Discovery Net
Unifying the World's Knowledge

Building a Knowledge Grid for the Pharmaceutical Industry

Prof. Yike Guo
Dept. of Computing
Imperial College

Founder & CEO
InforSense

Discovery Net Project

Discovery Net
Unifying the World's Knowledge

- ***Discovery Net*** is a £2.4 Million UK e-Science Pilot Project building the next-generation *Grid for Knowledge Discovery* by combining Grid-related *standards* with the *Kensington* platform.

- Discovery Net infrastructure is based on extending the ***KDE*** platform from InforSense for meeting grid computing requirements.

- ***Mission:*** Providing advanced grid infrastructure for creating, integrating, managing and leveraging intellectual properties of a modern organisation.

Discovery Net for the Enterprise

Discovery Net
Unifying the World's Knowledge

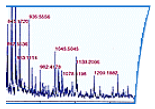
Corporate Information

Literature

Databases

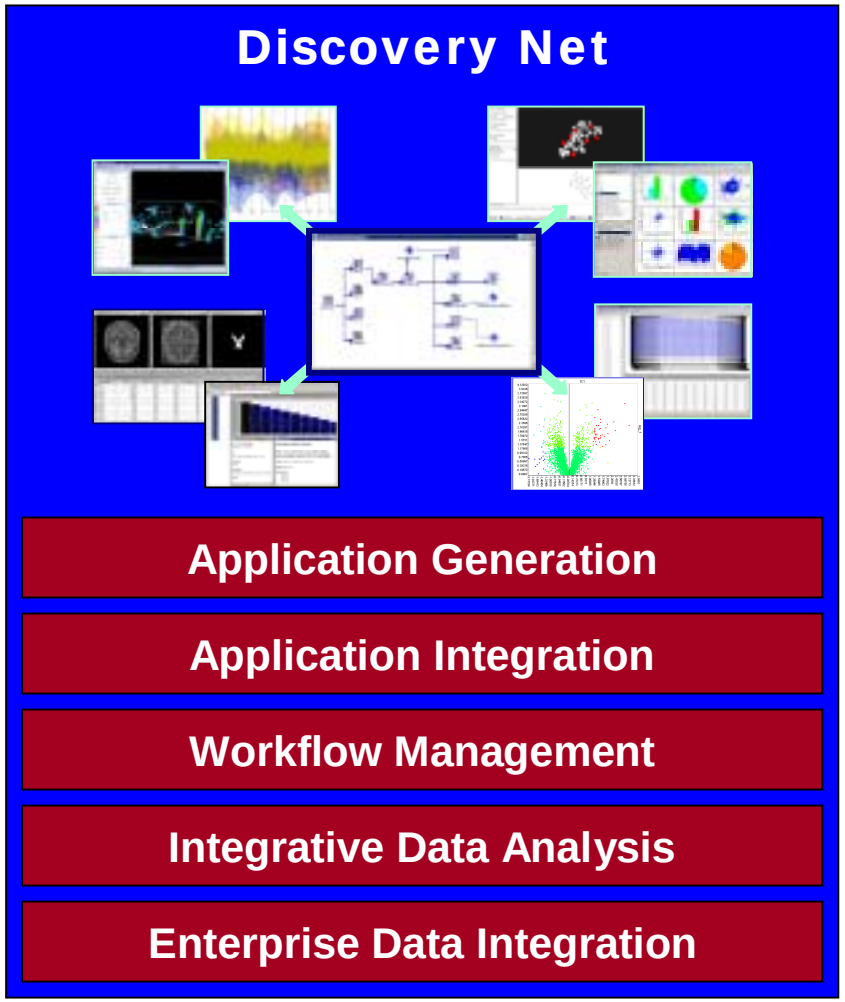
Operational Data

Images



Instrument Data

Workflow-based Discovery Services

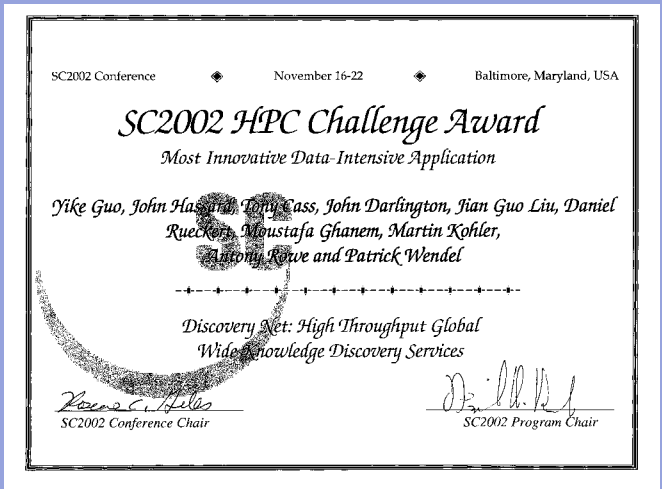


Corporate Intelligence

Technology Leadership

Discovery Net
Unifying the World's Knowledge

High Performance Computing Challenge


SC2002 Conference ◆ November 16-22 ◆ Baltimore, Maryland, USA

SC2002 HPC Challenge Award

Most Innovative Data-Intensive Application

Yike Guo, John Hassard, Tony Cass, John Darlington, Jian Guo Liu, Daniel Rueckert, Moustafa Ghanem, Martin Kohler, Antony Rowe and Patrick Wendel

Discovery Net: High Throughput Global Wide Knowledge Discovery Services

SC2002 Conference Chair

SC2002 Program Chair

180,000 clicks, 350 cut and pastes, 200 database accesses, 250 access to computation services will be done in one workflow and one click

