

Classy Data Analysis

Hansen | Nagda | Pimentel | Reyes | Tai



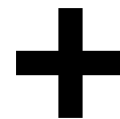


Challenge

We have a 1-D understanding of 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.

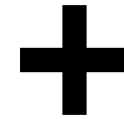
We need a more complex solution to find like-minded organizations.



Solution

... but NPOs are multi-faceted.

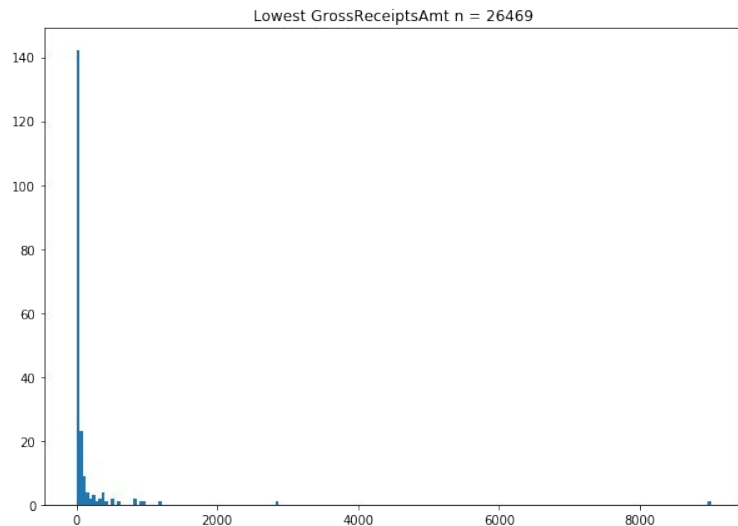
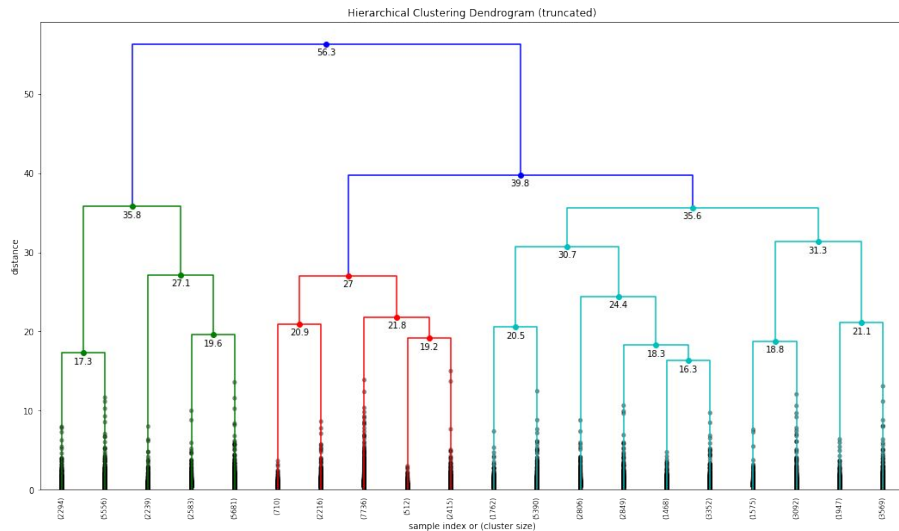
- Lay out a common “social space”, where the organizations that drive social change and potential donors can connect, find organizations and be offered recommendations, and where discovery of new causes - and events within causes (i.e. fundraisers) - could be facilitated.
- Use combination of government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





Results

Clustering models allow for a much wider semantic space.



680 financial clusters and 165 text clusters provide reassurance that each nonprofit will be aligned to a robust, yet more focused peer group.



Agenda

1. Introduction
2. Solution
3. Results for Text Feature Engineering
4. Results for Financial Feature Engineering
5. Next Steps
6. Insights



