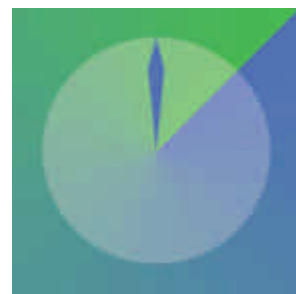


Building and Mining the HIV data cube

Elke Van Craenenbroeck

Luc Dehaspe



PharmaDM

Streamlining Drug Discoveries

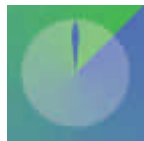
Acknowledgements

► Rega Institute

- ✓ Annemie Vandamme and co-workers
- ✓ Myriam Witvrouw and co-workers
- ✓ Christof Pannecouque and co-workers

► PharmaDM

- ✓ Henk Vandecasteele
- ✓ Wim Trekker
- ✓ Kurt De Grave



Overview

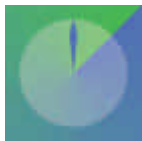
- ▶ Data cubes
- ▶ Data cube mining
- ▶ HIV data cube mining
- ▶ Demo



Data cubes

Data Warehouses - Multidimensional data model

- ▶ Data cube allows data (facts) to be modeled and viewed in multiple dimensions (not restricted to 3)
- ▶ Facts
 - ✓ Central theme, e.g. *sales*, represented by fact table
 - ✓ Facts are numerical measures
- ▶ Dimensions
 - ✓ Perspective on the data
 - E.g., organize sales wrt time, item, branch, location
 - ✓ Described in dimension table
 - E.g., dimension table for item contains attributes ItemName, Brand, Type
 - Dimension tables prespecified or automatically generated (adaptive)



Data cubes

- 2-D view of sales data for *AllElectronics* according to dimensions time and item

location = “Leuven”

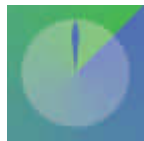
time (<i>quarter</i>)	item (<i>type</i>)			
	<i>Home entertainment</i>	<i>Computer</i>	<i>Phone</i>	<i>Security</i>
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580



Data cubes

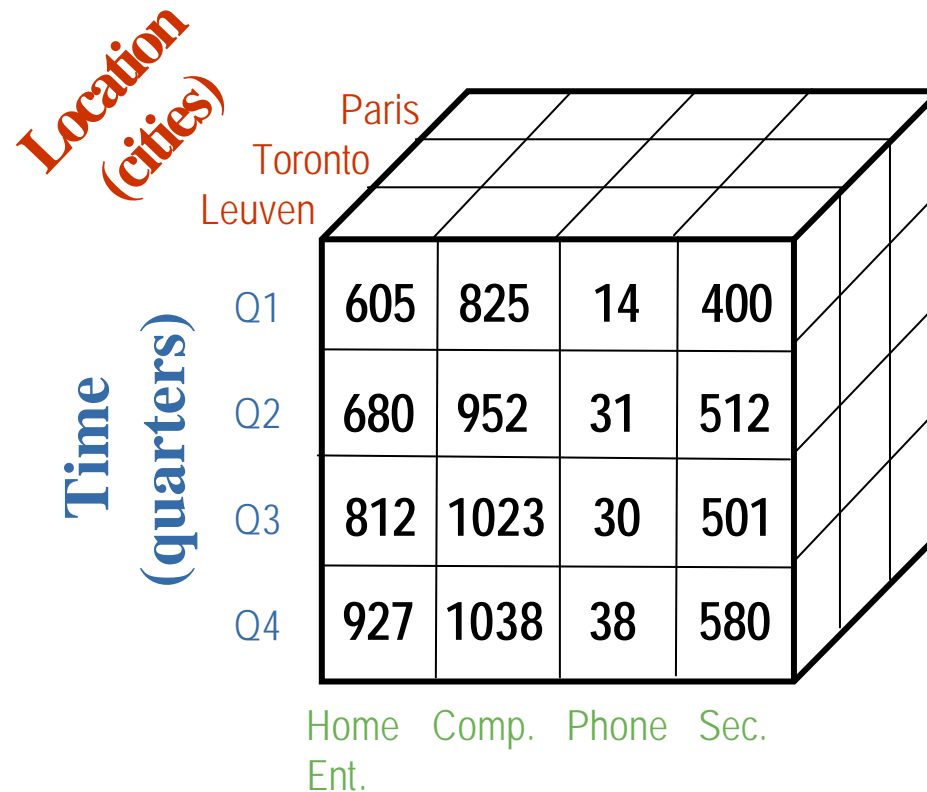
- 3-D view of sales data for AllElectronics according to dimensions time, item, location

	location = "Leuven"				location = "Toronto"				location = "Paris"			
	item (type)				item (type)				item (type)			
time	<i>Home ent.</i>	<i>Comp.</i>	<i>Phone</i>	<i>Sec.</i>	<i>Home ent.</i>	<i>Comp.</i>	<i>Phone</i>	<i>Sec.</i>	<i>Home ent.</i>	<i>Comp.</i>	<i>Phone</i>	<i>Sec.</i>
Q1	605	825	14	400	589	765	23	345	865	675	11	509
Q2	680	952	31	512	755	786	23	657	871	860	49	489
Q3	812	1023	30	501	979	287	14	987	909	768	16	678
Q4	927	1038	38	580	678	1002	24	867	787	1229	29	623

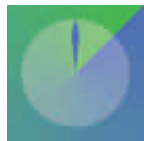


Data cubes

- ▶ Data-cube representation, according to dimensions time, item, location



cuboid



PharmaDM

Streamlining Drug Discoveries

Item (types)

Fact data: sales volume in €1000

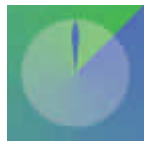
© PharmaDM - 2002

Heverlee June 3, 2002

Data cubes

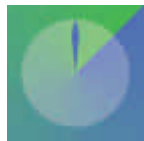
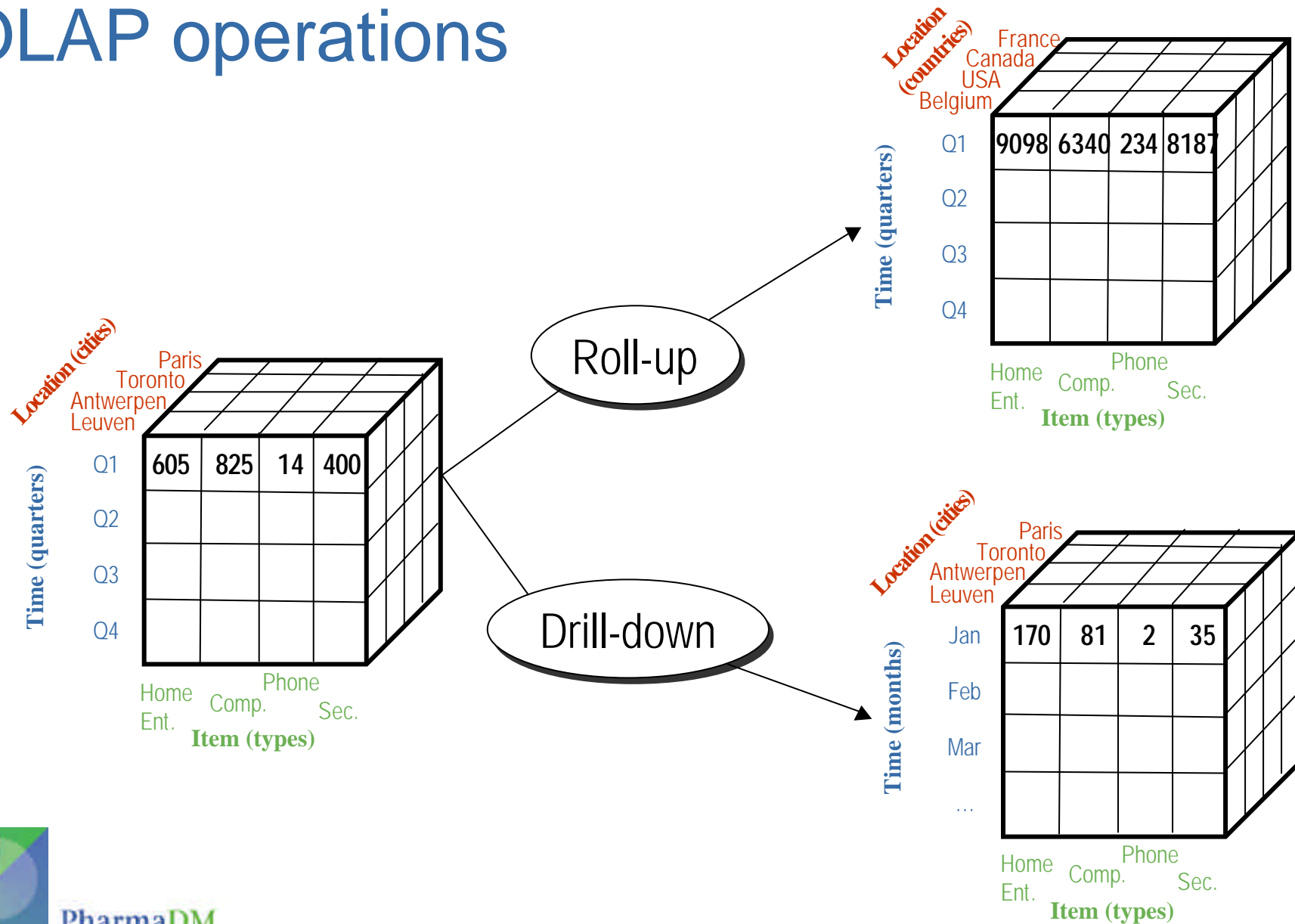
OLAP operations

