Dimensionality Reduction From Three Angles

Tamara Munzner
Department of Computer Science
University of British Columbia

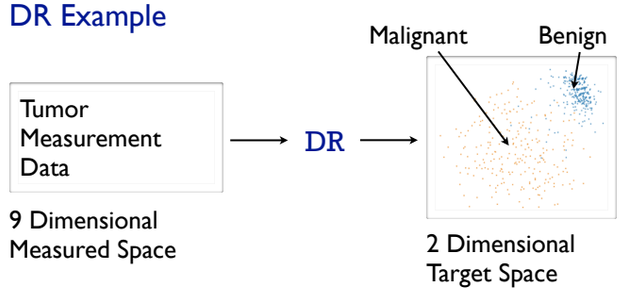
2014 SIAM Data Mining Workshop on Exploratory Data Analysis
26 Apr 2014

http://www.cs.ubc.ca/~tmm/talks.html#eda14

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
- what's the goal?
  - improve performance of downstream algorithm
    - avoid curse of dimensionality
  - data analysis
    - if look at the output: **visual data analysis**

DR Example



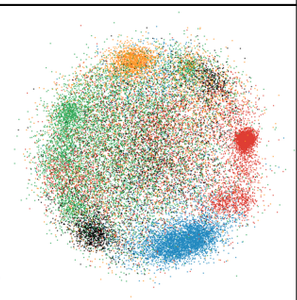
Angles of Attack

- invent algorithms
- build 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?
  - 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?



Glimmer

Multilevel MDS on the GPU

joint work with:
Stephen Ingram, Marc Olano

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

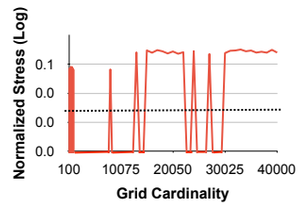
Glimmer: Multilevel MDS on the GPU.
Ingram, Munzner, Olano. IEEE TVCG 15(2):249-261, 2009.

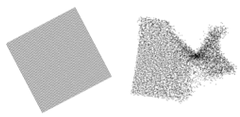
MDS: Multidimensional Scaling

- entire family of methods, linear and nonlinear
- classical scaling: minimize strain
  - Nystrom/spectral methods: O(N)
    - Landmark MDS [de Silva 2004], PivotMDS [Brandes & Pich 2006]
  - 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

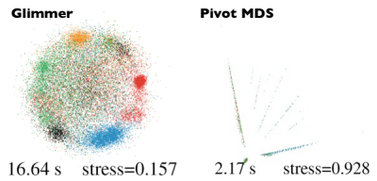


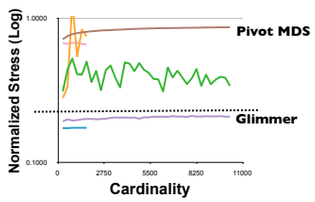


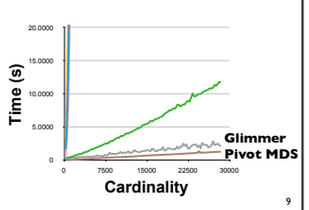
- Use as subsystem within new multilevel GPU alg with much better convergence properties

Sparse Dataset (docs): N=D=28K

- quality higher
- speed equivalent

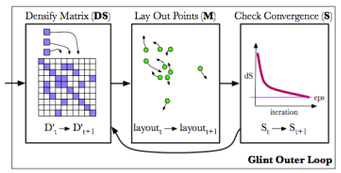






Methods and Outcomes

- methods
  - quantitative algorithm benchmarks: speed, quality
    - systematic comparison across 1K-10K instances vs 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.
Ingram, Munzner. Proc. SIGRAD 2012.

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
- how should we show people DR results?