Results

Deep learning encodings generate more meaningful

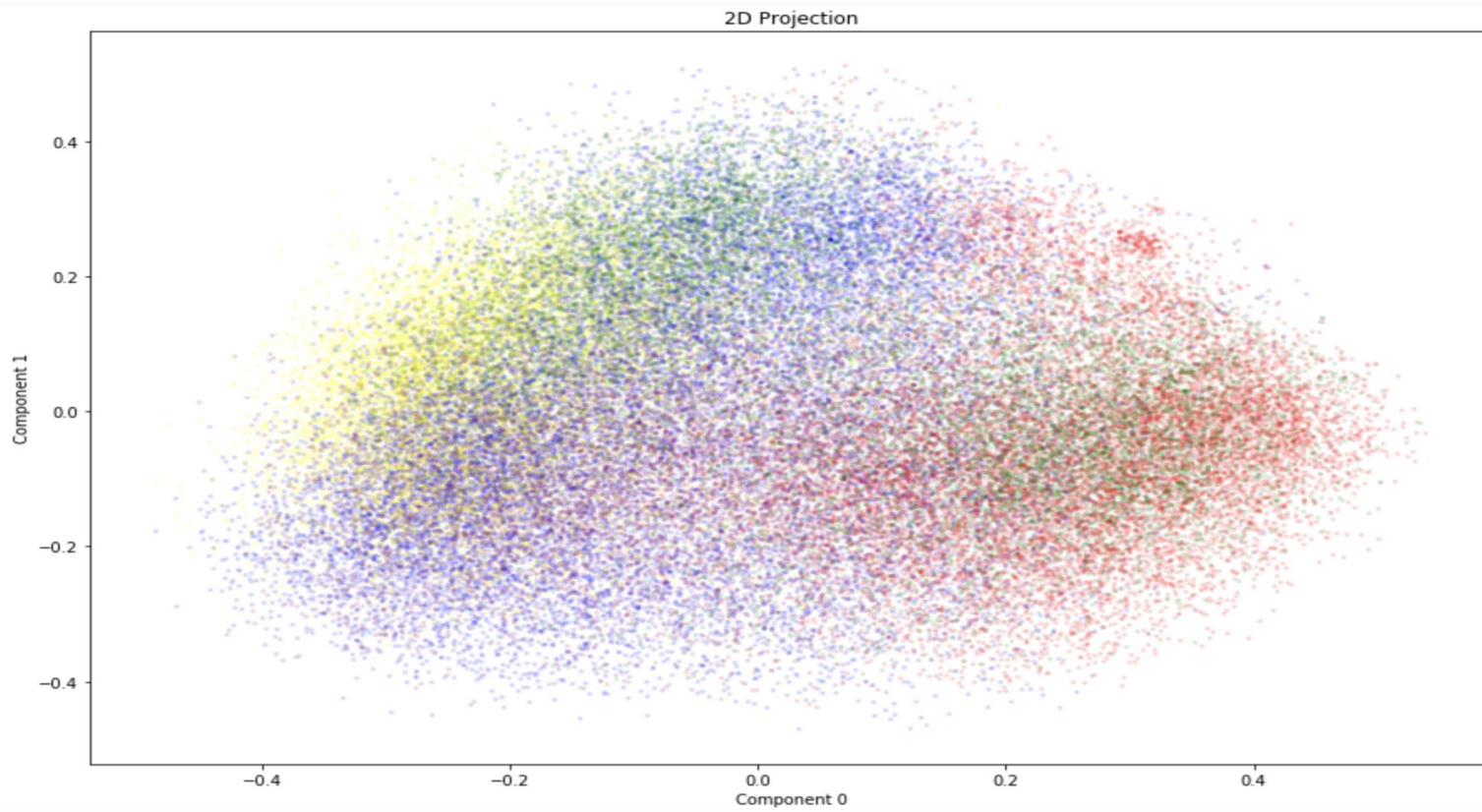
Sample 1:
AIDS TASK FORCE OF LAPORTE AND PORTER CO
Sample 2:
MAUI AIDS FOUNDATION INC
Sample 3:
SIEMPRE UNIDOS
Sample 4:
HEALTHHIV
Sample 5:
BOULDER COUNTY AIDS PROJECT
Sample 6:
WRITERS PLANNERS TRAINERS
Sample 7:
COMMUNITY OUTREACH MEDICAL CENTER
Sample 8:
CASCADE AIDS PROJECT
Sample 9:
CLINICNET
Sample 10:
UNION POSITIVA INC

Sample 1:
INDEPENDENCE FOR THE BLIND OF WEST FLORIDA INC
Sample 2:
JAMAICA OUTREACH PROGRAM INC
Sample 3:
RESTORING SIGHT INTERNATIONAL INC
Sample 4:
VIRGINIA LIONS EYE INSTITUTE FOUNDATION INC
Sample 5:
CALIFORNIA VISION FOUNDATION
Sample 6:
FOUNDATION FOR BLIND CHILDREN
Sample 7:
CENTRAL VALLEY CENTER FOR THE VISION AND HEARING IMPAIRED
Sample 8:
AMERICAN SOCIETY OF RETINA SPECIALISTS
Sample 9:
NATIONAL FOUNDATION FOR EYE RESEARCH
Sample 10:
HOUSTON EYE ASSOCIATES FOUNDATION

Prior to this project, the all twenty of these nonprofits would likely have been lumped into a much larger “Health” sector. Universal Sentence Embeddings provide much more focused results.



