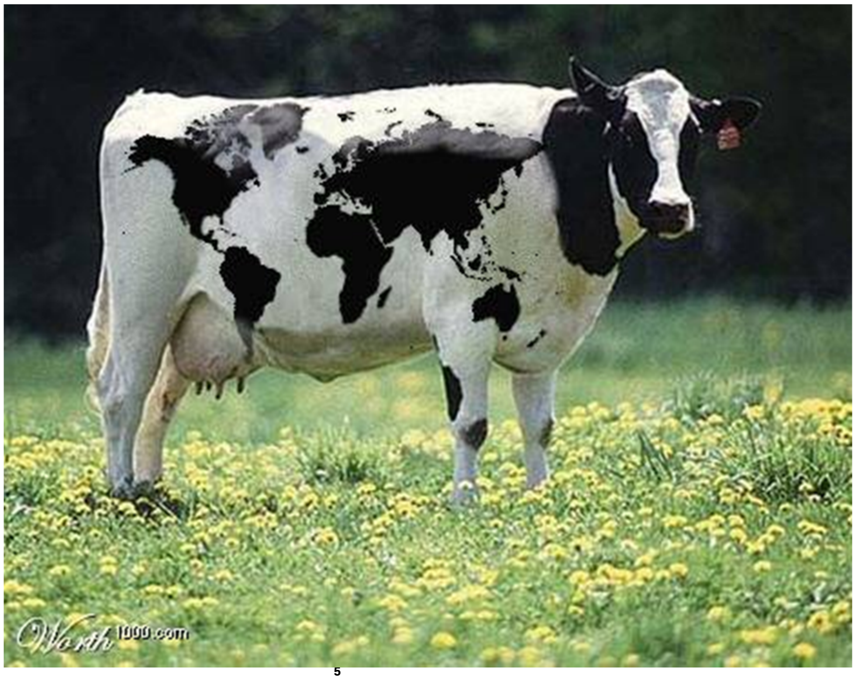
Lecture 7

1

2

3

4



5

6

Data Mining Knowledge Discovery: An Introduction and an Application

Acknowledgement

Gregory Piatetsky-Shapiro
KDnuggets
gregory@kdnuggets.com

8

Trends leading to Data Flood

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific Data: astronomy, biology, etc
  - Web, text, and e-commerce

9

Big Data Examples

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data

10

5 million terabytes created in 2002

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.
- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
- US produces ~40% of new stored data worldwide
- See

www.sims.berkeley.edu/research/projects/how-much-info-2003/

11

Largest databases in 2003

- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 3.3 Billion pages, ? TB
  - IBM WebFountain, 160 TB (2003)
  - Internet Archive (www.archive.org),~ 300 TB

12

Data Mining Application Areas

- Science
  - astronomy, bioinformatics, drug discovery, …
- Business
  - advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, …
- Web:
  - search engines, bots, …
- Government
  - law enforcement, profiling tax cheaters, anti-terror(?)

13

Assessing Credit Risk: Case Study

- Situation: Person applies for a loan
- Task: Should a bank approve the loan?
- Note: People who have the best credit don't need the loans, and people with worst credit are not likely to repay. Bank's best customers are in the middle

14

Credit Risk - Results

- Banks develop credit models using variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan
- Widely deployed in many countries

15

Successful e-commerce – Case Study

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought "**Advances in Knowledge Discovery and Data Mining**", also bought "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**"
- Recommendation program is quite successful

16

Genomic Microarrays – Case Study

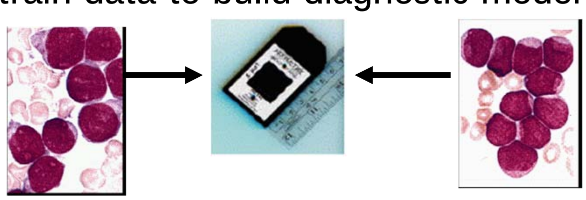
Given microarray data for a number of samples (patients), can we

- Accurately diagnose the disease?
- Predict outcome for given treatment?
- Recommend best treatment?

Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

ALL

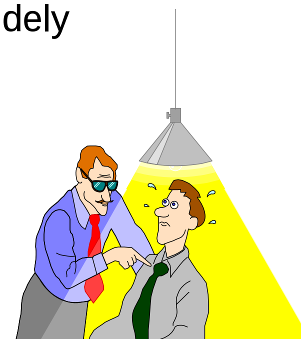


AML

Results on test data:
33/34 correct, 1 error may be mislabeled

Data Mining, Security and Fraud Detection

- Credit card fraud detection – widely done
- Detection of money laundering
  - FAIS (US Treasury)
- Securities fraud detection
  - NASDAQ KDD system
- Phone fraud detection
  - AT&T, Bell Atlantic, British Telecom/MCI
- "Total" Information Awareness – very controversial



19

Problems Suitable for Data-Mining

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

Privacy considerations important if personal data is involved

20

Big Data Example

21

Knowledge Discovery Definition

Knowledge Discovery in Data is the

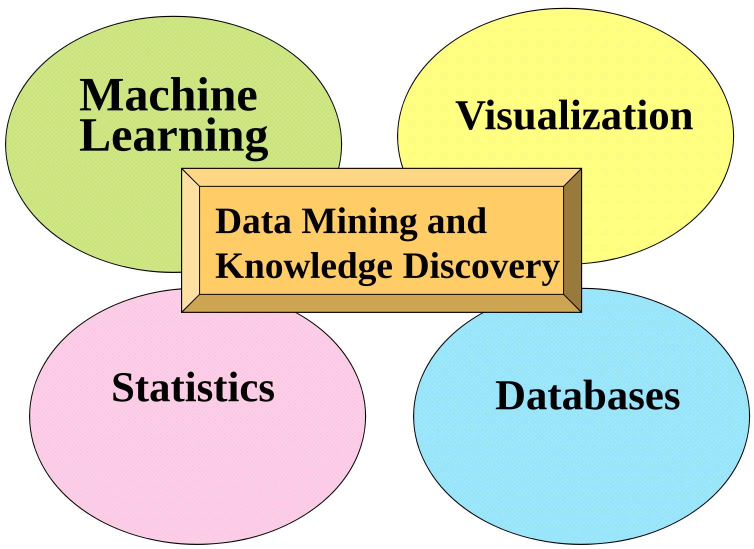
non-trivial process of identifying

- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining,* Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

22

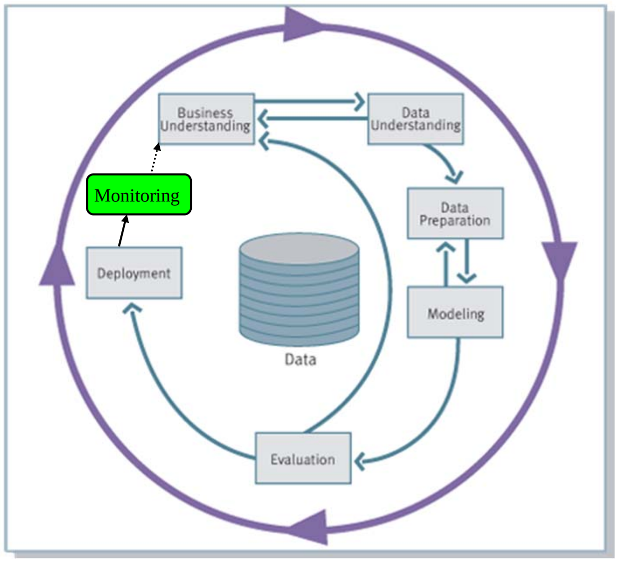
Related Fields



Statistics, Machine Learning and Data Mining

- Statistics:
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Knowledge Discovery Process flow, according to CRISP-DM



see www.crisp-dm.org for more information

25

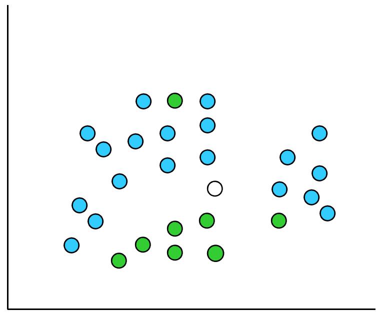
Major Data Mining Tasks

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection**: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- …

26

Data Mining Tasks: Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

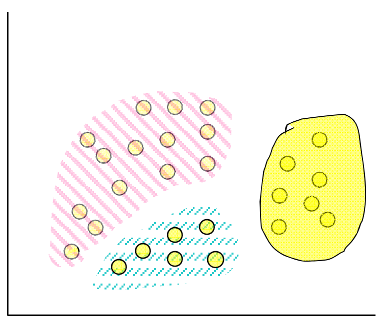


Many approaches:
Statistics,
Decision Trees,
Neural Networks,
...

