Machine Learning in Research Libraries

A Snapshot of Projects, Opportunities and Challenges

Liz Lorang, Nebraska-Lincoln
Harish Maringanti, Utah
John Wang, Notre Dame

CNI Fall 19

IMLS Planning Grant (LG-72-18-0221-18)

- How can Machine Learning (ML) better facilitate cross-disciplinary discovery?
- Where are we?
- Based on what we learned, what should we do next?

Convocate

- Interdisciplinary project
- Scholars, students, librarians, IT
- Controlled vocabularies
- ML potentials

Convocate website | CNI presentation | IFLA article

Grant Components

- Literature review - Done
- Environmental scan (Survey) - Done
- Series of workshops - Done
- Writer workshop (Addition) - Done
- White paper - In progress
- Open access book (Addition) - In progress

Grant Numbers

- 324 respondents
- 4 ML workshops
- 24 speakers
- 104 workshop Attendees
- 21 authors
- 19 chapters
- 1 author workshop

Challenge: Cross Disciplines

- University structural constraints
- Organic approach, or
- Institution level stimulus

Challenges and Opportunities

- Capacity to learn, experiment, and test
- Resources to support innovation
- ML curriculum for academics
- Commercial algorithms for scholarship
- Data-centric nature of the technology
- Collaboration between faculty and library

ALL U NEED

J. Willard Marriott Library
THE UNIVERSITY OF UTAH

Sheeko: A computational helper to describe digital images

Harish Maringanti
Associate Dean of IT & Digital Library

Vivek Srikumar
Assistant Professor, Computer Science

Dhanushka Samarakoon
Assistant Head of Software Development

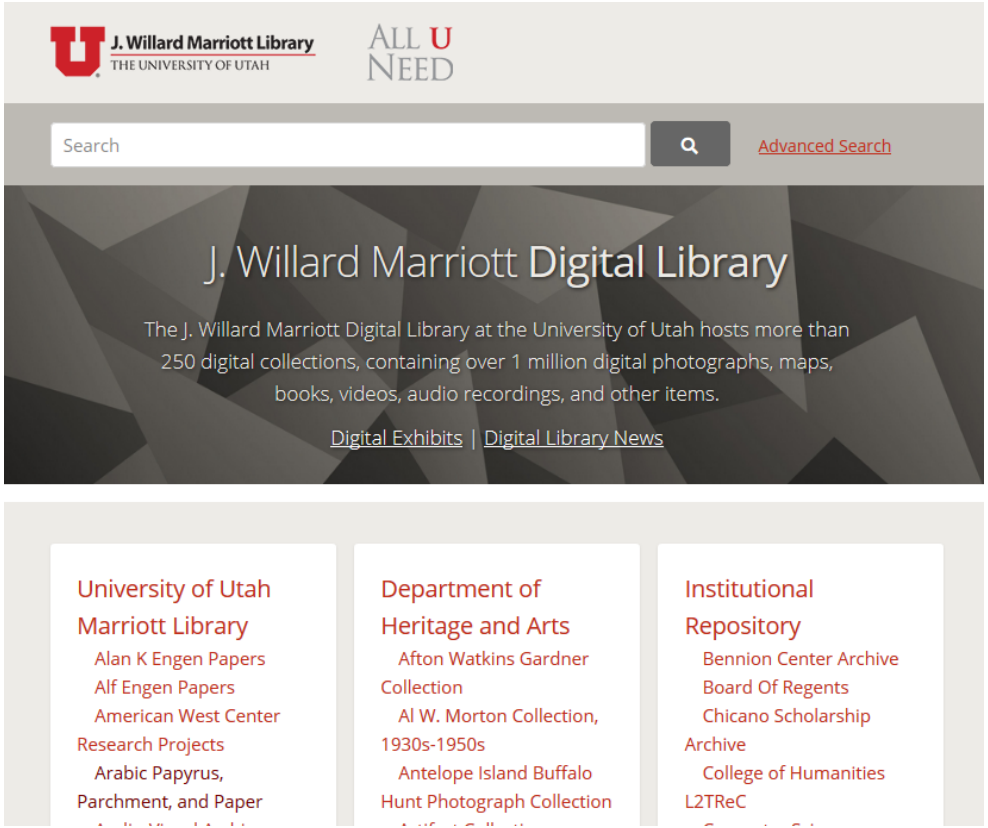
Bohan Zhu
Software Developer

Library Collections

Started in 2001

321 collections

470k+ photographs



https://collections.lib.utah.edu/

Goals

Expedite metadata creation

Enhance metadata and discovery experience for users

Address backlog issues in processing collections

Our approach - Machine Learning / Image Analysis

From the MIT Press Essential Knowledge book Machine Learning:

