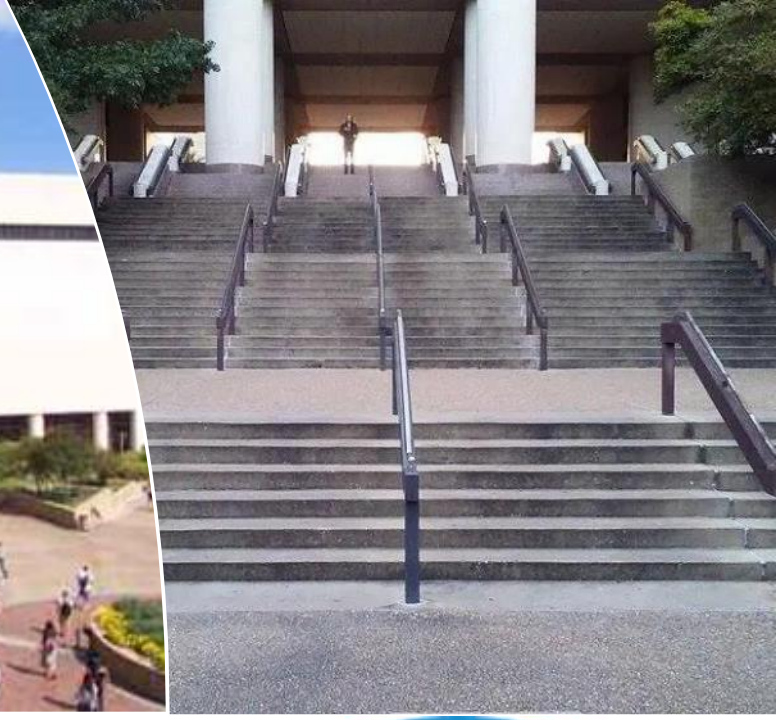
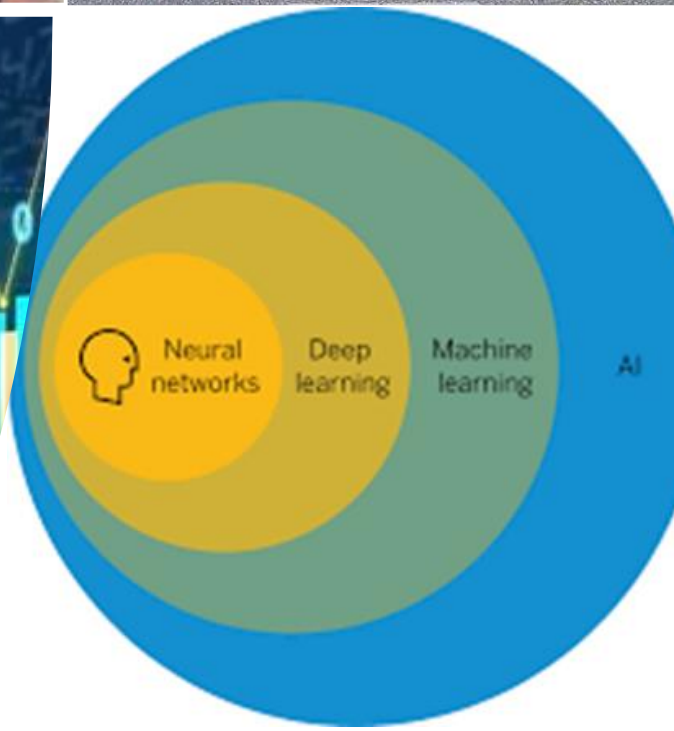
Steps Towards Building Library AI Infrastructures and Programs

(Research Data Repositories, Scholarly Research Ecosystems and AI Scaffolding)





Presented for New Horizons in AI for Libraries
IFLA Satellite Conference, Galway, Ireland
National University of Ireland, July 21, 2022

Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services
Texas State University Libraries, USA
July 2022, ruzwyshyn@txstate.edu

Texas State University Libraries

Large Academic Library system, ARL Library
Main campus Library and campus and other offsite campus libraries (Health Professions, Austin/Roundrock)

Texas State University, Undergraduate, Graduate and Doctoral Institution 40,000 Students

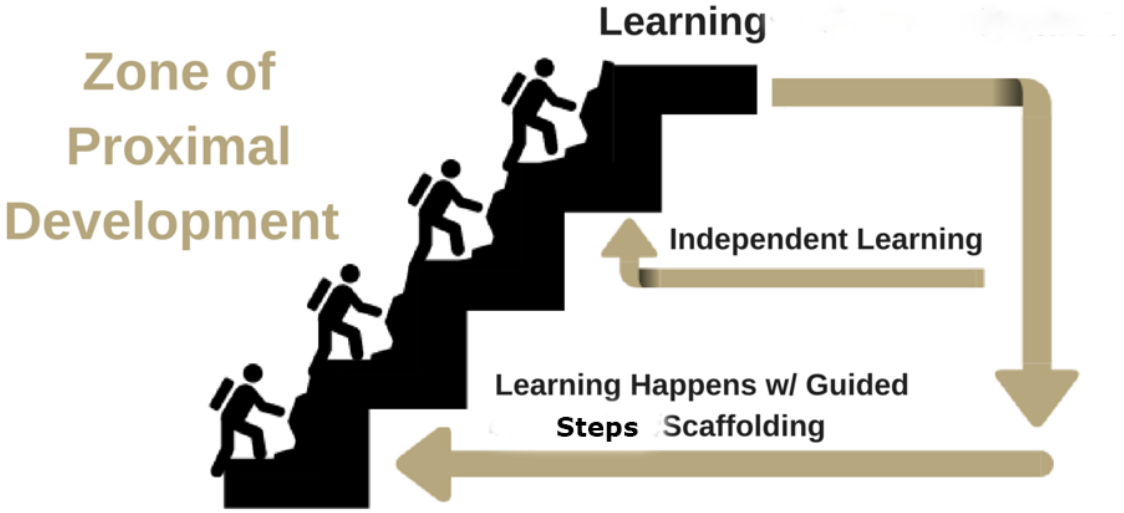
Carnegie Class II Doctoral University (Higher Research Activity)

Designated Emerging Research Institution (Texas)

Texas State Libraries,
Academic Research Library
ARL Library

Steps and Scaffolding Towards Building Library AI Infrastructures

Learning is Too Hard: Anxiety



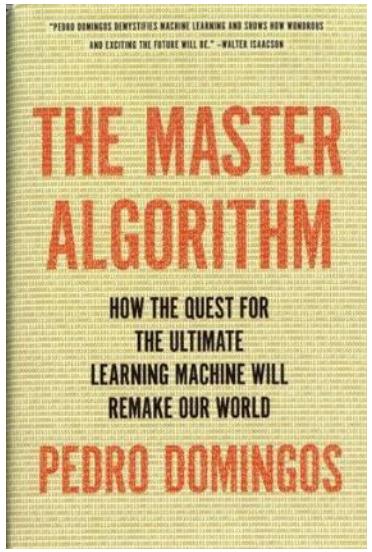
Concept Areas: Data Science, Machine Learning, AI, Information Science, Programming, IT Project Management

Learning is Too Easy: Boredom

AI Has Many Main Paradigms and Origins

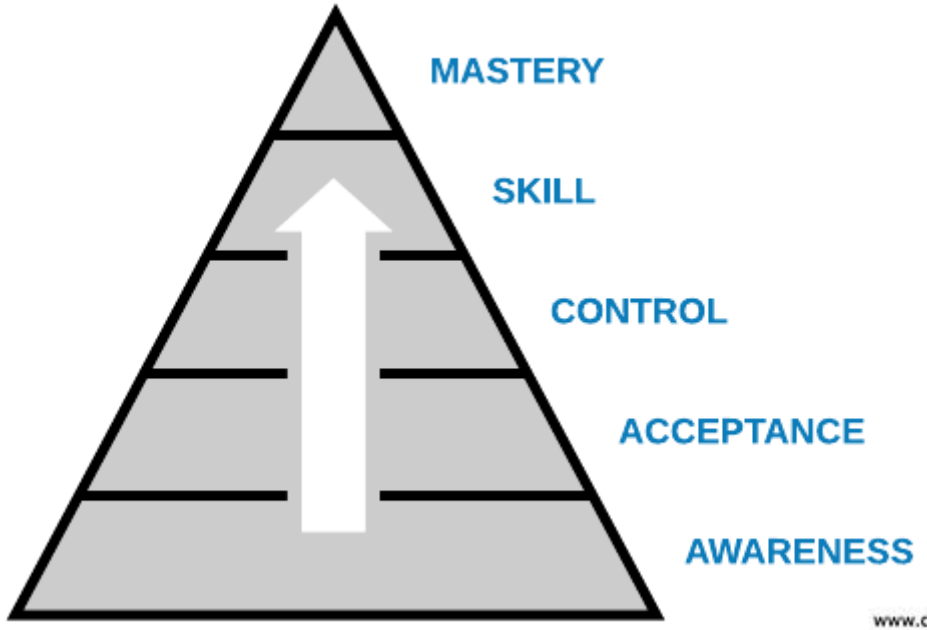
Algorithm Methods, Best Problem Types to Solve and Solution Methods for These Problems,
Dr. Pedros Domingos, University of Washington

AI Paradigm	Origin	Algorithm	Problem	Solution
Deep Learning Machine Learning	**Neuroscience (Neural Nets)**	**Back Propagation Neural Nets**	Complex Tasks, Hidden Patterns	Back propagation
Symbolic AI	Logic, Philosophy	Inverse Deduction	Knowledge Composition	Inverse Deduction
Bayesian Inference	Statistics, Probability Theory	Probabilistic Inference	Uncertainty	Probabilistic Inference
Evolutionary Computation	Evolutionary Biology (Complexity Theory	Genetic Algorithms	Structure Discovery	Genetic Programming
Reasoning by Analogy	Psychology	Kernel Machines (Support Vector Machines)	Similarity	Kernel Machines

