Dimensionality Reduction: From Several Angles

Tamara Munzner
Department of Computer Science
University of British Columbia

Quick Research Overview

- 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

Technique-driven: Dimensionality Reduction

- what is it?
  - map data from high-dimensional measured space into low-dimensional target space
- how can you tell when you need it?
  - could estimate true dimensionality

Problem-driven: Many Domains

Problem-driven: Genomics

Problem-driven: Genomics, Fisheries Sim

Evaluation: Dimensionality Reduction

Separate vs integrated views
Distortion impact on search/memory

Problem-driven: Focus+Context

Evaluation: Focus+Context

Technique-driven: Graph Drawing

Visualization: Sim

Dimensionality Reduction: Visually Reduced Data

Malignant
Benign

Dr. Michael Sedlmair, Matthew Brehmer, Stephen Ingram

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

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

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

Michael Sedlmair, Matthew Brehmer, Stephen Ingram

http://www.cs.ubc.ca/~mtn/talks.html#kelowna16
@tammamunzner

387164
386455
Motivation

- open questions
  - how are real people actually using DR tools/techniques?
  - does it match up with what we think/expect/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
- 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?
- why and how do people use DR?
- so... how do we answer these questions?
  - many validation methods to choose from!

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
- abstract tasks
  - naming synthesized dimensions
  - mapping synthesized dimension to original dimensions
  - verifying clusters
  - naming clusters
  - matching clusters and classes

Why Is Validation Difficult?

- four levels of design problems
  - different threats to validity at each level
  - 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
    - is it valid? idiom
    - how is it shown?
      - visual encoding idiom: how to draw
      - interaction idiom: how to manipulate
    - algorithm
      - efficient computation

Validation Solution: Methods From Many Fields

- anthropological
  - design
  - computer science
  - cognitive psychology
  - data analysis
  - visual encoding
  - interaction idiom
  - idiom
  - visual encoding idiom
  - efficient computation

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

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
  - you’re showing them the wrong thing
- idiom
- how is it shown?
- visual encoding idiom
- interaction idiom
- algorithm
- efficient computation

Why 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: Glimt
  - 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?

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: Glimt
  - 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?

Outline

- can we design better DR algorithms?
  - algorithm for GPU MDS: Glimmer
  - algorithm for MDS with costly distances: Glimt
  - 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?

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?

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

Methods and Outcomes

- methods
  - quantitative algorithm benchmarks: speed, quality
  - systematic comparison across 1K-10K instances vs 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

Glimmer

Multilevel MDS on the GPU

join work with
Stephen Ingram, Marc Olano

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

Glimmer MDS on the GPU

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

A Nested Model of Visualization Design and Validation

Who Might Use DR?
• DR in the Wild revealed broad set of users

Math / Stats
Data Knowledge

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.
joint work with:
Stephen Ingram
33
http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

An MDS Framework for Costly Distance Functions

Middle Ground Users
• middle ground users benefit from guidance

Math / Stats
Data Knowledge

Operator Space

Global Guidance
• which operations and in which order?

Local Guidance
• what to do with a given operator?

DimStillerWorkflows for Dimensional Analysis and Reduction
joint work with:
Stephen Ingram, Veronica Irvine, Melanie T ory, Steven Bergner, Torsten Möller
36

GlintAn MDS Framework for Costly Distance Functions
GlintAn MDS Framework for Costly Distance Functions.
joint work with:
Stephen Ingram
33
http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

Glint: An MDS Framework for Costly Distance Functions.
joint work with:
Stephen Ingram
33
http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

An MDS Framework for Costly Distance Functions

Middle Ground Users
• middle ground users benefit from guidance

Math / Stats
Data Knowledge

Operator Space

Global Guidance
• which operations and in which order?

Local Guidance
• what to do with a given operator?

DimStillerWorkflows for Dimensional Analysis and Reduction
joint work with:
Stephen Ingram, Veronica Irvine, Melanie T ory, Steven Bergner, Torsten Möller
36

GlintAn MDS Framework for Costly Distance Functions
GlintAn MDS Framework for Costly Distance Functions.
joint work with:
Stephen Ingram
33
http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

Glint: An MDS Framework for Costly Distance Functions.
joint work with:
Stephen Ingram
33
http://www.cs.ubc.ca/labs/imager/tr/2012/Glint/

An MDS Framework for Costly Distance Functions

Middle Ground Users
• middle ground users benefit from guidance

Math / Stats
Data Knowledge

Operator Space

Global Guidance
• which operations and in which order?

Local Guidance
• what to do with a given operator?

DimStillerWorkflows for Dimensional Analysis and Reduction
joint work with:
Stephen Ingram, Veronica Irvine, Melanie T ory, Steven Bergner, Torsten Möller
36
Methods and Outcomes

- usage scenarios: workflows
- identified several (preliminary DR/TVW 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!

Understanding User Task

- abstract: search involving spatial areas and estimation
- which grid cell has the most points of the target color

- domain-specific examples
  - “Where do people with high incomes live?”
  - “Does this area have high education levels?”

- 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
  - 2D landscapes (color only) better than 3D landscapes
    - result: yes

- significantly faster, no significant difference in accuracy

Cluster Separation

- simple idea

Visual Cluster Separation Measures

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

- goal: number captures whether human looking at layout sees something interesting
- if so, then computations are done, to refine clustering
- measures checked with user studies

- but our attempt to use for guidance showed problems

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

- qualitative method out of social science: coding
- open coding: gradually build/define code set
- axial coding: relationships between categories

- evaluating the measures
- metric aligns with human judgment?
- if not: what are the reasons?

Analysis Approach

- qualitative method out of social science: coding
  - open coding: gradually build/define code set
  - axial coding: relationships between categories

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

Methods and Outcomes

A Taxonomy of Cluster Separation Factors

High-Level Results

Centroid Failure Example

- big classes overspread small ones

Relevant Taxonomy Factors

Centroid: Mapping Assumptions Into Taxonomy

Related Work

- Scagnostics [Wilkinson et al. 2005]
  - mathematical description and algorithmic instantiation vs human perception

A Visualization System for an Environmental Sustainability Model

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

Reflections on QuestVis

- metaphor: horse race vs. music debut

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

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

Centroid: Mapping Assumptions Into Taxonomy

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 they need to do it at all

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

Pitfall Example: Premature Publishing

- technique-driven problem-driven

Must be first!

Am I ready?
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#kelowna16
• 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, Andraide 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