
# DATA MINING FOR HEALTHCARE MANAGEMENT

Prasanna Desikan
prasanna@gmail.com
Center for Healthcare Innovation
Allina Hospitals and Clinics
USA

Kuo-Wei Hsu
kuowei.hsu@gmail.com
National Chengchi University
Taiwan

Jaideep Srivastava
srivasta@cs.umn.edu
University of Minnesota &
Center for Healthcare Innovation
Allina Hospitals and Clinics
USA

# Outline

- Introduction
- Why Data Mining can aid Healthcare
- Healthcare Management Directions
- Overview of Research
  - Kinds of Data
  - Challenges in data mining for healthcare
  - Framework
  - Prominent Models
- Sample case study
- Summary and Future Directions

# INTRODUCTION

# Healthcare Management

"Health administration or healthcare administration is the field relating to leadership, management, and administration of hospitals, hospital networks, and health care systems."*

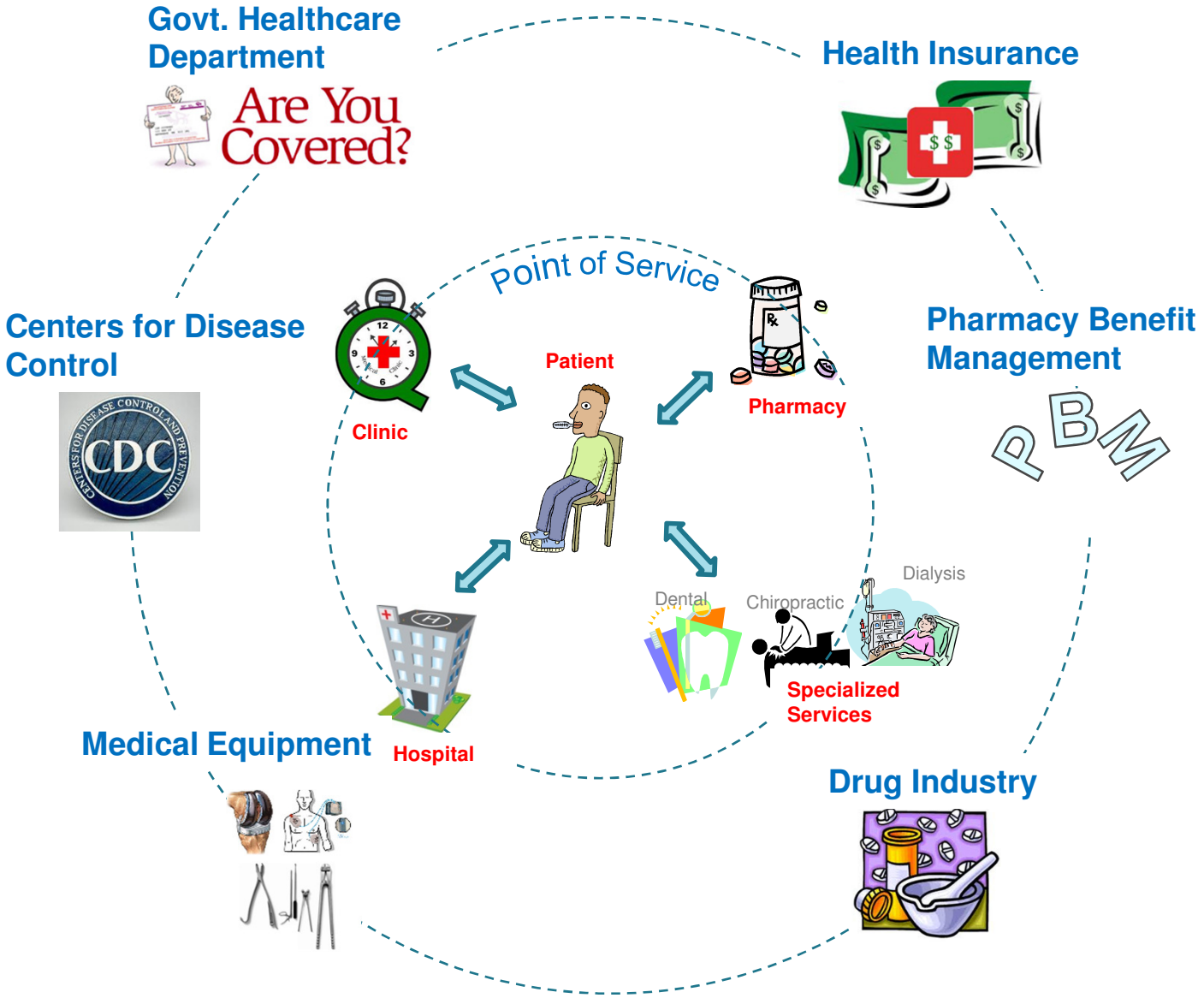
It is actually a broad area that could encompass:

- Healthcare Informatics
- Medical Device Industry
- Pharmaceutical Industry
- Hospital Management
- System Biology

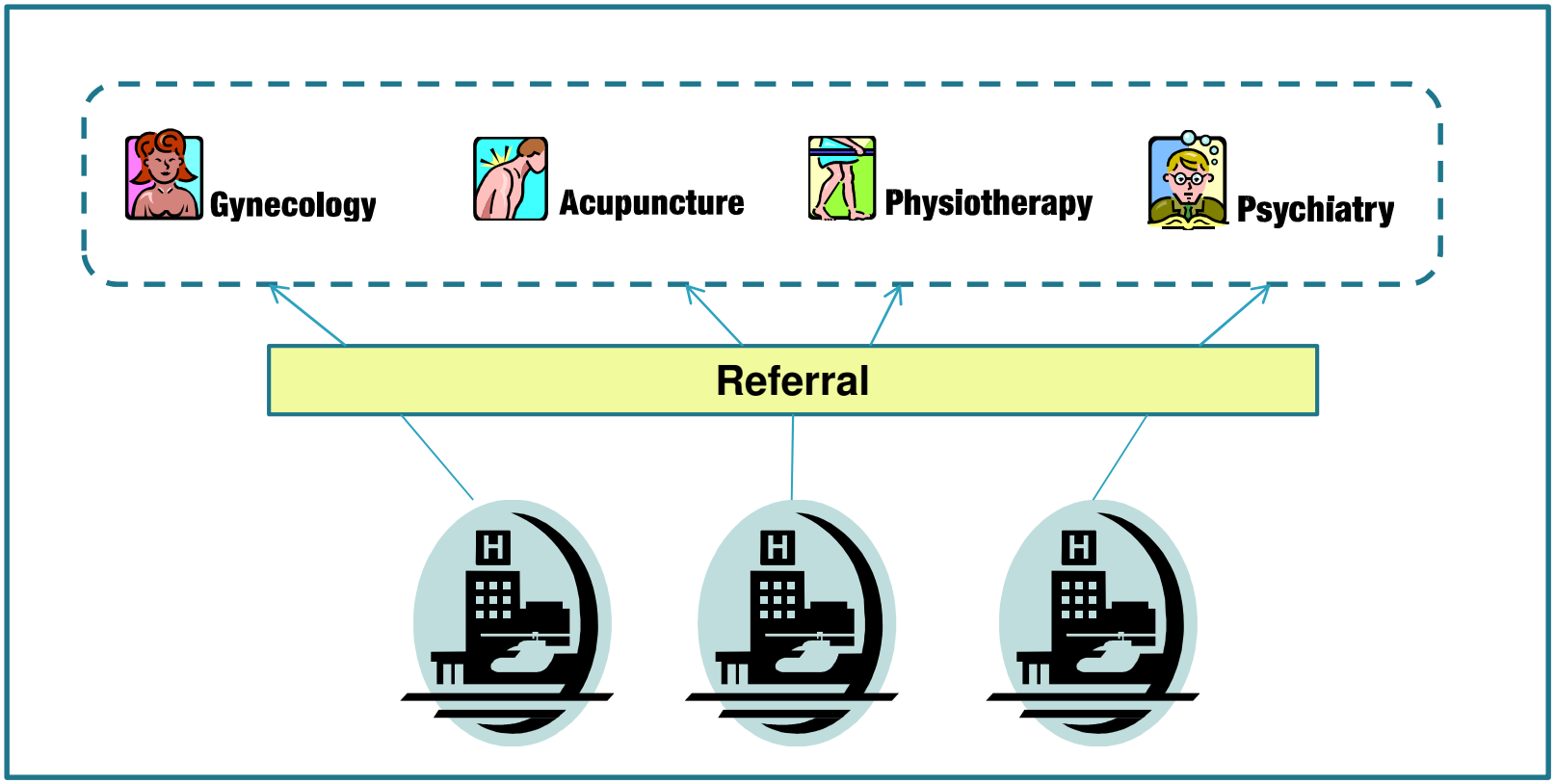
and many more….

*http://en.wikipedia.org/wiki/Healthcare_management

# Healthcare Ecosystem – A Perspective

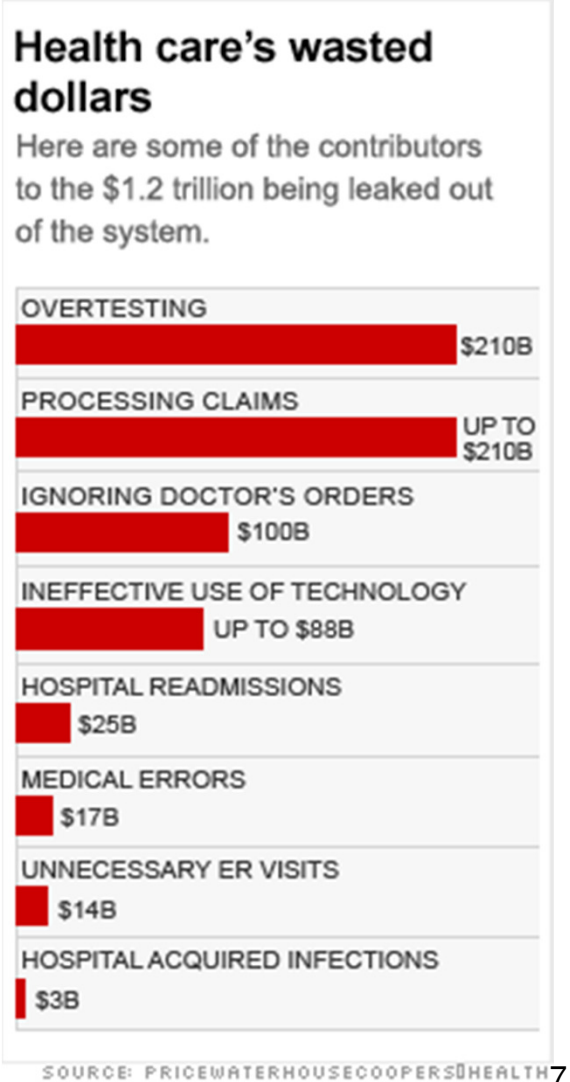


# Interface between Patients and Medical Services




# Motivation for Healthcare Management

- Problem:
  - "**Government health spending wastes a heck of a lot of money**," U.S. Vice President Joe Biden, 2/25/2010
  - "**Healthcare spending 17 percent of economy**", UPI.com, 2/4/2010
  - "**More than $1.2 trillion spent on health care each year is a waste of money**", CNNMoney.com, 8/10/2009



