

# Distributed Data Mining in Discovery Net

Dr. Moustafa Ghanem  
Department of Computing  
Imperial College London

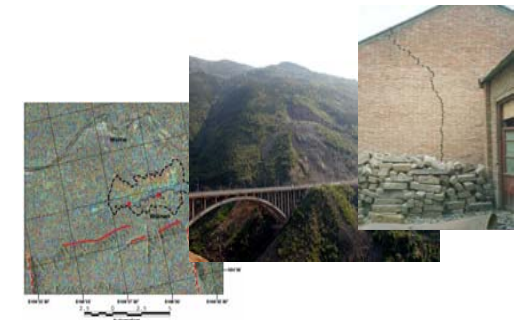
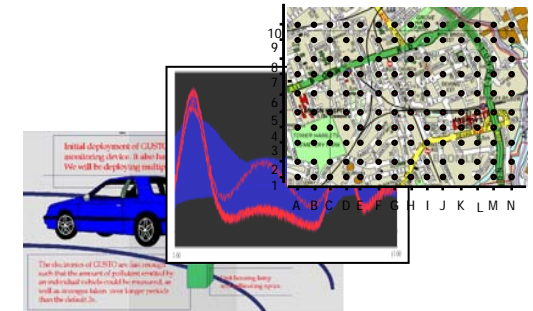
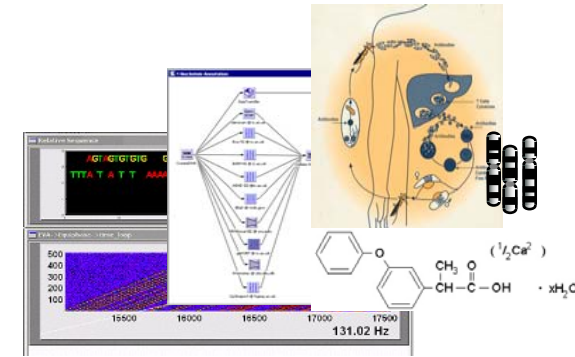
1. What is Discovery Net
2. Distributed Data Mining for Compute Intensive Tasks
3. Distributed Data Mining for Sensor Grids
4. Knowledge Discovery from Naturally Distributed Data Sources
5. What Do Scientists Really Want?

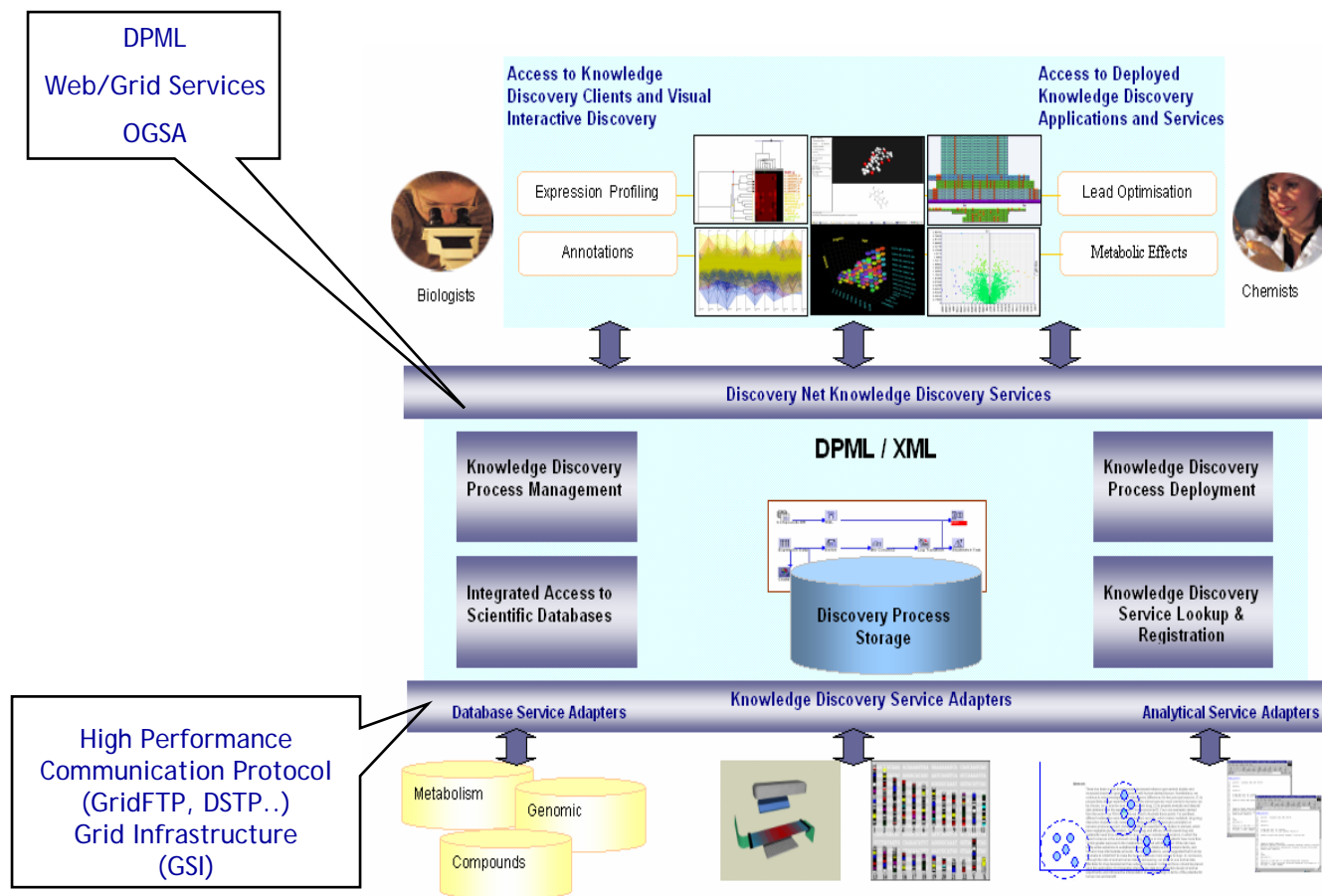
# 1. What is Discovery Net

## What is Discovery Net?

- Funding : One of the eight UK national e-science Pilot Projects funded by EPSRC (£2.2M)
- Start Oct 2001, End March 2005
- Goal :Construct the World's first Infrastructure for Global Knowledge Discovery Services
- Key Technologies:
  - Open Service Computing
  - High Throughput Devices and Real Time Data Mining
  - Real Time Data Integration & Information Structuring
  - Cross Domain Knowledge Discovery and Management
  - Discovery Workflow and Discovery Planning

- Life Sciences
  - High throughput genomics and proteomics
    - Distributed Databases and Applications
- Environmental Modelling
  - High throughput dispersed air sensing technology
    - Sensor Grids
- Real time geo-hazard modelling
  - Earthquake modelling through satellite imagery
    - High performance Distributed Computation





### D-Net Clients:

End-user applications and user interface allowing scientists to construct and drive knowledge discovery activities

### D-Net Middleware:

Provides services and execution logic for distributed knowledge discovery and access to distributed resources and services

### Computation & Data Resources:

Distributed databases, compute servers and scientific devices.

- Goal: Plug & Play
  - Data Sources,
  - Analysis Components &
  - Knowledge Discovery Processes

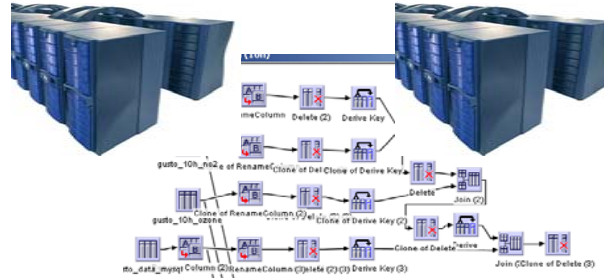
- Generic Data Mining
  - Classification, Clustering, Associations, ..
- Unstructured-Data Mining
  - Text Mining, Image Mining
- Domain-specific Mining
  - Bioinformatics, Cheminformatics, ..

- 2. Distribution of Compute Intensive Tasks
  - a. *Distributed Data Mining for Geo-hazard Prediction*

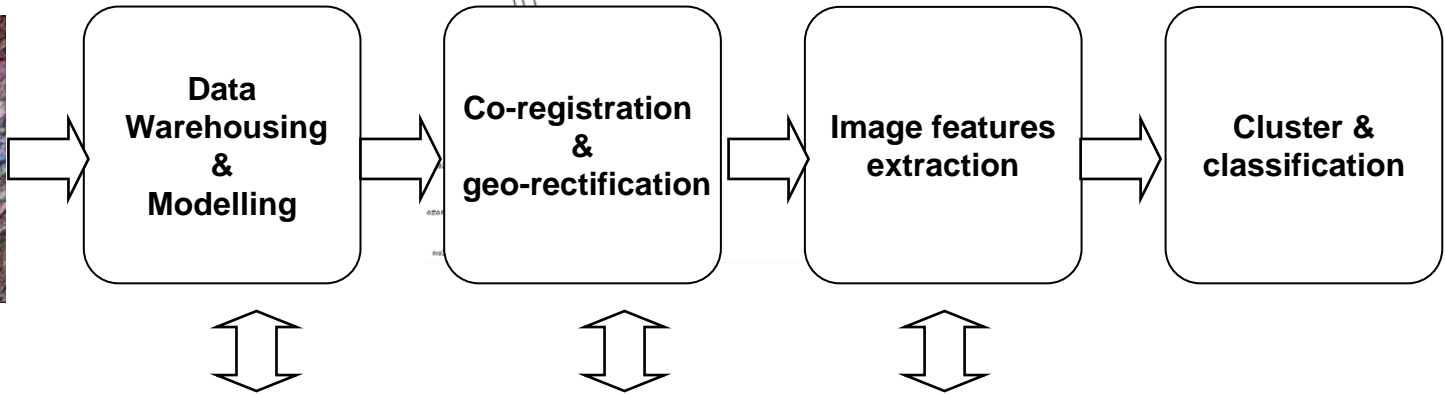
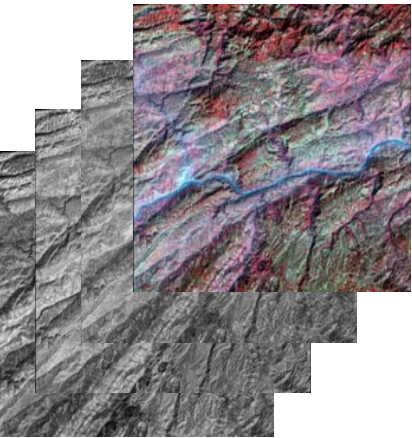


## Grid-based Geo-hazard Data Mining

- Grid-based HPC Computation
- Automatically co-register a stack of imagery layers at high precision and speed.

