InforSense
The Integrative Analytics Company

Open Discovery Workflows

Beyond Pipelining

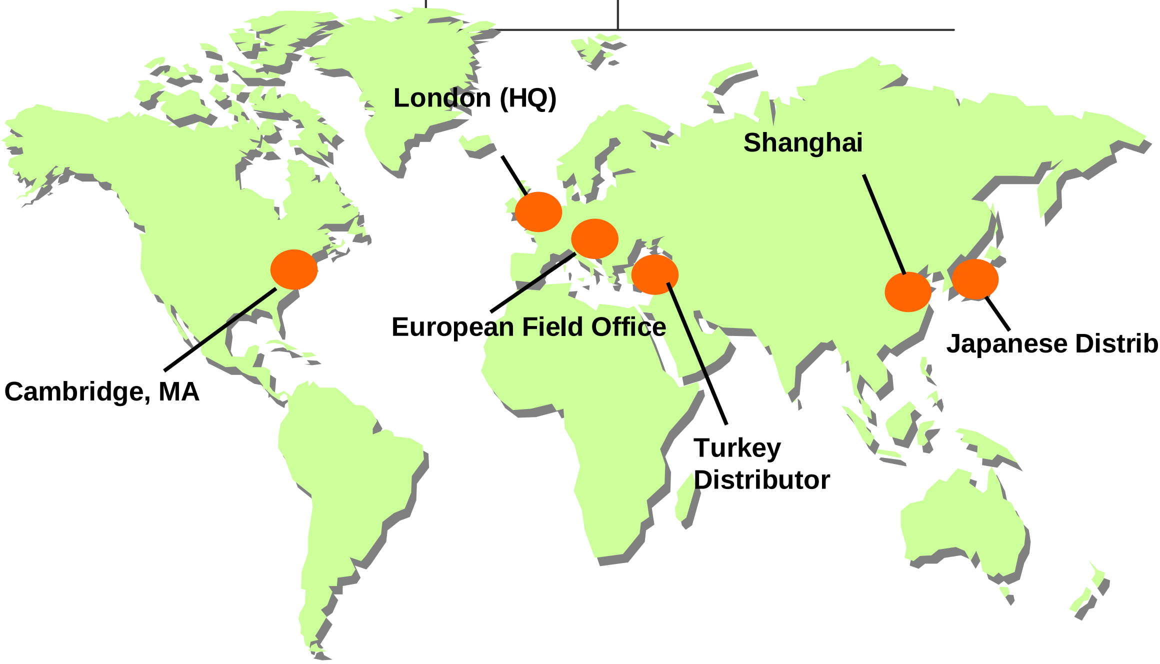
Jonathan Sheldon, CSO

(LifeScience and Healthcare)

InforSense Ltd

InforSense
The Integrative Analytics Company

InforSense Introduction



P&G

gsk

affibody

National Cancer Institute

AstraZeneca

Servier

GE

Bayer

Unilever

Berlex

Curis

Centre Genomique de Montreal
Montreal Genome Centre

Windber Research Institute

Boehringer Ingelheim

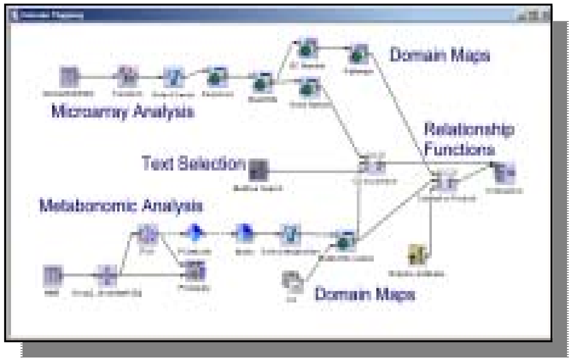
sanofi aventis

astellas
Leading Light for Life

InforSense
The Integrative Analytics Company

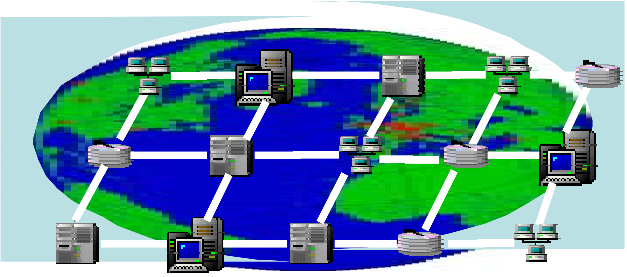
InforSense Technology Foundation

Workflow based portal building technology: Automatically deploying workflow composing services to an enterprise portal
BioWorld IT 2005 Best of the Show Award - Knowledge Management & Collaboration

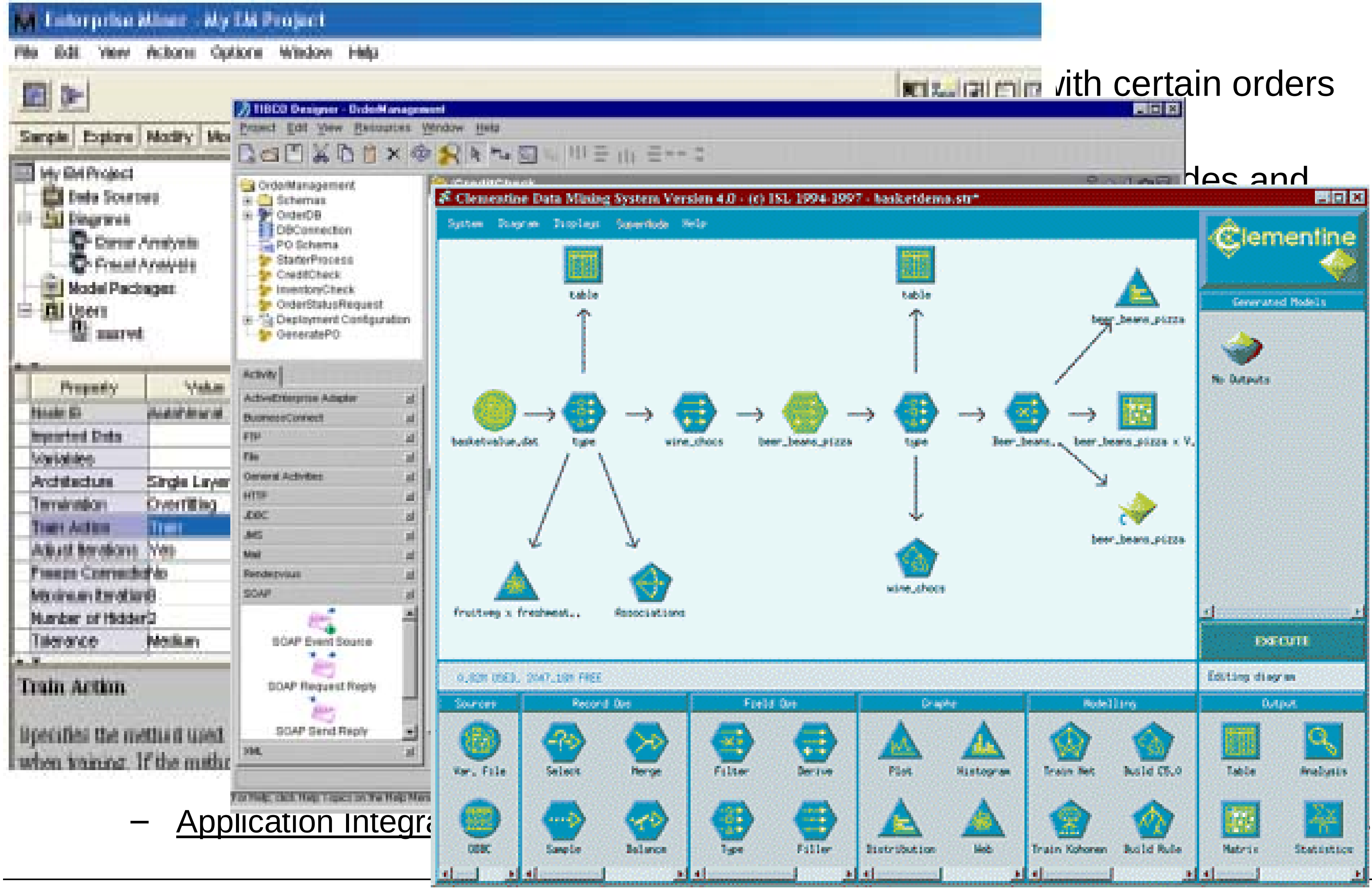


Workflow technology : Building applications by composing services
IEEE Supercomputing 2002 : Most Innovative Award

Web Services based informatics infrastructure: Harness multi-domain distributed resources
IEEE Supercomputing 1998 : Most Innovative Award



Workflow is not new…



Property	Value
Node ID	AutoNeural
Imported Data	
Variables	
Architecture	Single Layer
Termination	Overfitting
Train Action	Train
Adjust Iterations	Yes
Freeze Connections	No
Maximum Iterations	8
Number of Hidden	3
Tolerance	Medium

Train Action

Specifies the method used when training. If the meth

0.32M USED, 2047.18M FREE

vith certain orders

des and

– Application integra

InforSense
The Integrative Analytics Company

Building a workflow

1 Select:

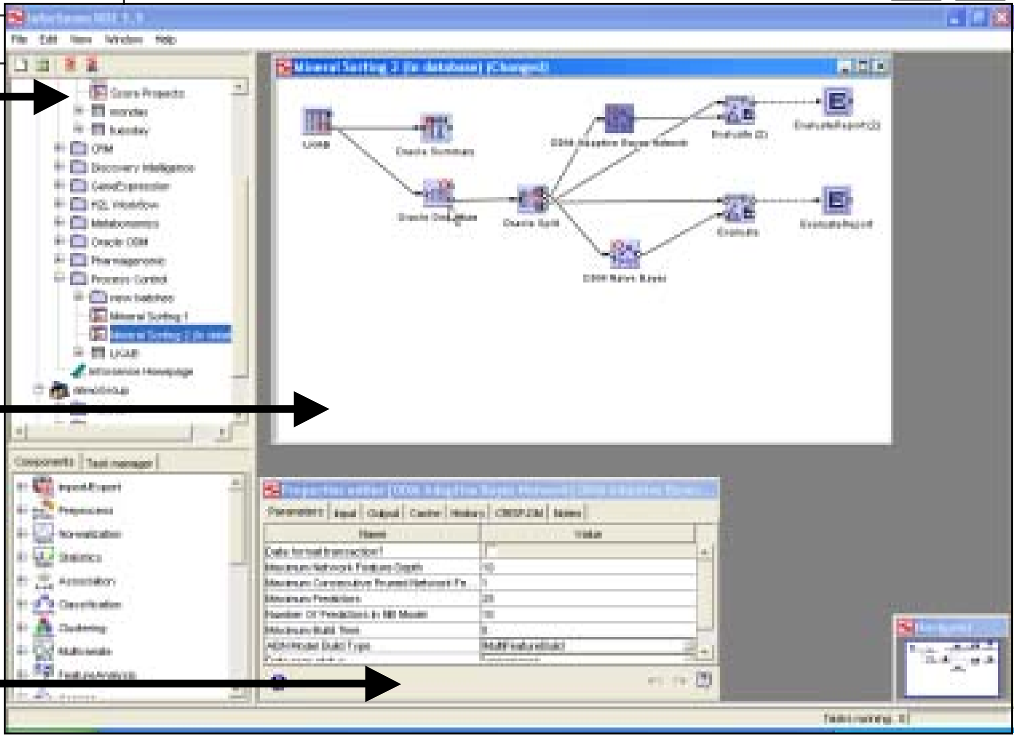
- Data Sources
- Data Mining / Statistics
- Data Processing / Transformation
- SQL Functions
- Interactive data visualization / reporting

