# Service-Oriented Distributed Data Analysis in Grids and Clouds

**Domenico Talia**

**ICAR-CNR & UNIVERSITY OF CALABRIA**

**Italy**

**talia@deis.unical.it**

**Joint work with : Eugenio Cesario, Marco Lackovic, Paolo Trunfio**

HPC 2010 – June 21-25, 2010 - Cetraro, Italy

# Goal

- Discuss a strategy based on the use of services for the design of **distributed knowledge discovery tasks and applications** on Cloud, Grids and large distributed systems.

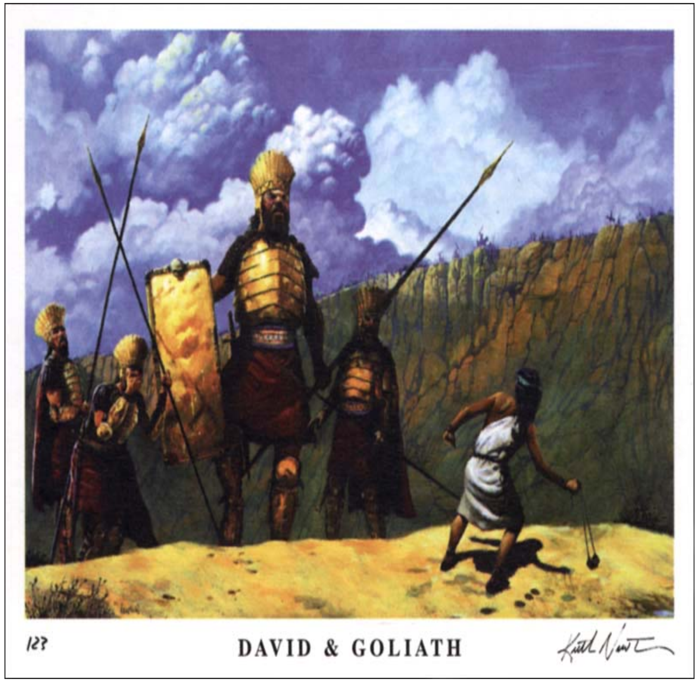
- Outline how **service-oriented knowledge discovery tasks** can be developed **as a collection of Grid/Web/Cloud services**.

- Present a **service-oriented framework for** composing and running **distributed data mining workflows**.

# Complex Big Problems

- Bigger and more complex problems must be solved by large scale distributed computing.

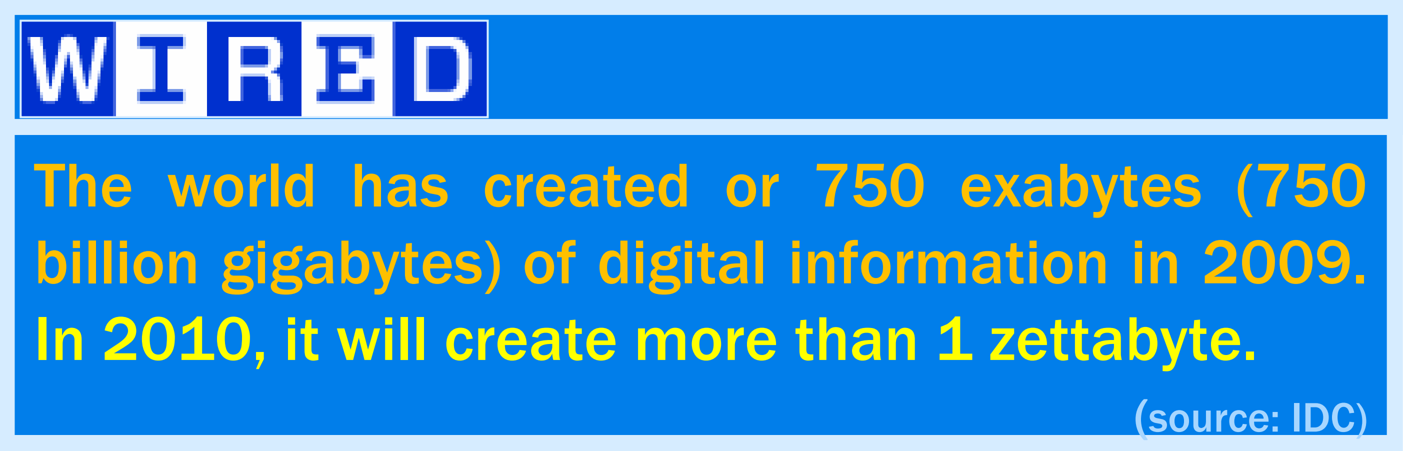
- DATA SOURCES are larger and larger and ubiquitous (Web, sensors, mobile devices, telescopes, …).



- The huge amount of DATA available today requires data analysys techniques to aid people to deal with it.

# Data Availability or Data Deluge?

- Today the information stored in digital data archives is enormous and its size is still growing very rapidly.



