PROCESS

Miriah Meyer
University of Utah
ADMINISTRIVIA
LAST TIME
Tufte’s integrity principles

Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity.

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented. *aka The Lie Factor*

Show data variation, not design variation.
Tufte’s design principles

- maximize the data-ink ratio
- avoid chart junk (sometimes)
- use multifunctioning elements
- layer information
- maximize the data density
  - shrink the graphics
  - maximize the amount of data shown (sometimes)
Williams’s design principles

Contrast

Repetition

Alignment

Proximity
critique & redesign exercise ...
Proportion of pupils gaining good GCSEs

blue
Weddington St Nicolas
Whitestone Bulkington
Bedworth North West
Abbey Wrenbrook
Camp Hill Gilley Common
Bede Poplar
Arbury Stockingford

brown
Shipston
Stratford upon Avon
Wellesbourne Kineton
Southam Feddon
Studley Hendley
Alcester Bidford

green
Dunchurch Division
Fosse
Eastlands Hillmorton
Earl Craven
Rugby Town West
Brownsover Benn Newbold

purple
Warwick Rural West
Kenilworth
Warwick
Whitnash
Warwick Rural East
North Leamington
South Leamington

red
North Warwickshire North
Nwarks West
Nwarks South
North Warwickshire East

- visualization design process
- types of research contributions
COMMENTS ON READINGS
visualization design process

types of research contributions
How do you select and foster a good collaboration?
target
translate
translate goals into data analysis tasks
structure and characterize data
create abstraction of problem
What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!

What they hear
blah blah GINGER blah blah blah, blah, blah, GINGER, blah, blah, GINGER, blah, blah, blah, blah, blah...
From patterns of conservation in gene expression profiles we want to visualize the evolutionary mechanisms that influence gene regulation.

What do you want to visualize?

What do you want to visualize?

hmmmm... ahHA!

hmmmm... ahHA!
From patterns of conservation in gene expression profiles we want to visualize the evolutionary mechanisms that influence gene regulation.

So, you want to characterize the differences in, and find similar groups of, time series that correspond to nodes in a binary tree.
data

- tabular
  - ordered
    - categorical
    - ordinal
    - quantitative
  - relational
  - spatial

this Thursday
4  **AN ANALYTIC TASK TAXONOMY**

The ten tasks from the affinity diagramming analysis are:

- Retrieve Value
- Filter
- Compute Derived Value
- Find Extremum
- Sort
- Determine Range
- Characterize Distribution
- Find Anomalies
- Cluster
- Correlate

Each of the tasks is presented in the following sections, along with a *pro forma* abstract [9] and example questions that serve as general models and examples of the tasks. These tasks are not meant to be a normative picture of user analytic activity, but rather to provide a vocabulary for discussion.
design visual encodings and interaction mechanisms to support data and task abstraction

transform data with appropriate computational methods

try many ideas!
Automating the Design of Graphical Presentations of Relational Information
MacKinlay, 1986

Semiology of Graphics
Bertin, 1967
try many ideas

space of possible solutions

consideration space

good solution

best solution

space of possible solutions

consideration space
RAPID PROTOTYPING

Prototyp Level

- Non interactive
- No/few features
- Not Usability tested
- Not stable
- Interactive
- Fully featured
- Usability Tested
- Stable

- Sketches
- LoFi PT
- HiFi PTs
- Beta Version
- Full System
ideation exercise . . .
target

translate

design

implement

bring your design to life

optimize algorithms
in early March

D3: Data-Driven Documents
Michael Bostock, Vadim Ogievetsky, Jeffrey Heer

ABSTRACT
Data-Driven Documents (D3) is a novel representation-transparent approach to visualization for the web. Rather than hide the underlying scenegraph within a toolkit-specific abstraction, D3 enables direct inspection and manipulation of a native representation: the standard document object model (DOM). With D3, designers selectively bind input data to arbitrary document elements, applying dynamic transforms to both generate and modify content. We show how representational transparency improves expressiveness and better integrates with developer tools than prior approaches, while offering comparable notational efficiency and retaining powerful declarative components. Immediate evaluation of operators further simplifies debugging and allows iterative development. Additionally, we demonstrate how D3 transforms naturally enable animation and interaction with dramatic performance improvements over intermediate representations.

MATERIALS AND LINKS
PDF (2.2 MB) | Software | Video | BibTeX Citation

CITATION
D3: Data-Driven Documents
Michael Bostock, Vadim Ogievetsky, Jeffrey Heer
PDF (2.2 MB) | Software | Video

http://vis.stanford.edu/papers/d3
target

translate

design

implement

problem characterization and abstraction 80%

visualization design 20%
ensure visualization system supports target users’ goals
in late March
no amount of brilliant design can overcome designing for the wrong thing
NESTED MODEL

what can go wrong, and how do we validate?

