Machine Learning + Libraries Summit Event Summary

LC Labs
Digital Strategy Directorate

Primary Event Organizers: Meghan Ferriter, Eileen Jakeway, Abigail Potter

Report Authors: Eileen Jakeway (Primary), Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, Kate Zwaard

February 13, 2020

LIBRARY
LIBRARY OF CONGRESS

Executive Summary

On Friday, September 20, 2019, the Library of Congress hosted the Machine Learning + Libraries Summit. This one-day conference convened 75 cultural heritage professionals (roughly 50 from outside the Library of Congress and 25 staff from within) to discuss the on-the-ground applications of machine learning technologies in libraries, museums, and universities. Hosting this conference was part of a larger effort to learn about machine learning and the role it could play in helping the Library of Congress reach its strategic goals such as enhancing discoverability of the Library's collections, building connections between users and the Library's digital holdings, and leveraging technology to serve creative communities and the general public. Further reports, project results, and analysis about the application of machine learning to Library of Congress collections are forthcoming.

This event summary includes more detailed information about the conference proceedings. It broadly summarizes recurring themes of discussion and compiles the outputs of the small group activities. We hope it serves as a point of entry into broader conversations around the challenges, opportunities, and actionable items concerning machine learning in cultural heritage.

The event was divided into three themes: 1) ongoing projects, 2) opportunities and challenges regarding partnerships & vendors, and 3) future applications. Each thematic strand included lightning talks,[1] a small group activity, and a whole group discussion. The final event agenda is included in Appendix A.

The goals of the conference were to

- survey the range of ongoing projects in the broader cultural heritage landscape;
- surface major possibilities and barriers for applying machine learning in a library setting;
- demonstrate the possibilities of machine learning for use at the Library of Congress to internal audiences.

Threads emerging from whole group discussion at the conference include

- ethics, transparency, and communication;
- access to resources;
- attracting interest in GLAM (galleries, libraries, archives, museums) datasets;
- building machine learning literacy;
- expanding machine learning user communities;
- operationalization;
- connecting machine learning and crowdsourcing;
- metrics for evaluation of vendors and projects; and
- copyright and implications for the use of content.

[1] Appendix C includes an abstract and representative slide for each of the 25 lightning talks presented.

The small group exercises were organized around the following themes:

- **Defining success in (machine learning) projects:** The materials gathered from this exercise suggested that, at first blush, results and outcomes were identified as keys to "success;" however, further discussion surfaced the ways subject expertise is essential to practically implementing machine learning
- **Takeaways for collaboration on (machine learning) projects:** Observations loosely fell under the broad themes of: 1) project management, 2) expectations management, 3) data, 4) resources, and 5) team composition.
- **Milestones for machine learning projects in the next 6 months, 1-2 years, and 3-5 years**: The most frequent asks were around developing educational programming for MLIS students and building technical literacies; the need for funding; and desired documentation to be created about best practices, use cases, and ethical considerations.

Table of Contents

INTRODUCTION

Machine learning is broadly defined as training computers to detect patterns across large datasets. It can be used, with varying degrees of error, to carry out library-related tasks including but not limited to identifying and extracting visual content, tagging images, identifying content type, and enhancing metadata. LC Labs hosted the Machine Learning + Libraries Summit to gain an understanding of how and why machine learning is applied in the cultural heritage field. One desired outcome from this gathering was to begin the process of identifying machine learning approaches with potential application at the Library of Congress.

The conference presented an opportunity to supplement a close collaboration between LC Labs and a research team[2] with a broader survey of insights from practitioners in other spaces such as museums, libraries, and creative computing. Areas of expertise represented at the Summit included metadata generation; generative art, music and text; semantic annotation of AV content; object classification; and enhancing search. The event included a balance of small group, interactive, output-focused exercises, short "lightning talk" presentations, and whole group discussion to both facilitate theoretical dialogue and solicit practical, tactical outcomes for moving forward in the near term.

LC Labs invited broad representation from the Library of Congress—not only to spread awareness of the use of machine learning technologies but also to demonstrate the connections with ongoing work and needs at the Library of Congress. This work fulfills a Digital Strategy goal to "expand applied research to help the Library understand how to leverage emerging technologies to help connect our users with our resources and content."

[2] In tandem with the organization of the conference, LC Labs began working with an outside research team to test the practical dependencies of a machine learning project using Library collections and data. LC Labs contracted with the Project AIDA team at the University of Nebraska-Lincoln for a collaborative research project. Researchers carried out the following tasks via machine learning:

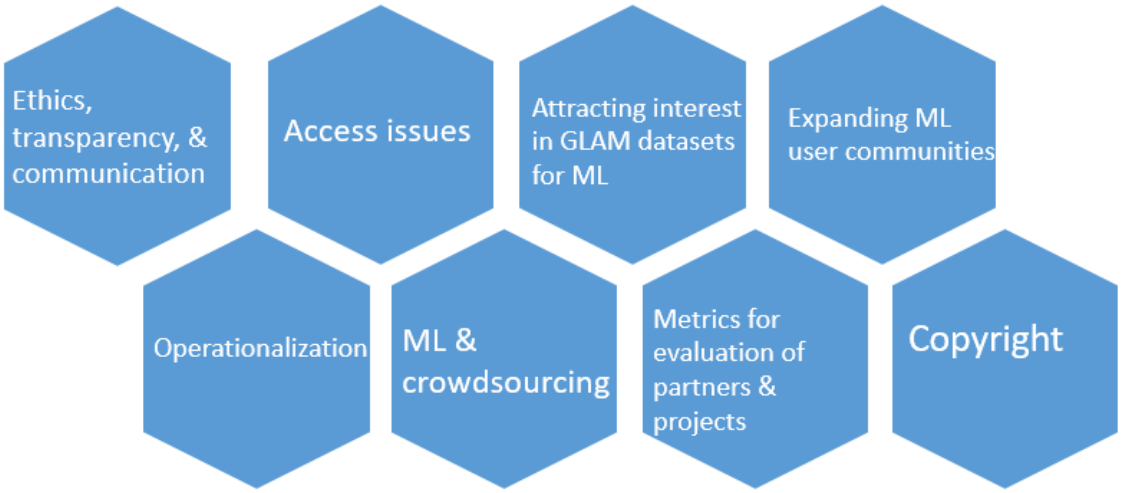
- Document Segmentation - segmenting image and text material from newspaper content
- Figure/Graph Extraction - find figures in newspaper content and extract text from figures
- Document Type Classification - classify handwritten vs. typed material
- Quality Assessment - analyze image quality of digitized MSS material
- Digitization type differentiation - recognize image digitized from microfilm

The final deliverables from these experiments—including curated datasets, project documentation, and white paper—will be available in 2020.

Whole-Group Discussion

Emergent Trends

Each session of lightning talks and interactive activities were followed by a whole-group discussion to debrief topics raised. The themes listed below reoccurred throughout the whole group discussions for each strand of the conference (projects—partnerships—future applications) based on the compiled notes from scribes present for the whole event.



Overview of major themes from discussion; organization of graphic mirrors order of themes discussed below.

Ethics, transparency and communication

A major thread running through the Machine Learning + Libraries Summit was that there is much more "human" in machine learning than the name conveys. This human involvement also brings with it the human subjectivities, biases, and distortions built into our information landscape. Machine learning, a technology that relies heavily on human judgment, taxonomies of categorization, and training data derived from society, may be more impacted than most technologies by issues of bias, making conversations about ethics all the more central. Transparency and communication were put forward as first steps to mitigate issues of bias often built into training datasets.

Another ethical consideration is that of the human labor involved in commercial machine learning programs. This arose as a factor to consider when assessing how/why commercial entities are often able to out-pace GLAM organizations with regards to technical advances in machine learning.

Access issues

Questions of "access" to machine learning coalesce around 1) resources (financial, computing, human) and 2) expertise. Both of these sets of resources are dictated by various factors; for this group, the most important considerations came down to the size of the institution, the buy-in of leadership, willingness to either fund in-house computer scientists or outsource labor to contractors or outside organizations. A lot of the resources extend beyond running the machine learning algorithms; in fact, a lot of the

conversation centered on "getting collections machine learning ready." This work relies on the labor of curators, catalogers, digital collections specialists, and developers working on digitizing, cataloging, cleaning, and providing access to these datasets.

Scale is another issue that ties in closely with access to resources. Summit participants discussed how a larger institutional scale necessitates more computing power and thus more funding and, by extension, is only a possibility for a select group of institutions. Smaller institutions might not face the challenge of absorbing sheer numbers of entities, such as machine-generated tags to enhance metadata, but might also have a harder time getting buy-in for machine learning technologies in the first place and/or getting their data machine-learning-ready.

Several times "funding models" were raised as being crucial to the conversation. More explicitly, participants called for establishing a cost model for machine learning infrastructure in order to approach senior leadership with evidence-based cost estimates.

Attracting interest in using GLAM datasets for machine learning

Throughout the Summit, many participants commented on the complexities of attracting users to conduct projects with machine learning. Although "machine learning" is typically associated with flashy, innovative, transformative, and futuristic problem-solving, operationalizing this technology is in fact pain-staking and labor-intensive. The main conversational threads focused around a) ways to incentivize ML projects and b) strategies for reaching user communities in libraries, computer science, and scholarly circles. A particular focus was placed on the role of GLAM organizations in leading a more general democratization of machine learning by building expanded literacies in technical *and* non-specialized communities.

Expanding machine learning user communities

Building on the theme above, in addition to attracting practitioners who can apply machine learning technologies to GLAM institutions' data, it is equally as important to attract researchers who may not yet know that their research project might benefit from the use of machine learning. After discussion, the challenge of building out this audience—to include users such as historians, anthropologists, and literary scholars—crystallized around the necessity for these users to articulate a well-defined research question that can be approached computationally.

