



AUTHORS ALLIANCE

The mission of Authors Alliance is to advance the interests of authors who want to serve the public good by sharing their creations broadly.

Education and Advocacy on

Fair Use | Free Expression | Research | Open Access and Open Data
Platform Integrity | Publisher and Tech Competition



authorsalliance.org/join



Text and Data Mining: Demonstrating Fair Use

- Mellon Foundation supported project
- Workshop series
- Report on text data mining research and legal barriers
- Want to talk? Reach out to us at info@authorsalliance.org



Road Map

- Text and data mining overview
- Copyright and licenses
- DMCA exemption for text data mining
- Key takeaways and open questions

text and data mining





Automated analytical techniques aimed at analyzing digital text and data in order to generate information that reveals patterns, trends, and correlations in that text or data.



But why?

"The discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise 'hidden' meanings." - Marti Hearst

Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks

ROBERT HOFFMANN, MARTIN KRALLINGER, EDUARDO ANDRES, JAVIER TAMAMES, CHRISTIAN BLASCHKE, AND ALFONSO VALENCIA [Authors Info & Affiliations](#)

SCIENCE'S STKE • 10 May 2005 • Vol 2005, Issue 283 • p. pe21 • DOI: 10.1126/stke.2832005pe21

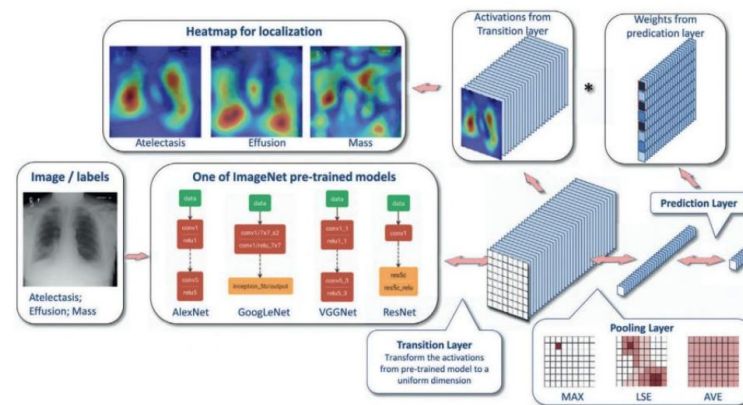


Figure 5.3 The overall flow-chart of our unified DCNN framework and disease localization process.

Yifan Peng , Zizhao Zhang , Xiaosong Wang , Lin Yang , & Le Lu , Chapter 5 - Text mining and deep learning for disease classification, Handbook of Medical Image Computing and Computer Assisted Intervention (Elsevier, 2019), <https://doi.org/10.1016/B978-0-12-816176-0.00010-7>



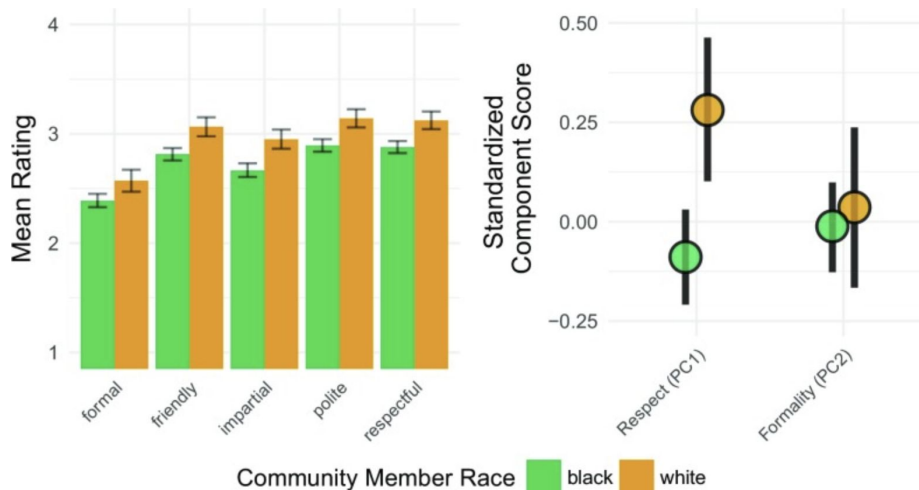
PNAS

Language from police body camera footage shows racial disparities in officer respect

