CSCS

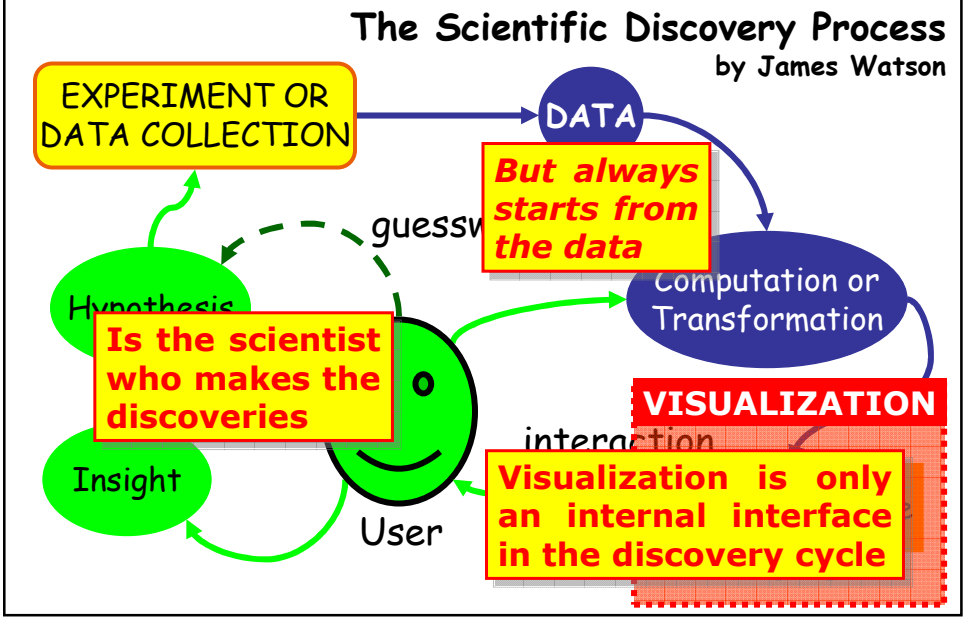
Swiss National Supercomputing Centre

Mario Valle

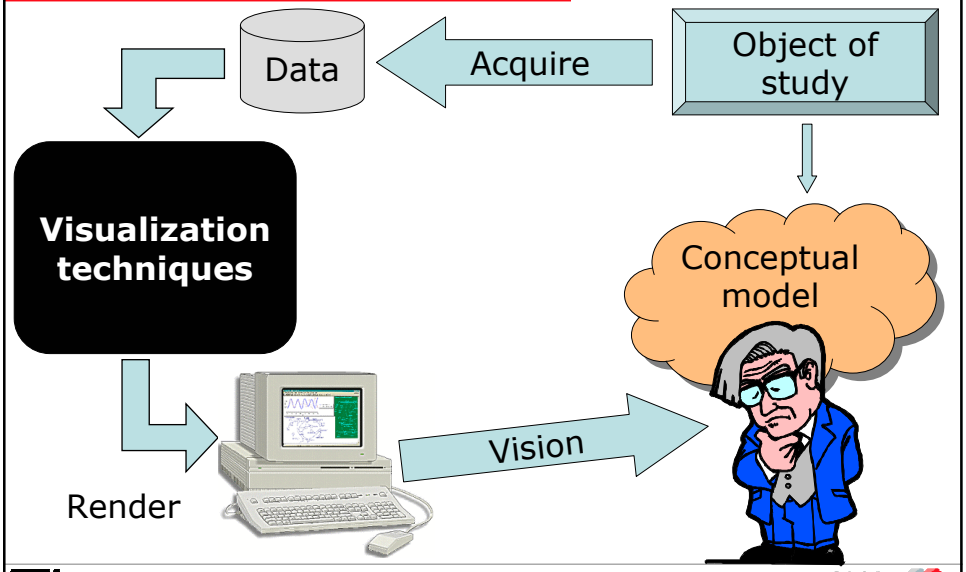
CECAM – 18-19/09/2008

Visualization issues for chemistry data

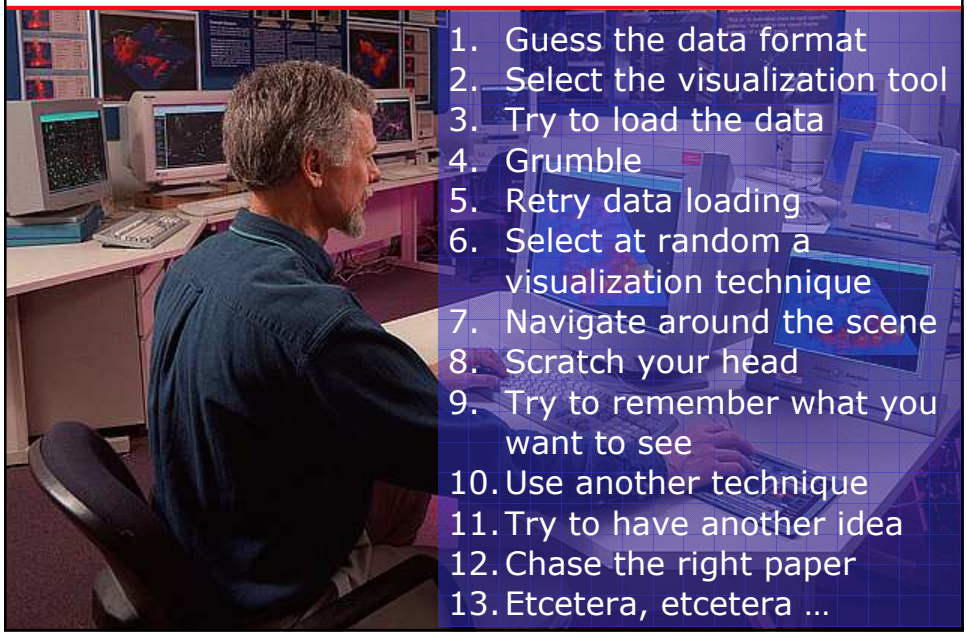
Visualization fosters discovery



The visualization process



The real visualization cycle



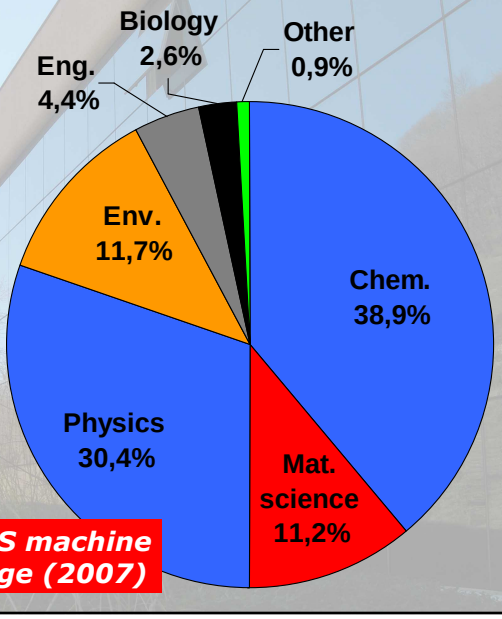
1. Guess the data format
2. Select the visualization tool
3. Try to load the data
4. Grumble
5. Retry data loading
6. Select at random a visualization technique
7. Navigate around the scene
8. Scratch your head
9. Try to remember what you want to see
10. Use another technique
11. Try to have another idea
12. Chase the right paper
13. Etcetera, etcetera …

Chemistry support at CSCS

In 2007 almost **exactly half** of the CSCS machines time has been assigned to chemistry related projects.

