Establishing a visualization design space
A case study in infectious disease genomic epidemiology

Anamaria Crisan

PhD Candidate, Computer Science

University of British Columbia

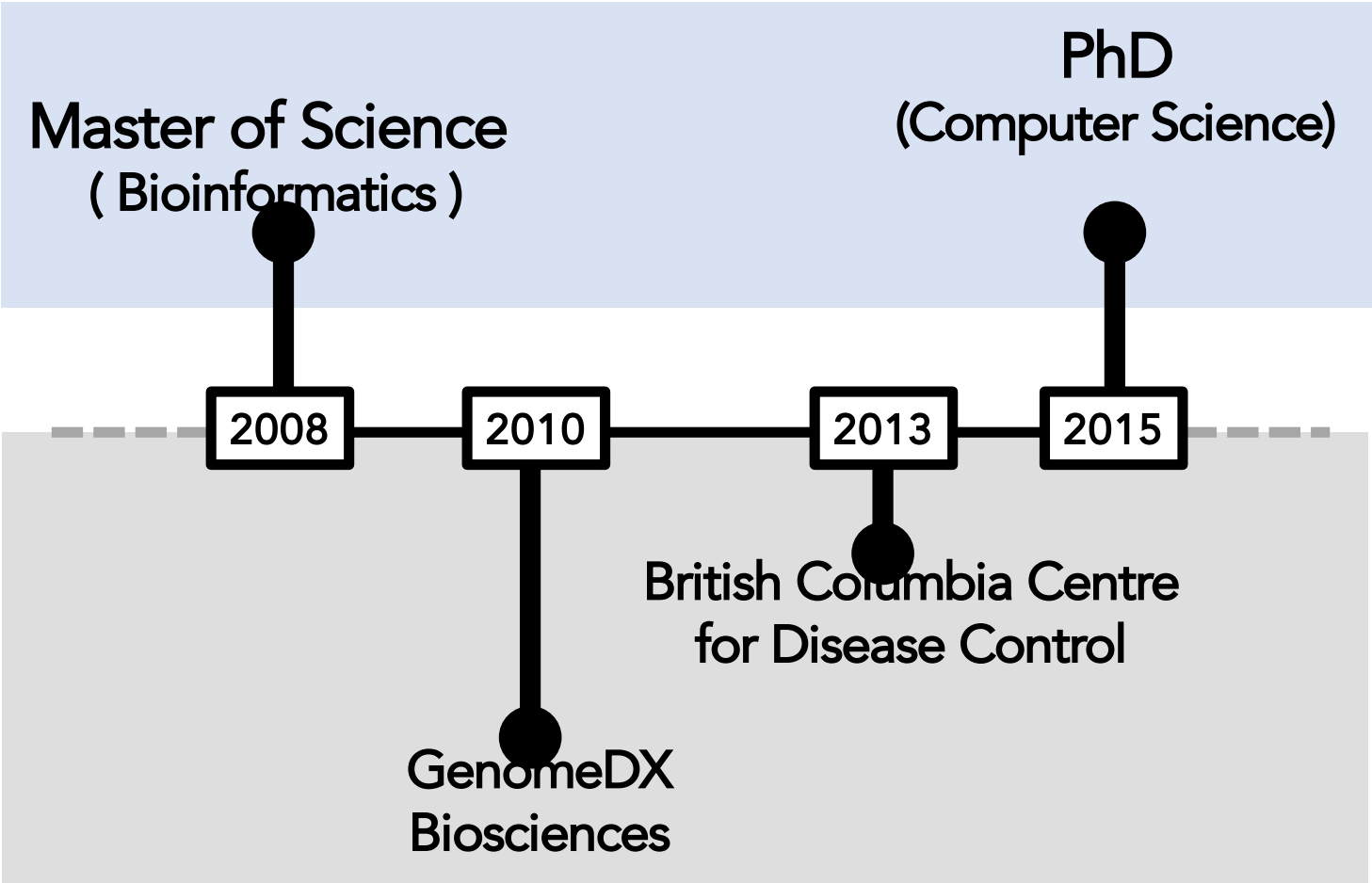
https://doi.org/10.1101/325290

@amcrisan *acrisan@cs.ubc.ca* *http://cs.ubc.ca/~acrisan*

PhD Candidate, Computer Science
University of British Columbia

Thesis: Visualizing Public Health Data

Advisors: Dr. Tamara Munzner
Dr. Jennifer Gardy



What we'll talk about

Why should we visualize data?

Thinking systematically about data visualization

GEViT: a Genomic Epidemiology Visualization Typology

Why should we visualize data?

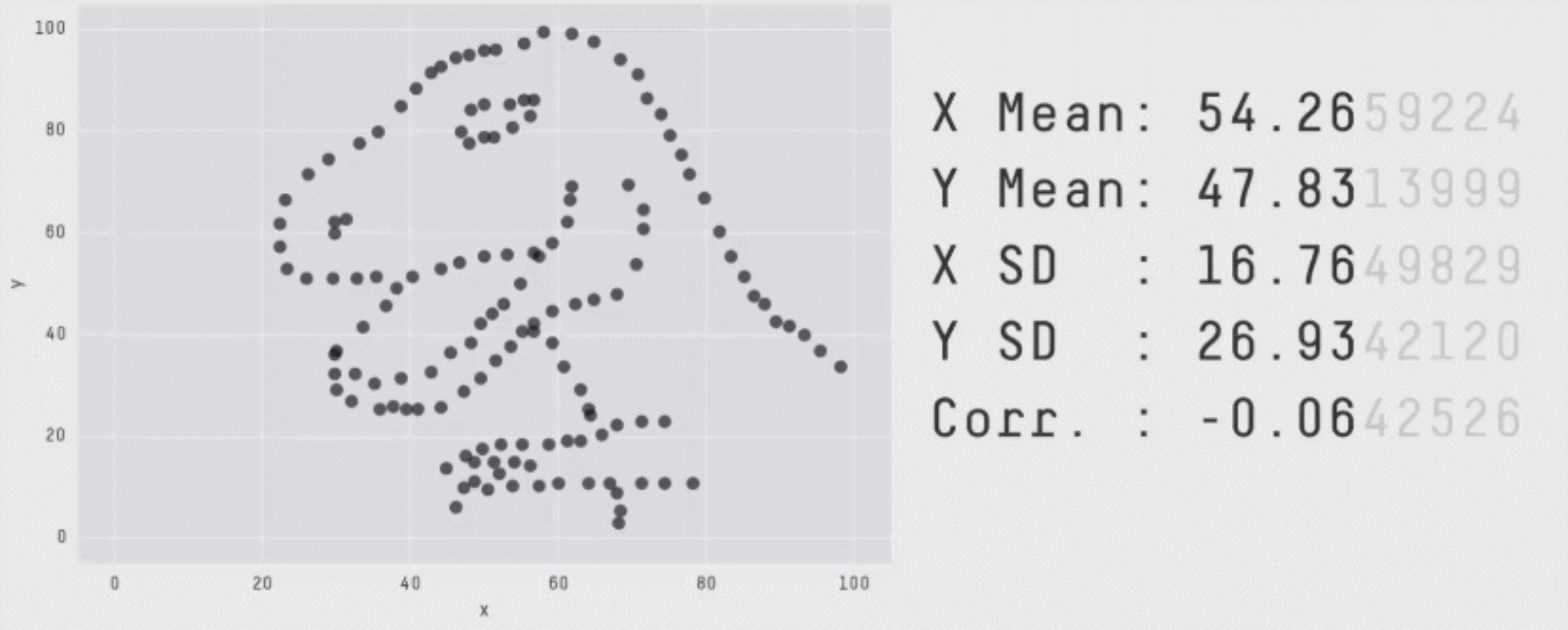
Statistics is not always the answer

```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

Autodesk Research (2017). *Same Stats, Different Graphs: https://www.autodeskresearch.com/publications/samestats*

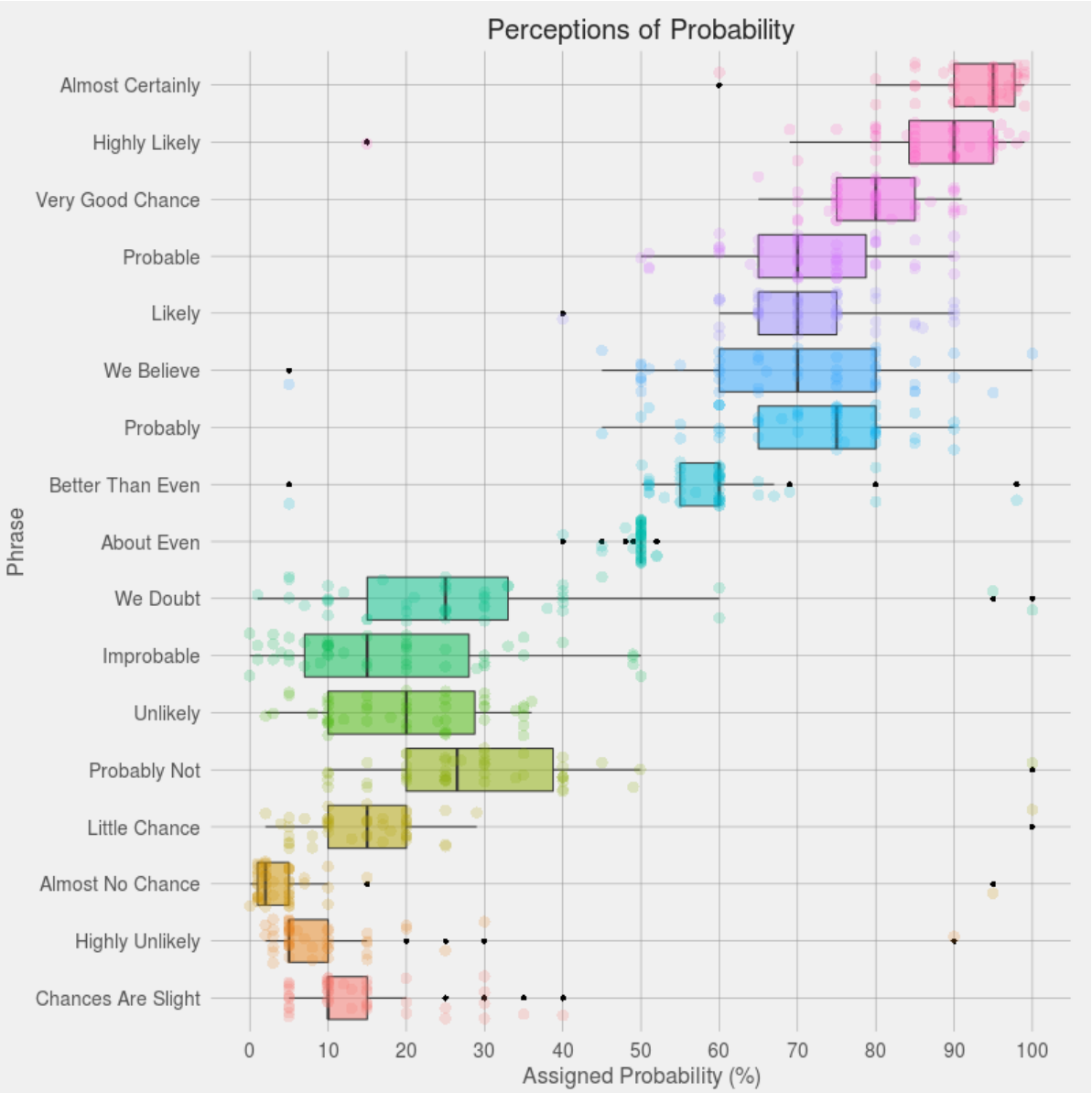
Statistics is not always the answer

Same stats, different graphs (*Datasaurus*)



Autodesk Research (2017). *Same Stats, Different Graphs: https://www.autodeskresearch.com/publications/samestats*

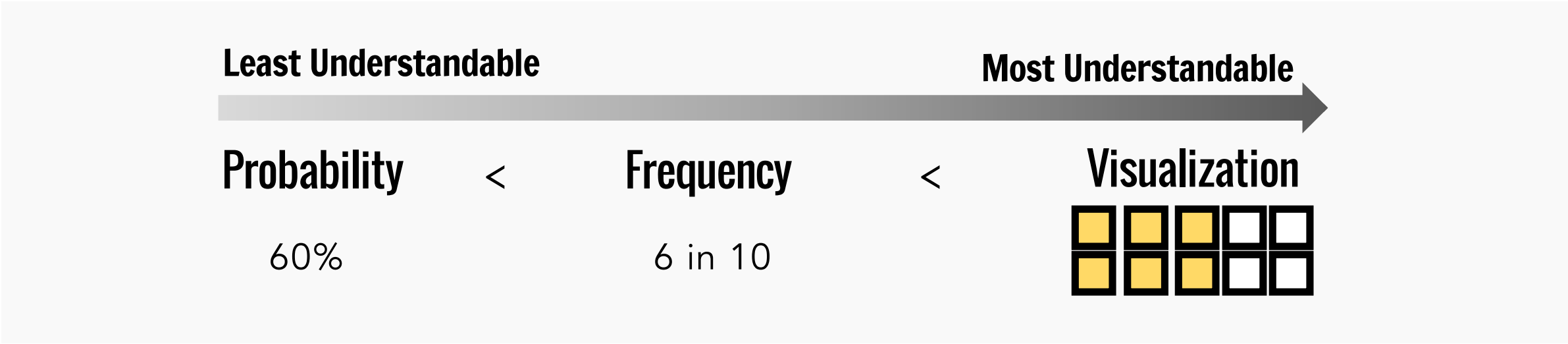
Humans interpret numerical information differently



It is not always easy to reason consistently with numbers

http://bit.ly/1FxtT2z

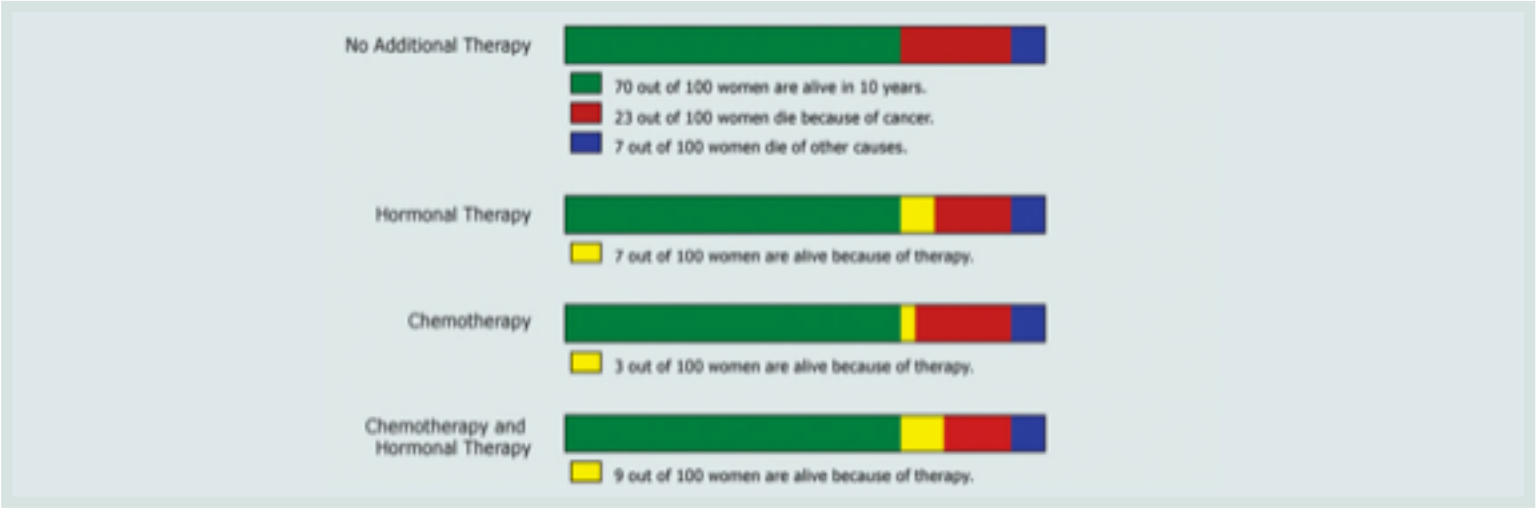
Humans interpret numerical information differently



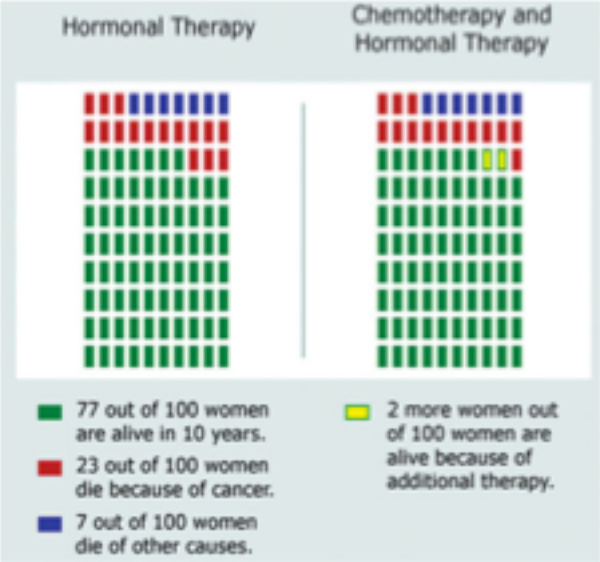
- Numeracy : the ability to reason with numbers
  - Individuals with low numeracy have a difficulty interpreting numbers and probabilities
  - Also true amongst educated professionals
- Visualization can make data more accessible to individuals with lower numeracy skills

Whiting (2015) "How well do health professionals interpret diagnostic information? A systematic review"

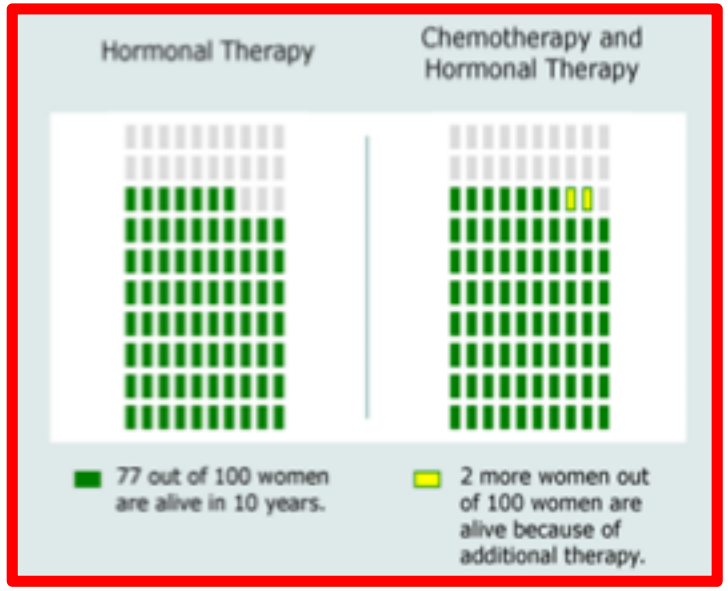
Need the *right* data visualization, not just *a* visualization



Alternative 1



Alternative 2



Zikmund-Fisher (2013). A demonstration of "less can be more" in risk graphics.

The state of visualization in public health

- Greater understanding that visualizing data is important
- Barriers for creating data visualizations are *lowering*
  - Many domain specialists (scientists, public servants) routinely create data visualizations
- Guidance on what makes a good data visualization is *absent*
  - Making the right data visualization for some specific context is harder problem
  - Few people outside of infovis read the infovis literature
- Lack of guidance = ineffective *ad hoc* solutions

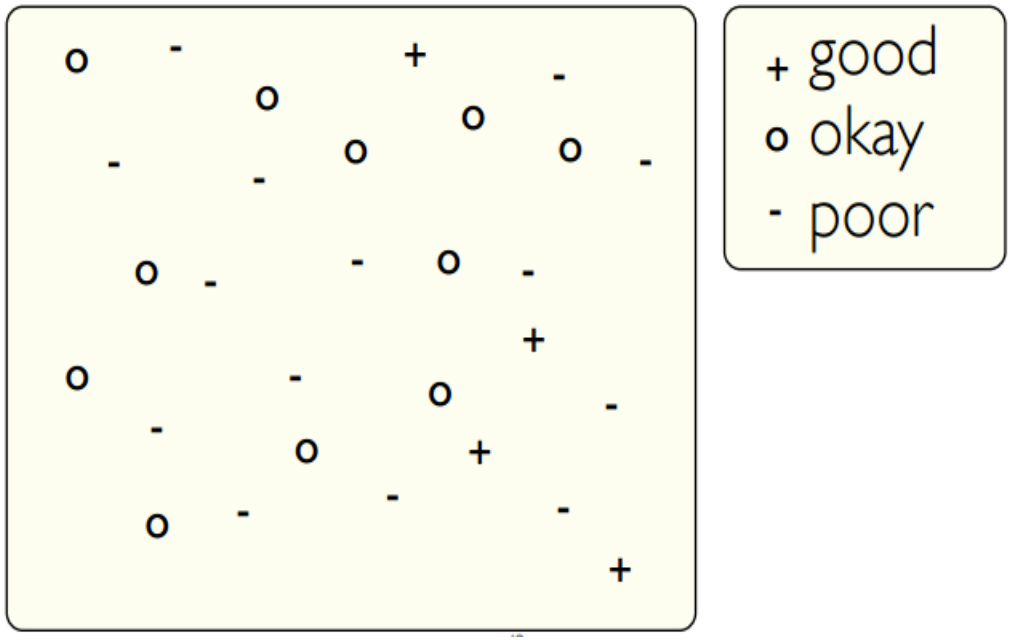
The state of visualization in public health

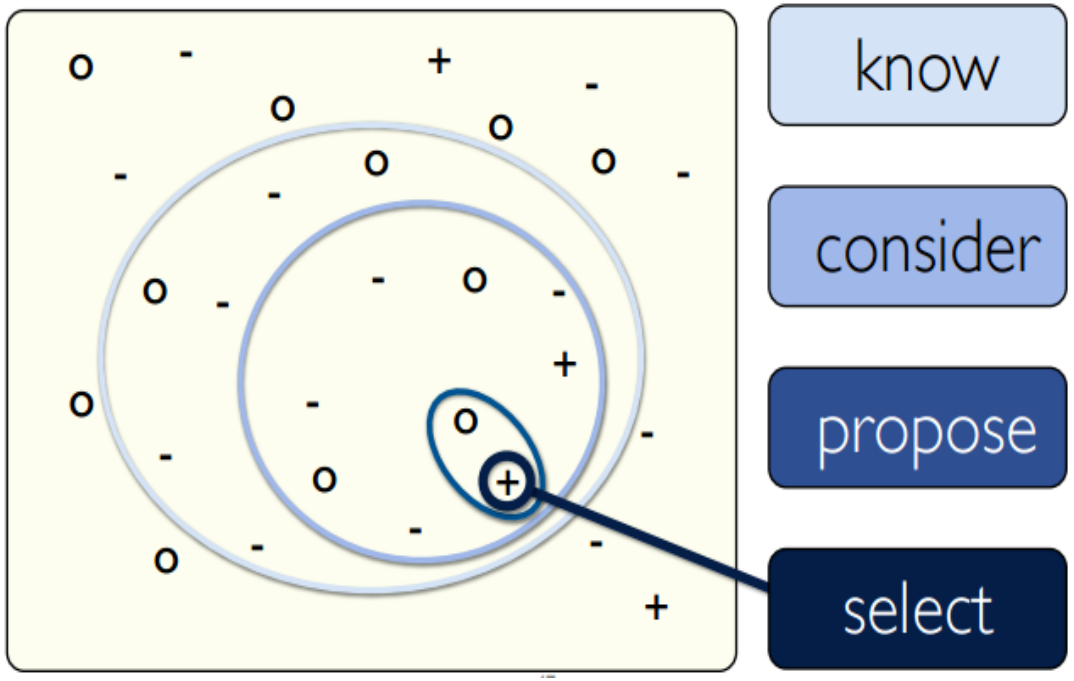
- Greater understanding that visualizing data is important
- Barriers for creating data visualizations are *lowering*