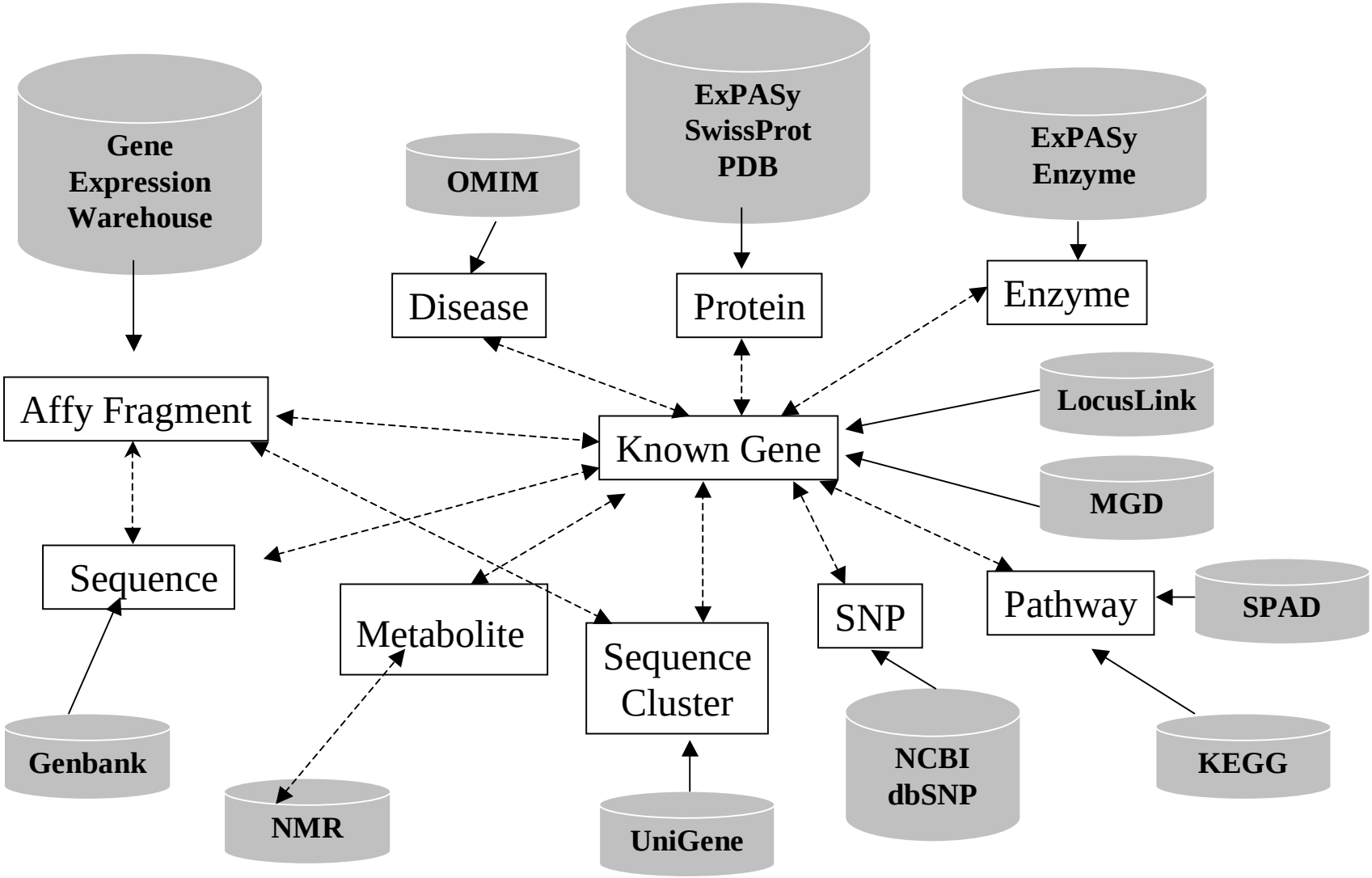
Recognized for outstanding text mining ability in an international competition organized by the ACM (Association for Computing Machinery)

Discovery Net workflows offer the best solution to integrate bioinformatics resources for integrated discovery activities.

A Life Science Collaborator

Data Integration: The First Step

Discovery Net
Unifying the World's Knowledge



Data Warehousing: Thinking by Following Schemas

Discovery Net
Unifying the World's Knowledge

- **Query to the data**
  - Which genes are expressed similarly to a particular gene XYZ?
  - Which genes are co-expressed in differing conditions ?
  - classification (of tumours, diseased tissues etc.): which patterns are characteristic for a certain class of samples, which genes are involved?
  - functional grouping of genes: Are changes clustered in particular classes?
  - metabolic pathway information: Is a certain pathway/route in a pathway affected?
  - disease information & clinical follow up: correlation to expression patterns.
  - phenotype information for mutants: Are there correlations between particular phenotypes and expression patterns?

