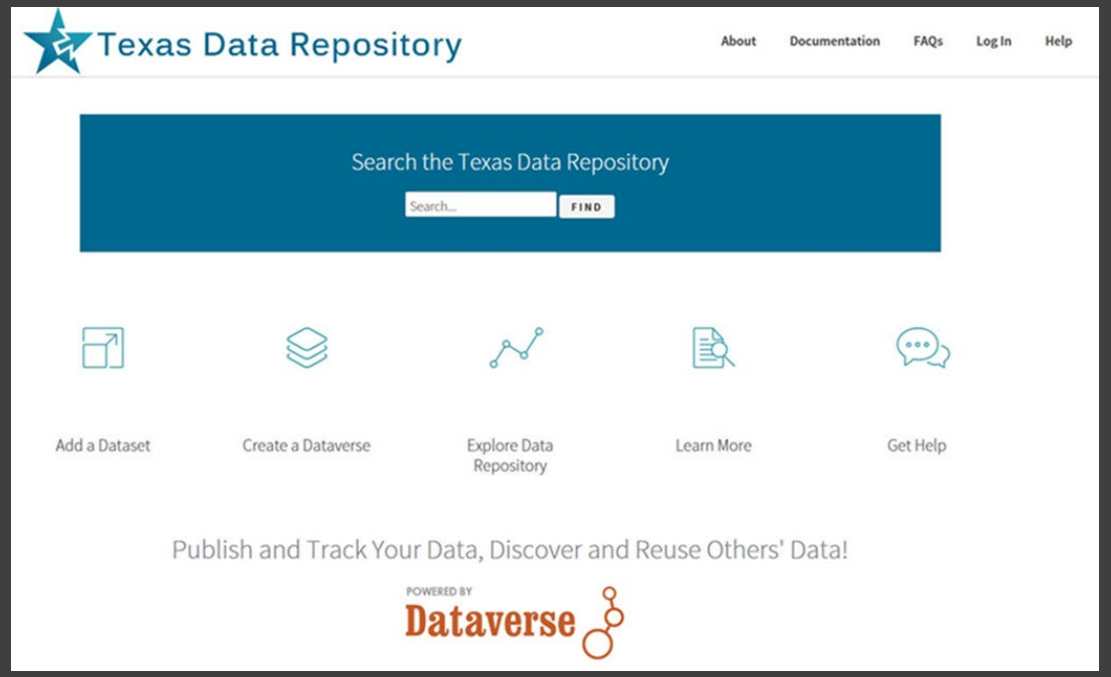
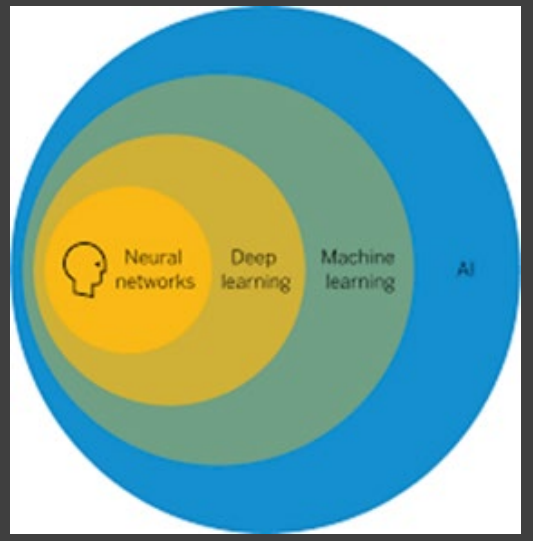
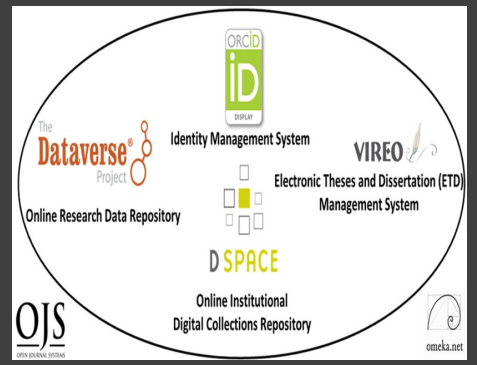
Digital Technology and Innovation in Library Research Services

From **Data and Scholarly Research Ecosystems** to **Artificial Intelligence and Global Discovery**





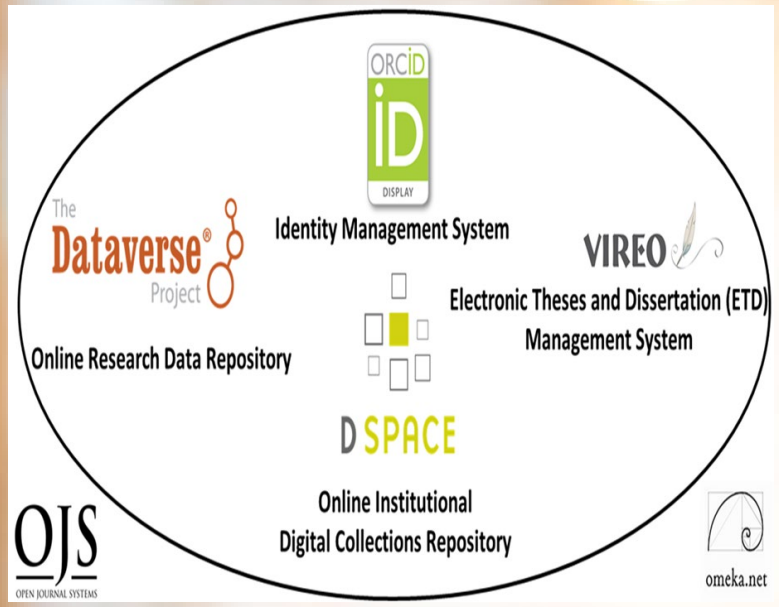


Session 1: Digital Transformation and Libraries
PCLIPK Library Policy Forum
National Library of Korea, November 3, 2022