DimStiller

Workflows for Dimensional Analysis and Reduction

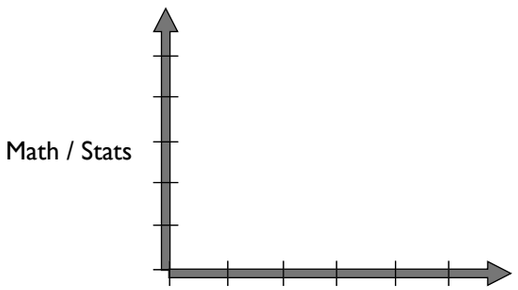
joint work with:
Stephen Ingram, Veronika Irvine, Melanie Tory, Steven Bergner, Torsten Möller

http://www.cs.ubc.ca/labs/imager/tr/2010/DimStiller/

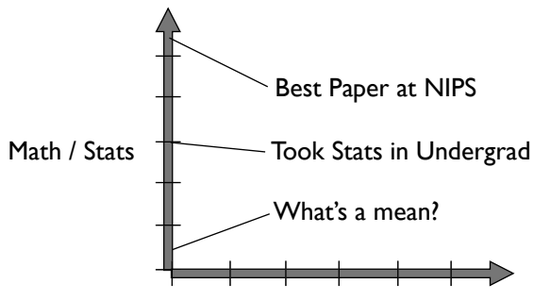
DimStiller: Workflows for dimensional analysis and reduction.
Ingram, Munzner, Irvine, Tory, Bergner, Moeller. Proc. VAST 2010, p 3-10.

Who Might Use DR?

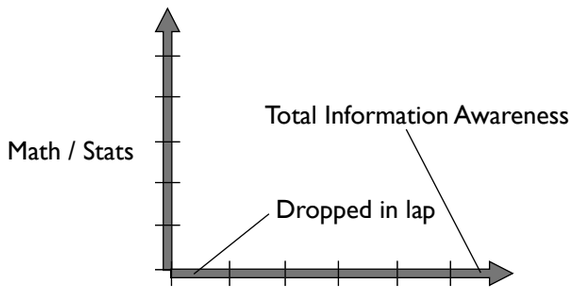
- DR in the Wild revealed broad set of users



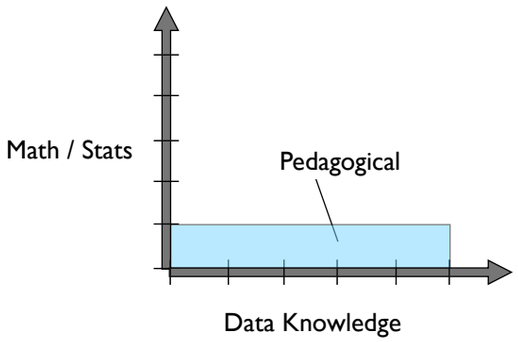
Who Might Use DR?



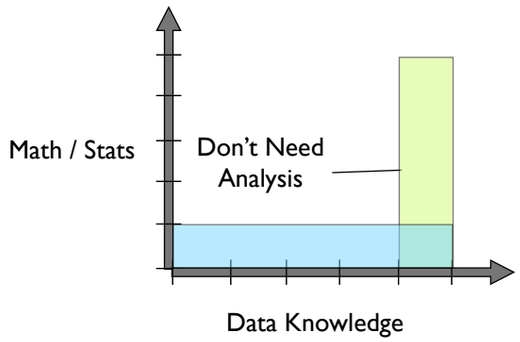
Who Might Use DR?



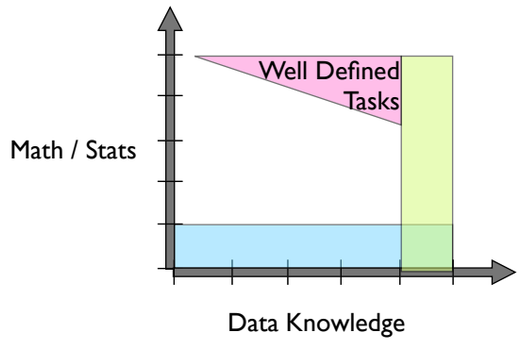
Who Might Use DR?



Who Might Use DR?

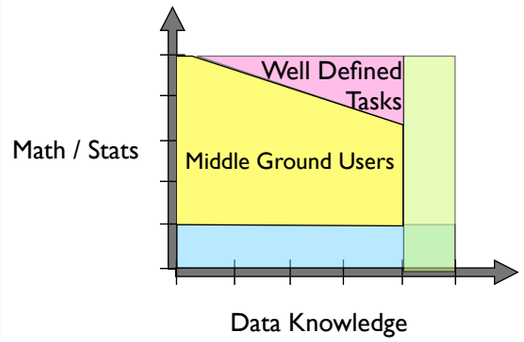


Who Might Use DR?

