Data Mining and Knowledge Discovery in Biomedical Images

Otman A. Basir
Pattern Analysis and Machine Intelligence Group
Department of Systems Design Engg,

University of Waterloo

Outline

- The Problem
- From Data to Knowledge
- Data Mining
- Knowledge Discovery
- Discovering Knowledge from Meta Information Sources
- The LORNET objective
- Conclusion

The problem

- rapidly expanding data collection
- manual analysis of data is unrealistic
- Data contains valuable information that can aid in decision making, and may reveal interesting trends
- Processing power and storage capacity are increasing, but the number of potential patterns to investigate grows exponentially of number of attributes
- →
- Need scalable and efficient algorithms to extract knowledge from data stores
- Need methods to distinguish interesting information from uninteresting information

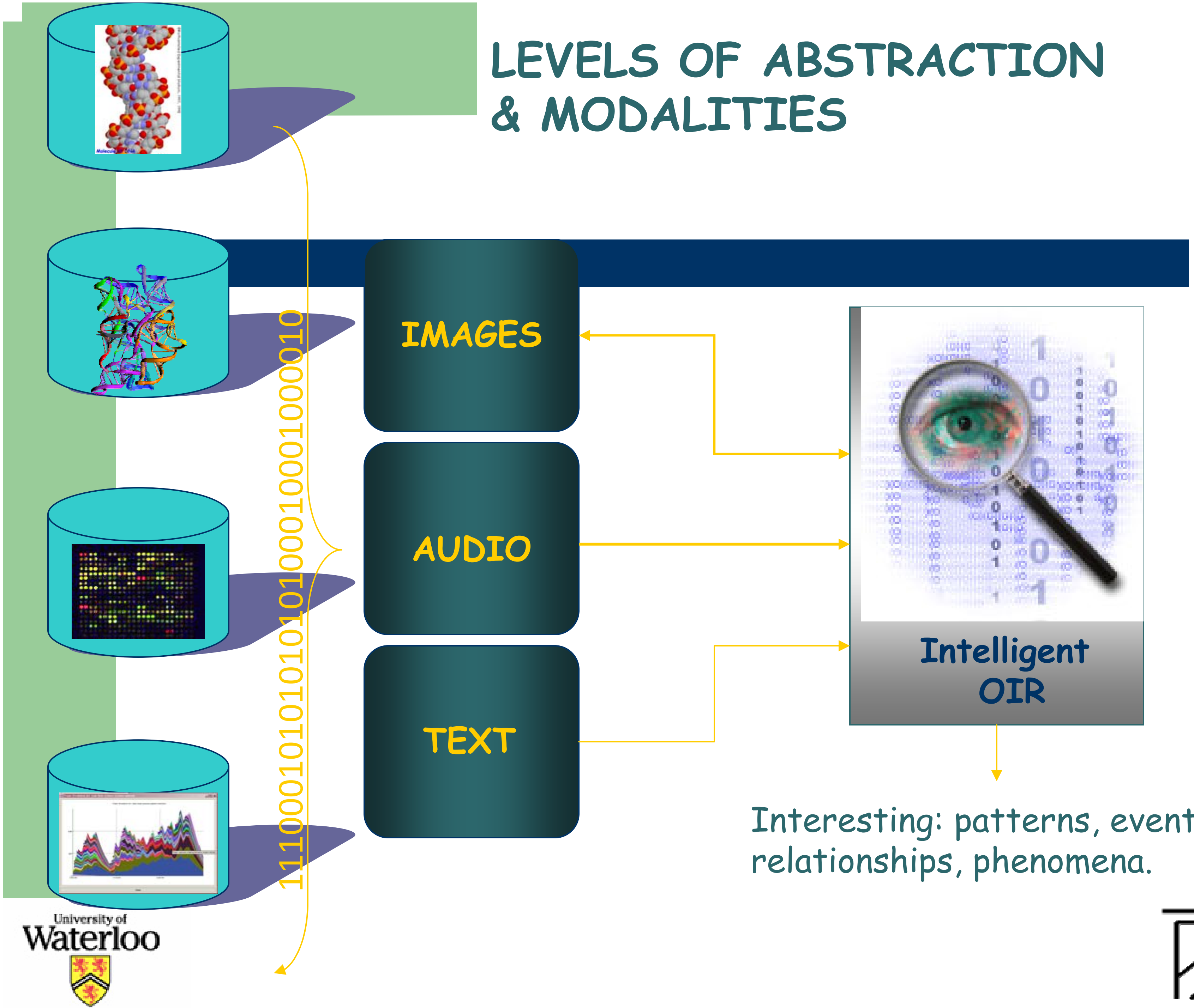
Bio-Informatics and DM

- Biological data are abundant and information rich
  - Multimodal
  - Can ambiguous
  - Data produced at different levels
    - molecules, cells, organs, organisms, populations
  - Data obtained from different channels
    - Structure: sequence, shape, energy,…
    - Function: gene expression, pathway, phenotypic and clinical data,…

The Problem

- The vast size and the multimodal nature of biomedical information and the knowledge it contained make it essential to introduce intelligent organization, interpretation, and retrieval methods.