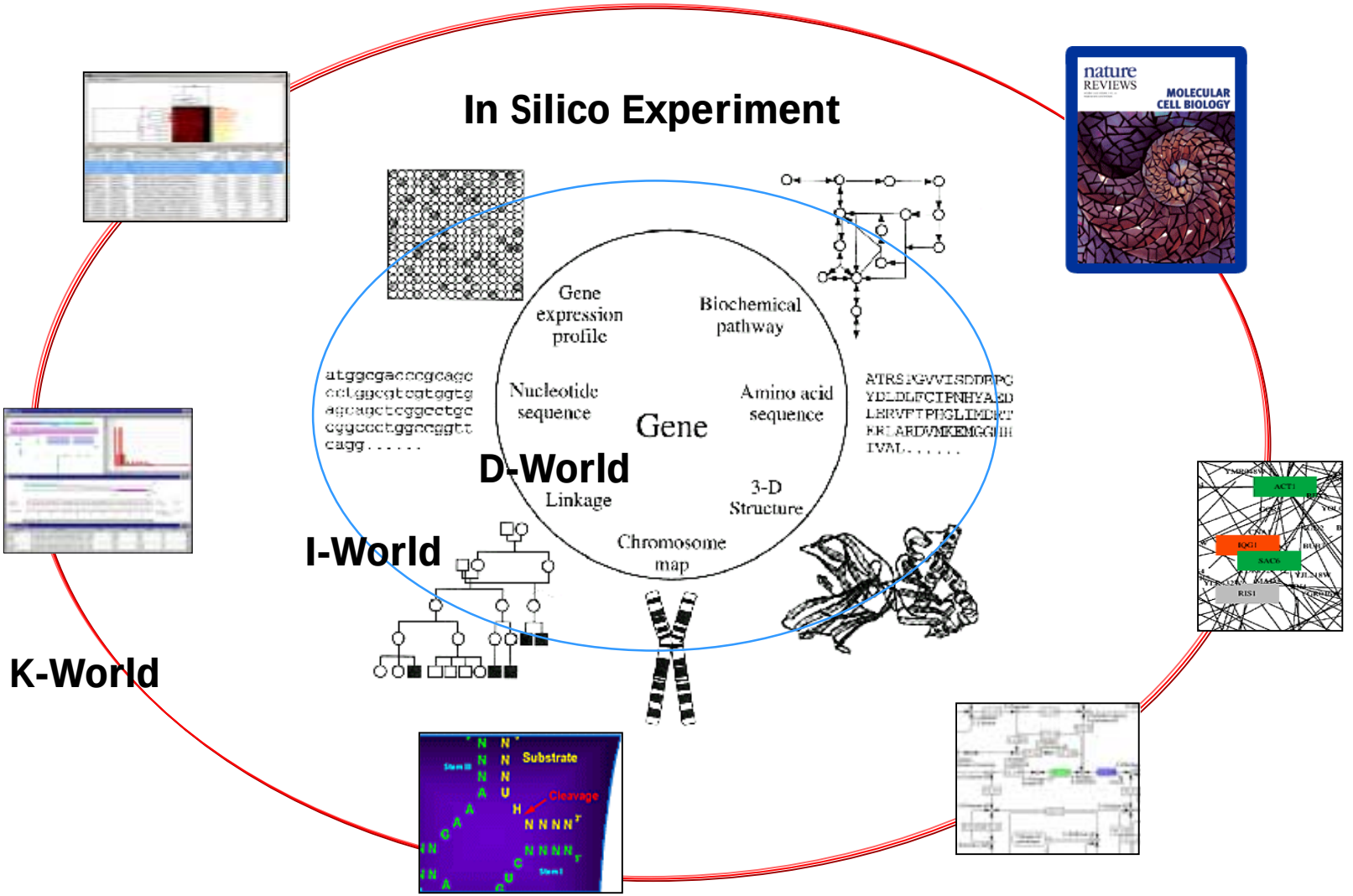
(Un)fortunately, Scientists Never Just Think Following Schemas

Discovery Net
Unifying the World's Knowledge

- Why those genes are co-expressed?
- What do their protein products do?
- What is the common regulatory motifs of a co-expressed gene set?
- Can we patent them?
- Do we know which metabolic pathway they are in? If there is no, can I synthesis one?
- Are there HTS results for any proteins in the pathway?
- Are there any compounds in the HTS library that hit selectively and consistently against those proteins?
- Which ones have good activity, availability and toxicity?

From Data Integration to Knowledge Grid

Discovery Net
Unifying the World's Knowledge



The IT Issues of Building a Knowledge Grid

Discovery Net
Unifying the World's Knowledge

- **Data Integration:**
  - Dynamic Real Time Link All Required Information
- **Application Integration:**
  - Easy Plug-in of Applications
  - Applications Integrated based on a “Discovery Logic”
  - Service-based Open Informatics
- **People Integration:**
  - Global-wide Discovery Groupware
  - Support a Collaborative Research Environment
- **Knowledge Integration:**
  - Multi-subjects and Multi-modality Integrative Analysis to Cross Validate and Annotate Related Discovery Work
  - Discovery Workflow to Present a “Discovery Logic”, to construct a “Discovery Plan”

