Investigating CRDC AI Data Readiness (AIDR)

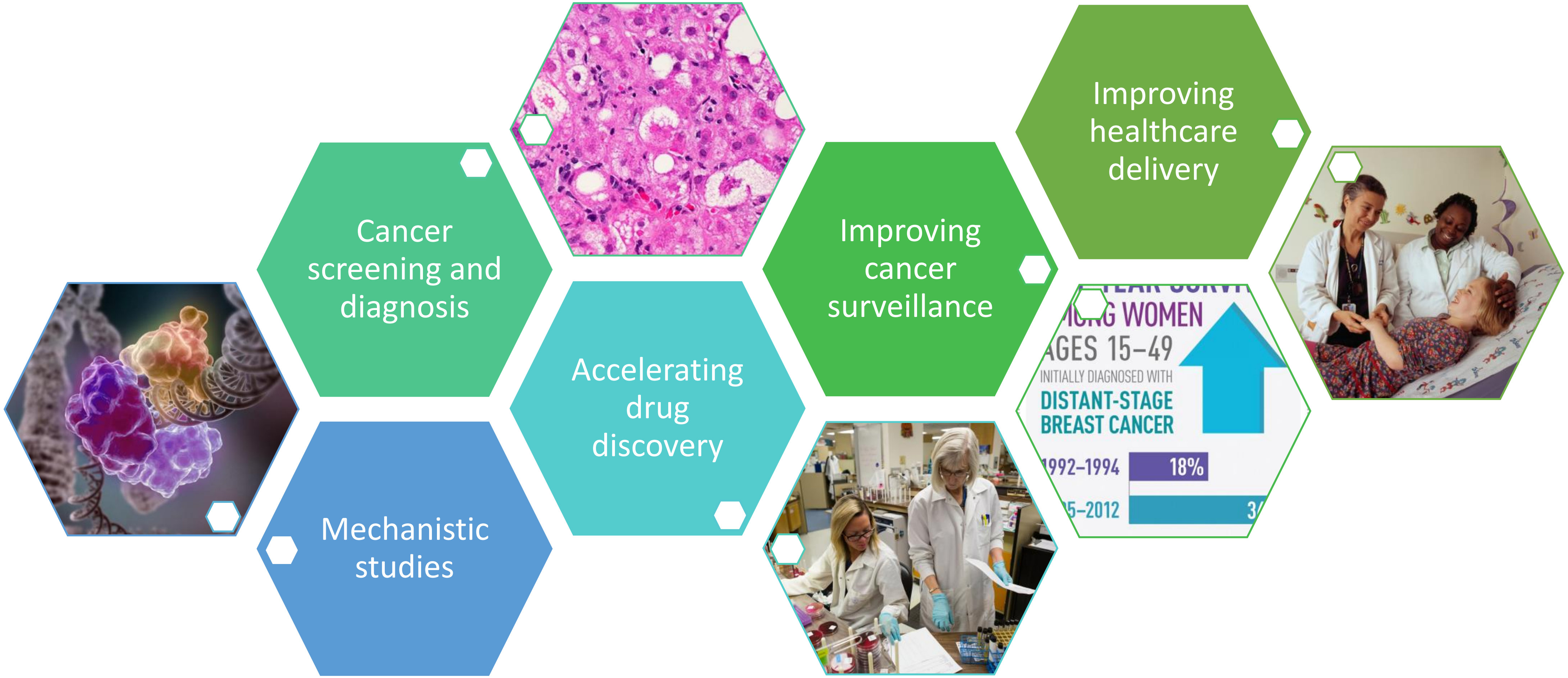
Emily J. Greenspan, Ph.D.

