Analytics

NLP for Social Good: Contrasting Approaches to Text Classification

Anna Bethke
Data Science Fellow & Project Lead
Delta Analytics

anna.bethke@intel.com
twitter: @data_beth

Sean McPherson
Data Science Fellow & Project Lead
Delta Analytics

sean@deltanalytics.org
twitter: @PhaseAnalytics

Agenda

- Intro Delta Analytics
- Teaching Machine Learning for Social Good
- Open Media Foundation
- Career Village
- Approaches to Text Classification

Analytics

Who We Are

Delta Analytics is...

An organization...

- Founded in 2013
- Based in SF Bay Area
- Dedicated to using Data for Change

A community of...

Professionals in

- Analytics
- Data Science
- Design
- Marketing

working across a wide variety of industries.

Who are...

Passionate about volunteering their skills to close the skill gap in the social sector and drive meaningful, sustainable impact.

Through...

Two programs

1. Providing pro-bono data consulting services to nonprofits.

2. Developing and teaching a machine learning curriculum to cultivate local capacity.

Since 2013...

Completed **29 projects** with nonprofits and social impact organizations

Beta tested an **open source curriculum** with **30 students** in Nairobi, Kenya

Recruited **70+ professionals** who have volunteered thousands of hours consulting and developing data for good resources

Donated over **15,000 hours**

$0 charged for services

Where our Fellows Work

Our Programs

# Delta Analytics runs two core programs

Data Service Grant

Bridging the nonprofit skill gap by enabling nonprofits to **accelerate their impact** through leveraging **skilled data professionals** on project opportunities.

Machine Learning for Good

Empower anyone, anywhere to leverage data for good in their communities by **building technical capacity** around the world.

Data Service Grant
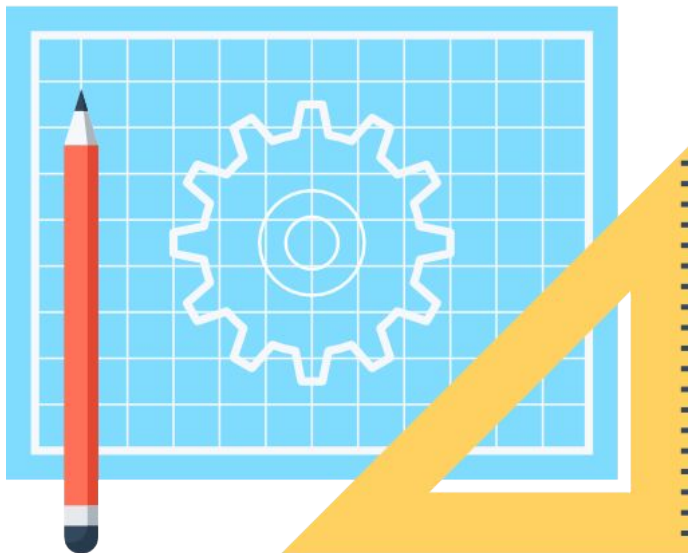
- Closing the skill gap through connecting nonprofits with skilled professionals on data projects

- 6 month service grants to 5-10 nonprofit organizations per year

- Each project is staffed with 3-5 data professionals

- Fellowship Program offers volunteers community and professional development

Data Service Grant - Types of Projects

	Analytics	Data Science	Data Tools
Problem Type	• Operational analytics • Impact evaluation	• Predictive application	• Improve existing data processes
Deliverable	• Dashboard • Report	• Models • Software Infrastructure	• Deployed tool • Potentially open source
Case Study	BUILD TECHNOSERVE	RAINFOREST CONNECTION	TBD

INTRO - DELTA	ML for Social Good	OMF	Career Village	Text Classification

Note: Case studies in appendix

Data Service Grant - Past Recipients

Data Service Grant - Current Projects

Why Machine Learning for Good?

1) Machine learning is an incredibly powerful tool

Like all powerful tools machine learning requires **responsibility & accountability**

We believe the face of data science has to be **more representative** of the world it serves

2) Delta has been working for the last 3 years to help non-profits around the world

Our **non-profit grant recipient program is small** and selective. We only work with 7-10 nonprofit grant recipients every year

There is an incredible amount of **unmet demand.** We cannot help all the non-profits and communities that reach out to us for help

Open Source Curriculum

Open source curriculum developed to train students with beginner coding proficiency

Our code is **Publically Available** and **Open Source**

Help us spread the word! bit.ly/social-good-ml

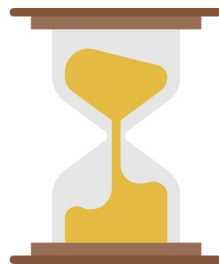| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Open Source Curriculum

All our curriculum is **public**. We welcome **comments** and have **more than 300 remote teachers** from all over the world.