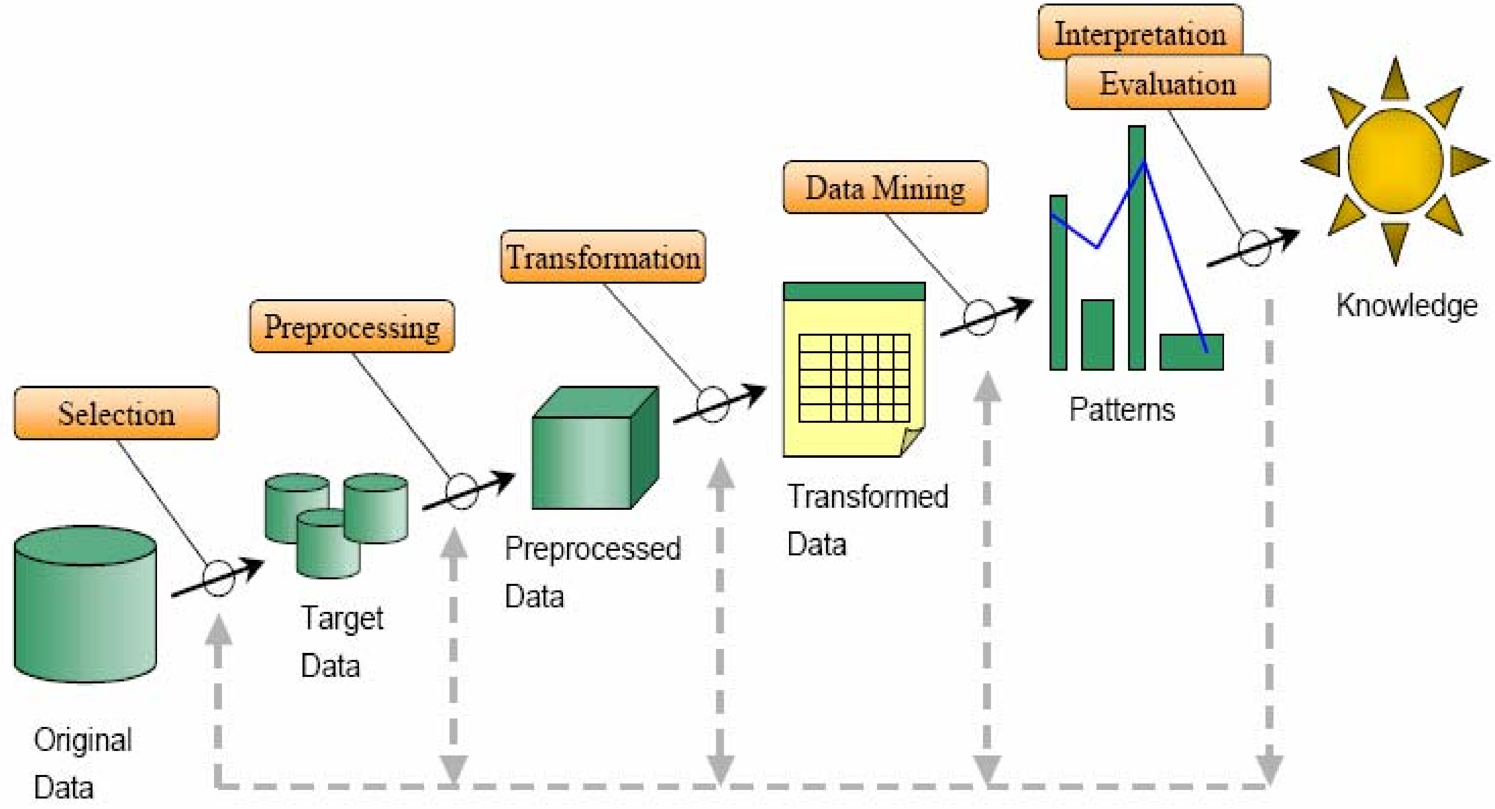
LEVELS OF ABSTRACTION & MODALITIES



Interesting: patterns, events, relationships, phenomena.

University of Waterloo

IOR: From data to knowledge



Ref: Yang and Wong

From Data to Knowledge

- *KDD is a process*

  Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable ***patterns*** *in* ***data***

- *Data Mining is a step*

  Data Mining is a step in the KDD process consisting of particular ***algorithms*** *that, under some acceptable computational efficiency limitations, produces a particular enumeration of* ***patterns*** *over data*

  - *Classification, Clustering*
  - *Regression, Summarization*
  - *Dependency modeling, Change/Deviation detection*

University of Waterloo

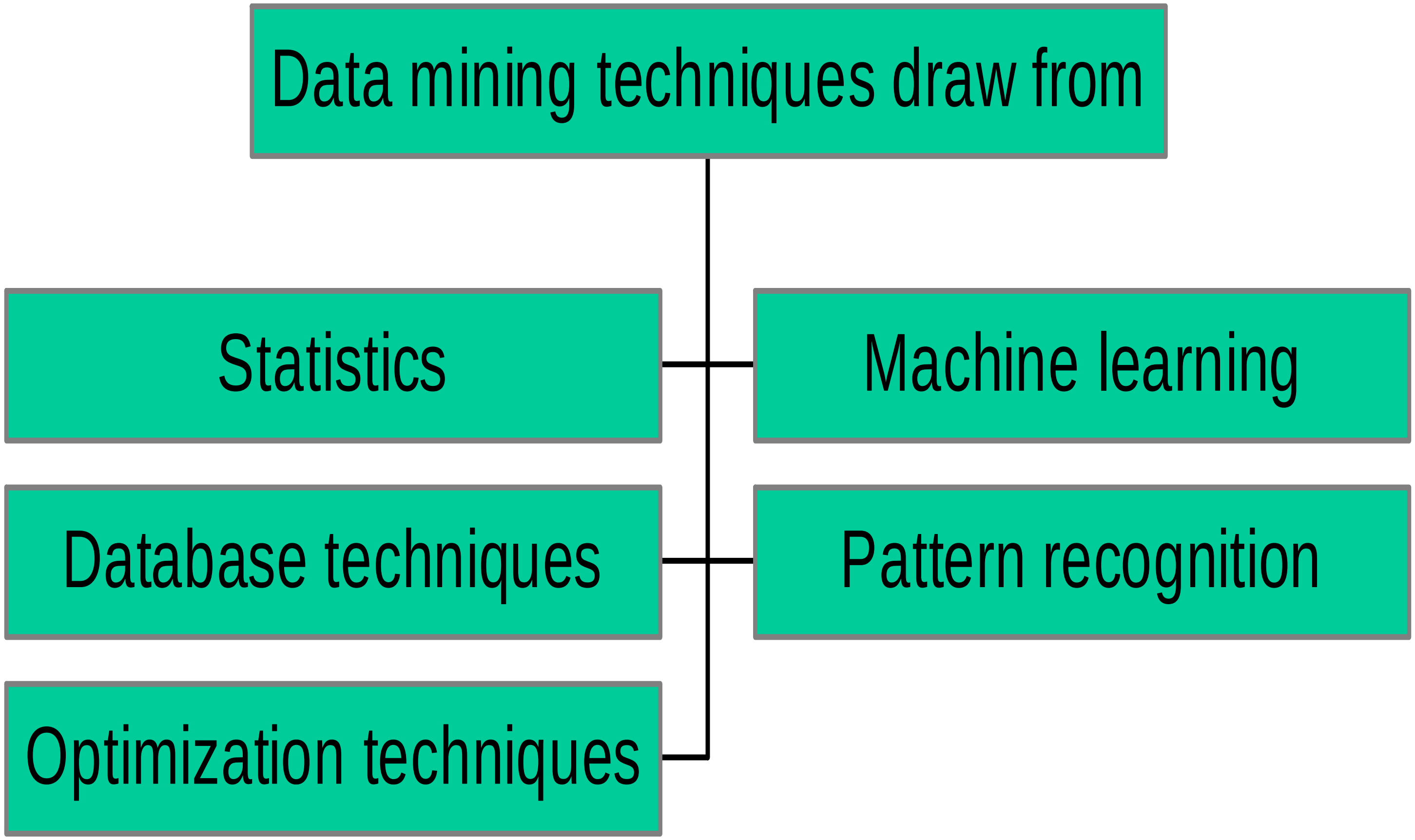
Data Mining: what and what not!

- Discover interesting patterns, relationships, and trends in data → knowledge
- designed for large data sets
- Scale is one of the characteristics that distinguishes data mining applications from traditional machine learning applications
- Data mining techniques will discover patterns in any data
- The patterns discovered may be meaningless
- User determines how to interpret the results
- cannot generate information that is not present in the data
- They can only find the patterns that are already there

Types of Mining

- Association Rule Mining
  - Initially developed for market basket analysis
  - Goal is to discover relationships between attributes
  - Uses include decision support, classification and clustering
- Classification and Prediction (Supervised Learning)
  - Classifiers are created using labeled training samples
  - Training samples created by ground truth / experts
  - Classifier later used to classify unknown samples
- Clustering (Unsupervised Learning)
  - Grouping objects into classes so that similar objects are in the same class and dissimilar objects are in different classes
  - Discover overall distribution patterns and relationships between attributes