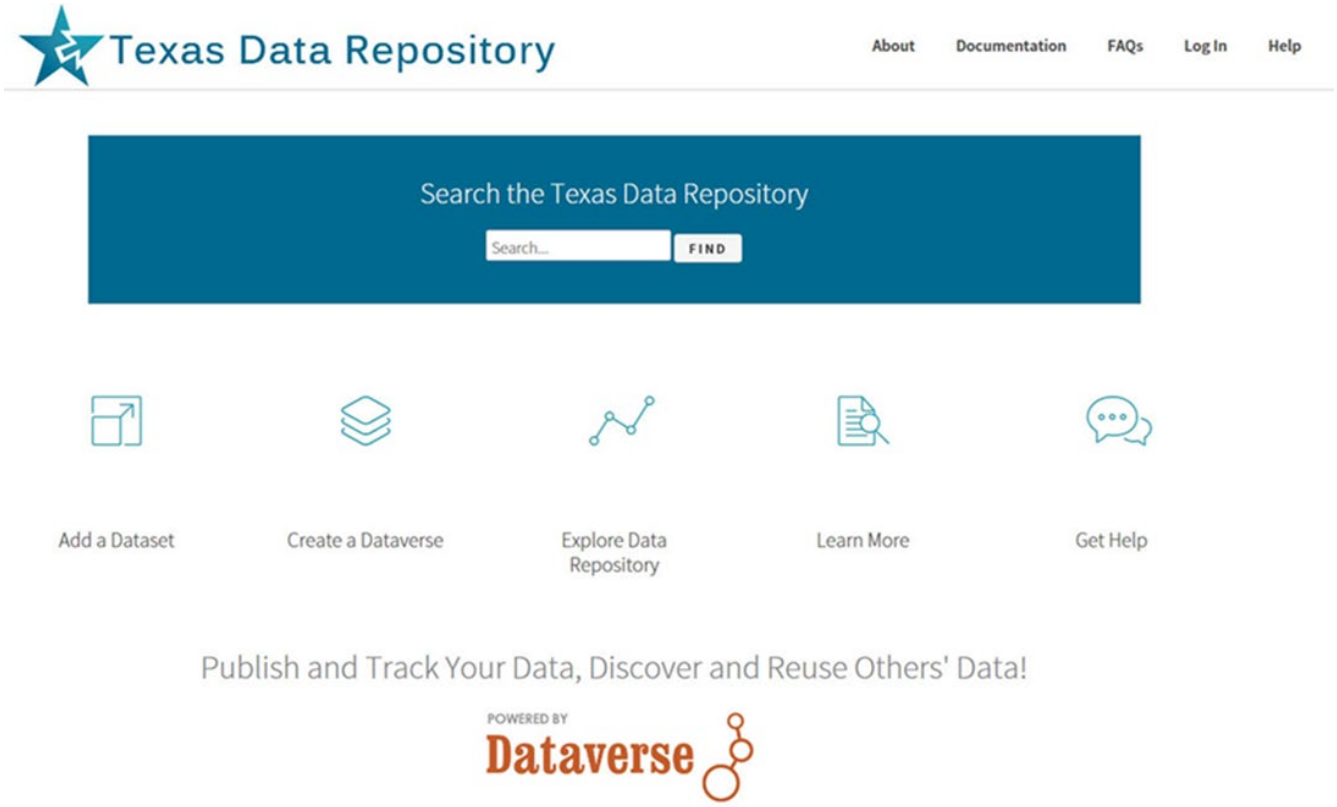
Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries, USA
ruzwyshyn@txstate.edu

The Physical Library is Transforming From Book Warehouse to Digital Ecosystem

- 90% of Materials Budgets are currently Electronic Resources, e-journals and e-books.
- Digital Literary, Specialized Software and Digital Technologies Are Ushering In New Paradigms
- Data Research Repositories, Scholarly Research Ecosystems and New Artificial Intelligence Models and Infrastructures are Bringing in the new millennia