You covered this today:

- ✓ What is machine learning?
- ✓ How do you define a research question?
- ✓ What are observations?
- ✓ What are features?
- ✓ What are outcome variables?
- ✓ Introduction to KIVA data

Data Cleaning → Time series

If we have observations over time, we need to do time series validation.

Ask yourself:
a. Is the data for the correct time range?
b. Are there unusual spikes in the data over time?

What should we do if there are unusual spikes in the data over time?

Link to all slides here: http://www.deltanalytics.org/curriculum.html

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Machine Learning for Good - Kenya Pilot

- Partnered with Moringa School in Nairobi, Kenya for a **30 student pilot**

- We taught a night class from 6:00 pm to 9:00pm 2x a week. We had work sessions every Saturday.

- Many of our students travelled for **more than an hour after work to attend class.**

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Machine Learning for Good - Morocco

In the next few months we will teach classes in Morocco and in the local bay area near San Francisco.

Université Ibn Zohr

L'UNIVERSITÉ DE L'EXCELLENCE

En tant qu'espace d'échange et d'ouverture d'esprit,
l'Université Ibn Zohr fait de l'excellence un levier de développement humain
respectueux de la diversité culturelle et humaine de la société marocaine

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Open Media Foundation

Delta Team

Anna Bethke
Head of AI4Good
Intel Corporation

Zong Huang
Research Analyst
The Brattle Group

Graham McAlister
Data Scientist
Esurance

Sydney Wong
Antitrust and Competition
Economics
Charles River Associates

Putting the power of the media and technology in the hands of the people.

SERVICES

We help socially-conscious organizations use the power of media to carry out their mission and vision.

Start a Project

EDUCATION

We offer a variety of classes to help you create your own media and make it more effective.

Take a Class

TOOLS

We provide access to state-of-the-art media tools and high-end technology resources at affordable rates.

Rent Equipment/Studios

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Open Media Project ^{0.2.2}

[Base url: open.ompnetwork.org/api]

https://raw.githubusercontent.com/omfound/ompgov-api/master/ompgov-api-0.2.2-swagger.yaml

OMP BETA API documentation for accessing government meeting data (city councils, state legislature, school boards) using the Open Media Project (OMP). You can learn more about the OMP at http://www.ompnetwork.org. Authentication is not currently required but an API key will be required in the future.

Contact the developer

Schemes

HTTPS ▾

session   Government meeting sessions   ⌄

GET	/sessions Fetches sessions

GET	/sessions/{sessionId} Find session by ID

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Enhancing Meeting Transparency

Create an automatic tagging system to enable:

- Easier searching
- Insight into key discussions
- Analysis and tracking of discussions

Career Village

Delta Team

Stephen Hsu
Data Science Student
University of San Francisco

Sean McPherson
Data Scientist
Phase Analytics

Seema Misha
Data Analyst
Varian Medical

Yiquin Yang
Data Science Student
University of California, Davis

CareerVillage.org

CareerVillage.org crowdsources answers to career questions posed by underserved youth. The key factor that matches volunteers to questions to answer **is** which tags are applied to the question. The goal of our project **is** to improve the precision **and** recall of the system that matches tags to questions.

How do I get started in the field of data science/big data/machine learning?

Title

I'm a third year computer science major, and I've been trying to learn more about data science/big data/machine learning. There aren't any classes offered at the college that I'm attending that cover these topics, and I'm a little lost as to how I would get started in any of these fields. What are some resources that I could use to learn more about these fields and get a taste of what they're like? Any help would be appreciated.

Text

Topics I follow

#computer-software #data-science #data-analysis #big-data #machine-learning #software-engineering #engineering #analytics #computer-engineering

Tags → **Tags** → #United States of America   #phd   #signal-processing   #electrical-engineering   #data-visualization   #data-science

Albert P.
Student   Los Angeles, California   1841 Karma   25 Questions
You are both in Californie.

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Contrasting Approaches to Text Classification

Text Classification Approaches

OMF

Data cleaning

Removing "Fluff"

Unsupervised text clustering – LDA

Rule-based label algorithm

Aggregate labels to session level

Career Village

Data cleaning

Label cleaning

Supervised text recommendation

Data augmentation

Real-time text recommendation

Data Cleaning Steps for NLP

Necessary

- Lowercase

Potentially Helpful

- Remove stopwords
  - The, and, a, etc.
- Remove non-letters
  - Numbers, characters, etc
- Lemmatization

Our Current Python Packages

- nltk
- spacy
- re
- gensim
- sklearn
- pandas
- numpy
- matplotlib

Removing the "Fluff"

- Noticed that there were several LDA labels that were too "fluffy" - parliamentary proceedings, people thanking/welcoming each other, etc
- We hand-labeled 10,000 captions as fluff or not fluff
- Then used the best (most predictive) labels to remove as much fluff from our data as possible

```
In [*]: generate_labels(df, 100)
```

```
2 / 100
reason we took this on other than a have really needed to be done as we have found in many he
arings over the years together about as that and it probably should is is i really wanted it
to be
```

```
Fluff? 1
```

```
In [7]:  # 1 is fluff (so meeting aspects w/o a clear topic)
         # 0 is not fluff
         # s is spanish
         # 42 is unknown
```

Did removing fluff improve the results?

