Introduction

- Attribution goes to the Building Legal Literacies for Text Data Mining Institute and the faculty
    - All information contained herein is taken from faculty lectures, transcripts, and slides; and from discussions amongst the faculty and participants
- An OER is forthcoming
- This does not cover international copyright or technological protection measures
- This is not legal advice and I am not a lawyer
- See UC Berkeley's Responsible Access Workflows for an excellent summation of copyright, licensing, privacy, and ethics considerations when undertaking TDM research

Copyright

Copyright

Copyright law is part of a legal system that covers both creation and use. It is filled with exceptions and exemptions to strike a balance between the exclusive rights granted to creators and the rights of users, including TDM researchers.

Copyright is

- A "bundle of rights"
- Limited economic monopoly for authors
- A system "to promote the progress of science and the useful arts"

Codification of copyright law within the United States

- US Constitution, article 1, section 8, clause 8
- Copyright Act of 1790
- Copyright Act of 1976
    - Current copyright law

To understand copyright, you must know the following

- Original/creative and fixed
    - Creators get copyright in works that are original, creative, and fixed in a tangible medium
    - Only works that are original and embody some "minimum amount of creativity" can be copyrighted
- No registration required
- Grant of rights to owners (the "bundle")
    - Reproduce the work copies
    - Prepare derivative works
    - Distribute copies
    - Perform the copyrighted work publicly
    - Display the copyrighted work publicly
- Wide range of protected works

    - Copyright is automatically applied to the following
        - Literary works
        - Musical works
        - Dramatic works
        - Pantomimes/choreographic works
        - Pictorial, graphical, and sculptural works
        - Motion pictures and audiovisual works
        - Sound recordings
        - Architectural works
- Long term of protection
    - 1790: 14 years + 14 years renewal
    - 1909: 28 years + 28 years renewal
    - 1976: life of the author + 50 years
    - Today: life of the author + 70 years
- Exemptions and limitations
    - Section 107, fair use
    - You cannot copyright the following
        - Slogans and logos (trademark law)
        - Processes, methods, and systems (patent law)
        - Proprietary formulas/recipes (trade secret law)
        - Raw data (un-copyrightable)
            - You cannot copyright a fact

Key takeaways

- Copyright law is part of the legal system that covers creation and use
- Copyright is meant to strike a balance between the rights of the copyright holders and users of those copyrighted works, such as TDM researchers
- The cycle of creation, dissemination, and expiration of rights into the public domain is a critical component of copyright law. Without this balance the system loses its value, or prevents the public from receiving the benefits of the bargain

Public domain

Public domain is the commons of material that is not protected by copyright. Anyone is free to use, copy, share, and remix public domain works. This includes

- Works for which copyright has expired
    - When a copyright for a work expires, the work enters the public domain
    - No copyright restrictions; anyone can do anything they want with the works, including activities that were formerly the "exclusive right" of the copyright holder
    - In 2020, all works first published in the US in 1924 or earlier are now in the public domain in the US
- Works for which copyright owners failed to comply with "formalities"

    - Copyright law used to require "formalities," and if they were not met the work entered the public domain
        - These formalities existed in some form through March 1989
    - Many works published between 1925-March 1989 may also be in the public domain
        - Cornell University Library Copyright Information Center, Copyright Term and the Public Domain in the United States (updated 2020)
        - University of California Berkeley Law, Is it in the Public Domain? A Handbook for Evaluating the Copyright Status of a Work Created in the United States Between January 1, 1923 and December 31, 1977 (created 2014)
- Works that are not copyrightable
    - Facts
    - Lists of ingredients, rules of a board game
    - Titles, phrases, slogans
    - Works created by the federal government

Public domain and TDM

- If a work or collection of works that you are working with is in the public domain, then copyright issues do not apply
    - You do not need to investigate whether accessing, using, and sharing these public domain materials is allowable

Key takeaways

- The public domain has nothing to do with what is readily available for public consumption. Just because something is on the internet does not mean it is in the public domain
- The public domain is important to the production of creativity. Authors need these essential building blocks with which to work
- If a work you are using is in the public domain, copyright issues do not apply

Licenses and copyright

A license is a grant of authorization from a copyright holder to exercise one of their exclusive rights.