  - Many domain specialists (scientists, public servants) routinely create data visualizations
- Guidance on what makes a good data visualization is *absent*
  - Making the right data visualization for some specific context is harder problem
  - Few people outside of infovis read the infovis literature
- Lack of guidance = ineffective *ad hoc* solutions
- **Our proposed solution: systematically create an explorable vis design space**
  - Shows what is possible with visual design
  - Can help make the search for good visualizations easier

Thinking systematically about visualization design

Design Spaces : A quick primer

Design spaces are made of visualization design choices or varying utility (+ 0 -)





Source: Sedlmair (2012) "Design Study Methodology"

We have some intuition on design choices

- All images below show chairs, but they have different designs
- All chairs can be used for a common task: sitting
- But – fundamentally, different chairs are suited for different contexts

Not suitable as an office chair (-)

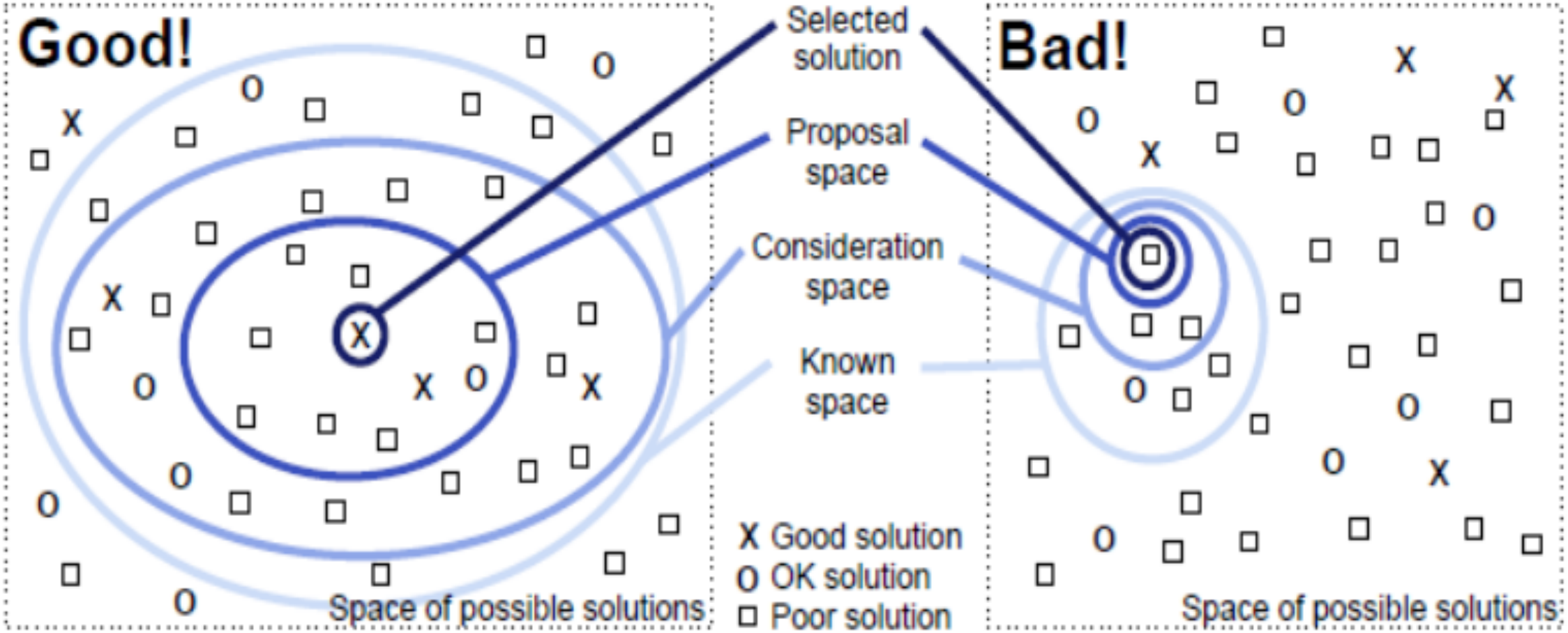
Suitable as an office chair (+,0)

Design Spaces : A quick primer

GOAL – nudge domain specialists toward better design choice solutions

(i.e. shop around and find the best chair)

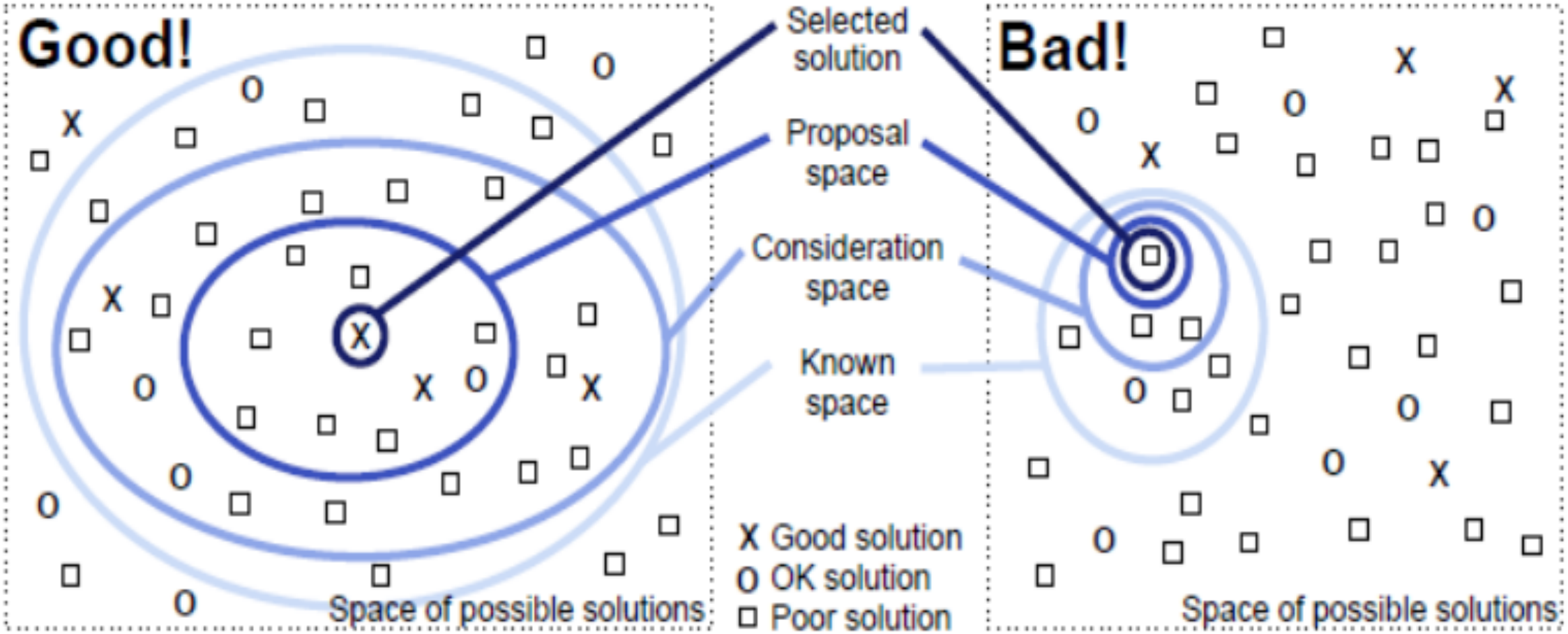
(i.e. choosing the nearest & cheapest chair)



Source: Munzner (2014) "Visualization Analysis and Design"

Design Spaces : A quick primer

BUT – how do we **systematically** describe design space to promote good exploration?

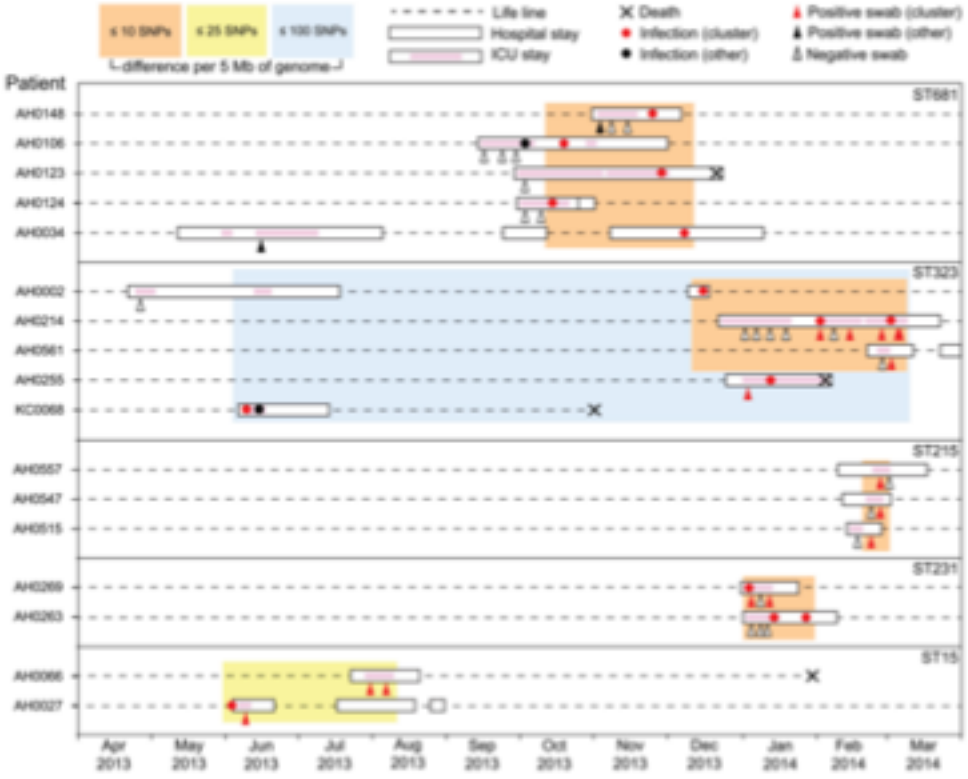


Source: Munzner (2014) "Visualization Analysis and Design"

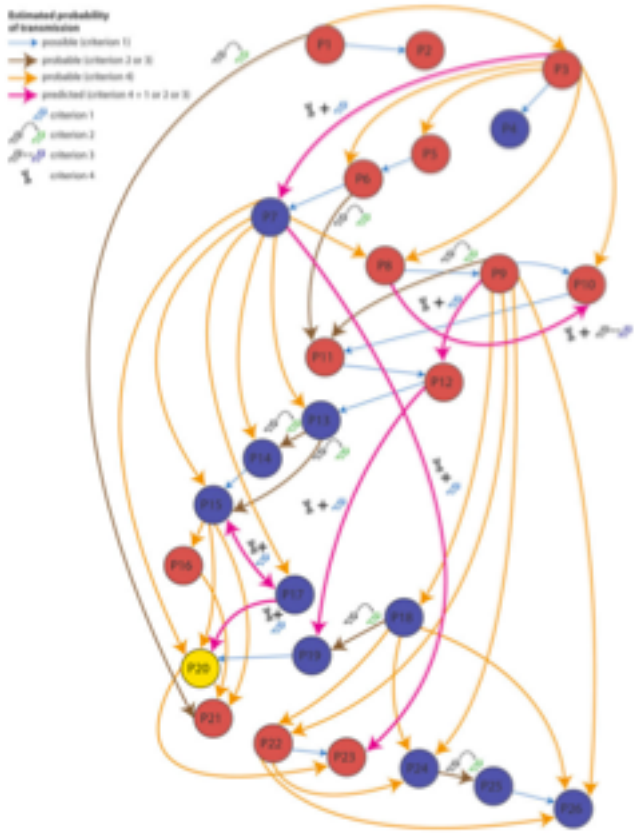
Considering a design space for public health

- There is considerable variability in public health visualization design
- *Example:* all visualizations below show a hospital transmission

Gorrie (2017)



Willman (2015)



Davis (2015)

Creating an explorable visualization design space

https://doi.org/10.1101/325290

An overview of our approach

- Can we capture the extent of variability in visualization design?

An overview of our approach

- Can we capture the extent of variability in visualization design?
- *Why* should we do this?
  - Initially, we were simply curious about how others visualize data
  - Then we realized how useful and powerful having the design space was

An overview of our approach

- Can we capture the extent of variability in visualization design?
- *Why* should we do this?
  - Initially, we were simply curious about how others visualize data
  - Then we realized how useful and powerful having the design space was
- *How* do we do this?
  - No methods in vis literature for systematic construction of a design space
  - Image classification missed important contextual data (no useful training data)

An overview of our approach

- Can we capture the extent of variability in visualization design?
