Semantic data integration generating value in drug discovery

ICIC 2006 Nimes
October 25, 2006

Dr. Daniel Keesman

Overview

- **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.

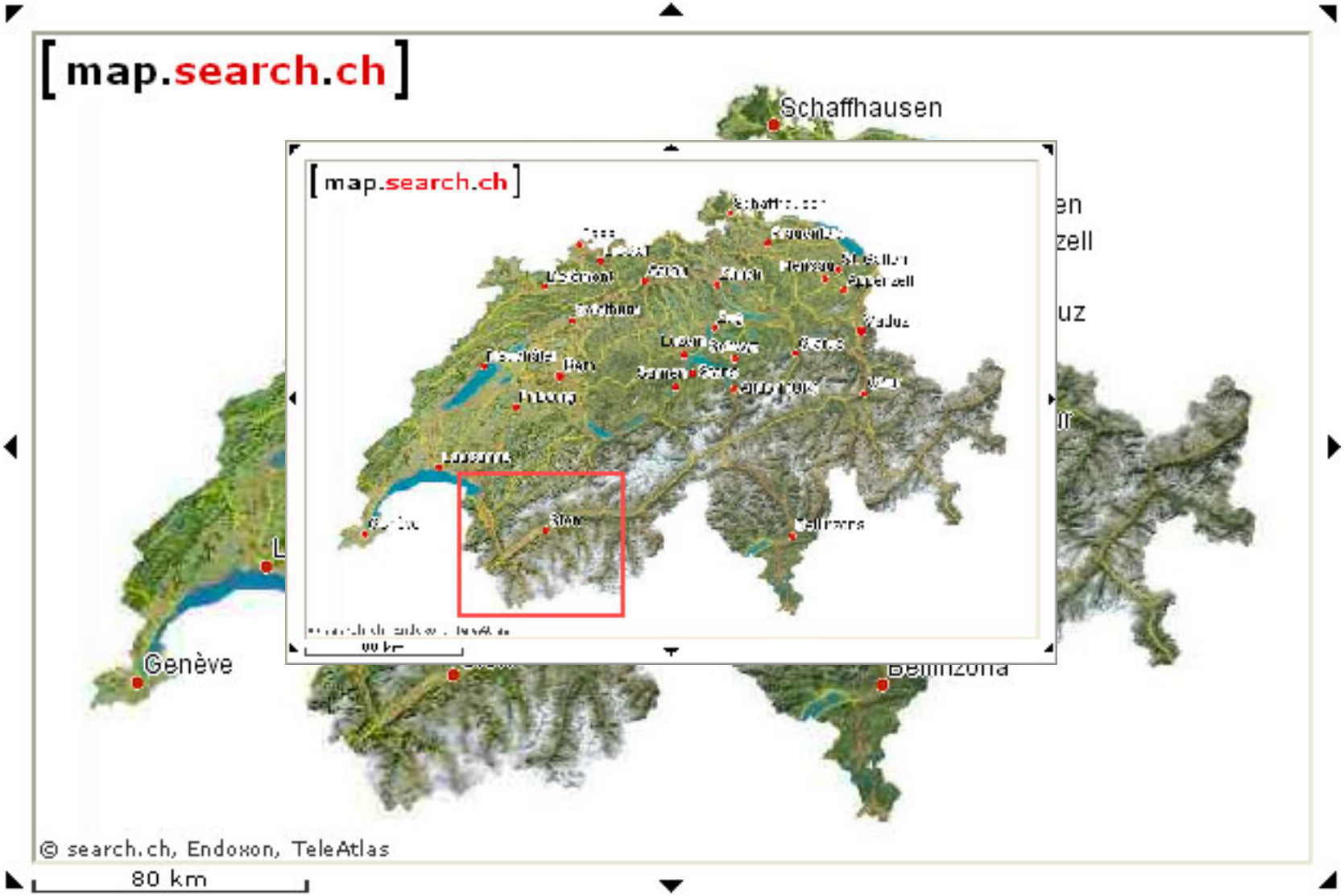
(Source: TopQuadrant)

