Avant-Garde HIV Research:

Harmonizing and Visualizing Patient Data

By: Sandy Law
s4law@eng.ucsd.edu

Introduction

- Human Immunodeficiency Virus
  - a virus that compromises the immune system
  - pandemic (35.3 million living with HIV in 2012 [1])
  - complex social factors influencing transmission
- “avant-garde” approach to HIV research
  - innovative, creative, novel
  - potentially open new avenues
  - HIV treatment and prevention

Exploratory Data Analysis

- Typically, researchers formulate hypothesis then gathers data
  - requires having the idea first
- EDA approach emphasizes data first
  - use data to suggest hypotheses
  - analyze with visual or statistical tools
  - may offer fresh perspectives

Scope

- UCSD AntiViral Research Center (AVRC)
- Studies from medical sites
  - San Diego and Tijuana areas
  - Amigo, Proyecto El Cuete, Mujer Mas Segura, Hombre Seguro, STAHR, STAHRII, Parejas
- Selection of approx. 50 fields
  - demographics, medical history, lifestyle etc.
  - drawing data from 9 unique sources

Previous Workflow

1. Studies ask patients questions and collect medical data
2. Medical sites export data
3. Email exchange with UCSD researchers
4. Manually converted and compiled
5. Stored in a flat table structure
6. Research is conducted
   - HIV sequencing

Motivation

1. Improve on an inefficient workflow
   a. reduce the need to handle raw data
   b. give access to more current data
   c. handle aggregating data from different sources
   d. remove email exchanges
2. Explore tools that aid EDA techniques
   a. intuitive, robust, multi-featured
   b. handle multidimensional, complex data
   c. let researchers to view HIV patient data in new ways

Aggregating Data

- Same questions, slightly different answers
- Ex: “What is your gender?”
  - Amigo Study, Female is stored as “0”
  - El Cuete, Female is stored as “2”
- A universal “codebook” detailing a data dictionary for these values
- Conversions for every data source, every data field

Changing Formats (Small)

- Smaller transitions as well
- a few field names changing, more answer options added, etc.
- Ex: Storing results of nucleic acid-based test
  - “result” prior to Dec. 2013
  - “nat_result” after
- Depending on collect date, appropriate value

Changing Formats (Large)

- AVRC - Approx. 5 years ago, “BBL” database structure abandoned
- Ex: “Symptomatic of Pharyngitis”
  - stored under “EHXRF2” in BBL format
  - “pharyngitis” in new format
- BBL data only partially moved to new
- Treat as two different sources of data

Maintaining Data Integrity

- Combining data from multiple sources
- similar yet different questions/answers
- dealt on case by case basis with care
- consult those who have experience

Ex: Sexual Orientation

1. “What is your sexual orientation?”
2. “Whom do you have sex with?”
3. “Do you think of yourself as heterosexual, homosexual or bisexual?”

Is the intention the same? A patient who identifies as heterosexual may still engage in homosexual intercourse.

Ex: Sexual Partners

1. total # of sexual partners
2. # of male sex and # of female
3. # of casual sex partners and is there a regular partner?

Are questions 2 and 3 the sum of their parts?
Unintentionally excluding certain possibilities?

