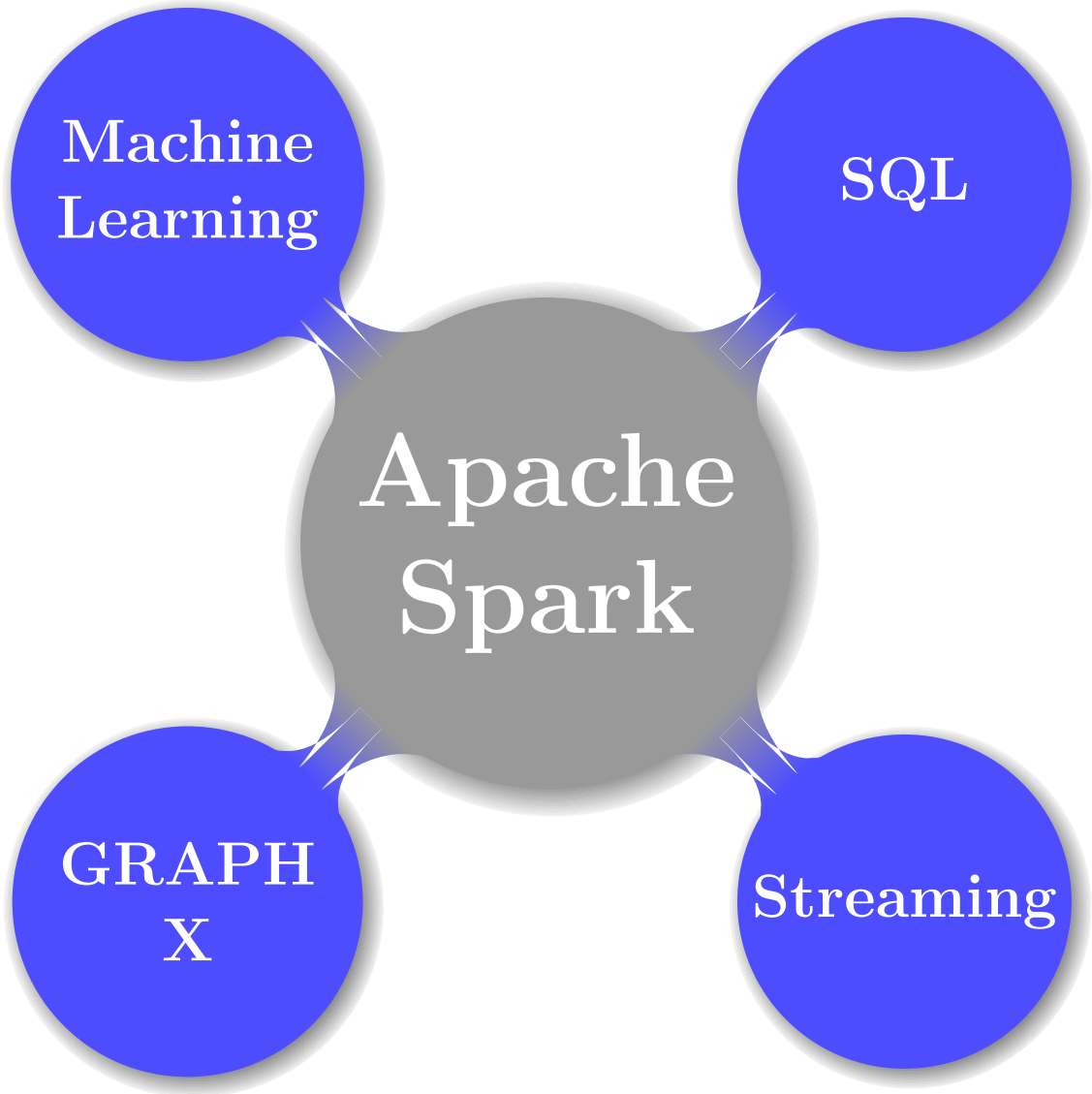
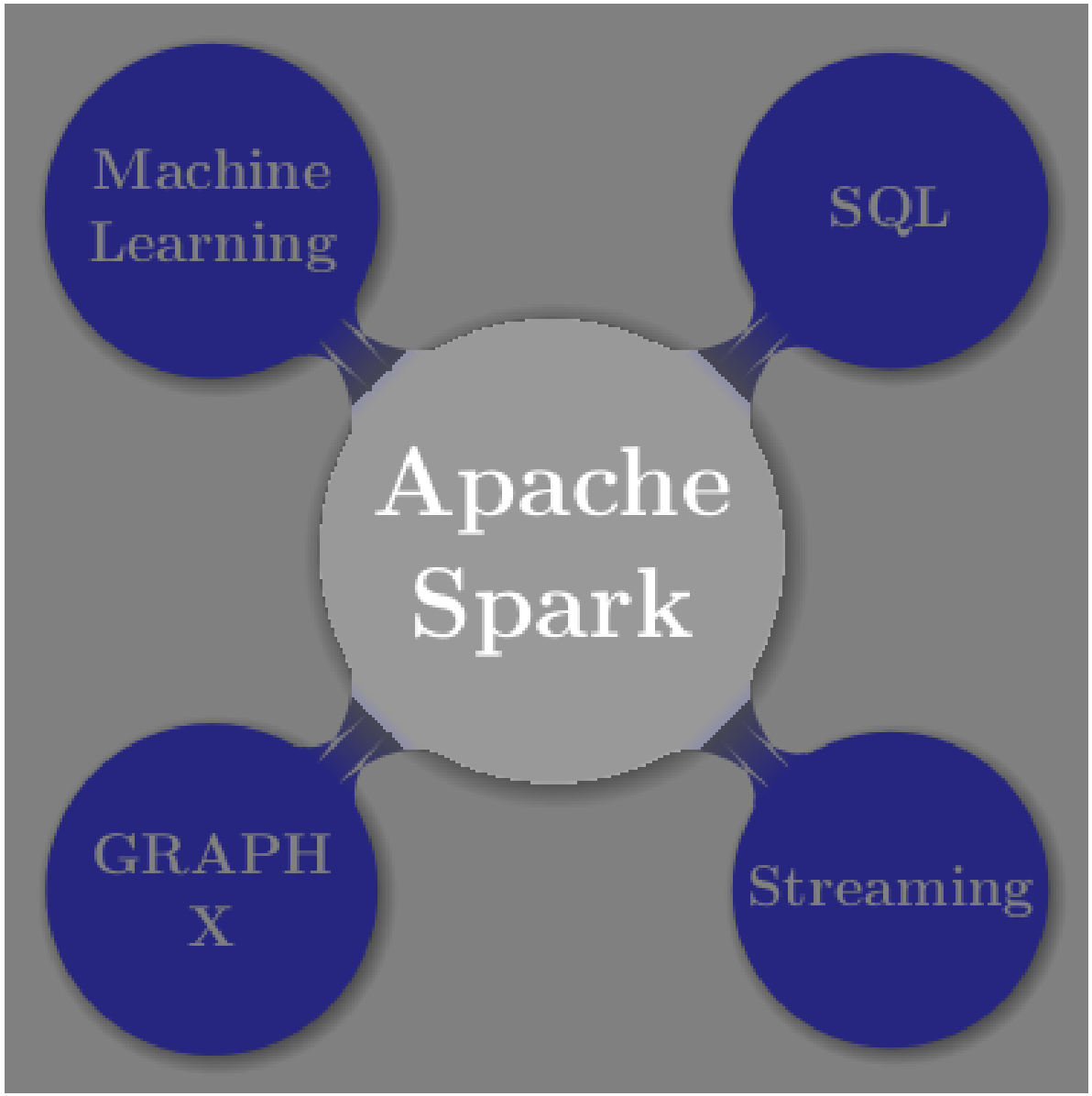
Overview on Spark activities

