Dimensionality Reduction From Several Angles

Tamara Munzner
Department of Computer Science
University of British Columbia

Outline
• can we design better DR algorithms?
– algorithm for GPU MDS: Glimmer
– algorithm for MDS with costly distances: Glint
• can we build a DR system for real people?
• how should we show people DR results?
• when do people need to use DR?

15

SF LARGE

A Nested Model of Visualization Design and Validation

Tamara Munzner
Department of Computer Science
University of British Columbia

A Nested Model of Visualization Design and Validation


Dimensionality Reduction

• what is it?
– map data from high-dimensional measured space into low-dimensional target space
• when to use it?
– when you can't directly measure what you care about
– true dimensionality of dataset conjectured to be smaller than dimensionality of measurements
– latent factors, hidden variables

Two-Year Cross-Domain Qualitative Study

• in the wild
– HCI term for work in the field with real users
– vs controlled lab setting
• interviewed two dozen high-dim data analysts
– across over a dozen domains and past several years
• final results coming soon
– taxonomy of abstract tasks for DR
– identified significant unmet user needs
• why and how do people use DR?
– overarching question weaving through projects in this talk
– preliminary results from study informed many of them

Questions: A Progression

• can we design DR algorithms/techniques that are better than previous ones?
• can we build a DR system that real people use?
– when do people need to look at DR output?
– how can we figure out what people need?
– how should people look at DR output?
– how can we tell if we're drawing the right picture?
– do metrics match up with human perception?
• why and how do people use DR?

Angles of Attack

• design algorithms
• design systems
• design tools to solve real-world user problems
• evaluate-validate all of these
• create taxonomies to characterize existing things
• benefits of multiple angles
– parallax view of what's important
– outcomes cross-pollinate

Questions and Answers

• can we design DR algorithms/techniques that are better than previous ones?
• can we build a DR system that real people use?
– when do people need to look at DR output?
– how can we figure out what people need?
– how should people look at DR output?
• why and how do people use DR?
– so... how do we answer these questions?
– many validation methods to choose from!

Matching Validation With Design Level

threat: wrong problem
validate: observe and interview target users
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: bad data/operation abstraction
validate: analyze computational complexity
threat: slow algorithm
validate: measure system time/memory
threat: implementation
validate: qualitative/quantitative result image analysis
[ test on any users, informal usability study ]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Where Do We Go From Here?

• no single paper includes all methods of validation
– pick methods based on angle of attack
• in this talk
– cover many different methods and kinds of questions they can help with answering

Even More Questions

• open questions
– how are real people actually using DR tools/techniques?
• does it match up with what we think/hope/assume?
– why are they using it?
• what are their goals and tasks, at abstract level?
– is it working?
• how do their goals match up with implicit assumptions behind different benchmarks?
• do current state of the art tools meet their needs?

Target Space

characterizing the problems of real-world users
abstracting into operations on data types

A Nested Model

of Visualization Design and Validation

Tasks and Challenges

joint work with:
Michael Sedlmair, Matthew Brehmer, Stephen Ingram

work in progress

Four Levels of Design and Validation

• four levels of design problems
– different threats to validity at each level

Dimensionality Reduction

• why do people use DR?
– improve performance of downstream algorithm
– avoid curse of dimensionality
– data analysis
– if look at the output: visual data analysis

Questions: A Progression

• can we design DR algorithms/techniques that are better than previous ones?
• can we build a DR system that real people use?
– when do people need to look at DR output?
– how can we figure out what people need?
– how should people look at DR output?
• why and how do people use DR?

Reduction In the Wild

Tasks and Challenges

joint work with:
Michael Sedlmair, Matthew Brehmer, Stephen Ingram

work in progress

(1)

Dimensionality Reduction

• why do people do DR?
– improve performance of downstream algorithm
– avoid curse of dimensionality
– data analysis
– if look at the output: visual data analysis

Questions and Answers

• can we design DR algorithms/techniques that are better than previous ones?
• can we build a DR system that real people use?
– when do people need to look at DR output?
– how can we figure out what people need?
– how should people look at DR output?
• why and how do people use DR?