Rob Voigt^{a,1}, Nicholas P. Camp^b, Vinodkumar Prabhakaran^c, William L. Hamilton^c, Rebecca C. Hetey^b, Camilla M. Griffiths^b, David Jurgens^c, Dan Jurafsky^{a,c}, and Jennifer L. Eberhardt^{b,1}

^aDepartment of Linguistics, Stanford University, Stanford, CA 94305; ^bDepartment of Psychology, Stanford University, Stanford, CA 94305; and ^cDepartment of Computer Science, Stanford University, Stanford, CA 94305

Contributed by Jennifer L. Eberhardt, March 26, 2017 (sent for review February 14, 2017; reviewed by James Pennebaker and Tom Tyler)



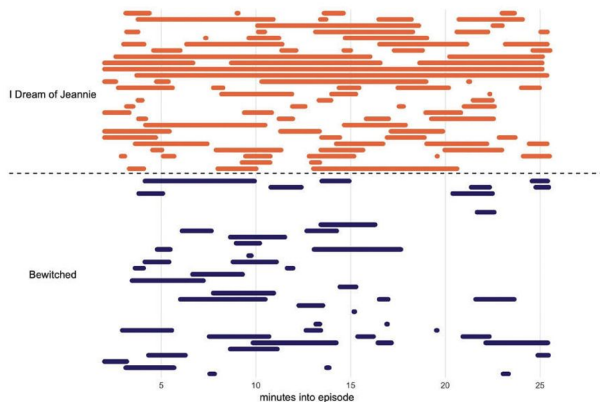


Fig. 2 A plot showing frames where the main female character had *not* been seen for at least 2 min. Each row depicts a particular episode, roughly 25 min in length, from the 1966 to 1967 seasons of *Bewitched* and *I Dream of Jeannie*. Faces were detected using the DVT (Arnold and Tilton, 2017)

Taylor Arnold & Lauren Tilton, Distant Viewing: Analyzing Large Visual Corpora (2019), <https://distantviewing.org/papers/2019-distant-viewing.pdf>



Fig. 3 The left column of this grid of photographs shows images selected from the FSA-OWI archive, a collection of documentary photography taken by the United States Government between 1935 and 1943. To the right of each image are the seven closest other images in the collection using the distant metric induced by the penultimate layers of the InceptionV3 neural network model (Szegedy *et al.*, 2015). Notice that each row detects images with a similar dominant object: horses, wooden houses, pianos, train cars, and cooking pots



Journal of Cultural Analytics

Underwood, Ted, David Bamman, and Sabrina Lee. 2018. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics* 3 (2). <https://doi.org/10.22148/16.019>.

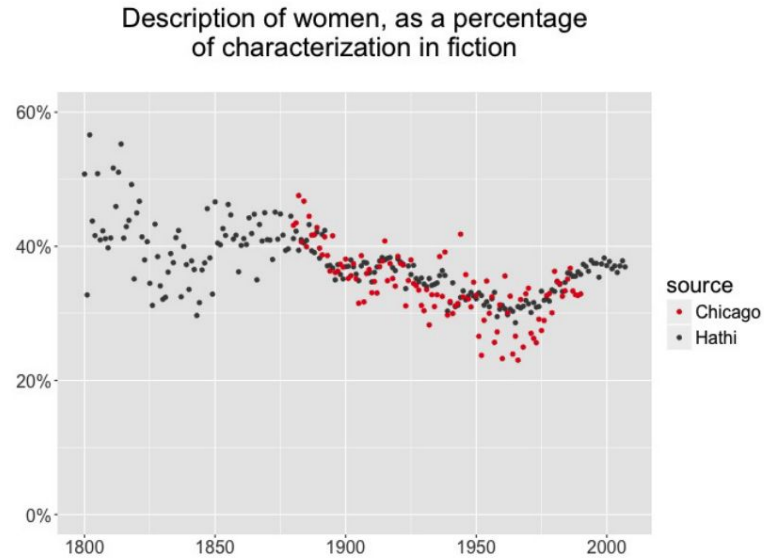


Figure 1. The percentage of words used in characterization that describe women.



prosecraft

linguistics for literature

(about 107 pages)