In a research library setting

- The license is to copy or display protected works on your computer
- Databases, eresources, and other electronic content is usually made available under a license either directly to the user or to the library on behalf of its users
- The license tells you which uses have been authorized, and authorization is often conditioned on the licensee doing certain things (like paying a fee)
- A license may also include promises by the institution or the user to not engage in certain uses, or only to use licensed content under certain circumstances

- At George Mason, you need to check the Text and Data Mining Sources infoguide for information on what resources can be text and data mined

Other settings

- You can ask the copyright owner for a license to use the work
- Read the license! Or talk to someone who has

A license can affect copyright issues in several ways

- Permitted uses are safe from copyright liability
- Not clearly permitted uses may need some additional justification
- Expressly forbidden uses may be off-limits even though copyright law would allow them (e.g., fair use) because that constitutes a breach of contract

Some works are available under public licenses that allow for specific uses of copyrighted works without the need to seek additional permission from the owner

- See Creative Commons licenses
- If works are made available under a public license or similar, these works might be used in ways that comply with the terms of the license

Fair use

Fair use is a doctrine that gives the user a right to exercise one of the exclusive rights of copyright without obtaining the permission of the copyright owner and without the payment of a license fee.

Four fair use factors

1. The purpose and character of the use
    1. Has the material been transformed by adding new meaning or expression?
    2. Was value added by creating new information, meaning, or understanding?
    3. When a work is used for a different purpose than the original, the factor will likely weigh in favor of fair use
2. The nature of the copyrighted work
    1. Was the copyrighted work that was used creative or factual in nature?
        1. The more factual the work, the more likely this factor will weigh in favor of fair use and vice versa
    2. Is the copyrighted work published or unpublished?
        1. If a work is unpublished, this factor is less likely to weigh in favor of fair use
3. The amount and substantiality of the portion used in relation to the copyrighted work as a whole
    1. Courts look at how much of the work was taken, both quantitatively and qualitatively
        1. Quantitatively, courts look at how much of the original work was used

        2. Qualitatively, some courts look at whether the "heart" of the work was taken (e.g., the essential bit of the work that is why people want to engage and acquire the work)
    2. The more that is taken the less likely the use is to be fair
4. The effect of the use on the potential market for or value of the copyrighted work
    1. Courts consider whether the use would hurt the market for the original work

Development of transformative fair use

- Use of any copyrighted materials is substantially more likely to pass fair use muster if the use is transformative
- A work is transformative if, according to the Supreme Court, it "Adds something new, with a further purpose or different character, altering the first with new expression, meaning or message"
- Courts have shifted towards collapsing the traditional four fair use factors to ask
    - Does the use transform the material, by using it for a different purpose?
    - Was the amount taken appropriate to the new purpose?

Key takeaways

- Fair use is for everyone and is useful to the TDM researcher, because TDM involves accessing, copying, and processing works that may be in copyright
- If there were no fair use, and copyright holders could forbid you from using copyrighted work without permission, this would vastly stifle free expression and scholarship
- Transformative fair use is the life and breath of scholarship, research, and teaching
- Fair use law is adaptable to various scenarios; the purpose of fair use is flexibility

Fair use and TDM

TDM researchers are fortunate to have a long and deep line of cases that provides fairly clear support for the kinds of things they do with in-copyright material. Core TDM research methods are well-suited for fair use.