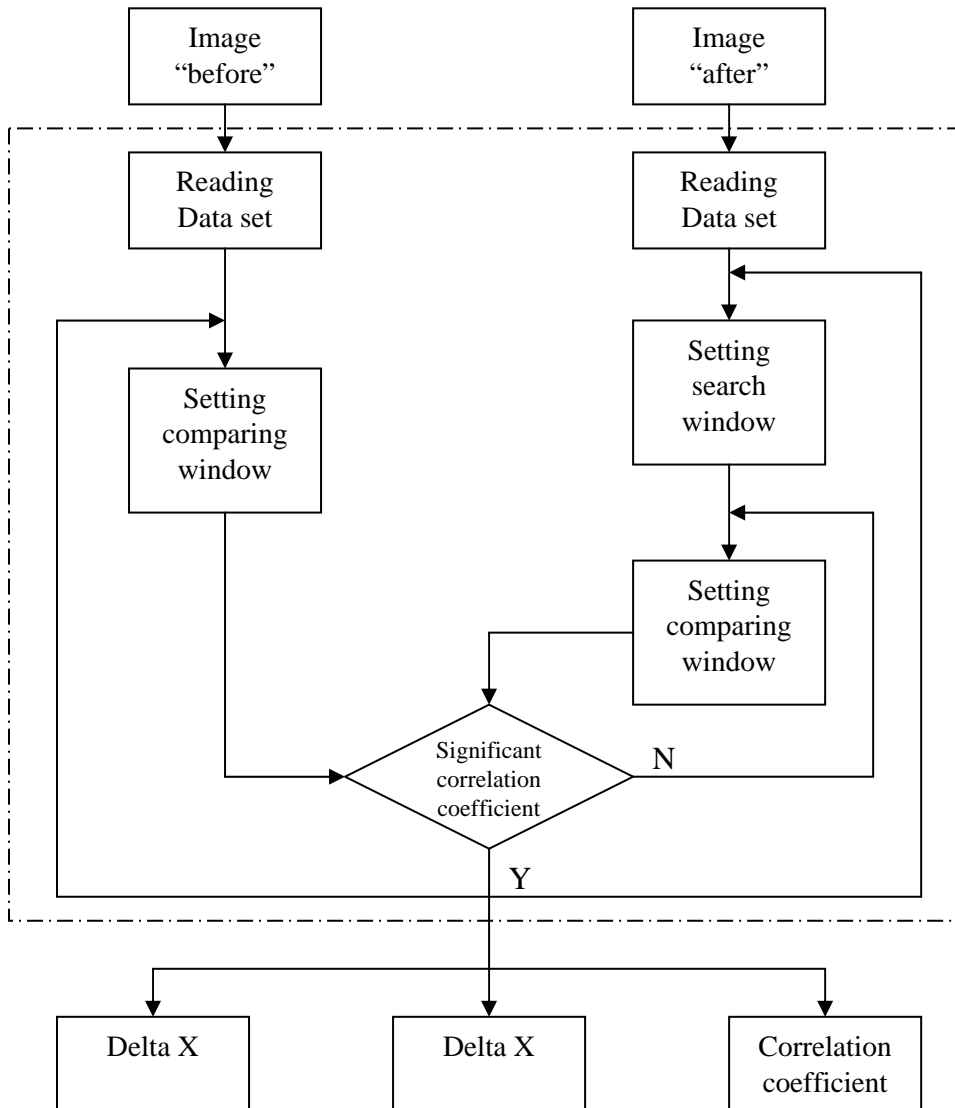


- Workflow to Coordinate Grid Computation



■ Grid-based Data Access and Integration

## Normalised cross-correlation (NCC) template algorithm



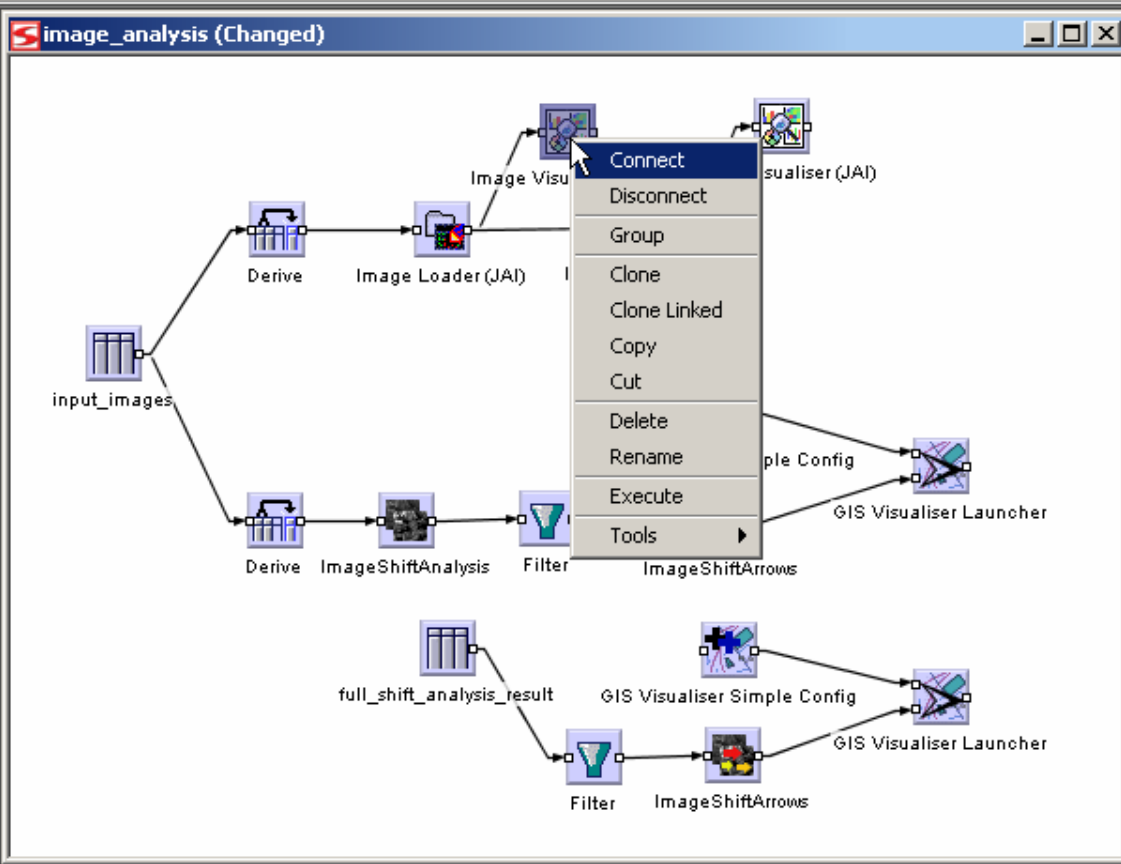
Operating on a remotely accessed MPI UNIX parallel computer through fast network with DNet interface. Slow but high accuracy: 24 processors 10 hours for one scene of Landsat-7 ETM+ Pan imagery data. The algorithm also run on GRID.

Userspace: //demo@localhost:1099

- demo
  - Geohazard
    - image\_analysis
    - view\_arrow\_file
  - gismaps
  - GM\_Royston
  - GUSTO
  - 15 min means
  - hourly\_mean
- services
  - Earth Sciences
  - GM Scenario
  - GUSTO
  - H2L Workflow
  - Plasma Physics
  - Volcano Plot
  - Interactive Browser Demo Fil...
  - Interactive Browser Demo Cr...

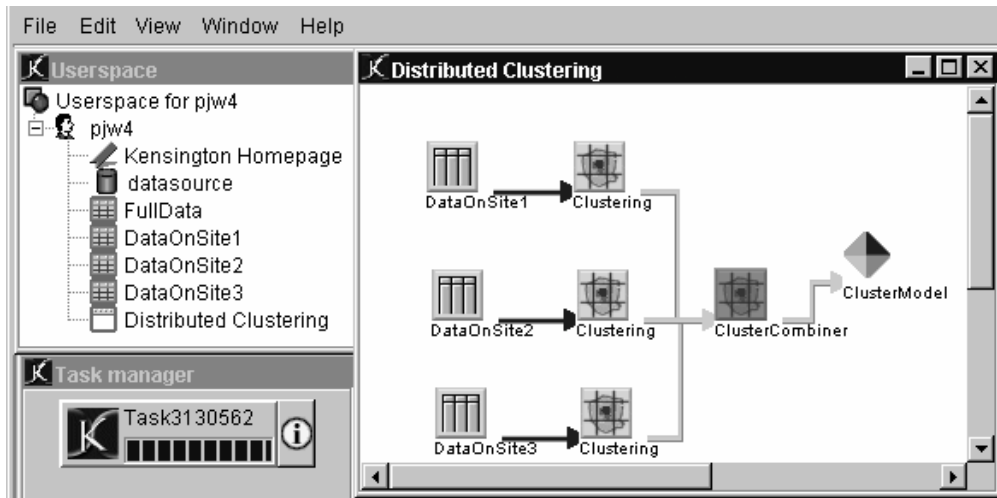
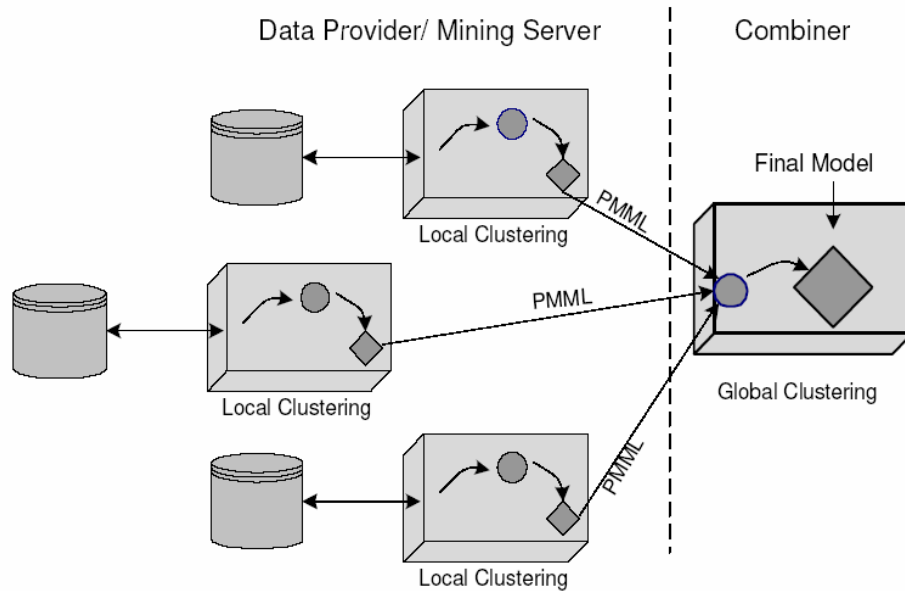
Components Task manager

- Import/Export
- Pre-process
- Normalization
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- FeatureAnalysis
- Assess
- Plasma Physics
- GeneSense



- 2. Distribution of Compute Intensive Tasks
  - b. *Distributed Clustering*

## Workflows for Distributed Data Clustering



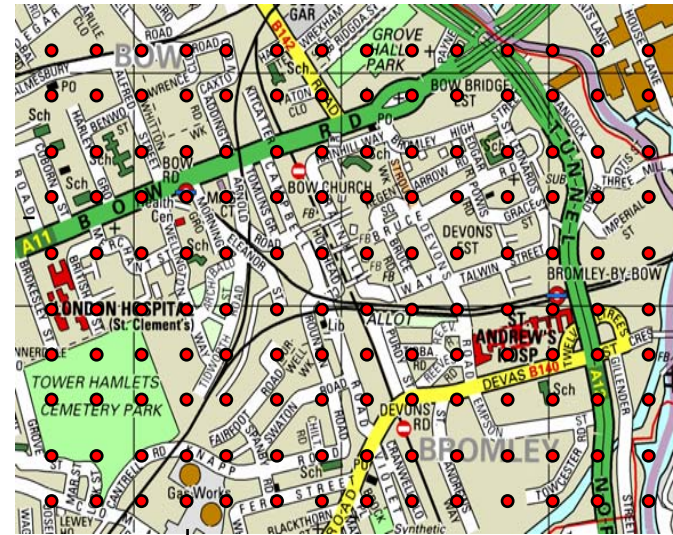
### 3. Distributed Mining over Sensor Grid Data

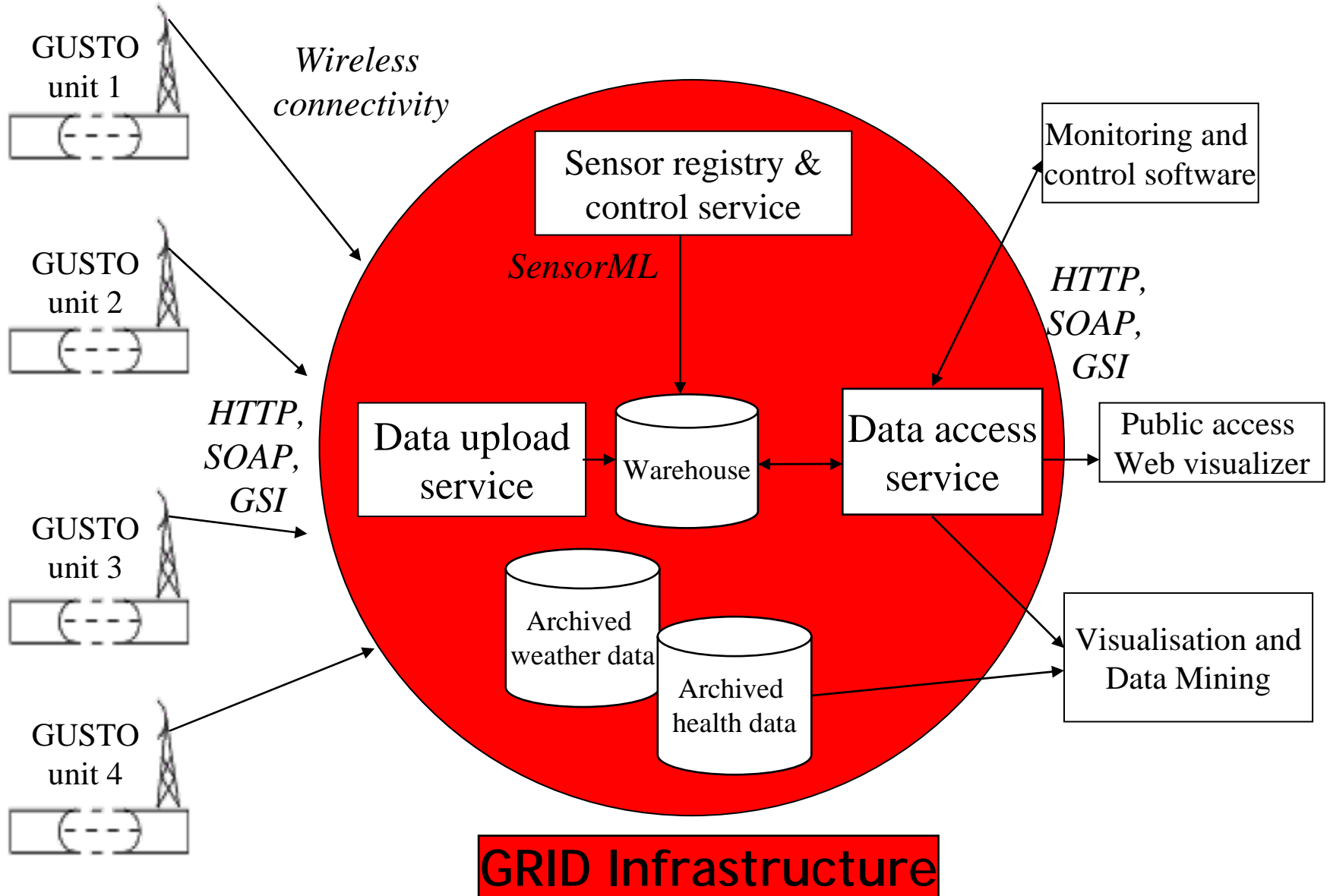
*Distributed Spatial Data Mining for Air Pollution Modelling*

