Data Mining and Knowledge Discovery in Healthcare and Medicine

Bijan Raahemi, Ph.D., P.Eng, SMIEEE
Associate Professor
Telfer School of Management and
School of Electrical Engineering and Computer Science
University of Ottawa

April 30, 2014

Agenda

- Data Mining
- Data Mining in Healthcare
- Research 1:
  - Identifying High-cost Patients using Data Mining Techniques and a Small Set of Non-trivial Attributes
- Research 2:
  - Brain-based Biomarkers for Depression Diagnosis
- Questions

Data Flood

Large amount of Data are being produced

- Retail industry
- Bank, business transactions
- Telecom industry, Mobile phone services
- Web, text, and e-commerce
  - Google searches Billions of pages, hundreds of TB
- Scientific data: astronomy, biology
- Healthcare
- …

- Data is growing at a phenomenal rate
  - 90% of data generated in the last 2 years

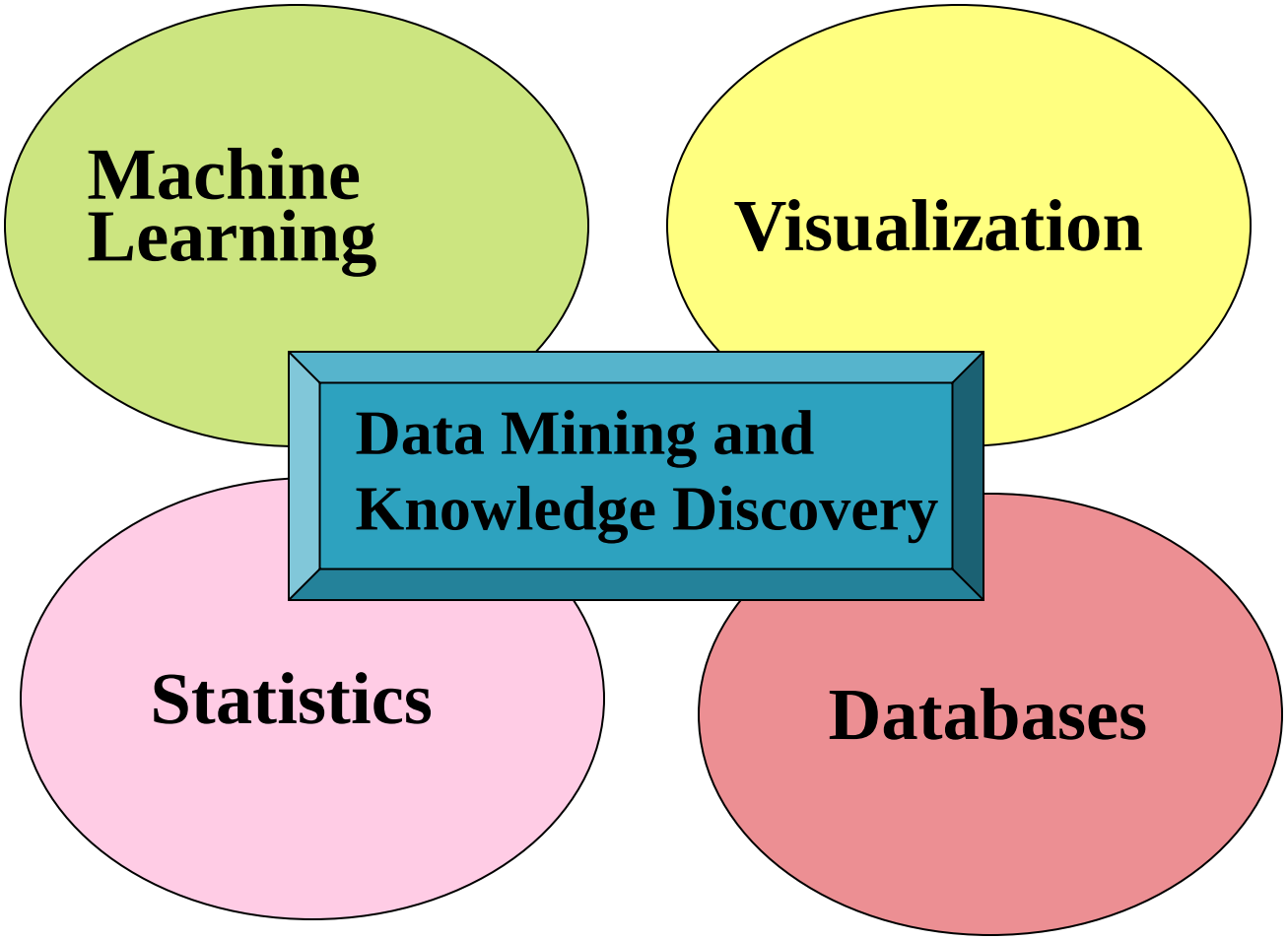
DATA MINING
UNCOVER HIDDEN INFORMATION

Data Mining is the exploration and analysis of a large quantities of data in order to discover meaningful patterns and rules.

- process of identifying *non-trivial*, *valid, novel,* potentially *useful,* and *understandable patterns* in data.