Ex: A Chlamydia Diagnosis

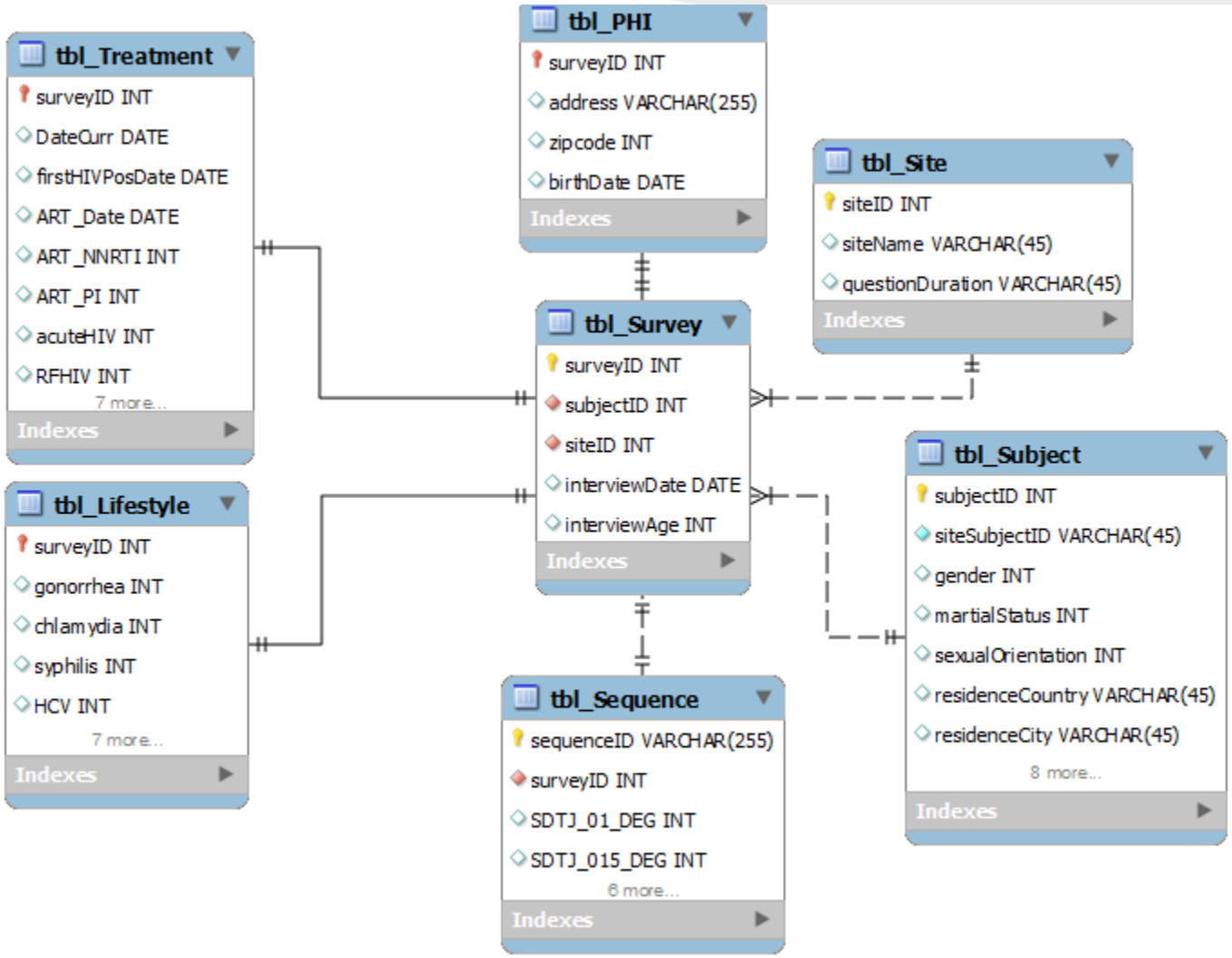
Questions whose answers are tied to a period of time. “Have you been diagnosed with chlamydia in the past x months?”

- All sources have different “x” value
- “no” at 3 months doesn’t mean “no” for 4
- “yes” at 6 months doesn’t mean “yes” for 3
- No tidy solution, couple (answer, time)

Implementation

- Agile-like development process
- periodic meetings and ongoing conversation
- capture appropriate collection of fields
- understanding workflow in detail
- accurate representation of data

