

Semantic data integration generating value in drug discovery

ICIC 2006 Nimes
October 25, 2006

Dr. Daniel Keesman

- + **Introduction**

- + **Tour of three examples of integration ...**
 - Target discovery and validation for holiday planning
 - Lead discovery by high throughput biomolecular screening
 - Target discovery and validation for drug discovery

- + **Along the way ...**
 - Challenges
 - Solutions
 - Technologies employed
 - Value generated

Levels of integration

+ Level 1 — technical integration

- Data can be aggregated across systems regardless of their platform, operating system, or location. The standards used for the Internet and the World Wide Web are the answer to this initial level of information integration (e.g. TCP/IP).

+ Level 2 — syntactic integration

- Information sources agree on a common syntax for exchanging such information. HTML and XML serve this purpose in most integration solutions today. Web services and service-oriented integration can provide a more formal framework for XML-based syntactic integration.

+ Level 3 — semantic integration

- A formal mapping of the meaning of terms from different information sources is developed and a knowledge model of the relationships among these terms is created. Built on top of the first two levels, this lets you move data in and out of systems while ensuring that the data is referring to the same thing.

(Source: Warner and Vap, Business Integration Journal)

Semantic technologies

- + Semantic technologies represent meaning through connectivity. The meaning of terms, or concepts, in the model is established by the way they connect to each other.
- + A semantic model expresses multiple viewpoints.
- + The models are consultable (accessible) by applications at runtime.
- + Semantic applications are thin because they work with *smart* data. All the business rules logic is held in the models shared across applications.

Traditional application and data integration approaches focus on moving and transforming data elements between systems.

Semantics is about combining and correlating data elements amongst themselves to understand relationships, spot trends, and create new information from new combinations.

(Source: TopQuadrant)

Integration example I

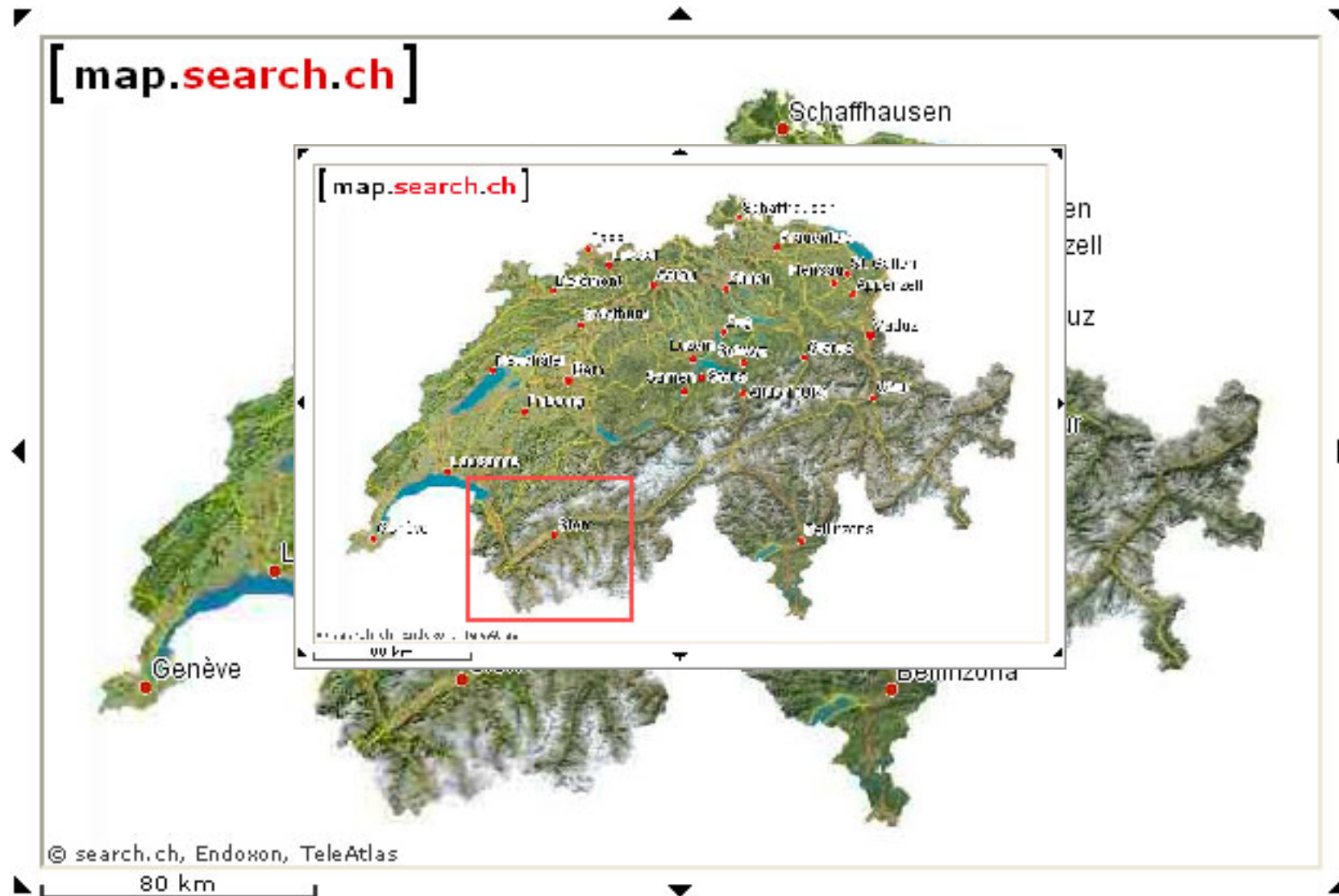
- + An example from daily life
- + Target discovery and validation for holiday planning



Integration example I



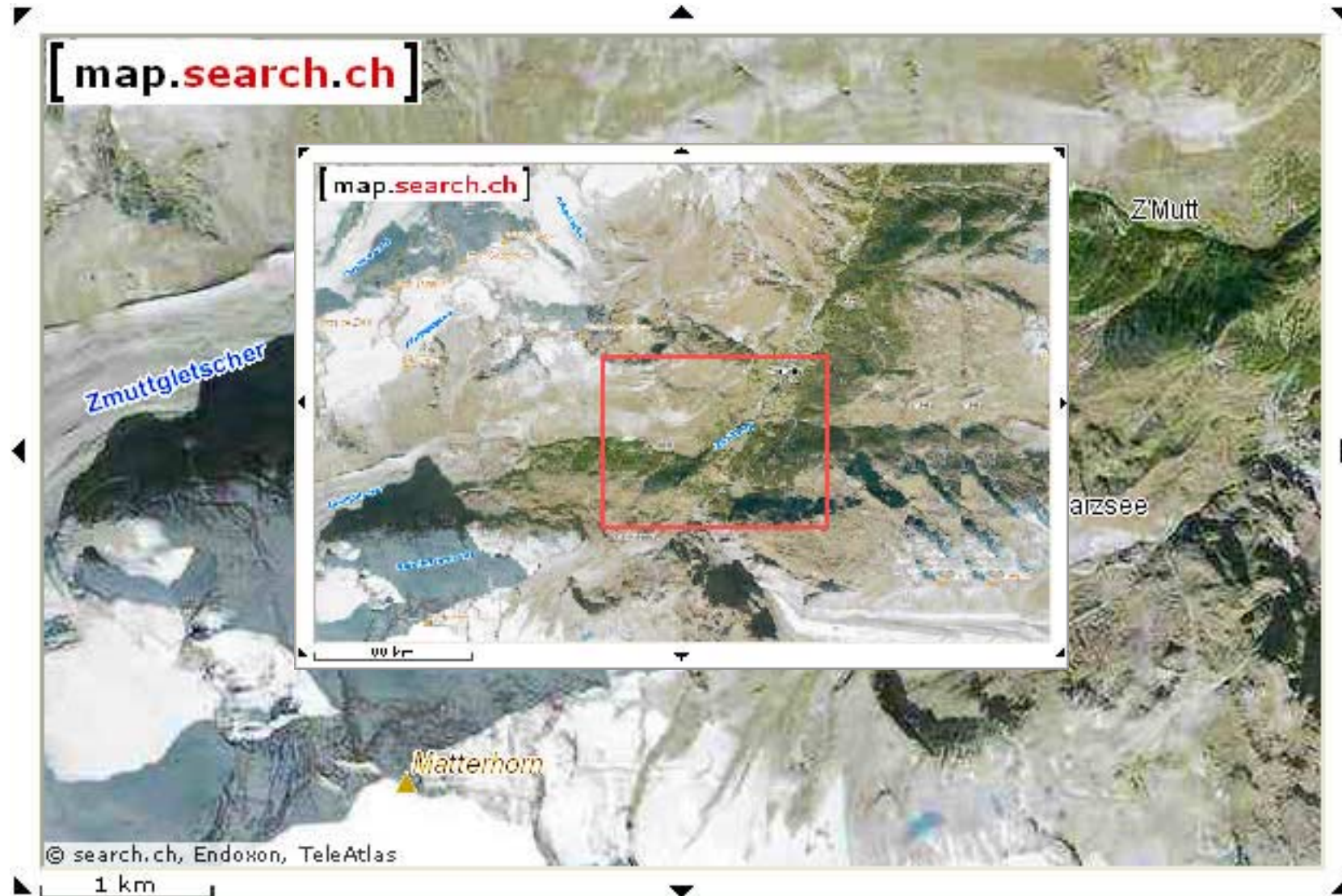
Integration example I



Integration example I



Integration example I



Integration example I



Integration example I

The screenshot shows a satellite map from map.search.ch with a yellow information popup. The popup contains the following text:

Cable car
Zermatt ZBAG-lz [SBB-departures](#)
[Line](#) [Departures](#) [Direction](#)
✓ 14:47 14:47 Furi
SBB Timetable to here [from here](#)
From: »
To: Zermatt ZBAG-lz

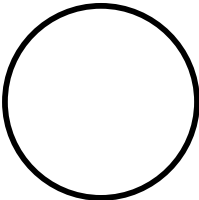
The map background shows a street network with labels like 'Schünmattstrasse', 'Blalweg', 'Wiesbodenweg', and 'Staldenstrasse'. A scale bar at the bottom left indicates 50 meters. The URL 'map.search.ch' is visible in the top left corner of the map area.