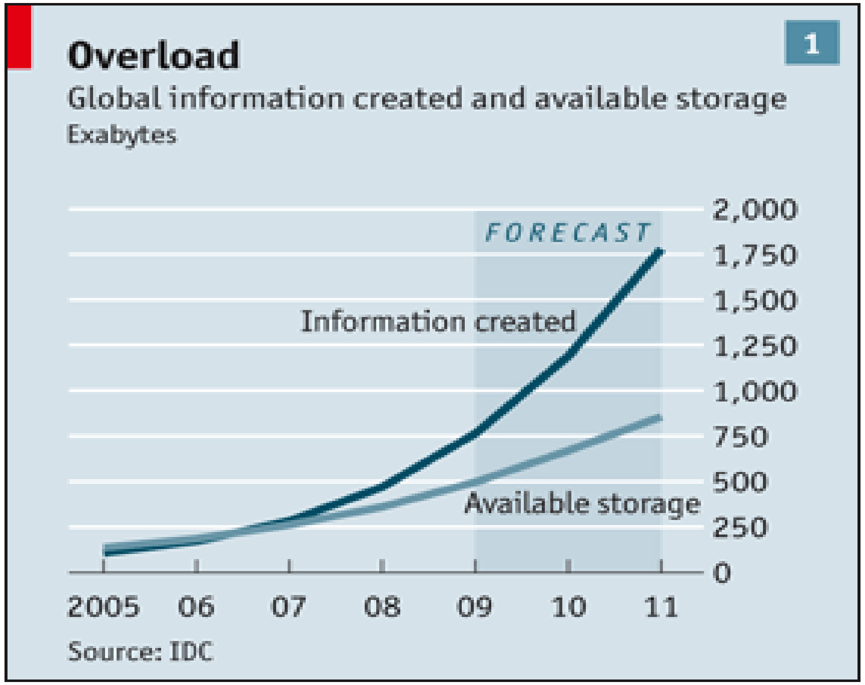
WIRED

The world has created or 750 exabytes (750 billion gigabytes) of digital information in 2009. In 2010, it will create more than 1 zettabyte.

(source: IDC)

# Data Availability or Data Deluge?

- Whereas until some decades ago the main problem was the **shortage of information**, the challenge now seems to be
  - the **very large volume of information** to deal with and
  - the **associated complexity** to process it and to extract significant and useful parts or summaries.



# Data Analysis

- Today our main problem is not only storing DATA, but it is analyse, mine, and process DATA for making it useful.

Source: The Economist

# Distributed Data Intensive Apps

- The use of computers (and associated digital data) changed our way to make discoveries and is improving both speed and quality of the scientific discovery processes.
- In this scenario HPC, Cloud and Grid systems provide an effective **computational support** for running **distributed data intensive applications** and for **knowledge discovery from large and distributed data sets**.
- Grid systems, HPC computers, and cloud computing systems demonstrated to be key technologies for e-Science. **They can be used in integrated frameworks through service interfaces**.

# Service-Oriented Distributed Data Mining

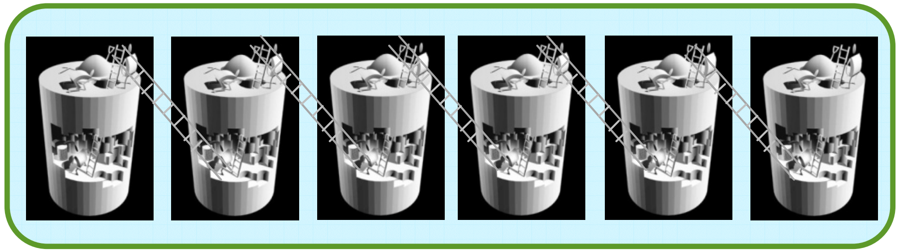
- **Knowledge discovery (KDD)** and **data mining (DM) are**:
  - Compute- and data-intensive processes/tasks
  - Often based on distribution of data, algorithms, and users.

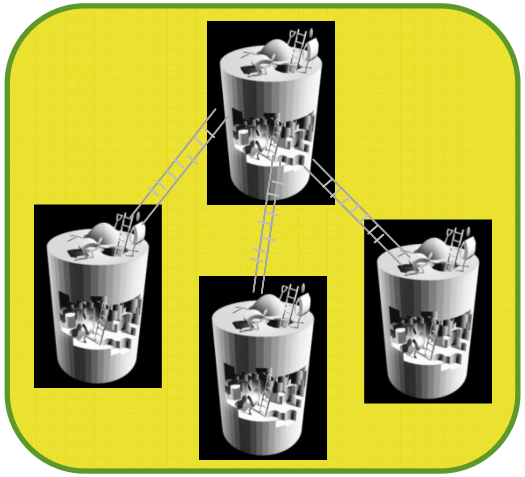
- Large scale service-oriented systems like Clouds and Grids integrate both distributed computing and parallel computing, thus they are **key infrastructures for high-performance distributed knowledge discovery**. (e.g., **Knowledge Grids**, **Data Analytics Clouds**)

- They also offer
  - security, resource information, data access and management, communication, scheduling, fault detection, …

