

# Online Data Research Repositories

From Research Data and Datasets to Artificial Intelligence and Discovery

ORCID iD  
DISPLAY

The Dataverse Project  
Online Research Data Repository

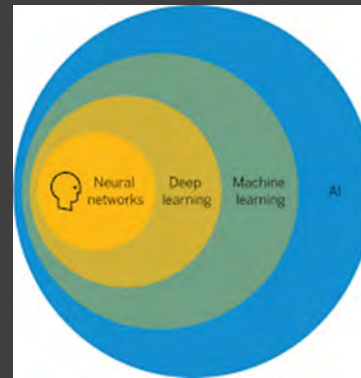
Identity Management System

VIREO  
Electronic Theses and Dissertation (ETD) Management System

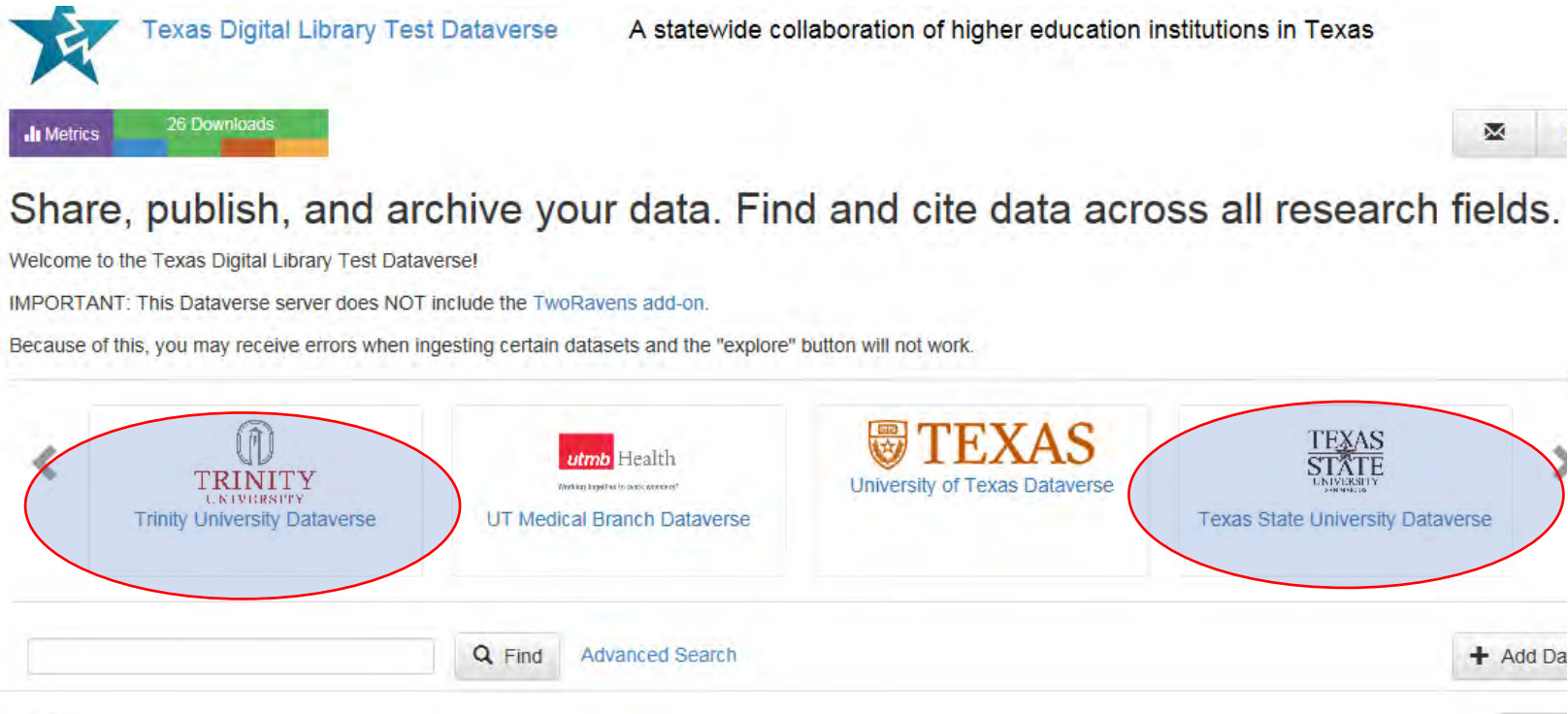
DSPACE  
Online Institutional Digital Collections Repository

OJS  
OPEN JOURNAL SYSTEMS

omeka.net



# What is an Online Data Research Repository?



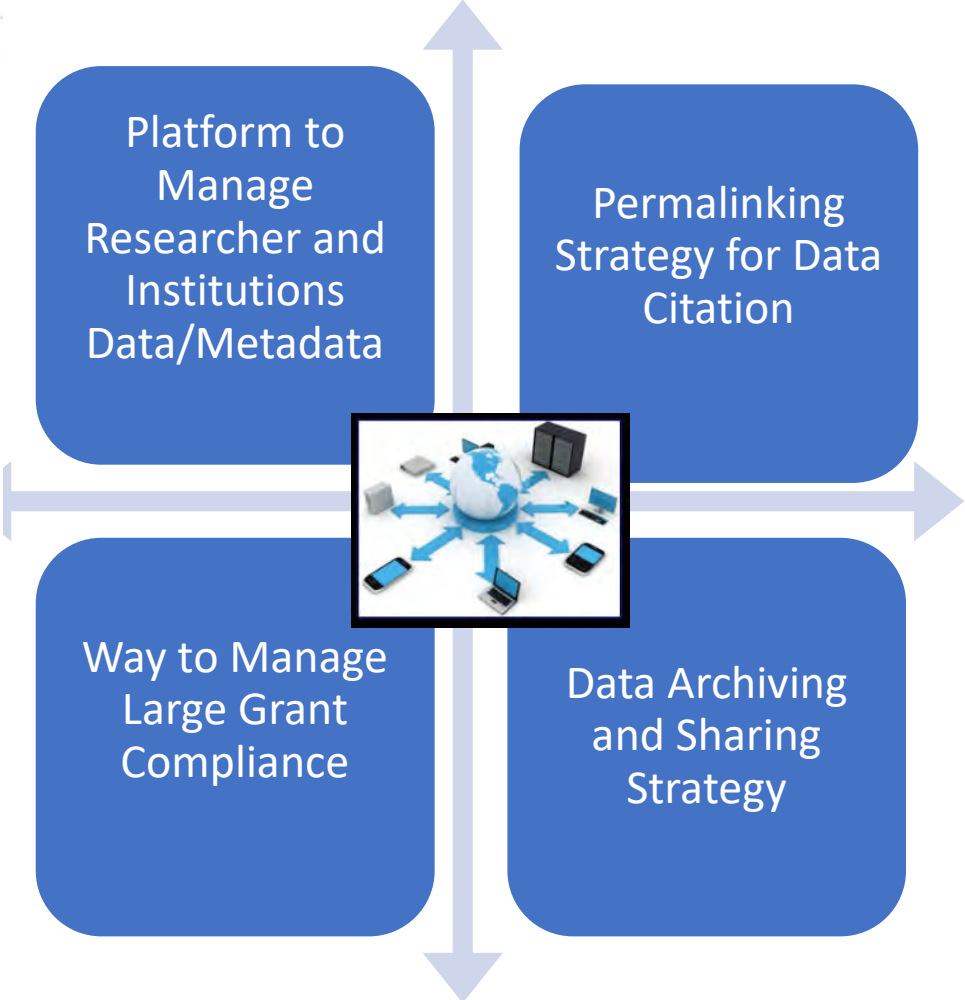
The screenshot shows the homepage of the Texas Digital Library Test Dataverse. At the top left is the logo for the Texas Digital Library Test Dataverse, a blue star with a white 'T' and 'D'. To its right is the text "Texas Digital Library Test Dataverse" and "A statewide collaboration of higher education institutions in Texas". Below this is a "Metrics" section showing "26 Downloads". The main heading reads "Share, publish, and archive your data. Find and cite data across all research fields." Below this is a welcome message and an important notice: "IMPORTANT: This Dataverse server does NOT include the TwoRavens add-on. Because of this, you may receive errors when ingesting certain datasets and the 'explore' button will not work." A row of four data repository logos is displayed, with the first and last ones circled in red: Trinity University Dataverse, UT Medical Branch Dataverse, University of Texas Dataverse, and Texas State University Dataverse. At the bottom, there is a search bar with "Find" and "Advanced Search" buttons, and an "Add Data" button.



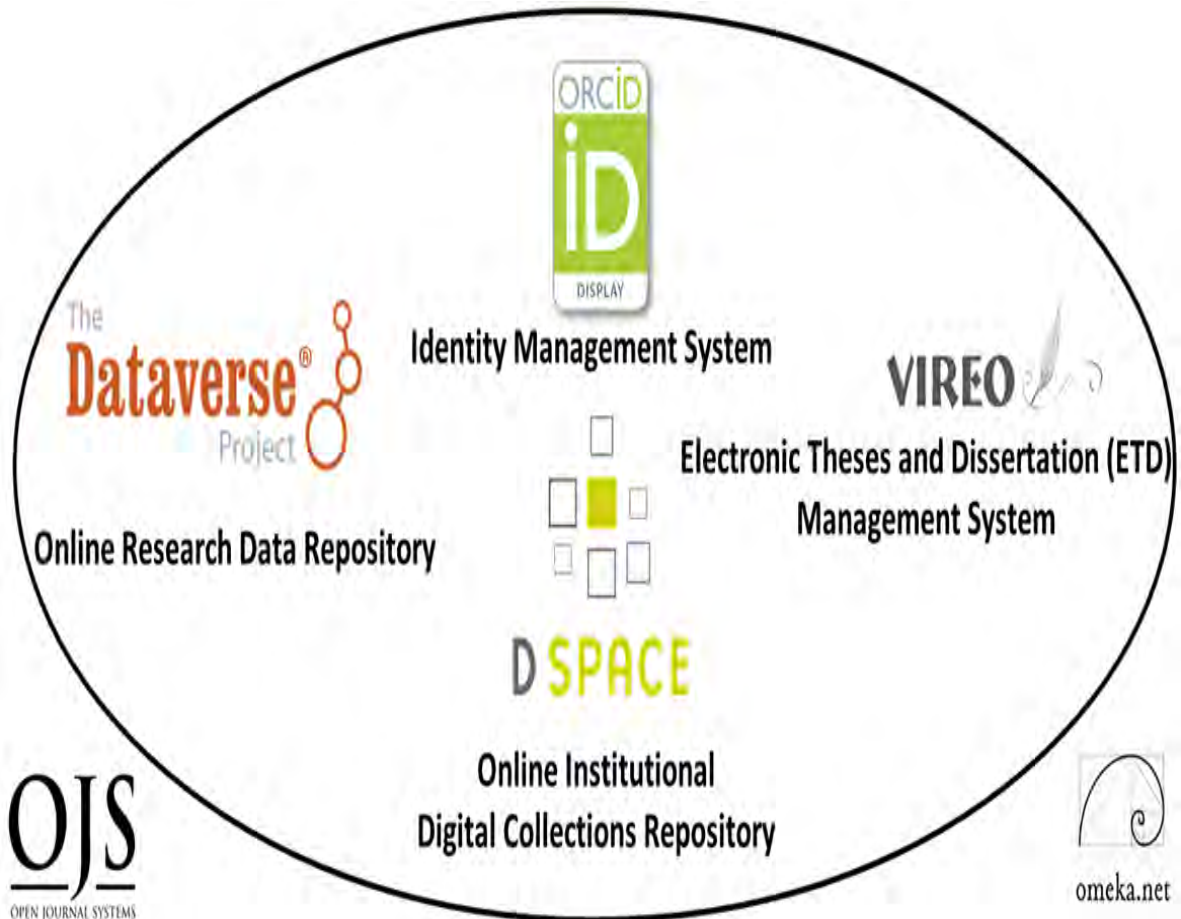
Texas Data Repository which is a shared repository of several Texas Universities leveraging technological cooperation and expertise among academic research libraries libraries, <https://dataverse.tdl.org>

# What is the Utility of An Online Research Data Repository?

The screenshot shows the Texas Data Repository website. At the top left is the logo with a star and the text "Texas Data Repository". To the right are navigation links: "About", "Documentation", "FAQs", "Log In", and "Help". Below this is a search bar with the text "Search the Texas Data Repository" and a "FIND" button. Underneath the search bar are five icons with labels: "Add a Dataset" (document with arrow), "Create a Dataverse" (stack of papers), "Explore Data Repository" (line graph), "Learn More" (document with magnifying glass), and "Get Help" (speech bubbles). At the bottom, it says "Publish and Track Your Data, Discover and Reuse Others' Data!" followed by the "POWERED BY Dataverse" logo. The URL <https://dataverse.tdl.org/> is at the bottom.



