Data, Design, and Engagement

Lessons from 30+
Data for Social Good Projects

National Academies of Sciences, Engineering, and Medicine
10 December 2018
About me

**Philosophy**
then
**Software Engineering**
then
**Data Science**

Co-founder of *DrivenData*

---

**Peter Bull**

Data Scientist for Social Good

@pjbull | @drivendataorg
CONCEPT TO CLINIC
A NEW KIND OF ML COMPETITION
Point Tally

Prediction

- New API endpoint to fetch dataset directory
- 140MB fresh repo size (without much code or data, not using LFS)
- Create a template for algorithm documentation
- Document how to update documentation
- Show "X candidates found" on the "Detect and select" page
- Accept/reject nodule candidates
- Show list of candidates nodules on "Detect and select" page
- Add ability to set right/left lung for candidate nodule
- Display metadata about an image prior to selection
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.
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

Activities
- Strategic Planning, Management and Accountability
- Supporting Research and Analysis
- Collaboration for Policy Development and Implementation
- Fostering Supportive Environment

Contributions
- Human Health Resource Strategy
- Internationally Educated Health Professional Initiative
- Core Contributions (e.g., CMIRPS)
- National Wait Times Initiative
- Patient Wait Time Guarantee Pilot Project Fund

Evidence and Support:
- Data collection methods and tools
- Benchmarks, indicators and reports for measuring health care system performance
- Research reports, databases, guidance materials
- Planning tools and strategies
- New approaches, models and best practices
- Policy research, options and advice
- Education and training sessions

Outputs
- Information Dissemination Mechanisms

Intermediate Outcomes
- Increased awareness of:
  - Health care system performance
  - New approaches, models and best practices

Immediate Outcomes
- Increased knowledge and understanding of:
  - Health care system performance
  - New approaches, models and best practices

Adoption of:
- New approaches, models and best practices

Ultimate Outcome
- Decreased barriers to health care system reform

New options:
- Health care system policies
  - Strategies

Improvements in health care system:
- Planning
  - Performance

Improved health care systems
We are budgeting our things. Beware there be dollars here. Expenditures. Monies for making students good.
<table>
<thead>
<tr>
<th>Use</th>
<th>Sharing</th>
<th>Reporting</th>
<th>Student_Type</th>
<th>Position_Type</th>
<th>Object_Type</th>
<th>Pre_K</th>
<th>Operating_Status</th>
<th>Object_Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>Teacher</td>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>Pre-K</td>
<td>Operating</td>
<td>NOLABEL</td>
</tr>
<tr>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>Teacher</td>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>NOLABEL</td>
<td>Pre-K</td>
<td>Operating</td>
<td>NOLABEL</td>
</tr>
</tbody>
</table>

**PETRO-VEND FUEL AND FLUIDS**

Regional Playoff Hosts

**SATELLITE COOK**

Capital Assets - Locally Defined Groups

**ITEMGH EXTENDED DAY**

Water and Sewage *

Food Services - Other Costs

**NON-CAPITALIZED AV Supp.- Materials**
# New Data

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

# Predictions!

<table>
<thead>
<tr>
<th>LABEL COLUMN</th>
<th>Function</th>
<th>Sharing</th>
<th>Object_Type</th>
<th>Student_Type</th>
<th>Position_Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOST LIKELY</td>
<td>Teacher Compensation</td>
<td>School Reported</td>
<td>Base Salary/Compensation</td>
<td>Poverty</td>
<td>Teacher</td>
</tr>
<tr>
<td>2ND LIKELY</td>
<td>Food Services</td>
<td>School on Central Budgets</td>
<td>Benefit</td>
<td>Unspecified</td>
<td>Coordinator/Manager</td>
</tr>
<tr>
<td>3RD LIKELY</td>
<td>NO_LABEL</td>
<td>NO_LABEL</td>
<td>Supplies/Materials</td>
<td>NO_LABEL</td>
<td>NO_LABEL</td>
</tr>
</tbody>
</table>
Saving about 300 staff-hours per year
Project Frame 1: Automation
Project Frame 2: Smarter Strategy
4 years of history
8,000 locations
315,000 violations
See the business you’d like to review?

1. Loui Loui
   - Cuisine: Seafood, Salad, Sandwiches
   - Price: $5
   - Location: Stoneham, MA 02180
   - Reviews: 29
   - Phone: (781) 435-2745

2. Pintxo Pincho Tapas Bar
   - Cuisine: Tapas Bars
   - Price: $5
   - Location: Woburn, MA 01801
   - Reviews: 189
   - Phone: (781) 932-1379

3. Giacomo’s Ristorante
   - Cuisine: Italian
   - Price: $5
   - Location: Boston, MA 02113
   - Reviews: 2137
   - Phone: (617) 523-8026

- **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

**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.

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

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)?
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
Design Thinking

Start here

DESIRABLE
  Human

VIABLE
  Business

FEASIBLE
  Technology
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, myoung, jfcrespo, dennison}@google.com
Google, Inc.
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?