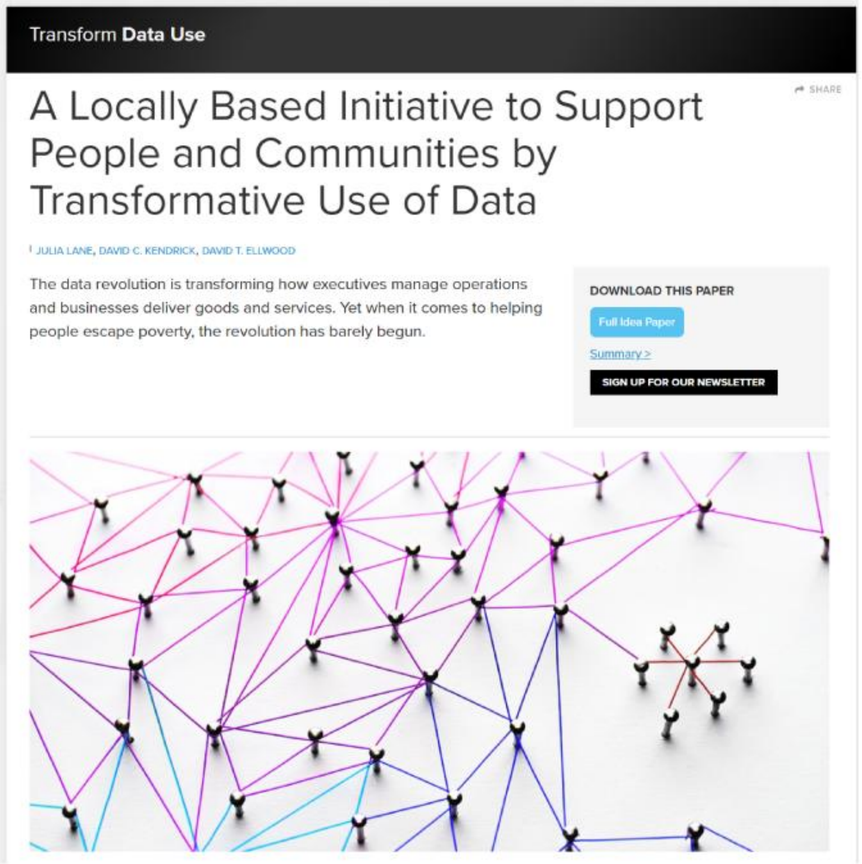
Making Data Great Again (or data, data, everywhere – we have to stop and think)

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



Transform **Data Use**

SHARE

A Locally Based Initiative to Support People and Communities by Transformative Use of Data

JULIA LANE, DAVID C. KENDRICK, DAVID T. ELLWOOD

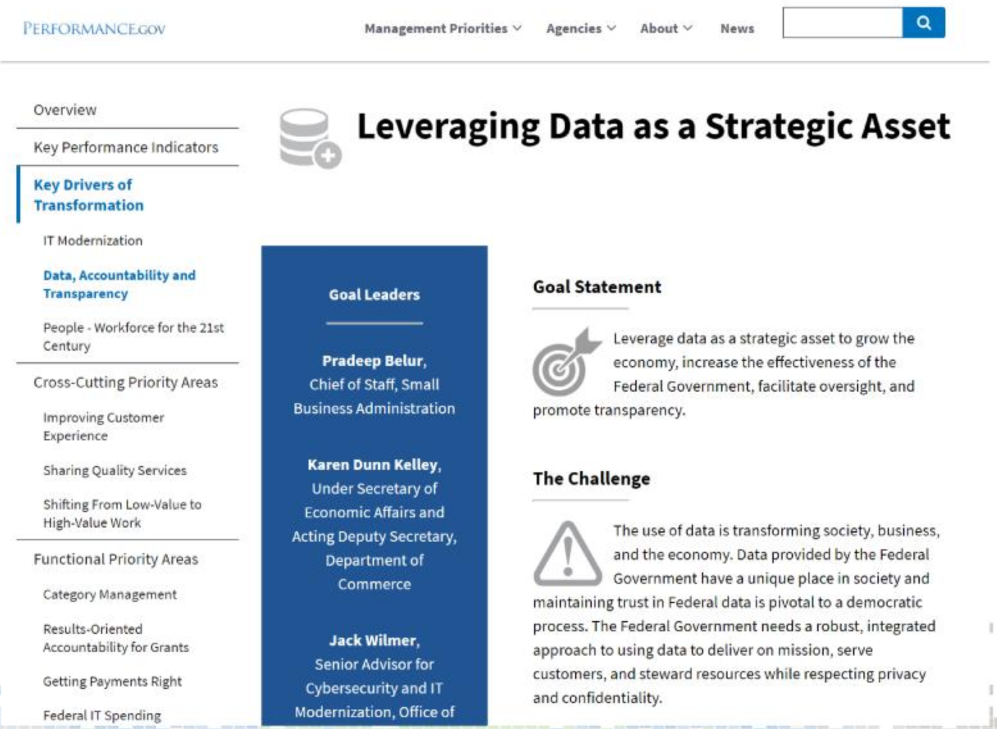
The data revolution is transforming how executives manage operations and businesses deliver goods and services. Yet when it comes to helping people escape poverty, the revolution has barely begun.

DOWNLOAD THIS PAPER

Full Idea Paper

Summary >

SIGN UP FOR OUR NEWSLETTER



- Overview
- Key Performance Indicators
- **Key Drivers of Transformation**
  - IT Modernization
  - **Data, Accountability and Transparency**
  - People - Workforce for the 21st Century
- Cross-Cutting Priority Areas
  - Improving Customer Experience
  - Sharing Quality Services
  - Shifting From Low-Value to High-Value Work
- Functional Priority Areas
  - Category Management
  - Results-Oriented Accountability for Grants
  - Getting Payments Right
  - Federal IT Spending

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.

Outline

Rethinking measurement

Operationalizing

A possible approach

- Human
- Technical

Next steps

Outline

Rethinking measurement

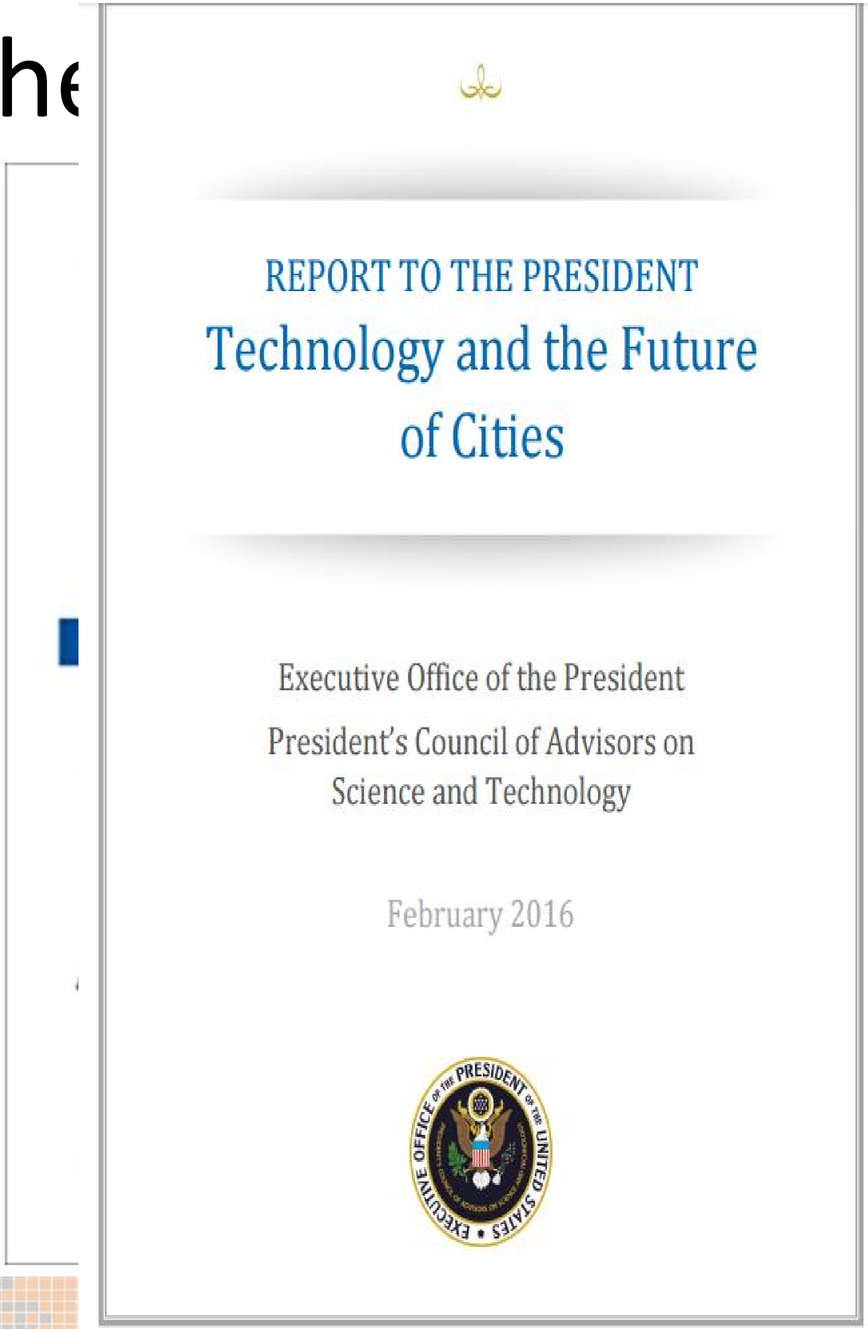
Operationalizing

A possible approach

- Human
- Technical

Next steps

Whe d?



