

# *Classy* Data Analysis

Hansen | Nagda | Pimentel | Reyes | Tai





# Agenda

1. Introduction
  - a. Solution
  - b. Purpose & Stakeholders
2. The Data
  - a. Data Pipeline
  - b. Communication Plan
3. Model Solutions
  - a. Text Modeling
  - b. Financial Modeling
  - c. Geographical
4. Product
5. Results & Conclusion



# Challenge

**We have a 1-D understanding of Non Profit Organizations (NPOs)...**

- More often than not, the services of an NPO span multiple sectors, e.g Health, Education, etc.
- Financially speaking, small local charities operate very differently from multi-million dollar organizations.

We need a more complex solution to find like-minded organizations.

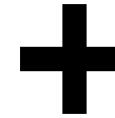




# Solution

**... but NPOs are multi-faceted.**

- Lay out a common “social space”, where the organizations that drive social change and potential donors can connect, and be offered recommendations, and where discovery of new causes - and events within causes (i.e. fundraisers) - could be facilitated.
- Use combination of government IRS form 990 (returns for nonprofits) data along with external textual information (i.e. social media) to create a robust semantic space.

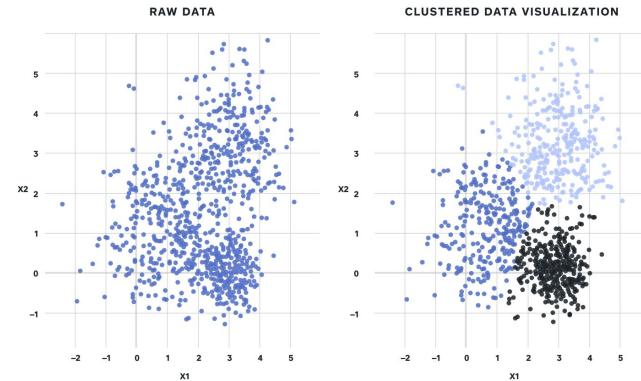




# Purpose

To connect NPOs with their peers and potential donors.

1. How do non-profit organizations (NPO) cluster in this semantic space?
  - a. Are there groups that were not easily apparent before?
2. How can data help non-profits in their fundraising and other efforts?
3. How does one nonprofit compare to another?
  - a. Benchmarking
    - i. Financial Performance
    - ii. Mission Statements
    - iii. Geographic location



The #1 online and mobile fundraising platform.



# The Stakeholders

- Classy Corporation
  - Use new social space to better engage with fundraising user base
  - Recommendations are more targeted with better similarity
- Non-Profit Researchers
  - Creating a new semantic space of non-profit organization can allow enhanced understanding of nonprofits' interrelationships rather than a wall of numbers.
- Other users
  - Publishing this model as an open source system allows other users to use this to create their own semantic space of non-profit organizations to develop new insights and recommendations





## Stories of Usage - #1

A very passionate supporter of the World Wide Fund for Nature (WWF) wants to reach out to other like-minded nonprofits to open discussions about creating a powerful “alliance” for nature conservancy. She desires to reach out to similarly well-funded organizations like the WWF and plans on personally meeting with the heads of such organizations.





## Stories of Usage - #2

A women's rights activist wants to get involved with the Women's March Foundation, but there are no marches being held in her hometown. She would like to find information about like-minded organizations operating within her rural community.



**WOMEN'S  
MARCH  
FOUNDATION**





## Stories of Usage - #3

A new non-profit organization serving hungry children in developing nations is struggling to make strides. The head of the organization, John, would like to grow his network and connect with heads from other local charities to gain advice and discuss fundraising strategies.

John wants to engage with charities that have raised funds for similar purposes and but that are also within driving distance.





# Agenda

1. Introduction
  - a. Solution
  - b. Purpose & Stakeholders
2. **The Data**
  - a. **Data Pipeline**
  - b. **Communication Plan**
3. Model Solutions
  - a. Text Modeling
  - b. Financial Modeling
  - c. Geographical
4. Product
5. Results & Conclusion



# Where Do We Get Our Data?



Open Data on AWS

Share any volume of data with as many people as  
you want



\* Forms 990 and 990-EZ which have been electronically filed with the IRS

# What Does The Data Look Like?

Form 990

Return of Organization Exempt From Income Tax	
Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)	
► Do not enter social security numbers on this form as it may be made public. ► Go to www.irs.gov/Form990 for instructions and the latest information.	
B For the 2017 calendar year, or tax year beginning , 2017, and ending , 20	
C Name of organization D Employer identification number	
D Doing business as E Telephone number	
F Name and address of principal officer: G Gross receipts \$	
H Is this a group return for subdivisions? Yes No I Are all subdivisions included? Yes No If "No," attach a list. (see instructions) J Group exemption number ►	
K Form of organization: Corporation Trust Association Other ► L Year of formation: M State of legal domicile:	
Part I Summary	
1 Briefly describe the organization's mission or most significant activities:  2 Check this box [ ] if the organization discontinued its operations or disposed of more than 25% of its net assets. 3 Number of voting members of the governing body (Part VI, line 1a) . . . . . 4 Number of non-voting members of the governing body (Part VI, line 1b) . . . . . 5 Total number of individuals employed in the current year 2017 (Part V, line 2a) . . . . . 6 Total number of volunteers (estimate if necessary) . . . . . 7a Total unrelated business revenue from Part VIII, column (C), line 12 . . . . . 7b Net unrelated business taxable income from Form 990-T, line 34 . . . . .	
Revenue Expenses	
8 Contributions and grants (Part VIII, line 1h) . . . . . 9 Program service revenue (Part VIII, line 2g) . . . . . 10 Investment income (Part VIII, lines 3, 4, and 7d) . . . . . Other revenue (Part VIII, column (A), lines 5, 6d, 8c, 9c, and 11e) . . . . . 12 Total revenue—add lines 8 through 11 must equal Part VIII, column (A), line 12 . . . . .	
13 Grants and similar amounts paid (Part IX, column (A), lines 1-3) . . . . . 14 Benefits paid to or for members (Part IX, column (A), line 4) . . . . . 15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5-10) . . . . . 16 Professional fundraising fees (Part IX, column (A), line 11e) . . . . . b Total fundraising expenses (Part IX, column (D), line 23) ► . . . . . 17 Other expenses (Part IX, column (A), lines 1a-11d, 11f-24e) . . . . . 18 Total expenses. Add lines 13-17 (must equal Part IX, column (A), line 25) . . . . . 19 Revenue less expenses. Subtract line 18 from line 12 . . . . .	
Beginning of Current Year End of Year	
20 Total assets (Part X, line 16) . . . . . 21 Total liabilities (Part X, line 26) . . . . . 22 Net assets or fund balances. Subtract line 21 from line 20 . . . . .	
Part II Signature Block	
Under penalties of perjury, I declare that I have examined this return, including accompanying schedules and statements, and to the best of my knowledge and belief, it is true, correct, and complete. Declaration of preparer (other than officer) is based on all information of which preparer has any knowledge.	
Sign Here ► Signature of officer Date Type or print name and title	
Paid Preparer Use Only ► Preparer's name Preparer's signature Date Check [ ] if PTIN self-employed Firm's name ► Firm's address ► Phone number ► Yes No	
May the IRS discuss this return with the preparer shown above? (see instructions) Yes No For Paperwork Reduction Act Notice, see the separate instructions. Cat. No. 11282Y Form 990 (2017)	