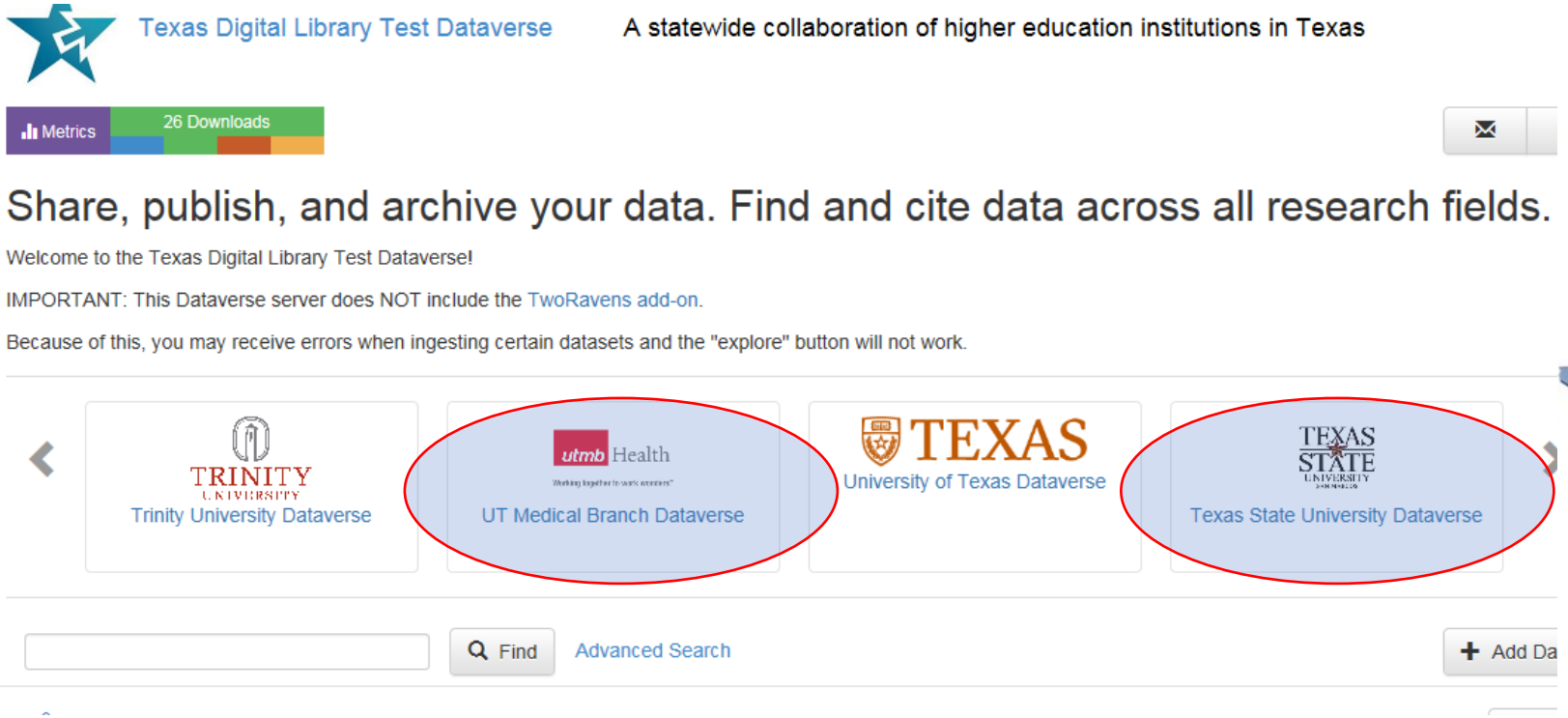
What is an Online Research Data Repository?



https://dataverse.tdl.org/

- Platform to Manage Researcher and Institutions Data/Metadata
- Permalinking Strategy for Data Citation
- Way to Manage Open Access, Funder Compliance and Open Science and Research
- Data Archiving and Sharing Strategy

What Does and Online Research Data Repository Do?



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

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

Many Types of Data Projects (Sizes)

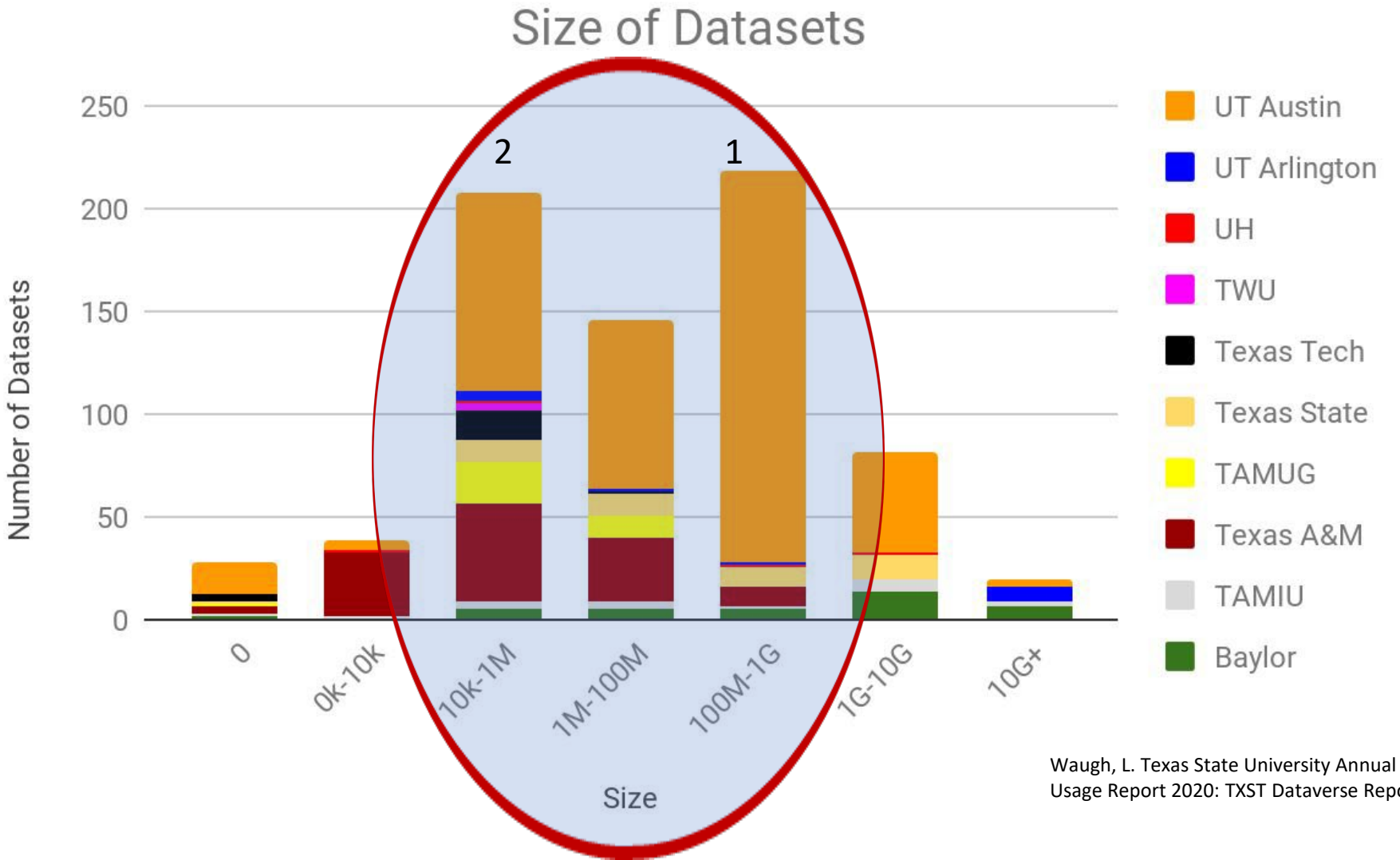
1) Normal range (<4GB Files <10GB Datasets)
Files/Data Fit on Server/Cloud, may be uploaded to the Data Repository, 4GB files, 10GB Datasets)

