# Seeing, Scatterplots and Shakespeare

**Michael Gleicher**
Department of Computer Sciences
University of Wisconsin Madison
(on sabbatical at INRIA Rhone-Alpes)
Team Imagine

**Happy (American) Thanksgiving!**

## Acknowledgements

This work is a collaboration with a great set of people!

**Students:**
Danielle Albers, Eric Alexander, Michael Correll, Adrian Mayorga
**Collaborators:**
Steve Franconeri, Jonathan Hope, Robin Valenza, Mike Witmore

This research in funded in part, by the National Science Foundation and the Andrew Mellon Foundation

Seeing, Scatterplots and Shakespeare

Michael Gleicher
Department of Computer Sciences
University of Wisconsin Madison
(on sabbatical at INRIA Rhone-Alpes)
Team Imagine

## Data Visualization

How do we use **pictures** to help **understand**, and **communicate** data?

## Data Visualization

Seeing

Scatterplots

Shakespeare

## Data Visualization

Seeing

Scatterplots

Shakespeare?

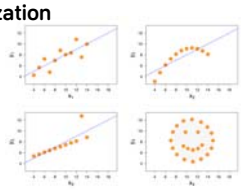
Perceptual Science:
How do people see?
How do we use this knowledge?

## Data Visualization

Seeing

Scatterplots

Shakespeare?



Re-Examine Basic Methods:
Consider foundations
Impact what really gets used!

## Data Visualization

Seeing

Scatterplots

Shakespeare?


## Shakespeare?

Literary Scholarship as a motivating domain

Literary Scholarship as a way of thinking

New ways of thinking -> New approaches

Until we take the time to learn about how the other side thinks, we can't really work **together**.

Once we learn how each other thinks, our ways of thinking can infuse each other's.

This is not just building tools for our friends.

It's a lot more fun and interesting


**One journal cover image leads to (at least) three challenges**

The axes are meaningless!

Explainers – crafted projections
VAST 2013

Can people interpret this?

Perception of average value in scatterplots
InfoVis 2013

The scatterplot has too many points!

Splatterplots – scalable scatterplots
TVCG 2013

## Visualizing English Print 1470-1800

What if you had access to all surviving books?

## Why?

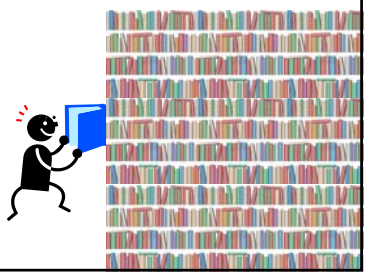
Consider larger collections of books

Consider language not content

See patterns across language

See small scale patterns in familiar texts

Be uncultured and still hang out with the cool kids



## Literary Scholarship...

The statistics are not the argument

Exemplars and Outliers
Go back to the sources

Arguments based on context and knowledge

Multiple viewpoints and lenses

# Seeing Shakespeare: Scalable Scholarship

Study Literature without Reading?

## Texts as Data?

Need to turn books into numbers

## Text tagging

Text

Vector

```
4, 0, 3.6, 4.7, 0, 3.4, …
```

Count of word Category 1 — Count of word Category 2 — Count of word Category 3

## Just counting

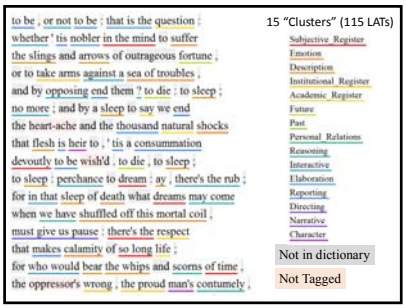
Words (phrases) have a type (tag)

Docuscope

Simple matching against a dictionary

Hand-built dictionaries

100-115 Categories, 12-17 Clusters (groups)




Texts

Vectors

```
4, 0, 3.6, 4.7, 0, 3.4, …
3, 2.4, 0, 4.2, 4.7, 5, …
1.5, 2.3, 0, 1.2, 6.2, …
…
```

## What about more sophisticated text analysis?

Simple methods are:

Easier to understand and explain

Focused on word usage without considering meaning

## What about more sophisticated text analysis?

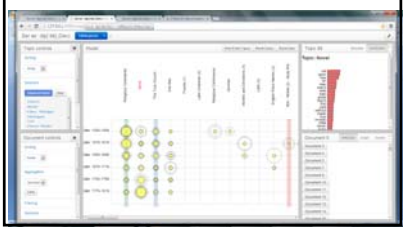
Topic modeling?

Yes, we're working on it too.

Eric Alexander, et al. Serendipity: turning topics back to texts. In preparation. 2013-2014.

## Serendipity: Exploring Topic Models

Eric Alexander, and others – in preparation, 2013-2014

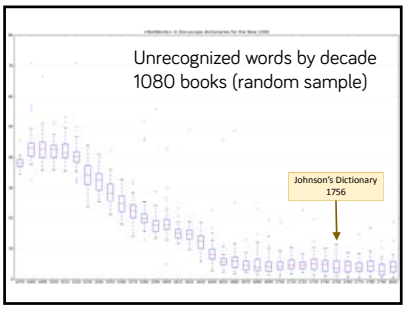


## Just counting

Words (phrases) have a type (tag)

Docuscope
Simple matching against a dictionary
Hand-built dictionaries

100-115 Categories, 12-17 Clusters (groups)




## How to look at 100+ dimensions?

## How to look at 100+ dimensions?

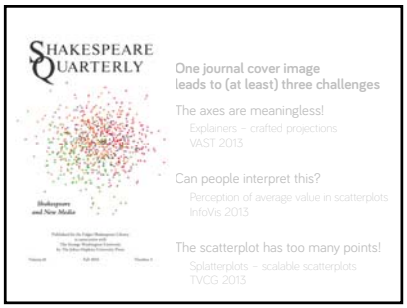
## How to support "Humanist" arguments?

