

Data, Design, and Engagement

Lessons from 30+
Data for Social Good Projects

About me

Philosophy

then

Software Engineering

then

Data Science

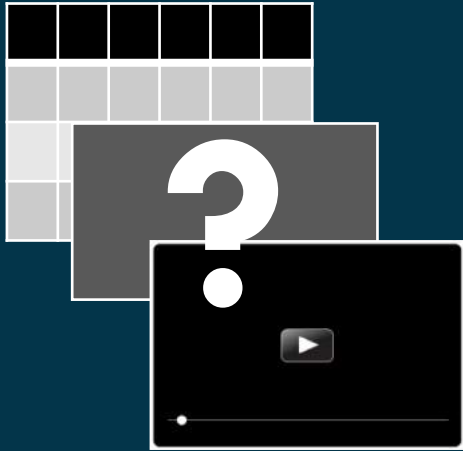
Co-founder of **DrivenData**

@pjbull | @drivendataorg



Peter Bull

Data Scientist for Social Good



DRIVEN DATA

Active Competitions



UNTIL EARLY 2018 **\$100,000**

Early detection of abnormal lung tissue saves lives. This is a new type of data challenge where competition-tested algorithms are just the beginning. Help build open software that brings

[LET'S GO! →](#)



2 WEEKS LEFT **\$50,000**

Cod, haddock, flounder - these iconic fish have supported New England's fishing fleets for generations. When they're not swimming around the depths of the North Atlantic Ocean,

dmytro
CURRENT LEADER [COMPETE →](#)



2 MONTHS, 1 WEEK LEFT

Using environmental data collected by various U.S. Federal Government agencies - from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric

Divyanshu Suri [COMPETE →](#)



4 MONTHS, 2 WEEKS LEFT

Can you predict whether a donor will return to donate blood given their donation history? Good data-driven systems for predicting donations and supply needs can improve the entire

rcking2
CURRENT LEADER [COMPETE →](#)



4 MONTHS, 3 WEEKS LEFT

The UN's Millennium Development Goals provide the big-picture perspective on international development. We can use these indicators to help understand where to

hristo.buyuklie...
CURRENT LEADER [COMPETE →](#)



5 MONTHS, 1 WEEK LEFT

Can you predict which water pumps are faulty? Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which waterpoints are functional, which need some repairs,

rnox
CURRENT LEADER [COMPETE →](#)



1 YEAR, 5 MONTHS LEFT

We're rebooting our first prized competition for fun and education! Budgets for schools and school districts are huge, complex, and unwieldy. It's no easy task to digest

marielgh
CURRENT LEADER [COMPETE →](#)

Completed Competitions



COMPETITION HAS ENDED **\$16,000**

Data on penguin populations are limited primarily due to the fact that most monitored colonies are nearby permanent research stations. This means that any remote populations are

ambarishg
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$1,000**

US presidential elections come but once every 4 years, and this one's a big one. There are lots of people trying to predict what will happen. Can you top them? In this challenge, you'll predict

tallmeasure
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$10,000**

Contribute to open, cutting-edge research on the use of wearables in promoting health and independence for seniors. Passive monitoring can help detect when something is wrong, all

Daniel_FG
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$15,000**

Your challenge is to develop a model that will predict the yield of Dar Si Hmad's fog nets for every day during an evaluation period, using historical data about meteorological conditions and

ulery
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$5,000**

Metis wants to know: can you identify a bee as a honey bee or a bumble bee? These bees have different behaviors and appearances, but given the variety of backgrounds, positions, and image

ea
[RESULTS →](#)



COMPETITION HAS ENDED **\$5,000**

The City of Boston regularly inspects every restaurant to monitor and improve food safety and public health. As in most cities, health inspections are generally random, which can increase

LilianaMedina
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$5,000**

Recent literature suggests that the demand for women's health care will grow over 6% by 2020. Given how rapidly the health landscape has been changing over the last 15 years, it's

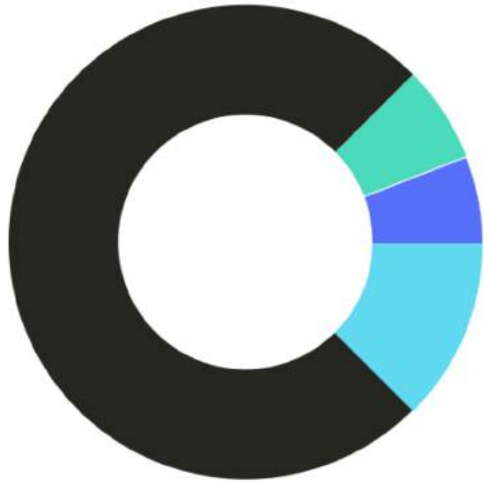
giba
1ST PLACE [RESULTS →](#)



COMPETITION HAS ENDED **\$7,500**

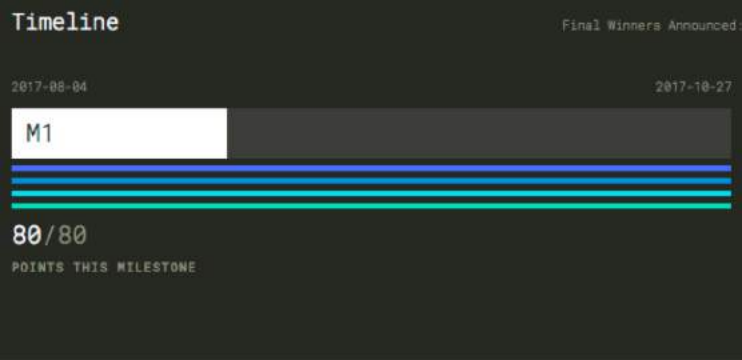
Budgets for schools and school districts are huge, complex, and unwieldy. It's no easy task to digest where and how schools are using their resources. Education Resource

quocnle
1ST PLACE [RESULTS →](#)



