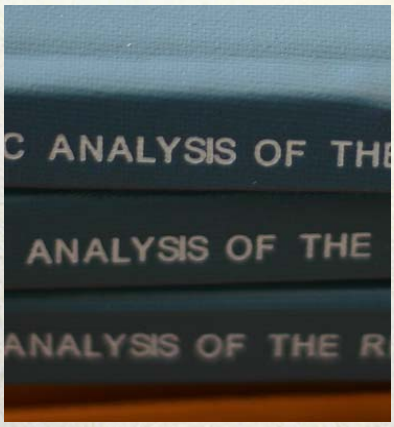
Data Analysis

Andrea Wiggins

IST 400/600

April 14, 2008

Data Analysis



- Data are collected, created, and kept for the purpose of analysis
- Without analysis, it's just a bunch of bits
- Data managers need familiarity with analysis practices

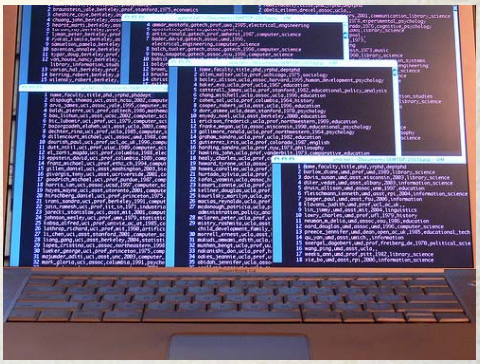
http://flickr.com/photos/techne/100055322/

Overview

- Types of data analysis
- Requirements for analysis
- Basic steps in data analysis
- Types of tools
- Scientific analysis workflows
- Types of analysis output

Requirements for Analysis

- Data
- Analysis design
- Analysis tools
- Computing resources to run the analysis
- Human expertise



http://www.flickr.com/photos/anikarenina/369089979/

Computational Resources

- The computing resources required for analysis will depend on scale and complexity
- Scale refers to the the data
- Complexity refers to the analysis

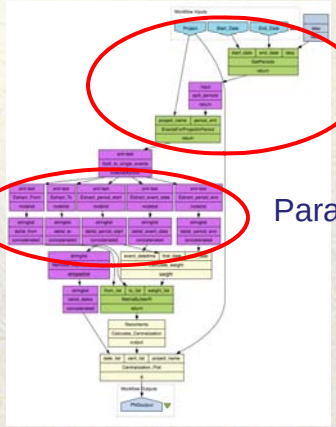
Scale of Analysis

- Physical locations of the data and the analysis machines
- Communication networks
- Volume of data, number of data streams
- Time to complete reflects size of data and complexity of analysis
  - Hours or days may be required

Complexity

- Both analytic and computational complexity are relevant
  - Some operations are “cheap” and others are “expensive”
  - Number of calculations required - every function is made of other functions
- Execution in serial versus parallel processing: how many tasks at once?

Serial & Parallel Processes



Small Scale Computing

- Regular microcomputers like your laptop
- Ordinary consumer PCs are able to do some significant computational work

http://www.flickr.com/photos/cayusa/431036565/

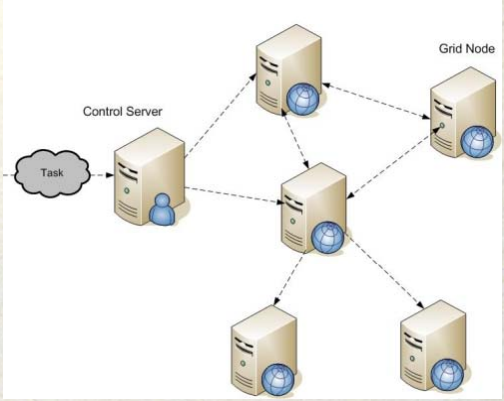
Moderate Scale Computing

- Relatively small, locally-managed clusters
  - Google's smallest cluster: 13 servers
  - Reservoir Simulation Joint Industry Project's cluster ->

http://www.cpge.utexas.edu/rsjip/

Macro Scale Computing

- High performance and grid computing
  - NYSGrid
  - TeraGrid
  - NEESGrid
  - Etc.



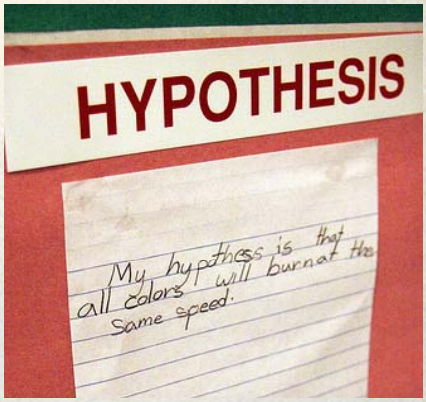
http://ajlopez.wordpress.com/2007/12/03/grid-computing-programming/

Types of Data Analysis

- Two primary types for quantitative data
  - EDA: exploratory data analysis
  - CDA: confirmatory data analysis
- Third type for non-numerical data
  - QDA: qualitative data analysis
    - Photographs, words, observations
    - Traditionally found in social sciences

Confirmatory Data Analysis

- Uses statistical tests to confirm or falsify hypotheses
- You know what you're looking for
- Analysis is usually carefully planned in advance



http://flickr.com/photos/activitystory/105110622/

Exploratory Data Analysis

- Methods used for data mining
  - Nontrivial knowledge discovery from data
- Looking at data to form hypotheses for CDA testing (on a different data set)
- Don't always know what you're looking for, analysis evolves over time
- Caution: sometimes you find what you're looking for, even if it isn't there!

