RESEARCH PLAN

The transmission of HIV involves often complex social and sexual interactions among people, which are rarely straightforward. Understanding such human and viral interactions will require a new analysis framework based on a combination of phylogenetic, socio-demographic and geographic data from HIV-infected and at-risk populations [30]. Analysis of these multi-dimensional data promises to uncover hidden social, sexual, geographic and virologic relationships, but new advanced visualization and interactive tools are needed to explore such links. We propose to investigate how web-based interactive visualization methods can help investigators understand the influence of external factors on HIV transmission networks. Using a novel approach based on Stanford University's Data Driven Documents—or D3 [2]—we will access real-time data, analyze correlations and gain insights into how epidemiologic, geographic and virologic data are associated in different dimensions, all in an effort to better inform and target prevention interventions for HIV.

1. Specific Aims and Hypotheses

Aim 1: To develop and validate an interactive and web-based visualization approach able to integrate and cohesively represent phylogenetic, clinical, geographic, and socio-demographic data from HIV-infected and at-risk populations as collected from the primary infection cohort, the UCSD Owen Clinic and the San Diego County Epidemiology Department. The interactive visualization aims to be flexible and reconfigurable, enabling exploratory data analysis and "what-if" investigations.

- Hypothesis 1A: Cutting-edge web-based data visualization techniques (i.e. D3) can be used to create visual clusters of interacting variables from individuals. Interactive visualization of this data will enable parallel analysis in multiple dimensions (e.g. sexual behavior, drug use, residence, demographics) and exploration of hidden connections between the many factors that influence the spread of HIV-infection in the population.
- Hypothesis 1B: The distribution over time of key events causing the growth of particular HIV phylogenetic clusters (i.e. transmission hotspots) can be automatically summarized by building interactive visualization of study variables (e.g. sexual behavior, risk venues, demographics) aligned on a timeline-based model.

Aim 2: To implement automatic mechanisms to identify, report and visualize abnormal or substantial growth (or decrease) of HIV-infected individuals in particular phylogenetic, geographic or socio-demographic clusters. The goal is to enable targeted interventions on particular identified risk characteristics.

- Hypothesis 2: Data from the San Diego Primary Infection Cohort (P.I. Little), the UCSD Owen Clinic (CFAR Clinical Core), Network Informed Prevention Study (PI. Smith) and the San Diego County Epidemiology (Epidemiologist Dr. Tweeten) can be effectively monitored in such a way to identify particular changes in the spread of HIV infections. Such changes can be graphically implemented through specific interactive visualizations to further and quickly explore the reasons behind particular growths and inform interventions.

2. Backgrounds and Significance

Proper medical management and good adherence to highly active antiretroviral therapy (HAART) enable the effective treatment of HIV infected individuals and markedly reduce their infectiousness [22,23,33]. However, HIV therapy is lifelong and the costs associated with the medications and monitoring are extremely high [5,12]. Preventing even one HIV transmission would introduce costs savings on the order of $500,000-$1,000,000 dollars [28]. Although prevention is a recognized key to reduce costs for HIV and, in the US, mass education campaigns and widely available testing have been deployed for years, HIV incidence has not decreased significantly [25]. Additionally, a population-wide preventative (i.e. 'Universal') approach would require a large expenditure to reach all individuals at risk. In this era of limited public health resources, targeted approaches that provide the maximum benefit for the cost are necessary. We believe that targeted models of prevention could be more effective to decrease HIV incidence, similar to the ring vaccination methods used in the Small Pox eradication campaign in the 1970s [10].

Current ongoing work at UCSD [30] has recently started to co-analyze virologic, epidemiologic, clinical, geographic, and socio-demiographic data from HIV-infected individuals to characterize the local transmission network and use the results to target HIV prevention resources. This ongoing work is transitioning HIV epidemiology from a passive monitoring tool to an active real-time surveillance system that has the potential to impact HIV transmission in high risk groups. The key of this approach is to integrate a variety of datasets across the San Diego region—San Diego Primary Infection Cohort, the UCSD Owen Clinic, Network Informed Prevention Study and the San Diego County Epidemiology Department—and build a HIPAA compliant database of the phylogenetic, socio-demographic and geographic characteristics of HIV infected individuals in the region, as well as update this repository in a real-time manner. While this extensive information should allow mapping of identified simple relationships, it will not allow for deep interrogation of complex relationships between highly dissimilar data, like phylogenetic, socio-demographic and geographic clustering. To more fully realize the potential of these already collected data, we propose to develop and validate methods for the visualization of such complex interactions. **Specifically, we propose to enhance the existing data collection methods with a novel web-based interactive visualization approach that enables researchers to interact with the data and explore in real-time the many facets of the local transmission network**. Our approach aims to enable new ways to visually and interactively investigate the complexities of the defined social and sexual networks, phylogenetic networks and geography of risk venues and residence. Investigators will be able for example to correlate an observed increase of HIV+ cases in real time with demographic, geographic and phylogenetic clustering variables.