26,814

TOTAL WORDS

4th PERCENTILE
of all the books in our library

83.94%

VIVIDNESS

53rd PERCENTILE
of all the books in our library

8.08%

PASSIVE VOICE

51st PERCENTILE
of all the books in our library

4.53%

ALL ADVERBS

99th PERCENTILE
of all the books in our library

1.31%

LY-ADVERBS

80th PERCENTILE
of all the books in our library

3.22%

NON-LY-ADVERBS

99th PERCENTILE
of all the books in our library



arXiv:2101.00027v1 [cs.CL] 31 Dec 2020

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao	Stella Biderman	Sid Black	Laurence Golding
Travis Hoppe	Charles Foster	Jason Phang	Horace He
Anish Thite	Noa Nabeshima	Shawn Presser	Connor Leahy

EleutherAI
contact@eleuther.ai

Abstract

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present *the Pile*: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.¹

1 Introduction

Recent breakthroughs in general-purpose language modeling have demonstrated the effectiveness of training massive models on large text corpora for downstream applications (Radford et al., 2019; Shoybi et al., 2019; Raffel et al., 2019; Rosset, 2019; Brown et al., 2020; Lepikhin et al., 2020). As the field continues to scale up language model training, the demand for high-quality massive text data will continue to grow (Kaplan et al., 2020).

The growing need for data in language modeling has caused most existing large-scale language models to turn to the Common Crawl for most or all of their data (Brown et al., 2020; Raffel et al., 2019). While training on the Common Crawl has been effective, recent work has shown that dataset di-

versity leads to better downstream generalization capability (Rosset, 2019). Additionally, large-scale language models have been shown to effectively acquire knowledge in a novel domain with only relatively small amounts of training data from that domain (Rosset, 2019; Brown et al., 2020; Carlini et al., 2020). These results suggest that by mixing together a large number of smaller, high quality, diverse datasets, we can improve the general cross-domain knowledge and downstream generalization capabilities of the model compared to models trained on only a handful of data sources.

To address this need, we introduce the Pile: a 825.18 GiB English text dataset designed for training large scale language models. The Pile is composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones. In addition to its utility in training large language models, the Pile can also serve as a broad-coverage benchmark for cross-domain knowledge and generalization ability of language models.

We introduce new datasets derived from the following sources: PubMed Central, ArXiv, GitHub, the FreeLaw Project, Stack Exchange, the US Patent and Trademark Office, PubMed, Ubuntu IRC, HackerNews, YouTube, PhilPapers, and NIH ExPorter. We also introduce OpenWebText2 and BookCorpus2, which are extensions of the original OpenWebText (Gokaslan and Cohen, 2019) and BookCorpus (Zhu et al., 2015; Kobayashi, 2018) datasets, respectively.

In addition, we incorporate several existing high-quality datasets: Books3 (Presser, 2020), Project Gutenberg (PG-19) (Rae et al., 2019), OpenSubtitles (Tiedemann, 2016), English Wikipedia, DM Mathematics (Saxton et al., 2019), EuroParl (Koehn, 2005), and the Enron Emails corpus (Kliant and Yang, 2004). To supplement these, we also in-

¹<https://pile.eleuther.ai/>



**is it legal?
copyright and licenses**



copyright protects original creative expression

17 U.S.C. § 102(a)

Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device

Art. I, Sec. 8: To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries

copyright's “exclusive rights”

17 U.S.C. § 106

The owner of a copyright . . . has the exclusive right to do and to authorize any of the following:

- To reproduce the copyrighted work;
- To prepare derivative works;
- To distribute copies of the work;
- To perform the work publicly;
- To display the work publicly.





