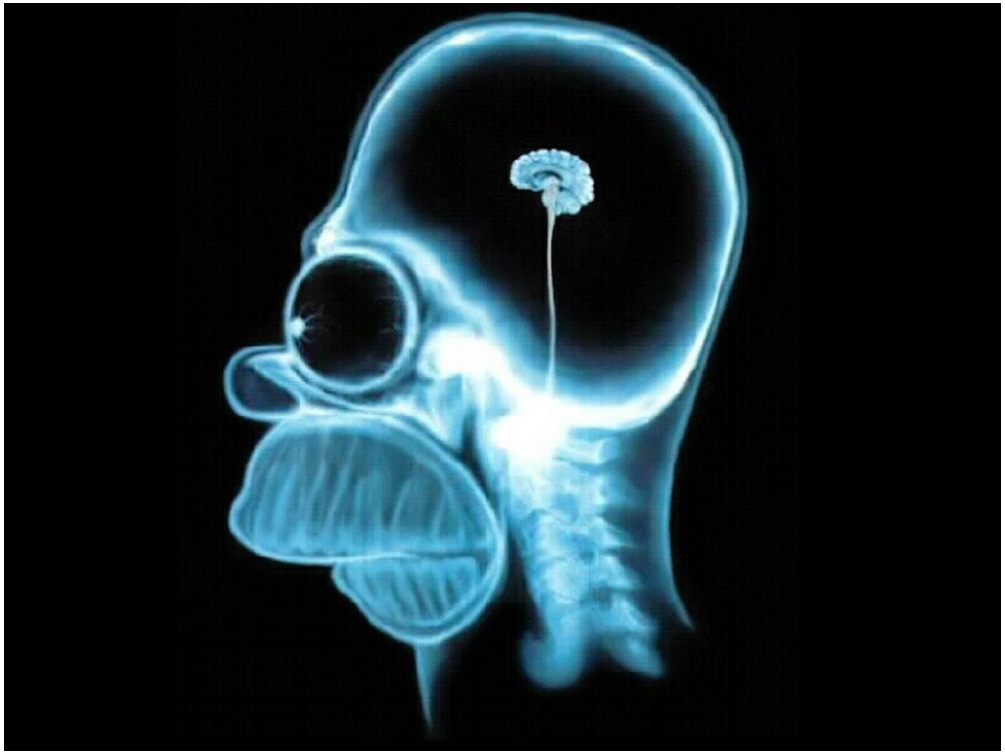
Leveraging High Performance Computing (HPC) for Preservation and Curation

David Minor
UC San Diego Library
(Formerly of San Diego Supercomputer Center)

Why us?

General statements



Traditional view of HPC

HPC centers

Tend to be:

- Focused on research outcomes
- Less focused on production services
- Driven by short/soft money
- Agile
- Have a range of diverse expertise and services

Curation and Preservation

Preservation and curation processes

Tend to be:

- Focused on production
- Driven by long-term institutional commitments
- Not agile
- Range of services and staff but a shared common customer focus

Connection points

- Behind the scenes:
  - Infrastructure
  - Common functions

- Outward-facing
  - Requirements for persistence
  - Evolving user expectations

Infrastructure

What was rarified HPC a few years ago is common now

- High speed networking within and among data centers
- Interfaces that can be grouped for performance
- Storage that's "smarter" and more "self-healing"
- And yeah lots of capacity

These map well to preservation and curation needs

Preservation / curation functions

Common functions

- Fixity and checksums
- Data replication
- Data tracking / registration
- The importance of data generally

These map well to HPC

Requirements for persistence

- Mandated by funders
- Expected by users
- Demanded by data center policies

- Still a challenge in HPC
  - Scale
  - Short-term funding model an issue
  - Concept of scratch as core function
  - Mindset of “we can just re-run the code”

Some examples

- Evolved out of SDSC
  - SRB/iRODS
  - NSF-based partnerships with NCAR and Maryland
  - Was designed for scale (300TBs in 2007)
  - Was an experiment!

- Now “owned” by UCSD Library
  - More closely matches goals and strategies
  - Better aligns with production mindset
  - IT BETTER FREAKING WORK

… but

- Infrastructure is a mix
  - Library-owned Islilon system
  - Staffing is transitioning to Library
  - All equipment sits in SDSC data center
  - Utilizes high-speed networks
  - Better connections to internal users being built

- Includes two HPC infrastructures in the mix
  - Chronopolis and Texas Advanced Computing Center
- Planning for petabyte-scale distributed storage
- We will have to tackle moving massive amounts of data thousands of miles

From the other direction

Large-scale, high-performance drivers for preservation and curation:

- EarthCube
- Research Data Alliance

- Many local campus efforts underway….

General conclusions

- Move toward ecosystem approaches
- Tying services and infrastructure from the start
- Our user bases are growing together