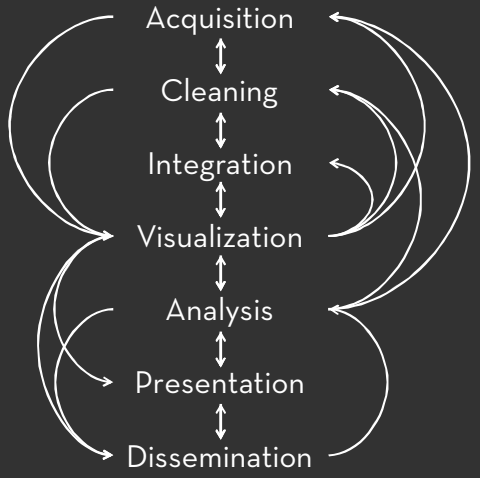
CS448G :: 20 Apr 2011

Data Integration

Jeffrey Heer Stanford University



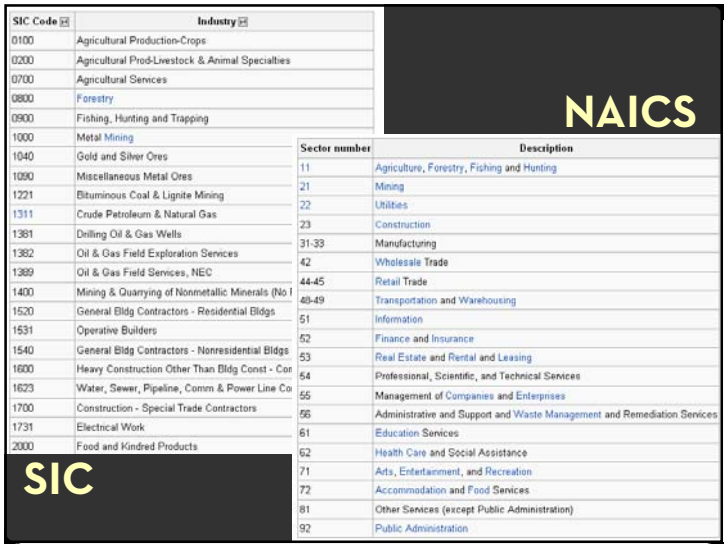
What is Data Integration?
Why is it important?

Data Integration

The problem of combining data **residing in different sources** and providing users with a **unified view** of these data.

Data integration is occurring with **increasing frequency** as the **volume** and **the need to share** existing data continues to grow.

Or, for our purposes: **how can analysts effectively leverage multiple data sources?**



SIC Code	Industry
0100	Agricultural Production-Crops
0200	Agricultural Prod-Livestock & Animal Specialties
0700	Agricultural Services
0800	Forestry
0900	Fishing, Hunting and Trapping
1000	Metal Mining
1040	Gold and Silver Ores
1090	Miscellaneous Metal Ores
1221	Bituminous Coal & Lignite Mining
1311	Crude Petroleum & Natural Gas
1381	Drilling Oil & Gas Wells
1382	Oil & Gas Field Exploration Services
1389	Oil & Gas Field Services, NEC
1400	Mining & Quarrying of Nonmetallic Minerals (No I
1520	General Bldg Contractors - Residential Bldgs
1531	Operative Builders
1540	General Bldg Contractors - Nonresidential Bldgs
1600	Heavy Construction Other Than Bldg Const - Cor
1623	Water, Sewer, Pipeline, Comm & Power Line Co
1700	Construction - Special Trade Contractors
1731	Electrical Work
2000	Food and Kindred Products

SIC

Sector number	Description
11	Agriculture, Forestry, Fishing and Hunting
21	Mining
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Education Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration

NAICS

How to treat this data?

