atragene
Research Informatics

Data Integration and Decision-Making For Biomarkers Discovery, Validation and Evaluation

D. POLVERARI, CTO
October 06-07 2008

© atragene 2008

Data integration definition and aims

Definition :

Data integration consists in combining multiple sources of information to be viewed as one

Aims :

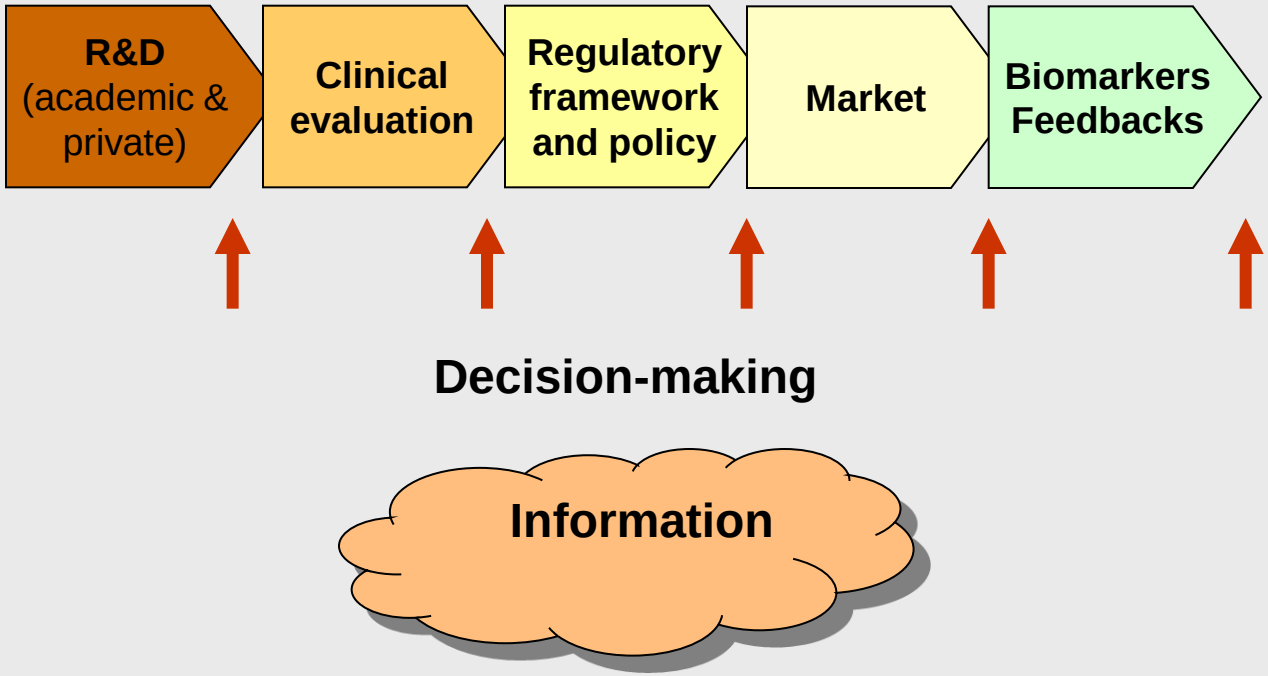
To Make data easily available for :