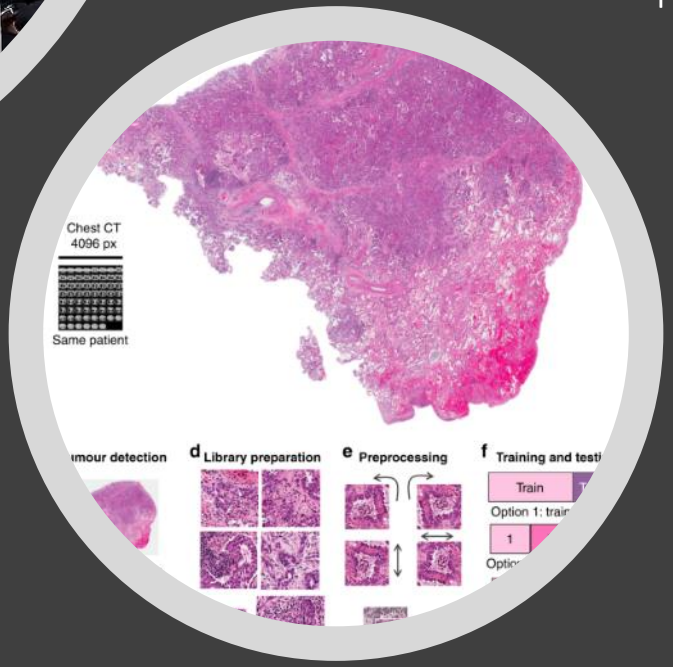


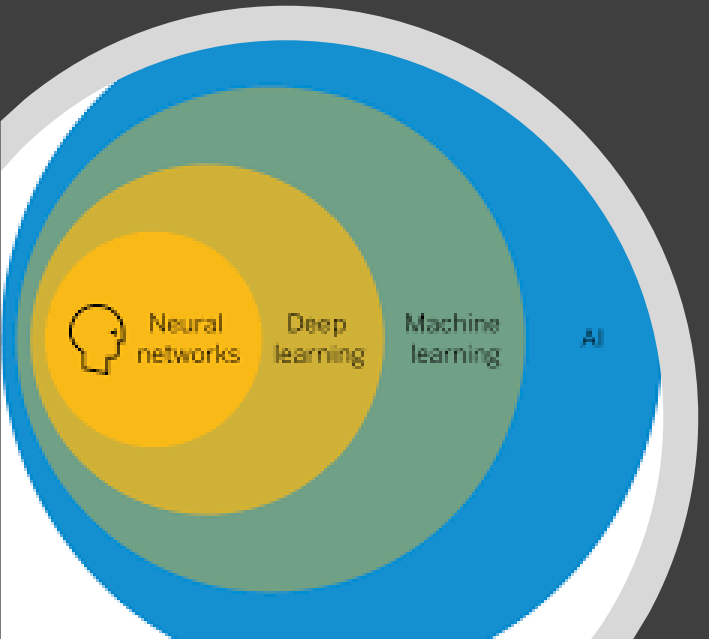
Laddered Processes Towards Building Library AI Awareness & Competencies

Awareness, Building Skills, Knowledge, Mastery



Multi-Year Process 2014-2022



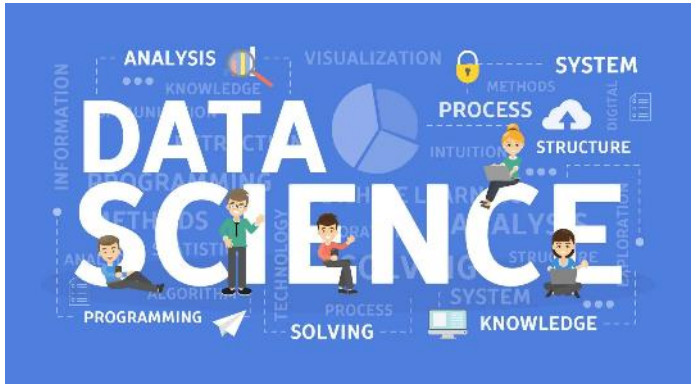


Last Ten Years Has Shown Incredible Progress of AI

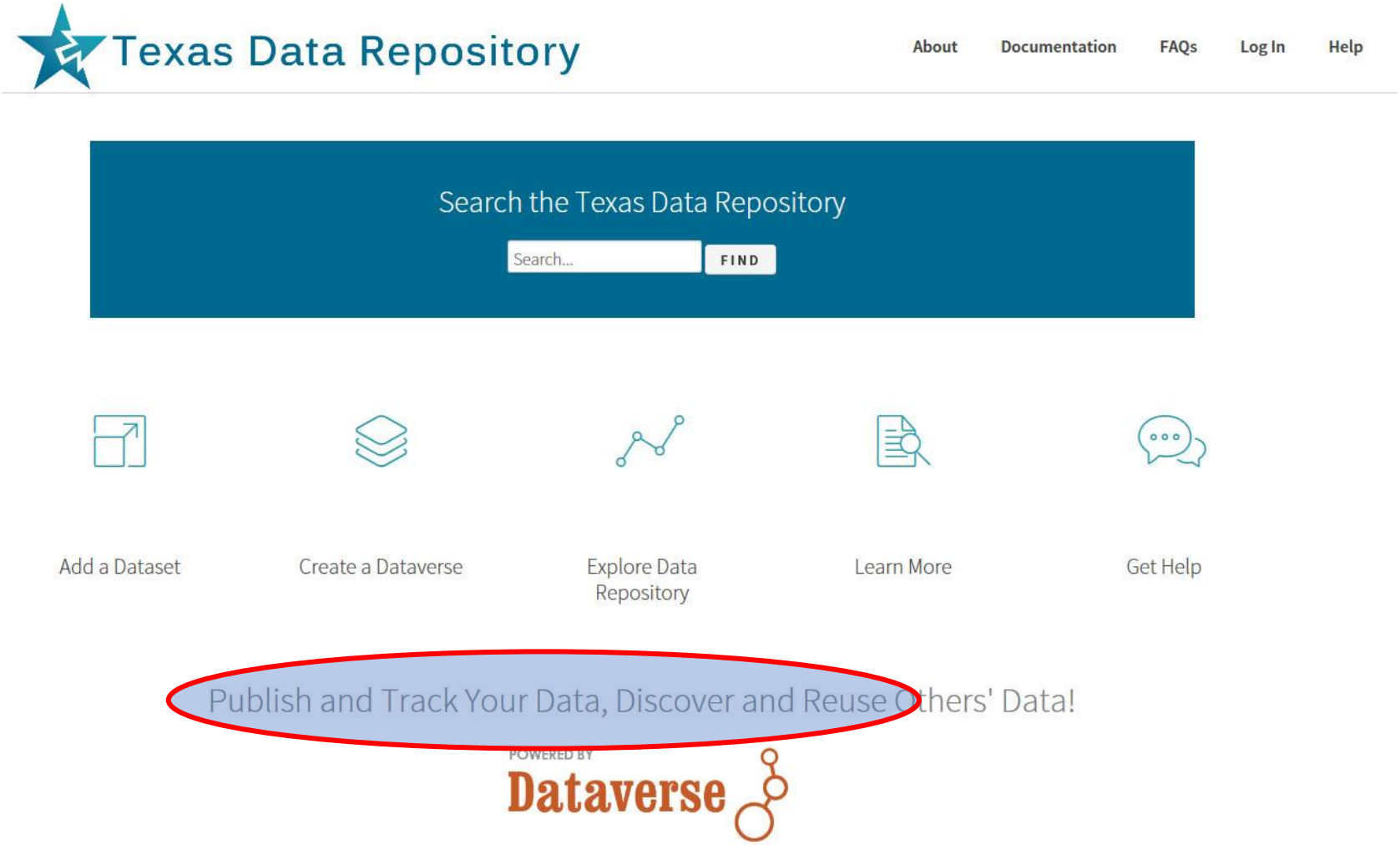
AI (Machine Learning (Deep Learning)) = Algorithms + Greater Computing Power + Large **Data Sets**

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

Clear Trajectory
in Libraries from
Data Collection
To Data Science ->
Data Repositories ->
Data Analytics ->
Data Visualization >
AI