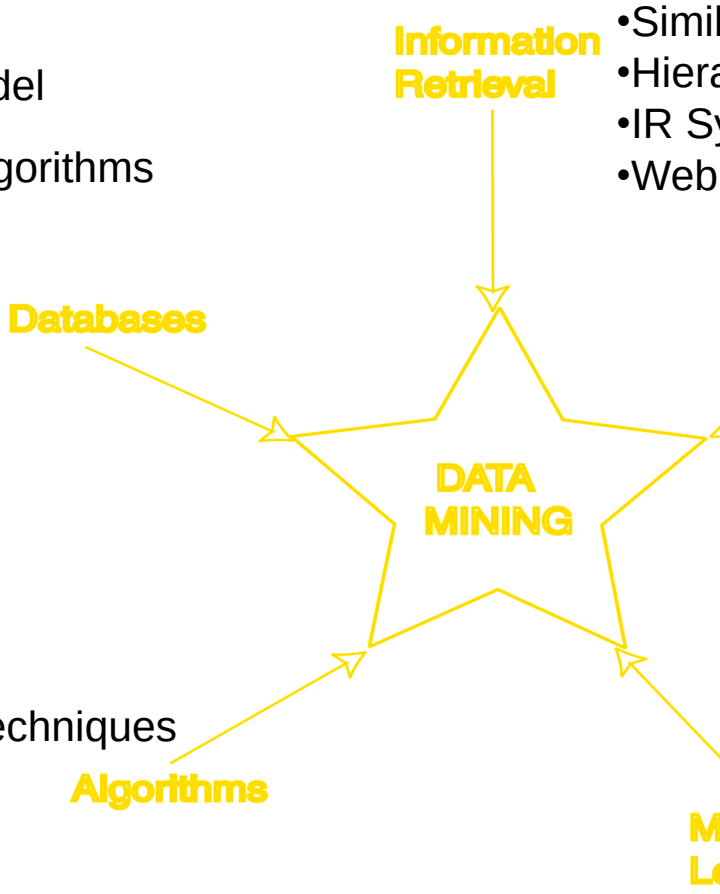
- Finding hidden information in a database

Related Fields



Source: [KDnuggets]

Related Fields (cont'd)



- Relational Data Model
- SQL
- Association Rule Algorithms
- Data Warehousing

- Similarity Measures
- Hierarchical Clustering
- IR Systems
- Web Search Engines

- Bayes Theorem
- Regression Analysis
- K-Means Clustering
- Time Series Analysis

- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Neural Networks
- Decision Tree Algorithms

Source: [Dunham]

Data Mining Techniques

- **Classification** : predefined groups
- **Clustering** – grouping data into categories with similar attributes. No predefined groups
- **Association Rules (Affinity Grouping)** – e.g. : people who buy product X, often buy product Y with probability Z%
- **Estimation** – e.g. : household income in a specific area
- **Prediction** – e.g. : predict stock market
- **Description and Profiling (Summarization)** – e.g. :“more men study engineering than women”

Applications of Data Mining

- Business:
  - Customer profiling/Segmentation
  - Target marketing
  - Market-Basket Analysis, Recommendation Program
  - Retention, Churn and hazard Analysis
  - Credit Risk
  - Fraud Detection

- Industries span a wide range

  Banking, telecom, retails sales, manufacturing, sports/entertainment

Applications of Data Mining (cont'd)

- Web:
  - search engines, bots
- Science
  - astronomy, bioinformatics, Genetics
- Government
  - law enforcement, profiling tax cheaters, counter terrorism

Applications of DM in Healthcare

(a) Clinical Medicine

- Using ANN on ECG for diagnosis AMI

(b) Public Health

- Detection of disease outbreaks and clustering patterns
- Prediction of SARS epidemics, its diagnosis and outcomes

(c) Policy/Planning

- Cost prediction in diseased or in general population
- Analyzing capacity problems using disease/ personal attributes

(d) Text mining

- Mining of medical literature or clinical data repositories

Research 1

Identifying High-cost Patients in General Population using Data Mining Techniques and a Small Set of Non-trivial Attributes

Seyed A. Izad Shenas.[1], Bijan Raahemi.[2], Craig Kuziemsky.[2]

[1].M.D., M.Sc. Health Systems, [2].Associate professors, University of Ottawa

KDD lab, University of Ottawa

Motivation

- **Rising health expenditure**

  US spent >$1T (14% GDP) in 1997, and $2.5T (17% GDP) in 2009 (CMS 2011, CDC 2009);

  Canada spent 11-12% GDP on health during 2008-2010 period (CIHI 2010)

- **Increasing cost of chronic diseases**

  75% of all costs goes to chronic diseases (DeVol & Bedroussian 2007; CDC 2009; CIHI 2010)

- **Aging population**

- **Existing methods**

  Mostly oriented to statistical modeling of disease-specific costs (Concaro 2009; Weinstein 2009; Khan 2006; Farley 2006)

