Interactive Tools for Data Transformation & Visualization

Jeffrey Heer  Stanford University

How much data (bytes) did we produce in 2010?

2010: 1,200 exabytes
10x increase over 5 years

Gantz et al, 2008, 2010
cabspotting.org
The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, ... because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google’s Chief Economist
The McKinsey Quarterly, Jan 2009
Data Wrangling (n):

A process of iterative data exploration and transformation that enables analysis.

The goal of wrangling is to make data useful:
- Map data to a form readable by downstream tools (database, stats, visualization, ...)
- Identify, document, and (where possible) address data quality issues.

---

DataWrangler

Transform History

Data Quality Meter

Suggested Transforms

Interactive Data Table

Declarative data transformation language
- **Tuple mapping** - split, merge, extract, delete
- **Lookups and joins** - e.g., FIPS code to US state
- **Reshaping** - e.g., cross-tabulation
- **Sorting, aggregation, etc.**

Informed by prior work in databases, namely Potter’s Wheel & SchemaSQL
Data Wrangler

Declarative data transformation language

+ Mixed-initiative interface for data transforms
  · **Select** data elements of interest
  · **Suggest** applicable transforms
  · Enable rapid **preview and refinement**

Comparative Evaluation

Compared Wrangler performance to Excel with 3 data cleaning tasks on small data sets.

Median completion time for Wrangler at least twice as fast in all tasks.

Skilled Excel users benefit proportionately!
How do people create visualizations?

**Chart Typology**
- Pick from a stock of templates
- Easy-to-use but limited expressiveness
- Prohibits novel designs, new data types

**Component Architecture**
- Permits more combinatorial possibilities
- Novel views require new operators, which requires software engineering.

**Efficiency**

**Expressiveness**

**Chart Typologies**
- Excel, Many Eyes, Google Charts

**Visual Analysis Languages**
- Tableau VizQL, ggplot2, HiVE

**Component Model Architectures**
- Improvise, Prefuse, Flare

**Graphics APIs**
- OpenGL, Java2D, GDI+, Processing

Today’s first task is not to invent wholly new [graphical] techniques, though these are needed. Rather we need most vitally to recognize and reorganize the essential of old techniques, to make easy their assembly in new ways, and to modify their external appearances to fit the new opportunities.


**Protovis**: A Declarative Language for Visualization

A graphic is a composition of data-representative marks.

with Mike Bostock & Vadim Ogievetsky
Protovis
Create customized visualizations using a declarative specification language.

var vis = new pv.Panel();
vis.add(pv.Bar)
data([1, 1.2, 1.7, 1.5, .7])
.bottom(10)
.width(20)
.height(function(d) d * 70)
.left(function() this.index * 25 + 20);
vis.render();

Protovis (http://protovis.org) – Declarative Visualization Specification
Exploiting Declarative Specification

Protovis has led to faster designs, less code
Job Voyager: 5x less code, 10x less dev time
Over 40,000 downloads and widely in use
Multiple implementations: JavaScript & Java
Behind-the-scenes optimization & parallelization
20x scalability over prior systems (in Java)

d3.js Data-Driven Documents
by Mike Bostock
sense.us

A Web Application for Collaborative Visualization of Demographic Data

with Fernanda Viégas and Martin Wattenberg
The great postmaster scourge of 1910?
Or just a bug in the data?

Voyagers and Voyeurs

Complementary faces of analysis

**Voyager** – focus on visualized data
Active engagement with the data
Serendipitous comment discovery

**Voyeur** – focus on comment listings
Investigate others’ explorations
Find people and topics of interest
Catalyze new explorations
Content Analysis of Comments

Feature prevalence from content analysis ($\min$ Cohen’s $\kappa = .74$)

High co-occurrence of Observation, Question, and Hypothesis

16% of sense.us comments and 10% of Many-Eyes comments reference data integrity issues.

Acquisition → Cleaning → Integration → Visualization → Analysis → Presentation → Dissemination

Acquisition → Cleaning → Integration → Visualization → Analysis → Presentation → Dissemination
Students & Collaborators
Mike Bostock
Jason Chuang
Sean Kandel
Diana MacLean
Vadim Ogievetsky
Joe Hellerstein, Andreas Paepcke
Fernanda Viégas, Martin Wattenberg

Interactive Tools for Data Transformation & Visualization

Jeffrey Heer  http://vis.stanford.edu