Case study: Authors Guild v. Google, Inc. (2014)

- Google made digital copies of millions of books from partner research libraries
- Made the resulting corpus searchable through its Google Books service
- Using Google Book Search, users could identify books that contained a desired word or phrase
- Search results showed limited snippets of the text so users could see their term in context and get a better sense of the result's relevance to their interest
- Authors Guild sued alleging infringement; Google argued that Book Search was fair use
- Second circuit court of appeals ruled in favor of fair use
- Compare your activities to the ones analyzed here will be helpful as you figure out how fair use might apply to your research
    - Uses in the Google Books case were
        - Copying millions of complete in-copyright books to create a search index

        - Displaying snippets of in-copyright text as search results to users in the public
        - Ngram graphs showing frequency of words and phrases in the corpus over time
    - These two practices--compiling works into a machine-readable corpus, and revealing relevant portions of the corpus to the public to substantiate the results of machine analysis--are likely to occur in many TDM projects
- Second circuit held that three key activities by Google were all "highly transformative"
    - Copying of the entire text of books to create a searchable index
    - Creating the ngrams tool to show frequency of words and phrases in the corpus over time
    - Display of snippets from books as part of the search process, to help users identify relevant search results
- Four factor test
    - Factor one: purpose and character of the use
        - TDM does not merely supersede the objective of the original work but "instead add[s] something new, with a further purpose or different character"
    - Factor two: nature of the copyrighted work
        - Court gave cursory treatment to this factor, saying that nothing influenced it one way or another
    - Factor three: amount and substantiality of the portion used
        - Copying entire works was "literally necessary" to achieve its purpose of having a reliable search functionality
    - Factor four: market effect for or value of copyrighted work
        - The snippets of in-copyright text that Google does display are not a competing substitute for the original works
        - Creation of the search index did not make any of the works available to consumers, so it had no direct market effect
- Second circuit held that Google Books service was a fair use; found that
    - "The purpose of Google's copying of the original copyrighted books is to make available significant information about those books," a different function from the original books
    - The amount copied was reasonable to enable the transformative use
    - The amount revealed to users was tailored to the legitimate transformative purpose and did not threaten to substitute for ordinary consumer purchase
    - The use would not cause any market harm to the original works

Three core uses that are likely to occur in most TDM research projects

- Copying to create a database for TDM
    - Creating a database/corpus for TDM analysis is a highly transformative purpose
    - The appropriate amount for this work is typically entire works
    - Creating a database has no market effect, is not a licensable "derivative work"
- Using derived data

    - Derived data: word frequency tables, ngram graphs, etc.
    - Using derived data does not infringe on the rights of copyright owner when they comprise unprotectable facts and ideas
    - Copyright only protects the expressive content of protected works
    - Infringement requires "substantial similarity"
- Publishing data sets
    - Further publishing of the data set requires a separate fair use analysis
    - Consider effects of release on traditional market
    - Consider the amount released
    - Consider the security measures in place to prevent the kinds of access that could create harm to the market for that work

Fair use myth-busting
Claim: You cannot rely on fair use if you ask for permission and are denied

- False
- You don't have to ask for permission or alert a copyright holder when a use of materials is protected by fair use
- If you do ask for permission, you can still claim fair use if your permission request is refused or ignored

Claim: An author cannot rely on fair use if they are using an entire copyrighted work

- False
- Amount of work copied is just one factor courts consider alongside the other factors
- Courts look at whether the amount used was reasonable in light of the purpose of the use
- Example: Google Books case

Claim: You cannot rely on fair use if you are using unpublished material

- False
- Congress amended the Copyright Act in 1992 to explicitly allow for fair use when using unpublished works

Claim: An author cannot rely on fair use if they are using highly creative copyrighted work

- False
- While courts do consider whether the copyrighted material used is primarily factual or creative under the second factor, "the nature of the work," this factor is rarely decisive on its own

Claim: An author cannot rely on fair use if they are making commercial use of a copyrighted work

- False
- Commercial uses can be fair use, and not all noncommercial uses will be fair use

Licensing

Licensing basics

A license is a "contract not to sue." A license or a contract is a legal interest created by a titleholder granting use privileges to some non-titleholder. Licenses can determine what a TDM researcher can do within legal bounds.

