Visualization for Hackers: Why It’s Tricky, and Where to Start

Tamara Munzner
Department of Computer Science
University of British Columbia

Hackers on Planet Earth (HOPE) X
19 July 2014, New York NY

http://www.cs.ubc.ca/~tmm/talks.html#hope14
Outline

• introduction
  – what’s vis anyway?

• LiveRAC
  – server logs: managed web hosting
    (with AT&T)

• Overview
  – text: visual document mining for journalists
    (with Associated Press)

• big picture and wrapup
Defining visualization (vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Why?...
Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

• many analysis problems ill-specified, not clear what questions to ask in advance – don’t need vis when fully automatic solution exists and is trusted

**Anscombe’s Quartet**

<table>
<thead>
<tr>
<th>Identical statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x mean</td>
<td>9</td>
</tr>
<tr>
<td>x variance</td>
<td>10</td>
</tr>
<tr>
<td>y mean</td>
<td>8</td>
</tr>
<tr>
<td>y variance</td>
<td>4</td>
</tr>
<tr>
<td>x/y correlation</td>
<td>1</td>
</tr>
</tbody>
</table>
Why use an external representation?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- external representation: replace cognition with perception
Analysis framework: Four levels, three questions

- **domain** situation
  - who are the target users?

- **abstraction**
  - translate from specifics of domain to vocabulary of visual design
  - **what** is shown? **data abstraction**
  - **why** is the user looking at it? **task abstraction**

- **idiom**
  - **how** is it shown?
    - **visual encoding idiom**: how to draw
    - **interaction idiom**: how to manipulate

- **algorithm**
  - efficient computation

---


Why analyze?

• huge design space
  – visual encoding: combinatorial explosion of choices
  – add interaction: even bigger
  – add data abstraction transformation: truly enormous

• most possibilities ineffective for particular task/data combination
  – implication: avoid random walk, be guided by principles

• analysis framework: scaffold to think systematically about design space
  – ensure that consideration space encompasses full scope of possibilities
  – improve chances that selected solution is good not mediocre
  – today’s focus: abstractions and idioms, what-why-how
Outline

• introduction
  – what’s vis anyway?

• LiveRAC
  – server logs: managed web hosting (with AT&T)

• Overview
  – text: visual document mining for journalists (with Associated Press)

• big picture and wrapup
LiveRAC

*Interactive Visual Exploration of System Management Time-Series Data*

**joint work with:**
Peter McLachlan, Eleftherios Koutsofios, Stephen North.

LiveRAC video

http://youtu.be/ld0c3H0VSkw
What: Data abstraction

• multidimensional table: time series data
  – key attributes
    • time
      – 50,000: 5-minute intervals over 6 months
      – multiscale levels of interest
    • devices
      – 4000
    • parameters
      – 20
      – ex: CPU usage, memory load, network traffic, alarms, ...
  – value attributes
    • parameter value for device at time point
      – quantitative
    • device groups
      – categorical

Tables

Attributes (columns)
Items (rows)
Cell containing value

Multidimensional Table

Attribute Types

Categorical
Quantitative
Ordered
Why: Tasks in domain language

• interpret network environment status
• report generation
• capacity planning
• event investigation/forensics
• coordination
  – between customers, engineering, ops
Why: Task abstraction

• browse and correlate across combinations of parameter, device, time
  – correlate alarm attribute with other parameter attribs
  – find trends across groups of devices
  – summarize over different time intervals
  – identify devices at or beyond parameter thresholds
  – identify critical parameter values
  – compare device behavior at specific event times
How: Facet

• facet: partition data into multiple views
  – juxtapose views side by side
  • same encoding, different data: small multiples
How: Juxtapose

• juxtapose linked views
  – linked highlighting
    • marker line tracks across views

Facet

➔ Juxtapose and Coordinate Views

➔ Share Encoding: Same/Different

➔ Linked Highlighting
How: Navigate

- semantic zooming
  - representation adapts to pixels available for object
    - many: superimposed line charts with full labeling
    - some: iconic line chart (sparkline)
    - few: color-coded box (heatmap)
How: Superimpose

• superimpose layers
  – vs juxtapose side by side
How: Reduce

• reduce data shown with complex combination of filtering and aggregation
  – embed focus+context in single view
  – distort geometry
    • metaphor: stretch and squish navigation
    • shape: rectilinear
    • foci: multiple
    • impact: global
How: Reordering

