A large-scale chemical data integration system

Gaia Paolini

Summary

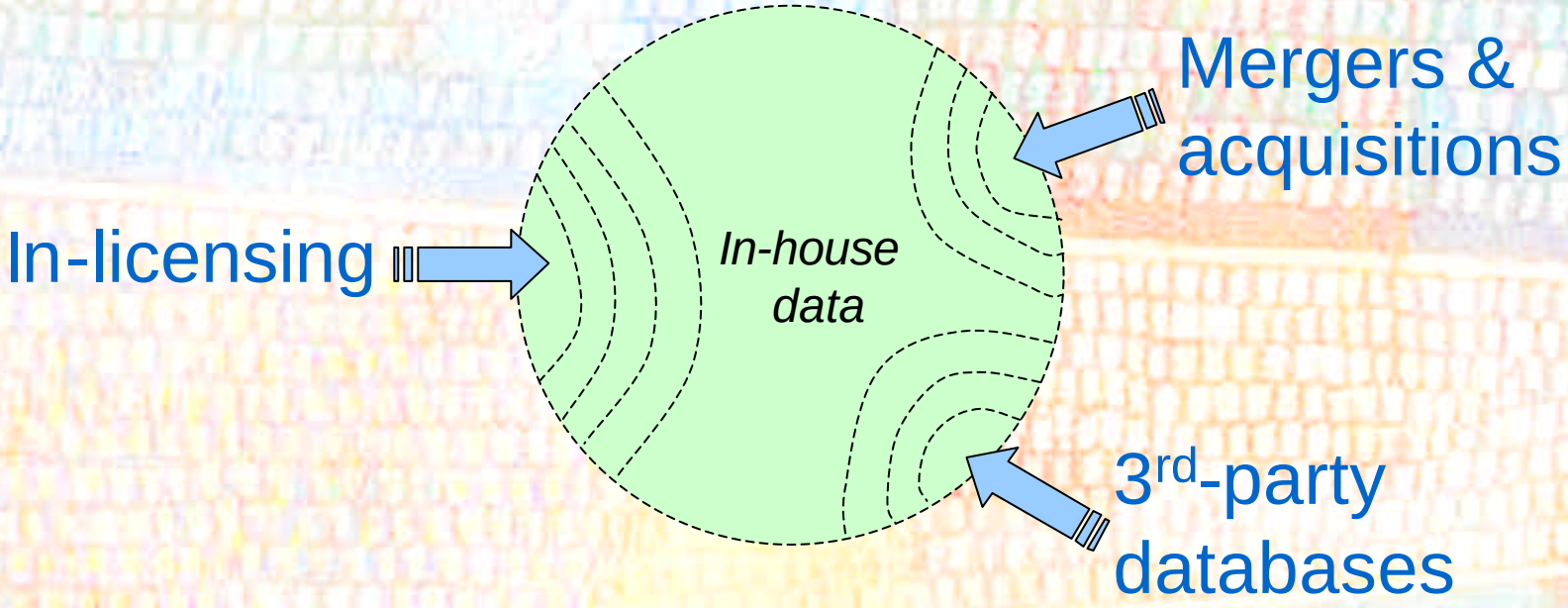
- Current situation
- Business case
- Aims
- The design process
- Functionality
- Applications

The Project

- A large chemical data warehouse to store and integrate Pfizer and third-party information using chemical structure as the natural entry point
- Millions of chemical structures
- Based on the DayCart Oracle Cartridge

Why Integrate?

- The need to integrate and mine disparate sources of data



Why Integrate?

- Data available to buy and integrate from external sources
- Need for active chemoinformatics research repository
- Opportunity to highlight connections
  - Chemical Properties
  - Structural similarities

Aims of the Data Warehouse

- Enable chemical/pharmaceutical data mining and knowledge discovery
- Store chemical structures and properties together with related entities
  - Biology
  - Portfolio
  - Inventory