License requirements (what makes a license valid)

1. Offer
    1. Parties make a promise to do some specified action in the future
2. Consideration
    1. Something of value is promised in exchange for the specified action or non-action
    2. It is the value that induces the parties to enter into the contract
3. Acceptance
    1. Offer has to be clearly accepted
    2. May be expressed through words, deeds, or performances as called for in the contract
4. "Meeting of the minds"
    1. Parties understood and agree to the basic substance and terms of the contract

Standard license provisions

- Parties
    - Name the correct parties
    - Good area to look for in case you need to contact the person/party
- Overview/purpose
    - This is an opportunity to tell parties (including third parties viewing the contract) what the contract is about
- Payment
- Date
    - Date of execution by each party is included so there is a time at which the parties become bound to the contract
- Signature

"Boilerplate" contracts

- Standard, and not typically heavily negotiated
- Repeated terms appearing in all kinds of contracts
- Most likely, the TDM project you are dealing with will have boilerplate language

Sections in licenses where TDM related clauses might reside

- Authorized uses or permitted uses
    - There is sometimes a section on non-permitted uses or restrictions
- Definitions
    - This section occasionally defines TDM
- Any sections listed as "intellectual property" or "copyright"

    - TDM related clauses usually live in these sections

Open and public licenses
Public licenses are "boilerplate," meaning that they are non-negotiated. These are licenses under which copyright holders may choose to release their works for use by the public without requiring special permission. Open licenses are a subset of public licenses.

- Provides a mechanism for copyright holders to grant to "the public" permissions to use their work
- A license between the copyright holder and the public for particular uses of a work that would otherwise be restricted
- Don't undo or modify copyright
- Rely on the copyright holders' exclusive economic and moral rights in order to do something interesting with their works: to choose not to make money from their creations, or to choose not to prevent redistribution or derivative works

Creative Commons

- Provide a bunch of options that modify the blanket permissions granted by the copyright holder
- These options can include
    - Every use be accompanied by an attribution (CC-BY)
    - Only non-commercial uses are allowed (CC-NC)
    - Disallow derivative works (translations, etc.) (CC-ND)
    - Many more
- Specifically designed to make copyright-protected content more open
- Are open because any use not strictly prohibited is actually allowed

Library e-resource licenses
Problematic terms in library e-resource licenses

- Data mining
    - *Subject to any content-specific restrictions, Customer and its Authorized Users may extract and compile data from locally-loaded copies of the Purchased Content for Customer's teaching, learning, and research purposes*
- Restrictions
    - *Except as expressly permitted above, Customer and its Authorized Users shall not: ... text mine, data mine or harvest metadata from the service*

Principles for negotiating licenses

- The right to read is the right to text mine; this is a right we should never willingly sign away

    - Some have advocated for an escape clause in licenses, along the lines of "notwithstanding the foregoing, nothing in this license should be read to prohibit fair use"
    - Since the courts typically rule in favor of fair use, this clause should allow for TDM
- Maintain the clear position that one of the primary affordances of electronic texts is the ability to read them with a computer--that is, to do TDM
    - Why spend so much money for an electronic copy if we can't use it differently than we would a print book?
- It is crucial to be prepared to walk away from negotiations if the terms aren't right

Model licenses

- Various research library consortia have developed and adapted as an expression of what the community considers reasonable expectations for licensing terms
- Important for several reasons
    - Lighten the load of drafting from scratch
    - Set general expectations that are shared by the community
- Examples
    - Center for Research Libraries
    - NERL - NorthEast Research Libraries
    - Because these and other model licenses originate in the academy, they are favorable to academic uses
    - Serve to put our vendors on notice that these are terms our community expect and will demand

Model licenses language

- Text and data mining
    - *Authorized Users may use the Licensed Materials to perform and engage in text and/or data mining activities for academic research*
- Escape clause
    - *Licensee and Authorized Users may make all use of the Licensed Materials as is consistent with United States copyright law, including its Fair Use Provisions*

Websites and other terms of use

A concern among TDM researchers is whether there are legal repercussions and/or breach of contract when scraping is inconsistent with website policies.

