A systematic method for surveying data visualizations and a resulting genomic epidemiology visualization typology: GEViT

Anamaria Crisan
PhD Candidate, Computer Science
University of British Columbia

https://doi.org/10.1101/325290

@amcrisan  acrisan@cs.ubc.ca  http://cs.ubc.ca/~acrisan
How do people choose a data visualization?

- Below are visualizations for hospital outbreaks
- Different data, different visualizations, different emphasis – WHY?

How do people choose a data visualization?

Could I conduct a *systematic review of data visualizations* used in infectious disease genomic epidemiology?
Our method to conduct a systematic review of data visualizations
An overview of our method

• Our method has two analysis phases:

- Literature Analysis
- Qualitative Analysis
- Quantitative Analysis
- Visualization Analysis
An overview of our method

- Some analyses are automated (🤖) and others are not (👤)

<table>
<thead>
<tr>
<th>Analysis Phase</th>
<th>Literature Analysis</th>
<th>Visualization Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Qualitative Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quantitative Analysis</td>
</tr>
</tbody>
</table>
An overview of our method

- Analysis phases answer **different** research questions

<table>
<thead>
<tr>
<th>Analysis Phase</th>
<th>Literature Analysis</th>
<th>Visualization Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Qualitative Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quantitative Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Research Question</th>
<th>WHY are researchers visualizing data?</th>
<th>HOW are researchers visualizing data?</th>
<th>HOW MANY examples of specific visualizations?</th>
</tr>
</thead>
</table>
## Applying our method

| Our Objective | Across the many topics of microbial gen epi research articles identify and enumerate the different kinds of visualizations that are used |
## Applying our method

### Literature Analysis Steps
- **Text mining of document corpus to identify topics**
- **Systematically sample papers with topics as strata**

### Our Objective

Across the many **topics** of microbial gen epi research **articles**
identify and enumerate the different kinds of visualizations that are used
Applying our method

Literature Analysis Steps
- Text mining of document corpus to identify topics
- Systematically sample papers with topics as strata

Our Objective
- Across the many topics of microbial gen epi research articles, identify and enumerate the different kinds of visualizations that are used

Visualization Analysis Steps
- Derived a code set to classify research figures (GEViT)
- Applied GEViT to collection of research figures
Applying our method

Literature Analysis Steps
- Text mining of document corpus to identify topics
- Systematically sample papers with topics as strata

Our Objective
- Across the many topics of microbial gen epi research articles, identify and enumerate the different kinds of visualizations that are used

Visualization Analysis Steps
- Derived a code set to classify research figures (GEViT)
- Applied GEViT to collection of research figures
- Applied descriptive statistics to derived code sets
# Key analysis steps & their results

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Acquisition &amp; Unsupervised Clustering</td>
<td>17,974</td>
<td>Article topic clusters&lt;br&gt;<code>WHY</code> are researchers visualizing data?</td>
</tr>
<tr>
<td>Limit to clusters of human pathogens</td>
<td>6,350</td>
<td></td>
</tr>
<tr>
<td>Stratified Sampling</td>
<td>221</td>
<td>801 figures &amp; 49 tables</td>
</tr>
</tbody>
</table>
# Key analysis steps & their results

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
</table>
| Article Acquisition & Unsupervised Clustering      | 17,974     | Article topic clusters
WHY are researchers visualizing data?               |
| Limit to clusters of human pathogens               | 6,350      |                                                                         |
| Stratified Sampling                                | 221        | 801 figures & 49 tables                                                |
| Iterative axial coding                             | 221        | A genomic epidemiology visualization typology (GEViT)
How are researchers visualizing data?              |
### Key analysis steps & their results

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Acquisition &amp; Unsupervised Clustering</td>
<td>17,974</td>
<td>Article topic clusters <strong>WHY</strong> are researchers visualizing data?</td>
</tr>
<tr>
<td>Limit to clusters of human pathogens</td>
<td>6,350</td>
<td></td>
</tr>
<tr>
<td>Stratified Sampling</td>
<td>221</td>
<td>801 figures &amp; 49 tables</td>
</tr>
<tr>
<td>Iterative axial coding</td>
<td>221</td>
<td>A genomic epidemiology visualization typology (GEViT) <strong>How</strong> are researchers visualizing data?</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>221</td>
<td>Current common visualization practices</td>
</tr>
</tbody>
</table>
A snapshot of our findings
17,974 articles on infectious disease gene epi.

