Dimensionality Reduction
From Several Angles

Tamara Munzner
Department of Computer Science
University of British Columbia

UBC-Okanagan, Kelowna BC
23 August 2016

http://www.cs.ubc.ca/~tmm/talks.html#kelowna16
@tamaramunzner
Quick Research Overview

- technique-driven work
- problem-driven work
- theoretical foundations
- evaluation

Diagram showing the connections between these concepts.
Technique-driven: Graph Drawing

TreeJuxtaposer

James Slack

Kristian Hildebrand

David Auber (Bordeaux)

TopoLayout
SPF
Grouse
GrouseFlocks
TugGraph
Evaluation: Graph Drawing

Dmitry Nekrasovski
Adam Bodnar
Joanna McGrenere (UBC)
Jessica Dawson (UBC)

Stretch and squish navigation
Search set model of path tracing
Technique-driven: Dimensionality Reduction

Stephen Ingram

Glimmer

DimStiller

Glint

QSNE
Evaluation: Dimensionality Reduction

Melanie Tory

Points vs landscapes for dimensionally reduced data

Michael Sedlmair (UVic)

Guidance on DR & scatterplot choices

Taxonomy of cluster separation factors
Problem-driven: Genomics

Aaron Barsky
(Microbio)

Jenn Gardy

Robert Kincaid
(Agilent)

Miriah Meyer
(Harvard)

Hanspeter Pfister

Cerebral

MizBee

MulteeSum, Pathline
Problem-driven: Genomics, Fisheries Sim

Joel Ferstay
(CBC Cancer)

Cydney Nielsen

Maryam Booshehrian
(SFU)

Torsten Moeller

Variant View

Vismon
Problem-driven: Many Domains

SessionViewer: web log analysis

Peter McLachlan
(AT&T Research)

Stephen North
(AT&T Research)

Heidi Lam
(Google)

Diane Tang
(Google)

LiveRAC: systems time-series
Evaluation: Focus + Context

Heidi Lam (UBC)

Ron Rensink

Robert Kincaid (Agilent)

Distortion impact on search/memory

Separate vs integrated views
Journalism

Matt Brehmer

Stephen Ingram

Jonathan Stray (Assoc Press)

Johanna Fulda (Sud. Zeitung)

Matt Brehmer

Overview

TimeLineCurator
Theoretical Foundations

- Visual Encoding Pitfalls
  - Unjustified Visual Encoding
  - Hammer In Search Of Nail
  - 2D Good, 3D Better
  - Color Cacophony
  - Rainbows Just Like In The Sky

- Strategy Pitfalls
  - What I Did Over My Summer
  - Least Publishable Unit
  - Dense As Plutonium
  - Bad Slice and Dice

Papers Process & Pitfalls

- Visual Encoding Pitfalls
  - Unjustified Visual Encoding
  - Hammer In Search Of Nail
  - 2D Good, 3D Better
  - Color Cacophony
  - Rainbows Just Like In The Sky

- Strategy Pitfalls
  - What I Did Over My Summer
  - Least Publishable Unit
  - Dense As Plutonium
  - Bad Slice and Dice

Design Study Methodology

Michael Sedlmair

Miriah Meyer

Matt Brehmer

Visualization Analysis & Design
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

• how can you tell when you need it?
  – could estimate true dimensionality
DR Example

Tumor Measurement Data

9 Dimensional Measured Space

DR

2 Dimensional Target Space

Malignant

Benign
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
Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences

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

http://www.cs.ubc.ca/labs/imager/tr/2014/DRVisTasks/

Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences
Brehmer, Sedlmair, Ingram, and Munzner.
Motivation

• open questions
  – how are real people actually using DR tools/techniques?
    • does it match up with what we think/hope/assert/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?

