Matthew J. Graham (Caltech)

Data

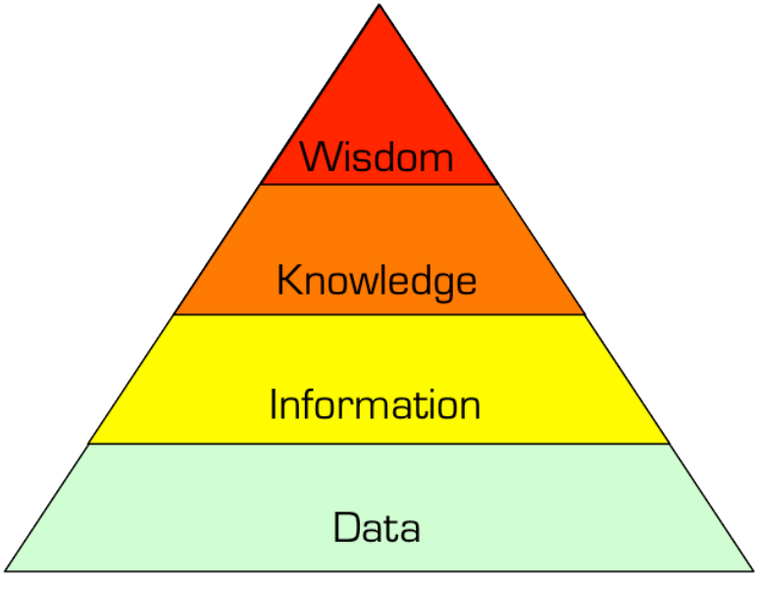
Jet Propulsion Laboratory
Keck Institute for Space Studies
Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology
Caltech

overview

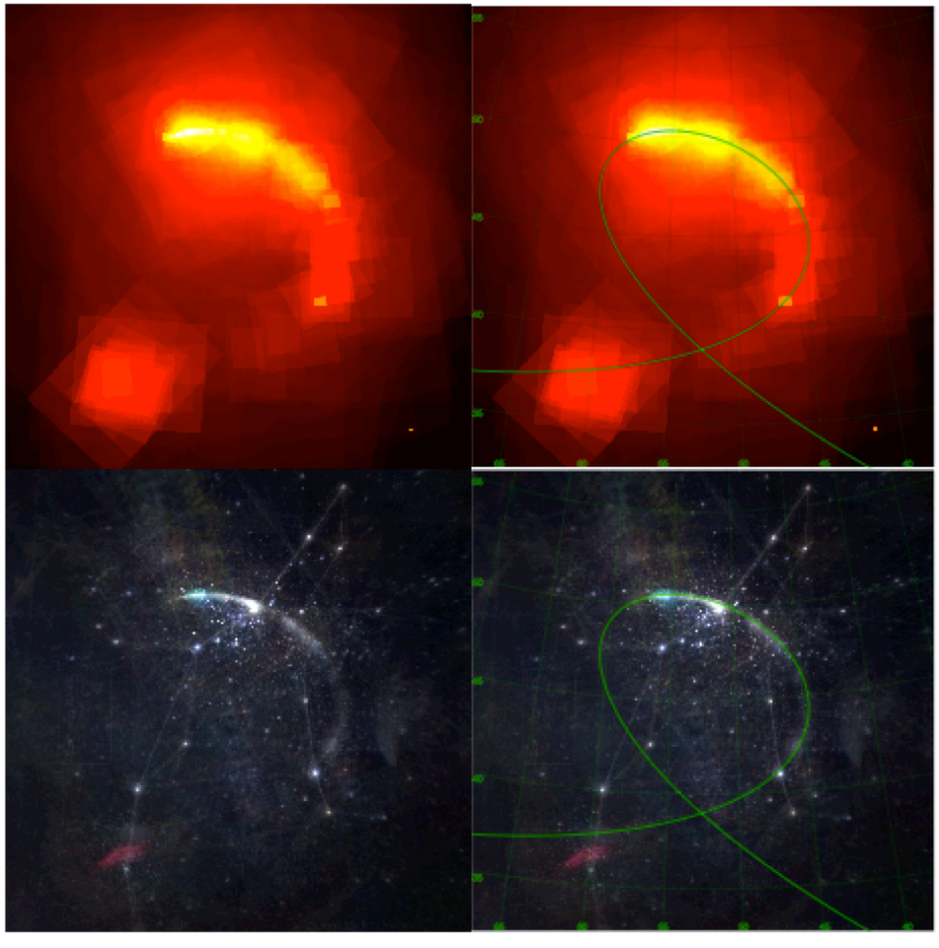
- data
- data models
- relational databases
- sql - the basics
- advanced sql
- alternative databases

what is data?

- Values of qualitative or quantitative variables belonging to a set of items
- Metadata is also data (and arguably more important)



the power of metadata



scale of data

1 Petabyte of data is …

- enough to store the DNA of the entire US population - and then clone them over twice
- over 2000 years of continuous MP3 playback
- 13.3 years of HD video
- one Tweet per person on the planet per day for 7 weeks

but

- 60 nights of LSST data
- 90s of SKA data

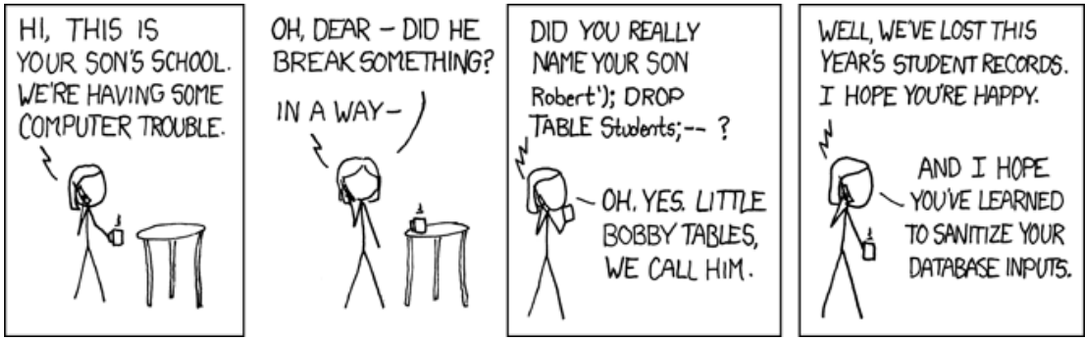
commonplace in the next decade

structured data

- Programmatically we deal with arrays, maps, lists, sets, queues, trees, graphs, …
- No unique solution to working with data - depends on the problem being addressed (and personal preference)
- Data can be regarded as structured or connected

what is a database?

- A structured collection of data residing on a computer system that can be easily accessed, managed and updated
- Data is organised according to a database model
- A Database Management System (DBMS) is a software package designed to store and manage databases



why use a dbms?

- data independence
- efficient and concurrent access
- data integrity, security and safety
- uniform data administration
- reduced application development time
- data analysis tools

scale of databases

- "DBs own the sweet spot of 1GB to 100TB" (Gray & Hey, 2006)
  - SQLite
  - MySQL, PostgreSQL
  - SQLServer, Oracle
  - *Hive/HadoopDB, SciDB, Redis, MonetDB, NuoDB