diXA - a **d**ata **i**nfrastructure for chemical **sa**fety

Jos Kleinjans
Dept of Toxicogenomics
Maastricht University

Big data in toxicogenomics

- The amount of data in our world has been exploding. And the ability to store, aggregate and combine data, and then use the results to perform deep analyses has become ever more accessible – *McKinsey & Co (2011)*

- Acquisition of the knowledge to develop high-throughput testing assays would involve the discovery of toxicity pathways and networks from vast amounts of data… Central repositories for -omics data … exist to a small extent … The scale of data storage and access envisioned by the committee is much larger - *Toxicity Testing in the 21st Century (2007)*

- The vision is a scientific community that does not waste resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make *use* of data. Data become an infrastructure that scientists can use on their way to new frontiers – *A Digital Agenda for Europe (2010)*

- But digital information is inherently fragile and often at risk of loss. Access to valuable digital materials tomorrow depends upon preservation actions taken today; and, over time, access depends on ongoing and efficient allocation of resources to preservation - *Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010)*
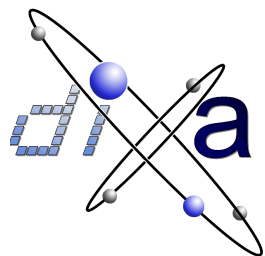
Main objectives of EU/FP7's diXa

- To further develop and adopt a robust and sustainable service infrastructure (e.g. data infrastructure and e-science environment) for harbouring multiplexed data sets as produced by past, current and future EU research projects on developing <u>non-animal tests</u> for predicting chemical safety as conducted by the research community of toxicogenomics

- To link this with other research communities maintaining globally available chemical/toxicological data bases and data bases on molecular data of human disease.

diXa's research hypothesis

- The main scientific concept which has initiated the diXa proposal, is
  - to create a large public data infrastructure of genomic signatures of drugs, industrial chemicals and cosmetics, and
  - to develop pattern-matching bioinformatics and biostatistics tools to detect similarities among these signatures,
  - in order to describe all biological states induced with a chemical exposure, in terms of genomic signatures relevant for the human situation *in vivo*.