- Sharing
- Analysis and indicators computation
- Visualization

Data integration aims to reduce decision-Making step duration

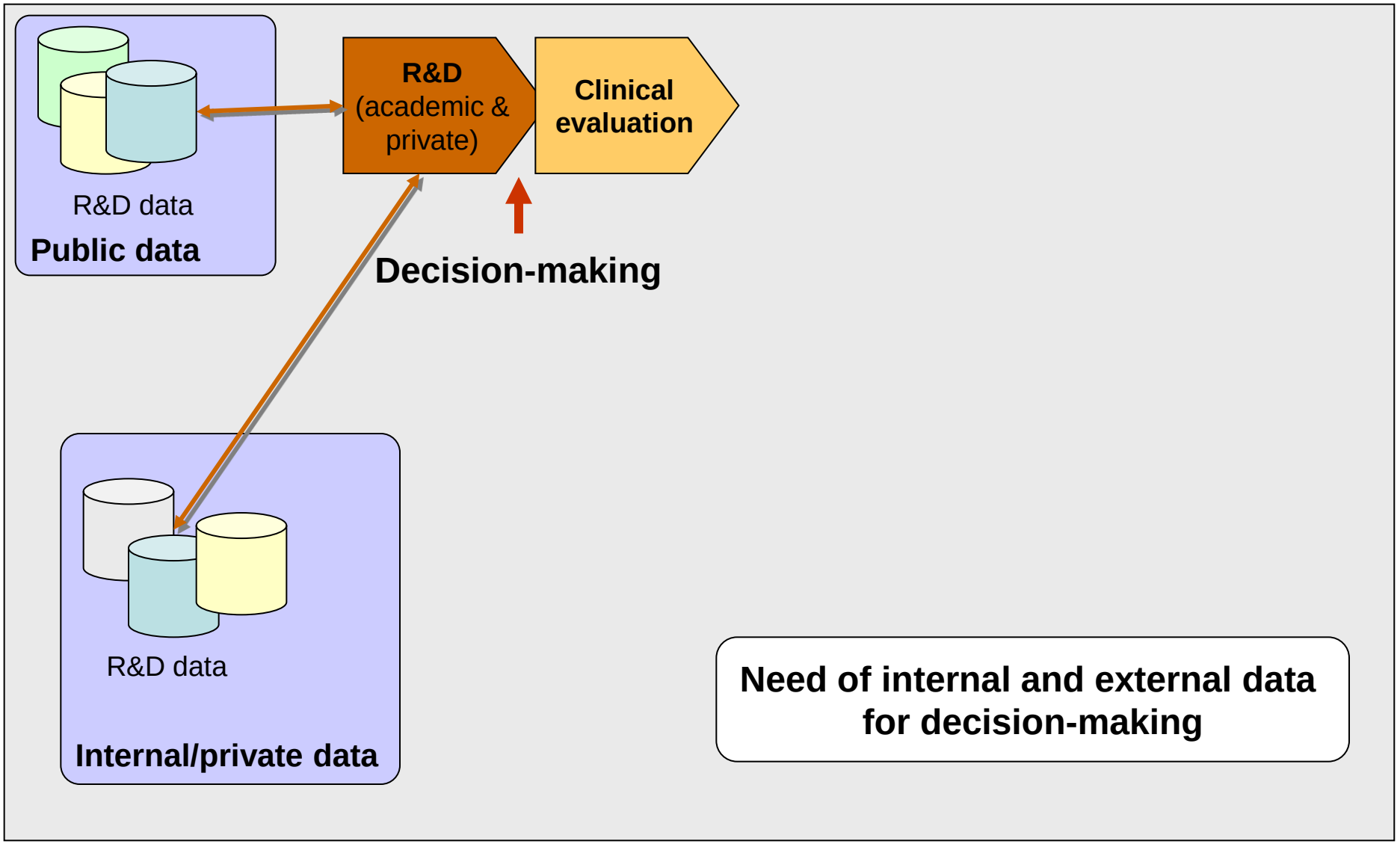
© atragene 2008

Biomarkers business chain : from R&D to market dataflow



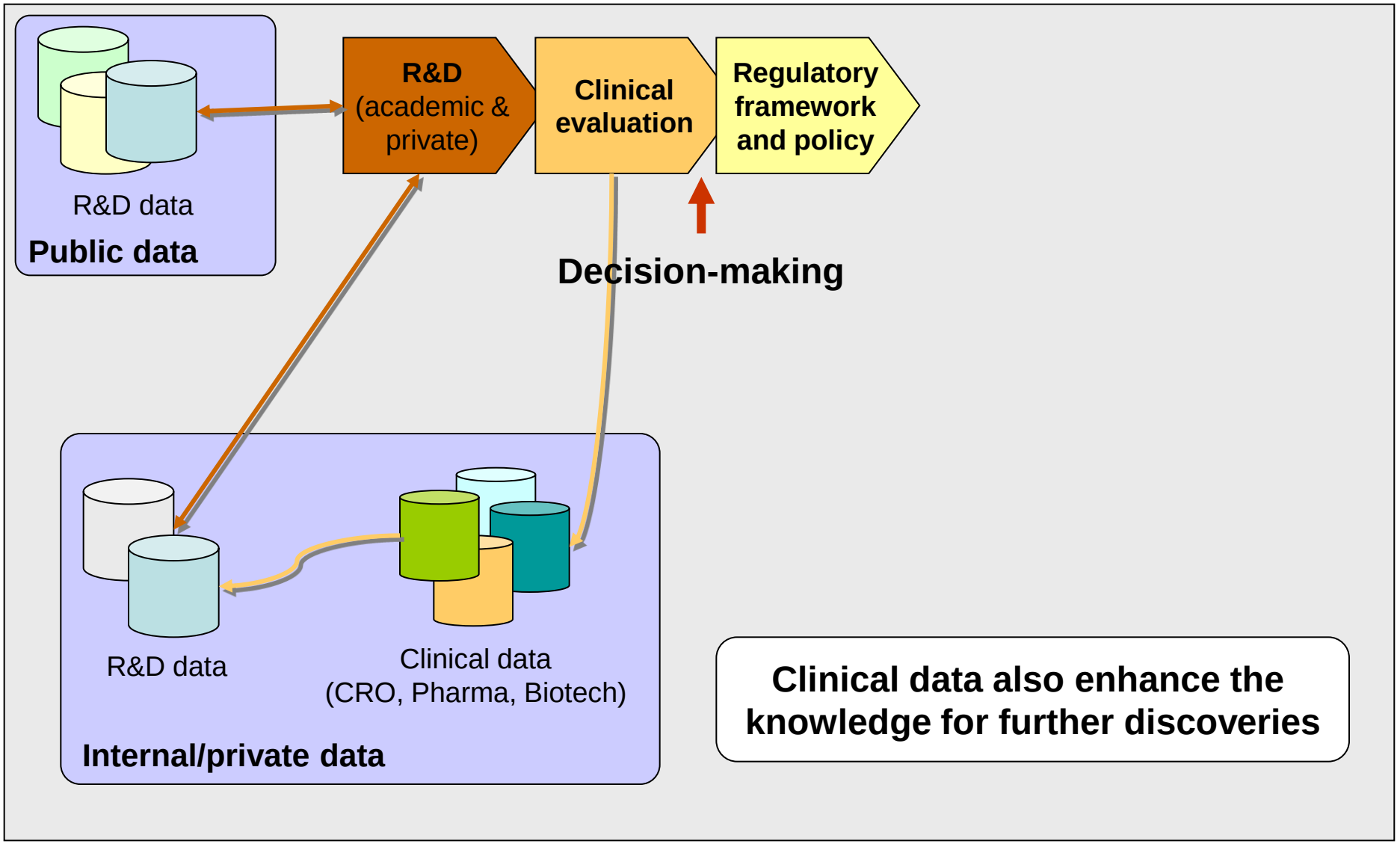
© atragene 2008

Biomarkers business chain : from R&D to market dataflow



