Tackling The Challenges of Big Data

Big Data Analytics

John Guttag

Professor

Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Analytics

Applications - Medical Outcomes

Introduction

John Guttag

Professor

Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

My Research Group

- Technical areas
  - Machine learning and data mining
  - Algorithm design and signal processing
  - Computer vision
  - Software systems
- Application areas
  - *Medical analytics*
  - Financial analytics
  - Sports analytics
  - Mobile telemedicine
- Principal funding
  - Quanta Computer, NSF, NIH NSERC, QCRI, MSR, MLB.com

© 2014 Massachusetts Institute of Technology

Medical Analytics

- Building predictive models using machine learning
- From signals to bio-markers
- Video-based monitoring & diagnoses





© 2014 Massachusetts Institute of Technology

The Good News

- Capacity to gather medically significant data growing quickly
- Better instrumentation (e.g., MRI machines, ambulatory monitors, cameras) generates more information/patient

Pictures removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

- More storage capacity allows information to be saved
- Economic and social forces creating more aggregation of data

© 2014 Massachusetts Institute of Technology

Problems Posed by Increase in Available Data

- Clinicians spending ever more time
  - Studying data about their patients
    *And still ignoring most of the data
  - Tracking onslaught of new medical data
    *And not keeping up

© 2014 Massachusetts Institute of Technology

Problems Posed by Increase in Available Data

- Researchers using analytical techniques that:
  - Don't scale to multi-modal data, thousands of variables, millions of patients
  - Designed to test hypotheses, not to uncover new knowledge
  - Require prohibitively expensive and time consuming clinical studies

© 2014 Massachusetts Institute of Technology

Using ML to Make Useful Predictions

- Predicting medical outcomes is hard
- It will be many years before biology provides answers
- But there is lots of "found data" that can be used to derive pretty good predictions
- One focus is predicting avoidable complications

© 2014 Massachusetts Institute of Technology

Accurate Predictions Can Help

- Better decisions about which treatments (if any) to choose
  - E.g., surgery or drugs
- Better management of medical providers
  - Match right patients with right providers, streamline utilization of resources
- Understanding what goes into making a good prediction can inform changes
  - E.g., prophylactic treatments

© 2014 Massachusetts Institute of Technology

The Big Data Challenge

- Too Big: images, videos, signals
  - 10's of millions of patients, billions of bits/patient
  - Parallelism will help, but not if we use n^3 algorithms
- *Too Hard*
  - Multiple modalities: signals, lab results, images, genomic, natural language …
  - Always incomplete, often incorrect, outcomes ambiguous
  - Ground truth often hard to come by
  - Site-to-site variation
- One saving grace: human physiology and medical practice changes a lot more slowly than financial markets
  - What we learn likely to be of long-lasting value

©John Guttag

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics
Applications - Medical Outcomes
Introduction

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Module: Big Data Analytics

John Guttag

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Analytics

Applications - Medical Outcomes

John Guttag

Professor

Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

Computationally-Generated Biomarkers

- Biomarker (broadly defined)
  - A characteristic that can be objectively measured and provides an indicator of normal or pathological processes or expected responses to a therapeutic intervention
- Computationally-Generated Biomarker
  - A biomarker generated by applying computation to medical data
- One current project
  - Cardiovascular risk stratification

 © 2014 Massachusetts Institute of Technology

Cardiovascular Risk Stratification

- Acute coronary syndrome (ACS) common: ~1.25M/ year in U.S.
  - 15% - 20% of these people will suffer cardiac-related death within 4 years
- Have lots of good treatments

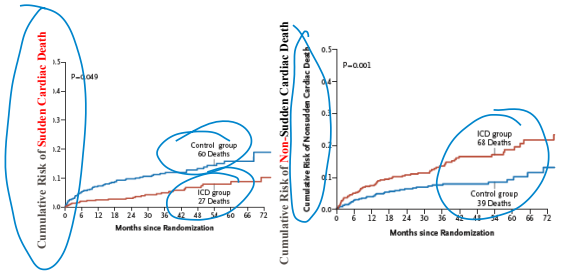
Eat better Get an ICD

Pictures removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

  - Choosing who should get which (or any) is hard
  - Fine-grained risk stratification the key

 © 2014 Massachusetts Institute of Technology

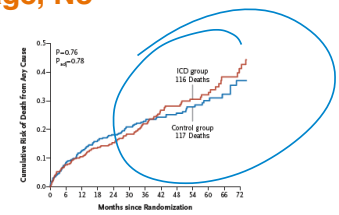
Do ICD's Save Lives?