- ▶ Each dimension contains multiple levels of abstraction defined by concept hierarchies
 - ✓ Users can view data from different perspectives
 - ✓ Data cube operations exist to materialize different views
 - ✓ OLAP provides user-friendly environment for interactive data analysis
- ▶ Operations
 - ✓ Roll-up
 - ✓ Drill-down
 - ✓ Slice and dice
 - ✓ Pivot (rotate)



Data cubes

OLAP operations



Data Cubes

Concept hierarchies

- ▶ Concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts

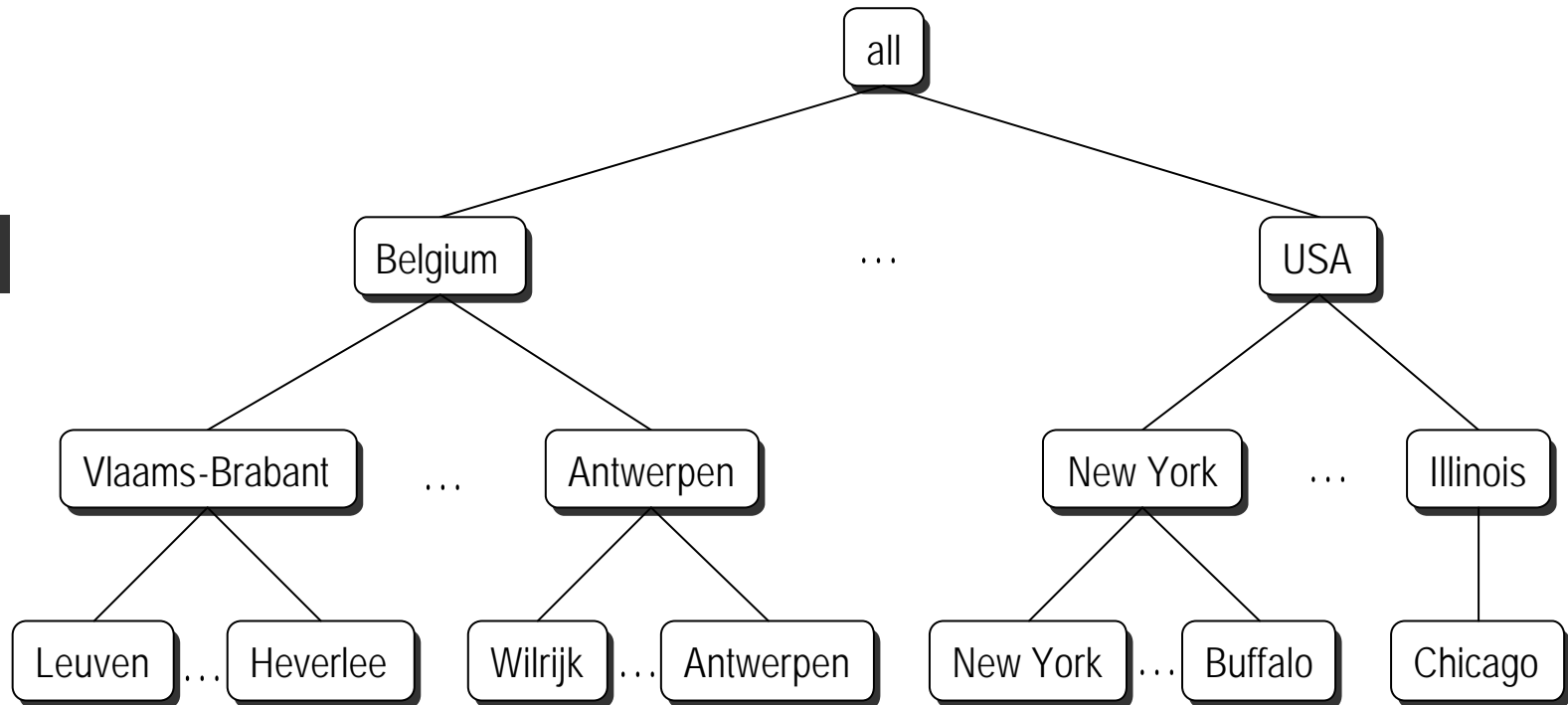
location

all

country

State

city



PharmaDM