Form 990-EZ

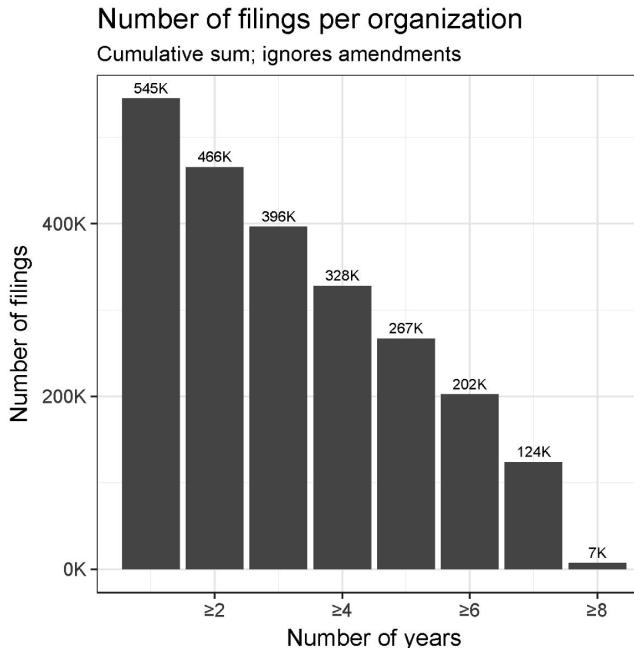
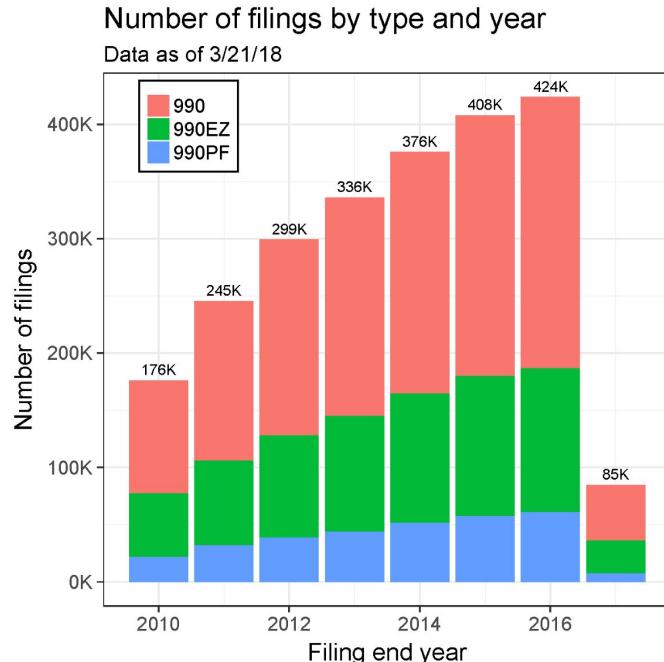
Return of Organization Exempt From Income Tax	
Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)	
► Do not enter social security numbers on this form as it may be made public. ► Go to www.irs.gov/Form990EZ for instructions and the latest information.	
B For the 2017 calendar year, or tax year beginning , 2017, and ending , 20	
C Name of organization D Employer identification number	
E Telephone number	
F Name and address of principal officer: G Gross receipts \$	
H Is this a group return for subdivisions? Yes No I Are all subdivisions included? Yes No If "No," attach a list. (see instructions) J Group exemption number ►	
K Form of organization: Corporation Trust Association Other ► L Year of formation: M State of legal domicile:	
Part II Summary	
1 Briefly describe the organization's mission or most significant activities:  2 Check this box [ ] if the organization discontinued its operations or disposed of more than 25% of its net assets. 3 Number of voting members of the governing body (Part VI, line 1a) . . . . . 4 Number of non-voting members of the governing body (Part VI, line 1b) . . . . . 5 Total number of individuals employed in the current year 2017 (Part V, line 2a) . . . . . 6 Total number of volunteers (estimate if necessary) . . . . . 7a Total unrelated business revenue from Part VIII, column (C), line 12 . . . . . 7b Net unrelated business taxable income from Form 990-T, line 34 . . . . .	
Revenue Expenses	
8 Contributions and grants (Part VIII, line 1h) . . . . . 9 Program service revenue (Part VIII, line 2g) . . . . . 10 Investment income (Part VIII, lines 3, 4, and 7d) . . . . . Other revenue (Part VIII, column (A), lines 5, 6d, 8c, 9c, and 11e) . . . . . 12 Total revenue—add lines 8 through 11 must equal Part VIII, column (A), line 12 . . . . .	
13 Grants and similar amounts paid (Part IX, column (A), lines 1-3) . . . . . 14 Benefits paid to or for members (Part IX, column (A), line 4) . . . . . 15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5-10) . . . . . 16 Professional fundraising fees (Part IX, column (A), line 11e) . . . . . b Total fundraising expenses (Part IX, column (D), line 23) ► . . . . . 17 Other expenses (Part IX, column (A), lines 1a-11d, 11f-24e) . . . . . 18 Total expenses. Add lines 13-17 (must equal Part IX, column (A), line 25) . . . . . 19 Revenue less expenses. Subtract line 18 from line 12 . . . . .	
Beginning of Current Year End of Year	
20 Total assets (Part X, line 16) . . . . . 21 Total liabilities (Part X, line 26) . . . . . 22 Net assets or fund balances. Subtract line 21 from line 20 . . . . .	
Part III Signature Block	
Under penalties of perjury, I declare that I have examined this return, including accompanying schedules and statements, and to the best of my knowledge and belief, it is true, correct, and complete. Declaration of preparer (other than officer) is based on all information of which preparer has any knowledge.	
Sign Here ► Signature of officer Date Type or print name and title	
Paid Preparer Use Only ► Preparer's name Preparer's signature Date Check [ ] if PTIN self-employed Firm's name ► Firm's address ► Phone number ► Yes No	
May the IRS discuss this return with the preparer shown above? (see instructions) Yes No For Paperwork Reduction Act Notice, see the separate instructions. Cat. No. 11282Y Form 990 (2017)	

Form 990-PF

Return of Private Foundation	
Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)	
► Do not enter social security numbers on this form as it may be made public. ► Go to www.irs.gov/Form990PF for instructions and the latest information.	
B For the 2017 calendar year or tax year beginning , 2017, and ending , 20	
C Employer identification number	
D Telephone number (see instructions)	
E If exemption application is pending, check here [ ]	
F If foreign organizations meeting the 85% test, check here [ ] G If exemption application is pending, check here [ ]	
H If private foundation status was terminated under section 507(d)(1)(A), check here [ ]	
I Check all that apply: Initial return Final return Amended return Address change Name change	
J Check type of organization: Section 501(c)(3) exempt private foundation Section 4947(a)(1) nonexempt charitable trust Other taxable private foundation	
K Fair market value of all assets at end of year (from Part II, col. (c), line 16) ► S	
L Accounting method: Cash Accrual	
M If the foundation is in a 60-month termination under section 507(b)(1)(B), check here [ ]	
Part I Analysis of Revenue and Expenses (The total of amounts in columns (b), (c), and (d) may not necessarily equal the amounts in column (a) (see instructions).)	
(a) Revenue and expenses per books (b) Net investment income (c) Adjusted net income (d) Disbursements for charitable purposes (cash basis only)	
1 Contributions, gifts, etc., received (attach schedule) 2 Check [ ] if the foundation is not required to attach Sch. B 3 Interest, dividends, savings and cash investments 4 Dividends and interest from securities 5 Gross rents 6 Net rental income or (loss) 6a Gross gain or (loss) from sale of assets on line 6a 7 Capital gain net income (from Part IV, line 2) 8 Net short-term capital gain 9 Income less expenses 10 Gross sales less returns and allowances b Less: Cost of goods sold c Gross profit or (loss) (attach schedule) 11 Other income (attach schedule) 12 Total. Add lines 1 through 11	
13 Compensation of officers, directors, trustees, etc. 14 Other employee salaries and wages 15 Professional fees and other payments to independent contractors 16 Legal fees (attach schedule) c Other professional fees (attach schedule) 17 Interest 18 Taxes (attach schedule) (see instructions) 19 Depreciation (attach schedule) and depletion 20 Occupancy 21 Travel, conference, and meetings 22 Printing, publications, postage, and shipping 23 Other expenses (attach schedule) 24 Total operating and administrative expenses. Add lines 13 through 23 25 Contributions, gifts, grants paid 26 Total expenses and disbursements. Add lines 24 and 25 27 Subtract line 26 from line 12: a Excess of revenue over expenses and disbursements b Net investment income (if negative, enter -0-) . . . . . c Adjusted net income (if negative, enter -0-) . . . . .	

## What Does The Data Look Like? - AWS XML

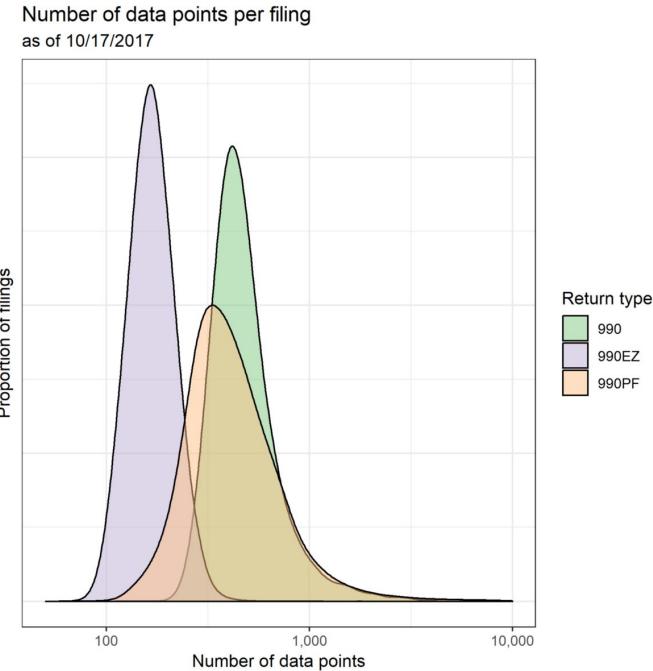
# What Does The Data Look Like?



Source: Applied Nonprofit Research, L.L.C.



# What Does The Data Look Like?



Source: Applied Nonprofit Research, L.L.C.

# What Does The Data Look Like? - Filer

. Any Non-Private Foundation

**990**

**Return of Organization Exempt From Income Tax**  
OMB No. 1545-0047  
**2017**  
**Open to Public Inspection**

Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)  
 Do not enter social security numbers on this form as it may be made public.  
 Go to [www.irs.gov/Form990](http://www.irs.gov/Form990) for instructions and the latest information.

