Data visualization strategies and tools for microbial genomic epidemiology

Anamaria Crisan
Vanier Canada Scholar & UBC Public Scholar
PhD Candidate, Computer Science
University of British Columbia

@amcrisan  acrisan@cs.ubc.ca  http://cs.ubc.ca/~acrisan
What we’ll talk about
Part I: Data Visualization Strategies & Tools

Part II: A brief (5 min) activity

Part III: Data Visualization Research in Practice
Part I: Data Visualization Strategies & Tools

Part II: A brief (5 min) activity

Part III: Data Visualization Research in Practice
Part I
Data visualization strategies and tools for microbial genomic epidemiology

Anamaria Crisan
Vanier Canada Scholar & UBC Public Scholar
PhD Candidate, Computer Science
University of British Columbia
PhD Candidate, Computer Science
University of British Columbia

Master of Science (Bioinformatics)

2008
GenomeDX Biosciences

2010

2013
British Columbia Centre for Disease Control

2015
PhD (Computer Science)
What we’ll talk about
Why should we visualize data?

How should we visualize data?

What datavis tools are available?
Why should we visualize data?
It is not always easy to reason consistently with numbers
Data Visualization is a Powerful Medium

Role of data visualization in the current paradigm of scientific research

= Communication
Do you have a research problem?

Yes.

Do all the Science!

But eventually you’ll have a problem, right?

No.

Inform the public!

Duh.

https://www.ratbotcomics.com/comics/pgrc_2014/1/1.html
Do you have a research problem?

Yes.

Do all the science!

Duh.

But eventually you’ll have a problem right?

No.

Inform the public!

Infographics are pretty

Maybe data visualization?
Yes.

No.

Do all the **Science!**

**Duh.**

**Inform**

the public!

**Did it work?**

**Infographics**

are pretty

**Problem?**

**Yes.**

**No.**

But eventually you’ll have a problem right?

**Maybe data Visualization?**
Yes.

Do all the Science!

Inform the public!

Different Infographics?

Did it work?

No :

Maybe data Visualization?

Problem?

Do you have a research Problem?

No.

But eventually you’ll have a problem right?
Did it work?

Yes! (maybe?)

No : ( )

Maybe data visualization?

Different infographics?

Inform the public!

Duh.

Do you have a research problem?

Yes.

No.

But eventually you’ll have a problem right?
Limitation #1: Missed Opportunity in Exploration

Missed Opportunity for Exploration

- Exploration is looking at your data, trying different analysis methods, assessing if there are outliers or missing data etc.
Limitation #1: Missed Opportunity in Exploration

Same stats, different graphs

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X Mean</td>
<td>54.2659224</td>
</tr>
<tr>
<td>Y Mean</td>
<td>47.8313999</td>
</tr>
<tr>
<td>X SD</td>
<td>16.7649829</td>
</tr>
<tr>
<td>Y SD</td>
<td>26.9342120</td>
</tr>
<tr>
<td>Corr.</td>
<td>-0.0642526</td>
</tr>
</tbody>
</table>

Limitation #1: Missed Opportunity in Exploration

Same stats, different graphs (Datasaurus)

Limitation #1: Missed Opportunity in Exploration

Opening up the machine learning black box
Limitation #1: Missed Opportunity in Exploration

Chihuahua or muffin?

Mop or sheep dog?
Limitation #1: Missed Opportunity in Exploration

Example: Trying to understand the black box

For instance, by combining feature visualization (what is a neuron looking for?) with attribution (how does it affect the output?), we can explore how the network decides between labels like Labrador retriever and tiger cat.

Several floppy ear detectors seem to be important when distinguishing dogs, whereas pointy ears are used to classify “tiger cat”.

Made with: JavaScript
Health data are complex to analyze and visualization
Limitations #2: Identifying the Appropriate Vis

Selecting the appropriate data visualization is challenging

- True for exploration & communication applications
Visualization Design ALSO matters

Design is a funny word. Some people think design means how it looks. But of course, if you dig deeper, it’s really how it works.

- Steve Jobs
Example: Communicating Survival Benefit of Cancer Therapy

Baseline Visualization

Alternative 1

Alternative 2

Zikmund-Fisher (2013). A demonstration of "less can be more" in risk graphics.
Example: Visualizing Arteries of the Heart for Surgery Planning

Made with: Processing
Example: Visualizing Arteries of the Heart for Surgery Planning

EXISTING STANDARD
Accuracy: 39%

REVISED VISUALIZATION
Accuracy: 91%

Borkin (2011). “Evaluation of Artery Visualizations for Heart Disease Diagnosis”

Made with: Processing
There are two aspects of visualizations to think about:

- Is it the appropriate visualization? How should we visualize data?
- How do you make a visualization? What datavis tools are available?
How should we visualize data?
Cross Cutting Disciplines in Information Visualization

- Human Perception & Cognition
- Computer Graphics
- Data Analysis
- Visualization Design & Analysis
Encoding and Decoding Information

Encoding
Data Transformation, Rendering

Data

Data Mapping

Visual Mapping

View

Decoding
Perception, Cognition

Knowledge

Cognition

Perception

R. Kosara (EagerEyes) – https://eagereyes.org/basics/encoding-vs-decoding
Putting it all Together for Visualization Design & Analysis

- Non-trivial to condense knowledge across all these areas
- Still an ongoing area of research
- I will try convey a simpler intuition about design & analysis
Guiding Principles for Visualizing your Data
Breaking Down a Visualization in Three Questions

Why? (Motivation)
Why do you need to visualize data?
How will you, or others, use the visualization?
Breaking Down a Visualization in Three Questions

Why? (Motivation)
Why do you need to visualize data?
How will you, or others, use the visualization?

What? (Data & Tasks)
What kind of data is being visualized?
What tasks are performed with the data?
Breaking Down a Visualization in Three Questions

**Why?** *(Motivation)*

Why do you need to visualize data?
How will you, or others, use the visualization?

**What?** *(Data & Tasks)*

What kind of data is being visualized?
What tasks are performed with the data?

**How?** *(Visual & Interactive Design)*

How do you make the visualization?
Is it the right visualization?

People tend to jump to this level and ignore why and what.
Design & Evaluation with Three Questions

**Design**

**Why?**
Does the visualization address the intended need?

**What?**
Are you using the right data, or deriving the right data?
Does the visualization support the tasks using that data?

**How?**
Are the visual & interactive choices appropriate for the data and tasks?
If interactive / computer based, is the visualization easy to use and reliable (i.e. doesn’t crash all the time)

**Evaluation**

**Why?**

**What?**

**How?**
Ideas from the research literature: the nested-model

**Why?**
- Domain situation: Observe target users using existing tools

**What?**
- Data/task abstraction

**How?**
- Visual encoding/interaction idiom: Justify design with respect to alternatives
- Algorithm: Measure system time/memory, Analyze computational complexity
- Analyze results qualitatively, Measure human time with lab experiment (lab study), Observe target users after deployment (field study)
- Measure adoption

**Design**

**Evaluation**

T. Munzner (2014) – Visualization Design and Analysis
Steps to Systematic Thinking in Data Visualization

Image Source: Valentin Antonucci via Pexels
Infovis (Information Visualization) research advocates an iterative process.

Design

- Domain Problem*
- Data + Task
- Visual + Interaction Design Choices
- Algorithm

Evaluation

*Domain Problem = Motivation

T. Munzner (2014) – Visualization Design and Analysis
An iterative approach to development allows us to get feedback before committing to ineffective design choices.
### Thinking Systematically about Data Visualization

<table>
<thead>
<tr>
<th>Domain Problem</th>
<th>Data + Task</th>
<th>Visual + Interaction</th>
<th>Design Choices</th>
<th>Algorithm</th>
</tr>
</thead>
</table>

1. Identify a relevant **problem** that effects you or a group of stakeholders

---

T. Munzner (2014) – Visualization Design and Analysis
Public Health Stakeholders

- Multidisciplinary decision making teams
  - More data & diverse data types = more informed decision making
  - BUT – different stakeholder abilities to interpret data & different needs
T. Munzner (2014) – Visualization Design and Analysis

2. Ask **what data** stakeholders use (is it available)?
3. Ask **what stakeholders do** with the data [tasks]
Data - Many Different Types of Data!

