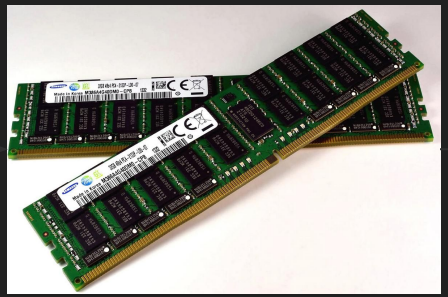
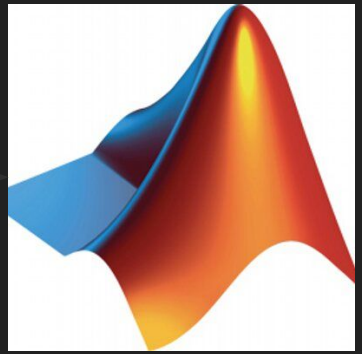
From Cosmos to Connectomes: The Evolution of Data-Intensive Science

By Randal Burns, Joshua T. Vogelstein, and Alexander S. Szalay

Presented by Ryan Marren

Summary: Moving past ‘Single Use Science’





- Worked quite well in the latter quarter of the 20th century.
- Cosmology stressed this paradigm, followed by Genetics
- Now Neuroscience is facing the same problem

Opportunity: Other fields have already overcome

- The Sloan Digital Sky Survey
- The first big open science initiative
- Inherently spatial data at large scales (~132 TB)
- User freedom to design queries (like Apple allowing external developers in App store)

- The Institute for Data-Intensive Engineering & Science
- Originators of SDSS
- Have launched ~20 open science initiatives since, expanding on SDSS (~10PB of data)

idies

Challenges: …but the brain is much different...

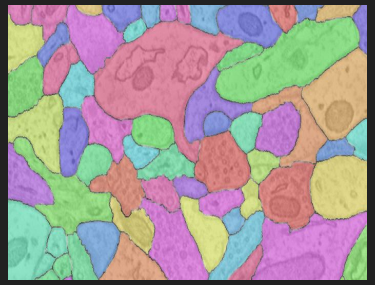
The Curse of Dimensionality:

- SDSS is inherently 2D, while connectomics data is *at least* 3D (multispectral, multi-resolution, time)

Scale
- SDSS ~132TB
- OCP already had multiple 100TB projects, 1PB expected in near future

Computer Vision
- 2D easier (simple stitching, color correction)
- 3D harder (3D stitching, tracking objects across layers)

Other Challenges



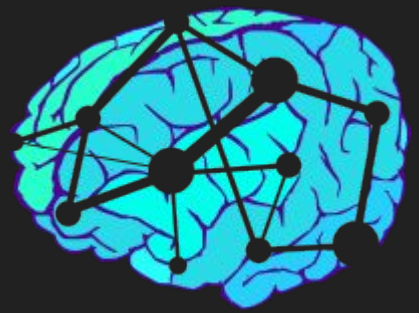
Write Intensive

- Many ways to represent and annotate connectomics data
- Analysis dependent on user input (read-write versus read only pipeline)

Heterogeneous Computation

- Many processes are computation specific (GPU acceleration for image processing, for example)
- “bring the computation to the data”
- data center + super computer

Actions: The Open Connectome Project



- Descendent of the SDSS
- “take raw, large, interesting neuroscience data sets and make them ready for wide-ranging analysis by anybody with interest (and internet).”

Results: Lessons Learned

Re-use is key!!

- Communities thrive when in the mindset of re-use from the outset
- Allows experts to jump in at any point in the pipeline

Smart data storage

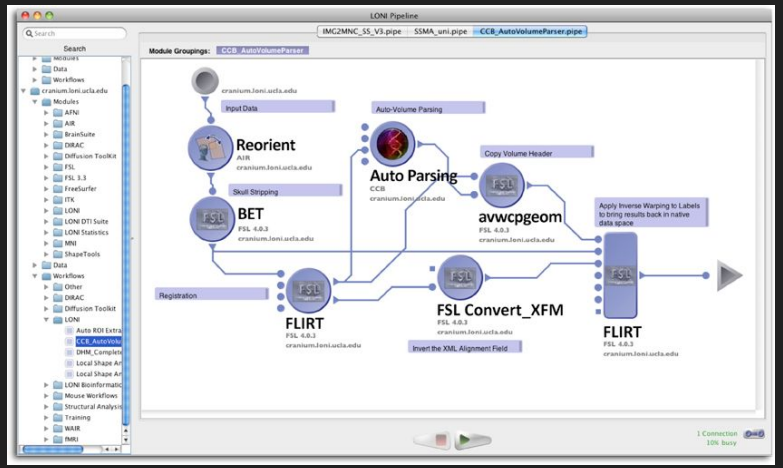
- Store data in a way that makes scientific inquiry easy
- e.g. spatial data storage in OCP and SDSS
- Ingest many file types

Results Continued...

```
2346  NEXT PAUSE
2347  DLY% = MDY% / 80
2350  X = 1
2360  Y = Y + 2
2370  STRING$ = "ALONE IN THE PARK
2380  FOR I = 1 TO  LEN (STRING$)
2390  CHAR$ =  MID$ (STRING$,I,1)
2400  GOSUB 9000
2410  NEXT I
2420  CHAR$ = " "
2430  DLY% = MDY%
2440  FOR I = 1 TO 3
2450  GOSUB 9000
2460  NEXT I
2530  REM  XX = 5
2540  Y = Y + 1
2545  DLY% = 0
2550  STRING$ = "I CALLED YOU A LI
2560  FOR I = 1 TO  LEN (STRING$)
2570  CHAR$ =  MID$ (STRING$,I,
```

Write modular, multi-use code

- Automatic, robust, scalable
- Most importantly: interoperable and gracefully failing
- If you do this, we can make cool things like the LONI pipeline



Future Challanges

Social Challanges

- Earning the trust of the connectomics community
- Good start (many field leaders contributing)

Technical Challenges

- Only a handful of researchers using data to fullest extent
- Many people just viewing a few images (no stress test)
- Data analysis is currently primative

You are the future!

Discussion

Pros

- Great argument for Open Science
- Historical perspective
- Why connectomics is possible, but why it is still difficult

Cons

- Few technical details
- Early in OCP life