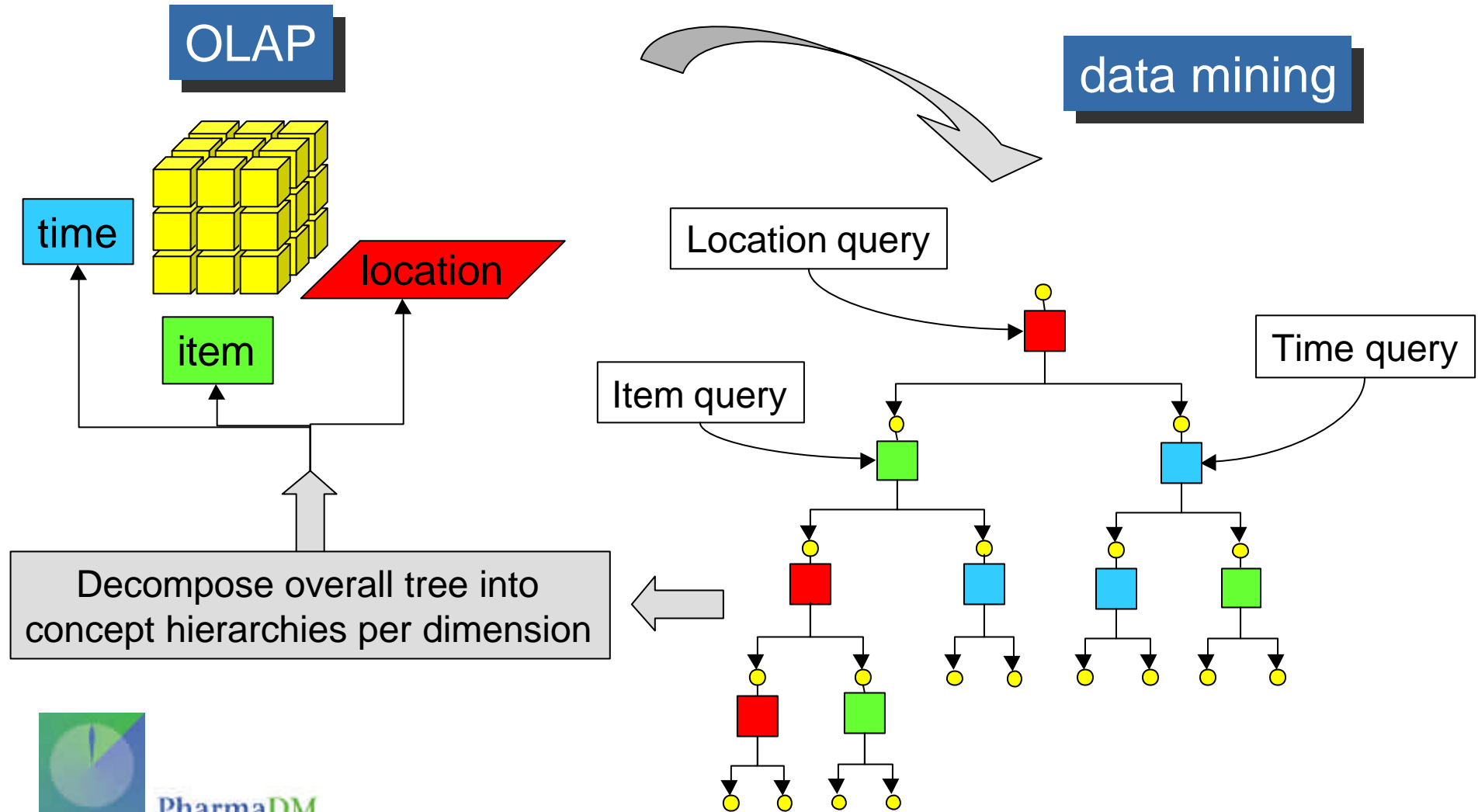
Streamlining Drug Discoveries

© PharmaDM - 2002

Heverlee June 3, 2002

Data cube mining

- Goal: generating concept hierarchies (\approx decision trees) automatically from data cube facts and dimension descriptions



Data cube mining

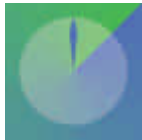
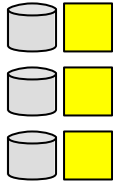
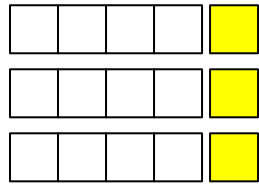
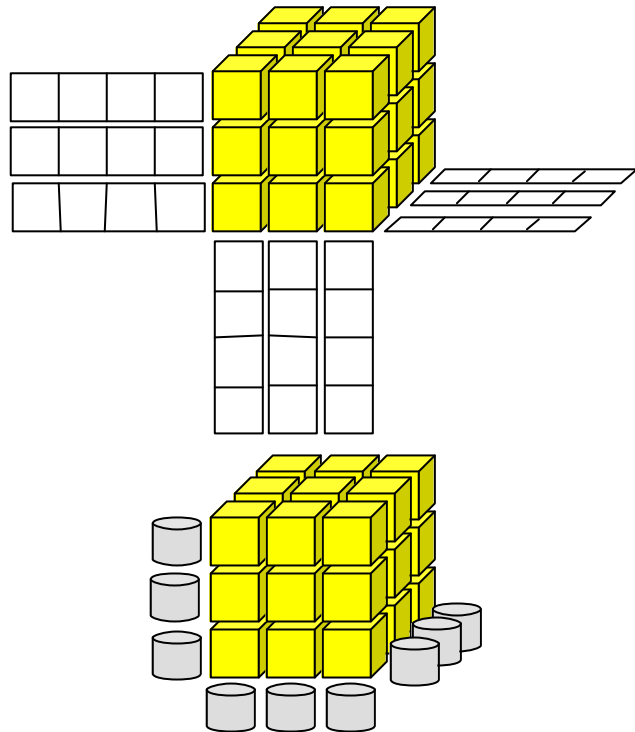
Data representation

propositional

relational

OLAP

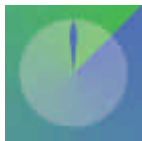
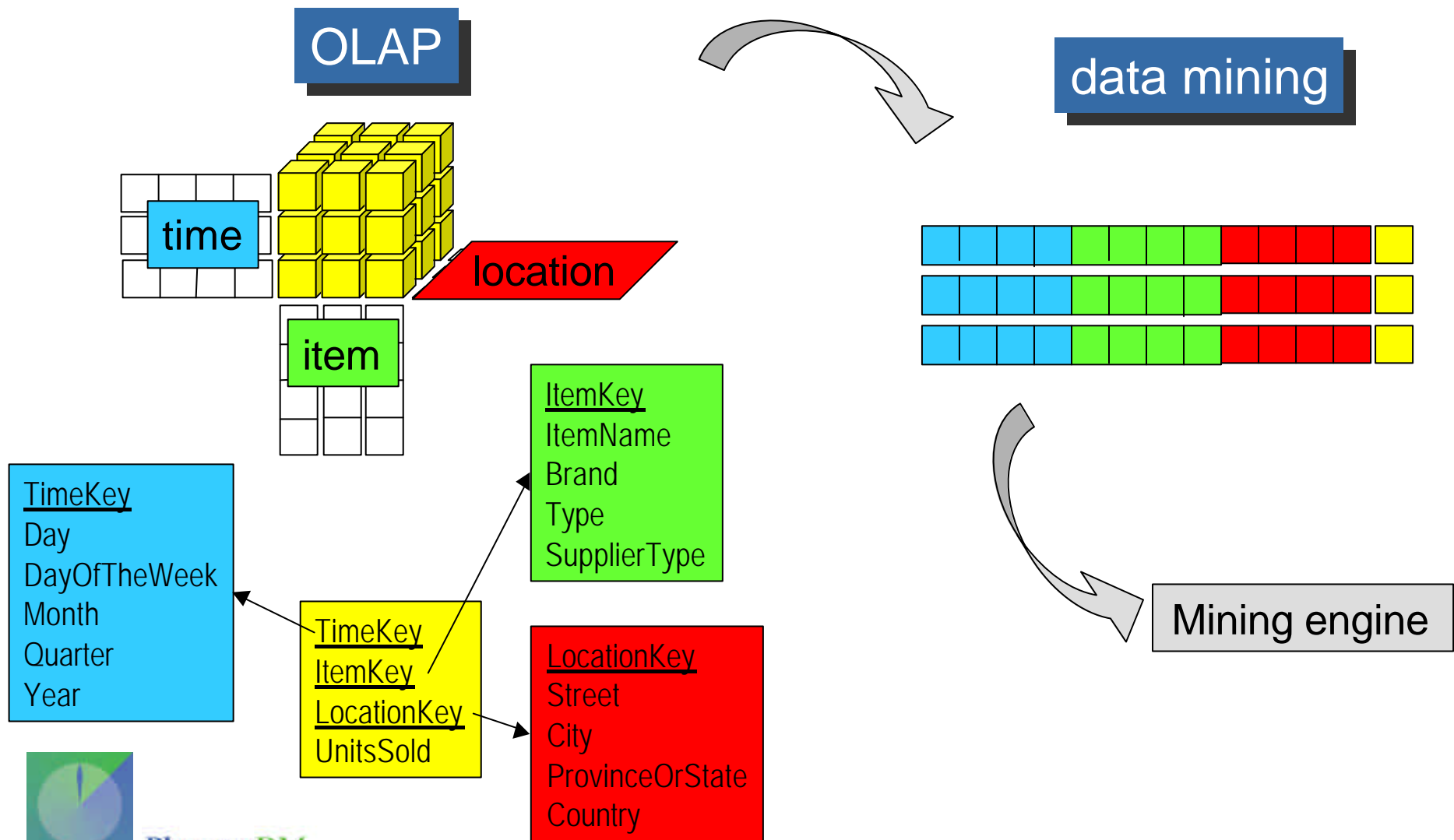
data mining



