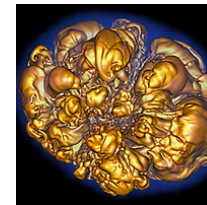
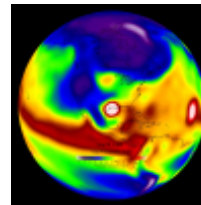
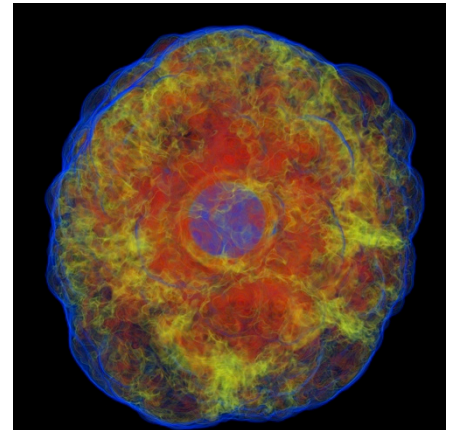
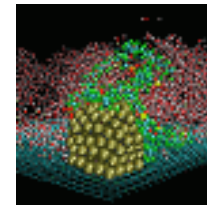
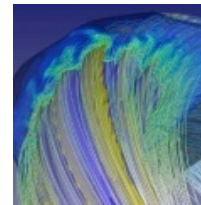
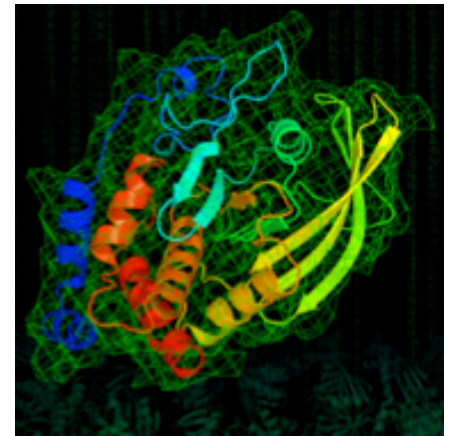
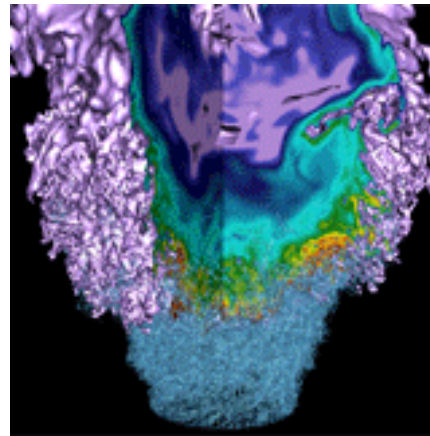