Innovation

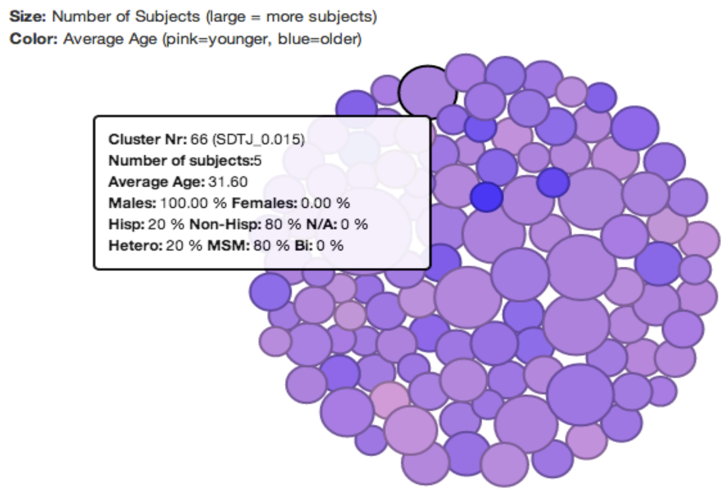
Analysis of a limited number of dimensions in terms of epidemiological data is already challenging, and often this results in the reduction of the ‘problem space’ to single dimensions [18], with outcomes quantifying the single variables rather then the relationship between them. Analyzing the prevalence or spread of HIV-infections across socio-demographical dimensions is even more challenging. Recent attempts [20] used spatial visualization to make sense of such data and showed promising results. However, these methods require the definition of a clear a-priori model for analysis that often is hard to develop when so many variables are at stake. This is where Exploratory Data Analysis (EDA) techniques come into play. As suggested by the mathematician and statistician John Tukey, often too much emphasis in statistics is placed on statistical hypothesis testing and more emphasis needs to be placed on using data to suggest hypotheses to test [37]. We feel that the multi-dimensional data that is being collected from HIV-infected individuals and at risk populations in San Diego perfectly exemplify this situation. **The innovation of our approach stems from enabling the explorations of multiple inter-related variables and data types and their correlation with spread of HIV-infection and the detection of anomalies in a graphical and interactive way**. This approach enables exploratory data analyses and “let the data suggest new hypotheses”. Only by enabling these loose, but flexible and multi-level analysis capabilities, can phenomena and relationships be analyzed and characterized.

3. Preliminary Studies and Feasibility

This project is situated at the intersection of epidemiology, molecular virology, data visualization and human-computer interfaces. We are confident that employing and integrating the techniques that we describe in this proposal will lead to a novel and important way to get access and investigate new correlations across the HIV transmission network on multiple dimensions. We are convinced of the feasibility of this project based on the following preliminary results and experiences.

- A) HIV Datasets: Data collection of the socio-demographic, geographic and phylogenetic data has already started. Currently more than 1000 data entries have been collected from four different sites. Every entry is characterized by a HIV sequence, phylogenetic clustering, demographic information (e.g. age, gender, race, ethnicity, marital status, sexual orientation, education level, birth country and city), sexually transmitted infections, sexual risk (e.g. transaction sex, commercial sex, number of partners), illicit drug use (e.g. needle sharing, use of heroin, meth, alcohol), clinical data (e.g. HIV stage at diagnosis, current CD4 levels, HAART use), risk venues (e.g. bathhouses, adult bookstores, bars, internet sites used to meet partners for sex or drug use), and geography (e.g. residence, border crossing). Data is being constantly collected at an annual rate of 30-50 HIV+ and 3000 HIV- high risk individuals and will continue to be collected for the duration of this project.