A For the 2017 calendar year, or tax year beginning	, 2017, and ending	. 20
B Check if applicable:		
<input type="checkbox"/> Address change	<input type="checkbox"/> Employer identification number	
<input type="checkbox"/> Name change	<input type="checkbox"/> Name of organization	
<input type="checkbox"/> Initial return	<input type="checkbox"/> Doing business as	
<input type="checkbox"/> Final return/revised return	<input type="checkbox"/> Number and street (or P.O. box if mail is not delivered to street address)	
<input type="checkbox"/> Amended return	<input type="checkbox"/> Room/suite	
<input type="checkbox"/> Application pending	<input type="checkbox"/> Telephone number	
C Name of organization		
D Employer identification number		
E Telephone number		
F Name and address of principal officer		
G Gross receipts \$		
H Is this a group return for subordinates? <input type="checkbox"/> Yes <input type="checkbox"/> No		
H(b) Are all subordinates included? <input type="checkbox"/> Yes <input type="checkbox"/> No		
If "No," attach a list. (see instructions)		
I Tax-exempt status: <input type="checkbox"/> 501(c)(3) <input type="checkbox"/> 501(c) ( ) (insert no.) <input type="checkbox"/> 4947(a)(1) or <input type="checkbox"/> 527		
J Website: ►		
K Form of organization: <input type="checkbox"/> Corporation <input type="checkbox"/> Trust <input type="checkbox"/> Association <input type="checkbox"/> Other ► L Year of formation: <input type="checkbox"/> State of legal domicile:		

**Part I Summary**

1 Briefly describe the organization's mission or most significant activities:		
2 Check this box ► if the organization discontinued its operations or disposed of more than 25% of its net assets.		
3 Number of voting members of the governing body (Part VI, line 1a)	3	
4 Number of independent voting members of the governing body (Part VI, line 1b)	4	
5 Total number of independent voting members for year 2017 (Part V, line 2a)	5	
6 Total number of volunteers (estimate if necessary)	6	
7a Total unrelated business revenue from Part VIII, column (C), line 12	7a	
b Net unrelated business taxable income from Form 990-T, line 34	7b	

**Activities & Governance**

	Prior Year	Current Year
8 Contributions and grants (Part VIII, line 1h)		
9 Program service revenue (Part VIII, line 2g)		
10 Investment income (Part VIII, column (A), lines 3, 4, and 7d)		
11 Total revenue—Part VIII, column (A), lines 5, 6a, 8c, 9c, 10a, and 11e)		
12 Total revenue—add lines 6 through 11 (must equal Part VIII, column (A), line 12)		

**Revenue**

	Beginning of Current Year	End of Year
13 Grants and similar amounts paid (Part IX, column (A), lines 1-3)		
14 Benefits paid to or for members (Part IX, column (A), line 4)		
15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5-10)		
16a Professional fundraising fees (Part IX, column (A), line 11e)		
b Total fundraising expenses (Part IX, column (D), line 25) ►		
17 Other expenses (Part IX, column (A), lines 11a-11d, 11f-24e)		
18 Total expenses. Add lines 13-17 (must equal Part IX, column (A), line 25)		
19 Revenue less expenses. Subtract line 18 from line 12		

**Expenses**

	Beginning of Current Year	End of Year
20 Total assets (Part X, line 16)		
21 Total liabilities (Part X, line 26)		
22 Net assets or fund balances. Subtract line 21 from line 20		

**Net Assets or Fund Balances**

**Part II Signature Block**  
Under penalties of perjury, I declare that I have examined this return, including accompanying schedules and statements, and to the best of my knowledge and belief, it is true, correct, and complete. Declaration of preparer (other than officer) is based on all information of which preparer has any knowledge.

**Sign Here**

► Signature of officer Date  
Type or print name and title

Paid Preparer Use Only  
Print/Type preparer's name Preparer's signature Date  If self-employed  
Firm's name Firm's EIN ►  
Firm's address Phone no.

May the IRS discuss this return with the preparer shown above? (see instructions)  Yes  No

For Paperwork Reduction Act Notice, see the separate instructions. Cat. No. 11282Y Form 990 (2017)

. Gross Receipts < \$50,000,  
\$50,000 < Gross Receipts < \$200,000  
And Total Assets < \$500,000

OMB No. 1545-1150

**2017**

**Open to Public Inspection**

**990-EZ**

**Short Form**

**Return of Organization Exempt From Income Tax**  
OMB No. 1545-1150  
**2017**  
**Open to Public Inspection**

Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except private foundations)  
 Do not enter social security numbers on this form as it may be made public.  
 Go to [www.irs.gov/Form990EZ](http://www.irs.gov/Form990EZ) for instructions and the latest information.

A For the 2017 calendar year, or tax year beginning	, 2017, and ending	. 20
B Check if applicable:		
<input type="checkbox"/> Address change	<input type="checkbox"/> Employer identification number	
<input type="checkbox"/> Name change	<input type="checkbox"/> Name of organization	
<input type="checkbox"/> Initial return	<input type="checkbox"/> Doing business as	
<input type="checkbox"/> Final return/terminated	<input type="checkbox"/> Number and street (or P.O. box if mail is not delivered to street address)	
<input type="checkbox"/> Amended return	<input type="checkbox"/> Room/suite	
<input type="checkbox"/> Application pending	<input type="checkbox"/> Telephone number	
C Name of organization		
D Employer identification number		
E Telephone number		
F Group Exemption Number ►		
G Accounting Method: <input type="checkbox"/> Cash <input type="checkbox"/> Accrual <input type="checkbox"/> Other (specify) ►		
H Check ► <input type="checkbox"/> if the organization is not required to attach Schedule B (Form 990, 990-EZ, or 990-PF).		

J Website: ►  
K Form of organization:  Corporation  Trust  Association  Other  
L Add lines 5b, 6c, and 7b to line 9 to determine gross receipts. If gross receipts are \$200,000 or more, or if total assets (Part II, column (B) below) are \$500,000 or more, file Form 990 instead of Form 990-EZ. ► \$

**Part I Revenue, Expenses, and Changes in Net Assets or Fund Balances** (see the instructions for Part I) ►  
Check if the organization used Schedule O to respond to any question in this Part I.

1 Contributions, gifts, grants, and similar amounts received	1
2 Prior year's revenue including government fees and contracts	2
3 Membership dues and assessments	3
4 Investment income	4
5a Gross amount from sale of assets other than inventory	5a
b Less: cost or other basis and sales expenses	5b
c Gain or (loss) from sale of assets other than inventory (Subtract line 5b from line 5a)	5c
6 Gaming and fundraising events	6
a Gross income from gaming (attach Schedule G if greater than \$15,000)	6a
b Gross income from fundraising events (not including \$ from fundraising events reported on line 1) (attach Schedule G if the sum of such gross income and contributions exceeds \$15,000)	6b
c Less: direct expenses from gaming and fundraising events	6c
d Net income or (loss) from gaming and fundraising events (add lines 6a and 6b and subtract line 6c)	6d

7a Gross sales of inventory, less returns and allowances	7a
b Less: cost of goods sold	7b
c Gross profit or (loss) from sales of inventory (Subtract line 7b from line 7a)	7c
8 Other revenue (describe in Schedule O)	8
9 Total revenue. Add lines 1, 2, 3, 4, 5c, 6d, 7c, and 8	9

10 Grants and similar amounts paid (list in Schedule O)	10
11 Benefit paid to or for members	11
12 Salaries, other compensation, and employee benefits	12
13 Professional fees and other payments to independent contractors	13
14 Occupancy, rent, utilities, and maintenance	14
15 Printing, publications, postage, and shipping	15
16 Other expenses (describe in Schedule O)	16
17 Total expenses. Add lines 10 through 16	17
18 Excess or (deficit) for the year (Subtract line 17 from line 9)	18
19 Net assets or fund balances at beginning of year (from line 27, column (A)) (must agree with end-of-year figure reported on prior year's return)	19

20 Other changes in net assets or fund balances (explain in Schedule O)	20
21 Net assets or fund balances at end of year. Combine lines 18 through 20	21



# SQL vs NoSQL

**Data structure is inconsistent across organizations.**

- Form Fields Change Over Time
- Fields Change Between Different Form Types (E.g 990, 990-EZ)