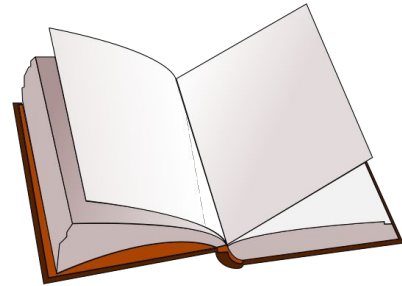
copyright does not protect facts, ideas or concepts

17 U.S.C. § 102(b)

in no case does copyright protection for an original work of authorship extend to any idea, procedure, **process**, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.



copyrighted works can
have unprotected facts,
ideas embedded in them



copyright can limit text data mining



- TDM research can implicate exclusive rights and Digital Millennium Copyright Act rules
- Permissible categories of works:
 - Public domain works
 - Licensed collections for TDM
 - Works under copyright, under a new exemption

fair use

17 U.S.C. § 107

Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, **teaching (including multiple copies for classroom use), scholarship, or research**, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

- (1)**the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
- (2)**the nature of the copyrighted work;
- (3)**the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- (4)**the effect of the use upon the potential market for or value of the copyrighted work.





1. purpose of the use
2. nature of the original
3. amount and substantiality
4. market effect



Google
Books



HATHI
TRUST

Authors Guild, Inc. v. Google, 721 F.3d 132 (2d Cir. 2015)

Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014)

13-4829-cv

UNITED STATES COURT OF APPEALS
FOR THE SECOND CIRCUIT

THE AUTHORS GUILD, BETTY MILES, JIM BOUTON, JOSEPH GOULDEN,
individually and on behalf of all others similarly situated
Plaintiffs-Appellants

HERBERT MITGANG, DANIEL HOFFMAN, individually and on behalf of all
other similarly situated, PAUL DICKSON, THE MCGRAW-HILL COMPANIES,
INC., PEARSON EDUCATION, INC., SIMON & SHUSTER, INC.,
ASSOCIATION OF AMERICAN PUBLISHERS, INC. CANADIAN
STANDARD ASSOCIATION, JOHN WILEY & SONS, INC., individually and
on behalf of all others similarly situated.
Plaintiffs

v.

GOOGLE, INC.

Defendant-Appellee

On Appeal from the United States District Court for the Southern District of New
York

**BRIEF OF DIGITAL HUMANITIES AND LAW SCHOLARS
AS *AMICI CURIAE* IN SUPPORT OF DEFENDANT-APPELLEES**

Jason M. Schultz*
Associate Professor of Clinical Law
NYU School of Law
245 Sullivan Street
New York, NY 10012
(212) 992-7365

Counsel for Amici Curiae

On the Brief:
Matthew Sag*
Professor
Loyola University of Chicago
School of Law

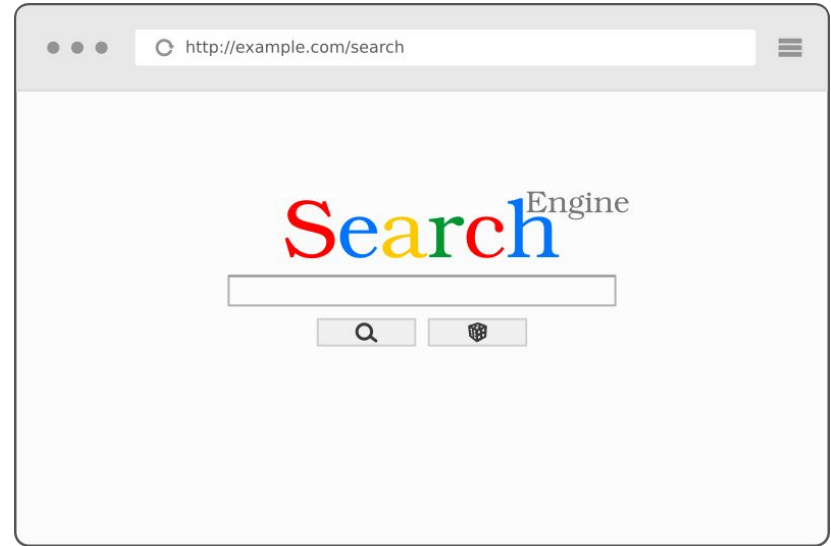
* Filed in their individual capacity and not on behalf of their institutions



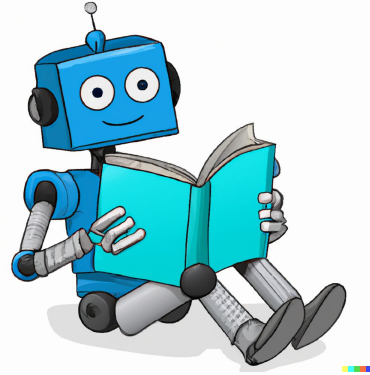
So what did the court say?