Some! It will likely be an iterative approach

- Topics are less fluffy
- Helped determine additional stopwords to remove
- Lot of "zoning" topics
- Some topics have more than one focus still

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Label Cleaning

Problem: Over 8000 labels!

- Fix slight variations - major vs majors
- Ignore irrelevant labels (details later)

Problem: ~5% of examples without labels

- Pseudo labeling

Train Model -> Generate Rec labels -> Retrain model w/ additional examples

Problem: Poorly labeled training data

- Hand labeled "Gold" training set
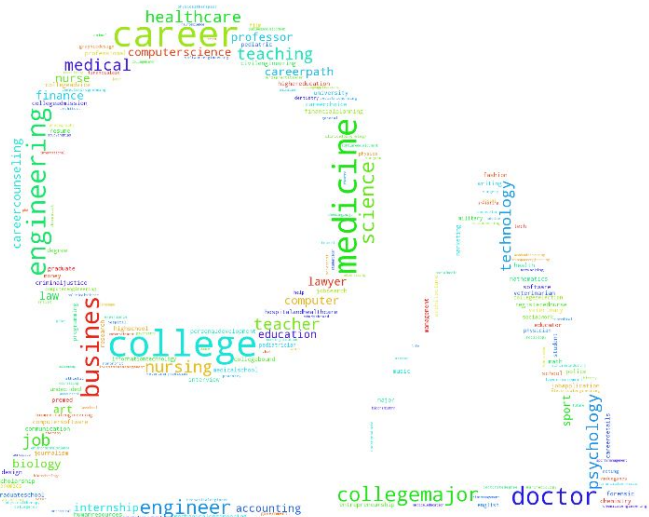
I want become a data analyst ? which is suitable subjects i need to study

m 10th class student, guide me for career...

#any

- What happens to the predictions as we train on more data?
    - First run: Train 8k, Validation 2k
    - Second run: Train 60k, Validation 5k
    - Third run: Train 220k, Validation 8k

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |
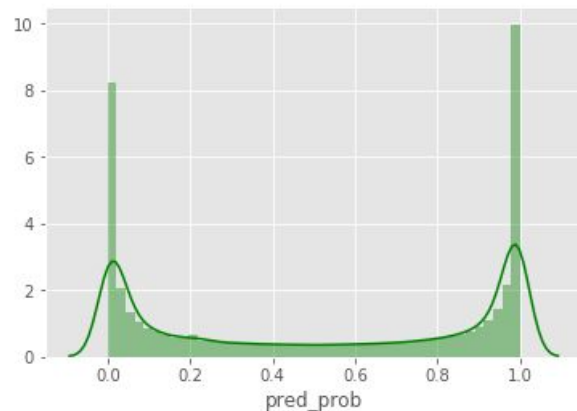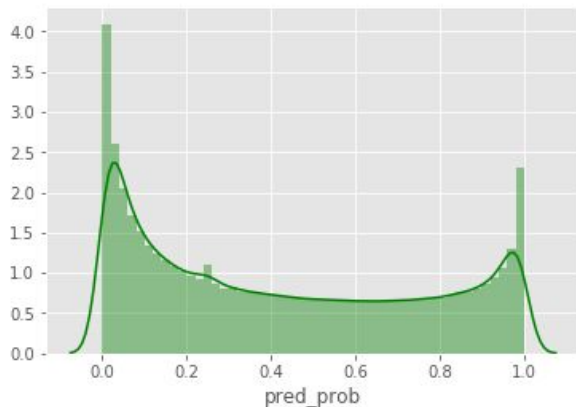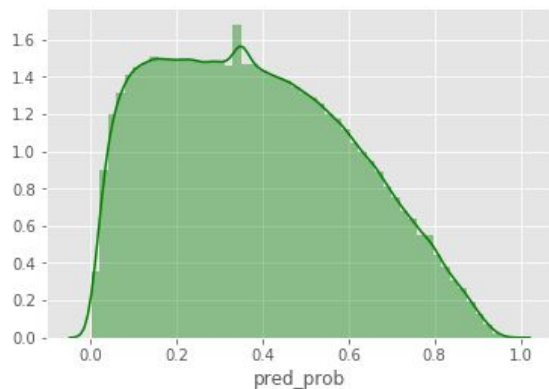
Original Topic Top Words