# The GUSTO Project - Update

(Generic UV Sensors Technologies & Observations)

- High throughput **open path** spectrometer system
- **Robust algorithm** for pollutant concentration **retrievals**
- Measures **SO<sub>2</sub>**, **NO**, **NO<sub>2</sub>**, **O<sub>3</sub>** & **Benzene** to ppb levels every few seconds
- Geared for **networking** of multiple GUSTO units within a **GRID Infrastructure**
- Can support Remote Sensing data for (contour) **mapping** of pollutants





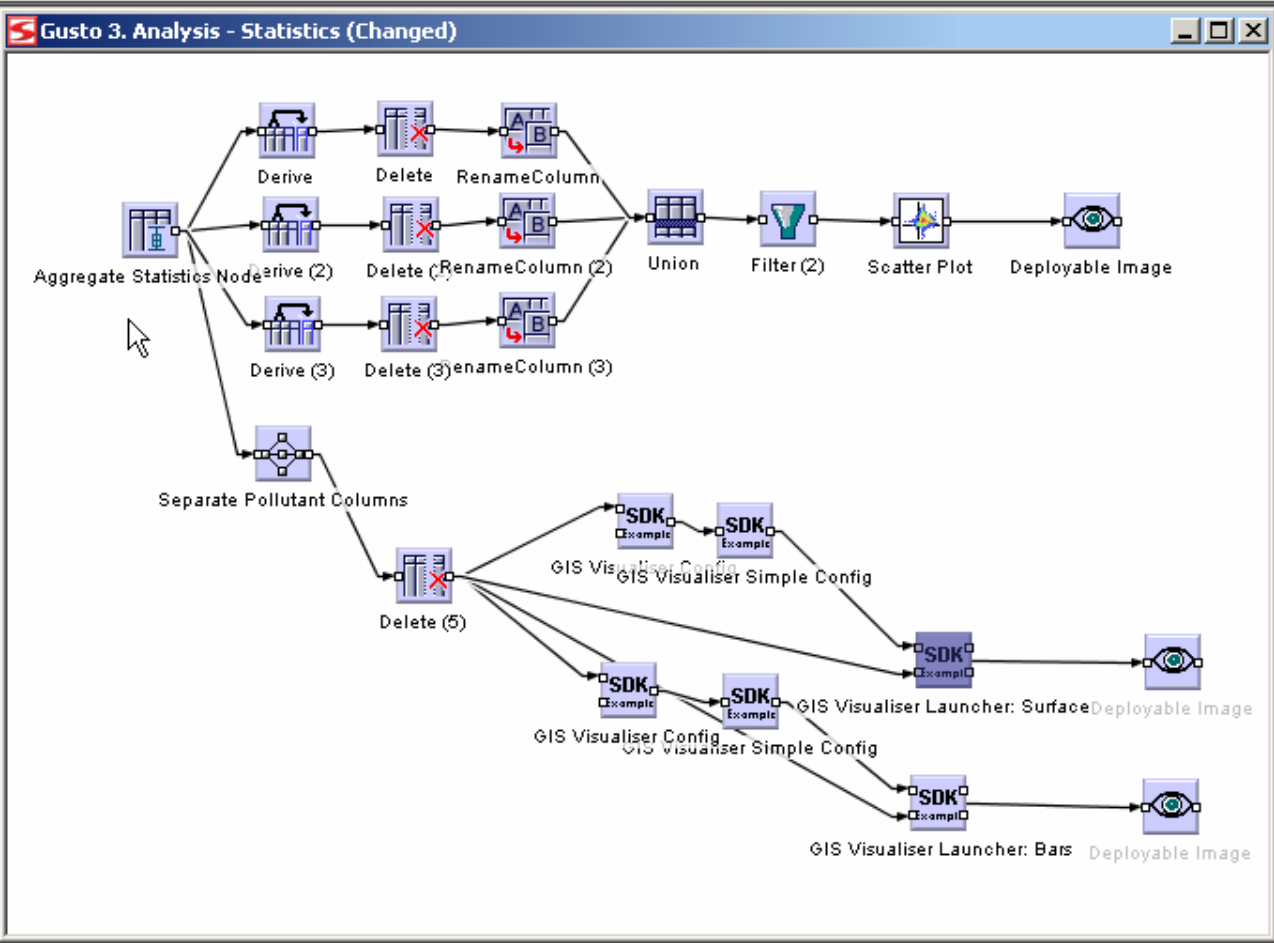


**GUSTO**

- data input
- Gusto 1. Data collection (
- Gusto 2. Visualisation (G
- Gusto 2. Visualisation (IB
- Gusto 3. Analysis - Statis**
- Gusto 4. Analysis - (incor
- 4pollutants\_normalised
- 4pollutants\_transposed
- gusto\_10h\_4pollutants
- gusto\_10h\_no
- gusto\_10h\_no2
- gusto\_10h\_ozone
- gusto\_10h\_so2
- gusto\_data\_mysql
- no2\_transposed
- no\_transposed
- ozone\_transposed
- Sensor\_coords

Components Task manager

- ImExport
- Preprocess
- Normalization
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Assess
- Oracle
- Oracle Text
- Oracle Stats
- GeneSense



**GUSTO**

- data input
- Gusto 1. Data collection (10h)
- Gusto 2. Visualisation (GIS) (Changed)**
- Gusto 2. Visualisation (IB)
- Gusto 3. Analysis - Statistics
- Gusto 4. Analysis - (incor)
- 4pollutants\_normalised
- 4pollutants\_transposed
- gusto\_10h\_4pollutants
- gusto\_10h\_no
- gusto\_10h\_no2
- gusto\_10h\_ozone
- gusto\_10h\_so2
- gusto\_data\_mysql
- no2\_transposed
- no\_transposed
- ozone\_transposed
- Sensor\_coords

**Components** | Task manager

- ImExport
- Pre-process
- Normalization
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Assess
- Oracle
- Oracle Text
- Oracle Stats
- GeneSense
- Oracle ODM

### Gusto 1. Data collection (10h)

### Gusto 2. Visualisation (GIS) (Changed)

**Label NO2**

**Label Ozone**

**Label SO2**

**Unionised Form, 4 Pollutants**

**gusto\_10h\_4pollutants**

**GIS Config - Bar Charts**

**GIS Config - Surface SO2**

**GIS - Background maps**

**GIS - Background maps**

**Surface SO2 & Bar Charts**

**Bar Charts**

**Bar Charts**

**Surface SO2 & Bar Charts**

**Surface SO2 & Bar Charts**

**DeployableImage**

**View Deployable Image**

**gusto\_10h\_no (2)**

**Derive (3)**

**Delete (4)**

**Pivot**

**no\_transposed**

**gusto\_10h\_so2 Copy of Derive (3) of Delete (4) of Pivot**

**so2\_transposed**

**gusto\_10h\_ozone (2) of Derive (3) of Delete (4) of Pivot**

**ozone\_transposed**

**gusto\_10h\_no2 (2) of Derive (3) of Delete (4) of Pivot**

**no2\_transposed**

**All 4 Transposed**

**Properties editor [GIS Visualiser Launcher] Surface SO2 & Bar Charts**

Parameters | Input | Output | Cache | History | CRISP-DM | Notes

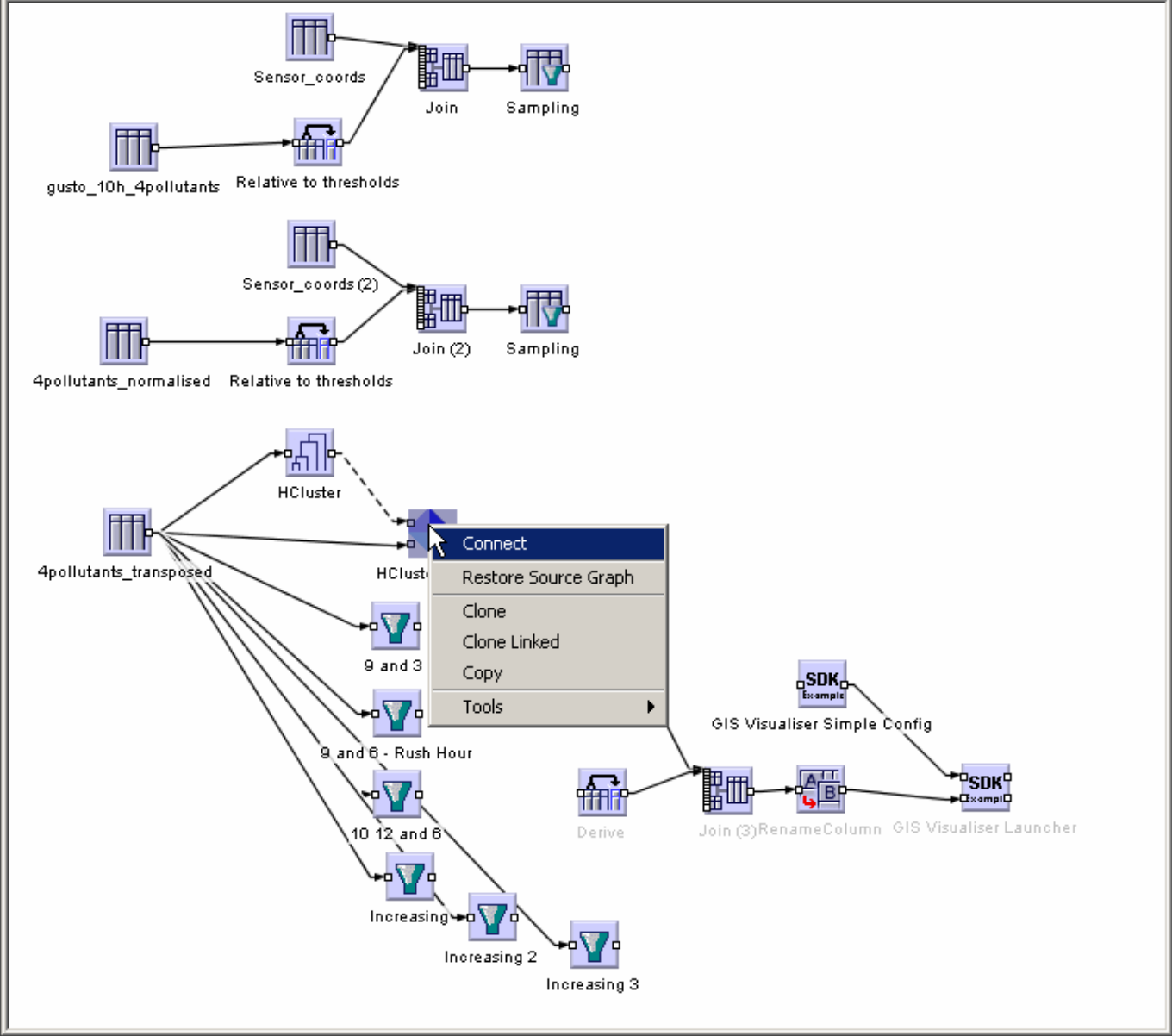
**GUSTO**

- data input
- Gusto 1. Data collection (
- Gusto 2. Visualisation (G
- Gusto 2. Visualisation (IB**
- Gusto 3. Analysis - Statis
- Gusto 4. Analysis - (incor
- 4pollutants\_normalised
- 4pollutants\_transposed
- gusto\_10h\_4pollutants
- gusto\_10h\_no
- gusto\_10h\_no2
- gusto\_10h\_ozone
- gusto\_10h\_so2
- gusto\_data\_mysql
- no2\_transposed
- no\_transposed
- ozone\_transposed
- Sensor\_coords

