



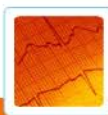
# Knowledge Discovery Mechanisms applied on Integrated Databases

## *Health-e-Child*

Yannis Ioannidis, University of Athens

010101  
101010  
110101





## Overview

- What is a Data Base (DB)?
- Knowledge Discovery (KD) in Data Bases
  - Application areas
  - Data Mining and Knowledge Discovery Process
- What is the Health-e-Child Project (HeC)?
  - HeC Objectives
    - Integrating Different Types of Data
    - Applying Knowledge Discovery Processes
    - Grid of Computers
- Knowledge Discovery on Integrated Data Bases and HeC
  - Issues & Challenges
  - Integrated Data Model
  - Killer Applications in HeC project
- Take Away



# What is a Data Base (DB)...

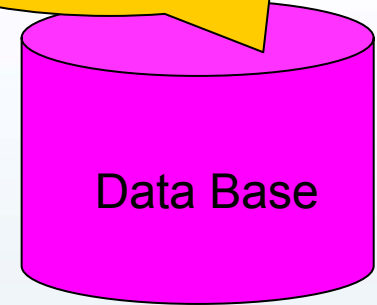
Definitions



# Data Base

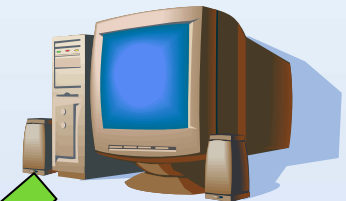
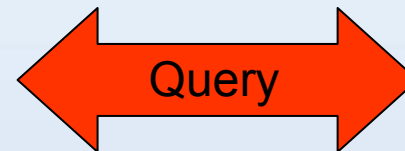
- **Data Base:** A structured collection of records or data that is stored in a computer system
- **Purpose:** A computer program or person can consult it to answer queries
- **How:** Using a query language!!
- **Why:** The records retrieved as answers to queries represent information that can be used to make decisions

Restaurants in Genoa: Location | Type of Food | Phone

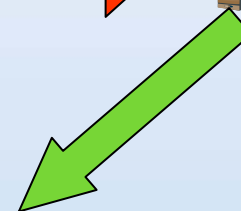


Example:

I am hungry...  
I want to eat Italian food..  
Where is the nearest restaurant?



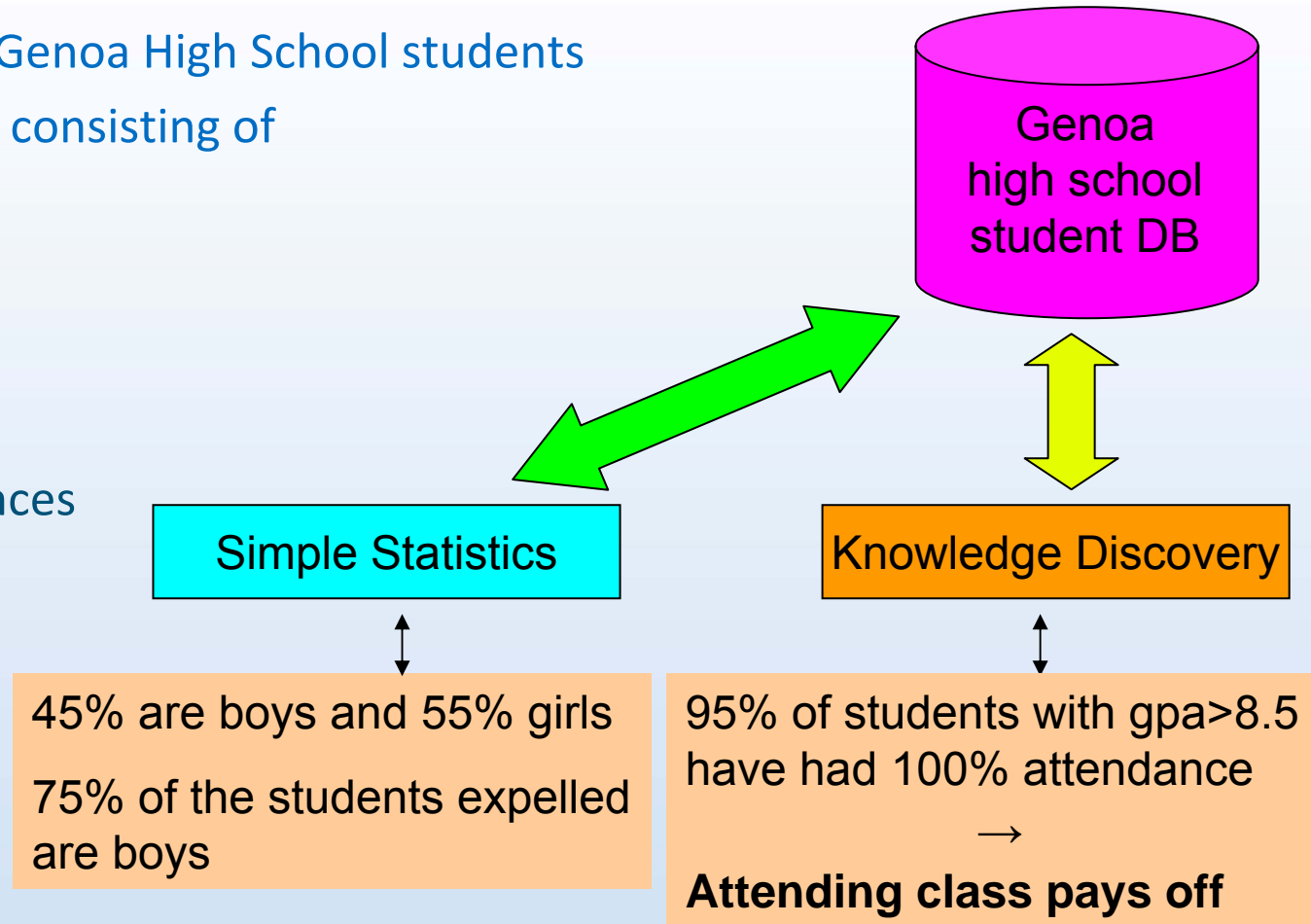
**Answer:** "There is a trattoria down the street..."