Munzner 2009
**problem**

threat: they don’t do that

validation: *immediate*, observe and interview users *downstream*, notice adoption rates
**abstraction**

threat: you’re showing them the wrong thing
validation: downstream, deploy and observe usage

Munzner 2009
**design**

threat: they way you show it doesn’t work
validation: *immediate*, justification with known principles *downstream*, qualitative or quantitative analysis of results; lab study measuring time and error
algorithm

threat: you’re code is too slow
validation: immediate, complexity analysis
downstream, benchmarks for system time and memory
avoid validation mismatch
- cannot validate encoding with system timings
- cannot validate abstraction with lab studies
target
translate
design
implement
validate

user-centered design
usability engineering
participatory design
design thinking
Pathline
A Tool for Comparative Functional Genomics Data

joint work with:
Bang Wong, Mark Styczynski, Tamara Munzner, Hanspeter Pfister
target
translate
design
implement
validate
functional genomics

how do genes work together to perform different functions in a cell?
functional genomics data

gene expression

molecular pathways
gene expression is ...

... the measured level of how much a gene is on or off
... a single quantitative value

biologists measure it ...

... for many genes
... in many samples (time points, tissue types, species)

visualized with heatmaps

[Wilkinson09] [Saldanha04] [Seo02] [Eisen98] [Gehlenborg10] [Weinstein08]

encode value with color
gene expression is …

... the measured level of how much a gene is on or off
... a single quantitative value

biologists measure it ...

... for many genes
... in many samples (time points, tissue types, species)

visualized with heatmaps

[Wilkinson09] [Saldanha04] [Seo02] [Eisen98]
[Gehlenborg10] [Weinstein08]

encode value with color
augmented with clustering
functional genomics data

gene expression

molecular pathways
the functioning of a cell is controlled by many interrelated chemical reactions performed by genes
glycolysis

TCA cycle

Pathways

www.genome.jp/kegg/
functional genomics

how do genes work together to perform different functions in a cell?

comparative functional genomics

how do the gene interactions vary across different species?
collaborators: Regev Lab at the Broad Institute

biology: metabolism in yeast

data: multiple genes
  multiple time points
  multiple related species
  multiple pathways

problem: existing tools can only look at a subset of this data

comparative functional genomics

how do the gene interactions vary across different species?
**metabolic pathways**

- 10 to 50 pathways of interest
- Inputs/outputs called metabolites
- **Directed graph**

**gene expression**

- 6000 genes and 140 metabolites
- 6 time points
- 14 species of yeast
- **3D table**

**similarity scores**

- Aggregate time series for a gene/metabolite over species
- Similarity of expression across species
- Aggregate: Pearson, Spearman, others

**phylogeny**

- Evolutionary relationship
- **Binary tree**

---

**Shortened form**

```
aggregate similarity scores
• 10 to 50 pathways of interest
• inputs/outputs called metabolites
• directed graph

6000 genes and 140 metabolites
• 6 time points
• 14 species of yeast
• 3D table

similarity scores
• aggregate time series for a gene/metabolite over species
• similarity of expression across species
• aggregate: Pearson, Spearman, others

phylogeny
• evolutionary relationship
• binary tree
```
tasks

- study expression data as a time series
- compare a limited number of time series
- compare similarity scores along a pathway(s)
- comparison of multiple similarity scores
metabolic pathways
similarity scores
similarity scores
similarity scores
gene expression
phylogeny
target
translate
design
implement
validate
encode quantitative values with spatial position
topological layout

linearized pathway

www.win.tue.nl/~mwestenb/genevis/
encode quantitative values with spatial position

curvemap
courtesy of M. Styczynski from JavaTreeview
jtreeview.sourceforge.net/
Pathline
linearized pathway representation
linearized pathway representation

common axes to compare similarity scores

- bars and circles
- visual layers for selective attention
- color-code gene direction
- multiple similarity scores
- multiple pathways
pathway to ordered list of nodes

unroll and cut  reinsert  shared coordinate frame and stylized marks
linearized pathway representation

putting it together . . .