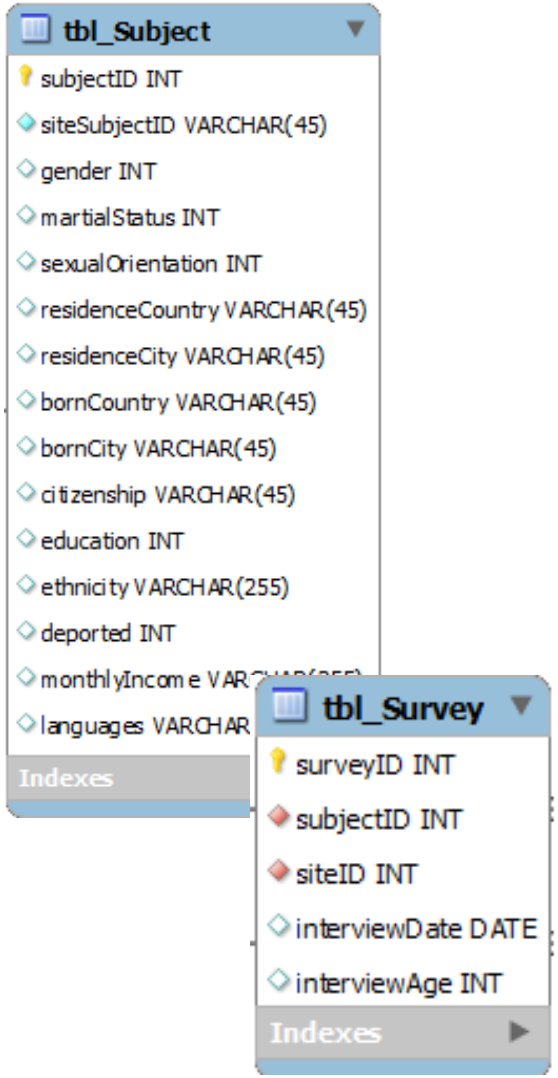
Database Schema



Relational Database

- Improve on flat table format
- logical separations to normalize data
- universal and uniform id

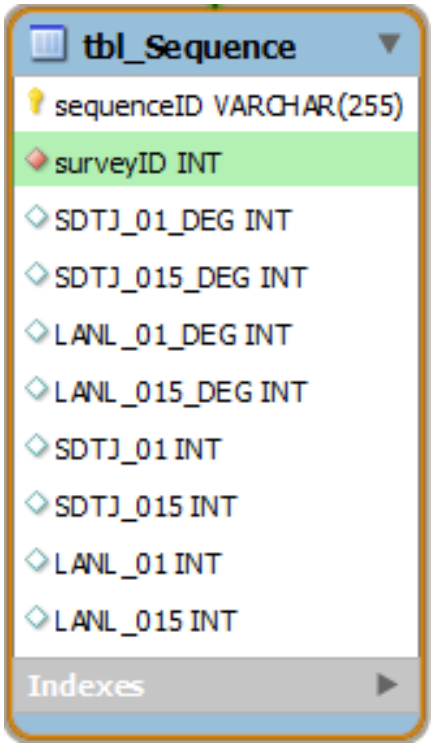
Dividing Subject & Survey



- Participating in multiple surveys
- Subject fields
  - birth country, gender, ethnicity, etc.
- Survey categories
  - lifestyle (homeless, heroin use, etc.)
  - treatment (cd4 levels, rash, etc)

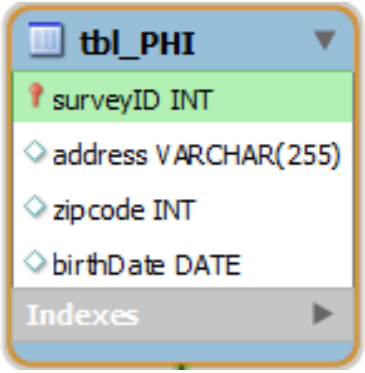
HIV Sequencing

- researchers generate IDs for HIV sequences
- identify different HIV strains
- ties sequence to the patient



Protected Health Information

- PHI is private patient data
  - higher standards of storage/management
  - protect privacy
  - zip codes, addresses, names, etc.
- Beyond our scope
- Avenue for future expansion



Scheduled Script

- AVRC database frequently exports data into CSV files on their server
- nightly Python script
  - join CSV files via Pandas.DataFrame [2]
  - Linux “diff” command compares it to most recent [3]
  - php script handles new/modified entries

Web Interface

Aggregate Sequence | Amigo ElCuete HombreSeguro MMS Parejas STAHR STAHRII | BBL AEH

Xataface

Search Aggregate Advanced Search

New

Import

Data Dictionary

List Records

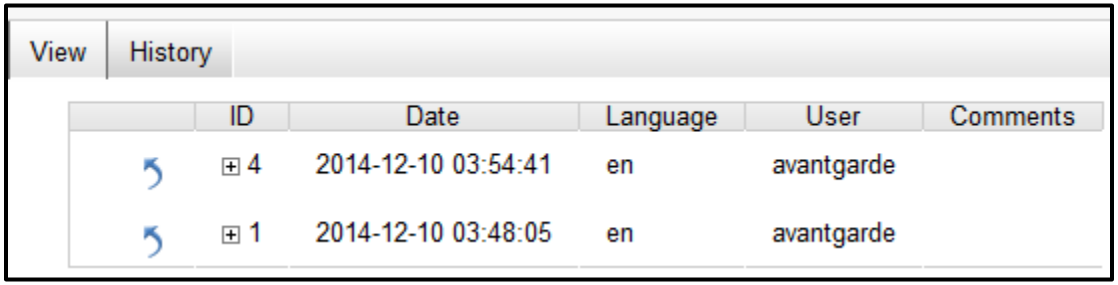
Export CSV Visualize Selected Visualize All

