National Human Services Data Consortium

2019 Spring Conference

Nashville, TN

April 15-17, 2019

Trust Building and Boundary Spanning:

Bridging Gaps in Data Collection and Analysis

Increasing Capacity & Building Connections: **Bridging to the Future**

Sponsored by the National Human Services Data Consortium

Center for Academic Community Engagement
10 Years
2008 - 2018

Dr. Ruth Kassel

Assoc Dir Academic Integration Siena College in Albany NY

Board Member-CARES of NY

(518) 782-6951

rkassel@siena.edu

CARES
OF NY, INC
ENDING HOMELESSNESS

Allyson Thiessen

Director of HMIS, serving 13 CoCs (25 Counties) for CARES of NY

Board member-NHSDC

(518) 489-4130 x103

athiessen@caresny.org

In 2016 Siena reached out to CARES of NY for a possible partnership

Siena was in need of big, messy data sets for analysis. Quickly.

CARES of NY had LOTS of big, messy data in dire need of analysis (because HMIS)

This was either going to go really well or really poorly, but we agreed we were up to giving it a try and hope for The Unicorn*…. a mutually beneficial relationship.*

We believe in unicorns!

Academic Community Engagement SPIn Program

Three Year Plan

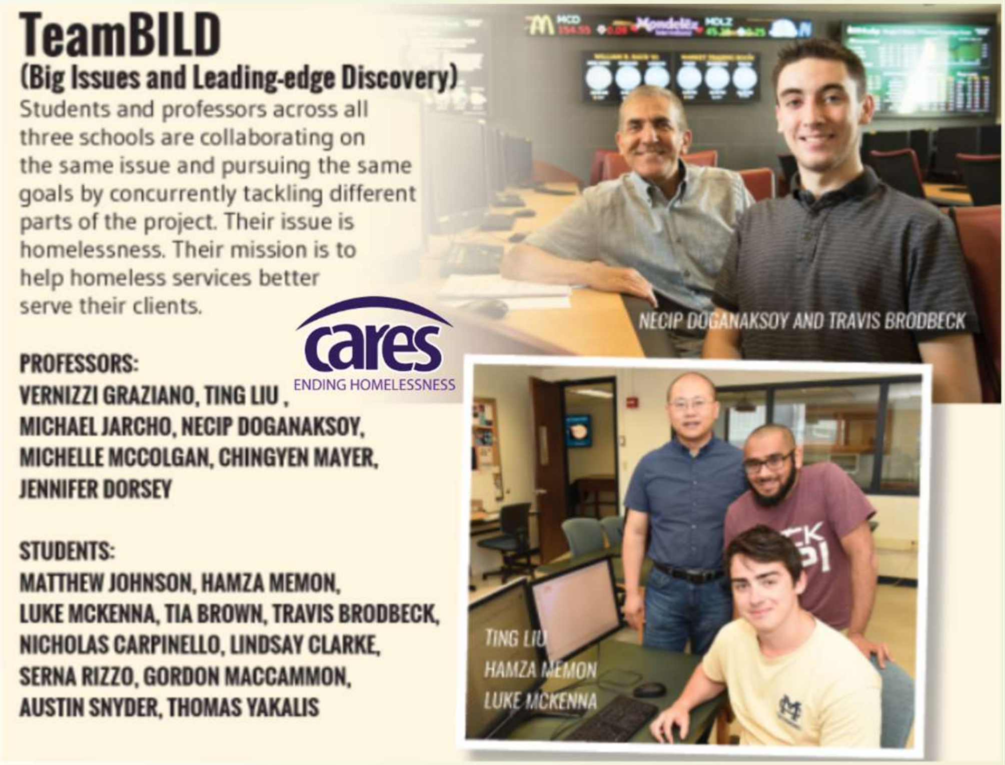
Y1- Relationship Building

Y2- Identity Building

Y-3 Sustainability

Center for Academic Community Engagement
10 Years
2008 - 2018

CARES
OF NY, INC
ENDING HOMELESSNESS



TeamBILD

(Big Issues and Leading-edge Discovery)

Students and professors across all three schools are collaborating on the same issue and pursuing the same goals by concurrently tackling different parts of the project. Their issue is homelessness. Their mission is to help homeless services better serve their clients.

cares
ENDING HOMELESSNESS

PROFESSORS:
VERNIZZI GRAZIANO, TING LIU , MICHAEL JARCHO, NECIP DOGANAKSOY, MICHELLE MCCOLGAN, CHINGYEN MAYER, JENNIFER DORSEY

STUDENTS:
MATTHEW JOHNSON, HAMZA MEMON, LUKE MCKENNA, TIA BROWN, TRAVIS BRODBECK, NICHOLAS CARPINELLO, LINDSAY CLARKE, SERNA RIZZO, GORDON MACCAMMON, AUSTIN SNYDER, THOMAS YAKALIS

NECIP DOGANAKSOY AND TRAVIS BRODBECK

TING LIU
HAMZA MEMON
LUKE MCKENNA

Year 1: Relationship Building (and growing pains)

Once all the paperwork was out of the way, the first year was spent getting to know each other and the dataset.

- Throwing ideas around
- Throwing ideas away
- Asking way more questions than could possible have answers for
- Getting to know each other and where our specific needs and skills fit into this partnership

First and biggest lesson: Stop “justing” people!

WHY DON’T THEY JUST….

IF THEY WOULD ONLY….

DON’T THEY REALIZE?????!!!!???

Researchers want to know:

Why don't users "just"…

fix the data?

do real-time data entry?

ask the clients more questions?

Community Members want to know:

Why can't the HMIS team "just"…

pull this report "real quick"

give me a quick answer to a complex question?

get HUD to change the regs? Programming? EVERYTHING?