C. Arnault, G. Barrand, J.E. Campagne, J. Peloton and S. Plaszczynski

LAL, Univ. Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, Orsay, France

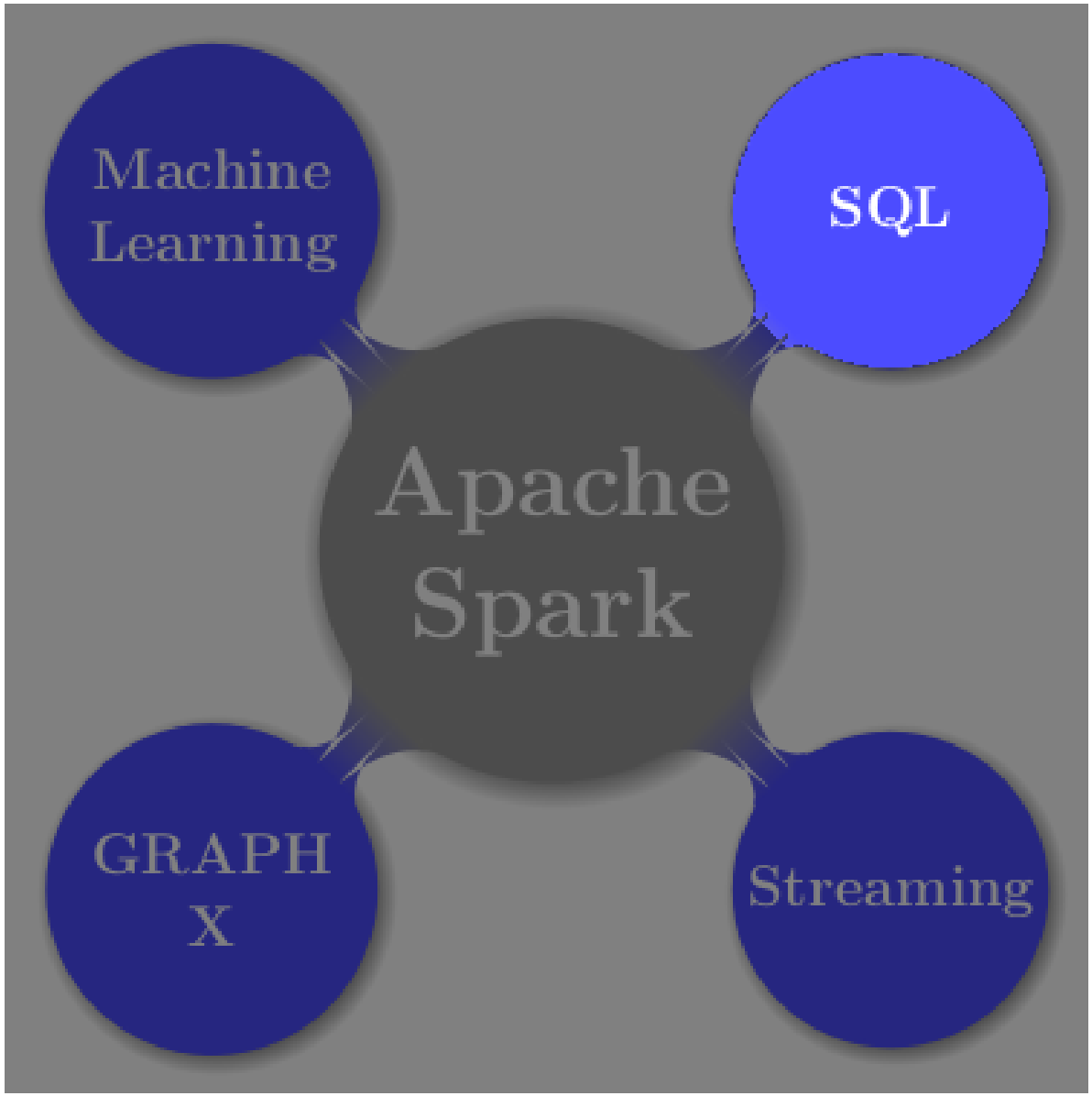
June 5, 2019


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

spark-fits

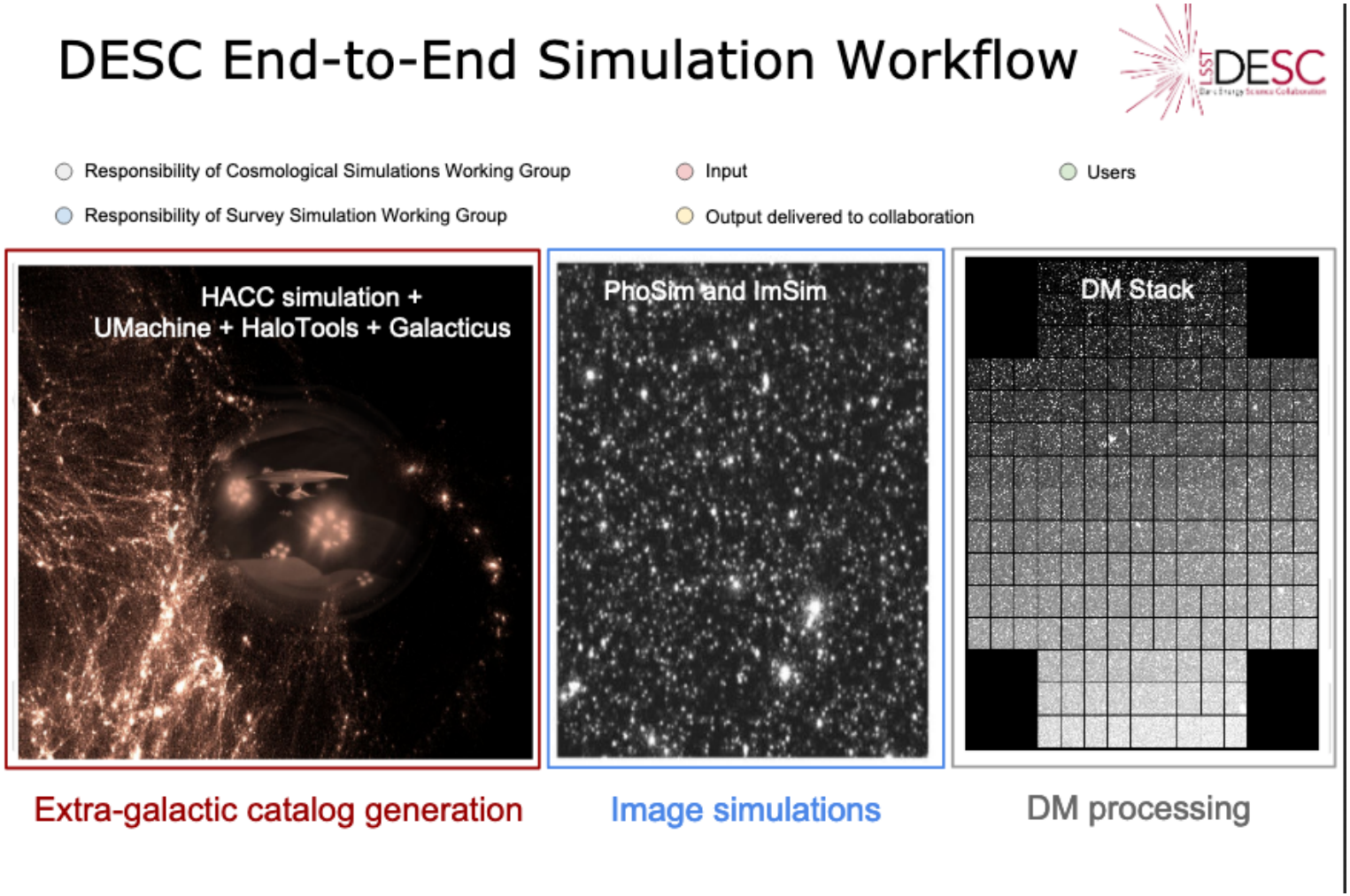
- BinTables OK: used, good performances, some issues reported+ cured, to be upgraded for `scala 2.12`
- Images available, no compression