		Survey ID	Site Subject ID	Site Name	Date of Interview	Baseline Age (in years)	Question Duration (in months)
		1	A025	Amigo	07/11/2011	44	4
		2	A053	Amigo	10/21/2011	36	4
		3	A188	Amigo	08/02/2012	27	4
		4	EC029	ElCuete	04/06/2011	34	6
		5	EC043	ElCuete	04/08/2011	35	6
		6	EC060	ElCuete	04/11/2011	33	6
		7	EC092	ElCuete	04/13/2011	30	6

Xataface, an open source PHP framework [4]

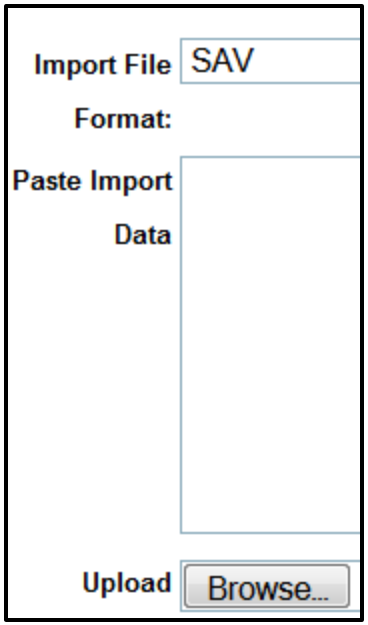
Administrators

- Control user access
- Permissions with fine granularity
  - assign roles
  - import, delete, add, change, etc.
- Monitor data
  - view/restore History
  - make changes, resolve anomalies



Importers

- Medical sites can upload their data
- IBM SPSS software data format
- Custom import filters
  - PSPP (open-source replacement) to handle format
  - preview data
  - data stored in site-specific table
  - conform data to our codebook
  - insert to our database



Researchers

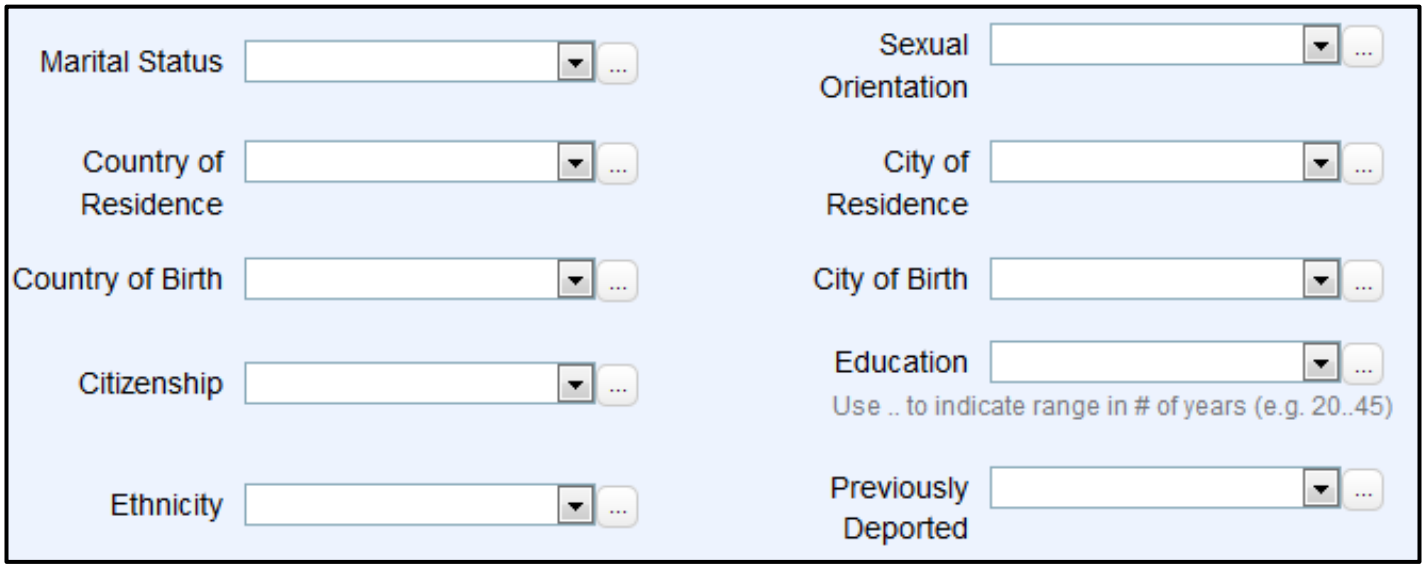
- A MySQL view showing sequence IDs and corresponding survey
  - export unfinished data points to a CSV
  - sequence ID filled in, imported into database
- Data dictionaries explaining values

```
Sequence:
CREATE VIEW Sequence AS (
   SELECT sur.surveyID, seq.sequenceID,
      seq.SDTJ_01_DEG, seq.SDTJ_015_DEG,
      seq.LANL_01_DEG, seq.LANL_015_DEG,
      seq.SDTJ_01, seq.SDTJ_015,
      seq.LANL_01, seq.LANL_015
   FROM tbl_Sequence seq
   RIGHT JOIN  tbl_Survey sur
   ON seq.surveyID = sur.surveyID
)
```