October 17, 2024

NIH NATIONAL CANCER INSTITUTE

Artificial Intelligence (AI)

A powerful tool that can be used across the cancer research continuum



Cancer screening and diagnosis

Mechanistic studies

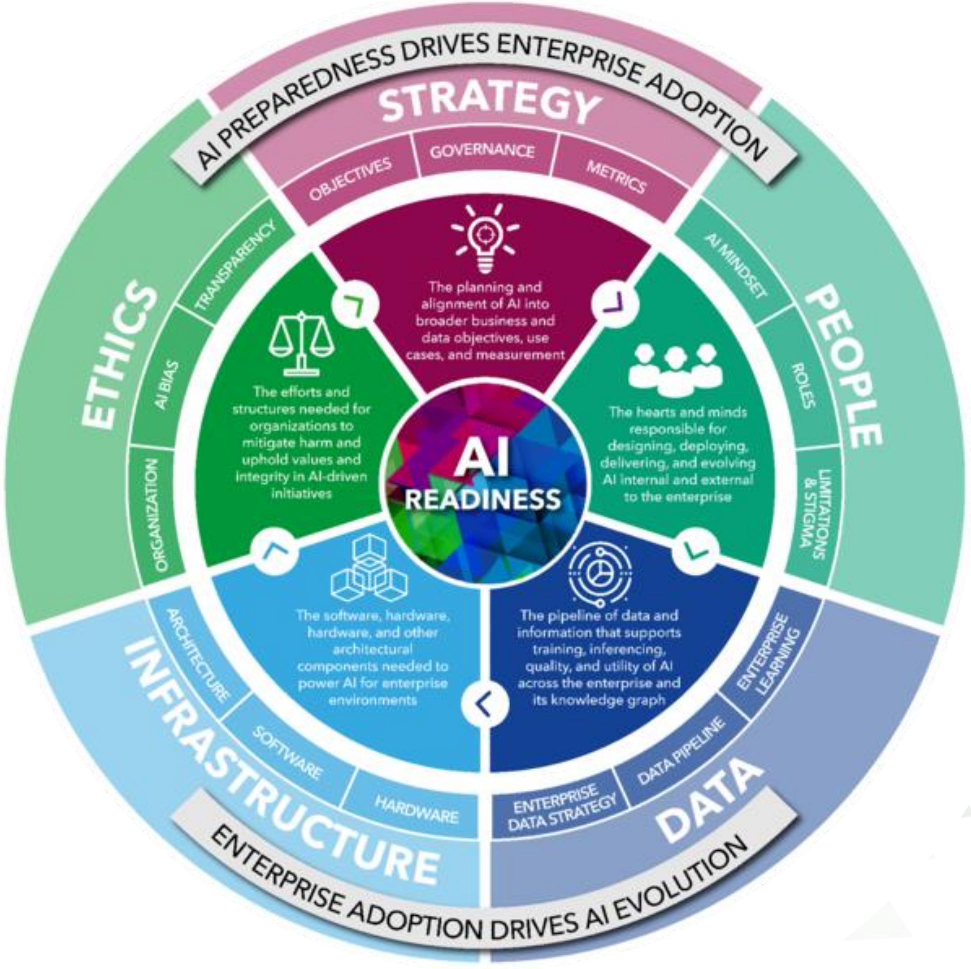
Accelerating drug discovery

Improving cancer surveillance

Improving healthcare delivery

CRDC for Enabling AI in Cancer Research

- **Data-centric AI**
  - Characterize, evaluate, and monitor the data underlying AI models
- **AI Data Readiness (AIDR)**
  - Expands FAIR principles to make data accessible for use in AI future applications
- **Data pre-processing for AI**
  - Labor intensive and inhibits democratization



The planning and alignment of AI into broader business and data objectives, use cases, and measurement

The hearts and minds responsible for designing, deploying, delivering, and evolving AI internal and external to the enterprise

The pipeline of data and information that supports training, inferencing, quality, and utility of AI across the enterprise and its knowledge graph

