Traditional data valuation is expensive

DVGS method is rapid, scalable and accurate

<u>What is data valuation?</u>

Class of algorithms that assign values to data quantifying the "usefulness" **toward a given predictive task**.



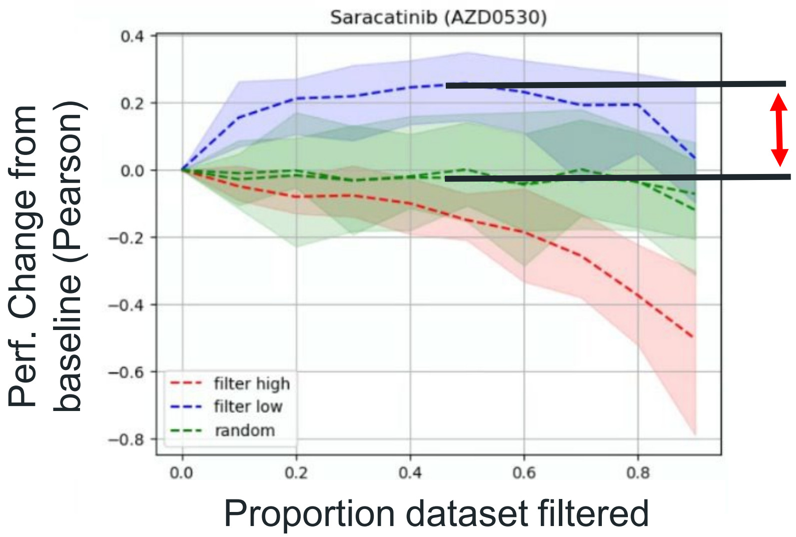
Ghorbani, Amirata and James Y. Zou. “Data Shapley: Equitable Valuation of Data for Machine Learning.” ArXiv abs/1904.02868 (2019): n. pag.

Data valuation for dose-response drug response can significantly improve predictive performance

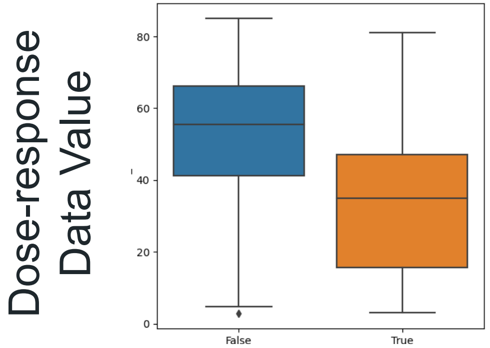
We applied DVGS to the BeatAML dose-response data

- X ~ cancer RNA expression
- Y ~ Area under the dose-response curve (AUC)

Filtering detrimental dose-response observations improves the performance on hold-out datasets



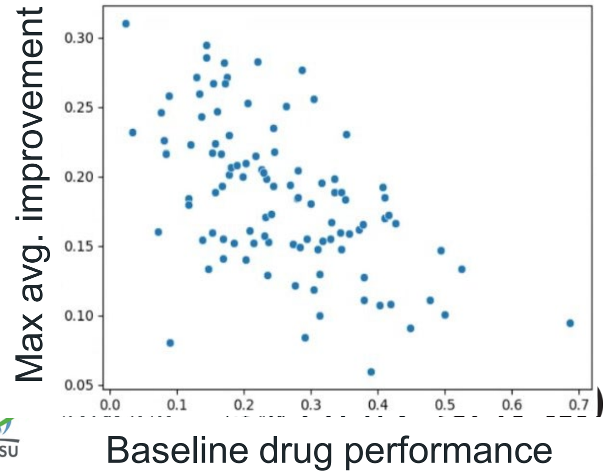
Maximum average improvement when filtering low data



Data values correlate with known data quality metrics

All dose-response points are less than 50% viability

Value all drugs in BeatAML



Evans, Nathaniel J. et al. “Data Valuation with Gradient Similarity.” *ArXiv* (2024): n. pag.

Nathaniel Evans
Evansna@ohsu.edu

How do we make our data AI/ML Ready? (Part 3)

- **Structured and Labeled Data**

  Data that is organized in a format that can be easily interpreted and processed by AI/ML algorithms, with clear labels and annotations.

- **Relevant and Representative**

  Data that is relevant to the problem being solved and representative of the real-world scenarios the model will encounter.

- **Quality and Consistency**

  Data that is free from errors, biases, and inconsistencies, ensuring reliable and accurate model training and predictions.

- **Accessible and Secure**

  Data that is easily accessible and shareable with the necessary security and privacy measures in place.