**Solution:**

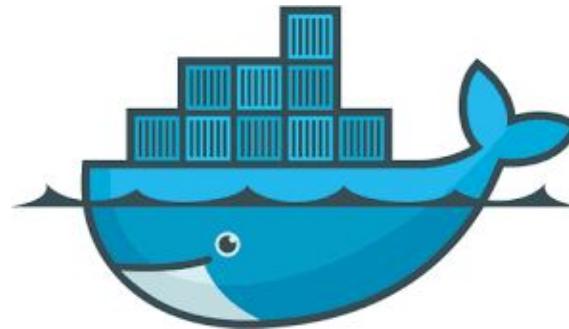




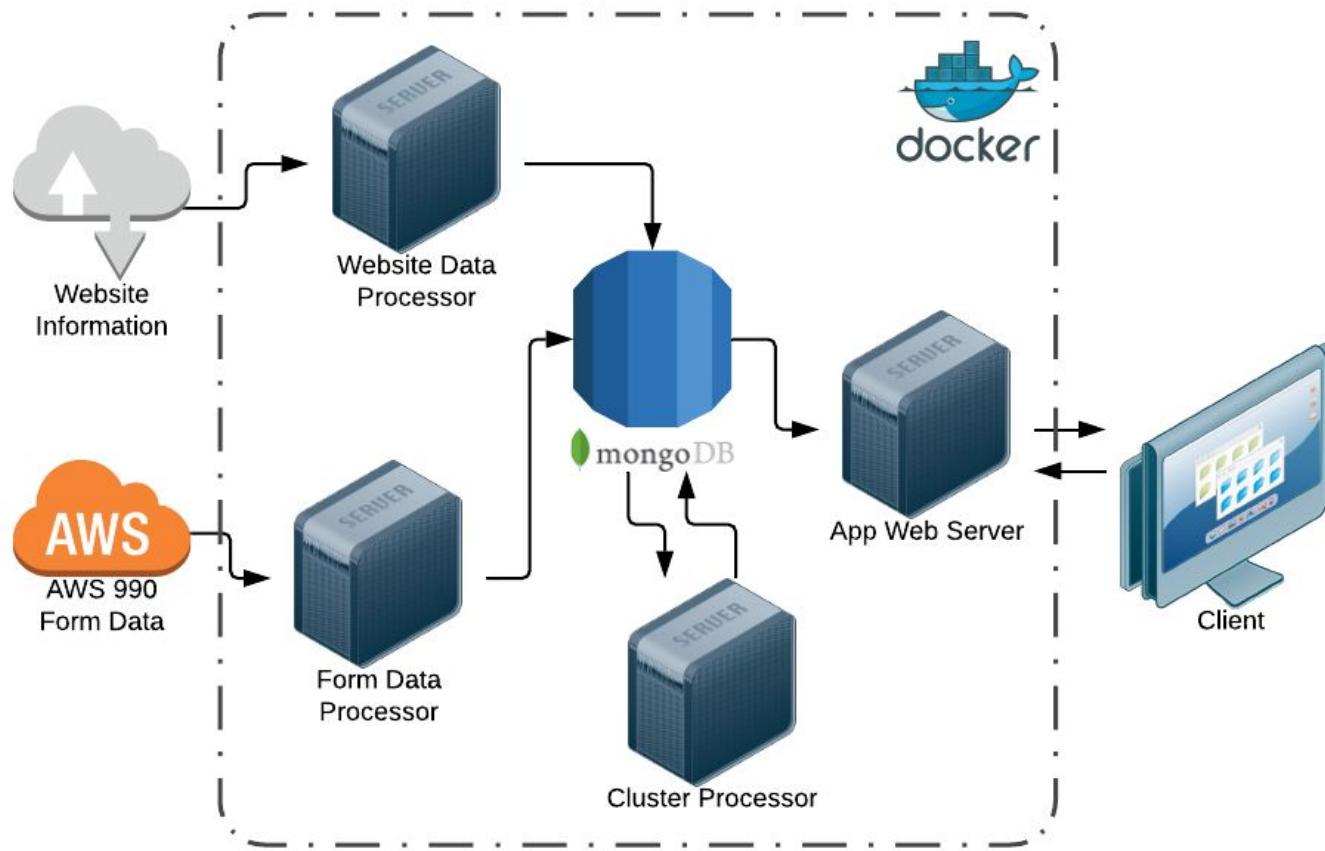
# Reproducibility

Docker containers allow us to scale.

- Created a complete docker environment for each module
- Allowed our code to be platform independent and easily scalable.



# Communication Plan



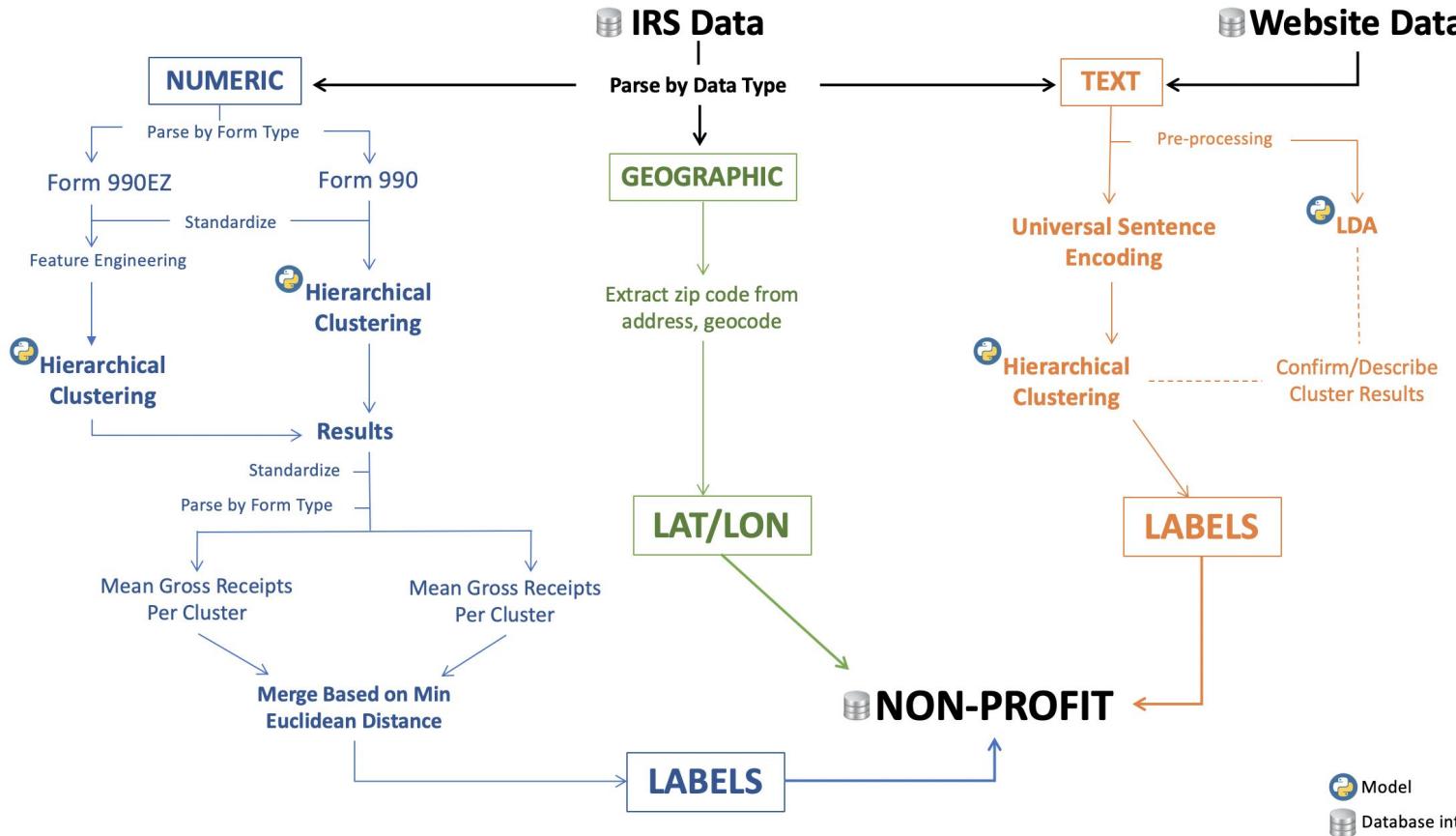


# Agenda

1. Introduction
  - a. Solution
  - b. Purpose & Stakeholders
2. The Data
  - a. Data Pipeline
  - b. Communication Plan
- 3. Model Solutions**
  - a. **Text Modeling**
  - b. **Financial Modeling**
  - c. **Geographical**
4. Product
5. Results & Conclusion

# Data Science Pipeline

The social space we sought to define is 3-dimensional.





# Text Modeling

## Validation / Processing Specs

**Preprocessing:** Lowercasing / Removal of Numbers + Stopwords + Punctuations / Lemmatization

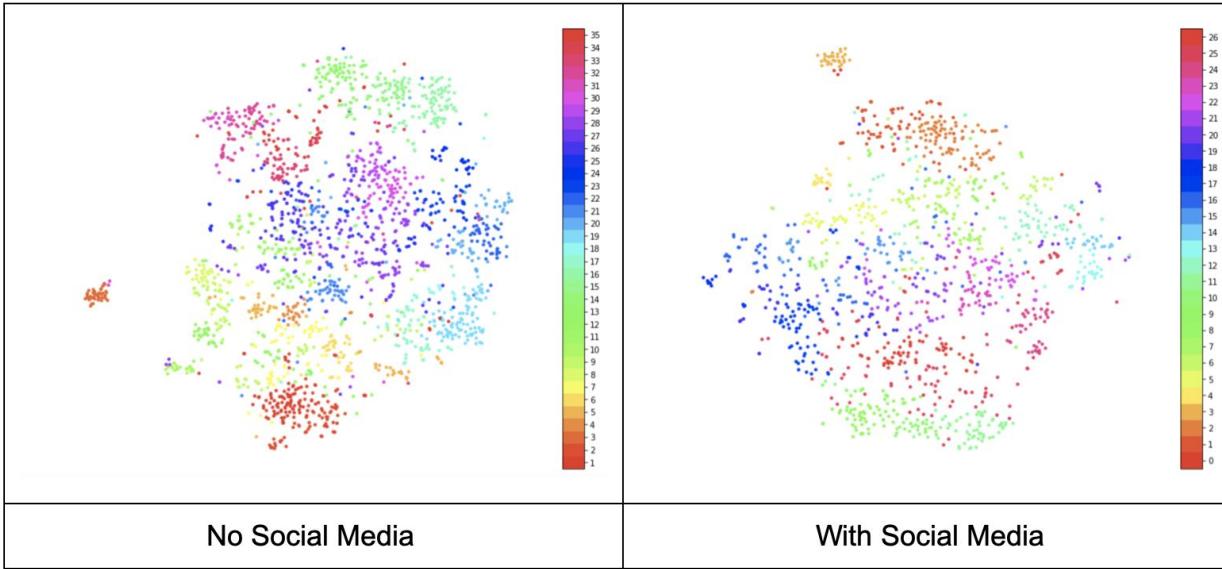
**Encoding:** Universal Sentence Encoder (512 output)

**Dimensionality Reduction:** PCA down to 200 features (95% variance) / Down to 70 features (84% variance)

**Hierarchical Clustering + Gap Statistic:** Linkage matrix generation and gap statistic scanning to find optimal number of clusters

# Significance of Results (t-SNE)

The addition of website data did not result in more robust clustering during the initial POC.



