Health Data Exchange (HDX): A GLOBAL Open Data, Software, and Learning Community Resource for R & D Acceleration and Dissemination

Javed Mostafa
Dean & Professor
University of Toronto
ischool.utoronto.ca

OUTLINE

Introduction

Why data exchange?

Major goals

System architecture

Sustainability

- FAIR
- Metadata Annotation

Summary

INTRODUCTION

In recent times the volume of digital data generated has skyrocketed to about 2.2 Exabytes per year

1 Exabyte = 1 Billion Gigabyte, 1 GB is about 4500 books, 200 pages each

Of the total volume about 30% of it is healthcare data – about 0.7 Exabyte ~ 1 Billion Gigabyte by 2025 (at current growth rate)

Health data generation grows at a faster pace than any other major sectors (e.g., manufacturing, financial, media and entertainment)

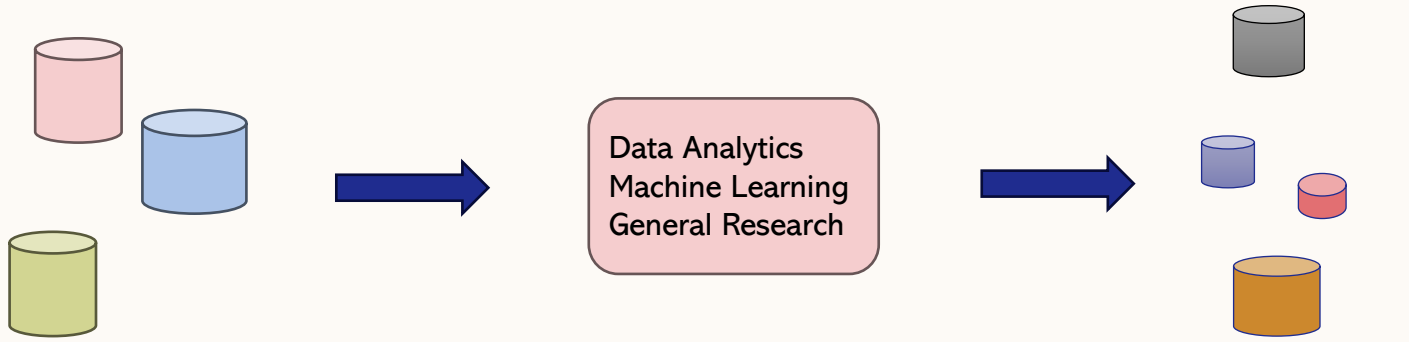
But, due to many security, IP, and bureaucratic reasons accessing and mining healthcare data is exceedingly difficult (even inside an organization)

https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion

WHY HEALTH DATA EXCHANGE (HDX)? REASON 1

Data analytics, machine learning, and a majority of research in biomedical and health sciences

Require ACCESS TO DATA and produce DATA



WHY HEALTH DATA EXCHANGE (HDX)? REASON 2

Very hard to gain access to data in a

- Consistent,
- Persistent,
- Standardized,
- Regulatory-governed, and
- Sustainable way

WHY HEALTH DATA EXCHANGE (HDX)? REASON 3

Post-analysis or post-research, data still has value

- Publication/dissemination
- Replication
- Training/pedagogy
- Follow-up or scale-up work
- Community- or team-building