- *Why* should we do this?
  - Initially, we were simply curious about how others visualize data
  - Then we realized how useful and powerful having the design space was
- *How* do we do this?
  - No methods in vis literature for systematic construction of a design space
  - Image classification missed important contextual data (no useful training data)
- *What we did:* developed a method for systematically constructing design spaces!
  - Influenced by literature reviews in medical literature
  - Added machine learning with qualitative methods
  - Applied this method to infectious disease genomic epidemiology (IDGE)

Our approach allows us to answer three different questions

Literature Analysis	WHY are researchers visualizing data?
Qualitative Data Visualization Analysis	HOW are researchers visualizing data, WHAT are they visualizing?
Quantitative Data Visualization Analysis	HOW MANY examples are there of specific visualizations?

An overview of our approach

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

An overview of our approach

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

An overview of our approach

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

An overview of our approach

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

Three major results from this work

- **Literature Analysis (Why):**
  - Understanding the structure of genomic epidemiology papers
  - Motivated intelligent sampling of data visualizations
  - Primary sources of our design space visualizations

Three major results from this work

- **Literature Analysis (Why):**
  - Understanding the structure of genomic epidemiology papers
  - Motivated intelligent sampling of data visualizations
  - Primary sources of our design space visualizations
- **Qualitative Analysis (What, How):**
  - Manually analyzed figures to classify elements of data visualizations
  - Generated a Genomic Epidemiology Visualization Typology (GEViT)

Three major results from this work

- **Literature Analysis (Why):**
  - Understanding the structure of genomic epidemiology papers
  - Motivated intelligent sampling of data visualizations
  - Primary sources of our design space visualizations
- **Qualitative Analysis (What, How):**
  - Manually analyzed figures to classify elements of data visualizations
  - Generated a Genomic Epidemiology Visualization Typology (GEViT)
- **Quantitative Analysis (How many):**
  - In IDGE it's nearly all trees and a surprising amount of tables

Developing and Operationalizing GEViT

How can we systematically describe images?

- What does GEviT do and <u>not</u> do?

GEViT provides a base

- Deliverables :
  1. Typology
  2. Interactive Gallery

GEViT does not evaluate

- Massive undertaking that would take many years
- Needs GEViT to conduct evaluations

- How can GEViT be used?