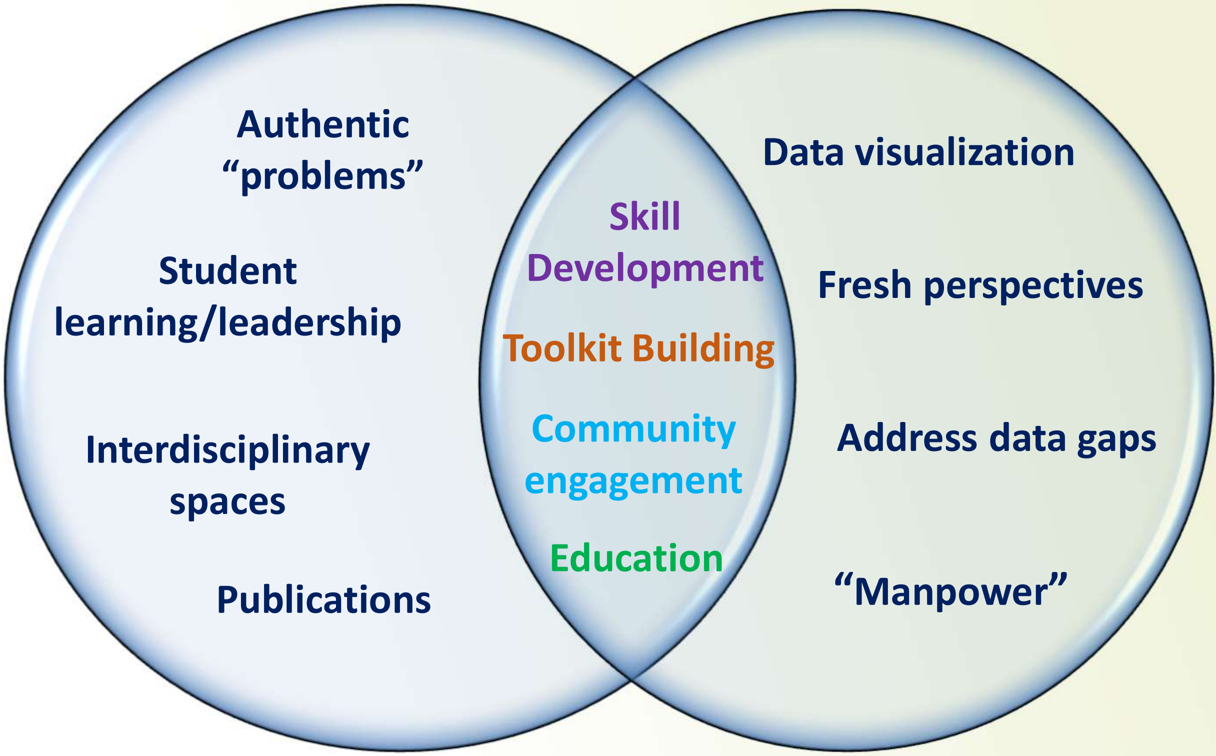
And, of course, the classic:

Why does the HMIS team even have an opinion? They're "JUST" data people…

Takeaway: collaboration can not begin until the "Justs" are out of the way.

Opportunities for authentic partnership

Academic Needs: Siena



Authentic “problems”

Student learning/leadership

Interdisciplinary spaces

Publications

Skill Development

Toolkit Building

Community engagement

Education

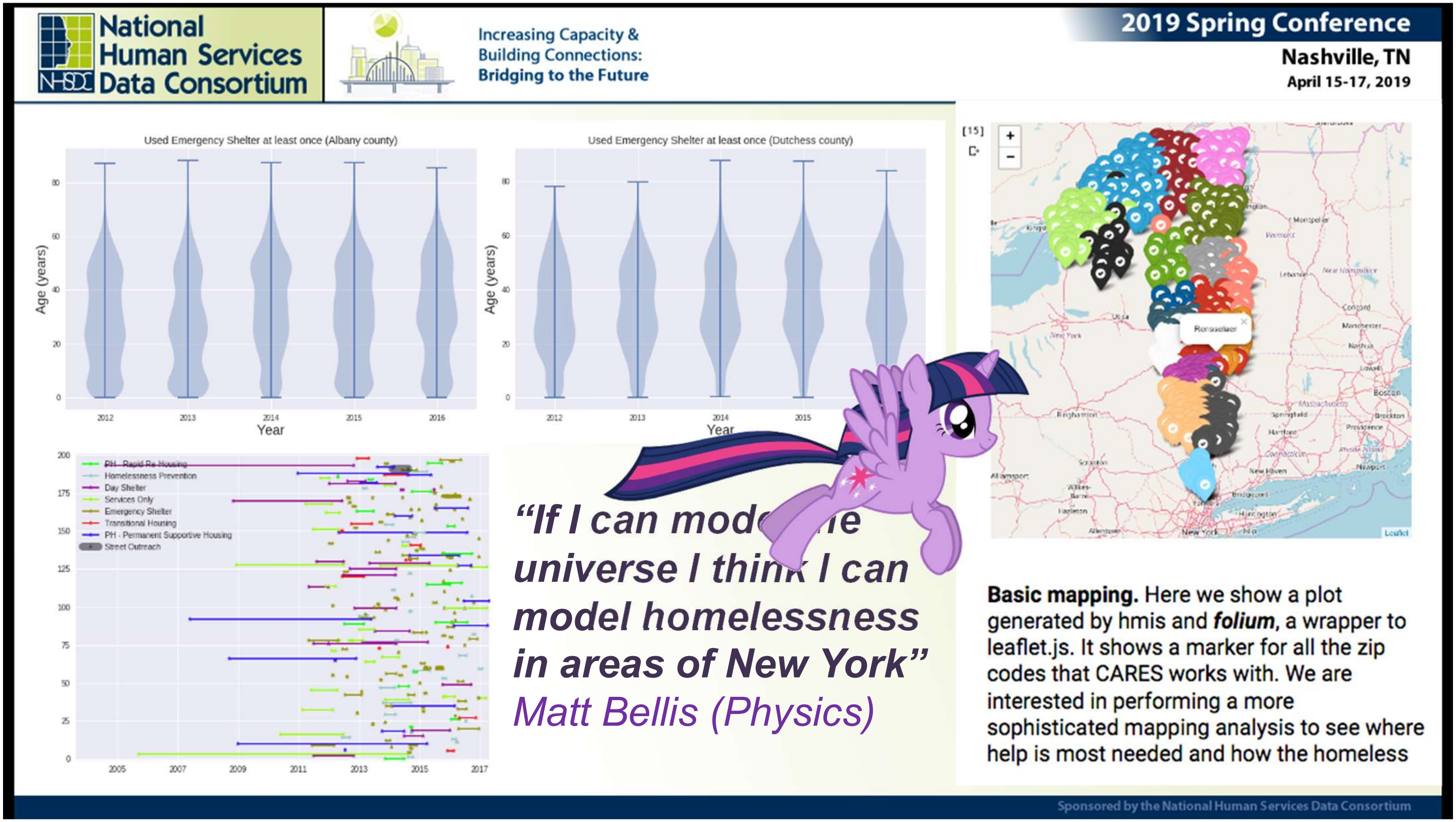
Data visualization

Fresh perspectives

Address data gaps

“Manpower”

HMIS Admin Needs: CARES of NY



National Human Services Data Consortium

Increasing Capacity & Building Connections: Bridging to the Future

2019 Spring Conference

Nashville, TN

April 15-17, 2019

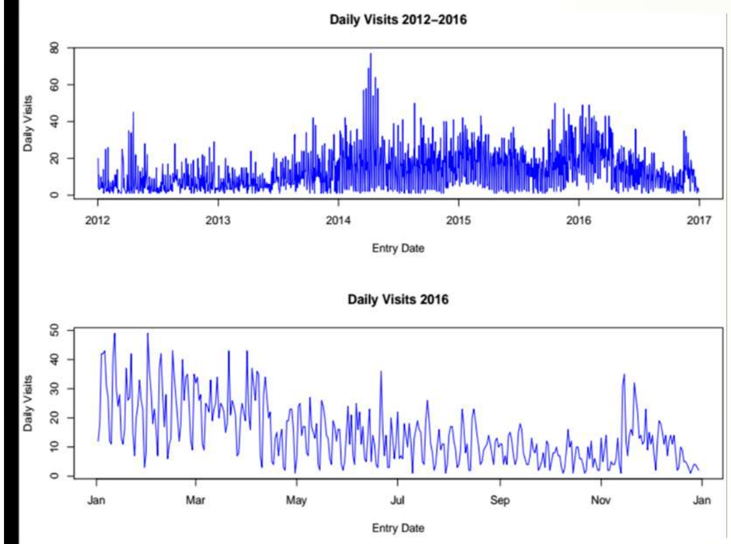
"If I can model the universe I think I can model homelessness in areas of New York"

