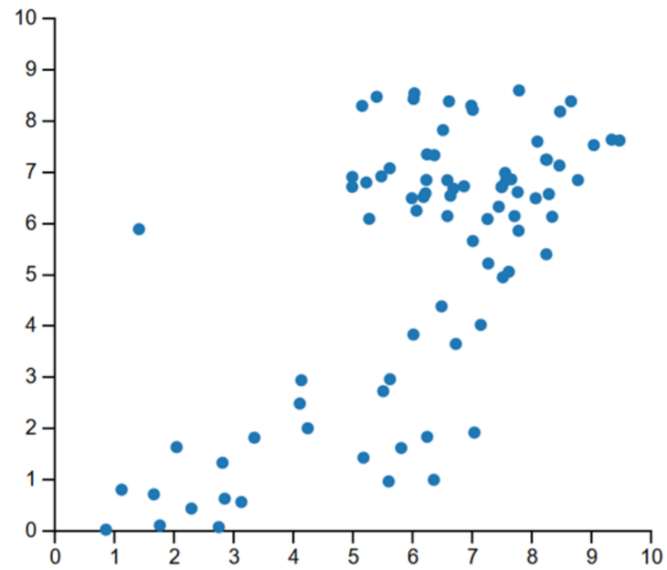
SCATTERPLOTS: TASKS, DATA AND DESIGN

A. Sarikaya and M. Gleicher
IEEE Transaction on Visualization and Computer Graphics

Presented By:
Shareen Mahmud

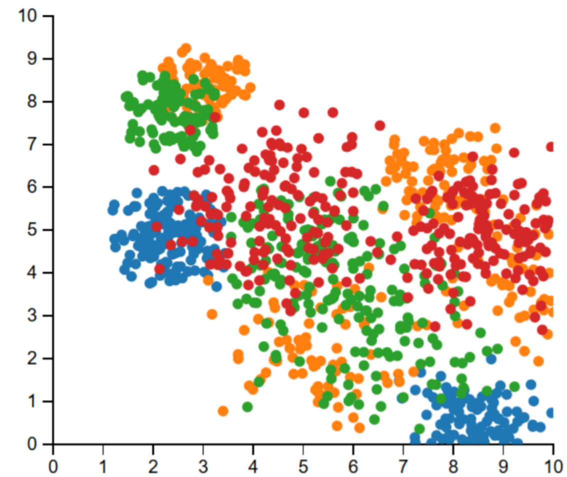
WHAT IS A TRADITIONAL SCATTERPLOT?

- Encodes two quantitative variables using the vertical and horizontal spatial position channels
- Each object in a dataset is represented with a point (mark)
- Effective in providing overviews, finding outliers, and judging correlation



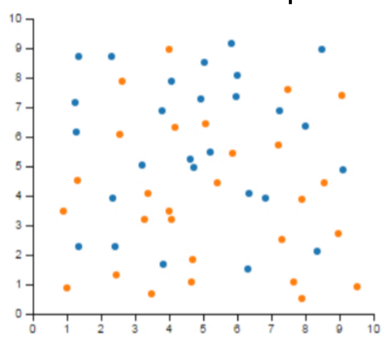
DOES IT FAIL?

- Yes! As data grows in scale, traditional scatterplots can become ineffective
- Overdraw is a concern where points overlap one another and masks points drawn under them.

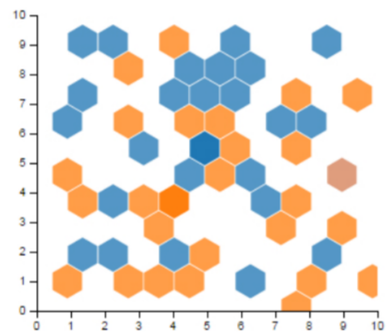


DIFFERENT DESIGNS SOLUTIONS

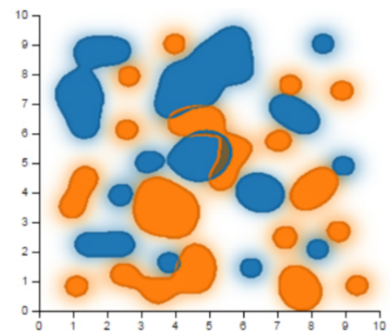
Traditional Scatterplot



Binned Scatterplot



Splatterplot



Designers have little guidance in how to select among choices. Which design to choose?

GOAL OF THE PAPER

- Help designers select scatterplot designs that are appropriate to their scenarios
- Identify factors that affect the appropriateness of scatterplot designs
- Create a framework based on the analysis goal and data characteristics

FACTORS THAT AFFECT THE DESIGN OF SCATTERPLOTS

- Analysis Tasks: What do viewers do with a scatterplot?
- Data Characteristics: How do they prompt changes in design?
- Design Decisions: What design variables need to be constructed?

