COLLEGE OF INFORMATION STUDIES

Challenges in Providing Access to the Digital Universe: Are Algorithms the Answer?

Aeolian Network Workshop # 3

Jason R. Baron
Professor of the Practice
University of Maryland
College of Information Studies
28 January 2022

Some opening thoughts….

Results of algorithms can be ”unnerving, unfair, unsafe, unpredictable, and unaccountable.”*

Accountability and transparency are key

Growing recognition of need for XAI in recordkeeping & archives (J. Bunn)

Archives exist to preserve records and to provide access to them

Archivists have an ethical duty to protect the privacy of living individuals by shielding personal information in an archives from disclosure

Can AI assist in furthering the archival mission with these competing priorities?

What's the alternative to using AI?

* Andrew Selbst & Solon Barocas, “The Intuitive Appeal of Explainable Machines,” Fordham L. Rev (2018), 87:1085

Trusting the algorithm

When I am using an algorithm to search a collection for responsive materials, how do I know it will obtain accurate results?

How well do AI classifiers do in segregating sensitive material within collections?

Are there dangers lurking in terms of privacy violations despite the use of perfect classifiers or parsers?

`

Propositions:

(1)The experience lawyers have with using machine learning goes a long way to demonstrating that algorithms used for relevance and the filtering of sensitive content can be successfully applied to a large archival repository

(2) We have no alternative but to rely on algorithmic ways if we wish to carry out a central mission of allowing access to the world’s accumulated knowledge.

Transformation from an era of government archives housing hard copy records…

…to the world of data stored on massive servers

A great percentage of unstructured data in public electronic repositories & archives are not from online sources

2019 Managing Government Records Directive

https://www.archivepolicy/m-19-21-transition-to-federal-records.pdf
s.gov/files/records-mgmt/

NARA will only accept born digital or digitized records after December 31, 2022 (no more paper)

Agencies since 2017 have been required to manage their email electronically (all email, both temporary and permanent in nature)

US National Archives holdings

- “NARA keeps only those Federal records that are judged to have continuing value—about 2 to 5 percent of those generated in any given year. By now, they add up to a formidable number, diverse in form as well as in content. There are approximately 13.28 billion pages of textual records; 10 million maps, charts, and architectural and engineering drawings; 44.4 million still photographs, digital images, filmstrips, and graphics; 40 million aerial photographs; 563,000 reels of motion picture film; 992,000 video and sound recordings; and 1,323 terabytes of electronic data. All of these materials are preserved because they are important to the workings of Government, have long-term research worth, or provide information of value to citizens”.
- Source: About the National Archives of the US, https://www.archives.gov/publications/general-info-leaflets/1-about-archives.html#:~:text=There%20are%20approximately%2013.28%20billion,video%20and%20sound%20recordings%3B%20and

Archival Universe

Hypothesis: by 2050, 99% of the US National Archives will consist of digital records

