Explainers: Expert Explorations with Crafted Projections

Michael Gleicher

University of Wisconsin – Madison

(on sabbatical at INRIA, Rhone-Alpes)

Warning: this presentation uses the Bariol font family that you have to pay for. I recommend it – it's cheap and beautiful

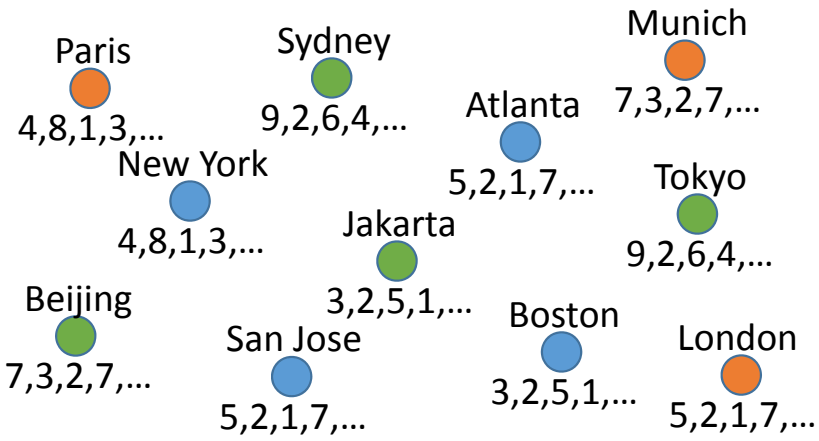
Explainers

An approach to explore high dimensional data:

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

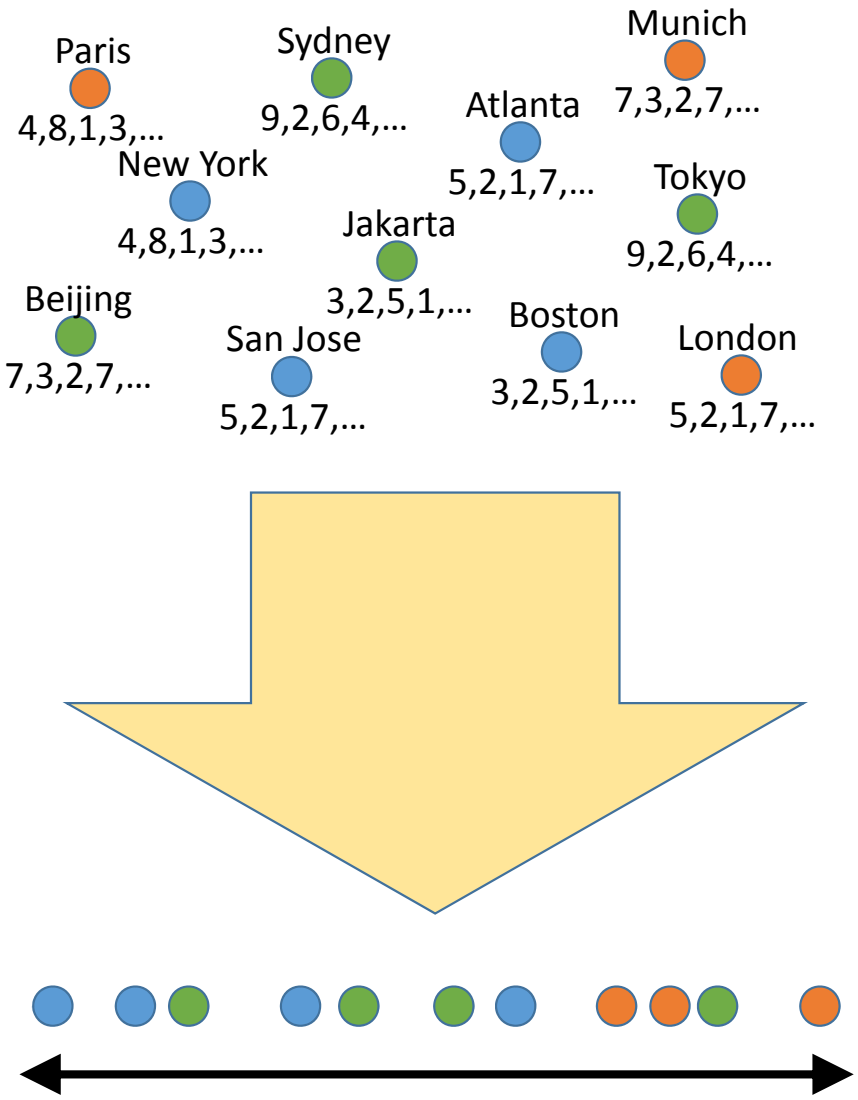
Explain **user**-defined concepts according to the data

Give the **user** control over tradeoffs



High Dimensional Data

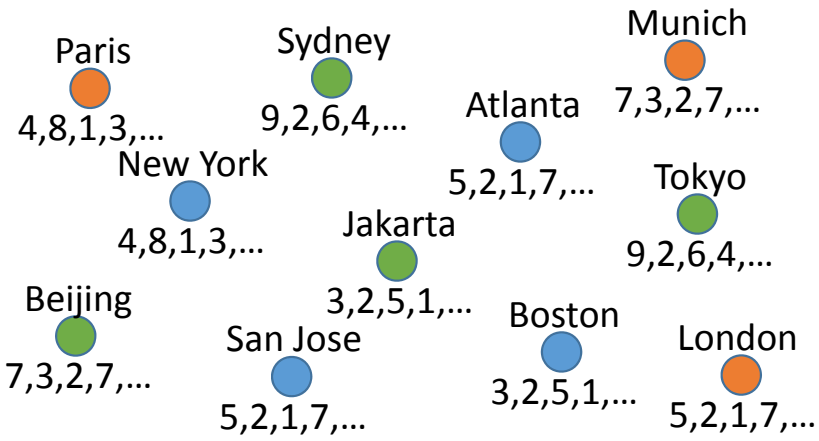
Objects have associated **Vectors**



Projections

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

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



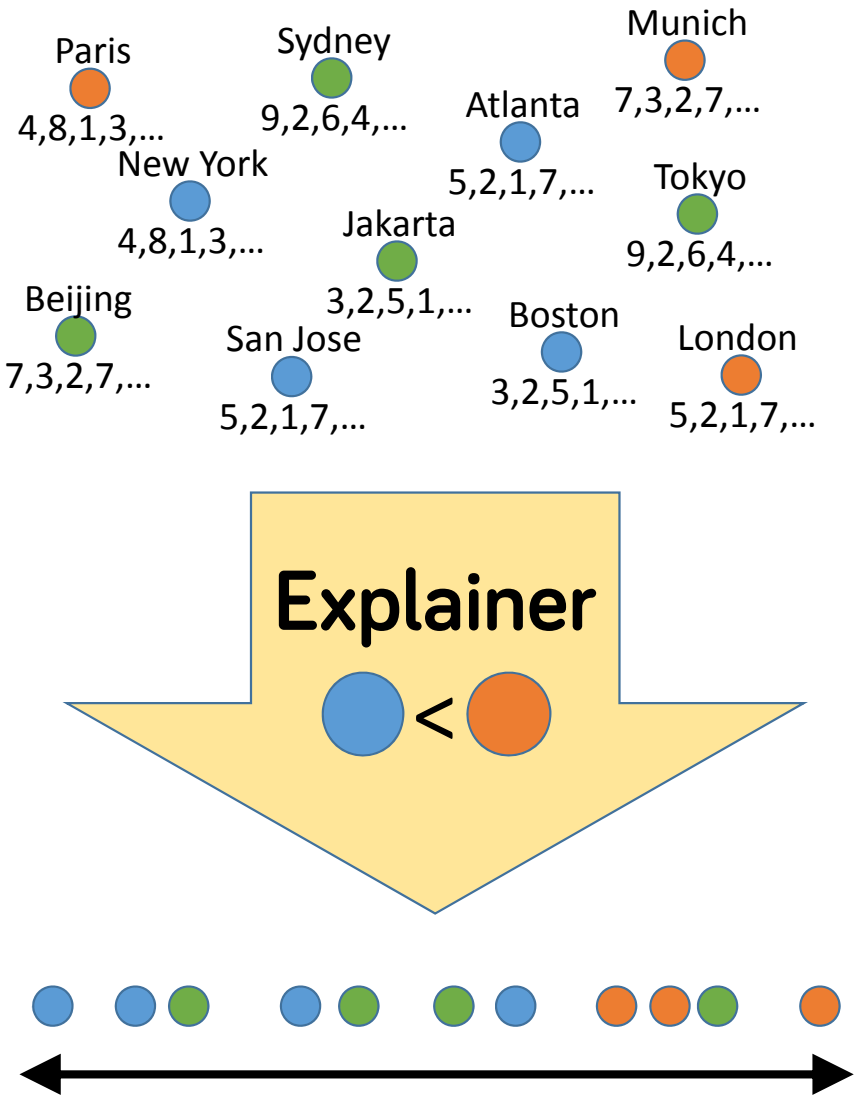
Specification

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?

