Dimensionality Reduction From Several Angles

Tamara Munzner

Department of Computer Science
University of British Columbia

University of Sydney, Sydney, Australia
9 June 2015

http://www.cs.ubc.ca/~tmm/talks.html#sydney15

@tamaramunzner

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

Estimating true dimensionality

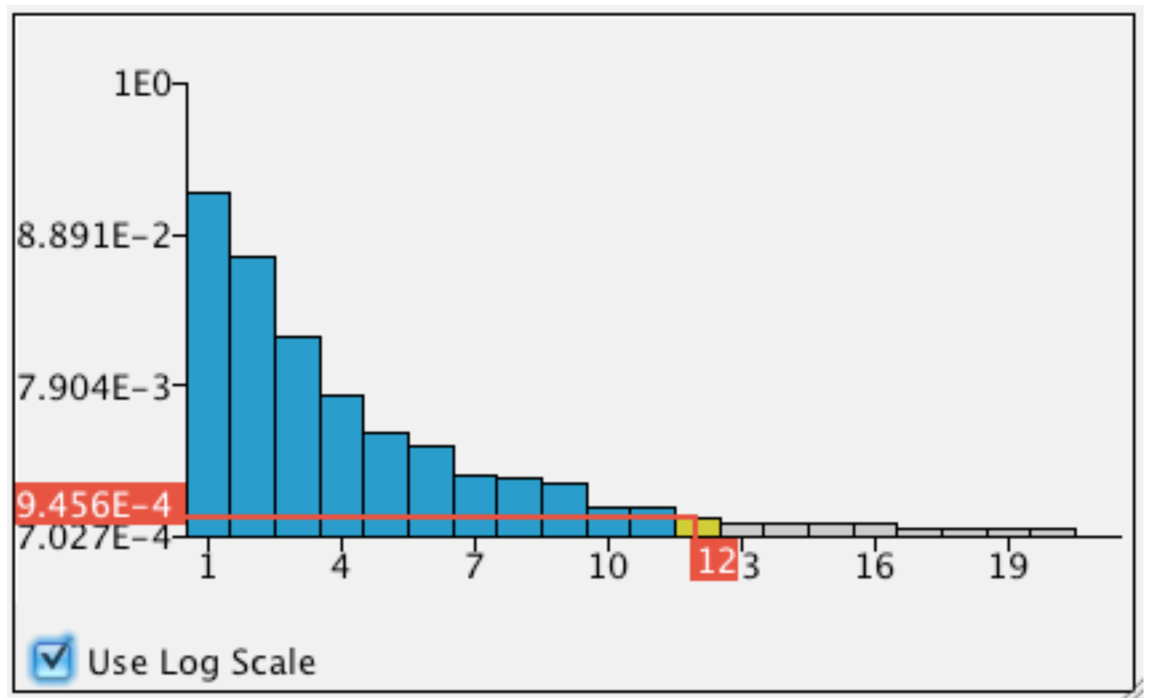
- error for low-dim projection vs high-dim projection
- no single correct answer; many metrics proposed
  - cumulative variance that is not accounted for
  - strain: match variations in distance (vs actual distance values)
  - stress: difference between interpoint distances in high and low dims

$$stress(D, \Delta) = \sqrt{\frac{\sum_{ij}(d_{ij} - \delta_{ij})^2}{\sum_{ij} \delta_{ij}^2}}$$

- D: matrix of lowD distances
- Δ: matrix of hiD distances δ_{ij}

Showing dimensionality estimates

- scree plots as simple way: error against # attribs



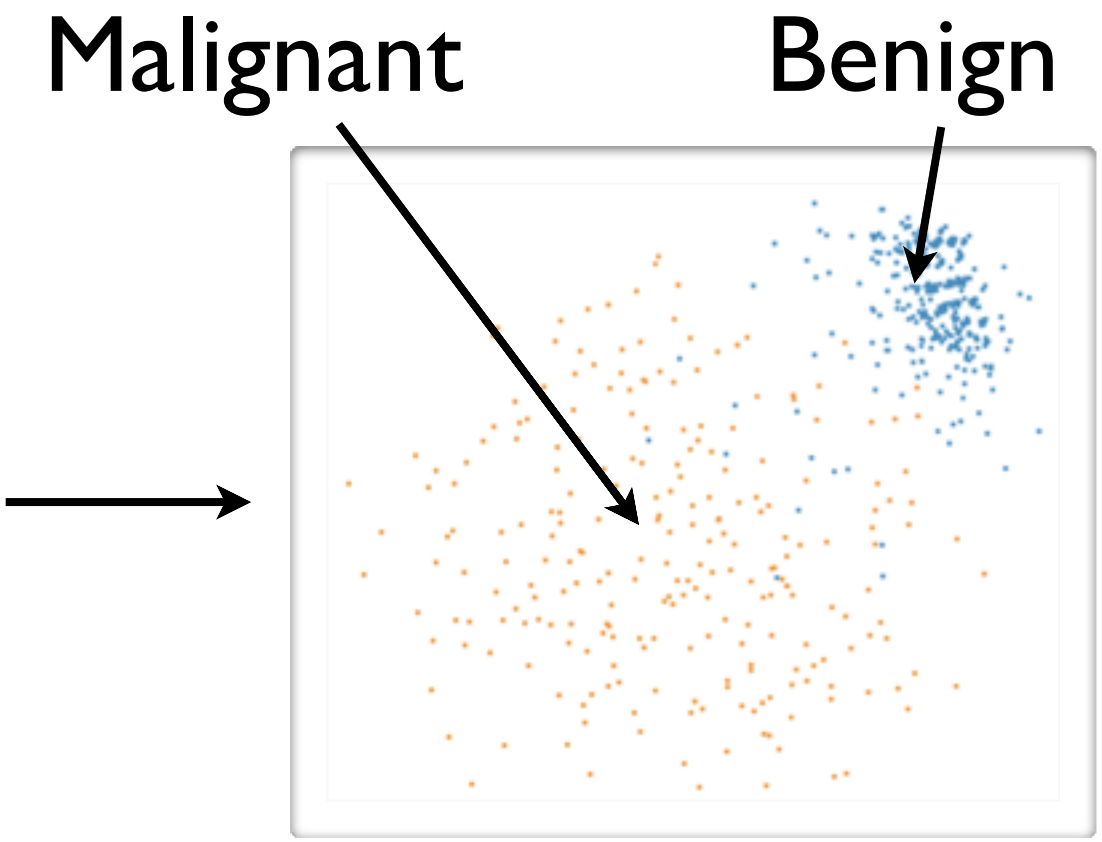
  – original dataset: 294 dims
  – estimate: almost all variance preserved with < 20 dims
  –

DR Example

Tumor
Measurement
Data

9 Dimensional
Measured Space

DR



2 Dimensional
Target Space

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.
Proc. Beyond Time & Errors: Novel Evaluation Methods For Information Visualization (BELIV) 2014, p.1-8.

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

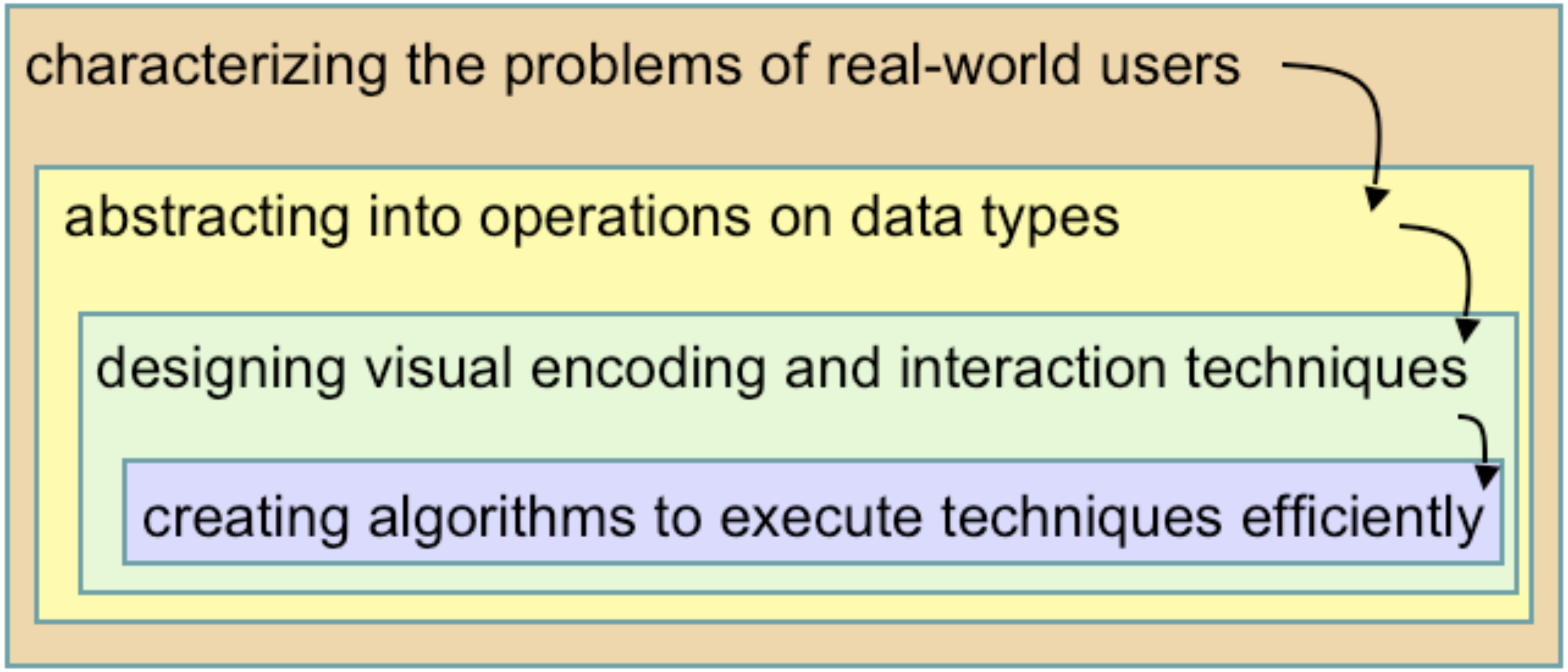
Two-Year Cross-Domain Qualitative Study

- 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

http://www.cs.ubc.ca/labs/imager/tr/2009/NestedModel/

A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).

Four Levels of Design and Validation

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

problem characterization:
you misunderstood their needs

data/task abstraction:
you're showing them the wrong thing

visual encoding / interaction techniques:
the way you show it doesn't work

algorithm:
your code is too slow

Nested Levels of Design and Validation

domain situation:

observe target users using existing tools

data/task abstraction:

encoding/interaction idiom:

justify design wrt alternatives

algorithm:

measure system time

analyze computational complexity

analyze results qualitatively

measure human time with lab experiment ("user study")

observe target users post-deployment ("field study")

measure adoption

- mismatch: cannot show idiom good with system timings
- mismatch: cannot show abstraction good with lab study

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

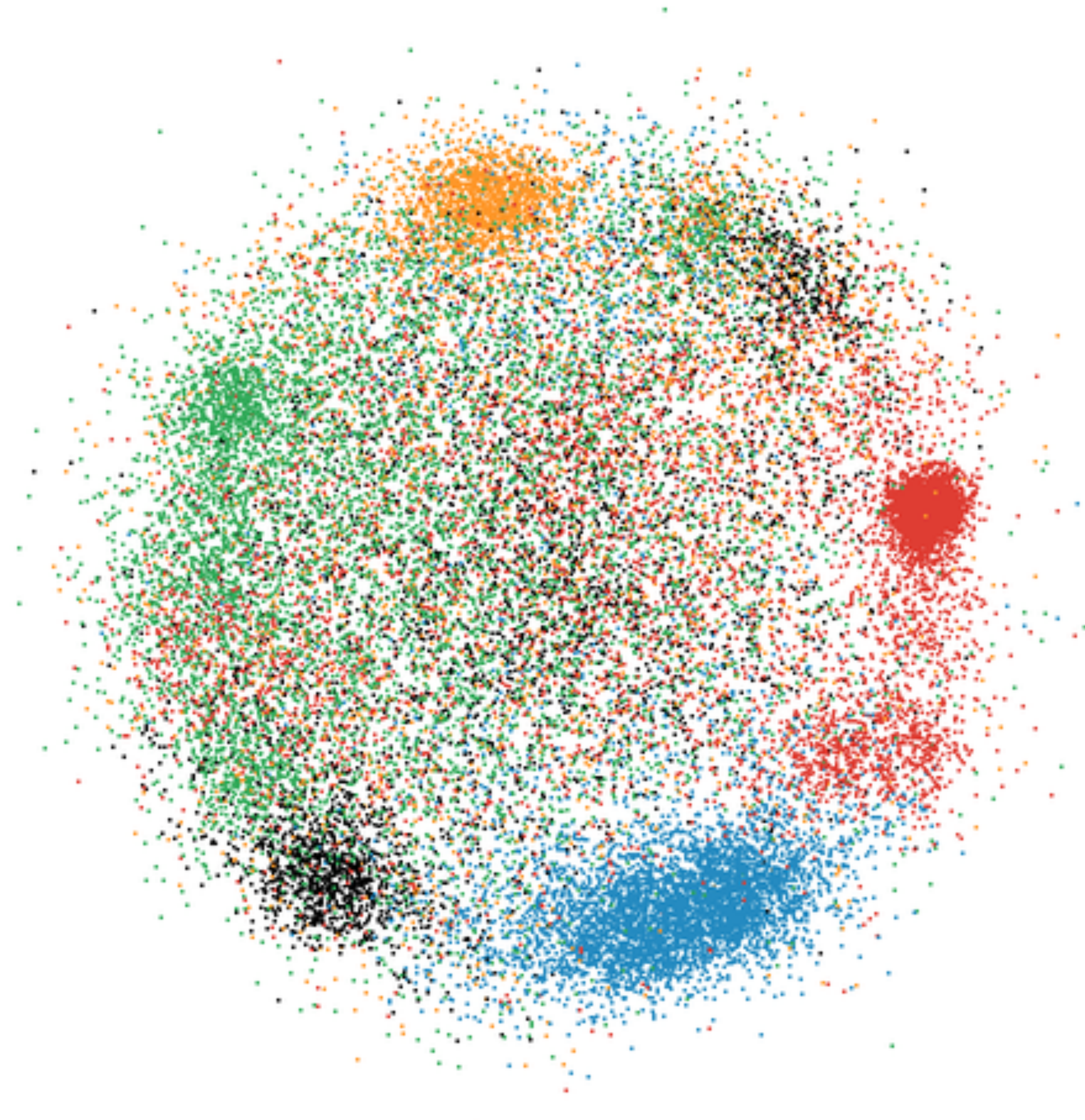
Outline

- 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

- 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/

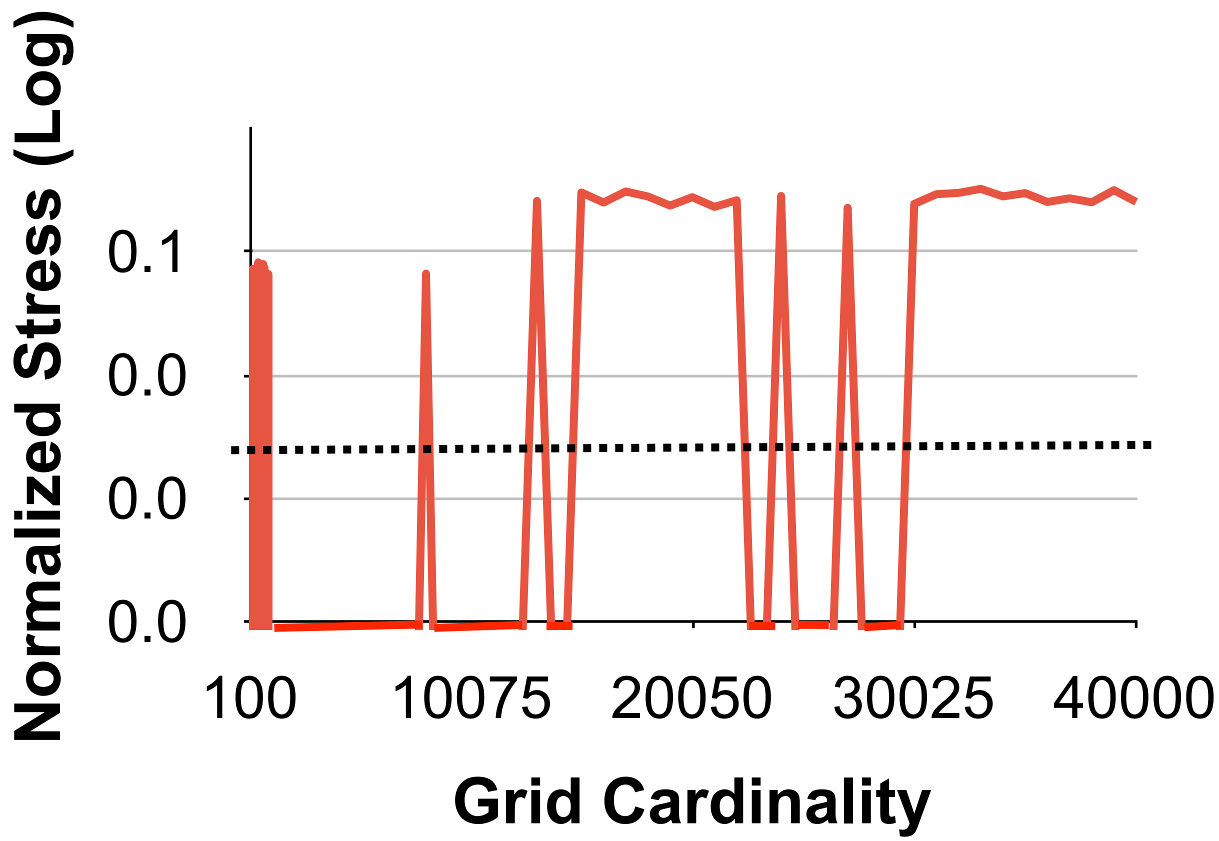
Glimmer: Multilevel MDS on the GPU.
Ingram, Munzner, Olano. IEEE TVCG 15(2):249-261, 2009.