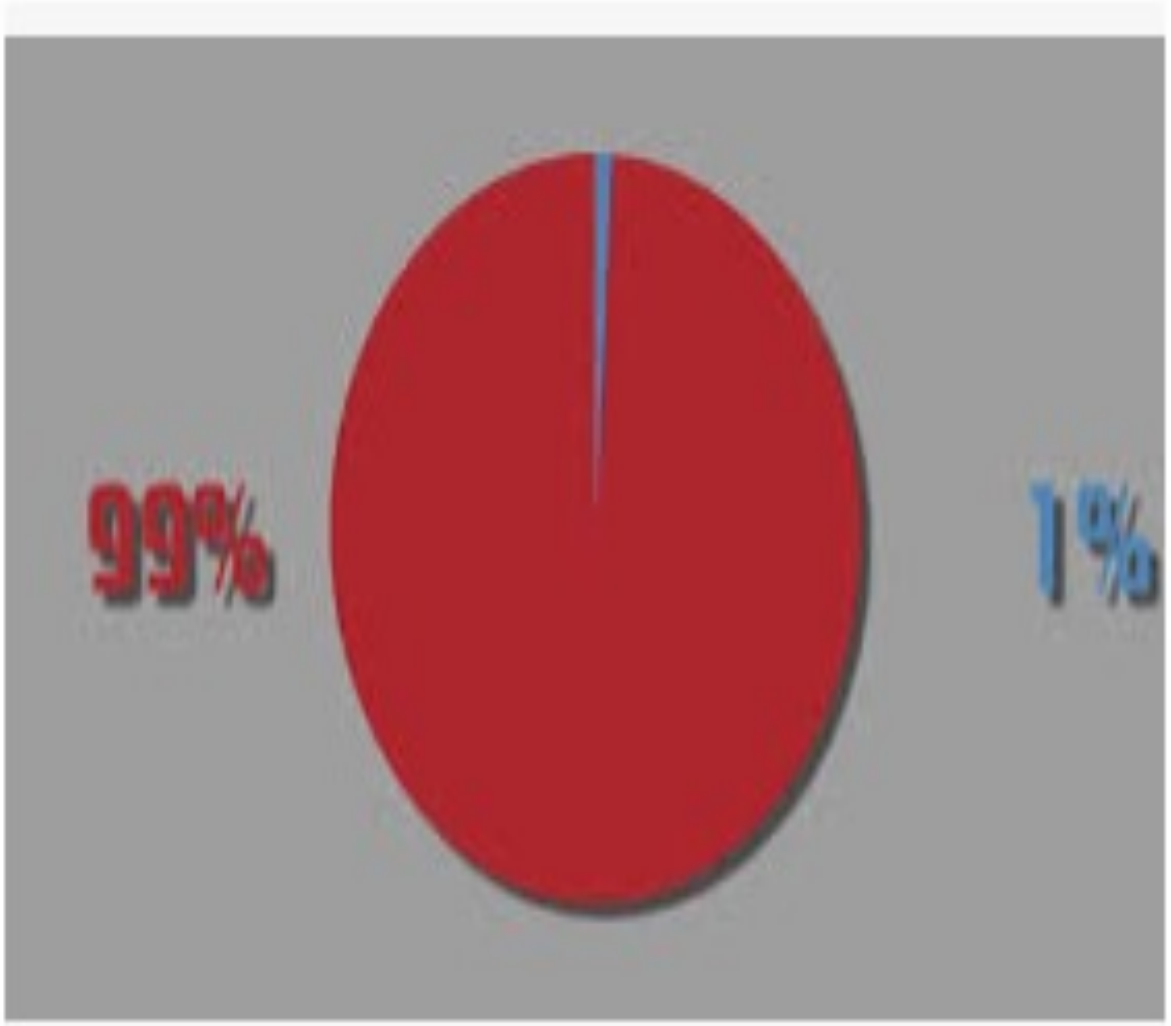
1 TB ~ 5 million pages (very conservative)
1,323 TB ~ 6.6 billion pages

13.28B pages (hard copy) is ~1% of

TB page equivalents:

6.6B pages (current) + (50B pages/yr x 28 yrs) (future) =1406.6B pages

And 99% of records are born digital, managed digitally, and soon they will be required to be transferred to archives in electronic or digital form



Archival Universe

If you don't like my number fill in your own.