SciDB @ NERSC



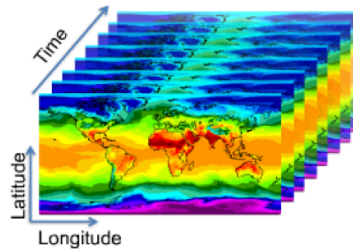
Yushu

Array Like Science Data

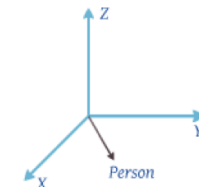
– More common than you think

Climate Simulation Output

-Terabytes of Output Per Run



Brain MRI Image

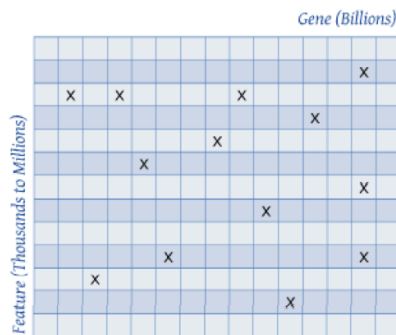


Many of people -> 4th Dimension

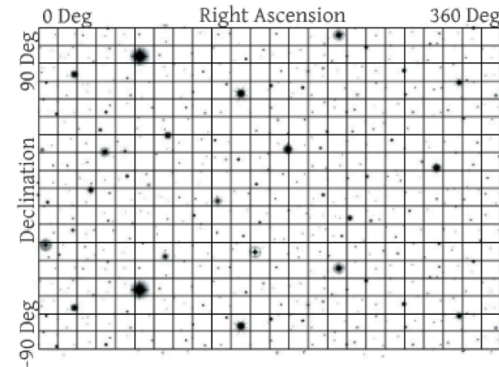


Gene Labeling

- Large Sparse Array



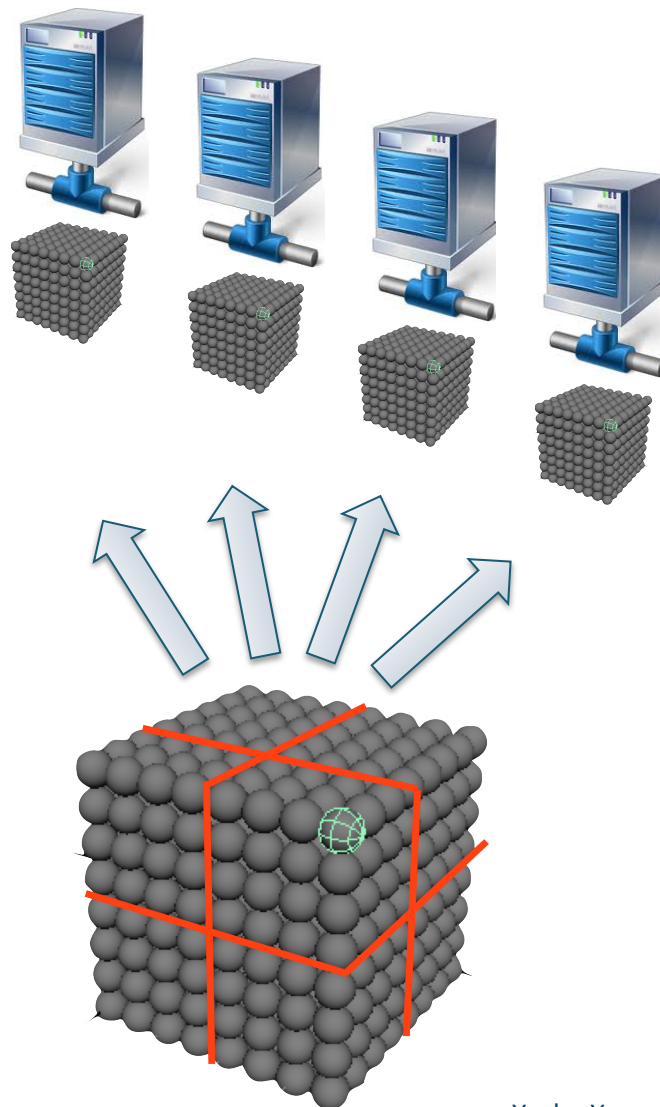
Catalog of Billions of Stars



SciDB, parallel processing without parallel programming

Everything in Arrays

- Locate an element at $O(\text{constant})$
- **Can be very sparse**
- Best for machine/simulation generated structure data
- Good for metadata too
- **Query-like language, auto-parallelization**
- **Do Calculations inside the DB**

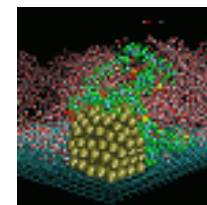
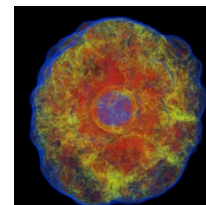
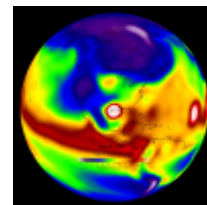
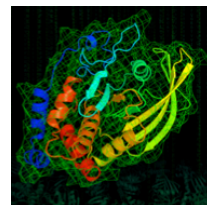
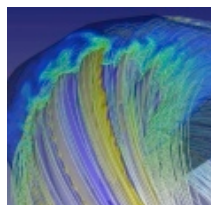
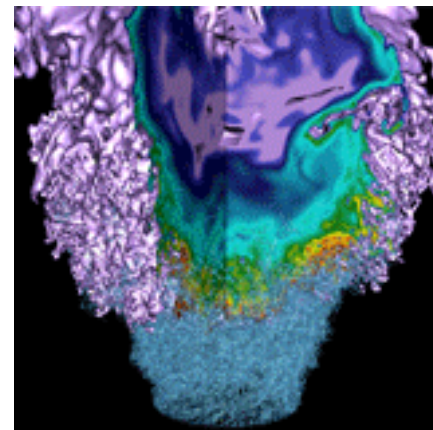


- **Partner up with Science Teams**
 - Hold their hands to load the 1st batch of data and implement the 1st major analysis operation
- **15+ Science Projects**
- **Complicated Workflows and Algorithms**
- **Multiple Science Domains:**
 - Astronomy, Climate, Bio-imaging, Genomic
- **Multiple Types of Data**
 - Spectrums, Images, Time Series
- **Large amount of data**
 - Normally 100GB-1TB, some has 5+TB

