Avant-Garde: Data Visualization Tool for HIV Epidemiology

By: Yu Xia
y2xia@eng.ucsd.edu

HIV Epidemic in San Diego - Tijuana Border

- Over 50 million U.S./Mexico border crossings each year
- In Tijuana, high rate of illicit drug usage, needle sharing, sex workers without protection all contribute to the growing prevalence of HIV infected patients
- 70% of the female sex workers have sex partners from the U.S.

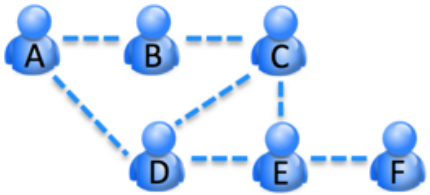
Researchers Want to Understand:

1. Are the San Diego and Tijuana epidemics connected?
2. What is the direction of viral migration across the border?
3. What are the risk factors driving the border HIV epidemic?

Molecular Epidemiology

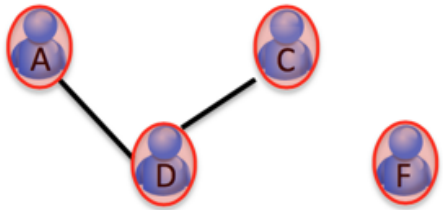
Using the relationships between viral sequence data sampled from infected individuals to characterize the spread of HIV

1) AACTGATCGC
2) AACAGATCGC



A Sexual Network: Real World Network

- Dashed lines indicate sexual contact



An HIV Transmission Network: What we see

- Solid lines indicate very similar viruses
- Red circles indicate HIV infected persons
- Other individuals are HIV negative so don't appear

Sanjay Mehta,
Molecular Epidemiology of HIV in San Diego and Tijuana. 2015

Exploratory Data Analysis

- Traditionally, researchers will come up with **hypothesis first** and then gathers the data, analyze it and verify the hypothesis.
- EDA emphasis on the **data first**. By using visual or statistical tools to analyze the data, researchers may find new perspectives or suggest hypotheses.

Project Goal

- Implement a data visualization tool to help medical researchers better understand and analyze the HIV patient data sets.
- Keep the tool independent from the data source, so it can support other usage scenarios
- Study and iteratively develop the new approach for the interactive visualization of phylogenetics, geographical data and socio-demographics

Project Scope

- Data set provided by UCSD AntiViral Research Center (AVRC)
- Include studies from multiple medical sites from San Diego and Tijuana areas
- Over 50 data variables are included: demographics, medical history, lifestyle, blood test, viral strain, etc

Data Set

- 1111 HIV sequences were collected from the collaborating studies (including 263 background sequences)

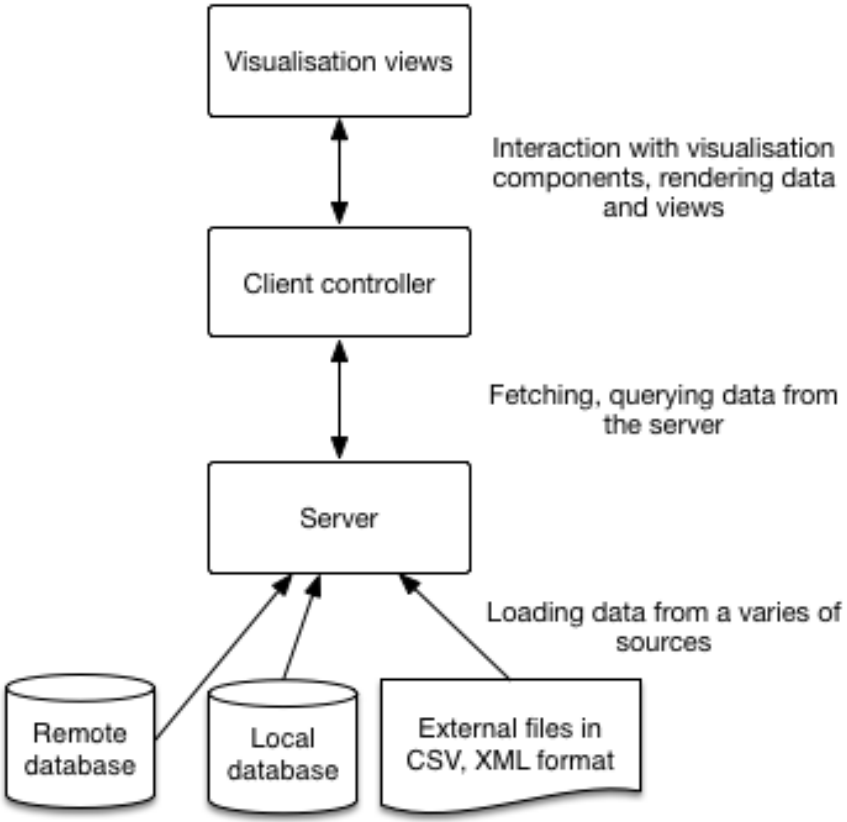
- The remaining 1107 sequences included – 76 from subjects residing in Tijuana 768 from subjects residing in San Diego – 263 background sequences from San Diego