Integration example I

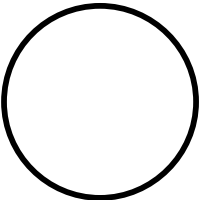


Vacation planning arena

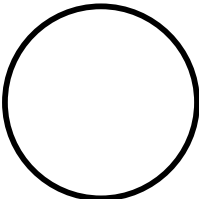
Topography



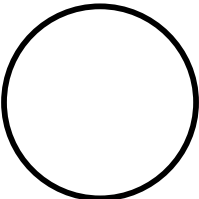
Cable cars



Hotels

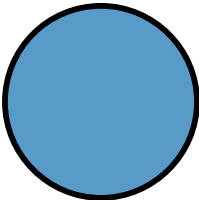


Train stations

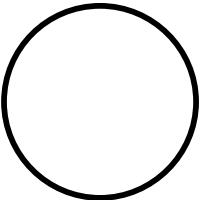


Vacation planning arena

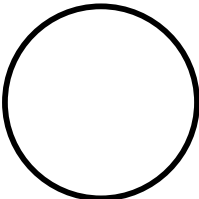
Topography



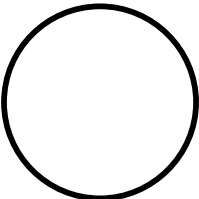
Cable cars



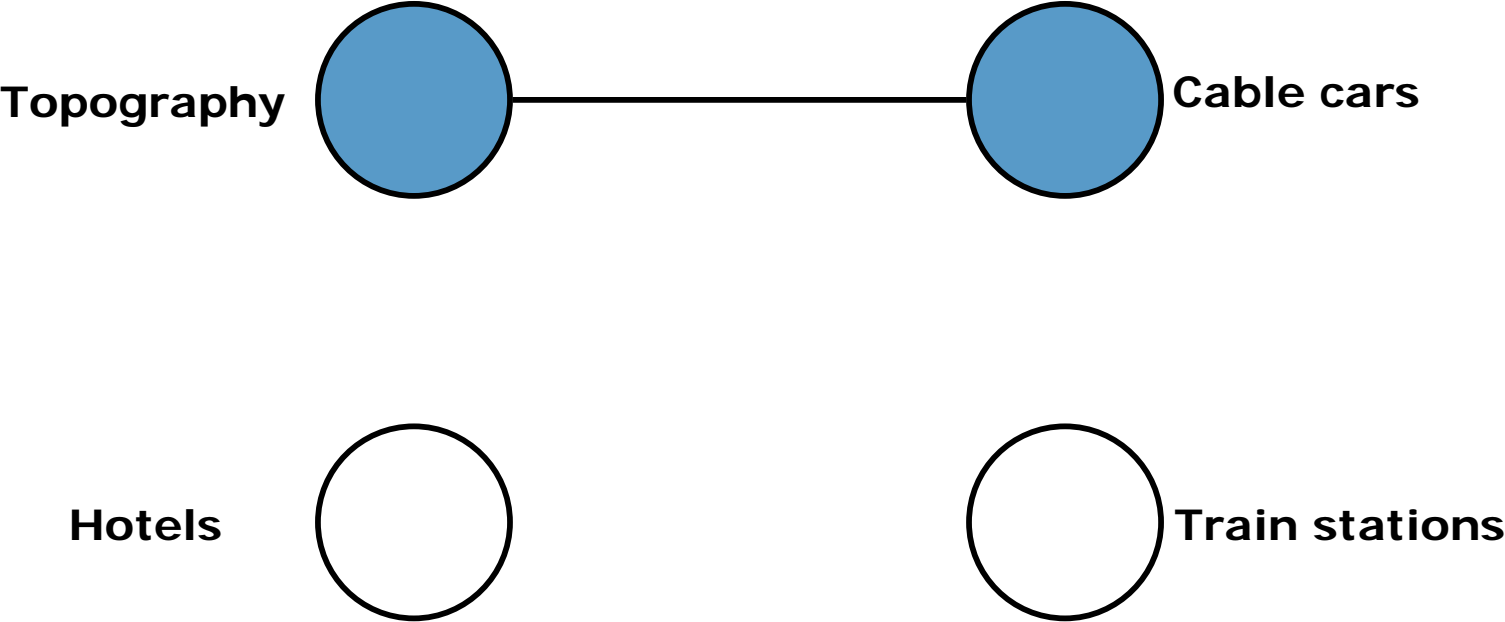
Hotels



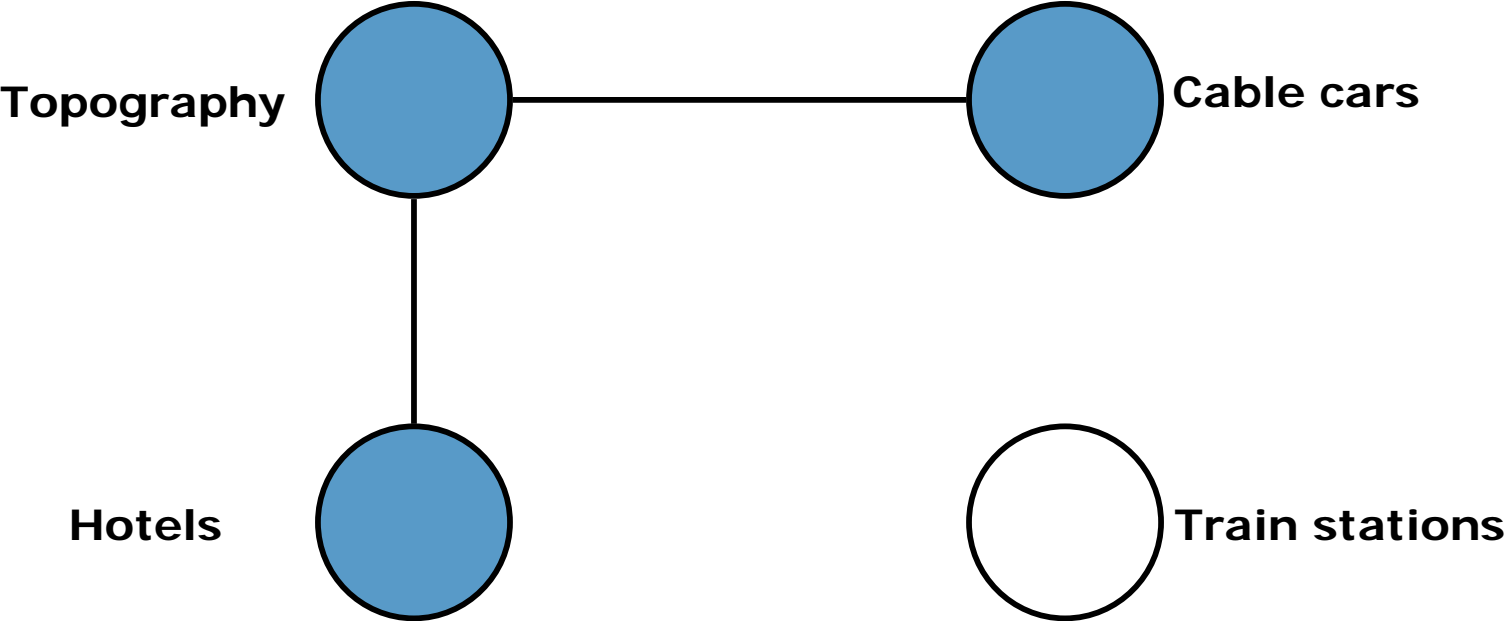
Train stations



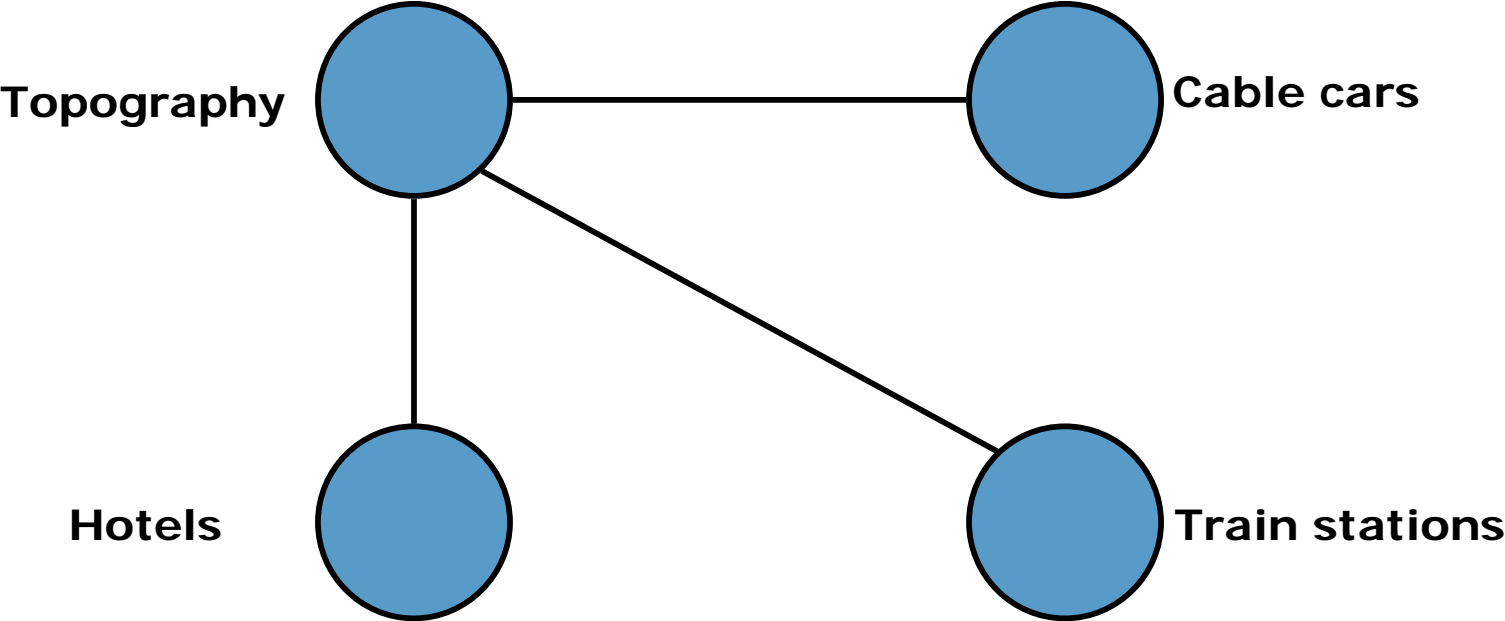
Vacation planning arena



Vacation planning arena

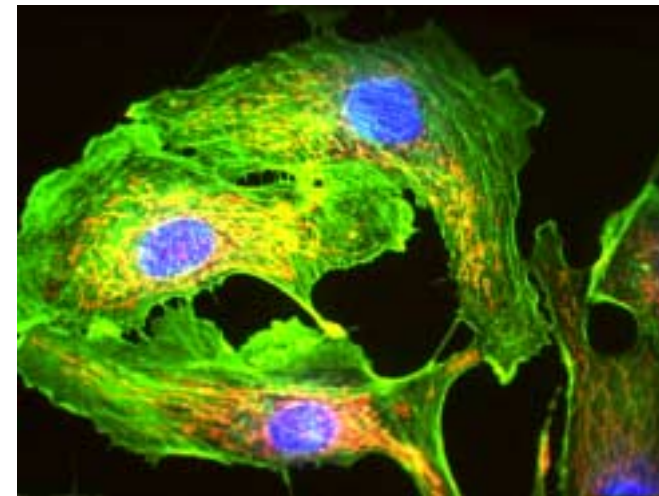
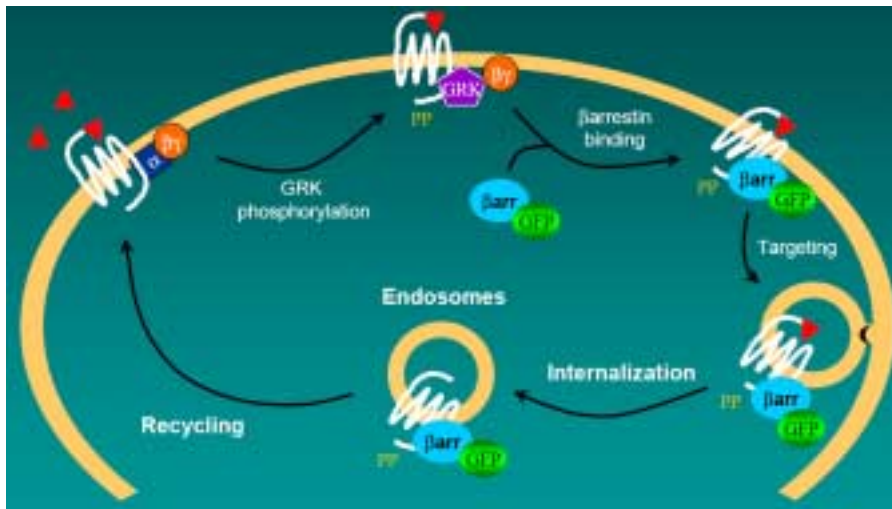


Vacation planning arena



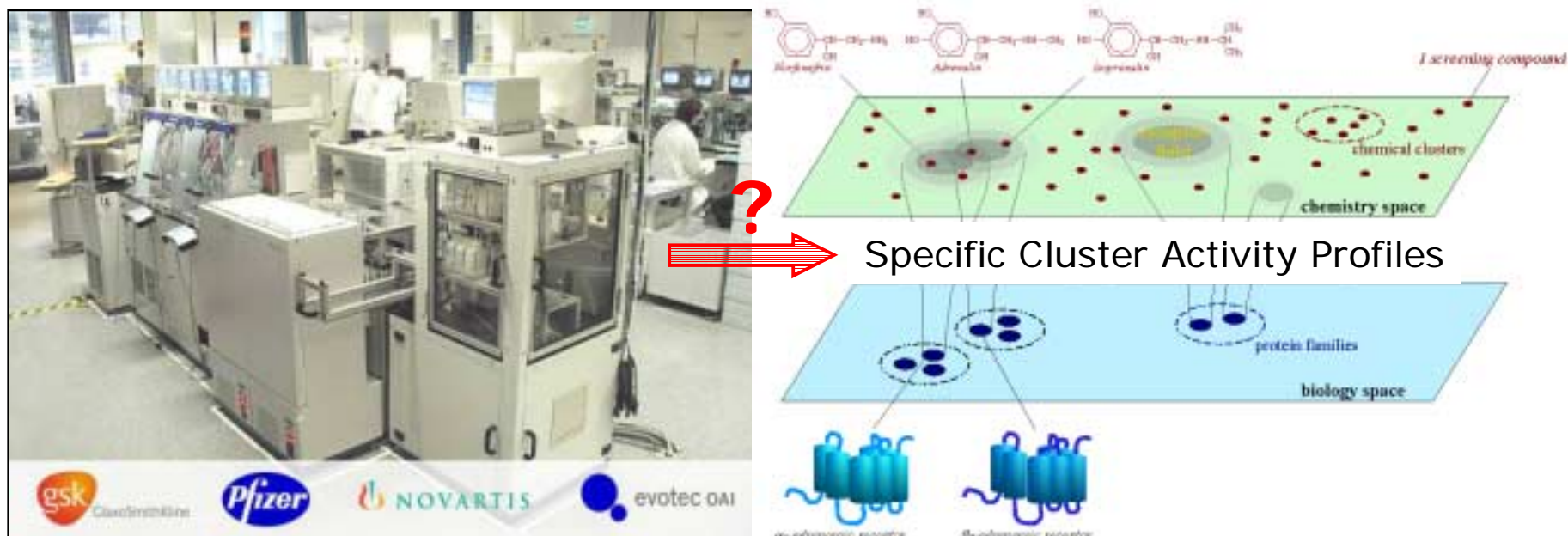
Integration example II

- + Lead discovery
- + High throughput, high content screening
 - In-depth biological assessment of entire libraries



Challenges in HTS Data Handling: Automated and Scalable Processing and Review

- + HTS: Screening 100'000s of compounds for bioactivity in 100s of - more or less defined - in vitro test systems
- + To: Systematically identify novel chemistries with a desired pharmacological activity



Analysis:

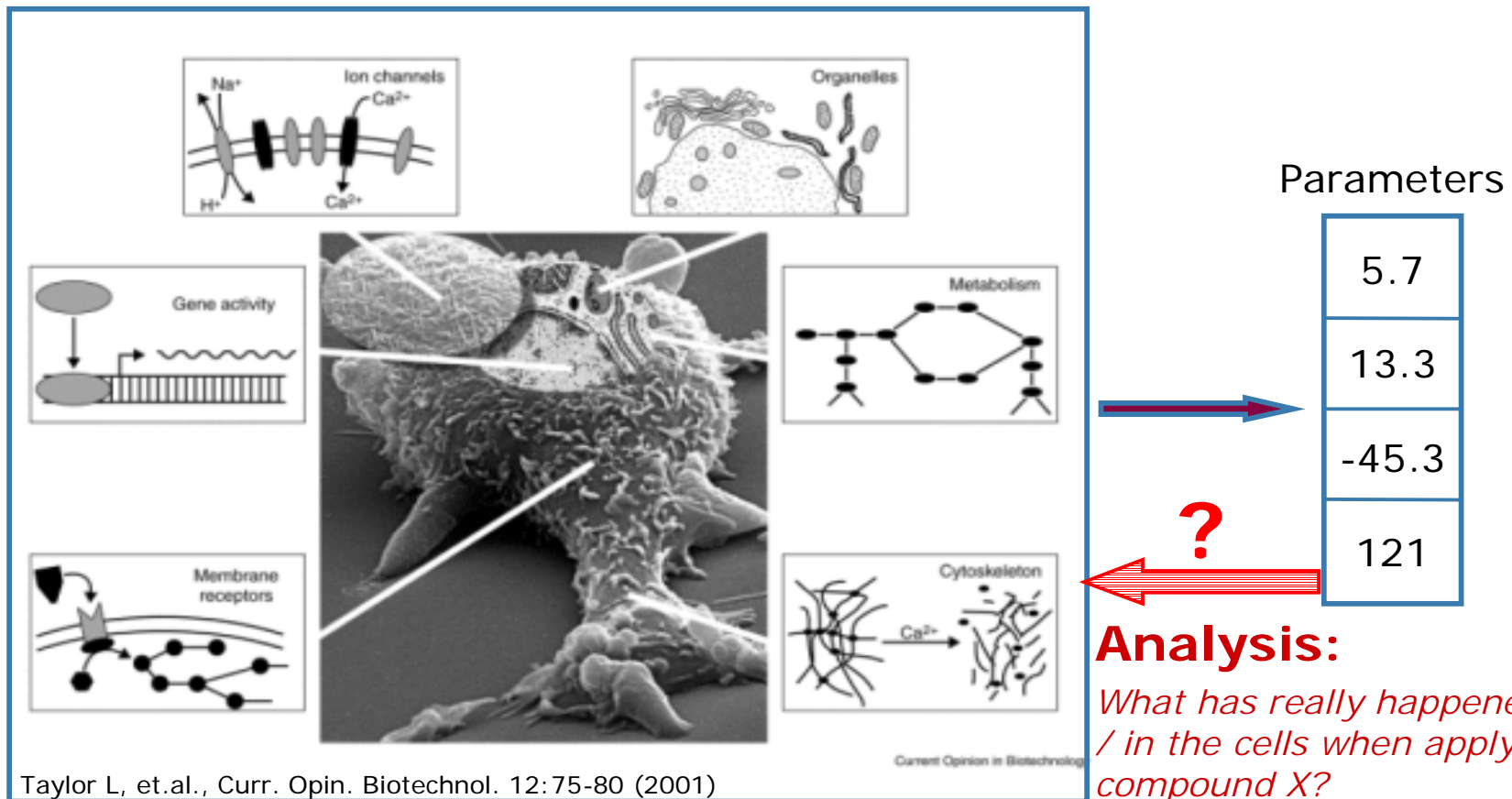
Which compounds were really specifically active and selective?

*Fisher HP & Heyse S,
Curr. Opin. Drug Disc.
Devel. 2005 8(3):334.*

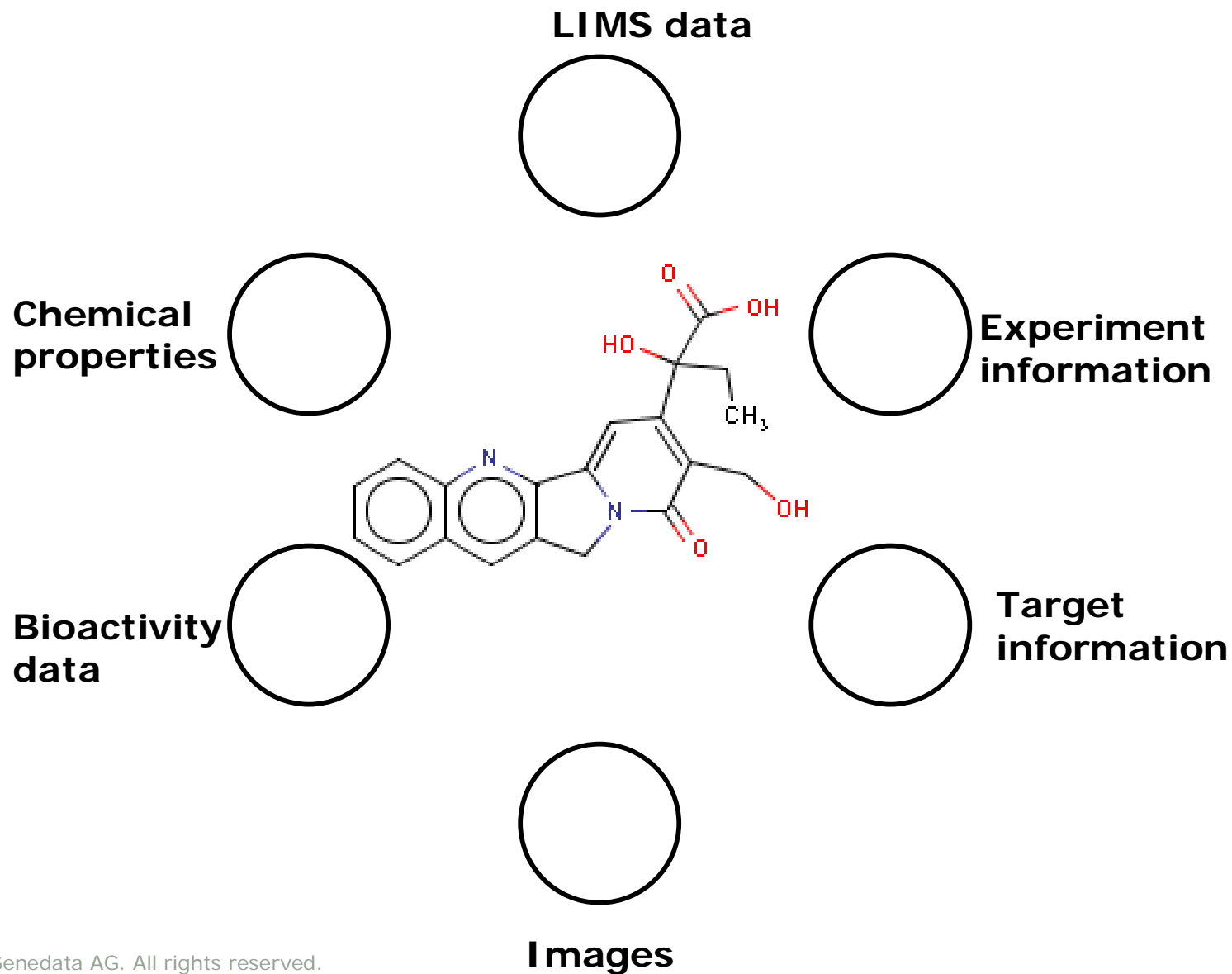
Challenge in HCS Data Handling: Interpretation of Complex, Large-scale Data

HCS: Probing complex systems, measuring few variables, on a large scale ...

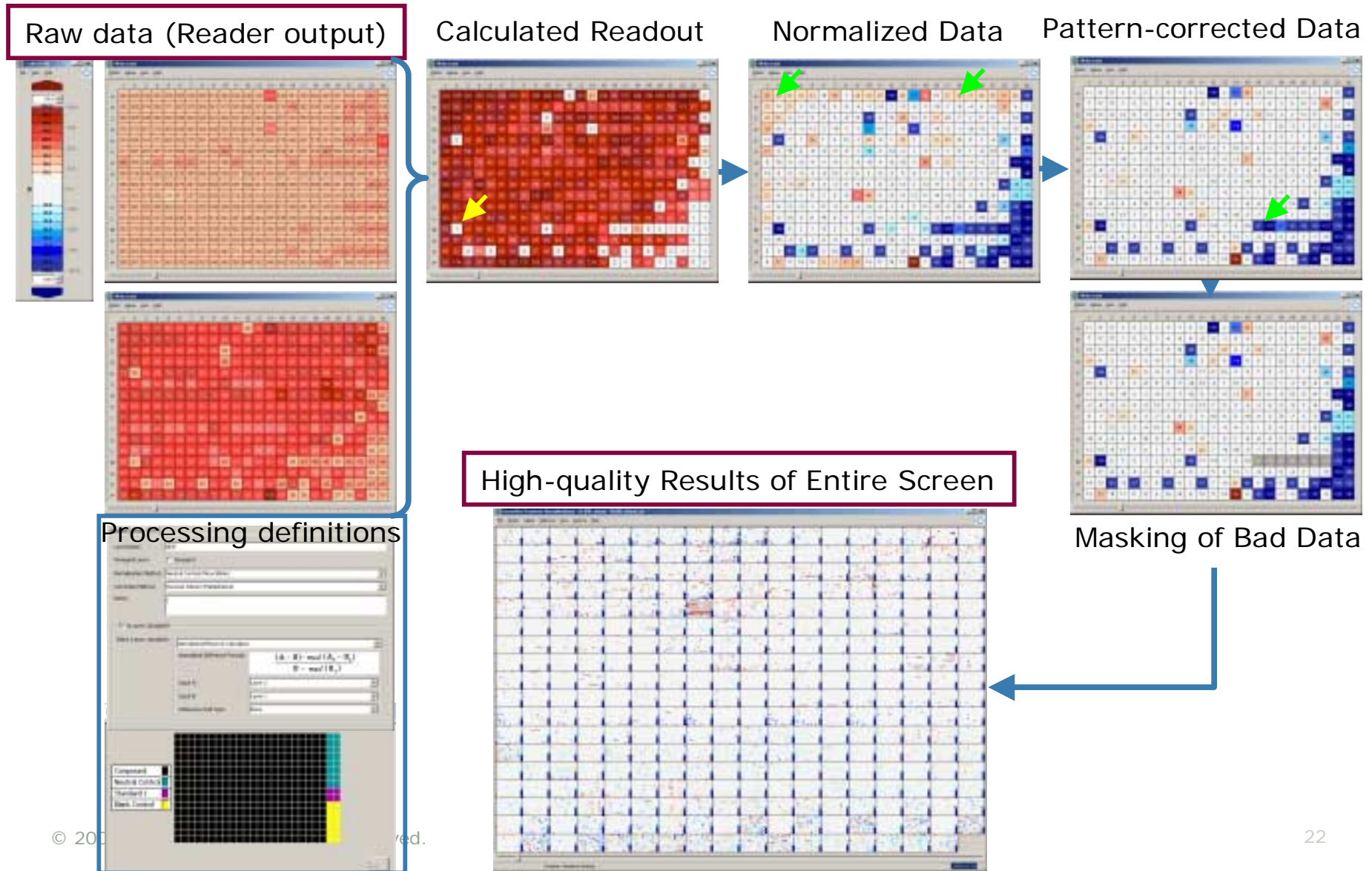
To: Identify novel chemistries with a defined pharmacological activity



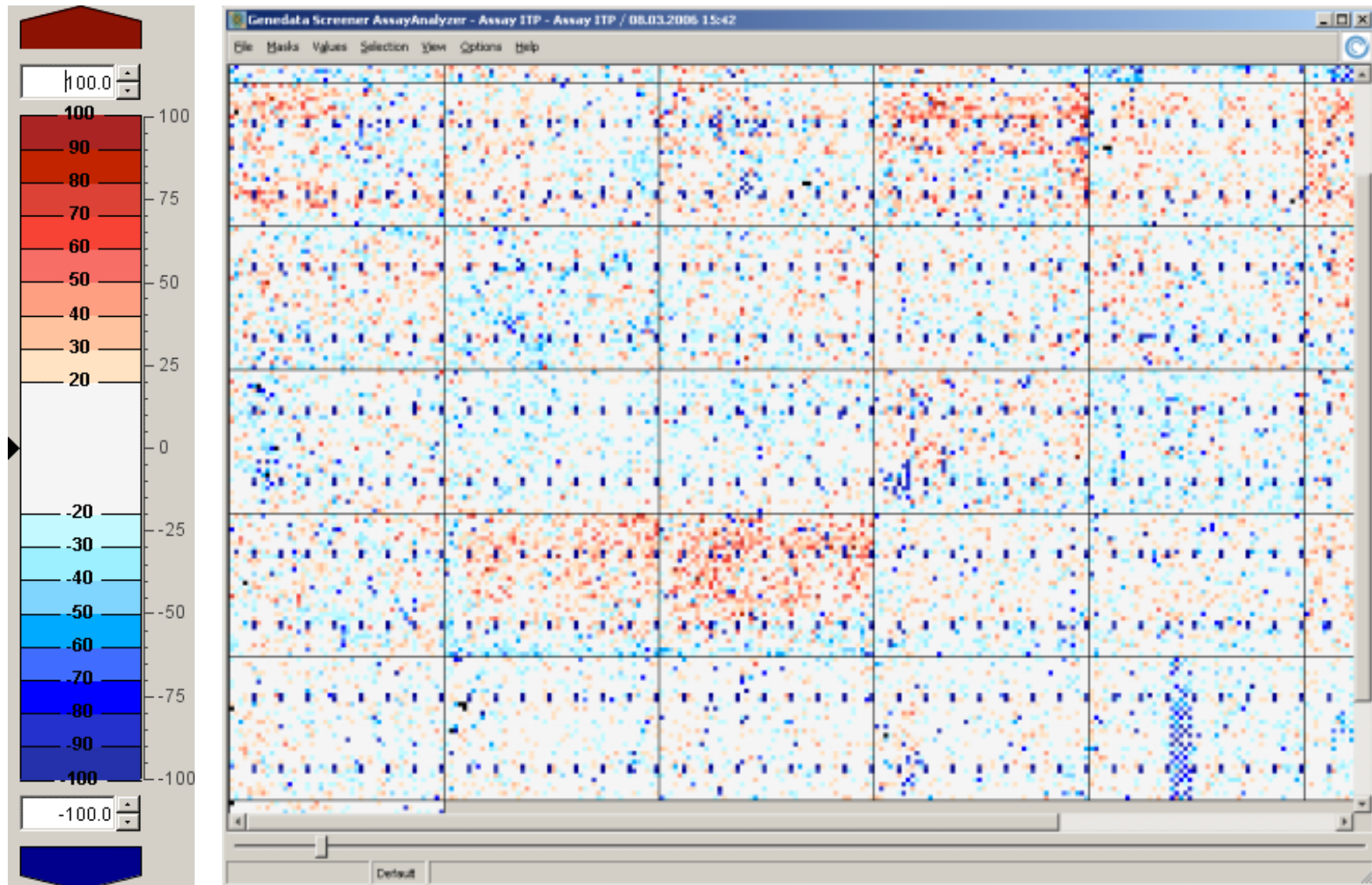
Lead discovery arena – Semantic integration of chemical data