The statistics are not the argument

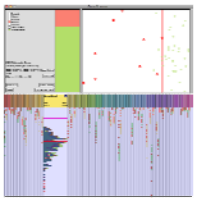
Exemplars and Outliers
Go back to the sources

Arguments based on context and knowledge

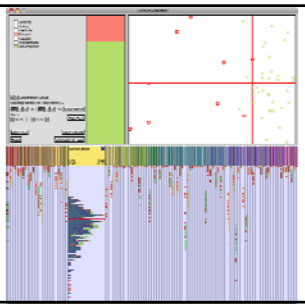
Multiple viewpoints and lenses



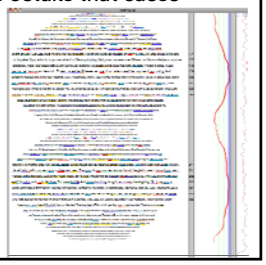

## Domain Specific Tools



Correll, Witmore, Gleicher. Exploring Collections of Tagged Text for Literary Scholarship. Computer Graphics Forum (Proceedings Eurovis). 2011.



## Back to the details that cause patterns




One journal cover image leads to (at least) three challenges

The axes are meaningless!

Explainers – crafted projections
VAST 2013

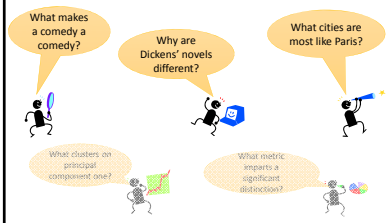
Can people interpret this?

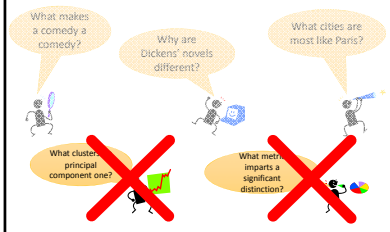
Perception of average value in scatterplots
InfoVis 2013

The scatterplot has too many points!

Splatterplots – scalable scatterplots
TVCG 2013

## How to look at 100+ dimensions?




## Explainers: Expert Explorations with Crafted Projections

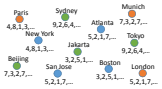
2013 IEEE VAST
Visual Analytics Science and Technology

Honorable Mention Award Winner

## Explainers

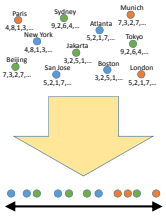
An approach to explore high dimensional data:

- Organize data according to **user**-defined concepts
- Explain **user**-defined concepts according to the data
- Give the **user** control over tradeoffs




High Dimensional Data

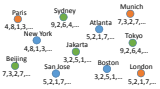
**Objects** have associated **Vectors**




Projections

**Functions** map **Vectors** to **Numbers**

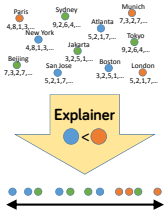
Produce a new **axis** or **dimension**
(or view)




Specification

User-defined concept
orange-ness
European-ness

like-the-marked-things-ness




**Explainers:**

Projections crafted to meet user specifications

With user control of tradeoffs between:

**Correctness:**

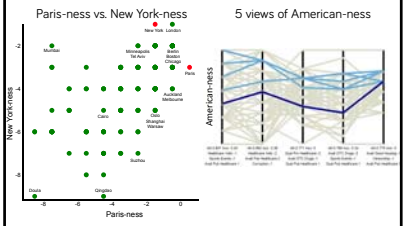
does it align with the user specification?

**Understandability:**

can the user interpret the mapping?

**Diversity:**

can we generate alternate mappings?




**Organize**

Relationships between points based on data and concepts

Rankings
Outliers
Extrema
Exemplars
Similarities

**Explain**

Relationships with the data connect concepts and variables

Where do the orderings come from?

Are variables correlated with concepts?

To make things concrete

## An Example: Shakespeare's Plays

**More Examples online!**

`http://graphics.cs.wisc.edu/Vis/Explainers`

Texts

Vectors

```
4, 0, 3.6, 4.7, 0, 3.4, …
3, 2.4, 0, 4.2, 4.7, 5, …
1.5, 2.3, 0, 1.2, 6.2, …
…
```

36 Plays = 36 Vectors

115 "Measurements" of each text = 115 dimensions

## genre

Categorization given by Shakespeare's contemporaries

Comedy | Tragedy | History

Category for plays written after that

Late Plays

## comedic-ness

A measure of how much of a comedy something is

It's the "stuff" comedies have more of
Where "stuff" has to be in the data

Tragedy
History
Late Plays
<
Comedy

**Organization:**

What is most/least comedic?

**Explanation:**

How is the word usage (measured stuff) different in comedies?


## A comedicness explainer

c = f (V)
c = comedicness
f = a function (**Explainer**) that maps from V to c
V = vector from a text (length 115)

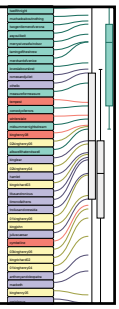
Choose f such that:

1. It is correct (meets specification)
2. It is understandable (simple)
3. We can have alternatives (other functions that meet 1 and 2) as well

## An Explainer

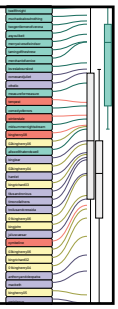
comedicness = M – I

f(V) = V[39] – V[42]



## Understanding the Visualization*

* The Visual Encoding is not a strong part. Suggestions are most welcomed!