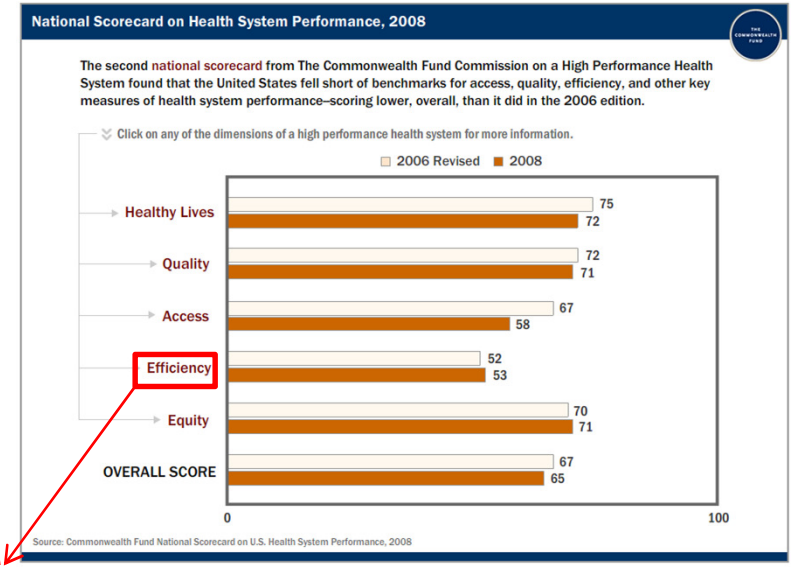
# Performance of the U.S. Health Care System Internationally

Exhibit ES-1. Overall Ranking

| Country Rankings |
|---|
| 1.00–2.33 |
| 2.34–4.66 |
| 4.67–7.00 |

| | AUS | CAN | GER | NETH | NZ | UK | US |
|---|---|---|---|---|---|---|---|
| OVERALL RANKING (2010) | 3 | 6 | 4 | 1 | 5 | 2 | 7 |
| Quality Care | 4 | 7 | 5 | 2 | 1 | 3 | 6 |
| Effective Care | 2 | 7 | 6 | 3 | 5 | 1 | 4 |
| Safe Care | 6 | 5 | 3 | 1 | 4 | 2 | 7 |
| Coordinated Care | 4 | 5 | 7 | 2 | 1 | 3 | 6 |
| Patient-Centered Care | 2 | 5 | 3 | 6 | 1 | 7 | 4 |
| Access | 6.5 | 5 | 3 | 1 | 4 | 2 | 6.5 |
| Cost-Related Problem | 6 | 3.5 | 3.5 | 2 | 5 | 1 | 7 |
| Timeliness of Care | 6 | 7 | 2 | 1 | 3 | 4 | 5 |
| Efficiency | 2 | 6 | 5 | 3 | 4 | 1 | 7 |
| Equity | 4 | 5 | 3 | 1 | 6 | 2 | 7 |
| Long, Healthy, Productive Lives | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Health Expenditures/Capita, 2007 | $3,357 | $3,895 | $3,588 | $3,837* | $2,454 | $2,992 | $7,290 |

Note: * Estimate. Expenditures shown in $US PPP (purchasing power parity).
Source: Calculated by The Commonwealth Fund based on 2007 International Health Policy Survey; 2008 International Health Policy S
Survey of Primary Care Physicians; Commonwealth Fund Commission on a High Performance Health System National Scorecard; and
OECD Health Data, 2009 (Paris: OECD, Nov. 2009).




**Reflects lack of leverage of Information technology**

- U.S. **is lagging in adoption of national policies** that promote primary care, quality improvement, and information technology.
- Health reform legislation addresses these deficiencies; for instance, the American Recovery and Reinvestment Act signed by President Obama in February 2009 included **approximately $19 billion to expand the use of health information technology.**
- The Patient Protection and Affordable Care Act of 2010 also will work toward realigning providers' financial incentives, **encouraging more efficient organization and delivery of health care**, and investing in preventive and population health.

***Key Takeaway***: Need for an efficient , organized and knowledge based decision support systems.

Source : http://www.commonwealthfund.org/Content/Publications/Fund-Reports/2010/Jun/Mirror-Mirror-Update.aspx. 4/29/2011

# WHY DATA MINING CAN AID HEALTHCARE

# Why Data Mining?

- Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc.
- The large amounts of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making.
- *Data mining*
  - *brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns*
  - *that provide healthcare professionals an additional source of knowledge for making decisions*
- The decisions rests with health care professionals.

# How does data mining help?

- Data mining is a collection of algorithmic ways to extract informative patterns from raw data
  - Data mining is purely data-driven; this feature is important in health care
- $y = \boldsymbol{f}(\boldsymbol{x})$
  - We have seen $\boldsymbol{x}$ (set of independent variables) and observed $y$ (dependent variable); data mining tells us something about the nature of $\boldsymbol{f}$
    - x = symptoms or test results, y = diseases;
    - x = treatments, y = symptom
  - It tells us "how"
    - How is x related to y? What function describes their relationship?
- $\boldsymbol{f}(x, y)$ = score, or $\boldsymbol{f}(x|y) = \mathbf{Pr}(x|y)$
- Data mining does not (directly) explain to us "why" – Why does x cause y?
  - It helps doctors/physicians (domain experts) figure that (causation) out
  - 'Descriptive/predictive model' vs. 'Causal model'

# Who does it benefit?

- Data mining can help
  - Healthcare insurers detect fraud and abuse,
  - Healthcare organizations make customer relationship management decisions,
  - Physicians identify effective treatments and best practices, and
  - Patients receive better and more affordable healthcare services.

# HEALTHCARE MANAGEMENT DIRECTIONS

# Key Dimensions in Healthcare Management [Koh05]

- Diagnosis and Treatment
- Healthcare Resource Management
- Customer Relationship Management
- Fraud and Anomaly Detection

# Diagnosis and Treatment

Medical decision support (to doctors) [Hardin2008]

- Analysis of digitized images of skin lesions to diagnose melanoma [Burroni 2004]
- Computer-assisted texture analysis of ultrasound images aids monitoring of tumor response to chemotherapy[Hub2000]
- Predicting the presence of brain neoplasm with magnetic resonance spectroscopy [Zellner2004]
- Analysis of digital images of tissue sections to identify and quantify senile plagues for diagnosing and evaluating the severity of Alzheimer's disease. [Hibbard1997]

# Diagnosis and Treatment

## Treatment plan (to patients)

- Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular treatment option
- Based on patients' profile, history, physical examination, diagnosis and utilizing previous treatment patterns, new treatment plans can be effectively suggested
- Examples
  - Onset, treatment and management of depression [Hadzic2010]
  - Treatment Decision Support Tool for Patients with Uterine Fibroids [Campbell2010]

# Healthcare Resource Management

- Using logistic regression models to compare hospital profiles based on risk-adjusted death with 30 days of non-cardiac surgery
- Neural network system to predict the disposition in children presenting to the emergency room with bronchiolitis
- Predicting the risk of in-hospital mortality in cancer patients with nonterminal disease

# Healthcare Resource Management

## Prediction of inpatient length of stay

- Effectively manage the resource allocation by identifying high risk areas and predicting the need and usage of various resources.
- For example, a key problem in the healthcare area is the measurement of flow of patients through hospitals and other health care facilities.
- If the inpatient length of stay (LOS) can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced.

# Customer Relationship Management

- CRM is to establish close customer relationships [Rygielski02]
  - The focus shifts away from the breadth of customer base (product-oriented view, mass marketing) to the depth of each customer's needs (customer-oriented view, one-to-one marketing)
- CRM is built on an integrated view of the customer across the whole organization [Puschmann01]
  - Customers have a fractured view of an enterprise; the enterprise has a splintered view of the customer
- Kohli et al. demonstrate a web-based Physician Profiling System (PPS) to strengthen relationships with physicians and improve hospital profitability and quality [Kohli01]

# Customer Relationship Management (cont.)

- Development of total customer relationship in healthcare includes several tenets [Berwick97]
  - "In a helping profession, the ultimate judge of performance is the person helped"
  - "Most people, including sick people, are reasonable most of the time"
  - "Different people have different, legitimate needs"
  - "Pain and fear produce anxiety in both the victim and the helper"
  - "Meeting needs without waste is a strategic and moral imperative"
- Some demographic characteristics (e.g. age, health status, and race) and institutional characteristics (e.g. hospital size) consistently have a significant effect on a patient's satisfaction scores [Young00]
- Chronic illnesses (e.g. diabetes and asthma) require self-management and a collaborative patient-physician relationship [Ouschan06]

# Customer Relationship Management (cont.)

- The principles of applying of data mining for customer relationship management in the other industries are also applicable to the healthcare industry.
- The identification of usage and purchase patterns and the eventual satisfaction can be used to improve overall customer satisfaction.
- The customers could be patients, pharmacists, physicians or clinics.
- In many cases prediction of purchasing and usage behavior can help to provide proactive initiatives to reduce the overall cost and increase customer satisfaction.

# Fraud and Anomaly Detection

- Bolton and Hand briefly discuss healthcare insurance fraud [Bolton02]
  - Examples of frauds:
    - Prescription fraud: claims for patients who do not exist
    - Upcoding: claims for a medical procedure which is more expensive or not performed at all
  - Examples of detection methods:
    - Neural networks, genetic algorithms, nearest neighbor methods
    - Comparing observations with those they have similar geodemographics

# Fraud and Anomaly Detection

- Data mining has been used very successfully in aiding the prevention and early detection of medical insurance fraud.
- The ability to detect anomalous behavior based on purchase, usage and other transactional behavior information has made data mining a key tool in variety of organizations to detect fraudulent claims, inappropriate prescriptions and other abnormal behavioral patterns.
- Another key area where data mining based fraud detection is useful is detection and prediction of faults in medical devices.