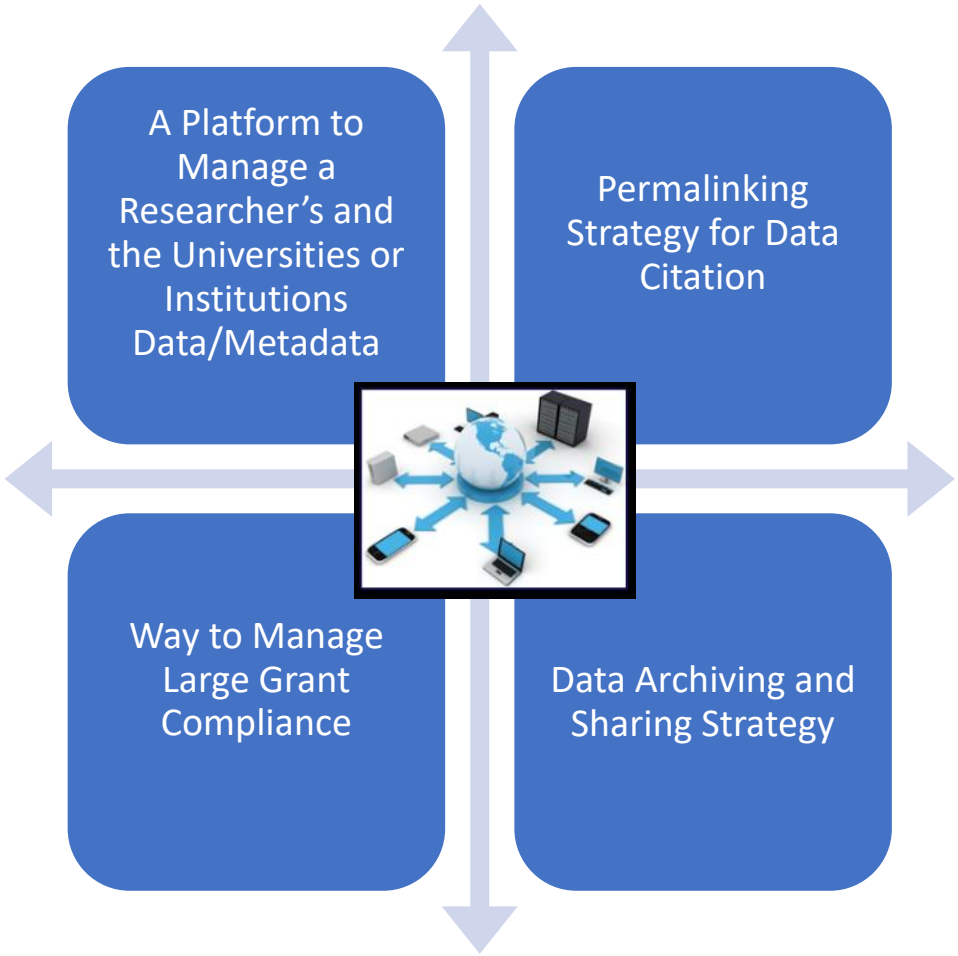


Begin with an Academic Data Research Repository

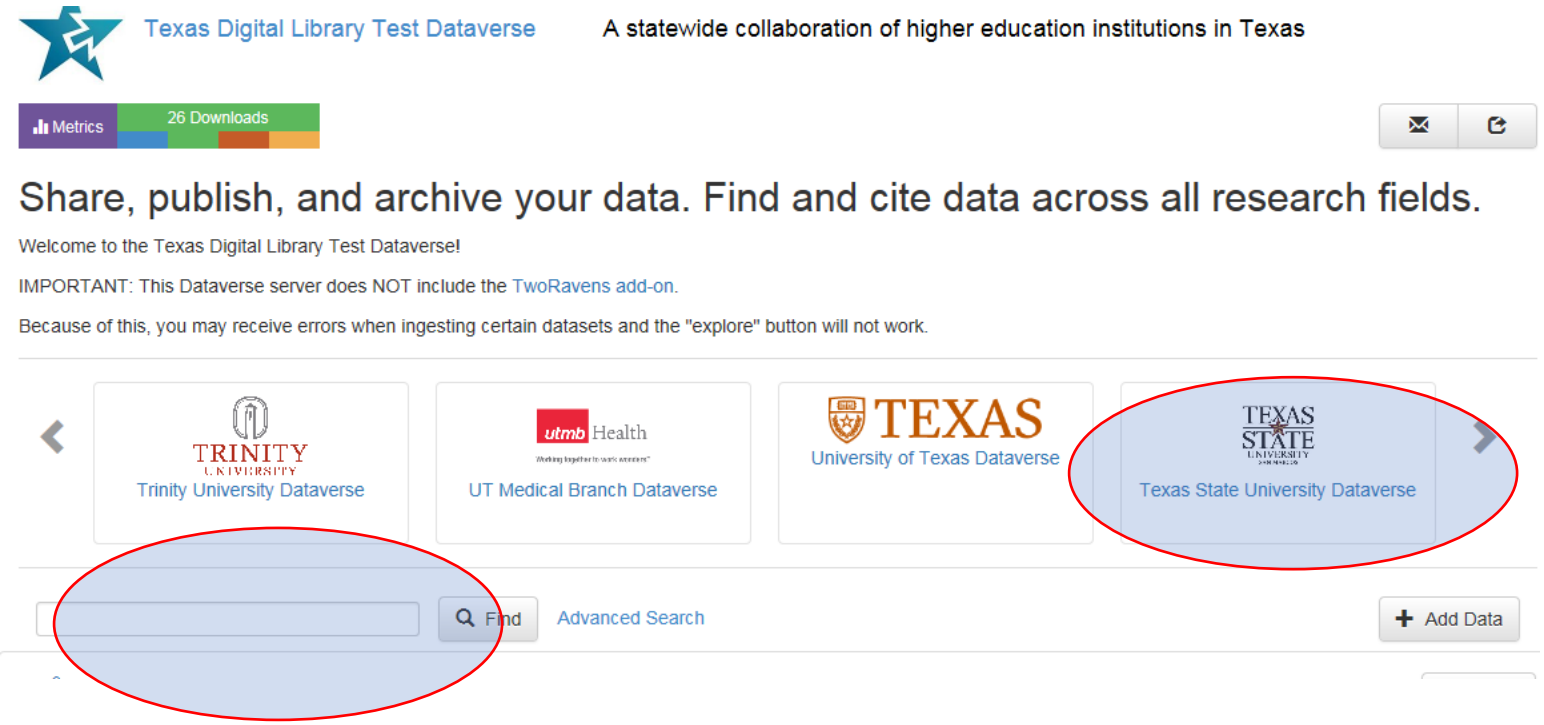


http://data.tdl.org

What is an Online Data Research Repository?



TEXAS RESEARCH DATA REPOSITORY



Texas Digital Library Consortium of 22 universities across Texas leveraging technological cooperation among academic libraries

Data Repositories Allow Building Skills For AI

Data Organization, Data Cleaning, Structured Data Citation, Sensitive Data and Metadata Schemas

OpenRefine is a powerful tool for working with messy data: cleaning it

Harvard Dataverse Network

Create Acc

REPLICATION DATA FOR: A MULTIVARIATE MODEL OF STRATEGIC ASSET ALLOCATION

hdl:1902.1/QBXRSFLBQJUNF:3:ZnYhHkZe2veTJAWaBDpPKA==

Version: 2 – Released: Thu Oct 03 16:46:32 EDT 2013

CATALOGING INFORMATION | Data & Analysis | Comments (0) | Versions

Data Citation

If you use these data, please add the following citation to your scholarly references. Why cite?

```
John Y. Campbell; Yeung L. Chan; and Luis Viceira, 2007, "Replication data for: A
Multivariate Model of Strategic Asset Allocation", http://hdl.handle.net/1902.1/QBXRSFLBQJ
UNF:3:ZnYhHkZe2veTJAWaBDpPKA== The Harvard Dataverse Network [Distributor] V2 [Version]
```

Citation Format Print

Original Publication

Results found in this publication can be replicated using these data.

Campbell, John Y.; Chan, Yeung Lewis; and Viceira, Luis M., 2003, "A multivariate model of strategic asset allocation," Journal of Finar Economics, Elsevier, vol. 67(1), pages 41-80: article available here

Publications

John Y. Campbell &Yeung Lewis Chan &Luis M. Viceira, 2001. "A Multivariate Model of Strategic Asset Allocation," NBER Working Pap National Bureau of Economic Research, Inc. article available here

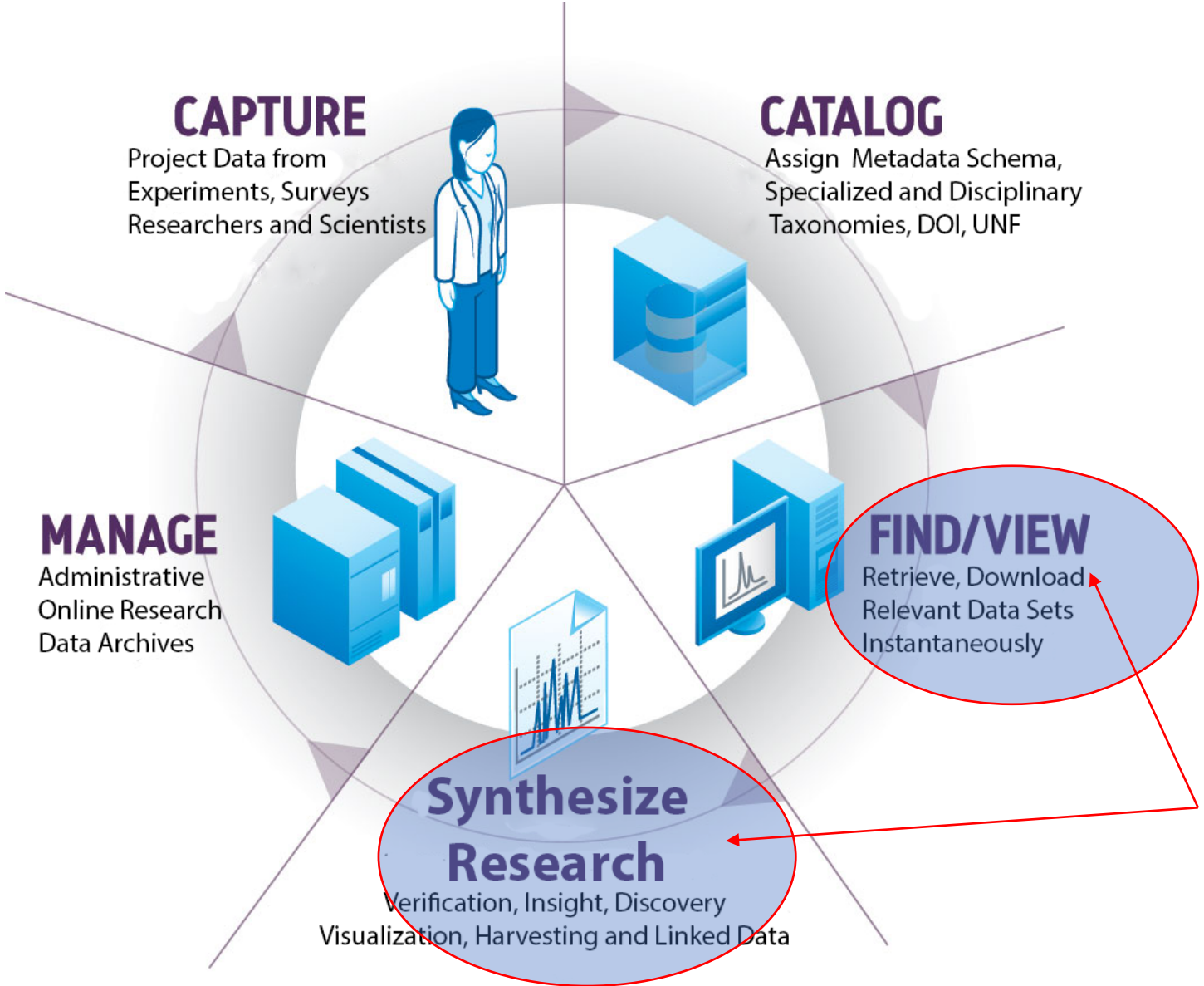
Campbell, John Y & Chan, Yeung Lewis & Viceira, Luis M, 2001. "A Multivariate Model of Strategic Asset Allocation," CEPR Discussior 3070, C.E.P.R. Discussion Papers. article available here

Data Citation Details

Title	Replication data for: A Multivariate Model of Strategic Asset Allocation
Study Global ID	hdl:1902.1/QBXRSFLBQJ
Authors	John Y. Campbell (Harvard University); Yeung L. Chan; and Luis Viceira
Producer	John Y. Campbell HARVARD Faculty of Arts and Sciences DEPARTMENT OF ECONOMICS
Production Date	2003
Funding Agency	National Science Foundation; Hong Kong RGC Competitive Earmarked Research Grant (HKUST 6965/01H); Division of Research of th Business School

The Research Data Repository Lifecycle

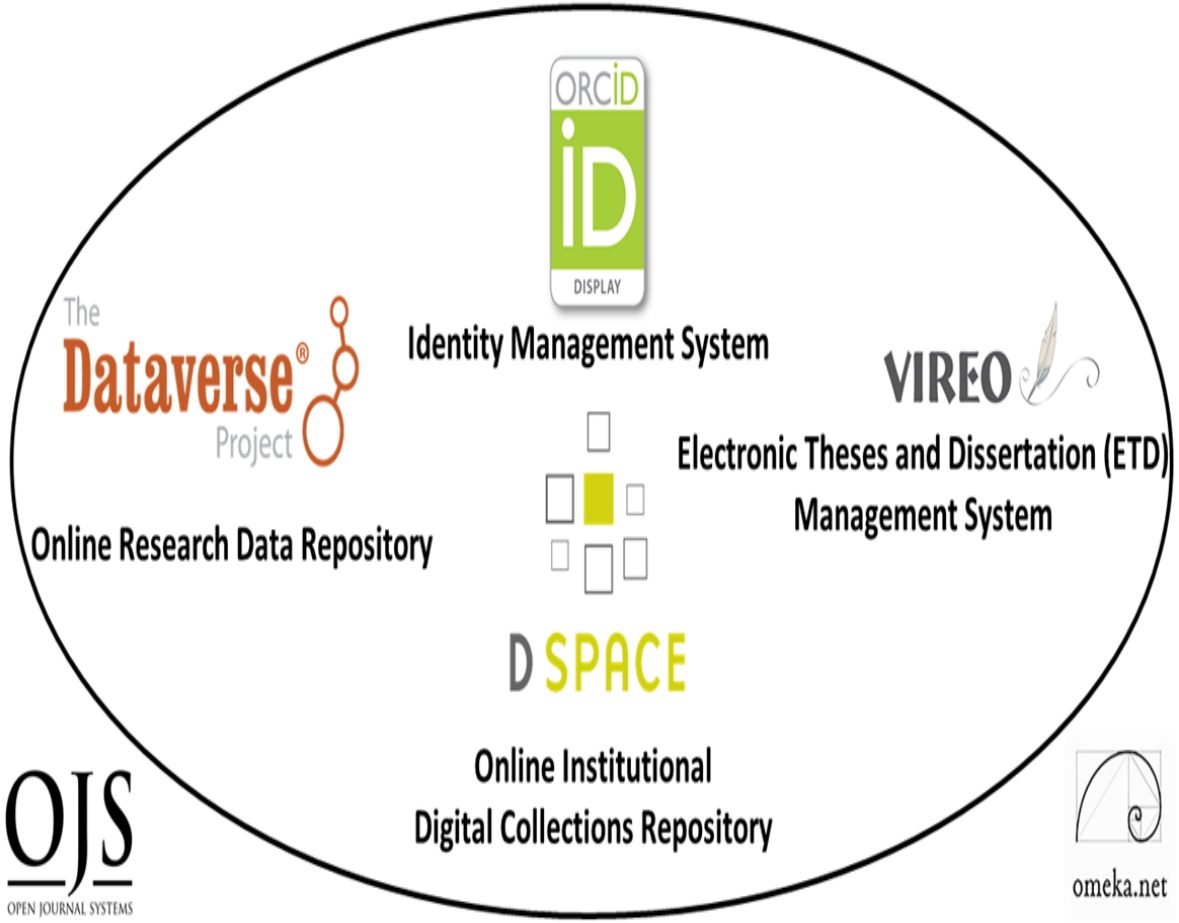
Setting Better Foundations & Organization for AI Infrastructures



Data Repository provides
Basic AI, Machine Learning, Open Science and Research Needs.