> Machine learning is a kind of AI, where the computer is programmed to optimize a performance criteria using examples or past experience. The machine does ***what the data tell it to***, not what a program tells it to.

https://chrisbourg.wordpress.com/2017/03/16/what-happens-to-libraries-and-librarians-when-machines-can-read-all-the-books/

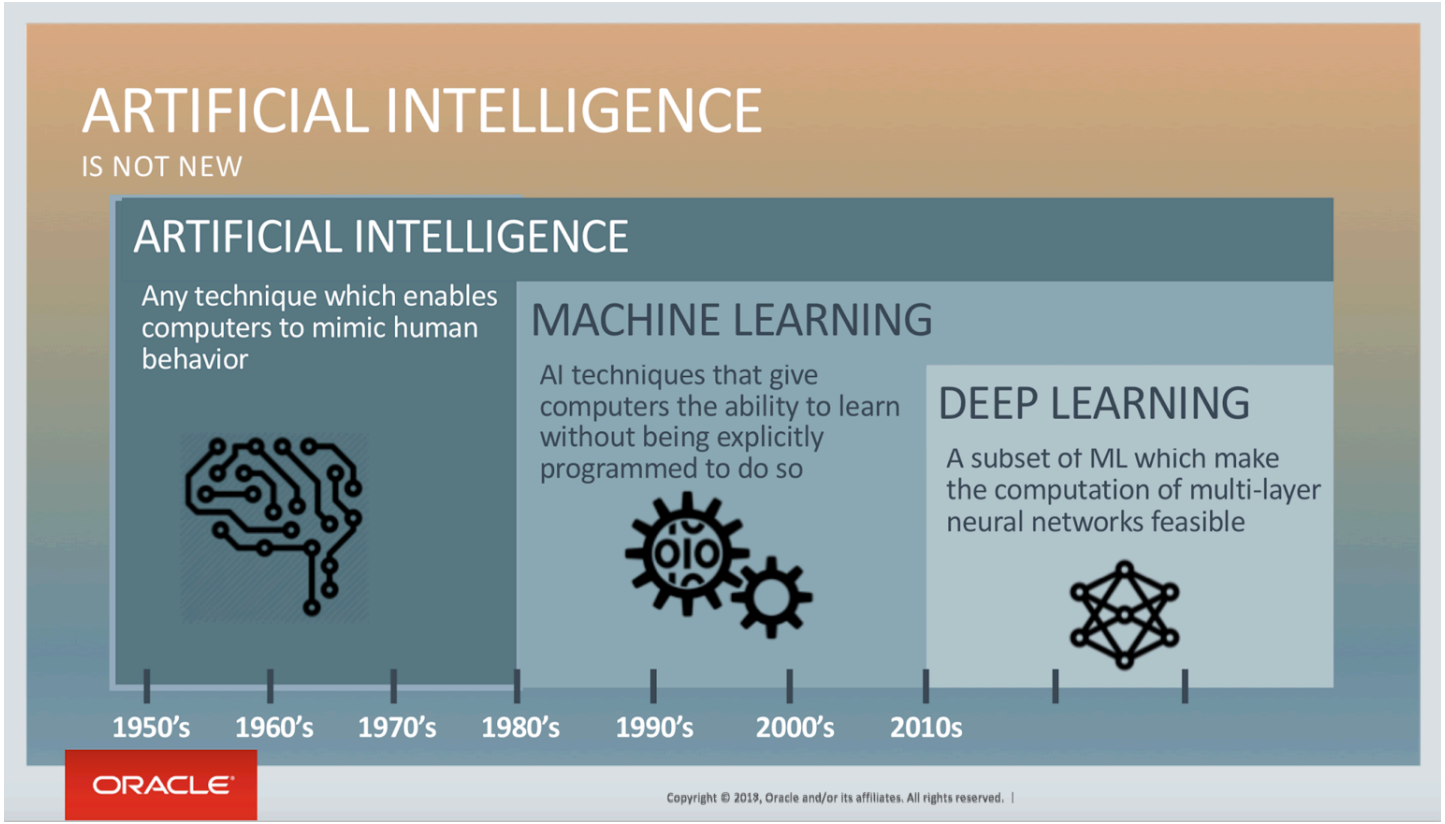
Other AI definitions

From Comparison of National Strategies to promote Artificial Intelligence:

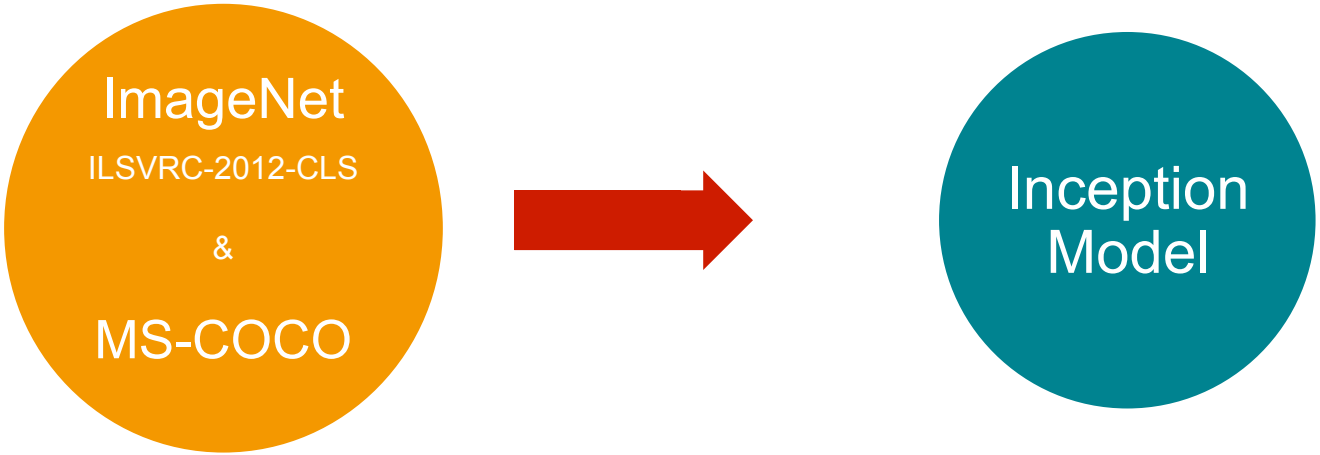
"what is referred to as AI changes with each major technological break-through, and the definition must therefore be periodically adjusted."

The only thing they have in common is an understanding of AI as a driving force in the digital revolution, which involves both potential and risk in terms of social, economic and, to some extent, security policy.