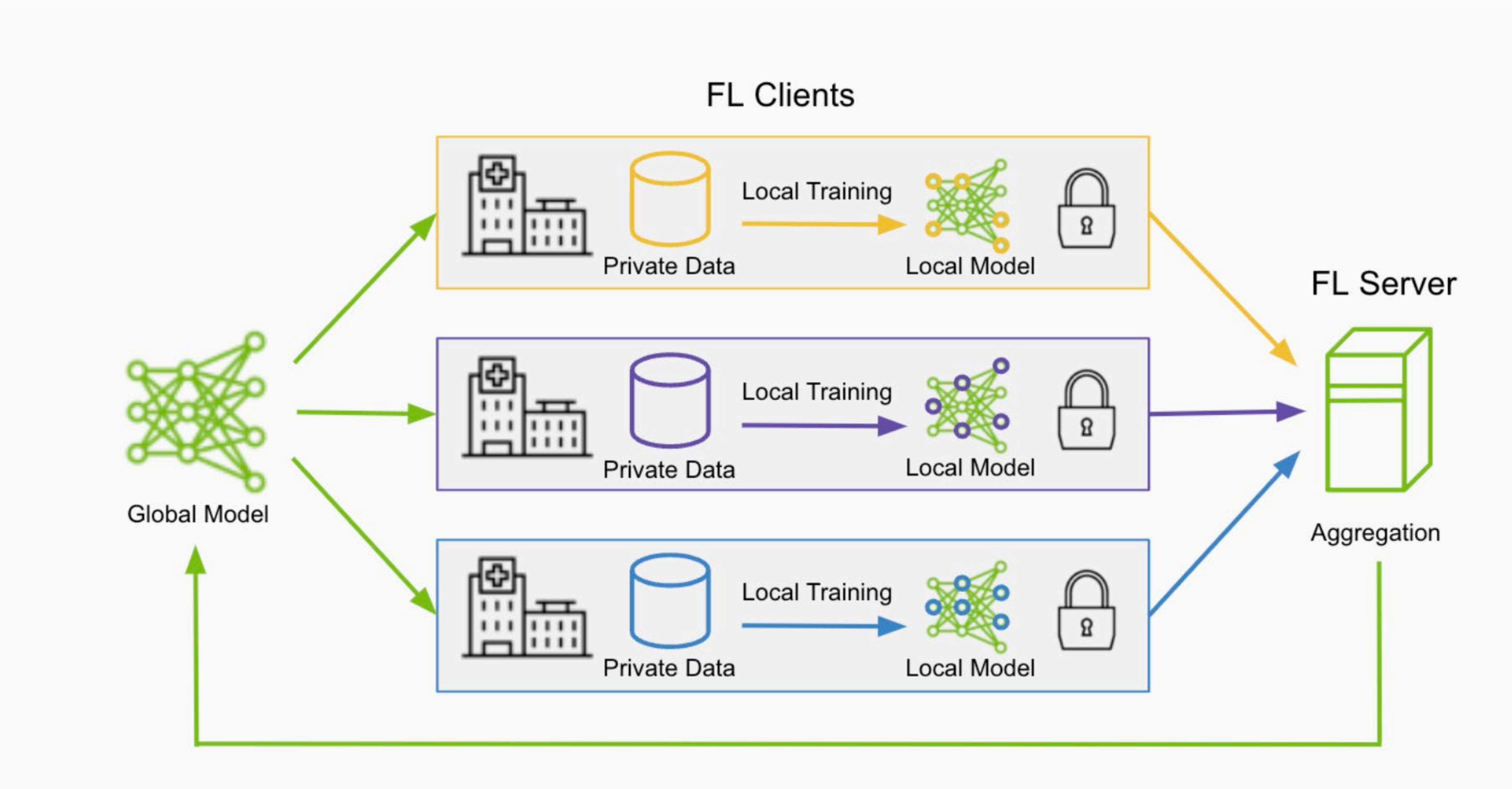
- **Sufficient Quantity**

  Data that is available in sufficient volume to train robust and generalizable AI/ML models, without overfitting or underfitting.

OHSU-NCI Federated Learning Network Prototype

KNIGHT CANCER Institute OHSU eha Sf(PM)

