Bradley Malin, PhD (b.malin@vanderbilt.edu)
Associate Professor of Biomedical Informatics, School of Medicine
Associate Professor of Computer Science, School of Engineering
Vanderbilt University
January 11, 2016
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 313

Office Hours: Upon Request
Contact: b.malin@vanderbilt.edu

Your Professor is Brad Malin
- BS, Molecular Biology
- MS, Computer Science (Data Mining)
- MPhil, Public Policy & Management
- PhD, Computer Science (Computation, Organizations & Society)

Faculty Member: DBMI (1st), EECS (2nd)
Directs: Health Data Science Center (hiplab.org)
Directs: Health Information Privacy Laboratory (hiplab.org)

Sample Research Areas
- Medical Record Access Control, Mining, and Modeling
- Anonymization of Medical & Genomic Data
- Service-Oriented Clinical Information Systems Design
- Big Data Record Linkage

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, Matlab, Scala, Excel)

More to come (projects, homeworks, etc.) – links will be available from front page of website
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
- Fast Graph-Based Approaches to Data Re-identification
- Building and Evaluating Clinical Text De-identification Tools
- Anonymization of clinical profiles / sets of diagnoses
### Final Projects

<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</td>
<td>March 20</td>
<td>5%</td>
</tr>
<tr>
<td>the project area and how you intend to address the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>research within the confines of this semester.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Status Report Presentation:</strong> Briefing for the</td>
<td>March 30</td>
<td>5%</td>
</tr>
<tr>
<td>class on project area and first phase of research.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Written Project Status Report:</strong> A summary of</td>
<td>April 3</td>
<td>10%</td>
</tr>
<tr>
<td>the progress you have made. (No more than 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pages).</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final Project Presentation:</strong> Showcase of</td>
<td>April 25</td>
<td>5%</td>
</tr>
<tr>
<td>research methods and results. (No more than 15</td>
<td>(last day of</td>
<td></td>
</tr>
<tr>
<td>minutes)</td>
<td>class)</td>
<td></td>
</tr>
<tr>
<td><strong>Final Project Report:</strong> This will be in the form</td>
<td>May 1 (in</td>
<td>25%</td>
</tr>
<tr>
<td>of a conference-style paper. It will summarize the</td>
<td>lieu of final)</td>
<td></td>
</tr>
<tr>
<td>research area, your methodology, experience, and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>contributions of your work.</td>
<td></td>
<td></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

### Security for 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

- **But Data Can Re-identify!**

- Can I see some anonymous data?

- **Ah! I know who this is!**
A Visual Perspective

**Utility**

**Privacy**

To ensure utility, you must reveal all the data.

---

A Visual Perspective

**Utility**

**Privacy**

To ensure privacy, you must not reveal any data.

---

A Visual Perspective

**Utility**

**Privacy**

Here Lives Data Privacy.

---

Privacy, Policy, & Preference

- Individuals want control over who can – AND CAN NOT – view their health-related records.

Privacy:

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

Policy:

- Hospital Records
- Yoda Data

Preference:

- Physicians
- Insurance
- Researchers
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 can not always control who gets, and has access to, our information
- Legally, however, data collectors may be required to maintain your privacy

Enabling Personal Control Through Technology

- Preventing Data Collection in Public Spaces
  - [http://www.appliedautonomy.com/see.html](http://www.appliedautonomy.com/see.html)

Beyond Policy and Informative Technology

- We can not 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/](http://webcams.vanderbilt.edu/))

Featheringill!

More Cameras

- [Peabody Library Terrace](http://peabody.vanderbilt.edu/about/webcams/peabody_library_ terrace_webcam.php)
- [Even More Camerass](http://peabody.vanderbilt.edu/about/webcams/index.php)
  - [http://webcams.vanderbilt.edu/kissam/](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

Auditing (Week 3)

- Medical Records & Audits
  - Who looked at my medical record?
  - Should they be looking?
  - How do we construct machine learning strategies that make sense?
Identifiability & A Whole Lot of Data
(Weeks 4 - 5)

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

De-identification & Scrubbing Narratives
(Week 6)

- 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 (Week 7)

- 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 (Week 8 - 9)

- If de-identification fails, can we provably protect identity?
  - Yes we can! Welcome to formal models of anonymization

  - We'll be looking at 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!!!
Ethical Reasoning and Privacy (Week 9)

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

Advanced Topics in Anonymization (Week 10 - 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

Image and Video Privacy (Weeks 12 - 13)

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

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?

Privacy Preserving Data Mining (Week 14)

- You have data. I have data. We all have data. How can we 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

Private Record Linkage & Information Retrieval (Week 14)

- Multiparty computation isn’t always efficient
- We’ll look at two special problems from earlier in the semester and see if they can be solved with relaxed assumptions of the “adversaries”
- We’ll define the conditions under which these solutions work and show why.
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/sglass/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