**Components** | Task manager

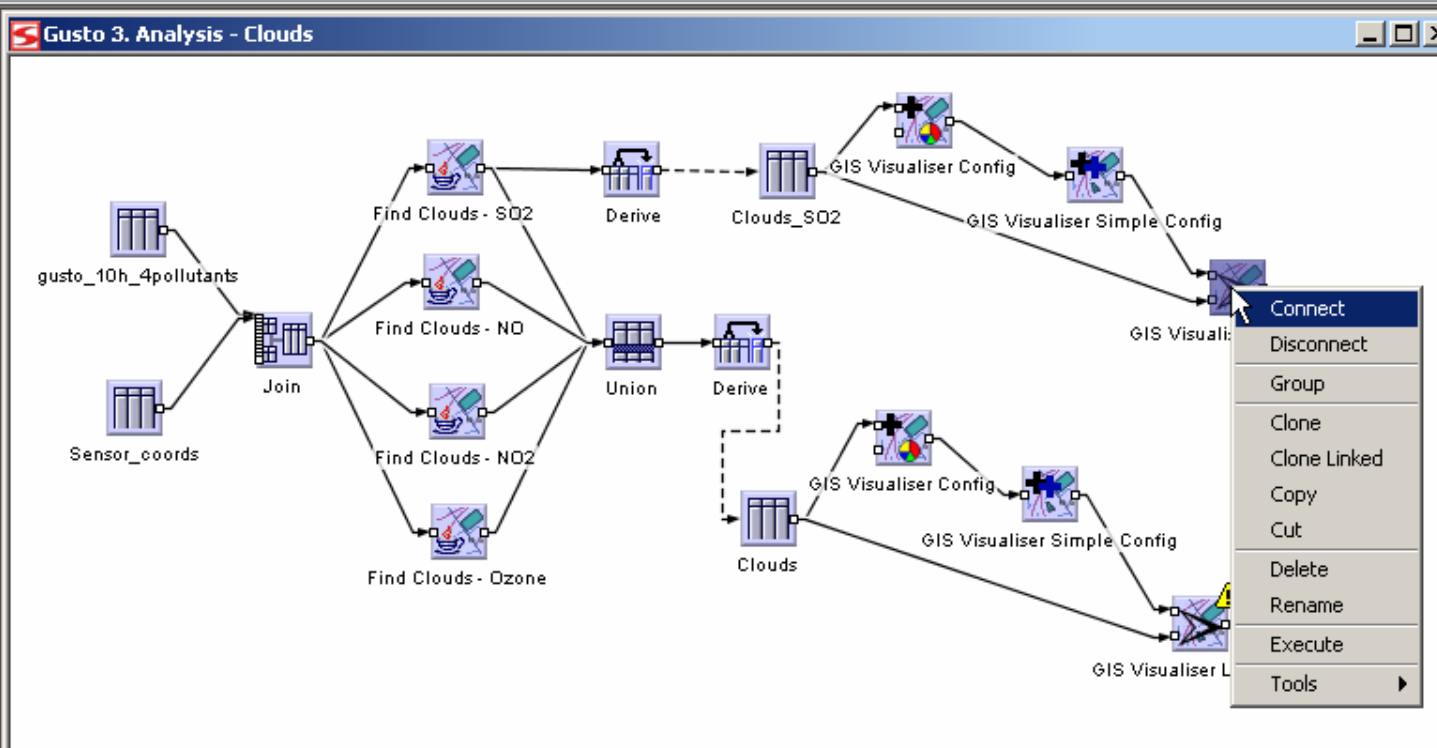
- ImExport
- Pre-process
- Normalization
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Assess
- Oracle
- Oracle Text
- Oracle Stats
- GeneSense

**Gusto 2. Visualisation (IB) (Changed)**



**GUSTO**

- data input
- Gusto 1. Data collection (...
- Gusto 2. Visualisation (G...
- Gusto 2. Visualisation (IB...
- Gusto 3. Analysis - Cloud**
- Gusto 3. Analysis - Statis...
- 4pollutants\_normalised
- 4pollutants\_transposed
- Clouds
- Clouds\_SO2
- gusto\_10h\_no
- gusto\_10h\_no2
- gusto\_10h\_ozone
- gusto\_10h\_so2
- gusto\_15min\_means
- gusto\_data\_mysql
- no2\_transposed
- no\_transposed



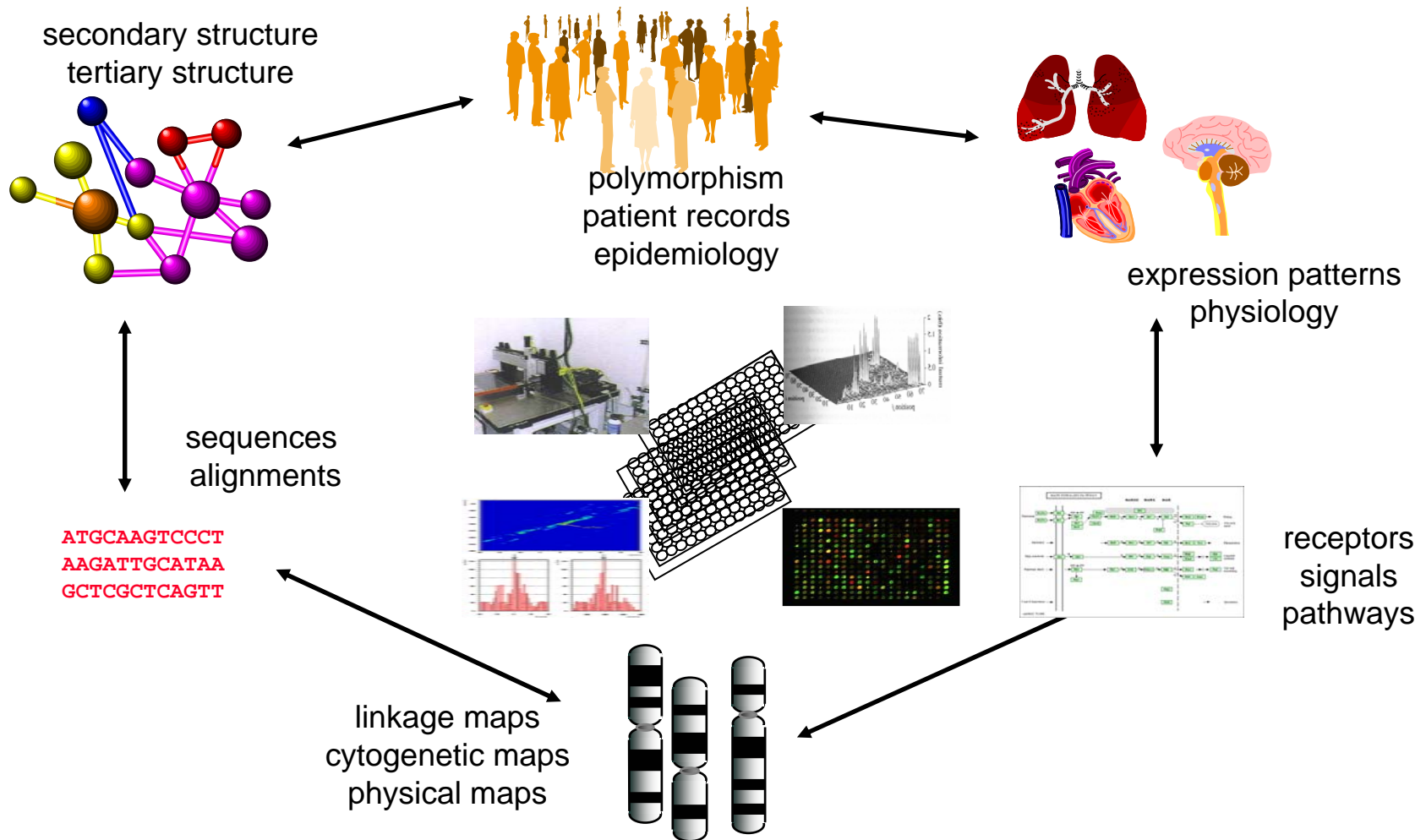
- Connect
- Disconnect
- Group
- Clone
- Clone Linked
- Copy
- Cut
- Delete
- Rename
- Execute
- Tools

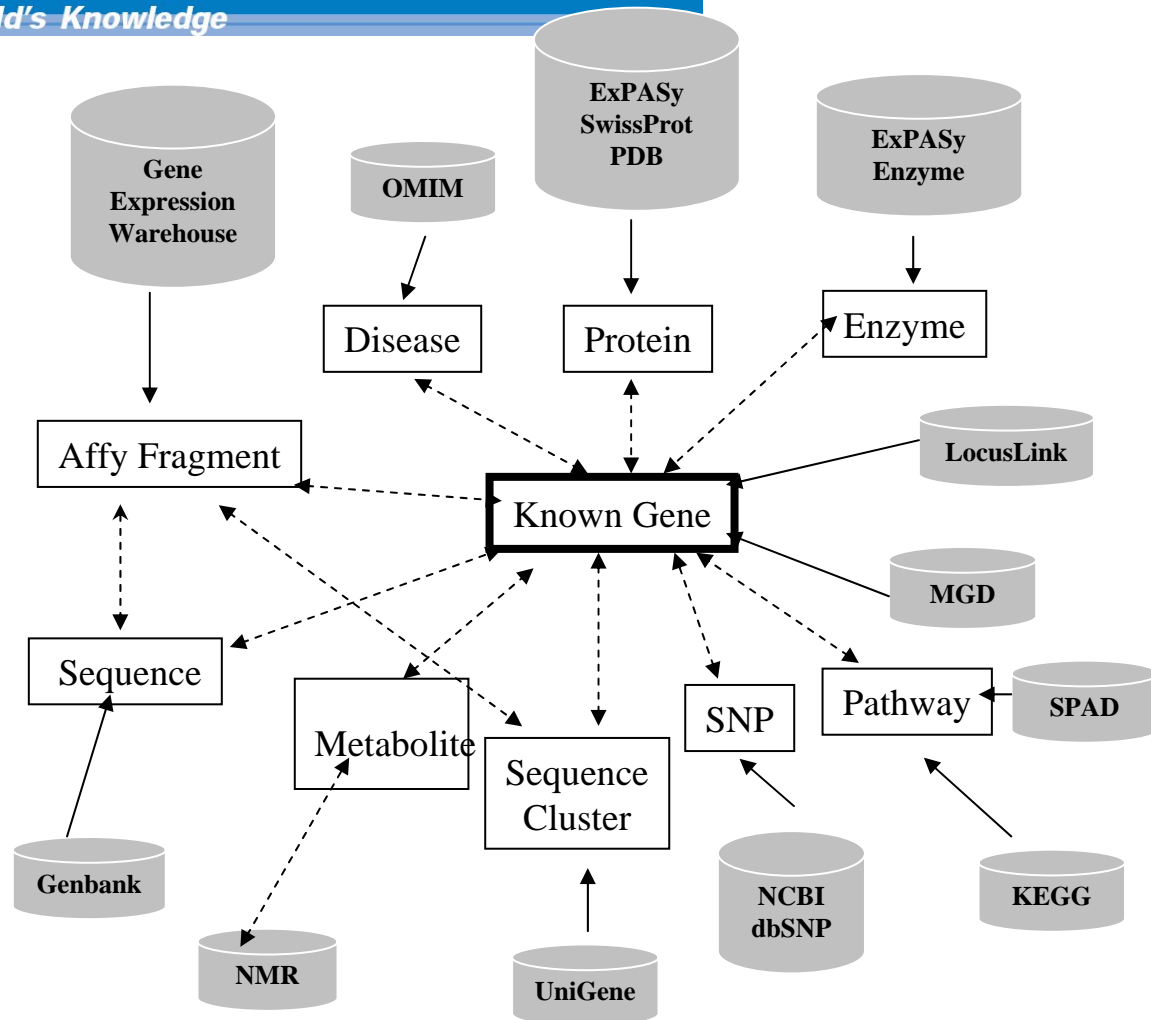
**Components** | Task manager

- Import/Export
- Pre-process
- Normalization
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Feature Analysis
- Assess
- Plasma Physics
- GeneSense