2 Connect:

- Connect data and components in GUI
- Workflow describes complex data processing and analysis

3 Execute:

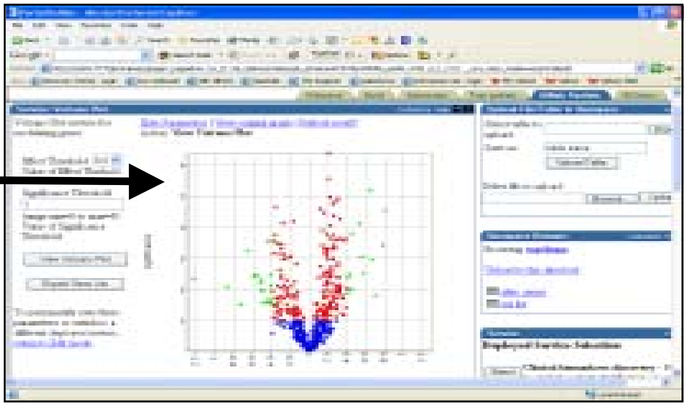
- “In database” processing & analytics



Name	Value
Data format transaction?	
Maximum Network Feature Depth	10
Maximum Consecutive Pruned Network Fe...	1
Maximum Predictors	25
Number Of Predictors In NB Model	10
Maximum Build Time	0
ABN Model Build Type	MultiFeatureBuild

4 Deploy:

- Define parameters of workflow to expose
- Publish as: portlet, web application, SOAP service, command line app

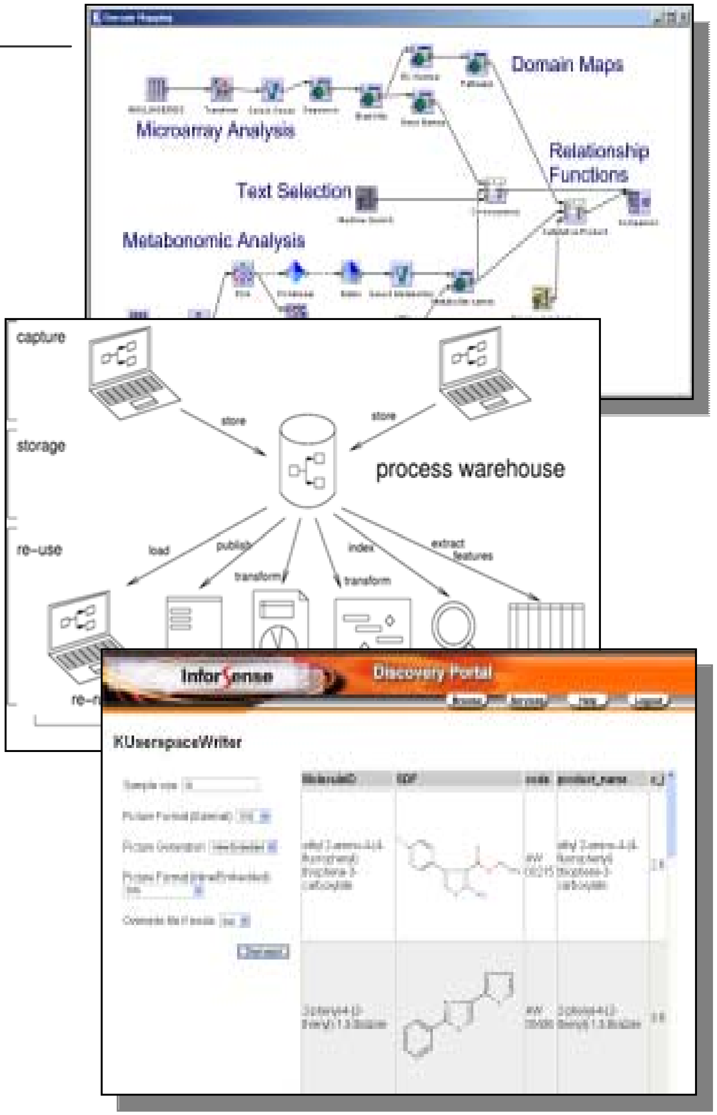


5 Warehouse:

- Store workflows to capture ad-hoc querying and establish best practices

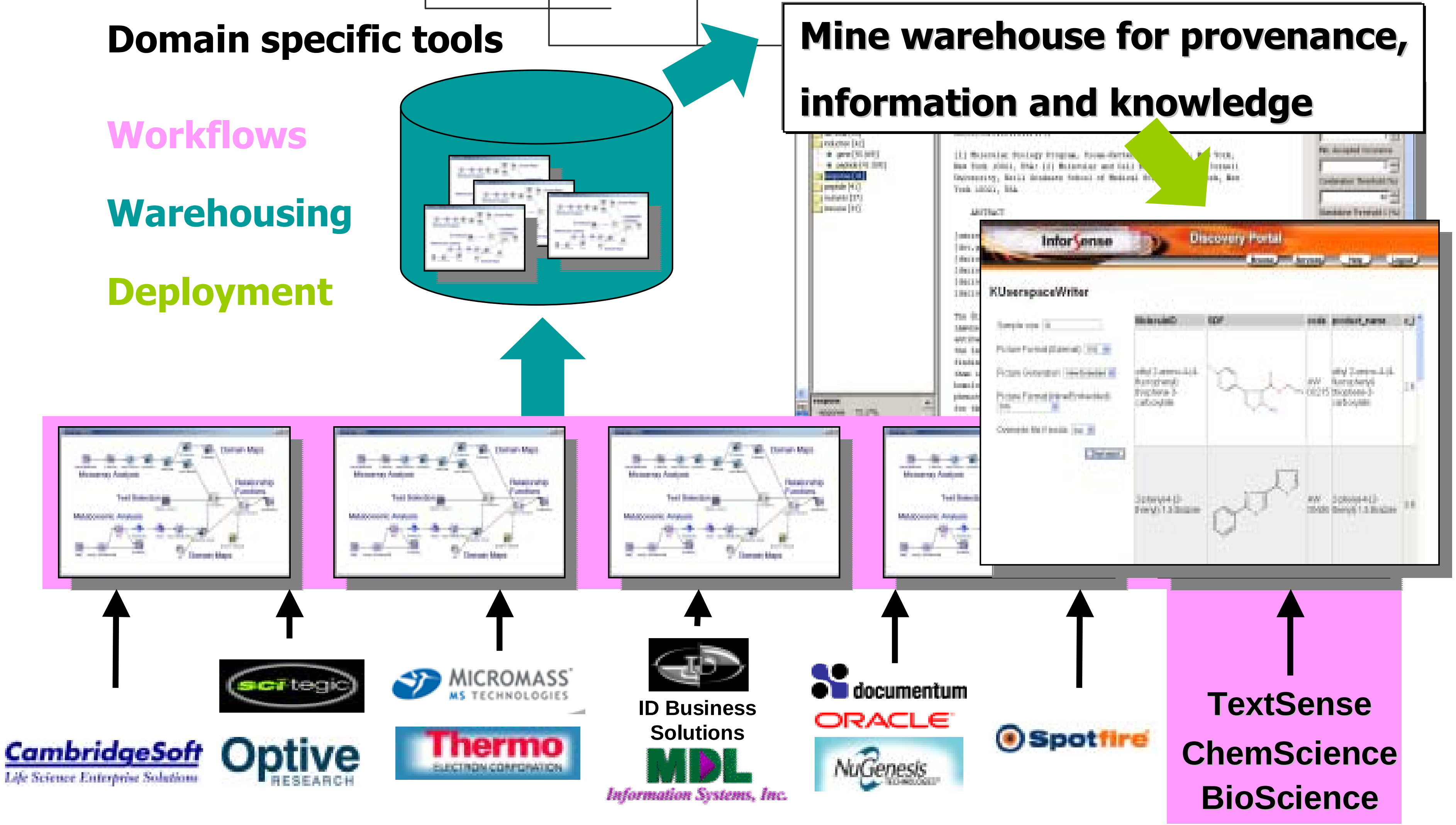
InforSense Workflow Life Cycle

- Constructing a ubiquitous workflow:
  by scientists
  - Integrate your information resources/software applications cross-domain
  - Support innovation and capture the best practice of your scientific research
- Warehousing workflows:
  for scientists
  - Manage discovery processes in your organisation
  - Construct an enterprise process knowledge bank
- Deployment workflow:
  to scientists
  - Turn your workflows into reusable applications
  - Turn every scientist into a solution builder