Computer Fraud and Abuse Act

- Bars any "unauthorized" access to any "protected computer"
- Courts have said this means essentially any machine connected to the internet

Twitter Developer policy is a good example of a robust, enforceable contract governing a commonly-used source of research data

- Twitter API makes it easy to retrieve massive amounts of data from Twitter
- Twitter tightly regulates how that data can be used and shared
- Twitter API Terms create a strong, enforceable contract by ensuring anyone who participates is required to clearly signal their assent
- Only permits access to those who have created an account and assented
- Twitter makes special allowances for scholarly use, but academics are also prohibited from sharing large corpora of full text tweets

Digitized library collections may also be governed by contracts, even public domain material. This includes

- HathiTrust
    - Created in partnership with Google
        - Limitations on reuse were part of that arrangement
    - HathiTrust and member libraries use terms of use restrictions to ensure that users don't do anything that would place them in breach of their agreement with Google
    - Additional terms govern HathiTrust Research Center's TDM tools and are designed to ensure users remain within the bounds of what fair use permits
- Adam Matthew
    - Common for these materials to be in the public domain, but they are rare and might not exist in digital form elsewhere
        - It is possible to keep them behind paywalls and monetize access
    - Typically requires users to agree not to download collections in bulk or share them publicly

Beyond the terms of the license

Other than the terms, ask

- Am I bound?
    - A contract requires both offer and acceptance
    - To accept a contract, you need adequate notice of its terms
- How does this relate to fair use?
    - Myth: fair use doesn't apply to licensed materials
    - Reality: fair use and contract are separate sources of authority
    - Remember, a license is authorization to exercise a copyrighted right, and when your use is fair, you don't need authorization
    - But, a contract is a legally enforceable promise, and you can promise not to exercise fair use rights
    - Everything depends on the specifics of the contract
    - Unless the language is clearly prohibiting TDM, fair use typically survives
        - "User shall not..." or "User shall not...without additional permission"
- What happens if I breach?
    - Remedies for (non-infringing) breach of a license are less severe than copyright
    - Breaching a license is not always infringement

    - Private enforcement is more likely (and can still be serious)
    - Most likely negative outcome is that the licensor can cut off access to the resource institution-wide
- Trespass to chattels?
    - Unreasonable interference with the ordinary use of someone's personal property
    - Trespass (an intentional interference with another person's lawful possession of property) to chattels (personal property, eg, a server)
    - Examples
        - A DDOS attack, which barrages a server with so many inquiries that the server becomes unusable for its ordinary purpose
        - Automated scraping or web harvesting that takes place in a time or manner that interferes with the vendor's ordinary use of the server
- Risk management
    - Consider permission
        - Reach out to copyright holder/licensor about getting additional or more specific permissions
        - Being told no doesn't hurt your fair use argument
    - Be polite
        - Good will can be won by scraping, downloading, or accessing content in ways that don't interfere with the ordinary use of a licensed resource
            - When scraping, don't hit the servers hard, especially during normal business hours
    - Welcome dialog
        - Be available and responsive when folks have concerns
        - If someone reaches out, don't ignore them
        - Be willing/able to take it down
        - If you share data, include a point of contact

Potential solutions

Short term solutions that might respect legal boundaries

- Non-consumptive/non-expressive research modes (eg, HathiTrust Research Center)
    - Hathi Trust Research Center and the Hathi Trust corpus allows for access to entities, sentiment scores, token counts, verb counts, etc. without violating terms of use
- Publishing metadata and extracted features (eg, mediacloud.org)
    - This allows users to find the full text content on their own, through their own licensing regime
- Remote access to compliant computer systems
    - A virtualized server residing in an institution and is bound by its licensing agreements but the data always stays on the hard drive of that compliant physical server
    - Enables collaboration across institutions and internationally because researchers could remote-in and access the data that stays on the server

- Publishing or sharing small validation sets
    - Random samples of larger corpora that are published under fair use provisions (if that would apply)
    - This would allow collaborators to develop and refine their algorithms that they want to run on the larger corpus
- Working face to face
    - Most campus licensing provisions for library materials have a cutout for visiting scholars
    - If you bring someone physically to your location they will have temporary access to the same content you have