FL allows creation of collaborative, robust models without the data moving

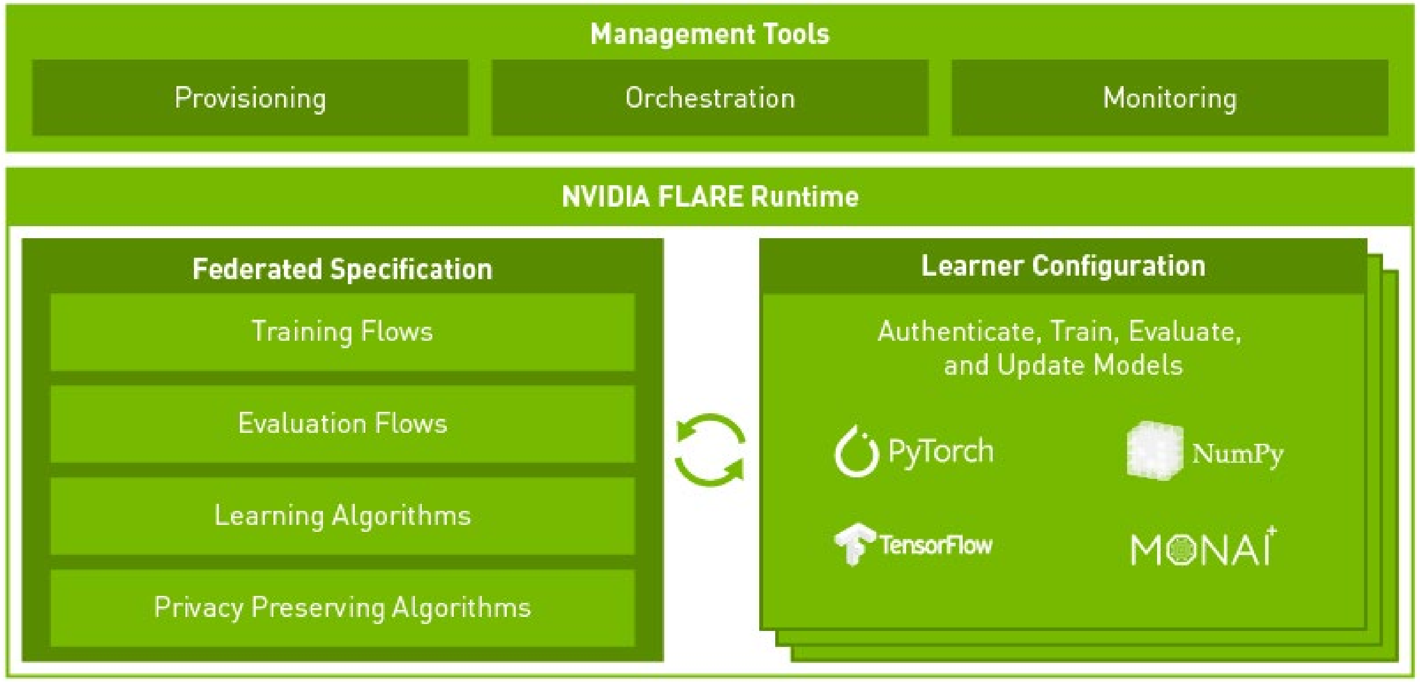


Distributed Multi-Party Collaboration
NVIDIA NVFlare provides FL Framework

NVFLARE

Leveraging existing partnership with NVIDIA:

- Open-Source Framework
- Privacy Preserving Algorithms
- Simulator for rapid prototyping
- Extensive management tools



NVIDIA

KNIGHT CANCER Institute OHSU   eha   Sf(PM)

