# Chapter 5.
# Knowledge Representation

Data Mining : A Knowledge Discovery Approach
Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan

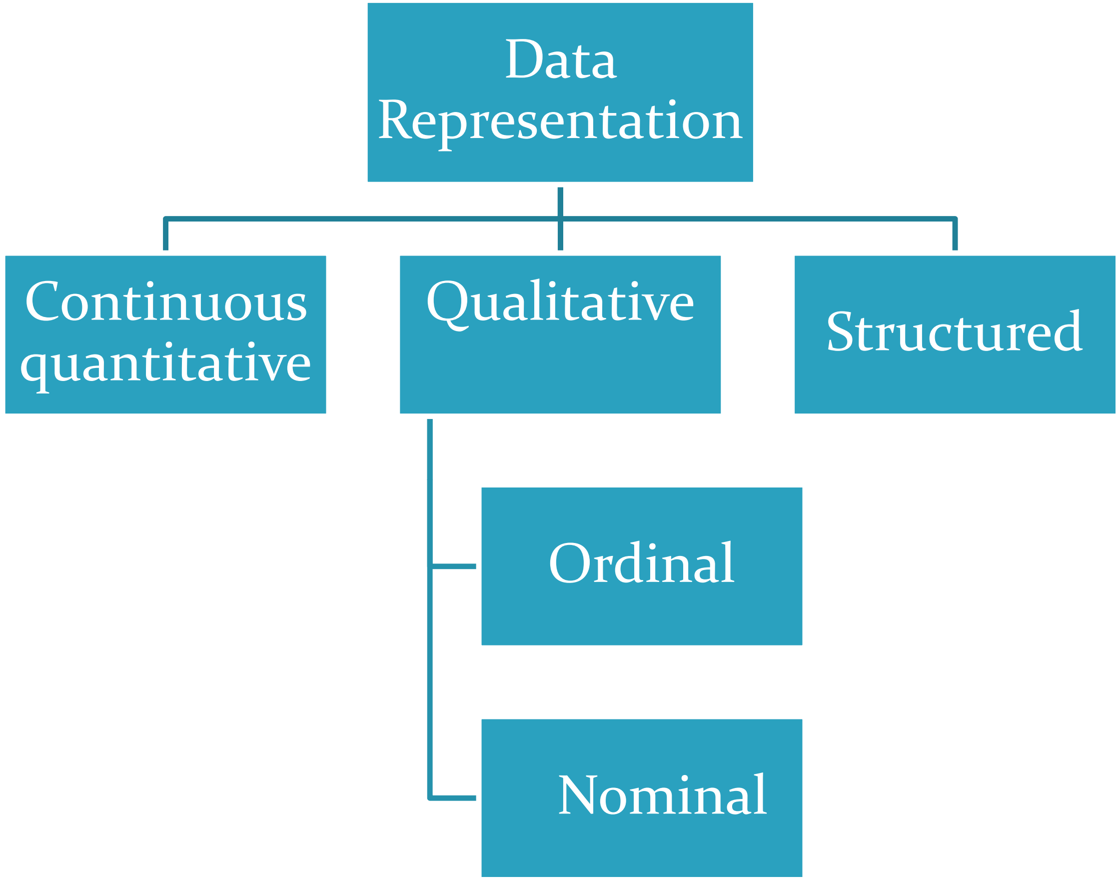
Nikhil K Pawanikar
Asst. Prof, Department of IT, University of Mumbai

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

# 5.1
# Data Representation and their Categories

- **The data available for data mining purposes are highly diversified**
- **several categories of data types (data variables)**
  - **Continuous quantitative data**
  - **Qualitative data**
  - **Structured data**


Data Representation
Continuous quantitative
Qualitative
Structured
Ordinal
Nominal

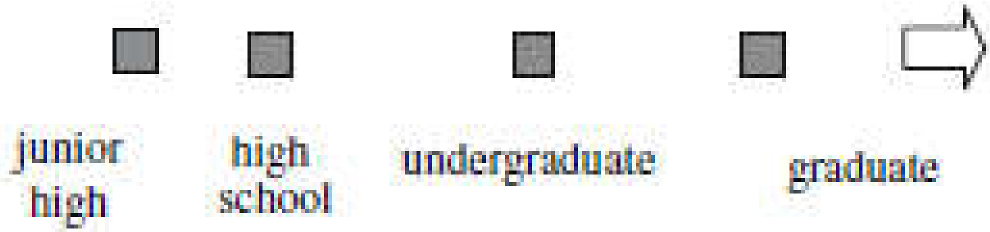
# Continuous quantitative data

- These data concern continuous variables
- They occur very often in relationship to physical phenomena as these data naturally generate such continuous variables.