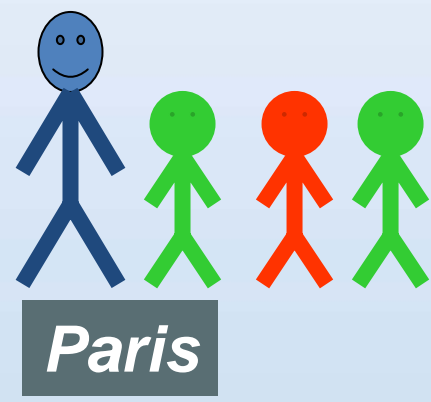
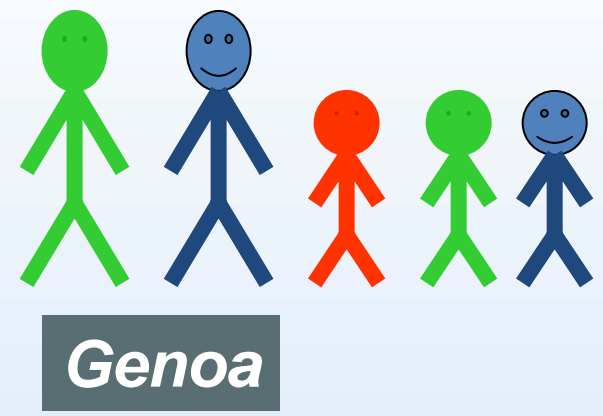
# Knowledge Discovery in Databases

- Database of Genoa High School students
- Data records consisting of
  - Age
  - Gender
  - Grades
  - Courses
  - Attendances
  - Behavior
  - ...



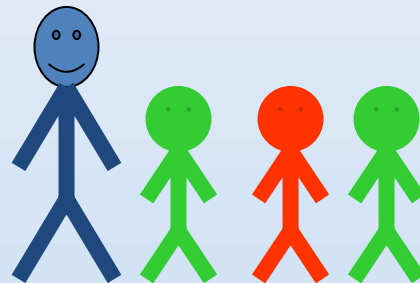
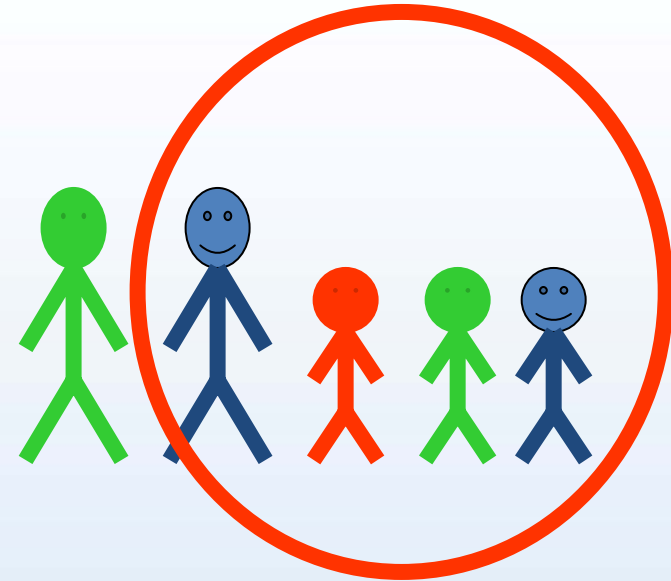
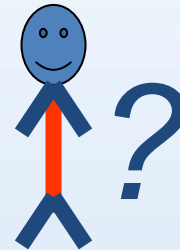
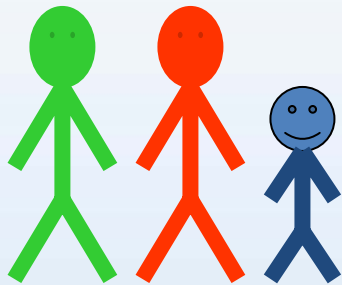


# Knowledge Discovery in Medical Databases





# Knowledge Discovery in Medical Databases





# Knowledge Discovery includes

## Machine Learning



## Visualization



### Data Mining and Knowledge Discovery

## Statistics



## Databases





## Knowledge Discovery Application Areas

- Science
  - astronomy, bioinformatics, drug discovery, ...
- Business
  - CRM (Customer Relationship management), fraud detection, eCommerce, manufacturing, sports/entertainment, telecoms, targeted marketing, ...
- Web
  - search engines, advertising, web and text mining, ...
- Government
  - surveillance (?!), crime detection, profiling tax cheaters, ...
- Health Care
  - pattern discovery in medical images, analysis of microarray (gene-chip) experimental data to relate to diseases, analyzing side-effects of drugs, and effectiveness of treatments; optimization of processes within the hospital, ...

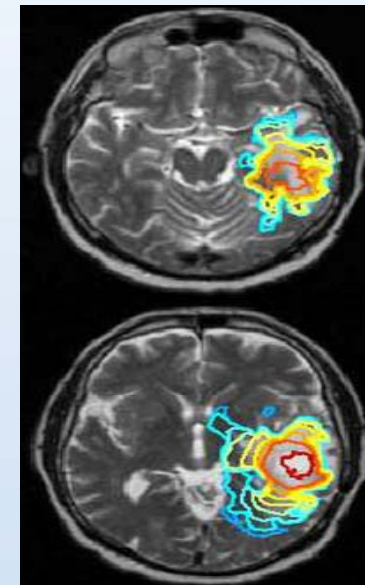
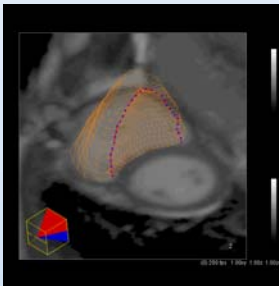
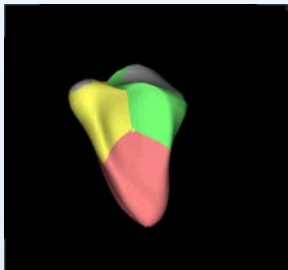