COUNTRY	YEAR	GDP (in USD)
Argentina	2005	$ 183,193,408,941
Argentina	2007	$ 260,789,095,459
Argentina	2009	$ 307,155,148,184
Brazil	2004	$ 663,760,000,000
Brazil	2006	$ 1,088,917,279,412
Brazil	2008	$ 1,652,632,229,228

Levels of Integration

Lightweight

No direct table "join" — plot data in the same space (map, time series) or side-by-side.

On-Demand

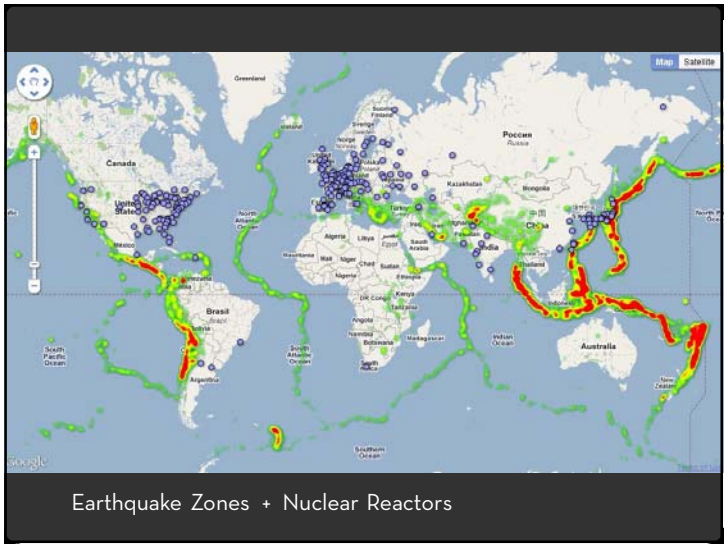
Ad-hoc data integration to enable analysis.

Premeditated

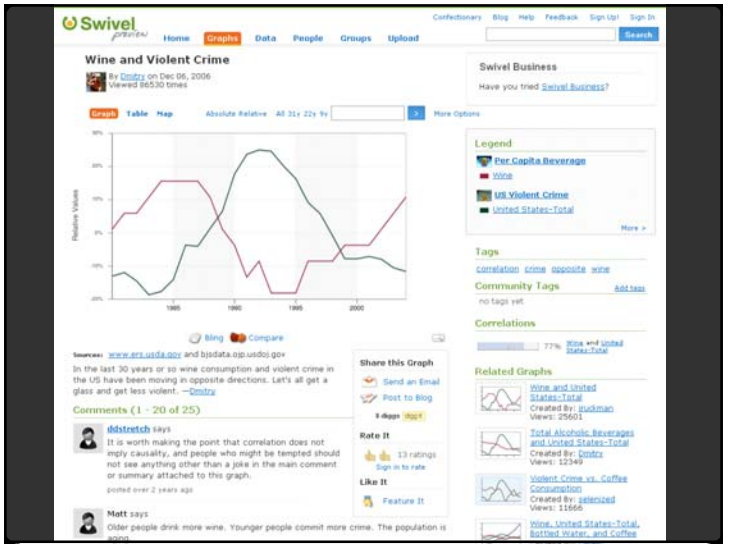
Planned integration to create data infrastructure — mediated schema or ETL (extract-transform-load)

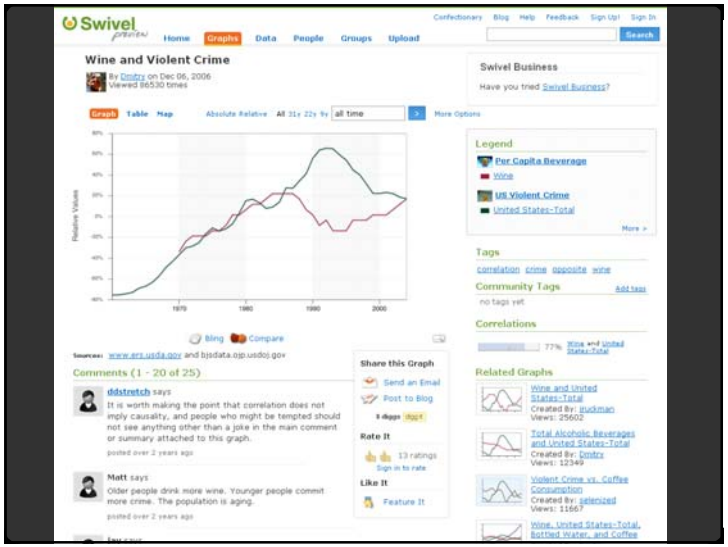
Note: these are my own, idiosyncratic category labels...