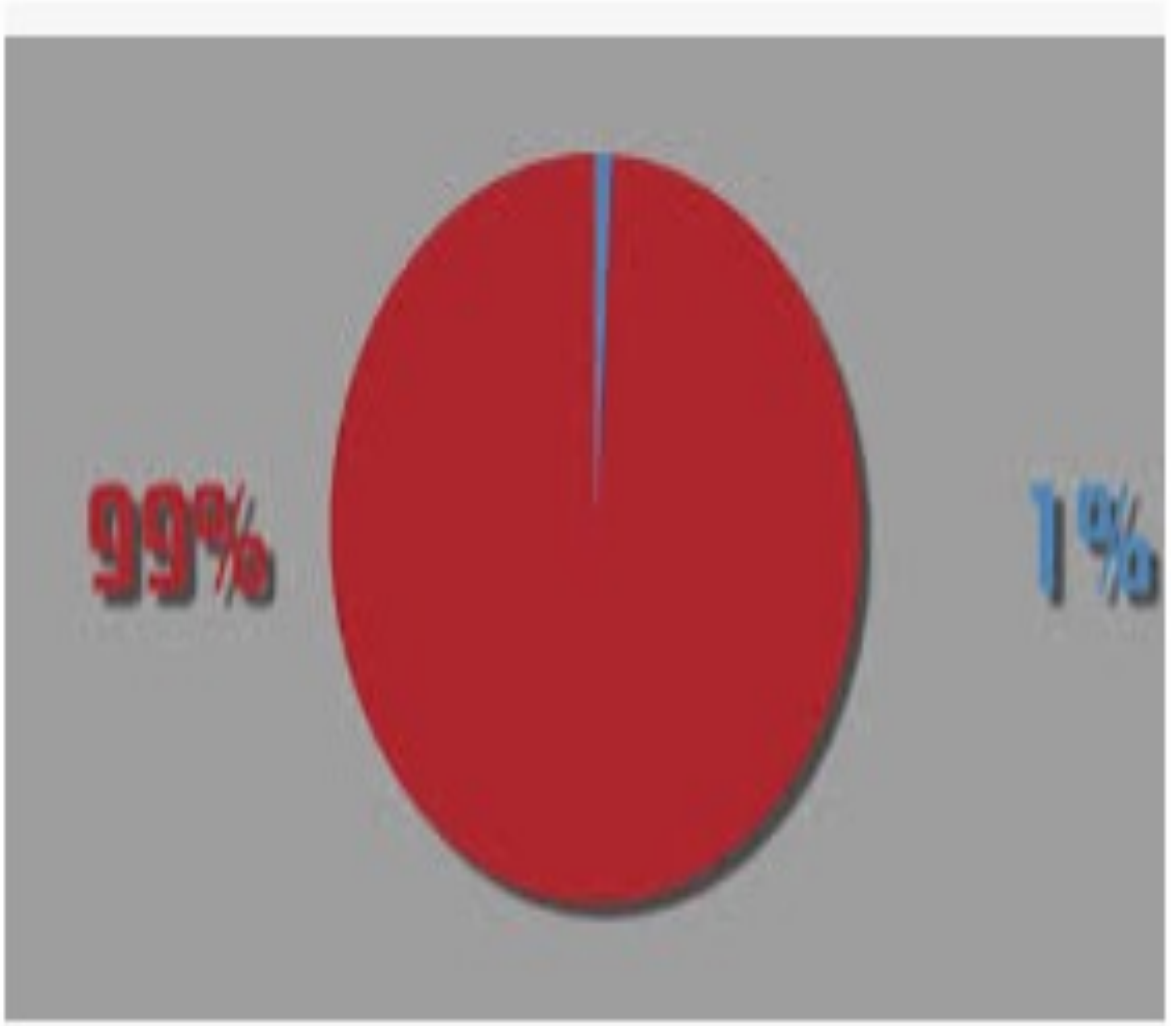
Hypothesis: by 2050, 99% of the US National Archives will consist of digital records

1 TB ~ 5 million pages (very conservative)
1,323 TB ~ 6.6 billion pages

13.28B pages (current pages in hard copy) is ~1% of

6.6B pages (current TB equivalent) + (50B pages/yr future estimate) x 28 yrs =1406.6B pages

99% of records are born digital, managed digitally, and soon they will be required to be transferred to archives in electronic or digital form



Looming Public Policy Issue: Failure to Provide Public Access to White House email

From Reagan to Obama
Over 500 million individual emails with attachments
(~2 billion pages)

WH email in NARA's legal custody that are open and available for public access*

• Reagan era Iran-Contra emails	5,000
• Supreme Court Justice-related emails	
• John Roberts-related emails	60,000
• Elana Kagan-related emails	75,000
• Brett Kavanaugh-related emails	170,000
• Other emails released in litigation or FOIA	~100,000
TOTAL RELEASED	~410,000
TOTAL ACCUMULATED WH EMAILS	~535,000,000 (thru Obama)
PERCENTAGE OPEN:	0.08%

* Note: figures represent distinct emails, not total pages. Page equivalents of emails & attachments > 2 billion (est.).