Operationalization

Another major thread at the Machine Learning + Libraries Summit was that of operationalization or putting machine learning to use in an everyday work setting. The challenges of operationalization centered on 1) scale and infrastructure 2) building cross-functional teams that can have a transformational impact and 3) institutional support, a topic that was echoed across all sessions.

Once again, scale is a major driver of challenges in operationalizing machine learning in large institutions. The sheer number of entities (labels, tags, metadata) generated as outputs of machine learning would require building infrastructures not only to handle all these inputs but also to display them compellingly to end users. For example, it would be equally as important to work with UX designers as cataloguers in order to reflect and represent which outputs are machine-learning-generated in our systems.

Secondly, there was discussion at the conference about various approaches to building the "right" team—what should be the balance of programmers to librarians to utility players? Should you grow the team around the project or build a versatile, agile team capable of convening the right people in the right place at the right time? While no definitive conclusions were reached, it was suggested that having cross-disciplinary teams that are in conversation with all the parts of the Library are necessary to move a project from a pilot or prototype to a fully operationalized practice. This also entails building models that are iterative and designed to respond to feedback, and establishing mechanisms for evaluation and assessment from people across the institution.

The final major factor to consider in the operationalization of machine learning is obtaining institutional buy-in and support from senior leaders. This draws on what was discussed above with regards to establishing evidence-based cost models, building ML literacy and more. Little more was discussed regarding *how* to obtain senior leaders' support. More work could be done to identify successful mechanisms in future conversations.

Machine learning and crowdsourcing

Over the course of the Summit, it was mentioned that crowdsourcing affords opportunities for machine learning. Crowdsourcing was often cited as a means of "prepping" data sets for computational and machine learning uses, as a way of creating the training data, and then as a mechanism for training machine learning models with labels and tags (creating human-segmented data). However, crowdsourcing was also a topic around which ethical considerations were foregrounded in conversation. In response to the question of volunteer fatigue, one of the attendees noted the importance of valuing volunteers' contributions and time through communication and transparency. She suggested that this would help avoid losing volunteers' interest and time.

Metrics for evaluation of partners and projects

The only concrete suggestion that came out of this theme was the expressed need to develop metrics for evaluation of a project before, during, and after it takes place. It was suggested to draw on the community of people in attendance at the Summit to create benchmarks for evaluation during the deployment of a project.

Copyright

Copyright appears to be a major source of concern for many of the participants in attendance working with large datasets. The only proposed solutions were looking for rights-cleared or public domain materials, using the DMCA safe harbor, and negotiating rights agreements during acquisition.

Questions for Consideration

In her lightning talk, Heather Yager, MIT Libraries, posited the following questions on behalf of all libraries operating in the age of artificial intelligence. It is worth calling them to attention again for consideration by readers.

- What is AI good at, right now? Where does it struggle?
- What is the role of data in AI/ML, and how can we procure, structure, document, and interpret data ethically for AI/ML use cases?

- What does the AI-enabled organization look like, in terms of skill sets, workforce, business processes, and services?
- How do libraries, as data stewards, work to debias datasets and promote an understanding of ethical application of AI among practitioners?
- How do we make good decisions about AI/ML tooling in our own tech environments, and how will we determine, strategically, what (and how) we build / select / use?

Shared Challenges Identified in Discussion

- Identifying datasets and projects that lend themselves to ML
- Identifying clear research questions that are rife for computational analysis
- Building cross-functional teams
- Receiving institutional support from senior leadership
- Lack of a clear roadmap for the use of machine learning in cultural heritage

GROUP EXERCISES

The following section contains an overview of three small-group exercises conducted at the conference. Each section provides an overview of the exercise, high-level data analysis of the artifacts generated by the exercise, and a visual representation that aims to summarize what various participants discussed in their groups.

The purpose of this lightweight analysis is not to rigorously interrogate the unstructured data collected through these exercises. Rather, it is to provide readers with insight into the thoughts and ideas circulating among attendees of the conference. We aim to provide a starting point for a broader conversation among an expanded audience of professionals who were not in the room for these exercises by collecting, recording and sharing these observations.

Session 1: Attributes of Successful Machine Learning Projects

The first hands-on activity of the day asked participants to record all characteristics that came to mind when thinking of "successful machine learning projects." The materials gathered from this exercise suggested that while results and outcomes were initially identified as keys to "success," further discussion surfaced the ways subject expertise is essential to practically implementing machine learning.

Attendees wrote the adjectives to describe "success" for machine learning projects on index cards and then placed them in the center of the table. Next, they paired up and picked 3-5 random cards from the pile. In those pairs, they sketched out how they would incorporate these traits into their project, institution, or context.

After transcribing the 712 words used to describe "successful machine learning projects", the primary report author used Voyant Tools to identify the most frequently appearing words in the corpus. They are: data (9) ; ml (9); results (7); collection (5); good (5). A frequency table of the top 20 terms made using Voyant Tools is included in Appendix B.

When nearly synonymous terms were combined under a single term, *project(s)* rose to one more mention than *collections* as seen below.

Data	9
ML	9
Results	7
Project(s)	6
Collections	5
Good	5

The second part of this activity documented how people would actualize these measures of success in their own organizational or professional context. The materials transcribed for this section contained 1,274 words; the top five commonly occurring terms are: data (20), ml (11), domain (10), collections (8), expertise (8). A full list of the top 20 terms made using Voyant Tools can also be found in Appendix B.

When like terms are folded under a single term, the numbers change significantly:

Data	20
Domain, expertise	18
ML, AI	16
Project(s)	15
Human(s)	9
Collections	8

When discussing implementation steps, domain-specific knowledge and expertise of staff was mentioned more often than ML and AI. Such cursory analysis may suggest that many of the event attendees brought focus to the importance of the content or subject matter being used and the process involved in preparing collections for machine learning use. However, there also remains a heavy focus on data—more specifically, ground truth and sharing datasets.

A side-by-side comparison reveals that there is significant overlap between the terms named as aspirational attributes and those presented as requirements for implementation. However, the points at which they diverge (results vs. domain/expertise) may point to a noteworthy difference between broad-strokes "success" and actual implementation; whereas the aspirational thinking exercise resulted in the discussion of concrete "results" and fleetingly nebulous quality of "good," the plans for actualizing these traits at home leaned heavily on the "expertise" of staff and the "human(s)" at the center of it all.

Exercise 1.1: Attributes of "successful machine learning projects"	
Data	9
ML	9
Results	7
Project(s)	6
Collections	5
Good	5

Exercise 1.2: Actualizing these attributes in your own context	
Data	20
Domain, expertise	18
ML, AI	16
Project(s)	15
Human(s)	9
Collections	8

Although "human(s)" did not make it into the top six traits for "success" picked for the exercise, it was the *second most commonly occurring* combined term in the set of cards written down as attributes of successful machine learning projects in the first part of the exercise but not chosen for the second exercise. See Appendix B for a more complete list. Had these cards been chosen, the resulting conversation may have turned out differently.

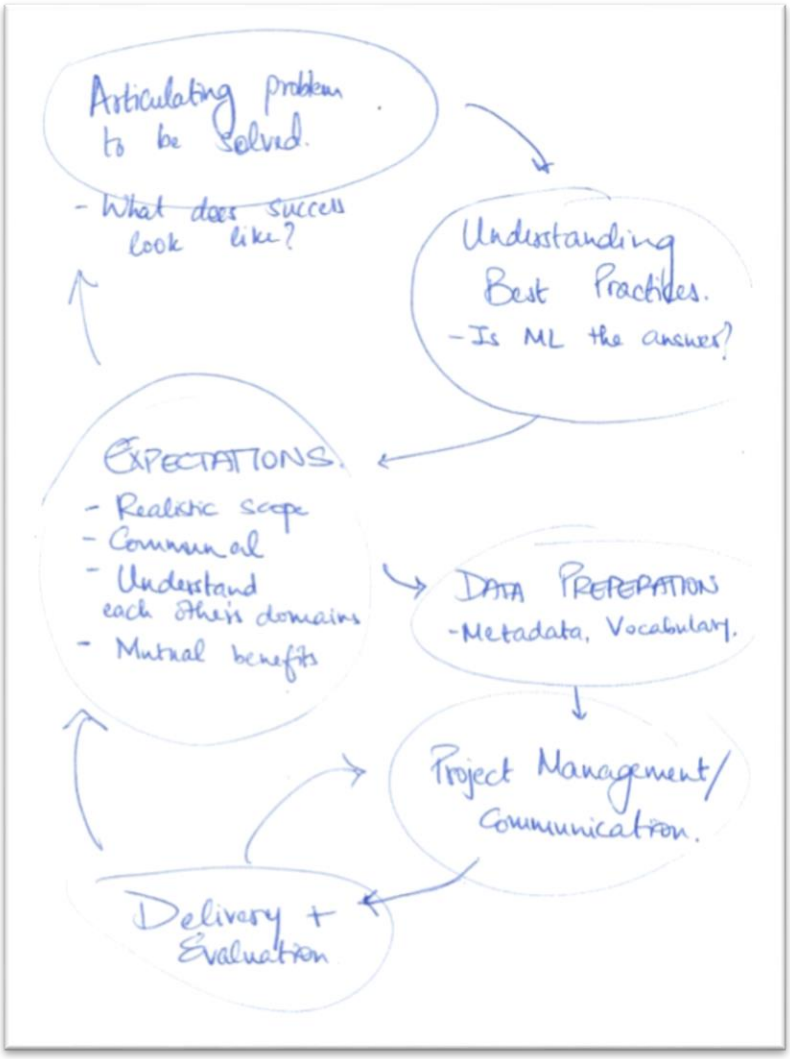
While there seems to be consensus around the importance of data, domain expertise, and human involvement in the machine learning process, defining success for machine learning projects remains challenging. This comes as no surprise as the measures of success defined by computer scientists, which often rely on metrics of "accuracy," are not always, or even often, in line with the considerations of librarians and cultural heritage professionals. These concerns coalesce largely around asking questions of the data themselves: is it representative? How is it biased? How can we make informed decisions about the results of machine learning models? How can we present those complexities to users?