• why and how do people use DR?
  – overarching question weaving through projects in this talk
  – preliminary results from study informed many of them
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

• **five abstract tasks**
  – naming synthesized dimensions
  – mapping synthesized dimension to original dimensions
  – verifying clusters
  – naming clusters
  – matching clusters and classes
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 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!
A Nested Model of Visualization Design and Validation

characterizing the problems of real-world users
abstracting into operations on data types
designing visual encoding and interaction techniques
creating algorithms to execute techniques efficiently

Analysis Framework: Four Levels, Three Questions

- **domain** situation
  - who are the target users?
- **abstraction**
  - translate from specifics of domain to vocabulary of vis
- **what** is shown? data abstraction
  - often don’t just draw what you’re given: transform to new form
- **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 Is Validation Difficult?

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

- **Domain situation**
  You misunderstood their needs

- **Data/task abstraction**
  You’re showing them the wrong thing

- **Visual encoding/interaction idiom**
  The way you show it doesn’t work

- **Algorithm**
  Your code is too slow
Validation Solution: Methods From Many Fields

- mismatch: algorithm benchmarks for idiom validation
- mismatch: lab study for abstraction validation

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
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
Outline

• can we design better DR algorithms?
• can we build a DR system for real people?
• how should we show people DR results?
• when do people need to use DR?
Outline

• can we design better DR algorithms?
  – algorithm for GPU MDS: Glimmer
  – algorithm for MDS with costly distances: Glint
  – algorithm for DR for sparse document data: QSNE

• can we build a DR system for real people?
• how should we show people DR results?
• when do people need to use DR?
Glimmer
Multilevel MDS on the GPU

joint work with:
Stephen Ingram, Marc Olano

http://www.cs.ubc.ca/labs/imager/tr/2008/glimmer/

MDS: Multidimensional Scaling

- entire family of methods, linear and nonlinear
- classical scaling: minimize strain
  - Nystrom/spectral methods: $O(N)$
    - limitations: quality for very high dimensional sparse data
- distance scaling: minimize stress
  - nonlinear optimization: $O(N^2)$
    - SMACOF [de Leeuw 1977]
    - force-directed placement: $O(N^2)$
      - Stochastic Force [Chalmers 1996]
      - limitations: quality problems from local minima
- Glimmer goal: $O(N)$ speed and high quality
Glimmer Strategy

- Stochastic force alg suitable for fast GPU port
  - but systematic testing shows it often terminates too soon

- Use as subsystem within new multilevel GPU alg with much better convergence properties
Sparse Dataset (docs): $N=D=28K$

- quality higher
- speed equivalent

16.64 s  stress=0.157

2.17 s  stress=0.928
Methods and Outcomes

• methods
  – quantitative algorithm benchmarks: speed, quality
    • systematic comparison across 1K-10K instances vs a few spot checks
  – qualitative judgements of layout 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
Glint

An MDS Framework for Costly Distance Functions

joint work with:
Stephen Ingram

http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

Glint: An MDS Framework for Costly Distance Functions.
Dimensionality Reduction for Documents with Nearest Neighbour Queries

joint work with:
Stephen Ingram

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

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

Math / Stats

Data Knowledge
Who Might Use DR?

- Best Paper at NIPS
- Took Stats in Undergrad
- What’s a mean?
Who Might Use DR?

- Math / Stats
- Total Information Awareness
- Dropped in lap

Data Knowledge
Who Might Use DR?

Data Knowledge

Math / Stats

Pedagogical
Who Might Use DR?

Math / Stats

Don’t Need Analysis

Data Knowledge
Who Might Use DR?

Math / Stats

Data Knowledge

Well Defined Tasks
Who Might Use DR?

- middle ground users benefit from guidance
Global Guidance

Operator Space

Sloppy, Misunderstood → Compact, Evocative
Global Guidance

PCA

Correlation

Variance

MDS

SPLOM

Operator Space

Sloppy, Misunderstood

Compact, Evocative

Filter

http://www.statmethods.net/advgraphs/images/corrgram3.png
http://en.wikibooks.org/wiki/File:Scree_plot_for_the_initial_dataset_Figure_36.jpg
http://www.iconfinder.com/icondetails/44818/400/data_filter_icon?r=1
http://www.personality-project.org/R/
Global Guidance

- which operations and in which order?