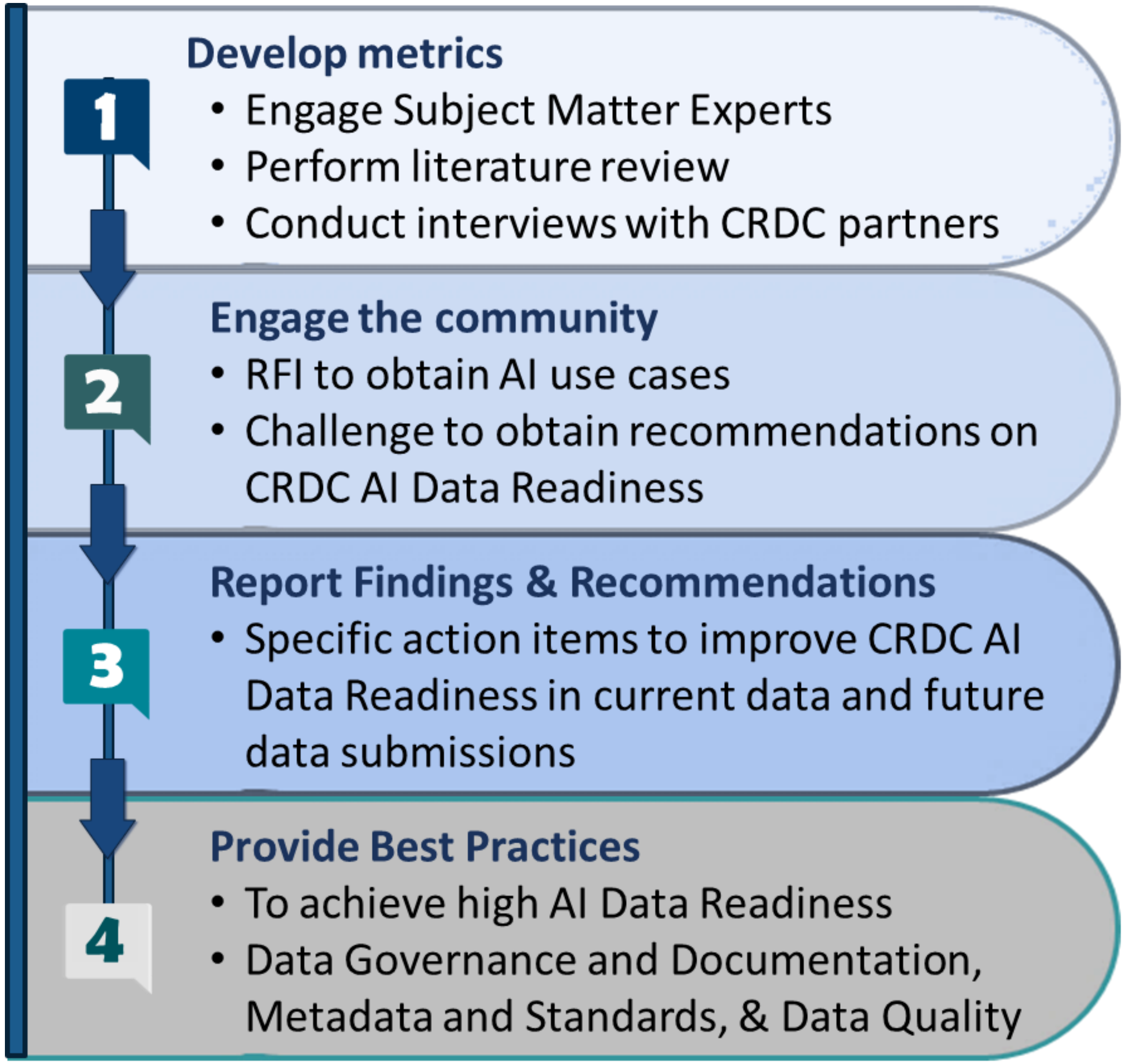
The software, hardware, hardware, and other architectural components needed to power AI for enterprise environments

The efforts and structures needed for organizations to mitigate harm and uphold values and integrity in AI-driven initiatives

https://www.vanderschaar-lab.com/dc-check/what-is-data-centric-ai/

Evaluating CRDC AIDR

- Current Data Commons Harmonization Activities
    - File standards
    - Data/Metadata standards
    - Uniform analytic pipelines
- Cross CRDC Harmonization In Progress
    - Improving data model & metadata
    - Unified portal for search



1 **Develop metrics**

- Engage Subject Matter Experts
- Perform literature review
- Conduct interviews with CRDC partners

2 **Engage the community**

- RFI to obtain AI use cases
- Challenge to obtain recommendations on CRDC AI Data Readiness

3 **Report Findings & Recommendations**

- Specific action items to improve CRDC AI Data Readiness in current data and future data submissions