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

MAJOR GOALS OF HDX

A phased design, development, and sustainable approach

AREAS OF FOCUS

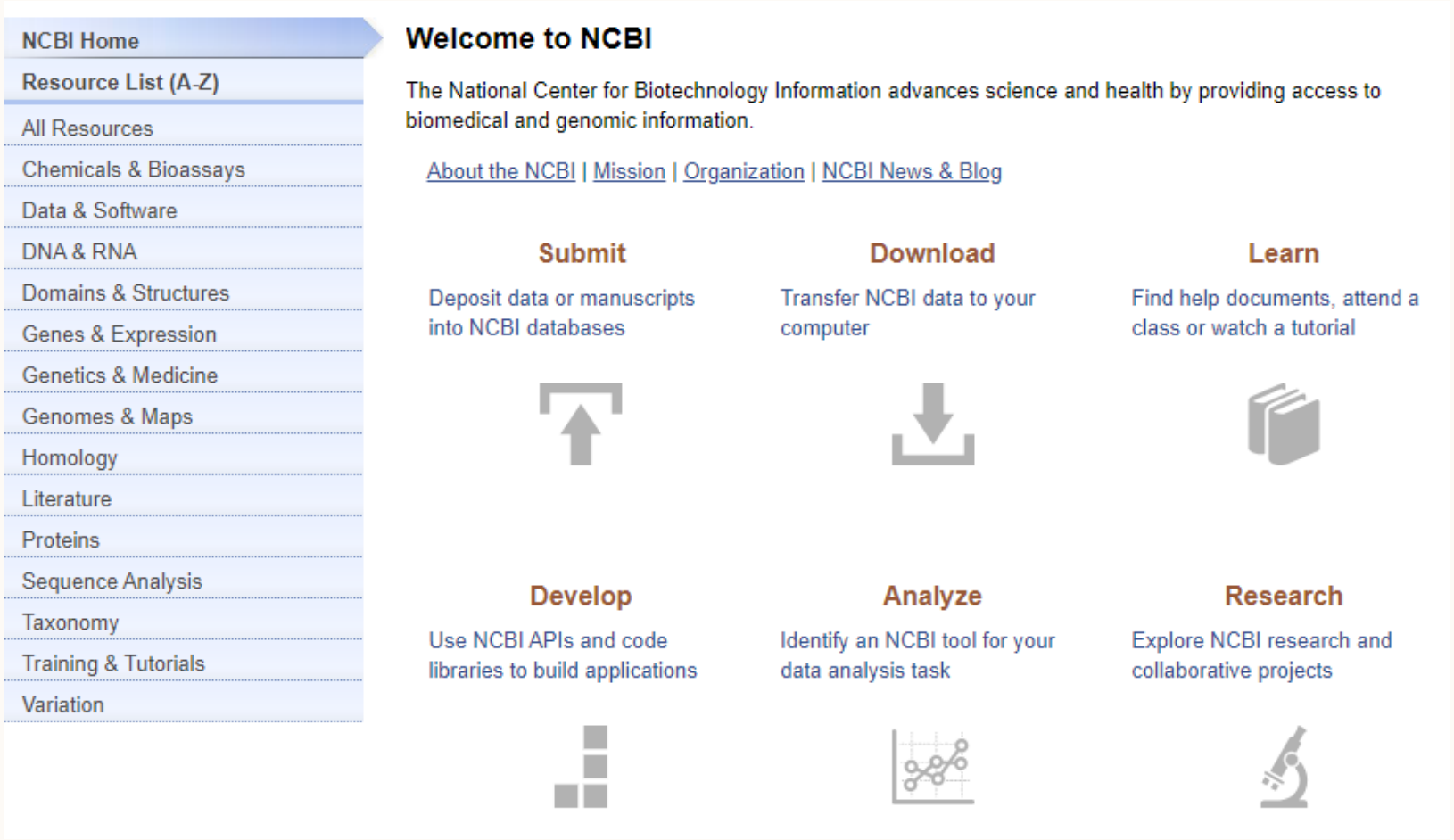
SUPPORT RESEARCHERS

- Findability of data
- Accessing data
- Using data
  - Support software pipeline and scalable computing
  - Support secure and audit-trackable use
  - Support dissemination of results
- Contribution of data from researchers
- Updating and retirement of data

SUSTAINABLE PLATFORM

- Community-building
  - Linking to publications
  - Linking publications
  - Linking researchers
- Training and pedagogy
- Partnerships with publishers, non- and for-profit organizations

EXAMPLE OF PEER RESOURCES: NCBI



https://www.ncbi.nlm.nih.gov/

EXAMPLE OF PEER RESOURCES: NATIONAL COVID COHORT COLLABORATIVE OR N3C

N3C Knowledge Store

Access and secure sharing platform for analysis templates and derived or ingested publicly available datasets.

N3C Concept Set Browser

Browse curated medical code sets, including peer-review validation information.

Data Catalog

Browse the most commonly accessed data tables as well as notional data for learning.