# What is Health-e-Child...

Project description

## Health-e-Child (HeC) Project

- A project for children.
- Three Paediatric Diseases will be studied
  - Heart diseases (Right Ventricular Overload, Cardiomyopathy)
  - Inflammatory diseases (Juvenile Idiopathic Arthritis)
  - Brain tumours (Gliomas)





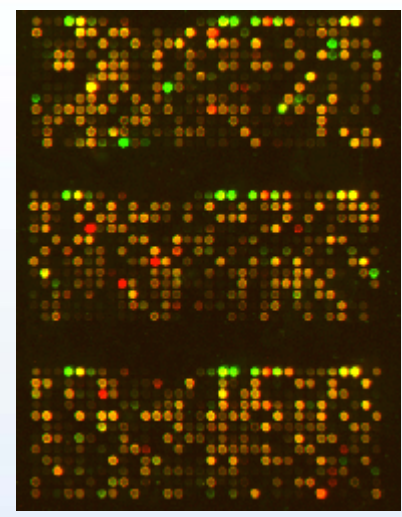
## Health-e-Child Objectives

- Create applications and services that will perform **knowledge discovery** in order to:
  - Help the doctor with the diagnosis of the disease
  - Improve the quality of care for the patients
  - Reduce the cost of treatment
  - Derive disease models
  - Present a Decision Support System helping the selection of treatment, medication, etc.
- Very nice!
- **How** are you going to do that?

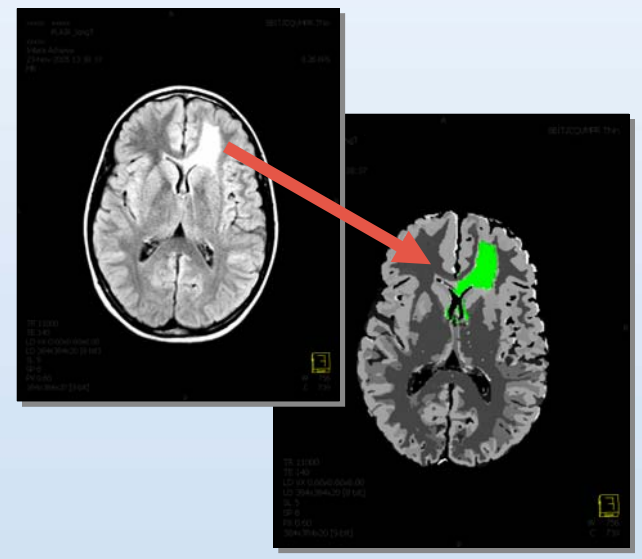


# By Integrating Different Types of Data

- Genomic and Molecular Data
  - Microarray technology
- Imaging Modalities (MRI, XRay, CT, US)
- Clinical Data
  - Forms of health data
  - Patient Records - Electronic Medical Records (EMR)

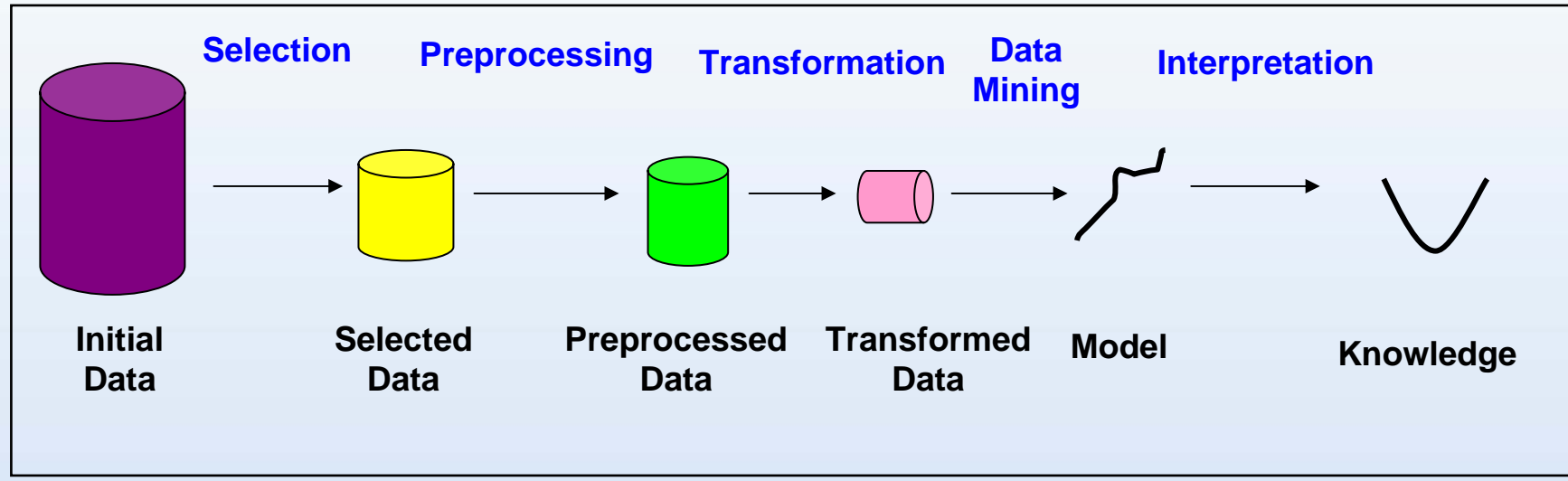


The image shows several overlapping screenshots of medical forms. The top-most form is titled 'DIAGNOSIS COLLECTING FORM-I' and includes fields for 'VisitNo:' and 'VisitDate:'. Below it are forms labeled 'DIAGNOSIS COLLECTING FORM-II', 'DIAGNOSIS COLLECTING FORM-III', and 'DIAGNOSIS COLLECTING FORM-IV'. At the bottom, a 'HISTOPATHOLOGY' form is visible. In the center, a 'BRAIN TUMOUR STUDY' interface is shown with a grid of buttons: 'Baseline Information', 'Enter Patient ID', 'Enter Visit', 'Diagnosis', 'Surgery', 'Histopathology', 'Follow-up Data', 'Observation Study', 'Chemotherapy', 'Radiotherapy', 'Diagnosis Collecting Form I-IV', and 'Relapse/Progression'.





