



Closing the Loop Between ML Research & Library Systems

Ryan Cordell

School of Information Sciences
University of Illinois Urbana-Champaign

ryancordell.org | @ryancordell

Search

All Formats

Search Loc.gov

GO

THE SIGNAL

Search this blog

GO

Print Subscribe Share/Save

About This Blog
ISSN 2691-672X

Categories

- Abbie Grotke
- At the Museum
- Brian Foo
- By the People Transcription Program
- Camille Salas
- Chad Conrady
- Charlotte Kostelic
- computational research
- Content Matters Interview
- Courtney McClellan
- crowdsourcing
- Data Librarianship
- Derek Miller
- Digital Content
- Digital Humanities
- digital preservation
- digital scholarship
- DPOE Interview
- Duff Johnson
- Education and Training
- Elizabeth Brown
- FADGI
- formats

Machine Learning + Libraries: A Report on the State of the Field

July 22, 2020 by Leah Weinryb-Grohsgal

Digital collections in libraries are vast—and growing, as we continue to digitize cultural heritage materials and acquire new born digital collections. At the same time, the use of machine learning and artificial intelligence has grown exponentially. At LC Labs, we explore how technology can help fulfill the Library of Congress’s [vision](#) that “all Americans are connected to the Library of Congress,” sharing this knowledge and learning from others. To this end, we are delighted to release a new report, “[Machine Learning + Libraries: A Report on the State of the Field](#),” by Dr. Ryan Cordell.

Dr. Cordell is an Associate Professor of English at Northeastern University and an experienced digital humanities practitioner. The report is part of the Library’s 2019 “Season of Machine Learning,” sponsored by LC Labs and the Digital Strategy Directorate. The practical applications of machine learning have great potential to improve access, discovery, and engagement in library collections. But, as Dr. Cordell points out, libraries are also in a position to lead broader conversations about responsible use of these technologies.

The report expertly frames machine learning and both its possibilities and its challenges. After defining machine learning and distinguishing it from the broader category of artificial intelligence (AI), Dr. Cordell describes the history of the use of these technologies in libraries and explores various approaches. Urging care and thoughtful implementation of machine learning, the report points out the risks and potential benefits for crowdsourcing, searchability of type and handwritten documents and audiovisual materials, collection management, conservation and preservation, and creative and artistic works. The report concludes with recommendations for responsibly implementing machine learning, improving access to data, developing infrastructure, and cultivating expertise.



[Madge Lessing, full length, on bicycle, facing left; holding musical horn to lips]. Photograph copyrighted by E. Chickering, c1898. Library of Congress Print s& Photographs Division. [//www.loc.gov/resource/cph.3b10346/](http://www.loc.gov/resource/cph.3b10346/)

Machine Learning + Libraries

A Report on the State of the Field

Ryan Cordell

Associate Professor
Northeastern University English Department
r.cordell@northeastern.edu

Commissioned by LC Labs
Library of Congress

Version Published: July 14, 2020

The majority of machine learning (ML) experiments in libraries stem from a simple reality: human time, attention, and labor will always be severely limited in proportion to the enormous collections we might wish to describe and catalog. ML methods are proposed as tools for enriching collections, making them more useable for scholars, students, and the general public. ML is posited as an aide to discoverability and serendipity amidst informational abundance. We might imagine, for example, patrons browsing automatically-derived topics of interest across a digital library comprising thousands or millions of texts—more texts, certainly, than typical constraints on labor or expertise would allow us to imagine labelling manually.

Ryan Cordell, "Machine Learning + Libraries" (2020)

How can the reliability of ML data and metadata be assessed, and how can probabilistic information be integrated with human-created information, or integrated into systems designed around hand-assigned categories, tags, summaries, and so forth? To phrase this central question in another way, the ML and libraries field must develop means to bridge a world that prioritizes expert data and metadata, created slowly, and a set of methods that generate useful but flawed data and metadata, more quickly and at a larger scale [...] ML derived data will likely never meet the gold standard for reliability, but could nonetheless enhance discoverability in some of the ways researchers have promised for decades.

Ryan Cordell, "Machine Learning + Libraries" (2020)

5.4.2. Pilot Integrated Interfaces for Communicating ML-Derived Data

As I describe in 4.5, much of the literature in ML and libraries is composed in the “perpetual future tense.” Projects are framed in terms of the discoveries they will spark by allowing users to browse or search library data in new ways. The applications outlined in Section 3 of this report exemplify this perpetual future tense. The many activities outlined under the heading of “discoverability” in 3.2— e.g. clustering, classification, metadata exaction—do not become really valuable until users can make use of them to explore. Currently there is a stark divide between most collections and the outputs of the research those collections enable, but this divide could be bridged as ML data is used to directly enrich collections. Projects such as “National Neighbors”¹⁵⁹ and “Neural Neighbors”¹⁶⁰ suggest what ML-integrated interfaces might look like, and the ways they might enable researchers, students, and other users to interact with collections in new ways, but even these interfaces do not directly tie into library catalog systems.

I strongly recommend that libraries strive not for polish in these interfaces, but instead for explicit communication of ML processes and decisions. In our interview, Benjamin Lee identified human-computer interaction (HCI) as one of the most fruitful areas for collaboration between libraries and computer scientists around machine learning. From the perspective of HCI, he argued, a researcher

should not wait for the data to be perfect, but instead present it as a pilot or prototype, learn from users, and refine from there.¹⁶¹ Where computer scientists bring expertise in ML methods, libraries understand information literacy and can help construct interfaces that communicate ML data to a wide range of prospective users.

ML-integrated interfaces should report, for instance, the confidence scores for relationships, annotations, or other metadata determined through an ML algorithm, and seek to make users more aware of the probabilistic basis of this data. The aim of such reporting is not simply to cast doubt—though skepticism is healthy in this domain—but to make presentations of ML data opportunities for cultivating literacy for ML and probabilistic methods. Helping users understand the confidence rating behind a particular label or category helps contextualize any claims they might make. A sense of ML’s limitations could, perhaps counterintuitively, serve to increase overall confidence in ML, because its claims will be understood as contextual and relational rather than totalizing. Through the cultivation of explaining, ML-integrated interfaces, the library will help meet the educational goals I outline in 5.5.1 below, centering its engagements with ML through pedagogy and outreach.