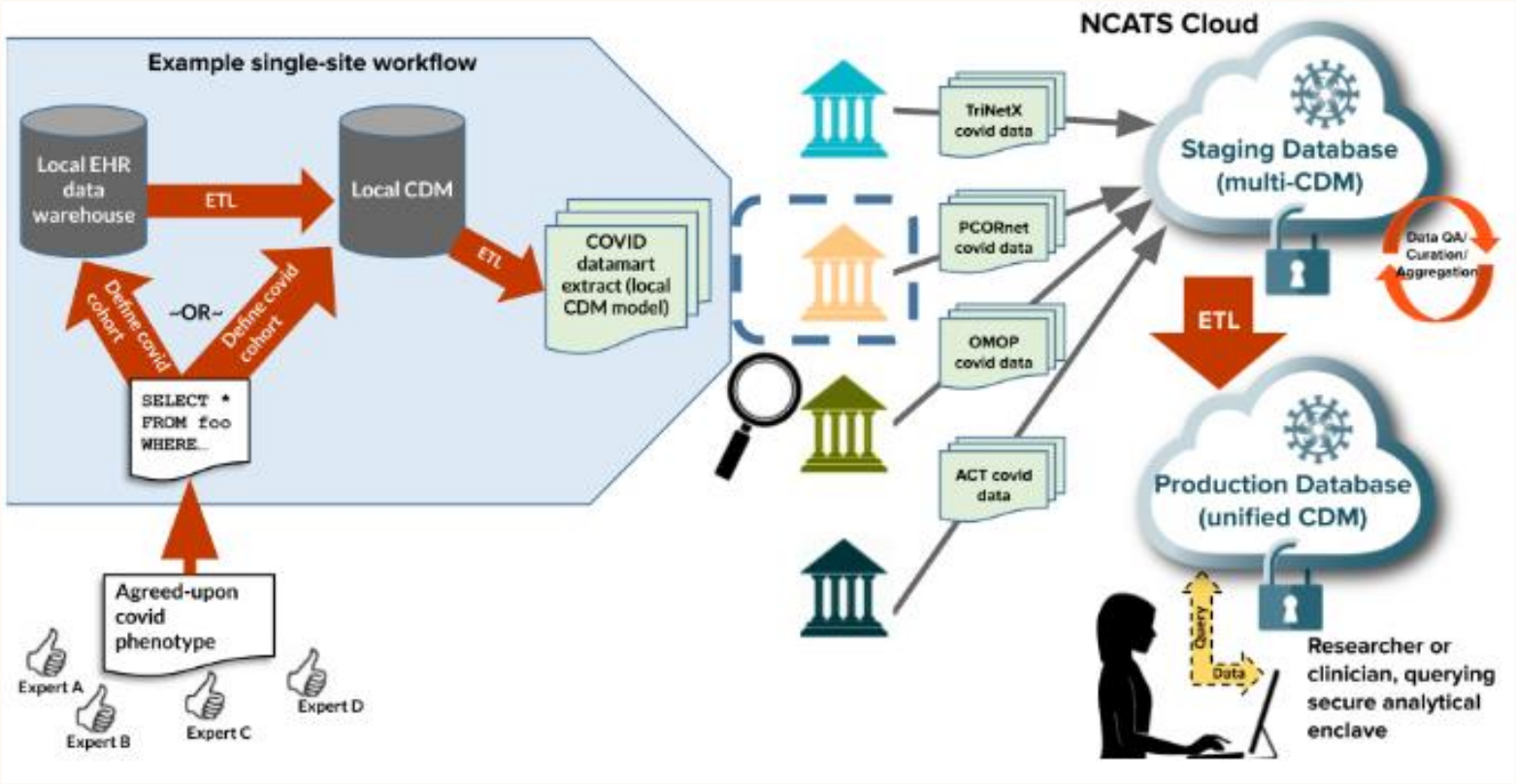
Community Notes

The N3C Community Notes application provides a place for users to asynchronously share documentation, tips, and links to other resources with the N3C community.

NCATS National COVID Cohort Collaborative

Importantly, since the shared data could be partly identifiable, the data is stored in a secure platform, with strictly regulated usage. Researchers who meet the requirements (https://ncats.nih.gov/n3c/about/applying-for-access) can apply for access to the data, for use only for COVID research and response planning.

DETAIL VIEW OF HEALTH DATA SHARING: USE CASE N3C



https://ncats.nih.gov/n3c/about/data-overview

GOALS FOR HDX

FINDABILITY

- Resources need to be found quickly and easily

ACCESSIBILITY

- Resources need to be searchable based on standardized terminologies and common words
- Retrieved results should be easily browsable and understandable

USE

- Results should be downloadable
- Tools should allow the user to analyze the data using scalable compute resource
- Output should be easily storable, recoverable, and sharable
- Allow contribution of new data
- Retirement of access and data