10/20 informative labels

```
Topic #0:  music license authority applause property one owner going dogs dog
Topic #1:  traffic going bike road one street theres were streets highway
Topic #2:  commission committee board going questions one want time meeting us
Topic #3:  design code standards santa one rules were guidelines going theres
Topic #4:  services health police community department human county officers also mental
Topic #5:  years year one time months last ago two were going
Topic #6:  plan urban process were renewal goals going one community plans
Topic #7:  housing land development growth affordable units were need density eugene
Topic #8:  county state city law federal government local states oregon lane
Topic #9:  homeless health insurance one going were cities community theyre also
Topic #10: yeah want going go right were one say something little
Topic #11: water food force going were task trees theres one things
Topic #12: public hearing record speak questions hearings applicant comment city comments
Topic #13: school kids community one years day students city thornton children
Topic #14: dollars million percent money year thousand fund hundred budget tax
Topic #15: were things going us want community make way lot kind
Topic #16: you mayor that followed mìxsica name good it want okay
Topic #17: motion second item vote council favor discussion move agenda meeting
Topic #18: property parking area building feet street site space south one
Topic #19: going session yeah meeting good one want time right last
```

New Topic Top Words

15/20 informative labels

```
Topic #0:  rights food marijuana vested human city tobacco retail choose make
Topic #1:  information community performance make way also public group going health
Topic #2:  feet right yeah one foot acres road side square street
Topic #3:  zoning property zone park neighborhood open space site properties one
Topic #4:  school neighborhood food students business community make also businesses want
Topic #5:  planning commission make zoning final decision plan zone review change
Topic #6:  yeah fireworks id one make said want going good go
Topic #7:  partnerships also city community park weve plan economic services partners
Topic #8:  tax yeah way make go city going pay right fee
Topic #9:  urban zoning zone growth land plan industrial boundary site residential
Topic #10: also windows lot one car camping city use traffic place
Topic #11: resolution item motion city calendar 2017 agreement agenda council approve
Topic #12: you followed next mayor claire okay vinis good alan here
Topic #13: fees waste fund water city impact funds fee one look
Topic #14: carbon services transportation impacts health also one treatment weve way
Topic #15: renewal urban district riverfront downtown chair city agency council vice
Topic #16: housing economic also public services development quiet airport county safety
Topic #17: code city application plan criteria zoning ordinance standards site section
Topic #18: dollars tax percent million hundred thousand sales ten six revenue
Topic #19: one day bus said make ago every also right go
```

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Supervised Label Recommendation

Goal: fast, simple to use, accurate recommendation system

- FastText by Facebook

Problem: Small training set

- Pretrained word vectors

Constraint: Real-time model deployed using AWS Lambda microservice

- 50 to 250 MB limit on uploads!

Adding A Rule Based Component

Problem: Some topics (like climate change) don't occur frequently enough to make their own automatic label, but are important and DO occur

Solution:

- Create a list of topics with distinct lists of associated words
    - Agriculture: chicken, egg, farming, agriculture, manure, xeriscape
    - Utilities: electricity, energy, fiber, power, solar, utilities
- Compare topic words to words in comments using word embeddings, to create topic probabilities

TODO: Determine best ways to aggregate automatic labels and rule-based labels
          Incorporate phrases instead of single words (?)

Data Augmentation

Label Frequency in Training Data
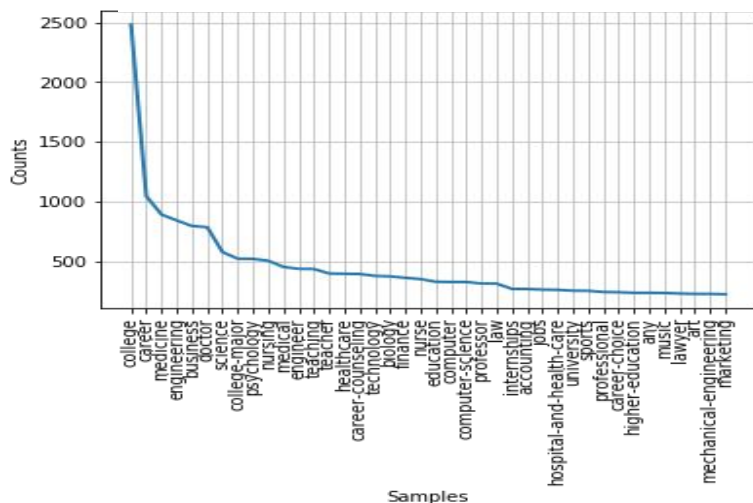
- Label frequency distribution shows strong imbalance in data
- Thesaurus-based data augmentation was performed which generates additional examples for infrequent labels by substituting synonyms of important keywords in example
- Solution: create additional example of useful labels, and allow classification model to ignore irrelevant labels with low frequency
  - #deathisapreferablealternative
  - FastText: -minCountLabel