# A Data Repository May Also Be Placed Within a Larger Digital Scholarship Research Ecosystem



## TWO PRIMARY COMPONENTS (Content)

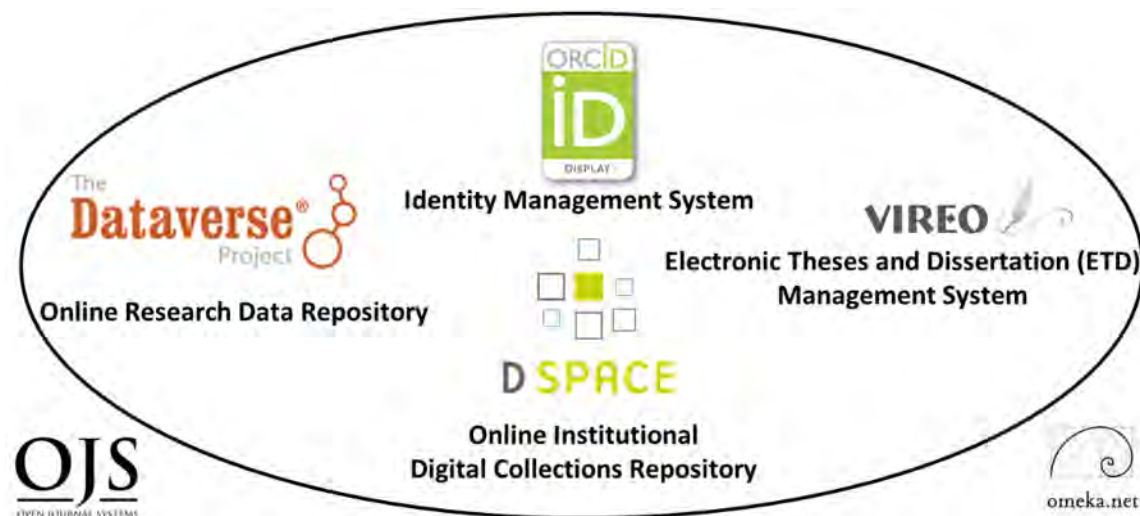
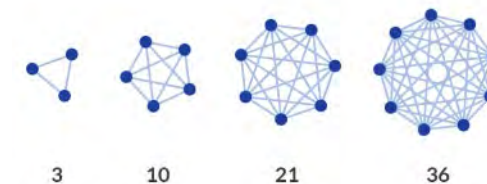
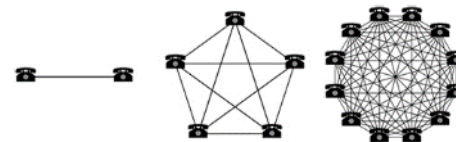
- **RESEARCH DATA REPOSITORY**
- **DIGITAL COLLECTIONS REPOSITORY**

## FOUR TERTIARY COMPONENTS (Communication)

- Electronic Thesis and Dissertation Management System
- Identity Management System
- Open Academic Journal Software
- User Interface/Content Management Software



# Collocating Open Source Digital Components in a Networked Research Ecosystem Enables Larger Connections and/or Network Effects



# Together These Digital Ecosystem Components Enable the Academic Research Cycle



Pragmatic Levels

# One Size Does Not Fit All for Various Data Research Repository Project Needs

## Many Types of Data Projects (Sizes)

### 1) Normal range (<4GB Files <10GB Datasets)

Files/Data Fit on Server/Cloud, may be uploaded to the Data Repository, 4GB files, 10GB Datasets)

### 2) Large Projects, Bigger Data <TB

(Data may require specialized university IT Support, i.e. terabyte/petabyte tape drives, Pointers, Checksums)

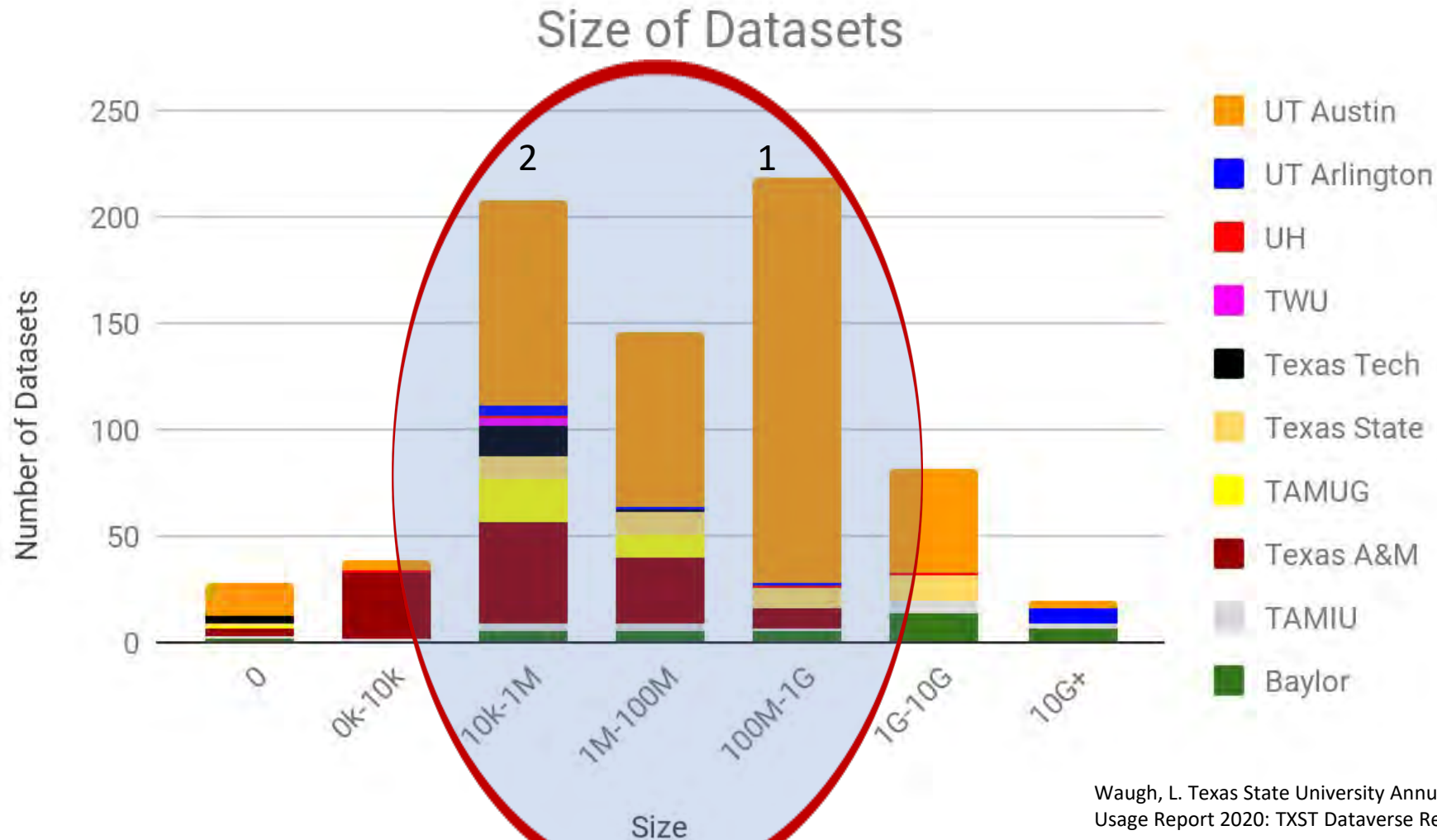
### 3) Huge Projects, Big Data

(Projects require consortial possibilities, national models, **Texas Advanced Computer Center TAAC**, LyraSIS, Duracloud, AWS S3, Custom Solutions)