COMMUNITY

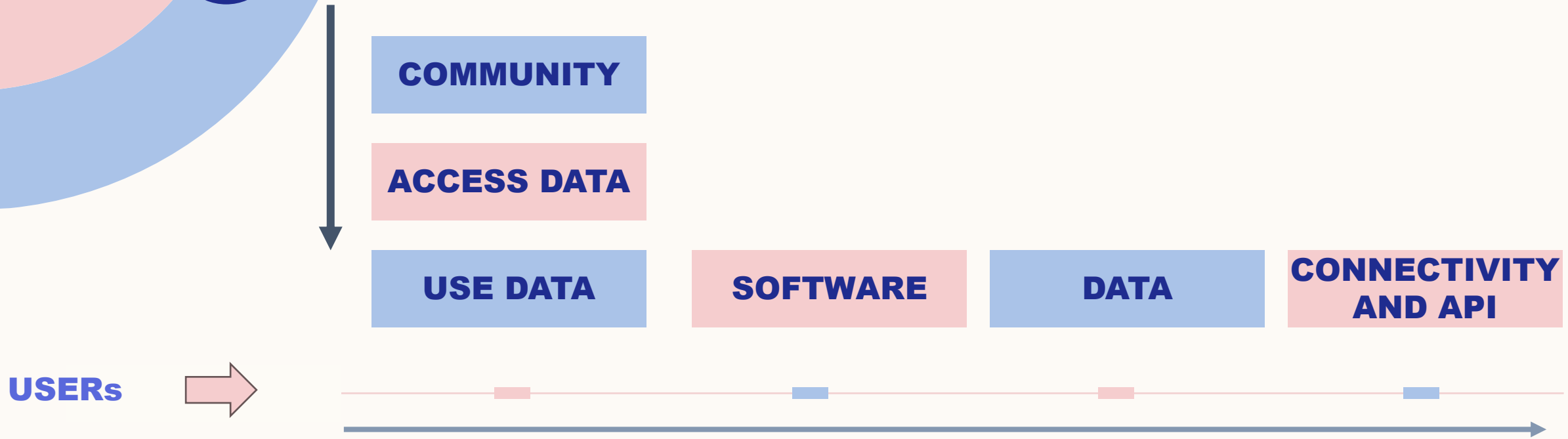
- Support finding the contributors /producers of the data and communicating with them
- Support tools for identifying teams and resources
- Provide tools for conducting meetings, access relevant resources, and store outcomes of meetings/
- Platform for collaborations

TRAINING/ LEARNING

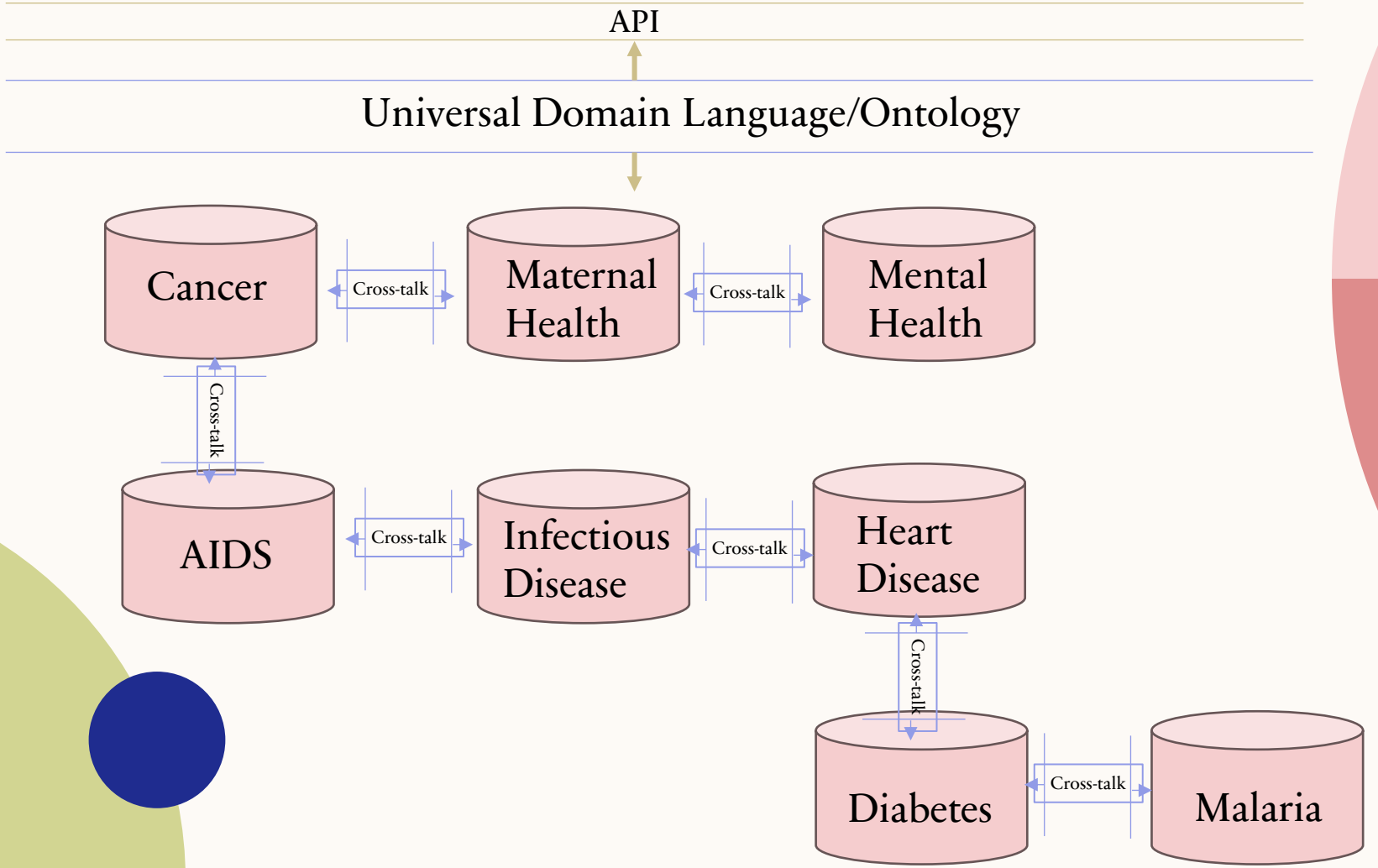
- Support tools for accessing and using data for teaching purposes
- Provide documentation and training materials and link appropriate data to the learning resources
- Provide a way to link external learning resources to internal data