Integration example I

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

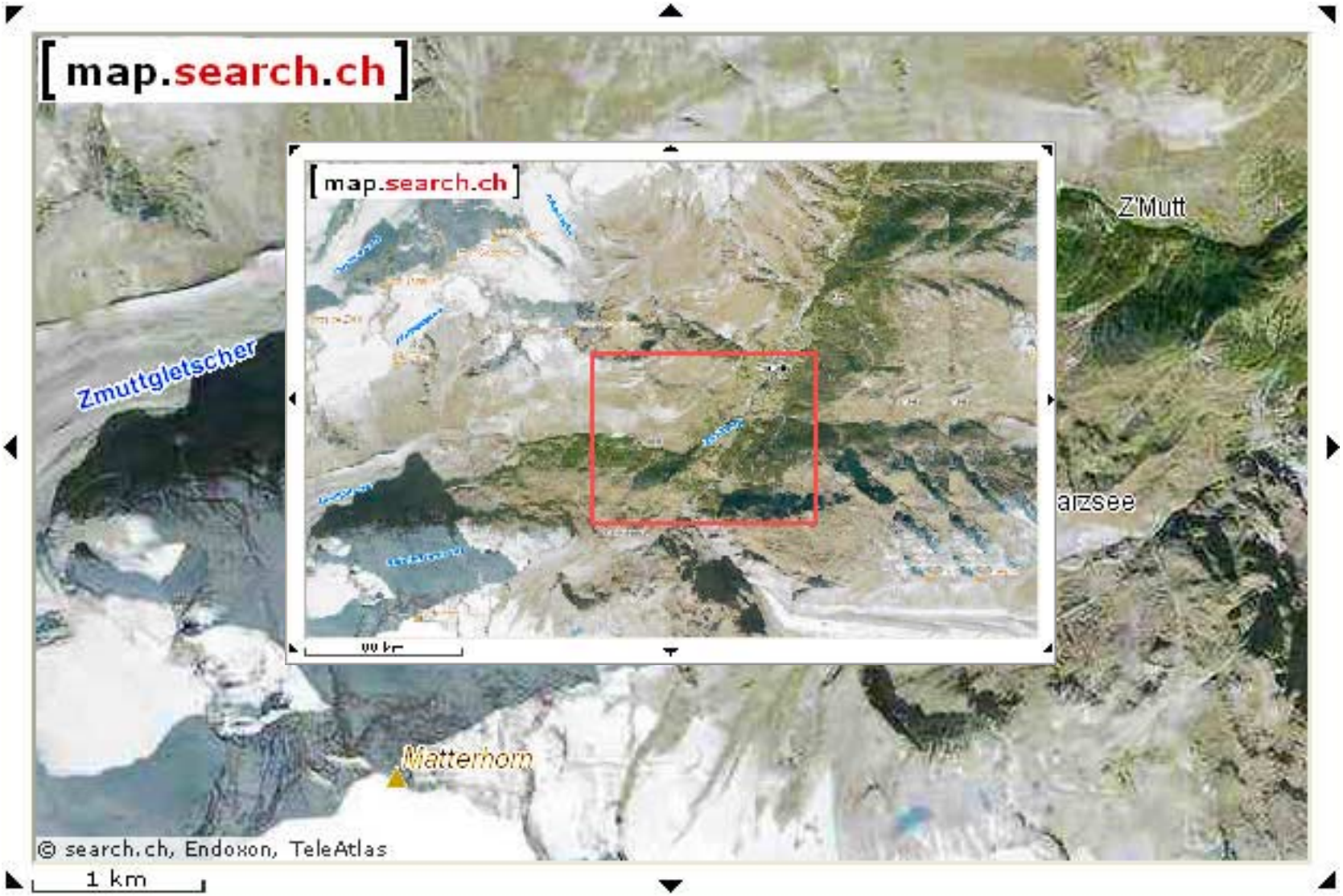
Integration example I

Integration example I



Integration example I

Integration example I



Integration example I

Integration example I

Integration example I

Vacation planning arena

Topography

Cable cars

Hotels

Train stations

Vacation planning arena

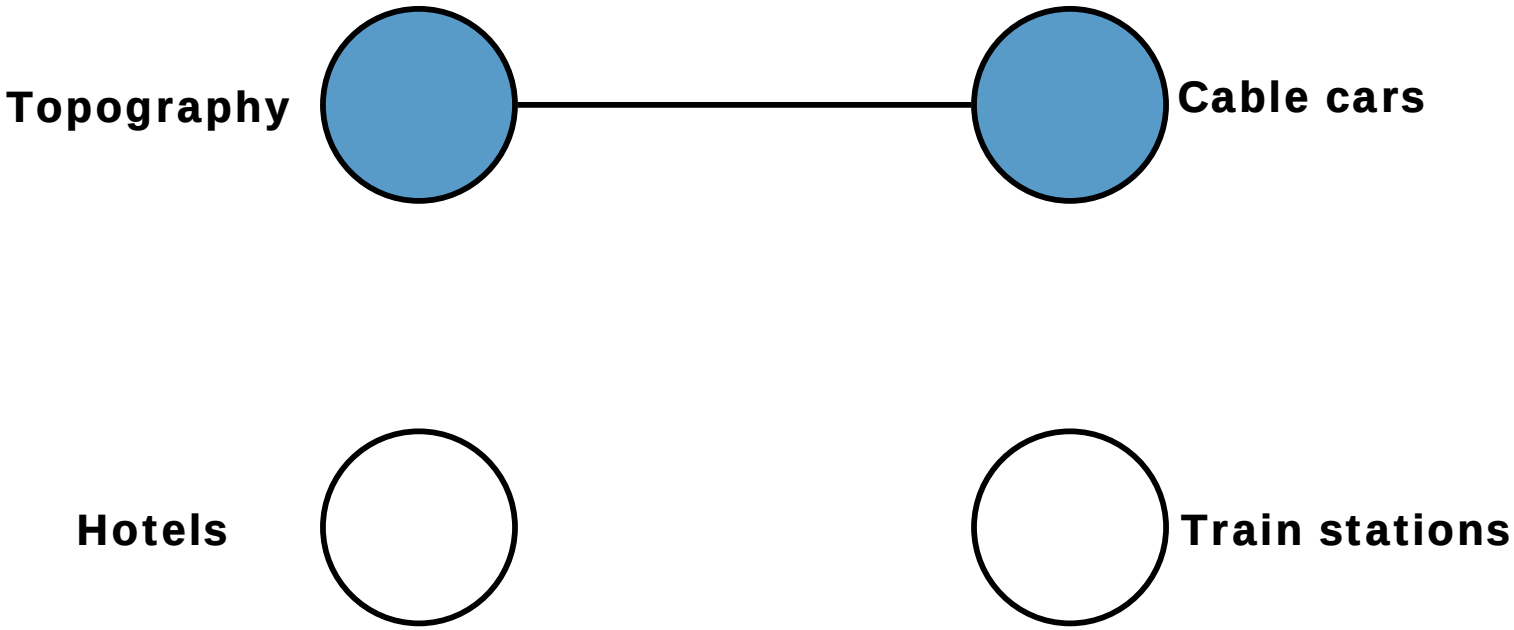
Topography

Cable cars

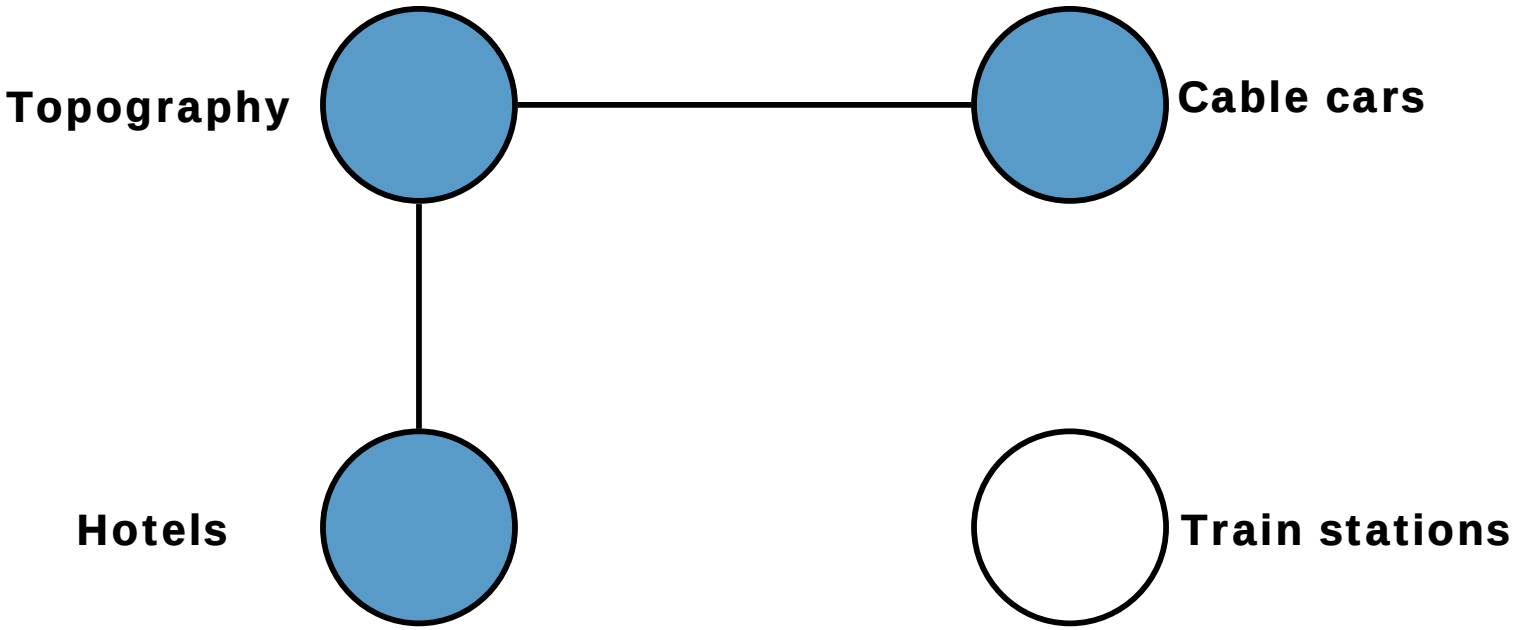
Hotels

Train stations

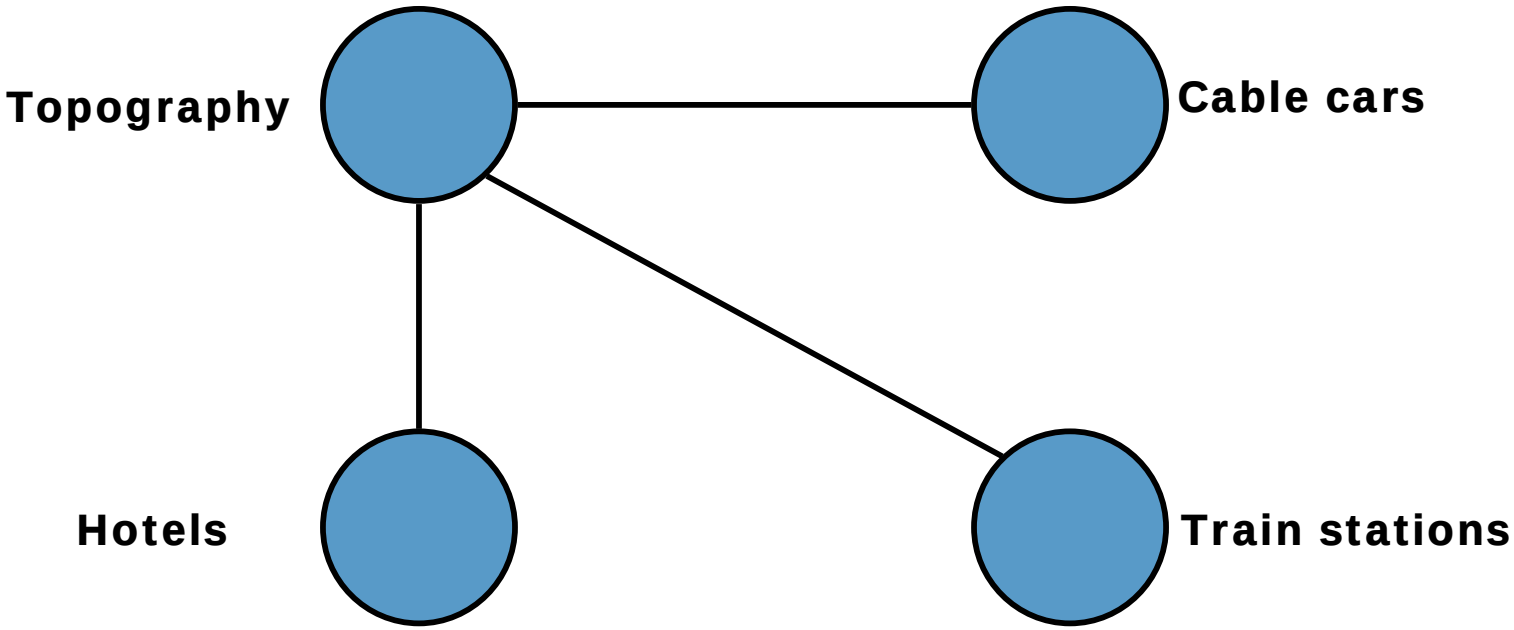
Vacation planning arena



Vacation planning arena

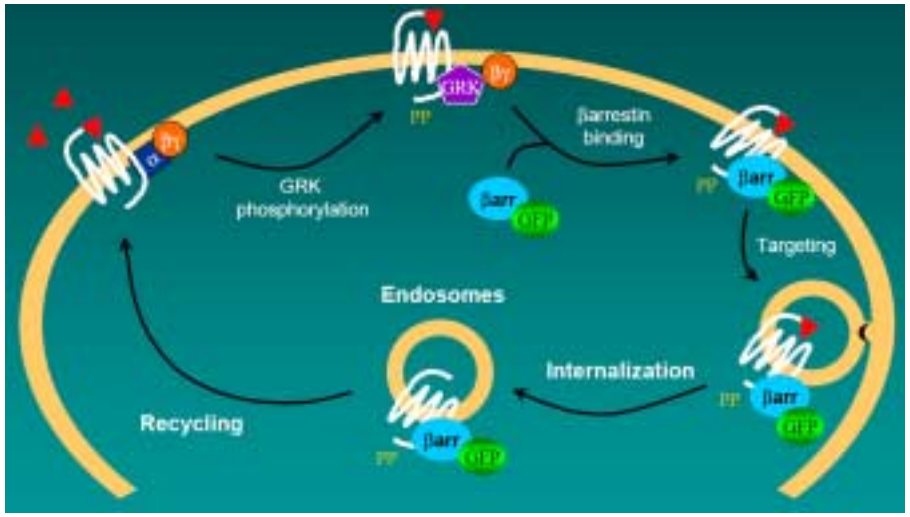


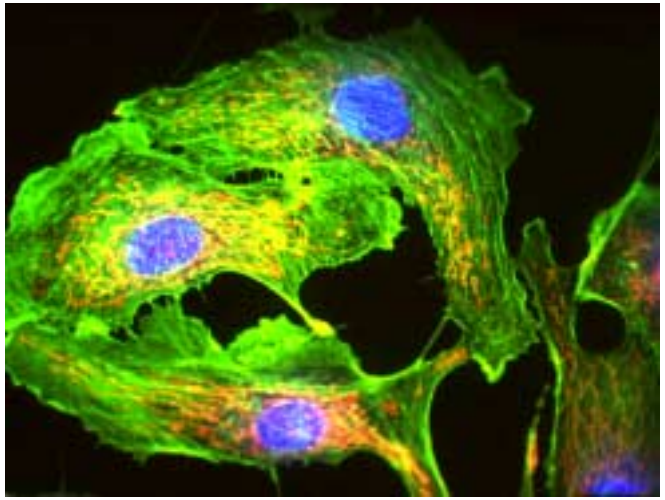
Vacation planning arena



Integration example II

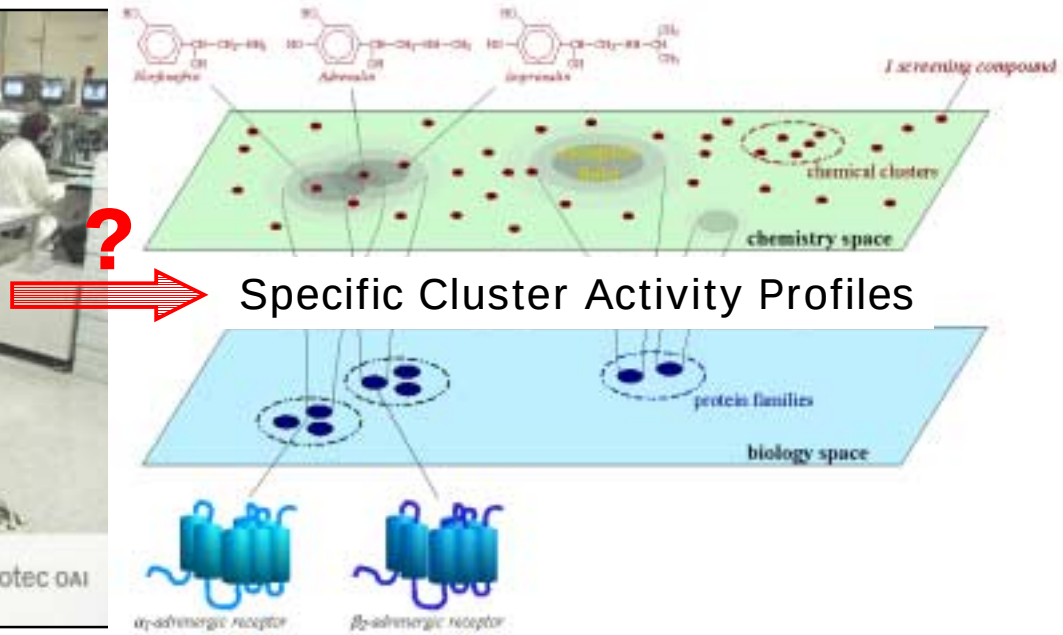
- **Lead discovery**
- **High throughput, high content screening**
  - In-depth biological assessment of entire libraries





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**



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

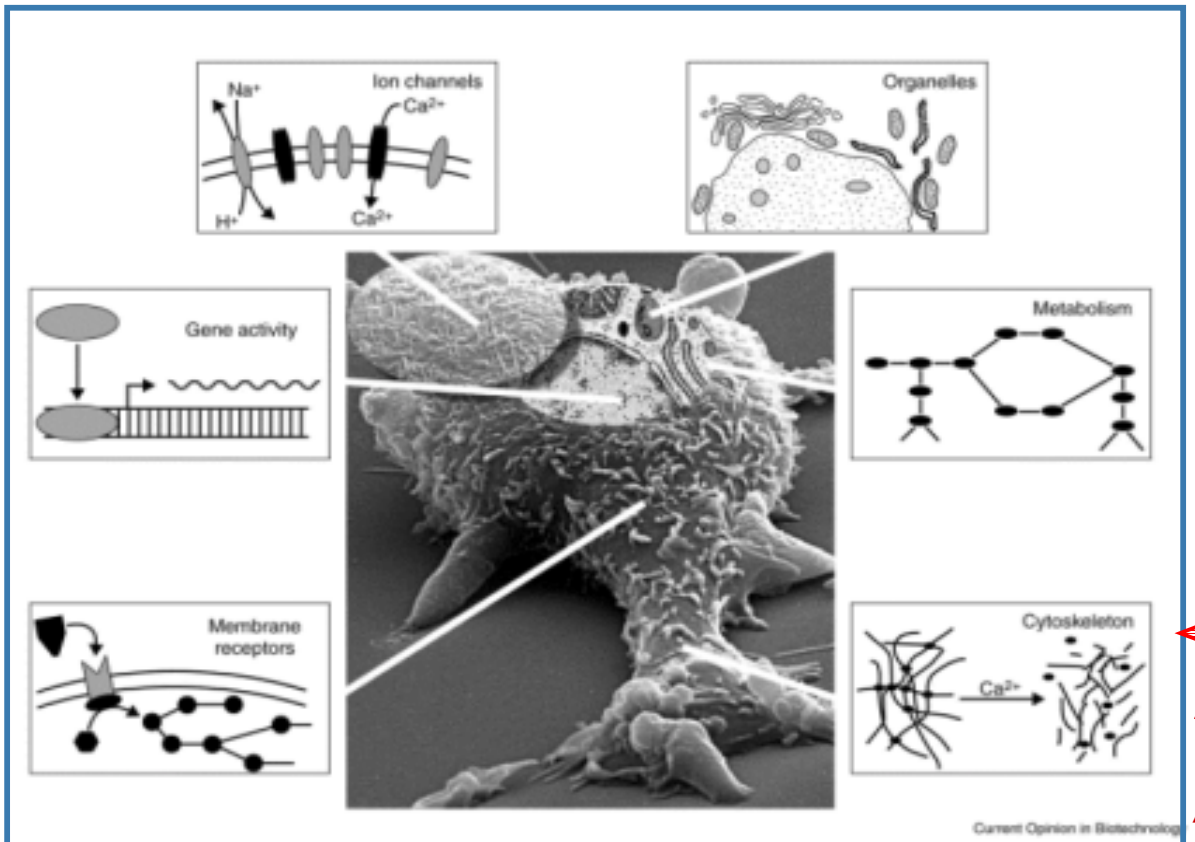
Analysis:

Which compounds were really specifically active and selective?

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



Taylor L, et.al., Curr. Opin. Biotechnol. 12:75-80 (2001)

Parameters

5.7
13.3
-45.3
121

?

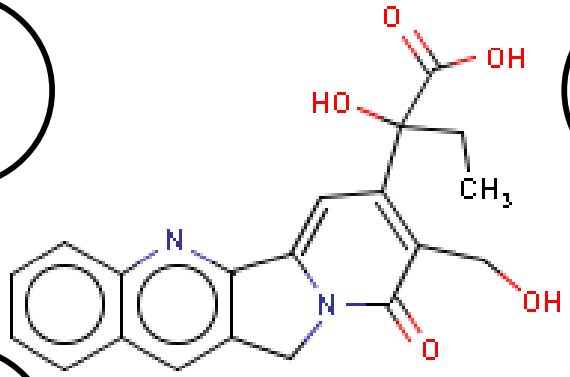
Analysis:

What has really happened to / in the cells when applying compound X?

Lead discovery arena – Semantic integration of chemical data

LIMS data

Chemical properties



Experiment information

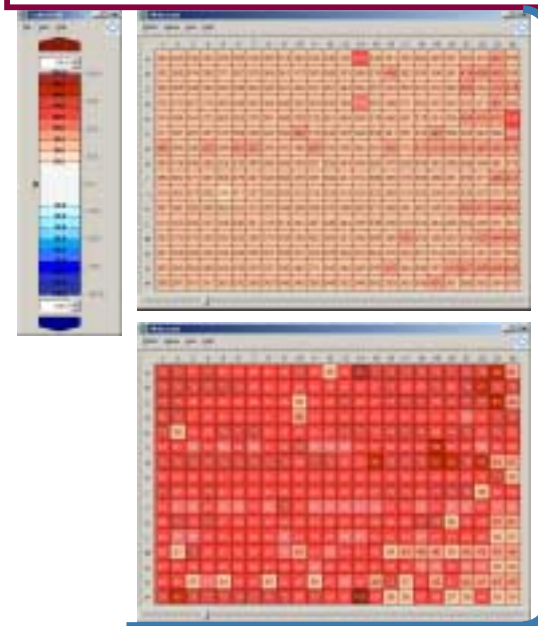
Bioactivity data

Target information

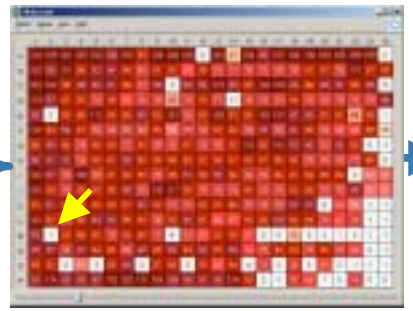
Images

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

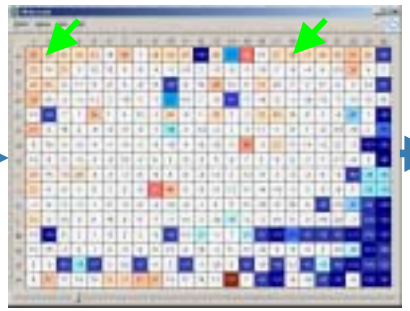
Raw data (Reader output)