Digital Scholarship Ecosystems, Foundations for AI

Six Open Source Software Components



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

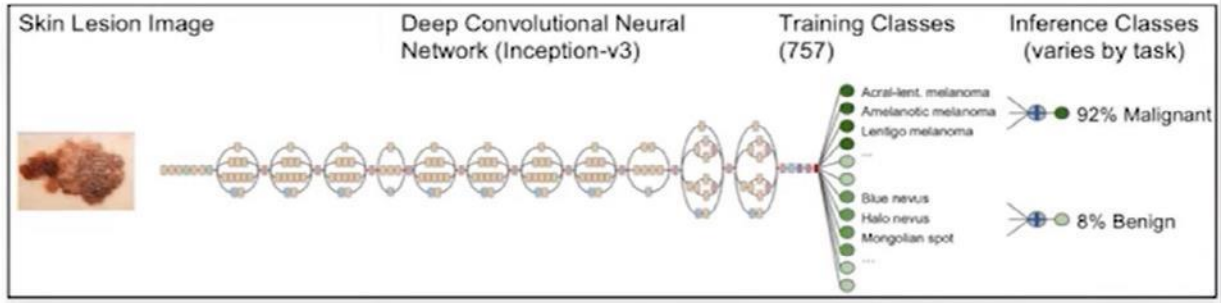
Dermatologist-level Classification of Skin Cancer with Deep Neural Networks,

Nature 2017, Andre Esteva, Brett Kupress, Sebastian Thrun et al.

AI Models, Deep Learning, Convolutional Neural Nets, Labeled Medical Data from Image Data Archives

Skin Cancer Diagnosis:

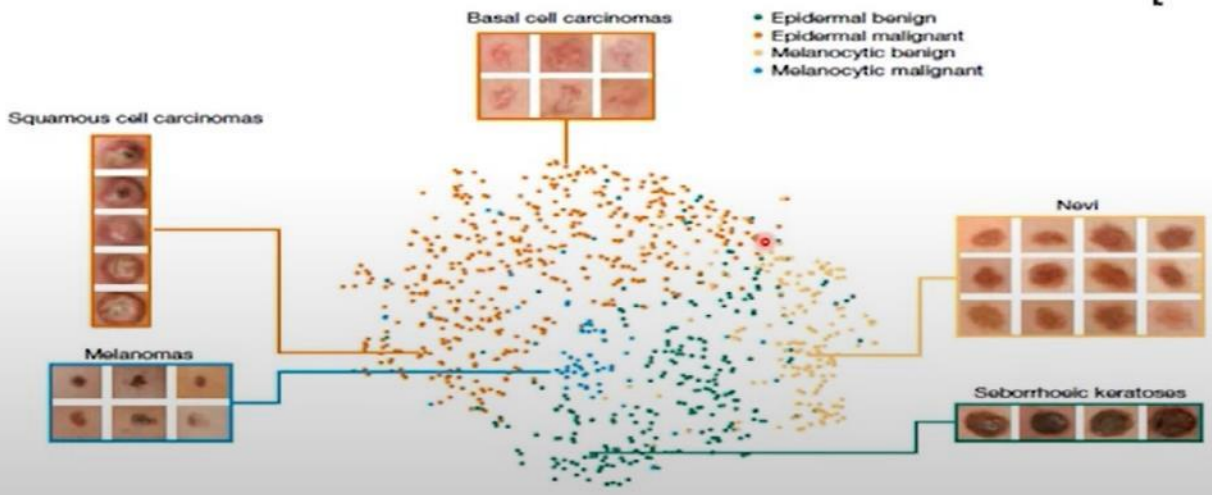
Trained on 1.4 M standard photographs
Retrained on 129,450 skin images
Deep net Inception v3 architecture
Outperforms doctors



[Esteva et al., *Nature* 2017]

[Video](#)

[Stanford Overview](#)



Open Science, **Data Research Repositories**, Discovery and AI

- **Harvard Dataverse Data Repository**
Dermatology Image Dataset,
Dr. Philip Tschandl,
Viennesse Dermatologist
Great Example of Open Science
- https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Sign Up | Log In

ViDIR Dataverse

(Medical University of Vienna)

Harvard Dataverse > ViDIR Dataverse >

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`).

Subject	Medicine, Health and Life Sciences; Computer and Information Science
Keyword	Dermatoscopy
Related Publication	Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 (2018). doi: 10.1038/sdata.2018.161
License/Data Use Agreement	Custom Dataset Terms

Files | Metadata | Terms | Versions

Search this dataset...

Filter by

File Type: All Access: All

Sort

1 to 6 of 6 Files

Download

HAM10000_images_part_1.zip
ZIP Archive - 1.3 GB
Published Jun 4, 2018
15,433 Downloads
MD5: 463...e46

HAM10000_images_part_2.zip
ZIP Archive - 1.3 GB
Published Jun 4, 2018
11,809 Downloads
MD5: da4...84b

HAM10000_metadata.tab
Tabular Data - 810.9 KB
Published Jan 29, 2021
5,938 Downloads
8 Variables, 10015 Observations UNF:6:WcXi...myQ==

Open Access Data Repository Metadata and Data for Download(Images)

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

2017 Stanford
Nature Deep Learning Cancer ID Article

2018 Harvard Dataverse Datarepository Upload
Open Source Viennesse Dermatalogical Image Library

November 2021
Dspace Repository Undergraduate Thesis BRAC University, Dhaka Bangladesh, Dept. of Computer Science and Engineering
Downloaded July 2022

BRAC University Institutional Repository

Digital Collections Repository

Dspace
http://dspace.bracu.ac.bd/

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

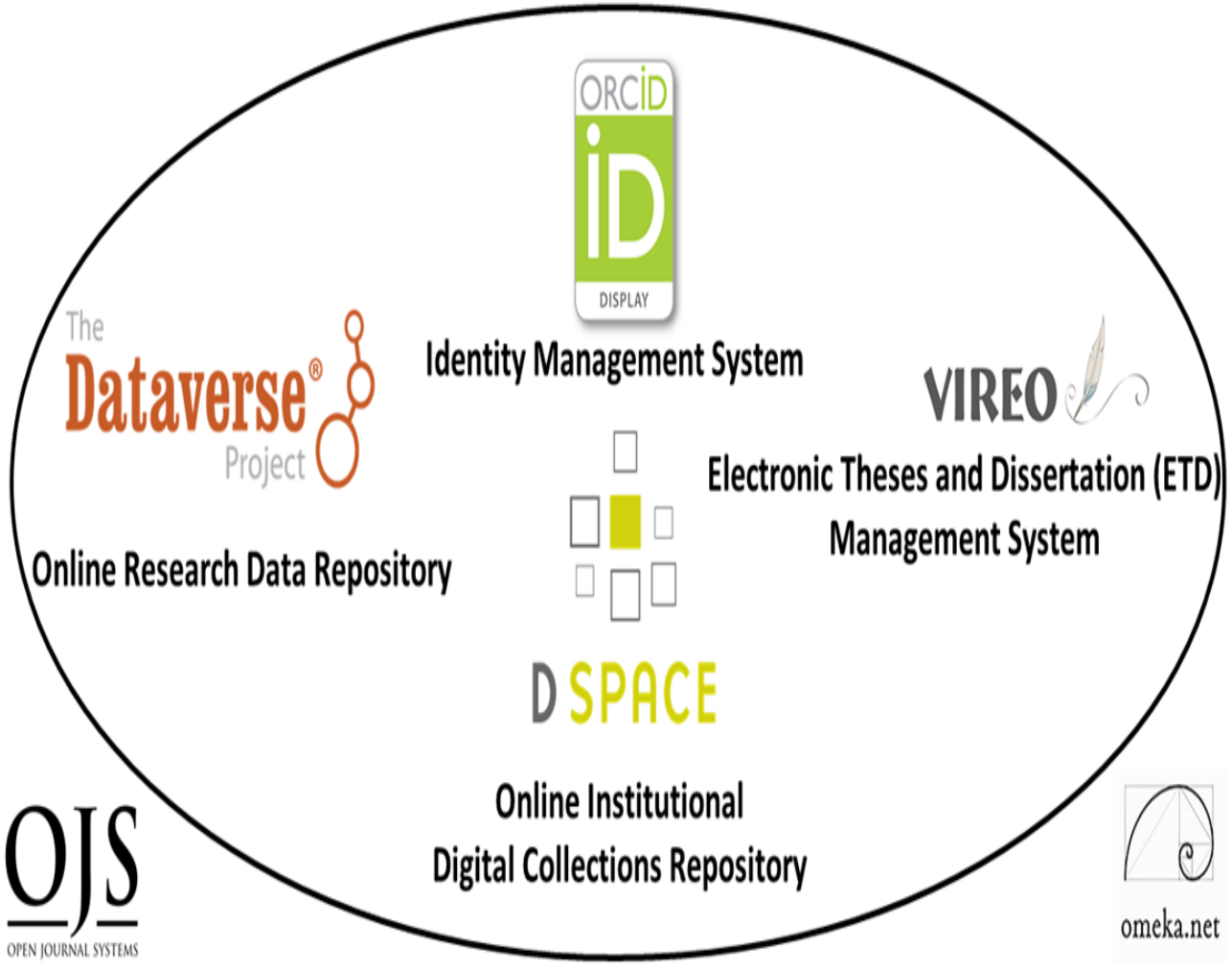
Search

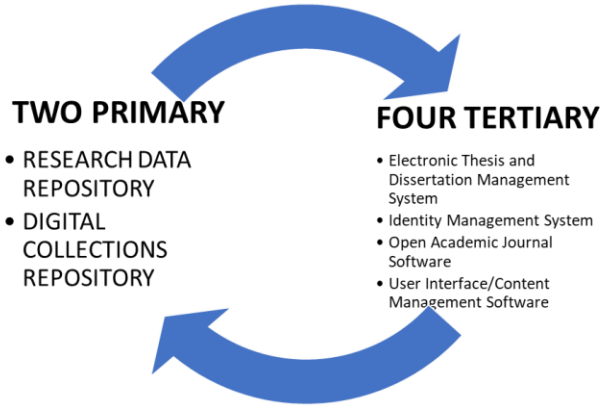
Search BracU IR
This Collection

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 Scholarship Ecosystem Centered on Research Data Repository and Collections Repository