5.4.2. Pilot Integrated Interfaces for Communicating ML-Derived Data

As I describe in 4.5, much of the literature in ML and libraries is composed in the “perpetual future tense.” Projects are framed in terms of the discoveries they will spark by allowing users to browse or search library data in new ways. The applications outlined in Section 3 of this report exemplify this perpetual future tense. The many activities outlined under the heading of “discoverability” in 3.2— e.g. clustering, classification, metadata exaction—do not become really valuable until users can make use of them to explore. Currently there is a stark divide between most collections and the outputs of the research those collections enable, but this divide could be bridged as ML data is used

to di
sugg
stud
direc

5.4.2. Pilot Integrated Interfaces for Communicating ML-Derived Data

I str
com
com
com

shot
user
unde
wide

ML-integrated interfaces should report, for instance, the confidence scores for relationships, annotations, or other metadata determined through an ML algorithm, and seek to make users more aware of the probabilistic basis of this data. The aim of such reporting is not simply to cast doubt—though skepticism is healthy in this domain—but to make presentations of ML data opportunities for cultivating literacy for ML and probabilistic methods. Helping users understand the confidence

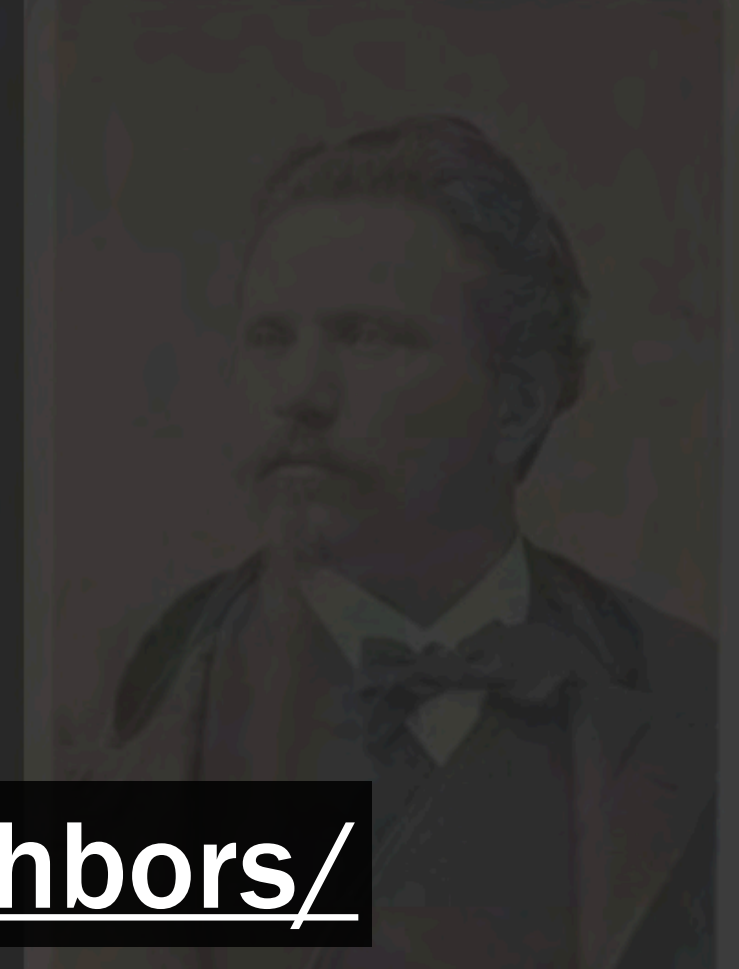
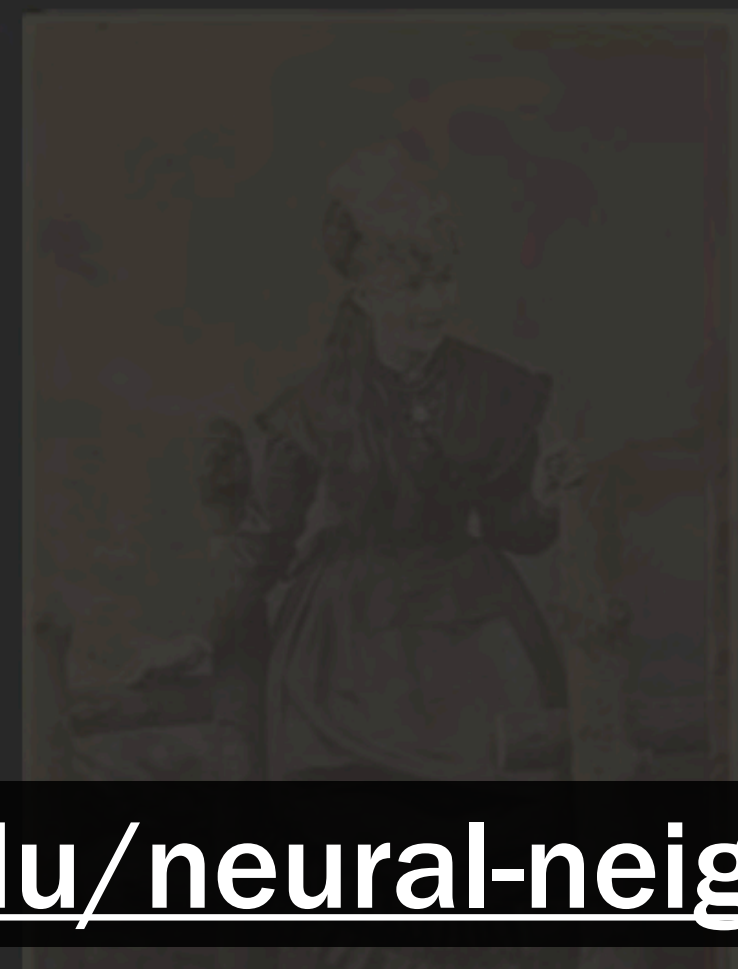
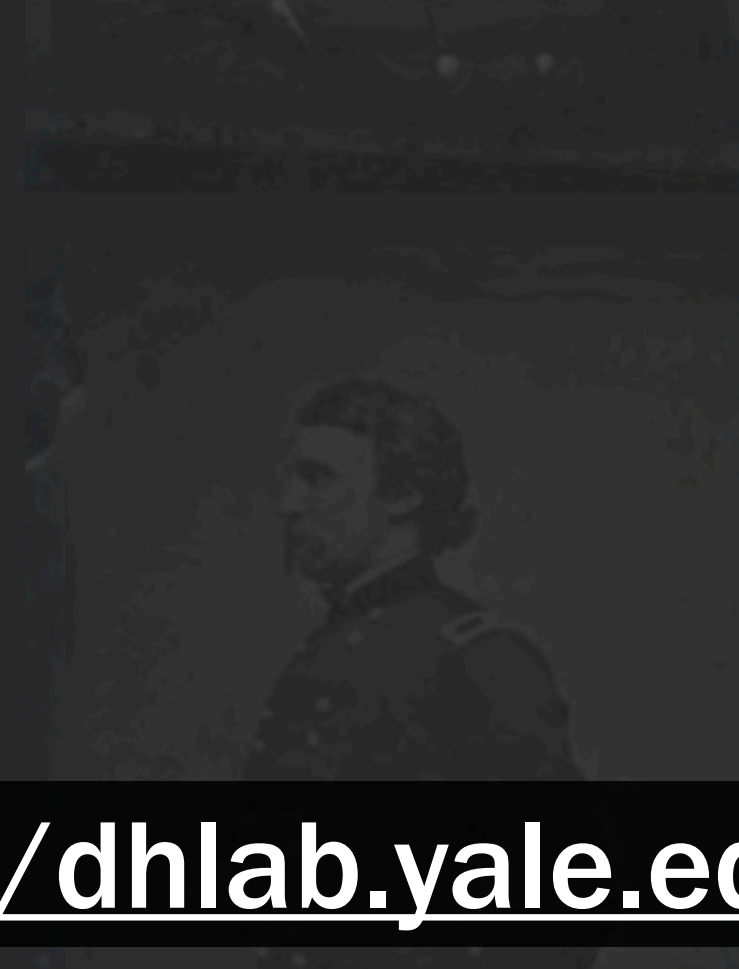
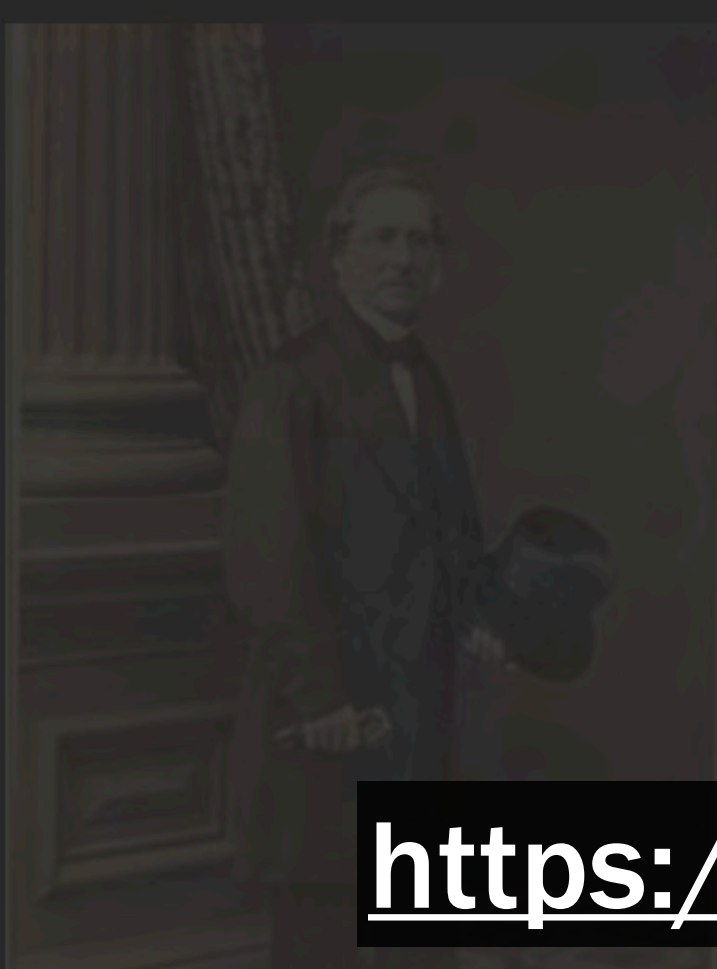
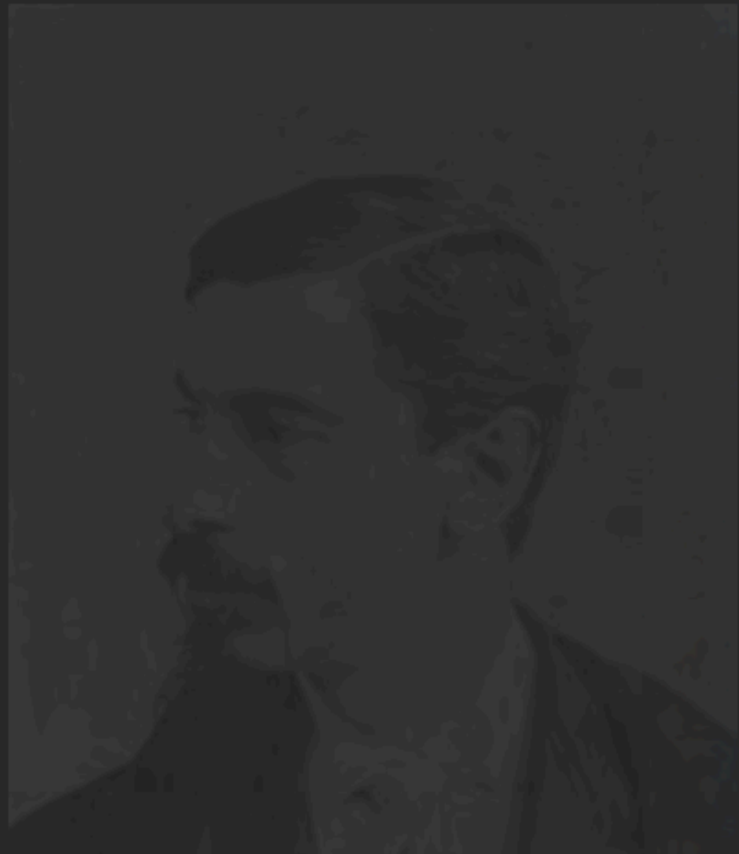
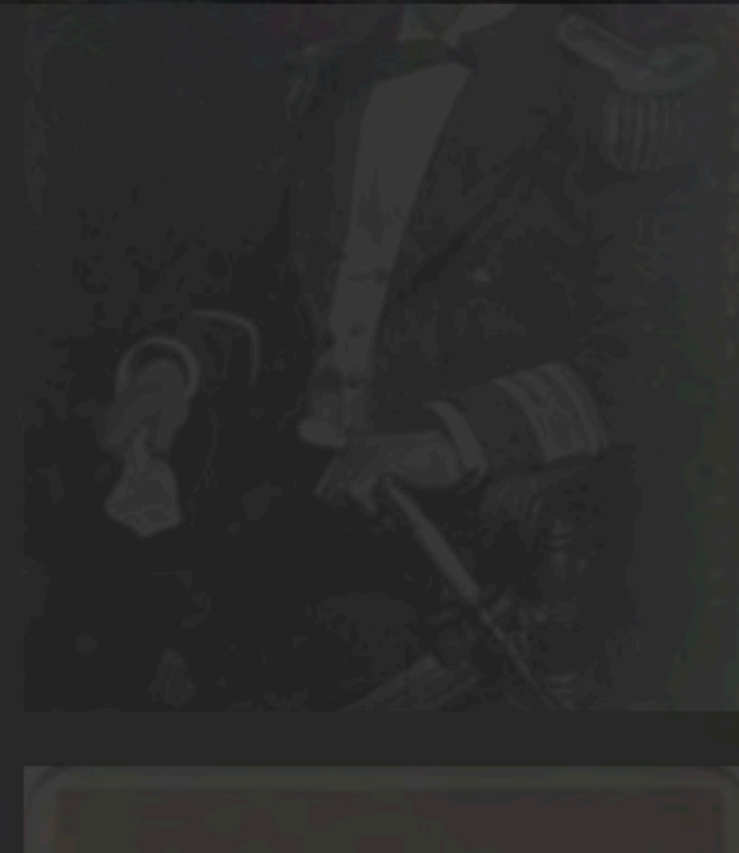
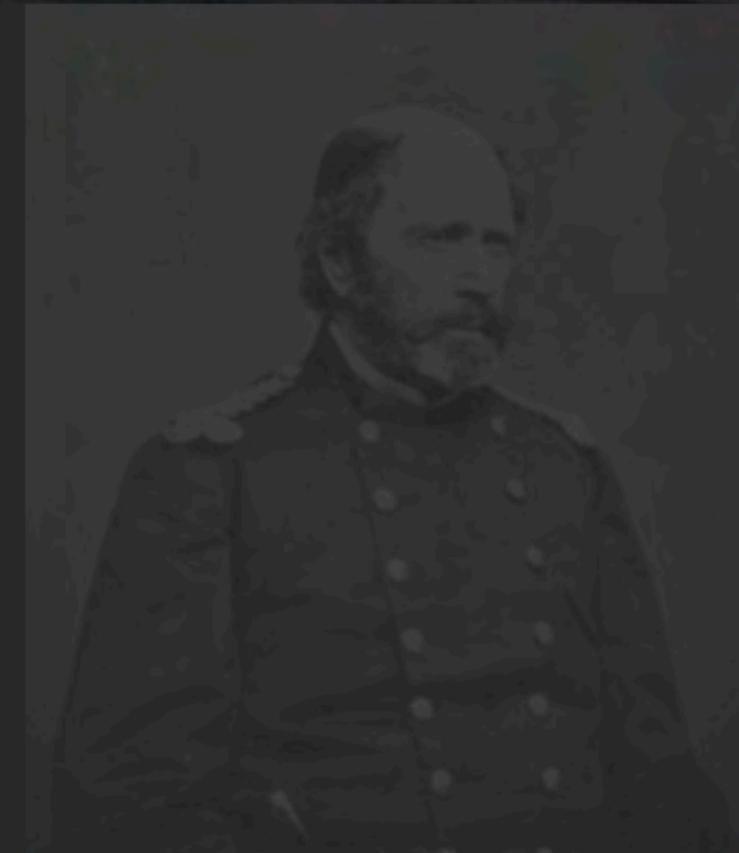
ML-integrated interfaces should report, for instance, the confidence scores for relationships, annotations, or other metadata determined through an ML algorithm, and seek to make users more aware of the probabilistic basis of this data. The aim of such reporting is not simply to cast doubt—though skepticism is healthy in this domain—but to make presentations of ML data opportunities for cultivating literacy for ML and probabilistic methods. Helping users understand the confidence rating behind a particular label or category helps contextualize any claims they might make. A sense of ML’s limitations could, perhaps counterintuitively, serve to increase overall confidence in ML, because its claims will be understood as contextual and relational rather than totalizing. Through the cultivation of explaining, ML-integrated interfaces, the library will help meet the educational goals I outline in 5.5.1 below, centering its engagements with ML through pedagogy and outreach.