Longer-term solutions

- Build more collaborative open data sets
    - Linguistic Data Consortium
        - Has a small, but reasonable, licensing fee
        - Gives access to full text data that can be shared
    - AWS Common Crawl
        - Freely available web content at a very large scale
        - Licensing might apply here, so check
    - WIkipedia
        - Large amount of content that can be mined and shared with no restrictions
- Advocate for better data agreements
    - Clearer terms, more expansive allowable uses for research purposes
    - Give us clearer boundaries so we know what we can and cannot do
    - Also respects the need for researchers to have relatively free and broad access to sensitive, in-copyright materials
- Empower "data ombudspersons"
    - Someone empowered to make a final decision for your campus of what you're allowed and not allowed to do with text and data
    - Would know the legal landscape, understand licensing, etc.

Privacy

Privacy law

US sources of privacy law

- Constitution
    - Constitution does not explicitly include the right to privacy
    - Supreme Court has found it implicitly grants a right to privacy against governmental intrusion in the First, Third, Fourth, and Fifth Amendments
    - These constitutional rights are not what we're dealing with in a TDM context
- Federal statutes

    - Includes the Children's Online Privacy Protection Act, Fair Credit Reporting Act, Family Educational Rights and Privacy Act, etc.
    - Any research you do that involves the kind of information at issue under any of these statutes is going to require IRB approval because it invokes "private information" under federal law
    - These statutes impose obligations on how such data should be collected, managed, and disclosed or not
- State statutes and common law (law derived from the court opinions themselves, apart from any statute that might exist)
    - General privacy laws created by states
    - These statutes and common law create a tort cause of action resulting from an unlawful invasion of privacy
    - Torts are a wrongful act or an infringement of a right (other than under contract) leading to civil legal liability
    - It is a civil, not a criminal, wrong that you could do to someone, and it's something you do that's an infringement of some non-contract based right that either statutes or common law have created

Prosser Torts

- Four torts we need to be aware of as TDM researchers
- These privacy torts vary by state, so there will be questions of which state's law applies
- These are four privacy torts most states recognize either through statute or common law rights

1. Intrusion upon seclusion
2. Public disclosure of embarrassing private facts
    1. Both of these require the invasion of something secret, secluded, or private
    2. Person must have had a reasonable expectation of seclusion or solitude in the invaded place or as to the particular topic or matter intruded upon
    3. Community standards are important for gauging whether a privacy violation has occurred
    4. These are the two Prosser torts you should be concerned about in the context of TDM research
3. Painting someone in a false light
    1. If you've published information widely (whereas defamation is to a single person)
    2. Publication identifies the plaintiff
    3. There is an element of fiction or falsity
    4. The falsity would be highly offensive to a reasonable person
    5. You were at fault in publishing the information
4. Appropriation of name or likeness
    1. Protects a person's exclusive use of his or her own identity
    2. Name or likeness means means the concept of a person's character

Exceptions to Prosser torts favorable to TDM researchers

- Right of privacy is not violated by comment or disclosures as relates to matters of legitimate public interest
    - Courts look at whether the facts you are seeking to disclose are of legitimate public concern and/or would be highly offensive to a reasonable person
- A person's death ends their right of privacy
- No privacy concerns if the people are not identifiable
- If someone has released the information themselves, whether by giving you permission or through social media, they cannot sustain a privacy tort claim

Key takeaways

- Work on your TDM project can proceed if the following is true
    - If the subject matter of the collections is no longer living, or
    - If subject matter is newsworthy or of public interest, or
    - If the subject matter is not identifiable, or
    - If the subject matter has released the information themselves, either by giving you permission or posting on social media

General Data Protection Regulation (GDPR)

Researchers are not guaranteed to be insulated from international privacy regulation when their data collection is conducted within the US, as it can be produced in countries across the world. The contexts in which people share information online should play an important role in the sharing and use of information, even if US privacy law doesn't cover it.