27

Data Mining Tasks: Clustering

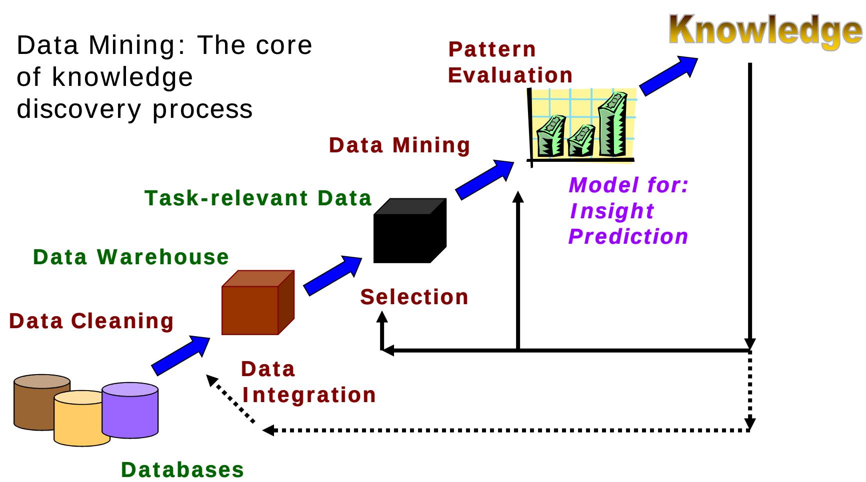
Find "natural" grouping of instances given un-labeled data



28

Knowledge Discovery Process

Data Mining: The core of knowledge discovery process



29

Summary:

- Technology trends lead to data flood
  - data mining is needed to make sense of data
- Data Mining has many applications, successful and not
- Knowledge Discovery Process
- Data Mining Tasks
  - classification, clustering, …

30

More on Data Mining and Knowledge Discovery

- KDnuggets
  - news, software, jobs, courses,...
  - www.KDnuggets.com
- ACM SIGKDD – data mining association
  - www.acm.org/sigkdd

31

Machine Learning: finding patterns

Finding patterns

- Goal: programs that detect patterns and regularities in the data
- Strong patterns ⇒ good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

33

Machine learning techniques

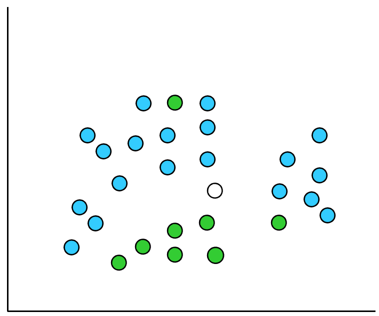
- *Algorithms for acquiring structural descriptions from examples*
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived
    (*may be even more important*)
- Methods originate from artificial intelligence, statistics, and research on databases

witten&eibe

34

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

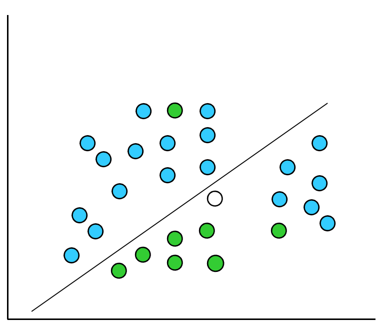


Many approaches:
Regression,
Decision Trees,
Bayesian,
Neural Networks,
...

Given a set of points from classes
what is the class of new point ○?