**Dataset Types**

- **Tables**
  - Items (rows)
  - Attributes (columns)
  - Cell containing value

- **Networks**
  - Nodes (items)
  - Links

- **Fields (Continuous)**
  - Grid of positions
  - Cell
  - Attributes (columns)
  - Value in cell

- **Geometry (Spatial)**
  - Positions

- **Multidimensional Table**
  - Attributes
  - Value in cell

- **Trees**

T. Munzner (2014) – Visualization Design and Analysis
Data - Don’t Just Visualize the Raw Data!

Example

Original (Raw) Data

Derived Data

Example when this advice is ignored

T. Munzner (2014) – Visualization Design and Analysis

XKCD
Tasks - How People Use the Data

Geographic Overview of Prostate Cancer
- Useful for epidemiologists and policy makers
- Supports surveillance tasks

In Washington in 2015, age-adjusted rate of new Prostate Cancer cases was 91.7 per 100,000 men (95% Confidence Interval: 88.7 - 94.8).
3,814 Prostate Cancer cases were reported.

Individual Prostate Cancer Risk
- Good for patients and doctors
- Supports treatment decision making tasks

Risk of prostate cancer if biopsy were to be performed
Based on the provided risk factors a prostate biopsy performed would have a:
- 4% chance of high-grade prostate cancer,
- 19% chance of low-grade cancer,
- 73% chance that the biopsy is negative for cancer.

About 2 to 4% of men undergoing biopsy will have an infection that may require hospitalization.

Please consult your physician concerning these results.

Source: Atlanta CDC
Source: http://riskcalc.org/PCPTRC/ (UT San Antonio)
Tasks - How People Use the Data

• Tasks can also change how the same data should be visualized
• Example: representing US electoral collage results

Standard Map

Cartogram
Tasks - How People Use the Data

• Tasks can also change how the same data should be visualized
• Example: representing US electoral collage results

Standard Map

Snakey Diagram
Tasks - How People Use the Data

• Tasks can also change how the same data should be visualized
• Example: representing US electoral collage results
How can we identify tasks and data?
My research: making a clinical report for tuberculosis

• Mixed methods approach to gathering data and tasks

Data Gathered

<table>
<thead>
<tr>
<th></th>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Design</td>
<td></td>
<td>Exploratory Sequential Model</td>
</tr>
</tbody>
</table>

Discovery

Information Gathering

- Expert Consults
- TB Workflow Map
- Task & Data Questionnaire

Design

Design & Evaluation

- Design Sprint
- Design Choice Questionnaire

Implement

Finalize Design

Organism

The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

Drug Susceptibility

Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

- □ No drug resistance predicted
- □ Mono-resistance predicted
- ✓ □ Mul-drug resistance predicted
- □ Extensive drug resistance predicted

Drug class | Interpretation | Drug Resistance Gene (Amino Acid Mutation)
--- | --- | ---
Ethambutol | No mutation detected | Susceptible
Pyrazinamide | No mutation detected | Isoniazid katG (S315T) First Line Resistant
Rifampin | rpoB (S531L) | Streptomycin No mutation detected
Ciprofloxacin | No mutation detected | Ofloxacin No mutation detected
Moxifloxacin | No mutation detected | Amikacin No mutation detected
Kanamycin | No mutation detected | Kanamycin耐药

Second Line Susceptible

- Capreomycin No mutation detected

General Information

- Patient Name: JOHN DOE
- Barcode: 12345678910
- Birth Date: 2000-01-01
- Patient ID: 12345678910
- Location: SOMEPLACE
- Sample Type: SPUTUM
- Sample Source: PULMONARY
- Sample Date: 2016-12-25
- Sample ID: A12345678
- Sequenced From: MGIT CULTURED ISOLATE
- Reporting Lab: LAB NAME
- Report Date/Time: 2017-01-01, 15:36
- Requested By: REQUESTER NAME
- Requester Contact: REQUESTER@EMAIL.COM

Summary

The specimen was positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.
<table>
<thead>
<tr>
<th>WGS equivalent</th>
<th>DIAGNOSIS TASKS</th>
<th>TREATMENT TASKS</th>
<th>SURVEILLANCE TASKS</th>
<th>TOTAL SCORE</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Identifier</td>
<td>Same</td>
<td>3 3 3 3 3</td>
<td>3 3 3</td>
<td>2 1 1 1 1 1</td>
<td>26</td>
</tr>
<tr>
<td>Sample Collection Date</td>
<td>Same</td>
<td>3 3 3 3 3</td>
<td>3 3 3</td>
<td>1 1 1 1 1 1</td>
<td>24</td>
</tr>
<tr>
<td>Sample Prior TB Results</td>
<td>Same</td>
<td>3 2 3 3 3</td>
<td>3 3 3</td>
<td>1 1 1 0 1</td>
<td>23</td>
</tr>
<tr>
<td>Speciation</td>
<td>Speciation</td>
<td>1 3 2 3 3</td>
<td>3 3 3</td>
<td>2 1 1 1 1</td>
<td>23</td>
</tr>
<tr>
<td>Sample Type (sputum, fine needle aspirate etc.)</td>
<td>Same</td>
<td>2 3 2 3 3</td>
<td>3 3 3</td>
<td>1 1 1 0 1</td>
<td>22</td>
</tr>
<tr>
<td>Culture results</td>
<td>NA</td>
<td>1 3 2 3</td>
<td>3 3 3</td>
<td>2 1 1 0 1</td>
<td>22</td>
</tr>
<tr>
<td>Sample Collection Site (lymph node, lung etc.)</td>
<td>Same</td>
<td>2 3 2 3 3</td>
<td>3 3 3</td>
<td>1 1 0 0 1</td>
<td>21</td>
</tr>
<tr>
<td>Acid Fast Bacilli Smear</td>
<td>Speciation</td>
<td>2 3 2 3 3</td>
<td>2 3 3</td>
<td>1 1 1 0 1</td>
<td>21</td>
</tr>
<tr>
<td>Resistotype</td>
<td>Predicted DST</td>
<td>0 2 3 1</td>
<td>3 3 2</td>
<td>2 1 1 1 1</td>
<td>19</td>
</tr>
<tr>
<td>Phenotypic DST</td>
<td>Predicted DST</td>
<td>0 2 3 2 2</td>
<td>3 3 2</td>
<td>1 1 1 0 1</td>
<td>18</td>
</tr>
<tr>
<td>Chest x-ray</td>
<td>NA</td>
<td>3 3 2 3</td>
<td>0 2 3</td>
<td>1 0 0 0 0</td>
<td>17</td>
</tr>
<tr>
<td>Report Release Date</td>
<td>Same</td>
<td>2 2 1 2</td>
<td>2 2 2</td>
<td>1 0 0 0 0</td>
<td>15</td>
</tr>
<tr>
<td>Requester IDs</td>
<td>Same</td>
<td>2 2 2 2</td>
<td>2 2 2</td>
<td>1 0 0 0 0</td>
<td>15</td>
</tr>
<tr>
<td>Interpretation or comments from reviewer</td>
<td>Same</td>
<td>2 2 1 2</td>
<td>2 2 3</td>
<td>1 0 0 0 0</td>
<td>15</td>
</tr>
<tr>
<td>Predicted DST</td>
<td>Predicted DST</td>
<td>0 2 2 1</td>
<td>3 3 2</td>
<td>1 0 1 0 0</td>
<td>15</td>
</tr>
<tr>
<td>MIRU-VNTR</td>
<td>SNPs</td>
<td>0 2 3 1</td>
<td>1 1 1</td>
<td>1 1 1 1 1</td>
<td>13</td>
</tr>
<tr>
<td>Cluster Assignment</td>
<td>Same</td>
<td>0 2 2 1</td>
<td>1 1 0</td>
<td>1 1 1 1 1</td>
<td>11</td>
</tr>
<tr>
<td>SNP Variant distance</td>
<td>SNPs</td>
<td>0 2 2 1</td>
<td>1 1 0</td>
<td>1 1 1 1 1</td>
<td>10</td>
</tr>
<tr>
<td>Phylogenetic Tree</td>
<td>Same</td>
<td>0 2 1 1</td>
<td>1 1 0</td>
<td>1 0 1 1 1</td>
<td>9</td>
</tr>
<tr>
<td>Reviewer ID</td>
<td>Same</td>
<td>1 1 1 1</td>
<td>1 1 1</td>
<td>1 0 0 0 0</td>
<td>8</td>
</tr>
<tr>
<td>TST results</td>
<td>Speciation</td>
<td>3 1 1 1</td>
<td>0 0 0</td>
<td>1 0 0 0 0</td>
<td>7</td>
</tr>
<tr>
<td>IGRA results</td>
<td>Speciation</td>
<td>3 1 1 1</td>
<td>0 0 0</td>
<td>1 0 0 0 0</td>
<td>7</td>
</tr>
<tr>
<td>Lab QC</td>
<td>WGS Specific</td>
<td>0 1 2 1</td>
<td>1 1 0</td>
<td>1 0 0 0 0</td>
<td>7</td>
</tr>
<tr>
<td>Spoligotype</td>
<td>SNPs</td>
<td>0 1 1 1</td>
<td>0 0 0</td>
<td>1 0 0 0 0</td>
<td>3</td>
</tr>
<tr>
<td>RFLP</td>
<td>SNPs</td>
<td>0 1 1 1</td>
<td>0 0 0</td>
<td>1 0 0 0 0</td>
<td>3</td>
</tr>
</tbody>
</table>
**Summary**

