KDD introduction

K. Gibert[1]

[1]Department of Statistics and Operation Research

Knowledge Engineering and Machine Learning group

Universitat Politècnica de Catalunya, Barcelona

Introduction

- Knowledge Society *[United Nations, 2005]*
  - Great need of getting knowledge from
    - **Data**
    - **Organizations**
    - **Natural, industrial or artificial phenomena**
  - Support complex decision making processes
- Enormous quantities of data to analyze
  - Boom Internet late 1990s *[Tim Berners-Lee, 1990], 1995 www free&global*
  - New technologies
  - Exponentially increasing

Web contains twice the written information of whole Humanity R. Baeza, CEDI 2010 Yahoo head of research

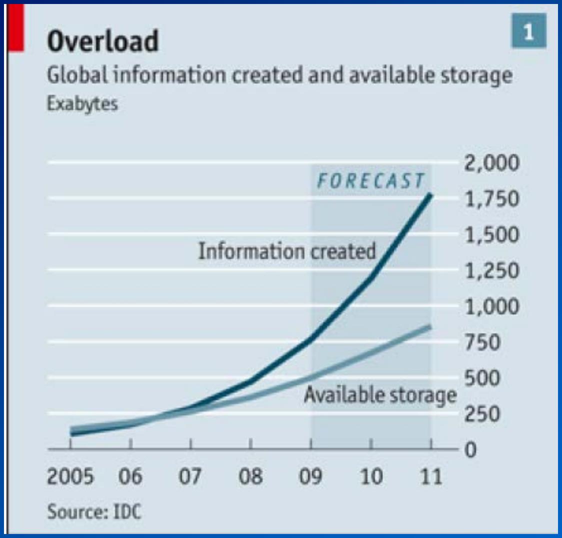
40 zettabytes 2020 (~45 trillion GB)

a 50-fold growth in a decade.

Int'l Data Corporation, USA des 2012

2.5 quintillion daily bytes (90% of it 2011-2013)

IBM report 2013



Source: IDC

K. Gibert

Introduction

- Knowledge Society *[United Nations, 2005]*
  - Great need of getting knowledge from
    - **Data**
    - **Organizations**
    - **Natural, industrial or artificial phenomena**
  - Support complex decision making processes
- Enormous quantities of data to analyze
  - Boom Internet late 1990s *[Tim Berners-Lee, 1990], 1995 www free&global*
  - New technologies
  - Exponentially increasing
- Classical data analysis is poor
  - Too much data
  - Phenomena too complex
- New approaches required

K. Gibert

Data Mining and Knowledge Discovery

- Interdisciplinary problem

  "*Non trivial identifying of valid, novel, potentially useful, ultimately understandable patterns in data*"