PharmaDM

Streamlining Drug Discoveries

Data cube mining



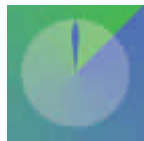
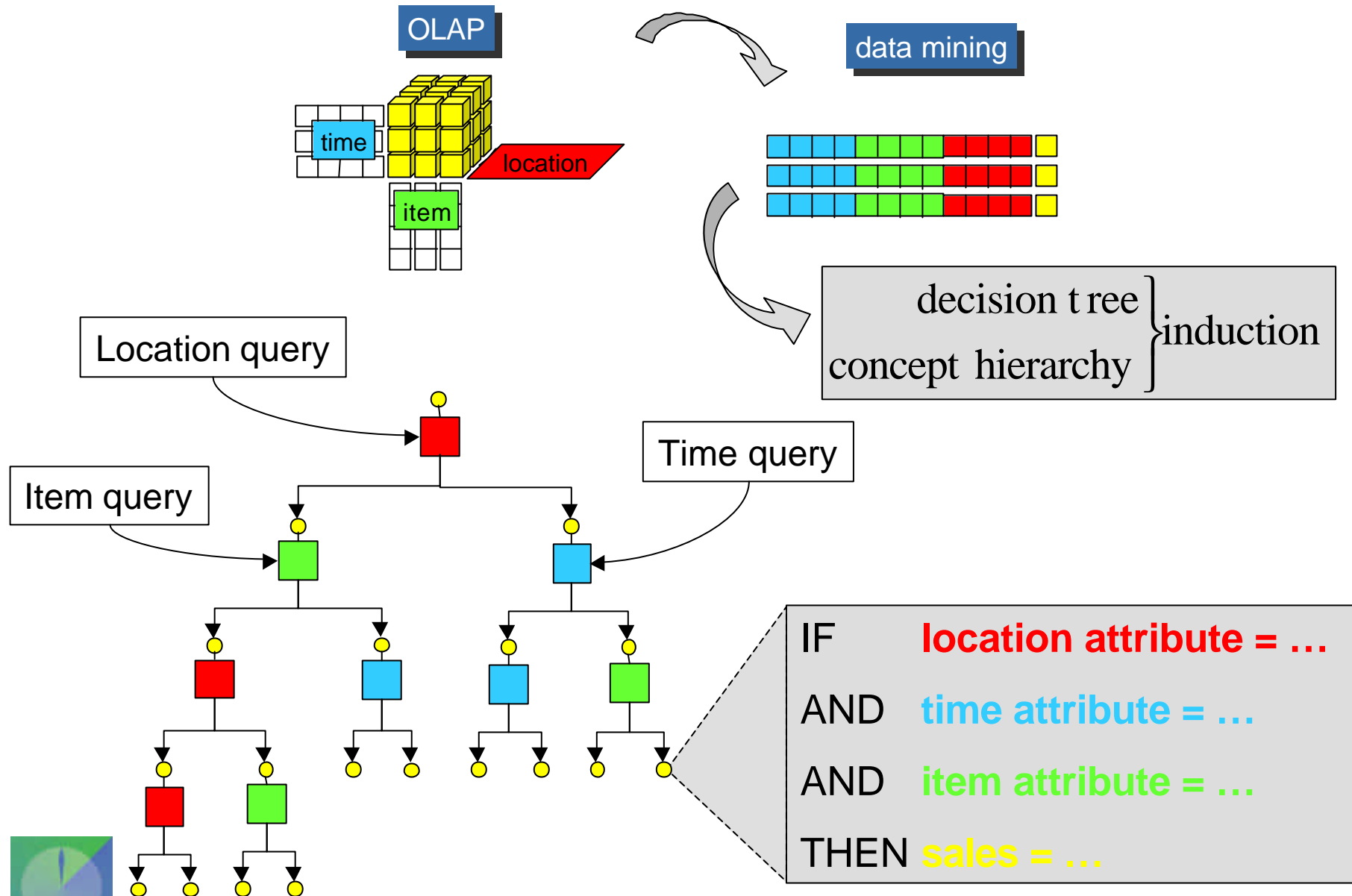
PharmaDM

Streamlining Drug Discoveries

© PharmaDM - 2002

Heverlee June 3, 2002

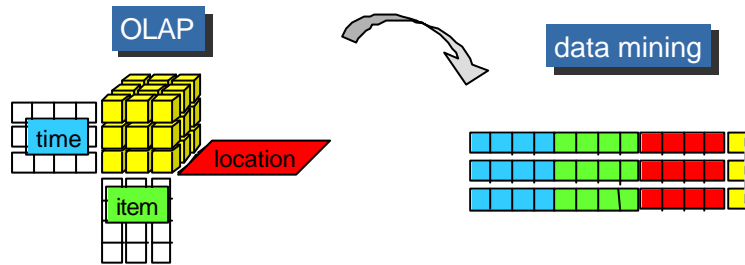
Data cube mining



PharmaDM

Streamlining Drug Discoveries

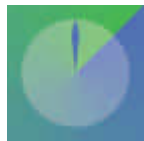
Data cube mining



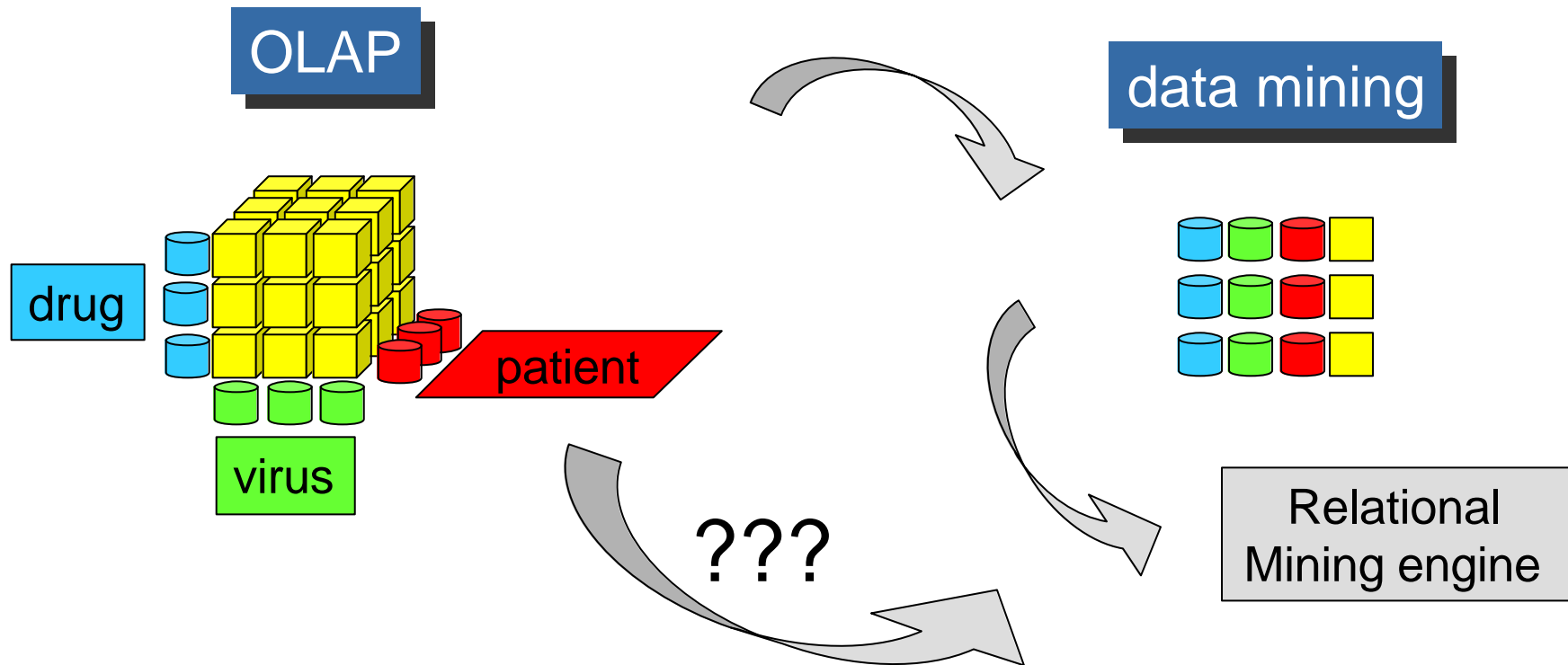
3×3×3 cube ⇒ 27 examples

- redundant storage
- redundant computation

	time	location	item	
1	1	1	item 1	location 1
	1	2	item 1	location 2
	1	3	item 1	location 3
	2	1	item 2	location 1
	2	2	item 2	location 2
	2	3	item 2	location 3
	3	1	item 3	location 1
	3	2	item 3	location 2
	3	3	item 3	location 3
2	1	1	item 1	location 1
	1	2	item 1	location 2
	1	3	item 1	location 3
	2	1	item 2	location 1
	2	2	item 2	location 2
	2	3	item 2	location 3
	3	1	item 3	location 1
	3	2	item 3	location 2
	3	3	item 3	location 3
...