# Present Sizes of Texas Data Repository Datasets

Most 1MB <1GB, Greater than 10 GB+ Rare



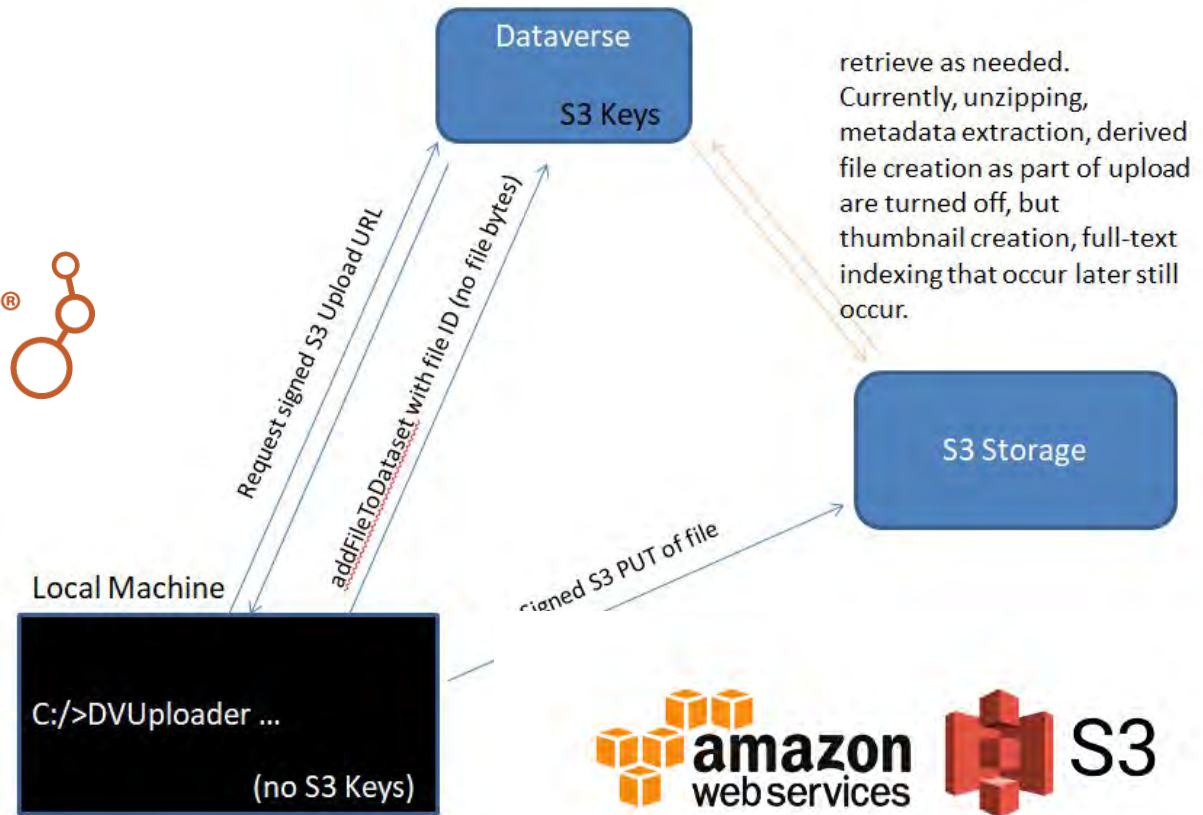


# Beta Prototyping Big & Bigger Data Options

2020-2022

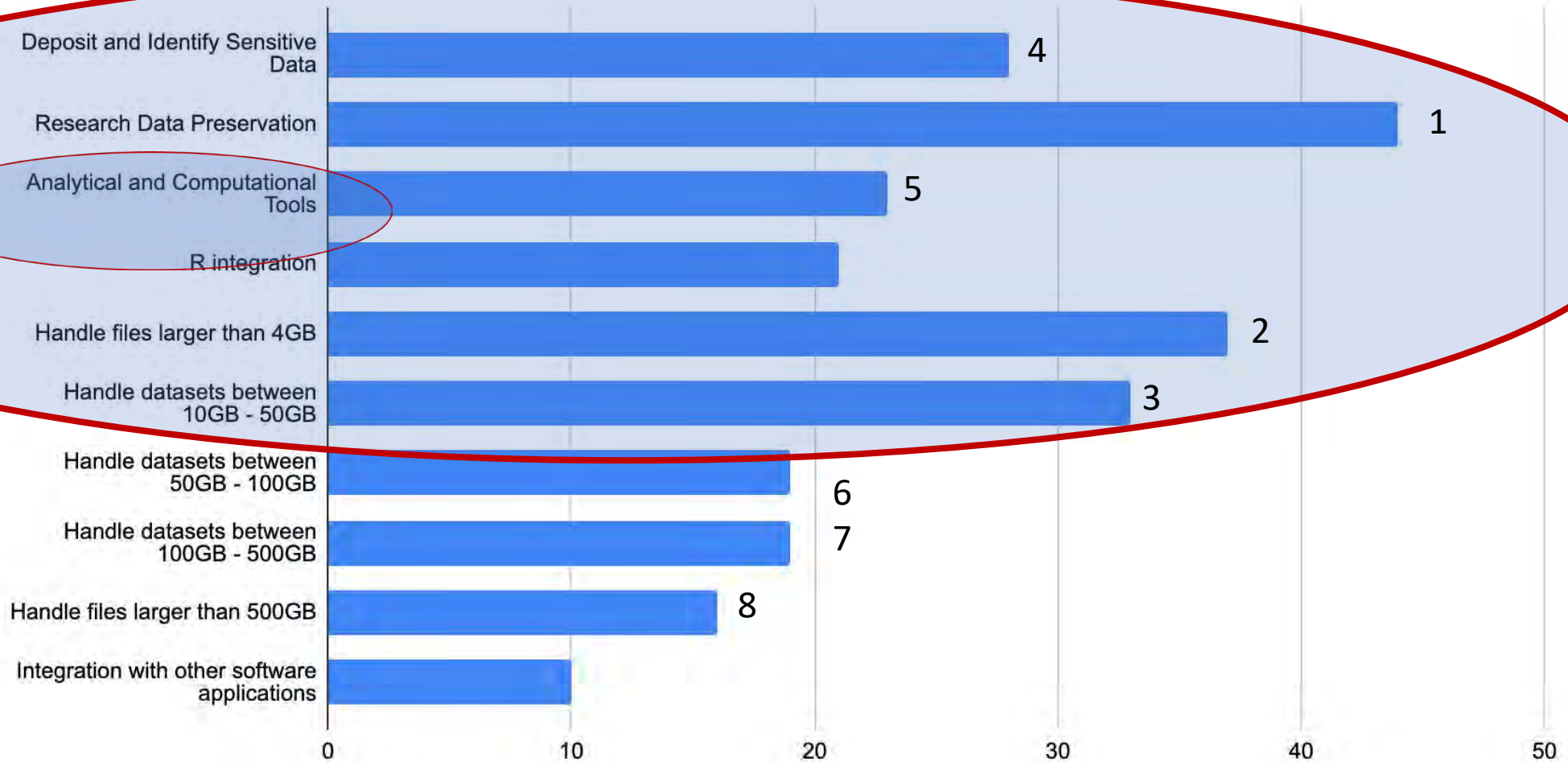


Up to 300 GB/dataset  
Fee Based Institutional Model 7.5/13.5 K/Year



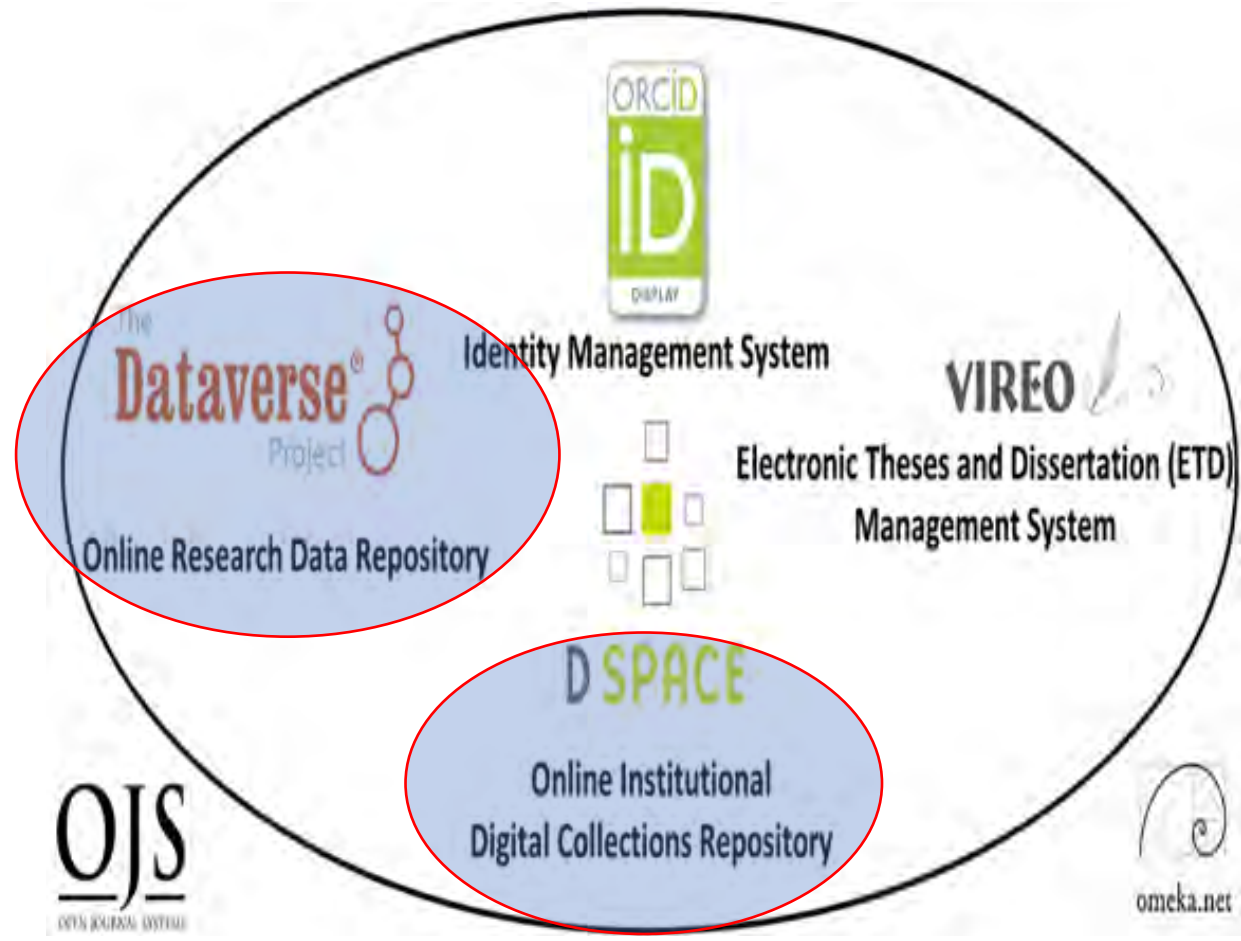
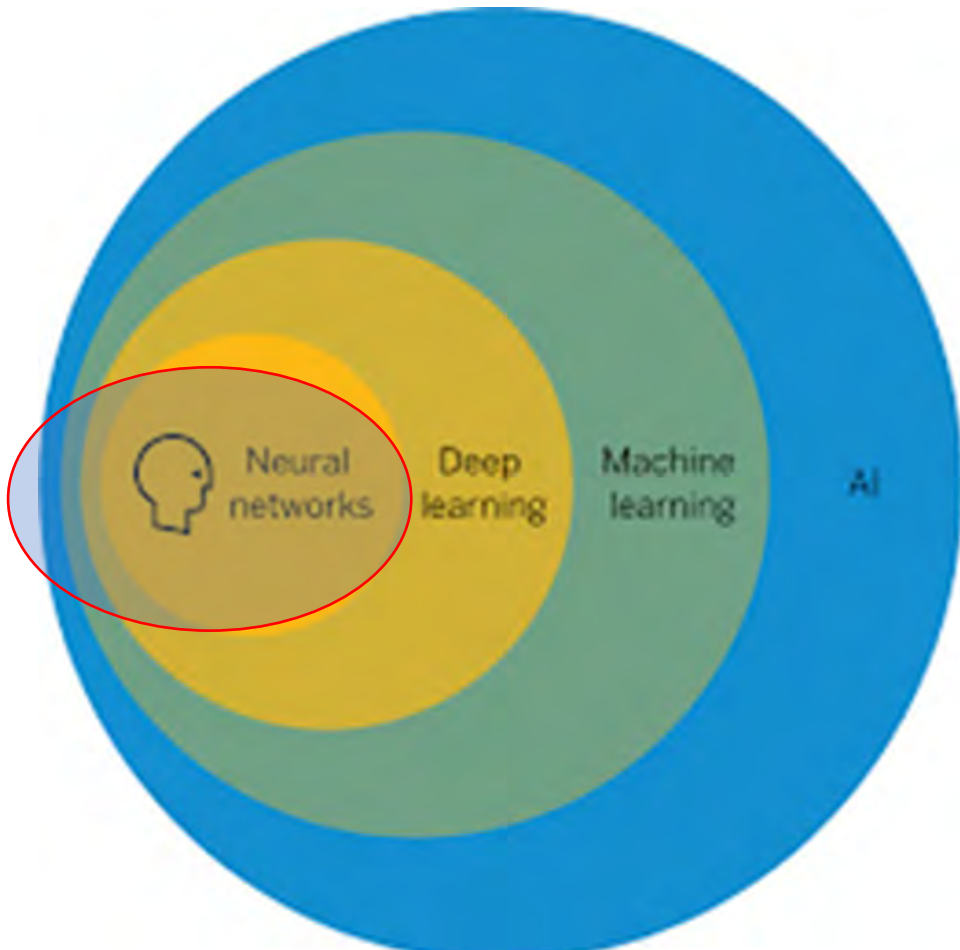
<20 GB Upload  
(Download Challenges)