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



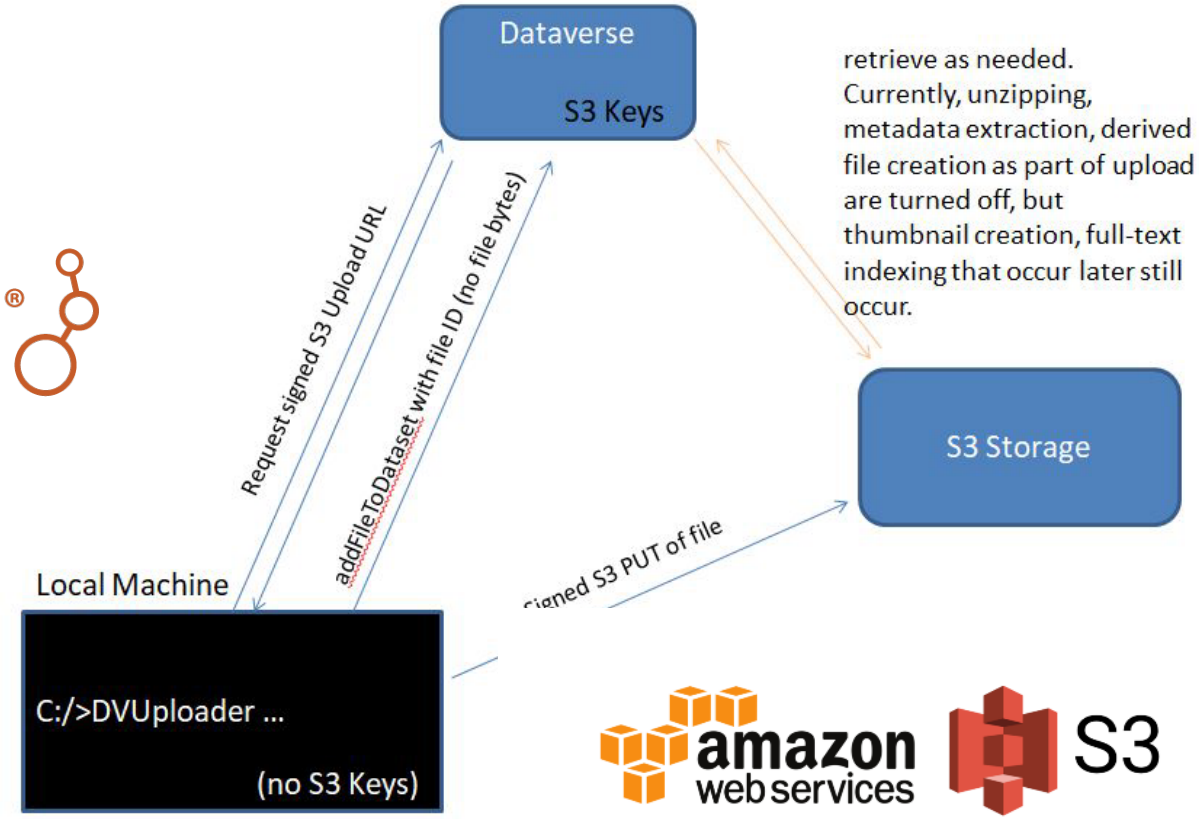
Waugh, L. Texas State University Annual Usage Report 2020: TXST Dataverse Repository. Texas State University

