Healthcare Analytics

CSE 40817
About Me
Healthcare Analytics

CSE 40817
2010-2018
What *is* Healthcare Analytics?
What *is* Healthcare Analytics? Informatics?
Defining a Field

- "Medical Informatics is the science and art of modeling and recording real-world clinical concepts and events into computable data used to derive actionable information, based on expertise in medicine, information science, information technology, and the scholarly study of issues that impact upon the productive use of information systems by clinical personnel." - S. Silverstein.

- "The study, invention and implementation of structures and algorithms to improve communication, understanding and management of medical information." - Homer Warner.

- “the science of using system-analytic tools … to develop procedures (algorithms) for management, process control, decision making and scientific analysis of medical knowledge." - Ted Shortliffe

- Studies the organization of medical information, the effective management of information using computer technology, and the impact of such technology on medical research, education, and patient care. The field explores techniques for assessing current information practices, determining the information needs of health care providers and patients, developing interventions using computer technology, and evaluating the impact of those interventions. This research seeks to optimize the use of information in order to improve the quality of health care, reduce cost, provide better education for providers and patients, and to conduct medical research more effectively." - Stephen Johnson
Even More Definitions

• “The field that concerns itself with the cognitive, information processing, and communication tasks of medical practice, education, and research, including the information science and the technology to support these tasks.” - Greenes and Shortliffe

• “biomedical informatics as the scientific field that deals with biomedical information, data and knowledge - their storage, retrieval and optimal use for problem solving and decision making.” - Shortliffe and Blois

• “…comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and health care.” Van Bemmel

• “[i]n medical informatics we develop and assess methods and systems for the acquisition, processing, and interpretation of patient data with the help of knowledge that is obtained in scientific research.” - Musen and van Bemmel
A Long History
Our Definition of the Course

“The interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving and decision making, motivated by efforts to improve human health.”

– American Medical Informatics Association
AMIA (2012)
So definitions are great, but definitions don’t explain why need healthcare analytics
Experiential Medicine
Treatment Variability
Evidence-Based Medicine
Where Does Evidence Come From
Paper-Based Paradigm
Digital Information
Digital Information
Information Overload

- **Genomic Data**: ~3GB
- **Estimated average hospital will generate 665 Tb of data.**
- **Monitors generate 1000+ Readings/sec**
- **Medical image archives are increasing by 20-40% annually.**
- **80% coding variability among diagnosis and lab tests**
- **4.9 Million remote monitoring devices**
- **X-Ray ~30MB**
- **The number of hospitals using health information technology has more than doubled in the last two years**
Emergence of Informatics

- Health Informatics
- Distributed Computing
- Advancing Architectures
- Data Processing
- Computer Science
- Hypothesis Testing
- Statistical Inference
- Experimental Design

Healthcare
- Biological Systems
- Clinical Factors
- Population Health

Biostatistics

Math and Statistics
- Hypothesis Testing
- Statistical Inference
- Experimental Design

Health Information Technology

Data Science and Machine Learning
Health Informatics

Clinical
- Disease Prediction
- Imaging Analysis
- Decision Support

Research
- Drug Discovery
- Population Profiling
- Clinical Trial Matching

Population Health
- Disease Surveillance
- Intervention Planning
- Resource Allocation

Healthcare
- Biological Systems
- Clinical Factors
- Population Health

Biostatistics

Health Information Technology

Math and Statistics
- Hypothesis Testing
- Statistical Inference
- Experimental Design

Computer Science
- Distributed Computing
- Advancing Architectures
- Data Processing

Data Science and Machine Learning

Administrative
- Fraud Detection
- Staffing
- Infection control
An Exploding Industry

106 Startups transforming healthcare with AI
Current Field According to AMIA

- Biomedical informatics (BMI)
- Education and research
- Methods, techniques, theories
  - Bioinformatics and structural (imaging) informatics
  - Health informatics (HI): clinical informatics and public health informatics
- Informatics in translational science:
  - Translational bioinformatics (TBI) and clinical research informatics (CRI)

Basic research

Applied research and practice

Molecules, cells, tissues, organs

Patients, individuals, populations, societies
Our Course
Analytics vs. Informatics
General Course Goals

This course will provide an overview to many of the concepts, techniques, and theories associated with analytics in the healthcare domain.

• **Data**
  – Identify and address common quality issues in health data
  – Utilize statistical methods to explore and compare data from electronic health records

• **Modeling**
  – Identify proper analytical techniques to address various research questions
  – Programmatically apply computational models to electronic health record data

• **Implications and Ethics**
  – Describe the ethical, and privacy implications of working with health data.
Two Key Areas

• **Descriptive Tasks**
  – Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that are able to summarize the underlying relationships in data. Descriptive tasks are often exploratory in nature and frequently require additional work to explain the results.