Capstone Email Policy

- Voluntary policy for US agencies to meet NARA 2016 mandate for management of email in electronic form
- Senior officials' emails deemed "permanent"; all other employees' emails retained for 7 years
- 200+ components of the federal government have elected to follow policy

Capstone email repositories

- Depending on the agency, senior official emails may already number in the hundreds of thousands or millions
- All other employees' emails may amount to millions, tens or hundreds of millions
- By 2030, billions of emails accumulated across entire Executive branch
- Capstone repositories in the future may be broadened to include other forms of electronic messaging, either by new legislation or simply by practice

Billions of digital e-mail records hiding in plain sight for for two reasons

Reason 1: A failure to date to use advanced search tools incorporating machine learning

- E-Discovery Timeline for Lawyers Recognizing/Using AI Methods
- TREC Legal Track 2016-2010
- Grossman & Cormack (2011)
- Da Silva Moore v Publicus (2012)
- Technology Assisted Review (TAR)

Reason 2: A need for archivists, records managers, and lawyers to filter sensitive content of all types from public records before access is granted

Categories of Personal Information (not exhaustive)

PII

- Names as metadata fields
- Social security numbers
- Telephone numbers
- Passport information
- Bank and financial information
- Credit card numbers
- Vehicle registrations
- Date of birth
- Height and Weight
- Asset information

Sensitive Personal Information

- Medical history
- Criminal history
- Sexual orientation
- Racial or ethnic origin
- Religious beliefs
- Political beliefs
- Mental health
- Genetic or biometric data

Spectrum of Difficulty (i.e., Trust)

Key to Unlocking Archives:
AI searches + Filtering for Sensitivities

Biases in Repositories of Unprecedented Scale Mean The Potential for Amplified Bias Due to AI

Archival silences refer to the erasure of archives, and histories of marginalized communities within traditional archival holdings. Institutional archives often collect histories of those in power and we must acknowledge that these holdings reflect the narratives of predominantly white creators. Assuming that collecting institutions are neutral does not benefit the groups who have been systematically left out of archival collections.

Ryerson University, “Archives and Special Collections”, https://learn.library.ryerson.ca/asc/silence-and-bias

Bias in the Capstone Universe of Email (negative legacy)

- Built into Capstone policies for email archiving is role-based appraisal
- Senior officials' communications deemed to be permanent
- But who are the "senior officials"?
  - Historically, a non-diverse population -- although in recent decades US Government agencies have been somewhat better in promoting persons of color into Senior Executive Service appointments
- Still, there is a bias in having a permanent collection of government email representing only what has bubbled up to the most senior officials, without an "in the trenches" view as well from less senior employees

Confirmation Bias in Searching

“The application of AI for search and retrieval has been critically discussed by historians. Problematising the use of full-text search within born-digital archives, Winters calls for new approaches from archival science and artificial intelligence that are more sensitive to archival hierarchy and context, avoiding the pitfall of only finding what is known (“confirmation bias”) in favour of seeking the unknown or surfacing gaps and absences in the data”.

Colavizza, G., et al., *Journal on Computing and Cultural Heritage*, 15:1 art. 4 (Dec 2021)
https://dl.acm.org/doi/10.1145/3479010

Keyword Searching & Confirmation Bias

- “[T]he tendency to search for, interpret, favor, and recall information in a way that confirms or supports one’s prior beliefs or values”.

  -- *Wikipedia, Confirmation Bias*

- Donald Rumsfeld: “unknown unknowns”
- Results in substantial over-inclusiveness (low precision) and under-inclusiveness (low recall)
  - Multiple words with single meaning (depressing recall %)
  - Single words with multiple meanings (depressing precision %)
- Hugely inefficient due to failure to rank order results

Judge Grimm writing for the U.S. District Court for the District of Maryland

"[W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI [electronically stored information], all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying on such searches for privilege review." ***Victor Stanley, Inc. v. Creative Pipe, Inc.,*** 250 F.R.D. 251 (D. Md. 2008); *see id., text accompanying nn. 9 & 10* (citing to TREC Legal Track research project)

TECHNOLOGY-ASSISTED REVIEW IN E-DISCOVERY CAN BE MORE EFFECTIVE AND MORE EFFICIENT THAN EXHAUSTIVE MANUAL REVIEW

By Maura R. Grossman* & Gordon V. Cormack† **

Cite as: Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

Defining "Technology Assisted Review" (TAR)

- A process for prioritizing or coding a collection of electronic documents using a computerized system that harnesses human judgments of one or more subject matter experts on a smaller set of documents and then extrapolates those judgments to the remaining document population.