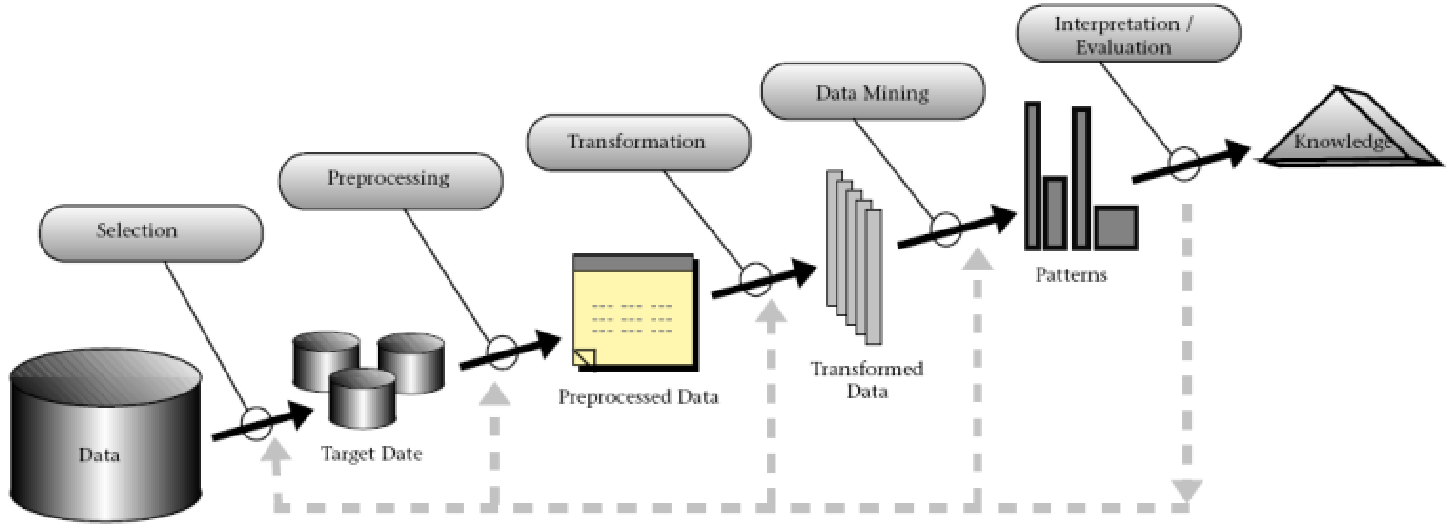
  [Fayyad 96]

  Chief Data Officer & Group Managing Director
  Barclays Bank
  Chairmann Oasis-500
  Yahoo Chief Data Officer&EVP (2004-2008)

- Starting:
  - 1989: First Int'l Workshop on KDD in IJCAI
  - 1994: First proceedings
  - August 1995: First Int'l Conference on KDD *(4000 submissions!!)*
  - 1996: First State of the art (Fayyad et al.)
  - 1997: Data Mining & Knowledge Discovery journal launch

Data Mining and Knowledge Discovery

- Knowledge Discovery System [Fayy96]:

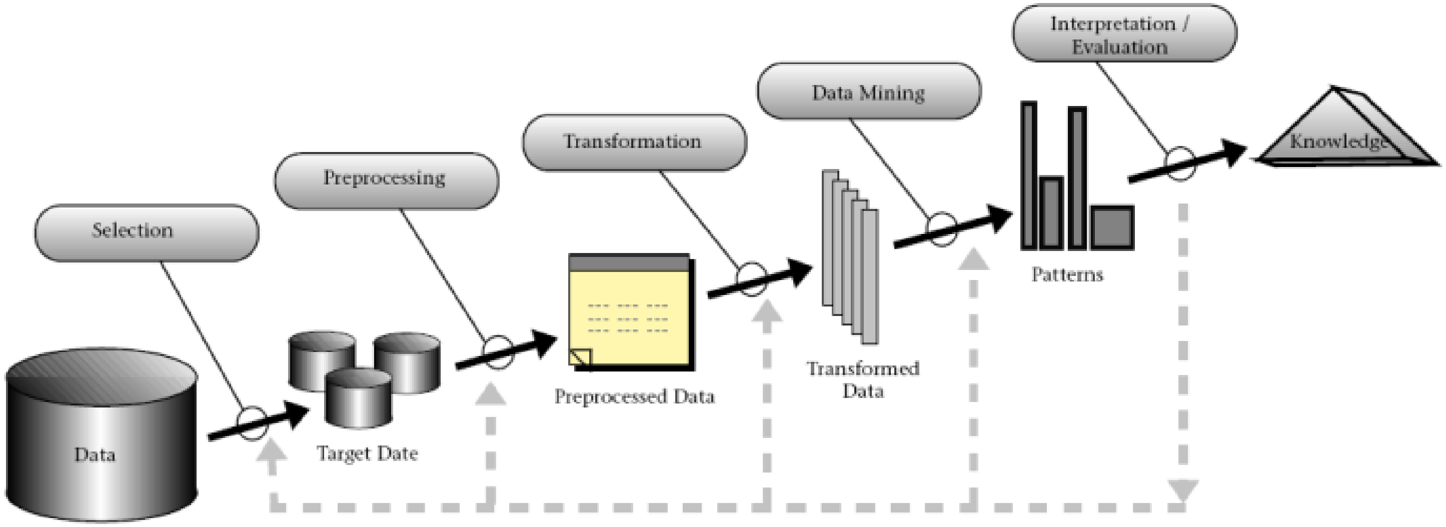


- Problem definition
- Data collection
- Data cleaning and preprocessing
- Dimensionality reduction
- DM technique choice
- Data mining
- Interpretation and discovered knowledge production

Terminological ambiguity
Data Mining vs KDD

Data Mining and Knowledge Discovery

- Knowledge Discovery System [Fayy96]:



- Very ambicious goals

Data Mining and Knowledge Discovery

- Banca d'Italia [1995]: *Built a KDD system for*
  - Daily update of the whole set of movements
  - Decide what and how to analyze
  - Select relevant results
  - Produce a daily 2-pages synthesis (natural language)

Daily support to main boss decision making

Data Mining and Knowledge Discovery

- Banca d'Italia
  - *Built a KDD system for*

Daily support decision making of the main boss

- Technological problems
  - Millions of movements per day
  - Time to transmit to the central server?
  - Time to update the database?
  - How to select and retrieve proper data to analyze from DB?
  - How to validate results and verify technical assumptions?

- Methodological problems
  - Which is important to analyze today?
  - Which is the proper Data Mining technique?
  - Which are relevant results?
  - How to express results for the main boss?

Data Mining and Knowledge Discovery

- Big Supermarket chains *(Wal-Mart, EEUU, 1992 [Kelly 1996])*
  - Daily update the datawarehouse with costumer's bill contents

    (20milions daily transactions [Babcock 1994])
  - Decide what/how to analyze: Habits *(Market Basket analysis [Brin 1997])*
  - Select relevant results
    - What is buyed more
    - Main associations between products

      30% of transactions containing beer also contain diapers???????

      2% of transactions contain both of these items [Agrawaal 1996]

New Knowledge

  - Analyze the pattern in depth
    - *Friday between 5 and 7 pm, Young customers, Males*
  - Understanding the pattern

    Just-married with small kid cannot meet friends in pub on Friday night for party
    Helps wife with the shopping (required things…. Diapers for the kid);
    Beer is his personal reward to spend Friday at home

What is the usefullness of elicitating this knowledge?

Data Mining and Knowledge Discovery

Knowledge is power!!

- What is the value of identifying that

  Young new fathers buy diapers and beer on

  - Acquiring strategic information
  - Capacity of planning actions

    Wal-Mart moved the beer next to the diapers and beer sales went up

  - Capacity of becoming PROACTIVE

    What about moving snacks (peanuts and pretzels) next to diapers?