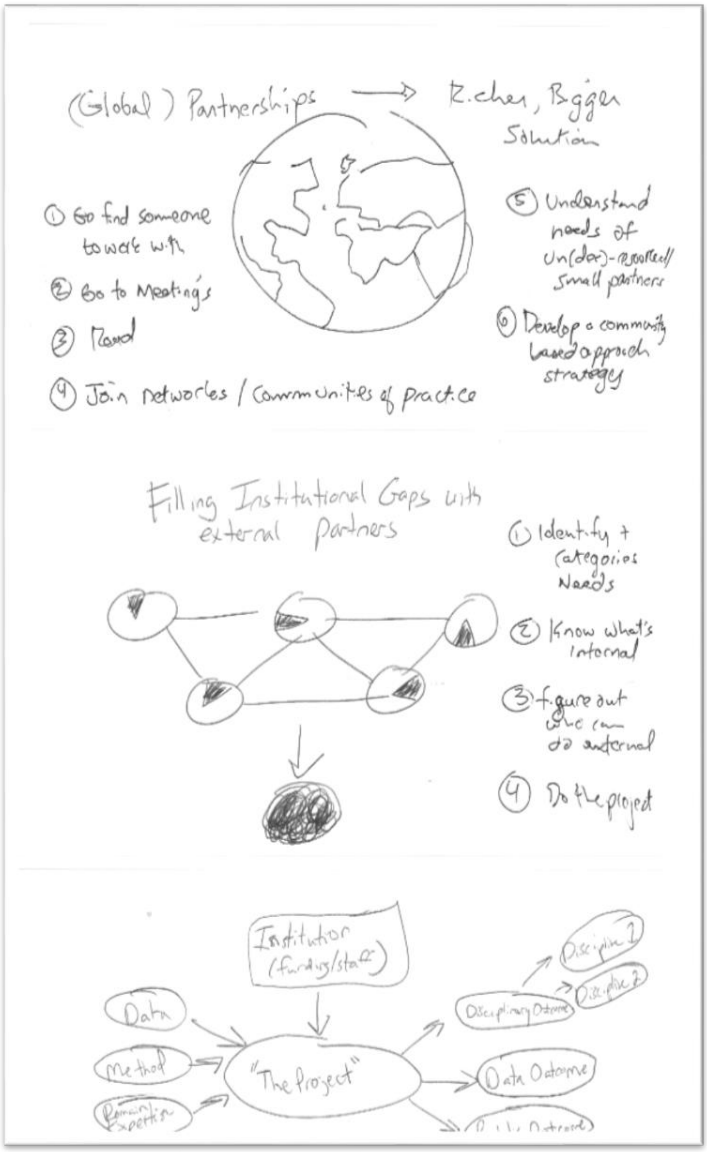
Session 2: Takeaways for Collaboration on (Machine Learning) Projects

The second activity focused on partnerships and collaborations on machine learning projects. The activity prompted participants to pair up and identify "big takeaways" they observed from entering into an agreement with another person or organization on a machine learning project. Participants were encouraged to think more generally about collaborative projects if they felt they did not know enough about machine learning-related collaborations specifically.

The following graphic represents a summary of themes that emerged from this exercise. Observations loosely fell under the broad themes of: 1) project management, 2) expectations management, 3) data, 4) resources, and 5) team composition. The more granular takeaways are included in the chart below.

Included below are select scanned artifacts made by groups to organize their takeaways into higher-level themes.





The artifacts above were produced by participants of the Machine Learning + Libraries Summit held at the Library of Congress on September 20, 2019.

The artifacts above were produced by participants of the Machine Learning + Libraries Summit held at the Library of Congress on September 20, 2019.

Session 3: Milestones for Machine Learning in 6 months, 2 years, 5 years

The final activity of the conference focused on identifying machine learning-related milestones for the represented institutions in the next 6 months, 2 years, and 5 years. Participants brainstormed ideas on their own and then, in small groups, discussed what actions need to be taken in order to reach said milestones.

The chart below encapsulates all of the proposed milestones and groups them into corresponding high-level themes of: education, funding, project management, communication, documentation to be created, project-specific goals, partnerships, broader community efforts, and advances in technology. The most commonly occurring asks were around developing educational programming for MLIS students and building technical literacies; the need for funding; and desired documentation to be created about best practices, use cases, and ethical considerations.

Milestones that appear in **bold** were raised multiple times by different participants for the same suggested timeframe. Milestones appearing multiple times but across different estimated timeframes are presented in the same color font below.

	6 months	1-2 years	3-5 years
Education	• **Formal training, education: data/AI literacy (for MLIS students)** • **Class to introduce students to Python programming, data management** • **Create tutorials + guidelines based on ML work using real world library data** • Sharing Jupyter notebooks	• Have curriculum, a sequence of courses, a set of teaching materials for MLIS students to learn machine learning • Rudimentary machine learning workshop	• Work with LIS programs around the world to develop, use and share teaching materials
Funding	• Seek funding • Get funding • Massive infusion of funds from funding agencies to smaller orgs • Seek out sponsors, build teams, clarify goals, and deliverables	• Free up resources/prioritize • Fundraising • Fund and complete an ML project	• ML business model with effective governance

	• Cost: operationalize/evaluate • More funding, more collaborations	
Project Management • Define problems to apply ML to, fitting priority problems • Review of tools/requirements and needs • Establish team • Pipeline from ML to human output with context • Defined + budgeted project proposal	• Evaluate • Establish and execute programme of activity; iterate if needed. • 2-5 funded research projects • 2-5 projects using services • Completing an ML project and ingesting it so it improves search	• **Testing** • **Test-improve-test-improve-communicate** • Production-ready, tested and vetted tooling that has been proven to work then partnerships • Human-computation results • Live, server-side interaction between user and AI model-- tweak results based on individual research question • Notable project results
Communication • Clear statement about LAM & AI values • Collection of ethical considerations for AI/ML applied to cultural heritage	• Guidelines for reducing bias in datasets • Implement standards on all released datasets and any that interact with the public	• Communicate results to users + community
Documentation to be created • Well-documented case studies • Examples of what's working/best practices • Literature review to what is unique to AI • Decide on standards for describing datasets (could be data nutrition labels)	• Draft of literature review created, case studies (based on original use cases) being developed • **Mechanism for sharing pre-trained ML models** • **Data repository of shared ML models**	• Study to determine effectiveness of standards, adjust as necessary • Co-created statement on research data use and reuse for Congress

Documentation, cont.	• **Index of all ML projects** • Guidelines for reducing bias in datasets • ML-ready datasets • 5-year plan • 2-year research study • ML strategy plan report • ML expertise: questions, use cases, priority 0-6 months, understanding ML funds • ML/AI referenced in strategies & annual report	
Project-specific	• Advances in DH cooperation toward African film history project • User selection features for pix plot • Full ML experiment with crowd.loc.gov • Label all wildlife photos • Complete ML pilot over crowd data	• High-accuracy models trained on "our" collection data that reflect critical and scholarly terminology • Pilot ML project w/crowd QA	• Every library with one or more ML'ed collection • ML + crowd integrated loc.gov platform for metadata creation, transcription of AV • ML supporting at least one item in our directional plan
Partnerships	• Identify partners (who can test in a production capacity) • Co-creation of research (partner with data creators) to develop ethical model/taxonomy of data	• Team-building (internal) • External: partnerships, MOU, recruiting	• Collaborations w/ FIAF and scholars in many countries re Africa film history project

Broader community	• Creation of a working group or community of practice w/lots of groups • Raise awareness of all ML issues; attend summit at Stanford to further discussions		• Office for AI assessment • Professional board for lobbying on ethics and IP law • Consortium to build best practices • Family historians can explain how ML affects the datasets they use
Advances in technology		• Advances in ML analysis/bodies in motion in motion pictures • Evaluation of available **tools to improve access to collections**--prototype? • **Tools to support access to collections** • Dev studio	• Infrastructure to support projects in Named Entity Recognition (NER), Optical Character Recognition (OCR), supervised ML • Production-ready tools • GLAM collections systems can easily ingest ML data to present it back in discovery systems • Usable open-source tools to use for AV

Near-Term Possibilities at the Library of Congress

Related Labs Activities

LC Labs is supporting projects to explore and evaluate potential machine-learning applications at the Library of Congress. The goal of these collaborations is to continue exploring the limitations and capabilities of machine learning technologies and to uncover the dependencies necessary for the Library of Congress and its users to benefit from it. These projects range from experimentation around visual extraction to data gathering about the state of the field more generally to testing various ML models on Library collections. Hopefully, this experience and synthesis will present other parts of the Library with sufficient evidence and questions for consideration when investigating machine learning projects of their own.

2019 Innovator in Residence Program

Ben Lee's 2019 Innovator in Residence project will apply machine learning to extract visual content from the Library's digital collections. His goal is to make these images available to users in an interactive visualization such as on a timeline or a map or searching by topic.

Brian Foo, another Innovator in Residence, has developed an algorithm capable of identifying segments in audiovisual content that appear to contain music. This automatic segmentation and labeling of videos has applications in enriching metadata, such as noting when an interviewee begins playing a song, and increasing access to these materials for users who are interested in knowing precisely where to look for certain song segments.