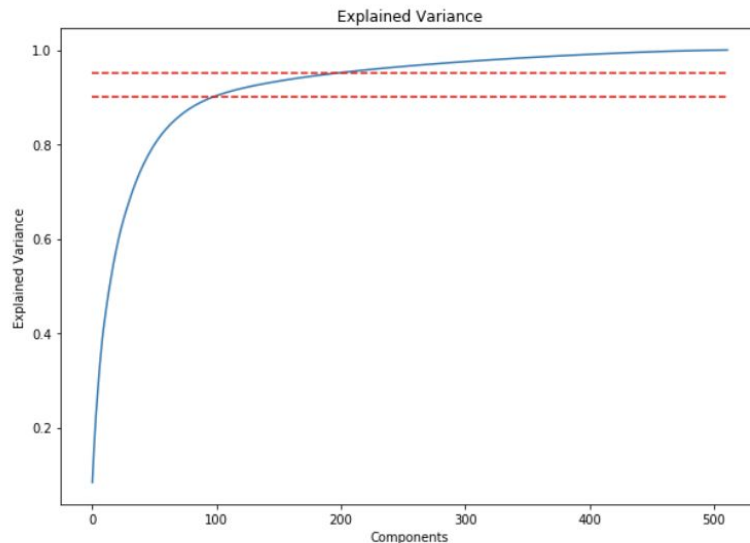
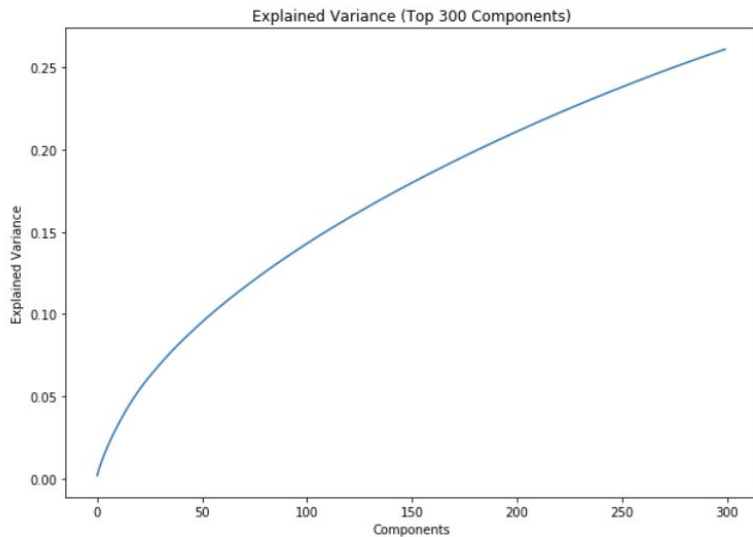
Text Data Clustering in 2-D





Significance of Results

Text clusters aim to *capture* variance.

