SDSS SYMPOSIUM ON DATA SCIENCE & STATISTICS
BEYOND BIG DATA: COLLABORATION IN SCIENCE, INDUSTRY, AND SOCIETY
VIRTUAL • JUNE 3–5, 2020

OPEN DATA VISUALIZATIONS AND ANALYTICS AS TOOLS FOR POLICY-MAKING

Loni Hagen, Ph.D.
Assistant Professor, School of Information
University of South Florida

USF UNIVERSITY OF SOUTH FLORIDA

MY BACKGROUND

- Constitution
- Law enforcement

Data life cycle

privacy and ethical concerns throughout

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.e26845b4

RELEVANT PUBLICATIONS

- Hagen, L. (2018). **Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?** *Information Processing & Management*, *54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006
- Hagen, L., Keller, T. E., Yerden, X., & Luna-Reyes, L. F. (2019). **Open data visualizations and analytics as tools for policy-making**. *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2019.06.004

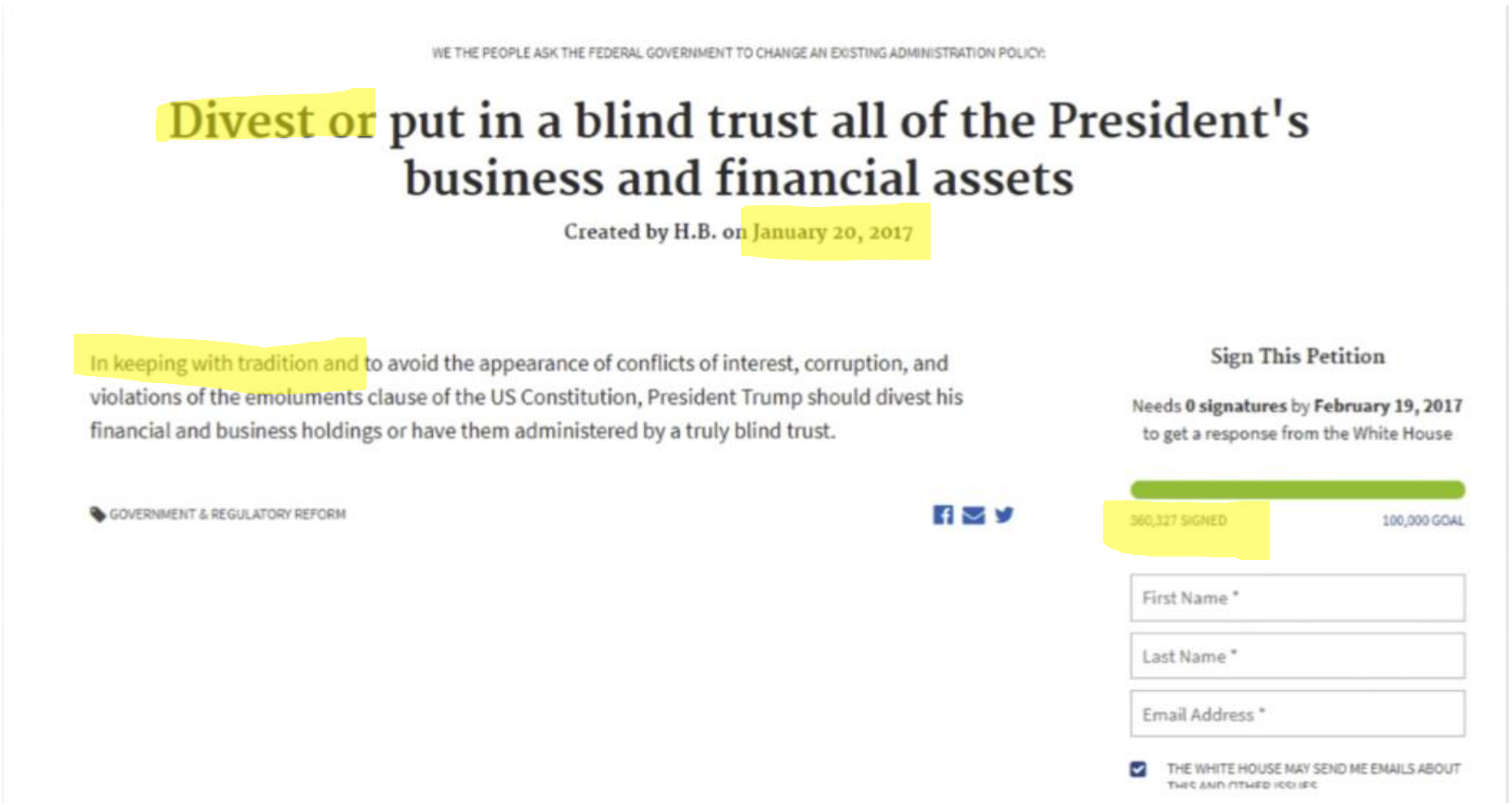
A GENERAL QUESTION:

- How to address barriers in adopting data science for policy decision making or for government organizations?

TWO TASKS

- 1. Topic modeling and visualization
- 2. Usability assessment for policy decision making

DATA AND METHODS



WE THE PEOPLE ASK THE FEDERAL GOVERNMENT TO CHANGE AN EXISTING ADMINISTRATION POLICY:

Divest or put in a blind trust all of the President's business and financial assets

Created by H.B. on January 20, 2017

In keeping with tradition and to avoid the appearance of conflicts of interest, corruption, and violations of the emoluments clause of the US Constitution, President Trump should divest his financial and business holdings or have them administered by a truly blind trust.

