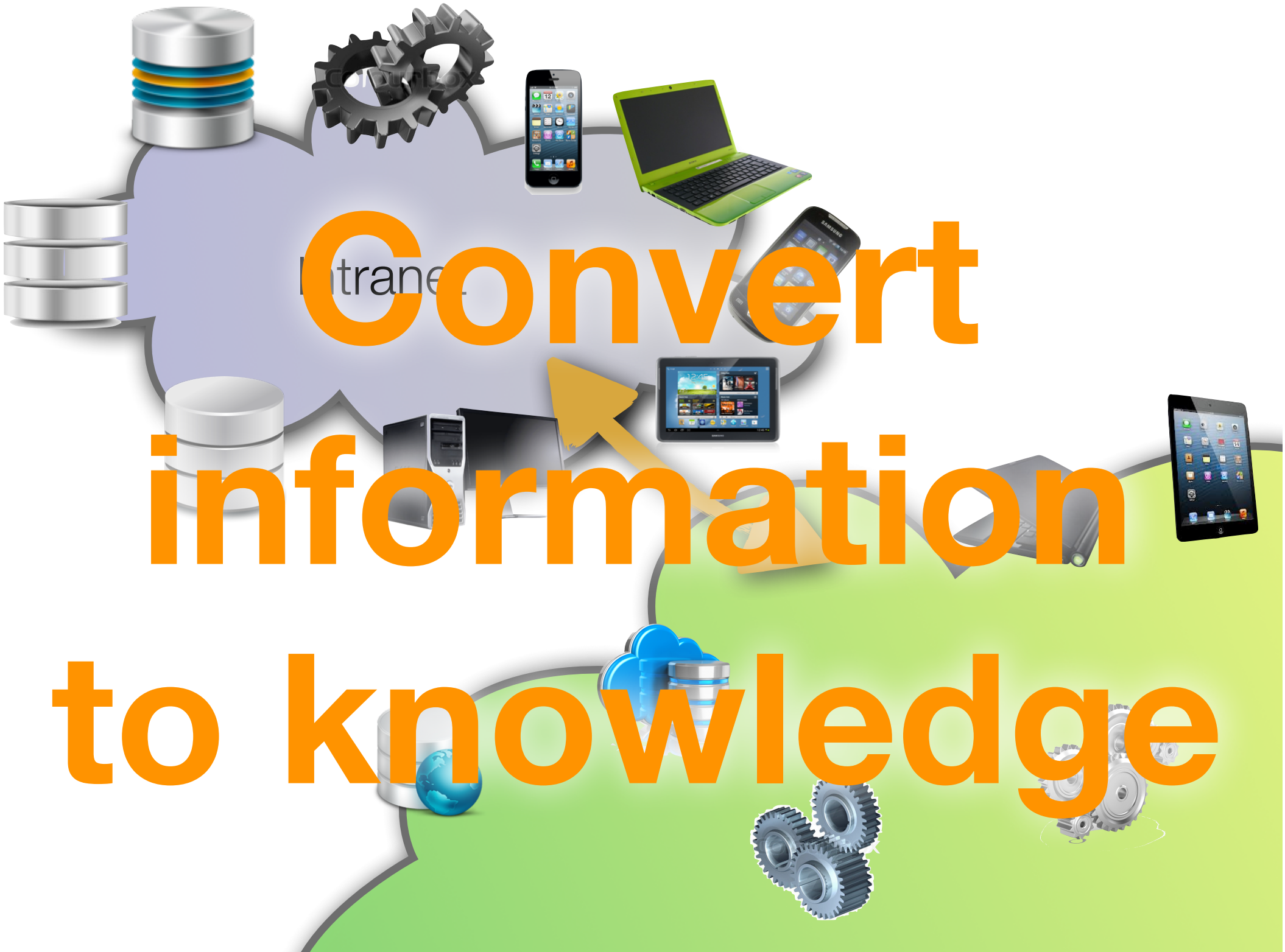
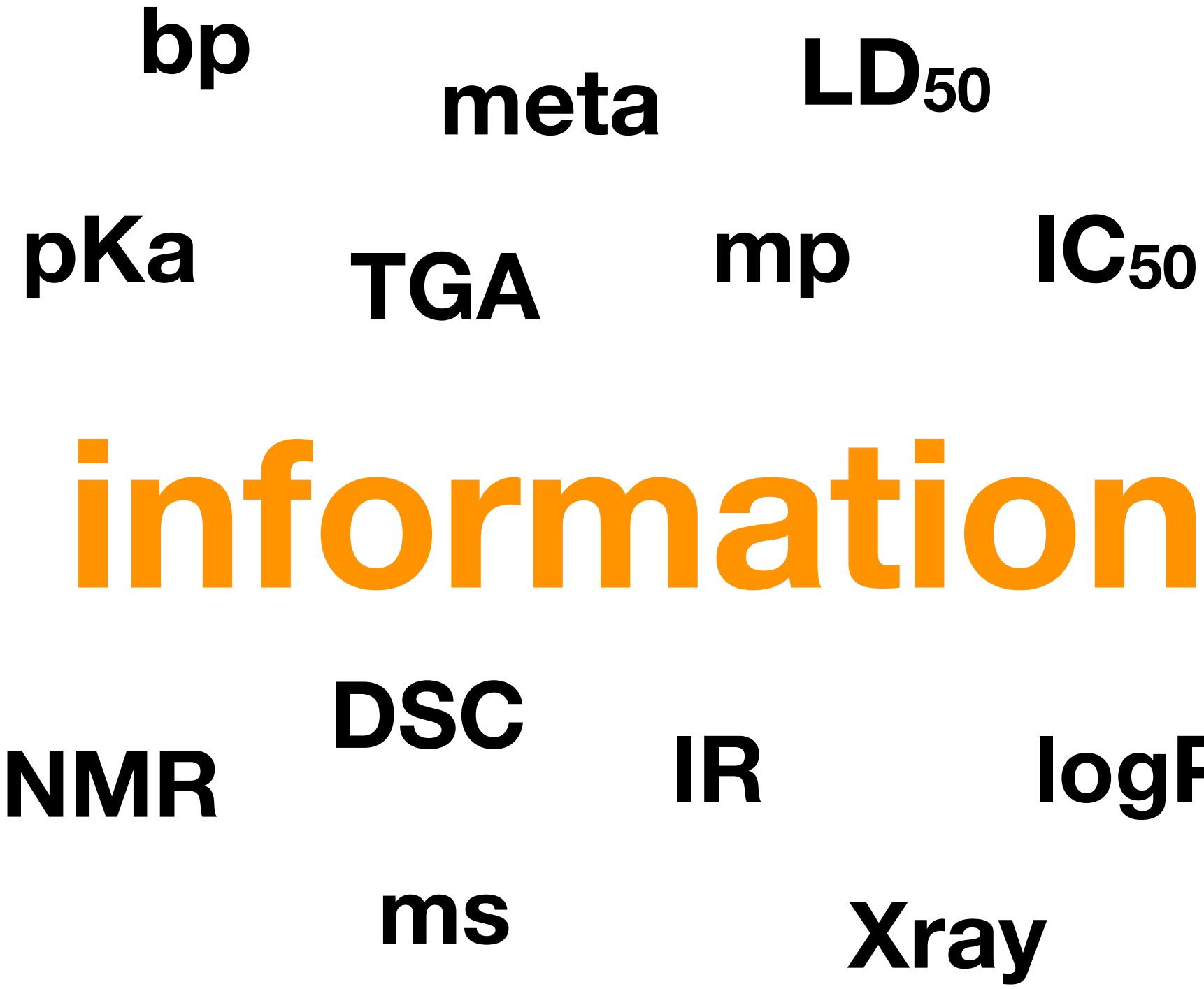
Why should we store chemical information ?