Matching Validation With Design Level

threat: wrong problem
validate: observe and interview target users
threat: ineffective encoding/interaction technique
validate: justify encoding/interaction design
threat: bad data/operation abstraction
validate: analyze computational complexity
threat: slow algorithm
validate: measure system time/memory
threat: implementation
validate: qualitative/quantitative result image analysis
[ test on any users, informal usability study ]
validate: lab study, measure human time/errors for operation
validate: test on target users, collect anecdotal evidence of utility
validate: field study, document human usage of deployed system
validate: observe adoption rates

Where Do We Go From Here?

• no single paper includes all methods of validation
– pick methods based on angle of attack
• in this talk
– cover many different methods and kinds of questions they can help with answering

Even More Questions

• open questions
– how are real people actually using DR tools/techniques?
• does it match up with what we think/hope/assume?
– why are they using it?
• what are their goals and tasks, at abstract level?
– is it working?
• how do their goals match up with implicit assumptions behind different benchmarks?
• do current state of the art tools meet their needs?

Target Space

characterizing the problems of real-world users
abstracting into operations on data types

A Nested Model

of Visualization Design and Validation

Tasks and Challenges

joint work with:
Michael Sedlmair, Matthew Brehmer, Stephen Ingram

work in progress

(1)
## MDS: Multidimensional Scaling

- entire family of methods, linear and nonlinear
- classical scaling: minimize strain
  - Nystrom/spectral methods: $O(N)$
- distance scaling: minimize stress
  - nonlinear optimization: $O(N^3)$
  - SMACOF [de Leeuw 1977]
  - force-directed placement: $O(N^k)$
  - Stochastic Force [Chalmers 1996]
- limitations: quality problems from local minima
  - Glimmer goal: $O(N)$ speed and high quality

### Methods and Outcomes

- methods
  - quantitative algorithm benchmarks: speed, quality
- outcomes
  - characterized kinds of datasets where technique yields quality improvements
- then what?
  - saw what real users could do with it after release
  - identified limitations

## MDS Speed on Distance Matrix Data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Author/Year</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic MDS</td>
<td>Torgersen '52</td>
<td>$O(N^3)$</td>
</tr>
<tr>
<td>SMACOF</td>
<td>de Leeuw '77</td>
<td>$O(N^3)$</td>
</tr>
<tr>
<td>Pivot MDS</td>
<td>Brandes '07</td>
<td>$O(kN)$</td>
</tr>
<tr>
<td>Glimmer</td>
<td>Ingram '09</td>
<td>$O(cN)$</td>
</tr>
<tr>
<td>LAMP</td>
<td>Joia '11</td>
<td>$O(kN)$</td>
</tr>
</tbody>
</table>

### MDS Algorithm Types

- Gradient-based Optimization: SMACOF
- Spectral/Analytic: Pivot MDS
- Force-Directed: Glimmer

## Costly Distances

- DR in the Wild revealed many real-world examples

### Methods and Outcomes

- methods
  - algorithm benchmarks
- outcomes
  - dataset characterization different from previous work
  - characterized distance metrics where architecture yields speed improvements
- then what?
  - keep talking to real users as way to discover more unmet needs

---

### Outline

- can we design better DR algorithms?
  - next: how do we get people to use DR properly?
  - move emphasis from solo algorithms to entire system
- can we build a DR system for real people?
  - system that provides guidance: DimStiller
- when do people need to use DR?
- how should we show people DR results?
- why and how do people use DR?
**DimStiller**

*Workflows for Dimensional Analysis and Reduction*

**joint work with:**
Stephen Ingram, Veronika Irvine, Melanie Tory, Steven Bergner, Torsten Möller


---

**Who Might Use DR?**

- DR in the Wild revealed broad set of users

**Methods and Outcomes**

- **methods**
  - usage scenarios: workflows
    - identified several (preliminary DRITW results)
    - built system to accommodate new ones as they're uncovered
  - **outcomes**
    - prototype system: "DR for the rest of us"

- **then what?**
  - who else needs guidance? not just end users!