CONCEPT TO CLINIC

A NEW KIND OF ML COMPETITION



Point Tally

Prediction

MVP	
@tdraebing	20
@relinakano	5
@musale	5
@anushthakalia	3
@foo-bar-baz-qux	3

Interface frontend

MVP	
@musale	5
@WGierke	3

Interface backend

MVP	
@WGierke	5
@poopingtonic	5

This repository Search Pull requests Issues Marketplace Explore

concept-to-clinic / concept-to-clinic

Unwatch 22 Unstar 131 Fork 62

Code Issues 32 Pull requests 11 Settings Insights

Filters is:issue is:open Labels Milestones New Issue

☐ 32 Open 22 Closed

Author Labels Projects Milestones Assignee Sort

☐ New API endpoint to fetch dataset directory

#77 opened 3 days ago by rahit 0 of 3

☐ 140MB fresh repo size (without much code or data, not using LFS)

community minor official structure-or-tooling

#59 opened 10 days ago by WGierke 5 of 5

2

☐ Create a template for algorithm documentation

community docs minor official

#58 opened 12 days ago by tdraebing

4

☐ Document how to update documentation

community docs medium official

#41 opened 26 days ago by pjbull 0 of 5 1-mvp

☐ Show "X candidates found" on the "Detect and select" page

feature interface-frontend minor official

#37 opened 26 days ago by lamby 1-mvp

☐ Accept/reject nodule candidates interface

feature interface-frontend official

#36 opened 26 days ago by lamby 1-mvp

☐ Show list of candidates nodules on "Detect and select" page

feature interface-frontend major official

#35 opened 26 days ago by lamby 1-mvp

☐ Add ability to set right/left lung for candidate nodule

feature interface-frontend major official

#34 opened 26 days ago by lamby 1-mvp

☐ Display metadata about an image prior to selection

feature interface-frontend major official

#33 opened 26 days ago by lamby 1-mvp

OPEN SOURCE PROJECTS

Cookiecutter Data Science

A logical, reasonably standardized, but flexible project structure for doing and sharing data science work.



Deon is a command line tool that allows you to easily add an ethics checklist to your data science projects.

DIRECT ENGAGEMENTS

BILL & MELINDA
GATES foundation





Bloomberg
Business

404

Error! Error! Error!

404. PAGE NOT FOUND

Unfortunately, this page does not exist. Please check your URL or return to the [Home Page](#).

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hardest work and the biggest need

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hardest work and the biggest need

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs

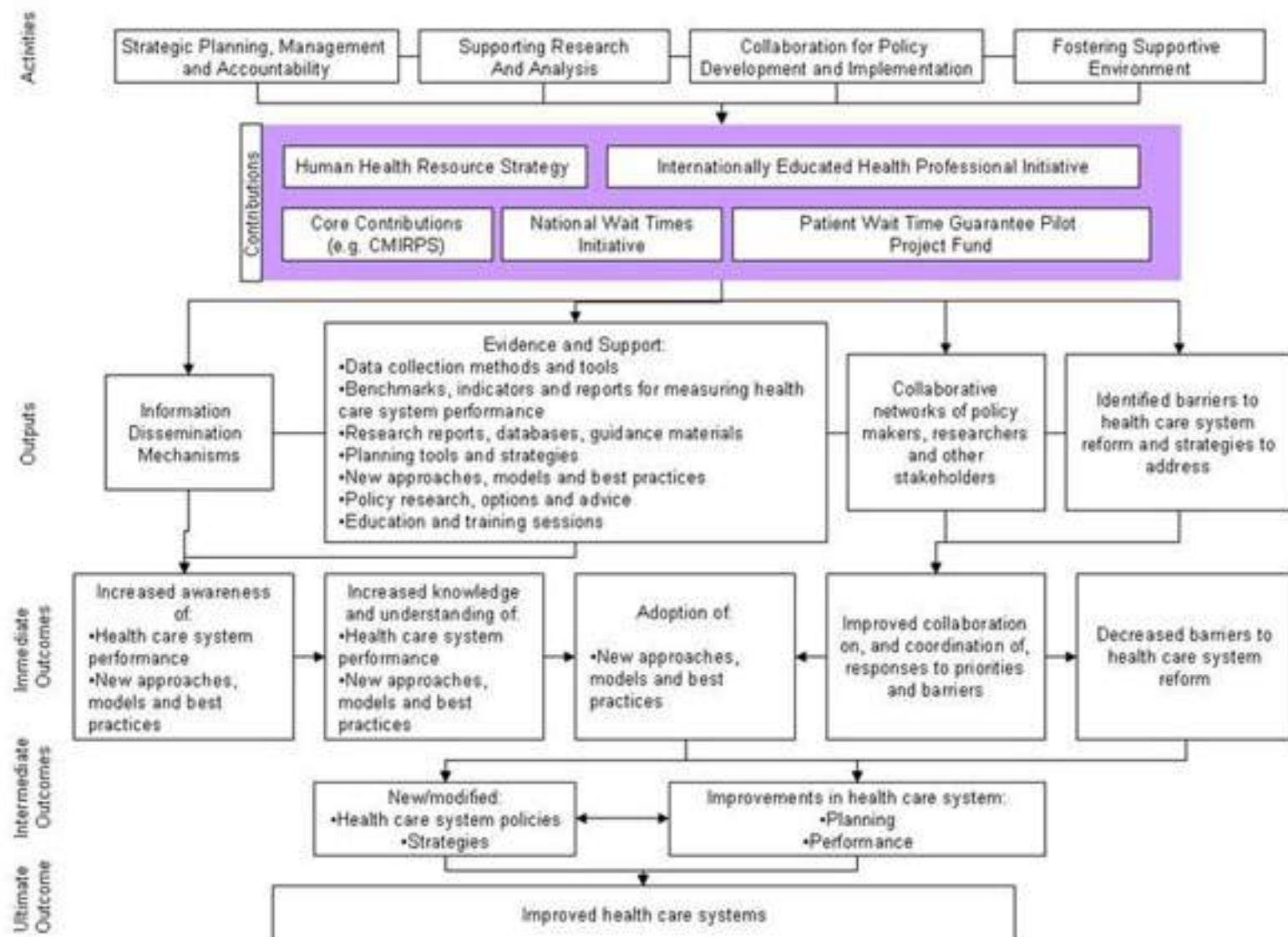
A question for social sector organizations:

How can you use data more effectively?

Answer:

Measure Impact!