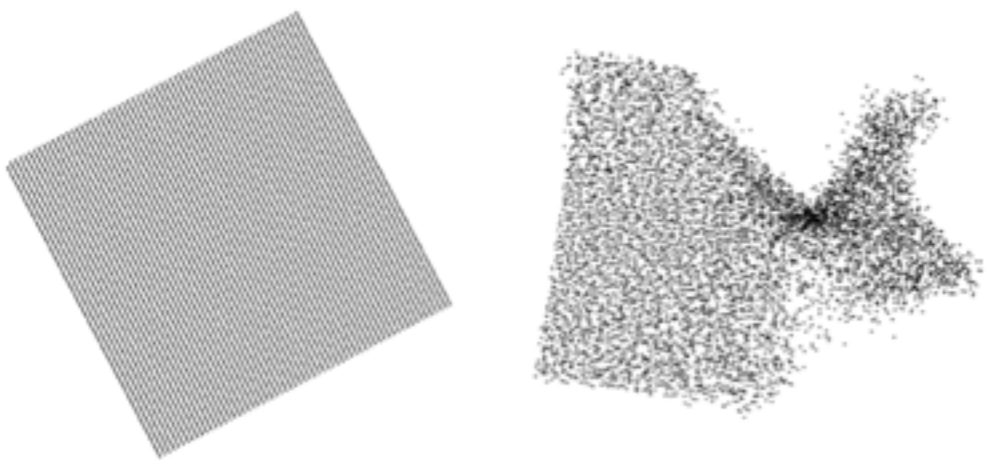
MDS: Multidimensional Scaling

- 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

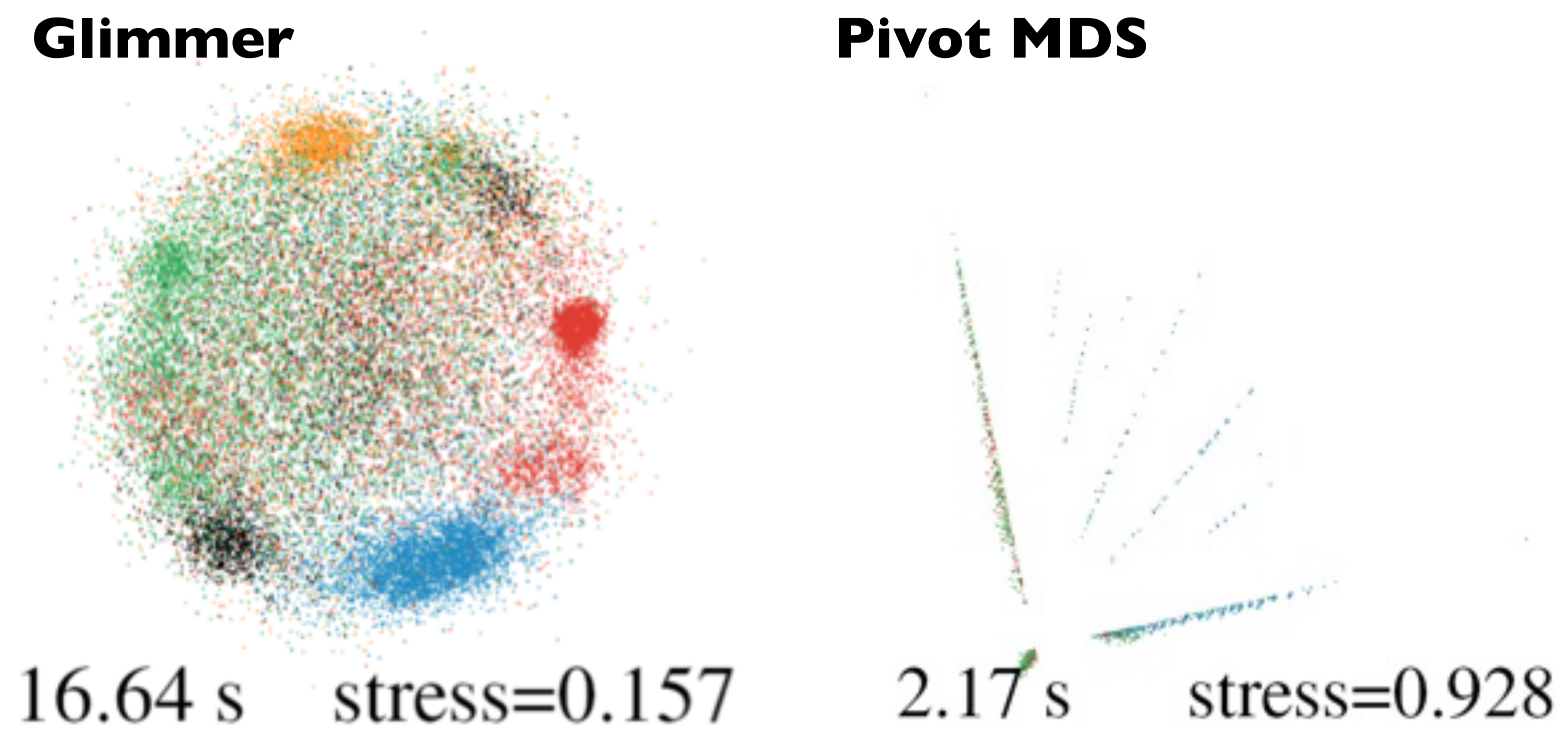


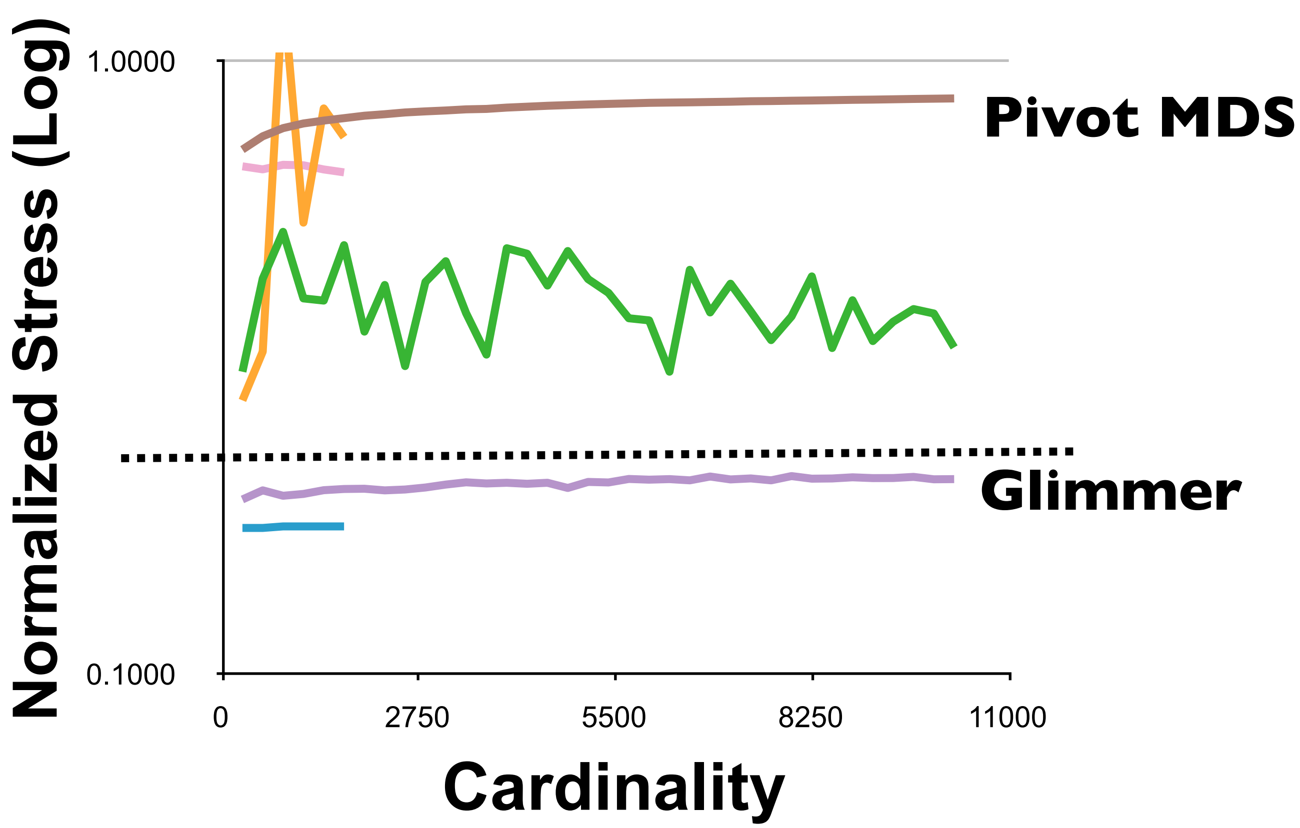


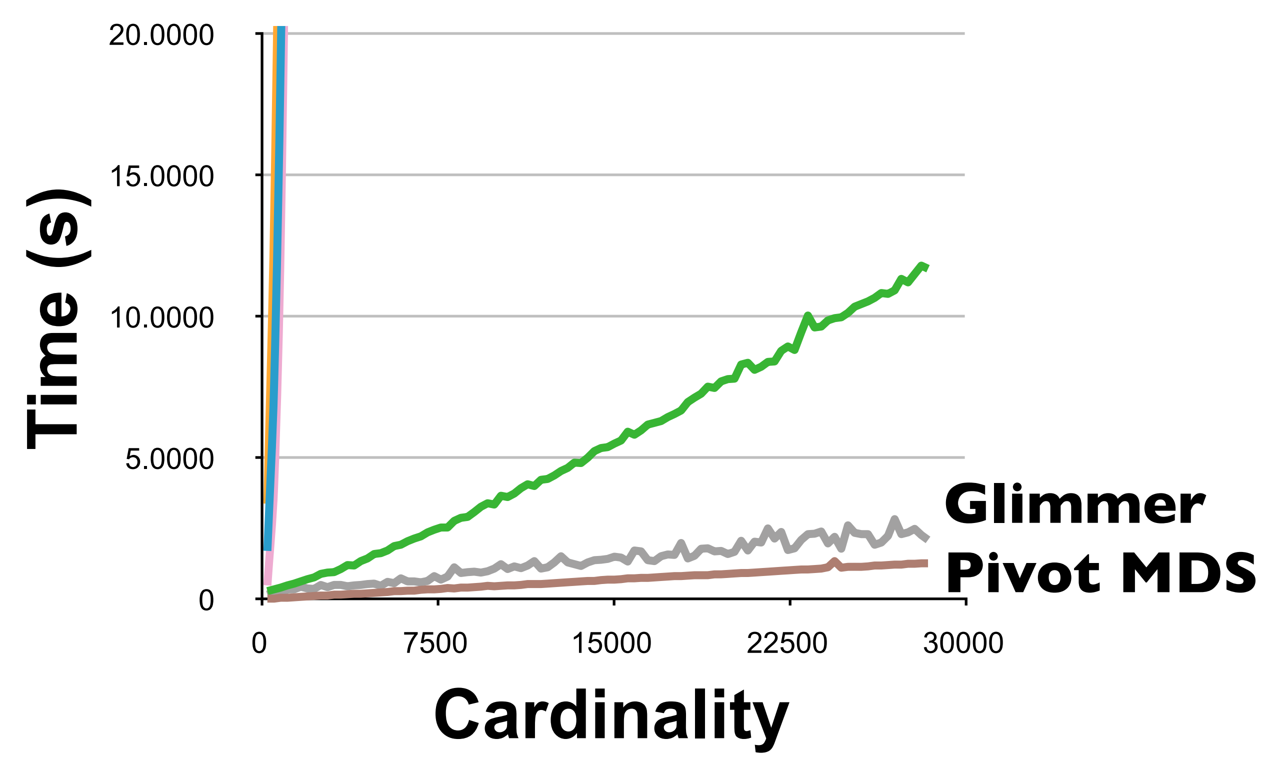
- Use as subsystem within new multilevel GPU alg with much better convergence properties

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

– quality higher
– speed equivalent

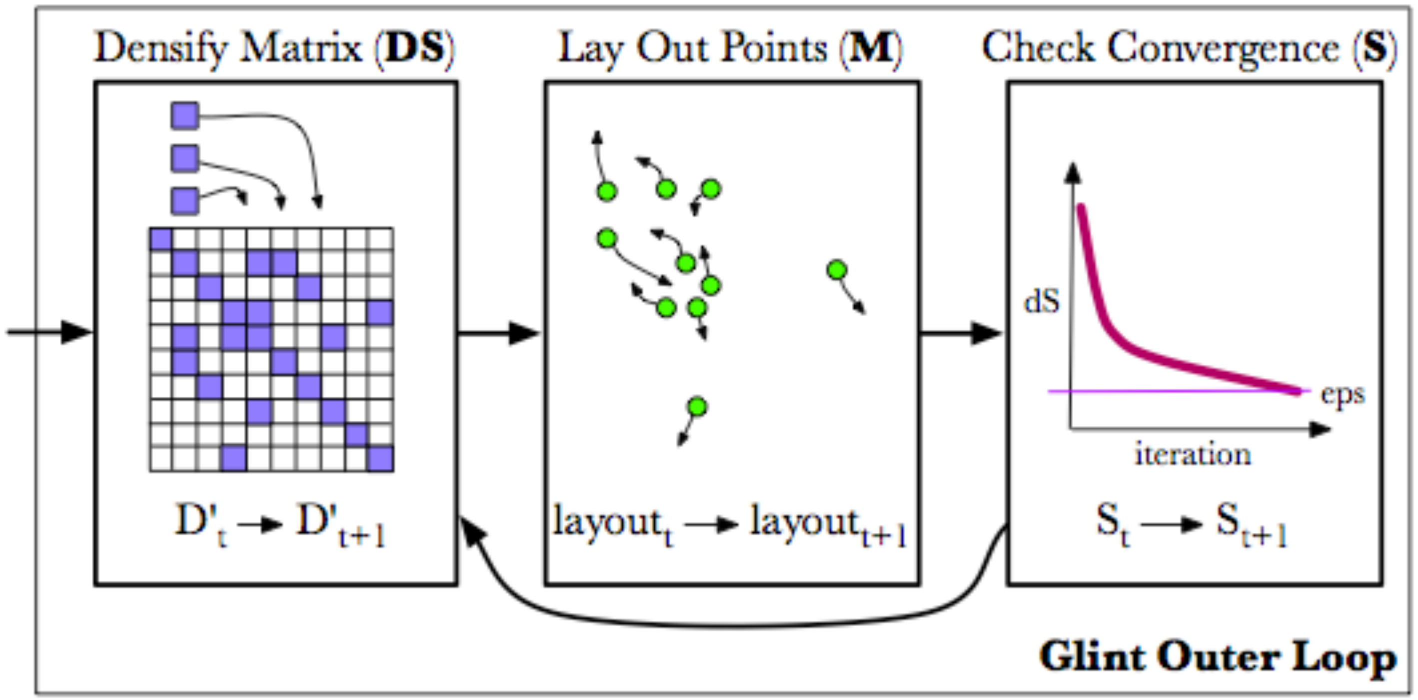






Methods and Outcomes

- 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.

MDS Algorithm Speeds

- newer algorithms linear, but...

Age

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

MDS Speed on Coordinate Data

`shuttle` benchmark
N = 43K
D = 9

Classic MDS | SMACOF

Hours
to Compute

Pivot MDS | Glimmer | LAMP

1 Second
to Compute

- time to calculate distance between two points
  - 0.00001 second

MDS Speed on Distance Matrix Data

`flickr` benchmark
N = 1925
d = EMD

Classic MDS	SMACOF	Pivot MDS	Glimmer	LAMP
Hours	Hours	>1 hour manual	Hours	Hours

- time to calculate distance between two points
  - 0.01 second

MDS Input: Coordinates vs Distances

High Dimensional Geometry → MDS → Low Dimensional Geometry

Coordinate Space → High Dimensional Geometry

Distance Matrix → High Dimensional Geometry

Coordinate Space → Low Dimensional Geometry

- some systems intrinsically require coordinates
  - fundamental to LAMP speedup approach
- some handle both
  - including Glimmer

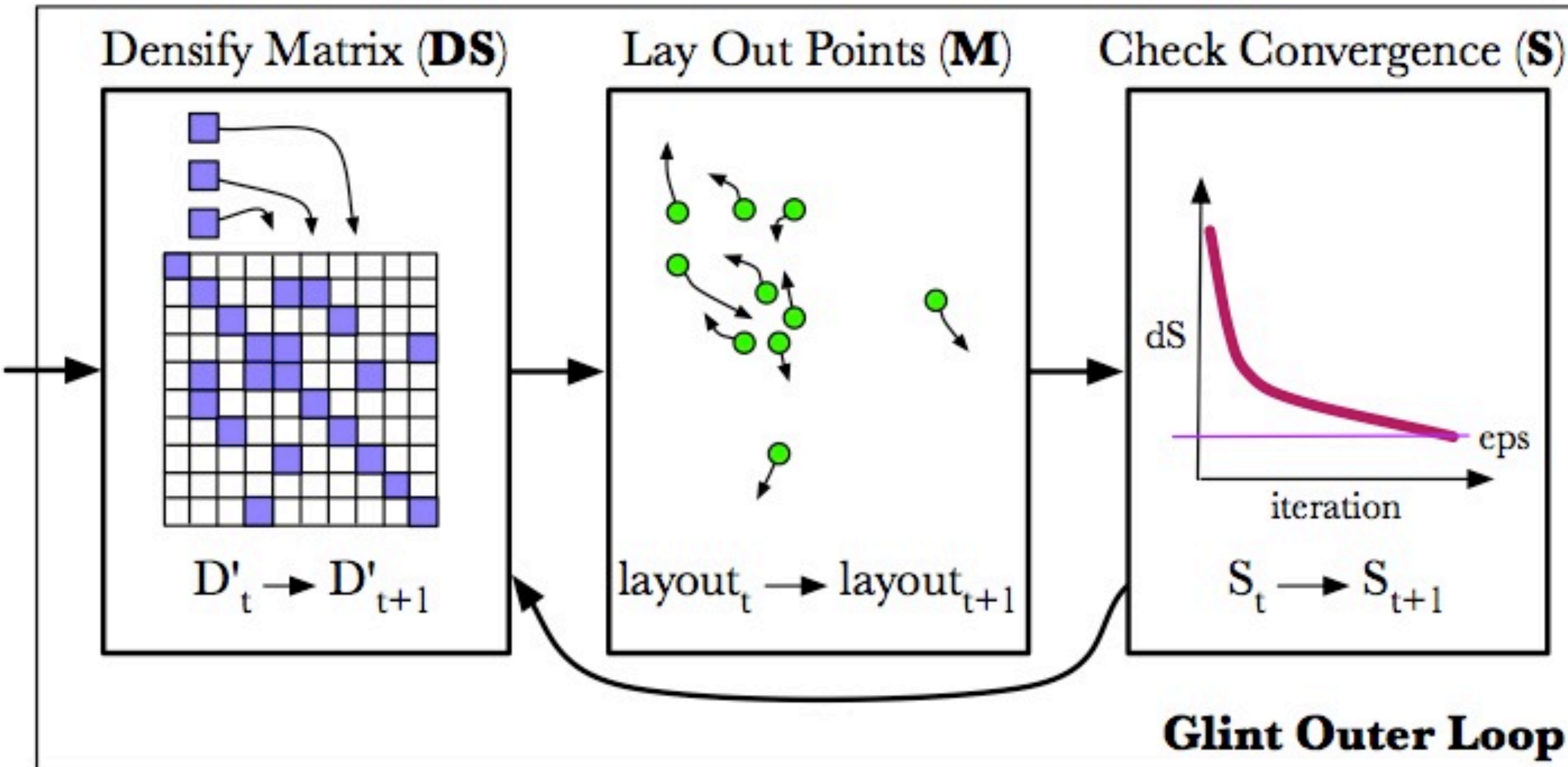
Costly Distances

- DR in the Wild revealed many real-world examples

	Distance function	Cost (seconds)
Cheap	Euclidean on 9-D data	0.00001
	Database Query	0.001
Costly	Earth Mover Distance	0.01
	Euclidean on 4M-D data	1.0
	Human-in-the-loop	10.0

Glint Framework

- calculate as few distances as possible, maintain quality
- three-stage architecture



Glint Instantiations

- framework accommodates broad spectrum of algorithm types
  - three instantiations provided

MDS Algorithm Type	Chosen Algorithm
Gradient-based Optimization	SMACOF
Spectral/Analytic	Pivot MDS
Force-Directed	Glimmer

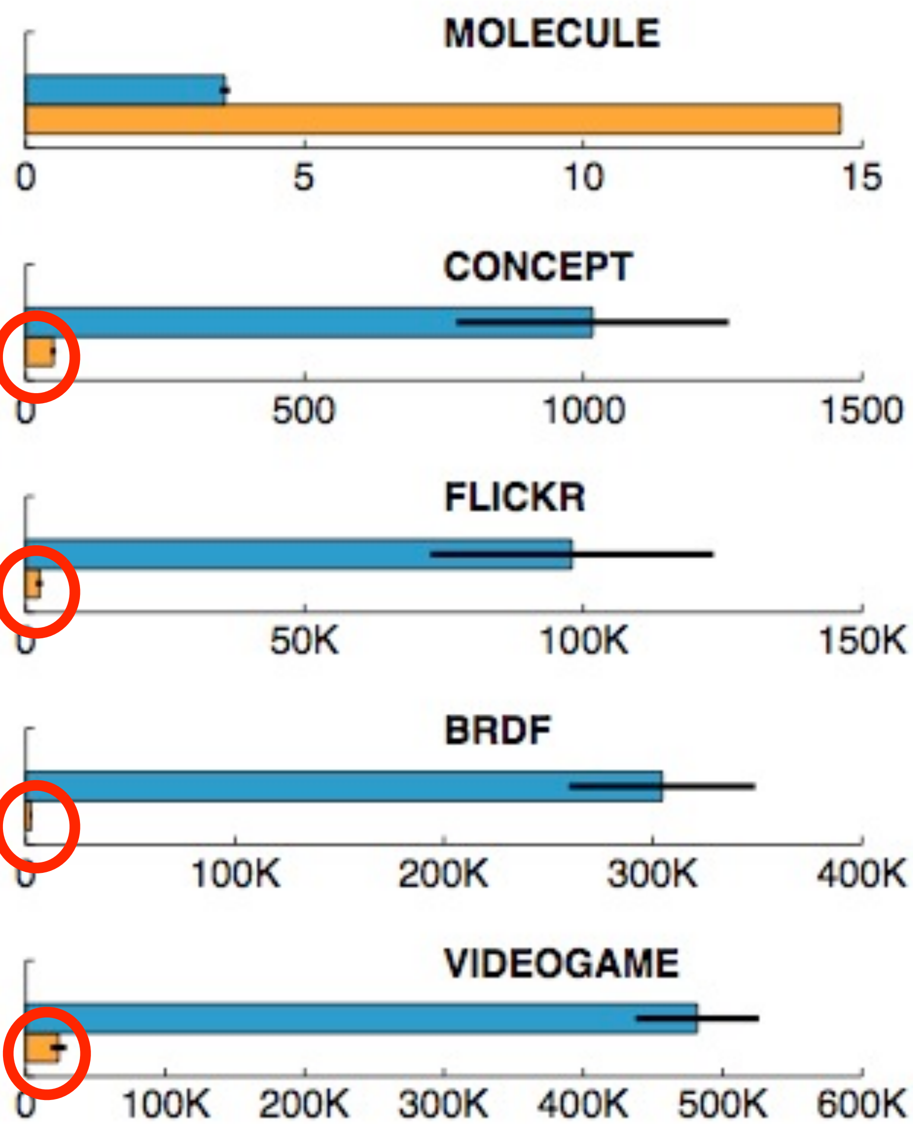
Force-Directed Instantiation Results

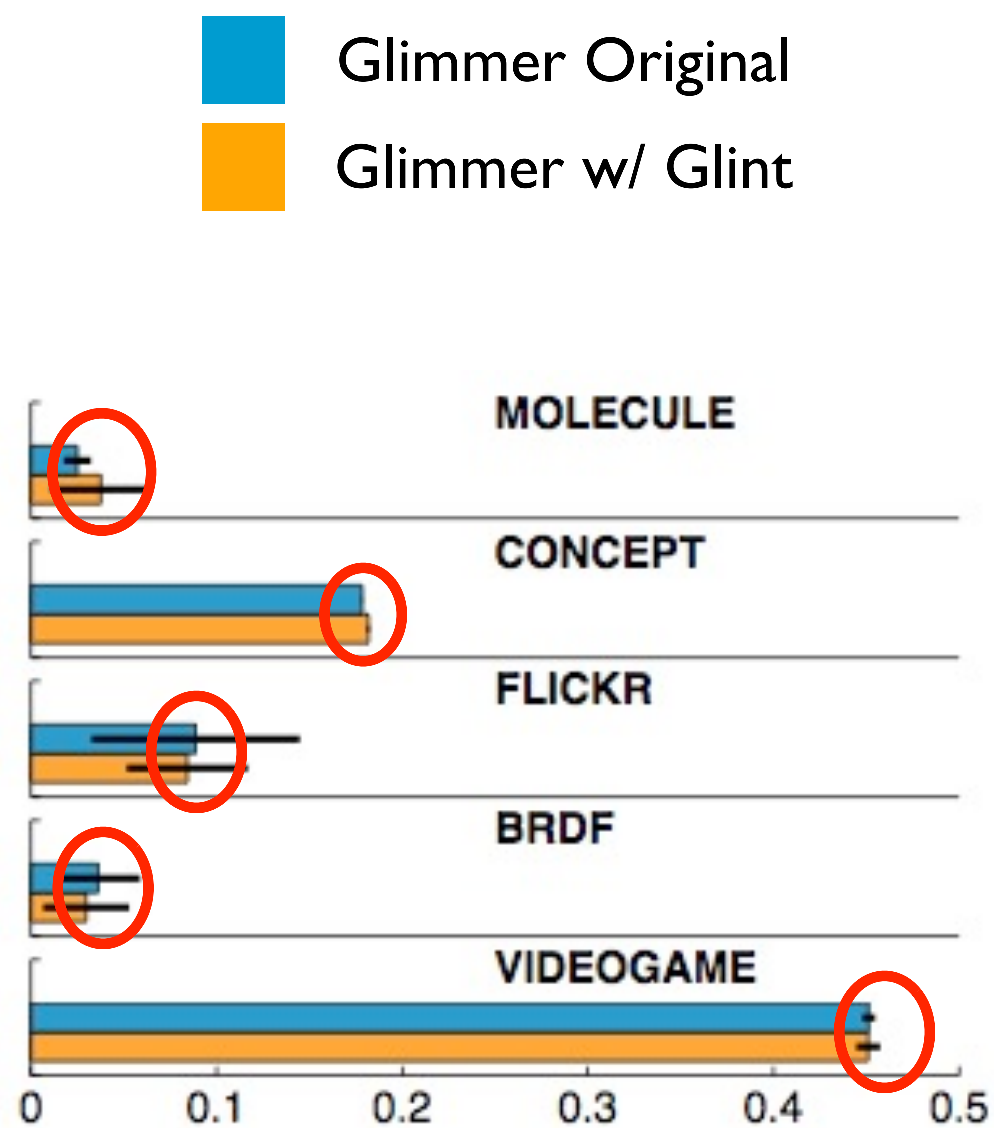
Speed

Quality

Cheap dist

Costly dists





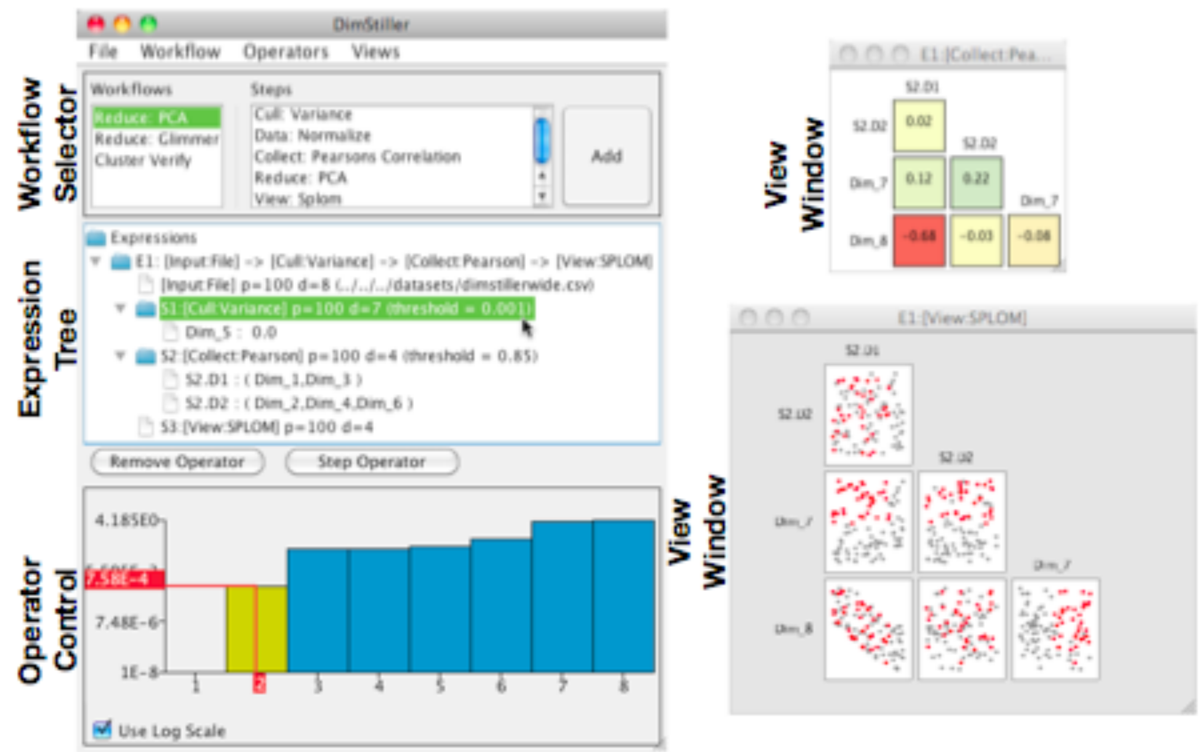
major speed improvements while quality maintained