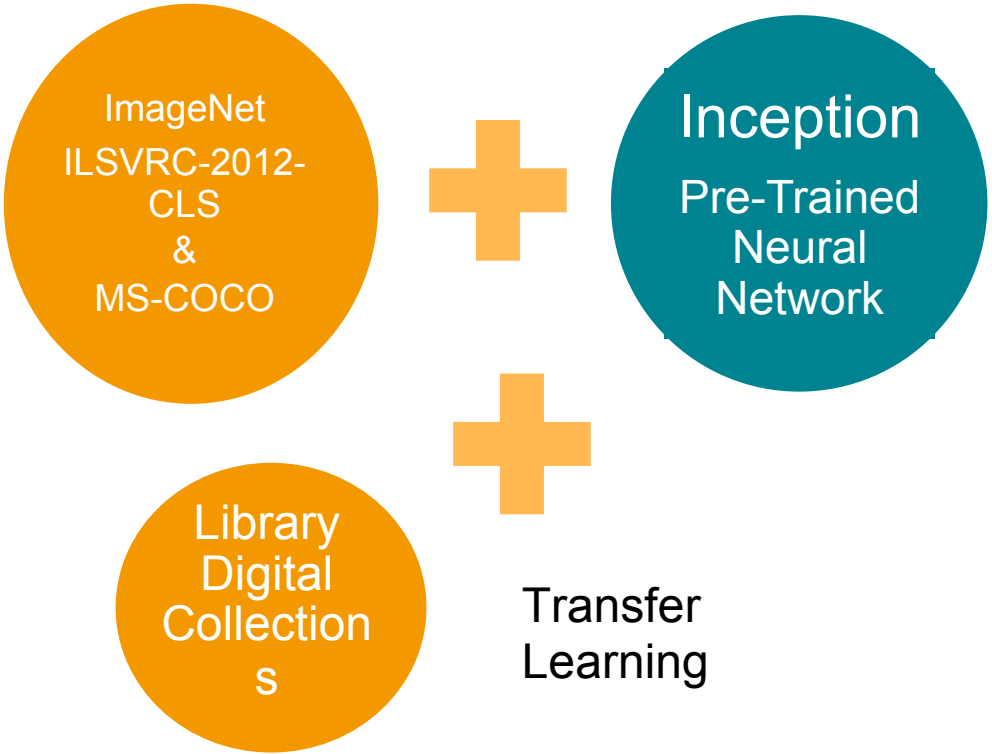
https://www.kas.de/documents/252038/4521287/Comparison+of+National+Strategies+to+Promote+Artificial+Intelligence+Part+1.pdf/397fb700-0c6f-88b6-46be-2d50d7942b83?version=1.1&t=1560500570070