- use spatial position for similarity scores
- topology is secondary
Pathline
curvemap
curvemap

inspired by heatmaps

- base visual unit is a curve
- filled, framed line charts to enhance shape perception
- rows are species
- columns are genes/metabolites
- overlays to enhance trends
curvemap

inspired by heatmaps

- base visual unit is a curve
- filled, framed line charts to enhance shape perception
- rows are species
- columns are genes/metabolites
- overlays to enhance trends
Demo
target
translate
design
implement
validate
case study

- qualitative research method
- in-depth study of individual or group
- real-world setting
- description and interpretation
whole genome duplication

both genes

one gene
www.pathline.org
LESSONS LEARNED

- process supports efficient development
- collaborators’ time commitment is front loaded
- rapid prototyping is essential
LESSONS LEARNED

- process supports efficient development
- collaborators’ time commitment is front loaded
- rapid prototyping is essential
- put off coding as long as possible
contributions

- **Pathline**
  - multiple genes, time points, species, and pathways

- **linearized pathway representation**

- **curvemap**

- **tool deployment**
  - open source
  - used daily by several collaborators
- visualization design process
- types of research contributions
5. PAPER TYPES

A Visweek paper typically falls into one of five categories: technique, system, design study, evaluation, or model. We briefly discuss these categories below. Although your main paper type has to be specified during the paper submission process, papers can include elements of more than one of these categories. Please see “Process and Pitfalls in Writing Information Visualization Research Papers” by Tamara Munzner for more detailed discussion on how to write a successful Visweek paper.

Technique papers introduce novel techniques or algorithms that have not previously appeared in the literature, or that significantly extend known techniques or algorithms, for example by scaling to datasets of much larger size than before or by generalizing a technique to a larger class of uses. The technique or algorithm description provided in the paper should be complete enough that a competent graduate student in visualization could implement the work, and the authors should create a prototype implementation of the methods. Relevant previous work must be referenced, and the advantage of the new methods over it should be clearly demonstrated. There should be a discussion of the tasks and datasets for which this new method is appropriate, and its limitations. Evaluation through informal or formal user studies, or other methods, will often serve to strengthen the paper, but are not mandatory.

System papers present a blend of algorithms, technical requirements, user requirements, and design that solves a major problem. The system that is described is both novel and important, and has been implemented. The rationale for significant design decisions is provided, and the system is compared to documented, best-of-breed systems already in use. The comparison includes specific discussion of how the described system differs from and is, in some significant respects, superior to those systems. For example, the described system may offer substantial advancements in the performance or usability of visualization systems, or novel capabilities. Every effort should be made to eliminate external factors (such as advances in processor performance, memory sizes or operating system features) that would affect this comparison. For further suggestions, please review “How (and How Not) to Write a Good Systems Paper” by Roy Levin and David Redell, and “Empirical Methods in CS and AI” by Toby. 
L4: Data

REQUIRED READING
Data Principles

Many aspects of a visualization design are driven by the kind of data that we have at our disposal: what kind of data do we need to look at? What information can we figure out from the data itself, versus the meanings that we must be told explicitly? What high-level concepts will allow us to split datasets apart into general and useful pieces? What kind of attributes does the data have to begin with, and what kinds of derived data might we compute in order to draw a more effective picture?

This chapter approaches these questions with a taxonomy of visualization data types and semantics that meshes well with the principles in Part II and methods in Part III. Figure 2.1 shows the big picture, which will be expanded on in the rest of the chapter.

The chapter begins with a discussion of dataset types, then makes a distinction between semantics and types, and continues with attribute types and attribute semantics. It returns to datasets with semantics. The chapter then covers derived attributes and spaces. It concludes by relating this taxonomy of data principles to the idea of the data abstraction level of the nested design model.

2.1 Dataset Types
Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases

Chris Stolte, Diane Tang, and Pat Hanrahan

Abstract—In the last several years, large multidimensional databases have become common in a variety of applications such as data warehousing and scientific computing. Analysis and exploration tasks place significant demands on the interfaces to these databases. Because of the size of the data sets, dense graphical representations are more effective for exploration than spreadsheets and charts. Furthermore, because of the exploratory nature of the analysis, it must be possible for the analysts to change visualizations rapidly as they pursue a cycle involving first hypothesis and then experimentation. In this paper, we present Polaris, an interface for exploring large multidimensional databases that extends the well-known Pivot Table interface. The novel features of Polaris include an interface for constructing visual specifications of table-based graphical displays and the ability to generate a precise set of relational queries from the visual specifications. The visual specifications can be rapidly and incrementally developed, giving the analyst visual feedback as they construct complex queries and visualizations.

Index Terms—Database visualization, database analysis, visualization formalism, multidimensional databases.

1 INTRODUCTION

In the last several years, large databases have become common in a variety of applications. Corporations are creating large data warehouses of historical data on key aspects of their operations. International research projects such as the Human Genome Project [20] and Digital Sky Survey [31] are generating massive databases of scientific data.

...