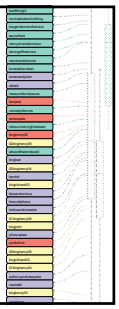


## Understanding the Visualization*

Objects in rank order

Color by specified class

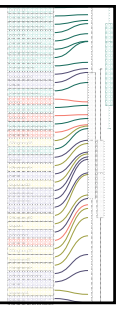
* The Visual Encoding is not a strong part. Suggestions are most welcomed!



## Understanding the Visualization*

Lines connect rank (left) to value (position on number line)

* The Visual Encoding is not a strong part. Suggestions are most welcomed!

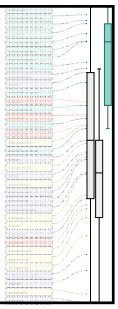


## Understanding the Visualization*

Box Plots show class separation

Left: all data
Right: each class of interest

* The Visual Encoding is not a strong part. Suggestions are most welcomed!

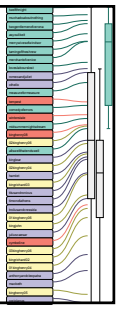


## An Explainer

comedicness = M – I

f(V) = V[39] – V[42]

Tradeoff:
**simple** (linear, 2 variables, unit coefficients)
but
**5 "wrong"**



## 5 Wrong

False Positives

False Negative

## Wrong?

Interesting Outliers
"Romeo and Juliet" is pretty comedic

Ambiguous Classifications
Late Plays are called Tragi-Comedies

Near-Misses
A tiny shift, and this would be different

| | |
|---|---|
| M – I | (5 wrong) |
| C – B – I | (4 wrong) |
| C – I – 10 M | (1 wrong) |
| 31 D – 100 M – 3 A | (none wrong) |

"standard" L1 SVM (none wrong, reasonable margin)

25.3698 Q + 11.8823 U + 6.9492 F + 5.4897 A + 4.1489 P - 3.3765 N + 2.6392 D + 2.0172 F - 1.5404 I + 1.1864 R - 0.7958 C + 0.7272 D

## What's Understandable*?

Simple form (linear vs. non-linear,…)
**A+B** vs. $e^{-wk(A,B)}$
Parsimony (few variables)
**A+B** vs. W+X+Y+Z
Simple Coefficients (small integers)
**A – 2B** vs. 1.235 A – 4.327 B
Familiar Variables
**A + B** vs Q + W

## Simpler Functions

Easier to Understand → More likely to lead to Theory

Less Expressive → Less Likely to be Accurate

## Tradeoffs

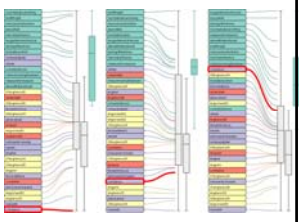
Give the **user** control over the tradeoffs

## Diversity

F + Q – I
C – M – I
P + N + D

**Same:**
Correctness
Simplicity

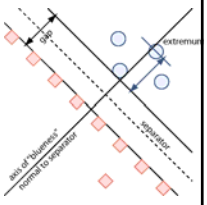
**Different:**
Explanations
Orderings



## How to find functions?

Optimization problem
Minimize
amount "wrong"
"Cost" of function

Support Vector Machine (SVM)




## How to Implement Explainers?

**Fancy Math**
Encode tradeoffs into the SVM
Solve for the best tradeoffs
Adjust parameters to tune

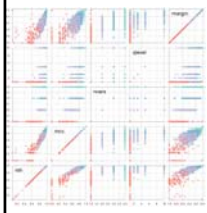
Solve one big optimization problem

*Not a standard SVM, so needs a slow and finicky solver*

**Brute Force**
Sample space of variable sets
Solve an SVM for each
Sort and filter to find interesting ones

Solve many small optimization problems

*Generates a diverse and interesting exploration of tradeoffs*

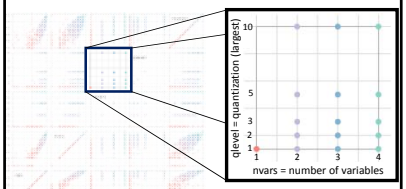


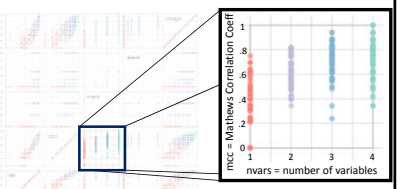
**Finding the interesting explainers**

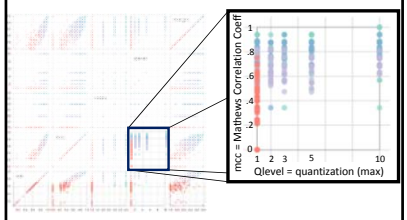
2667 Explainers generated

Rank-by-Feature or "Scagnostics" style analysis

2667 points – each an explainer
5 properties of each






## Isn't this just...

Some prior approaches to help situate our work

**Explainers**

Organize data according to user-defined concepts
Explain user-defined concepts according to data

**Explainers add:**
User-defined concepts
Control over tradeoffs
Connection between concepts and variables
Generation of alternatives

**Dimensionality Reduction**
e.g. PCA, CCA, IsoMap, ... - standard statistical and ML practices

Organize data according ~~to user-defined concepts~~
~~Explain user-defined concepts according to data~~

**Explainers add:**
User-defined concepts
Control over tradeoffs
Connection between concepts and variables
Generation of alternatives


M. Witmore and J. Hope. The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre. Shakespeare Quarterly, 61(3) 357-390.