- B) Data Driven Documents (D3): D3 [2] is a mature technology that is widely used to build interactive web-based information visualizations. The core philosophy behind D3, aims to bind input data to arbitrary web documents elements and enable to both generate and modify content. D3 comes with a range of libraries, in particular *d3js* (D3 JavaScript) [3] that are aimed to quickly prototype scalable and dynamic web-based visualizations. Following a design-by-example paradigm [13], D3 presents thousands of examples on freely consultable code repositories such as the D3 Gallery [4]. Examples come from academia, industrial labs, and the press with the New York Times often publishing D3-based visualizations [32].



Size: Number of Subjects (large = more subjects)
Color: Average Age (pink=younger, blue=older)

Cluster Nr: 66 (SDTJ_0.015)
Number of subjects:5
Average Age: 31.60
Males: 100.00 % **Females:** 0.00 %
Hisp: 20 % **Non-Hisp:** 80 % **N/A:** 0 %
Hetero: 20 % **MSM:** 80 % **Bi:** 0 %

Figure 1: Prototype D3 interactive visualization representing the clusters of HIV+ individuals' data collected to date.

- C) Preliminary Prototype: Based on the data collected to date, we built a preliminary D3 prototype to study the feasibility of our approach. The interactive visualization (Figure 1) automatically creates visual clusters represented by adjacent circles of different sizes and colors. Size represents the number of subjects pertaining to the same group, color the average age. Supplementary descriptive information about the cluster (e.g. average age, ethnicity, sexual orientation) can be accessed by hovering with the cursor on one of the clusters. Visualization elements can be tailored to the specific needs or preference of the user. We envision this prototype visualizations to be used in a variety of situations: for instance, if an investigator wants to know how many Hispanic HIV+ men who have sex with men visited a particular venue in the past three months, then the visualization tool could graph these elements over time, phylogenetic clustering, clinical factors, etc.

4. Research Design and Methods

Good information visualization can elicit new insights on data not otherwise visible [6,34,35,36]. A wonderful example is John Snow's map of London of 1854 that was able to stop an infamous cholera outbreak, by visualizing the correlation between cholera cases and the water supply network linked to a bad water pump [9]. Adding interactivity capabilities to visualizations enables further explorations and quick hypothesis testing, such as the ones presented by Hans Rosling with Gapminder [26] who introduced animated statistics in many settings, but most importantly in public health [11,27].

As part of this project we will develop a range of interactive visualizations of the collected socio-demographic, geographic and phylogenetic data. The overall goal is to build a fully functional web-based, HIPAA compliant system that allows researchers to get real-time access to flexible and customizable visualization of the collected data. The research design of the proposed project maps the hypotheses and aims defined in the previous section. The project will follow a combination of the human centered design lifecycle [15] and the Rapid Applications Development (RAD) [21] approaches. To iteratively develop the proposed platform following a contextual and participatory design [1,7,16,29], we plan to actively involve researchers across multiple disciplines through the CFAR Scientific Focus Group platforms. In this way, agile software development and active user involvement in the design of the proposed prototype–both critical for the success of this project–will be adequately addressed.

Aim 1: In the first part of the project we will be **study and iteratively develop the new approach for the interactive visualization of phylogenetics, geographical data and socio-demographics**. This will be based on D3 [2] open source software. This aim will deliver three initial core prototype visualizations:

a) Multi-layered descriptions of phylogenetic clusters: Summaries and descriptive statistics of phylogenetic clusters will be visualized through pie charts when hovering cursor on specific clusters. Multiple layers of data can then be visualized side-by-side in a pop-up window. We will use brushing techniques [17,31] to highlight how data visualized in one dimension (e.g. sex) will be distributed in another (e.g. age).

b) Network connectivity: Every cluster will be represented by a graph of the phylogenetic distance of each subject. Nodes represent individuals, the length of the edges between nodes represents their phylogenetic distance. We will use the TN93 algorithm based transmission network reconstruction approach developed by Dr. Kosakovsky Pond and Dr. Joel Wertheim [19,24] to calculate the nodes' distance, and force-directed graph visualization to place nodes [14].

c) Time-based dynamic update of clusters: The basic cluster-based visualization showed in Figure 1 will be augmented with a time component that will enable explorations of how clusters change over time. We will superimpose the interactive visualization described above on a dynamic timeline that can be operated (e.g. dragged, zoomed) to show data at a specific time in the past. The timeline will also be animated as in [27].