Creating diXa's data infrastructure

* by integrating TGX data from FP6/FP7 projects
* by linking with chemical/mol. medicine data bases
* thus enabling cross-platform, cross-study integrations
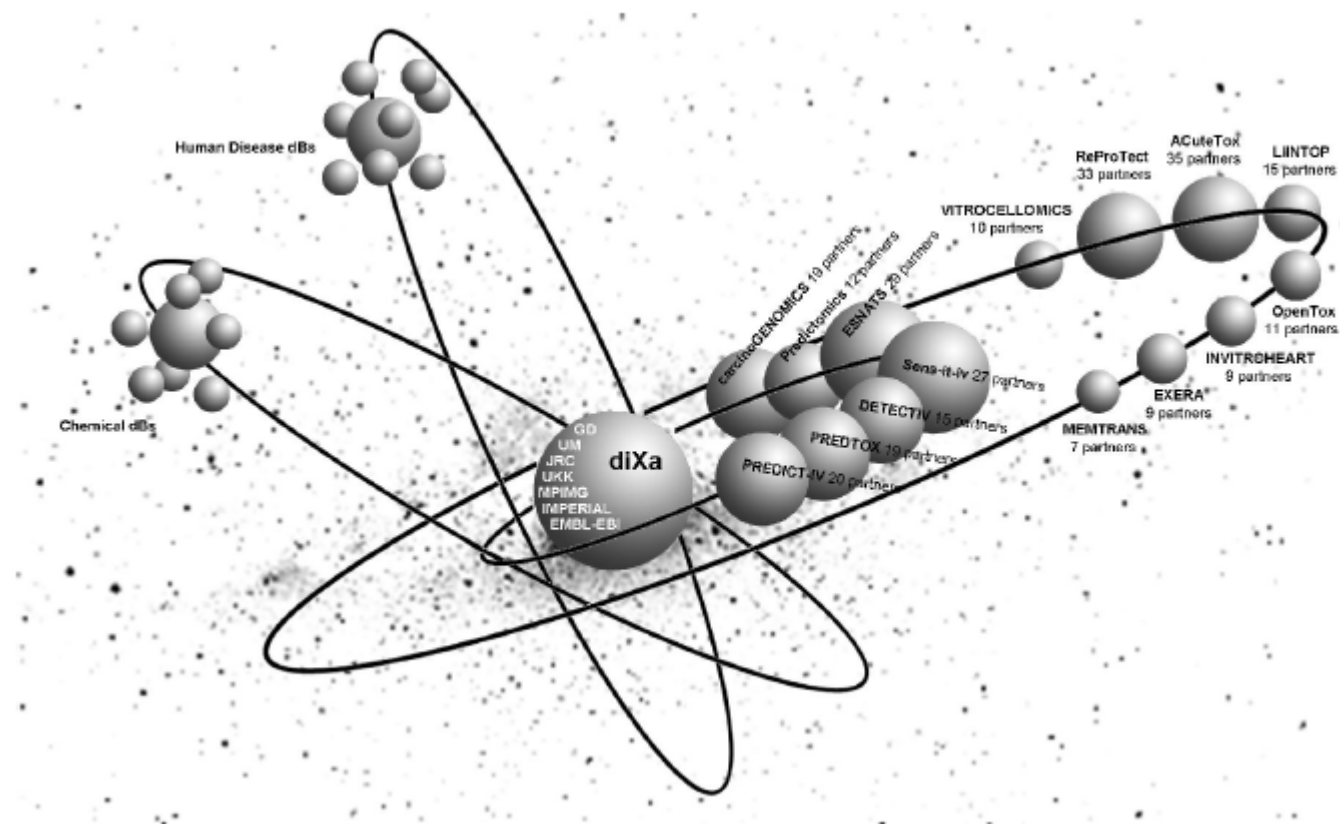