Simulations Data Access Layer

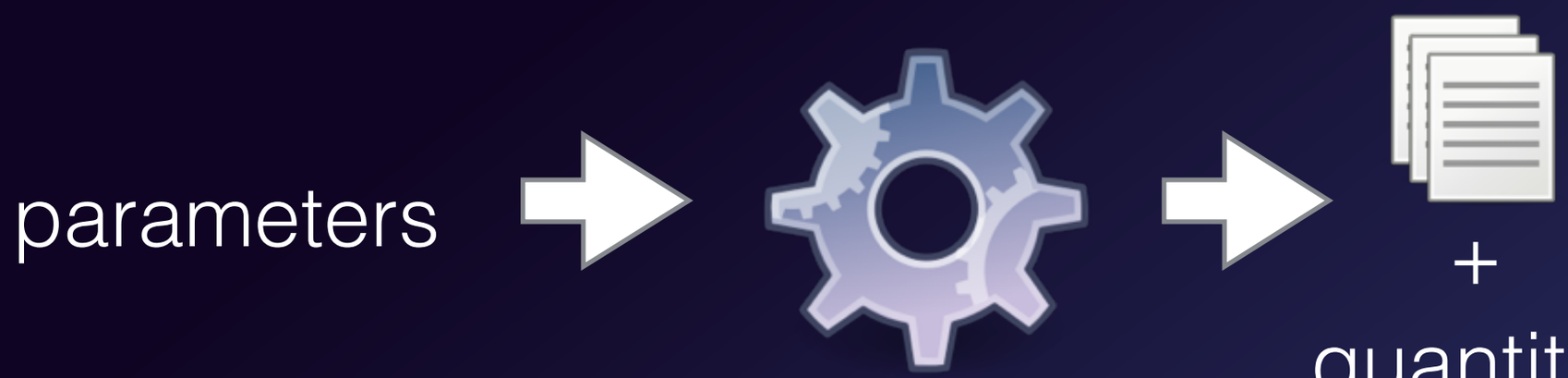
Theory IG

david.languignon@obspm.fr

2015.06.18 - Sesto - IVOA Interoperability meeting

What are we talking about ?

Of course it’s not that simple



Of course it’s not that simple

can be

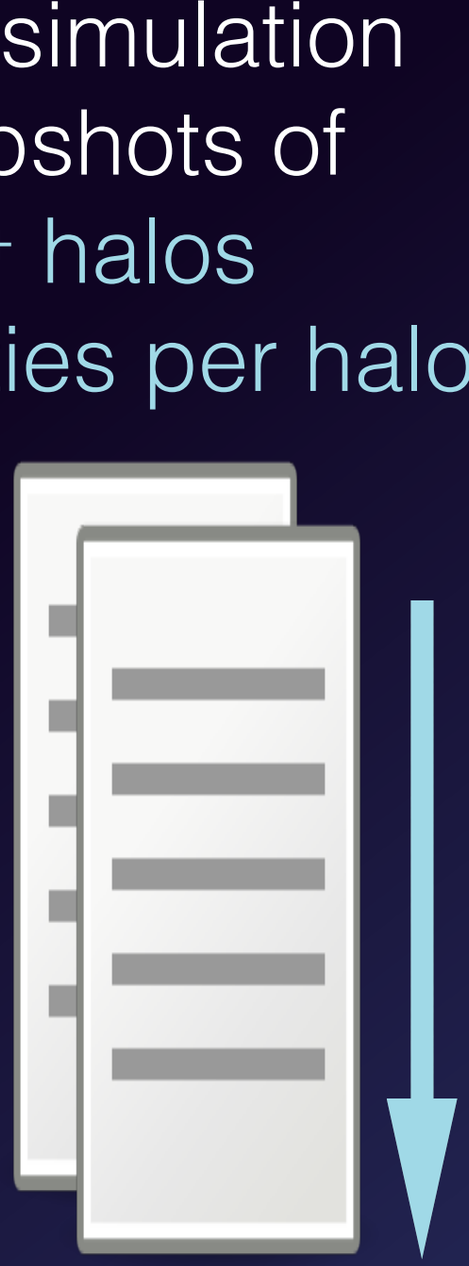
- theoretical spectra (ISM, galaxies)
- time dependant cubes (cosmology, MHD)
- trajectories (planetology)
- chemical structures (Interstellar clouds)
- catalogs (dark matter halos)

+

huge <> small
centralised <> distributed
lots of objects <> lots of properties

Typical use-cases

nbody simulation
2 snapshots of
10^{5+} halos
20 properties per halos



Catalog

Grid

10^{2+} chemical structure simulations
1 cloud per simulation
10^{5+} properties per cloud

So what ?