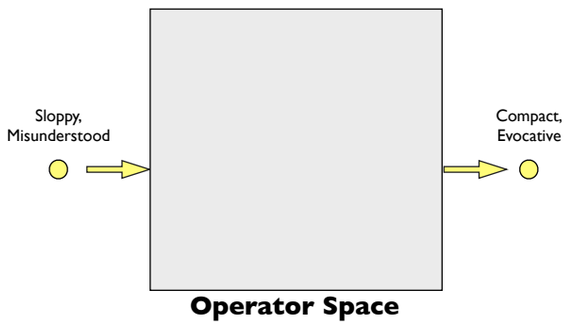


Who Might Use DR?

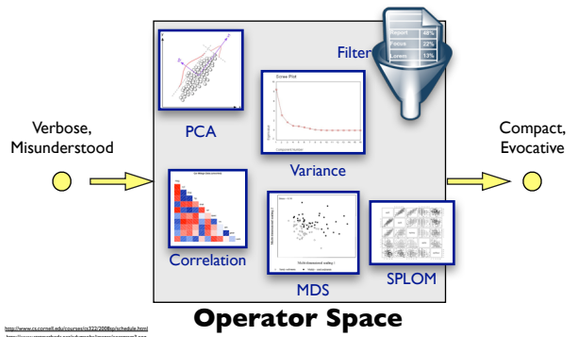
- middle ground users benefit from guidance



Global Guidance

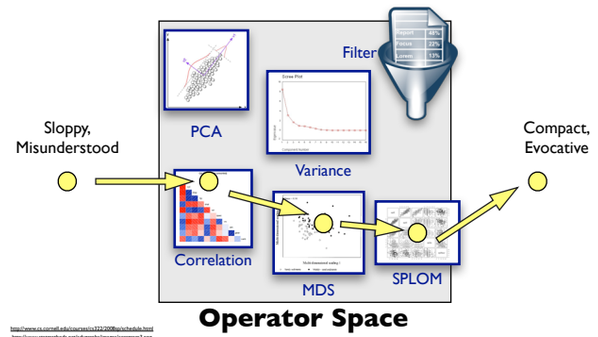


Global Guidance

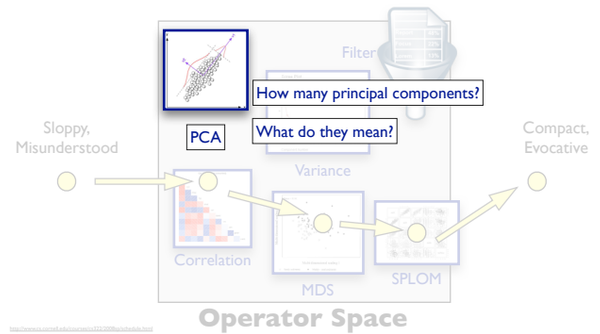


Global Guidance

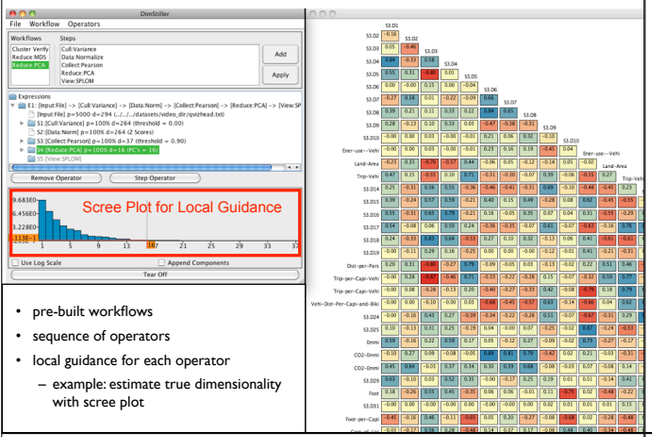
- which operations and in which order?



Local Guidance



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 field study 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!