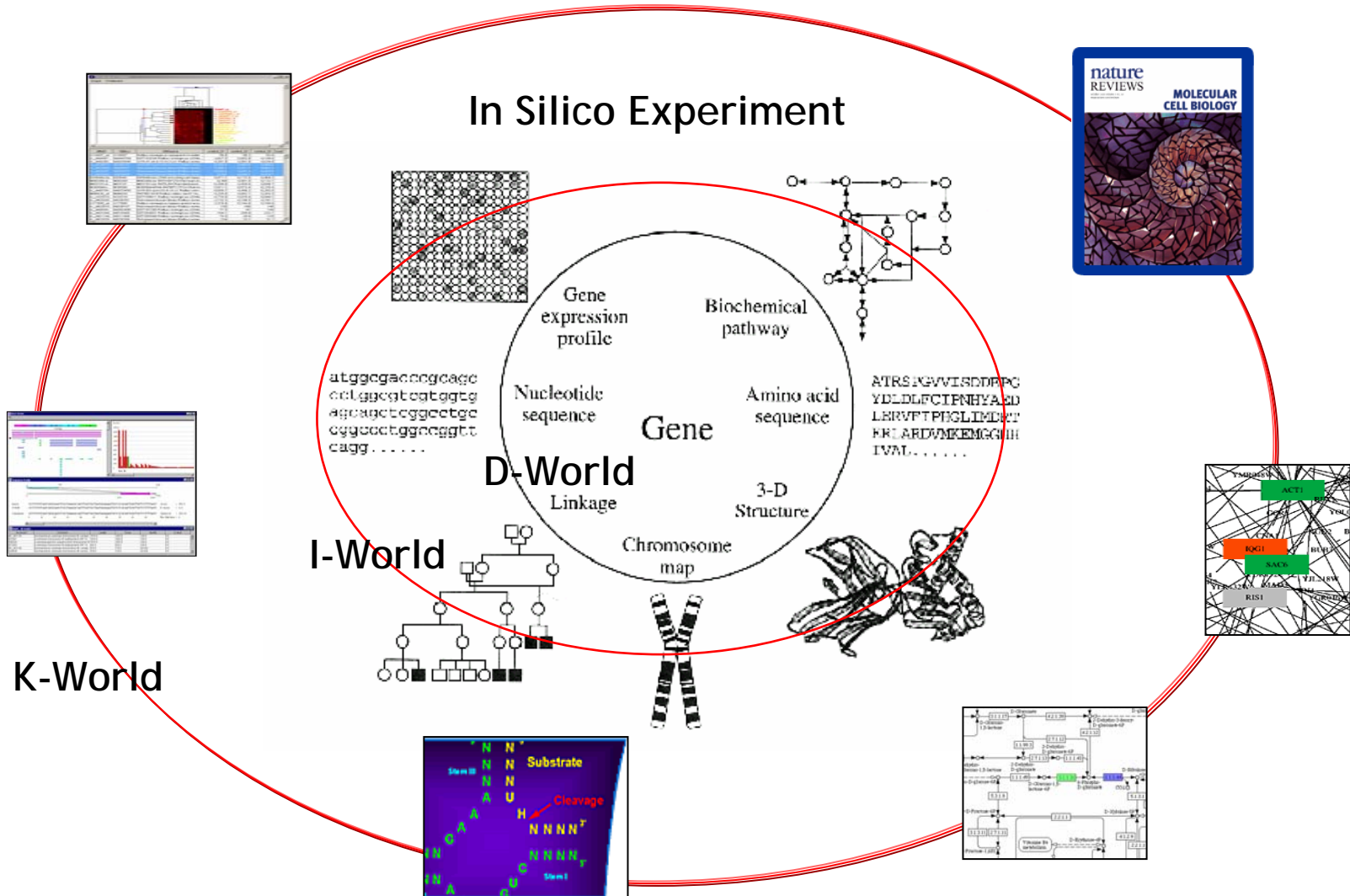
## 4. Knowledge Discovery from Naturally Distributed Data Sources

*Distributed Data Mining in Life Sciences*

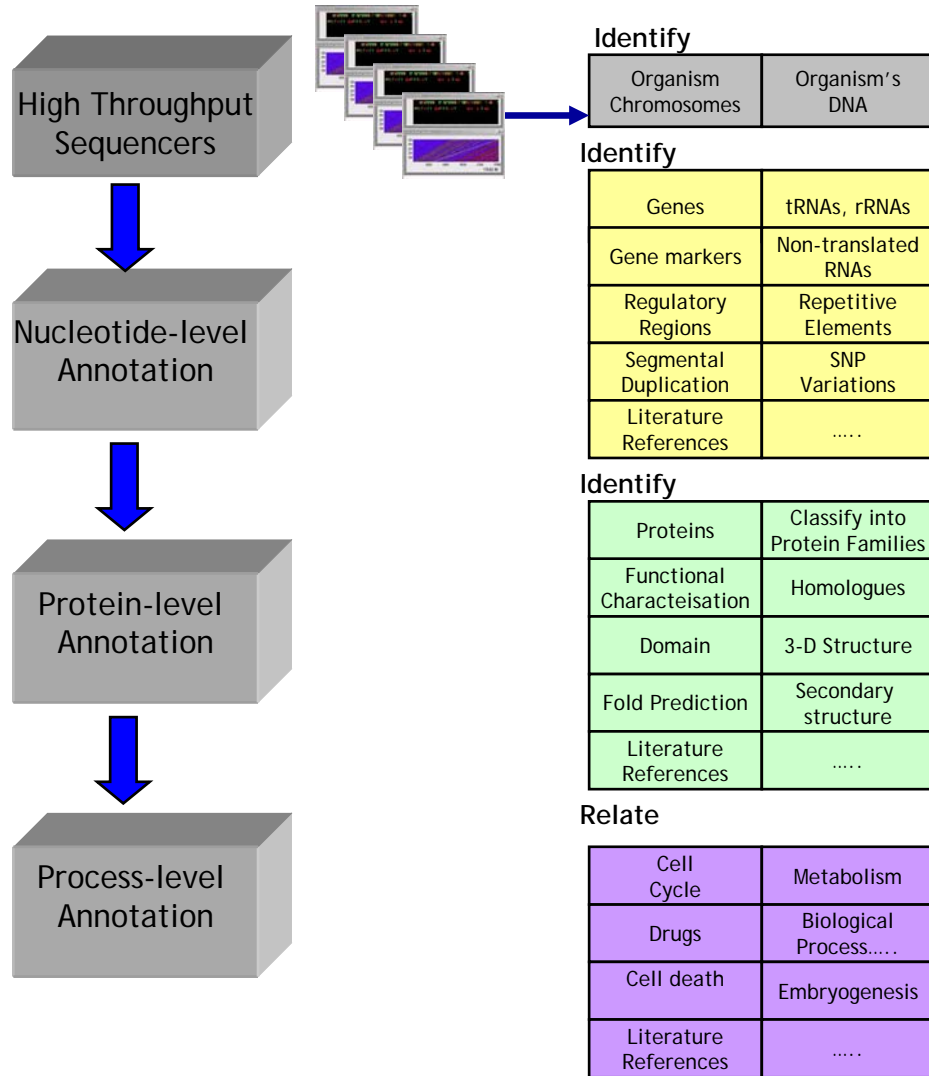




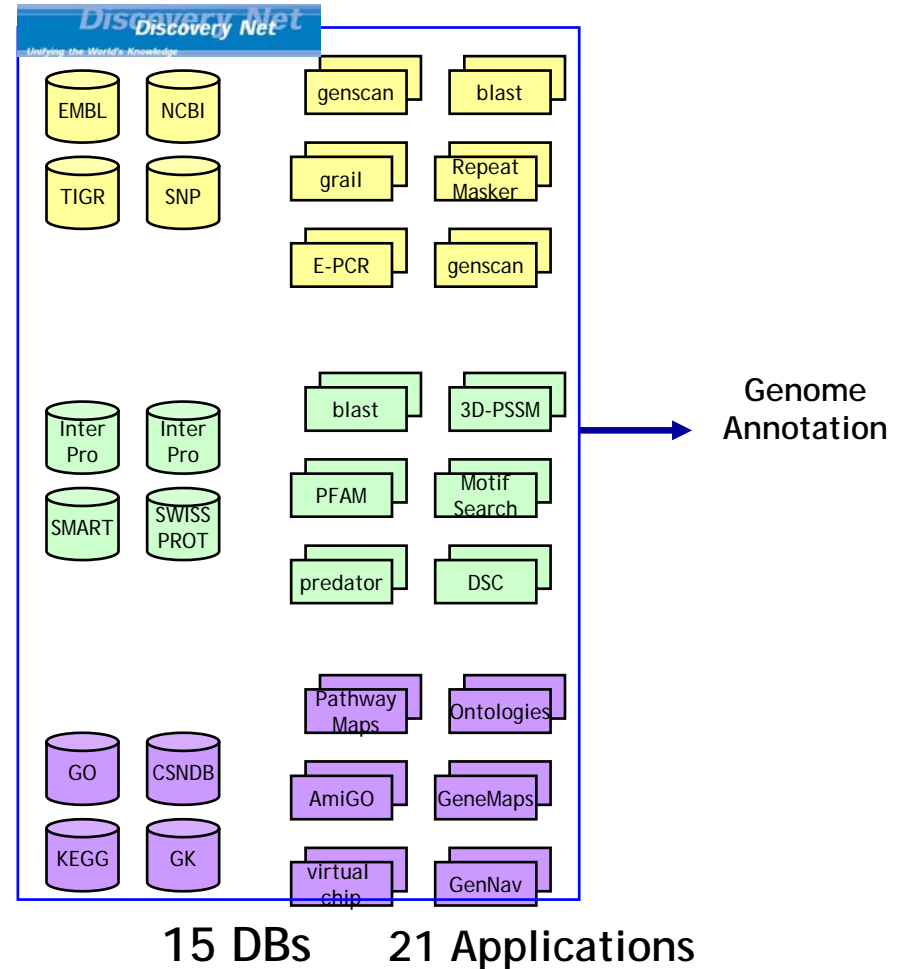
- Given a collection of microarray generated gene expression data, what kind of questions the users wish to pose.
- Design an integration schema?







### D-Net based Global Collaborative Real-Time Genome Annotation



SC2002 Conference



November 16-22

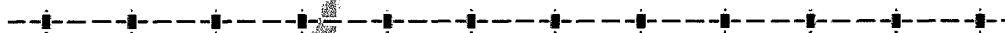


Baltimore, Maryland, USA

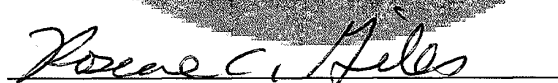
# *SC2002 HPC Challenge Award*

*Most Innovative Data-Intensive Application*

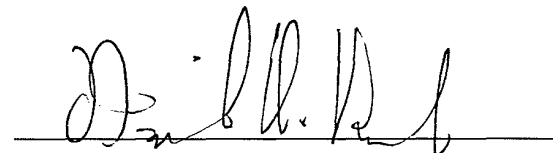
*Yike Guo, John Hassard, Tony Cass, John Darlington, Jian Guo Liu, Daniel  
Rueckert, Moustafa Ghanem, Martin Kohler,  
Antony Rowe and Patrick Wendel*



