Clinical Text Analysis

Challenges and Opportunities
Natural Language Processing (NLP) is an expansive and highly complex topic. As such, we will cover a general set of NLP techniques as they pertain to clinical notes.

*We have a great course here at ND about NLP if you’re interested in the topic*

- **Wednesday:**
  - Structural
    - Segmentation and Tokenization
    - Stop words
    - Punctuation and Numbers
  - Morphological Processing
    - Spelling
    - abbreviations
    - Stemming and Lemmatization
  - From Text to Feature Vectors
    - BoW Representation

- **Today:**
  - Improvements to BoW representation
    - TF-IDF / N-grams
  - Modeling Data
    - Similarity
    - Machine Learning Approaches
  - Bonus Content:
    - UMLS
    - State of the Art: Word Embedding
Data Representation

“This patient will likely be readmitted”

```
[1, 0, 1, 0, 1, 0, 1, 1, ...]
```

# Raw Counts and Beyond

<table>
<thead>
<tr>
<th>Sentence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
<td></td>
</tr>
<tr>
<td>Patient has broken arm</td>
<td></td>
</tr>
<tr>
<td>Patient was mobile. Patient discharged</td>
<td></td>
</tr>
</tbody>
</table>
Raw Counts and Beyond

<table>
<thead>
<tr>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
</tr>
<tr>
<td>Patient has broken arm</td>
</tr>
<tr>
<td>Patient was mobile. Patient discharged</td>
</tr>
</tbody>
</table>

First, all unique terms are selected across all documents/ sentences to create a **vocabulary**

- Unique Terms:  
  - Patient  
  - Has  
  - Diabetes  
  - Broken  
  - Arm  
  - Healed  
  - Discharged
**Bag of Words Representation (BoW)**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Patient</th>
<th>Has</th>
<th>Diabetes</th>
<th>Broken</th>
<th>Arm</th>
<th>Healed</th>
<th>Discharged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient has broken arm</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient was mobile. Patient</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>discharged</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- A unique index is provided to each word in the vocabulary.
- Each sentence is then represented by a vector that is as long as the number of distinct terms in our vocabulary.
  - At each index in this list, we mark how many times the given word appears in our sentence.
Extending Counts

- A problem with utilizing term counts is that highly frequent (non stop word) terms can dominate the document
  - With larger counts, remember our eye blink ex.

- However these terms may not contain as much information to differentiate the documents as rarer but specific terms.

- One approach is to rescale the value for a term by how often it appears across documents
  - In this way frequent terms in a document that are frequent across all documents are penalized.
Term Frequency – Inverse Document Frequency (TF-IDF)

This approach to scoring is called: *Term Frequency – Inverse Document Frequency*,

- **Term Frequency**: Score of frequency of the term in the current document.
  - Slightly refined from pure counts:
    - As every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones.
    - In term frequency the count is divided by the document length to normalize

- **Inverse Document Frequency**: Score of rarity of term across all documents

\[ W_{t,d} = tf_{t,d} \times \log \left( \frac{N}{df_t} \right) \]

- \( df_t = \# \text{ documents with term } t \)
- \( tf_{t,d} = \# \text{ occurrences of } t \text{ in } d \)
- **TF-IDF Example:**
  - Sentence 1: The infant is admitted to the NICU.
  - Sentence 2: The patient is admitted to the MICU.

<table>
<thead>
<tr>
<th>Word</th>
<th>TF</th>
<th>IDF</th>
<th>TF*IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A 1/7</td>
<td>B 1/7</td>
<td></td>
</tr>
<tr>
<td>The</td>
<td>1/7</td>
<td>1/7</td>
<td>Ln(2/2) = 0</td>
</tr>
<tr>
<td>Infant</td>
<td>1/7</td>
<td>0</td>
<td>Ln(2/1) = 0.3</td>
</tr>
<tr>
<td>Patient</td>
<td>0</td>
<td>1/7</td>
<td>Ln(2/1) = 0.3</td>
</tr>
<tr>
<td>Is</td>
<td>1/7</td>
<td>1/7</td>
<td>Ln(2/2) = 0</td>
</tr>
<tr>
<td>Admitted</td>
<td>1/7</td>
<td>1/7</td>
<td>Ln(2/2) = 0</td>
</tr>
<tr>
<td>To</td>
<td>1/7</td>
<td>1/7</td>
<td>Ln(2/2) = 0</td>
</tr>
<tr>
<td>The</td>
<td>1/7</td>
<td>1/7</td>
<td>Ln(2/2) = 0</td>
</tr>
<tr>
<td>NICU</td>
<td>1/7</td>
<td>0</td>
<td>Ln(2/1) = 0.3</td>
</tr>
<tr>
<td>MICU</td>
<td>0</td>
<td>1/7</td>
<td>Ln(2/1) = 0.3</td>
</tr>
</tbody>
</table>
Bag of Words Model