Lightweight Integration



Earthquake Zones + Nuclear Reactors





Lightweight Integration

Most data mash-ups: maps, time series

Or, simply plot data side-by-side

Relatively cheap & easy – offsets the actual "integration" work to perception / cognition

Challenges:

Find the right data, format (or geocode) it, query it, and then properly visualize it.

Integration “On Demand”

Web Integration (Dontcheva et al 2007)

(a) Create extraction rules for web content.

(b) Use extracted content from one site to query “extractable” pages from another site.

Vispedia (Chan et al 2008)

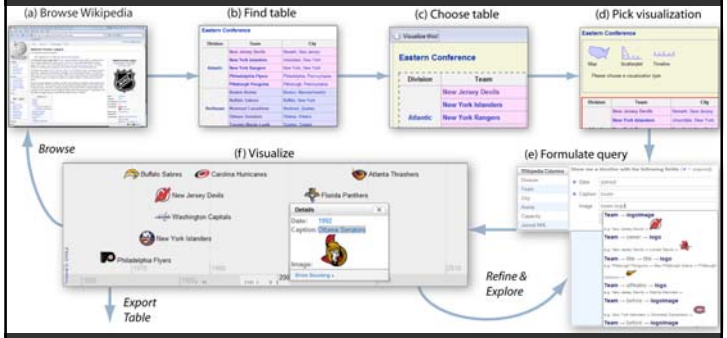


Tableau Data Blending

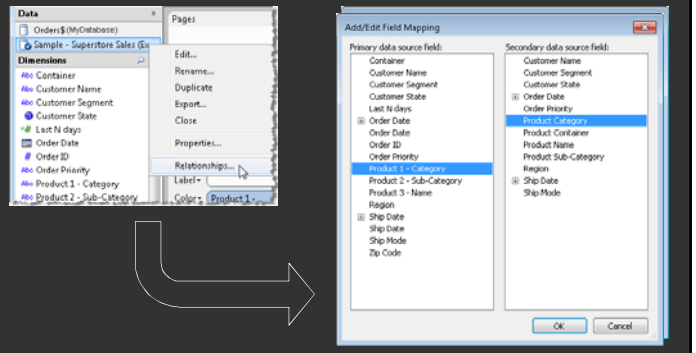
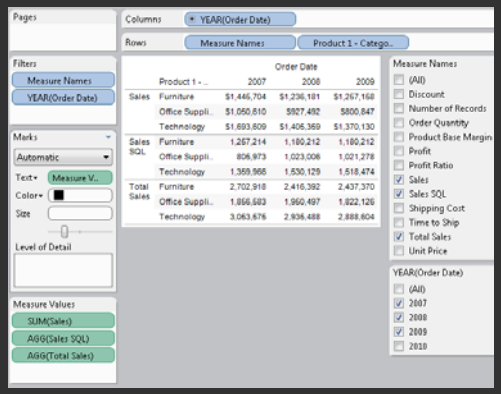


Tableau Data Blending



Premeditated Integration

Premeditated Integration

This is the "classical" DB approach

GOAL: comprehensive, 100% correct, robust data

Approaches:

Mediated Schema – query multiple data sources in a logically unified form

Extract-Transform-Load – take the contents of one database, transform them appropriately, and load them into a unified target database.