Figure 2: Health Care Policy Contribution Program – Logic Model



IMPACT MEASUREMENT

Collection Plan

Data Collection

Data Storage

Data Analysis

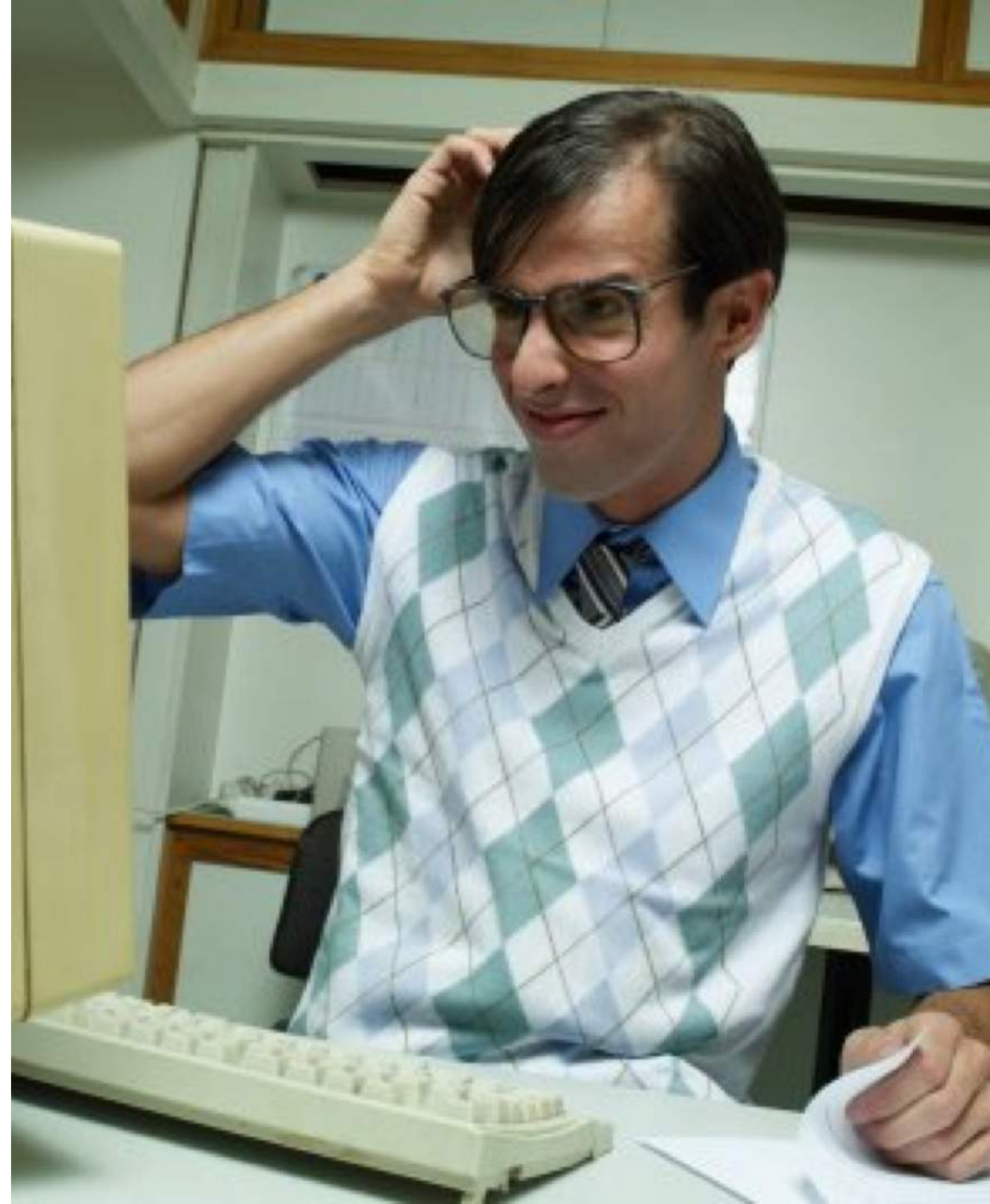
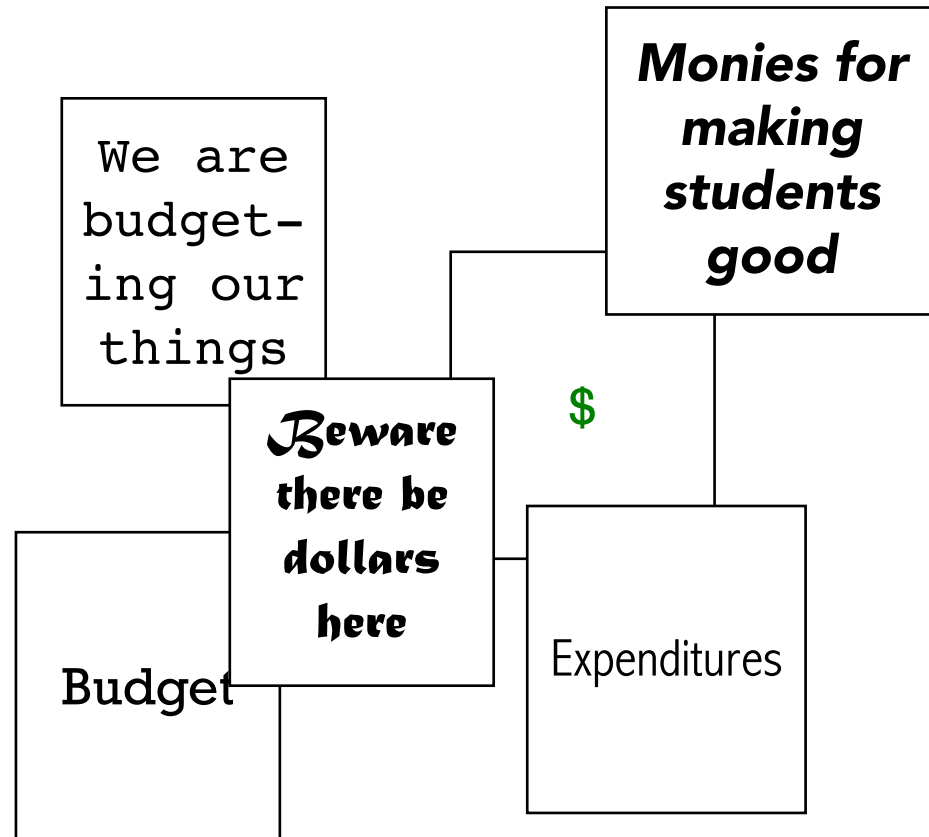
Results