# Significance of Results (LDA)

Topic modeling verified the validity of our text clusters.

Topic 0:

youth club program provide sport activity recreational community soccer service

Topic 1:

public access library support organization state free people service medium

Topic 2:

museum organization inc facility program community food god providing literacy

Topic 3:

community church child christian training ministry service provide resource local

Topic 4:

community theatre member medium production radio young produce company resident

Topic 5:

water organization program environmental fair event florida restoration youth institute

Topic 6:

animal adoption promote provide humane shelter rescue dog pet help

Topic 7:

land conservation community organization park public center local energy preserve

Topic 8:

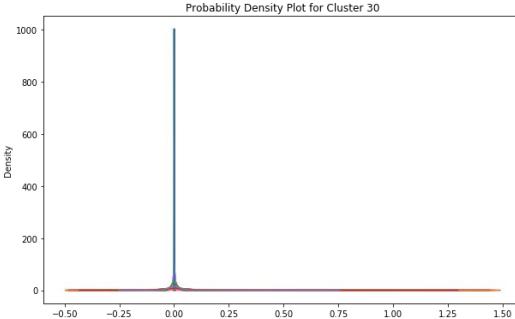
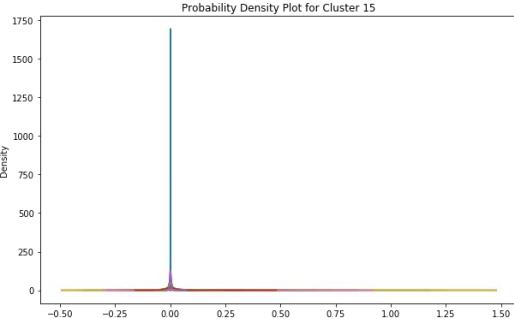
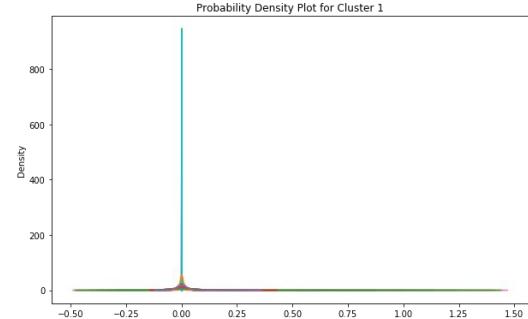
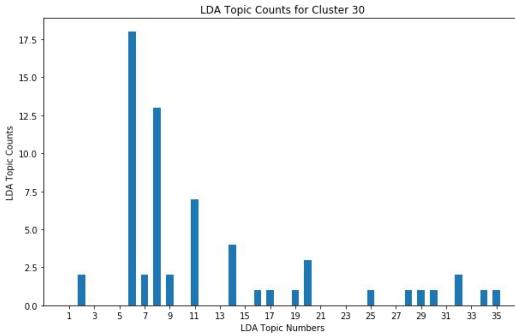
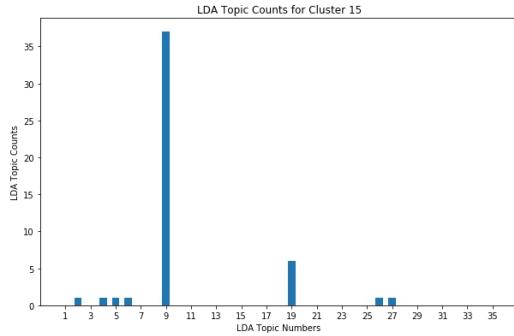
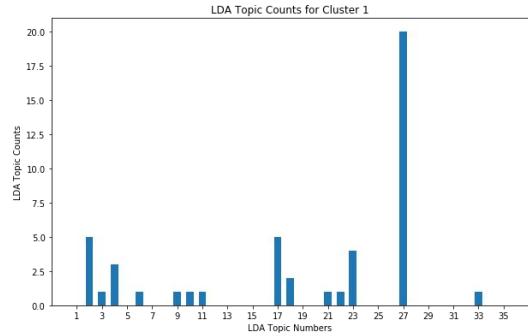
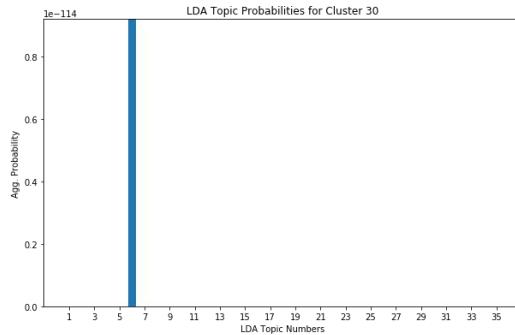
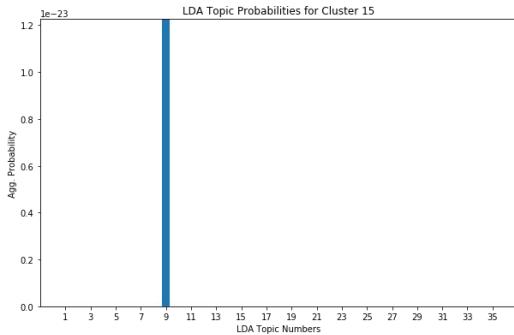
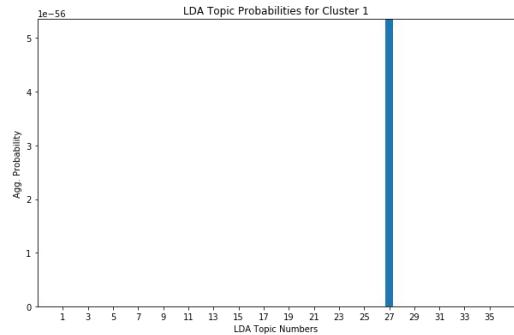
music camp community summer center art performance education experience inc

Topic 9:

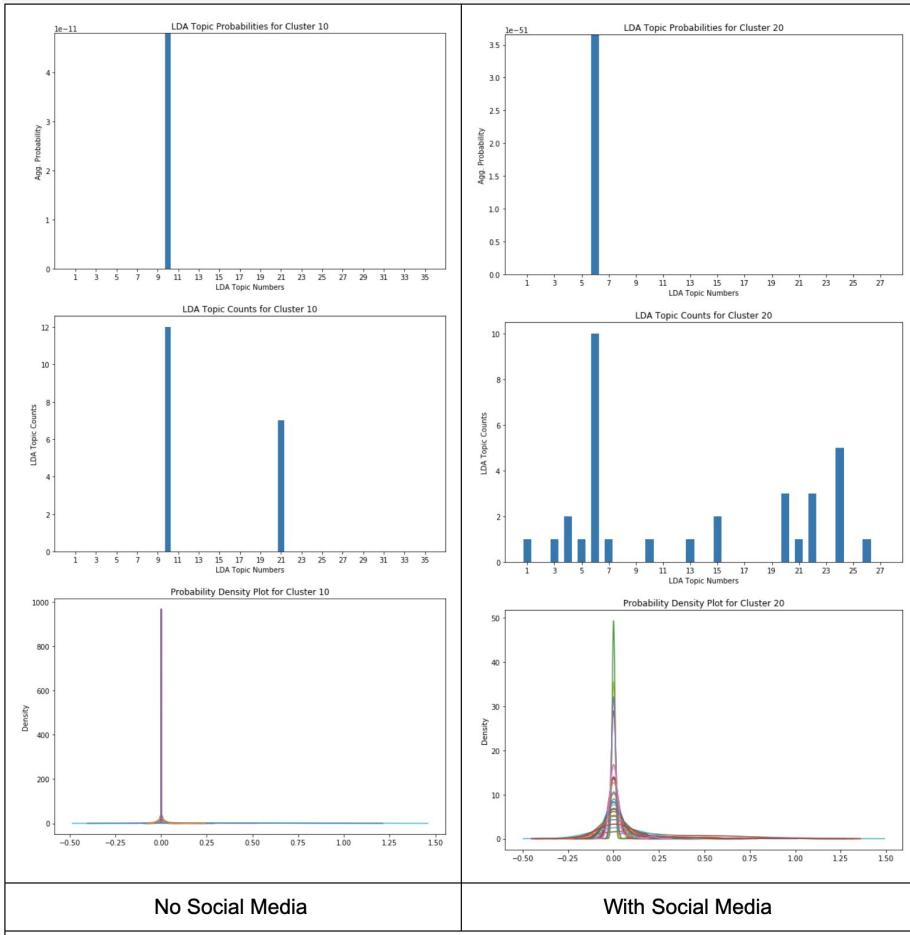
fire kid new provide county volunteer senior child family inc

Topic 10:

community development economic foundation organization program improve research life project



- It appears that the simple corpus yielded tighter clusters than the website-augmented version.
- This observation is well-aligned with the quality of the web-scraped text. Advanced web-scraping techniques can be a subject for future explorations.