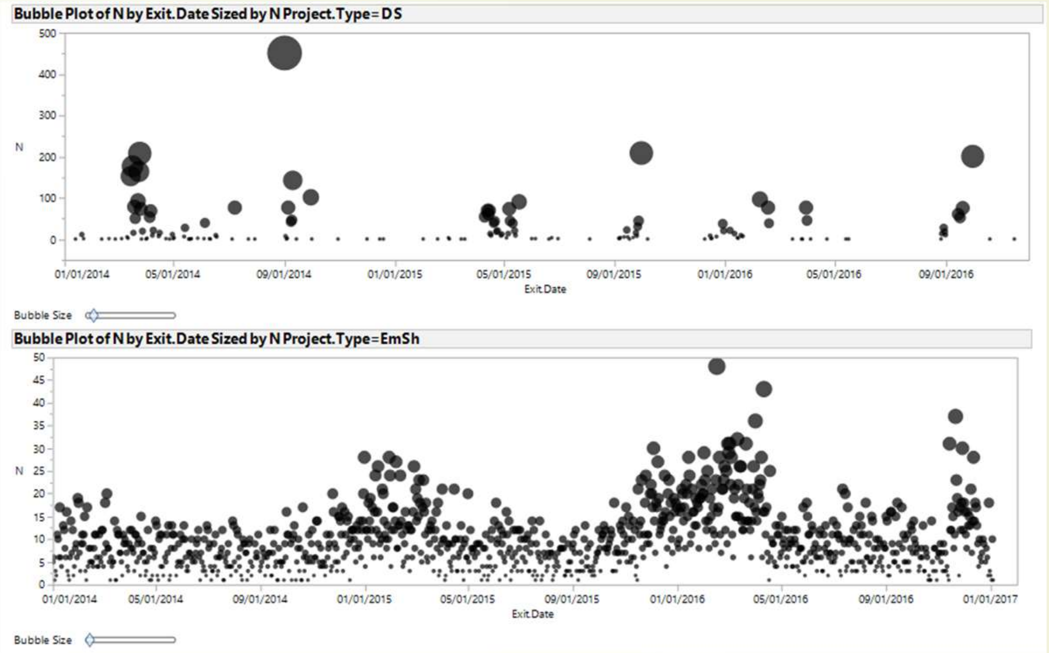
Matt Bellis (Physics)

Basic mapping. Here we show a plot generated by hmis and ***folium***, a wrapper to leaflet.js. It shows a marker for all the zip codes that CARES works with. We are interested in performing a more sophisticated mapping analysis to see where help is most needed and how the homeless

Sponsored by the National Human Services Data Consortium

EDA…Data Cleaning



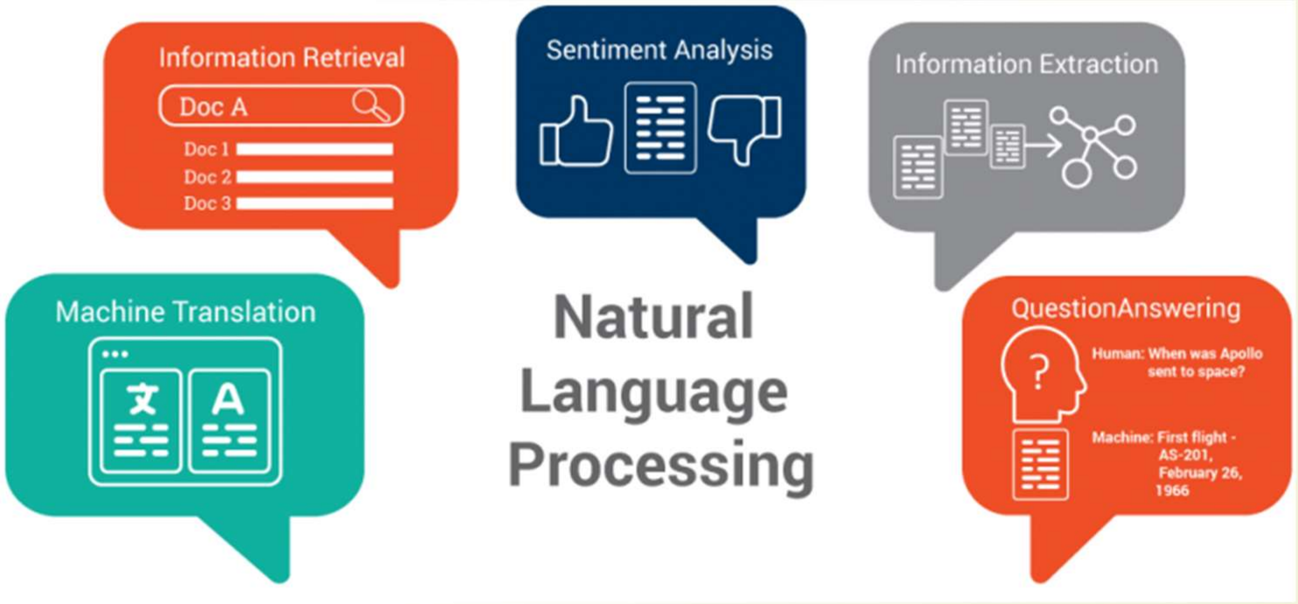


"This is a lot like problem solving a mechanical issue at GE; a lot of messy data that we need to organize to solve a problem."

Necip Doganaksoy & Travis Brodbeck (Accounting)

“Can I can do natural language processing work on homelessness?”

Ting Liu (Computer Science)



Deduplication, Anonymization, Thick data

Deduplication & its Impact on Data Quality

Mark Eliseo and Michael Lostritto
Mentor: Dr. Ting Liu

SIENA college
The education of a lifetime

Background Information
Importance of Deduplication

What is the Issue?

- Homelessness in Upstate NY
- Data which holds sensitive information belonging to various clients of homeless shelters
- Sensitive information is important to these clients because they are at a point in their life where such data is necessary to begin to improve their situation.
- Accurate data is necessary to further research into the issue of homelessness and create a solution.
- CARES is a non profit organization which helps support the effort to end homelessness by providing assistance to community homeless services
- Dataset given by CARES contains approximately 80,000 entries, some of which are repeating sets due to human error
- In order to improve the quality of the dataset given and address the overall issue of homelessness, the method of deduplication was introduced.