Machine Learning State-of-the-Field Report Forthcoming

Dr. Ryan Cordell, associate professor at Northeastern University, attended the Machine Learning + Libraries Summit and has been contracted to write a comprehensive report detailing the state of the field of machine learning in cultural heritage. His report is scheduled to be published in March of 2020.

University of Nebraska-Lincoln Final Deliverables

The Project AIDA team will be releasing a prototype, final report, code, and documentation in spring 2020 for members of the public and Library staff.

Other possible applications at the Library of Congress

Preservation

Machine learning may be used to assist with assessing collection management and preservation challenges. An initial project would investigate the use of image segmentation to automate the extraction of call numbers from photos of the stacks and, from that data, create a "heat map" visualizing the stacks most in need of intervention.

Rare Materials

Preliminary findings from the UNL project suggest that trained machine learning models are able to extract visual content from handwritten manuscript materials as well as typed documents.

External Engagement

AI for LAM Slack

This communication channel is a way of joining an informal community of practice around the use of AI in libraries, archives, and museums.

Ethics in ML & GLAM group

Spurred by their conversations at the Summit, several attendees met up in New York City to begin drafting a statement of values around the application of machine learning to collections.

CONCLUSION

While a number of cultural heritage organizations have had promising results with early machine learning projects, the application of this technology to library collections remains experimental. As the number of applications continue to grow and technology continues to improve, it will require collections readiness, building institutional capacity, cross-functional collaboration, and attention to ethics. In the short term, this may mean incorporating workflows to make digital collections ready for computational and machine learning use.[3] In the long term, it may mean expanding staffing structures to build capacity or collaborating more transparently with partners or vendors to document where datasets come from and how they are structured.

LC Labs, which falls under the auspices of the Digital Strategy Directorate in the Office of the Chief Information Officer, will continue to identify collections that may be suitable for machine learning use, collect evidence about the results of our ongoing ML-related projects, and share those results and recommendations. It is essential to the work of this team to involve stakeholders from across the Library of Congress in these exploratory conversations and collaborations as investigation of these technologies and approaches continues.

In order to complement this work, we look to our peers and colleagues to 1) share the outcomes of their projects, 2) communicate their specific needs and requirements for collaboration, and 3) envision a range of tasks they would like to see performed that may benefit from computer assistance.

Questions, comments, or responses stemming from the Machine Learning + Libraries Summit or this report can be sent via email to LC-Labs@loc.gov .

[3] Recommendations for supporting the computational readiness of digital collections has already been outlined in the 2017 Digital Scholarship Working Group Report.

Appendix A

Machine Learning + Libraries Summit Agenda & Participants

MACHINE LEARNING + LIBRARIES SUMMIT

Library of Congress, Madison Building, Washington, D.C.
Montpelier Room
September 20, 2019

8:00 – 8:30	Registration and coffee, meet and greet
8:30 – 8:45	Welcome, Overview of Meeting Goals
9:00 – 11:45	**Strand 1: Recent & Ongoing Projects** *Lightning Talks • Table Discussions • Whole Group Debrief* Topics may include: AV materials, scalability, communicating results to users, discoverability, ethics & ML, inherent biases.
11:55 – 12:55	Lunch provided
1:00 – 3:45	**Strand 2: Partnerships** *Lightning Talks · Table Discussions · Whole Group Debrief* Topics may include: interoperability, commercial solutions, professional ethics, frameworks for partnership.
3:45 – 4:00	Networking/Coffee Break
4:05 – 5:15	**Strand 3: Horizon** *Lightning Talks · Closing Discussion* Topics may include: ethics questions, future technologies, funding, action plans.
5:15 – 5:30	Closing Remarks, What We Learned & Next Steps

75 participants attended the Machine Learning + Libraries Summit. Collectively, this group represented the following institutions:

- American Museum of Natural History
- AVP
- British Library
- Carnegie Museum of Art
    - Data Science Lab
- Digital Public Library of America
    - Digitization Program Office
- Fitzwilliam Museum, University of Cambridge
- Georgia Tech University
- Goodly Labs
- HathiTrust/ Research Center
- Hoover Institution Library & Archives, Stanford University
- IIIF Consortium
- Independent Research Artist(s)
- Indiana University Bloomington
- Library of Congress
- Media Ecology Project, Dartmouth College
- MIT Libraries
- National Archives & Records Administration
- National Endowment for the Humanities
- Northeastern University
- OCLC Research
- Old Dominion University
- Rutgers University
- Smithsonian Institution
- Stanford Libraries
- UC Berkeley
- United States Holocaust Memorial Museum
- University of Nebraska-Lincoln
- University of Notre Dame
- University of Pittsburgh
- University of Texas Libraries
- University of Utah
- University of Washington Computer Science and Engineering
- Virginia Tech University
- WGBH Media Library & Archives
- Yale Digital Humanities Lab
- Zooniverse

Appendix B

Frequency Chart: Successful machine learning projects (created using Voyant Tools)

	Term	Count
1	data	9
2	ml	9
3	results	7
4	collection	5
5	good	5
6	time	5
7	clean	4
8	saves	4
9	scale	4
10	classification	3
11	human	3
12	idea	3
13	just	3
14	metadata	3
15	new	3
16	outcomes	3
17	project	3
18	projects	3
19	success	3
20	tasks	3

Frequency Chart: Actualizing attributes of successful machine learning projects (created using Voyant Tools)

	Term	Count
1	data	20
2	ml	11
3	domain	10
4	collections	8
5	expertise	8
6	project	8
7	use	8
8	need	7
9	projects	7
10	scale	7
11	better	6
12	goal	6
13	results	6
14	useful	6
15	ai	5
16	human	5
17	library	5
18	new	5
19	set	5
20	time	5

Frequency Chart: Attributes of successful machine learning projects *not* chosen for actualization exercise

	Term	Count	Tr
1	data	14	
2	project	8	
3	results	8	
4	new	7	
5	human	6	
6	ml	6	
7	training	5	
8	easy	4	
9	good	4	
10	machine	4	
11	needs	4	
12	problem	4	
13	real	4	
14	time	4	
15	corpus	3	
16	end	3	
17	humans	3	
18	insight	3	
19	learning	3	
20	potential	3	

Appendix C

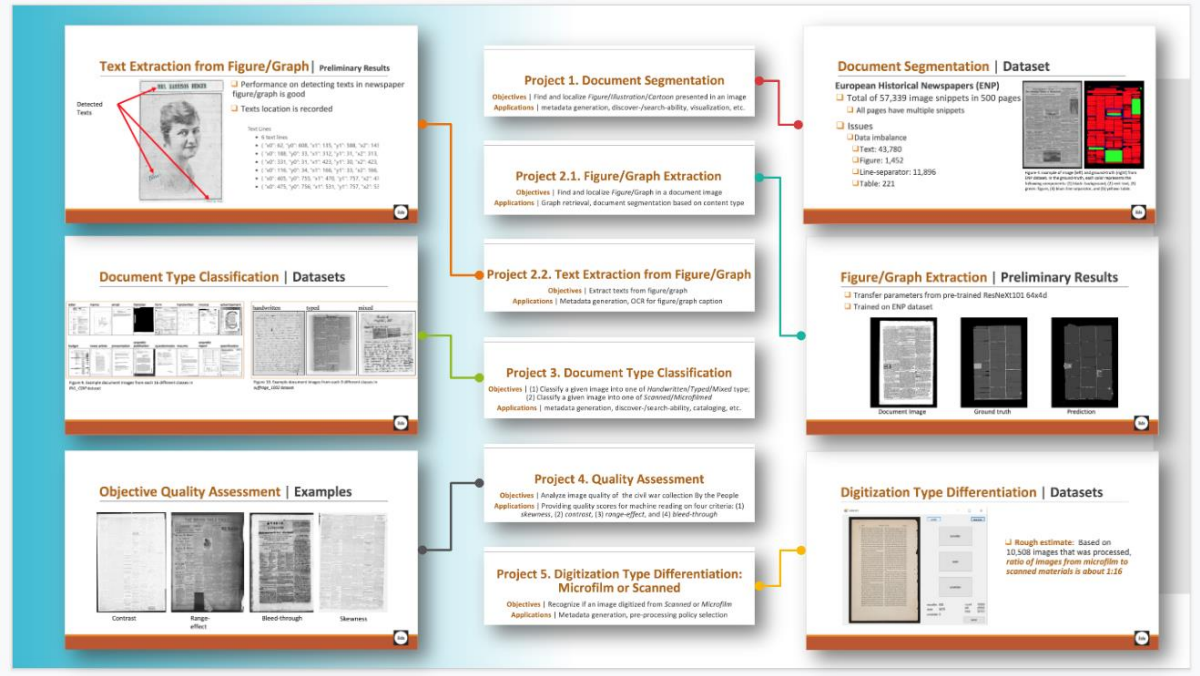
Lightning Talk Presentations

LC Labs selected 25 projects to give lightning talks representing diverse collections/content type, tasks/approaches, results, and users/audiences. A brief description of each talk is included below. Presentation slides are available here. They act as a non-comprehensive survey of work combining machine learning and cultural heritage across the United States.