Scope

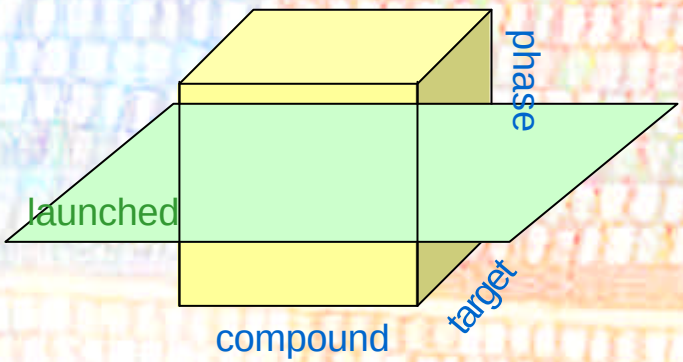
- Data warehouse
- Common consolidated set of data
- Repository of *selected fields* from Pfizer and third-party data
- Source independent
- Chemo-centric : indexed on structure not compound ID
- Emphasis on data integration rather than front end client application

Requirements

- Unique chemical structure indexing
  - Multiple and hierarchical tautomeric and stereochemical indexing
- Integrate internal and external data
  - Indexed by chemical structure
- Integrate chemo- and bio-informatics communities
  - Fit-for-purpose model architecture
  - Uses corporate dictionaries to standardise entities
  - Create connections and synonym tables

What do we want from our data?

- Data should be easy to
  - access
  - compare
  - exchange
  - manipulate



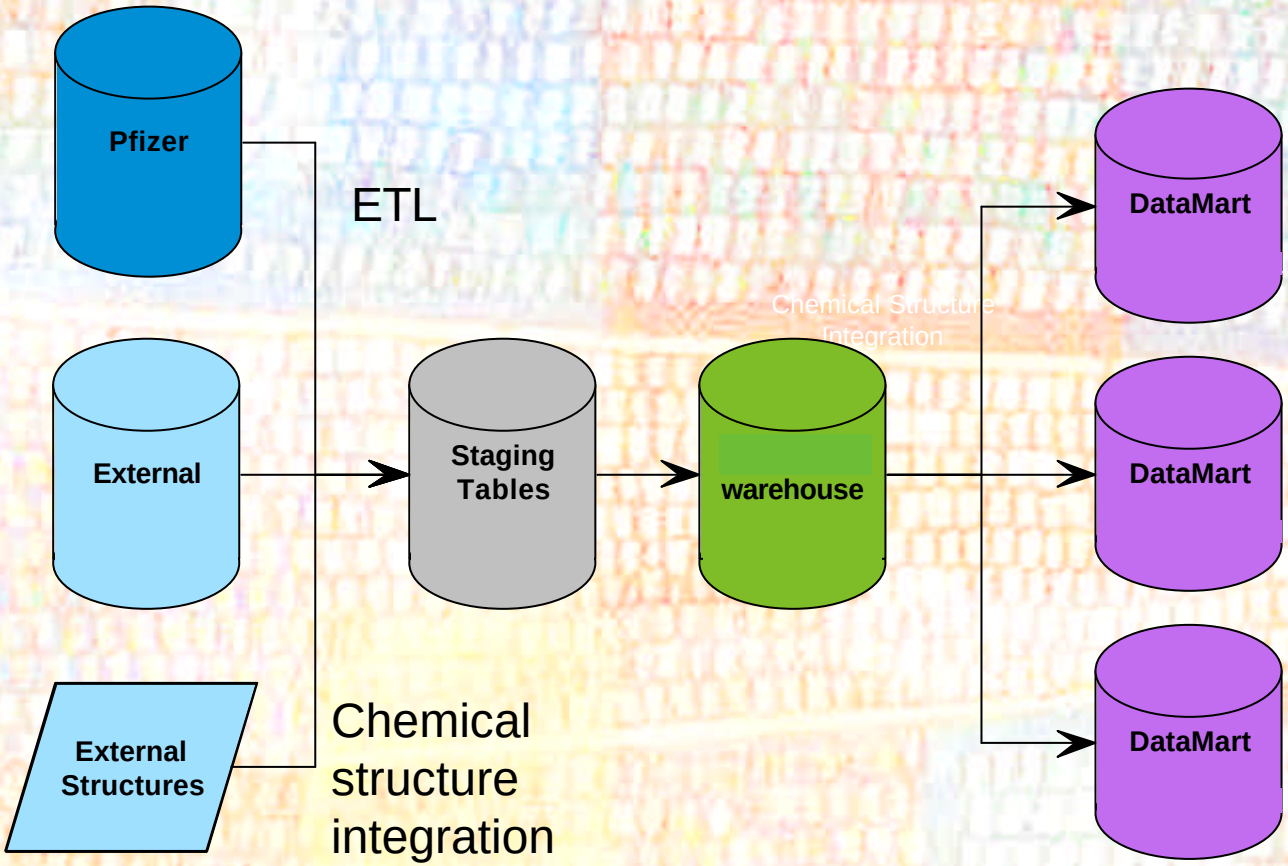
Why data integration?