  - Concise descriptions to discuss data visualizations
  - Understand what visualizations are common and possible
  - Get ideas for data visualization design

[WHY] Literature Analysis

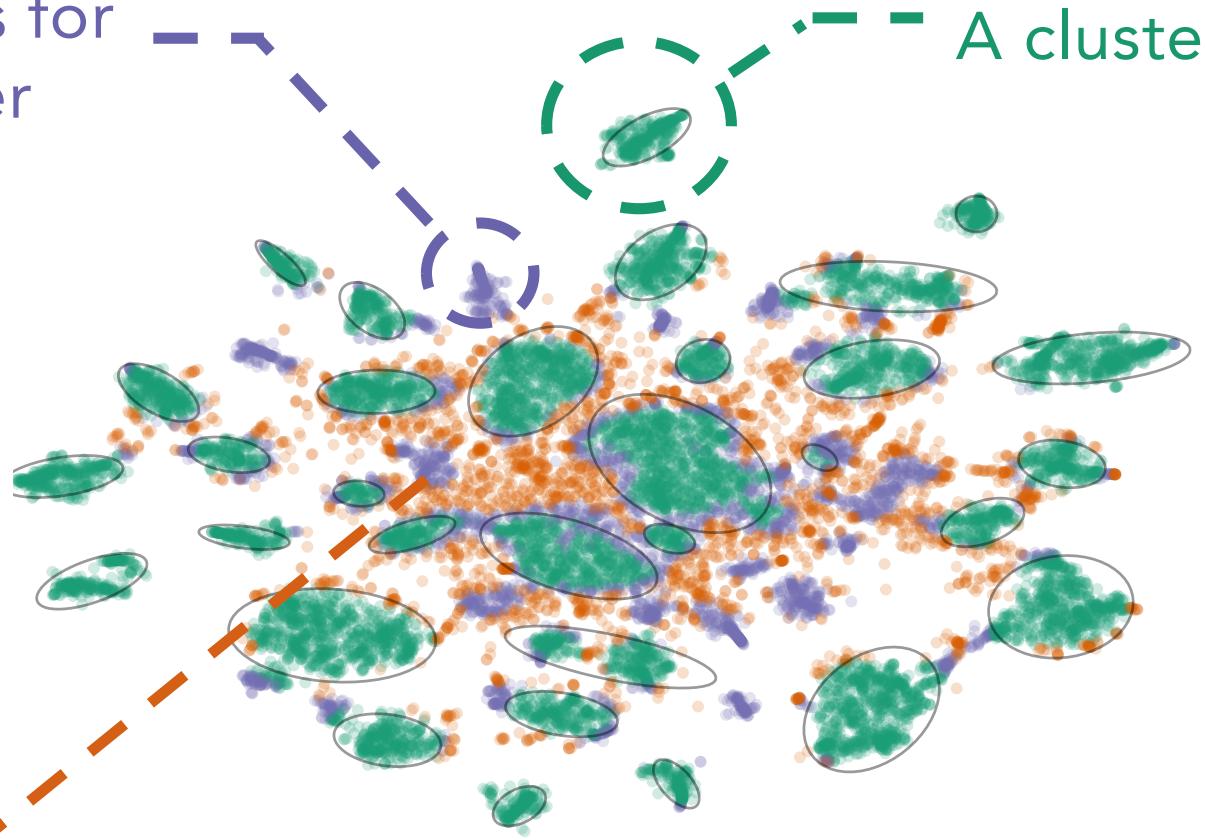
Setting up the visualization design space

Overview of exploration and sampling of the document corpus

17,974	All documents
15,315	Removal of “very noisy” articles

Article acquisition & unsupervised clustering

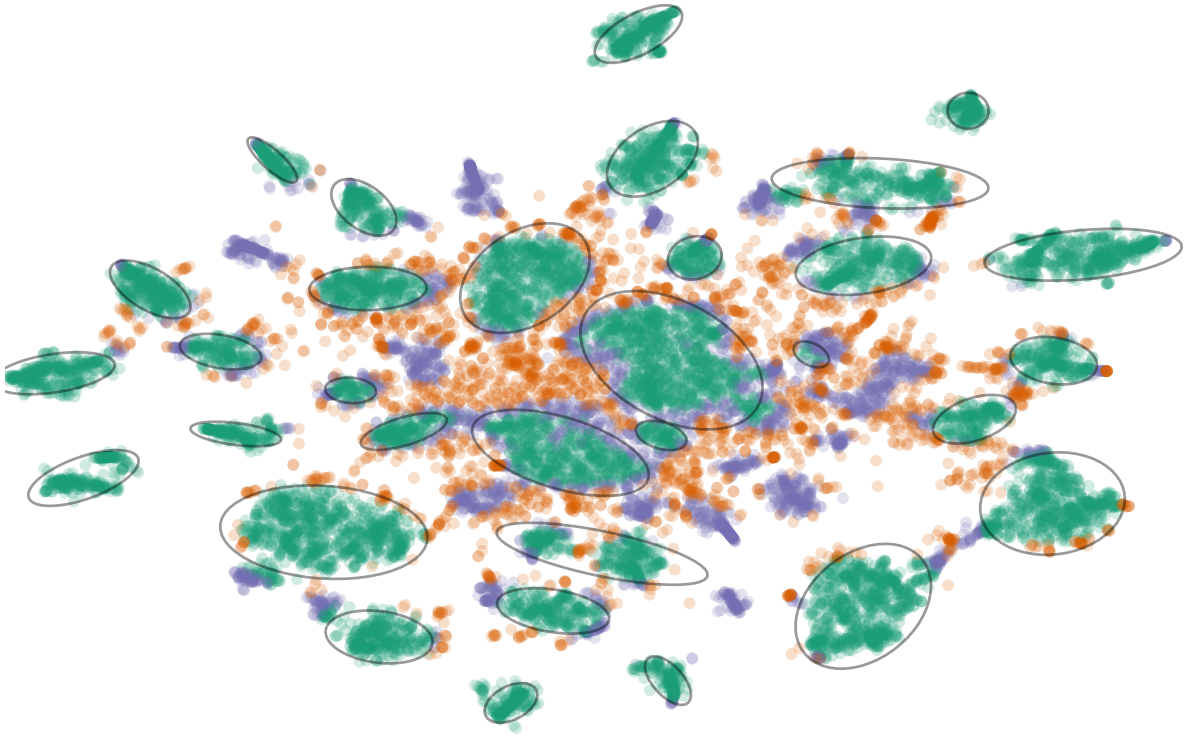
Literature analysis: discovering topics in the document corpus



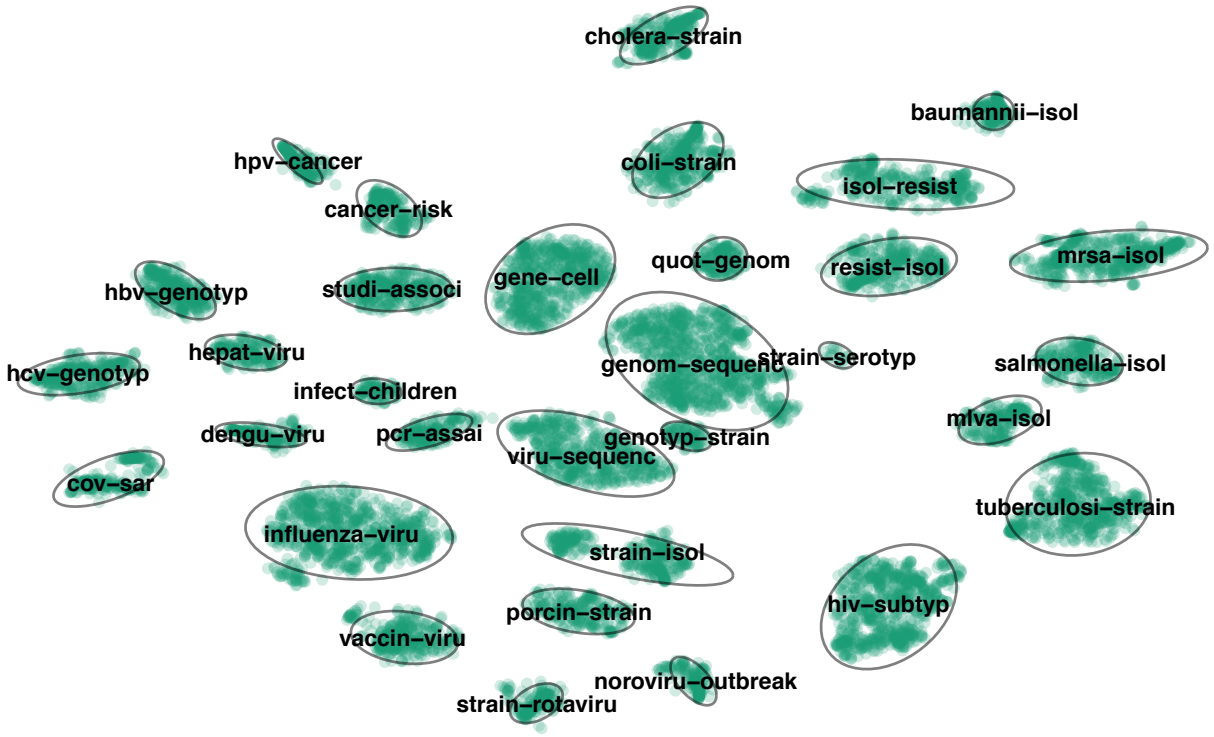
N = 15,315 (very noisy articles removed)

Literature analysis: discovering topics in the document corpus

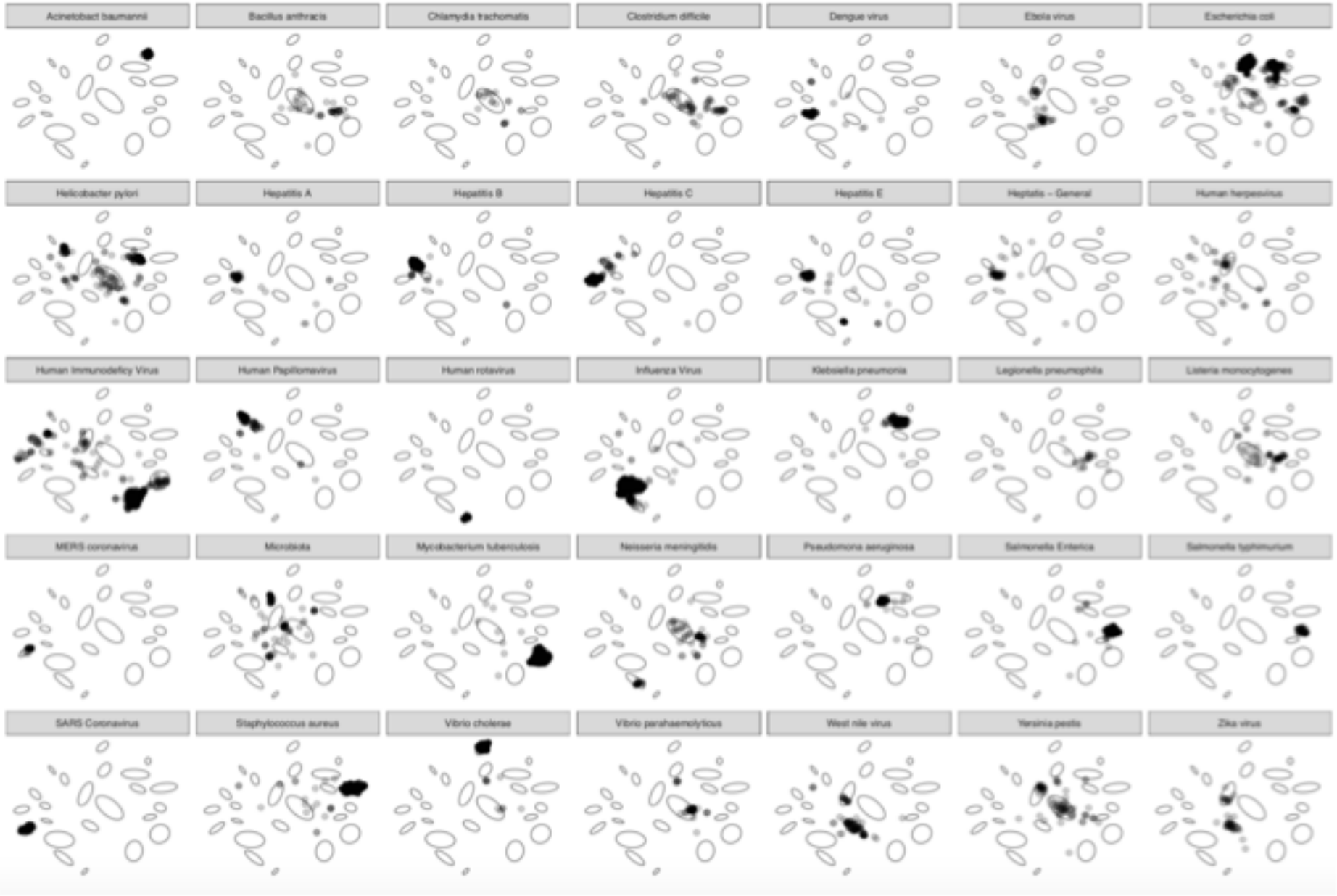
Topic clustering results



Clusters for sampling



Literature analysis: verifying the results



Overview of exploration and sampling of the document corpus

17,974	All documents
15,315	Removal of "very noisy" articles
9,551	Articles about human pathogen
6,350	Common human pathogen articles

Article acquisition & unsupervised clustering

Limit to common human pathogens
Apply *a priori* concepts

Linking pathogen clusters to a priori concepts

- Wanted to include addition public health concepts
- Topics assessed *a priori*, assigned to common terms between clusters

Molecular Biology Concepts

- Characterization
- Diversity
- Drug Resistance
- Genome
- Genotype
- mBio
- Phylogeny
- Reservoir
- Vector

Epidemiology Concepts

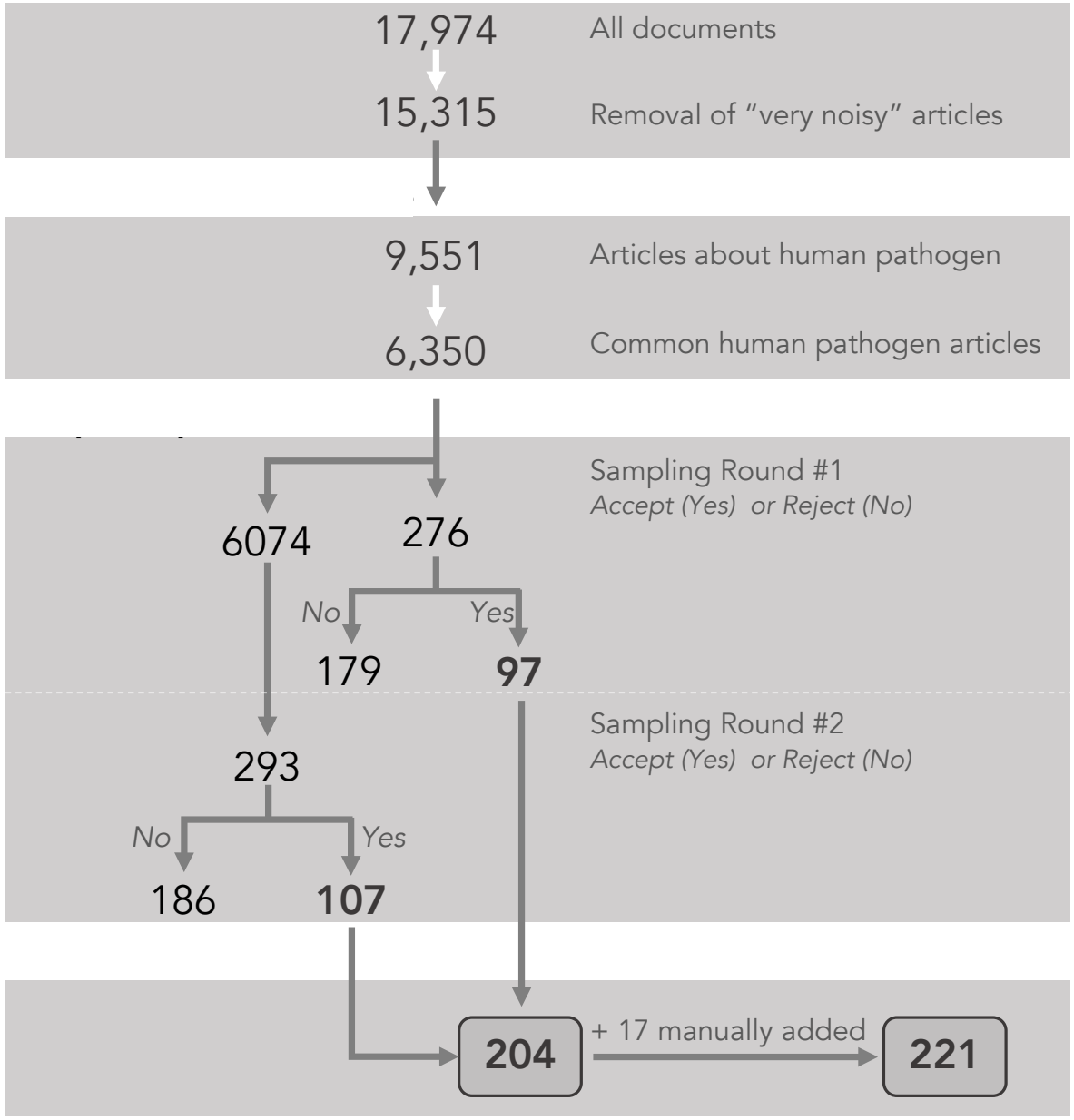
- Cluster
- Geography
- Outbreak (International / Community/ Hospital)
- Surveillance
- Transmission
- Vaccine

Medical Concepts

- Clinical
- Cancer
- Diagnosis
- Outcome
- Treatment

23 a priori concepts in total

Overview of exploration and sampling of the document corpus



Article acquisition & unsupervised clustering

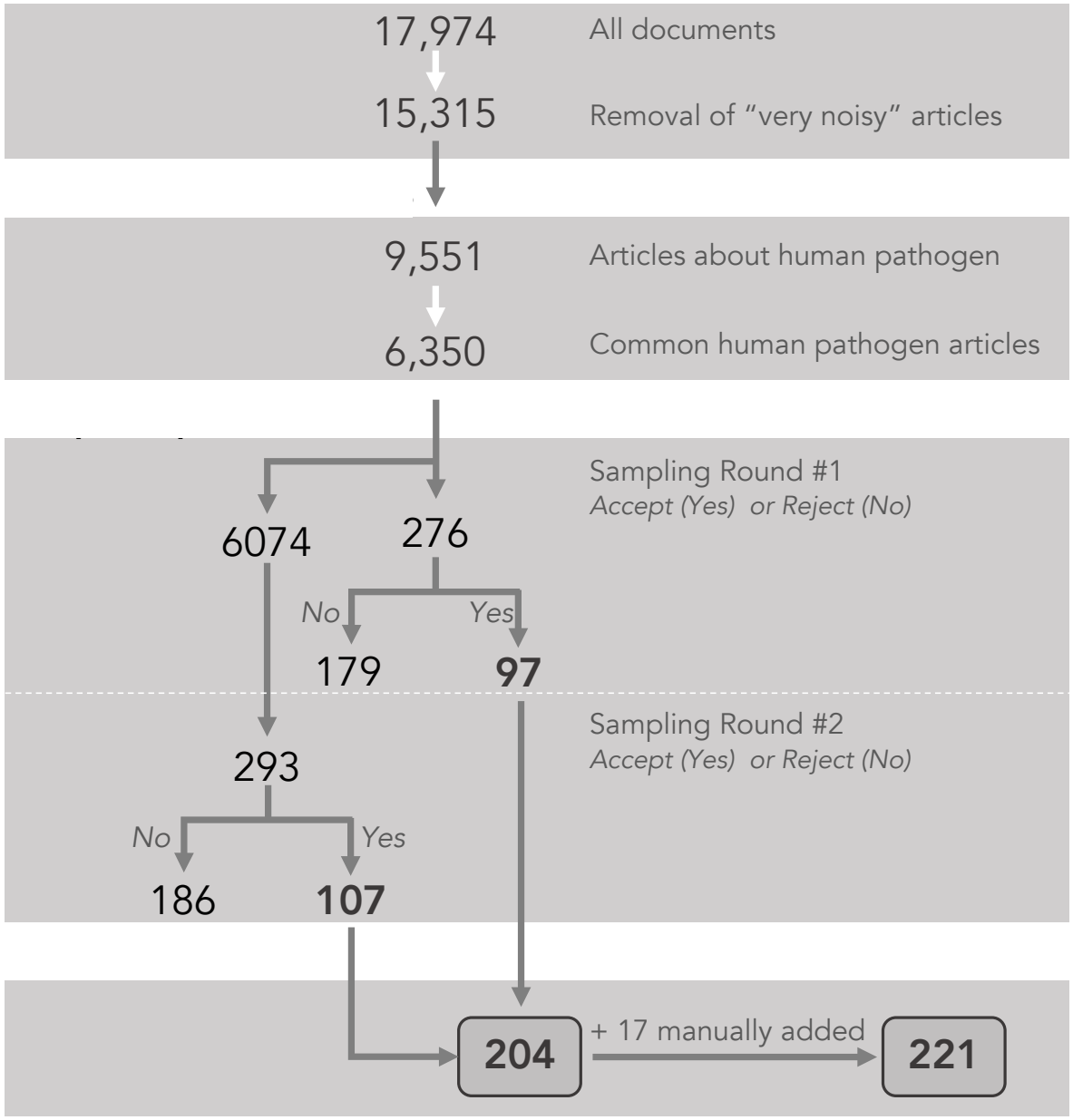
Limit to common human pathogens
Apply a priori concepts

Sampling of articles (two rounds)

- Random stratified sampling
  - Strata: pathogen, *a priori* concepts

Yielded 801 figures and 49 tables

Overview of exploration and sampling of the document corpus



Article acquisition & unsupervised clustering

Limit to common human pathogens
Apply a priori concepts

Sampling of articles (two rounds)

- Random stratified sampling