- Also referred to as "supervised or active machine learning" or "computer-assisted review"

Source: Adapted from Grossman-Cormack Glossary of Technology Assisted Review, v. 1.0 (Oct 2012)

Manual Review vs TAR

- Impossibility of reviewing digital collections for responsive or sensitive content
- Manual review misses a substantial number of responsive or privileged documents
- TAR methods (with people in the loop)
  - find at least as many responsive or privileged documents as people do
  - make fewer errors
  - are vastly more efficient
- Source: M. Grossman & G. Cormack (Richmond J.)

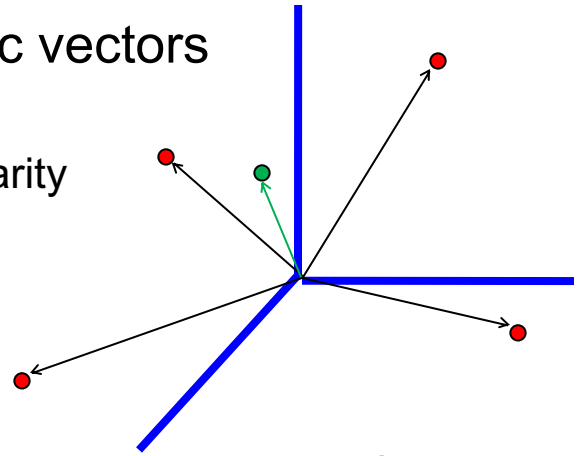
TAR’s advantages

- Identifies patterns in relationships between terms and concepts
  - Through supervised learning the algorithm works with training data to map unseen examples
- Principle that words used in same contexts tend to have similar meanings
- Technique obtains high levels of recall
- Rank orders documents by relevance
- Vastly reduces time and resources on task
- Most recently, ”continuous active learning” (TAR 2.0) reduces training set to bare minimum (arguably reducing training bias).

Source: M. Grossman & G. Cormack (Richmond J.)

Vector Space Model

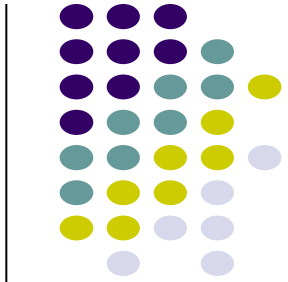
- The Model
  - documents represented as vectors in N-dimensional space where N is number of ‘terms’ in the document set
    - term is usually a word (stem); but might be phrase or thesaurus class
    - terms are weighted based on frequency and distribution of occurrences
  - information need is natural language text mapped in same space
  - matching is similarity between query and doc vectors
    - example similarity: cosine of angle between vectors
    - allows documents to be ranked by decreasing similarity



- Pros and Cons
  - good: less brittle than pure Boolean
  - bad: less transparency---depending on weights, a doc with few query terms can be ranked higher than a doc with many

From Ellen Voorhees, Georgetown 2009

Vector Similarities



- Document-Document similarity
  - docs are similar to the extent they contain the same terms
  - doc pairs with maximal similarity detects duplicates
  - document clustering
    - *cluster hypothesis*: "Closely associated documents tend to be relevant to the same requests."
    - thus, do retrieval based on returning whole clusters since usually much more information in doc-doc comparison than doc-query
- Term-Term similarity
  - terms are similar to the extent the occur in the same documents
  - term clustering
    - query expansion
    - provide bottom-up description of document set

	T1	T2	T3	T4	…
D1	5	0	33	0	…
D2	0	0	8	0	…
D3	1	4	0	2	…
D4	0	3	0	4	…
D5	0	1	0	0	…
D6	5	3	2	0	…
…					

From Ellen Voorhees, Georgetown 2009

Judicial endorsement of predictive analytics in document review by Judge Peck in *da Silva Moore v. Publicis Groupe* (SDNY 24 Feb. 2012)

- This opinion appears to be the first in which a Court has approved of the use of computer-assisted review. . . . What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review . . . Computer-assisted review can now be considered judicially-approved for use in appropriate cases.