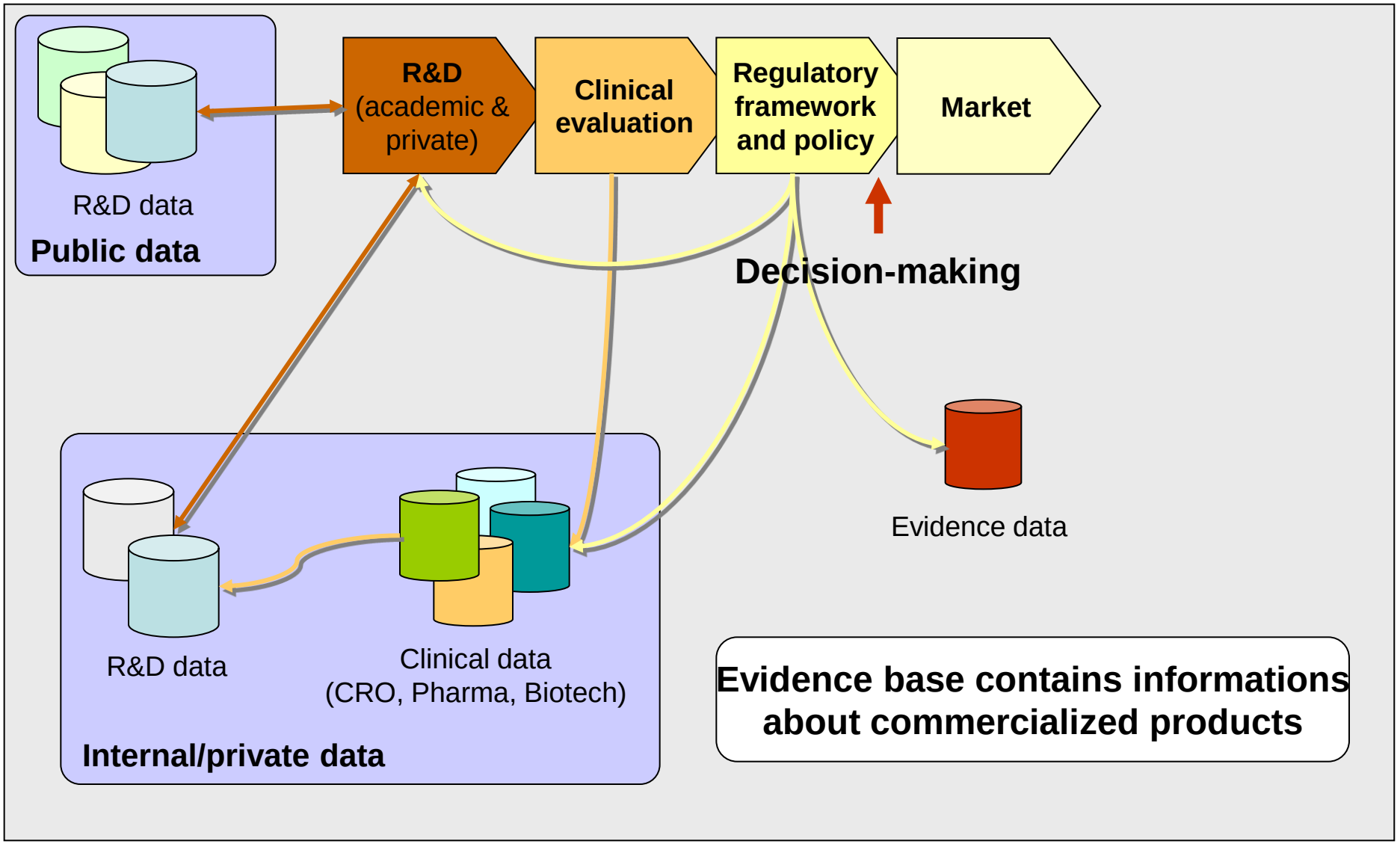
© atragene 2008

Biomarkers business chain : from R&D to market dataflow



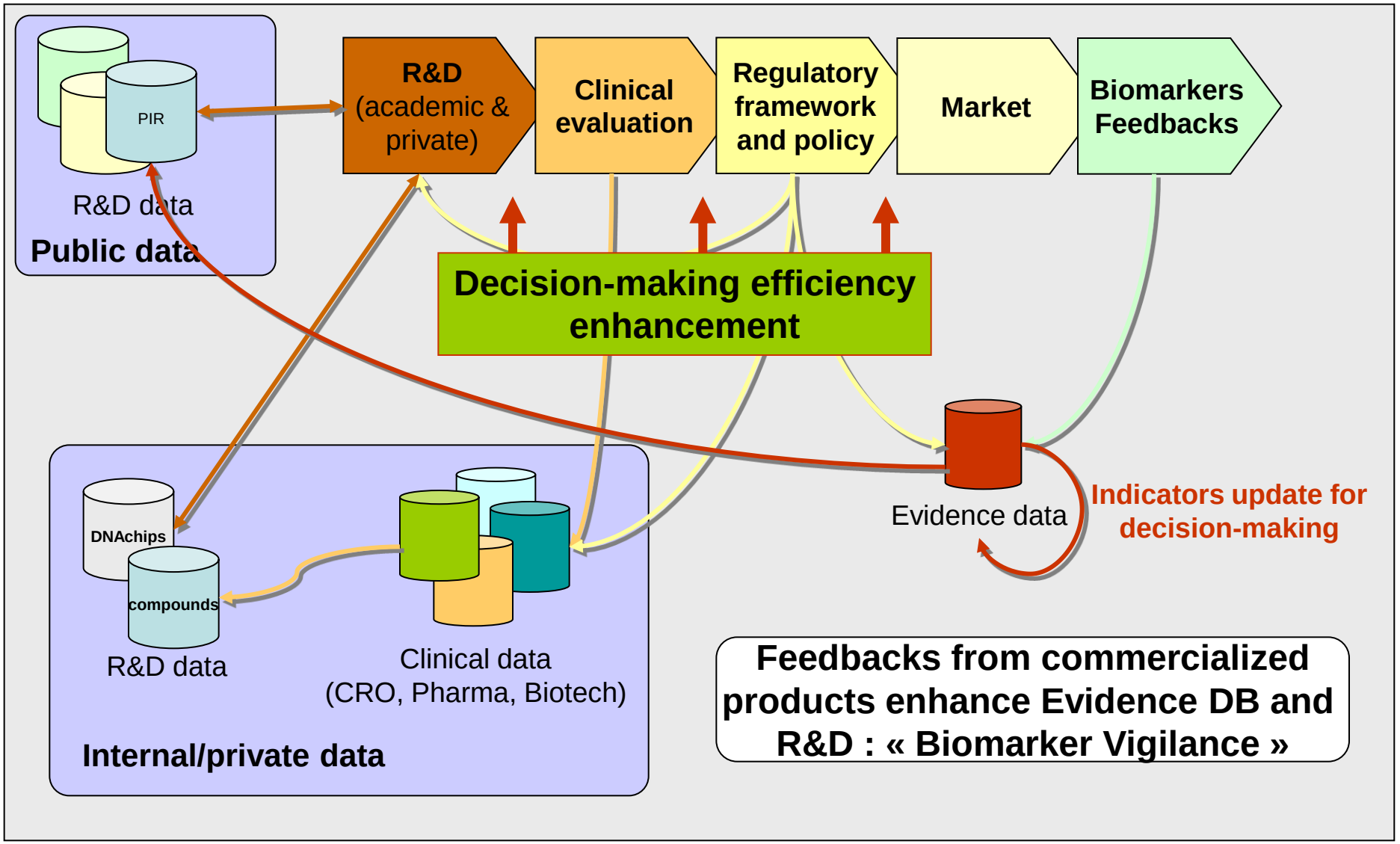
© atragene 2008

Biomarkers business chain : from R&D to market dataflow



© atragene 2008

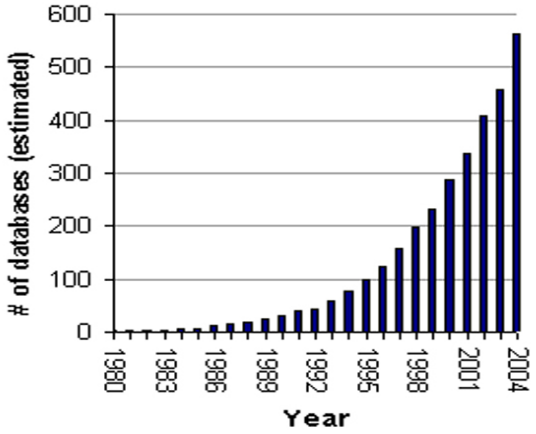
Biomarkers business chain : from R&D to market dataflow



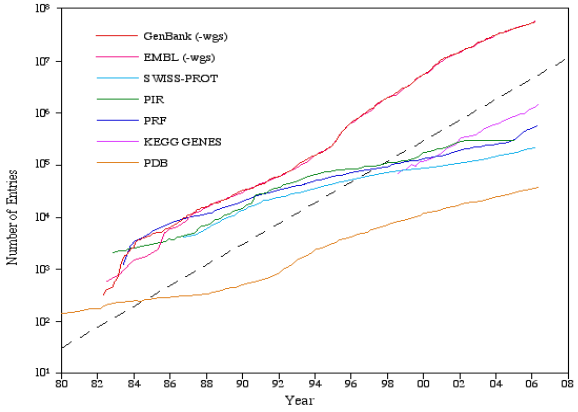
© atragene 2008

Data integration issues

Data volume and sources diversity (more than 1000 sources only in R&D)



Number of Database published in Medline
Wren et al. BMC Bioinformatics 2005 6(Suppl 2):S2



Entries growth in several life-science databases
http://www.genome.ad.jp

High frequency data update

Data quality (error rates measures, curation workflow)