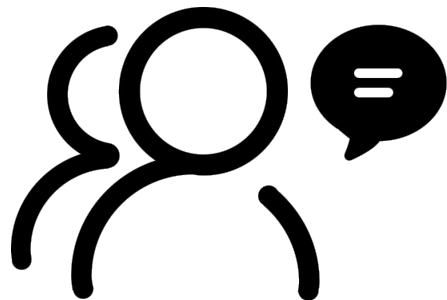


Explained variance of TFIDF embeddings and Tensorflow embeddings



Next Steps for Modeling

Text Features

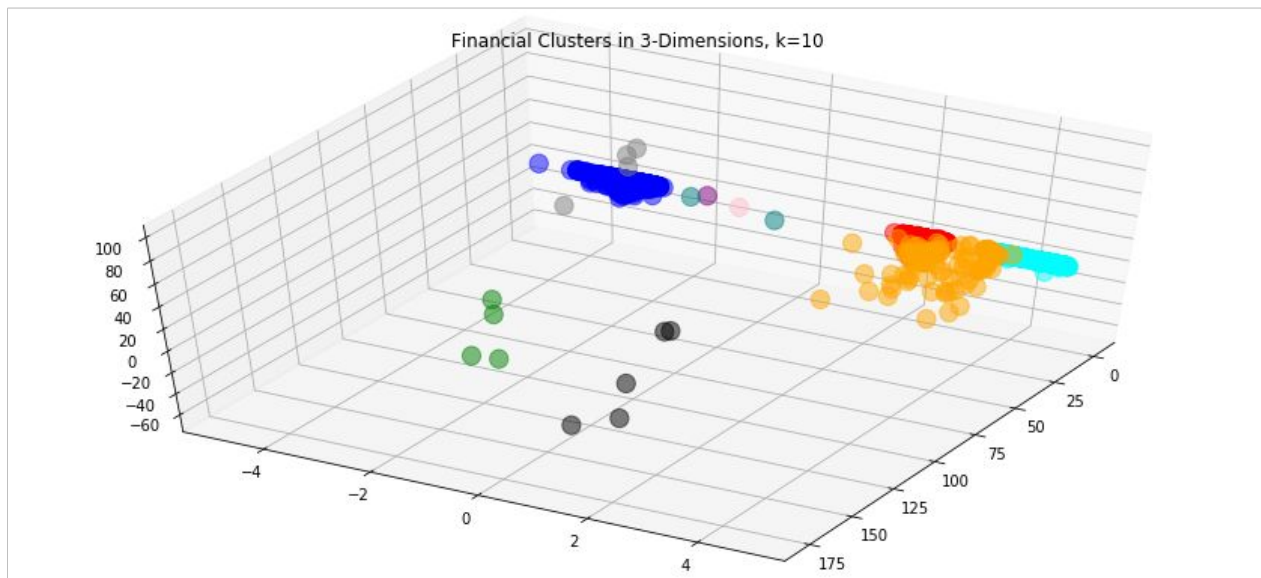


- Other text embedding methodologies (i.e. LDA)
- Ensembling clustering models
- Website data
 - Companies' more-detailed and description "About Us" sections
 - Tax forms contain high level details, lacking specifics
- Social media data
 - Similar motivation to website data with additional perk of (potentially implementing) online cluster updates
 - Web Scraping of Social Media Profiles from Websites



Results

ML algorithms generate more meaningful groupings.

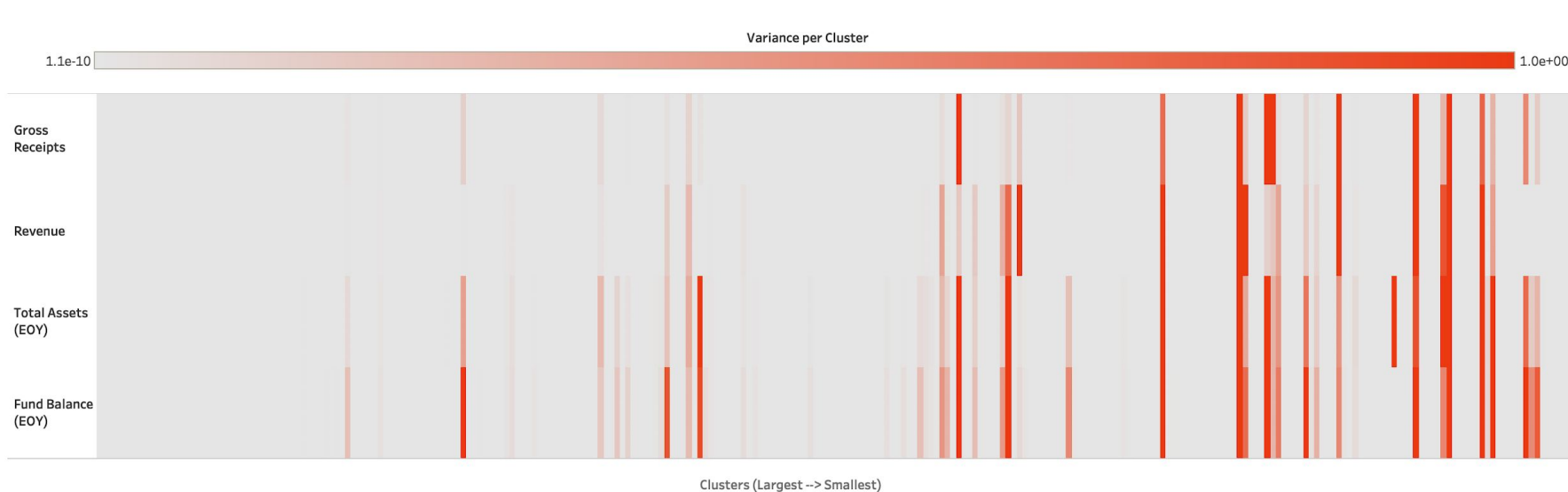


Form 990s and 990EZs provide copious amounts of financial data points. Our feature engineering process helps to block out the noise and generate information out of a wall of numbers.



