What Shakespeare Taught Us About Visualization and Data Science

Michael Gleicher
Department of Computer Sciences
University of Wisconsin Madison
Acknowledgements

Visualizing English Print Project

Co-Pis
Michael Witmore (Folger Library)
Jonathan Hope (Strathclyde University)
Robin Valenza (UW English)

VEP Students and Staff
Eric Alexander, Deidre Stuffer, Erin Winter, Hao Fu, Joe Kohlmann, ...
Jonathan’s students (Heather, Beth, …)
Robin’s students (Mattie, Cathy, …)

Other Vis Collaborators

My Students (past and present)
Danielle Szafir, Michael Correll, Alper Sarikaya, Kendall Park, …

Perception and HCI Collaborators
Steve Franconeri, Bilge Mutlu, Chuck Hansen, and their students

Other Domain Collaborators
George Phillips, Dave O’Connor, Colin Dewey, …
Lesson 1

It’s a team effort

Lesson 1b

I’m not the Shakespeare Expert
So... why am I here?
Michael Gleicher
Visual Computing Group

Human Graphics Interaction
authoring pictures, videos, animations

Human Robot Interaction
robots!

Human Data Interaction
visualization, visual analytics, interactive learning
Data Visualization

Visual Analytics

Human Data Interaction
Lesson 2

Visualization is not (just) about making pictures
Data Science is not (just) about doing statistics
They are (also) about Human-Data Interaction
What Shakespeare Taught Us About Visualization and Data Science

Michael Gleicher
Department of Computer Sciences
University of Wisconsin Madison
Caveats
This is not a talk about Shakespeare

But he makes for a catchy title

It is a talk about a collaboration with literary scholars

With the goal of getting beyond Shakespeare
(Early Modern Period 1470-1700)
(Or is that 1470-1660? Or 1800?)
This is not a talk about Digital Humanities

This is about a collaboration with literature scholars.

What can the rest of us learn?
Why experience with Literary Scholarship?

Why not experience with Biochemistry, Virology, Robotics, … ?
It’s nice to have an application that people are interested in
Why experience with Literary Scholarship?

Some lessons we could have gotten elsewhere

but this project made them more clear

Some lessons came from unique aspects of the project

but I think they are more general

Some lessons come from the unique collaboration

Humanist* thinking in data analysis!

* I dislike the term “Humanist” because of what it implies about the rest of us. But it is how “they” self-identify.
Mike’s Mantra

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
Working with Literary Scholars
When shall we three meet againe?
In Thunder, Lightning, or in Raine?
Some Lessons from “Humanist” Thinking

*They got along fine before “data-centric” thinking*

Importance of **exemplars** and **outliers**

Importance of going back to the **source** (specific passages)

Value of **multiple** points of view

Contextualized arguments with lots of **background**

**Editing** and **curation** are scholarly activities
Visualizing English Print
c 1470-1800 (1700)

What if you had access to all surviving books?
Visualizing English print (VEP)

Large collections of texts becoming available to scholars
How to enable (traditional) literary scholars to use it?

Problem Factory:
What questions should/could one ask?
How do you ask those questions?
What tools to provide to help?

What can we learn about Visualization / Data Science?
Why? (what is in it for literature)

Consider larger collections of books beyond the “canon”

See language develop (independent of content)

See developments over decades

See patterns too subtle for people to pick up
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

And these lead to many more...
Vis research success stories

Text Viewer
Correll et al, EuroVis 2011

TextDNA
Szafir et al, EuroVis 2016

Serendip
Alexander et al, VAST 2014

Explainers
Gleicher, VAST 2013

Topic Model Comparison
Alexander et al, VAST 2015

Sketch-based search
Correll et al, VAST 2016

Splatterplots
Mayorga & Gleicher 2013
Sarikaya & Gleicher 2015

And others…
What is visualization research?
VEP Vis success stories (some of each kind)

Text Viewer
Correll et al, EuroVis 2011

TextDNA
Szafir et al, EuroVis 2016

Serendip
Alexander et al, VAST 2014

Explainers
Gleicher, VAST 2013

Topic Model Comparison
Alexander et al, VAST 2015

Sketch-based search
Correll et al, VAST 2016

Splatterplots
Mayorga & Gleicher 2013
Sarikaya & Gleicher 2015

And others…
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)
Problem: Scatterplot with way too many points!

Solution: Splatterplot

This is pretty!
But when should we use it?
Lots of choices
Need actionable advice
Scatterplots are common
Lots of designs
How do we choose?
**Task oriented analysis**

Ask Dr. Scatterplot!

VEP Vis success stories (some of each kind)

Text Viewer
Correll et al, EuroVis 2011

TextDNA
Szafir et al, EuroVis 2016

Serendip
Alexander et al, VAST 2014

Explainers
Gleicher, VAST 2013

Topic Model Comparison
Alexander et al, VAST 2015

Sketch-based search
Correll et al, VAST 2016

Splatterplots
Mayorga & Gleicher 2013
Sarikaya & Gleicher 2015

And others…
But are there bigger lessons?

