

# Visualizing Public Health Data

Anamaria Crisan, MSc

PhD Candidate

University of British Columbia



@amcrisan



acrisan@cs.ubc.ca



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

# I will attempt to make two points

- The differences between clinical medicine and public health and the vis implications
- Provide an overview of the state of visualization within a public health domain

# Public Health and Clinical Medicine

## **Clinical Medicine**

- Targets individual patients

## **Public Health**

- Targets populations

## **Clinical Medicine**

- Targets individual patients
- Diagnosis and treatment focused

## **Public Health**

- Targets populations
- Prevention and control focused

## **Clinical Medicine**

- Targets individual patients
- Diagnosis and treatment focused
- Interventions are typically pharmaceutical interventions

## **Public Health**

- Targets populations
- Prevention and control focused
- Pharmaceutical, but also more commonly environmental and behavioral interventions

## Clinical Medicine

- Targets individual patients
- Diagnosis and treatment focused
- Interventions are typically pharmaceutical interventions
- Decision makers typically siloed specialists (doctors, nurses, etc.)

## Public Health

- Targets populations
- Prevention and control focused
- Pharmaceutical, but also more commonly environmental and behavioral interventions
- Decision makers diverse, not necessarily specialists (community leaders, etc.)

# Clinical Medicine vs. Public Health

## Clinical Medicine

- Targets individual patients
- Diagnosis and treatment focused
- Interventions are typically pharmaceutical interventions
- Decision makers typically siloed specialists (doctors, nurses, etc.)

*Example:*

**Treating lung cancer patient**

## Public Health

- Targets populations
- Prevention and control focused
- Pharmaceutical, but also more commonly environmental and behavioral interventions
- Decision makers diverse, not necessarily specialists (community leaders, etc.)

**Anti-smoking campaign**

*Both use data, even the same data, in different ways*

# Visualization consumers in clinical medicine

- Currently data vis tends to emphasize clinical medicine applications and targets clinicians, researchers, and patients

Clinicians



Researchers



Patients



# Visualization consumers in public health

- Public Health has much more multidisciplinary decision making teams
  - More data & diverse data types = more informed decision making
  - BUT – different stakeholder abilities to interpret data & different needs
- Gap: few vis applications for public health