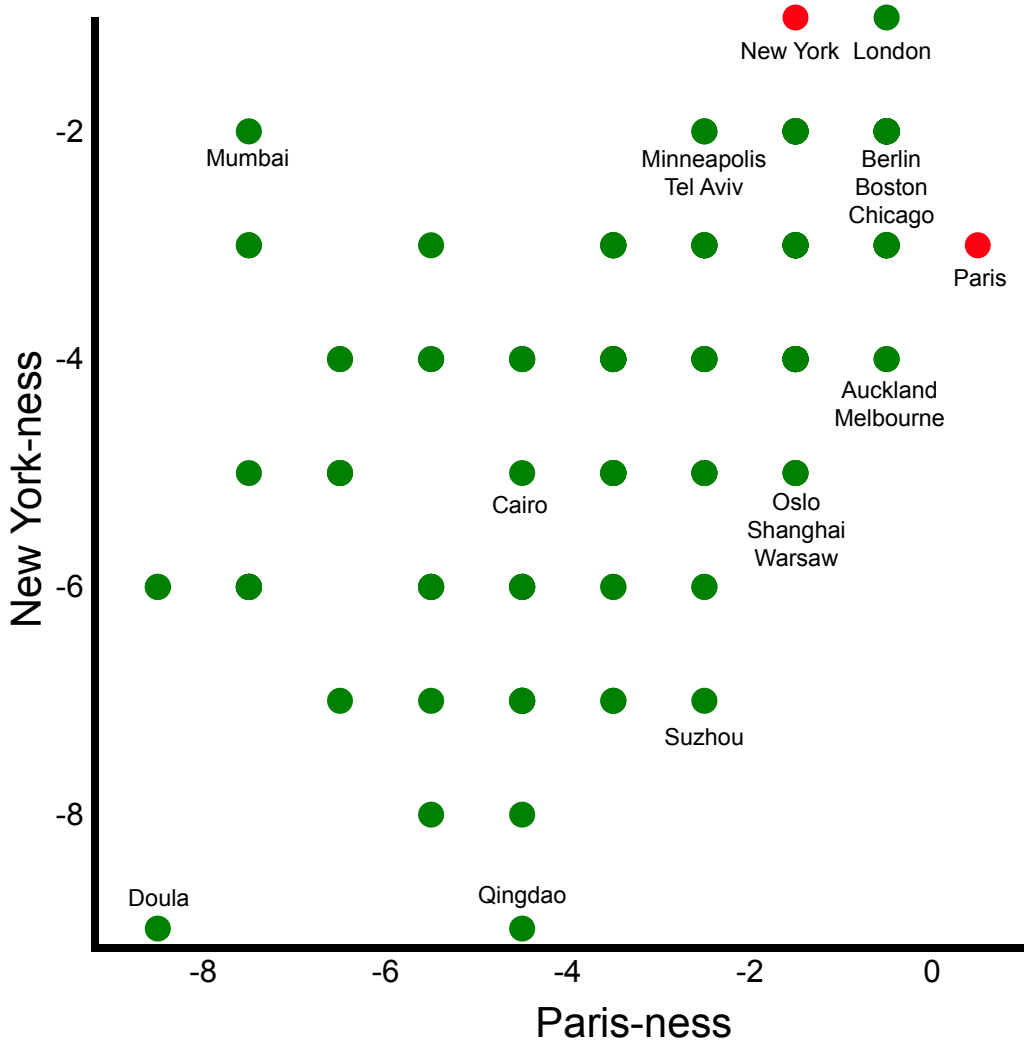
Understandability:

can the user interpret the mapping?

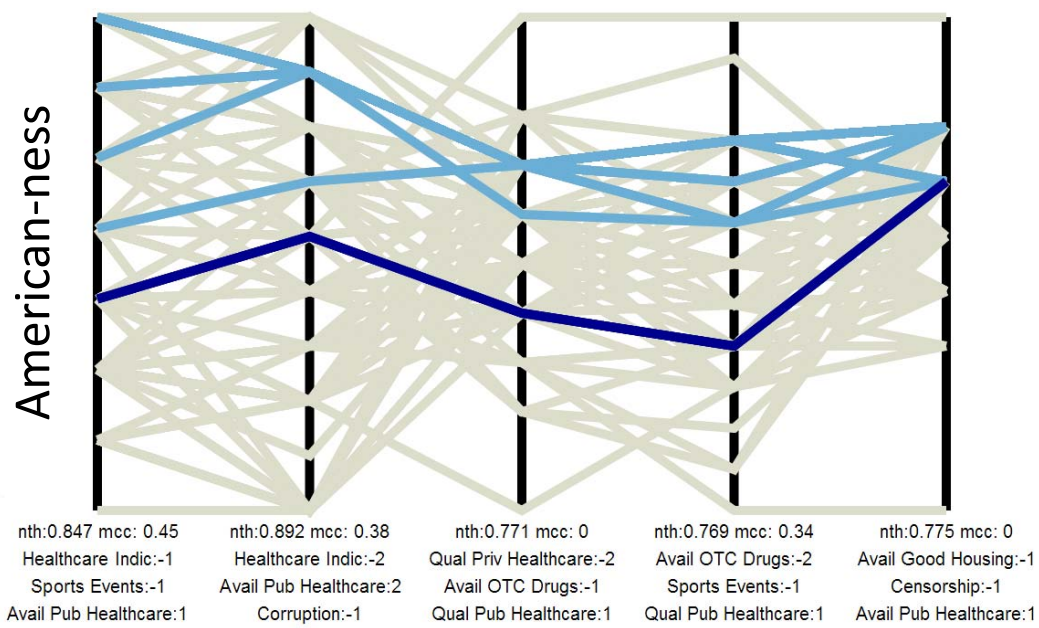
Diversity:

can we generate alternate mappings?

Paris-ness vs. New York-ness



5 views of American-ness



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?

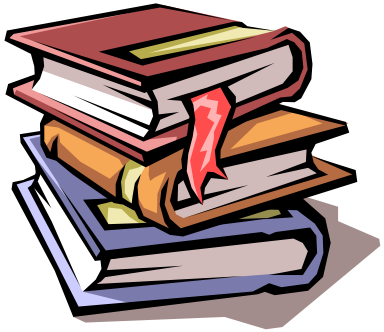
To make things concrete

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

What measurements?

Count words of each type

Words (phrases) have a type (tag)

Docuscope (from CMU)
Simple matching
Hand-built dictionaries

115 Categories

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 ,

genre

Categorization given by Shakespeare's contemporaries

Comedy **Tragedy** **History**

Category for plays written after that

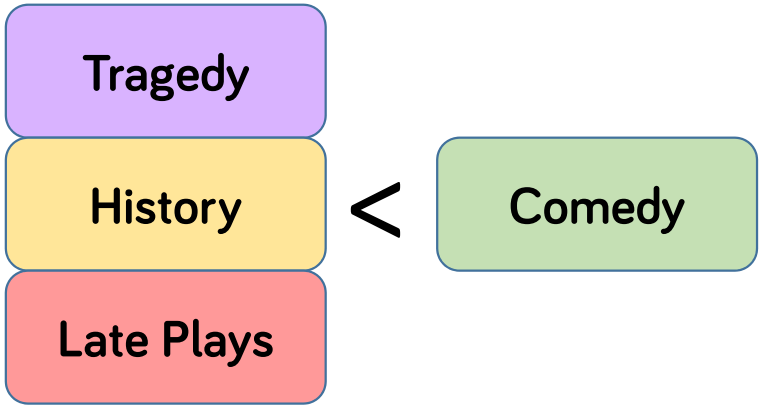
Late Plays

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



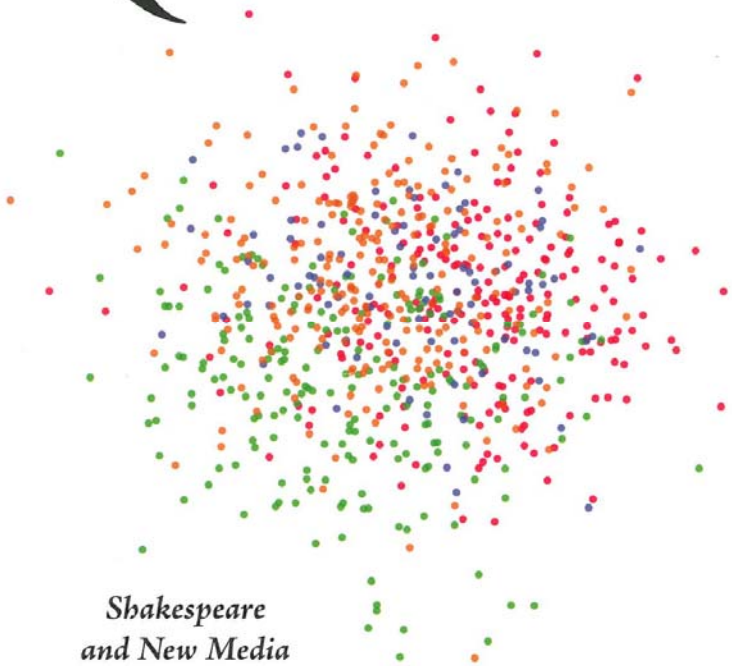
Organization:

What is most/least comedic?