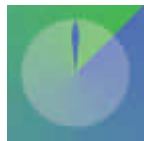
HIV data cube mining



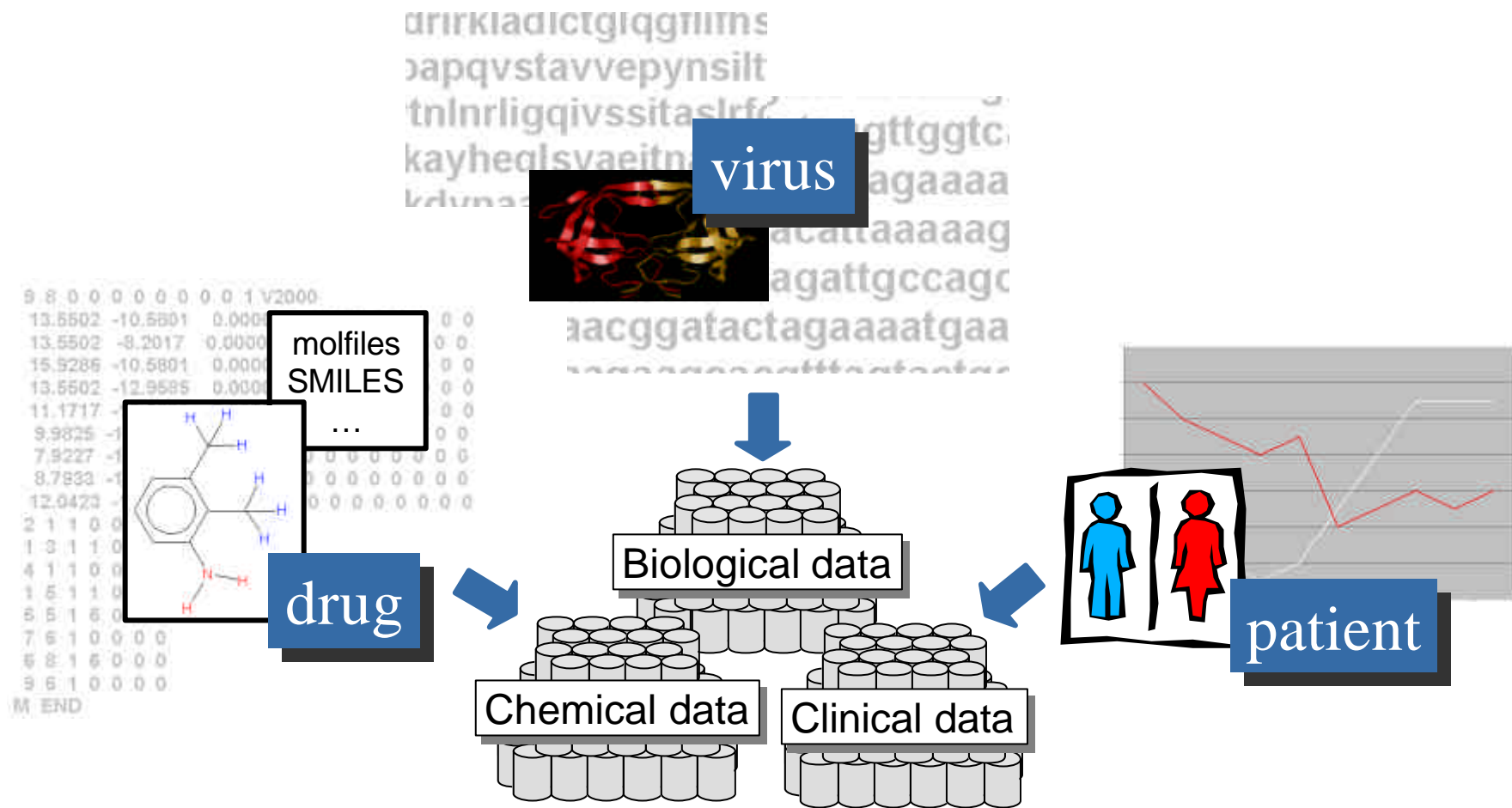
- ▶ massively redundant storage
- ▶ massively redundant computation

Project: PharmaDM – Dep. CS – Rega Institute

- ▶ Integration of life science data
- ▶ Integrated analysis using combinations of data mining techniques
- ▶ Application to HIV (virology) research



HIV research: types of data



HIV research

biological

- *In vitro* screening results of potential drugs (activity, toxicity)
- Genotype of (resistant) virus strains

chemical

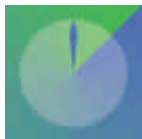
- Structural info on potential drugs
- Physico-chemical properties

clinical

- patient medical history
- therapy response and failure

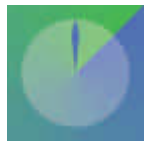
Objectives:

- Insights in structure-activity relationships (SAR) and structure-toxicity relationships
- Insights in molecular origin of viral resistance
- Insights in optimisation of HIV therapy



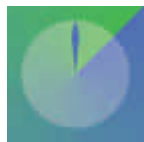
Example: antiviral activity

- ▶ chemical – biological dimension
 - ▶ 8 compounds, 8 virus strains
 - ✓ Outcome: resistant, susceptible
 - ▶ Information to use for prediction
 - ✓ Virological: mutations
 - ✓ Compound info
 - molecular structure
 - mechanism of action



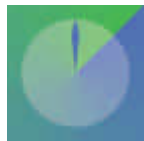
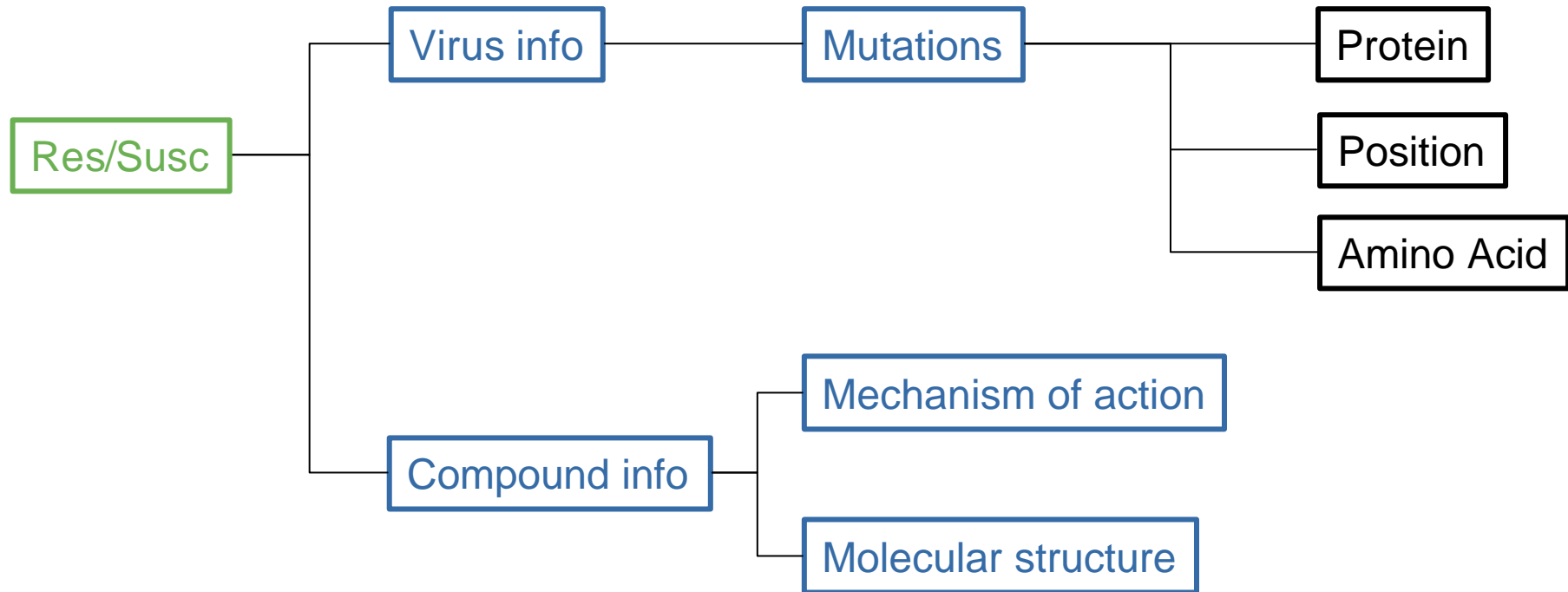
	Strain1	Strain2	Strain3	Strain4	Strain5	Strain6	Strain7	Strain8
Comp1	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible
Comp2	Susceptible	Resistant	Susceptible	Susceptible	Resistant	Susceptible	Resistant	Susceptible
Comp3	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant
Comp4	Susceptible	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible
Comp5	Susceptible	Resistant	Susceptible	Susceptible	Resistant	Susceptible	Resistant	Susceptible
Comp6	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	Susceptible	Susceptible
Comp7	Susceptible	Resistant	Susceptible	Susceptible	Resistant	Susceptible	Resistant	Susceptible
Comp8	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant

- Resistant
- Susceptible

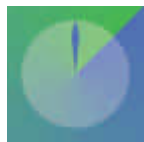
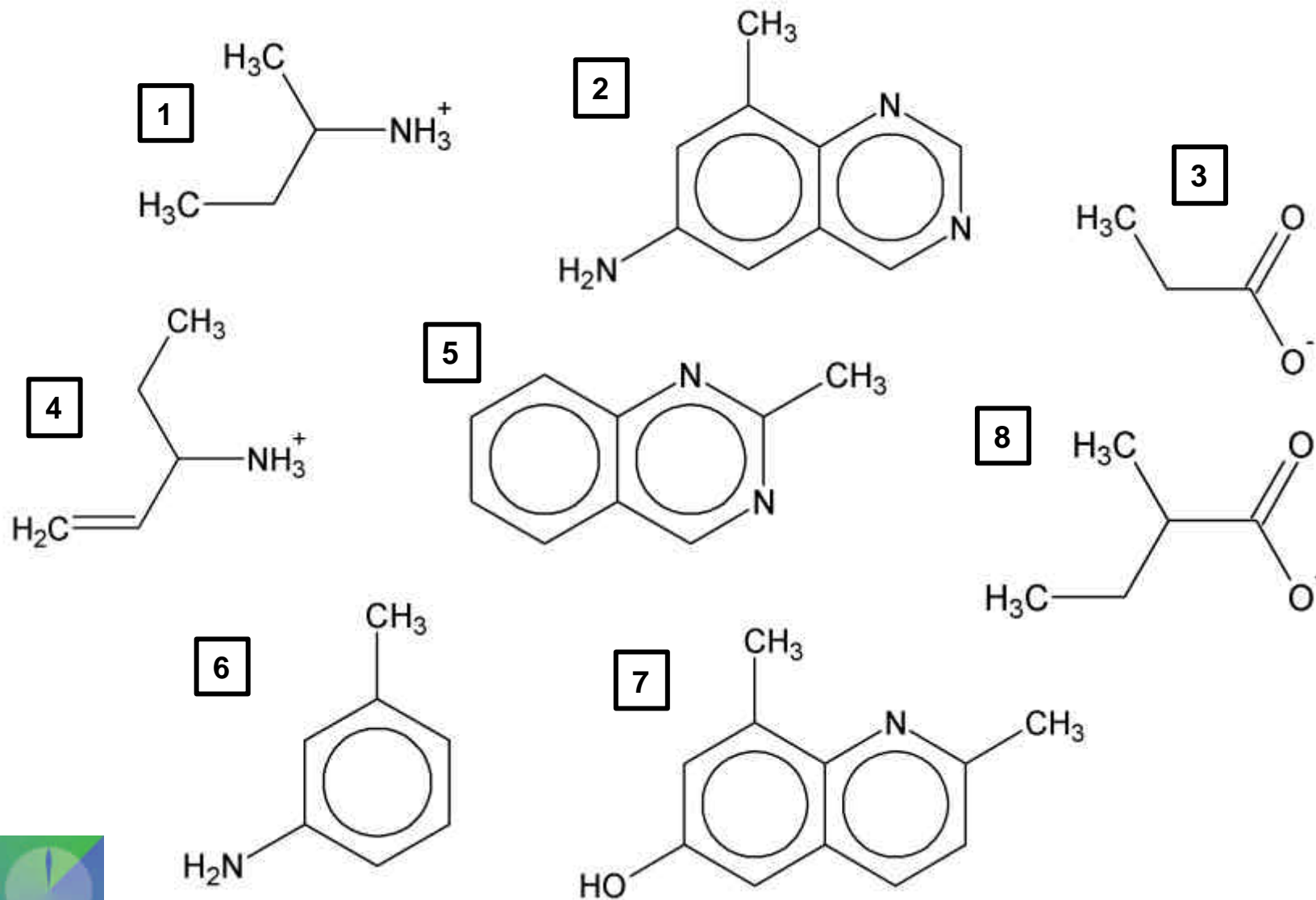


Example: antiviral activity

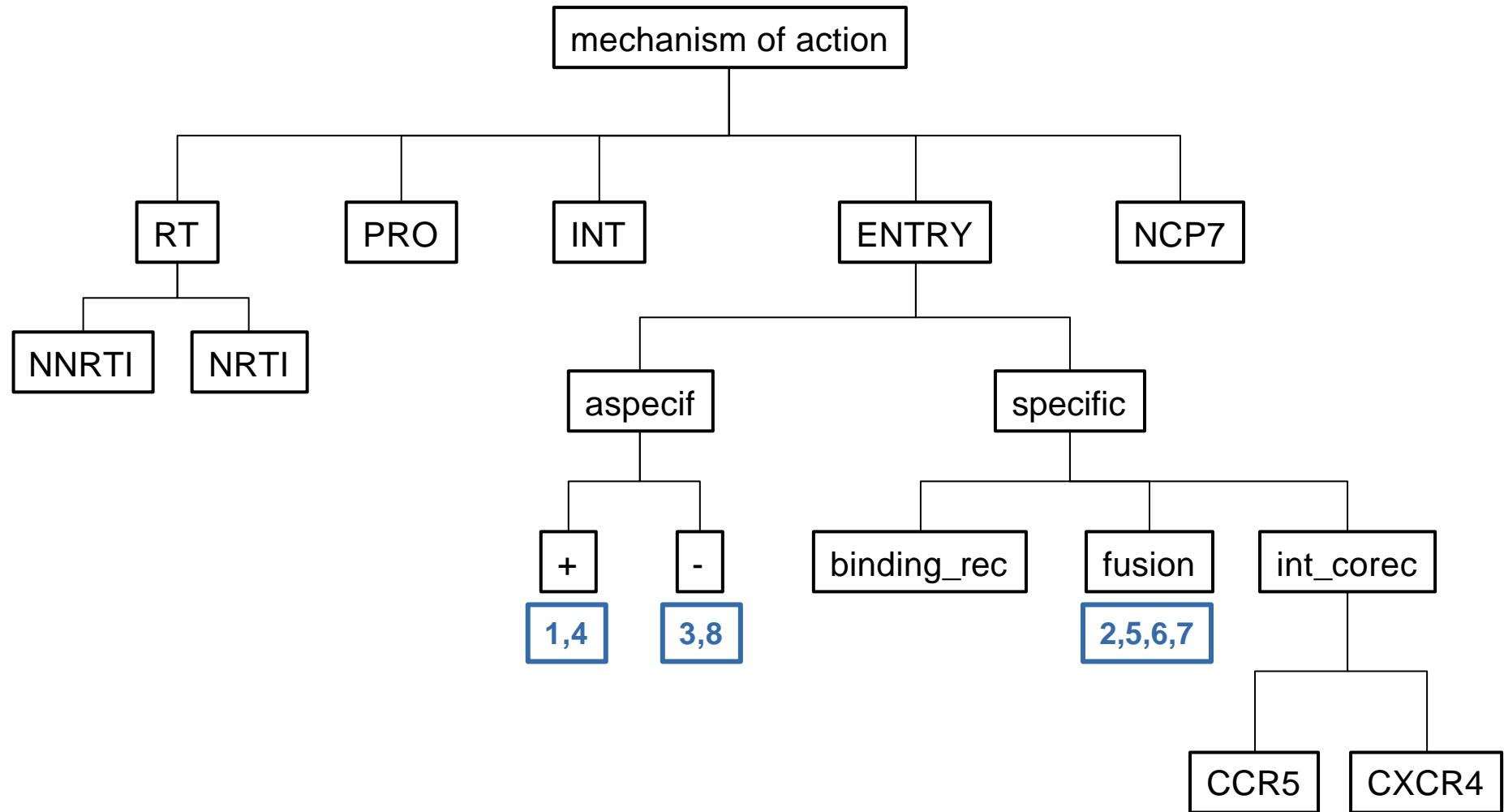
- chemical – biological dimension



Example: antiviral activity: compound info



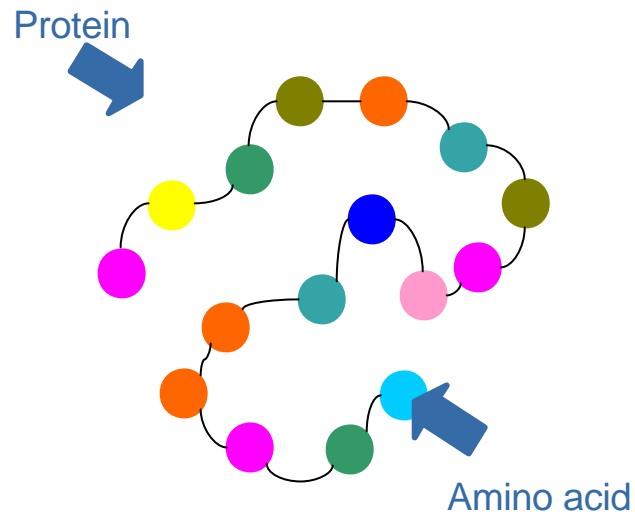
Example: antiviral activity: compound info