domain expert
data scientist

domain expert
data scientist

DATA SCIENCE

Education Resource Strategies



	Use	Sharing	Reporting	Student_Type	Position_Type	Object_Type	Pre_K	Operating_Status	Object_Type
Instruction	Instruction	School Reported	School	NO_LABEL	Teacher	NO_LABEL	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Base Salary/Compensation	Non PreK	PreK-12 Operating	PERSONAL
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Substitute Compensation	NO_LABEL	PreK-12 Operating	TEACHER
Instruction	O&M	School Reported	School	Unspecified	Custodian	Benefits	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Special Education	Non-Position	Supplies/Materials	NO_LABEL	PreK-12 Operating	EDUCATION
Instruction	O&M	School on Central Budgets	Non-School	Unspecified	Coordinator/Manager	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	EQUIPMENT
Instruction	Instruction	School Reported	School	NO_LABEL	NO_LABEL	Supplies/Materials	NO_LABEL	PreK-12 Operating	SUPPLIES
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Base Salary/Compensation	Non PreK	PreK-12 Operating	PERSONAL
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	REGULAR
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Operating, Not PreK-12	OTHER
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Substitute Compensation	Non PreK	PreK-12 Operating	Salaries
Instruction	Instruction	School Reported	School	Poverty	NO_LABEL	Supplies/Materials	NO_LABEL	PreK-12 Operating	SUPPLIES
Instruction	O&M	Leadership & Management	Non-School	Unspecified	Other	Base Salary/Compensation	Non PreK	PreK-12 Operating	REGULAR
Instruction	O&M	School Reported	School	Unspecified	Non-Position	Rent/Utilities	NO_LABEL	PreK-12 Operating	ELECTRICITY
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	Non-Capitalized
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	Purchases
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	SALARIES
Instruction	Instruction	School Reported	School	Poverty	Teacher	Base Salary/Compensation	Non PreK	PreK-12 Operating	REGULAR
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Operating, Not PreK-12	BONUS
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Substitute Compensation	Non PreK	PreK-12 Operating	PERSONAL
Instruction	Instruction	School Reported	School	Poverty	Teacher	Base Salary/Compensation	Non PreK	PreK-12 Operating	PERSONAL
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	EQUIPMENT
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	O&M	Shared Services	Non-School	Unspecified	Other	Other Compensation/Stipend	Non PreK	PreK-12 Operating	Extra Duty
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Poverty	Teacher	Base Salary/Compensation	Non PreK	PreK-12 Operating	REGULAR
Instruction	O&M	Shared Services	Non-School	Unspecified	Other	Other Compensation/Stipend	Non PreK	PreK-12 Operating	Extra Duty
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	Purchases
Instruction	O&M	Leadership & Management	Non-School	Unspecified	Coordinator/Manager	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	SALARIES
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	SALARIES
Instruction	O&M	School Reported	School	Unspecified	Other	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	OTHER
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Pupil Service	School Reported	School	Poverty	Non-Position	Supplies/Materials	NO_LABEL	PreK-12 Operating	GENERAL
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Other Compensation/Stipend	Non PreK	PreK-12 Operating	PERSONAL
Instruction	Instruction	School Reported	School	Poverty	Teacher	Other Compensation/Stipend	Non PreK	PreK-12 Operating	PERSONAL
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Substitute Compensation	NO_LABEL	PreK-12 Operating	OTHER
Instruction	Instruction	School Reported	School	Poverty	Teacher	Other Compensation/Stipend	NO_LABEL	PreK-12 Operating	EXTRA DUTY
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Other Compensation/Stipend	NO_LABEL	PreK-12 Operating	ADDITIONAL
Instruction	O&M	School Reported	School	Unspecified	Custodian	Other Compensation/Stipend	Non PreK	PreK-12 Operating	Extra Duty
Instruction	Pupil Service	Shared Services	Non-School	Unspecified	Teacher	Other Compensation/Stipend	NO_LABEL	PreK-12 Operating	ADDITIONAL
Instruction	Instruction	School Reported	School	NO_LABEL	NO_LABEL	Supplies/Materials	NO_LABEL	PreK-12 Operating	SUPPLIES
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Substitute Compensation	NO_LABEL	PreK-12 Operating	OTHER
Instruction	O&M	School Reported	School	Unspecified	Custodian	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	PERSONAL
Instruction	NO_LABEL	NO_LABEL	Non-School	NO_LABEL	NO_LABEL	Supplies/Materials	NO_LABEL	PreK-12 Operating	SUPPLIES
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	EMPLOYED
Instruction	ISPD	Leadership & Management	Non-School	Special Education	Other	Other Compensation/Stipend	NO_LABEL	PreK-12 Operating	WORKSHEET
Instruction	Instruction	School Reported	School	ELL	Teacher	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Other Compensation/Stipend	Non PreK	PreK-12 Operating	Salaries
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	SALARIES
Instruction	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	NO_LABEL	Non-Operating	Other Non-Capitalized
Instruction	O&M	School Reported	School	Unspecified	Other	Benefits	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	O&M	School Reported	School	Unspecified	Custodian	Base Salary/Compensation	Non PreK	PreK-12 Operating	PERSONAL
Instruction	Instruction	School Reported	School	Unspecified	Teacher	Benefits	NO_LABEL	PreK-12 Operating	EMPLOYED
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Base Salary/Compensation	NO_LABEL	PreK-12 Operating	SALARIES
Instruction	NO_LABEL	School Reported	School	NO_LABEL	NO_LABEL	Contracted Services	NO_LABEL	PreK-12 Operating	OTHER
Instruction	Instruction	School Reported	School	NO_LABEL	Teacher	NO_LABEL	NO_LABEL	PreK-12 Operating	CONTRACTED
Instruction	Instruction	School Reported	School	Unspecified	Substitute	Benefits	NO_LABEL	PreK-12 Operating	CONTRACTED

PETRO-VEND FUEL AND FLUIDS

Regional Playoff Hosts

Capital Assets - Locally Defined Groups

SATELLITE COOK

ITEMGH EXTENDED DAY

Water and Sewage *

Food Services - Other Costs

NON-CAPITALIZED AV

Supp.- Materials

New Data

Regular *	TEACHER	Special Instruction	TCHR, SCNDY MATH	Certificated Employees Salaries And Wages	IN OTH CERTIFICATED PERSON	1	Disadvantaged Youth *	\$104,928.19	Title I - Disadvantaged Children/Targeted Assistance	BUILDING ALLOCATIONS
-----------	---------	---------------------	------------------	---	----------------------------	---	-----------------------	--------------	--	----------------------

Predictions!

LABEL COLUMN	Function	Sharing	Object_Type	Student_Type	Position_Type
MOST LIKELY	Teacher Compensation	School Reported	Base Salary/Compensation	Poverty	Teacher
2 ND LIKELY	Food Services	School on Central Budgets	Benefit	Unspecified	Coordinator/Manager
3 RD LIKELY	NO_LABEL	NO_LABEL	Supplies/Materials	NO_LABEL	NO_LABEL