• **Predictive Tasks**
  – The objective of these tasks is to learn (model) a relationship between attributes with the intent to predict the value a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.
Course Topics

Preliminaries
Data Understanding
Data Preprocessing
Statistical Methods
Classification & Regression
Validation & Interpretation
Advanced Topics
Course Topics

1. Data Understanding
   Types of data; classes and attributes; interactions among attributes; relative distributions; summary statistics; data visualization

2. Data Preprocessing
   Addressing noise and outliers; Standardizing data; sampling data; feature selection and using principal components to eliminate attributes

3. Statistical Methods
   Relationships between variables; Uncertainty and confidence intervals; Distributions; 1, 2, and 3 group parametric and non-parametric statistical tests; multiple comparisons post-hoc tests
Course Topics

Classification and Regression
Types of data; information and uncertainty; classes and attributes; interactions among attributes; relative distributions; summary statistics; data visualization

Validation and Interpretation
Concepts and methods of validation and testing data; Cross-validation; effect sizes; sensitivity analysis; performance metrics (error, ROC curves, lift curves)

Advanced Topics
Population health data; Ethics and human subjects research; Touch on: Physiological waveforms; Networks; State of the art machine learning models
What this Course Is Not and What it Will Be

- Biostatistics
- Health Informatics
- Math and Statistics
- Computer Science
- Data Science and Machine Learning
- Healthcare

Biological Systems
Clinical Factors
Population Health

Distributed Computing
Advancing Architectures
Info Processing

Biostatistics
Hypothesis Testing
Statistical Inference
Experimental Design

Health Information Technology

Data Science and Machine Learning

30
What this Course Is Not and What it Will Be

**Biostatistics**
- Hypothesis Testing
- Statistical Inference
- Experimental Design

**Math and Statistics**
- Hypothesis Testing
- Statistical Inference
- Experimental Design

**Computer Science**
- Distributed Computing
- Advancing Architectures
- Info Processing

**Health Information Technology**

**Healthcare**
- Biological Systems
- Clinical Factors
- Population Health

**Data Science and Machine Learning**
Preliminaries

It is not... Bioinformatics
It is not... Image Analysis
It is not... Data Science / ML

\[ \delta^{(l-1)} = \frac{\partial e(w)}{\partial s_j^{(l-1)}} \]
\[ = \sum_{j=1}^{d^{(l)}} \frac{\partial e(w)}{\partial s_j^{(l)}} \times \frac{\partial s_j^{(l)}}{\partial x_i^{(l-1)}} \times \frac{\partial x_i^{(l-1)}}{\partial s_i^{(l-1)}} \]
\[ = \sum_{j=1}^{d^{(l)}} \delta_j^{(l)} \times w_{ij}^{(l)} \times \theta'(s_i^{(l-1)}) \]
Will Be: Hands On Experience

Python

NumPy
SciPy
matplotlib

pandas

\[ y_{it} = \beta x_{it} + \mu_i + \varepsilon_{it} \]