- **Unique features of MEPS Survey database**

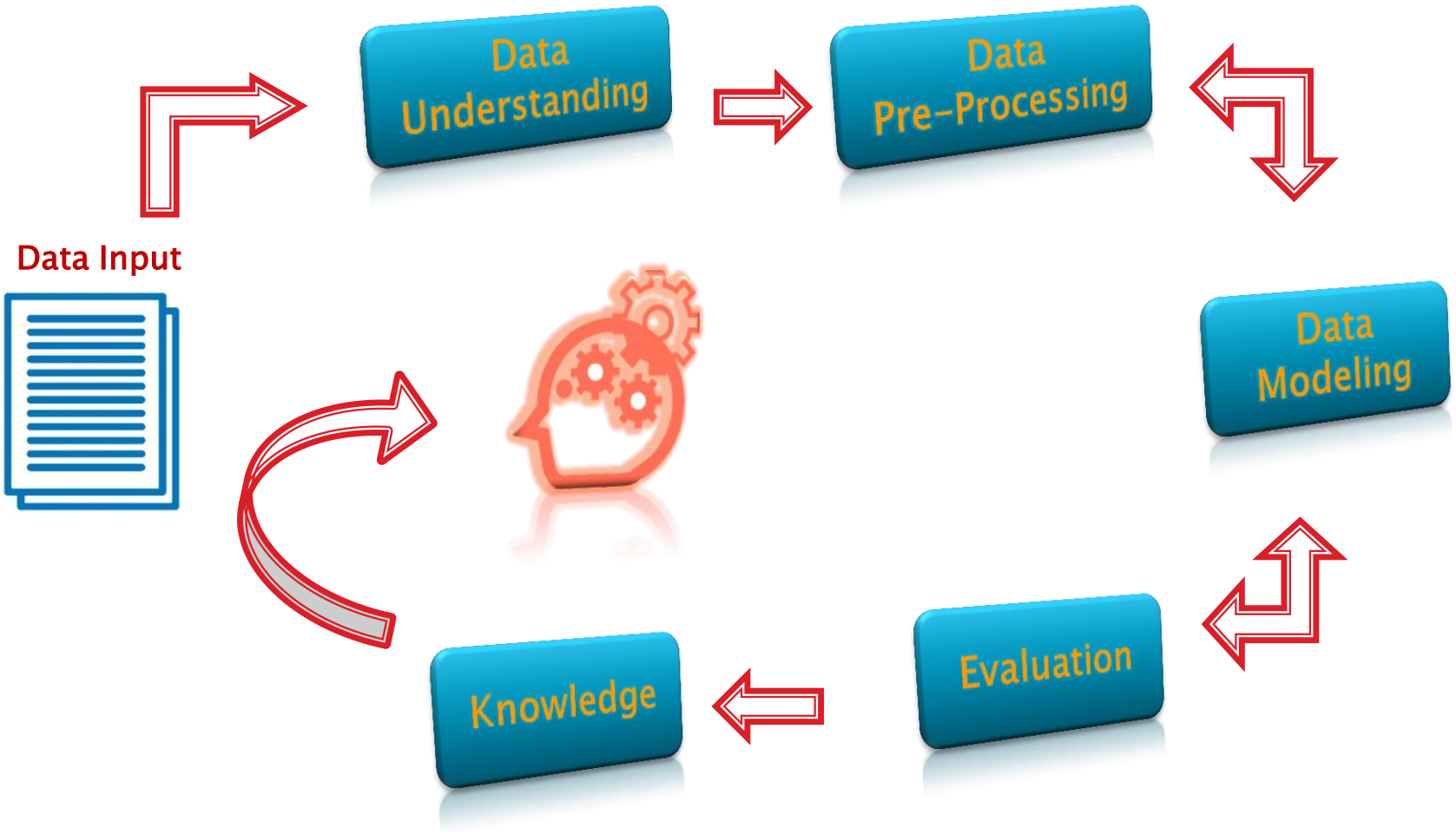
Research Objectives

1. Build ***predictive models*** including Decision Tree (DT) and Artificial Neural Network (ANN) to identify high-cost/low-cost patients.