Education Resource Strategies

Saving about
300 staff-hours
per year



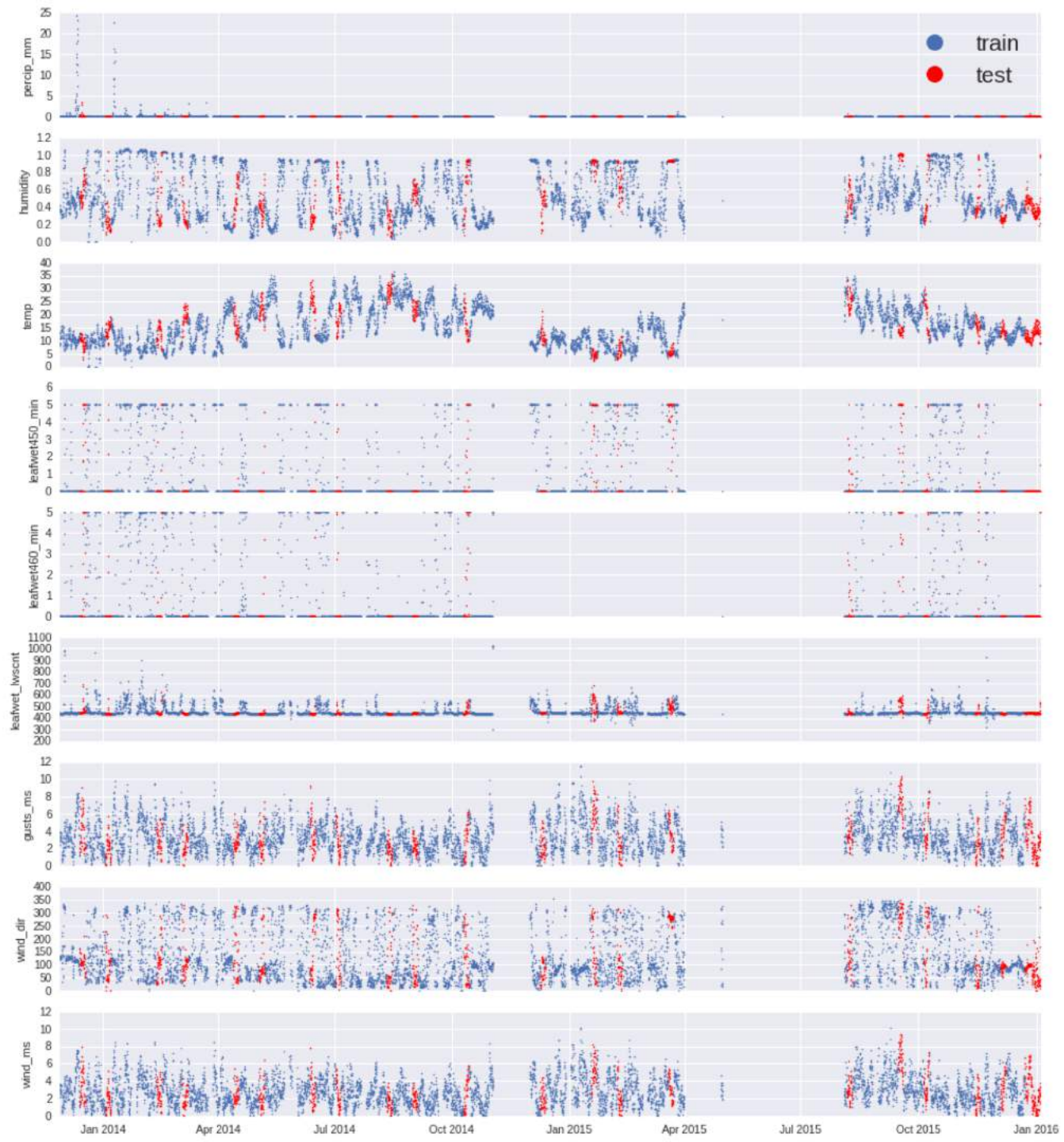
Project Frame 1: Automation

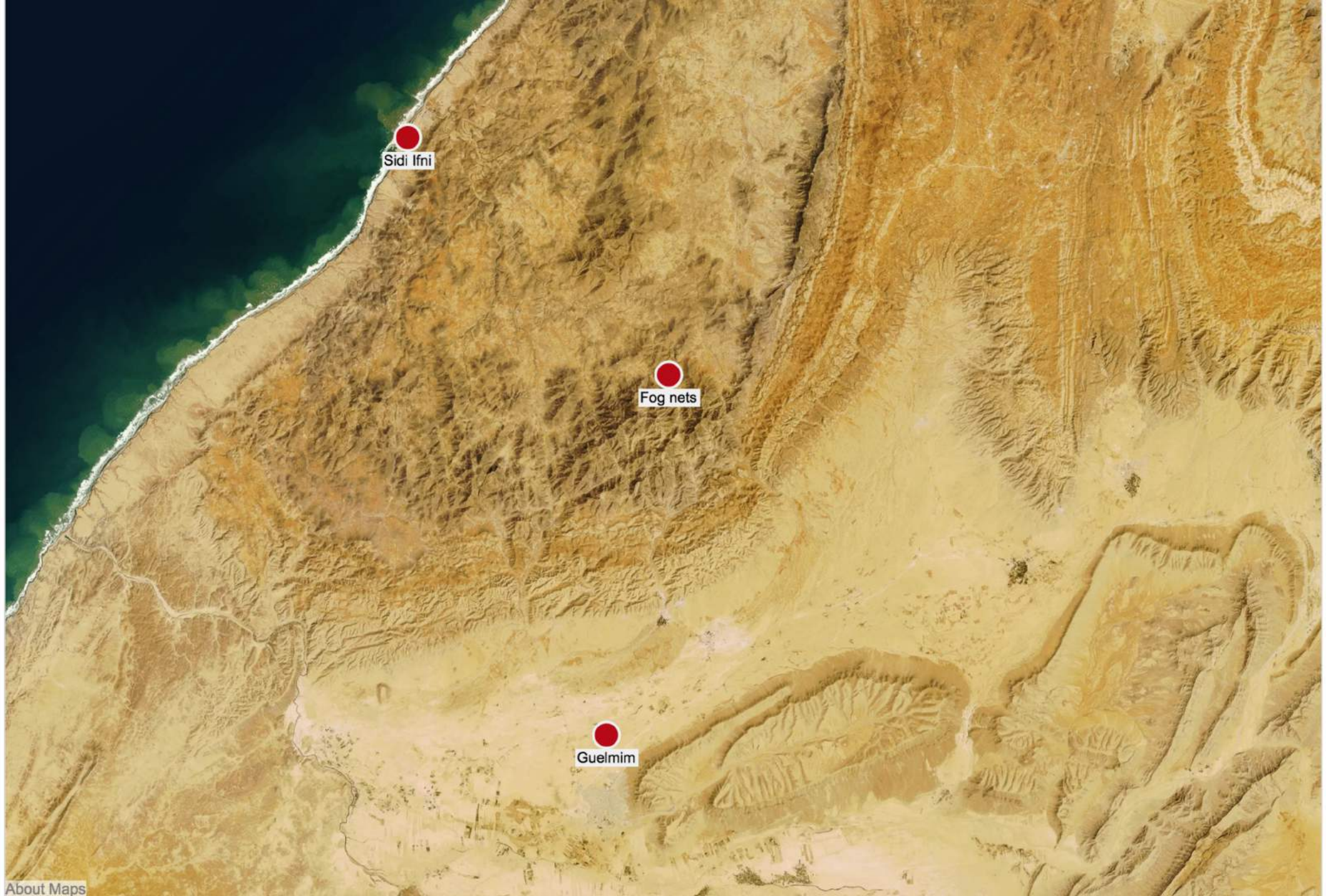






Dar Si Hmad
دار سي حماد





Sidi Ifni

Fog nets

Guelmim

Project Frame 2: Smarter Strategy



A dark, grayscale map of a city, likely New York City, showing a dense network of streets and water bodies. Numerous small, bright yellow dots are scattered across the map, representing specific locations of violations. The dots are more densely clustered in certain areas, such as the upper right and lower center, while other areas have fewer dots.

