DS-6 meeting – Leicester 02/24/2005

Giuseppe Longo^{1,4,5}

Ciro Donalek^{1,6}, Antonino Staiano^{1,4} Roberto Tagliaferri^{2,5}, Angelo Ciaramella^{2,5}

- 1 Dept. Of Physical Sciences, University of Napoli
- 2 Dept. Of Mathematics, University of Salerno
- 3 Caltech Dept. Of Astronomy
- 4 INAF –Napoli Unit
- 5 INFN Napoli Unit
- 6 Caltech, Pasadena, USA

&

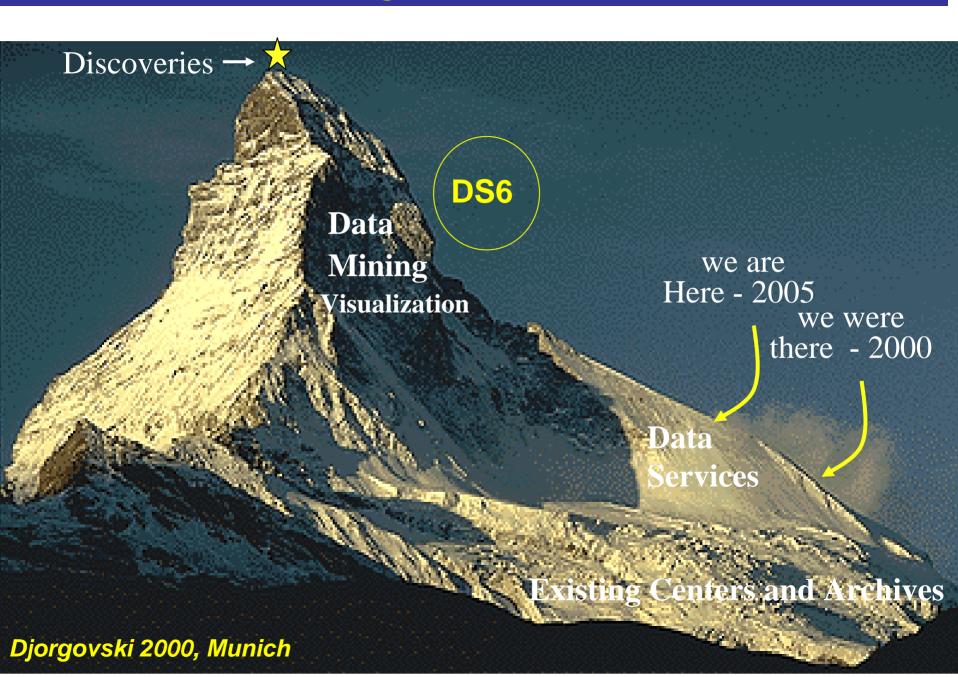
U. Becciani

INAF-Catania

C. Gheller

CINECA - Bologna

DS-6 meeting – Leicester 02/24/2005



What is data mining?

The capability to extract useful information from large data sets

Guided by a priori knowledge	Based on statistical properties
Accurate	Less accurate
Reliable	More exciting
Socially acceptable	Socially questioned

Feature selection

Clustering

Pattern and trend recognition

Feature selection

From an ocean of parameters, exctract only those which are relevant to a given task

compression of data sets, computational speed

Clustering:

the mathematical word for most of the tasks performed by the astronomers in interpreting data

Pattern and trend recognition:

All the other tasks done by the astronomers

IMPLEMENTATION OF AN AUTOMATIC CLASSIFIER (S/U) FOR MULTIBAND SURVEY DATA

- 1. Construction of the "a priori" knowledge data archive (first: existing data; after: integration with simulations)
- 2. Implementation of a feature selection supervised tool
- 3. Application to test data sets (Sloan, DPOSS, ESO archive data)
- 4. Generalization for usage on a distributed architecture

CDS ("a priori" knowledge)
Edimburgh, Portsmouth (SDSS data)
Napoli (DPOSS and A.I. know how)
Edimburgh, Leicester, Portsmouth, Trieste (GRID know-how)
Catania (Visualization tools)