Approaches to Data Mining



Data mining techniques draw from

Statistics

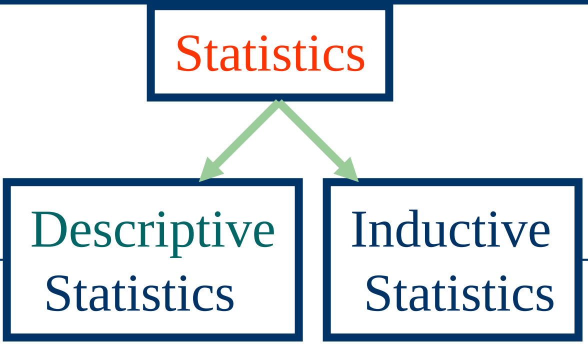
Machine learning

Database techniques

Pattern recognition

Optimization techniques

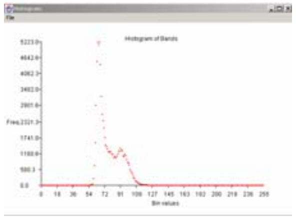
Statistics

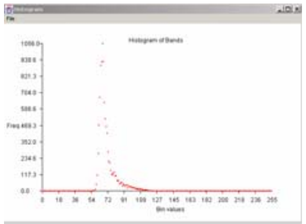


Data
Description
77766621127w123
65`12`3uyhw`19ew8t27ew

Describe data

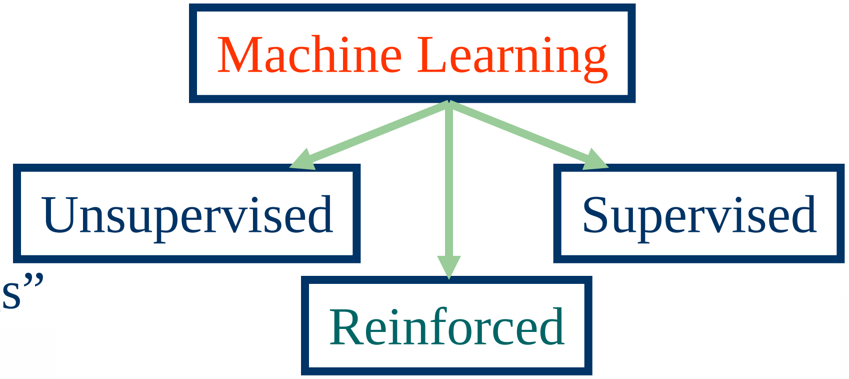
Make forecast
and inferences





Are two sample sets
identically distributed ?

Machine Learning

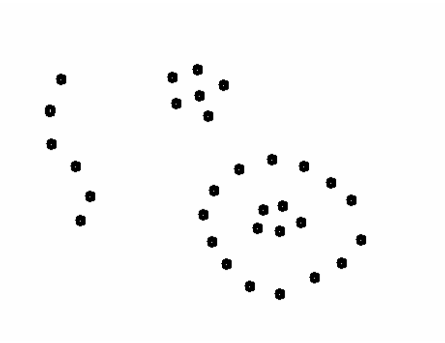

Machine Learning

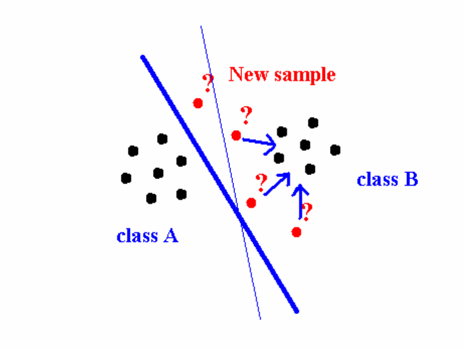
Unsupervised

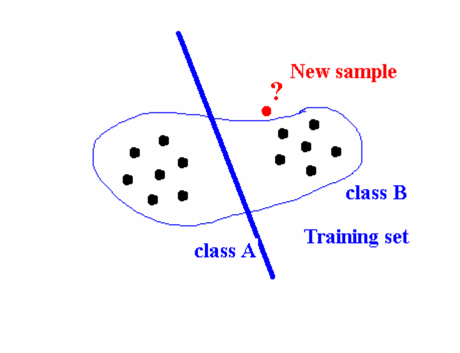
Reinforced

Supervised

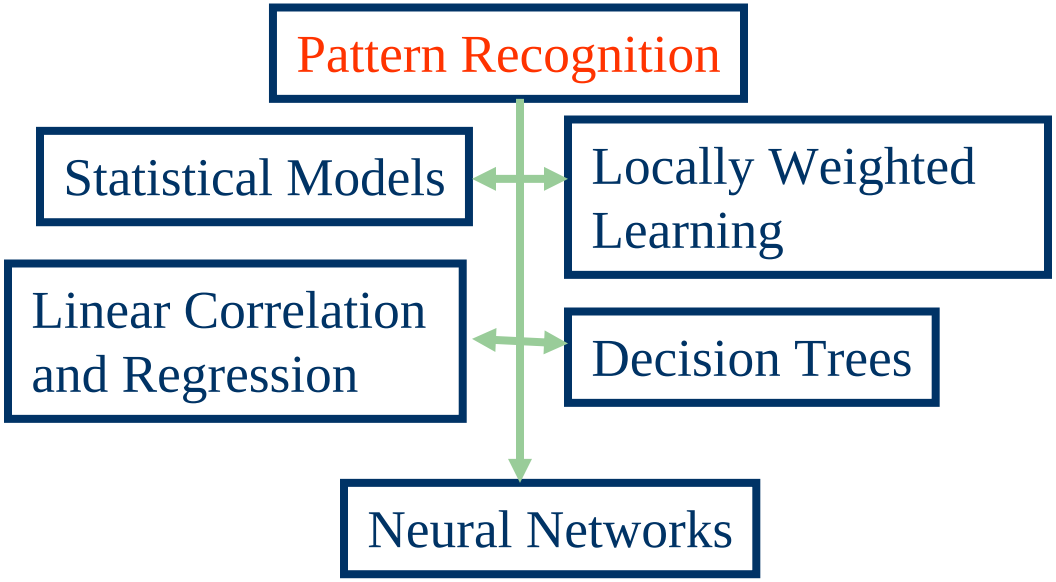
“Natural groupings”







Pattern Recognition



Pattern Recognition

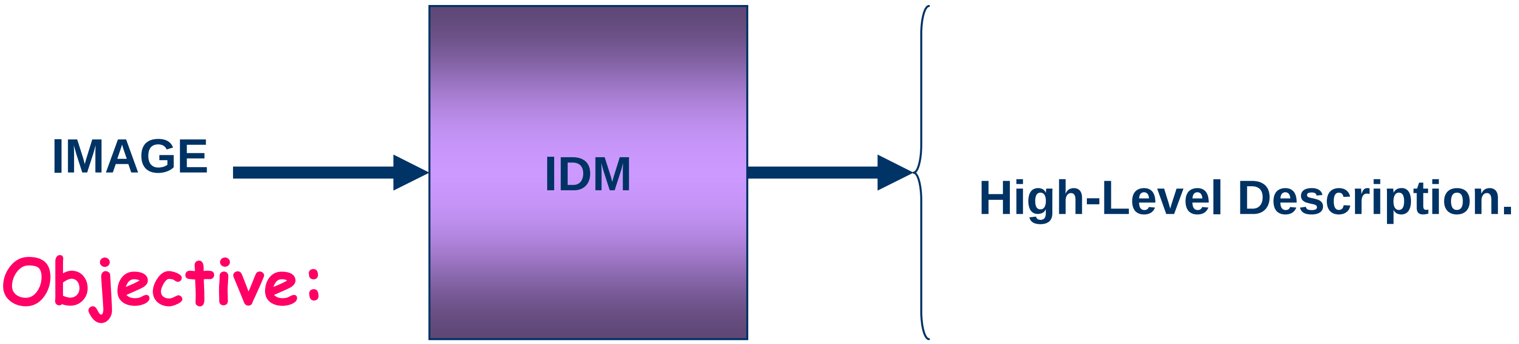
Statistical Models

Locally Weighted Learning

Linear Correlation and Regression

Decision Trees

Neural Networks



Objective:

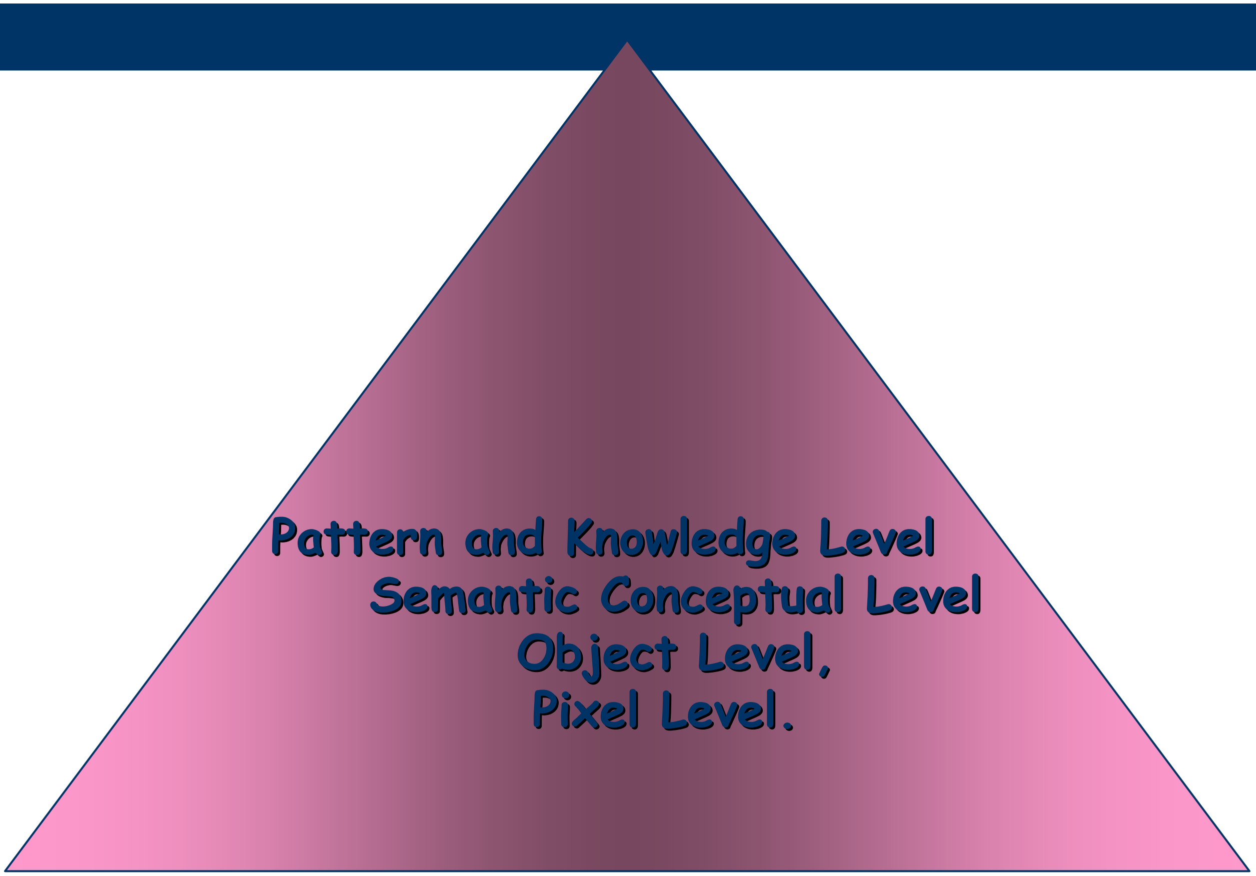
- automatically extract semantically meaningful information (knowledge) from images.

Challenges

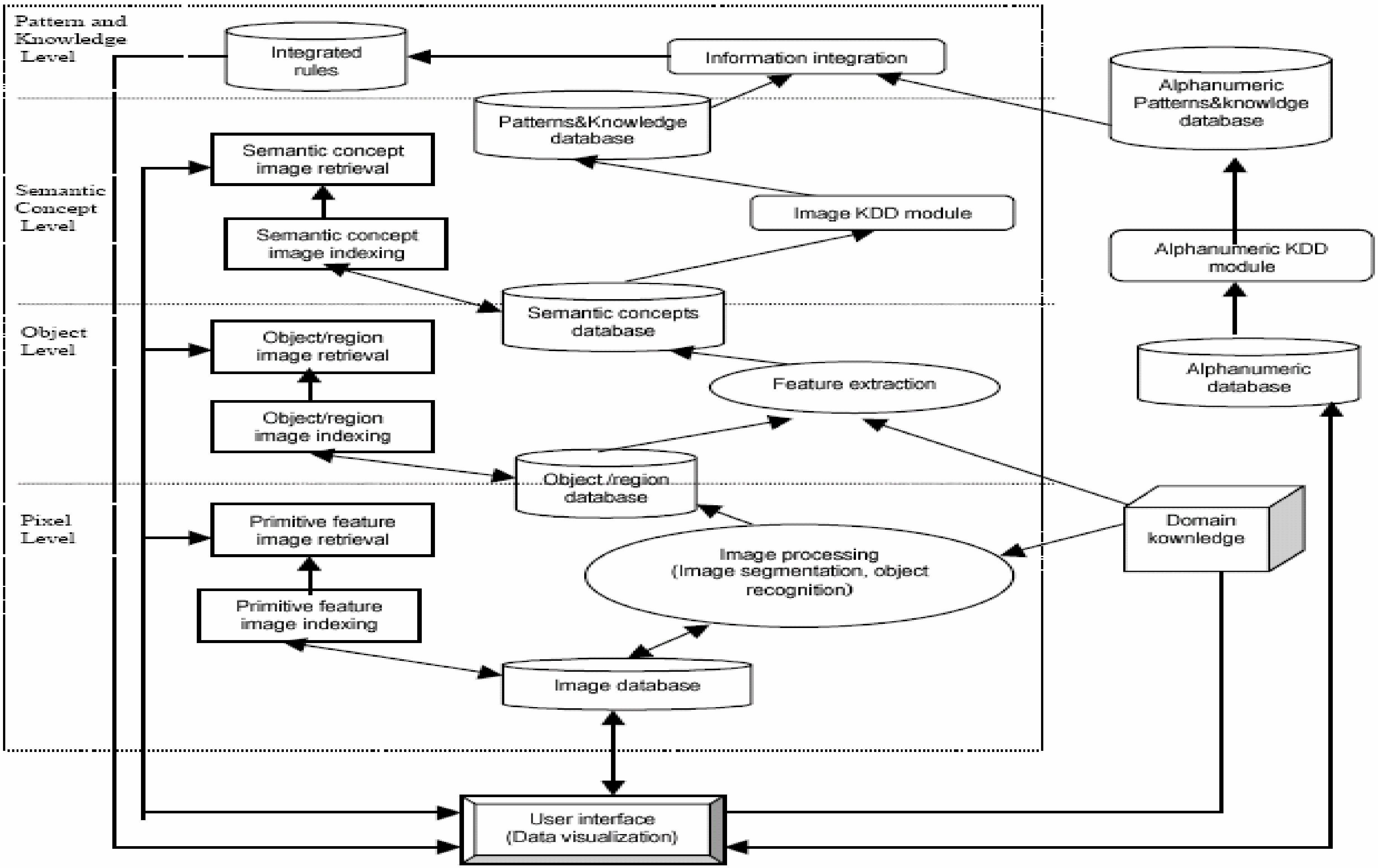
- Determine how low-level, pixel representation contained in a raw image or image sequence can be processed to identify high-level spatial objects and relationships.
- Unlike traditional data mining, considerable complexity is associated with image data mining mainly due to the difficulty of having a proper representation of image information.

University of Waterloo

IDM from Images
Levels of Abstraction



Pattern and Knowledge Level
Semantic Conceptual Level
Object Level,
Pixel Level.



Pattern and Knowledge Level

Semantic Concept Level

Object Level

Pixel Level

Integrated rules

Information integration

Alphanumeric Patterns&knowldge database

Patterns&Knowledge database

Semantic concept image retrieval

Semantic concept image indexing

Image KDD module

Alphanumeric KDD module

Semantic concepts database

Object/region image retrieval

Object/region image indexing

Feature extraction

Alphanumeric database

Object /region database

Primitive feature image retrieval

Primitive feature image indexing

Image processing (Image segmentation, object recognition)

Domain kownledge

Image database

User interface (Data visualization)

Steps

- Image Pre-processing
- Feature Extraction
- Feature Selection
- Image Understanding
- Pattern Discovery
- Pattern Association
- Knowledge Rep.
- Knowledge Discovery
- Ontological Rep of VO
- Testing

From image data to knowledge

Discover patterns in, from, and about images to:

- •Be able to retrieve them on content and utilize their contents to Optimize decisions.