4 years of history
8,000 locations
315,000 violations

See the business you'd like to review?



1. Loui Loui

★★★★★ 29 reviews

\$\$ Seafood, Salad, Sandwiches

101 Main St
Stoneham, MA 02180
(781) 435-2145



2. Pintxo Pincho Tapas Bar

★★★★★ 189 reviews

\$\$ Tapas Bars

385 Main St
Woburn, MA 01801
(781) 932-1379

★ Write a Review

★ Write a Review



3. Giacomo's Ristorante

★★★★★ 2137 reviews

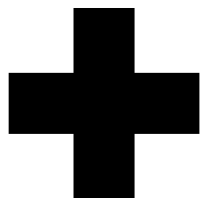
\$\$ Italian

North End
355 Hanover St
Boston, MA 02113
(617) 523-9026

★ Write a Review

★ Write a Review

- **RATINGS**
- **REVIEWS**
 - Date of the review
 - Free form text
- **CATEGORIES**
 - Cuisine
 - Format
- **NEIGHBORHOODS**
- **CHECKINS**



Project Frame 3: Smarter Use of Scant Resources

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hard work the sector needs

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs

Social sector organizations exist for the public good, which demands a higher attention to data ethics.

Data collection

Data security: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?

WHERE THINGS HAVE GONE WRONG

Personal and financial data for more than 146 million people was stolen in Equifax data breach.

Modeling

Explainability: Can we explain in understandable terms a decision the model made in cases where a justification is needed?

WHERE THINGS HAVE GONE WRONG

Patients with pneumonia with a history of asthma are usually admitted to the intensive care unit as they have a high risk of dying from pneumonia.

Given the success of the intensive care, neural networks predicted asthmatics had a low risk of dying and could therefore be sent home. Without explanatory models to identify this issue, patients may have been sent home to die.

Deployment

Redress: Have we discussed with our organization a plan for response if users are harmed by the results (e.g., how does the data science team evaluate these cases and update analysis and models to prevent future harm)?

WHERE THINGS HAVE GONE WRONG

In Arkansas, an algorithm was used to allocate home care under a Medicaid waiver program to determine how frequently a caretaker would visit a patient.

However due to coding errors, people with diabetes or cerebral palsy were confronted with cuts to their care and with no clear mechanism for redress.

Eventually a court case was brought by the ACLU.

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hard work the sector needs

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs

A memo to big tech

The techlash against Amazon, Facebook and Google—and what they can do

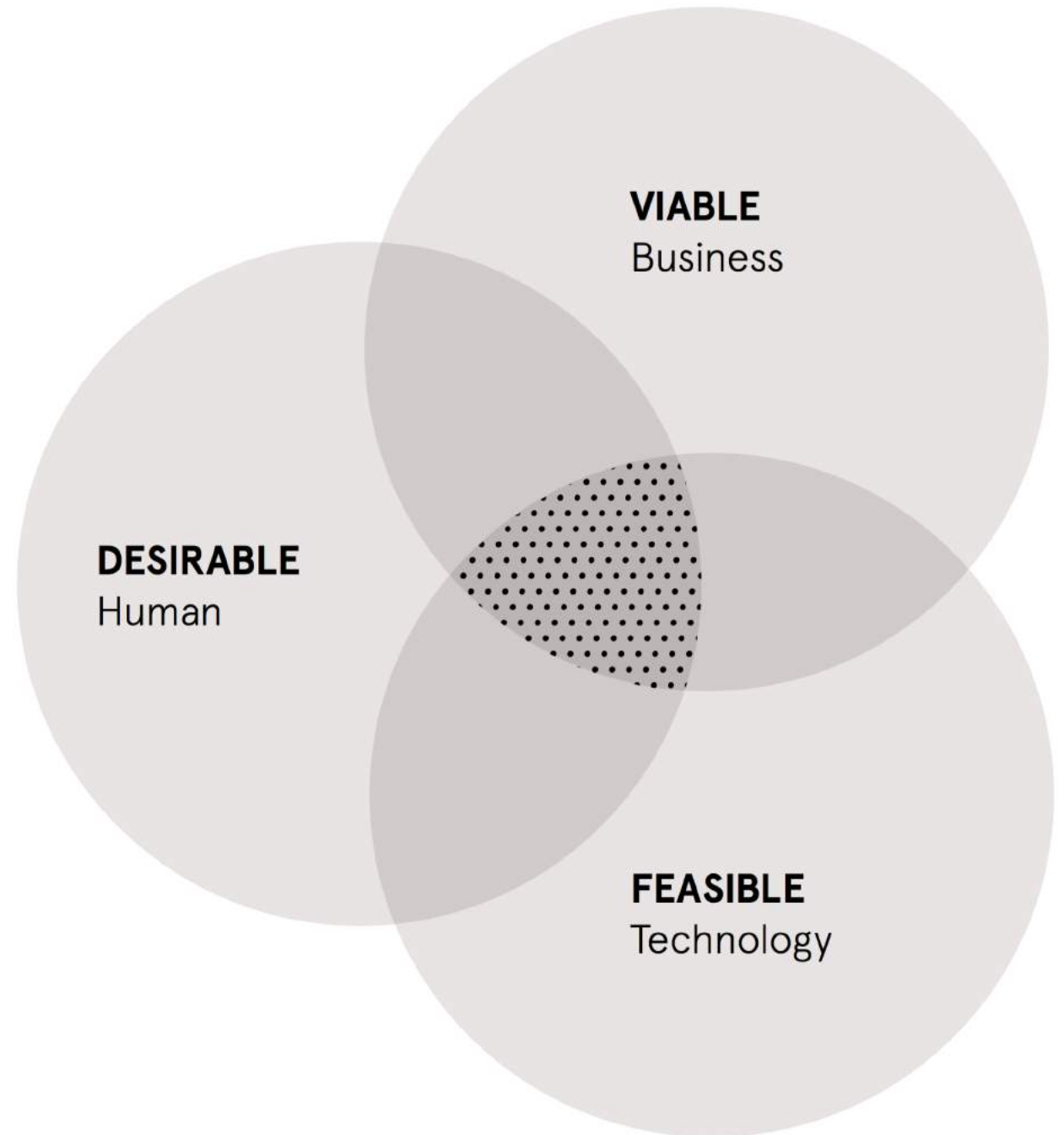
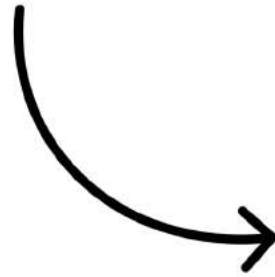
Which antitrust remedies to welcome, which to fight