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

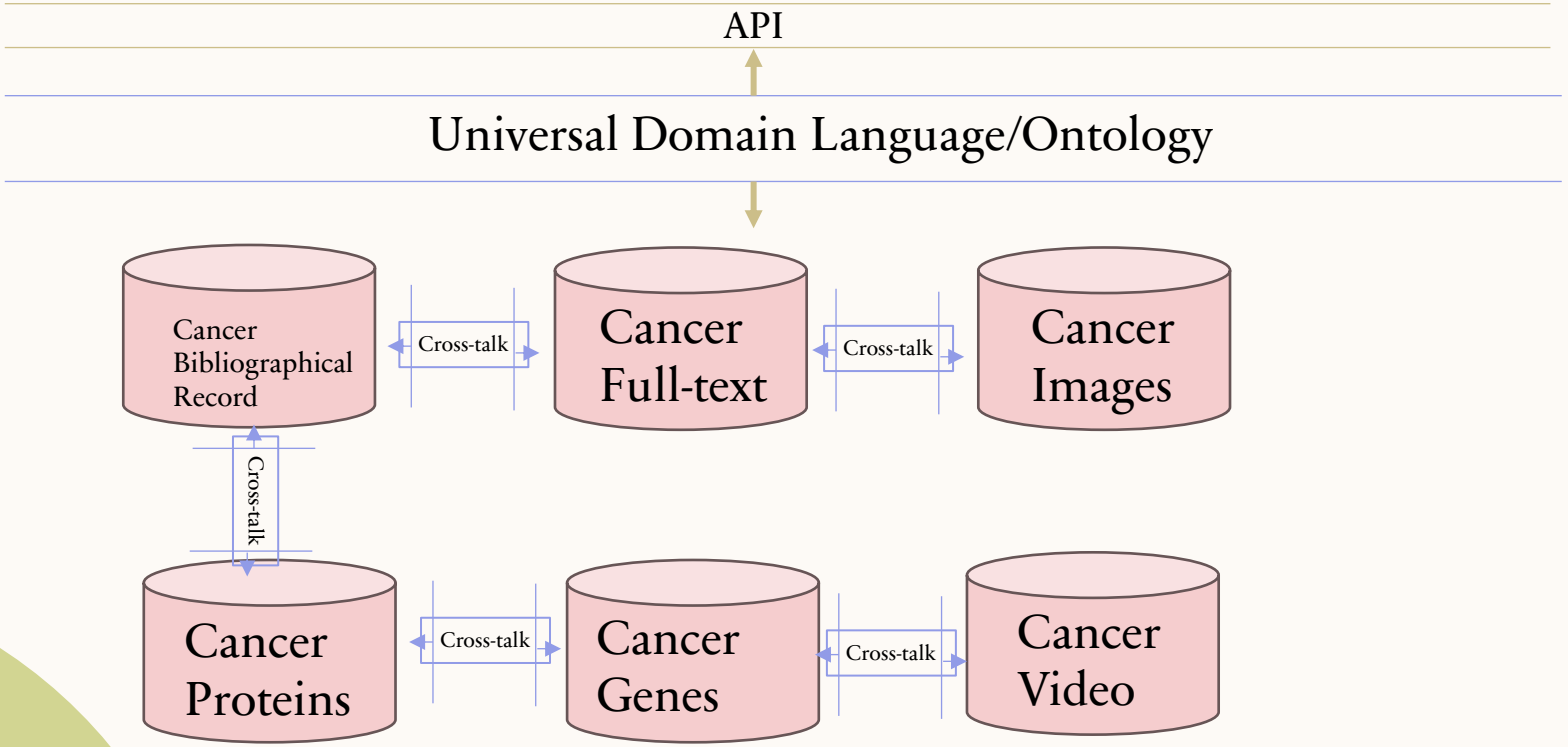
HDX ARCHITECTURE



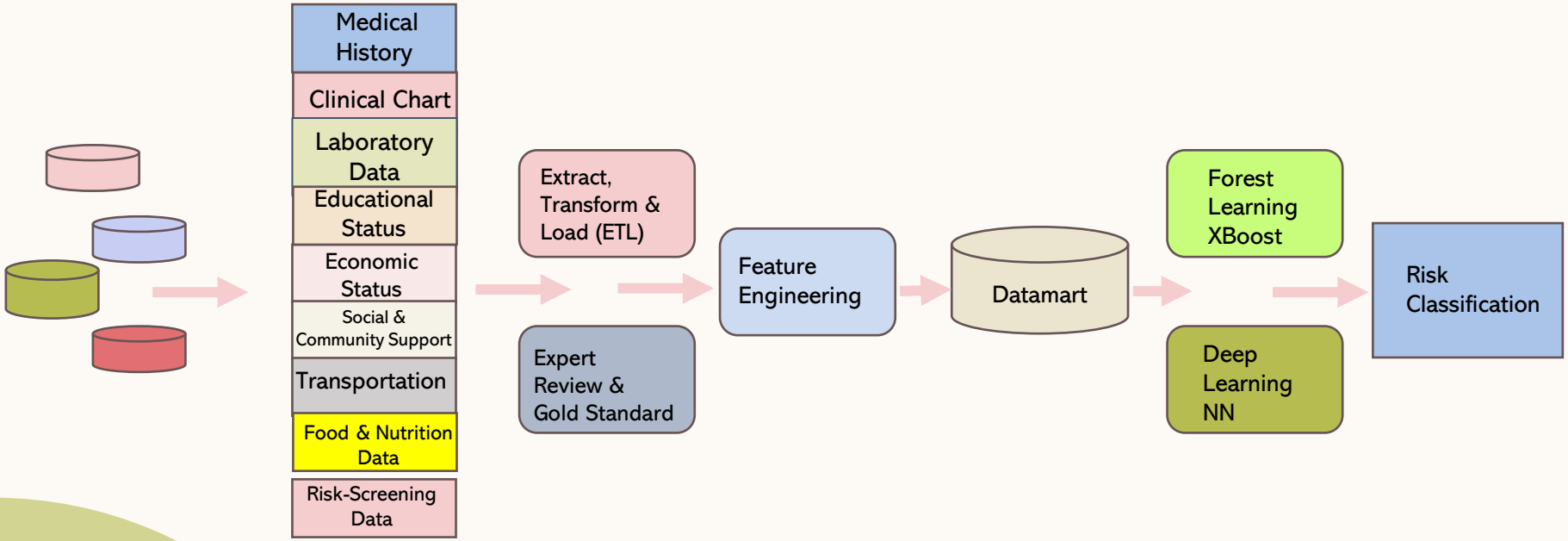
HDX ARCHITECTURE: DATA CATALOG I



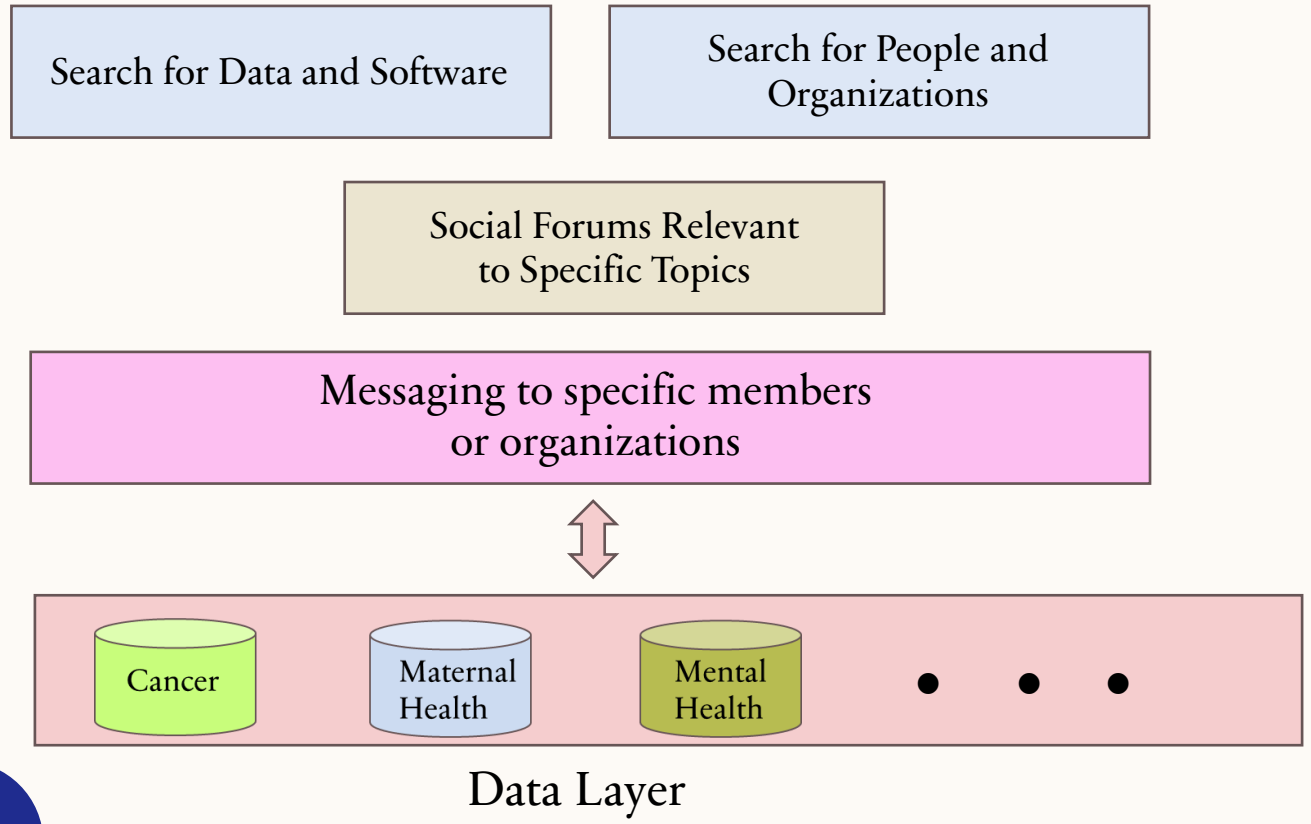
HDX ARCHITECTURE: DATA CATALOG II



HDX ARCHITECTURE: SOFTWARE PIPELINE I



HDX ARCHITECTURE: COMMUNITY



SUSTAINABILITY I

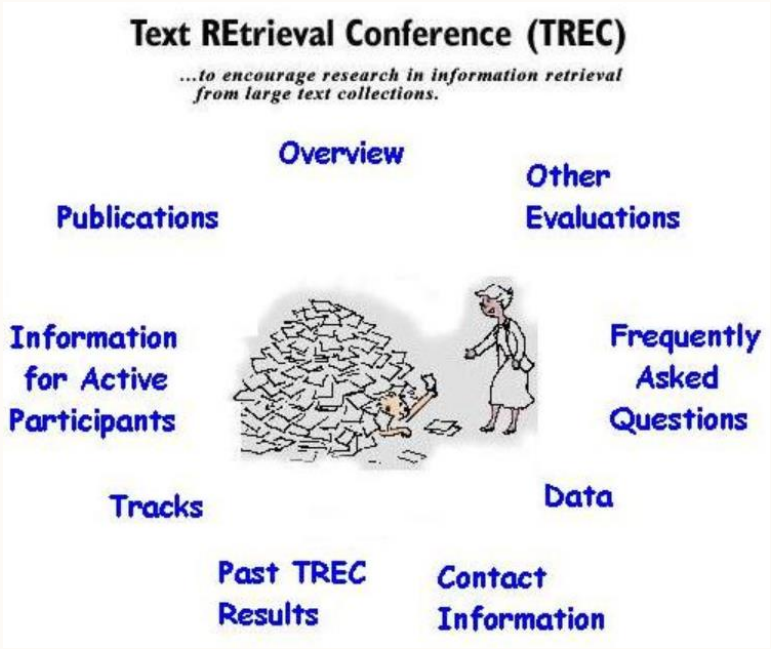
USER COMMUNITY

- Provide organizational and domain-based directories for the user community to find collaborators
- Support discovery of other researchers and experts who have association with data sets (produced the data or used the same data set before)
- Allow users to identify relationships between domains and members of the research/expert community
- Support functions such as posting of blogs/comments and securely message members of the community

SUSTAINABILITY II

USER COMMUNITY

- Create "Gold Standards" for different standardized machine learning predictive models and classifiers
- Support "Friendly but Rigorous Competitions" on an annual basis to promote awareness and develop closer bonds among the users of data and software
  - A nice model is US NIST TReC Data Sets and the user community which NIST nurtures



trec.nist.gov

SUSTAINABILITY III

PARTNERSHIPS & BUSINESS MODEL

- Partner with research centers/universities, publishers, NGOs, non-profit and for-profit research organizations, and networks of current data hubs (e.g., NCBI)
- Shared IPs with partners on advanced data access and manipulation tools
- Establish a non-profit status for HDX but have a solid business model

FINDABILITY, ACCESSIBILITY, INTEROPERABILITY, AND REUSABILITY (FAIR)

Developed by group of scientists and practitioners from organizations worldwide [2016]*

Easy to appreciate if you consider the original emphasis was on "machine" discovery, access, and use of data

- Finding and discovery: Improving visibility of content using identifiers
- Accessibility: Terminology and ontology standardization
- Interoperability: Semantic integration
- Reusability: Securing data and use of secure protocols