  - stack: in contact with DM (@Washington), wait for new Butler
  - commisioning?


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

Data analysis

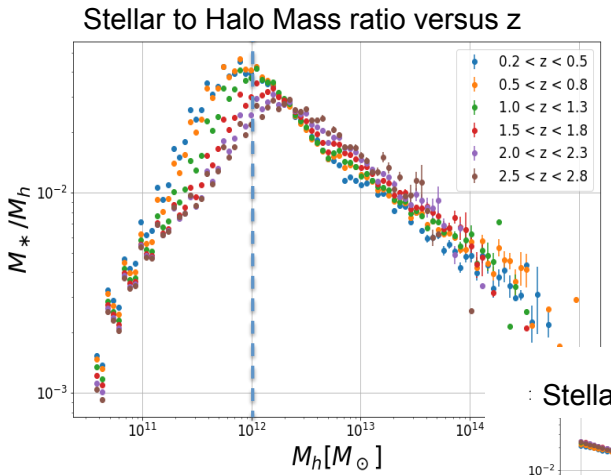
- on CoLoRe LSST10Y (6 10^9 gals) : histograms in $\simeq 10s$, tomographic shells in $\simeq 30s$
- ArXiv:1807.03078 + notebook https://github.com/astrolabsoftware/1807.03078
- R&D on distributed 2-point function computation (Ahmed)