# Examples of Research in Data Mining for Healthcare Management.

| Researching topic | Researching institute | Dataset |
|---|---|---|
| Healthcare data mining: **predicting inpatient length of stay** | School of Information Management and Engineering, Shanghai University; Harrow School of Computer Science | Geriatric Medicine department of a metropolitan teaching hospital in the UK. |
| Designing **Patient-Specific Seizure Detectors** From Multiple Frequency Bands of Intra-cranial EEG Using Support Vector Machines | The Center for Computational Learning Systems (CCLS) and The Columbia University Medical School (CUMC) | Columbia University Medical School has collected approximately 30 TB of intra-cranial EEG recordings. |
| **Classification, Treatment and Management of Alzheimer's Disease** Using Various Machine Learning Methods | MGR University, Chennai; UVCE , Bangalore,; Defence Institue of Advanced Technology Pune | National Institute on Aging, USA. |

# OVERVIEW OF RESEARCH

# Kinds of Data

- “An EHR is an electronic version of a patient’s medical history, that is maintained by the health- care provider over time, and includes all of the key administrative clinical data relevant to that person’s care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, medical images and radiology reports. “ [Maglogiannis09]
- HL7 (Health Level Seven)
  - Developed to improve health informatics interoperability
  - Working in the 7th layer, application layer, of the Open Systems Interconnection model
  - www.hl7.org

# EMR and EHR

- Electronic Medical Record (EMR): It contains patient information that is stored and retrieved locally in a stand-alone system used by a provider
- Electronic Health Record (EHR): It contains patient information that is stored and retrieved in systems used by all providers who care about the patient
- Many use EMR and EHR interchangeably
- Examples of applying data mining on EMR/EHR:
  - Ludwick and Doucette study the adaption of EMR in primary care [Ludwick09]
  - Cerrito works on EMR from an Emergency Department [Cerrito07]
  - Buczak et al. works on disease surveillance on EMR [Buczak09]

# Health Level 7 (HL 7)

- HL7 provides standards for interoperability that improve care delivery, optimize workflow, reduce ambiguity and enhance knowledge transfer among healthcare providers [HL7]

**The V2.4 Message**

```
MSH|^~\&|GHH LAB|ELAB-3|GHH OE|BLDG4|200202150930||ORU^R01|CNTRL-3456|P|2.4<cr>
PID|||555-44-4444||EVERYWOMAN^EVE^E^^^^L|JONES|19620320|F|||153 FERNWOOD DR.^
^STATESVILLE^OH^35292||(206)3345232|(206)752-121||||AC555444444||67-A4335^OH^20030520<cr>
OBR|1|845439^GHH OE|1045813^GHH LAB|15545^GLUCOSE|||200202150730|||||||||
555-55-5555^PRIMARY^PATRICIA P^^^^MD^^|||||||||F||||||444-44-4444^HIPPOCRATES^HOWARD H^^^^MD<cr>
OBX|1|SN|1554-5^GLUCOSE^POST 12H CFST:MCNC:PT:SER/PLAS:QN||^182|mg/dl|70_105|H|||F<cr>
```

**The V3 Message**

```
<recordTarget>
   <patientClinical>
   <id root="2.16.840.1.113883.19.1122.5" extension="444-22-2222"
       assigningAuthorityName="GHH Lab Patient IDs"/>
   <statusCode code="active"/>
     <patientPerson>
       <name use="L">
          <given>Eve</given>
          <given>E</given>
          <family>Everywoman</family>
       </name>
       <asOtherIDs>
          <id extension="AC555444444" assigningAuthorityName="SSN"
              root="2.16.840.1.113883.4.1"/>
       </asOtherIDs>
     </patientPerson>
   </patientClinical>
</recordTarget>
```

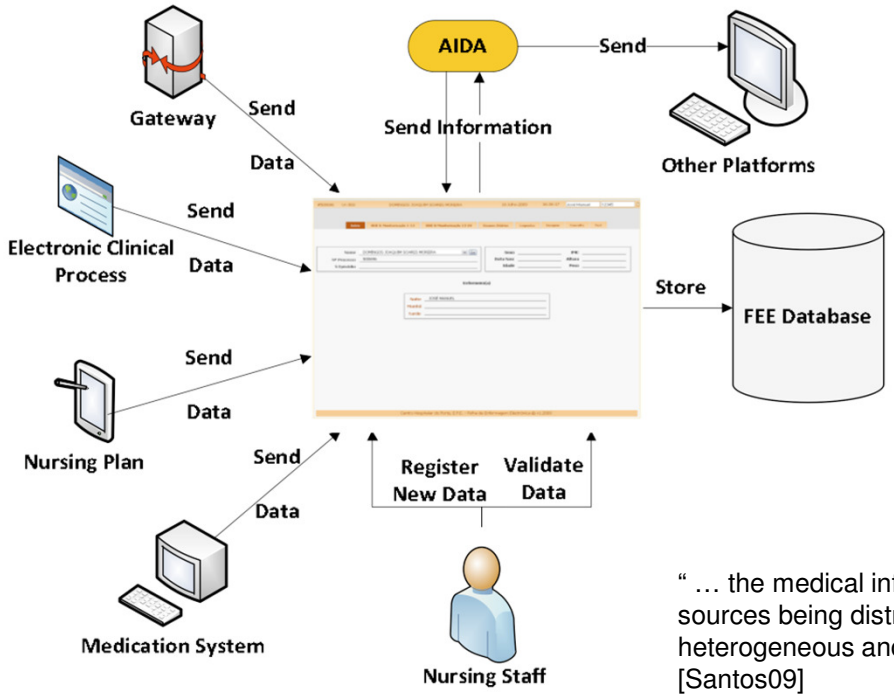
# Kinds of Data

**Table 2.** Electronic Health Records (EHR) data modalities [Maglogiannis09]

| Digital Data | Contrast / Resolution (No. of samples per second x bits per sample) | Data Size |
|---|---|---|
| Demographic Data | | ~ 100 KB |
| Clinical Data (Biosignals) | | ~ 100 KB / incident |
| Digital audio stethoscope (Heart Sound) | 10000 x 12 | ~ 120 kbps |
| Electrocardiogram ECG | 1250 x 12 | ~ 15 Kbps |
| Electroencephalogram EEG | 350 x 12 | ~ 10 Kbps |
| Electromyogram EMG | 50000 x 12 | ~ 600 Kbps |
| Ultrasound, Cardiology, Radiology | 512x512x8 | 256 KB (image size) |
| Magnetic resonance image | 512x512x8 | 384 KB (image size) |
| Scanned x-ray | 1024x1250x12 | 1.8 MB (image size) |
| Digital radiography | 2048x2048x12 | 6 MB (image size) |
| Mammogram | 4096x4096x12 | 24 MB (image size) |
| Compressed and full motion video (telemedicine) | - | 384 kbps to 1.544 Mb/s (speed) |

# Kinds of Data

- Electronic Nursing Record (ENR): “While improving health care practices and patient care, it also provides easily and rapidly available data for a decision support system in real time.” [Santos09]




“ … the medical information sources being distributed, heterogeneous and complex …” [Santos09]

**Fig. 1 – ENR: Information Sources**

# Kinds of Data

- Data warehouse for integration of “evidence-based” data sources [Stolba06]

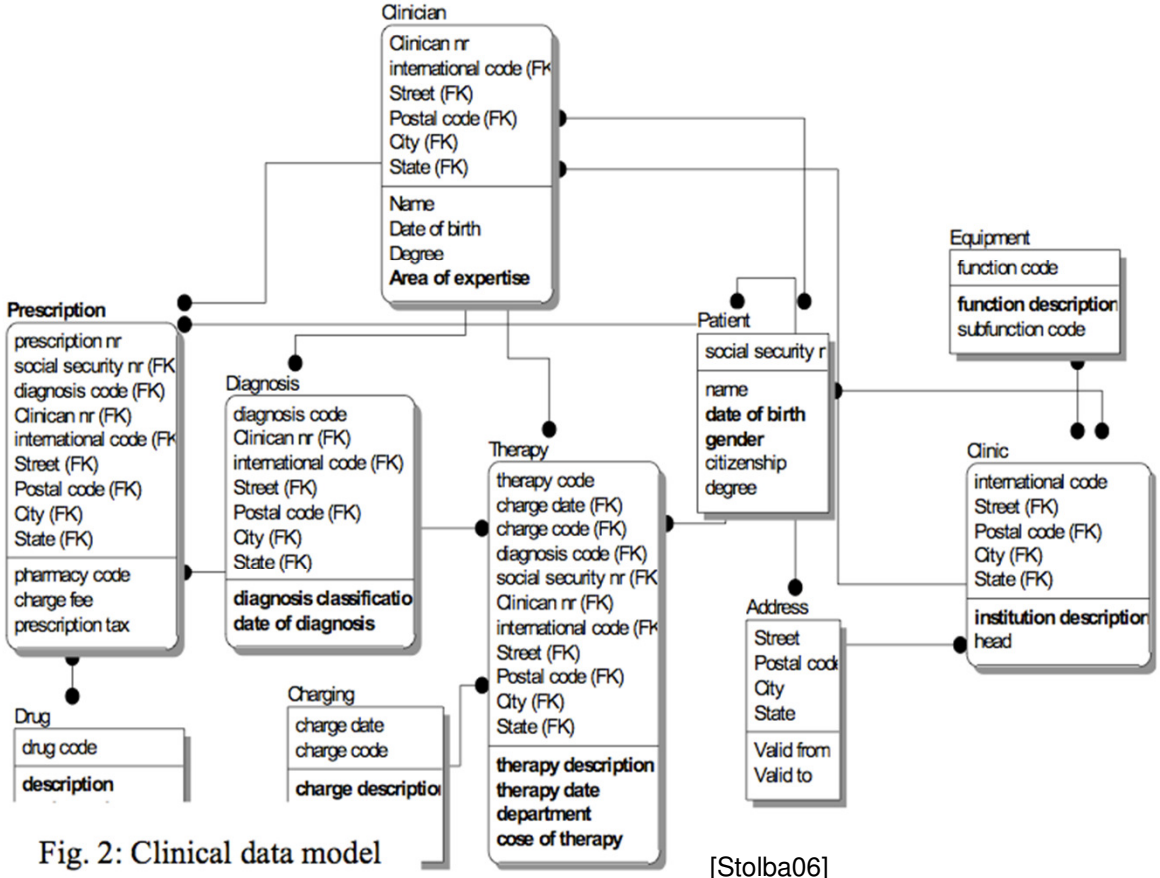



Fig. 2: Clinical data model

[Stolba06]

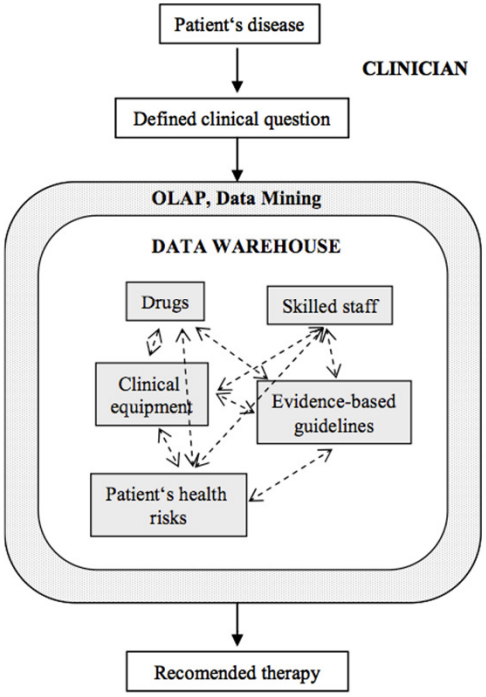



Fig. 3: Data warehouse at the point of care

[Stolba06]

# Challenges in Data Mining for Healthcare

- Data sets from various data sources [Stolba06]
- Example 1: Patient referral data can vary extensively between cases because structure of patient referrals is up to general practitioner who refers the patient [Persson09]
- Example 2: Catley et al. use neural networks to predict preterm birth on a heterogeneous maternal population [Catley06]
- Example 3: “Traditional clinical-based prognosis models were discovered to contain some restrictions to address the heterogeneity of breast cancer” [Ahmad09]