GOVERNMENT & REGULATORY REFORM

Sign This Petition

Needs **0 signatures** by **February 19, 2017** to get a response from the White House

360,327 SIGNED 100,000 GOAL

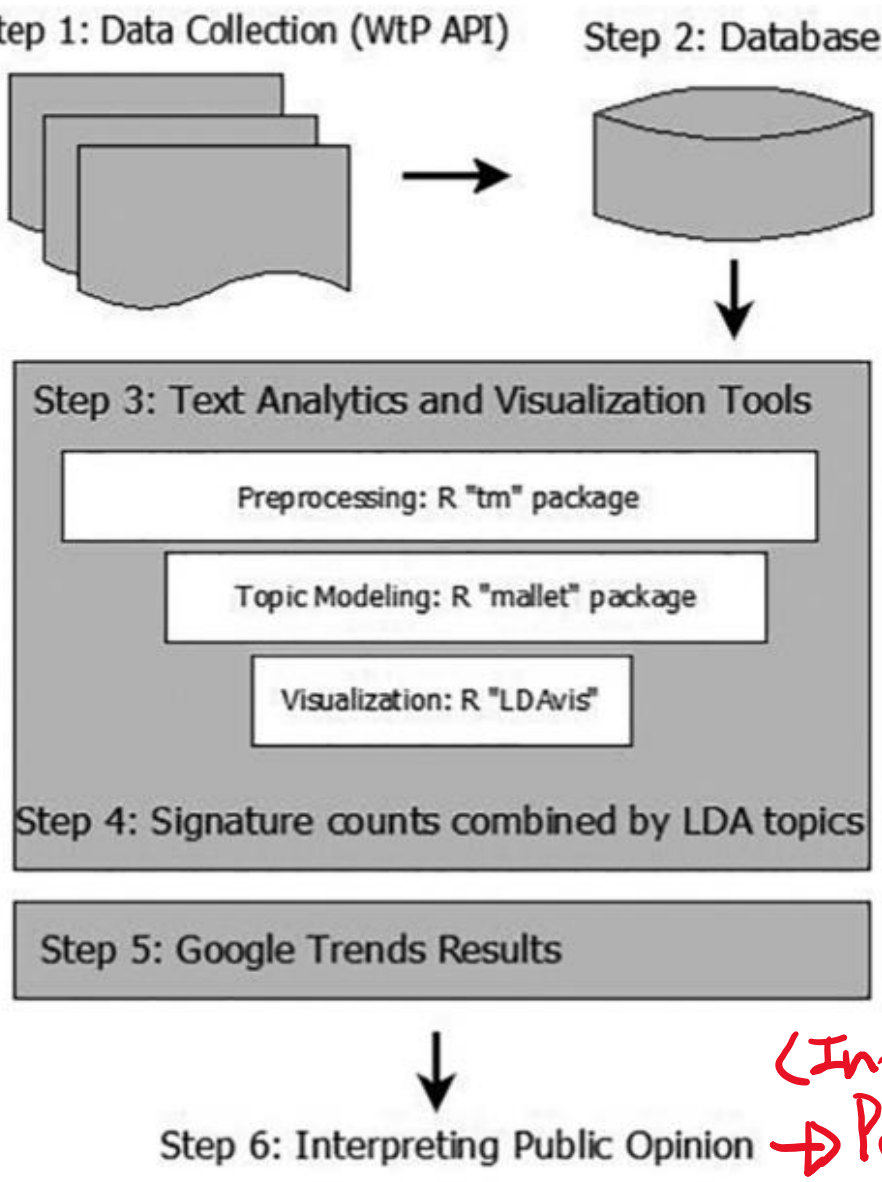
First Name *

Last Name *

Email Address *

THE WHITE HOUSE MAY SEND ME EMAILS ABOUT

ID	Date_created	Title	Body	SignatureCount	Issues
1301	9/22/2011	Stop Animal Homelessness at Its Roots	Every year in the United States, an estimated 6 to 8 million lost, abandoned, or unwanted dogs and cats enter animal shelters and nearly half of these animalsxmany of them healthy, young, and adoptablexmust be euthanized because there are too many animals and not enough good homes. This tragedy occurs because people don't spay and neuter their animals and because greedy breeders continue to churn out more puppies. Because all dogs and cats are precious and because no more animals need to be bred when so many others go without hope of being adopted, PETA is calling for a mandatory spay-and-neuter law until all dogs and cats in the United States have a home to call their own. Sign the petition calling for a mandatory spay-and-neuter law to help end the animal overpopulation crisis.	11,786	Energy & Environment
1326	9/22/2011	Grant voters the ability to vote for the President of the United States by dissolving the electoral college.	The elections of 1824, 1876, 1888, and 2000 produced an Electoral College winner who did not receive the plurality of the nationwide popular vote - that is, the American people did not get the President democracy should have selected. Due to the way electoral votes are allocated, candidates have a strong incentive to focus their campaigns on "swing" states with many voters, such as Florida, and neglect states such as Texas which do not swing. Due to necessary rounding errors when allocating votes, members of a sparsely populated state effectively accrue more voting power than members of a well-populated state. This increases the electoral power of members of certain states while reducing it for others on an ongoing basis. We beg our leaders to dissolve this system and let us vote.	29,311	c("Civil Rights & Equality", "Government & Regulatory Reform")