Medical  
Health  
Officers



Clinicians



Nurses



Researchers



Community  
Leaders



Patients



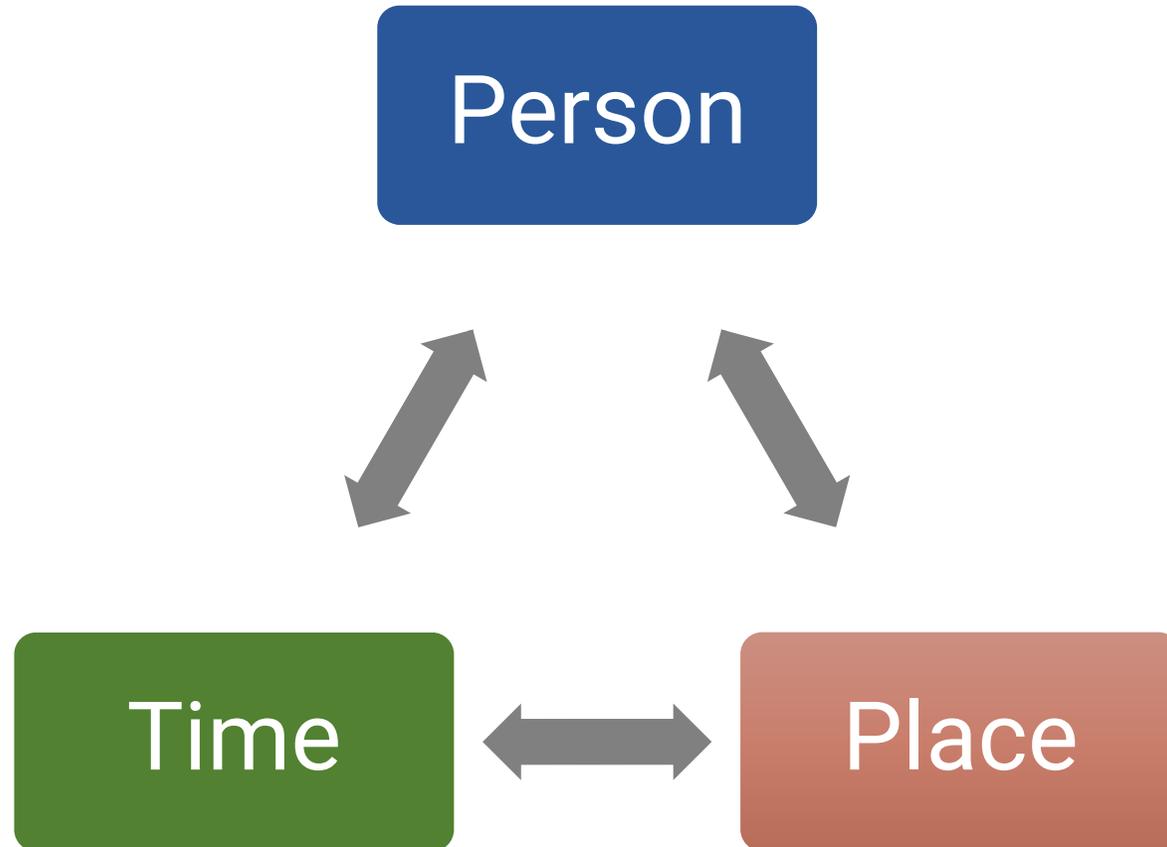
Politicians



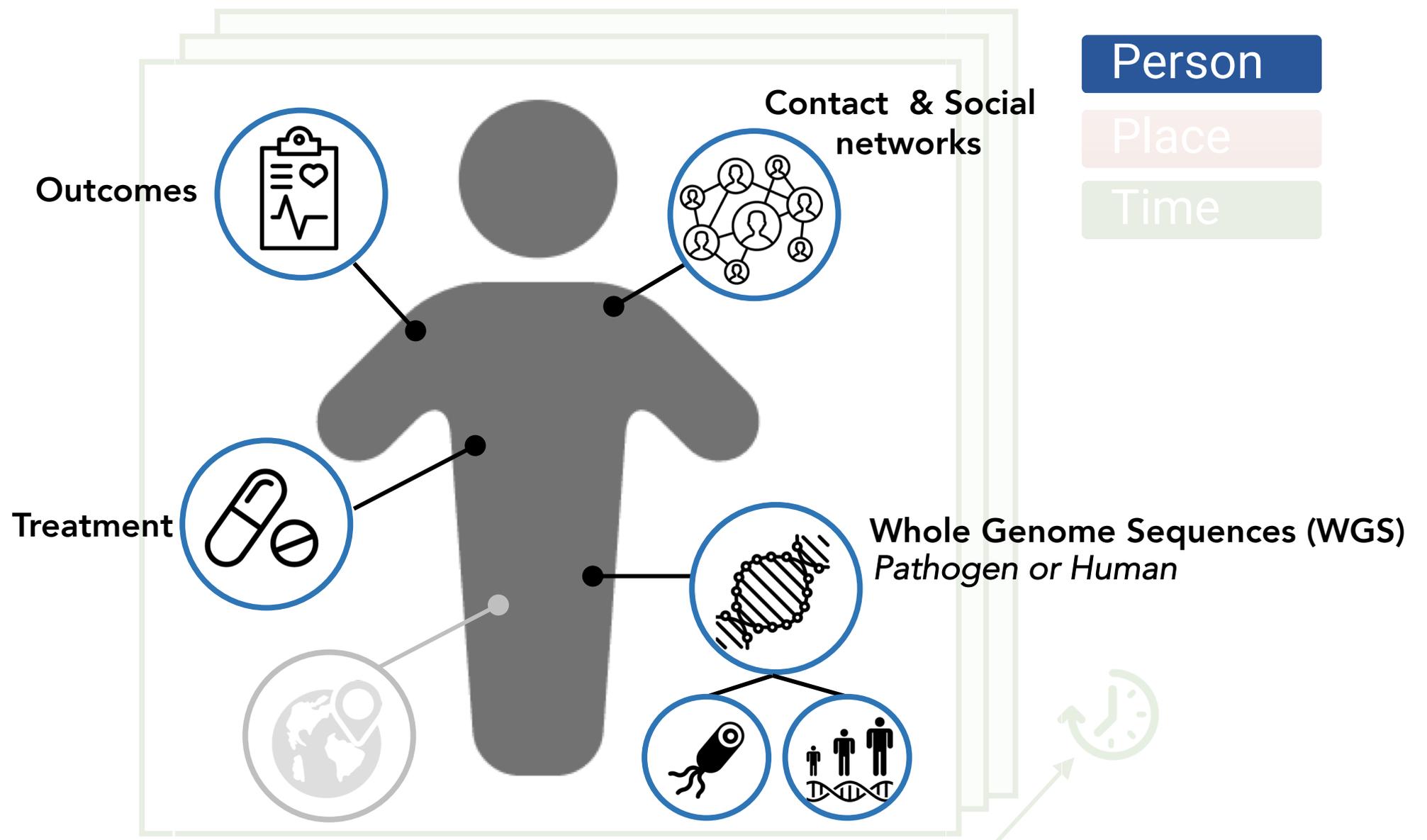
# What are Public Health Data?