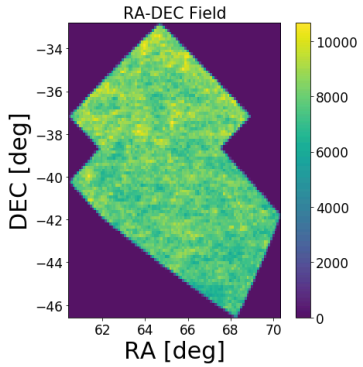
DC2

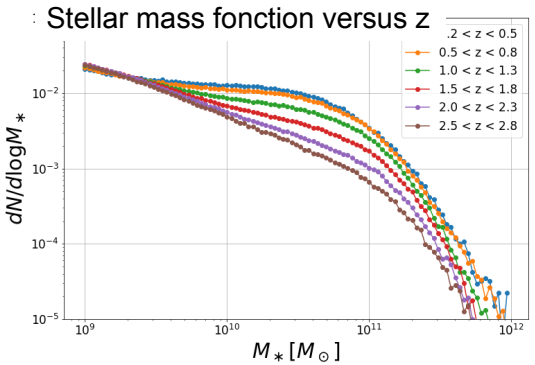


K.Heitman Lyon 2018

cosmoDC2







Volume: $(4.2\text{Gpc})^3$
Résolution m=2.610^9 M_{sun}
~10^{12} particules

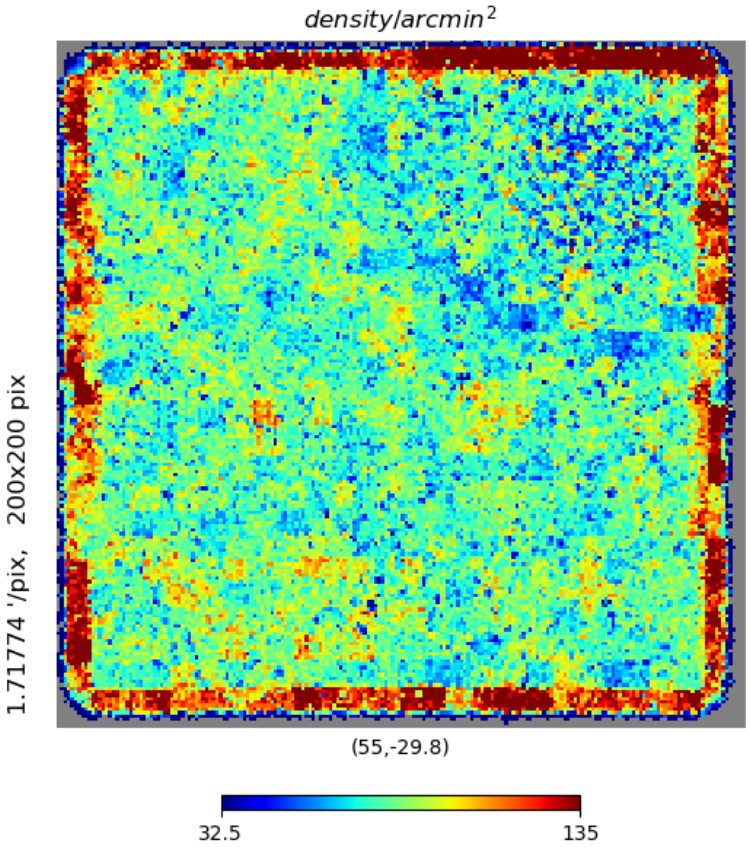
1er catalogue de galaxies: 440 deg^2, z<3
(plus tard: 5000 deg^2)

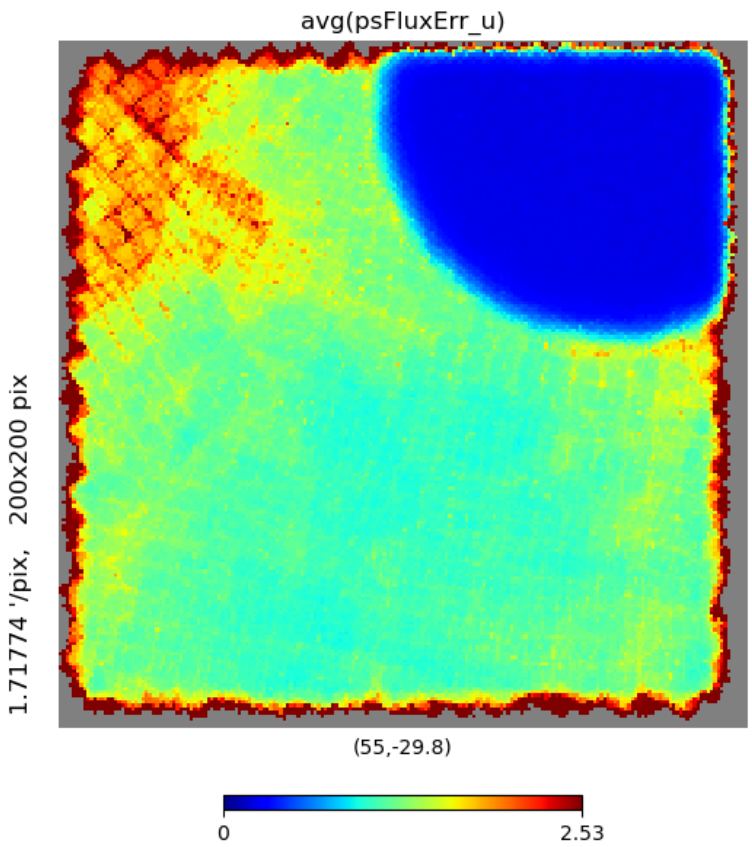
J.E Campagne 2

see also JP talk.