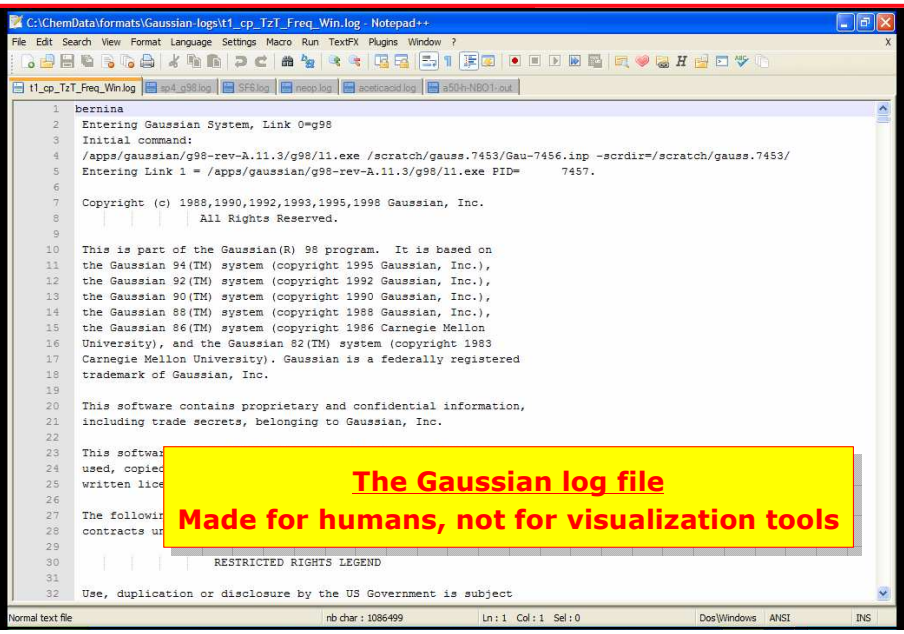
The visualization group works in **strict cooperation with the users.**

Not only for day-to-day support, but acting also as a **catalyst** for introducing innovative techniques and methods that **help their work.**



CSCS machine usage (2007)

The archetypal problem...



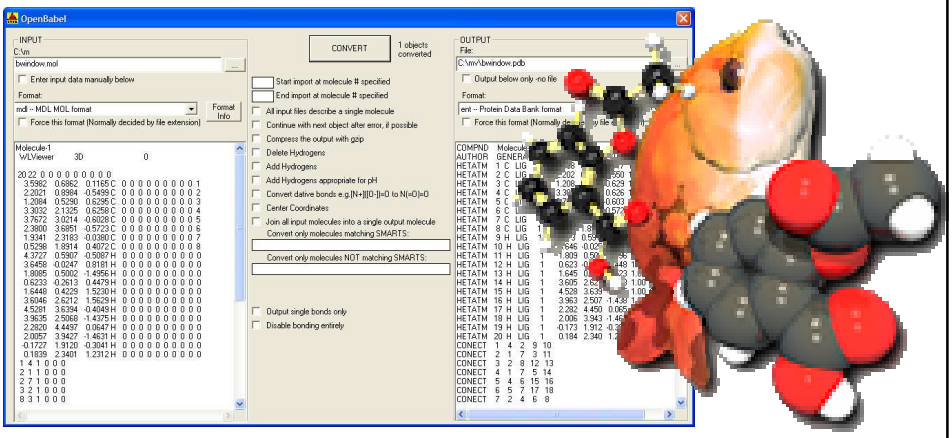
A puzzle for visualization tools

```
HEADER    1 Configuration at time step     4800.00
ATOM      1 N    ASP     1      14.927  -1.261 -21.608 -0.52  0.00
ATOM      2 HN   ASP     1      14.887  -2.262 -21.481  0.29  0.00
ATOM      3 CA   ASP     1      14.856  -0.745 -22.986  0.04  0.00
ATOM      4 HA   ASP     1      15.901  -0.688 -23.290  0.09  0.00
ATOM      5 CB   ASP     1      14.029  -1.678 -23.900 -0.03  0.00
ATOM      6 HB1  ASP     1      13.045  -1.215 -23.986 -0.01  0.00
ATOM      7 HB2  ASP     1      14.469  -1.602 -24.894 -0.01  0.00
ATOM      8 CG   ASP     1      13.834  -3.180 -23.648  0.80  0.00
ATOM      9 OD1  ASP     1      13.554  -3.809 -24.700 -0.80  0.00
ATOM     10 OD2  ASP     1      13.887  -3.809 -22.529 -0.80  0.00
ATOM     11 C    ASP     1      14.276   0.676 -23.013  0.54  0.00
ATOM     12 O    ASP     1      14.394   1.340 -24.055 -0.58  0.00
ATOM     13 N    ILE     2      13.776   1.127 -21.895 -0.41  0.00
ATOM     14 HN   ILE     2      13.822   0.443 -21.154  0.27  0.00
ATOM     15 CA   ILE     2      13.807   2.541 -21.468 -0.06  0.00
ATOM     16 HA   ILE     2      14.643   3.012 -21.984  0.09  0.00
ATOM     17 CB   ILE     2      12.558   3.270 -21.979  0.13  0.00
ATOM     18 HB   ILE     2      12.273   2.957 -22.983  0.02  0.00
ATOM     19 CG2  ILE     2      11.342   3.013 -21.077 -0.32  0.00
ATOM     20 HG21 ILE     2      11.191   1.934 -21.102  0.09  0.00
ATOM     21 HG22 ILE     2      10.416   3.334 -21.554  0.09  0.00
ATOM     22 HG23 ILE     2      11.478   3.383 -20.061  0.09  0.00
ATOM     23 CG1  ILE     2      12.735   4.862 -22.021 -0.04  0.00
ATOM     24 HG11 ILE     2      13.725   5.027 -22.447  0.02  0.00
ATOM     25 HG12 ILE     2      12.796   5.238 -21.000  0.02  0.00
ATOM     26 CD1  ILE     2      11.683   5.605 -22.779 -0.07  0.00
ATOM     27 HD1  ILE     2      11.179   4.915 -23.456  0.02  0.00
ATOM     28 HD2  ILE     2      12.118   6.412 -23.369  0.02  0.00
ATOM     29 HD3  ILE     2      10.989   6.004 -22.040  0.02  0.00
```