Open Workflow:
The Discovery Net Solution

Discovery Net
Unifying the World's Knowledge

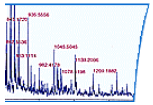
Scientific Information

Literature

Databases

Operational Data

Images

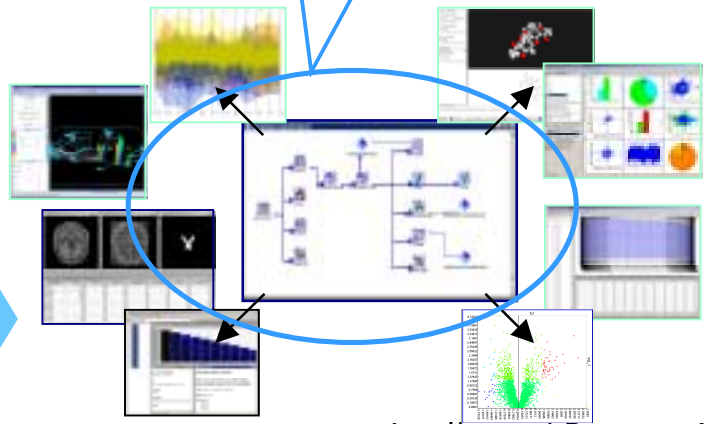


Instrument Data

Discovery Workflow

Real Time Data Integration

Discovery Services



Dynamic Application Integration

Intellectual Property Management

Scientific Discovery

Using Distributed Resources

Workflow

Discovery Net
Unifying the World's Knowledge

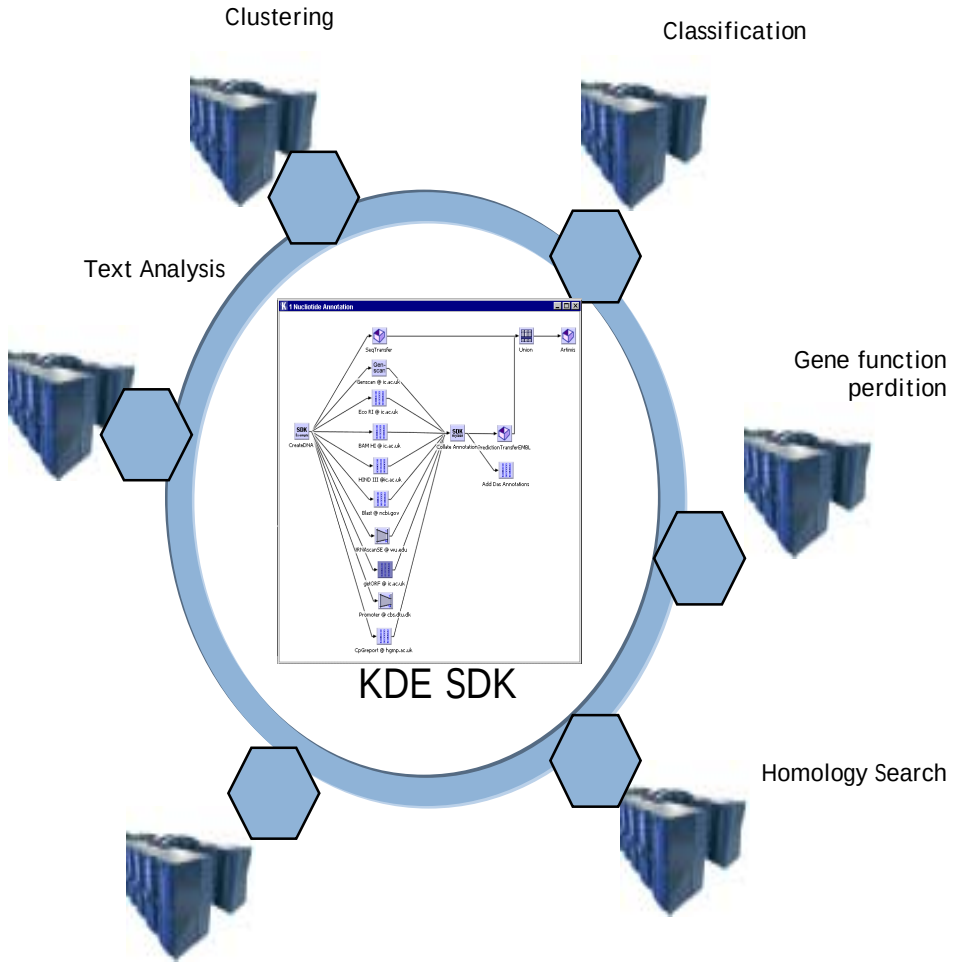
- **Workflow not only provides**
  - provenance information
  - operation logic
- **But also provides**
  - means for information/application/service composition
  - strategy for distributed computation (in a grid environment)
  - model for defining and deployment of new service
- **In one word, workflow represents the Knowledge of Action.**
- **That is why we call the workflows in discovery informatics as Discovery Plans.**

Workflow for Service Composition

- **Workflow = Discovery Planning by Service Composition**

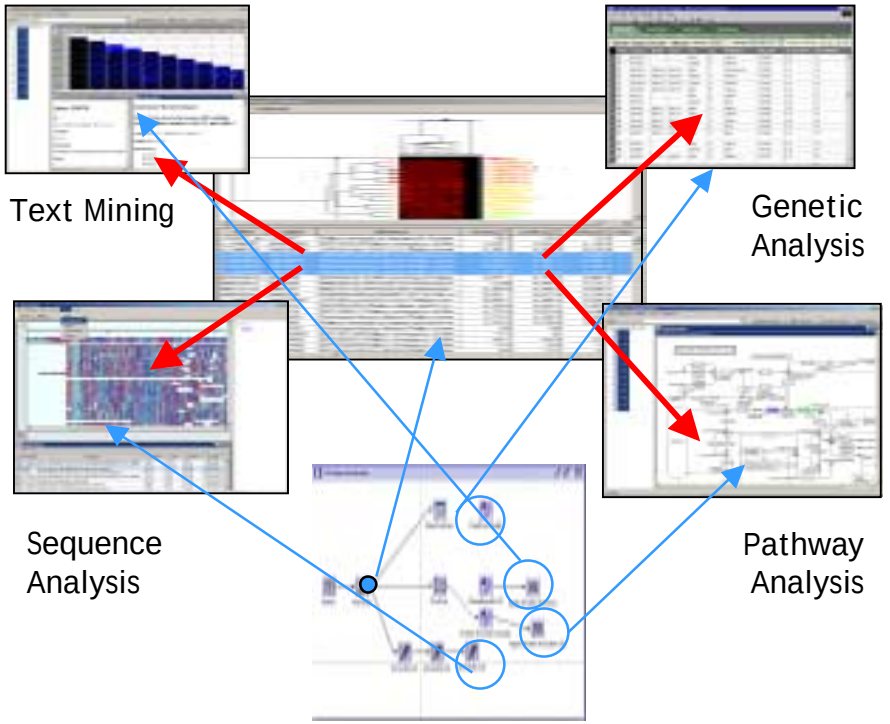
- **Towards a Workflow-based Dynamic Service Integration:**
  - Knowledge Servers: allow users to register, locate and execute (bioinformatics/chemoinformatics) applications as services.
  - Execution Servers: allow users to control the execution of component services in distributed environments
  - Open Grid Service Architecture:

    OGSA-compliant architecture supporting global wide application/service integration.