Significance of Results

Financial clusters aim to *reduce* variance.

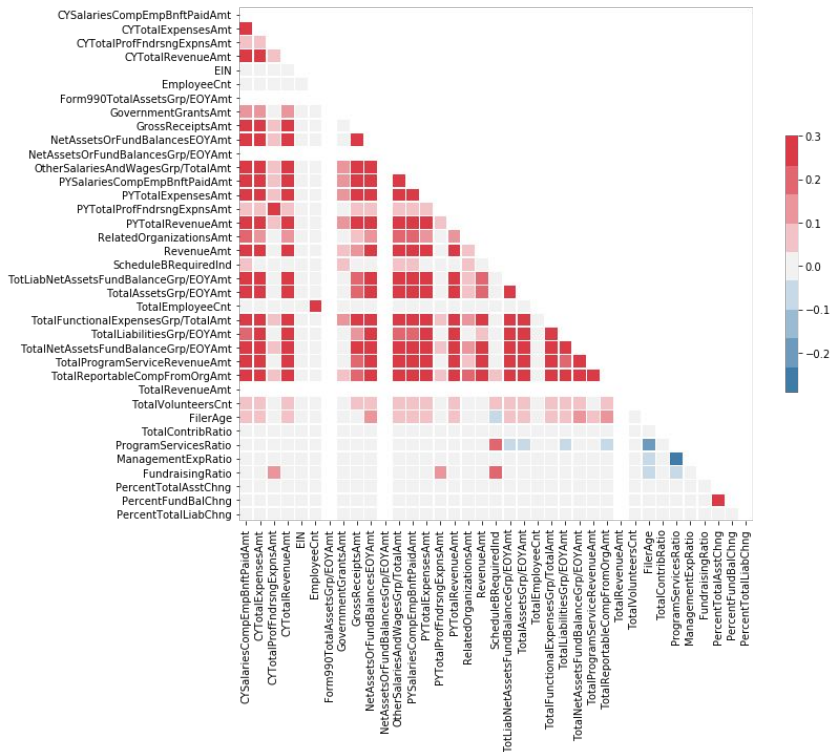


The largest clusters have very low variance in these key fields. This is an indication of “good” groupings. As the cluster size decreases, variance tends to increase, indicating that we may need to add more data to ensure similarity within groups.



Correlation of Financial Data

- To reduce variance in clusters, examined if financial features were related
- Found that there are key ratios that can be created to explain correlated information.

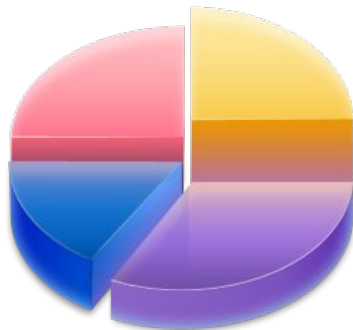
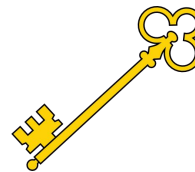




Next Steps for Modeling

Feature engineering is key for financial data.

- Organizations have key metrics for financial standing, need to add these to separate Organizations. Need to account for size and scale of organization
 - Ex: Ratio of expenses spent on fundraising
- More strategic approach to choosing thresholds and partitioning 990 data
 - Split by Form Type (already split by Gross Receipts implicitly)



The Forms

Any Non-Private Foundation

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

Private Foundations

990 Return of Organization Exempt From Income Tax

OMB No. 1545-0047

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.

2017 Open to Public Inspection

Department of the Treasury Internal Revenue Service

A For the 2017 calendar year, or tax year beginning 2017, and ending 20

B Check if applicable:

☐ Address change

☐ Name change

☐ Initial return

☐ First return/return after

☐ Amended return

☐ Application pending

C Form of organization:

☐ Sole business

☐ Other business

D Employer identification number

E Room/suite

F Telephone number

G Gross receipts \$

H If this is a group return for additional filers, check the following:

☐ Yes ☐ No

I Tax-exempt status: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

J Website: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

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 VII, line 1a) 3

4 Number of independent voting members of the governing body (Part VII, line 1b) 4

5 Total number of individuals employed in calendar year 2017 (Part V, line 2g) 5

6 Total number of volunteers (estimate if necessary) 6

7a Total unrelated business revenue from Part VIII, column (c), line 12 7a

7b Net unrelated business taxable income from Form 990-T, line 34 7b

Revenue

8 Contributions and grants (Part VIII, line 1h) 8

9 Program service revenue (Part VIII, line 2g) 9

10 Investment income (Part VIII, column (A), lines 3, 4, and 7d) 10

11 Other revenue (Part VIII, column (A), lines 5, 6d, 8c, 9c, 10c, and 11e) 11

12 Total revenue—add lines 8 through 11 (must equal Part VIII, column (A), line 12) 12

Expenses

13 Grants and similar amounts paid (Part IX, column (A), lines 1–3) 13

14 Benefits paid to or for members (Part IX, column (A), line 4) 14

15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5–10) 15

16a Professional fundraising fees (Part IX, column (A), line 11e) 16a

16b Total fundraising expenses (Part IX, column (D), line 25) 16b

17 Other expenses (Part IX, column (A), lines 11a–11d, 11f–24e) 17

18 Total expenses. Add lines 13–17 (must equal Part IX, column (A), line 25) 18

19 Revenue less expenses. Subtract line 18 from line 12 19

20 Total assets (Part X, line 16) 20

21 Total liabilities (Part X, line 26) 21

22 Net assets or fund balances. Subtract line 21 from line 20 22

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 _____

Preparer's signature _____ Date _____

Check ☐ if self-employed

Print name _____ 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. 11982Y Form 990 (2017)

990-EZ Short Form Return of Organization Exempt From Income Tax

OMB No. 1545-1102

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.

2017 Open to Public Inspection

Department of the Treasury Internal Revenue Service

A For the 2017 calendar year, or tax year beginning 2017, and ending 20

B Check if applicable:

☐ Address change

☐ Name change

☐ Initial return

☐ First return/return after

☐ Amended return

☐ Application pending

C Form of organization:

☐ Sole business

☐ Other business

D Employer identification number

E Room/suite

F Telephone number

G Gross receipts \$

H If this is a group return for additional filers, check the following:

☐ Yes ☐ No

I Tax-exempt status: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

J Website: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

K Form of organization: ☐ Corporation ☐ Trust ☐ Association ☐ Other

L Year of formation

M State of legal domicile

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 ☐

Revenue

1 Contributions, gifts, grants, and similar amounts received 1

2 Program service revenue including government fees and contracts 2

3 Membership dues and assessments 3

4 Investment income 4

5a Gross amount from sales of assets other than inventory 5a

5b Less: cost or other basis and sales expenses 5b

5c Gain or (loss) from sales of assets other than inventory (Subtract line 5b from line 5a) 5c

6 Gaming and fundraising events

6a Gross income from gaming (attach Schedule G if greater than \$15,000) 6a

6b Gross income from fundraising events (not including contributions) 6b

6c Less: direct expenses from gaming and fundraising events 6c

6d 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

7b Less: cost of goods sold 7b

7c 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 Benefits 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

Expenses

13 Grants and similar amounts paid (Part IX, column (A), lines 1–3) 13

14 Benefits paid to or for members 14

15 Salaries, other compensation, and employee benefits 15

16a Professional fees and other payments to independent contractors 16a

16b Total fundraising expenses (Part IX, column (D), line 25) 16b

17 Other expenses (Part IX, column (A), lines 11a–11d, 11f–24e) 17

18 Total expenses. Add lines 13–17 (must equal Part IX, column (A), line 25) 18

19 Revenue less expenses. Subtract line 18 from line 12 19

20 Total assets (Part X, line 16) 20

21 Total liabilities (Part X, line 26) 21

22 Net assets or fund balances. Subtract line 21 from line 20 22

990-PF Return of Private Foundation or Section 4947(a)(1) Trust Treated as Private Foundation

OMB No. 1545-0052

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.

2017 Open to Public Inspection

Department of the Treasury Internal Revenue Service

A For the 2017 calendar year, or tax year beginning 2017, and ending 20

B Check if applicable:

☐ Address change

☐ Name change

☐ Initial return

☐ First return/return after

☐ Amended return

☐ Application pending

C Form of organization:

☐ Sole business

☐ Other business

D Employer identification number

E Room/suite

F Telephone number (see instructions)

G Gross receipts \$

H If this is a group return for additional filers, check the following:

☐ Yes ☐ No

I Tax-exempt status: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

J Website: ☐ 501(c)(3) ☐ 501(c)(4) ☐ 501(c)(29) ☐ 527

K Form of organization: ☐ Corporation ☐ Trust ☐ Association ☐ Other

L Year of formation

M State of legal domicile

Part I Analysis of Revenue and Expenses (The total of amounts in columns (b), (c), and (d) may not necessarily equal the amount in column (a) (see instructions).)