REPORT TO THE PRESIDENT

Technology and the Future of Cities

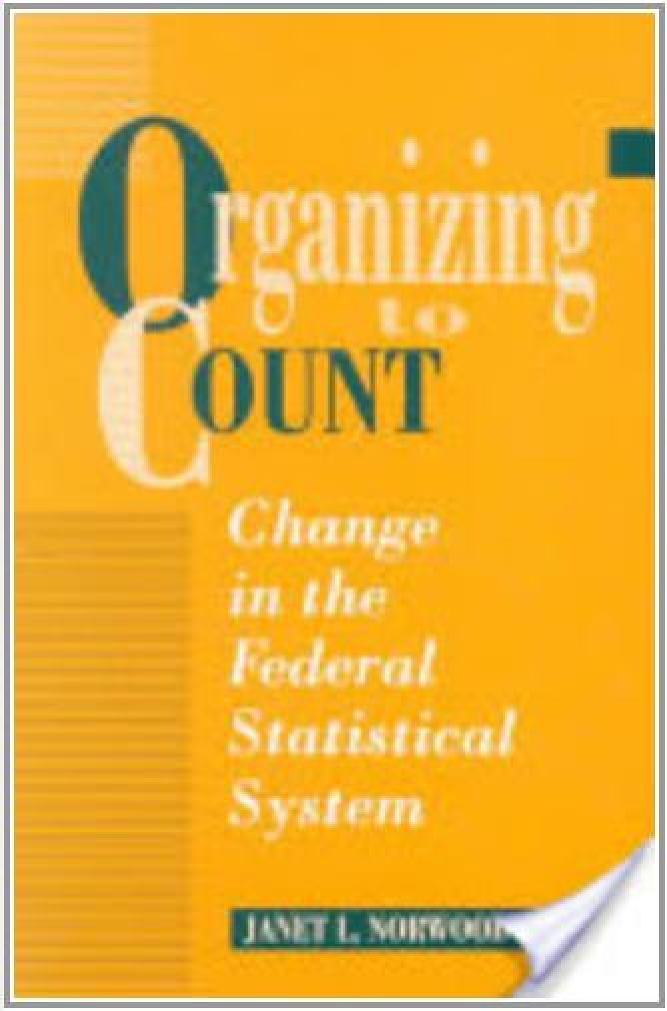
Executive Office of the President

President's Council of Advisors on Science and Technology

February 2016

Demand in previous century

- Great Depres
- Wartime eco
- Colin Clark, S ard Stone



Organizing to Count

Change in the Federal Statistical System

JANET L. NORWOOD

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

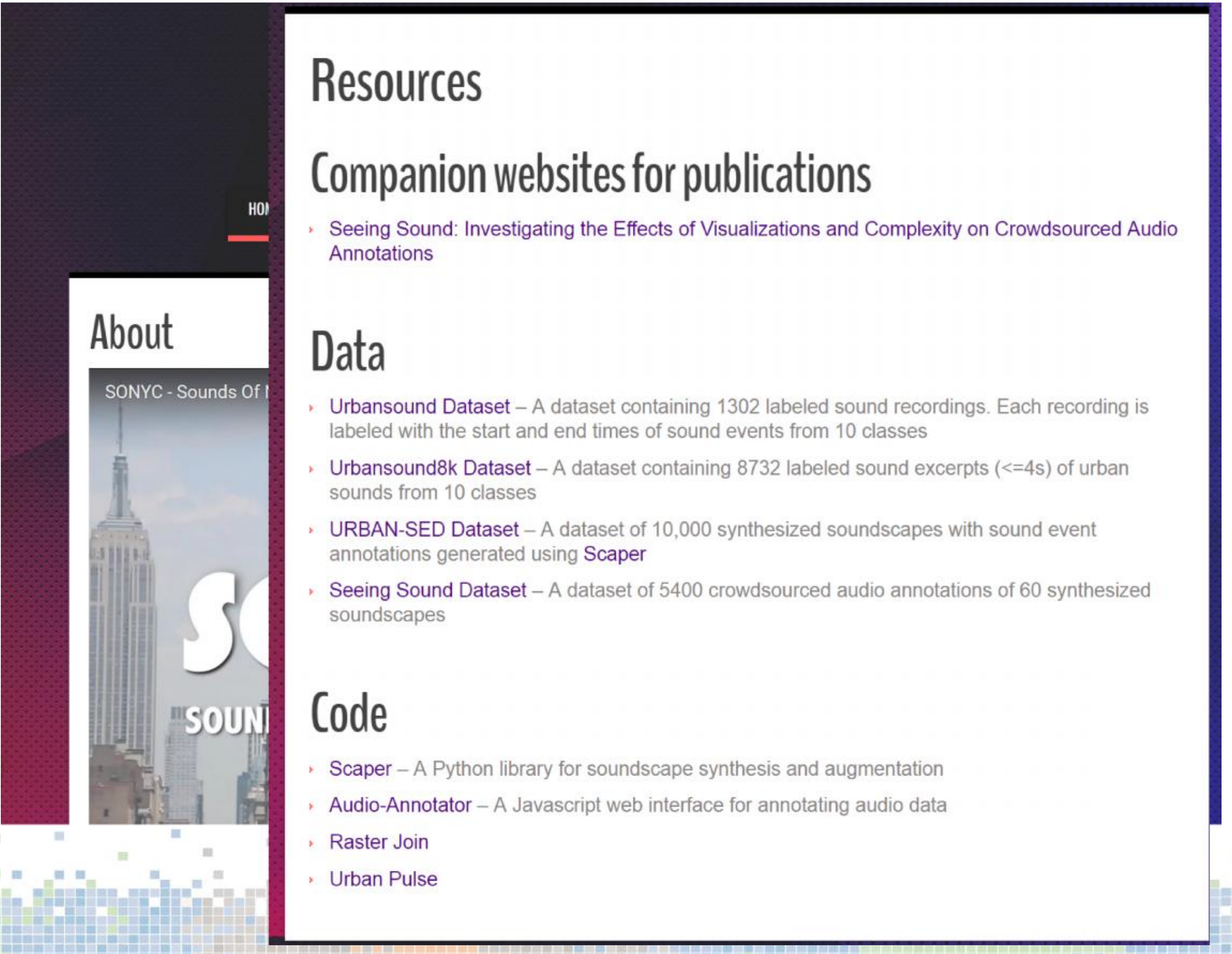
Operationalizing

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

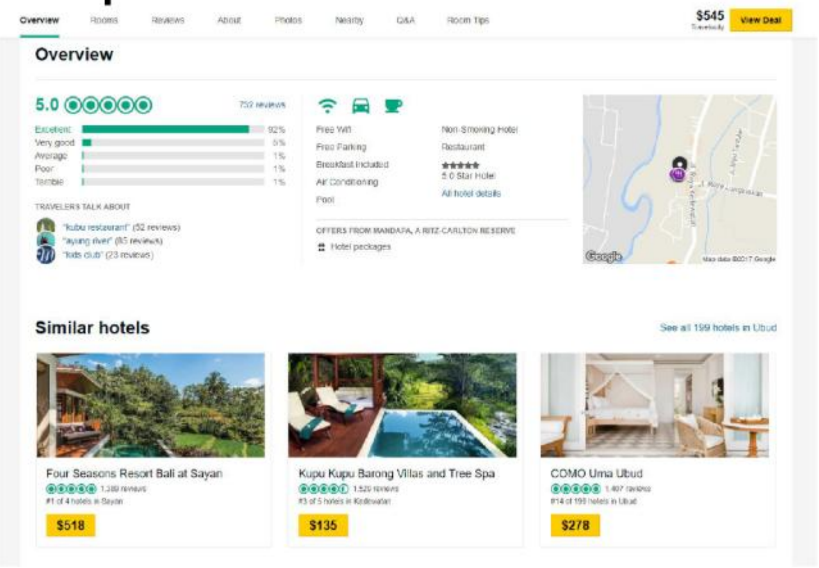
What is needed?

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

How do we trade off?

Collection, documentation, Curation

TripAdvisor



Overview Rooms Reviews About Photos Nearby Q&A Room Tips

$545 View Deal

Overview

5.0 ●●●●● 752 reviews

Excellent	92%
Very good	5%
Average	1%
Poor	1%
Terrible	1%

TRAVELERS TALK ABOUT

"kubu restaurant" (52 reviews)
"ayung river" (85 reviews)
"kids club" (23 reviews)

Free Wifi
Free Parking
Breakfast Included
Air Conditioning
Pool

Non-Smoking Hotel
Restaurant
★★★★★
5.0 Star Hotel
All hotel details

OFFERS FROM MANDAPA, A RITZ-CARLTON RESERVE

Hotel packages

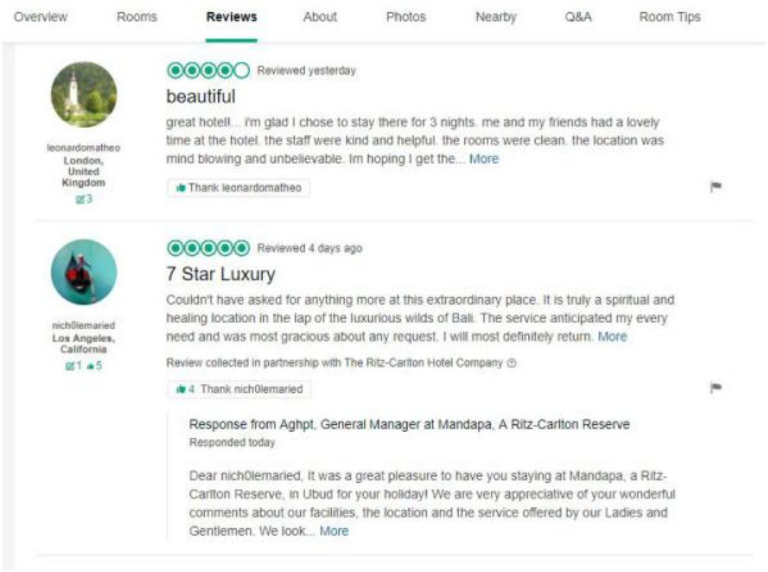
Similar hotels

See all 199 hotels in Ubud

Four Seasons Resort Bali at Sayan
●●●●● 1,389 reviews
#1 of 4 hotels in Sayan
$518

Kupu Kupu Barong Villas and Tree Spa
●●●●◐ 1,520 reviews
#3 of 5 hotels in Kedewatan
$135

COMO Uma Ubud
●●●●● 1,407 reviews
#14 of 199 hotels in Ubud
$278



Overview Rooms Reviews About Photos Nearby Q&A Room Tips

leonardomatheo
London, United Kingdom
3

●●●●○ Reviewed yesterday

beautiful

great hotell... i'm glad I chose to stay there for 3 nights. me and my friends had a lovely time at the hotel. the staff were kind and helpful. the rooms were clean. the location was mind blowing and unbelievable. Im hoping I get the... More

Thank leonardomatheo

nich0lemaried
Los Angeles, California
1 5

●●●●● Reviewed 4 days ago

7 Star Luxury

Couldn't have asked for anything more at this extraordinary place. It is truly a spiritual and healing location in the lap of the luxurious wilds of Bali. The service anticipated my every need and was most gracious about any request. I will most definitely return. More

Review collected in partnership with The Ritz-Carlton Hotel Company

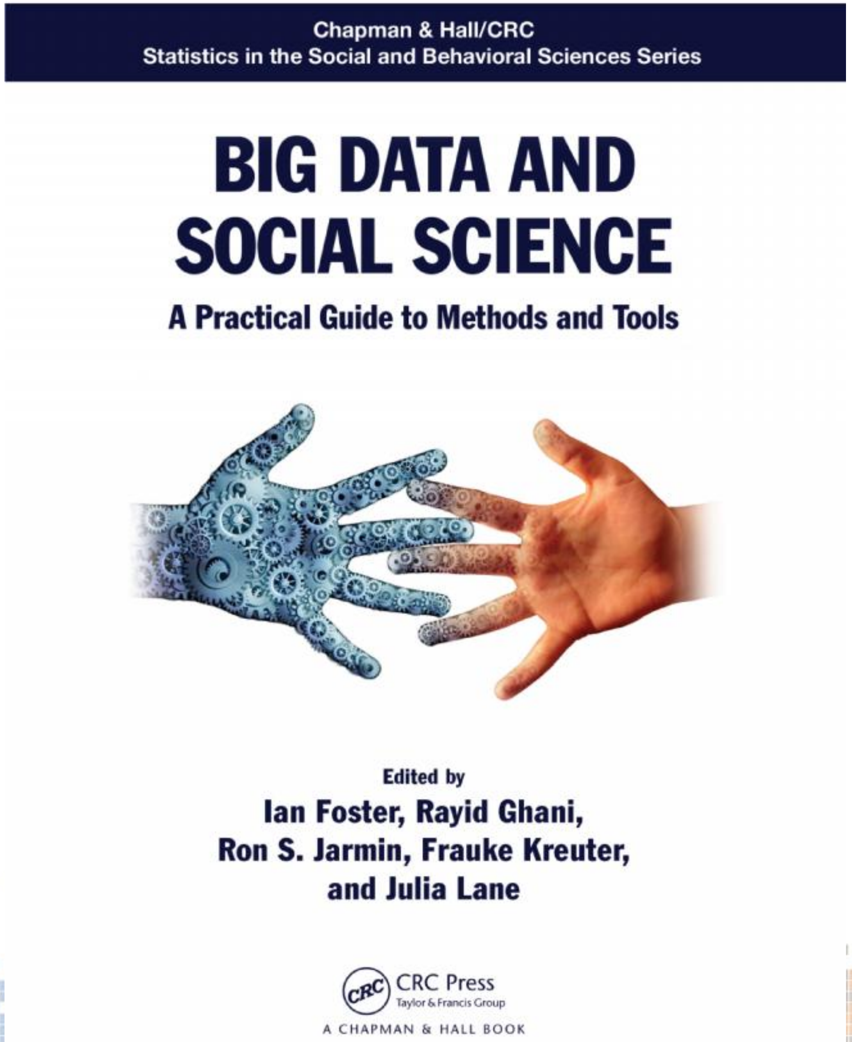
4 Thank nich0lemaried

Response from Aghpt, General Manager at Mandapa, A Ritz-Carlton Reserve
Responded today

Dear nich0lemaried, It was a great pleasure to have you staying at Mandapa, a Ritz-Carlton Reserve, in Ubud for your holiday! We are very appreciative of your wonderful comments about our facilities, the location and the service offered by our Ladies and Gentlemen. We look... More

New skills

- Framing question
- Webscraping/APIs
- Data Management
- Linkage
- Machine Learning
- Text Analysis
- Graph Analysis
- Visualization
- Inference
- Privacy and Confidentiality



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

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

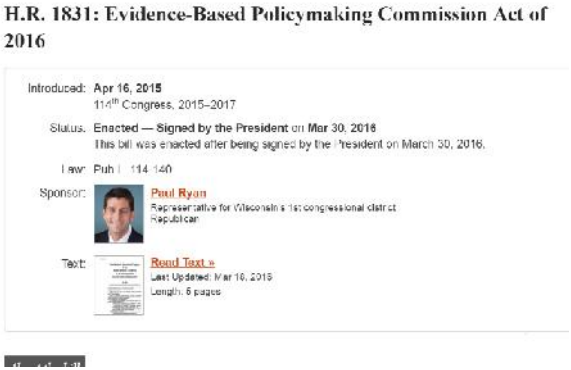
Next steps

"If HP only knew what HP knows, we would be much more profitable"

(former CEO Lew Platt)

al

Federal level



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

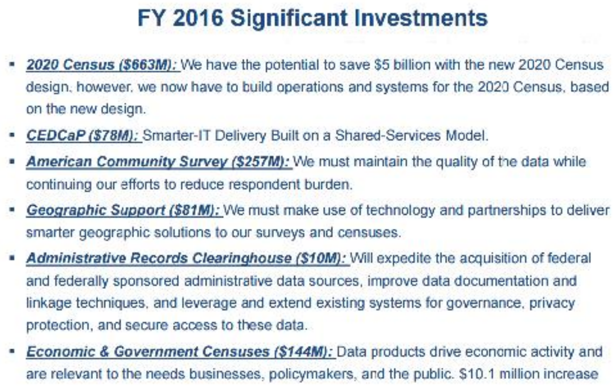
Introduced: **Apr 16, 2015**
114th Congress, 2015–2017