• Regardless of which representation (Count, TF-IDF) is used, the resulting vector model it is called a “bag”, as information about the order or structure of words is discarded.
  – We simply know if the word appeared in the document or not

• Now it should be clear why correctly extracting and matching terms in the preprocessing steps is so important
  – If two terms have capitalizations, tenses, misspelling, etc. they will each have a unique feature potentially misrepresenting data
What About Compound Phrases

• We use multiple terms to describe scenarios everyday
  – New York is neither *New* or a *York*

• Clinically this can be very meaningful
  – Benign tumor
  – Malignant tumor

• Some role in negation as well
  – Patient is not good
  – Patient is very good
N-Grams

- As a solution we can extend our BoW model to look at groupings of sequential terms.
- The number of terms is defined by $N$
  - However the BoW model is constructed in the same way, where each n-gram is treated as a unique term and the count/tf-idf value is recorded in the vector

\[
\begin{align*}
N = 1 : & \text{This is a sentence} \\
N = 2 : & \text{This is a sentence} \\
N = 3 : & \text{This is a sentence}
\end{align*}
\]
### Bi(2)-Gram Representation Example

**Sentence:**
- while walk, accidentally fall knee hit ground, near left eye. her fall observe, patient profess loss consciousness, recall entire event.

<table>
<thead>
<tr>
<th>walk</th>
<th>observe</th>
<th>('eye', 'her')</th>
<th>('accidentally', 'fall')</th>
</tr>
</thead>
<tbody>
<tr>
<td>accidentally</td>
<td>patient</td>
<td>('her', 'fall')</td>
<td>('fall', 'knee')</td>
</tr>
<tr>
<td>fall</td>
<td>profess</td>
<td>('fall', 'observe')</td>
<td>('hit', 'ground')</td>
</tr>
<tr>
<td>knee</td>
<td>loss</td>
<td>('observe', 'patient')</td>
<td>('ground', 'near')</td>
</tr>
<tr>
<td>hit</td>
<td>consciousness</td>
<td>('patient', 'profess')</td>
<td>('left', 'eye')</td>
</tr>
<tr>
<td>ground</td>
<td>recall</td>
<td>('profess', 'loss')</td>
<td>('near', 'left')</td>
</tr>
<tr>
<td>near</td>
<td>entire</td>
<td>('loss', 'consciousness')</td>
<td>('left', 'eye')</td>
</tr>
<tr>
<td>left</td>
<td>event</td>
<td>('consciousness', 'recall')</td>
<td>('recall', 'entire')</td>
</tr>
<tr>
<td>eye</td>
<td>('while', 'walk')</td>
<td>('recall', 'entire')</td>
<td>('entire', 'event')</td>
</tr>
<tr>
<td>fall</td>
<td>('walk', 'accidentally')</td>
<td>('recall', 'entire')</td>
<td>('entire', 'event')</td>
</tr>
</tbody>
</table>
### Bi(2)-Gram Representation Example

- It should be clear that even at an order of 2 the feature space can explode for even small vocabularies
- More, as we increase N (tri-grams, etc.), more low frequency terms are created
  - As a result thresholds are often used
    - Terms must appear in at least X% of documents/ sentences