In order to continuously assess the effectiveness of the developed prototypes we will employ an iterative model and involve clinical and translational research stakeholders. We already have agreements from key members of CFAR (Drs. Davey Smith, Sara Gianella, Sanjay Mehta, Matthew Strain and Susan Little). The web-based nature of our prototypes will make it easy to elicit contextual feedback on key areas of the visualizations. To facilitate contextual feedback we will work with flexible web-based annotation tools, such as Notable [38].

Aim 2: The second part of the project will focus on **identifying and visualizing data over time when specific important changes happen**. This will be based on three main components: (1) A flexible data analytic framework to enable researchers to issue structured search query to the data currently visualized; (2) A user-based filter system able to combine a list of arbitrary parameters for any of the available dimensions (e.g. difference in cluster size > 50 during two consecutive months); (3) A dynamic visualization engine able to map user-based filters to the analytic framework and present a list of annotations on the timeline when the particular conditions specified in the filters are satisfied.

By clicking on the dynamically generated time-based annotations, users will be able to further analyze the selected data at that specific time-point and test new hypotheses that drove those particularly important changes. Selected parameters can be saved and applied to new data being added in real-time, enabling the identification of similar conditions also in the future. For instance, investigators who identify particular growth in HIV+ clusters in specific areas of San Diego and correlate it with drug use, could instruct the system to monitor future increases of drug use, automatically visualize those episodes, and be notified when this happens.

Future Directions.

The current collected data is one of the best HIV-related data collections in the world, but it could be better. For example, consistent geographic data are still missing for some participants, limiting the development of geo-based interactive visualization. However, these data are now being added to the assembled datasets. Their upcoming availability will allow us to develop geo-specific visualization and map network data on a regional map of San Diego enabling further understanding of the local impact of HIV incidence. With more comprehensive data we plan to finalize visualization prototypes and build a solid infrastructure that could be deployed to the broader CFAR community and used to explore additional factors in the context of the local transmission network. We also plan to extend data collection to more sites beyond the San Diego area and test if the developed framework is scalable, while making our infrastructure available to a wider audience.

The identification of key moments in time is an important feature needed to enable targeted interventions in a timely manner. We plan to augment this with automatic algorithmic and machine learning components able to discover patterns and detect anomalies. Creating automatic alerts would benefit researchers that would be notified directly without the need to explicit poll the data and execute manual queries.

We also plan to exploit the visualization facilities of Calit2 at UCSD such as the VROOM [8] and our experience working with large visualization interfaces, to study how multidimensional data such as the ones analyzed here can profit of visualization and interaction on high-resolution wall displays.

Finally, this CFAR Development grant will not only develop an important HIV-related project that will be useful to many CFAR investigators. It will also develop Dr. Nadir Weibel, an Assistant Research Scientist in the Department of Computer Science at UCSD, into a new and promising HIV investigator.