(a) Revenue and expenses per source

1 Contributions, gifts, grants, etc. received (attach schedule) 1

2 Check ☐ if the foundation is not required to attach Sch. B 2

3 Interest on savings and temporary cash investments 3

4 Dividends and interest from securities 4

5a Gross rents 5a

5b Net rental income or (loss) 5b

6a Net gain or (loss) from sale of assets not on line 10 6a

6b Gross sales price for all assets on line 6a 6b

7 Capital gain net income (from Part IV, line 2) 7

8 Net short-term capital gain 8

9 Income modifications 9

10a Gross sales less returns and allowances 10a

10b Less: cost of goods sold 10b

10c Gross profit or (loss) (attach schedule) 10c

10d Other income (attach schedule) 10d

12 Total. Add lines 1 through 11 12

13 Compensation of officers, directors, trustees, etc. 13

14 Travel, conferences, and meetings 14

15 Pension plans, employee benefits 15

16a Legal fees (attach schedule) 16a

16b Accounting fees (attach schedule) 16b

16c Other professional fees (attach schedule) 16c

17 Interest 17

18 Depreciation (attach schedule) and depletion 18

20 Occupancy 20

21 Travel, conferences, and meetings 21

22 Printing and publications 22

23 Other expenses (attach schedule) 23

24 Total operating and administrative expenses. Add lines 13 through 23 24

25 Contributions, gifts, grants paid 25

26 Subtract line 26 from line 12 26

27 Excess of revenue over expenses and disbursements 27

28 Net investment income (if negative, enter -0-) 28

29 Adjusted net income (if negative, enter -0-) 29

(b) Net investment income

(c) Adjusted net income

(d) Disbursements for charitable purposes (attach schedule)

Changes to Data Pipelines & Solutions Architecture

- Many websites reference their facebook, twitter, linkedin, youtube accounts and these can be invaluable to gather 360 degree information about a company. Extracting Social Media from their websites for the organizations will allow the exploration of social aspects of the organization.
- Data needs to be collected in a text warehouse that can easily be transformed

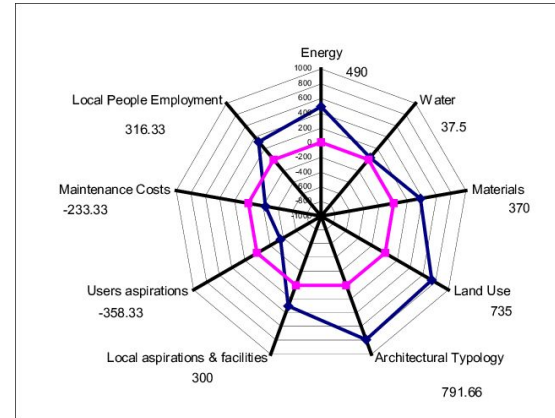




Next Steps in Product Design

Measures of “similarity” now chosen by end-user.

- Decide how to display the information to the user
 - Bring together clusters in meaningful ways for user specifications
- Create visualizations
- Radial graph to describe clusters





Modeling text and financial data separately...

- Improves accuracy of *both* models
- Enhances user experience by providing more options to the end user
- See interesting weakly related groups with just text from form 990

The inclusion of more features...

- through addition of social media and/or website data will provide even more information about the who, what, when, where, why and how of the nonprofits, and therefore create stronger links between them
- through feature engineering of financial data should unveil currently overlooked patterns to produce more definitive clusters

