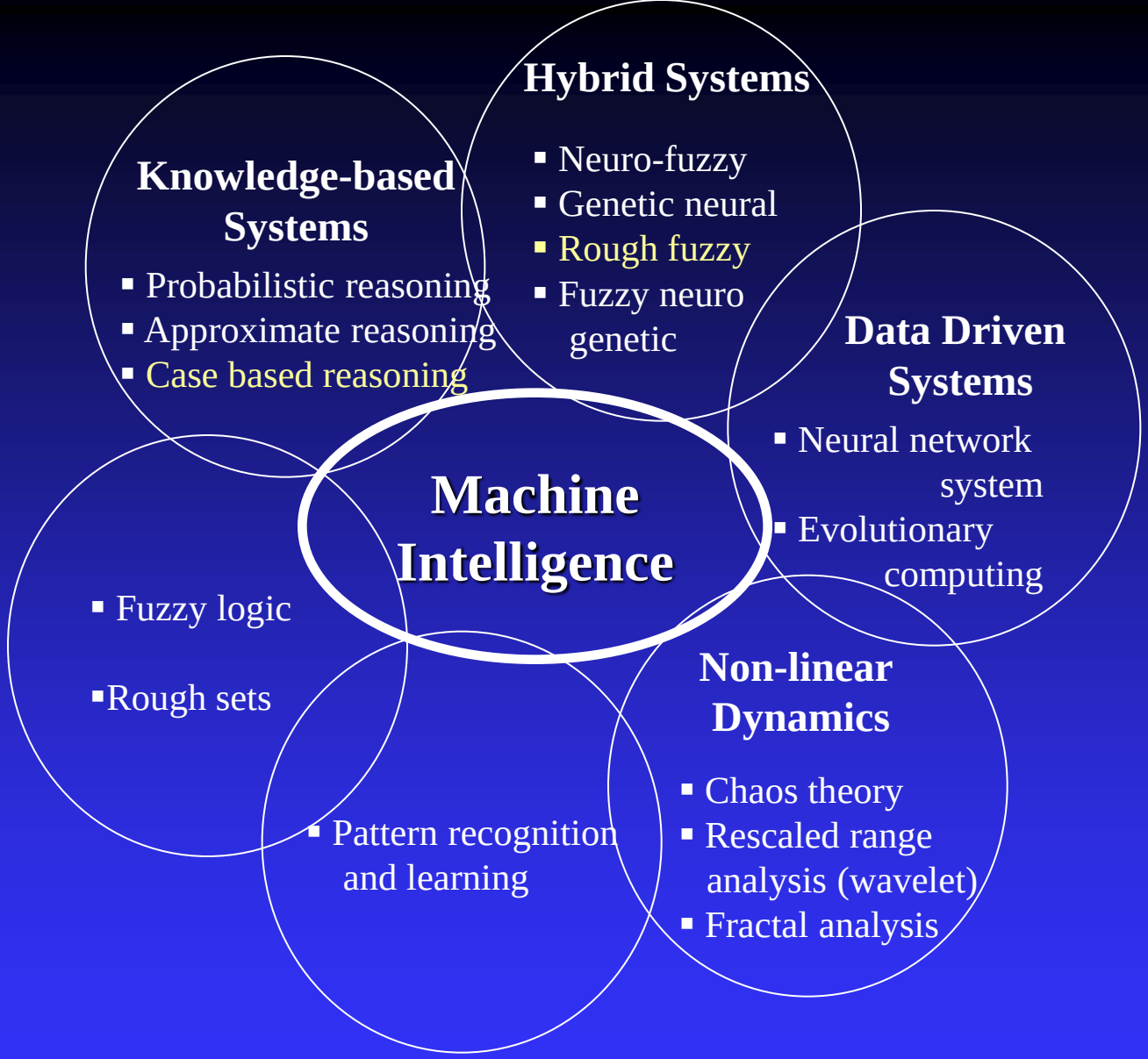
1st Mediterranean Conf. on Pattern Recogntion and Artificial Intelligence, Tebessa, Algeria, Nov 22, 2016

Granular Data Mining and Rough-fuzzy Computing: Data to Knowledge and Big Data Issues

Sankar K. Pal
Indian Statistical Institute
Calcutta
http://www.isical.ac.in/~sankar
&
Data Mining Services Ltd (DMS), U.K.

Contents

- Machine Intelligence & DM
- Fuzzy sets and uncertainty analysis
- Rough sets and information granules
  - Example: Case mining
  - Other applications
- Rough-fuzzy computing: Significance
- Generalized rough sets and entropy
  - Object extraction & video tracking
- Bioinformatics: Gene and miRNA selection
- Challenging issues
- Relevance to Big Data



Machine Intelligence: A core concept for grouping **various advanced technologies** with **Pattern Recognition and Learning**

IAS are physical embodiments of Machine Intelligence

Pattern Recognition and *Machine Learning* principles applied to a very large (both in size and dimension) heterogeneous database

≡ *Data Mining*

Data Mining + Knowledge Interpretation

≡ Knowledge Discovery

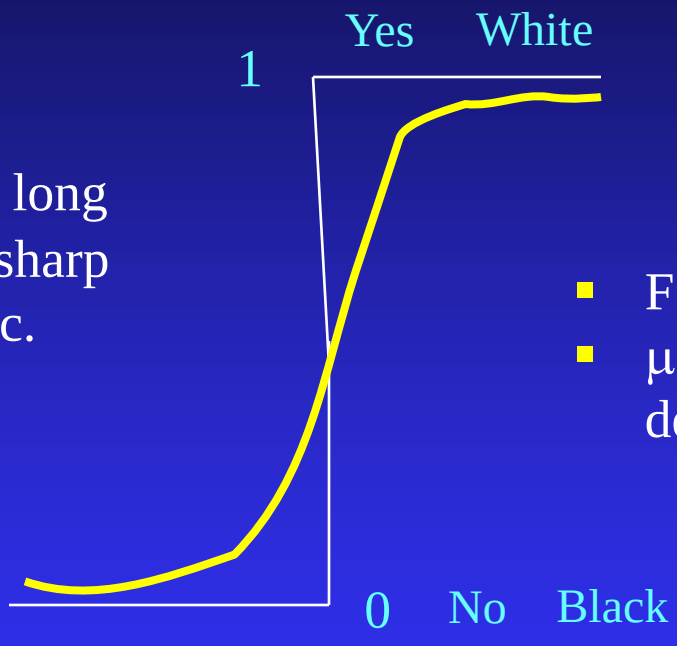
Process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

Fuzzy Sets:
Flexibility & Uncertainty Analysis

(Lotfi Zadeh, *Inform. Contro1*, 1965)

Classical Set $\mu \in \{0,1\}$ **Hard**

Fuzzy Set $\mu \in [0,1]$ **Soft**



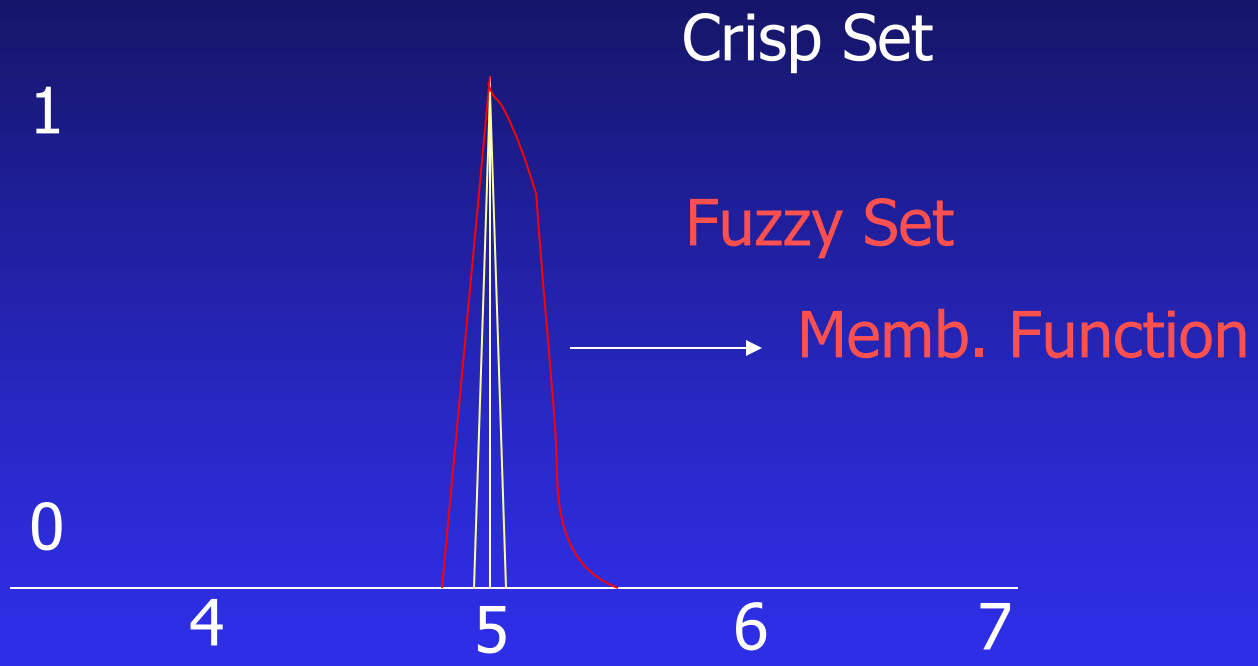
Examples : tall man, long street, large number, sharp corner, very young, etc.

- FS are nothing but μ-functions
- μ-functions are context dependent

$$A = \{(\mu_A(x), x) : \text{for all } x \in X\}$$

$\mu_A(x)$: Membership function : degree of belonging of x to A or degree of possessing some imprecise property represented by A

Meeting at 5 PM



- Fuzzy Sets are nothing but Membership Functions
- Membership Function: Context Dependent

Characteristics of FS

- FS is a *Generalization* of classical set theory

  ⇒ Greater flexibility in capturing faithfully various aspects of incompleteness or imperfection in a situation

- Flexibility is associated with the Concept of μ

  ⇒ As $\mu \uparrow$ Amount of Stretching the Concept $\downarrow$

  ⇒ FS are *Elastic*, Hard sets are inelastic

- $p(x)$: Concerns with the no. of occurrences of x
- $\mu(x)$: Concerns with the compatibility (similarity) of x with an imprecise concept

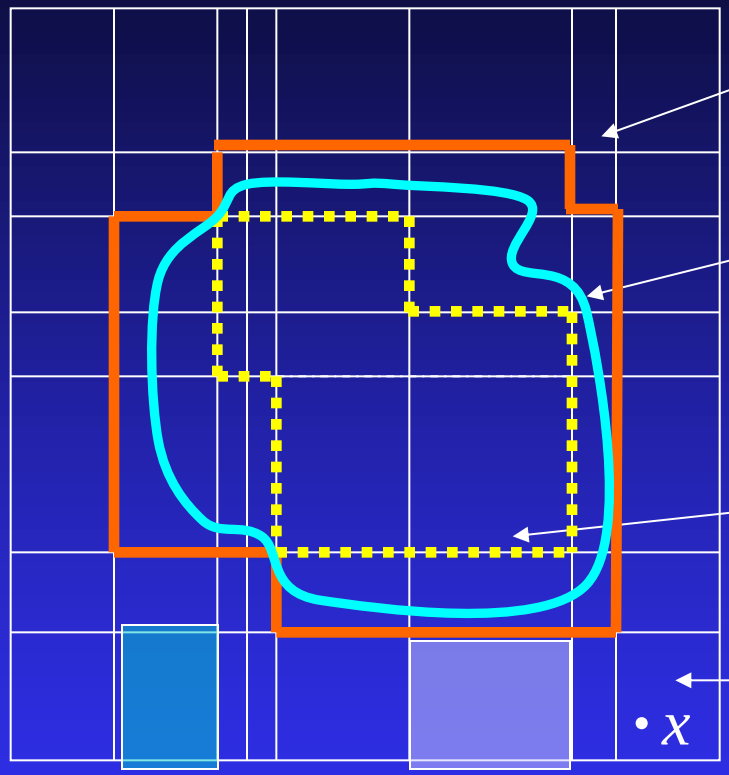
- Concept of Flexibility & Uncertainty Analysis (overlapping data/ concept/ regions)

Rough Sets and Granular Computing

Rough Sets

Z. Pawlak 1982, Int. J. Comp. Inf. Sci.

$\Omega_B \subseteq U$



$[x]_B$ = set of all points belonging to the same granule as of the point x in feature space Ω_B.

$\Rightarrow$ $[x]_B$ is the set of all points which are *indiscernible* with point x in terms of feature subset B.