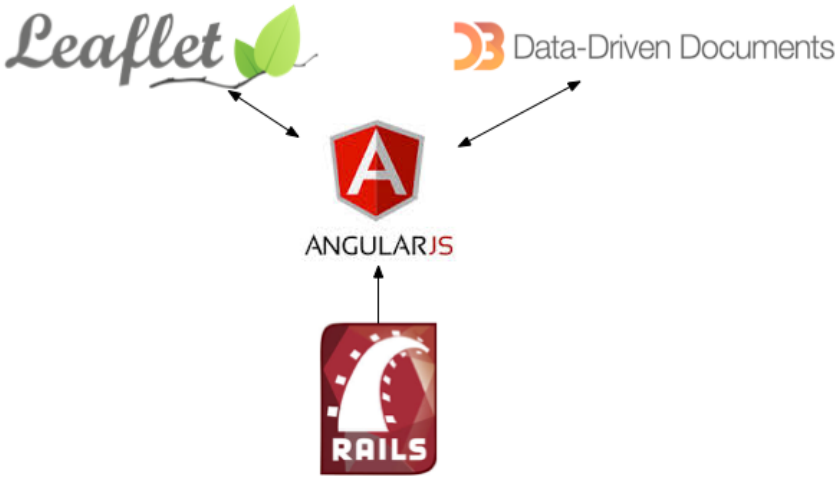
Data Variables Types

- Unique identifier fields (pid)
- Categorical fields (gender, ethnicity, etc)
- Numerical value fields (HIV loads)
- Date fields (enrollment date)
- Geographical fields (address, Zip code)

Demo

System Architecture





Server-side Functionalities

- Data loading
  - Loading data files in CSV, JSON format into the DB
  - Loading from from local DB directly
  - Interfacing with remote DB
- Data querying
  - Take client-side data query request and return filtered results
- Data rendering
  - Different visualization components require data in different structures, we prepared corresponding data templates for the server-side to use