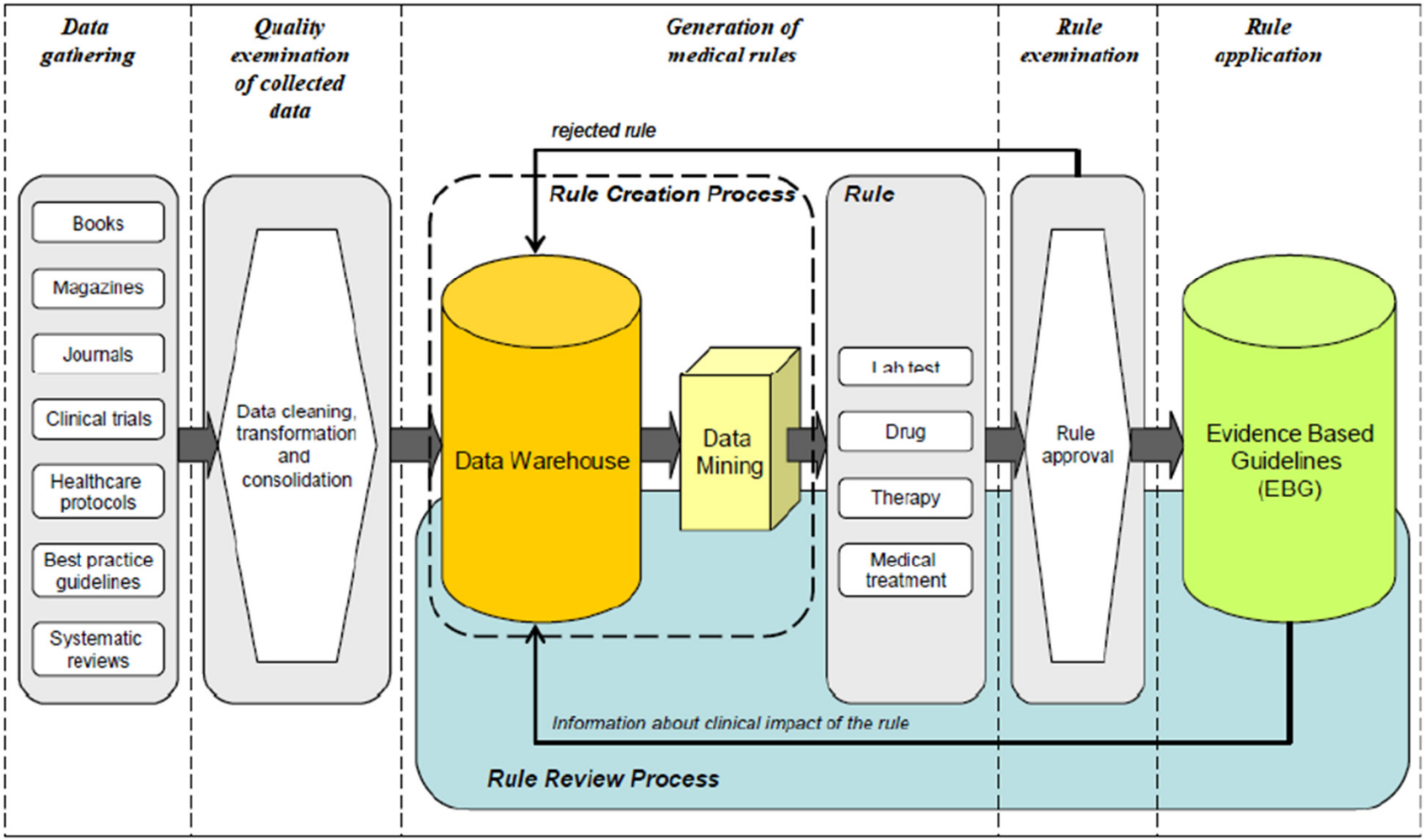
# Challenges in Data Mining for Healthcare

- Data from heterogeneous sources present challenges [Kwiatkowska07]
  - Sampling bias: “Clinical studies use diverse collecting methods, inclusion criteria, and sampling methods”
  - Referral bias: “Data represent a preselected group with a high prevalence of disease”
  - Selection bias: “Clinical data sets include patients with different demographics”
  - Method bias: “Predictors have varied specifications, granularities, and precisions”
  - Clinical spectrum bias: “Patient records represent varied severity of a disease and co-occurrence of other medical problems”

# Challenges in Data Mining for Healthcare

- Missing values, noise, and outliers
- “Cleaning data from noise and outliers and handling missing values, and then finding the right subset of data, prepares them for successful data mining” [Razavi07]
- Transcription and manipulation of patient records often result in a high volume of noise and a high portion of missing values [O’Sullivan08]
- “Missing attribute values can impact the assessment of whether a particular combination of attribute-value pairs is significant within a dataset” [Laxminarayan06]

# Typical Data Mining Framework




[Stolba06]

# Prominent Models

- O'Sullivan et al. propose to incorporate formalized external expert knowledge in building a prediction model for asthma exacerbation severity for pediatric patients in the emergency department [O'Sullivan08]
- The secondary knowledge source identified as relevant for our retrospective asthma data is the Preschool Respiratory Assessment Measure (PRAM) asthma index

**Table 3.** Decision Trees built on PRAM and non-PRAM sets

| Set | Size | Sens | Spec | Acc | AUC |
|---|---|---|---|---|---|
| Entire | 362 | 73 | 63 | 69 | 69 |
| **PRAM Set** | **147** | **93** | **96** | **95** | **98** |
| **Non-PRAM Set** | **206** | **89** | **53** | **74** | **77** |

[O'Sullivan08]

(Children's Hospital of Eastern Ontario, Ottawa, Canada)
[O'Sullivan08]

# Prominent Models

- Palaniappan and Awang demonstrate a web-based Intelligent Heart Disease Prediction System (IHDPS) to use medical profiles to predict the likelihood of patients getting a heart disease [Palaniappan08]
  - Techniques: Decision trees, naïve Bayes, neural networks
    - Each has its unique strength

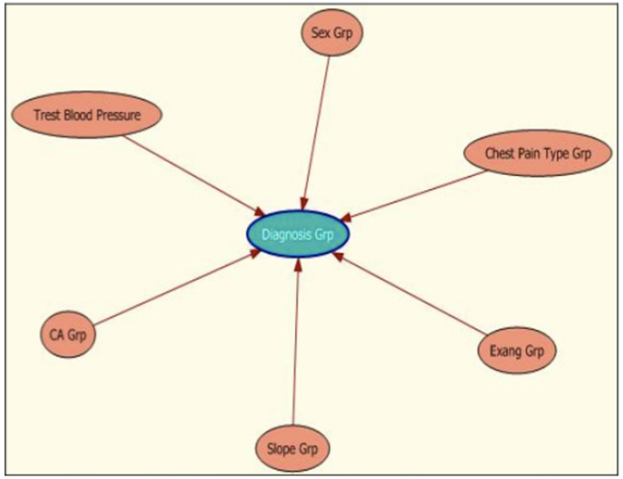



*Figure 7. Decision Trees dependency network* [Palaniappan08]

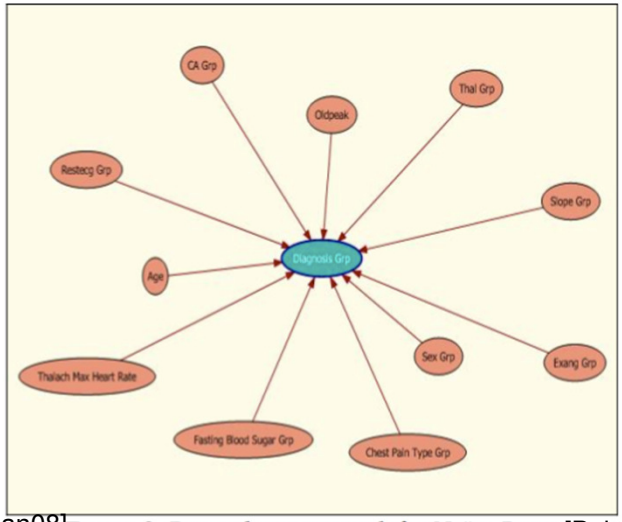



*Figure 8. Dependency network for Naïve Bayes* [Palaniappan08]

# Prominent Models

- Persson and Lavesson investigate prediction models for patient referrals [Persson09]
  - "A patient referral contains information that indicates the need for hospital care and this information is differently structured for different medical needs"

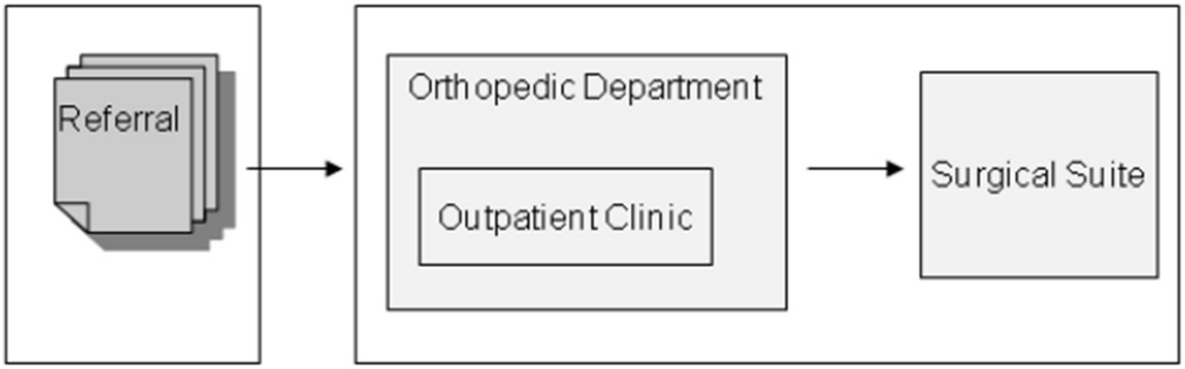



Fig. 2. A conceptual view of the studied healthcare process. The referrals from general practitioners are submitted to the hospital care.

# Prominent Models

- Kuttikrishnan et al. propose a system to assist clinicians at the point of care [Kuttikrishnan10]
  - Knowledge base: Rules and associations of compiled data
  - Inference engine: Combination of rules and patient's data
  - Mechanism to communicate: System-user interaction
  - Technique: Neural networks

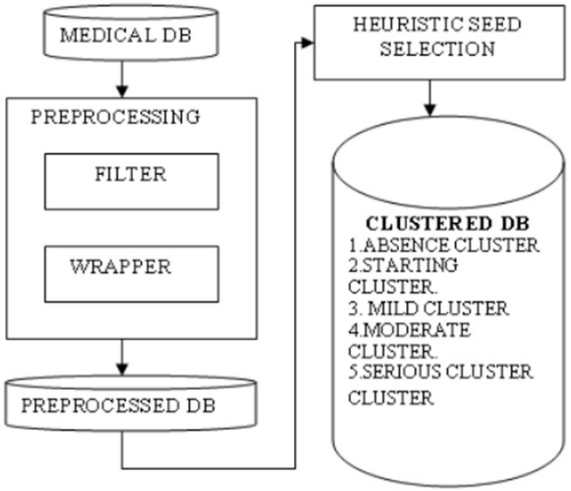



Fig. 1. Block Diagram [Kuttikrishnan10]

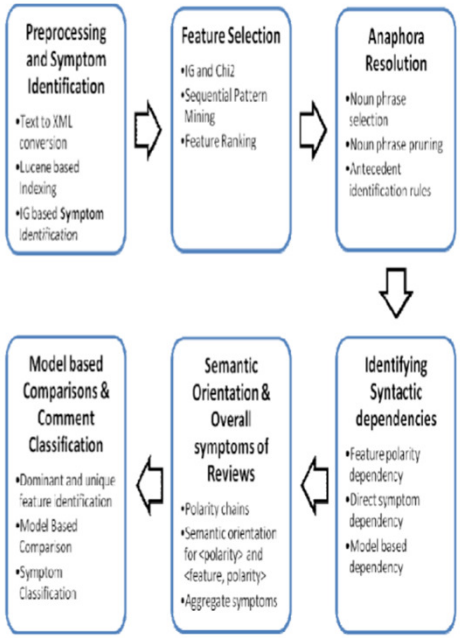



Fig. 9. Process flow [Kuttikrishnan10]