# What are Public Health Data?

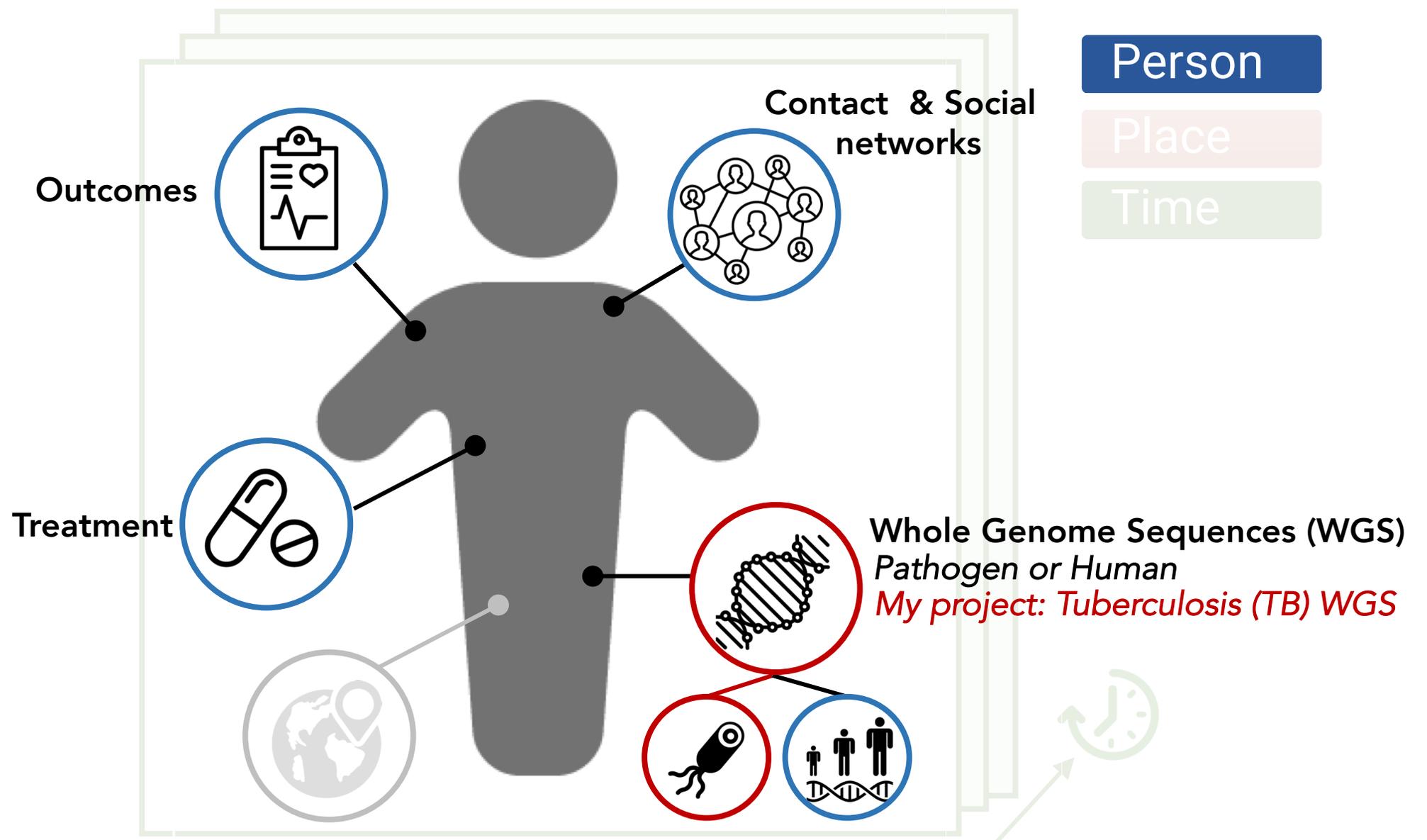
## The Epidemiological Trinity



# What are Public Health Data?



# What are Public Health Data?



# What are Public Health Data?

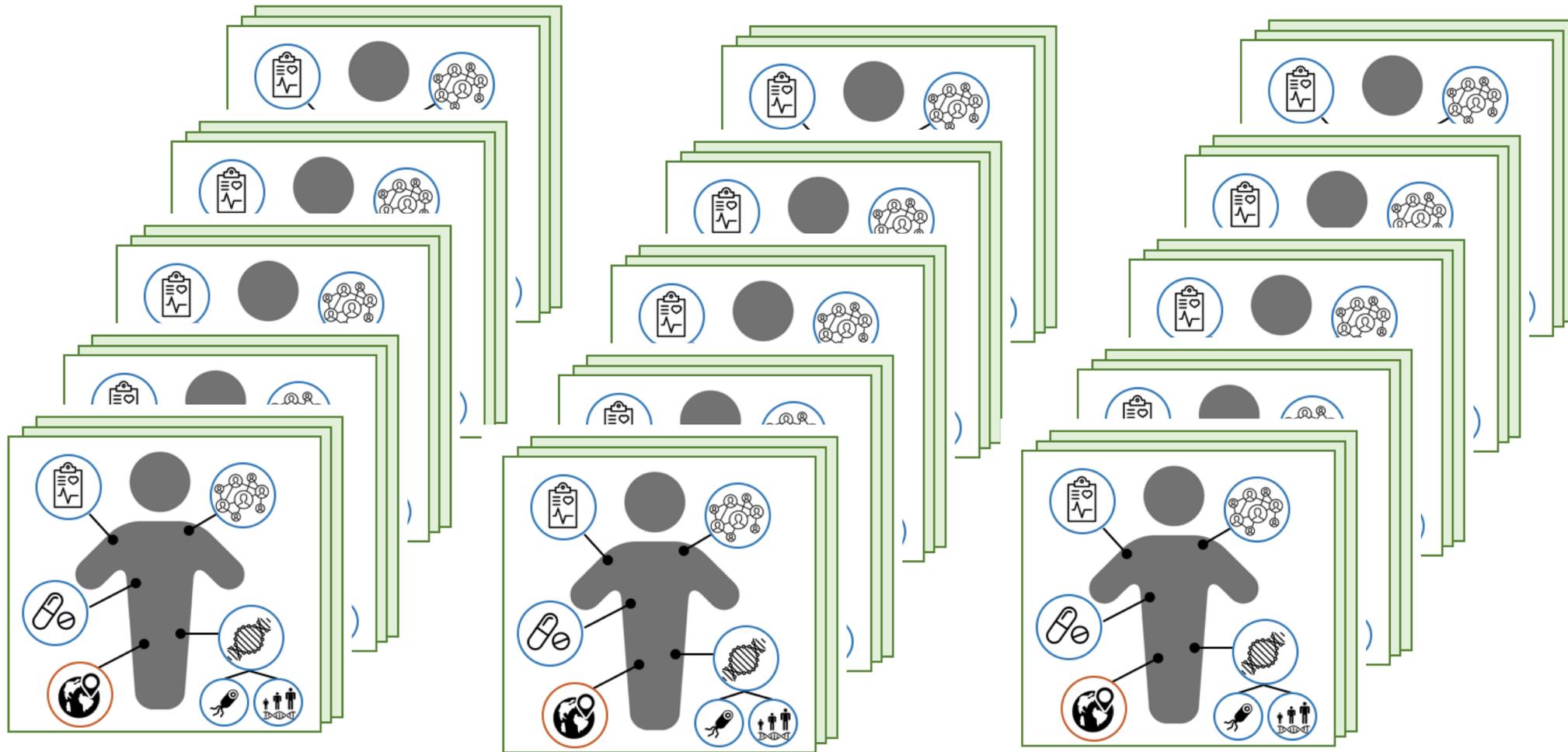


# What are Public Health Data?



# What are Public Health Data?

Via EHRs data are passively collected about entire populations over time