InforSense
The Integrative Analytics Company

Ubiquitous workflows, warehouses and deployment engines



Domain specific tools

Workflows

Warehousing

Deployment

Mine warehouse for provenance, information and knowledge

TextSense
ChemScience
BioScience

Experiment & design

Enumeration & descriptors

Lab & analytical data

Compound & activity data

Database & documents

Visualization

InforSense analytical systems

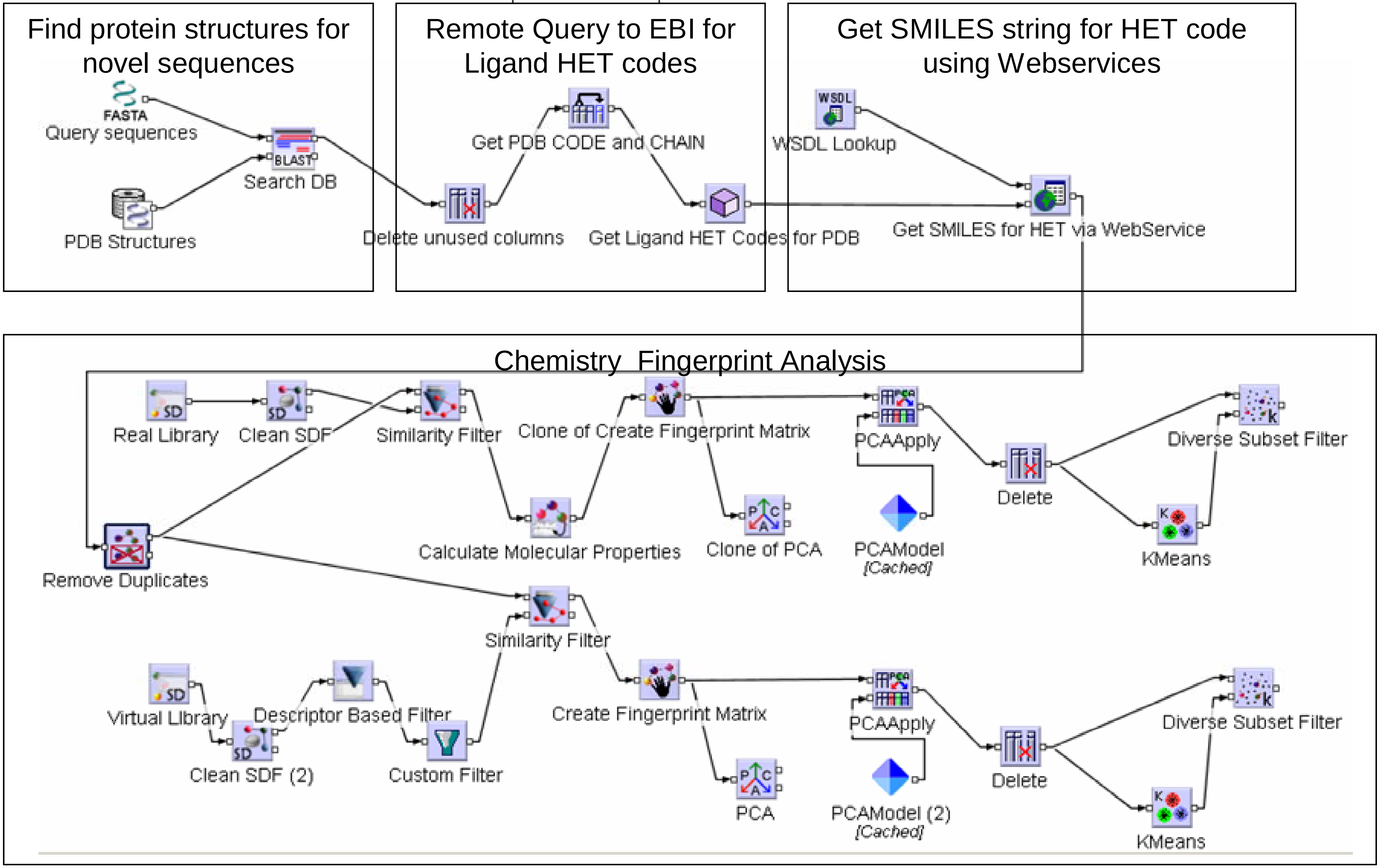
InforSense
The Integrative Analytics Company

There is something new about the use of workflow in scientific informatics…

- Workflow is an “machine executable protocol” for in silico scientific activities
- Thus, a scientific workflow needs to be :
  - **Ubiquitous** : scientific data are heterogamous and scientific activities are rich in variety
  - **Integrative** : resources (data, software, instruments) need to be dynamically combined for scientific study
  - **Personal** : scientists need total freedom in defining and building their creative scientific process
  - **Interactive** : building a scientific workflow with a “ trial and error” approach
  - **Open** : workflow should be able to integrate components from any vendors
  - **Reusable** : workflow should be reused by others in various form
  - **Manageable** : workflow represents process knowledge so it evolves overtime

InforSense
The Integrative Analytics Company

Ubiquitous : Manage the complete workflow of a project in one framework



Find protein structures for novel sequences

Remote Query to EBI for Ligand HET codes

Get SMILES string for HET code using Webservices

Chemistry Fingerprint Analysis

Integrative : Dynamic cross domain integration of databases & applications

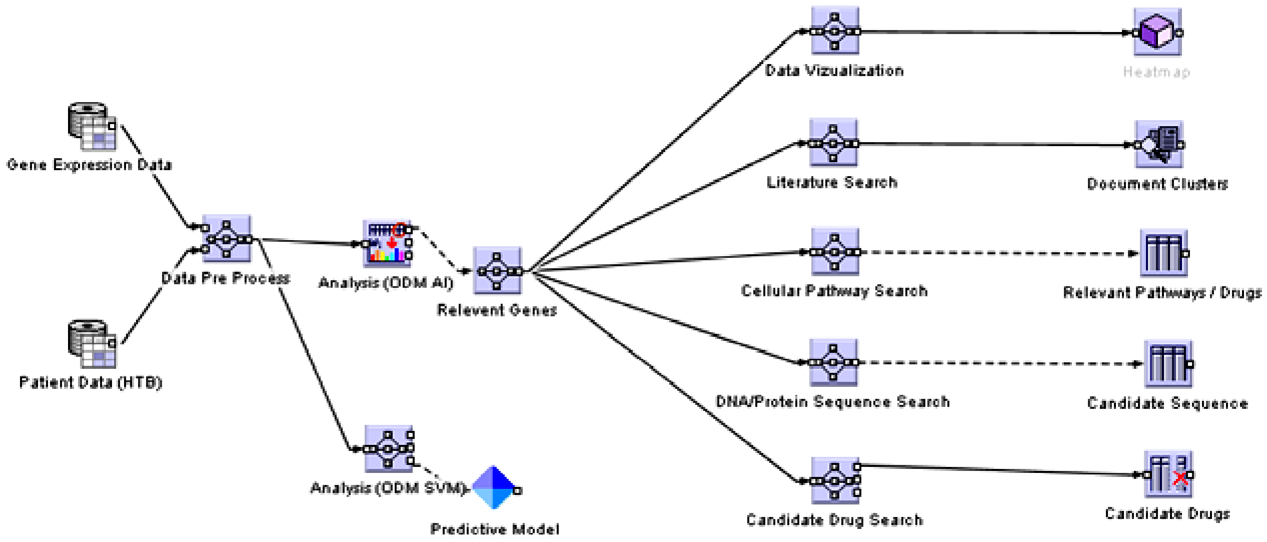
- **Used Modules in KDE**
  - 60 nodes representing 5 x DBs[1] and Oracle components[2], 3 x 3rd party applications[3], 4 x public DBs[4] and 1 web service[5]
  - Gain: 27 times faster (10 vs. 270 Man Days), reusable and sharable

The Workflow

Data Acquisition

Multiple Types of Analysis

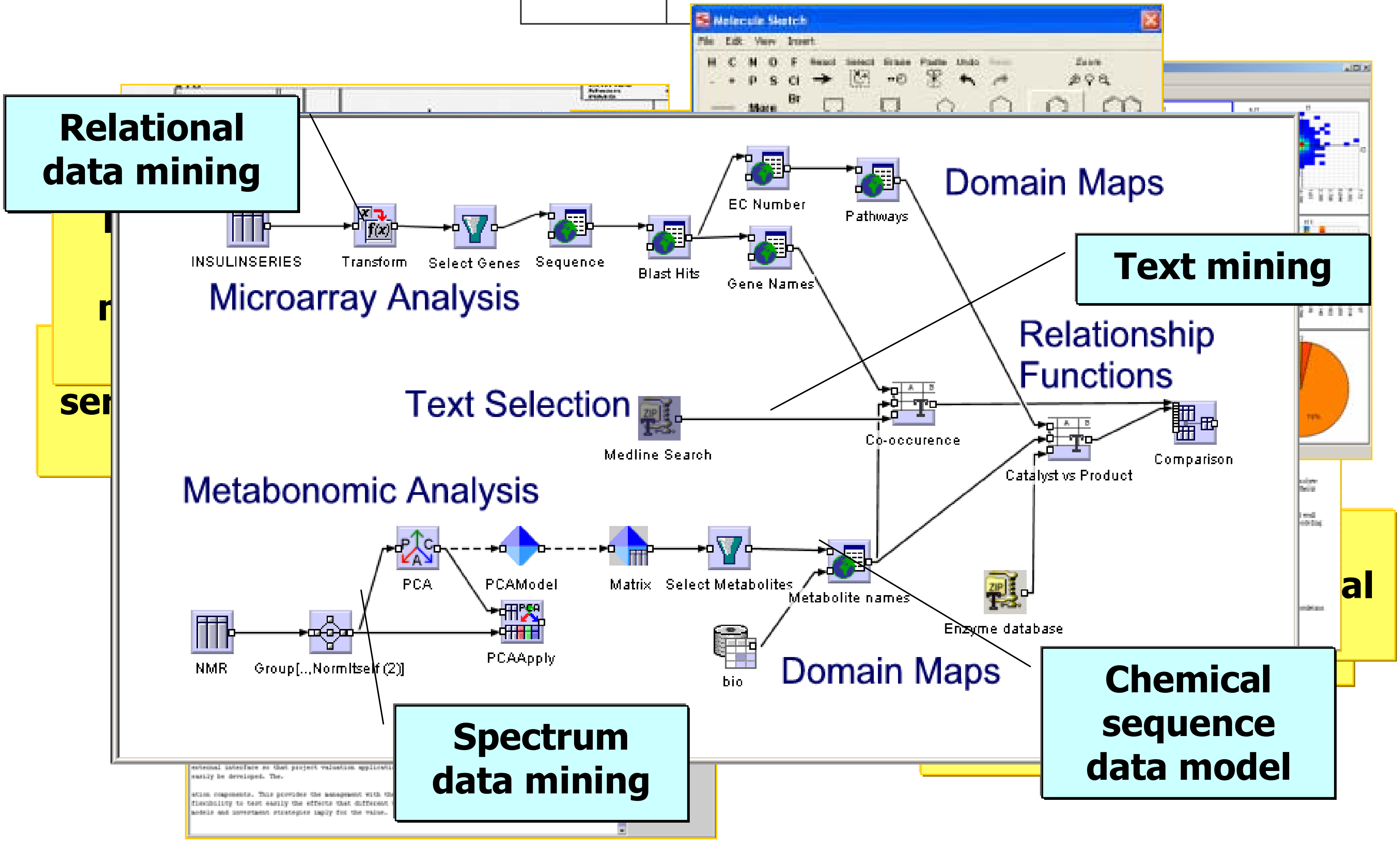
Relevant Results



[1] Gene Expression, Patient, [2] ODM (SVM, AI, BLAST, TEXT), [3] Chemical Editor, Blast Viewer, HeatMap, [4] ChemBank, PubMed, NCBI, Beilstein, [5] KEGG