Recent & Ongoing Projects

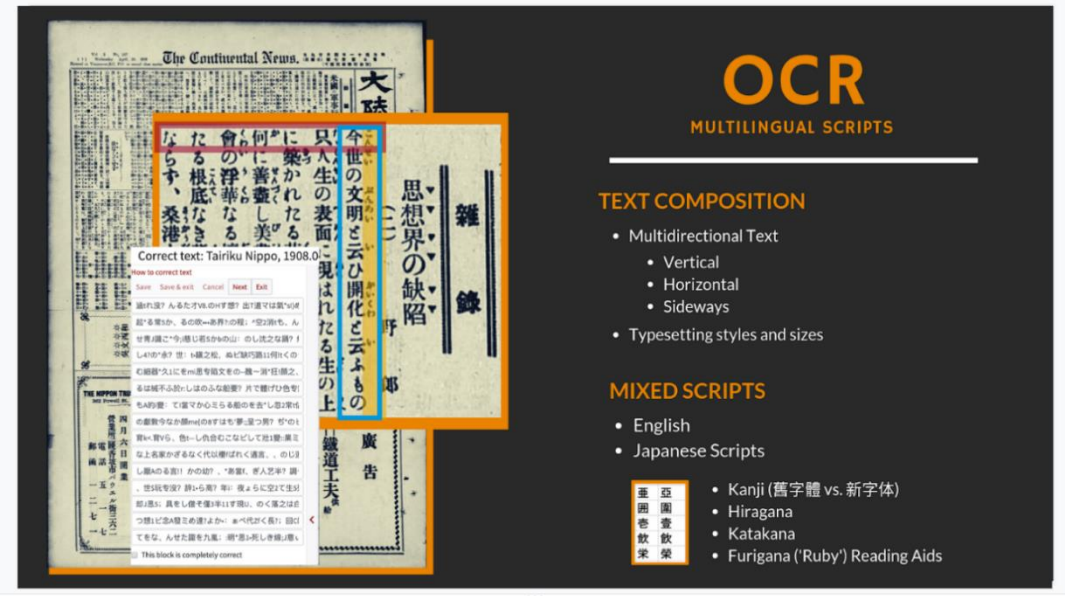
Leen-Kiat Soh, University of Nebraska-Lincoln: "Digital Libraries, Intelligent Data Analytics, and Augmented Description"

- 5-month applied research project with goals to develop and investigate the viability of textual and image-based data analytics approaches to support discovery, understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections.



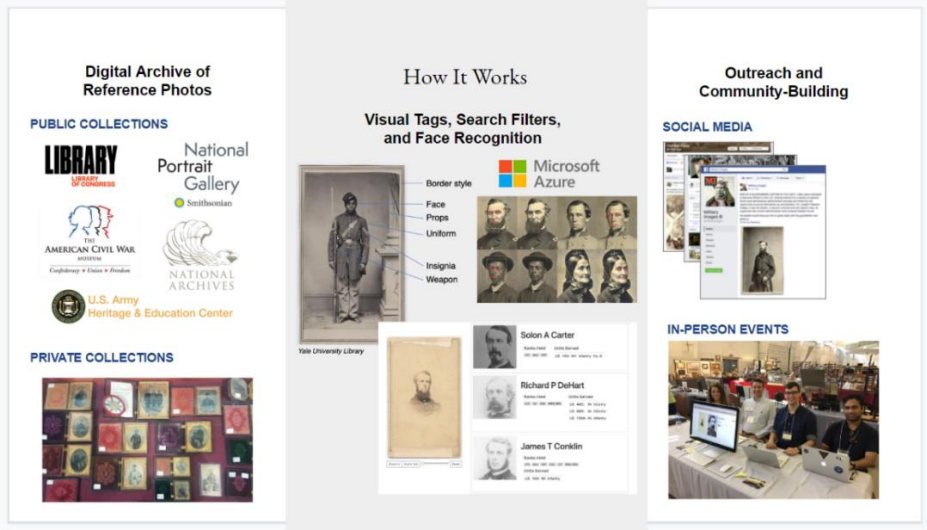
Thu-Phuong 'Lisa' Nguyen, Hoover Institution Library & Archives, Stanford University: "Hoji Shinbun Digital Collection: Newspapers of the Japanese Diaspora (1868-1945)"

- Focused on the successes and challenges of mass digitization through collaborative international partnerships and surfacing multilingual/multi-directional content through OCR as demonstrated through the Hoji Shinbun Digital Collection case study using page-level segmentation and article segmentation.



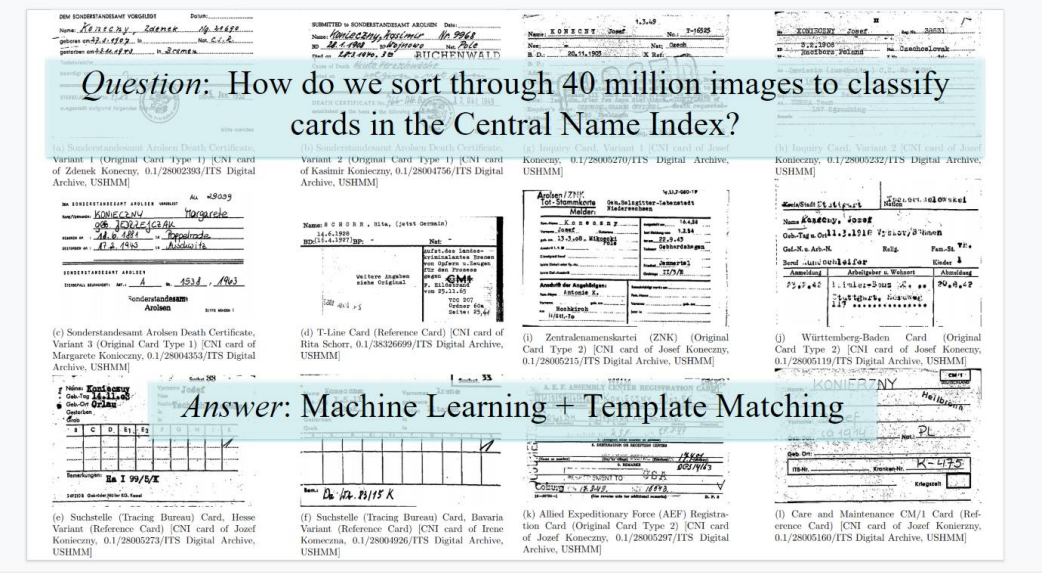
Kurt Luther, Virginia Tech University: "Combining Crowdsourcing and Face Recognition to Identify Historical Portraits"

- Civil War Photo Sleuth (www.civilwarphotosleuth.com) is a free public website that combines crowdsourcing and AI-based face recognition to identify unknown soldiers in American Civil War-era photos.



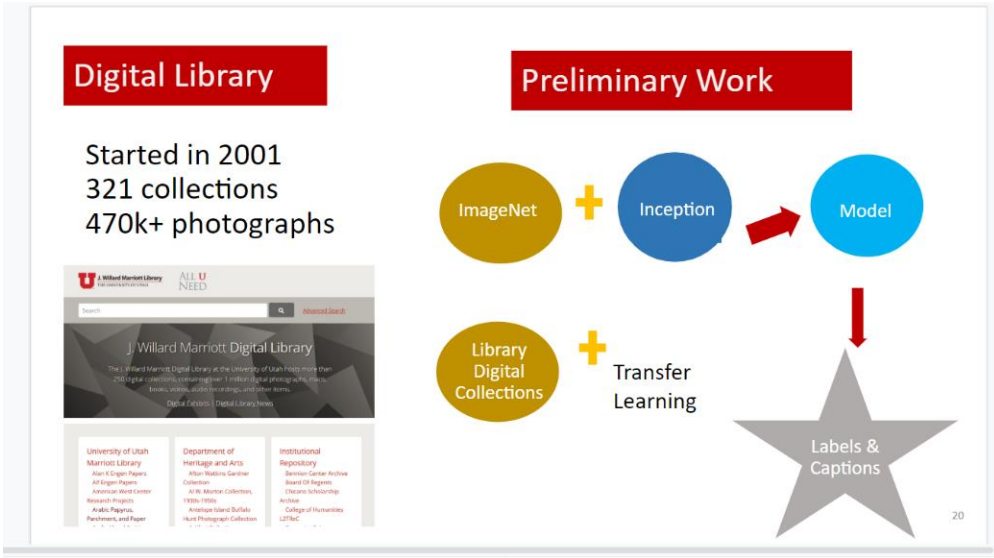
Benjamin Charles Germain Lee, Michael Haley Goldman, United States Holocaust Memorial Museum: "The International Tracing Service and Machine Learning"

- This project at the USHMM used machine learning to sort through 40 million images of cards contained in the Central Name Index of the International Tracing Service. Ben used template matching and machine learning to automate the retrieval of the cards making reference to death certificates.



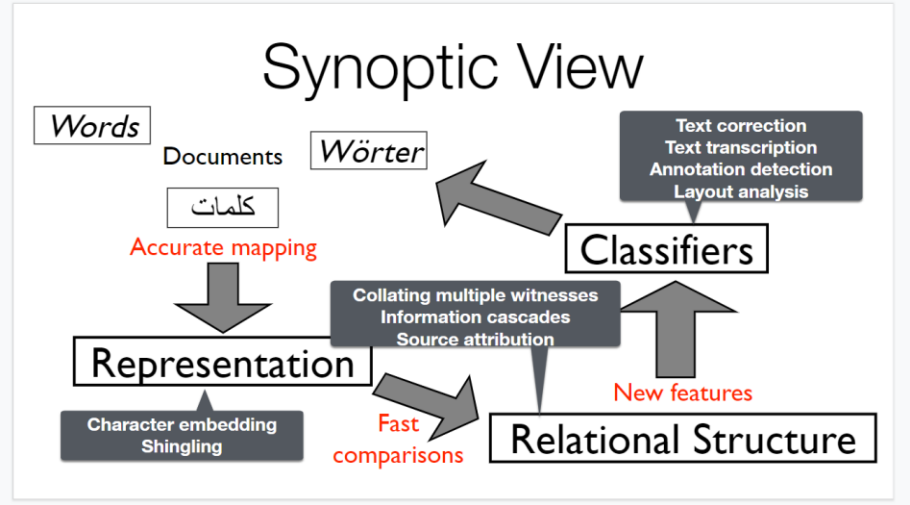
Harish Maringanti, University of Utah: "Sheeko: A computational helper to describe digital images"

- This project attempts to use machine learning to generate metadata for the library archives at the University of Utah. Project goals are to enhance the discovery experience for users, expedite metadata creation, and address backlog issues in processing collections.