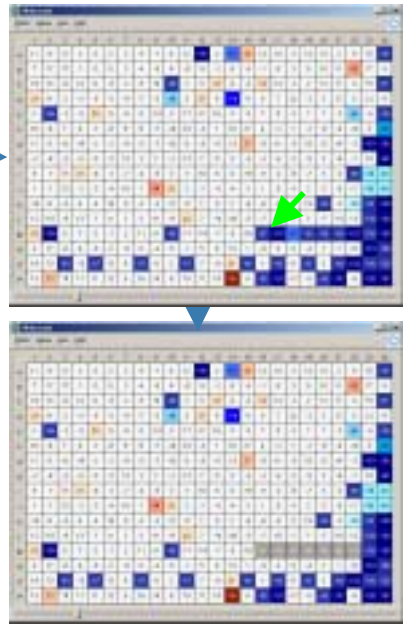
Calculated Readout



Normalized Data

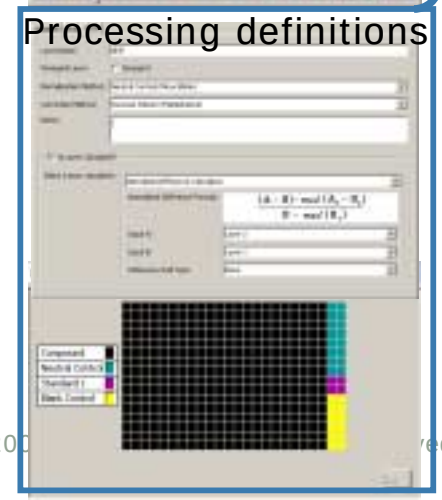


Pattern-corrected Data

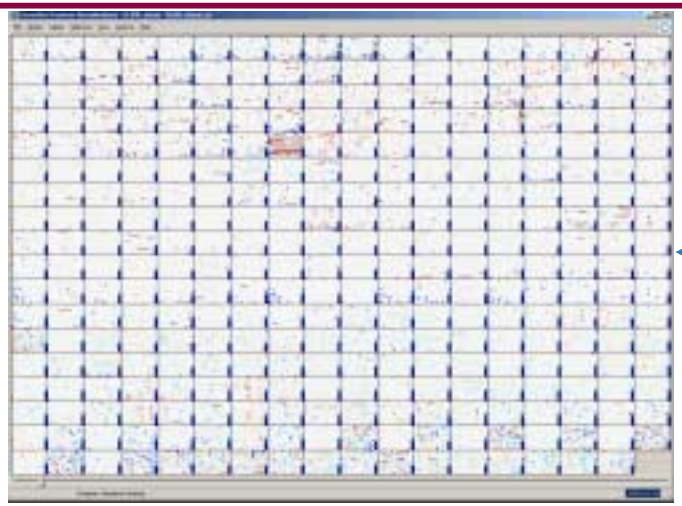


Masking of Bad Data

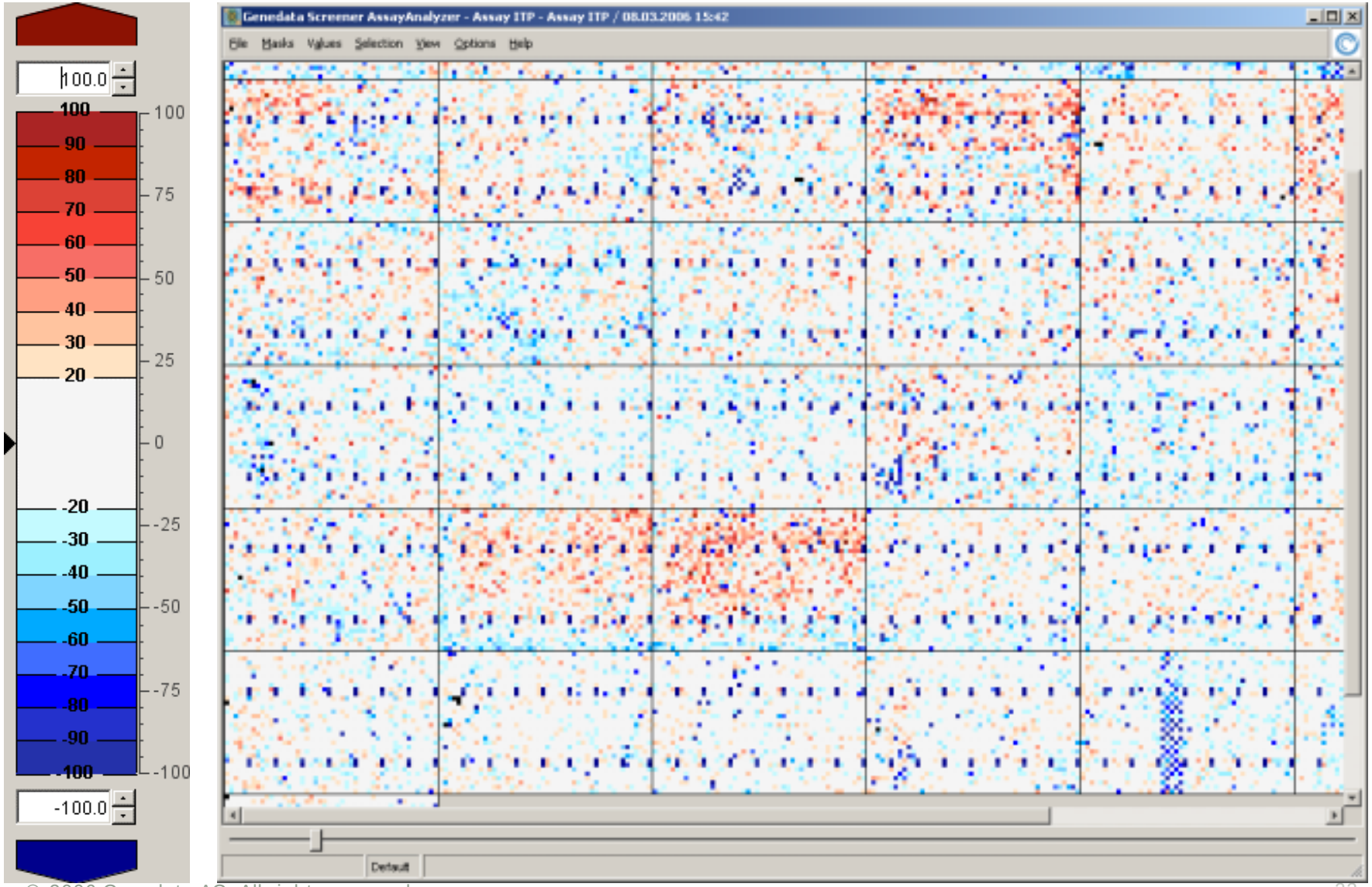
Processing definitions



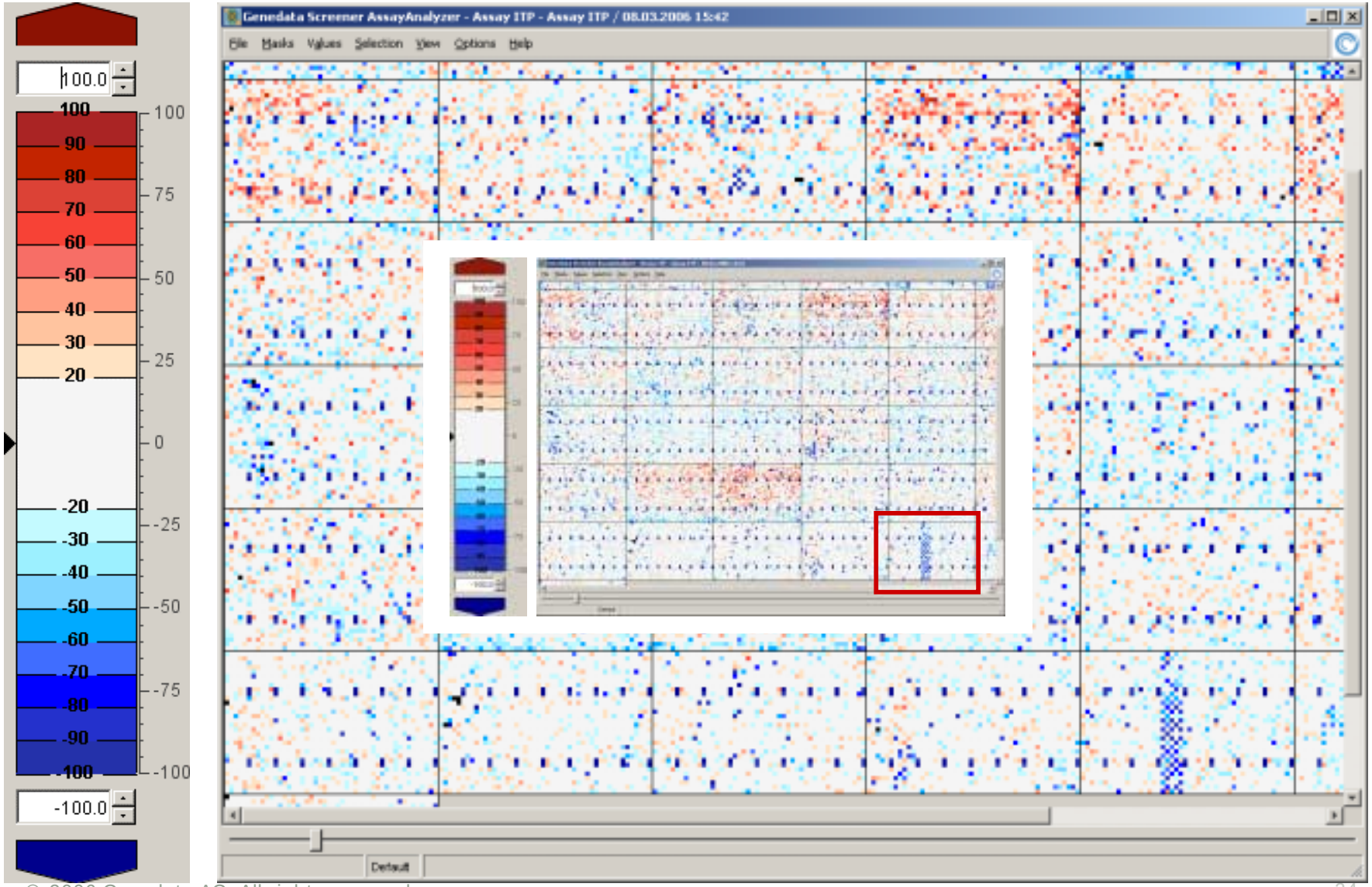
High-quality Results of Entire Screen



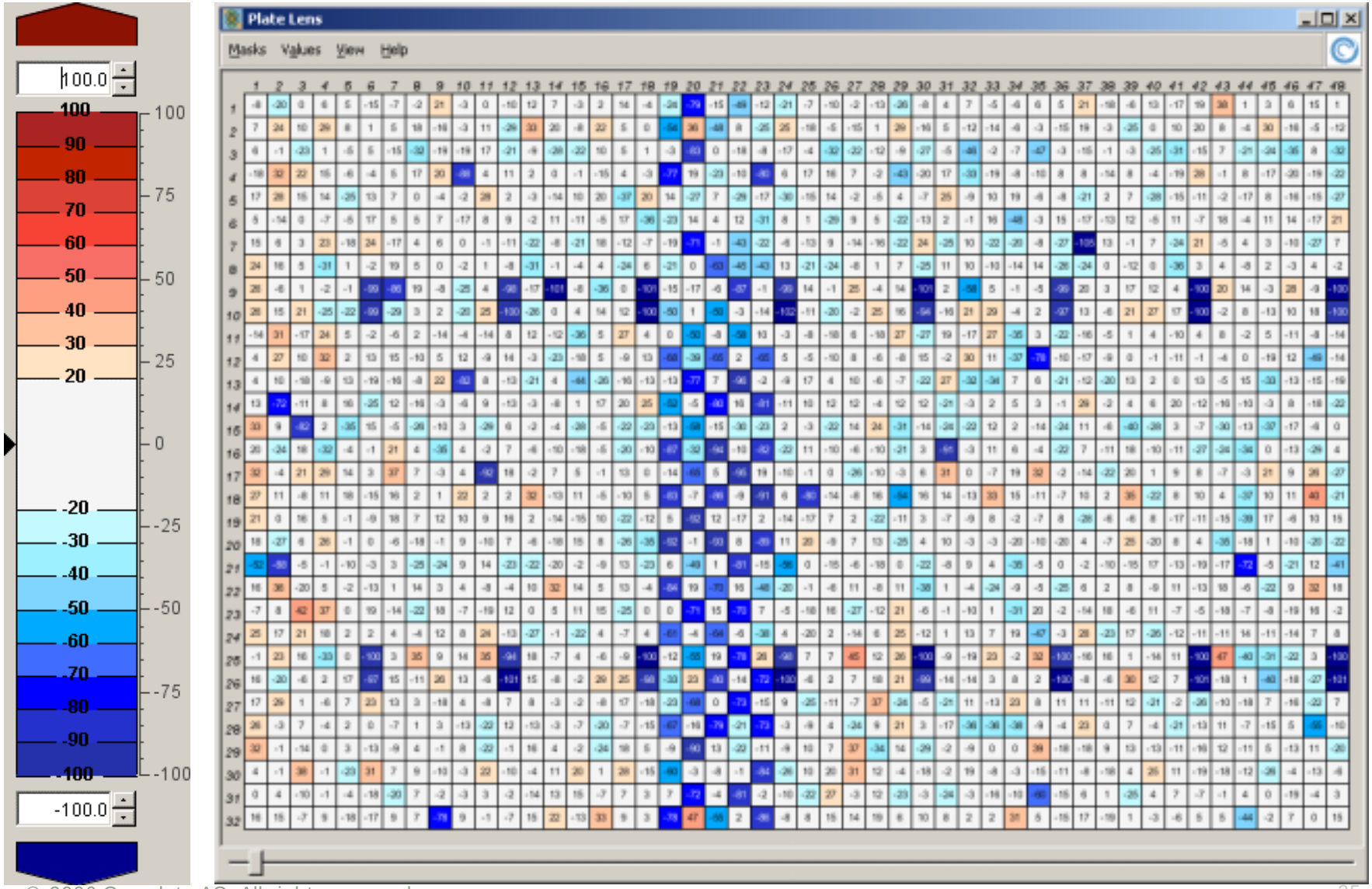
Integration example II



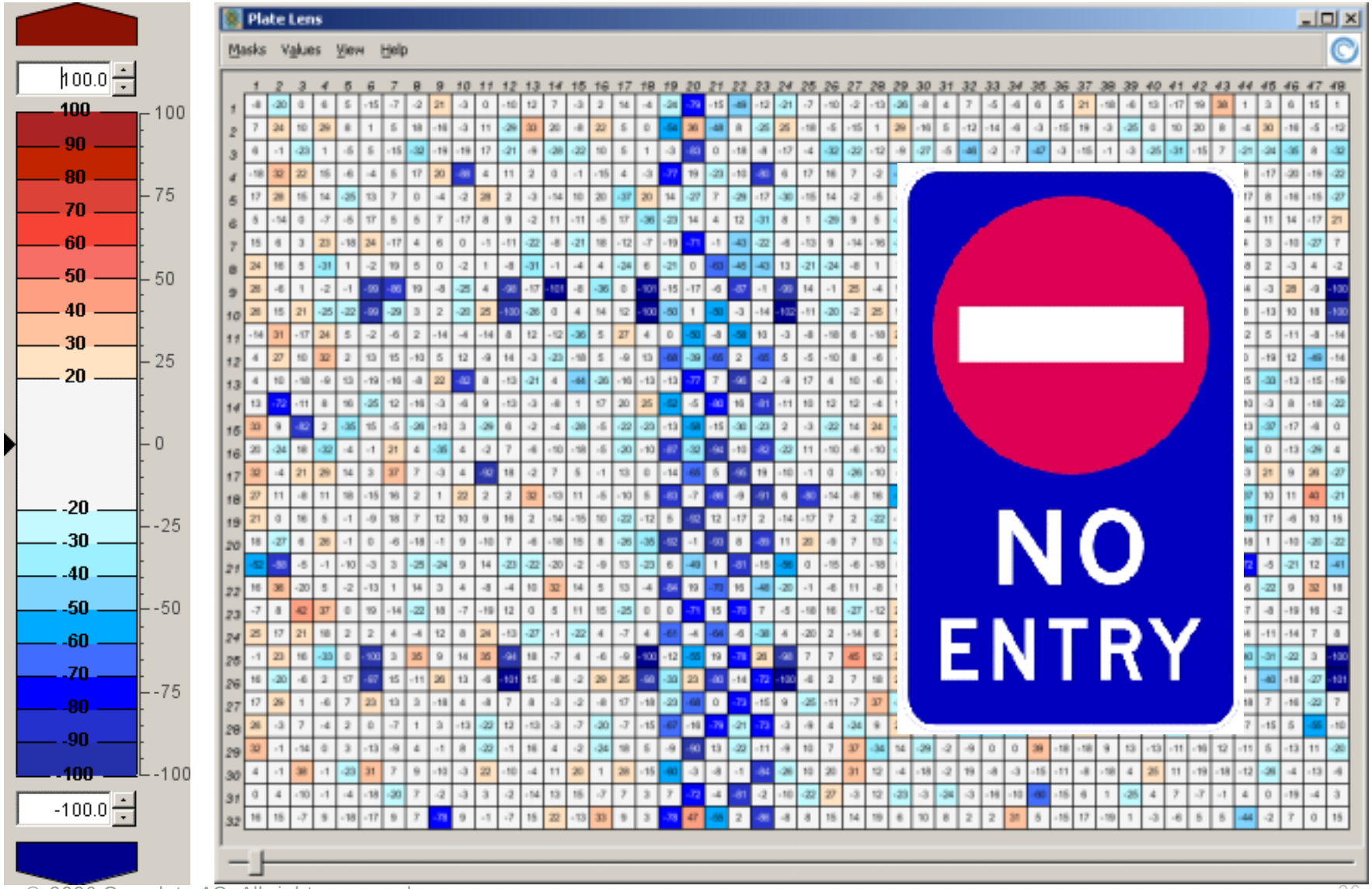
Integration example II



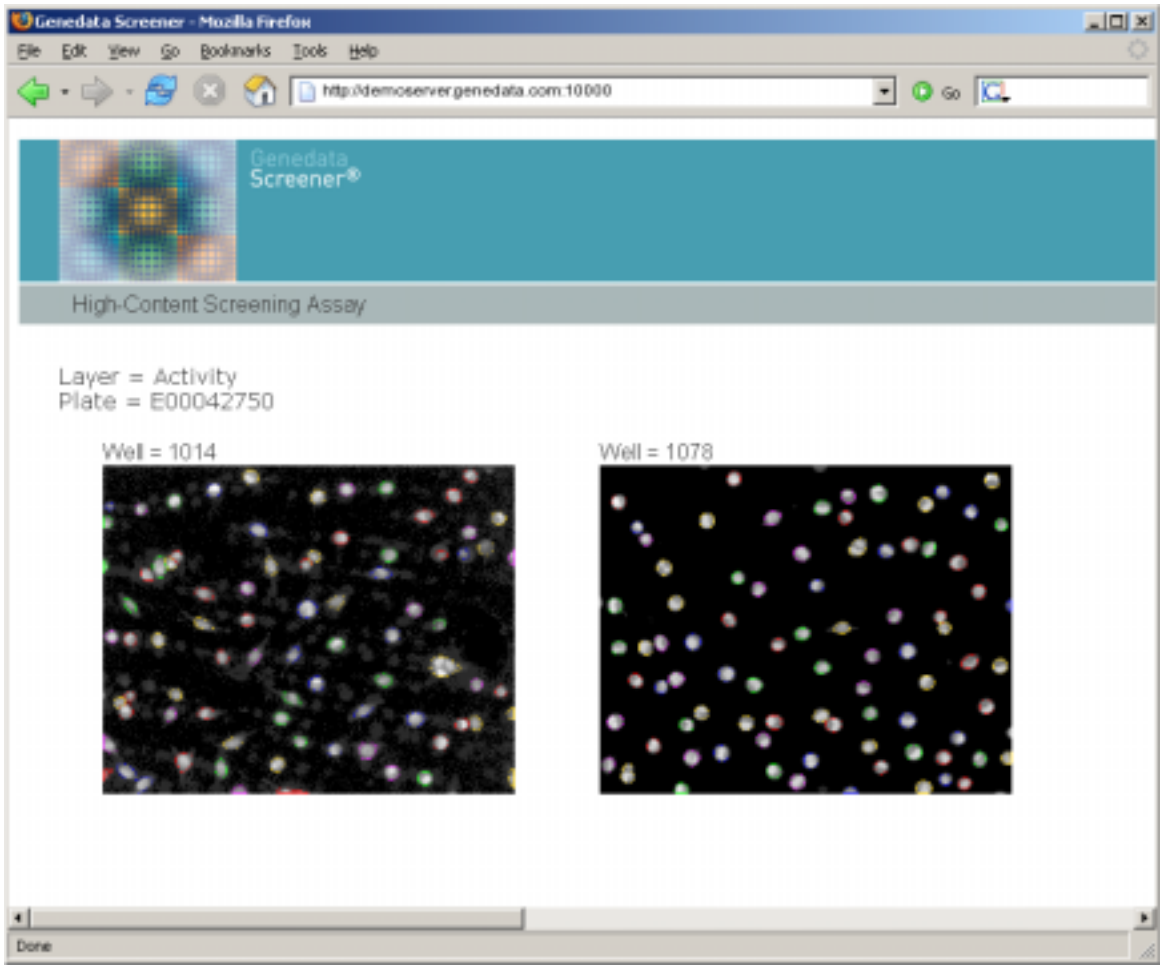
Integration example II



Integration example II



Integration example II



Integration example II

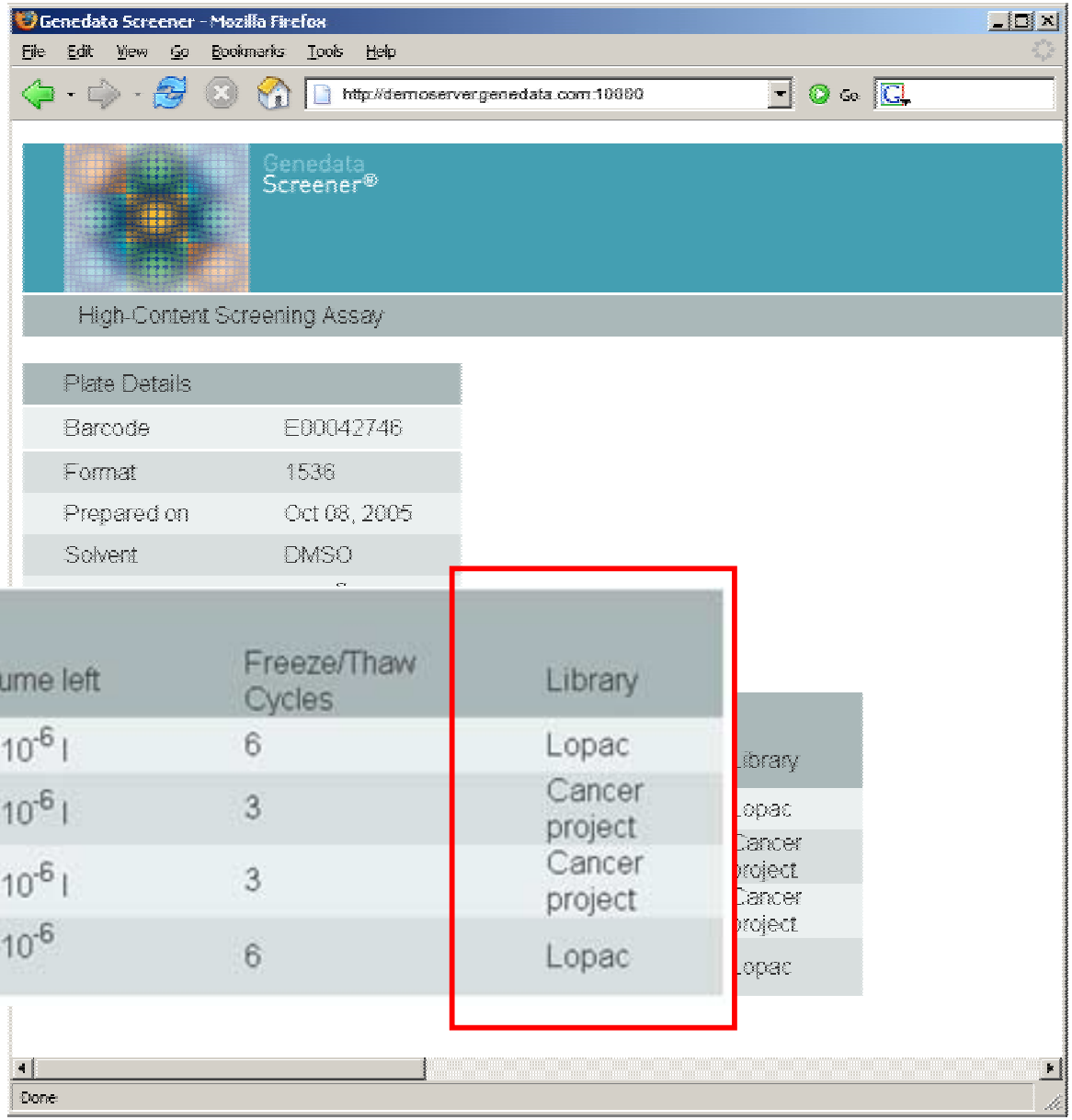