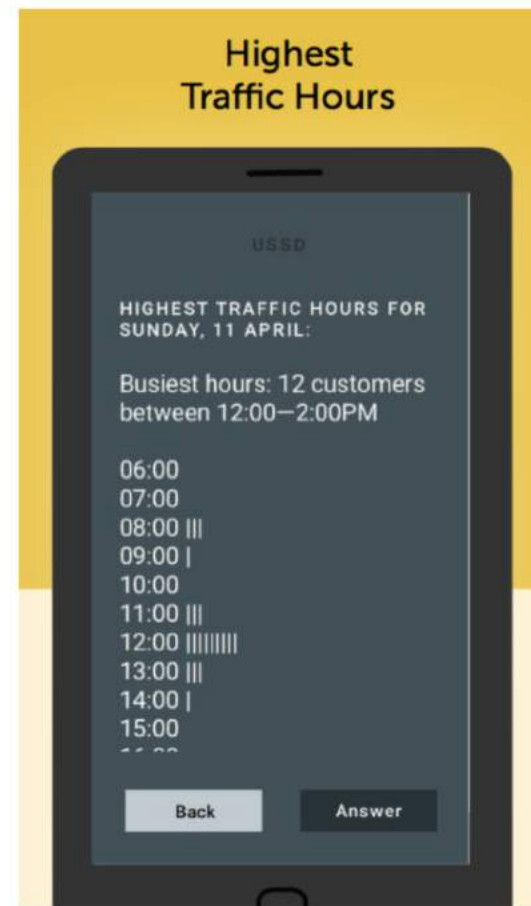
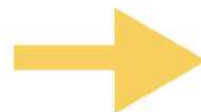
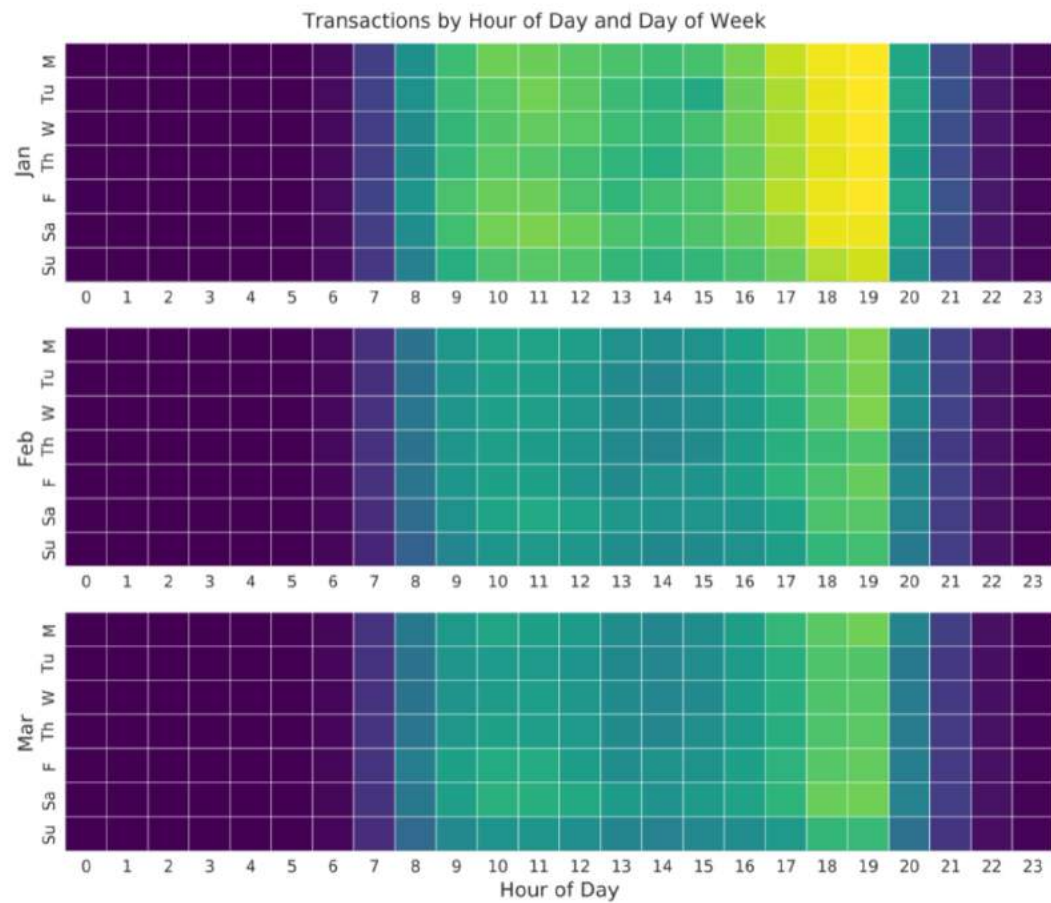


Ryan Olbrysh, AFP, EPA, Facebook, Amazon, Google

Design Thinking

Start here





core activities of the

Human Centered Data Scientist

1. Go to the field and observe data being generated.
2. Design with, not for, by iterating on prototypes.
3. Assess outcomes both quantitatively and qualitatively.
4. Be honest about what doesn't work and learn from it.