Methods and Outcomes

- methods
  - algorithm benchmarks
- outcomes
  - dataset characterization different from previous work motivated by needs of real-world users
  - 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

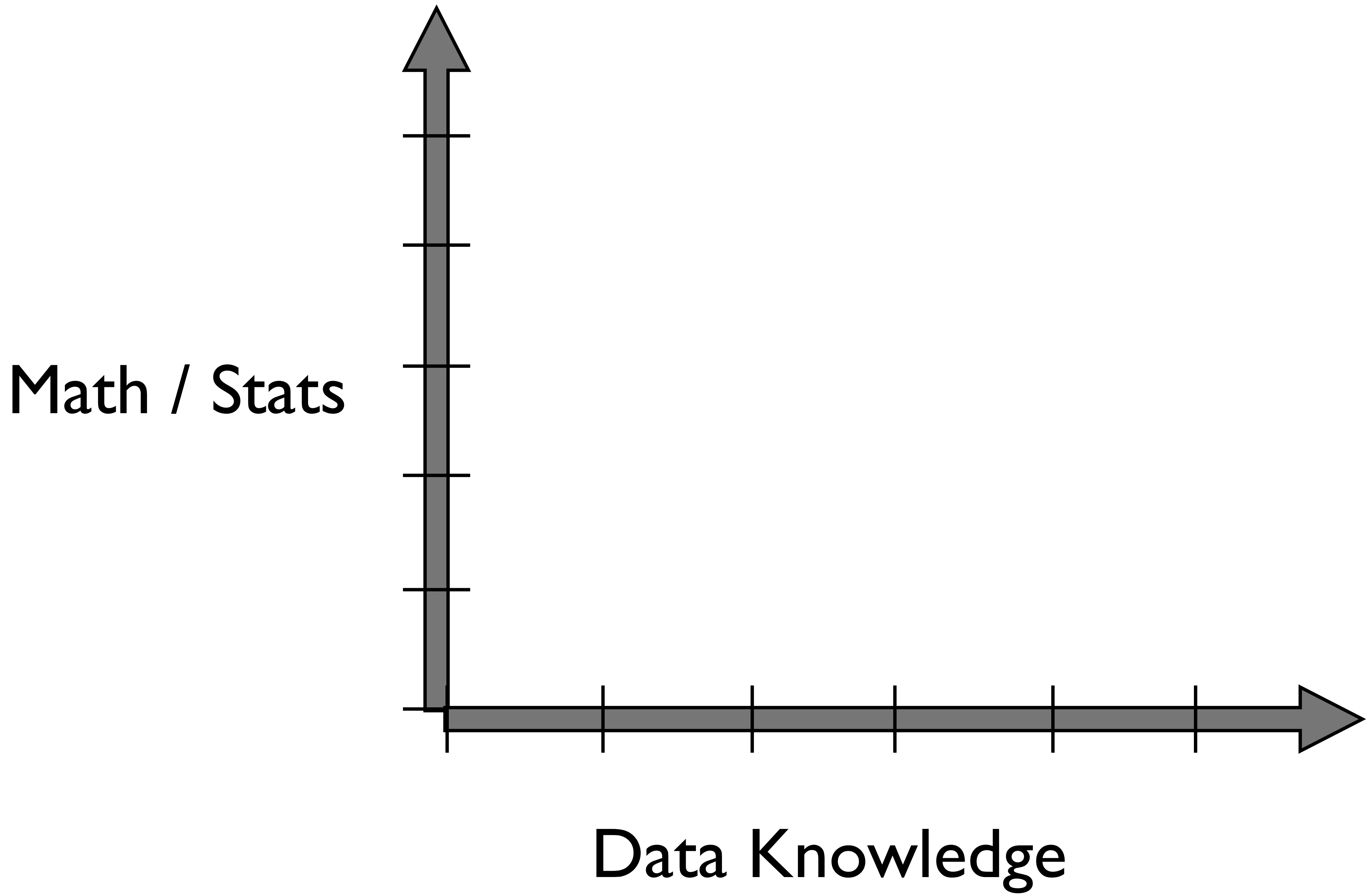
joint work with:
Stephen Ingram, Veronika Irvine, Melanie Tory, Steven Bergner, Torsten Möller

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

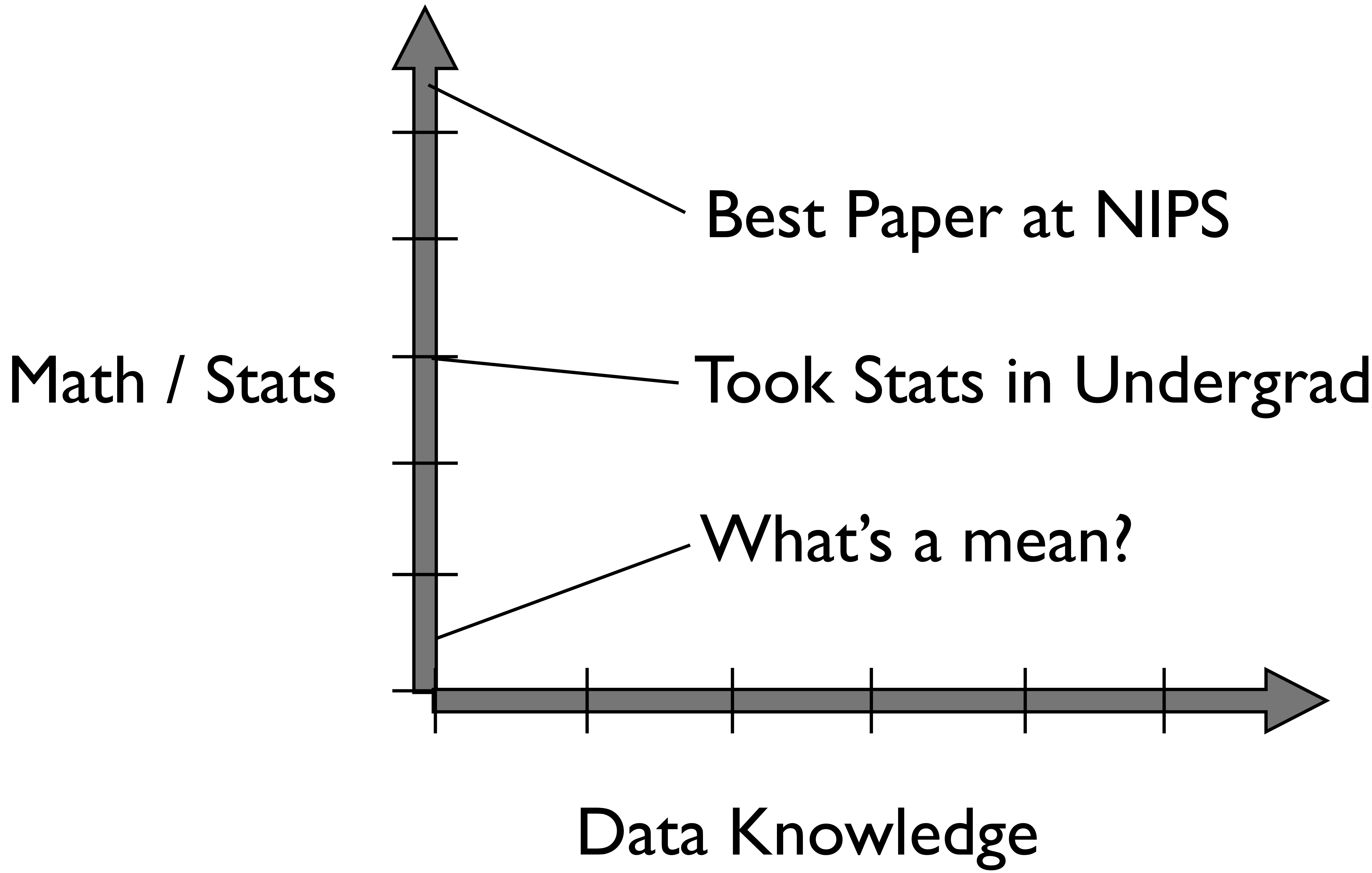
DimStiller: Workflows for dimensional analysis and reduction.
Ingram, Munzner, Irvine, Tory, Bergner, Moeller. Proc. VAST 2010, p 3-10.

Who Might Use DR?

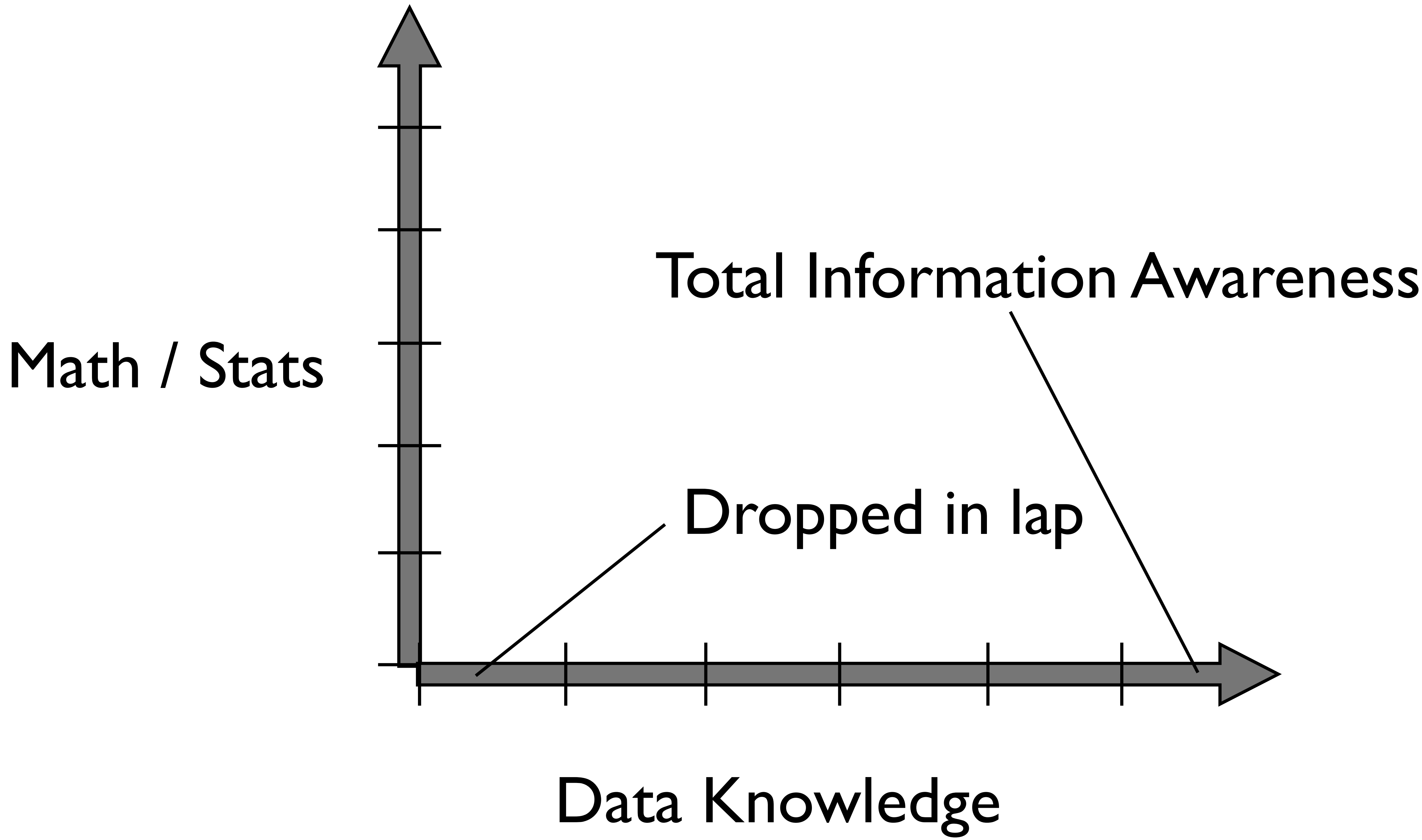
- DR in the Wild revealed broad set of users



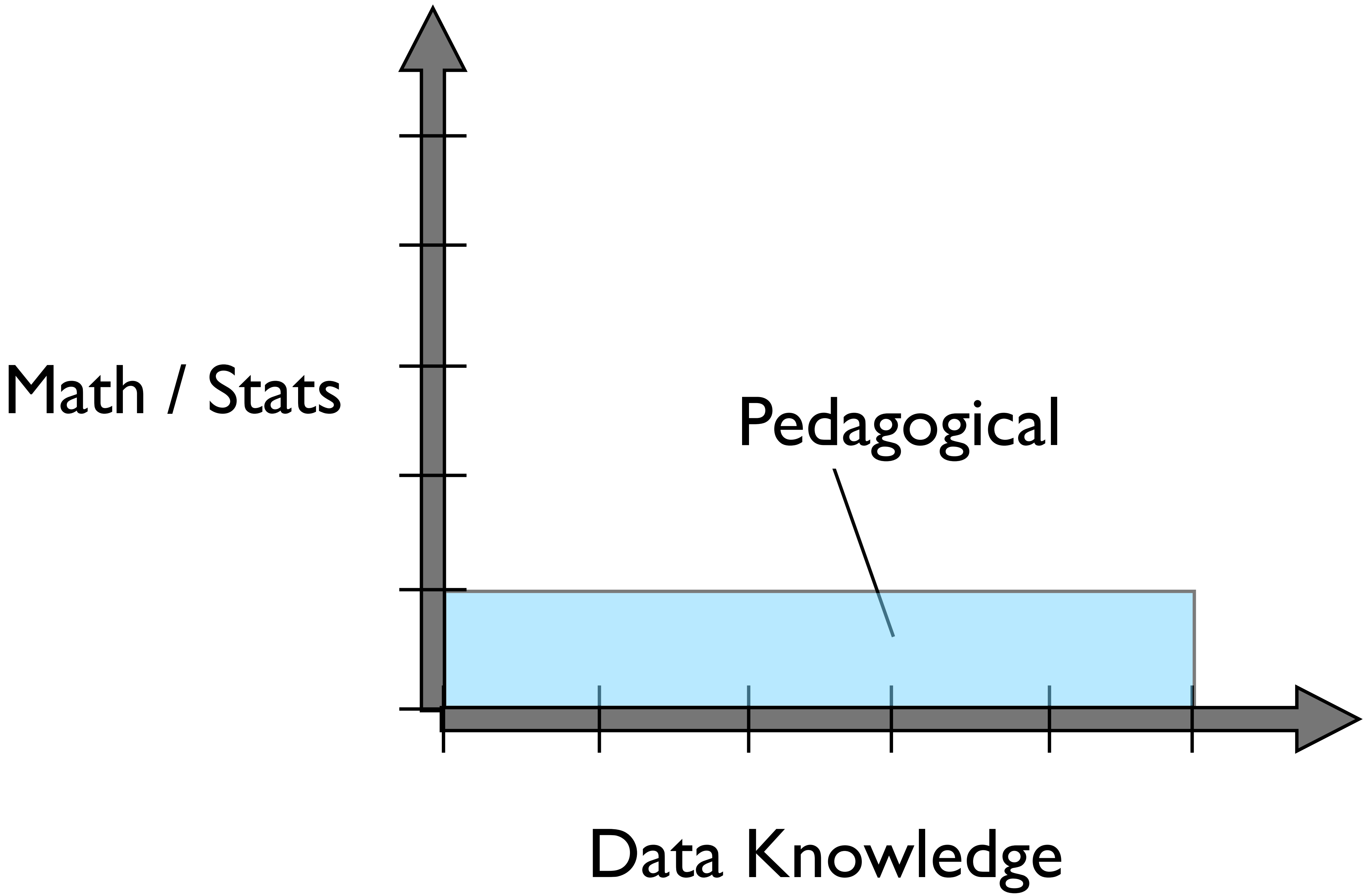
Who Might Use DR?



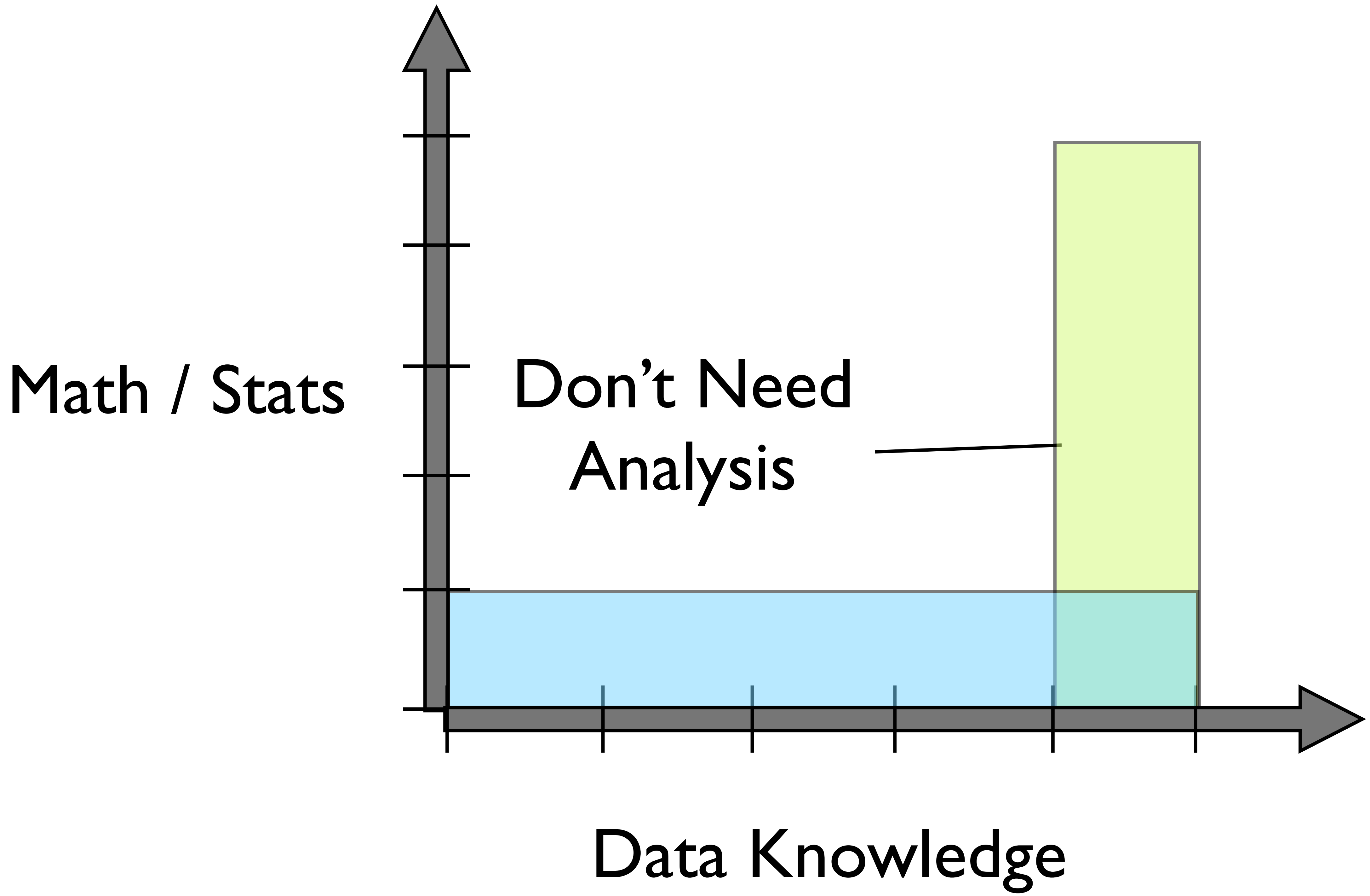
Who Might Use DR?