<Interview> → Policy decision makers

Fig. 2. Framework of the visual analytics of topic modeling.

mallet

LDAvis

TOPIC MODELING AND VISUALIZATION

TOPIC MODELING OUTPUTS

Topic ID	Label	Topic words
4	Cancer Disease**	health care cancer disease research medical treatment patients
5	Election Clinton**	vote investigation election clinton investigate people federal party
6	Prison Sentence**	justice years prison case life trial court release
7	Terrorism Syria**	war terrorist people stop government terrorism genocide syria
8	Guns Firearms**	law amendment gun rights states laws ban weapons
9	Children Gender	children child women sex law sexual parents rights
10	Religion*	rights government religious human freedom god religion church
11	National Holiday*	day national american house holiday white awareness world
12	Water Park Energy**	water national energy park land oil areas gas
13	Police & *BLM* **	police law officers violence enforcement officer black death
14	Internet Companies*	internet service information access companies small business government
15	Students School Education**	students school education schools student public children college
16	Ukraine Russia*	ukraine russian russia puerto sanctions japan ukrainian rico
17	Visa Immigration**	visa immigration united states status family green home
18	Military Veterans**	military service members veterans soldiers war army forces
19	White Anti Genocide**	white anti genocide countries whites racist word code
20	Http & China*	http www org chinese people human china world
21	Animal*	animals animal dogs wild hong dog kong horses
22	Secession*	states united government state america people powers nature
23	Vehicle & FAA**	vehicles safety vehicle faa aircraft air cars flight
24	Medal Award*	medal honor freedom award presidential game team american
25	Food Labeling**	food fda products foods health safe labeling ban
26	Marijuana**	marijuana drug cannabis medical schedule hemp states substances
27	Ebola & TPP*	ebola trans media trump trade partnership people protect
28	FDA & Blood	fda blood life india drug sri sikhs drugs
29	Mcllellan	mcclellan act iran veterans toxic nuclear congress health
30	Charly Wingate	charly robbery pardon vietnam max retrial wingate circumcision

TOPIC MODELING OUTPUTS

ntext	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
1	0.000507	0.000535	0.000568	0.000929	0.000822	0.000688	0.002259	0.00066	0.000512	0.000773	0.000517	0.000328
2	0.000486	0.000513	0.000544	0.000891	0.068791	0.43588	0.124573	0.000633	0.000491	0.000742	0.000496	0.000315
3	0.00045	0.000474	0.000503	0.000824	0.642029	0.00061	0.064876	0.000586	0.000454	0.000686	0.000458	0.000291
4	0.0005	0.000527	0.00056	0.000916	0.00081	0.02864	0.002228	0.000651	0.000505	0.000762	0.000509	0.000324
5	0.017113	0.000623	0.000661	0.001083	0.000957	0.000802	0.06872	0.595556	0.000597	0.066988	0.000602	0.000383
6	0.00045	0.000474	0.000503	0.000824	0.000729	0.00061	0.039727	0.000586	0.000454	0.038409	0.000458	0.000291
7	0.000668	0.000705	0.000748	0.001225	0.001083	0.000906	0.002977	0.00087	0.000675	0.001019	0.000681	0.000433
8	0.0342	0.38702	0.000672	0.001101	0.017773	0.000815	0.053075	0.000782	0.000607	0.084913	0.000612	0.000389
9	0.000737	0.433535	0.000825	0.001351	0.001194	0.001	0.106321	0.00096	0.000744	0.124769	0.000751	0.000477
10	0.000514	0.000542	0.000576	0.000943	0.058366	0.000698	0.635149	0.00067	0.00052	0.000784	0.000524	0.000333
11	0.000439	0.000463	0.061821	0.000804	0.025243	0.000595	0.001954	0.000571	0.000443	0.000669	0.000447	0.000284
12	0.000486	0.000513	0.000544	0.476914	0.000788	0.00066	0.002167	0.000633	0.000491	0.109547	0.041297	0.000315
13	0.000486	0.000513	0.000544	0.109697	0.000788	0.00066	0.002167	0.000633	0.000491	0.000742	0.408515	0.000315
14	0.000694	0.000732	0.000777	0.001272	0.001125	0.000942	0.003092	0.000904	0.000701	0.001058	0.000707	0.000449
15	0.000514	0.000542	0.000576	0.000943	0.000833	0.000698	0.002292	0.561612	0.00052	0.000784	0.000524	0.000333
16	0.000694	0.000732	0.000777	0.001272	0.234017	0.000942	0.003092	0.000904	0.000701	0.001058	0.000707	0.000449
17	0.000581	0.000613	0.000651	0.001065	0.000942	0.000789	0.00259	0.000757	0.000587	0.065899	0.016846	0.000376
18	0.000493	0.386589	0.000552	0.000904	0.000799	0.000669	0.112502	0.000642	0.014286	0.000752	0.000502	0.000319
19	0.000507	0.142327	0.000568	0.100184	0.000822	0.000688	0.144051	0.00066	0.000512	0.000773	0.000517	0.000328
20	0.00048	0.000506	0.000537	0.000879	0.000778	0.000651	0.699881	0.000625	0.000485	0.000732	0.000489	0.000311
21	0.000486	0.000513	0.000544	0.000891	0.000788	0.00066	0.233378	0.000633	0.000491	0.000742	0.000496	0.000315

TOPIC VISUALIZATION

- http://localhost:13454/session/file486053111c83/index.html#topic=18&lambda=1&term=
- Github: https://github.com/lonihagen/Topic-Modeling