NEJM Oct. 2009

© 2014 Massachusetts Institute of Technology

On Average, No



- Too many: Potentially risky, always expensive (~$50k)
  - 90% of recipients received < 0 medical benefit
- Too few: 100's of deaths/day potentially avoidable
- Need to do a better job of matching patients with therapies

© 2014 Massachusetts Institute of Technology

Approaches to Identifying High Risk Cases

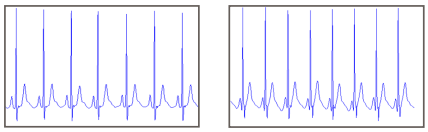
- Clinical characteristics
  - E.g., gender or high blood pressure
- Traditional biomarkers
  - E.g., cholesterol levels
- Echocardiography
  - E.g., LVEF
- Electrocardiography (ECG)
- We combine all of these
  - Innovation on use of ECG

Jason Crow/The Human Face of Big Data

© 2014 Massachusetts Institute of Technology

Morphological Variability

- Detect signs of small and transient electrical instability in myocardium by detecting minor differences in shape of normal appearing heart beats
- Invisible to human eye



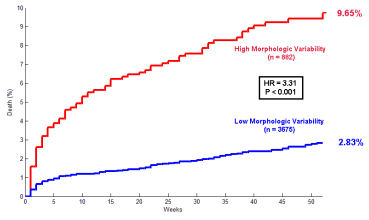
Patient 19919 (died)
(MV > 2 x high risk threshold)

Patient 1593 (survived)
(MV < 1/2 x high risk threshold)

© 2014 Massachusetts Institute of Technology

Data from MERLIN TIMI-36 dataset (~1B beats)

– About 4,500 ACS patients
– 1 year follow-up for cardiovascular death (193 events)



© 2014 Massachusetts Institute of Technology

MV and Other Risk Variables

Multivariate Analysis of Patients with *LVEF > 40%**

Parameter	Adjusted Hazard Ratio (*)	95% Confidence Interval	P Value
MV	**2.31**	**1.28 – 4.16**	**0.005**
BNP	1.94	1.10 – 3.43	0.022
BMI	0.93	0.54 – 1.62	0.805
CrCl	1.48	0.86 – 2.56	0.159

* Also adjusted for age, hypertension, diabetes, hypercholesterolemia, prior MI, prior angina, ST changes, cardiac markers

Z. Syed. C. Stultz, B. Scirica B, and J. Guttag, "Computationally generated cardiac biomarkers for risk stratification following acute coronary syndrome," *Science Translational Medicine*, September 2011.

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics
Applications - Medical Outcomes
Section 2

THANK YOU

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics

John Guttag
Professor
Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics
Applications - Medical Outcomes

John Guttag
Professor
Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

Healthcare-Associated Infections (HAIs)

- 1.7 million people per year get an infection during a hospital stay
- Contribute to approx. 4% of all deaths in the US
- Estimated costs of $20 Billion per year

© 2014 Massachusetts Institute of Technology

HAIs Remain Stubbornly Prevalent

THE WALL STREET JOURNAL.

"Medicare Shift Fails to Cut Hospital Infections"

--

Oct. 10th 2012

SCIENTIFIC AMERICAN™

"Hospitals Fail to Take Simple Measures to Thwart Deadly Infections, Survey Says"

-- April 8th, 2013

The New York Times

"More Aggressive Action Urged to Curb Hospital Infections"

-- May 29th, 2013

© 2014 Massachusetts Institute of Technology

Why?

- Temporal dynamics of risk factors are not well understood
- Global risk factors do not take into account institutional differences (institutional risk factors)
- Goal: use machine learning to build data-driven hospital-specific risk stratification models for healthcare-associated infections

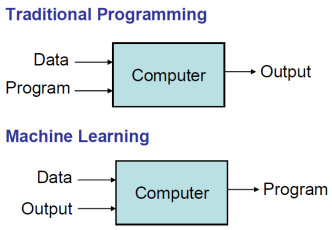
© 2014 Massachusetts Institute of Technology

What Is Machine Learning?

- Automating automation
  - Computer programs can automatically follow rules
  - How do we determine these rules automatically in the first place?
- Machine learning focuses on getting computers to program themselves
  - Let the data do the work
  - Automatically generate programs that create the right outputs from data

© 2014 Massachusetts Institute of Technology

What Is Machine Learning?



© 2014 Massachusetts Institute of Technology

How Are Things Learned?

- Memorization
  - Accumulation of individual facts
  - Limited by
    - *Time to observe facts
    - *Memory to store facts
- Generalization
  - Deduce new facts from old facts
  - Limited by accuracy of deduction process
    - *Essentially a predictive activity
    - *Assumes that the past predicts the future

© 2014 Massachusetts Institute of Technology

Machine Learning Methods

- Many different ones, all try to learn a model that is a generalization of examples
- Supervised: given a set of feature/label pairs, find a rule that predicts the label associated with a previously unseen input
- Unsupervised: given a set of feature vectors (without labels) group them into "natural clusters"
- All have three components
  - Representation of the model
  - Metric for assessing goodness of model
  - Optimization method for learning the model

© 2014 Massachusetts Institute of Technology

More Not Always Better

- Irrelevant information can slow things down
  - Volume of search space (i.e., possible combinations of features) increases exponentially with extra dimensions
- Irrelevant information can lead to a bad model
  - Fit the noise rather than the signal
  - Especially when dimensionality large relative to number of examples