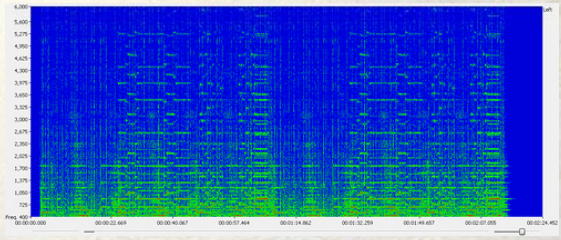
Qualitative Data Analysis

http://flickr.com/photos/valix/939388335/

- Most common in social sciences, where data sets are usually smaller
- Uses a variety of methods to analyze non-numerical data
- Many qualitative analysis methods are difficult or impossible to automate

Context of Analysis

- Scientific inquiry
- Business intelligence
- Monitoring
  - Carefully planned regular reporting
  - As-needed ad hoc analysis



http://flickr.com/photos/makou0629/1145908929/

Data Analysis: Design

- Examine (some of) the raw data
  - Especially important with data meshing, when multiple different data sources are used together
- Design the analysis & instantiate it
  - Test existing hypotheses
  - Explore data to form hypotheses
  - Use databases & analysis tools

Data Analysis: Prepare Data

- Clean the data - preprocessing
  - Remove “noise”
- Sample the data
  - Select the portion to use for analysis
- Validate the analysis
  - Use a subsample to check your analysis
  - Do the results make sense?
  - Can you check intermediate values?

Data Analysis: Revise & Run

- Revise the analysis & test again
  - Also known as debugging
  - Good idea to compare manually and automatically computed results when possible to verify that everything works
  - Repeat as needed
- Run the full analysis when ready

Data Analysis: Save Output

- Save/export the analysis results, artifacts, and appropriate metadata
  - Data selection criteria, sample, analysis design version
  - When analysis was run, by whom
  - System details
    - Time to run, exceptions
  - Other relevant details dictated by your context of inquiry

Data Analysis: Use

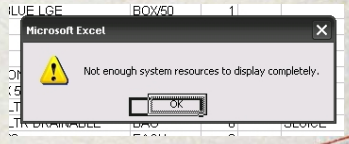
- Write up the results
  - Often requires returning to the raw data, analyzed data, and other information about the analysis
- Questions always arise…
  - Something looks out of place, doesn't make sense, can't possibly be true
  - Double-check everything: results, analysis records, analysis metadata

Very Important Details

- Data formats
  - Format/s of raw data in source/s
  - Format/s required for analysis
  - Format/s of outputs: image, csv, statistics, descriptive text, etc.
- Data manipulation
  - Moving from source to analysis to usable results, without losing/abusing anything

Data Analysis Tools

- Vary by preferences, skills, demands of the data
- Custom solutions
  - Collections of small modular scripts
  - Customized vendor software installations
- Stats packages
  - SPSS, SAS, R
- eScience tools



http://flickr.com/photos/alistairmcmillan/2102898220/

Open Science Movement

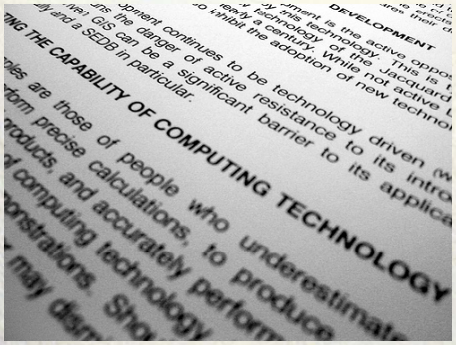
- Not just open data, also open analysis: Open Notebook Science

From Jean-Claude Bradley in Nature Proceedings : doi:10.1038/npre.2007.39.1 : Posted 11 Jun 2007

Analysis Workflows

- Scientists "need access to tools and services that help ensure that metadata are automatically captured or created in real-time" - *Cyberinfrastructure Vision for 21st Century Discovery*
- Taverna Workbench demo video
  - Example of a scientific workflow analysis tool, used for genetics - and social science!
  - http://floss.syr.edu/Presentations/TavernaDemoRedux.m4v

Analysis Outputs

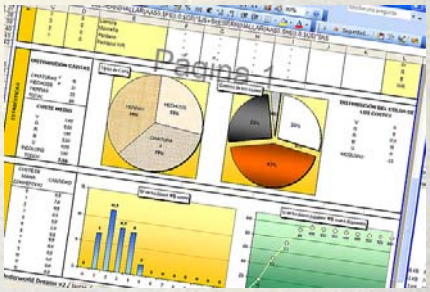


http://flickr.com/photos/thiru/278930492/

- Most analysis starts as numbers and ends up as words
  - Scholarly articles
  - White papers
  - Technical reports
- Visualizations
  - More on Wednesday

Dashboard Reports

- At-a-glance reports for regular, ongoing monitoring
- Uses many visualizations
- Usually intended for managers & executives



http://flickr.com/photos/jauladeardilla/345883088/

Concluding Thoughts

- Understanding how data is used will help you manage it better
- Planning ahead makes data analysis go more smoothly
- Data analysis almost never goes perfectly
- Analysis is the fun part of research, when discoveries are made

Questions for Discussion

- What can data managers contribute to data analysis?
- What are some of the factors that are relevant to designing data analysis?
- How is metadata relevant to designing data analysis?
- How is metadata relevant to reporting data analysis results?