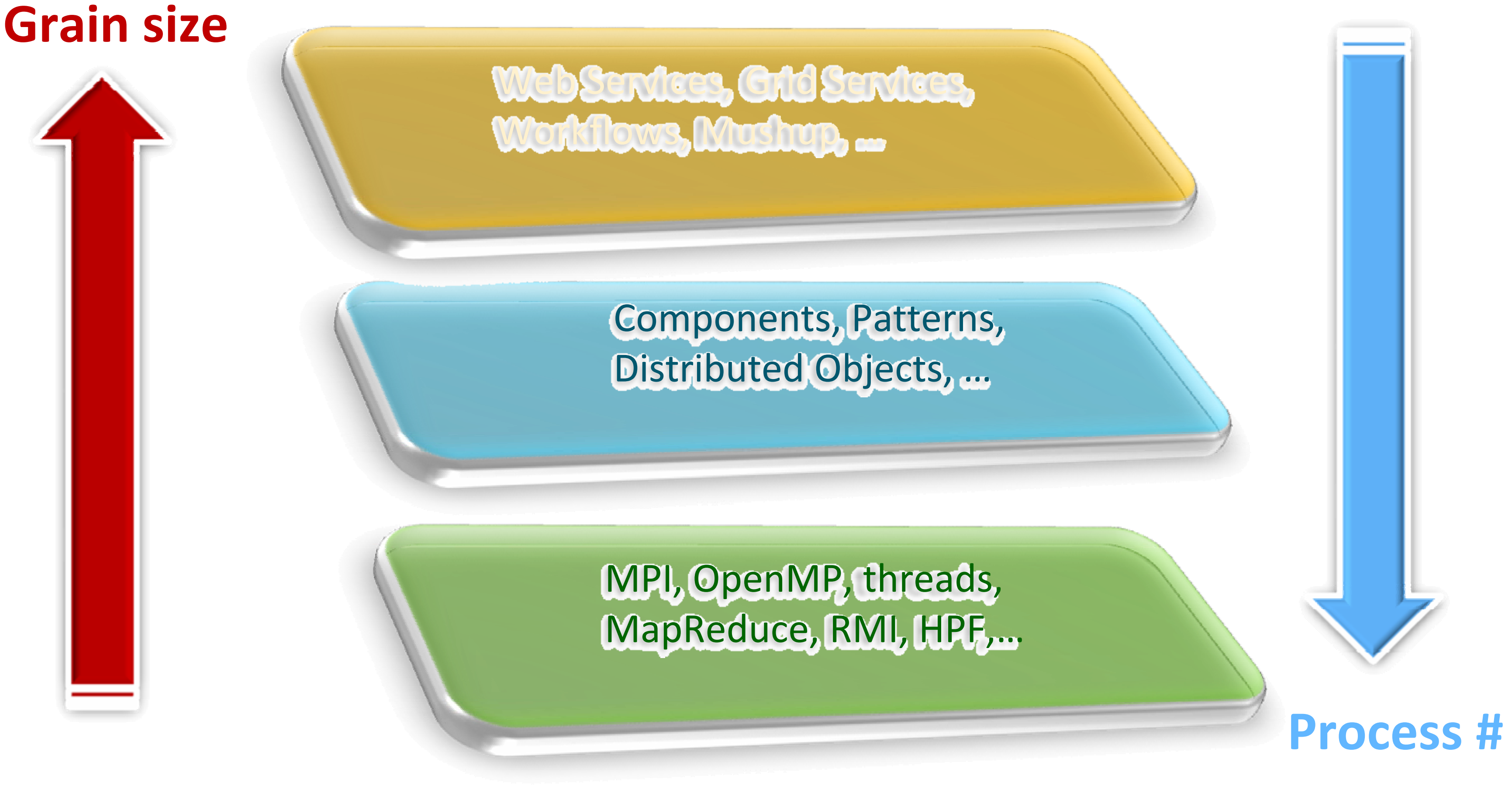
# Distributed Data Analysis Patterns

- **Data parallelism? Task parallelism?**
- **Managing data dependencies**
- **Dynamic task graphs/workflows** (data dependencies)
- **Dynamic data access** involving large amounts of data
- **Parallel data mining** and/or **Distributed data mining**
- Programming **distributed mining operations/taks/patterns**





# Programming Levels



Grain size

Web Services, Grid Services, Workflows, Mushup, ...

Components, Patterns, Distributed Objects, ...

MPI, OpenMP, threads, MapReduce, RMI, HPF,...