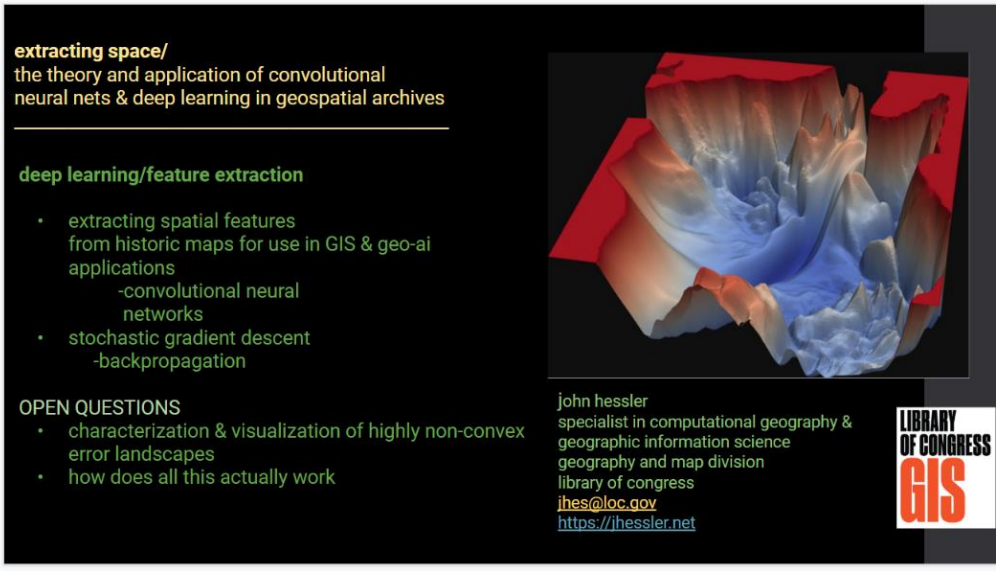
David Smith, Northeastern University: "Networked Texts: Improved Inference by Exploiting Relational Structure"

- The Networked Texts project analyzes the sources and structure of large heterogeneous collections of noisily digitized historical newspapers and books. These techniques not only suggest approaches to cataloguing these collections but also provide another source of training data for document layout and transcription models and the analysis of readers' annotations in books.



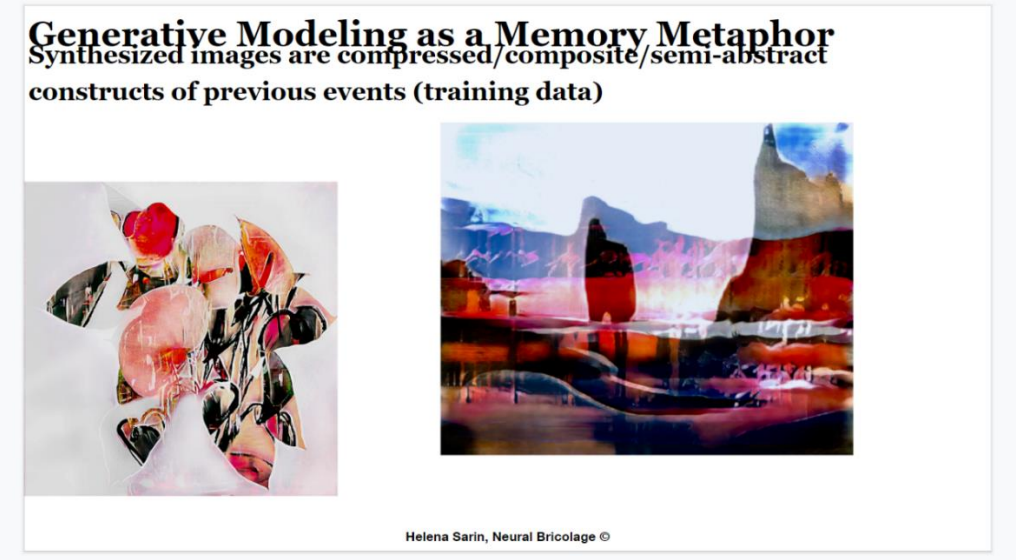
John Hessler, Library of Congress: "Extracting Space: the theory and application of convolutional neural nets and deep learning in geospatial archives"

- John Hessler's work uses deep learning to extract spatial features from historic maps in the Library's collections. He emphasized the importance of having a deeper understanding of how neural networks work.



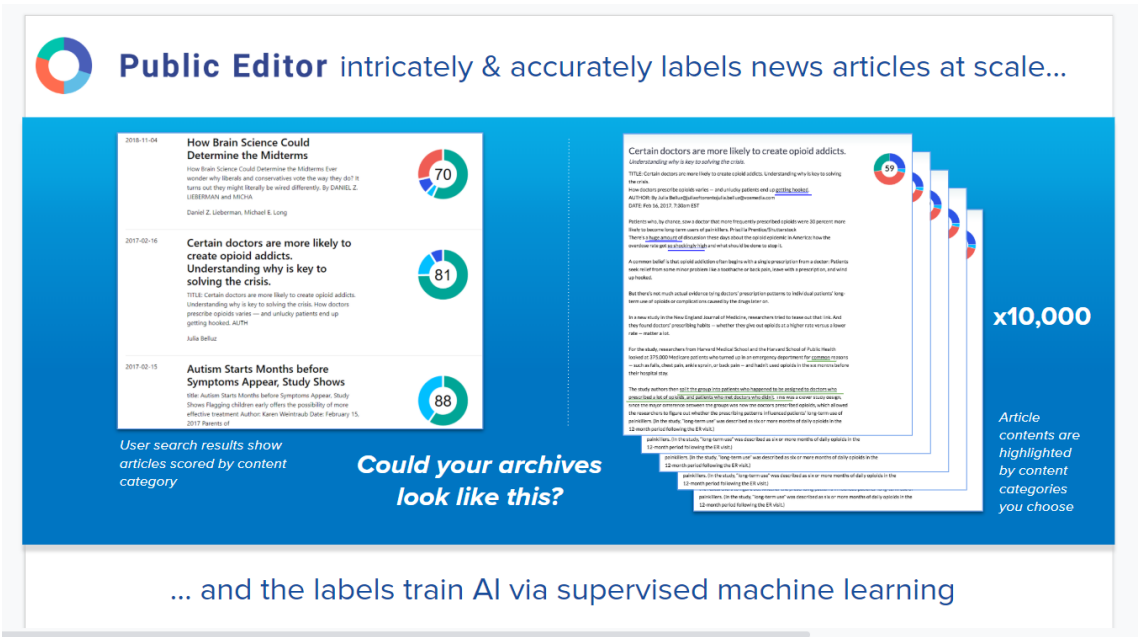
Helena Sarin, Neural Bricolage: "Visual Truth in the AI Era: Generative Models"

- Sarin applies machine learning models to see the world differently. By using her own artwork as training data, Sarin uses generative models to produce new art which is a "new, partial construct different from its sources."



Nick Adams, TagWorks: "Supervised Benchmarking will transform humanities"

- TagWorks is a project that decomposes a scholar's analytical expertise into multiple simplified data labeling interfaces that guide non-experts in applying labels to documents, images, video, and 3-D holograms of objects. The purpose is to crowd-source the labeling of datasets to train machine learning algorithms to similarly label the rest of the archive.



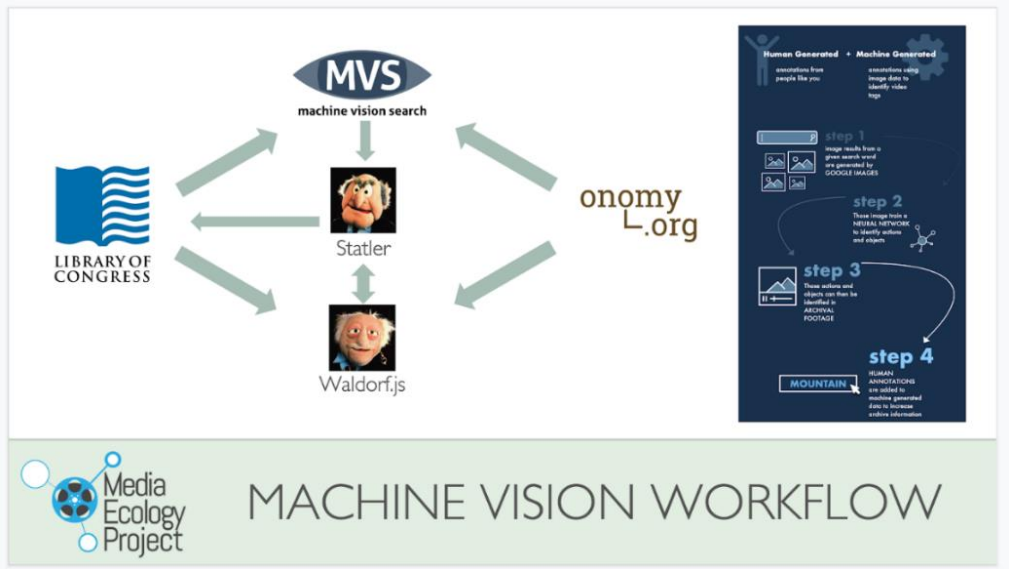
Ross Goodwin, Data Artist: “AI & Artists”

- Goodwin’s talk centered on the idea that we cannot trust artificial intelligence until humans can produce data that reflects the best parts of ourselves. He argues that there is a need to have artists using AI technologies.