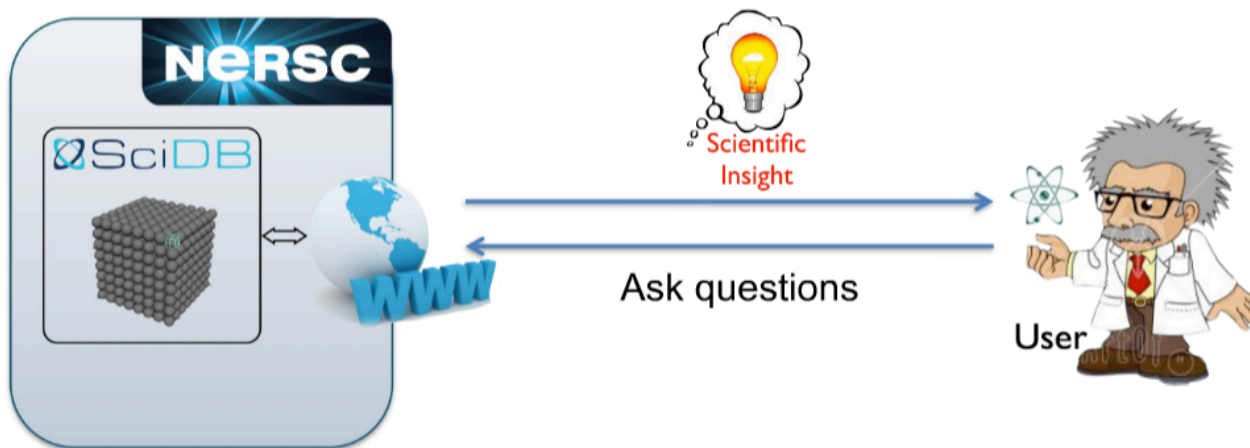
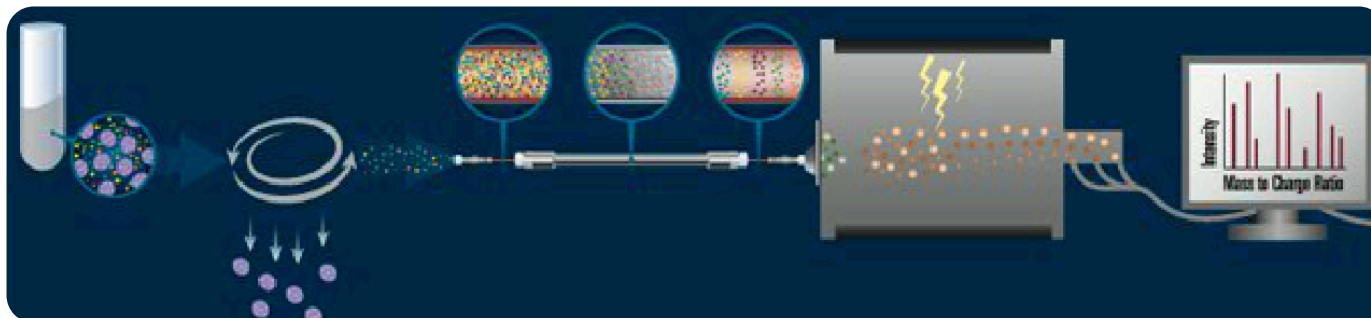
Types of Data Suitable for SciDB

- **Imaging data: digital pictures from light sources or telescopes**
- **Time series data collected from sensors**
- **Spectral data**
- **Graph-like structures that represent relations between entities (sparse matrix)**