Object catalog (DPDD)

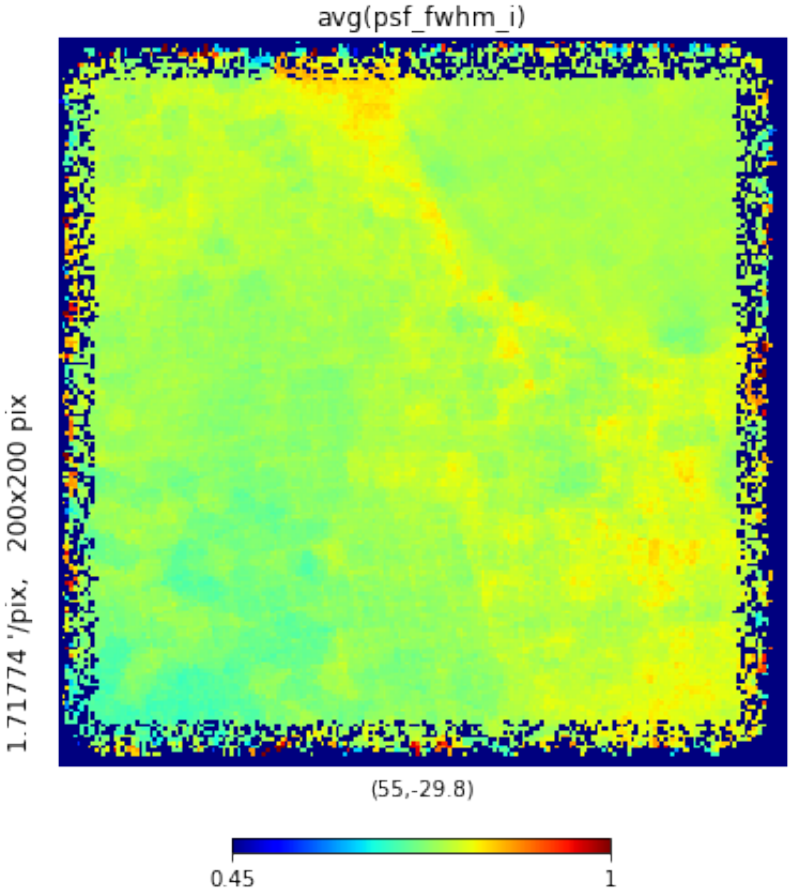
(very simple+efficient with Spark) Run1.2p