2. Evaluate ***the accuracy*** of the predictive models.

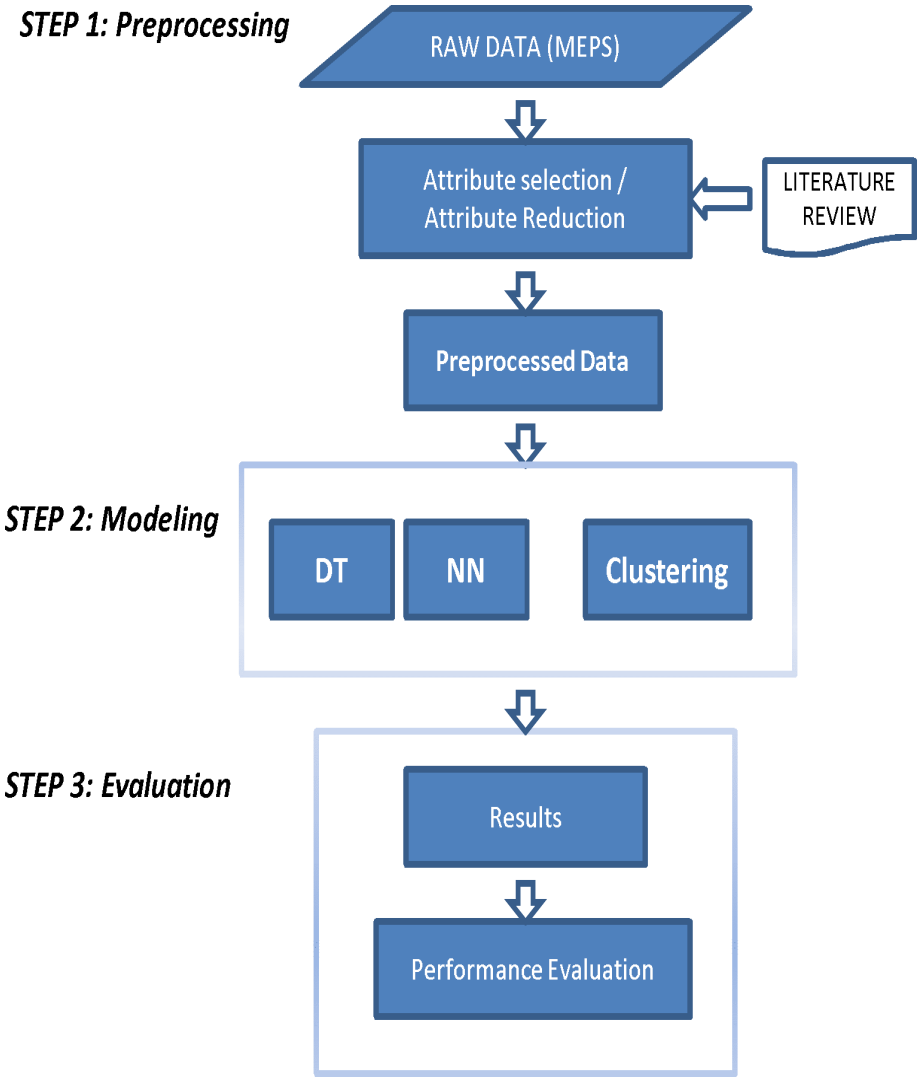
3. Identify the ***small set of non-trivial attributes*** which can reasonably predict the high-cost population.

Methodology (based on CRISP-DM process model)



Research Methodology

- Quantitative Modeling
- Tools
  - IBM SPSS Modeler 14.1
  - IBM SPSS Statistics
  - Minitab 16
  - MS Access
- Attribute selection
- Record partitioning / balancing
- Evaluation



The Dataset

Medical Expenditure Panel Survey (MEPS)

- Conducted by the American Agency for Healthcare Research and Quality (AHRQ)
- A panel/year, 5 interview rounds
- 30,000 records/year
- Household Component: 1600-1800 attributes
- HC 2006-2008: >90,000 records
- Released in ASCII and SAS file
- Available at http://meps.ahrq.gov/mepsweb/

Household Component Includes:

- demographics
- health status and health conditions,
- medical events by type (hospital including inpatient, outpatient, and ER; office-based; dental; home health; prescribed medications;
- other events including glasses, ambulance, equipment
- date and other details of an event
- charges and payments by source for each event
- employment profile,
- health insurance profile,
- income,
- access to care,
- level of satisfaction

Data Preparation: Targets

- TOTEXP1 (current yr)

Percentile	Expenditures ($)	Expenditures (%)
Top 5%	59,575,116	45
Top 10%	80,914,106	61
Top 20%	104,156,078	78
All Records (100%)	133,147,351	100

- TOTEXP2 (next yr)

Percentile	Expenditures ($)	Expenditures (%)
Top 5%	67,365,197	48
Top 10%	88,896,390	64
Top 20%	113,365,014	81
All Records (100%)	140,149,629	100

- 6 targets

Target Title	Target Year	High-cost Threshold
TOTEXP1-95	Current Year	> 95 Percentile
TOTEXP1-90	Current Year	> 90 Percentile
TOTEXP1-80	Current Year	> 80 Percentile
TOTEXP2-95	Next Year	> 95 Percentile
TOTEXP2-90	Next Year	> 90 Percentile
TOTEXP2-80	Next Year	> 80 Percentile

Building Models

Classification

All 5 modules (39 attributes)	DT / ANN
All 5 modules (39 attributes)	DT
Each module separately	C5.0 / CHAID
Combinations of modules	CHAID