<table>
<thead>
<tr>
<th>PMID</th>
<th>YearPub</th>
<th>Journal</th>
<th>Authors</th>
<th>Title</th>
<th>Abstract</th>
<th>PMCID</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>27607836</td>
<td>2016</td>
<td>PLoS negle Antonanion, I Bacillus cereus Biovar Ant</td>
<td>Through full genome analyses c PMC5015827</td>
<td></td>
<td>10.1371/journal.pntd.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26819311</td>
<td>2016</td>
<td>Journal of Bodewes, R</td>
<td>Spatiotemporal Analysis of influenza A viruses are major pm</td>
<td></td>
<td>10.1128/JVI.03046-15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26735036</td>
<td>2016</td>
<td>Journal of Verheghe, F</td>
<td>Prevalence and Genetic Di Since the first description of livi</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>25113057</td>
<td>2014</td>
<td>Annals of Chen, Sheng</td>
<td>Spread of carbapenemase BACKGROUND: The rapid emergence</td>
<td></td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>22732256</td>
<td>2014</td>
<td>Journal of Arduino, Son</td>
<td>Transposons and Integrants: Multiple transposons, integrants NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>22386850</td>
<td>2014</td>
<td>Journal of Infection, I</td>
<td>Rivero-Piore Genetic diversity of comm With the recent detection of MI NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>19487205</td>
<td>2009</td>
<td>Philosophi Sloat, P M A; HIV decision support: from Human immunodeficiency virus</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19000628</td>
<td>2008</td>
<td>Virology</td>
<td>Chen, Rubing Frequent inter-species tra Revealing the factors that shape PM</td>
<td></td>
<td>10.1016/j.virology.2008.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18564686</td>
<td>2008</td>
<td>The Souther Tiwari, Hare</td>
<td>Molecular typing of clinica Molecular typing of total 84 Sta NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>17617184</td>
<td>2007</td>
<td>Clinical microkiwalec, M</td>
<td>Hospital outbreak of vancomycin A resistant outbreak caused by var NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>28042011</td>
<td>2017</td>
<td>Journal of Deuereg, Application of next genera Current molecular diagnostics c NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26598368</td>
<td>2015</td>
<td>Journal of Box, Allison</td>
<td>Functional Analysis of Bac The classical and El Tor biotype PM</td>
<td></td>
<td>10.1128/JB.000747-15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22502605</td>
<td>2012</td>
<td>Letters in</td>
<td>Alilabad, N H Molecular diversity of CTX AIMS: The objective of this stud NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>21643168</td>
<td>2011</td>
<td>BMC geno Chen, Yunsal</td>
<td>Comparative genomic ana BACKGROUND: Vibrio parahectori PM</td>
<td></td>
<td>10.1186/1471-2164-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21177926</td>
<td>2010</td>
<td>The Journ Lorusso, Ales</td>
<td>Genetic and antigenic cha Prior to the introduction of the PM</td>
<td></td>
<td>10.1099/vir.0.027557-0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20375036</td>
<td>2010</td>
<td>The Journ Post, Virginio</td>
<td>Evolutionary type of ABAr-type ge OBJECTIVES: To determine if m NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>19809504</td>
<td>2009</td>
<td>PLoS one</td>
<td>Khiabanian, I Reassortment patterns in Three human influenza pandem PM</td>
<td></td>
<td>10.1371/journal.pone.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18815309</td>
<td>2009</td>
<td>Virologie</td>
<td>Salloum, Sha</td>
<td>Escape from HLA-B*08-01 The inherent sequence diversity PM</td>
<td>10.1128/JVI.00997-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17651134</td>
<td>2007</td>
<td>FEBS miC Chen, Yunsal</td>
<td>Genetic variation of capsule Both NRT36S and AS are NAG-1 NA</td>
<td>NA</td>
<td></td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>15213324</td>
<td>2004</td>
<td>Proceeding Holden, Matti</td>
<td>Complete genomes of two Staphylococcus aureus is an im PM</td>
<td></td>
<td>10.1073/pnas.0402521</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
17,974 articles following unsupervised clustering

Initial Topic clustering results

t-SNE followed by hdbscan

Final set of topic clusters

Articles appear to cluster around pathogens

A cluster of articles

Unclustered Articles

Too few articles for a reliable cluster
Why is this finding relevant?