Judge Peck's *da Silva Moore* decision

Judicial acceptance of results of searches without the need for an evidentiary hearing

So long as there are robust results, *reasonableness* standard met

Some disagreement in the profession but Judge Peck's rule has been adopted as the norm – few cases exist resulting in strong challenges to how ML/TAR methods conducted

Computer Science > Information Retrieval

[Submitted on 14 Nov 2020]

Providing More Efficient Access To Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege

Jason R. Baron, Mahmoud F. Sayed, Douglas W. Oard

Test Collection:
Records of White House Assistant to the President for Domestic Policy Elena Kagan, from the Clinton Presidential Library online database

Using a Classifier to Segregate Out Sensitive Content

- We show that when classifiers are trained and tested under consistent conditions it is possible to design classifiers that achieve F_1 measures between 70% and 83% (i.e., if tuned so that precision and recall were equal, we would expect that between 70% and 83% of the exempt material would be found, and that the same fraction of the content identified by the classifier as exempt would truly be exempt).
- We study the effects of differences between reviewers, between the materials held by different custodians, and within the topical content of the records being classified to identify which differences pose the greatest challenge for current text classifiers.
- We control for the effect of document type and recognizable characteristics of content items to study classifier effectiveness on the content and document types that human reviewers find most difficult.
- We suggest directions for future work, identifying a need to model contextual factors that require access to evidence beyond the boundaries of specific documents.
- We introduce a new freely distributable test collection that is annotated for the deliberative process privilege under exemption 5 of the FOIA.[11]

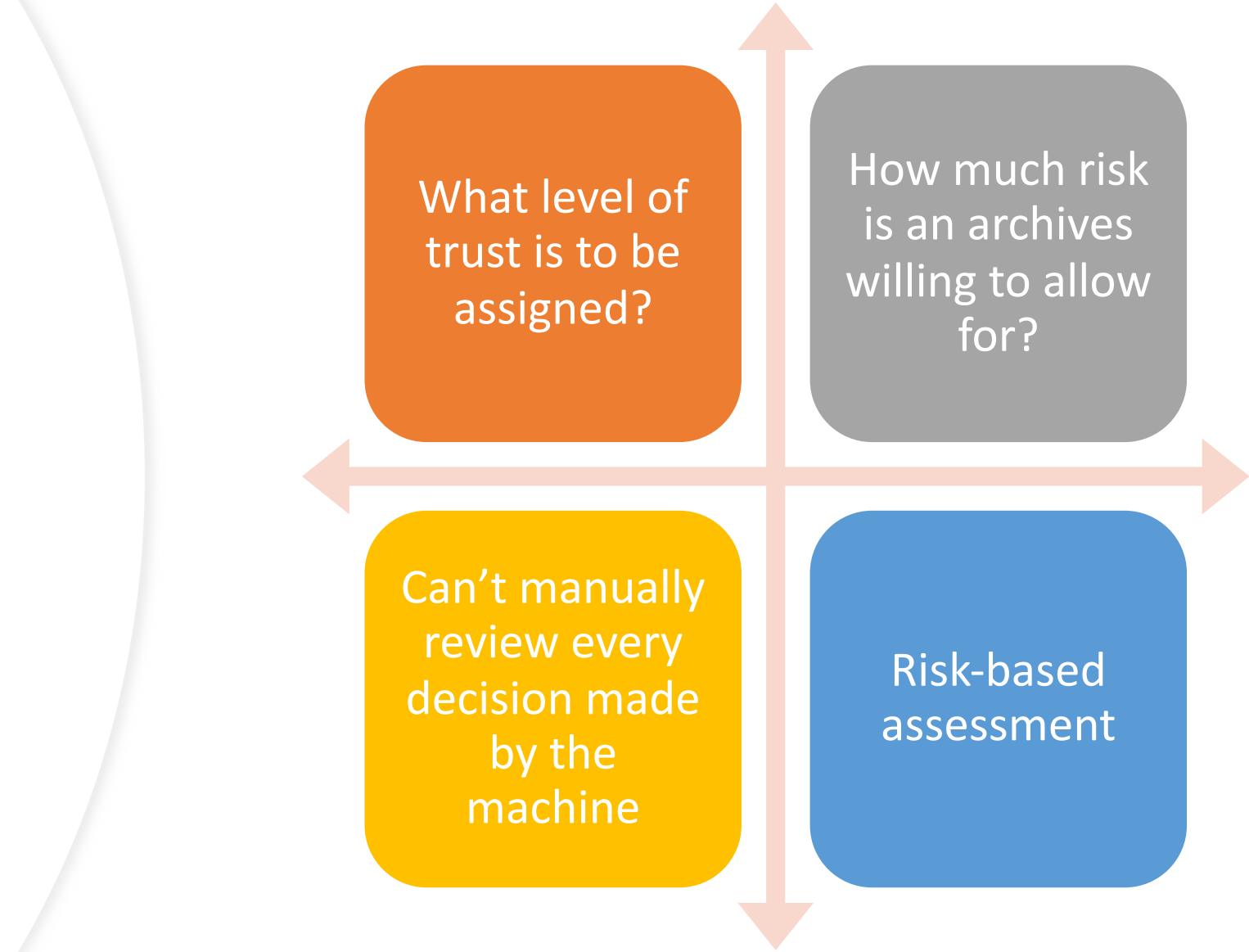
J.R. Baron, M. Sayed & D. Oard, Journal on Computing and Cultural Heritage, 15:1, article 5: 1-19 (2022), https://dl.acm.org/doi/abs/10.1145/3481045 (pre-print at https://arxiv.org/abs/2011.07203)