- change spatial arrangement
  - resort by selected attribute
  - check for correlations between aligned attribute columns
    - ex: high load without high CPU, maybe I/O bound
Importance of arranging space: Underlying definitions

• marks
  – geometric primitives

• channels
  – control appearance of marks
Channels: Expressiveness types and effectiveness rankings

**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**

- Spatial region
- Color hue
- Motion
- Shape

- Spatial position channels best in both cases
  - high accuracy

- more on channel rankings: hour-long talk

*Visualization Principles*

http://www.cs.ubc.ca/~tmm/talks.html#networkbio12
Algorithms

- back end: SWIFT server
- front end: PRISAD rendering
  - separate threads for render vs server update
  - guaranteed visibility of semantically important marks even when squished small
- sublinear rendering: $O(p)$ where $p = \text{pixel count}$
  - scalable for $n$ of millions
  - generic framework
    - time series charts, gene sequences, trees

Outline

• introduction
  – what’s vis anyway?

• LiveRAC
  – server logs: managed web hosting
    (with AT&T)

• Overview
  – text: visual document mining for journalists
    (with Associated Press)

• big picture and wrapup
Overview

The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists

joint work with:
Matt Brehmer, Stephen Ingram, Jonathan Stray

http://www.cs.ubc.ca/labs/imager/tr/2014/Overview/
https://www.overviewproject.org/
Origin story: WikiLeaks meets Glimmer

- WikiLeaks: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  - conjecture that existing label classification falls short of showing all meaningful structure in data
    - friendly action, criminal incident, ...
  - had some NLP, needed better vis tools

- Glimmer: multilevel dimensionality reduction algorithm
  - scalability to 30K documents and terms

What: Data and task abstraction

• derive data to transform text into visualizable dataset
  – from documents to high-dimensional table
    • bag of words model
      – attribute: any word that appears across entire collection
      – document/item: word counts (sparse)
  – from high-dimensional table to low-dimensional table
    • synthesize new dimensions that capture most of high-dim proximity structure
    • find clusters of items in lowD space
      – discover: generate or verify

→ Produce

→ Derive

→ Tables

→ Data and Dataset Types

→ Consume

→ Discover
Dimensionality reduction for document datasets

Task 1

In HD data
Out 2D data

What?
⇒ In High dimensional data
⇒ Out 2D Data

Why?
⇒ Produce
⇒ Derive

Task 2

In 2D data
Out Scatterplot, Clusters & points

What?
⇒ In 2D data
⇒ Out Scatterplot
⇒ Out Clusters & points

Why?
⇒ Discover
⇒ Explore
⇒ Identify

How?
⇒ Encode
⇒ Navigate
⇒ Select

Task 3

In Scatterplot, Clusters & points
Out Labels for clusters

What?
⇒ In Scatterplot
⇒ In Clusters & points
⇒ Out Labels for clusters

Why?
⇒ Produce
⇒ Annotate

• more on DR: hour-long talk *Dimensionality Reduction from Several Angles*
  [http://www.cs.ubc.ca/~tmm/talks.html#linz14](http://www.cs.ubc.ca/~tmm/talks.html#linz14)
Overview video (version 1)

http://www.cs.ubc.ca/labs/imager/tr/2012/modiscotag
What/Why/How interplay

- **why:** understand clusters
- **what:** derive data of full cluster hierarchy
  - explore space of possible clusterings
- **how:** show cluster hierarchy
  - arrange space: node-link
- **how:** support tagging clusters/docs
  - following or cross-cutting hierarchy!
    - simple annotation
    - progress tracking
    - user-defined semantics
How: Idiom design decisions

• **facet: juxtapose linked views**
  – linked color coding
    • cluster hierarchy tree
    • DR scatterplot
    • tags
  – reading text/keywords
    • cluster list
    • doc reader

**Juxtapose and Coordinate Views**

- Share Encoding: Same/Different
- Share Data: All/Subset/None
- Linked Highlighting

**Why?**

**How?**

**What?**

- facet: juxtapose linked views
  - linked color coding
  - cluster hierarchy tree
  - DR scatterplot
  - tags
  - reading text/keywords
    - cluster list
    - doc reader

**Identity Channels: Categorical Attributes**

- Spatial region
- Color hue
- Motion
- Shape
Algorithm

• version 1
  – fast cluster hierarchy construction for sparse data
  – research prototype by PhD student
  – positive initial assessment from AP Caracas bureau chief
    • barrier to adoption: difficult install/load process
Algorithm