Database Design Decisions

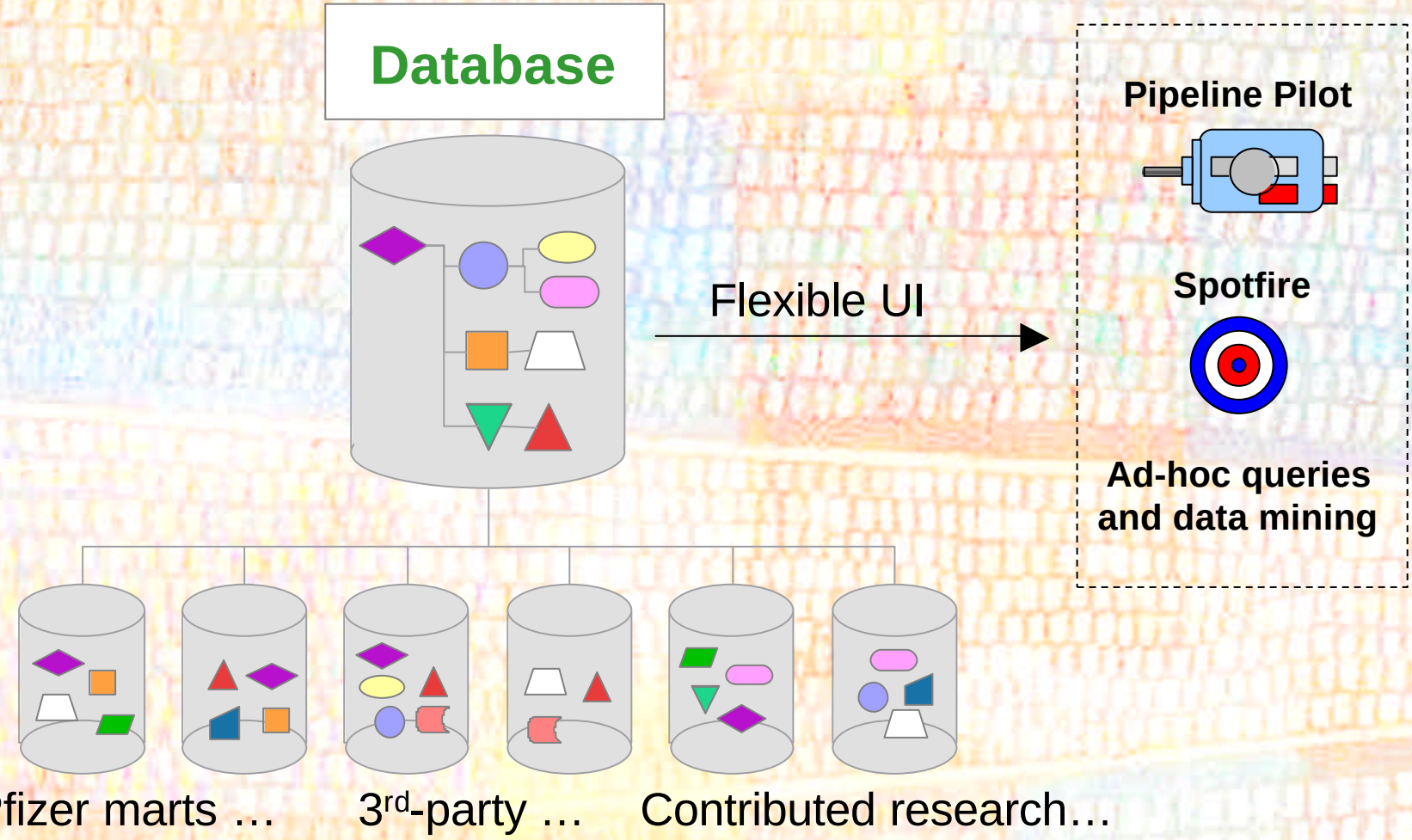
- Central data warehouse
- Selective data integration
- Focus on chemical structure
- SMILES representation in DayCart
- Flexible compound wiring