Method
Deduplication

- Constructed an algorithm in JAVA (algorithm2.java) that was capable of sorting through multiple fields of information accurately
- In order to merge the repeating information in the dataset that involved data quality we:
  1. Checked each entry for their quality
  2. If there were more than two entries with good quality data we tested the frequency of each entry.
  3. If separate entries contained equal frequencies we decided to use the most recent entry that was inputted into the dataset.
- In order to merge data that could not be disputed we used any input that was true
  - The process of deduplication involved the construction and execution of custom methods in JAVA
  - Below in figure 1.1 is an example of a part of one method created to sort out information regarding client veteran status

Figure 1.1

Data
How well did deduplication improve data quality?

- Human errors that were corrected within agorithm2.java include:
  - Instances where families were checking into homeless shelters and some members were given entirely new ID's due to misrecognition. The presence of new ID's creates repeated sets of data within the dataset for each client in the family thus tainting the overall quality of data.
  - Situations where clients report wrong information to shelters intentionally or unintentionally
    - An example could be a client accidentally sharing the wrong Social Security Number (SSN)
  - Conflicting records in which one client enrolled in a shelter but they enroll again, causing a new input of information (repeating fields of data)
  - Information silos are formed because shelters refuse to share their records with one another thus creating multiple records for a single client who entered multiple shelters.
  - Data quality was improved through the deletion of useless redundant information. The quality of data increased through the accuracy of the client information fields as well.
  - The figures below demonstrate how algorithm2.java fulfilled its purpose in deduplicating client information belonging to the data.

Figure 1.2

The example output in Figure 1.2 involves made up information that mimics the type of information fields that are in the actual CARES dataset. This example displays how algorithm2.java successfully merged multiple fields of information regarding the race of clients into one set of information.

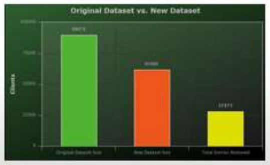
The output shown in figure 1.3 displays information (made to mimic CARES dataset) which contains more sensitive client information. Algorithm2.java successfully sorted through the "good" data and the "bad" data to merge it into one set of accurate information. Figure 1.3 also displays how it used frequency to sort through which data would be best to merge.

Figure 1.3

Results
Data Quality & Dataset Length

- Individual person's information is more complete
- The overall size of the dataset was decreased by 31%
- Repeating information belonging to clients was successfully condensed into a single row of data.
- The accuracy of client information was improved.

Figure 1.4



Conclusion

- This project has been a success in its ability to reduce the amount of human error associated with the way data is handled thus improving data quality.
- It will help support the fight against homelessness in the area through enhancing the accuracy of client information as well as through the enhancement of research on the subject.
- If data quality continues to increase amongst local homeless shelter datasets then:
  - Data quality improvement can be expanded to reach a vast variety of places.
  - The effort to fight homelessness can evolve to support homeless people across the NYS and eventually Nationally.

Acknowledgement
We would like to acknowledge CARES for the service they provide within the community, and we would like to acknowledge CURCA for assistance in furthering research in improving data quality and deduplication.

Build an Online Homeless Database

Dr. Ting Liu[1], Luis Concepcion-Bido[1], Caleb Ryor[1], and Travis Brodbeck[2]
[1]Computer Science, Siena College, Albany, NY [2]Accounting, Siena College, Albany, NY

SIENA*college*
The education of a lifetime

Abstract

Organizations that help for homeless people usually are not willing to share the data because of sensitive personal information. Their data isn't stored the most optimally either. We plan to build a nationwide database that can be shared with organizations/researchers to help find new approaches to the issue of homelessness through proper storage of data and analysis.

Introduction to the Siena Homelessness Project

The main hurdle with the data we were faced with was the sheer amount of it needed to be handled. Everything was being stored in CSV files which provided plenty of function based issues as well as space issues causing Excel to perform slowly. All the personal information given to us was double hashed so we could still track information regarding a given individual. The data itself consists of various site and client data from homeless shelters all across the Albany area all of which is collected and managed by CARES, Inc., our data provider and community partner.

Solution: Database

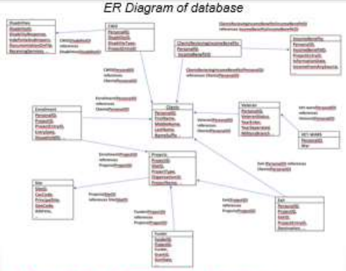
Utilizing a SQL database for the sake of storing all the information provided to us gave us much more freedom as it enabled there to no longer be a limit on space as well as having more flexible options in how we could record this data. The next steps that needed to be made before the database could be complete however involved optimizing where certain data was being stored for even more ease of access.

Example of Data Entries in Table

Building an SQL Homeless Database

As pictured below, the first step to creating said database involved us forming an ER diagram to visualize what was being created. Each of the 9 CSV files provided became its own table with the exception of the Clients table being split to hold Veteran and VeteranStatus information elsewhere for the sake of storage optimization. Important strides were made to create code to deduplicate data within specific tables within the database to make queries provide more accurate analysis as well as just further improving upon the database itself. Fortunately the data itself was sorted well enough for there not to be much more resorting to be done on our part, but the small changes that we made were paramount to the data involved to be analyzed properly at all.

ER Diagram of database



SQL Queries for Statistical Analysis

With the creation of a SQL database we are able to run various queries to parse through the data to find various figures for all kinds of issues. Here are a few examples of what we are able to currently do:

- Compare Veteran and Clients tables to find percentage of veterans
- Sort IDs within database to find percentage of youth, adults, and elderly within the system
- Find percentage of those with a given recorded disability

More statistical analysis is planned but much of what we can do is reliant on optimizing our tables for less complicated queries as well as continuing further with cleaning and deduplicating the data we store.