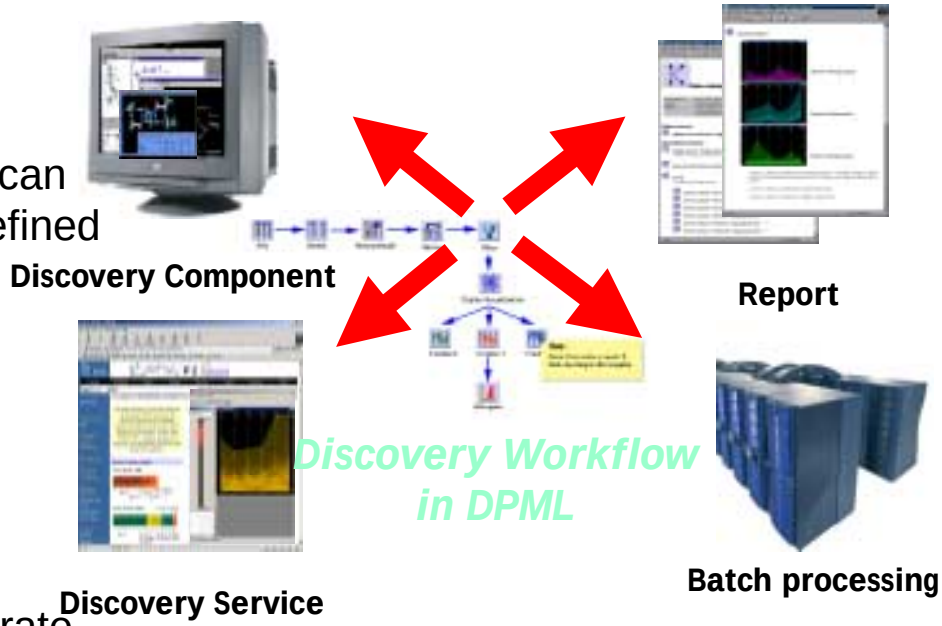
Workflow for Knowledge Integration

- **Knowledge Unification = Dynamically construction of schemas to organise related cross-domain analysis results and background knowledge**
- **Towards a Knowledge Schema Framework for integrative discovery**
  - **A Mechanism of Indexing, Annotating Discovery Results:** An querying and browsing system for discovered knowledge
  - **Discovery Plan based Knowledge Management:** An abstraction structure of discovery plans to organise enterprise level discovery activities—subjects, projects, experiments ….



Workflow as Deployable IP Assets

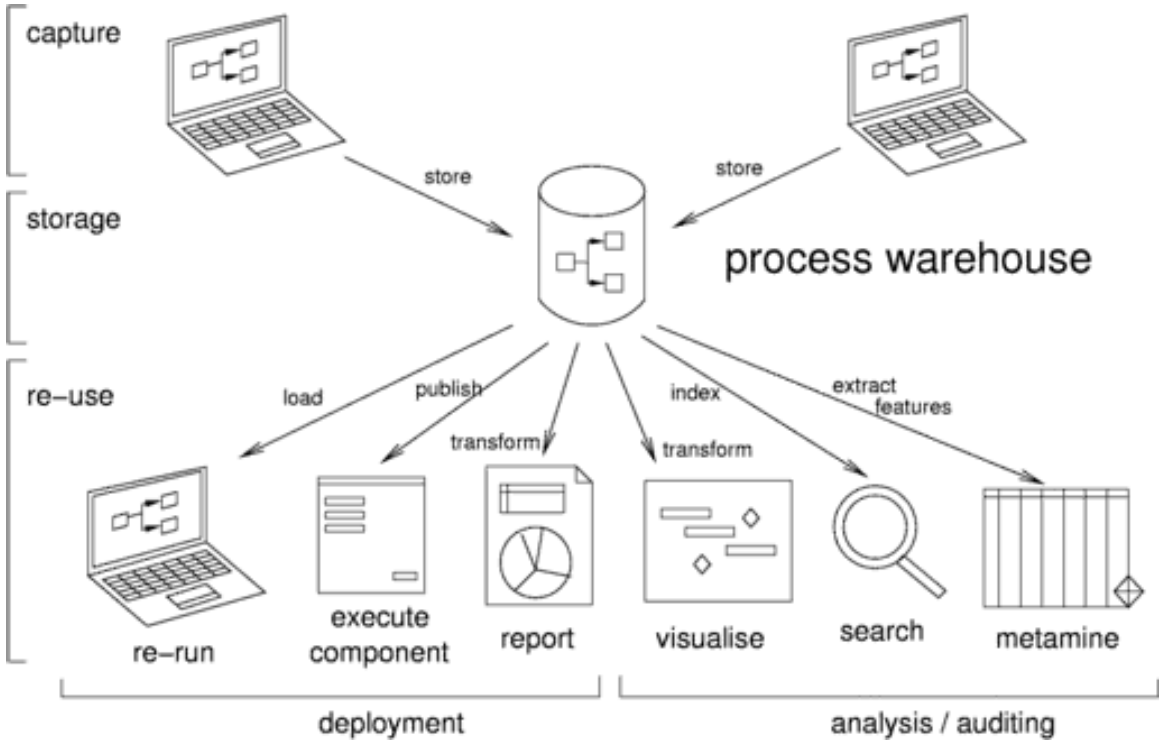
- **Workflow Deployment = On-demand rapid application construction :a Dynamic Deployment of Knowledge Discovery Services:**
  - **Discovery Plans as Collaborative Intellectual Property:** Discovery Plans can be stored, reused, audited, analysed, refined and deployed in various forms
  - **Deployment Engine :** allows users to transform workflows into software applications
  - **My Discovery:** each company will generate its own discovery software reflecting the company's own discovery logic.



Workflow Storage:
Enterprise Intellectual Property Management

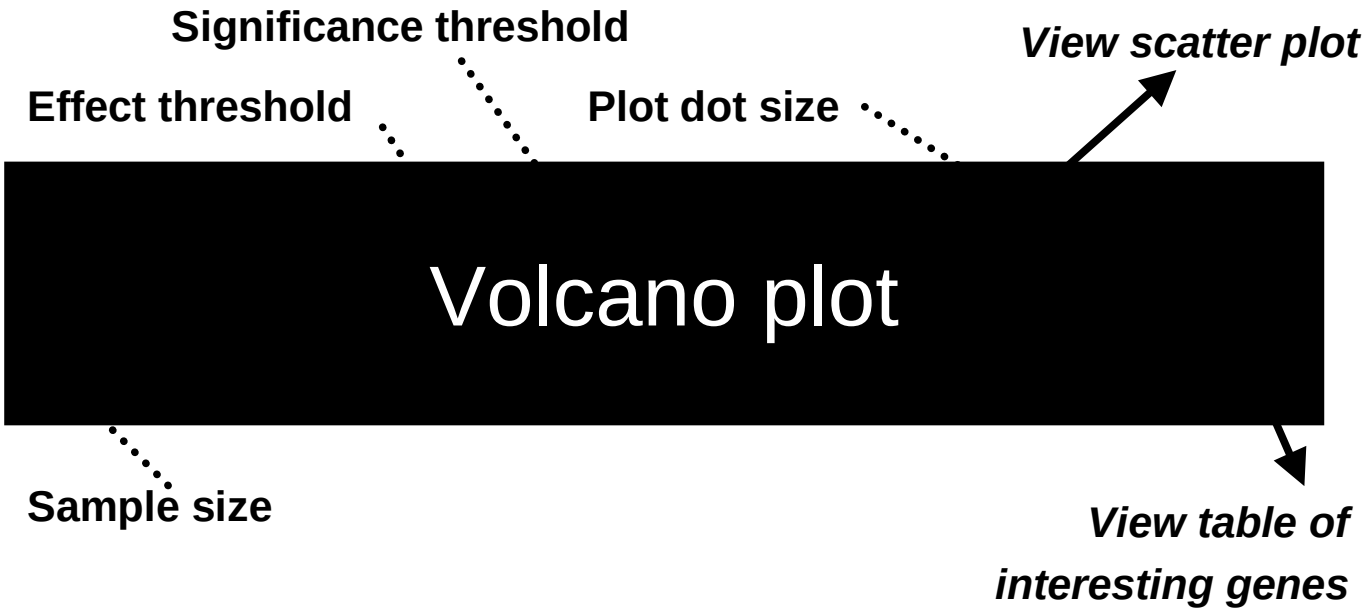
Discovery Net
Unifying the World's Knowledge