METADATA ANNOTATION

FINDABLE

- Unique identifiers and metadata for locating data

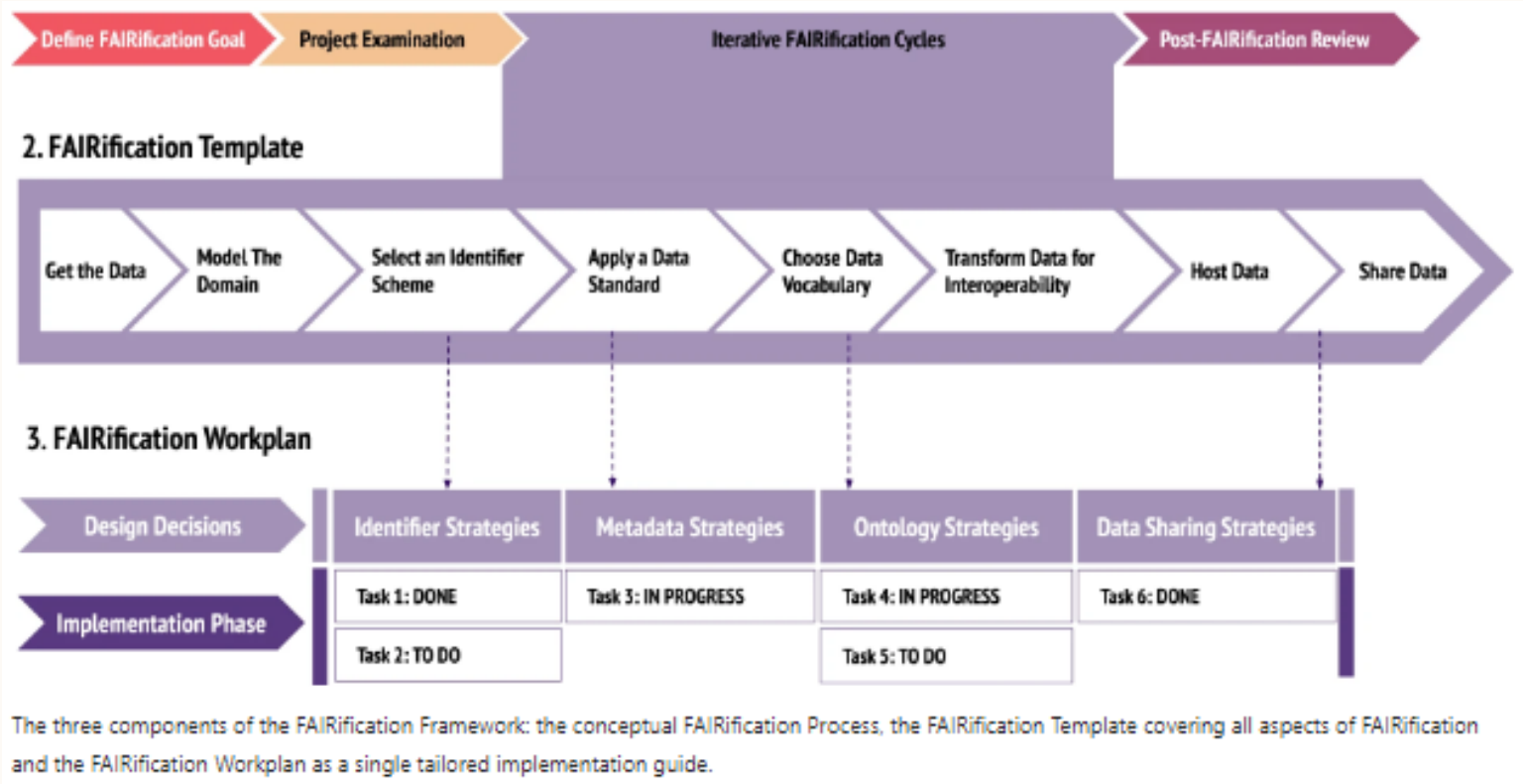
ACCESSIBLE

- Metadata freely available and accessed in a standardized way

INTEROPERABLE AND REUSABLE

- Source, origin, destination, and usage of data over time and across disciplines
- Secure and reliable access

FAIRIFICATION PROCESS



The three components of the FAIRification Framework: the conceptual FAIRification Process, the FAIRification Template covering all aspects of FAIRification and the FAIRification Workplan as a single tailored implementation guide.

https://www.nature.com/articles/s41597-023-02167-2

SUMMARY

- Data, software, and people information need to be aggregated on a single platform
- Findability, accessibility, manipulability, and sociability are key dimensions
- Must support pipelining of software to promote usage and replication
- Sustainability can be ensured through community and partnership building
- Must support FAIR principles

THANK YOU

Javed Mostafa
dr.javedm@utoronto.ca
ischool.utoronto.ca