Administrative Data Research Facility and Metadata

Julia Lane

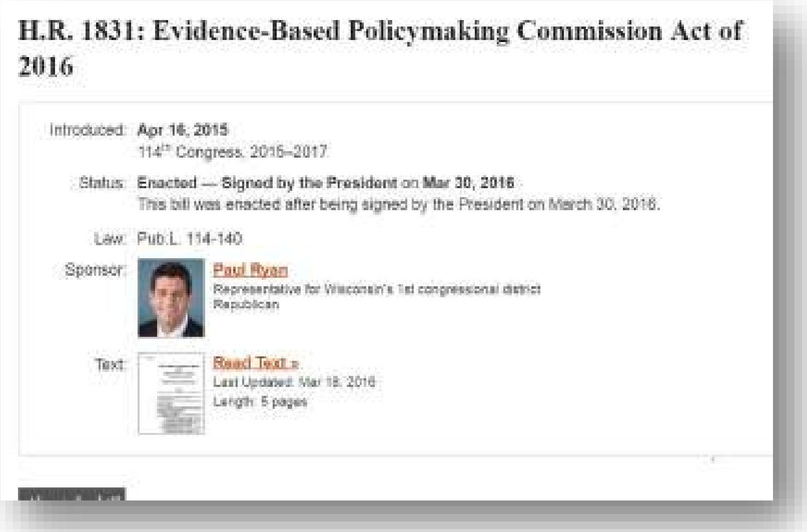
New York University

Key challenges to be solved with metadata – particularly for federal statistical system

- Limited internal capacity
- Security
- Legal mandates surrounding access and use
- Data sharing issues
  - cost
  - burden
  - data quality
  - data documentation
  - risk of bad analysis

jupyter

Context



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

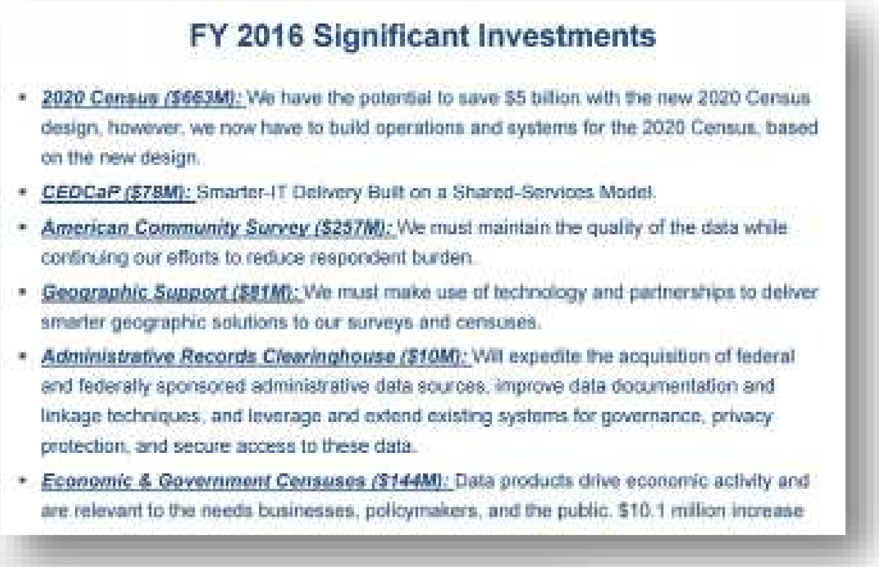
Introduced: **Apr 16, 2015**
114th Congress, 2015–2017

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

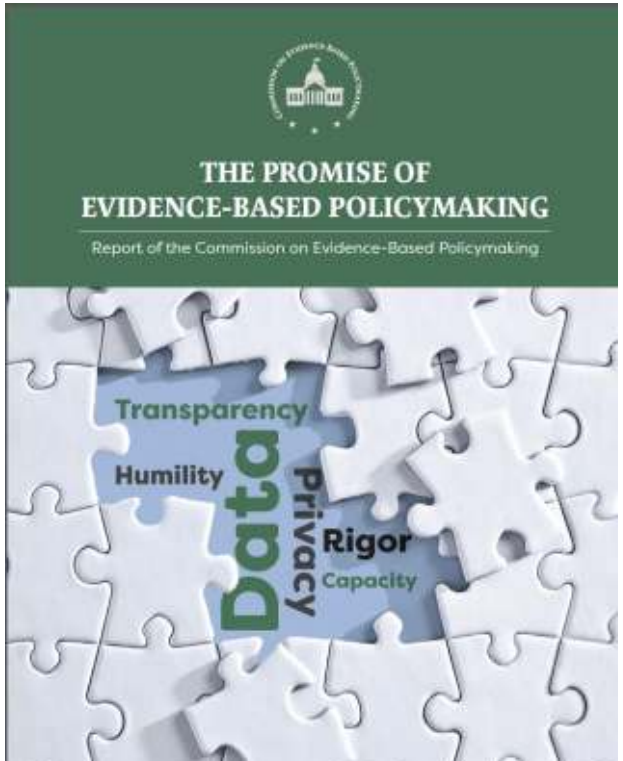
Sponsor: Paul Ryan
Representative for Wisconsin's 1st congressional district
Republican

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



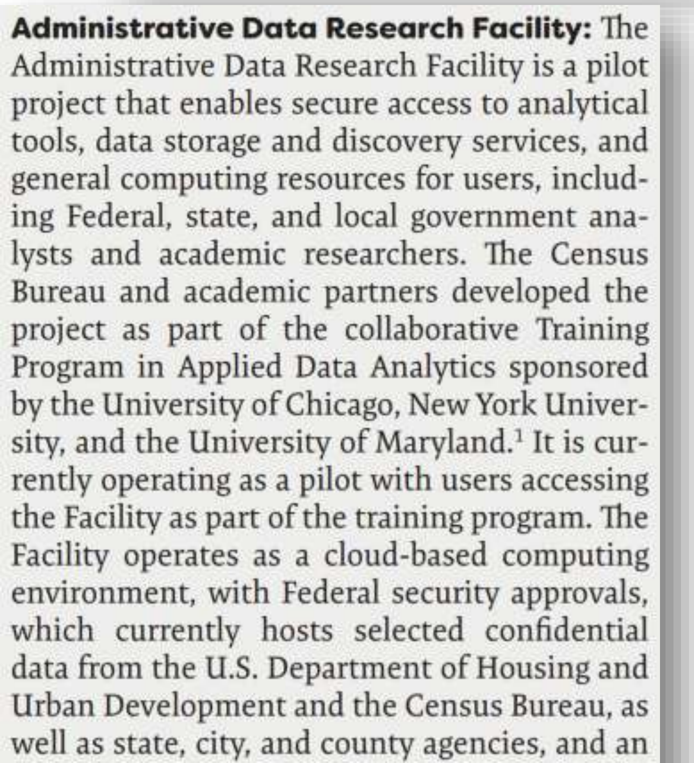
FY 2016 Significant Investments

- ***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 ($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