The specimen was positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

**Organism**

The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

**Drug Susceptibility**

Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

<table>
<thead>
<tr>
<th>Drug class</th>
<th>Interpretation</th>
<th>Drug</th>
<th>Resistance Gene (Amino Acid Mutation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Line</td>
<td>Susceptible</td>
<td>Ethambutol</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pyrazinamide</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Resistant</td>
<td>Isoniazid</td>
<td>katG (S315T)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rifampin</td>
<td>rpoB (S531L)</td>
</tr>
<tr>
<td>Second Line</td>
<td>Susceptible</td>
<td>Streptomycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ciprofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moxifloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Amikacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kanamycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Capreomycin</td>
<td>No mutation detected</td>
</tr>
</tbody>
</table>

□ No drug resistance predicted
□ Mono-resistance predicted
☑ Multi-drug resistance predicted
□ Extensive drug resistance predicted
4. Explore if other visualizations have addressed this **problem** and set of **tasks & data**

5. Implement **your own solution** *(remember this include interaction!)*

---

T. Munzner (2014) – Visualization Design and Analysis
Marks & Channels: Basic Building Blocks

Mark:
Basic Graphical Element
(basic building block)

Channel:
Controls the appearance of marks

T. Munzner (2014) – Visualization Design and Analysis
Marks Vary in their Effectiveness

Example

**Pie Chart**
- Angle & Area

**Bar Chart**
- Position
- Common Scale

Perception 🎨 and Cognition 🧠 Matter Too!

Original Visualization

Visualization as seen by color blind person
(color blindness (deuteranopia) impacts men more often)

Perception 🕍 and Cognition 🧠 Here too!

Colour scales also impact interpretation!
Perceptual research from Liu et al (2018)

Figure 1: Colormaps under study. We evaluate four single-hue, three perceptually-uniform multi-hue, a diverging, and a rainbow colormap(s). We divide them into (a) assorted, (b) single-hue and (c) multi-hue groups, with two colormaps repeated across groups for replication.

Figure 4: Error rate by colormap for each study. Plots depict bootstrapped means, with 50% (thick) and 95% (thin) CIs. (a) Assorted colormaps. Viridis excels in accuracy while jet is the most error-prone. (b) Single-hue colormaps. Though slightly faster, blues and greens have overlapping confidence intervals with the slower colormaps, oranges and greys. (c) Multi-hue colormaps. Multi-hue colormaps have comparable accuracy within group. The per-colormap average error rate of magma is higher as it contains degenerate cases.

Liu et al. (2018) - Somewhere Over the Rainbow: An Empirical Assessment of Quantitative Colormaps
Marks & Channels: ggplot2 example

```r
ggplot (data = mpg, aes(x = display, y = cty, colour = class)) + geom_point()
```

Channel: Position  
Channel: Colour

Mark: Point

**Note:** Generally in ggplot2 aesthetics refer to channels and geoms refer to marks, but there are complex geoms that aren’t simple marks but chart types (i.e. geom_density) and there are aesthetics that have little to do with the visual channels directly (i.e. group)

https://rpubs.com/hadley/ggplot-intro
Marks & Channels: Tableau example

Marks:
- Circle

Channels:
- Continent
- Country

Sheet 7
- Continent
- Gdp Per Cap
Linking Data to Mark and Channels to Make Visualizations

Data → Marks & Channels → Visualization
Linking Data to Mark and Channels to Make Visualizations

Data to viz
https://www.data-to-viz.com/

Chart Chooser
Examples from my own research

How do people visualize data?
My research: surveying visualizations in genomic epidemiology

http://gevit.net


OXFORD BIOINFORMATICS
Examples from my own research

How can we help people visualize data?
My research: simplifying the creation of data visualizations

#specify individual charts

phyloTree_chart<-specify_base(chart_type = "phylogenetic tree",data="tree_dat")
epicurve<-specify_base(chart_type = "histogram",data="tab_dat",x = "month")
map_chart<-specify_base("geographic map",data="tab_dat",lat = "latitude",long = "longitude")

#specify a combination

colour_combo<-specify_combination(combo_type = "color_linked", base_charts = c("phyloTree_chart","map_chart","epicurve"),link_by="country")

#plot the result

plot(colour_combo)
# Analyze different 
data types automatically

```
harmon_obj <- data_harmonization(tab_dat, tree_dat, genomic_dat, all_spatial)
```

# Create specifications 
that compile to minCombinr
```
component_specs <- get_spec_list(harmon_obj)
```

# plot the result one view at a time
```
plot_view(component_specs, view_num=1)
```
4. Explore if other visualizations have addressed this problem and set of tasks
5. Implement your own solution (part or all of that solution could be a new algorithm)
6. Test **multiple alternatives** (including new ones you develop) with stakeholders

7. Gather **qualitative & quantitative** evaluation data
1. Identify a relevant problem that effects you or a group of stakeholders
2. Ask what data stakeholders use (is it available)?
3. Ask what stakeholders do with the data [tasks]
4. Explore if other visualizations have addressed this problem and set of tasks & data
5. Implement your own solution (vis and/or algorithm)
6. Test multiple alternatives (including new ones you develop) with stakeholders
7. Gather qualitative & quantitative evaluation data
What data visualization tools are available?
Data Visualization Tools to Get You Started
Tools & Libraries for data visualization

Lisa Charlotte Rost has an excellent blog post about this: [http://bit.ly/2gRGx1J](http://bit.ly/2gRGx1J)
I am presenting her figures here
Tools & Libraries for data visualization

Lisa Charlotte Rost has an excellent blog post about this: http://bit.ly/2gRGx1J

Analysis vs Presentation
Tools & Libraries for data visualization

Lisa Charlotte Rost has an excellent blog post about this: http://bit.ly/2gRGx1J

Extent of Flexibility

How easy/hard it is to make data visualizations (including custom/novel visualizations)
Tools & Libraries for data visualization

Lisa Charlotte Rost has an excellent blog post about this: [http://bit.ly/2gRGx1J](http://bit.ly/2gRGx1J)

### Static vs Interactive

<table>
<thead>
<tr>
<th>Static</th>
<th>Web - Interactive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Apps</strong></td>
<td><strong>ILLUSTRATOR, NODEBOX, EXCEL, POLESTAR, RAW</strong></td>
</tr>
<tr>
<td><strong>Charting Libraries</strong></td>
<td><strong>GGPLOT2, MATPLOTLIB, R, SEABORN, BOKEH, PROCESSING</strong></td>
</tr>
</tbody>
</table>
Tools & Libraries for data visualization

Lisa Charlotte Rost has an excellent blog post about this: http://bit.ly/2gRGx1J

“There are no perfect tools, just good tools for people with certain goals”

See a detailed table here: http://bit.ly/2DeWPwV
Tools & Libraries for data visualization

Another take with commonly used tools: https://bit.ly/2SgrOzS
Don’t forget that pen and paper is an option too!

Dear Data Project
(Lupi & Posavec)
Datavis tools for (Microbial) Genomics
IGV Browser for all your genomic needs

https://software.broadinstitute.org/software/igv/
The classic UCSC genome browser

https://genome.ucsc.edu
GenVisR: Human Genomes in R

https://academic.oup.com/bioinformatics/article/32/19/3012/2196360
Variant Viewer: Human Genomes

http://www.cs.ubc.ca/labs/imager/tr/2013/VariantView/
Microreact: Microbial Genomics

https://microreact.org/
GenGIS: Microbial Genomics (Made in Canada!)

http://kiwi.cs.dal.ca/GenGIS/Main_Page
Nextstrain: Microbial Genomics

https://nextstrain.org/ebola
Wrapping up
DATA VISUALIZATION IS NOT JUST AN ART PROJECT
Key take-aways from this talk

- **Visualizations of data are useful**
  - Helpful in instance of low numeracy
  - Can used in communication and exploration

- **But.. visualization design also matters**
  - Many different alternatives, important to test

- **It’s possible to think systematically about visualizations**
  - Many disciplines cross cut information visualization research
  - At the minimum think “Why”, “What”, “How”

- **Encode data well so that others can decode it later**

- **Data visualization is a research process with open and interesting problems**
Additional Resources

- **Books to consider:**
  - Visualization Design and Analysis by Tamara Munzner (more technical)

- **Online resources:**
  - Distill Publication: https://distill.pub/
  - UW Interactive Data Lab: https://medium.com/@uwdata
  - Data stories podcast: http://datastori.es/

- **Inspiration:**
  - Information is Beautiful: https://informationisbeautiful.net/
  - Visualization WTF (examples of what not to do): http://viz.wtf/
Data visualization strategies and tools for microbial genomic epidemiology

Anamaria Crisan
Vanier Canada Scholar & UBC Public Scholar
PhD Candidate, Computer Science
University of British Columbia
Part I: Data Visualization Strategies & Tools

Part II: A brief (5 min) activity

Part III: Data Visualization Research in Practice
How many ways can we visualize these numbers?

