Creating explorable visualization design spaces:
An example from 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

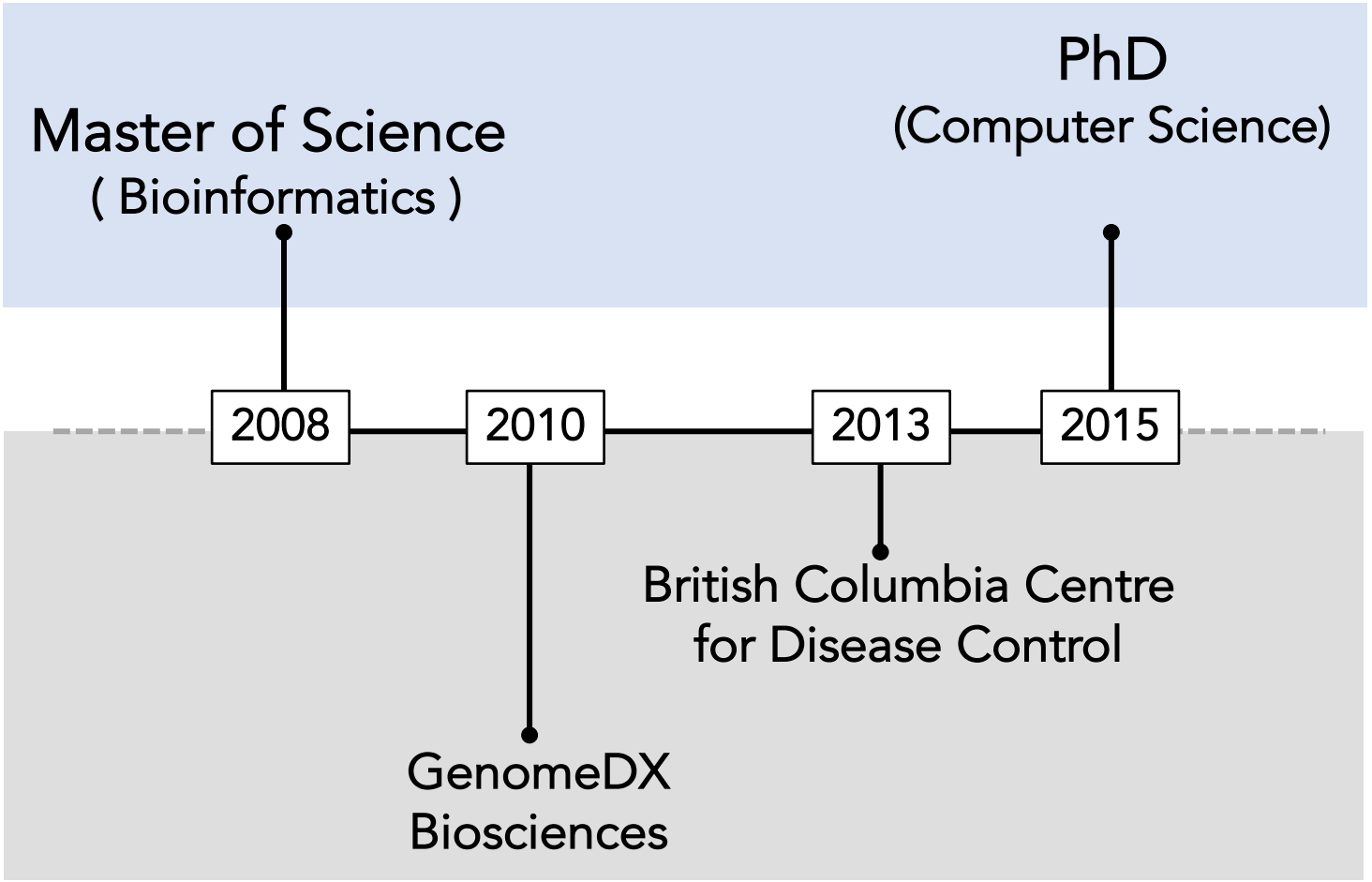
http://cs.ubc.ca/~acrisan

Hello World!

PhD Candidate, Computer Science
University of British Columbia

Thesis: Visualizing Public Health Data

Advisors: Dr. Tamara Munzner
Dr. Jennifer Gardy



I am graduating soon!

What we'll talk about

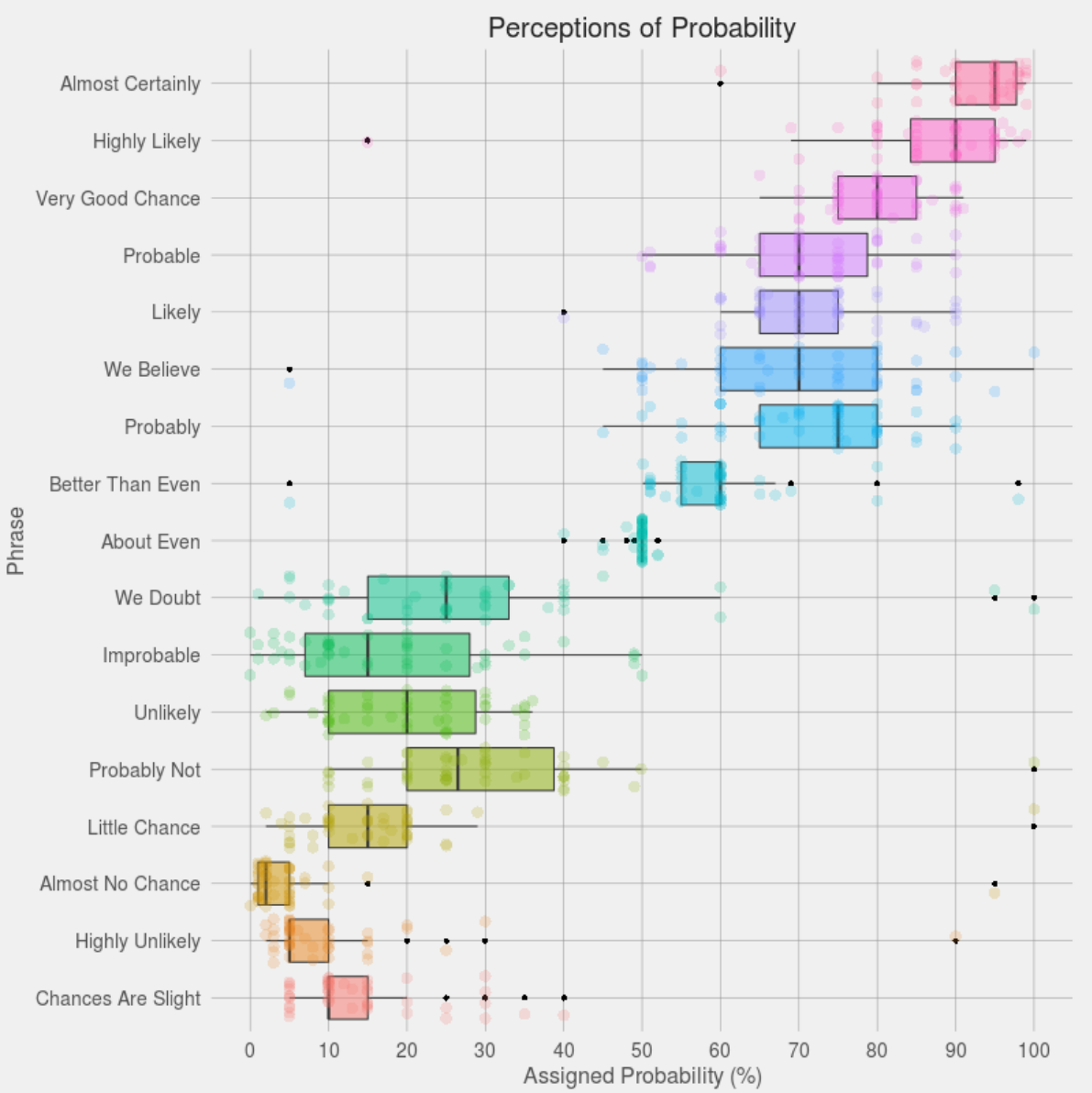
Why should we visualize data?

Thinking systematically about data visualization

GEViT: a Genomic Epidemiology Visualization Typology

Why should we visualize data?

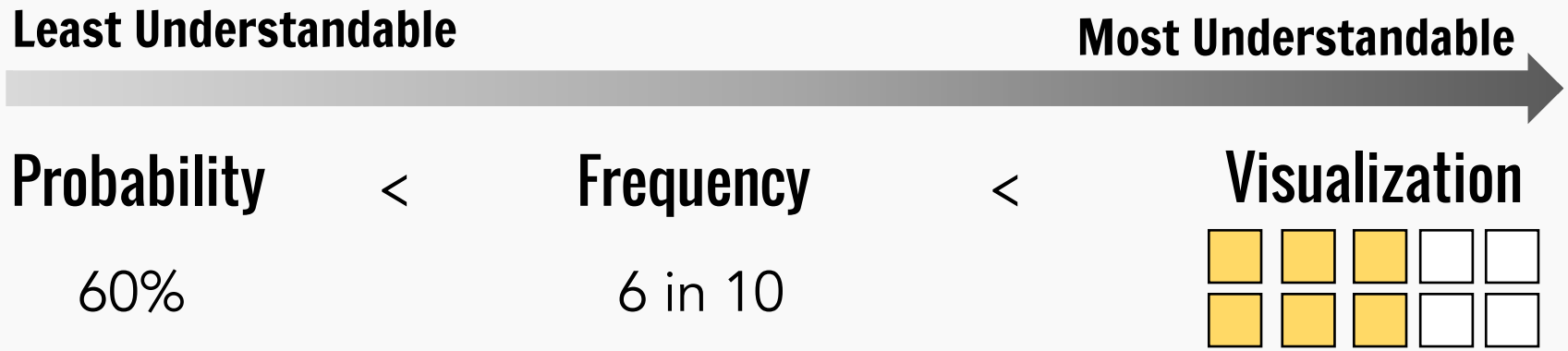
Humans interpret numerical information differently



It is not always easy to reason consistently with numbers

http://bit.ly/1FxtT2z

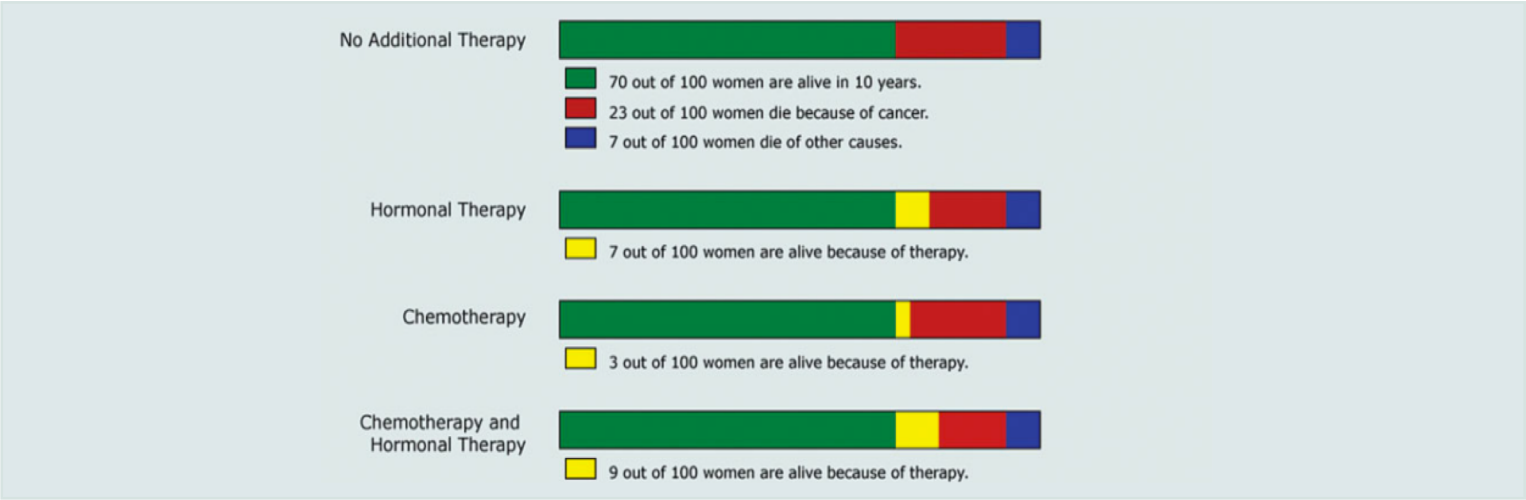
Humans interpret numerical information differently



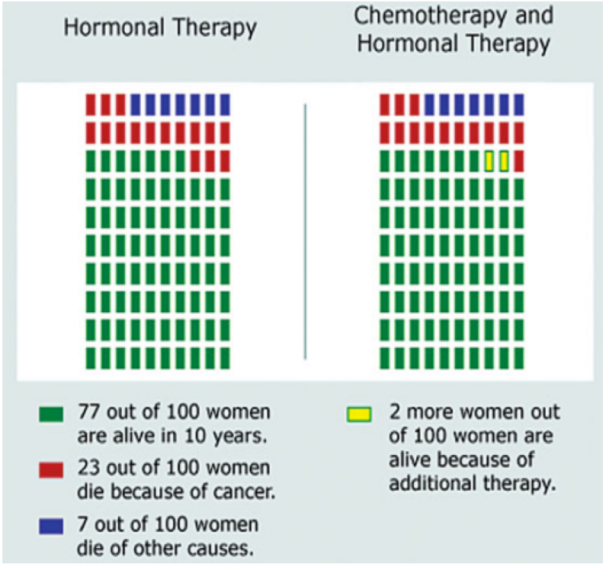
- Numeracy : the ability to reason with numbers
  - Individuals with low numeracy have a difficulty interpreting numbers and probabilities
  - Also true amongst educated professionals
- Visualization can make data more accessible to individuals with lower numeracy skills

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

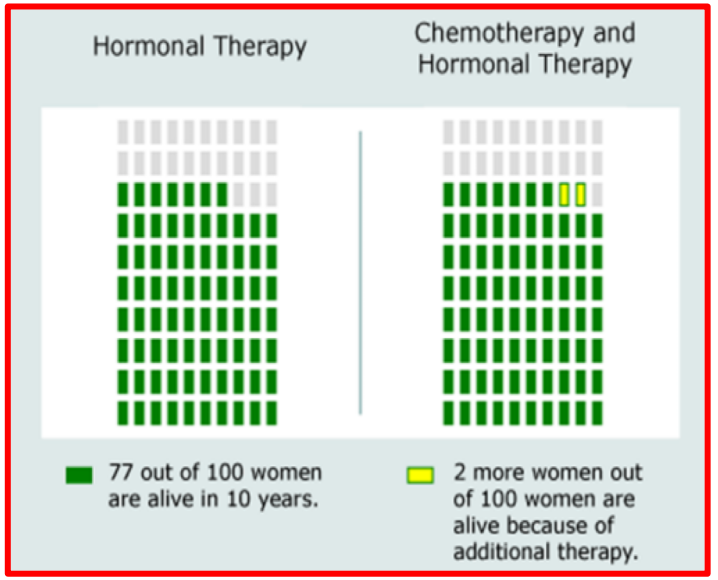
Need the *right* data visualization, not just *a* visualization