# **The State of Data Vis in Public Health**

# The state of visualization in public health

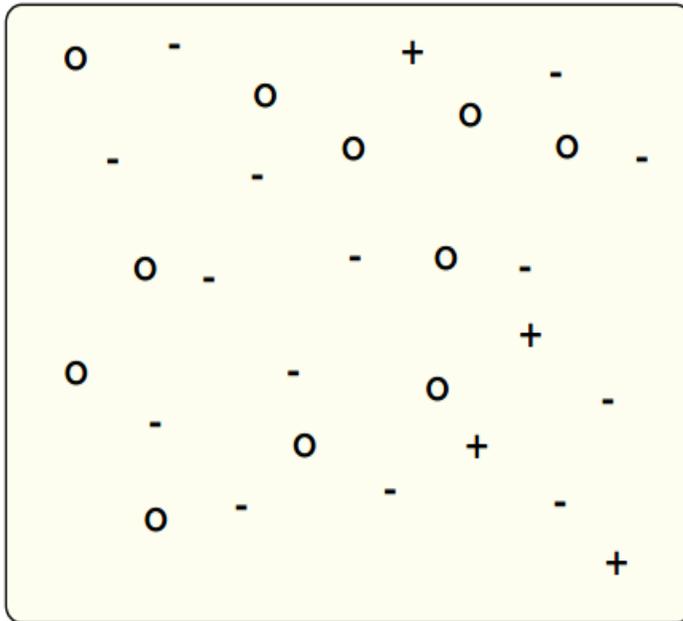
- Barriers for creating data visualizations are lowering
  - Many domain specialists (scientists, public servants) routinely create data visualizations
- Guidance on what makes a good data visualization is absent
  - Domain specialists don't read the vis literature
- Lack of guidance = inefficient unsupervised exploration of vis design space
  - "Hit or Miss" ad hoc design solutions