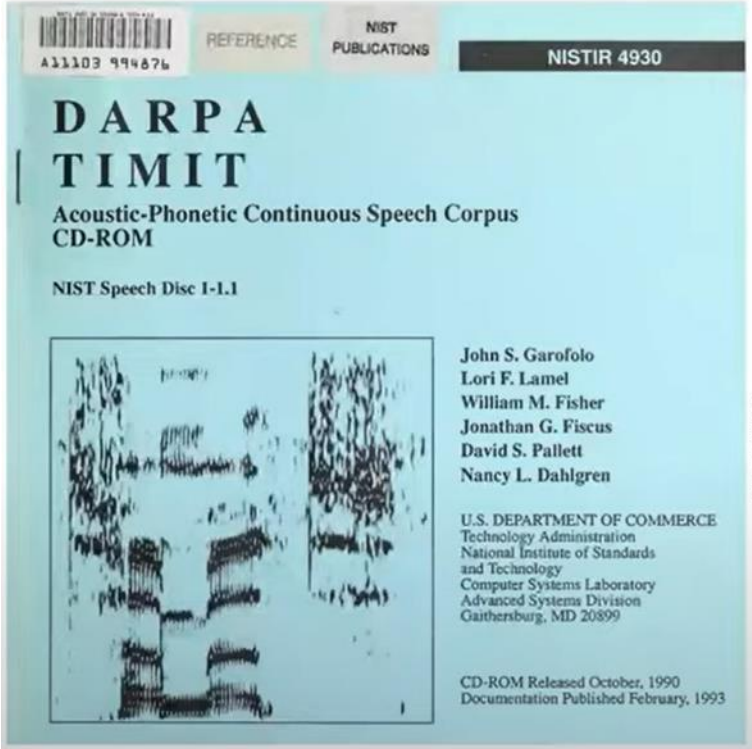
Many Useful Data Science Skills for AI Will Be Learned Here

Metadata Schemas
Data Organization
Data Cleaning
Data Classification
Creating Dataset Benchmarks
Standardization of Data

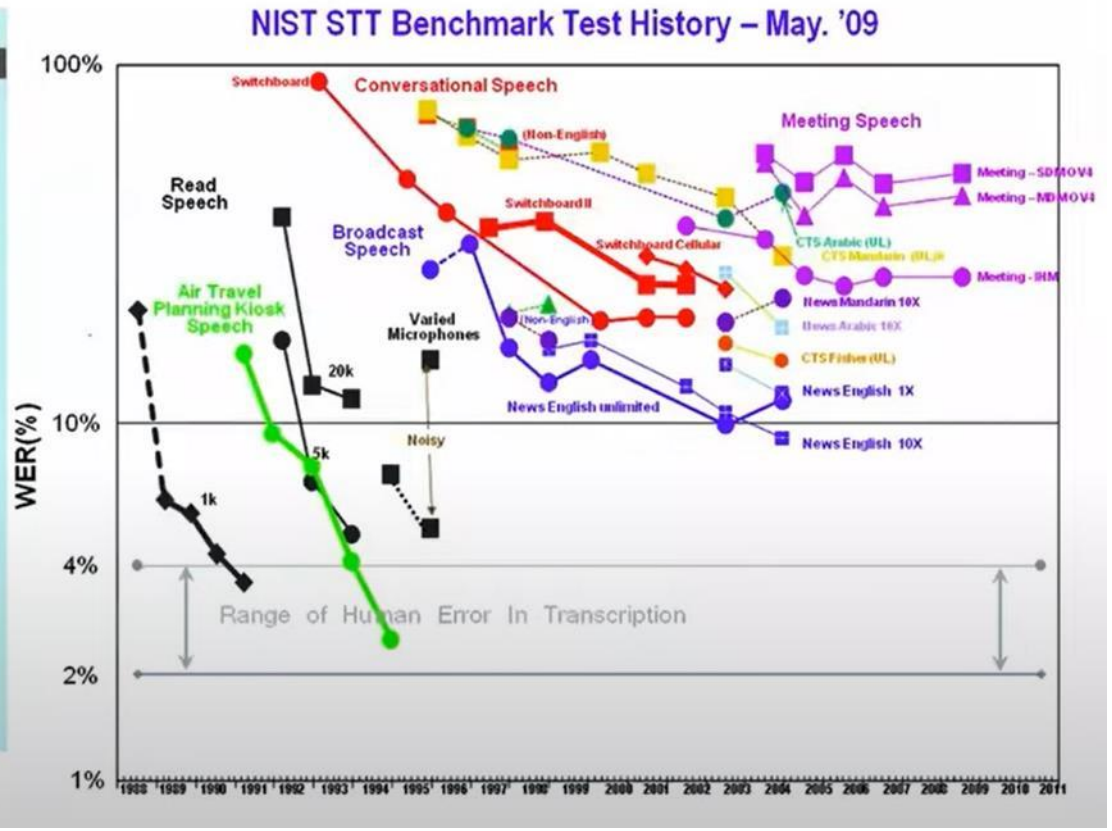
Data Repositories Allow Benchmarking

Measurement Leads to Comparison , Improvement and Progress of Human Knowledge

TIMIT



Many attribute to PM Charles Wayne



Imagenet
(Most Famous
large visual database)

Alexnet, 2012
Convolutional
Neural Network
Model
15% Error Rate

Human Resource Infrastructures Part I (Working Teams)

Future Hires

Machine Learning/Deep Learning/AI Specialist/ Data Scientist and/or AI Librarian (working with the data)

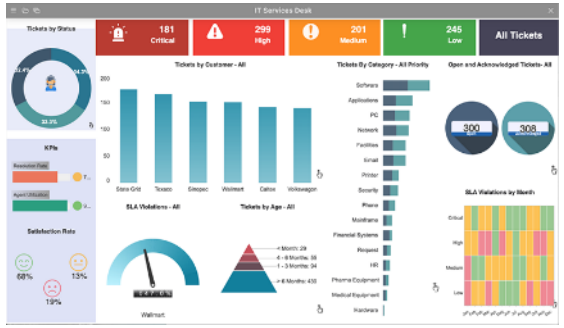
Data Visualization and Analytics Specialist (Tableau, Bayesia, Power BI)

Committee for Data Repository Workflows & Policies

Onsite Staff Skills
Metadata Cataloger
Data Repository Faculty/Student Liaison
Subject Liaisons (Outreach)

Current Staff
Digital Collections Librarian
(**Texas State Data Repository Librarian**
Dataverse/Publications Repository: D-Space)

DATA VISUALIZATION & ANALYTICS SPECIALIST



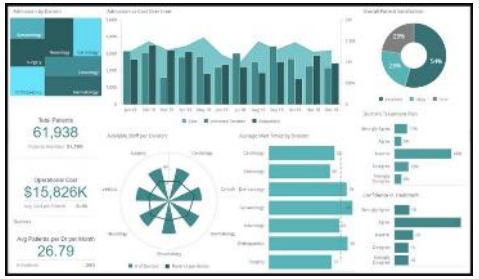
Texas State University Libraries is seeking Data Visualization & Analytics Specialist to provide library-wide support for data visualization and data analytics projects to support data-driven decision making and finding insights. This position requires a higher level of technology expertise and specialized knowledge to gather, manage, and analyze data and report complex data in easy-to-understand information visualizations.

RESPONSIBILITIES: Develop and maintain a data visualization and analytics strategy. Develop strategies to clean and normalize data for use in further analysis. Utilize data visualization strategies to report and present analytics and answer questions related to data analytics and data visualization. Pursue professional development activities to improve knowledge, skills, and abilities and perform special projects and other duties as needed.

QUALIFICATIONS:

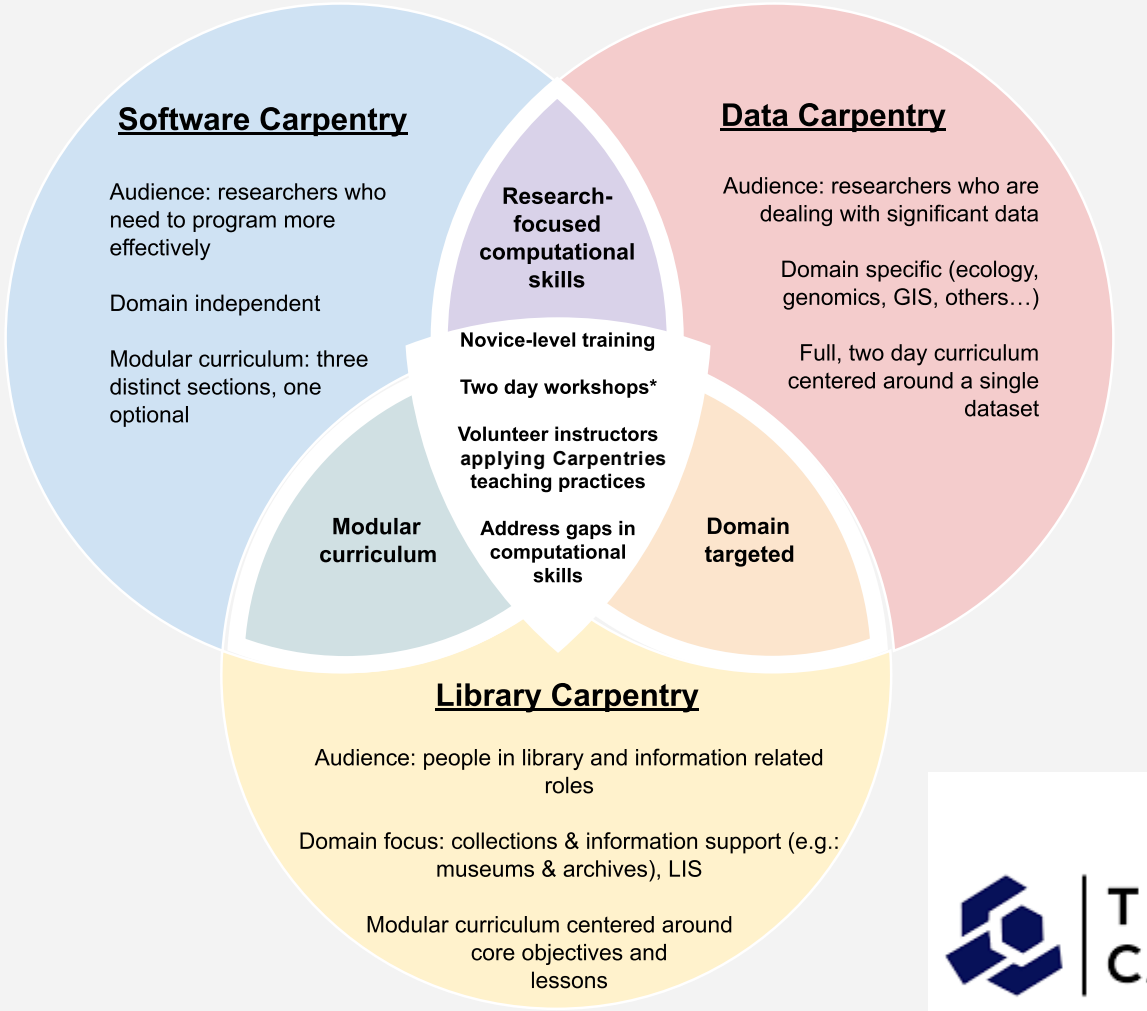
- **Required**: Ability to read, analyze, and understand data in a variety of formats; strong written, oral, and interpersonal skills, including ability to work effectively in a team; knowledge of data visualization applications such as PowerBI, Tableau or others; analytical skills; proficiency with Microsoft Excel; ability to utilize analytics/visualization tools in new, creative, and effective ways.