A total of 776 pieces of Shakespeare's plays from the First Folio, each piece consisting of 1000 words, **rated on two scaled PCs (1 and 4)**. The cumulative proportion of variation accounted for by the first four principal components is 12.33 percent, with component 1 accounting for 3.83 percent and component 4 accounting for 2.35 percent.

**Machine Learning Classification Techniques**

~~Organize~~ data according to user-defined concepts
~~Explain user-defined concepts according to data~~

**Explainers add:**
User-defined concepts
Control over tradeoffs
Connection between concepts and variables
Generation of alternatives

**User-Driven Spatializations**
e.g. Semantic Interaction (Endert++), LAMP (Paulovich++), Star Coordinates (Kandogan), ...

Organize data according to user-defined concepts
~~Explain user-defined concepts according to data~~

**Explainers add:**
User-defined concepts
Control over tradeoffs
Connection between concepts and variables
Generation of alternatives

## More to do . . . (current limitations)

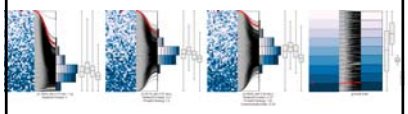
**User Experience**
- Visualizations
- Interactive Specification

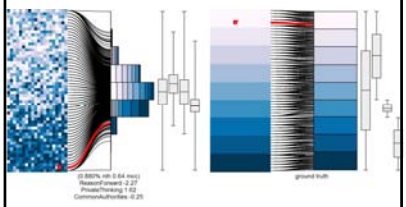
**Theory**
- Understanding understandability tradeoffs
- Statistical significance in negative results

**Scalability**
- More variables (redundancy)
- More objects
- More complex relationships

## 1080 Books

20 books per decade 1580-1800
Are books before 1680 different than books after 1710?






## Key Ideas

**User-defined** concepts
Multiple goals: **organize** and **explain**
User-control over **tradeoffs**: correctness, simplicity, diversity
**Alternative** viewpoints

Details:
Types of Simplicity
Implementation with SVM

**Explainers**

An approach to exploration and discovery in high dimensional data that

- **organizes** data according to **user-defined concepts**
- **explains** these **user-defined concepts** in terms of the data
- and generates **alternative viewpoints**

using machine learning techniques and providing control over **tradeoffs**.

**More Examples Online!**
`http://graphics.cs.wisc.edu/Vis/Explainers`

**Acknowledgements:**
This work would not have been possible without my fantastic domain collaborators, and the optimization and machine learning wizards in my department.
This work is supported in part by the Andrew Mellon Foundation through the "Visualizing English Print" project. This work is supported in part by NSF Awards IIS-1162037, CMMI-0941013, and DRL-1247262.


One journal cover image leads to (at least) three challenges

The axes are meaningless!
Explainers – crafted projections
VAST 2013

Can people interpret this?
Perception of average value in scatterplots
InfoVis 2013

The scatterplot has too many points!
Splatterplots – scalable scatterplots
TVCG 2013

## What can you do with too many points?

## Which Color Point is Higher on Average?

Gleicher, M., Correll, M., Nothelfer, C. and Franconeri, C. "Perception of Average Value in Multiclass Scatterplots." InfoVis 2013

## How did you do that?

## Visual Aggregation

People can extract summary statistics

Which Ones?
Efficiently?
Accurately?
How?

What can we do with it?
Why should we use it?

## Visual Aggregation

**Empirical Understanding**

Averages in Time Series
Correll, et al. CHI 2012

Tagged Text
Correll, et al. CHI 2013

Scatterplot Averages
Gleicher, et al. InfoVis 2013

Other statistics in Time Series
Albers, et al. CHI 2014(*)

**Practical Application**

Sequence Surveyor (Genetics)
Albers, et al. InfoVis 2011

LayerCake (Virus mutations)
Correll, et al. BioVis 2011

Molecular Surface Experiments
Sarikaya, et al. (in prep)

Decision Making
Correll, et al. (in prep)

## Ask the Turkers!

Crowdsource participants on Amazon's Mechanical Turk service

Careful Design to get valid results

Measure accuracy not speed

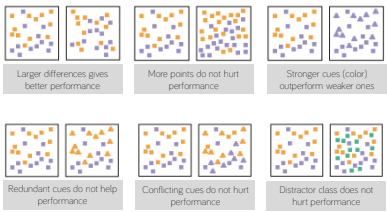
More like real tasks

Adjust hardness – not time allowed

## Scatterplot experiments

- Between subjects study
  - Series of experiments (one per condition)
  - 32 participants per condition
  - Established relations between conditions
- Within subjects study (repeated measures)
  - Reconfirm consults
  - 32 participants per experiment
  - Run pairs of conditions for key results
- Other experiments have similar designs

## Key Results

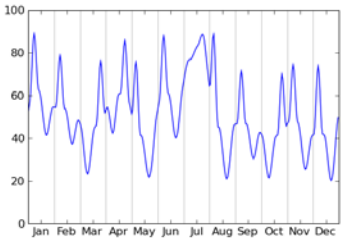


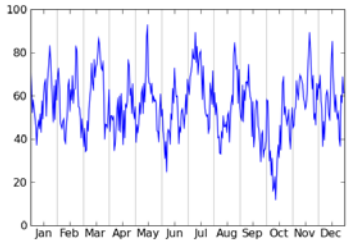

Gleicher, M., Correll, M., Nothelfer, C. and Franconeri, C. "Perception of Average Value in Multiclass Scatterplots." InfoVis 2013

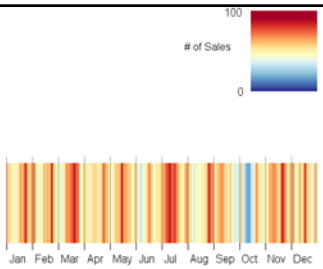
## Time Series?