  - Strata: pathogen, *a priori* concepts

Basis for further analysis

Yielded 801 figures and 49 tables

Qualitative Analysis

Manually analyzed paper figures to classify elements of data visualizations, derived GEViT

Developing 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 : GEViT, a hierarchical code set with separate taxonomies for:
  - Chart Types
  - Chart Combinations
  - Chart Enhancements

Chart Type: a foundational element of all data visualizations

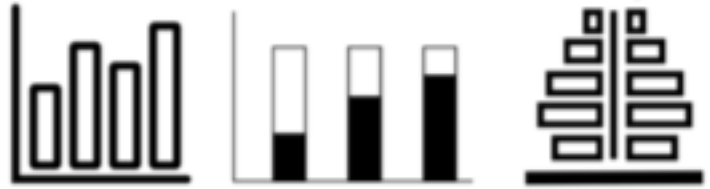
- Self explanatory what chart types are…
- Manually classified every single type of chart in every figure
- Report only what we found in the sample document corpus
- There were six classes of charts types
  - Charts also had special chart types (i.e. epidemic curve is a special case of bar chart)

Chart Type: a foundational element of all data visualizations

Common Statistical Charts

Bar Chart

Standard *Stacked* *Divergent*



Special Cases
- Epidemic Curve
- Diversity Chart
- LefSe Plot

Line Chart

Special Cases
- Bootscan
- Kaplan-Meier
- Skyline Plot

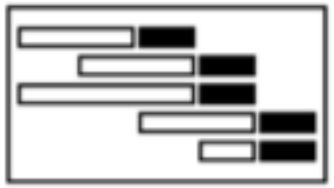
Scatter Plot

Special Cases
- Root-to-tip
- Ordination Plot
- Q-Q plot

Pie Chart

Venn Diagram

Timeline



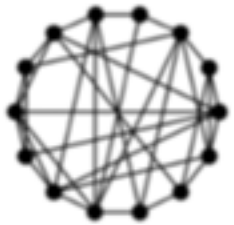
Distribution Plot

Histogram *PDF* *Boxplot* *Swarm Plot*

Chart Type: a foundational element of all data visualizations

Relational Charts

Node-link



Special Cases

- *eBurst*
- Social network
- Molecular network
- Minimum Spanning Tree

Flow Diagram

Chord Diagram

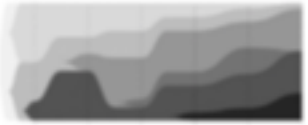
Sankey Diagram

Area Charts

Streamgraph*

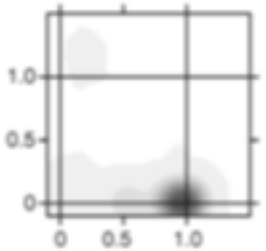
Absolute

Relative



Heatmap

Density Plot*



Spatial Charts

Geographic Map

Choropleth Map

Interior Map

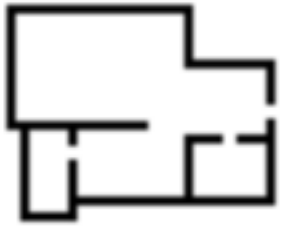


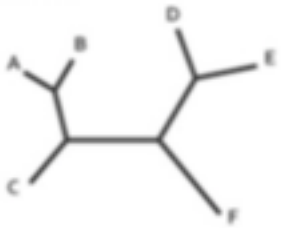
Chart Type: a foundational element of all data visualizations

Tree Chart Types

Phylogenetic Tree

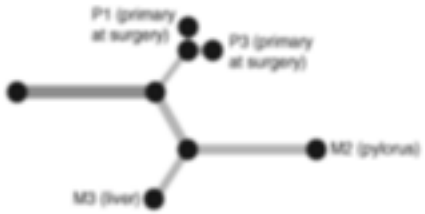
Rooted (Radial & Linear)

Unrooted (Radial & Linear)



Dendrogram

Clonal Tree

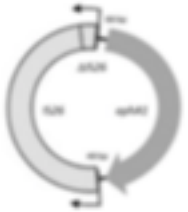


Genomic Chart Types

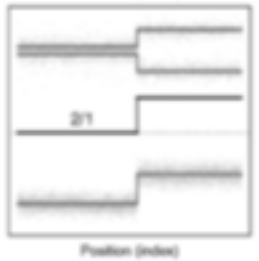
Genomic Structure

Linear

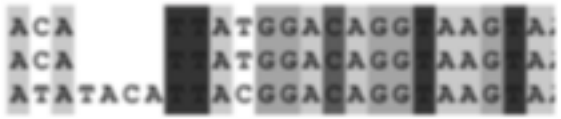
Radial



Variation Profile



Alignment



Sequence Logo Plot