Approximations of the set $X \subseteq U$ w.r.t feature subset B

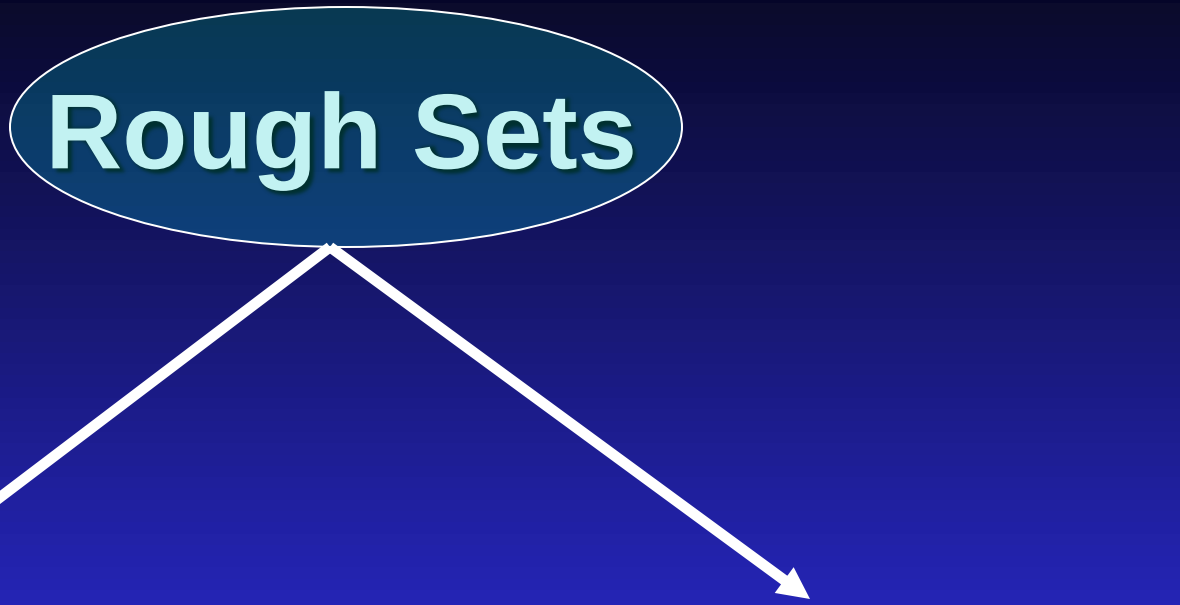
B-lower: $\underline{B}X = \{x \in U : [x]_B \subseteq X\}$ Granules definitely belonging to *X*

B-upper: $\overline{B}X = \{x \in U : [x]_B \cap X \neq \phi\}$ Granules definitely and possibly belonging to *X*

If $\underline{B}X = \overline{B}X$, *X* is *B-exact* or *B-definable*

Otherwise it is *Roughly definable*

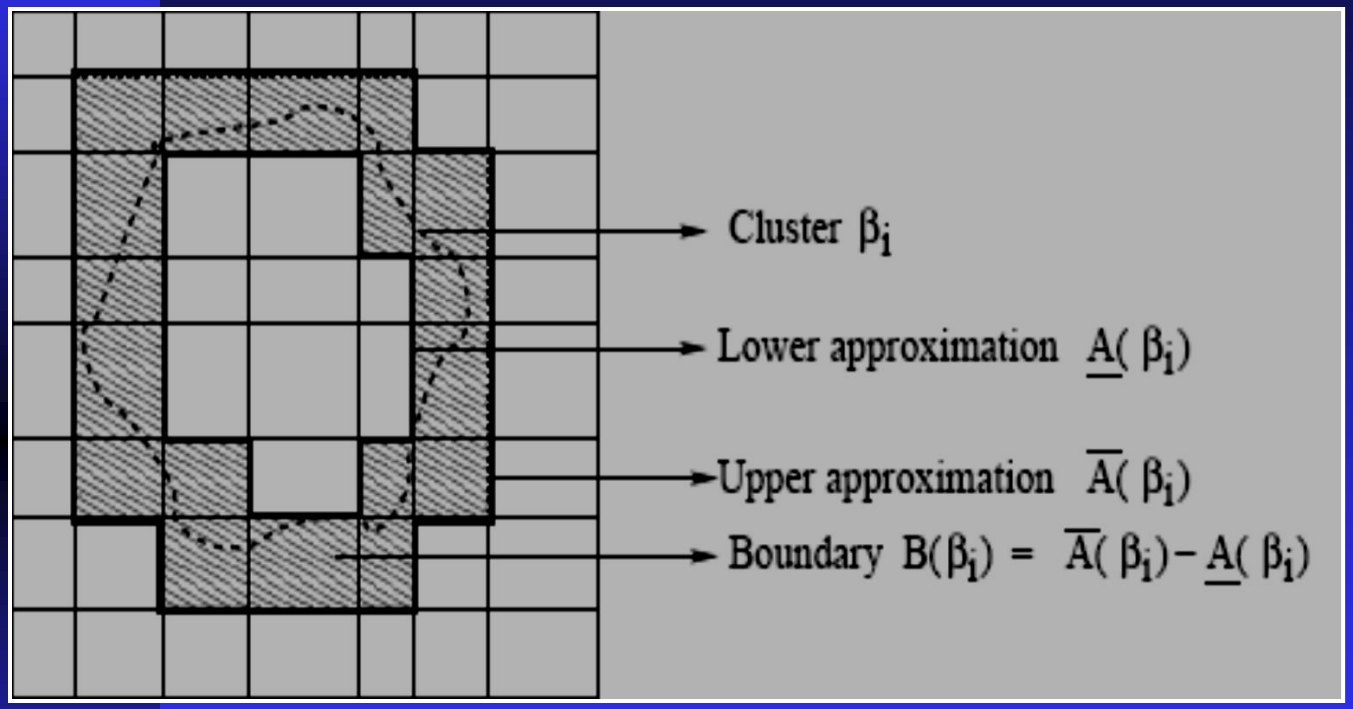
➡ Rough Sets are Crisp Sets, but with rough description



Two Important Characteristics

IEEE Trans. Syst., Man and Cyberns. Part B, 37(6), 1529-1540, 2007

Cluster definition using rough lower & upper approx



Roughness in β :
1 - |lower|/ |upper|

- Sets and Granules can *either or both* be fuzzy (in real life)
  - Lower and upper approximate regions could be fuzzy

➡ Generalized Rough Sets – Stronger model of uncertainty handling

(uncertainty due to overlapping regions + granularity in domain)

- Before I describe the Generalized Rough Sets and example applications of Granular Mining, let me explain the
  - Concept of granules
  - *f*-information granules & Case mining
  - Significance of rough granules
  - Relevance of Rough-Fuzzy computing in SC paradigm

Granulation: Natural clustering

- Replacing a fine-grained universe by a coarse-grained one, more in line with human perception

- L.A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic", *Fuzzy Sets and Systems*, 90, 111–127, 1997

- Rough set theory concerns with a granulated domain
  (crisp set defined over a crisp granulated domain)
  Z. Pawlak, Rough sets, *Int. J. Comp. Inf. Sci*, 11(5), 341-356, 1982

Concept of Granules

- Clump of indiscernible objects/points (similar objects, not discriminable by given attributes/relation)

- *Examples:* Granules in
  - Age: *very young, young, not so old,…*
  - Direction: *slightly left, sharp right,…*
  - School: each *class/section*
  - Image: *regions* of similar colors, gray values

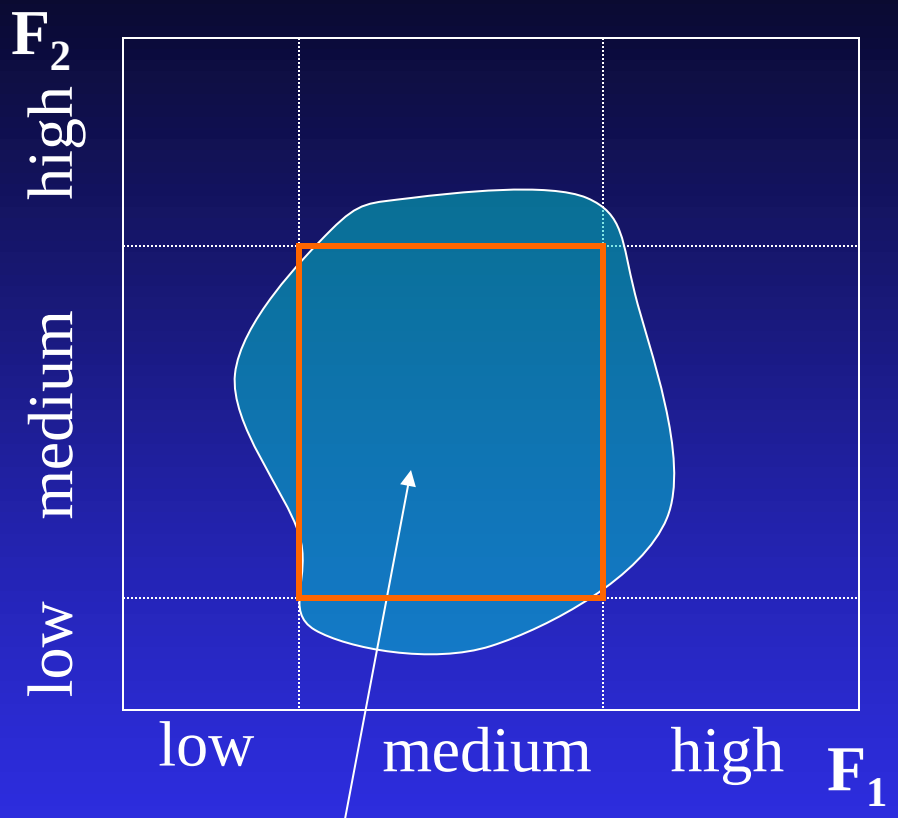
    e.g., max diff of 6 gray levels (Weber's law)

- Granules may be Crisp or Fuzzy (overlapping) **$**

Concept of -

f- Information Granules using Rough Rules

Information Granules and Rough Set Theoretic Rules



Rule $\leftarrow M_1 \wedge M_2$

- Rule provides crude description of the class using granule

Note:

- Rule characterizing the granule □ can be viewed as the Case or Prototype representing the class/ concept/ region

- Elongated objects need multiple rules/ granules
- *Unsupervised:* No. of granules is determined automatically

- Cases (prototypes) are granules, not sample points ➡ case generation, NOT selection

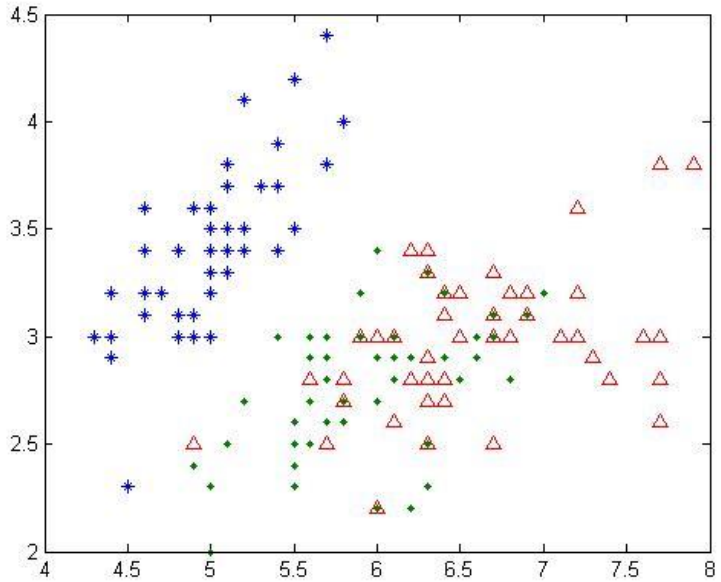
Note:

- All the features may not appear in rules
  - ➡ Dimensionality reduction
- Depending on topology, granules of different classes may have different dimensions ➡ Variable dimension reduction
  - Less storage requirement
  - Fast retrieval
- Suitable for mining data with large dimension and size

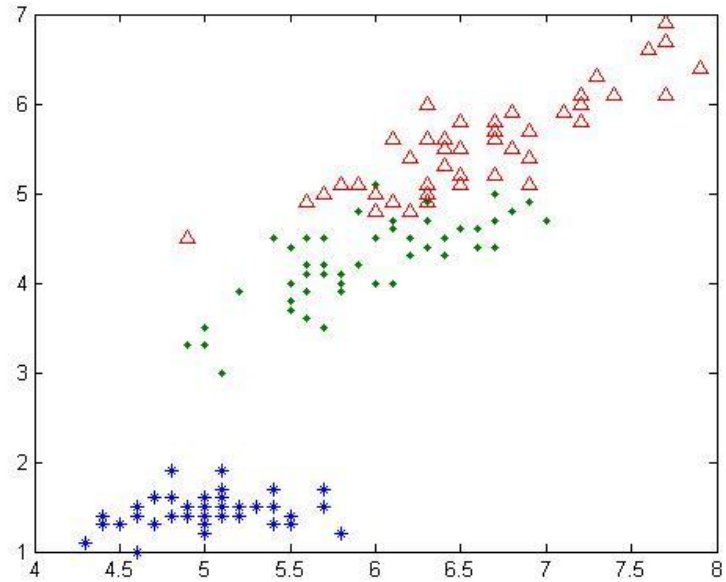
IEEE Trans. Knowledge Data Engg., 16(3), 292, 2004