Challenges

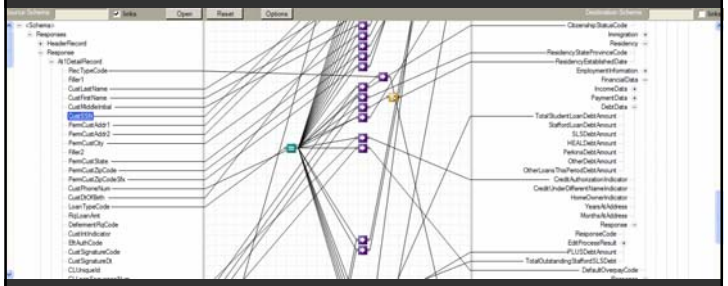
Schema Matching

How do the variables in one data source relate to the variables in the other? May require transformations of attributes.

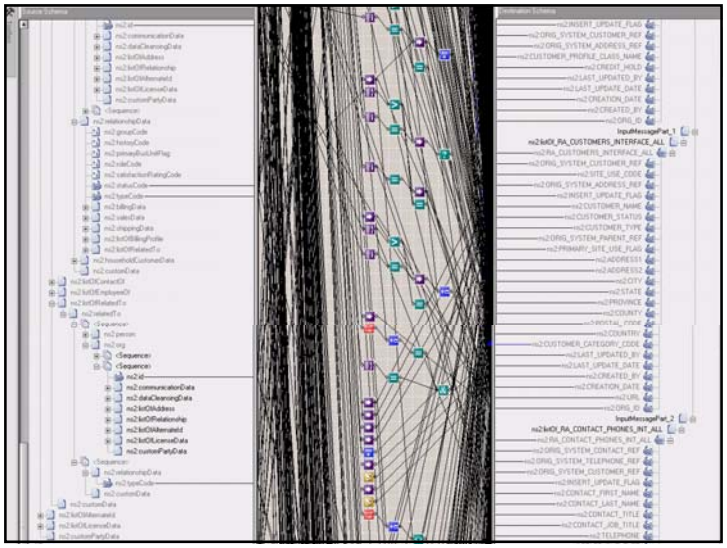
Entity Resolution

How do I know when values in one column reference the same entities as another? How do I convert values appropriately?

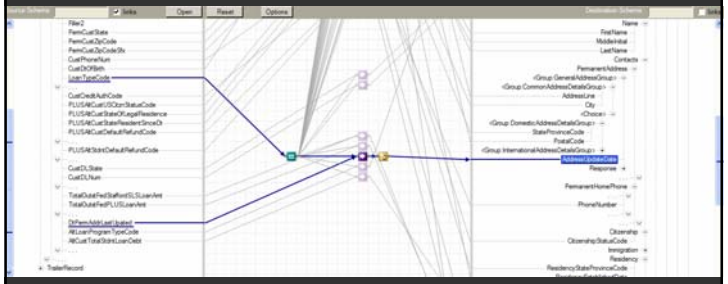
Visual Schema Mapping (Robertson et al)



XML schemas mapped by a data flow graph.
Nodes represent transformation operators.