Develop a Web Interface

Goals for web interface development

- Help visitors understand issues in homelessness data through short articles
- Serve as hub connecting organizations and researchers to share data
- Generate and showcase statistics from current datasets using SQL queries
- Be a center for Siena College data analysis tools with examples to bring in more clients

Screenshot of current frontpage

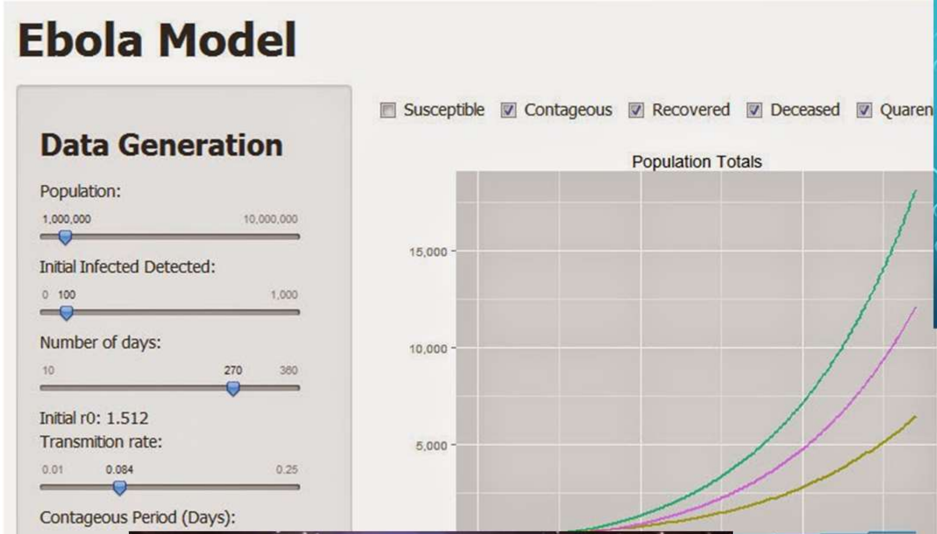
Future Work

- Generate more code for data deduplication
- Work with more companies to get them involved in the project for more datasets
- Talk with other parts of project on campus to showcase/host more tools on webpage

Acknowledgements

Research funded by the Center for Undergraduate Research and Creative Activities (CURCA) at Siena College. Data provided by CARES of NY, Inc., a non-profit organization that provides administrative support to community homeless services, working to create systems of care that end and prevent homelessness (caresny.org). Data is collected by the Homeless Management Information System (HMIS), a locally admininistered data collection and reporting software (hudexchange.info/programs/hmis/)

Stochastic Optimization



Ebola Model

Data Generation

Population: 1,000,000 10,000,000

Initial Infected Detected: 0 100 1,000

Number of days: 10 270 360

Initial r0: 1.512

Transmition rate: 0.01 0.084 0.25

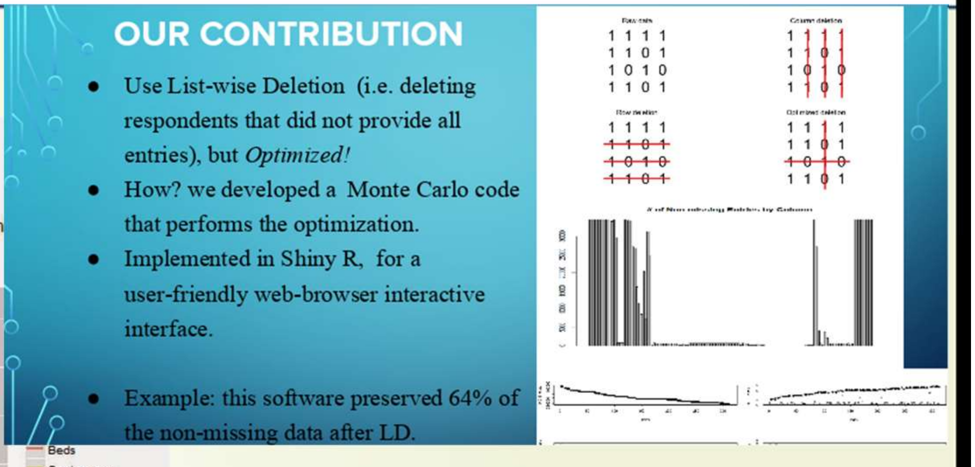
Contageous Period (Days): 0

Moralitity Ra 0

Social 'adapt

□ Susceptible ☑ Contageous ☑ Recovered ☑ Deceased ☑ Quaren

Beds
Contageous
Deceased
Quarentined
Recovered



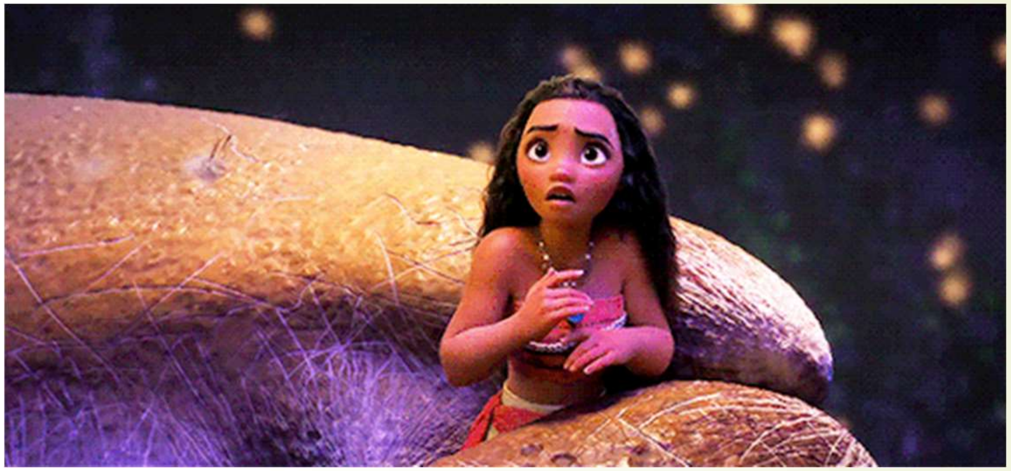
OUR CONTRIBUTION

- Use List-wise Deletion (i.e. deleting respondents that did not provide all entries), but *Optimized!*
- How? we developed a Monte Carlo code that performs the optimization.
- Implemented in Shiny R, for a user-friendly web-browser interactive interface.
- Example: this software preserved 64% of the non-missing data after LD.

I have been looking for a place to test out new tools for optimizing incomplete data

Graziano Vernizzi (Physics)

But…whenever you want to use software, there's a cost……



Siena:

- Creates community among faculty
- Re-energizes faculty and students
- Elevates the reputation for community engagement
- Helps get community engagement classifications

CARES of NY :

- Identifies and addresses gaps in data quality
- Research opportunities using HMIS data
- Excitement about USING data
- Educating on true state of homelessness
- Connects community organizations with students