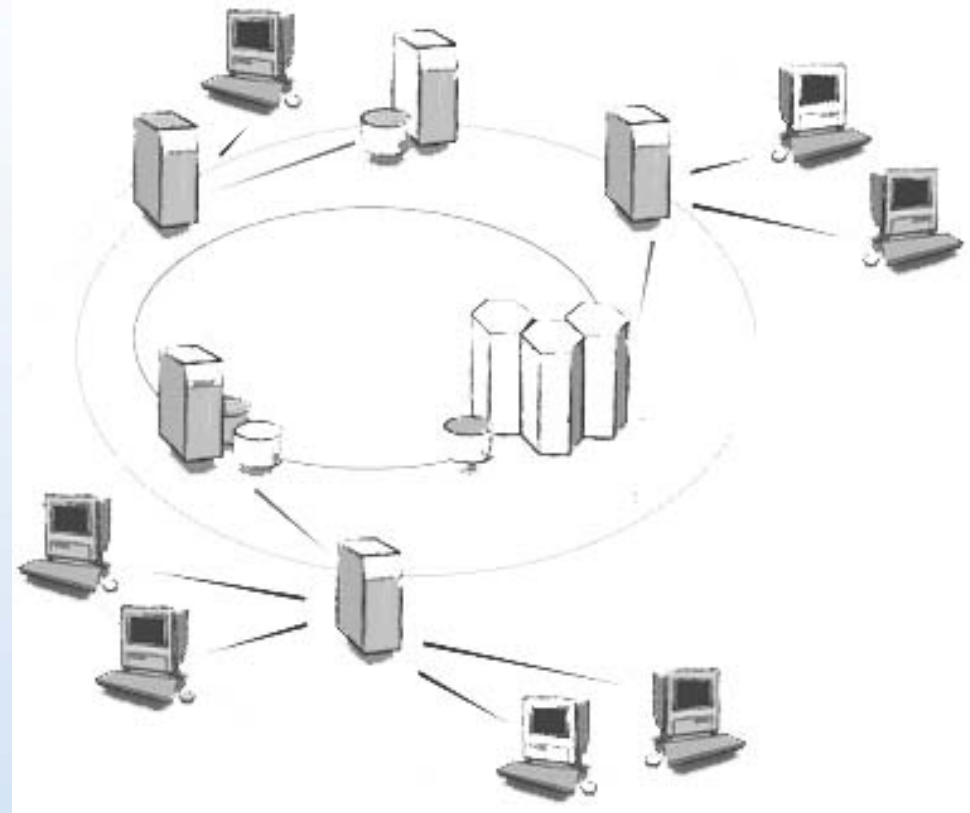
# By Applying Knowledge Discovery Processes

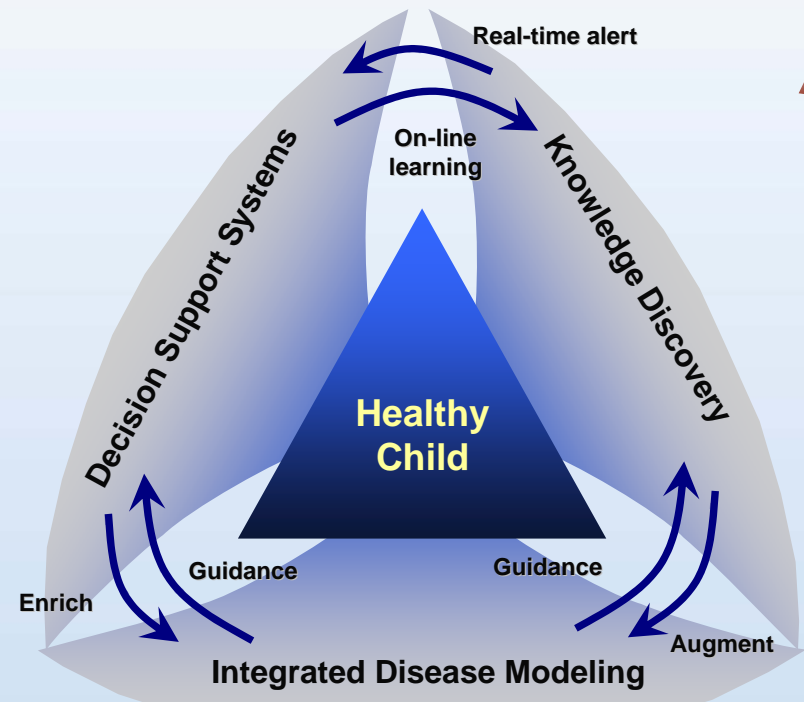
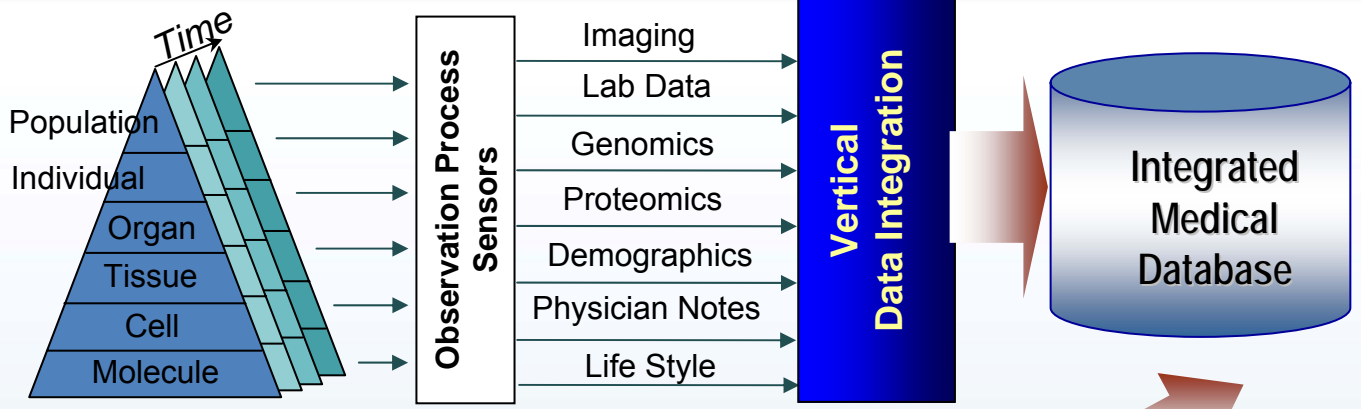