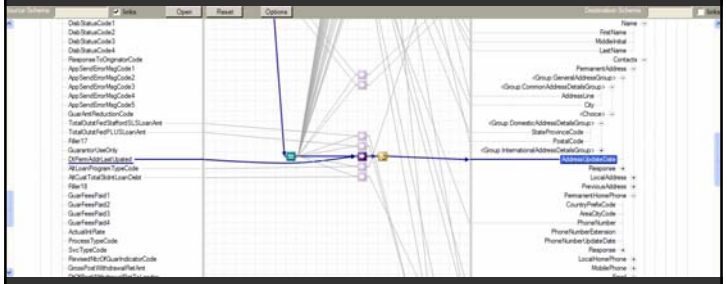


Visual Schema Mapping



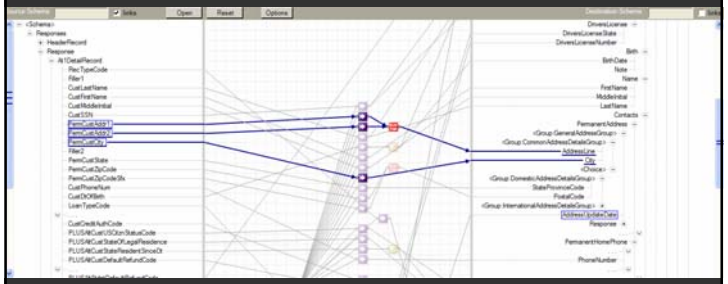
Coalesced trees for Focus + Context

Visual Schema Mapping

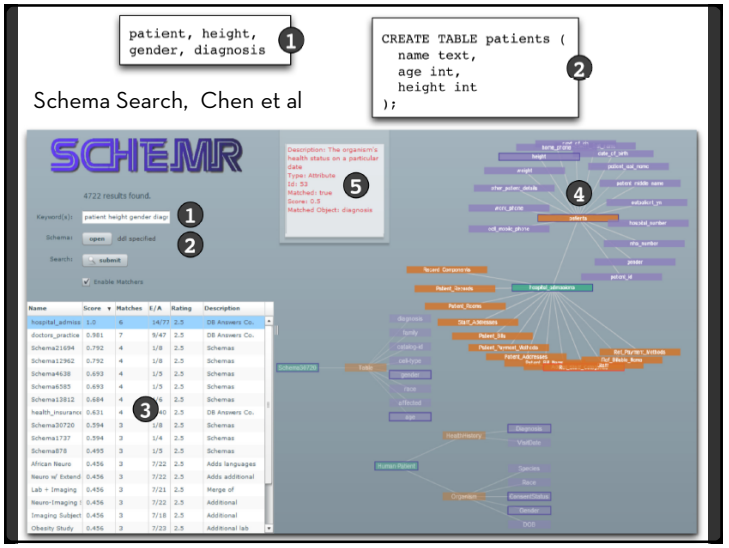


Auto-scrolling; Bent edges to avoid crossings

Visual Schema Mapping



Multiple selections; Search w/ auto-complete



Other Ideas

Pay-As-You-Go Integration (Jeffrey, Franklin, Halevy)

Data integration is costly and requires human judgment. We'd like to focus efforts to have the biggest impact (e.g. amortize costs).

IDEA: Use "on-demand" integration to evolve a data store towards (or beyond?) the service level of a "premeditated" scheme.

Other Ideas

Crowd-Sourced Databases (Marcus et al, CIDR 2011)

Data integration is "AI-Complete." Can we have humans perform these tasks instead?

IDEA: Incorporate *crowd operators* directly into a structured query language. Can express filter and join criteria as crowdsourced tasks (e.g., MTurk HITs). For example: "Is there a flower in this image?"

Develop *crowd-aware* query optimization techniques.

Conclusions

Data integration is a hard (*AI-complete*) problem.

Involving *humans in-the-loop* may provide substantial improvement to analysis. How might interactive tools *reduce the cost*?

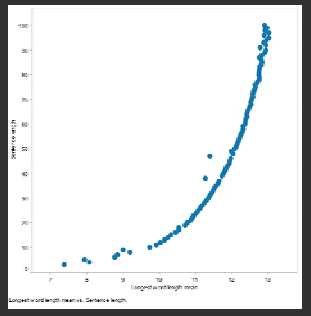
On-demand solutions that *"satisifice"* may be good enough to enable analysis. Can they also amortize over time in a *social process*?