© 2014 Massachusetts Institute of Technology

Back to Infections

- Challenges
  - Temporal aspect – changes over time
  - High dimensionality – thousands of variables
  - Paucity of cases – relatively rare events
  - Institutional differences – global models fail
  - Ground truth is not available – not all patients are tested

© 2014 Massachusetts Institute of Technology

Approach

- Use data-driven models to gain insight into the temporal aspects of the problem
- Use transfer learning based techniques to
  - Leverage data from other hospitals while
  - Incorporating hospital-specific features
- Topic of next section

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics
Applications - Medical Outcomes

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics

John Guttag

Professor

Massachusetts Institute of Technology

PROFESSIONAL EDUCATION

CSAIL

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data

Big Data Analytics

Applications - Medical Outcomes

John Guttag

Professor

Massachusetts Institute of Technology

© 2014 Massachusetts Institute of Technology

Predicting *Clostridium difficile*

- Bacteria that takes over the gut
- Transmitted through the mouth
- Causes severe diarrhea, intestinal diseases
- Treatment: metronidazole and oral vancomycin
- 20% of cases relapse within 60-days
- 178,000/year in U.S.
  - Approx. same as invasive breast cancer

© 2014 Massachusetts Institute of Technology

Risk Factors

Time Invariant

- Collected at the time of admission
- *e.g., admission complaint, previous admissions, home meds*

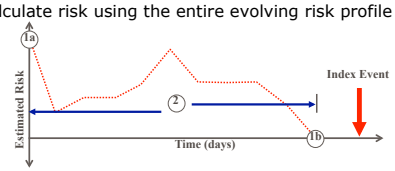
Time Varying

- Changes during the hospitalization
- *e.g., current meds, current procedures, current location, hospital conditions*

© 2014 Massachusetts Institute of Technology

Incorporating Time

- Typical Approaches in Clinical Literature
  - Calculate risk using only a snapshot of patient's state
    a. At time of admission [Tanner et al., 2009]
    b. n days before index event [Dubberke et al., 2011]
- Our Approach
  - Calculate risk using the entire evolving risk profile



© 2014 Massachusetts Institute of Technology

The Data

- EMR from a single large urban hospital in the US
- In-patients stays >=7 days from a single-year
- Population: ~10,000 hospital admissions
  - ~200 positive cases

© 2014 Massachusetts Institute of Technology

Variables Considered

- **Data available at the time of admission**
  - Age
  - Sex
  - Admission Complaint
  - Admission Procedure
  - Data from Hospital Admissions/Visits in last 90 days
  - Medical History
- **Data collected during the hospital stay**
  - Medications
  - Locations of the Patient
  - Procedures
  - Lab Results
  - Vitals
  - Devices
  - Staff

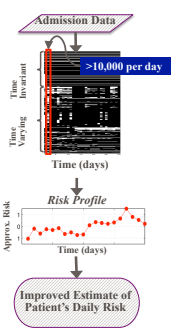
EHR Data
Not Billing Data

© 2014 Massachusetts Institute of Technology

Two Step Approach

① Time-Series Extraction: Inferring Risk Profile

② Time-Series Classification: Stratifying Patients Based on Risk Profile



© 2014 Massachusetts Institute of Technology

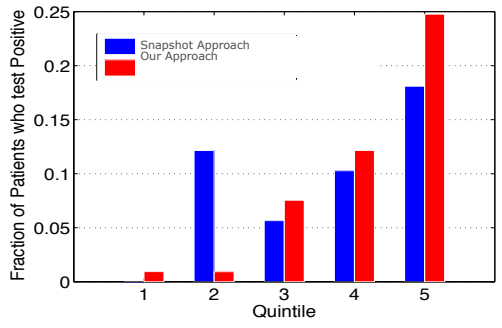
An Experiment

- Defining the index event
  - Positive examples: day of positive test result
  - Negative examples: midpoint of admission
    *Not day of discharge, that is easier but irrelevant
- Task: predict positive event two days in advance

Approach	AUROC (95% CI)	F-Score (95% CI)
Snapshot Approach	0.69 (0.61-0.77)	0.28 (0.19-0.38)
RP+Similarity$_{DTW}$	0.76 (0.69-0.82)	0.31 (0.22-0.41)
RP+HMM$_{posterior}$	0.76 (0.70-0.82)	0.30 (0.21-0.41)
RP+Features	0.79 (0.73-0.85)	0.37 (0.24-0.49)

© 2014 Massachusetts Institute of Technology

Results



© 2014 Massachusetts Institute of Technology

Wrapping Up

- Data analysis has transformed many fields
  - Biology
    - *From molecules, to genetics, to systems
  - Finance
    - *It takes a computer to lose $10,000,000/minute
  - Sports

Pictures removed due to copyright restrictions.
We are sorry for any inconvenience this may cause.

- It's time to transform medicine!

© 2014 Massachusetts Institute of Technology

Tackling The Challenges of Big Data
Big Data Analytics
Applications - Medical Outcomes

THANK YOU

PROFESSIONAL EDUCATION

CSAIL

© 2014 Massachusetts Institute of Technology