Alternative 1



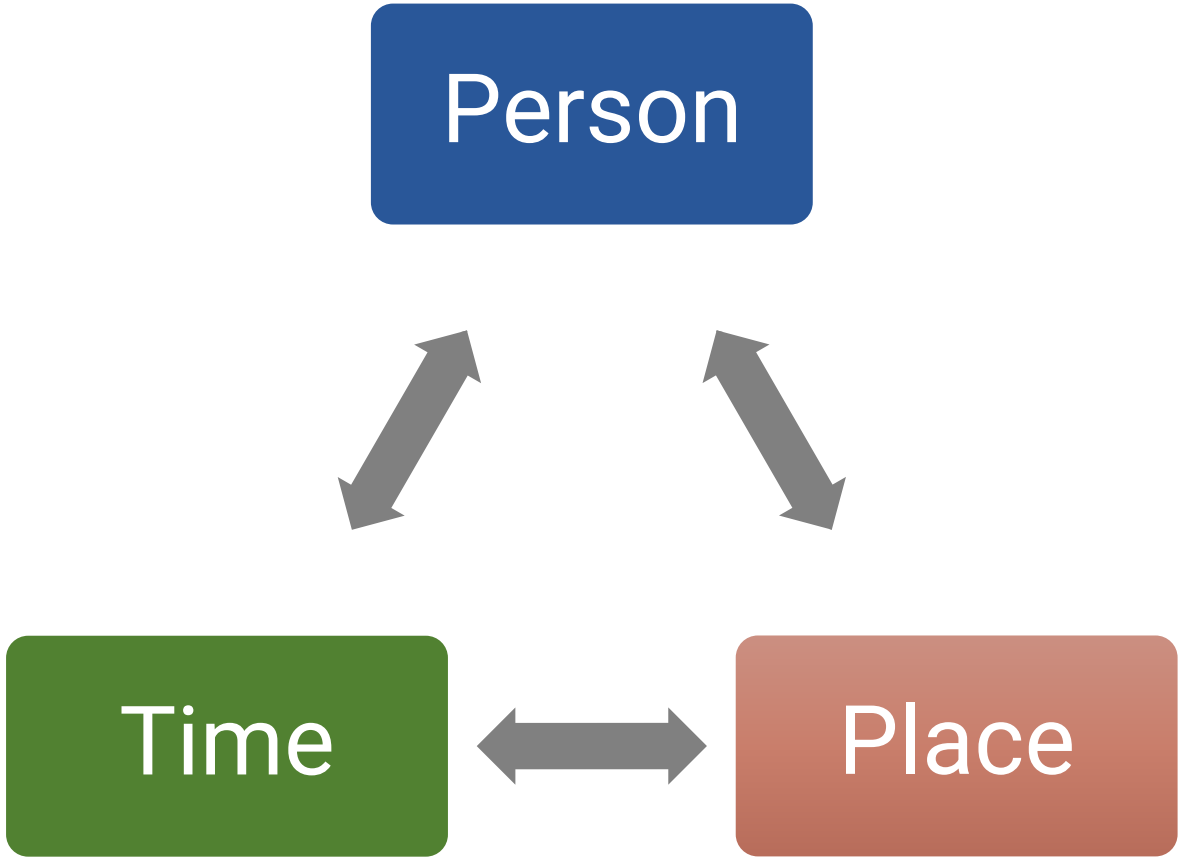
Alternative 2



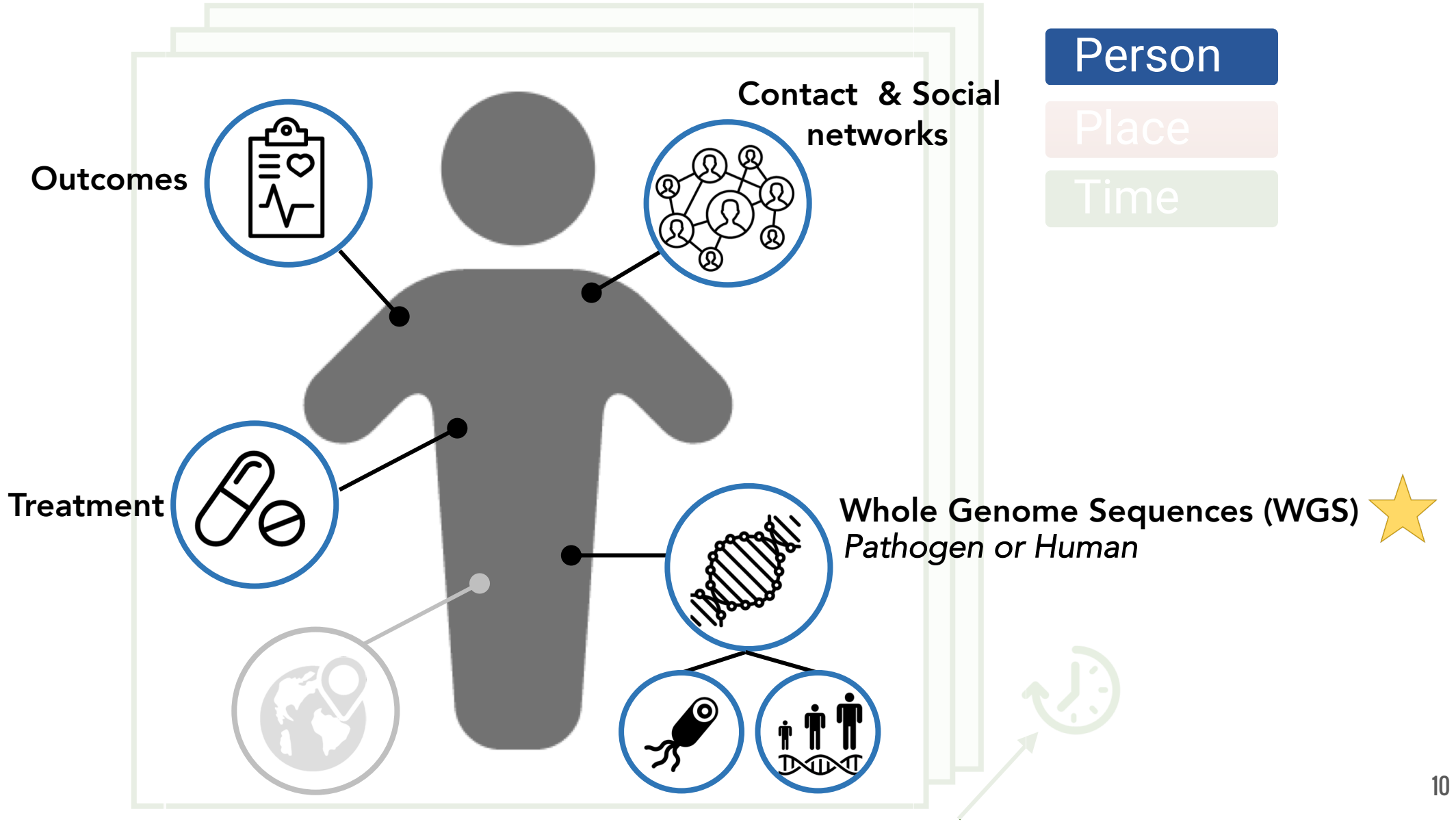
Zikmund-Fisher (2013). A demonstration of "less can be more" in risk graphics.

Public health data are complex

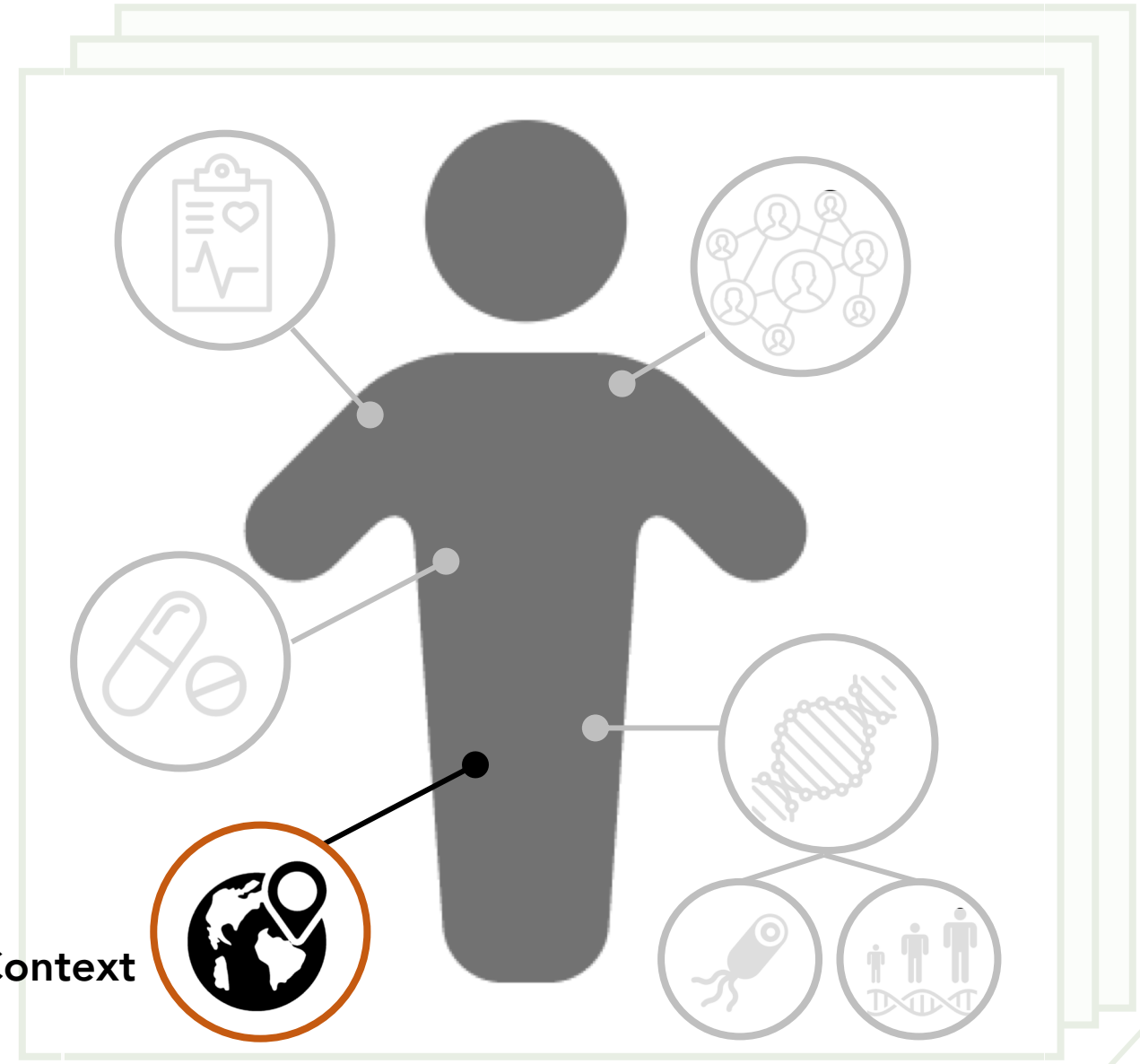
- Technology is impacting the kinds and amount of data collected
- Traditional epidemiological triad (below) now more complex



Public health data are complex



Public health data are complex



Person

Place

Time

Location
Geographic Context

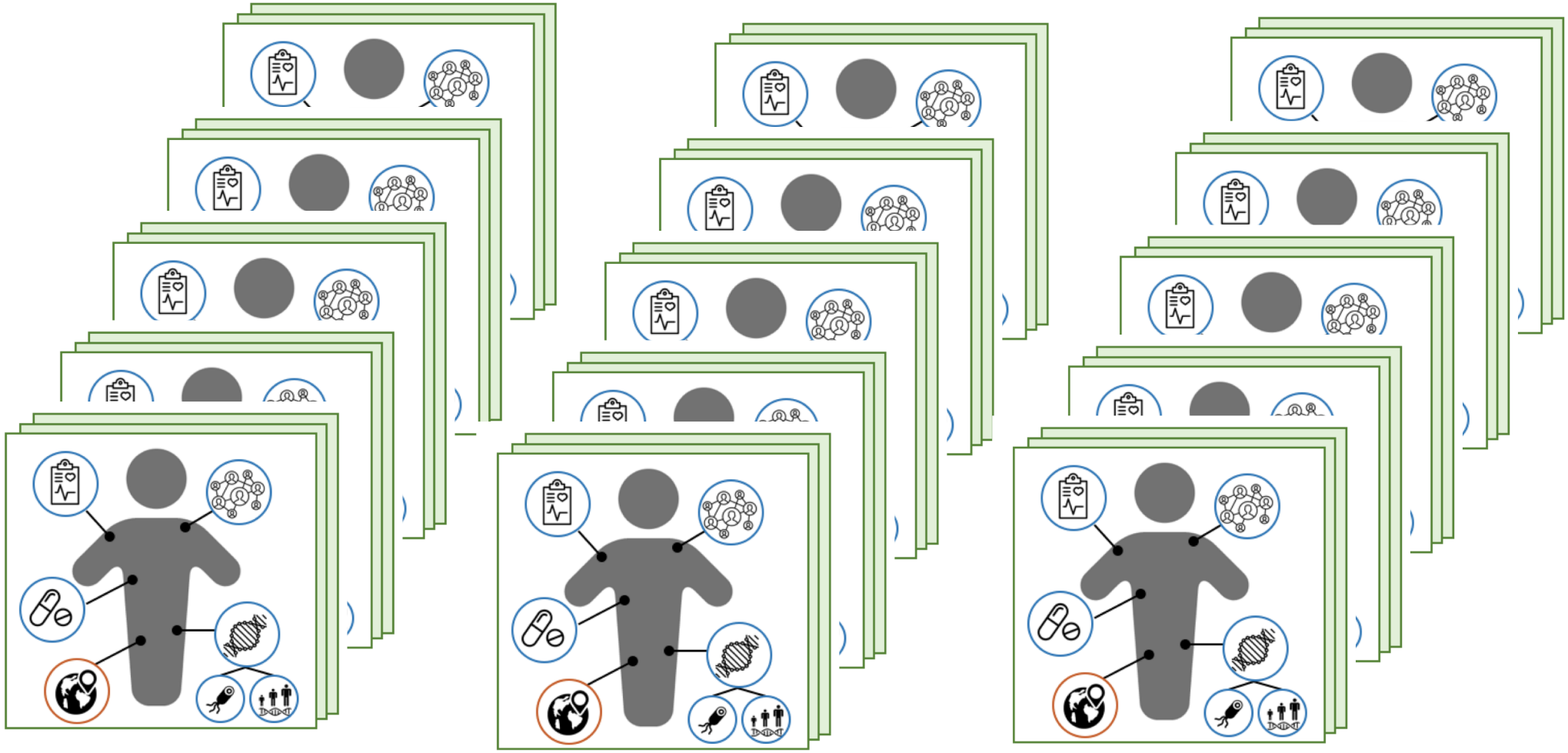
Public health data are complex

Person

Place

Time

Complex data are collected from whole populations



Data is used by many different decisions makers

- **In theory:** more decision makers, more data, better decisions
- **In reality:** different decision maker needs, different numerical literacy
- **Hypothesis:** data visualization can support with decision making

Medical Health Officers

Clinicians

Nurses

Researchers

Community Leaders

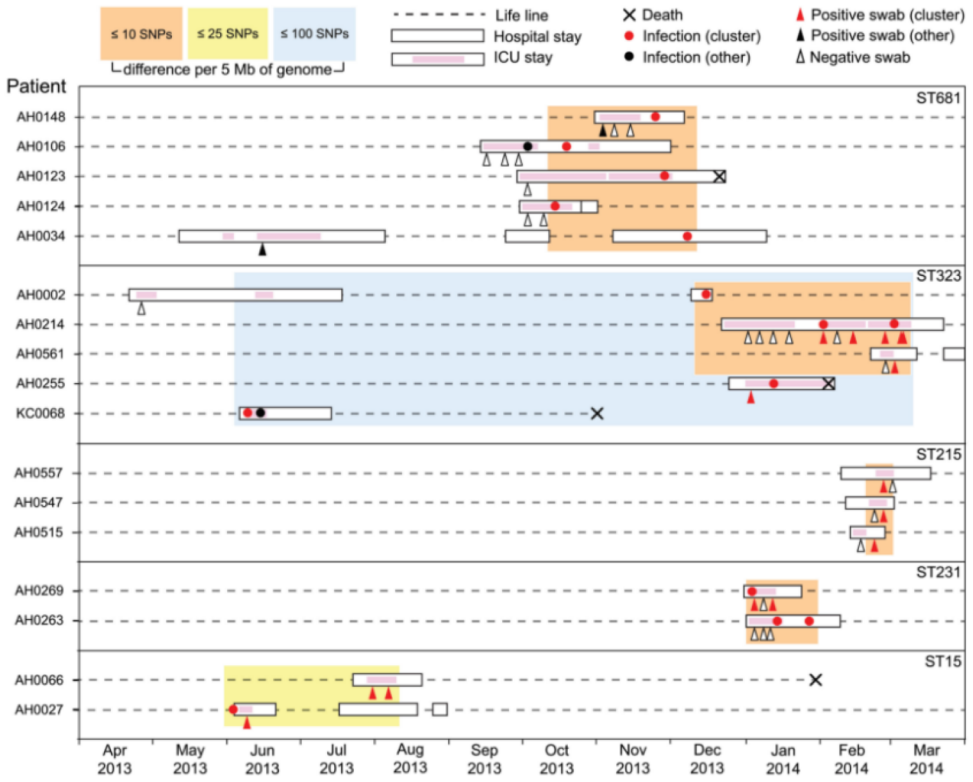
Patients

Politicians

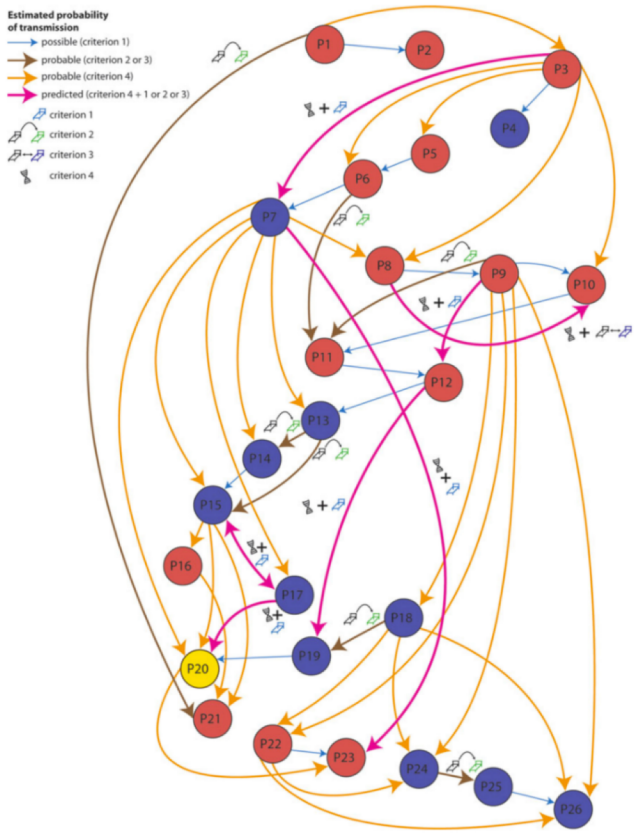
Public health data visualization is highly variable

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

Gorrie (2017)



Willman (2015)



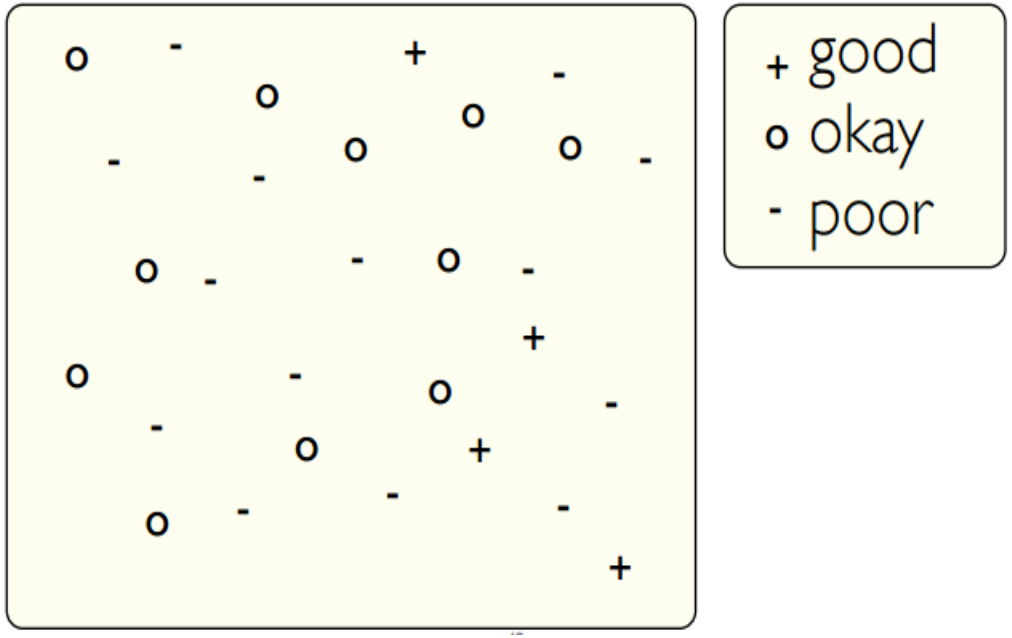
Davis (2015)

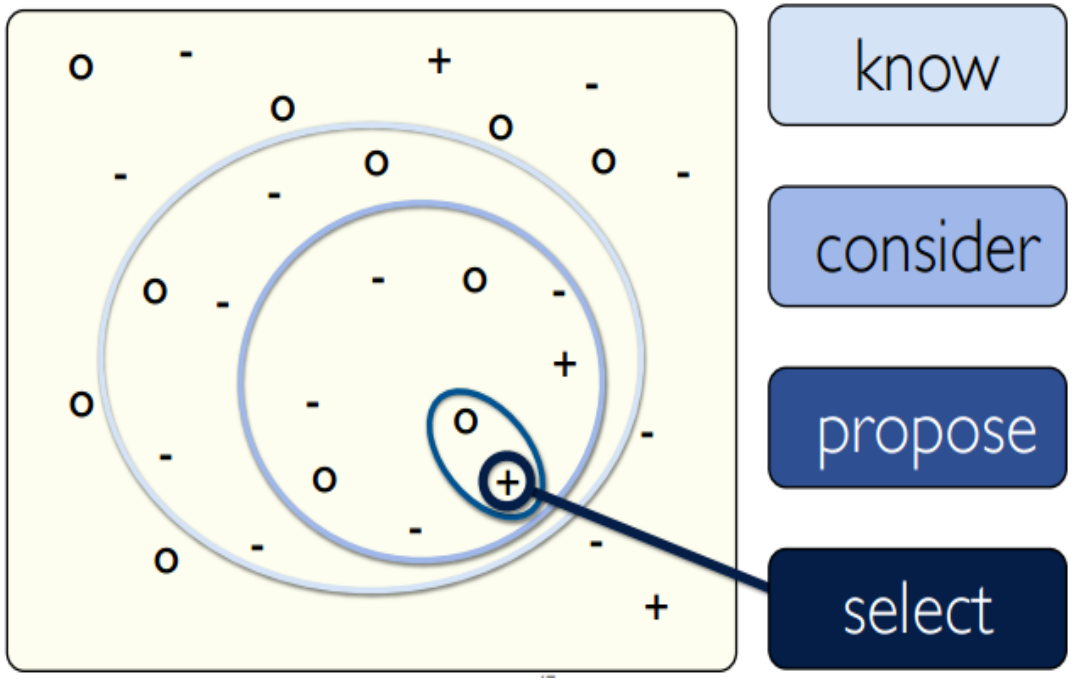
Could I conduct a **systematic review of data visualizations** used in public health genomic epidemiology?