General Data Protection Regulation (GDPR)

- Enacted in European Union in 2016; enforceable beginning in 2018
- Deals with protection of privacy and the collection and management of data
- Core of GDPR is personal data
- Information that allows a living person to be directly or indirectly identified from available data
    - Can include person's name, location, etc.; can also be an IP address
- Aims to give people greater control over their personal data
- Enacts technical measures that dictate how businesses and other entities process personal data of EU citizens
- Businesses and data controllers are required to enable safeguards to protect user data so that datasets are not public by default and cannot be used to identify people
- Can apply to entities that are based outside of the US
- Social media companies and other organizations that provide products and services to EU citizens is directly affected by these principles

GDPR data protection and accountability principles

1. Processing must be lawful, fair, and transparent to the data subject
2. Processing must only be for the legitimate purposes specified explicitly to the data subject at the time of collection

3. It should collect and process only as much data as absolutely necessary
4. Personal data must be kept accurate and up to date
5. Processors can only store personally identifying data for as long as needed to fulfill the specified purpose
6. Processing must be done in a way that ensures appropriate security, integrity, and confidentiality of the data
7. Data controller is responsible for being able to demonstrate compliance with these principles

GDPR users' rights

1. The right to be informed
2. The right of access
3. The right to rectification
4. The right to erasure
    1. Can be invoked when
        1. Personal data no longer necessary in relation to the purpose for which they were collected or processed
        2. When the data subject withdraws consent
        3. When personal data has been unlawfully processed
        4. When personal data must be erased to comply with a legal obligation in the EU or Member State law
        5. A few others
5. The right to restrict processing
6. The right to data portability
7. The right to object
8. Other rights in relation to automated decision making and profiling

GDPR safety values for TDM research

- Provisions will not apply under certain circumstances
    - Exercising the right of freedom of expression and information
    - Reasons of public interest in the area of public health
    - Archiving purposes related to scientific, historical, and statistical research
    - For the establishment, exercise, or defense of legal claims

Ethics

Sensitive data

Two ethics questions to ask when working with data

1. Public data used in TDM contexts might not be protected by privacy statutes. To what degree of care should be treat that data? Essentially, what should we do with data that is not private but might be sensitive?
    1. Average user's expectation of privacy doesn't always match with the law's definition of privacy

    2. For international research projects, protections will vary greatly depending on which national data protection regulation applies
    3. Many creators of data do not anticipate secondary uses of that data
        1. Novelists might not expect their works to be converted to data
2. When should you impose an ethical framework, and how do you balance that framework with truth-seeking, the public interest, and free expression?

Public but sensitive

- The use of the information for research purposes is potentially stripped of important narrative
- This can cause personal harm to the author and in some cases perpetuate harm to historically marginalized communities
    - Discovery replays a colonial paradigm, where content is imagined as unhinged from peoples and cultures and free for the taking (Kimberly Christen)

Structural racism and power balances

- Sensitive information might have been created under unequal power structures
- Underprivileged groups might lack the knowledge of how information about them will be used or have the ability to intervene in that usage
- World Intellectual Property Organization has tried to develop international frameworks to protect communities from having their traditional knowledge exploited
- Have also tried to protect them from overstudy and from not receiving the benefits of the research in some meaningful way
- Intellectual property laws and human rights frameworks are focused on individual rights, not group rights
- Without a move towards group rights it is not entirely possible for marginalized communities to have freedom to create, use, and enjoy knowledge assets

Obtaining consent

Consent-based model

- Even if consent is given for re-use in terms of service for a social media site, because the details are often buried in lengthy terms of service, users are likely unaware that they have consented to human subjects research through their use of a social networking platform alone
- How can it actually be considered consent if someone cannot have even conceived of the yet-to-be-determined queries a researcher might run relying on their personal data?
- For this reason scholars are moving away from the consent-based paradigm to a harms-based paradigm
- A consent-based research paradigm is not conducive to TDM