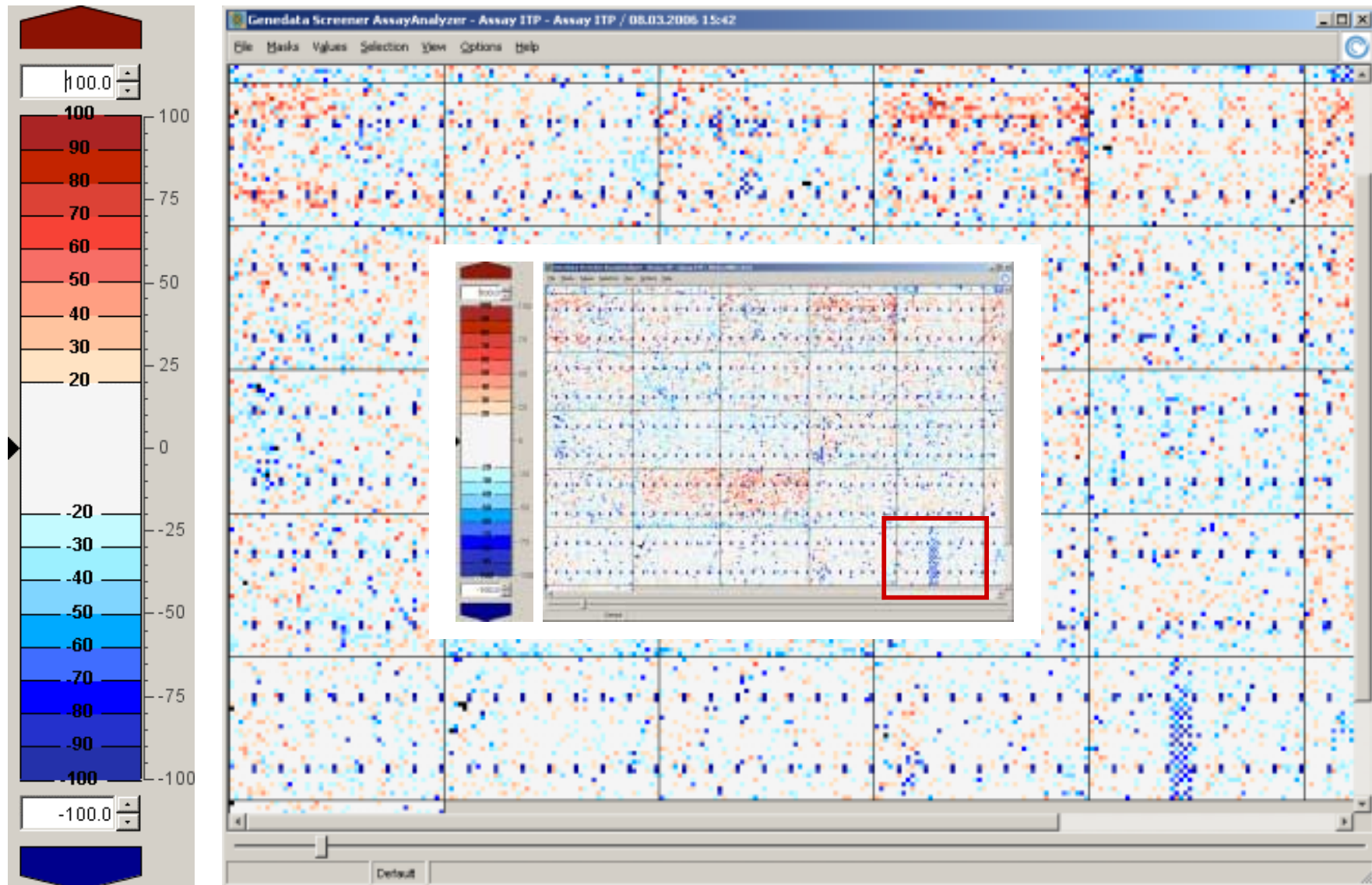
Solution: Automated System Coupling Process-Specific Business Logic with Rapid Review



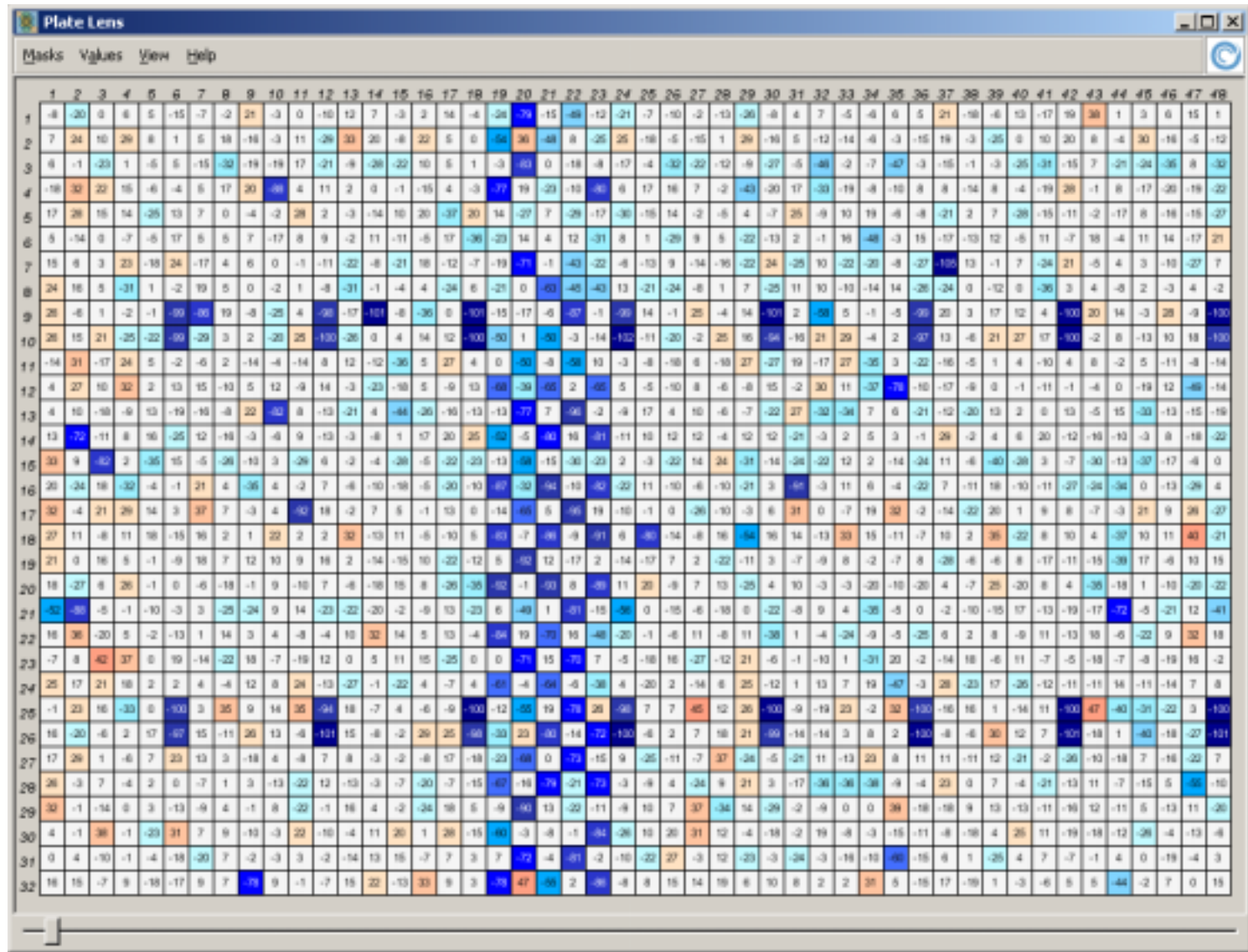
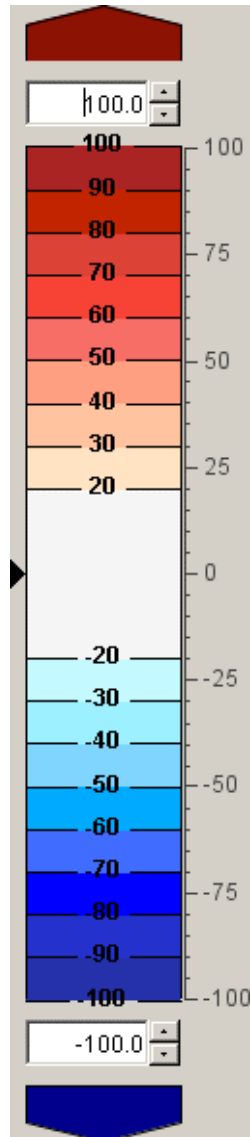
Integration example II



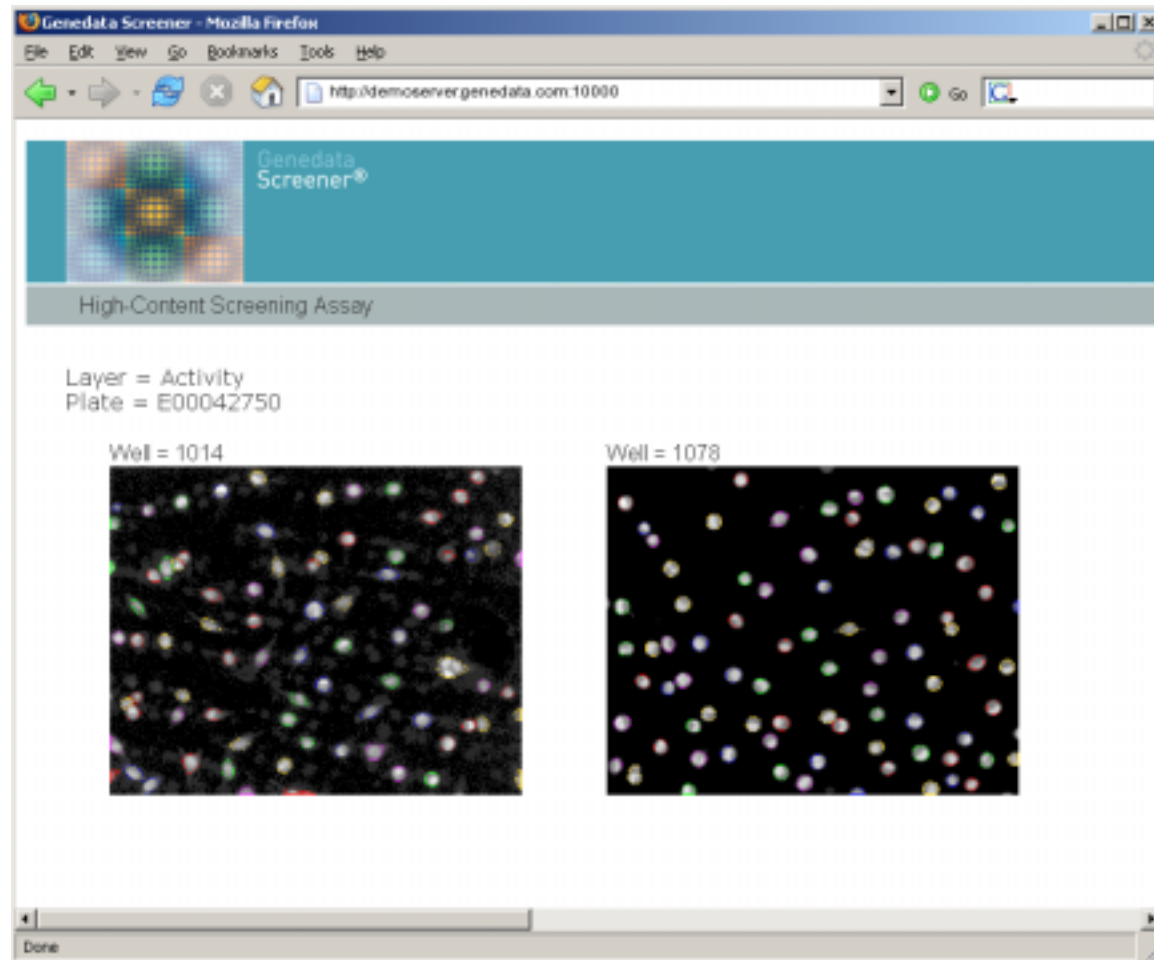
Integration example II



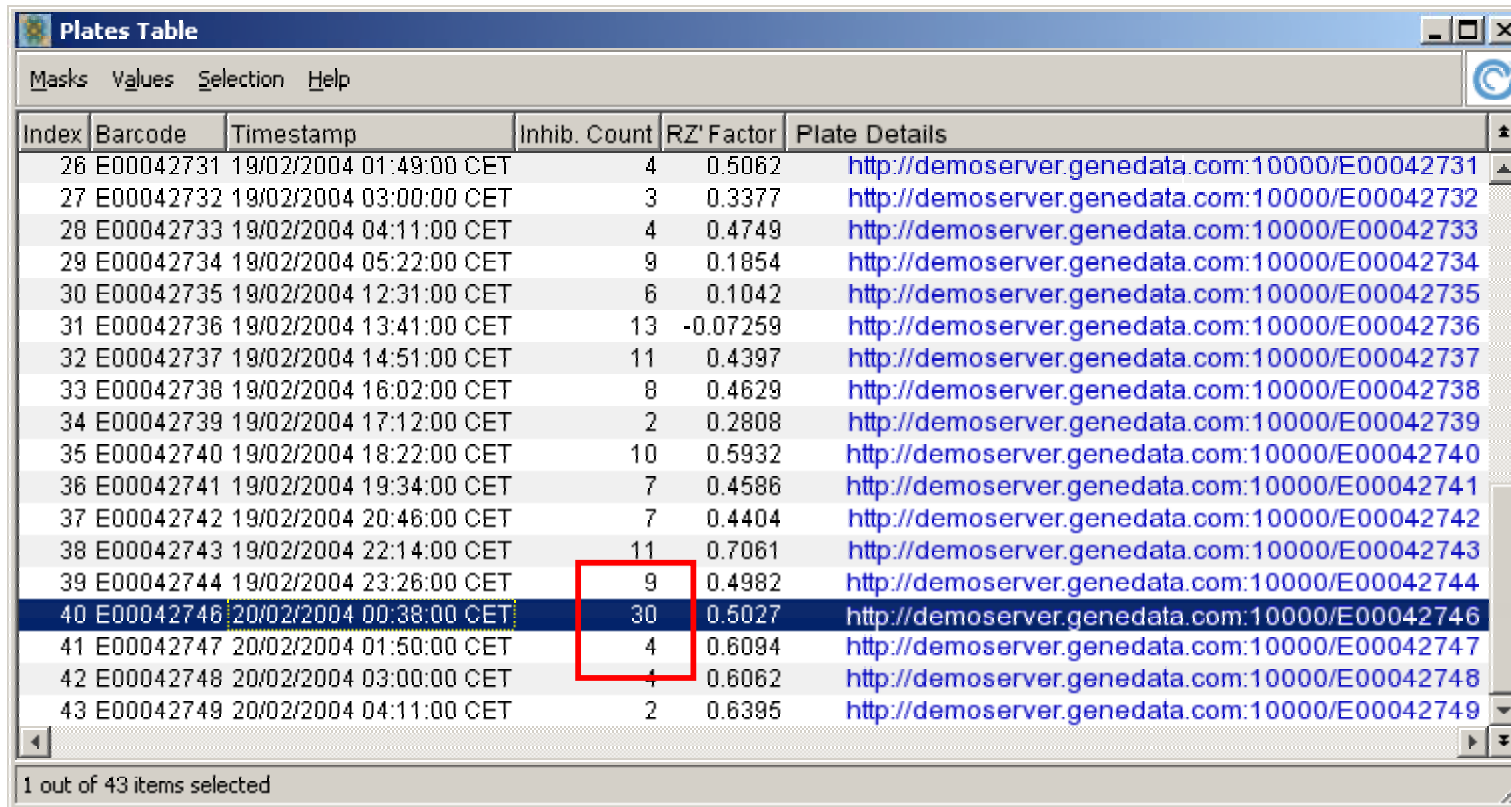
Integration example II



Integration example II



Integration example II



Index	Barcode	Timestamp	Inhib. Count	RZ' Factor	Plate Details
26	E00042731	19/02/2004 01:49:00 CET	4	0.5062	http://demosever.genedata.com:10000/E00042731
27	E00042732	19/02/2004 03:00:00 CET	3	0.3377	http://demosever.genedata.com:10000/E00042732
28	E00042733	19/02/2004 04:11:00 CET	4	0.4749	http://demosever.genedata.com:10000/E00042733
29	E00042734	19/02/2004 05:22:00 CET	9	0.1854	http://demosever.genedata.com:10000/E00042734
30	E00042735	19/02/2004 12:31:00 CET	6	0.1042	http://demosever.genedata.com:10000/E00042735
31	E00042736	19/02/2004 13:41:00 CET	13	-0.07259	http://demosever.genedata.com:10000/E00042736
32	E00042737	19/02/2004 14:51:00 CET	11	0.4397	http://demosever.genedata.com:10000/E00042737
33	E00042738	19/02/2004 16:02:00 CET	8	0.4629	http://demosever.genedata.com:10000/E00042738
34	E00042739	19/02/2004 17:12:00 CET	2	0.2808	http://demosever.genedata.com:10000/E00042739
35	E00042740	19/02/2004 18:22:00 CET	10	0.5932	http://demosever.genedata.com:10000/E00042740
36	E00042741	19/02/2004 19:34:00 CET	7	0.4586	http://demosever.genedata.com:10000/E00042741
37	E00042742	19/02/2004 20:46:00 CET	7	0.4404	http://demosever.genedata.com:10000/E00042742
38	E00042743	19/02/2004 22:14:00 CET	11	0.7061	http://demosever.genedata.com:10000/E00042743
39	E00042744	19/02/2004 23:26:00 CET	9	0.4982	http://demosever.genedata.com:10000/E00042744
40	E00042746	20/02/2004 00:38:00 CET	30	0.5027	http://demosever.genedata.com:10000/E00042746
41	E00042747	20/02/2004 01:50:00 CET	4	0.6094	http://demosever.genedata.com:10000/E00042747
42	E00042748	20/02/2004 03:00:00 CET	4	0.6062	http://demosever.genedata.com:10000/E00042748
43	E00042749	20/02/2004 04:11:00 CET	2	0.6395	http://demosever.genedata.com:10000/E00042749