4 **Provide Best Practices**

- To achieve high AI Data Readiness
- Data Governance and Documentation, Metadata and Standards, & Data Quality

Request for Information Summary

NCI CRDC AI use cases to inform an assessment of data readiness

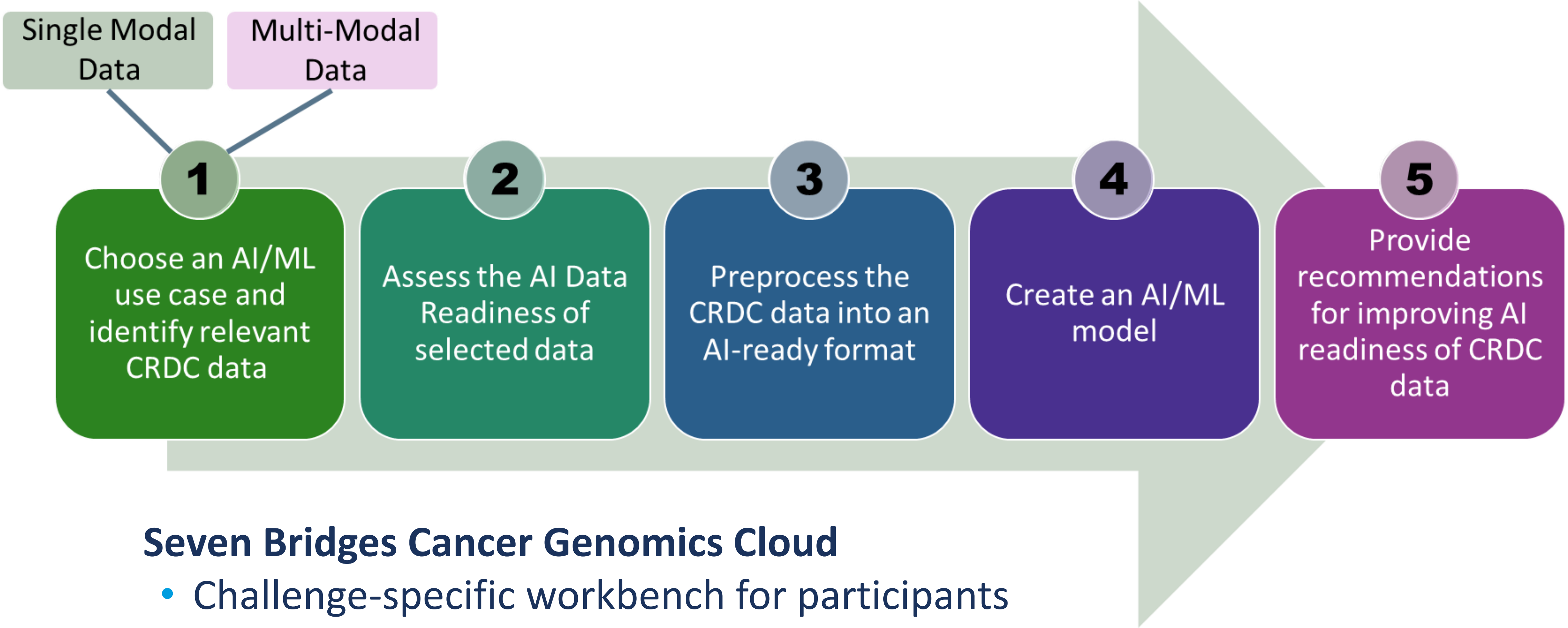
- **Purpose: Solicit broad community input on**
  - AI-readiness of data across multiple CRDC components
  - AI use cases for a CRDC AIDR Challenge
- **Example Questions:**
  - Identify the AI use case(s) that leverages CRDC or other cancer data
  - Describe high priority data types/elements for the use case
  - Describe any bias that you are aware of in the data for the use case
  - Describe data barriers/challenges encountered & improvement areas
- **Responses:**
  - Acellus Health, Certara, FDA, Jackson Lab & LBNL, MD Anderson, Northeastern, UMass Amherst, UPenn, Velsera

