Automating Metadata Documentation: A New Initiative

(OR DATA, DATA, EVERYWHERE – WE HAVE TO STOP AND THINK)

Julia Lane

New York University

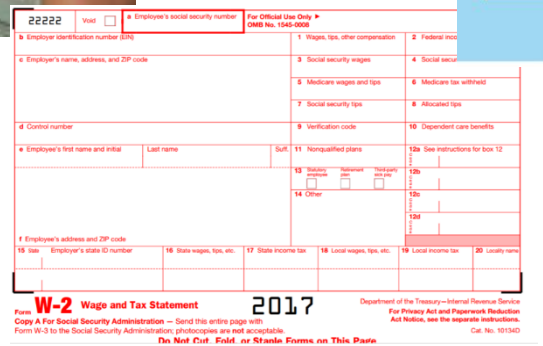
(with more contributors than I can count - but notably Frauke Kreuter, Rayid Ghani, Clayton Hunter, Brian Granger, Stefan Bender, Christian Hirsch, Hendrik Doll, Stu Feldman, Drew Gordon, Jonathon Morgan, Ekaterina Levitskaya, Daniel Castellani, Rafael Alves, Ophir Frieder, Jordan Boyd Graber, Evgeny Klochikhin, Christian Herzog, Ian Mulvaney, Alan Maloney, Daniel Hook)

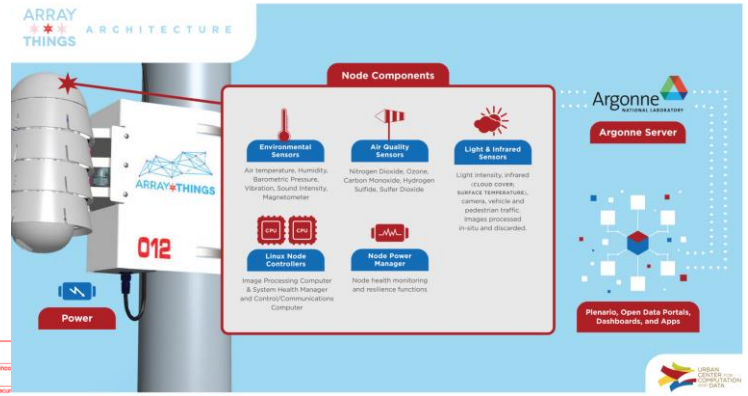
Takeawys

- Massive supply of new granular data
- New demand - local/city/regions
- Data science offers engagement and feedback tools
- Statistical skills critical
- Need to rethink data infrastructure

Data availability is burgeoning; curation, access and reproducibility is not…we need new measurement, tools and institutions

Much, much more data





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

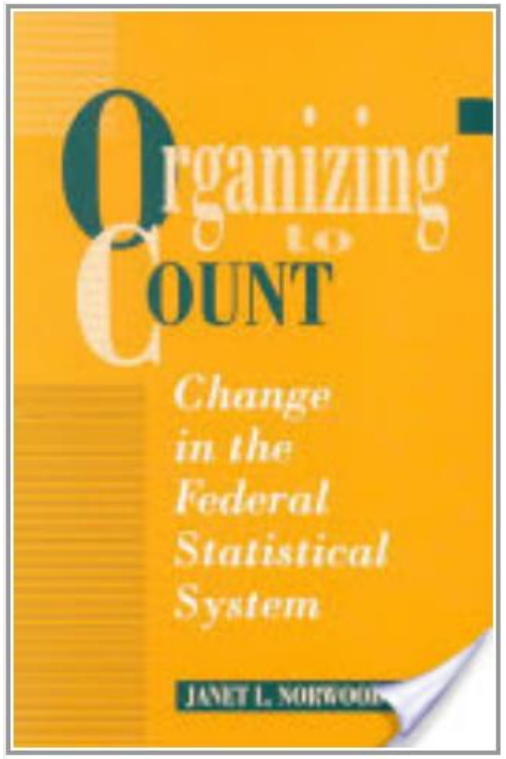
- Scaper – A Python library for soundscape synthesis and augmentation
- Audio-Annotator – A Javascript web interface for annotating audio data
- Raster Join
- Urban Pulse