Operator Space

http://www.statmethods.net/advgraphs/images/corrgram3.png
http://en.wikibooks.org/wiki/File:Scree_plot_for_the_initial_dataset_Figure_36.jpg
http://www.iconfinder.com/icondetails/44818/400/data_filter_icon?r=1
http://www.personality-project.org/R/
Local Guidance

• what to do with a given operator?

Sloppy, Misunderstood

Compact, Evocative

Operator Space

PCA

Correlation

Variance

How many principal components?

What do they mean?

MDS

SPLOM

Filter
DimStiller

- pre-built workflows
- sequence of operators
- local guidance for each operator
  - example: estimate true dimensionality with scree plot
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!
• can we design better DR algorithms/techniques?
• can we build a DR system for real people?

– next: more guidance about visual encoding

• how should we show people DR results?
  – visual encoding guidance for system developers: Points vs Landscapes
  – visual encoding guidance for metric developers wrt human perception: Visual Cluster Separation Factors

• when do people need to use DR?
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

Themescape: [http://www.k-n-o-r-z.de/publ/example/retriev1.htm]

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 × faster, 5-14 × more accurate
  - 2D landscapes (color only) better than 3D landscapes (color + height redundantly encoded)
    - result: yes
    - significantly faster, no significant difference in accuracy
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...
A Taxonomy of Visual Cluster Separation Factors

joint work with:
Michael Sedlmair, Andrada Tatu, Melanie Tory

http://www.cs.ubc.ca/labs/imager/tr/2012/VisClusterSep/
Cluster Separation

• simple idea
Visual Cluster Separation Measures

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

Sips et al.: Selecting good views of high-dimensional data using class consistency [EuroVis 2009]

Tatu et al.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data [VAST 2009]
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

  Tatu et al.: Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data [AVI 2010]

• but our attempt to use for guidance showed problems

Good!

No!
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
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
Centroid Measure

Centroid: 93

Good!

Bad!
Analysis Approach

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


• evaluating the measures
  – metric aligns with human judgement?
  – if not: what are the reasons?
Qualitative Analysis I: Cluster Separation Factors

- outlier
- shape
- split
- equidistant points
Analysis Approach

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


• evaluating the measures
  – metric aligns with human judgement?
  – if not: what are the reasons?

• building taxonomy of factors from reasons

• mapping measure failures onto taxonomy
A Taxonomy of Cluster Separation Factors

<table>
<thead>
<tr>
<th>Within-Class Factors</th>
<th>Between-Class Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Count</strong></td>
<td>few</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>small</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>sparse</td>
</tr>
<tr>
<td><strong>Clumpiness</strong></td>
<td>equidistant</td>
</tr>
<tr>
<td><strong>Outlier</strong></td>
<td>none</td>
</tr>
<tr>
<td><strong>Shape</strong></td>
<td>narrow</td>
</tr>
<tr>
<td><strong>Isotropy</strong></td>
<td>round</td>
</tr>
<tr>
<td><strong>Centroid</strong></td>
<td>evocative</td>
</tr>
<tr>
<td><strong>Variance of Count</strong></td>
<td>similar</td>
</tr>
<tr>
<td><strong>Variance of Size</strong></td>
<td>similar</td>
</tr>
<tr>
<td><strong>Variance of Density</strong></td>
<td>similar</td>
</tr>
<tr>
<td><strong>Mixture</strong></td>
<td>random</td>
</tr>
<tr>
<td><strong>Split</strong></td>
<td>contiguous</td>
</tr>
<tr>
<td><strong>Variance of Shape</strong></td>
<td>similar</td>
</tr>
<tr>
<td><strong>Inner-Outer Position</strong></td>
<td>non-existent</td>
</tr>
<tr>
<td><strong>Class Separation</strong></td>
<td>full overlap</td>
</tr>
</tbody>
</table>
High-Level Results