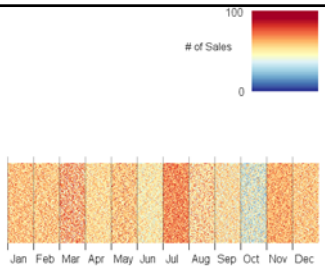
Correll, Albers, Franconeri, Gleicher. Comparing Averages in Time Series Data. CHI 2012.

## Which month has the highest average?



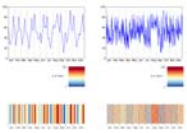

## Conditions

Linegraphs:

Regular or 1D permuted

Colorfields:

Regular or woven

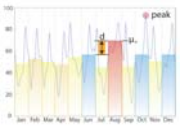


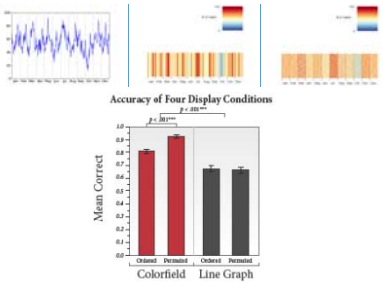
## Conditions

Noise of signal

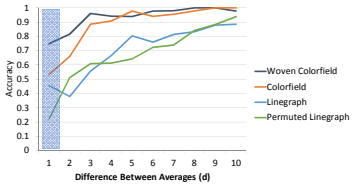
Gap between winner and distractor

Number of distractor months



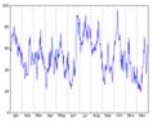




## Results

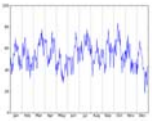



## Conditions

d = 10



d = 2



## What besides averages?

Albers, Correll, Gleicher. Task-Driven Design Variables for Time Series Visualization. (under review)

## Things You Might Care About

1. **Maxima:** Which month had the day with the highest sales for the year?
2. **Minima:** Which month had the day with the lowest sales for the year?
3. **Range:** Which month had the largest range of values?
4. **Average:** Which month had the highest average sales for the year?
5. **Spread:** Look at the average sales from each month. Which month had the sales which were the most spread out from their monthly average?
6. **Outliers:** Which month had the most unusual (outlier) sales days?

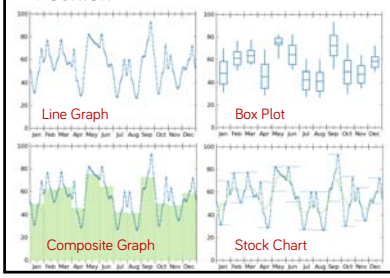
## Things That Might Matter

What **visual variable** do we use to encode the data? Position, color, size, alpha, …?

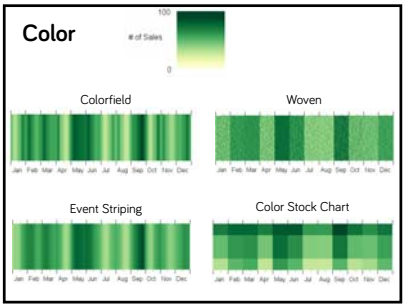
What **derived data** do we want to explicitly show? Extrema, averages, medians, modes, quartiles, …?

**How** should we compute this derived data? Discretely, continuously, simplified, complex, …?

## Position




## Color



## Results

| Encoding | Maxima | Minima | Range | Average | Spread | Outliers |
|---|---|---|---|---|---|---|
| Line graph | 87.5% | 78.9% | 74.2% | 27.7% | 49.8% | 36.7% |
| Modified Stock Chart | 68.3% | 86.1% | 89.8% | 56.3% | 26.7% | 54.0% |
| Box Plot | 75.0% | 83.8% | 86.5% | 68.8% | 85.9% | X |
| Composite Graph | 93.8% | 88.3% | 77.0% | 49.9% | 53.8% | 33.6% |
| Colorfield | 56.4% | 56.4% | 48.8% | 40.5% | 57.8% | 31.3% |
| Color Stock Chart | 68.9% | 75.4% | 68.8% | 79.3% | X | X |
| Woven Colorfield | 18.8% | 18.8% | 6.3% | 18.9% | 18.9% | 6.3% |
| Event Striping | 31.7% | 78.4% | 68.8% | 12.5% | 43.9% | 68.8% |

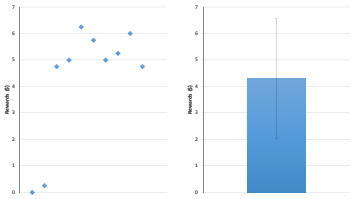
## Why Visual Aggregation?
### Why not just give them the answer?

1. You may not know what the viewer wants
2. Some aggregate properties might be complicated
3. You can't show all properties

4. It gives the viewer more information
5. Doing "work" might force them to think about things

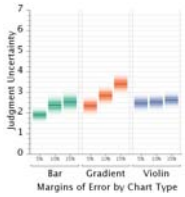
## Can better visualizations lead to better decision making?

## Should you take the bet?




Work with Michael Correll, in progress

## Results




## Quantity Estimation in Visualizations of Tagged Text

Correll, Alexander, Gleicher. CHI 2013

## More blue or purple?

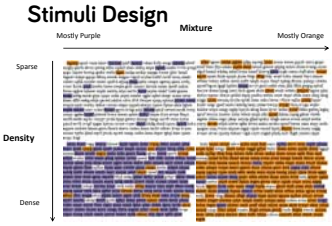
"I don't think they play at all fairly," Alice began, in rather a complaining tone, "and they all quarrel so dreadfully one can't hear one's-self speak—and they don't seem to have any rules in particular; at least, if there are, nobody attends to them—and you've no idea how confusing it is all the things being alive; for instance, there's the arch I've got to go through next walking about at the other end of the ground—and I should have croqueted the Queen's hedgehog just now, only it ran away when it saw mine coming!"