InforSense
The Integrative Analytics Company

Interactive : Building Complex Discovery Process with Visualisation



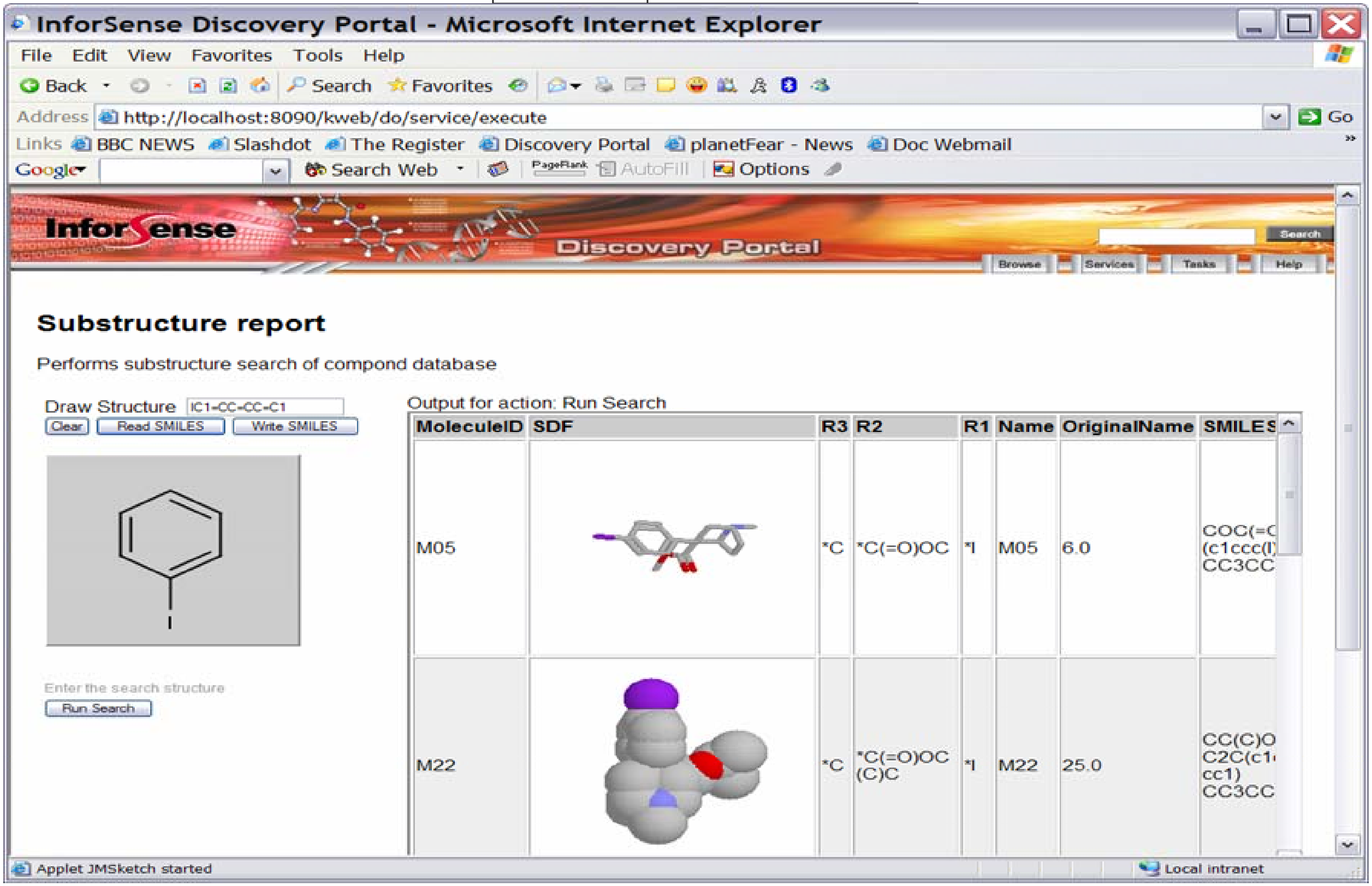
Relational data mining

Text mining

Chemical sequence data model

Spectrum data mining

Personal : Building and deploying personal applications



InforSense Discovery Portal - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8090/kweb/do/service/execute
Go
Links BBC NEWS Slashdot The Register Discovery Portal planetFear - News Doc Webmail
Google Search Web PageRank AutoFill Options
InforSense Discovery Portal
Search
Browse Services Tasks Help

Substructure report

Performs substructure search of compond database

Draw Structure IC1=CC=CC=C1

Clear
Read SMILES
Write SMILES
I

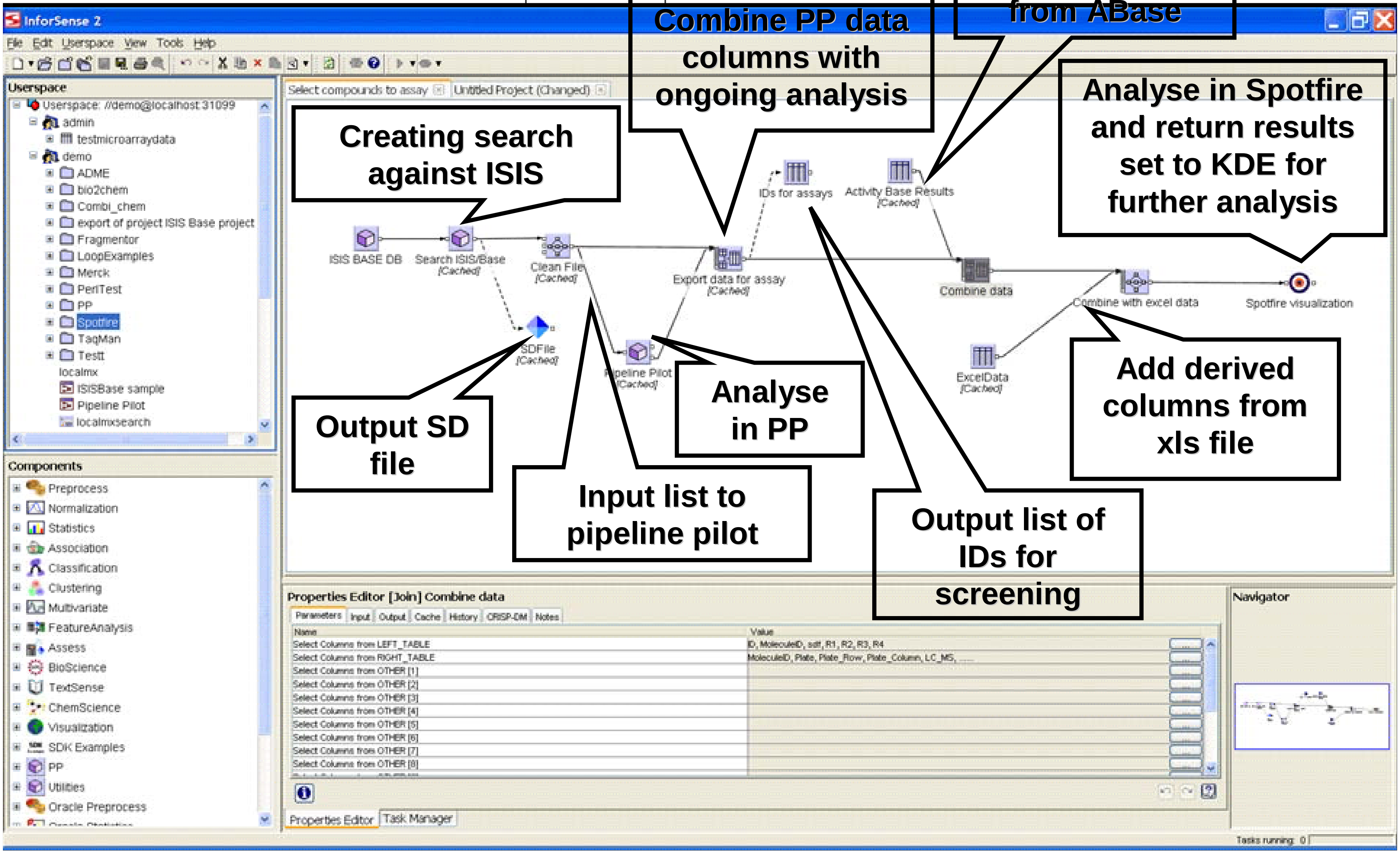
Enter the search structure

Output for action: Run Search

MoleculeID	SDF	R3	R2	R1	Name	OriginalName	SMILES
M05		*C	*C(=O)OC	*I	M05	6.0	COC(=C (c1ccc(I) CC3CC
M22		*C	*C(=O)OC (C)C	*I	M22	25.0	CC(C)O C2C(c1 cc1) CC3CC

Applet JMSketch started
Local intranet

Open : Working with any vendor



Creating search against ISIS

Combine PP data columns with ongoing analysis

Retrieve data from ABase

Analyse in Spotfire and return results set to KDE for further analysis

Output SD file

Input list to pipeline pilot

Analyse in PP

Output list of IDs for screening

Add derived columns from xls file

Properties Editor [Join] Combine data

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

Name	Value
Select Columns from LEFT_TABLE	ID, MoleculeID, sdf, R1, R2, R3, R4
Select Columns from RIGHT_TABLE	MoleculeID, Plate, Plate_Row, Plate_Column, LC_MS, ...
Select Columns from OTHER [1]	
Select Columns from OTHER [2]	
Select Columns from OTHER [3]	
Select Columns from OTHER [4]	
Select Columns from OTHER [5]	
Select Columns from OTHER [6]	
Select Columns from OTHER [7]	
Select Columns from OTHER [8]	

Properties Editor | Task Manager

Some other must have features…

- Higher level orchestration of workflows
  - conditional execution, loops..
- ‘Useful’ error reporting
  - exactly what has broken !
- Cached results
  - resilience, debugging
- Asynchronous jobs
  - long running processes

InforSense
The Integrative Analytics Company

Some other must have features…

- In database execution
  - speed, security
- Audit trial
  - who, what, when
- Meta data level consistency & error checking
  - construction of workflows with compatible data types
- Support for SOA
  - web services
- Enterprise infrastructure
  - cluster support, server admin tools

InforSense
The Integrative Analytics Company

Case Study:
Library design

Collaboration with GSK Cheminformatics

InforSense
The Integrative Analytics Company

History of the “Industrial” pharma (1)

- **Historically**
  - – Low volume
    - 30-50 cmpds/yr/chemist: 10,000s assay wells/yr
  - – Low information diversity
    - scientists generally dealt with limited types of data
  - – Reductionist approach
    - limited information per experiment
      - – Interpretation critical for next step
    - scientists required:
      - – simple systems to assist in information monitoring
      - – decision making resides with the scientist

gsk Slides from their e-science presentation at the EPSRC, April 2005.