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 spaces and 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 (-)

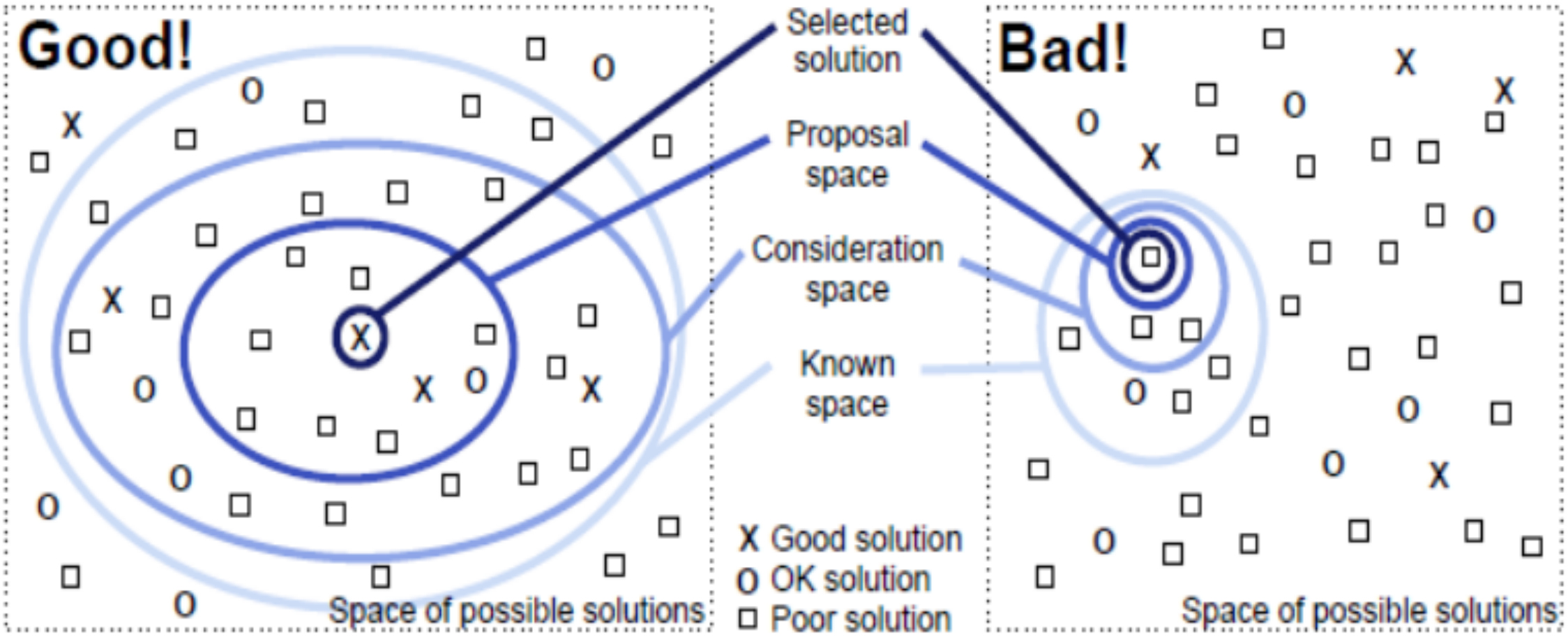
Suitable as an office chair (+,0)

Design Spaces : A quick primer

GOAL – nudge domain specialists toward better design choice solutions

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

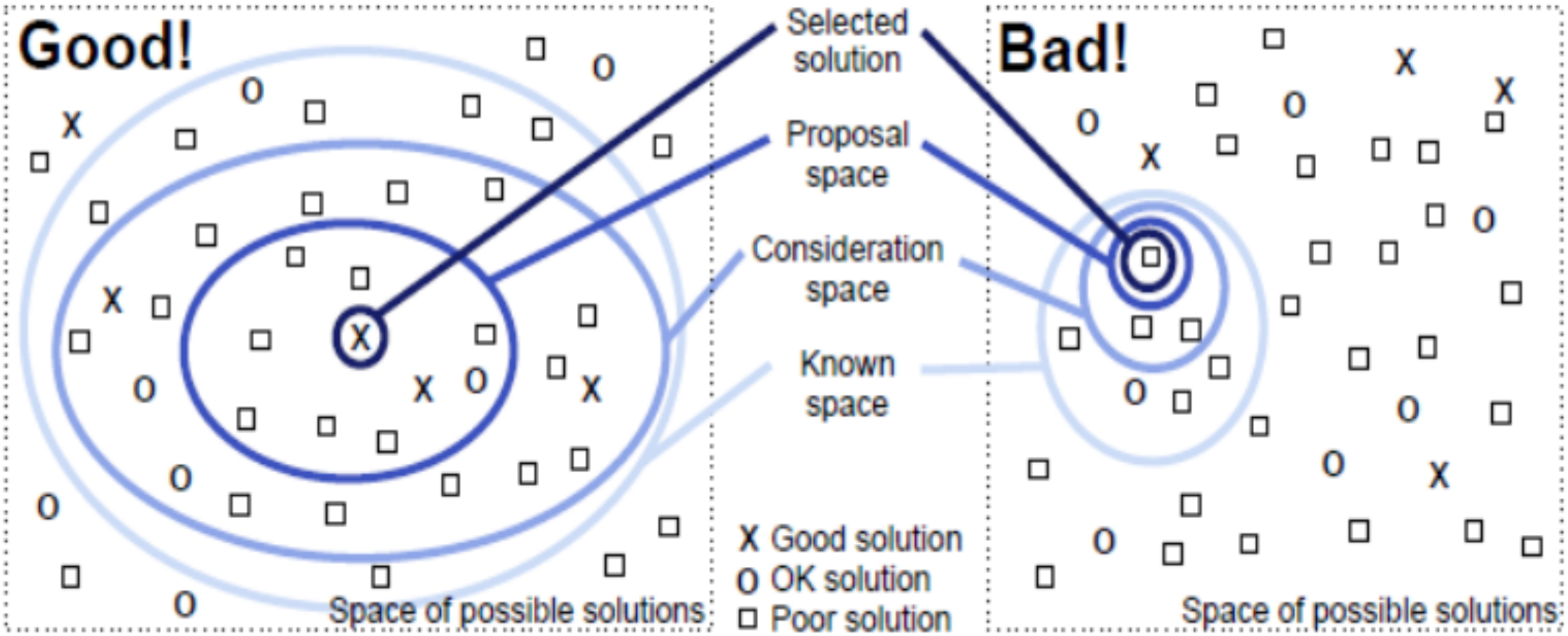
(i.e. choosing the nearest & cheapest chair)



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

Design Spaces : A quick primer

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

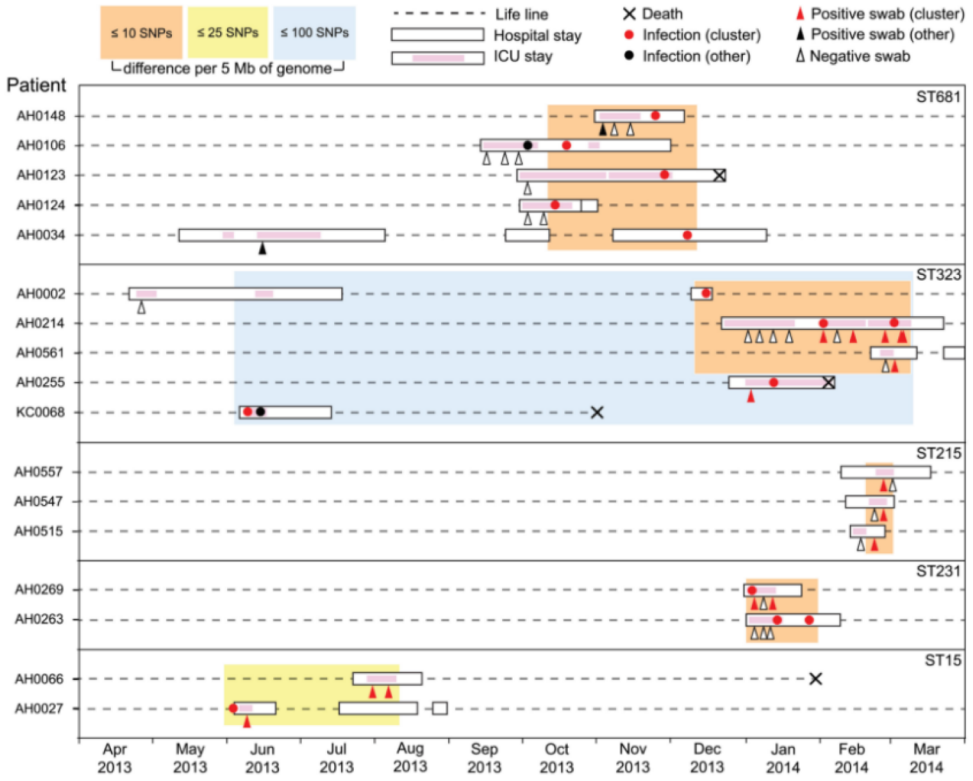


Source: Munzner (2014) "Visualization Analysis and Design"

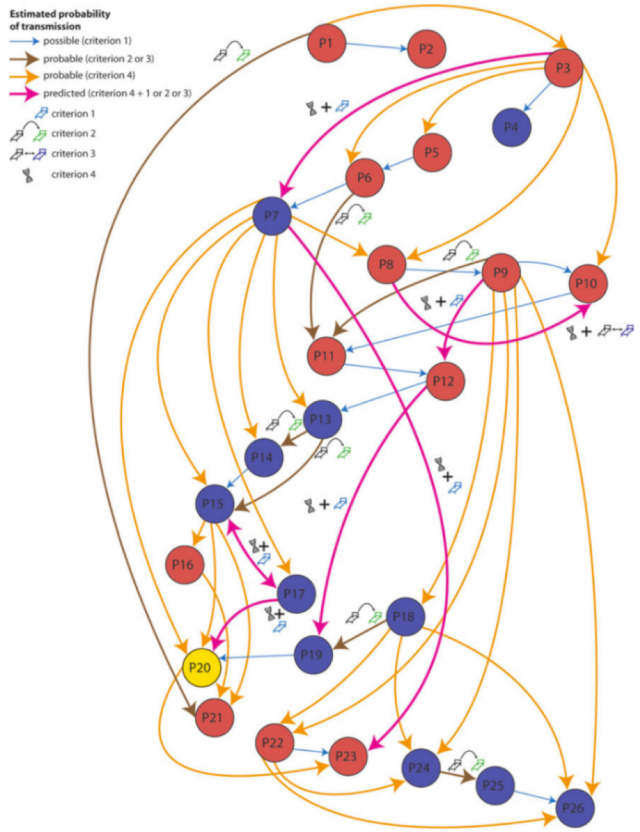
Considering a design space for public health

- **These are options in a visualization design space**
- **How do we describe and compare these visualization options?**

Gorrie (2017)



Willman (2015)



Davis (2015)

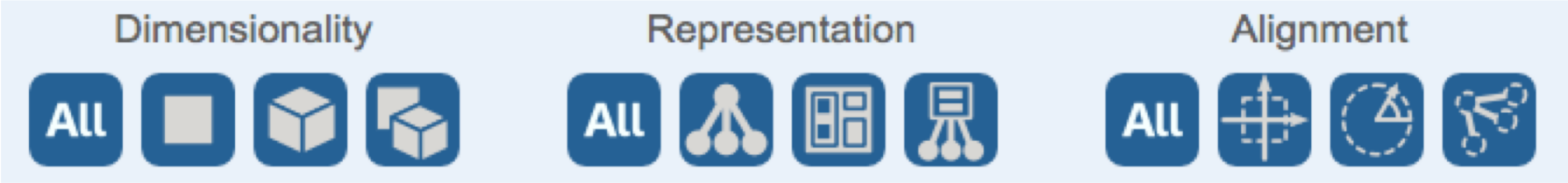
Design spaces are not a new idea

- **Design spaces are useful to understand what is possible**
  - Exists in architecture, computer science, and other disciplines

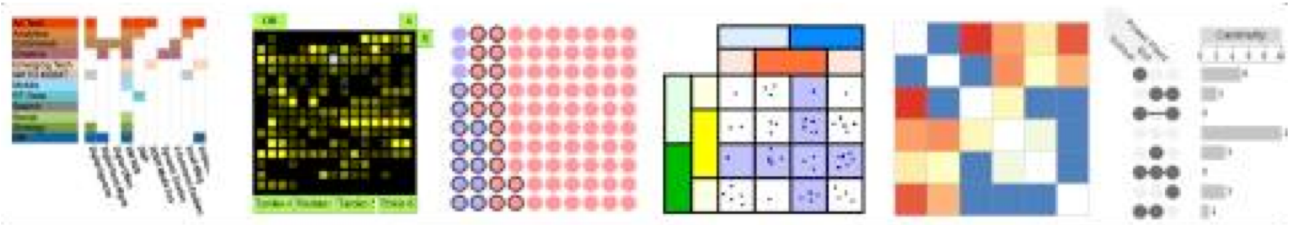
Design spaces are not a new idea

- **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**

treevis.net



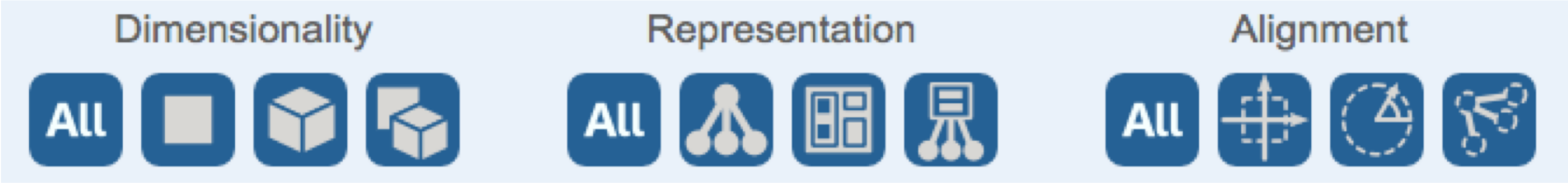
Setviz.net



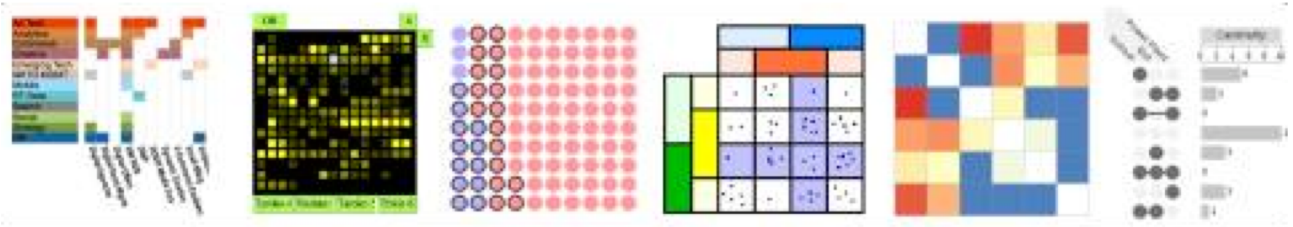
Design spaces are not a new idea

- **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 – no systematic and reproducible method exists for creating design spaces**
  - Examples below were constructed by author curation

treevis.net



Setviz.net

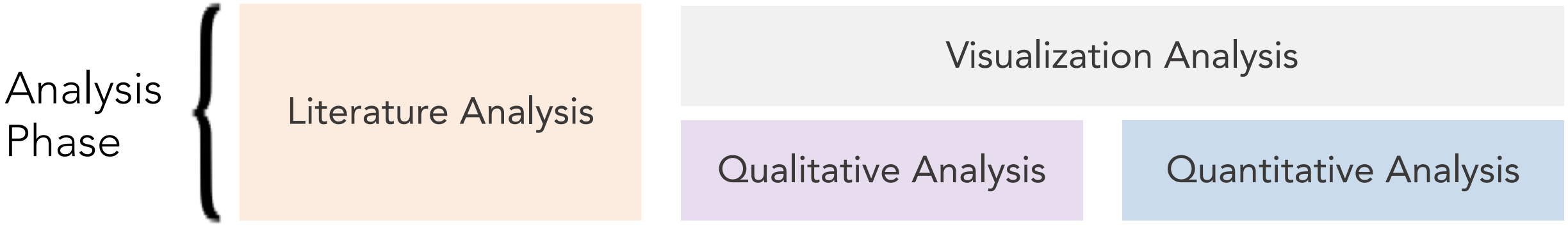


Creating an explorable visualization design space

https://doi.org/10.1101/325290

Our systematic method for survey data visualizations

Our method has two analysis phases:

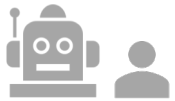


Our systematic method for survey data visualizations

Some analyses are automated () and others are not ()

Our systematic method for survey data visualizations

Analysis phases answer different research questions

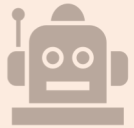
Analysis Phase	Literature Analysis	Visualization Analysis	
		Qualitative Analysis	Quantitative Analysis
			
Research Question	*WHY* are researchers visualizing data?	*HOW* are researchers visualizing data?	*HOW MANY* examples of specific visualizations?