## Contributions

**Empirical analysis** of comparison in tagged text.

**Design considerations** to improve comparison.

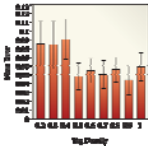
## Stimuli Design



## Results

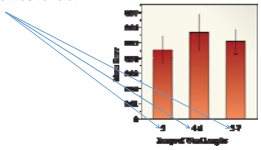
## Density

9 samples of density, random mixtures



## Word Length

3 uniform distributions of word length

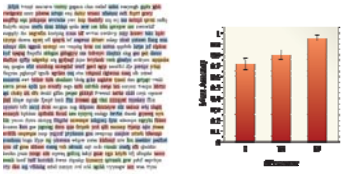


## Multi-category Comparison

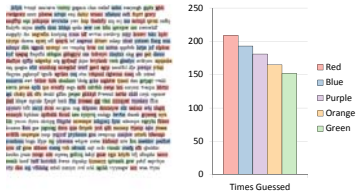
Which color is most common?

## d parameter

## d parameter



## Color bias




## Asymmetric Area

Which type of word is most common?

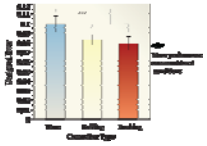
## Area Corrections

Padding

T r a c k i n g

## Area Corrections

## Results



to be , or not to be ; that is the question ;
whether ' tis nobler in the mind to suffer
the slings and arrows of outrageous fortune ,
or to take arms against a sea of troubles ,
and by opposing end them ? to die ; to sleep ;
no more ; and by a sleep to say we end
the heart-ache and the thousand natural shocks
that flesh is heir to , ' tis a consummation
devoutly to be wish'd . to die , to sleep ;
to sleep ; perchance to dream ; ay , there's the rub ;
for in that sleep of death what dreams may come
when we have shuffled off this mortal coil ,
must give us pause ; there's the respect
that makes calamity of so long life ;
for who would bear the whips and scorns of time ,
the oppressor's wrong , the proud man's contumely ,

15 "Clusters" (115 LATs)

- Subjective_Register
- Emotion
- Description
- Institutional_Register
- Academic_Register
- Future
- Past
- Personal_Relations
- Reasoning
- Interactive
- Elaboration
- Reporting
- Directing
- Narrative
- Character

Not in dictionary

Not Tagged

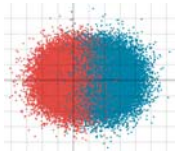
One journal cover image leads to (at least) three challenges

The axes are meaningless!
Explainers – crafted projections
VAST 2013

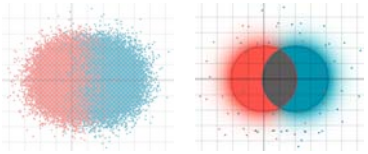
Can people interpret this?
Perception of average value in scatterplots
InfoVis 2013

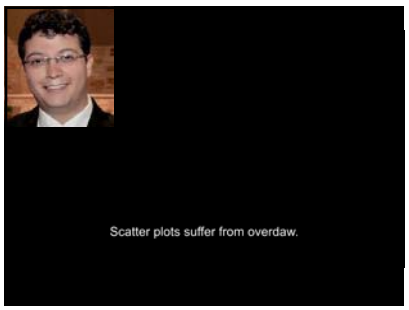
The scatterplot has too many points!
Splatterplots – scalable scatterplots
TVCG 2013

## Problem: Scatterplot with too many points!

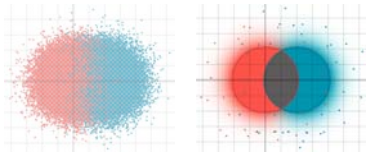


## Solution: Splatterplot

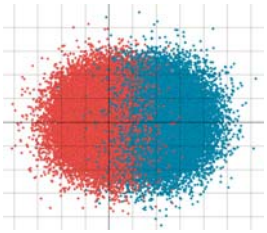


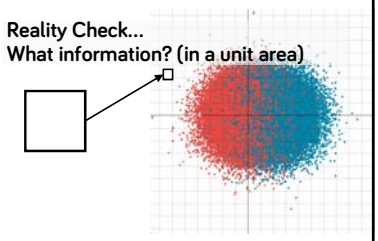



## Solution: Splatterplot

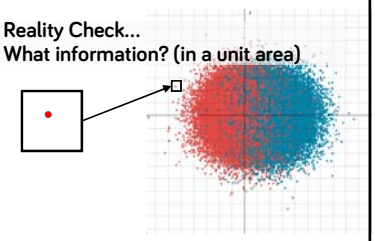


what if you have lots of points?