Keyword search is ungenerous: it demands a query, discourages exploration, and withholds more than it provides. This paper argues instead for *generous interfaces* that better match both the ethos of collecting institutions, and the opportunities of the contemporary web. Generous interfaces provide rich, navigable representations of large digital collections; they invite exploration and support browsing, using overviews to establish context and maintain orientation while revealing detail at multiple scales. Generous interfaces use multiple, fragmentary representations to reveal the complexity and diversity of cultural collections, and to privilege the process of interpretation. While they draw on techniques and models established in information retrieval and visualisation, generous interfaces emphasise process, pleasure and thoughtful engagement rather than the functional satisfaction of an information need.

Michael Whitelaw, "Generous Interfaces for Digital Cultural Collections" (2015)

Are we designing libraries that activate imaginations—both their users' imaginations and those of the expert practitioners who craft and maintain them? Are we designing libraries emancipated from what I'll shortly demonstrate is often experienced as an externally-imposed, linear and fatalistic conception of time? Are we at least designing libraries that dare to try, despite the fundamental paradox of the Anthropocene era we live in—which asks us to hold unpredictability and planetary-scale inevitability simultaneously in mind? How can we design digital libraries that admit alternate futures—that recognize that people require the freedom to construct their own, independent philosophical infrastructure, to escape time's arrow and subvert, if they wish, the unidirectional and neoliberal temporal constructs that have so often been tools of injustice?

Bethany Nowviskie, "Speculative Collections" (2016)