Data heterogeneity

© atragene 2008

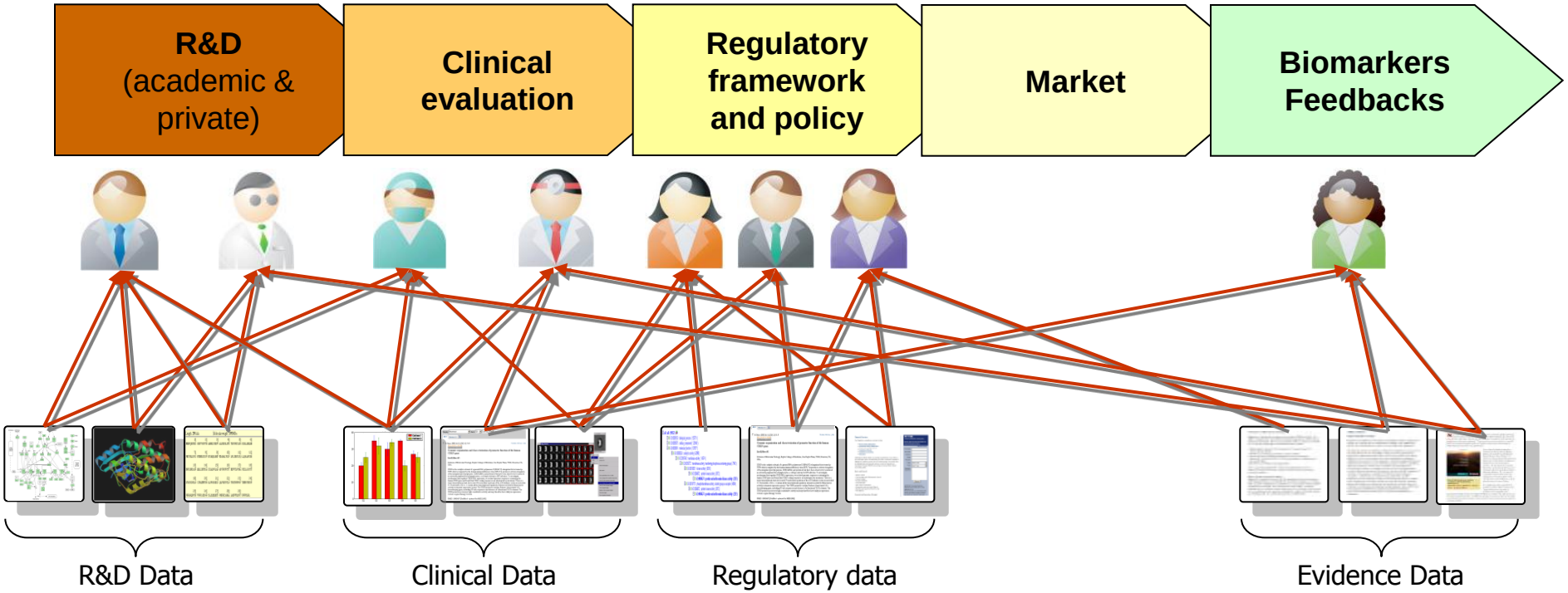
Data integration issues

Data heterogeneity :

- Many concepts : genes, proteins, diseases, pathways, Social, financial, ethical …,
- Many data formats (objects databases, relational databases, Text files, XML, …),
- Terms diversity (eg. gene names synonymy => ATP2B1 = PMCA1),
- Semantic diversity (eg. Population genetics : gene means allele),
- Syntax diversity (eg. Pr, prot, proteins, …)

© atragene 2008

Data integration issues



Each user is different !

Each one needs common and specifics data in the decision making process.

© atragene 2008

atragene
Research Informatics

How to access and share data to enhance decision-making efficiency ?

© atragene 2008

Integration strategies : data warehouse

Build Model : description of data relations and semantic

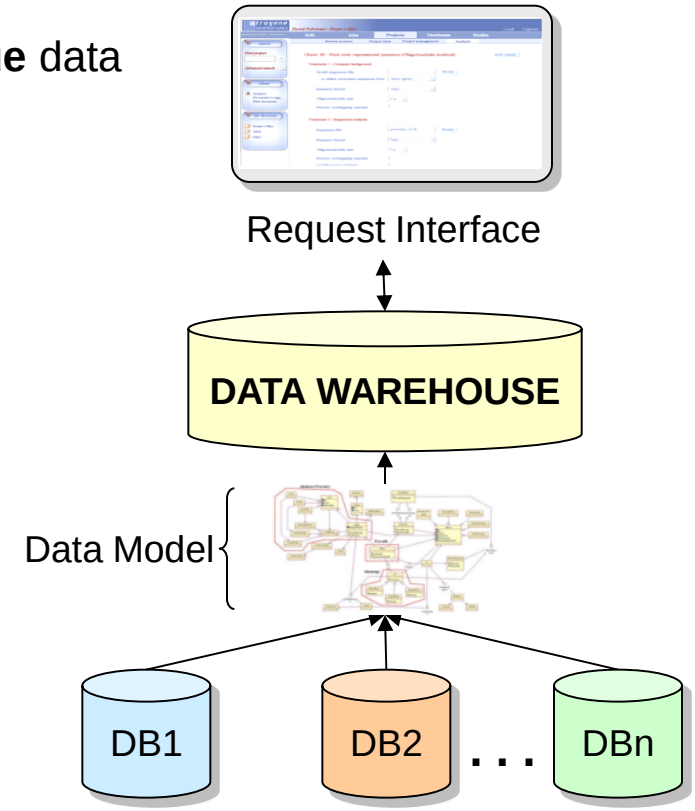
All the data are extracted, transformed, loaded in this **unique** data model

Data Warehouse main advantages :

- Centralized data storage and maintenance,
- Data extraction is simplified.