1 out of 43 items selected

Integration example II

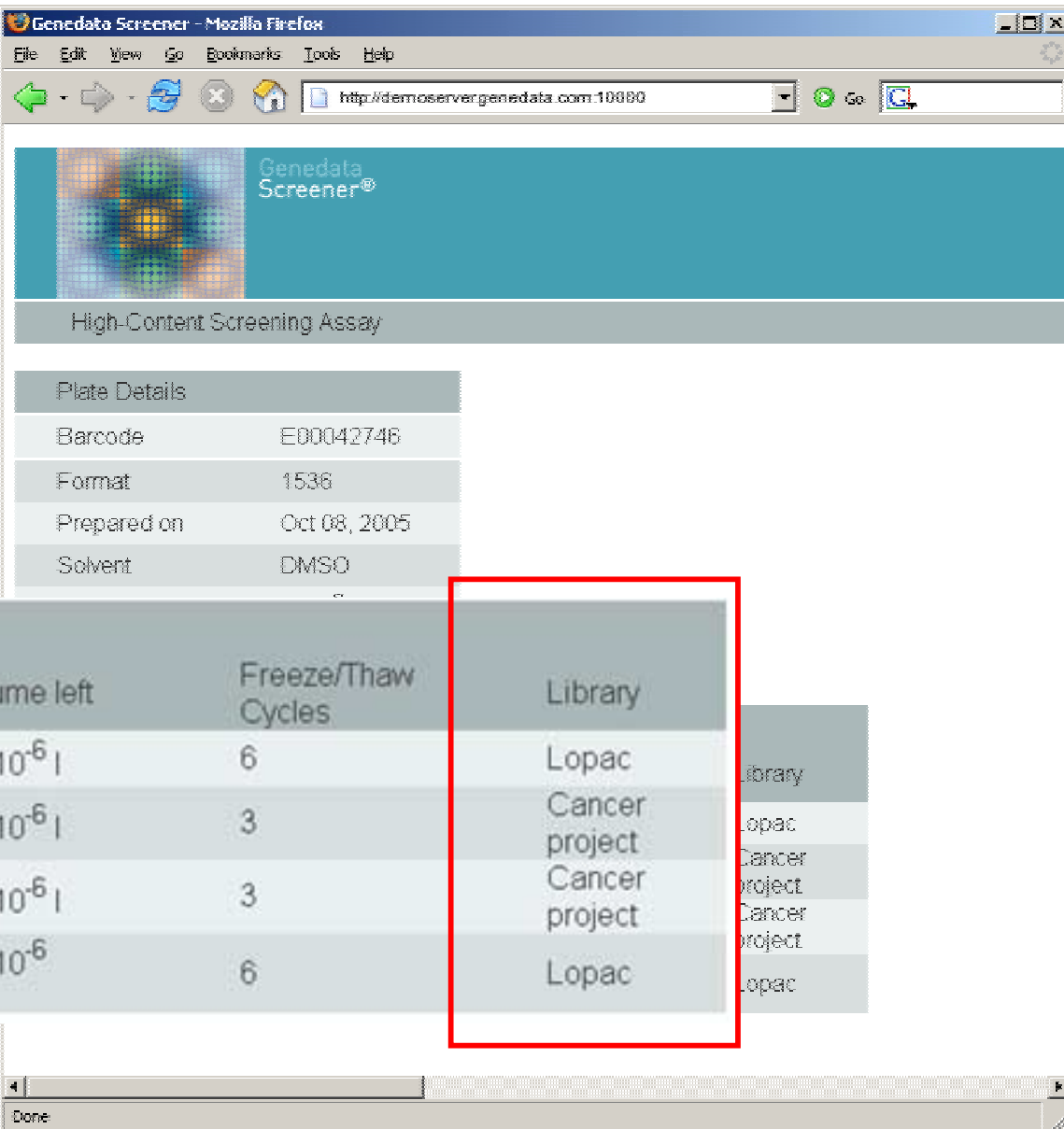


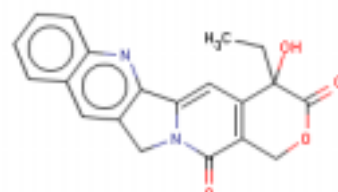
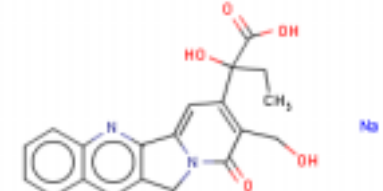
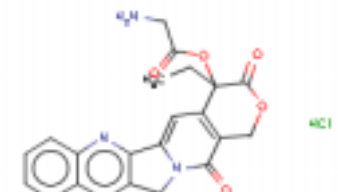

Plate Details

Barcode	E00042746
Format	1536
Prepared on	Oct 08, 2005
Solvent	DMSO

Motherplates

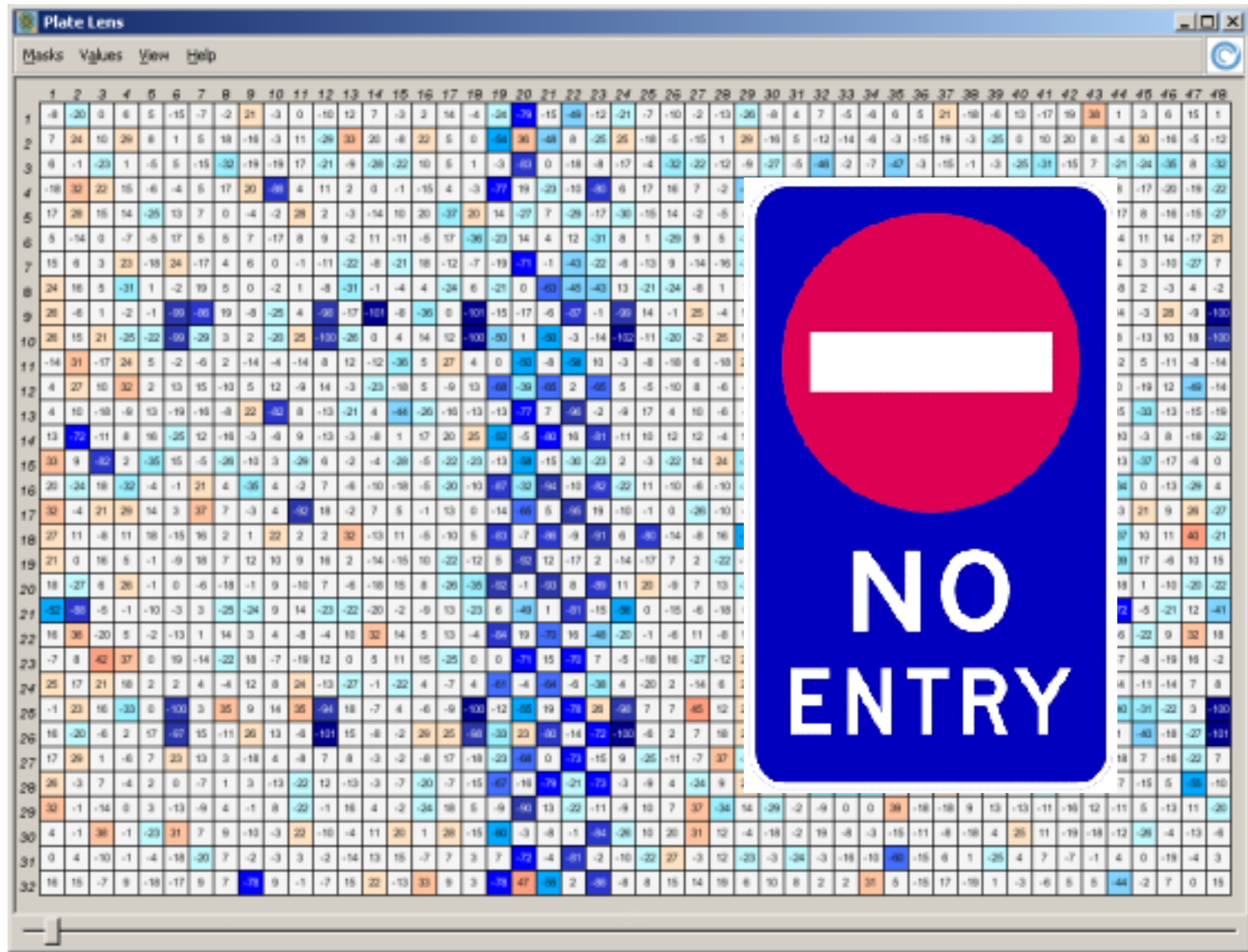
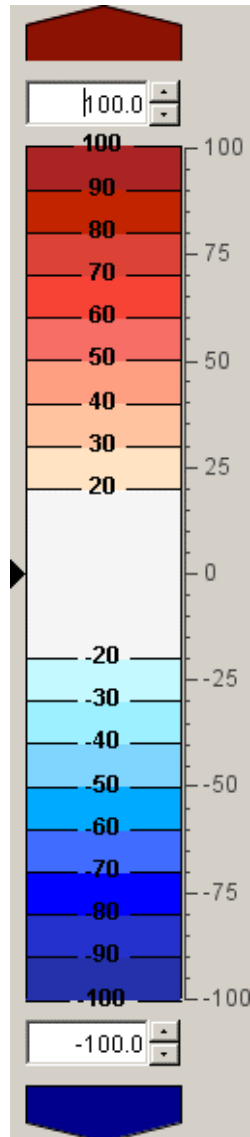
Barcode	Volume left	Freeze/Thaw Cycles	Library
M493821	10*10 ⁻⁶ l	6	Lopac
M789325	50*10 ⁻⁶ l	3	Cancer project
M789326	50*10 ⁻⁶ l	3	Cancer project
M483597	12*10 ⁻⁶ l	6	Lopac

Integration example II

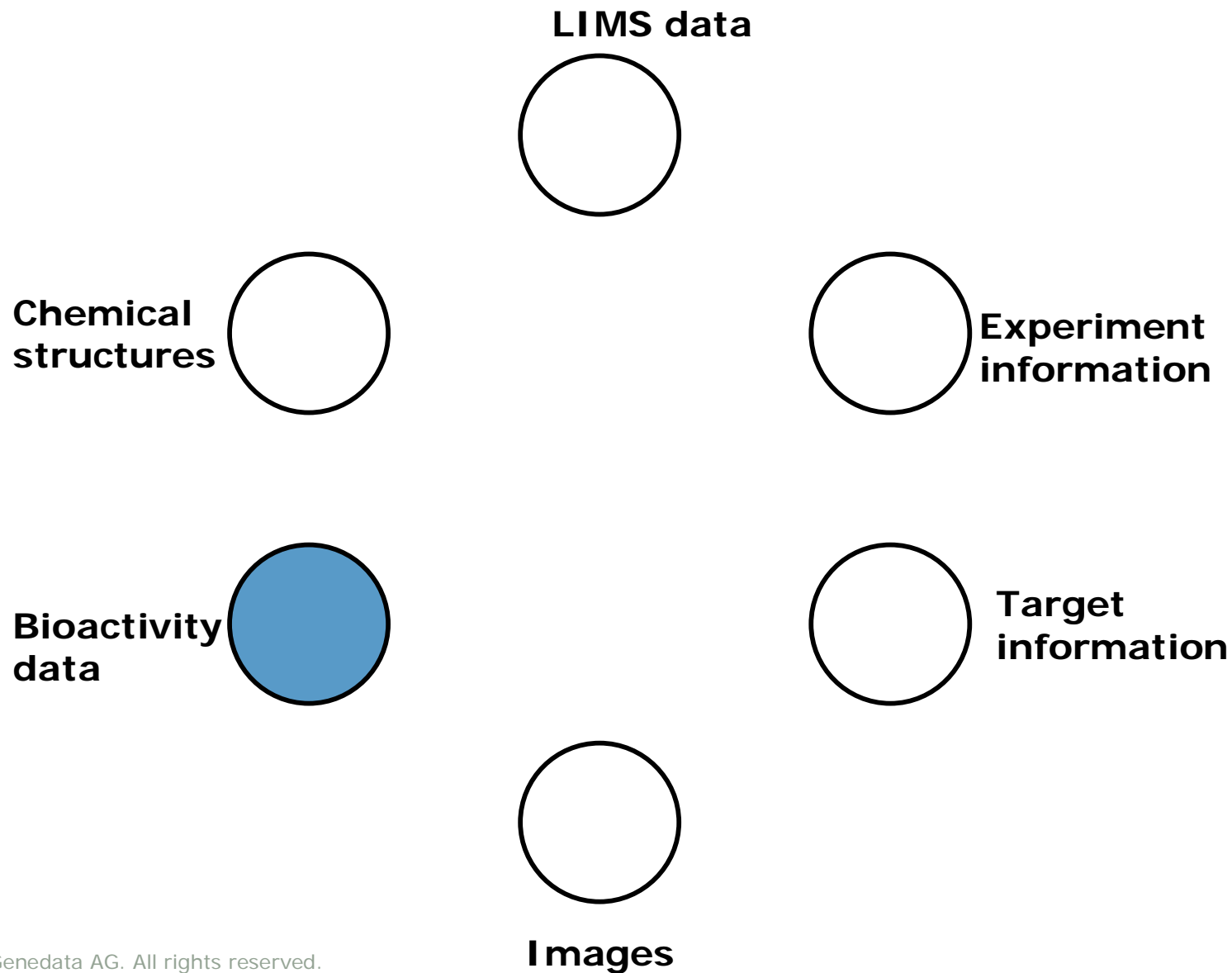
ID	Structure
NCL_NSC_094600	
NCL_NSC_100880	
NCL_NSC_606985	
	

4 items

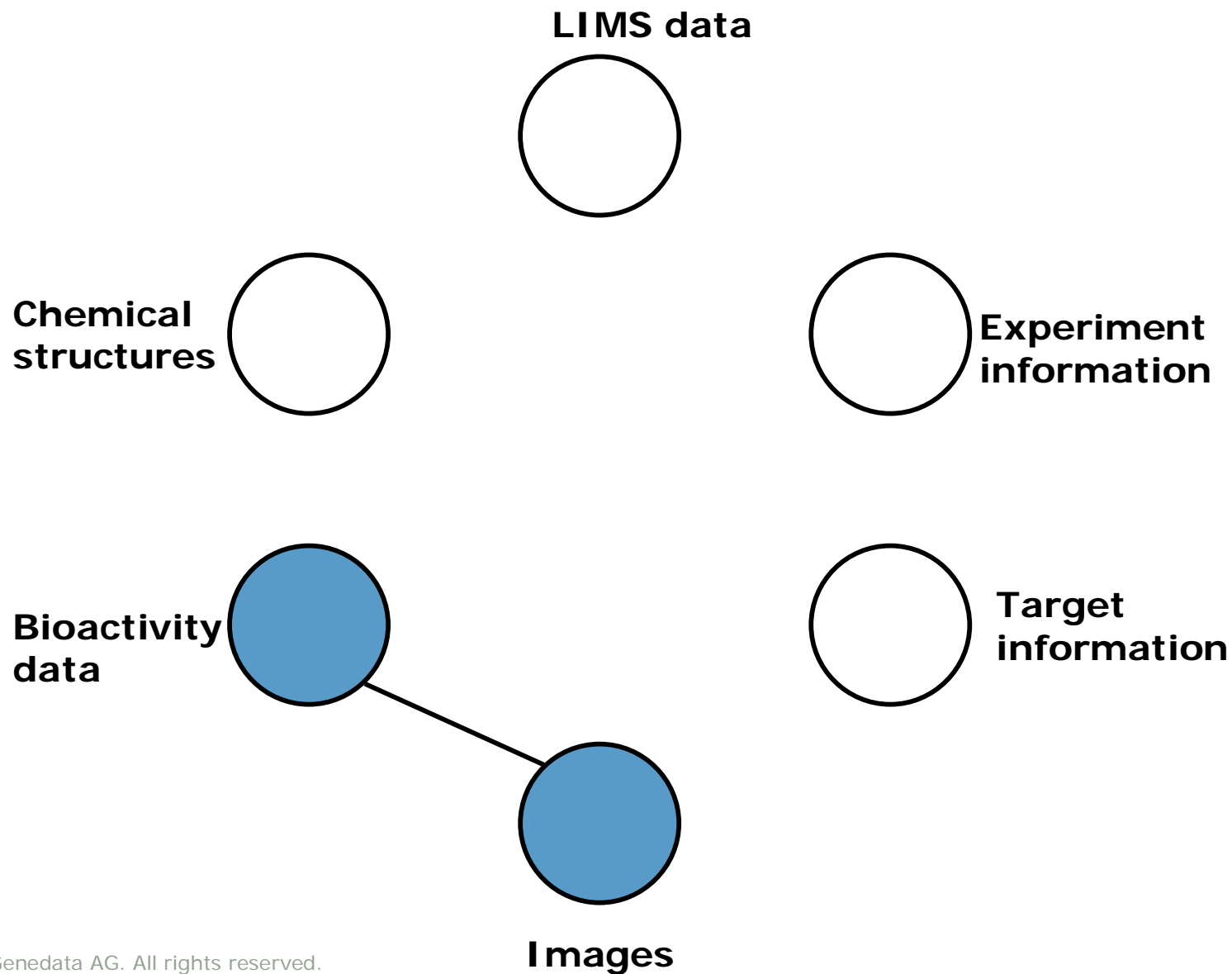
Integration example II



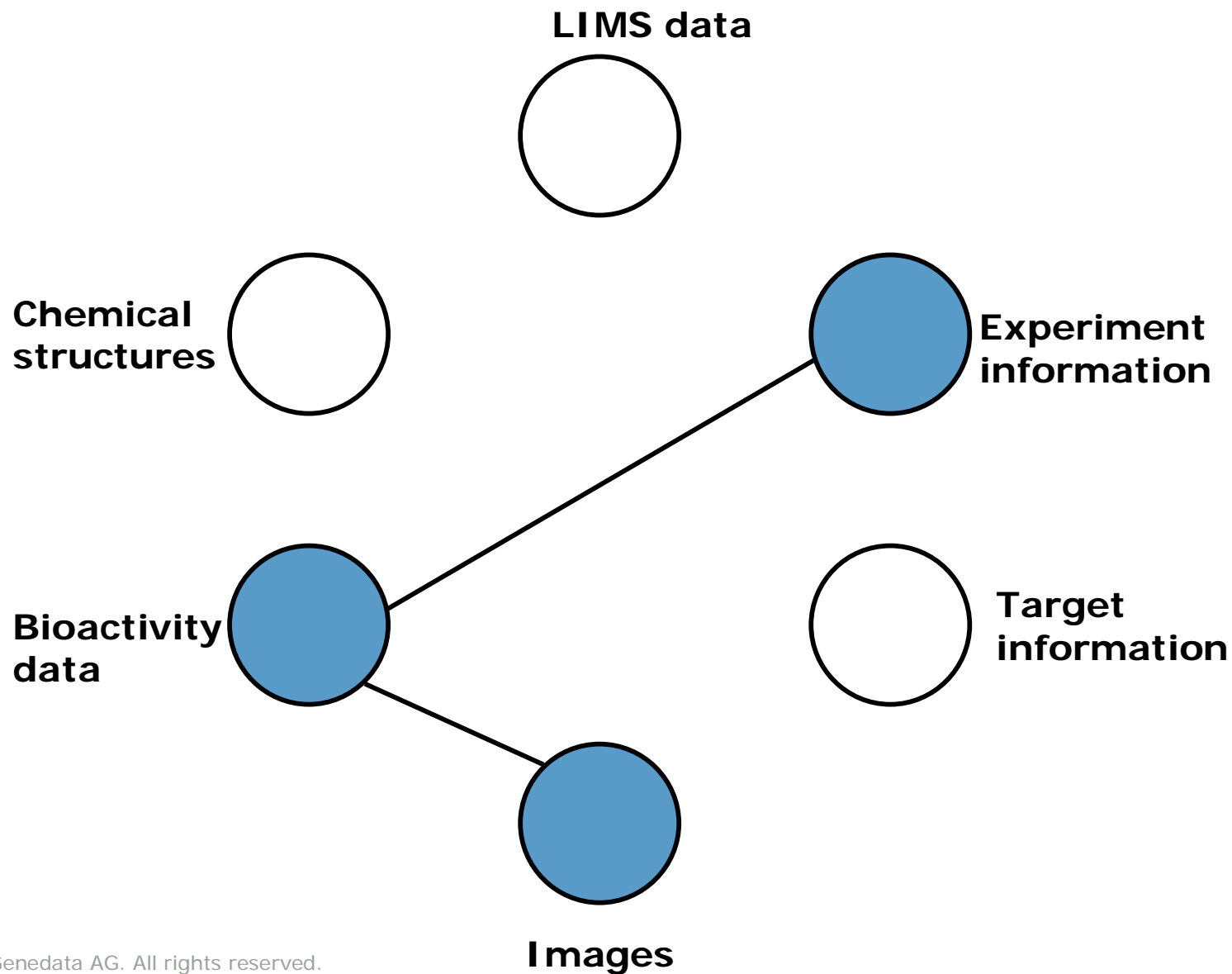
Lead discovery arena – Semantic integration of chemical data