Example: IRIS data case generation

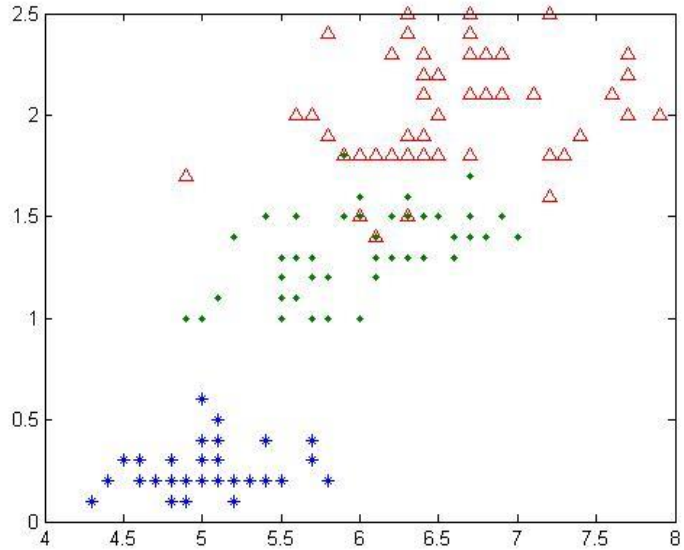
Three flowers: Setosa, Versicolor and Virginica
No of samples: 50 from each class
Features: sepal length, sepal width, petal length, petal width



(a)



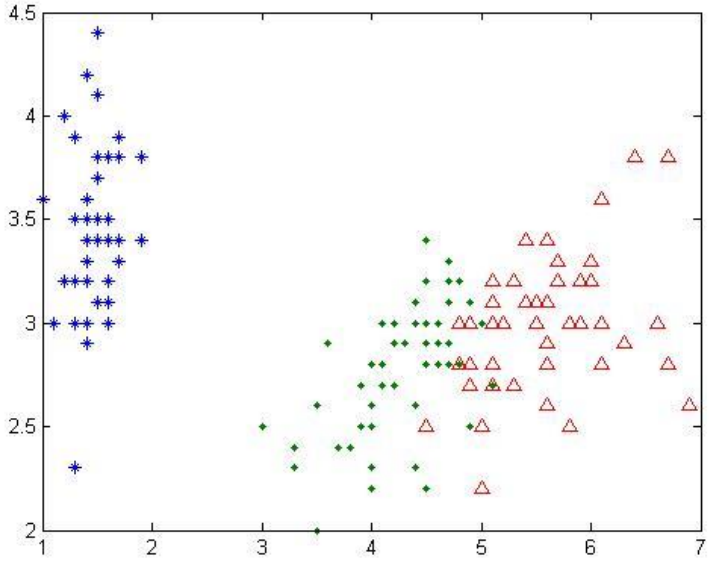
(b)



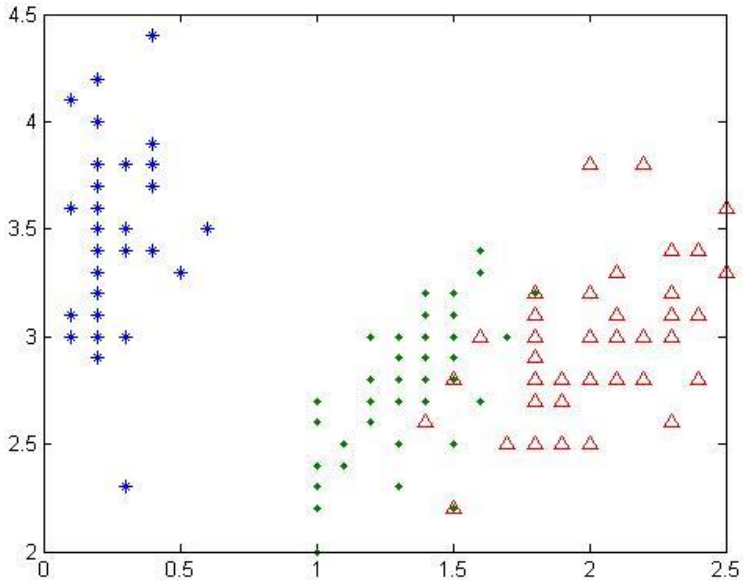
(c)

Iris Folowers: Setosa, Versicolor and Virginica

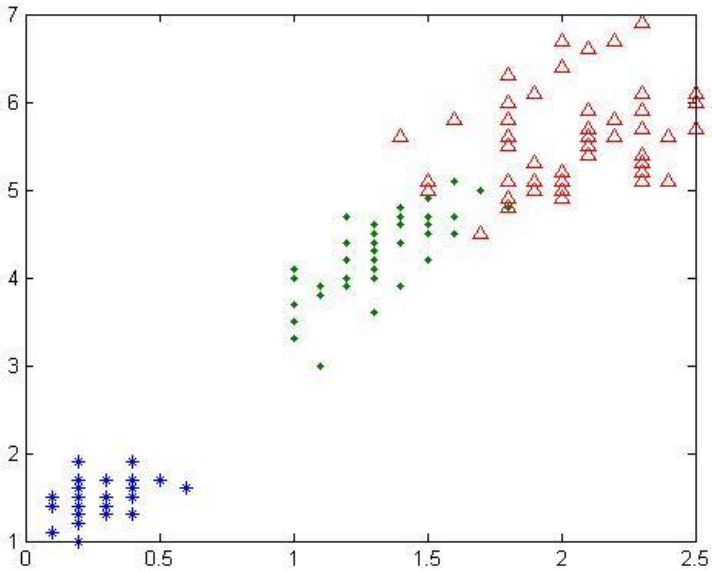
(a) Sepal L- Sep W
(b) Sepal L – Petal L
(c) Sepal L – Petal W



(a)



(b)

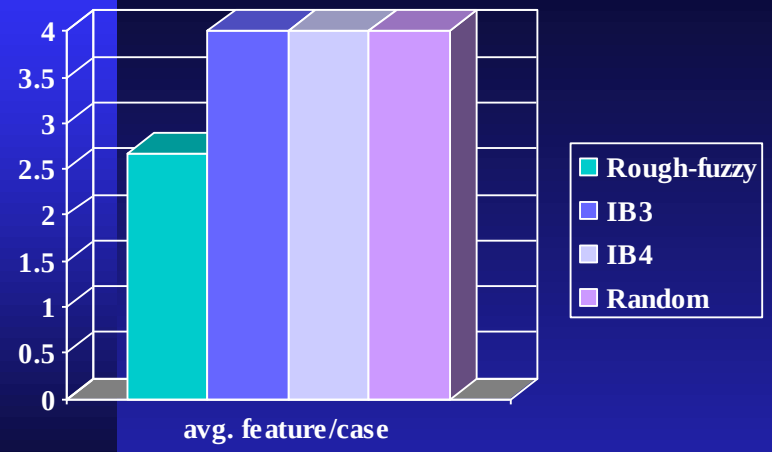


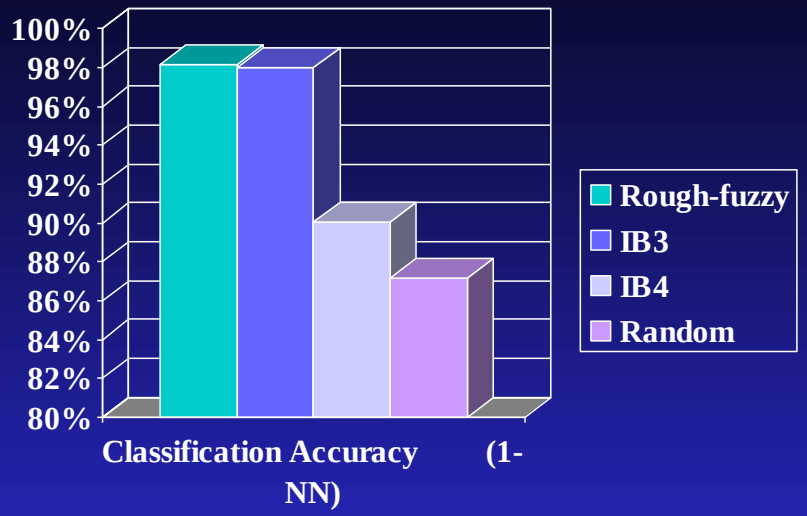
(c)

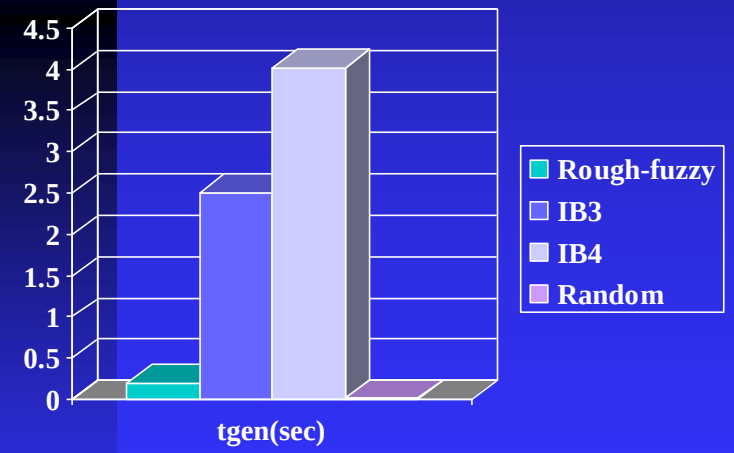
Iris Folowers: Setosa, Versicolor & Virginica

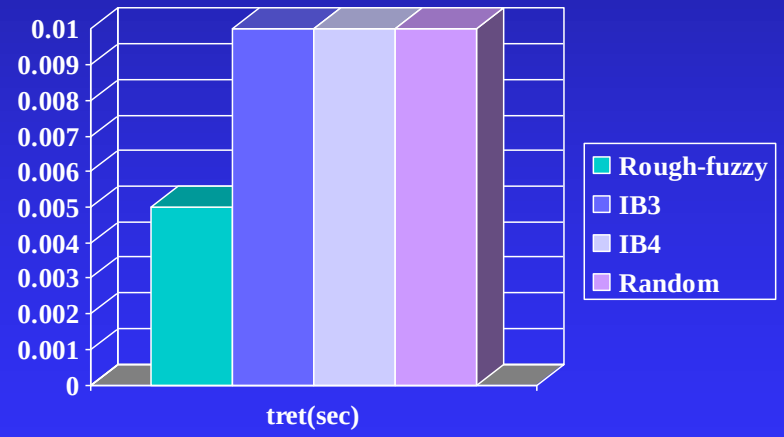
(a) Petal L - Sepal W
(b) Petal W - Sepal W
(c) Petal W - Petal L

Iris Flowers: 4 features, 3 classes, 150 samples









Number of cases = 3 (for all methods)

Granular Computing (GrC): Computation is performed using *information granules* and not the data points (objects)

- Information compression
- Computational gain

➢ Suitable for Mining Large Data

- Rough set theory enriched GrC research

Applications of Rough Granules

- Case Based Reasoning (evident is sparse)
- Prototype generation and class representation
- Clustering & Image segmentation (k selected autom)
- Case representation and indexing
- Knowledge encoding
- Dimensionality reduction
- Data compression and storing
- Granular information retrieval

Certain Issues

- Selection of granules and sizes
- Fuzzy granules
- Granular *fuzzy computing*
- *Fuzzy granular* computing

Concept of -

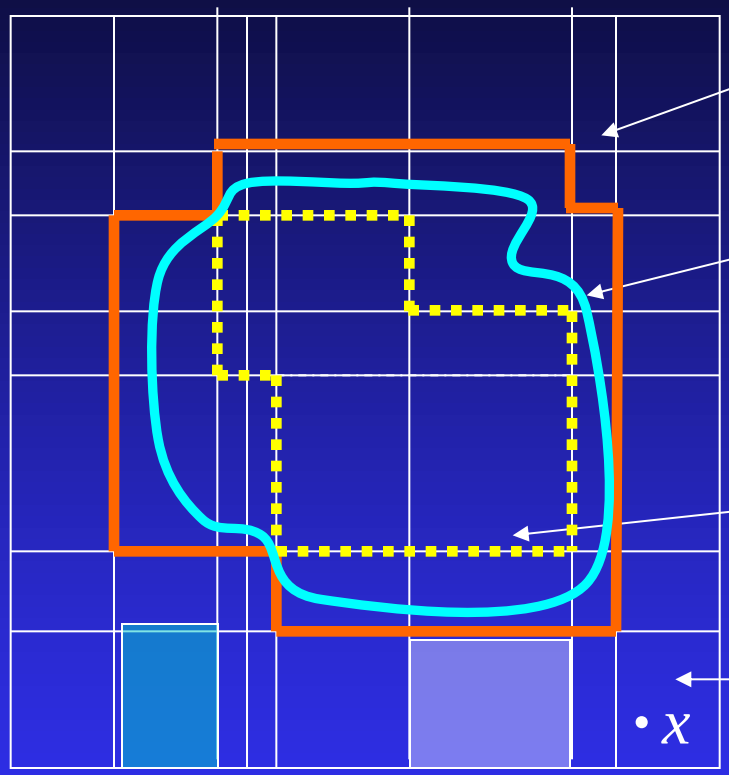
Generalized Rough Sets

Incorporate fuzziness in set & granules of rough sets

IEEE Trans. Syst, Man and Cyberns. Part B, 39(1), 117-128, 2009

Generalized Rough Sets

$\Omega_B \subseteq U$



- In practice, the Set and Granules, either or both, could be Fuzzy.