Data Warehouse main drawbacks :

- Model complexity
- Model weakness
- Available queries limited by designed model (Sql technology)



Ex : Archimed (hôpital universitaire de Genève)

Data Warehouses aren't new, they used to be found in financial and industrial fields for years

© atragene 2008

Example of datawarehouse Integration in a french bank

- Only several million customers
- A model containing about 100 tables
- 10 years old and about 20 major model updates implying ca. 1 to 3 months activity stop each
- Query limited to simple extraction, or predefined queries
- Analysis studies can’t be done directly through the dataware
- New analysis projects (commercial efficiency or customer comportemantal study) need at least ten to fifteen persons for one year
- Datawarehouses are now used for storage purpose

Data Warehouses strategy is restricted to well-known data with no/few concepts and model evolution

© atragene 2008

Integration strategies : databases federation

Sources used as Multiple autonomous remote databases

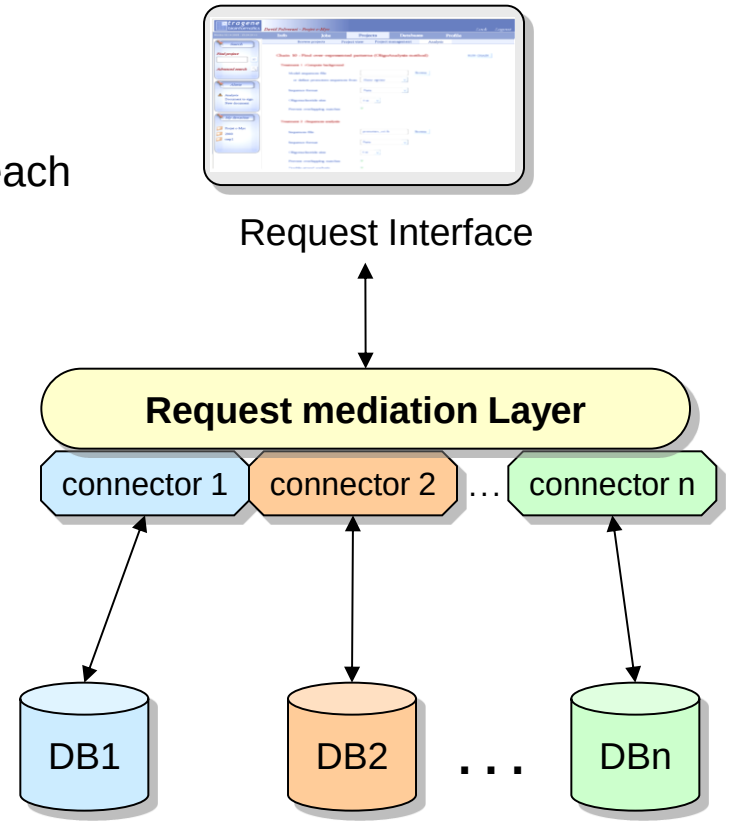
User queries are rewritten in sub-queries in the language of each data source by the mediator

Federation main advantages :

- Data sources stay under each team responsibility
- Data sources stay independent (Remote or locally)
- In theory, a new source = a new connector
- User can import his own data

Federation main drawbacks :

- Queries are still limited by mediator semantic
- Need developers to extend either connectors or mediator
- Queries may be time consuming



Ex : SRS, Tropical Biominer, etc.

© atragene 2008

Exemple of federated databases : SRS©, BioWidsom Ltd

- Sequence Retrieval System, a web based application dedicated to sequence data integration

- Lot of connectors available for life sciences data sources

- A user-friendly query interface

- Creating new connectors implies good development skills

- User queries are imposed by the mediator which cannot be modified

- Integration of your own data is tricky

- need to handle multiple technologies to extend base capabilities

- Not optimized for large volume queries

DBs Federation is really more adapted than datawarehouse but still too rigid to face heterogeneous data and user specificity

© atragene 2008

Integration strategies : Decison support system (DSS)

DSS are rapidly growing in industrial and financial field

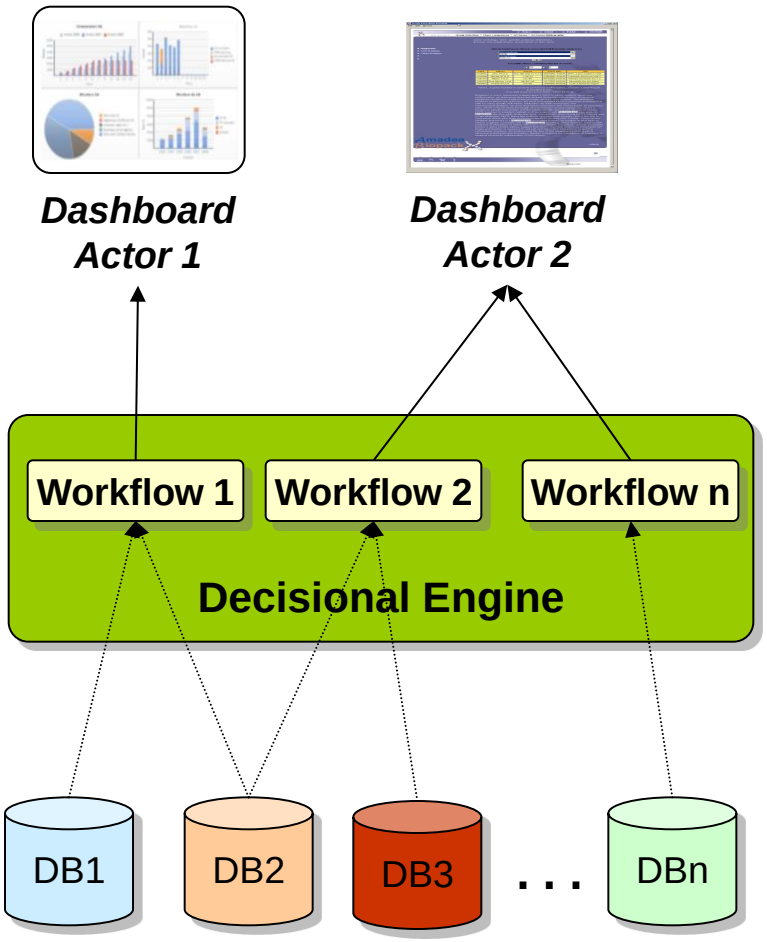
Like federative databases the data sources stay independent