# The state of visualization in public health

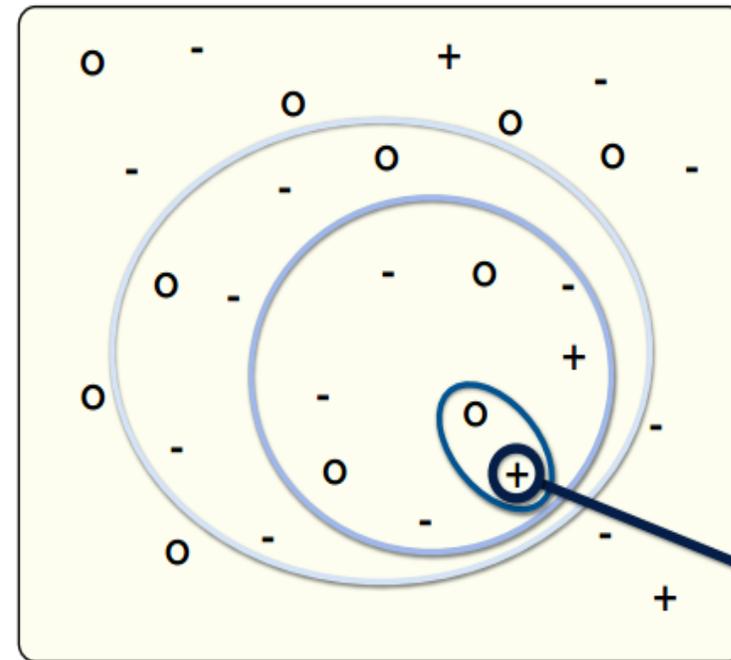
- Barriers for creating data visualizations are *lowering*
  - Many domain specialists (scientists, public servants) routinely create data visualizations
- Guidance on what makes a good data visualization is *absent*
  - Domain specialists don't read the vis literature
- Lack of guidance = inefficient unsupervised exploration of vis design space
  - "Hit or Miss" ad hoc design solutions
- **Our proposed solution: systematically create an explorable vis design space**