Beta Prototyping Big & Bigger Data Options

2020-2022

DRYAD

The Dataverse® Project

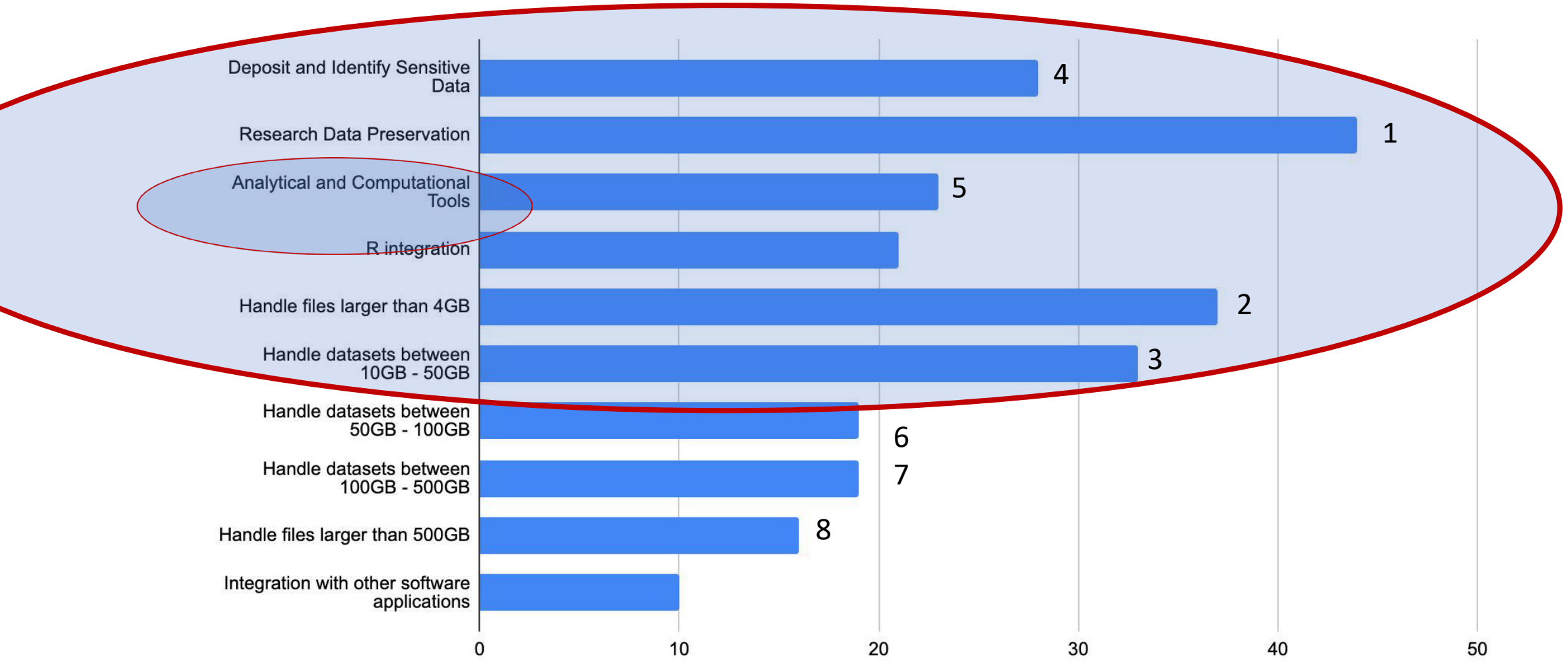


Up to 300 GB/dataset
Fee Based Institutional Model 7.5/13.5 K/Year

TACC
TEXAS ADVANCED COMPUTING CENTER

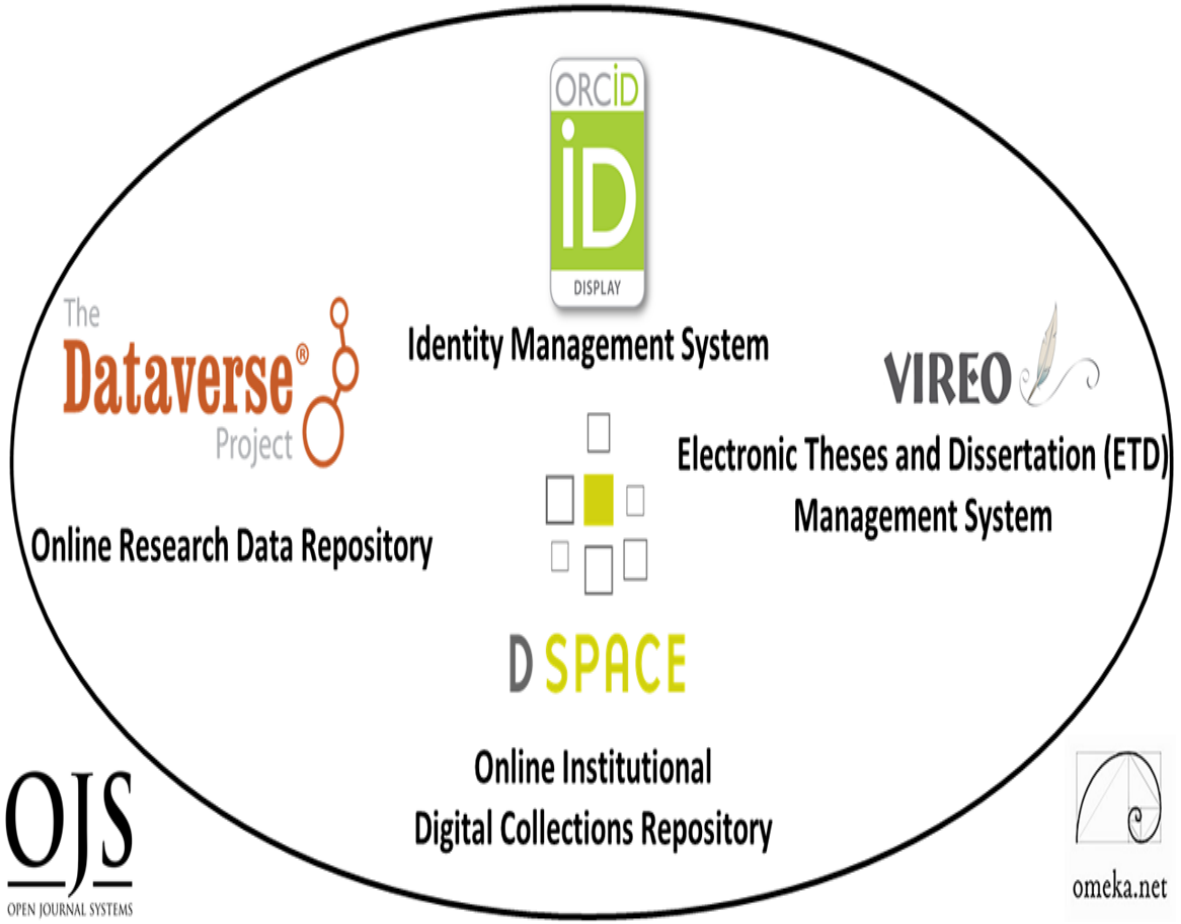
<20 GB Upload
(Download Challenges)