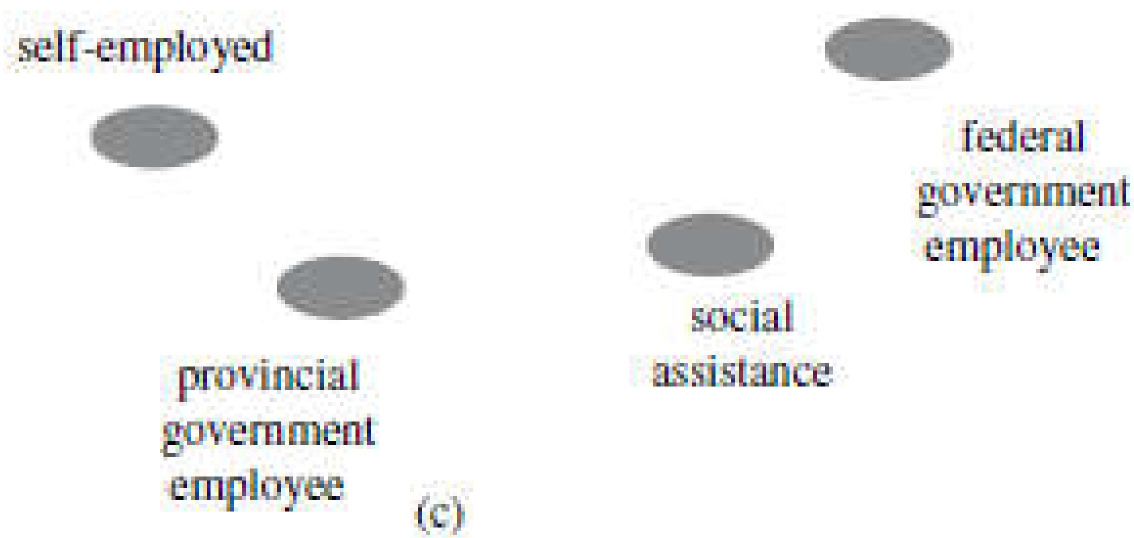
Pressure [kPa]

# Qualitative data

- typically assume some limited number of values
- either a linear organization which could be established (**ordinal qualitative data**)
- or no such order could be formed (**nominal qualitative data**)
- Qualitative data could also be the result of a transformation of original continuous quantitative data, subject to some generalization



