Data to Decision

DRAFT

Course URL:
http://pinformatics.tamhsc.edu/courses/phpm631

Outline

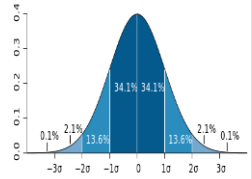
- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

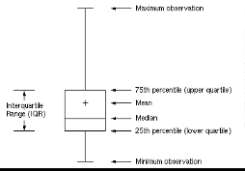
Outline

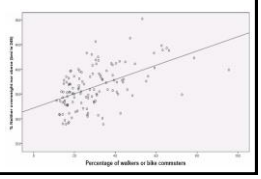
- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

Numerical Data : distribution

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

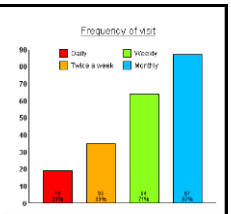


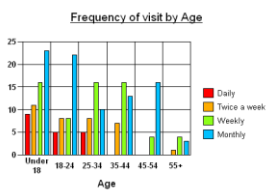


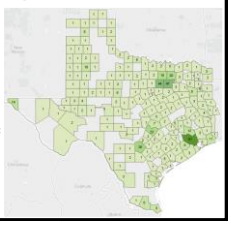


Categorical Data

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

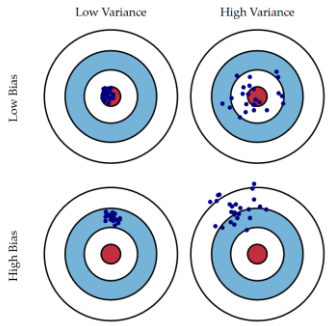






Bias and Variance

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



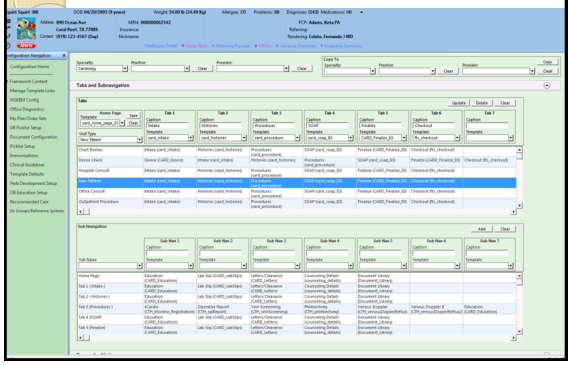
Outline

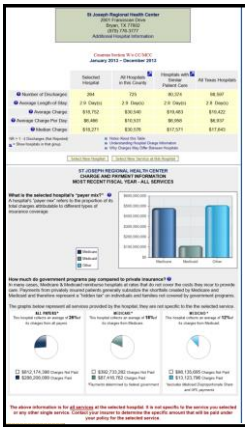
- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

Operational vs Decision Support Systems

- Operational Systems
  - Support day to day transactions
  - Contain current, "up to date" data
  - Examples: EMR, customer orders, inventory levels, 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/
- http://www.txpricepoint.org/Report.aspx?DRG=766&FacilityID=0410500

Outline

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

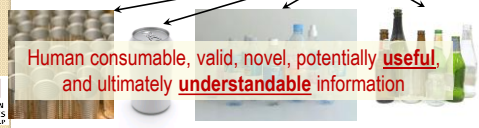
What is Data Science?

- 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

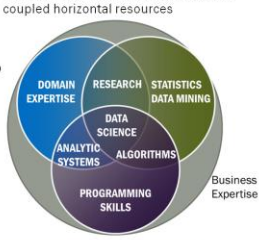
Population Informatics Research Group

NIST Big Data

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.

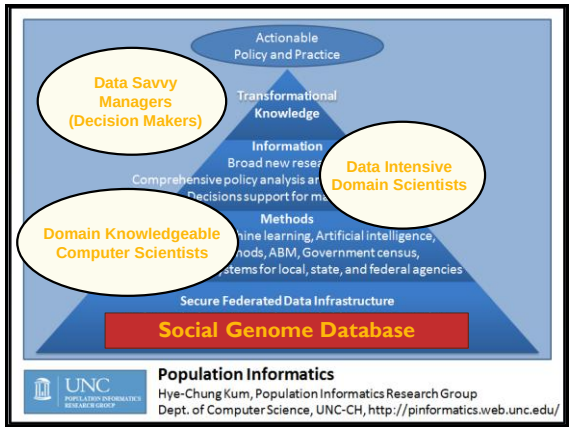
Big Data refers to digital data volume, velocity and/or variety whose management requires scalability across coupled horizontal resources



The Virtuous Cycle of Data to Decision & Action

1. Identify the business problem
2. Transform data into information using data mining techniques
3. Act on the information
4. Measure the results

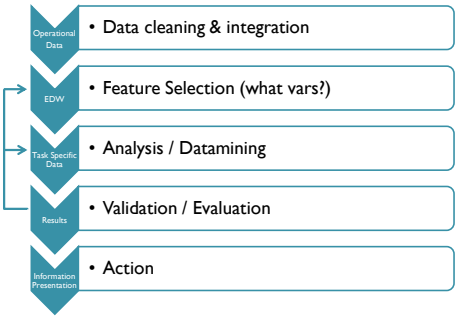
Population Informatics Research Group



UNC Population Informatics Research Group

Population Informatics
Hye-Chung Kum, Population Informatics Research Group
Dept. of Computer Science, UNC-CH, http://pinformatics.web.unc.edu/