• version 1
  – fast cluster hierarchy construction for sparse data
  – research prototype by PhD student
  – positive initial assessment from AP Caracas bureau chief
    • barrier to adoption: difficult install/load process

• version 2
  – web deployment, DocumentCloud integration, usability
    • many months of engineering
      – Knight Foundation funding to the rescue!
    • published story by unaffiliated reporter: police corruption in Tulsa
Algorithm: Spinoff series

• dimensionality reduction for huge text collections
  – great algorithm problem in its own right!
  – QSNE: fast and high-quality DR for millions of documents
    • key feature: handle sparseness appropriately

[Dimensionality Reduction for Documents with Nearest Neighbor Queries. Ingram and Munzner. Neurocomputing (Special Issue on Visual Analytics using Multidimensional Projections), to appear 2014.]

http://www.cs.ubc.ca/labs/imager/tr/2014/QSNE/
Path to adoption

• even more rounds of what/why/how interplay
  – which views needed? what should they show? how should they show it?
  – usability and utility

• version 3
  – published story: VP candidate Ryan asked for federal help even as championed cuts
  – published story: gun control debate

• version 4
  – followup investigation: government corruption in Texas
  – published story: police corruption in New York (*Pulitzer prize finalist!*)
Overview v4 video

- versions 3 and 4
  - no DR scatterplot
  - tree arrangement emphasizing nodes not links
  - combined doc/cluster viewer

http://vimeo.com/71483614
Why: Task abstractions revisited

- what’s in this collection? (of leaked docs)
  - generate hypothesis
  - summarize clusters
  - explore clusters

- locate evidence (within FOIA dump)
  - verify hypothesis
  - identify clusters/documents
  - locate clusters/documents

- prove non-existence of evidence
  - even harder!
  - exhaustive reading vs filtering out irrelevant

Now what?

• continuing adoption
  – food stamp distribution delays in North Carolina
  – credit card agreements allow repossession
  – this week
    • The Brilliance of Louis C.K.'s Emails: He Writes Like a Politician

• continuing development
  – Knight Foundation funds v5
    • named entity recognition
    • plugin API

https://www.overviewproject.org/
http://overview.ap.org/
Outline

• introduction
  – what’s vis anyway?

• LiveRAC
  – server logs: managed web hosting
    (with AT&T)

• Overview
  – text: visual document mining for journalists
    (with Associated Press)

• big picture and wrapup
Visualization Analysis & Design

http://www.cs.ubc.ca/~tmm/vadbook
Encode

Arrange

→ Express

→ Separate

Order

→ Align

→ Use

Map from categorical and ordered attributes

→ Color

→ Hue

→ Saturation

→ Luminance

Size, Angle, Curvature, ...

→ Shape

→ + \ ● \ □ \ △

→ Motion

Direction, Rate, Frequency, ...

Manipulate

Change

Juxtapose

Filter

Select

Partition

Aggregate

Navigate

Superimpose

Embed

How?

What?

Why?
Channels: Expressiveness types and effectiveness rankings

**Magnitude Channels: Ordered Attributes**

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)

**Identity Channels: Categorical Attributes**

- Spatial region
- Color hue
- Motion
- Shape
Four levels of design

• inverse cases: technique-driven vs. problem-driven work
  – both useful, but learning curve to switch between
Design Study Methodology

Reflections from the Trenches and from the Stacks

joint work with:
Michael Sedlmair, Miriah Meyer

http://www.cs.ubc.ca/labs/imager/tr/2012/dsm/
Design Studies: Lessons learned after 21 of them (+more)
Methodology for Problem-Driven Work

• definitions

• 9-stage framework

• 32 pitfalls and how to avoid them
Wrapup

• two systems analyzed
  – LiveRAC, Overview

• analysis framework big ideas
  – what: data abstraction
    • characterize and derive data
  – why: task abstraction
    • translate from domain-specific to generic
  – how: visual encoding and interaction idioms
    • separate from questions of algorithm design

– scaffolding for thinking systematically about full design space
  • describing existing systems helps with generating new ones
More Information

• this talk
  http://www.cs.ubc.ca/~tmm/talks.html#hope14

• papers, videos, software, talks, courses
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm

• book  (to appear Oct 2014)
  http://www.cs.ubc.ca/~tmm/vadbook

• acknowledgements
  – funding: AT&T, Knight Foundation, NSERC
  – talk feedback: Matt Brehmer