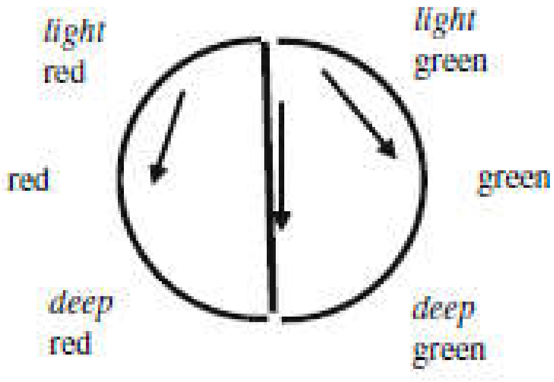

**ordinal qualitative**

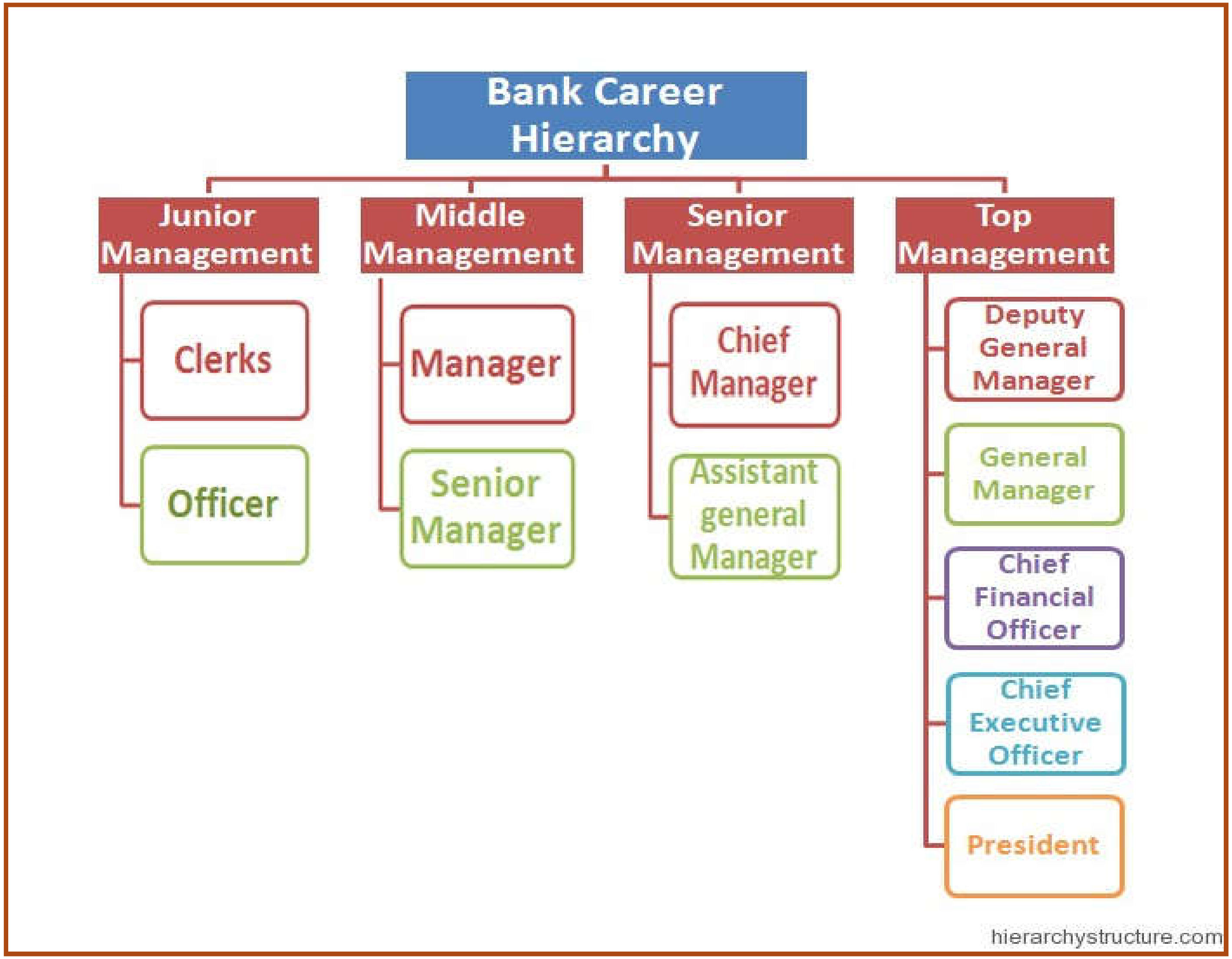



**nominal qualitative**

# Structured data

- These data become more complex as we build a **hierarchy of specialized** concept
- This process gives rise to tree structures whose nodes represent the values encountered in the problem while the links (edges) indicate relationships between them




Bank Career Hierarchy
Junior Management
Clerks
Officer
Middle Management
Manager
Senior Manager
Senior Management
Chief Manager
Assistant general Manager
Top Management
Deputy General Manager
General Manager
Chief Financial Officer
Chief Executive Officer
President
hierarchystructure.com

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. **Categories of Knowledge Representation**
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

# 5.2
# CATEGORIES OF KNOWLEDGE REPRESENTATION

- The basic concept or definition for knowledge representation is “Organizing available knowledge and representing it in a understandable manner, as well as dealing with the knowledge acquired through data mining.

- There are some main categories for Knowledge Representation, we call them as Knowledge Representation Schemes.

  - **Theses categories are-**
    - RULES
    - GRAPHS
    - NETWORKS

## RULES

- The most general form of the Rules is,

IF condition THEN conclusion

- In this for of knowledge Representation, ‘condition’ & ‘conclusion’ are the two descriptors of knowledge.

- There are chances where conclusion depends upon one or more conditions ,
  - For example,

IF $\text{condition}_1$ *and* $\text{condition}_2$ *and* …. *and* $\text{condition}_n$ THEN conclusion

## ❖ TYPES OF RULES

### ❑ GRADUAL RULES

- In gradual (graded) rules, rather than expressing an association between condition and conclusion ,we capture the trend within information granules and hence the condition and conclusion parts will each contain a term referring to that trend.

"the higher the values of the condition, the higher the values of the conclusion"

- For example,

"the higher the income, the higher the taxes"

Here, Income is a condition and taxes are conclusion but the value of taxes depends upon value of income so we can see the trends within information granules.

## ❑ <u>QUANTIFIED RULES</u>

➢ This is the rule for knowledge representation in which the condition has the quantitative effect on the conclusion.

➢ For example,

If any person is checking the sale for non-veg items for a year in the Mumbai region then that person will come to know that sale is decreased in the period of july-august due to it is time of "shravan" when people do not eat non-veg

➢ In this scenario, condition is not a quantity but it has quantitative effect.

## ❑ANALOGICAL RULES

- Here the rules focus on levels of similarity (closeness, resemblance, etc.) between pairs of items standing in the condition and the conclusion.

$$\text{IF similarity}(A_i, A_j) \quad \text{THEN similarity}(B_i, B_j)$$

- Here, Ai,Aj and Bi,Bj are two different entities

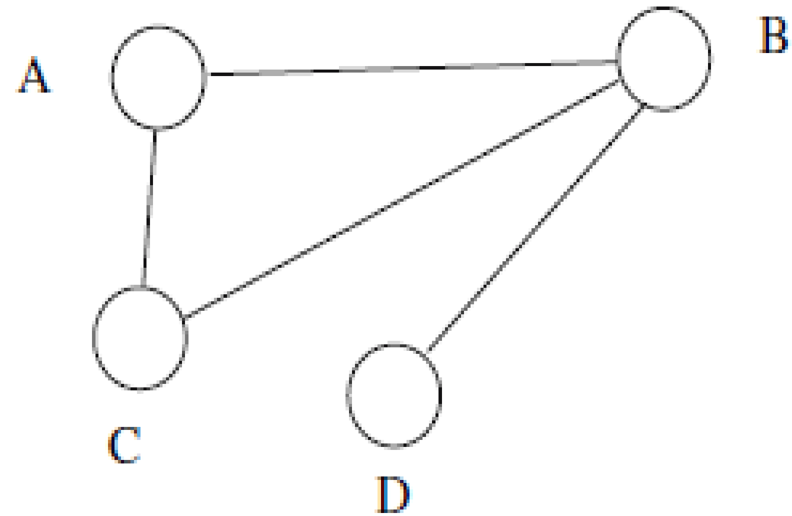
- For example,

  If there is hike in prices of diesel and petrol, then there will be hike in food items prices and number of commuters for public transport.

In this example, condition and conclusion have the similar effect.

# ❖ GRAPHS

- Graphs are fundamental constructs representing relationships between concepts.
- Concepts can be anything that representing a set of data or a single entity etc.
- Concepts are represented as nodes of a graph.
- Graphs are helpful in visualizing a collection of concepts and presenting key relationships (dependencies) between them.
- They are highly appealing to designers as well as to users of developed systems.
- The linkages (associations, dependencies, etc.) between the concepts are represented as the edges of the graph.