Client-side Components

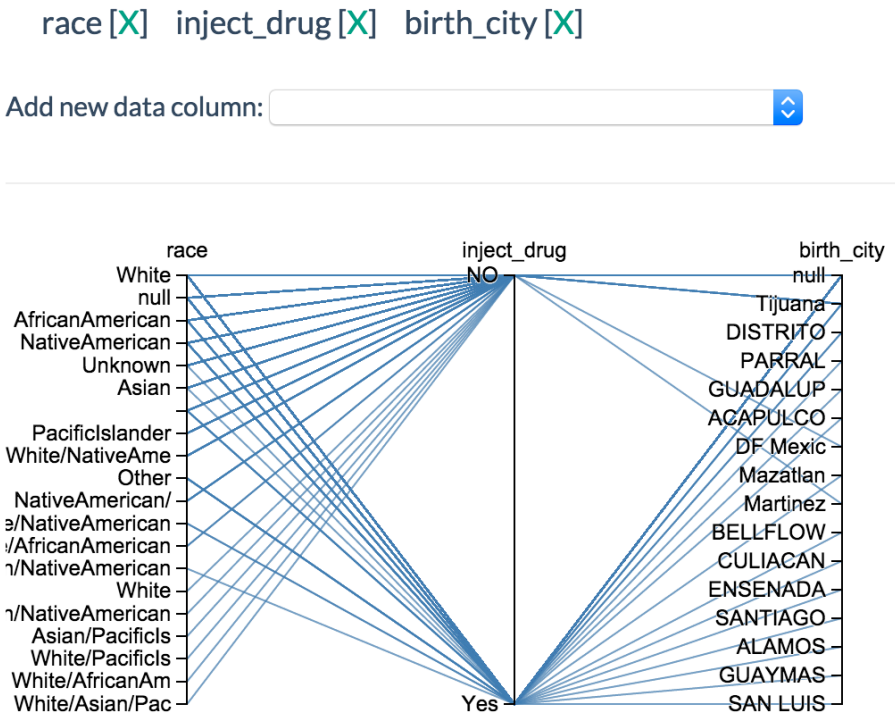
- General visualization components
- Time-based visualization components
- Heat map
- Data table
- Visualization path

Circle Packing

- Visualize large amounts of hierarchically structured data
- Easier to comprehend visually
- Display all hierarchies at once

Parallel Coordinate

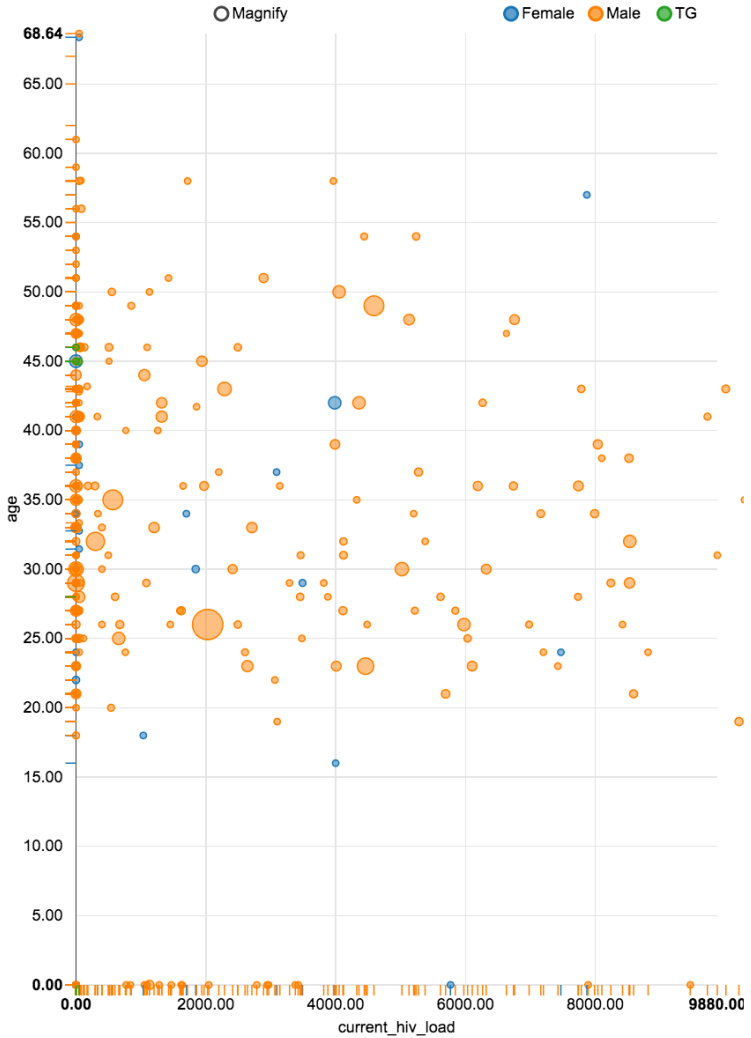
- Scalable framework to visualize multidimensional data
- Increase of the data dimensions corresponds to the addition of extra axes
- Data brushing to filter the data