Process #

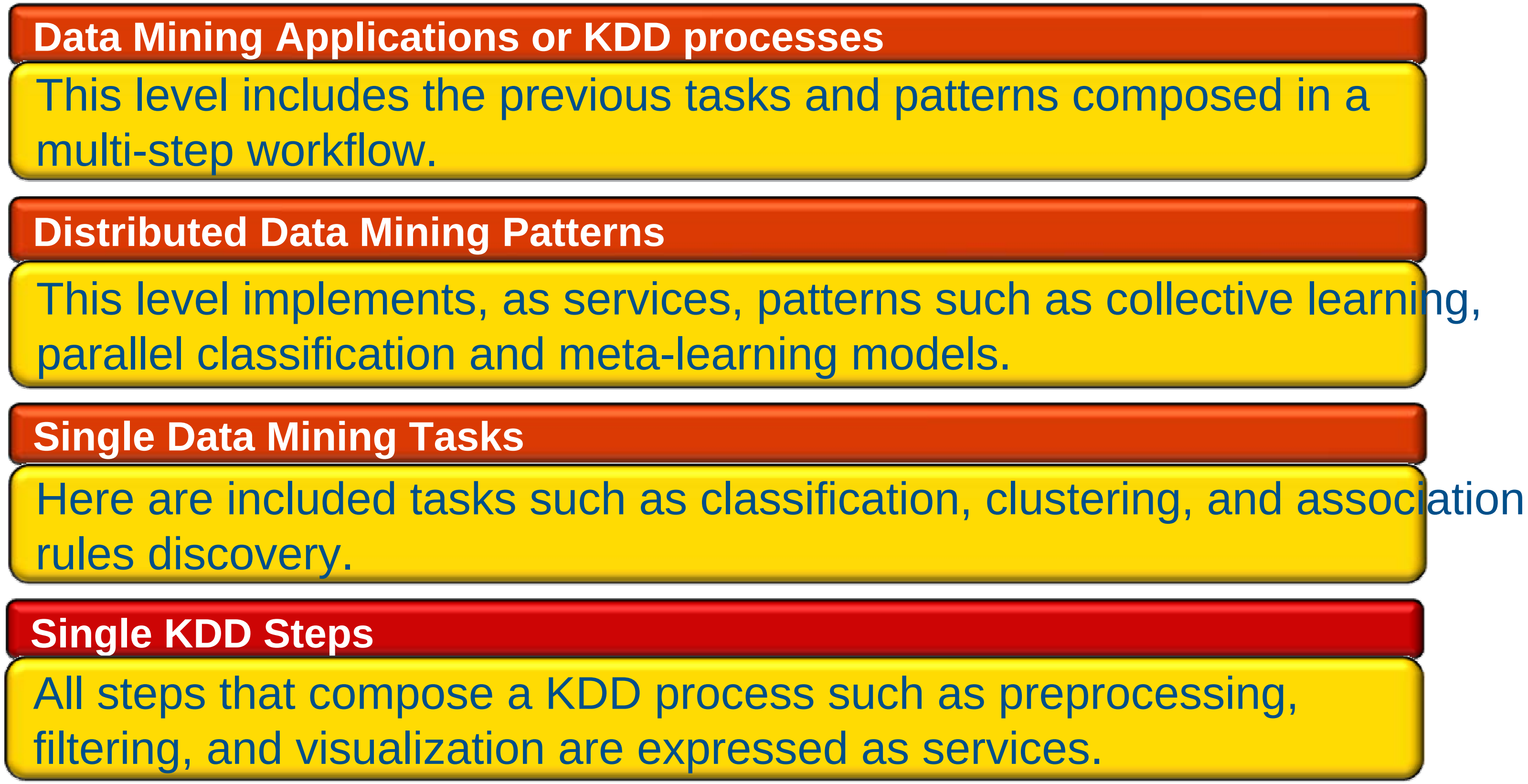
# Services for distributed data mining

- Exploiting the SOA model it is possible to define **basic services for supporting distributed data mining tasks/applications** in large scale distributed systems for science and industry (for example: from a private Cloud to InterClouds).

- Those services can address all the aspects that must be considered in data mining and in knowledge discovery processes
  - data selection and transport services,
  - data analysis services,
  - knowledge models representation services, and
  - knowledge visualization services.

# Collection of Services for Distributed Data Mining

- It is possible to design services corresponding to



**Data Mining Applications or KDD processes**

This level includes the previous tasks and patterns composed in a multi-step workflow.

**Distributed Data Mining Patterns**

This level implements, as services, patterns such as collective learning, parallel classification and meta-learning models.

**Single Data Mining Tasks**

Here are included tasks such as classification, clustering, and association rules discovery.

**Single KDD Steps**

All steps that compose a KDD process such as preprocessing, filtering, and visualization are expressed as services.

# Data mining services

- This collection of data mining services implements an

**Open Service Framework for Distributed Data Mining**

- Allowing developers to program distributed KDD processes as a composition of single and/or aggregated services available over a service-oriented infrastructure.

- Those services should exploit other basic Grid/Cloud services for data transfer, replica management, data integration and querying.

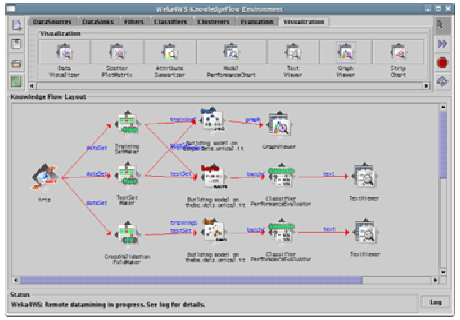
# Data mining services

- By exploiting the Web/Grid/Cloud services features it is possible to develop **data mining services accessible every time and everywhere** (remotely and from small devices).

- This approach can result in
  - Service-based distributed data mining applications
  - Data mining services for communities/virtual organizations.
  - Distributed data analysis services on demand.
  - A sort of **knowledge discovery eco-system** formed of a large numbers of decentralized data analysis services.

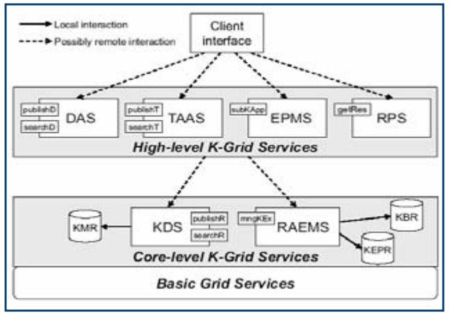
# Data mining services

- Service-based systems we developed
  - Weka4WS



Weka4WS: Remote datamining in progress. See log for details.

  - KNOWLEDGE GRID



  - Mobile Data Mining Services

# S-O Distributed Data Mining Workflows

- **DIS3GNO** is a visual framework for programming and running service-oriented data mining workflows in the **KNWOLEDGE GRID**.

- The **KNWOLEDGE GRID** is a system providing services to execute distributed data mining tasks or KDD processes as services.

- **DIS3GNO** supports all the phases of a distributed knowledge discovery process, including composition, execution, and results visualization.

# S-O Distributed Data Mining Workflows

- A data mining workflow is a graph in which
  - **nodes** typically represent data sources, filtering tools, data mining algorithms, and visualizers, and
  - **edges** represent execution dependencies among nodes.