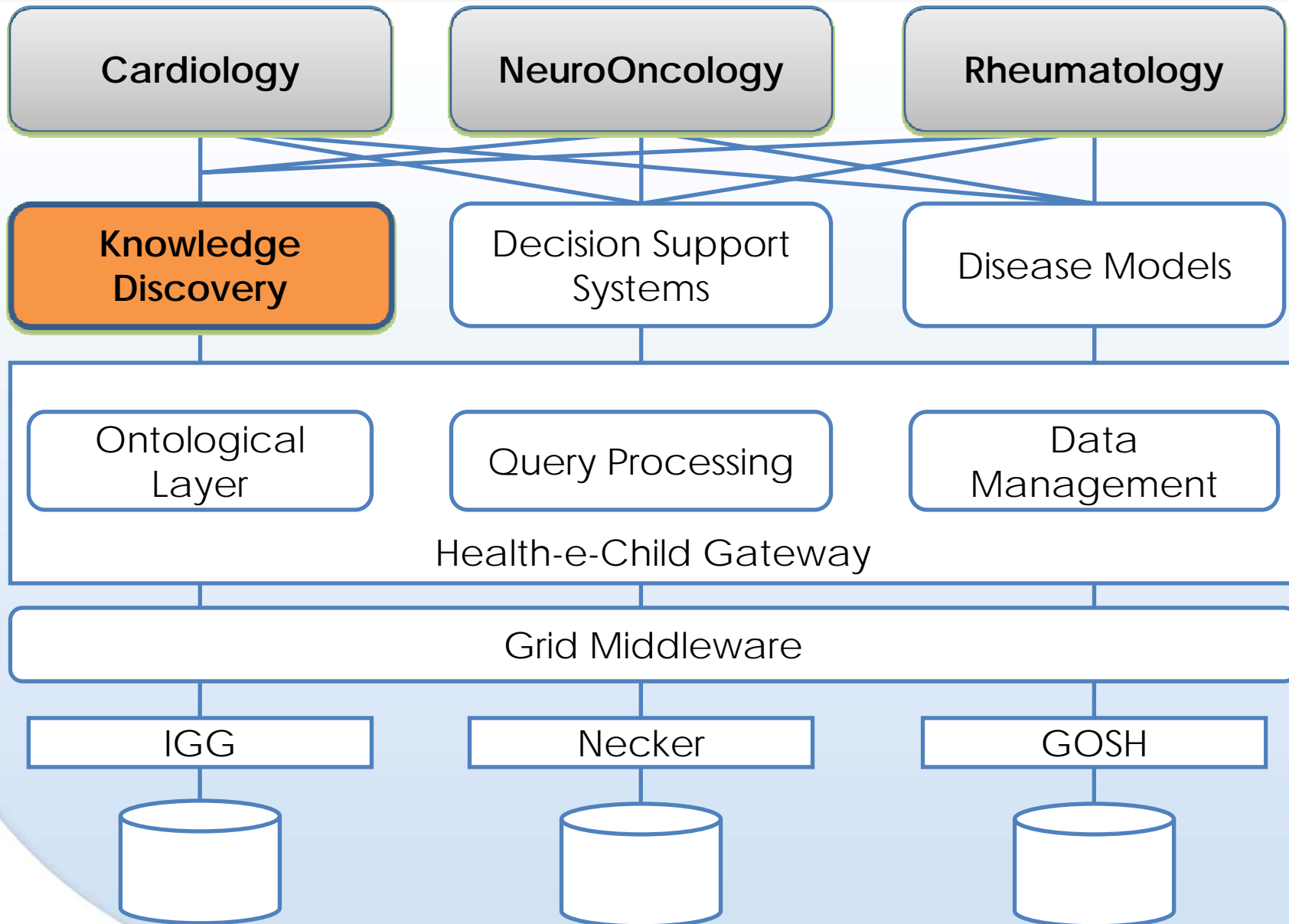


## Using a Grid of Computers

- Goals:
  - Link/Connect different organizations
  - Allow doctors to access all hospitals → databases
  - Ease computational complexity of such applications.









# Knowledge Discovery on Integrated Data Bases and Health-e-Child...

...