Partnerships & Vendors: Opportunities & Challenges

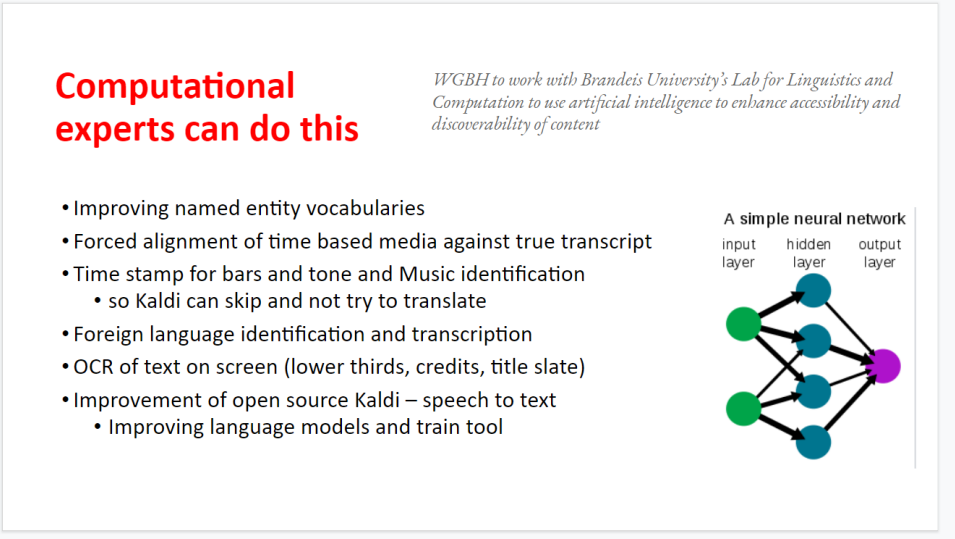
Mark Williams, John Bell, Dartmouth Media Ecology Project: "Semantic Annotation Tool"

- The Media Ecology Project's NEH-funded Semantic Annotation Tool (SAT) enables the creation of granular time-based annotations of moving image videos. The project is also funded to begin to create interface and semantic description strategies that would make moving image files accessible to blind and low-vision users.



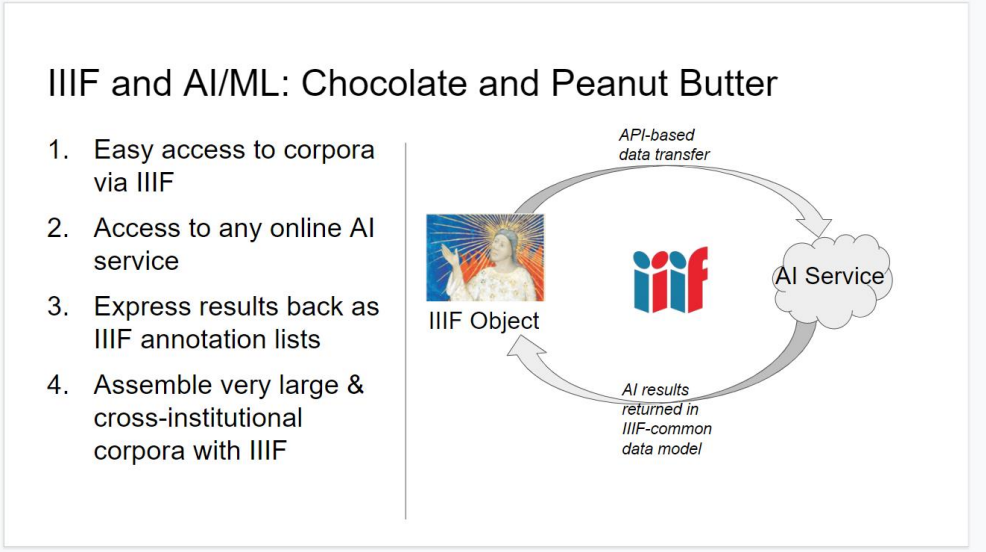
Karen Cariani, WGBH Media Library and Archives: "Speech to Text for Audiovisual Materials"

- Cariani's talk discussed how archivists need computer scientists to help with improving tools and ML but also that archivists have the data set/collections needed to train computational tools for learning and improvement. WGBH's current project will use OCR of text on video to help identify speakers, copyright info, verify program titles, funders, and credit rolls.



Josh Hadro, IIIF, and Tom Cramer, Stanford Libraries: "IIIF and AI/ML: Chocolate and Peanut Butter"

- Hadro highlighted the interoperability of using the IIIF standard for accessing cross-institutional corpora for use with machine learning. Cramer disussed the need to identify and establish channels for concrete exchange among libraries, archives, and museums for practical developments and application of artificial intelligence.

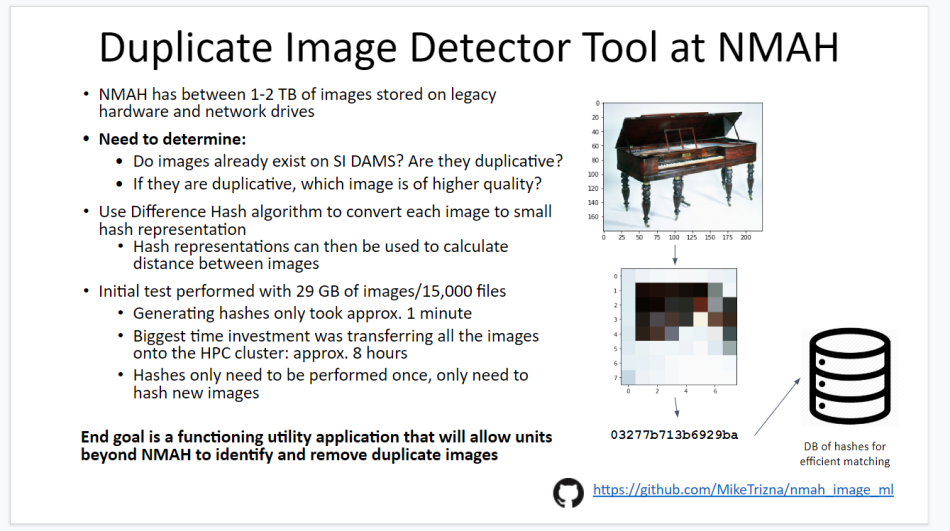


Mia Ridge, British Library: "Challenges in operationalizing data science"

- Ridge touched on three main kinds of challenges: scale, operational and interdisciplinary, and copyright. A larger scale requires new worflows and quickly grows expensive, operationalizing raises the question of producing public-facing infrastructure, and copyright involves negotiating complex rights issues.

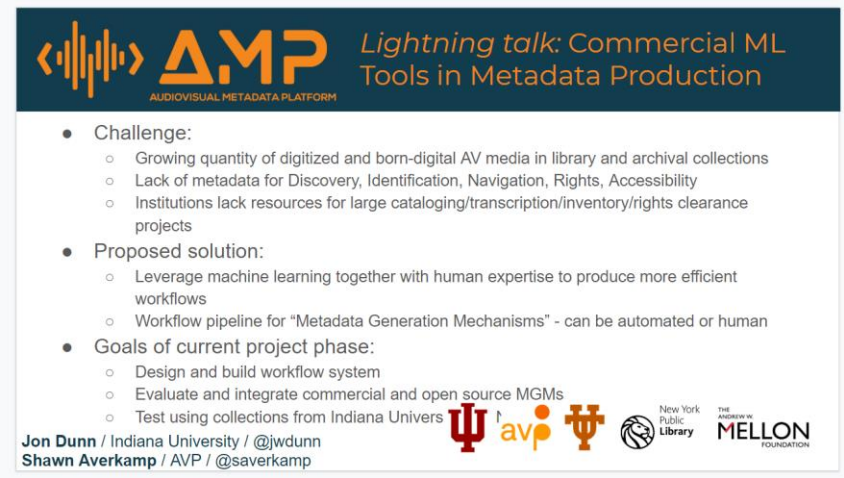
Rebecca Dikow, Corey DiPietro, Mike Trizna, Smithsonian: "Machine Learning at the Smithsonian"

- Dikow, DiPietro, and Trizna shared findings from three projects at the Smithsonian: ongoing work at the Data Science Lab, vision processing at the National Museum of American History (NMAH), and duplicate image detector tool at NMAH.