• In your head, on paper, or computer, sketch out as many examples as you can to visualize the following to numbers:

75  37
How many ways can we visualize these numbers?

- In your head, on paper, or computer, sketch out as many examples as you can to visualize the following numbers:

  75  37

*example:*
How many ways can we visualize these numbers?

- In your head, on paper, or computer, sketch out as many examples as you can to visualize the following to numbers:

  75  37

some solutions:
Part I: Data Visualization Strategies & Tools

Part II: A brief (5 min) activity

Part III: Data Visualization Research in Practice
Part III
Data visualization strategies and tools for microbial genomic epidemiology

Anamaria Crisan
Vanier Canada Scholar & UBC Public Scholar
PhD Candidate, Computer Science
University of British Columbia

@amcrisan  acrisan@cs.ubc.ca  http://cs.ubc.ca/~acrisan
The characteristics of data

Volume
Amount of data
The characteristics of data

Volume
- Amount of data

Variety
- Kinds of data
### The characteristics of data

<table>
<thead>
<tr>
<th>Volume</th>
<th>Variety</th>
<th>Veracity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of data</td>
<td>Kinds of data</td>
<td>Reliability of data</td>
</tr>
</tbody>
</table>

![Volume Chart](chart1.png)

![Variety Chart](chart2.png)

![Veracity Chart](chart3.png)
The characteristics of data

**Volume**
Amount of data

**Variety**
Kinds of data

**Veracity**
Reliability of data

**Velocity**
Speed of acquisition
How do we bridge the gap from data to insights and actions?
How do we bridge the gap from data to insights and actions?

Data Science & Data Visualization

- Explore
- Transform
- Visualize
- Model

Insights

Action
Doctoral research: visualizing complex & heterogenous data

Data Science & Data Visualization

- Transform
- Explore
- Model
- Visualize

Insights

Action


**Understand**: what data do stakeholders need to perform their tasks?

**The challenge**: limited stakeholder time, complex & restricted data access

**My strategy**: partner with stakeholders on a high-value project, gather necessary evidence for data & tasks

**Collaboration context**: evidence-based redesign of a clinical report with Public Health England

---

### MYCOBACTERIUM TUBERCULOSIS GENOME SEQUENCING REPORT

<table>
<thead>
<tr>
<th>Patient Name</th>
<th>JOHN DOE</th>
<th>Barcode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Date</td>
<td>2000-01-01</td>
<td>Patient ID 12345678910</td>
</tr>
<tr>
<td>Location</td>
<td>SOMEPLACE</td>
<td>Sample Type: SPUTUM</td>
</tr>
<tr>
<td>Sample Source</td>
<td>PULMONARY</td>
<td>Sample Date: 2016-12-25</td>
</tr>
<tr>
<td>Sample ID</td>
<td>A12345678</td>
<td>Sequenced From: MGT CULTURED ISOLATE</td>
</tr>
<tr>
<td>Reporting Lab</td>
<td>LAB NAME</td>
<td>Report Date/Time: 2017-01-01, 15:36</td>
</tr>
<tr>
<td>Requested By</td>
<td>REQUESTER NAME</td>
<td>Requester Contact: <a href="mailto:REQUESTER@EMAIL.COM">REQUESTER@EMAIL.COM</a></td>
</tr>
</tbody>
</table>

### Summary

The specimen was positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

### Organism

The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

### Drug Susceptibility

Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

<table>
<thead>
<tr>
<th>Drug class</th>
<th>Susceptible</th>
<th>Resistance Gene (Variant Allele)/Drug Resistance Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Line</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Susceptible</td>
<td>Ethambutol</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Pyrazinamide</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Resistant</td>
<td>Isoniazid</td>
<td>katG (S315T)</td>
</tr>
<tr>
<td></td>
<td>Rifampin</td>
<td>rpoB (S531L)</td>
</tr>
<tr>
<td><strong>Second Line</strong></td>
<td>Susceptible</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Streptomycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Ciprofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Ofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Moxifloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Amikacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Kanamycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Capreomycin</td>
<td>No mutation detected</td>
</tr>
</tbody>
</table>

---

**Crisan A**, McKee G, Munzner T, Gardy JL
“Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory”

**PEERJ** (2018)

**Crisan A**, Gardy JL, Munzner T
“On regulatory and organizational constraints in visualization design and evaluation”

**BELIV’16** – an IEEE VIS affiliated methods workshop
Understand: gathering evidence through mixed-methods

- Integrating mixed-methods (MM) research into design study methodologies

**Discover**
*Information Gathering*
- Expert Consults
- TB Workflow Map
- Task & Data Questionnaire

**Design**
*Design & Evaluation*
- Design Sprint
- Design Choice Questionnaire

**Deploy**
*Finalize Design*

---

**Data Gathered**
- Qualitative
- Quantitative

**MM Study Design**
- Exploratory Sequential Model
- Embedded Model

---

**MYCOBACTERIUM TUBERCULOSIS GENOME SEQUENCING REPORT**

- **Patient Name**: JOHN DOE
- **Barcode**: 
- **Birth Date**: 2000-01-01
- **Patient ID**: 12345678910
- **Location**: SOMEPLACE
- **Sample Type**: SPUTUM
- **Sample Source**: PULMONARY
- **Sample Date**: 2016-12-25
- **Sample ID**: A12345678
- **Sequenced From**: MGIT CULTURED ISOLATE
- **Reporting Lab**: LAB NAME
- **Report Date/Time**: 2017-01-01, 15:36
- **Requested By**: REQUESTER NAME
- **Requester Contact**: REQUESTER@EMAIL.COM

**Summary**
The specimen was positive for **Mycobacterium tuberculosis**. It is resistant to isoniaizd and rifampin. It belongs to a cluster, suggesting recent transmission.

**Organism**
The specimen was positive for **Mycobacterium tuberculosis**, lineage 2.2.1 (**East-Asian Beijing**).

**Drug Susceptibility**
Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

□ No drug resistance predicted
□ Mono-resistance predicted
✓□ Mul-drug resistance predicted
□ Extensive drug resistance predicted

**Drug class**
**Interpretation**
**Drug Resistance Gene** (Amino Acid Mutation)
- Ethambutol: No mutation detected
- Pyrazinamide: No mutation detected
- Isoniazid: katG (S315T) first line resistant
- Rifampin: rpoB (S531L)
- Streptomycin: No mutation detected
- Ciprofloxacin: No mutation detected
- Ofloxacin: No mutation detected
- Moxifloxacin: No mutation detected
- Amikacin: No mutation detected
- Kanamycin: No mutation detected

- Second Line Susceptible
- Capreomycin: No mutation detected
Understand: gathering evidence through mixed-methods

- Integrating mixed-methods (MM) research into design study methodologies

Discover

*Information Gathering*

Data Gathered

<table>
<thead>
<tr>
<th>Data Gathered</th>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert Consults</td>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>TB Workflow Map</td>
<td></td>
<td>☐</td>
</tr>
<tr>
<td>Task &amp; Data Questionnaire</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MM Study Design

Exploratory Sequential Model
**Discover: what data is used for different tasks**

### Expert Consults Participants

<table>
<thead>
<tr>
<th>Public Health Role</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician</td>
<td>2</td>
</tr>
<tr>
<td>Nurse</td>
<td>1</td>
</tr>
<tr>
<td>Laboratorian</td>
<td>2</td>
</tr>
<tr>
<td>Researcher</td>
<td>0</td>
</tr>
<tr>
<td>Surveillance</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7</strong></td>
</tr>
</tbody>
</table>

Semi-structure interviews
Qualitative Data

### Online Task & Data Survey Participants

<table>
<thead>
<tr>
<th>Public Health Role</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician</td>
<td>7</td>
</tr>
<tr>
<td>Nurse</td>
<td>3</td>
</tr>
<tr>
<td>Laboratorian</td>
<td>3</td>
</tr>
<tr>
<td>Researcher</td>
<td>1</td>
</tr>
<tr>
<td>Surveillance</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>17</strong></td>
</tr>
</tbody>
</table>