- Every scientist publishes the data with
  - his own, custom metadata format
  - sometimes without metadata at all…
- Discovery and search among that material
  - must be re-done from scratch for each new publisher
  - sometimes for each project of the same publisher…

Common format

- How do I describe my numerical simulations project so that it brings value to the users ?

  - Simulation Data Model

Where we are now

project

protocol (aka code)

experiments

Where we are now

- No way to know if there exists projects modelling the data/ process I deal with
- No way to know if such a project produced the data in the specific configuration I need
- No way to know, then, if I can access this data, and how

Time & money spent thinking & designing custom solutions, not re-usable, not interoperable

What it could be

- Standard protocol
- Interoperable services
- Re-usable components

Maximum ROI for simulation projects, observational missions & data publishing projects funding

And… put scientist back to research instead of struggling with data

Simulations Data Access Layer

- Discover if the kind of models you need exists
  - SimDAL Repository
- Search for interesting datasets in a particular project
  - SimDAL Search
- Access raw data & data cube cutouts
  - SimDAL Data Access

Simulation Repository

- Store simulation projects metadata, in standard IVOA format
- (intended) Centralised repository, with basic text search
- Very simple implementation
  - give access to SimDM xml serialisation files

Simulation Search

consider several experiments using code “Ac”

pressure
density

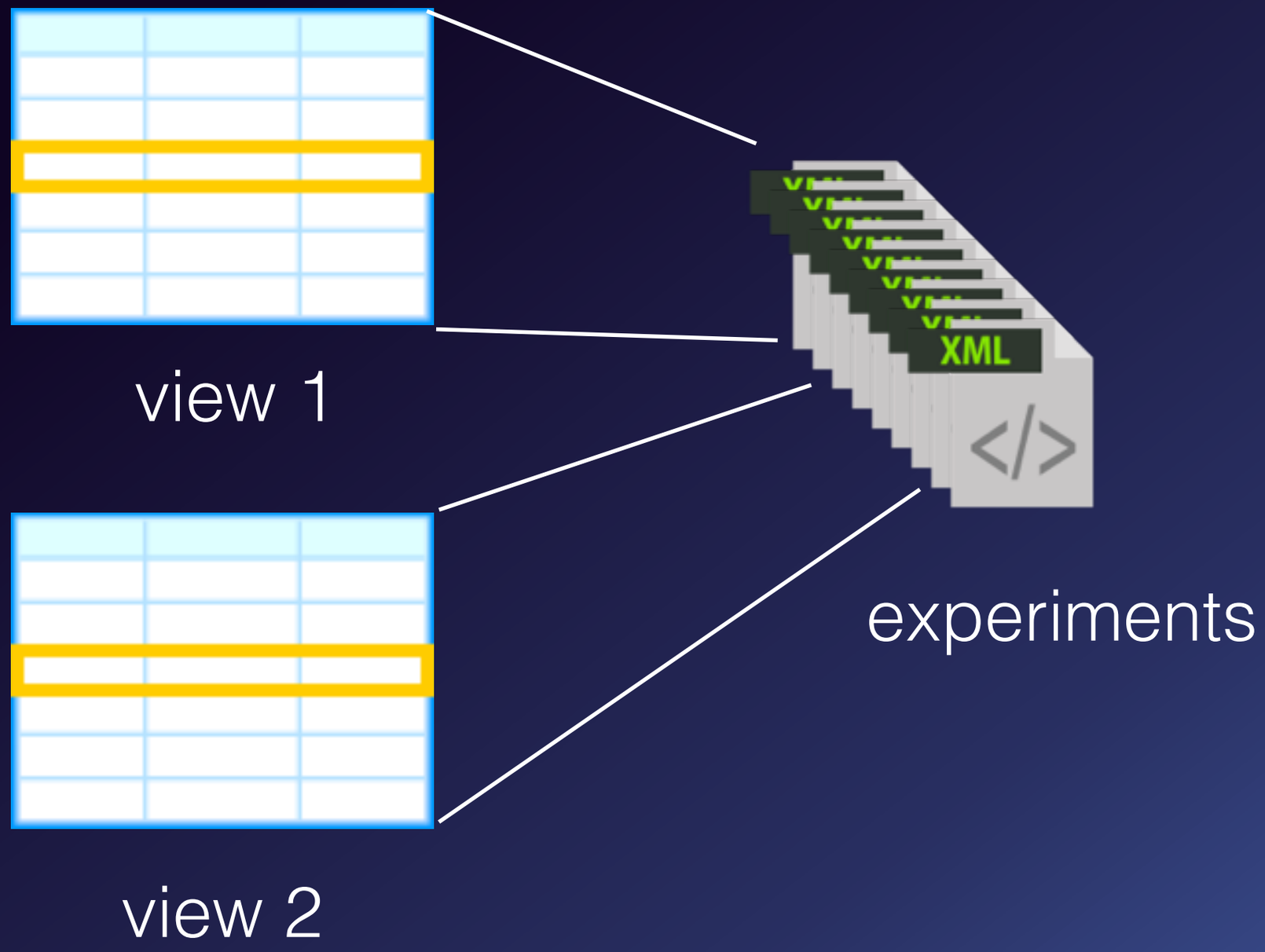
code Ac

pmass
pvelocity

Simulation Search

experiments

Simulation Search



Simulation Search

view 1

dataset	run	pressure	density	pmass	pvelocity
d1	r1	1	1	0	10
d2	r1	1	1	2	9
d3	r1	1	1	3	7
d1	r2	2	1	7	2

Simulation Search

View schema

dataset	run	pressure	density	pmass	pvelocity
d1	r1	1	1	0	10
d2	r1	1	1	2	9
d3	r1	1	1	3	7
d1	r2	2	1	7	2

view 1

view	column	…	utype	doc	datatype
view1	dataset		simdm://	…	text
view1	pressure		simdm://	…	double
view1	pmass		simdm://	…	double
…					

view 1 schema

