Outline: Health Care Data Use

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - What is Data Science?
  - What is Big Data?
  - Overview of Data Mining
- Understanding Data
Operational vs Decision Support Systems

- Operational Systems
  - Support day to day transactions
  - Contain current, “up to date” data
  - Examples: EMR, customer orders, inventory levels, payroll, bank account balances

- Decision Support Systems
  - Support strategic decision making
  - Contain historical, “summarized” data
  - Examples:
    - Clinical support: what treatment is best?
    - Population health
    - Management support: performance summary, market segmentation

Operational Application: EMR
DSS (Reports)

- PricePoint
  - for consumers
- http://www.txpricepoint.org/

Outline: Health Care Data Use

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - What is Data Science?
  - What is Big Data?
  - Overview of Data Mining
- Understanding Data
What is Data Science?

- Other words
  - Knowledge Discovery & Data mining (KDD)
  - Business Intelligence / Business Analytics
- Collecting and refining information from many sources
- Analyzing and presenting the information in useful ways
- So people can make better business decisions

Data Science
Knowledge Discovery & Data mining (KDD)

Big Data: operational data  KDD: Clean, Merge, Reprocess

Human consumable, valid, novel, potentially useful, and ultimately understandable information
Data Science Definition (Big Data less consensus)

- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery hypothesis, and analytical hypothesis analysis.
- A **Data Scientist** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific method process through each stage in the big data lifecycle.

Big Data refers to digital data volume, velocity and/or variety whose management requires scalability across coupled horizontal resources.

---

Data to Decision

Hierarchy of Data Science & Different Types of Data Scientists

*Hye-Chung Kum, Population Informatics Research Group*
KDD Process

- Operational Data
  - Data cleaning & integration

- EDW
  - Feature Selection (what vars?)

- Task Specific Data
  - Analysis / Datamining

- Results
  - Validation / Evaluation

- Information Presentation
  - Action

Video

- TX Mental Health Landscape (2:46)
  - [https://www.youtube.com/watch?v=8dPqQt0yXJA](https://www.youtube.com/watch?v=8dPqQt0yXJA)

- Wealth Inequality (1:30)
  - [https://www.youtube.com/watch?v=QPKKQnijnsM](https://www.youtube.com/watch?v=QPKKQnijnsM)

- Good managers know how to build the data story!
Outline: Health Care Data Use

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - What is Data Science?
  - What is Big Data?
  - Overview of Data Mining
- Understanding Data

Properties of BIG DATA: 4V

- Volume: constantly generating
- Velocity: constantly changing
- Variety: expressed in many ways
- Veracity: lots of errors
- (Value)

EXAMPLE: the INTERNET!
What do you do to find information/knowledge on the Internet?
The Big Data Problem - Nutshelled
Michael Franklin (UC Berkley)

Something’s gotta give:

- Time
- Money
- Quality (precision)

Massive, Diverse and Growing Data

AMPLab: Integrating Three Key Resources

- Algorithms
  - Machine Learning, Statistical Methods
  - Prediction, Business Intelligence
- Machines
  - Clusters and Clouds
  - Warehouse Scale Computing
- People
  - Crowdsourcing, Human Computation
  - Data Scientists, Analysts
Outline: Health Care Data Use

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - What is Data Science?
  - What is Big Data?
  - Overview of Data Mining
- Understanding Data

What is Data Mining?

- Using a combination of artificial intelligence, machine learning, and statistical analysis to analyze data
- and discover useful patterns that are “hidden” there
Business uses of data mining: Essentially five tasks

- Classification: Group data into predetermined categories
  - Classify credit applicants as low, medium, high risk
  - Classify insurance claims as normal, suspicious
- Estimation: Estimate probability of an event through models built from previous data
  - Estimate the probability of a direct mailing response
  - Estimate the potential cohort size for a clinical trial
- Prediction: Predict an outcome based on input based on models built from previous data
  - Predict which customers will leave within six months
  - Predict which patient will return to the ED
- Affinity Grouping: Group people based on similar characteristics
  - Find out what books to recommend to Amazon.com users
  - Find treatment regime that was successful for similar patient
- Description
  - Help understand large volumes of data by uncovering interesting, useful, and actionable patterns

---

Job market of data scientists

- Statisticians will be the next sexy job
  - Google Chief Economist Hal Varian
- Shortage of 190,000 data scientists by the year 2019
  - McKinsey Global Institute
Outline: Health Care Data Use

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - What is Data Science?
  - What is Big Data?
  - Overview of Data Mining
- Understanding Data

Applications in Health

- A March 2014 poll from MeriTalk and EMC found that 63 percent of healthcare executives in the federal government believe that big data will improve population health management

- Examples
  - Manage population health
    - Accountable Care Organizations (ACO)
  - Clinical decision support
  - Cohort identification for clinical trials
  - Medical fraud detection
Bias and Variance

http://scott.fortmann-roe.com/docs/BiasVariance.html

- precise but not valid?
- What is real data like?
- Adjust for bias
- Take into account variance

Numerical Data: distribution

- Mean
- Standard Deviation
  - How dispersed
- Range: Max/Min
- Median (percentile)
- Scatter Plot: 2 vars
Categorical Data

- Tabulation
- Cross tabulation
  - 2 variables
- GIS: maps

Take Away I
What is Data Science? KDD Process

- Data cleaning & integration
- Feature Selection (what vars?)
- Analysis / Datamining
- Validation / Evaluation
- Action
Take Away II
What is Big Data?

- 4 Vs of Big Data
  - Volume: lots of data
  - Velocity: constantly generating & changing
  - Variety: expressed in many ways
  - Veracity: lots of errors
  - (Value)

- Big Data Problems
  - Time
  - Money
  - Quality (Precision)

- Three Resources: AMP
  - Algorithm
  - Machine
  - People

Take Away III
Business uses of data mining: Essentially five tasks

- Classification
  - Classify credit applicants as low, medium, high risk
  - Classify insurance claims as normal, suspicious

- Estimation
  - Estimate the probability of a direct mailing response
  - Estimate the potential cohort size for a clinical trial

- Prediction
  - Predict which customers will leave within six months
  - Predict which patient will return to the ED

- Affinity Grouping
  - Find out what books to recommend to Amazon.com users
  - Find treatment regime that was successful for similar patient

- Description
  - Help understand large volumes of data by uncovering interesting, useful, and actionable patterns
How do you get good with data?

- Sorry, no short cuts. Build experience.
- In this course, start you out.
  - Tableau / Excel
  - SQL
  - Assignment 1
  - Labs

Reminder: due next two weeks

- Lab 1: most of you should be done during class
- Assignment 1: submit on E-campus day before class
  - Week one: progress report
  - Week two: Final Tutorial
- Readings: Chapters 1 & 2
- Quiz 1 (E-campus: posted on Tues)
  - Practice quiz
- Group presentation emails