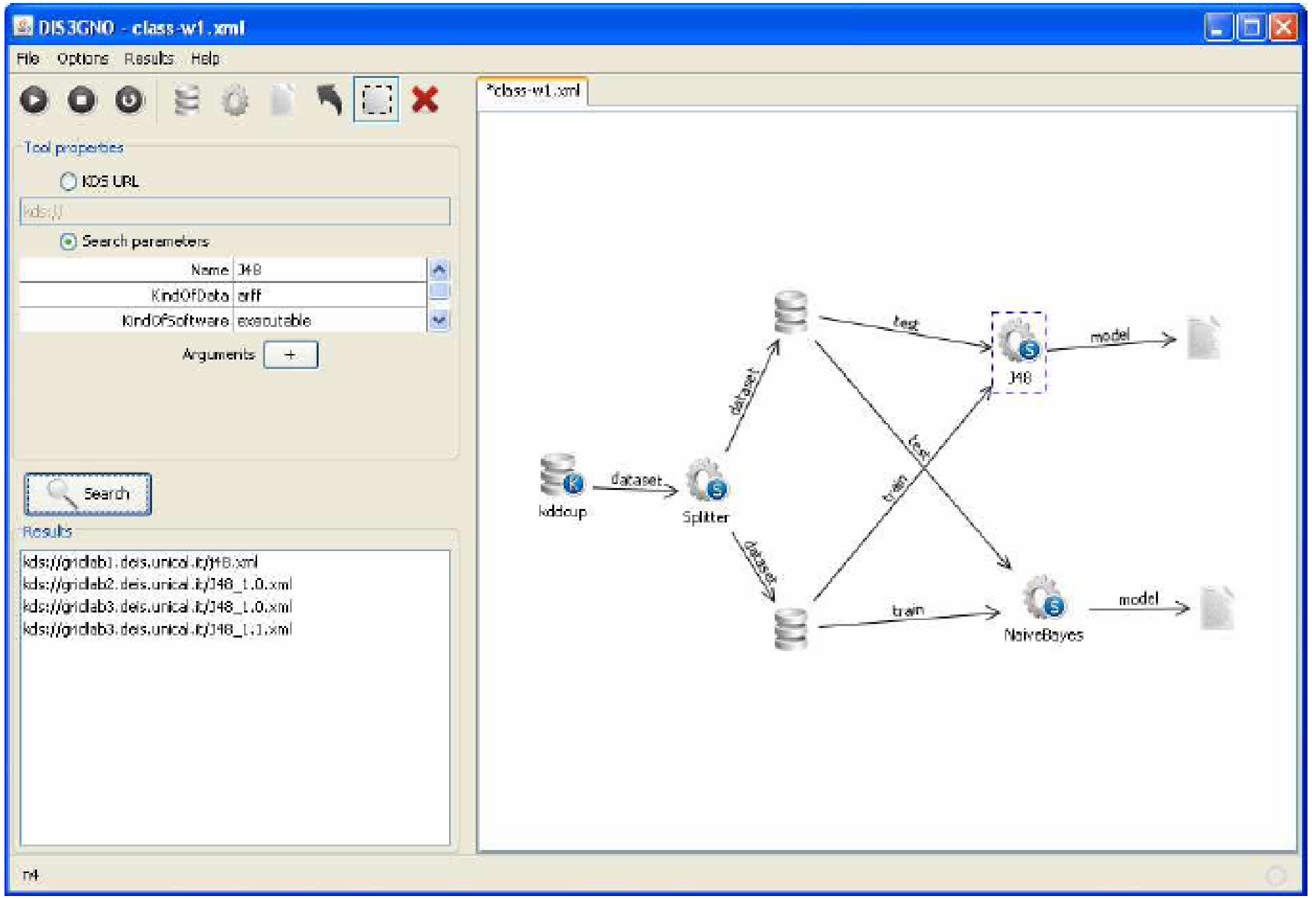
- **DIS3GNO** supports all the phases of a distributed knowledge discovery process, including composition, execution, and results visualization.

- **Each node is a service**.

# S-O Distributed Data Mining Workflows

- The workflow concept plays a fundamental role in the KNWOLEDGE GRID at different levels of abstraction.
- A client application submits a distributed data mining application to the KNWOLEDGE GRID by describing it through an XML workflow formalism (conceptual model).
- The conceptual model describes data and tools to be used, with or without specifying information about their location or implementation.

# DIS3GNO: A Visual Framework



○ KDS URL

kds://

● Search parameters

| | |
|---|---|
| Name | J48 |
| KindOfData | arff |
| KindOfSoftware | executable |

Arguments +

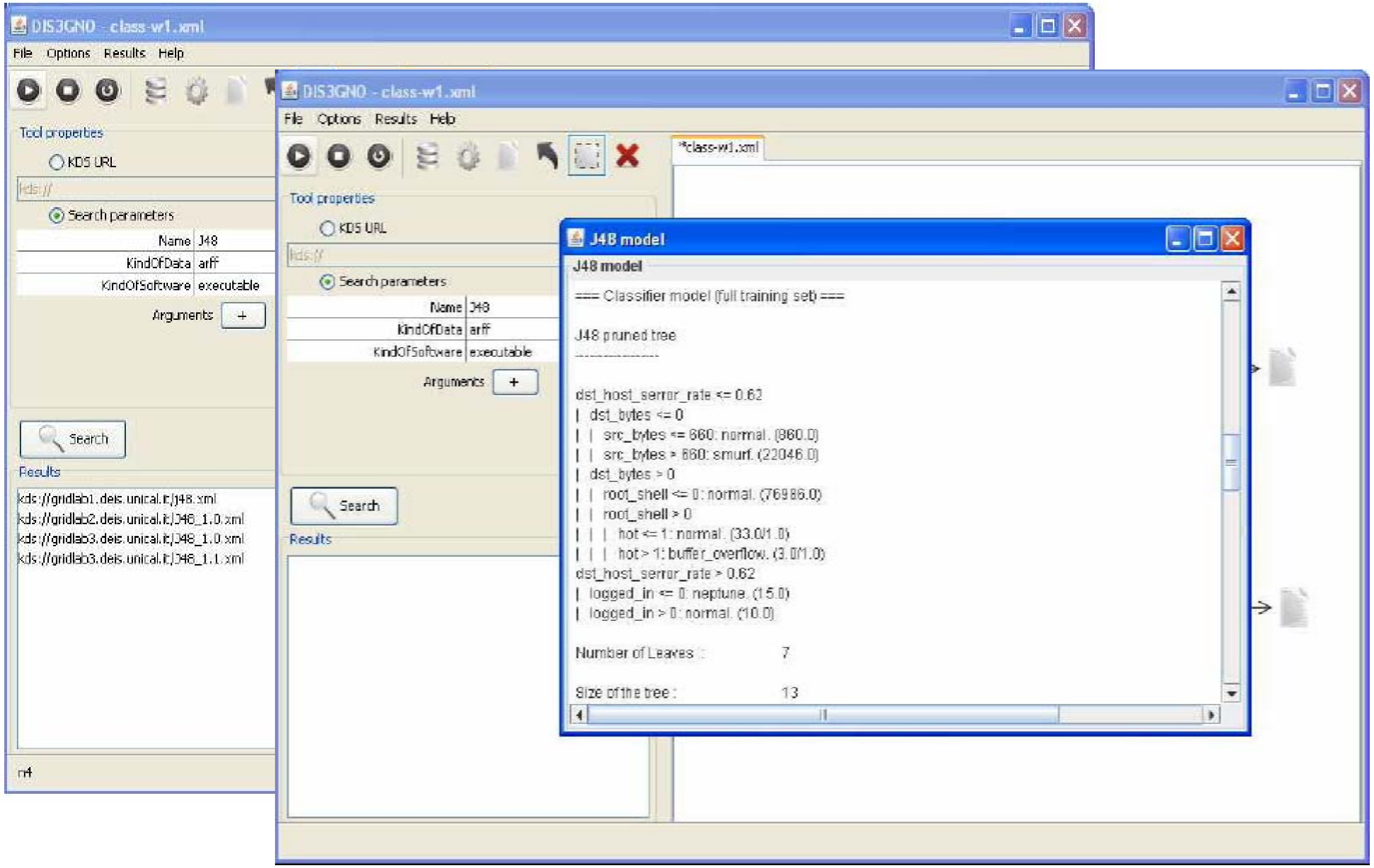
```
kds://gridlab1.deis.unical.it/J48.xml
kds://gridlab2.deis.unical.it/J48_1.0.xml
kds://gridlab3.deis.unical.it/J48_1.0.xml
kds://gridlab3.deis.unical.it/J48_1.1.xml
```