# Text Validation (Local LDA)

Problem with Global LDA is that it assumes exactly one topic per cluster

- LDA (`n_topics`) is set to (`n_clusters`) from gap statistic
- However, what if hierarchical clustering + gap statistic failed to separate topics?
  - For example, one cluster contains 2 topics or more
- Local LDA approach fits an LDA model (`n_topics = 10`) **for each cluster**
- For each cluster, we ask:
  - In the 10-dimensional LDA topic space, how many regions of density are there?



# Topic Model Validation Algorithm

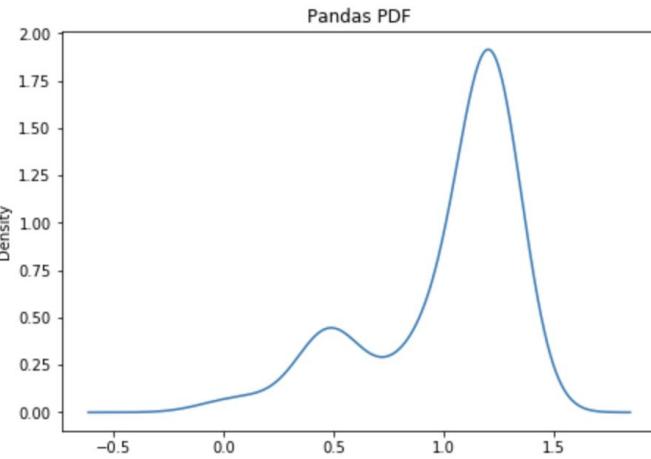
1. Cluster documents using hierarchical clustering and the gap statistic
2. For each cluster:
  - a. Fit an LDA model on the cluster documents, **setting the number of topics to be equal to 10 (arbitrarily)**
  - b. Transform cluster documents into the 10-dimensional topic space
  - c. Construct a pairwise-distance matrix shaped  $(N, N)$  where  $N$  is the number of documents in a given cluster (various distance metrics can be used)
  - d. For each row or equivalent column in the pairwise distance matrix,  
construct a probability density plot of distance values
    - i. Find the number of local maxima (above a specified threshold) for  
each probability density plot
    - ii. Aggregate statistics for number of local maxima (mean, standard  
deviation)



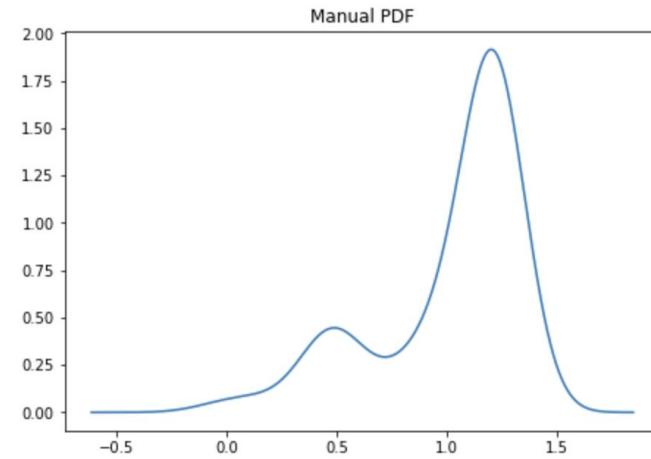
## We Claim...

“For a sample belonging to a well-defined cluster in N-dimensional space, its distance to all other samples within that cluster will be close to a uniform distribution. In a PDF representation, this distance will have a prominent peak indicating a high likelihood.”