ANALYSIS TASKS

- Gathered 23 model tasks from various vis literature to capture what viewers do with scatterplots
- Four data visualization experts performed an open card sort where tasks were grouped together based on their similarity
- Refined the categories post hoc to generate a complete picture of the task space

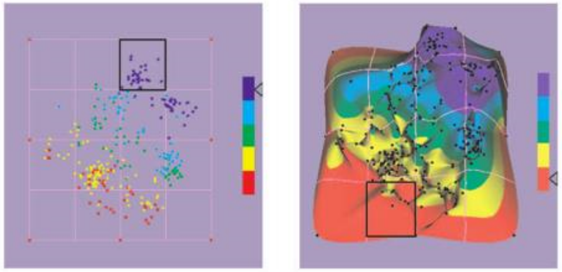


Fig. 2 Example trials from our experiment. Target levels are 5 (blue) in the left example and 1 (red) in the right example. Correct answers are highlighted with black outlines.

Task: Which section of the graph has the most dots of [this] color?

M. Tory, et al. Spatialization design: Comparing points and landscapes. IEEE Transactions on Visualization and Computer Graphics, 13(6): 1262–1269, 2007.

ANALYSIS TASKS

- A final list of 12 tasks split into 3 categories

  Object Centric

  Browsing

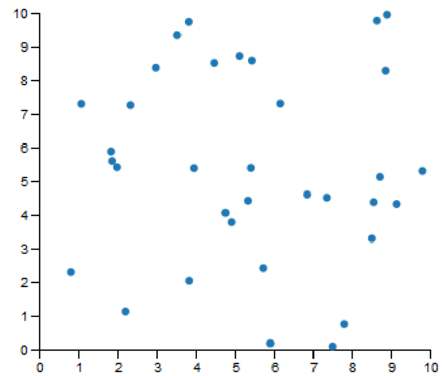
  Aggregate Level

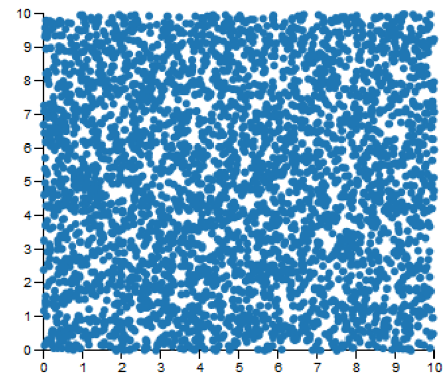
- A combination of these tasks can be used as building blocks to achieve an analysis goal

	# Task	Description
object-centric	1 Identify object	Identify the referent from the representation
	2 Locate object	Find a particular object in its new spatialization
	3 Verify object	Reconcile attribute of an object with its spatialization (or other encoding)
	4 Object comparison	Do objects have similar attributes? Are these objects similar in some way?
browsing	5 Explore neighborhood	Explore the properties of objects in a neighborhood
	6 Search for known motif	Find a particular known pattern (cluster, correlation)
	7 Explore data	Look for things that look unusual, global trends
aggregate-level	8 Characterize distribution	Do objects cluster? Part of a manifold? Range of values?
	9 Identify anomalies	Find objects that do not match the 'modal' distribution
	10 Identify correlation	Determine level of correlation
	11 Numerosity comparison	Compare the numerosity/density in different regions of the graph
	12 Understand distances	Understanding a given spatialization (e.g., relative distances)