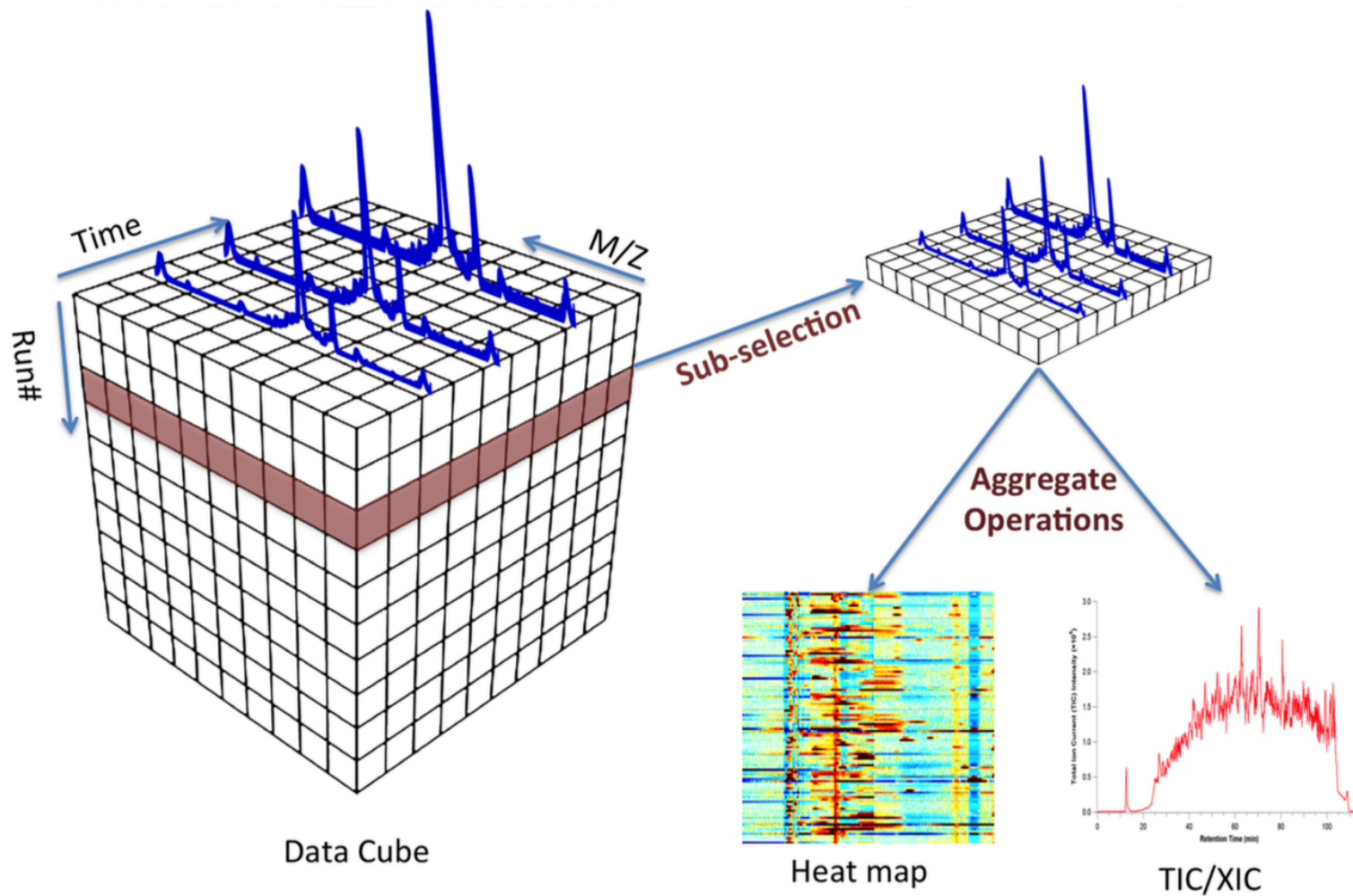
Examples



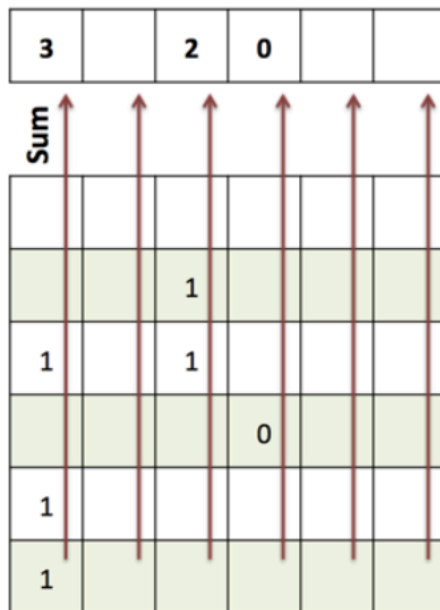
MetAtlas (LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY)



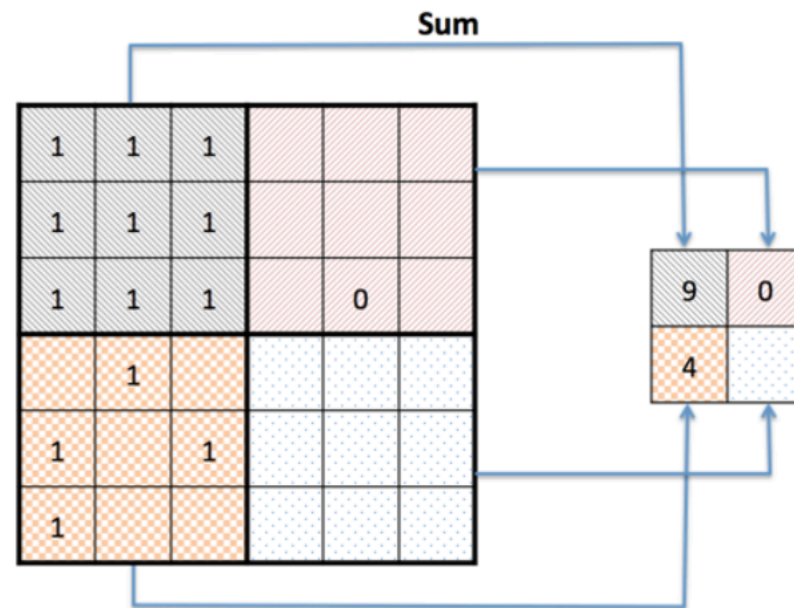
MetAtlas (LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY)