# Design Spaces : A quick primer

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



+ good  
o okay  
- poor



know

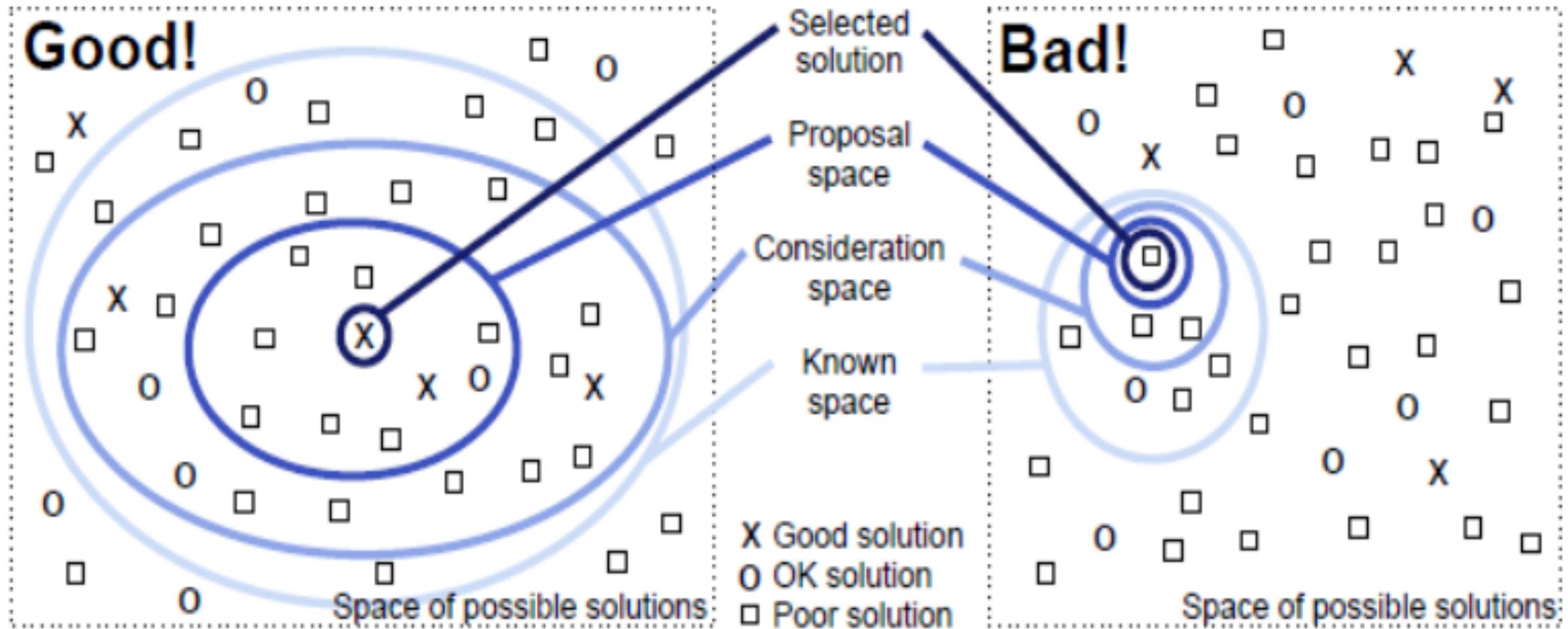
consider

propose

select

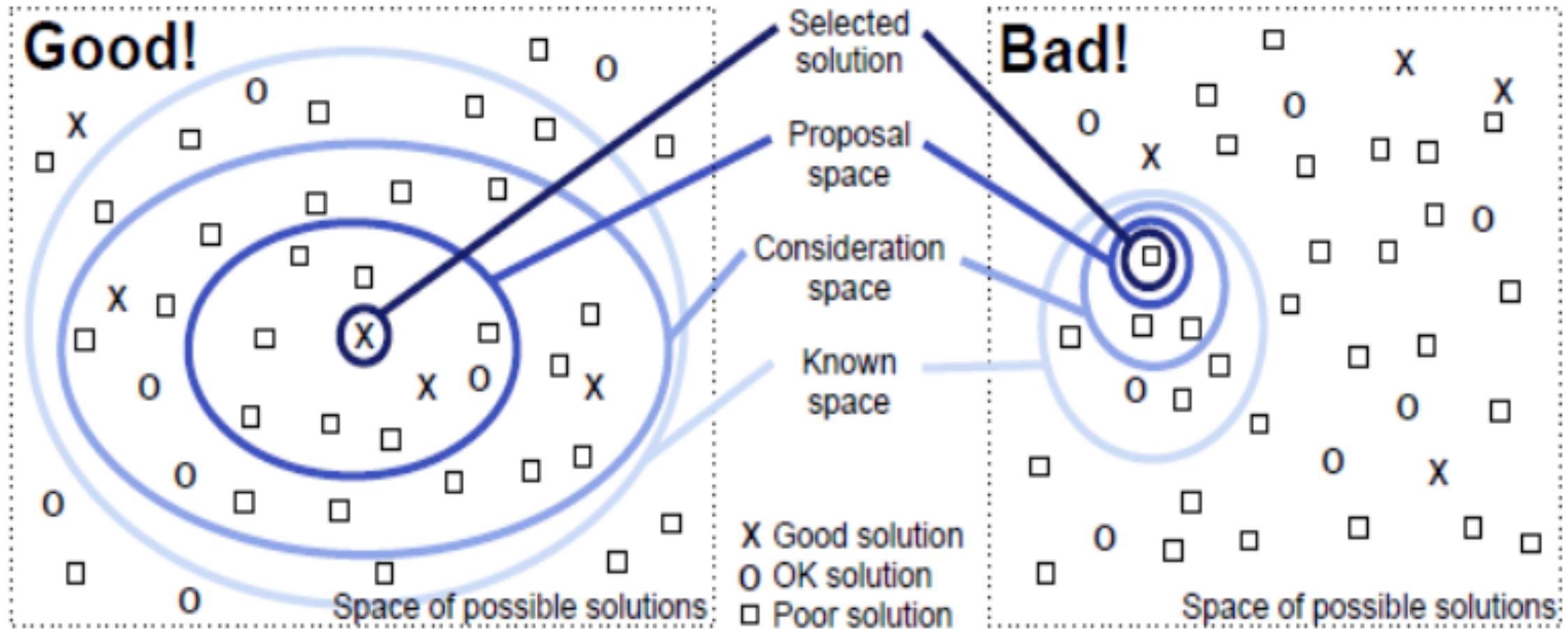
# Design Spaces : A quick primer

GOAL – nudge domain specialists toward better design choice solutions



# Design Spaces : A quick primer

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



# Constructing a design space

- Our observation: there's a lot of figures in research papers, let's study them!
- Challenge: methods for systematic assessment of data visualizations don't exist
  - Systematic matters! Shows the good, the bad, and the common
  - Existing studies (setvis, treevis, vishealth) are not systematic reviews of specialist's domain
- We combined methods from epidemiology with infovis to construct a design space

# Our approach allows us to answer three different questions

- Scope: Infectious Disease Genomic Epidemiology literature
- Objective: Identify and enumerate the kinds of visualizations generated for different topics of infectious disease genomic epidemiology

# Our approach allows us to answer three different questions

- Scope: Infectious Disease Genomic Epidemiology literature
- Objective: Identify and enumerate the kinds of visualizations generated for different topics of infectious disease genomic epidemiology

Literature  
Analysis

Qualitative Data  
Visualization Analysis

Quantitative Data  
Visualization Analysis

# Our approach allows us to answer three different questions

- Scope: Infectious Disease Genomic Epidemiology literature
- Objective: Identify and enumerate the kinds of visualizations generated for different topics of infectious disease genomic epidemiology

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?

Catalogue

Figure

Only show images with the following tags (select to activate):

Missed Opportunity

✓ Good Practice

Good Practice

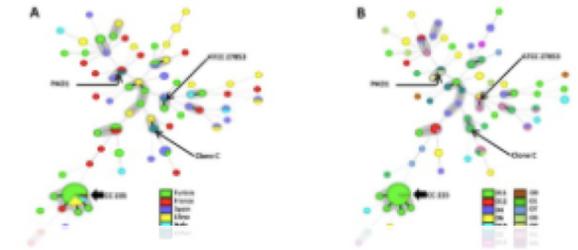


Figure 1. Minimal Spanning Tree (MST) analysis of *Phlebotomus perniciosus* strains based on MLST data. Each circle corresponds to an ST. The area of each circle corresponds to the number of isolates. The relationships between strains are indicated by the connections between the nodes and the lengths of the branches joining them. Black lines connecting pairs of STs indicate that they differ in one allele (black lines), two alleles (grey), or three or more alleles (colored). Only accessions (STs) that belong to the same clonal complex (clonal complex were defined from this collection, and CC213 was the predominant). Four MST graphs were generated separately based on the following associations: A: ST vs countries, B: ST vs serogroup, C: ST vs mating occurrence and D: ST vs enzootic.