History of the “Industrial” pharma (2)

- **What happened in the last 5 years?**
  - “Industrialisation” - Application of “principles of industrialisation” to drug discovery
    - high volume
      - 10,000 cmpd/yr/chemist/100+ million wells/yr
  - Biology revolution
    - Human genome
      - “system biology” – holistic view and interpretation
      - high content data --- images
      - multiple result types from each experiment – bio-markers, pathways
  - Knowledge integration
    - scientific discipline integration
  - Scientists required
    - complex systems, algorithms, statistics…….
    - decision making shared between systems and scientists
    - “Informatics” essential – partnership not service

InforSense
The Integrative Analytics Company

How has pharma IT tackled the transition?

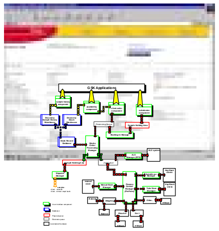
- Business as usual
  - problem centric view
    - build applications
    - integrate applications
- Educate scientists in the realms of IT
  - "Now I need to be an IT expert alongside chemistry, biology, genetics, robotics, engineering ……"
  - interesting time scale - generations
- Technology is our saviour!
  - client server, web services, java, C#, Corba, OO programming, extreme programming, grid computing, …..

gsk

What are the results?

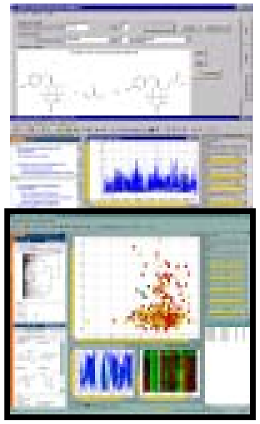
Chemistry

samples

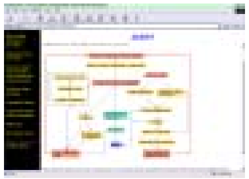


screening

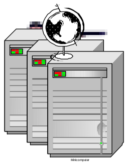
data



“library” design



infrastructure

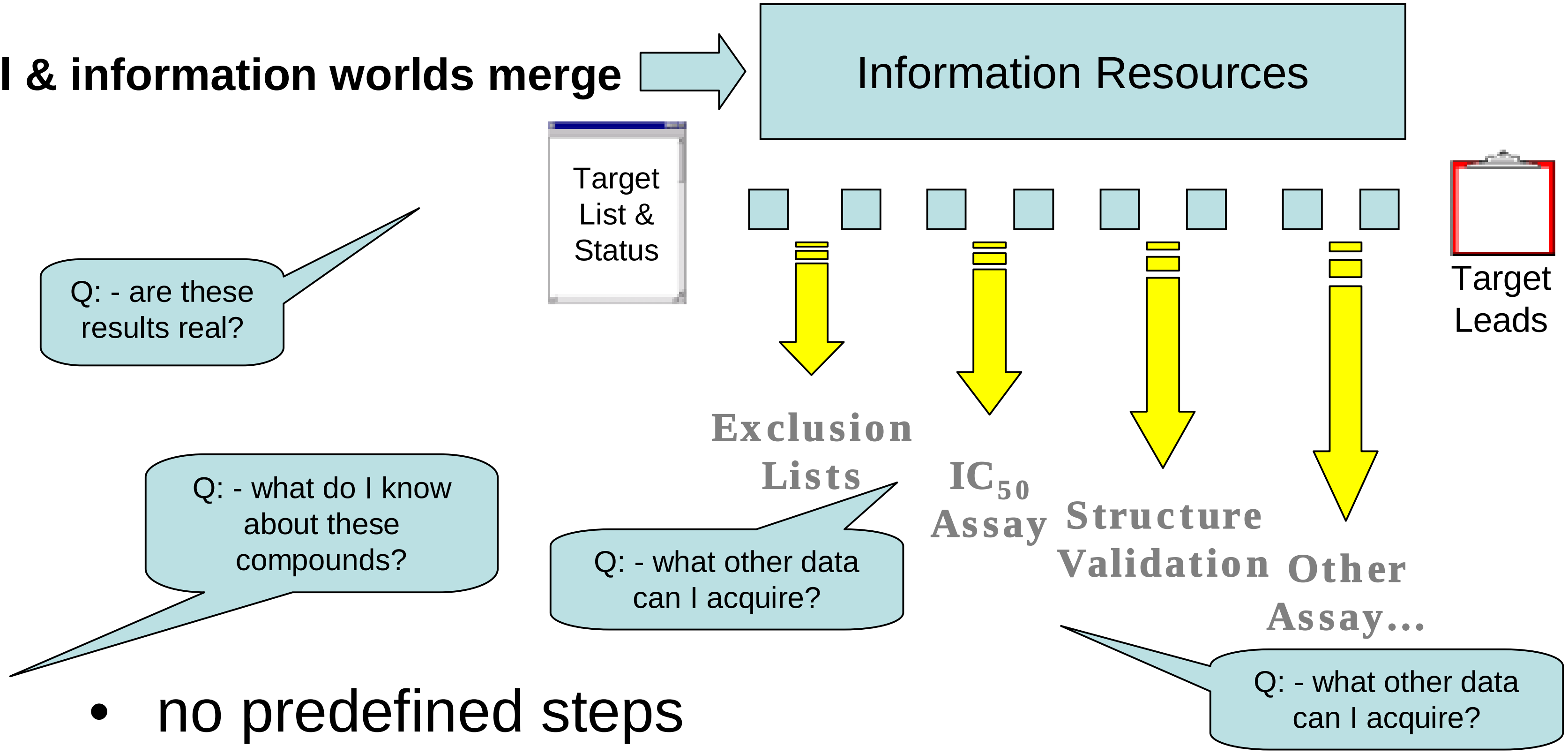


- “islands” or “silos” of process & data
  - *complex integration problem*
    - *“spaghetti” joins our worlds - unsustainable - cost*
- *control with “IT”*
  - *mismatch in cycle time to change*
  - *engineered out serendipity*
  - *service role reversed*

InforSense
The Integrative Analytics Company

Doodling in knowledge and experiment space

physical & information worlds merge


Information Resources
Target List & Status
Target Leads
Exclusion Lists
IC_{50} Assay
Structure Validation
Other Assay...
Q: - are these results real?
Q: - what do I know about these compounds?
Q: - what other data can I acquire?
Q: - what other data can I acquire?

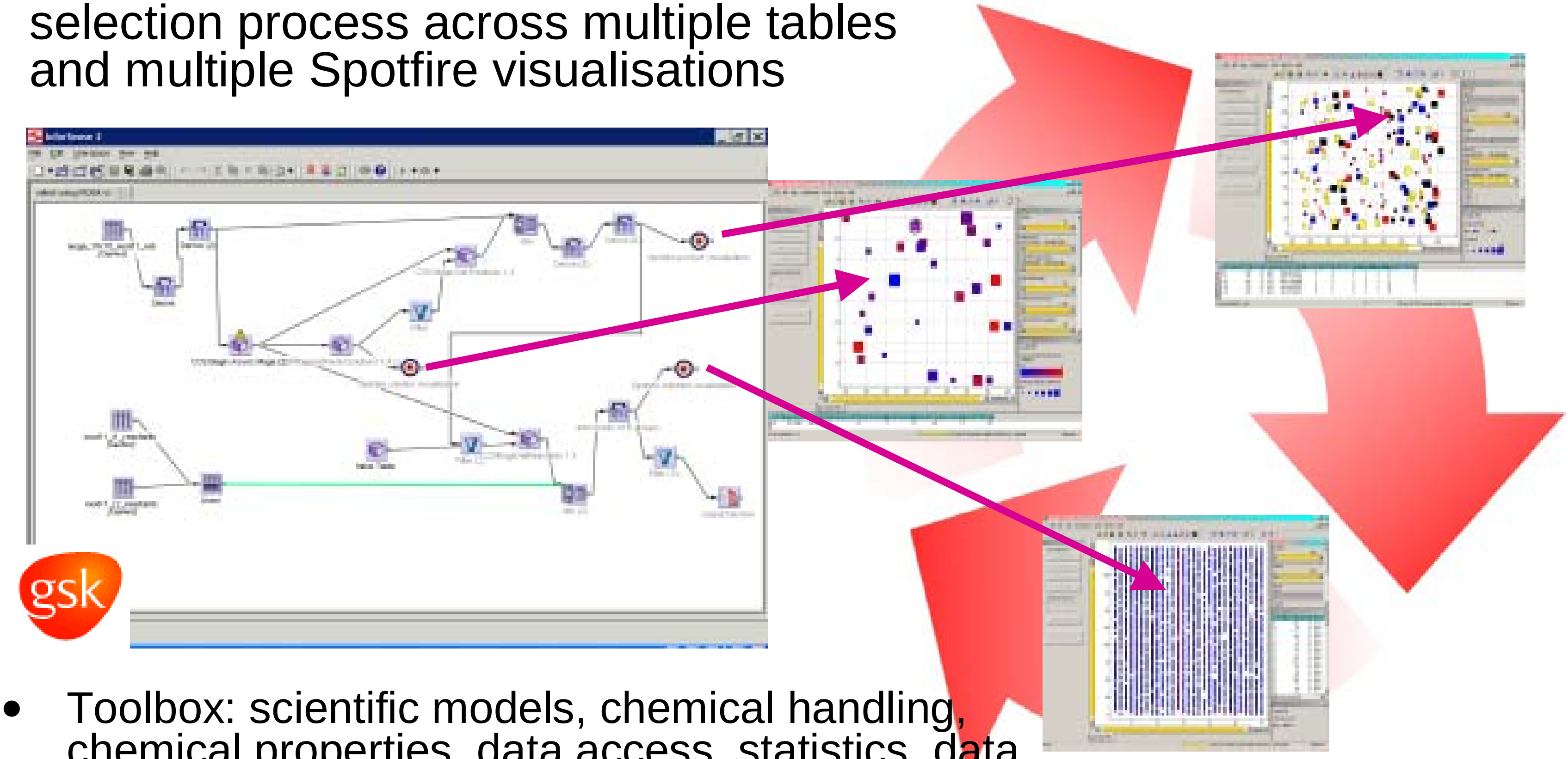
- no predefined steps
  - capture what was done, don't restrict what can be done?
    - don't restrict the non-obvious

