Making Data Great Again
(or data, data, everywhere – we have to stop and think)

Julia Lane
New York University
Key ideas

• Economy has changed substantially => new measures necessary
• Enormous potential with new data
• Statistical agencies have new role
• We need to build new demand-driven institutions – local plus federal
• We need to stop and think
A Locally Based Initiative to Support People and Communities by Transformative Use of Data

The data revolution is transforming how executives manage operations and businesses deliver goods and services. Yet when it comes to helping people escape poverty, the revolution has barely begun.

Overview

Key Performance Indicators

Key Drivers of Transformation

Leveraging Data as a Strategic Asset

Goal Statement

The use of data is transforming society, business, and the economy. Data provided by the Federal Government have a unique place in society and maintaining trust in Federal data is pivotal to a democratic process. The Federal Government needs a robust, integrated approach to using data to deliver on mission, serve customers, and steward resources while respecting privacy and confidentiality.
Outline

Rethinking measurement
Operationalizing
A possible approach
- Human
- Technical
Next steps
Outline

Rethinking measurement
Operationalizing
A possible approach
- Human
- Technical
Next steps
REPORT TO THE PRESIDENT
Technology and the Future of Cities

Executive Office of the President
President’s Council of Advisors on Science and Technology

February 2016
Demand in previous century

- Great Depression
- Wartime economic planning
- Colin Clark, Simon Kuznets, Richard Stone
Demand now

• Economic activity?
  – GDP
  – Resiliency
  – Sustainability
  – Mobility

• Units?
  – Networks
  – Neighborhood
  – Country
Rethinking measures

• New products
  – Services
  – Intangible assets
  – Technology/robots

• New people
  – Immigration
  – Globalization

• New boundaries
  – Local
  – Regional
  – Cross national
Outline

Rethinking measurement

Operationalizing

A possible approach
  - Human
  - Technical

Next steps
Data collection

Resources

Companion websites for publications

- Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations

Data

- Urbansound Dataset – A dataset containing 1302 labeled sound recordings. Each recording is labeled with the start and end times of sound events from 10 classes
- Urbansound8k Dataset – A dataset containing 8732 labeled sound excerpts (<=4s) of urban sounds from 10 classes
- URBAN-SED Dataset – A dataset of 10,000 synthesized soundscapes with sound event annotations generated using Scaper
- Seeing Sound Dataset – A dataset of 5400 crowdsourced audio annotations of 60 synthesized soundscapes

Code

- Scaper – A Python library for soundscape synthesis and augmentation
- Audio-Annotator – A Javascript web interface for annotating audio data
- Raster Join
- Urban Pulse
What is needed?

• Timeliness?
• Closeness to core measure?
• Coverage?
• Geographic detail
• Longitudinal Consistency

How do we trade off?
Collection, documentation, Curation

TripAdvisor
New skills

- Framing question
- Webscraping/APIs
- Data Management
- Linkage
- Machine Learning
- Text Analysis
- Graph Analysis
- Visualization
- Inference
- Privacy and Confidentiality
Or, as computer scientists put it

- Understand “Business” problem
- Map to Machine Learning problem
- Understand the data
- Explore and Prepare the data
- “Feature” Development
- Method Selection
- Evaluation
- Deployment
Outline

Rethinking measurement
Operationalizing
A possible approach
- Human
- Technical
Next steps
“If HP only knew what HP knows, we would be much more profitable”

(former CEO Lew Platt)
Federal level

H.R. 1831: Evidence-Based Policymaking Commission Act of 2016

Introduced: April 18, 2015
114th Congress, 2015–2017

Status: Enacted — Signed by the President on Mar 30, 2016.
This bill was enacted after being signed by the President on March 30, 2016.