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hard work the sector needs

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs



THE DATA CAPACITY GAP

140k – 180k

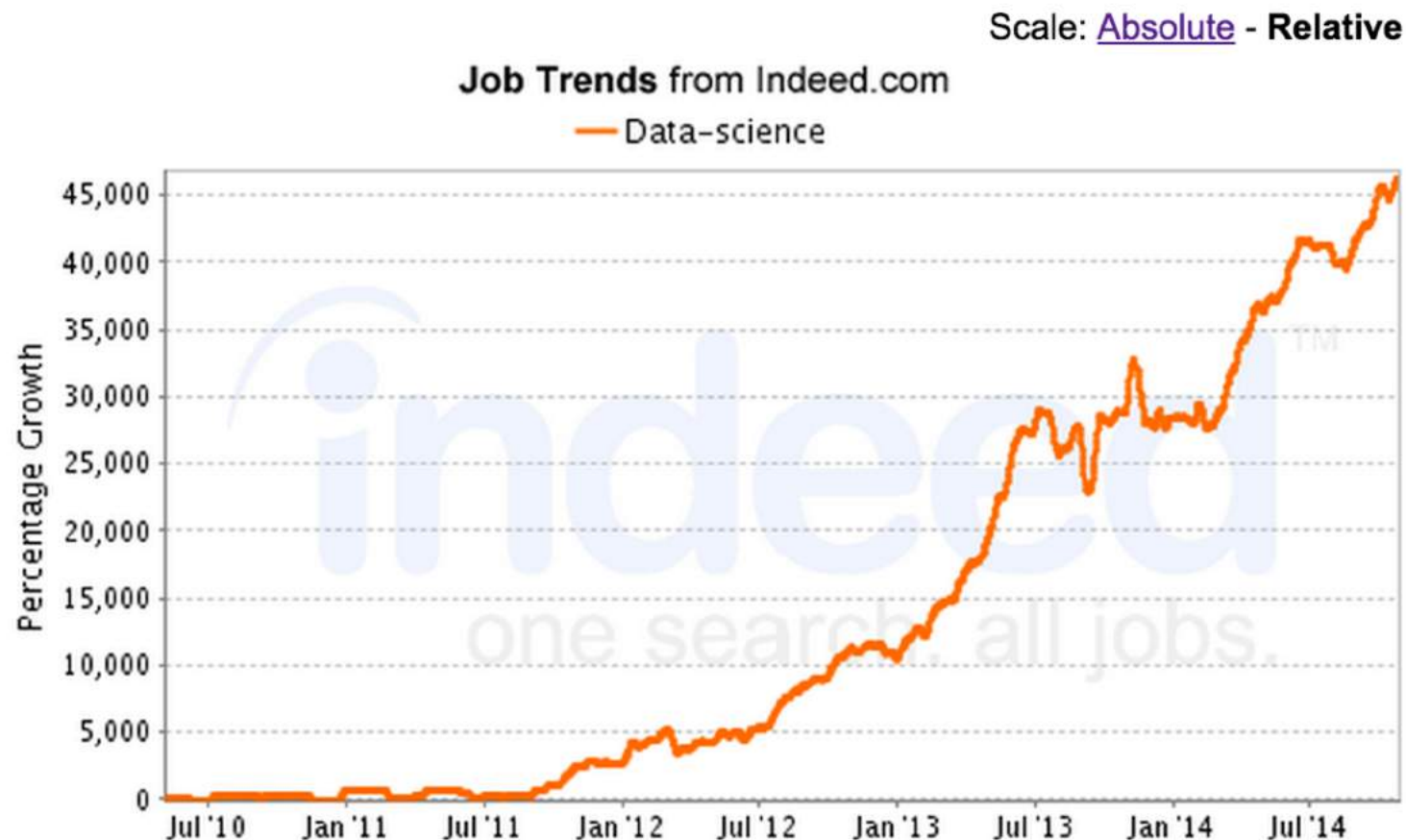
shortage of data
scientists

\$118,700

average salary of
data scientist

\$135,000

average salary
executive director
US nonprofit



Machine Learning: The High-Interest Credit Card of Technical Debt

**D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov,
Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young**
`{dsculley, gholt, dgg, edavydov}@google.com`
`{toddphillips, ebner, vchaudhary, mwyoung}@google.com`
Google, Inc

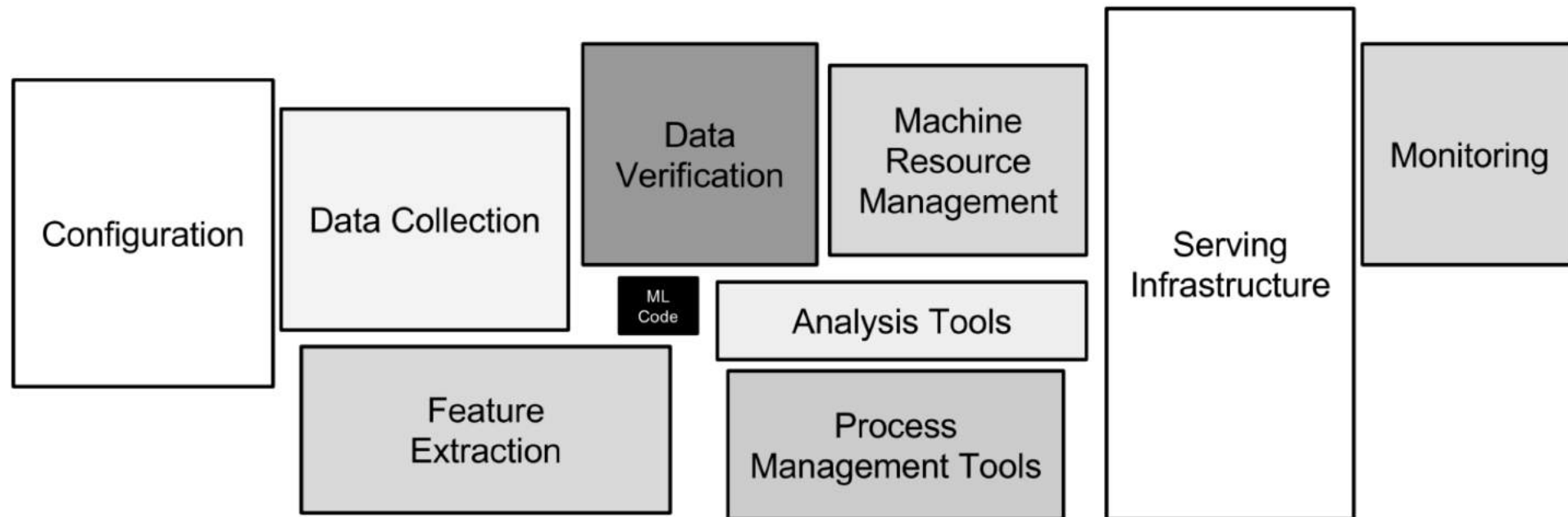


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley, gholt, dgg, edavydov, toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner, vchaudhary, mwyong, jfcrespo, dennison}@google.com
Google, Inc.

Data for Impact Field Guide

Finding a Project

Identifying and scoping the project is the hardest work and the biggest need

Launching a Project

The bar for data ethics is higher in the social sector

Running a Project

We need a process that builds trust between technologists and the social sector

Wrapping up a Project

The capacity gap is wide and jeopardizes handoffs

Thanks!

Any questions?