---

**Global Guidance**

- **which operations and in which order?**

- **Well Defined Tasks**

**Local Guidance**

- **what to do with a given operator?**

**Who Might Use DR?**

- middle ground users benefit from guidance

---

**Spatialization Design**

*Comparing Points and Landscapes*

**joint work with:**
Melanie Tory, David W. Sprague, Fuqu Wu, Wing Yan So

Information Landscapes
• 2D or 3D landscape from set of DR points
  – height based on density
• oddly popular choice in DR
  – despite known occlusion/distortion problems with 3D
  – assertions: pattern recognition, spatial reasoning, familiar

Understanding User Task
• abstract: search involving spatial areas and estimation
  Estimate which grid cell has the most points of the target color
• domain-specific examples
  "Where in the display are people with high incomes?"  
  "Does this area also have high education levels?"
  "Does this area correspond to a particular work sector?"
• non-trivial complexity yet fast response time
• frequent subtask in pilot test of real data analysis

Lab Study: Test Human Response Time and Error
• hypotheses
  – points are better than landscapes
    • result: yes!
    • much better: 2:4 x faster, 5:14 x more accurate
  – 2D landscapes (color only) better than 3D landscapes
    • result: yes
    • significantly faster, no significant difference in accuracy

Cluster Separation
• simple idea
  [Diagram of full overlap, partial overlap, adjacent, separate, distant]

Visual Cluster Separation Measures
• Many cluster separation measures proposed for
  semi-automatic guidance in high-dim data analysis

816 Dataset Instances
• 75 datasets
  – 31 real, 44 synthetic
  – pre-classified
• 4 DR methods
  – PCA
  – Robust PCA
  – Glimmer MDS
  – t-SNE
• 3 visual encoding methods
  – 2D scatterplots, 3D scatterplots,
  2D SPLOMs
  – color-coded by class

User vs. Data Study
• user study
  – previous work on validating cluster measures
    • many users, few datasets
    • missing: dataset variety
• data study
  – few users, many datasets

Centroid Measure
[Diagram of Centroid: 93]

Qualitative Analysis I: Cluster Separation Factors
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Analysis Approach
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Centroid: 93

A Taxonomy of Cluster Separation Factors

Methods and Outcomes
• methods
  – lab study: controlled experiment
• outcomes
  – prescriptive advice at visual encoding level
    • avoid 3D landscapes
• then what?
  – yet more guidance from user studies? not so fast...

High-Level Results

Analysis Approach
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Visual Cluster Separation Measures
• goal: number captures whether human looking at
  layout sees something interesting
• after computation is done, not to refine clustering
• measures checked with user studies

Visual Cluster Separation Measures
• goal: number captures whether human looking at
  layout sees something interesting
• after computation is done, not to refine clustering
• measures checked with user studies

User vs. Data Study
• user study
  – previous work on validating cluster measures
    • many users, few datasets
    • missing: dataset variety
• data study
  – few users, many datasets

Qualitative Analysis I: Cluster Separation Factors
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Analysis Approach
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Centroid: 93

A Taxonomy of Cluster Separation Factors

Methods and Outcomes
• methods
  – lab study: controlled experiment
• outcomes
  – prescriptive advice at visual encoding level
    • avoid 3D landscapes
• then what?
  – yet more guidance from user studies? not so fast...

High-Level Results

Analysis Approach
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Visual Cluster Separation Measures
• goal: number captures whether human looking at
  layout sees something interesting
• after computation is done, not to refine clustering
• measures checked with user studies

Visual Cluster Separation Measures
• goal: number captures whether human looking at
  layout sees something interesting
• after computation is done, not to refine clustering
• measures checked with user studies

User vs. Data Study
• user study
  – previous work on validating cluster measures
    • many users, few datasets
    • missing: dataset variety
• data study
  – few users, many datasets

Qualitative Analysis I: Cluster Separation Factors
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Analysis Approach
• qualitative method out of social science: coding
  – open coding: gradually build/refine code set
  – axial coding: relationships between categories

Centroid: 93

A Taxonomy of Cluster Separation Factors

Methods and Outcomes
• methods
  – lab study: controlled experiment
• outcomes
  – prescriptive advice at visual encoding level
    • avoid 3D landscapes
• then what?
  – yet more guidance from user studies? not so fast...