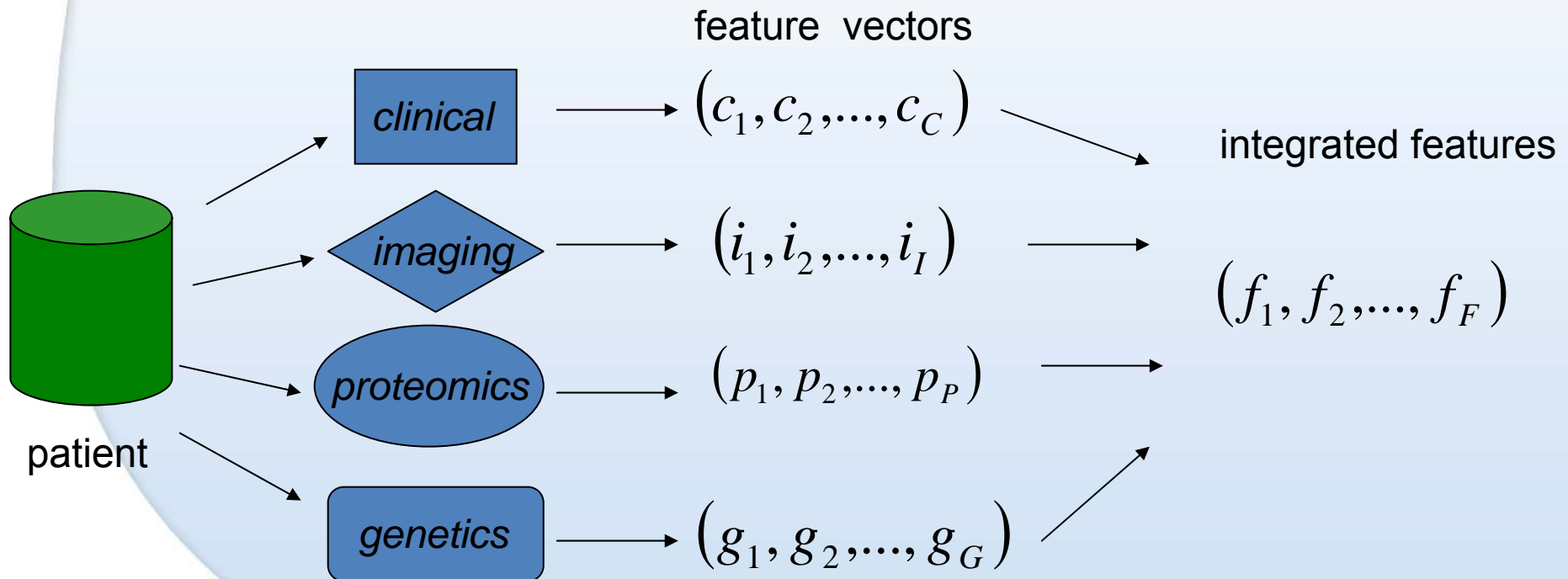


## Issues and Challenges

- Unique features of biomedical data
  - Heterogeneous, distributed, and complex data
  - High dimensionality (especially in Genomics/Proteomics)
  - Lack of standardisation
  - Changing data and knowledge (Temporal Evolution)
  - Biomedical data not characterized mathematically
  - Missing, inconsistent, and noisy data
  - Ambiguous result interpretation
  - Data ownership, privacy, and security concerns
  - Only few samples available (currently in HeC)

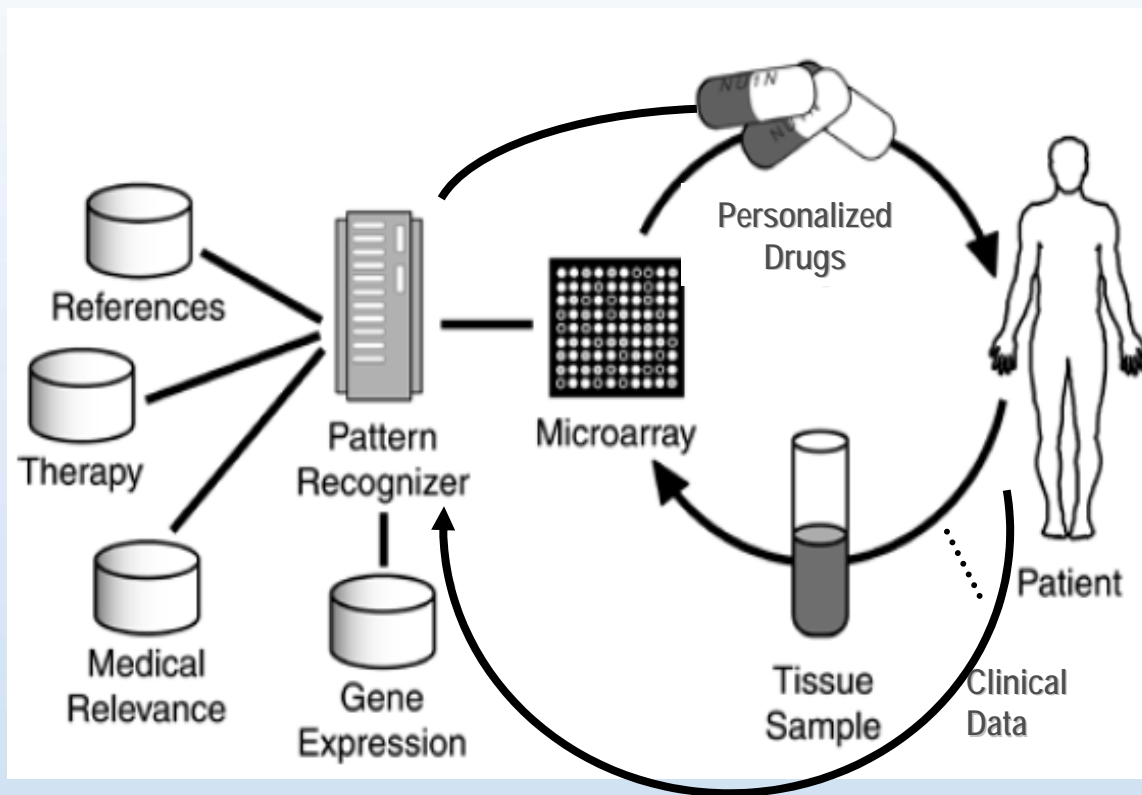
## A simple approach → Using a single heterogeneous vector

- For each patient, measurements from different data sources are combined into a single feature vector (→ heterogeneous data).
  - This is a straightforward solution
  - Many techniques perform well with heterogeneous data



## Ultimate Goals of HeC project

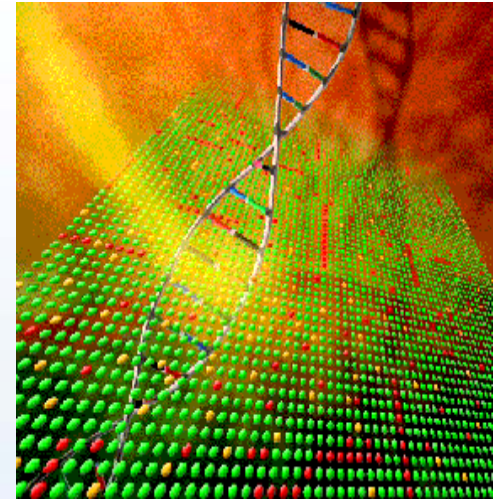
- Discover patient trends in vertically-integrated data
- Personalized medicine (custom, just-in-time delivery of medications tailored to patient's condition)





**Take Away...**

A long but exciting way to go!