Programming a data mining workflow as a graph of services and run them in parallel.

# DIS3GNO : A Visual Framework

- DIS3GNO is the user front-end for two main KNOWLEDGE GRID operations:
    - *Metadata management.* DIS3GNO provides an interface to publish and search metadata about data and tools.
    - *Design and Execution management.* DIS3GNO provides an environment to design and execute distributed data mining applications as workflows, through the interaction with the execution services of the KNOWLEDGE GRID.

# DIS3GNO: A Visual Framework



| | |
|---|---|
| Name | J48 |
| KindOfData | arff |
| KindOfSoftware | executable |

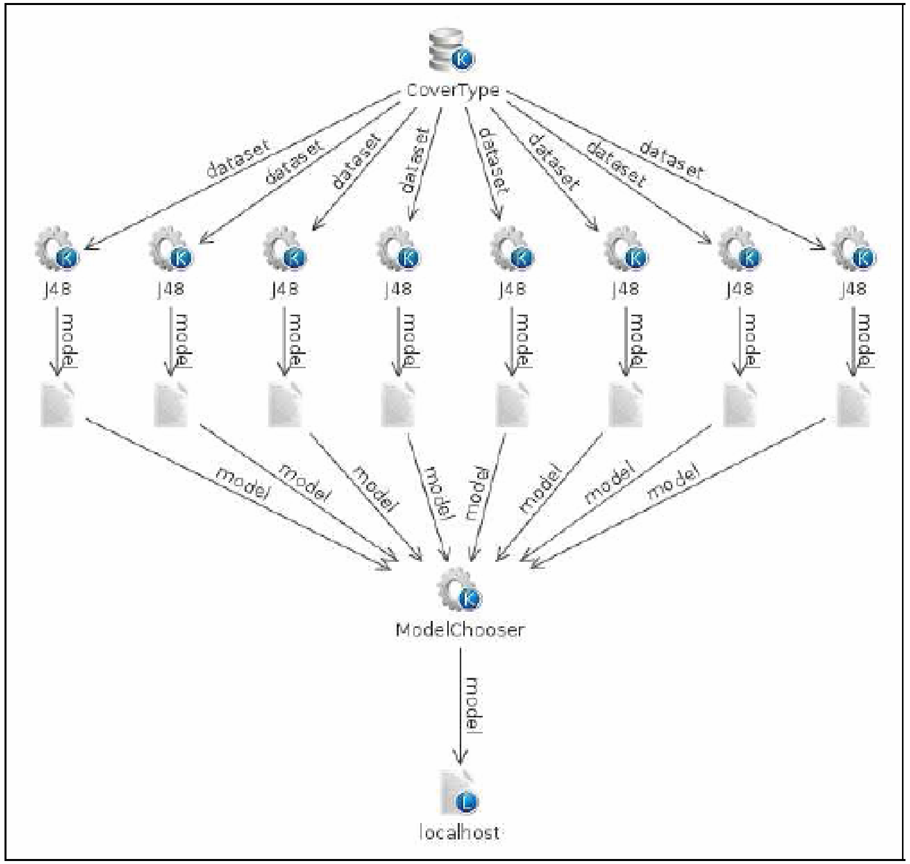
```
kds://gridlab1.deis.unical.it/j48.xml
kds://gridlab2.deis.unical.it/J48_1.0.xml
kds://gridlab3.deis.unical.it/J48_1.0.xml
kds://gridlab3.deis.unical.it/J48_1.1.xml
```

J48 model

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

dst_host_serror_rate <= 0.62
|   dst_bytes <= 0
|   |   src_bytes <= 660: normal. (860.0)
|   |   src_bytes > 660: smurf. (22046.0)
|   dst_bytes > 0
|   |   root_shell <= 0: normal. (76986.0)
|   |   root_shell > 0
|   |   |   hot <= 1: normal. (33.0/1.0)
|   |   |   hot > 1: buffer_overflow. (3.0/1.0)
dst_host_serror_rate > 0.62
|   logged_in <= 0: neptune. (15.0)
|   logged_in > 0: normal. (10.0)

