

## Data Privacy and Biomedicine Syllabus

### Data Privacy in Biomedicine (BMIF-380 / CS-396)

**Instructor: Bradley Malin, Ph.D. (b.malin@vanderbilt.edu)**

**Semester: Spring**

**Time: Tuesdays & Thursdays, 2:35 – 3:50pm**

**Location: Featheringill Hall, Room 313**

**Website: <http://people.vanderbilt.edu/~b.malin/BMIF380/index.html>**

**Office Hours: TBD**

### DESCRIPTION

The integration of information technology into biomedical environments has enabled unprecedented advances in the collection, storage, analysis, and rapid dissemination of patient-specific data to physicians and researchers. Given the potential wealth of such detailed databases for further advances in healthcare, many organizations share, or anticipate sharing, their collections for various purposes related to quality assurance, public health, and research. However, in the face of today's complex networked environments, many organizations are finding it increasingly difficult to share biomedical data due to concerns about patient privacy and anonymity. For instance, how can we share patient-specific data without revealing the identity of the patient? Security practices, such as role based access control and encrypted communications ensure authentication and secure communications, but they do not prevent the leakage of inferences from the data after it has been accessed or transmitted. Thus, this course is concerned with the analysis and protection of data privacy. The goal of this course is to introduce students to the computational challenges, as well as formal privacy protection solutions, for data privacy in healthcare and biomedical research environments. The topology of data privacy is a highly interdisciplinary landscape and material in this course will touch on issues and methodologies from bioinformatics, cryptography, data mining, databases, distributed systems, law, machine learning, medical informatics, policy, and statistics.

### OBJECTIVES

After this course, students will be able to analyze data privacy issues from three non-exclusive perspectives:

1. *Data Detectives*: Oftentimes data is shared with false beliefs about privacy and data protection. From this perspective students will learn how seemingly private information, can be learned using automated strategies.
2. *Data Protectors*: Students will learn how to construct privacy protection technologies that provide formal computational guarantees of privacy in data collection and sharing.
3. *Technology Policy Designers*: Computational models provide a basis for protection, but in order to implement such technology in the real world, it must support, and not circumvent, existing policy specification. From this perspective, students will learn how to develop privacy protection solutions which complement policy regulations.

### PREREQUISITES

*Required:* Students are expected to have proficiency in designing and writing software programs. There is no programming language requirement for this class,

## Data Privacy and Biomedicine Syllabus

though experience with object orientation is beneficial.

*Recommended:* Students should be comfortable with learning about basic statistics, data structures, and algorithm analysis. When appropriate, quantitative and computational methodology will be reviewed. Knowledge of, and prior experience with, security principles is NOT a prerequisite for this course.

### GRADING

Criteria	Percent of Grade
Project	50%
<i>(Initial Proposal)</i>	(5%)
<i>(Status Report)</i>	(15%)
<i>(Final Report &amp; Presentation)</i>	(30%)
Homework Assignments (3 assignments, 10% each)	30%
Reading Summaries	10%
Class Participation	10%
	100%

**Required Reading Assignments:** There is no primary textbook for this course. Reading assignments will be selected from various periodicals. Students will be required to read and submit brief summaries of assigned readings. Your summaries should be no longer than one page in length. Readings will be made available online or as in-class handouts at least one lecture before they are due. Your summaries will be graded on a {check-minus, check, check-plus} scale.

- ✓- : You skimmed the assigned reading and barely understood, or summarized, its meaning and implications.
- ✓: You demonstrated that you read the material by providing a reasonable account of its contents, its strengths, and weaknesses.
- ✓+ : You provided a critical assessment of the reading and show insight regarding the reading's topic.

Email your summaries to [b.malin@vanderbilt.edu](mailto:b.malin@vanderbilt.edu) before the beginning of class.

**Project:** In lieu of a final exam, each student must complete an independent project on a data privacy issue in biomedicine. Projects should investigate a topic of interest to the student, and must demonstrate analysis and critical thinking in data privacy. The project will require a significant commitment and contribute to a substantial part of the final grade. A list of sample project topics will be made available and reviewed in class.

**Honesty Policy:** From the Vanderbilt Student Handbook, "HONESTY is a commitment to refrain from lying, cheating, and stealing. Recognizing that dishonesty undermines community trust, stifles the spirit of scholarship, and threatens a safe environment, we expect ourselves to be truthful in academic endeavors, in relationships with others, and in pursuit of personal development." You are permitted, and encouraged, to discuss homework assignments with other students. However, you must do your own work and submit your own solutions.

## Data Privacy and Biomedicine Syllabus

### TOPIC AND SCHEDULE OVERVIEW (*Tentative and Subject to Change*)

#### **Week 0 (Jan 10): Course Overview and Introduction to Data Collection and Privacy**

What is data privacy? How does it relate to data security principles, such as authorization, access control, and authentication?

#### **Week 1 (Jan 15, 17): Policy, De-identification, Re-identification**

In the first part of this week, we will review the various ideologies, legal, and policy precedents for privacy in modern healthcare environments and society. Who collects medical information and when do patients have control over their privacy? Can policy and specification of privacy protections be automated?