# What New Data Repository Features Would Users Like to See in 2022?



# Last Five Years Has Shown Incredible Progress of, Analytical Computational Tools, Particularly, AI

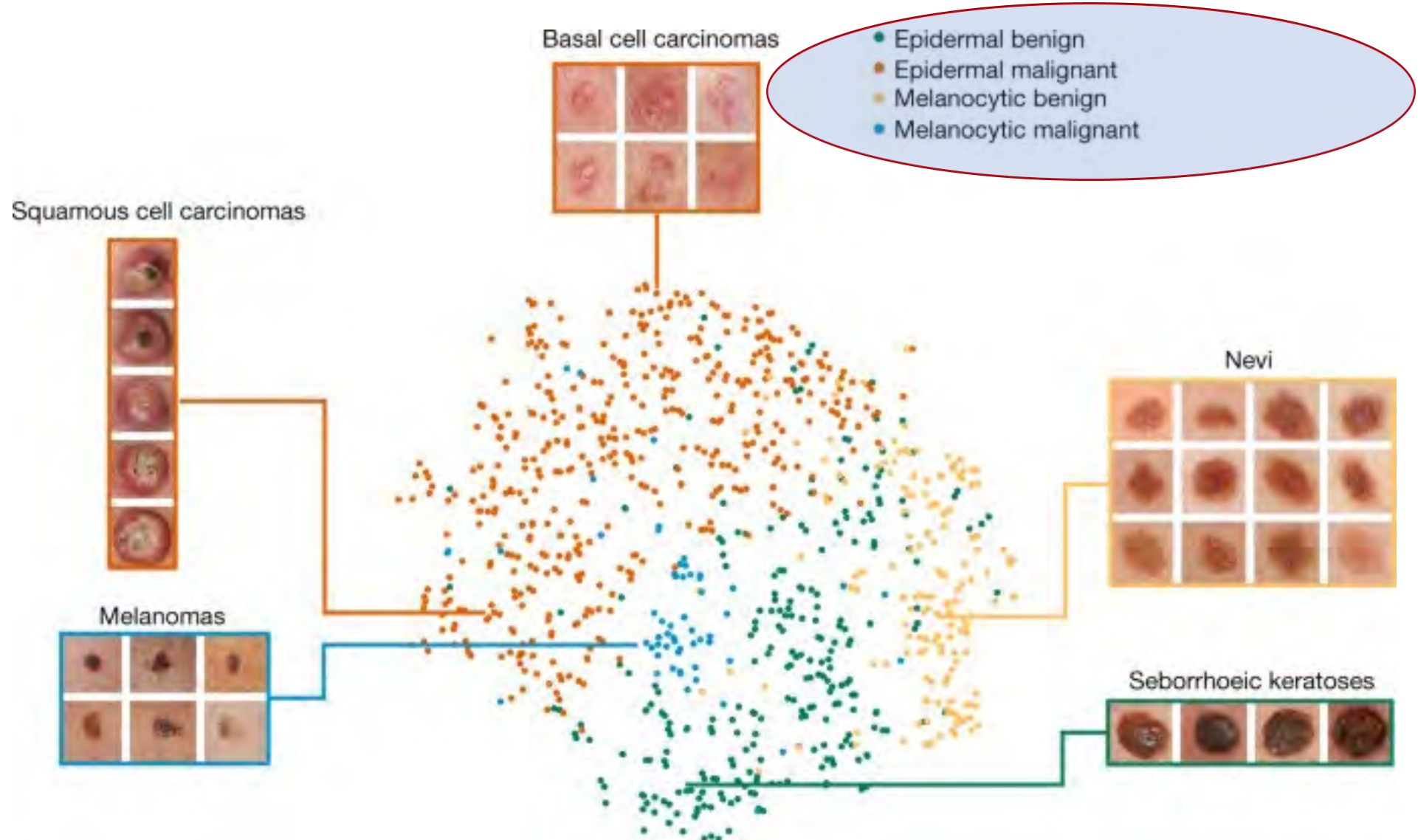
Machine Learning, Deep Learning, Computer Vision, Object Recognition, Cancer Detection





# Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks

2017, Nature, Esteva, Thrun et Al





# Data Research Repository Upload

Open Science Dermatology Image Dataset, Dr. Philip Tschandl

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>

## The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions

Version 3.0



Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", <https://doi.org/10.7910/DVN/DBW86T>, Harvard Dataverse, V3.  
UNF:6:APKSSDGVDhwPBWzsStU5A== [fileUNF]

Cite Dataset ▾

[Learn about Data Citation Standards.](#)

Access Dataset ▾

Contact Owner

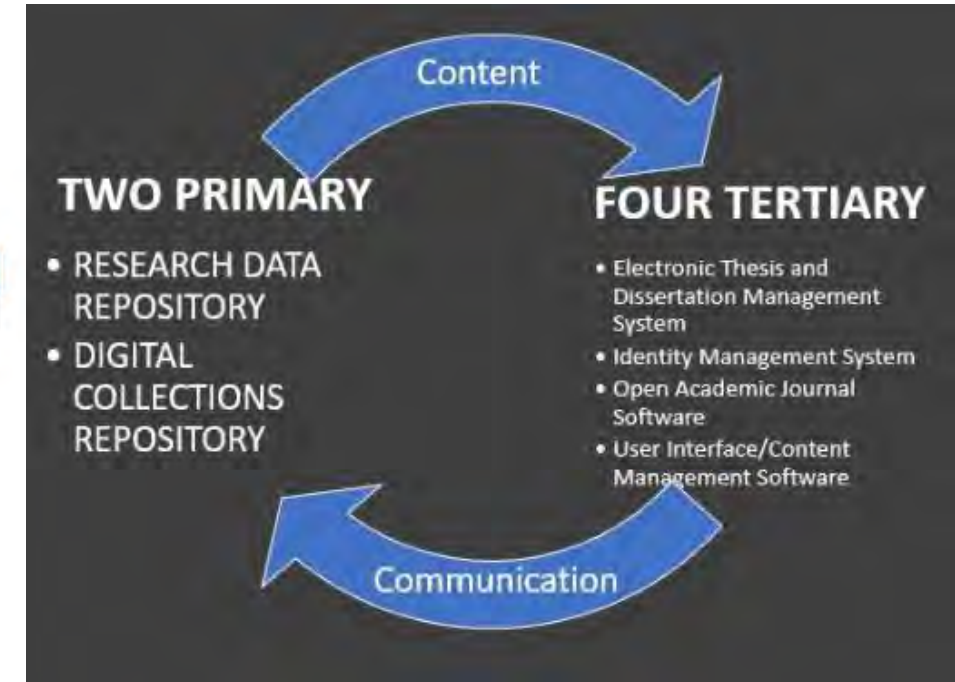
Share

Dataset Metrics ⓘ

58,334 Downloads ⓘ

### Description ⓘ

Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease ( *akiec* ), basal cell carcinoma ( *bcc* ), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, *lck* ), dermatofibroma ( *df* ), melanoma ( *mel* ), melanocytic nevi ( *nv* ) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, *vasc* ).



# An efficient deep learning approach to detect skin Cancer



View/Open

20341030, 19141024,  
16141014\_CSE.pdf (2.208Mb)

## Date

2021-09

## Publisher

Brac University

## Author

Islam, Ashfaqu  
Khan, Daiyan  
Chowdhury, Rakeen Ashraf

## Metadata

Show full item record

## URI

<http://hdl.handle.net/10361/15932>

## Abstract

Each year, millions of people around the world are affected by cancer. Research shows that the early and accurate diagnosis of cancerous growths can have a major effect on improving mortality rates from cancer. As human diagnosis is prone to error, a deep-learning based computerized diagnostic system should be considered. In our research, we tackled the issues caused by difficulties in diagnosing skin cancer and distinguishing between different types of skin growths, especially without the use of advanced medical equipment and a high level of medical expertise of the diagnosticians. To do so, we have implemented a system that will use a deep-learning approach to be able to detect skin cancer from digital images. This paper discusses the identification of cancer from 7 different types of skin lesions from images using CNN with Keras Sequential API. We have used the publicly available HAM10000 dataset, obtained from the Harvard Dataverse. This dataset contains 10,015 labeled images of skin growths. We applied multiple data pre-processing methods after reading the data and before training our model. For accuracy checks and as a means of comparison we have pre-trained data, using ResNet50, DenseNet121, and VGG11, some well-known transfer learning models. This helps identify better methods of machine-learning application in the field of skin growth classification for skin cancer detection. Our model achieved an accuracy of over 97% in the proper identification of the type of skin growth.

## Keywords

Cancer detection; Convolutional neural networks; Image classification; Deep learning

## LC Subject Headings

Machine learning; Cognitive learning theory (Deep learning)

## Description

This thesis is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering, 2021.

## POLICY GUIDELINES

- BracU Policy
- Publisher Policy

Search



Search BracU IR

This Collection

## BROWSE

All of BracU Institutional Repository

Communities & Collections

By Issue Date

Authors

Titles

Subjects

This Collection

By Issue Date

Authors

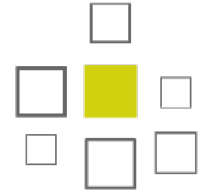
Titles

Subjects

## MY ACCOUNT

Login

Register



DSpace

Digital Collections  
Repository

Dspace  
<http://dspace.bracu.ac.bd/xmlui/handle/10361/15932>

BRAC University  
Libraries  
Institutional  
Repository



- Table of Contents
- List of Figures
- List of Tables
- Nomenclature
- Introduction
- Related Work
- Different Types of Skin Cancer
- Dataset Description**
- Dataset Pre-processing
- Model Training
- Model Building and Evaluation by CNN Model using Keras Sequential API
- Model Building and Evaluation using RESNET50
- Model Building and Evaluation using DENSENET121
- Model Building and Evaluation using VGG11
- Conclusion
- Bibliography

# An Efficient Deep Learning Approach to Detect Skin Cancer

by

Ashfaqul Islam

20341030

Daiyan Khan

19141024

Rakeen Ashraf Chowdhury

16141014

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
September 2021

**The Progress of Knowledge Through Global Open Science & Network Possibilities**

**2017 Stanford Nature Deep Learning Cancer ID Article**

**2018** Viennesse Doctor in Austria uploaded Dermatological Image Library to **Harvard Dataverse Data repository**

**2021 (November) Undergrad Thesis Published in Dspace Repositor**  
BRAC University, Dhaka Bangladesh, Dept. of Computer Science and Engineering

**All Downloaded July 2022 Texas, USA for Dublin IFLA Big Data Presentation**

# Questions & Comments

Ray Uzwyshyn, [ruzwyshyn@txstate.edu](mailto:ruzwyshyn@txstate.edu)  
<http://rayuzwyshyn.net>





## References and Further Resources

*Artificial Intelligence. Machine Learning. Neural Networks. Future Technology.* Bloomberg Businessweek Canada. 2022.  
<https://www.youtube.com/watch?v=ypVHymY715M>

Cann, A., Dimitriou, K. Hooley, T. 2011. *Social Media: A Guide for Researchers.* Research Information Network. University of Derby, UK.

Chan-Park, C. and Sare, L. Waugh, S. 2022. *Results of the Texas Data Repository User Survey, 2022.* Texas Conference on Digital Libraries Presentation.

ColdFusion (2018). *Why Deep Learning Now?* (Documentary Overview). [https://www.youtube.com/watch?v=b31yDNB\\_cil](https://www.youtube.com/watch?v=b31yDNB_cil)

Echle et al. Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. *British Journal of Cancer.* November 2020.  
<https://www.nature.com/articles/s41416-020-01122-x>

Esteva, A., Thrun, S. et al. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, Volume 542 (February 2, 2017). pp. 115-119.  
doi:10.1038/nature21056

Fridman, Lev. *MIT Deep Learning and Artificial Intelligence Lectures.* <https://deeplearning.mit.edu/> 2022.