35

Classification: Linear Regression



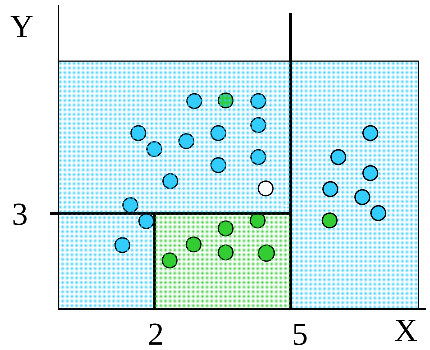
- Linear Regression

  $w_0 + w_1 x + w_2 y >= 0$

- Regression computes w_i from data to minimize squared error to 'fit' the data
- Not flexible enough

36

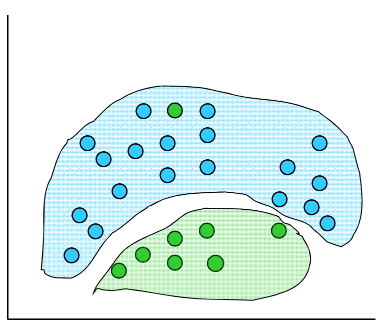
Classification: Decision Trees



if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

37

Classification: Neural Nets



- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

38

The weather problem

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no

Given past data,
Can you come up
with the rules for
Play/Not Play ?

What is the game?

39

The weather problem

- Conditions for playing golf

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

witten&eibe

40

Weather data with mixed attributes

- Some attributes have numeric values

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```


Predicting CPU performance

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Music Example

43

BP based Neural Networks : Data Mining

Issues Related to Data

- Nature of Data : *Source, Utility, Behaviour, Description*
- *Source:* Online/Offline, from Static/Dynamic Systems
- *Utility*: Analysis, Design, Diagnostics
- *Behaviour:* Discrete/Continuous
- *Description:* Quantitative/Qualitative

45

Issues Related to Data (Contd.)

- Are they sparse or dense?
- Are they in raw or clean form?
- Are they representative of the application domain?
- Are they noisy?
- Do they contain missing data?
- Scientific data :

  Insight (novelty detection, anomalies etc.)

  Predictive Model (Neural networks)

46

Problem Domain

Aim was to find any **novelty in the dataset** and to establish an **accurate mapping** between **propeller configuration parameters** and its **performance parameters**

Data Acquisition and Neural Networks Model for Data Mining

- USN series data of marine propeller design:

  Denny, S. B., Puckette, L. T., Hubble, E. N., Smith, S. K. and Najarian, R. F. (1989), A new usable propeller series, Marine Technology, 26, 3, 173-191.

- Neural Networks Model:
  - *BP Based Network*
- Prediction Evaluation:
  - *Resubstitution*
  - *Bootstrap,*
  - *Cross-Validation*
  - *Hold-out*

Data Variables

The experimental design data of **301 samples** cover parameters like

- ✓ **T**hrust coefficient K_T,
- ✓ Torque coefficient K_Q,
- ✓ Efficiency (η) versus
- ✓ Advance coefficient J for various values of pitch diameter ratio (P/D),
- ✓ Expanded area ratio (EAR),
- ✓ Number of blades (z) and
- ✓ Cavitation number (σ).

49

Data Accuracy

- **Propeller rps (n): ±1 rpm (±1/1000 = 0.1% full scale**; max rpm is not given, therefore this figure is an approximation assuming engine rpm to be 1500 and given 1.5:1 reduction gear-box).
- **Ship speed (*V*): ±0.1 knot (±0.1/20 = 0.5% full scale**; max ship velocity is not given therefore this figure is an approximation).
- **Thrust (*T*): ±0.25 lb (±0.5% full scale**; for lower T values, the relative error is larger).
- **Torque (*Q*): ±64.8 in.-lb (±0.2% full scale**; for lower *Q* value, the relative error is larger).