<table>
<thead>
<tr>
<th>walk</th>
<th>observe</th>
<th>('eye', 'her')</th>
<th>('accidentally', 'fall')</th>
</tr>
</thead>
<tbody>
<tr>
<td>accidentally</td>
<td>patient</td>
<td>('her', 'fall')</td>
<td>('fall', 'knee')</td>
</tr>
<tr>
<td>fall</td>
<td>profess</td>
<td>('fall', 'observe')</td>
<td>('knee', 'hit')</td>
</tr>
<tr>
<td>knee</td>
<td>loss</td>
<td>('observe', 'patient')</td>
<td>('hit', 'ground')</td>
</tr>
<tr>
<td>hit</td>
<td>consciousness</td>
<td>('patient', 'profess')</td>
<td>('ground', 'near')</td>
</tr>
<tr>
<td>ground</td>
<td>recall</td>
<td>('profess', 'loss')</td>
<td>('near', 'left')</td>
</tr>
<tr>
<td>near</td>
<td>entire</td>
<td>('loss', 'consciousness')</td>
<td>('left', 'eye')</td>
</tr>
<tr>
<td>left</td>
<td>event</td>
<td>('consciousness', 'recall')</td>
<td></td>
</tr>
<tr>
<td>eye</td>
<td>('while', 'walk')</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fall</td>
<td>('walk', 'accidentally')</td>
<td>('recall', 'entire')</td>
<td></td>
</tr>
</tbody>
</table>
Using the Representations

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Patient</th>
<th>Has</th>
<th>Diabetes</th>
<th>Broken</th>
<th>Arm</th>
<th>Healed</th>
<th>Discharged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient has broken arm</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient was mobile. Patient discharged</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
### Special Case: Similarity

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Patient</th>
<th>Has</th>
<th>Diabetes</th>
<th>Broken</th>
<th>Arm</th>
<th>Healed</th>
<th>Discharged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient has broken arm</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient was mobile. Patient discharged</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- One of the most common questions surrounding text representations is, “How similar are these sentences / documents”?

- However, determining similarity is not always so easy
  - Documents can have vastly different lengths, which can skew the BoW Euclidian distance based simply on magnitude.
### Cosine Similarity

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Patient</th>
<th>Has</th>
<th>Diabetes</th>
<th>Broken</th>
<th>Arm</th>
<th>Healed</th>
<th>Discharged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient has Diabetes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient has broken arm</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patient was mobile.</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Patient discharged</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- More commonly, we turn to the cosine similarity, which offers a normalized comparison of the feature vector.
  - Captures both magnitude and direction.

\[
similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}
\]
Modeling Text Data

• Beyond clusters (similarity groupings) It should be clear that the BoW representation provides the same type of feature vector we have worked with all semester
  – Rather than different factors (age, sex, gender), each feature is a distinct term entity (Unigram or n-gram)
  – Can use either counts or TF-IDF, as each term has a unique continuous value

• As a result we can execute many of the modeling or comparison techniques we have discussed throughout this class.
  – Remember that each row in the transformed data represents a unique **document**, thus we may need special considerations depending on the research question
    • Aggregate all notes from one individual into a single document, etc.
Influential Terms

- While TF-IDF lets us identify the most important terms within a document, linear models offer a nice framework to identify important terms between documents or outcomes.

- Although we have mostly covered linear models in this class due to their interpretability, there exist a number of more appropriate models with respect to text analysis:
  - Support Vector Machines
  - Naive Bayes
  - Random Forests
Advanced Topics

STATE OF THE ART
Unified Medical Language System®

• Provided by the National Library of Medicine
  – UMLS is designed to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health.

  – Comprised of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
Unified Medical Language System®

- Made up of 3 primary tools
  - Metathesaurus®
  - Semantic Network
  - The SPECIALIST Lexicon
The Metathesaurus®

- Contains over one million biomedical concepts from over 100 source vocabularies, organized by concept or meaning

- Primary purpose is to map between coding systems (taken from established vocabularies such as SNOMED, ICD-9-CM, and MeSH)
  - Links alternative names and views of the same concept and identifies useful relationships between different concepts.
Preferred terms (Concepts)

Collection of terms in the concept:

- disease; Hodgkin
- Hodgkins disease
- Hodgkin Disease
- Hodgkin's disease, unspecified
- Hodgkin's disease, unspecified type
- Hodgkin's disease (clinical)
- Hodgkin's disease NOS, unspecified site
- Hodgkin's disease NOS (disorder)
- Hodgkin's sarcoma (clinical)
- Hodgkin's sarcoma NOS
- Hodgkin's sarcoma of unspecified site
- Hodgkin's sarcoma of unspecified site (disorder)
- Hodgkin's sarcoma-unspec. Site
- Hodgkin lymphoma
- Lymphogranuloma, Malignant
- Lymphogranulomatosis
- Lymphogranulomatosis, malignant
- Lymphomas Hodgkin's disease

- Hodgkin's Disease
Semantic Network

• Provide consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts.