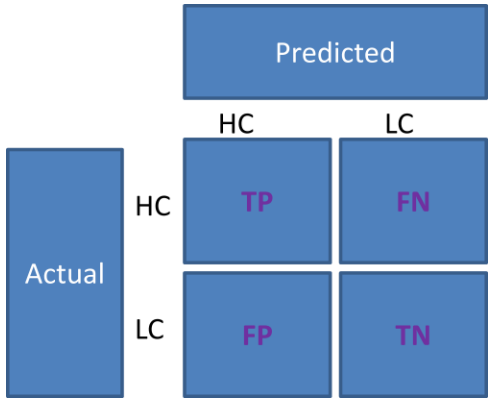
Clustering

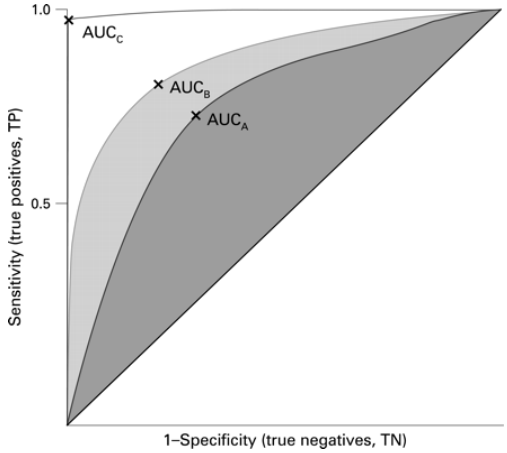
K-Means on top attributes	K-MEANS

Performance evaluation

DT / ANN

- Sensitivity (Sn) $TP/(TP+FN)$
- Specificity (Sp) $TN/(TN+FP)$
- Correctness accuracy $(TP+TN)/(TP+FN+TN+FP)$
- G-Mean (trade-off) $\sqrt{Sn * Sp}$
- Area under ROC curve (AUC)





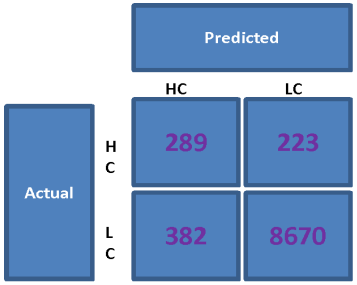
K-Means Clustering

- Silhouette measure

Results

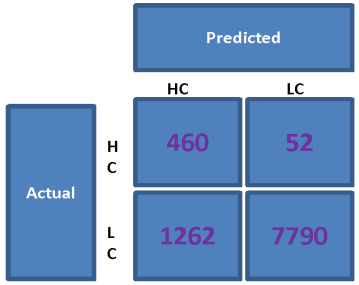
Performance measures

C5.0



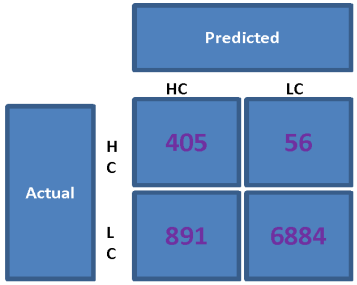
Sensitivity: 56%
Specificity: 96%
Accuracy: 94%
G-Mean: 73%
AUC: 0.816

CHAID



Sensitivity: 90%
Specificity: 86%
Accuracy: 86%
G-Mean: 88%
AUC: 0.946

NN



Sensitivity: 87%
Specificity: 86%
Accuracy: 76%
G-Mean: 86%
AUC: 0.956

- High accuracy comes from Sp not Sn.
- AUC / G-mean are better trading-off Sp vs. Sn.

AUC / G-MEAN superior to Accuracy

Results

C5.0 / CHAID Models

Demographics

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	67.0	0.784	56	68	62
Chaid	69.3	0.717	62	70	66

Health

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	75.8	0.772	69	76	72
Chaid	73.8	0.789	72	74	73

Preventive

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	70.6	0.787	65	71	68
Chaid	70.6	0.743	68	71	70

PrioC

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	77.3	0.799	61	78	69
Chaid	75.2	0.772	67	76	71

Visits

Model	Accuracy	AUC	ST	SP	G-MEAN
C5.0	91.9	0.765	67	93	79
Chaid	88.3	0.944	87	88	87
NN	88.5	0.954	90	88	89

CHAID, better than C5.0

CHAID: Combination of Attribute

	Accuracy	G-Mean	AUC
DHealth	78 (74)*	74 (74)	0.804 (0.801)
DPreventive	69 (69)	71 (70)	0.752 (0.763)
DPrioC	75 (71)	71 (70)	0.784 (0.768)
Dvisits	87 (88)	86 (87)	0.937 (0.944)

*Numbers in parentheses show the performance Measures for the combination models when they use all attributes (large set).