- **Workflow Storage**



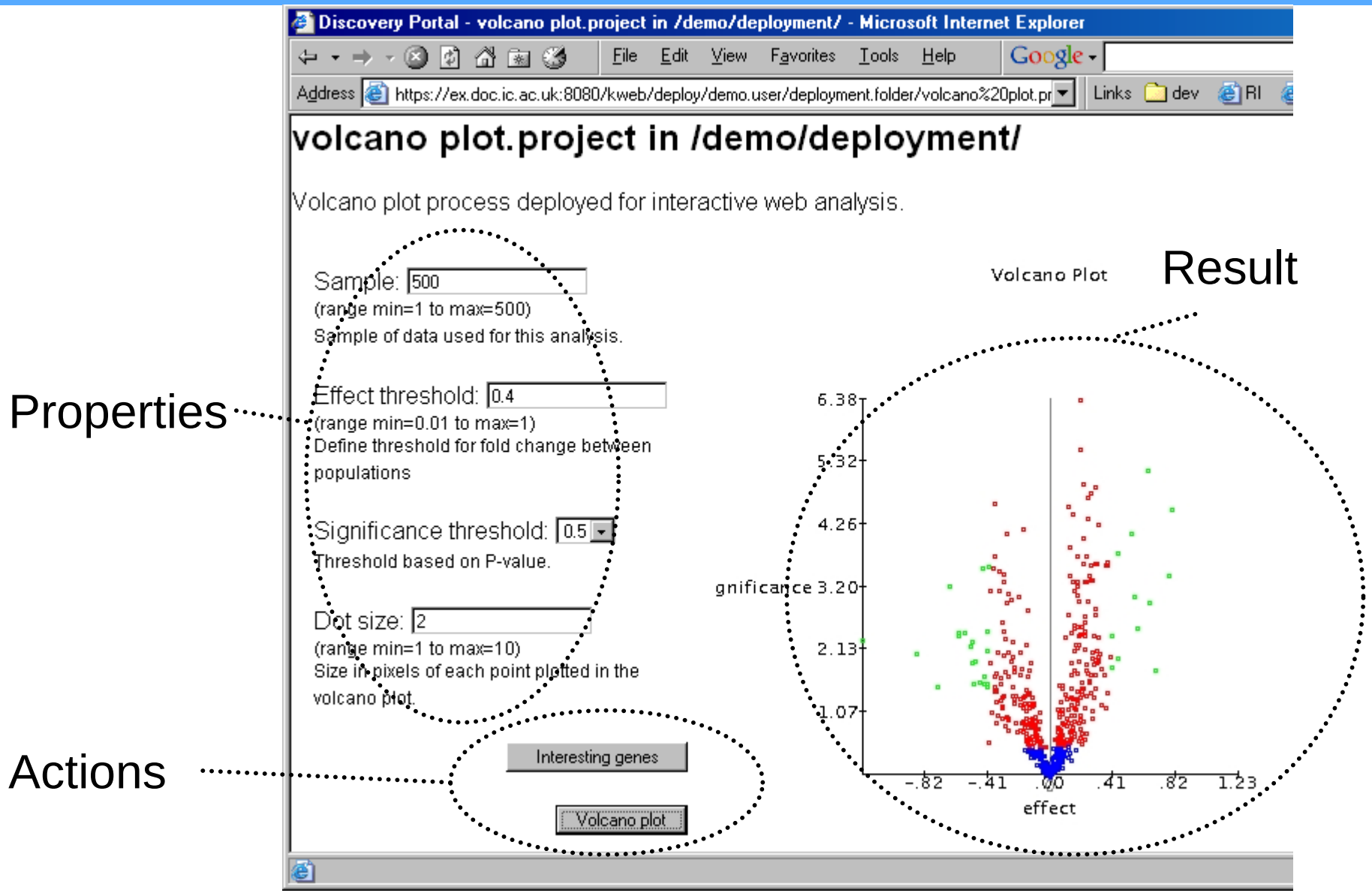
Workflow Deployment

- **Define properties**



- **Define actions**
- **Deploy**

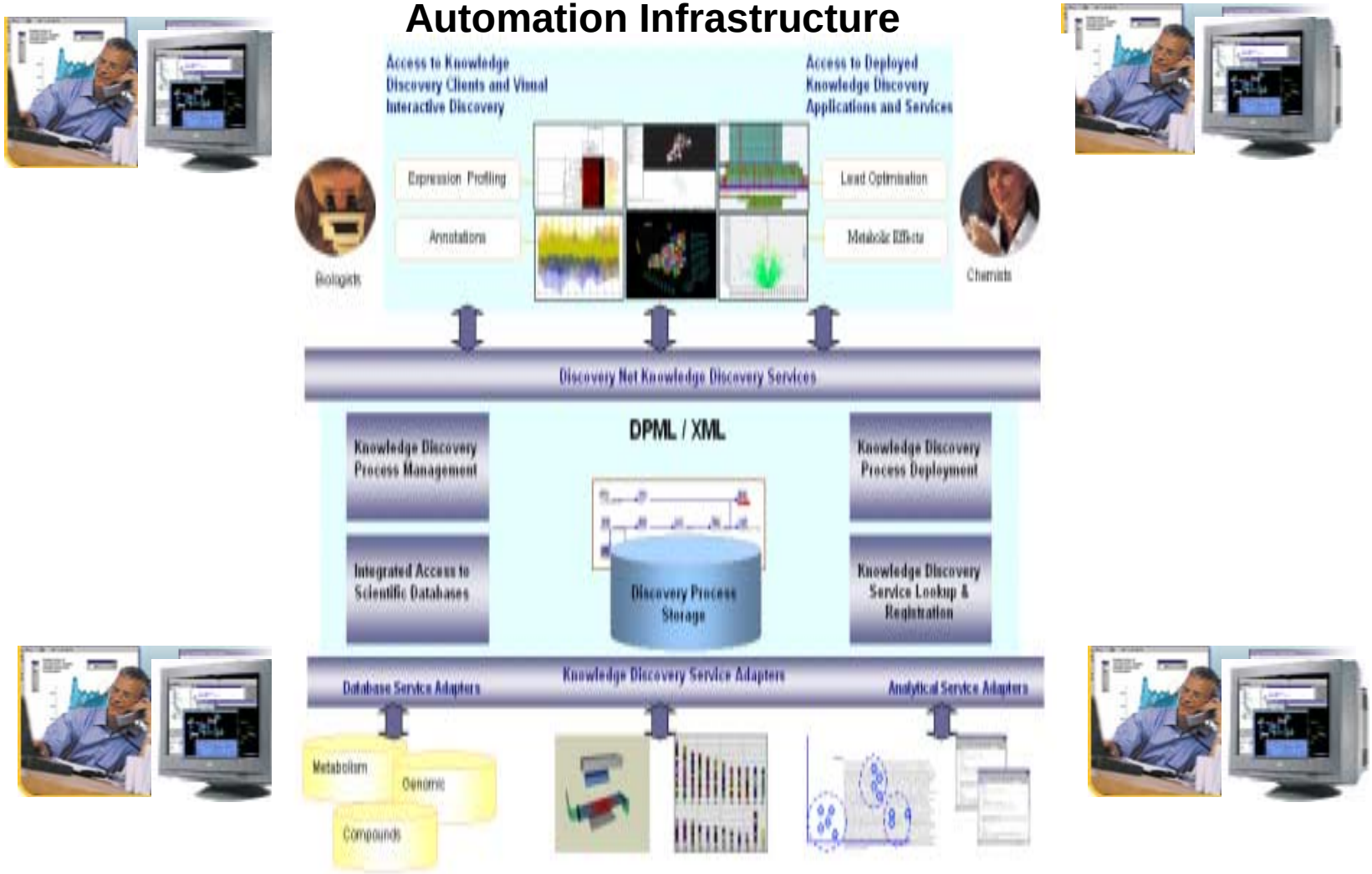
Deploying Workflow as New Service



Discovery Net-based Enterprise Knowledge Grid

Discovery Net
Unifying the World's Knowledge

Enterprise Discovery Automation Infrastructure



Case Study: SC2002 HPC Challenge

Discovery Net
Unifying the World's Knowledge

High Throughput Sequencers