Plates Table

Masks Values Selection Help

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

1 out of 43 items selected

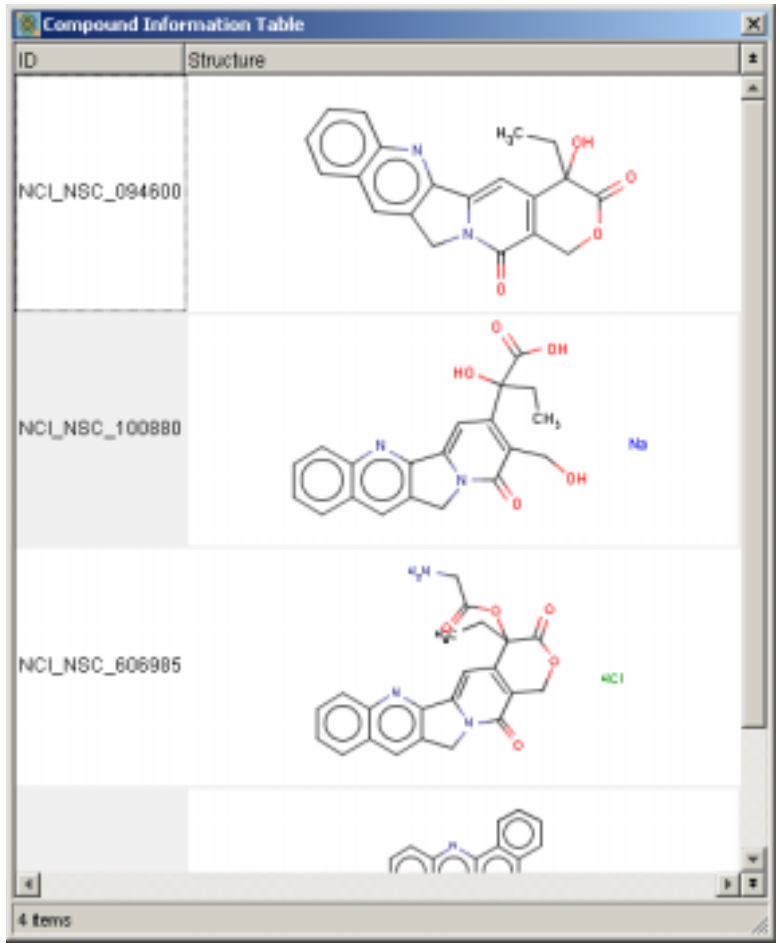
Integration example II



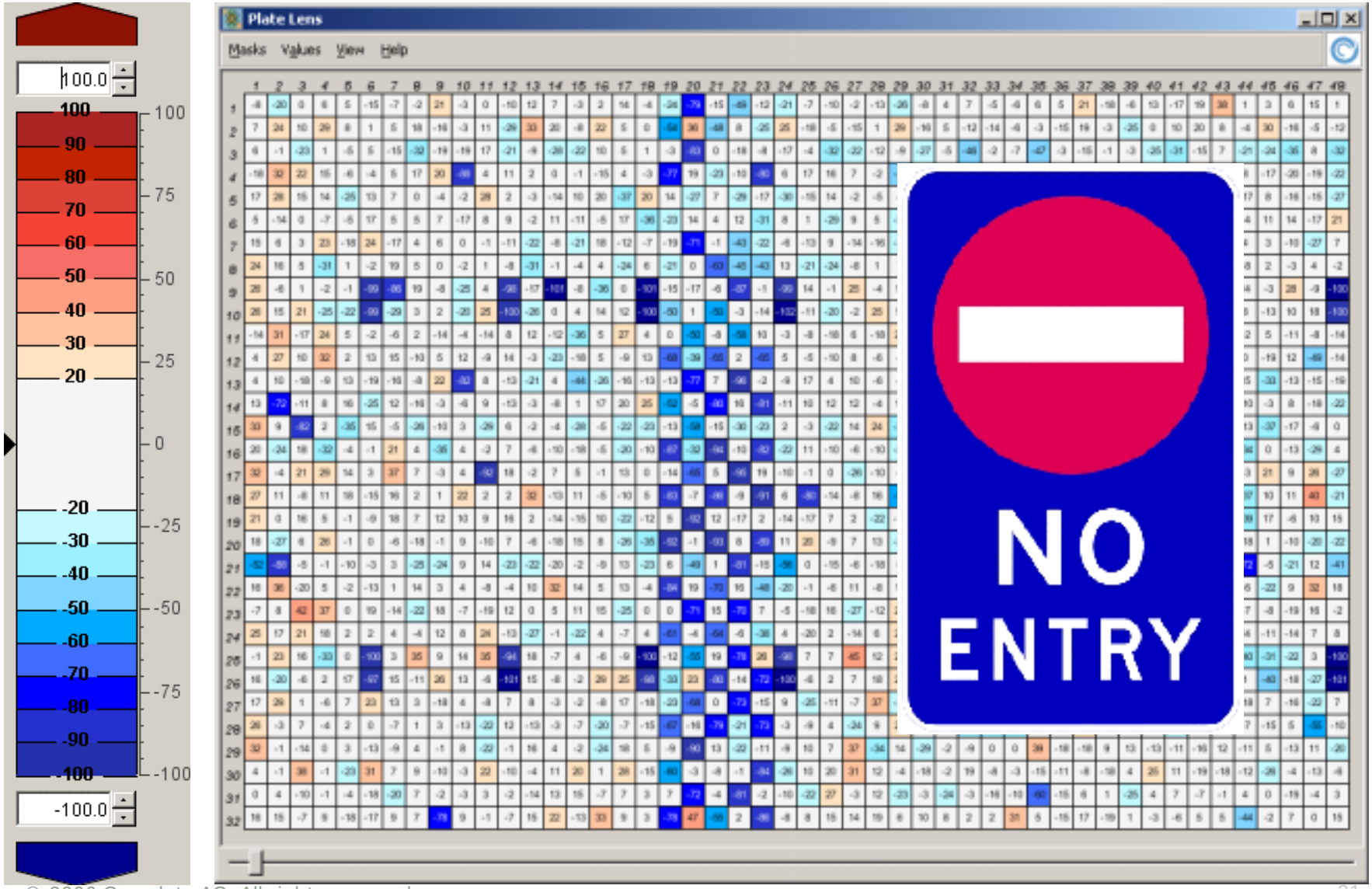
Motherplates

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	$12*10^{-6}$	6	Lopac

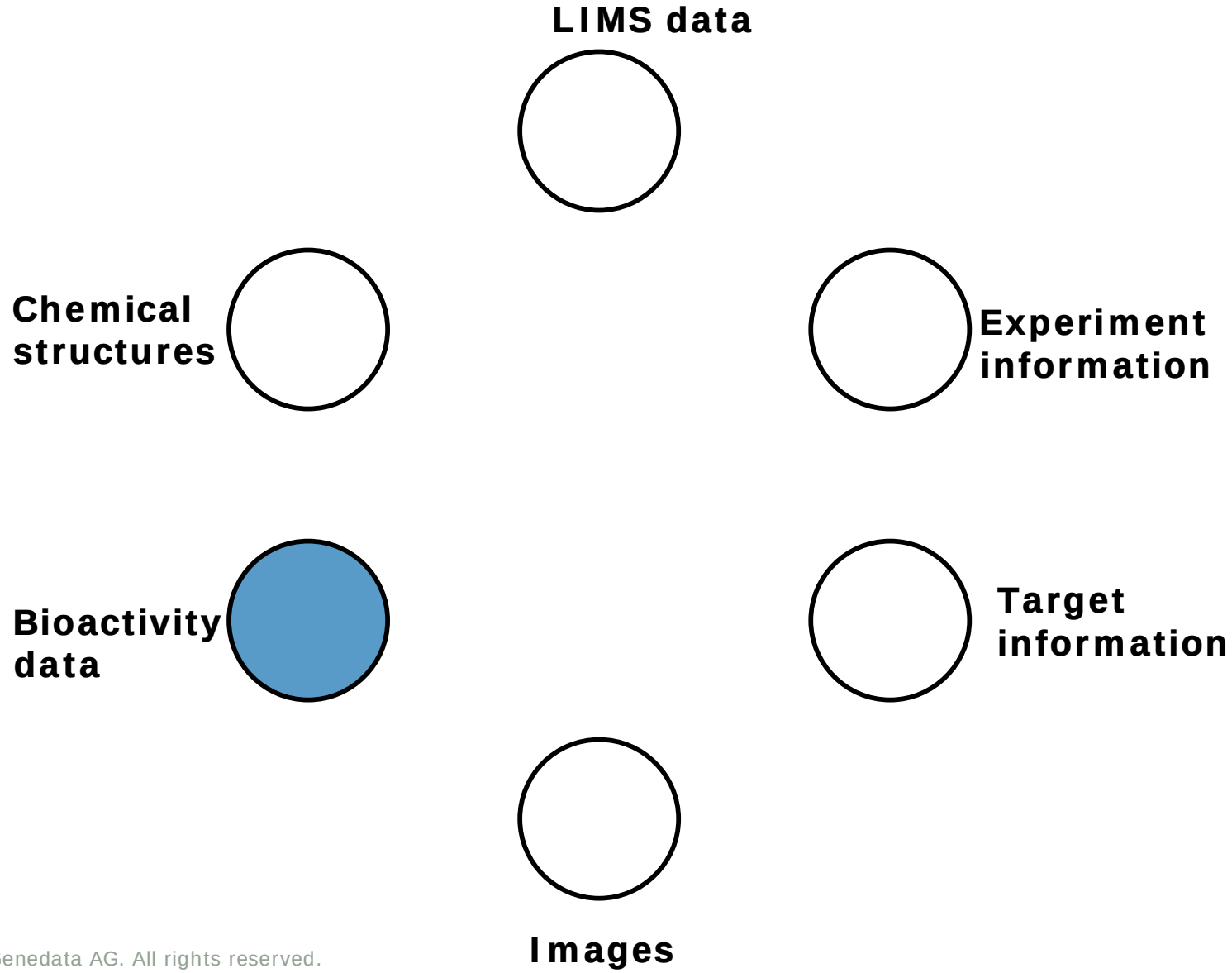
Integration example II



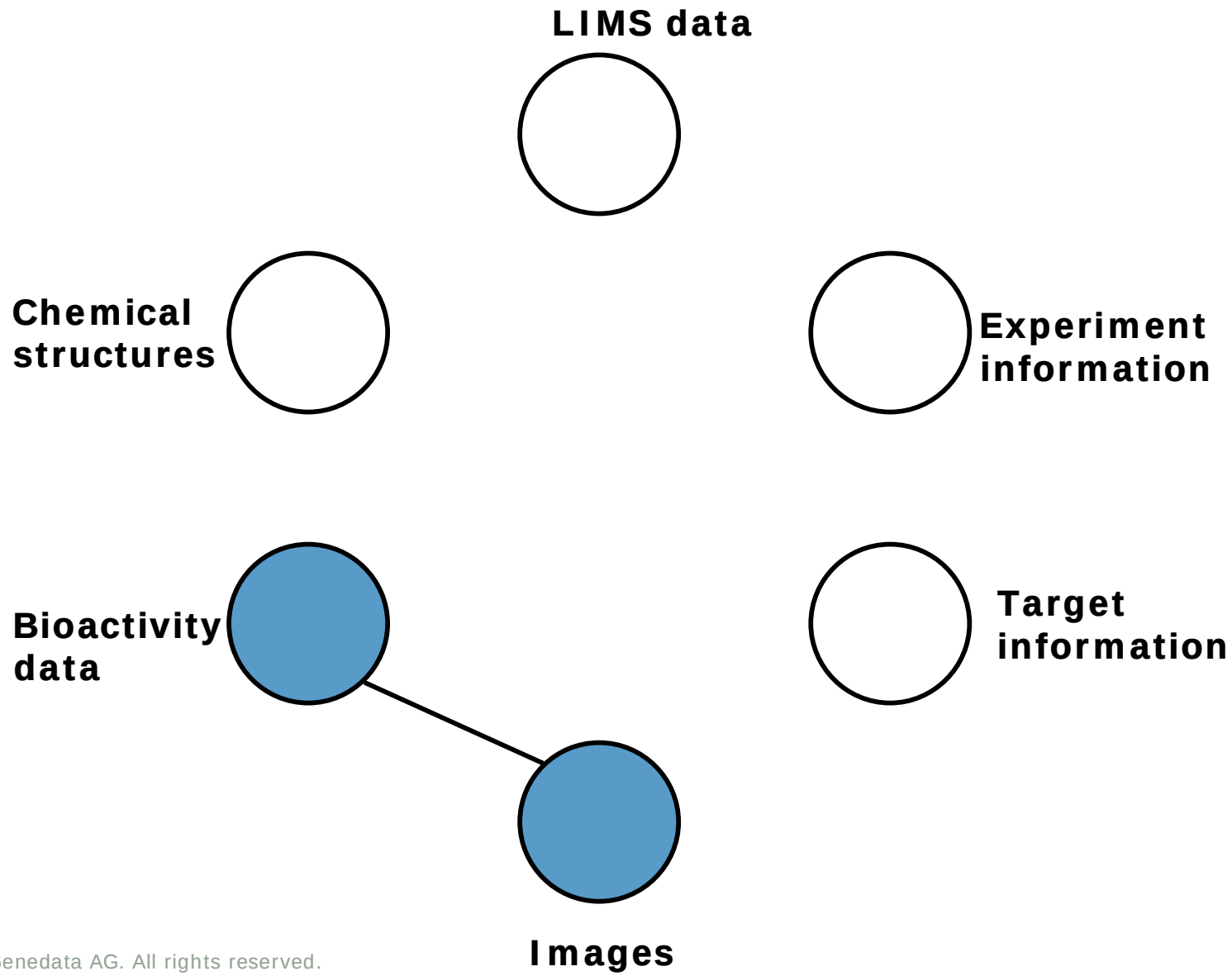
Integration example II



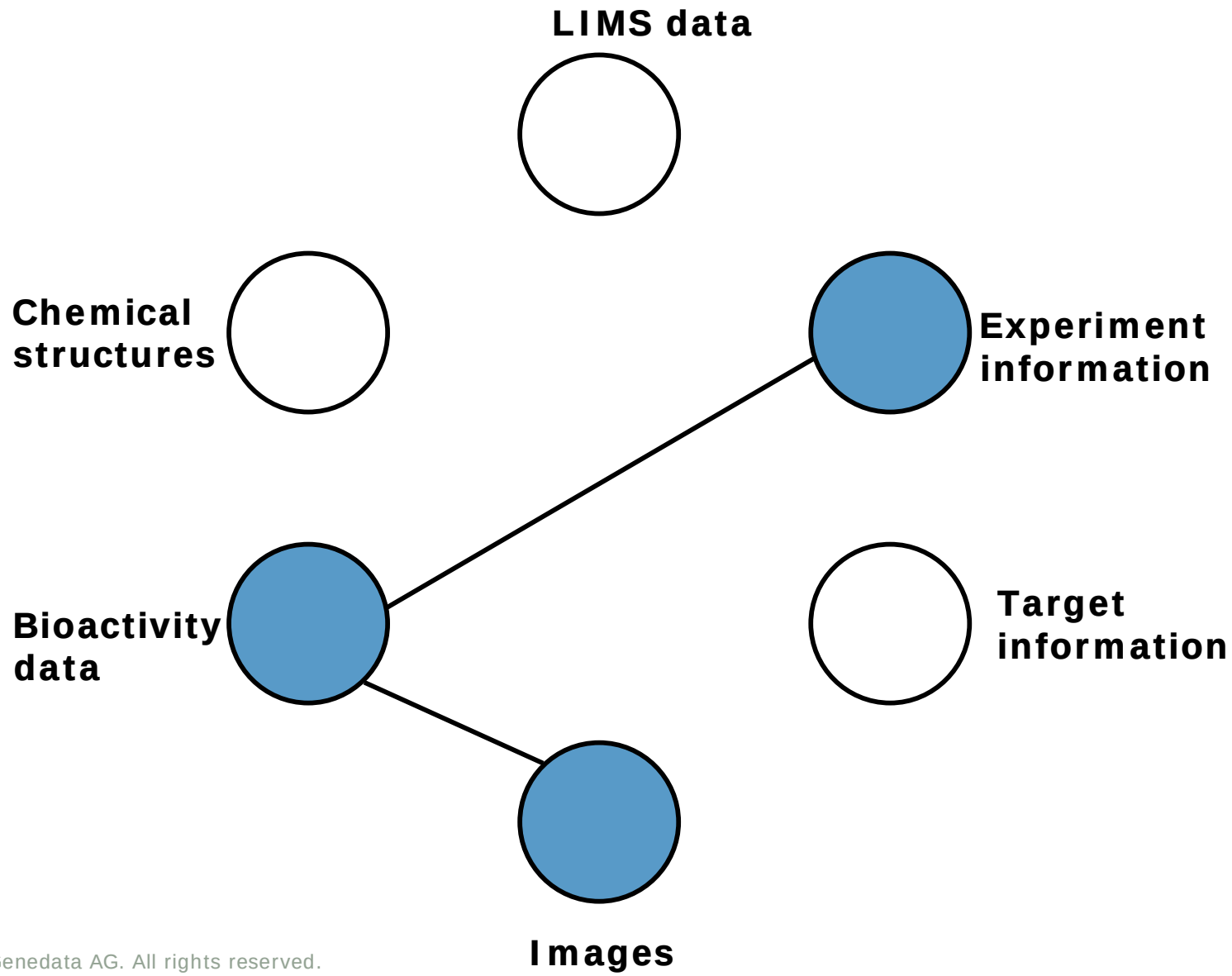
Lead discovery arena – Semantic integration of chemical data



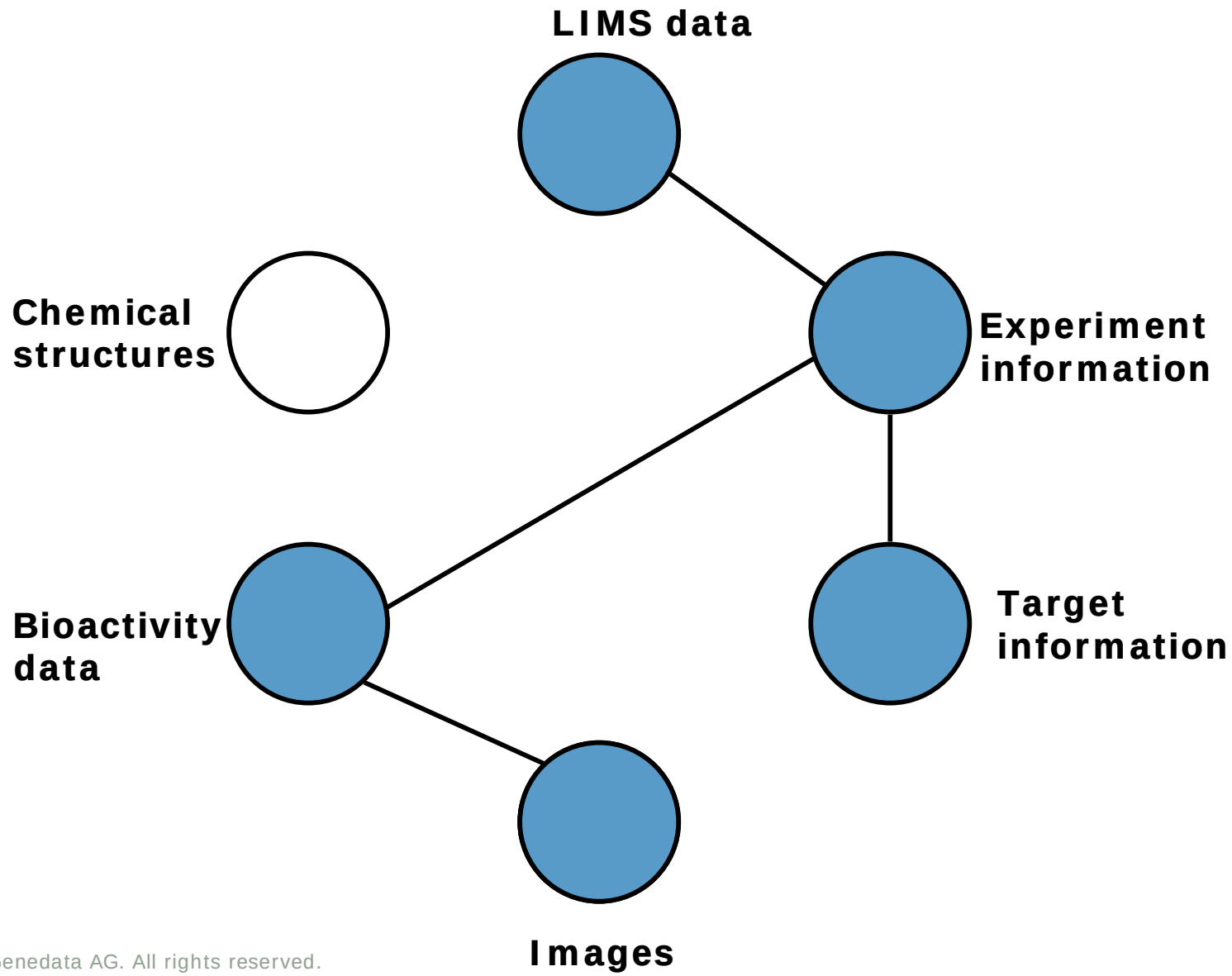
Lead discovery arena – Semantic integration of chemical data