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

Aggregating Comment Topics Up

- Take most confident comment labels and aggregate their probabilities to the agenda and session level

Session 1: zoning, urban planning

Session 2: roads, agriculture, budget

Session 3: housing, audience participation

| INTRO - DELTA | ML for Social Good | OMF | Career Village | Text Classification |

OMF

Visualizing a Sessions Topics

Budget, Building, Public Services

Real-Time Text Recommendation

System recommends labels as user types

Model trained on complete title + text

Problem: Unexpected results for short or incomplete entries

- Train with incomplete records and #incomplete label
- Engineer system to recognize incomplete entries

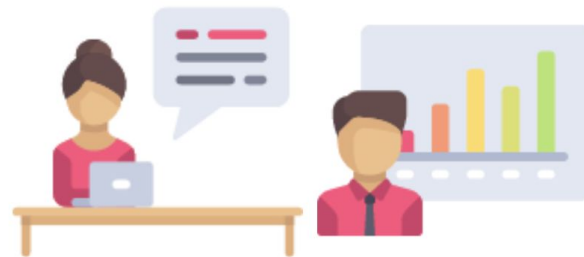Spread the word, let's build technical capacity all over the world.

How to get involved with Delta teaching?

In the Bay Area? Apply to be a volunteer on the teaching team. Applications open later this year!

Contribute to our code base remotely: bit.ly/social-good-ml

Join our slack community and become a beta tester to give feedback on our curriculum and field questions from remote students around the world.

Open Media Foundation

CareerVillage.org

Analytics

Questions?

Appendix

We believe data is powerful, and anybody should be able to use it.

Delta Teaching Leads

Sara Hooker

Director of Educational Outreach

Amanda Su

Project Management & Vision

Dr. Brian Spiering

Coding Lead

Hannah Song

Curriculum Lead

Our goal is simple: empower students with very little if any exposure to machine learning to leverage data for good in their communities.
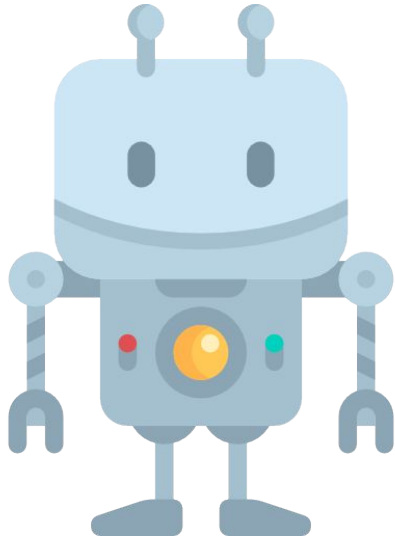
We started this journey for two reasons.

1. Machine learning is an incredibly powerful tool.

Like all powerful tools machine learning requires responsibility & accountability.

We believe the face of data science has to be more representative of the world it serves.

2. Delta has been working for the last 3 years to help non-profits around the world with their data.

Our non-profit grant recipient program is small and selective. We only work with 7-10 nonprofit grant recipients every year.

There is an incredible amount of unmet demand. We cannot help all the non-profits and communities that reach out to us for help.

What have we done so far?

In-person courses

Last year, we travelled to Nairobi, kenya and taught a machine learning for good course.

Graduation ceremony at Moringa School, Nairobi.

Many of our students travelled for more than an hour after work to attend class.

We taught a night class from 6:00 pm to 9:00pm 2x a week. We had work sessions every Saturday.
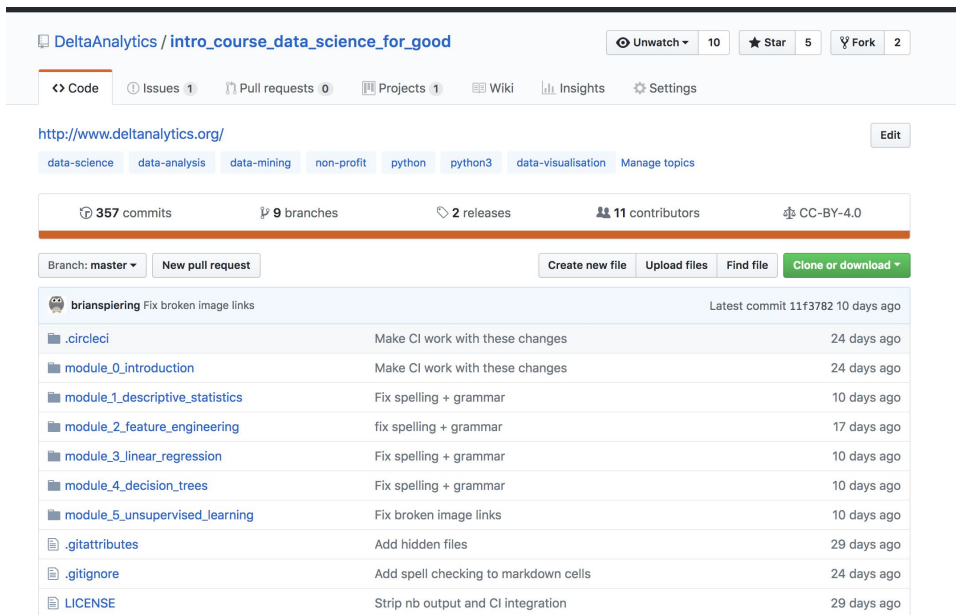
We emphasize teaching using local data to empower community insights.