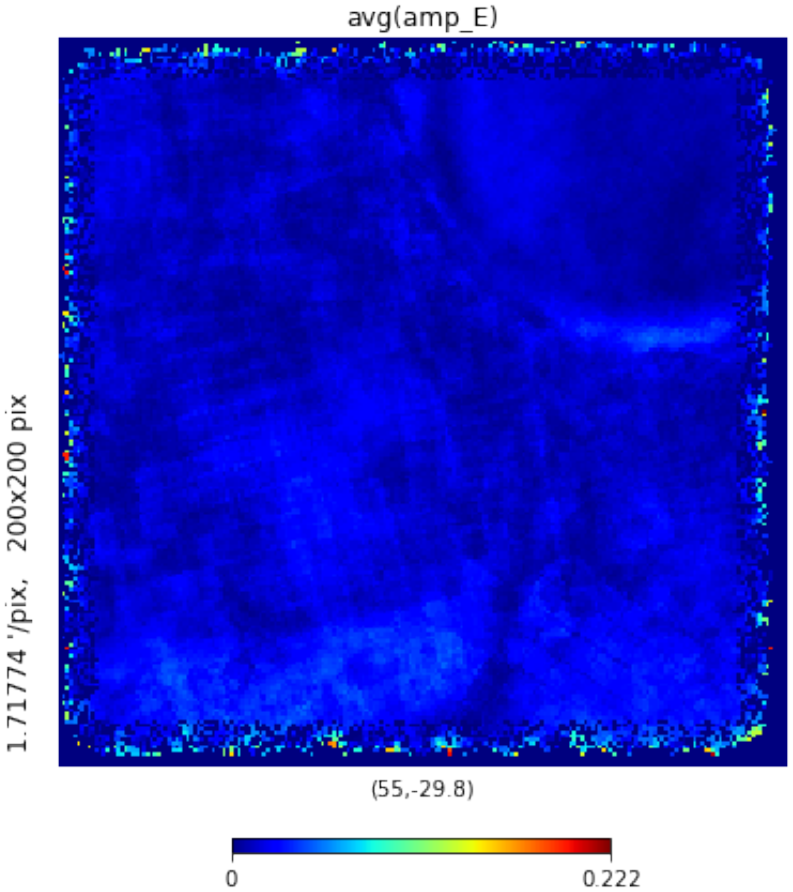




https://github.com/LSSTDESC/DC2production/issues/323

Syst. maps for 3 × 2pt TF

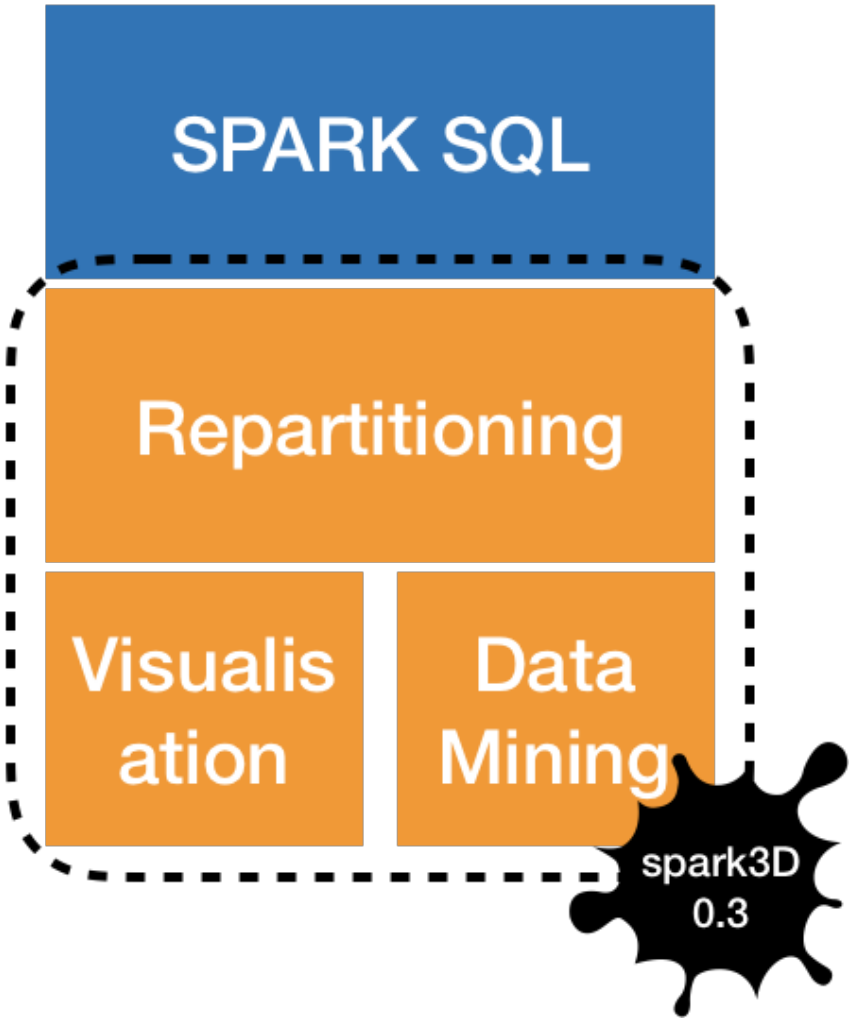




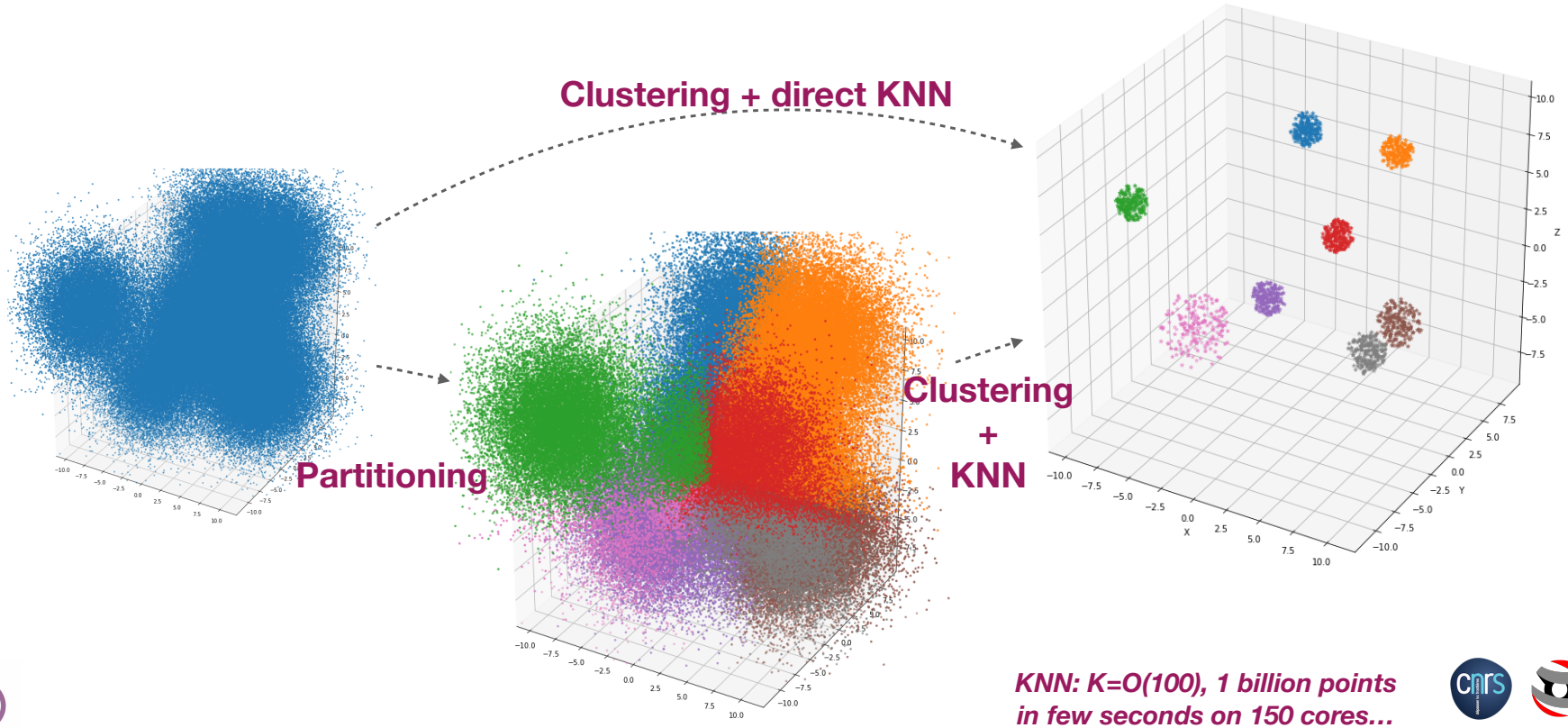
https://github.com/LSSTDESC/TXPipe/blob/master/examples/Syst3x2pt.ipynb

+ `TXSelector` sparkization

Spark-3D


SPARK SQL
Repartitioning
Visualis
ation
Data
Mining
spark3D
0.3

spark3D: K Nearest Neighbours



KNN: K=O(100), 1 billion points in few seconds on 150 cores...