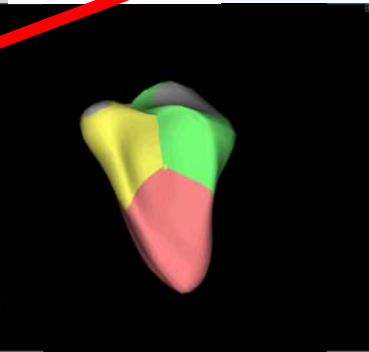
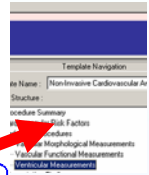
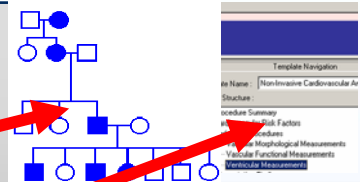
# KDD and Integrated Biomedical Databases

Some Examples in HeC

# Example Disease: Right Ventricular Overload

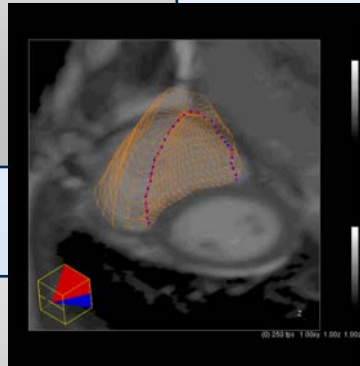
## Clinical Data

- Demographic, history & familial
- Lifestyle
- Clinical notes
- ECG



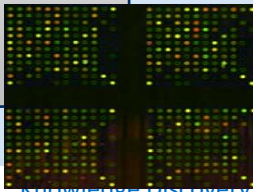
## Imaging Data

- 2D/3D Echo
- Tissue Doppler
- MRI



## Genetic Data

- Karyotyping
- Array-CGH



## Clinical Features

- prolonged PR interval in electrocardiogram
- systolic ejection murmur on auscultation

## Anatomical Features

- Hyperkinetic RV muscle
- Increased RV-LV ratio
- Ventricular septum defect
- Thickening (hypertrophy) of the RV muscle

## Genetic Features

- candidates for gene mutations are e.g. 4p13-q12, 6p21.3, 1p31-p21, 3p25, 6q21-q23.2, 5q34

training and specific patient data

## Decision Support

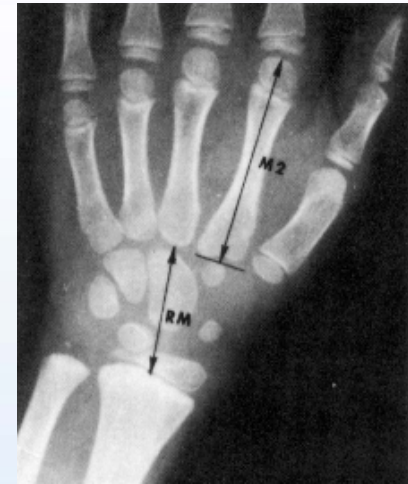
- prediction of type and timing of treatment
- classification of RV overload
- retrieval of similar cases