• The Semantic Network contains 133 semantic types and 54 relationships
  – Major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts / ideas

https://semanticnetwork.nlm.nih.gov/
SPECIALIST Lexicon

• A general English lexicon that includes biomedical vocabulary.
  – Coverage The lexicon entry for each word or term records the syntactic, morphological, and orthographic information

• Words often have several inflected forms.
  – The verb "treat", for example, has three inflectional variants: "treats" the third person singular present tense form, "treated" the past and past participle form, and "treating" the present participle form.

• Multi-word terms in the Metathesaurus may have word order variants in addition to their inflectional and alphabetic case variants.
  – The Lexical Tools allow the user to abstract away from this sort of variation.
State-of-the-Art: Word Embedding

- Looking to take the next step beyond discrete words, researchers are now looking to semantics
- Significantly more challenging problem
  - Relies on the context of words
  - Made possible by the massive amounts of text now available.
State-of-the-Art: Word Embedding

- Given a vector of fixed length:
  - Uses neural networks to learn a representation to predict the word given it’s context.
  - Does so by looking at word orderings:
    - Continuous Bag of Words, Skim Gram, etc.

![Diagram of Word Embedding]

Given a 1-hot vector of fixed length, it uses a neural network to learn a representation that predicts the word given its context. This is done by looking at word orderings, such as Continuous Bag of Words or Skim Grams. The diagram illustrates this process with a neural network structure, where the input vector is processed through hidden and output layers to predict the word. The learned weight matrix is shown for the example of the word "ants," which is represented as a 1-hot vector. The multiplication of the 1-hot vector with the learned weight matrix results in the word embedding vector [10 12 19].
Recent Example

GRAM: Graph-based Attention Model for Healthcare Representation Learning Edward Choi, et al. Knowledge Discovery and Data Mining (KDD) 2017

http://mp2893.com/scatterplot/dpm8_glove.html
Human Subjects Research and the Institutional Review Board (IRB)
A Little Bit of (Unfortunate) History

- Research hasn’t always been carried out so ethically...
  - 40-year-long U.S. Public Health Service Syphilis Study at Tuskegee
  - Willowbrook Experiments
  - Sadly, many more…. (Jewish Chronic Disease Hospital Case, San Antonio contraceptive study, etc.)

- Resulted in legislation calling for regulations to protect human subjects, and for a National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research to examine ethical issues related to human subject research.

- The Commission's final and most influential report, *The Belmont Report*, elucidates three basic ethical principles that investigators must balance when conducting research with human subjects: respect for persons, beneficence and justice.
Regulation History

- Nuremberg Code of 1947
- Declaration of Helsinki 1964
- PHS Memo 1966
- National Research Act of 1974
  - Established the modern IRB system for regulating research involving human subjects.
- Belmont Report 1979
The Institutional Review Board (IRB) is an independent committee established to review and approve research involving human subjects.

**BENEFICENCE**
- Risk/Benefit Analysis
- Data Safety
- Experimental Design
- Qualifications of PI

**JUSTICE**
- Subject selection
- Inclusion/exclusion
- Recruitment

**RESPECT FOR PERSONS**
- Informed consent
- Surrogate consent
- Assent
- Privacy & Confidentiality
- Vulnerable Populations
The Common Rule

• In 1991, 16 federal agencies formally adopted the core of these regulations in a common Federal Policy for the Protection of Human Subjects – Known as the Common Rule

• Dictates much of how we review and regulate institutional research today
The IRB in Terms of the Common Rule

5 CFR 46.111 and 21 CFR 56.111

• The risks to human research subjects are minimized by using procedures that are consistent with sound research design and that do not unnecessarily expose the research participants to risk, and whenever appropriate, by using procedures already being performed on subjects for diagnosis or treatment purposes.