Nucleotide-level Annotation

Protein-level Annotation

Process-level Annotation

Identify

Organism Chromosomes	Organism's DNA

Identify

Genes	tRNAs, rRNAs
Gene markers	Non-translated RNAs
Regulatory Regions	Repetitive Elements
Segmental Duplication	SNP Variations
Literature References	…..

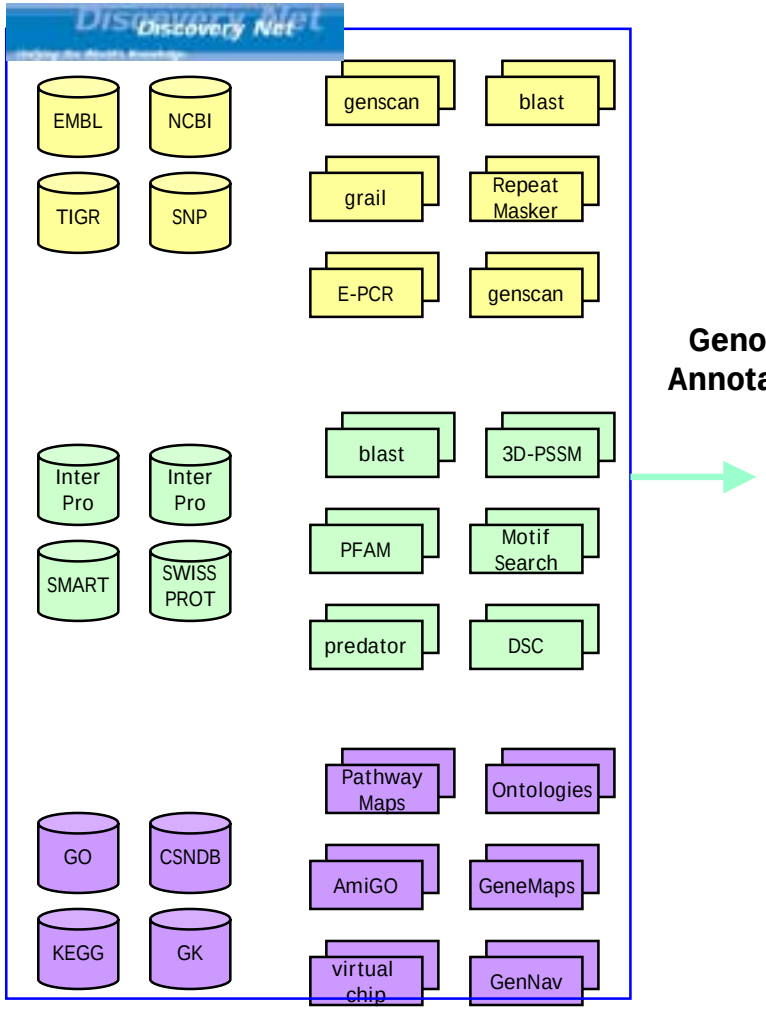
Identify

Proteins	Classify into Protein Families
Functional Characteisation	Homologues
Domain	3-D Structure
Fold Prediction	Secondary structure
Literature References	…..

