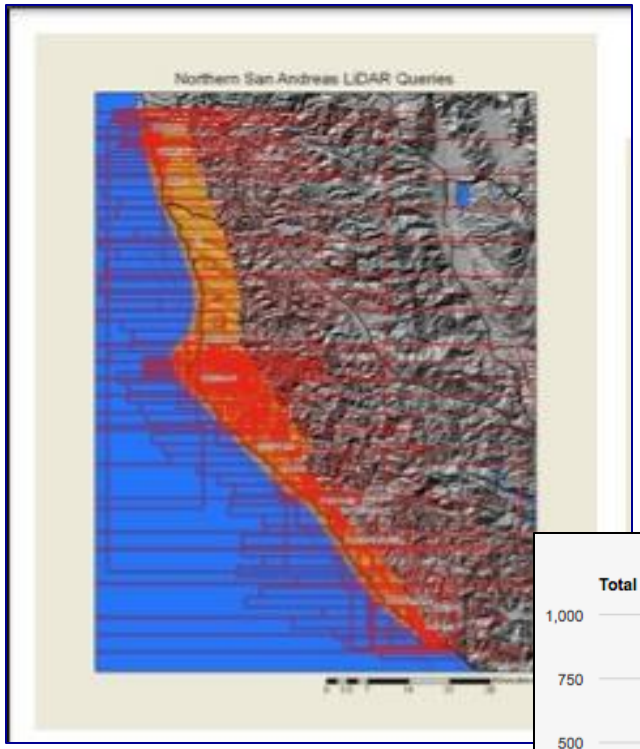
Data Hot Spots

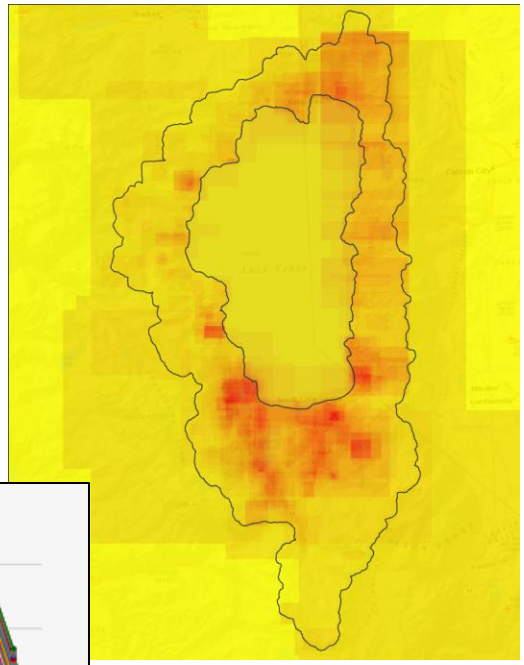
Chaitan Baru
San Diego Supercomputer Center
UC San Diego

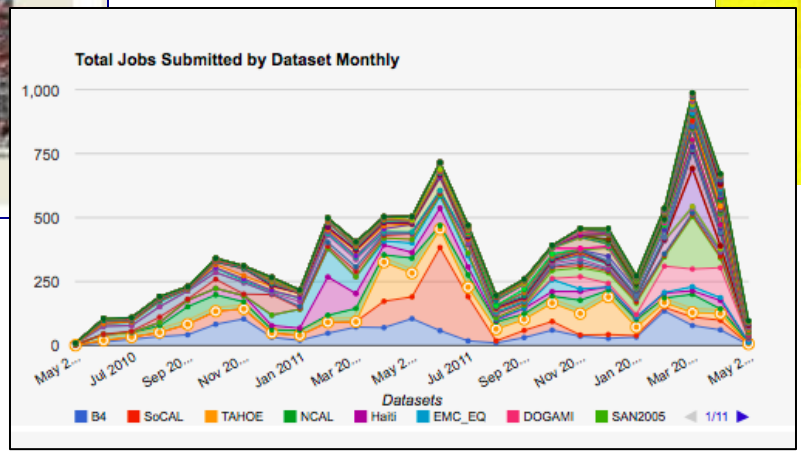
Data Hot Spots: Chaitan Baru, SDSC

- Our example is based on remote-sensing, geospatial data, airborne LiDAR
  - Multi-terabyte datasets
  - Collections with 10's to 100's of such data
  - And only increasing (repeat scans)

Hot regions and hot time periods







Benchmarking-related issues

- How does the hot spot phenomenon vary across genres?
- Big data systems may benefit with a good storage class hierarchy (disk to SSD to RAM)
  - Implicit but also explicit directives?
- Use different storage/processing approaches for a dataset for hot vs non-hot
  - E.g. Hadoop vs dbms, or other?
- May also be related to the “kinetic data” issue