50

Data Extraction

- The experimental data are displayed in **many graphs** that include all test instances described by the dimensionless parameters. We used **data extracted** from the original graphical data and given to us by **Neocleous and Schizas (1995).**

- *Neocleous, C C. and Schizas, C. N. (1995), Artificial neural networks in marine propeller design, In Proceedings of ICNN'95 - International Conference on Neural Networks, IEEE Computer Society, NY, 2, 1098-1102*

51

Model Input and Output

- **Input Parameters**
  - ✓ Advance coefficient J
  - ✓ Pitch diameter ratio (P/D),
  - ✓ Expanded area ratio (EAR),
  - ✓ Number of blades (z) and
  - ✓ Cavitation number (σ).
- **Output Parameters**
  - ✓ Thrust coefficient K_T,
  - ✓ Torque coefficient K_Q,
  - ✓ Efficiency (η)

52

Model Selection and Parameters

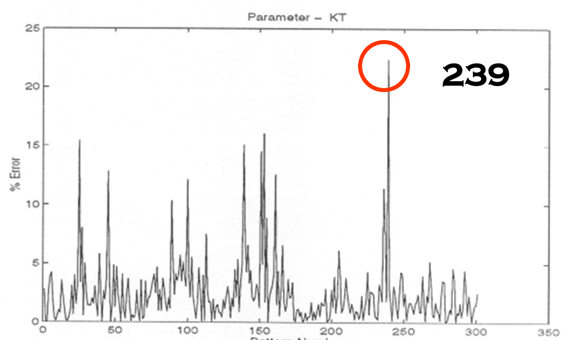
- Multilayer Perceptron-Improved backpropagation
- MATLAB 5.2 + Neural Networks Tool Box
- Architecture 5-30-30-3
- Learning Rate = 0.02
- Sigmoidal Activation Function
- SSE = 0.5
- No network or parameter optimization

53

Insight into Data- Anomaly Detection

K_T:

Data Entry Error



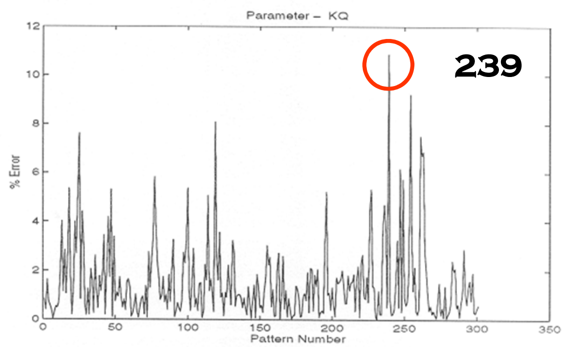
The actual value of the pattern =0.063 and the neural network predicted as 0.073

54

Insight into Data- Anomaly Detection

K_Q:

Data Entry Error

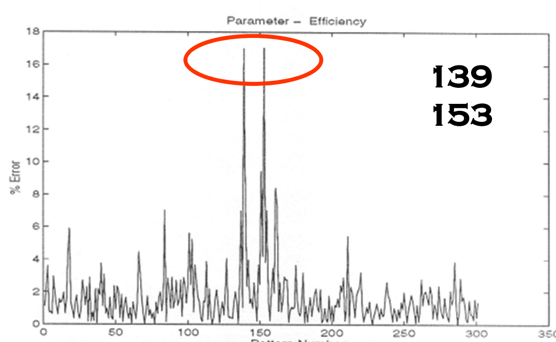


The actual value of the pattern =0.155 and the neural network predicted as 0.166

55

Insight into Data- Anomaly Detection

Efficiency (η)



139: Data Entry Error
153:Best Fit Curve

56

Prediction Accuracy

Exercise	K_T	K_Q	η
Resubstitution	**5.92**	**3.46**	**3.41**
.632 Bootstrap (l=92)			
Mean	**6.81**	**3.90**	**3.59**
Standard Deviation	**0.69**	**0.32**	**0.23**
Leave-one-out	**8.04**	**4.21**	**3.62**
K-Fold Analysis			
Mean	**7.91**	**4.57**	**4.18**
Standard Deviation	**0.36**	**0.28**	**0.12**
Hold-Out			
Mean	**9.79**	**5.79**	**4.60**
Standard Deviation	**1.46**	**0.70**	**0.26**

57

Summary

- Better ***insight*** about the data set and could trace down discrepancies in the data so that data entry errors could be corrected.
- As a universal approximator model, neural networks model has performed very well as a ***predictive model***
- ***Good data quality*** leads to build model in a single iteration
- Easy to build ***Decision Support System*** using neural networks model
- Requirement of simultaneously ***cleaning and learning*** model

58

Recent Development

59