Welcome to Data Privacy in Biomedicine

For CS: You’re sitting in 8396-02
For Informatics: You’re sitting in 7380

When: Mondays and Wednesdays, 3:10 – 4:25pm
Where: Featheringill Hall, Room 211

Office Hours: Upon Request

Contact: b.malin@vanderbilt.edu

http://www.hiplab.org/courses/BMIF380

More to come (projects, homeworks, etc.) – links will be available from front page of website

Course Objectives

- After this course, you should be able to analyze data privacy from three non-exclusive perspectives:
  - Data Detectives: Understand how seemingly private information can be discovered (or exploited) using automated strategies.
  - Data Protectors: Construct privacy protection technologies that provide formal computational guarantees of privacy in disclosed databases.
  - Technology Policy Designers: Develop privacy protection technologies that complement policy regulations.

Expectations

- You are expected to be competent in an object-oriented programming language (Java, C++, Python, ...)
- You are expected to have a working knowledge of the Internet, word processing, and basic databases (Access, Oracle, MySQL, PostGres) and analysis tools (R, Python, Matlab, Scala, Julia, Excel, ...)

Beyond Expectations

- You have experience in
  - information security
  - data structures, algorithms, and statistics
  - public policy and legal frameworks
Grading

- This is a research-oriented course. There are no exams.
- A substantial portion of your grade will be based on your “final” project.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>% of Total Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Project</td>
<td>50%</td>
</tr>
<tr>
<td>Homework Assignments</td>
<td>30%</td>
</tr>
<tr>
<td>Reading Summaries</td>
<td>10%</td>
</tr>
<tr>
<td>Class Participation</td>
<td>10%</td>
</tr>
</tbody>
</table>

Homework Policy

- Unless the assignment calls for a group project, please do your own homework.
- You can discuss the homework with other students, including the ways in which you approach the solutions to the questions, but the final submission must be your own.
- Do not plagiarize without proper attribution – not even in your reading summaries, which leads me to…

Reading Summaries

- There is no textbook for this course.
- Assigned readings will be available the lecture before it is due (at the latest).
- Your summaries should be no more than 1 page in length
- Summaries will be graded on a {✓, ✓, ✓} scale
  - ✓: You skimmed the reading and barely understood its meaning
  - ✓: You read the reading and provided a reasonable account of its contents
  - ✓+: You demonstrated critical reasoning and insight regarding the topic
- Submit summaries to b.malin@vanderbilt.edu before class

Final Projects

- Your project should be an independent study on a data privacy issue, with relationship to the area of biology, medicine, or health more generally
- You may design your own project or choose from a predefined set of topics (will be available on the course website later in the semester)
- Do not be afraid to discuss your project ideas with the instructor!

Sample Topics

- Access Control Frameworks for Distributed Medical Record Systems
- Surveillance of Electronic Medical Record Accesses for Suspicious Behavior
- Evaluation and Design of Privacy Technologies for Personal Health Records (See Microsoft HealthVault Initiative)
- Finding & Relating Publicly Available Repositories of Person Specific Biomedical Information
- Building and Evaluating Clinical Text De-identification Tools
- Anonymization of clinical profiles / sets of diagnoses
- Applications of big data frameworks to sanitizing clinical data
- Applications of security frameworks (e.g., Blockchain)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Due Date</th>
<th>% of Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Proposal</strong>: A one-pager that describes the project area and how you intend to address the research within the confines of this semester. This will be broken down into a several phases.</td>
<td>March 12</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Status Report Presentation</strong>: Briefing for the class on project area and first phase of research. (No more than 5 minutes)</td>
<td>March 25</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Written Project Status Report</strong>: A summary of the progress you have made. (No more than 4 pages)</td>
<td>March 31</td>
<td>10%</td>
</tr>
<tr>
<td><strong>Final Project Presentation</strong>: Showcase of research methods and results. (No more than 15 minutes)</td>
<td>April 22 (last day of class)</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Final Project Report</strong>: This will be in the form of a conference-style paper. It will summarize the research area, your methodology, experience, and contributions of your work.</td>
<td>April 28 (in lieu of final)</td>
<td>25%</td>
</tr>
</tbody>
</table>
**Why Do We Need A Course on Privacy?**

- **Authentication**: login with password, tokens, keys
- **Authorization**: permission and role-based models to read/write data
- **Encryption**: to avoid eavesdropping during transmission and storage

---

**Data Privacy Definitions** (paraphrase Sweeney)

- The study of **computational solutions** for releasing data such that a) the data is practically useful (utility) while b) the aspects of the subjects of the data are not revealed (privacy).

- **Privacy Protection** ("data protectors"):
  - release information such that entity-specific properties (e.g. identity) are controlled
  - restrict what can be learned

- **Data Linkage** ("data detectives")
  - combining disparate pieces of entity-specific information to learn more about an entity

---

**A Visual Perspective**

- **Utility** vs **Privacy**
To ensure utility, you must reveal all the data.