Islam, A., Khan, D. and Chowdhury, R. 2021. *An Efficient Deep Learning Approach to Detect Skin Cancer Undergraduate Thesis.* BRAC University DSpace Institutional Repository, 2021. Available: <http://dspace.bracu.ac.bd/xmlui/handle/10361/15932>

Kleinveltdt, Lynn. Smarter high education learning environments through AI: What this means for academic libraries. *Trends and Issues in Library Technology: Special Issue on AI:* June 2022. pp. 12-15. <https://repository.ifla.org/handle/123456789/1940>

Mitchell, Tom. 2022 *Where on Earth is AI Headed?* Carnegie Mellon. <https://www.youtube.com/watch?v=ij9vqTb8Rjc>

Peters, T. and Waugh, L. Larger Data Storage Report: Research Data Management Initiatives and Planning, January 2022. Texas State University Libraries (Unpublished White Paper)

Texas Data Repository 2022. <https://dataverse.tdl.org/>

Tschandl, Phillip et al. *Human-computer Collaboration for Skin Cancer Recognition.* *Nature Medicine*, 22 June 2020, 1229-1234. See:  
<https://www.nature.com/articles/s41591-020-0942-0>.

Uzwyshyn, R. 2022. **Online Research Data Repositories and Digital Scholarly Ecosystems: From Research Data and Datasets to Artificial Intelligence and Discovery.** IFLA WLIC 2022. Dublin, Ireland. DOI: 10.13140/RG.2.2.12354.86728

---. 2022. Steps Towards Building Library AI Infrastructures: Research Data Repositories, Scholarly Research Ecosystems and AI Scaffolding. *New Horizons in Artificial Intelligence in Libraries* (IFLA Satellite Conference), National University of Ireland, Galway, IR. DOI: 10.13140/RG.2.2.21120.30728

---. 2021. Frameworks for Long Term Digital Preservation Infrastructures. *Computers in Libraries.* September 2021. pp.4-8.

Uzwyshyn, R. 2020. *Developing an Open-Source Digital Scholarship Ecosystem.* ICEIT2020. St. Anne's College Oxford, United Kingdom. February 2020. Available at:  
[https://www.researchgate.net/publication/336923249\\_Developing\\_an\\_Open\\_Source\\_Digital\\_Scholarship\\_Ecosystem](https://www.researchgate.net/publication/336923249_Developing_an_Open_Source_Digital_Scholarship_Ecosystem).

- - -. 2020. *Open Digital Research Ecosystems: How to Build Them and Why.* *Computers in Libraries*, (40) 8. November 2020. [https://www.researchgate.net/publication/345956074\\_Online\\_Digital\\_Research\\_Ecosystems\\_How\\_to\\_Build\\_Them\\_and\\_Why](https://www.researchgate.net/publication/345956074_Online_Digital_Research_Ecosystems_How_to_Build_Them_and_Why)

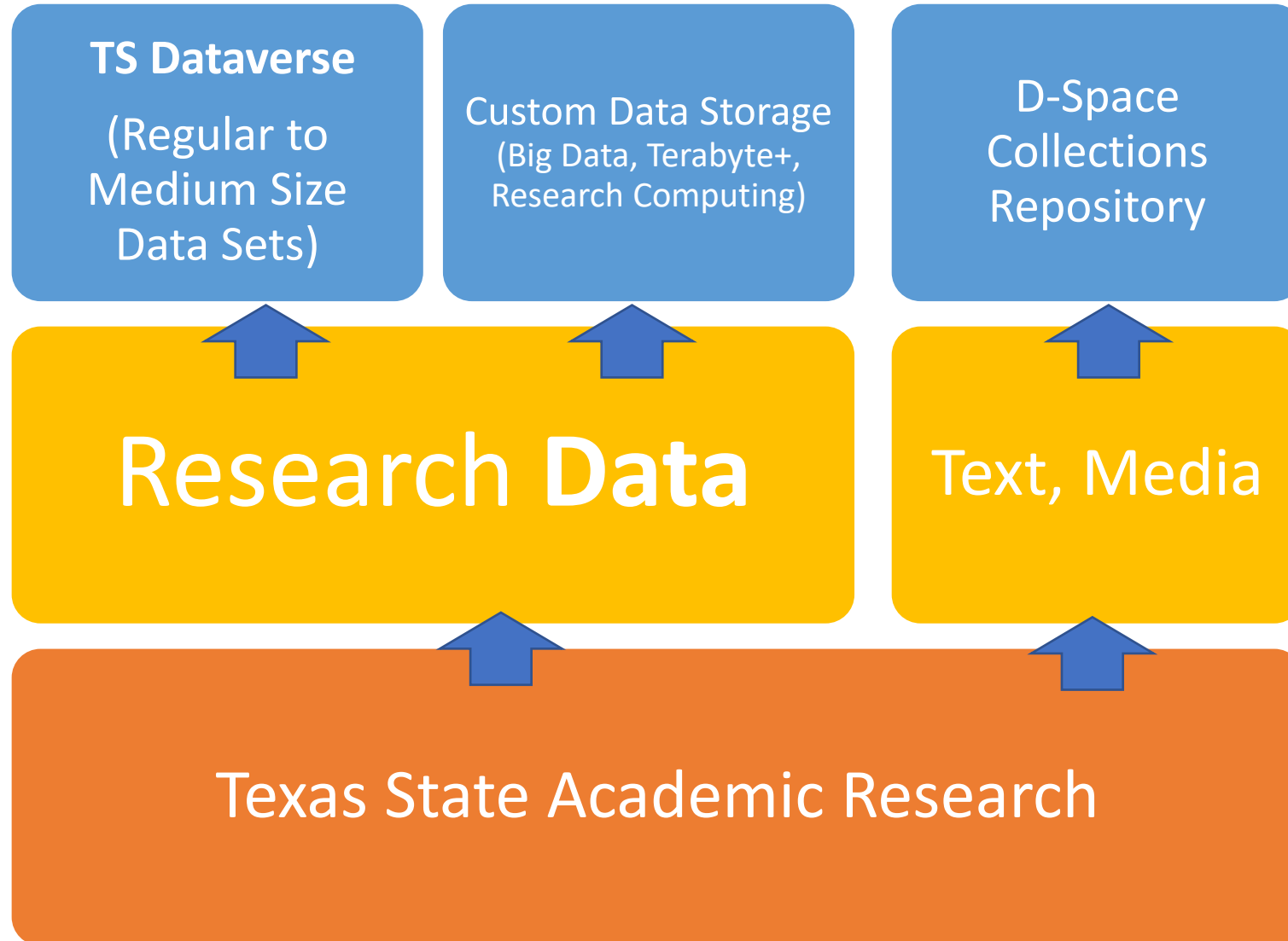
---. 2016. Online Research Data Repositories: The What, When Why and How. *Computers in Libraries.* 36:3, April 2016. pp. 18-21.  
<http://rayuzwyshyn.net/TXU2016/OnlineDataResearchRepositoriesUzwyshyn.pdf>

Waugh, L. *Texas State University Annual Usage Report 2020.* TXST Dataverse Repository. Texas Conference on Digital Libraries Presentation. Texas State University.





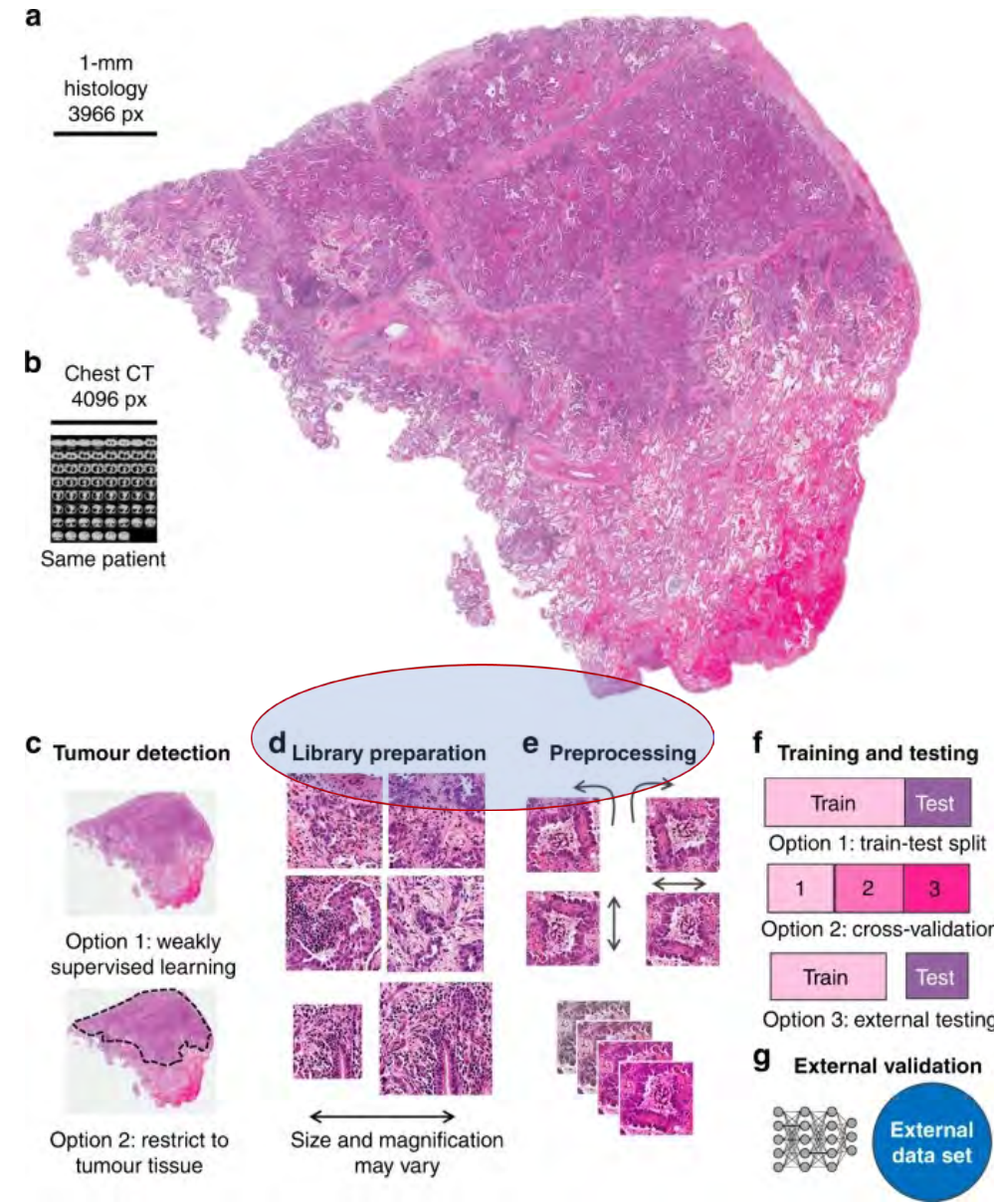
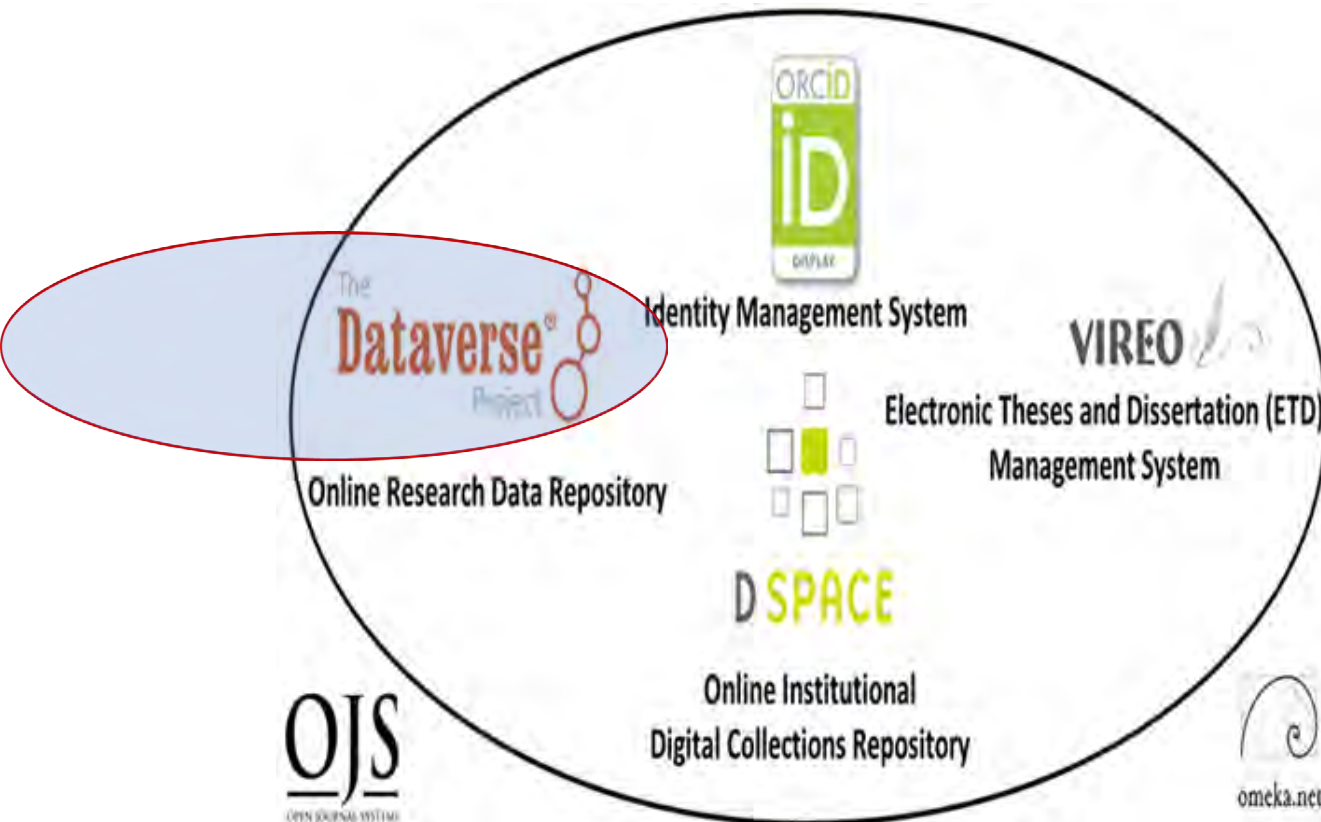
# Texas State Repositories Architecture