Some Primitives

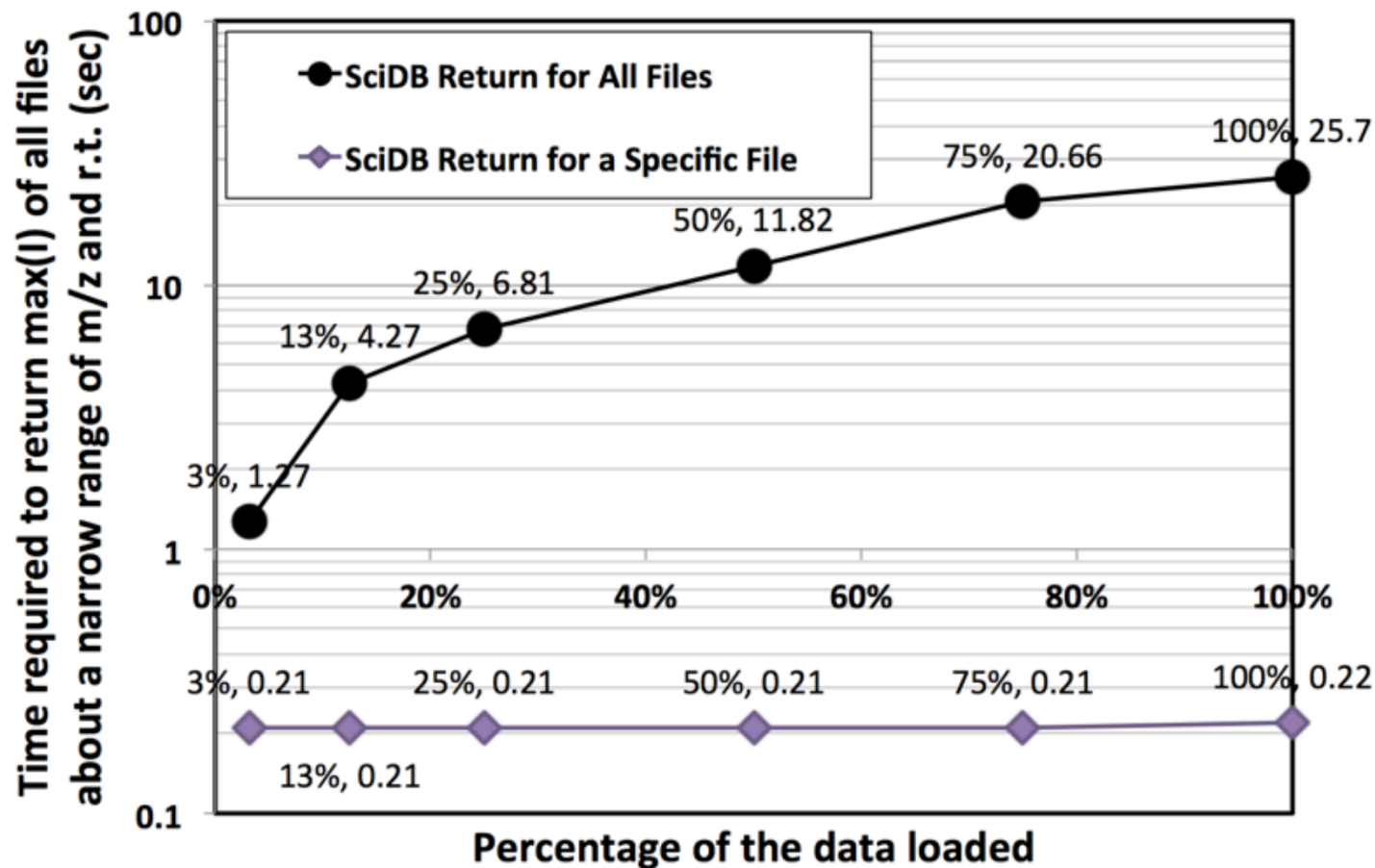


Aggregate along one dimension

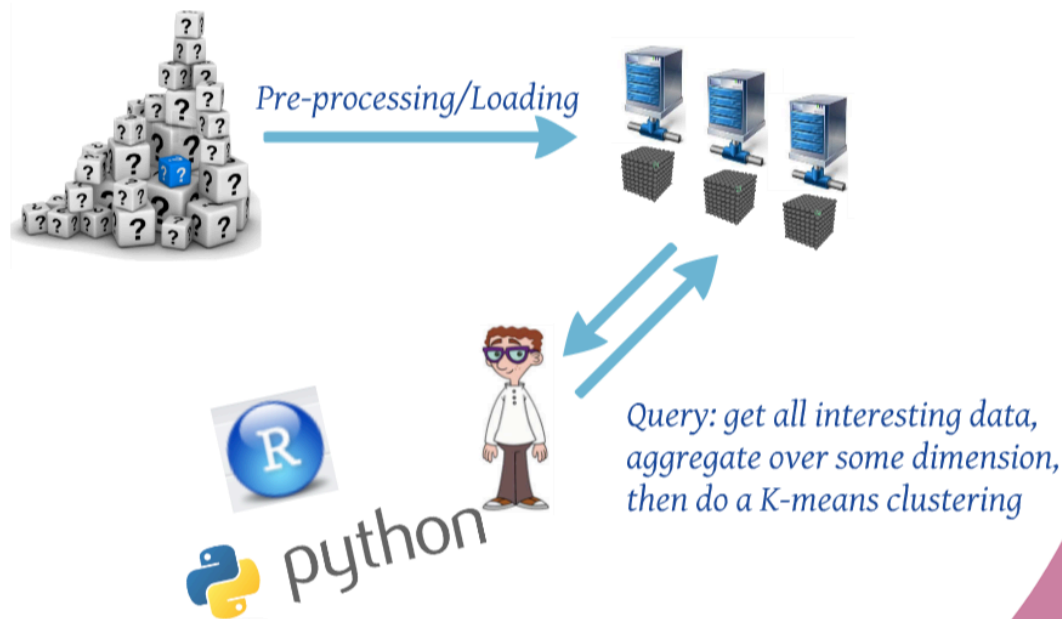
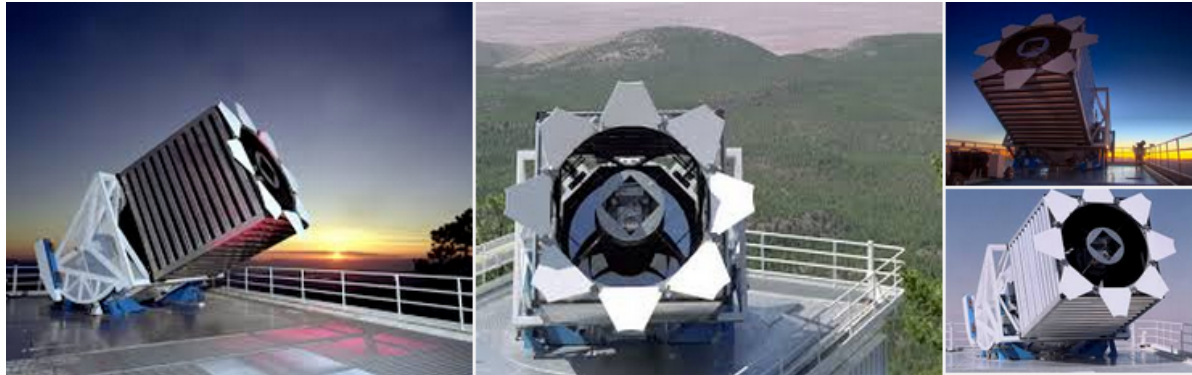


Aggregate by re-gridding

Benchmark of MetAtlas Workload