⇨ Generalized Rough Set

⇨ Stronger Paradigm for Uncertainty Handling

IEEE Trans. Syst, Man and Cyberns. Part B, 39(1), 117-128, 2009

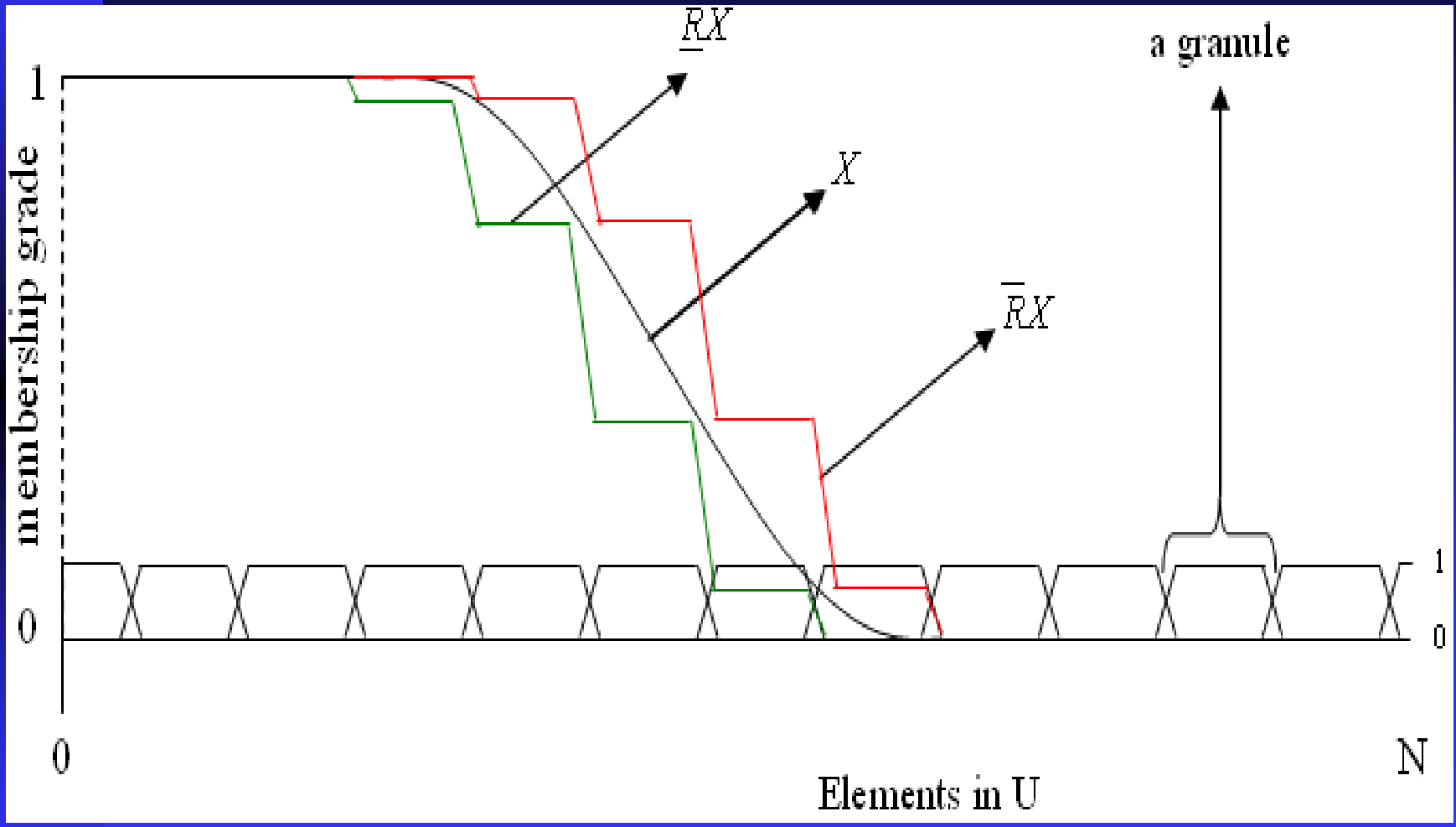
Incorporate Fuzziness in Set & Granules

- *X* is a **crisp** set & Granules have **crisp** boundaries - rough set of *X*
- *X* is a **fuzzy** set & Granules have **crisp** boundaries - rough-fuzzy set of *X*
- *X* is a **crisp** set & Granules have **fuzzy** boundaries – fuzzy-rough set of *X*
- *X* is a **fuzzy** set & Granules have **fuzzy** boundaries – fuzzy rough-fuzzy set of *X*

Generalized Rough Sets

R is an equivalence relation

X is a ***fuzzy*** set & Granules have ***fuzzy*** boundaries



The pair $< \underline{R}X, \overline{R}X >$ is referred to as the fuzzy rough-fuzzy set of *X*.

Roughness Measure

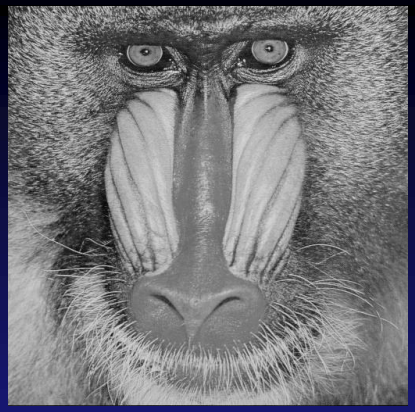
- $$\rho_R(X) = 1 - \frac{|\underline{RX}|}{|\overline{RX}|}$$

  ➡ a measure of inexactness of X

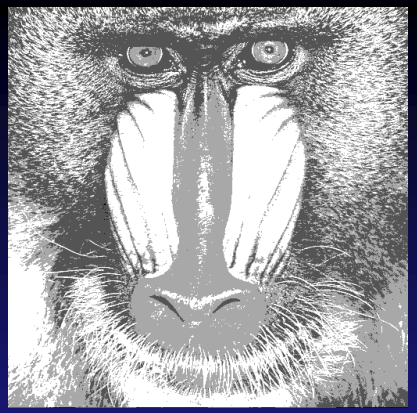
- $\underline{RX}$ and $\overline{RX}$ are the lower and upper approxs. of X

Segmentation Results

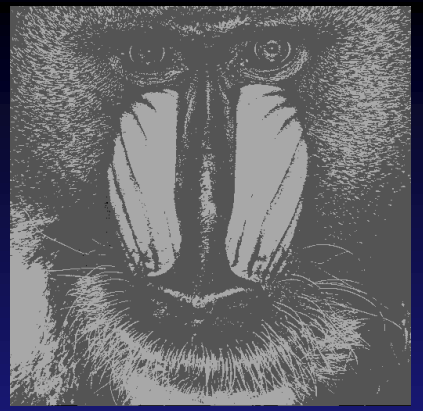
- *Effect of granules* **(ω = 6)**



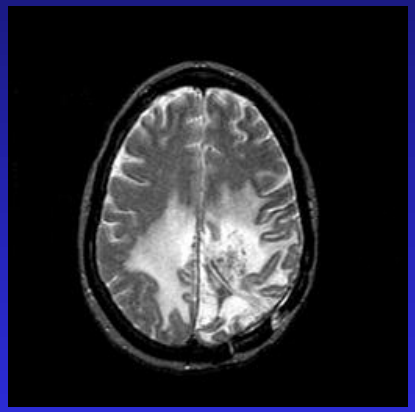
Baboon image



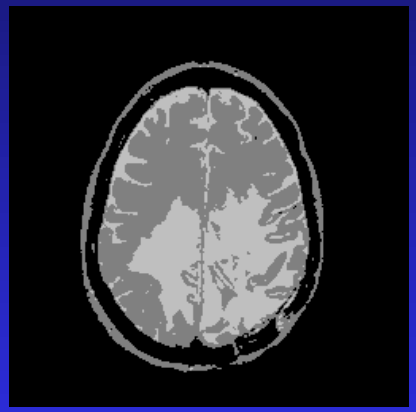
Proposed r-f entropy



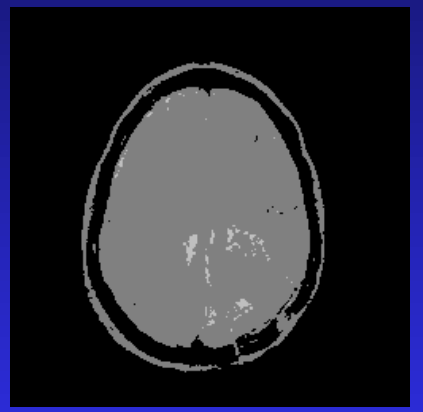
Fuzzy entropy



Brain MR image



Proposed r-f entropy



Fuzzy entropy

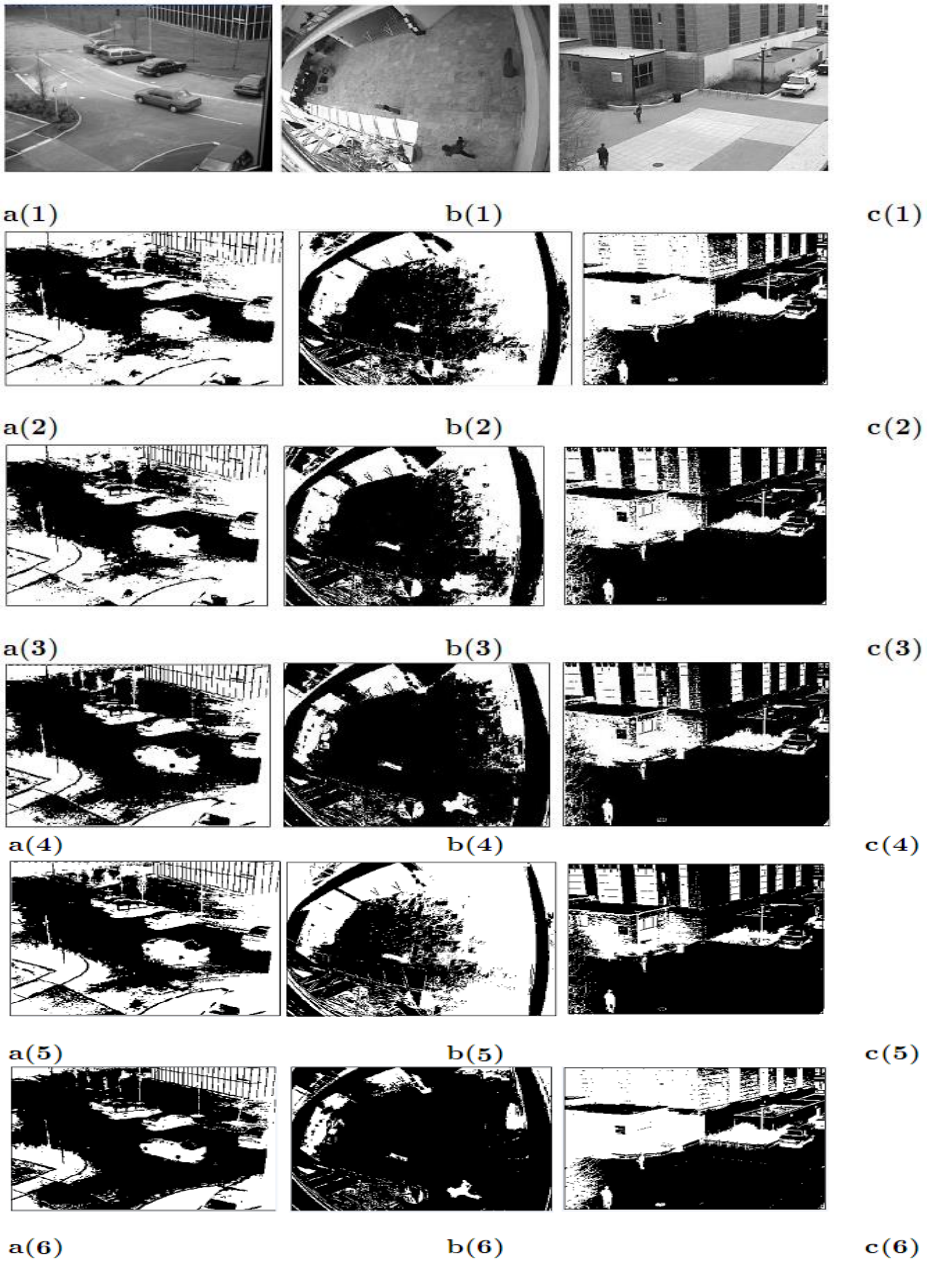
Remote sensing image

Proposed r-f entropy

Fuzzy entropy

- So far we considered granules of equal size
- Next, consider granules of **unequal** size