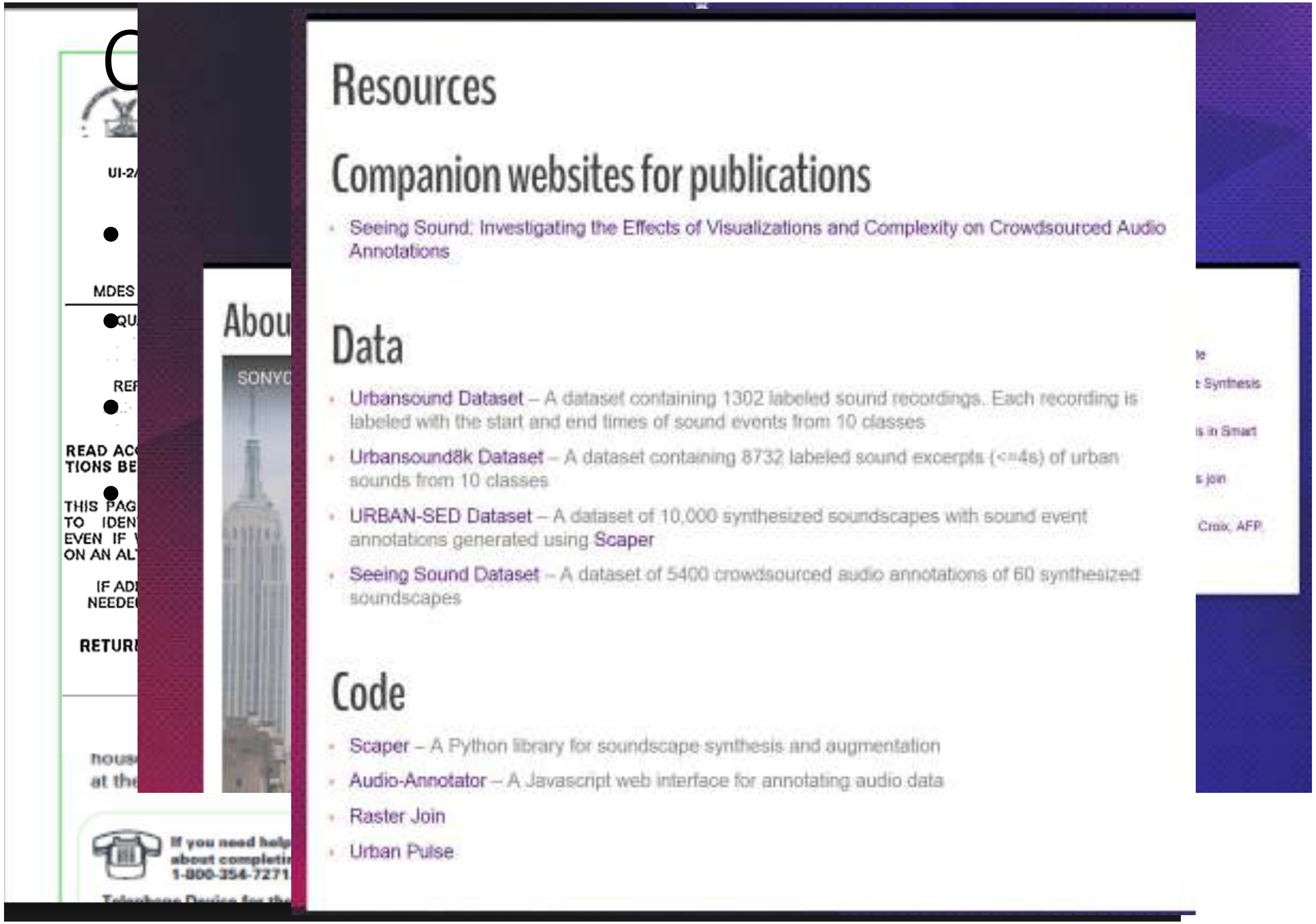


THE PROMISE OF
EVIDENCE-BASED POLICYMAKING

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



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

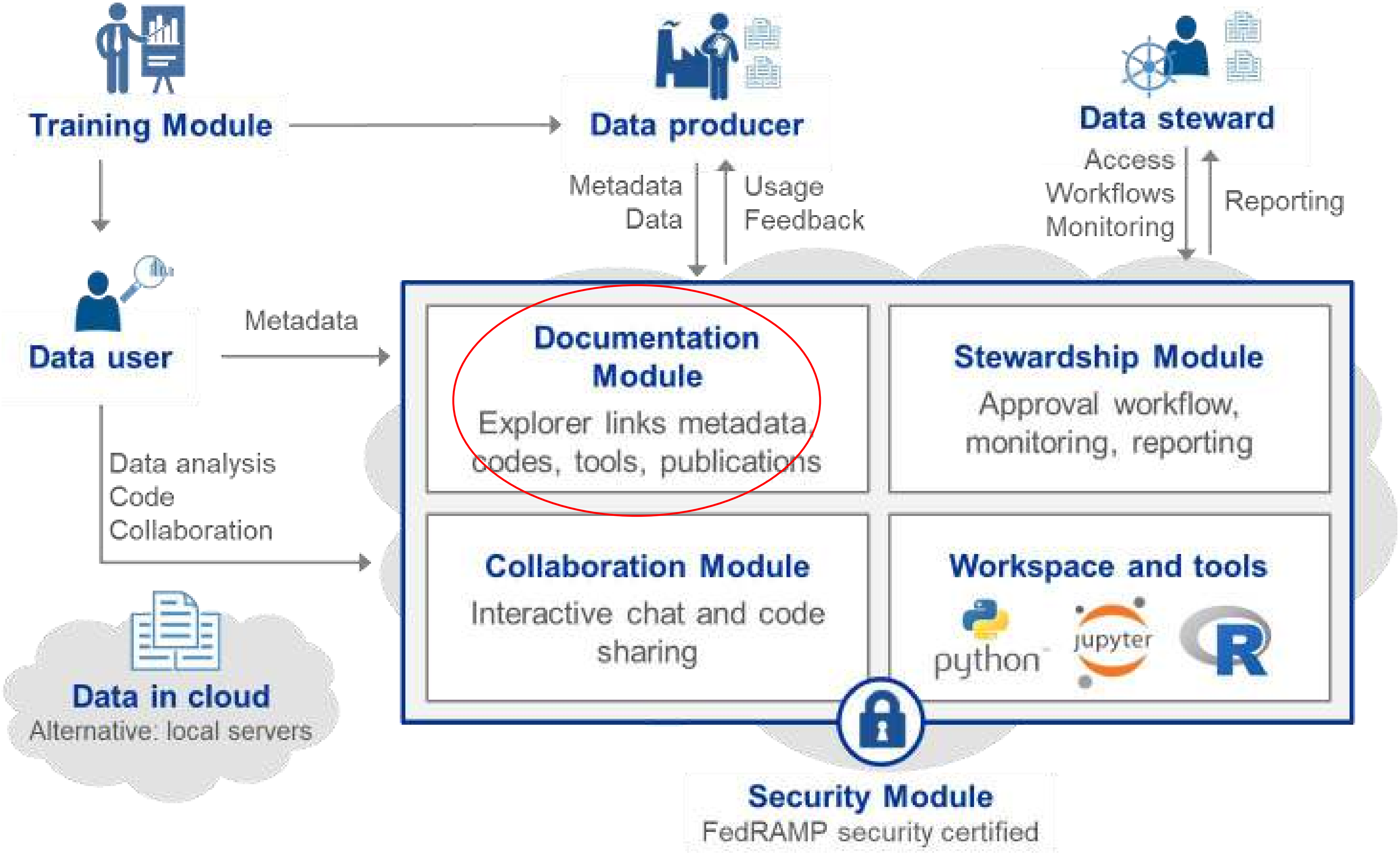
Build technical environment

Users:
- Federal, state and local data owners
- Analysts and researchers
- Federal, state and local program managers

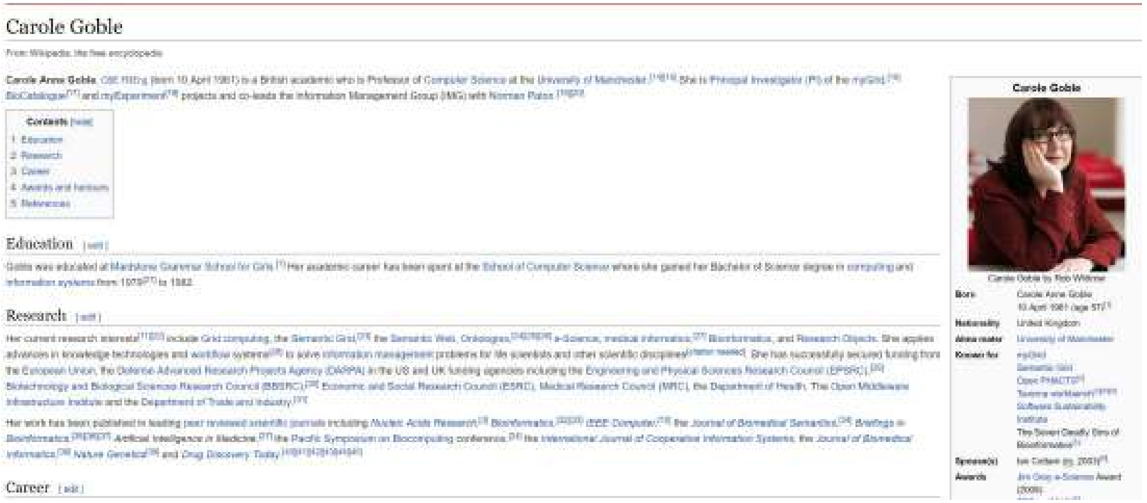
Technical Needs:
- Management and Secure Stewardship
- Access, Discovery and Collaboration
- Analysis and Dissemination

Secure
Reusable
Scalable
Extensible
Interoperable

Functional characteristics



Inspiration



Carole Goble

From Wikipedia, the free encyclopedia

Carole Anne Goble, CBE FREng (born 10 April 1961) is a British academic who is Professor of Computer Science at the University of Manchester.[1][6][3] She is Principal Investigator (PI) of the myGrid,[10] BioCatalogue[11] and myExperiment[12] projects and co-leads the Information Management Group (IMG) with Norman Paton.[13][14]

Contents [hide]
1 Education
2 Research
3 Career
4 Awards and honours
5 References

Education [edit]

Goble was educated at Maidstone Grammar School for Girls.[1] Her academic career has been spent at the School of Computer Science where she gained her Bachelor of Science degree in computing and information systems from 1979[1] to 1982.