There are many core measurement issues

- New products
  - Services
  - Intangibles
- New people
  - Immigration
  - Globalization
- New boundaries
  - Local
  - Regional
  - Cross national
- New organizations
  - Firms
  - Platforms
  - Networks
- New tradeoffs
  - Timeliness?
  - Closeness to core measure?
  - Coverage?
  - Geographic detail
  - Longitudinal Consistency
- New privacy constraints

Skills and institutional gaps

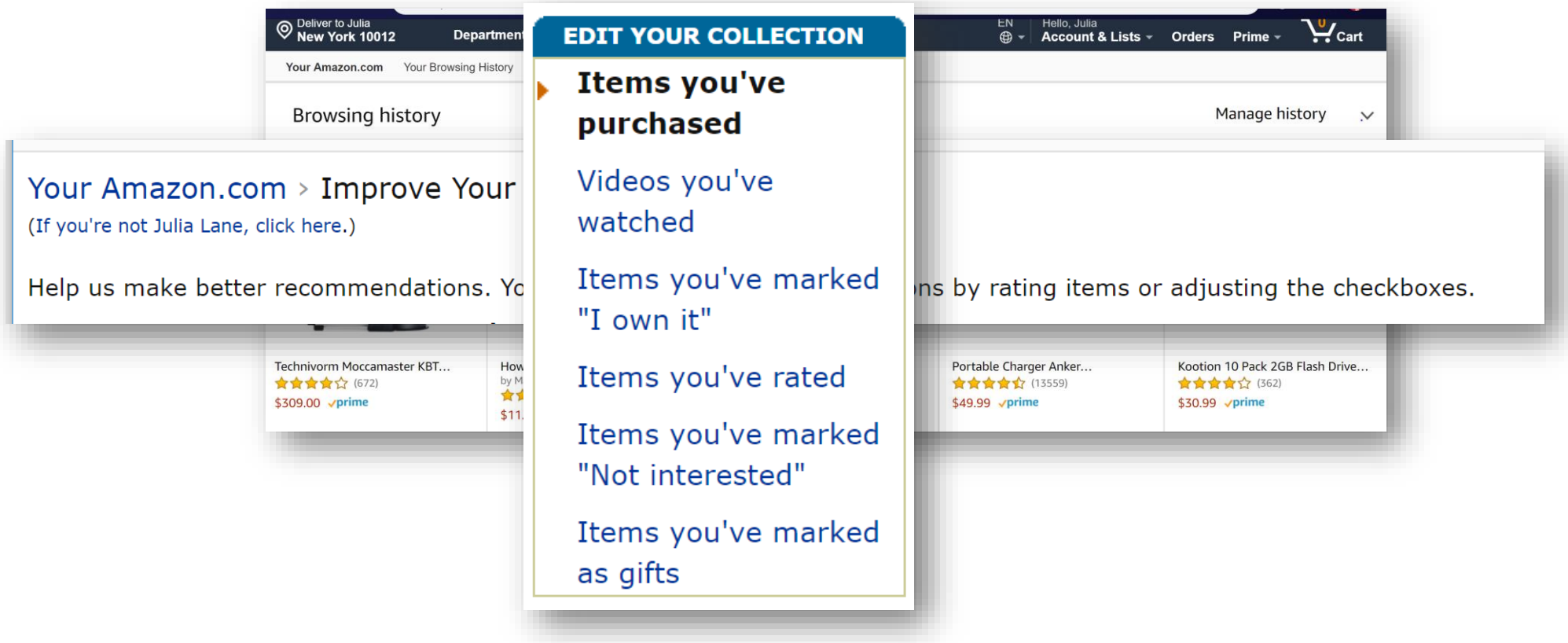
- The past



- The future

Vision

Private Sector businesses collect, curate and disseminate data Search and Discovery is automated



Other sciences have done the same

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.[14][15] She is Principal Investigator (PI) of the myGrid,[16] BioCatalogue[17] and myExperiment[18] projects and co-leads the Information Management Group (IMG) with Norman Paton.[19][20]

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[21] to 1982.