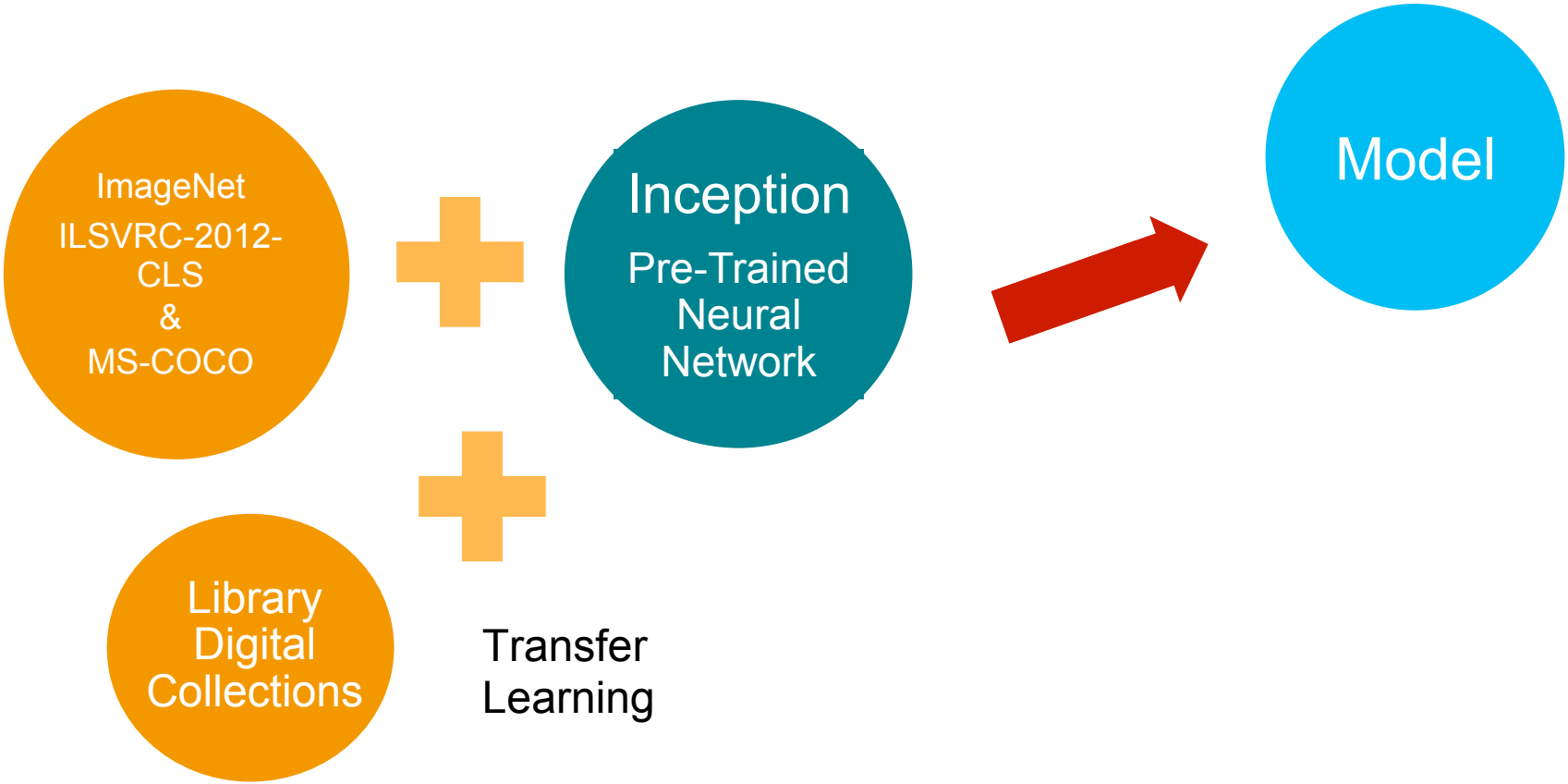


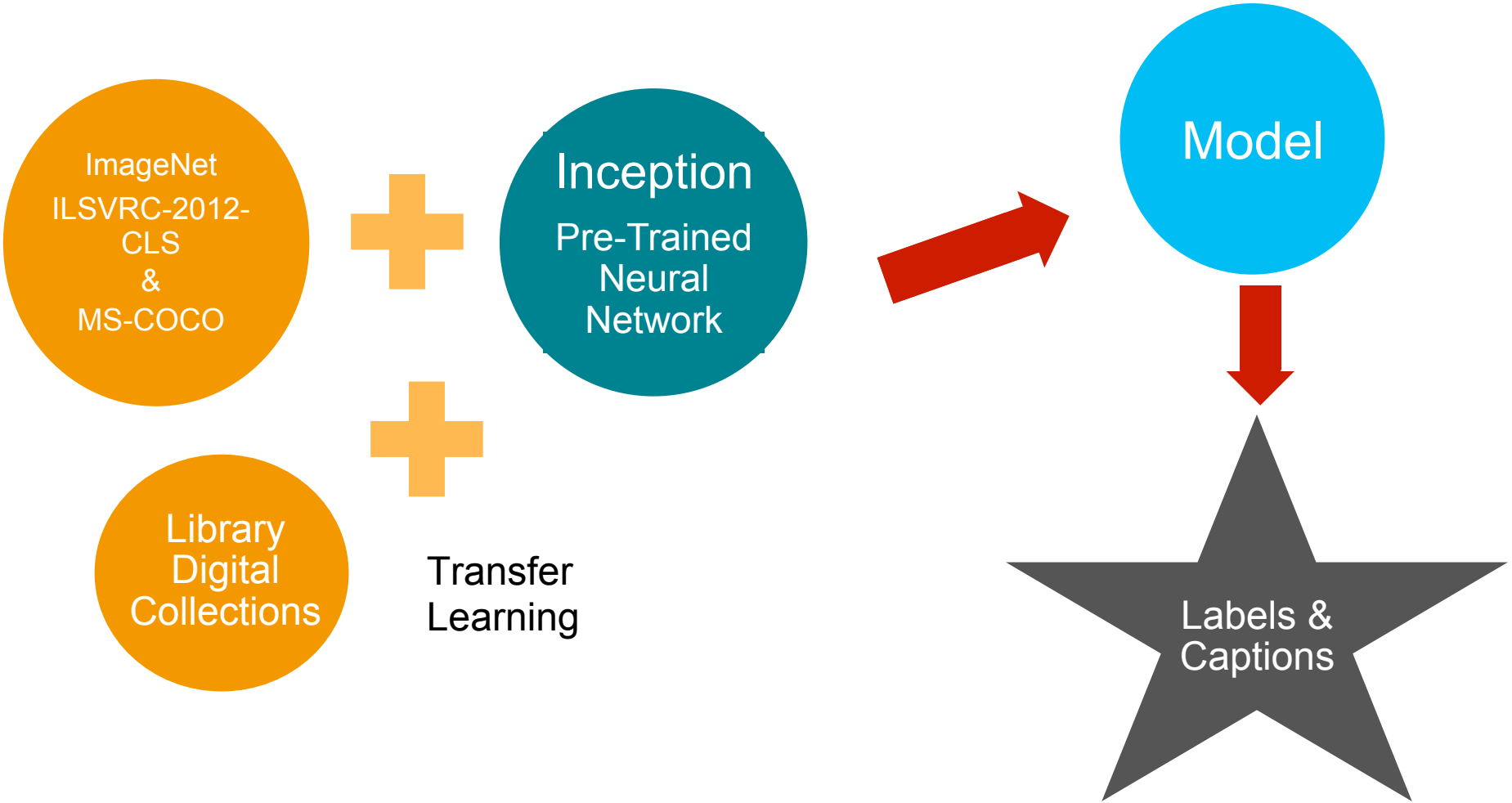
https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning


ImageNet
ILSVRC-2012-CLS
&
MS-COCO
Inception
Model


Inception
Pre-Trained
Neural
Network
Library
Digital
Collections
Labels &
Captions


ImageNet
ILSVRC-2012-CLS
&
MS-COCO
Inception
Pre-Trained Neural Network
Library Digital Collections
Transfer Learning


ImageNet
ILSVRC-2012-CLS
&
MS-COCO
Inception
Pre-Trained Neural Network
Model
Library Digital Collections
Transfer Learning


ImageNet
ILSVRC-2012-CLS
&
MS-COCO
Inception
Pre-Trained
Neural
Network
Model
Library
Digital
Collections
Transfer
Learning
Labels &
Captions

Cambell's Ferry

https://collections.lib.utah.edu/ark:/87278/s6319vk5

suspension bridge
pier
worm fence
snake fence
snake-rail fence
Virginia fence
viaduct
steel arch bridge

Malecon, Cuba

- a couple of people that are on a surfboard
- a couple of people that are sitting on a surfboard

Student group portrait

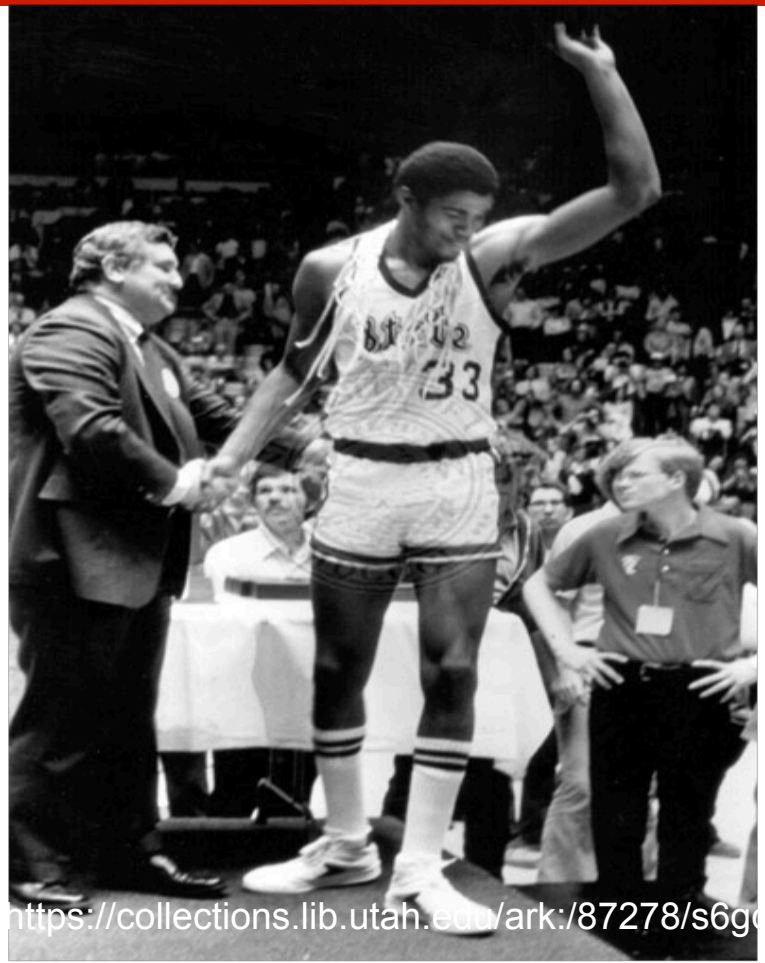
- a black and white photo of a group of people