- Support decision making through informed-decision

  Buying department

  Marketing department

Important economic implications

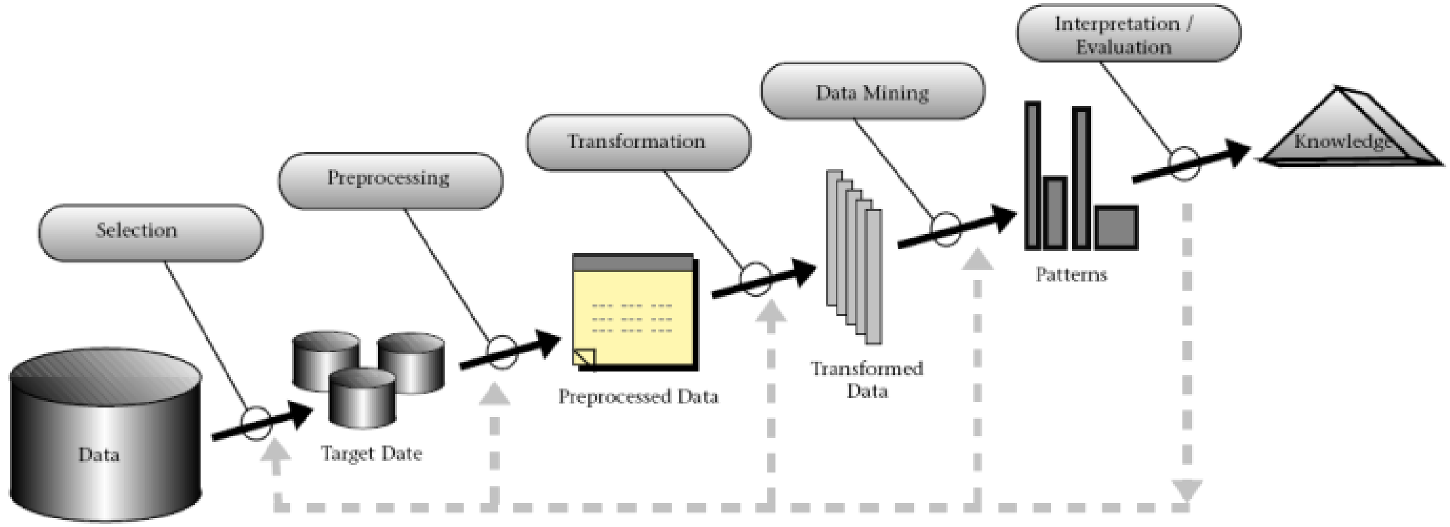
- From then on: apparently misplaced things in stores

Data Mining and Knowledge Discovery

- From then on: apparently misplaced things in stores

Data Mining and Knowledge Discovery

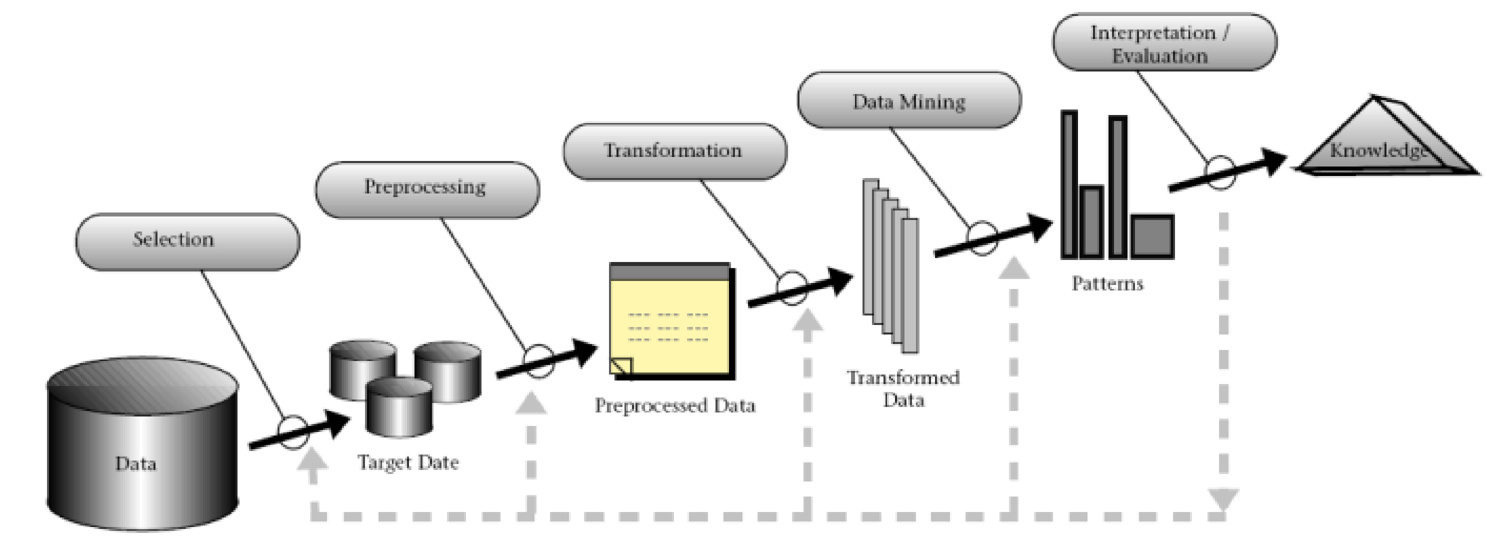
- Knowledge Discovery System [Fayy96]:



- Very ambicious goals

Data Mining and Knowledge Discovery

- Knowledge Discovery System [Fayy96]:



- Very ambicious goals
- No complete system on yet
  - Connection to DataWarehouses
  - Tools to assist preprocessing
  - Collection of data mining techniques *(AMD, NN, IR, AssR, Reg...)*
  - Tools on automatic reporting phase
  - Manual process management and knowledge production

K. Gibert

Data Mining and Knowledge Discovery

- New paradigm proposed by Fayyad

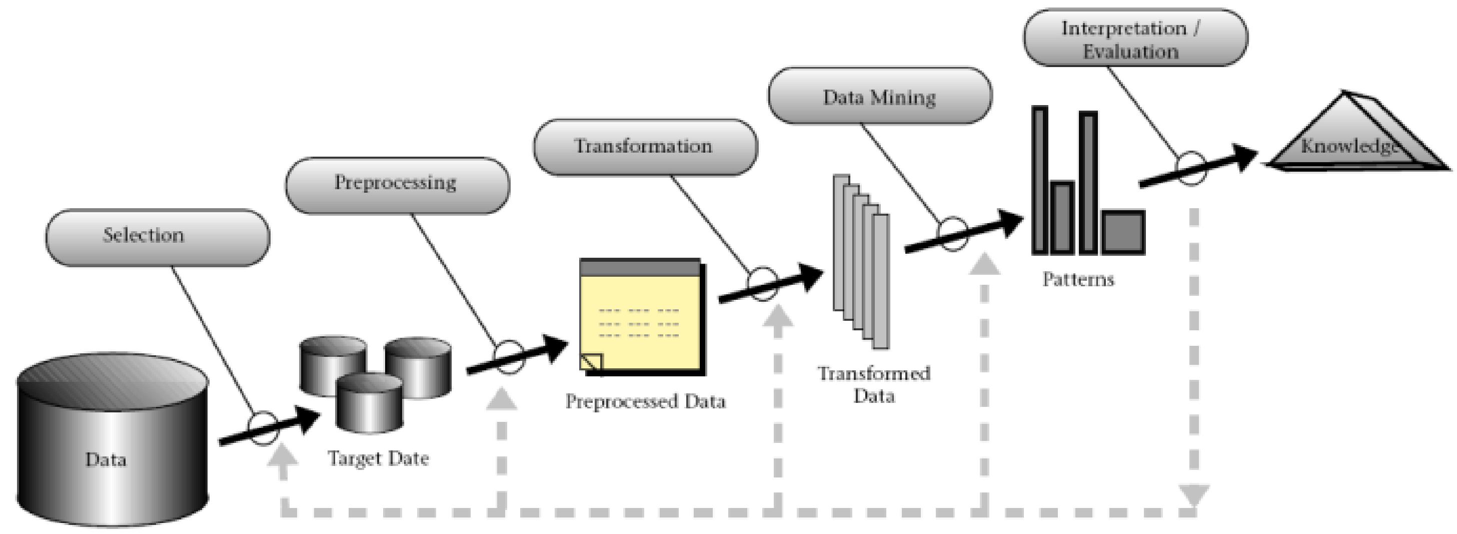
> "*Most previous work on KDD has focussed on [...] data mining step. However, the other steps are of considerable importance for the successful application of KDD in practice""*
>
> **[Fayyad 96]**

- Include prior and posterior analysis in KDD
  - Requires Great efforts in real applications
    - Specially in medical systems (uncertainty, imprecise, multi-scaled,..)
  - Time consuming, difficult (no standard methodology stablished)
  - Expert interaction required
  - Domain-dependent?