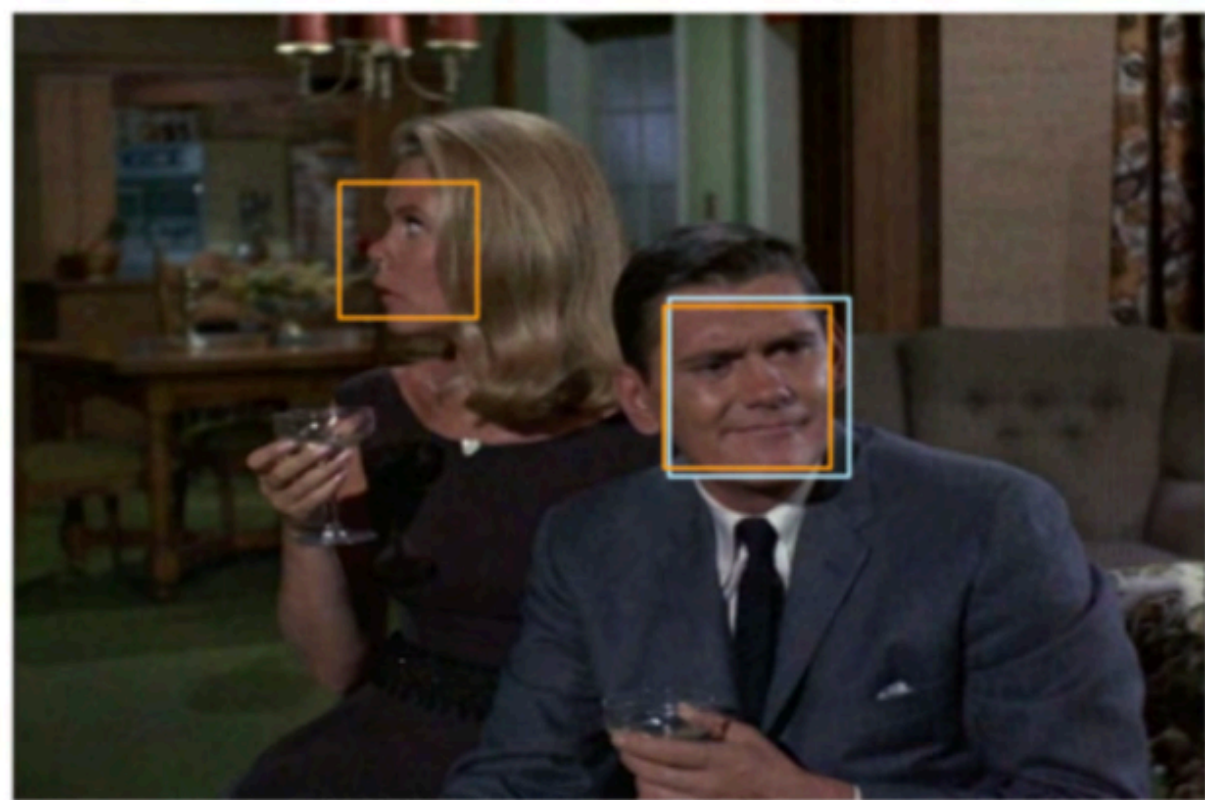
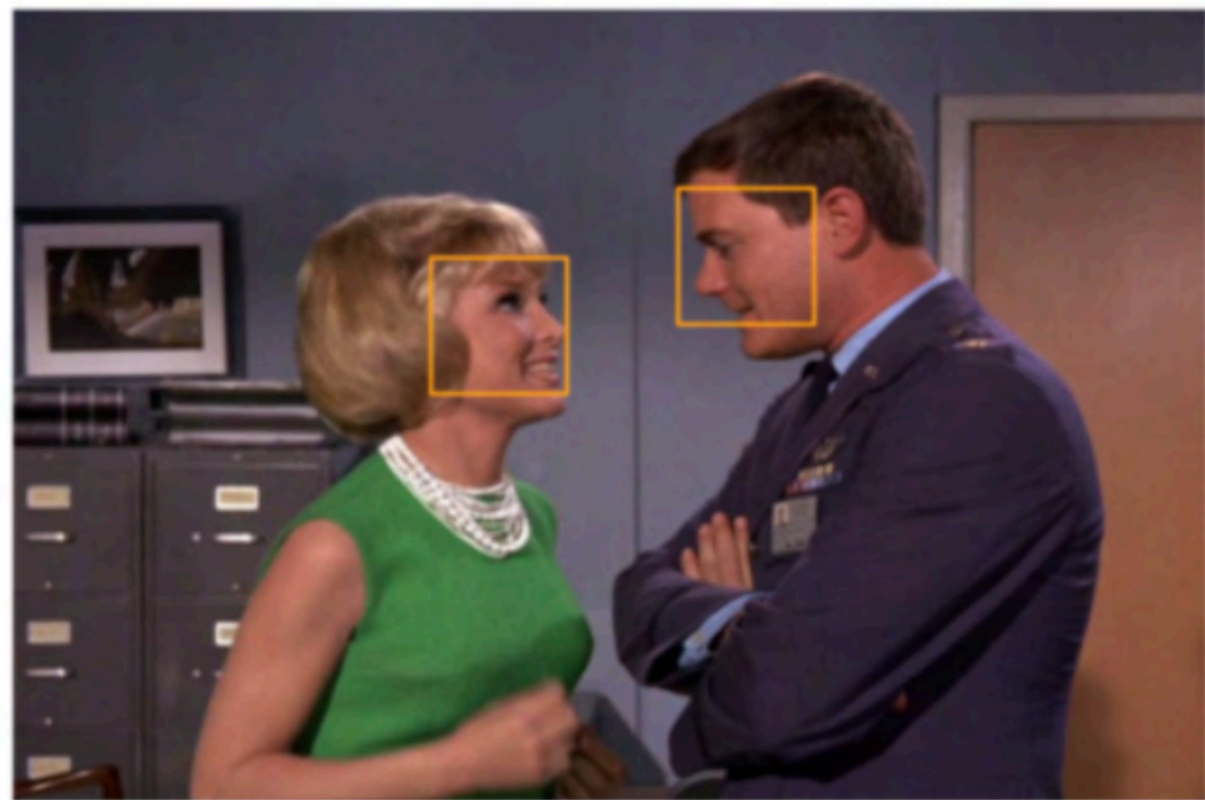
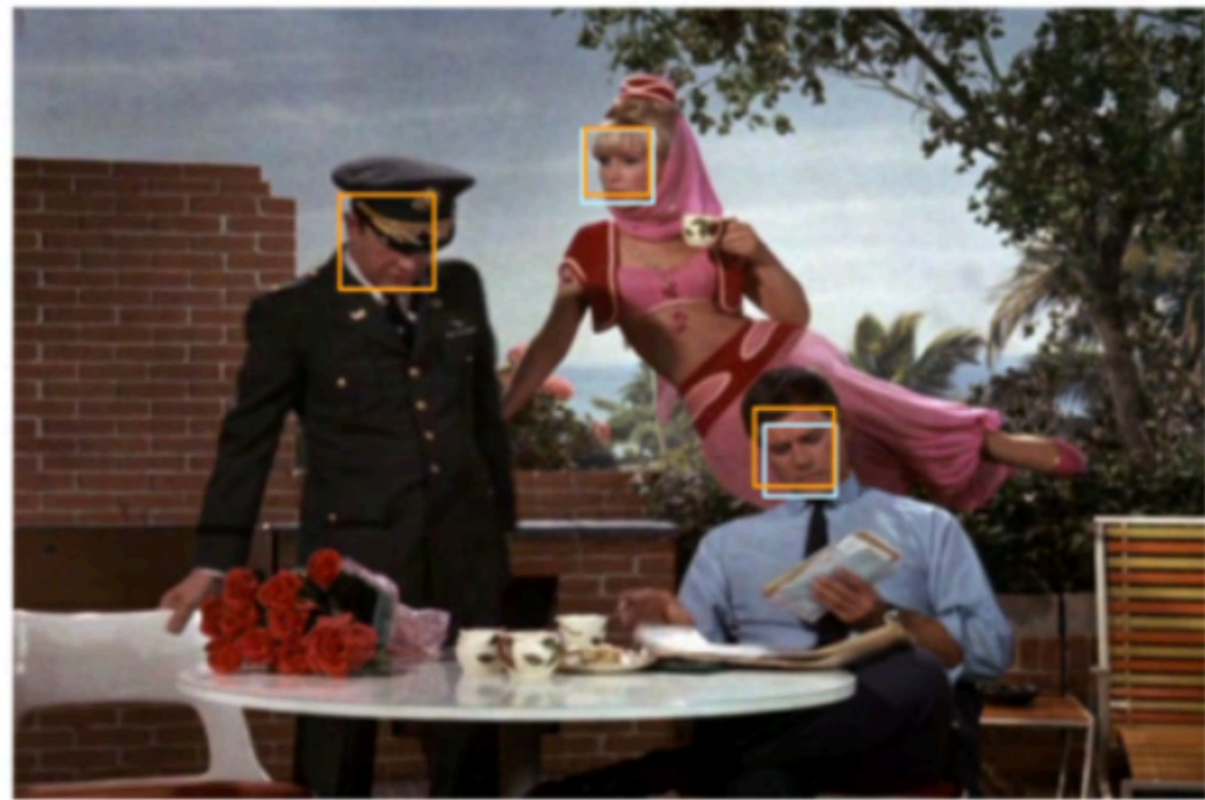


Figure 1: Faces detected using a HOG detector (blue) and a neural network (orange) from screen shots of I Dream of Jeannie and Bewitched.