- Why should I guess the atom type from the atom name?
- Why should I force one unrelated field to contain a user defined value?
- Why a trajectory should be stored in an *ad-hoc* manner?

Offered with compliments to you by your friendly PDB file format

Universal converter/reader?



Open Babel is a project designed to interconvert between many chemistry file formats

Unfortunately it fall short in various areas

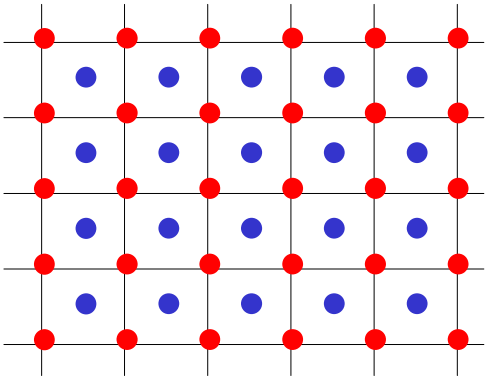
1st humble request

Provide us with a file format:

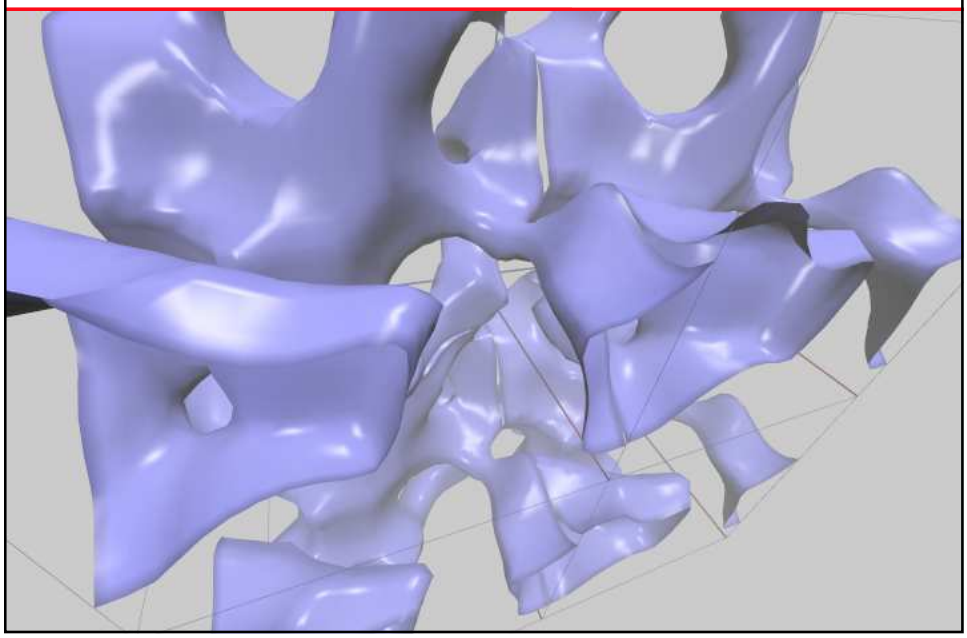
- Easily parseable
- That supports all the information the users need
- That supports all the information the user do not know yet they need

Where are the values defined?