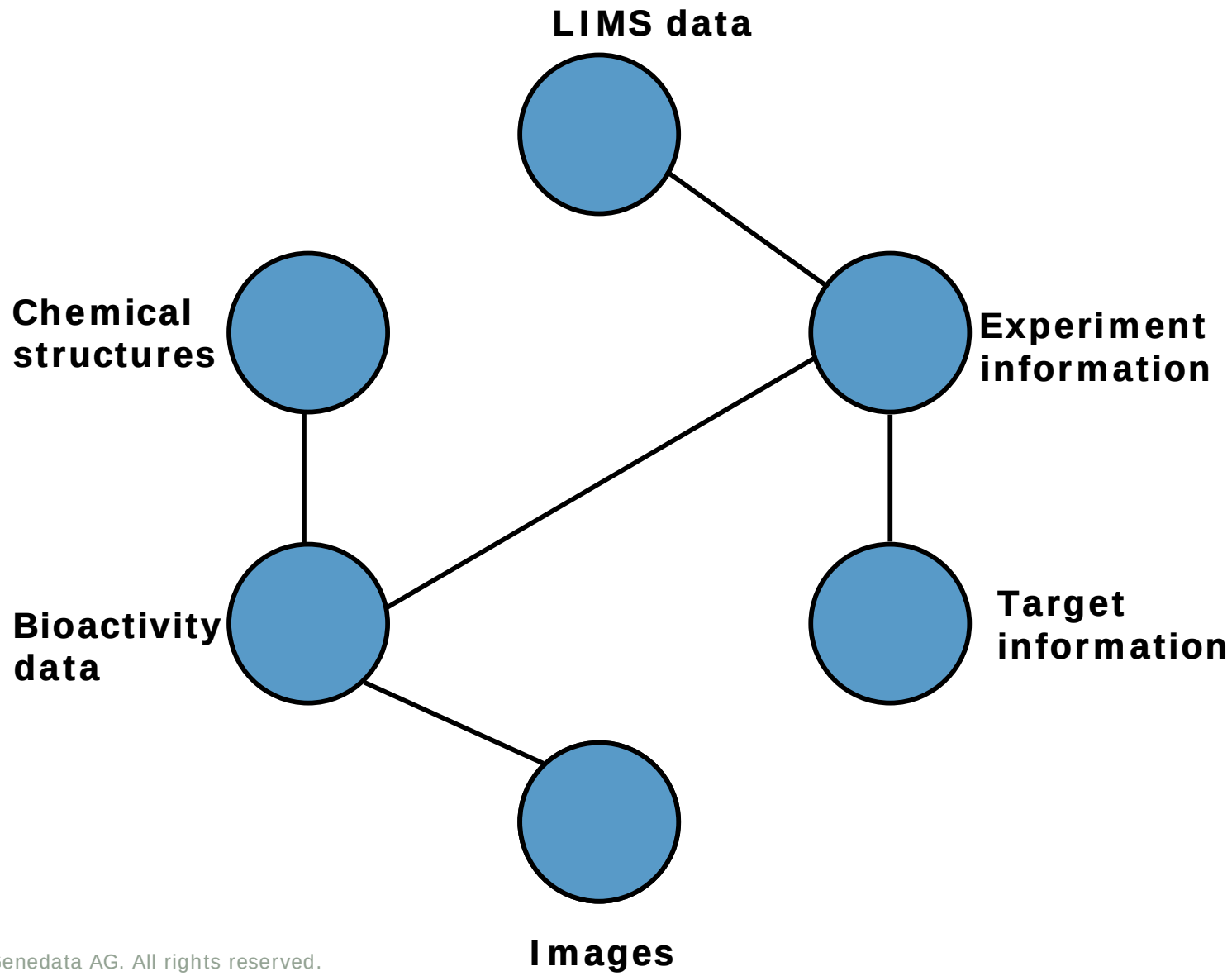
Lead discovery arena – Semantic integration of chemical data



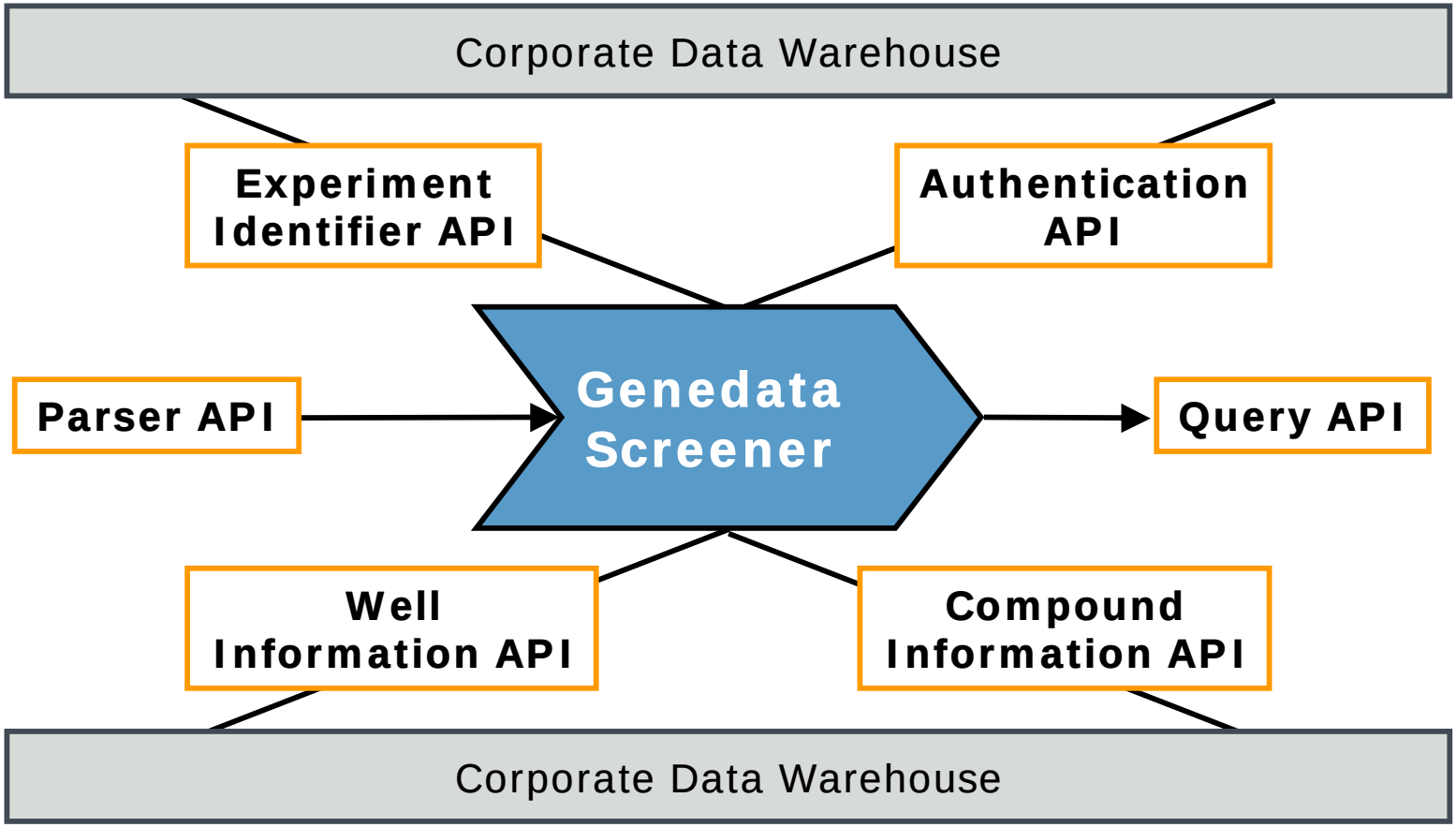
Lead discovery arena – Semantic integration of chemical data



Lead discovery arena – Semantic integration of chemical data

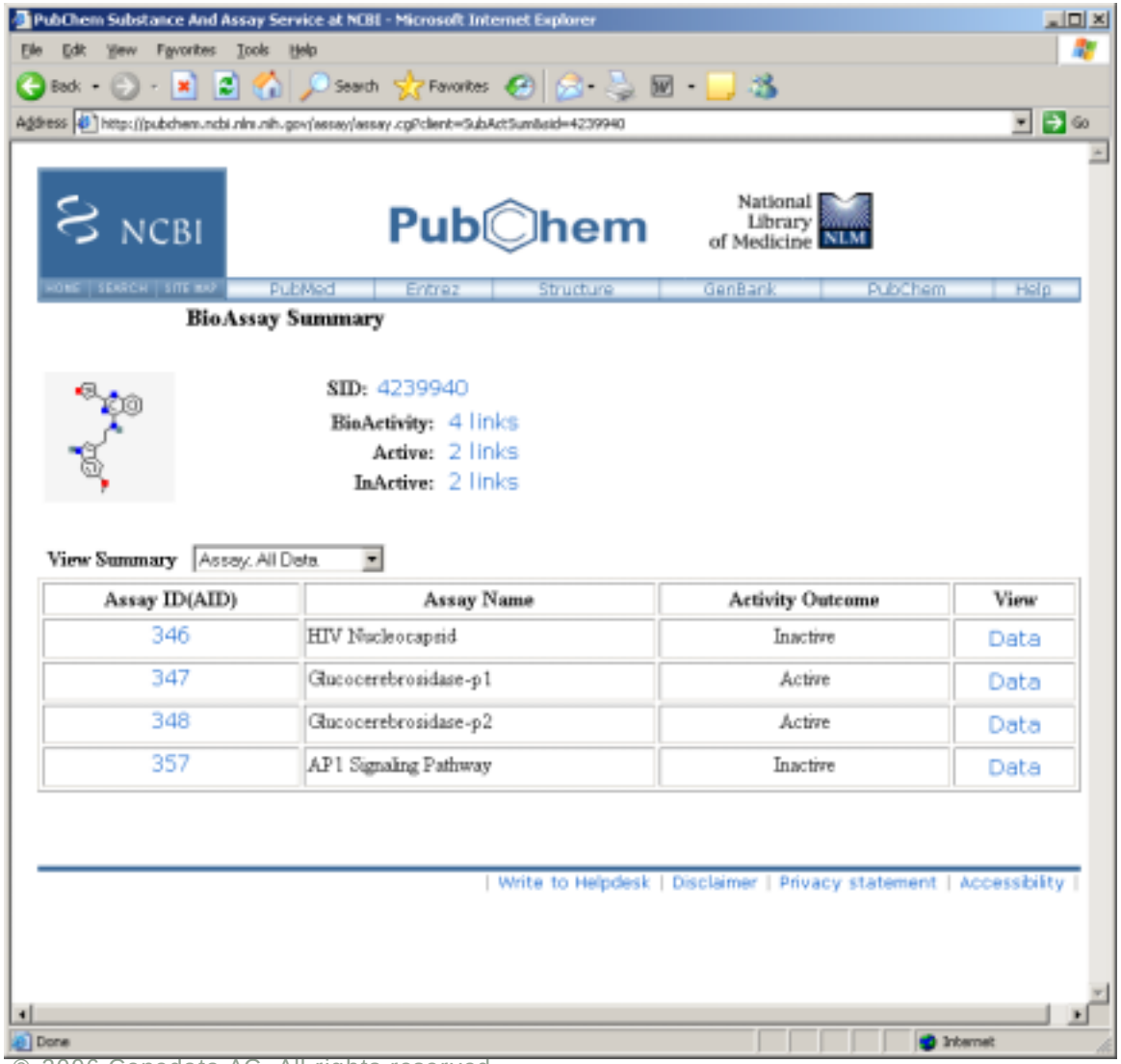


Semantic integration of specific experimental data with Corporate Information

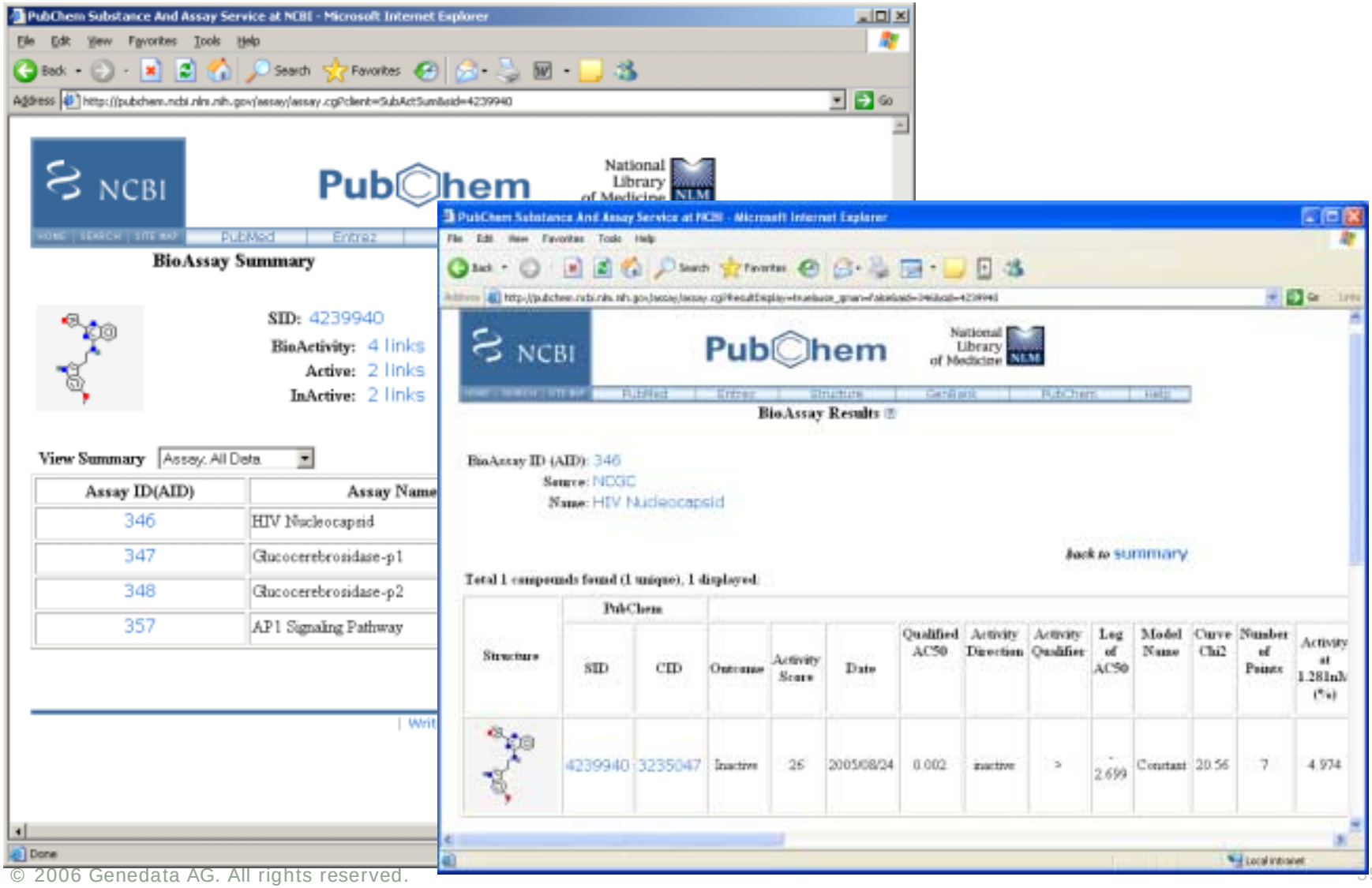


Easy integration with corporate-wide IT infrastructure increases convenience and facilitates decision making process