Scatter Plot

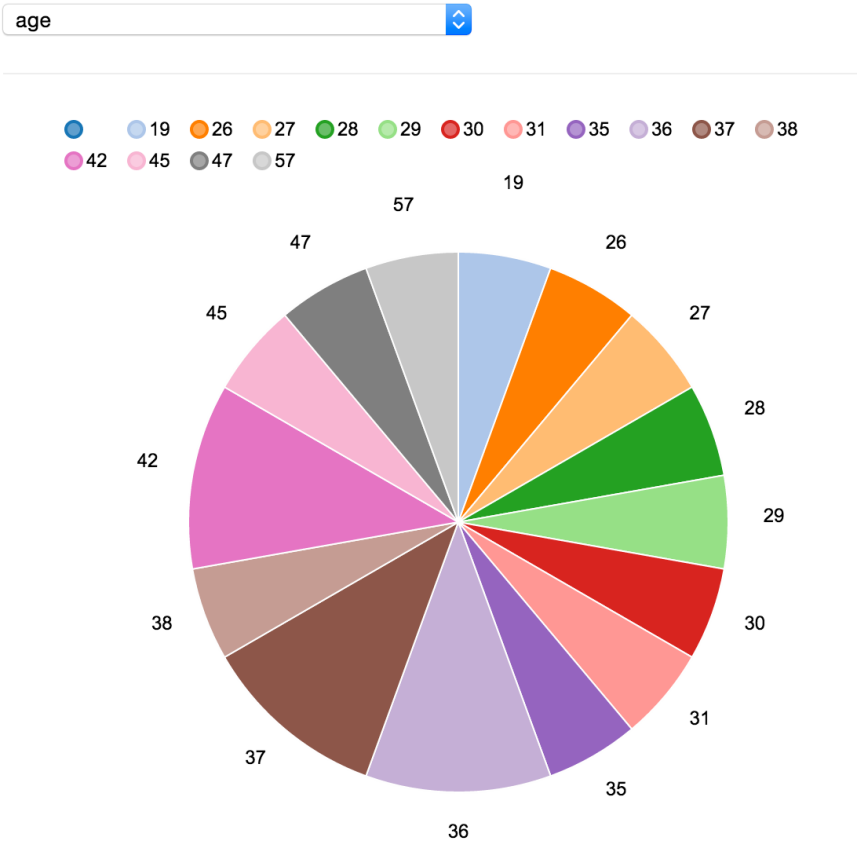
Plot data points on a horizontal and a vertical axis to show how much one variable is affected by another