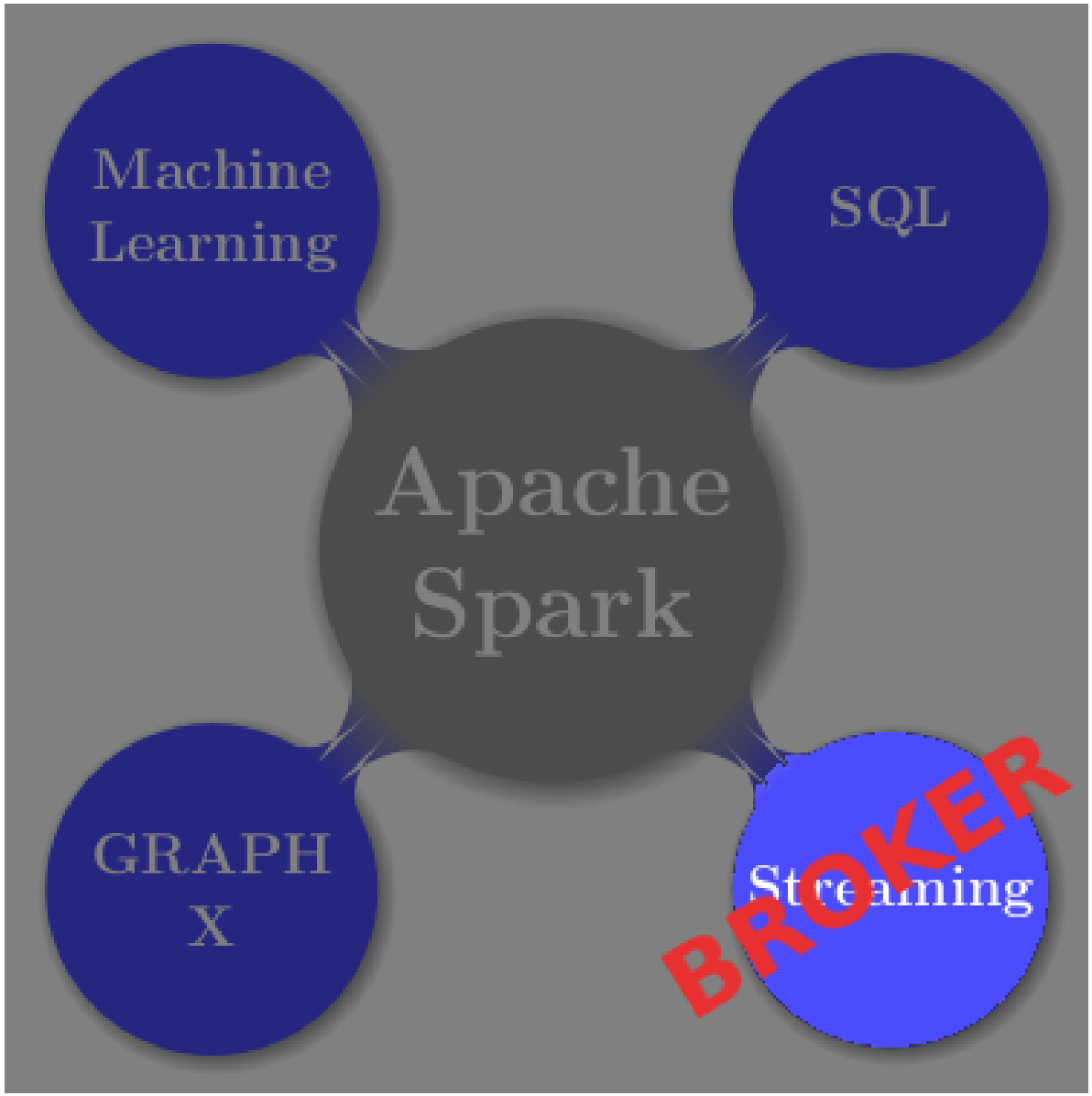
clustering algorithm =`k-Means` →`DBSCAN` : recoded in `scala`