Example Comparison



Original

Otsu's thresholding

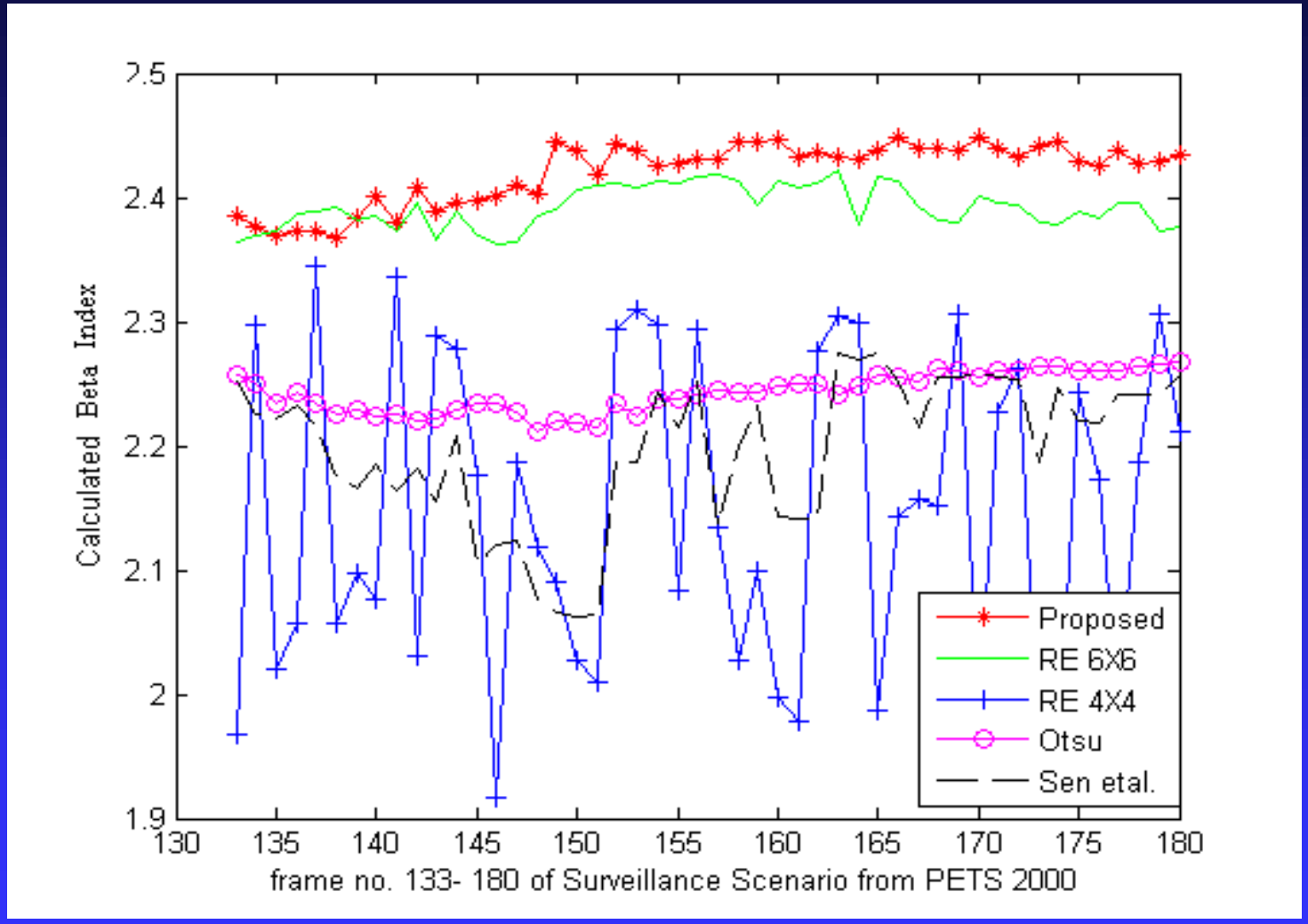
RE with 4x4 granule

RE with 6x6 granule

Rough-fuzzy with crisp set and 6x6 granule

Proposed methodology

Variation of *β-Index* over sequence '**a**'



- Homogeneous granules of unequal size reduce the formation of spurious segments → Reduce abrupt change of index-value over frames.

Video Tracking

- Spatial segmentation on each frame

+

- Temporal segmentation based on 3 previous frames

- Granules of Unequal Size
  - *quad-tree decomposition* (spatial)
  - merits over fixed size for Video Tracking

Other RE (6x6) - SP

Proposed

Otsu

RE (6x6) - PUM

RE (4x4) - PUM

IEEE Trans. Cybernetics (to appear)

❖ Use Neighborhood Granules in RGB-D:

3-d spatio-temporal, 2-d spatio-color, 1-d color

❖ Granules of Arbitrary Shape (natural)

- Use granules, instead of pixels, for O/B partition
- Granules obtained from *Lower Approx. Object Model Granular level* rule based decision
- Automatic updation of rule base with flow graph

Deals with ambiguous tracking situation (unsupervised)

(overlapping, newly appeared object, merged with similar color)

Extraction of Temporal Information (in D-space)

Current Frame f_t

f_{t-1} f_{t-2} f_{t-3}

(abs

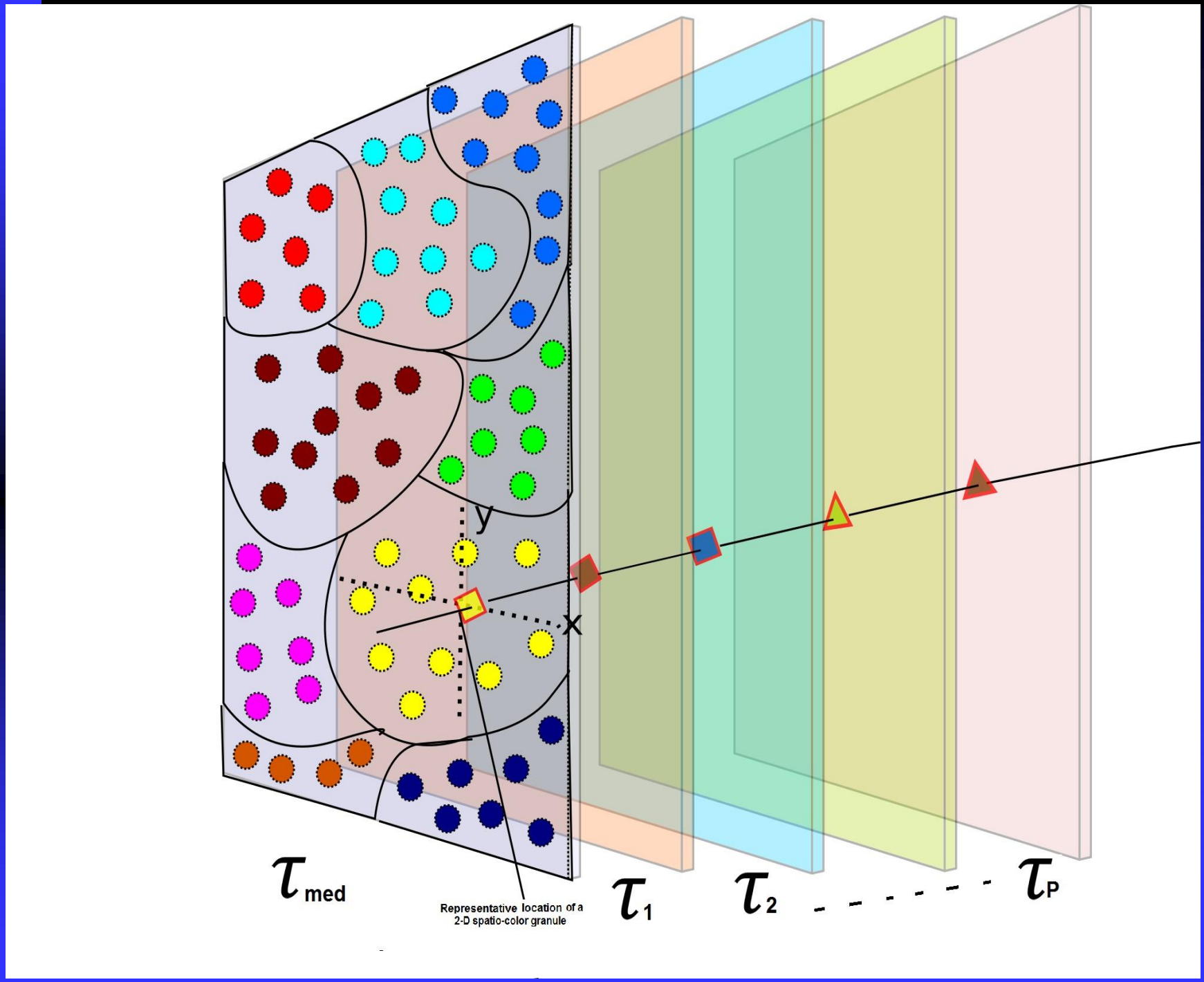
τ

UNION

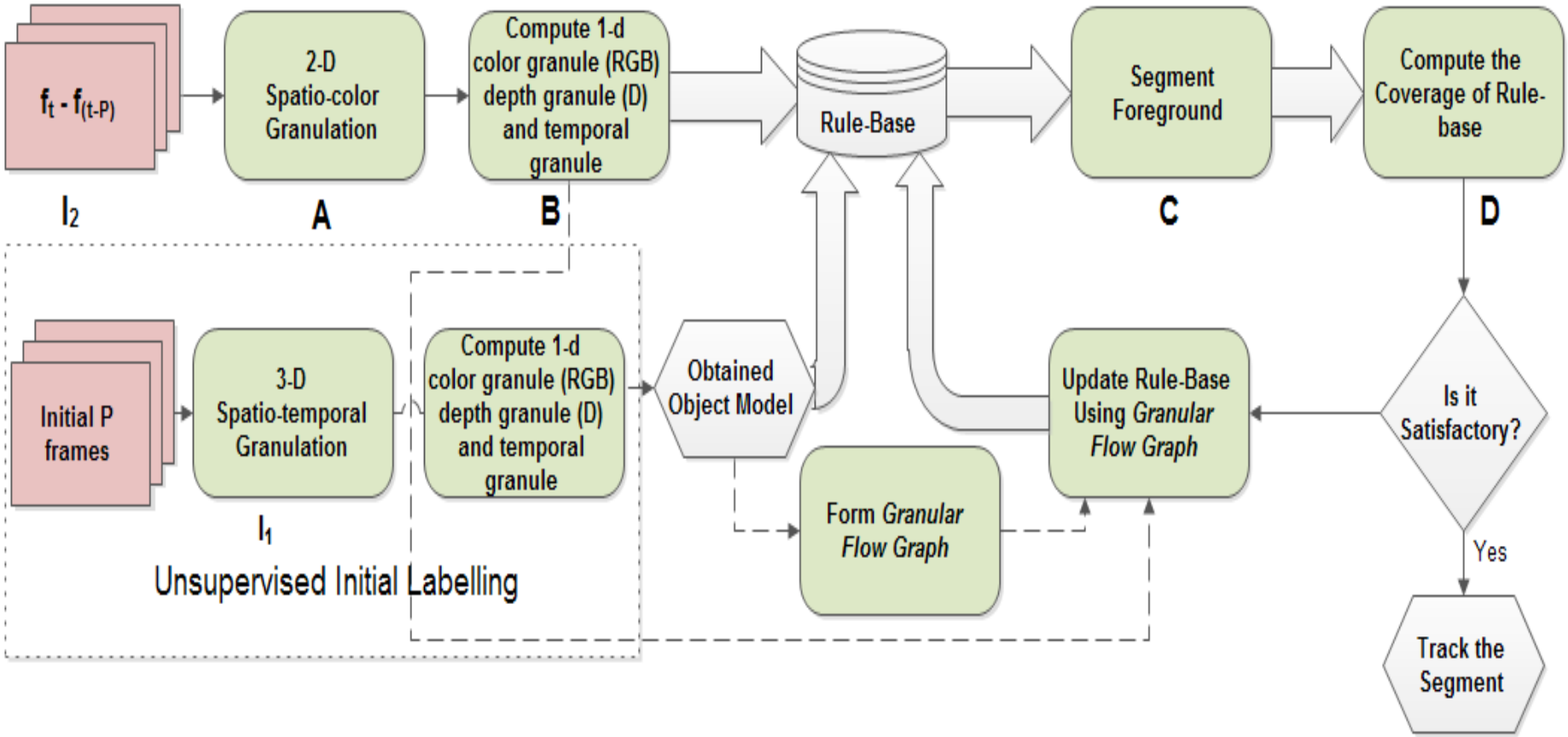
INTERSECTION

Intersection & $|f_t - f_P|$: Lower & Upper approx. of object in temporal domain

- $\cap\{\tau_p \forall p \in P\}$ denotes the common moving regions of τ_p over P frames ➡ estimation of Lower approximation of the moving object(s) (O_{low}) in temporal domain ➡ Object model in D-space
- $\cup\{ O_{low} , \tau_P\} = \cup\{ O_{low} , | f_t - f_P|\}$ (basically τ_P) denotes estimation of Upper approximation of the moving object(s) (O_{up}) in temporal domain
- Values of the features (RGB-D, Temporal) contained in the set O_{low} are the core values of the object model
- Values of the features in the set $\{O_{up} - O_{low}\}$, i.e., the boundary region, determine the extent to which the values in the object model are allowed


y
x
τ_{med}
Representative location of a 2-D spatio-color granule
τ_1
τ_2
τ_P

Methodology



Without Flow Graph Based Adaptation

Newly appeared object: **Fails**

(Frames per sec = 15, P = 6)