gsk

- **Objective:**
  - Dynamic cheminformatics infrastructure
- **Specification:**
  - Access to world-class scientific algorithms and tools
  - Access to disparate data sources, multiple locations
  - An intuitive GUI for medicinal chemists – not just IT experts
  - Build and deploy “just-in-time” applications
  - Capture and dissemination of “best practice”
- **Solution:**
  - InforSense KDE + IOE open SOA, integrating services, data, tools
  - OWPN partner integrations
  - Spotfire collaboration -> integrative visual analytics

Library design @ GSK

KDEs Visual Network orchestrates the selection process across multiple tables and multiple Spotfire visualisations



- Toolbox: scientific models, chemical handling, chemical properties, data access, statistics, data visualisation, ...
- Scientists can doodle in chemical space
  - Capture how scientists made decisions
- New algorithms, data sources added in < 1 hour

InforSense
The Integrative Analytics Company

Case Study:
Mapping the evolution of SARS

Collaboration with the Shanghai Center for Bioinformation Technology (SCBIT)

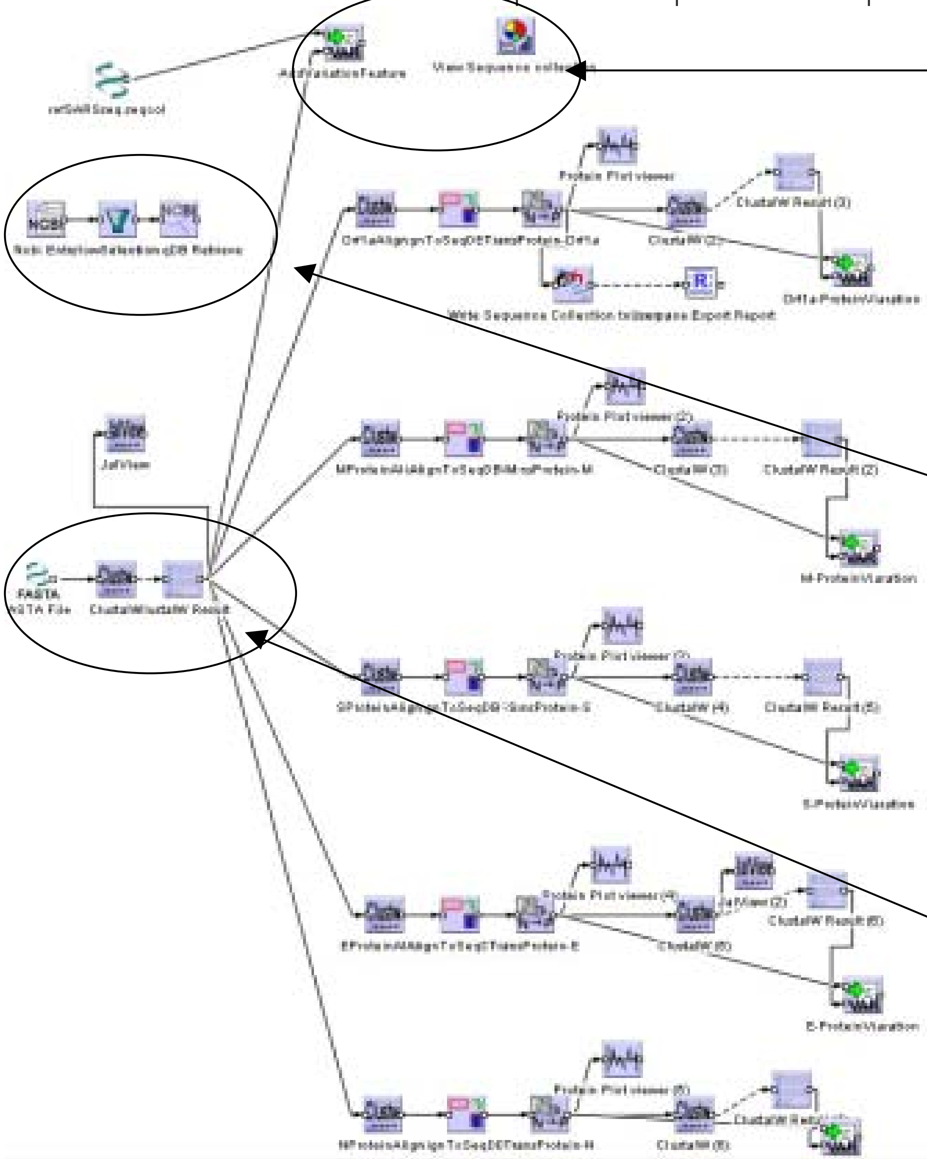
SARS

- By 2003, SARS was rapidly developing into a pandemic
- The WHO needed to quickly find a way to:
  - Rapidly establish the relationship between genomic variations and the biology of SARS
  - Characterise genomic variations (deletions, single nucleotide variations) during the epidemic
  - Rapidly aggregate and analyse multiple data sources using many different tools, requiring different analysis techniques, from multiple research groups world-wide, in multiple formats on multiple platforms
  - Publish and inform research and treatment groups world-wide to enable appropriate treatment regimes for each patient

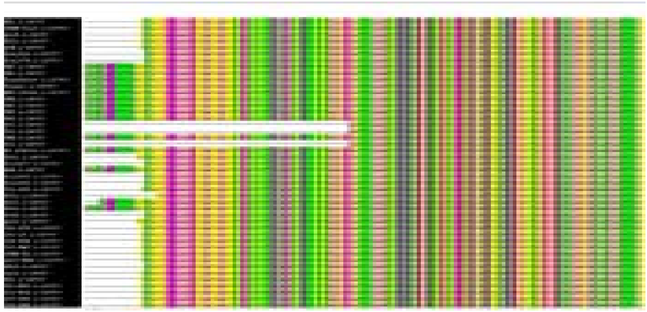
InforSense

The Integrative Analytics Company

SARS: Genomic Analysis



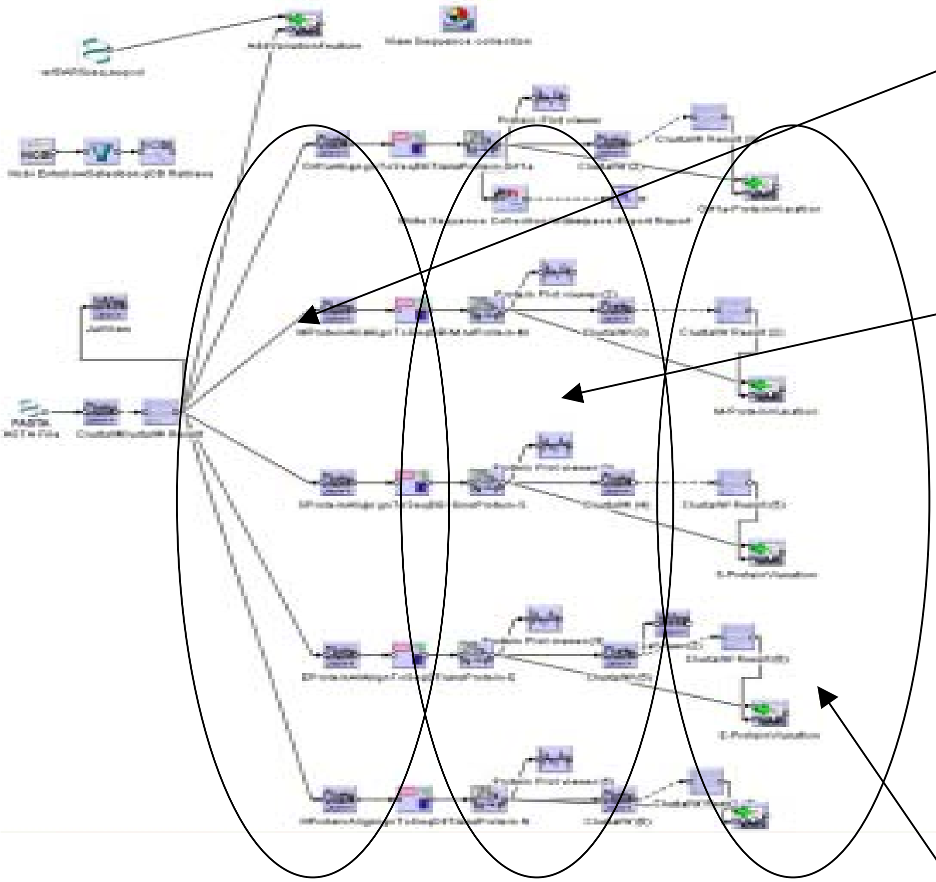
Examining the variations in different strains