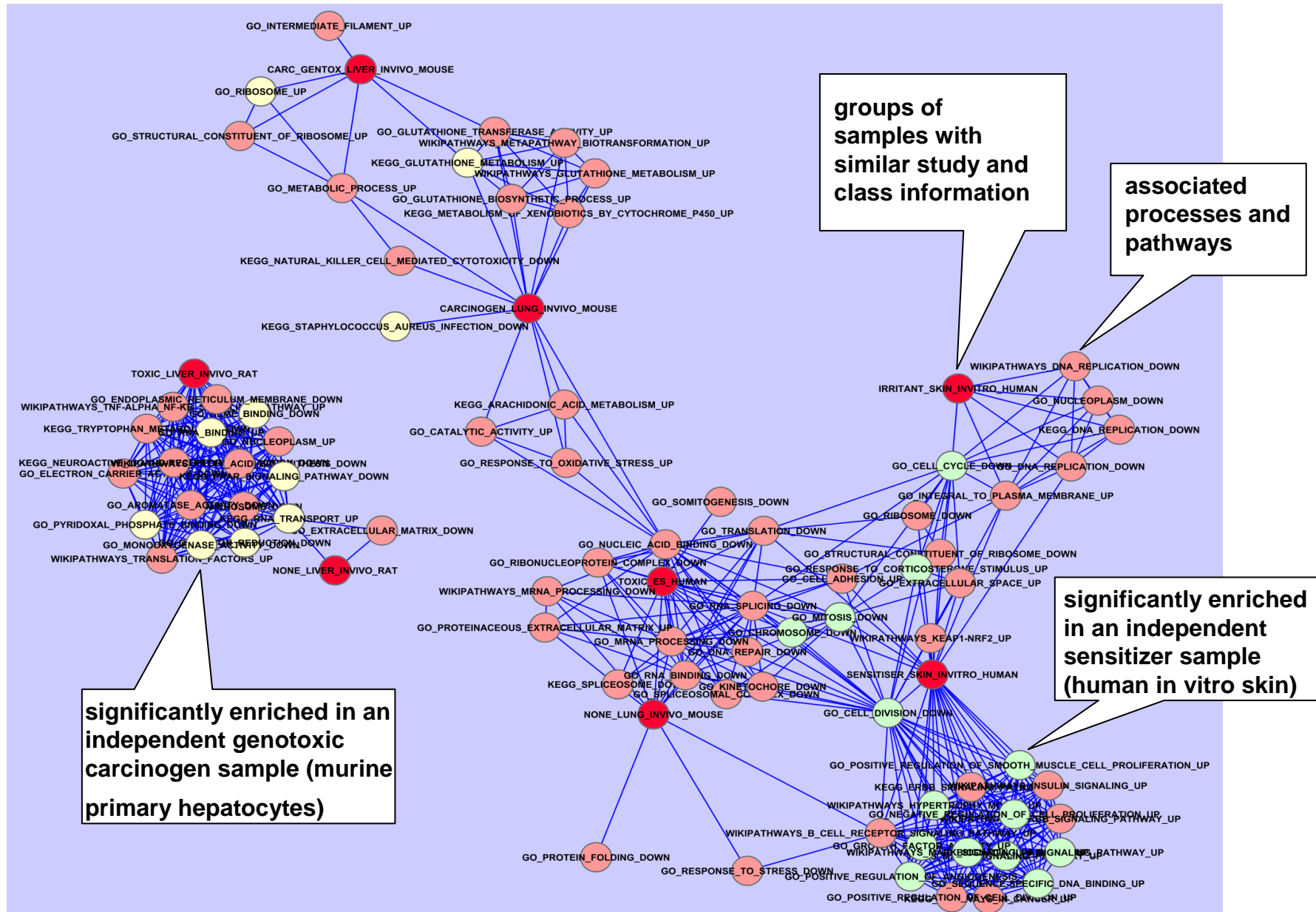
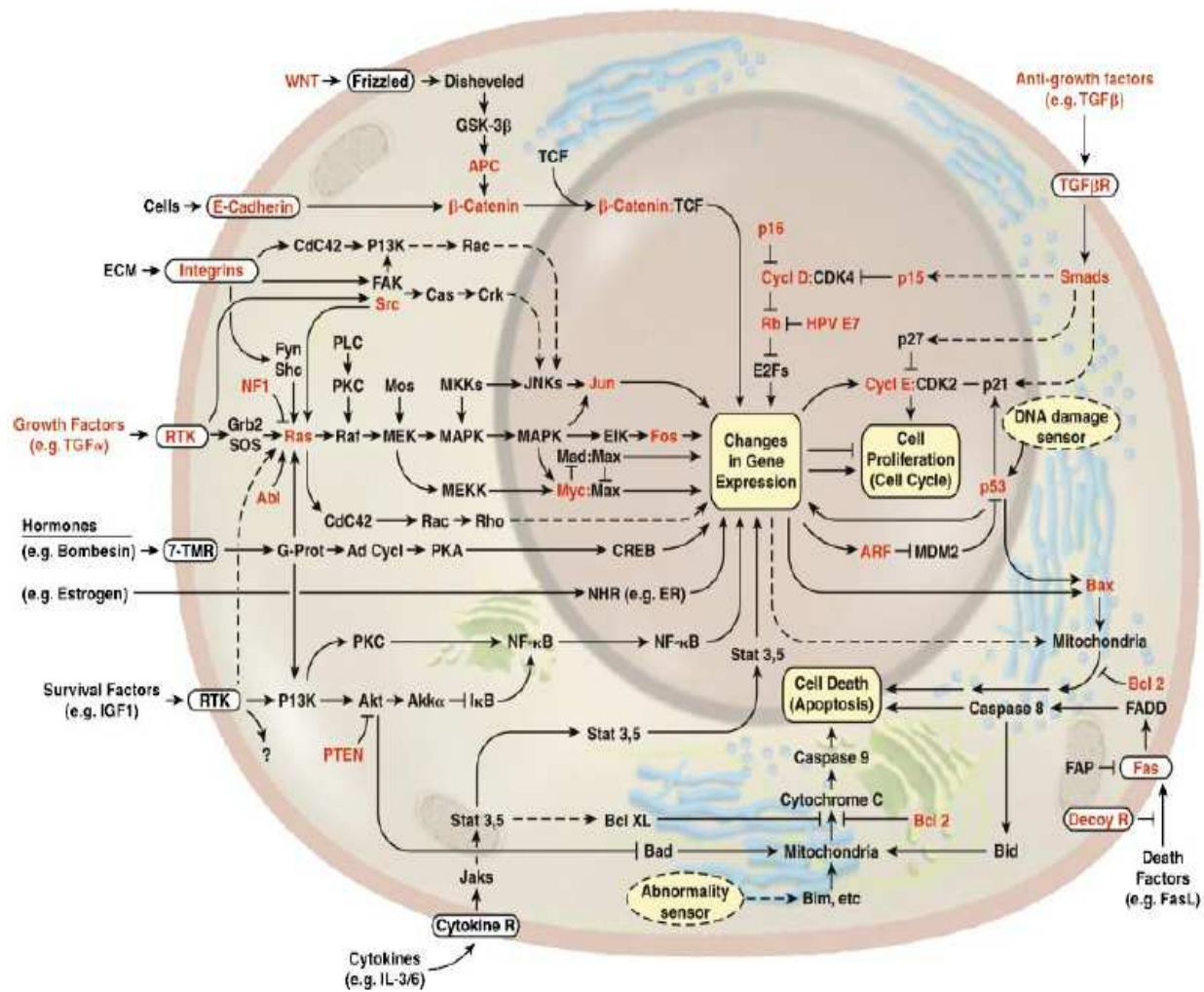# Full