# Prominent Models

- De Toledo et al. discuss models of outcome prediction for subarachnoid hemorrhage (SAH) [DeToledo09]
  - Techniques: C4.5, fast decision tree learner, partial decision trees, repeated incremental pruning to produce error reduction, nearest neighbor with generalization, and ripple down rule learner
    - The best classifier is the C4.5 algorithm

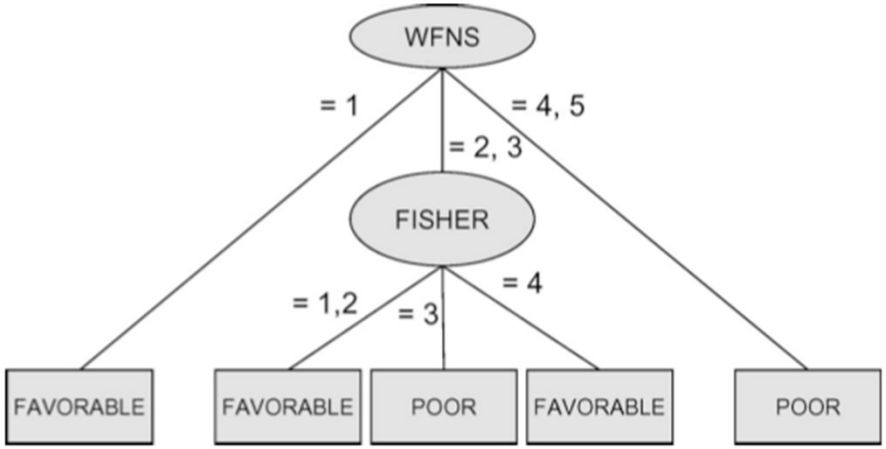



Fig. 1. Final classifier: C4.5 decision tree, dichotomized outcome. [DeToledo09]

# Prominent Models

- Razavi et al. discuss a model to predict recurrence of breast cancer [Razavi07]
  - "Identifying high-risk patients is vital in order to provide them with specialized treatment"
  - Technique: Decision tree

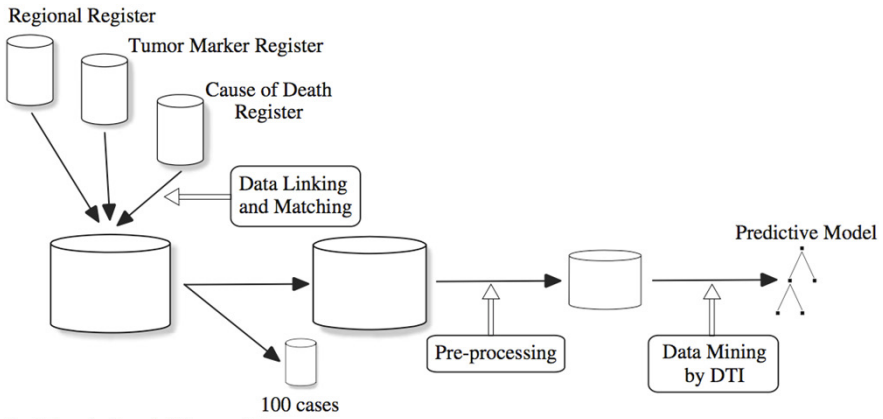



**Fig. 1** Steps leading to building a predictive model

**Table 4** Confusion matrix showing predictions done by oncologists and the decision tree in comparison with the real outcomes (no reply from oncologist 1 for one of the cases 1)

| | Real outcomes | Oncologist 1 | | Oncologist 2 | | DTI | |
|---|---|---|---|---|---|---|---|
| | | No rec | Rec | No rec | Rec | No rec | Rec |
| No recurrence | 81 | 79 | 2 | 70 | 11 | 78 | 3 |
| Recurrence | 19 | 17 | 1 | 8 | 11 | 15 | 4 |

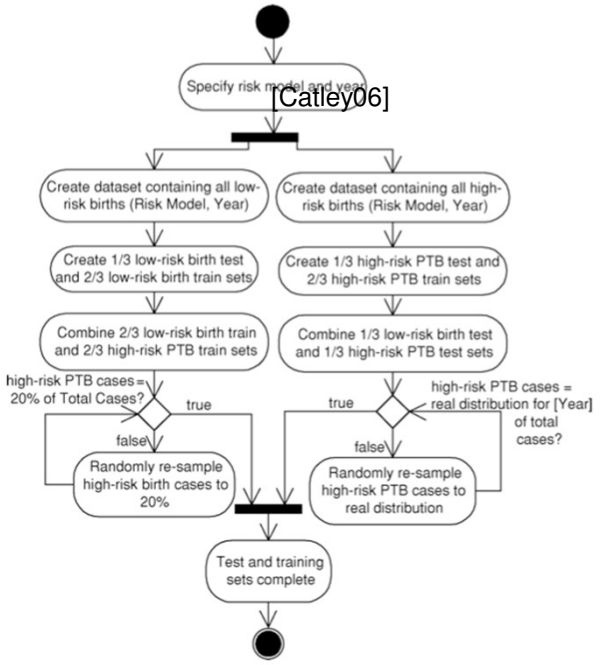
*No rec*: No recurrence; *Rec*: Recurrence; *DTI*: Decision Tree Induction.

# Prominent Models

- Catley et al. propose a screen tool to early prediction of preterm birth [Catley06]
  - Current procedure uses costly and invasive clinical testing

8 obstetrical variables were selected as being nonconfounding for predicting PTB:

- maternal age,
- number of babies this pregnancy,
- number of previous term babies,
- number of previous preterm babies,
- parity (total number of previous children),
- baby's gender,
- whether mother has intention to breastfeed,
  maternal smoking after 20 weeks gestation.




[Catley06]

Fig. 2. Flow of control for creating the artificial high-risk PTB model.

[Catley06]

**Some results:**

- Previous term birth is not a good indicator of future preterm birth.
- Maternal intention to breastfeed has minimal impact on results.

# Prominent Models

- Ahmad et al. propose to integrate clinical and microarray data for accurate breast cancer prognosis [Ahmad09]
  - "Breast cancer patients with the same diagnostic and clinical prognostics profile can have markedly different clinical outcomes"

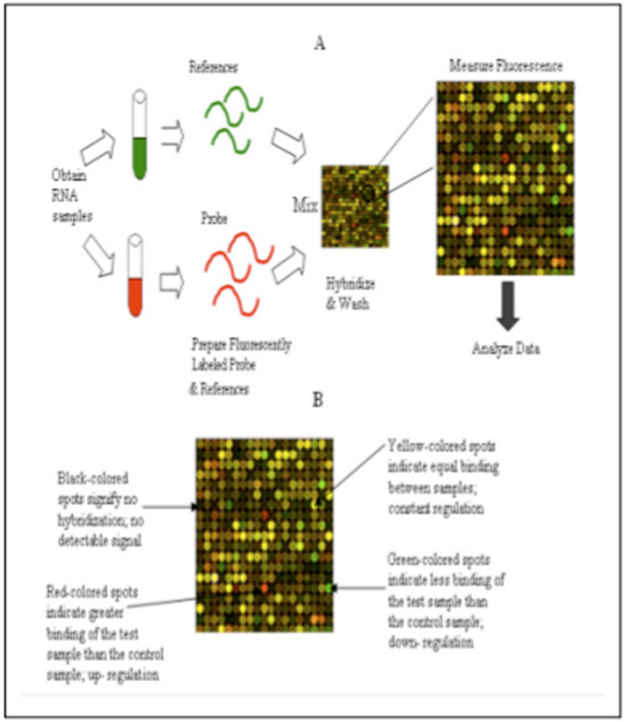
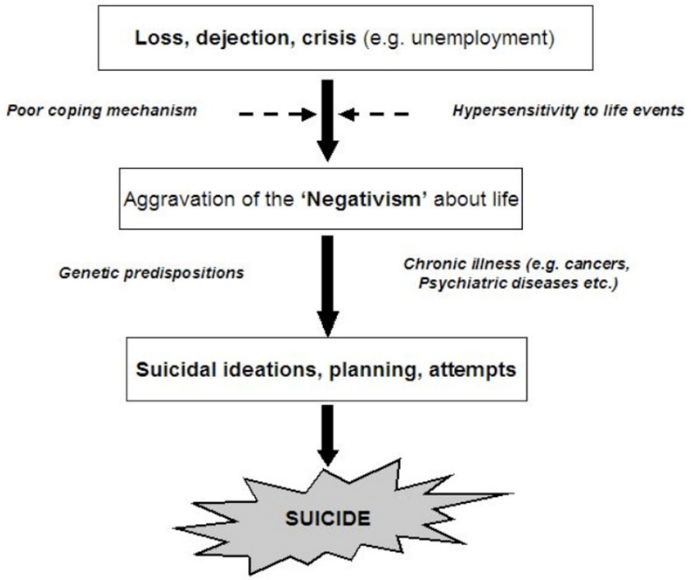



**Fig. 3.** Microarray experiment [Ahmad09]

# Prominent Models

- Chattopadhyay et al. discuss data-processing techniques for suicidal risk evaluation [Chattopadhyay08]
  - "Noise is eliminated incorporating the expertise of psychiatrists and psychologists case-by-case"
  - "Missing values are filled up with the most common value of the corresponding attributes"
  - "Correlation analysis has been done to identify which two data within an attribute are statistically similar"




[Chattopadhyay08]

**Fig. 1.** Roles of environment and neurobiological factors behind suicide

# Prominent Models

- Laxminarayan et al. propose a modified association rule mining technique to extract patterns from sequence-valued attributes such as sleep-related data by aggregating set of events that occur in the window [Laxminarayan06]
  - "Association mining may also be useful for selection of variables prior to using logistic regression"

$$\text{bmi-class} = \text{obese} \rightarrow \text{snore} = \text{yes}$$
$$[\text{Confidence} = 91.7\%; \text{Support} = 40.9\%;$$

$$\text{gasping} = \text{yes} \rightarrow \text{snore} = \text{yes}$$
$$[\text{Confidence} = 93.2\%; \text{Support} = 33.9\%;$$

[Laxminarayan06]

# Prominent Models

- McGarry et al. investigate graph-mining techniques for protein-protein interaction for diabetes research [McGarry07]
  - "The difficulties of complexity encountered by the pharmaceutical industry when developing the necessary assays for drug discovery have proved this conclusively, that there is no simple or direct link from genome to drugscular relationships." [McGarry07]

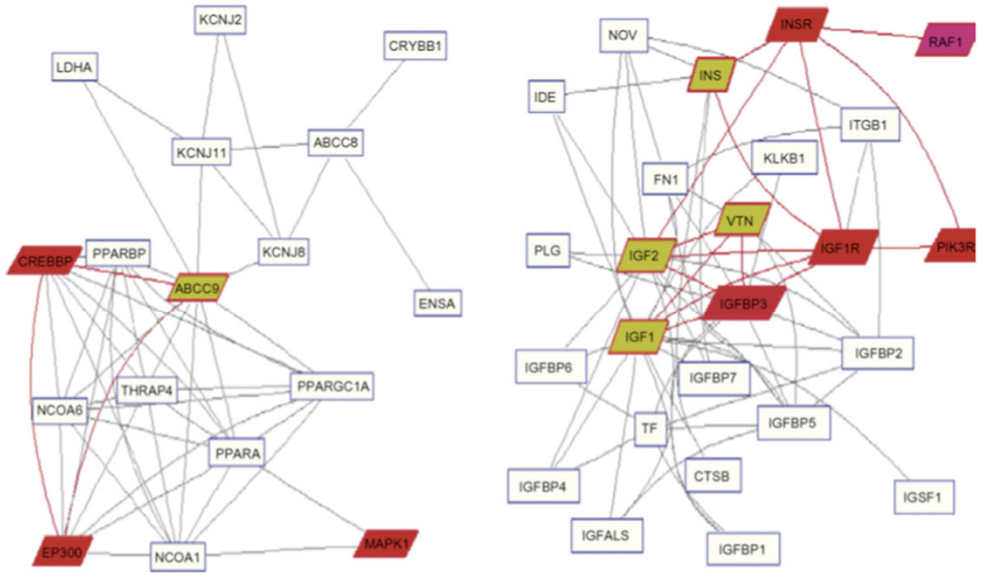



**Figure 10** Interaction diagrams for KCNJ11 (left) and IGF1 (right), proteins identified as essential using centrality measures are represented as colored parallelograms with nonessential proteins represented as squares.