- purpose of the use
 - highly transformative; new purpose
- nature of the original
 - mixed but not very important
- amount and substantiality
 - appropriate given purpose
- market effect
 - not competing substitutes
 - did care about security so as to not affect original market





Copyright Meets Generative AI



- October 2015: *Google Books* decision
- November 2022: OpenAI introduces ChatGPT
- November 2022: First major lawsuit against an AI company (OpenAI, Github, and Microsoft)
- February 2023: Copyright Office issues opinion letter on registration in AI-generated works
- March 2023: Copyright Office launches AI initiative
- February 2024: More than a dozen copyright lawsuits filed



Three big questions

1. Are generative AI outputs protected by copyright?
2. Is it permissible to use copyrighted works for use as training data?
3. Are AI outputs infringing?





publishing text data mining research

- Like text data mining itself, publishing TDM research is generally a fair use (scholarship)
 - But: consider what you are reproducing from underlying works
- But if TDM research does not comply with copyright laws or licenses, there can be consequences for published works
 - Ex) In 2021, a paper on COVID-19 vaccine hesitancy was retracted because TDM researchers breached terms of service



licenses can limit text data mining

- Licenses can limit a researcher's ability to conduct text and data mining research
 - By forbidding text and data mining
 - By forbidding an activity necessary to conduct text data mining
- Licensed collections for text and data mining come with their own set of limitations

Amazon Kindle

Limitations. You may not remove or modify any proprietary notices or labels on the Kindle Content. In addition, you may not **attempt to bypass, modify, defeat, or otherwise circumvent any digital rights management system or other content protection or features used as part of the Service.**



Sample Insert Language for Stand-Alone TDM License

GRANT OF LICENSE: Licensee and Authorized Users may conduct TDM on the Licensed Materials for non-profit scholarly, research, or educational purposes. Licensee and Authorized Users may utilize and share the TDM Outputs, or the analysis or derived data from conducting TDM, in their scholarly work and make such TDM Outputs, analysis, or results available for use by others, except to the extent that doing so would substantially reproduce or redistribute the original Licensed Materials for third parties, or create a product for use by third parties that would substitute for the Licensed Materials.



digital locks



technical protection measures (TPMs)



- Watermarks
- Digital Rights Management (DRM)
- Content Scramble
- Advanced Access Content System

Digital Millennium Copyright Act



A) **No person shall circumvent a technological measure** that effectively controls access to a work protected under this title.

B) **The prohibition** contained in subparagraph (A) **shall not apply to persons** who are users of a copyrighted work which is in a particular class of works, if such persons are, or are likely to be in the succeeding 3-year period, **adversely affected** by virtue of such prohibition **in their ability to make noninfringing uses** of that particular class of works under this title, as determined under subparagraph (C).

C) . . . [*a complex process through which the U.S. Copyright Office will issue regulations allowing users to circumvent technical protection measures.*]



UNITED STATES COPYRIGHT OFFICE



Petition for New Exemption Under 17 U.S.C. § 1201

8th Triennial Rulemaking

Please submit a separate petition for each proposed exemption.

NOTE: Use this form if you are seeking to engage in activities not currently permitted by an existing exemption. If you are seeking to engage in activities that are permitted by a current exemption, instead of submitting this form, you may submit a petition to renew that exemption using the form available at <https://www.copyright.gov/1201/2021/renewal-petition.pdf>.

If you are seeking to expand a current exemption, we recommend that you submit both a petition to renew the current exemption, and, separately, a petition for a new exemption using this form that identifies the current exemption, and addresses only those issues relevant to the proposed expansion of that exemption.

ITEM A. PETITIONERS AND CONTACT INFORMATION

Please identify the petitioners and provide a means to contact the petitioners and/or their representatives, if any. The "petitioner" is the individual or entity proposing the exemption.