Research [edit]

Her current research interests[11][22] include Grid computing, the Semantic Grid,[23] the Semantic Web, Ontologies,[24][25][26] e-Science, medical informatics,[27] Bioinformatics, and Research Objects. She applies advances in knowledge technologies and workflow systems[28] to solve information management problems for life scientists and other scientific disciplines.[citation needed] She has successfully secured funding from the European Union, the Defense Advanced Research Projects Agency (DARPA) in the US and UK funding agencies including the Engineering and Physical Sciences Research Council (EPSRC),[29] Biotechnology and Biological Sciences Research Council (BBSRC),[30] Economic and Social Research Council (ESRC), Medical Research Council (MRC), the Department of Health, The Open Middleware Infrastructure Institute and the Department of Trade and Industry.[31]

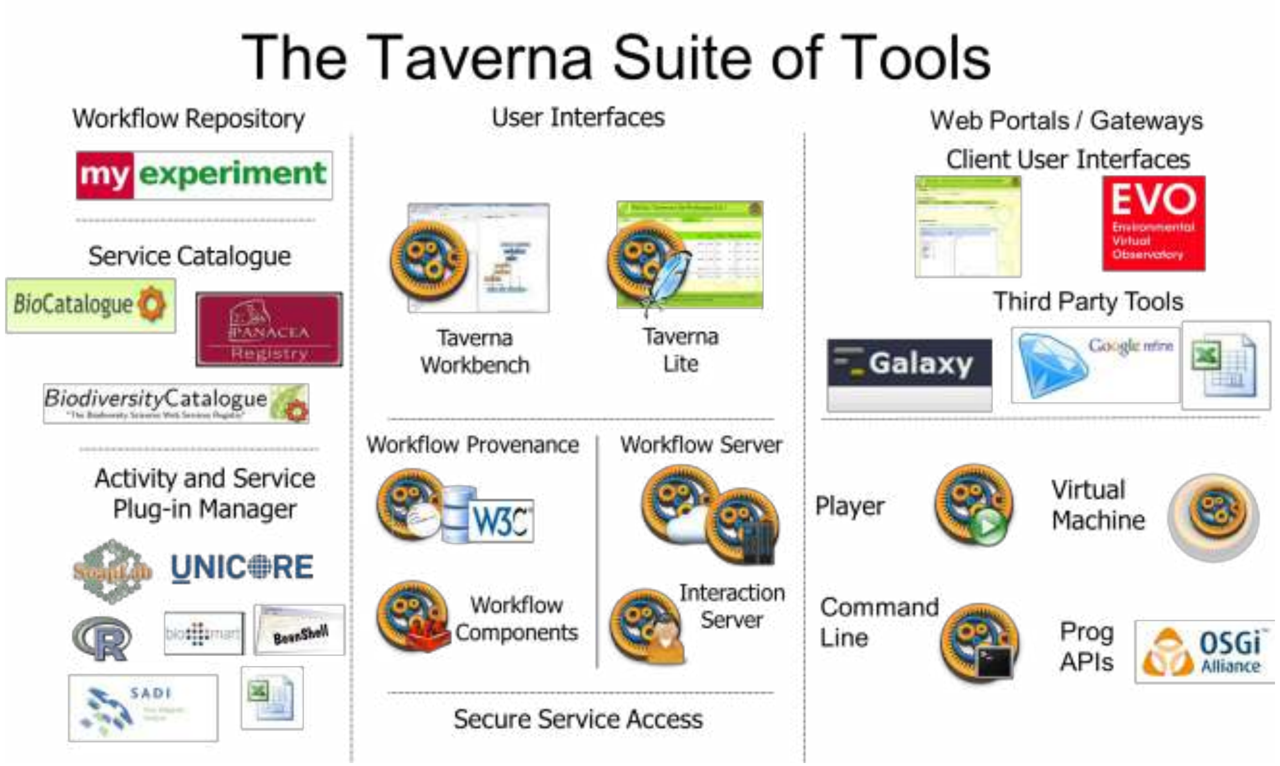
Her work has been published in leading peer reviewed scientific journals including *Nucleic Acids Research*,[3] *Bioinformatics*,[32][33] *IEEE Computer*,[15] the *Journal of Biomedical Semantics*,[34] *Briefings in Bioinformatics*,[35][36][37] *Artificial Intelligence in Medicine*,[27] the Pacific Symposium on Biocomputing conference,[24] the *International Journal of Cooperative Information Systems*, the *Journal of Biomedical Informatics*,[38] *Nature Genetics*[39] and *Drug Discovery Today*.[40][41][42][43][44][45]

Career [edit]

Carole Goble

Carole Goble by Rob Widmer

Born	Carole Anne Goble 10 April 1961 (age 57)[1]
Nationality	United Kingdom
Alma mater	University of Manchester
Known for	myGrid Semantic Grid Open PHACTS[1] Taverna workbench[1][2] Software Sustainability Institute The Seven Deadly Sins of Bioinformatics[1]
Spouse(s)	Ian Cottam (m. 2003)[1]
Awards	Jim Gray e-Science Award (2008)



RESEARCH

- Github

Isidor Ni

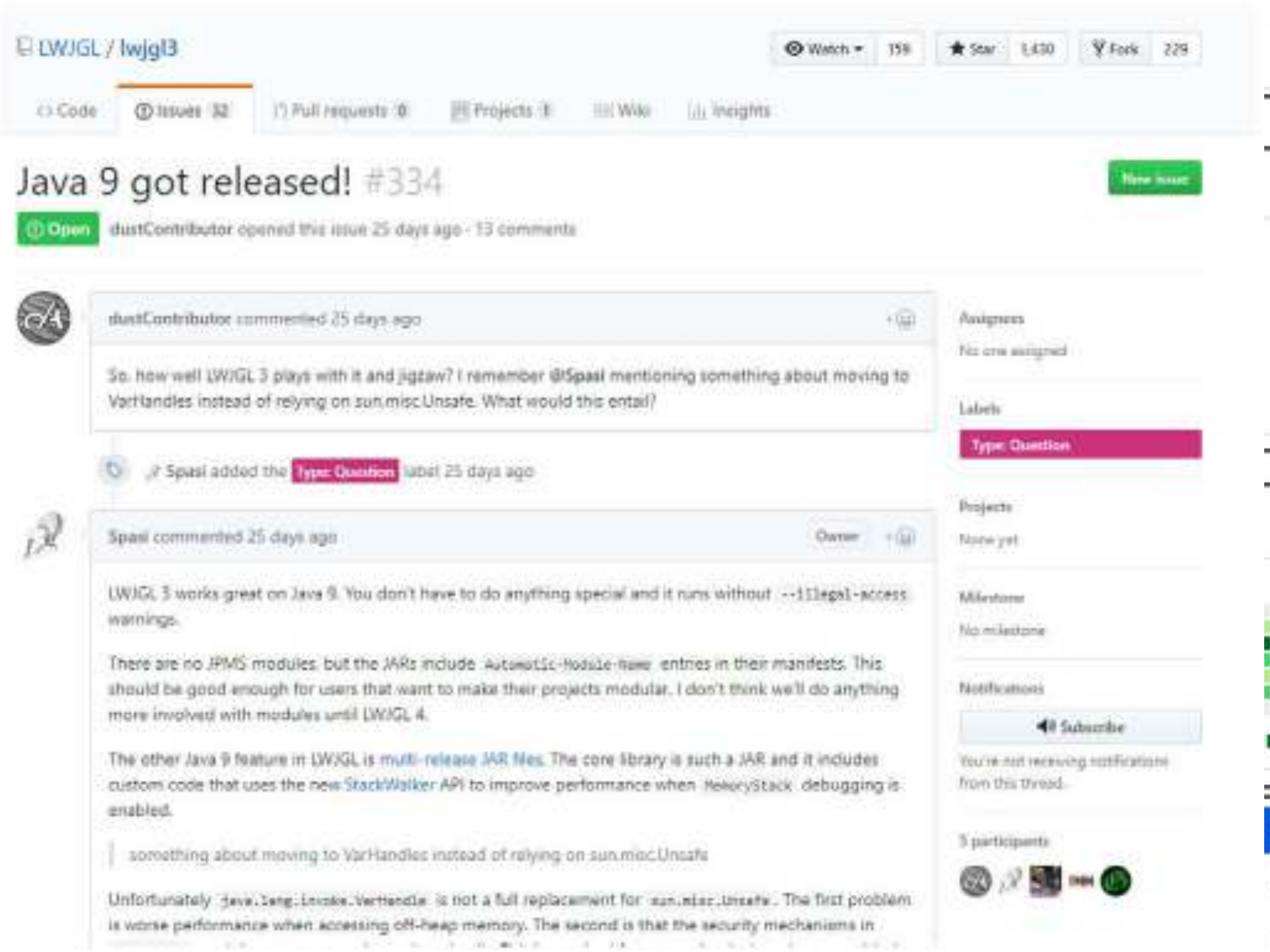
isidorn

VSCode

Block or report u

Microsoft

Zurich



LWJGL / lwjgl3

Java 9 got released! #334

Open dustContributor opened this issue 25 days ago · 13 comments

dustContributor commented 25 days ago

So, how well LWJGL 3 plays with it and jigzaw? I remember @Spasi mentioning something about moving to VarHandles instead of relying on sun.misc.Unsafe. What would this entail?

Spasi added the Type: Question label 25 days ago

Spasi commented 25 days ago

LWJGL 3 works great on Java 9. You don't have to do anything special and it runs without `--illegal-access` warnings.

There are no JPMS modules, but the JARs include `Automatic-Module-Name` entries in their manifests. This should be good enough for users that want to make their projects modular. I don't think we'll do anything more involved with modules until LWJGL 4.

The other Java 9 feature in LWJGL is multi-release JAR files. The core library is such a JAR and it includes custom code that uses the new StackWalker API to improve performance when `MemoryStack` debugging is enabled.