The mediation layer is replaced by the decisional layer :

- A workflow designer
- User and general workflows (connectors and treatment),
- Workflow execution engine

DSS main advantages :

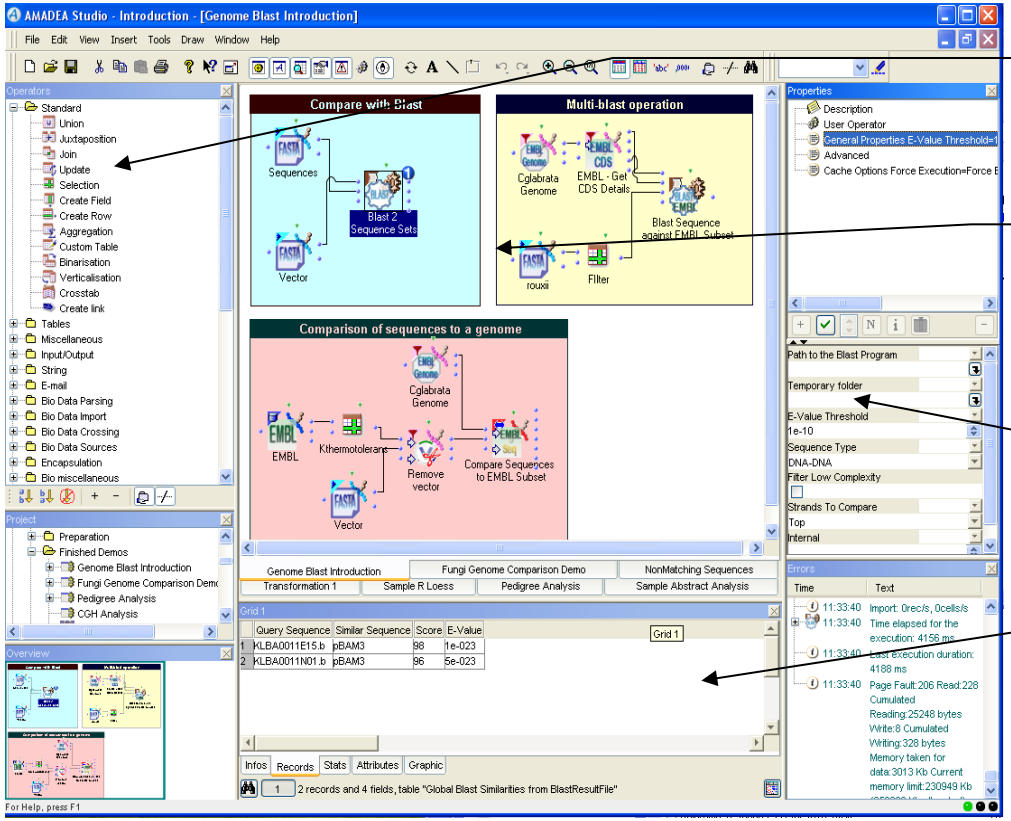
- Graphically created workflow,
- Experts don't need informatician to design solution,
- Integration time is reduced by 3 to 30 times
- Optimized for massive amounts of data,
- Now, Mediators can be user specific
- No semantic model (Data Heap)



DSS platform for Biomarkers

© atragene 2008

DSS interface example : Amadea™



connectors and operators list

Complex workflow built by simple connector/operator drag and drop steps

Connectors / Operator parameters

Real-Time visualization of workflow modification (even with milion records)

With the authorization of Dr. Cécile Fairhead from Pasteur Institute, Paris

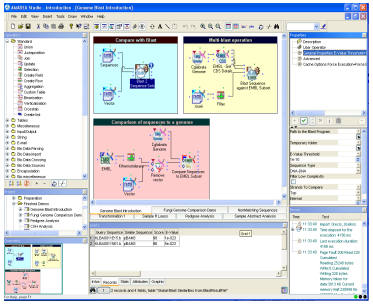
DSS allows Business experts to explore their data and to create integration workflows alone in few clics !