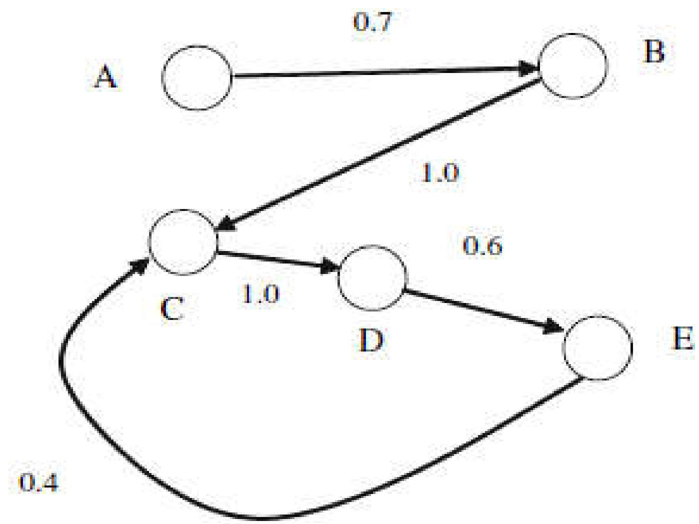



An example graph representing dependencies between concepts.

- Here, A,B,C and D are various concepts present in the data.
- A,B,C,D are the nodes the linkages among them are shown in the graph.
- The linkages are A-B, A-C, B-C, B-D

# ❖ **DIRECTED GRAPHS**

- There are many variations and augmentations of generic graphs.
- First, Directed Graphs
  - Their links can indicate directional relationships between the nodes.
  - In directed graphs, we can show Chaining effect as well as Looping effect.
- Second, graphs can also show numeric quantification of the links that indicate the strength of dependency between the concepts.




A directed graph showing a chaining effect (A-B-C) and a loop (C-D-E-C) in the graph.

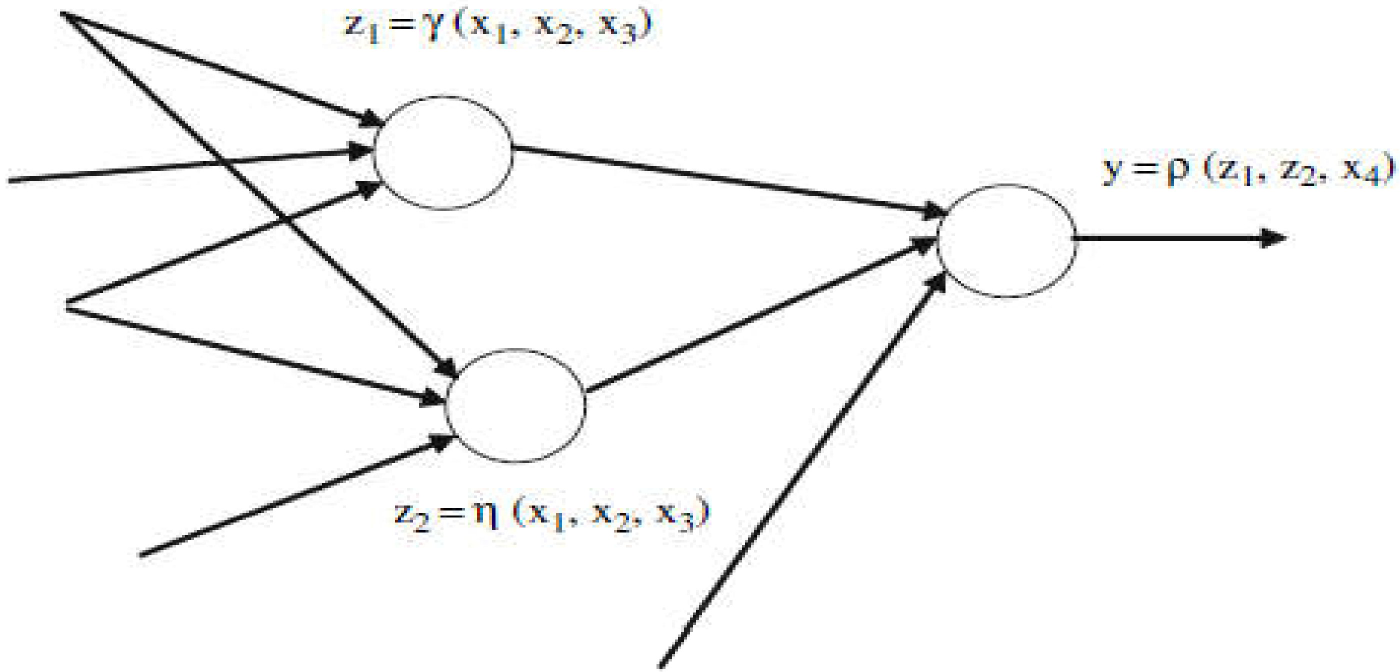
- In this graph, the linkages among the nodes are directed.
- We can show chaining effect as (A-B-C) and looping effect as (C-D-E-C).
- The numeric quantifications are also showed in above graph indicating strength of dependency between the concepts.

## ❖ NETWORKS

- the Network not only represents the knowledge itself but also contains the underlying processing being realized at the local level.

- Networks can be regarded as generalized graphs i.e. the directed graphs which are connected to each other to form a network.

- Networks can be built in a hierarchical structure where a node at a higher conceptual level unfolds into a collection of nodes at a lower level.




- X1,x2,x3 are some values coming to the z1 and z2 nodes
- Y is a major process which is accepting the values from z1 and z2 and processing it.
- Z1 and z2 are the underlying processes x1,x2,x3 are variables for the same process.