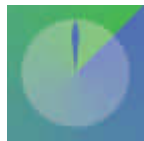
Example: antiviral activity: virus info

► Virus strains

✓ Mutations



Strains	Protein	Mutations
1	gp120	F250A
2	gp41	A190F, T200D
3	gp120	A280E, T285D
4	-	-
5	gp41	A190F
6	gp41	S150G
7	gp41	A190F, T200D
8	gp120	S279E, S284D



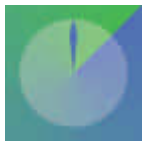
A190F
in gp41

NO

YES

Strain1 Strain6 Strain3 Strain4 Strain8 Strain2 Strain7 Strain5

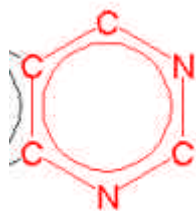
	Strain1	Strain6	Strain3	Strain4	Strain8	Strain2	Strain7	Strain5
Comp1	Red	Green	Green	Green	Green	Green	Green	Green
Comp2	Green	Green	Green	Green	Green	Red	Red	Red
Comp3	Green	Green	Red	Green	Red	Green	Green	Green
Comp4	Green	Green	Green	Red	Green	Green	Green	Green
Comp5	Green	Green	Green	Green	Green	Red	Red	Red
Comp6	Green	Red	Green	Green	Green	Green	Green	Green
Comp7	Green	Green	Green	Green	Green	Red	Red	Red
Comp8	Green	Green	Red	Green	Red	Green	Green	Green



PharmaDM

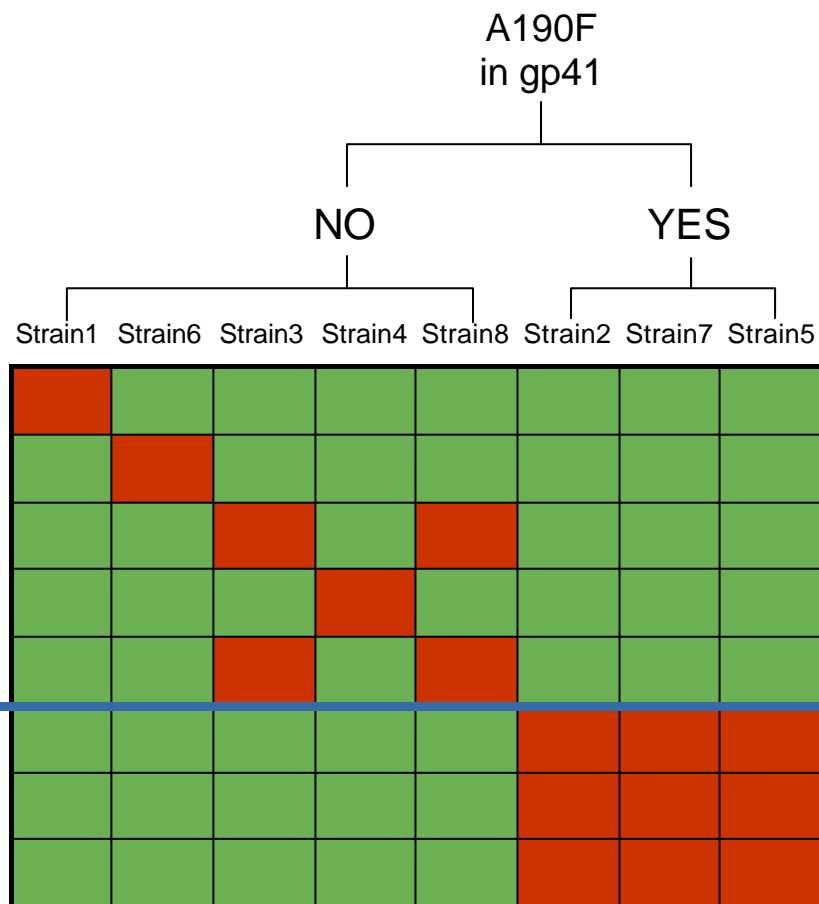
Streamlining Drug Discoveries

Hetero-aromatic ring present



NO

YES



PharmaDM

Streamlining Drug Discoveries

A190F
in gp41

NO

YES

Mutation to
acidic AA

YES

NO

Strain3 Strain8 Strain1 Strain4 Strain6 Strain2 Strain7 Strain5

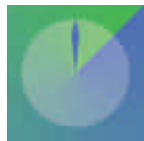
Hetero-aromatic
ring present

NO

YES

Comp1
Comp6
Comp3
Comp4
Comp8
Comp2
Comp7
Comp5

Comp1	Green	Green	Red	Green	Green	Green	Green	Green
Comp6	Green	Green	Green	Green	Red	Green	Green	Green
Comp3	Red	Red	Green	Green	Green	Green	Green	Green
Comp4	Green	Green	Green	Red	Green	Green	Green	Green
Comp8	Red	Red	Green	Green	Green	Green	Green	Green
Comp2	Green	Green	Green	Green	Green	Red	Red	Red
Comp7	Green	Green	Green	Green	Green	Red	Red	Red
Comp5	Green	Green	Green	Green	Green	Red	Red	Red