walk-through from data sources to data consumers
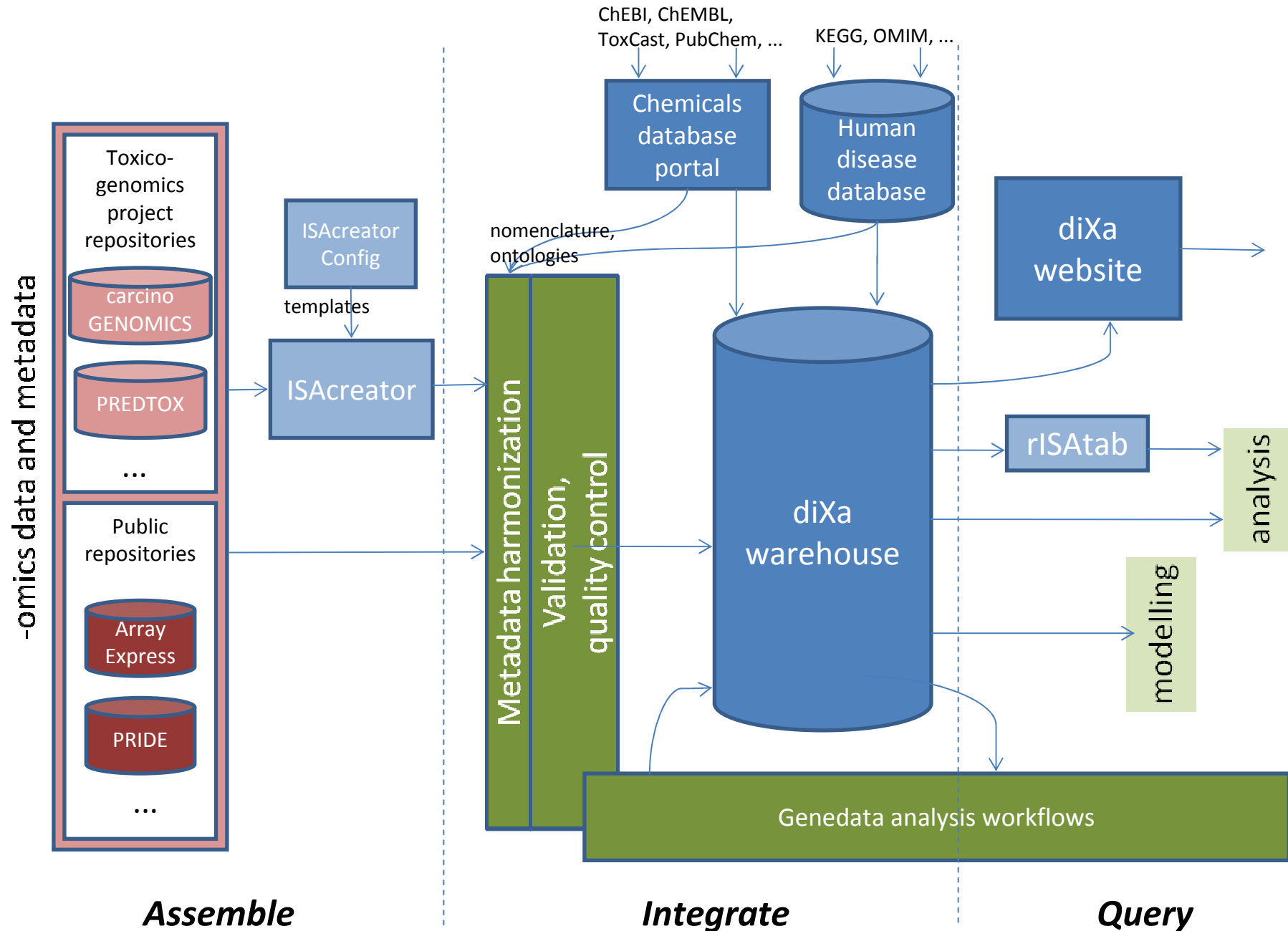
Pathway analysis