Using a Classifier to Segregate Out Sensitive Content

- We show that when classifiers are trained and tested under consistent conditions it is possible to design classifiers that achieve F_1 measures between 70% and 83% (i.e., if tuned so that precision and recall were equal, we would expect that between 70% and 83% of the exempt material would be found, and that the same fraction of the content identified by the classifier as exempt would truly be exempt).
- We study the effects of differences between reviewers, between the materials held by different custodians, and within the topical content of the records being classified to identify which differences pose the greatest challenge for current text classifiers.
- We control for the effect of document type and recognizable characteristics of content items to study classifier effectiveness on the content and document types that human reviewers find most difficult.
- We suggest directions for future work, identifying a need to model contextual factors that require access to evidence beyond the boundaries of specific documents.
- We introduce a new freely distributable test collection that is annotated for the deliberative process privilege under exemption 5 of the FOIA.[11]

J.R. Baron, M. Sayed & D. Oard, Journal on Computing and Cultural Heritage, 15:1, article 5: 1-19 (2022), https://dl.acm.org/doi/abs/10.1145/3481045 (pre-print at https://arxiv.org/abs/2011.07203)

After the AI search…. What level of human review should be undertaken?

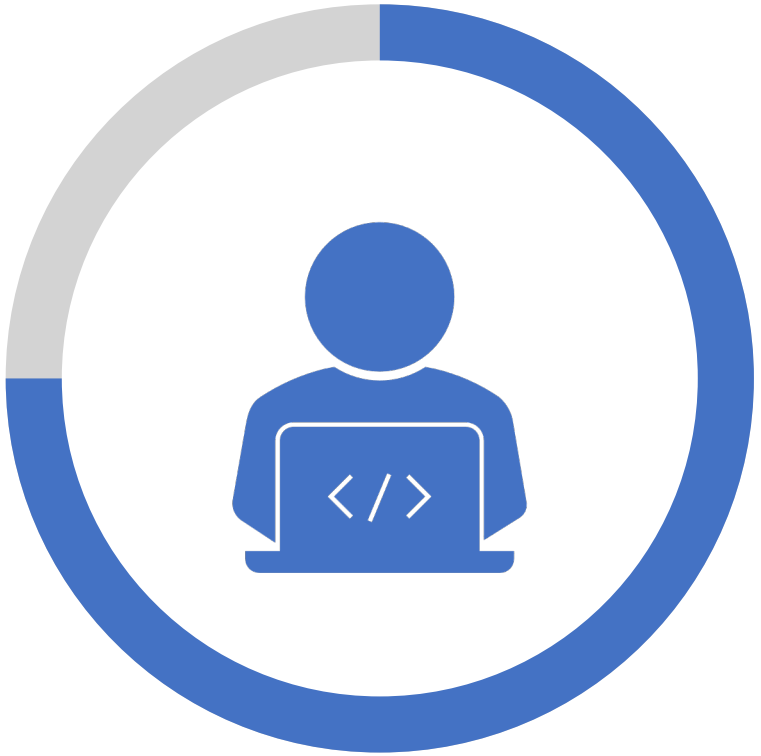


De-Anonymization

- If I have applied privacy filters of some kind (e.g., redactions), what kind of AI methods exist to nevertheless de-anonymize privacy related material in collections?