Number of Leaves  :     7

Size of the tree :      13
```

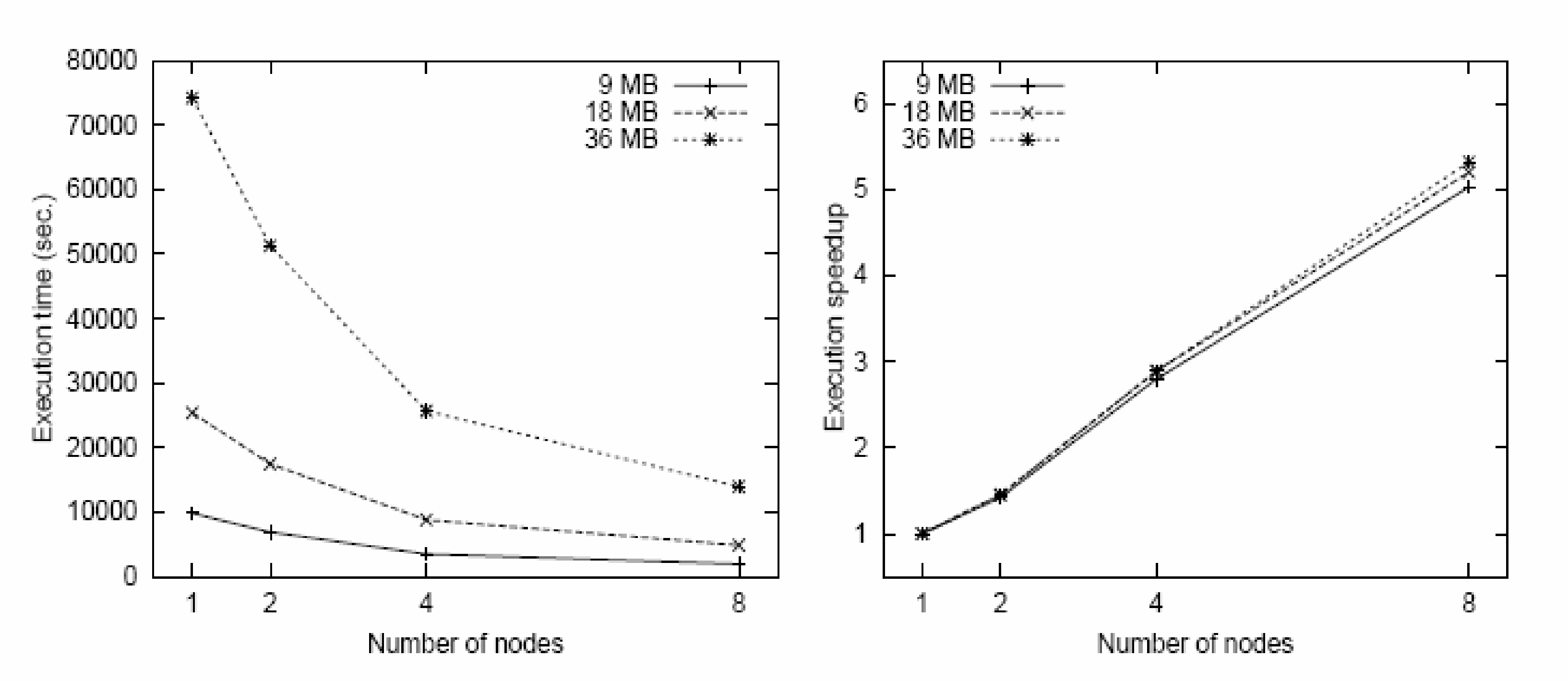
Workflow running and results visualization after workflow completion.

# Data Mining Workflows with DIS3GNO



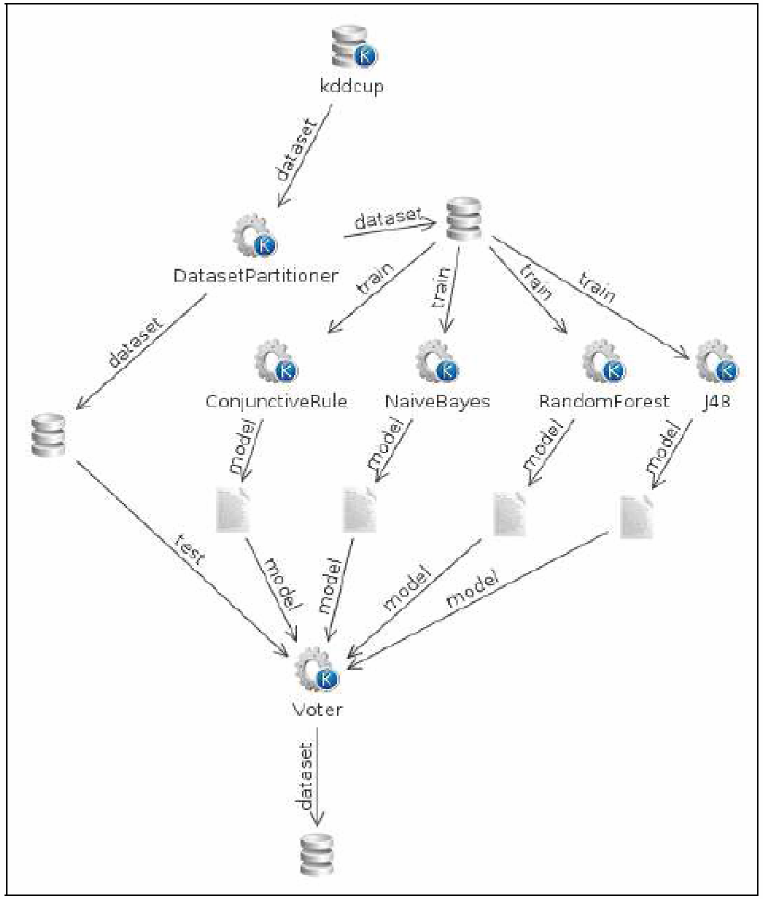
Eight similar classifiers in parallel produce different classifications (using different parameters) of the same dataset. The best classification is selected by the ModelChooser node.

# Performance Results



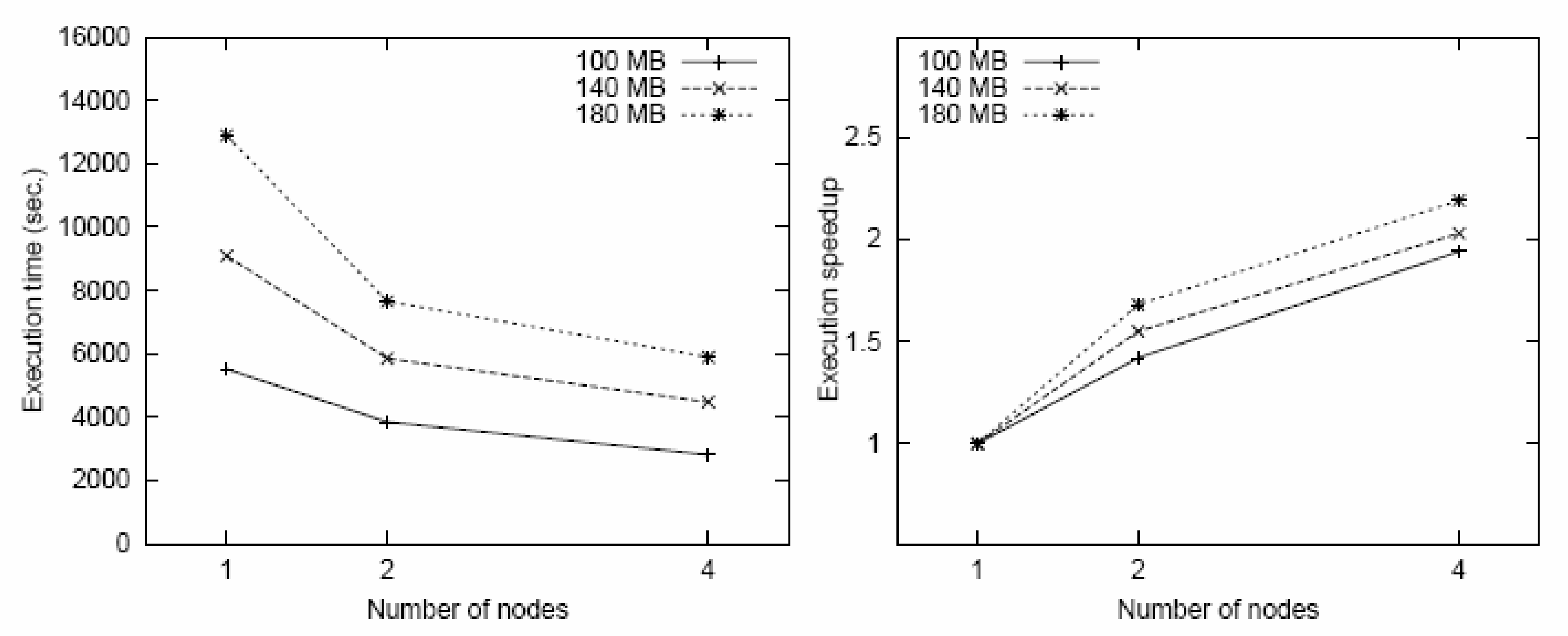
Execution time and speedup with different dataset sizes. With the 36 MB dataset, time is reduced from 21 hours to 3,5 hours.

# Data Mining Workflows with DIS3GNO



In an ensemble learning application four different classifiers in parallel produce 4 classifications from 4 different training sets. The best classification is selected by voting.

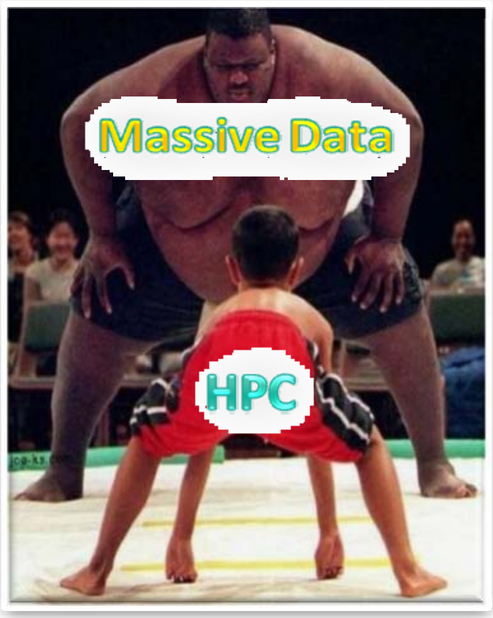
# Performance Results



Execution time and speedup with different dataset sizes. The overall execution time is bound to the execution time of the slowest algorithm, thus limiting the total speedup.

# Summary

- New HPC infrastructures allow us to attack new problems, BUT require to solve more challenging problems.
- New models, frameworks, and environments are required
  - Data is becoming a BIG player, programming data analysis applications and services is a must.
  - New ways to efficiently compose different models and paradigms are needed.
  - The service-oriented approach can be a viable integration paradigm.



- In a long-term vision, **pervasive collections of data analysis services** and applications must be **accessed and used as public utilities**.
- We must be ready for managing with this scenario.

# QUESTIONS?

grid.deis.unical.it