VOTable
(header only)

Simulation Search

View query

select dataset **where** pmass > 2 and pmass < 5

dataset	run	pressure	density	pmass	pvelocity
d1	r1	1	1	0	10
d2	r1	1	1	2	9
d3	r1	1	1	3	7
d1	r2	2	1	7	2

Simulation Search

View design

- Views are abstractions, can have infinite number of columns
  - can be implemented however you want !
    - rdbms (mapped to flat tables), document db, xml file
- Views have flat table oriented simple query language

We can describe objects with any number of properties

The user is always exposed to easy to query flat tables

Simulation Search

View design

- No longer stuck with relational db columns number limits

DB engine	Max columns*
mysql	4096
oracle	1000
sql server	30 000
postgres	250 - 1600

*src: wikipedia

- The underlying implementation can be designed to fit the expected number of columns.

Unlocks critical use cases involving huge columns number

Simulation Data Access

- **Easy access to dataset raw data**
  - sync ressource with links towards files
- **Easy access to cube subset -cutout-**
  - user can ask for dataset schema (same as Search)
  - user can query dataset cutout through async resource (same simple query language as Search)

Simulation Data Access

Cutout

Simulation Data Access

- Dataset id found in SimDAL Search
- Dataset is exposed through the same view abstraction than in Search

select object **where** y > 25

object	x	y	z	rv	charge
o1	12	0	12	12	23
o2	34	100	234	7	12
o3	45	23	2	3	14
o4	21	29	45	7	15

IVOA integration

- All resource responses are VOTable
- Async resources comply with the uws standard
- Built on top of the SimDM standard
- DALi & VOSI

IVOA integration, specificities

- REST interface with hypermedia control
- Stream pagination system based on REST / VOTable
- The view abstraction, semantically close to TAP/TAP_SCHEMA
- Cutout queries are json documents, posted to a resource

Take away

- Answers 3 main use cases, through 3 components
  - discover a project of interest
  - search inside a project for interesting datasets
  - access a dataset subset and/or raw material
- Built on top of SimDM
- Use existing IVOA standards & best practices