Number of loans on KIVA each year
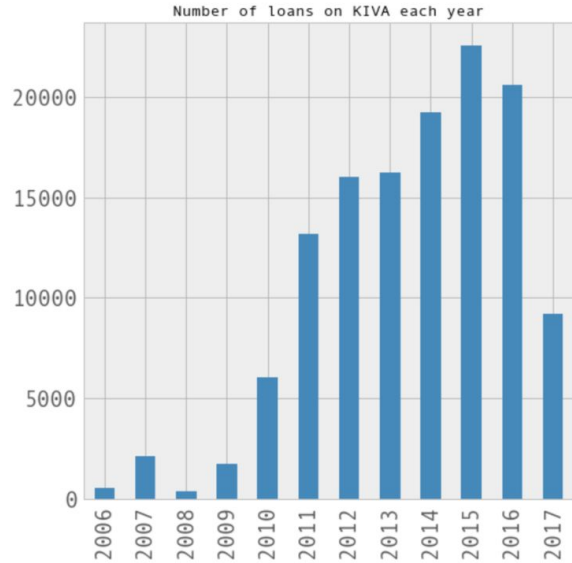
In our pilot program in Kenya we used ~127,000 microfinance loans over 11 years.

Democratize code base and curriculum content.

Our code is Publically Available and Open Source (huge effort from our coding team).

bit.ly/social-good-ml

Help us spread the word!

All our curriculum is public. We welcome comments and have more than 300 remote teachers from all over the world.

You covered this today:

- ✓ What is machine learning?
- ✓ How do you define a research question?
- ✓ What are observations?
- ✓ What are features?
- ✓ What are outcome variables?
- ✓ Introduction to KIVA data

Data Cleaning	Time series

If we have observations over time, we need to do time series validation.

Ask yourself:
a. Is the data for the correct time range?
b. Are there unusual spikes in the data over time?

What should we do if there are unusual spikes in the data over time?

Link to all slides here: http://www.deltanalytics.org/curriculum.html

In the next few months we will teach classes in Morocco and in the local bay area in San Francisco.

Université Ibn Zohr

L'UNIVERSITÉ DE L'EXCELLENCE

En tant qu'espace d'échange et d'ouverture d'esprit, l'Université Ibn Zohr fait de l'excellence un levier de développement humain respectueux de la diversité culturelle et humaine de la société marocaine

Spread the word, let's build technical capacity all over the world.

How to get involved with Delta teaching?

In the Bay Area? Apply to be a volunteer on the teaching team. Applications open later this year!

Contribute to our code base remotely: bit.ly/social-good-ml
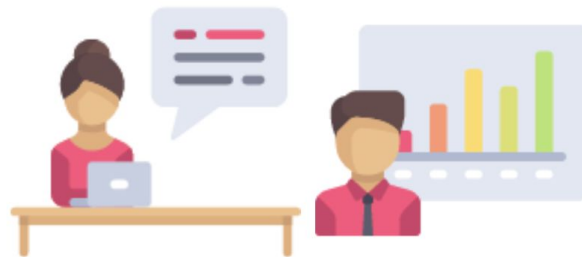
Join our slack community and become a beta tester to give feedback on our curriculum and field questions from remote students around the world.

Apply for a Grant

APPLY

2019 Grant Applications open through September 30th.

Apply here
http://www.deltanalytics.org/2018-data-grant.html

QUESTIONS?