Lead Discovery Portal - PubChem

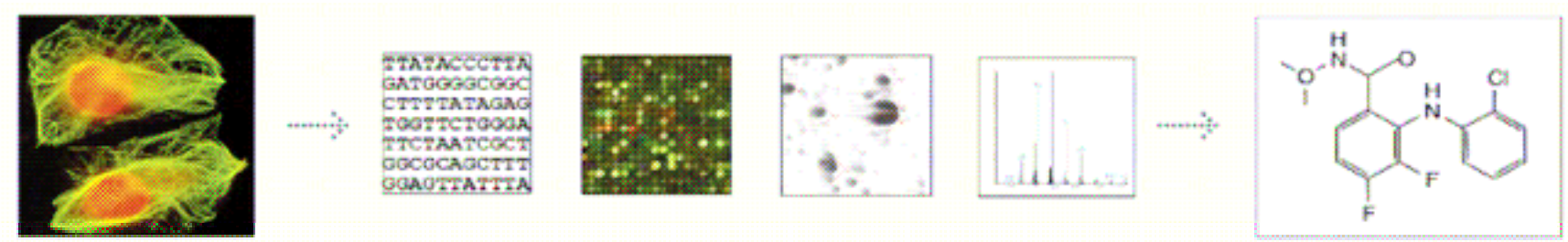


Lead Discovery Portal - PubChem

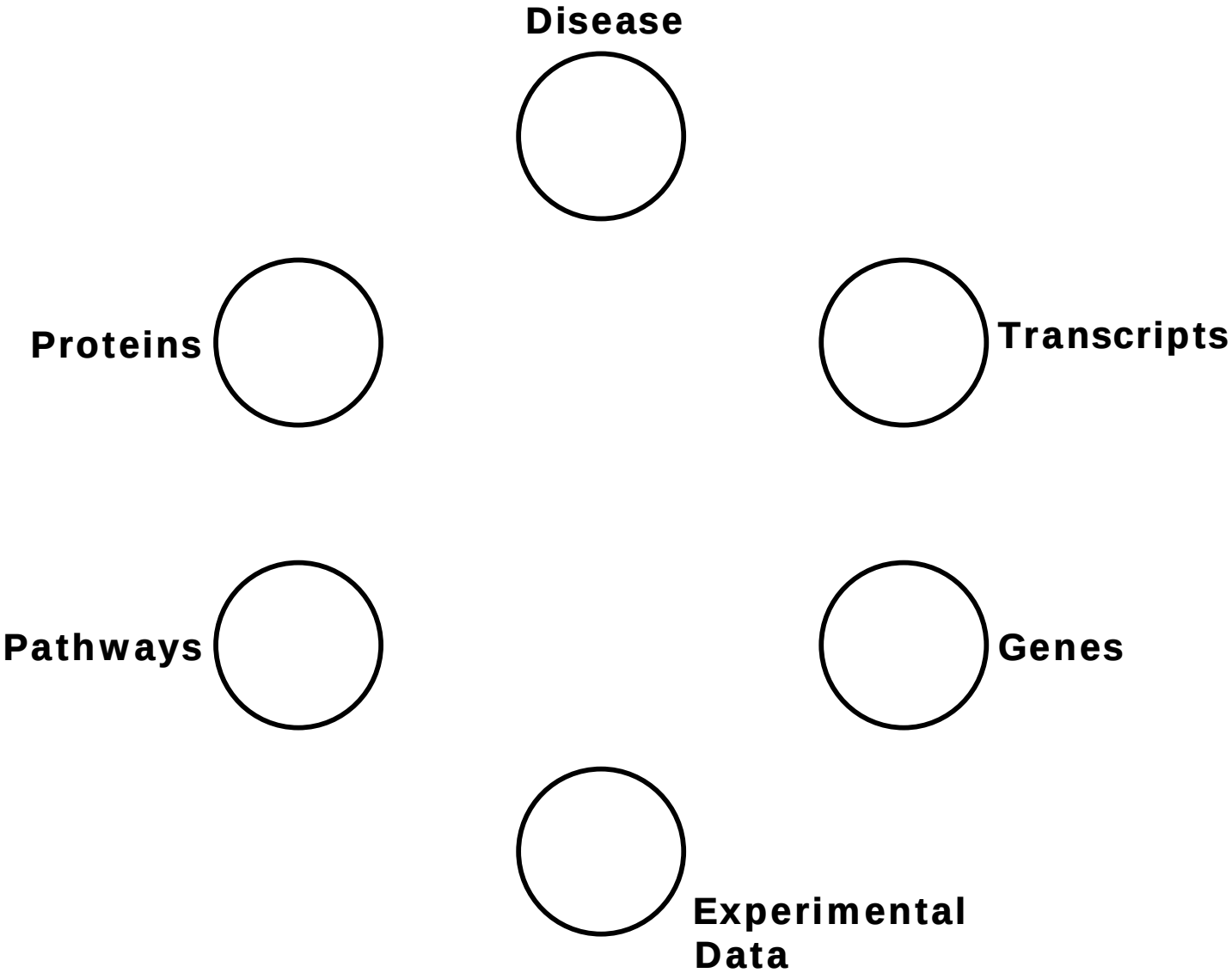


Integration example III

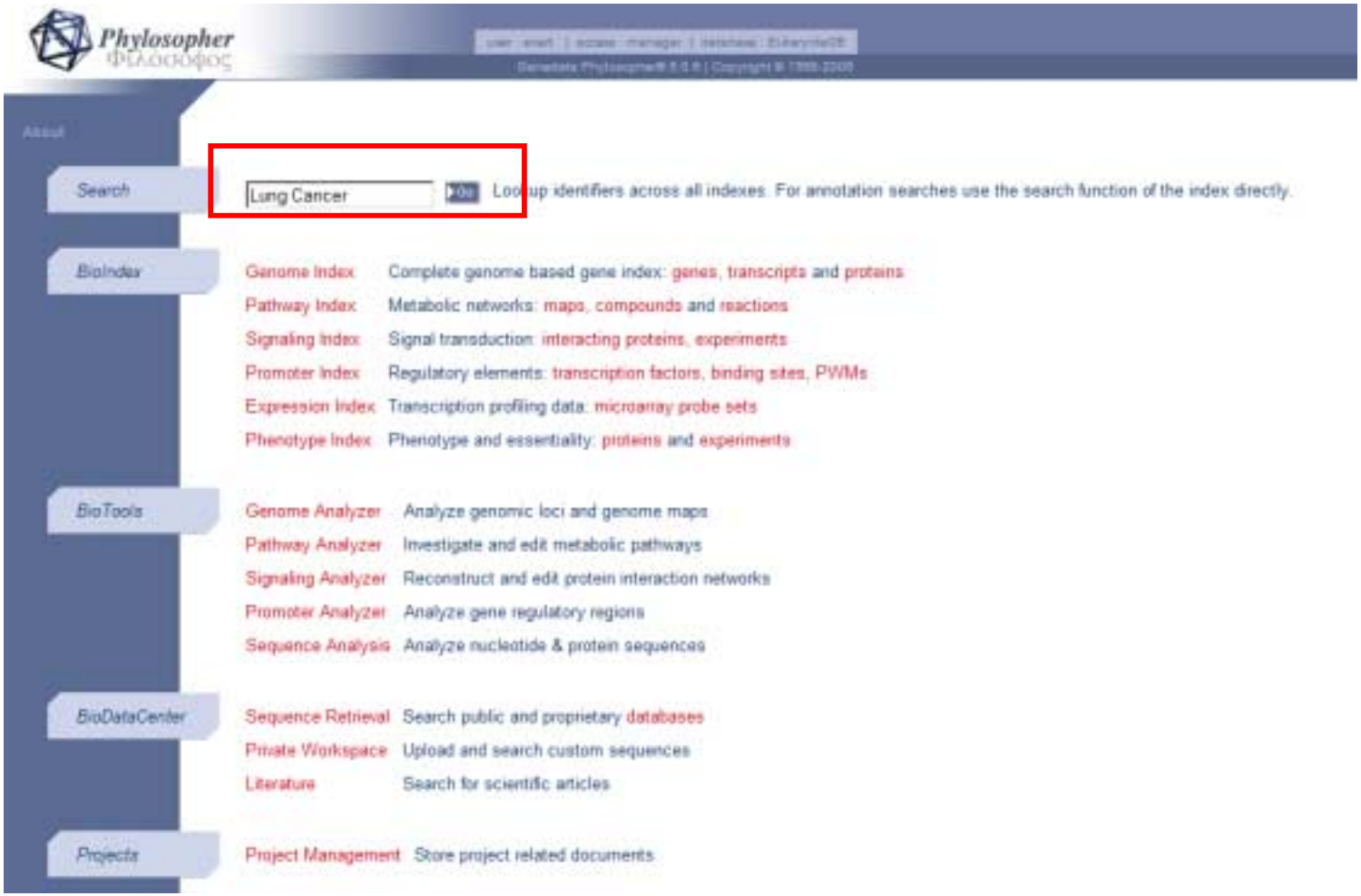
- **Target Discovery using …**
  - Public and proprietary data
  - Integrating data from different ‘omics technologies



Target discovery arena – Semantic integration of biological data

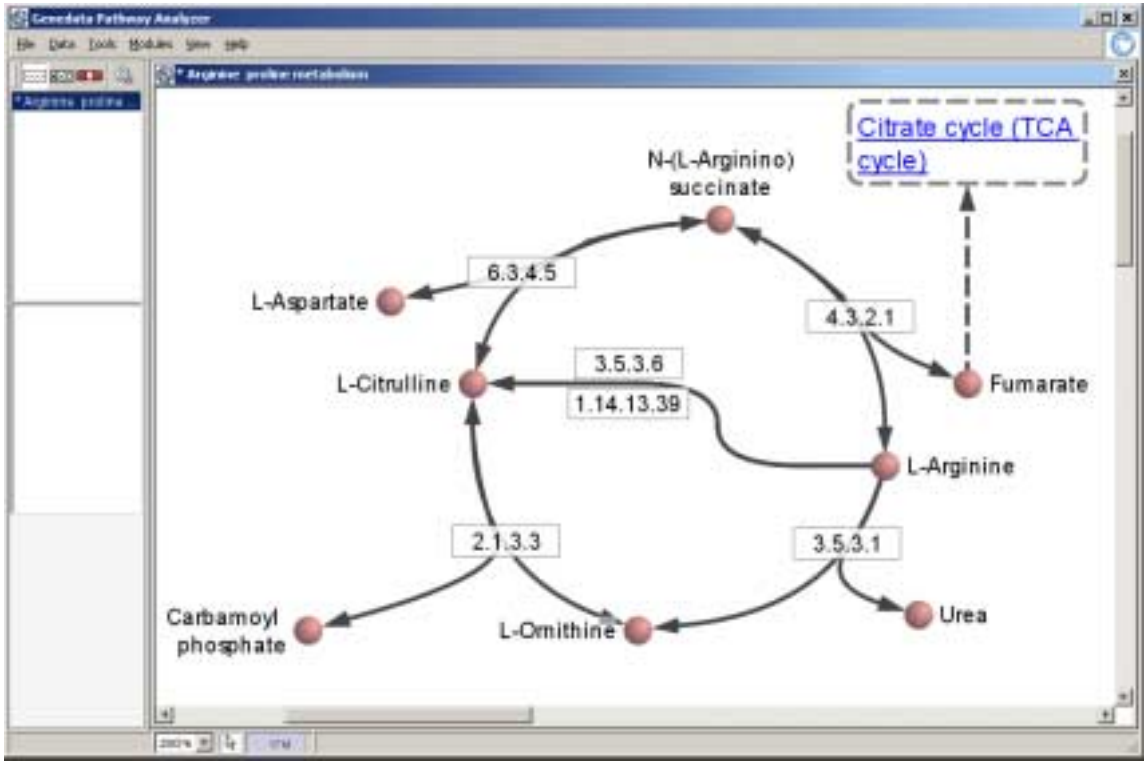


Start with a disease

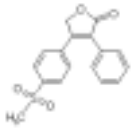


Find related proteins

Investigate pathways



=



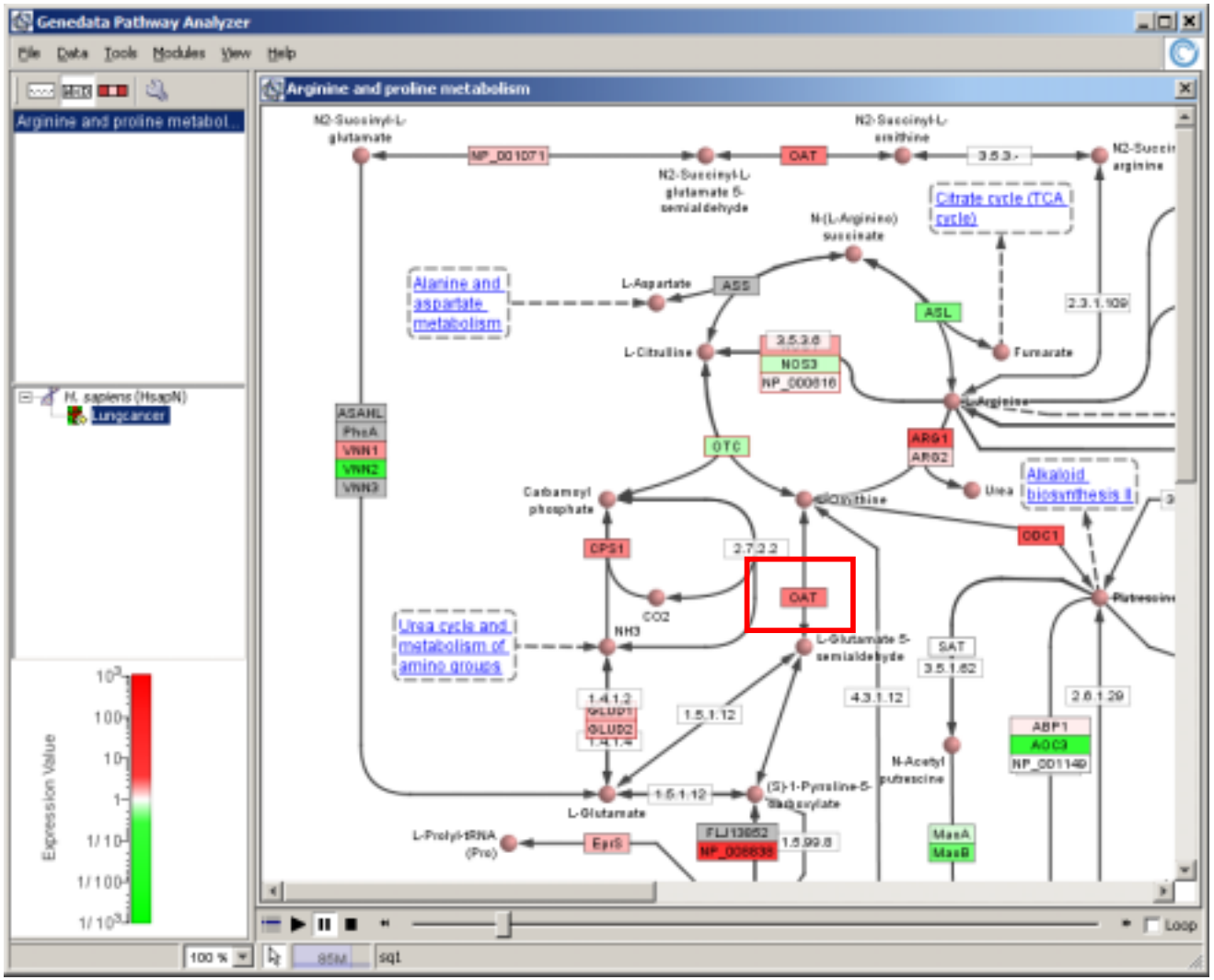
(Metabolite or compound)

1.1.1.83 =

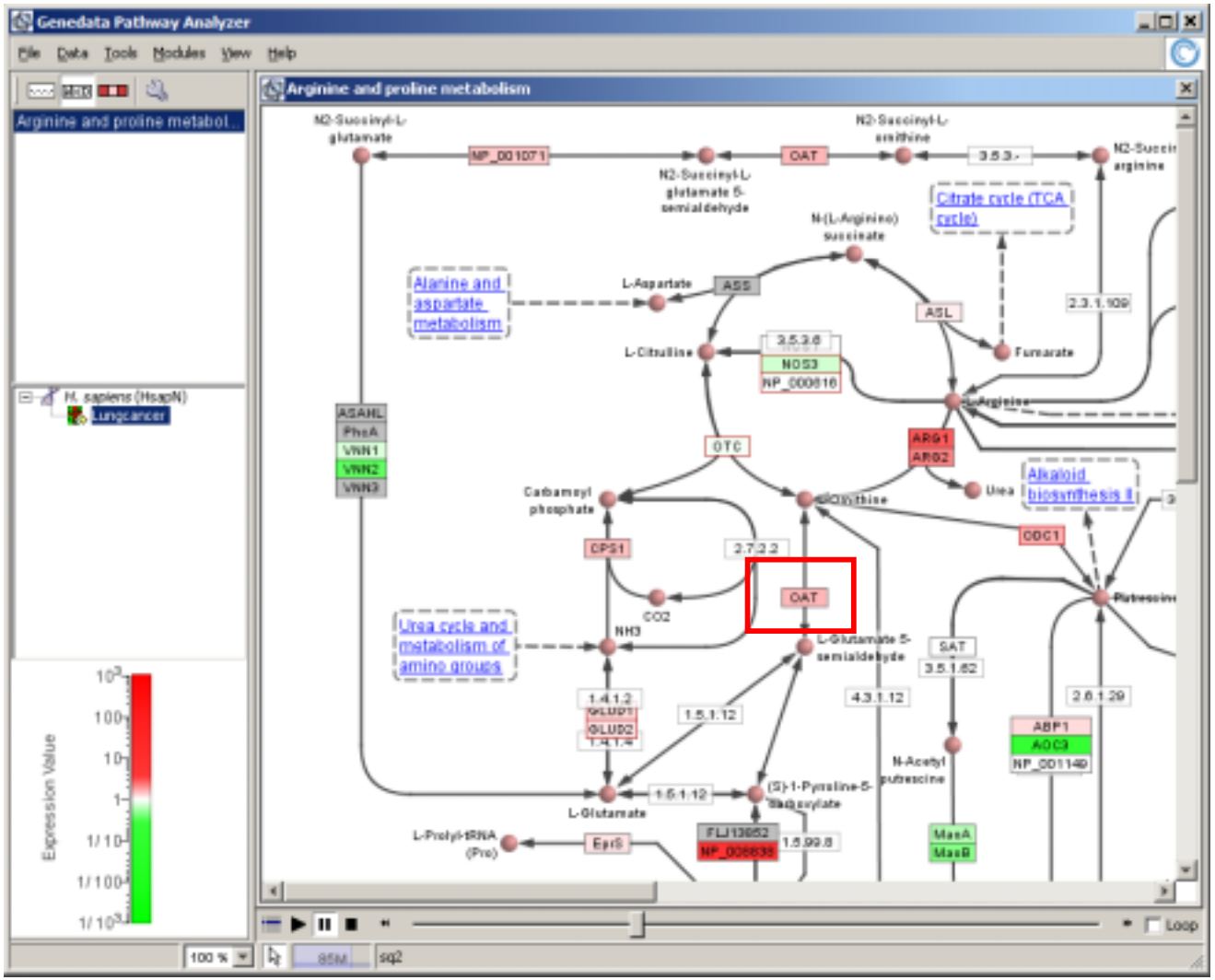
1 $C_{15}H_{24}N_2O_{17}P_2$ (UDPglucose) + 1 $C_{15}H_9O_5R_2$ (Anthocyanidin) ⟷ 1 $C_9H_{14}N_2O_{12}P_2$ (UDP) + 1 $C_{21}H_{19}O_{12}R_2$ (Anthocyanidin-3-O-D-glucoside)

(Biochemical reaction)