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

Law: Pub.L. 114-140

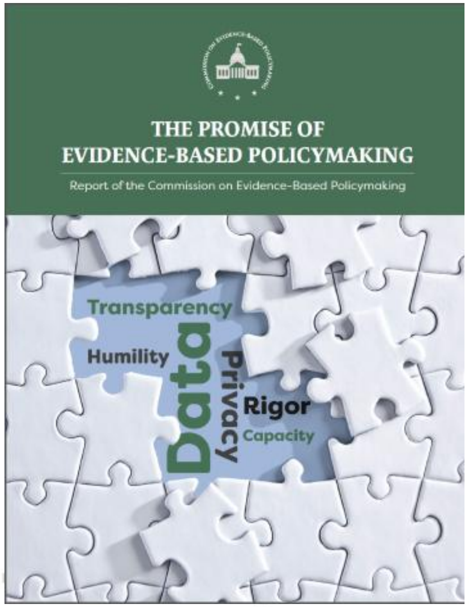
Sponsor: Paul Ryan
Representative for Wisconsin's 1st congressional district
Republican

Text: Read Text »
Last Updated: Mar 16, 2016
Length: 6 pages



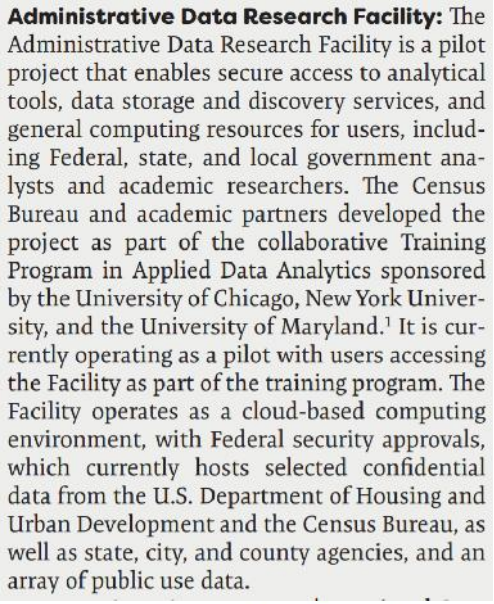
FY 2016 Significant Investments

- ***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 ($78M):*** 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 ($10M):*** 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



THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking



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.[1] 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.

A number of barriers

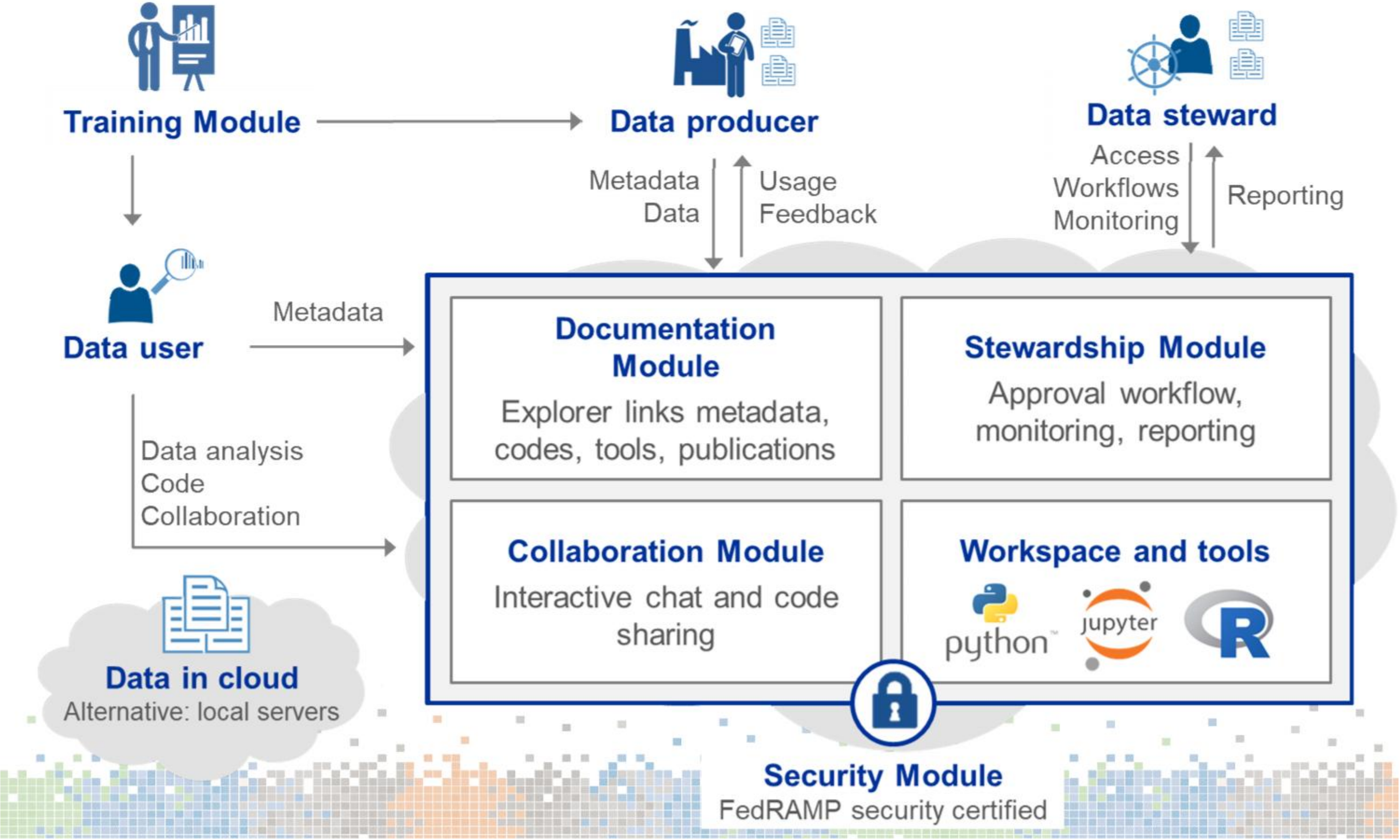
Technical

- cost
- burden
- data quality

Human

- data documentation
- risk of bad analysis
- legal mandates surrounding data access and use
- Workforce capacity

A possible approach



Training Module

Data producer

Data steward

Metadata
Data

Usage
Feedback

Access
Workflows
Monitoring

Reporting

Data user

Metadata

Data analysis
Code
Collaboration

Documentation Module
Explorer links metadata, codes, tools, publications

Stewardship Module
Approval workflow, monitoring, reporting

Collaboration Module
Interactive chat and code sharing

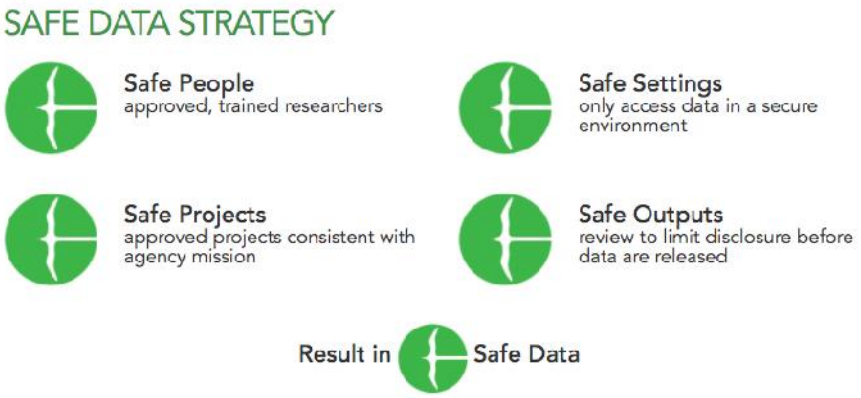
Workspace and tools

Data in cloud
Alternative: local servers

Security Module
FedRAMP security certified

Our approach

Secure **computing & analytics** platform



SAFE DATA STRATEGY

Safe People
approved, trained researchers

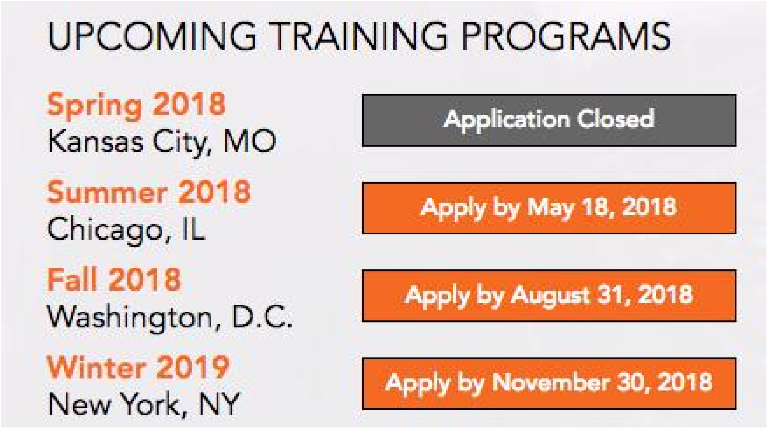
Safe Settings
only access data in a secure environment

Safe Projects
approved projects consistent with agency mission

Safe Outputs
review to limit disclosure before data are released

Result in Safe Data

Analytics **training** programs



UPCOMING TRAINING PROGRAMS

Spring 2018 Kansas City, MO	Application Closed
Summer 2018 Chicago, IL	Apply by May 18, 2018
Fall 2018 Washington, D.C.	Apply by August 31, 2018
Winter 2019 New York, NY	Apply by November 30, 2018