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



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

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



Digital & Web Services : University Libraries : Texas State University (txstate.edu)

TWO PRIMARY COMPONENTS
(Content)

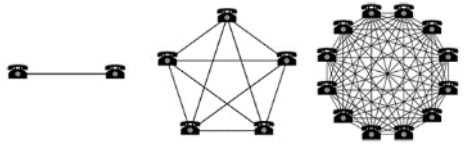
- **RESEARCH DATA REPOSITORY**
- DIGITAL COLLECTIONS REPOSITORY

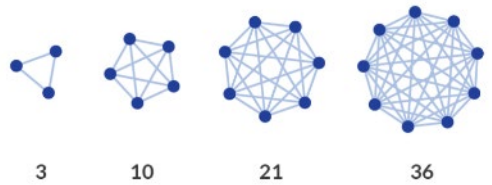
FOUR TERTIARY COMPONENTS
(Communication)

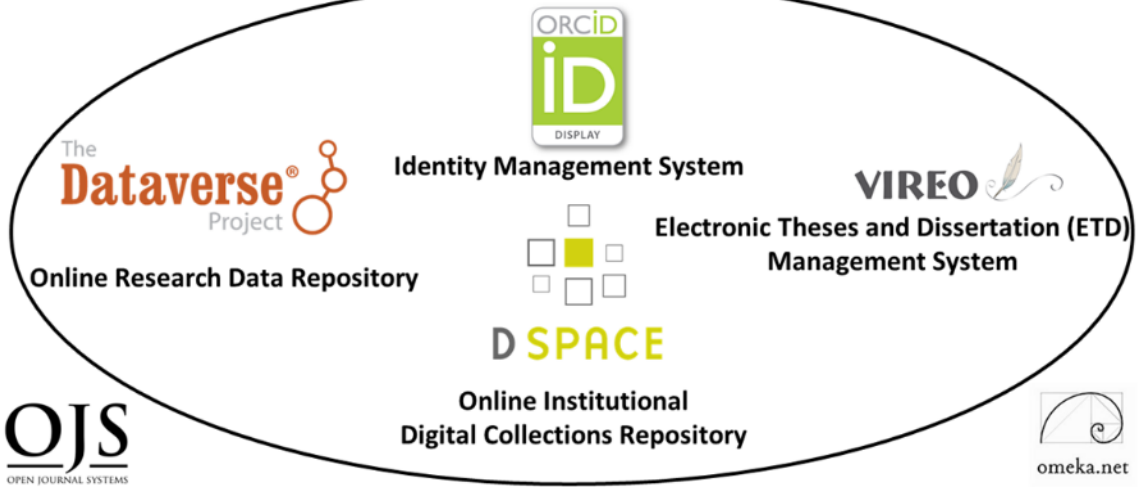
- Electronic Thesis and Dissertation Management System (VIREO)
- Identity Management System (ORCID)
- Open Academic Journal Software (OJS3
- User Interface/Content Management Software (OMEKA)

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

(Sum of Connections Becomes Greater than the Parts)







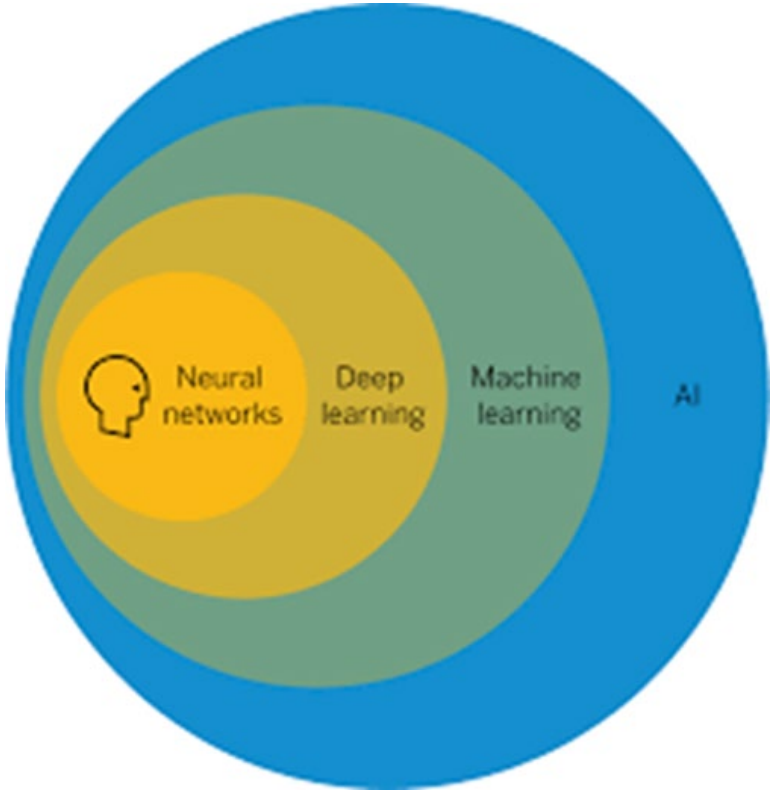
Together These Digital Ecosystem Components Enable the Academic Research Cycle