Applying our method to public health data visualizations

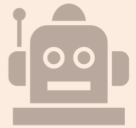
Our Objective	Across the many topics of microbial gen epi research articles identify and enumerate the different kinds of visualizations that are used

Applying our method to public health data visualizations

Literature Analysis Steps



Text mining of document corpus to identify topics



Systematically sample papers with topics as strata

Our Objective

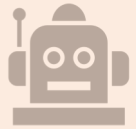
Across the many topics of microbial gen epi research articles

identify and enumerate the different kinds of visualizations that are used

Applying our method to public health data visualizations

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

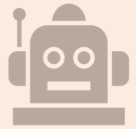
Derived a code set to classify research figures (GEViT)

Applied GEViT to collection of research figures

Applying our method to public health data visualizations

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

The novelty of our approach

- **Why is this approach different?**
  - Uses machine **and** human knowledge to expose a design space
  - Systematically and reproducibly samples across possible visualizations
  - Develops typology for visualization analysis

The novelty of our approach

- **Why is this approach different?**
  - Uses machine **and** human knowledge to expose a design space
  - Systematically and reproducibly samples across possible visualizations
  - Develops typology for visualization analysis
- **What does this approach reveal?**
  - Current common practices in visualization design for gen epi.
  - Practitioner driven, not expert driven, visualizations
  - Good, bad, and absent visualization practices

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
- Need GEViT to conduct evaluations

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
- Need GEViT to conduct evaluations

- **How can GEViT be used?**
  - Use GEViT to systematize analysis of visualizations
  - Understand what visualizations are common and possible
  - Get ideas for data visualization design

A snapshot of our findings

Photo by Tirachard Kumtanom from Pexels

Key analysis steps and their results

Analysis Step	# articles	results
Article Acquisition & Unsupervised Clustering	17,974	*Article topic clusters* *WHY* are researchers visualizing data?
Limit to clusters of human pathogens	6,350	
Assign *a priori* concepts	6,350	*23 a priori concepts*
Stratified Sampling	221	*801 figures & 49 tables (for qualitative analysis)*

Key analysis steps and their results

Analysis Step	# articles	results
Article Acquisition & Unsupervised Clustering	17,974	*Article topic clusters* *WHY* are researchers visualizing data?
Limit to clusters of human pathogens	6,350	
Assign *a priori* concepts	6,350	*23 a priori concepts*
Stratified Sampling	221	*801 figures & 49 tables (for qualitative analysis)*
Iterative axial coding	221	*A genomic epidemiology visualization typology (GEViT)* *HOW* are researchers visualizing data?

Key analysis steps and their results

Analysis Step	# articles	results	
Article Acquisition & Unsupervised Clustering	17,974	*Article topic clusters* *WHY* are researchers visualizing data?	1
Limit to clusters of human pathogens	6,350		
Assign *a priori* concepts	6,350	*23 a priori concepts*	
Stratified Sampling	221	*801 figures & 49 tables (for qualitative analysis)*	2
Iterative axial coding	221	*A genomic epidemiology visualization typology (GEViT)* *HOW* are researchers visualizing data?	3
Descriptive Statistics	221	*Current common visualization practices*	4

1 Discover article topics & assign *a priori* concepts

17,974 articles in their **original format**

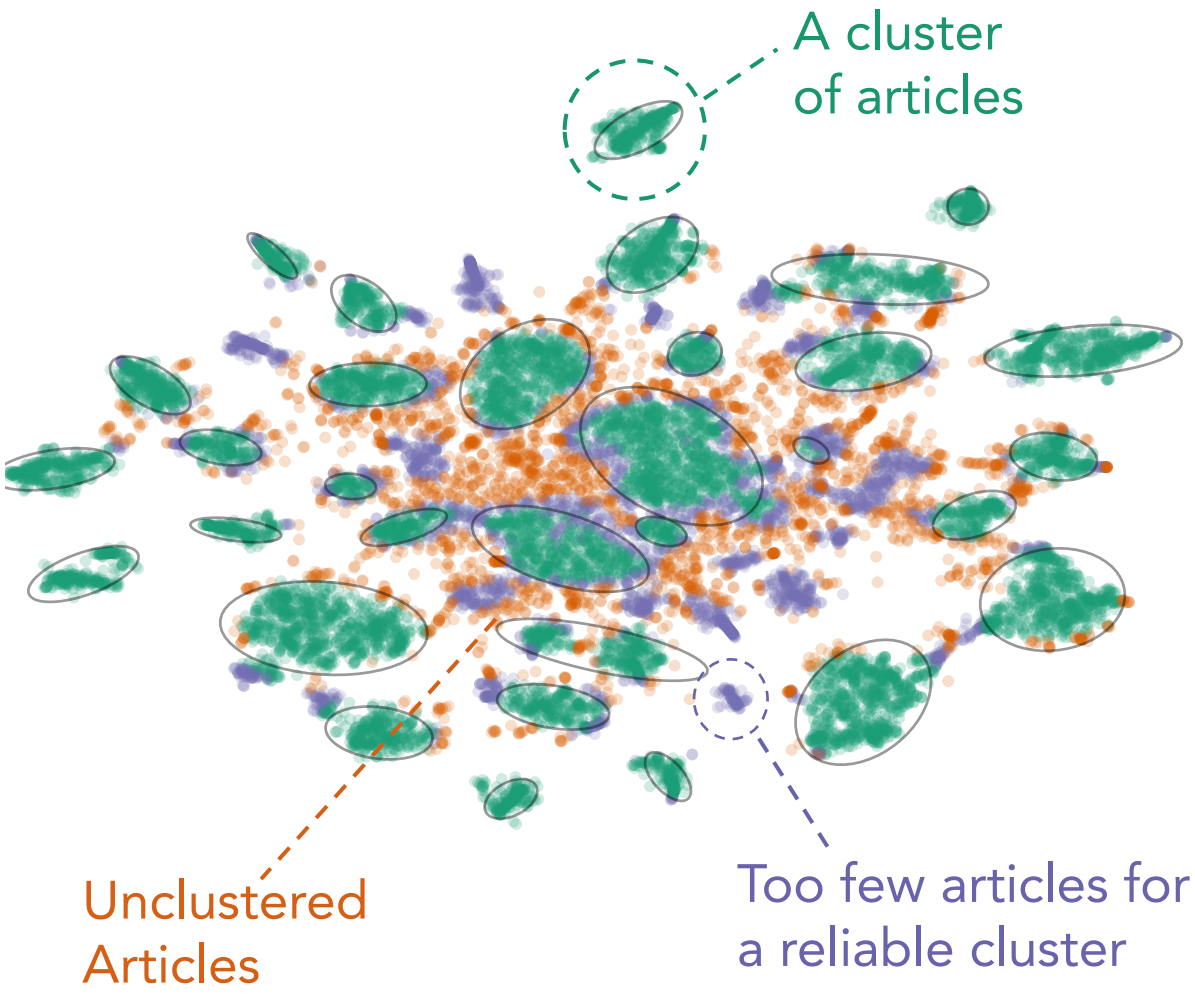
Used in text analysis

PMID	YearPub	Journal	Authors	Title	Abstract	PMCID	DOI
27746398	2016	Clinical mi	Pì©rez-Lago,	A novel strategy based on	OBJECTIVE: Molecular epidemic	NA	NA
27607836	2016	PLoS negle	Antonation, I	Bacillus cereus Biovar Ant	Through full genome analyses c	PMC5015827	10.1371/journal.pntd.0(
26819311	2016	Journal of	Bodewes, Rc	Spatiotemporal Analysis o	Influenza A viruses are major pa	PMC4836327	10.1128/JVI.03046-15
26758561	2016	The Journa	Chen, Liang-J	Diversity and evolution of	The wide circulation of novel av	NA	NA
26735033	2016	Journal of	Verhegghe, I	Prevalence and Genetic Di	Since the first description of live	NA	NA
25113057	2014	Annals of (	Chen, Sheng;	Spread of carbapenemase	BACKGROUND: The rapid emer	PMC4236511	10.1186/s12941-014-0(
23938759	2013	Philosophi	Gray, Rebecc	Evolutionary analysis of he	Reconstructing the transmissio	PMC3758194	10.1098/rstb.2013.0168
22723256	2012	Journal of	Arduino, Son	Transposons and integrons	Multiple transposons, integrons	NA	NA
22386850	2012	Infection, g	Rivero-Pì©re	Genetic diversity of comm	With the recent detection of MI	NA	NA
19487205	2009	Philosophi	Sloot, P M A;	HIV decision support: from	Human immunodeficiency virus	NA	NA
19000628	2008	Virology	Chen, Rubing	Frequent inter-species tra	Revealing the factors that shap	PMC2633721	10.1016/j.virol.2008.10
18564686	2008	The Southe	Tiwari, Hare	Molecular typing of clinica	Molecular typing of total 84 Sta	NA	NA
17617184	2007	Clinical mi	Kawalec, M;I	Hospital outbreak of vancc	A mixed outbreak caused by var	NA	NA
28042011	2017	Journal of	Deurenberg,	Application of next genera	Current molecular diagnostics c	NA	NA
26598368	2015	Journal of	Box, Allison I	Functional Analysis of Bac	The classical and El Tor biotype	PMC4719448	10.1128/JB.00747-15
22502605	2012	Letters in a	Aliabad, N H	Molecular diversity of CTX	AIMS: The objective of this stud	NA	NA
21645368	2011	BMC geno	Chen, Yansh	Comparative genomic ana	BACKGROUND: Vibrio parahaer	PMC3130711	10.1186/1471-2164-12
21177926	2010	The Journa	Lorusso, Ales	Genetic and antigenic cha	Prior to the introduction of the 2	PMC3133703	10.1099/vir.0.027557-0
20375036	2010	The Journa	Post, Virginia	Evolution of AbaR-type ge	OBJECTIVES: To determine if m	NA	NA
19809504	2009	PloS one	Khiabanian, I	Reassortment patterns in	Three human influenza pandem	PMC2752997	10.1371/journal.pone.0
18815309	2008	Journal of	Salloum, Sha	Escape from HLA-B*08-res	The inherent sequence diversity	PMC2583685	10.1128/JVI.00997-08
17784948	2007	BMC geno	Menard, Aym	Architecture of Burkholder	BACKGROUND: The Burkholderi	PMC2194791	10.1186/1471-2164-8-3
17651134	2007	FEMS micı	Chen, Yansh	Genetic variation of capsu	Both NRT36S and A5 are NAG-S	NA	NA
15213324	2004	Proceeding	Holden, Matt	Complete genomes of two	Staphylococcus aureus is an imp	PMC470752	10.1073/pnas.0402521
28348549	2017	Frontiers i	Katz, Lee S;G	A Comparative Analysis of	Modern epidemiology of foodbo	PMC5346554	10.3389/fmicb.2017.00

17,974 clustered articles

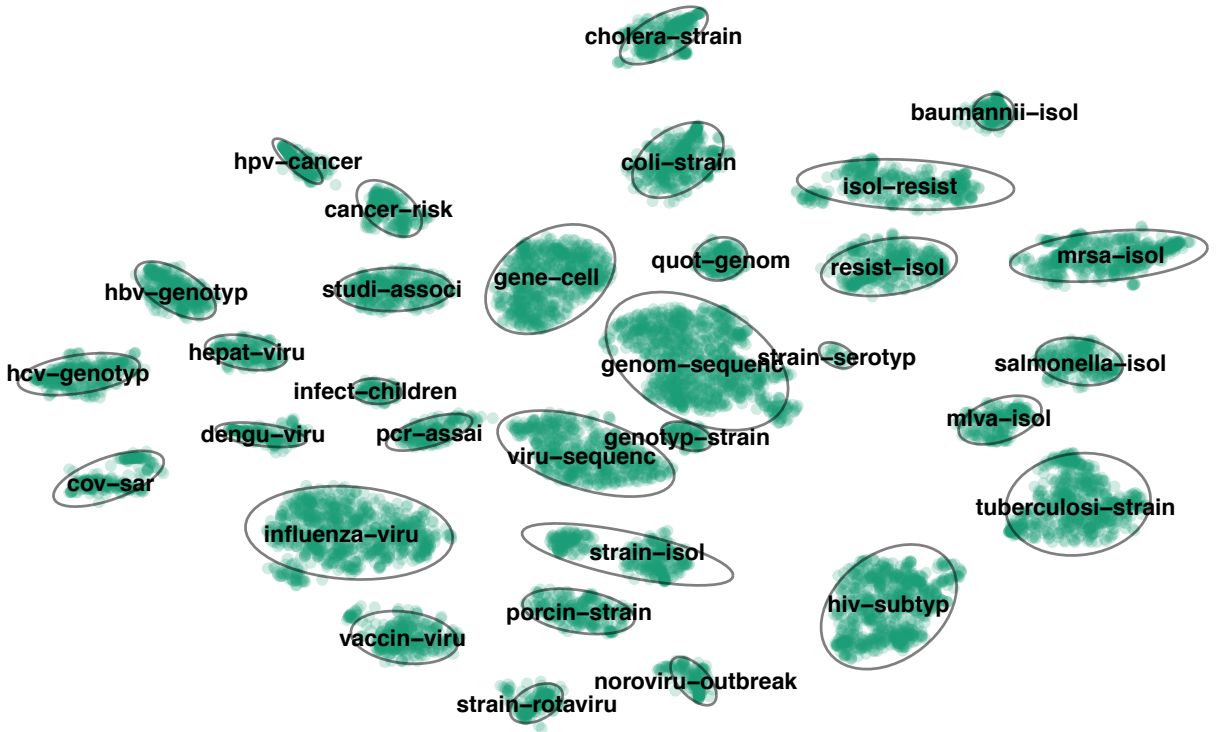
Initial Topic clustering results

t-SNE followed by hdbscan

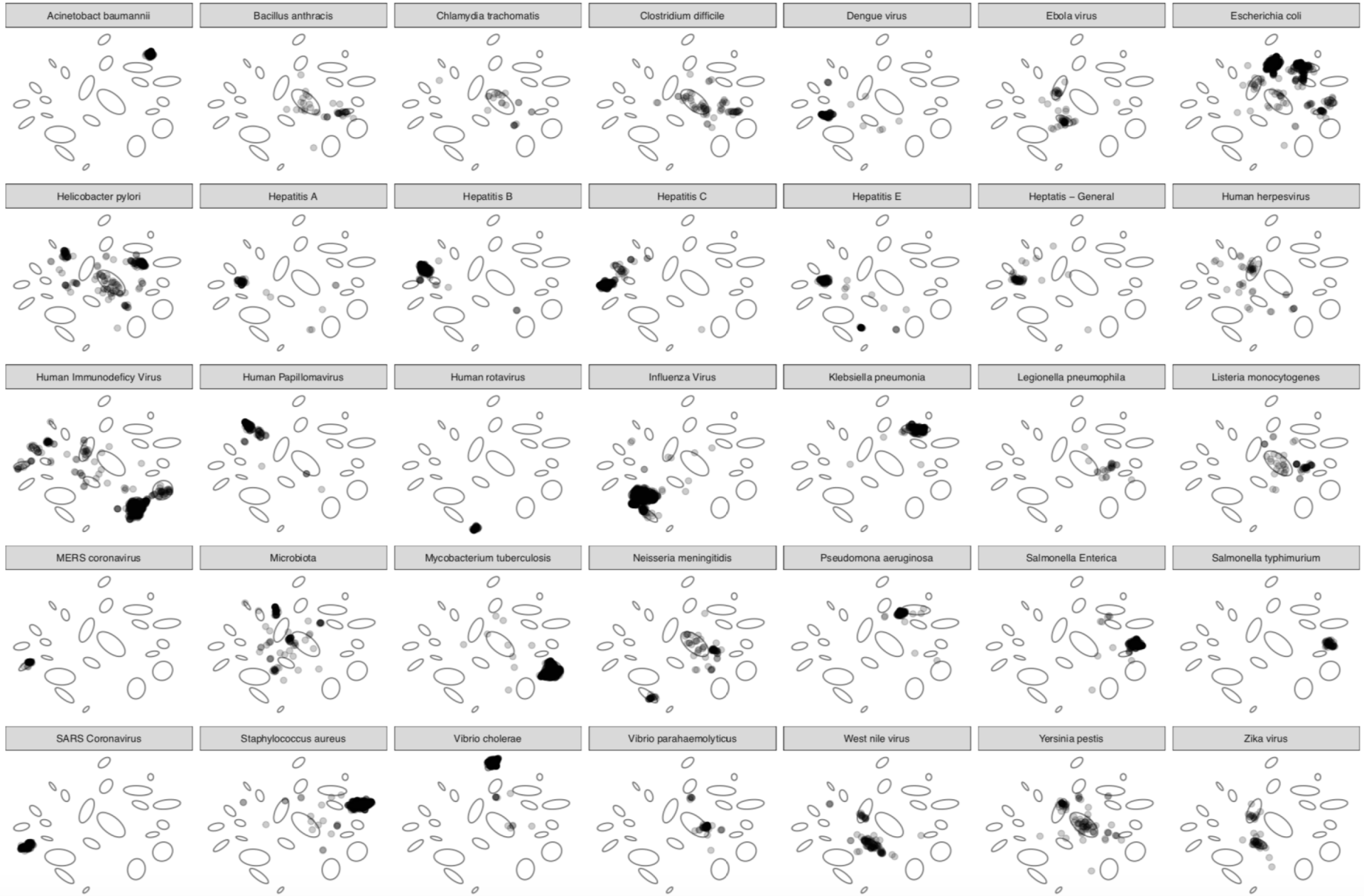


Final set of topic clusters

Articles appear to cluster around pathogens



Articles cluster around pathogens: independent validation



Using cluster results to assign *a priori* concepts

- *A priori* concepts = public health concepts developed by experts
  - 23 *a priori* concepts in total
- Assignment of a *priori* concepts to documents
  - Extract frequent terms that occur within and between clusters
  - Assign *a priori* concept to terms

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

Why are these finding relevant?