El Cuete (6 Months)	
GENDER:	1-Male, 2-Female, 3-Transexual, 8-Refuse to Answer
BORNTJ:	1-Yes, 0-No, 7-Don't Know, 8-Refuse to Answer
BRTHCTRY:	0-Mexico, 1-United States, 2-Other, 7-Don't Know, 8-Refuse to Answer

Researchers

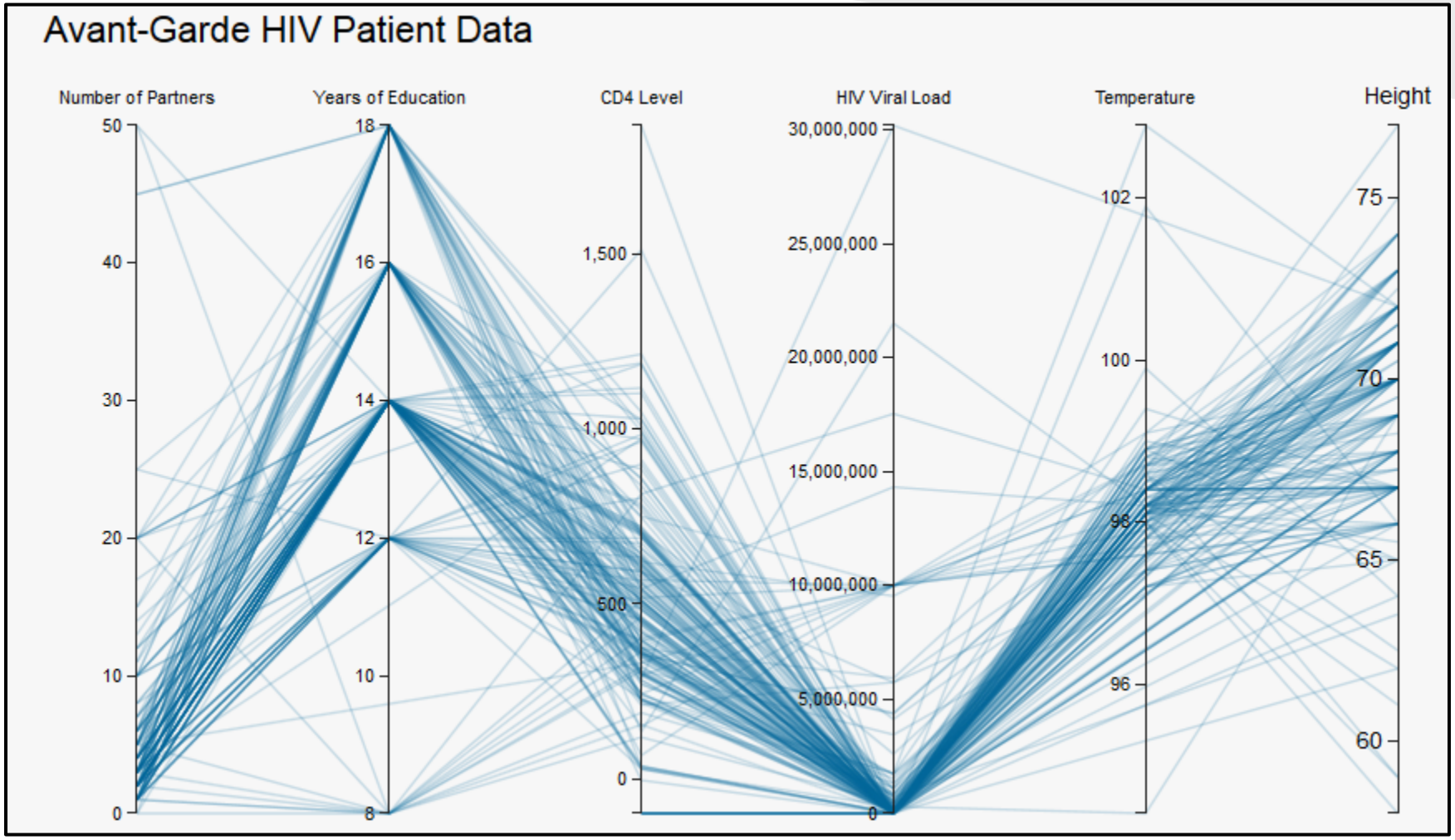
- View an aggregation of the data
- Search to narrow data sets
- Export and/or visualize