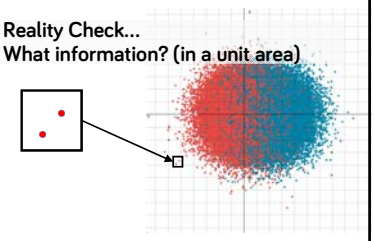




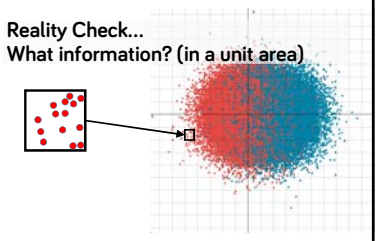
Reality Check…
What information? (in a unit area)



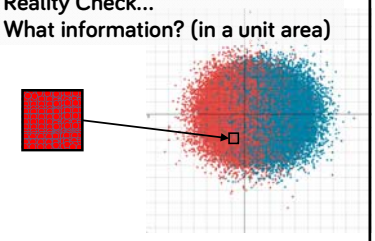
Reality Check…
What information? (in a unit area)



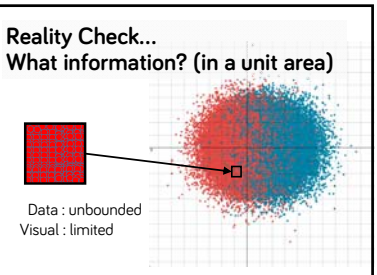
Reality Check…
What information? (in a unit area)



Reality Check…
What information? (in a unit area)



Reality Check…
What information? (in a unit area)



Reality Check…
What information? (in a unit area)

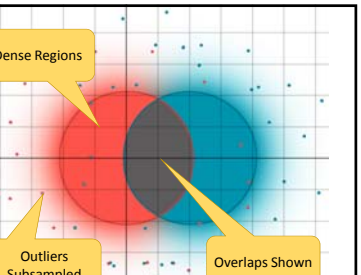
Data : unbounded
Visual : limited

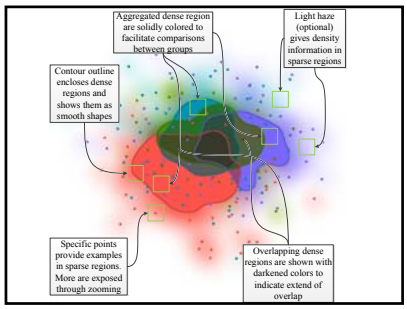
### Bounded Information Density

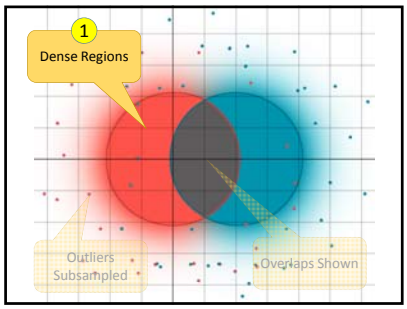
In a Unit of Area:
Amount of data is **unbounded**
Amount you can see is **limited**

Need to **limit** the amount shown

**Choose** what to display by **abstracting** the data





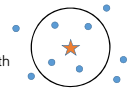




## Kernel Density Estimation (KDE)

Count how many points near every position

Weight by distance

Size of kernel (circle) is the bandwidth



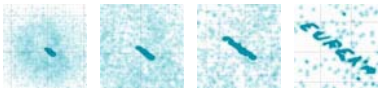
Creates smooth fields

## Screen Space KDE

Parameters based on perceptual properties

Independent of data

Does the right thing when you zoom

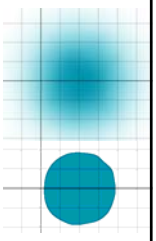


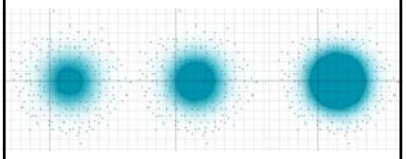
## Discrete dense regions

Threshold

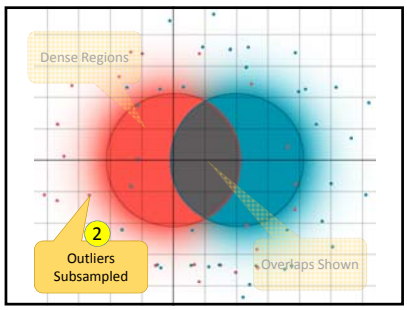
**Why? (single set case)**
Dynamic range of density may be high
and hard to encode
At some point, it's just "dense"
Crisp boundaries are better visually