> something about moving to VarHandles instead of relying on sun.misc.Unsafe

Unfortunately `java.lang.invoke.VarHandle` is not a full replacement for `sun.misc.Unsafe`. The first problem is worse performance when accessing off-heap memory. The second is that the security mechanisms in

Assignees

No one assigned

Labels

Type: Question

Projects

None yet

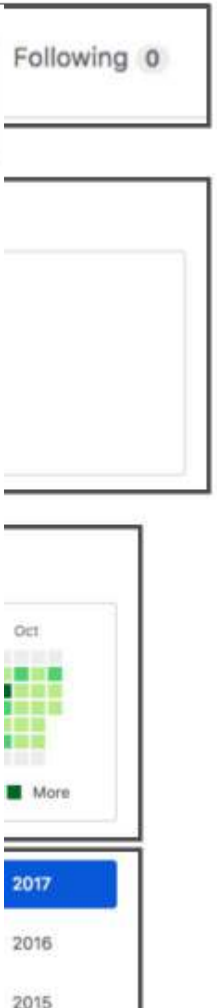
Milestone

No milestone

Notifications

You're not receiving notifications from this thread.

5 participants



Created 56 commits in 3 repositories

Microsoft/vscode 53 commits

Microsoft/vscode-node-debug 2 commits

Microsoft/vscode-generator-code 1 commit

Microsoft/vscode 53 commits

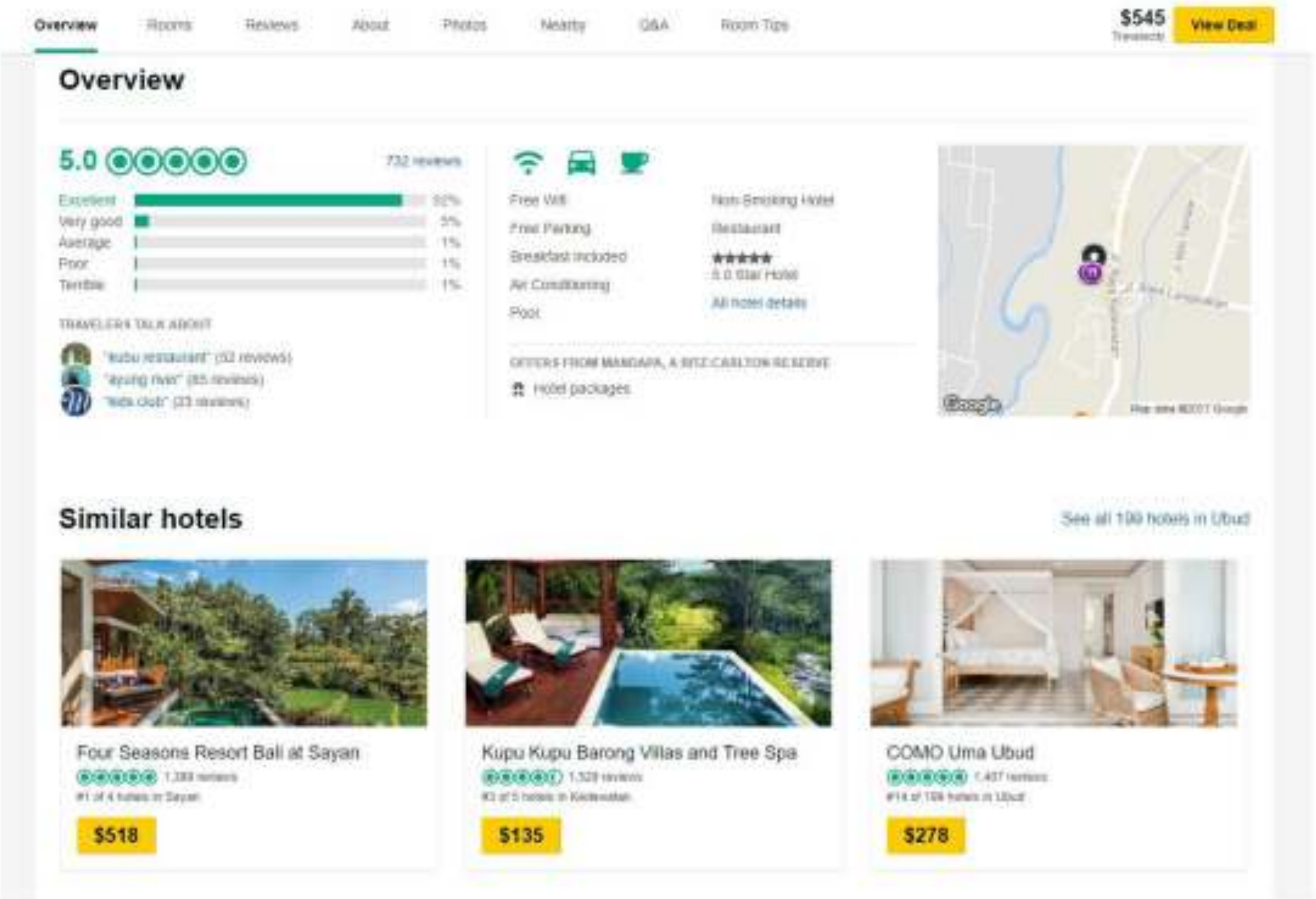
Microsoft/vscode-node-debug 2 commits

RESEARCH

- Github
- Data.world
- Pinterest
- TripAdvisor

ttja
Lincoln
Unit
Kingd
11

Linds
6



Overview

5.0 ●●●●● 732 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
●●●●● 1,283 reviews
#1 of 4 hotels in Sayan
$518

Kupu Kupu Barong Villas and Tree Spa
●●●●◐ 1,328 reviews
#3 of 5 hotels in Kedewatan
$135