- **Provides an overview of all the articles**
  - Wayfinding to support downstream decisions

Why are these finding relevant?

- **Provides an overview of all the articles**
  - Wayfinding to support downstream decisions
- **Different data visualizations for different pathogens?**
  - Different transmission routes, mortality, etc.
  - Should there be different visualizations? Yes, No, Maybe?

Why are these finding relevant?

- **Provides an overview of all the articles**
  - Wayfinding to support downstream decisions
- **Different data visualizations for different pathogens**
  - Different transmission routes, mortality, etc.
  - Should there be different visualizations? Yes, No, Maybe?
- **Adds systematicity that is absent in author curation**
  - Visualizations sampled can be good or bad
  - Can think statistically about the distribution of visual designs

Why are these finding relevant?

- **Provides an overview of all the articles**
  - Wayfinding to support downstream decisions
- **Different data visualizations for different pathogens**
  - Different transmission routes, mortality, etc.
  - Should there be different visualizations? Yes, No, Maybe?
- **Adds systematicity that is absent in author curation**
  - Visualizations sampled can be good or bad
  - Can think statistically about the distribution of visual designs

We created an R tool based upon the literature analysis steps - Adjutant!

https://doi.org/10.1093/bioinformatics/bty722 </> https://github.com/amcrisan/adjutant

2 Sample articles to gather figures & (missed opportunity) tables

Why sample articles?

- **Automatic image recognition has limits**
  - No imagenet or well curated training data for scientific figures
  - Also, not clear what to optimize for

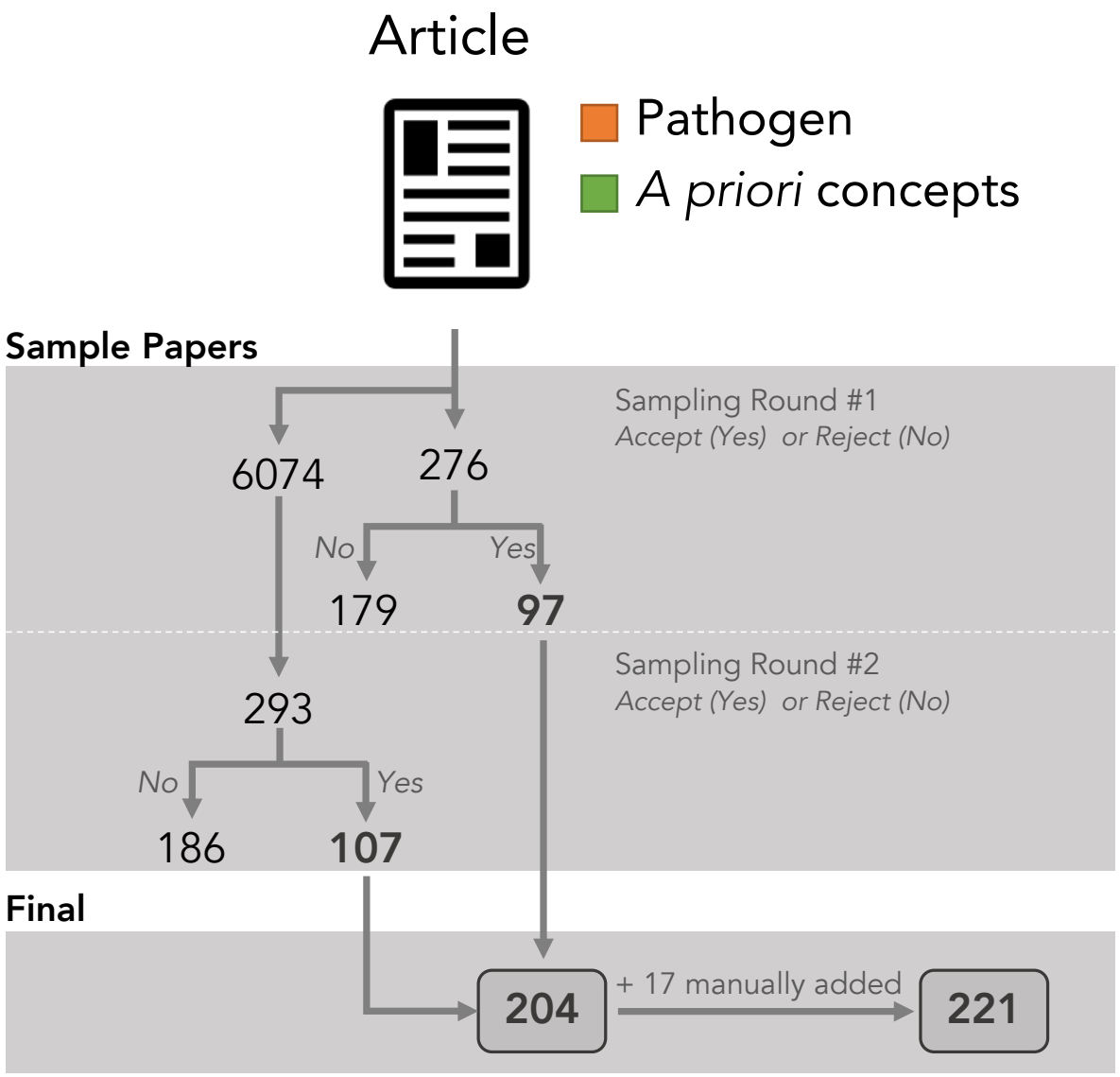
Why sample articles?

- **Automatic image recognition has limits**
  - No imagenet or well curated training data for scientific figures
  - Also, not clear what to optimize for
- **Qualitative analysis enabled deeper analysis**
  - No set visual representation for data – it's context dependent
  - Contextual factors important, but difficult, to obtain
  - Image recognition does not capture context

Why sample articles?

- **Automatic image recognition has limits**
  - No imagenet or well curated training data for scientific figures
  - Also, not clear what to optimize for
- **Qualitative analysis enabled deeper analysis**
  - No set visual representation for data – it's context dependent
  - Contextual factors important, but difficult, to obtain
  - Image recognition does not capture context
- **Since qualitative analysis necessary, we had to sample**
  - Difficult to review nearly 18,000 articles manually

Sampling procedure



Sampling of articles (two rounds)

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

Basis for further analysis

Yielded 801 figures and 49 tables

3 Developing GEViT:
A genomic epidemiology visualization typology

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

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

Developing GEViT

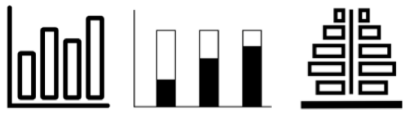
- **Input:** 801 figures, 49 tables
- **Used qualitative coding techniques to analyze research figures**
  - Multiple rounds of classifying and codifying elements of figures
  - Used figures from sample papers to derive codes
- **Figures in the same paper were analyzed separately**
  - Multi-part figures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

Chart types: a basic building block

- Chart types we observed in our analysis

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

Distribution Plot
Histogram | PDF | Boxplot | Swarm Plot

Pie Chart

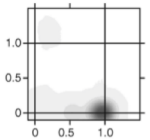
Venn Diagram

Colour Charts

Category Stripe

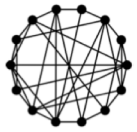
Heatmap

Density Plot*



Relational Charts

Node-link

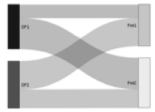


Special Cases
- *eBurst*
- Social network
- Molecular network
- Minimum Spanning Tree

Flow Diagram

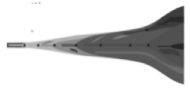
Chord Diagram

Sankey Diagram

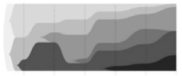


Temporal Charts

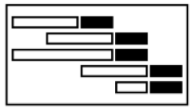
Streamgraph*
Absolute



Relative



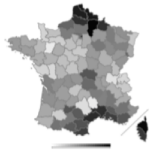
Timeline



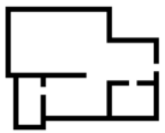
Spatial Charts

Geographic Map

Choropleth Map



Interior Map



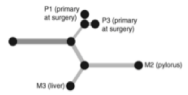
Tree Charts

Phylogenetic Tree
Rooted (Linear & Radial)

Unrooted
(Linear & Radial)

Dendrogram

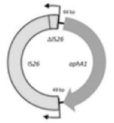
Clonal Tree*



Genomic Charts

Genomic Map
Linear

Radial



Alignment

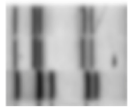
Composition Plot

Sequence Logo Plot

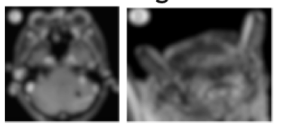
Other Charts

Table

Image
Gel Image



General Image



Miscellany

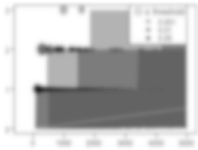
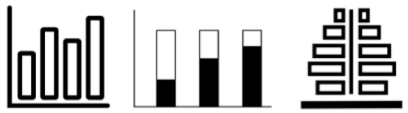


Chart types: a basic building block

- Most tools just support common statistical charts

Common Statistical Charts

Bar Chart
Standard | Stacked | Divergent



Special Cases
- Epidemic Curve
- Diversity Chart
- LefSe Plot

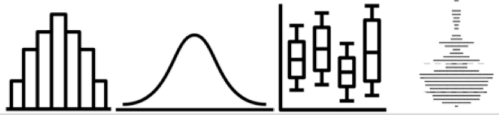
Line Chart

Special Cases
- Bootscan
- Kaplan-Meier
- Skyline Plot

Scatter Plot

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

Distribution Plot
Histogram | PDF | Boxplot | Swarm Plot



Pie Chart

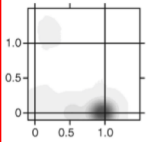
Venn Diagram

Colour Charts

Category Stripe

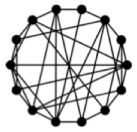
Heatmap

Density Plot*



Relational Charts

Node-link

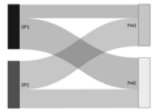


Special Cases
- *eBurst*
- Social network
- Molecular network
- Minimum Spanning Tree

Flow Diagram

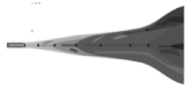
Chord Diagram

Sankey Diagram

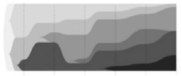


Temporal Charts

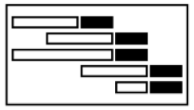
Streamgraph*
Absolute



Relative



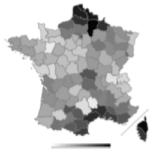
Timeline



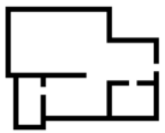
Spatial Charts

Geographic Map

Choropleth Map

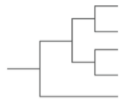


Interior Map



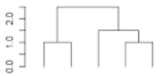
Tree Charts

Phylogenetic Tree
Rooted (Linear & Radial)



Unrooted
(Linear & Radial)

Dendrogram

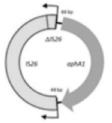


Clonal Tree*

Genomic Charts

Genomic Map
Linear

Radial



Alignment

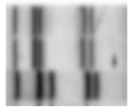
Composition Plot

Sequence Logo Plot

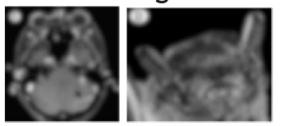
Other Charts

Table

Image
Gel Image



General Image



Miscellany

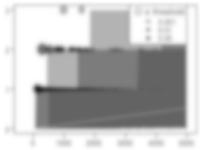
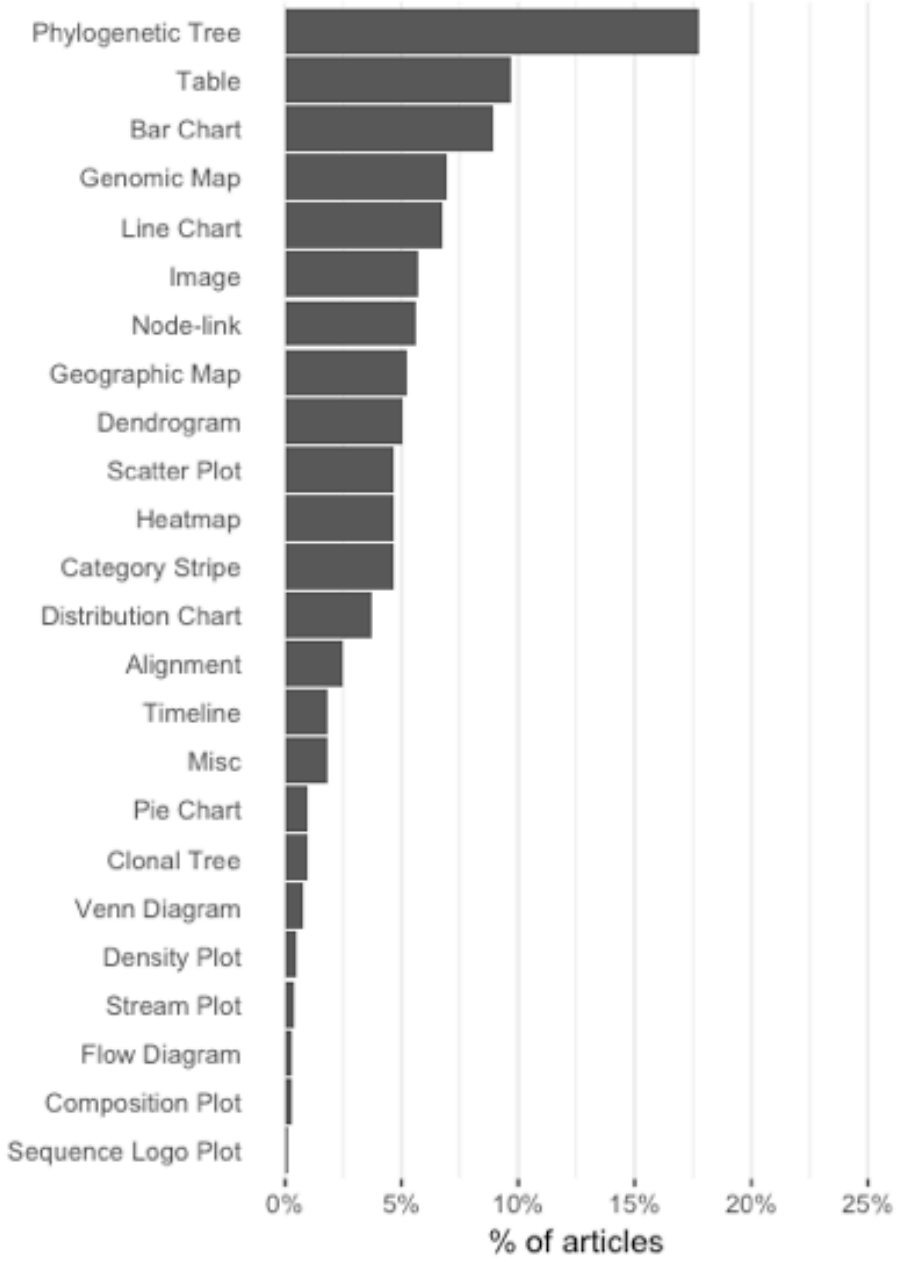