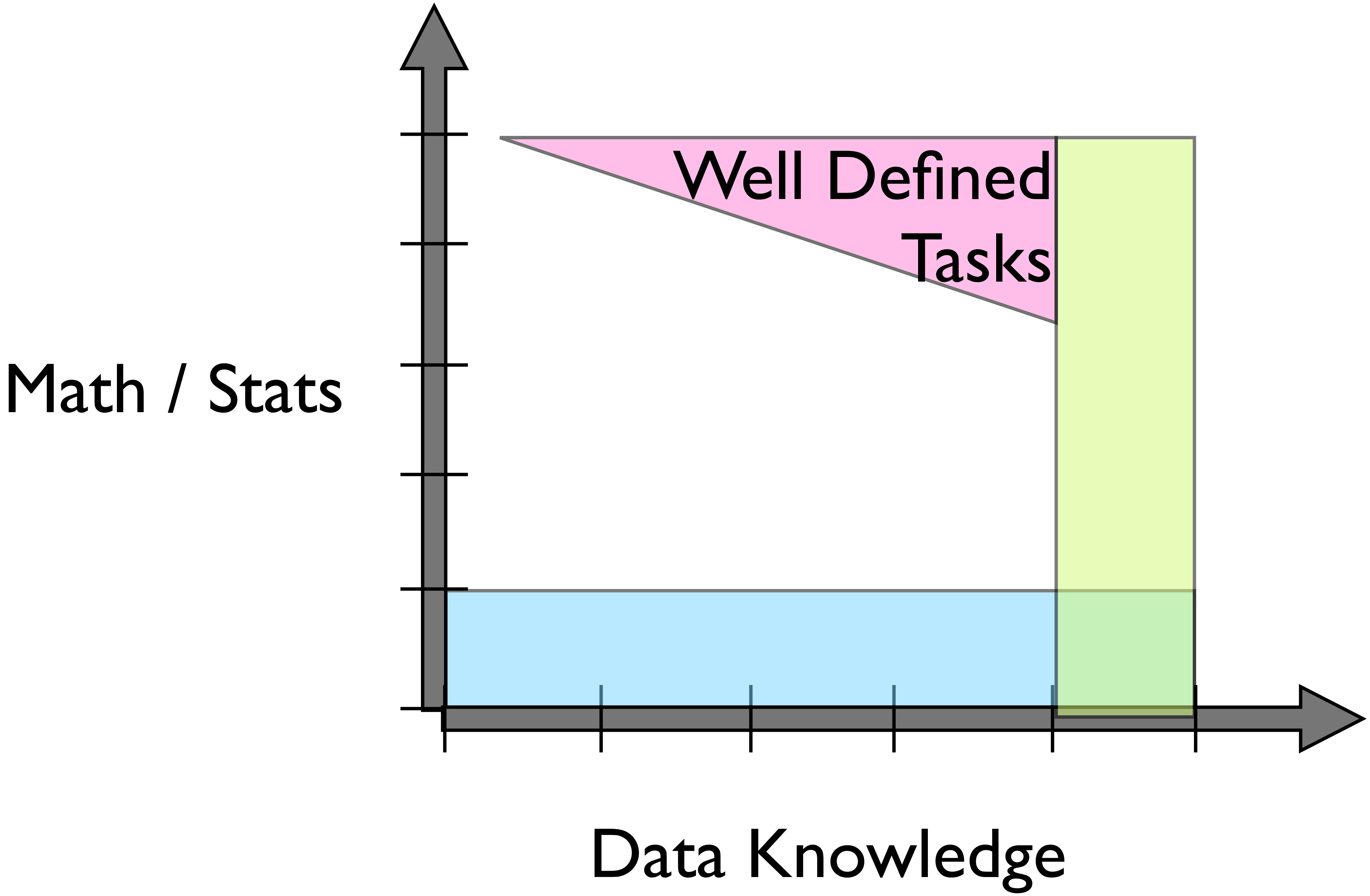
Who Might Use DR?



Who Might Use DR?

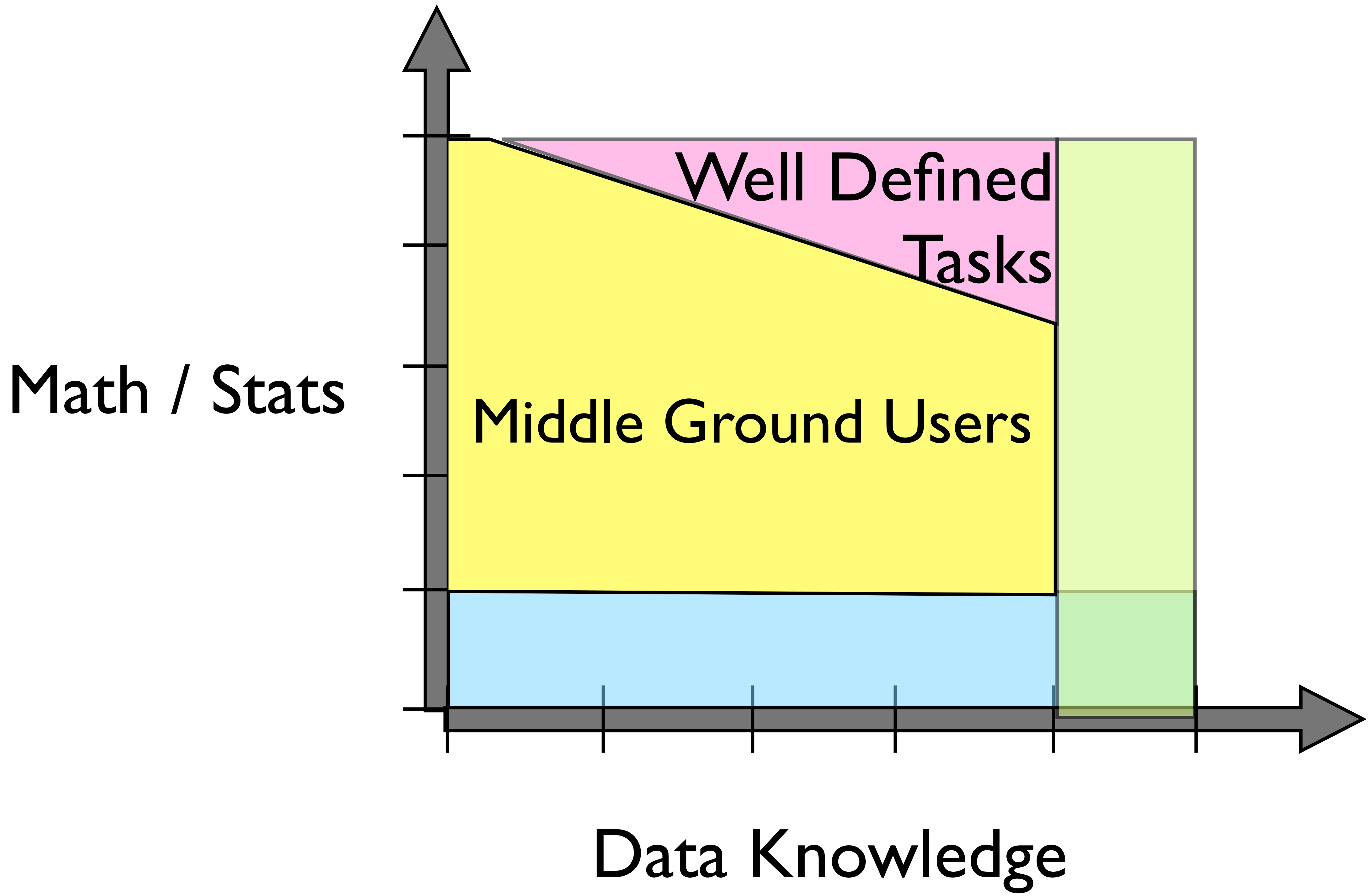


Who Might Use DR?

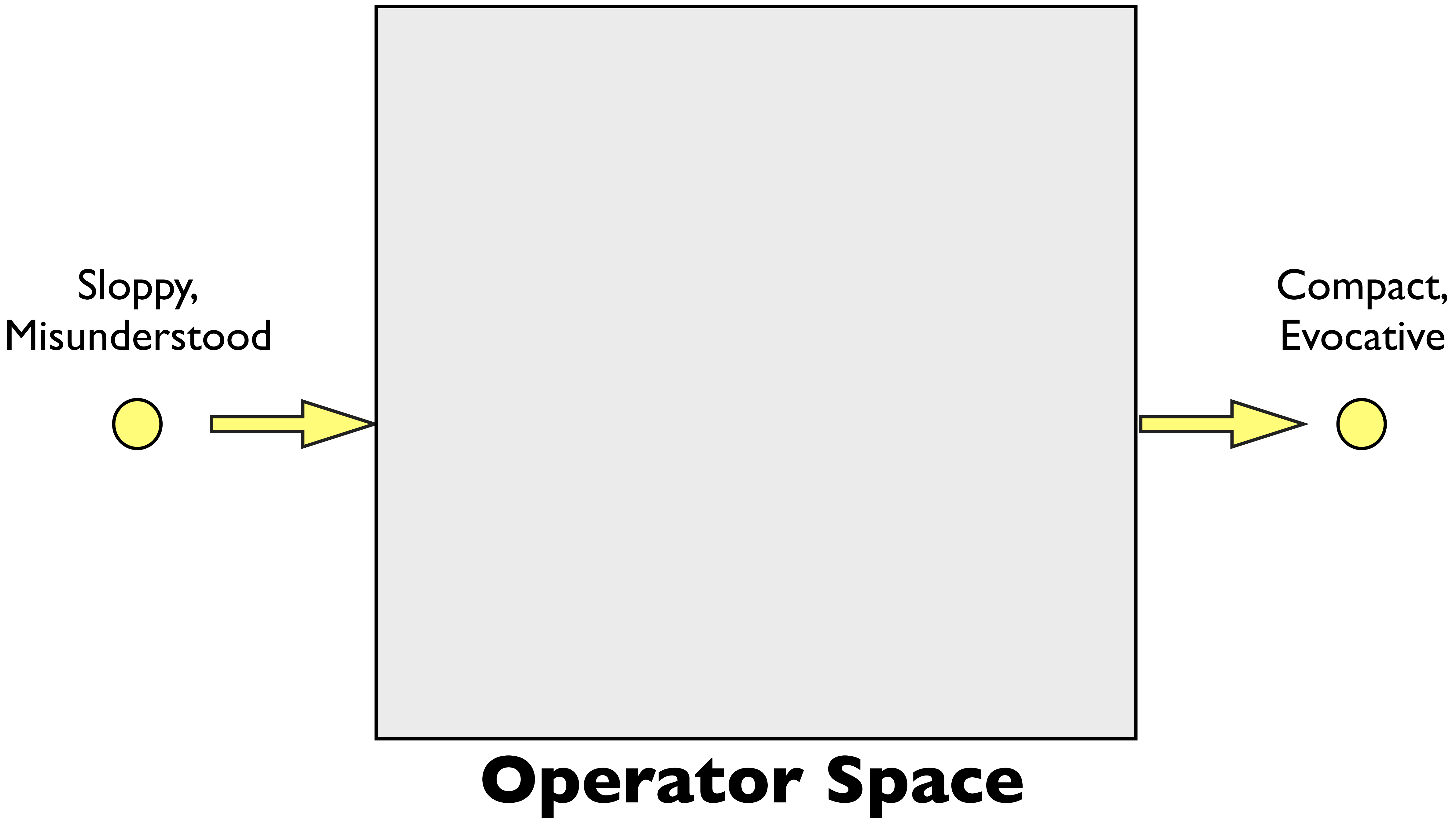


Who Might Use DR?

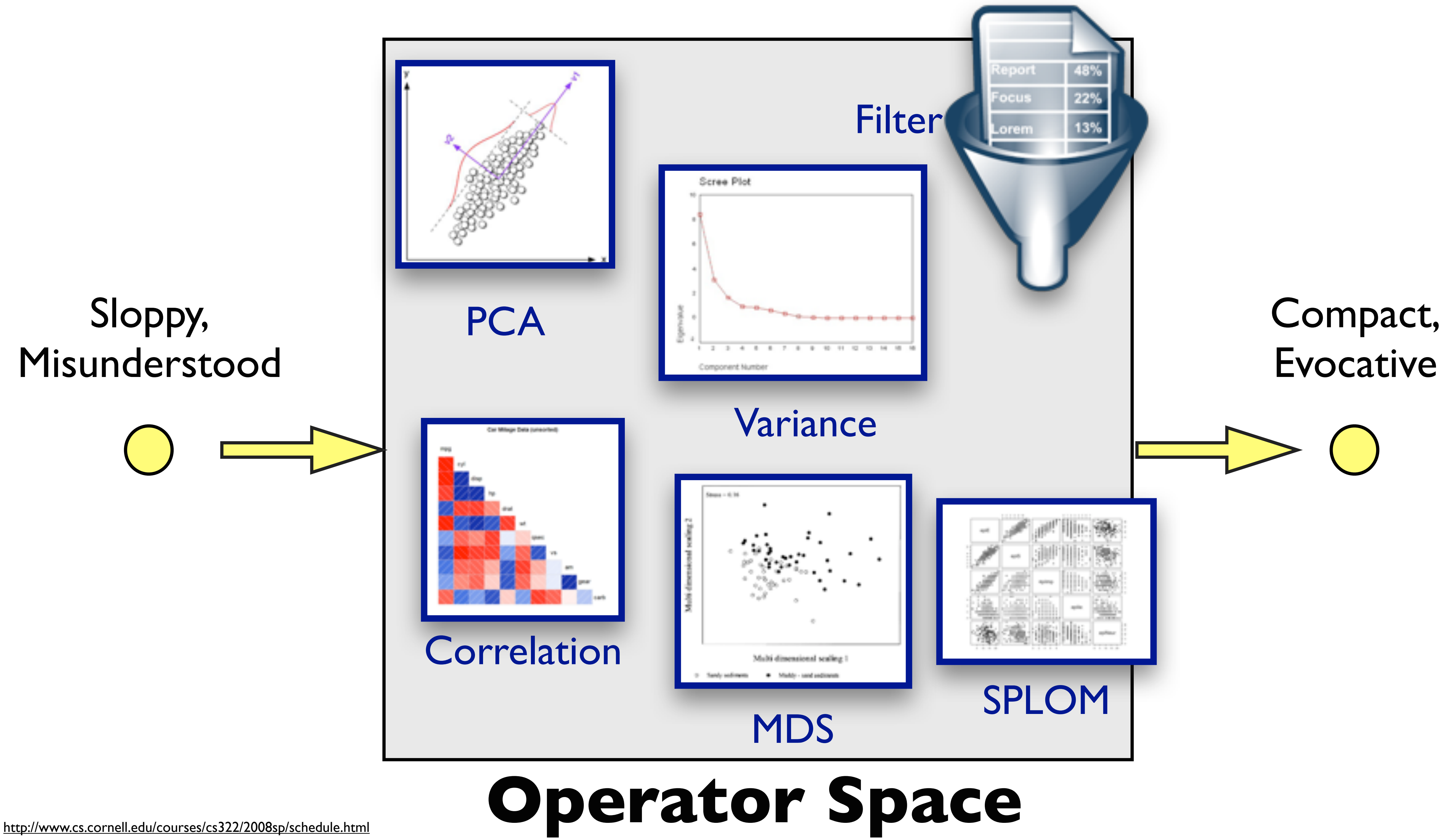
- middle ground users benefit from guidance



Global Guidance



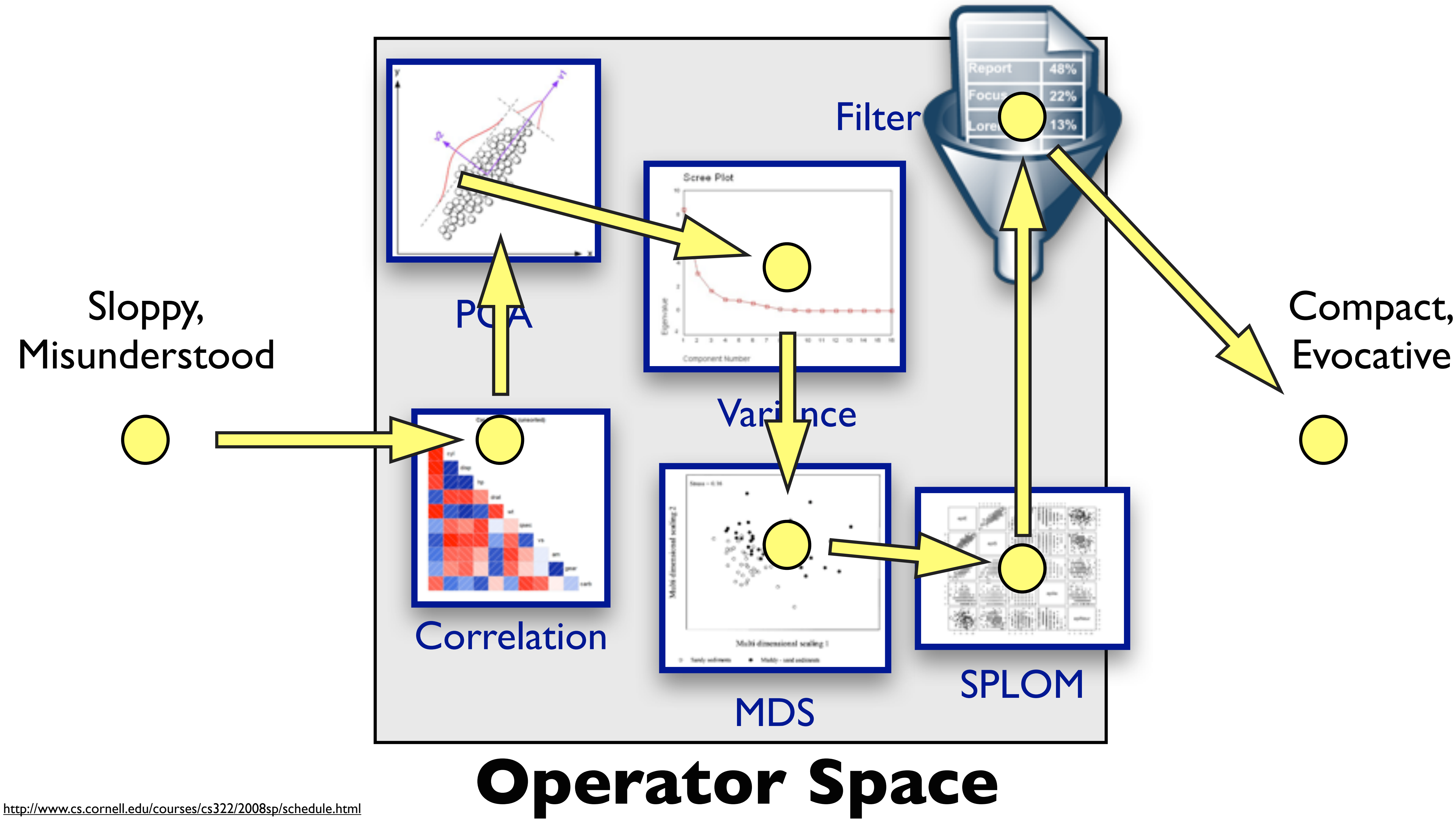
Global Guidance



http://www.cs.cornell.edu/courses/cs322/2008sp/schedule.html
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.scielo.cl/scielo.php?pid=S0716-078X2001000200019&script=sci_arttext
http://www.iconfinder.com/icondetails/44818/400/data_filter_icon?r=1
http://www.personality-project.org/R/

Global Guidance

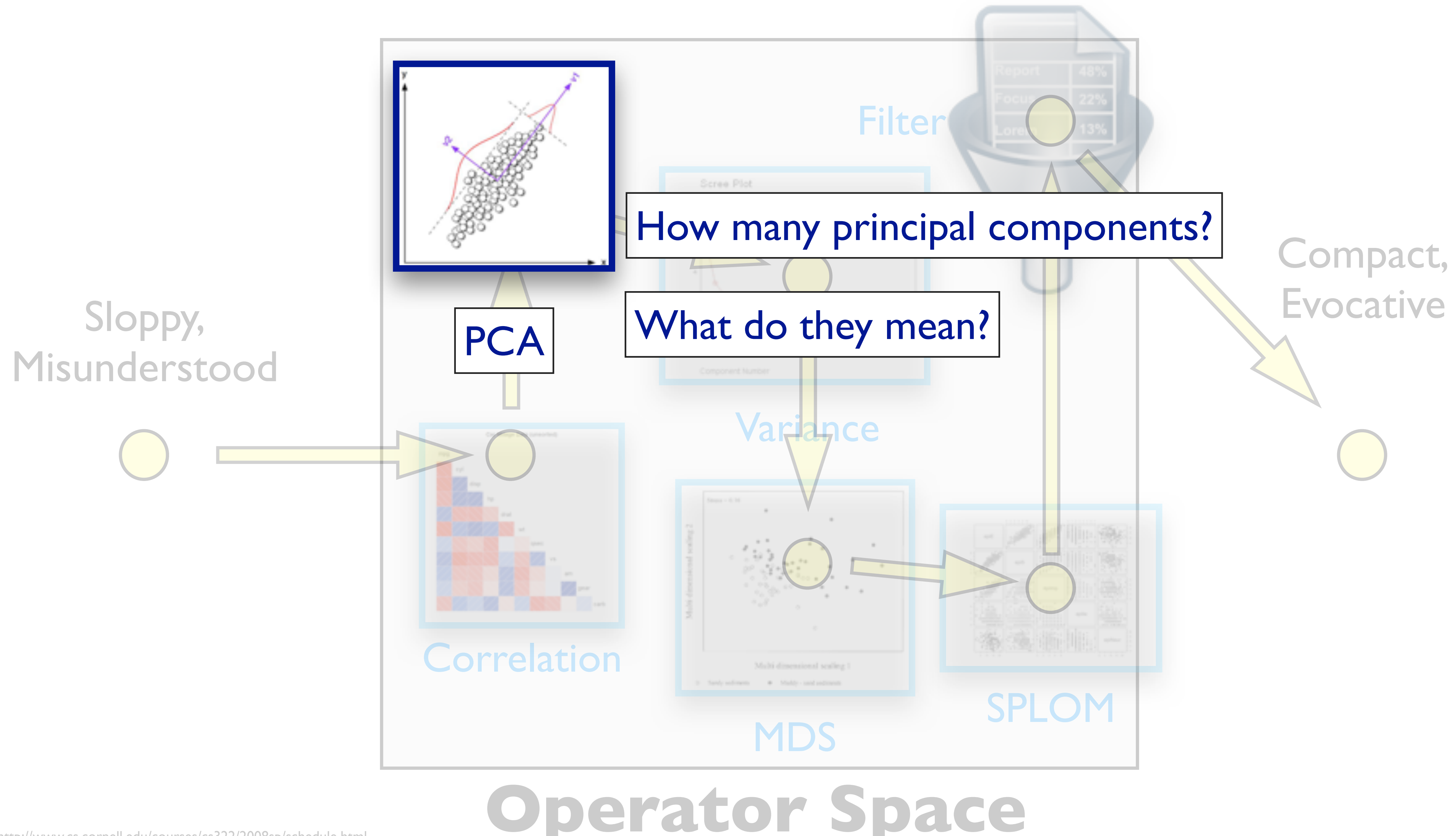
- which operations and in which order?



http://www.cs.cornell.edu/courses/cs322/2008sp/schedule.html
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.scielo.cl/scielo.php?pid=S0716-078X2001000200019&script=sci_arttext
http://www.iconfinder.com/icondetails/44818/400/data_filter_icon?r=1
http://www.personality-project.org/R/