Visualization

- EDA approach: visualize data sets
- D3.js - JavaScript library [5]
  - Data-Driven Documents
  - components used to build variety of graphs
- Parallel coordinates [6]
  - “transforms multivariate relations into 2-D patterns” - Alfred Inselberg [7]
  - researchers choose a data set, represent them in a parallel coordinates graph

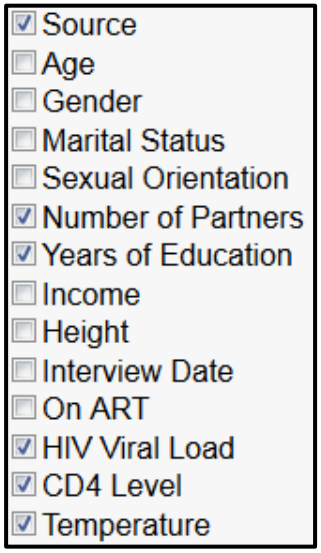
Parallel Coordinates



Interactivity

Maximizing effectiveness of EDA methods by increasing interactivity.

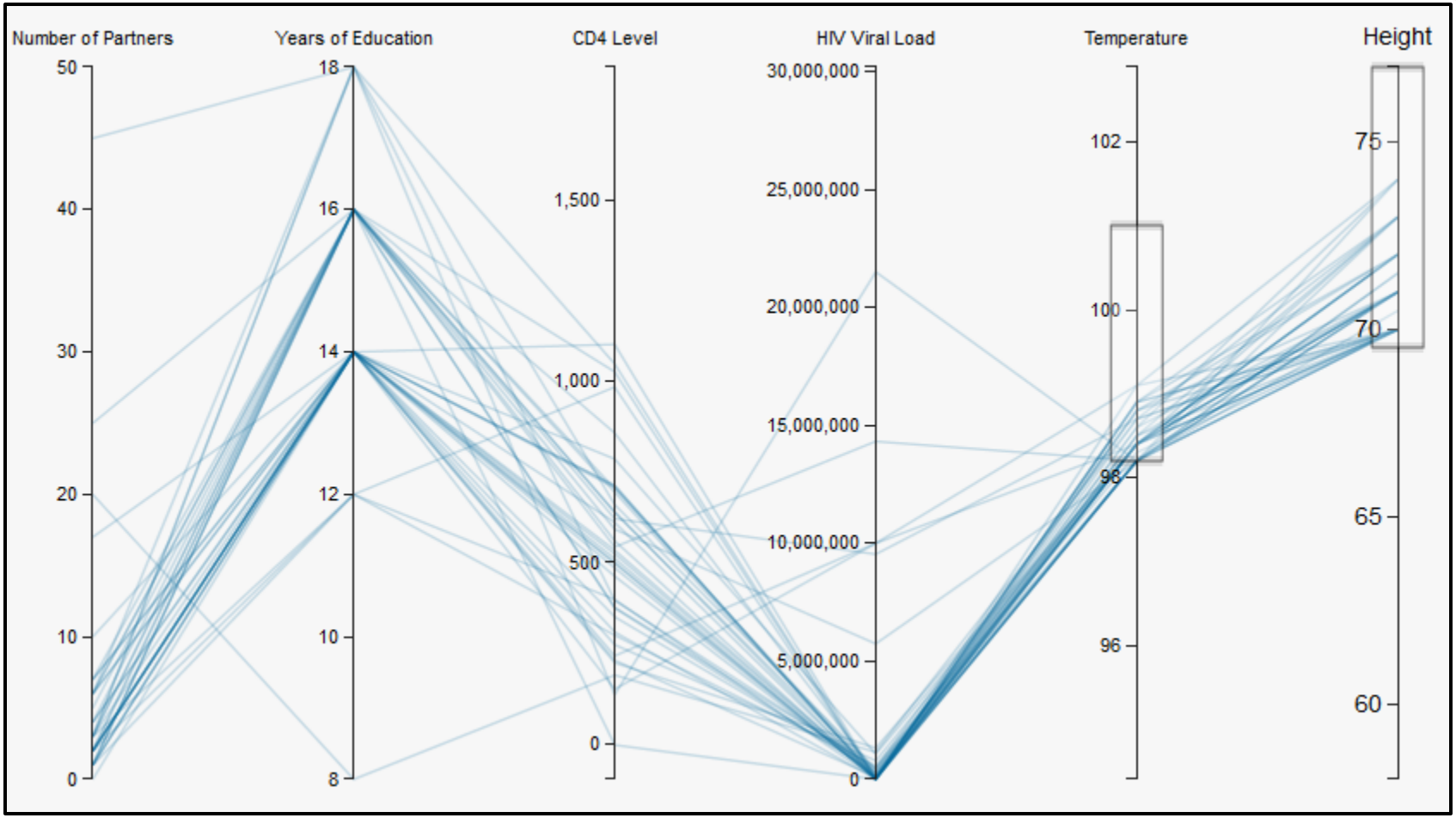
- Adding and removing axes
- Rearranging by dragging left or right
- Brushing