COMO Uma Ubud
●●●●● 1,487 reviews
#14 of 199 hotels in Ubud
$278

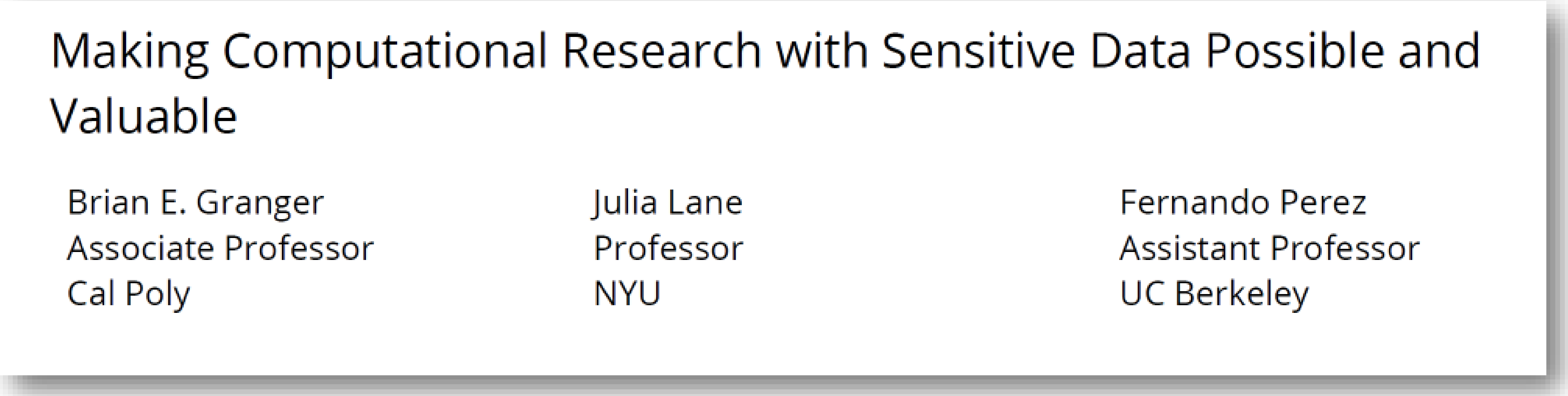
View Collection

eadership
300 Readers

ior Reviewer
5 Reviews

Terrible 0

Points
0
1
300 points to go



Making Computational Research with Sensitive Data Possible and Valuable

Brian E. Granger
Associate Professor
Cal Poly

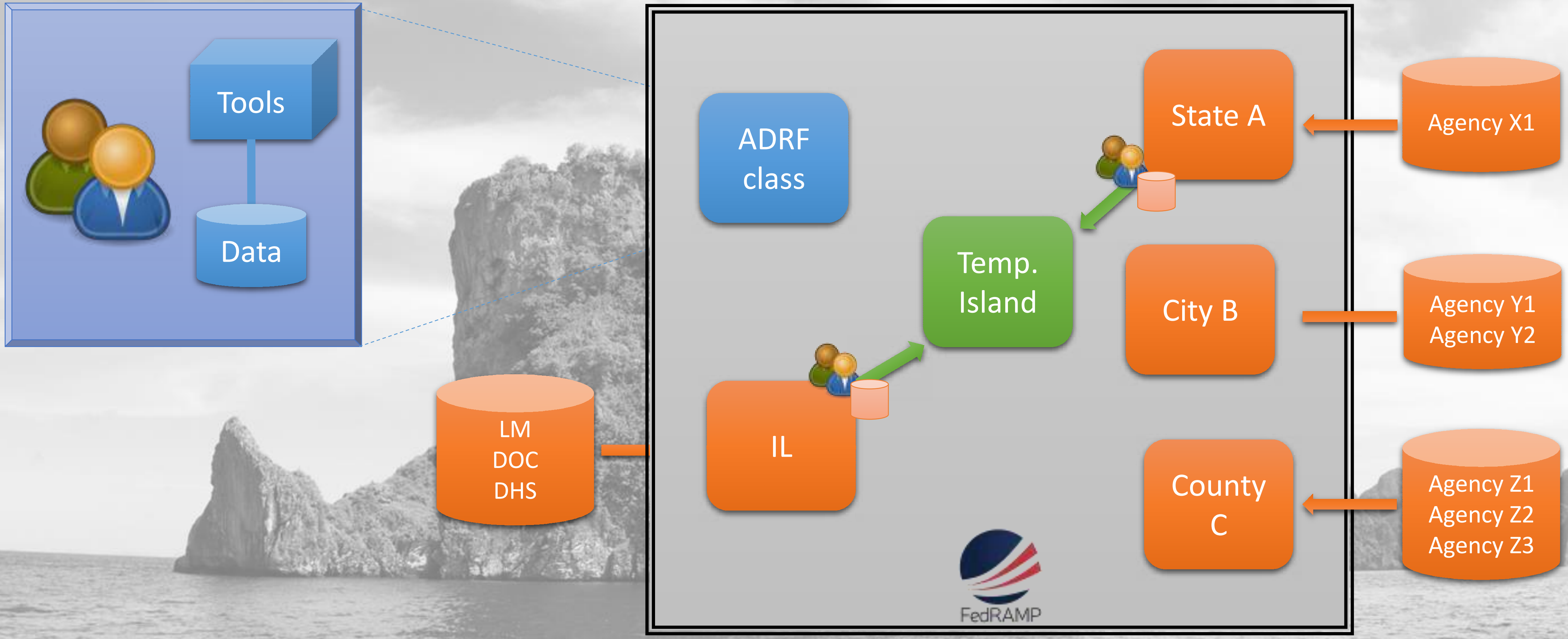
Julia Lane
Professor
NYU

Fernando Perez
Assistant Professor
UC Berkeley

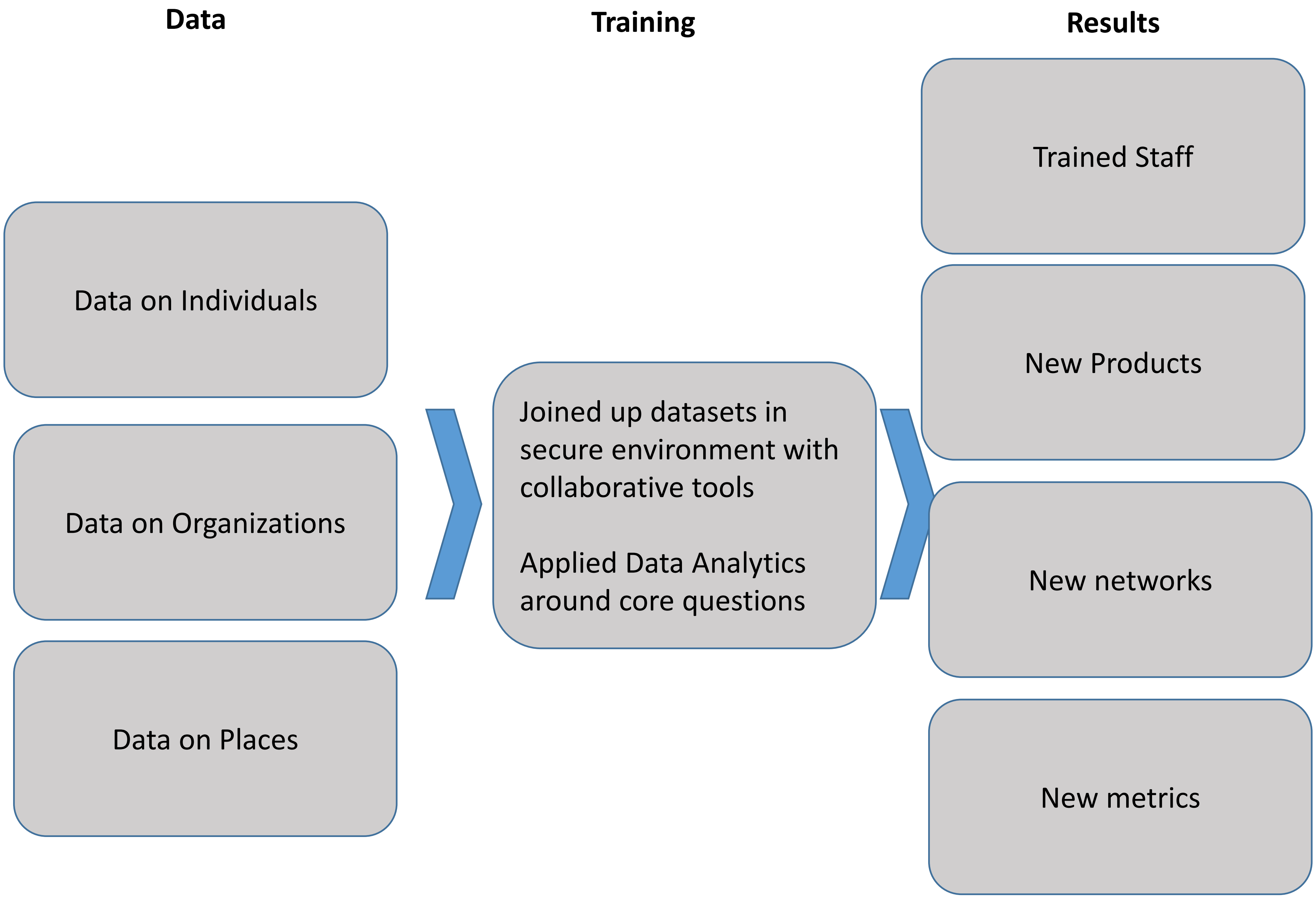
Alfred P. Sloan Foundation

SCHMIDT FUTURES

Overdeck Family Foundation


Tools
Data
ADRF
class
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



Data

Training

Results

Data on Individuals

Data on Organizations

Data on Places

Joined up datasets in secure environment with collaborative tools

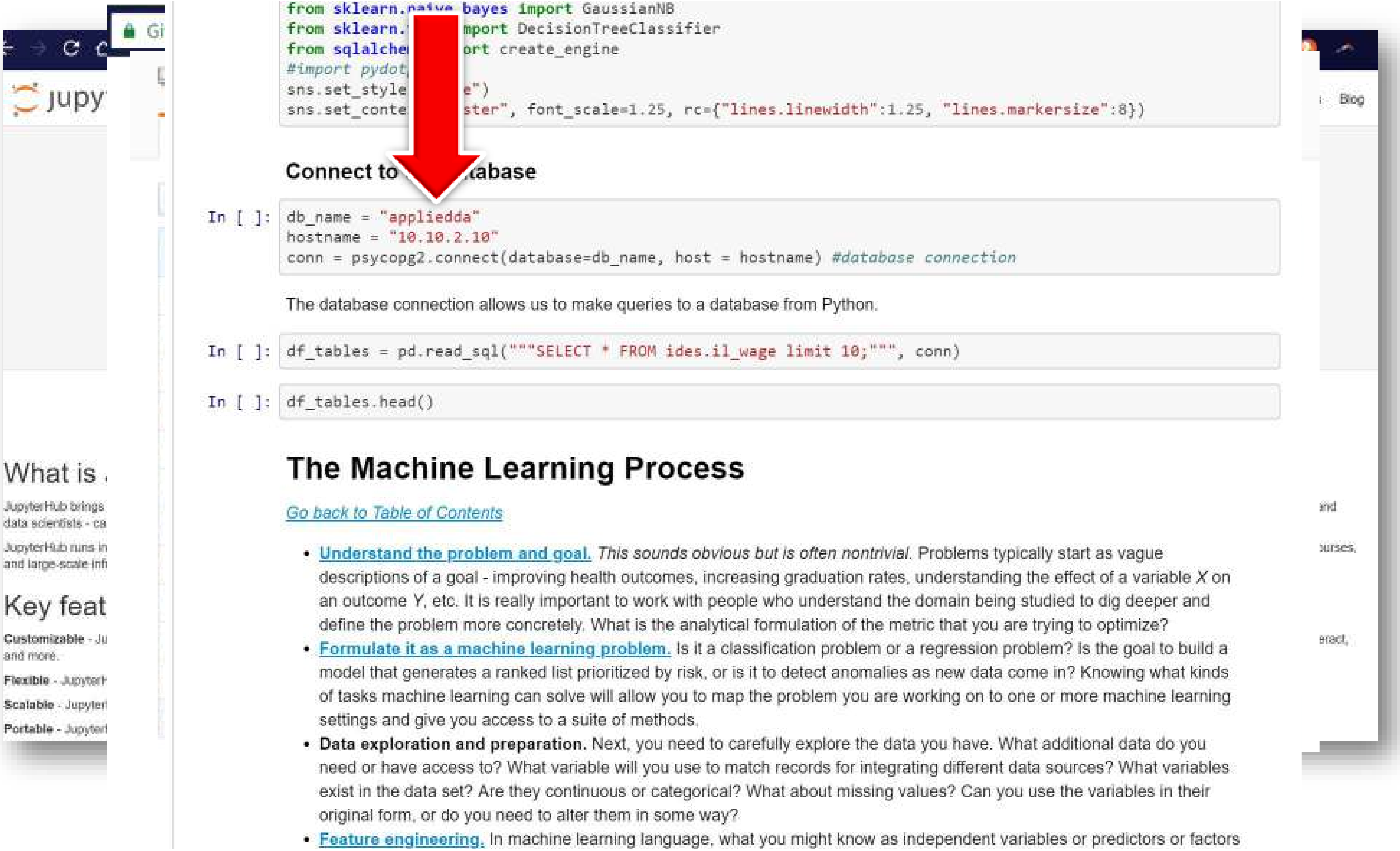
Applied Data Analytics around core questions

