IJCAI-09

Workshop on Machine Learning and AI Applications in Astrophysics and Cosmology

Pasadena, California,
July 16 - 17, 2009

A Modern Scientific Discovery Process

Data Gathering (e.g., from sensor networks, telescopes…)

↳ **Data Farming:**

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability
} Database Technologies

↳ **Data Mining** (or Knowledge Discovery in Databases):

Pattern or correlation search
Clustering analysis, automated classification
Outlier / anomaly searches
Hyperdimensional visualization

Key Technical Challenges

↳ **Data Understanding**

↳ **New Knowledge**

Key Methodological Challenges

+feedback

Information Technology ➔ New Science

- The information volume grows exponentially

Most data will never be seen by humans

➡ The need for data storage, network, database-related technologies, standards, etc.

- Information **complexity** is also increasing greatly

Most data (and data constructs) cannot be comprehended by humans directly

➡ The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery …

- We need to create ***a new scientific methodology*** to do the 21st century, computationally enabled, data-rich science…
- ML and AI will be essential components of the new scientific toolkit

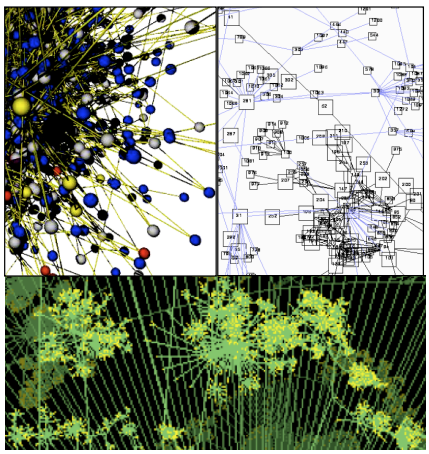
The Key Challenge: Data Complexity

Or: The Curse of Hyper-Dimensionality

1. **Data mining algorithms scale very poorly:**

   N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

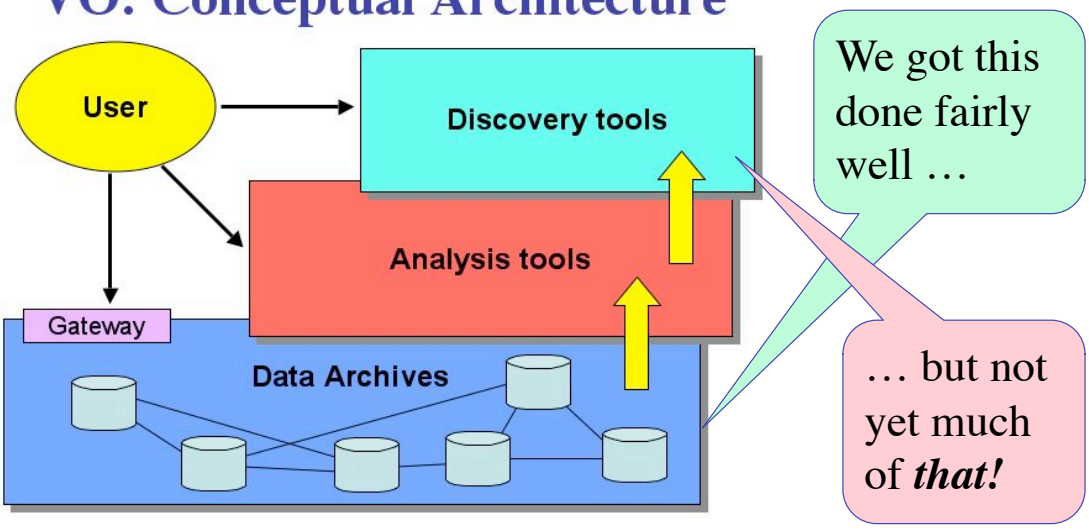
   – Clustering ~ N log N → N^2, ~ D^2
   – Correlations ~ N log N → N^2, ~ D^k ($k \geq 1$)
   – Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)



2. **Visualization in >> 3 dimensions**

- The complexity of data sets and interesting, meaningful constructs in them is *exceeding the cognitive capacity of the human brain*
- We are biologically limited to perceiving D ~ 3 - 10(?)
- Visualization is a bridge between data and human intuition/understanding

VO: Conceptual Architecture



Virtual Observatory framework today is a ***data grid*** of astronomy – but it needs to become also the ***discovery space***

KDD / ML / AI tools are essential if the VO is to fulfill its scientific potential and its intended role

The Roles for Machine Learning and Machine Intelligence in CyberScience:

- **Data processing:**
  - Object / event / pattern classification
  - Automated data quality control (glitch/fault detection and repair)
- **Data mining, analysis, and understanding:**
  - Clustering, classification, outlier / anomaly detection
  - Pattern recognition, hidden correlation search
  - Assisted dimensionality reduction for hyperdim. Visualisation
  - (Adaptive) workflow control in Grid/Cloud-based apps
  - Human-computer interface / collaboration / synergy
- **Data farming and data discovery:** semantic web, and beyond
- **Code design and implementation:** from art to science?

About This Workshop:

- Discussion is the main thing – don't be shy!
- Problems, challenges, new ideas – not a recap of the past results (except sparingly, as an illustration)
- We can change the agenda as needed, responding to the flow of ideas, discussions
- Initiating new collaborations and projects would be great

Some logistics:

- Dinner tonight
- Proceedings?