Can we trust the algorithm?

- Of course algorithms have biases based on their inputs.
- Of course algorithms do not do a perfect job on information retrieval tasks (in e-discovery over 75% recall is considered great)
- Of course algorithms contain difficult to explain features
- But . . .



The Right Questions To Ask:

- First, can we trust the algorithm with humans in the loop more than we trust humans alone to perform the same information retrieval tasks?
- Second, can humans alone even perform the same tasks?
- Based on my experience, the answer to the first question is yes.
- And to the second question, no.

Dark Archives

Dark Archives

- Sensitive content consisting of personal information closed between 100 and 110 years (less the age of the individual, if known). If the age of the individual is not known, for minors it is closed for the whole period, and for those deemed to be over the age of sixteen, 80 or 94 years.
- Source: Moss, Michael S. and Gollins, Tim J. (2017) "Our Digital Legacy: an Archival Perspective," Journal of Contemporary Archival Studies", vol.4, art. 3, http://elischolar.library.yale.edu/jcas/vol4/iss2/3
- Rule at the US National Archives is a presumption of closure for 75 years from date of creation of the document

“More Product, Less Process”

Mark A. Greene and Dennis Meissner, “More Product Less Process: Revamping Traditional Archival Processing,” *American Archivist* 68:2 (2005)

- Waiting for every individual named in records to pass away before opening vast amounts of records is a strategy, but in my view not a very good one.
- We should continue to promote ML/TAR methods that efficiently find responsive records and accurately segregate personal content as best we can
- We should adopt a risk-based model for tolerating errors (inadvertent release of personal info / PII).

Duty of Technological Competence

US lawyers must understand the benefits and risks associated with technology.

American Bar Association Rule 1.1, Comment 8

Lawyers have an affirmative duty (1) to be proficient in the technology they use in the representation of a client; and (2) to consider technology that may improve the professional services the lawyer provides to his or her clients.

Should A Similar Rule Apply to Archival Professionals Handling Algorithms?

UNDERSTANDING TECHNOLOGY TO PROVIDE FOR BETTER ACCESS & FOR FILTERING FOR SENSITIVE INFORMATION

UNDERSTANDING AI BIAS

Letting go of a reluctance to embrace the AI world

Impenetrable

…but think of the black box as a “gift” to archivists to improve access to vast digital collections.

COLLEGE OF
INFORMATION
STUDIES

Jason R. Baron
University of Maryland
College of Information Studies
jrbaron@umd.edu

References

- Office of Management and Budget & US National Archives, M-19-21, “Transition to Electronic Records” (2019), https://www.archives.gov/files/records-mgmt/policy/m-19-21-transition-to-federal-records.pdf
- US National Archives, “White Paper on the Capstone Approach and Capstone GRS” (2015), https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf
- TREC Legal Track, https://trec-legal.umiacs.umd.edu
- M. Grossman & G. Cormack,”Technology-Assisted Revew in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review,” *Richmond J. of Law and Technology* 17:3 (2011), https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1344&context=jolt

References (con’t)

- Da Silva Moore v. Publicus Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012), https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2011cv01279/375665/96/
- N. Pace & L. Zakaras, “Where the Money Goes: Understanding Litigant Expenditures for Electronic Discovery, RAND Corporation (2012), https://www.rand.org/pubs/monographs/MG1208.html
- J.R. Baron, M. Sayed, & D. Oard, “Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege”, *Journal on Computing & Cultural Heritage,* 15:1, article 5:1-19 (2022), https://dl.acm.org/doi/abs/10.1145/3481045 (pre-print at https://arxiv.org/abs/2011.07203)