Acknowledgements

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

Students:
Danielle Albers, Eric Alexander, Michael Correll, Adrian Mayorga, Alper Sarikaya, …

Collaborators:
Steve Franconeri, Jonathan Hope, Robin Valenza, Mike Witmore, …

This research is 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
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?
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
Traditional Scholarship: Close Reading

Scalable? Thousands of books?
Traditional Scholarship: Close Reading Scalable? Thousands of books?
Seeing Shakespeare: Scalable Scholarship
Study Literature without Reading?
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
Texts as Data?

Need to turn books into numbers
Text tagging

Text → Vector

Counts:
- Count of word Category 1: 4
- Count of word Category 2: 0, 3.6, 4.7
- Count of word Category 3: 0, 3.4, ...

Diagram shows a book representing text input and an arrow pointing to a vector with calculated counts for different word categories.
Just counting

Words (phrases) have a type (tag)

Docuscope
Simple matching against a dictionary
Hand-built dictionaries

100-115 Categories, 12-17 Clusters (groups)
Unrecognized words by decade

1080 books (random sample)

Johnson’s Dictionary
1756
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.

Serendipity: Exploring Topic Models
Just counting

Words (phrases) have a type (tag)

**Docuscope**

Simple matching against a dictionary
Hand-built dictionaries

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

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
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
Domain Specific Tools

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?
What makes a comedy a comedy?

Why are Dickens’ novels different?

What cities are most like Paris?

What clusters on principal component one?

What metric imparts a significant distinction?
What makes a comedy a comedy?

Why are Dickens’ novels different?

What cities are most like Paris?

What clusters on principal component one?

What metric imparts a significant distinction?
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

- Paris: 4,8,1,3,...
- New York: 4,8,1,3,...
- Beijing: 7,3,2,7,...
- Sydney: 9,2,6,4,...
- Jakarta: 3,2,5,1,...
- Atlanta: 5,2,1,7,...
- Tokyo: 9,2,6,4,...
- Munich: 7,3,2,7,...
- San Jose: 5,2,1,7,...
- Boston: 3,2,5,1,...
- London: 5,2,1,7,...
- New York: 4,8,1,3,...
- Tokyo: 9,2,6,4,...
- Munich: 7,3,2,7,...
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

Explain
Relationships with the data connect concepts and variables

Rankings
Outliers
Extrema
Exemplars
Similarities

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

**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

\[ f(V) = V^{39} - V^{42} \]

M = Predicted Future

I = Inclusiveness
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, …)
\[ A + B \text{ vs. } e^{-\omega k(A,B)} \]

Parsimony (few variables)
\[ A + B \text{ vs. } W + X + Y + Z \]

Simple Coefficients (small integers)
\[ A - 2B \text{ vs. } 1.235 A - 4.327 B \]

Familiar Variables
\[ A + B \text{ vs. } Q + W \]

* Based on intuitions – really need some science here.
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
Does this hold over all Early Modern Drama?
How to find functions?

Optimization problem
Minimize
amount “wrong”
“Cost” of function

Support Vector Machine (SVM)
### How to Implement Explainers?

<table>
<thead>
<tr>
<th>Fancy Math</th>
<th>Brute Force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encode tradeoffs into the SVM</td>
<td>Sample space of variable sets</td>
</tr>
<tr>
<td>Solve for the best tradeoffs</td>
<td>Solve an SVM for each</td>
</tr>
<tr>
<td>Adjust parameters to tune</td>
<td>Sort and filter</td>
</tr>
<tr>
<td>Solve one big optimization problem</td>
<td>Solve many small optimization problems</td>
</tr>
<tr>
<td></td>
<td>Generates a diverse and interesting exploration of tradeoffs</td>
</tr>
</tbody>
</table>

*Not a standard SVM, so needs a slow and finicky solver*
Isn’t this just...

Some prior approaches to help situate our work
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

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.
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-094103, 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)
- Molecular Surface Experiments
  - Sarikaya, et al. EuroVis 2014
- Decision Making
  - Correll, et al. (submitted)
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**

- Larger differences gives better performance.
- More points do not hurt performance.
- Stronger cues (color) outperform weaker ones.
- Redundant cues do not help performance.
- Conflicting cues do not hurt performance.
- Distractor class does not hurt performance.

Gleicher, M., Correll, M., Nothelfer, C. and Franconeri, C. “Perception of Average Value in Multiclass Scatterplots.” InfoVis 2013
Time Series?