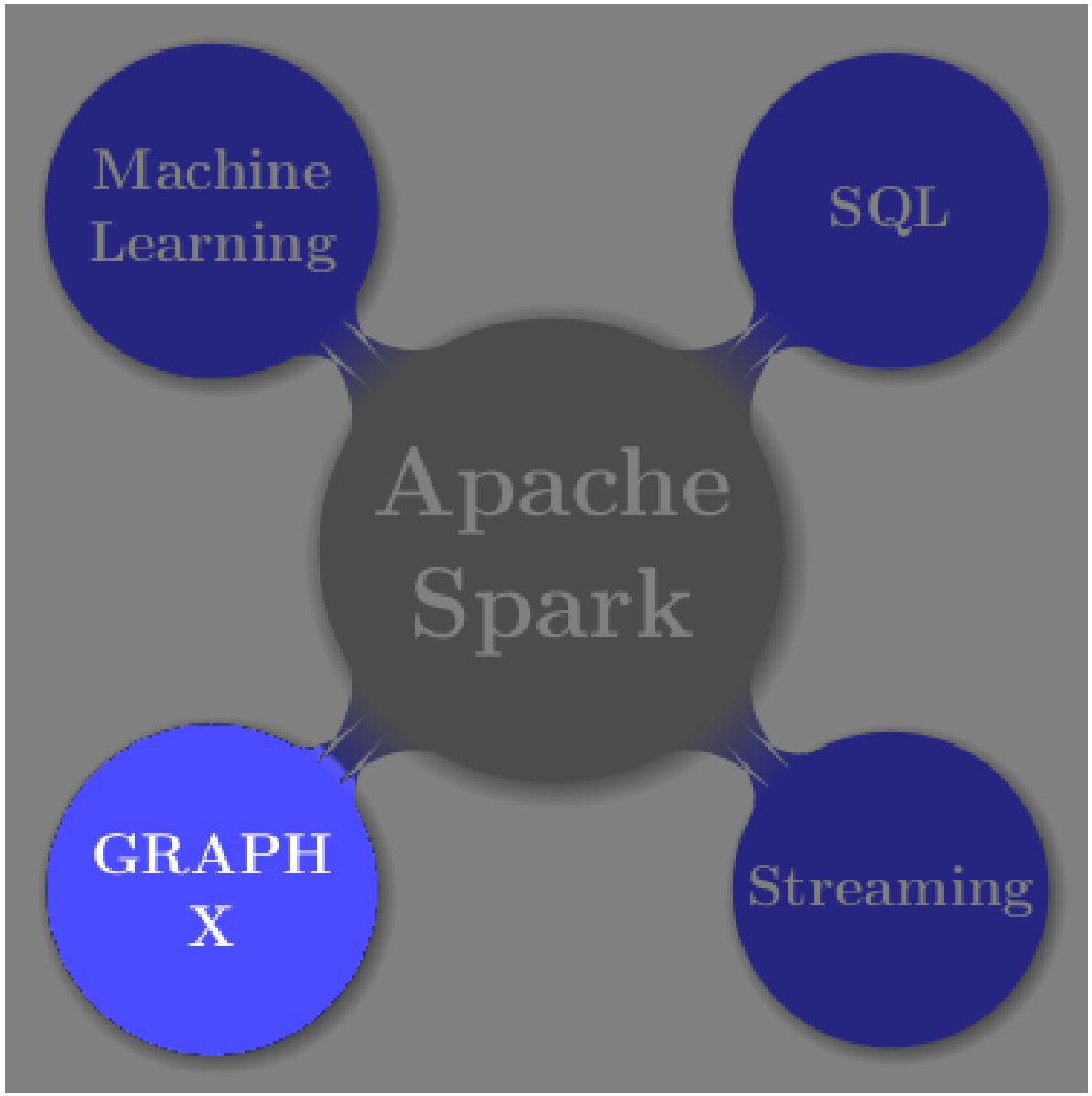
Catalogs cross-match (join)

- 2D simple with `HEALPix` indexing →DPDD×Truth (redshift)
- 3D not that simple: key is `Spark` 3D partitionning (octree)

performances are great/amazing/awesome/...

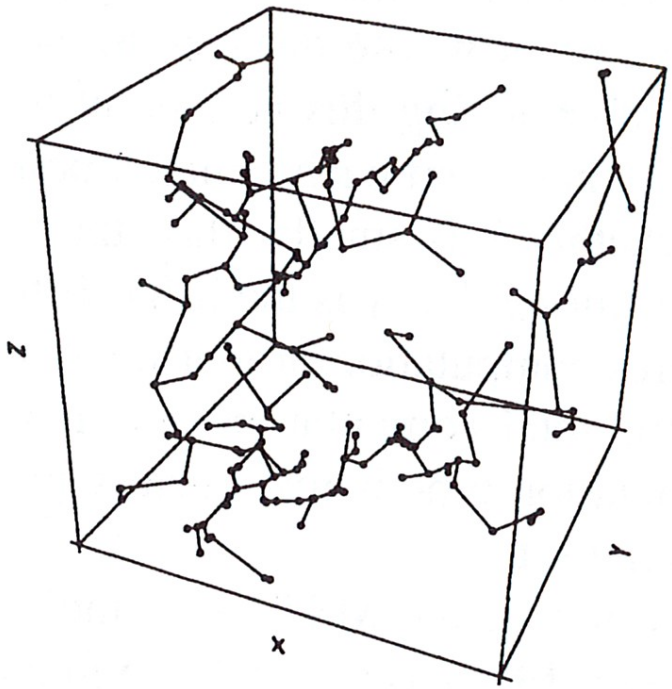
→paper (multi-cell analysis)


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming
BROKER


Machine Learning
SQL
Apache Spark
GRAPH X
Streaming

Beyond 2pt statistics

- shape statistics
- topological characteristics
- Minimum Spanning Tree



Formation

- DC2-production/DC2-analysis notebooks
- `AstroLab` tutorials
- `LSSTDESC/desc-spark`
- Ecole IN2P3 d'informatique (programation fonctionnelle,16-20 sep), ecole Euclid (19-31 aout)

Visualisation

- *inexlib* software : core(C++)+ `OpenGL` rendering + `python` (`SWIG`)
- `Linux, Mac, Windows,iOS, Android` : laptops, mobiles, tablets, wall of screens...
- client/server mode OK
- bi-directional ports access with `openssh`
- galaxy colors, sizes (ongoing ellipticities)
- standalone (but interfaced to `Spark`)

Wall of screen