Changes of number of signatures per topic by time

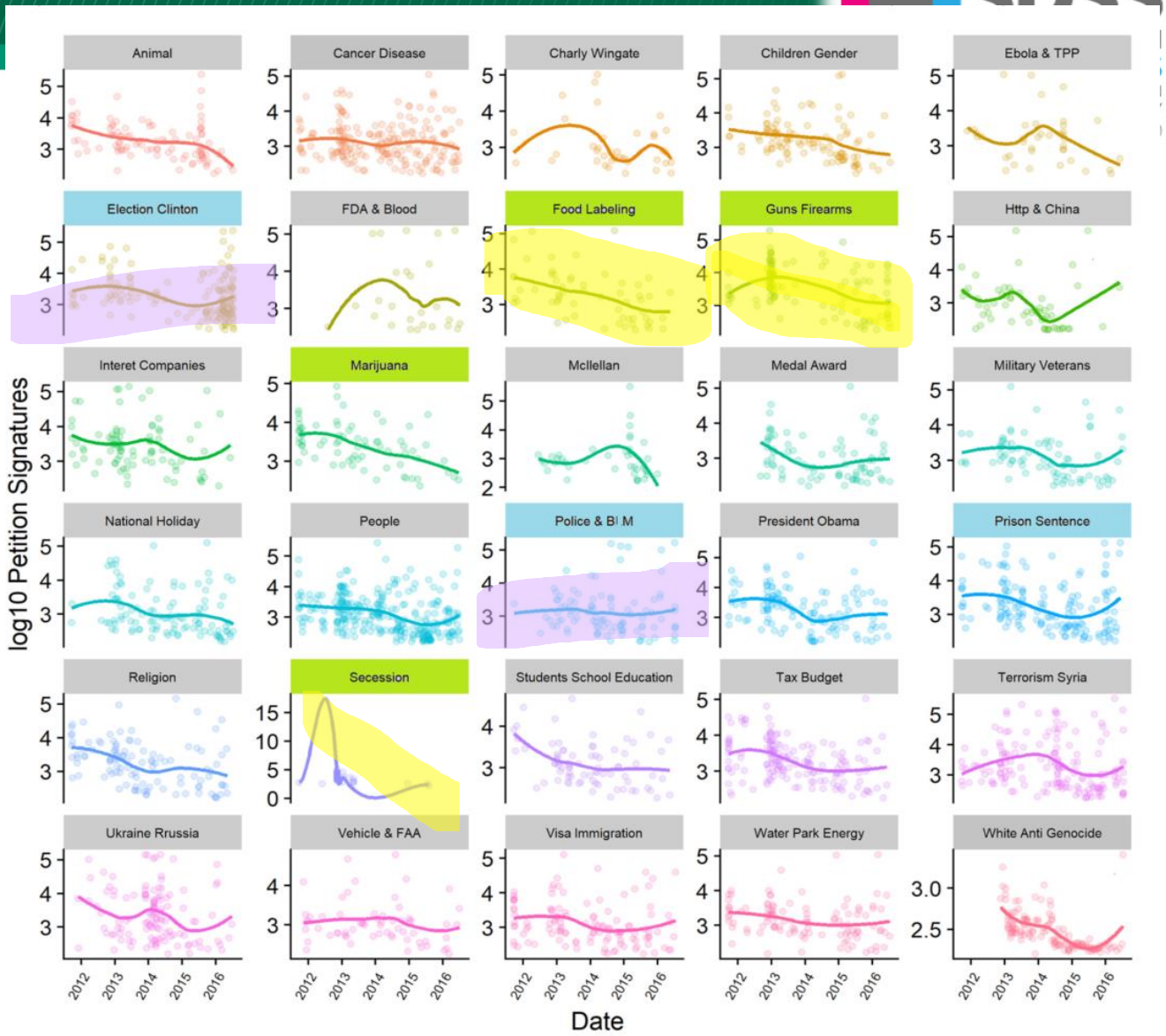


Fig. 6. Changes of number of signatures per topic by time.