- **Preferred:** Degree in information science, applied statistics, business analytics, computer science or another quantitative or data visualization field; experience with SQL or other query language; experience with R, Python, statistical analysis languages, predictive analytics, and/or AI software.



Learning Paths: From Data to Carpentries

Foundational Coding and Data Science Skills for researchers Worldwide



DATA CARPENTRY

Libraries Can Host Carpentry Workshops

https://carpentries.org/

THE CARPENTRIES

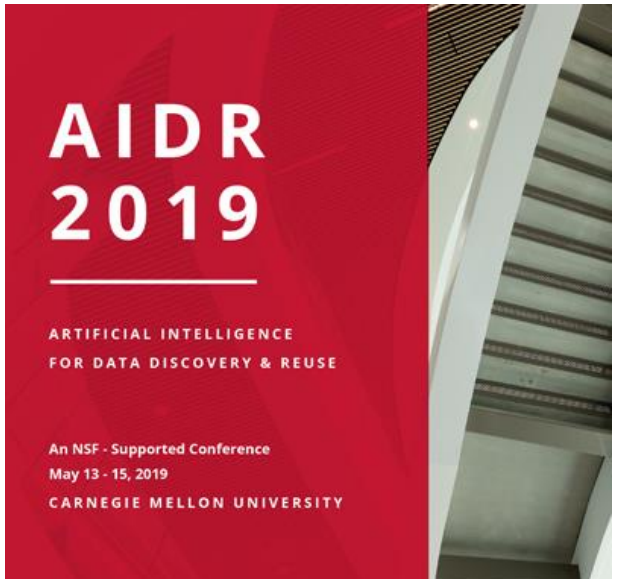
Conferences and Learning

Library IT and Digital Services May Be Getting Interested in AI

Fantastic Futures
2nd International Conference on AI for Libraries, Archives and Museums
Stanford Libaries (2019)

Artificial Intelligence for Data Discovery & ReUse & Open Science Symposium (2020)

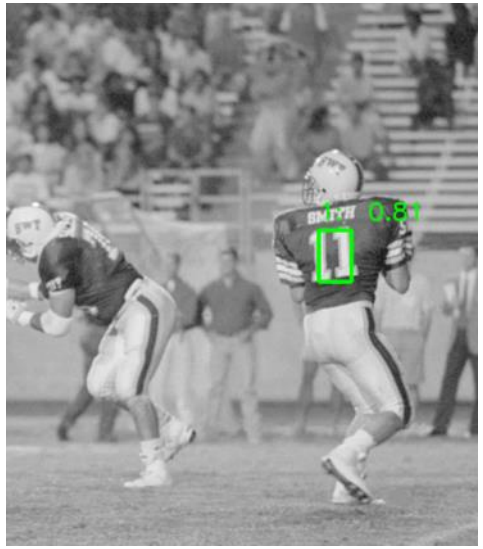
Texas Conference on Digital Libraries, Patrice Andre Prud'homme (TCDL) Oklahoma State, Computers in Libraries, Yale Art History,Pixplot (Image Categorization, CNI, Compters in Libraries

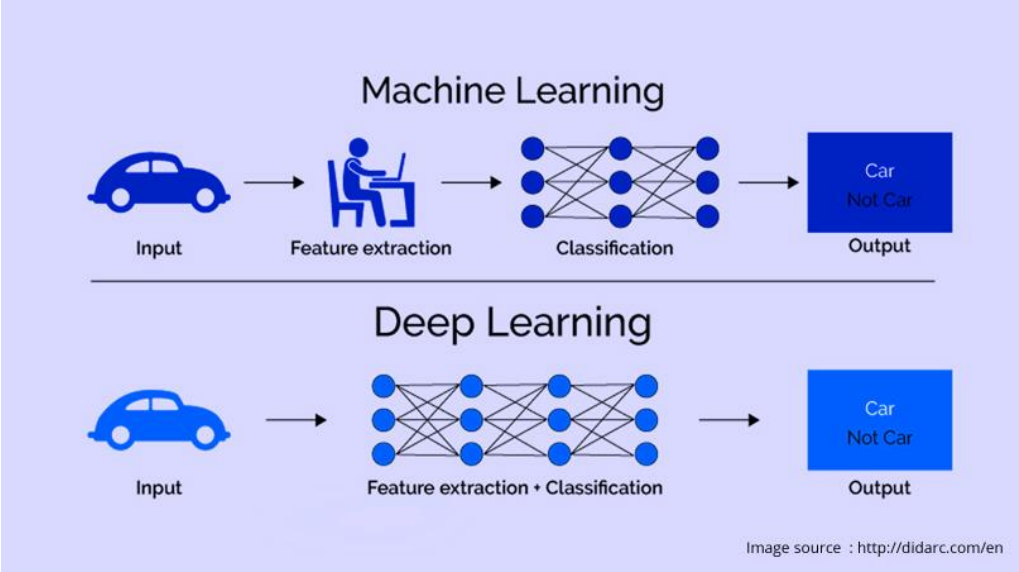


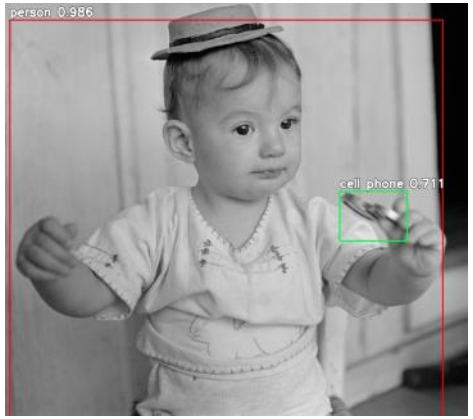
Digital and Web Services R&D & Learning

Convolutional Neural Nets and Deep Learning Models

- Processing Power
  (Compute)
- Python
- Video Cards
  (NVIDIA GPU's)
- Pretrained Models
- ResNet, YOLO, COCO
  (200k labeled images, 80 categories)
- University Archives
  San Marcos Public
  Newspaper Image Negatives
  90 years of digitization
  800, 000 images)

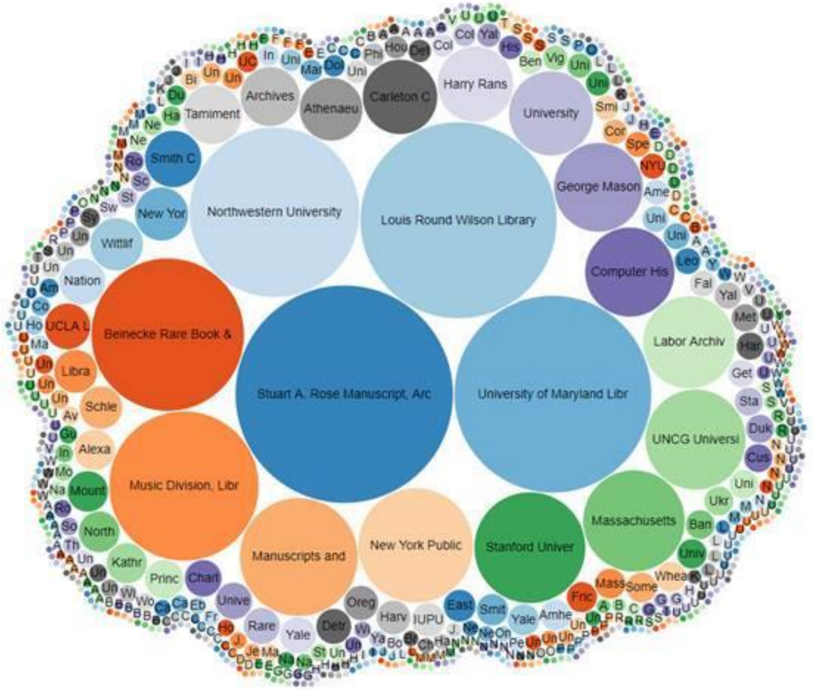


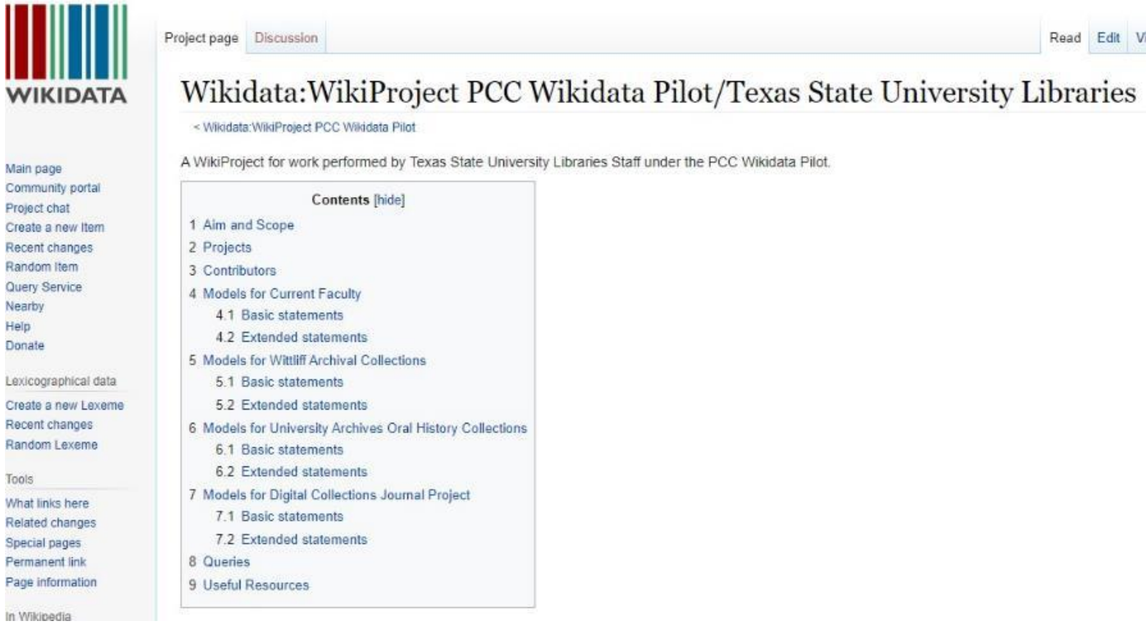




Cataloging, Metadata, Wikidata, Semantic Web, AI

- Metadata Services Cataloger
- Crosswalking between Systems
- Successful Linked Data Project
  Wikidata Semantic Web Project: Faculty, Oral History Collections, Wittliff Archives, OJS Journals
  (Data that is Machine Readable, ie. Google etc)
  Moving from MARC Silos to Online (Wikidata)
- Learning Many Data Science Skills, Data Models, Data Batch & Cleaning Tools: OpenRefine, Quickstatements,
  Python (7 staff)





Wikidata:WikiProject PCC Wikidata Pilot/Texas State University Libraries

< Wikidata:WikiProject PCC Wikidata Pilot

A WikiProject for work performed by Texas State University Libraries Staff under the PCC Wikidata Pilot.

