SAND2008-2621P

An Ever-Growing Avalanche

Large Data and Visualization

April 25, 2008

Andy Wilson
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Working with Large Data

- There is no One True Schema. Adapt or die.
- Self-describing formats are good!
  - XML, SQL, OO databases
- Loading all the data is hopeless.
- Try to load the *structure* of the data.
  - Maximum expressiveness, minimum size
  - For documents: terms, entities and concepts
  - For transaction, travel and communication records: people, places, organizations and relationships
- Smart loading still requires human guidance.

Guidelines for Mastering Scale

- A large graph is just a blob. Sense-making requires abstractions.
  - Hierarchies, clusters, patterns, communities, centrality
  - Outliers are still important!
  - Allow drill-down to original data

- Analysis tools should help you ask questions.
  - Reference librarian instead of card catalog
  - Prefer searching to browsing

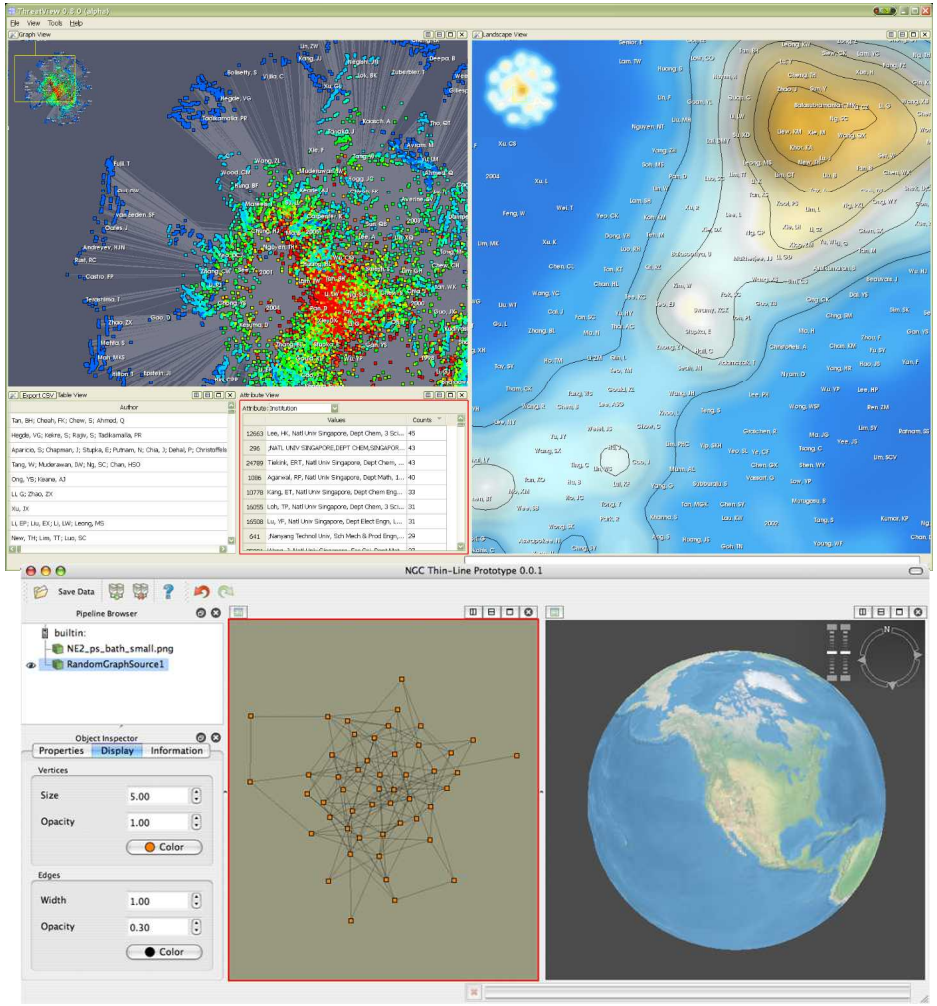
Map of Science

K. Boyack and W. B. Paley

Sandia National Laboratories

Insight Through Correlation

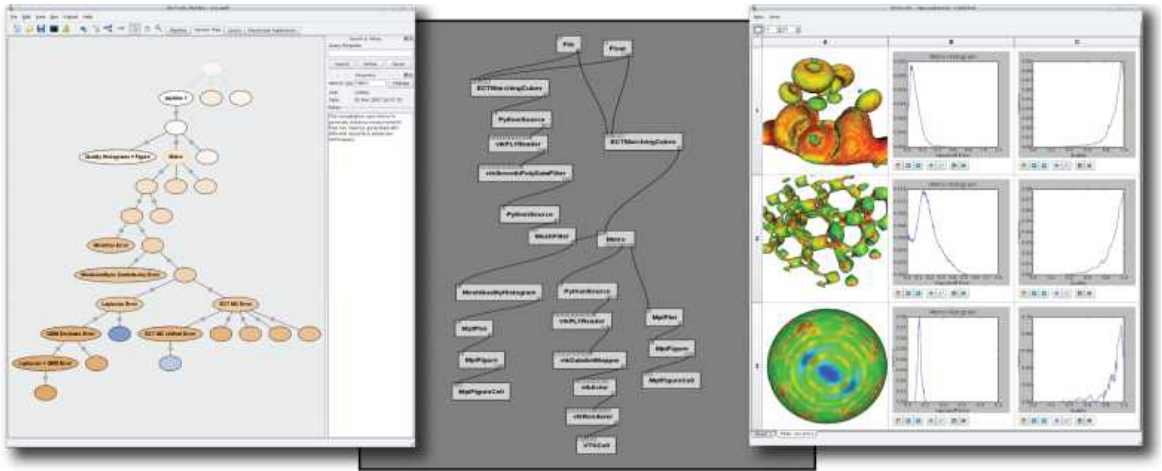
- Different views show different properties
- “Aha” moments happen when everything falls into place
  - This is easier when you can see all the pieces of the puzzle
  - Overlay, side-by-side, small multiples
  - It also helps to see the interactions between views
  - Patterns in one = patterns in another?



Sandia National Laboratories

Sandboxes

- A sandbox is an environment for “what if” and for constructing arguments.
  - Quick access to exploratory tools
  - “How are these actors connected?”
- The system’s job is to track workflow and allow for annotations and assumptions.
  - Backtrack and replay
  - Retract or change assumptions
  - Embody a complete chain of reasoning



VisTrails
www.vistrails.org

Sandia National Laboratories