5. Literature Cited

1. Beyer, H. and Holtzblatt, K. Contextual design. *interactions 6*, 1 (1999), 32–42.
2. Bostock, M., Ogievetsky, V., and Heer, J. D3; Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2301–2309.
3. Bostock, M., Ogievetsky, V., and Heer, J. D3 JavaScript Library. http://d3js.org.
4. Bostock, M., Ogievetsky, V., and Heer, J. D3 Design Gallery. http://github.com/mbostock/d3/wiki/Gallery.
5. Bozzette, S.A., Joyce, G., McCaffrey, D.F., et al. Expenditures for the care of HIV-infected patients in the era of highly active antiretroviral therapy. *The New England journal of medicine 344*, 11 (2001), 817–823.
6. Card, S.K., MacKinlay, J.D., and Schneiderman, B. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.
7. Clemensen, J., Larsen, S.B., Kyng, M., and Kirkevold, M. Participatory design in health sciences: using cooperative experimental methods in developing health services and computer technology. *Qualitative health research 17*, 1 (2007), 122–130.
8. Defanti and Schulze, J.P. Immersive Visualization Laboratory - Infrastructure. http://ivl.calit2.net/wiki/index.php/Infrastructure.
9. Fekete, J.-D., Van Wijk, J.J., Stasko, J.T., and North, C. The value of information visualization. In *Information Visualization*. Springer, 2008, 1–18.
10. Fenner, F., Henderson, D.A., Arita, I., Jezek, Z., and Ladnyi, I.D. Smallpox and its eradication. *World Health Organization; Geneva*, (1988).
11. Gapminder Foundation. Wealth and Health of Nations. http://www.bit.ly/RRjy6C.
12. Gebo, K.A., Fleishman, J.A., Conviser, R., et al. Contemporary Costs of HIV Health Care in the HAART Era. *AIDS (London, England) 24*, 17 (2010), 2705–2715.
13. Hartmann, B., Doorley, S., and Klemmer, S.R. Hacking, Mashing, Gluing: Understanding Opportunistic Design. *IEEE Pervasive Computing 7*, 3 (2008), 46–54.
14. Holten, D. and Van Wijk, J.J. Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum 28*, 3 (2009), 983–990.
15. International Organization for Standardization. ISO 9241-210:201: Ergonomics of human-system interaction - Part 210: Human-centered design for interactive systems. 2010.
16. Johnson, C.M., Johnson, T.R., and Zhang, J. A user-centered framework for redesigning health care interfaces. *Journal of biomedical informatics 38*, 1 (2005), 75–87.
17. Keim, D.A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 1–8.
18. Kilmarx, P.H. Global epidemiology of HIV. *Current opinion in HIV and AIDS 4*, 4 (2009), 240–246.
19. Kosakovsky Pond, S.L. TN93 Distance Matrix Generator. http://github.com/spond/TN93.
20. Lòpez-De Fede, A., Stewart, J.E., Hardin, J.W., Mayfield-Smith, K., and Sudduth, D. Spatial visualization of multivariate datasets: an analysis of STD and HIV/AIDS diagnosis rates and socioeconomic context using ring maps. *Public health reports (Washington, D.C.: 1974) 126 Suppl 3*, (2011), 115–126.
21. Millington, D. and Stapleton, J. Developing a RAD standard. *IEEE Software 12*, 5 (1995), 54–55.
22. Moore, R.D. and Chaisson, R.E. Natural history of HIV infection in the era of combination antiretroviral therapy. *AIDS (London, England) 13*, 14 (1999), 1933–1942.
23. Palella, F.J., Jr, Delaney, K.M., Moorman, A.C., et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *The New England journal of medicine 338*, 13 (1998), 853–860.
24. Pond, S.L.K. and Muse, S.V. HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*. Springer, 2005, 125–181.
25. Prejean, J., Song, R., Hernandez, A., et al. Estimated HIV Incidence in the United States, 2006–2009. *PLoS ONE 6*, 8 (2011), e17502.
26. Rosling, H., Rosling, O., and Rosling Rönnlund, A. Gapminder Foundation. http://www.gapminder.org.
27. Rosling, H. and Zhang, Z. Health advocacy with Gapminder animated statistics. *Journal of Epidemiology and Global Health 1*, 1 (2011), 11–14.
28. Schackman, B.R., Gebo, K.A., Walensky, R.P., et al. The lifetime cost of current human immunodeficiency virus care in the United States. *Medical care 44*, 11 (2006), 990–997.

29. Schuler, D. and Nacmioka, A. *Participatory Design: Principles and Practices*. Lawrence Erlbaum.
30. Smith, D. NIH NIDA - AvantGarde: Molecular epidemiology for HIV prevention for drug users and other risk groups. 2012.
31. Swayne, D.F., Lang, D.T., Buja, A., and Cook, D. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis 43*, 4 (2003), 423–444.
32. The New York Times. Four Ways to Slice Obama's 2013 Budget Proposal. http://www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html.
33. Torres, R.A. and Barr, M. Impact of combination therapy for HIV infection on inpatient census. *The New England journal of medicine 336*, 21 (1997), 1531–1532.
34. Tufte, E. *Beautiful Evidence*. Graphics Press, 2006.
35. Tufte, E.R. and Graves-Morris, P.R. *The visual display of quantitative information*. Graphics press Cheshire, CT, 1983.
36. Tufte, E.R. Envisioning information. *Optometry & Vision Science 68*, 4 (1991), 322–324.
37. Tukey, J.W. Exploratory data analysis. *Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass 1*, (1977).
38. Zurb Inc. Notable - Better Interfaces Through Faster Iteration. http://www.notableapp.com.