With Flow Graph Based Adaptation

Newly appeared object: **Succeeds**

IEEE Trans, Syst., Man and Cyberns, Part B, 40(3), 741-752, 2010
IEEE Trans. Knowledge & Data Engineering, 22(6), 854-867, 2010

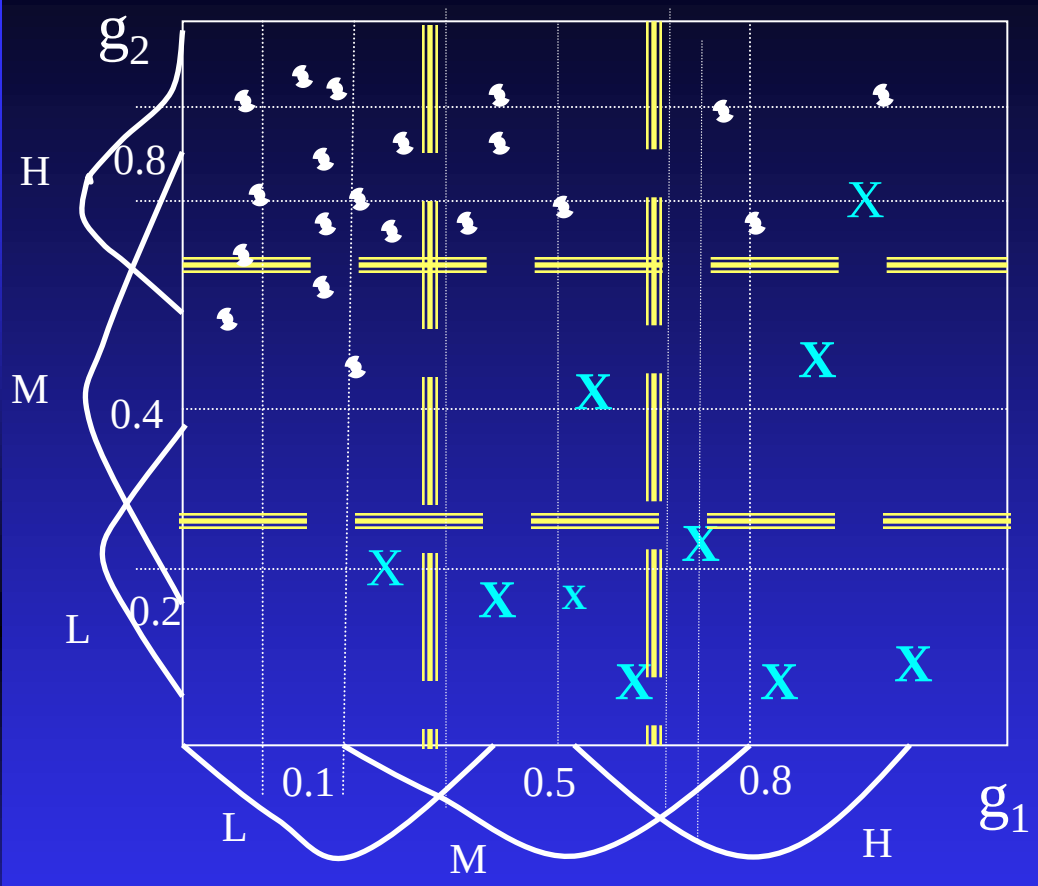
Example:

f-Information Measure and Gene Selection from Microarray Data

- Small sample, large dimension
- Granules (CI) model *Low, Medium & High* for overlapping classes ➡ Fuzzy Approx. Space (FEPM)

- Selection principle → *Maximization* of relevance of a gene w.r.t. decision attribute and *Minimization* of redundancy w.r.t. other genes
- Merits: FEPM based *density approximation approach vs.* those of Discretization and Parzen-window for computing Entropy and Mutual, V & Chi-square information

Example:



FEPM $M_{g_1g_2}$

- Attribute sets P & Q = g_1 & g_2
- Joint freq of segment (P_i,Q_j)

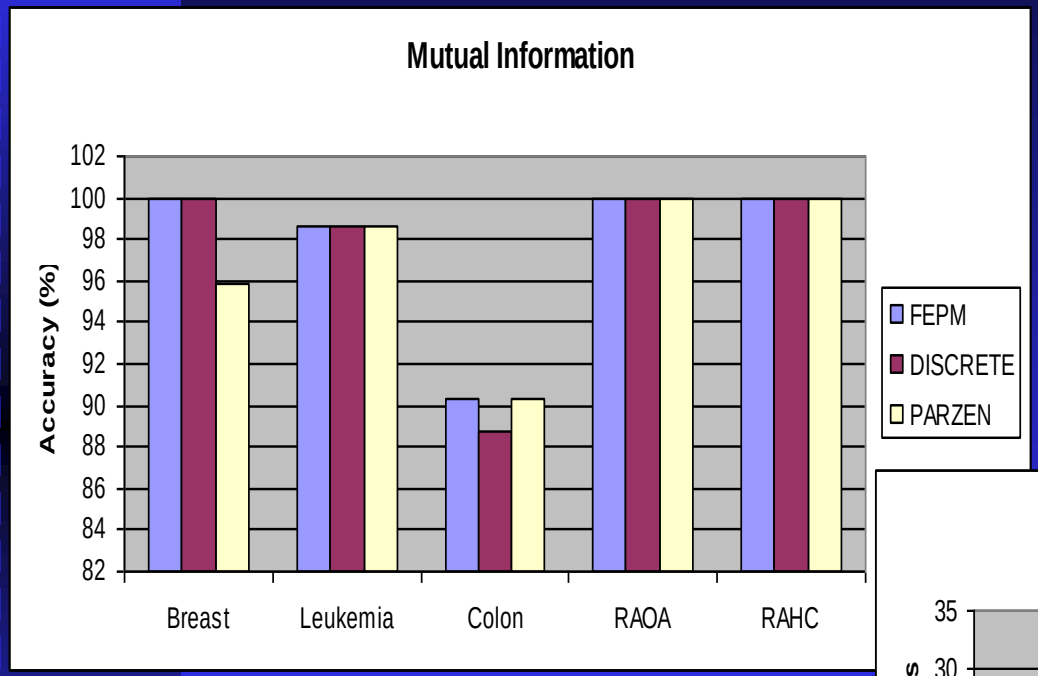
$$\lambda_{P_iQ_j} = \frac{1}{n}\sum_{k=1}^{n}(m_{ik}^{\mathbb{P}} \cap m_{jk}^{\mathbb{Q}}).$$

Each sample, in g_1-g_2 granulated space, is characterized by a 9-dimensional membership vector, each component corresponds to one of the nine fuzzy partitions *L-L, L-M, L-H, .., M,H, .., H-H.* Membership in *(L,H)* segment -

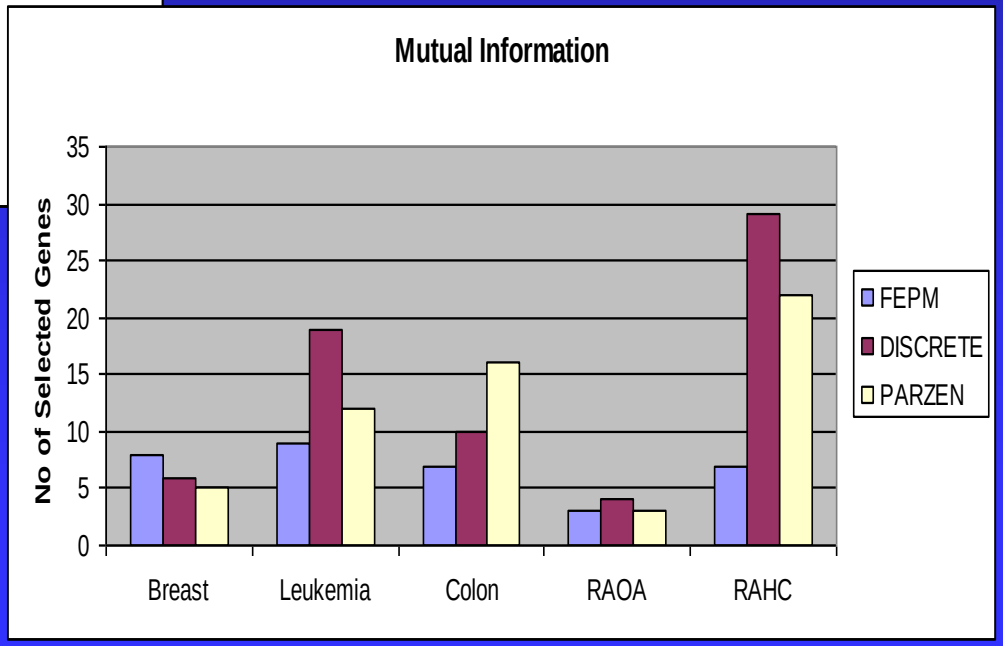
$$m^{g_1g_2}{}_{LH} = m^{g_1}{}_{L} \cap m^{g_2}{}_{H}$$

IEEE Trans. Syst., Man and Cyberns., Part B, 40(3), pp. 741-752, 2010

Gene Selection from Microarray Data



Genes/Samples
7129/49
7070/72
2000/62
18000/51
26000/50



- Maximization of relevance and minimization of redundancy
- Maximum 50 genes selected
- Highest accuracy obtained and the # genes required to obtain that are plotted
- SVM based classification accuracy (leave-one-out method)

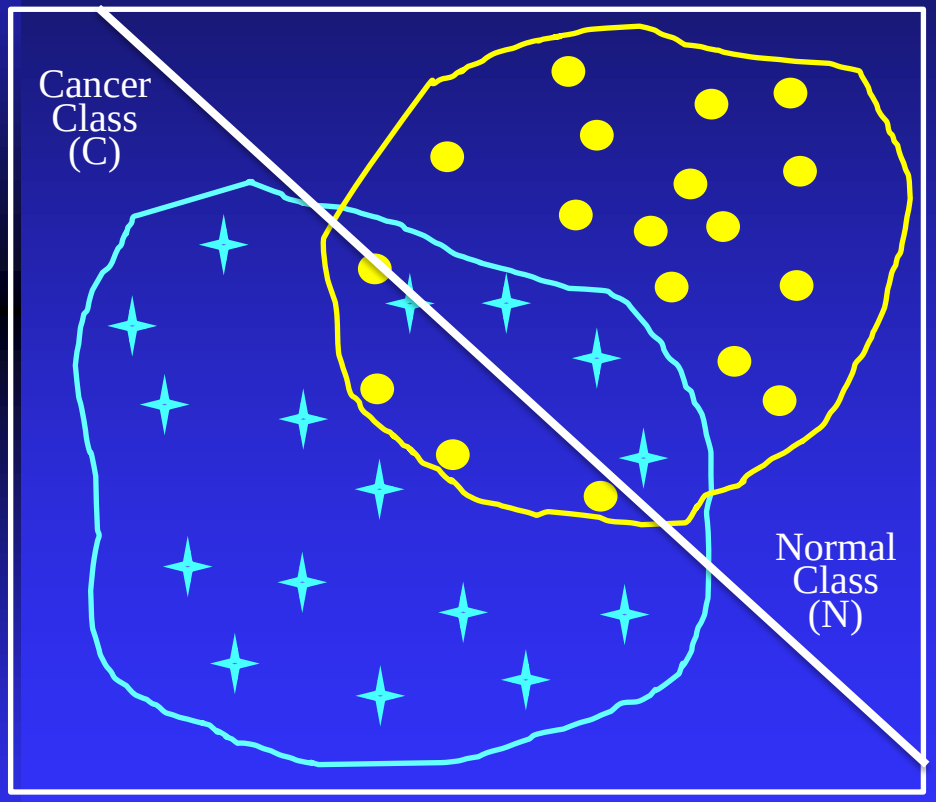
➢ *Higher or same accuracy with lower number of genes, except breast, in case of FEPM*

Example:

miRNA Ranking in Cancer: Fuzzy-Rough Entropy & Histogram Based Patient Selection

- Set is crisp & Granules are fuzzy: *Fuzzy-rough entropy*
- *Fuzzy Lower Approx.* of N & C classes: used to find prob. (relative frequency) of definite and doubtful regions for entropy computation
- *Entropy* minimization implies higher *Relevance* of a miRNA
- *Top 1% miRNAs provides significant improvement over entire set in terms of F-score*
- *Superiority over related methods*