Gaussian Cube, PLT, CHGCAR are unclear on this point



Where are the values defined?

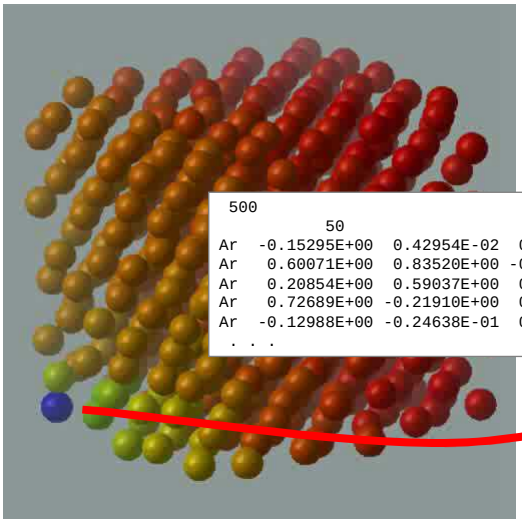


2nd humble request

Provide us with a file format:

- With a well defined semantic
- That does not play dirty tricks to spare a few bytes
- That accepts user contributions to clarify corner cases

What is wrong with this data?



```
 500
           50
Ar  -0.15295E+00  0.42954E-02  0.67474E-01  0.30441E+00  0.78083E+00
Ar   0.60071E+00  0.83520E+00 -0.81022E-01  0.41648E+00
Ar   0.20854E+00  0.59037E+00  0.81560E+00  0.33224E+00
Ar   0.72689E+00 -0.21910E+00  0.53502E+00  0.28482E+00
Ar  -0.12988E+00 -0.24638E-01  0.16533E+01  0.30124E+00
 . . .
```

What sounds strange (or interesting…) with this data?

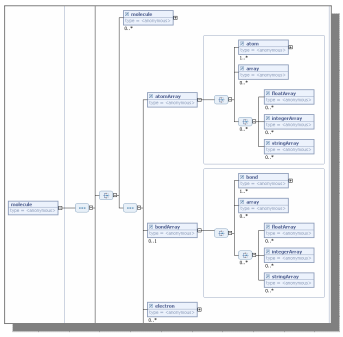
A bug in my XYZ data reader…

3rd humble request

Provide us with methods to write the files from custom (one-off) code:

- Simple to use (at the level of a bunch of "printf")
- Flexible
- Simple to use (already said that?)

But there are well built formats



CML

CIF

```
data_global
_chemical_name 'Calcite'
loop_
_publ_author_name
'Graf D L'
_journal_name_full "American Mineralogist"
_journal_volume 46
_journal_year 1961
_journal_page_first 1283
_journal_page_last 1316
_publ_section_title
;
Crystallographic tables for the rhombohedral carbonates
;
_chemical_formula_sum 'Ca C O3'
_cell_length_a 4.9900
_cell_length_b 4.9900
_cell_length_c 17.0615
_cell_angle_alpha 90
_cell_angle_beta 90
_cell_angle_gamma 120
_cell_volume 367.916
_symmetry_space_group_name_H-M 'R -3 c'
loop_
_symmetry_equiv_pos_as_xyz
'x,y,z'
'2/3+x,1/3+y,1/3+z'
'1/3+x,2/3+y,2/3+z'
'x,x-y,1/2+z'
'2/3+x,1/3+x-y,5/6+z'
'1/3+x,2/3+x-y,1/6+z'
```