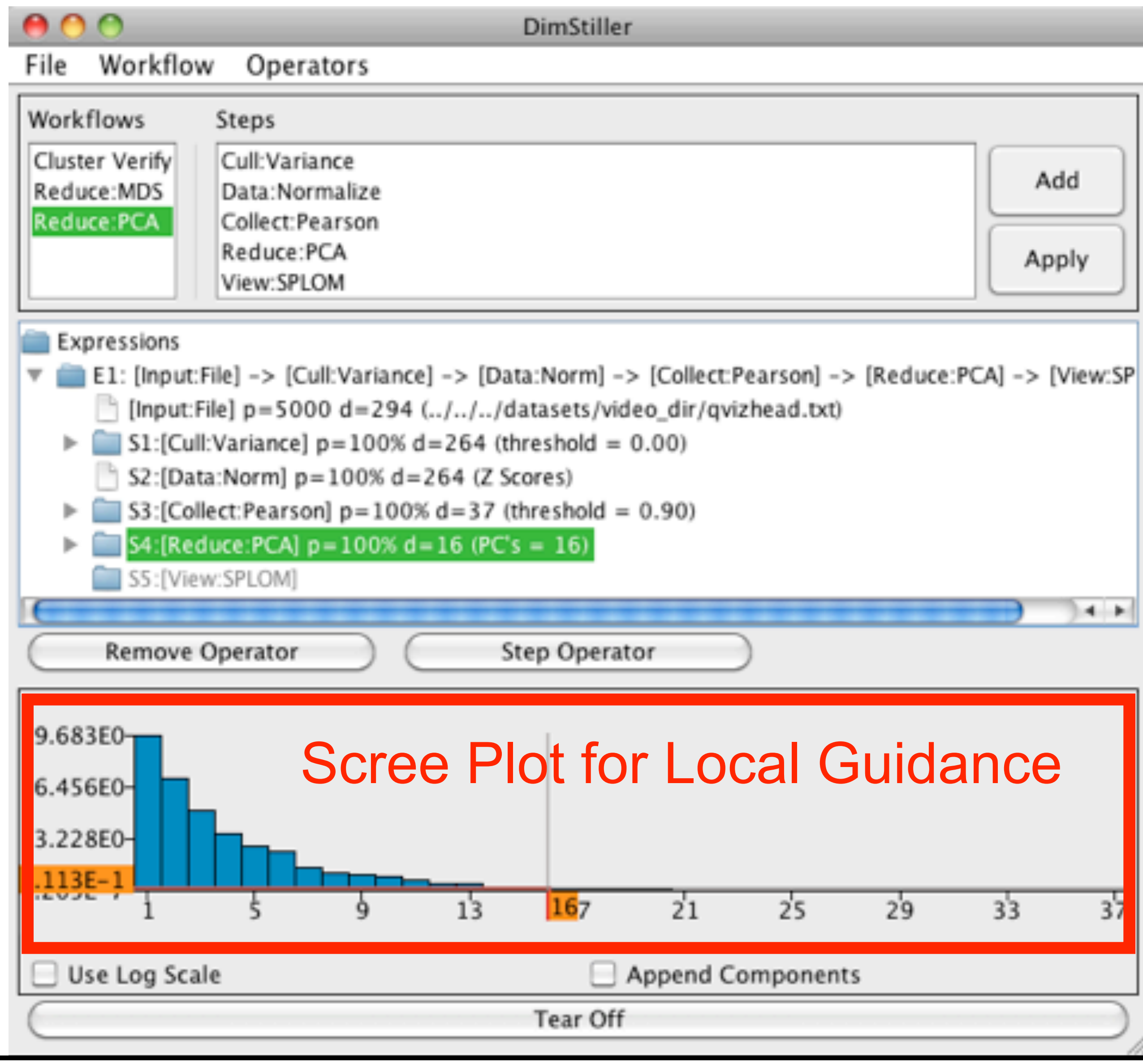
Local Guidance

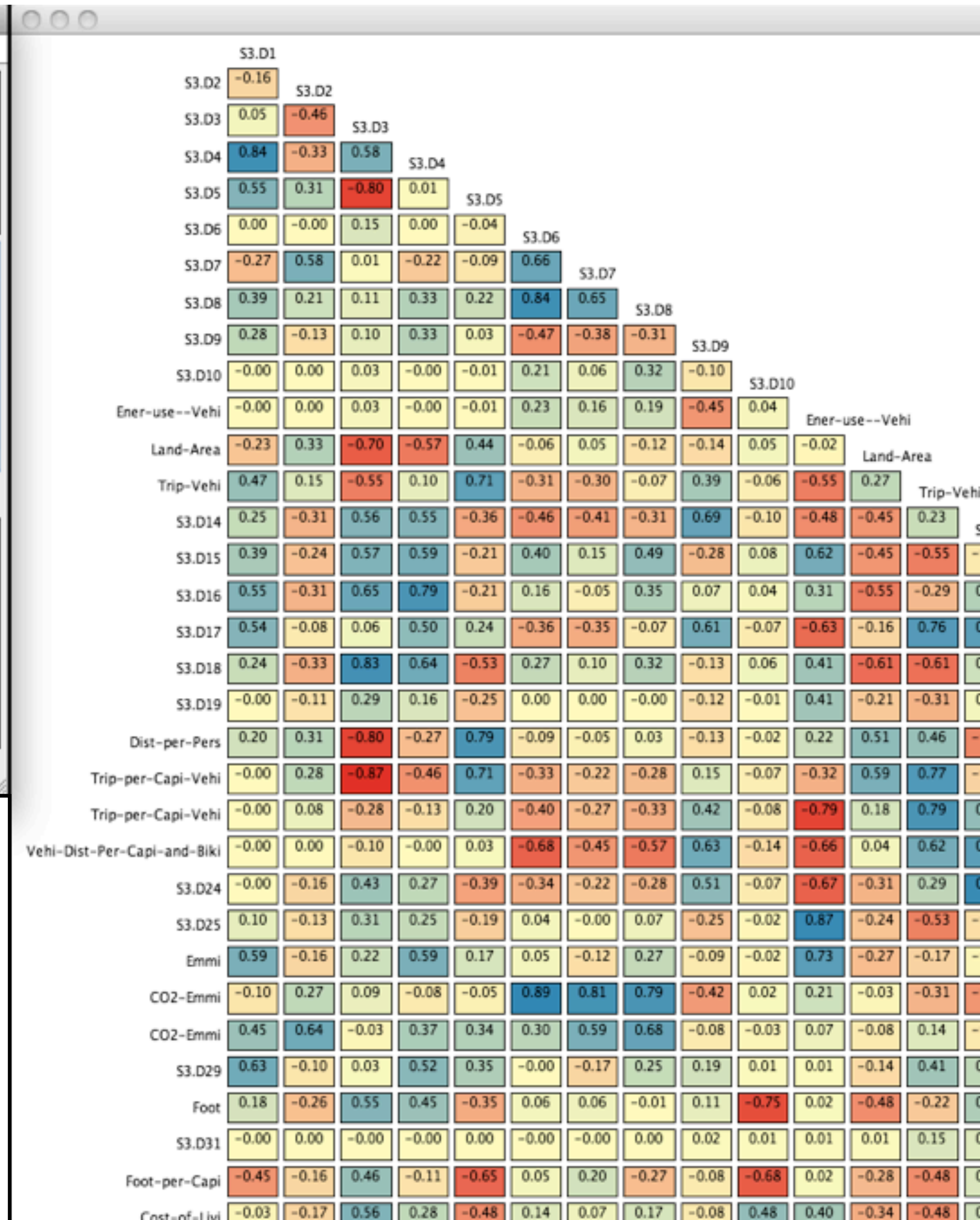
- what to do with a given operator?



http://www.cs.cornell.edu/courses/cs322/2008sp/schedule.html
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.scielo.cl/scielo.php?pid=S0716-078X2001000200019&script=sci_arttext

DimStiller





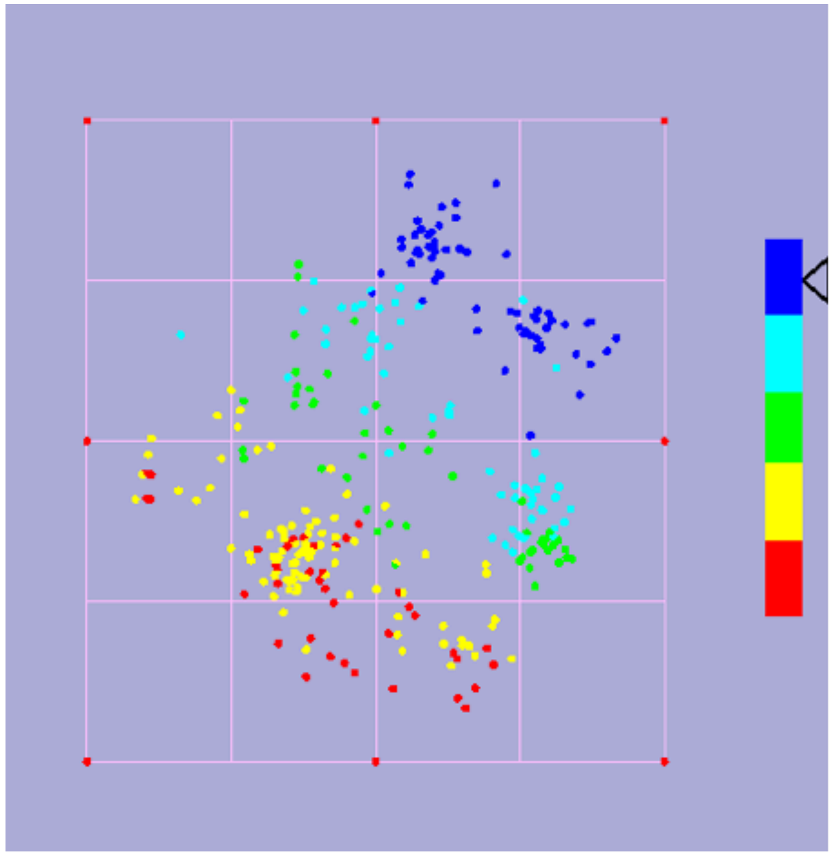
- pre-built workflows
- sequence of operators
- local guidance for each operator
  - example: estimate true dimensionality with scree plot

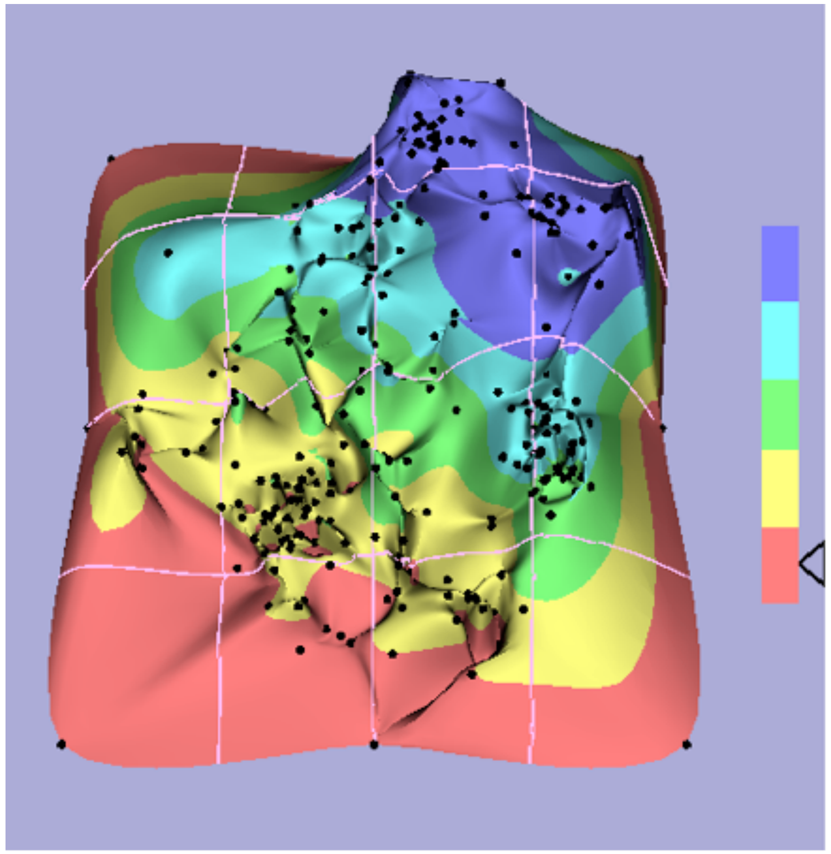
Methods and Outcomes

- 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!

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 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:

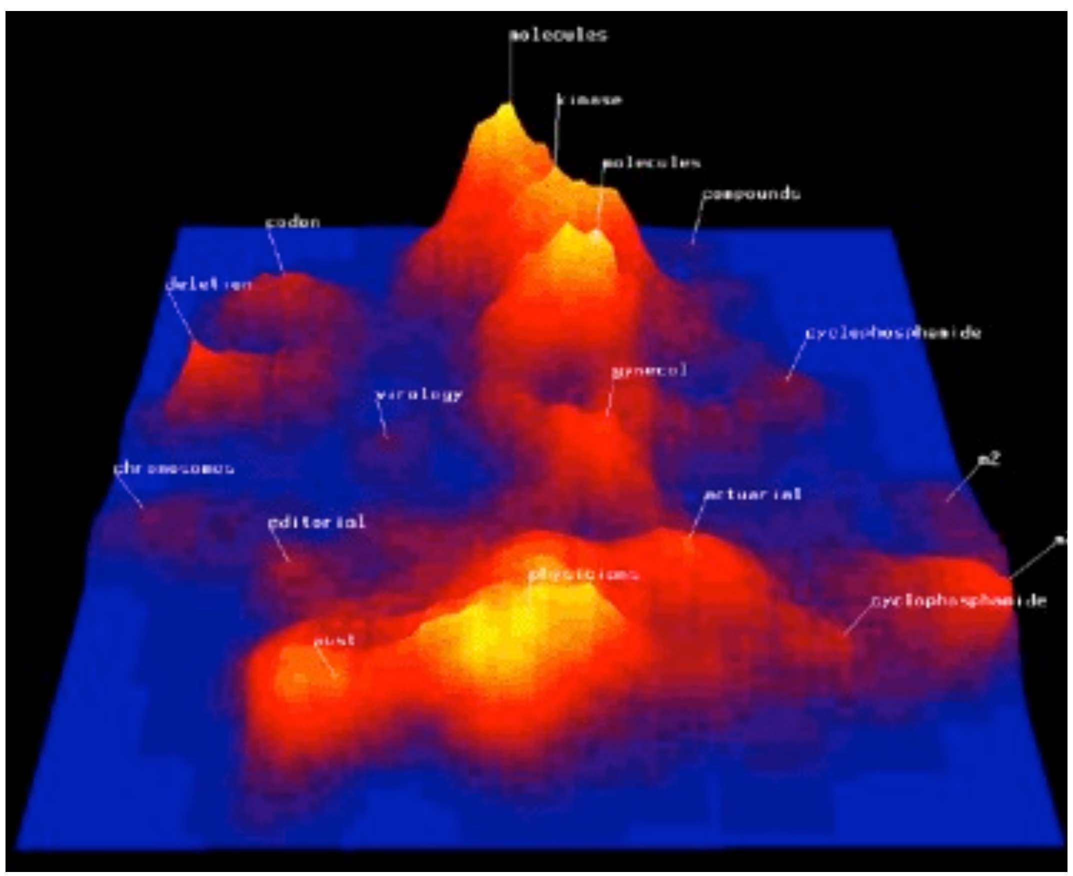
Melanie Tory, David W. Sprague, Fuqu Wu, Wing Yan So

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).

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



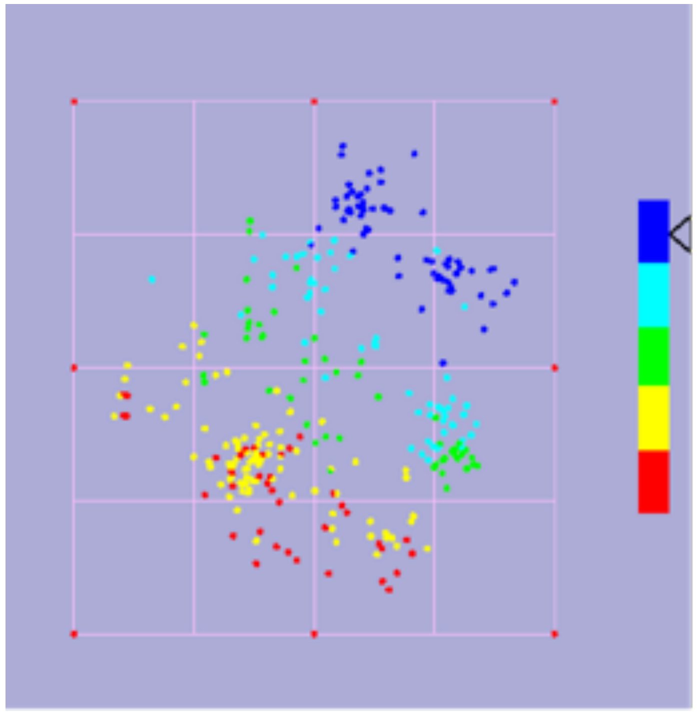
Themescape:
[http://www.k-n-o-r-z.de/publ/example/retriev1.htm]

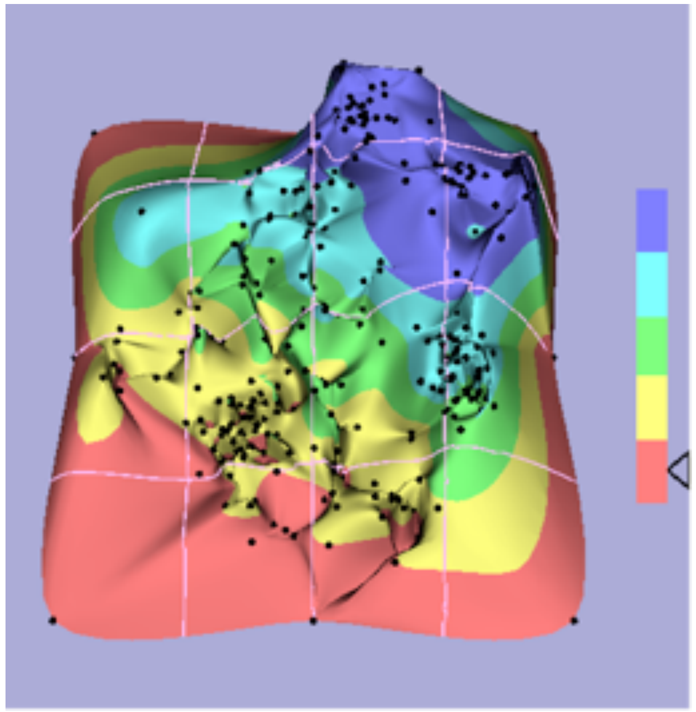
[Guide to MicroPatent Aureka 9 ThemeScape]

Understanding User Task

- abstract: search involving spatial areas and estimation

  Estimate which grid cell has the most points of the target color



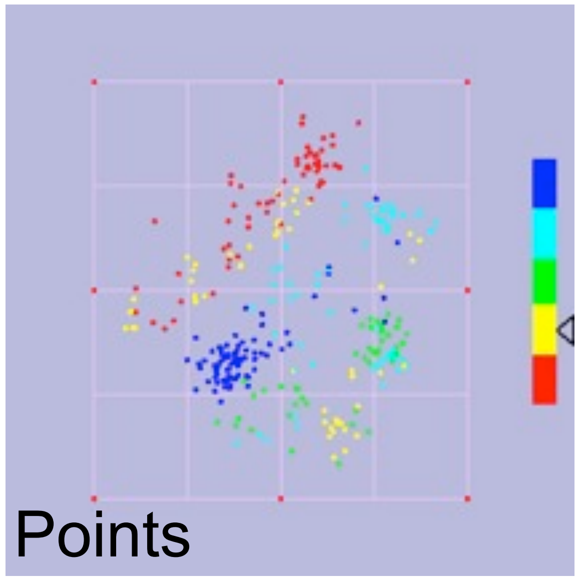


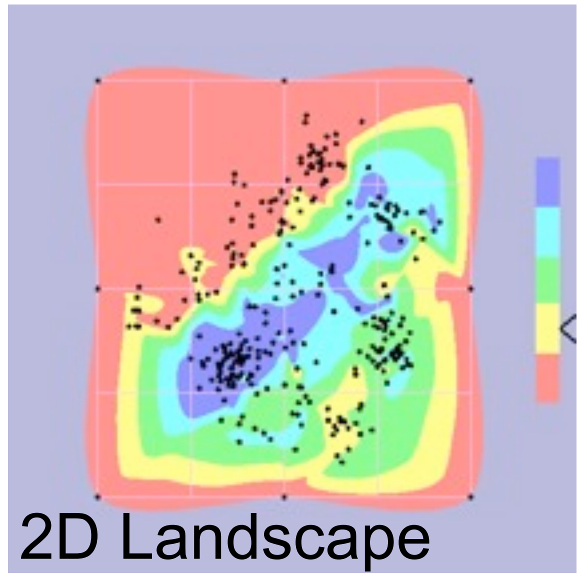
- domain-specific examples

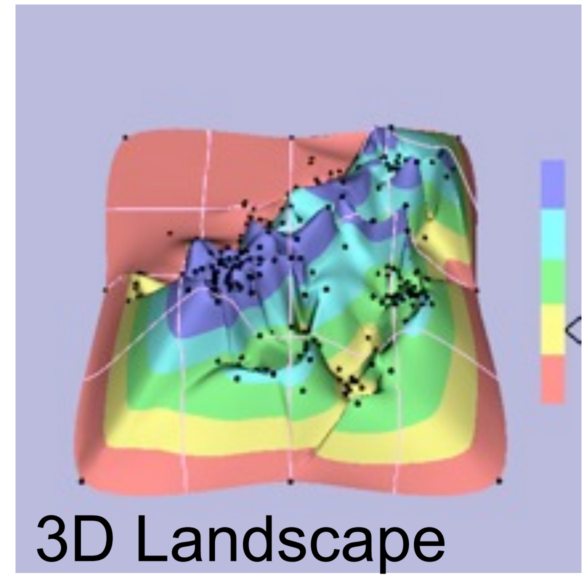
  "Where in the display are people with high incomes?"
  "Does this area also have high education levels?"
  "Does this area correspond to a particular work sector?"