RFI: https://grants.nih.gov/grants/guide/notice-files/NOT-CA-24-017.html

RFI Use Cases for Challenge

#	Category	Purpose: “Build an AI/ML model to…”
1	Cancer Risk	Assess the risk of an individual developing a specific cancer type
2	Prevention	Distinguish early/pre-cancer from advanced disease
3	Diagnosis	Distinguish amongst different cancer subtypes
4	Diagnosis	Classify cancer cells versus healthy cells in a specific tissue
5	Prognosis	Predict survival in metastatic cancer
6	Prognosis	Predict cancer recurrence
7	Prognosis	Assess the risk of a tumor progression from benign to malignant, localized to metastasized, or one stage to another
8	Treatment	Predict the efficacy of a single or combination therapy
9	Treatment	Understand the relationship between the tumor microenvironment or immune response and cancer progression

AIDR Challenge Process


Single Modal Data
Multi-Modal Data
1
Choose an AI/ML use case and identify relevant CRDC data
2
Assess the AI Data Readiness of selected data
3
Preprocess the CRDC data into an AI-ready format
4
Create an AI/ML model
5
Provide recommendations for improving AI readiness of CRDC data

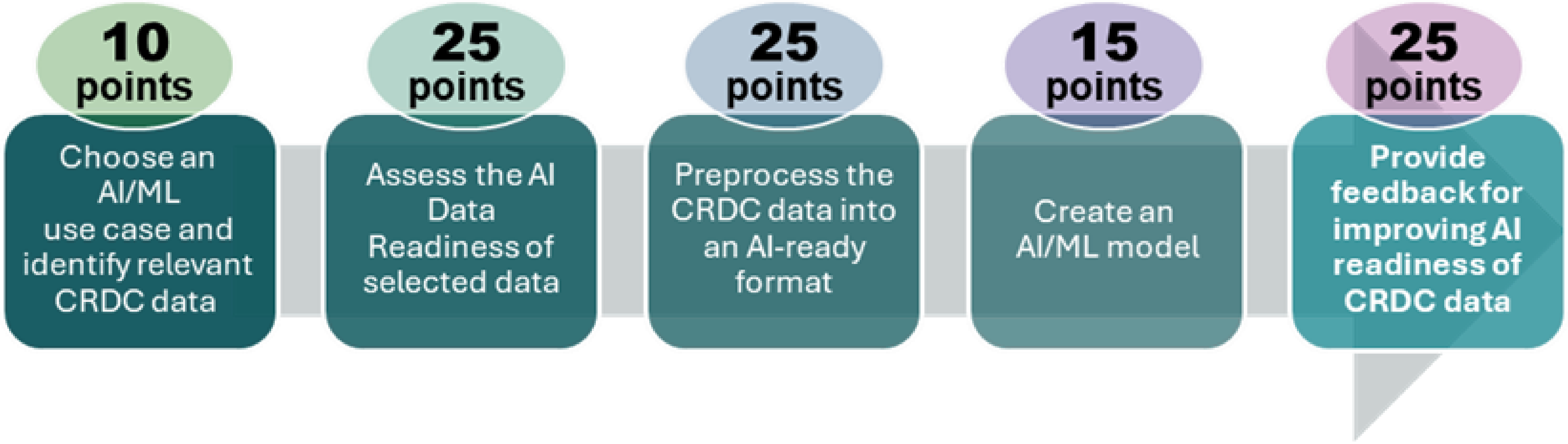
Seven Bridges Cancer Genomics Cloud

- Challenge-specific workbench for participants
- $300/1000 cloud credit per participant/team

Challenge.gov: https://www.challenge.gov/?challenge=ai-data-readiness-nci-challenge

AIDR Challenge Judging Criteria

Data-centric challenge, the AI model(s) were assessed for functionality not their performance quality



10 points
Choose an AI/ML use case and identify relevant CRDC data

25 points
Assess the AI Data Readiness of selected data

25 points
Preprocess the CRDC data into an AI-ready format

15 points
Create an AI/ML model

25 points
Provide feedback for improving AI readiness of CRDC data

AIDR Challenge Engagement

- **Registration Stats**
  - 50 groups registered to participate
- **Submission Stats**
  - Data accessed from 4 Data Commons
  - All teams used open access data
  - Single Modal = 14 projects