Which month has the highest average?
Conditions

Linegraphs:
Regular or 1D permuted

Colorfields:
Regular or woven
Accuracy of Four Display Conditions

\[ p < 0.001^{**} \]

\[ p < 0.001^{***} \]

<table>
<thead>
<tr>
<th>Colorfield</th>
<th>Line Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered</td>
<td>Permutated</td>
</tr>
</tbody>
</table>

Mean Correct

\[ 0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.6 \quad 0.7 \quad 0.8 \quad 0.9 \quad 1.0 \]
Results

The graph shows the accuracy as a function of the difference between averages (d). The graph includes four lines:

- Woven Colorfield (dark blue)
- Colorfield (orange)
- Linegraph (light blue)
- Permuted Linegraph (green)

The x-axis represents the difference between averages, while the y-axis shows accuracy. The shaded region highlights a particular range on the x-axis.
Conditions

\[ d = 10 \]  \quad \text{and} \quad  \[ d = 2 \]
What besides averages?

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

Line Graph

Box Plot

Composite Graph

Stock Chart
## Results

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Maxima</th>
<th>Minima</th>
<th>Range</th>
<th>Average</th>
<th>Spread</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line graph</td>
<td>87.5%</td>
<td>78.9%</td>
<td>74.2%</td>
<td>47.7%</td>
<td>48.8%</td>
<td>36.7%</td>
</tr>
<tr>
<td>Modified Stock Chart</td>
<td>88.7%</td>
<td>96.1%</td>
<td>91.8%</td>
<td>56.3%</td>
<td>39.7%</td>
<td>34.0%</td>
</tr>
<tr>
<td>Box Plot</td>
<td>75.0%</td>
<td>93.8%</td>
<td>88.5%</td>
<td>68.8%</td>
<td>85.0%</td>
<td>X</td>
</tr>
<tr>
<td>Composite Graph</td>
<td>93.0%</td>
<td>88.3%</td>
<td>77.0%</td>
<td>85.9%</td>
<td>53.8%</td>
<td>33.6%</td>
</tr>
<tr>
<td>Colorfield</td>
<td>59.4%</td>
<td>56.6%</td>
<td>48.8%</td>
<td>60.5%</td>
<td>57.8%</td>
<td>31.3%</td>
</tr>
<tr>
<td>Color Stock Chart</td>
<td>69.9%</td>
<td>73.4%</td>
<td>64.8%</td>
<td>70.3%</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Woven Colorfield</td>
<td>43.0%</td>
<td>45.7%</td>
<td>38.7%</td>
<td>77.7%</td>
<td>71.3%</td>
<td>23.0%</td>
</tr>
<tr>
<td>Event Striping</td>
<td>61.7%</td>
<td>59.4%</td>
<td>44.1%</td>
<td>52.3%</td>
<td>42.2%</td>
<td>66.8%</td>
</tr>
</tbody>
</table>
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
6. Uncertainty of perception maps to uncertainty in data
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
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
Dense Regions

Outliers Subsampled

Overlaps Shown
Specific points provide examples in sparse regions. More are exposed through zooming.

Aggregated dense region are solidly colored to facilitate comparisons between groups.

Contour outline encloses dense regions and shows them as smooth shapes.

Light haze (optional) gives density information in sparse regions.

Overlapping dense regions are shown with darkened colors to indicate extend of overlap.
Dense Regions

Outliers Subsampled

Overlaps Shown
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
Dense Regions

Outliers Subsampled

Overlaps Shown

2
Subsample sparse regions
To Haze or not to Haze?
Edges

Strokes

Clear Clutter

Both require distance to region
Contours?
Complicated with multiple groups
Dense Regions

Outliers Subsampled

Overlaps Shown

3
Multiple Groups

Compute densities independently

Color per group

Pick distinctive colors
Colors for combinations

Multi-variate color encoding?
Map $\mathbb{R}^n$ to a color
Colors for combinations

Multi-variate color encoding?
Map $\mathbb{R}^n$ to a color

Colors for set combinations
Map $2^\binom{n}{2}$ 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?

Clutter vs. Number of Points Visible

- Scatter plot
- Splatterplot
- Scatter plot trend
- Splatterplot trend

Number of Points Visible

Clutter

0 5 10 15 20 25 30

20,000 40,000 60,000 80,000 100,000
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!

<table>
<thead>
<tr>
<th>Theory</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand Visual Density</td>
<td>WebGL implementation</td>
</tr>
<tr>
<td>Consider other tradeoffs</td>
<td>Massive Data Handling</td>
</tr>
<tr>
<td>Other Types of Data</td>
<td>Evaluation (see InfoVis paper)</td>
</tr>
<tr>
<td>3D (volumes)</td>
<td>Non-GPU version for my laptop 😞</td>
</tr>
</tbody>
</table>
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-094103, 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

gleicher@cs.wisc.edu