Multiple choice questionnaire
Quantitative data
**Discover** : quantified consensus for data used for some tasks

<table>
<thead>
<tr>
<th>Data</th>
<th>WGS equivalent</th>
<th>DIAGNOSIS TASKS</th>
<th>TREATMENT TASKS</th>
<th>SURVEILLANCE TASKS</th>
<th>TOTAL SCORE</th>
<th>cat.</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Identifier</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sample Collection Date</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Patient Prior TB Results</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Speciation</td>
<td>Speciation</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sample Type (sputum, fine needle aspirate etc.)</td>
<td>Same</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Culture results</td>
<td>NA</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sample Collection Site (lymph node, lung etc.)</td>
<td>Same</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Acid Fast Bacilli Smear</td>
<td>Speciation</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Resistotype</td>
<td>Predicted DST</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Phenotypic DST</td>
<td>Predicted DST</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Chest x-ray</td>
<td>NA</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Report Release Date</td>
<td>Same</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Requester IDs</td>
<td>Same</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Interpretation or comments from reviewer</td>
<td>Same</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Predicted DST</td>
<td>Predicted DST</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MIRU-VNTR</td>
<td>SNPs</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cluster Assignment</td>
<td>Same</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SNP/variant distance</td>
<td>SNPs</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Phylogenetic Tree</td>
<td>Same</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reviewer ID</td>
<td>Same</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TST results</td>
<td>Speciation*</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IGRA results</td>
<td>Speciation*</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lab QC</td>
<td>WGS Specific</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Spoligotype</td>
<td>SNPs</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RFLP</td>
<td>SNPs</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Consensus among participants:
- 3 (>75%)
- 2 (50% - 25%)
- 1 (25% - 50%)
- 0 (<25%)
**Discover**: for some tasks, stakeholders don’t know what data to use

- A surprising finding: limited consensus for data used in surveillance tasks

<table>
<thead>
<tr>
<th>Data</th>
<th>WGS equivalent</th>
<th>DIAGNOSIS TASKS</th>
<th>TREATMENT TASKS</th>
<th>SURVEILLANCE TASKS</th>
<th>TOTAL SCORE</th>
<th>Consensus among participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Diagnose Latent TB</td>
<td>Diagnose Active TB</td>
<td>Reactive vs New Infection</td>
<td>Choose Meds</td>
<td>Choose Tx</td>
</tr>
<tr>
<td>Patient Identifier</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Sample Collection Date</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Patient Prior TB Results</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Speciation</td>
<td>Speciation</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Sample Type (sputum, fine needle aspirate etc.)</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Culture results</td>
<td>NA</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Sample Collection Site (lymph node, lung etc.)</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Acid Fast Bacilli Smear</td>
<td>Speciation</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Resistotype</td>
<td>Predicted DST</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Phenotypic DST</td>
<td>Predicted DST</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Chest x-ray</td>
<td>NA</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Report Release Date</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Requester IDs</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Interpretation or comments from reviewer</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Predicted DST</td>
<td>Predicted DST</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MIRU-VNTR</td>
<td>SNPs</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Cluster Assignment</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>SNP/variant distance</td>
<td>SNPs</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Phylogenetic Tree</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Reviewer ID</td>
<td>Same</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>TST results</td>
<td>Speciation</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>IGRA results</td>
<td>Speciation</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Lab QC</td>
<td>WGS Specific</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Spoligotype</td>
<td>SNPs</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>RFLP</td>
<td>SNPs</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Degree of Consensus:
- 3 (90%) - very high
- 2 (70%) - high
- 1 (50%) - moderate
- 0 (<25%) - low

<table>
<thead>
<tr>
<th>Core Score</th>
<th>cat.</th>
<th>% agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>3</td>
<td>(&gt;75%)</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>(50% - 25%)</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>(25% - 50%)</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>(&lt;25%)</td>
</tr>
</tbody>
</table>
Understand: gathering evidence through mixed-methods

- Integrating mixed-methods (MM) research into design study methodologies

**Data Gathered**
- Qualitative
- Quantitative

**MM Study Design**
- Exploratory Sequential Model
- Embedded Model

**Discover**
- Information Gathering
  - Expert Consults
  - TB Workflow Map
  - Task & Data Questionnaire

**Design**
- Design & Evaluation
  - Design Sprint
  - Design Choice Questionnaire

**Organism**
The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

**Drug Susceptibility**
Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Interpretation</th>
<th>Drug Resistance Gene (Amino Acid Mutation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethambutol</td>
<td>Susceptible</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td></td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>First Line Resistant</td>
<td>katG (S315T)</td>
</tr>
<tr>
<td>Rifampin</td>
<td></td>
<td>rpoB (S531L)</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>Susceptible</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>Susceptible</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Ofloxacin</td>
<td>Susceptible</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>Susceptible</td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Amikacin</td>
<td></td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Kanamycin</td>
<td></td>
<td>No mutation detected</td>
</tr>
<tr>
<td>Capreomycin</td>
<td>Second Line Susceptible</td>
<td>No mutation detected</td>
</tr>
</tbody>
</table>
**Design** : creating and testing alternative report designs

**Design Sprint**
- Creative session
- Making prototypes

**Design Choice Questionnaire**
- Multiple Choice questionnaire
- Quantitative & Qualitative data

---

<table>
<thead>
<tr>
<th>Public Health Role</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician</td>
<td>13</td>
</tr>
<tr>
<td>Nurse</td>
<td>5</td>
</tr>
<tr>
<td>Laboratorian</td>
<td>3</td>
</tr>
<tr>
<td>Researcher</td>
<td>8</td>
</tr>
<tr>
<td>Surveillance</td>
<td>8</td>
</tr>
<tr>
<td>Other*</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>54</strong></td>
</tr>
</tbody>
</table>
Mycobacterium Whole Genome Sequencing Report from MGIT Positive Samples

Not for diagnostic use

01/02/1915

## Sample Details

<table>
<thead>
<tr>
<th>Location</th>
<th>Date received in Lab</th>
<th>Specimen ID</th>
<th>Run date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford</td>
<td>123456789</td>
<td>123456-79aab-910abr-15243hg</td>
<td></td>
</tr>
</tbody>
</table>

## Organism Identification

<table>
<thead>
<tr>
<th>Predicted/closest match</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBCOMP/microti</td>
<td></td>
</tr>
<tr>
<td>TBCOMP</td>
<td>100%</td>
</tr>
<tr>
<td>TBCOMP/TB</td>
<td>96.77%</td>
</tr>
<tr>
<td>TBCOMP/tuberculosis-canetti</td>
<td>35.71%</td>
</tr>
<tr>
<td>MACCOMP</td>
<td>21.21%</td>
</tr>
</tbody>
</table>

## Sample/Sequencing Quality

<table>
<thead>
<tr>
<th>Total reads (~millions)</th>
<th>Mapped %</th>
<th>No reads mapped (~millions)</th>
<th>Coverage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.73</td>
<td>99.47</td>
<td>4.7</td>
<td>91.99</td>
</tr>
</tbody>
</table>

## Resistance Summary

<table>
<thead>
<tr>
<th>INH</th>
<th>Rif</th>
<th>EMB</th>
<th>Pza</th>
<th>Qui</th>
<th>Sm</th>
<th>Ag</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

## Resistotype

<table>
<thead>
<tr>
<th>Drug</th>
<th>Mutation</th>
<th>Nucleotides</th>
<th>Support (ACGT)</th>
<th>Source – (R/Total)</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>INH</td>
<td>katG_A727T</td>
<td>GCC→ACC</td>
<td>(160/0/1/0) (0/164/0/0) (0/167/0/0)</td>
<td>Unclassified</td>
<td>UNK</td>
</tr>
</tbody>
</table>
Mycobacterium Whole Genome Sequencing Report from MGIT Positive Samples

Not for diagnostic use 01/02/1915

Sample Details

<table>
<thead>
<tr>
<th>Location</th>
<th>Date received in Lab</th>
<th>Specimen ID</th>
<th>Run date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford</td>
<td></td>
<td>123456789</td>
<td>01/01/19150115</td>
</tr>
<tr>
<td>Local Lims</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUID</td>
<td></td>
<td>123456-79aab-910abr-15243hg</td>
<td></td>
</tr>
</tbody>
</table>

Organism Identification

Predicted/closest match

- TBCOMP/microti: 100%
- TBCOMP: 100%
- TBCOMP/TB: 96.77%
- TBCOMP/tuberculosis: 35.71%
- MACCOMP: 21.21%

Sample/Sequencing Quality

<table>
<thead>
<tr>
<th>Total reads (~millions)</th>
<th>Mapped %</th>
<th>No reads mapped (~millions)</th>
<th>Coverage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.73</td>
<td>99.47</td>
<td>4.7</td>
<td>91.99</td>
</tr>
</tbody>
</table>

Resistance Summary

<table>
<thead>
<tr>
<th>INH</th>
<th>Rif</th>
<th>EMB</th>
<th>PZA</th>
<th>QUI</th>
<th>SM</th>
<th>AG</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Resistotype

<table>
<thead>
<tr>
<th>Drug</th>
<th>Mutation</th>
<th>Nucleotides</th>
<th>Support (ACGT)</th>
<th>Source – (R/Total)</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>NH</td>
<td>katG_A727T</td>
<td>GCC-&gt;ACC</td>
<td>(160/0/160)</td>
<td>Unclassified</td>
<td>UNK</td>
</tr>
</tbody>
</table>

Summary

The specimen was positive for Mycobacterium tuberculosis. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

Organism

The specimen was positive for Mycobacterium tuberculosis, lineage 2.2.1 (East-Asian Beijing).