Sponsor: Paul Ryan
Representative for Wisconsin’s 1st congressional district
Republican

Text:
Read Text
Last Updated: Mar 18, 2016
Length: 5 pages

FY 2016 Significant Investments

- **2020 Census ($663M)**: We have the potential to save $5 billion with the new 2020 Census design, however, we now have to build operations and systems for the 2020 Census, based on the new design.
- **CEDCaP ($17M)**: Smarter IT Delivery Built on a Shared-Services Model.
- **American Community Survey ($257M)**: We must maintain the quality of the data while continuing our efforts to reduce respondent burden.
- **Geographic Support ($81M)**: We must make use of technology and partnerships to deliver smarter geographic solutions to our surveys and censuses.
- **Administrative Records Clearinghouse ($15M)**: Will expedite the acquisition of federal and federally sponsored administrative data sources, improve data documentation and linkage techniques, and leverage and extend existing systems for governance, privacy protection, and secure access to these data.
- **Economic & Government Censuses ($144M)**: Data products drive economic activity and are relevant to the needs businesses, policymakers, and the public. $10.1 million increase

Administrative Data Research Facility: The Administrative Data Research Facility is a pilot project that enables secure access to analytical tools, data storage and discovery services, and general computing resources for users, including Federal, state, and local government analysts and academic researchers. The Census Bureau and academic partners developed the project as part of the collaborative Training Program in Applied Data Analytics sponsored by the University of Chicago, New York University, and the University of Maryland. It is currently operating as a pilot with users accessing the Facility as part of the training program. The Facility operates as a cloud-based computing environment, with Federal security approvals, which currently hosts selected confidential data from the U.S. Department of Housing and Urban Development and the Census Bureau, as well as state, city, and county agencies, and an array of public use data.
REPORT TO THE PRESIDENT
Technology and the Future of Cities

Executive Office of the President
President’s Council of Advisors on Science and Technology

February 2016
A number of barriers

Technical
• cost
• burden
• data quality

Human
• data documentation
• risk of bad analysis
• legal mandates surrounding data access and use
• Workforce capacity
A possible approach

Documentation Module
Explorer links metadata, codes, tools, publications

Stewardship Module
Approval workflow, monitoring, reporting

Collaboration Module
Interactive chat and code sharing

Workspace and tools
- Python
- Jupyter
- R

Security Module
FedRAMP security certified

Data in cloud
Alternative: local servers

Data analysis
Code
Collaboration

Metadata
Usage Feedback

Data user

Data producer

Training Module

Data steward
Access
Workflows
Monitoring

Reporting
Outline

Rethinking measurement
Operationalizing
A possible approach
- Human
- Technical
Next steps
Human approach

• Work with trusted partners
• Create value proposition
  – Develop products of value to data owners
  – Build workforce capacity
• Build metadata documentation automatically
Results: Over 40 Confidential Datasets

<table>
<thead>
<tr>
<th>Federal (6)</th>
<th>States (12)</th>
<th>Cities (15)</th>
<th>Counties (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census (LEHD and ACS)</td>
<td>Labor (Wage records, QCEW, UI claims)</td>
<td>NYPD</td>
<td>King County Transportation, Human Services</td>
</tr>
<tr>
<td>HUD (Housing Choice Voucher Program,</td>
<td>Human Services (TANF, SNAP)</td>
<td>Chicago PD</td>
<td>Mecklenburg County Corrections</td>
</tr>
<tr>
<td>Public Housing, Project-based Section 8, and the Section 202/811 Programs)</td>
<td>Corrections (admissions and exits)</td>
<td>NYC Labor, Human Services, Corrections, Homeless, VocEd</td>
<td></td>
</tr>
<tr>
<td>Revenue (Business tax)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Team work
Networks: >90 govt agencies; >200 participants
What our **participants** say about the program

"Love the Jupyter notebooks!! ... I love how the code snippets and explanations are set up in the Jupyter notebooks. The format of going through it individually and discussing questions/challenges in your group, with the experts available when needed, worked really well for my learning style."

I could see our agency benefiting potentially from something like this in that, as the system builds out and collects additional resources/datasets that impact criminal justice system practices, this may be an option for a place for us to look for the results of studies using evidence based practices.