Good Practice

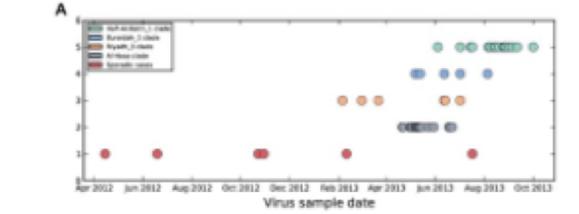


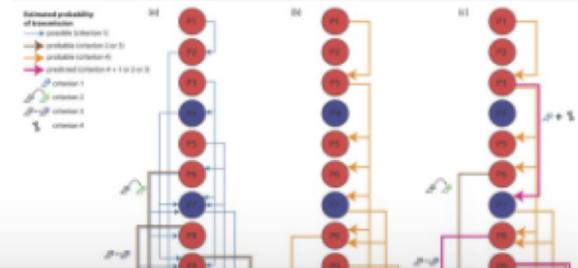
FIG 4. Distribution of NGS-GVP data over time and space. (A) All available NGS-GVP genomes were stratified by phylogenetic clade (see Fig. 1) and plotted by virus sample date. The length of each data row determined as the difference in days between the first and last observed sample of that virus and labeled the following values: All data (A), April 2012 to 22 June 2012 (B), April, 2012 to 10 June 2012 (C), 10 June 2012 to 1 July 2012 (D), 1 July 2012 to 1 Aug 2012 (E), 1 Aug 2012 to 1 Sept 2012 (F), 1 Sept 2012 to 1 Oct 2012 (G), 1 Oct 2012 to 1 Nov 2012 (H), 1 Nov 2012 to 1 Dec 2012 (I), 1 Dec 2012 to 1 Jan 2013 (J), 1 Jan 2013 to 1 Feb 2013 (K), 1 Feb 2013 to 1 Mar 2013 (L), 1 Mar 2013 to 1 Apr 2013 (M), 1 Apr 2013 to 1 May 2013 (N), 1 May 2013 to 1 June 2013 (O), 1 June 2013 to 1 July 2013 (P), 1 July 2013 to 1 Aug 2013 (Q), 1 Aug 2013 to 1 Sept 2013 (R), 1 Sept 2013 to 1 Oct 2013 (S), 1 Oct 2013 to 1 Nov 2013 (T), 1 Nov 2013 to 1 Dec 2013 (U), 1 Dec 2013 to 1 Jan 2014 (V). All available NGS-GVP genomes were stratified by phylogenetic clade (see Fig. 1) and plotted by the case location. Clades are indicated by the small black circles, and regional viruses by larger circles colored according to phylogenetic clade.

Good Practice

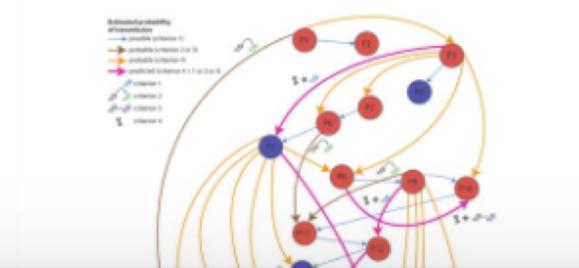


Figure 1. Distribution of influenza A (H1N1) and (H7N9) viruses, Guangdong Province, China, 2013-2014. Shading indicates percentage of environmental swab specimens from live poultry markets in each region that were positive for each influenza subtype by reverse transcription PCR. Circles indicate locations of human cases, larger circles indicate higher numbers of cases.

Good Practice



Good Practice



NA

Use the different filters below to navigate the GEVIT Gallery. To get more information about each filter click the **i** icon

Paper Lookup (PMID):

Visualization Context



Pathogen:

Topic:

Data (from figure captions):

Visualization Graphical Properties



Chart Type

Special Chart Type

[https://amcrisan.shinyapps.io/gevit\\_gallery/](https://amcrisan.shinyapps.io/gevit_gallery/)  
*Unpublished & still some work to be done so please don't distribute*

# Implications of our findings

- **Surprise finding: a lot of data in data visualizations were not visualized!**
- Pedagogical implications :
  - Can we give people more complex vis applications when their vis skills are kind of low?
  - How can we improve vis literacy?
  - I think a design space is a useful discussion tool
- Software develop implications:
  - Discussion of a design space in bioinformatics development
  - GEViT is resource to provide alternative designs
  - Alternative designs also see gaps in the where vis research is needed
- Human-in-the-loop implications:
  - Need to think beyond image recognition problems
  - Might be premature to apply AI methods (no good training data)