*Discovery Net: High Throughput Global  
Wide Knowledge Discovery Services*

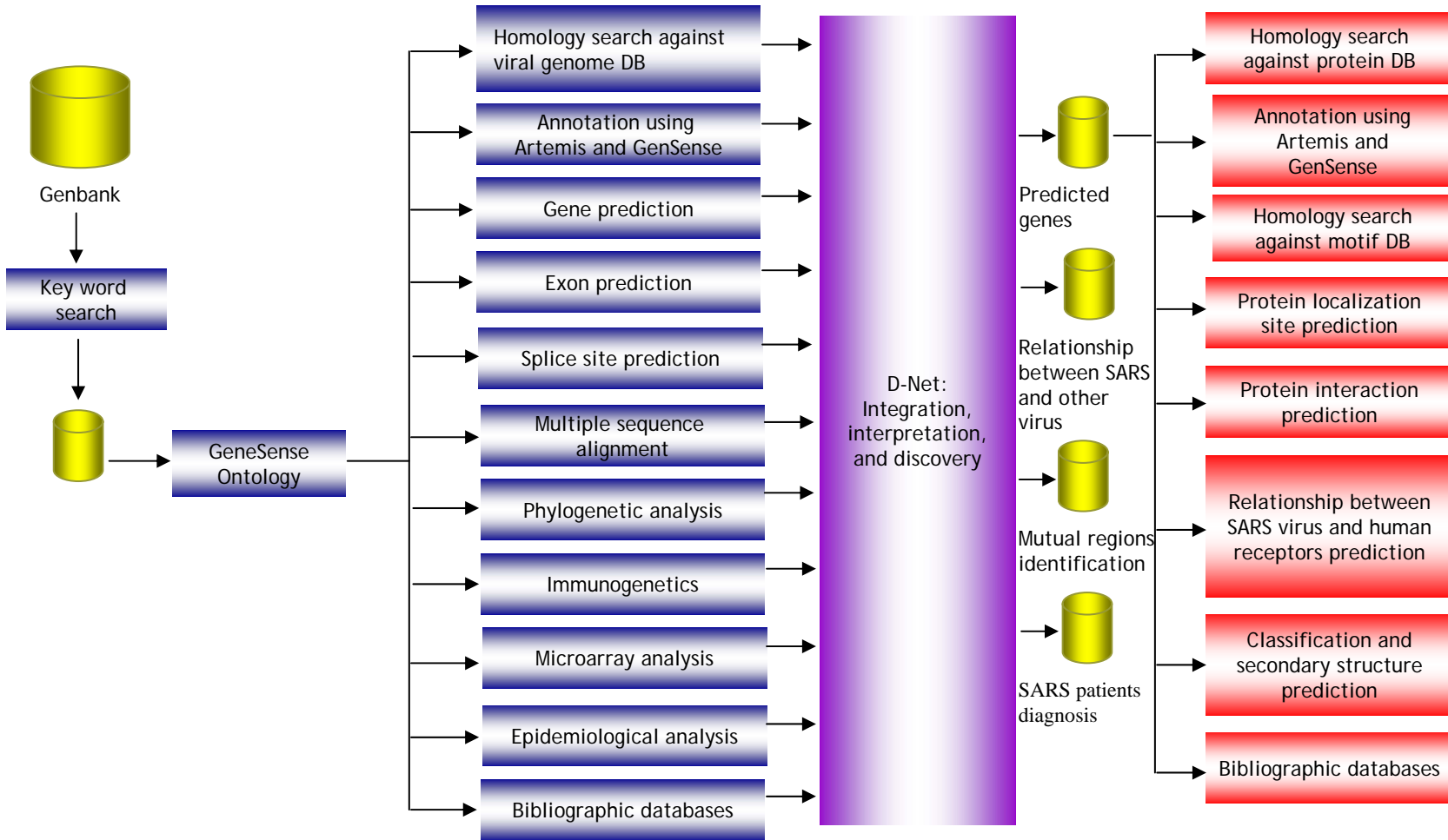


*SC2002 Conference Chair*

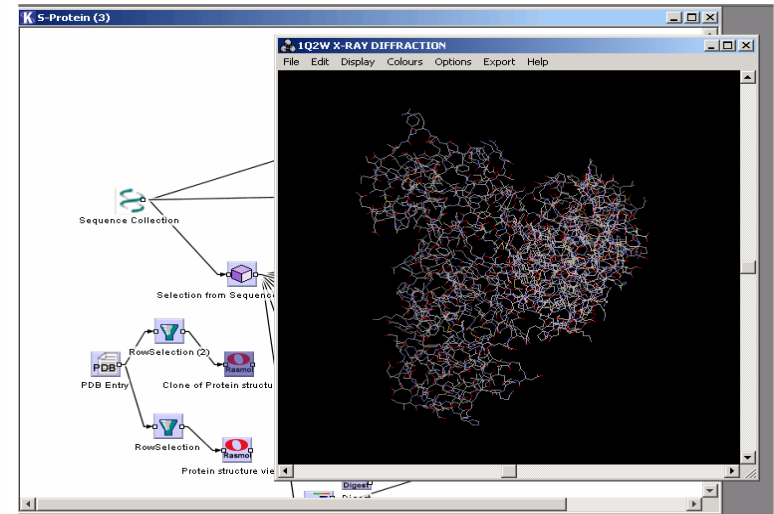
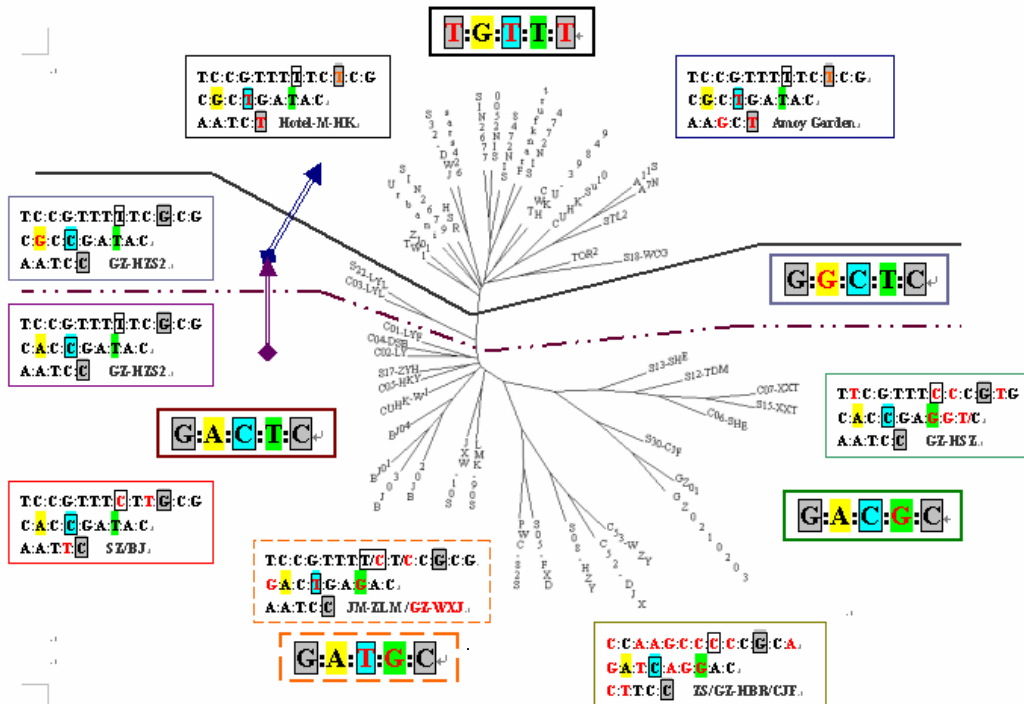


*SC2002 Program Chair*

## Discovery Net in Action: China SARS Virtual Lab

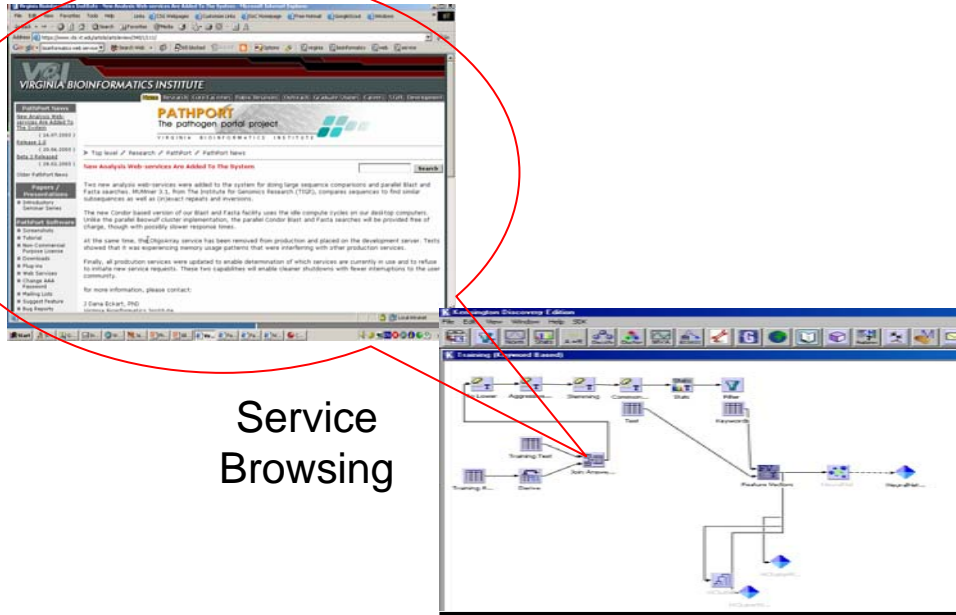


## Discovery Net in Action: SARS Virus Mutation Analysis



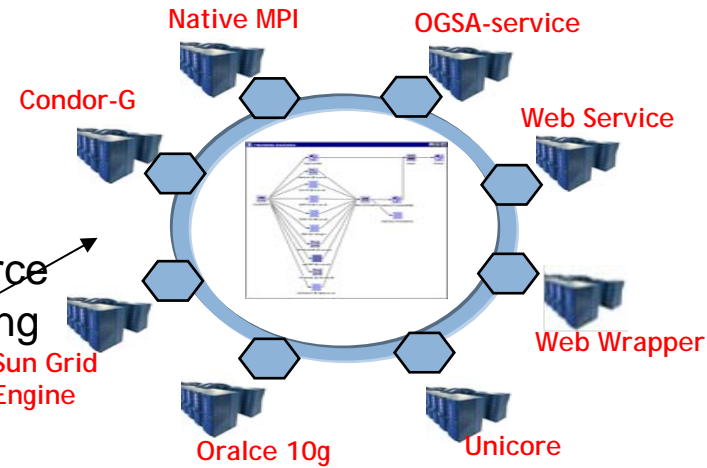
5. What do Scientist Really Want?  
*Does it really work?*

# Towards Compositional Grid Services



Service Browsing

Resource Mapping

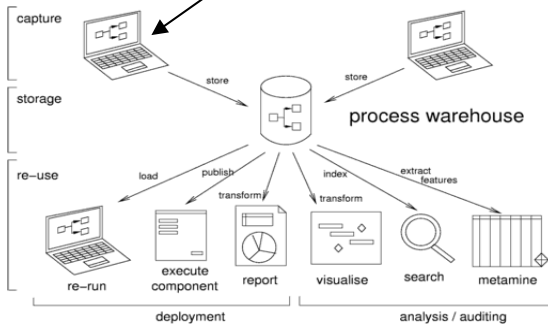


Workflow Execution  
A compositional GRID

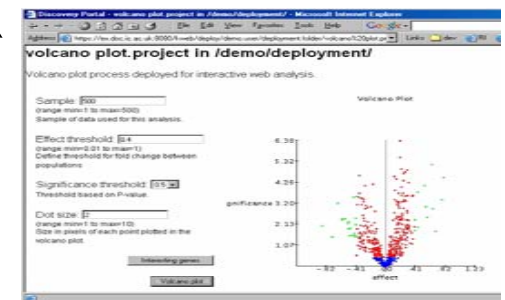
## Workflow Authoring Composing services

Workflow Warehousing

Service Abstraction



Workflow Management  
Collaborative Knowledge Management



Workflow Deployment:  
Grid Service and Portal

The screenshot displays the Discovery Net Service Composition environment. The main workspace, titled "Untitled Project (1) (Changed)", contains two service components: "seq\_id" and "Web Service". The "seq\_id" component is represented by a grid icon, and the "Web Service" component is represented by a "WS" icon. A mouse cursor is positioned over the "seq\_id" component.

On the left side, there is a file explorer showing the project structure under "Userspace: //demo@localhost:3000". The structure includes folders for "demo", "ClustaWDEMO", "Deployment", "Sars Prediction", "SARS\_Protein\_Alignments", and "Web Services". Under "Web Services", there are sub-components like "PathPort service (2)", "PathPort service", "test WebService node - S", "gff", "seq\_id", and "Stock\_symbols". Below these are "complete", "alldata", "SARS deployed workflow", "SARS Prediction", "SARS1", and "SARS3".

At the bottom left, a "Components" palette lists various tools and services, including "BioScience\_Nucleotide\_Tools", "BioScience\_Protein\_Tools", "BioScience\_Remote\_Query", "BioScience\_Visualisations", "BioScience\_Tools", "WebServices", "Oracle", "Grid", "Oracle Stats", and "GeneSense". The "WebServices" component is currently selected.

At the bottom center, a "Properties editor" window is open, displaying the text "Properties editor".

At the bottom right, a "Navigator" window is visible, showing a small diagram of the service composition.

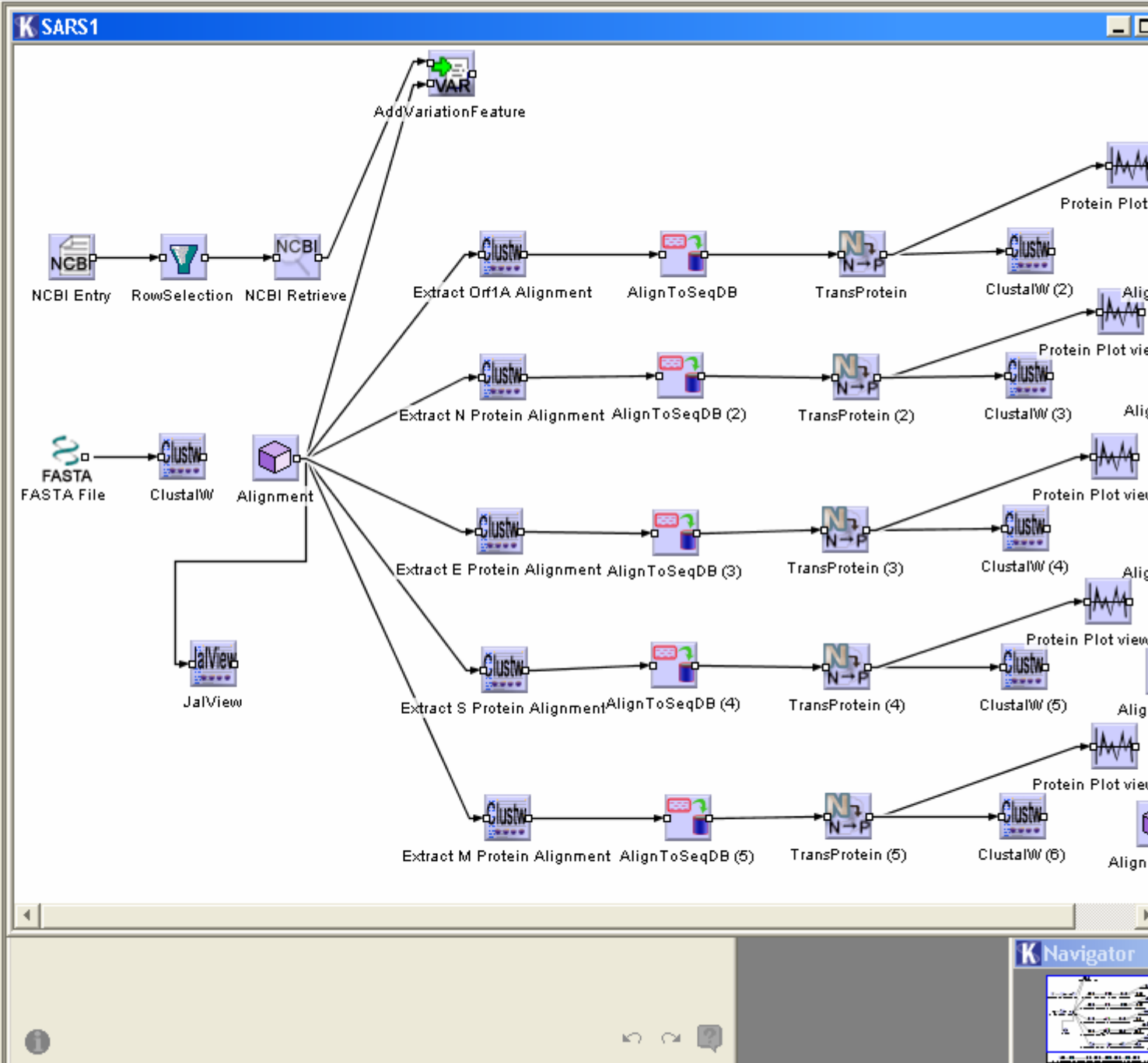
The Windows taskbar at the bottom shows the Start button, several application icons, and the system tray with the time "3:17 PM". The taskbar also displays the current directory path: "C:\cygdrive\c\Kensi...".