Explanation:

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

SHAKESPEARE QUARTERLY



Shakespeare and New Media

Published for the Folger Shakespeare Library
in association with
The George Washington University
by The Johns Hopkins University Press

Volume 61 Fall 2010 Number 3

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)

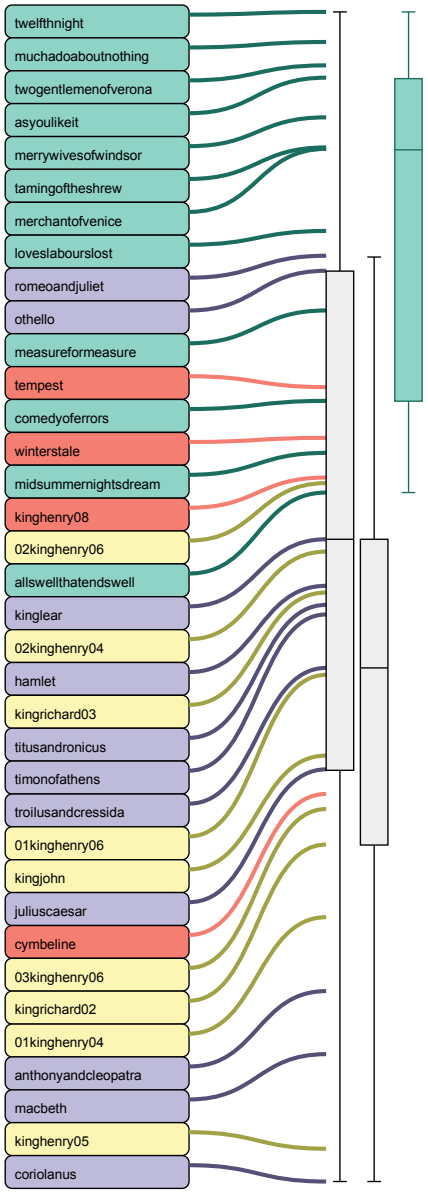
Choose f such that:

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

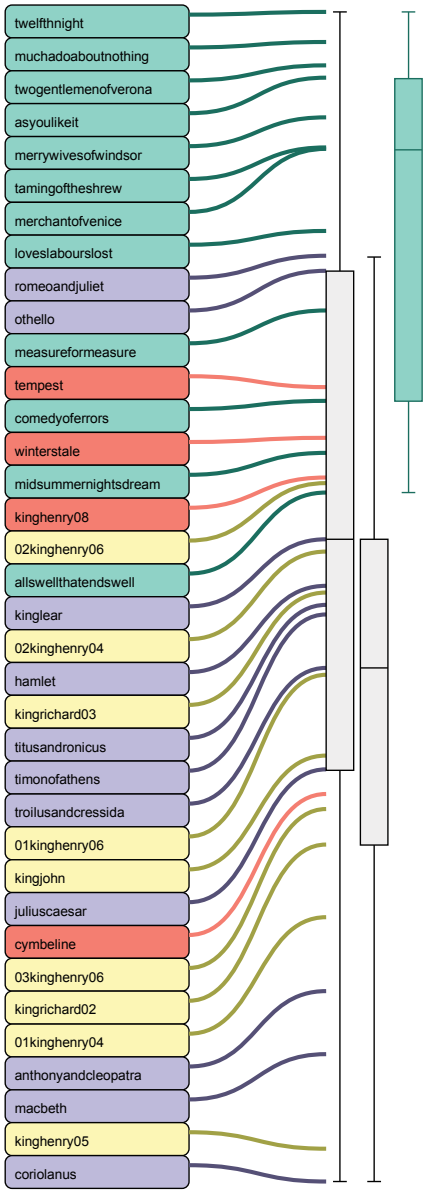
An Explainer

comedicness = M − I

$$f(V) = V[39] - V[42]$$



Understanding the Visualization*

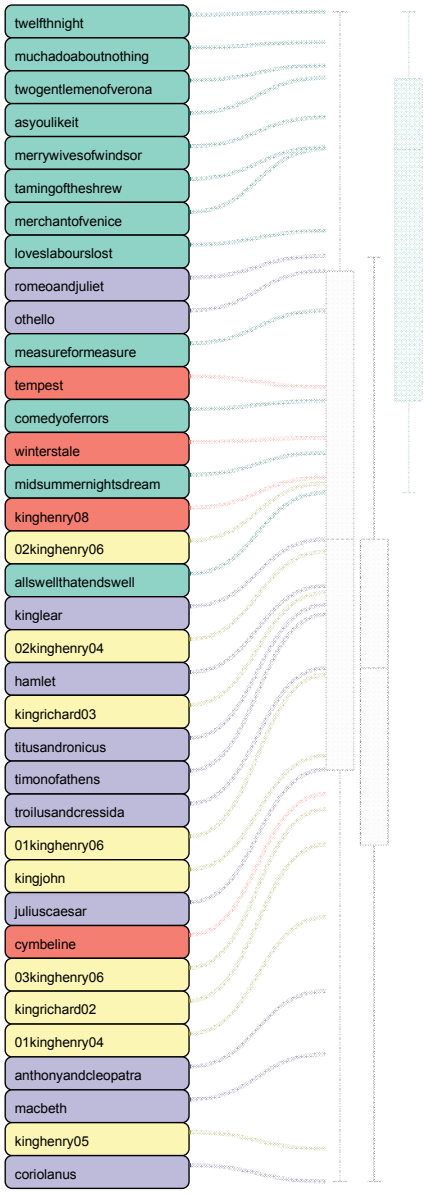


* The Visual Encoding is not a strong part. Suggestions are most welcomed!

Understanding the Visualization*

Objects in rank order

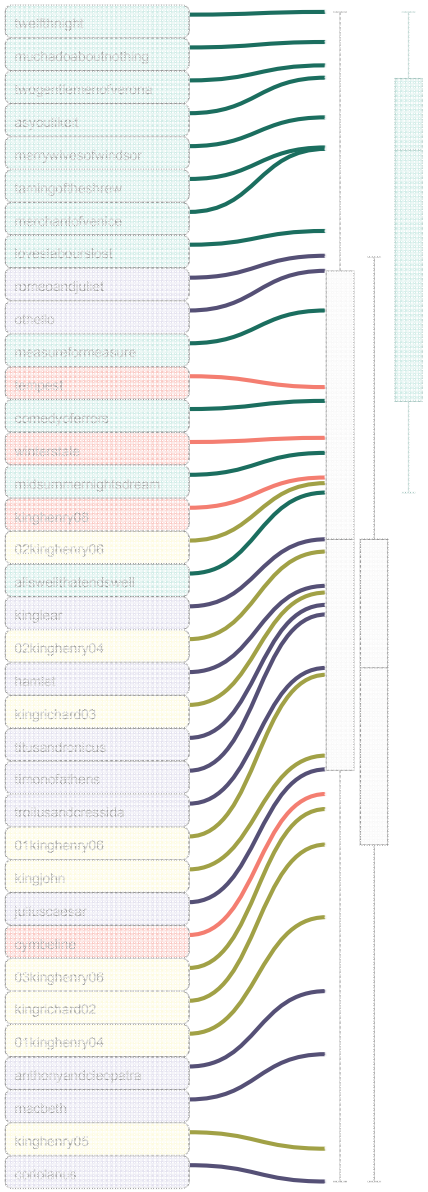
Color by specified class



* 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)

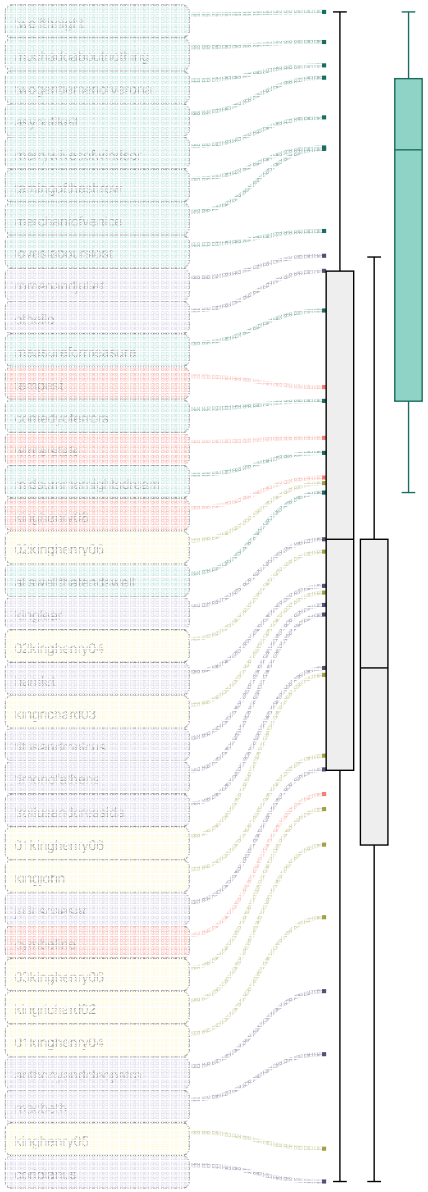


* The Visual Encoding is not a strong part. Suggestions are most welcomed!