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation

3. **Granularity of Data and Knowledge Representation Schemes**

4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

- **Information granules** that support **human-centric computing.**

- **human-centricity** - all faculties of a computing systems that facilitate interaction with human users

- This facilitation could be either by improving and/or enhancing the quality of communication of findings or by accepting user inputs that can be realized in a more flexible and friendly manner

- Information granules tend to absorb human endeavors.

  - We (humans) usually cast any task in a certain conceptual framework of basic entities

  - Using this framework, we formulate generic concepts adhering to some level of abstraction, carry out processing, and communicate the results to the external environment.

  - Ex: Processing of an Image by human beings compared to computers

  - Interpretation of ECG signal as certain discrete symbols

- Depending on a specific problem and who the user is, the size of information granules (time intervals) could vary dramatically.
- For a designer of high-speed digital systems the temporal information granules are nanoseconds and microseconds.
- These commonly encountered examples are convincing enough

  to lead us believe that

  (a) information granules are key components of knowledge representation and processing,

  (b) the level of granularity of information granules (their size) becomes crucial to the problem description, and

  (c) there is no universal level of granularity of information; the size of granules is problem specific and user dependent

- The challenge - to develop a computing framework within which all these representation and processing efforts could be formally realized.
- The common platform emerging within this contextcomes under the name **granular computing**, an emerging paradigm of information processing.

- Granular Computing is innovative in several ways:

1. It identifies essential commonalities between surprisingly diversified problems and technologies, which can be cast into a unified framework that we usually refer to as a granular world.

2. With the emergence of the unified framework of granular processing we get a better grasp on the role of interaction between various formalisms and can visualize the way in which they communicate.

3. It brings together the existing formalisms of set theory (interval analysis), fuzzy sets, rough sets, etc. under the same roof by clearly showing that in spite of their visibly distinct underpinnings, they also possess fundamental commonalities.

4. By building upon the commonalities of the existing formal approaches, Granular Computing helps build heterogeneous and multifaceted models of processing of information granules with technologies that

5. Granular Computing fully uses a notion of variable granularity, whose range may cover detailed numeric entities and very abstract and general information granules. It looks at the aspects of compatibility of information granules and communication mechanisms between the granular worlds.

- To view Granular Computing from a historic perspective we can acknowledging that a number of

  fundamental mechanisms of granulation have come from

  - ✓ interval analysis
  - ✓ fuzzy sets
  - ✓ uncertain variables
  - ✓ rough sets.

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes

4. **Sets and Interval Analysis**

5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

- Sets are rooted in a two-valued logic using the fundamental predicate of membership "belongs to" or "element of" ($\in$).

- Interval analysis depends upon a collection of intervals in the line of real's [a,b],[c,d].

- Sets are fundamental manifestations of abstraction whose role is to organize knowledge about the world and to form the most suitable perspective from which that knowledge could be described.

- The most essential manipulation of sets is carried out through basic operations such as union, intersection, complement, exclusive or, and the like.

## ALGEBRIAC OPERATION'S

- Union of A and B

$(A \cup B)(x) = \max(A(X),(B(X))$

- Intersection of A and B

$(A \cap B)(x) = \min(A(X),(B(X))$

- Complement of A $= 1 \quad A(x)$

## Algebraic processing in interval analysis:

- Addition

$$[a,b]+[c,d] \quad = \quad [a+c,b+d]$$

- Multiplication

$$[a,b] \ [c,d] = [\min(ac, ad, bc, bd), \max(ac, ad, bc, bd)]$$

- Division

$$[a,b]/[c,d] = [\min(a/d, a/c, b/c, b/d), \max(a/d, a/c, b/c, b/d)]$$

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis

5. **Fuzzy Sets as Human-Centric Information Granules**

6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

- Fuzzy sets allows a concept of **partial membership** so that elements that are "typical" to the concept and those of borderline character can be discriminated.

The **membership function** $\mu_A(x)$ is defined as

$$\mu_A(x) = \begin{cases} 1 & \text{iff } x \in A, \\ 0 & \text{iff } x \notin A. \end{cases}$$

- One cannot specify a single, well-defined element that forms a solid border between full belongingness and full exclusion. Warm weather, high speed, fast car are examples of information granules that are fuzzy in nature.

- Fuzzy sets, with their soft or gradual transition boundaries, are ideal to capture the notion of partial membership.
- A fuzzy set A defined in X is characterized by its membership function:

A : X --->[0,1]

where A(x) denotes a degree of membership of x in A.