Chart types: a basic building block



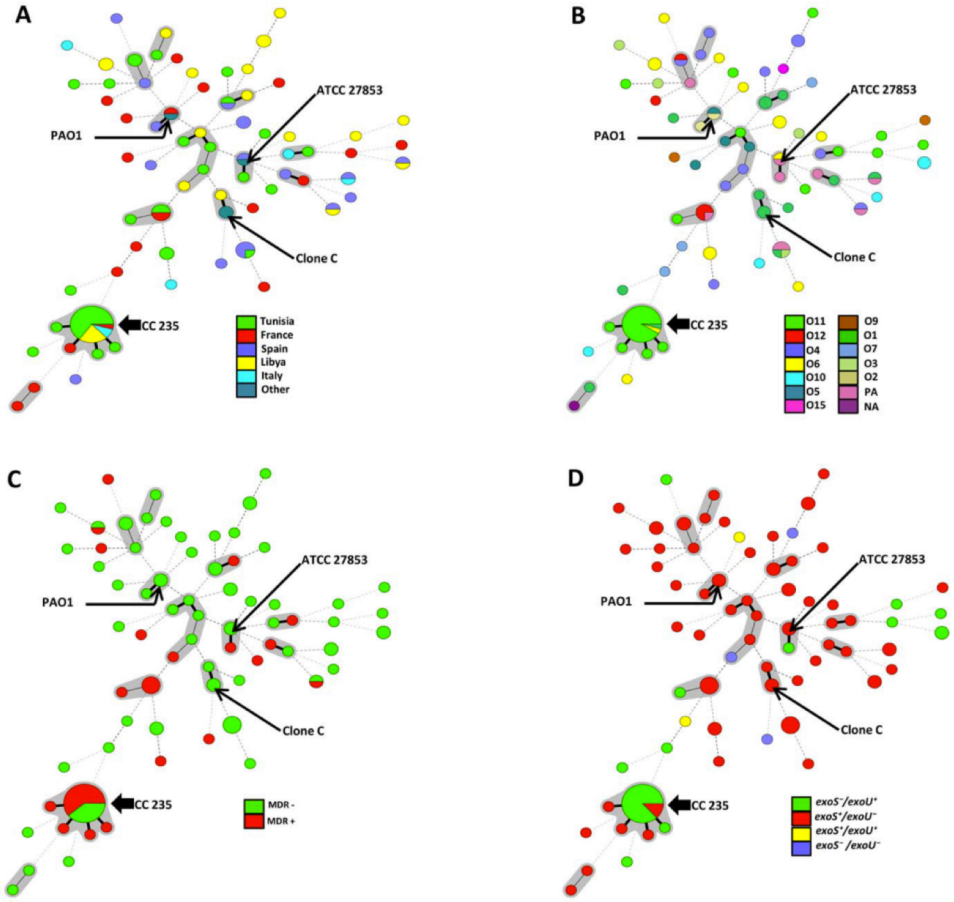
Current Common Practice

- 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

Chart Combinations: showing even more data

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

Example: Same chart type, different metadata



Example: Two chart types together

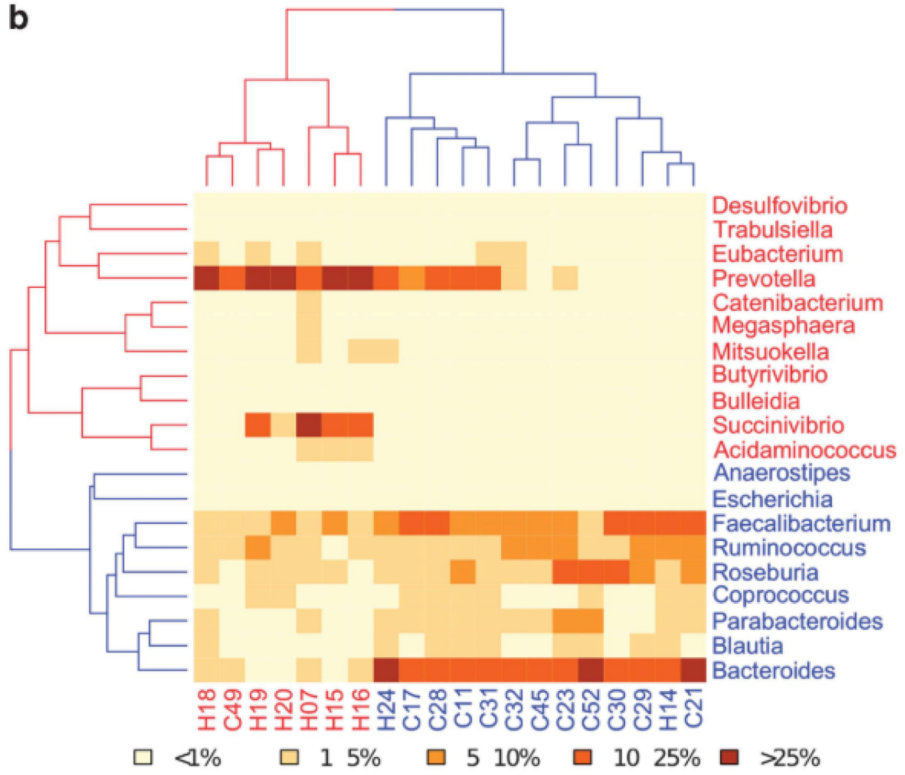


Chart Combinations: showing even more data

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 Types *Linked*	Many	Many	Visual, but not spatial	
Many Types *General*	Many	Many	NA	
Complex Combinations	Many	Many	Context dependent	

Chart Combinations: showing even more data

Combination Type	# of chart types	# of charts	Linkage type	*Example*	Current Practice
Simple	1	1	NA		40.1% of all figures
Composite	Many	1	Spatially Aligned		20.3%
Small Multiples	1	Many	Chart Type & Data		17.3%
Many Types *Linked*	Many	Many	Visual, but not spatial		13.5%
Many Types *General*	Many	Many	NA		8.8%
Complex Combinations	Many	Many	Context dependent		11.9%

Chart Enhancements: overlaying metadata

- **Mark = basic graphical element (line, point, area)**
- **Enhancement = adding marks or re-encoding marks of the base chart type**
- **Built upon design language in information visualization (infovis) literature**

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

Chart Enhancements: overlaying metadata

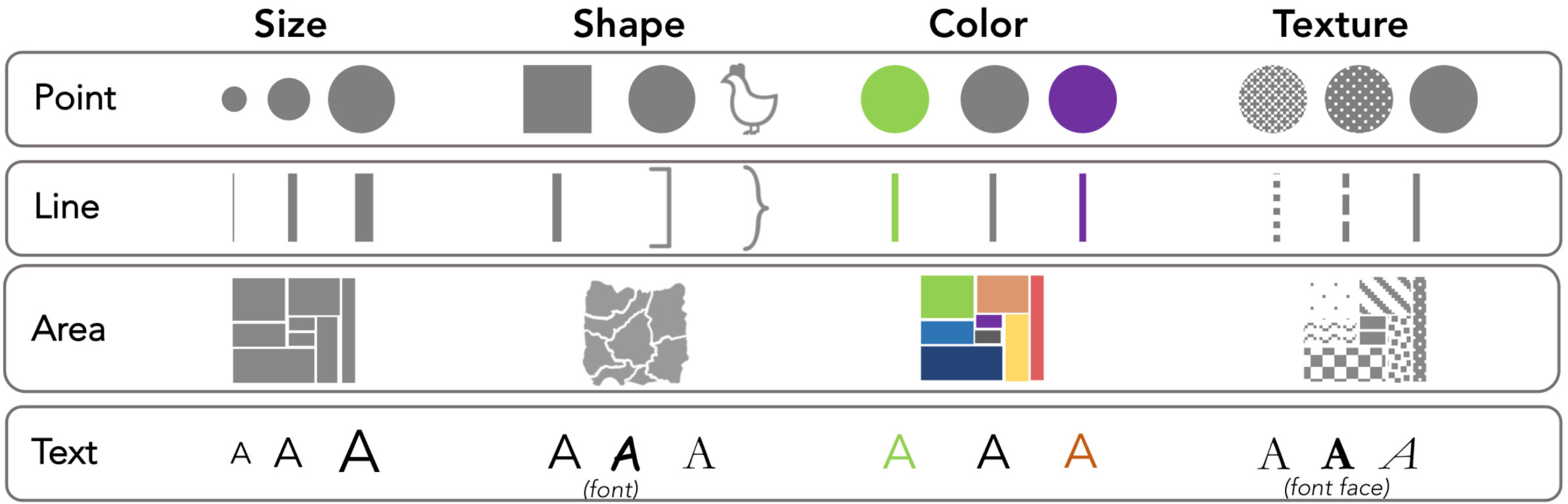


Chart Enhancements: overlaying metadata

- **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.

Current Practice 59.6% of all figures | 45.6% | 33.6%

Chart Enhancements: overlaying metadata

Base Chart

Tree

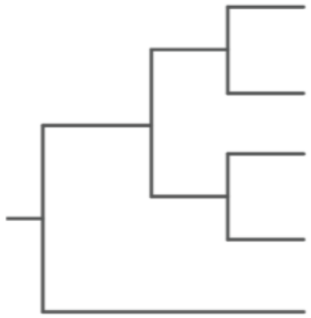
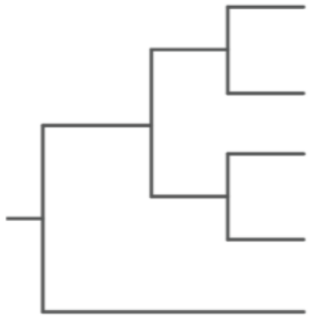


Chart Enhancements: overlaying metadata

Structured Enhancements

Base Chart
Tree



Re-encode Marks
Line: *color*

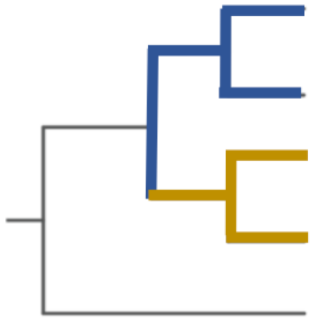
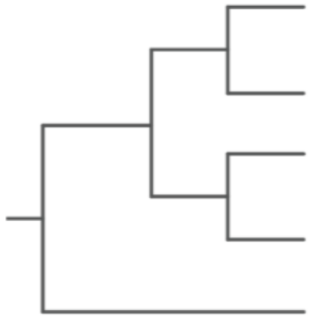


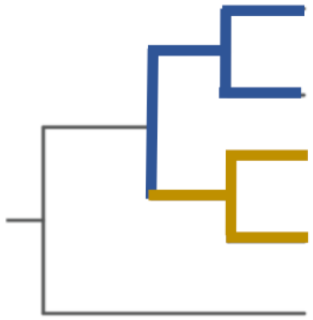
Chart Enhancements: overlaying metadata

Structured Enhancements

Base Chart
Tree



Re-encode Marks
Line: *color*



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

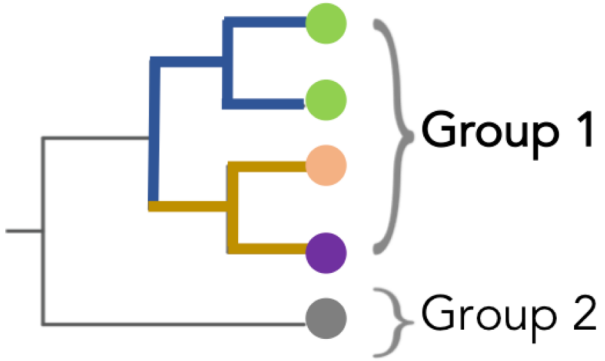


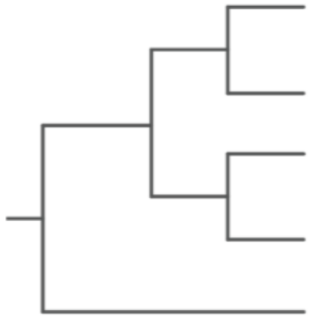
Chart Enhancements: overlaying metadata

Structured Enhancements

Unstructured Enhancements

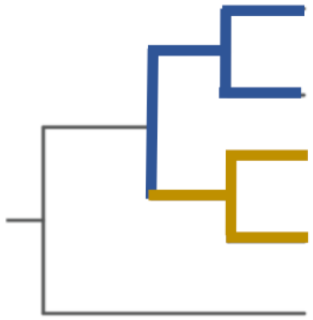
Base Chart

Tree



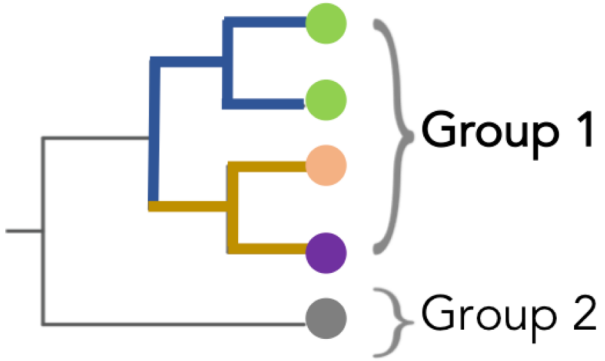
Re-encode Marks

Line: *color*



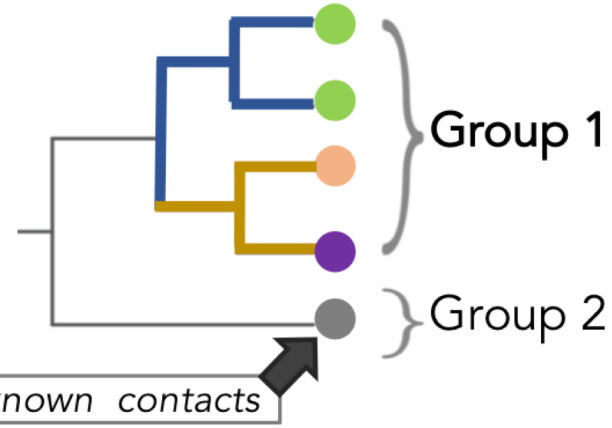
Add Marks

Point: *color*; line; text: *font face*



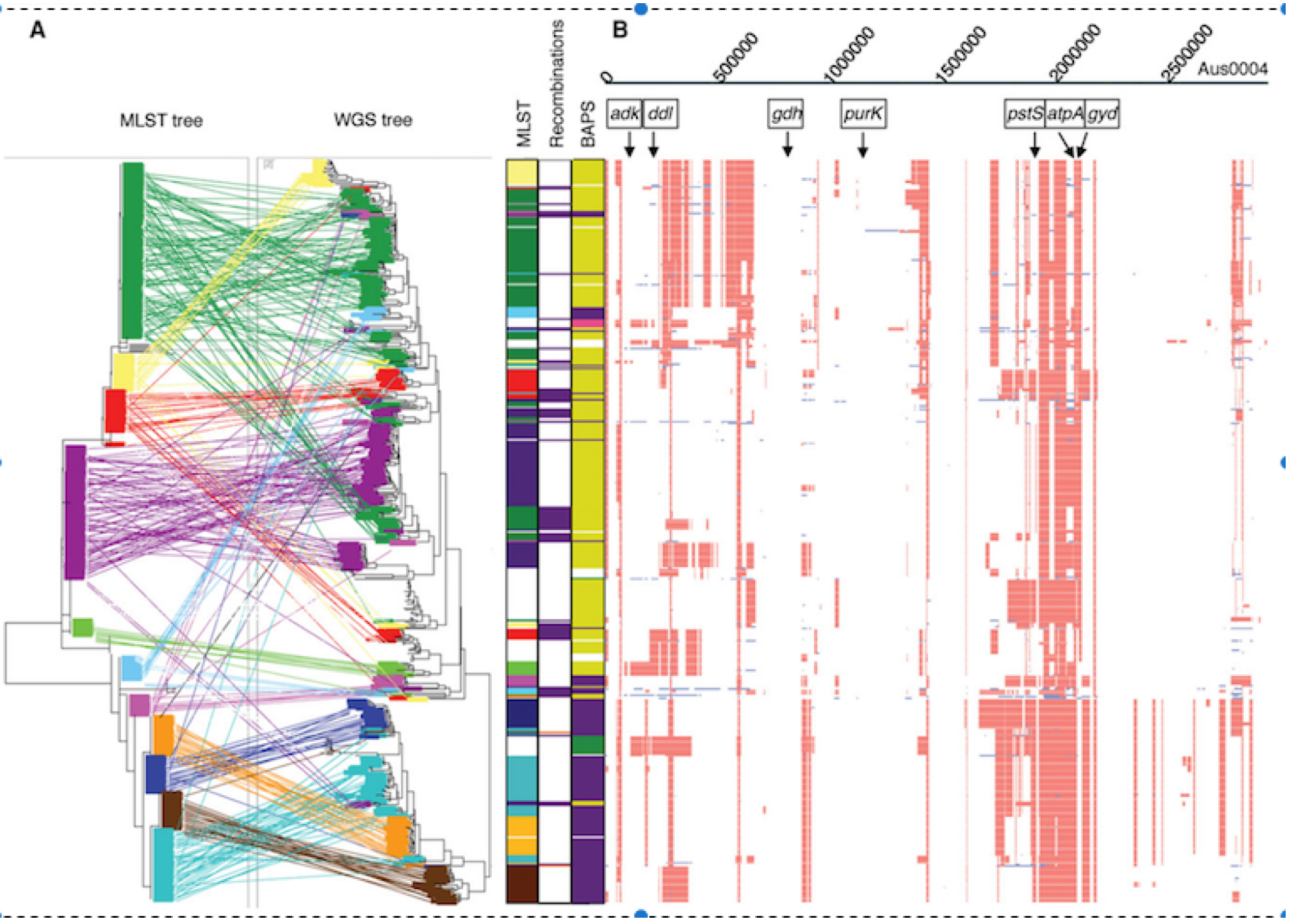
Add Annotation

Arrow, text



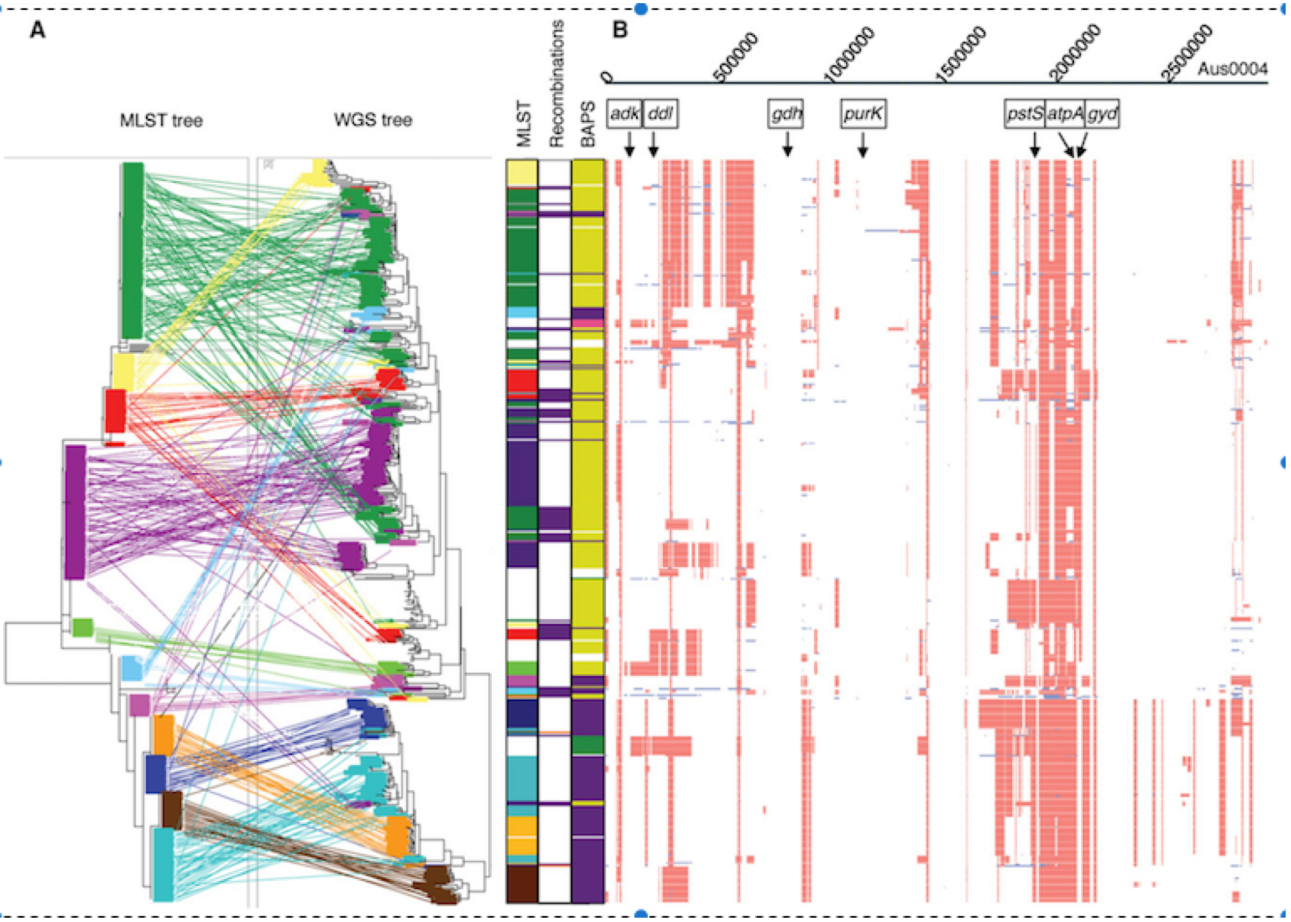
GEViT in action!