Chart Type: a foundational element of all data visualizations

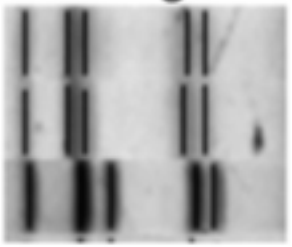
Other Charts

Table

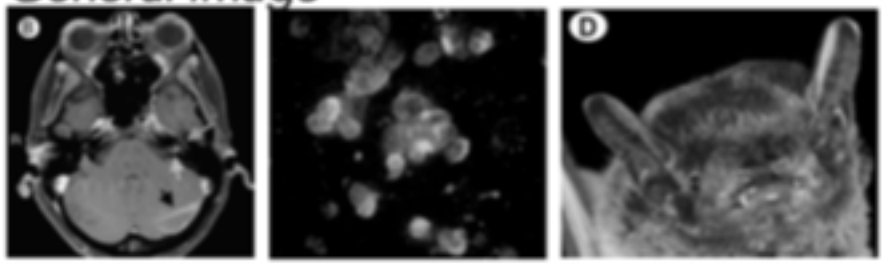
Category Stripe

Image

Gel Image



General Image



Unclassified

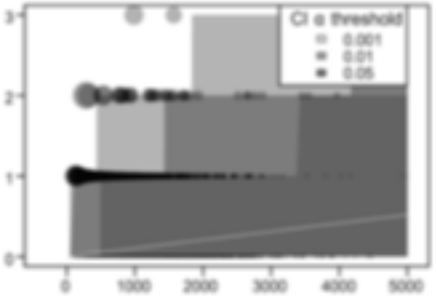
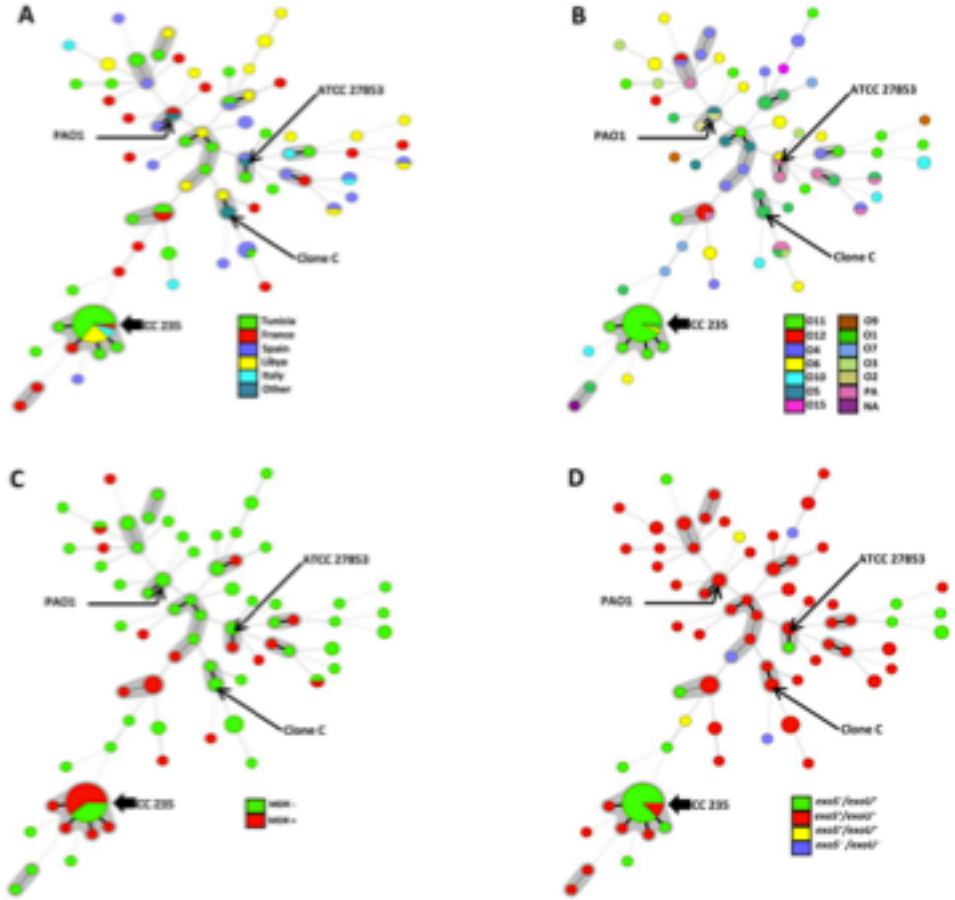


Chart Combinations: showing different aspects of the data

- Observed that charts were combined in a specific, consistent pattern
- We classified every single of chart combinations within a figure

Example: Same chart type, different metadata



Example: Two chart types together

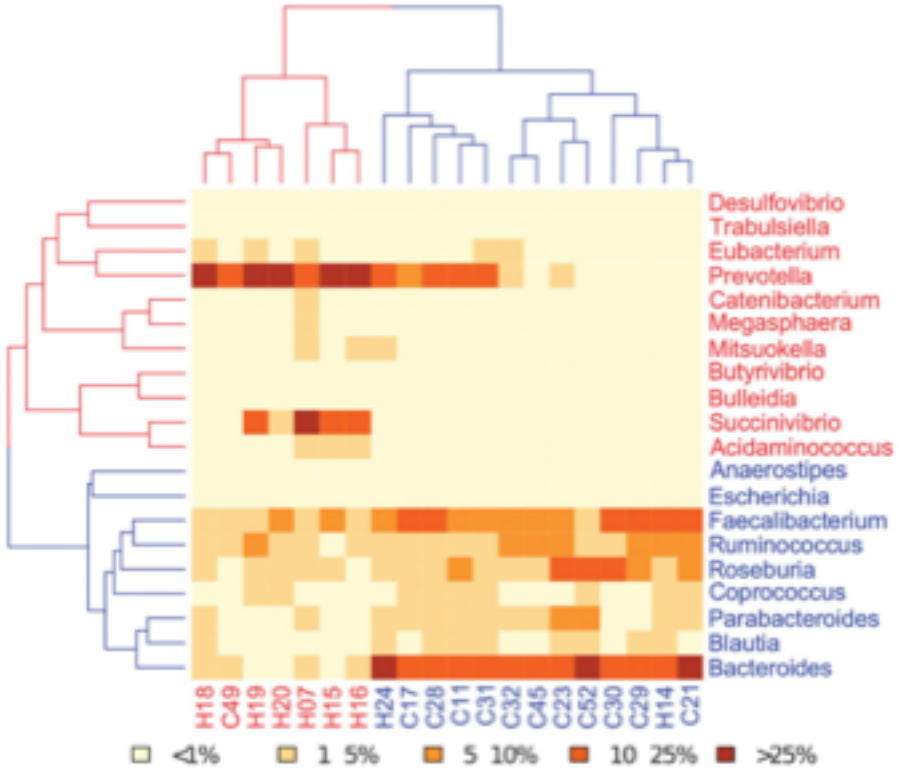
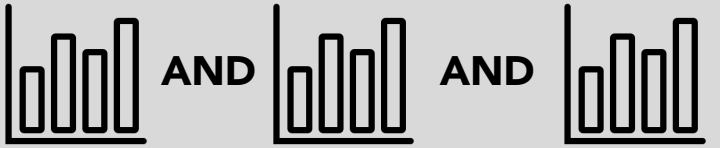
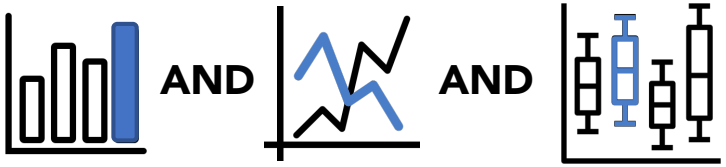


Chart Combinations: showing different aspects of the data

Combination Type	# of chart types	# of charts	Linkage type	*Example*
Simple	1	1	NA	
Composite	Many	1	Spatially Aligned	
Small Multiples	1	Many	Chart Type & Data	
Many Type *Linked*	Many	Many	Visual, but not spatial	
Many Type General	Many	Many	NA	
Complex Combination	Many	Many	Context dependent	

Enhancements : overlaying additional metadata

- Mark = basic graphical element (line, point, area)
- Enhancement = adding marks or re-encoding marks of the base chart type

Add Marks

Adding Additional Marks to base chart type

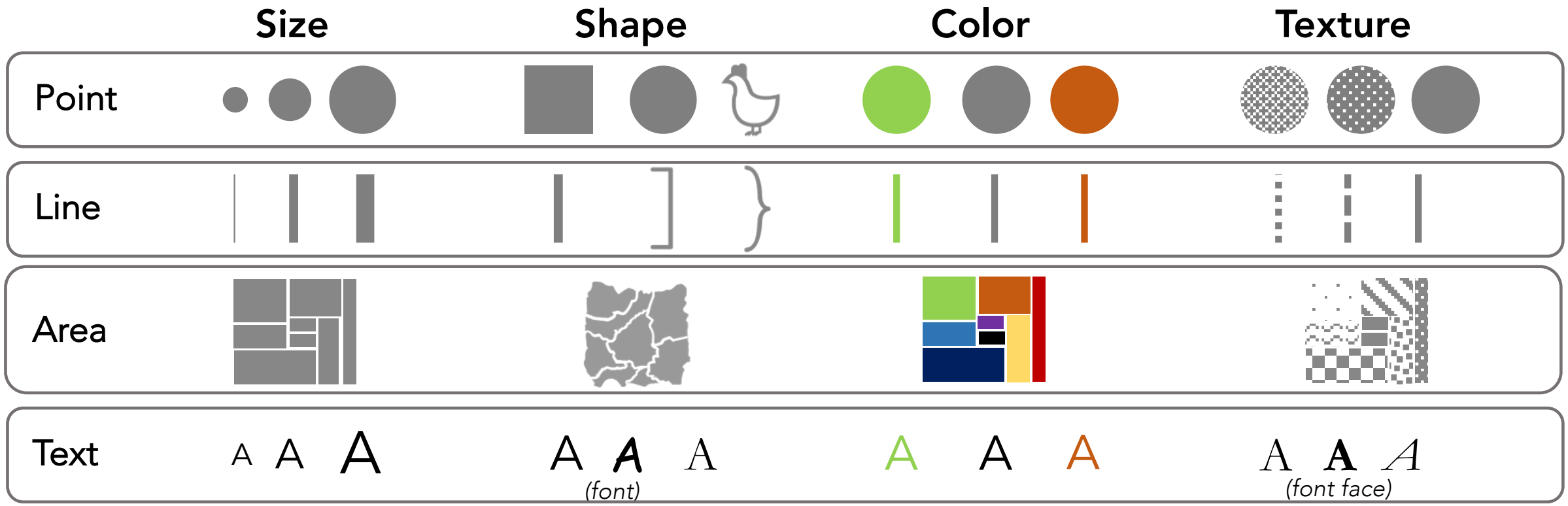
- Point
- Line
- Area Mark
- Text
- Glyph

Re-encode Marks

Re-encode existing marks via channels

- Size
- Shape
- Color
- Texture
- Font

Enhancements : overlaying additional metadata



Special Cases : Containment Mark; Connection Mark; Glyph

Enhancements : overlaying additional metadata

- Structured Enhancement: Encodings are added/changed on many/all marks
- Unstructured Enhancement: Encoding are added/changed to one or a few marks

Structured enhancement

Unstructured enhancement

Add Marks

Adding Additional Marks to base chart type

- Point
- Line
- Area Mark
- Text
- Glyph

Re-encode existing marks

Re-encode existing marks via channels

- Size
- Shape
- Color
- Texture
- Font Face (specific to text)

Add Annotation

Manually Adding annotations

- Same as added marks, but include arbitrary ink too

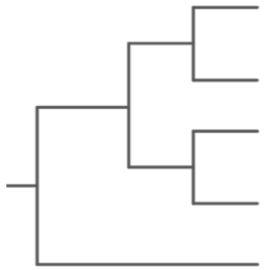
Note: Sometimes the line between adding a mark and adding an annotation is very subtle.