Central Data Warehouse

- Data is decoded, loaded, cleaned and mapped



Selective data integration



Data Integration

- Consolidated, homogeneous set of data:
  - One index for every entity
  - One unit of measure for every property
- We can:
  - Highlight connections between entities
  - Create new connections
  - Filter on properties
  - Interface to other databases

Chemo-centric design

- Every entity and property is connected to a chemical structure
- Seamless integration of different data sources
  - Can measure how a data source enriches chemical space
- Consistent modelling of tautomers and stereo-isomers
  - Easy to apply hierarchical order (e.g. parent-child)
  - Any (and multiple) grouping of structures allowed
- Intuitive application of chemo-informatics methods

DayCart Oracle Cartridge

- SMILES chemical representation
- Structure comparison, transformation, manipulation
- Fast data retrieval

Pfizer

DayCart: Chemical Representation

- SMILES syntax support
  - Compact, linear representation
  - Self contained language
  - Computer friendly & searchable
- No proprietary data types!

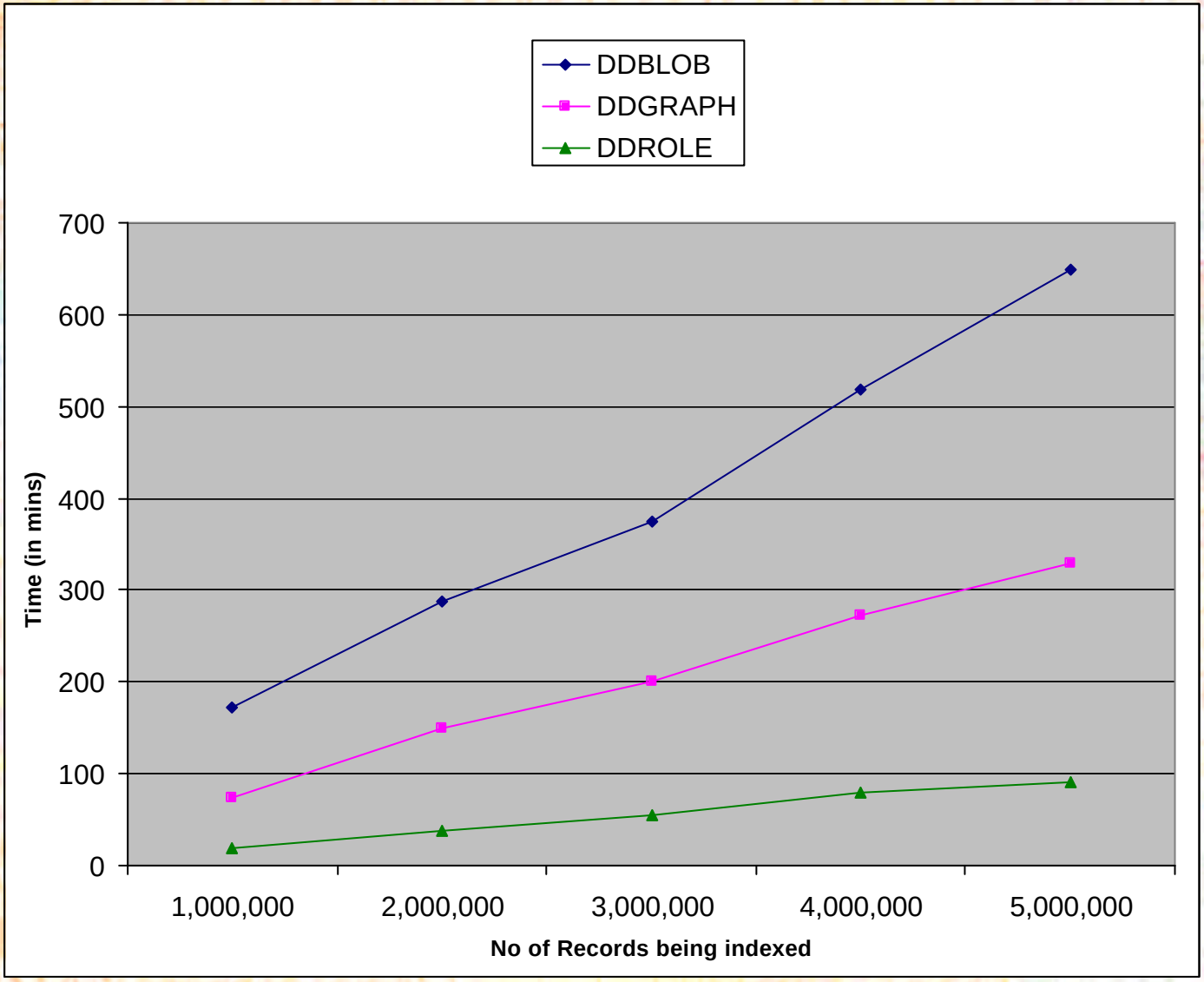
DayCart: Functions for Chemical Information