DATA CHARACTERISTICS

Data characteristics can influence the design of an appropriate scatterplot

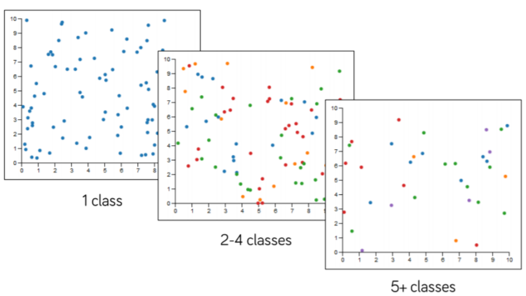




HTTPS://ALPER.DATAV.IS/ASSETS/PUBLICATIONS/SCATTERPLOTS/SCATTERPLOT-TALK.PDF

DATA CHARACTERISTICS

List of design affecting data characteristics collected from the literature



Data Attribute	Possible Values	Relevant Work
Class label	No class label, 2-4 classes, 5+ classes	Elliott and Rensink [2015], Gramazio et al. [2014], Sips et al. [2009]
Num. of points	Small (<10), medium (10–100), large (100–1000), very large (>1000)	Cottam et al. [2013], Gleicher et al. [2013], Keim et al. [2010], Mayorga and Gleicher [2013], Tory et al. [2007]
Num. of dimensions	Two continuous, two derived, or >2 dimensions	Best et al. [2006], Chan et al. [2010], Sedlmair et al. [2013]
Spatial nature	Dimensions do/do not map to spatial position	MacEachren [1995], Montello et al. [2003]
Data distribution	Random, linear correlation, overlap, manifolds, clusters	Bertini et al. [2011], Li et al. [2008], Rensink and Baldridge [2010], Sedlmair et al. [2013], Sips et al. [2009], Tatu et al. [2010], Dang and Wilkinson [2014], Wilkinson et al. [2005]

DESIGN DECISION

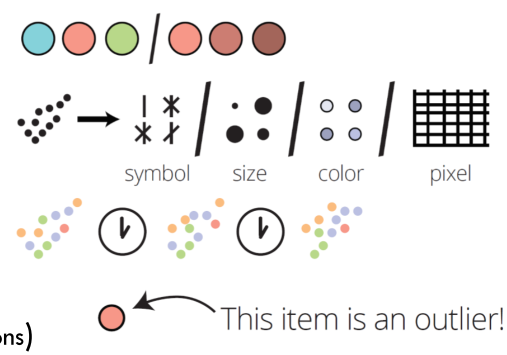
- Identified design decisions by applying a keyword ("scatter") search methodology on 3040 vis papers.
- Clustered the design choices into 4 groups

Point Encoding (Example: Color)

Point Grouping (Example: Binning)

Point Position (Example: Animation)

Graph Amenities (Example: Annotations)



- Interaction Intent

Cluster	Design Choice	Example
Point Encoding	Color	
	Size	
	Symbols	
	Outline	
	Opacity	
	Texture	
	Depth of Field	
	Blurriness	
Point Grouping	Representation Type	implicit explicit
	Positional Binning	symbol size color pixel
	Polygon Enclosure	convex hull statistical density
	Shape Abstraction	
Point Position	Subsampling	
	Displacement	
	Animation	
	Projection	
	Zooming	
Graph Amenities	Grid Lines	
	Axis Ticks	
	Legend	Series 1 Series 2
	Trend Lines	linear nonlinear
	Annotations	This item is an outlier!

DESIGN SPACE TO EVALUATE APPROPRIATENESS OF DESIGN STRATEGIES

	#	Task
object-centric	1	Identify object
	2	Locate object
	3	Verify object
	4	Object comparison
browsing	5	Explore neighborhood
	6	Search for known motif
	7	Explore data
aggregate-level	8	Characterize distribution
	9	Identify anomalies
	10	Identify correlation
	11	Numerosity comparison
	12	Understand distances

Cross product of these three is huge!

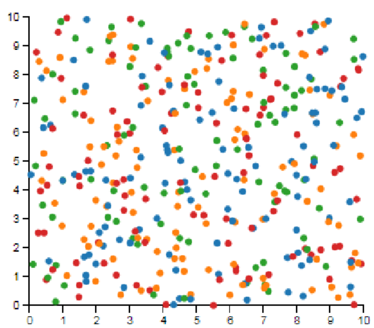
Leads to over 4300 discrete scatterplot scenarios

Data Attribute	Possible Values
Class label	No class label, 2-4 classes, 5+ classes
Num. of points	Small (<10), medium (10–100), large (100–1000), very large (>1000)
Num. of dimensions	Two continuous, two derived, or >2 dimensions
Spatial nature	Dimensions do/do not map to spatial position
Data distribution	Random, linear correlation, overlap, manifolds. clusters

Cluster	Design Choice	Example
Point Encoding	Color	
	Size	
	Symbols	
	Outline	
	Opacity	
	Texture	
	Depth of Field	
	Blurriness	
Point Grouping	Representation Type	
	Positional Binning	
	Polygon Enclosure	
	Shape Abstraction	
Point Position	Subsampling	
	Displacement	
	Animation	
	Projection	
	Zooming	
Graph Amenities	Grid Lines	
	Axis Ticks	
	Legend	
	Trend Lines	
	Annotations	

A SLICE OF THE SPACE: TASK & DESIGN STRATEGIES

- Framework illustrated with a 2D slice of the entire grid (60 out of 4300 grids)
- Entire set of tasks and design strategies
- Data characteristics fixed to "large" number of points and classes with an unstructured distribution of data