• Provides an overview of the articles
  • Wayfinding for downstream analyses
Why is this finding relevant?

• Provides an overview of the articles
  • Wayfinding for downstream analyses

• Different data visualizations for different pathogens
  • Different transmission routes, mortality, etc.
  • Should there be different visualizations? Yes, No, Maybe?
Why is this finding relevant?

• Provides an overview of the articles
  • Wayfinding for downstream analyses

• Different data visualizations for different pathogens
  • Different transmission routes, mortality, etc.
  • Should there be different visualizations? Yes, No, Maybe?

• Sample across topic clusters for qualitative analysis
  • Assumption: Different topics will have different data visualizations
  • Sample across topics for better coverage
Why is this finding relevant?

• Provides an overview of the articles
  • Wayfinding for downstream analyses

• Different data visualizations for different pathogens
  • Different transmission routes, mortality, etc.
  • Should there be different visualizations? Yes, No, Maybe?

• Sample across topic clusters for qualitative analysis
  • Assumption: Different topics will have different data visualizations
  • Sample across topics for better coverage

Paper Alert!
We made a literature analysis into an R tool - Adjutant!

https://doi.org/10.1093/bioinformatics/bty722
https://github.com/amcrisan/adjutant
Describing visualization design through GEViT

- **Input:** 801 figures, 49 tables
- **Used qualitative coding techniques to analyze research figures**
  - Multiple rounds of classifying and codifying elements of figures
  - Used figures from sample papers to derive codes
Describing visualization design through GEViT

- **Input:** 801 figures, 49 tables
- Used qualitative coding techniques to analyze research figures
  - Multiple rounds of classifying and codifying elements of figures
  - Used figures from sample papers to derive codes
- Figures in the same paper were analyzed separately
  - Multi-part figures were analyzed together
Describing visualization design through GEViT

- **Input:** 801 figures, 49 tables
- **Used qualitative coding techniques to analyze research figures**
  - Multiple rounds of classifying and codifying elements of figures
  - Used figures from sample papers to derive codes
- **Figures in the same paper were analyzed separately**
  - Multi-part figures were analyzed together
- **Result: A Genomic Epidemiology Visualization Typology (GEViT)**
  - GEViT is a hierarchical code set with separate taxonomies for:
    - Chart Types
    - Chart Combinations
    - Chart Enhancements
Describing visualization design through GEViT

- What does GEViT do and not do?

**GEViT provides a base**

- A Visualization Typology for visual design
  1. Chart Types
  2. Chart Combinations
  3. Chart Enhancements

- An Interactive Gallery

**GEViT does not evaluate**

- Massive undertaking that would take many years
- Needs GEViT to conduct evaluations
Chart types: the basic building blocks

Common Statistical Charts
- Bar Chart: Standard, Stacked, Divergent, Special Cases (Epidemic Curve, Diversity Chart, LEfSe Plot)
- Line Chart: Special Cases (Bootscan, Kaplan-Meier, Skysline Plot)
- Scatter Plot: Special Cases (Root-to-tip, Ordination Plot, Q-Q plot)

Distribution Plot
- Histogram, PDF, Boxplot, Swarm Plot

Colour Charts
- Pie Chart, Venn Diagram, Category Stripe, Heatmap, Density Plot*

Relational Charts
- Node-link: Special Cases (eBurst, Social network, Molecular network, Minimum Spanning Tree)

Flow Diagram
- Chord Diagram, Sankey Diagram

Temporal Charts
- Streamgraph*: Absolute, Relative

Spatial Charts
- Geographic Map, Choropleth Map, Interior Map

Tree Charts
- Phylogenetic Tree: Rooted (Linear & Radial)

Genomic Charts
- Genomic Map: Linear, Radial
- Alignment, Composition Plot

Dendrogram, Clonal Tree*

Sequence Logo Plot

Other Charts
- Table, Image, Gel Image, General Image, Miscellany
Current Common Practice