# Combining Data Centered Research Ecosystems + Artificial Intelligence

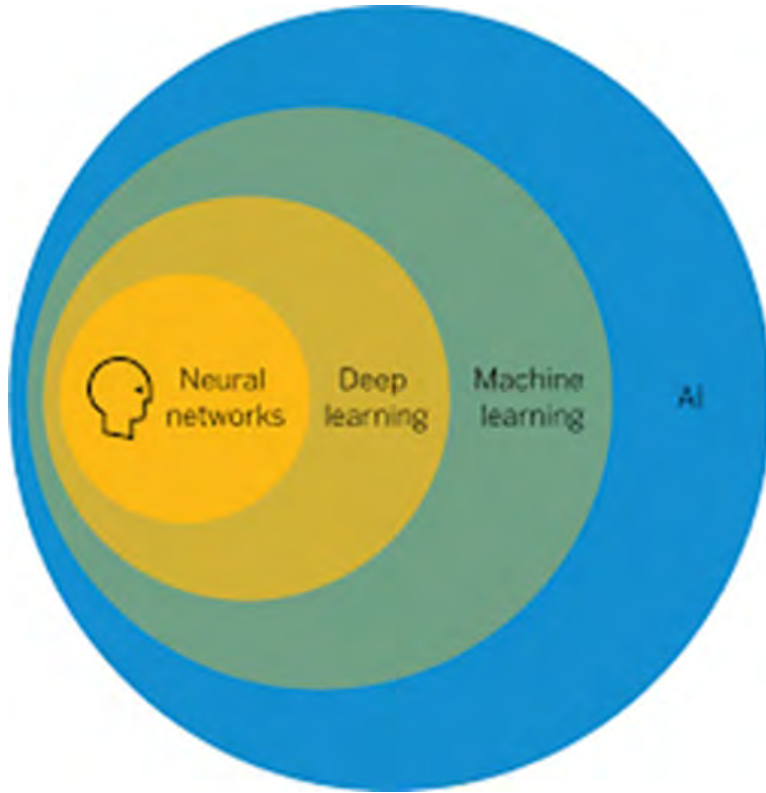
(Many New Possibilities for Global Open Science, New Insights and NewDiscovery)



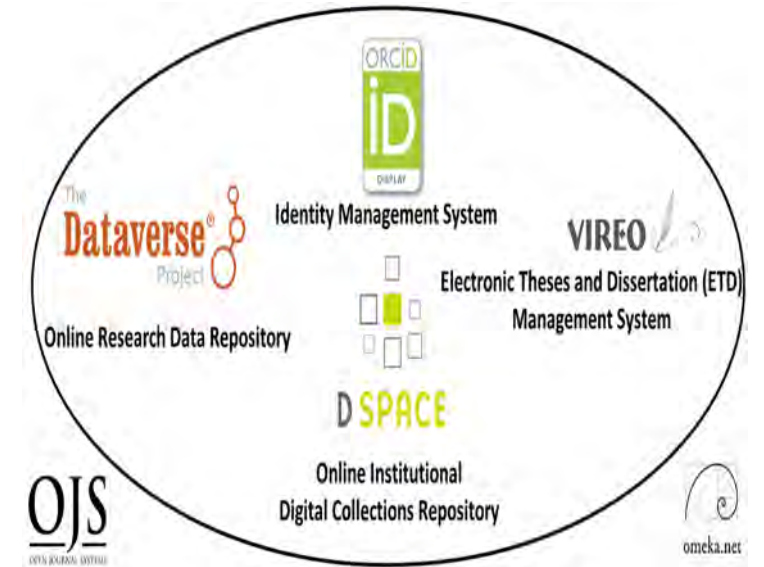
Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. British Journal of Cancer, Echle et al. November 2020

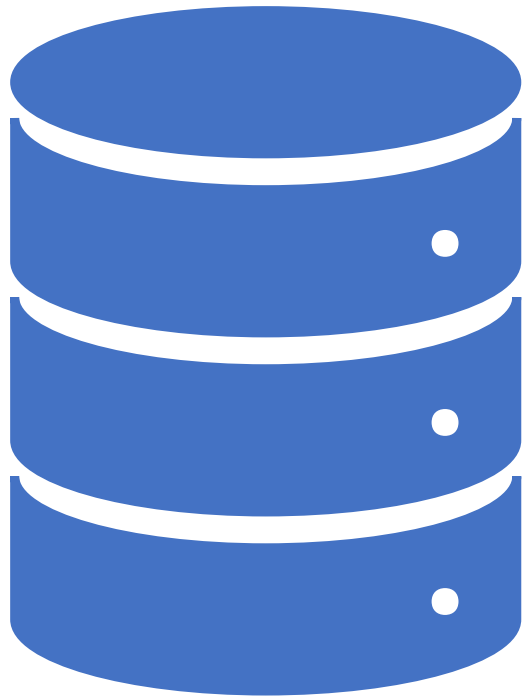
# Last Five Years Has Shown Incredible Progress of, Analytical Computational Tools, Particularly, AI

Artificial Intelligence (Machine Learning (Deep Learning)) = Better Algorithms + Greater Computing Power + Large Data Sets



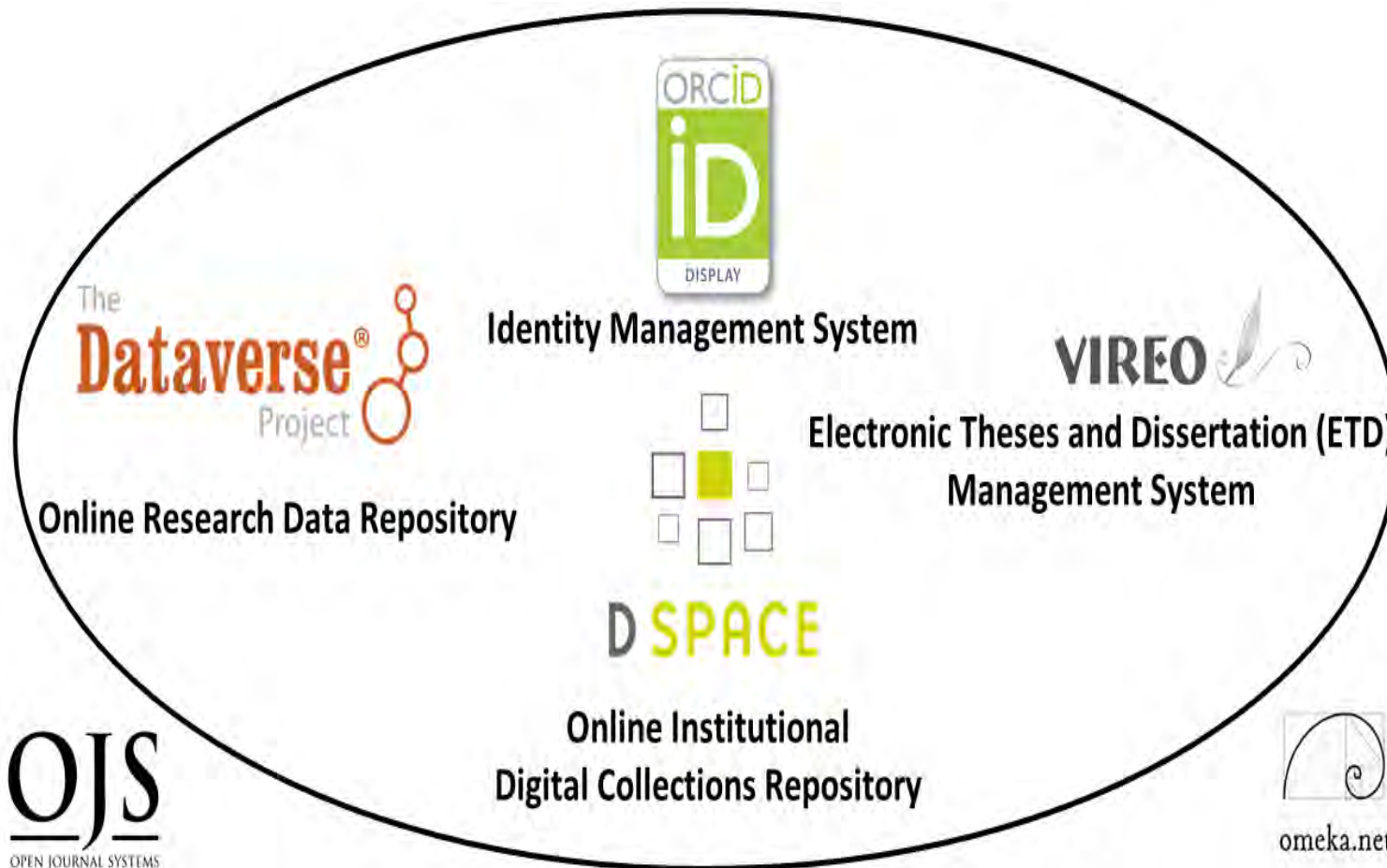
- **Computer Vision (Facial/Object Recognition Cancer Cell Detection )**
- Natural Language Processing (Speech to Text, Translation)
- Cybersecurity, Fraud Detection
- Conversational Chatbots & Robotic Agents
- Strategic Reasoning (AlphaGo)