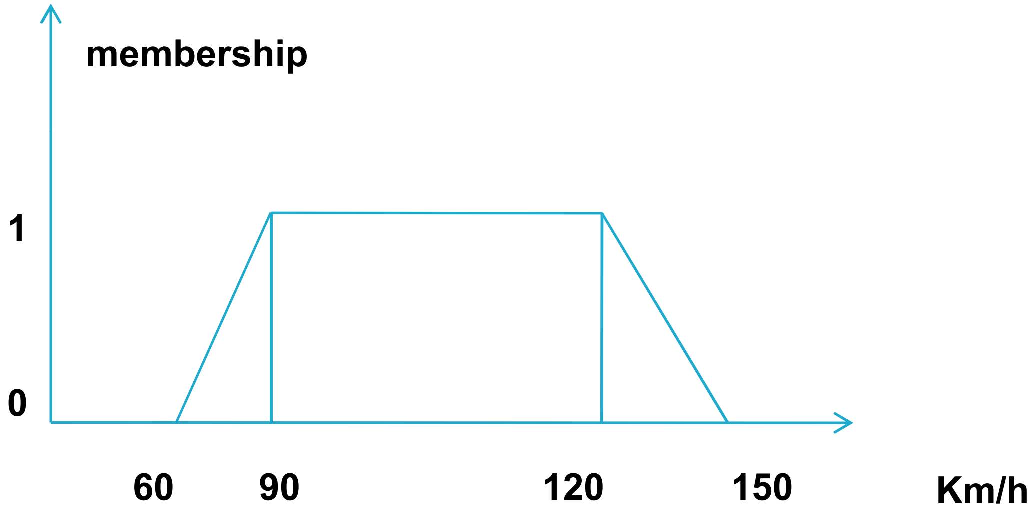


**Table 5.3.** Generic descriptors used in the characterization of fuzzy set $A$.

| Notion and definition | Description |
|---|---|
| $\alpha$-cut $A_\alpha = \{x \mid A(x) \geq \alpha\}$ | Set induced by some threshold consisting of elements belonging to $A$ to an extent not lower than $\alpha$. By choosing a certain threshold, we convert $A$ into the corresponding set representative. $\alpha$-cuts provide important links between fuzzy sets and sets. |
| Height of $A$, $\text{hgt}(A) = \sup_x A(x)$ | Supremum of the membership grades; $A$ is normal if $\text{hgt}(A) = 1$. Core of A is formed by all elements of the universe for which $A(x)$ attains 1. |
| Support of A, $\text{supp}(A) = \{x \mid A(x) > 0\}$ | Set induced by all elements of $A$ belonging to it with nonzero membership grades |
| Cardinality of $A$, $\text{card}(A) = \int_{\mathbf{X}} A(x)dx$ (assumed that the integral does exist) | Counts the number of elements belonging to $A$; characterizes the granularity of $A$. Higher $\text{card}(A)$ implies higher granularity(specificity) or, equivalently, lower generality |

- A family of fuzzy sets defined in **X** is denoted as F(**X**).
- Fuzzy sets are provided in the form of **membership functions** either in their continuous or discrete format.
- Fuzzy sets defined in the line of real numbers (**R**) whose membership functions satisfy several appealing properties such as

    (a) **unimodality**,

    (b) **continuity**, and

    (c) **convexity**, are referred to as **fuzzy numbers**.

## Operations on Fuzzy sets

- Fuzzy sets defined in the same space are combined logically through logic operators of intersection, union, and complement.

$A \cap B : (A \cap B)(x) = T\ (A, B) = \min(A, B)$

$A \cup B : (A \cup B)(x) = T\ (A, B) = \max(A, B)$

$A^{-} : \quad A^{-}\ (x) = 1\ \ A(x)$

## Fuzzy Relations

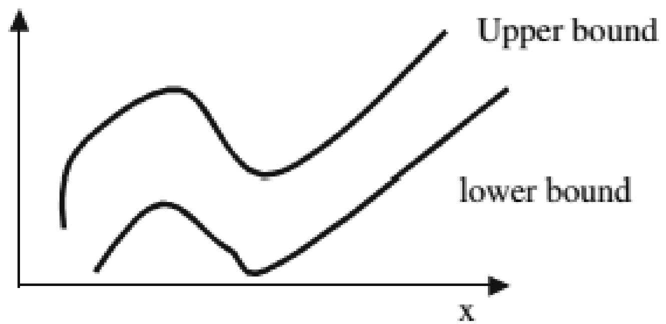
- Fuzzy relations are defined in Cartesian products of some spaces and represent composite concepts.
- For instance, the notion "high price and fast car" can be represented as a fuzzy relation R defined in the Cartesian product of price and speed.

$$R: X \times Y \rightarrow [\mathbf{0}, \mathbf{1}] \text{ with}$$

X and Y being the corresponding spaces of price and speed.

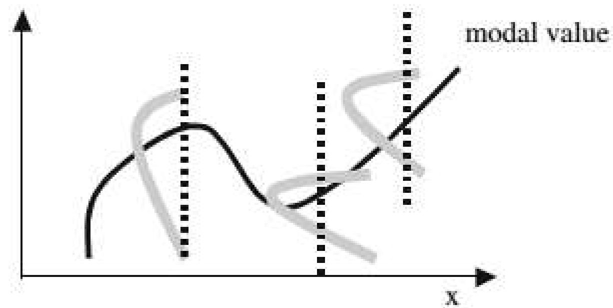
# Second-Type Fuzzy Sets

- Fuzzy sets are constructs with membership grades in the unit interval.
- Second type fuzzy sets are fuzzy sets whose membership grades are fuzzy sets defined in [0,1].
- Here we depart from individual numeric membership grades and acknowledge that the degrees of membership themselves could exist in interval-valued fuzzy sets.



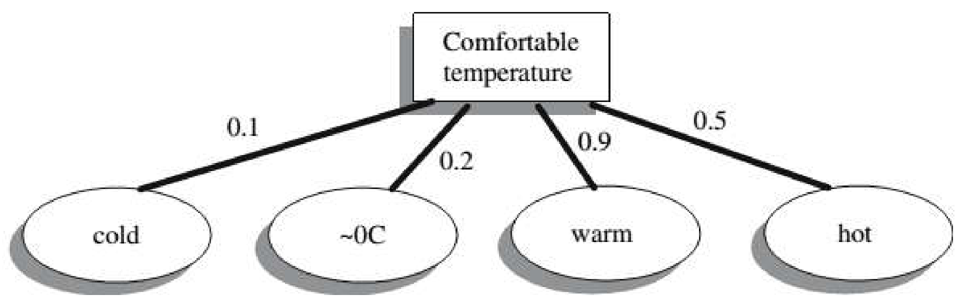
(a)

# Fuzzy Sets of Order-2




- Fuzzy sets of order 2 are another conceptual extension of fuzzy sets, where we define a certain fuzzy set over a universe of several reference fuzzy sets.