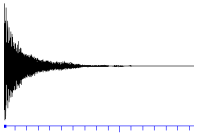
Luc Patiny
luc@patiny.com

ZAKODIUM


Intranet
Convert
information
to knowledge


bp
meta
LD50
pKa
TGA
mp
IC50
information
DSC
NMR
IR
logP
ms
Xray


physical property
QSAR
spectra
bond length
fragmentation
toxicity
illness
QC
reporting
knowledge



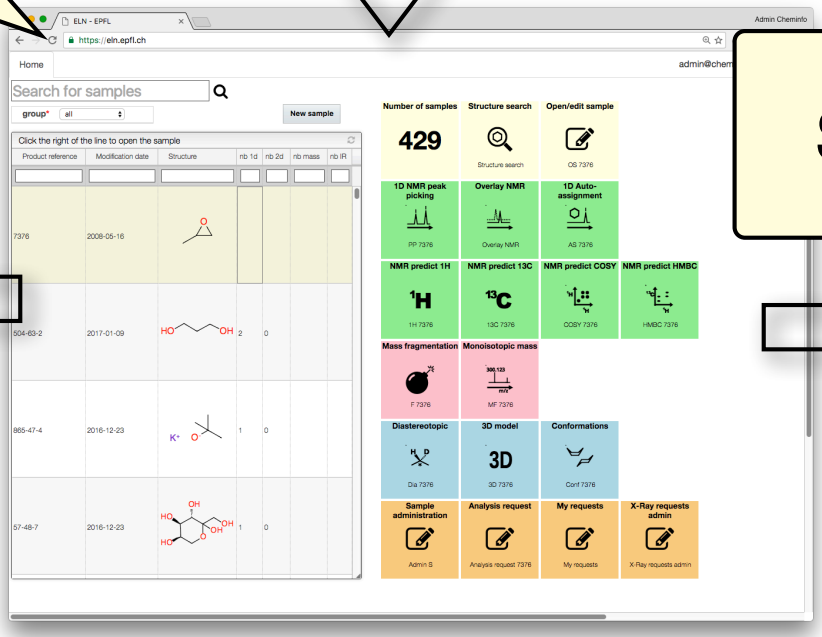
DB

STORE

www.c6h6.org

RETRIEVE

PROCESS



SHARE

Image analysis
Mass analysis
Similarity search
Predictions

EXTRACT KNOWLEDGE

…mental part
Original data
R
Sharing data
Zenodo
…

Store “correctly” the information

- Perennial and reprocessable !
- Only use **ALIVE** published **DESCRIBED** format
  - no proprietary format
  - no PDF, word, excel, …
  - no complex strings like 1H NMR (CDCl3): δ = 0.84 (3 H, t, J = 7.4 Hz, CH3), 0.94 (3 H, t, J = 7.4 Hz, CH3), 1.23 [6 H, q, J = 6.9 Hz, P(O)OCH2CH3], 1.51 (4 H, m, CH2), 2.20 (1 H, sextet, J = 6.6 Hz, CH), 3.80 (3 H, s, OCH3), 4.01 [4 H, m, P(O)OCH2CH3], 4.63 (1 H, d, JH,P = 17.1 Hz, NCHP), 6.88 (2 H, d, J = 8.6 Hz, CH).

Store “correctly” the chemical information

- bp : sign - low - high - pressure
  - 100-120 (20 torr) ; 10 (15 mmHg) ; >300 ; …
- chemical structures: molfile
- spectra: JCAMP-DX, netCDF, mzXML
- images: TIFF, PNG (no JPG)
- biological results with corresponding protocols

- Traceability, flexibility : noSQL database

Description of a Jcamp file

```
##TITLE= TP1
##JCAMPDX= 5.0          $$ Bruker NMR JCAMP-DX V1.0
##DATA TYPE= NMR Spectrum
##DATA CLASS= XYDATA
##XUNITS= HZ
##YUNITS= ARBITRARY UNITS
##XFACTOR= 0.545023283394666
##YFACTOR= 1
##FIRSTX= 4464.28571428571
##LASTX= 0
##DELTAX= -0.545023283394666
##MAXY= 361742655
##MINY= -541762
##NPOINTS= 8192
##FIRSTY= -162325
##XYDATA=(X++(Y..Y))
   8191.00000000       -162325       -149109       -131504        -88302
   8187.00000000       -201586       -173329       -175079       -155694
   8183.00000000       -145034       -102259       -179947        -80706
   8179.00000000       -106990       -116682       -168235       -114949
...
##END=
```

- LDR : Labeled Data Records
- 80 characters (or less) on one line
- Starts with Data Labels
- unlimited number of lines

Demonstrations

- **STORE**
- **RETRIEVE**
  - Find a product from homepage
  - Search by structure
  - Lipinski search
- **PROCESS**
  - NMR: assign / compare / predict / search
  - Mass: High resolution
  - IR
  - Chromatogram and GC/LC ms
  - Image analysis
- **Extract knowledge: PCA**
- **ELN**
- **EXPORT**
  - Report
  - Zenodo

www.C6H6.org

Advanced visualization platform

- Solves scientific problems in the browser
- Pure HTML5 / javascript
- Easily reusable javascript library
- Research, service and teaching
- Over 150 tools in 3 years

Thanks !

ZAKODIUM

ChemCalc

- Alain Borel (Library EPFL)
- Laure Menin
- Michael Krompiec
- Marek Noga

NMRdb

- Julien Wist
- Andrès Castillo
- Andrès Bernal

script, visualizer, components

- Norman Pellet
- Michaël Zasso
- Daniel Kostro
- Jefferson Hernández
- Miguel Asencio
- Julien Wist
- Andrès Castillo