DHealth	
RTHLTH	45
AGE	28
ANYLIM	22
SEX	5

DPrioC	
PCCOUNT	81
AGE	12
SEX	4
REGION	3

DPreventive	
CHOLCK	47
AGE	32
BOWEL	15
SEX	5

DVisits	
IPDIS	47
OBTOT	45
AGE	8
SEX	1

ALL
PCCOUNT
CHOLCK
IPDIS
RTHLTH
OBTOT
AGE
ANYLIM
BOWEL
SEX
REGION

Reduced attributes

/

Age / Sex shared

Recommendations

Health planners

- Survey design
- Disease management programs

Policy makers

- Resource allocation
- Policy evaluation

Insurance firms

- Customization of insurance plans
- Client screening

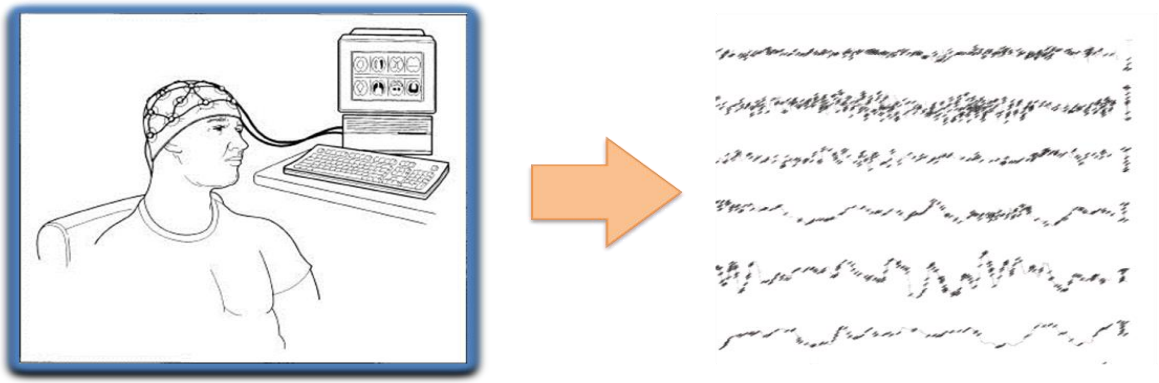
Research 2

Brain-based Biomarkers for Depression Diagnosis

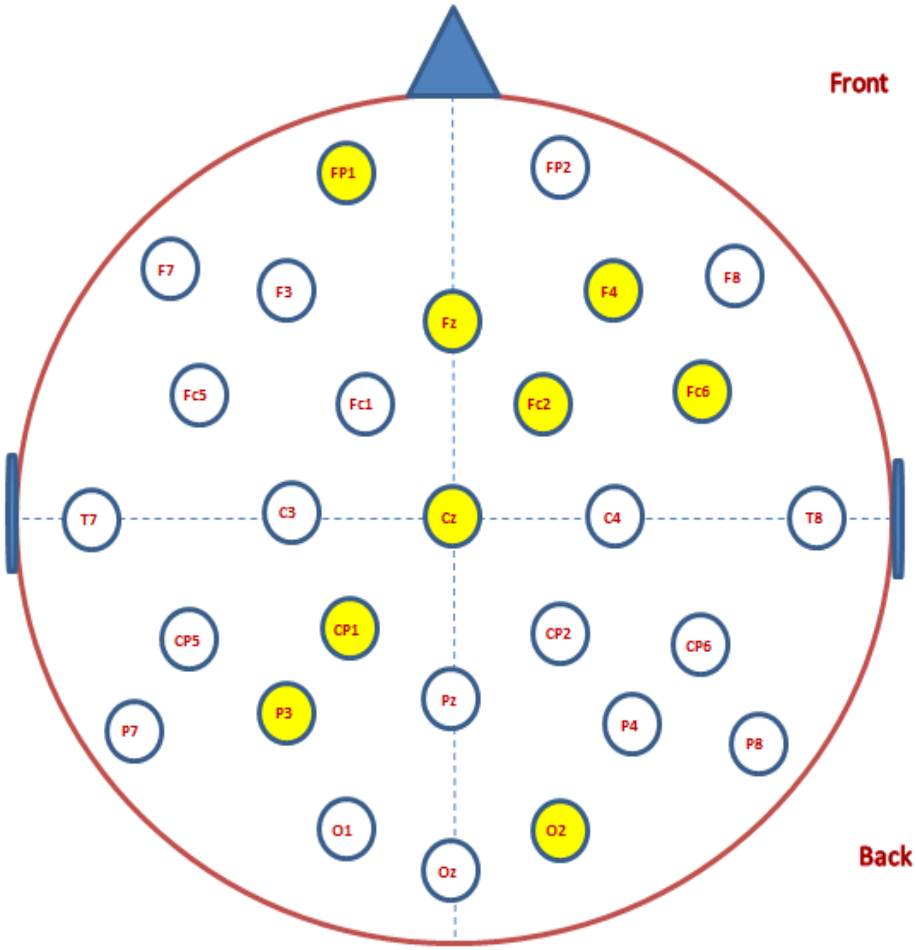
F. Alazab, B. Raahemi
KDD lab, University of Ottawa

Introduction

- Electroencephalogram (**EEG**): A record of the electric activity from the scalp, obtained with the aid of an array of electrodes.
- **EEG** measures voltage fluctuations resulting from ionic current flows within the neurons of the brain.
- **EEG** is used in cognitive science, cognitive psychology, and psychophysiological research



28 Probes


Front
FP1
FP2
F7
F3
Fz
F4
F8
Fc5
Fc1
Fc2
Fc6
T7
C3
Cz
C4
T8
CP5
CP1
CP2
CP6
P3
Pz
P4
P7
P8
O1
O2
Oz
Back

EEG for Depression Diagnoses

Input Data

- EEG signals record the electric activities of the brain by attaching **28 electrodes** on the scalp to collect **four frequency bands** (Alpha, Beta, Delta and Theta) from **Mastiod** and **Cz** sites, during the eyes closed **EC** and eyes opened **EO.**
- In this study, we focused on analysing the four bands collected from **96 persons, 43 healthy controls and 53** depressed patient before medication.