## Knowledge Discovery

- classification of subtypes
- genotype/phenotype correlation

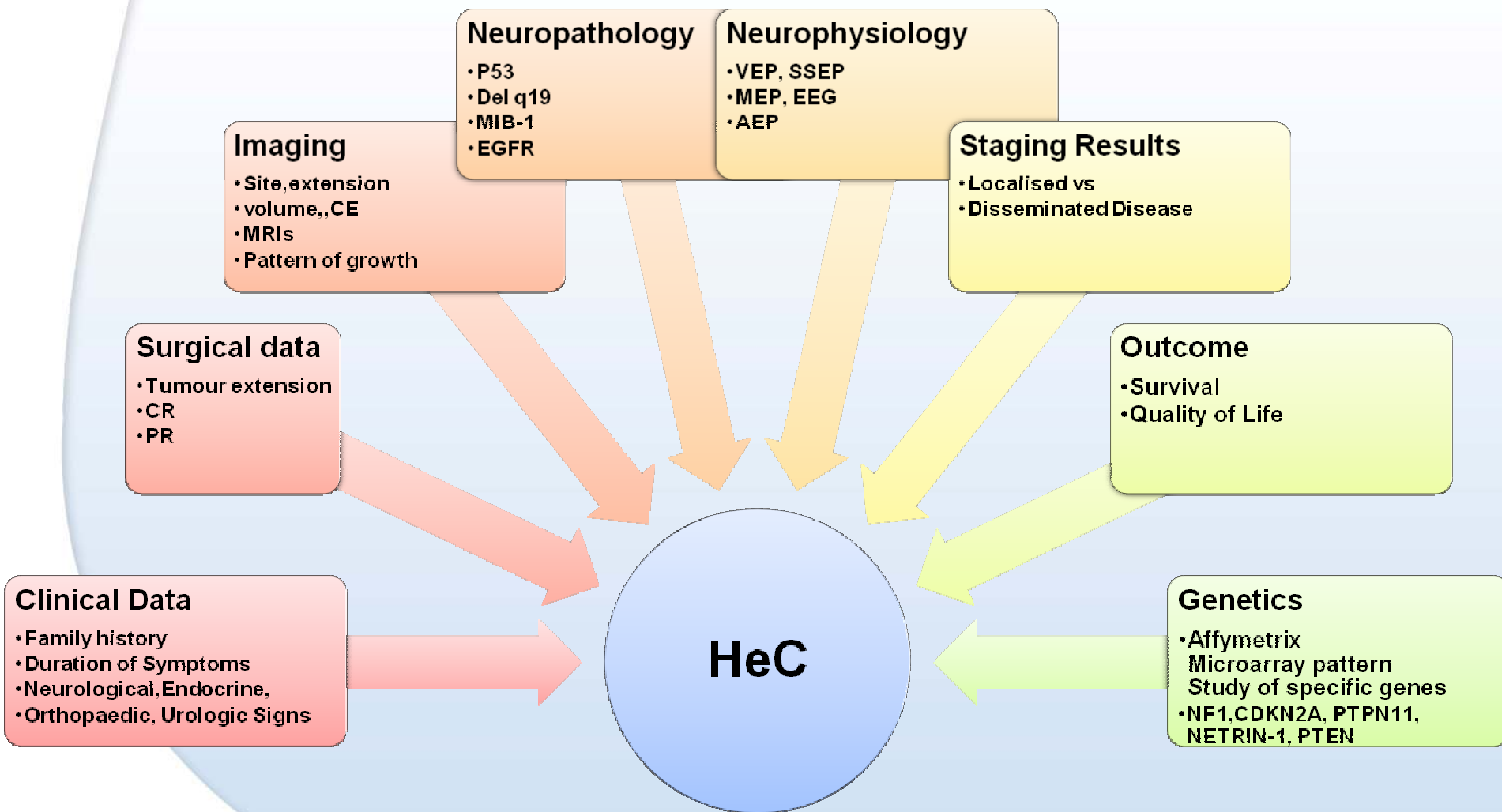
## Knowledge Discovery In Rheumatology

- Research goals
  - Identify correlations between gene combinations and particular diseases
  - Improve current classification of disease subtypes
  - Predict the disease outcome / evolution early
  - Automatically suggest drug prescriptions, e.g., to stop/slow down disease evolution



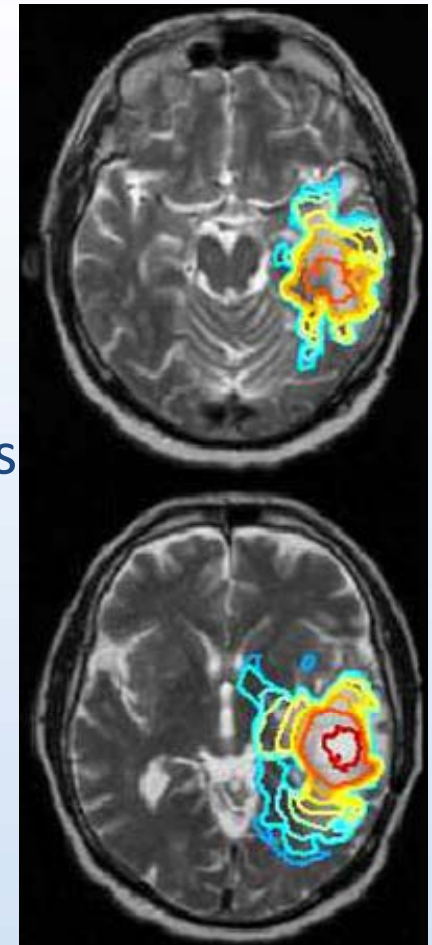


## Example: Data Sources in Brain Tumour Study



## Knowledge Discovery in Brain Tumours

- Research goals
  - Find prognosis and correlate it with tumour origin site
  - Suggest treatment strategy
  - Predict outcome and correlate it with age or genetics or ...
  - Provide more precise classification of diseases
  - Meta-analysis of published information



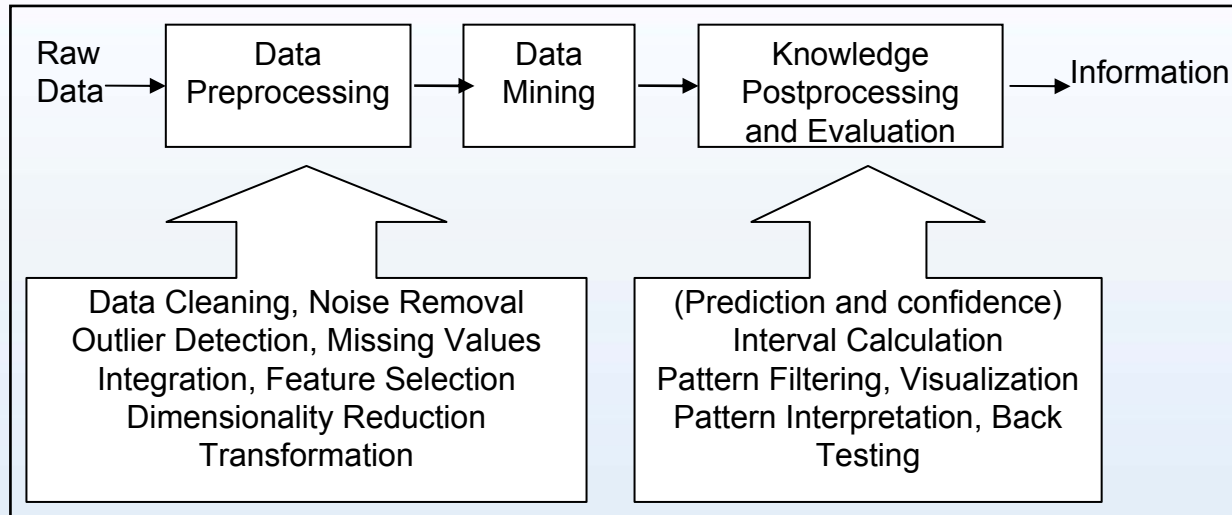


## Issues and Challenges

- Data cleaning, preprocessing, and semantic integration of heterogeneous, distributed biomedical databases over the Grid
- Similarity search and comparison in biomedical data
- Diversity of data mining techniques
  - Association analysis: co-occurring bio-sequences or other correlated patterns
  - Frequent pattern-based cluster analysis
  - Visual data mining
  - Privacy preserving mining
  - Data mining over images
- Quality of Answer – Quality of Service
- Medical knowledge sharing and differential diagnosis

# Data Mining and Knowledge Discovery Process

- Simplified model of KDD (Knowledge Discovery in Databases) process



- Five-step KDD process