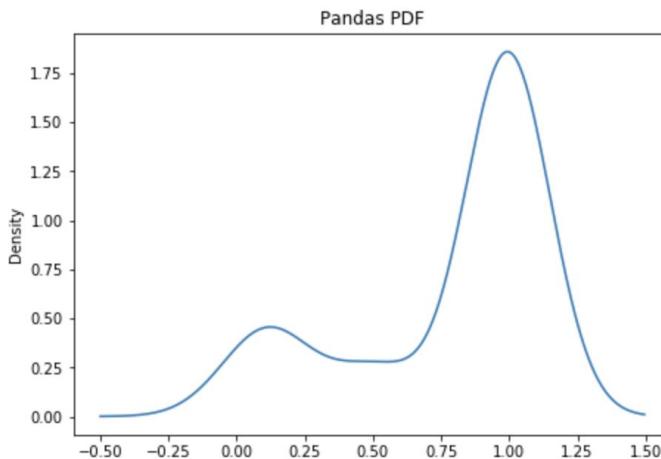
## Euclidean



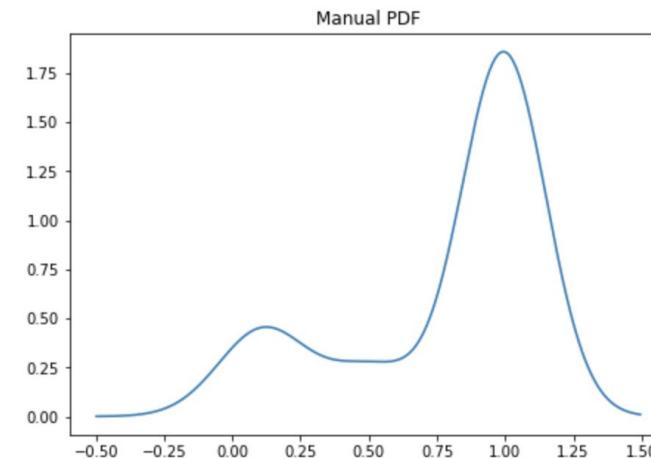
Number of local maxima: 1



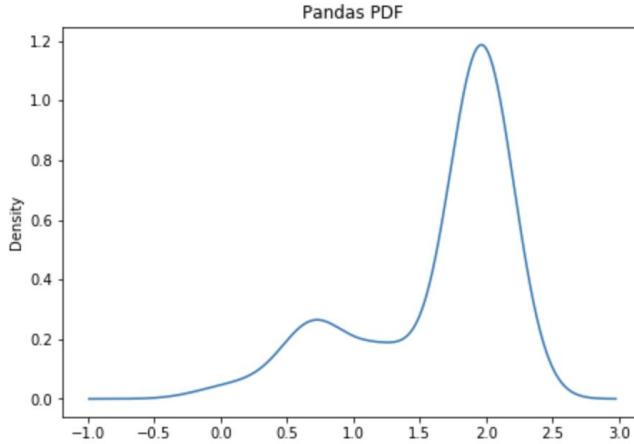
## Cosine



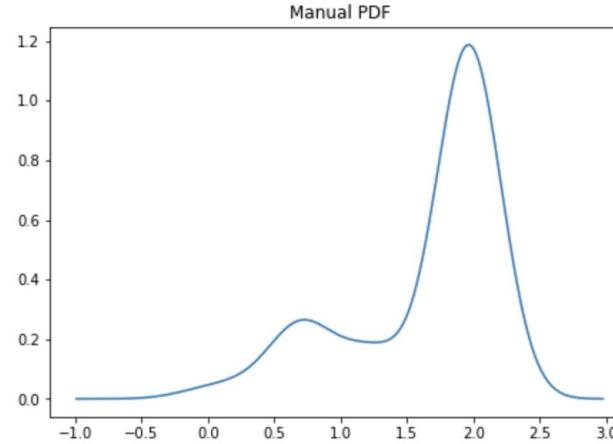
Number of local maxima: 1



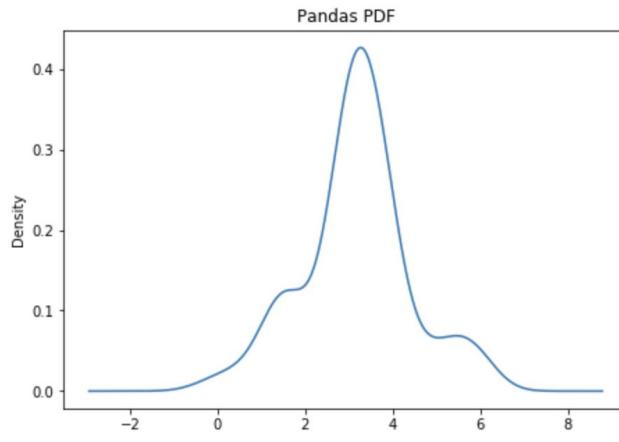
Cityblock



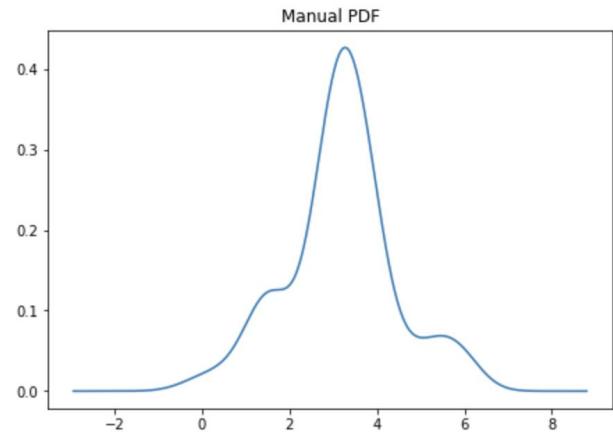
Number of local maxima: 1



Mahalanobis



Number of local maxima: 1





# Numeric Modeling

## Validation / Processing Specs

**Preprocessing:** Data cleaning, feature engineering, standardization

**Parsing:** Split data by form type and model separately

**Dimensionality Reduction:** PCA down to 5 features (99% variance)

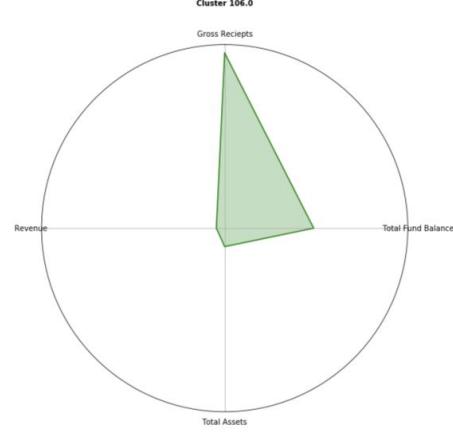
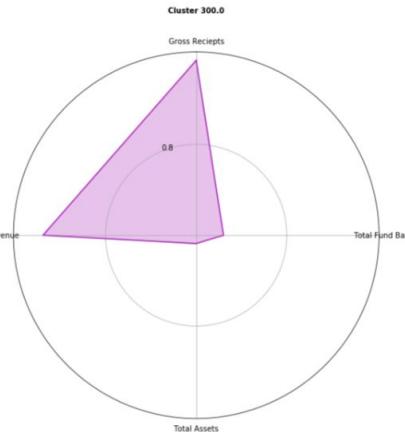
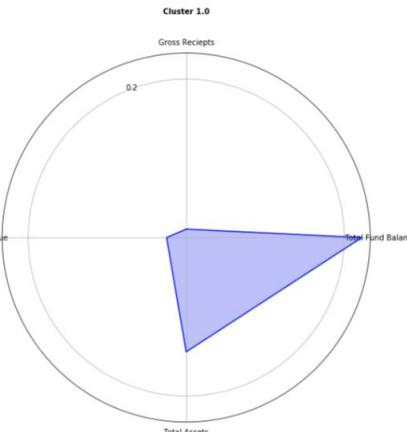
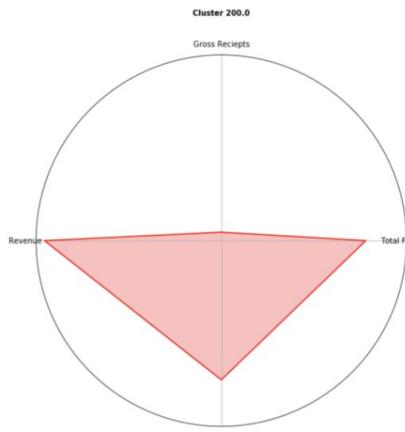
**Hierarchical Clustering + Gap Statistic:** Linkage matrix generation and gap statistic scanning to find optimal number of clusters

**Cluster Merging:** Merge 990EZ clusters with nearest 990 cluster



# Model Validation

Our aim is to maintain diversity across clusters....

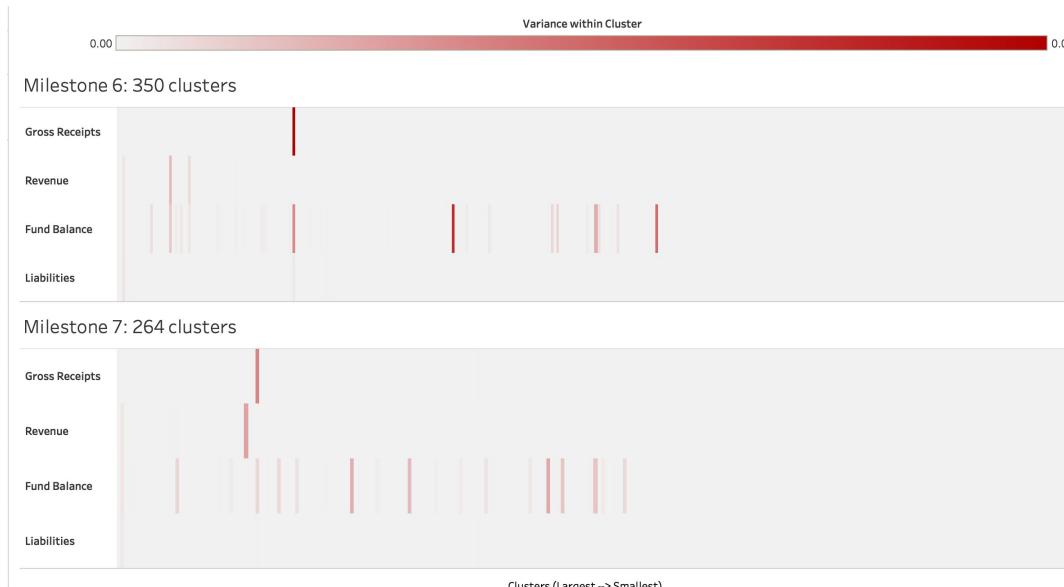


Radar plots similar to those above allow us to describe individual clusters in a visual way. We can also use these plots to verify the diversity of our clusters.



# Model Validation

... while decreasing variance within clusters.

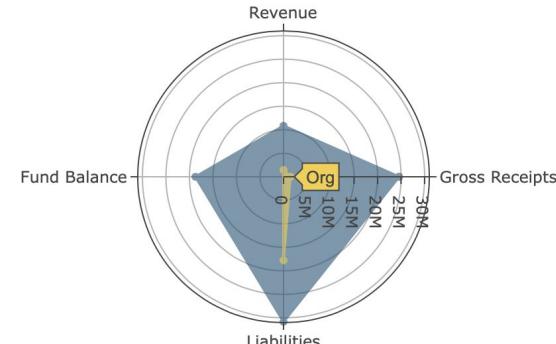
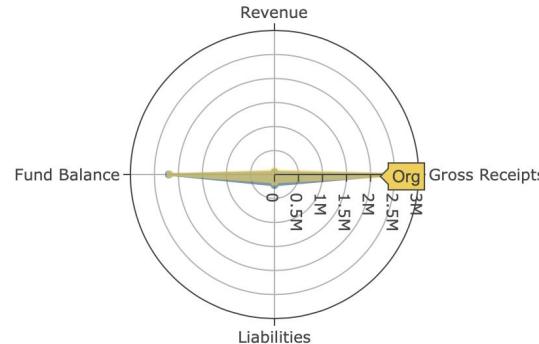
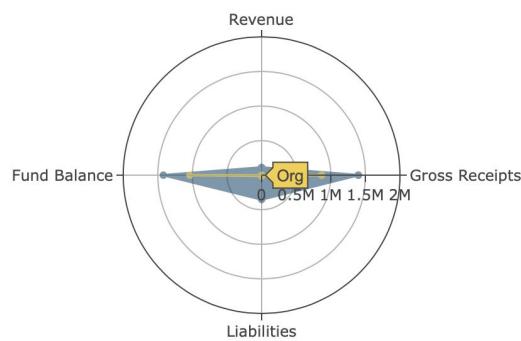


The introduction of new features such as % changes from BoY to EoY and funding allocation ratios cut the number of clusters in half and reduced the number of clusters containing only a single nonprofit.



# Results

**Financial performance of organizations is usually proportional to the mean of the cluster it belongs to.**

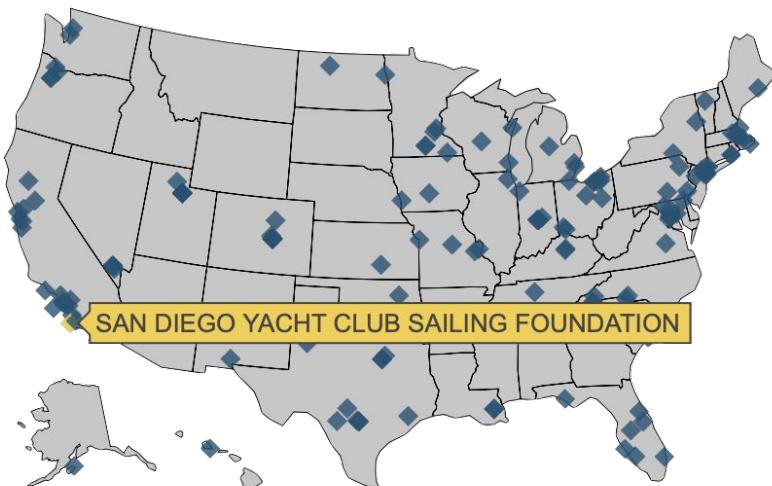


Radar plots comparing orgs to their peer groups show consistencies in shape, but not magnitude, indicating that the ratios between finance metrics were considered more characteristic of financial status than the dollar amounts themselves.