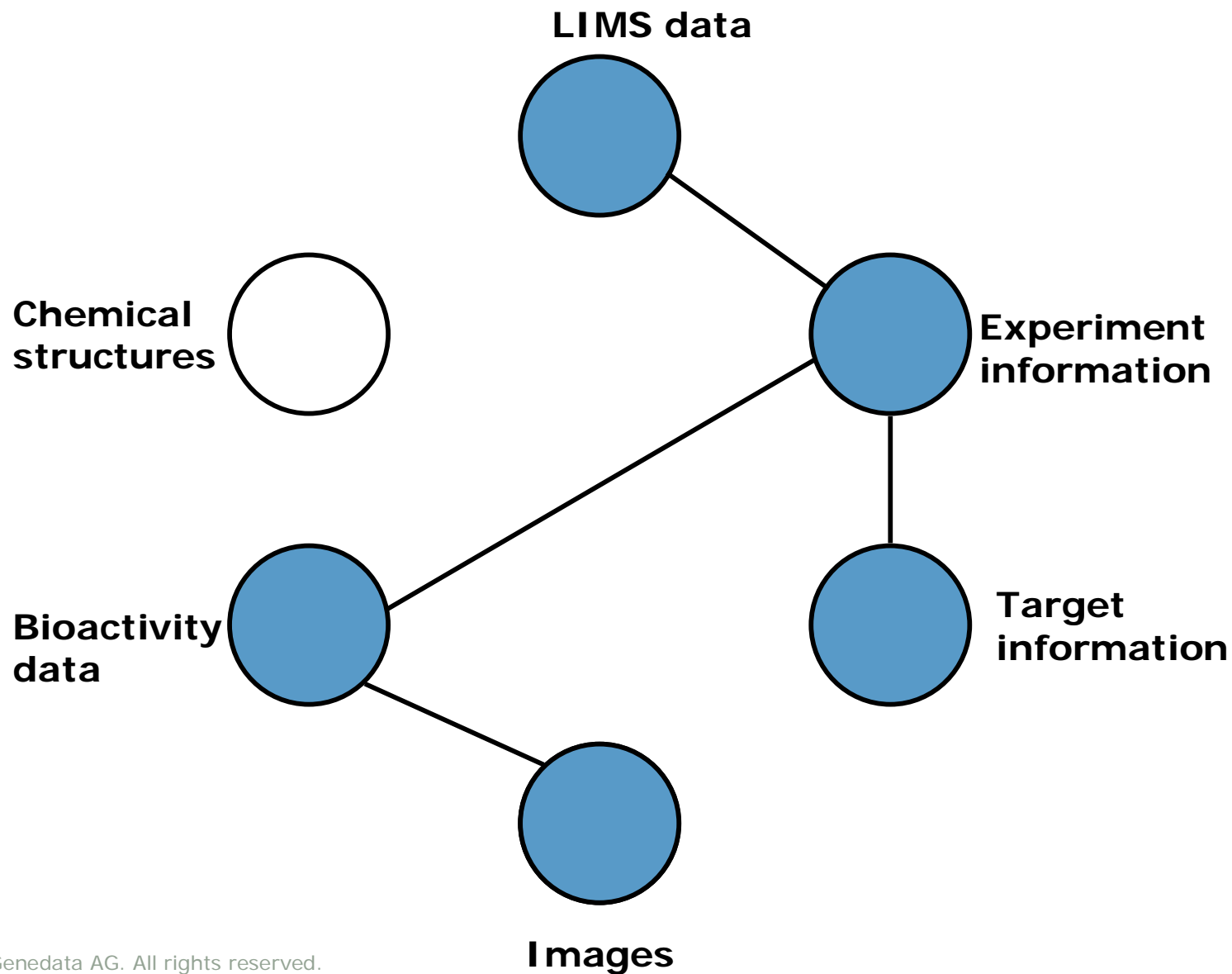
Lead discovery arena – Semantic integration of chemical data



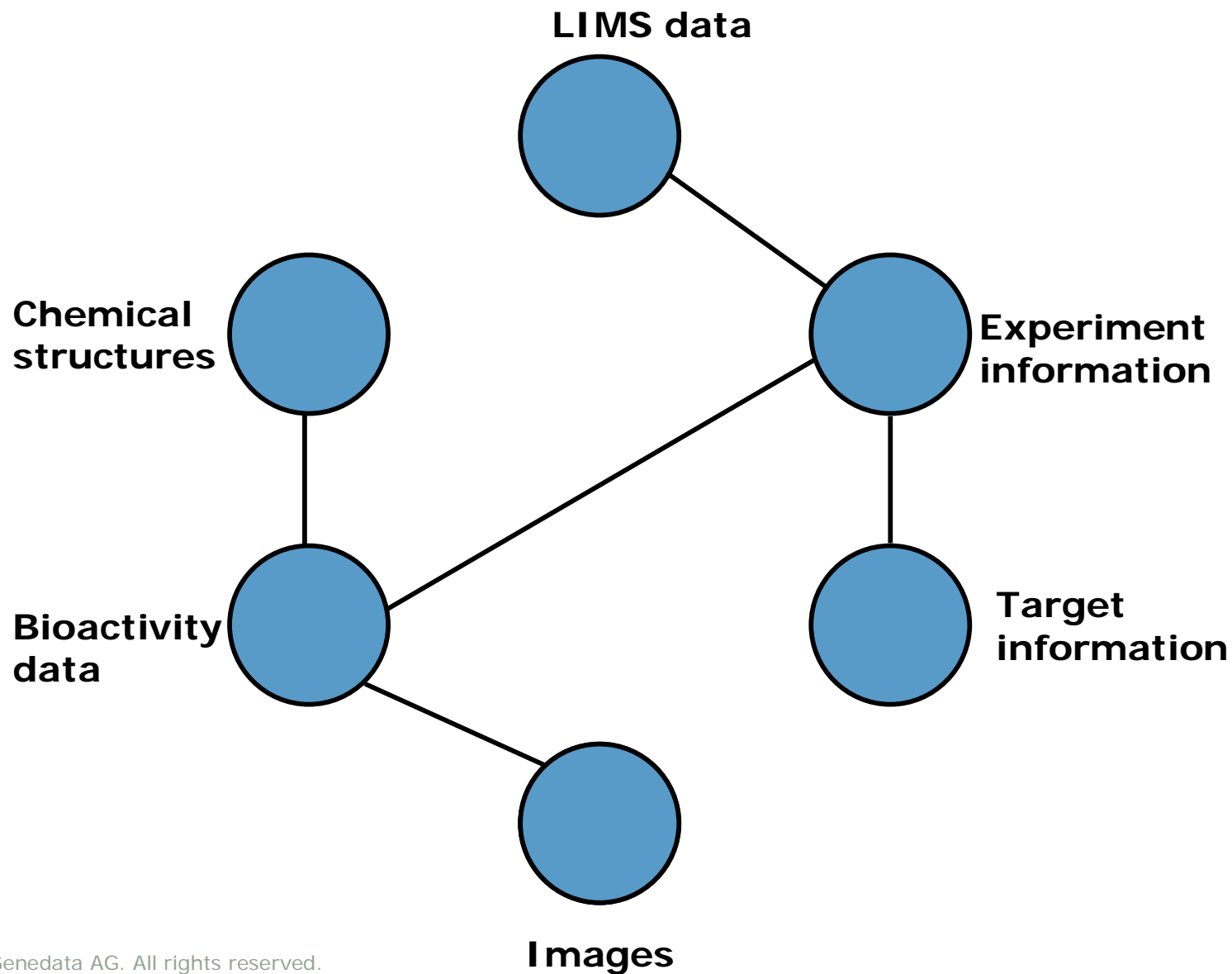
Lead discovery arena – Semantic integration of chemical data



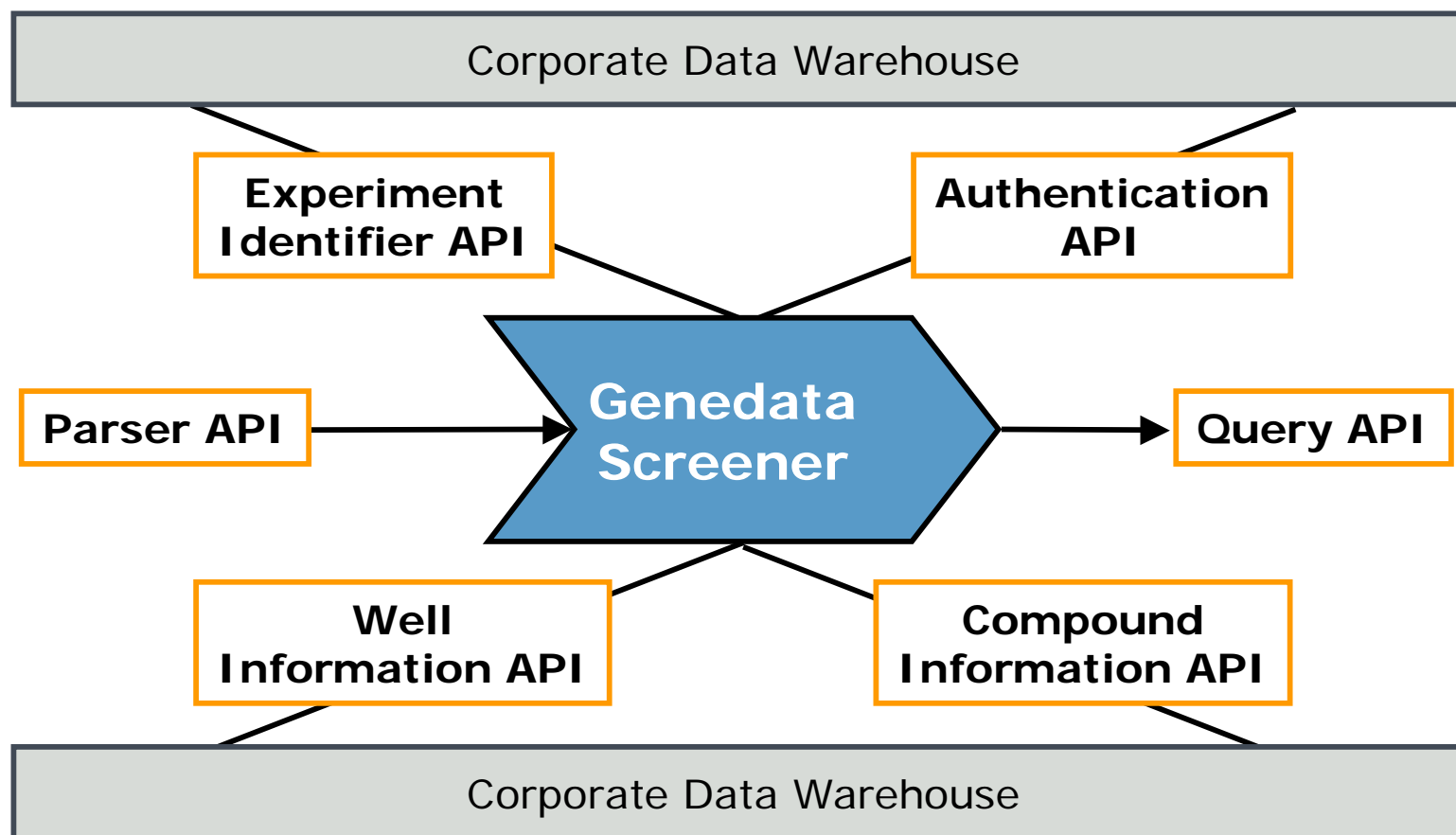
Lead discovery arena – Semantic integration of chemical data



Lead discovery arena – Semantic integration of chemical data

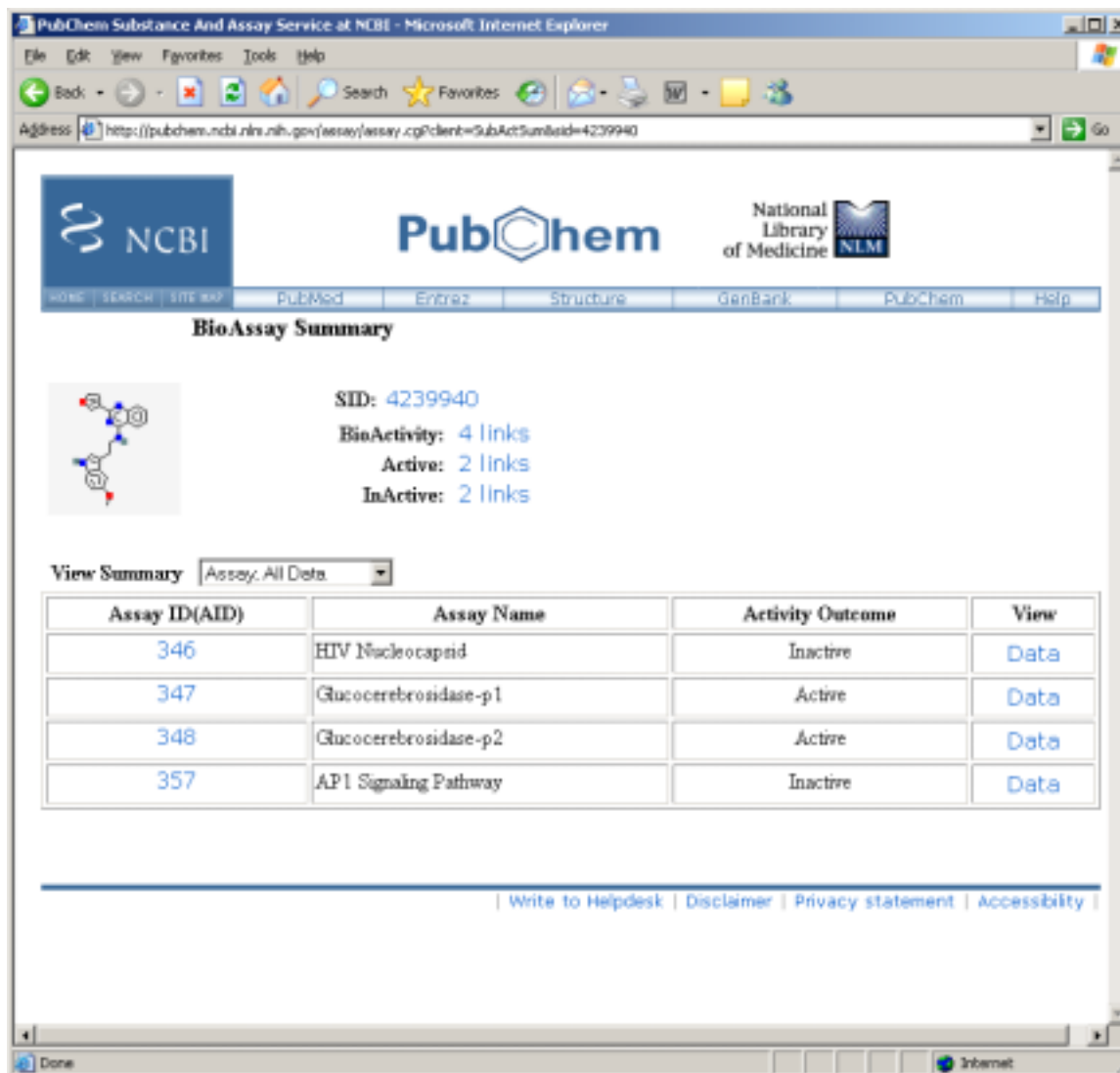


Semantic integration of specific experimental data with Corporate Information



Easy integration with corporate-wide IT infrastructure increases convenience and facilitates decision making process

Lead Discovery Portal - PubChem




PubChem Substance And Assay Service at NCBI - Microsoft Internet Explorer

Address: <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?clnt=SubActSum&sid=4239940>

NCBI PubChem National Library of Medicine NLM

HOME | SEARCH | SITE MAP | PubMed | Entrez | Structure | GenBank | PubChem | Help

BioAssay Summary



SID: 4239940
BioActivity: 4 links
Active: 2 links
InActive: 2 links

View Summary

Assay ID(AID)	Assay Name	Activity Outcome	View
346	HIV Nucleocapsid	Inactive	Data
347	Glucocerebrosidase-p1	Active	Data
348	Glucocerebrosidase-p2	Active	Data
357	AP1 Signaling Pathway	Inactive	Data

[Write to Helpdesk](#) | [Disclaimer](#) | [Privacy statement](#) | [Accessibility](#)

Lead Discovery Portal - PubChem

The image shows two overlapping browser windows from the PubChem website. The left window displays the 'BioAssay Summary' for assay ID 4239940, including a chemical structure and a table of assay details. The right window displays the 'BioAssay Results' for the same assay, showing a table of results for one compound.