- non-trivial complexity yet fast response time
- frequent subtask in pilot test of real data analysis

Lab Study: Test Human Response Time and Error



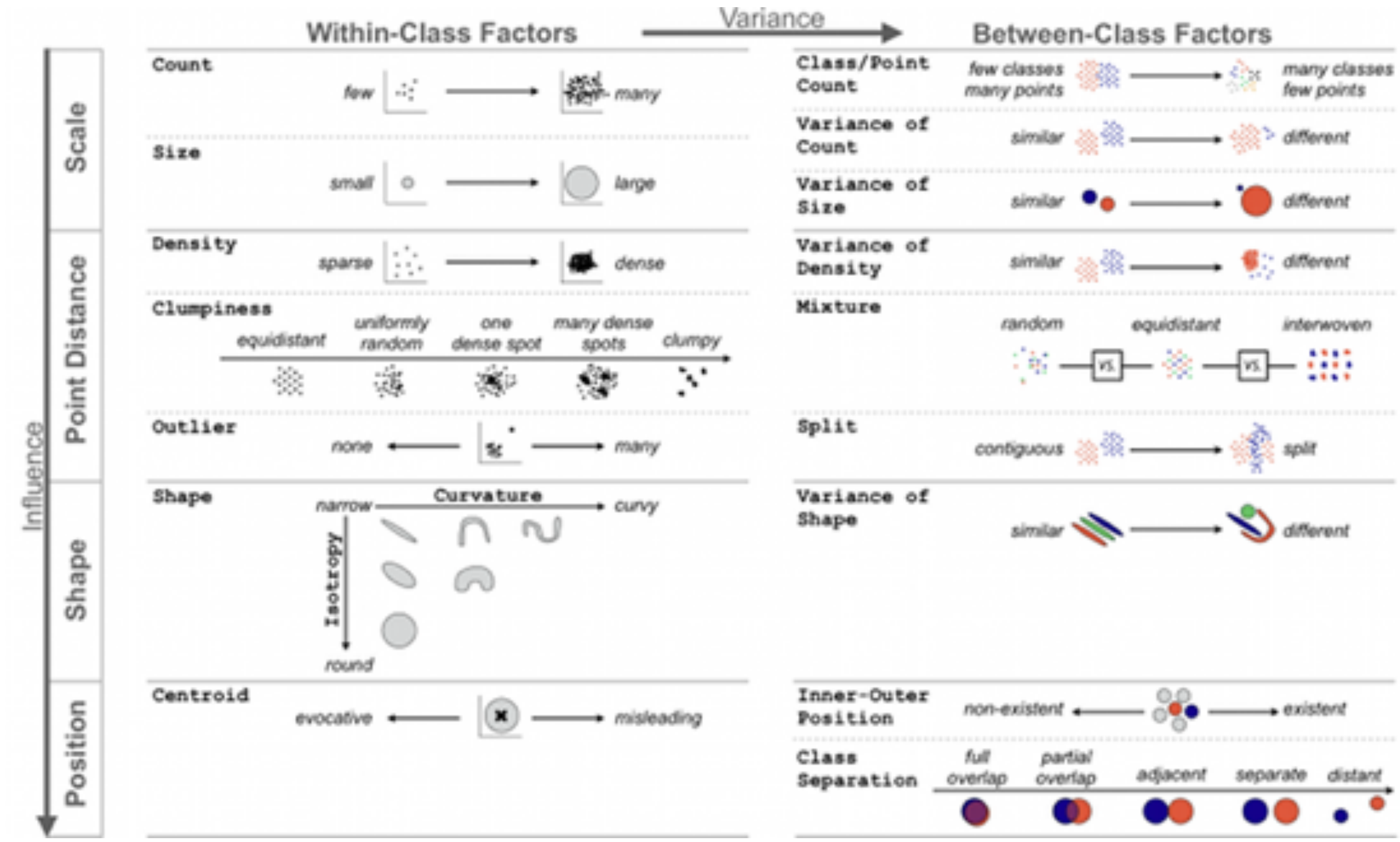




- hypotheses
  - points are better than landscapes
    - result: yes!
    - much better: 2-4 × faster, 5-14 × more accurate
  - 2D landscapes (color only) better than 3D landscapes (color + height redundantly encoded)
    - result: yes
    - significantly faster, no significant difference in accuracy

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:

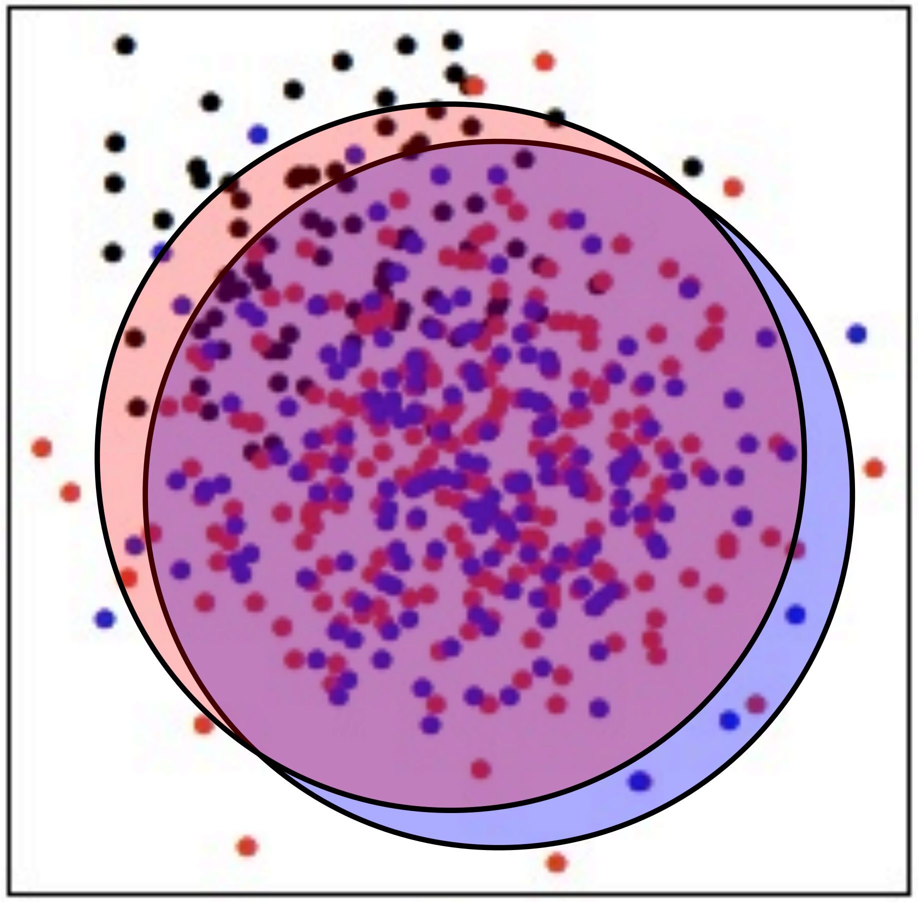
Michael Sedlmair, Andrada Tatu, Melanie Tory

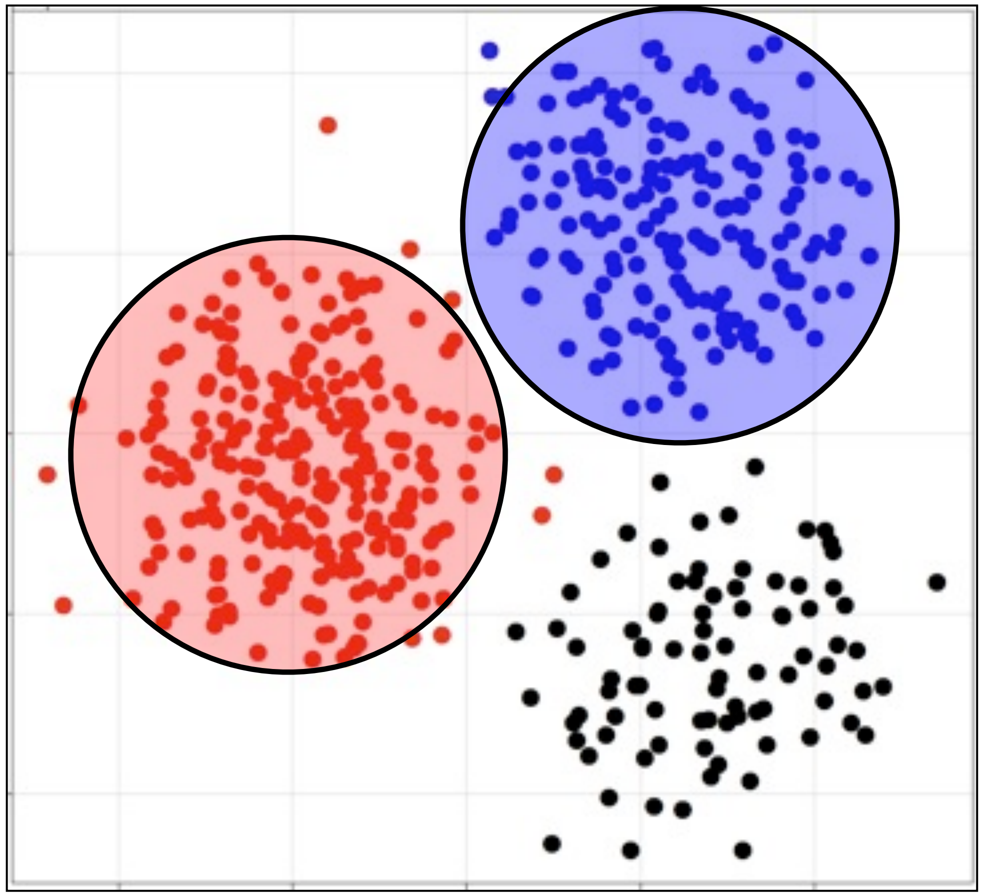
http://www.cs.ubc.ca/labs/imager/tr/2012/VisClusterSep/

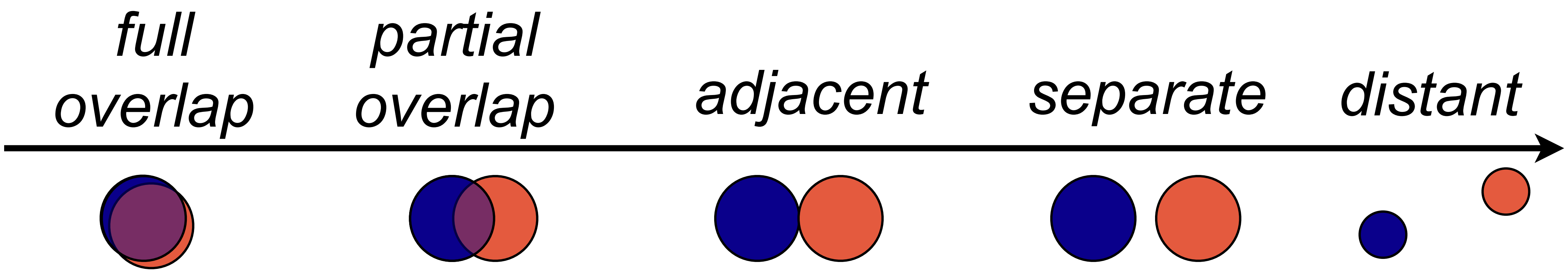
A Taxonomy of Visual Cluster Separation Factors.
Sedlmair, Tatu, Munzner, Tory. Computer Graphics Forum 31(3):1335-1344, 2012 (Proc. EuroVis 2012).

Cluster Separation

- simple idea





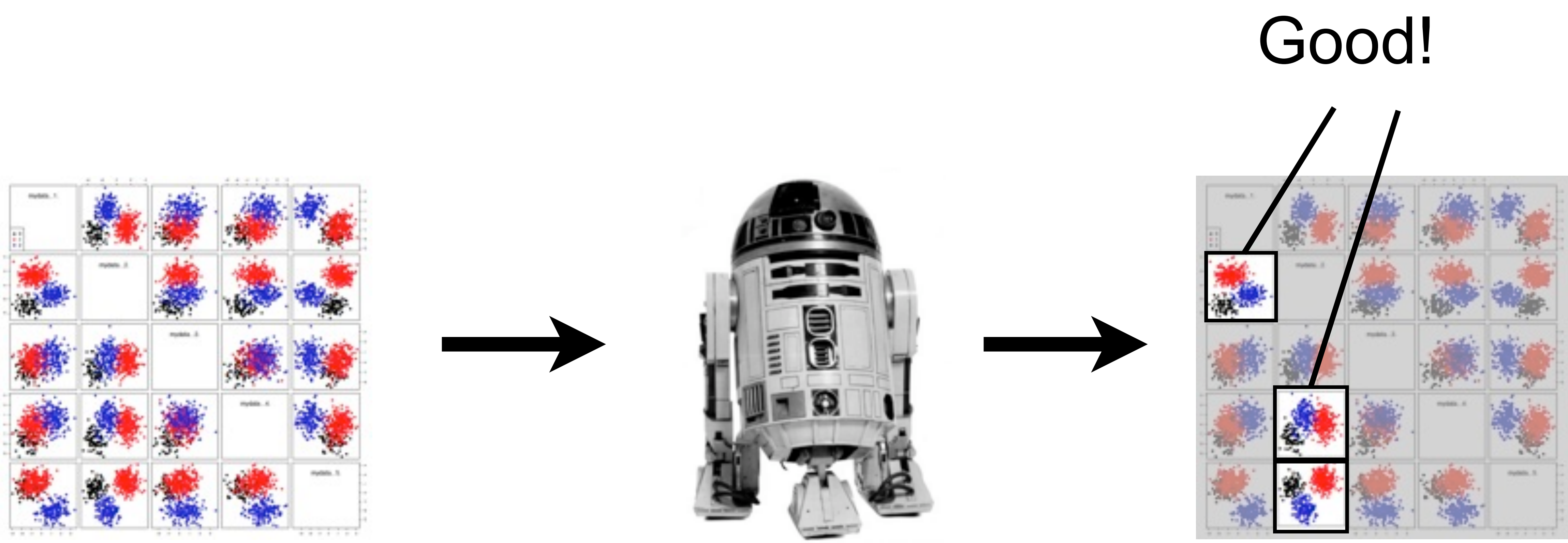


Visual Cluster Separation Measures

- 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]

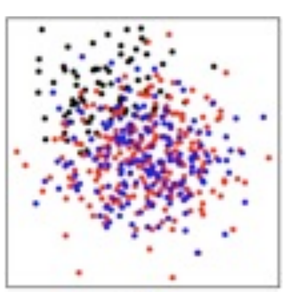


Visual Cluster Separation Measures

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

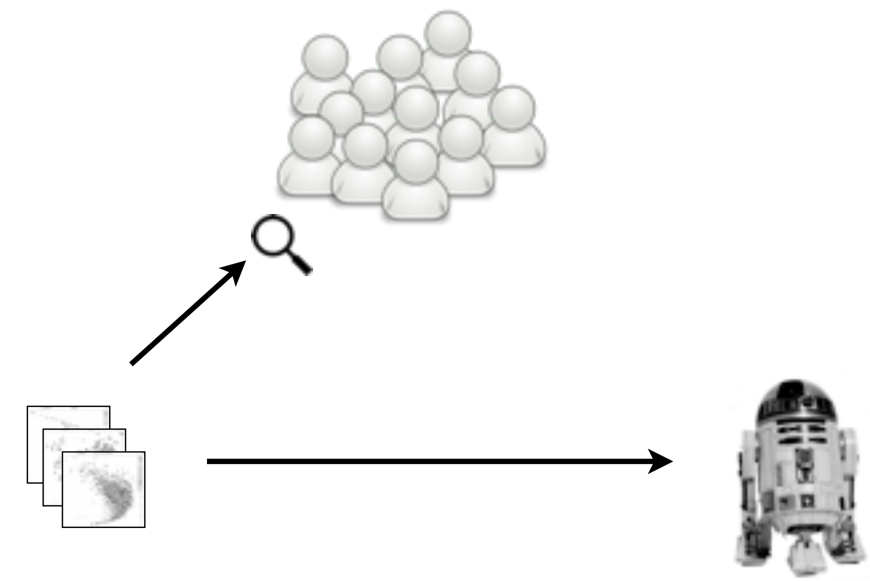
  Tatu et al.: Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data [AVI 2010]

- but our attempt to use for guidance showed problems

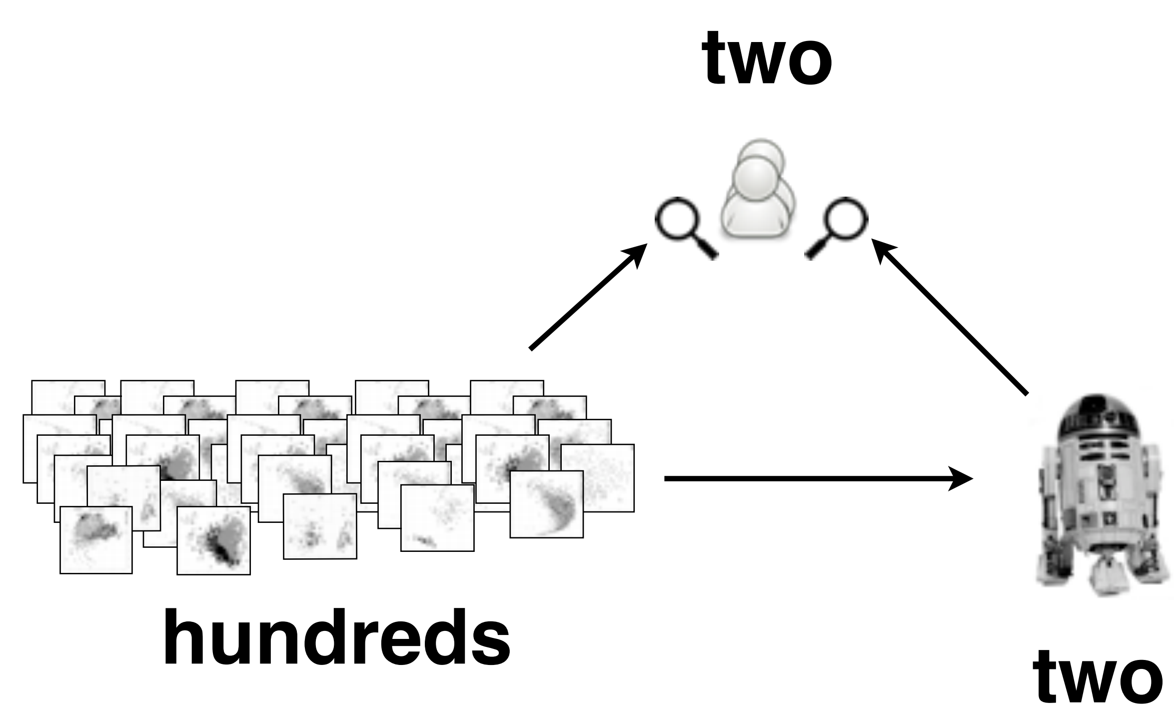
 →  → Good!

User vs. Data Study

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

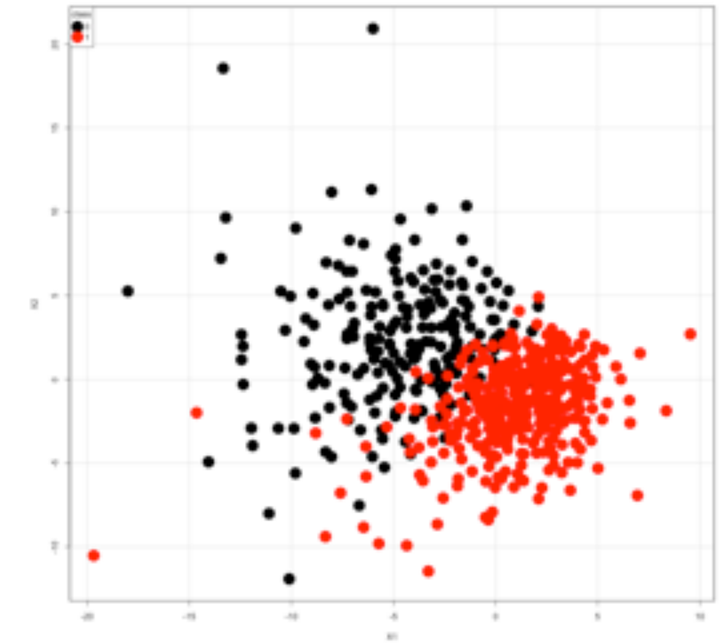


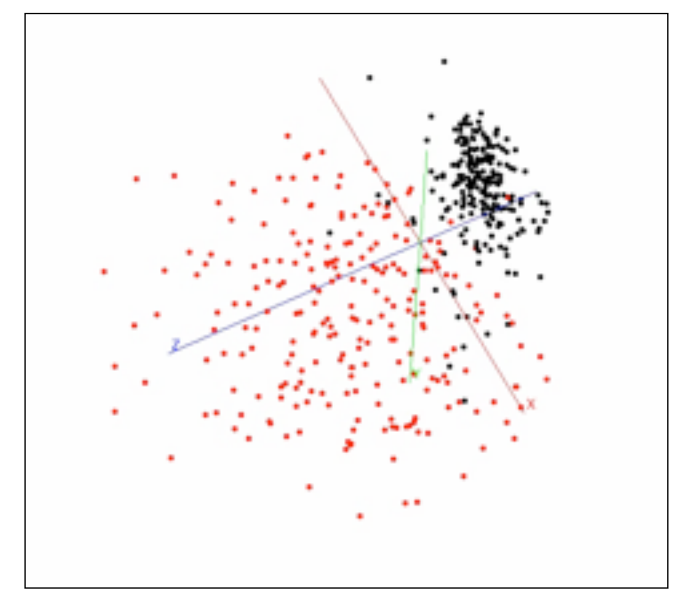
- data study
  - few users, many datasets