# Geographic

This space is driven by intuition and interpretability.



The geographic space is simply defined by the physical distances between organizations.

- Geocode zip codes to obtain latitude & longitude
- Calculate Haversine distance



# Agenda

1. Introduction
  - a. Solution
  - b. Purpose & Stakeholders
2. The Data
  - a. Data Pipeline
  - b. Communication Plan
3. Model Solutions
  - a. Text Modeling
  - b. Financial Modeling
  - c. Geographical
- 4. Product**
5. Results & Conclusion



# Product

**Our product is an interactive visualization of the 3-dimensional social space.**

## MOTIVATION:

A **single**, easy-to-understand and **interactive** visual that answers the question:

“Which organizations are **most similar to an organization that I care about?**”

While tying together the three main descriptive categories of nonprofit organizations (text, finance, geography)



# Visualization

**Dimensionality reduction is necessary for human interpretability.**

- Text Data - 200 features after embedding + PCA
- Financial Data - 34 features
- Geographical Data - 2 features (latitude / longitude)

To represent these 3 categorical dimensions in a human-interpretable space, we collapse these 3 data sources into a single feature, each:

Distance to centroid!

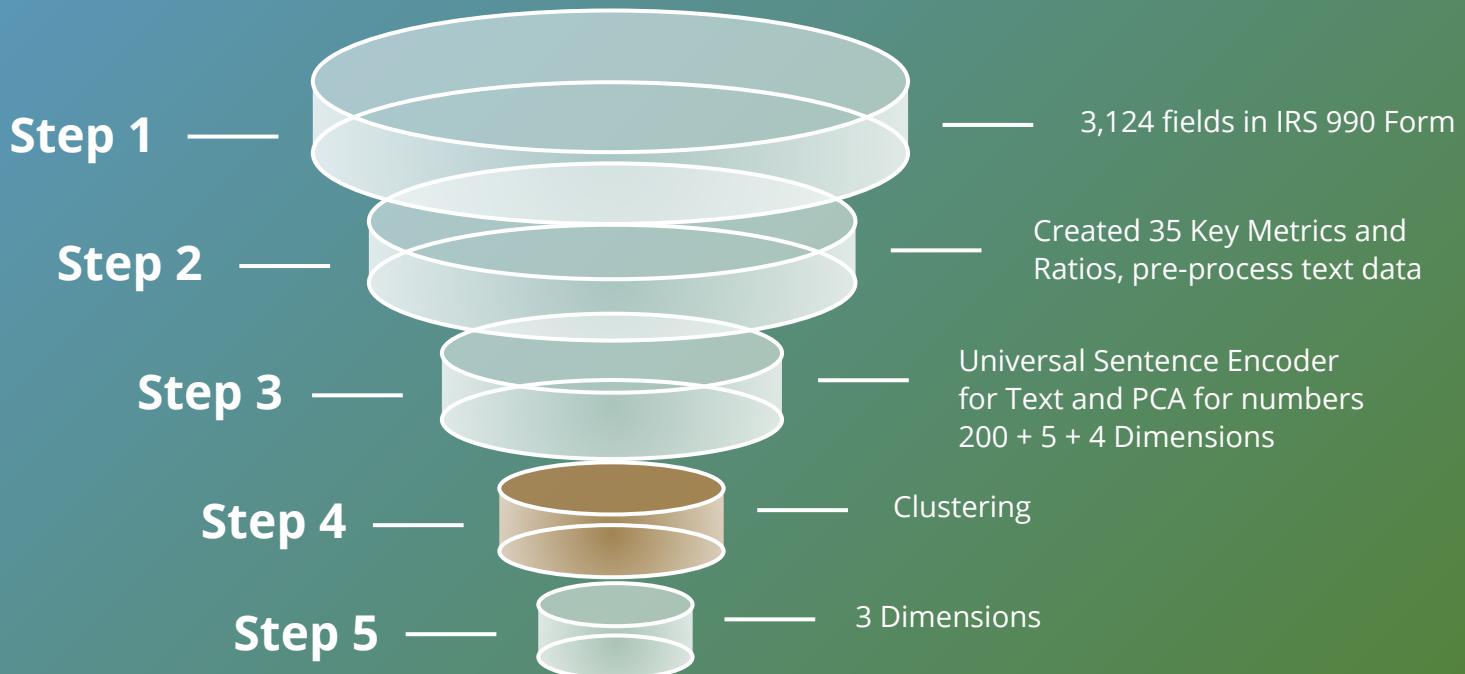


# User Interaction

## How it Works:

1. Select an organization (Org. A)
2. Identify Org. A's cluster in “text space” and find all samples belonging to that cluster (Set A)
3. Identify Org. A's cluster in “financial space”, find all samples belonging to that cluster (Set B), and create the union of Set A and Set B (Set C)
4. For all samples in Set C:
  - a. Calculate their distance in “text space” to the cluster centroid of Org. A
  - b. Calculate their distance in “financial space” to the cluster centroid of Org. A
  - c. Calculate their geographical distance from the location of Org. A
5. Plot distances on a 3-axis visualization (each axis for the 3 metrics in Step 4)

# IRS 990 Form Feature Reduction



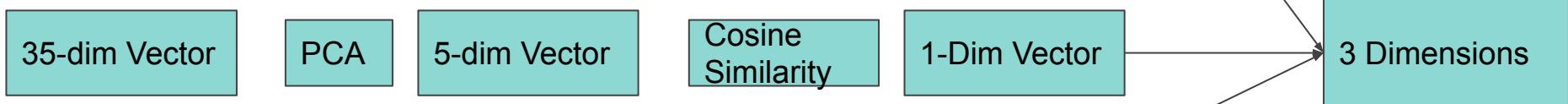


# Dimensionality Reduction Explained

Text Data: Universal Sentence Encoder



Financial Data



Geo Data: Lat, Long

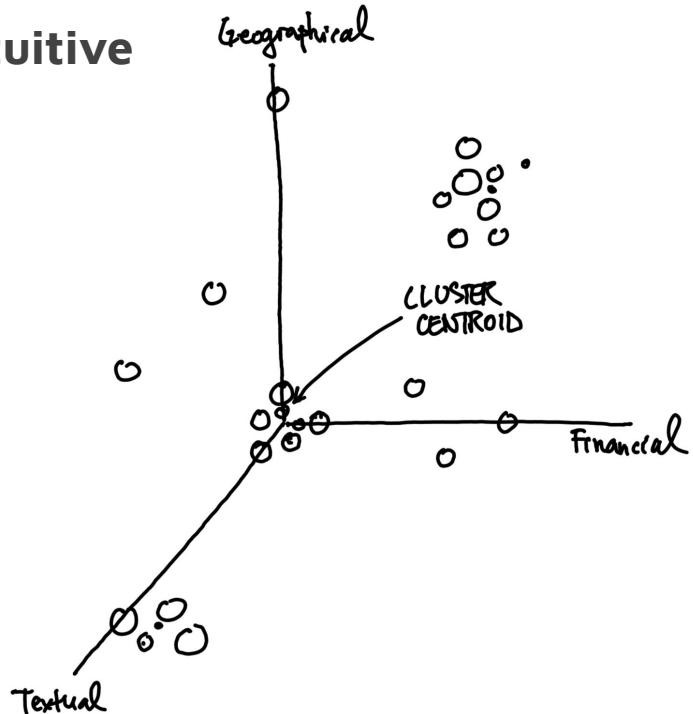




# Visualization

**3-Dimensional scatter plot is most intuitive visual representation.**

- Organizations close to the origin are most typical of Org. A's peer group
- Organizations close to each other that are far away from the origin are close to each other and represent pockets of similar organizations that are loosely related to Org. A



# Visualization Demo



# Agenda

1. Introduction
  - a. Solution
  - b. Purpose & Stakeholders
2. The Data
  - a. Data Pipeline
  - b. Communication Plan
3. Model Solutions
  - a. Text Modeling
  - b. Financial Modeling
  - c. Geographical
4. Product
- 5. Results & Conclusion**



# Business Potential

01



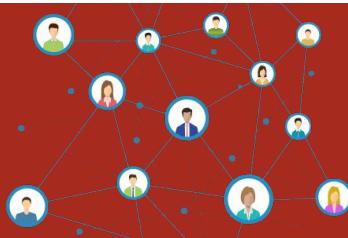
- Able to reach out to similar purpose organizations that are in similar financial standing as WWF

02



- Able to find similar purpose organization as Women's March that is nearby for participation

03



- John searches for the same zip code as his organization to find NPOs who have successfully raised funds for hunger



# Conclusions



- Organizations can be explored using 3 numerical values
- Separating organizations into useful groups
- Useful to model finance and purpose independently
- Non Profit Organizations are unique and challenging

# Questions?