Depict data in four dimensions X axis, Y axis, group color, shape size



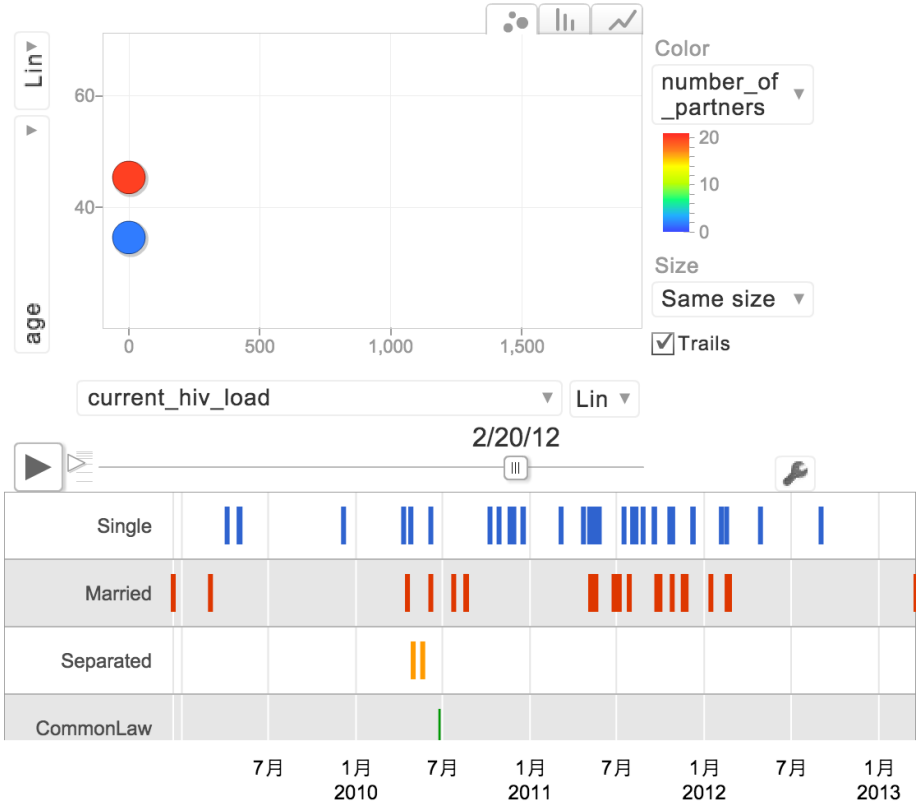
Pie Chart

- Circular chart divided into sectors, illustrating relative magnitudes or frequencies
- Can visualize data set under different data columns
- Can hide certain data value within the chart



Motion Chart & Timeline Chart

Time-based visualization components can help researchers to understand how certain data set column's value change across the time.



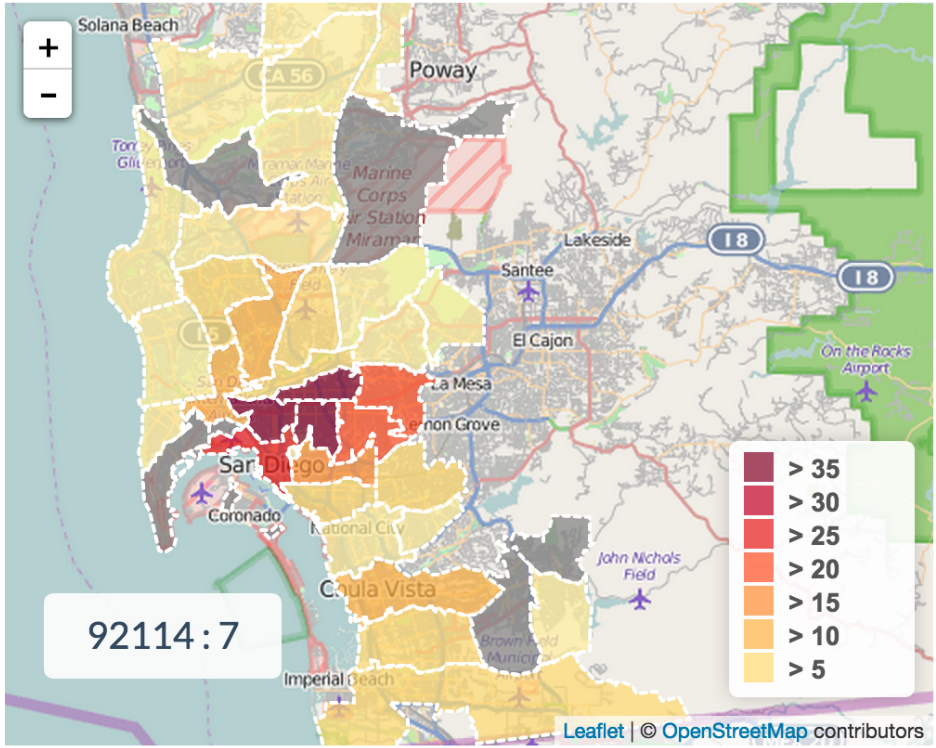
Force-Directed-Graph

Force-directed-graph is used to visualize clustered data relationship.