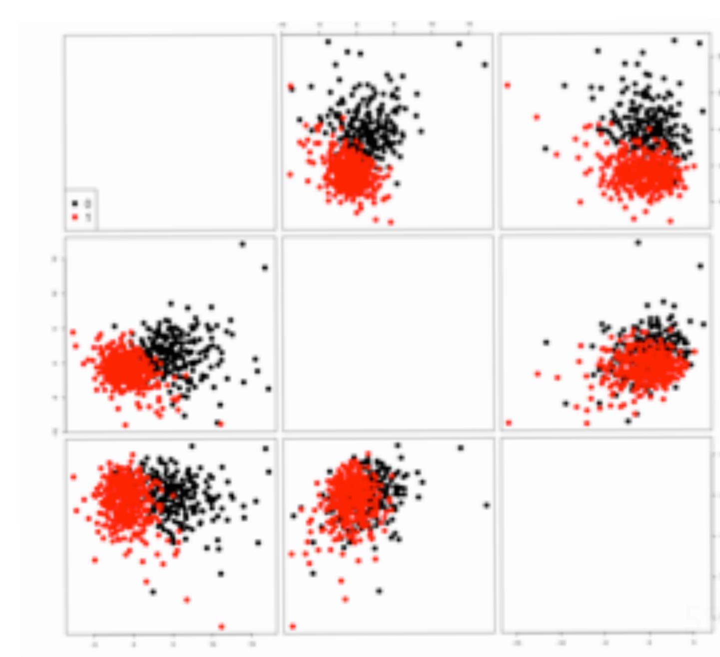


816 Dataset Instances

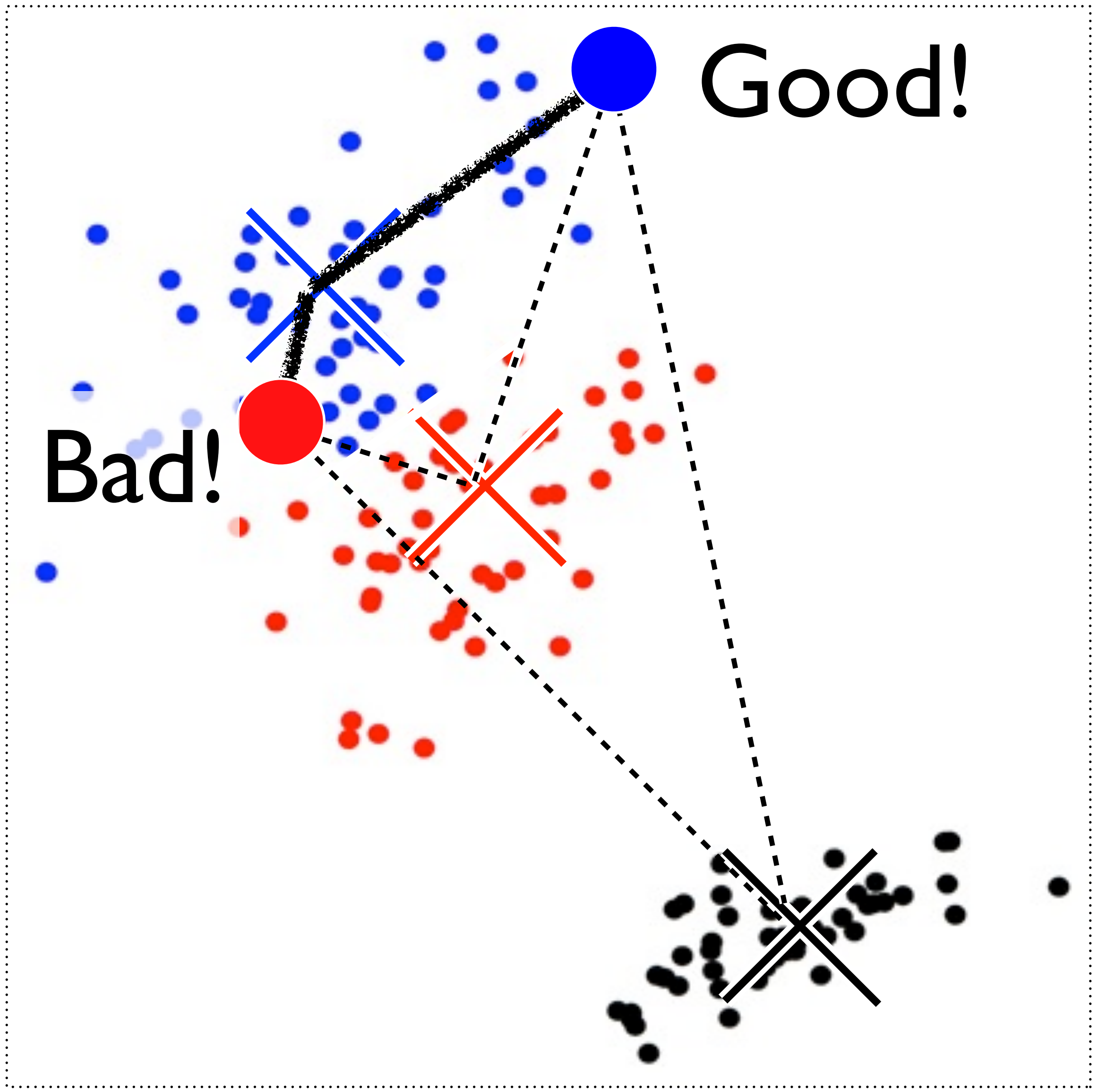
- 75 datasets
  - 31 real, 44 synthetic
  - pre-classified
- 4 DR methods
  - PCA
  - Robust PCA
  - Glimmer MDS
  - t-SNE
- 3 visual encoding methods
  - 2D scatterplots, 3D scatterplots, 2D SPLOMs
  - color-coded by class







Centroid Measure



Centroid: 93

Analysis Approach

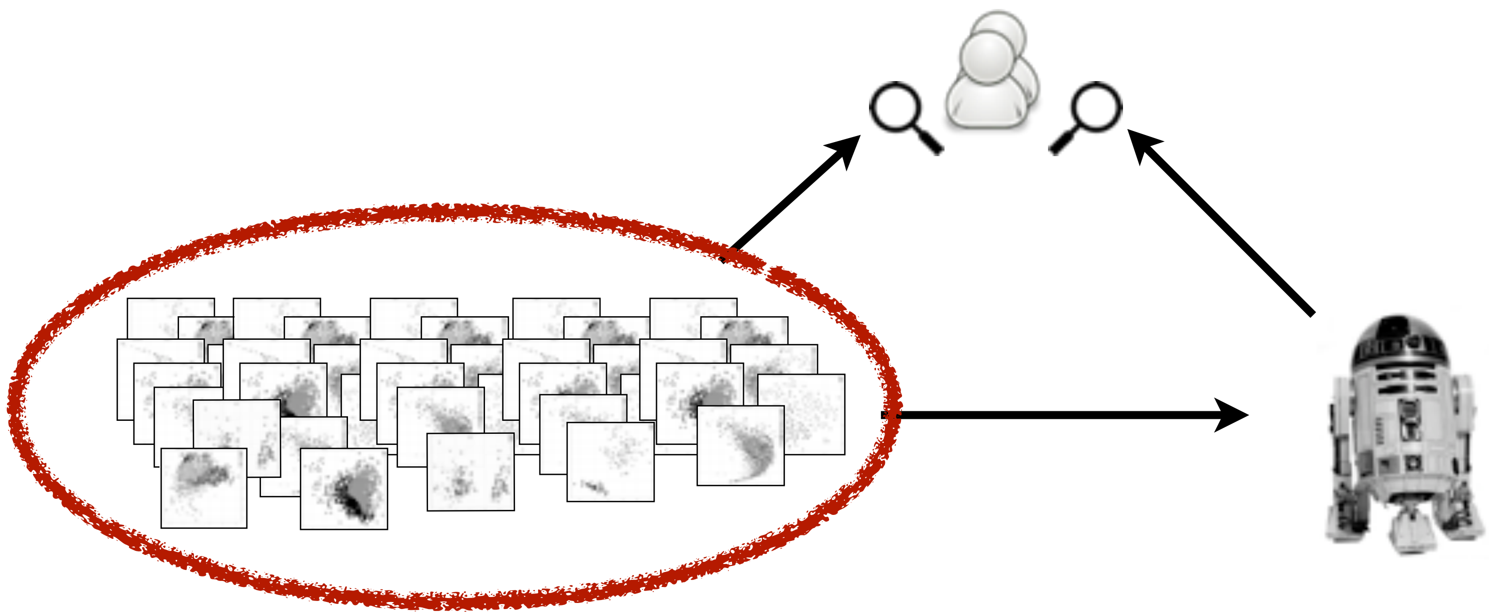
- qualitative method out of social science: coding
  - open coding: gradually build/refine code set
  - axial coding: relationships between categories

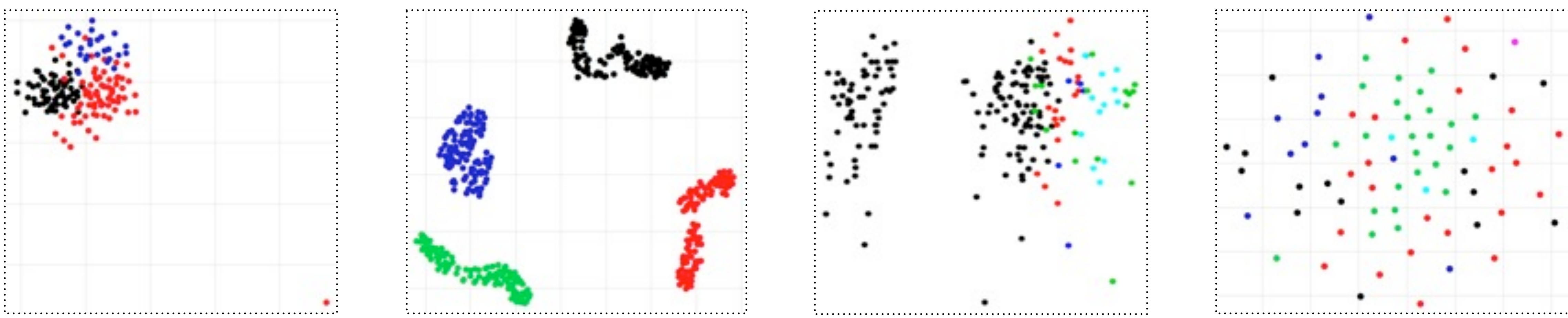
    Charmaz, K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. 2006.

    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?

Qualitative Analysis I: Cluster Separation Factors





outlier **shape** **split** **equidistant points**

Analysis Approach

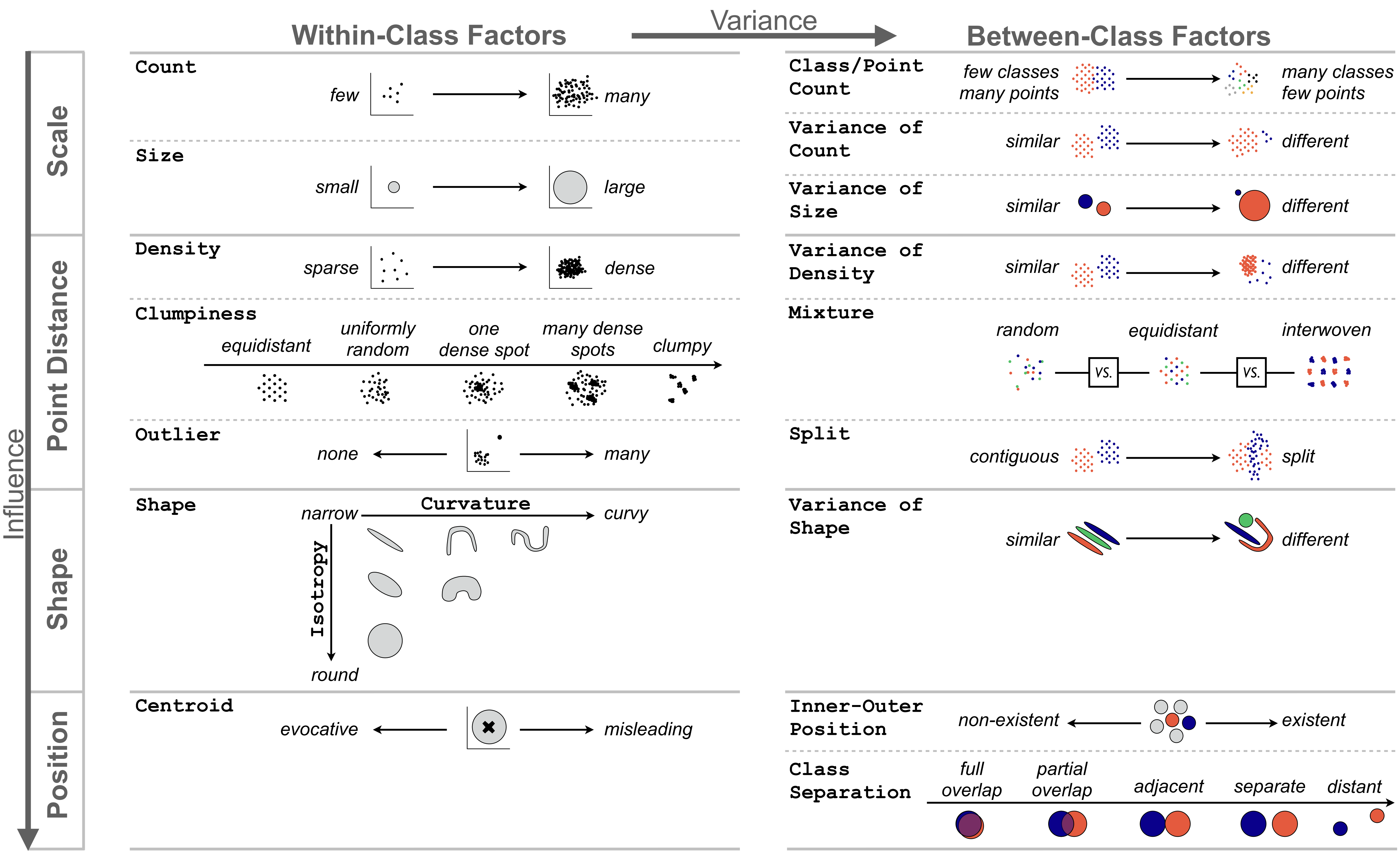
- qualitative method out of social science: coding
  - open coding: gradually build/refine code set
  - axial coding: relationships between categories

    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

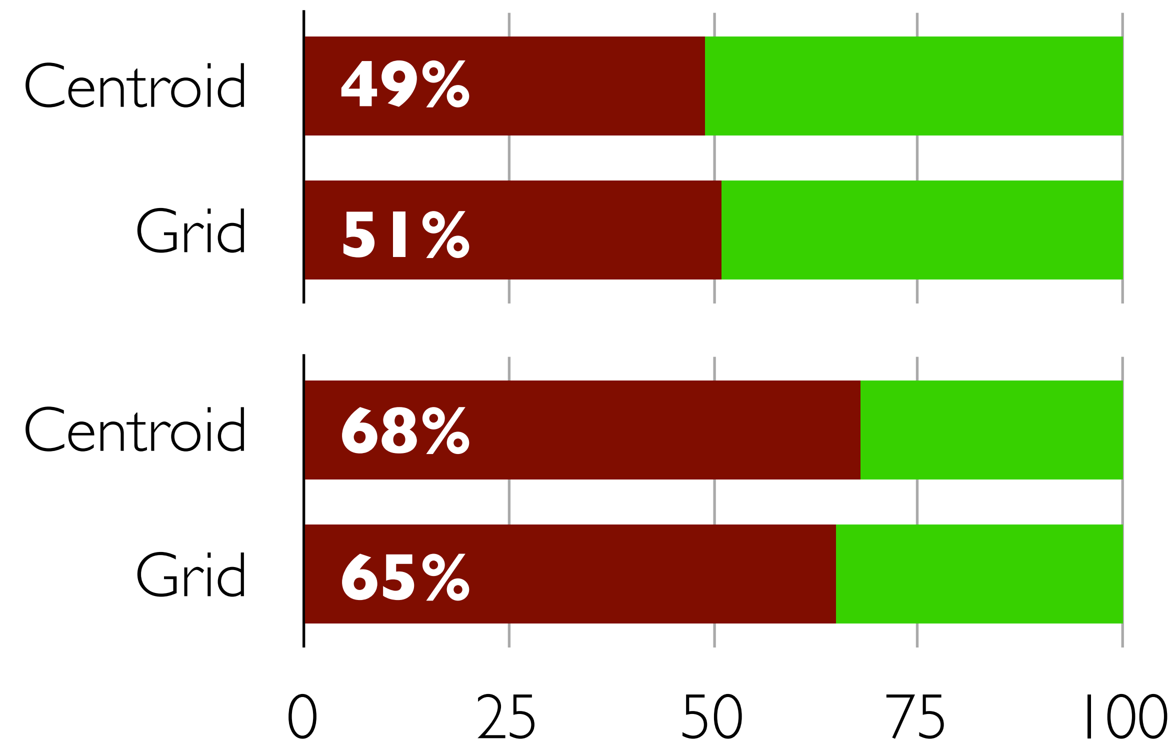


High-Level Results

Failure cases
Ok

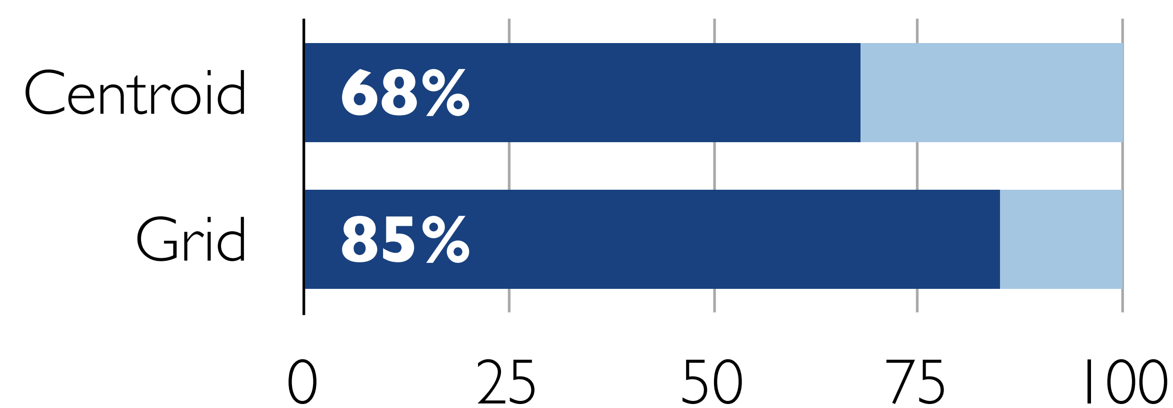
All (816)

Only real (296)