I learned a lot from this
Three interesting aspects of VEP

Literary scholarship – and ways of thinking
Literary scholars – with their needs and abilities
300 years of old books – variety and quality
300+ years of books

A period of intense changes:
cultural, historical, intellectual, political, …

*How does the printed record reflect this?*

The language changed a lot
Printing changed a lot
The book business changed a lot
Libraries changed a lot
300+ years of books

Every known surviving book?

Huge historical efforts
to catalog known books
to microfilm known books
to transcribe known books

Text Creation Partnership (TCP)
Hand-keyed about 60,000 books into SGML/XML
Sources of variation

Variety of topics
Authors’ thoughts and ideas
Authors’ words
Printers’ practicality
Librarians’ taking care of books
Transcribers translating books

Modern processing tools handling things correctly
It’s naïve to spell naïve with an i

Did the author use an ï?
Did the printer have it?
Did the dots get smudged?
Did the microfilming make it look like a dust mark?
Did the transcriber see it?
Did they encode it weirdly that year?
Did our Unicode pipeline mess it up?
Did our spelling standardizer get it wrong?
One form of variation can obscure another
But...
One person’s noise is another’s data

spelling obscures ideas
or
observe the development of spelling

rotting pages obscure content
or
preserved books tell us what was valued
Data Wrangling

Make data convenient for analysis
Clean away “unwanted” variance
Leave enough signal

Hard choices!
Data Wrangling Lessons

Everyone knows Data Wrangling is a big deal, time consuming, ...

Getting good meta-data requires investment

Transparency of data wrangling is valuable

Comprehensibility – the stakeholders need to understand

Curation and editing is part of scholarship. These are decisions – make them wisely!
A view of Data Wrangling (or cleanup)

It’s about variation:

variation is what we’re interested in

in order to expose the variation you care about
you need to clean away the variation you don’t

Curation and editing are part of scholarship. These are decisions – make them wisely!
A case study...
Counting “the”

It is interesting at scale
Look for correlations across huge amounts of literature!

It is transparent
we can explain – and check – every step

It’s still hard to get right…
Just because counting is easy – doesn’t mean anything else is
Some literature theory...

Writing about things in the world (Extra-Subjective)

vs.

Writing about abstract things (Intra-Subjective)

How does this play out over 300 years, 60,000 documents, ... ?

How does this play out over plays?

about 1500 documents- that we have lots of meta-data on
Extra-Subjective vs. Intra-Subjective

Is there an easy way to measure it?

Statistical analysis:
The primary variance in the collection is correlated with labels (science, plays)

The main source of this variance is the most common word
Count “the”? 

Use definite articles to refer to objects in the world

**Science** uses it more than **Plays**

We only have (extensive) labels for **science** and **plays**

Most things are **unlabeled**
Count “the” ?
23% “the” – really?

A plain and easy rule to rigge any ship by the length of his masts, and yards, without any further trouble

23% “the”?

Stay One time the length of the Mast wanting

Lifts Three times the length of the Shroudes.

Clugarnets Three times the length of the Shroudes.

Buntlines Two times the length of the Mast from the Deck to the Crostrees.
AN EXACT LIST OF THE FRENCH Fleet and Commanders For this Present YEAR

51% “the” if you count differently

51% of “words” – only counting things we recognize as words
Why is this hard?

We need to carefully treat each book in a uniform way
Which words do we extract off the pages?

We need to carefully treat each word in a uniform way
How to deal with bad characters and spelling variation?

We need to analyze in a uniform and transparent way
We need good meta-data

The scholars (and their audience) must understand each step!
We need to get back to the sources!
Big lesson:

Data Science is a process
Challenges are everywhere!
Big lesson:

We must help all users understand all phases of the process
Big lesson:

We must help all users understand all phases of the process
Towards Comprehensibility in Modeling

The other title for this talk
Data Science is a Process

There are lots of steps
Many provide challenges
Many provide opportunities

Identify Problem
Gather Data
Clean Data
Abstract
Design
Build
Evaluate/Validate
Interpret
Act
Disseminate
Can we automate this process? Make it possible for subject experts?
Data Driven Discovery of Models

Smart systems to do steps

Smart user experience to guide users

Some steps seem really hard

Identify Problem
Gather Data
Clean Data
Abstract
Design
Build
Evaluate/Validate
Interpret
Act
Disseminate
Problem Definition

This will be a description of the problem for the user to review. This problem is very interesting. It is perhaps the most interesting of all the problems. Can you think of a problem more interesting than it?
Workflow View
Show steps in process
Show progress and next steps
Connect to Visualizations
Allow for skipping (guess for me)
Allow for review (go back)
## Comprehensibility across the Process

<table>
<thead>
<tr>
<th>Who?</th>
<th>Why?</th>
<th>Where?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stakeholders</td>
<td>Reason for Wanting</td>
<td>Phase of the <strong>Process</strong></td>
</tr>
<tr>
<td>Developers</td>
<td>Improve Performance</td>
<td>Identify Problem</td>
</tr>
<tr>
<td>Data Scientists</td>
<td>Build Theory</td>
<td>Gather Data</td>
</tr>
<tr>
<td>Domain Experts</td>
<td>Extend/Characterize</td>
<td>Clean Data</td>
</tr>
<tr>
<td>Audience</td>
<td>Build Trust</td>
<td>Abstract</td>
</tr>
<tr>
<td>Subjects</td>
<td>Actionability</td>
<td>Design</td>
</tr>
<tr>
<td></td>
<td>Convince</td>
<td>Build</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interpret</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disseminate</td>
</tr>
</tbody>
</table>