Retrieval of publicly available knowledge

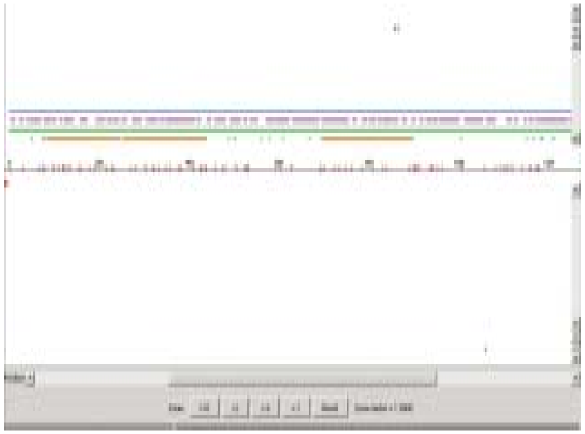
Alignment performed on the Grid

SARS: Proteomic Analysis



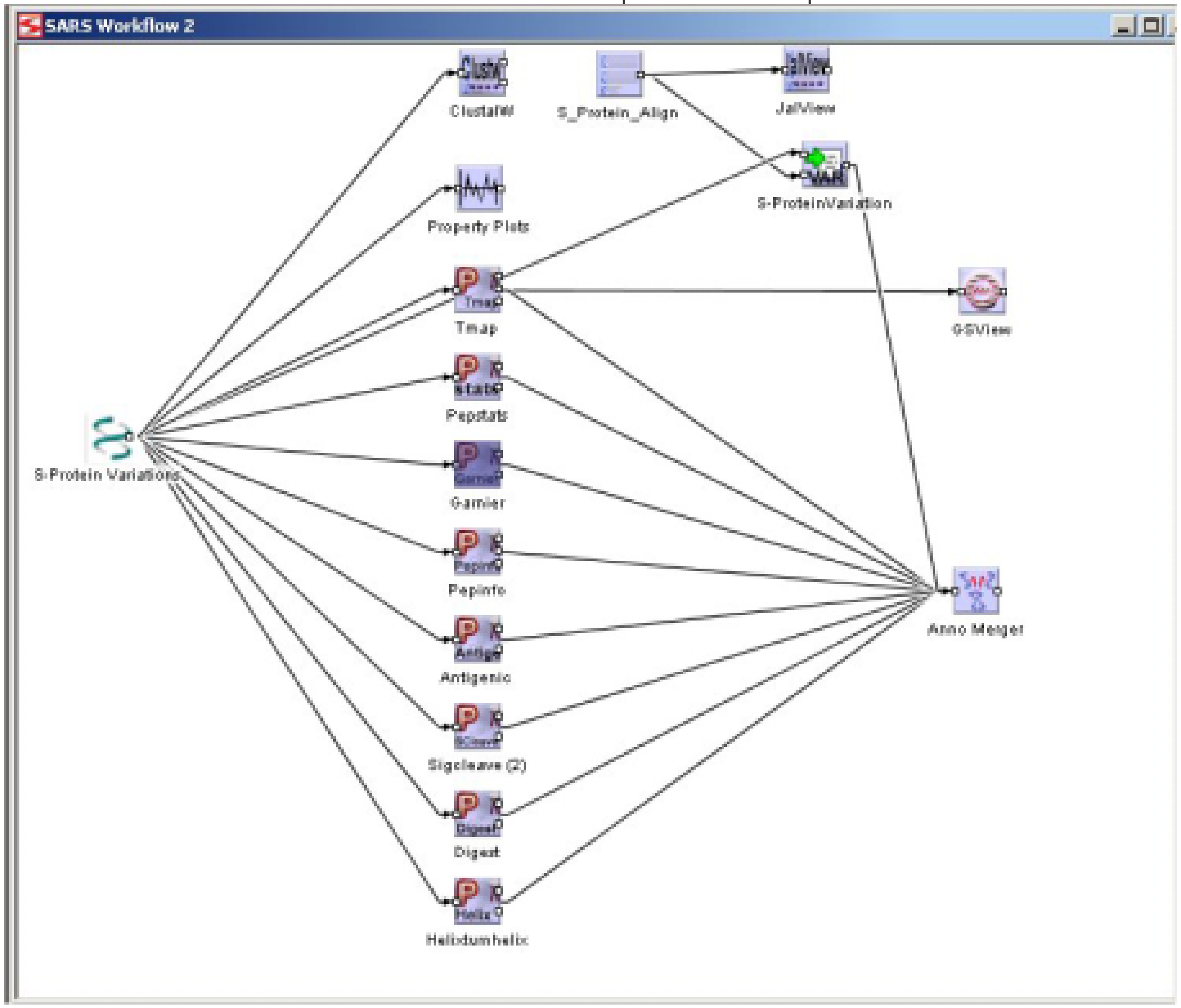
Isolating interesting genomic regions

Identifying relevant protein sequences



Observing the variations in the resulting protein

SARS: Protein Annotation



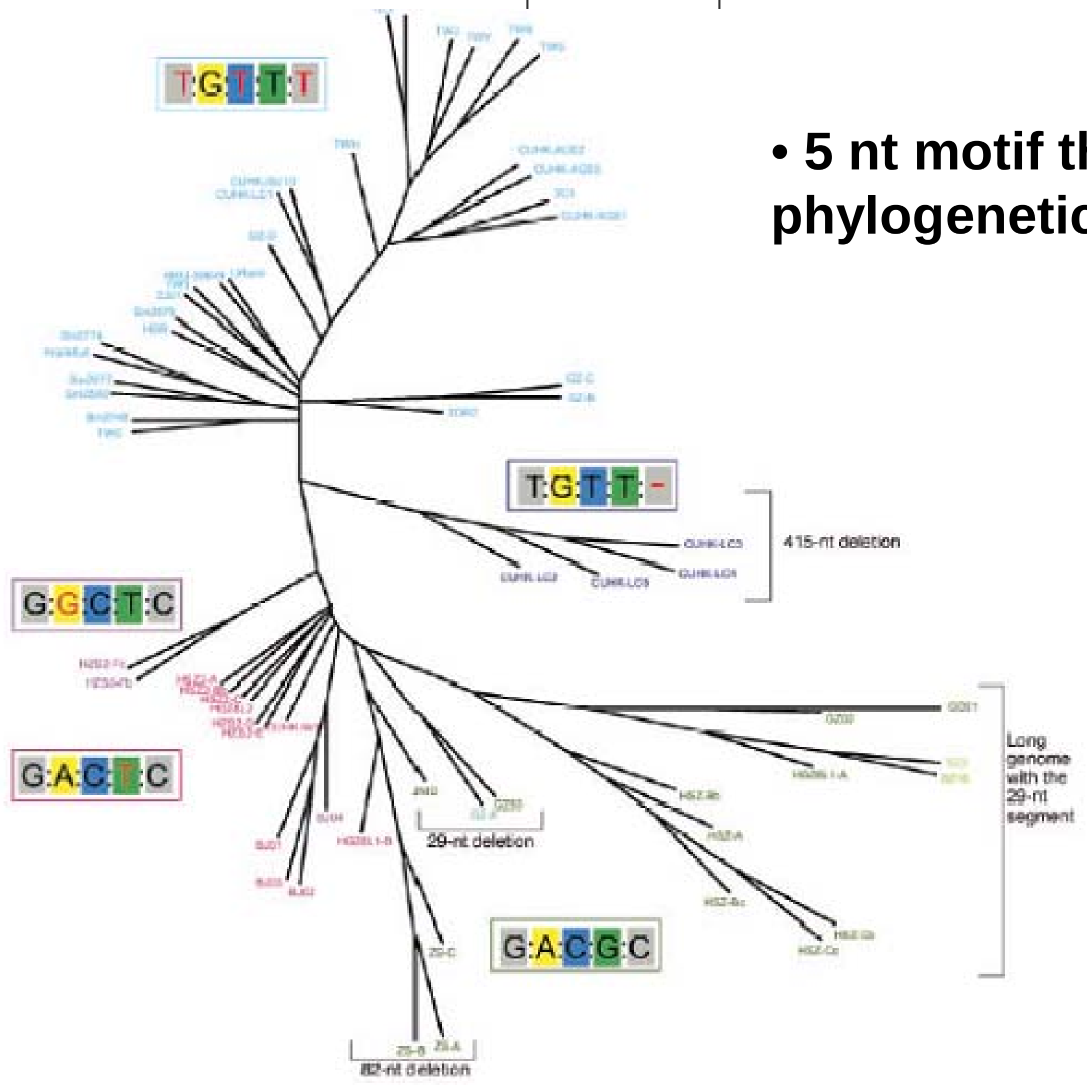
- Annotation using applications from EMBOSS package

- Execution on different physical resources

- Annotations merged

InforSense
The Integrative Analytics Company

SARS: Genotype Clustering

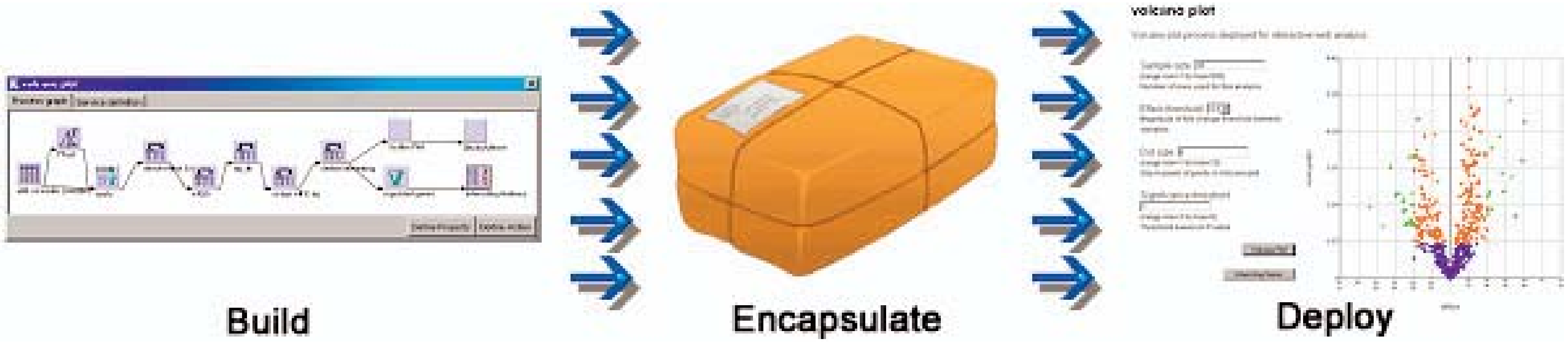


- **5 nt motif that characterises the phylogenetically related genotypes**

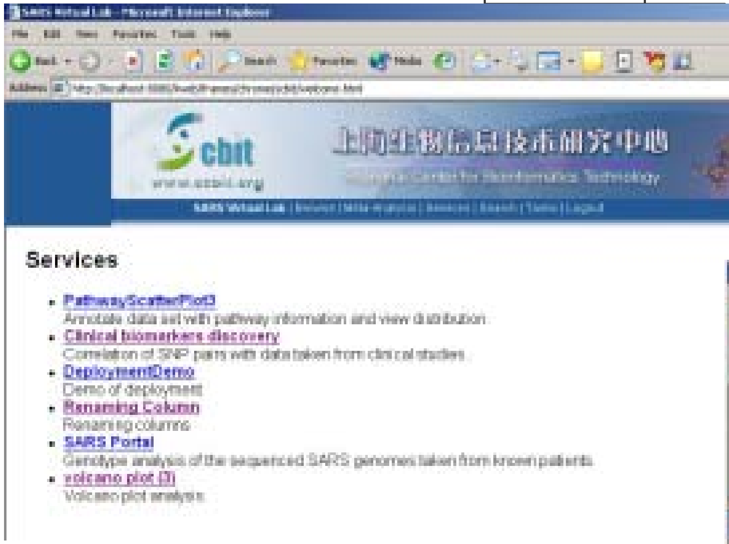
He J-F *et al* (2004)
Science **303**, 1666-1669