- Vestment
- Hoopskirt
- Crinoline
- Suit
- Clothing

Unidentified gymnast

- and white photo shows gymnast a gymnastics meet performing a routine the balance beam

Magic Johnson after his Michigan State team beat Larry Bird's Indiana State team for the NCAA Championship in 1979. The final four were held at the Special Events Center.



https://collections.lib.utah.edu/ark:/87278/s6g

- a black and white photo of a group of people playing tennis
- black and white photo shows a group of gymnasts cheering
- basketball

Challenges

1. Domain Adaption

A model that is trained to optimize predictive accuracy on one domain (e.g., general web images captured in cell phone cameras) may not be well-suited for another domain (e.g., archival scans of black and white photographs).

President Wheatlake

- a man and a woman sitting in front of a laptop computer. (p=0.000110)
- a man and a woman sitting at a table with a laptop. (p=0.000096)

https://collections.lib.utah.edu/ark:/87278/s6z06f6j

Challenges

1. Domain Adaption

A model that is trained to optimize predictive accuracy on one domain (e.g., general web images captured in cell phone cameras) may not be well-suited for another domain (e.g., archival scans of black and white photographs).

2. Diverse captioning

Generating captions and labels from multiple perspectives to improve discoverability of content. Augment human metadata generation process with machine learning output.

Telephone Lines

ML generated Caption: Black and white photo of horse drawn carriage;

https://collections.lib.utah.edu/ark:/87278/s66b607z

Challenges

1. Domain Adaption

A model that is trained to optimize predictive accuracy on one domain (e.g., general web images captured in cell phone cameras) may not be well-suited for another domain (e.g., archival scans of black and white photographs).

2. Diverse captioning

Generating captions and labels from multiple perspectives to improve discoverability of content. Augment human metadata generation process with machine learning output.

3. Socially and Culturally responsible

Kabuki Theater

https://collections.lib.utah.edu/ark:/87278/s6bk1vgg

- a room filled with lots of different types of luggage
- a group of people standing around a room with a lot of luggage

Role of Libraries

Partner in generating use-inspired cases for research[1]

Partner in generating datasets[2]

"…datasets themselves may be of limited use in an AI context without an investment in labeling and curation."

Have a seat at the table to engage in responsible Operations[1]

1. https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html
2. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

AI4LAM is an international, participatory community focused on advancing the use of artificial intelligence in, for and by libraries, archives and museums.

Conferences

1st International Conference on AI for Libraries, Museums, and Archives

December 5, 2018 at National Library of Norway

2nd International Conference on AI for Libraries, Museums, and Archives

December 4-6, 2019 at Stanford University

http://ai4lam.org

Thank You!

ALL U NEED

J. Willard Marriott Library
THE UNIVERSITY OF UTAH

INVESTIGATING MACHINE LEARNING AND LIBRARIES AT THE UNIVERSITY OF NEBRASKA-LINCOLN . . . AND BEYOND

Elizabeth Lorang
University Libraries
University of Nebraska-Lincoln

liz.lorang@unl.edu | projectaida.org | elizabethlorang.com

Project Team

University of Nebraska-Lincoln
Elizabeth Lorang, co-director
Leen-Kiat Soh, co-director
Yi Liu, research assistant
Chulwoo (Mike) Pack, research assistant

University of Virginia
John O'Brien
Sarah Berkowitz
Worthy Martin

Image Analysis for Archival Discovery (Aida)

The Aida research team explores applications of image analysis and machine learning in digital libraries of historic materials. We're especially interested in what we might learn from the millions of digital images that librarians, archivists, and others are creating as they digitize the cultural record. We're intrigued by the questions that machine learning approaches might help to surface in these collections and about our professional practices—and also by the questions our collections and professional practices might help to surface about machine learning.