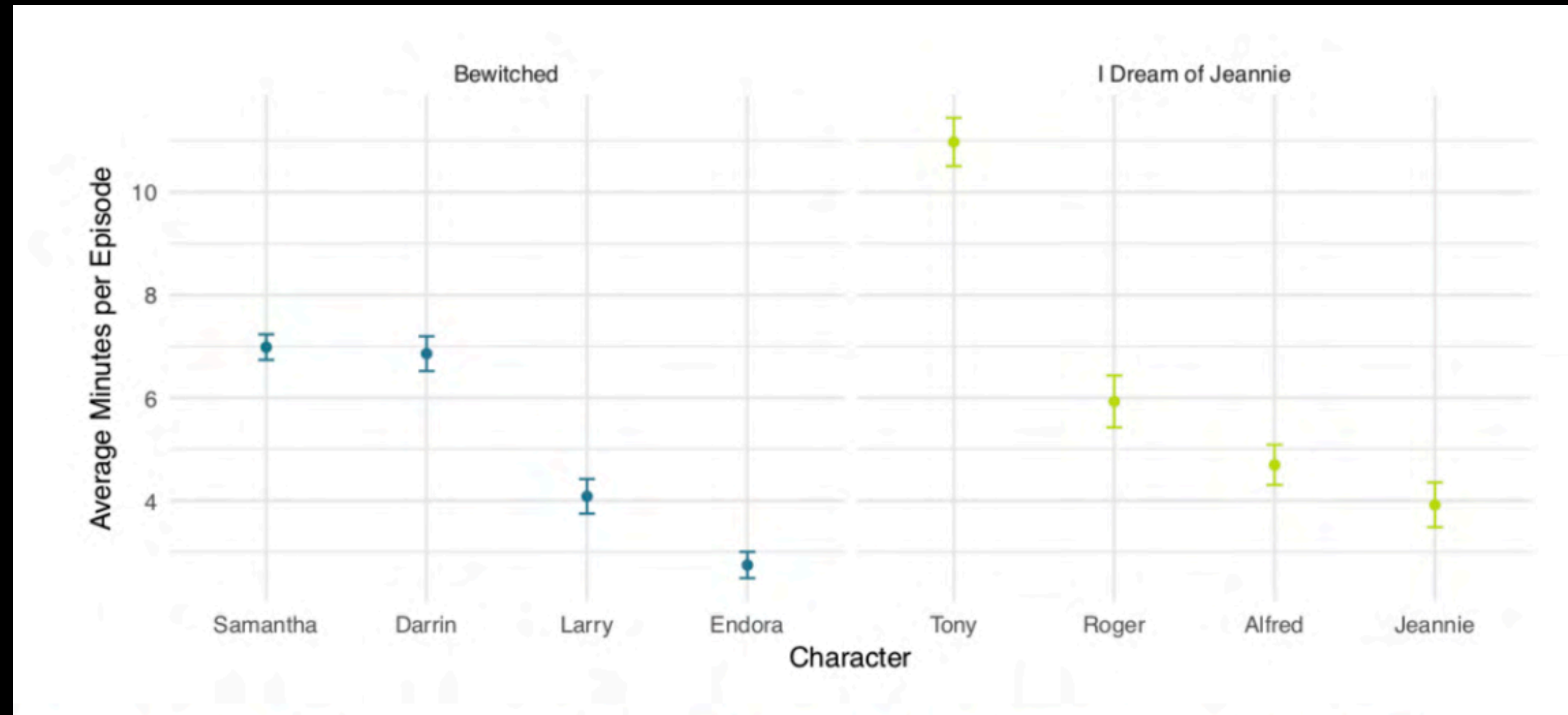


Figure 5: Average minutes per episode for which a character is visible. Error bars with 95% confidence intervals for the mean of each group.

A Pattern History of Jazz

Patterns (653)

[-1, -1, -1, -2, -2, -1] (127)

Sort by:

Frequency Value LEP

Contains large jumps (at least a fourth)

Search Pattern

Filter by Performer:

Filter by Length

Length Comparison

<= == >=

Pattern Type:

All
 Repetitive
 Diatonic
 Chromatic
 Arpeggio
 Mixed

Huron Contour:

All

Info

[Listen & See](#)

[Instances](#)

[Stats](#)

[Timeline](#)

[General Stats](#)

[Help](#)

Dig That Lick

This is an interactive visualization of the pattern history of jazz solo improvisation developed by the [Dig that Lick Project](#). The patterns were extracted with help of the **melpat** application from the [MeloSpySuite](#) developed by the [Jazzomat Research Project](#).

The data was gathered by calculating pattern partitions of selected artists (Bob Berg, Michael Brecker, Clifford Brown, Steve Coleman, John Coltrane, Miles Davis, Dave Liebman, Charlie Parker, Sonny Rollins, Woody Shaw, and Wayne Shorter) and then searching for these patterns in the WJD. These patterns partition were calculated for a minimum length of six intervals occurring at least three times in three different solos. Exceptions are the Bob Berg patterns where the condition was minimum length of seven with two instances in at least two different solos. Another exception are the Charlie Parker patterns which were taken not from the WJD but from the Omnibook with the condition of at least six intervals occurring at least ten times in ten different solos.

Currently, the data set contains 11630 instances of 653 distinct patterns. For the future, it is planned to add more patterns by other artists.

Current Version: **0.9**

Design & Coding: **Klaus Frieler**

Powered by [Shiny](#).

ISMIR 2018 Paper: [Two web applications for exploring melodic patterns in jazz solos](#).

(c) 2018 **Dig That Lick Project**

Copyright notice: All audio snippets included here are solely used for educational and scientific purposes.

Have fun!

https://jazzomat.hfm-weimar.de/pattern_history/

Transcribe. Collaborate. Share...

...and benefit from cutting edge research in Handwritten Text Recognition!

[Download version 1.11](#)[Download version 1.11 for Mac](#)[Wiki » How-to guide \(pdf\) »](#)

Scholars

Are you transcribing historical documents? Handwritten or printed, from the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and reliable way? And do you appreciate to get support from automated tools such as Handwritten Text Recognition and Layout Analysis?

[View details »](#)

Archives

Are you responsible for large collections of handwritten and printed documents? Do you believe that digitisation paves the way to realise new opportunities to access, enrich and explore archival material? And are you open to involve humanities scholars and volunteers so that they can work with these documents in an effective way – producing data which can also be integrated in your repository?

[View details »](#)

Volunteers

Are historical letters, postcards, manuscripts or medieval documents fascinating for you? Do you enjoy deciphering handwriting – this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to scholarship and science?

[View details »](#)