Gorrie (2017)



GEViT in action!

Gorrie (2017)



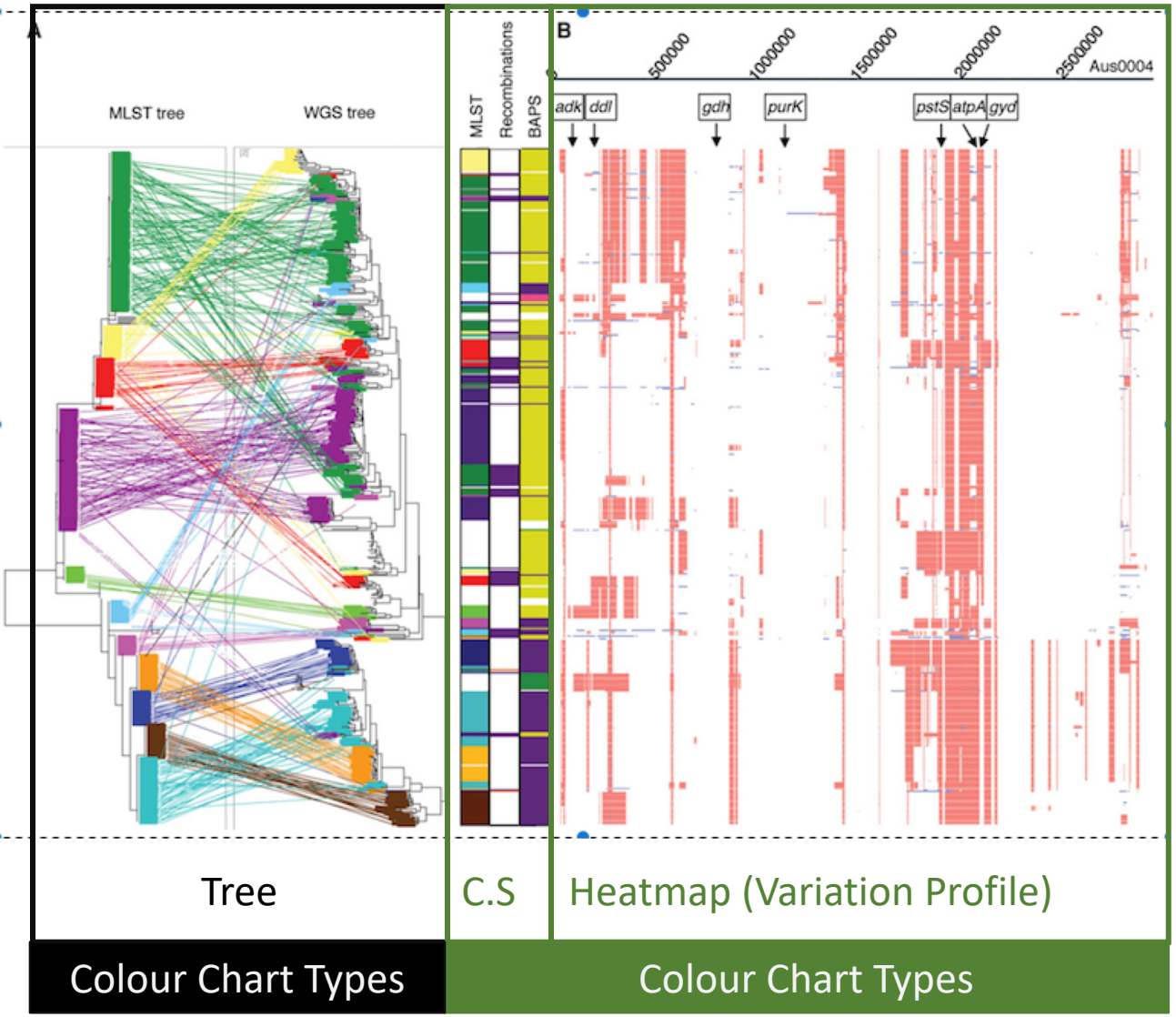
Visualization Breakdown

Literature Analysis ***(why)***

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

GEViT in action!

Gorrie (2017)



Tree | C.S | Heatmap (Variation Profile)

Colour Chart Types | Colour Chart Types

Visualization Breakdown

Literature Analysis ***(why)***

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

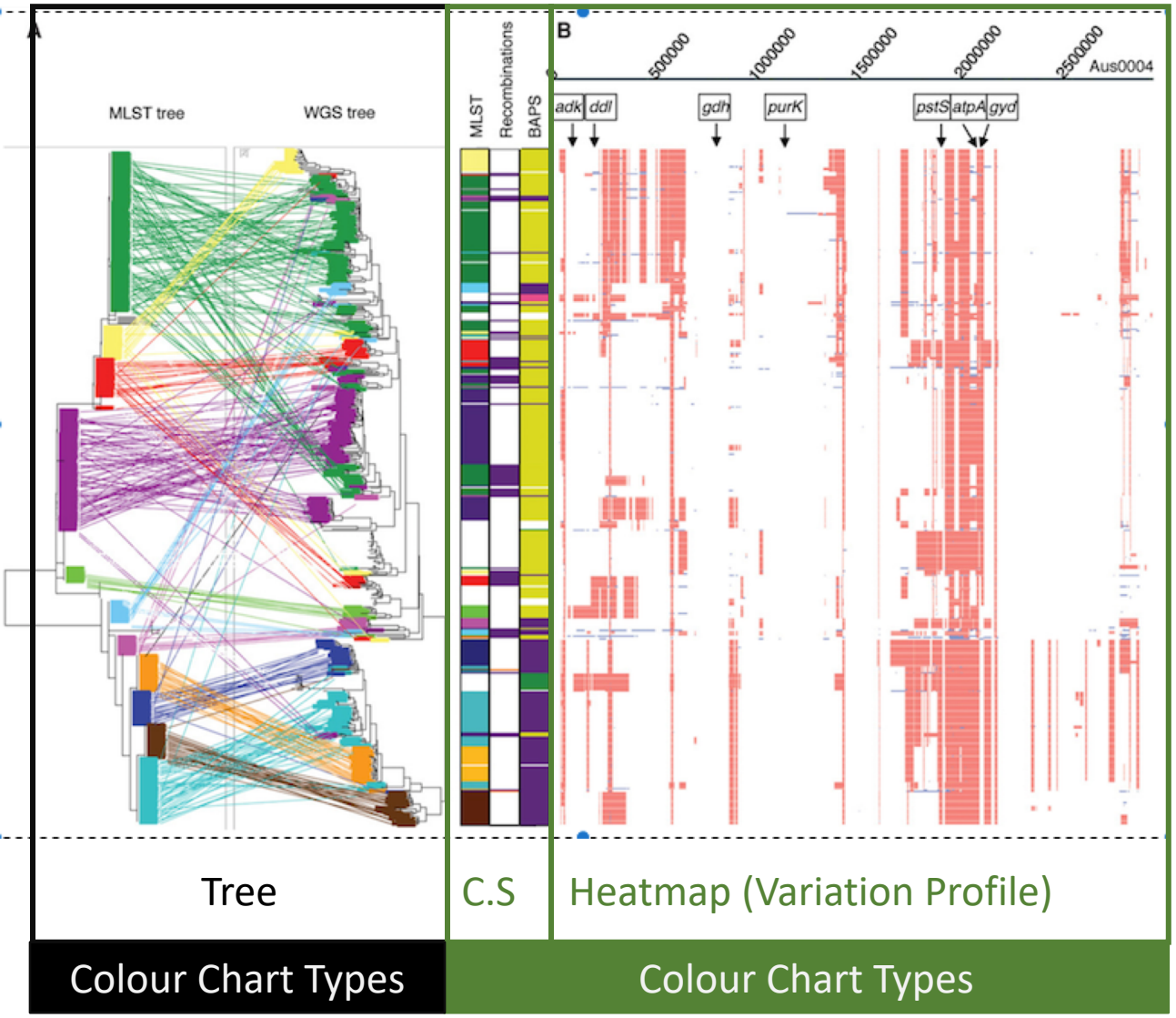
Visualization Components ***(how)***

Chart Type	Tree (Rooted Phylogenetic Tree) Category Stripe Heatmap (Variation Profile)

Colour chart types are common statistical charts that rely on color in their design

GEViT in action!

Gorrie (2017)



Visualization Breakdown

Literature Analysis ***(why)***

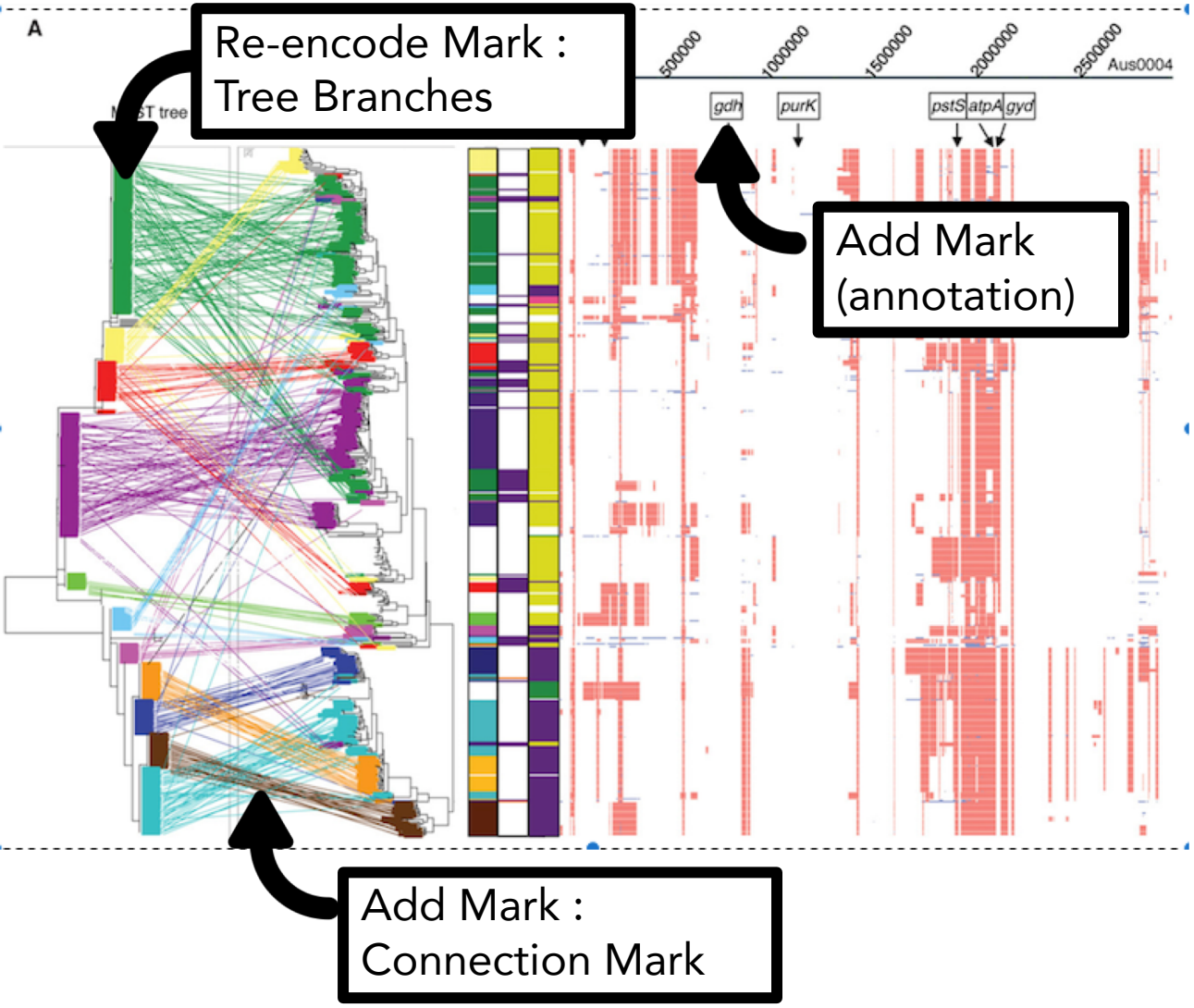
- ***Pathogen****: Enterococcus faecium*
- ***A priori concepts:*** *control; genome; outbreak; drug resistance; phylogeny; genotype*

Visualization Components ***(how)***

Chart Type	Tree (Rooted Phylogenetic Tree) Category Stripe Heatmap (Variation Profile)
Chart Combination	Composite *(spatially aligned)*

GEViT in action!

Gorrie (2017)



Visualization Breakdown

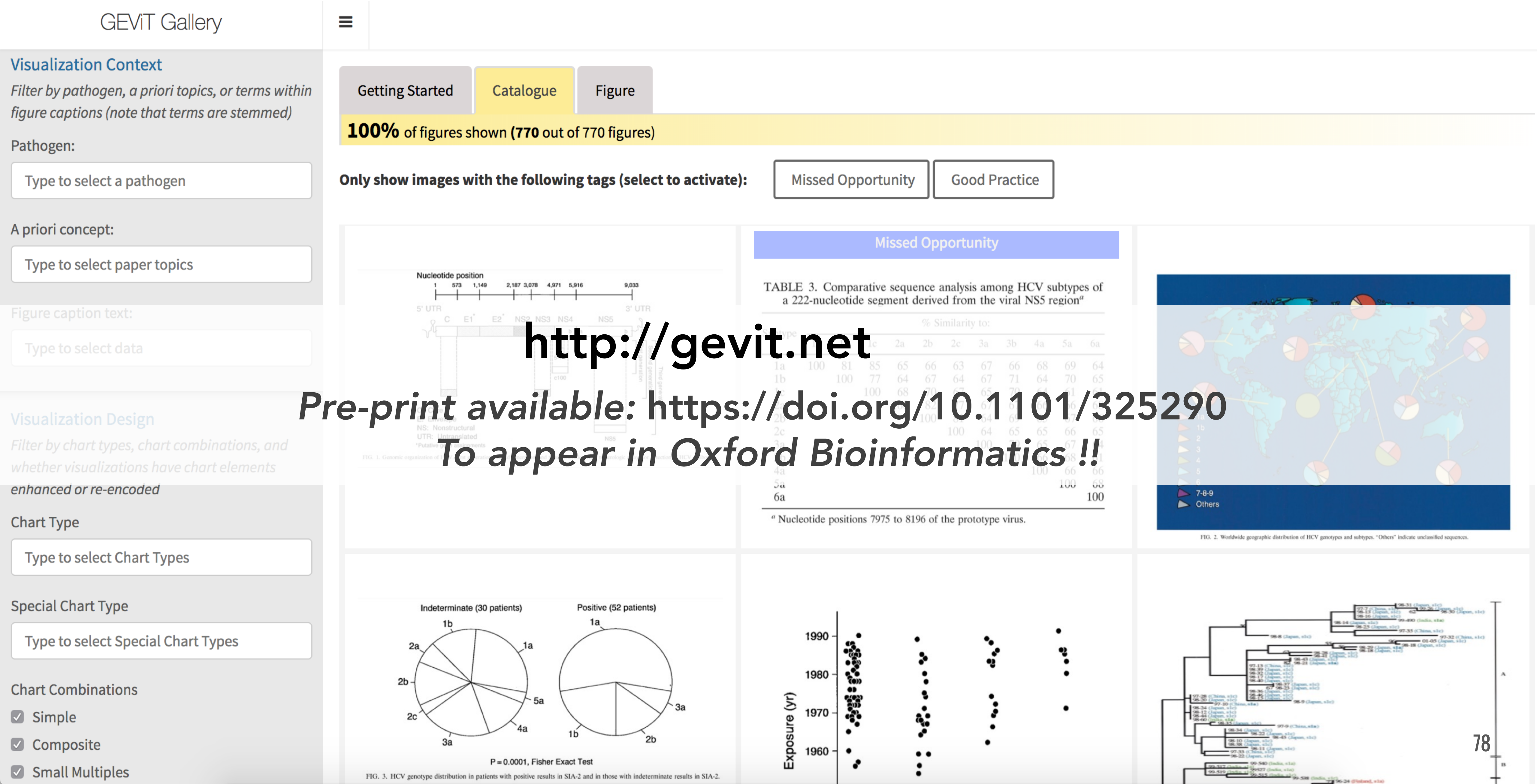
Literature Analysis ***(why)***

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

Visualization Components ***(how)***

Chart Type	Tree (Rooted Phylogenetic Tree) Category Stripe Heatmap (Variation Profile)	
Chart Combination	Composite *(spatially aligned)*	
Chart Enhancement	Re-encode Marks	Tree – *branches*
	Add Marks	Tree - *Connection Marks*
	Add Mark *(unstructured)*	Heatmap – *Textboxes*

GEViT in action!



http://gevit.net

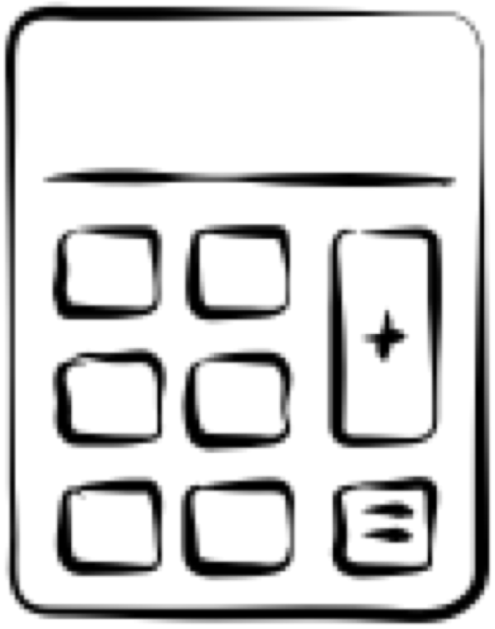
Pre-print available: https://doi.org/10.1101/325290

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

4 Current Common Visualization Practices



Descriptive statistics have been sprinkled throughout the results

Let’s look at the big picture

An appraisal of the design space

- **Wide variety of visualization quality** 🧐
  - Only possible to assess this with systematic approach

An appraisal of the design space

- **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

An appraisal of the design space

- **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
- **Visualized data shows wide variety of design space used** 👏
  - Some use chart types, combinations, and enhancements better than others
  - Still, maybe community can't see the forest for the tree.. so to speak..