# Big, Bigger Data and Big Data

# What are the General Common Characteristics for a Data Repository and Digital Scholarship Ecosystem?



Open Source Software



Active Developer Communities



Customizable Components

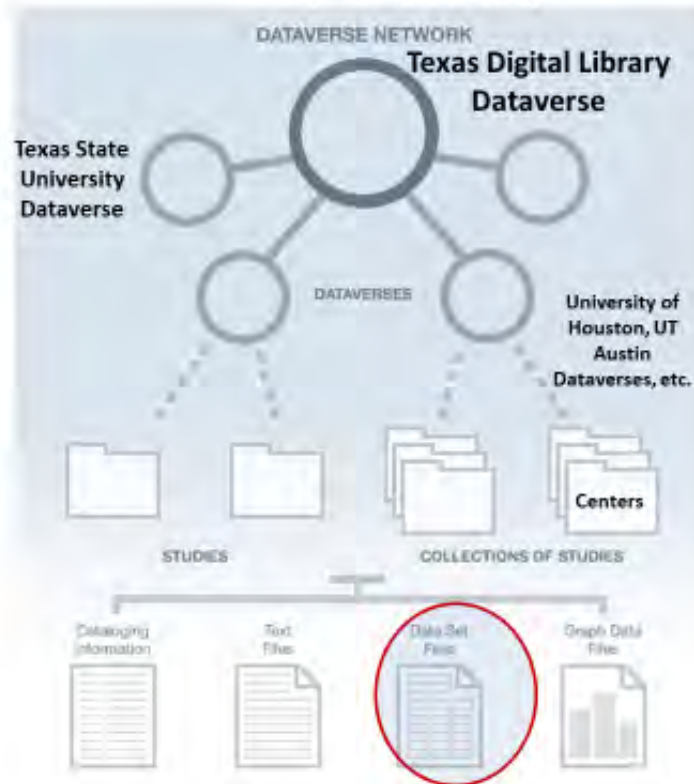


omeka.net



# Texas State University Dataverse: Can be configured as Single Instance or as a Consortial Model

## Dataverse Architecture (Consortial)



(Texas Aggregates Various Individual Universities through the Texas Digital Library)



<https://dataverse.tdl.org/>

# The Progress and Potential of AI, Discovery, Data and Big Data Ecosystems for Libraries and Research Institutions



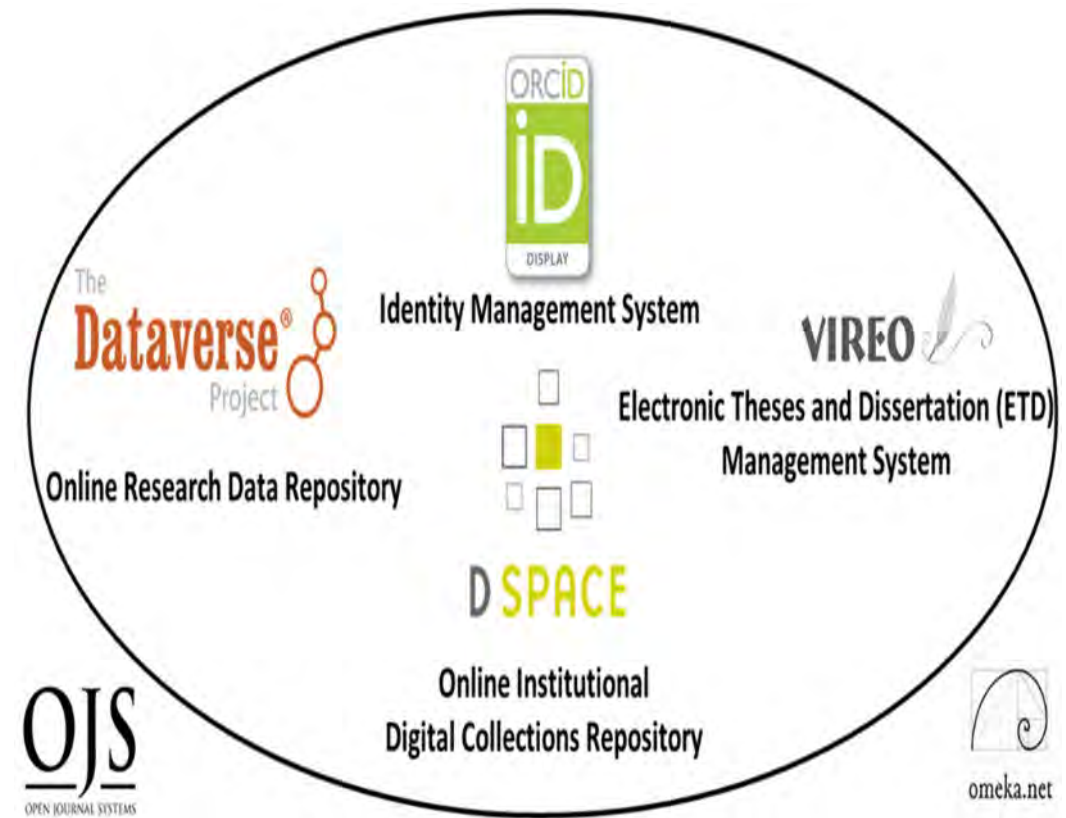
2017 Stanford  
Nature Deep Learning  
Cancer ID Article

2018 Viennesse Doctor  
uploaded Dermatological Image  
Library to Harvard Dataverse  
Data repository

2019 Global Open Science  
Through Network Possibilities

2021 (November)  
Dspace Repository  
Undergraduate Thesis  
BRAC University, Dhaka  
Bangladesh, Dept. of  
Computer Science and  
Engineering

Downloaded July 2022  
Texas, USA



- Table of Contents
- List of Figures
- List of Tables
- Non-refereed
- Introduction
- Related Work
- Different types of Skin Cancer
- Dataset Description
- Dataset Pre-processing
- Model Training
- Model Building and Evaluation by CNN Model using keras Sequential API
- Model Building and Evaluation using RESNET50
- Model Building and Evaluation using DENSENET121
- Model Building and Evaluation using VGG11
- Conclusion
- Bibliography

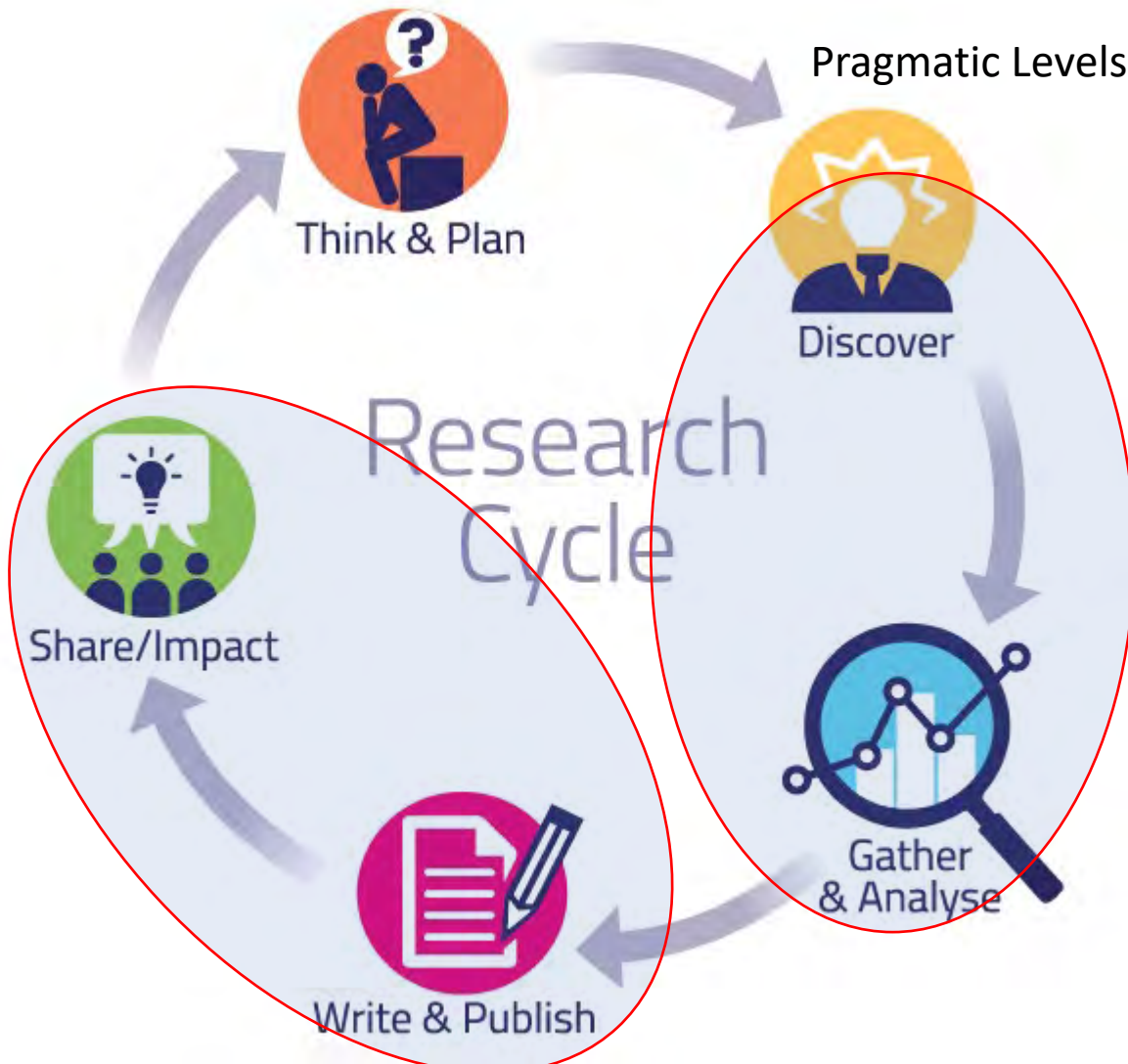
An Efficient Deep Learning Approach to Detect Skin Cancer

by  
Abdullah Islam  
20341306  
Doljan Khan  
19141024  
Rakun Ashraf Chowdhury  
16111011

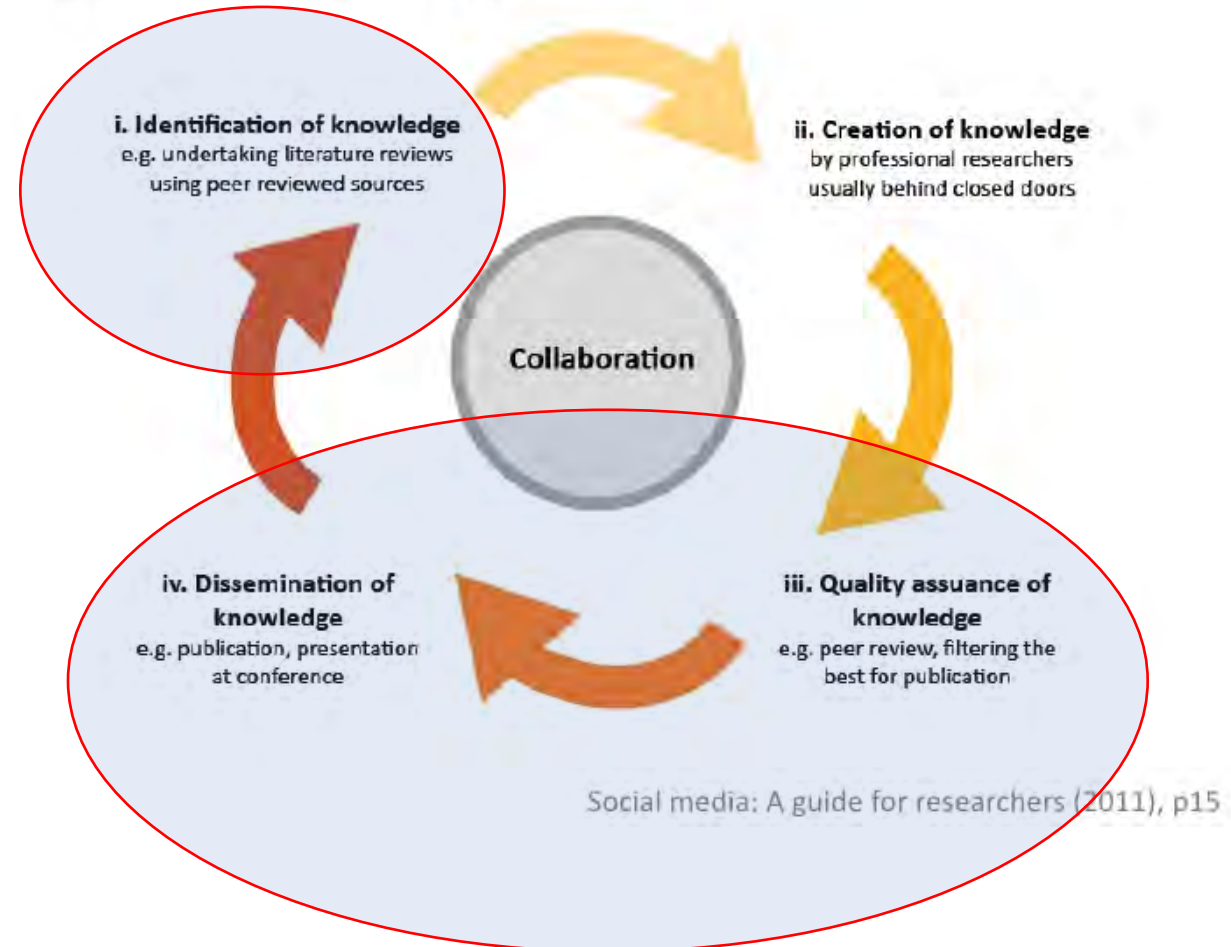
A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
September 2021