• The risks to human research subjects are reasonable in relation to the anticipated benefits (if any) to the individual, and the importance of the knowledge that may be expected to result.

• The selection of human subjects for research participation is equitable.

• Human research subjects are adequately informed of the risks and benefits of research participation and the procedures that will be involved in the research; and that informed consent is obtained from each prospective human research subject, or his/her legally authorized representative, in accordance with, and to the extent required by federal regulations and IRB policies.

• Research plan, when appropriate, makes adequate provisions for monitoring the data collected to ensure the safety of human research subjects.

• There are adequate provisions to protect the privacy of human research subjects and to maintain the confidentiality of research data.

• Appropriate additional safeguards have been included in the study to protect the rights and welfare of human research subjects who are likely to be vulnerable to coercion or undue influence – (e.g., children, prisoners, pregnant women, decisionally impaired persons, or economically or educationally disadvantaged persons).
Human Subjects Research

- Although this seems like an obvious question, not all research requires IRB approval
  - We must ask, is this Human Subjects Research

- Animal research is also highly regulated, but that is outside the scope of this class
Human Subjects Research

- To begin, we must define a “Human Subject” as a living individual, about whom an investigator conducting research obtains:
  - Data through intervention or interaction with the individual:
    - Note that a human subject is not simply any person with whom the investigator interacts- the interaction must result in data about the person, themselves.
    - Intervention and interaction can include procedures done actively in person (interviews), passively (observation), and those done electronically (phone calls, emails, and electronic surveys). Intervention may also include manipulation of a subject’s environment.

  OR

  - Identifiable private information (We will come back to this)
    - **Private data** is information about characteristics or behavior in a context in which a person can reasonably expect no observation or recording is taking place. Examples may include a medical record, bank account information, and similar types of information. However, it may also include personal beliefs or responses, which a subject would not otherwise publicize
    - **Identifiable information** allows an observer to determine the identity of the source, whether that is through identifiers like name, date of birth, address, phone number, and more, or it may be through the use of a code or other data maintained to linking research information to the source.
Is it Research?

• We also define Research as:
  – A systematic investigation,
  – Designed to contribute to generalizable knowledge

  – **Systematic investigation** means the researcher has identified a plan to collect data in order to answer a research question. This can include both data collection through communication with subjects, or secondary analysis of already-collected data.

  – When an investigation is designed to contribute to **generalizable knowledge**, it is intended to provide general conclusions, which can be applied beyond a single individual, population, internal program or organization.
Levels of IRB Review

- If the research does qualify as Human Subjects Research, the IRB must review it
  - However all studies are not the same
  - Broken into 3 levels

Full Board
- More than “minimal risk” to subjects
- Not covered under other review categories
  - Example: interventions involving physical or emotional discomfort or sensitive data

Expedited
- Not greater than minimal risk
- Fits one of the 9 Expedited Review Categories
  - Examples: Collection of biospecimens by noninvasive means, Research with existing documents/record collected for non-research purposes in which subjects are identifiable

Exempt
- Less than “minimal risk”
- Fits one of the 6 Exempt Categories
  - Example: Research with de-identified records, anonymous surveys

*Defined by federal regulation (45 CFR 46)
**Exempt Categories**

There exist 6 categories of research that allow the study to be classified as exempt, and not require continued oversight.

<table>
<thead>
<tr>
<th>EXEMPT CATEGORIES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDUCATION</strong></td>
</tr>
<tr>
<td>1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.</td>
</tr>
<tr>
<td><strong>TESTS, SURVEYS, INTERVIEWS</strong></td>
</tr>
<tr>
<td>2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.</td>
</tr>
<tr>
<td><strong>PUBLIC OFFICIALS</strong></td>
</tr>
<tr>
<td>3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (b)(2) of this section, if: (i) the human subjects are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.</td>
</tr>
<tr>
<td><strong>EXISTING DATA</strong></td>
</tr>
<tr>
<td>4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.</td>
</tr>
<tr>
<td><strong>DEMONSTRATION PROJECTS</strong></td>
</tr>
<tr>
<td>5. Research and demonstration projects which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) Public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.</td>
</tr>
<tr>
<td><strong>FOOD</strong></td>
</tr>
<tr>
<td>6. Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed or (ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture.</td>
</tr>
</tbody>
</table>
There exist 7 (really 9) categories of research that allow for expedited review of the study.