Crisp Classes and Fuzzy Granules of Patients



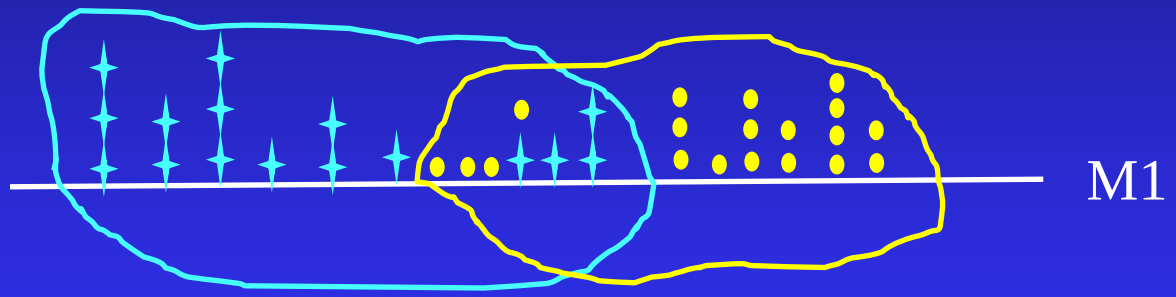
Normal patient

Normal granule

Cancer patient

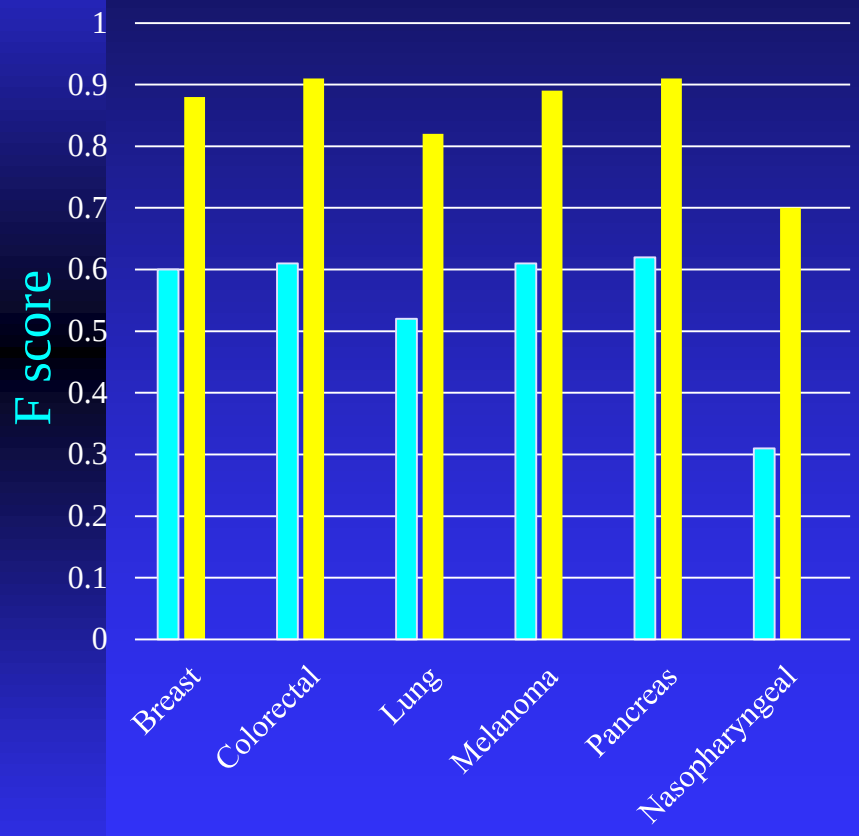
Cancer granule

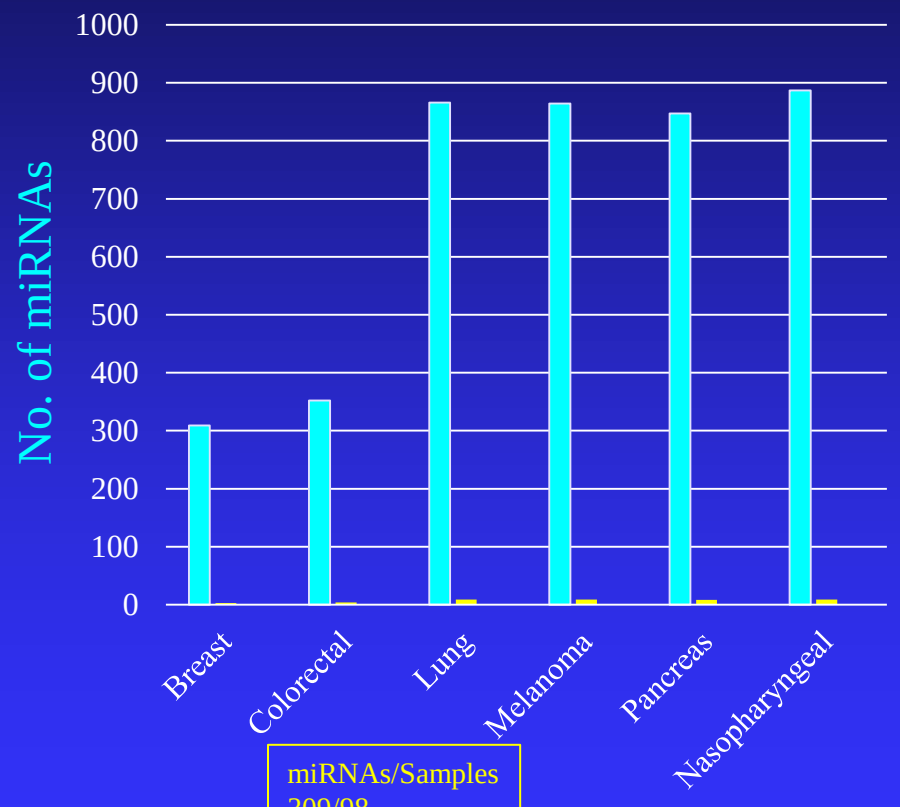
Granulation w.r.t. each miRNA
(granules in one dimension)



M1 ➡ a miRNA

Results: Relevance (All 1% selected)

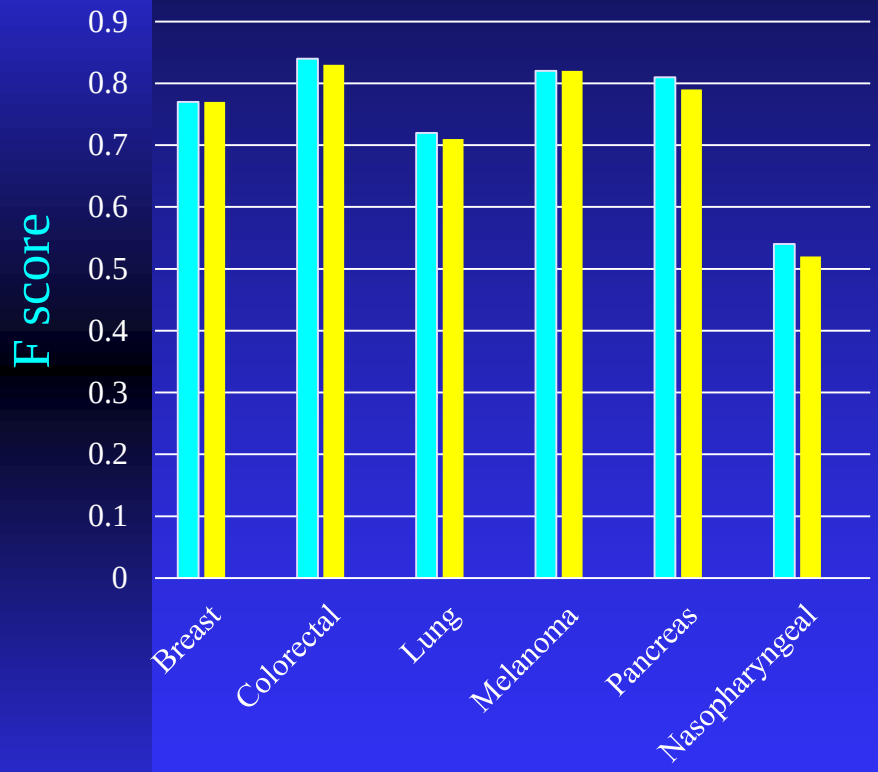


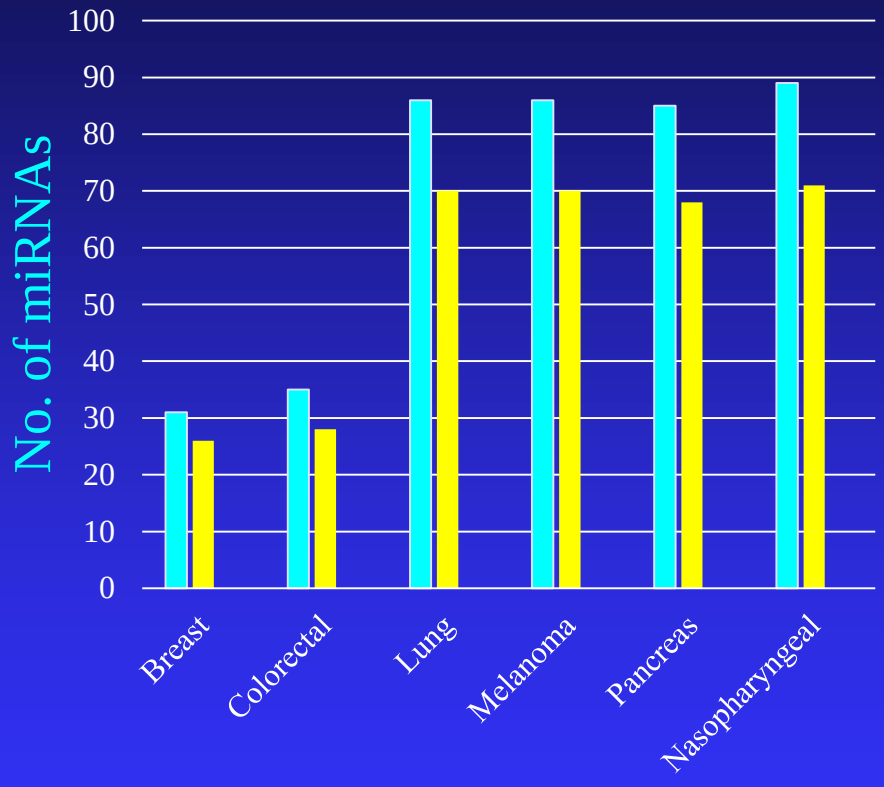


miRNAs/Samples
309/98
352/66
866/36
864/57
847/158
887/50

- Classification using SVM

Results: Redundancy

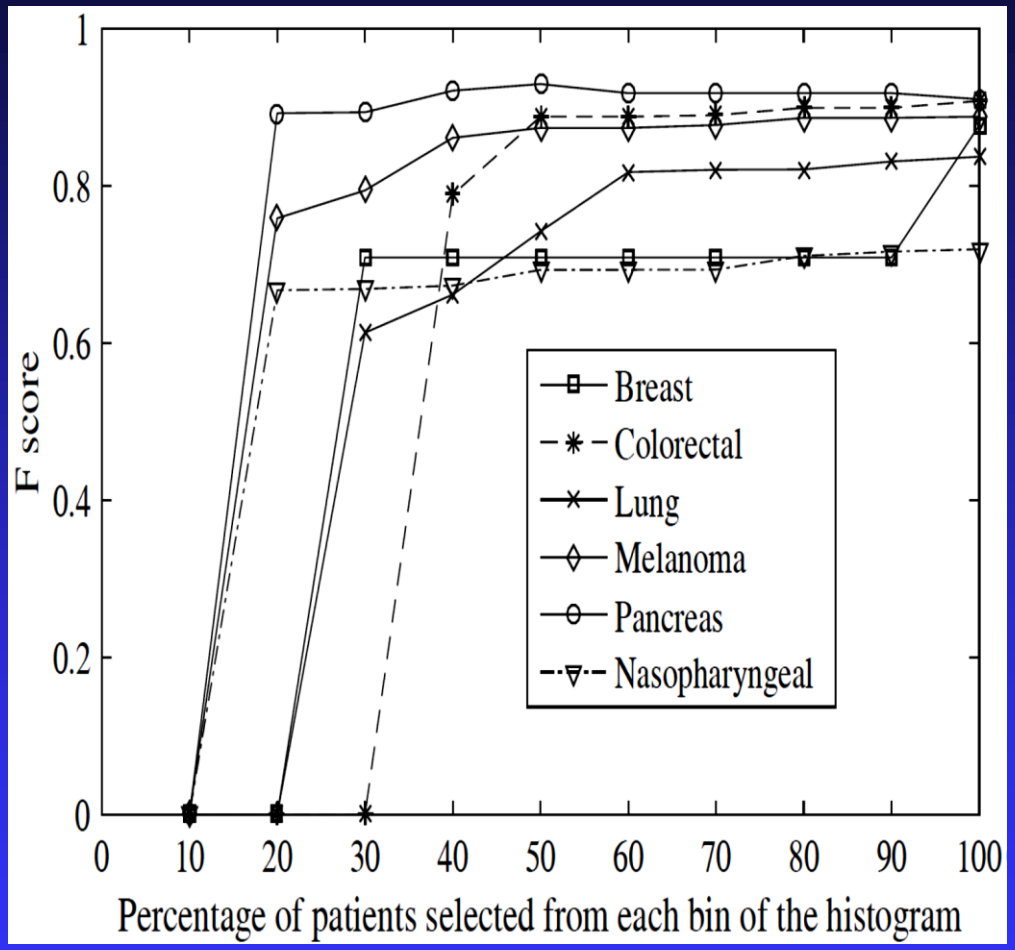




(10% selected 20% removed from them)

- Classification using SVM

Results: HFREM (histogram based selection of patients)



F score by selecting different percentages of patients from √n (n = # each categoty paients) bins of the N and C histograms

Information Sciences, 314, 100-117, 2015

Example:

Fuzzy-rough Community (virtual group) in Social Networks

- High volume, dynamic, complex
- Granules model *f-Relations/ Interactions* of actors

Summary

- Different Machine learning tools
- Data: Videos, Gene expression and social networks

Where are these leading to ?