Output

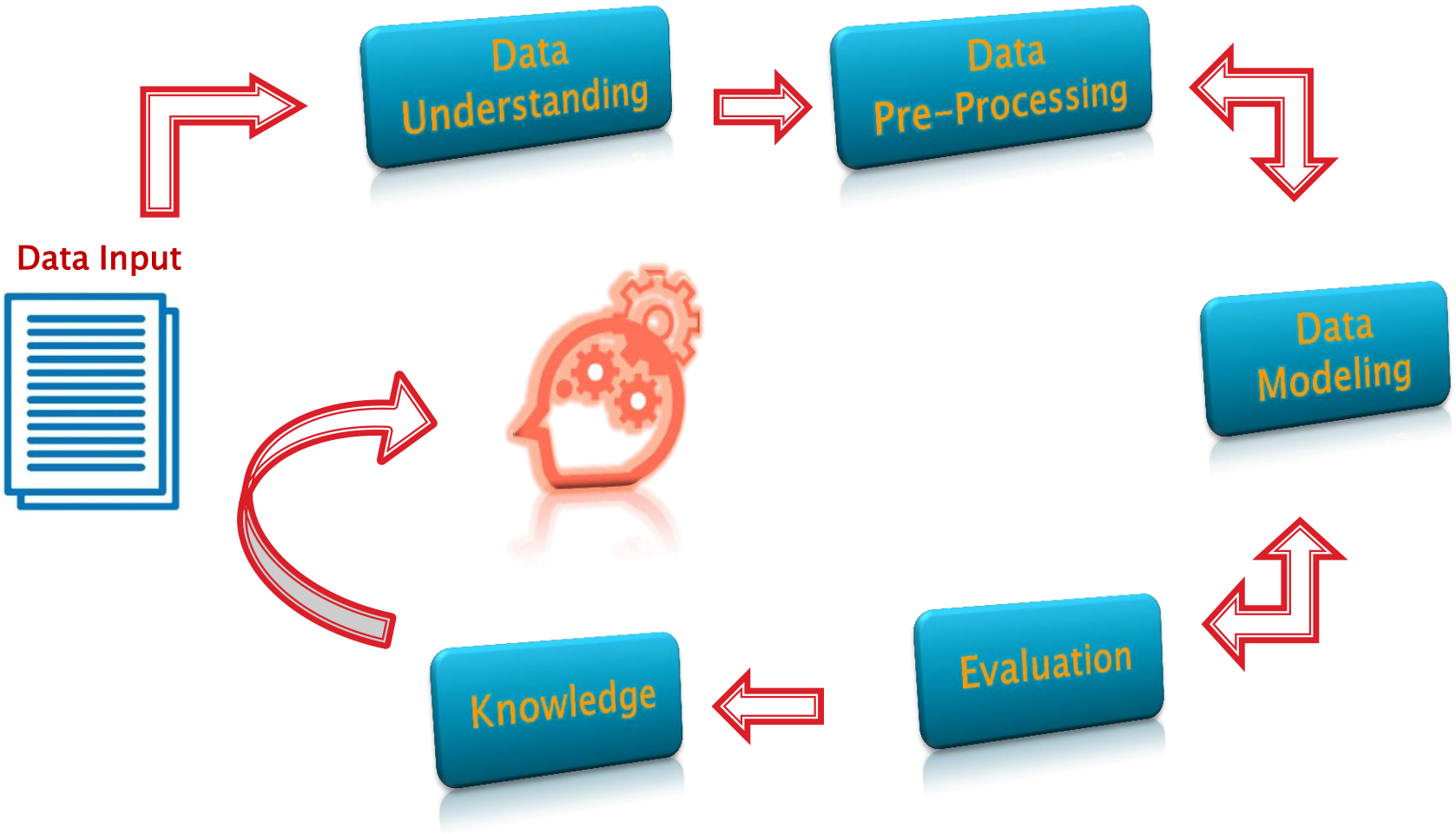
- Determine whether a patient is depressed based on the input of EEG signals**.**

Objectives

- Applying **Data Mining** methods to:

  - ✓ Reduce Features by identifying the most significant signals
  - ✓ Map Features by discriminating them based on the class
  - ✓ Build a Predictive Model to identify depressed patients based on their EGG signals

Methodology (based on CRISP-DM process model)