Many privacy regulations and policies protect patient privacy through the “de-identification” of data. This week, we will look into what de-identification entails and how it relates to “anonymity”. In the second part of this week, students will learn how to characterize uniqueness in data, both at elemental and population levels of granularity.

#### **Week 2 (Jan 22, 24): Availability of Personal Information and Identifiers**

Personal information is available in many different resources both on- and offline. Where is this information? How do we automatically capture and organize it for privacy assessments? This week we will look at various information repositories, such as vital records and statistics (including birth records, death records, marriage records, court documents) and the Social Security Death Index. We will also discuss issues such as the potential of unique numbers for persistent patient identifiers and the history of the Social Security Number.

#### **Week 3 (Jan 29, 31): Record Linkage**

This section of the course will present concepts and methodology associated with the linkage of data in disparate databases. Methods will be drawn from deterministic and probabilistic frameworks. We will also discuss how linkage methods can be automated and their application within electronic medical record systems.

#### **Week 4 (Feb 5, 7): Trails and Graph-Based Approaches to Privacy**

People leave information behind in many different organizations. Simple methods of data protection and de-identification appear sufficient to protect information from privacy compromise. However, simple automated strategies can be constructed to link information across databases. This week will discuss the evolution of the “trail” re-identification attack, which will be used to illustrate a formal model of data re-identification. At this point, we will explore how graph-based modeling can be applied to represent privacy problems in general.

#### **Week 5 (Feb 12, 14): Text Scrubbing**

A large amount of information collected in biomedical setting is in the form of free text: doctor’s notes, laboratory reports, discharge reports, and more. How can we de-identify text information? Can we ever achieve “anonymized” text? This section of the course will review various methods and software for the discovery and replacement of personal identifiers.

#### **Week 6 (Feb 19, 21): Formal Models of Anonymity**

This week will look into formal models of anonymity protection, such as  $k$ -map and  $k$ -anonymity. We will also look at ways in which formal models can be satisfied through

## Data Privacy and Biomedicine Syllabus

computational transformations of data, such as generalization, suppression, and aggregation. We will study the computational complexity of the anonymizing data with minimal changes to the data and review various heuristics and strategies for data protection.

### **Week 7 (Feb 26, 28): Genomic and Familial Databases**

As high-throughput technologies become further ingrained in the clinical environment, the collection and sharing of biological information, such as DNA data, is becoming more common. In this section of the course, we will investigate ways in which patient identity in genomic data is protected, how it is re-identified and how it can be formally protected. The latter part this week will be dedicated to data privacy issues associated with the collection and sharing of information for population-based investigations. Does this information pose a threat to privacy and if so, how can we formally prevent such a threat?

### **SPRING BREAK - March 4, 6**

### **Week 8 (March 11, 13): Ethical and Legal Issues in Medical Data Privacy**

In the first part of this week, Prof. Ellen Wright-Clayton will provide an overview and specific examples of legal issues and cases in medical information privacy.

Then, in the second part of this week, we'll consider some of the ethical issues in the application of data privacy technologies. Simply because you can build a re-identification technology, doesn't mean that you should use it, does it?

### **Week 9 (March 18, 20): Image and Video Privacy**

Video is increasingly used for monitoring and surveillance in health care environments, such as managed care facilities. This week we will investigate several procedures and principles for removing personally identifying features, e.g., an individual's face, from video streams. We will also investigate how images, e.g., the picture of a face, are a special case of video streams, can be protected using formal models of anonymity.

### **Week 10 (March 25, 27) Privacy Preserving Biosurveillance and Geospatial Information // Project Status Report Presentations**

Public health and epidemiology require geographic information regarding the presence of clinically interesting cases to detect potential outbreaks and bioterrorist activities. However, the sharing of geographic and spatiotemporal information may lead to re-identification. This week we will investigate various approaches by which such information may be protected during data sharing.

The second part of this week will be dedicated to student projects. Students will write a short summary of their problem statement, initial research design, and make a short presentation on the status of their projects for an in-class evaluation.

### **Week 11 (April 1, 3): Secure Multiparty Computation and its Applications to Distributed Data Mining**

The traditional application of cryptography is framed from a two-party viewpoint in which two participants, Alice and Bob, exchange information, such as a patient's medical record, over an unsecured channel. An extension to the traditional model is secure multiparty computation (SMC), which is concerned with the interaction of two or more

## Data Privacy and Biomedicine Syllabus

participants that need to exchange information to construct a result without revealing private information. This week we will look at how secure multiparty computation is used in the context of data mining across a set of data holders with different privacy constraints.

### **Week 12 (April 8, 10): Private Record Linkage**

How can we integrate information on people with revealing their identifying information. This question has been studied for years in biomedical environments and beyond. Various solutions have been proposed, including one-way hashing, keyed encryption, and oblivious transfers. In this week's lectures, we will review how obscured personal identifiers can be compared.

### **Week 13 (April 15, 17): Trail Anonymization / Student Final Presentations**

In this section of the course, students will learn how to integrate formal privacy models with private record linkage to thwart re-identification problems. To demonstrate this approach, we will study how the trail re-identification problem can be addressed in a real world environment.

The final lecture will be dedicated to students' presentations on their final projects.