Scientists

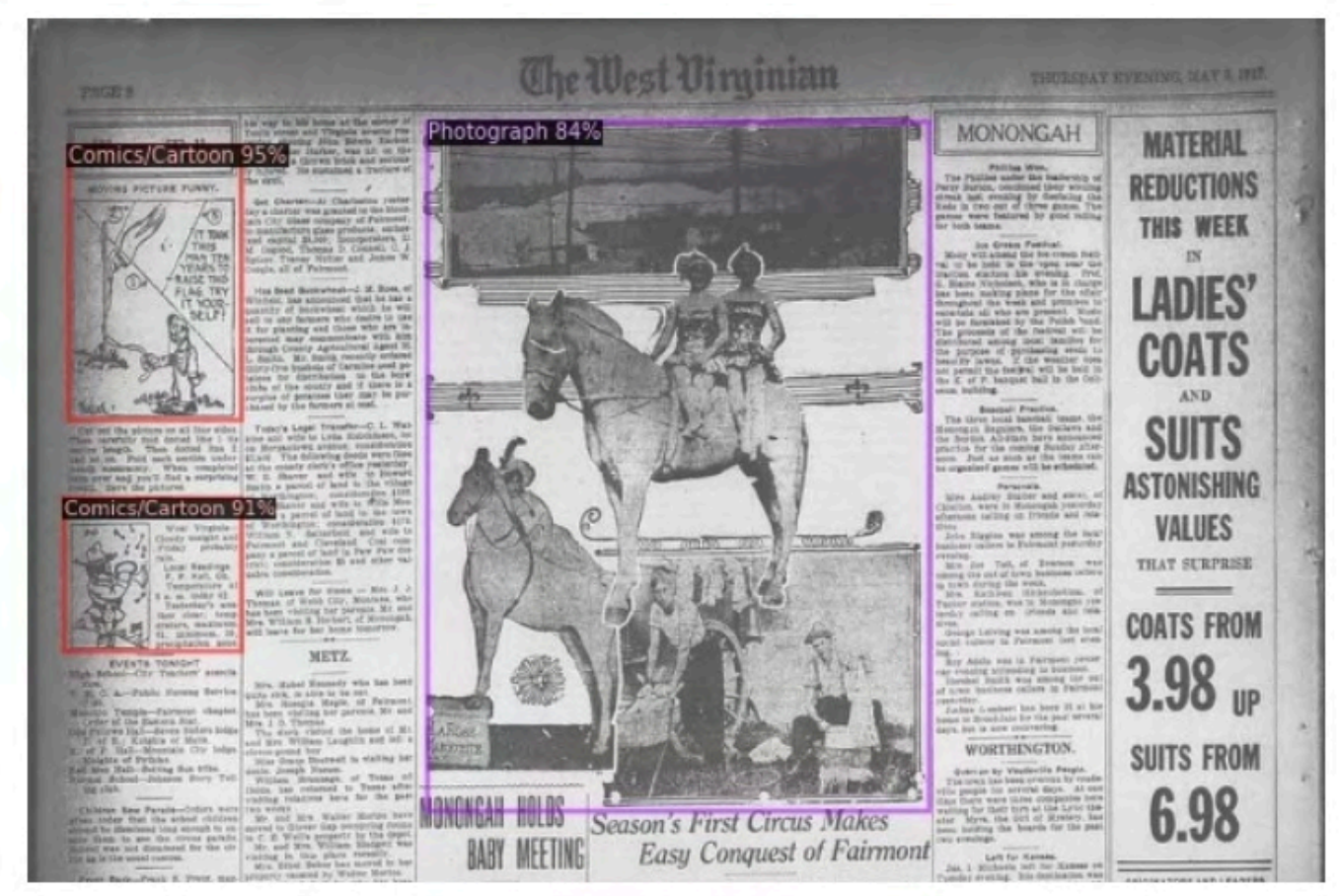
Are you a computer scientist and working in the fields of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and publishing? And you would really enjoy to get some reference data in a well-acknowledged format, such as PAGE?

[View details »](#)

<https://transkribus.eu/Transkribus/>

« Experiments

Newspaper Navigator



Get Data!

Explore the visual and textual content within the Chronicling America digitized newspaper collection in new ways using machine learning!

Coming Summer 2020

What is Newspaper Navigator?

Newspaper Navigator is a project by Ben Lee currently under development during his time as an Innovator-in-Residence at the Library of Congress. The first stage of Newspaper Navigator is to extract content such as photographs, illustrations, cartoons, and news topics from the Chronicling America newspaper scans and corresponding OCR using emerging machine learning techniques. The second stage is to reimagine an exploratory search interface over the collection in order to enable a wide range of people to navigate the collection according to their interests. With Newspaper Navigator, Ben hopes to engage the American public, enable new digital humanities and cultural heritage research, and advance computer science research.

For more information on *Newspaper Navigator*, feel free to consult [Ben's repo](#), which includes code, a whitepaper, and demos, all of which are being regularly updated.

<https://labs.loc.gov/work/experiments/newspaper-navigator/>

Try these neural network-generated recipes at your own risk.

Stuck in a rut in the kitchen? Tired of preparing sandwiches the same old way?

Machine learning can help!

I trained a [neural network](#) on over 30,000 examples of cookbook recipes, and it learned to produce new recipes of its own. You can learn more about the training process, and watch it learn to generate new recipes [here](#).

They aren't good recipes, though. In fact, almost all of them are terrible. [I made one of them once](#), and now I still cringe at the faintest whiff of horseradish. [SuperDeluxe made another of them](#), but at least they are professionals and were wise enough not to eat any.

Here for your entertainment I give you several more recipes the neural network has generated, with the caveat that if you should try to prepare or, god forbid, actually consume one of these, I am absolutely not responsible for the consequences.

Small Sandwiches

dish, chili, lemon, salads, seafood

½ cup shortening

1 cup snow peas and cut into ¼ inch cubes

1 1 inch

15 oz peach halves, remaining posting

1 salad dressing

½ cup barley

2 large bones sliced chicken salmon:

--FILLING--

1 cup minced season tomatoes

2 cup hot water

¼ cup vegetable oil

2 cup all-purpose flour

2 tablespoon the seasoned salt

1 cup margarine, melted

1 lb jumbo shrimp

1 or freshly ground black pepper

1 up

1 thai shrimp; finely chopped

1 garlic clove, minced

Mix all ingredients except cheese and process 1 hour. Pour over ribs.

Cover and bake for 30-35 minutes. Serve with warm milk and marinade distributed; prepare the bottoms.

Watch the end of the fillets to the heat and set in a bowl and heat at this low for 5 minutes, until softened. Top with a little the next 2 ingredients; spoon the one day, 1 ½ hours. Take an and inverting it and turn the center. Let cool in the pan on wire rack. To serve cooking time: It is been ribsotro. while the serving is alternatively rich will puree in the miquinally preparing gravy. They should seal.

Yield: 4 servings

Beothurtreed Tuna Pie

<https://aiweirdness.com/>



DONATE

RACIAL JUSTICE REQUIRES ALGORITHMIC JUSTICE. SUPPORT THE MOVEMENT

CONTACT

LIBRARY

SPOTLIGHT

ABOUT

TAKE ACTION

TECHNOLOGY SHOULD SERVE ALL OF US. NOT JUST THE PRIVILEGED FEW.

Join the Algorithmic Justice League in the movement towards equitable and accountable AI.

Enter your email

JOIN THE MOVEMENT

OUR LIBRARY

RESEARCH

ART/FILM

POLICY/ADVOCACY

MULTIMEDIA

EDUCATION

MEDIA

PRESS

<https://www.ajlunited.org/>



English



[R]ecent scholarship has warned that much of this technical work treats problematic features of the status quo as fixed, and fails to address deeper patterns of injustice and inequality. While acknowledging these critiques, we posit that computational research has valuable roles to play in addressing social problems — roles whose value can be recognized even from a perspective that aspires toward fundamental social change...Computing research can serve as a *diagnostic*, helping us to understand and measure social problems with precision and clarity. As a *formalizer*, computing shapes how social problems are explicitly defined—changing how those problems, and possible responses to them, are understood. Computing serves as *rebuttal* when it illuminates the boundaries of what is possible through technical means. And computing acts as *synecdoche* when it makes long-standing social problems newly salient in the public eye.