Google trends and topic signature counts

platform specific impact

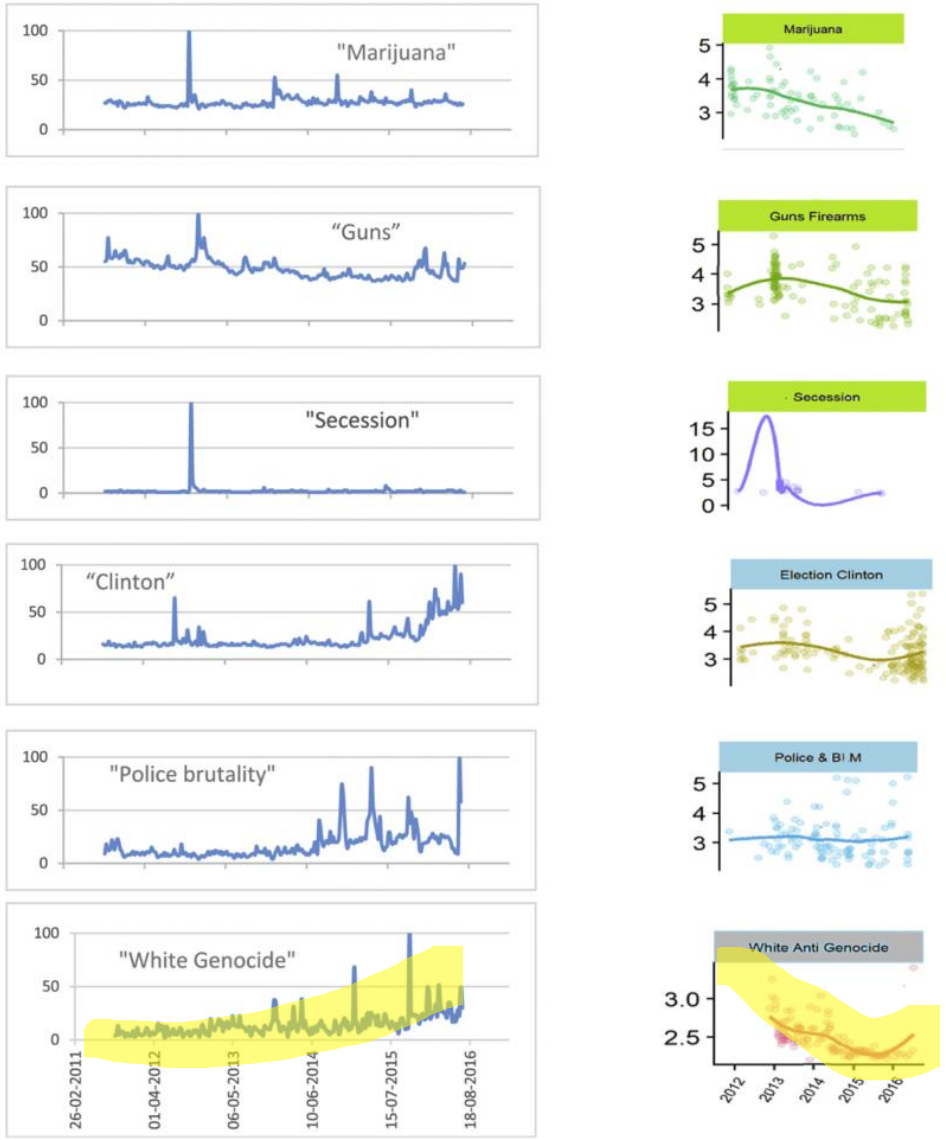


Fig. 7. Google trends and topic popularity.

INTERVIEWS

- Relevance of the topics
- Interpretability
- Learnability
- Utility
- Skills needed to be able to produce or apply these tools
- Potential of social media to influence policy conversations

Overview of main responses from experts.

Topic	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
Expert background	Legislative Director (Policy Analyst)	Principal Data Scientist	Former Technology Director at PwC	Trustee for the NYS Higher Education Services analyzing policy impacts	County Legislature Representative	Communications Specialist. User of data in news.
Relevance of the topics	Only the Ukraine Russia, secession topics are relevant.	Some are relevant.	Some are relevant.	Some are relevant.	Many topics are big issues in real life.	Some are more relevant to general public interest than the other.
Interpretability	Visualization of topics over time are more meaningful than a simple topic modeling presentation itself.	Interesting topics. Some topic signatures seem to match the general interest over time, like Clinton & election.	Likes the data visualization of signatures over time, but he would like to see it across longer time frame to get a more information. Some topics are very vague.	Change of the number of the signatures match the change of the public interest in politics in real life.	He felt the results are interesting, but it could have been little confusing without some extra explanations. The search interest do not necessarily align with the number of the signatures.	It catches my eyes, but not very self-explanatory. More interested in the results with big increase or decrease.
Learnability	Very user friendly. If implemented into their domain, only need minor training to use it but need more training to understand the mechanism behind it.	Nice design. Very easy to interact with.	Easy to interact with it but not clear what are the insights that can be drawn from these results.	Easy to understand and interact with the interface.	It will be confusing without explanation.	It is pretty self-explanatory after the introduction. It is easy to interact with the interface.
Utility	Good to analyze feedback from residents automatically. May also be helpful to analyze some controversial issues (only to a certain degree).	It will be useful to track longitudinal change if it can be proved representing general publics.	Data visualization help easily communicate insights with people with different levels of technical backgrounds.	It will be helpful to dealing with large amount of qualitative data. Google trend results may better represent general public interest rather than petition signatures.	Helpful for elected officials and policy makers to get to know the specific people's concern and attitude towards certain issues.	Help interact with data, and easier to extract information from the visualization.
Skills needed to be able to produce or apply these tools	Make sure that the data are from expert sources that can represent the general publics. Need to be trained to understand the mechanics behind the scenes.	Need solid technical skills to put things together. Also need technical training to use the tools.	How to use the tool to effectively communicate with clients and let them understand the information contained in the visualization.	Critical thinking, be reflective, good at math, computer science and technology.	Data analysis skill, programming skills.	Data analytic skill.
Potential of social media to influence policy conversations	Social media and petitions sites definitely play a role, but it cannot represent the general public because of access issues.	Not sure how much it will have impact on legislatures or policy making.	Skeptical because they only represent small groups of people who are either far-left or far-right.	Not sure if social media and petition can represent the general public.	It's good to collect feedback and interact with people. Not sure how it will affect the policy making, may raise awareness.	Petitions won't necessarily lead to any change in policy making. Inappropriate contents of the petitions make people view them as non-high quality reference.