(1) Authors Alliance
Brianna Schofield, Executive Director
brianna@authorsalliance.org

Represented by:
Samuelson Law, Technology & Public Policy Clinic
UC Berkeley, School of Law
Catherine Crump, Director
Gabrielle Daley, Clinical Teaching Fellow
Jason Francis and Alistair McIntyre, Clinical Law Students
[crrump@clinical.law.berkeley.edu](mailto:crcrump@clinical.law.berkeley.edu)

(2) American Association of University Professors
Risa Lieberwitz, AAUP General Counsel, rlieberwitz@aaup.org
Aaron Nisenson, AAUP Senior Counsel, anisenson@aaup.org
Nancy Long, AAUP Associate Counsel, nlong@aaup.org

(3) Library Copyright Alliance

Represented by:
Jonathan Band
policybandwidth
jband@policybandwidth.com

37 CFR 201.40(b)(5)

(i) Literary works, excluding computer programs and compilations that were compiled specifically for text and data mining purposes, distributed electronically where:

(A) The circumvention is undertaken by a researcher affiliated with a nonprofit institution of higher education, or by a student or information technology staff member of the institution at the direction of such researcher, solely to deploy text and data mining techniques on a corpus of literary works for the purpose of scholarly research and teaching;

(B) The copy of each literary work is lawfully acquired and owned by the institution, or licensed to the institution without a time limitation on access;

(C) The person undertaking the circumvention views the [contents](#) of the literary works in the corpus solely for the purpose of verification of the research findings; and

(D) The institution uses effective security measures to prevent further dissemination or downloading of literary works in the corpus, and to limit access to only the persons identified in [paragraph \(b\)\(5\)\(i\)\(A\)](#) of this section or to researchers or to researchers affiliated with other institutions of higher education solely for purposes of collaboration or replication of the research.

(ii) For purposes of [paragraph \(b\)\(5\)\(i\)](#) of this section:

(A) An institution of higher education is defined as one that:

(1) Admits regular students who have a certificate of graduation from a secondary school or the equivalent of such a certificate;

(2) Is legally authorized to provide a postsecondary education program;

(3) Awards a bachelor's degree or provides not less than a two-year program acceptable towards such a degree;

(4) Is a public or other nonprofit institution; and

(5) Is accredited by a nationally recognized accrediting agency or association.

(B) The term "effective security measures" means security measures that have been agreed to by interested copyright owners of literary works and institutions of higher education; or, in the absence of such measures, those measures that the institution uses to keep its own highly confidential information secure. If the institution uses the security measures it uses to protect its own highly confidential information, it must, upon a reasonable request from a copyright owner whose work is contained in the corpus, provide information to that copyright owner regarding the nature of such measures.



Who can use this exemption?

- Researcher affiliated with a nonprofit institution of higher education*
- Student or staff member working at direction of such researcher
- Researchers at other institutions, for purposes of collaboration or verification only



What materials can you use?

- Motion pictures on protected DVDs, BluRay, digital download
- Literary works distributed electronically, but NOT
 - Computer programs
 - Compilations specifically made for TDM
- Copies must be owned by the college/university or licensed without time limit



Where and when?