- Exact match
- Substructure
- Similarity
- Tautomers
- Salts
- Stereochemistry

DayCart: Indexes

- Four (domain) indexes
  - DDBLOB: substructure, similarity
  - DDGRAPH: tautomers, stereochemistry
  - DDROLE: salts
  - DDEXACT: exact match
- Essential for performance
- Trade-off data-load/index building
- Partitioning? (Next version)

DayCart: Indexes



DayCart: VCS_normalize

- Transform structures according to database rules encoded in SMIRKS
- Apply internal business rules
- Standardize structures
- Performance?

Applications

- Perform large-scale data mining
  - Accelerate exploration of new ideas at project inception
- Repository for chemo-informatics knowledge
  - Advanced research database for computational chemists

Example Query: chemical toolbox

Find all *screens* and *compounds* tested against each *target*

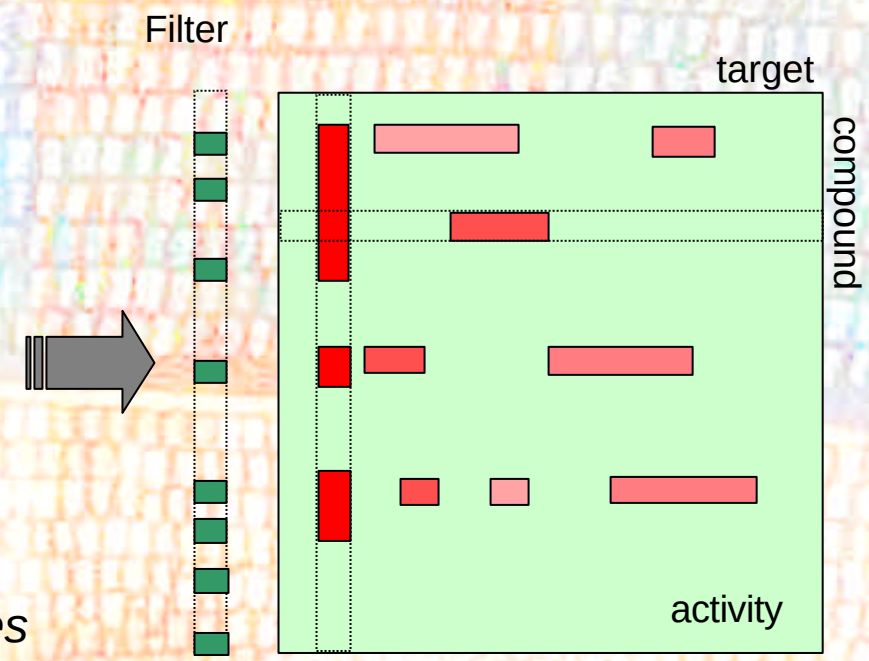
Find all *activity* results & rank compounds

Filter out non druggable compounds

Select available compounds

Filter out non-selective compounds

Select top ten representative diverse *structures*



"Show me the most potent, selective tools for each target, available in-house"

Acknowledgements







Pfizer

Thank you!