- Phylogenetic trees most common (not a surprise)
  - Hard for interpret (a finding from our prior work)
- Tables in figure next most common (a surprise)
  - Also, a lot of data as text
- Limited range of chart types used
- Anecdotal observation: common statistical charts most widely supported by software
# Chart combinations: showing even more data

<table>
<thead>
<tr>
<th>Combination Type</th>
<th># of chart types</th>
<th># of charts</th>
<th>Linkage type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td><img src="chart1.png" alt="Example" /></td>
</tr>
<tr>
<td>Composite</td>
<td>Many</td>
<td>1</td>
<td>Spatially Aligned</td>
<td><img src="chart2.png" alt="Example" /></td>
</tr>
<tr>
<td>Small Multiples</td>
<td>1</td>
<td>Many</td>
<td>Chart Type &amp; Data</td>
<td><img src="chart3.png" alt="Example" /></td>
</tr>
<tr>
<td>Many Types Linked</td>
<td>Many</td>
<td>Many</td>
<td>Visual, but not spatial</td>
<td><img src="chart4.png" alt="Example" /></td>
</tr>
<tr>
<td>Many Types General</td>
<td>Many</td>
<td>Many</td>
<td>NA</td>
<td><img src="chart5.png" alt="Example" /></td>
</tr>
<tr>
<td>Complex Combinations</td>
<td>Many</td>
<td>Many</td>
<td>Context dependent</td>
<td><img src="chart6.png" alt="Example" /></td>
</tr>
</tbody>
</table>
# Chart Combos: current common visualization practices

<table>
<thead>
<tr>
<th>Combination Type</th>
<th># of chart types</th>
<th># of charts</th>
<th>Linkage type</th>
<th>Example</th>
<th>Current Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td><img src="chart.png" alt="Simple Example" /></td>
<td>40.1% of all figures</td>
</tr>
<tr>
<td>Composite</td>
<td>Many</td>
<td>1</td>
<td>Spatially Aligned</td>
<td><img src="chart.png" alt="Composite Example" /></td>
<td>20.3%</td>
</tr>
<tr>
<td>Small Multiples</td>
<td>1</td>
<td>Many</td>
<td>Chart Type &amp; Data</td>
<td><img src="chart.png" alt="Small Multiples Example" /></td>
<td>17.3%</td>
</tr>
<tr>
<td>Many Types Linked</td>
<td>Many</td>
<td>Many</td>
<td>Visual, but not spatial</td>
<td><img src="chart.png" alt="Many Types Linked Example" /></td>
<td>13.5%</td>
</tr>
<tr>
<td>Many Types General</td>
<td>Many</td>
<td>Many</td>
<td>NA</td>
<td><img src="chart.png" alt="Many Types General Example" /></td>
<td>8.8%</td>
</tr>
<tr>
<td>Complex Combinations</td>
<td>Many</td>
<td>Many</td>
<td>Context dependent</td>
<td><img src="chart.png" alt="Complex Combinations Example" /></td>
<td>11.9%</td>
</tr>
</tbody>
</table>

*Note: The percentages represent the proportion of figures using each combination type.*
Chart enhancements: overlaying metadata
Chart enhancements: overlaying metadata

Chart Enhancement Examples

Base Chart

Re-encode Marks
Line: color

Add Marks
Point: color; line; text: font face

Add Annotation
Arrow, text

No known contacts
Chart enhancements: overlaying metadata

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Shape</th>
<th>Color</th>
<th>Texture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point</td>
<td>. . .</td>
<td>□ . . .</td>
<td>. . . . . .</td>
<td>. . . . . .</td>
</tr>
<tr>
<td>Line</td>
<td></td>
<td>I I I</td>
<td>I I I I I I</td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>A A A</td>
<td></td>
<td>A A A A A</td>
<td>A A A A A</td>
</tr>
</tbody>
</table>

**Current Practice**  
>80% of all figures have some enhancement
GEViT in action

Gorrie (2017)
GEViT in action

Gorrie (2017)

Visualization Breakdown

Literature Analysis (why)
- **Pathogen**: Enterococcus faecium
- **A priori concepts**: control; genome; outbreak; drug resistance; phylogeny; genotype
GEViT in action

Gorrie (2017)

Visualization Breakdown

**Literature Analysis** *(why)*
- **Pathogen**: Enterococcus faecium
- **A priori concepts**: control; genome; outbreak; drug resistance; phylogeny; genotype