Year 2: Identity Building

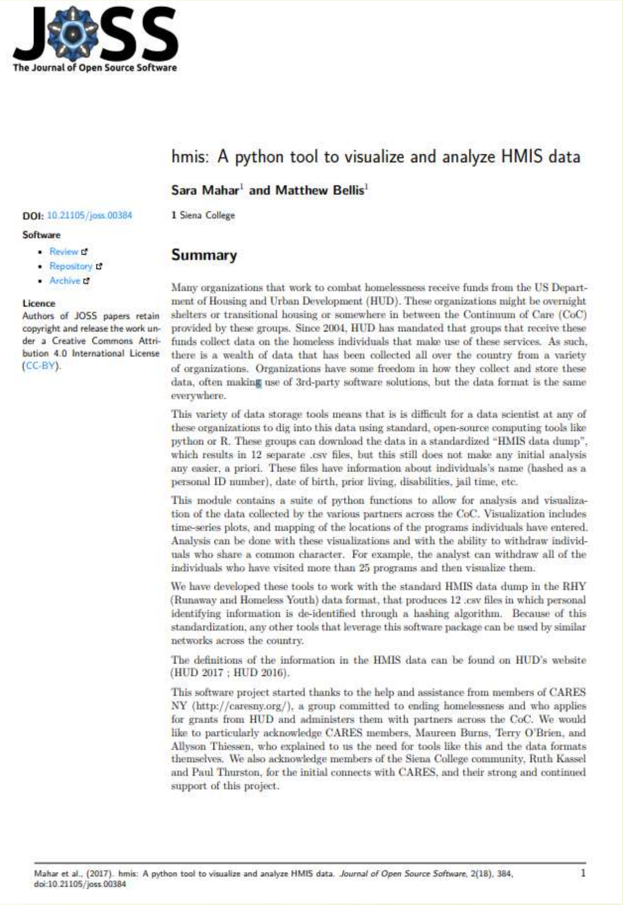
Accomplishments

- Clear delineation of roles
- Better understanding of commitment
- First publications
- Data for Good Exchange poster
- Redefining wants and expectations

Our first publication!

Journal of Open Source Software (JOSS) article on the HMIS Visualization suite developed in Year 1.

JOSS 10.21105/joss.00384



JOSS
The Journal of Open Source Software

hmis: A python tool to visualize and analyze HMIS data

Sara Mahar[1] **and Matthew Bellis**[1]

1 Siena College

DOI: 10.21105/joss.00384

Software

- Review
- Repository
- Archive

Summary

Many organizations that work to combat homelessness receive funds from the US Department of Housing and Urban Development (HUD). These organizations might be overnight shelters or transitional housing or somewhere in between the Continuum of Care (CoC) provided by these groups. Since 2004, HUD has mandated that groups that receive these funds collect data on the homeless individuals that make use of these services. As such, there is a wealth of data that has been collected all over the country from a variety of organizations. Organizations have some freedom in how they collect and store these data, often making use of 3rd-party software solutions, but the data format is the same everywhere.

This variety of data storage tools means that is is difficult for a data scientist at any of these organizations to dig into this data using standard, open-source computing tools like python or R. These groups can download the data in a standardized "HMIS data dump", which results in 12 separate .csv files, but this still does not make any initial analysis any easier, a priori. These files have information about individuals's name (hashed as a personal ID number), date of birth, prior living, disabilities, jail time, etc.

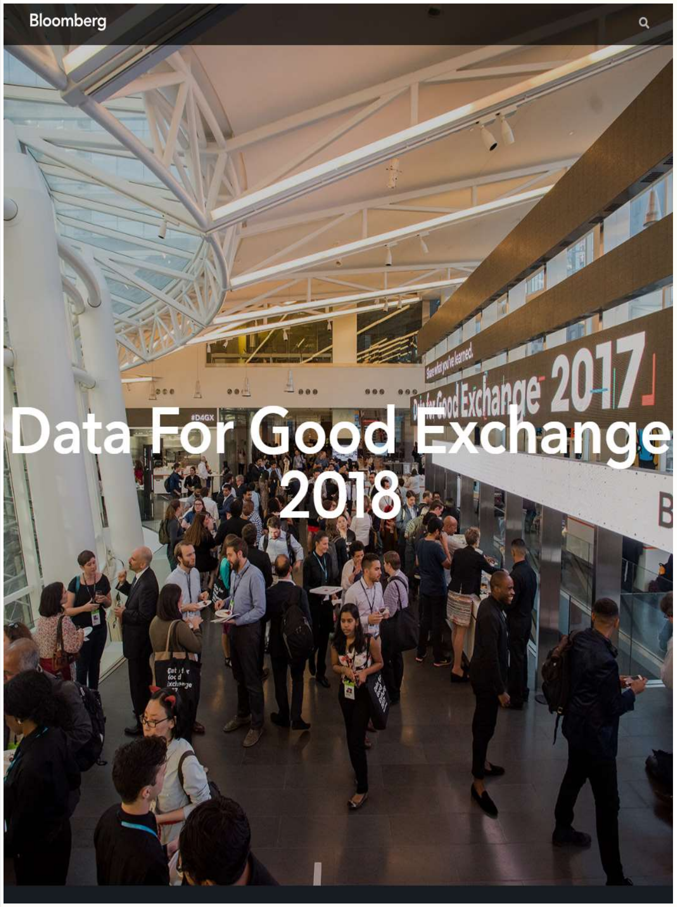
This module contains a suite of python functions to allow for analysis and visualization of the data collected by the various partners across the CoC. Visualization includes time-series plots, and mapping of the locations of the programs individuals have entered. Analysis can be done with these visualizations and with the ability to withdraw individuals who share a common character. For example, the analyst can withdraw all of the individuals who have visited more than 25 programs and then visualize them.

We have developed these tools to work with the standard HMIS data dump in the RHY (Runaway and Homeless Youth) data format, that produces 12 .csv files in which personal identifying information is de-identified through a hashing algorithm. Because of this standardization, any other tools that leverage this software package can be used by similar networks across the country.

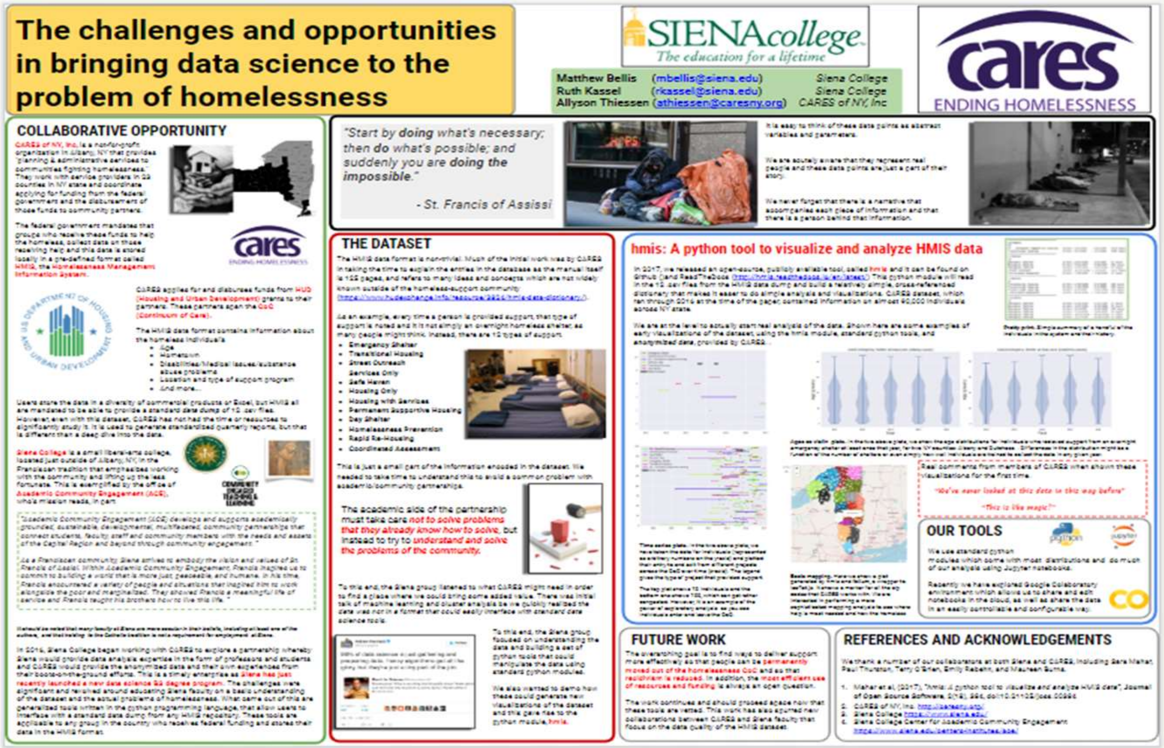
The definitions of the information in the HMIS data can be found on HUD's website (HUD 2017 ; HUD 2016).