Danielle Fulmer  
Director of Business Analytics

Katy Fitzgerald  
Management Analyst
Outline

Rethinking measurement
Operationalizing
A possible approach
  - Human
  - Technical
Next steps
Conceptual Framework: User Needs
Components

Component 1: Security
Component 2: Data Discovery
Component 3: Data Stewardship
Component 4: Collaboration
Component 5: Training
Conceptual Framework: Security from the beginning

Federal Risk and Authorization Management Program
Provides a standardized cloud-based approach:
- security assessment
- authorization
- continuous monitoring

ADRF Status
May: Successful readiness assessment
June: Census Authority to Test
July: Title 13 Census data ingested
September: Full Assessment
February 2018: Census Authorization to Operate
June 2018: HHS Authorization to Operate in process
ADRF SaaS
Components

Component 1: Security
Component 2: Data Discovery
Component 3: Data Stewardship
Component 4: Collaboration
Component 5: Training
Related to data you’ve viewed

New data similar to data you've used

What others have done with similar data (recipes)

Recipes like yours

Thank you Charlie Catlett
Data Discovery

• Step 1: Create the set of corpora and metadata (computer science technology)
• Step 2: Figure out how you learn from it and automate it (machine learning techniques)
• Step 3: Gamification – recognize and emphasize patterns (with human curation)
Implementation: Search and Discovery
Components

Component 1: Security
Component 2: Data Discovery
Component 3: Data Stewardship
Component 4: Collaboration
Component 5: Training
Components

Component 1: Security
Component 2: Data Discovery
Component 3: Data Stewardship
Component 4: Collaboration
Component 5: Training
Implementation: Collaboration

Data Provider → User

Data Stewardship → Projects Module → ADRF Explorer

ADRF Staff

REST API

DF Admin

ADRF Secure Border

Sync and Control

LDAP (Auth)

Workspaces + Analytical Tools
hi folks - for anyone using IDHS data in their projects we have a bit more info on programs to help welfare recipients find stable jobs (thanks to Susan H for posing question and Rick Hendra for a great response!) - this doc will also be linked on the class website: https://docs.google.com/document/d/1GTnuPAWxxtw3CUncX238cWwVbz6FAdhI5O1pXsuNgg/edit?usp=sharing

clayton.hunter 11:48 AM
shared this file:

Job assistance programs for welfare recipients

Job assistance programs for welfare recipients

Question posed:
We are trying to add some context to our project and I wondered if you had a contact person at the Illinois DRS that could help fill in some questions about programs available to TANF/benefit recipients. I looked on the DRS website and while they do have a lot of information, there’s not much on programs available to help recipients move to stable jobs. For instance, there’s a program called EPIC directed towards SNAP recipients, but I haven’t found much else.

Response from Richard Hendra, MDRD:
Yes, we have very specific guidance as we worked on this particular issue there. The ESA evaluation had a site in Chicago that was focused on providing TANF recipients with stable jobs. The short term report here had more detail about the program, the implementation and the interim effects. Note that the UI data had major coverage issues with the segment of the TANF caseload that we were working with. The final results are in this giant report. I’d suggest the interim (shorter) report. We used various measures of employment stability. A common measure is the extent to which individuals worked in 4 consecutive
Components

Component 1: Security
Component 2: Data Discovery
Component 3: Data Stewardship
Component 4: Collaboration
Component 5: Training
“If you work in social science and would like to explore the power of big data, this book is clearly for you...This book is complete and comprehensive. It covers all necessary steps to finish a big data project; collecting raw data, cleaning and preprocessing data, applying various modeling tools to analyze the data, evaluating results, protecting privacy, and addressing ethical problems...All the important topics concerning big data are covered, making this book a good reference that you should always keep on your desk.” (2017) Book Reviews, *Journal of the American Statistical Association*, 112:518, 878-882, DOI: 10.1080/01621459.2017.1325629
Content Example: Machine Learning