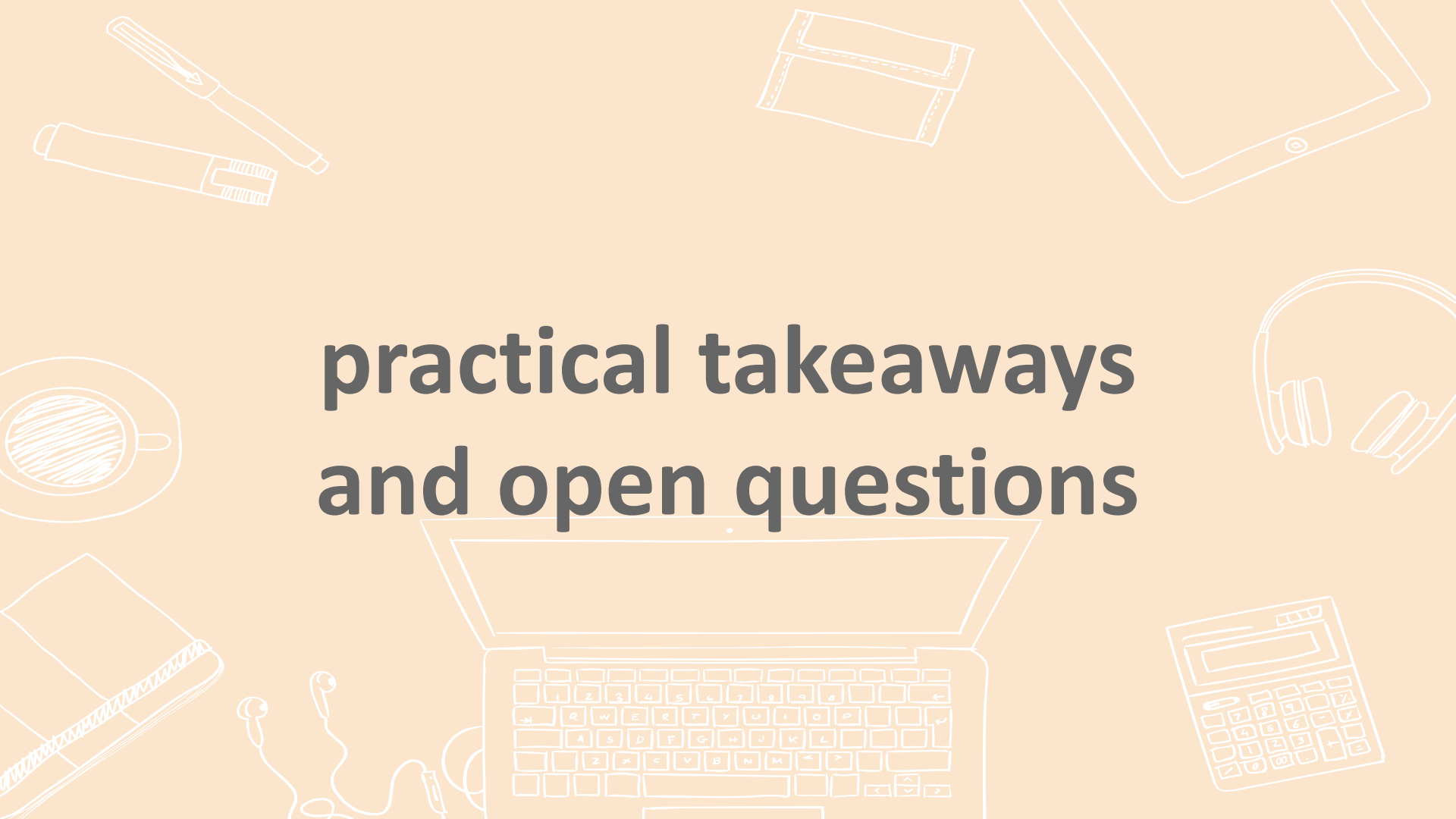
- Access only with “effective security measures” to prevent further downloading
 - A jointly agreed standard between the institution and the rightsholder (hasn’t happened) OR
 - Internal standard to protect “highly confidential information”



Update: Renewing and Expanding the Exemption in 2024

- Streamlined renewal process
- Renewal and expansion petitions submitted in September
- Reply (opposition) comments due Feb 20
- Final decision likely fall 2024





practical takeaways and open questions

copyright and licenses

- Not everything is protected by copyright (but even unprotected materials can be protected by DRM)
- You can get permission in lots of cases
- Fair use comes into play when you don't have permission
- TDM for academic research has a strong basis in the law





copyright and licenses

- What your outputs look like will be important:
 - are you reproducing extensive expressive text (protected by copyright)?
 - Or are you just providing information about expressive text (not protected by copyright)?

DMCA and digital locks

- Most researchers doing TDM work within universities are covered by the exemption
- Only applies to motion pictures and literary works (but not computer programs or compilations made for TDM use)
- Security requirements, especially for research data sharing, are potentially complicated





open questions

- What about other types of works such as music, visual works, video games, streaming services?
- Does fair use protect TDM using a corpus that wasn't legal when created (e.g., SciHub?)
- How do we navigate licenses?



Question for discussion

Imagine you are interested in learning about depictions of philosophers in modern, popular culture. You have a \$100,000 grant to investigate.

- what materials would you be interested in looking at?
- what are the barriers to building a corpus?
- how does the law influence what you do?



Questions?



Scan to join!



Question for discussion

- Think about your own potential research questions
 - What is hard about building or using an existing corpus?
 - How do legal or licensing limitations limit the scope of your research?
- How, at Brown, would you navigate the security and collaboration restrictions?