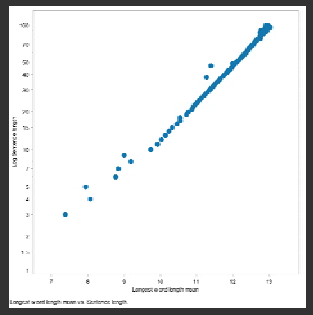
Assignment 2 Highlights*

*** Not all assignments have been graded yet. Don't be devastated if you don't see yours here!**

Sentence Length & Word Complexity On Wikipedia

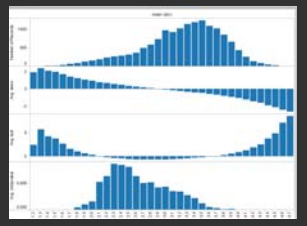
Philip Guo



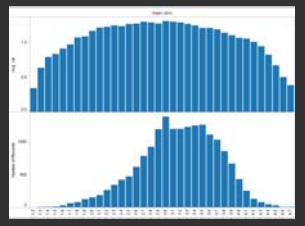


Characterizing Netflix Movie Rating Distributions

Andrea Zvinakis



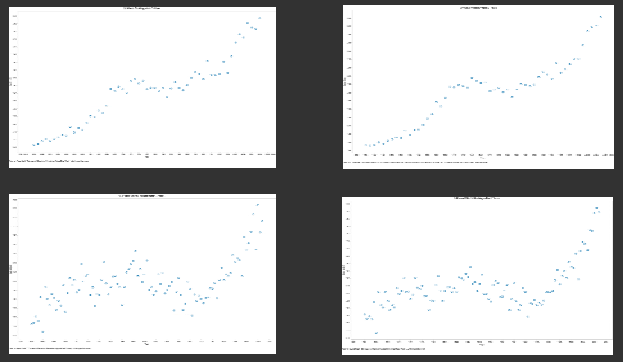
Average rating compared with number of movies, average skew, average kurtosis, and average chi-squared p-value, from top to bottom. This dataset is filtered for movies with chi-squared p-values of <0.05.



Average rating compared with average variance (above) and number of movies (below)

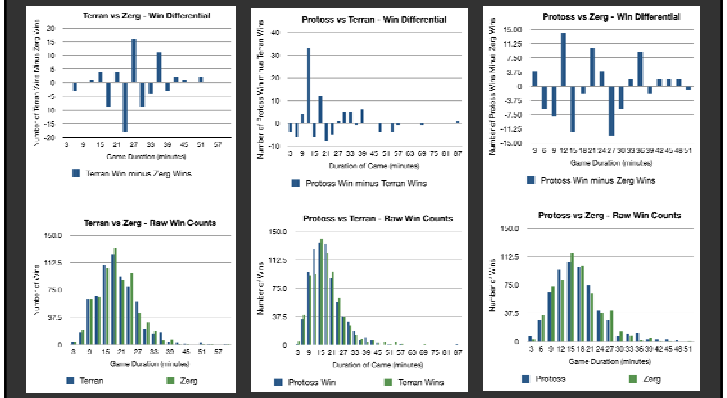
Trending Topic Rates Over Time

Sean Kandel



SC2 Replay Analysis

Salman Ahmad



Next Week

Two Guest Speakers

MON: Jock Mackinlay, Tableau Software

WED: Jeff Hammerbacher, Cloudera

Would you like to join either for lunch? Let us know and we will accommodate who we can.

Reading: Managing Big Data

*Reading responses **due on Wed** morning*

Final Project

Final Project

Initiate an interactive analysis research project.

- **4/25** Draft Abstract – 1-paragraph description
- **5/2** Proposal – abstract, stakeholders & related work
- **5/4** Initial Presentation – sketch, diagram & feedback
- **5/18** Initial Prototype – skeletal but runnable & testable
- **6/1** Advanced Prototype – (nearly) complete
- **6/6** Final Presentation & 4-Page Paper

You may work in teams of **1-3 people**.
Full deliverable details available online.

Discussants

Salman Ahmad
Andrew Pariser
Jonathan Nation