Data Understanding

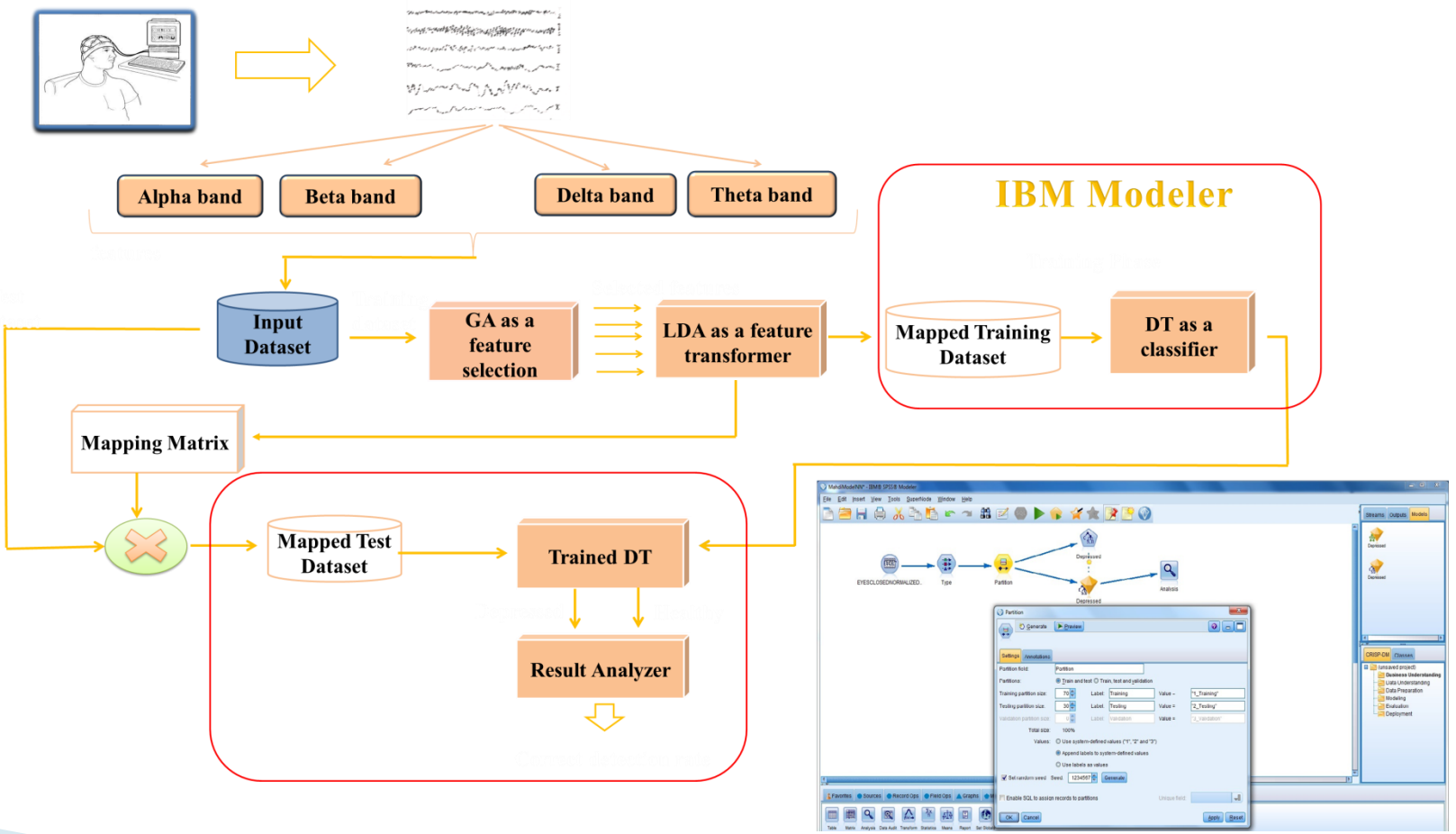
- **Mastoid Bone**

- 4 frequency Bands, 28 features per band, and a target class for Eyes Opened and Eyes Closed status.
  - Alpha at (8-11 Hz) , (11-14 Hz) and the total of both.
  - Beta
  - Theta
  - Delta

- **Cz Bone**

- 4 frequency Bands, 28 features per band, and a target class for Eyes Opened and Eyes Closed status.
  - Alpha at (8-11 Hz) , (11-14 Hz) and the total of both.
  - Beta at (8-11 Hz) , (11-14 Hz) and the total of both.
  - Theta
  - Delta

Building Models



Experiment 1: Bands Analyzed Individually

- Eyes Opened & closed result for Mastoid bone

Bands	Number of Features	Accuracy %	FP% Healthy	FP% Depressed
Alpha	29 features of EC	75%	4	19
	12 features (GA+LDA)	**83%**	**5**	**8**
	29features of EO	52%	1	44
	15 features (GA+LDA)	**80%**	**9**	**10**
Beta	29 features of EC	65%	26	7
	10 features (GA+LDA)	**78%**	**9**	**12**
	29 features of EO	80%	10	9
	11 features (GA+LDA)	**77%**	**9**	**13**
Delta	29 features of EC	52%	1	45
	18 features (GA+LDA)	**80%**	**8**	**11**
	29 features of EO	62%	11	25
	10 features (GA+LDA)	**74%**	**21**	**4**
Theta	29 features of EC	75%	4	19
	13 features (GA+LDA)	**83%**	**5**	**8**
	29 features of EO	72%	15	12
	10 features (GA+LDA)	**82%**	**4**	**13**

Experiment 2: Bands Analyzed Together

Alpha (8-11Hz) , Beta, Delta & Theta (Total samples 91)

- **Eyes Closed (Mastoid Reference)-**

Methods	Accuracy	FP% – Healthy	FP% – Depressed
112 raw features of EC	84%	6	9
60 features processed by GA+LDA	94%	4	2

- **Eyes Opened (Mastoid Reference)-**

Methods	Accuracy	FP% – Healthy	FP% – Depressed
112 raw features of EO	88%	8	3
58 features processed by GA+LDA	94%	2	4

Conclusions

- We found that **combining the four bands together during the analysis showed a higher accuracy** comparing to individual analysis, meaning those bands complement each other for better identifying depressed patients.

- According to the results, references generated from the Mastoid Bone showed higher accuracy compared with Cz references.

Thank You !

Questions?