  - After good prior analysis, proper data mining easy

Data Mining and Knowledge Discovery

- Knowledge Discovery System [Fayy96]:



- Wide scope approach

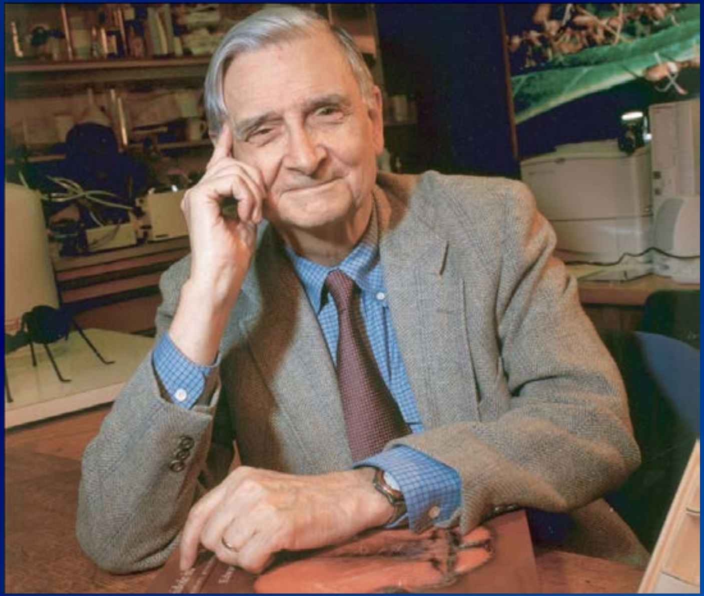
  - Also interesting to better know very complex small datasets

Multidisciplinariety

Combination or hybridation of techniques

K. Gibert

Data Mining and Knowledge Discovery



A balanced perspective cannot be acquired by studying disciplines in pieces; the consilience among them must be pursued. Such unification will be difficult to achieve. But I think it is inevitable. Intellectually it rings true, and it gratifies impulses that arise from the admirable side of human nature. To the extent that the gaps between the great branches of learning can be narrowed, diversity and depth of knowledge will increase.

[E.O. Wilson 1998]

Biologist, Harvard U, EEUU

Twice Pulitzer; Times (1995) 25 most influencial people in America

Data Mining and Knowledge Discovery

The Elephant and the blind Men (Ancient India)

[Puchala 1971]

An elephant came to a small town (had ever seen one)

Ancient council (5 blind men) went to feel the elephant with their hands.

Later on, they sat down and began to discuss their experiences.