StatsModels
Statistics in Python

Jupyter
Will be: A chance to use real health data
Will Be Interactive
# Syllabus and Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Friday</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Semester Start</td>
<td>Welcome - Course Overview / What is Healthcare Analytics?</td>
<td>Types of Healthcare Data / Introduction to EMR data</td>
<td>Assignment 1 Out Fri</td>
</tr>
<tr>
<td>(Aug 20-24)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>Data Understanding</td>
<td>Summary Statistics / Visualization</td>
<td>Lab - Introduction to MIMIC / EMR Data</td>
<td>Assignment 1 Due Fri</td>
</tr>
<tr>
<td>(Aug 27-31)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 3</td>
<td>Data Prep - Noise and Outliers</td>
<td>Data Prep – Reduction and Transformation</td>
<td>Lab – New Data Example</td>
<td></td>
</tr>
<tr>
<td>(Sept 3-7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 4</td>
<td>Statistics – Review basics</td>
<td>Statistics – common statistical techniques</td>
<td>Statistics – Advanced topics</td>
<td>Assignment 2 Out Fri: Stats</td>
</tr>
<tr>
<td>(Sept 10-14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 5</td>
<td>Quiz 1 / Linear Models 1 (Regression)</td>
<td>Linear Models 1 (Regression)</td>
<td>Lab – model</td>
<td></td>
</tr>
<tr>
<td>(Sept 17-21)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 6</td>
<td>Linear Models 2 (Logistic Regression)</td>
<td>Linear Models 3 (Mixed Effects)</td>
<td>Lab – ICU Mortality Risk</td>
<td>Assignment 2 due Mon</td>
</tr>
<tr>
<td>(Sept 24-28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 7</td>
<td>Temporal Modeling 1 (Time Series)</td>
<td>Temporal Modeling 1 (Time Series)</td>
<td>Lab – Time Series</td>
<td></td>
</tr>
<tr>
<td>(Oct 01-05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 8</td>
<td>Temporal Modeling 2 (Survival Analysis)</td>
<td>Temporal Modeling 2 (Survival Analysis)</td>
<td>Lab Survival Analysis Project Proposal Due</td>
<td>Assignment 3: Out Mon Linear and Temporal Modeling</td>
</tr>
<tr>
<td>(Oct 08-12)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fall Break</td>
<td>Fall Break</td>
<td>Fall Break</td>
<td>Fall Break</td>
<td></td>
</tr>
<tr>
<td>(Oct 15-19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Syllabus and Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Friday</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 9</td>
<td>Natural Language Processing</td>
<td>Natural Language Processing <strong>Quiz 2</strong></td>
<td>Lab: Clinical Notes</td>
<td></td>
</tr>
<tr>
<td><strong>Oct 22-26</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 10</td>
<td>Validation</td>
<td>Sensitivity Analysis</td>
<td>Lab – Sensitivity Analysis</td>
<td><strong>Assignment 3 due Mon</strong></td>
</tr>
<tr>
<td><strong>Oct 29-Nov 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 11</td>
<td>Advanced Topics: Neural Networks</td>
<td>Advanced Topics: Network analysis</td>
<td>Advanced Topics: TBD</td>
<td><strong>Milestone 1 Due Fri</strong></td>
</tr>
<tr>
<td><strong>Nov 05-09</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 12</td>
<td>Population Health Data</td>
<td>International Health: Intro</td>
<td>Quiz 3 Lab: Project Check In</td>
<td><strong>Assignment 4 Out Fri</strong></td>
</tr>
<tr>
<td><strong>Nov 12-16</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 13</td>
<td>Data Privacy and Ethics - IRB</td>
<td>Thanksgiving Holiday</td>
<td>Thanksgiving Holiday</td>
<td></td>
</tr>
<tr>
<td><strong>Nov 19-23</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 14</td>
<td>MIMIC Case study</td>
<td>MIMIC Case study</td>
<td>MIMIC Case study</td>
<td><strong>Milestone 2 Wed</strong></td>
</tr>
<tr>
<td><strong>Nov 26-30</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 15</td>
<td>Waveforms</td>
<td>Course Review Project Questions</td>
<td>Reading Day</td>
<td><strong>Assignment 4 due Mon</strong></td>
</tr>
<tr>
<td><strong>Dec 03-07</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Course Project

The primary deliverable of this course will be a semester-long course project, bringing together and apply the various topics covered in class.

• The goal of the project is to go through the complete analytics process (sometimes called the knowledge discovery process) to answer one or more questions about a topic of your own choosing.

• Done in groups of 2-3, the project will be comprised several deliverables throughout the semester.

• More information is available on the webpage.
Grading

• 45% Class Project
  – The 45% will be split across several deliverables throughout the semester.

• 20% Assignments (4 assignments, at 5% each)

• 15% Quizzes
  – As this course will cover an extensive number of topics related to the field of healthcare analytics, we will have 3 (non-cumulative) quizzes throughout the semester (3 quizzes at 5% each)

• 10% Case Study
  – The course will conclude with a case study for students to work through the process of answering a clinical question. The process will include identifying and extracting relevant data in an EMR, cleaning it, and analyzing it in such a manner to provide a justifiable answer to the question.

• 10% Participation
  – Attendance during the week and engagement in the Friday lab sessions will contribute to the students overall participation grade.

No Final or Midterm 😊
Textbook

- No required textbook
- Some useful references:
  - Health Analytics / Modeling:
    - Health Data Analytics
    - Clinical prediction Models
  - Toolset:
    - Python for Data Analysis (2nd edition)
Course Logistics

• Lecture:
  – 10:30-11:20 am (Monday / Wednesday / Friday), DeBartlo Hall 125

• Office hours: 384 Nieuwland Science Hall
  – Wednesday 1 PM to 3 PM
  – Thursday 10 AM to 12 PM
  – By Appointment (please feel free to email for alternative times)

• Course Website (slides, assignments, project information):
  – https://www3.nd.edu/~kfeldman/cse40817.fa18/www/home.php

• Course Communications
  – Slack (Invites to come out this week)
  – Sakai (Assignment submission)
Next Class – Types of Health Data

[Diagram showing various types of health data attributes, including aggregate, administrative, wellness, clinical, demographic, and omics attributes.]

- Aggregate Attributes
  - Financial
  - Payments
  - Claims
  - Connected Devices
  - Cost
- Administrative Attributes
  - Agency for Healthcare Research and Quality
  - Centers for Medicare and Medicaid
- Wellness Attributes
  - Social Media
  - Network
  - Diagnoses
  - Procedures
  - Lab Exams
  - Medications
  - Documentation
  - Imaging
  - Patient Satisfaction
- Clinical Attributes
  - Unstructured
  - Structured
- Demographic Attributes
  - Genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - Lipidomics
- Omics Attributes