Problem Formulation

Exercise 2

We have only scratched the surface of what we can do with our model. We've only tried one classifier (Logistic Regression), and there are plenty more classification algorithms in sklearn. Let's try them!

```python
clfs = {'RF': RandomForestClassifier(n_estimators=50, n_jobs=-1),
        'ET': ExtraTreesClassifier(n_estimators=10, n_jobs=-1, criterion='entropy'),
        'LR': LogisticRegression(penalty='1l', C=1e5),
        'SGD': SGDClassifier(loss='log'),
        'GB': GradientBoostingClassifier(learning_rate=0.05, subsample=0.5, max_depth=6, n_estimators),
        'NB': GaussianNB()}

sel_clfs = ['RF', 'ET', 'LR', 'SGD', 'GB', 'NB']

max_p_at_k = 0
for clfNM in sel_clfs:
    clf = clfs[clfNM]
    clf.fit(X_train, y_train)
    print clf
    y_score = clf.predict_proba(X_test)[1,1]
```

Four Key Points:
- Define
- Prepare
- Define
- Build
Products: Corrections and Employment

Table 1 summarizes the median time spent in different states for each cluster.

Table 1. Median Time Spent in Each State by Cluster

Table 4. Recidivism Rates by Cluster

<table>
<thead>
<tr>
<th></th>
<th>At least one incident of recidivism</th>
<th>At least one technical violation</th>
<th>Technical violations as a percent of recidivism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full cohort</td>
<td>53%</td>
<td>31%</td>
<td>60%</td>
</tr>
<tr>
<td>Primarily incarcerated</td>
<td>41%</td>
<td>26%</td>
<td>65%</td>
</tr>
<tr>
<td>Intermittent employment</td>
<td>66%</td>
<td>39%</td>
<td>58%</td>
</tr>
<tr>
<td>Unemployed after initial incarceration</td>
<td>23%</td>
<td>14%</td>
<td>61%</td>
</tr>
<tr>
<td>Intermittent incarceration</td>
<td>99%</td>
<td>66%</td>
<td>67%</td>
</tr>
<tr>
<td>Working after incarceration</td>
<td>43%</td>
<td>21%</td>
<td>49%</td>
</tr>
</tbody>
</table>

Figure 2 Cluster Analysis: Five clusters were identified from the trajectories.
Tailored and Customizable Metrics

Fig. 2: Dashboard metrics (left) and industry subsets (right)
The dashboard can visualize different metrics (left) – including QWI metrics developed in in the context of the Census LEHD program –, subsetting the data by different industries (right).
Comparing Employment Dynamics Across Borders

Fig. 3: Comparing total earnings with Illinois border counties

The dashboard can include border counties from the states that provide data to the ADRF.
Outline

Rethinking measurement
Operationalizing
A possible approach
- Human
- Technical
Next steps
Identify key priority areas

• Crime/justice
• Social problems (opioids)
• Education
• Workforce/economic development

Leveraging Data as a Strategic Asset

Goal Leaders

Pradeep Belur,
Chief of Staff, Small Business Administration

Karen Dunn Kelley,
Under Secretary of Economic Affairs and Acting Deputy Secretary, Department of Commerce

Jack Wilmer,
Senior Advisor for Cybersecurity and IT Modernization, Office of

Goal Statement

Leverage data as a strategic asset to grow the economy, increase the effectiveness of the Federal Government, facilitate oversight, and promote transparency.

The Challenge

The use of data is transforming society, business, and the economy. Data provided by the Federal Government have a unique place in society and maintaining trust in Federal data is pivotal to a democratic process. The Federal Government needs a robust, integrated approach to using data to deliver on mission, serve customers, and steward resources while respecting privacy and confidentiality.
Key ideas

• Economy has changed substantially => new measures necessary
• Enormous potential with new data
• Statistical agencies have new role
• We need to build new demand-driven institutions – local plus federal
• We need to stop and think
Comments welcome

• Julia Lane
• Julia.lane@nyu.edu