An appraisal of the design space

- **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
- **Visualized data shows wide variety of design space used** 👏
  - Some use chart types, combinations, and enhancements better than others
  - Still, maybe community can't see the forest for the tree.. so to speak..
- **Current visualizations will not scale for big data** 😞

An appraisal of the design space

- **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
- **Visualized data shows wide variety of design space used** 👏
  - Some use chart types, combinations, and enhancements better than others
  - Still, maybe community can't see the forest for the tree.. so to speak..
- **Current visualizations will not scale for big data** 😞
- **Many visualizations not understandable by other public health professionals** 😩
  - Our prior work indicates trees are hard to interpret

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 even 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 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

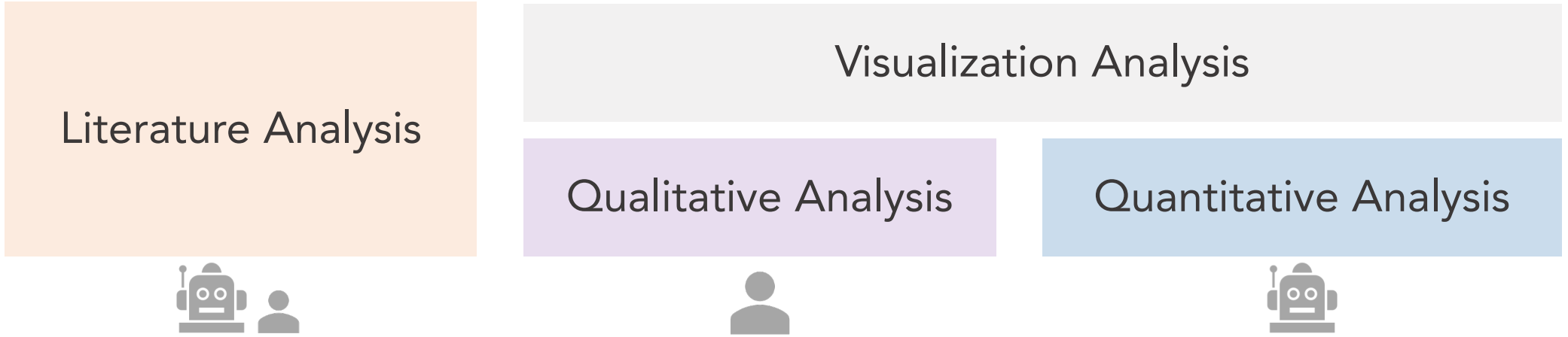
Implications of our research findings

- **Need to move away from *ad hoc* visualization development**
  - Need awareness of design space
  - 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)
  - Design space variance tells you easy/hard it for a community to adapt new data vis
  - Source of inspiration for researchers

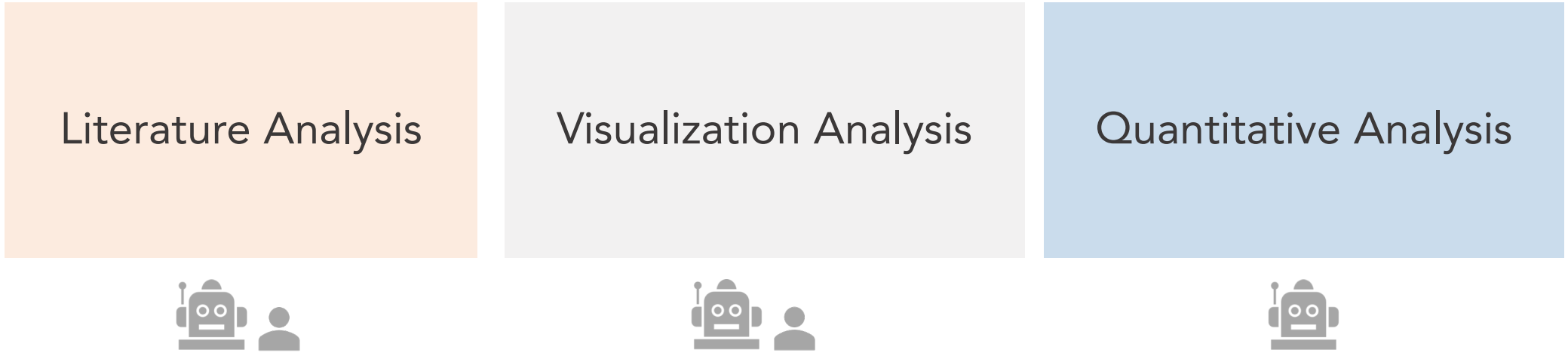
Next steps

Automating design space creation

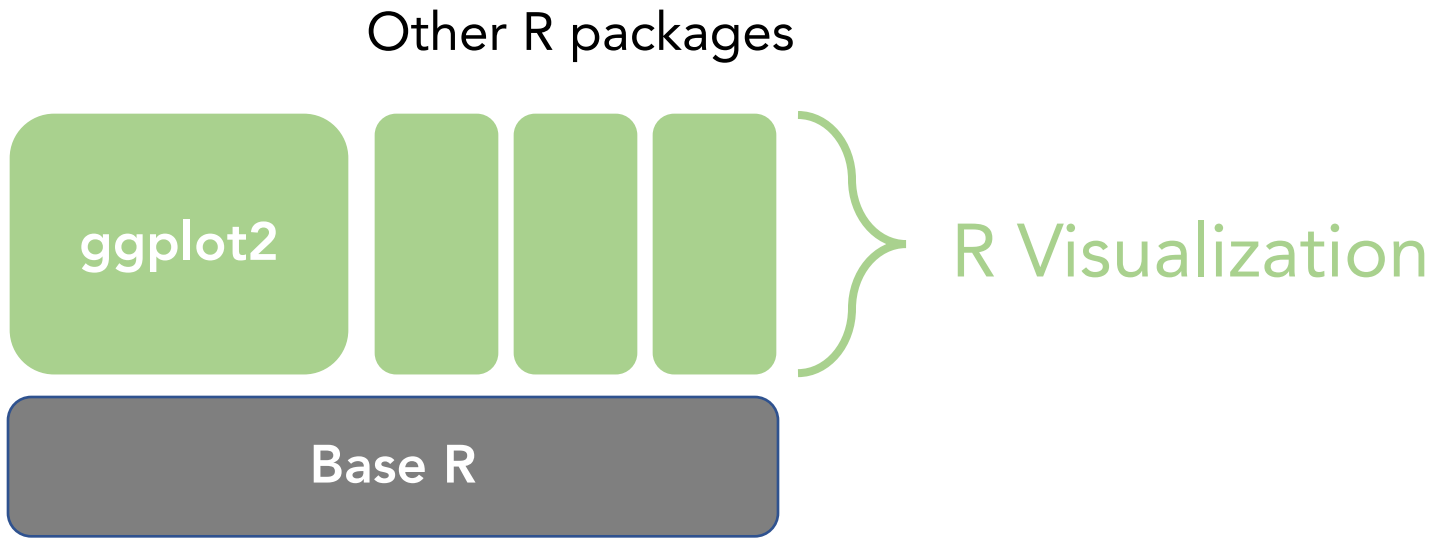
Current: Still needs (wo)man power to construct design space



Future: Human-in the loop, but driven by automated methods

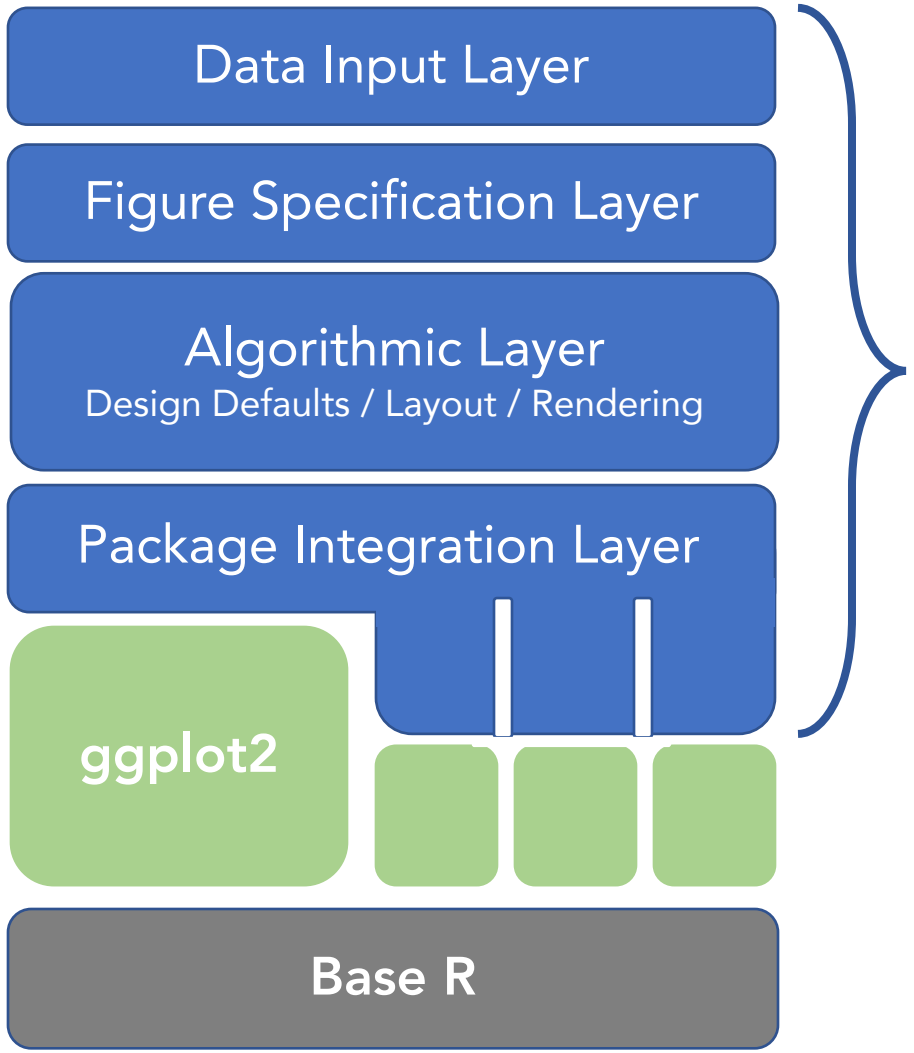


Using R infrastructure as a based



Building data visualization tools with GEViT

Creating visualizations with simple and minimal syntax

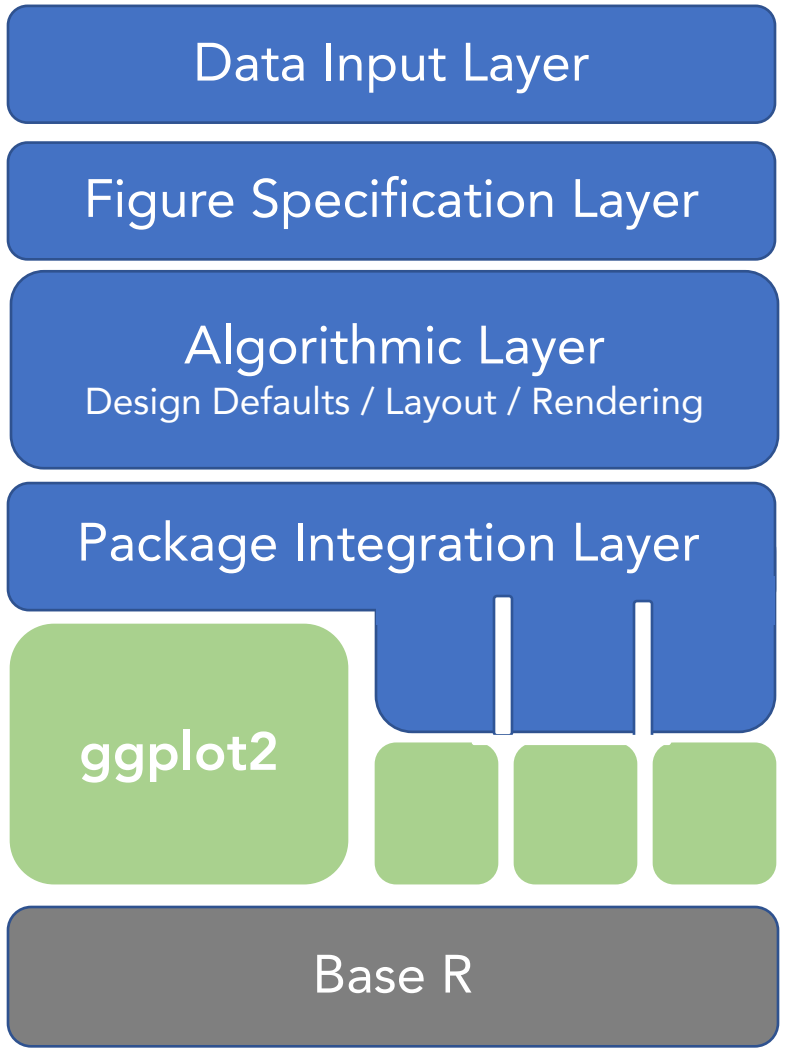


GEViT API

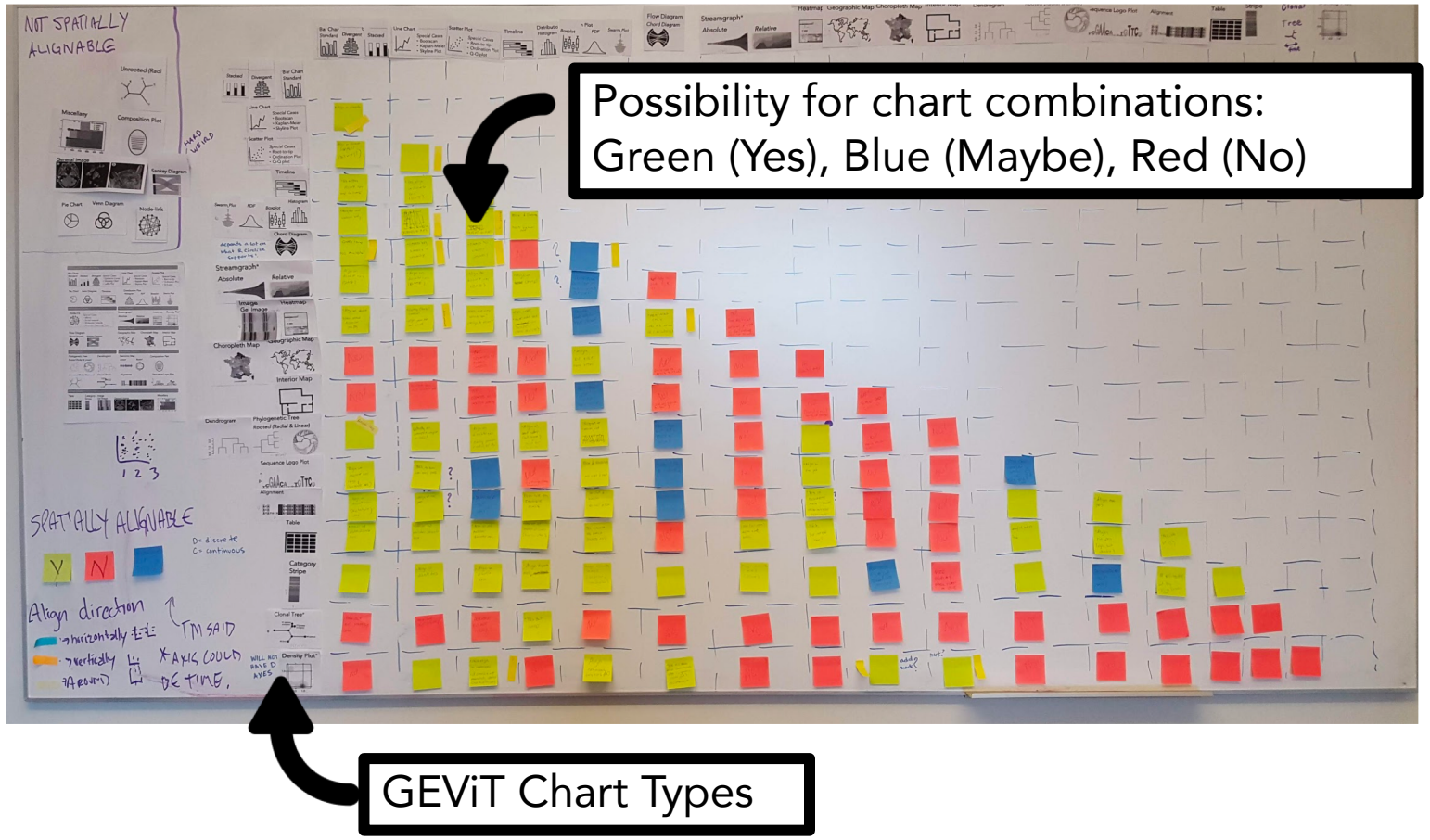
- GEViT design space findings drive figure specifications
- Specifications drive plotting, layout, and rendering functions

Building data visualization tools with GEViT

Development process reveals many interesting challenges



Designing chart combinations layout algorithm



Thanks!

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

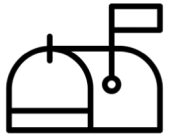
I am graduating soon!
I am on the job market this year!

Pre-Print + Other Stuff

https://doi.org/10.1101/325290

https://amcrisan.github.io/talks

http://gevit.net



acrisan@cs.ubc.ca

Go forth and analyze!

Icons by noun project and images from pexels

Creating explorable visualization design spaces:
An example from 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]