Userspace: //demo@localhost:3000

- demo
  - ClustaMDEMO
  - Deployment
  - Sars Prediction
  - SARS\_Protein\_Alignments
  - Web Services
  - complete
    - alldata
    - SARS deployed workflow
    - SARS Prediction
    - SARS1**
    - SARS3
    - SARS4
    - SARS5
    - WSDL test
    - sProtein
    - gff2

Components Task manager inforsense.net

- BioScience\_Alignment
- BioScience\_Nucleotide\_Tools
- BioScience\_Protein\_Tools
- BioScience\_Remote\_Query
- Annotation Based Search
- COGSearch
- NCBI Entry
- NCBI Retrieve**
- PDB Entry
- PSORTSearch
- SMARTSearch
- SwissProt SeqDB Retrieve
- SwissProt&TrEMBL Entry



K Navigator

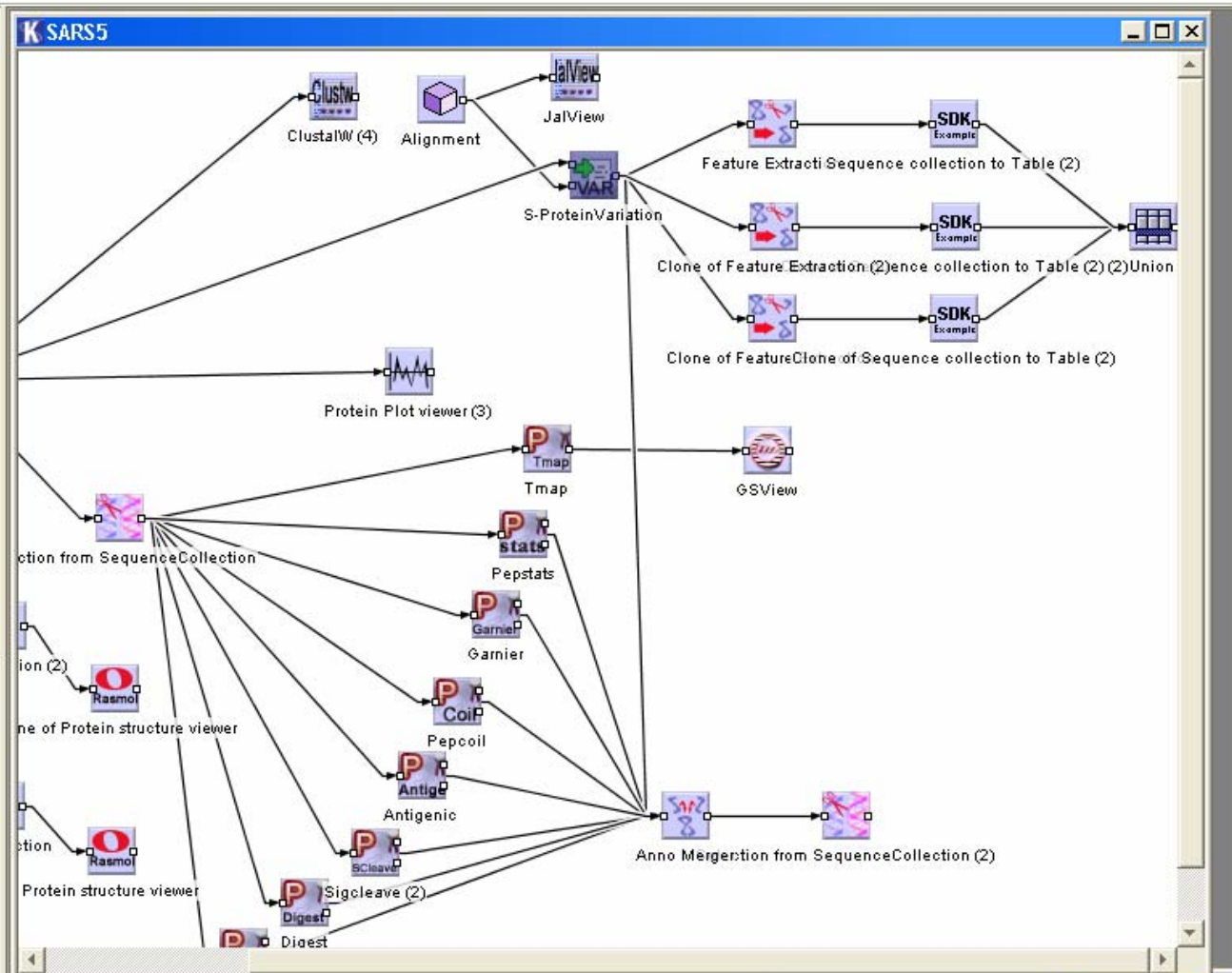


Userspace: //demo@localhost:3000

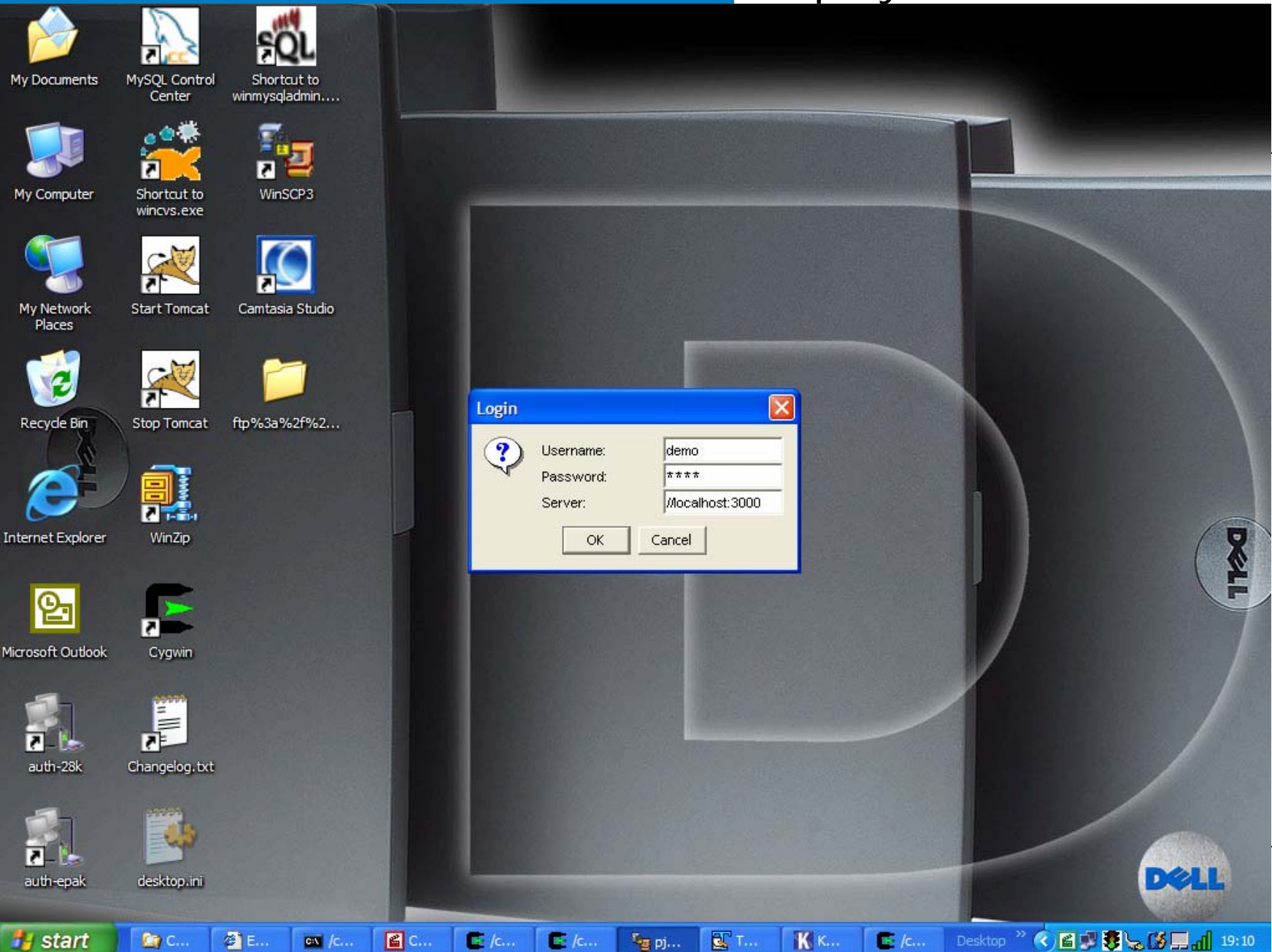
- demo
  - ClustaWDEMO
  - Deployment
  - Sars Prediction
  - SARS\_Protein\_Alignments
  - Web Services
  - complete
    - alldata
    - SARS deployed workflow
    - SARS Prediction
    - SARS1
    - SARS3
    - SARS4
    - SARS5**
    - WSDL test
    - sProtein
  - gff2

Components Task manager inforsense.net

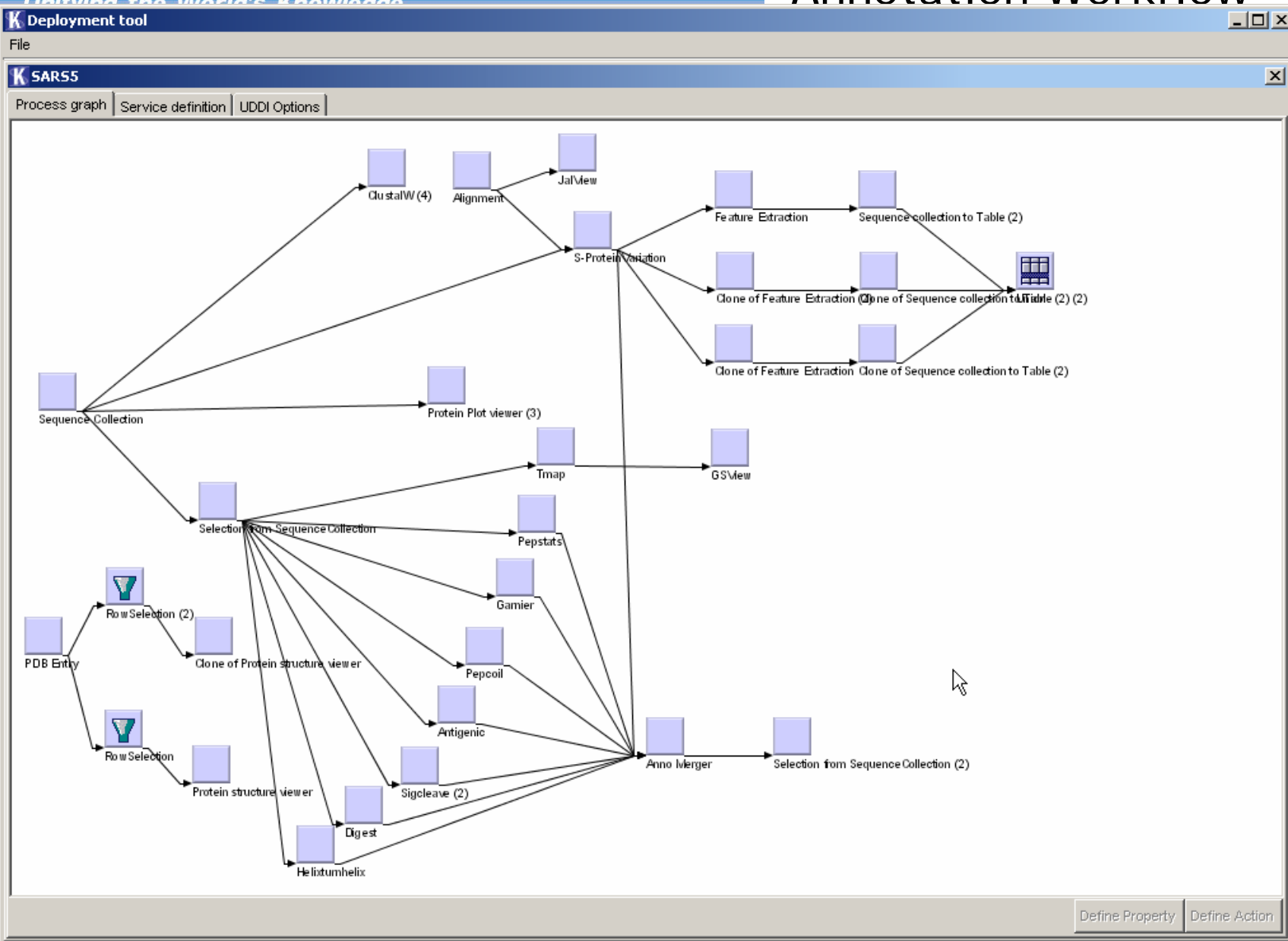
- ImExport
- Preprocess
- Normalisation
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Assess
- BioScience