- A genetic pathway is the set of interactions occurring between a group of genes which depend on each other's individual functions in order to make the aggregate function of the network available to the cell – *Wikipedia*

- Pathway finding tools – and there are many – are built on existing pathway models and apply an algorithmic approach to identify most significantly perturbed pathways, to a given data set, and then use a statistical scoring mechanism

- Strongpoints are integration and visualization

- Problems relate to data consistency of reference dBs, and quality of underlying knowledge bases (chemical, toxicological, genomics, phenotypic, pathological)

- Relevant TGX reference dBs: CEBS, ArrayExpress, Comparative TGX dB, GEO

# NTC-wide associations between toxicity class, compounds and processes

groups of samples with similar study and class information

associated processes and pathways

significantly enriched in an independent sensitizer sample (human in vitro skin)

significantly enriched in an independent genotoxic carcinogen sample (murine primary hepatocytes)
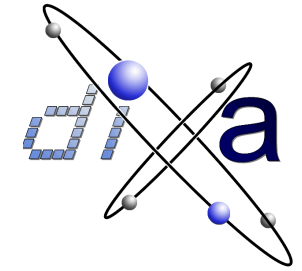
Model construction

Hanahan, Weinberg (2001/2011) Cell.

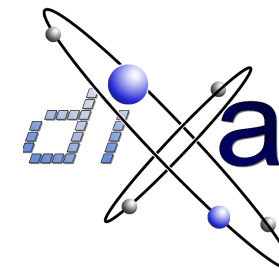# Visualization of the potential use of diXa's end-product
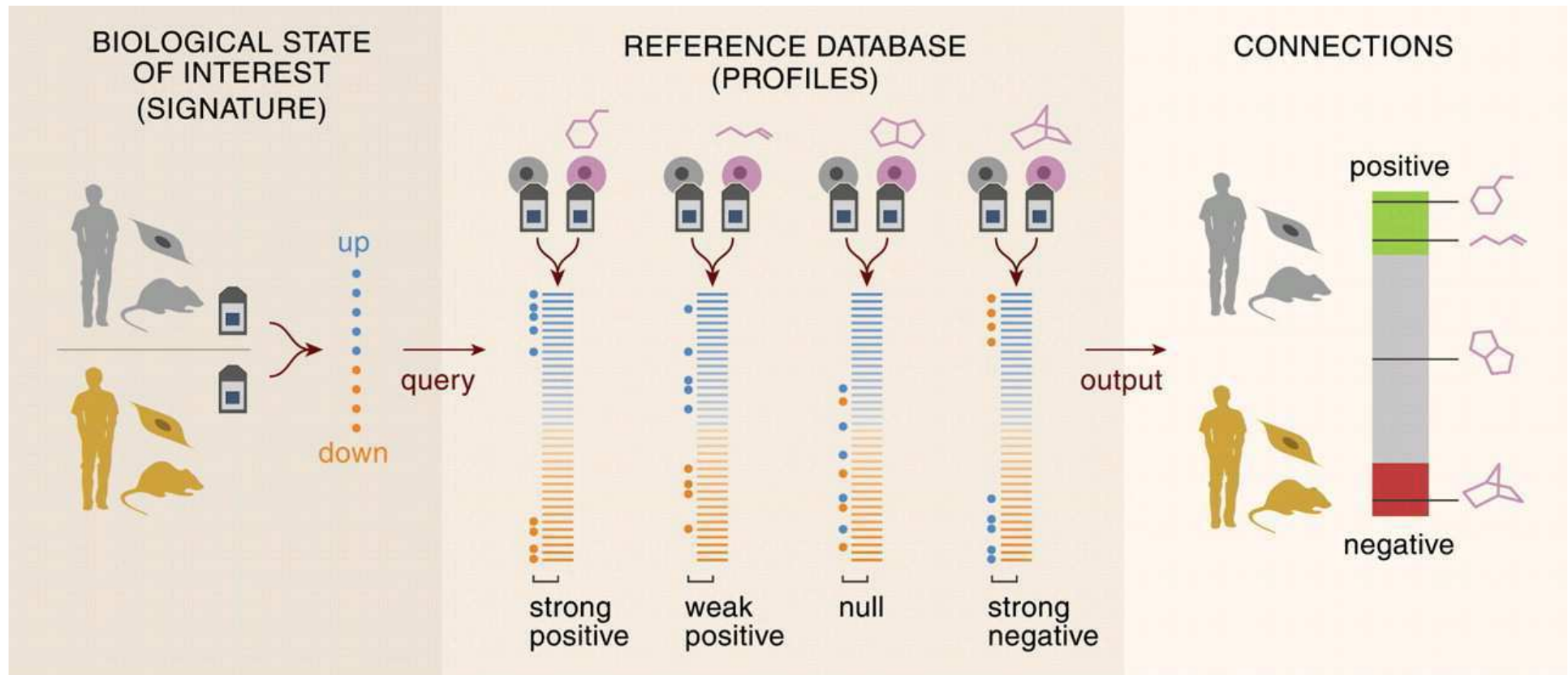
J. Lamb et al. The Connectivity Map: using gene-expression signatures …Science. 2006

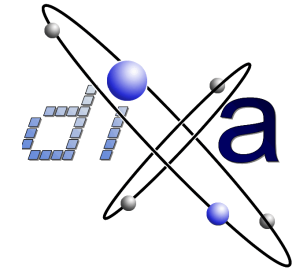**Pilot on perturbed pathways in liver toxicity**

Overview of CP-CSA activities

- Networking activities
  - fully functional, web-based, openly accessible and sustainable e-infrastructure for capturing toxicogenomics data from relevant EU FP6/FP7 projects
  - linking to available data bases holding chemico/physico/toxicological information
  - linking to data bases on molecular medicine
- Servicing support from this Collaborative Data Infrastructure
  - clear communication channels with the TGX research community
  - deliver commonly agreed core service support
    - by providing SOPs for seamless data sharing
    - by offering quality assessments and newly developed tools and techniques for data management
    - by offering access to the diXA infrastructure for toxicogenomics scientists
  - all supported by hands-on training.
- Joint research initiative
  - cross-platform integrative statistical analyses
  - cross-study meta-analyses
  - systems modeling for predicting chemical safety as alternative to animal tests
- Thus contributing to the vision expressed by Riding the Wave and Europe's Digital Agenda

the diXa consortium

Partner nr	Partner	Expertise
1	UM	Coordination/management Data infrastructure Content Data analysis
2	EMBL-EBI	Data infrastructure Content Dissemination
3	MPIMG	Data analysis. modelling
4	GD	Data infrastructure Servicing
5	ICL	Data infrastructure Data analysis, modeling
6	JRC	Data infrastructure Content
7	UKK	Content

THANK YOU !!

www.diXa-fp7.eu