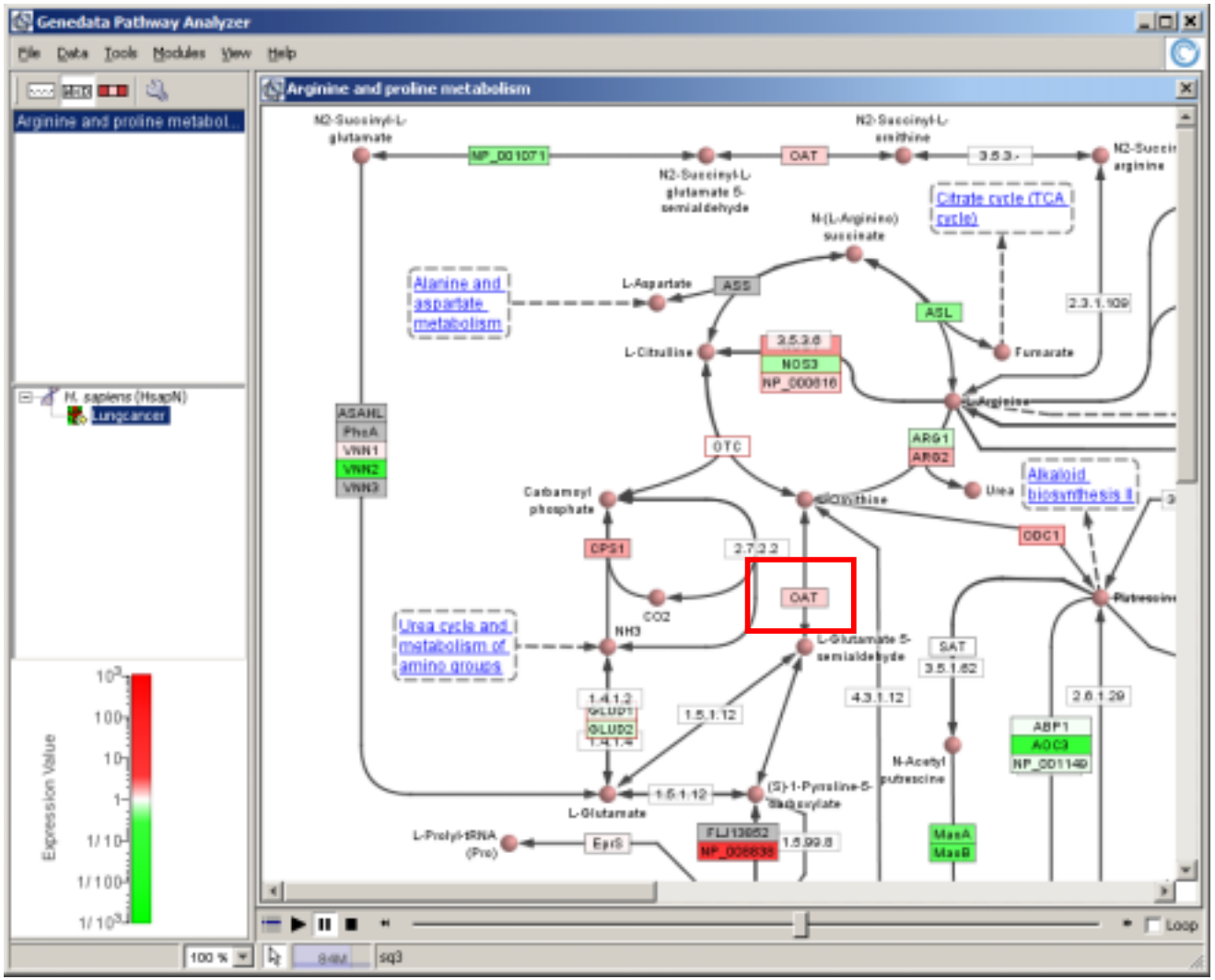
Integrate experimental data



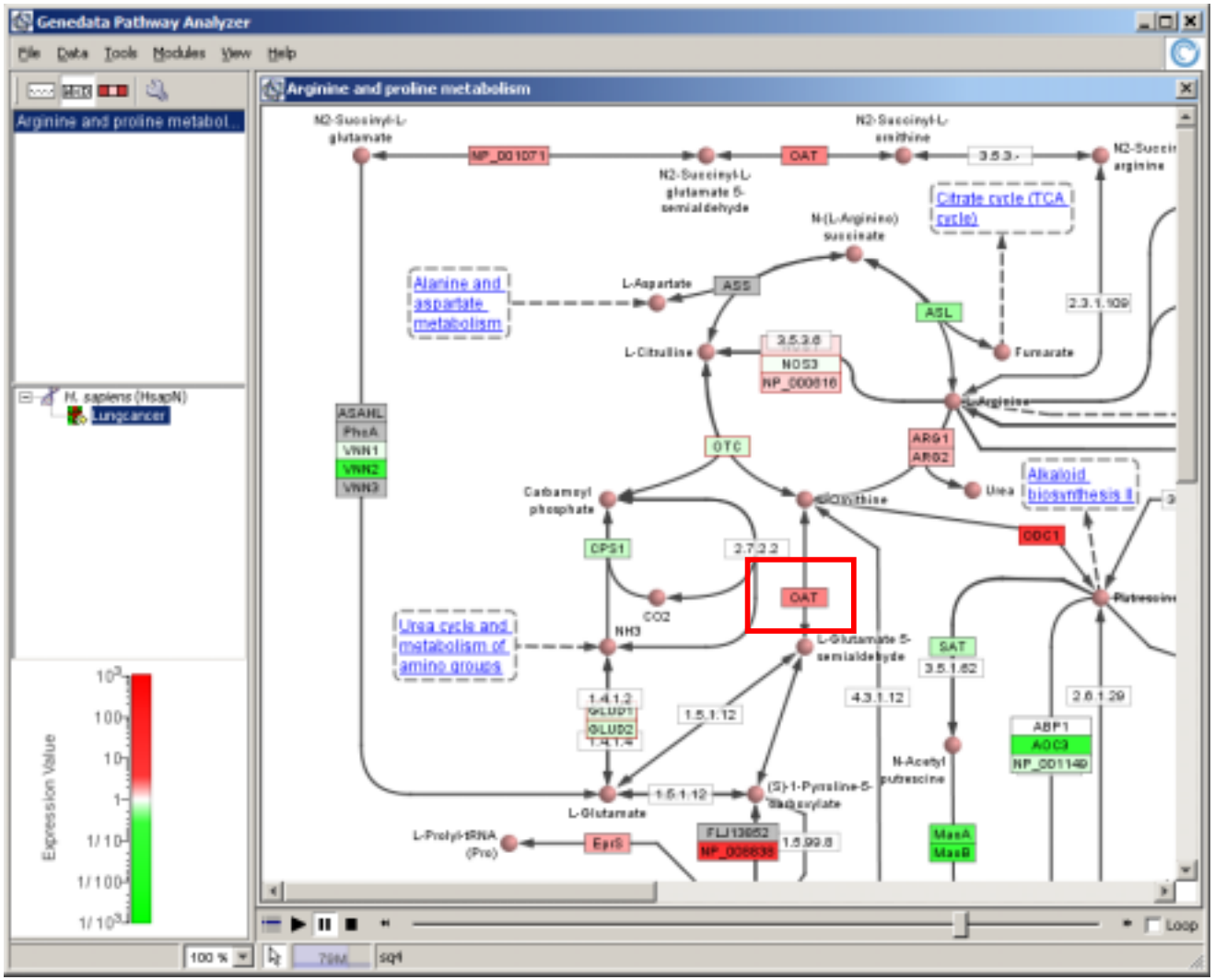
Integrate experimental data



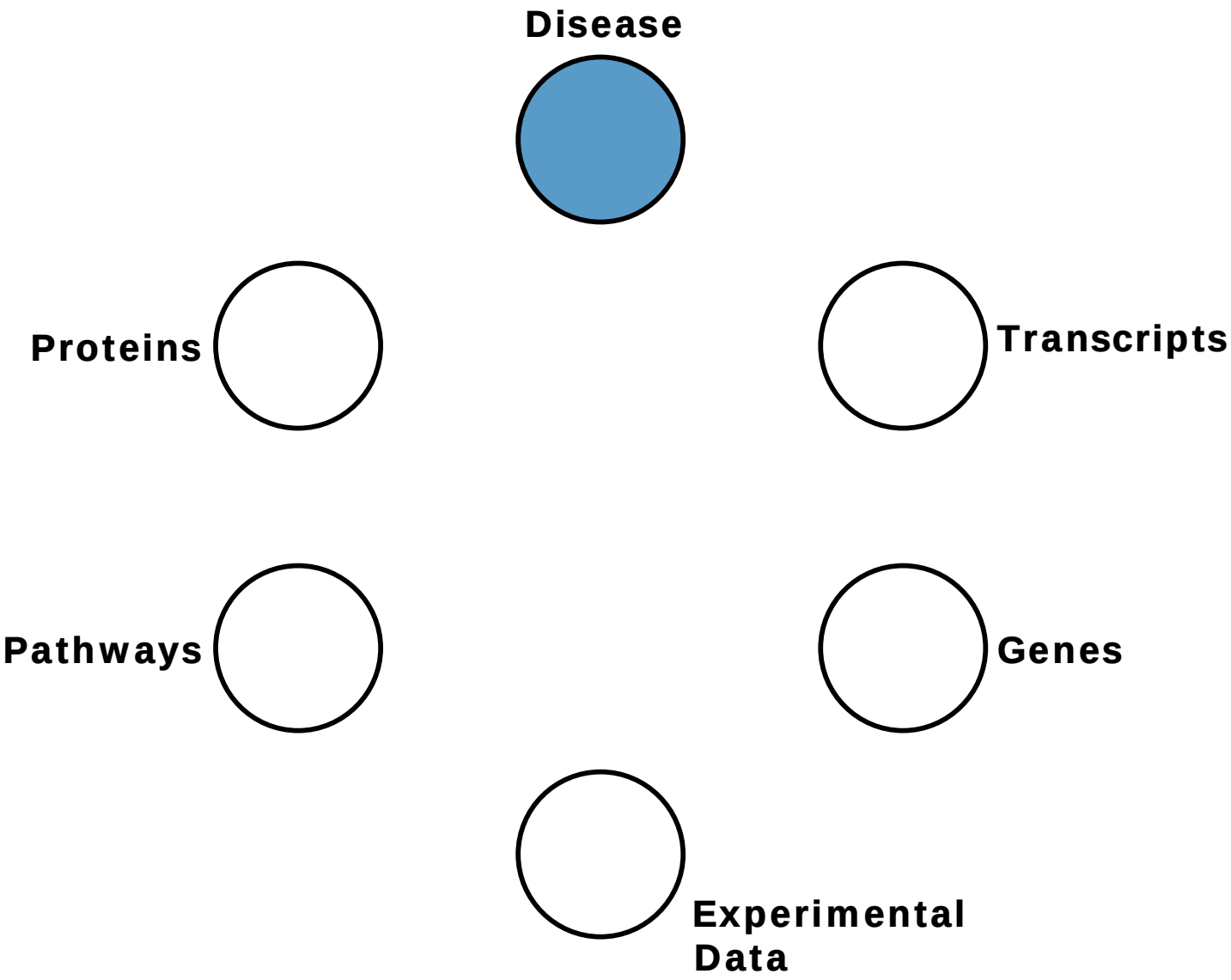
Integrate experimental data



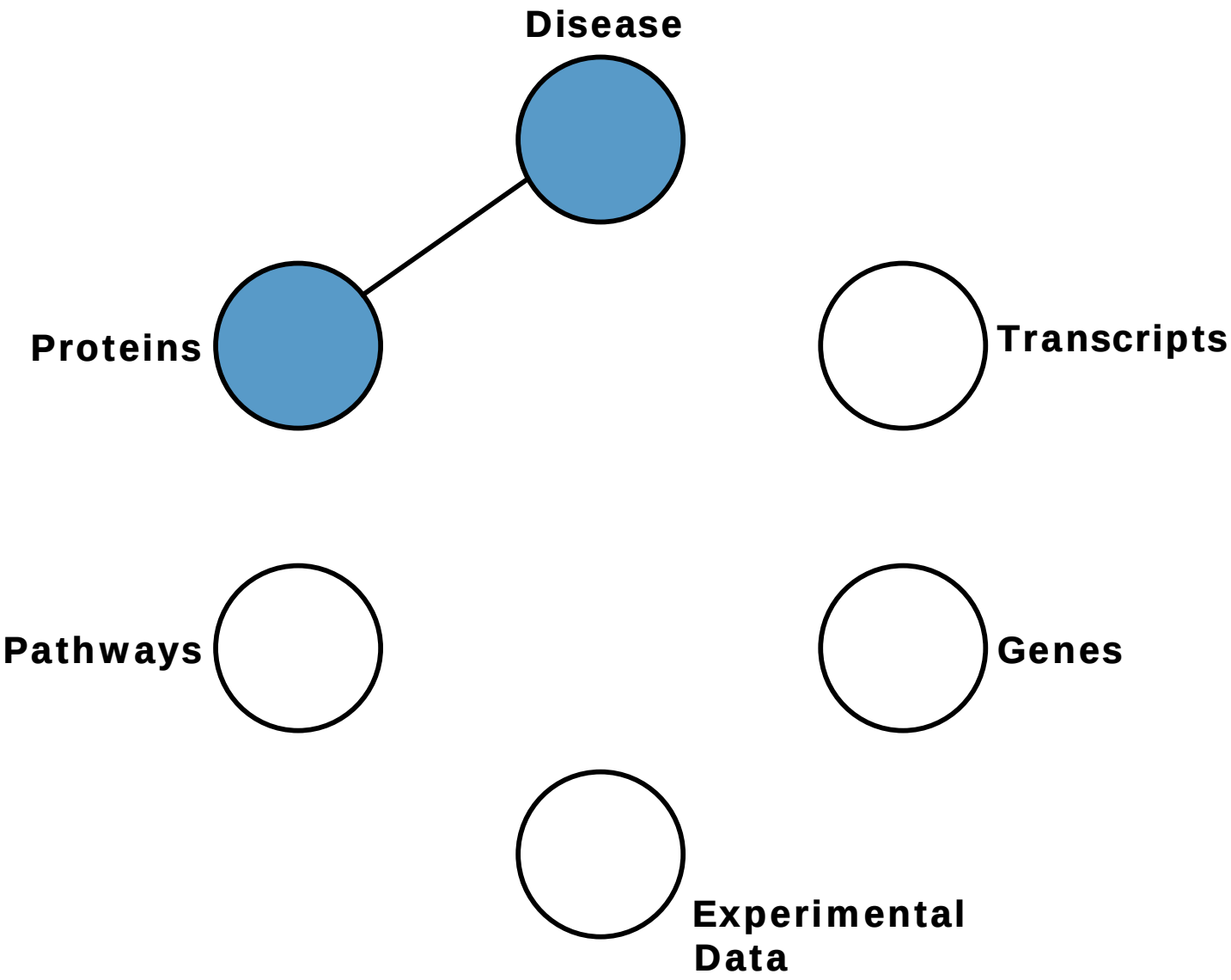
Integrate experimental data



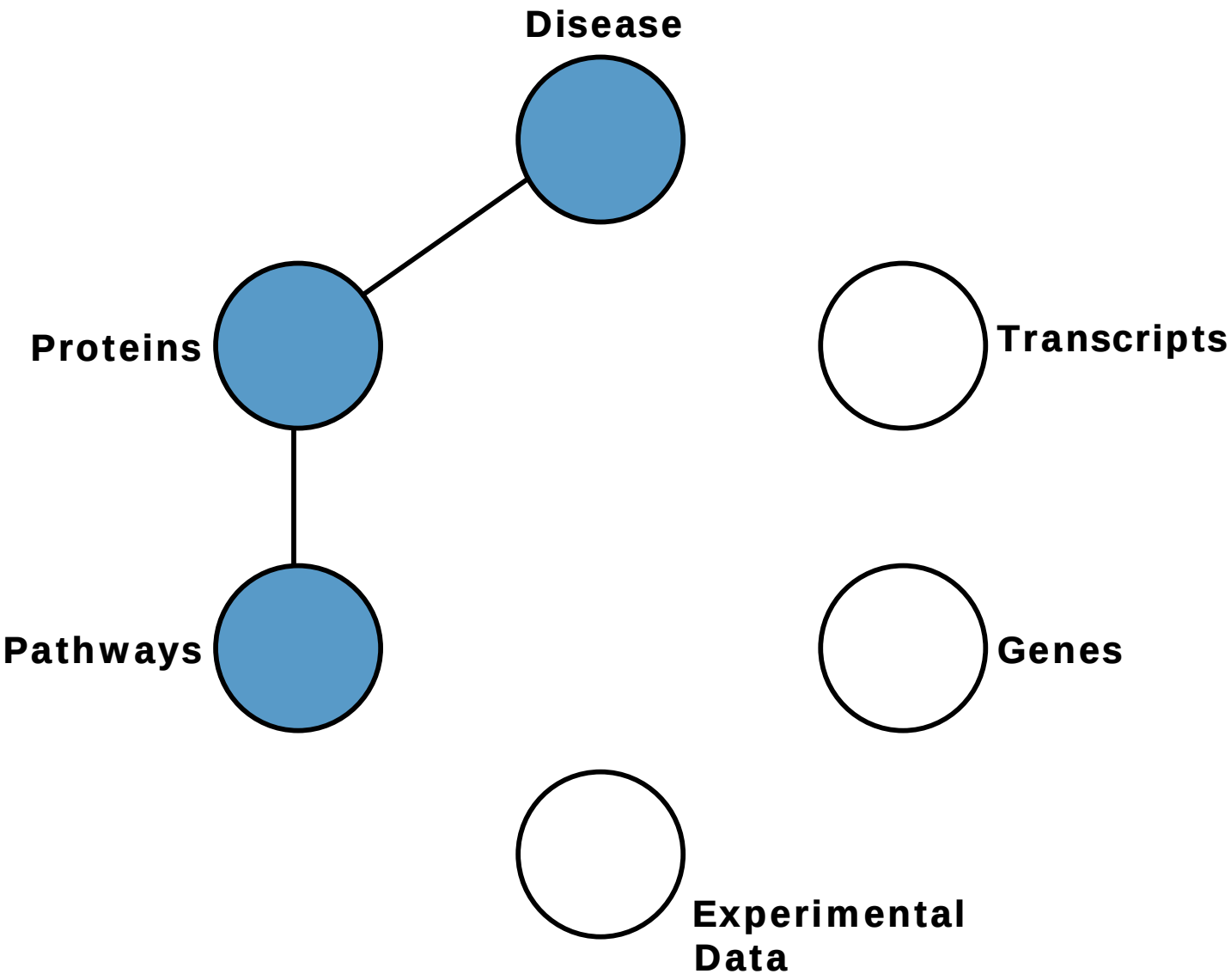
Target discovery arena – Semantic integration of biological data