© atragene 2008

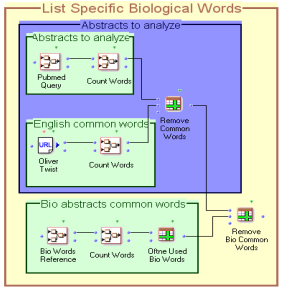
HKIS example

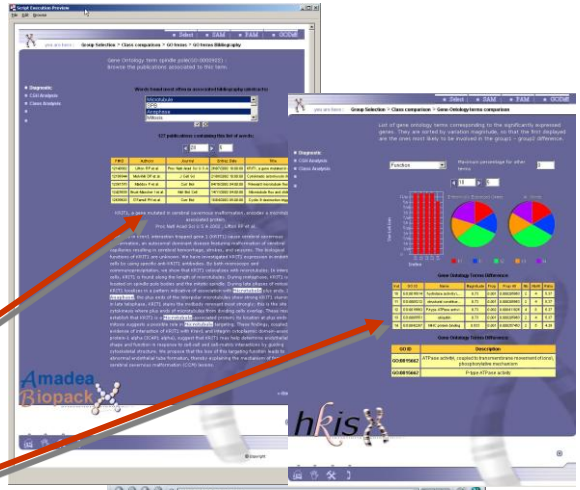
HKIS : a European **R&D** and **clinical** network on cancer (Fr, Ger, Ita) using data from genomics, proteomics, bibliographics, clinical trials etc.

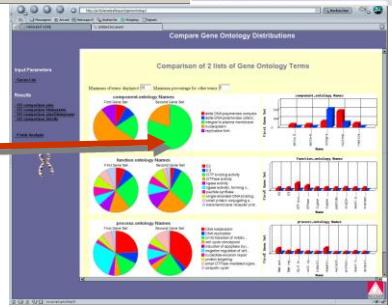
c.a. few weeks from data integration to HKIS platform deployment



Data integration and analysis wokflows design







Web deployment to share data and analysis workflow

DECISION

© atragene 2008

Recommandations

- Data providers side:
  - Data have to stay stored in **independent databases** maintained, enriched and verified by each responsible team,
  - Data providers have to structure their databanks based on an **existing ontology**,
  - **Links between data** in other databases should be available if they exist,
  - **Data must be downloadable** in a readable format without using a query interface,

- DSS side:
  - DSS must make data available, **whatever their model and semantic are**, in order to allow users to integrate easily new data resources,
  - DSS Workflows should be created **without coding skills** (Graphically edited) and allow easy integration of new analytics tools,
  - **Analytical workflows** must be **easily shared** and used on-line via web portals,
  - Results visualization and indicators have to be simple and **adapted to the user** discipline (researchers, jurists, etc)

© atragene 2008

atragene
Research Informatics

Contacts & Informations :

contact@atragene.com

Tel. +33 (0)1 77 01 80 65

Fax : +33 (0)1 45 21 17 35

© atragene 2008

Integration strategies : Decison support system (DSS)

DSS are rapidly growing in industrial and financial field

Like federative databases the data sources stay independent

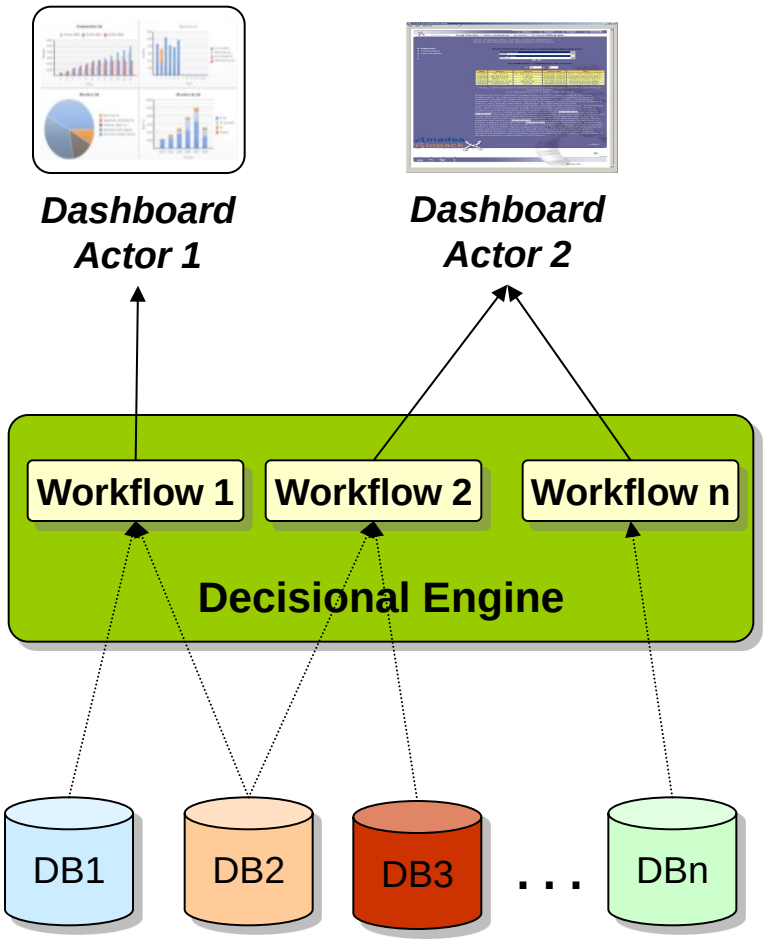
The mediation layer is replaced by the decisional engine :

-**Specific data** integration,

-Analysis workflows **“on-the-fly”** design

DSS main advantages :

- Graphically created workflow,
- Experts don’t need informatician to design solution,
- Integration time is reduced by 3 to 30 times
- Optimized for massive amounts of data,
- Now, Mediators can be user specific
- No semantic model (Data Heap)



DSS platform for Biomarkers

© atragene 2008