Similar experience with CFD data

CFX

1. Proprietary
2. Difficult to read
3. Format not documented

CGNS cfd data standard

1. Standard
2. Covers everything
3. Readers/Writers available

cfd data standard

Similar experience with CFD data

1. Proprietary
2. Difficult to read
3. Format not documented

1. Standard
2. Covers everything
3. Readers/Writers available

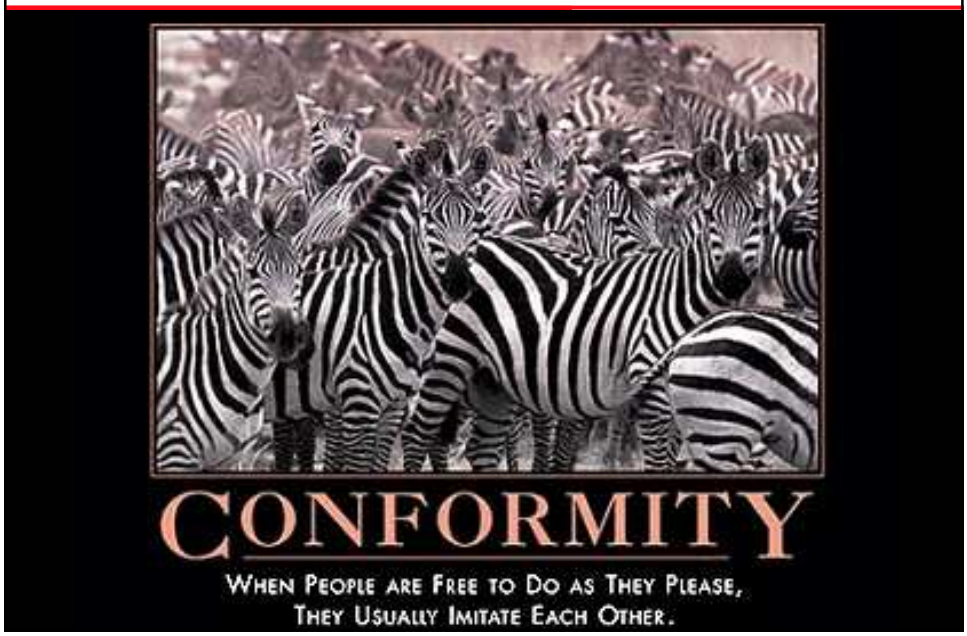
1. Need post conversion
2. Designed by committee
3. Always something missing
4. Low quality implementations

4th humble request

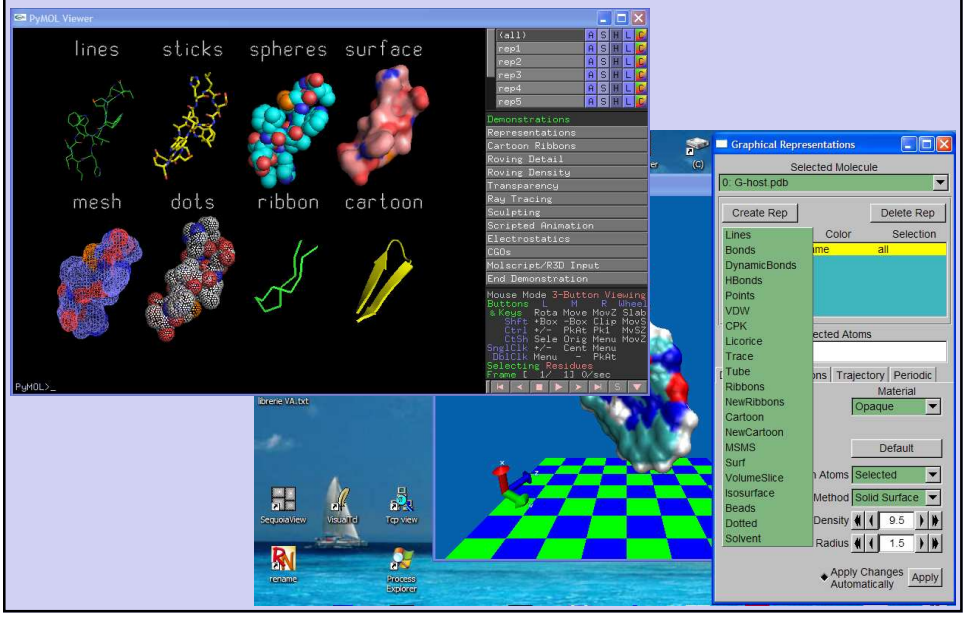
Give us a file format that provides a read/write library:

- Simple to use
- That supports all the information the user (do not know yet) needs
- That can be easily extended

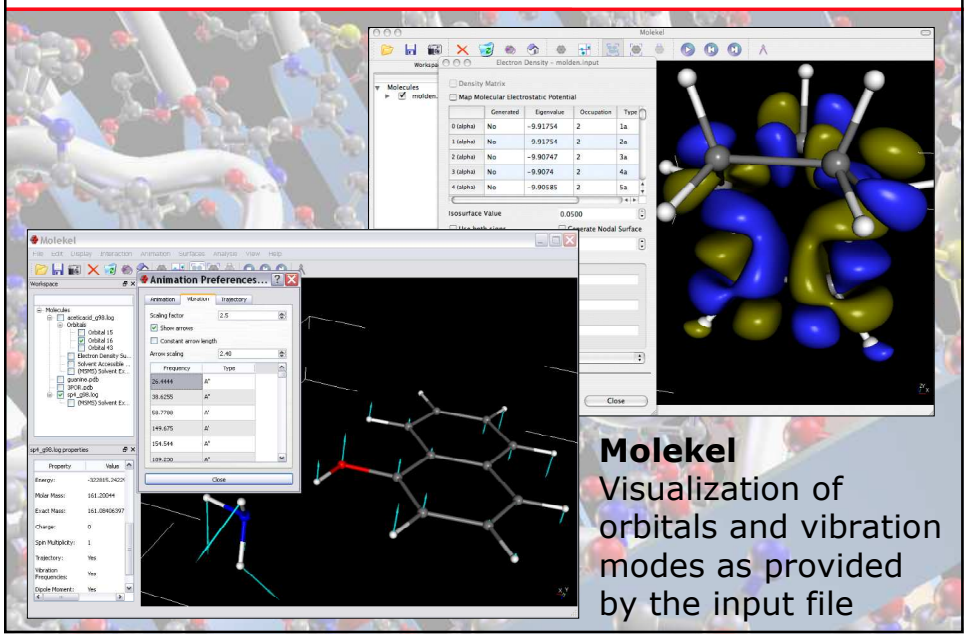
Representational conformism



Lots of tools, same representations



Show what the data provide



Molekel
Visualization of orbitals and vibration modes as provided by the input file

Extend the visual language!

The limits of my language

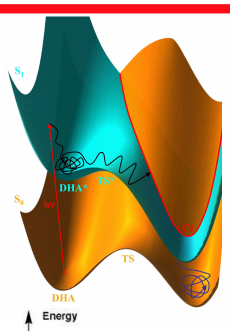
are the limits of my mind.

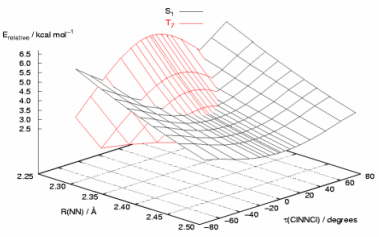
All I know

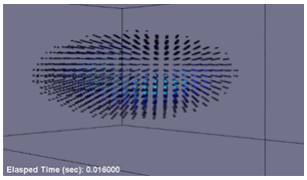
is what I have words for.