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

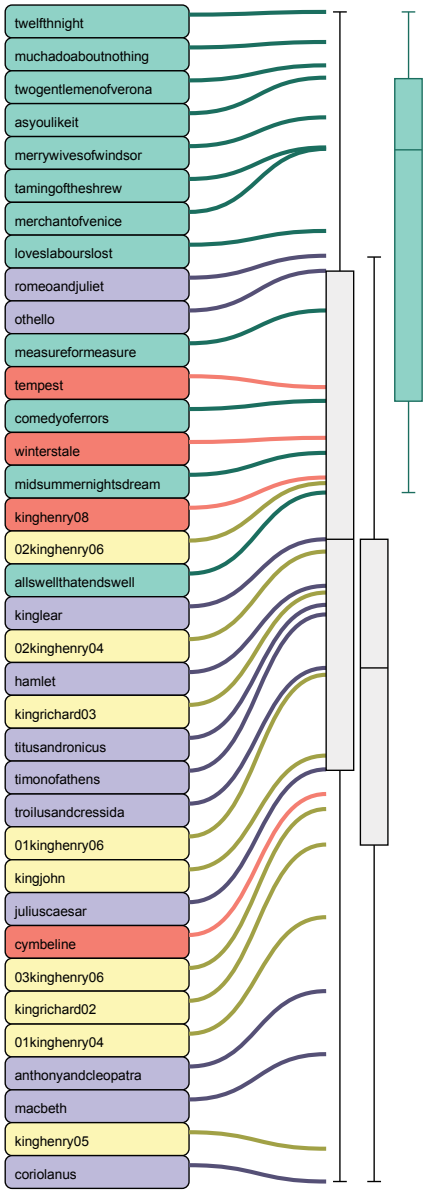
$$f(V) = V[39] - V[42]$$

Tradeoff:

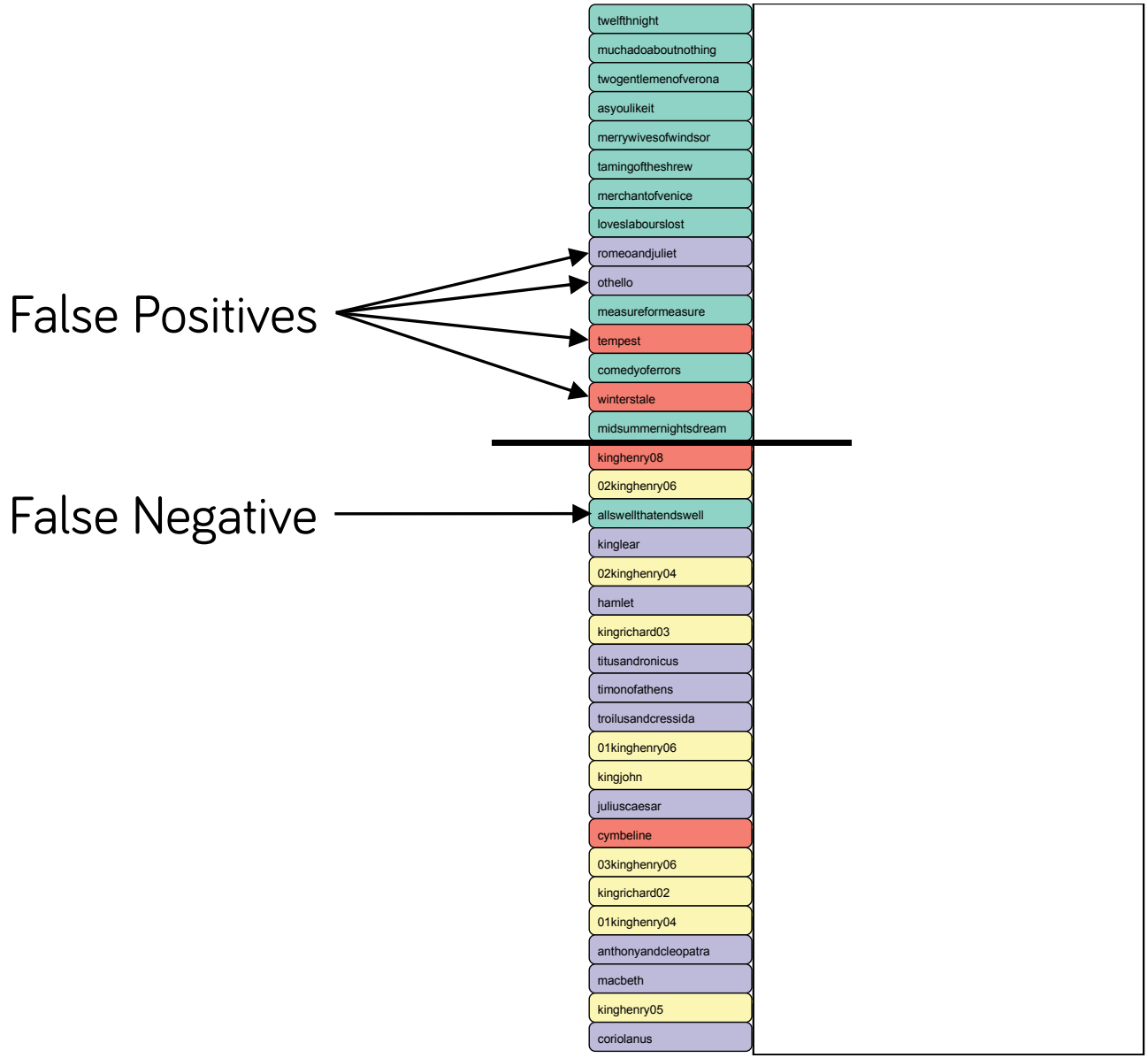
simple (linear, 2 variables, unit coefficients)

but

5 “wrong”



5 Wrong


False Positives
False Negative
twelfthnight
muchadoaboutnothing
twogentlemenofverona
asyoulikeit
merrywivesofwindsor
tamingoftheshrew
merchantofvenice
loveslabourslost
romeoandjuliet
othello
measureformeasure
tempest
comedyoferrors
winterstale
midsummernightsdream
kinghenry08
02kinghenry06
allswellthatendswell
kinglear
02kinghenry04
hamlet
kingrichard03
titusandronicus
timonofathens
troilusandcressida
01kinghenry06
kingjohn
juliuscaesar
cymbeline
03kinghenry06
kingrichard02
01kinghenry04
anthonyandcleopatra
macbeth
kinghenry05
coriolanus

Wrong?