This software project started thanks to the help and assistance from members of CARES NY (http://caresny.org/), a group committed to ending homelessness and who applies for grants from HUD and administers them with partners across the CoC. We would like to particularly acknowledge CARES members, Maureen Burns, Terry O'Brien, and Allyson Thiessen, who explained to us the need for tools like this and the data formats themselves. We also acknowledge members of the Siena College community, Ruth Kassel and Paul Thurston, for the initial connects with CARES, and their strong and continued support of this project.

Mahar et al., (2017). hmis: A python tool to visualize and analyze HMIS data. *Journal of Open Source Software*, 2(18), 384, doi:10.21105/joss.00384 1



2nd Publication: Bloomberg Poster



The challenges and opportunities in bringing data science to the problem of homelessness

Matthew Bellis (mbellis@siena.edu) Siena College
Ruth Kassel (rkassel@siena.edu) Siena College
Allyson Thiessen (athiessen@caresny.org) CARES of NY, Inc

COLLABORATIVE OPPORTUNITY

CARES of NY, Inc. is a not-for-profit organization in Albany, NY that provides "planning & administrative services to communities fighting homelessness." They work with service providers in 23 counties in NY state and coordinate applying for funding from the federal government and the disbursement of those funds to community partners.

The federal government mandates that groups who receive these funds to help the homeless, collect data on those receiving help and this data is stored locally in a pre-defined format called HMIS, the Homelessness Management Information System.

CARES applies for and disburses funds from HUD (Housing and Urban Development) grants to their partners. These partners span the CoC (Continuum of Care).

The HMIS data format contains information about the homeless individual's

- Age
- Hometown
- Disabilities/Medical issues/substance abuse problems
- Location and type of support program
- And more...

Users store the data in a diversity of commercial products or Excel, but HMIS all are mandated to be able to provide a standard data dump of 12 .csv files. However, even with this dataset, CARES has not had the time or resources to significantly study it. It is used to generate standardized quarterly reports, but that is different than a deep dive into the data.

Siena College is a small liberal-arts college, located just outside of Albany, NY, in the Franciscan tradition that emphasizes working with the community and lifting up the less fortunate. This is exemplified by the office of Academic Community Engagement (ACE), who's mission reads, in part

> "Academic Community Engagement (ACE) develops and supports academically grounded, sustainable, developmental, multifaceted, community partnerships that connect students, faculty, staff and community members with the needs and assets of the Capital Region and beyond through community engagement."
>
> As a Franciscan community, Siena strives to embody the vision and values of St. Francis of Assisi. Within Academic Community Engagement, Francis inspires us to commit to building a world that is more just, peaceable, and humane. In his time, Francis encountered a variety of people and situations that inspired him to work alongside the poor and marginalized. They showed Francis a meaningful life of service and Francis taught his brothers how to live this life."

It should be noted that many faculty at Siena are more secular in their beliefs, including at least one of the authors, and that training in the Catholic tradition is not a requirement for employment at Siena.

In 2016, Siena College began working with CARES to explore a partnership whereby Siena would provide data analysis expertise in the form of professors and students and CARES would provide the anonymized data and their own experiences from their boots-on-the-ground efforts. This is a timely enterprise as Siena has just recently launched a new data science BS degree program. The challenges were significant and revolved around educating Siena faculty on a basic understanding of the dataset and the actual problems of homelessness. What came out of this are generalized tools written in the python programming language, that allow users to interface with a standard data dump from any HMIS repository. These tools are applicable to any group in the country who receives federal funding and stores their data in the HMIS format.

> "Start by doing what's necessary; then do what's possible; and suddenly you are doing the impossible."
>
> - St. Francis of Assissi

It is easy to think of these data points as abstract variables and parameters.

We are acutely aware that they represent real people and these data points are just a part of their story.

We never forget that there is a narrative that accompanies each piece of information and that there is a person behind that information.

THE DATASET

The HMIS data format is non-trivial. Much of the initial work was by CARES in taking the time to explain the entries in the database as the manual itself is 122 pages, and refers to many ideas and concepts which are not widely known outside of the homeless-support community (https://www.hudexchange.info/resource/3824/hmis-data-dictionary/).

As an example, every time a person is provided support, that type of support is noted and it is not simply an overnight homeless shelter, as many people might think. Instead, there are 12 types of support:

- Emergency Shelter
- Transitional Housing
- Street Outreach
  Services Only
- Safe Haven
- Housing Only
- Housing with Services
- Permanent Supportive Housing
- Day Shelter
- Homelessness Prevention
- Rapid Re-Housing
- Coordinated Assessment

This is just a small part of the information encoded in the dataset. We needed to take time to understand this to avoid a common problem with academic/community partnerships.

The academic side of the partnership must take care *not to solve problems that they already know how to solve*, but instead to try to *understand and solve the problems of the community.*

To this end, the Siena group listened to what CARES might need in order to find a place where we could bring some added value. There was initial talk of machine learning and cluster analysis but we quickly realized the data was not in a format that could easily interface with standard data science tools.

To this end, the Siena group focused on understanding the data and building a set of python tools that could manipulate the data using standard python modules.

We also wanted to demo how these could generate new visualizations of the dataset and this gave rise to the python module, hmis.

hmis: A python tool to visualize and analyze HMIS data

In 2017, we released an open-source, publicly available tool, called hmis and it can be found on Github (and ReadTheDocs (http://hmis.readthedocs.io/en/latest/). This python module will read in the 12 .csv files from the HMIS data dump and build a relatively simple, cross-referenced dictionary that makes it easier to do simple analysis and visualizations. CARES dataset, which ran through 2016 at the time of the paper, contained information on almost 90,000 individuals across NY state.

We are at the level to actually start real analysis of the data. Shown here are some examples of early visualizations of the dataset, using the hmis module, standard python tools, and anonymized data, provided by CARES.