DISCUSSION

- Positive feedback on interpretability and relevance
- Recommendations
  - Granular topics with stance or sentiment
  - Topic marked with meaningful labels
  - More locally meaningful datasets
- Take away:
  - User centric tool development
  - Work practices re-design
  - Resource limitation: open source, training, technical capacity

Hagen, L., Seon Yi, H., Pietri, S., & E. Keller, T. (2019). **Processes, Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami.** *20th Annual International Conference on Digital Government Research on - Dg.o 2019*, 1–10. https://doi.org/10.1145/3325112.3325212

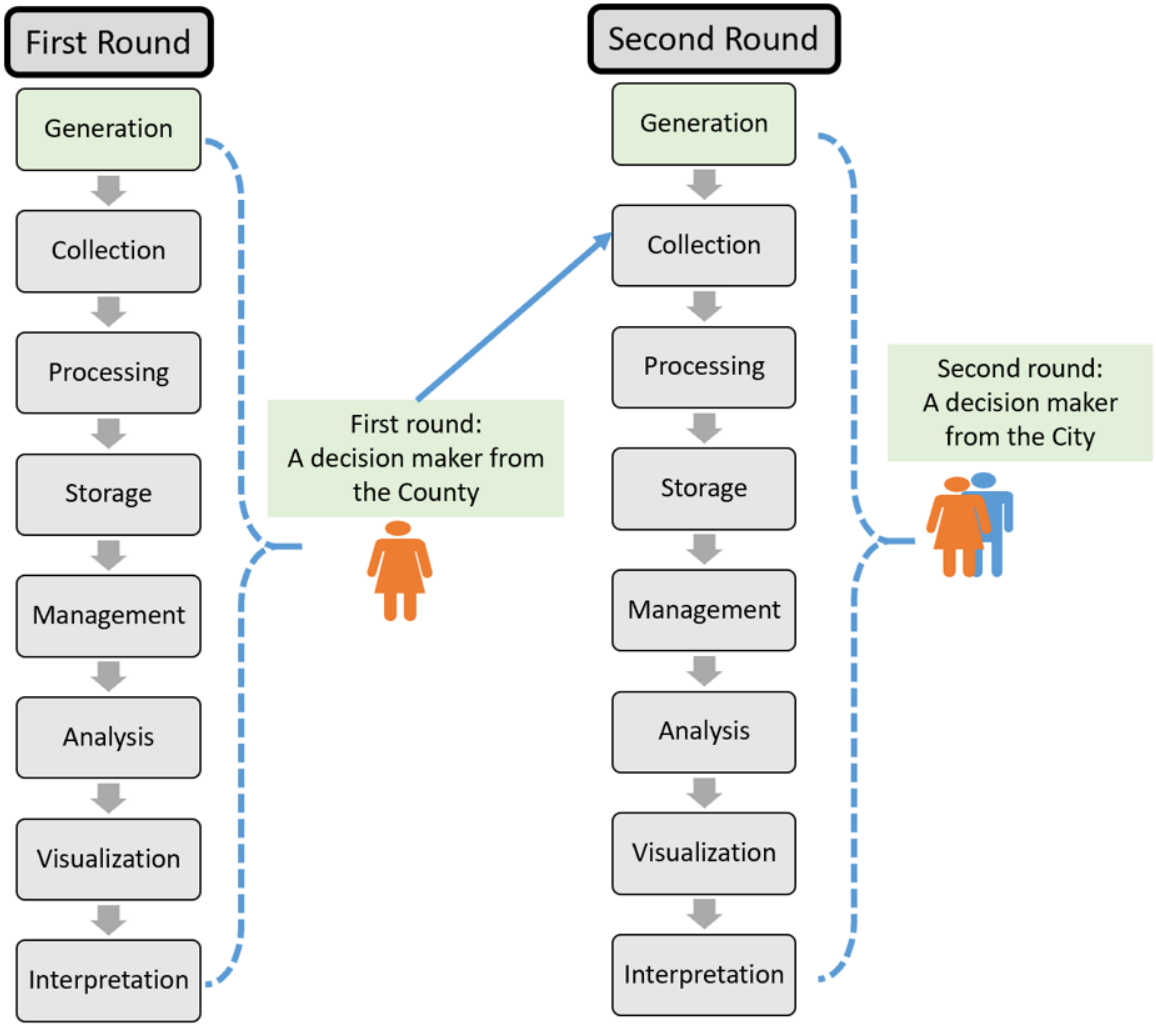


Figure 2. Supervised Data Science Strategy

DATA IS BIASED

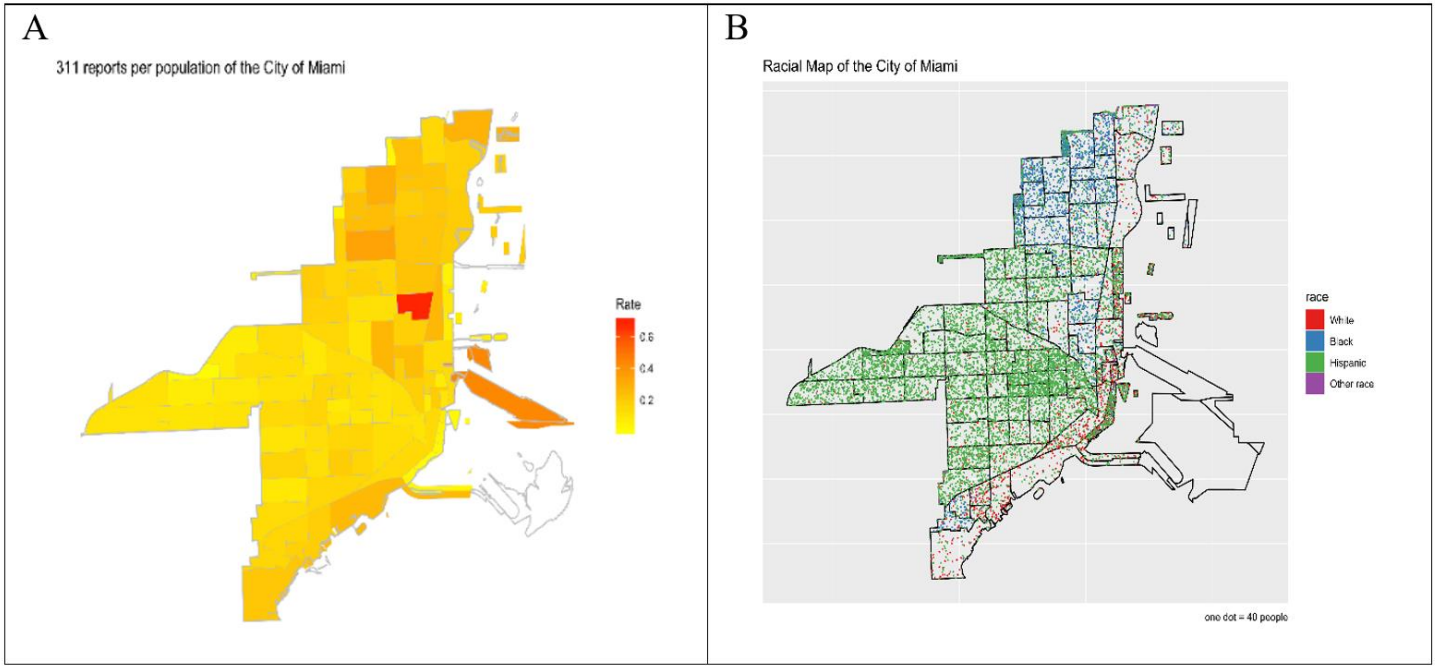
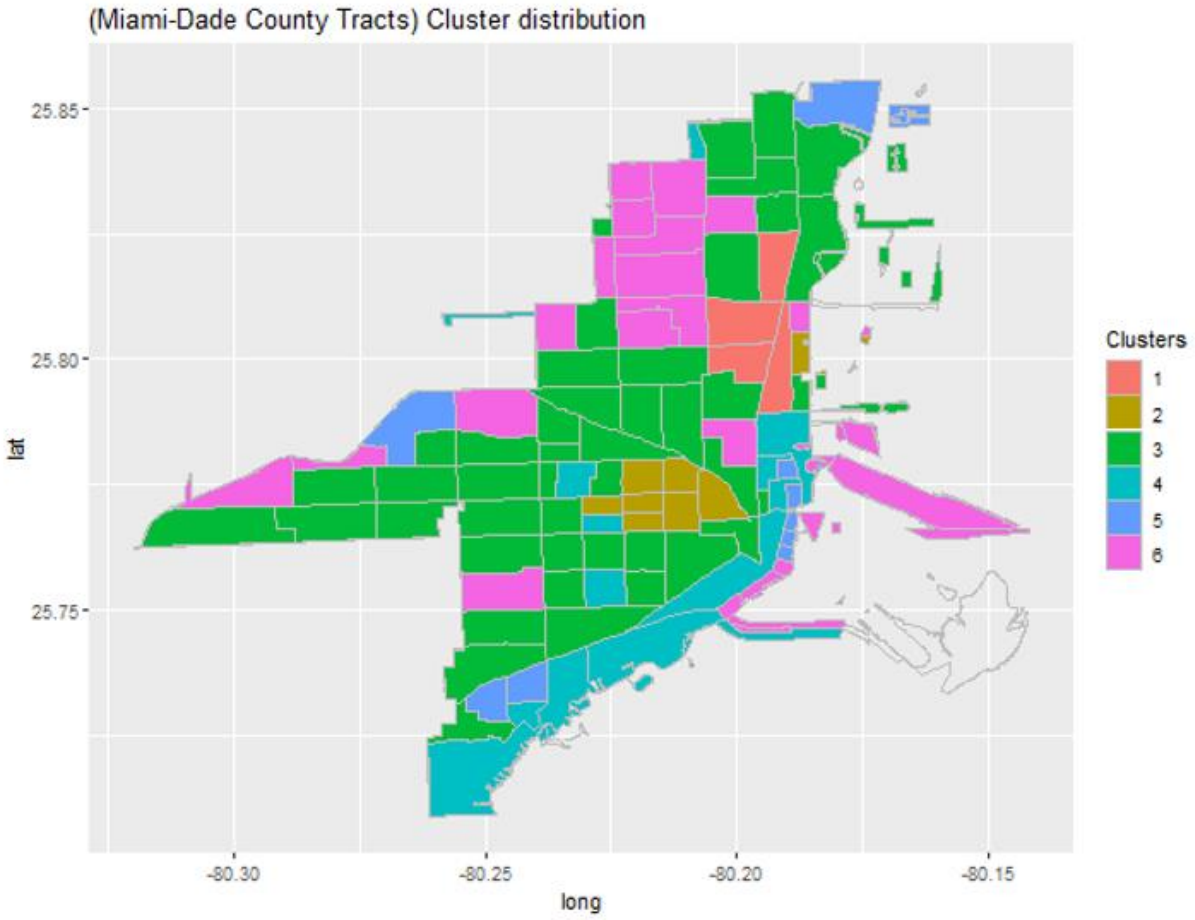
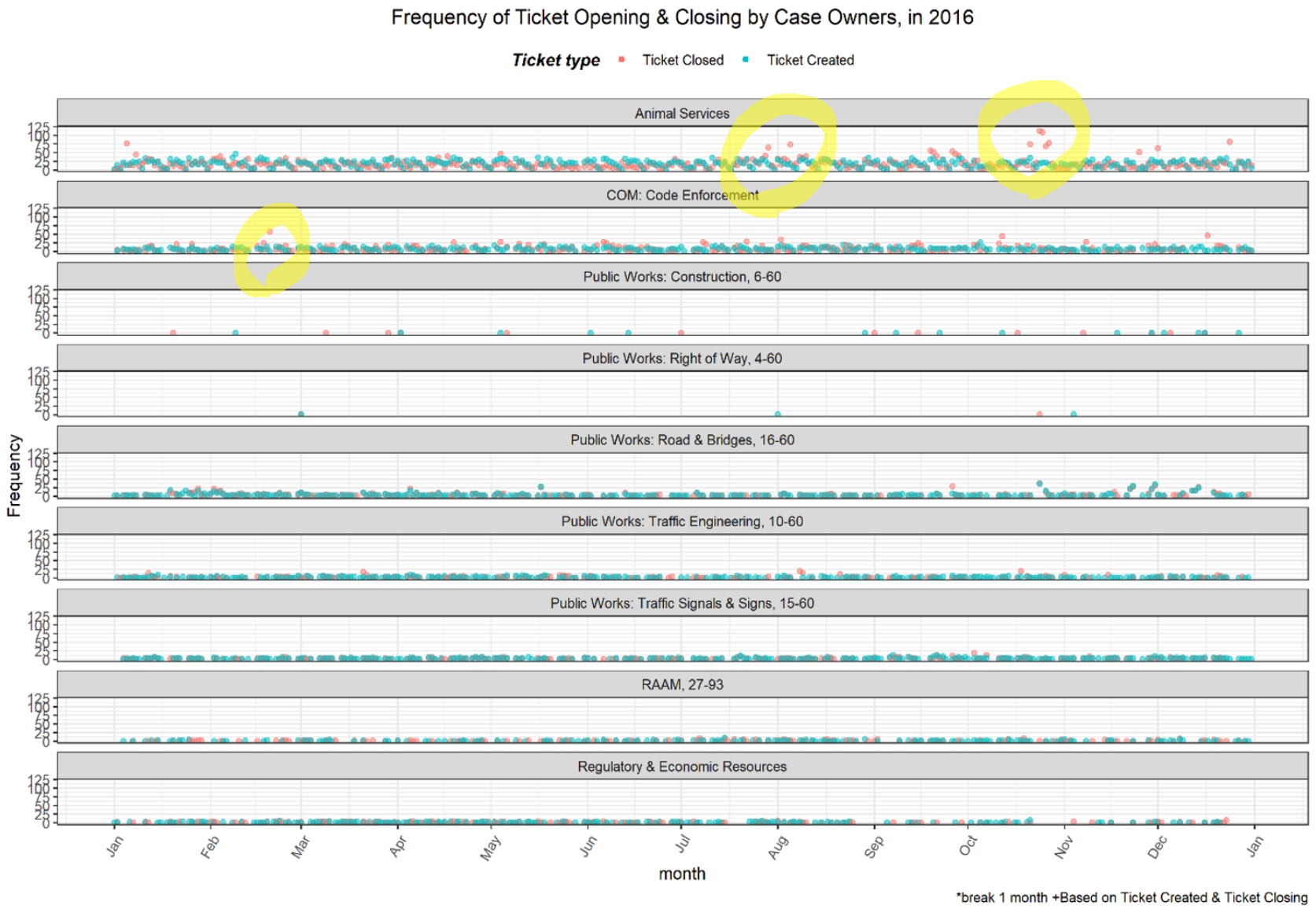