Outline

Rethinking measurement

Operationalizing

A possible approach

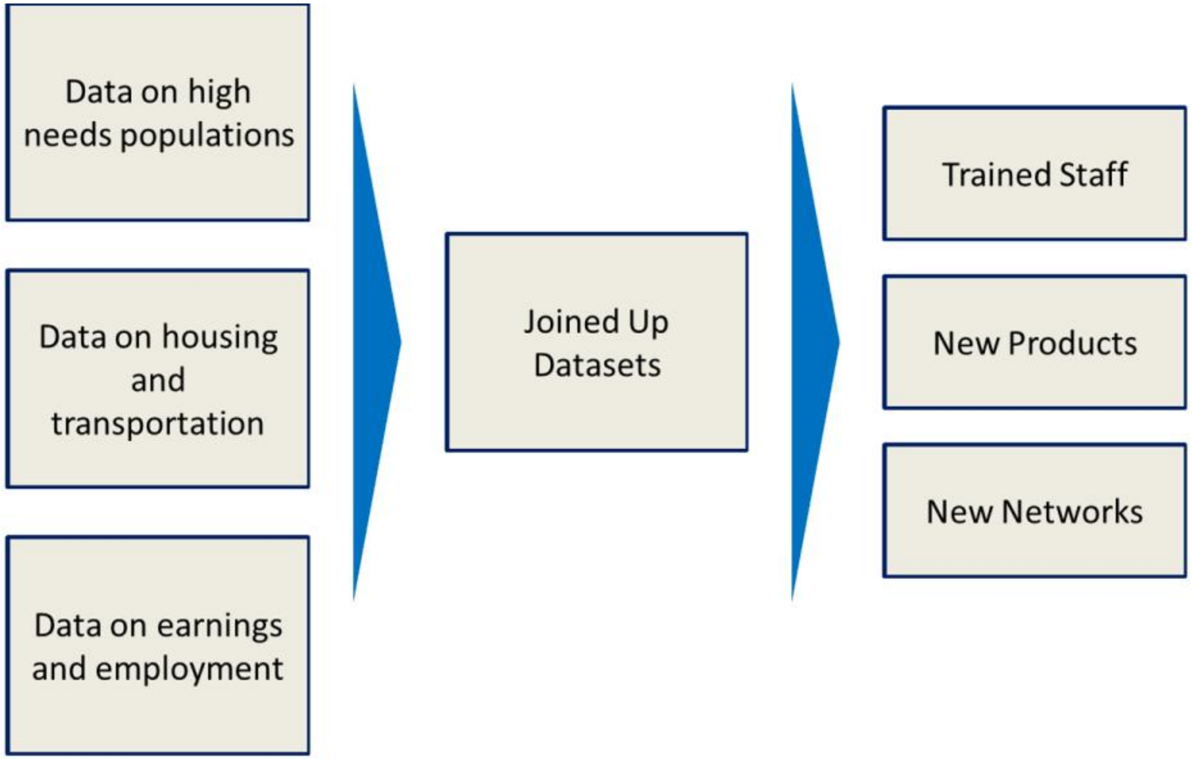
- Human
- Technical

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

Specifics



Data on high needs populations

Data on housing and transportation

Data on earnings and employment

Joined Up Datasets

Trained Staff

New Products

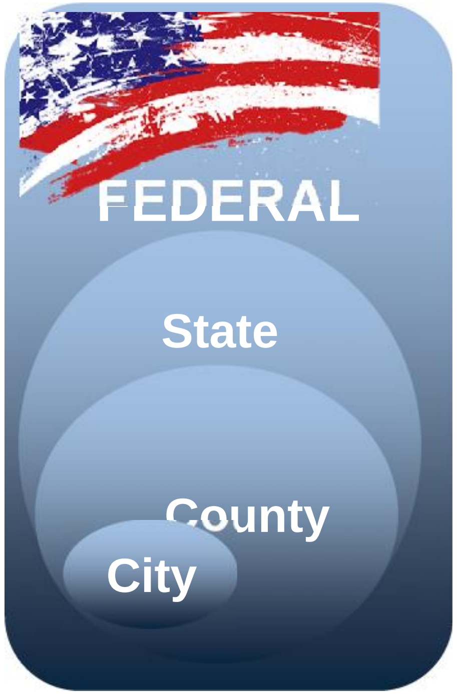
New Networks

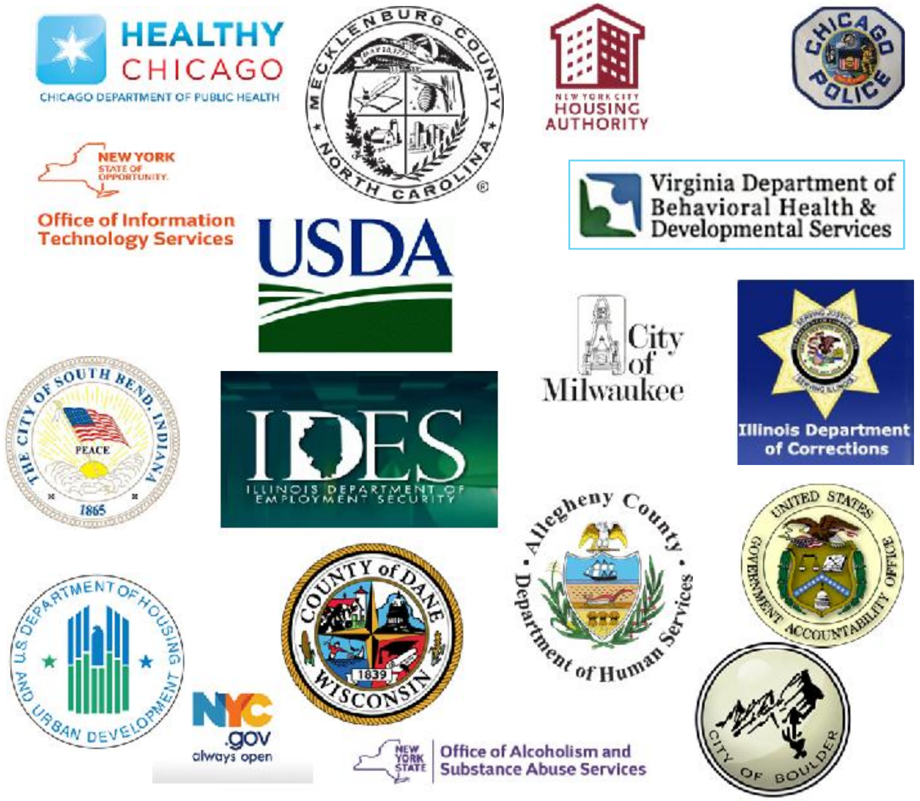
Results: Over 40 Confidential Datasets

Federal (6)	States (12)	Cities (15)	Counties (9)
Census (LEHD and ACS)	Labor (Wage records, QCEW, UI claims)	NYPD	King County Transportation, Human Services
HUD (Housing Choice Voucher Program, Public Housing, Project-based Section 8, and the Section 202/811 Programs)	Human Services (TANF, SNAP)	Chicago PD	Mecklenburg County Corrections
	Corrections (admissions and exits)	NYC Labor, Human Services, Corrections, Homeless, VocEd	
	Revenue (Business tax)		

Team work

Networks: >90 govt agencies; >200 participants




HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH
MECKLENBURG COUNTY NORTH CAROLINA
NEW YORK CITY HOUSING AUTHORITY
CHICAGO POLICE
NEW YORK STATE OF OPPORTUNITY.
Office of Information Technology Services
USDA
Virginia Department of Behavioral Health & Developmental Services
City of Milwaukee
Illinois Department of Corrections
THE CITY OF SOUTH BEND, INDIANA
PEACE
1865
IDES
ILLINOIS DEPARTMENT OF EMPLOYMENT SECURITY
Allegheny County • Department of Human Services •
UNITED STATES GOVERNMENT ACCOUNTABILITY OFFICE
U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
COUNTY of DANE WISCONSIN 1839
NYC .gov always open
NEW YORK STATE Office of Alcoholism and Substance Abuse Services
CITY OF BOULDER

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."

Danielle Fulmer
Director of Business Analytics

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.

Katy Fitzgerald
Management Analyst

Outline

Rethinking measurement

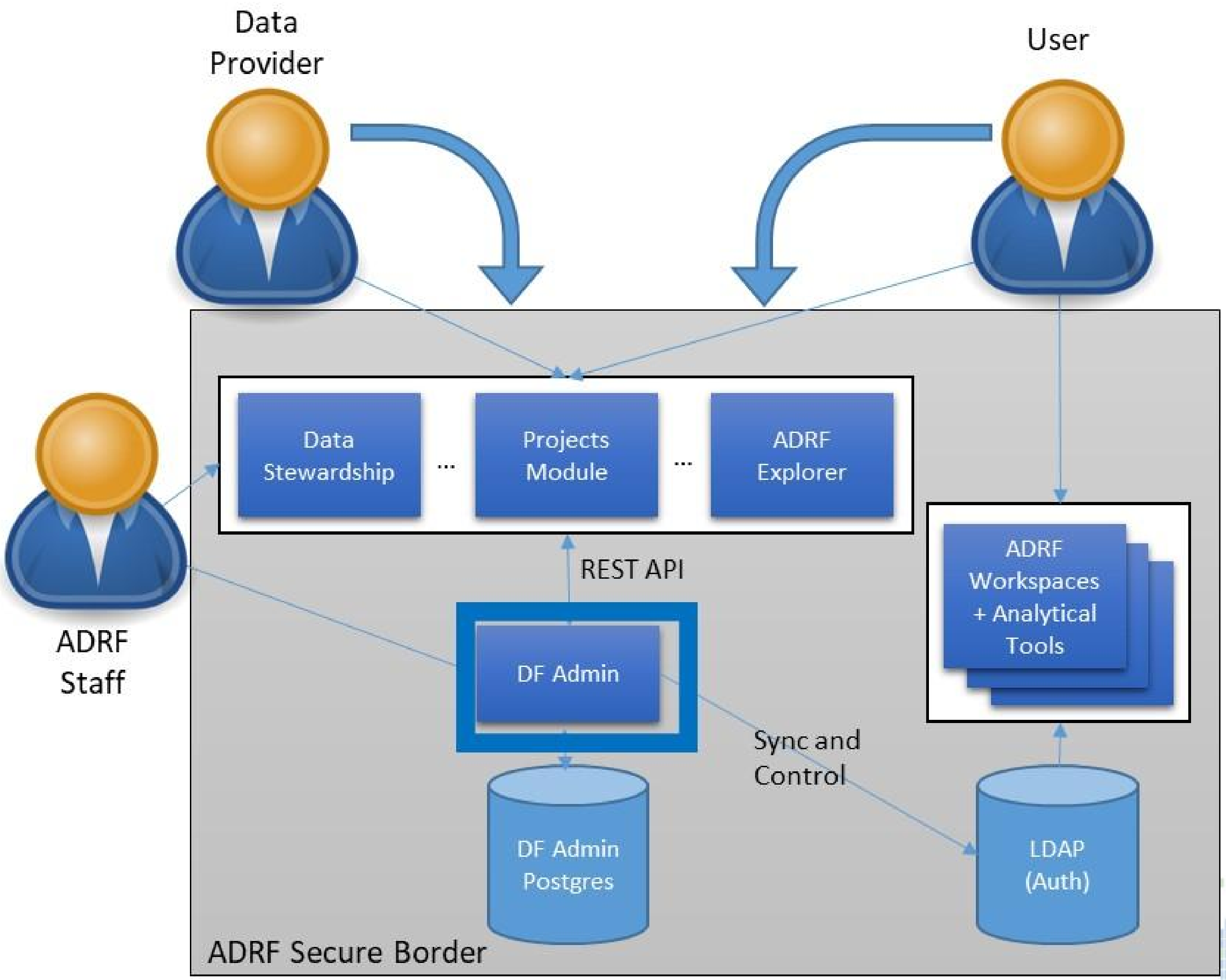
Operationalizing

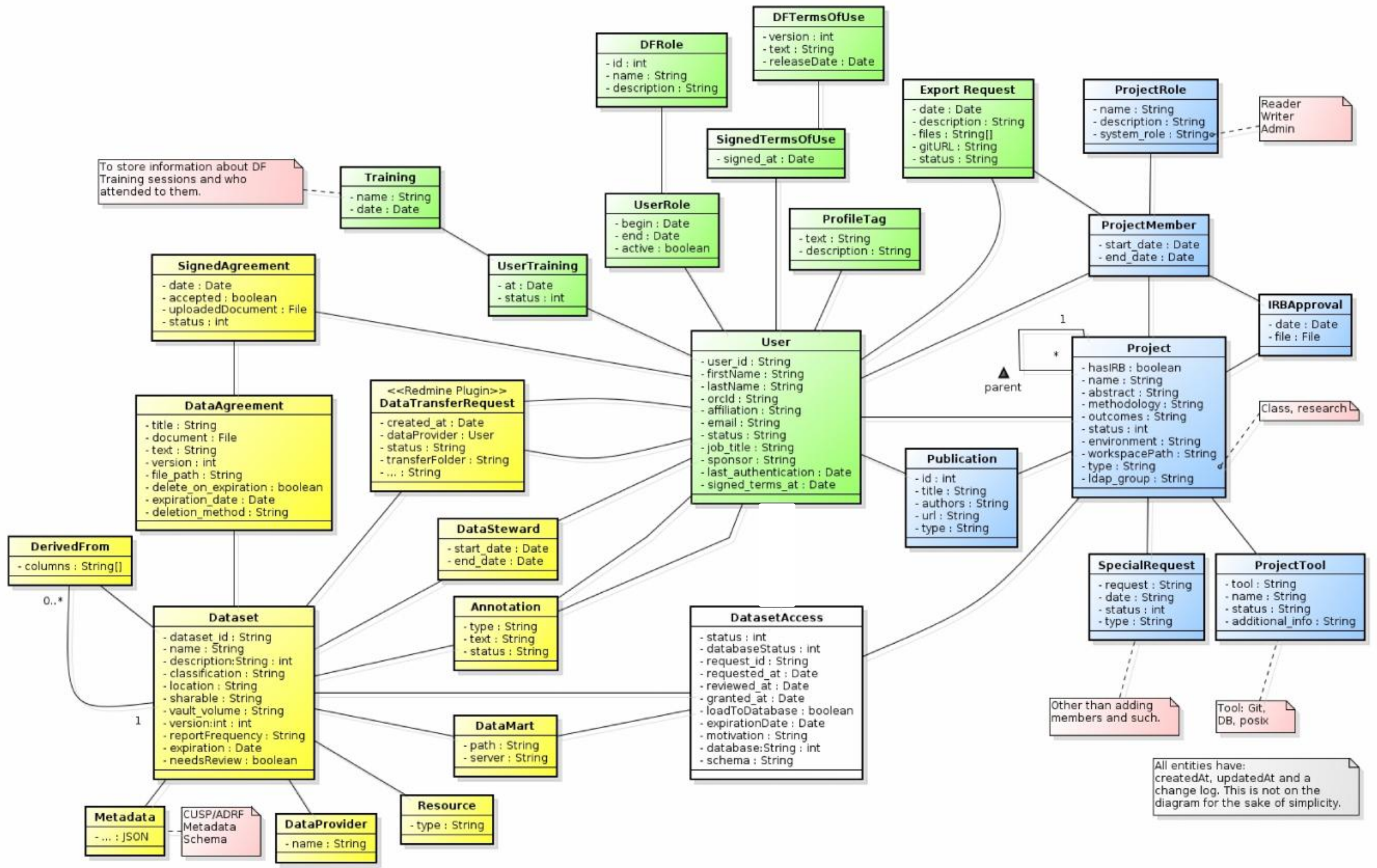
A possible approach

- Human
- Technical

Next steps

Conceptual Framework: User Needs





DFRole
- id : int
- name : String
- description : String

DFTermsOfUse
- version : int
- text : String
- releaseDate : Date

SignedTermsOfUse
- signed_at : Date

Export Request
- date : Date
- description : String
- files : String[]
- gitURL : String
- status : String

ProjectRole
- name : String
- description : String
- system_role : String

Reader
Writer
Admin

To store information about DF Training sessions and who attended to them.