High-Level Results
Methods and Outcomes
- methods
  - qualitative data study
  - we encourage more work along these lines
- outcomes
  - taxonomy to understand current problems
  - measures
  - taxonomy to advise future development
  - measures, techniques, systems
- then what?
  - from how to help them do DR better to understanding when they need to do it at all

Outline
- how can we design better DR algorithms/techniques?
- how can we build a DR system for real people?
- how should we show people DR results?
- next: continue figuring out what people need
- when do people need to use DR?
  - sometimes they don’t: QuestVis
  - how to figure out when they do or don’t: Design Study Methodology

Hammer Looking for a Nail
- wrong task abstraction: they didn’t need DR!
  - goal mismatch
  - discussion of issues and behavior change from general public
  - not data analysis to understand exact relationships between input and output variables
  - this failure case was one of motivations for nested model
- how can we tell what users actually need?
  - talking to users: necessary but not sufficient
  - we now have some answers!
  - we have proposed a methodology for problem-driven research
    - design studies: build vis tools to solve user problems
    - DR is one of many possible techniques that might be used

Methods and Outcomes
- methods
  - introspection on lessons learned as authors and reviewers
  - extensive literature search
- outcomes
  - prescriptive methodology advice
    - here’s a way to do design studies
    - avoid these pitfalls
- exhortation
  - meta/how-to/reflection papers are worth doing
  - thinking about methods and methodologies is fruitful for any flavor of research!

Centroid: Mapping Assumptions Into Taxonomy
- centroid only reliable if
  - round-ish clusters
  - not more than one dense spot
  - no outliers
  - similar sizes & number of points
  - rarely true for real datasets

Design Studies
- long and winding road with many pitfalls
  - reflections after doing 21 of them
- many successes, a few failures, many lessons learned

How To Do Design Studies
- definitions
- 9-stage framework
- 32 pitfalls and how to avoid them

Work in Progress
- DR in the Wild
  - end point: stay tuned
- DR for journalism
  - Overview project http://overview.ap.org
  - funded by Knight Foundation, collaboration with Stray@AP
    - starting point: Glimmer meets WikiLeaks
      - led us to identify and address more urgent real-world analysis needs
      - iterative rounds of development, deployment, adoption
    - end point: stay tuned
      - Pulitzer Prize finalist story used Overview for data analysis
        (Adam Playford, Nowadays For Their Eyes Only)

Conclusions
- cross-fertilization from attacking DR through different methodological angles
  - scratching original itch often leads to problems that are important and high impact
  - outcomes of evaluation informs how to build
  - grappling with issues of building what studies to run
  - taxonomy creation informs what to build: unsolved problems
- finding mismatches
  - between principles and practice
  - between practice and needs
  - need parallax view of principles, practices, and needs!
• further info
  – http://www.cs.ubc.ca/~tmm/talks.html#charlotte14
  – http://www.cs.ubc.ca/group/infovis

• acknowledgements
  – funding: NSERC Strategic Grant
  – joint work: all collaborators
    - Aaron Barsky, Steven Bergner, Matthew Brehmer, Stephen Ingram, Veronika Irvine, Miriah Meyer, Torsten Möller, Marc W. Sprague, Melanie Tory, Michael Sedlmair, Wing Yan So, Andrada Tatu, Marc Williams, Fuqu Wu
  – feedback on this talk
    - Matthew Brehmer, Joel Ferstay, Stephen Ingram, Torsten Möller, Michael Sedlmair, Jessica Dawson

• hiring opportunity
  – Stephen Ingram (DimStiller, Glimmer, Glint) will finish postdoc soon
  – http://www.cs.ubc.ca/~sfingram
  – available for hacker-analyst job in industry or research lab