Outline

- 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 metric developers wrt human perception: Visual Cluster Separation Factors
  - (for system developers: Scatterplot and DR Technique Choices)
  - (visual encoding guidance for system developers: Points vs Landscapes)

A Taxonomy of

Visual Cluster Separation Factors

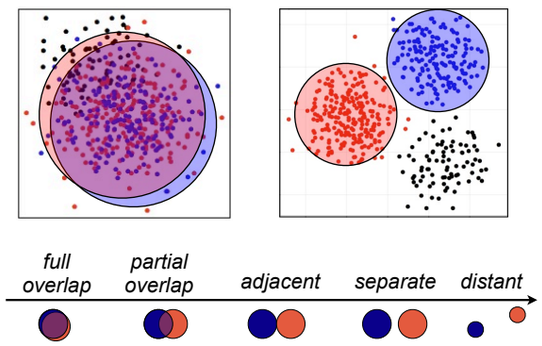
joint work with:
Michael Sedlmair, Andrada Tatu, Melanie Tory

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

A Taxonomy of Visual Cluster Separation Factors.
Sedlmair, Tatu, Munzner, Tory. Computer Graphics Forum 31(3):1335-1344, 2012 (Proc. EuroVis 2012).

Cluster Separation

- simple idea

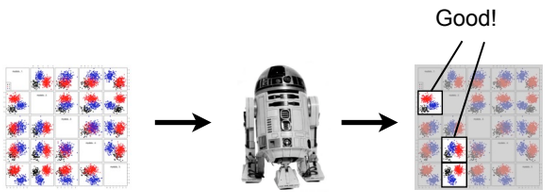


Visual Cluster Separation Measures

- 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]

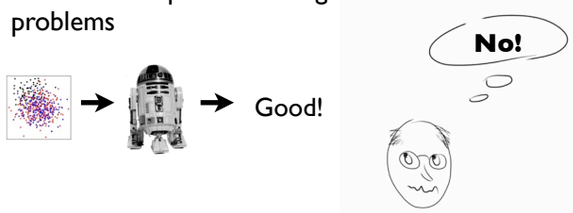


Visual Cluster Separation Measures

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

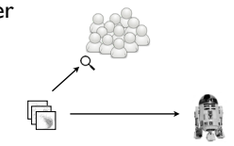
Tatu et al.: Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data [AVI 2010]

- but our attempt to use for guidance showed problems

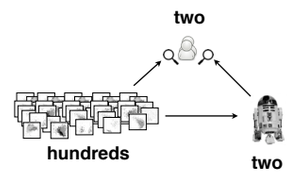


User vs. Data Study

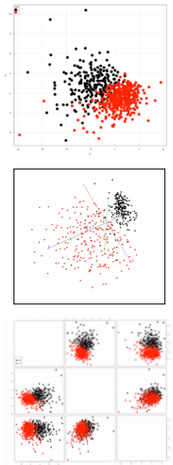
- user study
  - previous work on validating cluster measures
  - many users, few datasets
  - missing: dataset variety



- 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

Analysis Approach

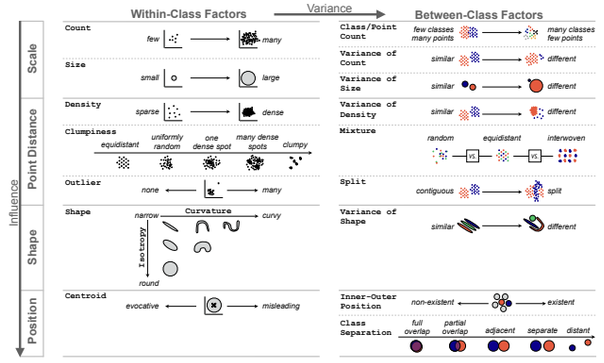
- qualitative method out of social science: coding
  - open coding: gradually build/refine code set
  - axial coding: relationships between categories

Charmaz, K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. 2006.

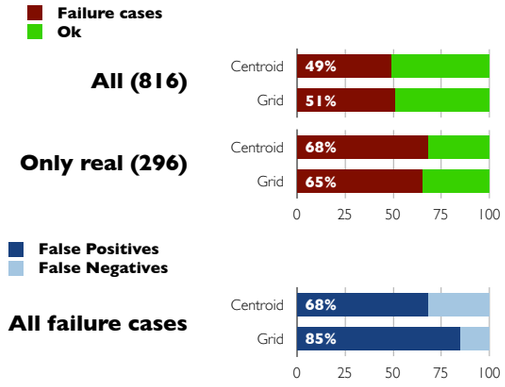
Furniss, D., Blandford, A., Curzon, P. and Mary, Q. (2011). Confessions from a grounded theory PhD: experiences and lessons learnt. Proc. ACM CHI 2011, p 113-122.