Who? Why? Where?
(for How? and What? See the paper)

<table>
<thead>
<tr>
<th>Who?</th>
<th>Why?</th>
<th>Where?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stakeholders</td>
<td>Reason for Wanting</td>
<td>Phase of the <strong>Process</strong></td>
</tr>
<tr>
<td>Developers</td>
<td>Improve Performance</td>
<td>Identify Problem</td>
</tr>
<tr>
<td>Data Scientists</td>
<td>Build Theory</td>
<td>Gather Data</td>
</tr>
<tr>
<td>Domain Experts</td>
<td>Extend/Characterize</td>
<td>Clean Data</td>
</tr>
<tr>
<td>Audience</td>
<td>Build Trust</td>
<td>Abstract</td>
</tr>
<tr>
<td>Subjects</td>
<td>Actionability</td>
<td>Design</td>
</tr>
<tr>
<td></td>
<td>Convince</td>
<td>Build</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evaluate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interpret</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disseminate</td>
</tr>
</tbody>
</table>

Some phases for comprehensibility

Data Wrangling
understand what data the data is and what it can do

Model Building
understand classifiers to build theory / identify items

Validation Experiments
understand experiments to identify items / build trust
Towards Comprehensibility in Modeling

This is not just comprehensible models
genre

Categorization given by Shakespeare's contemporaries

- Comedy
- Tragedy
- History

Category for plays written after that

- Late Plays
data

Count words

Count kinds of words – using dictionaries
- Docuscope (Igarashi and Kaufer)
- Lists of phrases for each “Language Action Type”
- Simple and Transparent dictionary matching
  (once the dictionaries are made)

Fancier methods (Topic Models, Machine Learning, …)
Is “the” good enough?

Histories and Tragedies are more external
Comedies are more internal

The second most common word, “I”, might be better (it is for internal)

600 plays of Shakespeare’s era
Neither is great, maybe try combining them?
Mathematical Models

“the” – “I” --- pretty good
("the” + “to”) – (“I” + “a”) --- very good

How do we come up with such equations?
What do we do with them?

Machine Learning Classification:
Make up these function so that one “class” is high
The other classes are low
Comprehensibility in modeling?

They may never understand Support Vector Machines (or choose your favorite buzzword)

But they might understand what it does…
Use Models for Insight

We know the answers!
Teaching the machine can help us:

Can you pick apart known groups?

Does the data capture the concepts?
Language features vs. genre

Can we quantify and organize? (comedic-ness)
Comedicness

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?
Use **Simple** Models for Insight

Create models for comprehensibility

See what you get wrong

Look “inside” models to see how they work

comedicness = M − I

f(V) = V[39] − V[42]

M = Predicted Future
I = Inclusiveness

36 plays of Shakespeare, Docuscope Features
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
Simpler Functions

Easier to Understand
More likely to lead to Theory

Less Expressive
Less Likely to be Accurate
(but more likely to overfit)
Tradeoffs

Give the user control over the tradeoffs

But how do we help them make informed choices?
Paths to model usability?

Interpretable models
simplify the models so they can be understood

Examinable models
look inside the models and hope you understand

Instance-based explanations
pick some decisions and try to understand them

Experiment/Outcome Examination
look at the right input/outputs from the black box
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
Use Simple Models for Insight

Create models for comprehensibility

See what you get wrong

Look “inside” models to see how they work

Lesson

Good tools for exploring your results are useful.
Is literature unique?

Large library has evolved
Understanding different kinds of variation
Bringing in other knowledge
Look at specific examples and outliers
Drill into details
Learn from validation experiments

Get stakeholders involved in all data science phases by helping them understand
Some Lessons from “Humanist” Thinking

They got along fine before “data-centric” thinking

Importance of exemplars and outliers
Importance of going back to the source (specific passages)

Value of multiple points of view

Contextualized arguments with lots of background
Editing and curation are scholarly activities
Diversity

F + Q – I
C – M – I
P + N + D

Same:
Correctness
Simplicity

Different:
Explanations
Orderings
Maybe Literature Scholars aren’t so weird?

They got along fine before “data-centric” thinking

Importance of exemplars and outliers
Importance of going back to the source (specific passages)
Value of multiple points of view
Contextualized arguments with lots of background
Editing and curation are scholarly activities
OK, some things are weird
But, this might be a lesson...

Don’t just give “them” your methods

If you listen carefully,
they might have things to teach you
Thanks!
To you for listening.
To my students and collaborators.
To the NSF, NIH, DARPA and Mellon Foundation for funding.

What Shakespeare taught us about (Visual) Data Science

Michael Gleicher
University of Wisconsin Madison
gleicher@cs.wisc.edu