Interesting Outliers

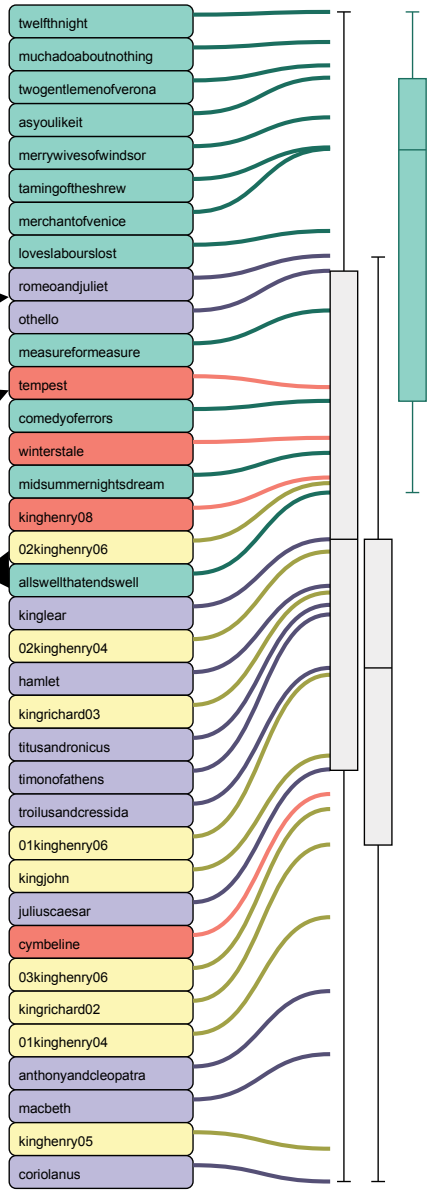
"Romeo and Juliet" is pretty comedic

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. $e^{-\omega k(A,B)}$

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:

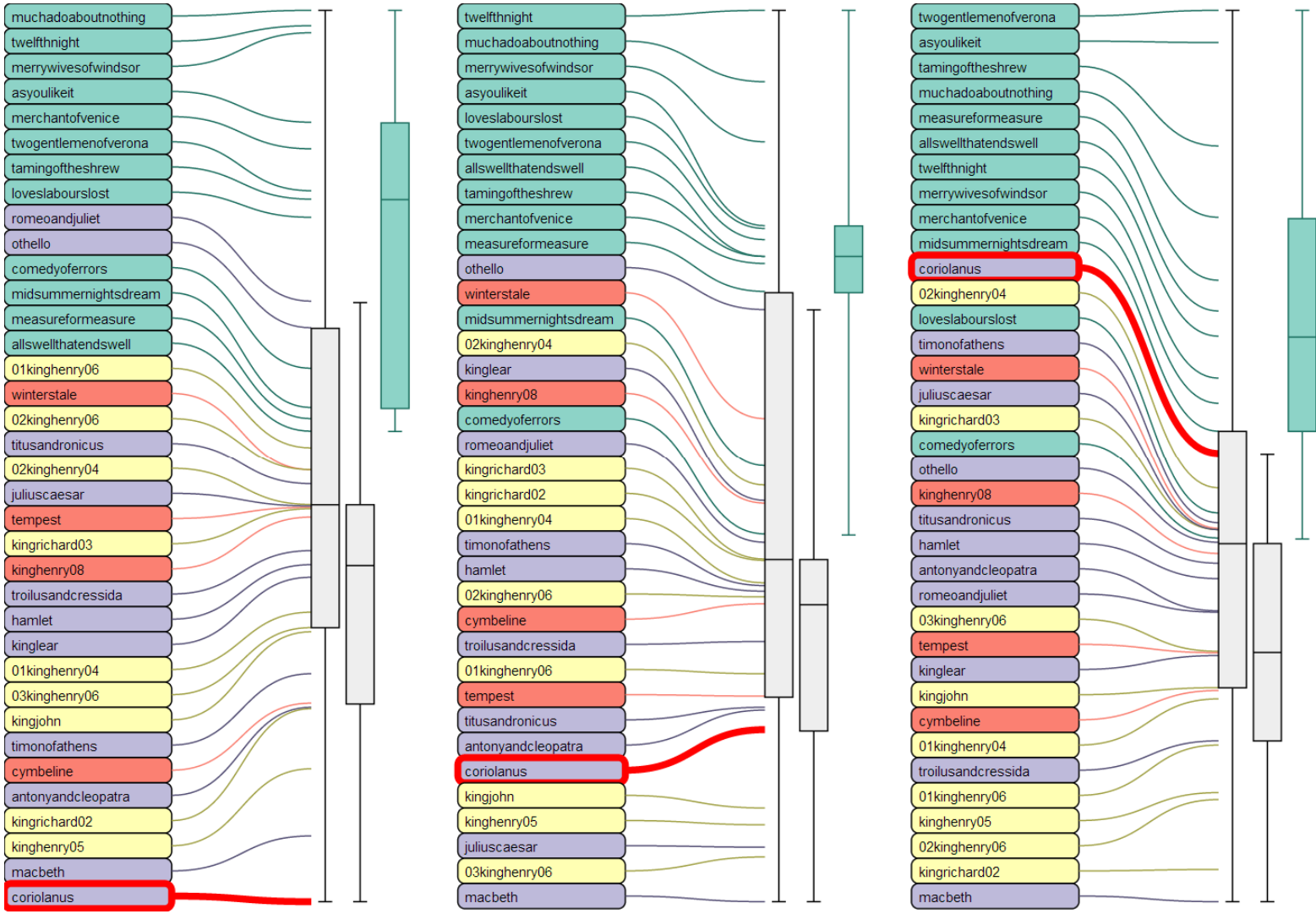
Correctness

Simplicity

Different:

Explanations

Orderings



How to find functions?