Drug Susceptibility

- Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.
- □ No drug resistance predicted
- □ Mono-resistance predicted
- ✓ Multi-drug resistance predicted
- □ Extensive drug resistance predicted

<table>
<thead>
<tr>
<th>Drug class</th>
<th>Interpretation</th>
<th>Drug</th>
<th>Resistance Gene (Amino Acid Mutation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Line</td>
<td>Susceptible</td>
<td>Ethambutol</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Pyrazinamide</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Resistant</td>
<td>Isoniazid</td>
<td>katG (S315T)</td>
</tr>
<tr>
<td></td>
<td>Resistant</td>
<td>Rifampin</td>
<td>rpoB (S331L)</td>
</tr>
<tr>
<td>Second Line</td>
<td>Susceptible</td>
<td>Streptomycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Ciprofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Ofloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Moxifloxacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Amikacin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Kanamycin</td>
<td>No mutation detected</td>
</tr>
<tr>
<td></td>
<td>Susceptible</td>
<td>Capreomycin</td>
<td>No mutation detected</td>
</tr>
</tbody>
</table>
**Design results**: using evidence to inform the design

- Visual hierarchy that follows a clinical narrative
  - Grouping of common data elements (gestalt)
  - Judicious use of emphasis for “at-a-glance” read
  - Prioritize reading flow for clinical tasks
- LaTeX report that is programmatically generated

---

**MYCOBACTERIUM TUBERCULOSIS GENOME SEQUENCING REPORT**

- **Patient Name**: JOHN DOE
- **Barcode**: 12345678910
- **Birth Date**: 2000-01-01
- **Patient ID**: 12345678910
- **Location**: SOMEPLACE
- **Sample Source**: PULMONARY
- **Sample Date**: 2016-12-25
- **Sample ID**: A12345678
- **Sequenced From**: MGIT CULTURED ISOLATE
- **Reporting Lab**: LAB NAME
- **Report Date/Time**: 2017-01-01, 15:36
- **Requested By**: REQUESTER NAME
- **Requester Contact**: REQUESTER@EMAIL.COM

**Summary**

The specimen was positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

**Organism**

The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

**Drug Susceptibility**

- **Drug class Interpretation Drug Resistance Gene (Amino Acid Mutations)**
  - **First Line Susceptible**
    - Ethambutol
    - Pyrazinamide
  - **Resistant**
    - Isoniazid: katG (S315T)
    - Rifampin: rpoB (S531L)
  - **Not Resistant**
    - Streptomycin: No mutation detected
    - Ciprofloxacin: No mutation detected
    - Ofloxacin: No mutation detected
    - Moxifloxacin: No mutation detected
    - Amikacin: No mutation detected
    - Kanamycin: No mutation detected
    - Capreomycin: No mutation detected

- **Second Line Susceptible**
  - Capreomycin
  - Ethionamide
  - Kanamycin
  - Ofloxacin
  - Pyrazinamide
  - Rifampin
  - Streptomycin
  - Thiophanate methyl

---

- Design results: using evidence to inform the design
- Visual hierarchy that follows a clinical narrative
- Grouping of common data elements (gestalt)
- Judicious use of emphasis for “at-a-glance” read
- Prioritize reading flow for clinical tasks
- LaTeX report that is programmatically generated
**Design results**: using evidence to inform the design

- We tested alternative designs to come up with the final report

### Control Design
(Oiginal report)

<table>
<thead>
<tr>
<th>Drug Susceptibility</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoniazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

### Alternative 1

<table>
<thead>
<tr>
<th>Drug Susceptibility</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Prediction</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

- Based on predicted antibiotic mutations, the individual has multidrug resistant TB

### Alternative 2

<table>
<thead>
<tr>
<th>Drug Susceptibility</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>Prediction</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

- Based on predicted antibiotic mutations, the individual has multidrug resistant TB

---

**Alternatives Generated in 'Design Sprint' Step**
**Design results**: example of a typical result from our study

**A - Control**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isonazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

**B - Alternative**

**Drug Susceptibility**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isonazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Extreme Drug-resistant (XDR)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely Drug-resistant (XDR)</td>
<td></td>
</tr>
</tbody>
</table>

**C - Alternative**

**Drug Susceptibility**

<table>
<thead>
<tr>
<th>Drug</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isonazid</td>
<td>Resistant</td>
</tr>
<tr>
<td>Rifampin</td>
<td>Resistant</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>Resistant</td>
</tr>
<tr>
<td>Pyrazinimde</td>
<td>Resistant</td>
</tr>
</tbody>
</table>

**Clinician**

**Non- Clinician**

**Rescaled rank score**

**Comments from respondents:**

“the check boxes provide an at-a-glance result”

“tick boxes may cause confusion when clinicians read XDR without realizing that option is not selected.”
**Design results**: whole reports were actually confusing

- Asked participants to evaluate isolated components (previous) and whole reports

“None are especially good (see previous comments on individual parts)”

- Participant Comment
**Design results**: we did this for many design choices

- **Generally, alternative designs preferred**
  - in 12 out of 14 comparisons to control

- **Designs should promote patient safety & precise interpretability**
  - Abbreviations should be avoided
  - Debate about prioritizing susceptible vs. resistant drugs

- **Clinically actionable data to be given priority**
  - Surveillance tasks aren’t clinically actionable

- **Sometimes we didn’t provide good alternatives**
Deploy: report delivers insights that lead to actions

- Our report is used by Public Health England and several other global public health organizations, and public health software systems
- Community consensuses: we’re tackling a difficult and underappreciated part of biomedical research
Do **YOU** have to go through all this effort for every report?
It depends on what you want to achieve

- Broad data collection can be used for other projects
  - We were also collecting data for future software projects
  - Stayed tuned for more details!

- At the very least test alternative designs
  - If you can’t do a Discovery stage (time, people, budget) at least to the Design stage
  - Check in with stakeholders to avoid ad hoc design issues

- Bioinformaticians: you should use human-centered design for your tools!
  - Not command line ≠ user friendly
  - If you didn’t test it with even one user it’s not “user friendly” or “intuitive”
  - Report design is a very simple example of how to use these methods
Understand: community strategies for data visualization

- **Understand**: the good, the bad, and the common datavis solutions already in use
- **Strategy**: systematically mine and describe visualization solutions
- **Output**: a *domain prevalence* visualization design space

---

Crisan A., Gardy JL, Munzner T
"A systematic method for surveying data visualizations and a resulting genomic epidemiology visualization typology: GEViT"
OXFORD BIOINFORMATICS (2018)
Understand: community strategies for data visualization

- Different visualizations for hospital outbreaks – which choice is right?

Gorrie (2017)  
Willman (2015)  
Davis (2015)
Understand: community strategies for data visualization

• Is it a good idea to visualize data like this?

Gorrie (2017)
Can I come up with a method to systematically review data visualizations?

What will I find by trying this method out on microbial genomic epidemiology research papers?
Understand: an overview of our systematic method

- Our method has two analysis phases:
  - Literature Analysis
  - Qualitative Analysis
  - Quantitative Analysis
  - Visualization Analysis
**Understand: an overview of our systematic method**

- Some analyses are automated (🤖) and others are manual (👤).