BioAssay Summary (Left Window):

SID: 4239940
 BioActivity: 4 links
 Active: 2 links
 InActive: 2 links

View Summary: Assay, All Data

Assay ID(AID)	Assay Name
346	HIV Nucleocapsid
347	Glucocerebrosidase-p1
348	Glucocerebrosidase-p2
357	AP1 Signaling Pathway

BioAssay Results (Right Window):

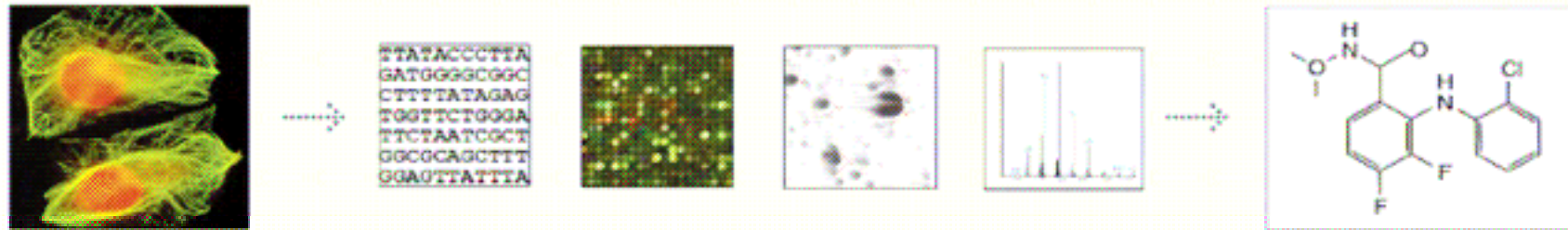
BioAssay ID (AID): 346
 Source: NCGC
 Name: HIV Nucleocapsid

Total 1 compounds found (1 unique), 1 displayed:

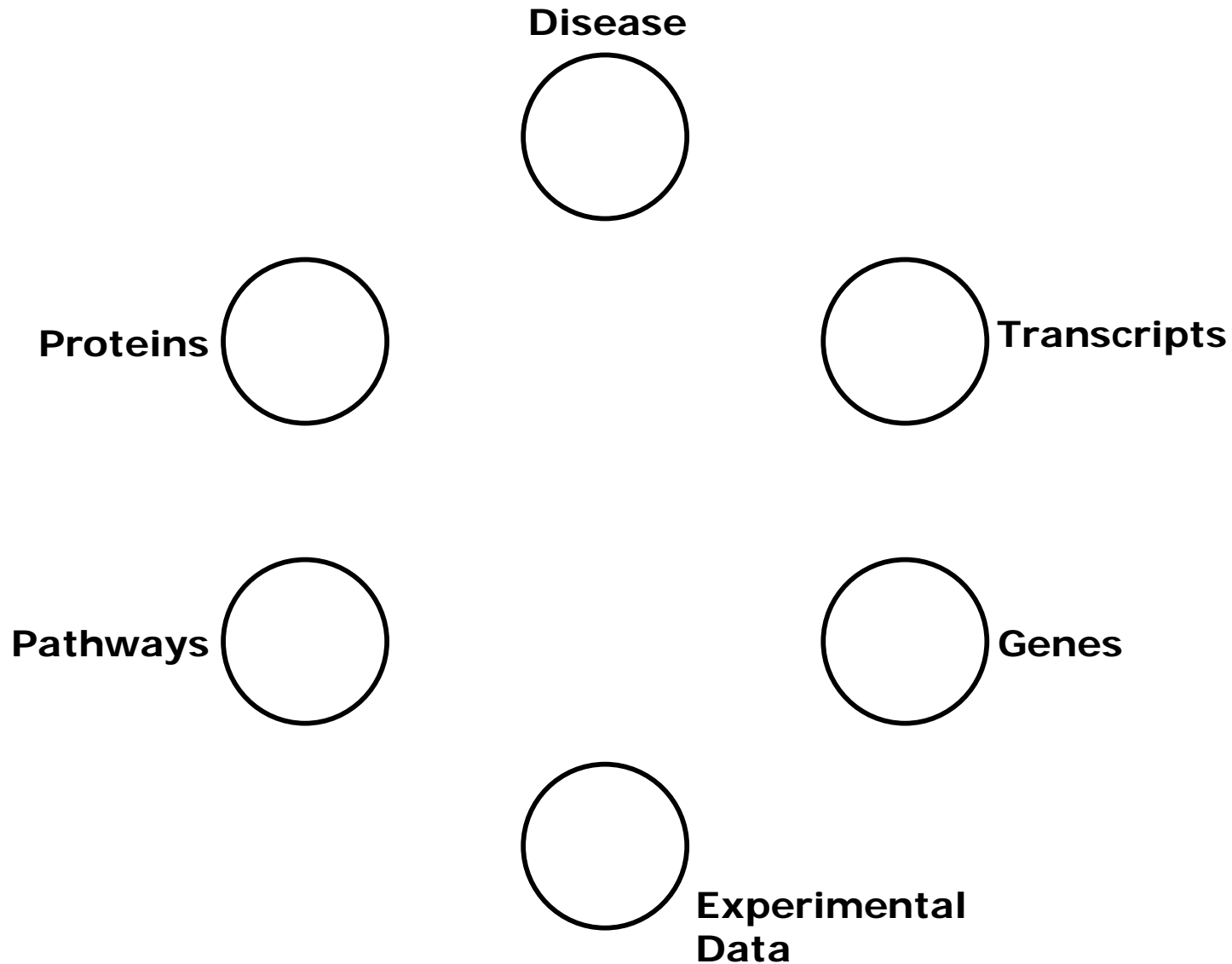
Structure	PubChem					Qualified ACS0	Activity Direction	Activity Qualifier	Log of ACS0	Model Name	Curve Ch2	Number of Points	Activity at 1.28InM (%)
	SID	CID	Outcome	Activity Score	Date								
	4239940	3235047	Inactive	25	2005/08/24	0.002	inactive	>	2.699	Constant	20.56	7	4.974

Integration example III

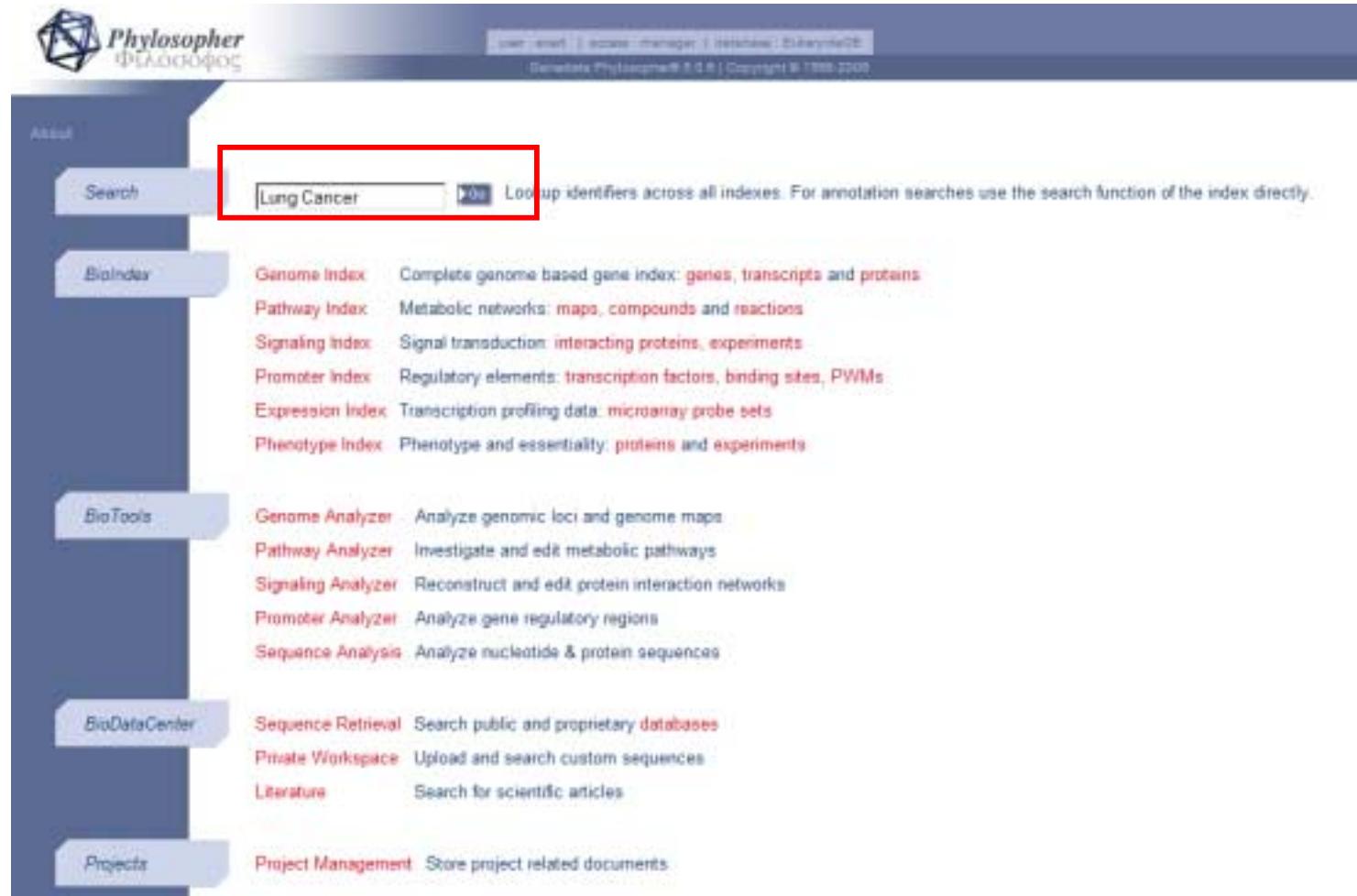
- + **Target Discovery using ...**
 - Public and proprietary data
 - Integrating data from different 'omics technologies



Target discovery arena – Semantic integration of biological data



Start with a disease



Phylosopher ΦΙΛΟΣΟΦΟΣ

user: guest | access: manage | settings: BioIndexDB
Genedata Phylosopher® 2.0.0 | Copyright © 1999-2006

About

Search Lookup identifiers across all indexes. For annotation searches use the search function of the index directly.

BioIndex

- Genome Index**: Complete genome based gene index: [genes](#), [transcripts](#) and [proteins](#)
- Pathway Index**: Metabolic networks: [maps](#), [compounds](#) and [reactions](#)
- Signaling Index**: Signal transduction: [interacting proteins](#), [experiments](#)
- Promoter Index**: Regulatory elements: [transcription factors](#), [binding sites](#), [PWMs](#)
- Expression Index**: Transcription profiling data: [microarray probe sets](#)
- Phenotype Index**: Phenotype and essentiality: [proteins](#) and [experiments](#)

BioTools

- Genome Analyzer**: Analyze genomic loci and genome maps
- Pathway Analyzer**: Investigate and edit metabolic pathways
- Signaling Analyzer**: Reconstruct and edit protein interaction networks
- Promoter Analyzer**: Analyze gene regulatory regions
- Sequence Analysis**: Analyze nucleotide & protein sequences

BioDataCenter

- Sequence Retrieval**: Search public and proprietary [databases](#)
- Private Workspace**: Upload and search custom sequences
- Literature**: Search for scientific articles

Projects

- Project Management**: Store project related documents

Find related proteins

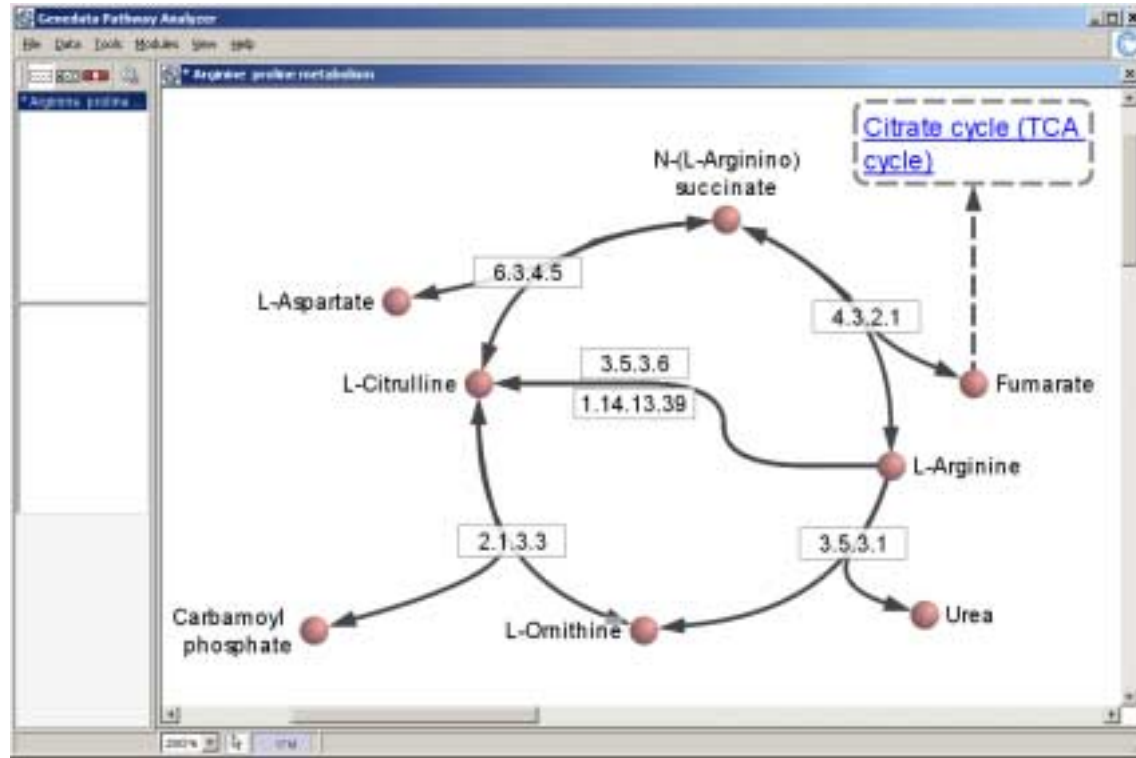
Lung Cancer (proteins 1 of 578)

<< >> Show entry 1 with 1 items on page sorted by Accession >Go >?

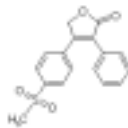
All	Protein ID		PFAM Motifs				Pathway Maps			Probes	
	Name	Description	Location	Score	P-Value	Map ID	Description	Analyzer	Chip	Probe ID	
<input type="checkbox"/>	cust OAT OAT OAT	PF00202	Aminotransferase class III	49 ... 438	623.2	1.8E-184	MAP00220	Urea cycle and metabolism of amino groups	View Map	HG-U133A	201599_at
						MAP00330	Arginine and proline metabolism	View Map	HG-U95Av2	36536_at	
						marie x	new Arginine and proline metabolism	View Map			

New Search Report Modify Group Selection Group Probes Group Pathways Functional Analysis Protein Analyzer >Go

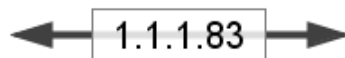
Investigate pathways



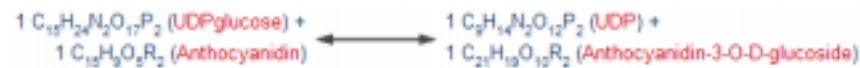
=



(Metabolite or compound)

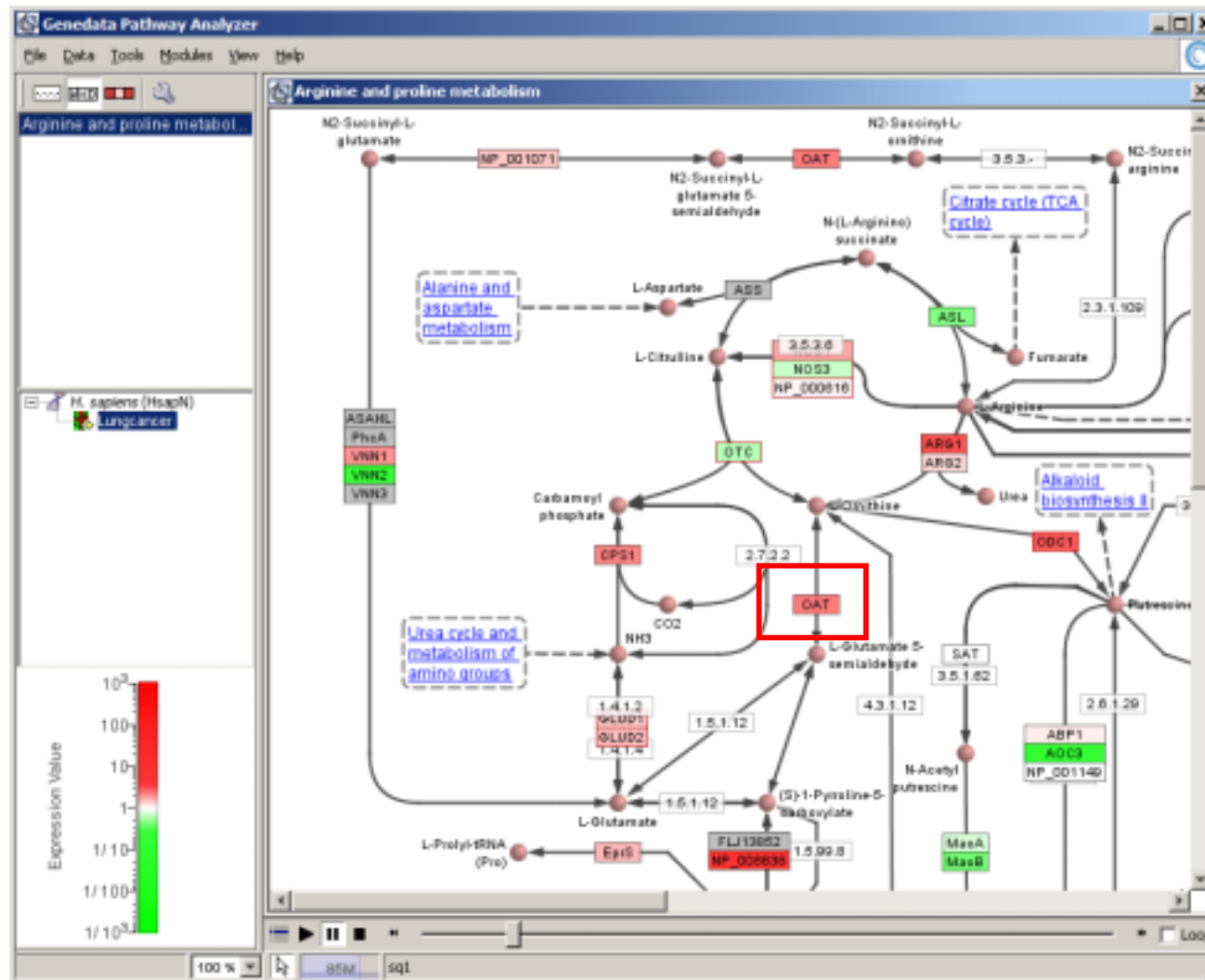


=

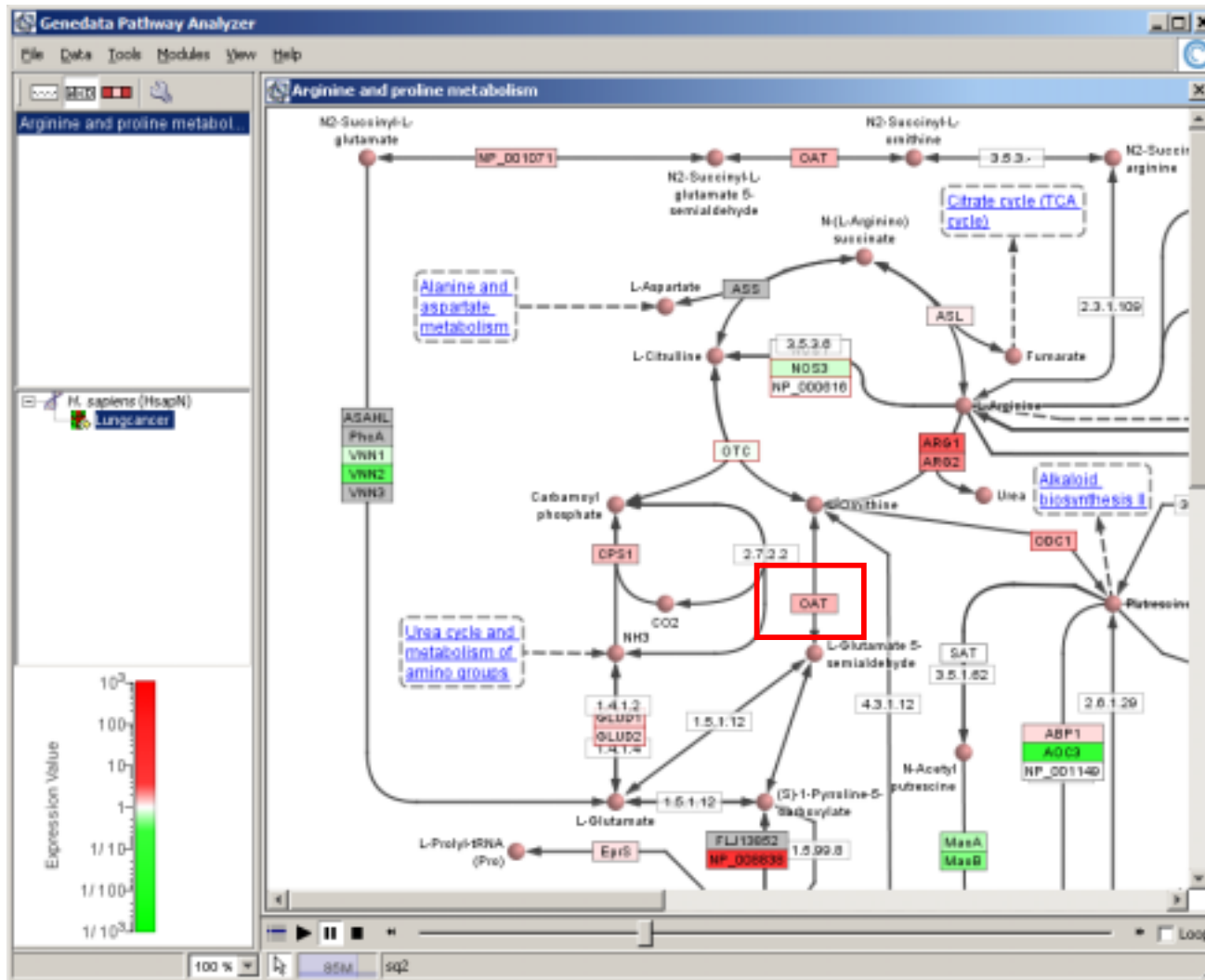


(Biochemical reaction)