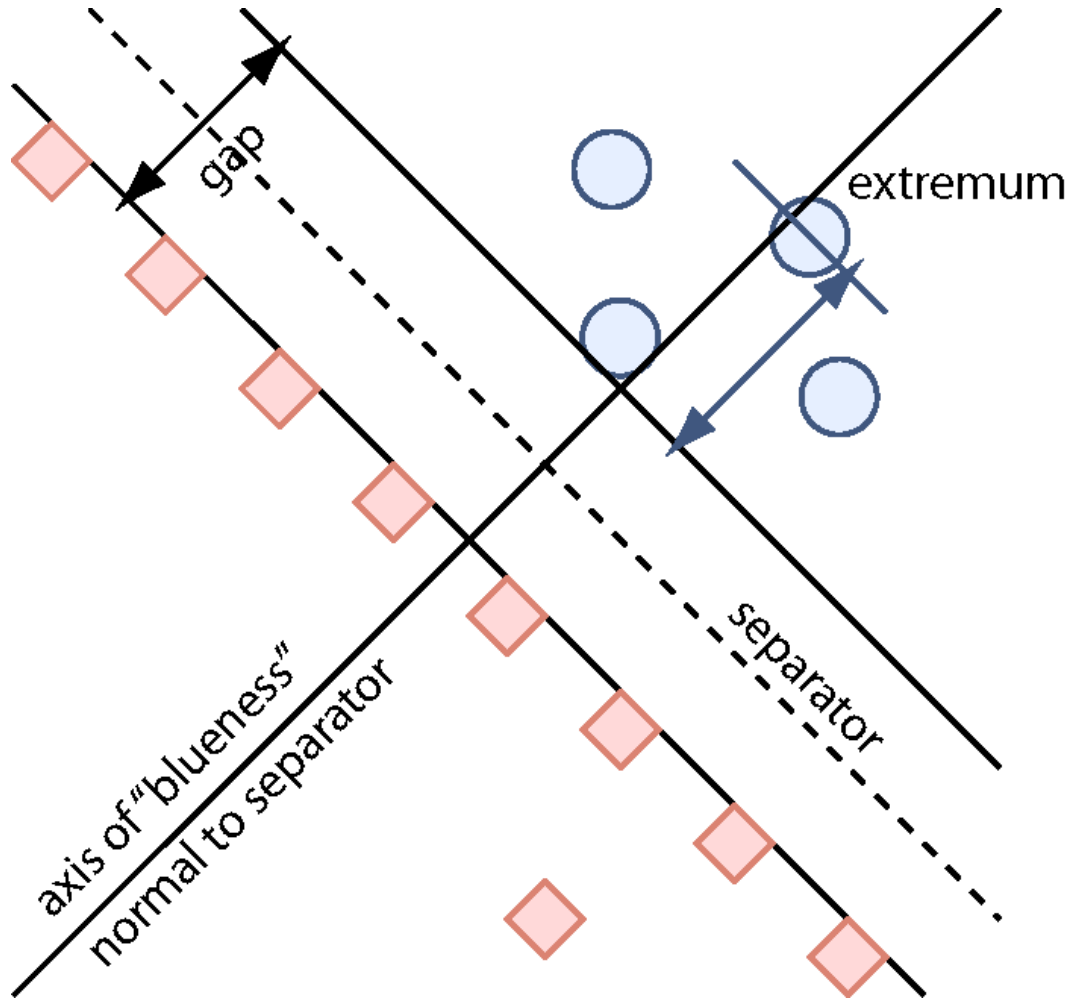
Optimization problem

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

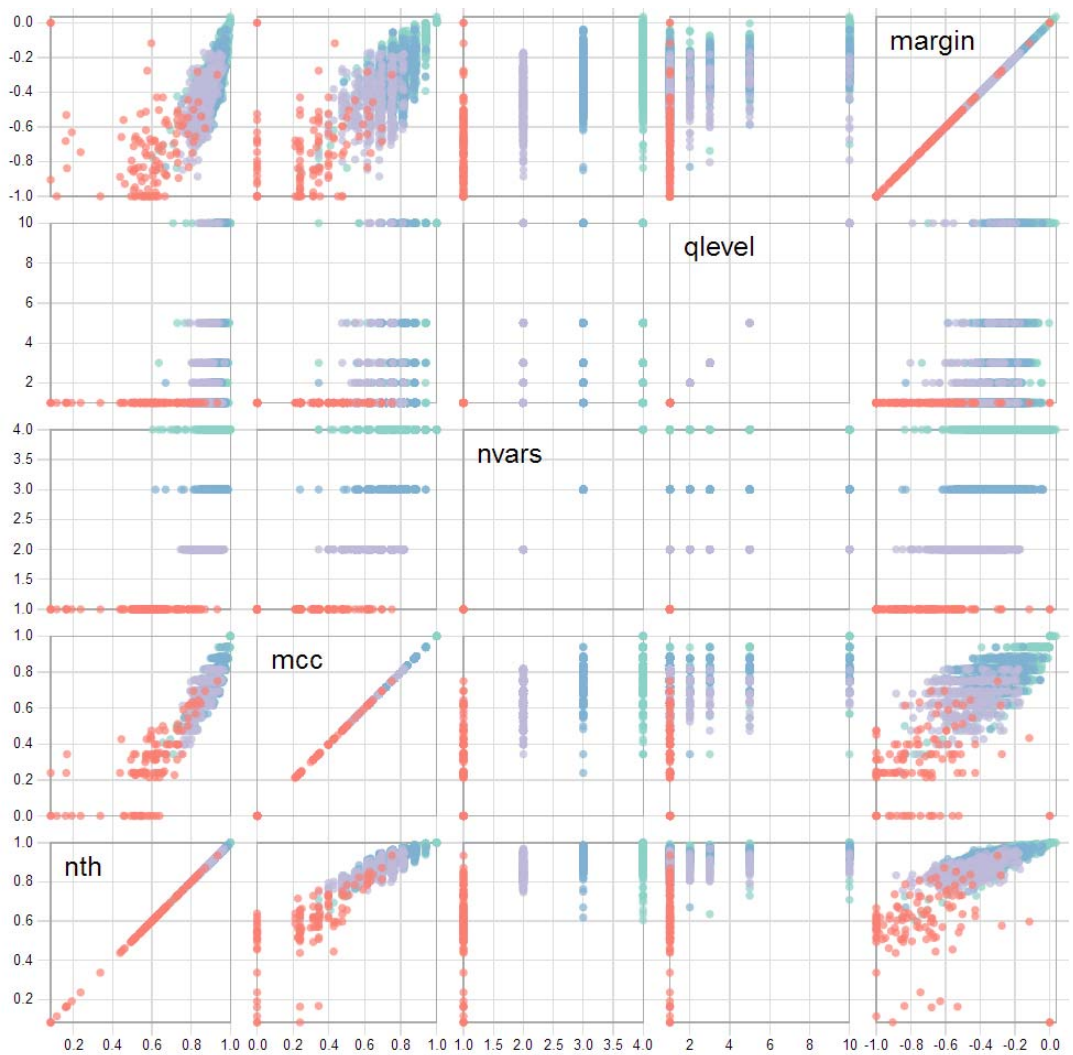
Sample space of variable sets

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



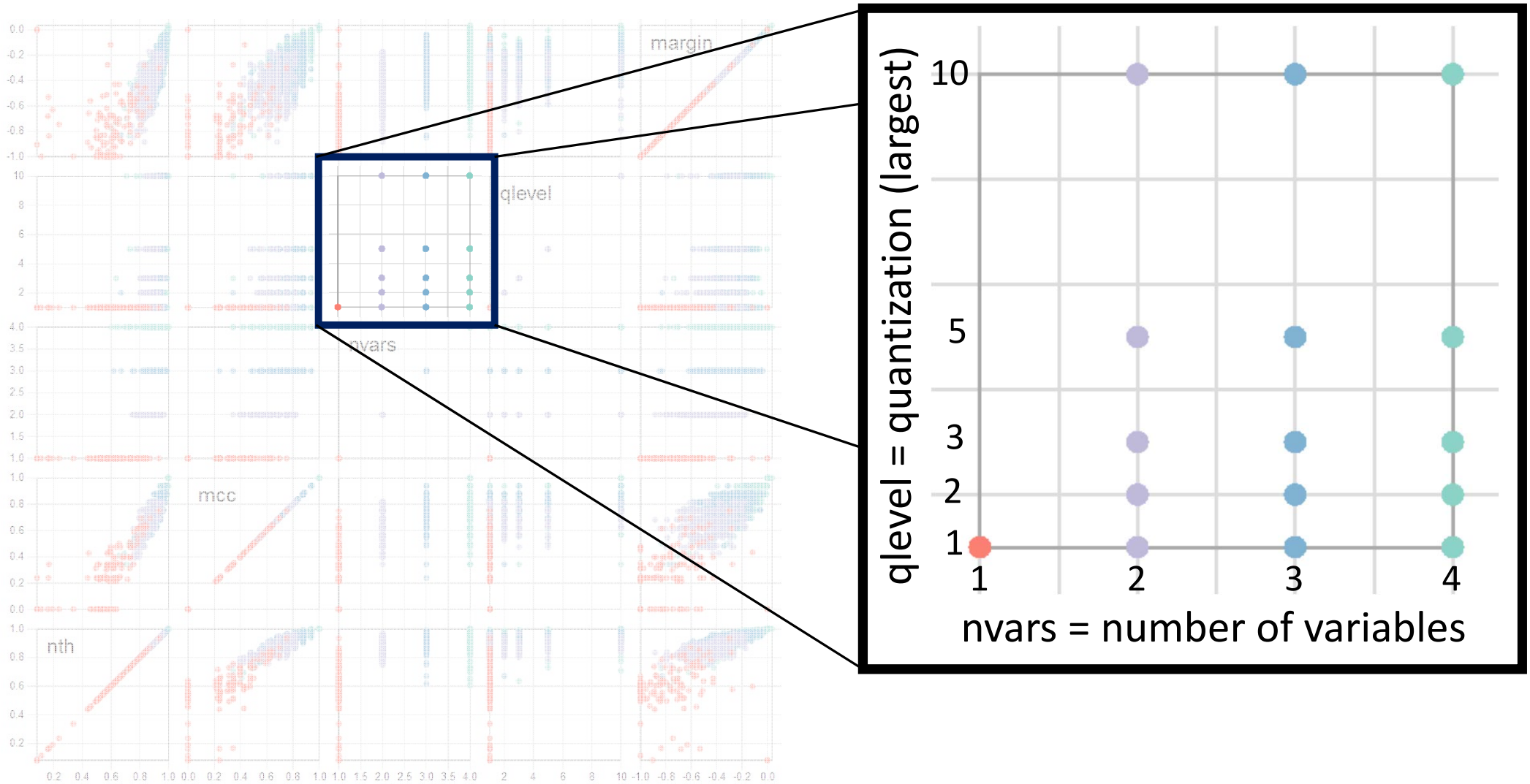
Finding the interesting explainers

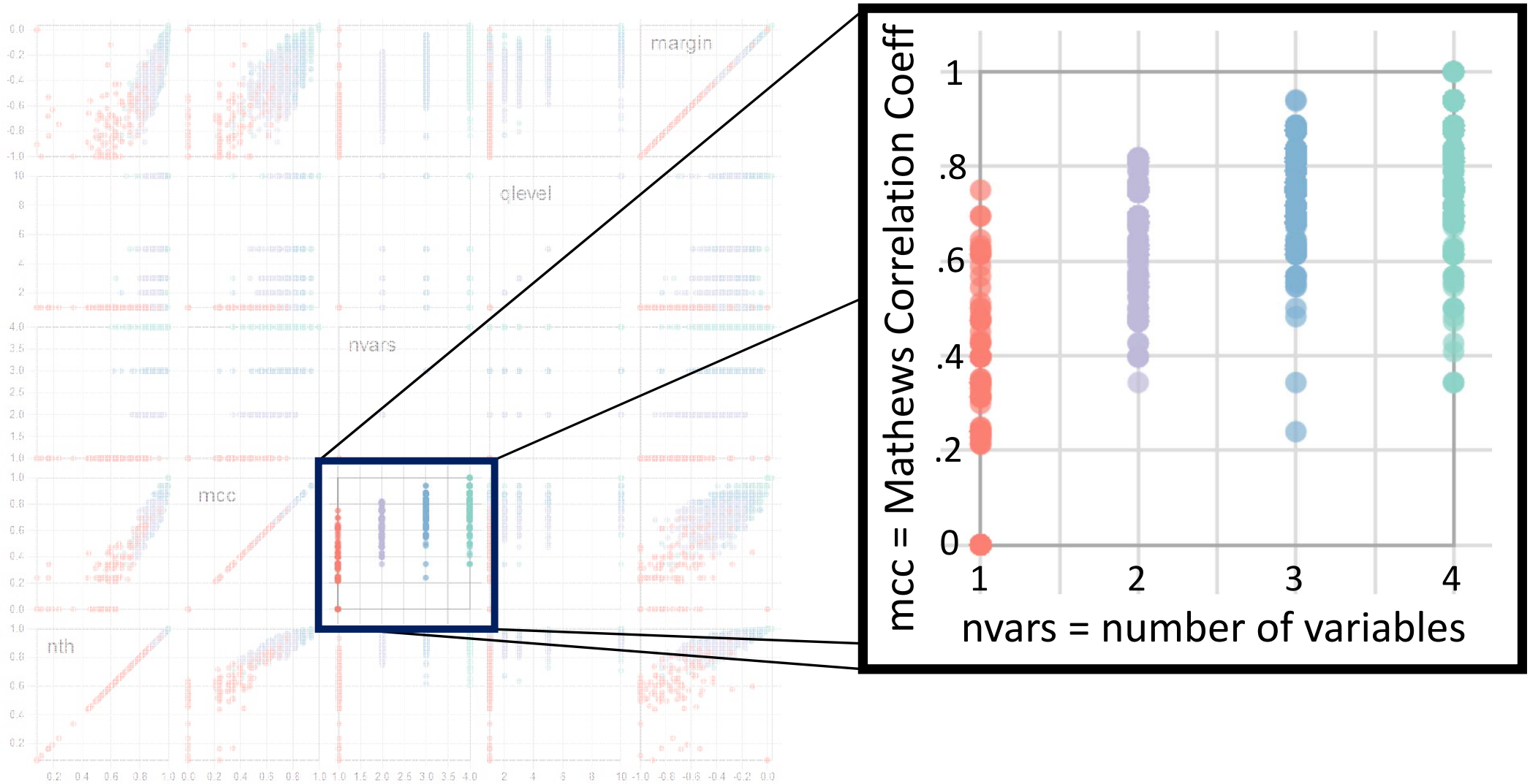
2667 Explainers generated

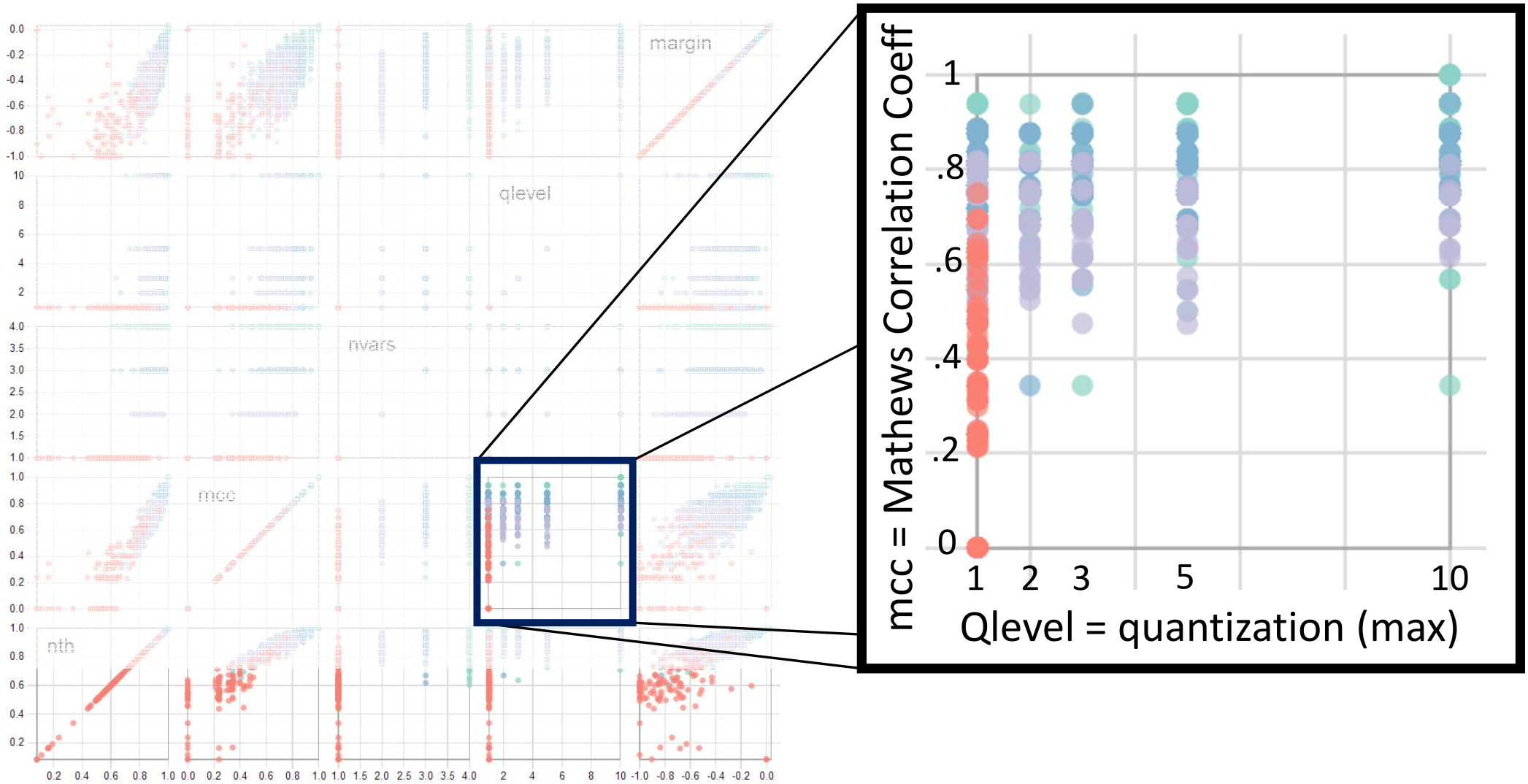
Rank-by-Feature or “Scagnostics” style analysis

2667 points – each an explainer