# **Additional Slides**

# An overview of our results so far

- **Literature Analysis:** Understanding the structure of genomic epidemiology papers promotes systematicity via intelligent sampling
  - Total sample ~18K papers on genomic epidemiology
  - Defined strata by pathogen (document structure) and a priori concepts (domain knowledge)
  - Literature analysis stratified sampling yielded ~850 figures for analysis from 221 papers
- **Qualitative Analysis:** Developed GEViT, a Genomic Epidemiology Visualization Typology
  - Developed a typology to systematically described charts using three descriptive axes: chart types, chart combinations, and chart enhancements
- **Quantitative Analysis:** It's nearly all phylogenetic trees, across all pathogens and concepts, but there's also a lot of tables and plain text
- **Surprising general conclusion:** most data is these data visualizations are not visualized

*“Identify and enumerate the kinds of visualizations generated per topic of i.d. genomic epidemiology”*

*“Identify and enumerate the kinds of visualizations generated per **topic** of i.d. genomic epidemiology”*

1

Text mining analysis of document corpus

*“Identify and enumerate the kinds of visualizations generated per topic of i.d. genomic epidemiology”*

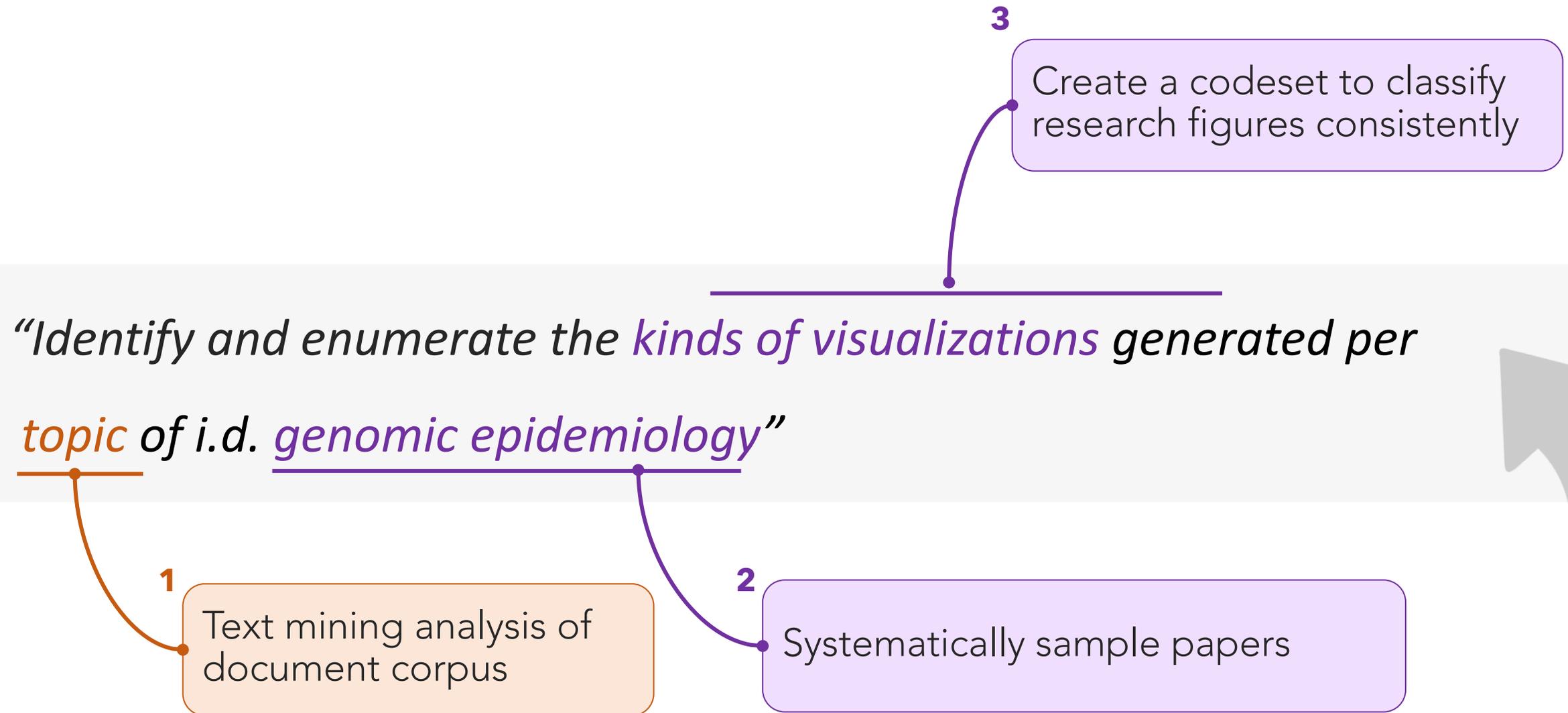
1

Text mining analysis of document corpus

2

Systematically sample papers

# An overview of our approach



# An overview of our approach

4

Apply code set to research figures

3

Create a codeset to classify research figures consistently

*“Identify and enumerate the kinds of visualizations generated per topic of i.d. genomic epidemiology”*

1

Text mining analysis of document corpus

2

Systematically sample papers

# An overview of our approach