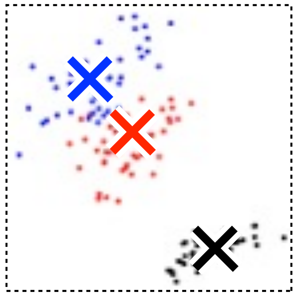
False Positives
False Negatives

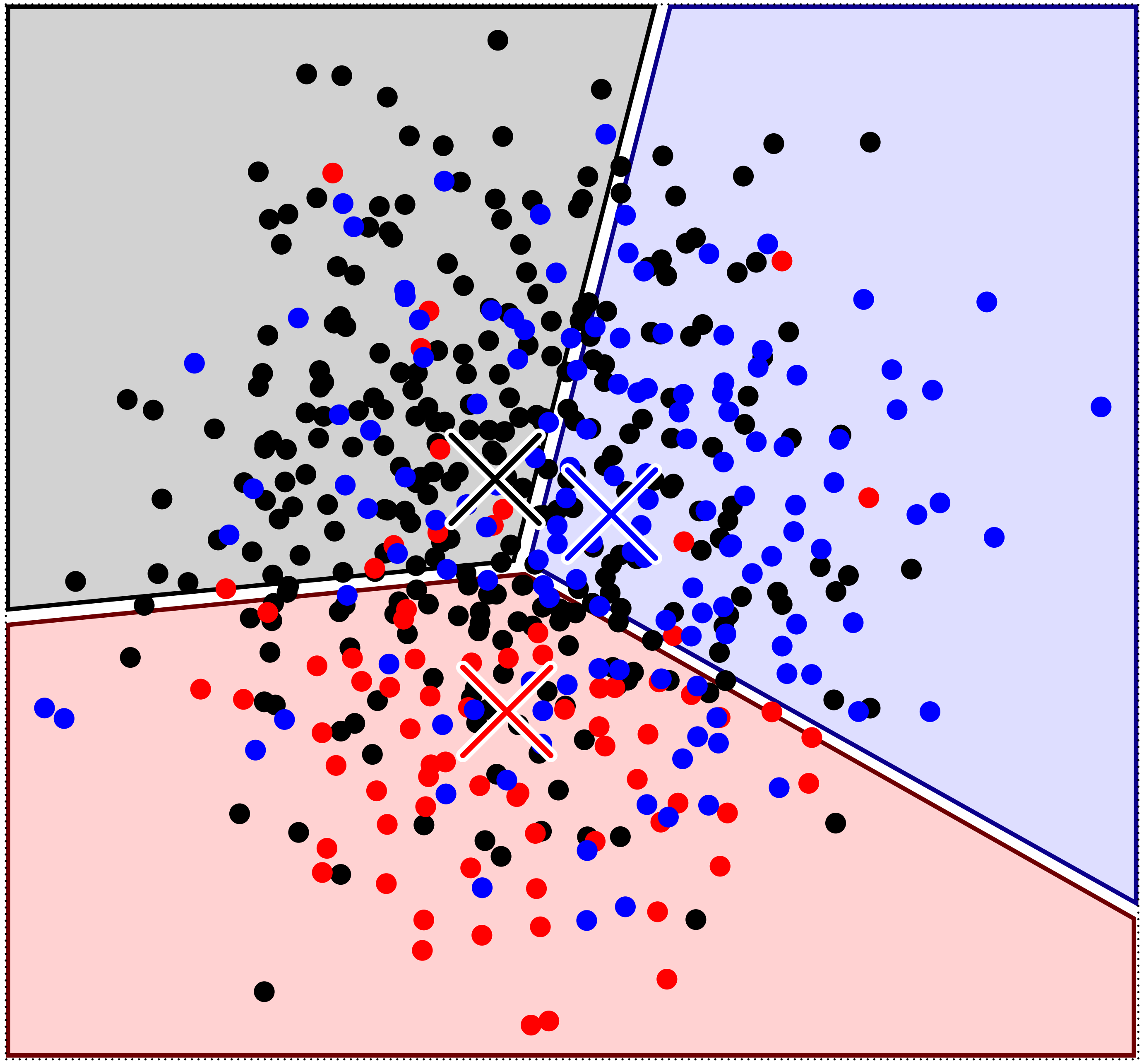
All failure cases



Centroid Failure Example

- big classes overspread small ones

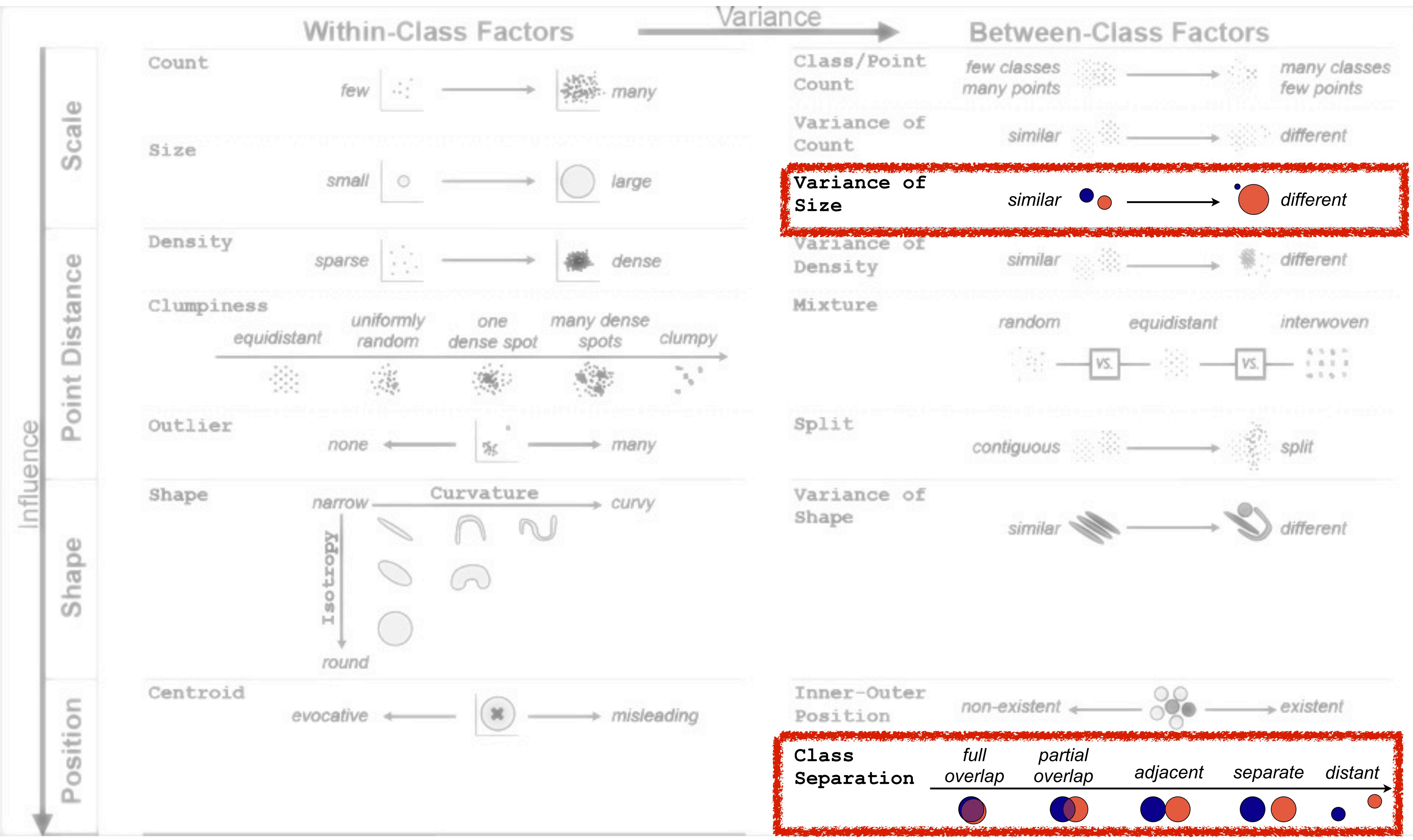




Red: **77 (Good)**
Problem: **FP**

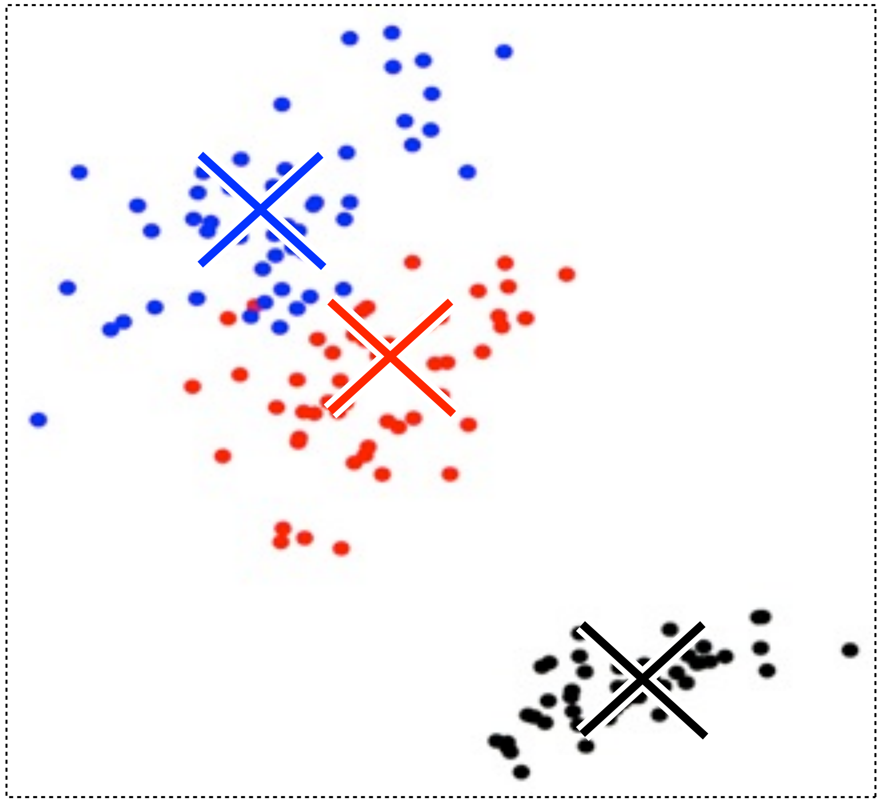
Data: Gaussian, synthetic
DR: MDS

Relevant Taxonomy Factors



Centroid: Mapping Assumptions Into Taxonomy

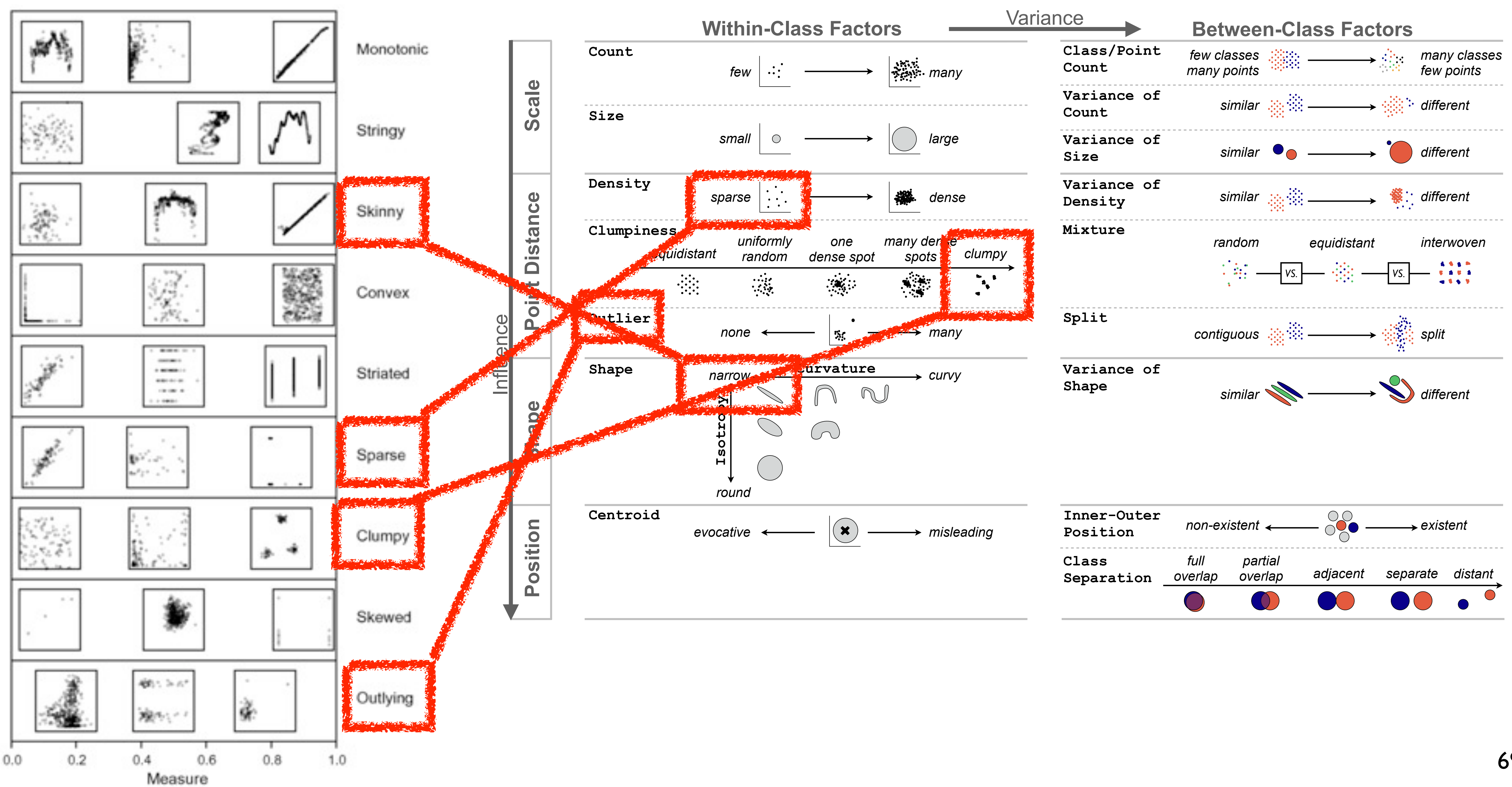
- centroid only reliable if
  - round-ish clusters
  - not more than one dense spot
  - no outliers
  - similar sizes & number of points



- rarely true for real datasets

Related Work

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



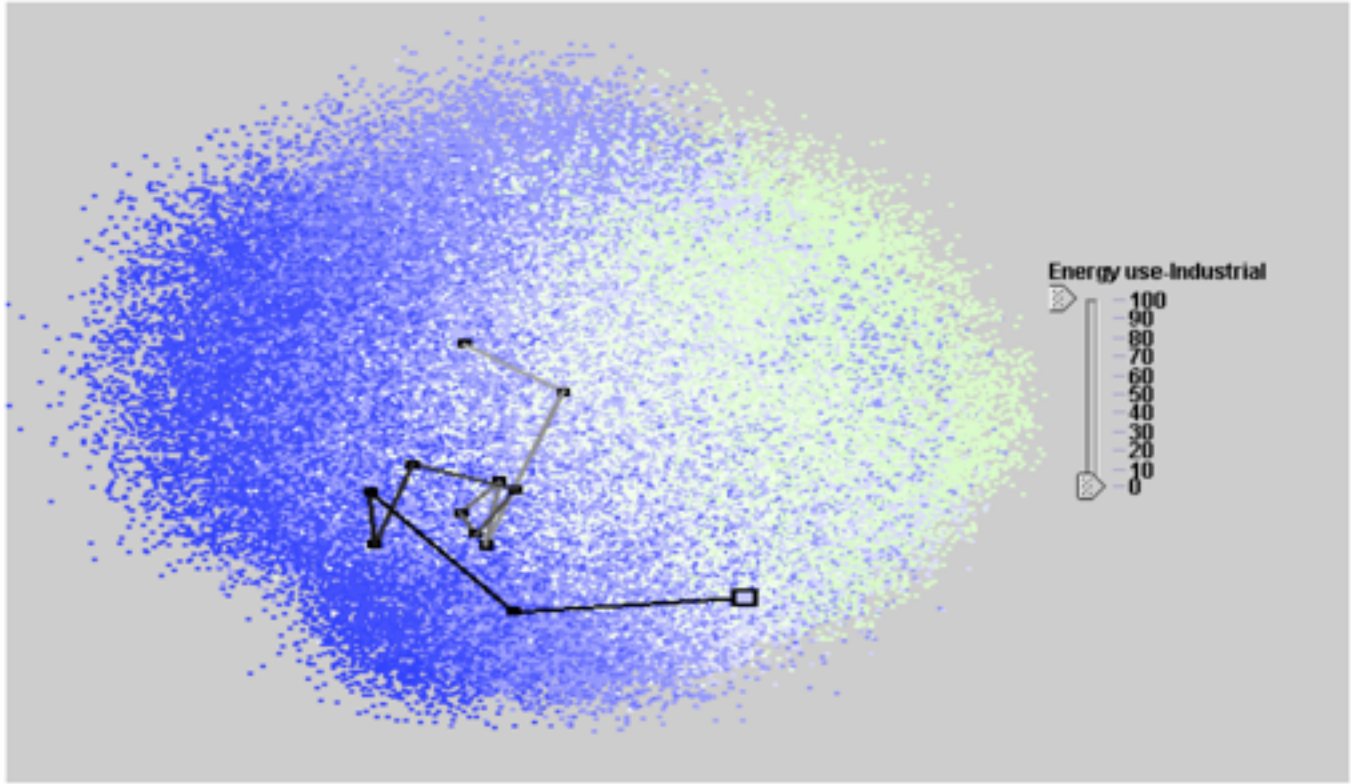
Methods and Outcomes

- 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/

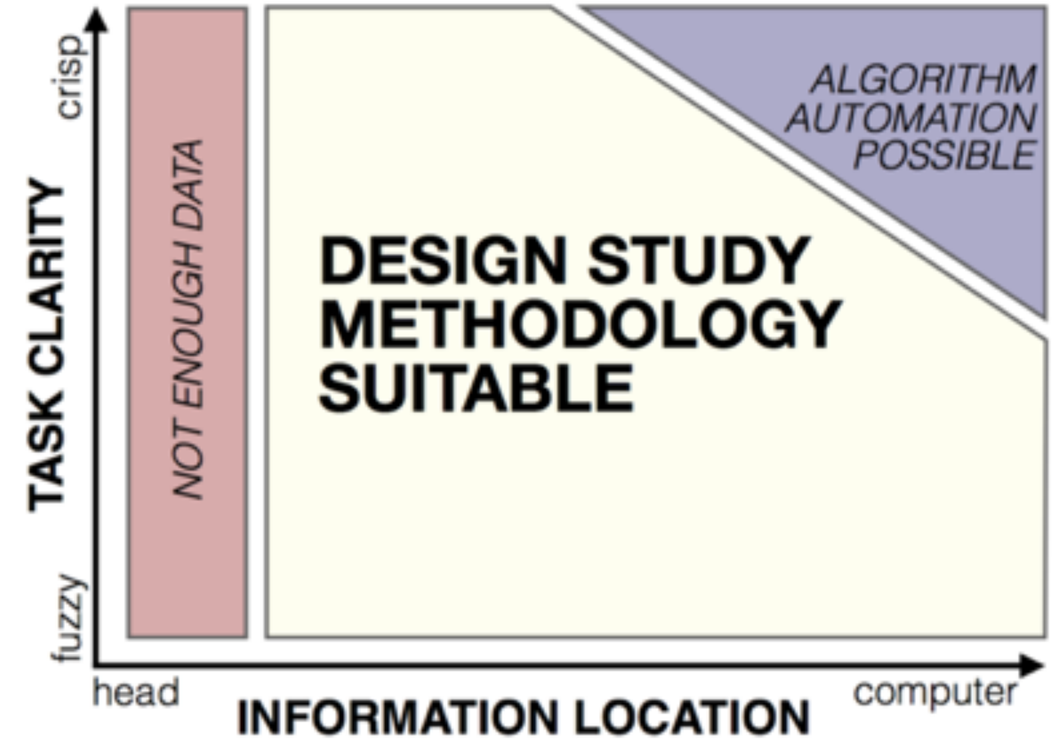
Reflections on QuestVis: A Visualization System for an Environmental Sustainability Model
Munzner, Barsky, Williams.
Scientific Visualization: Interactions, Features, Metaphors. Dagstuhl Follow-Ups 2, 2011, Chapter 17, p 240–259.

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

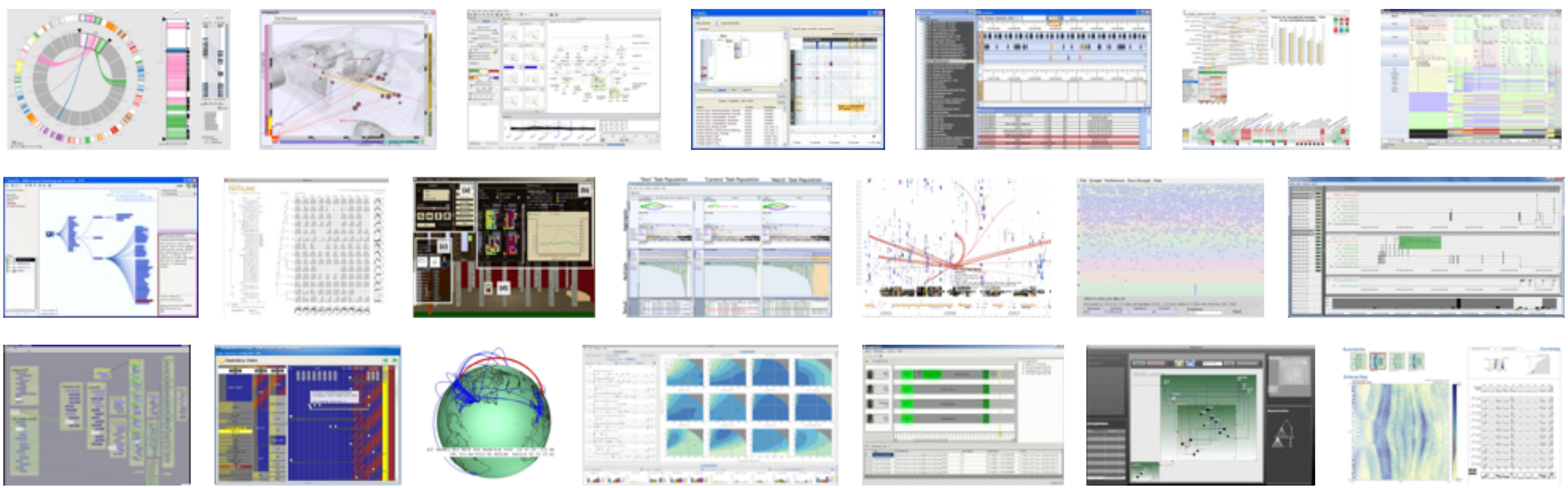
joint work with:
Michael Sedlmair, Miriah Meyer

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

Design Study Methodology: Reflections from the Trenches and from the Stacks.
Sedlmair, Meyer, Munzner. IEEE TVCG 18(12): 2431-2440, 2012 (Proc. InfoVis 2012).

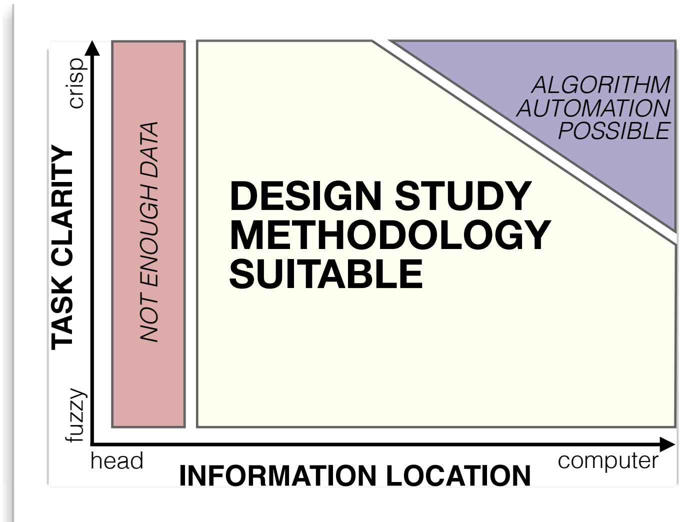
Design Studies

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

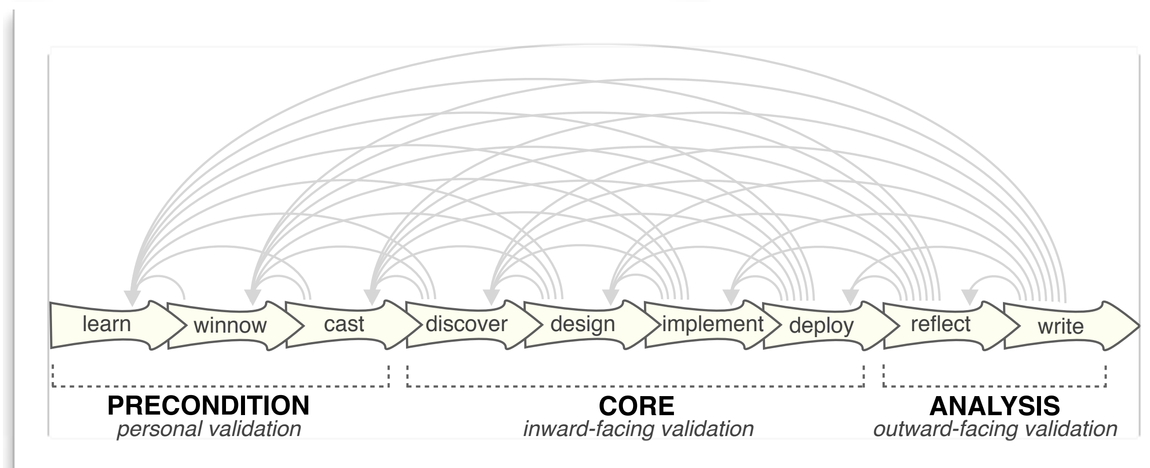


How To Do Design Studies

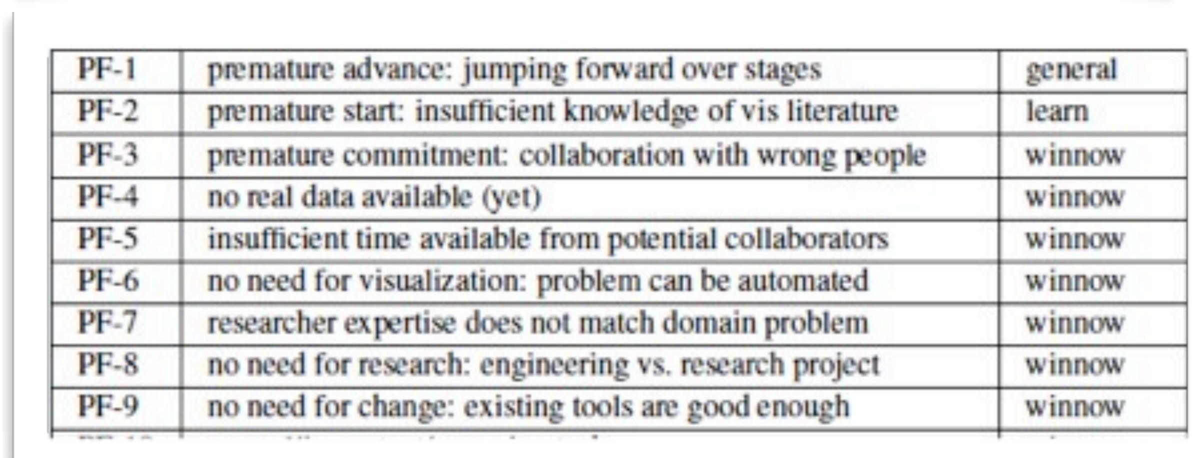
- definitions



- 9-stage framework



- 32 pitfalls and how to avoid them



PF-1	premature advance: jumping forward over stages	general
PF-2	premature start: insufficient knowledge of vis literature	learn
PF-3	premature commitment: collaboration with wrong people	winnow
PF-4	no real data available (yet)	winnow
PF-5	insufficient time available from potential collaborators	winnow
PF-6	no need for visualization: problem can be automated	winnow
PF-7	researcher expertise does not match domain problem	winnow
PF-8	no need for research: engineering vs. research project	winnow
PF-9	no need for change: existing tools are good enough	winnow

Pitfall Example: Premature Publishing

technique-driven

Must be first!

problem-driven

Am I ready?

http://www.prlog.org/10480334-wolverhampton-horse-racing-live-streaming-wolverhampton-handicap-8-jan-2010.html

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

- further info
  - http://www.cs.ubc.ca/~tmm/talks.html#sydney15
  - 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 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