5 properties of each


margin
qlevel
nvars
mcc
nth
qlevel = quantization (largest)
10
5
3
2
1
1
2
3
4
nvars = number of variables


margin
qlevel
nvars
mcc
nth
mcc = Mathews Correlation Coeff
1
.8
.6
.4
.2
0
1
2
3
4
nvars = number of variables


margin
qlevel
nvars
mcc
nth
mcc = Mathews Correlation Coeff
Qlevel = quantization (max)

Some prior approaches to help situate our work

Isn’t this just...

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~~

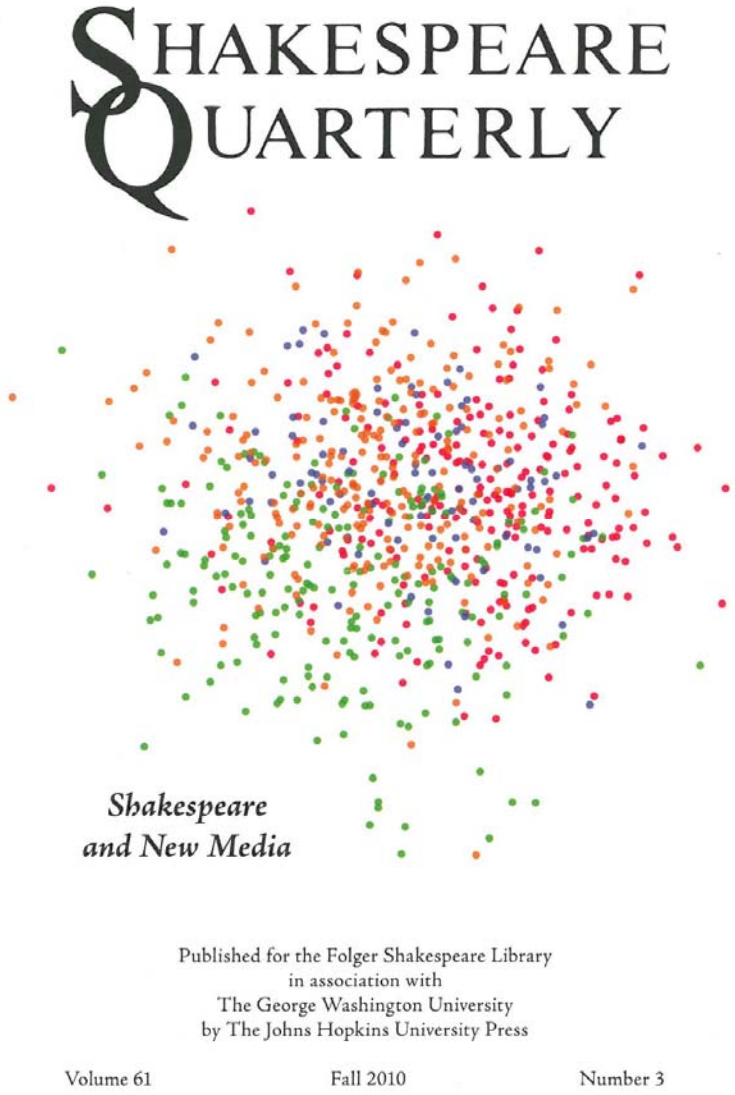
Explainers add:

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

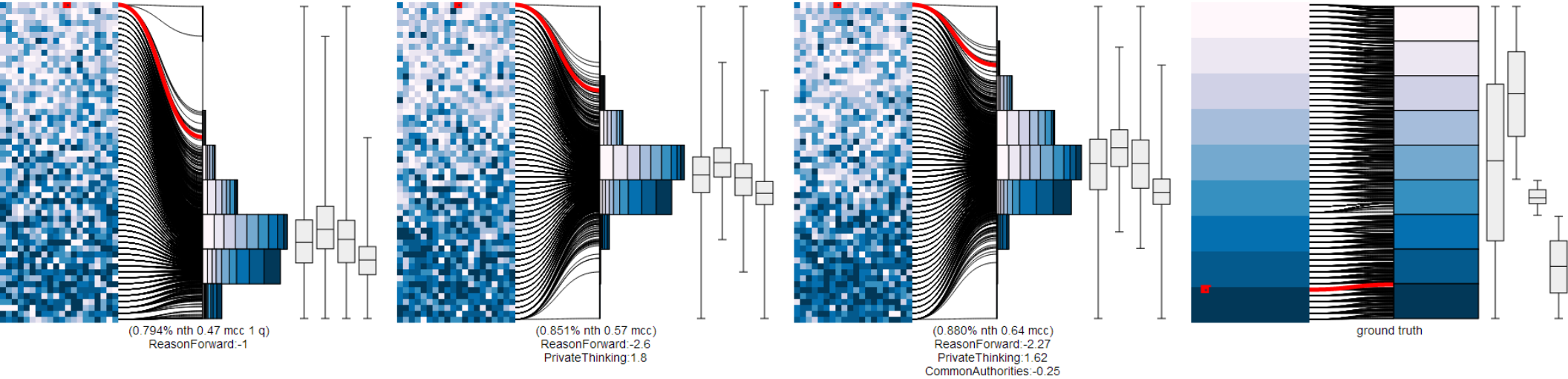
Scalability

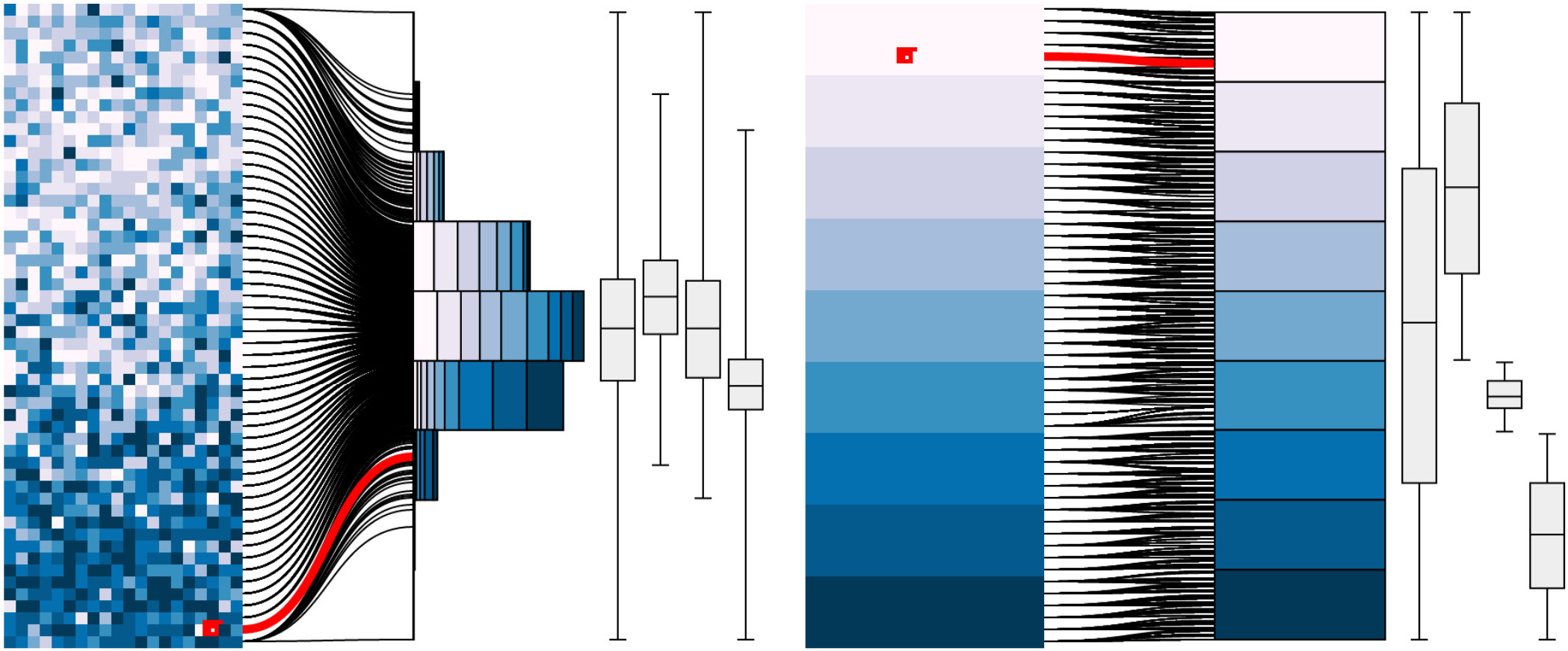
- More variables (redundancy)
- More objects
- More complex relationships

1080 Books

20 books per decade 1580-1800

Are books before 1680 different than books after 1710?




(0.880% nth 0.64 mcc)
ReasonForward:-2.27
PrivateThinking:1.62
CommonAuthorities:-0.25
ground truth

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**

helps **explain** 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.

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

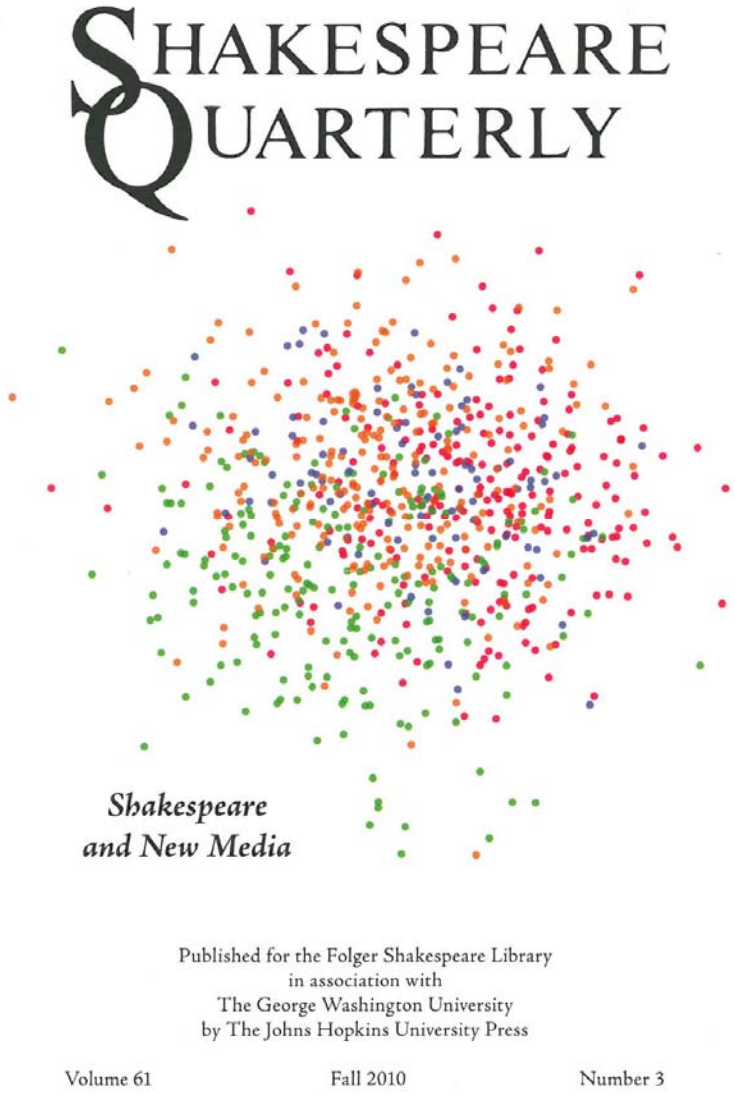
More Examples Online!

`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 scatterplot has too many points!

- Splatterplots – a method for dense scatterplots
- TVCG 2013 – Thursday afternoon

But you can get an average sense anyway…

- Perception of average value in scatterplots
- InfoVis 2013 – Tuesday afternoon

The axes are meaningless!

- Explainers – crafted projections
- VAST 2013 – Wednesday morning