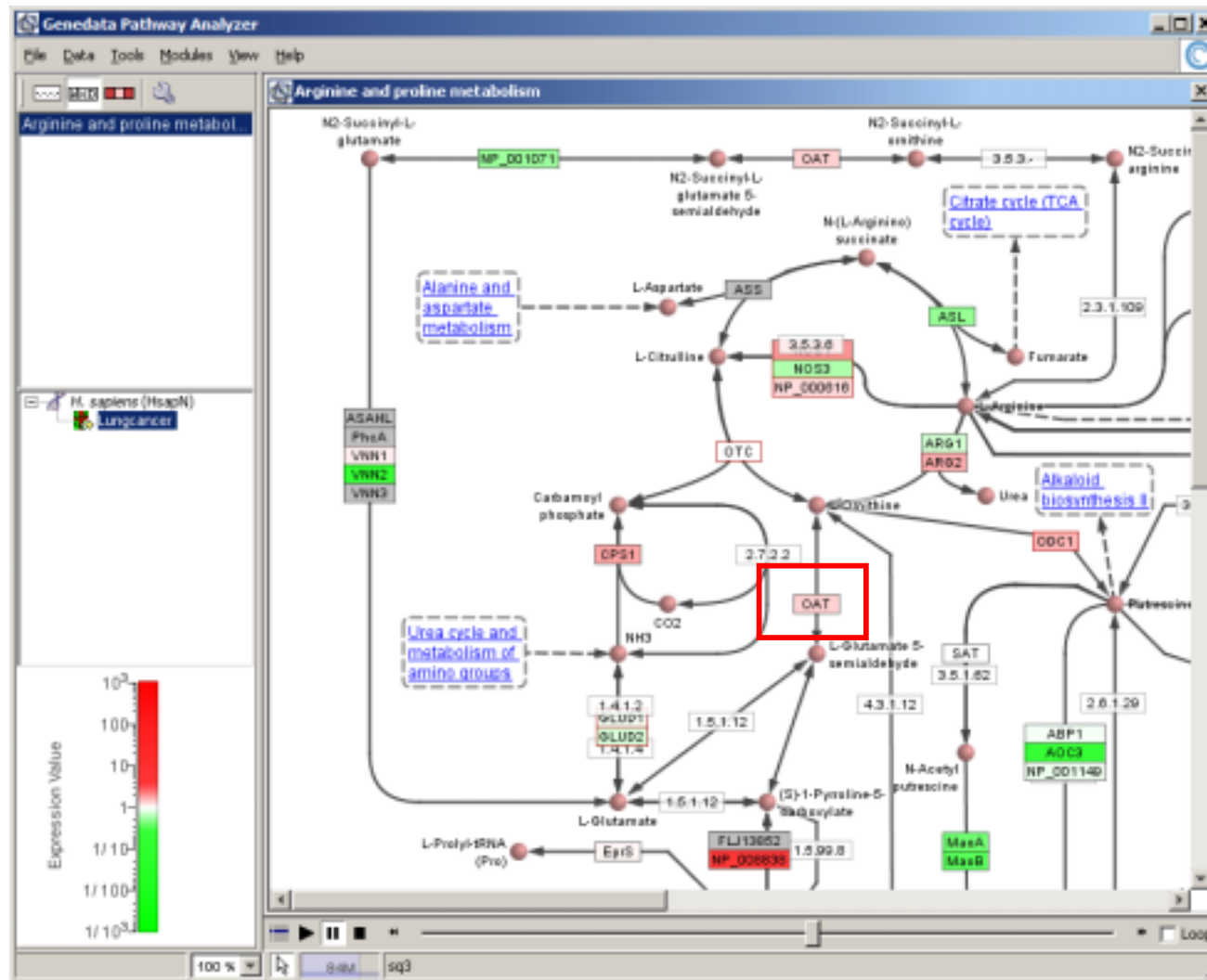
Integrate experimental data



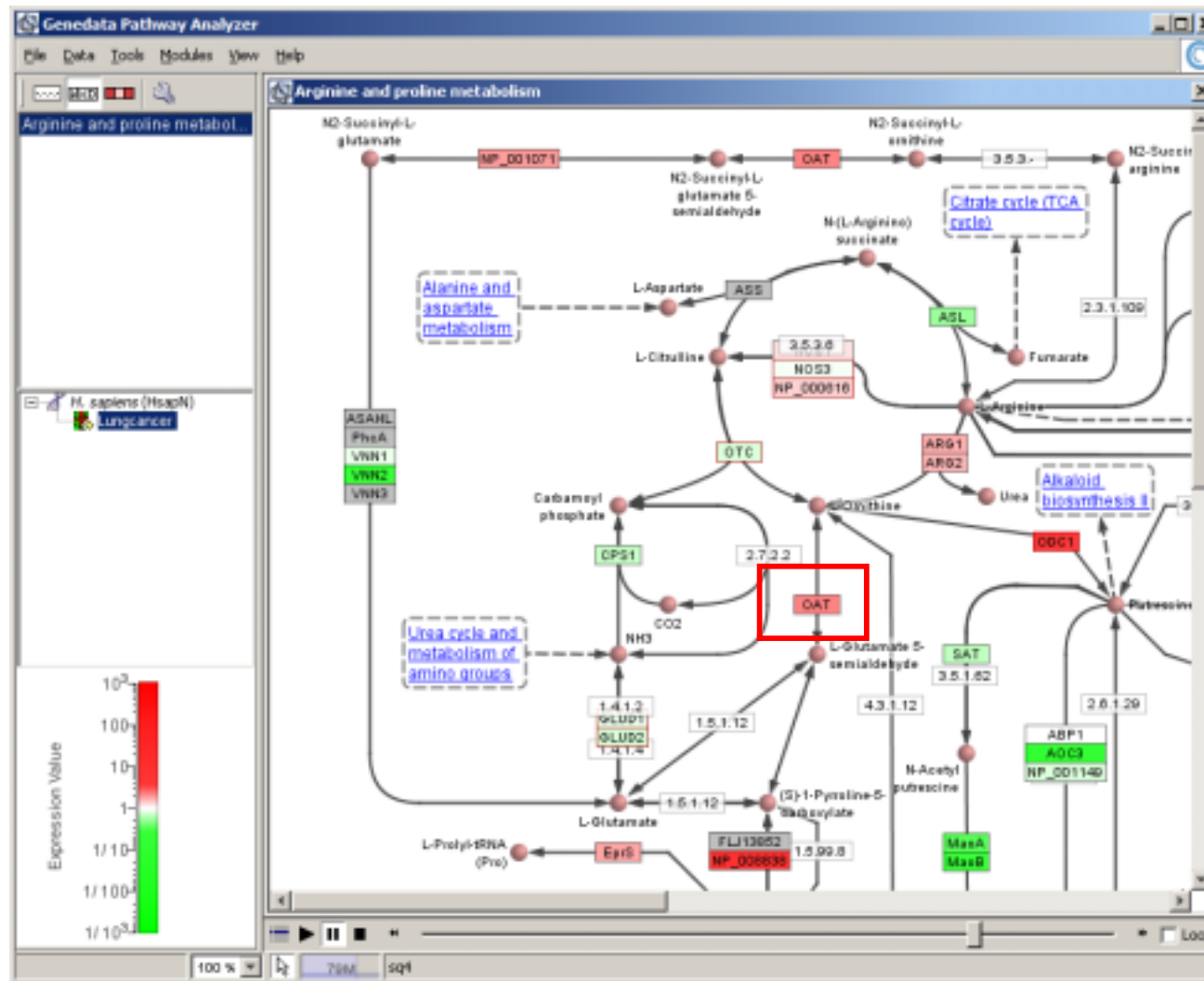
Integrate experimental data



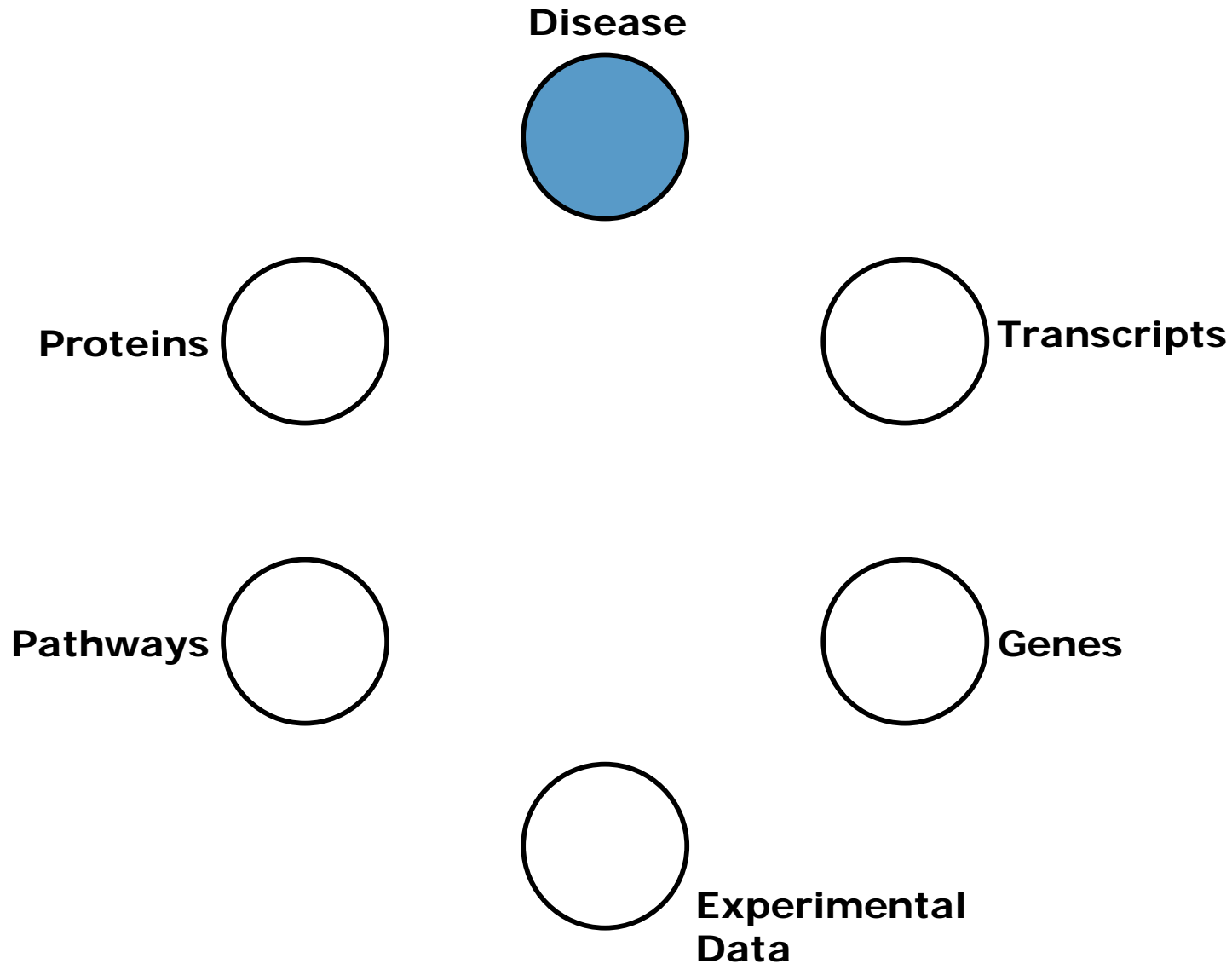
Integrate experimental data



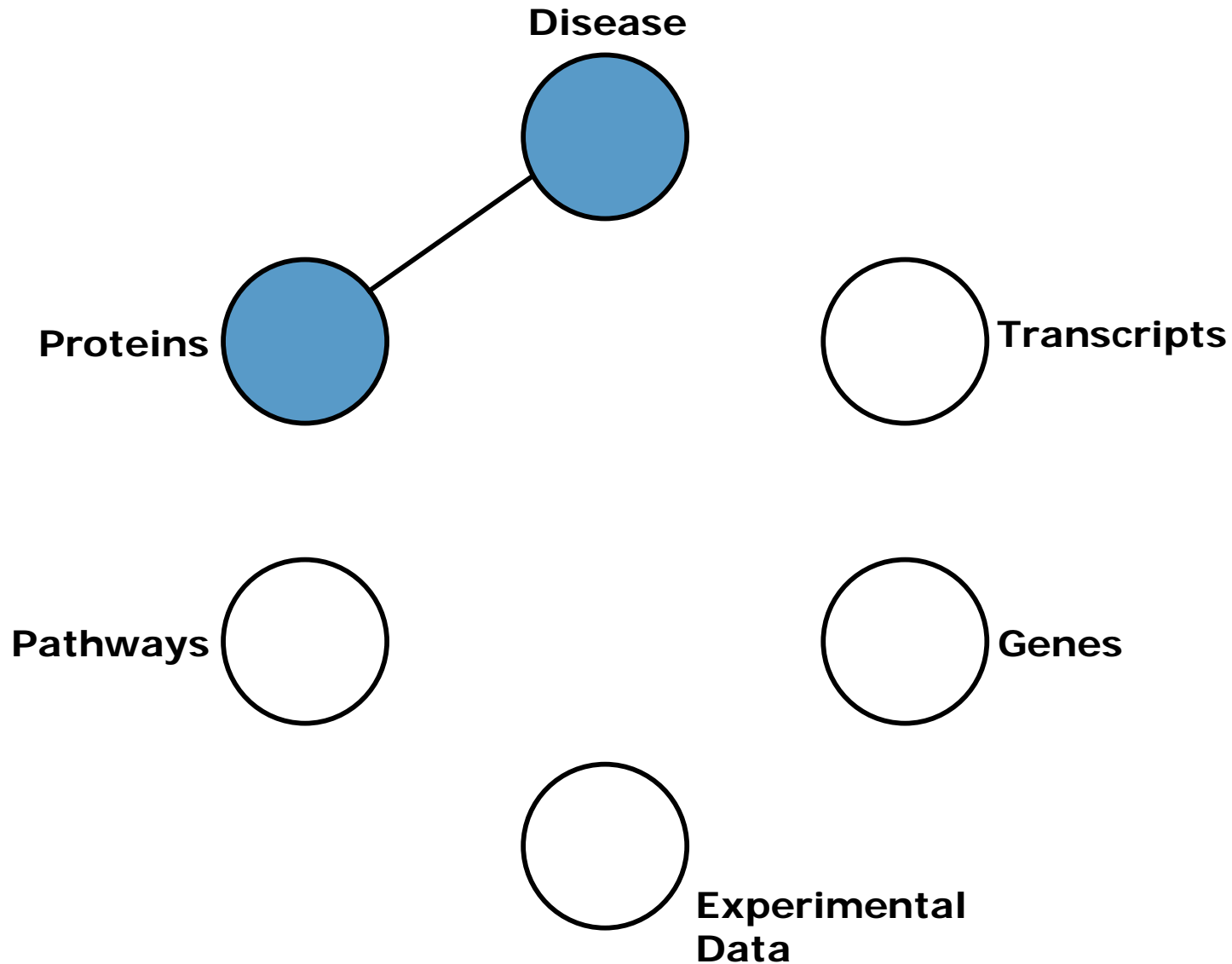
Integrate experimental data



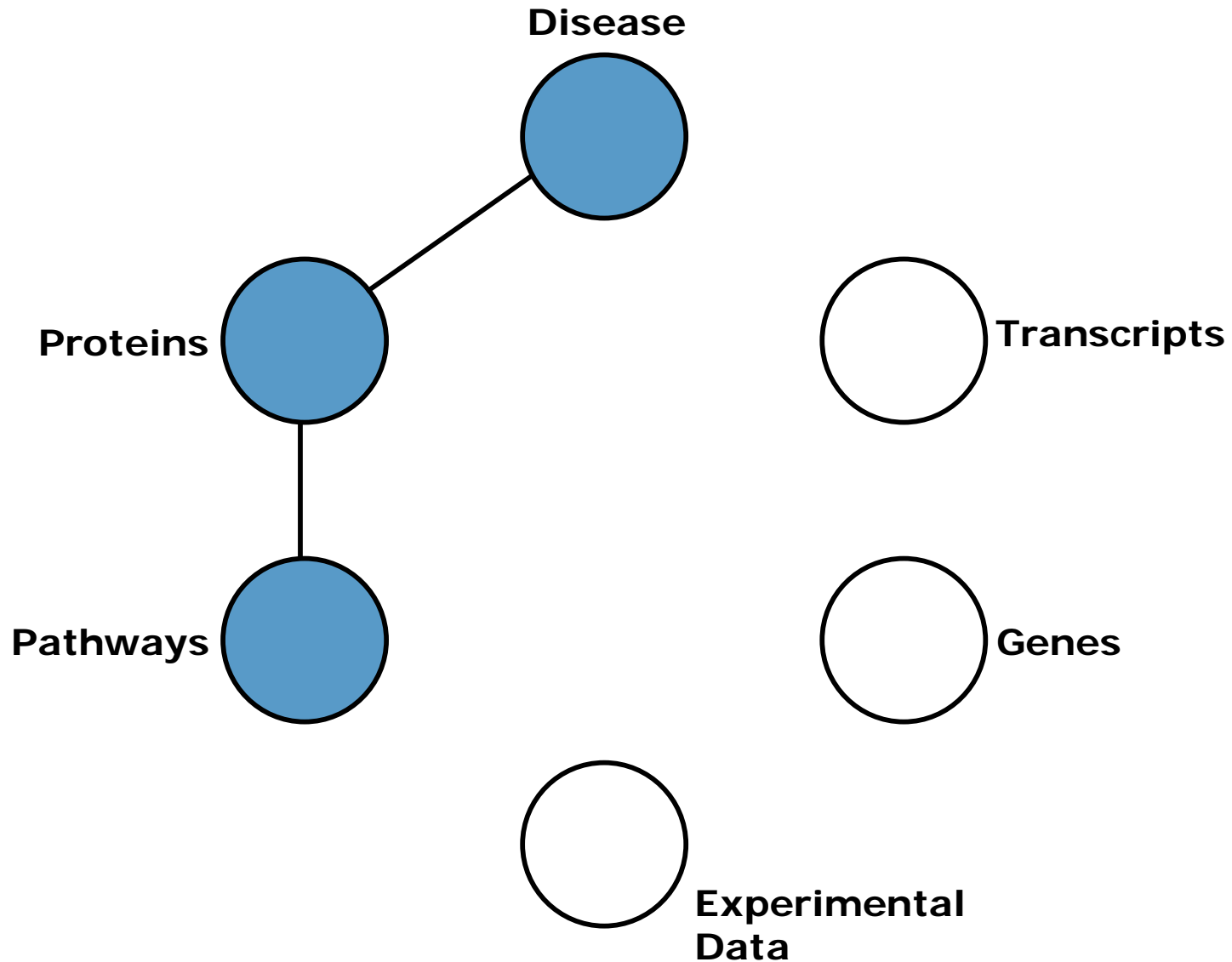
Target discovery arena – Semantic integration of biological data