- For instance, the term comfortable temperature can be defined by a collection of generic terms such as cold temperature, around zero, warm, hot etc



# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules

6. **Shadowed Sets**

7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets

7. **Rough Sets**

8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets

**8. Characterization of Knowledge Representation Schemes**

9. Levels of Granularity and Perception Perspectives
10. The Concept of Granularity in Rules

- The most essential facets worth considering are the following:

– expressive power of the model

– computational complexity and associated tradeoffs vs. flexibility of the knowledge representation scheme as well as its scalability

– designer and user familiarity with the specific knowledge representation scheme

– effectiveness of creating such a model on the basis of the existing domain knowledge and experimental data

– character of information granulation and the level of specificity of the information granules to be used in the constructs of knowledge representation.

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes

9. **Levels of Granularity and Perception Perspectives**

10. The Concept of Granularity in Rules

- Information granules are regarded as building blocks that capture the essence of the problem and facilitate its handling

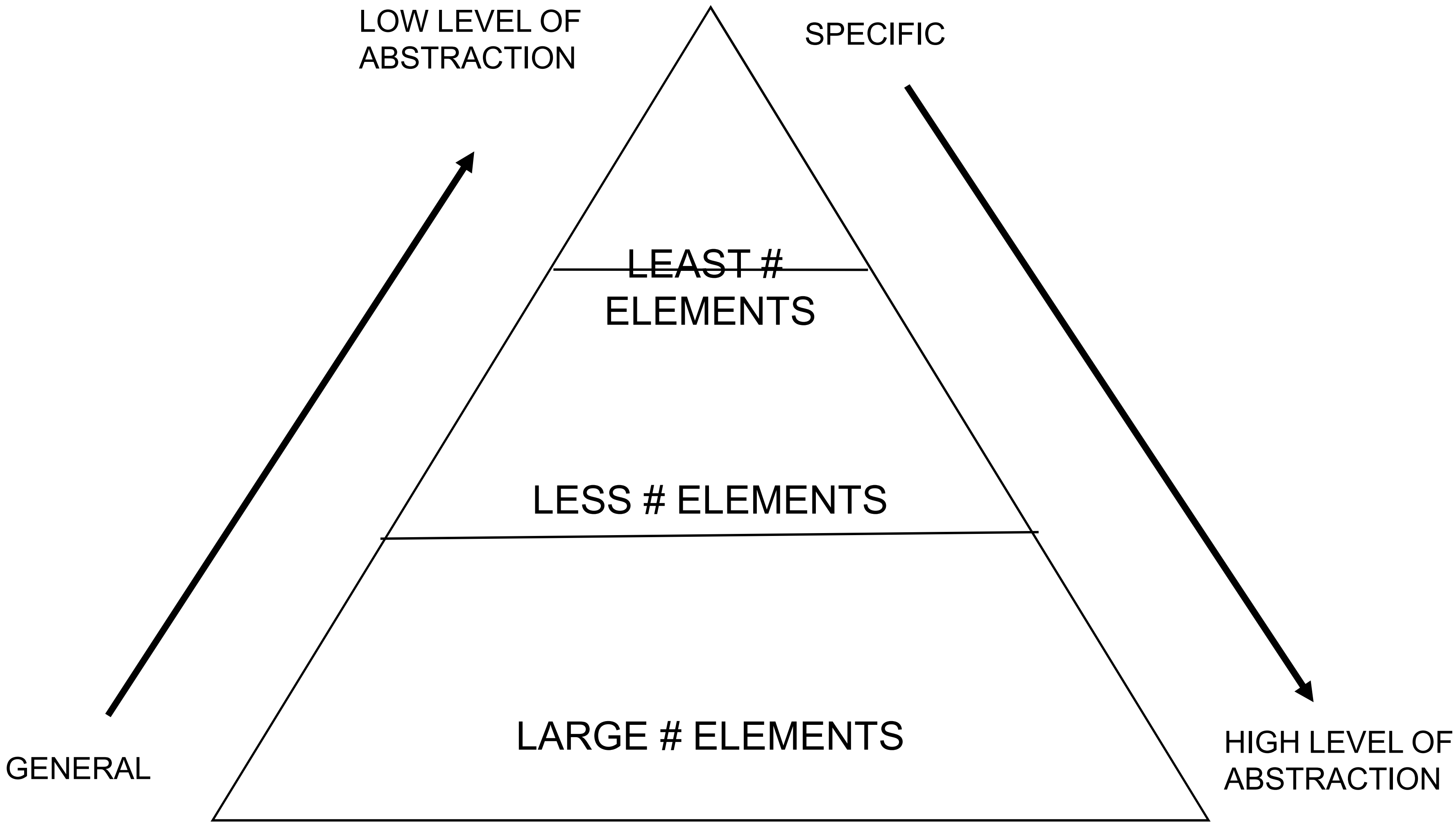
- By delivering a certain mechanism ofabstraction, information granules can be a flexible means to deal with the required level ofspecificity/generalityof the problem description

- In general granularity is associated with the number of elements or concepts that are included in the form of some information granule.