Faculty
Students
Researchers
Industry

Pragmatic Levels

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

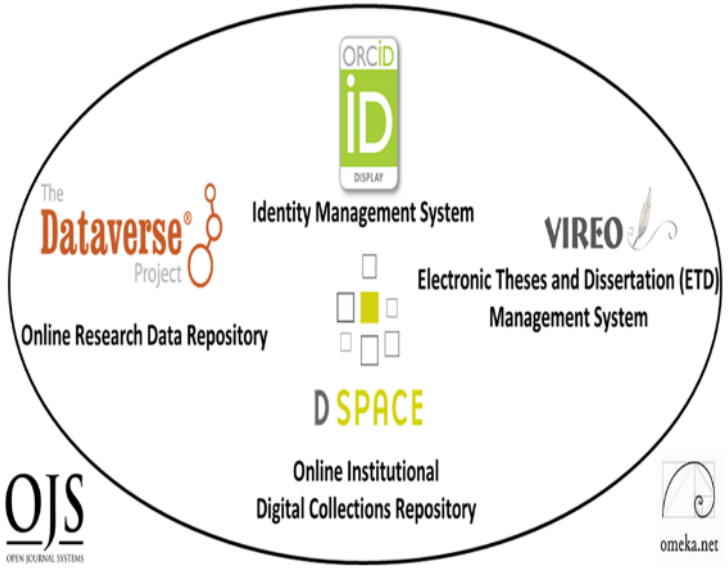
Artificial Intelligence (Machine Learning (Deep Learning)) = Better Algorithms +
Greater Computing Power +
Large Data Sets



Artificial Intelligence Models

- **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)