Training
- name : String
- date : Date

UserRole
- begin : Date
- end : Date
- active : boolean

ProfileTag
- text : String
- description : String

ProjectMember
- start_date : Date
- end_date : Date

SignedAgreement
- date : Date
- accepted : boolean
- uploadedDocument : File
- status : int

UserTraining
- at : Date
- status : int

IRBApproval
- date : Date
- file : File

User
- user_id : String
- firstName : String
- lastName : String
- orcId : String
- affiliation : String
- email : String
- status : String
- job_title : String
- sponsor : String
- last_authentication : Date
- signed_terms_at : Date

Project
- hasIRB : boolean
- name : String
- abstract : String
- methodology : String
- outcomes : String
- status : int
- environment : String
- workspacePath : String
- type : String
- ldap_group : String

Class, research

DataAgreement
- title : String
- document : File
- text : String
- version : int
- file_path : String
- delete_on_expiration : boolean
- expiration_date : Date
- deletion_method : String

<<Redmine Plugin>>
DataTransferRequest
- created_at : Date
- dataProvider : User
- status : String
- transferFolder : String
- ... : String

Publication
- id : int
- title : String
- authors : String
- url : String
- type : String

DataSteward
- start_date : Date
- end_date : Date

DerivedFrom
- columns : String[]

SpecialRequest
- request : String
- date : String
- status : int
- type : String

ProjectTool
- tool : String
- name : String
- status : String
- additional_info : String

Annotation
- type : String
- text : String
- status : String

Dataset
- dataset_id : String
- name : String
- description:String : int
- classification : String
- location : String
- sharable : String
- vault_volume : String
- version:int : int
- reportFrequency : String
- expiration : Date
- needsReview : boolean

DatasetAccess
- status : int
- databaseStatus : int
- request_id : String
- requested_at : Date
- reviewed_at : Date
- granted_at : Date
- loadToDatabase : boolean
- expirationDate : Date
- motivation : String
- database:String : int
- schema : String

Other than adding members and such.

Tool: Git, DB, posix

DataMart
- path : String
- server : String

All entities have: createdAt, updatedAt and a change log. This is not on the diagram for the sake of simplicity.

Resource
- type : String

Metadata
- ... : JSON

CUSP/ADRF Metadata Schema

DataProvider
- name : String

Components

Component 1: Security

Component 2: Data Discovery

Component 3: Data Stewardship

Component 4: Collaboration

Component 5: Training

Conceptual Framework: Security from the beginning

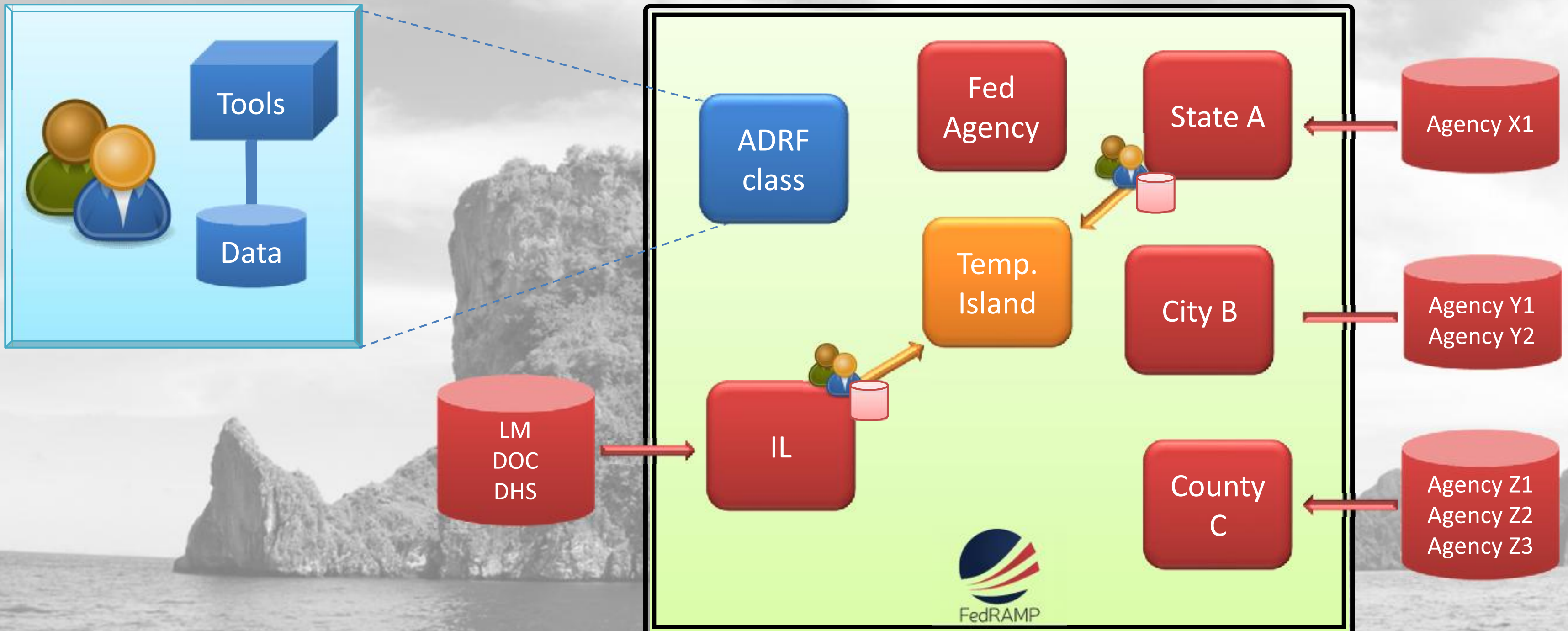
Federal Risk and Authorization Management Program

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


Tools
Data
ADRF class
Fed Agency
State A
Agency X1
Temp. Island
City B
Agency Y1
Agency Y2
LM
DOC
DHS
IL
County C
Agency Z1
Agency Z2
Agency Z3
FedRAMP

ADRF SaaS

Components

Component 1: Security

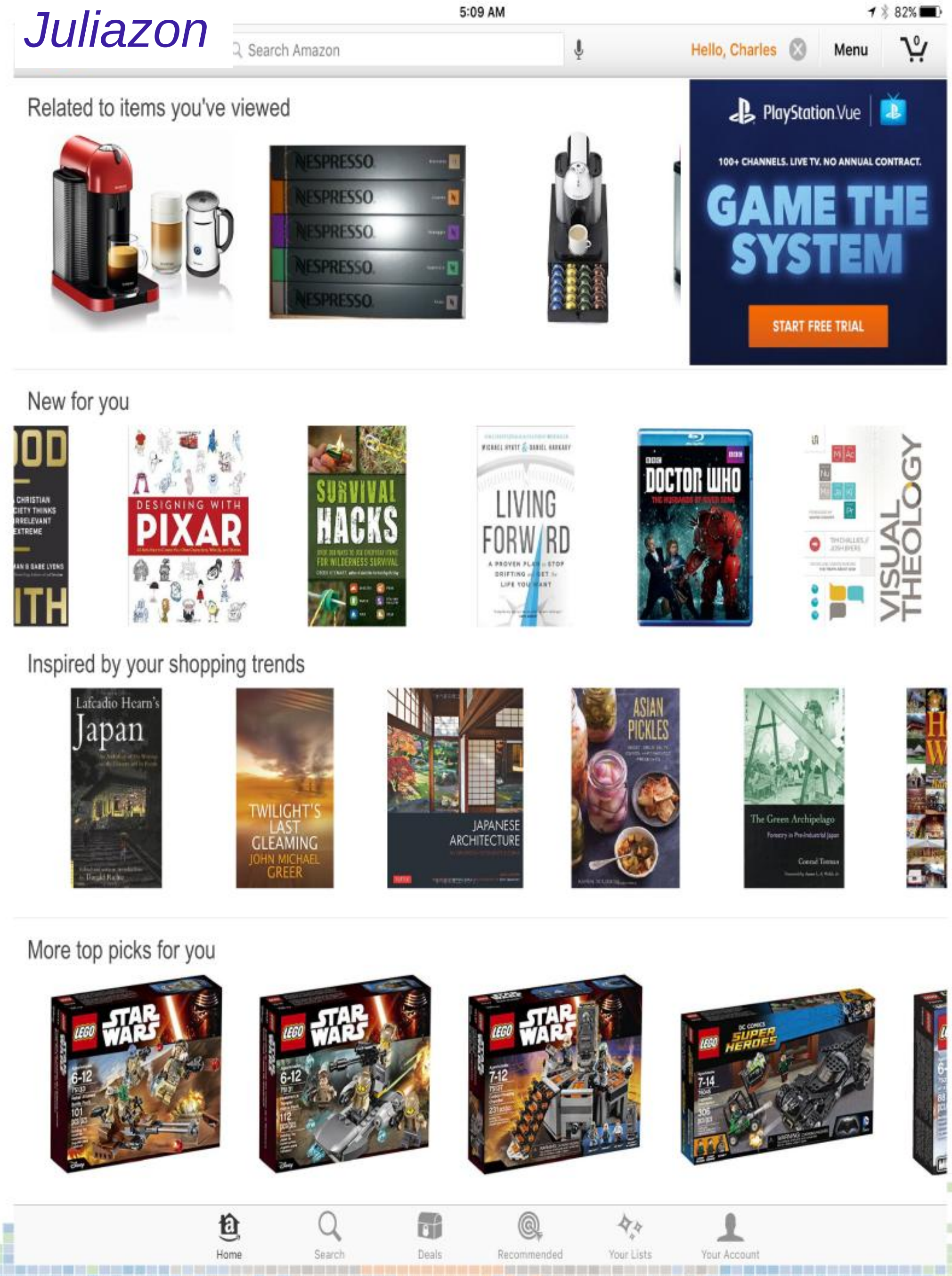
Component 2: Data Discovery

Component 3: Data Stewardship

Component 4: Collaboration

Component 5: Training

Juliazon



Related to data you've viewed

New data similar to data you've used

What others have done with similar data (recipes)

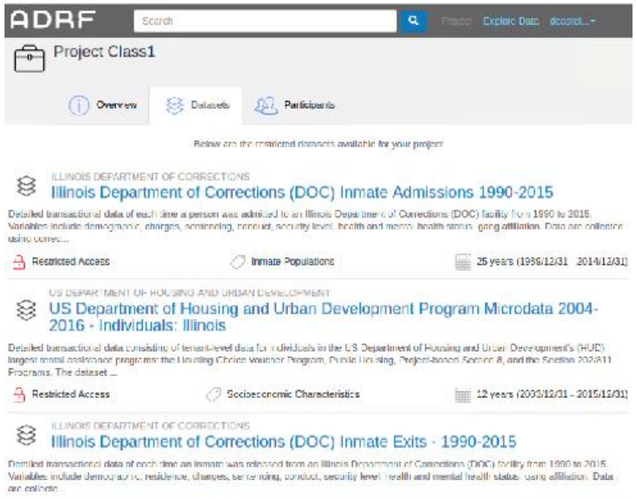
Recipes like yours

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



ADRF | Search | Project | Explore Data | dcostel...

Project Class1

Overview | Datasets | Participants

Below are the restricted datasets available for your project

ILLINOIS DEPARTMENT OF CORRECTIONS
Illinois Department of Corrections (DOC) Inmate Admissions 1990-2015

Detailed transactional data of each time a person was admitted to an Illinois Department of Corrections (DOC) facility from 1990 to 2015. Variables include demographic, charges, sentencing, conduct, security level, health and mental health status, gang affiliation. Data are collected using correc...

Restricted Access | Inmate Populations | 25 years (1989/12/31 2014/12/31)

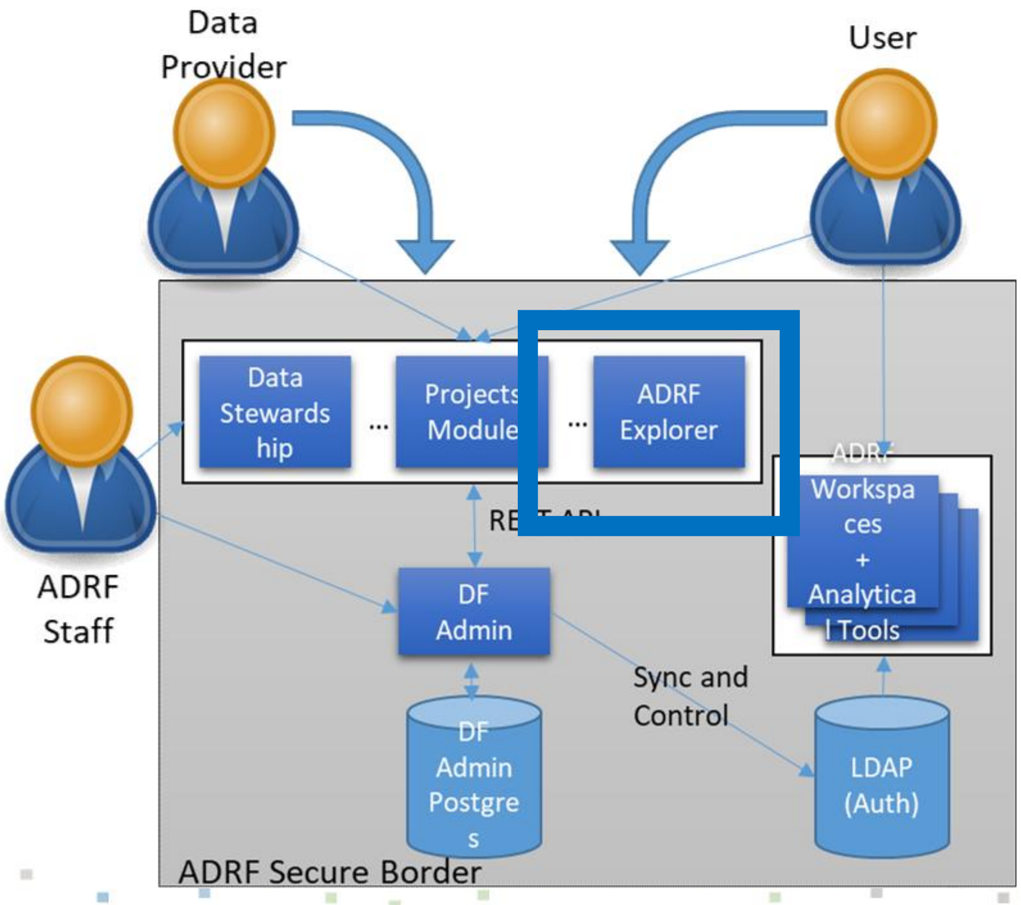
US DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
US Department of Housing and Urban Development Program Microdata 2004-2016 - Individuals: Illinois

Detailed transactional data consisting of tenant-level data for individuals in the US Department of Housing and Urban Development's (HUD) largest rental assistance programs: the Housing Choice Voucher Program, Public Housing, Project-based Section 8, and the Section 202/811 Programs. The dataset ...