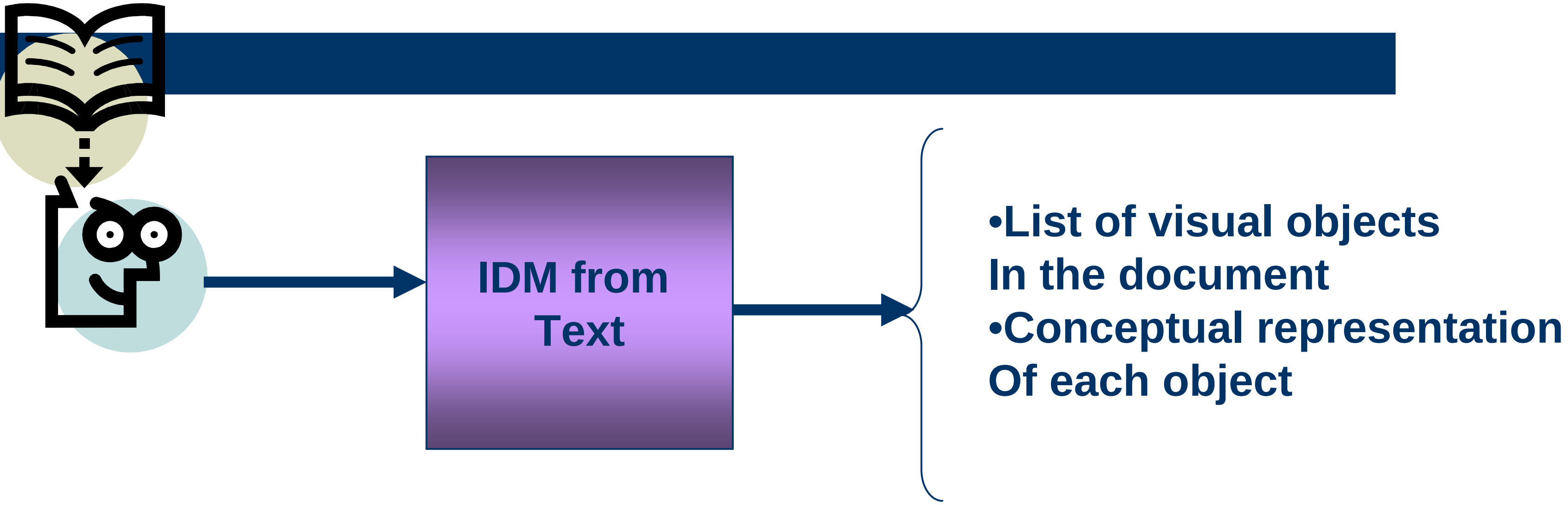
Approach:

- •Develop automatic pattern recognition techniques to identify and characterize patterns in image.

In doing so we ask:
Is there other sources of information pertaining to the content of the image that can help in gaining efficient and effective insight to image?

University of Waterloo

IDM from Text



IDM from Text

•List of visual objects
In the document
•Conceptual representation
Of each object

- To extract semantically meaningful information (knowledge) about images from text (body, caption, etc.)

Steps:

Association image & text

- Knowledge-from-text Rep
- Knowledge Extraction
- Knowledge Discovery
- Conceptual-Level matching of visual objects

Semantics and Images

- **Ontologies**
- The source of semantic interoperability problems is *semantic heterogeneity,* i.e., the different conceptualizations underlying representations of real world phenomena.
- *Ontologies* provide the means to capture and communicate such conceptualizations.
- With ontology we try to determine the various types and categories of objects and relations in all realms of being.
- From an information systems and artificial intelligence perspective, *ontologies* are content theories, identifying classes of objects and relations that exist in an area addressed by an information system

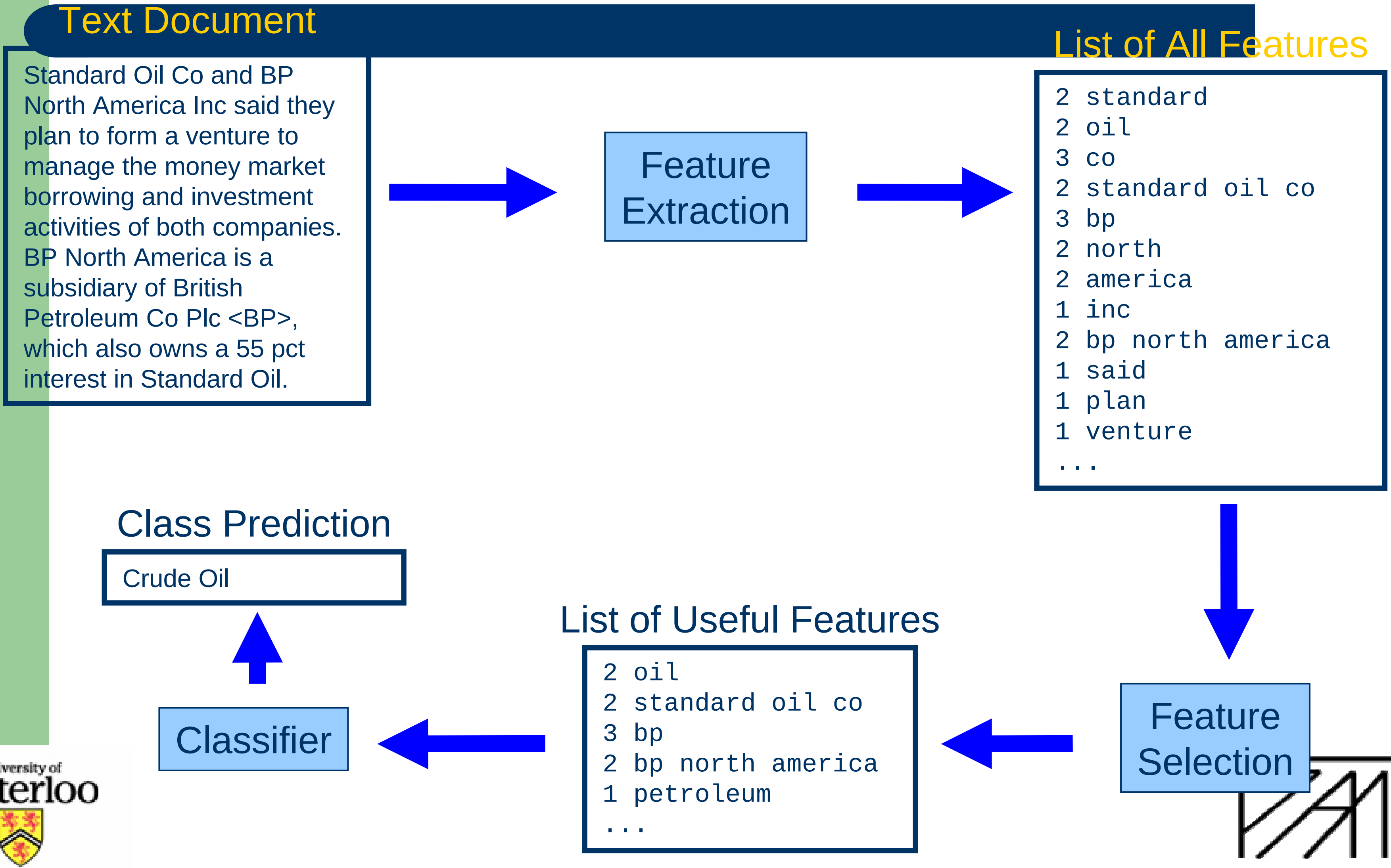
Ontology based approach

- *Semantic parsing:* After finding a focus of interest, an image, the semantic ontology model together with image instance data can be used in finding out relations between the selected image and other images in the collection.