Our current and recent efforts include "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project" (Library of Congress), "Extending Image Analysis for Archival Discovery" (IMLS, LG-71-16-0152-16), and "Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914" (subaward on IMLS, LG-00-17-0104-17).

Code & Data

Code developed for our project is available via our GitHub organization page.

Data generated for our project are made available through appropriate

Application of the Image Analysis for Archival Discovery Team's First-Generation Methods and Software to the Burney Collection of British Newspapers[1]

Elizabeth Lorang, Yi Liu, Chulwoo Pack, Leen-Kiat Soh, Delaram Rahimighazikalayeh, and John O'Brien

May 2019

1. Background

The purpose of the study presented and analyzed here is to explore the generalizability of the Image Analysis for Archival Discovery (Aida) team's approaches across newspaper corpora. Up until this study, we have focused our training and testing data on U.S. newspapers of the nineteenth century. The current study, "Application of the Image Analysis for Archival Discovery Team's First-Generation Methods and Software to the Burney Collection of British Newspapers," is the first test of our approaches—methods and software—to a different newspaper corpus, specifically the *17th and 18 Century Burney Newspapers Collection*. This study stands as the first complete attempt at applying Aida's software and methods to non-*Chronicling America* newspapers, as a step toward understanding the potential of our approaches across digitized historic newspapers. In taking this step, our goals were (1) to test how well the software and a classifier model developed on *Chronicling America* newspapers performed on newspapers from a different corpus, a corpus that represents both a different geographical region and time period as well as newspapers digitized at an early stage in newspaper digitization history; (2) to explore whether classification results would be improved by training a new classifier model on Burney Collection images. Overall, we sought to explore how robust and extensible the first-generation Aida approach is and to better understand which parts of our methods might be brought over to new corpora "as is," and which may need to be calibrated for specific contexts.

Summer of Machine Learning Collaboration with the University of Nebraska-Lincoln

September 16, 2019 by Meghan Ferriter

This is a guest post by Eileen Jakeway, an Innovation Specialist on the LC Labs team.

Below, Eileen is in conversation with Dr. Elizabeth Lorang, Dr. Leen-Kiat Soh, and doctoral candidates Mike Pack and Yi Liu. They are members of a research team from the University of Nebraska-Lincoln collaborating with the Library of Congress on applying machine learning algorithms to Library collections for processing, metadata generation, and enhancing discoverability.

What, in your opinion, was the most promising outcome of the five machine learning projects you worked on this summer?

Soh: To me, [it] was the explorative nature of the five projects informed by insights from analyzing the data and by hands-on practical concerns from the Library.

Pack: I was excited about the fact that a set of features extracted by a deep learning model could deal with several tasks, such as classification and segmentation. Also, the fact that transfer learning (i.e., knowledge transfer) reduces training time makes it worthwhile to delve further into what deep representation can do, such as clustering document images.

Liu: The most fruitful part to me was also the exploration of transfer learning. The transferred knowledge could help us train and find performance sweet spots much faster than training from scratch. The best example is the project for digitization type differentiation, for which the training process took only three iterations to reach 90% accuracy!

Lorang: The further we made it into the summer, the more excited I got about the potential for the experiments we were conducting to inform thinking about the collections within the Library of Congress, to potentially help shape internal

What would socially and culturally responsible machine learning application and development look like in cultural heritage digital libraries, and how might we achieve such a vision?

--Liz Lorang, Harish Maringanti, John Wang

Currently imagining a series of investigations that will lead to the development of

- sustained technical and social analyses;
- guidelines, frameworks, and knowledgebases;
- critical code and tool studies;
- collections assessments, code and datasets; and
- interviews exploring archivists' and digital librarians' knowledge of machine learning, including how they think about issues of bias.

liz.lorang@unl.edu | projectaida.org | elizabethlorang.com