**Information is thrown away!**



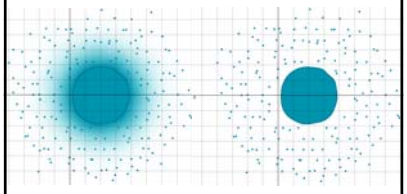


**Information is thrown away!**
Interactive control of threshold
Encode sparse regions differently




## Subsample sparse regions

## To Haze or not to Haze?

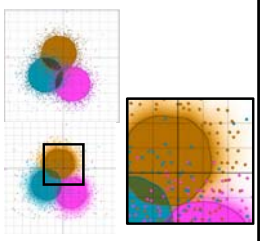


## Edges

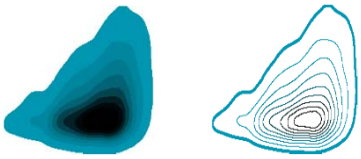
Strokes

Clear Clutter

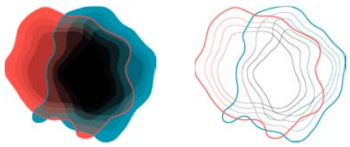
Both require distance to region

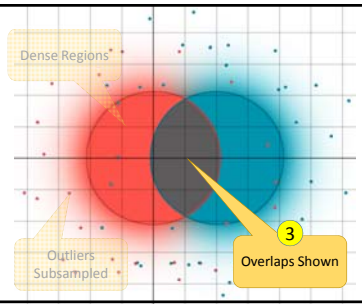


## Contours?



## Complicated with multiple groups



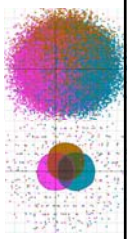



## Multiple Groups

Compute densities independently

Color per group

Pick distinctive colors



## Colors for combinations

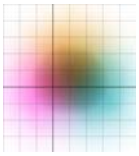
**Multi-variate color encoding?**

Map $R^n$ to a color

## Colors for combinations

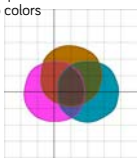
**Multi-variate color encoding?**

Map $R^n$ to a color



**Colors for set combinations**

Map $2^n$ set combinations to colors



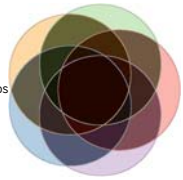
## Color Blending

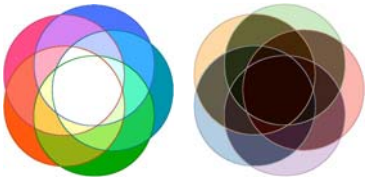
Encode sets with color

Hue = set

Lightness = number of overlaps

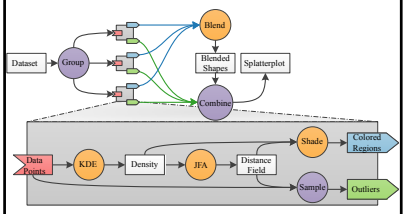
See evaluation in paper





# Implementation

Interactivity is critical!

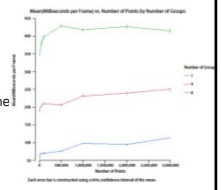



## Performance: Use the GPU

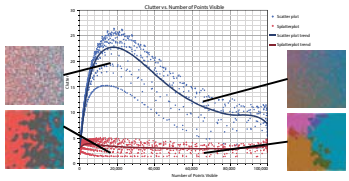
Draw points
Filter (convolution) for KDE
Jump Flood for distances

Render each set and combine

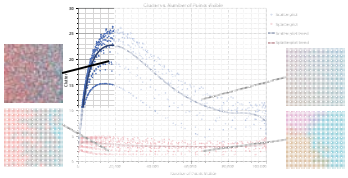
Lots of points – fast
Lots of groups – less fast



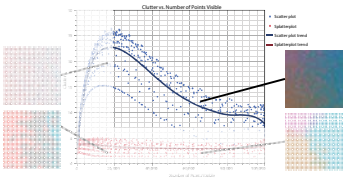
## Quantifying Perceptual Density?



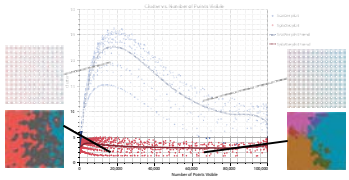
## Quantifying Perceptual Density



## Quantifying Perceptual Density



## Quantifying Perceptual Density

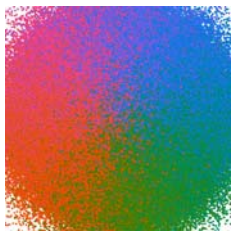


## Other ideas?

There is plenty of "related work" in research in practice

**Key Novelties in Splatterplots**

Choose abstractions to understand set relationships
Screen space density estimates
Dual Encodings



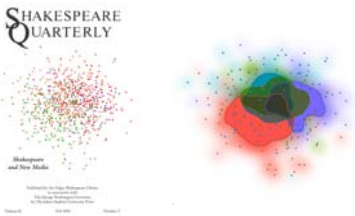
subsample?

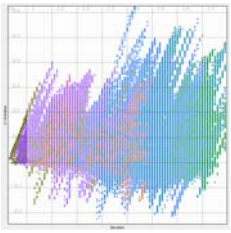
histograms and KDEs

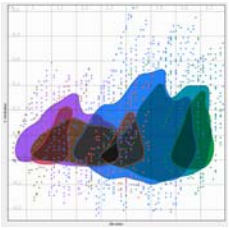
Splatterplot!

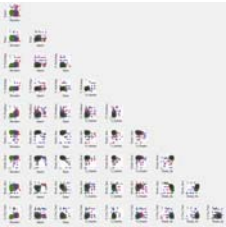
Real (or realistic) Examples

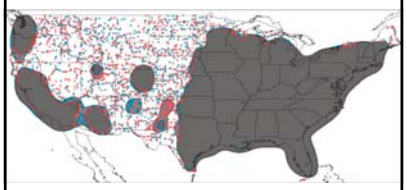
The synthetic data is pretty but...









## More to do!

| Theory | Practice |
|---|---|
| Understand Visual Density | WebGL implementation |
| Consider other tradeoffs | Massive Data Handling |
| Other Types of Data | Evaluation (see InfoVis paper) |
| 3D (volumes) | Non-GPU version for my laptop ☹ |

## Splatterplots

Scalable Display of Scatter Data

Bounded visual complexity

Screen space density estimation

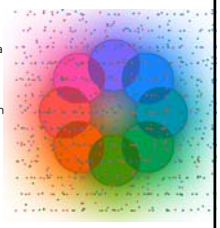
Dual encodings

GPU Implementation

**Acknowledgements**

This work is supported in part by the Andrew Mellon Foundation through the "Visualizing English Print" project

This work is supported in part by NSF Awards IIS-1162037, CMMI-0941013, and DRL 1247262




**One journal cover image leads to (at least) three challenges**

The axes are meaningless!

Explainers – crafted projections
VAST 2013

Can people interpret this?

Perception of average value in scatterplots
InfoVis 2013

The scatterplot has too many points!

Splatterplots – scalable scatterplots
TVCG 2013

## Data Visualization

Inspiration comes from a number of directions:

Seeing

Scatterplots

Shakespeare

## Thanks!

To you for listening.
To the organizers for inviting me
To my students and collaborators.
To the NSF and Mellon Foundation for funding.

# Seeing, Scatterplots and Shakespeare

**Michael Gleicher**
University of Wisconsin Madison
(on sabbatical at INRIA Rhone-Alpes)
`gleicher@cs.wisc.edu`