Restricted Access | Socioeconomic Characteristics | 12 years (2003/12/31 - 2015/12/31)

ILLINOIS DEPARTMENT OF CORRECTIONS
Illinois Department of Corrections (DOC) Inmate Exits - 1990-2015

Detailed transactional data of each time an inmate was released from an Illinois Department of Corrections (DOC) facility from 1990 to 2015. Variables include demographic, residence, charges, sentencing, conduct, security level, health and mental health status, gang affiliation. Data are collecte...



Components

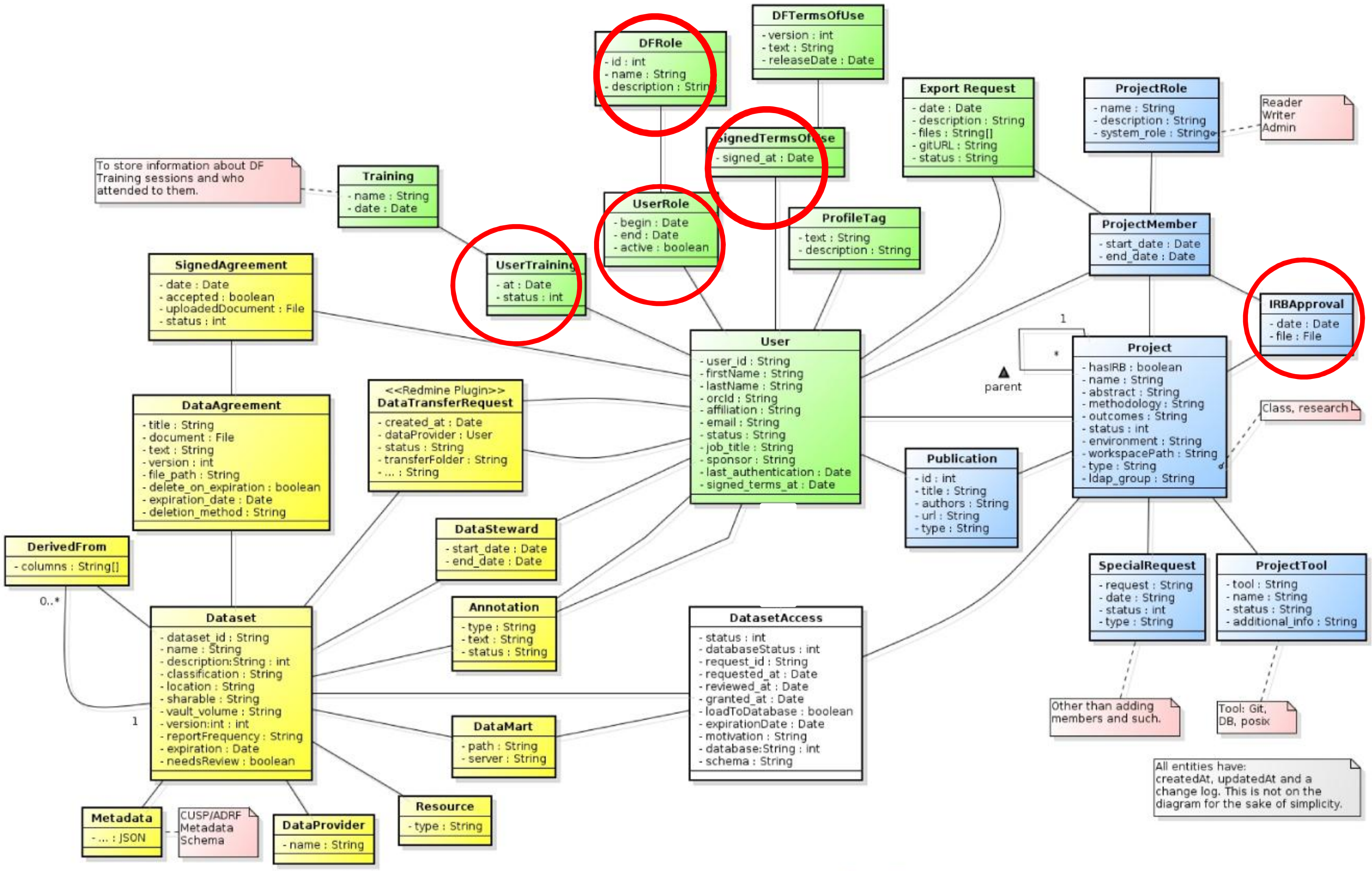
Component 1: Security

Component 2: Data Discovery

Component 3: Data Stewardship

Component 4: Collaboration

Component 5: Training



To store information about DF Training sessions and who attended to them.

DFRole
- id : int
- name : String
- description : Strin

DFTermsOfUse
- version : int
- text : String
- releaseDate : Date

Export Request
- date : Date
- description : String
- files : String[]
- gitURL : String
- status : String

ProjectRole
- name : String
- description : String
- system_role : String

Reader
Writer
Admin

Training
- name : String
- date : Date

SignedTermsOfUse
- signed_at : Date

UserRole
- begin : Date
- end : Date
- active : boolean

ProfileTag
- text : String
- description : String

ProjectMember
- start_date : Date
- end_date : Date

SignedAgreement
- date : Date
- accepted : boolean
- uploadedDocument : File
- status : int

UserTraining
- at : Date
- status : int

IRBApproval
- date : Date
- file : File

User
- user_id : String
- firstName : String
- lastName : String
- orcId : String
- affiliation : String
- email : String
- status : String
- job_title : String
- sponsor : String
- last_authentication : Date
- signed_terms_at : Date

Project
- hasIRB : boolean
- name : String
- abstract : String
- methodology : String
- outcomes : String
- status : int
- environment : String
- workspacePath : String
- type : String
- ldap_group : String

Class, research

DataAgreement
- title : String
- document : File
- text : String
- version : int
- file_path : String
- delete_on_expiration : boolean
- expiration_date : Date
- deletion_method : String

<<Redmine Plugin>>
DataTransferRequest
- created_at : Date
- dataProvider : User
- status : String
- transferFolder : String
- ... : String

Publication
- id : int
- title : String
- authors : String
- url : String
- type : String

DerivedFrom
- columns : String[]

DataSteward
- start_date : Date
- end_date : Date

SpecialRequest
- request : String
- date : String
- status : int
- type : String

ProjectTool
- tool : String
- name : String
- status : String
- additional_info : String

Dataset
- dataset_id : String
- name : String
- description:String : int
- classification : String
- location : String
- sharable : String
- vault_volume : String
- version:int : int
- reportFrequency : String
- expiration : Date
- needsReview : boolean

Annotation
- type : String
- text : String
- status : String

DatasetAccess
- status : int
- databaseStatus : int
- request_id : String
- requested_at : Date
- reviewed_at : Date
- granted_at : Date
- loadToDatabase : boolean
- expirationDate : Date
- motivation : String
- database:String : int
- schema : String

Other than adding members and such.

Tool: Git, DB, posix

DataMart
- path : String
- server : String

All entities have: createdAt, updatedAt and a change log. This is not on the diagram for the sake of simplicity.