Relate

Cell Cycle	Metabolism
Drugs	Biological Process…..
Cell death	Embryogenesis
Literature References	…..

Real- Time Genome Annotation



Genome Annotation

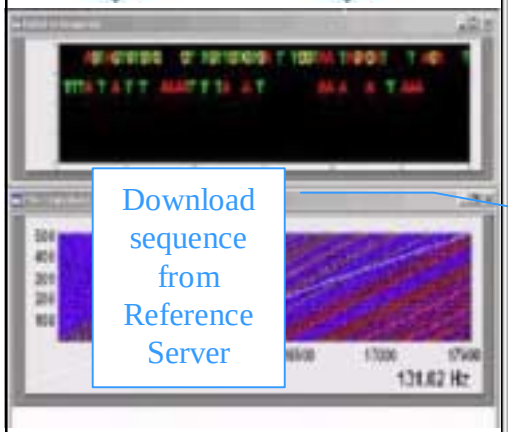
15 DBs

21 Applications

Workflow for SC2002 Competition

Discovery Net

Interactive Editor & Visualisation



Download sequence from Reference Server

Inter Pro

SMART

KEGG

EMBL

NCBI

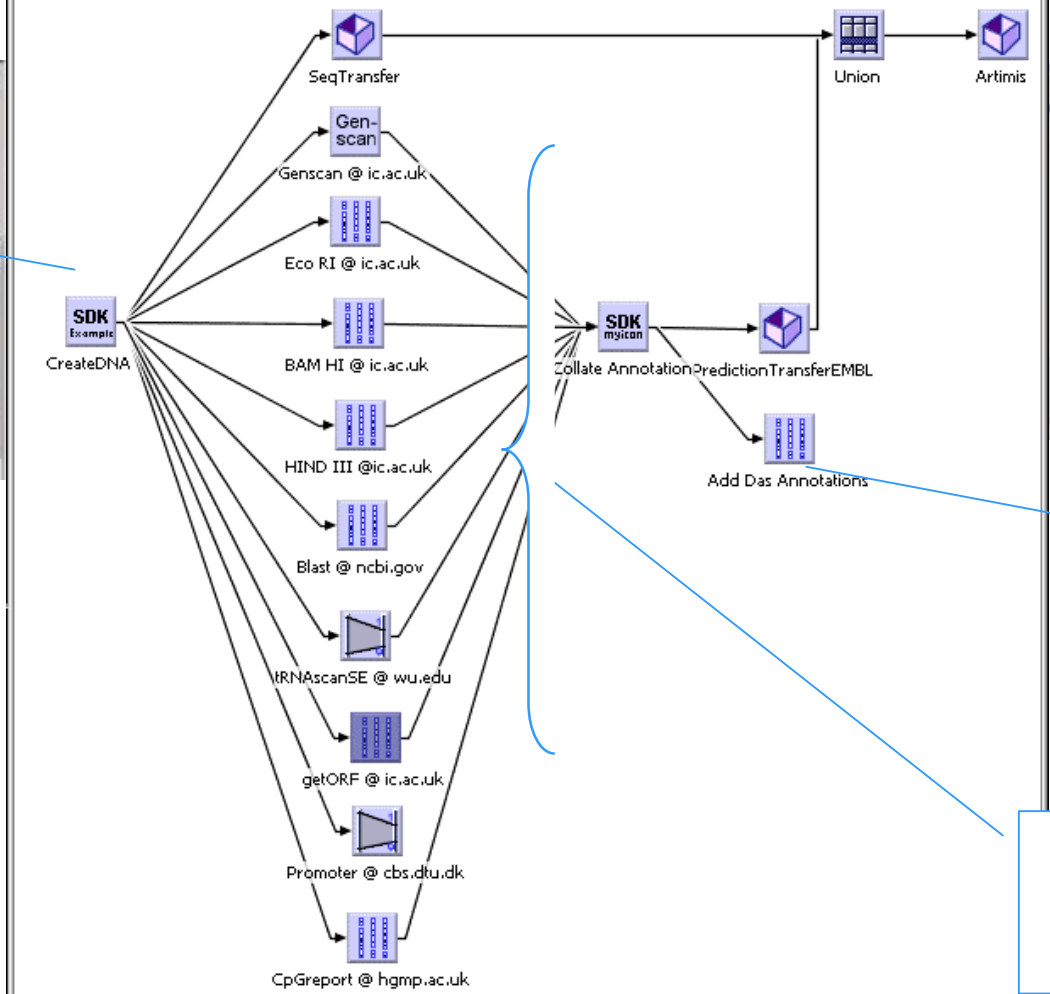
SWISS PROT

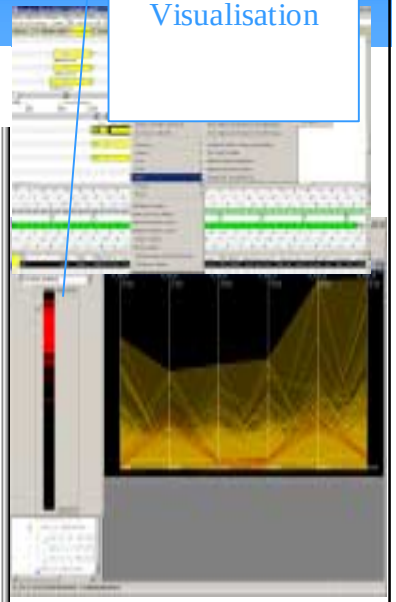
TIGR

SNP

GO

Nucleotide Annotation Workflows





Save to Distributed Annotation Server

Execute distributed annotation workflow

Discovery Net Vision