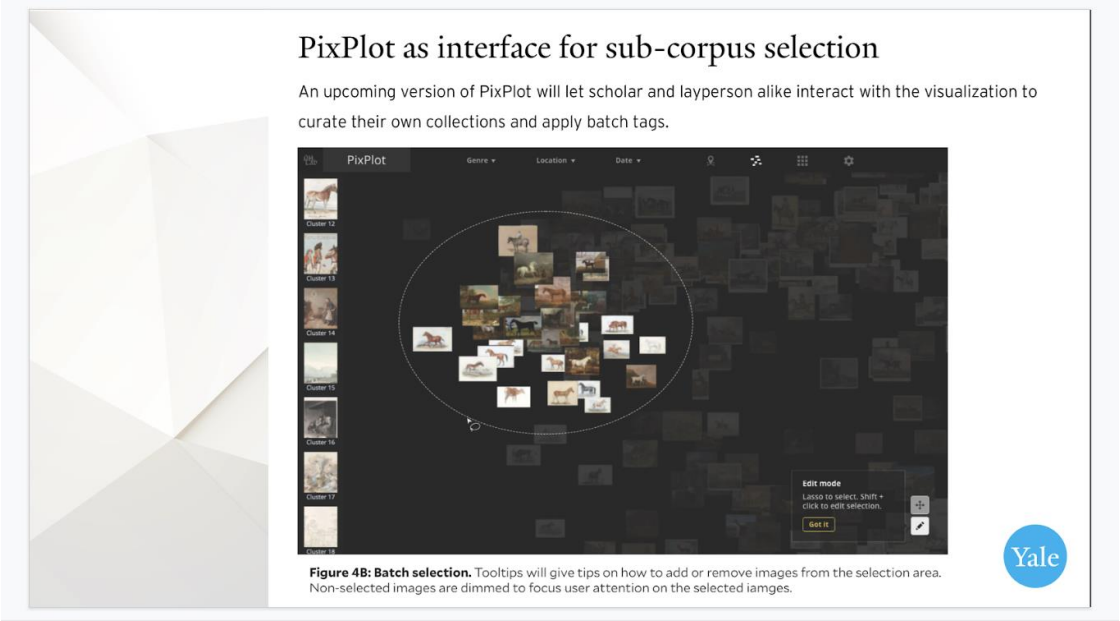
Jon Dunn, Indiana University, Shawn Averkamp, AVP: "Commercial ML Tools in Metadata Production"

- Indiana University is working with partners to design and build an open source software system known as AMP (Audiovisual Metadata Platform). AMP will enable the creation of workflows that incorporate both machine learning-based tools (commercial and open source) and human expertise to enable more efficient generation of metadata for digital audio and moving image resources supporting discovery, identification, navigation, and rights determination.



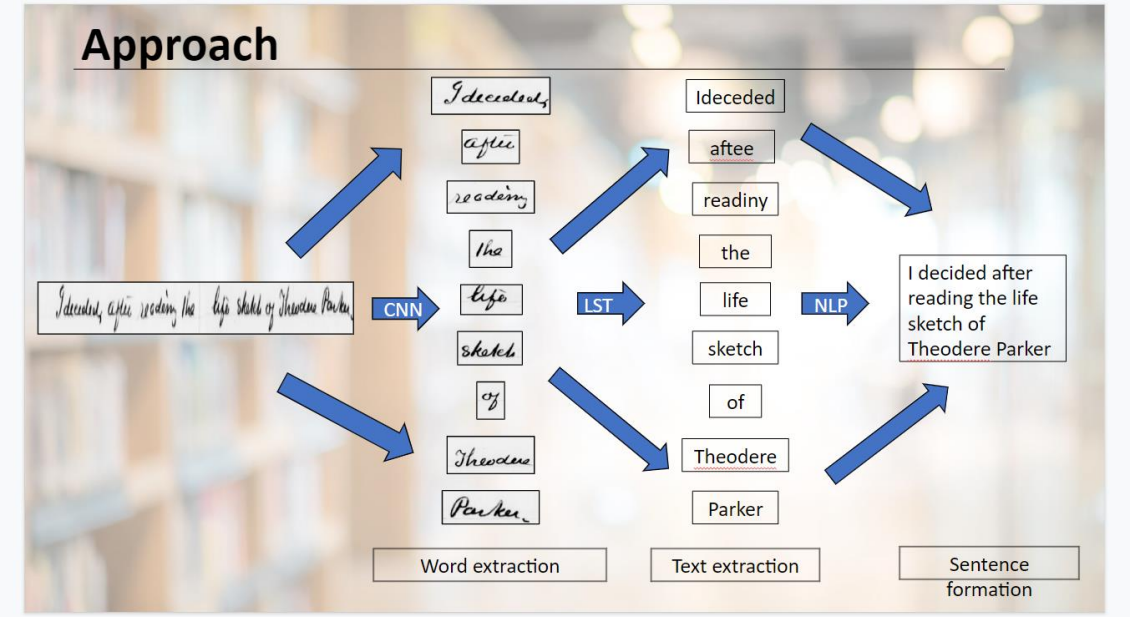
Peter Leonard, Yale Digital Humanities Lab: "PixPlot"

- PixPlot uses machine learning algorithms to group collections of images by their visual similarity allowing scholars to experience and search these items in a new way. Similar images appear proximate to one another.



Anishi Mehta, Georgia Tech: "Digitizing Historical Documents via Deep Learning: A Proof of Concept Study"

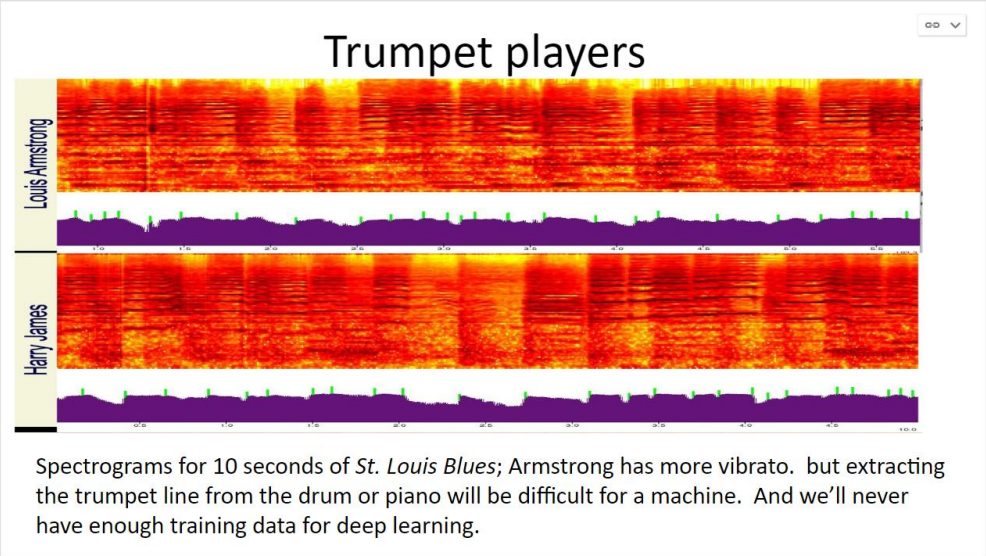
- Mehta presented preliminary results from her project testing handwriting recognition in historical documents.



Future Applications

Michael Lesk, Rutgers University: "Metadata for sound and pictures"

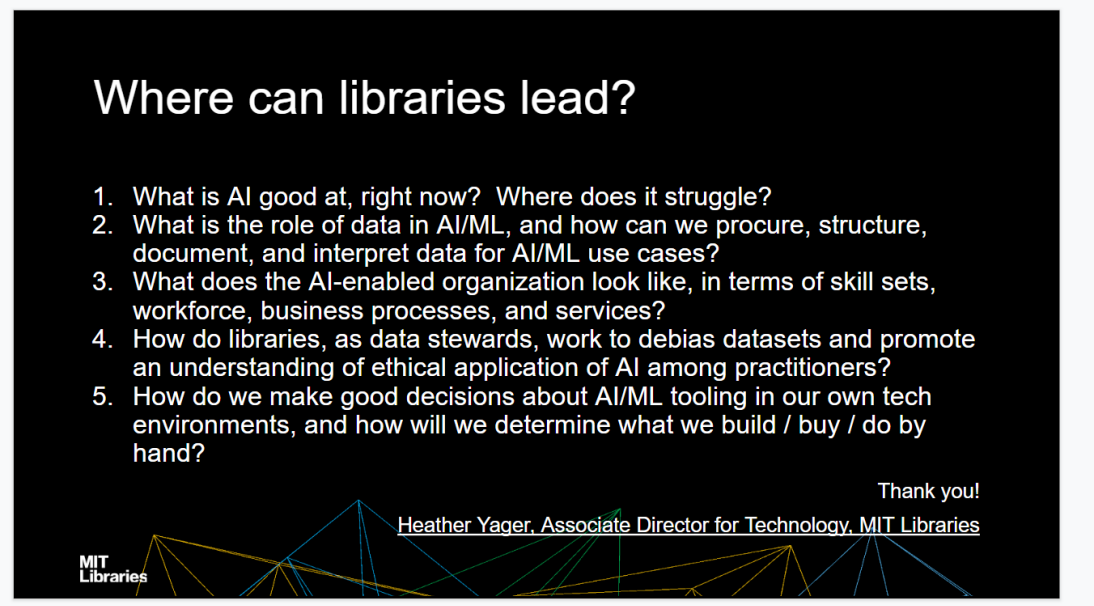
- Lesk discussed a project working on machine recognition of trumpet players in the vinyl jazz recordings at Rutgers.



Spectrograms for 10 seconds of *St. Louis Blues*; Armstrong has more vibrato. but extracting the trumpet line from the drum or piano will be difficult for a machine. And we'll never have enough training data for deep learning.

Heather Yager, MIT Libraries, "The future of work"

- Yager presented on the need to redesign workflows in order to make use of the strengths and limitations of artificial intelligence. Her discussion of bias in datasets led her to pose the question: where can libraries lead? See "questions for consideration" for a complete list.



Audrey Altman, Digital Public Library of America (DPLA): "Machine Learning at DPLA"

- DPLA has been working for the past couple of years to build out a new ETL system capable of supporting ML inquiry into their dataset of over 30 million library metadata records. A current topic-modeling pilot project aims to improve "more like this" recommendations for end-users, and to provide our contributing libraries with information about their institutions' topical coverage.

Hannah Davis, Research Artist and Generative Composer: "A Dataset is a Worldview"

- Using her projects on sentiment analysis and sonification, Davis showcased the subjectivities present in datasets and called for the recognition that "a dataset is a worldview."