Phrase Based Clustering

- Feature extraction
- Feature selection
- Classification and Clustering
- Use of phrases

Text Classification



Text Document

Standard Oil Co and BP North America Inc said they plan to form a venture to manage the money market borrowing and investment activities of both companies. BP North America is a subsidiary of British Petroleum Co Plc <BP>, which also owns a 55 pct interest in Standard Oil.

Feature Extraction

List of All Features

```
2 standard
2 oil
3 co
2 standard oil co
3 bp
2 north
2 america
1 inc
2 bp north america
1 said
1 plan
1 venture
...
```

Feature Selection

List of Useful Features

```
2 oil
2 standard oil co
3 bp
2 bp north america
1 petroleum
...
```

Classifier

Class Prediction

Crude Oil

Feature Extraction

- Each unique word or phrase in the document is considered a feature
- The result is a list of all words in the document
- Extraction Steps:
  - Turn all letters to lower case
  - Break words up by white space (space, tab, newline)
  - Remove punctuation marks
  - Optionally combine some adjacent words into phrases

Feature Selection

- The objective is to select features that are useful for the decision about the class
- Popular Techniques
  - Information Gain
    - Base on information theory, selects features with the most information about the classes
  - Chi Squared
    - Use chi squared test to measure statistical significance between the feature and the class
  - Correlation Feature Selection
    - Measures the correlation between the set of features and the class

Classification

- The objective is to select features that are useful for the decision about the class
- Popular Techniques
  - K Nearest Neighbour
    - Finds the K nearest documents and makes a decision based on these.
  - Naïve Bayes
    - Builds probability of each class and selects the one with the highest probability (the features are considered independent)
  - Support Vector Machine
    - Constructs a decision boundary in a multidimensional space that separates the classes.

Phrase-based Features

- Phrases: more informative features than individual words → local context matching
- Represent sentences rather than words
- Facilitate phrase-matching between documents
- Achieves accurate document pair-wise similarity
- Avoid high-dimensionality of vector space model

Phrase Extraction Model

- Merge a bigram that tends to co-occur together and replace it with a new symbol
- Co-occurrence is measured with Mutual Information
- Store the merge as a rule in a table
- Repeat until no more bigrams can be merged

Weight	Bigram	Symbol	Expanded phrase
7.68	<new, york>	w_1	`new york`
5.26	<stock, exchange>	w_2	`stock exchange`
8.51	< w_1, w_2>	w_3	`new york stock exchange`

University of Waterloo

Example

Initial sequence

trading	on	the	new	york	stock	exchange	closed

Calculate the Mutual Information for all the bigrams

Combine <new, york>

trading	on	the	new york	stock	exchange	closed

Calculate the Mutual Information for all the bigrams

Combine <stock, exchange>

trading	on	the	new york	stock exchange	closed

Calculate the Mutual Information for all the bigrams

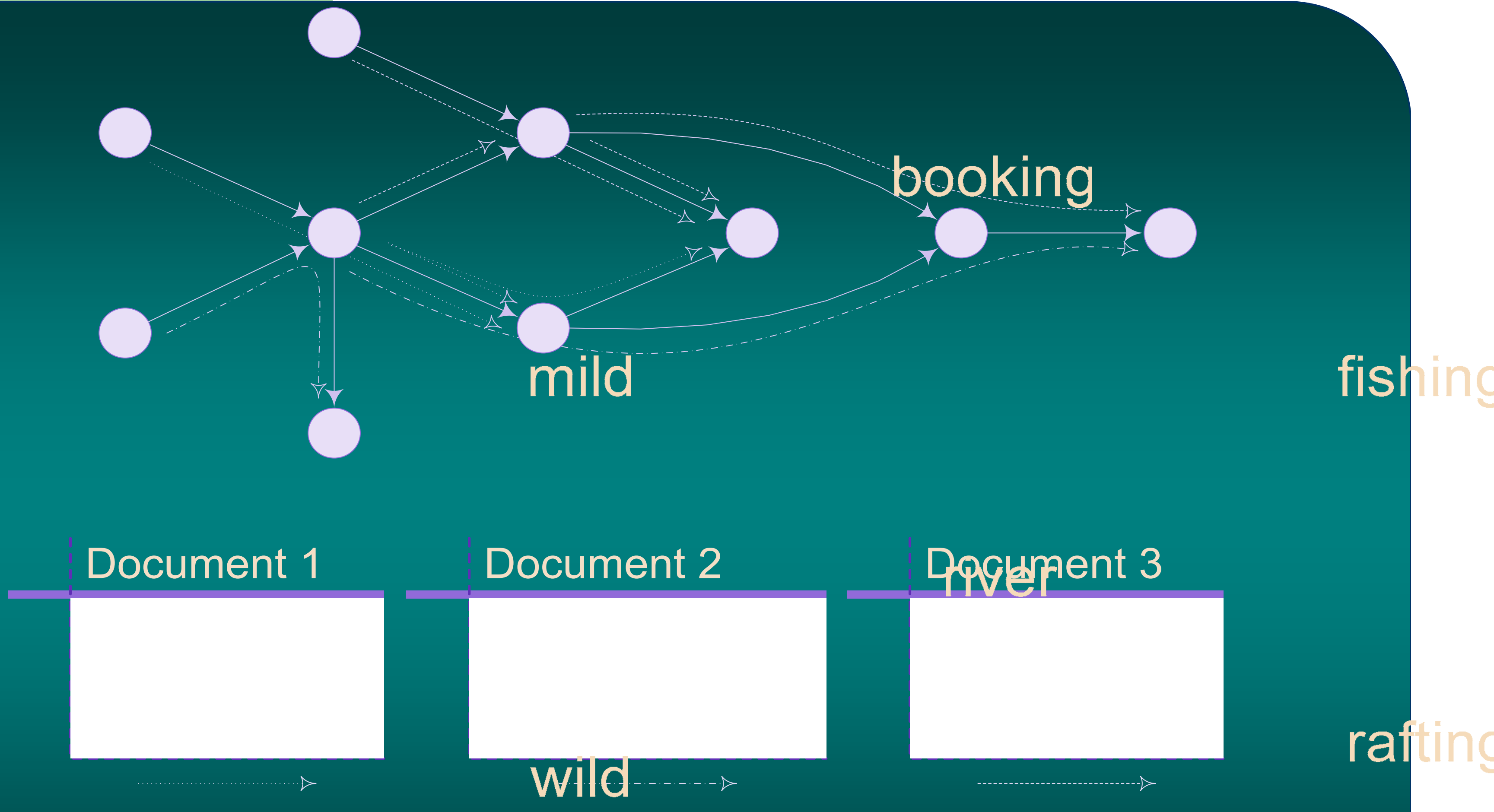
Combine <new york, stock exchange>

trading	on	the	new york stock exchange	closed

Phrase-based Document Indexing

- **Document Index Graph Structure**
  - A model based on a *digraph* representation of the phrases in the document set
  - Nodes correspond to unique terms
  - Edges maintain phrase representation
  - A phrase is a path in the graph
  - The model is an inverted list (terms → documents)
  - Nodes carry term weight information for each document in which they appear
  - Shared phrases can be matched efficiently