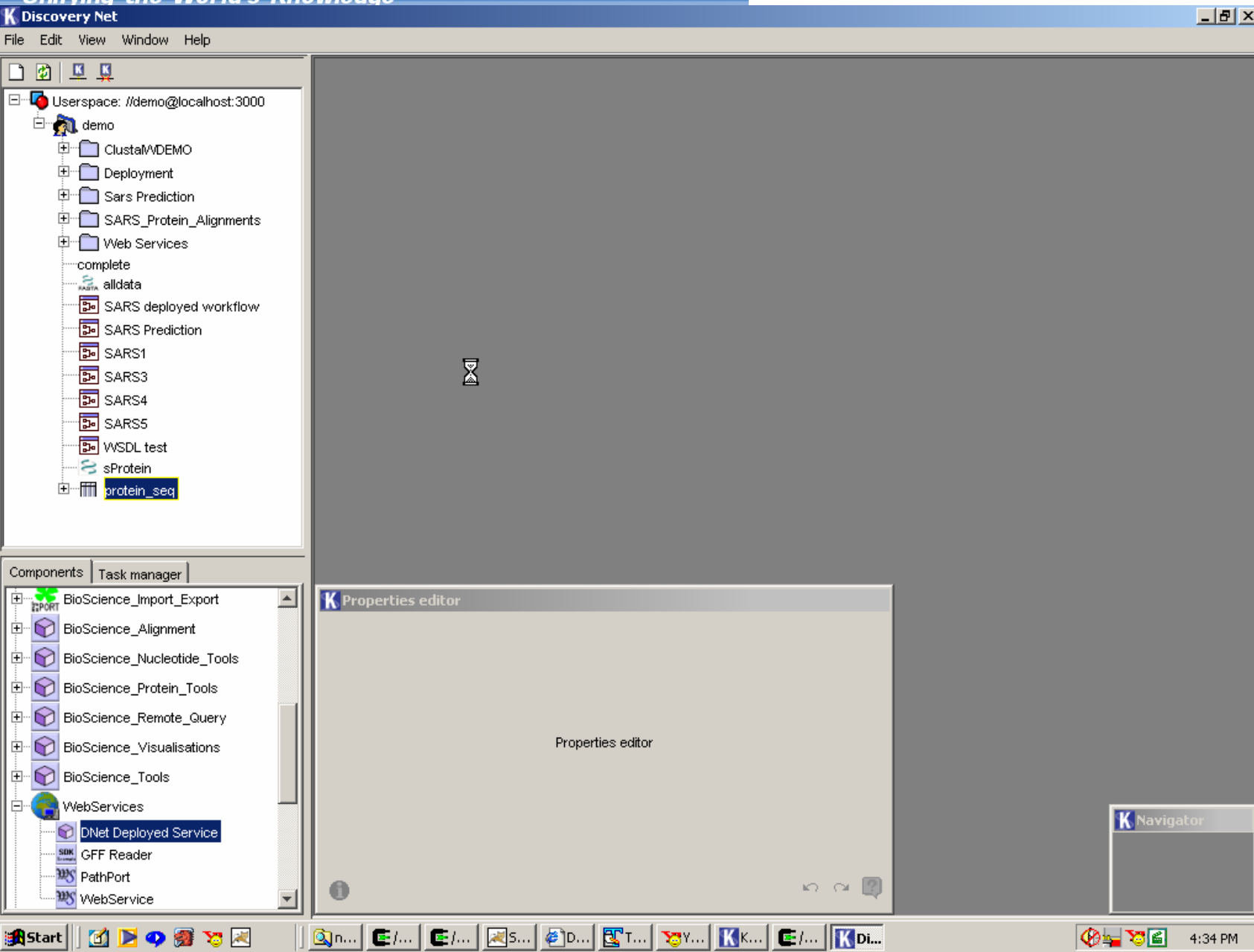
K Navigator



## Deploying Protein Annotation Workflow



The screenshot shows a Microsoft Internet Explorer browser window displaying a login page for the SARS Virtual Lab. The browser's address bar shows the URL `http://localhost:8080/kweb/scbit.html`. The login form is titled "SARS Virtual Lab - Login" and features the SCBIT logo (www.scbit.org) at the top. Below the logo, there are three input fields: "Username", "Password", and "Server". The "Server" field is pre-filled with `//localhost:3000`. A "Login" button is positioned to the right of the "Server" field. The browser's taskbar at the bottom shows several open applications, including "SARS Vir...", and the system clock indicates the time is 5:13 PM.

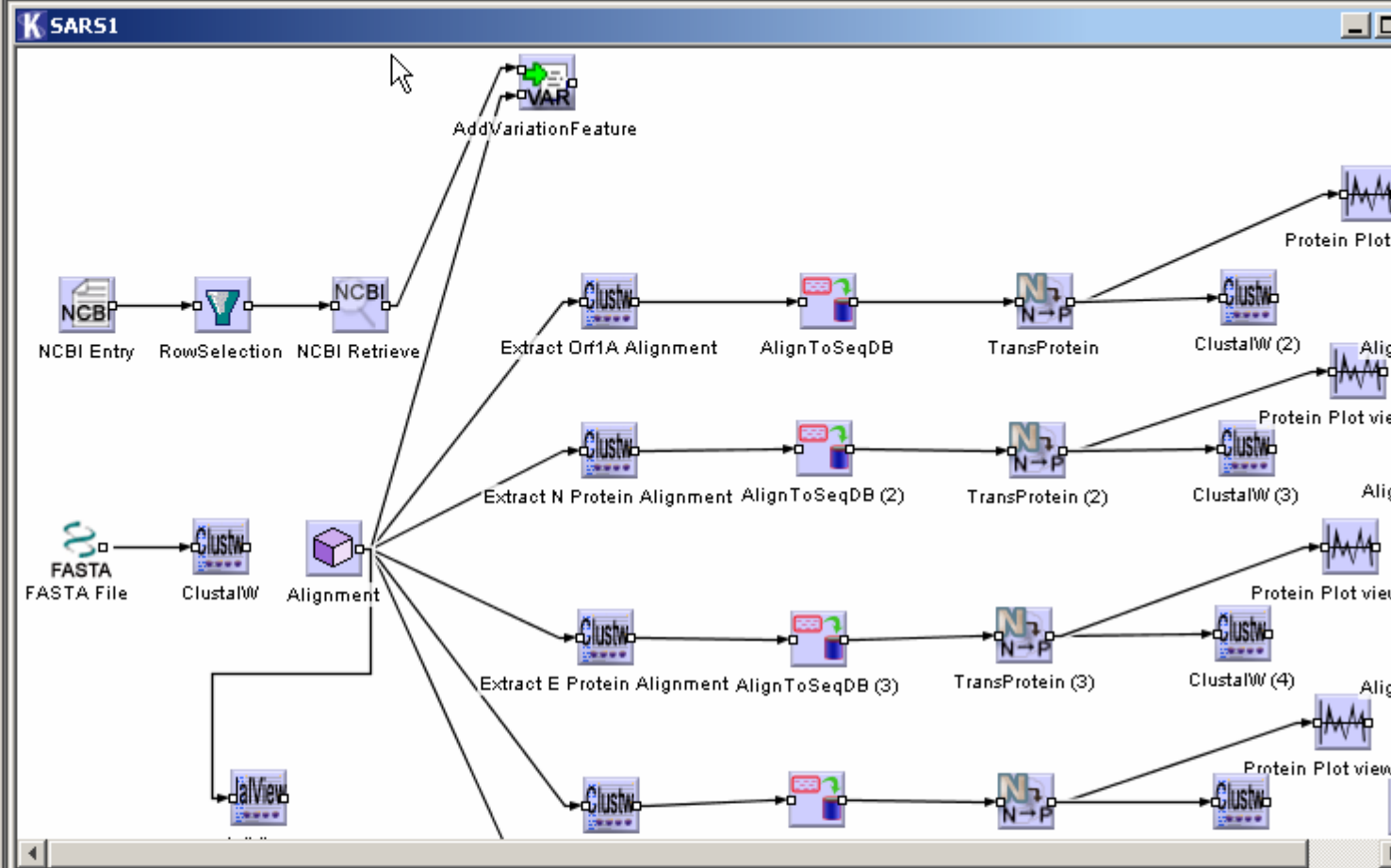


Userspace: //demo@localhost:3000

- demo
  - ClustaWDEMO
  - Deployment
  - Sars Prediction
  - SARS\_Protein\_Alignments
  - Web Services
  - complete
  - alldata
    - SARS deployed workflow
    - SARS Prediction
    - SARS1**
    - SARS3
    - SARS4
    - SARS5
    - WSDL test
    - sProtein
    - protein\_seq

Components Task manager inforsense.net

- Normalisation
- Statistics
- Association
- Classification
- Clustering
- Multivariate
- Assess
- BioScience
- BioScience\_Import\_Export
- BioScience\_Alignment
- BioScience\_Nucleotide\_Tools



### K Properties editor

Properties editor

### K Navigator

Unifying the World's Knowledge

The screenshot displays a Microsoft Internet Explorer browser window with the address `http://localhost:8080/kweb/frames/chrome/scbit/welcome.html`. The page header features the SCBIT logo and the text "上海生物信息技术研究中心" (Shanghai Center for Bioinformatics Technology). A navigation bar includes links for "SARS Virtual Lab", "Browse", "Meta-Analysis", "Services", "Search", "Tasks", and "Logout".

The main content area shows a workflow diagram starting from a "Sequence Collection" node. This node branches into several paths:

- One path goes through "ClustalW (4)" and "Alignment" to "JalView".
- Another path goes through "S-Protein Variation", which further branches into "Feature", "Clone o", and "Clone o".
- A third path goes through "Protein Plot viewer (3)".
- A fourth path goes through "Tmap" to "GSView".
- A fifth path goes through "Selection from Sequence Collection", which then branches into "Row Selection (2)", "Clone of Protein structure viewer", "Pepstats", "Garner", and "Pepcoil".



On the left side of the browser window, there is a vertical sidebar with a search input field and several buttons, including "View result" and "View". The bottom status bar shows the address `http://localhost:8080/kweb/search` and "Local intranet".

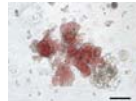
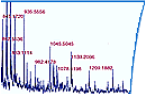
### Scientific Information



In Real Time

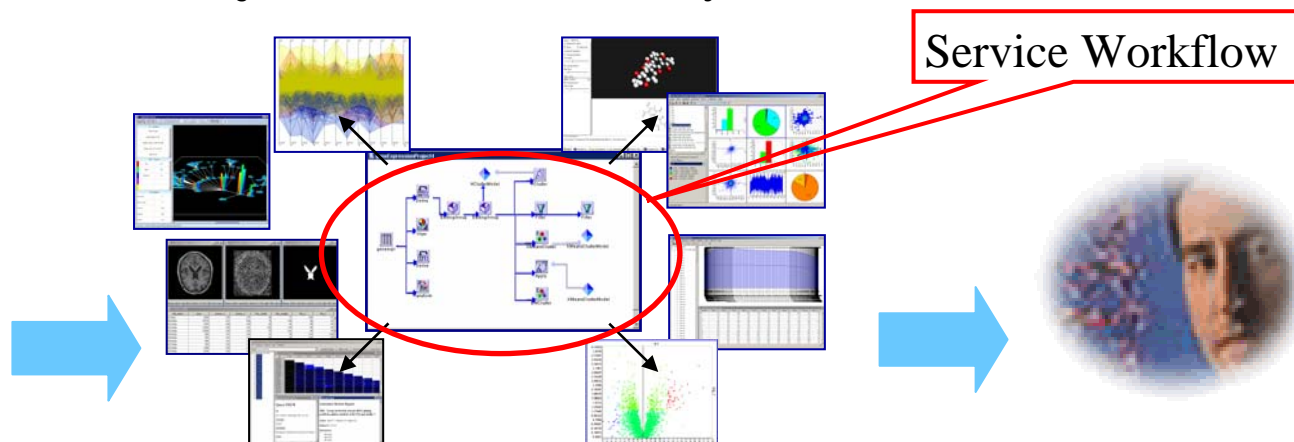
### Scientific Discovery

- Literature 
- Databases 
- Operational Data 

Count	Percent	Year
100	24.1%	1998
92	22.5%	1999
89	22.0%	2000
84	20.9%	2001
55	13.5%	2002
- Images 
- Instrument Data 

Real Time Data Integration

Discovery Services



Dynamic Application Integration

Integrative Knowledge Management

Using Distributed Resources