Northwestern IRB
## Existing or Prospective Data Collection

5. Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis). (NOTE: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(4). This listing refers only to research that is not exempt.)

## Recordings

6. Collection of data from voice, video, digital, or image recordings made for research purposes.

## Group Behavior/Surveys, Interviews, Focus Groups, QA

7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies. (NOTE: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(2) and (b)(3). This listing refers only to research that is not exempt.)

## Other Caveats

1. Research activities must present no more than minimal risk to human subjects
2. The expedited review procedure may not be used where identification of the subjects and/or their responses would reasonably place them at risk of criminal or civil liability or be damaging to the subjects= financial standing, employability, insurability, reputation, or be stigmatizing, unless reasonable and appropriate protections will be implemented so that risks related to invasion of privacy and breach of confidentiality are no greater than minimal.
3. The categories in this list apply regardless of the age of subjects, except as noted.
4. The expedited review procedure may not be used for classified research involving human subjects.
## Sensitivity, Data, and Identification

<table>
<thead>
<tr>
<th>Category 2: Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.</th>
<th>Key Differences</th>
<th>EXPEDITED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exemption applies if:</strong></td>
<td><strong>Exemption applies if:</strong></td>
<td><strong>Category 7:</strong> Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.</td>
</tr>
<tr>
<td>- Data is identifiable but not sensitive, or</td>
<td>- Data is sensitive but de-identified.</td>
<td></td>
</tr>
<tr>
<td>- Children are only undergoing educational tests or observation of public behavior without investigator interaction</td>
<td>- Data is identifiable, sensitive and measures are in place to protect confidentiality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Children are to be surveyed, interviewed, or observed with investigator interaction</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category 4: Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.</th>
<th>Exemption applies if:</th>
<th>Category 5: Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for non-research purposes (such as medical treatment or diagnosis). (NOTE: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(4). This listing refers only to research that is not exempt.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exemption applies if:</strong></td>
<td><strong>Exemption applies if:</strong></td>
<td></td>
</tr>
<tr>
<td>- All data is existing, and recorded without identifiers or codes</td>
<td>- Data is existing and identifiers retained, or</td>
<td></td>
</tr>
<tr>
<td><strong>Expedited applies if:</strong></td>
<td>- Data will be prospectively collected for non-research purposes (Identifiers can be kept)</td>
<td></td>
</tr>
<tr>
<td>- Data is existing and identifiers retained, or</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Data will be prospectively collected for non-research purposes (Identifiers can be kept)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*(NOTE: Some research in this category may be exempt from the HHS regulations for the protection of human subjects. 45 CFR 46.101(b)(2) and (b)(3). This listing refers only to research that is not exempt.)*
Privacy and Confidentiality

• Privacy:
  – Having control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others.

• Confidentiality
  – Methods used to ensure that information obtained by researchers about their subjects is not improperly divulged.
Confidentiality

• Confidentiality and anonymity are not the same
  – Anonymous means no one, anywhere, ever can identify individual subjects

• Names are not the only identifiers
  – Brings us to the notion of Protected health information (PHI)

• PHI is any information in the designated record set that can be used to identify an individual and that was created, used, or disclosed in the course of providing a health care service such as diagnosis or treatment.
18 Formal HIPAA Identifiers

- Names
- All geographical subdivisions smaller than a State
  - Including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code,
- All elements of dates (except year) for dates directly related to an individual:
  - Including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- Phone numbers
- Fax numbers
- Electronic mail addresses
- Social Security numbers
- Medical record numbers
- Medical record beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code
  - This does not mean the unique code assigned by the investigator to code the data

https://kb.iu.edu/d/ayyz
Next Class

• Case Study!
  – Groups of 4-6
    • These are the same groups you will be in for assignment 4

  – Intended to be done in class MWF next week
    • You will have distinct goals to meet regarding:
      – Data exploration / Cleaning
      – Modeling