Contents [hide]

1 Aim and Scope
2 Projects
3 Contributors
4 Models for Current Faculty
  4.1 Basic statements
  4.2 Extended statements
5 Models for Wittliff Archival Collections
  5.1 Basic statements
  5.2 Extended statements
6 Models for University Archives Oral History Collections
  6.1 Basic statements
  6.2 Extended statements
7 Models for Digital Collections Journal Project
  7.1 Basic statements
  7.2 Extended statements
8 Queries
9 Useful Resources

IDEA Institute on Artificial Intelligence

(Recommendation Letter, July, 2022)

- Week Long Fellows Program at University of Texas Austin (20 Fellows)
- Onboarding, Institute, Library Centered AI, Final Project
- Networking with National Library AI Experts and Other Fellows

IDEA INSTITUTE ON ARTIFICIAL INTELLIGENCE

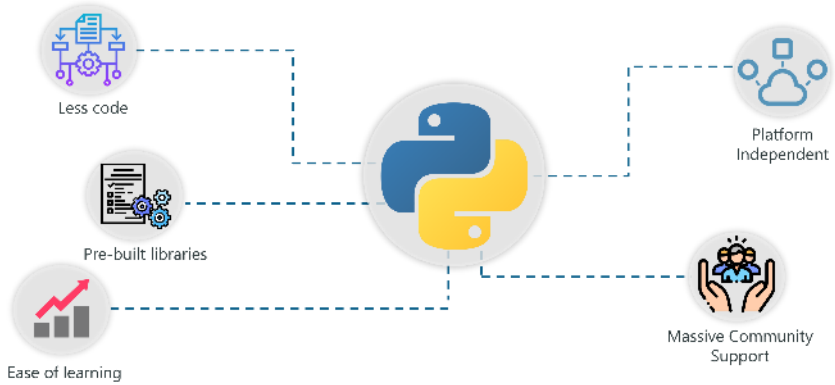
- AI challenges and opportunities, Ethical considerations and guidelines
- UX-Human/AI Interaction Lifecycle
- Existing library, archive, and museum projects
- AI project planning
  - Project Design
  - Data collection, classification, and transformation
  - Roles and implementation
- Python Basics, Python for Machine Learning
- APIs and bibliometrics
- AI in search and discovery
- Machine learning and coding
- Harvesting, evaluating, and training data sets for use in AI
- Conversational AI – Theoretical foundations
- Conversational AI – applications
- Linked open data Machine learning for text with topic modeling and clustering

INSTITUTE of Museum and Library SERVICES

Steps Towards AI: Learning Python, Spring 2022

- Hi Ray,

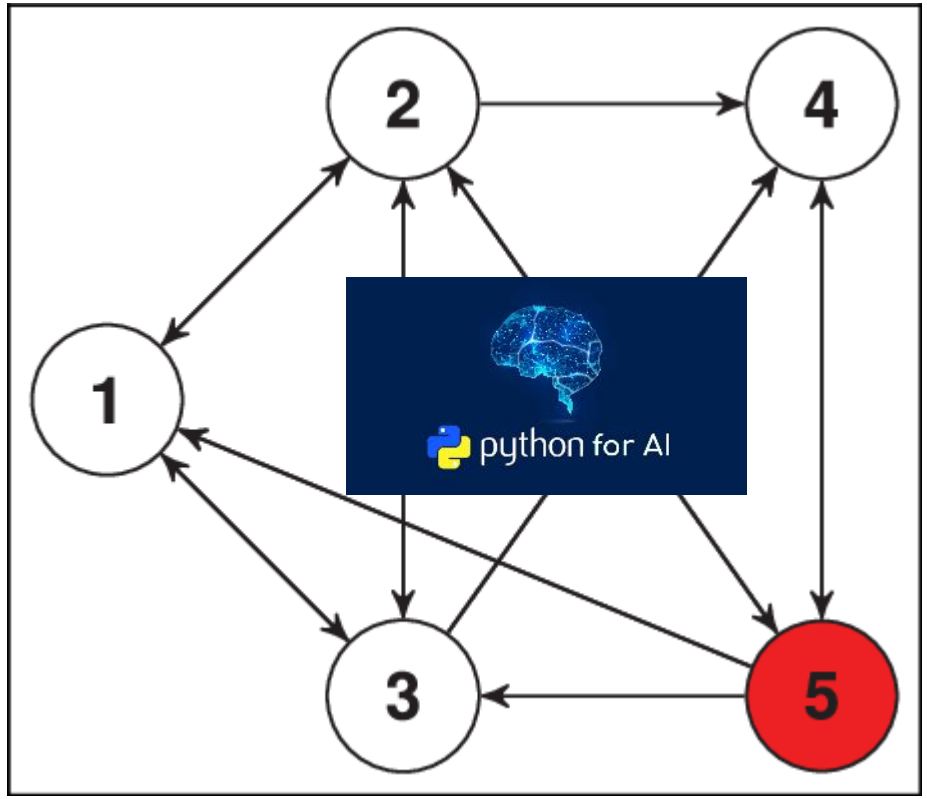
I wanted to let you know that we've started a Teams Group for myself, Carol, Alex and Amanda named "Python learners" so we could share tips and help each other on our various learning paths in an encouraging, safe space. ☺ ---Mary



Courses: Getting up to Speed with Python, Python and Machine Learning
Why Python for AI? – Artificial Intelligence with Python

Carol, Library Management System (LMS) Usage Data Insights

Mary, Metadata AI Extraction



Amanda, Collections Budget, Insight and Analytics

Alexandria Collections Analytics, Data Visualization

Todd/Jason Image Recognition Neural Nets Part II?

Ocelot Chatbot Administrator

AVP University Librarian Dissolves Research and Information Outreach Services (Reference & Subject Librarians)

↓

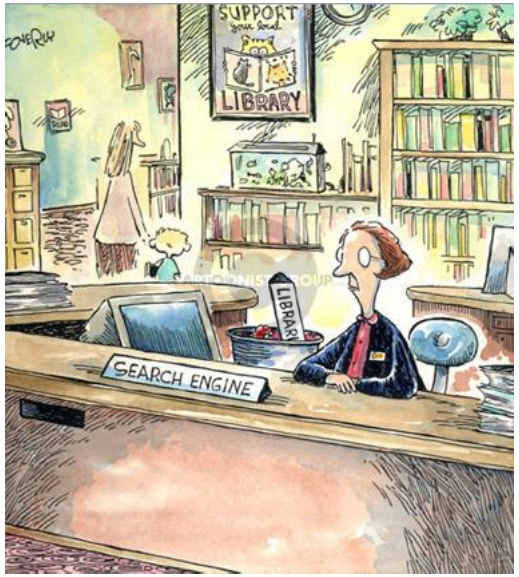
University IT Adopts New Ocelot Chatbot Infrastructure

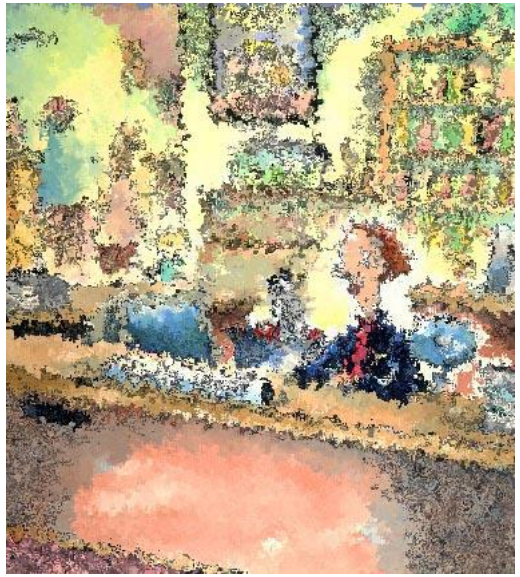
↓

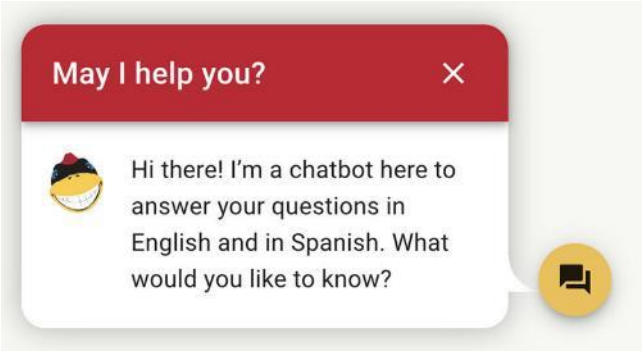
Digital and Collection Services receives New Libraries Chatbot Administrator

↓

Future Natural Language Processing R&D (GPT3-4, DeepMind Gopher)

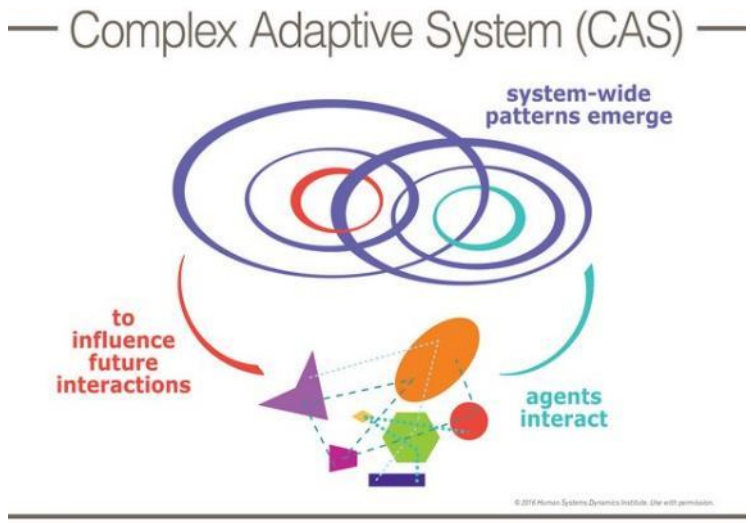


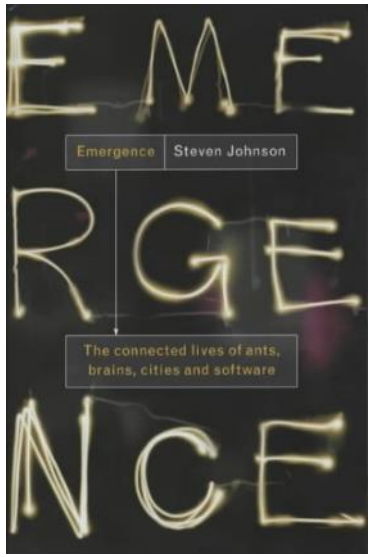




Emergence, Chaos Theory Complexity, Genetic Algorithms

Libraries are Complex Adaptive Non-Linear Dynamic Systems





TXU AI WORKING GROUP, (AIWG)

Forming an Artificial Intelligence Working Group (AIPG)
Digital Preservation Infrastructure Working Group 2017-2021 (DPWG)
Antecedent Models: Online Data Research Repository State Working Group 2014-2017, ODRWG)
(Texas State University Libraries Example, Enthusiastic Motivated Chair)

Purpose: An AI Working Group starts conversations, provides, direction, responsibility and accountability for:

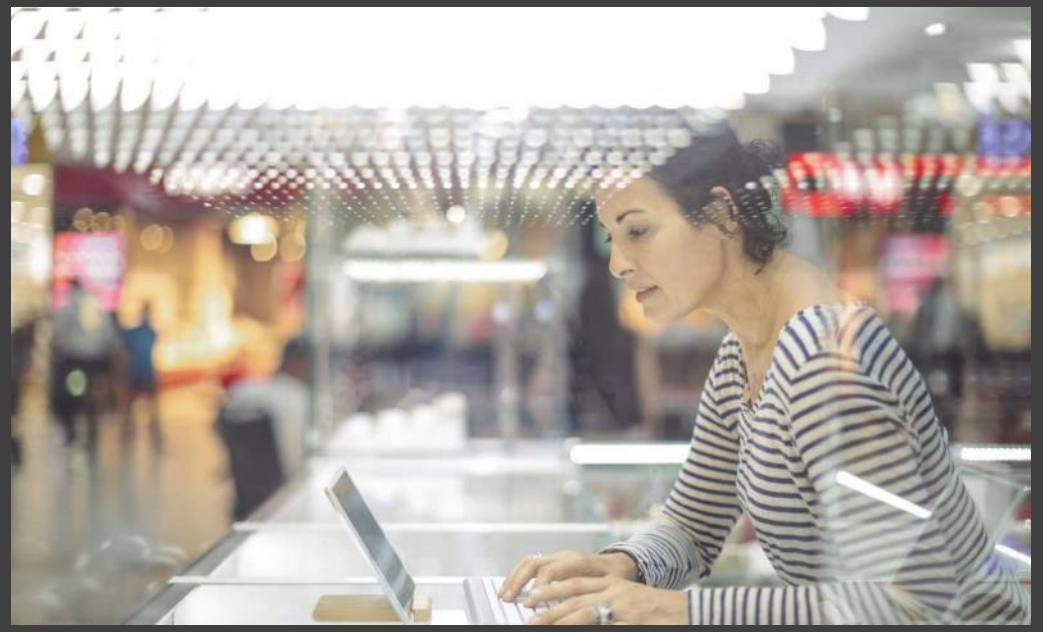
1) Artificial Intelligence Project, Policy Ethics related discussions
2) Later Oversight

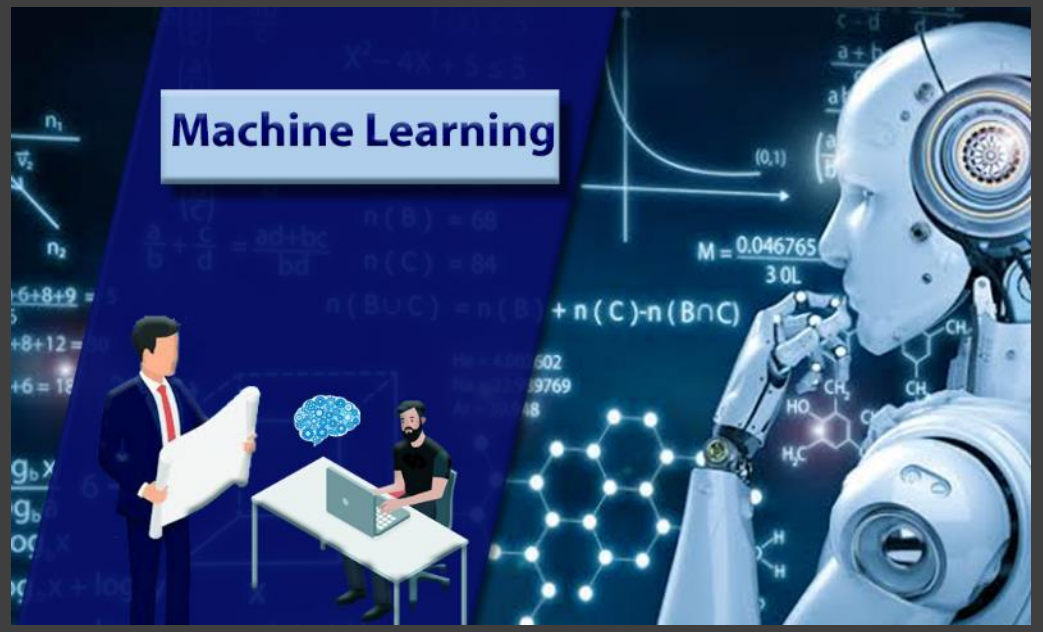
Membership: Metadata/Cataloging, Acquisitions, Digital and Web Services, Special Collections, New Tech/GIS Lab (Alkek1), and Research and Information Services - New Chatbot Admnistrator. (9-10 Staff).

Next Steps: A Graduate Student or Two

- Any University Engineering School or Computer Science DepartmentWill Have Graduate Students and Courses Like This:
- E5331 – **Machine Learning for Engineering Applications** (2019-2021) [*Listed Fall – 2022*]
- EE4331 – **Intro to ML for Engineering Applications** (2020-2021) [*Listed Fall – 2022*]

  Form a Relationship with the Electrical Engineering Professor and Hire His Graduate Student to Help with your team as part of their final project, graduate theses or part time Research Assistant to provide computational assistance and resources. These can be Masters Candidates or good advanced undergraduate students.





Future Steps: AI Postdoctoral Fellows and Permanent Library AI Positions

Postdoctoral Fellowship Program offers recent AI/Machine Learning related Ph.D. graduates the chance to develop research tools, resources, and services while exploring new career opportunities and opening Library possibilities.

https://www.clir.org/ https://www.clir.org/global/

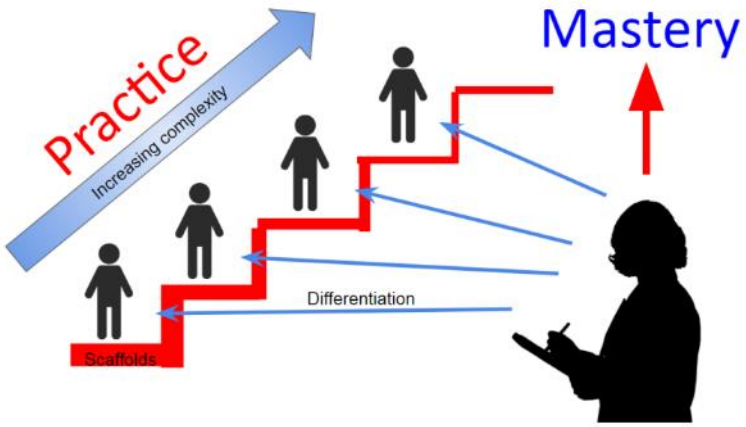
Postdoctoral AI Fellows work with library staff, faculty and graduate students on library related projects that forge and strengthen connections among library collections, archives, special collections digital technologies, and their current AI research and skills.

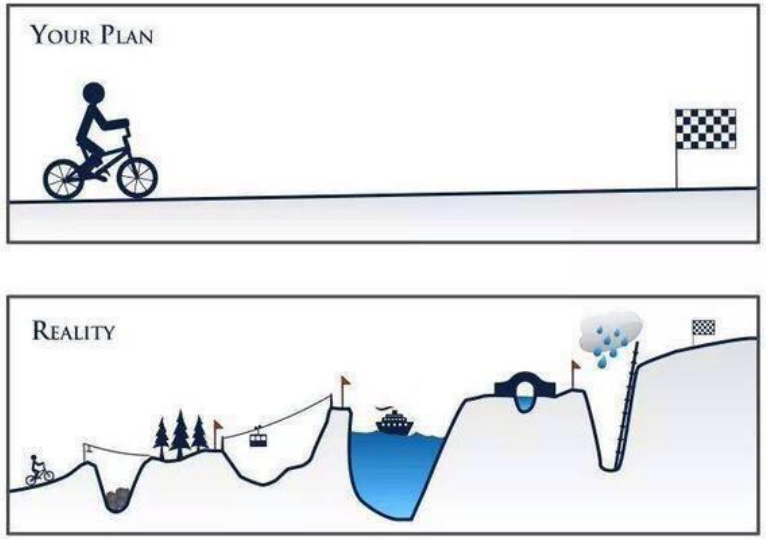
CLIR POSTDOCTORAL FELLOWSHIP PROGRAM

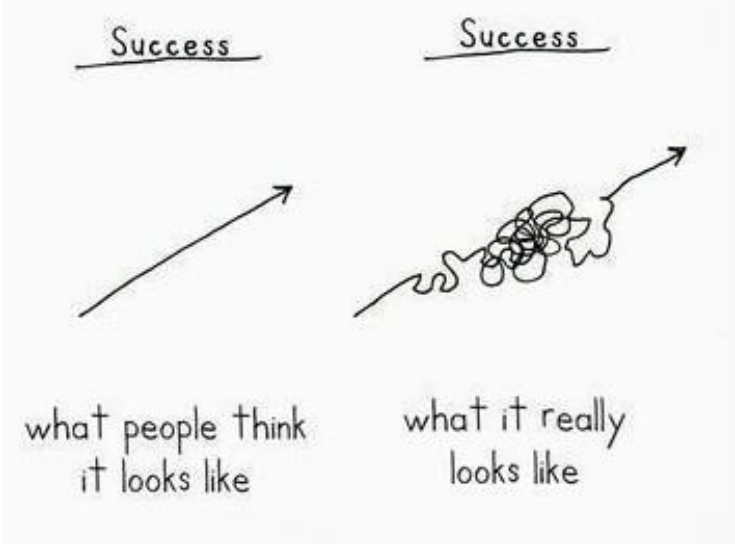
Hub for AI Research in Archives

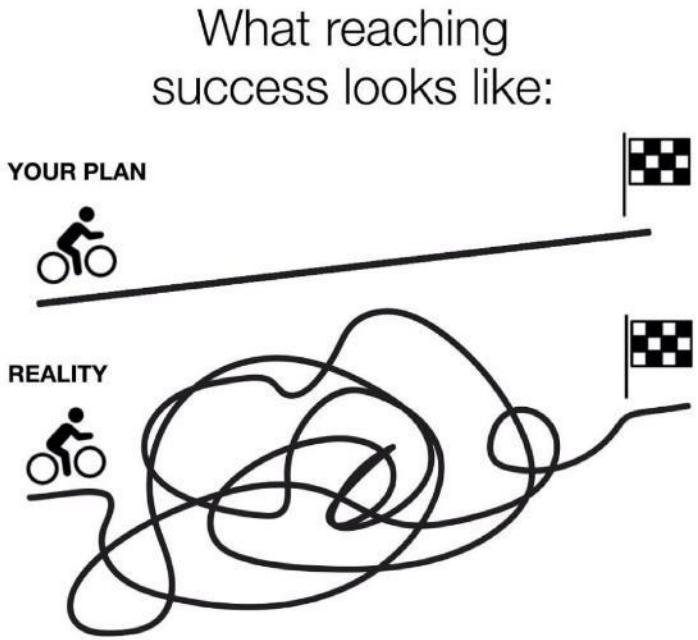
https://postdoc.clir.org/ https://haira.clir.org/blog/

Steps and Ideas For Scaffolding Towards Library AI Projects and Foundational Infrastructure Success









Questions/Comments

- Ray Uzwyshyn, Ph.D. MBA MLIS
  Director, Collections and Digital Services
  Texas State University Libraries, USA
  ruzwyshyn@txstate.edu , http://rayuzwyshyn.net
  July 2022

Texas State Repositories Architecture