Metadata
- ... : JSON

CUSP/ADRF Metadata Schema

DataProvider
- name : String

Resource
- type : String

Components

Component 1: Security

Component 2: Data Discovery

Component 3: Data Stewardship

Component 4: Collaboration

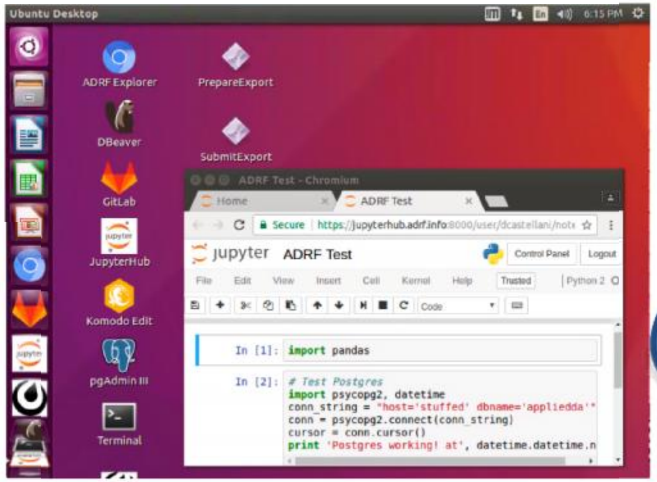
Component 5: Training

Implementation: Collaboration

PostgreSQL

python

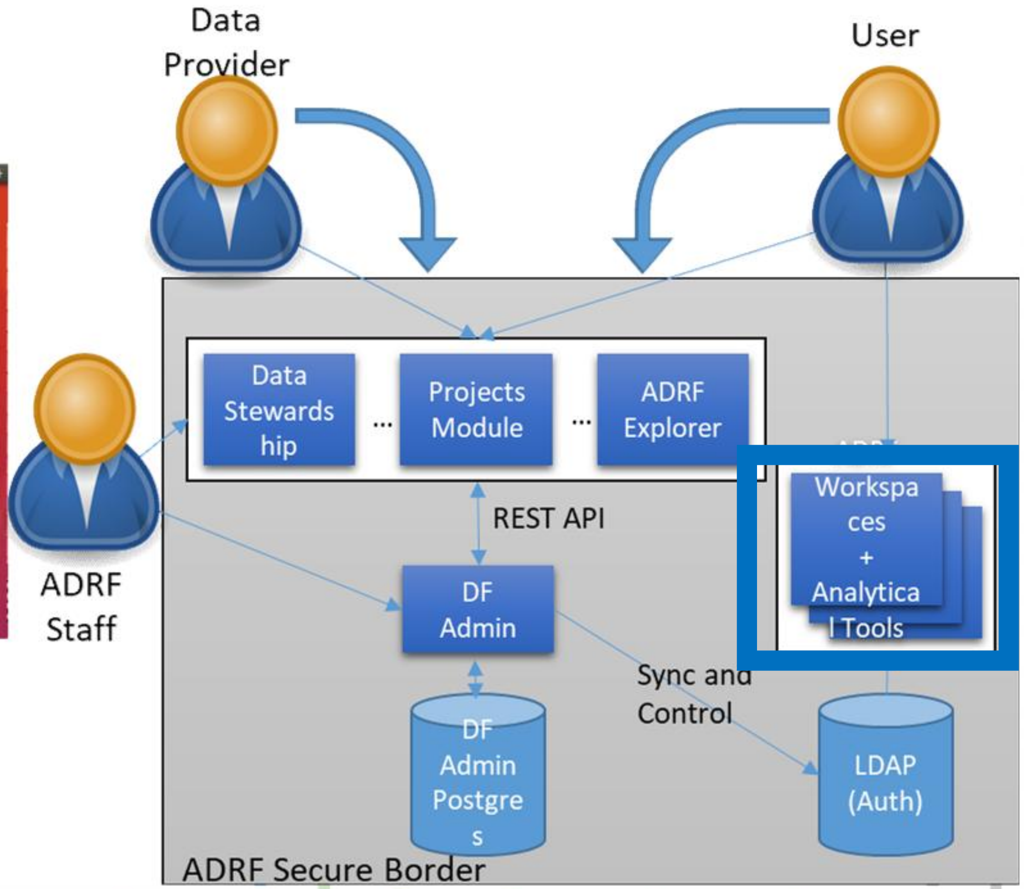
jupyter

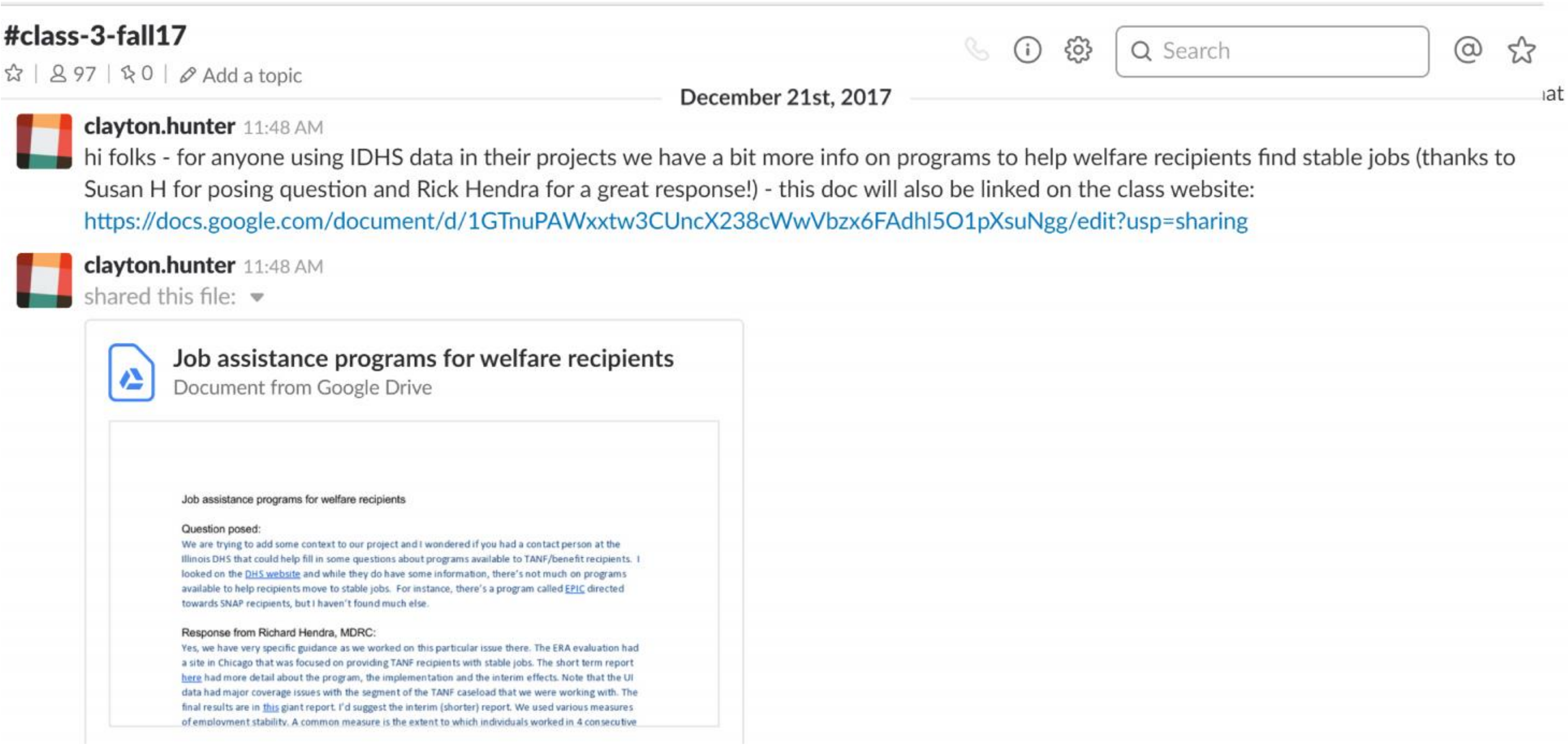


```
In [1]: import pandas

In [2]: # Test Postgres
        import psycopg2, datetime
        conn_string = "host='stuffed' dbname='appliedda'"
        conn = psycopg2.connect(conn_string)
        cursor = conn.cursor()
        print 'Postgres working! at', datetime.datetime.n
```

Https





December 21st, 2017

clayton.hunter 11:48 AM

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/1GTnuPAWxxtw3CUncX238cWwVbzx6FAdhl5O1pXsuNgg/edit?usp=sharing

clayton.hunter 11:48 AM

shared this file:

Job assistance programs for welfare recipients

Document from Google Drive

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 DHS that could help fill in some questions about programs available to TANF/benefit recipients. I looked on the DHS website and while they do have some 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, MDRC:

Yes, we have very specific guidance as we worked on this particular issue there. The ERA 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

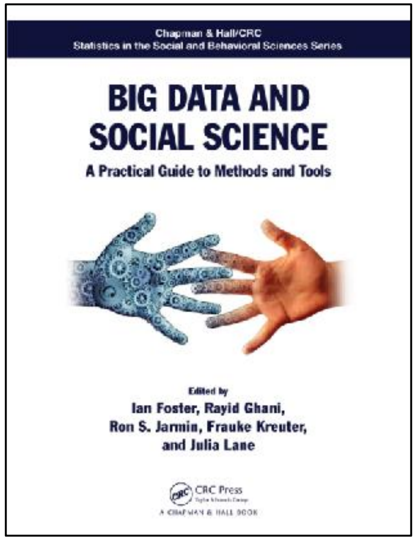
Component 2: Data Discovery

Component 3: Data Stewardship

Component 4: Collaboration

Component 5: Training

Textbook



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

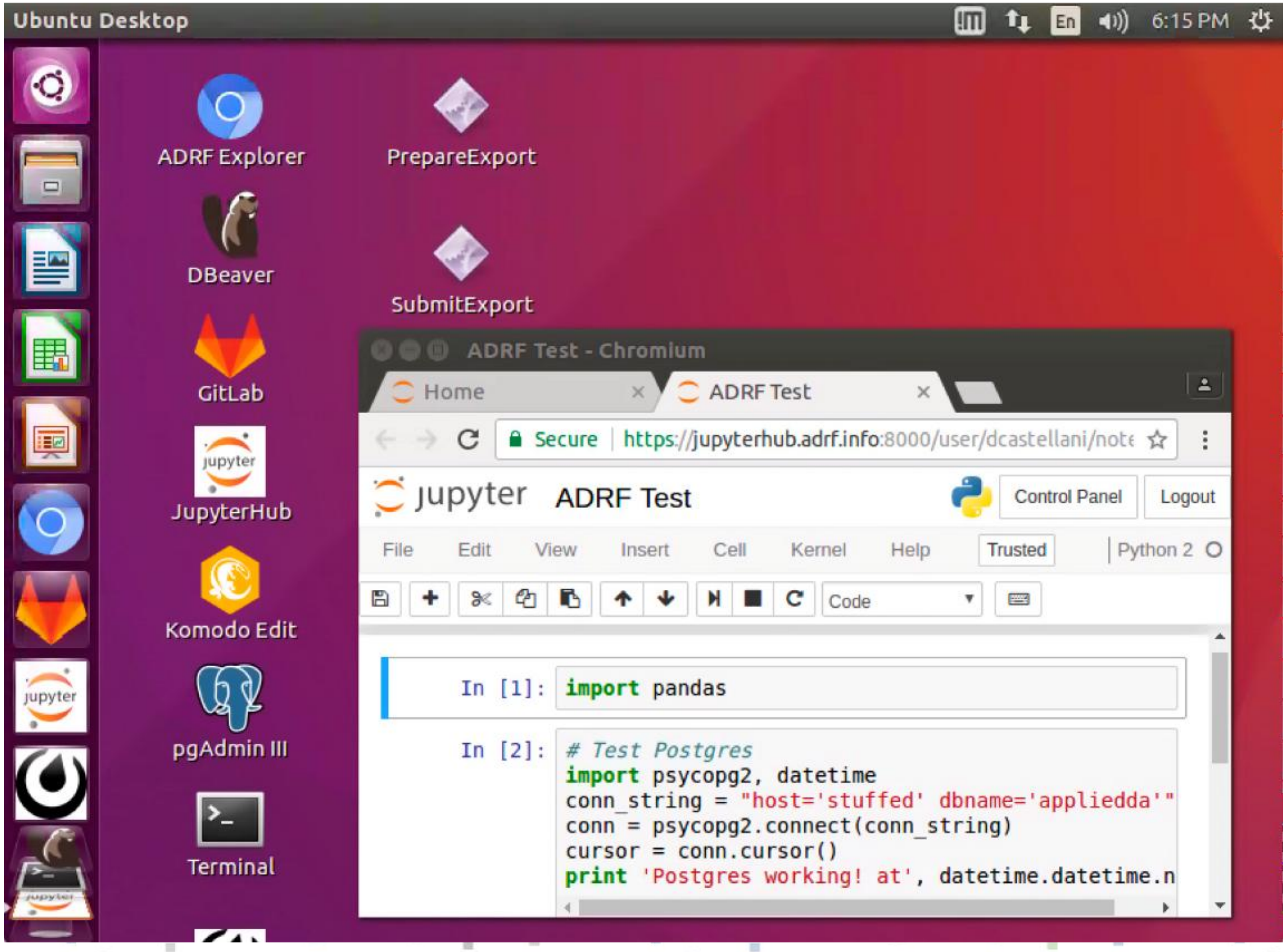
BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools

Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

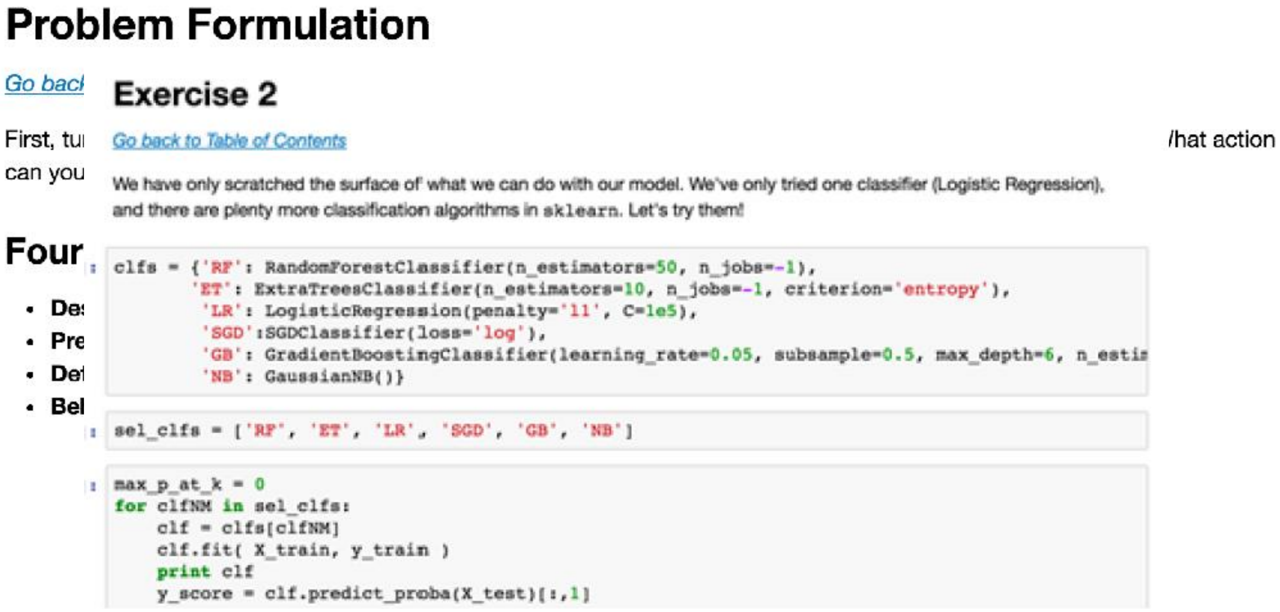
“If you work in social science and would like to explore the power of big data, this book is clearly for you…This book is complete and comprehensive. It covers all necessary steps to finish a big data project; collecting raw data, cleaning and preprocessing data, applying various modeling tools to analyze 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



```
In [1]: import pandas

In [2]: # Test Postgres
        import psycopg2, datetime
        conn_string = "host='stuffed' dbname='appliedda'"
        conn = psycopg2.connect(conn_string)
        cursor = conn.cursor()
        print 'Postgres working! at', datetime.datetime.n
```

Content Example: Machine Learning



Problem Formulation

Go bacl

First, tu /hat action

can you

Four

- De:
- Pre
- Def
- Bel

Exercise 2

Go back to Table of Contents

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!

```
clfs = {'RF': RandomForestClassifier(n_estimators=50, n_jobs=-1),
       'ET': ExtraTreesClassifier(n_estimators=10, n_jobs=-1, criterion='entropy'),
        'LR': LogisticRegression(penalty='l1', C=1e5),
        'SGD':SGDClassifier(loss='log'),
        'GB': GradientBoostingClassifier(learning_rate=0.05, subsample=0.5, max_depth=6, n_estim
        '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]
```