[McGarry07]

# SAMPLE CASE STUDY

# Case Study1: Data Mining Based Decision Tool for Evaluating Treatment Choices for Uterine Fibroids [Campbell2010]

## Objectives

- Can data mining techniques be applied to data collected from patients with uterine fibroids in order to predict a treatment choice?
- Which data mining method is most successful in predicting the treatment decisions?

## Background – Uterine Fibroids

- Non cancerous tumor in muscular layer of the uterus
- 30-40% of women diagnosed
- Symptoms (vary significantly)
  - 50% of fibroids are asymptomatic; Heavy and painful menstruation; Abdominal discomfort; Painful sexual intercourse; Urinary frequency and urgency; Infertility
- Size and symptoms may subside after menopause

# Treatment Options

| Procedure | Hysterectomy | Myomectomy | Uterine Artery Emolization | Hormone Therapy | Watchful Waiting |
|---|---|---|---|---|---|
| Description | Surgical removal of the uterus involves hospital stay and lengthy recovery period. | Removal of one or more of the fibroids with open ab. Surgery or laparoscopic or endoscopic techniques. | The uterine artery is blocked with small particles; the fibroid is starved of its blood supply. | A drug treatment that causes fibroid shrinkage. | No treatment. On going monitoring. |
| Advantages | Permanent solution because the uterus is removed. | Preserves the uterus and cervix. | Symptom relief with shorter hospital stay than hysterectomy or myomectomy. | Non-surgical, conservative method of fibroid treatment. | Sometimes fibroid symptoms diminish with menopause. |
| Disadvantages | Reproductive potential is lost. Other side effects possible. Recovery time of several weeks. | Reoccurrence of fibroid symptoms possible if new fibroids grow. | Risks include radiation, menopause, serious infection, bleeding, embolization of other organs, hysterectomy, loss ovarian function. | Temporary. Causes menopausal symptoms. May result in rapid return of symptoms after treatment. | Fibroids may continue to grow with an increase in symptoms. |
| Return to Normal Activity | 28-56 Days | 4-44 Days | 10 Days | ____ | ____ |
| Hospital Time | 2-5 Days | 1-3 Days | 1 Day | 0 | 0 |
| Procedure Time | 1.5-3 Hours | 1-3 Hours | .75-2 Hours | 0 | 0 |

# Treatment Decision

- Patient ultimately makes the decision
- Many factors contribute to decision process
  - Pain, age, discomfort, sexual side effects, desire for children, etc.
- Research has not revealed any particular treatment as a "best practice"
  - There may not be a consistent way to measure "success"
- Human analysis of data for decision making can be flawed
  - Personality, anecdotal information
- Data mining could be used to direct women towards successful treatment options
  - Which treatments have women like you been happy with in the past? Which have they been unhappy with?

# Data

- Given that we do not have long term satisfaction scores, the scope of the project has been limited to the question:
  - "Which treatment have women like you chosen to pursue in the past?"
- Data
  - 171 Patients
  - 70 attributes (many redundant or metadata)
  - Data from a survey taken by patients identified as having recently made a decision about uterine fibroids
  - 8 different clinics in Minnesota between Feb 2007 and Feb 2008

# Data

2. In the past few months, have you been bothered by any of the following symptoms that may be caused by uterine fibroids? Please fill in the number that describes how much these symptoms have been a bother to you.

| | 0=Not bothersome or no symptoms | | | | | | | | | | 10=Extremely bothersome |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| a. Heavy bleeding | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| b. Pelvic pressure or pain | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| c. Urinary frequency | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| d. Infertility | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

3. What has been the overall impact of fibroid-related symptoms on your everyday activities in the past few months?

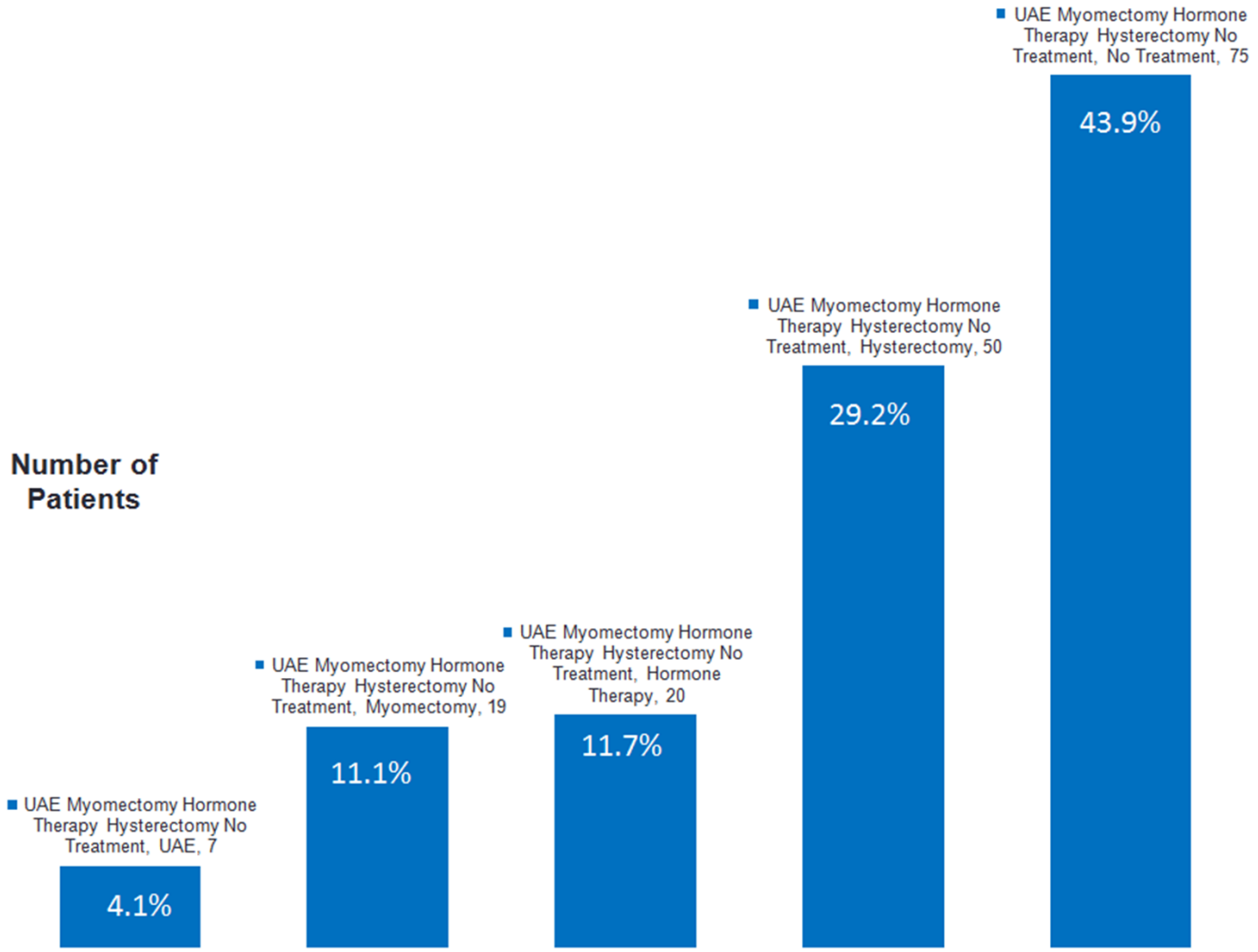
| 0=No interference or no symptoms | | | | | | | 10=Complete interference | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

# Data

7. Thinking about your most recent treatment decision for uterine fibroids, please rate each of the outcomes.

| How important was it for you to: | 0=Not at all important | | | | | | | | | | 10=Very important |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| a. Avoid taking medication | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| b. Do something right away to relieve symptoms | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| c. Keep your ability to have a baby | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| d. Minimize the amount of time you would spend recuperating from a treatment | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| e. Improve your sexual functioning | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| f. Have a permanent treatment | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| g. Have a treatment with a low failure rate | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

# Treatment Choices



# One Vs. Rest Scheme

- Improved accuracy
- Rare class problem
- Global accuracy does not indicate performance from minority class perspective

- Example:
  - In predicting hormone treatment:

| A | B | Classification |
|---|---|---|
| 51 | 0 | A = Not Hormone Therapy |
| 7 | 0 | B = Hormone Therapy |

  - 87.931 accuracy
  - **Recall (minority class perspective) = 0**

# Top 15 attributes as ranked by infogain for each treatment

| Rank | Hormone Therapy | | Hysterectomy | | Myomectomy | | No Treatment | | UAE | | | Key | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Attr. | Val | Attr. | Value | Attr. | Value | Attr. | Value | Attr. | Va | | Attribute | |
| 1 | PAM | >6 | PPT | >9 | PHB | >7 | PRR | <6 | BI | >2 | | PAM | Prefer to avoid medication |
| 2 | PAM | >1 | PPT | =10 | PHB | >8 | PRR | <7 | BI | >1 | | PPT | Prefer permanent treatment |
| 3 | PPT | <10 | PPT | >7 | PHB | >9 | PRR | <8 | DUR | 3 | | QH | Answer to hysterectomy quiz question |
| 4 | QH | * | PPT | >8 | PHB | =10 | PRR | <5 | BI | >3 | | QT | Answer to treatment quiz question |
| 5 | PAM | >7 | PPT | >3 | PHB | >6 | IFD | <6 | PSR | 7 | | BB | Bothered by bleeding |
| 6 | PAM | >4 | PPT | >5 | BUF | 8 | PRR | <9 | IFD | 5 | | QN | Number quiz questions correct (5 total) |
| 7 | PAM | >5 | IFD | >4 | PHB | >4 | IFD | <2 | BB | 4 | | IFD | Interferes with daily activities |
| 8 | PAM | >2 | IFD | >6 | Race | * | PRR | <1 | BI | >1 | | PHB | Prefer to have a baby |
| 9 | QT | * | BB | 8 | PHB | >5 | IFD | =0 | BI | >8 | | MB | Score of most bothering symptom |
| 10 | PAM | >3 | MB | 7 | BB | 1 | IFD | <1 | BI | >9 | | PRR | Prefers rapid relief |
| 11 | PAM | >8 | PRR | >8 | PHB | >3 | PRR | <5 | PSR | 6 | | BUF | Bothered by urinary frequency |
| 12 | BB | 8 | IFD | >2 | DUR | 2 | MB | <4 | BB | 3 | | DUR | Duration |
| 13 | QN | 4 | PPT | >6 | PHB | >1 | IFD | <3 | DUR | 1 | | BI | Bothered by infertility |
| 14 | PAM | =10 | IFD | >1 | AGE | <50 | MB | <5 | BI | >7 | | PSR | Prefer a short recuperation |
| 15 | AGE | 35 | PRR | >6 | PHB | >2 | BB | <1 | AGE | 50 | | AGE | Age of the patient |