SARS: Delivering the knowledge

- **Access functions via the web**
- **Rapidly deploy workflows as web applications**
  - ***No programming required!***



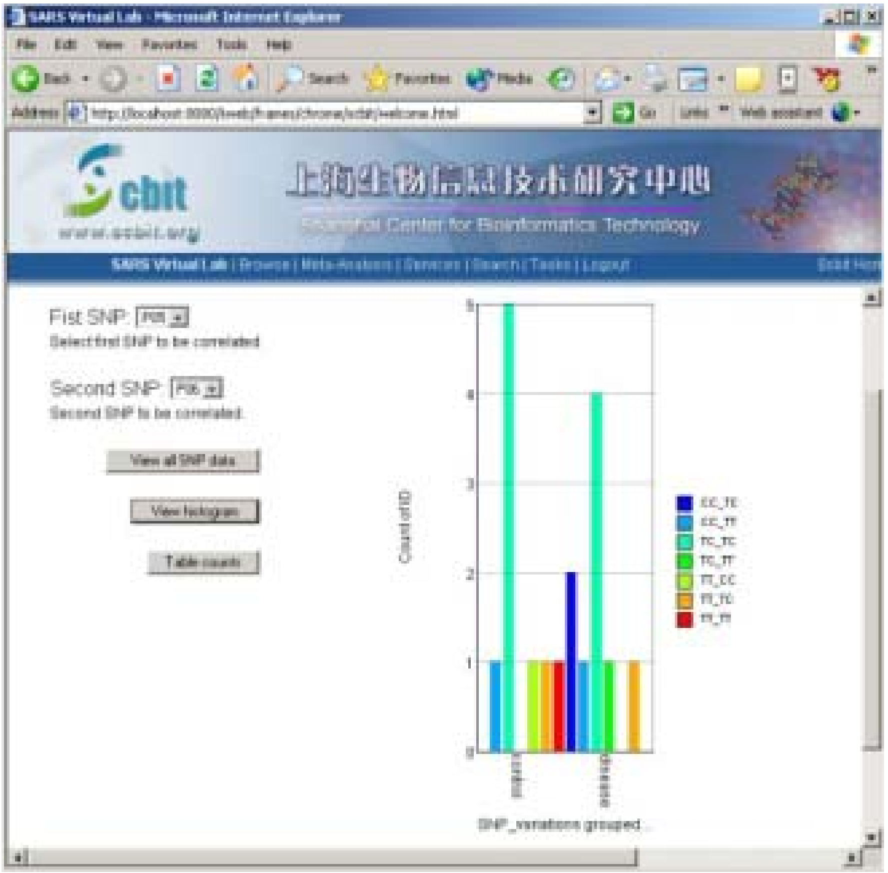
SARS: Delivering the knowledge


SARS Virtual Lab - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links Web assistant
cbit
www.scbit.org
上海生物信息技术研究中心
Shanghai Center for Bioinformatics Technology
SARS Virtual Lab | Browse | Meta-Analysis | Services | Search | Tasks | Logout

Services

- PathwayScatterPlot3
  Annotate data set with pathway information and view distribution
- Clinical biomarkers discovery
  Correlation of SNP pairs with data taken from clinical studies
- DeploymentDemo
  Demo of deployment
- Renaming Column
  Renaming columns
- SARS Portal
  Genotype analysis of the sequenced SARS genomes taken from known patients
- volcano plot (3)
  Volcano plot analysis

- **Updated as new datasets and analytical techniques become available**
- **Virtual Grid based research environment**


SARS Virtual Lab - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://localhost:8080/web/frames/chrome/scbit/welcome.html
Go Links Web assistant
cbit
www.scbit.org
上海生物信息技术研究中心
Shanghai Center for Bioinformatics Technology
SARS Virtual Lab | Browse | Meta-Analysis | Services | Search | Tasks | Logout
Count of ID
0
1
2
3
4
5
SNP_variations grouped
CC_TC
CC_TT
TC_TC
TC_TT
TT_CC
TT_TC
TT_TT

Fist SNP: [P05 ▾]
Select first SNP to be correlated

Second SNP: [P06 ▾]
Second SNP to be correlated.

View all SNP data

View histogram

Table counts

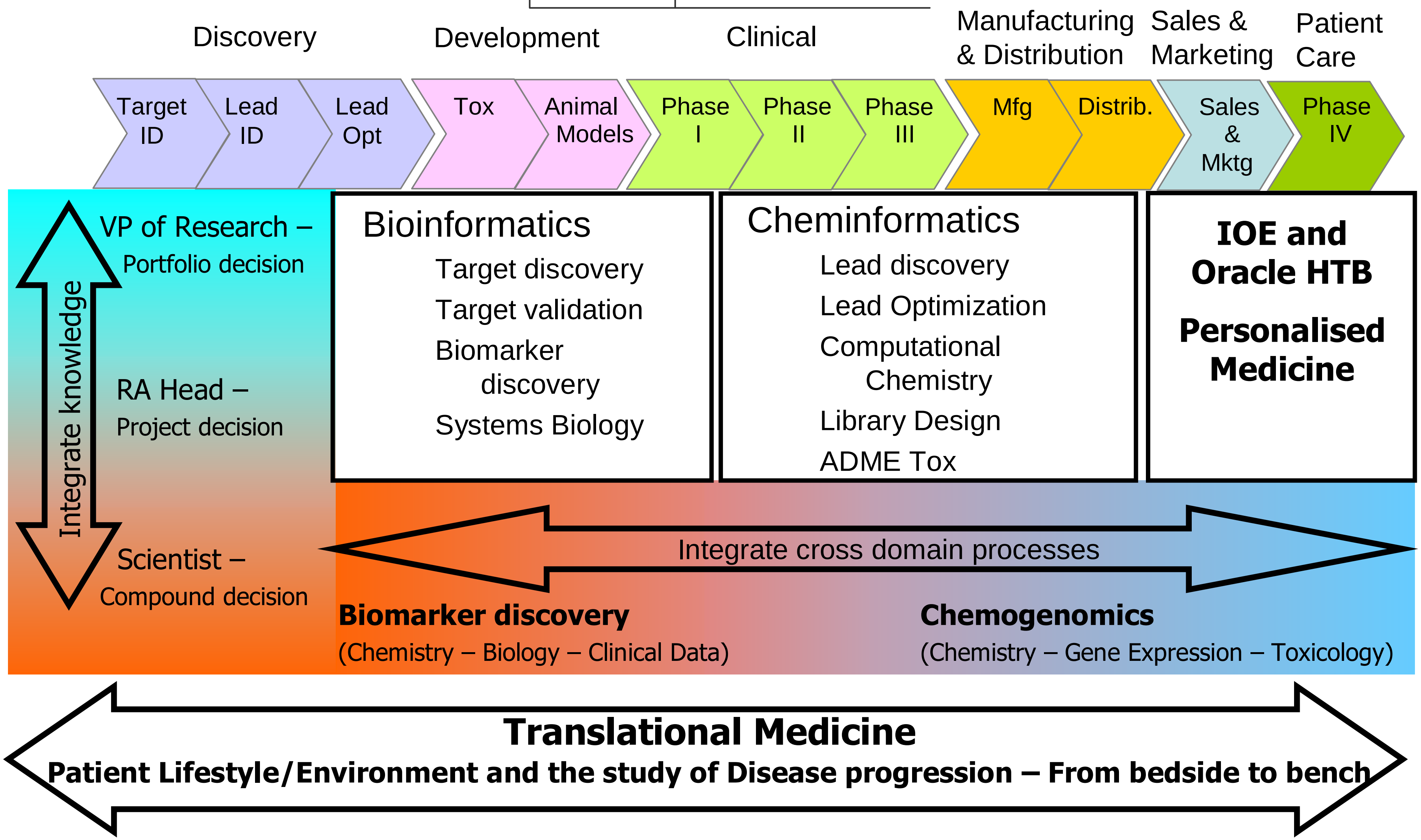
InforSense
The Integrative Analytics Company

SARS: What InforSense provided

- The ability to rapidly integrate multiple, disparate data services, available on the net
- The ability to integrate over 200 required compute services, including grid service infrastructure for compute intensive tasks
- Enable the knowledge and insights gained to be rapidly and automatically deployed as new web services and portals
- Minimise the investment required in tools and services – maximising the reuse of existing resources

InforSense
The Integrative Analytics Company

One integration platform from Discovery to Clinic, and bench to boardroom



Discovery | Development | Clinical | Manufacturing & Distribution | Sales & Marketing | Patient Care

Target ID → Lead ID → Lead Opt → Tox → Animal Models → Phase I → Phase II → Phase III → Mfg → Distrib. → Sales & Mktg → Phase IV

Integrate knowledge

VP of Research – Portfolio decision

RA Head – Project decision

Scientist – Compound decision

Bioinformatics

Target discovery
Target validation
Biomarker discovery
Systems Biology

Cheminformatics

Lead discovery
Lead Optimization
Computational Chemistry
Library Design
ADME Tox

IOE and Oracle HTB

Personalised Medicine

Integrate cross domain processes

Biomarker discovery
(Chemistry – Biology – Clinical Data)

Chemogenomics
(Chemistry – Gene Expression – Toxicology)

Translational Medicine

Patient Lifestyle/Environment and the study of Disease progression – From bedside to bench