Ludwig Wittgenstein,
Philosophical Investigations, 1953

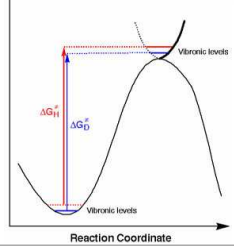
Quantum chemistry viz support

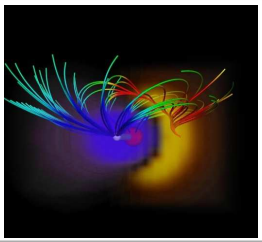


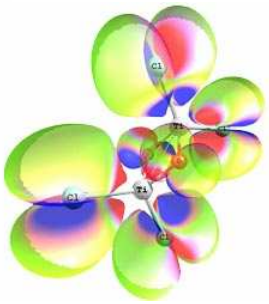




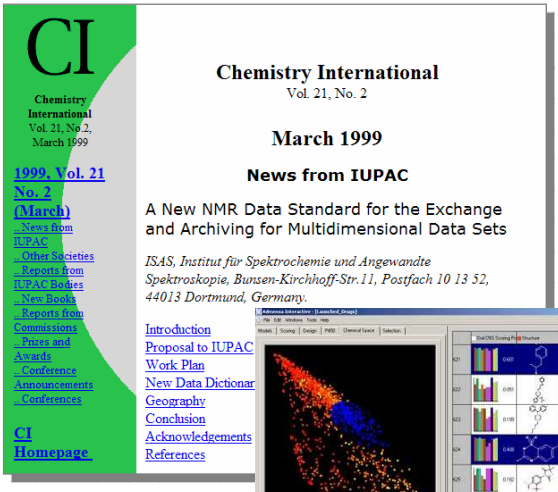
Dance of the Probability Fluid
Robert Wyatt (U. Texas)





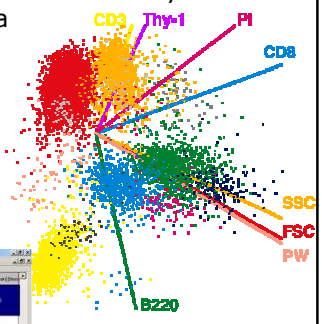


Multidimensional data support

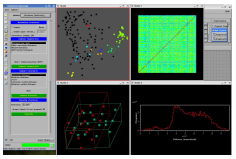


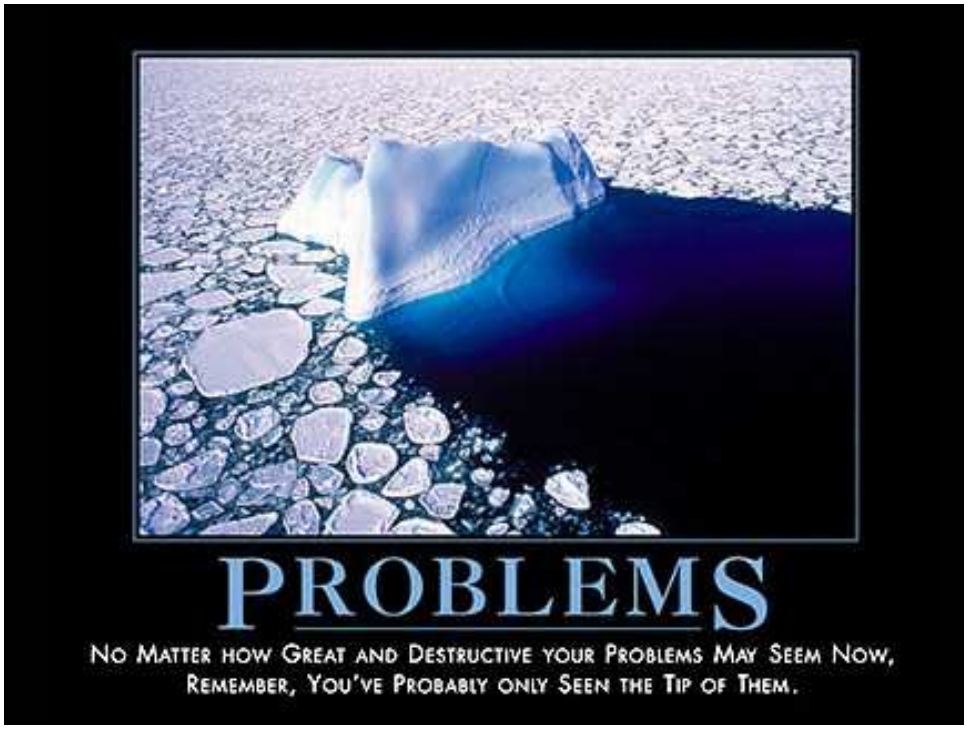
ADMEnsa Interactive

Biplot representation of blood chemistry data



Crystal fingerprinting





Summarizing our requests

When design a new data format, remember, it will be used (also) for visualization; and therefore:

- The users will outsmart you
- The visualization people will populate your nightmares
- The code developers will try not to use your libraries


But, always...
...we need your help!
Thanks!
BTW, I'm mvalle@cscs.ch