Number of Partners	Years of Education	CD4 Level	HIV Viral Load	Temperature	Height
4	16	592	5742130	98.2	71.5
2	14	364	9571	Data Unavailable	Data Unavailable
2	14	786	1077629	98.4	70
5	8	509	10214	98.1	74
2	14	611	29375	98.2	70
2	14	379	372604	98	59
10	18	270	1530180	98	70
4	14	886	11503	98.1	69
15	16	301	785926	98.2	69
5	16	428	1621	97.6	69
Data Unavailable	8	364	77015	97.8	67
2	8	464	97808	98.1	67
3	14	282	1084	97.2	67

Brushing

Selected areas of each axis to filter data.



Results

- New database structure
  - MySQL views to tailor displays to researcher needs
  - minimalizing impact of adding fields & tables
- Web interface
  - tie together all aspects
  - regulating access to data
  - automatic harmonizing across sources
- Parallel Coordinates
  - show HIV patient data from new perspective
  - interactivity to facilitate EDA approach

Conclusion

Two motivations: improve workflow and explore tools to aid Exploratory Data Analysis.

1. Created web-based tool handling logistical issues that were in the way of HIV research.
2. Provided the first of many visualizations that may help the formation of new HIV research ideas.

Future Work

- Putting the system into production
- More feedback from medical sites and researchers
- Expand our scope (sites and fields)
- More visualizations variety
- More interactivity & intuitiveness within them

Acknowledgements

- Professor Nadir Weibel
- Dr. Sanjay Mehta
- Christy M. Anderson

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

References

[1] UNAIDS Global Report 2013 Fact Sheet. http://www.unaids.org/en/resources/campaigns/globalreport2013/factsheet

[2] API Reference > pandas.DataFrame. http://pandas.pydata.org/pandas-docs/dev/generated/pandas.DataFrame.html

[3] About.com Linux. http://linux.about.com/library/cmd/blcmdl1_diff.htm

[4] Hannah, Steve. Xataface. http://xataface.com/wiki/about

[5] D3 Data-Driven Documents. http://d3js.org/

[6] Chang, Kai. Parallel Coordinates. https://syntagmatic.github.io/parallel-coordinates/

[7] Inselberg, Alfred. *Multidimensional Detective*. Tel Aviv University, Israel. http://web.cs.ucdavis.edu/~ma/ECS289H/papers/Inselberg1997.pdf

The End

Researcher Feedback