- =, <, or > relationships are inferred with expert knowledge but not indicated with an infogain feature
- Example: A patient who scored their preference to have a baby as 7 or greater will be more likely to choose myomectomy for a treatment

# Top 3 performing algorithm-ensemble-attribute combinations by 3 different metrics

| | | Treatments | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Performance Metric | Rank | UAE | | Myomectomy | | Hormone | | Hysterectomy | | No Treatment | |
| | | 4.10% | | 11.10% | | 11.70% | | 29.20% | | 43.90% | |
| | | Algorithm | Metric | Algorithm | Metric | Algorithm | Metric | Algorithm | Metric | Algorithm | Metric |
| F-Measure | 1 | A NB (30) | 0.5258 | A NB (30) | 0.4106 | A NB (25) | 0.4831 | B J48 (15) | 0.691 | B J48 (15) | 0.7781 |
| | 2 | A NB (35) | 0.5093 | A NB (25) | 0.4039 | S NB (50) | 0.4696 | S J48 (15) | 0.6862 | B J48 (20) | 0.7772 |
| | 3 | S NB (30) | 0.4163 | A NB (25) | 0.4024 | B NB (50) | 0.4695 | B SC (15) | 0.6842 | S J48 (15) | 0.7757 |
| Accuracy (%) | 1 | A NB (35) | 95.15 | A J48 (50) | 87.97 | A MP (35) | 87.68 | B JR (35) | 79.63 | B J48 (15) | 78.55 |
| | 2 | A NB (30) | 94.86 | A SC (50) | 87.61 | A SC (30) | 87.14 | A JR (35) | 78.99 | B J48 (20) | 78.55 |
| | 3 | A J48 (35) | 94.42 | A JR (100) | 87.25 | A SC (25) | 87.14 | B MP (15) | 78.7 | S J48 (15) | 78.12 |
| Area Under ROC | 1 | A NB (30) | 0.9709 | B NB (30) | 0.8399 | S NB (25) | 0.8571 | S NB (35) | 0.87 | S NB (15) | 0.852 |
| | 2 | A NB (25) | 0.9683 | B NB (35) | 0.8382 | S NB (30) | 0.8553 | B NB (25) | 0.8694 | B NB (15) | 0.8519 |
| | 3 | B NB (35) | 0.9651 | S NB (30) | 0.8375 | B NB (25) | 0.855 | S NB (15) | 0.8647 | B NB (20) | 0.849 |

- Abbreviation = Ensemble + Algorithm + (#attributes used)

Ensemble Techniques

- A = Adaboost
- B = Bagging
- S = Simple (no ensemble)

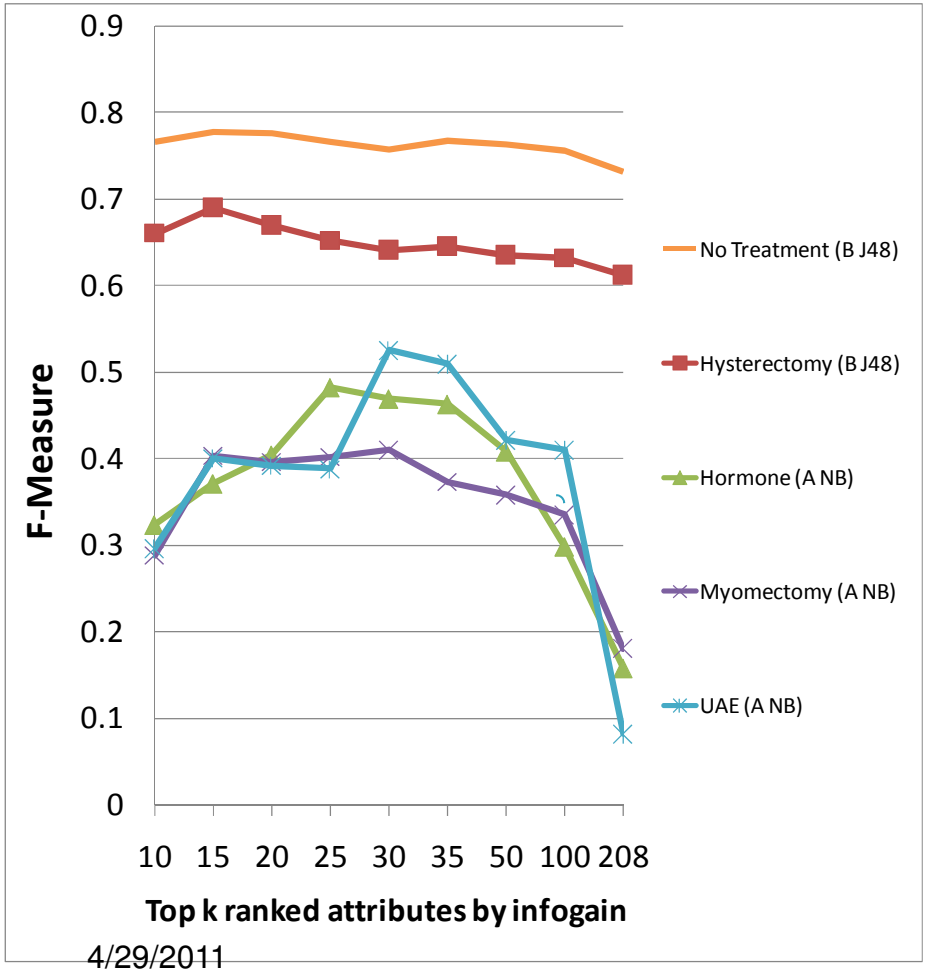
Algorithms

- NB = Naïve Bayes
- SC = Simple Cart
- MP = Multiperceptron
- JR = Jrip

# Observations and Conclusions

**Observations**

- Worse performance with too few or too many attributes
- Peak is at greater attribute number for rare classes
- Rare classes or Naïve Bayes more sensitive to number of attributes




## Conclusions

- Classifiers are making predictions that are better than random and would improve with more data
- Which classification algorithms are most successful in predicting the treatment decisions for the patients?
- Rare Classes:
  - Adaboost-Naïve Bayes (25-30 attributes ) selected
- Common Classes
  - Bagging-J48 (15 attributes)

# SUMMARY AND FUTURE DIRECTIONS

# Summary

- An introduction to healthcare management and the motivation to study this field and its impact on current research and market trends.
- Research discussion
  - types of available data,
  - the challenges involved
  - prominent models.
- A sample case study is presented to demonstrate how a certain application challenge can be addressed and the value of using data mining as a tool.

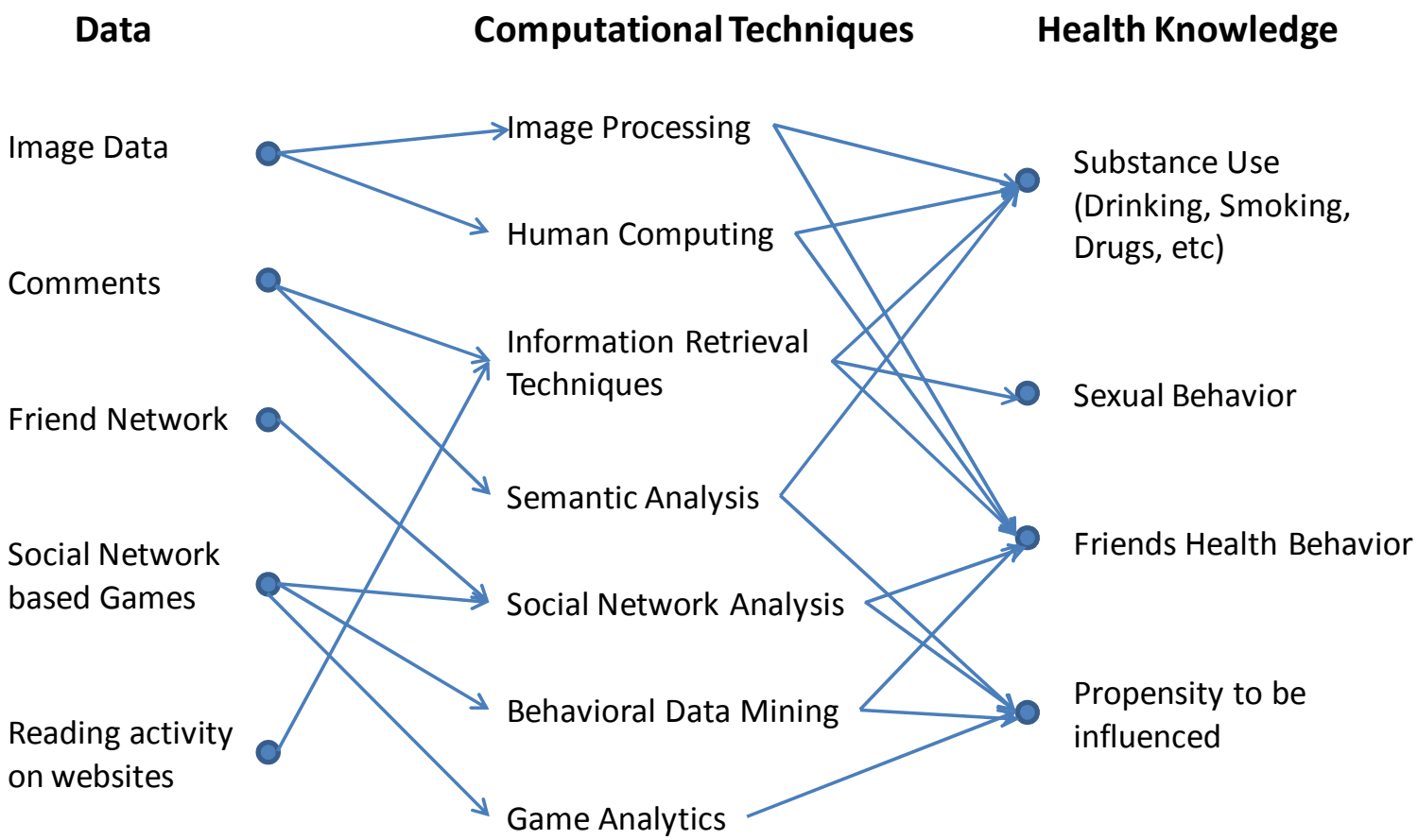
# Future Directions

- Data Storage and Access
- Data collection and analysis
- Integration of models from other domains
- Theoretical and Applied Research.