Enhancements : overlaying additional metadata

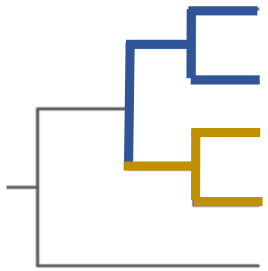
Structured Enhancements

Unstructured Enhancements

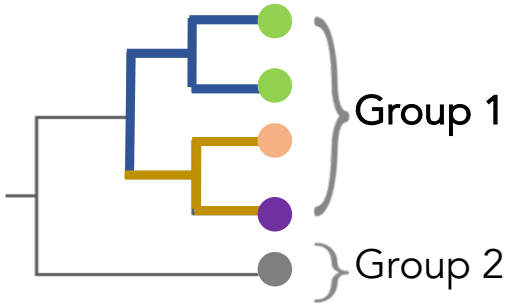
Base Chart
Tree



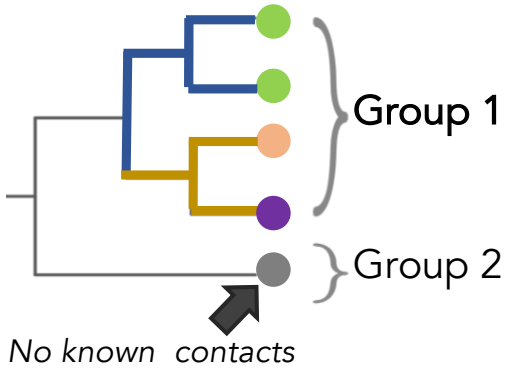
Re-encode Marks
Line - color



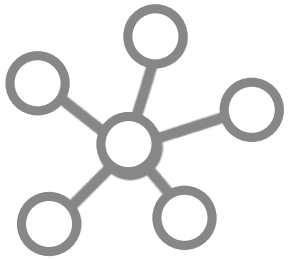
Add Marks
Point, line, & text



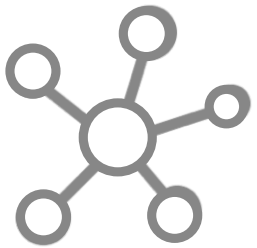
Add Annotation
Arrow, text



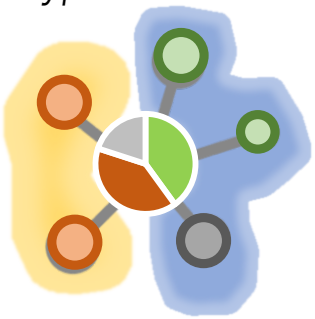
Node-link



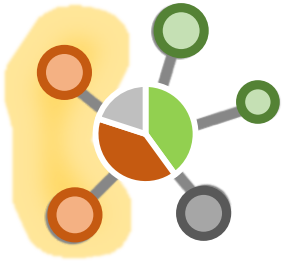
Point - size



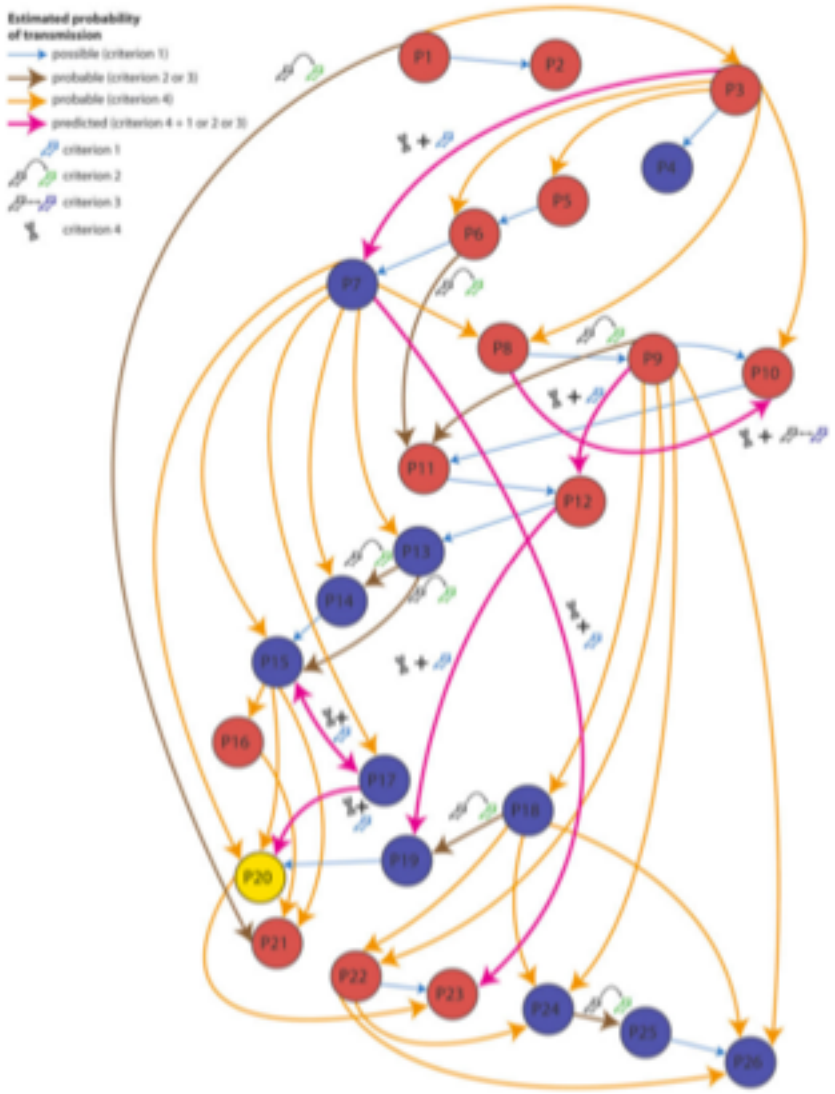
Glyph - Pie Chart



Containment Mark



GEViT in action



Visualization Breakdown

Visualization Context (why)

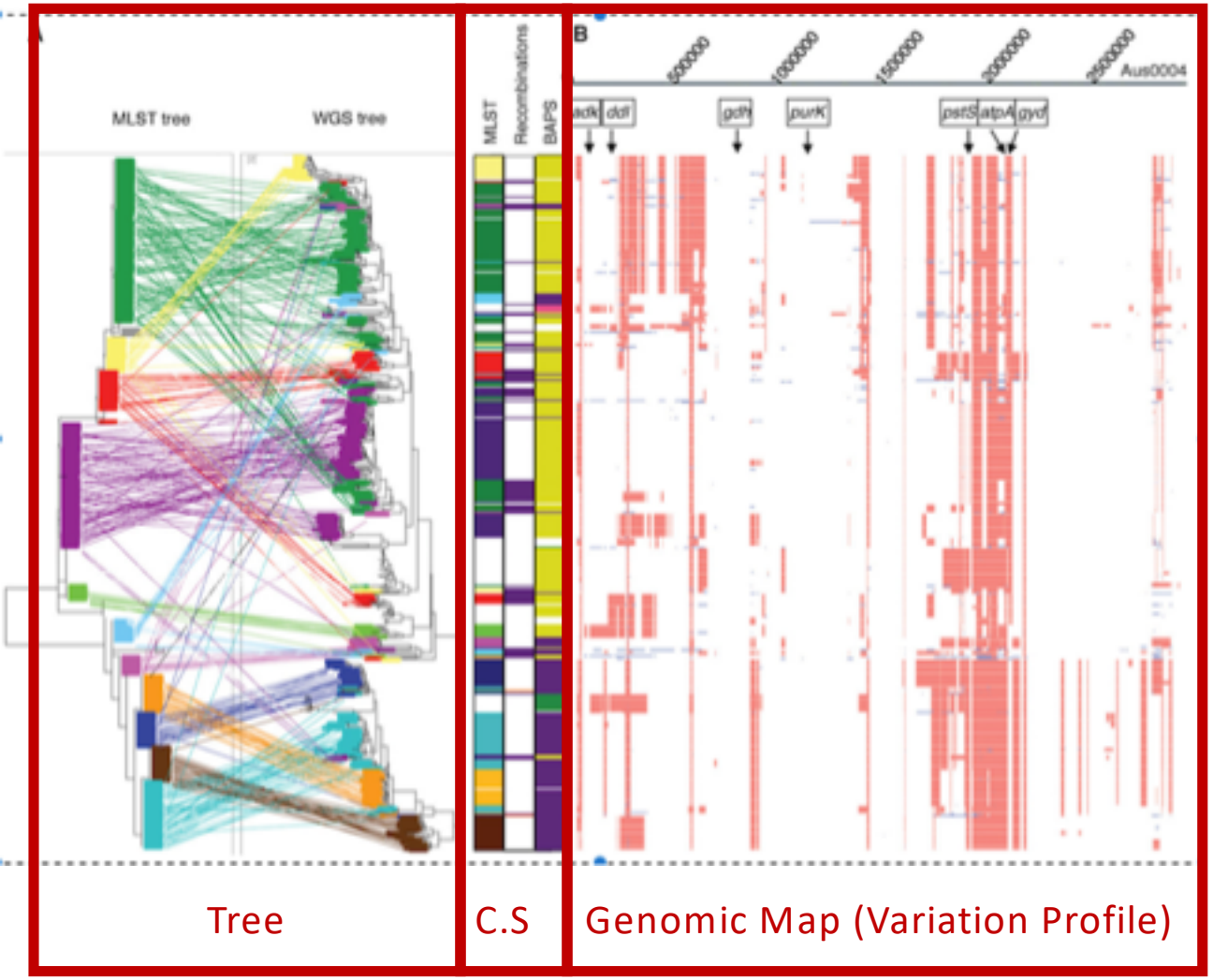
- *Pathogen: Pseudomonas aeruginosa*
- *Concepts: outbreak; control; genome; phylogeny*

Visualization Components (what, how)

Chart Type	Node-link	
Chart Combination	Simple	
Chart Enhancement	Re-encode marks	- Line (colour) - Node (colour)
	Add Marks	Text
	Add Annotations	Icons

GEViT in action

Gorrie (2017)



Visualization Breakdown

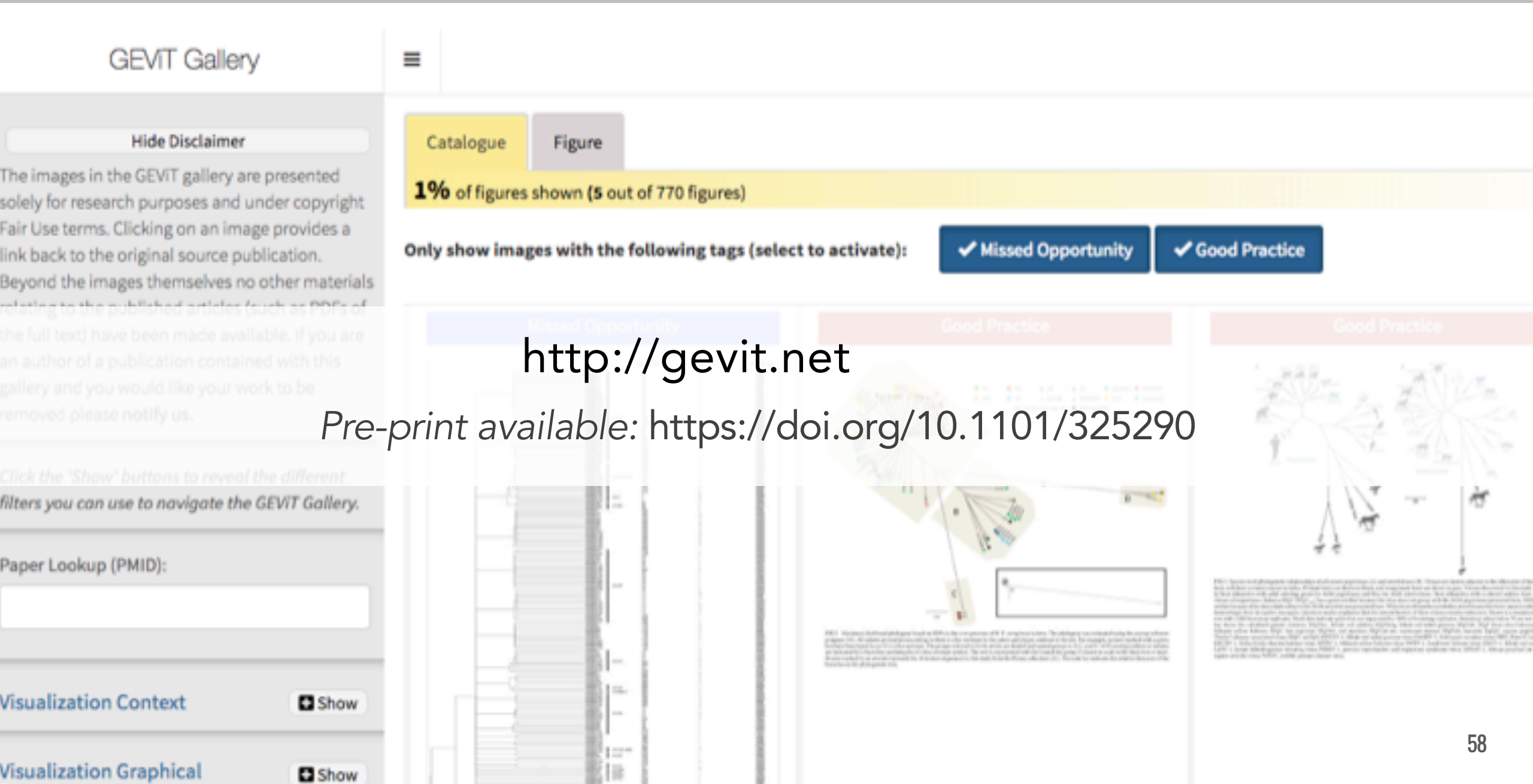
Visualization Context (why)

- *Pathogen: Enterococcus faecium*
- *Concepts: control; genome; outbreak; drug resistance; phylogeny; genotype*

Visualization Components (what, how)

Chart Type	Tree (Rooted Phylogenetic Tree) Category Stripe Heatmap (Variation Profile)	
Chart Combination	Composite *(spatially aligned)*	
Chart Enhancement	Re-encode marks	Tree – *color node branches*
	Add Marks	Tree - *Connection Marks*
	Add Annotations	Genomic Map – *Textboxes*

Putting it all together in the GEViT Gallery



http://gevit.net

Pre-print available: https://doi.org/10.1101/325290

Findings only a systematic approach could detect

- Wide variety of visualization quality 🧐
  - Only possible to assess this with systematic approach
- Most data in a data visualizations are NOT actually visualized 😱
  - Over reliance on tables and text labels
  - Shows lack of visualization design space knowledge
- Current visualizations will not scale for big data 😞
- Many visualizations not understandable by other public health professionals 😩
  - In prior work we conducted a study with public health stakeholders and beyond common statistical charts, stakeholders don't know how to interpret the visualization

GEViT helps to systematically classify images

- What does GEviT do and <u>not</u> do?