DustOff Workflow



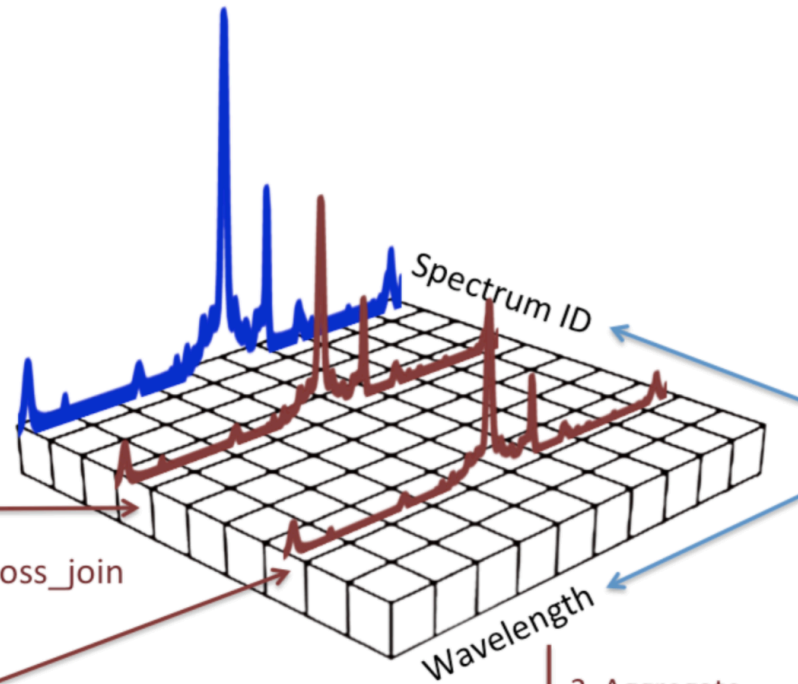
DustOff In SciDB

Collection of Spectrums

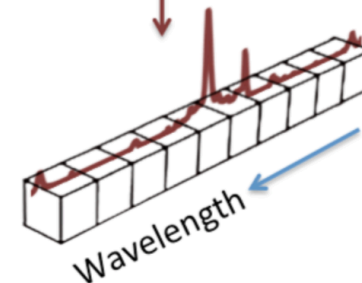
1. Sub-selection:
filter all spectra with
"meta data 1=100"