- 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

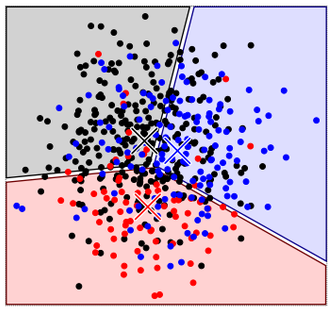


High-Level Results



Centroid Failure Example

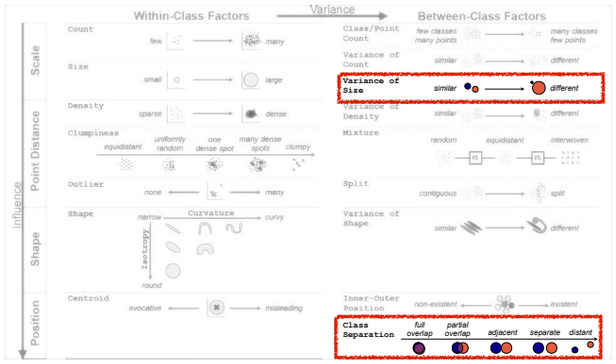
- big classes overspread small ones



Red: **77 (Good)**
Problem: **FP**

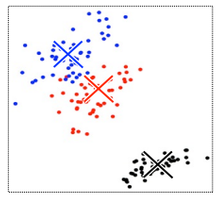
Data: Gaussian, synthetic
DR: MDS

Relevant Taxonomy Factors



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



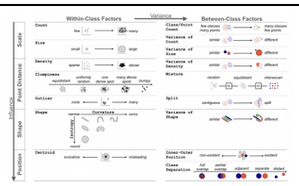
- rarely true for real datasets

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

Empirical Guidance on

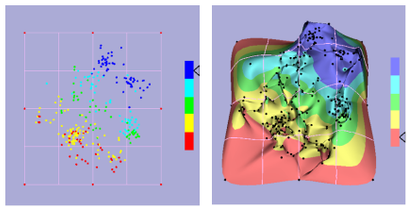
Scatterplot and Dimension Reduction Technique Choices



joint work with:
Michael Sedlmair, Melanie Tory

http://www.cs.ubc.ca/labs/imager/tr/2013/ScatterplotEval/

Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices.
Sedlmair, Munzner, Tory. IEEE TVCG 19(12):2634-2643 (Proc. InfoVis 2013).



Spatialization Design

Comparing Points and Landscapes

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

http://webhome.cs.uvic.ca/~mtory/publications/infovis2007.pdf

Spatialization Design: Comparing Points and Landscapes.
Tory, Sprague, Wu, So, and Munzner.
IEEE TVCG 13(6):1262–1269, 2007 (Proc. InfoVis 07).

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?
  - elsewhere: continue figuring out what people need
- (when do people need to use DR?)

Work in Progress

- DR in the Wild
  - multi-year cross-domain qualitative field study
- 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 unmet real-world analysis needs
      - new technique developed, deployed, adopted
    - ending point: stay tuned...

Conclusions

- cross-fertilization from attacking DR through different methodological angles
  - scratching own itches to find high-impact problems
    - 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

- further info
  - this talk: http://www.cs.ubc.ca/~tmm/talks.html#eda14
  - long version: http://www.cs.ubc.ca/~tmm/talks.html#utah13
  - http://www.cs.ubc.ca/group/infovis
    - papers, videos, open-source software (including Glimmer and DimStiller)
- acknowledgements
  - funding: NSERC Strategic Grant
  - joint work: all collaborators
    - Steven Bergner, Matthew Brehmer, Stephen Ingram, Veronika Irvine, Torsten Möller, Marc Olano, Michael Sedlmair, Andrada Tatu
  - 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
    - in fall 2014 after postdoc