Contact us at inquiry@deltanalytics.org

High Level Steps

- Data cleaning
- Removing "Fluff"
- Unsupervised text clustering – LDA
- Rule-based label algorithm
- Aggregate labels to session level

```
In [8]: captions_data[(captions_data['session_id'] == 254)]['caption']
Out[8]: 235      welcome everyone to the october twelfth city c...
        1676     transactions and to consult with counsel conce...
        4511     to state this general subject of the session a...
        5395     session and i welcome the audience back into t...
        12751    i think the people who are getting answer the ...
        14499    you want me ask my question yeah it's like whe...
        15926    us why okay oh i didn't pull it because i obje...
        17103    a very important thing and it's also a lot of ...
        18065    measure does and and and and what the money's ...
        19274    about the project so and susan peggy from a we...
        22035    approval on that i'm gonna let sue fahey who i...
        23454    compelling good evening so and what we're requ...
        25093    filter replacements we started that with the 1...
        26222    on either replacing shoring up or doing major ...
        27554    basis we're anticipating in the current market...
        29022    perspective with the savings we're anticipatin...
        30168    and so so this does not include anything assoc...
        31962    that we have set forth but these projects are ...
        33123    system right right this is yeah birth quakes a...
        34388    that doesn't have two sources of water all of ...
        35364    days of water about with our depends on the se...
        36724    going to see happen this is a little side note...
        37890    sitting there and and and then restored that a...
        39079    job nice nicely done so so the benefits are th...
        40435    decrease in rates actually because of that and...
```

Donate

FUNDING OUR PROGRAMS

We are currently raising funds for our 2018 programs. If you are interested in supporting our Data for Good mission, please consider contributing.

Delta Analytics is a registered 501(c)3 non-profit organization. Your contribution is tax deductible.

Donate here
https://www.paypal.com/us/fundraiser/charity/1392375

| BACKGROUND | PROGRAMS | GET INVOLVED |

Data Service Grant - BUILD
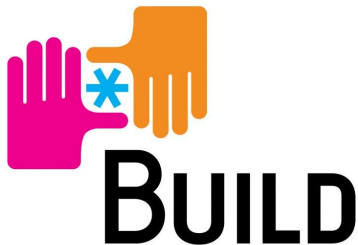
ANALYTICS   EDUCATION

BUILD is an Oakland, CA based education organization that provides an four year entrepreneurship program for at risk high school students to reduce the likelihood of dropping out, increase college acceptance, and improve career readiness.

Problem Statement

What factors, in particular BUILD program elements, determine student success?

Solution

Integrate BUILD's student demographic, college enrollment, and student participation across national programs to find factors which correlate with student college enrollment and persistence.

BACKGROUND   PROGRAMS   GET INVOLVED

Data Service Grant - BUILD

ANALYTICS   **EDUCATION**

Deliverables

- Regression analyses
- Project documentation
- Report summarizing findings

Graph 1
Students first school (476 students)

Enrolled in 2yr schools (50%)   Enrolled in 4yr schools (50%)

Graph 2
Success of students enrolled in 2yr schools
(235 students)

Success = 1
(2yr only)
67%

Success = 3
(transferred to 4yr)
33%

Graph 3
Time taken to transfer from 2yr to 4yr
(77 students)

Within 1 year
26%

Within 2 years
69%

Within 4 years
5%

Variable Importance chart:
- Site
- Pct Attended
- Not In Another College Prep Program
- Has Older Sibling In Build
- Gender
- Ethnicities Report Formula
- End Of 12Th Cum Weighted District Gpa
- Completed E 4

x-axis: Variable Importance (0, 10, 20, 30)

BACKGROUND   **PROGRAMS**   **GET INVOLVED**

Data Service Grant - Technoserve

ANALYTICS   ECONOMIC DEVELOPMENT

TECHNOSERVE is an Washington DC based international development organization that works to lift communities out of poverty through developing the capacity of individuals, promoting connections among local businesses, and improving the business environment.

Problem Statement

How does the financial performance of Technoserve assisted business differ based on owner gender, size of business, and sector?

Solution

Develop a data cleaning methodology for financials collected across different international markets and a data pipeline into an analytics dashboard to summarize and automate reporting.