Spectrum ID	Meta Data 1	Meta Data 2
0	2	"a"
1	2	"b"
2	5	"c"
3	100	"d"
4	1	"e"
5	23	"string"
6	5	...
7	100	...
...	...	...

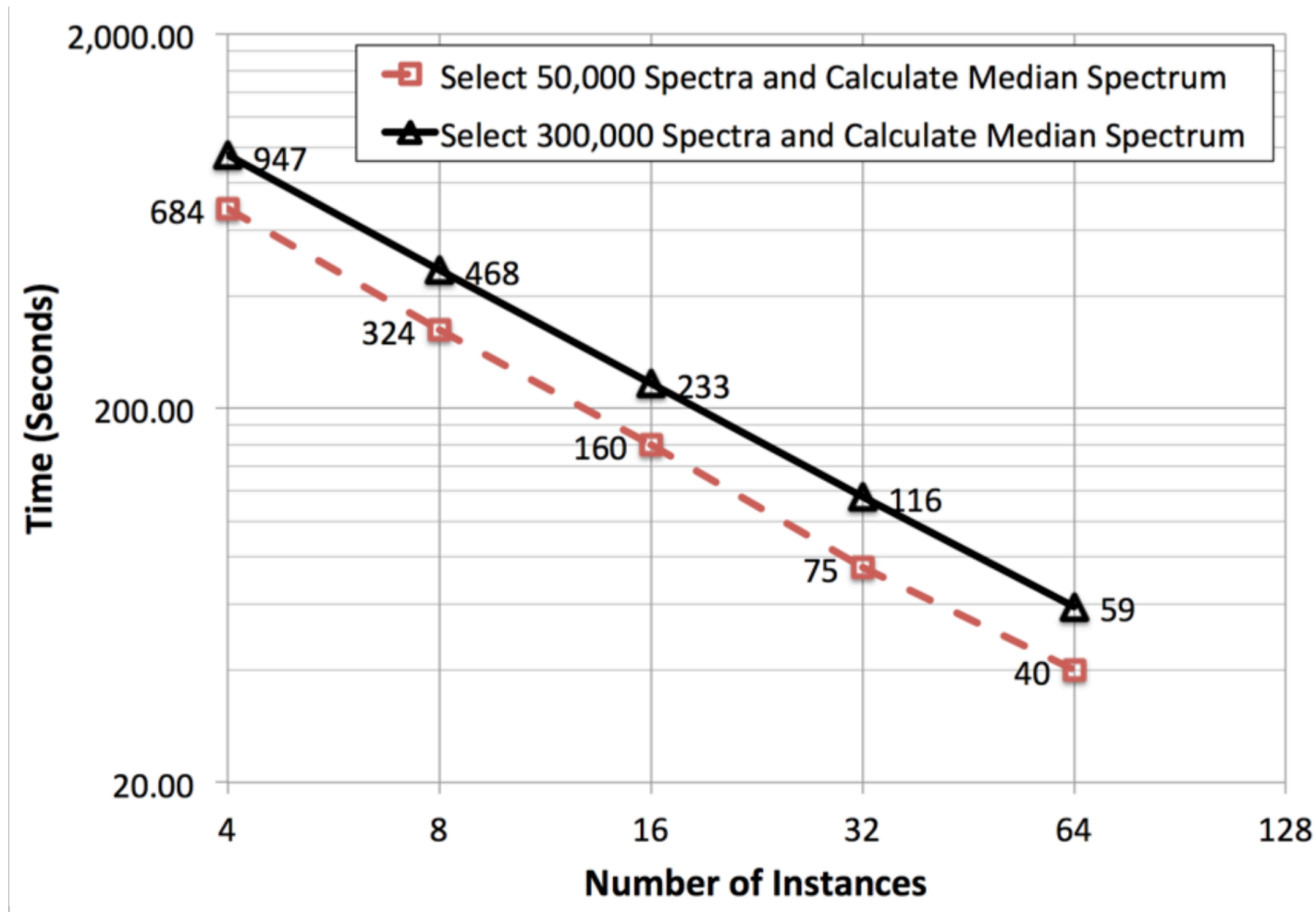
2. cross_join



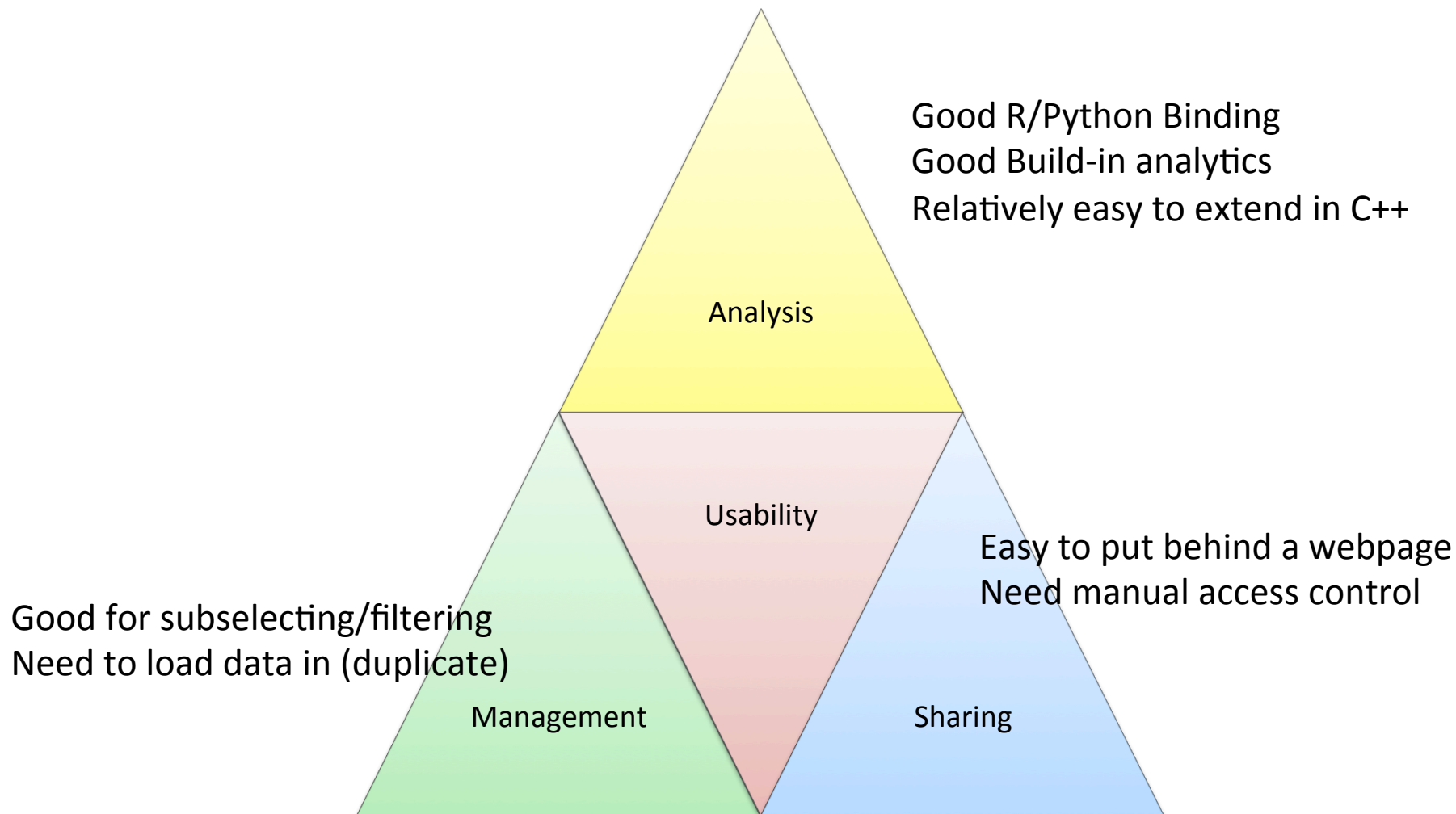
2. Aggregate
Median Calculation



DustOff Scaling



Strength/Weaknesses of SciDB



When to Use SciDB on Your Data

- 10+GB Data (and will grow)
- Looks like an array (dense or sparse)
- Write Once (or accumulate slowly) and read a lot
- Lots of Filtering and Aggregating
- Want to do Joins like SQL
- Do most of calculation inside the Database
- Linear Algebra on your Data

When not so use SciDB

- SciDB is for Analysis, NOT Transactions
- For crunching through large data to return 1 small result, but NOT to return millions of small results at high throughput

New SciDB Service at NERSC



- **Dedicated Servers**
- **Production Ready**
- **To Request SciDB:**
 - <https://www.nersc.gov/users/science-gateways/science-database-request-form/>