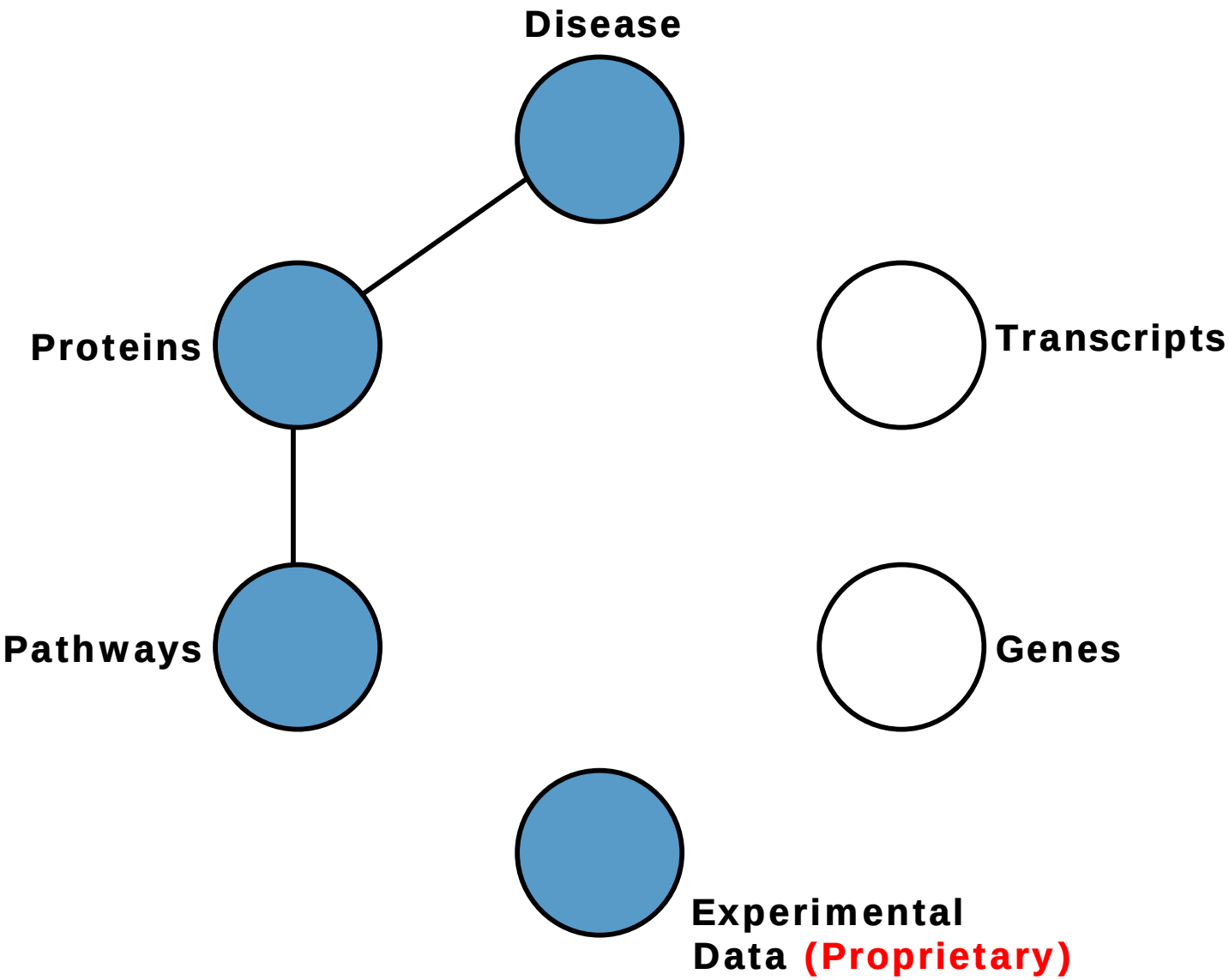
Target discovery arena – Semantic integration of biological data



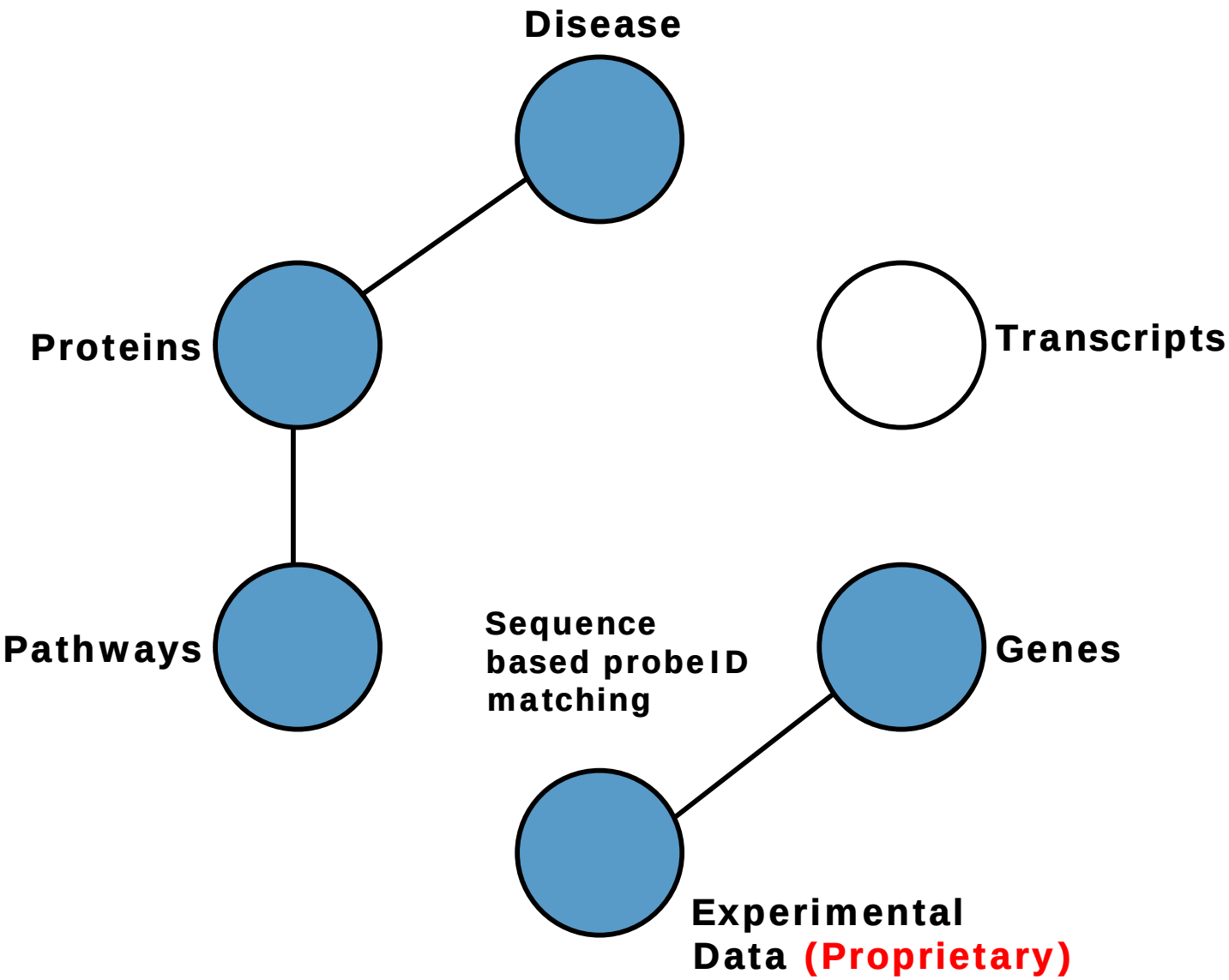
Target discovery arena – Semantic integration of biological data



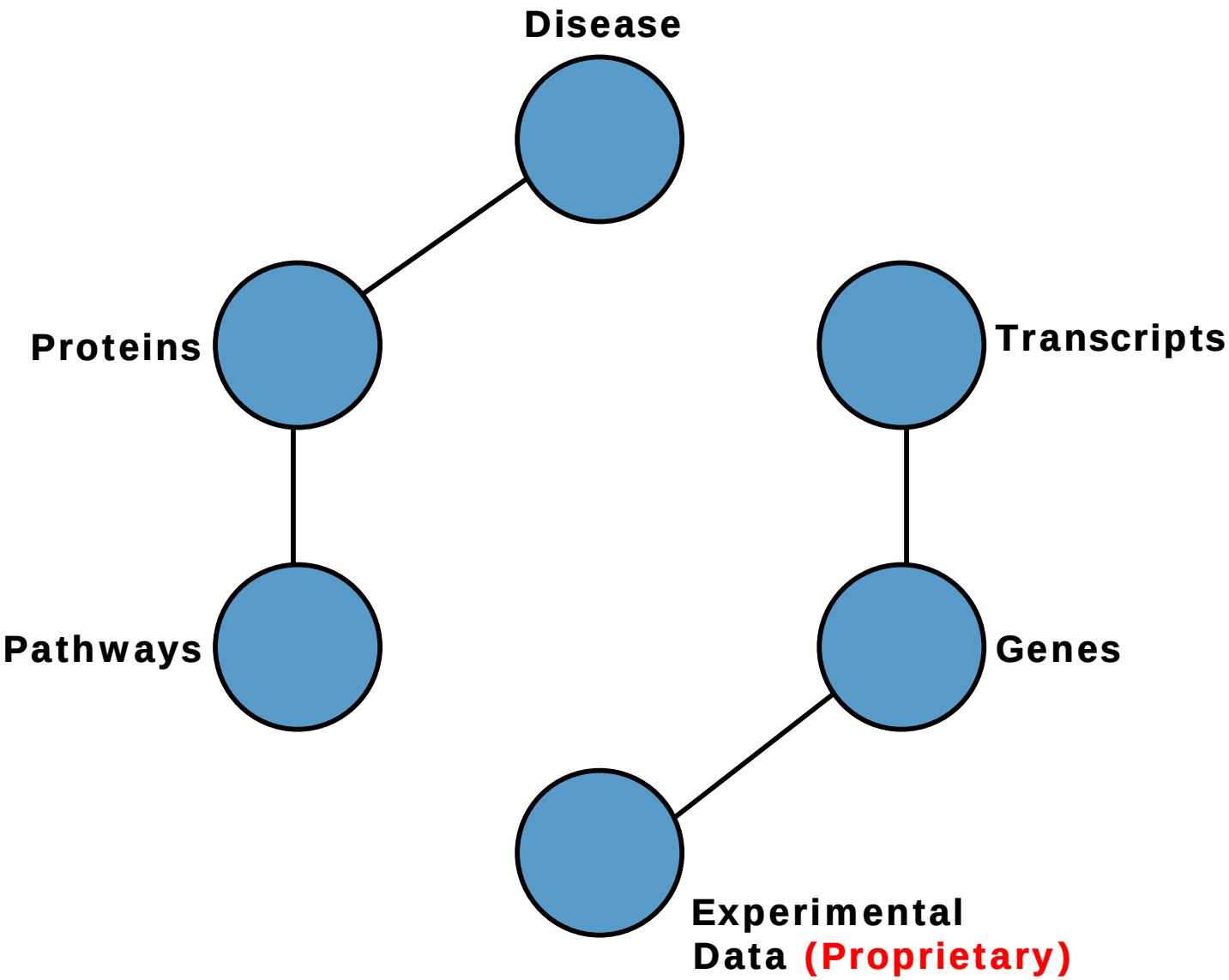
Target discovery arena – Semantic integration of biological data



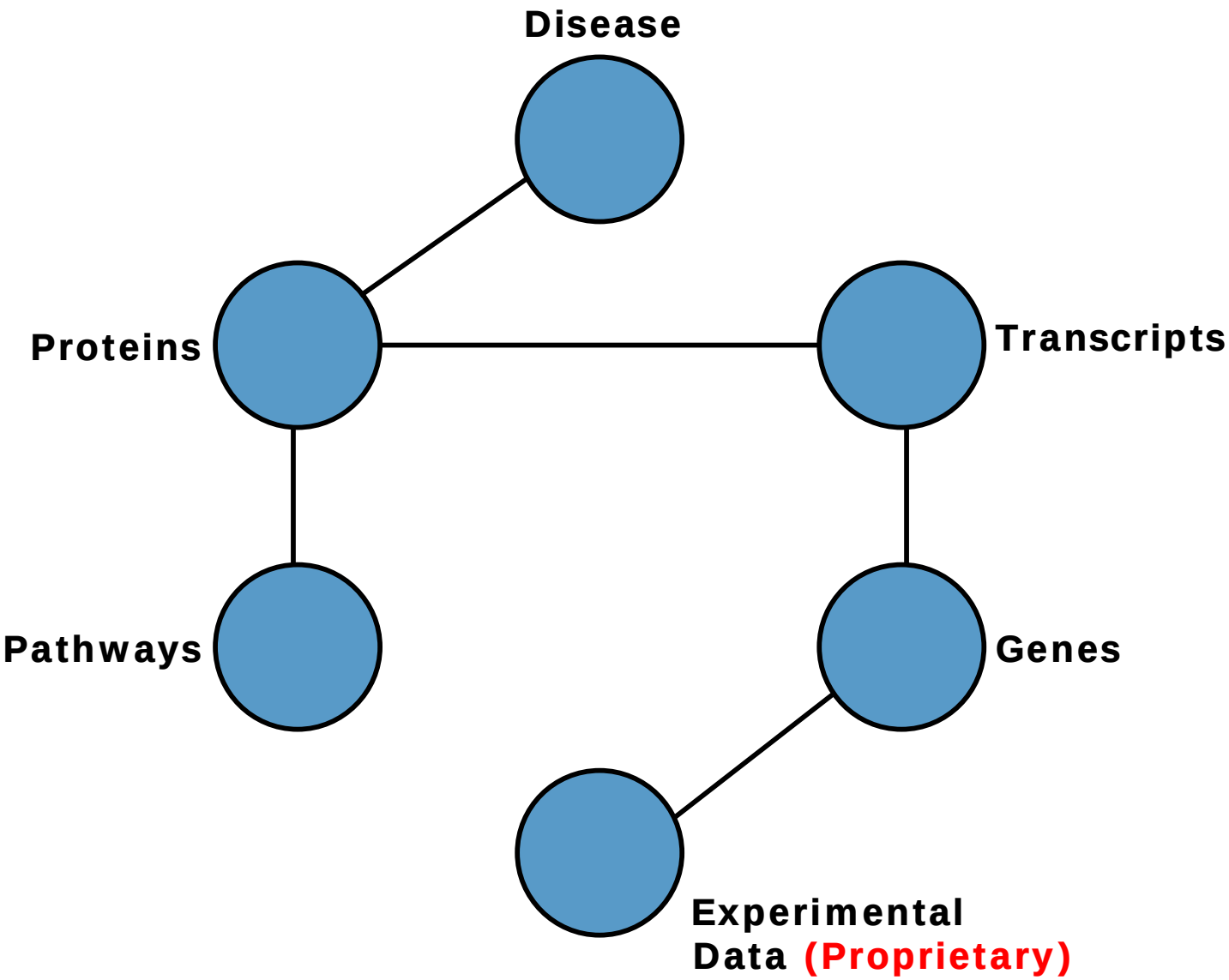
Target discovery arena – Semantic integration of biological data



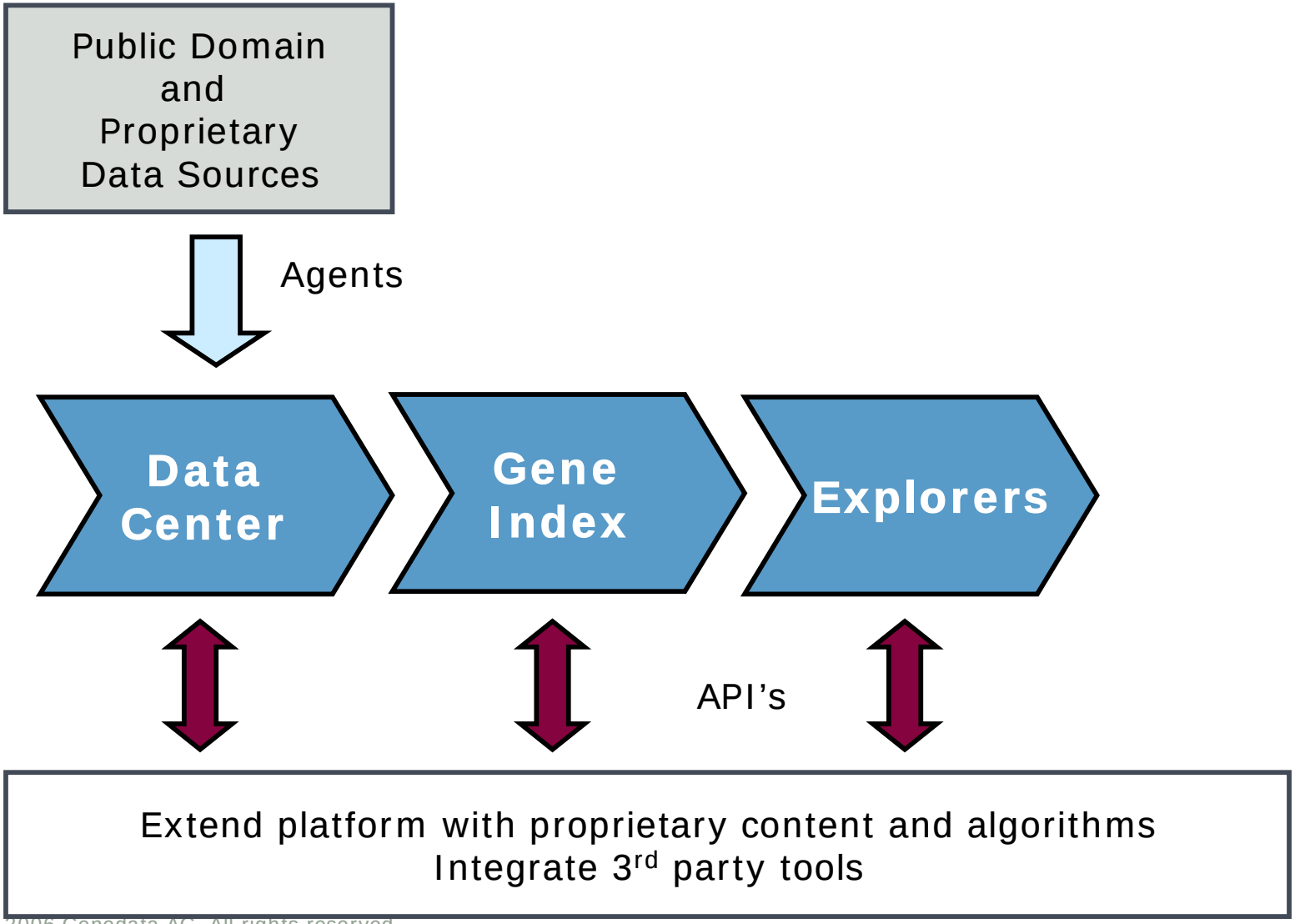
Target discovery arena – Semantic integration of biological data



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