GEViT provides a base

- Deliverables :
  1. Typology
  2. Interactive Gallery

GEViT does not evaluate

- Massive undertaking that would take many years
- Needs GEViT to conduct evaluations

- How can GEViT be used?
  - Concise descriptions to discuss data visualizations
  - Understand what visualizations are common and possible
  - Get ideas for data visualization design

The importance of our findings

Implications of our research findings

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

Implications of our research findings

- Need to move away from ad hoc visualization development
  - Need awareness of design space
  - Need to know what is possible, common, and event 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 research findings

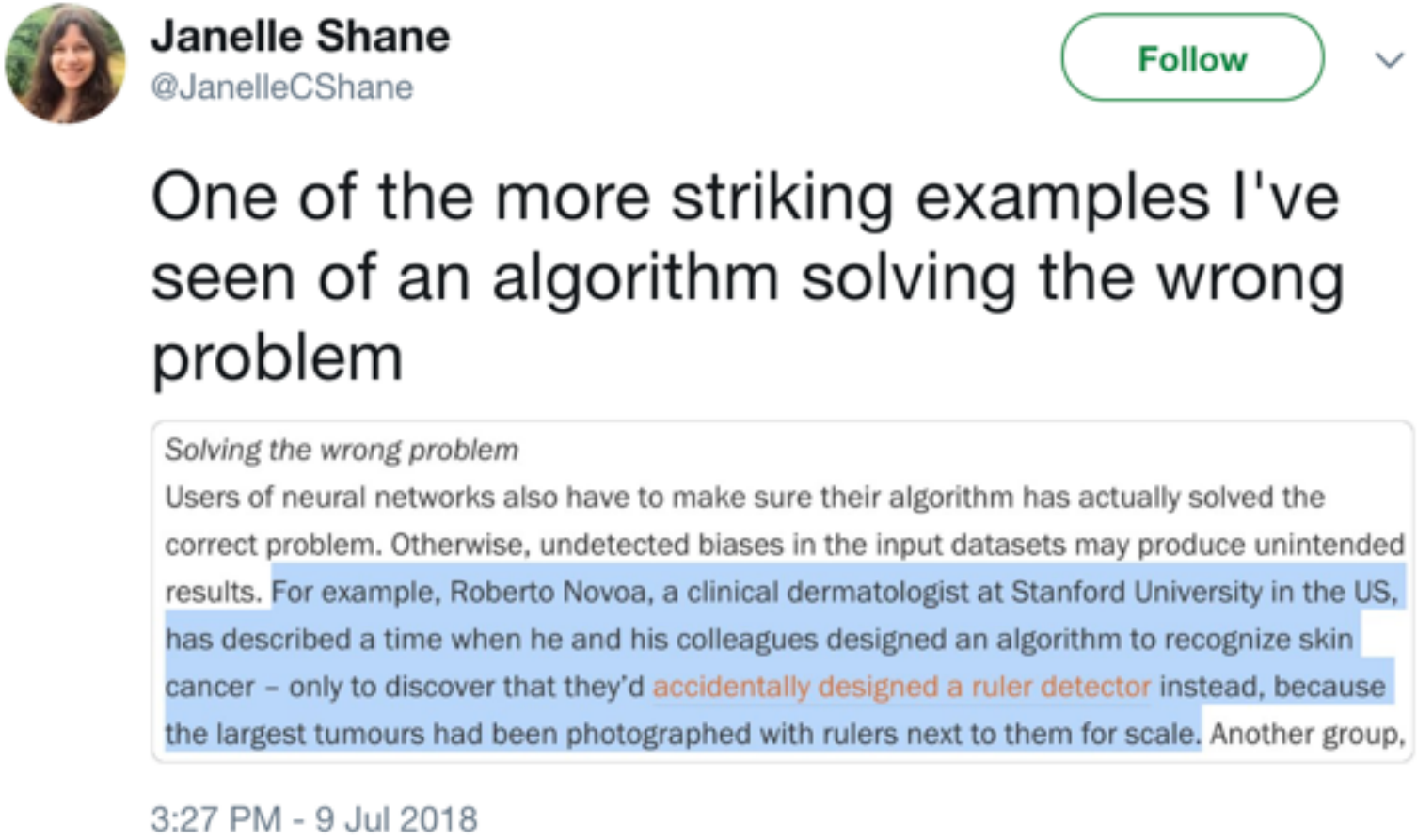
- Need to move away from ad hoc visualization development
  - Need awareness of design space
  - Need to know what is possible, common, and event absent
- Implications for bioinformatics and data visualization tool development
  - Need tools that support complexity and expressivity in visual design
  - Provides design alternatives for bioinformaticians to explore and test
- Implications for education
  - GEViT as a teaching tool (I am already doing this)
  - Design space variance tells you easy/hard it for a community to adapt new data vis
  - Source of inspiration for researchers

Next Steps: operationalizing GEViT further

- Using GEViT helping some public health stakeholders make better visualizations
  - Applications for general public are complex, better stick to technical stakeholders for now
- Shannah Fisher (undergraduate summer research student) helping to create an R package implementation of GEViT
  - Below : sorting out composite algorithm code

Next Steps: automating design space construction

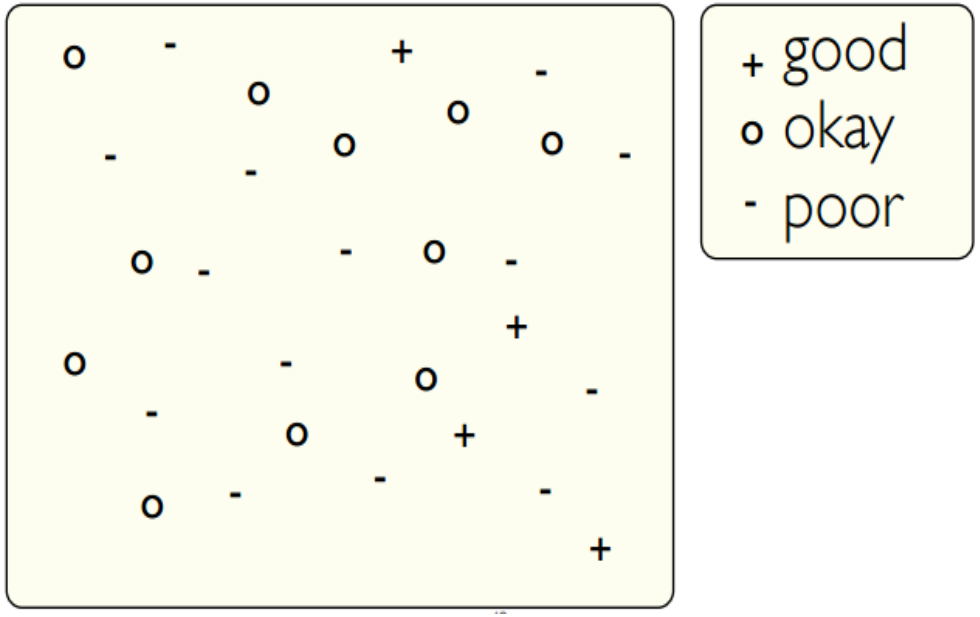
- With GEViT on hand, we can look to automation more
- We'll keep the human-in-the-loop – injection of domain knowledge is essential
- Perils of premature automation:



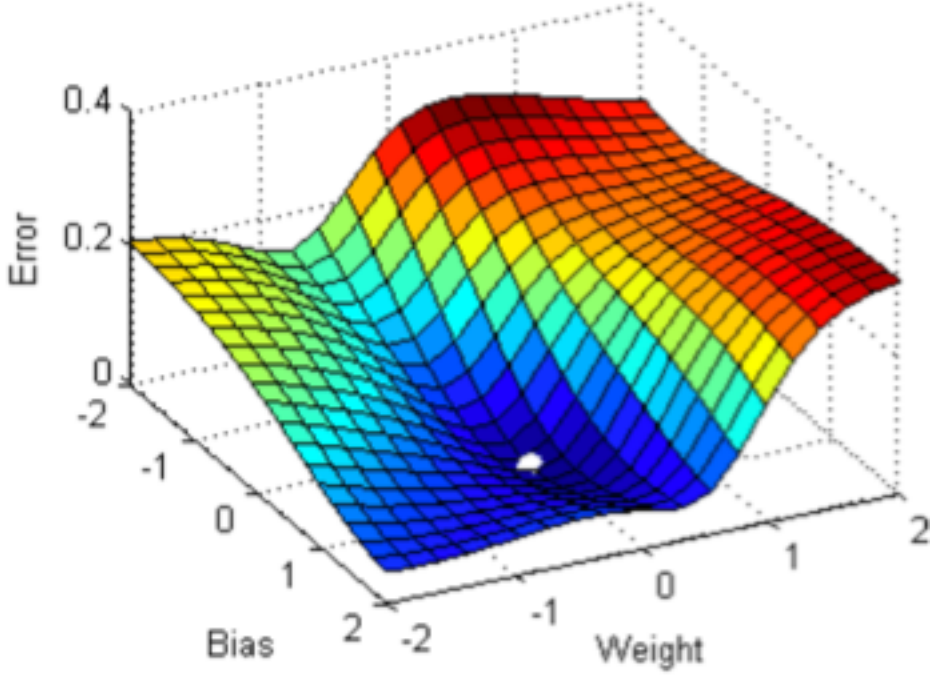
Next Steps: learning and suggesting good data visualizations

- Thinking of visualizations as *visual models*
- Can we leverage statistical techniques for data visualizations?
- Can we transform data visualization into a model selection problem?

Visualization Design Space



Model Error Surface



Establishing a visualization design space
A case study in infectious disease genomic epidemiology

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*

Literature Analysis

Approach	Literature Search	Data Clean-up	Unsupervised Clustering	Linking to *a priori* Topics	Sampling
Data	Pubmed Central *Titles & Abstracts*	Document corpus	Tidytext corpus, Document term matrix	Tidytext corpus Document corpus	Document corpus
Methods	Query Pubmed through R	Extract 1-gram, Remove stop words, Remove numbers, remove common words, Calculate td_idf metric	rTSNE, HBSCAN (search for optimal hbscan params) Name clusters by two most common names	Manual annotations	Sample per topic (per pathogen, see results) Manually assess appropriateness, re-sample for rejected
Packages	`risemed`, `parseJSON`	`tidytext, snowballC, dplyr, Stringr`	`rTSNE, hdbscan`	-	-
Output	Document corpus	Tidytext corpus, Document term matrix	add cluster to document corpus **[a result]**	add cross-cutting topic to document corpus **[a result]**	Sampled document corpus Spreadsheet keep/reject (reason)

Qualitative and Quantitative Analysis

Approach	Figure Extraction (including captions)	Axial Coding	Gallery Development	Quantitative Analysis
Data	Sampled Document Corpus *+ some manual additions*	Figure (and table) corpus	Sampled Document Corpus Figure & Tables Code set	Sampled Document Corpus Annotated Figures & Tables
Methods	Manual extract figures & some tables from PDF Optical character recognition for figure captions	Manual, lots of group discussion and iterative refinement	Prototype development	Univariate & Bivariate Descriptive Statistics
Packages	`tesseract`	-	`shiny`	`dplyr;ggplot`
Output	Figures & some tables with captions as text	Code set for: basic chart types, chart combinations, and chart annotations **[a result]**	Annotated Figures & Tables Browseable gallery **[results]**	Descriptive Statistics **[a result]**

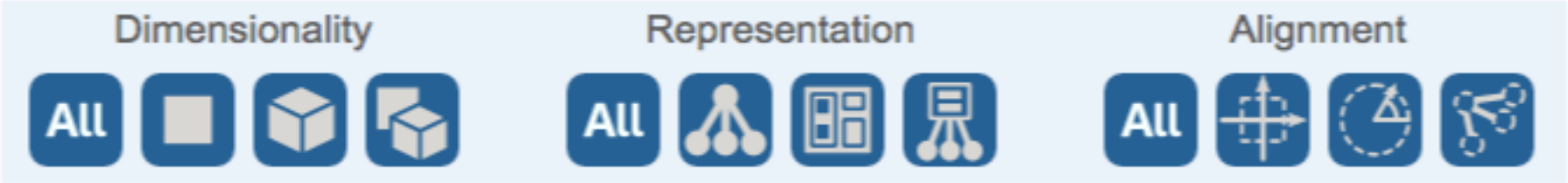
Design spaces are not a new idea

- Design spaces are useful to understand what is possible
  - Exists in architecture, computer science, and other disciplines
- Visualization researchers talk quite a bit about design spaces

Design spaces are not a new idea

- Design spaces are useful to understand what is possible
  - Exists in architecture, computer science, and other disciplines
- Visualization researchers talk quite a bit about design spaces
- YET – for visualizations, no systematic method exists for creating design spaces
  - Example of design space exploration tools below were not systematically constructed

treevis.net



Setviz.net