To ensure privacy, you must not reveal any data.

Here Lives Data Privacy

Privacy, Policy, & Preference
- Individuals want control over who can – AND CAN NOT – view their health-related records.

Yoda's Preferences
- Physicians = Yes
- Insurance = Yes
- Researchers = No

Data Collection, Policy, and Privacy
- Can design technology to:
  - Standardize policy specification
  - Inform about data collection
  - Address specific privacy concerns in data sharing
    - Anonymity
    - Confidentiality
    - Solitude
Beyond Policy and Informative Technology

- We cannot always control who gets, and has access to, our information.
- Legally, however, data collectors may be required to maintain your privacy.

Data Collection occurs everywhere, everyday, in all different forms (http://webcams.vanderbilt.edu/)

Featheringill!

https://engineering.vanderbilt.edu/atrium-cam/

More Cameras

Peabody Library Terrace
http://peabody.vanderbilt.edu/about/webcams/peabody_library_terrace_webcam.php

Even More Cameras!
http://peabody.vanderbilt.edu/about/webcams/index.php
http://webcams.vanderbilt.edu/kissam/

Biomedical Information

- Not quite in the public
- But... information is shared for various purposes in various contexts
- How do you protect privacy of corresponding individuals?
Schedule

- Let’s look at the syllabus.

Privacy Policy & the Law (Week 1)

- Privacy ideologies & frameworks
  - Who gets to collect information?
  - When is health information shared?
  - How is health information reused and why?

Access Control & Roles (Week 2)

- Access Control
  - Who gets to see information when?
  - Roles, job functions, & permissions
  - Formal representations

- What Constitutes a “Good” Role
  - Representation of organizational behavior
  - Grouping users based on legacy knowledge provided by system administrators

Martin Luther King Day

- No class on January 21

Identifiability & A Whole Lot of Data (Week 3)

- How can we suppress “identifiers” from data?
  - Structured data (e.g., database tables)
- Why does clinical treatment context influence de-identification strategies?

Identifiability & A Whole Lot of Data (Week 4)

- When can we find what was suppressed?
  - Using sample & population statistics to identify
  - Techniques to compute distinguishability

- It’s hard to protect health information… simply because there’s so much data
  - And there’s a lot more data than what’s in the medical record!
  - We’ll use Social Security Numbers as an example
Auditing (Weeks 5 and 6)

- Medical Records & Audits
  - Who looked at my medical record?
  - Should they be looking?
  - How do we construct machine learning strategies that make sense?

De-identification & Scrubbing Narratives (Weeks 6 and 7)

- How can we detect and suppress "identifiers" from unstructured data (e.g., clinical narratives)?
- Welcome to the wonderful world of natural language processing

Record Linkage (Weeks 7 and 8)

- Given all the data, how can we link it?
  - Look at "deterministic" methods, such as rules
    - What are the idiosyncrasies in the health domain that allow them to work? Enable them to fail?
  - Look at probabilistic methods based on frequentist and Bayesian statistics

Anonymization (Weeks 8 and 9)

- If de-identification fails, can we provably protect identity?
  - Yes we can! Welcome to formal models of anonymization
  - We’ll start with k-based models
    - Guarantee every shared record corresponds to at least k people
    - Efficient algorithms to achieve this goal
    - Various "types" of data

Spring Break!!!

- No class on March 4 or 6

Moving Beyond Anonymity (Weeks 9 - 11)

- Hiding in a crowd doesn’t always protect sensitive knowledge
  - How to design algorithms to protect against homogeneity and inference attacks.
  - We’ll look at the identifiability concerns associated high-dimensional data with a focus on:
    - genetic information
      - Statistical methods for identification
      - Strategies for anonymization of DNA data
Ethical Reasoning and Privacy (Week 10)

- When should you publish on privacy ad vulnerabilities?
- Should you disseminate re-identification software or findings?

Privacy Preserving Data Mining (Week 12)

- You have data. I have data. We all have data. How can be combine data to reveal results, but no individual records?
  - We’ll look at cryptographic methods for secure multiparty computation.
- Consider “horizontal” (different people different place) vs. “vertical” (same person, different place) partitioned data systems

Epidemiology and Geospatial Privacy (Week 13)

- Location data is shared for various purposes, but too much granularity can lead to identification
- How does identification occur?
- What anonymization strategies work for geocoded and spatial data? When?

Social Media (Week 13)

- How do people share health information online?
- Who do they share information about?
- Why do they share it?
- How can we automatically find such information?

Image and Video Privacy (Week 14)

- Images are everywhere in healthcare
- Video is becoming more prevalent
- How can we remove identifiers from JPEG, MPEG and other multimedia?

Final Project Presentations! (Week 15)

- The students are in control
- You’ll be graded by a committee of special reviewers
Readings for Next Lecture

  - http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm

- Department of Health and Human Services Summary of the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA)

- Optional
    - http://content.healthaffairs.org/content/28/2/416.full.pdf+html