Trained Staff

New Products

New networks

New metrics



```
from sklearn.naive bayes import GaussianNB
from sklearn.          mport DecisionTreeClassifier
from sqlalche          ort create_engine
#import pydot
sns.set_style          e")
sns.set_conte          ster", font_scale=1.25, rc={"lines.linewidth":1.25, "lines.markersize":8})
```

Connect to   tabase

```
In [ ]: db_name = "appliedda"
        hostname = "10.10.2.10"
        conn = psycopg2.connect(database=db_name, host = hostname) #database connection
```

The database connection allows us to make queries to a database from Python.

```
In [ ]: df_tables = pd.read_sql("""SELECT * FROM ides.il_wage limit 10;""", conn)
```

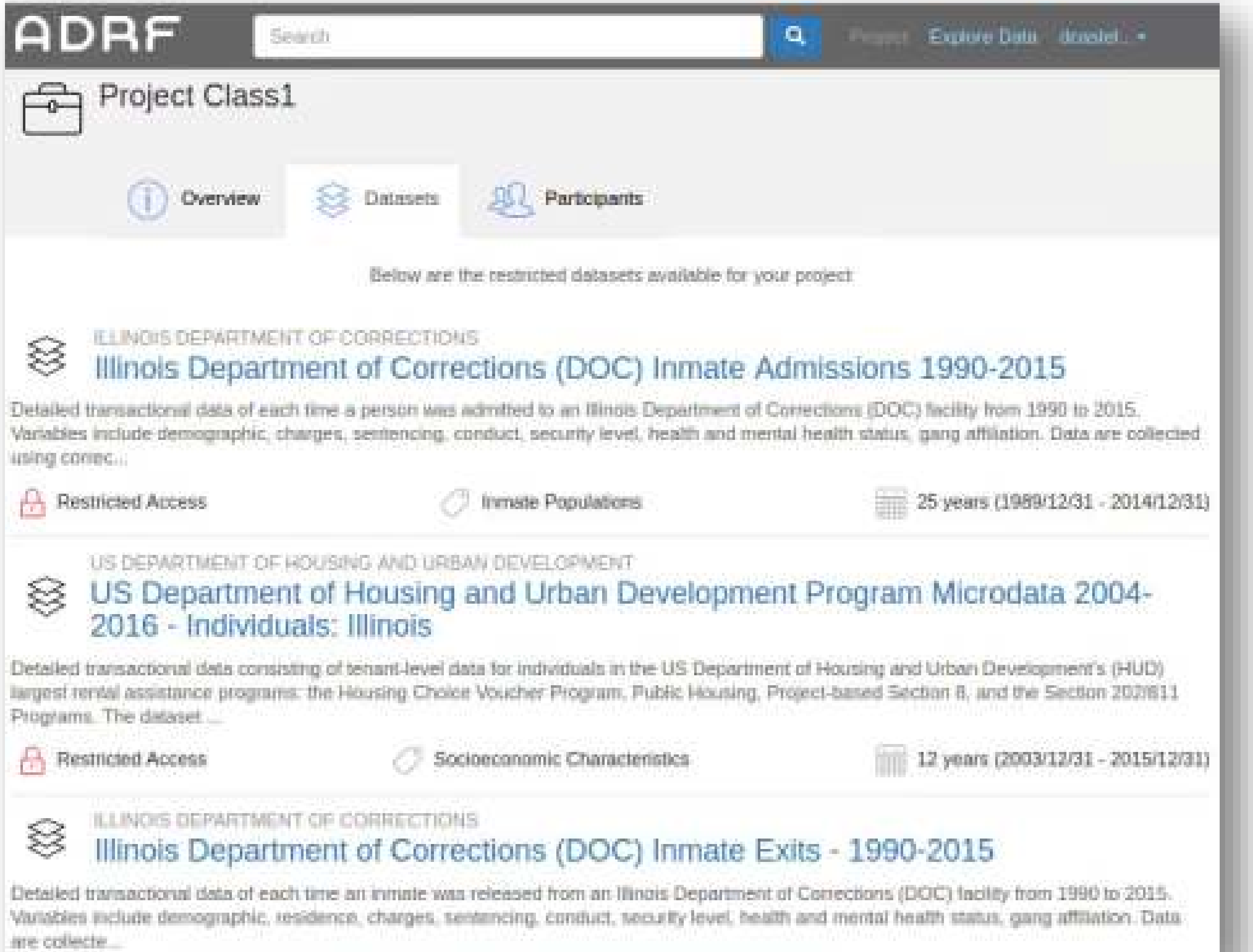
```
In [ ]: df_tables.head()
```

The Machine Learning Process

Go back to Table of Contents

- **Understand the problem and goal.** *This sounds obvious but is often nontrivial.* Problems typically start as vague descriptions of a goal - improving health outcomes, increasing graduation rates, understanding the effect of a variable *X* on an outcome *Y*, etc. It is really important to work with people who understand the domain being studied to dig deeper and define the problem more concretely. What is the analytical formulation of the metric that you are trying to optimize?
- **Formulate it as a machine learning problem.** Is it a classification problem or a regression problem? Is the goal to build a model that generates a ranked list prioritized by risk, or is it to detect anomalies as new data come in? Knowing what kinds of tasks machine learning can solve will allow you to map the problem you are working on to one or more machine learning settings and give you access to a suite of methods.
- **Data exploration and preparation.** Next, you need to carefully explore the data you have. What additional data do you need or have access to? What variable will you use to match records for integrating different data sources? What variables exist in the data set? Are they continuous or categorical? What about missing values? Can you use the variables in their original form, or do you need to alter them in some way?
- **Feature engineering.** In machine learning language, what you might know as independent variables or predictors or factors

Search and Discovery



Project Class1

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 ...

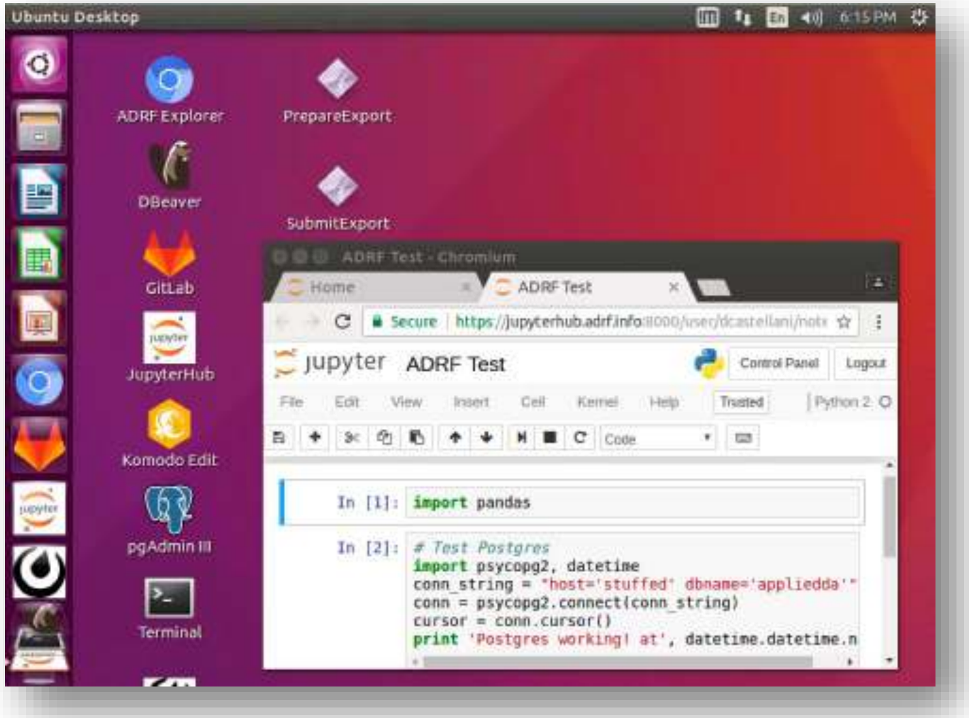
Restricted Access | Socioeconomic Characteristics | 12 years (2003/12/31 - 2015/12/31)