- 19 projects submitted for judging

- Multi-Modal = 5 projects
  - 2 used data from both GDC and PDC
  - 3 used data from a single Data Commons

AIDR Challenge Winners

Winning submissions satisfied challenge requirements and provided the most in-depth and insightful feedback regarding CRDC AI Data Readiness

$50,000 Total Prizes

Single Modal Data

1st Place: $15,000

Ruvos Health (Entity)
Jennifer Blasé (Lead)

2nd Place: $5,000

Agnes McFarlin (Individual)
No Affiliation

Multi-Modal Data

1st Place: $20,000

Abhishek Jha (Team)
Elucidata

2nd Place: $10,000

BAMF Health (Entity)
Jeff Van Oss (Lead)

AIDR Winners – Single Modal Data

2nd

Agnes McFarlin (Individual)

No Affiliation

Use Case: Identify cancerous lung nodules in DICOM images without the presence of annotated slides for reference

Commons: IDC **Study:** National Lung Screening Trial **Data Types:** CT DICOM images

AI Data Readiness Metrics:

- Data class imbalance, labeling, missingness, diversity, anonymization

Recommendations:

- Co-locate metadata with data and patient records
- Provide more instructions on how to use the REST API for specific queries
- Remove records of non-existent patients

AIDR Winners – Single Modal Data

1st

Ruvos (Entity)

Jennifer Blasé (Lead)

Use Case: Gene expression-based prediction of treatment response in ovarian cancer

Commons: GDC **Study:** TCGA Ovarian **Data Types:** RNA-Seq, Clinical Biospecimen

AI Data Readiness Metrics:

- Data quality, accessibility, quantity

Recommendations:

- Documentation on how to process the files
- Create Unified Modeling Language diagram showing the database schemas
- Provide examples of queries and use cases

AIDR Winners – Multi-Modal Data

2nd

BAMF Health (Entity)

Jeff Van Oss (Lead)

Use Case: Predict Von Hippel-Lindau (VHL) mutations in kidney tumors using radiomic features

Commons: GDC, IDC **Studies:** TCGA Kidney **Data Types:** CT, BAM, Somatic mutation

AI Data Readiness Metrics:

- Data comprehensiveness, completeness, size, variety of sources

Recommendations:

- Implement and enforce robust data quality assurance processes
- Standardize data formats, integrate relevant metadata for comprehensiveness
- Continuously monitor data quality and incorporate feedback loop

AIDR Winners – Multi-Modal Data

1st

Abhishek Jha (Team)

Elucidata

Use Case: Distinguish primary tumor from normal solid tissue in lung squamous cell carcinoma using transcriptomics and proteomics data

Commons: GDC, PDC **Studies:** CPTAC Lung SCC **Data Types:** RNA-Seq, Proteomics

AI Data Readiness Metrics:

- Access, class imbalance, missing data, confounding variables, normalization

Recommendations:

- Provide download on a per sample basis, a consistent schema for metadata
- Add more case/sample identifiers in file names and API query results
- Develop a common portal for querying and visualizing data across DCs

RFI and Challenge Recommendations

- Allow download of smaller subsets of data files*
- Co-locate metadata with data and patient records
- Provide examples of AI queries and use cases
- Adopt a federated learning framework for integration of de-identified CRDC data with multi-institutional data that provides clinical context
- **Across Data Commons**
  - Standardize data formats, naming conventions, and metadata schemas*
  - Develop a common portal for querying and visualizing data*
  - Adopt a schema crosswalk for discovery between different metadata standards*
- **Across NCI Cloud Resources**
  - Develop “resource packages” with data/toolsets for specific research activities

*** Already in progress**

Acknowledgements

- Booz Allen AIDR Team
  - Carolyn Hetzer
  - Abdullah Awaysheh
  - Lucy Han
  - Elise Berning
  - Anna Fernandez
  - Zeke Maier
  - Nirmal Agarwal
  - Jason DeChancie

- CBIIT Federal Team
  - Emi Casas-Silva
  - Granger Sutton
- NIH Office of Data Science Strategy
- RFI AIDR Respondents
- CRD AIDR Challenge Participants & Winners