KDD Process



Population Informatics Research Group

Thomas Davenport
Competing on Analytics

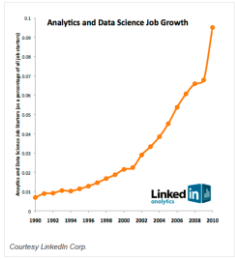
- Skill set for good data scientists
  - IT & Programming skills
  - Statistical skills
  - Business skills:
    - Understand pros/cons of decisions & actions
    - Communication skills
    - Excel / PowerPoint
  - Intense curiosity: the most important skill or trait. "a desire to go beyond the surface of a problem, find the question at its heart, and distill them into a very clear set of hypothesis that can be tested"

Population Informatics Research Group

Data science teams need people with the skills and curiosity to ask the big questions (oreilly)

- Technical expertise: the best data scientists typically have deep expertise in some scientific discipline.
- Curiosity: a desire to go beneath the surface and discover and distill a problem down into a very clear set of hypotheses that can be tested.
- Storytelling: the ability to use data to tell a story and to be able to communicate it effectively.
- Cleverness: the ability to look at a problem in different, creative ways.
- Health is a very important domain
  - Team lead: good questions, good interpretation & implications
- http://radar.oreilly.com/2011/09/building-data-science-teams.html

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

New Era in Science
Big Data Science

- Data is the new raw material of business: an economic input almost on par with capital and labor.(Microsoft's Craig Mundie)
- Those who can harness the power of data will lead the next century and drive innovation in commerce, scientific discovery, healthcare, finance, energy, government, and countless other fields.
- Students who learn to be a data science will be in high demand.

Outline

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

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

Sample Applications

- Direct Marketing
  - identify which prospects should be included in a mailing list
  - Clinical trial recruitment: cohort identification
- Market segmentation
  - identify common characteristics of customers who buy same products
  - Profile common characteristics in homogeneous patient group (Billings)
- Customer churn
  - Predict which customers are likely to leave your company for a competitor
  - Potentially Preventable Readmissions to ED
- Market Basket Analysis
  - Identify what products are likely to be bought together
  - Care coordination: common services for a condition
- Insurance Claims Analysis
  - discover patterns of fraudulent transactions (medical fraud)
  - compare current transactions against those patterns

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

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

Outline

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

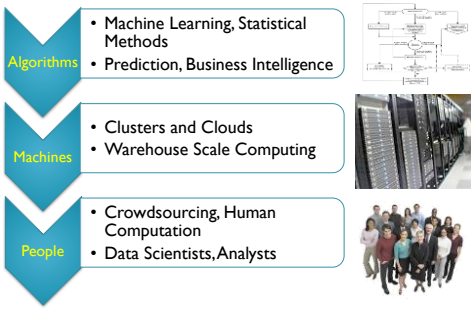
Properties of BIG DATA : 4V

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

EXAMPLE: the INTERNET!
What do you do to find information/knowledge on the Internet?

The Big Data Problem – Nutshelled
Michael Franklin (UC Berkley)

AMPLab
Integrating Three Key Resources



Outline

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

What is a data warehouse? (Enterprise Data Warehouse=EDW)

- A collection of data from multiple sources
  - within the company
  - outside the company
- Usually includes data relevant to the entire enterprise
- Usually includes summary data and historical data as well as current operational data
- Usually requires "cleaning" and other integration before use
- Therefore, usually stored in separate databases from current operational data

What is a data mart?

- A subset of a data warehouse focused on a particular subject or department

Data Warehousing considerations

- What data to include?
- How to reconcile inconsistencies?
- How often to update?

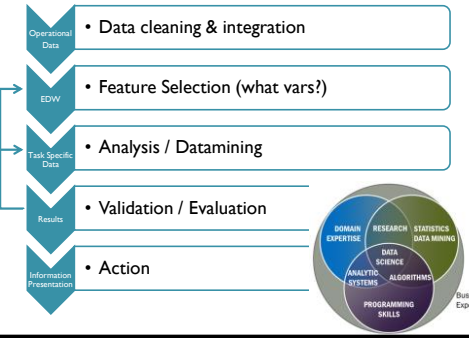
Outline

- Operational vs. Decision Support Systems
- What is Data Science/Business Intelligence
  - Overview of Data Mining
  - What is Big Data?
- Data Warehouses
- Case Studies
- Understanding Data
  - Data Mining Methods (Optional)

Case Studies

- Specialized Program for High Utilizers in One Hospital Network
  - innovative and effective High Alert emergency department overutilization program
  - "SETON's High Alert Program is a specialized, focused case management program. In this program, we develop individualized care plans based on the particular circumstances of individuals. These care plans can be electronically pushed into future clinical encounters to ensure the individualized plan is followed, resulting in consistent care directed at the specific patient."
  - Dr. Christopher Ziebell (Emergency Department Medical Director at the University Medical Center Brackenridge)
  - Dr. Ziebell serves on the Board of Trustees of SETON Healthcare and the Board of Managers of Emergency Service Partners; is Executive Director of Hospital Physicians in Clinical Research; and, chairs the EMS Steering Committee of the Travis County Medical Society.

Take Away I
What is Data Science? KDD Process



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