### Analysis Phase

- **Literature Analysis**
- **Qualitative Analysis**
- **Visualization Analysis**
- **Quantitative Analysis**
Understand: an overview of our systematic method

- Analysis phases answer **different** research questions

<table>
<thead>
<tr>
<th>Analysis Phase</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Analysis</td>
<td>WHY are researchers visualizing data?</td>
</tr>
<tr>
<td>Qualitative Analysis</td>
<td>HOW are researchers visualizing data?</td>
</tr>
<tr>
<td>Quantitative Analysis</td>
<td>HOW MANY examples of specific visualizations?</td>
</tr>
</tbody>
</table>
Explore: apply our method to genomic epidemiology

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Acquisition &amp; Unsupervised Clustering</td>
<td>17,974</td>
<td>Article topic clusters</td>
</tr>
<tr>
<td>Limit to clusters of human pathogens</td>
<td>6,350</td>
<td>WHY are researchers visualizing data?</td>
</tr>
<tr>
<td>Random Stratified Sampling (clusters as strata)</td>
<td>221</td>
<td>801 figures</td>
</tr>
</tbody>
</table>
Explore: apply our method to genomic epidemiology

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Acquisition &amp; Unsupervised Clustering</td>
<td>17,974</td>
<td>Article topic clusters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WHY are researchers visualizing data?</td>
</tr>
<tr>
<td>Limit to clusters of human pathogens</td>
<td>6,350</td>
<td></td>
</tr>
<tr>
<td>Random Stratified Sampling (clusters as strata)</td>
<td>221</td>
<td>801 figures</td>
</tr>
<tr>
<td>Iterative &amp; axial coding</td>
<td>221</td>
<td>A genomic epidemiology visualization typology (GEViT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How are researchers visualizing data?</td>
</tr>
</tbody>
</table>
## Explore: apply our method to genomic epidemiology

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th># articles</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article Acquisition &amp; Unsupervised Clustering</td>
<td>17,974</td>
<td>Article topic clusters <strong>WHY</strong> are researchers visualizing data?</td>
</tr>
<tr>
<td>Limit to clusters of human pathogens</td>
<td>6,350</td>
<td></td>
</tr>
<tr>
<td>Random Stratified Sampling <em>(clusters as strata)</em></td>
<td>221</td>
<td>801 figures</td>
</tr>
<tr>
<td>Iterative &amp; axial coding</td>
<td>221</td>
<td><strong>A genomic epidemiology visualization typology (GEViT)</strong> <strong>How</strong> are researchers visualizing data?</td>
</tr>
<tr>
<td>Descriptive Statistics</td>
<td>221</td>
<td><strong>Current common visualization practices</strong></td>
</tr>
</tbody>
</table>
Understand: an overview of our systematic method

- Analysis phases answer **different** research questions

**Literature Analysis**

**WHY** are researchers visualizing data?

**Visualization Analysis**

**Qualitative Analysis**

**HOW** are researchers visualizing data?

**Quantitative Analysis**

**HOW MANY** examples of specific visualizations?
18K articles on genomic epidemiology

<table>
<thead>
<tr>
<th>PMID</th>
<th>Year</th>
<th>Pub</th>
<th>Journal</th>
<th>Authors</th>
<th>Title</th>
<th>Abstract</th>
<th>PMCID</th>
<th>DOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>27746398</td>
<td>2016</td>
<td>Clinical</td>
<td>PI Drèz-Lago</td>
<td>A novel strategy based on OBJECTIVE: Molecular epidemic NA</td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1371/journal.pntd.0003900</td>
</tr>
<tr>
<td>26819311</td>
<td>2016</td>
<td>Journal of</td>
<td>Bodewes, R Spatiotemporal Analysis o Influenza A viruses are major pi PMCS363327</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/jvi.03046-15</td>
</tr>
<tr>
<td>25113057</td>
<td>2014</td>
<td>Annals of</td>
<td>Chen, Sheng Spread of carbapenemase BACKGROUN The rapid emer PCMC236651</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s12941-014-0111</td>
</tr>
<tr>
<td>23938759</td>
<td>2013</td>
<td>Philosophi</td>
<td>Gray, Rebec Evolutionary analysis of he Reconstructing the transmission PCMC375181</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/st.2013.0161</td>
</tr>
<tr>
<td>22723256</td>
<td>2012</td>
<td>Journal of</td>
<td>Arduino, Son Transposons and integrans Multiple transposons, integrans NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>22386850</td>
<td>2012</td>
<td>Journal of</td>
<td>Infection, J Rivero-PI Drè Genetic diversity of comm With the recent detection of MI NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>19487205</td>
<td>2009</td>
<td>Philosophi</td>
<td>Slight, P M A HIV decision support: from Human immunodeficiency virus NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>19000628</td>
<td>2008</td>
<td>Virology</td>
<td>Chen, Rubing Frequent inter-species tra Revealing the factors that shap PMCM233721</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>18564687</td>
<td>2008</td>
<td>South Tiwari, Hare Molecular typing of clinical Molecular typing of total 84 Sta NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
<td></td>
</tr>
<tr>
<td>17617184</td>
<td>2007</td>
<td>Clinical mi</td>
<td>Kawalec, Mc Hospital outbreak of vanccA mixed outbreak caused by var NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>26598368</td>
<td>2015</td>
<td>Journal of</td>
<td>Box, Allison I Functional Analysis of Bac The classical and El Tor biotype PMCM479448</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>22502065</td>
<td>2012</td>
<td>Letters in</td>
<td>Aliabad, N H Molecular diversity of CTX AIMS: The objective of this stud NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>21645368</td>
<td>2011</td>
<td>BMC geno</td>
<td>Chen, YuanS Comparative genomic ana BACKGROUN Vibrio paraaer aPMC313071</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>21177922</td>
<td>2010</td>
<td>The Journal</td>
<td>Lorussao, Ales Genetic and antigenic cha Prior to the introduction of the : PMCM333703</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>20375036</td>
<td>2010</td>
<td>The Journal</td>
<td>Post, Virginia Evolution of AbaR-type ge OBJECTIVES: To determine if m NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>18815309</td>
<td>2008</td>
<td>Journal of</td>
<td>Salloum, Shi Escape from HLA-B*08-re The inherent sequence diversity PMCM258368</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>17784948</td>
<td>2007</td>
<td>BMC geno</td>
<td>Chen, YuanS Comparative analysis of Modern epidemiology of foodx PMCM346545</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>17651134</td>
<td>2007</td>
<td>FEMS micro</td>
<td>Chen, YuanS Genetic variation of capsu Both NRT36S and AS are NAG-1 NA</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
</tr>
<tr>
<td>15213324</td>
<td>2006</td>
<td>Proceeding Holdent, Matti Complete genomes of twc Staphylococcus aureus is an imp PMCM470752</td>
<td></td>
<td></td>
<td></td>
<td>NA</td>
<td>10.1128/s0292-010-0021</td>
<td></td>
</tr>
</tbody>
</table>
Articles clustered primarily by pathogen

Final Topics Clustering Results
t-SNE & hdbscan

- Strategic subset: sample fixed # of articles from each cluster
  - Random stratified sampling
- Could capture variability in visualizations across pathogens – if it existed
- Sampling resulted in 221 articles that yielded 801 figures
Adjutant: unsupervised topic discovery for literature reviews

- People liked the literature review method, so I made it into an R package!
Understand: an overview of our systematic method

• Analysis phases answer **different** research questions

<table>
<thead>
<tr>
<th>Analysis Phase</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature Analysis</td>
<td>WHY are researchers visualizing data?</td>
</tr>
<tr>
<td>Quantitative Analysis</td>
<td>HOW are researchers visualizing data?</td>
</tr>
<tr>
<td>Qualitative Analysis</td>
<td>HOW MANY examples of specific visualizations?</td>
</tr>
</tbody>
</table>

Visualization Analysis

- WHY are researchers visualizing data?
- HOW are researchers visualizing data?
- HOW MANY examples of specific visualizations?
Qualitative analysis: how are figures constructed

- Extract figures from sample articles
- Figures are not tied to any one specific tool
  - Includes post-processing made on figures
- Used iterative and axial coding from qualitative methods
  - All figures analyzed separately
  - Multipart figures (i.e. Fig 3A, 3B) analyzed together
- Result: GEViT, a genomic epidemiology visualization typology
  - Taxonomic hierarchy describing chart types, chart combinations, and chart enhancements
GEViT chart types: the basic building blocks