**Visualization Components** *(how)*

<table>
<thead>
<tr>
<th>Chart Type</th>
<th>Tree (Rooted Phylogenetic Tree)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Category Stripe</td>
</tr>
<tr>
<td></td>
<td>Heatmap (Variation Profile)</td>
</tr>
</tbody>
</table>
GEViT in action

Gorrie (2017)

Visualization Breakdown

**Literature Analysis** *(why)*

- **Pathogen**: Enterococcus faecium
- **A priori concepts**: control; genome; outbreak; drug resistance; phylogeny; genotype

**Visualization Components** *(how)*

<table>
<thead>
<tr>
<th>Chart Type</th>
<th>Tree (Rooted Phylogenetic Tree)</th>
<th>Category Stripe</th>
<th>Heatmap (Variation Profile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart Combination</td>
<td>Composite <em>(spatially aligned)</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GEViT in action

Gorrie (2017)

Visualization Breakdown

Literature Analysis (why)
- **Pathogen**: Enterococcus faecium
- **A priori concepts**: control; genome; outbreak; drug resistance; phylogeny; genotype

Visualization Components (how)

<table>
<thead>
<tr>
<th>Chart Type</th>
<th>Tree (Rooted Phylogenetic Tree) Category Stripe Heatmap (Variation Profile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart Combination</td>
<td>Composite (spatially aligned)</td>
</tr>
<tr>
<td>Chart Enhancement</td>
<td>Re-encode Marks Tree – branches</td>
</tr>
<tr>
<td></td>
<td>Add Marks Tree - Connection Marks</td>
</tr>
<tr>
<td></td>
<td>Add Mark (unstructured) Heatmap – Textboxes</td>
</tr>
</tbody>
</table>
GEViT in action

Pre-print available: https://doi.org/10.1101/325290
To appear in Oxford Bioinformatics!!
Why is GEViT relevant?

- GEViT provides a common way to describe visualizations
- Adds systematicity and formalism to visual analysis & comparisons
- Summarizes current common visualization practices
Impact of our findings
Implications of our findings

- Need to move away from ad hoc visualization development
  - Need awareness of visualization alternatives
  - Need to know what is possible, common, and even absent
Implications of our findings

- Need to move away from ad hoc visualization development
  - Need awareness of visualization alternatives
  - Need to know what is possible, common, and even absent

- Implications for bioinformatics and data visualization tool development
  - Need tools that support complexity and expressivity in visual design
  - Provides design alternatives for bioinformaticians to explore and test
Implications of our findings

- Need to move away from ad hoc visualization development
  - Need awareness of visualization alternatives
  - Need to know what is possible, common, and even absent

- Implications for bioinformatics and data visualization tool development
  - Need tools that support complexity and expressivity in visual design
  - Provides design alternatives for bioinformaticians to explore and test

- Implications for education
  - GEViT as a teaching tool (I am already doing this)
  - Getting people from current practice to best practice
  - Inspiration for researchers
Next Steps
Next Steps: further automating our analysis

Current: Still needs (wo)man power to construct design space

Future: Human-in the loop, but driven by automated methods
Next Steps: using GEViT to create visualizations

Creating visualizations with simple and minimal syntax

- Data Input Layer
- Figure Specification Layer
- Algorithmic Layer
- Design Defaults / Layout / Rendering
- Package Integration Layer
- ggplot2
- Base R

GEViT API

- Uses GEViT terminology
- Goal: Rapid, expressive, and reproducible figure generation

R Visualization
Next Steps: using GEViT to create visualizations

Shannah Fisher
GEViT API developer
Undergraduate student @ UBC
Thank you!

Dr. Jennifer Gardy
Dr. Tamara Munzner

+ UBC infoVis group
Kimberly Dextras-Romagnino, Madison Elliott, Shannah Fisher, Micheal Opperman, and Zipeng Liu

+ Reviewers and many others that have provided feedback on this work

I am graduating soon!
I am on the job market this year!

Pre-Print + Other Stuff

https://doi.org/10.1101/325290

https://amcrisan.github.io/talks

http://gevit.net

acrisan@cs.ubc.ca

Icons by noun project and images from pexels

Go forth and analyze!