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson, "Roles for Computing in Social Change" (2020)

As disciplines primarily concerned with documentation collection and information categorization, archival studies have come across many of the issues related to consent, privacy, power imbalance, and representation among other concerns that the ML community is now starting to discuss. While ML research has been conducted using various benchmarks without questioning the biases in datasets, motives associated with the institutions collecting them, and how these traits shape downstream tasks, archives have...institutional and procedural structures in place that regulate data collection, annotation, and preservation that ML can draw from.

Eun Seo Jo and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning" (2019)

25 Clusters with Genre Probabilities



Jonathan Fitzgerald, English PhD



This text was likely reprinted in at least **193 other newspapers**

Likely Chronicling America reprints:

A gentleman claiming to be a "friend to the human race," and who keeps the run of facts, figures, and babies, has just laid before "an enquiring world" the following statistics:
"The whole number of languages spoken in the world amount to 3,064; 587 in Europe, 936 in Asia, 276 in Africa, and 1,264 in America. The inhabitants of our globe profess more than 1,000 different religions. The number of men is about equal to the number of women. The average of human life is about 33 years. One quarter part die

Rutland County Herald
Rutland, VT (1853-05-14)
[similarity scores](#)

Interesting Statistics.
A gentleman who keeps the run of facts, figures, and babies, has just laid before "an inquiring world," the following statistics:
The whole number of languages spoken in the world amount to 3,064; viz, 587 in Europe, 937 in Asia, 276 in Africa, and 1,264 in America.— The inhabitants of our globe profess more than 1,000 different religions. The number of men is about equal to the number of women. The average of human life is about 33 years. One fourth die previous to the age of 7

Sumter Banner
Sumterville, SC (1853-07-19)
[similarity scores](#)

FACTS IN HUMAN LIFE.
The number of languages spoken in the world, amounts to 3,064—587 in Europe, 896 in Asia, 276 in Africa, and 1,264 in America. The inhabitants of the globe profess more than 1,000 different religions. The number of men is about equal to that of the women. The average of human life is about 28 years. One-quarter die previous to the age of 7 years; one-half before reaching 17; and those who pass this age, enjoy a felicity refused to one-half the human species. To every 1,000 persons, only one reaches 100 years of life; to every 100, only 6 reach the age of 65; and not more than one in 500 live to 80 years of age.
There are on the earth, 1,000,000,000 inhabitants; and of these, 333,333,333 die every year; 91,334 every day; 3,780 every hour, and 60 every minute, or 1 every second. These losses are about balanced by an equal number of births. The married

Anti-Slavery Bugle
Lisbon, OH (1854-06-10)
[similarity scores](#)

Statistical Facts.
The number of languages spoken in the world amount to 3064. The inhabitants of the globe profess more than 1000 different religions. The number of men is about equal to the number of women. The average of human life is about 33 years. One quarter die previous to the age of seven years; and one-half before reaching seventeen; and those who pass this age enjoy a felicity refused to one-half the human species. Of every 1000 persons only one reaches 100 years of life; of every 100 only six reach the age of 65; and not more than one in 500 lives to 80 years of age. There are on the whole earth 1,000,000,000 inhabitants, and of these 33,333,533 die every year: 91,824

Placer Herald
Auburn, CA (1857-05-02)
[similarity scores](#)

[browse 189 more potential reprints...](#)

From the *Viral Texts Project*
Learn more about reprint detection at viraltxts.org

The daily Cairo bulletin., July 28, 1882, Image 2

About [The daily Cairo bulletin. \(Cairo, Ill.\) 1878-1???](#)

Image provided by: University of Illinois at Urbana-Champaign Library, Urbana, IL

Image viewer interface showing a newspaper page with various advertisements and text columns.

Navigation: Image: 2 of 4. Page All Pages Issues All Issues. Text PDF JP2 (4.5 MB)

Facts in Human Life.
There are three thousand sixty-four languages in the world, and its inhabitants profess more than one thousand religions. The number of men is about equal to the number of women. The average of life is about thirty-three years. One quarter die previous to the age of seventeen. To every one thousand persons only one reaches one hundred years of life; to every one hundred only six reach the age of sixty-five, and not more than one in five hundred lives to eighty years of age. There are on the earth one billion inhabitants; of these thirty-three million, thirty-three thousand thirty-three die every year, ninety-one thousand eight hundred and twenty-four every day, three thousand seven hundred and thirty every hour, and sixty every minute or one every second. The married are longer lived than the single, and above all those who observe a sober and industrious conduct. Tall men live longer than short ones. Women have more chances of life in their favor previous to fifty years of age than men have, but fewer afterwards. The number of marriages is in the proportion of seventy-five to every one thousand individuals.— Marriages are more frequent after equinoxes—that is, during the months of June and December. Those born in the spring are generally of a more robust constitution than others. Births are more frequent by night than by day, also deaths. The number of men capable of bearing arms is calculated at one-fourth of the population.

An Empire Success.
It has been proved by the most reliable testimony that Thomas' Electric Oil is an entire success in curing the most inveterate cases of rheumatism, neuralgia, lame back, and wounds of every description.
Paul G. Schuh Agent.

It is simply marvelous how quickly constipation, biliousness, sick headache, fever and ague, and malaria, are cured by "Seller's Liver Pills."

The invalid finds in "Dr. Lindsey's Blood Searcher" nature's great restorer. It is wonderful. Sold by all druggists.

A Grinning Death's Head
is scarcely more abhorrent to a refined observer, than a row of discolored teeth made visible by a smile. Correct the hideous blemish with delightful and healthful Sozodont, which whitens yellow teeth, imparts ruddiness and hardness to colorless, unhealthy gums, and a floral balminess to the breath. The feminine mouth becomes wondrously attractive in consequence of its use. Leading actresses and cantatrices regard it as incomparable.

THE Grand Central Hotel, 667 Broadway New York city, is one of the finest, if not

WHEAT STOCKS
\$10
\$20
\$50
\$100

INCREASE YOUR CAP
Investors of all amounts in Great Stocks as fully and extensively as possible. Our successful firm established in 1850. Sent weekly, dividends and profits. Dividends paid monthly or quarterly. Share. Address: MERRIAM, 141 & 143 LaSalle St., Chicago, Ill.
We want a local agent in every town. Excellent inducements. Good pay to a responsible, enterprising man. Write for terms.

FRANK TOOMEY,
AGENT FOR THE SALE OF THE GENUINE
BAXTER STEAM ENGINE
—Colt's Disc Engine—
Horizontal, Vertical and Marine Engines and Boilers.
YACHT ENGINES A SPECIALTY.
FARM ENGINES, MACHINISTS' TOOLS, STEAM AND MACHINERY OF ALL KINDS, BELTING, SHAFTING, Pulleys and General Supplies.
No. 131, North Third Street,



Closing the Loop Between ML Research & Library Systems

Ryan Cordell

School of Information Sciences
University of Illinois Urbana-Champaign

ryancordell.org | @ryancordell