# Together These Digital Ecosystem Components Enable the Academic Research Cycle



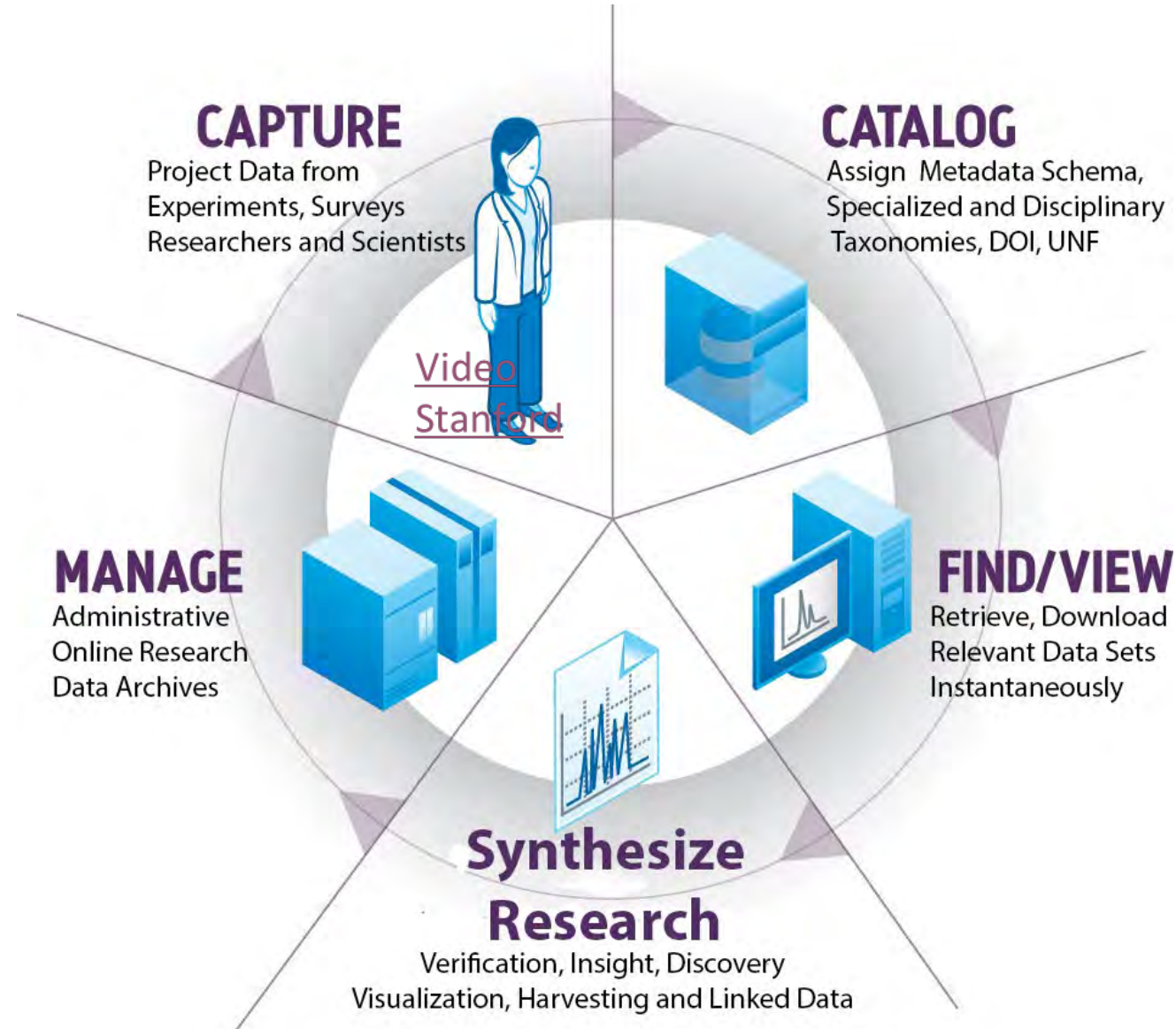
## The academic research cycle





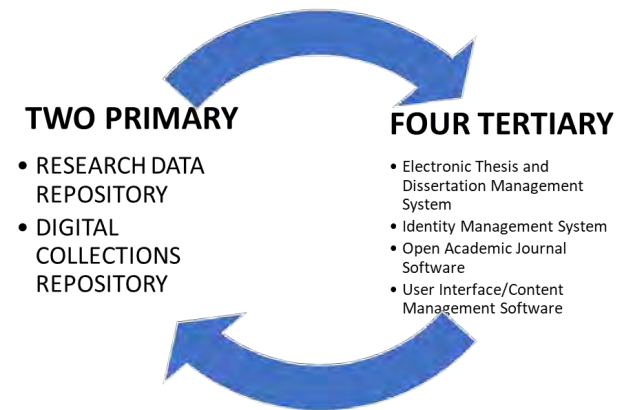
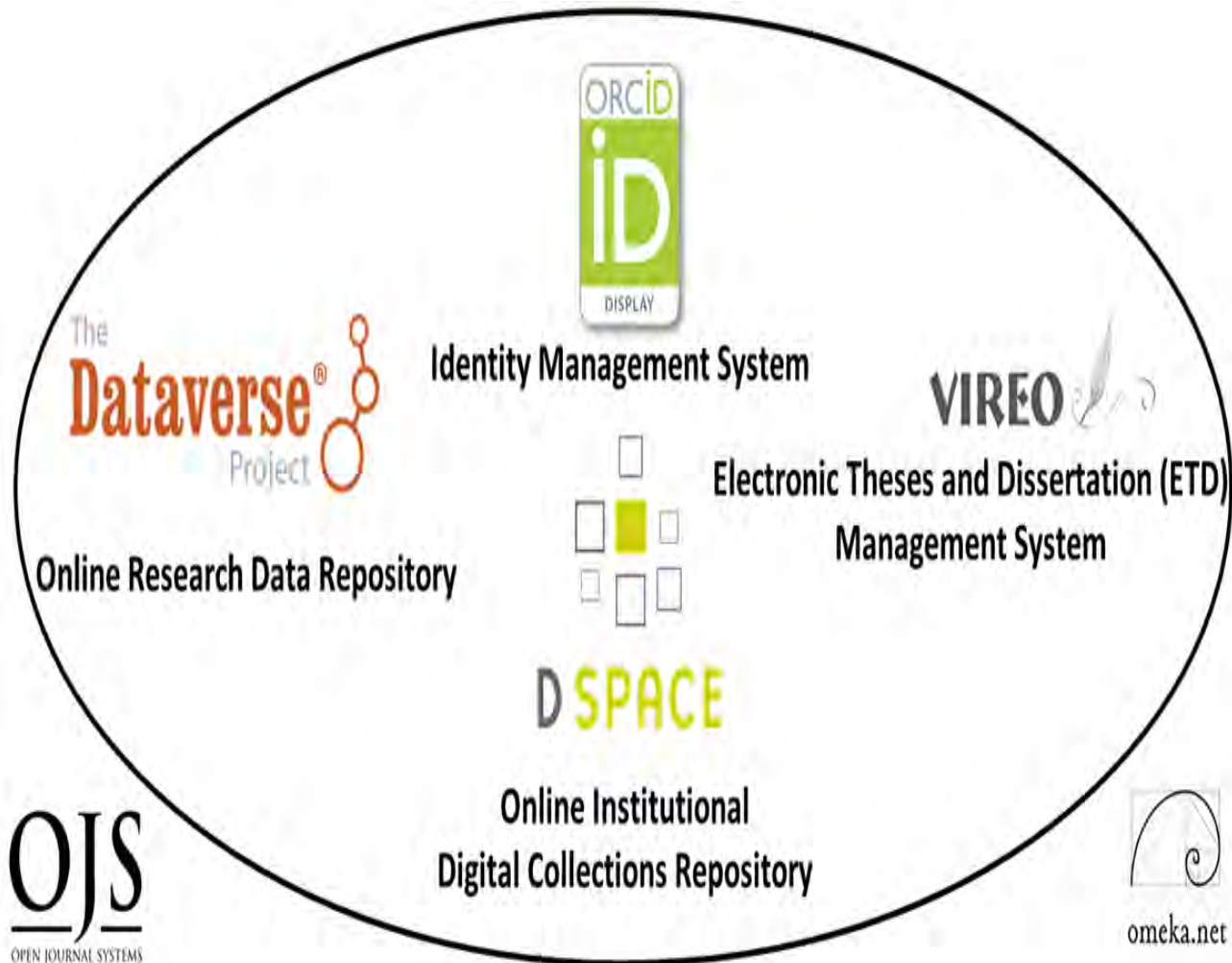
# The Research Data Repository Lifecycle

Setting Better Foundations & Organization for AI Infrastructures

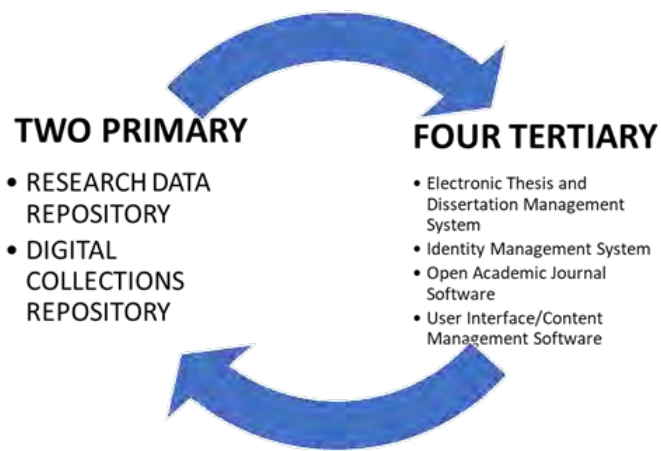




# Digital Scholarship Ecosystem Centered on Research Data Repository and Collections Repository



# Questions Comments



Ray Uzwyshyn, Ph.D. MBA MLIS  
Director, Collections and Digital Services  
Texas State University Libraries  
[ruzwyshyn@txstate.edu](mailto:ruzwyshyn@txstate.edu)  
<http://rayuzwyshyn.net>





