- Failure cases
- Ok

### All (816)

<table>
<thead>
<tr>
<th></th>
<th>Centroid</th>
<th>Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure</td>
<td>49%</td>
<td>51%</td>
</tr>
<tr>
<td>Ok</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Only real (296)

<table>
<thead>
<tr>
<th></th>
<th>Centroid</th>
<th>Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure</td>
<td>68%</td>
<td>65%</td>
</tr>
<tr>
<td>Ok</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### All failure cases

<table>
<thead>
<tr>
<th></th>
<th>Centroid</th>
<th>Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives</td>
<td>68%</td>
<td>85%</td>
</tr>
<tr>
<td>False Negatives</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Centroid Failure Example

- big classes overspread small ones

Red: 77 (Good)
Problem: FP
Data: Gaussian, synthetic
DR: MDS
### Relevant Taxonomy Factors

#### Within-Class Factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>few, many</td>
</tr>
<tr>
<td>Size</td>
<td>small, large</td>
</tr>
<tr>
<td>Density</td>
<td>sparse, dense</td>
</tr>
<tr>
<td>Clumpiness</td>
<td>equidistant, uniformly random, one dense spot, many dense spots, clumpy</td>
</tr>
<tr>
<td>Outlier</td>
<td>none, many</td>
</tr>
<tr>
<td>Shape</td>
<td>narrow, curvy</td>
</tr>
<tr>
<td>Curvature</td>
<td>round, isotropy</td>
</tr>
<tr>
<td>Centroid</td>
<td>evocative, misleading</td>
</tr>
</tbody>
</table>

#### Between-Class Factors

<table>
<thead>
<tr>
<th>Class/Point Count</th>
<th>few classes, many points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance of Count</td>
<td>similar, different</td>
</tr>
<tr>
<td>Variance of Size</td>
<td>similar, different</td>
</tr>
<tr>
<td>Variance of Density</td>
<td>similar, different</td>
</tr>
<tr>
<td>Mixture</td>
<td>random, equidistant, interwoven</td>
</tr>
<tr>
<td>Split</td>
<td>contiguous, split</td>
</tr>
<tr>
<td>Inner-Outer Position</td>
<td>non-existent, existent</td>
</tr>
<tr>
<td>Class Separation</td>
<td>full overlap, partial overlap, adjacent, separate, distant</td>
</tr>
</tbody>
</table>

#### Variance

- Variance of Size: similar, different

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
Related Work

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
Reflections on

**QuestVis**

A Visualization System for an Environmental Sustainability Model

joint work with:
Aaron Barsky, Matt Williams

http://www.cs.ubc.ca/labs/imager/tr/2011/QuestVis/

Reflections on QuestVis: A Visualization System for an Environmental Sustainability Model
Munzner, Barsky, Williams.
Application Domain: Sustainability

• user data: sustainability simulation model
  – high-dimensional inputs/outputs
• our decision: show relationship between input choices and output indicators with linked views including DR layout
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 as one of many possible techniques that might be used
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

• 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
Pitfall Example: Premature Publishing

**technique-driven**

Must be first!

**problem-driven**

Am I ready?


http://www.alaineknipes.com/interests/violin_concert.jpg
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!
Conclusions

• cross-fertilization from attacking DR through different methodological angles
  – scratching own itches often leads to problems that are important and high impact
    • outcomes of evaluation informs how to build
    • grappling with issues of building informs 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!
Thanks and Questions

• this talk
  – http://www.cs.ubc.ca/~tmm/talks.html#kelownan16

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

• book: Visualization Analysis & Design
  – http://www.cs.ubc.ca/~tmm/vadbook

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