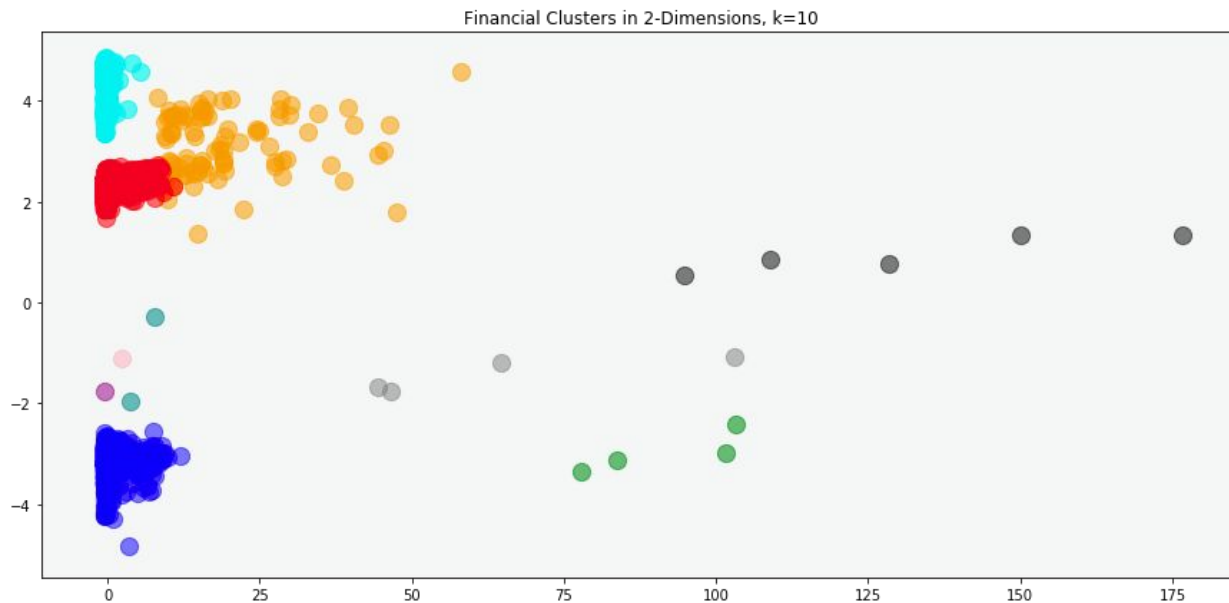


Questions?



Results

ML algorithms generate more meaningful groupings.



Form 990s and 990EZs provide copious amounts of financial data points. Our feature engineering process helps to block out the noise and generate information out of a wall of numbers.

Requirements

Audience:

Data Science and Product Teams

Main points to be made

- How accurate/significant are the results?
- What are the main insights so far?
- What step in product design do you recommend based on these results?
- How will this affect your data pipelines and solution architecture so far?
- What are next steps for modeling based on the progress and why?

Don't forget

to include your team, problem definition and data definitions in the beginning of the presentation. Think story lines in the captions!



The Data - Where From?

Open Data on AWS

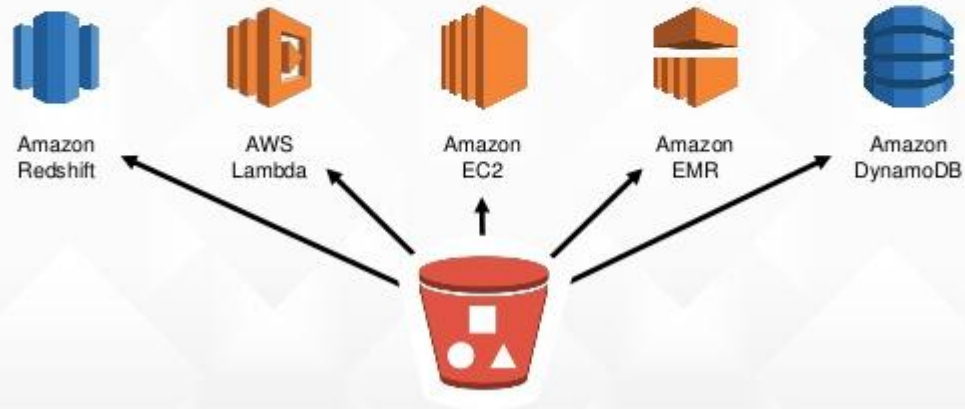
Share any volume of data with as many people as
you want

The banner features a teal background with white clouds and a faint grid pattern. The text 'Open Data on AWS' is prominently displayed in white. Below it, the phrase 'Share any volume of data with as many people as you want' is written in a smaller white font. In the bottom right corner, there is a small logo for the IRS Form 990, which includes the IRS seal and the text 'IRS Form 990'.

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

The Data - Open Data on AWS





Making data open on AWS enables more innovation by making data available for rapid access to our flexible and low-cost computing resources.



The Data - Data Integration Framework



The Data - Looks Like? MongoDB

 (1) 042662873	{ 10 fields }	Object
 _id	042662873	String
 DLN	93493243000066	String
 EIN	042662873	String
 FormType	990	String
 LastUpdated	2017-01-11T22:15:15	String
 ObjectId	201612439349300006	String
 OrganizationName	ELKS BUILDING CORP OF NORWOOD	String
 SubmittedOn	2017-01-04	String
 TaxPeriod	201603	String
 URL	https://s3.amazonaws.com/irs-form-990/...	String

The Data Pipeline Components