- The terms granularity and specificity are reciprocal: an increase in one of them leads to a decrease in the other.

  i.e Generality - The larger the number of components, the higher the level of abstraction, in other words, the level of granularity gets lower.

  Specificity - If only a few elements are contained in a certain information granule, we say that this granule is very specific, in other words, the level of granularity gets lower.


LOW LEVEL OF
ABSTRACTION
SPECIFIC
LEAST #
ELEMENTS
LESS # ELEMENTS
LARGE # ELEMENTS
GENERAL
HIGH LEVEL OF
ABSTRACTION

- Since granularity is reflective of the number of elements within the information granule, for quantifying we need some sort of counting of the number of elements.
- In the case of sets, the process is straightforward(belongingness)
- In the case of other formalisms such as fuzzy sets or rough sets, the process is not so simple

**Table 5.5.** Computing the granularity of information granules expressed as fuzzy sets and rough sets.

| **Information granules** | **Granularity** |
|---|---|
| Fuzzy sets | σ-count, for given fuzzy set $A$, its σ-count is determined as $\int_X A(x)dx$ (we assume that the integral does exist). In the case of a finite space **X**, the integral is replaced by a sum of the membership grades |
| Rough sets | cardinality of the lower and upper bound, card $(A_+)$, card$(A_-)$; the difference between these two describes roughness of $A$ |

# Topics to be covered:

1. Data Representation and their Categories: General Insights
2. Categories of Knowledge Representation
3. Granularity of Data and Knowledge Representation Schemes
4. Sets and Interval Analysis
5. Fuzzy Sets as Human-Centric Information Granules
6. Shadowed Sets
7. Rough Sets
8. Characterization of Knowledge Representation Schemes
9. Levels of Granularity and Perception Perspectives

10. **The Concept of Granularity in Rules**

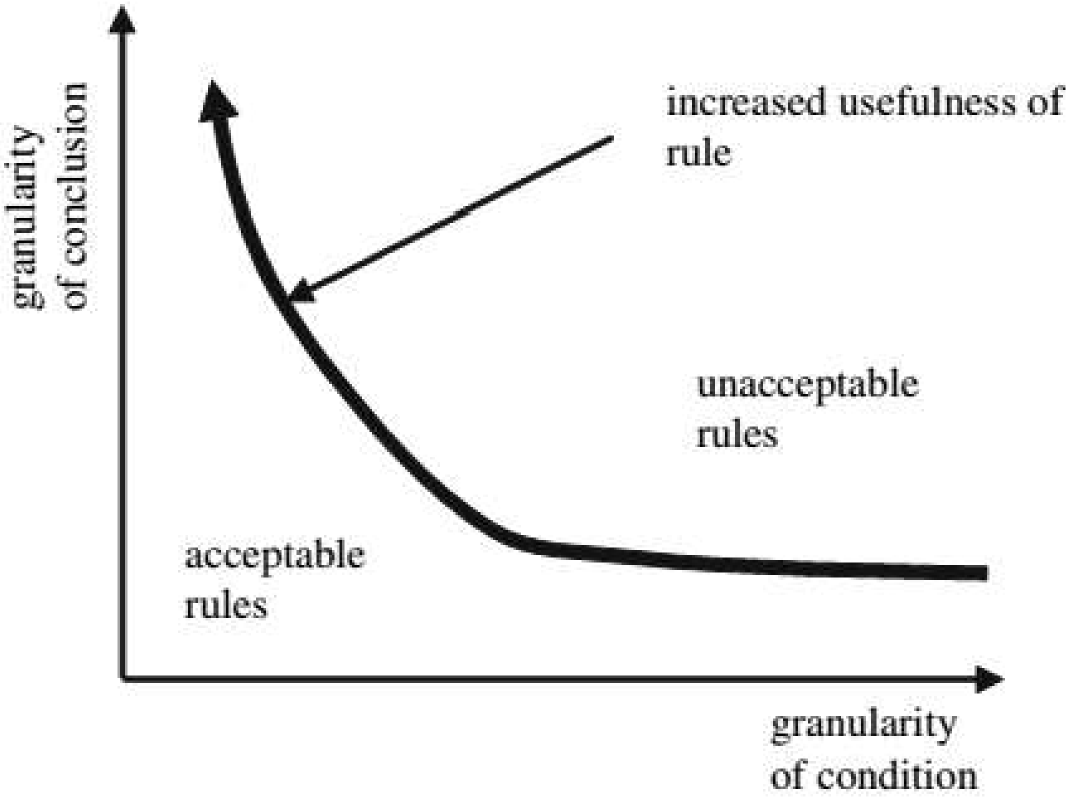
- Rules are generic constructs of knowledge representation whose components are information granules.
- By virtue of their very nature, they tend to capture general relationships between variables (attributes).
- To make meaningful dependencies, information granules have to be expressed at the level of abstract entities.
- For Ex,consider the following rules

  "if high temperature then a substantial sale of cold drinks"

  "if temperature of 28C then sale of 12,566 bottles of cold drinks"

- Bottom line : Rules come hand in hand with information granules

# Use of granular information in rule-based computing



- “If A then B” where A and B are represented as numeric intervals in the space of real numbers.
- A low level of granularity of the condition associated with a high level of granularity of the conclusion describes a rule of high relevance

# END OF CHAPTER