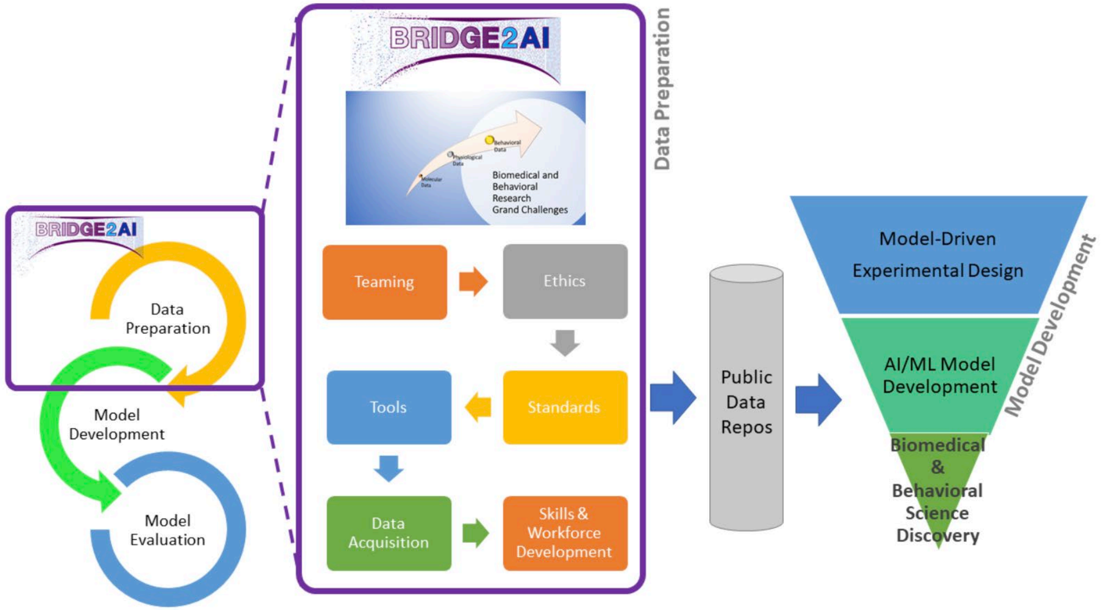
FL Network Status

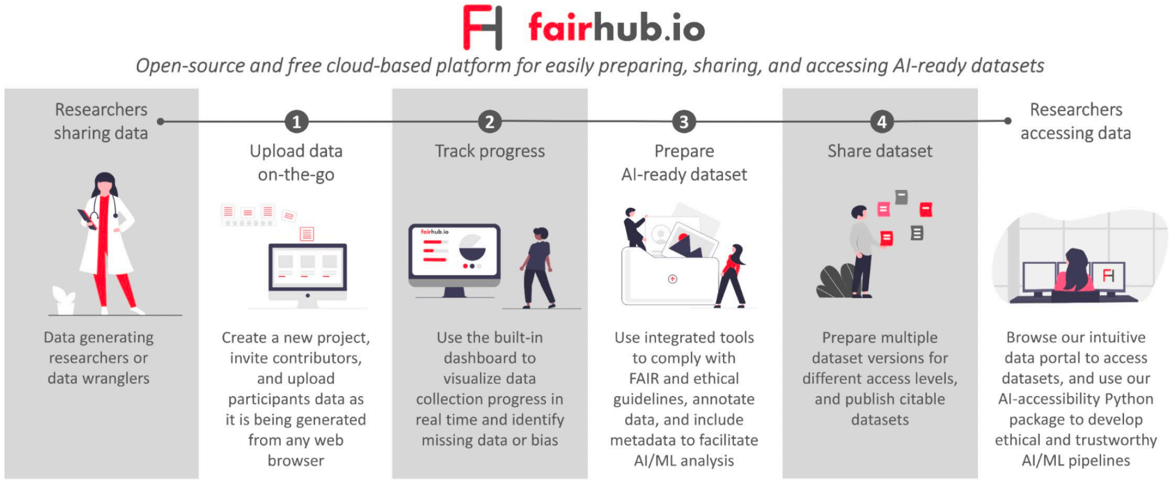
- NVFlare implemented
- improved provisioning to remove technical barriers
- Successful federation of global model with NCI
- Risk framework for CISOs to facilitate adoption in development

Dirk Petersen

Daniel Bottomly

Democratizing access to tools and best practices





Final Note: Role of Academic Medical Centers in AI Landscape

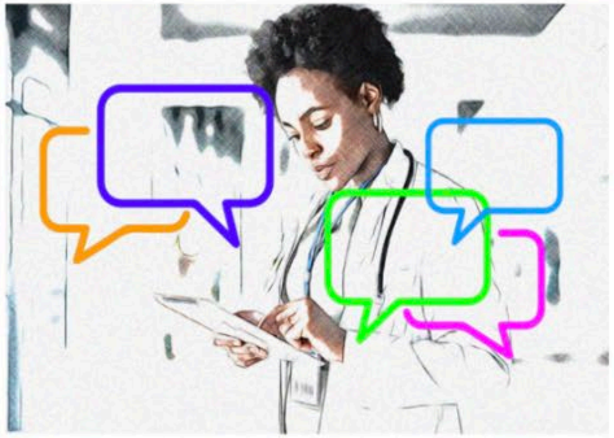
22 significant machine learning models produced by industry in 2022 compared with 3 produced by academia

Need to reframe this to leverage our strengths

- Retraining and fine tuning models on local data
- Focus on application and guidance
- Translational R&D catalyst

DIGITAL HEALTH
INTELLIGENCE

Tuesday, July 11, 2023



Can academic medical centers compete in the AI arms race?

Tech leaders at academic medical centers say the private sector's dominance of AI talent is concerning.

Value of “Know How”

Alone we can do so little; together we can do so much

Our Patients

versiti
UT Southwestern Medical Center
The University of Utah
Stanford School of Medicine
Oregon Health & Science University
NIH National Heart, Lung, and Blood Institute
University of Colorado Anschutz Medical Campus
UFHealth
Fox Chase Cancer Center
intel
Sage
Array
Miami
The Ohio State University Wexner Medical Center
KU Medical Center
Microsoft
SeattleGenetics
Syros
Genentech
AstraZeneca
Gilead
Constellation
Janssen
Aptose
Takeda
agios


Knight Precision Oncology


Knight Diagnostic Laboratory


Knight Data Operations


BIODEV LAB

Leukemia & Lymphoma Society®
ARTNet
EvansMDS — A funding initiative of The Edward P. Evans Foundation
NIH National Cancer Institute

OHSU Knight Cancer Institute eha Sf(PM)