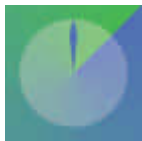
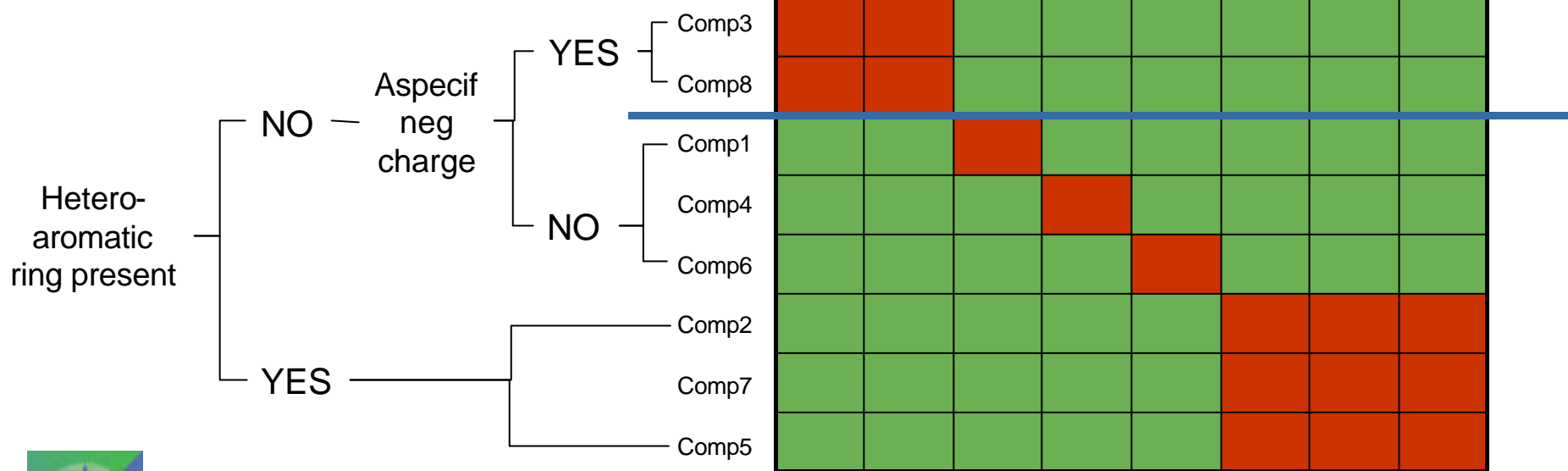
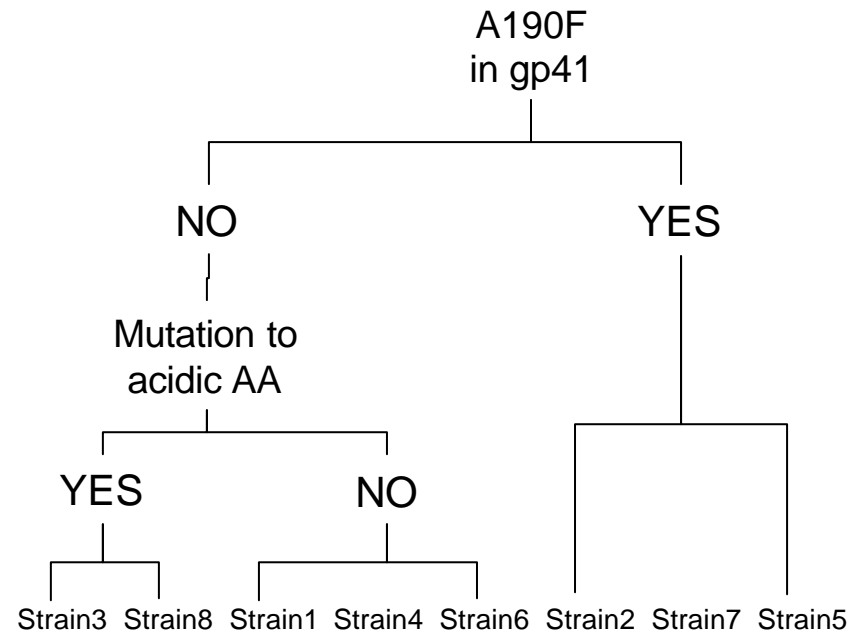


PharmaDM

Streamlining Drug Discoveries

© PharmaDM - 2002

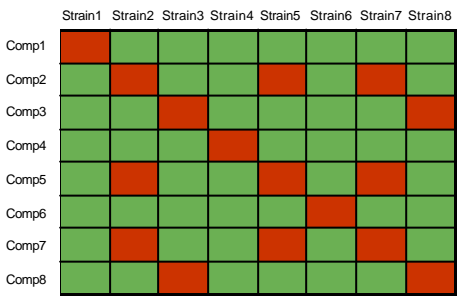
Heverlee June 3, 2002



DMax™ Demo

8 compounds

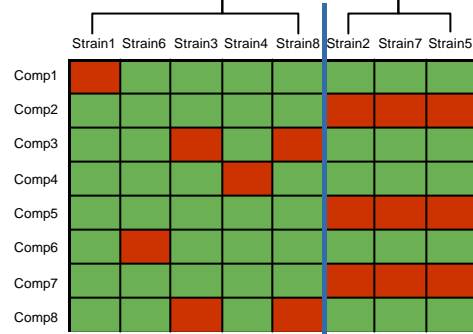
8 virus strains



- Resistant
- Susceptible

A190F in gp41

NO YES

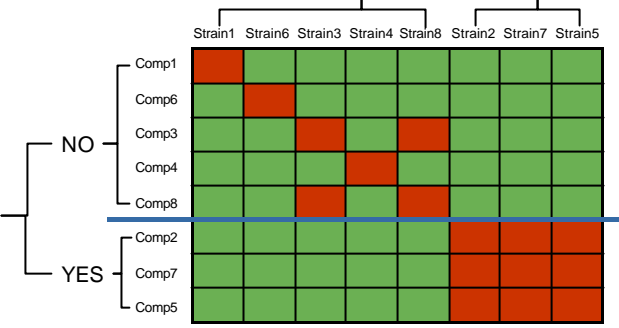


Hetero-aromatic ring present



A190F in gp41

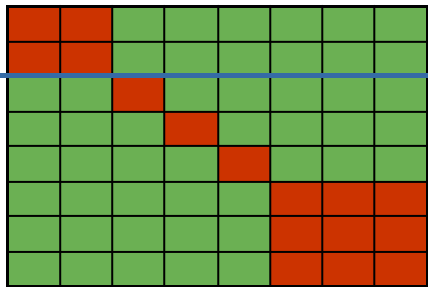
NO YES



A190F in gp41

NO YES

Mutation to acidic AA
YES NO



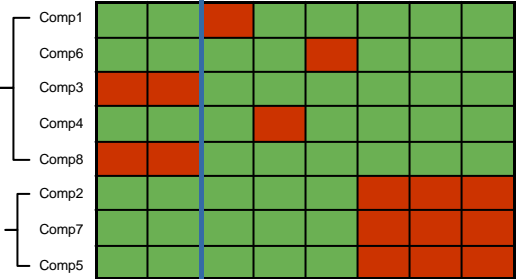
Hetero-aromatic ring present
NO - Aspecific neg charge
YES

Hetero-aromatic ring present

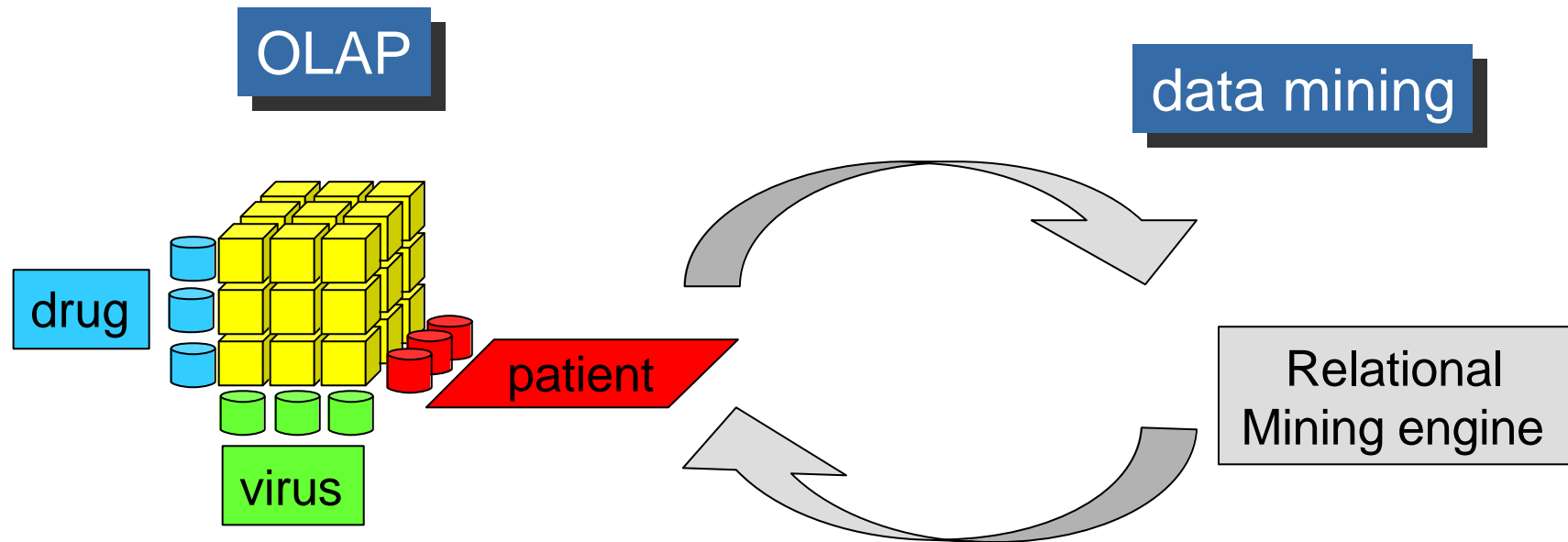
A190F in gp41

NO YES

Mutation to acidic AA
YES NO



Conclusion



- ▶ mining cube-like representation
 - ✓ more efficient storage
 - ✓ more efficient computation
- ▶ decision trees can be decomposed into concept hierarchies