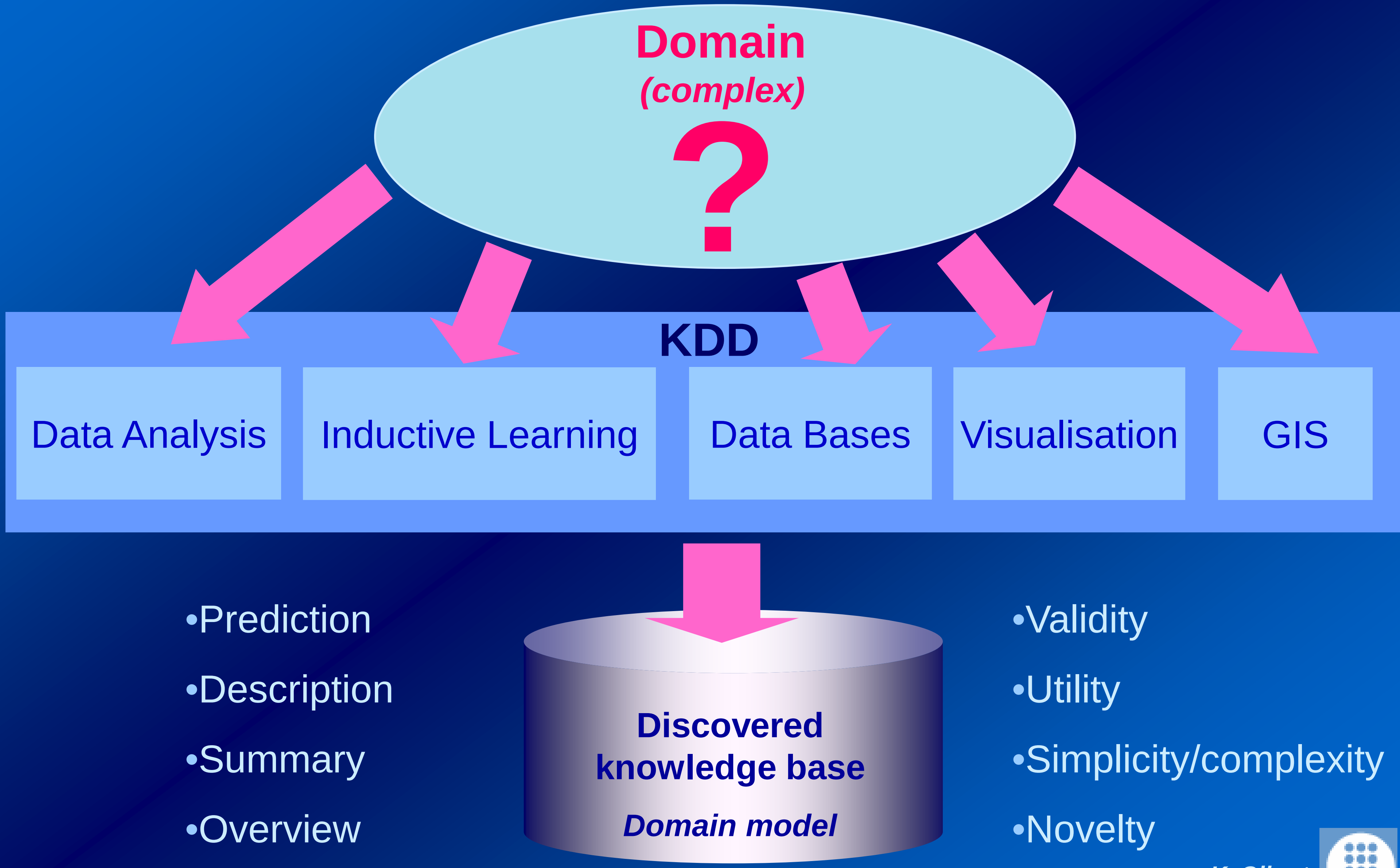
- One who touched the trunk and feld like a thick tree branch.
- Another who touched the tail feld like a snake or rope.
- Another who touched the leg, feld like a pillar.
- Another who touched the ear, said like a huge fan
- Another who touched the side, said like a wall.

Consilience
Multidisciplinariety

All had different partial views of the same reality

Putting all partial views together, the complete view could emmerge

Data Mining and Knowledge Discovery



- Prediction
- Description
- Summary
- Overview

- Validity
- Utility
- Simplicity/complexity
- Novelty

K. Gibert

Artificial Intelligence and Statistics

Interdisciplinar research field

- Starting:
  - 1985: Douglas Fisher and Bill Gale (AI&Stats Society)
  - 1986: First Int'l Conference on AI & Stats
- Main goals:
  - Promote communication between AI and Statistics communities

→ ***"We feel that there is great potential for development at the intersection of Artificial Intelligence, Computational Science and Statistics"***

Cheeseman and Oldford 94.

- Improve research in problems common to both

  (Data Mining and Knowledge Discovery, ...)

KDD uses

Decission Support

- Improving complex decision making
- Intelligent Decision Support Systems

Bussiness intelligence

Also useful to cope with complexity

Domains

- Marketing
- Bussiness
- Research

Medical, industrial, environmental applications