Document Index Graph



adventures

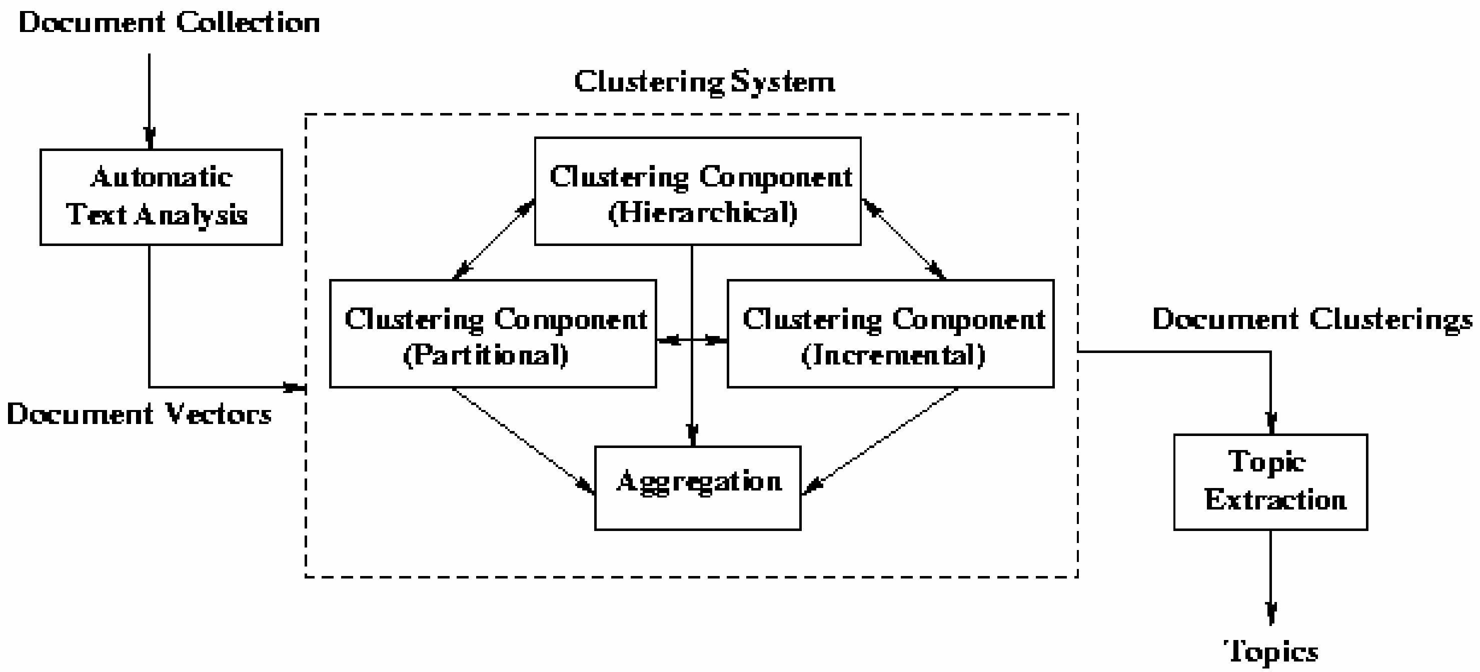
Pattern Discovery using Clustering Aggregation

- Cluster Ensemble
- Clusters of Meta Objects
  - Multiple distributed taxonomies
  - Info. Sources: different sets of meta data, different re-use/assembling scenarios
  - Base clusters are based on partial views

University of Waterloo

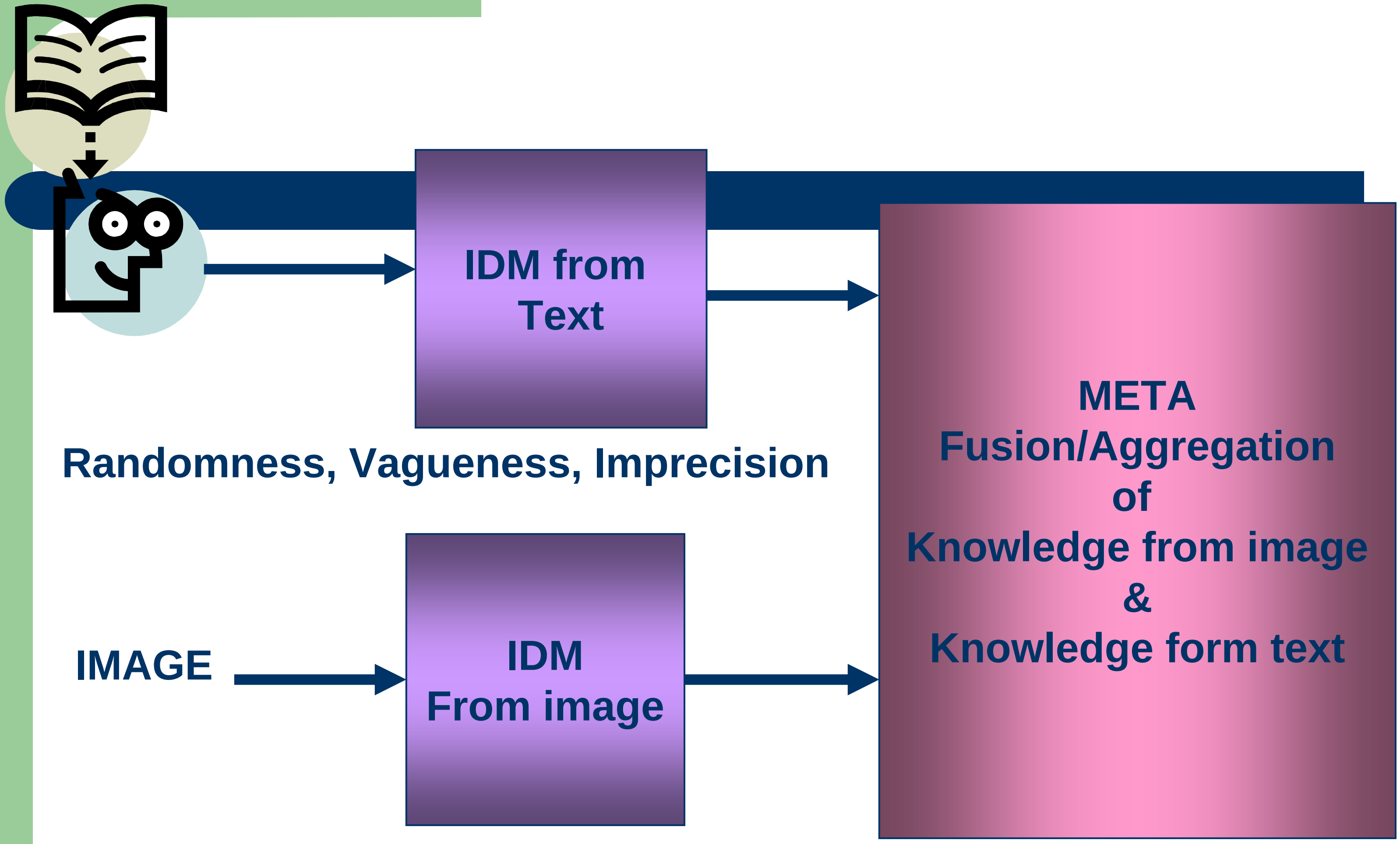
- Combining of multiple clustering
  - Mining complex web of relationships
  - Innovative machine learning techniques for integrating multiple objects clustering
  - Discovery of combined multi-view clusters.