Heatmap

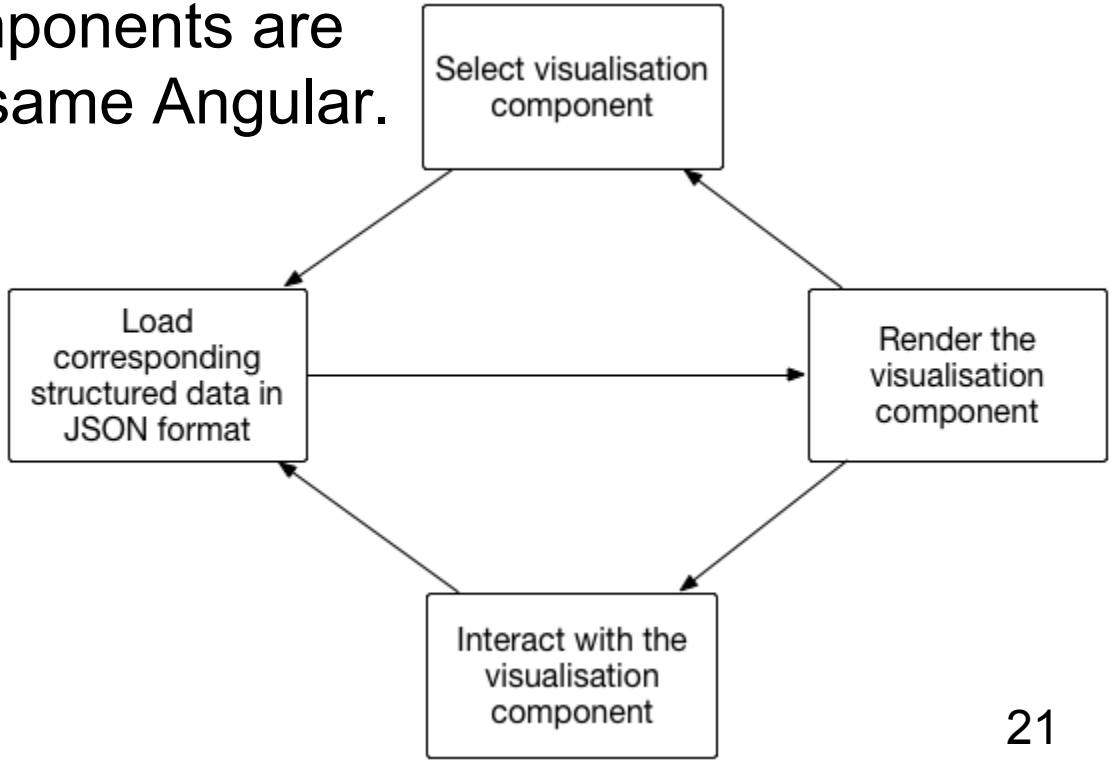
Show the data set's geographical distribution on the map

Customizable and extensible in terms of the overlay tiles, map interactions, coloring, legends.

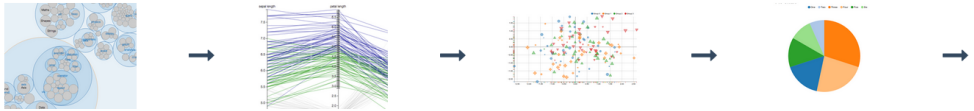


Client-side Interaction

All the visualization components are connected through the same Angular.js controller



Visualization Path



Visualisation components in the current path

Select next visualization

Path Name

Save Current Path

Save current visualisation path

Current visualiation: pie

Path info

{"cluster_id":"15"}

- Visualization path enables the users to leverage different visualization components to analyze the same data set
- Save the path for future reuse or share with colleague

Permission Control

- User management for general auth control
- Configurable per dataset rule based on user’s access level

Extensibility

Although we only implemented four general visualization components, it is quite easy to build additional ones.

- Implement your d3.js visualization as a Angular.js directive
- Wrap your visualization code as part of the nvd3.js components

Evaluation

We have deployed the application on production environment and invited researchers to use it and provide feedback.

Usage Feedback

- Intuitive to use
- Visualization path is useful, but still need enhancement
- Raw data need better pre-processing
- Improve on data selection and filtering
- Support more visualization components

Future Work

- Consolidate the code base
- Incorporate more visualization components
- Explore how we can enhance the tool to be plug and play against different data set
- Handling massive data set, what is the issues with the current implementation and how we solve them.

Acknowledgements

- Professor Nadir Weibel
- Dr. Sanjay Mehta from AVRC
- Sandy Law
- Yingyan Hua

For continuous aid, advice and feedback during the project development process.

The End