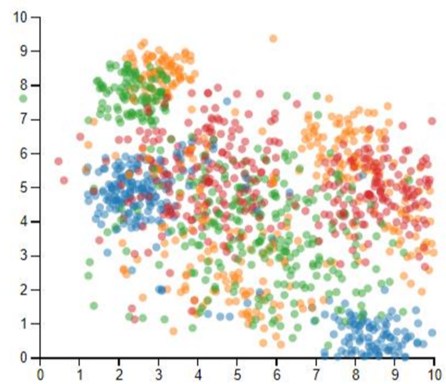


Task	A Point encoding	B Point position	C Point grouping	D Interaction intent	E Graph amenities
1 Identify object	✔	✔	✧	✔	✔*
2 Locate object	✔	✧	✧	✔	✔
3 Verify object	✔	✔*	✧	✔	✔
4 Compare objects	✔	✔	✧	✔	✔
5 Explore neighborhood	✔	✔	✔	✔	✔
6 Search for motif	✔	✔	✔	✔	✔*
7 Explore data	✔	✔	✔	✔	✔
8 Characterize distribution	✔	✔	✔	✧	✔
9 Find anomalies	✧	✔*	✧	✔*	✔
10 Identify correlation	✘	✘	✔	✘	✔
11 Characterize numerosity	✘	✘	✔	✘	✘
12 Characterize distances	✔*	✔	✔*	✔*	✔

✔ general support
✔* support in particular situations
✧ requires concurrent support from other encodings
✘ no improvement to task support

USING THE FRAMEWORK

- Difficult to support aggregate level tasks such as identifying anomalies, correlations and object density with point encoding and position (9A-11B)



Task	A Point encoding	B Point position	C Point grouping	D Interaction intent	E Graph amenities
1 Identify object	✔	✔	✧	✔	✔*
2 Locate object	✔	✧	✧	✔	✔
3 Verify object	✔	✔*	✧	✔	✔
4 Compare objects	✔	✔	✧	✔	✔
5 Explore neighborhood	✔	✔	✔	✔	✔
6 Search for motif	✔	✔	✔	✔	✔*
7 Explore data	✔	✔	✔	✔	✔
8 Characterize distribution	✔	✔	✔	✧	✔
9 Find anomalies	✧	✔*	✧	✔*	✔
10 Identify correlation	✘	✘	✔	✘	✔
11 Characterize numerosity	✘	✘	✔	✘	✘
12 Characterize distances	✔*	✔	✔*	✔*	✔

✔ general support
✔* support in particular situations
✧ requires concurrent support from other encodings
✘ no improvement to task support

USING THE FRAMEWORK

- Point grouping hurts object-centric tasks (1C-4C, 9C, 12C)
- However, by compositing point encoding, point position and interaction intent, object centric tasks can be supported.

Task	A Point encoding	B Point position	C Point grouping	D Interaction intent	E Graph amenities
1 Identify object	✔	✔	✧	✔	✔*
2 Locate object	✔	✧	✧	✔	✔
3 Verify object	✔	✔*	✧	✔	✔
4 Compare objects	✔	✔	✧	✔	✔
5 Explore neighborhood	✔	✔	✔	✔	✔
6 Search for motif	✔	✔	✔	✔	✔*
7 Explore data	✔	✔	✔	✔	✔
8 Characterize distribution	✔	✔	✔	✧	✔
9 Find anomalies	✧	✔*	✧	✔*	✔
10 Identify correlation	✘	✘	✔	✘	✔
11 Characterize numerosity	✘	✘	✔	✘	✘
12 Characterize distances	✔*	✔	✔*	✔*	✔

✔ general support
✔* support in particular situations
✧ requires concurrent support from other encodings
✘ no improvement to task support

WHAT-WHY-HOW ANALYSIS

Idiom	Scatterplots (Framework)
What: Data	Vis literature; papers
What: Derived	Table with Tasks, Data characteristics, Design choices
Why: Tasks	Compare design strategies
How: Encode	Multidimensional table, Color highlighting, marks to denote appropriateness of design decisions
How: Reduce	Dimensionality Reduction/Slicing
Scale	4300 scatterplot scenarios

STRENGTH AND LIMITATIONS

- Strengths

-First to identify scenarios specific to scatterplot design

-Provides scope to discover potential areas for future innovation in scatterplot design

- Provides a good reference point for designers to get started with scatterplot design

- Limitation

-Infeasible to present the high dimensional grid. Data characteristics were restricted

-Focuses on single scatterplot design. Multi scatterplot tasks were discarded

-Misses the evaluation component is the study. How useful did designers find this framework to be?

REFERENCES

Paper: https://alper.datav.is/assets/publications/scatterplots/scatterplots-preprint.pdf

Slides: https://alper.datav.is/assets/publications/scatterplots/scatterplot-talk.pdf

Project Page: http://graphics.cs.wisc.edu/Vis/scattertasks/