Products: Corrections and Employment

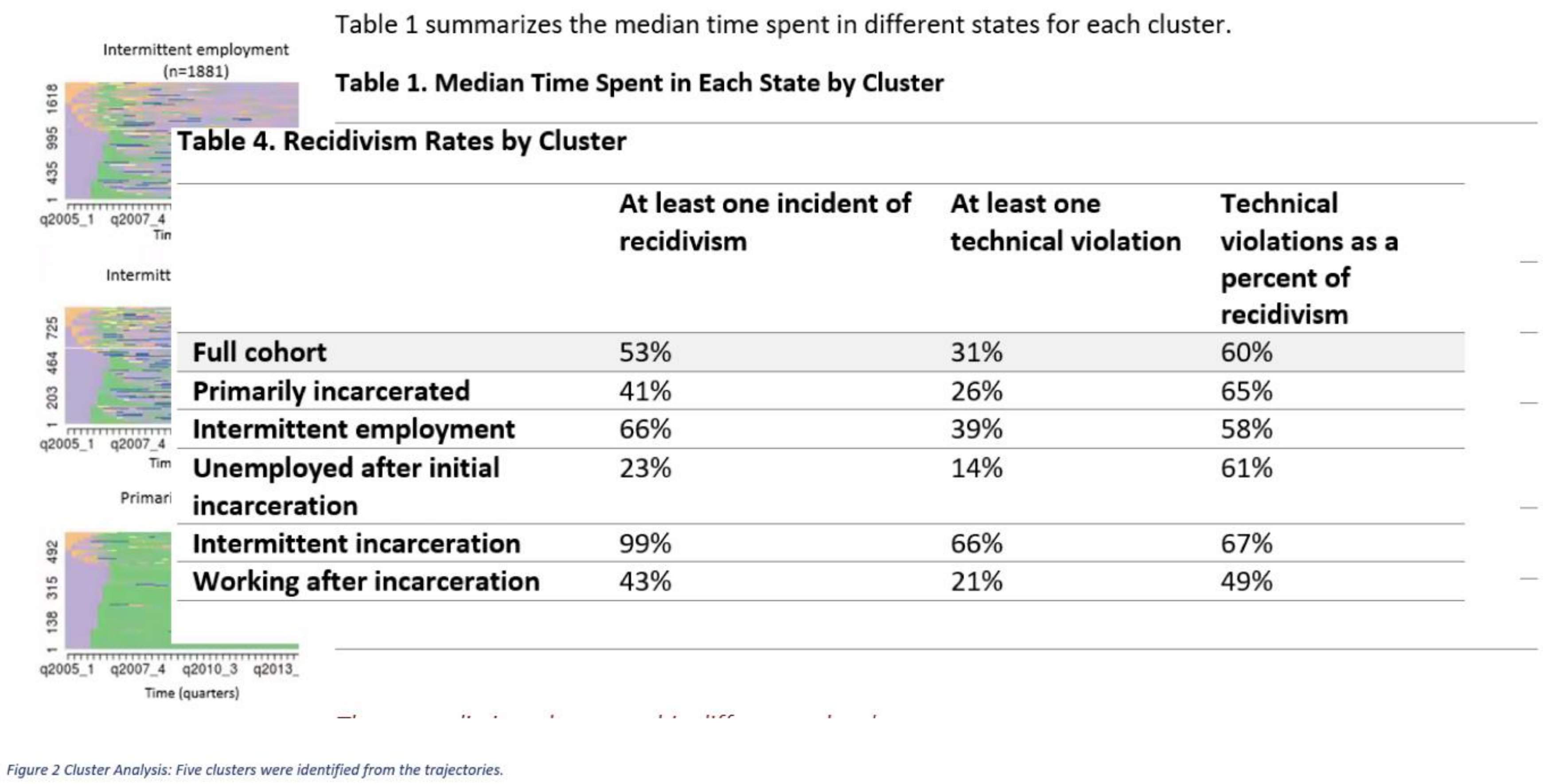


Table 1 summarizes the median time spent in different states for each cluster.

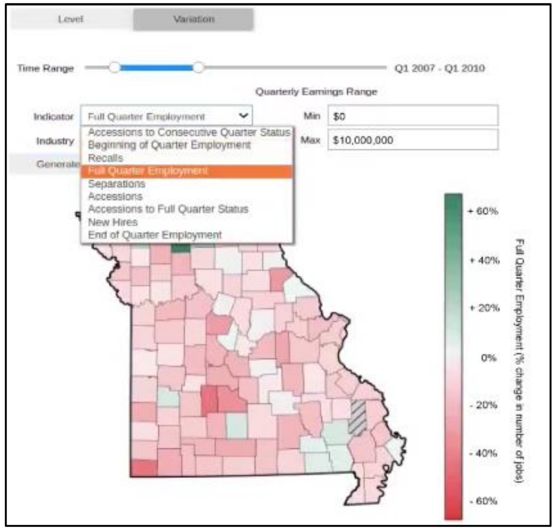
Table 1. Median Time Spent in Each State by Cluster

Table 4. Recidivism Rates by Cluster

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

Figure 2 Cluster Analysis: Five clusters were identified from the trajectories.

Tailored and Customizable Metrics



Time Range Q1 2007 - Q1 2010

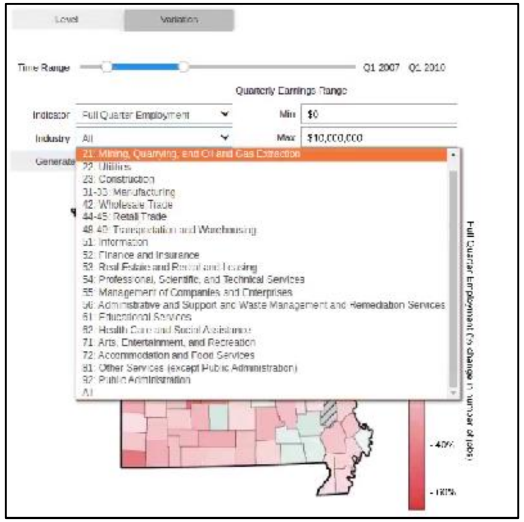
Quarterly Earnings Range

Indicator: Full Quarter Employment

Min: $0

Max: $10,000,000

Level
Variation
Industry
Generate
Accessions to Consecutive Quarter Status
Beginning of Quarter Employment
Recalls
Full Quarter Employment
Separations
Accessions
Accessions to Full Quarter Status
New Hires
End of Quarter Employment
+ 60%
+ 40%
+ 20%
0%
- 20%
- 40%
- 60%
Full Quarter Employment (% change in number of jobs)



Time Range Q1 2007 Q1 2010

Quarterly Earnings Range

Indicator: Full Quarter Employment

Industry: All

Min: $0

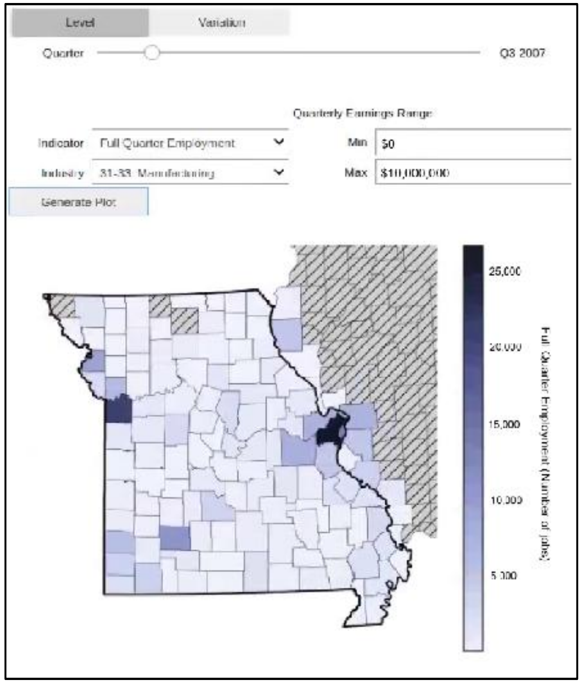
Max: $10,000,000

Level
Variation
Generate
21: Mining, Quarrying, and Oil and Gas Extraction
22: Utilities
23: Construction
31-33: Manufacturing
42: Wholesale Trade
44-45: Retail Trade
48-49: Transportation and Warehousing
51: Information
52: Finance and Insurance
53: Real Estate and Rental and Leasing
54: Professional, Scientific, and Technical Services
55: Management of Companies and Enterprises
56: Administrative and Support and Waste Management and Remediation Services
61: Educational Services
62: Health Care and Social Assistance
71: Arts, Entertainment, and Recreation
72: Accommodation and Food Services
81: Other Services (except Public Administration)
92: Public Administration
All
- 40%
- 60%
Full Quarter Employment (% change in number of jobs)

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



Quarter Q3 2007

Quarterly Earnings Range

Indicator: Full Quarter Employment

Industry: 31-33 Manufacturing

Min: $0

Max: $10,000,000

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

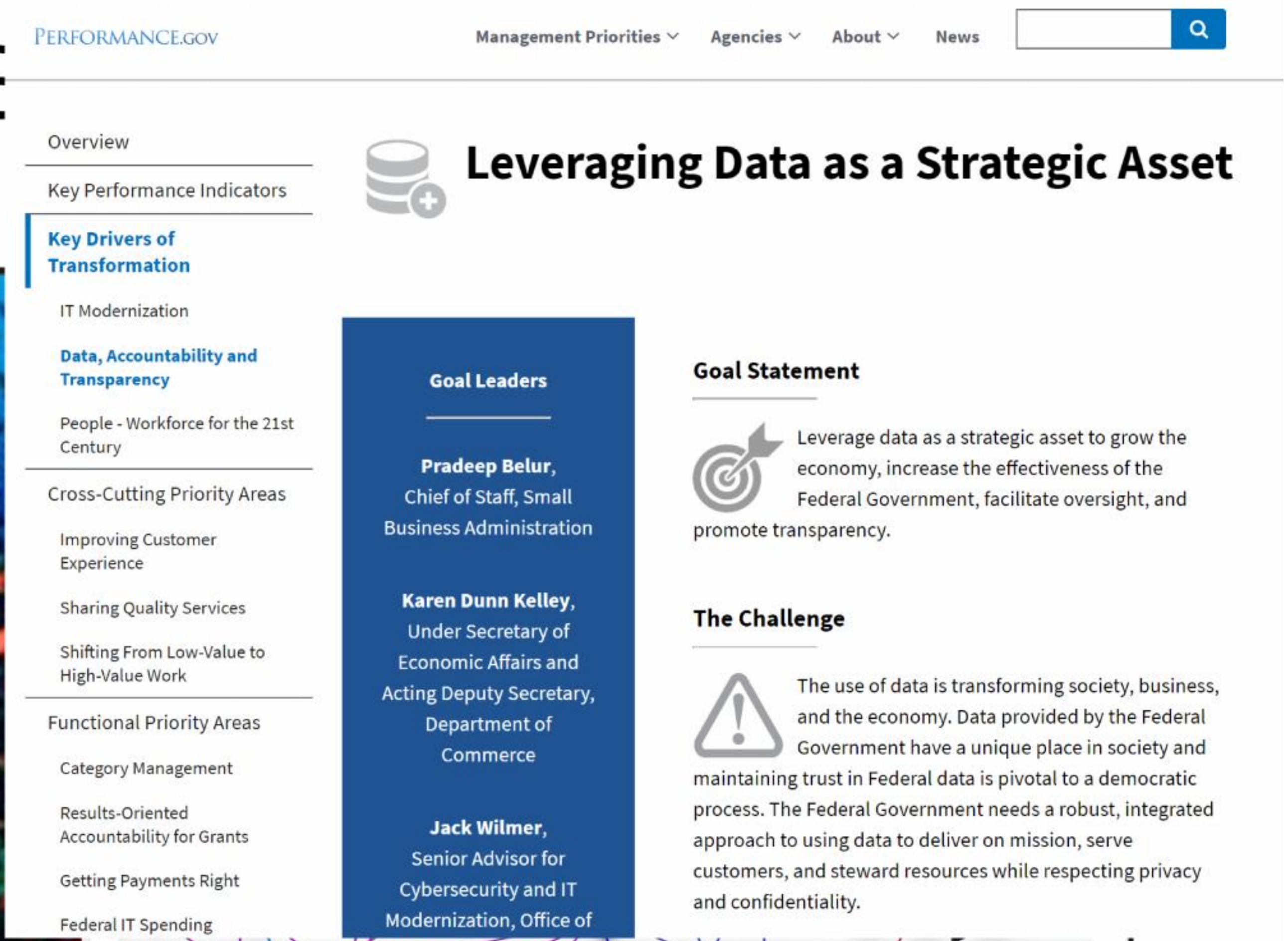
E

Identi

- Cri
- Soc
- Edu
- Wo

Build

- Ag
- Pro
- Citizen en



Overview

Key Performance Indicators

Key Drivers of Transformation

IT Modernization

Data, Accountability and Transparency

People - Workforce for the 21st Century

Cross-Cutting Priority Areas

Improving Customer Experience

Sharing Quality Services

Shifting From Low-Value to High-Value Work

Functional Priority Areas

Category Management

Results-Oriented Accountability for Grants

Getting Payments Right

Federal IT Spending

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