Pretty print. Simple summary of a handful of the individuals in the system and their history.

Time series plots. In the time series plots, we have taken the data for individuals (represented as arbitrary numbers on the y-axis) and plotted their entry to and exit from different projects across the CoC over time (x-axis). The legend gives the type of project that provided support. [illegible]

Ages as violin plots. In the two above plots, we show the age distributions for individuals who received support from an example emergency shelter at least once that year, for two NY counties: [illegible]

Basic mapping. Here we show a plot generated by hmis and folium, a [illegible] leaflet.js. [illegible]

Real comments from members of CARES when shown these visualizations for the first time.

"We've never looked at this data in this way before"

"This is like magic!"

OUR TOOLS

We use standard python modules which come with most distributions and do much of our analysis using Jupyter notebooks.

Recently we have explored Google Colaboratory environment which allows us to share and edit notebooks in the cloud, as well as share the data in an easily controllable and configurable way.

FUTURE WORK

The overarching goal is to find ways to deliver support more effectively so that people can be permanently moved out of the homelessness CoC and so that recidivism is reduced. In addition, the most efficient use of resources and funding is always an open question.

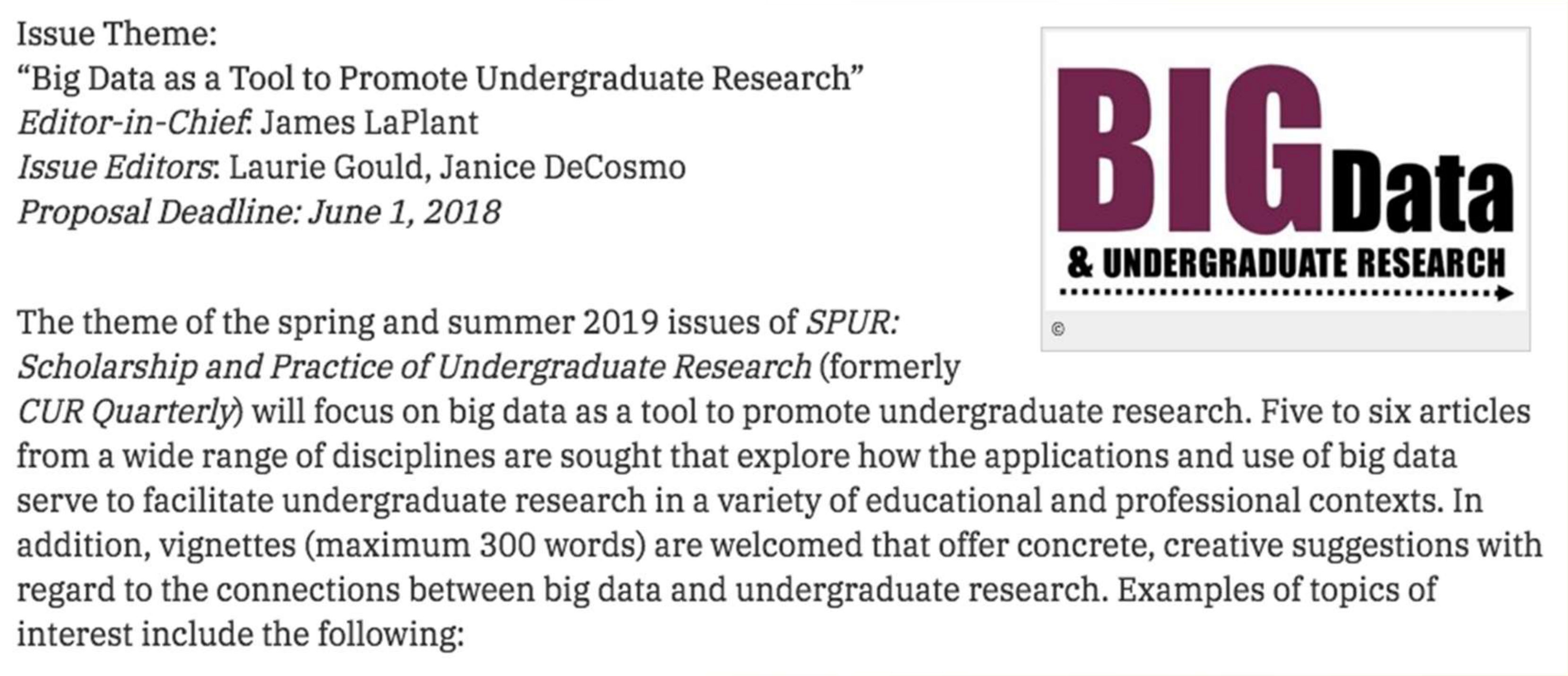
The work continues and should proceed apace now that these tools are vetted. This work has also spurred new collaborations between CARES and Siena faculty that focus on the data quality of the HMIS dataset.

REFERENCES AND ACKNOWLEDGEMENTS

We thank a number of our collaborators at both Siena and CARES, including Sara Maher, Paul Thurston, Terry O'Brien, Emily Rabenn, and Maureen Burns.

1. Maher et al, (2017), "hmis: A python tool to visualize and analyze HMIS data", Journal of Open Source Software, 2(18), 384, doi:10.21105/joss.00384
2. CARES of NY, Inc. http://caresny.org/
3. Siena College https://www.siena.edu/
4. Siena College Center for Academic Community Engagement https://www.siena.edu/centers-institutes/ace/

3rd Publication – Council of Undergrad Research Journal



Issue Theme:
"Big Data as a Tool to Promote Undergraduate Research"
Editor-in-Chief: James LaPlant
Issue Editors: Laurie Gould, Janice DeCosmo
Proposal Deadline: June 1, 2018

The theme of the spring and summer 2019 issues of *SPUR: Scholarship and Practice of Undergraduate Research* (formerly *CUR Quarterly*) will focus on big data as a tool to promote undergraduate research. Five to six articles from a wide range of disciplines are sought that explore how the applications and use of big data serve to facilitate undergraduate research in a variety of educational and professional contexts. In addition, vignettes (maximum 300 words) are welcomed that offer concrete, creative suggestions with regard to the connections between big data and undergraduate research. Examples of topics of interest include the following:

Year 3: Sustainability

- Pilot Project
- Economic Research
- Case Notes Analysis
- Deep Dive for DSS

How to SPIn your own partnership

Understand who the connectors are:

- Undergrad research center
- Community engagement center
- Outreach and volunteer center
- Don’t limit yourself to the social sciences

Be prepared to:

- Give a sample of the data (to evaluate: structure, size, messiness)
- Answer ALL THE QUESTIONS
- Get excited! It's contagious.
- Accommodate academic timelines
- Be flexible in what you want/need/expect
- Give as much as you get: time, energy, and enthusiasm
- Let this be student led; you'll be AMAZED at what you get (good and bad!)
- Work on projects that are low urgency but high importance

Do you believe in unicorns?