ILLINOIS DEPARTMENT OF CORRECTIONS

Illinois Department of Corrections (DOC) Inmate Exits - 1990-2015

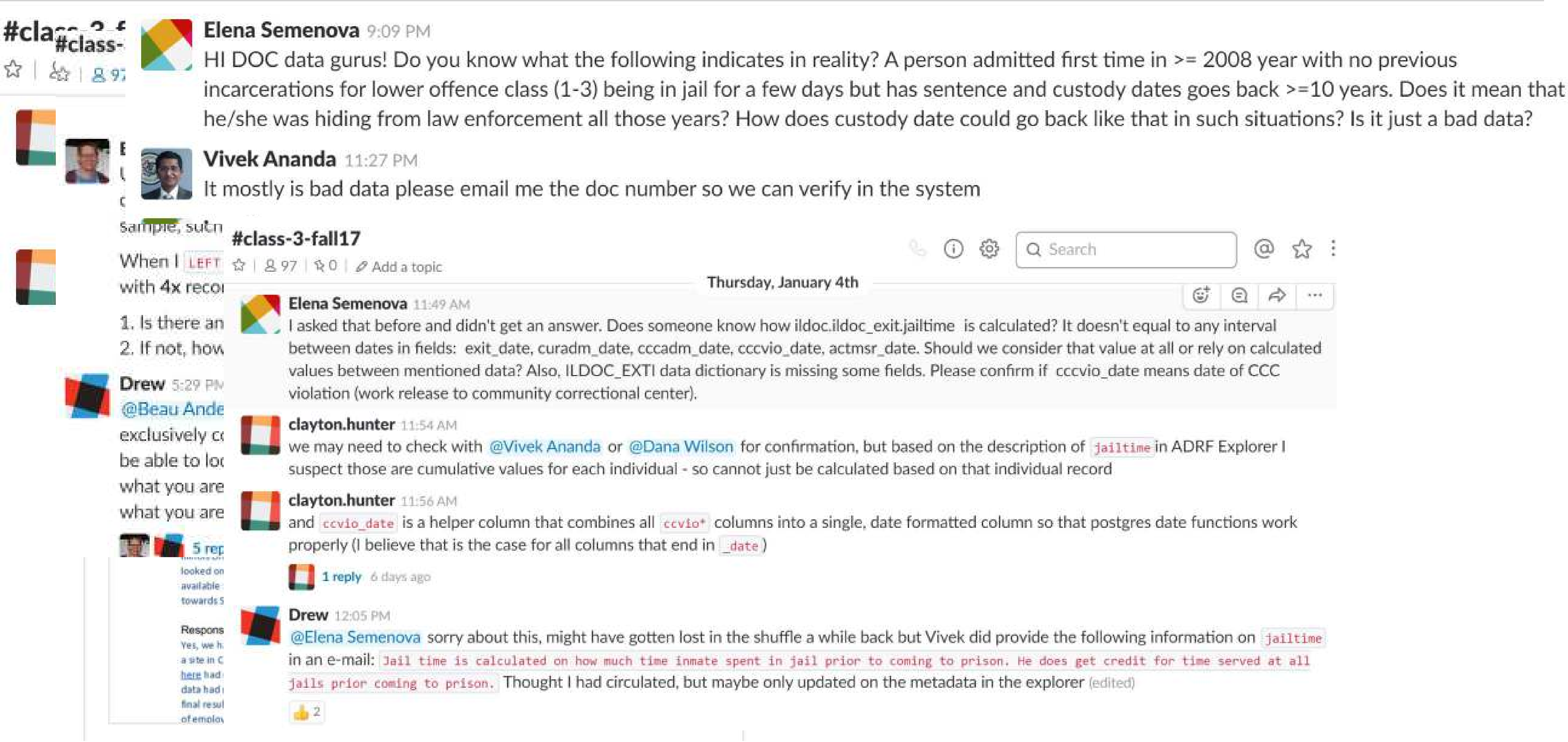
Detailed transactional data of each time an inmate was released from an Illinois Department of Corrections (DOC) facility from 1990 to 2015. Variables include demographic, residence, charges, sentencing, conduct, security level, health and mental health status, gang affiliation. Data are collecte...

Collaboration



```
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
```



#class-3-f #class-

Elena Semenova 9:09 PM

HI DOC data gurus! Do you know what the following indicates in reality? A person admitted first time in >= 2008 year with no previous incarcerations for lower offence class (1-3) being in jail for a few days but has sentence and custody dates goes back >=10 years. Does it mean that he/she was hiding from law enforcement all those years? How does custody date could go back like that in such situations? Is it just a bad data?

Vivek Ananda 11:27 PM

It mostly is bad data please email me the doc number so we can verify in the system

sample, such

When I LEFT

with 4x recor

1. Is there an

2. If not, how

Drew 5:29 PM

@Beau Ande

exclusively c

be able to lo

what you are

what you are

5 rep

looked on

available

towards 5

Respons

Yes, we h

a site in C

here had

data had

final resul

of employ

#class-3-fall17

☆ | 97 | 0 | Add a topic

Thursday, January 4th

Elena Semenova 11:49 AM

I asked that before and didn't get an answer. Does someone know how ildoc.ildoc_exit.jailtime is calculated? It doesn't equal to any interval between dates in fields: exit_date, curadm_date, cccadm_date, cccvio_date, actmsr_date. Should we consider that value at all or rely on calculated values between mentioned data? Also, ILDOC_EXTI data dictionary is missing some fields. Please confirm if cccvio_date means date of CCC violation (work release to community correctional center).

clayton.hunter 11:54 AM

we may need to check with @Vivek Ananda or @Dana Wilson for confirmation, but based on the description of `jailtime`in ADRF Explorer I suspect those are cumulative values for each individual - so cannot just be calculated based on that individual record

clayton.hunter 11:56 AM

and `ccvio_date` is a helper column that combines all `ccvio*` columns into a single, date formatted column so that postgres date functions work properly (I believe that is the case for all columns that end in `_date`)