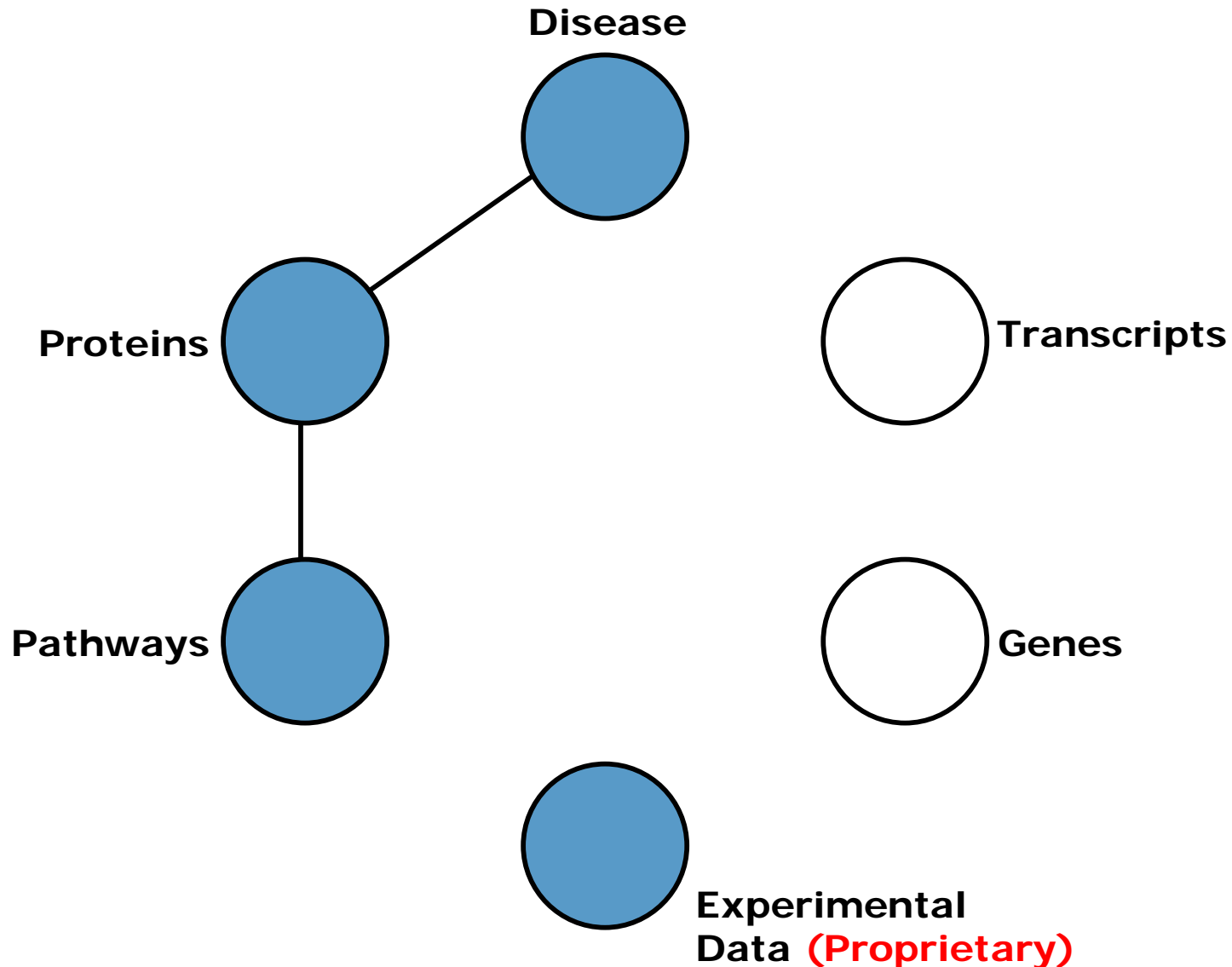
Target discovery arena – Semantic integration of biological data



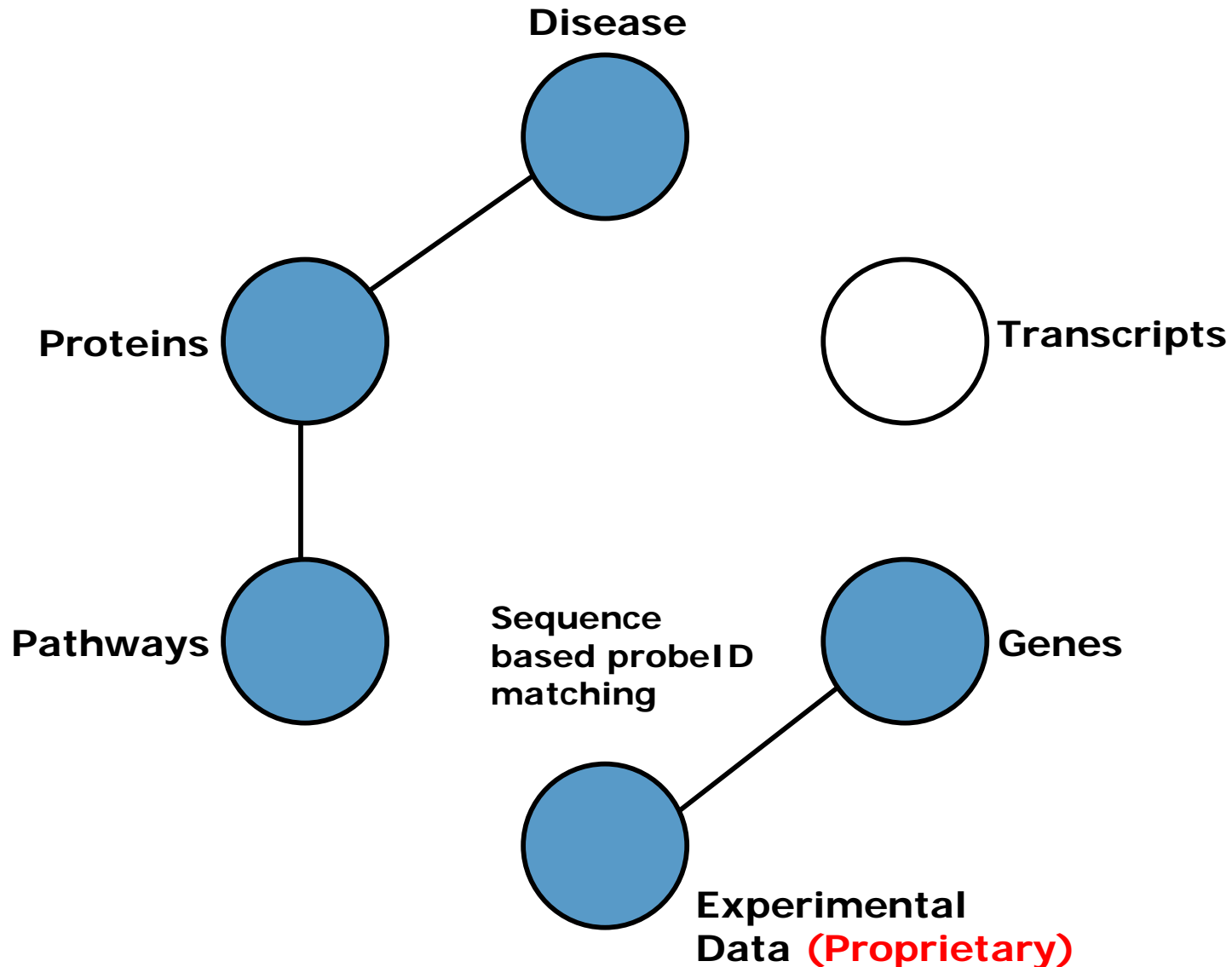
Target discovery arena – Semantic integration of biological data



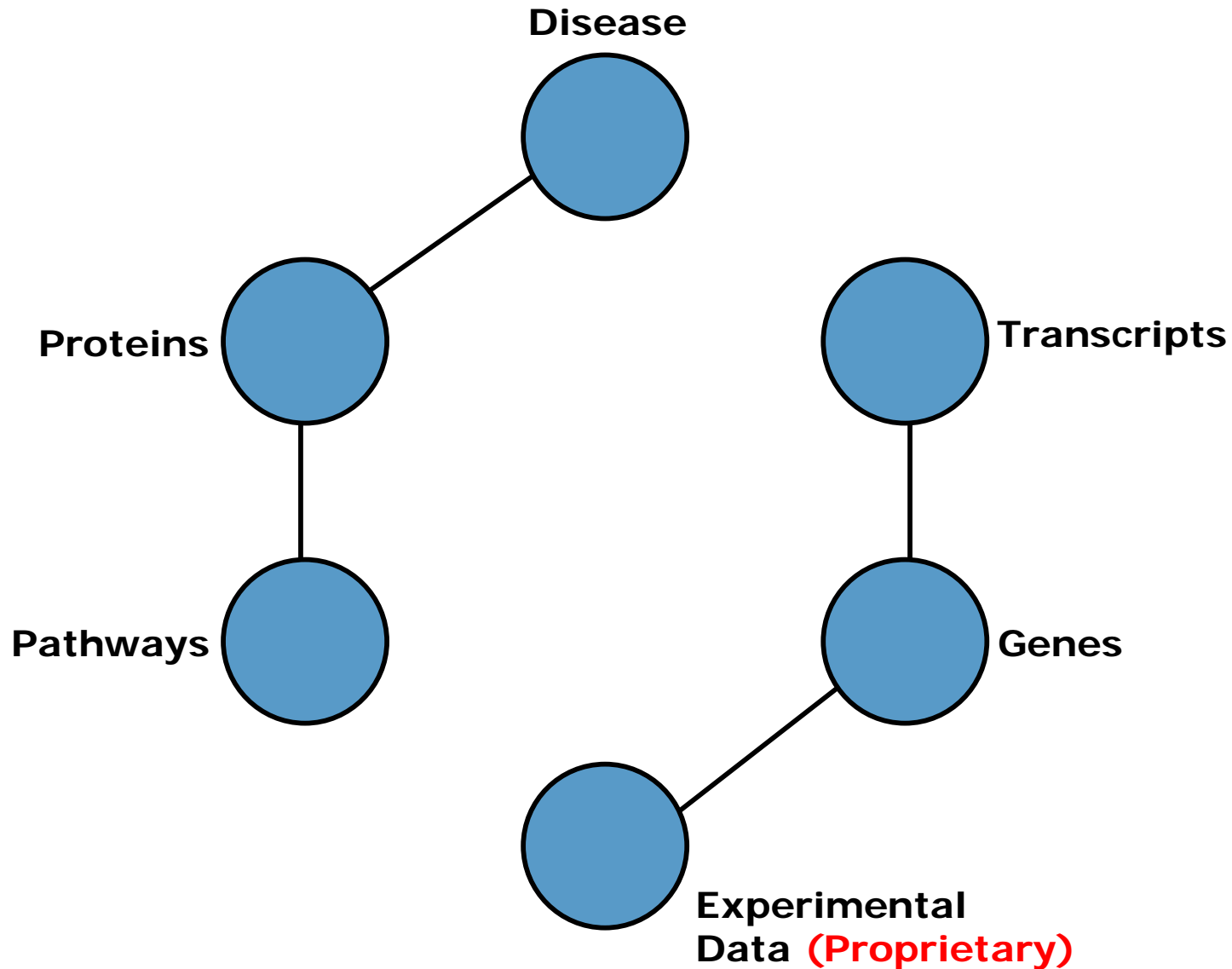
Target discovery arena – Semantic integration of biological data



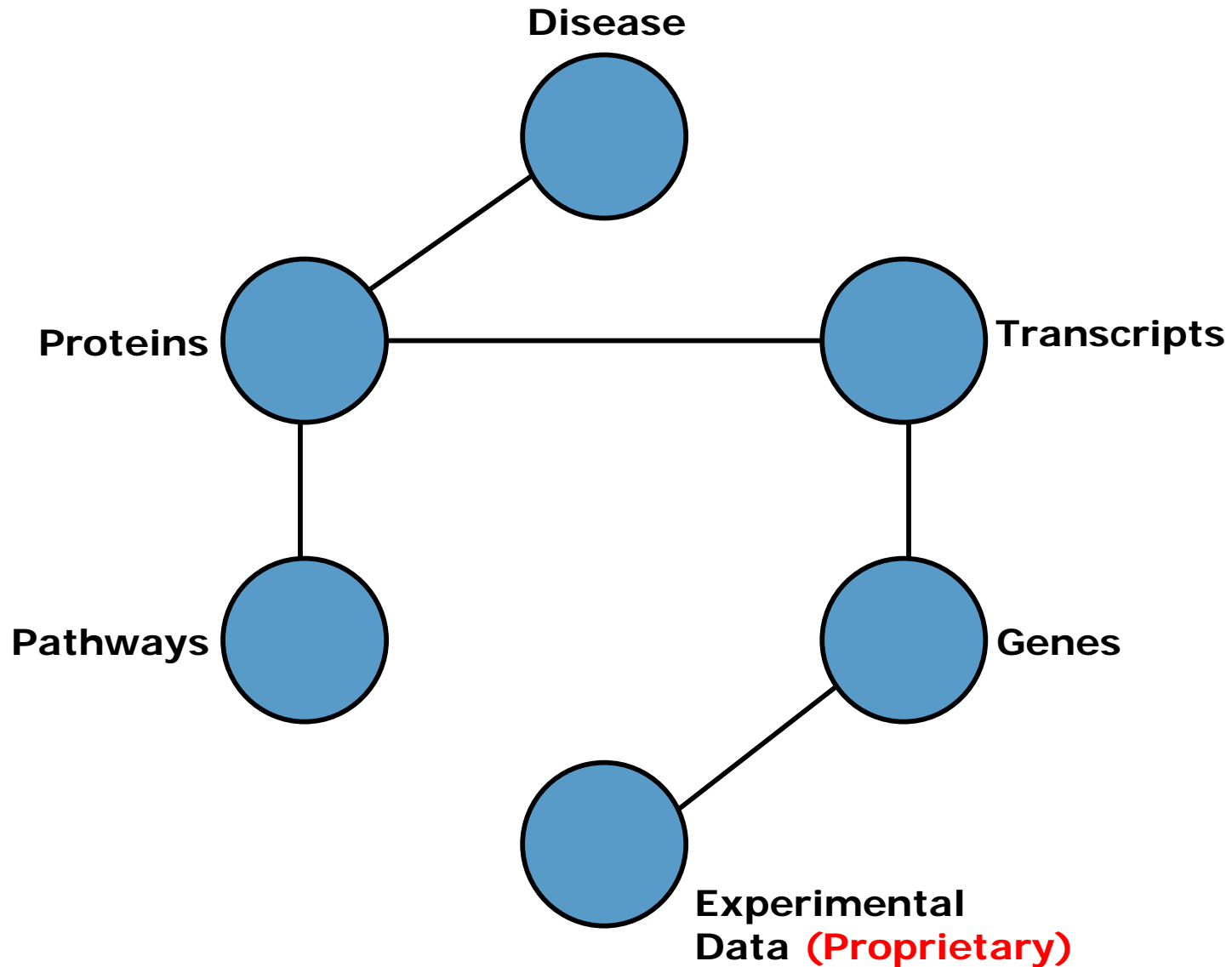
Target discovery arena – Semantic integration of biological data



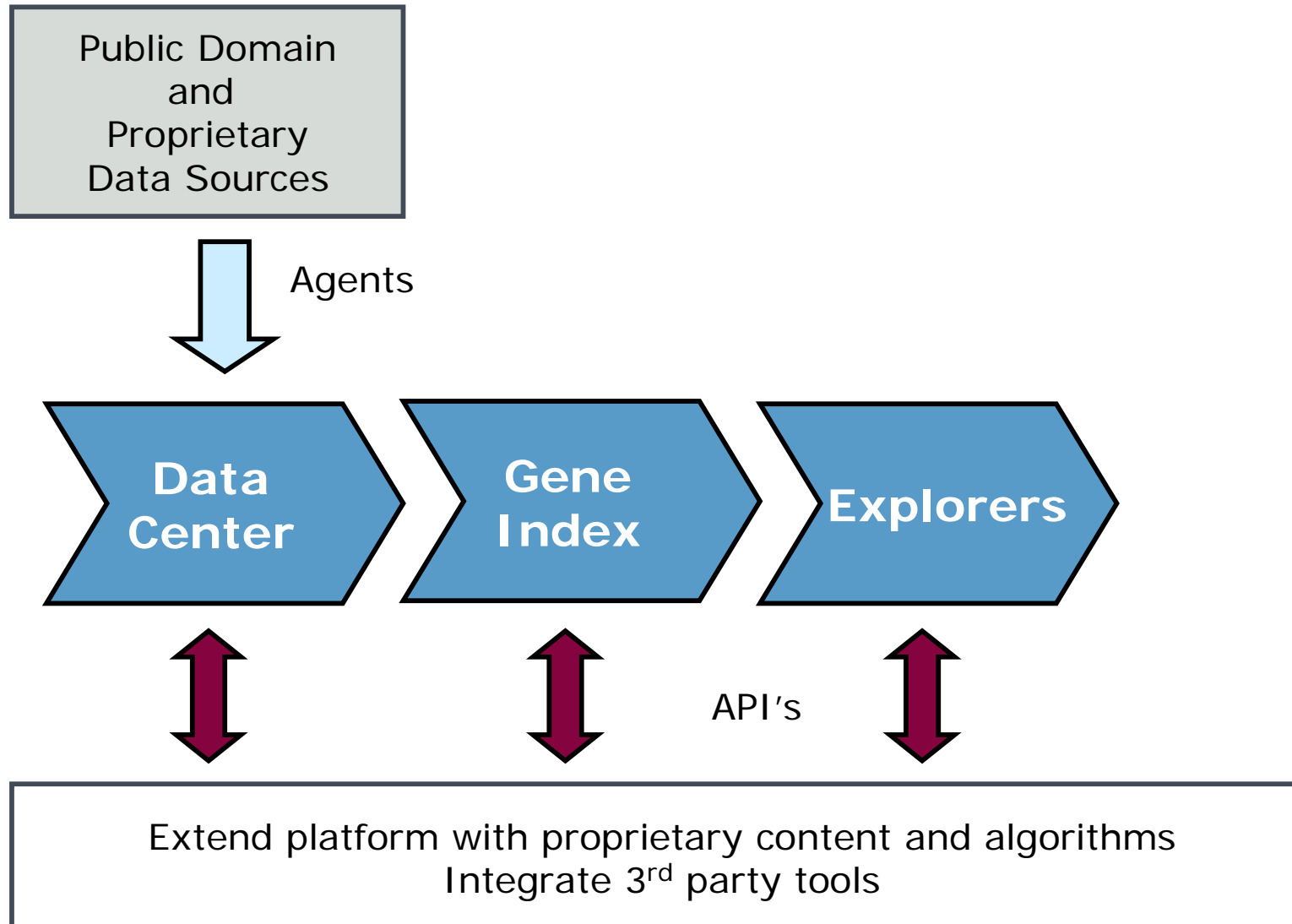
Target discovery arena – Semantic integration of biological data



Target discovery arena – Semantic integration of biological data



Agents and APIs



Conclusions (I)

- + **Examples show the advantage of integration**
 - See the entire picture
 - Drill down to answer specific questions
 - Bring data from different sources together in a single portal
 - Allow researches to work on a shared, consistent and current data base

- + **Semantic Integration requires**
 - Key Software technologies with open, public APIs
 - Very good understanding of (scientific) data
 - Very good understanding of processes and people involved
 - Standard database technologies to supply information
 - Agents to automatically update content

Conclusions (II)

- + **Integration of data remains the key (and the challenge) to deliver value to researchers and their processes**
- + **Semantic integration brings together data contextually, which would otherwise remain disconnected and therefore unused or under utilized**
- + **Discovery IT environments are built up piece by piece according to current and future user needs and need to be open and flexible**



Thank you

Genedata AG
Maulbeerstrasse 46
4016 Basel, Switzerland
Tel. +41 (0) 61 697 6099
Fax +41 (0) 61 697 7244

www.genedata.com