Figure 4. A. 311 requests per person and B. geographical display of racial composition
Note: color of a dot represents each race and one dot explains 40 residents in the tracts.

A. Geographical Presentation of Six Clusters based on category distribution of 311 requests




Frequency of Ticket Opening & Closing by Case Owners, in 2016
Ticket type
Ticket Closed
Ticket Created
Animal Services
COM: Code Enforcement
Public Works: Construction, 6-60
Public Works: Right of Way, 4-60
Public Works: Road & Bridges, 16-60
Public Works: Traffic Engineering, 10-60
Public Works: Traffic Signals & Signs, 15-60
RAAM, 27-93
Regulatory & Economic Resources
Frequency
0
25
50
75
100
125
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
month
*break 1 month +Based on Ticket Created & Ticket Closing

Data Quality Issues

Data life cycle

generation → collection → processing → storage → management → analysis → visualization → interpretation

Local 311 data

Local decision makers

privacy and ethical concerns throughout

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.e26845b4

INSTEAD OF HUMAN TESTING,

- Human-supervised data science
- Human in the loop data science

THANK YOU! QUESTIONS?

lonihagen@usf.edu

Twitter: @lonihagen

SDSS
SYMPOSIUM ON
DATA SCIENCE & STATISTICS
BEYOND BIG DATA: COLLABORATION
IN SCIENCE, INDUSTRY, AND SOCIETY
VIRTUAL • JUNE 3–5, 2020

UNIVERSITY of
SOUTH FLORIDA