<table>
<thead>
<tr>
<th>Common Statistical Charts</th>
<th>Tree Charts</th>
<th>Genomic Charts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar Chart</td>
<td>Phylogenetic Tree</td>
<td>Genomic Map</td>
</tr>
<tr>
<td>Standard</td>
<td>Rooted (Linear &amp; Radial)</td>
<td>Linear</td>
</tr>
<tr>
<td>Stacked</td>
<td>Unrooted (Linear &amp; Radial)</td>
<td>Radial</td>
</tr>
<tr>
<td>Divergent</td>
<td>Alignment</td>
<td>Composition Plot</td>
</tr>
<tr>
<td>Special Cases</td>
<td>Clonal Tree*</td>
<td></td>
</tr>
<tr>
<td>Epidemic Curve</td>
<td>Sequence Logo Plot</td>
<td></td>
</tr>
<tr>
<td>Diversity Chart</td>
<td>Other Charts</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>LeafPlot</td>
<td>Table</td>
<td>Gel Image</td>
</tr>
<tr>
<td>Line Chart</td>
<td>Image</td>
<td>General Image</td>
</tr>
<tr>
<td>Scatter Plot</td>
<td>Special Cases</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>Special Cases</td>
<td>Phylogenetic Tree</td>
<td>Genomic Map</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>Root-tip-to-tip</td>
<td>Linear</td>
</tr>
<tr>
<td>Kaplan-Meier</td>
<td>Ordination Plot</td>
<td>Radial</td>
</tr>
<tr>
<td>Skyline Plot</td>
<td>Q-Q plot</td>
<td>Composition Plot</td>
</tr>
<tr>
<td>Colour Charts</td>
<td>Relational Charts</td>
<td>Flow Diagram</td>
</tr>
<tr>
<td>Pie Chart</td>
<td>Node-link</td>
<td>Chord Diagram</td>
</tr>
<tr>
<td>Venn Diagram</td>
<td>Special Cases</td>
<td>Sankey Diagram</td>
</tr>
<tr>
<td>Category</td>
<td>EBurst</td>
<td></td>
</tr>
<tr>
<td>Stripe</td>
<td>Social network</td>
<td></td>
</tr>
<tr>
<td>Heatmap</td>
<td>Molecular network</td>
<td></td>
</tr>
<tr>
<td>Density Plot*</td>
<td>Minimum Spanning Tree</td>
<td></td>
</tr>
<tr>
<td>Spatial Charts</td>
<td>Geographic Map</td>
<td>Choropleth Map</td>
</tr>
<tr>
<td>Flow Diagram</td>
<td>Interior Map</td>
<td></td>
</tr>
<tr>
<td>Streamgraph*</td>
<td>Absolute</td>
<td>Relative</td>
</tr>
<tr>
<td>Temporal Charts</td>
<td>Timeline</td>
<td></td>
</tr>
<tr>
<td>Node-link</td>
<td>Spatial Charts</td>
<td>Geographic Map</td>
</tr>
<tr>
<td>Flow Diagram</td>
<td>Choropleth Map</td>
<td>Interior Map</td>
</tr>
<tr>
<td>Chord Diagram</td>
<td>Sankey Diagram</td>
<td></td>
</tr>
</tbody>
</table>
GEViT chart types: current common practices

- A tree is the most common chart type
  - Not a surprise
- A lot of data is in text and not visualized
GEViT chart *combinations* : showing data together

- **Baseline** is just a single chart

<table>
<thead>
<tr>
<th>Combination Type</th>
<th># of chart types</th>
<th># of charts</th>
<th>Linkage type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td><img src="chart.png" alt="Example" /></td>
</tr>
</tbody>
</table>

**Current Practice**

40.1% of all figures

Example of a simple chart

![Example of a simple chart](chart.png)

Associated with social milieu

Number of cases

Year of diagnosis
GEViT chart combinations: showing data together

- Users also combined individual charts to tell more aspects of the data.

<table>
<thead>
<tr>
<th>Combination Type</th>
<th># of chart types</th>
<th># of charts</th>
<th>Linkage type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td>OR</td>
</tr>
<tr>
<td>Spatially Aligned</td>
<td>Many</td>
<td>1</td>
<td>Horizontal or Vertical</td>
<td>AND</td>
</tr>
<tr>
<td>Small Multiples</td>
<td>1</td>
<td>Many</td>
<td>Chart Type &amp; Data</td>
<td>AND</td>
</tr>
<tr>
<td>Color Aligned</td>
<td>Many</td>
<td>Many</td>
<td>Visual, but not spatial</td>
<td>AND</td>
</tr>
<tr>
<td>Unaligned</td>
<td>Many</td>
<td>Many</td>
<td>NA</td>
<td>AND</td>
</tr>
</tbody>
</table>

Current Practice:
- 40.1% of all figures
- 20.3%
- 17.3%
- 13.5%
- 8.8%
GEViT chart combinations: showing data together

- Users also combined individual charts to tell show more aspects of the data

<table>
<thead>
<tr>
<th>Combination Type</th>
<th>Spatially Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td># of charts</td>
<td>1</td>
</tr>
<tr>
<td># of chart types</td>
<td>Many</td>
</tr>
<tr>
<td>Linkage</td>
<td>Horizontal and vertical alignment</td>
</tr>
</tbody>
</table>

Example of a spatially aligned combination
GEViT chart *combinations* : showing data together

- Users also *combined* individual charts to tell show more aspects of the data

<table>
<thead>
<tr>
<th>Combination Type</th>
<th>Small Multiples</th>
<th>Example of a small multiples combination</th>
</tr>
</thead>
<tbody>
<tr>
<td># of charts</td>
<td>Many</td>
<td></td>
</tr>
<tr>
<td># of chart types</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Linkage</td>
<td>Chart type &amp; data</td>
<td></td>
</tr>
</tbody>
</table>
GEViT chart combinations: showing data together

- Users also combined individual charts to tell show more aspects of the data

<table>
<thead>
<tr>
<th>Combination Type</th>
<th>Color Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td># of charts</td>
<td>Many</td>
</tr>
<tr>
<td># of chart types</td>
<td>Many</td>
</tr>
<tr>
<td>Linkage</td>
<td>Visual linkage (color)</td>
</tr>
</tbody>
</table>

Example of a color aligned combination
GEViT chart combinations: showing data together

- Users also combined individual charts to tell more aspects of the data.

<table>
<thead>
<tr>
<th>Combination Type</th>
<th># of chart types</th>
<th># of charts</th>
<th>Linkage type</th>
<th>Example</th>
<th>Current Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>1</td>
<td>1</td>
<td>NA</td>
<td></td>
<td>40.1% of all figures</td>
</tr>
<tr>
<td>Spatially Aligned</td>
<td>Many</td>
<td>1</td>
<td>Horizontal or Vertical</td>
<td></td>
<td>20.3%</td>
</tr>
<tr>
<td>Small Multiples</td>
<td>1</td>
<td>Many</td>
<td>Chart Type &amp; Data</td>
<td></td>
<td>17.3%</td>
</tr>
<tr>
<td>Color Aligned</td>
<td>Many</td>
<td>Many</td>
<td>Visual, but not spatial</td>
<td></td>
<td>13.5%</td>
</tr>
<tr>
<td>Unaligned</td>
<td>Many</td>
<td>Many</td>
<td>NA</td>
<td></td>
<td>8.8%</td>
</tr>
</tbody>
</table>
GEViT chart enhancements: adding metadata to chart types

**Current Practice**

>80% of all figures have some enhancement
GEViT chart enhancements: adding metadata to chart types
GEViT chart enhancements: adding metadata to chart types

Chart Enhancement Examples

- **Base Chart**: Tree
  - **Re-encode Marks**: Line: color
  - **Add Marks**: Point: color, line; text: font face
  - **Add Annotation**: Arrow, text

**Group 1**, **Group 2**, **No known contacts**
Understand: now we could systematically describe visualizations

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?
GEViT in action: knowledge translation via the GEViT Gallery


OXFORD BIOINFORMATICS
What does GEViT do and not do?

**GEViT provides a base**
- A Visualization Typology for visual design
  - Chart Types
  - Chart Combinations
  - Chart Enhancements
- An Interactive Gallery

**GEViT does not evaluate**
- Massive undertaking that would take many years
- GEViT provides platform for future evaluation research
minCombinR: simplified visual generation in R

```r
# specify individual charts
phyloTree_chart <- specify_base(chart_type = "phylogenetic tree", data = "tree_dat")
epicurve <- specify_base(chart_type = "histogram", data = "tab_dat", x = "month")
map_chart <- specify_base("geographic map", data = "tab_dat", lat = "latitude", long = "longitude")

# specify a combination
colour_combo <- specify_combination(combo_type = "color_linked", base_charts = c("phyloTree_chart", "map_chart", "epicurve"), link_by = "country")

# plot the result
plot(color_combo)
```
# Analyze different data types automatically
harmon_obj <- data_harmonization(tab_dat, tree_dat, genomic_dat, all_spatial)

# Create specifications that compile to minCombinr
component_specs <- get_spec_list(harmon_obj)

# Plot the result one view at a time
plot_view(component_specs, view_num = 1)
Part III
Data visualization strategies and tools for microbial genomic epidemiology

Anamaria Crisan
Vanier Canada Scholar & UBC Public Scholar
PhD Candidate, Computer Science
University of British Columbia