Relevance to CTP

- *Fuzzy (F)-Granularity* characteristics of Computational Theory of Perceptions (CTP) can be modeled using *Fuzzy-Rough* computing concept

Promising Future Research Problem

Relevance to BIG Data handling

Big-Data is

- High volume (scalable), high velocity (dynamic), high variety (heterogeneous) information
- Usually involves a collection of data sets so large and complex that it becomes difficult to process using conventional data analysis tools
- Requires exceptional technologies to efficiently process within *tolerable elapsed of times*

NEED completely new forms of processing to enable enhanced decision making and knowledge discovery

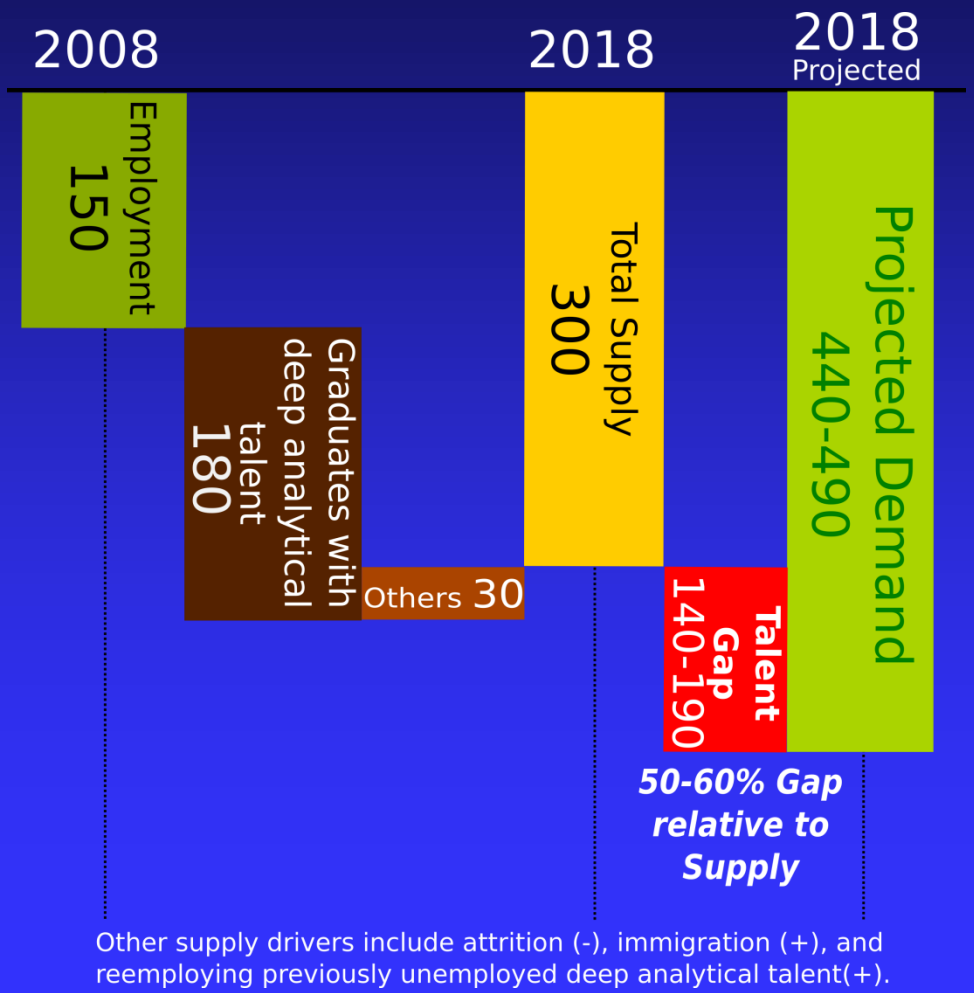
- New approaches – challenges, techniques, tools & architectures to solve new problems

BIG Data Objective:

- To develop complex procedures running over large-scale, enormous-sized data repositories for
  - **extracting useful knowledge** hidden therein,
  - **discovering new insights** from Big Data, and
  - **delivering accurate predictions** of various kinds as required in **tolerable elapsed of time**.

Predicted lack of talent for Big-Data related technologies in USA

- Supply and demand of deep analytical talent by 2018 (in Thousand People)



- 2018 supply: 300 thousand
- 2018 projected demand: 440-490 thousand

SOURCE: US Bureau of Labor Statistics; US Census; Dun and Bradstreet; Company interview; McKinsey Global Institute analysis

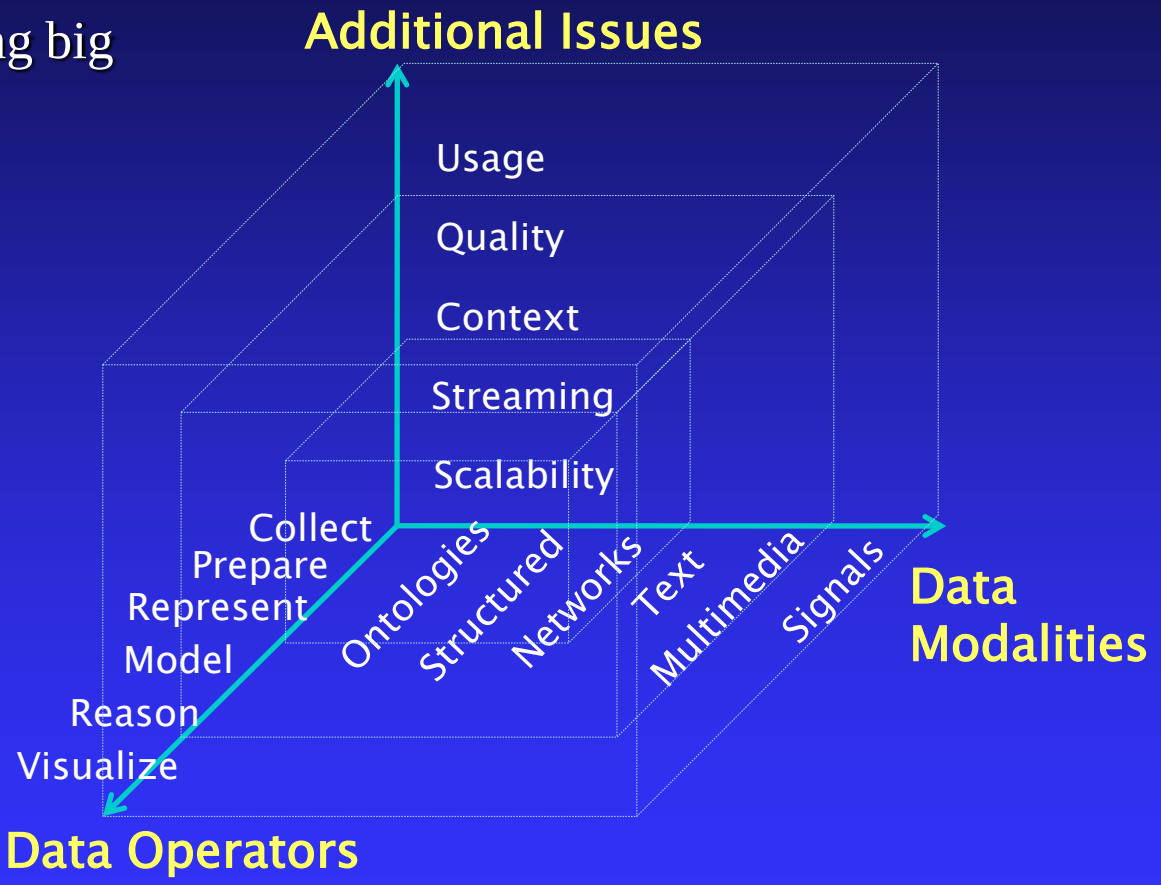
In India

Source: Analytics Special Interest Group, set up by NASSCOM

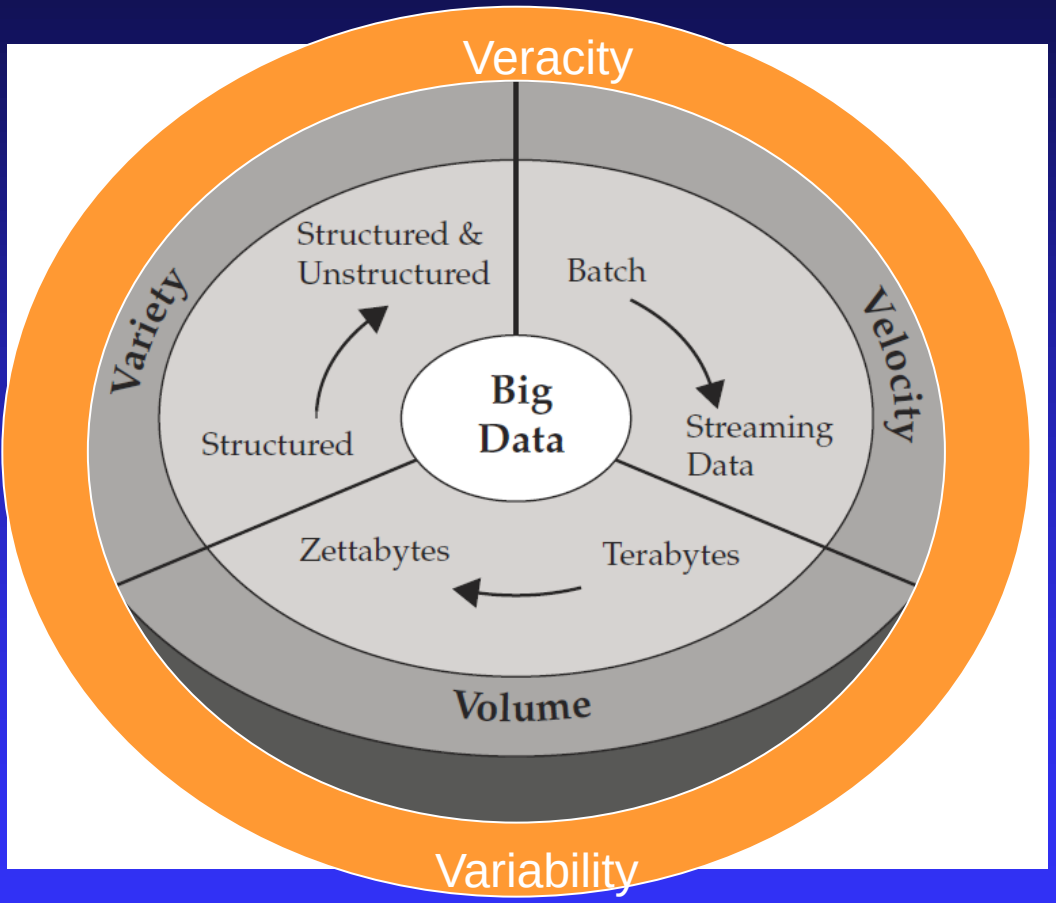
There will be a shortage of about 200,000.00 data scientist in India over the next few years

Dealing with big data: Issues

- Research areas for handling big data include:
  - IR,
  - PR,
  - DM, KDD,
  - ML,
  - NLP,
  - SemWeb,
  - …



Dealing with big data
(Handling challenges lying with all Vs)



Terabyte: $10^{12} = 1000^4$
Zettabyte: $10^{21} = 1000^7$

➡ Demands a revolutionary change *both* in Research Methodologies and Tools

Example: PR (till 80’s) –-> DM (since late Ninties)

- New approaches developed for different tasks of PR to handle DM problems (large data both in size and dimension)
- Example: Feature Selection - where instead of clustering samples in conventional PR, you cluster features themselves in DM

Soft Computing: Roles

- Uncertainty handling
- Learning
- Reasoning
- Searching and optimization

- Efficient, Flexible & Robust methodologies

Dealing with big data: Tasks

- Tasks like:
  - Data size and feature space adaptation
  - Feature selection/ extraction in Big data
  - Uncertainty modeling in learning, sample selection, and classification/ clustering on Big data
  - Granular computing (a clump of objects…)
  - Distributed learning techniques in uncertain environment
  - Uncertainty in cloud computing
  - -
  - -

(Where SC methodologies can be used, in general)

- Soft computing paradigm appears to have a strong promise in developing methodologies

- However, the existing computational intelligence techniques may need be *completely re-hauled* to handle the challenging issues -

  - Large scale Simulation tasks: Need hundreads of Parameter tuning
  - Computational complexity: Nlog(N), or N rather than N^2
  - Incorporate high-level intelligence to gain insight - involve human in the process of computing

❖ Computational aspects and scalability issues - not much addressed by SC community

In conclusion -

Without “Soft Computing”, Machine Intelligence and Data Mining Research Remains **Incomplete.**

- Without *Soft Computing and Granular Mining*, Big Data processing/analysis may remain incomplete

Acknowledgement

- Students and younger colleagues/ collaborators
- Raja Ramanna Fellowship of Govt. of India

Thank You!!

Giant Panda from Chendu: Life is so..o good with bamboo shoots