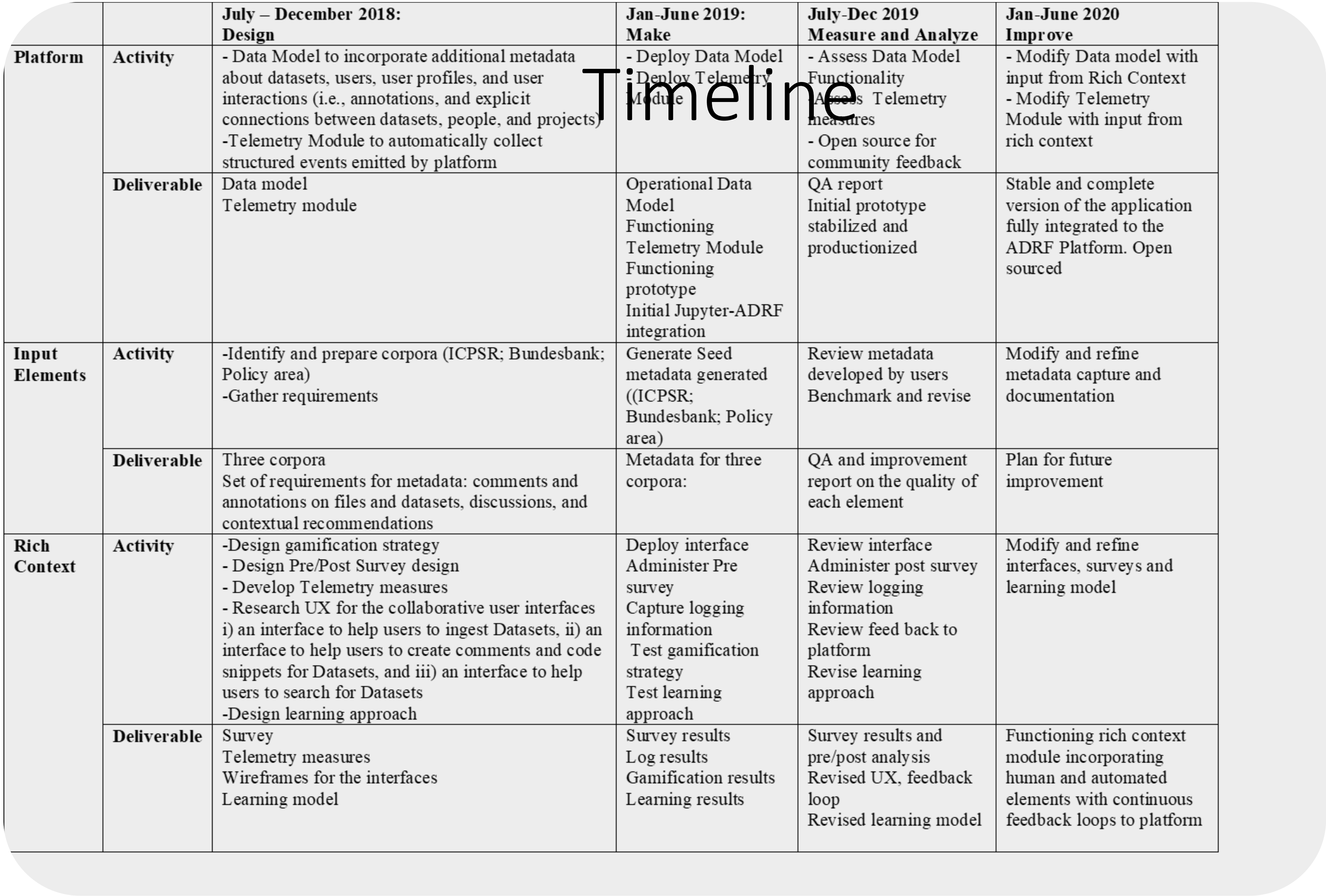
1 reply 6 days ago

Drew 12:05 PM

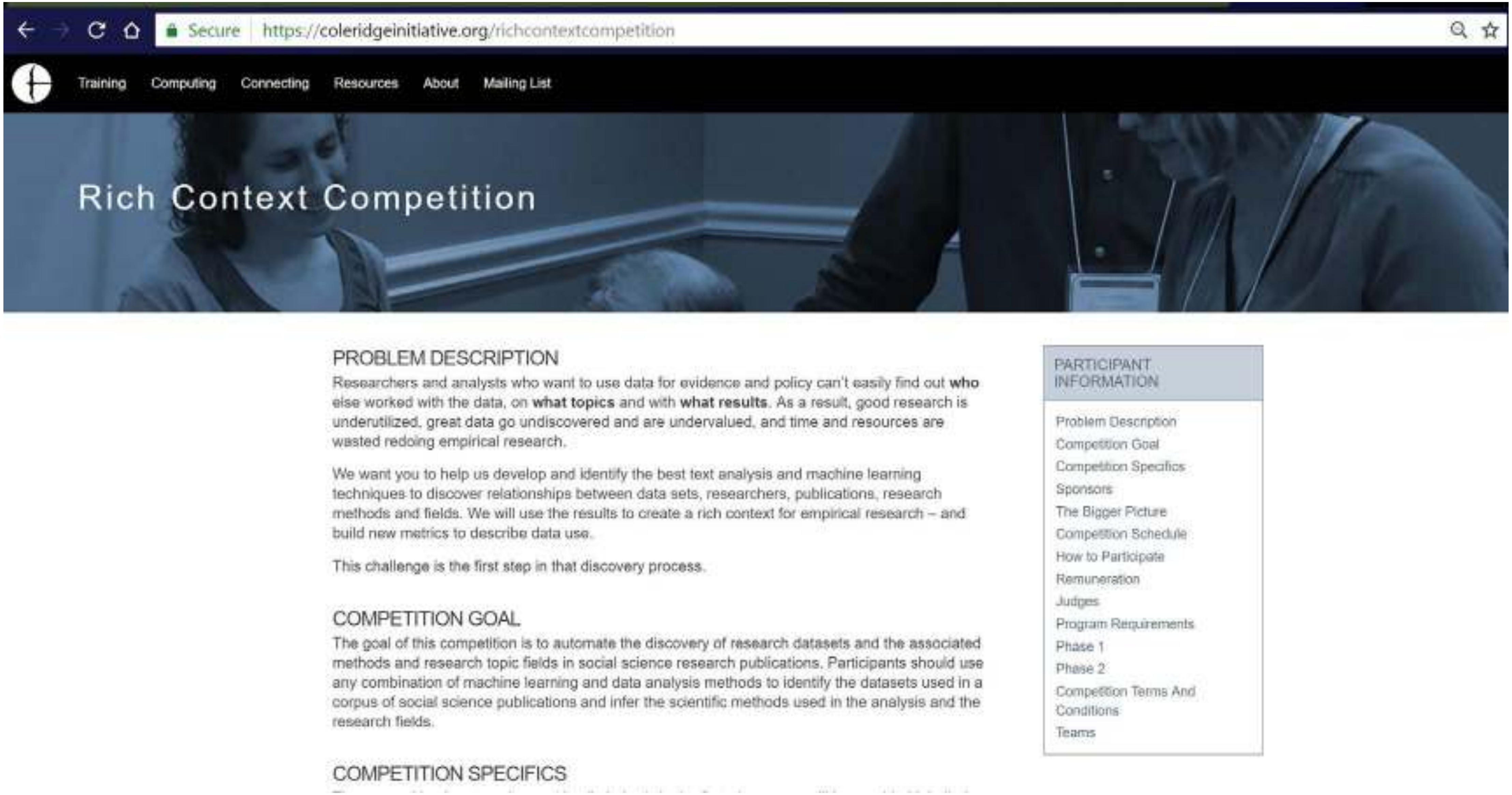
@Elena Semenova sorry about this, might have gotten lost in the shuffle a while back but Vivek did provide the following information on `jailtime` in an e-mail: `Jail time is calculated on how much time inmate spent in jail prior to coming to prison. He does get credit for time served at all jails prior coming to prison.` Thought I had circulated, but maybe only updated on the metadata in the explorer (edited)

👍 2

Timeline



		July – December 2018: **Design**	**Jan-June 2019:** **Make**	**July-Dec 2019** **Measure and Analyze**	**Jan-June 2020** **Improve**
Platform	**Activity**	- Data Model to incorporate additional metadata about datasets, users, user profiles, and user interactions (i.e., annotations, and explicit connections between datasets, people, and projects) -Telemetry Module to automatically collect structured events emitted by platform	- Deploy Data Model - Deploy Telemetry Module	- Assess Data Model Functionality - Assess Telemetry measures - Open source for community feedback	- Modify Data model with input from Rich Context - Modify Telemetry Module with input from rich context
	Deliverable	Data model Telemetry module	Operational Data Model Functioning Telemetry Module Functioning prototype Initial Jupyter-ADRF integration	QA report Initial prototype stabilized and productionized	Stable and complete version of the application fully integrated to the ADRF Platform. Open sourced
Input Elements	**Activity**	-Identify and prepare corpora (ICPSR; Bundesbank; Policy area) -Gather requirements	Generate Seed metadata generated ((ICPSR; Bundesbank; Policy area)	Review metadata developed by users Benchmark and revise	Modify and refine metadata capture and documentation
	Deliverable	Three corpora Set of requirements for metadata: comments and annotations on files and datasets, discussions, and contextual recommendations	Metadata for three corpora:	QA and improvement report on the quality of each element	Plan for future improvement
Rich Context	**Activity**	-Design gamification strategy - Design Pre/Post Survey design - Develop Telemetry measures - Research UX for the collaborative user interfaces i) an interface to help users to ingest Datasets, ii) an interface to help users to create comments and code snippets for Datasets, and iii) an interface to help users to search for Datasets -Design learning approach	Deploy interface Administer Pre survey Capture logging information Test gamification strategy Test learning approach	Review interface Administer post survey Review logging information Review feed back to platform Revise learning approach	Modify and refine interfaces, surveys and learning model
	Deliverable	Survey Telemetry measures Wireframes for the interfaces Learning model	Survey results Log results Gamification results Learning results	Survey results and pre/post analysis Revised UX, feedback loop Revised learning model	Functioning rich context module incorporating human and automated elements with continuous feedback loops to platform



Rich Context Competition

PROBLEM DESCRIPTION

Researchers and analysts who want to use data for evidence and policy can't easily find out **who** else worked with the data, on **what topics** and with **what results**. As a result, good research is underutilized, great data go undiscovered and are undervalued, and time and resources are wasted redoing empirical research.

We want you to help us develop and identify the best text analysis and machine learning techniques to discover relationships between data sets, researchers, publications, research methods and fields. We will use the results to create a rich context for empirical research – and build new metrics to describe data use.

This challenge is the first step in that discovery process.

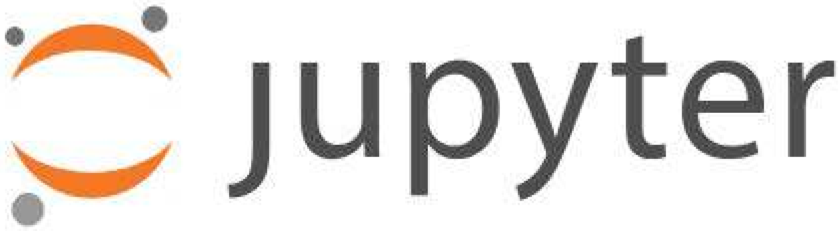
COMPETITION GOAL

The goal of this competition is to automate the discovery of research datasets and the associated methods and research topic fields in social science research publications. Participants should use any combination of machine learning and data analysis methods to identify the datasets used in a corpus of social science publications and infer the scientific methods used in the analysis and the research fields.

COMPETITION SPECIFICS

Key challenges to be solved with metadata – particularly for federal statistical system

- Limited internal capacity
- Security
- Legal mandates surrounding access and use
- Data sharing issues
  - cost
  - burden
  - data quality
  - data documentation
  - risk of bad analysis



Comments and questions?

- If interested in contributing – contact me at
- Julia.lane@NYU.EDU
- More info at https://coleridgeinitiative.org and http://jupyter.org