Clustering Aggregation



Stages for Combining Clusterings

- Distributed clustering of the objects
- Measure co-associations between objects based on their co-clustering - Voting
- Development of combination rules based on shared co-associations – Shared Nearest Neighbours (binary votes, weighted votes, and other approaches are being developed)



IDM from Text

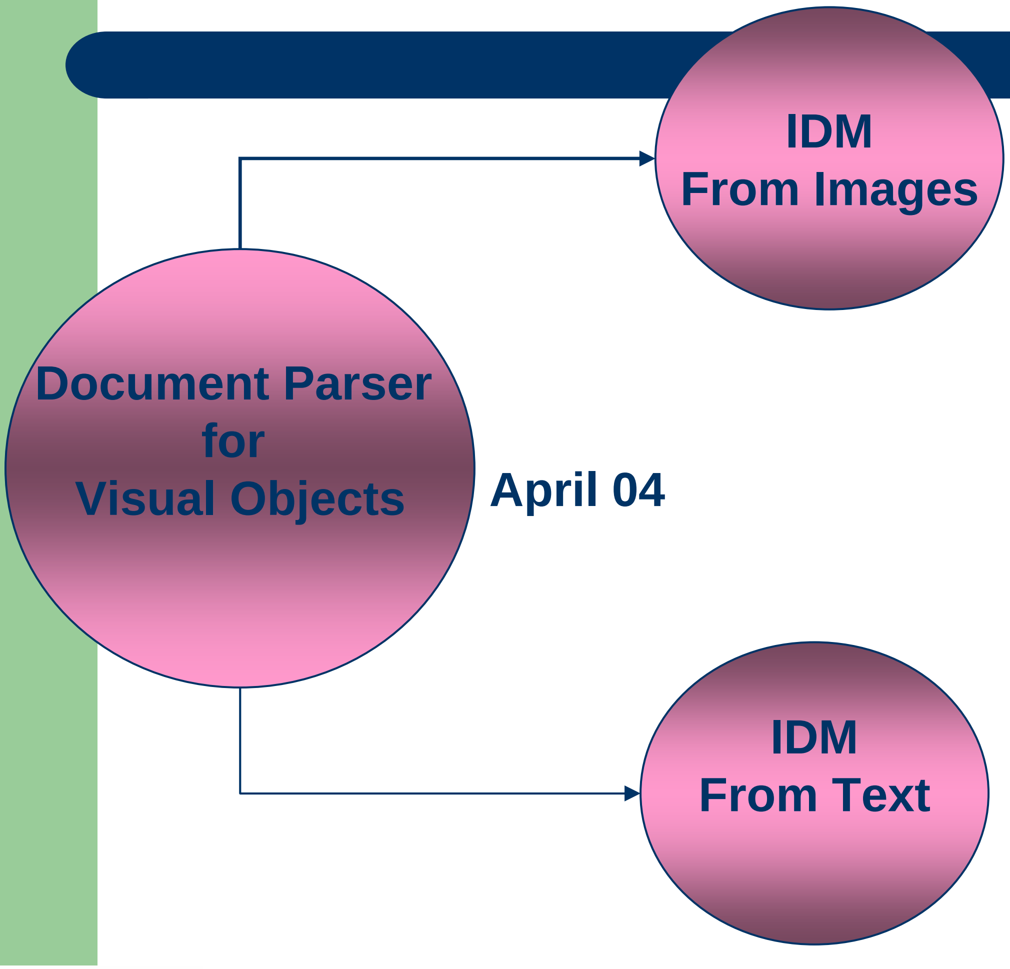
Randomness, Vagueness, Imprecision

IMAGE

IDM From image

META Fusion/Aggregation of Knowledge from image & Knowledge form text

Strategies:



Document Parser for Visual Objects

IDM From Images

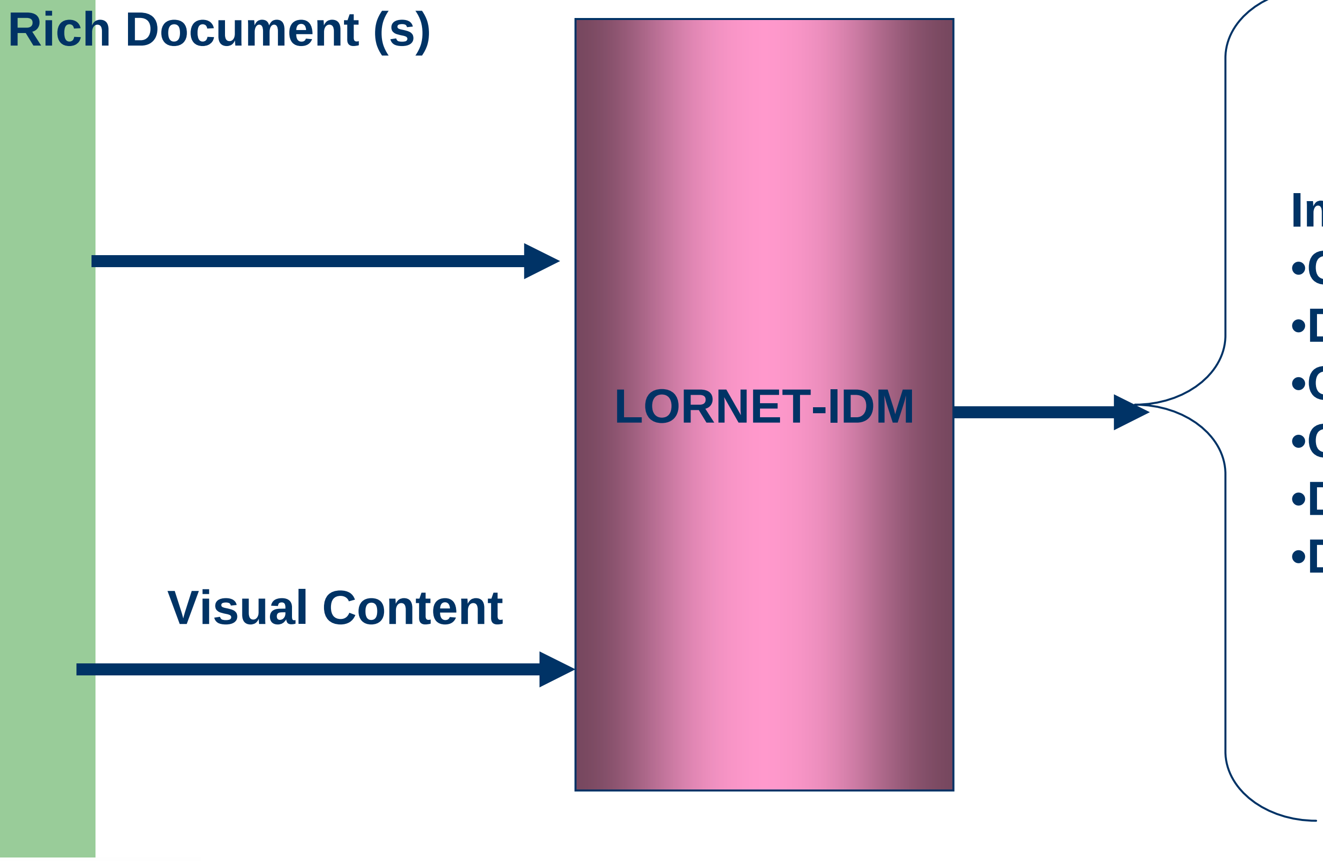
IDM From Text

April 04

- Image Pre-processing
- Feature Extraction
- Feature Selection
- Image Understanding
- Pattern Discovery
- Pattern Association
- Knowledge Rep.
- Knowledge Discovery
- Ontological Rep of VO
- Testing

Association image & text

- Knowledge-from-text Rep
- Knowledge Extraction
- Knowledge Discovery
- Conceptual-Level matching of visual objects



Rich Document (s)

Visual Content

LORNET-IDM

Image Derived:
- •Conceptual Rep.
- •Document Similarity
- •Contextual Cues
- •Conceptual Indexing
- •Document Summary
- •Document Clustering

Conclusion

- Critical need for scalable, effective and efficient algorithms to discover knowledge in images
- To explore contents of images fully and automatically we have to utilize other sources about them
- Ontologies based techniques are promising to include non-visual and visual context.