# Data Storage and Access

- Ensuring standard formats evolve and are adapted readily.
- Knowledge extraction by integrating data from various sources and formats.
- Storage and Access Mechanisms
  - Handling Dynamic Schema Changes
  - Flexible querying mechanisms

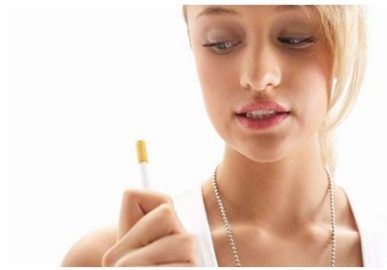
# Data collection and analysis –e.g. Collecting data from Social networking sites



# Social Ecological Model

| Societal | Organizational | Interpersonal | Intrapersonal |
|---|---|---|---|

Smoke-free workplace policy → Smoke-free school → Reduced exposure to smoking peers → Smoking isn't that cool



Smoking isn't that cool ← Fewer friends begin smoking ← No tobacco sponsor-ship of youth events ← Reduced community tolerance for cig-arette smoking

# Challenges and Issues to incorporate Social Network Data

- Technical challenges
  - Data extraction
    - Social Network Aggregation: Gathering information from various social networks and combining them at a single location
    - De-duplication: Identifying a user from multiple sources as the same users
  - Image processing
    - The limitation of automatic image annotation to capture and understand health related behavior
  - Social network analysis
    - Scalability issues : the large scale datasets that we should analyze
- Logistical and ethical issues
  - Some sites have strict privacy controls, difficult to obtain data
  - Data on publicly visible sites were not intended to be used for research purposes
    - Necessary to obtain consent?
    - If so, how to obtain consent from a large, complete network?
    - Parental vs. adolescent consent

# Conducting theoretical and applied research in this emerging field

- Scientific Research – use publicly available datasets such as UCI datasets
  - Scientific breadth
  - Replicable by other researchers
- Some publicly available datasets
- UCI Machine Learning Repository
- KDD Cup 2008 -Siemens (*Requires registration*)
- MIT-BIH Arrhythmia Database
- ECML/PKDD discovery challenge dataset.
- Healthcare Cost and Utilization Project (H-CUP)
- HIV Prevention Trials Network - Vaccine Preparedness Study/Uninfected Protocol Cohort
- National Trauma Data Bank (NTDB)
- Behavioral Risk Factor Surveillance System (BRFSS)
- Link to National Public Health Data Sets

(http://www-users.cs.umn.edu/~desikan/pakdd2011/datasets.html)

# Conducting research with high impact

- High Impact Research – use private datasets from collaborations
  - Need industry collaboration
  - Not easily replicable
  - Scientific breadth and validity may not be to a great extent
  - Basic research question may arise from collaboration
- Develop right interactive partnership. What can be obtained?
  - Access to data,
  - Understand what the problems are, and
  - You have people with power to implement ideas and help evaluate effectiveness

Healthcare Industry ↔ Academia Collaboration

# Key References


- [Ahmad09] Farzana Kabir Ahmad, Safaai Deris and Nor Hayati Othman, “Imperative Growing Trends toward Applying Integrated Data in Breast Cancer Prognosis”, MASAUM Journal of Basic and Applied Sciences, Vol.1, No.1 August 2009
- [Berwick97] DM Berwick, “The total customer relationship in health care: broadening the bandwidth”, Jt Comm J Qual Improv. 1997 May;23(5):245-50
- [Bolton02] Richard J. Bolton and David J. Hand, “Statistical Fraud Detection: A Review”, Statistical Science 2002, Vol. 17, No. 3, 235–255
- [Burroni2004] Burroni M, Corona R, Dell’Eva G, et al. Melanoma computer-aided diagnosis: reliability and feasibility study. Clin Cancer Res 2004;10:1881–1886.
- [Buczak09] Buczak, A.L.; Moniz, L.J.; Feighner, B.H.; Lombardo, J.S.; Mining electronic medical records for patient care patterns. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2009
- [Campbell2010] Kevin Campbell, N. Marcus Thygeson and Stuart Speedie. Exploration of Classification Techniques as a Treatment Decision Support Tool for Patients with Uterine Fibroids*;* Proceedings of International Workshop on Data Mining for HealthCare Management, PAKDD-2010.
- [Catley06] Christina Catley, Monique Frize, C. Robin Walker, and Dorina C. Petriu, “Predicting High-Risk Preterm Birth Using Artificial Neural Networks”, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 3, JULY 2006
- [Cerrito07] P.B. Cerrito: Mining the Electronic Medical Record to Examine Physician Decisions, Studies in Computational Intelligence (SCI) 48, 113–126 (2007)
- [Chattopadhyay08] S. Chattopadhyay, P. Ray, H.S. Chen, M.B. Lee and H.C. Chiang, “Suicidal Risk Evaluation Using a Similarity-Based Classifier”, ADMA 2008, LNAI 5139, pp. 51–61, 2008

# Key References


- [DeToledo09] Paula de Toledo, Pablo M. Rios, Agapito Ledezma, Araceli Sanchis, Jose F. Alen, and Alfonso Lagares, “Predicting the Outcome of Patients With Subarachnoid Hemorrhage Using Machine Learning Techniques”, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 13, NO. 5, SEPTEMBER 2009
- [Doniger02] Scott Doniger, Thomas Hofmann, Joanne Yeh. Journal of Computational Biology. December 2002, 9(6): 849-864
- [Gulera05] Inan Gulera, Elif Derya Ubeyli, “ECG beat classifier designed by combined neural network model”, Pattern Recognition 38 (2005) 199 – 208
- [Hardin2008] J. Michael Hardin and David C. Chhieng, Clinical Decision Support Systems: Theory and Practice, Eta S. Berner (Editor), Springer Verlag, Health Informatics Series,44-63; 2008.
- [Hadzic2010] Maja Hadzic, Fedja Hadzic and Tharam Dillon et al. Mining of patient data: towards better treatment strategies for depression. International Journal of Functional Informatics and Personalised Medicine, 2010
- [Hub2000] Huber S, Medl M,Vesely M, Czembirek H, Zuna I, Delorme S. Ultrasonographic tissue characterization in monitoring tumor response to neoadjuvant chemotherapy in locally advanced breast cancer. J Ultrasound Med 2000;19:677–686.
- [Hibbard1997] Hibbard LS, McKeel DW Jr. Automated identification and quantitative morphometry of the senile plaques of Alzheimer’s disease. Anal Quant Cytol Histol 1997;19:123–138.

# Key References


- [Jiang06] Zheng Jiang, Kazunobu Yamauchi, Kentaro Yoshioka, Kazuma Aoki, Susumu Kuroyanagi, Akira Iwata, Jun Yang, and Kai Wang. 2006. Support Vector Machine-Based Feature Selection for Classification of Liver Fibrosis Grade in Chronic Hepatitis C. J. Med. Syst. 30, 5 (October 2006), 389-394
- [Koh05] Koh HC, Tan G. " Data mining applications in healthcare". J Healthc Inf Manag. 2005;19(2):64-72.
- [Kohli01] Rajiv Kohli, Frank Piontek, Tim Ellington, Tom VanOsdol, Marylou Shepard, Gary Brazel, "Managing customer relationships through E-business decision support applications: a case of hospital-physician collaboration", Decision Support Systems, Volume 32, Issue 2, December 2001, Pages 171-187
- [Kuttikrishnan10] Murugesan Kuttikrishnan, Indumathi Jeyaraman, Manjula Dhanabalachandran, "An Optimised Intellectual Agent Based Secure Decision System for Health Care", International Journal of Engineering Science and Technology, Vol. 2(8), 2010, 3662-3675
- [Kwiatkowska07] Mila Kwiatkowska, M. Stella Atkins, Najib T. Ayas, and C. Frank Ryan, "Knowledge-Based Data Analysis: First Step Toward the Creation of Clinical Prediction Rules Using a New Typicality Measure", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 11, NO. 6, NOVEMBER 2007
- [Laxminarayan06] Parameshvyas Laxminarayan, Sergio A. Alvarez, Carolina Ruiz, and Majaz Moonis, "Mining Statistically Significant Associations for Exploratory Analysis of Human Sleep Data", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 3, JULY 2006

# Key References


- [Ludwick09] Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform. 2009 Jan;78(1):22-31
- [Maglogiannis09] Ilias Maglogiannis, Introducing Intelligence in Electronic Healthcare Systems: State of the Art and Future Trends. Artificial Intelligence, LNAI 5640, pp. 71 – 90, 2009
- [McGarry07] Ken McGarry, James Chambers, and Giles Oatley. 2007. A multi-layered approach to protein data integration for diabetes research. Artif. Intell. Med. 41, 2 (October 2007), 129-14
- [O'Sullivan08] Dympna O'Sullivan, William Elazmeh, Szymon Wilk, Ken Farion, Stan Matwin, Wojtek Michalowski, and Morvarid Sehatkar, "Using Secondary Knowledge to Support Decision Tree Classification of Retrospective Clinical Data", MCD 2007, LNAI 4944, pp. 238–251, 2008
- [Ouschan06] Robyn Ouschan, Jillian Sweeney, Lester Johnson, "Customer empowerment and relationship outcomes in healthcare consultations", European Journal of Marketing, Vol. 40 Iss: 9/10, pp.1068 – 1086, 2006
- [Palaniappan08] Palaniappan, S.; Awang, R., "Intelligent heart disease prediction system using data mining techniques", IEEE/ACS International Conference on Computer Systems and Applications, 2008. AICCSA 2008
- [Penny09] Kay Penny and Thomas Chesney, "Mining Trauma Injury Data with Imputed Values", Statistical Analysis and Data Mining 2: 246 – 254, 2009
- [Persson09] Persson, M.; Lavesson, N., "Identification of Surgery Indicators by Mining Hospital Data: A Preliminary Study", 20th International Workshop on Database and Expert Systems Application, 2009. DEXA '09

# Key References


- [Puschmann01] Puschmann, T.; Alt, R., “Customer relationship management in the pharmaceutical industry”, Proceedings of the 34th Annual Hawaii International Conference on System Sciences, 2001
- [Razavi07] Amir R. Razavi & Hans Gill & Hans Ahlfeldt, Nosrat Shahsavar, “Predicting Metastasis in Breast Cancer: Comparing a Decision Tree with Domain Experts”, J Med Syst (2007) 31:263–273
- [Rygielski02] Chris Rygielski, Jyun-Cheng Wang, David C. Yen, “Data mining techniques for customer relationship management”, Technology in Society 24 (2002) 483–502
- [Santos09] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, J. Neves, A. Silva, F. Rua, M. Salazar, C. Quintas, and A. F. Cabral. 2009. Nursing information architecture for situated decision support in intensive care units. In Proceedings of the 9th WSEAS international conference on Applied informatics and communications (AIC'09), Nikos E. Mastorakis, Metin Demiralp, Valeri Mladenov, and Zoran Bojkovic (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 491-496
- [Stolba06] N. Stolba, A. Tjoa: "The Relevance of Data Warehousing and Data Mining in the Field of Evidence-Based Medicine to Support Healthcare Decision Making"; ICCS 2006
- [Young00] Young, Gary J.; Meterko, Mark; Desai, Kamal R., “Patient Satisfaction With Hospital Care: Effects of Demographic and Institutional Characteristics”, Medical Care: March 2000 - Volume 38 - Issue 3 - pp 325-334
- [Zellner2004] Zellner BB, Rand SD, Prost R, Krouwer H, Chetty VK.A cost-minimizing diagnostic methodology for discrimination between neoplastic and non-neoplastic brain lesions: utilizing a genetic algorithm. Acad Radiol 2004;11:169–177