BACKGROUND   PROGRAMS   GET INVOLVED
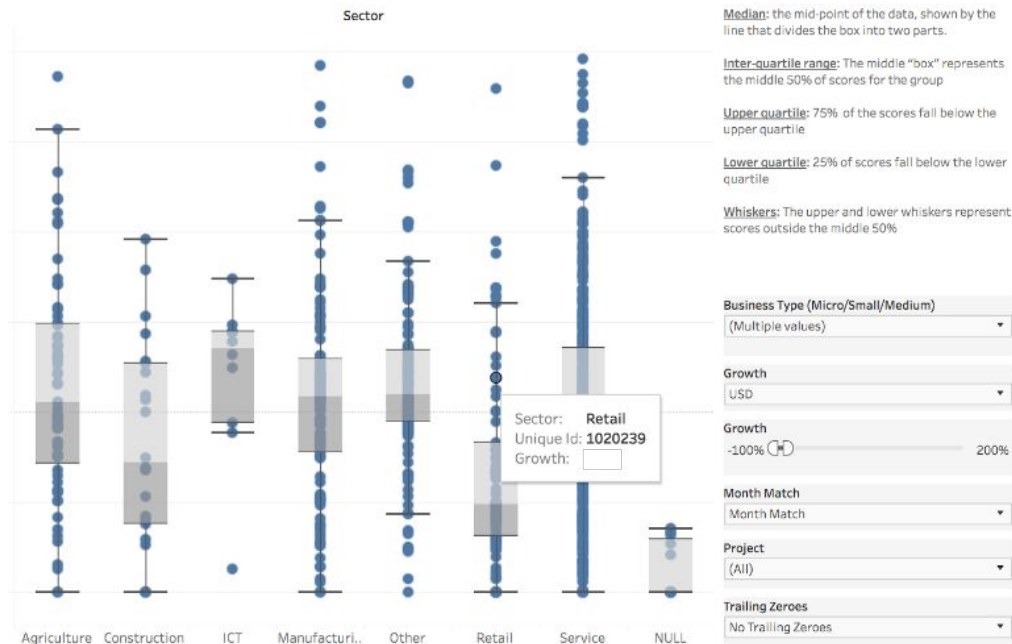
Data Service Grant - Technoserve

TECHNOSERVE

Deliverables

- Data cleaning and aggregation methodology
- Dashboards integrated with cleaned financial data
- Report summarizing findings

Sector

Median: the mid-point of the data, shown by the line that divides the box into two parts.

Inter-quartile range: The middle "box" represents the middle 50% of scores for the group

Upper quartile: 75% of the scores fall below the upper quartile

Lower quartile: 25% of scores fall below the lower quartile

Whiskers: The upper and lower whiskers represent scores outside the middle 50%

Sector: Retail
Unique Id: 1020239
Growth:

Business Type (Micro/Small/Medium)
(Multiple values)

Growth
USD

Growth
-100%   200%

Month Match
Month Match

Project
(All)

Trailing Zeroes
No Trailing Zeroes

Agriculture Construction ICT Manufacturi.. Other Retail Service NULL

Data Service Grant - Rainforest Connection

RAINFOREST CONNECTION is a San Francisco based environmental organization that seeks to leverage mobile technologies to monitor and protect remote forests by enabling real time interventions.

Problem Statement

How can Rainforest connection better use deep learning models to improve its accuracy in detecting and locating areas of the rainforest experiencing illegal logging?

Solution

Rewrite the software architecture for how RFCx builds and trains its models. Build models to estimate directionality of illicit activity.

BACKGROUND PROGRAMS GET INVOLVED
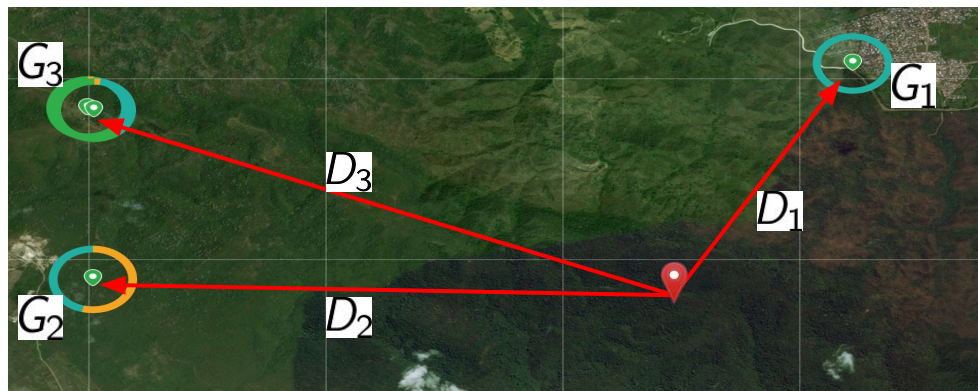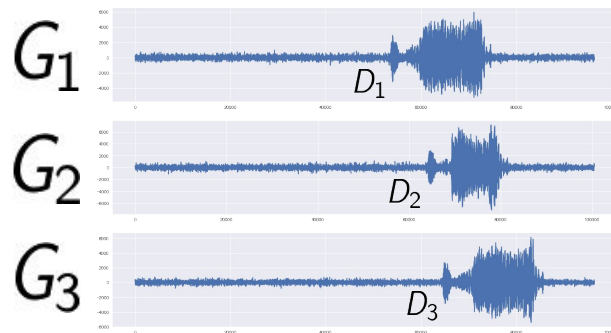
Data Service Grant - Rainforest Connection

RAINFOREST CONNECTION

DATA SCIENCE   ENVIRONMENT

Deliverables

- Command line code base to streamline model training
- Methodology to train directionality models

BACKGROUND   PROGRAMS   GET INVOLVED