Harms-based model

- This paradigm is based on treating "harm"

- The current regulatory framework is based off of the Common Rule and provides protections for clinical human subjects research
    - Term for the Federal Policy for the Protection of Human Subjects, which outlines the criteria and mechanisms for institutional review board (IRB) review of human subjects research
    - Outlines the basic provisions around informed consent and assurances of compliance
    - These are implemented by Institutional Review Boards
- Common Rule and TDM
    - Common rule is often inapplicable to TDM methods
    - Many TDM projects in the humanities fall outside the Common Rule's reach where there is no direct interaction with subjects or studies that involve subjects' private, identifiable information
    - That means that institutions are not required to oversee the research at all, even if there are ethical concerns
- Common Rule heavily influenced by the Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research
    - Core principles
        - Respect for persons, beneficence, and justice
    - Three areas of application
        - Informed consent
        - Assessment of risks and benefits
        - Selection of subjects

Ethical frameworks

Imagine you have the capacity to help someone in need, and helping them would not diminish your own capacity. Should you provide the help?

- Deontologist: Would recognize an obligation to help in accordance with a moral rule such as "do unto others as you would have them do unto you"
- Virtue ethicist: Would act based on the fact that helping the person would be charitable or benevolent
- Utilitarian: Consequences of helping will maximize well-being for the greatest number of people

These normative ethical frameworks pose problems for TDM research

- Each of these frameworks places an emphasis on moral responsibility and the agency of the individual
- Moral agency assumes free will
- There might not be free will of the participants for a few reasons
    - Power imbalances complicate the idea of choice or free will
    - Data could be used beyond the purpose for which it was originally intended
    - "Distributed morality" of big data: ethics of data use in a networked framework might be dependent on the morality of other actors in that network

Alternative ethics frameworks

- Ethics of care
    - Premised on relationships and care as a virtue
    - Recognizes uneven power relationships
    - Builds into research design an account for who possesses power or authority in a situation
    - Enables a progression from account for the rights and obligations of individuals to the rights and obligations of groups
    - Focus on relationships can make it very challenging to apply, especially for large data sets
- Risk-benefit analysis
    - Advocates for big data researchers and review boards to incorporate systematic risk-benefit assessments
        - Evaluate the benefits that would accrue to society as a result of a research activity
        - Intended uses of the data involved
        - Privacy threats and vulnerabilities associated with the research activity
        - Potential harms to human subjects as a result of the inclusion of their information in the data
    - Decision about whether to proceed with the research based on those balanced factors is not a yes/no decision

Kinds of harm (Dixon & Quirke)

- Psychological
    - Refers to participants' well-being, and inclusive of things like distress, embarrassment, stress, and betrayal of trust
- Physical
    - Includes physical pain, injury, and death
- Legal
    - Includes legal implications from exposure
- Social
    - Includes damage to relationships, social standing, or reputation, and would include impacts on personal and employment relationships through the disclosure of information

Key takeaways

- Ethics is not a one-time consideration
    - "Ethical issues permeate and unfold beyond the research design stage and throughout the entire research process" (Dixon & Quirke)
- When doing TDM research, consider different ethical frameworks and the different types of harm, and how best to minimize or weigh that harm

Strategies

Strategies to account for ethics

- Consult journal publication or professional association guidelines
- Develop local best practices
    - Conduct decision-making within your research group
- Impose access controls
    - Including user registration to view, publishing only data visualizations or extractions, etc.
- Undertake community engagement to consult with affected populations, and ensure that benefit reverts back to the communities
- Seek IRB involvement/approval, even if none is technically required
    - This can overwhelm IRBs and slow down the research process, so structural changes at your institution might be needed
- Adopt a new ethics/privacy paradigm
    - Moving from consent-based to harm-avoidance
    - Develop a balancing test that you like

To enact these strategies

- Regulations might need to be changed
- Policies of review boards might need to be revised to adopt definitions for terms such as privacy, confidentiality, security, and sensitivity
- Start with your own research and research teams at GMU