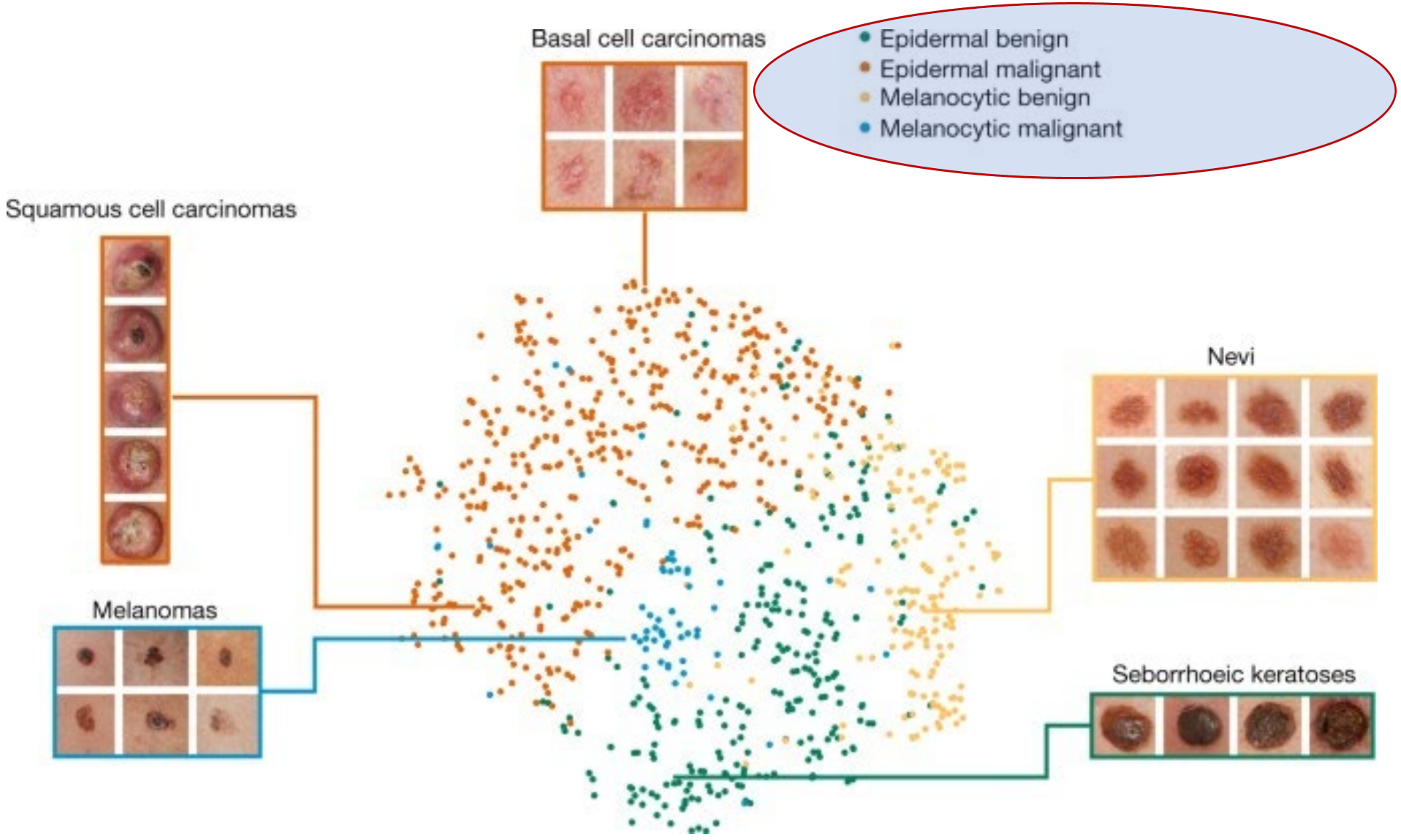
+



New Digital Library Ecosystem Possibilities

Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks

2017, Nature, Esteva, Thrun et Al



[Video Stanford]

Data Research Repository Upload

Open Science Dermatology Image Dataset, Dr. Philip Tschandl, University of Vienna
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.

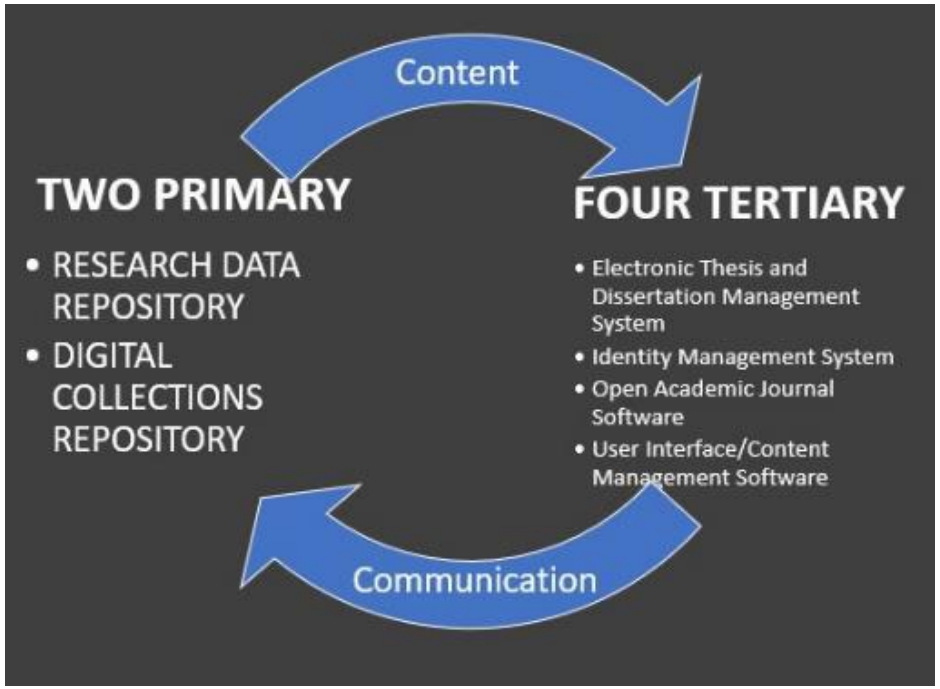
Access Dataset ▾ | Contact Owner | Share

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, `bkl`), dermatofibroma (`df`), melanoma (`mel`), melanocytic nevi (`nv`) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, `vasc`).



- 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 Dermatalogical 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

BRAC UNIVERSITY
Inspiring Excellence
Institutional Repository

BracU IR / School of Data and Sciences (SDS) / Department of Computer Science and Engineering (CSE) / Thesis & Report, BSc (Computer Science and Engineering) / View Item

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, Ashfaqul
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)

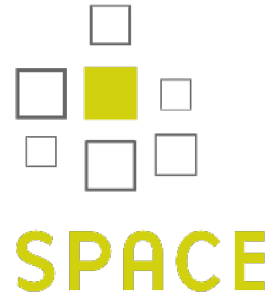
Description
This thesis is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering, 2021.

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



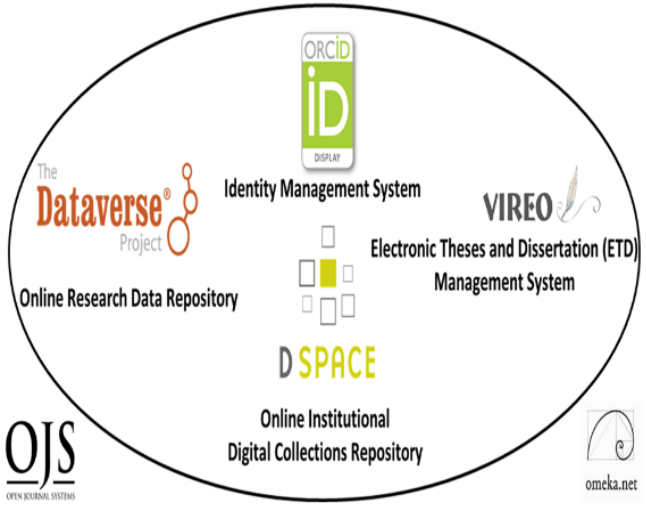
Digital Collections Repository

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

BRAC University Libraries Institutional Repository
Chaka Bangladesh

Questions & Comments

Ray Uzwyshyn, ruzwyshyn@txstate.edu
http://rayuzwyshyn.net



Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries, USA

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=b3IyDNB_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

Kleinveldt, 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.

- - -. 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

---. 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.