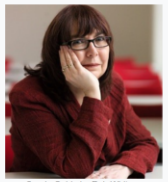
Research [edit]

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*,[10] 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 Whitrow

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[2] Taverna workbench[3][4][4] Software Sustainability Institute The Seven Deadly Sins of Bioinformatics[5]
Spouse(s)	Ian Cottam (m, 2003)[6]
Awards	Jim Gray e-Science Award (2008)

The Taverna Suite of Tools

Workflow Repository

myexperiment

Service Catalogue

BioCatalogue

PANACEA Registry

BiodiversityCatalogue "The Biodiversity Sciences Web Services Registry"

Activity and Service Plug-in Manager

SoapLab

UNICORE

R

biomart

BeanShell

SADI

User Interfaces

Taverna Workbench

Taverna Lite

Workflow Provenance

W3C

Workflow Server

Workflow Components

Interaction Server

Secure Service Access

Web Portals / Gateways

Client User Interfaces

EVO Environmental Virtual Observatory

Third Party Tools

Galaxy

Google refine

Player

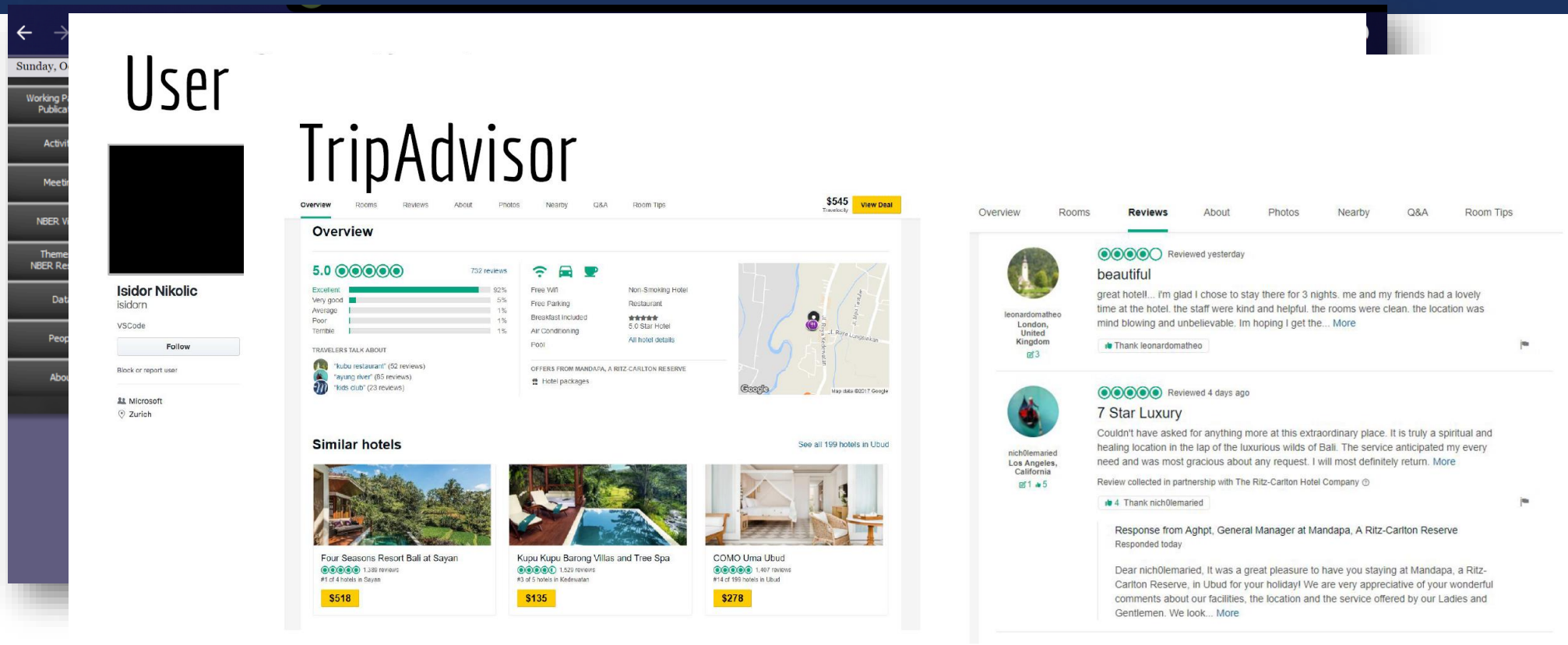
Virtual Machine

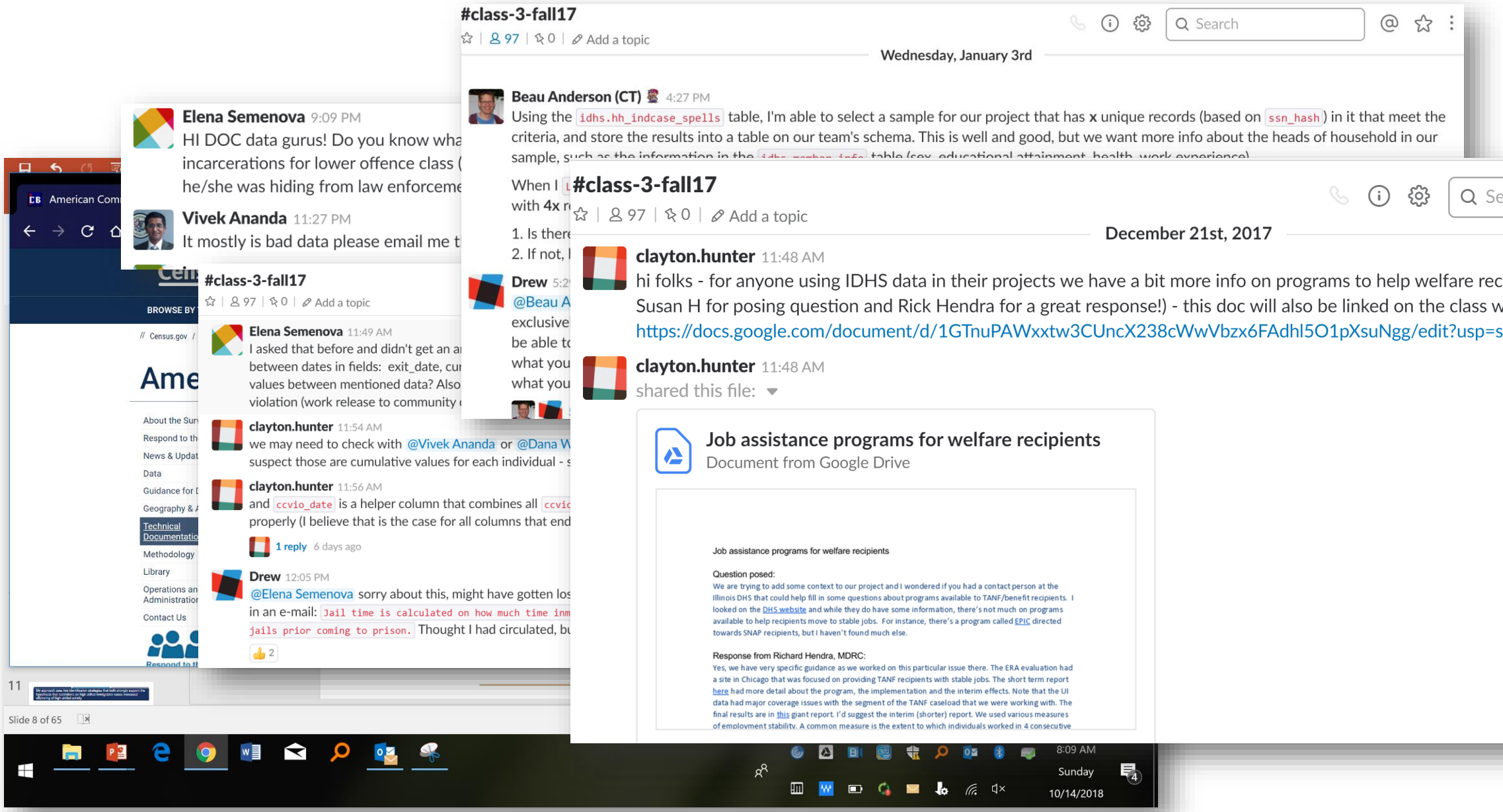
Command Line

Prog APIs

OSGi Alliance

Vision for Search and Discovery

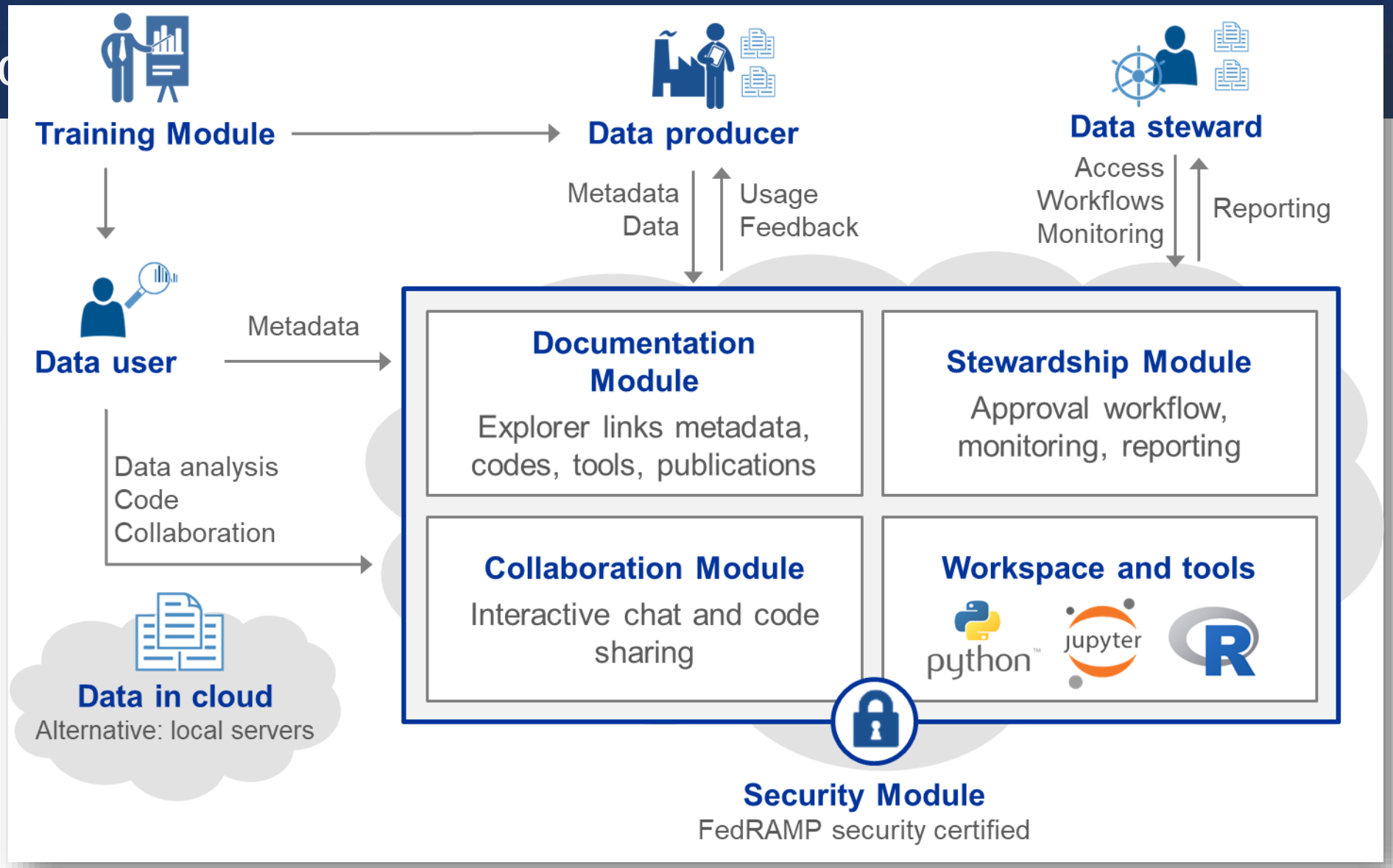


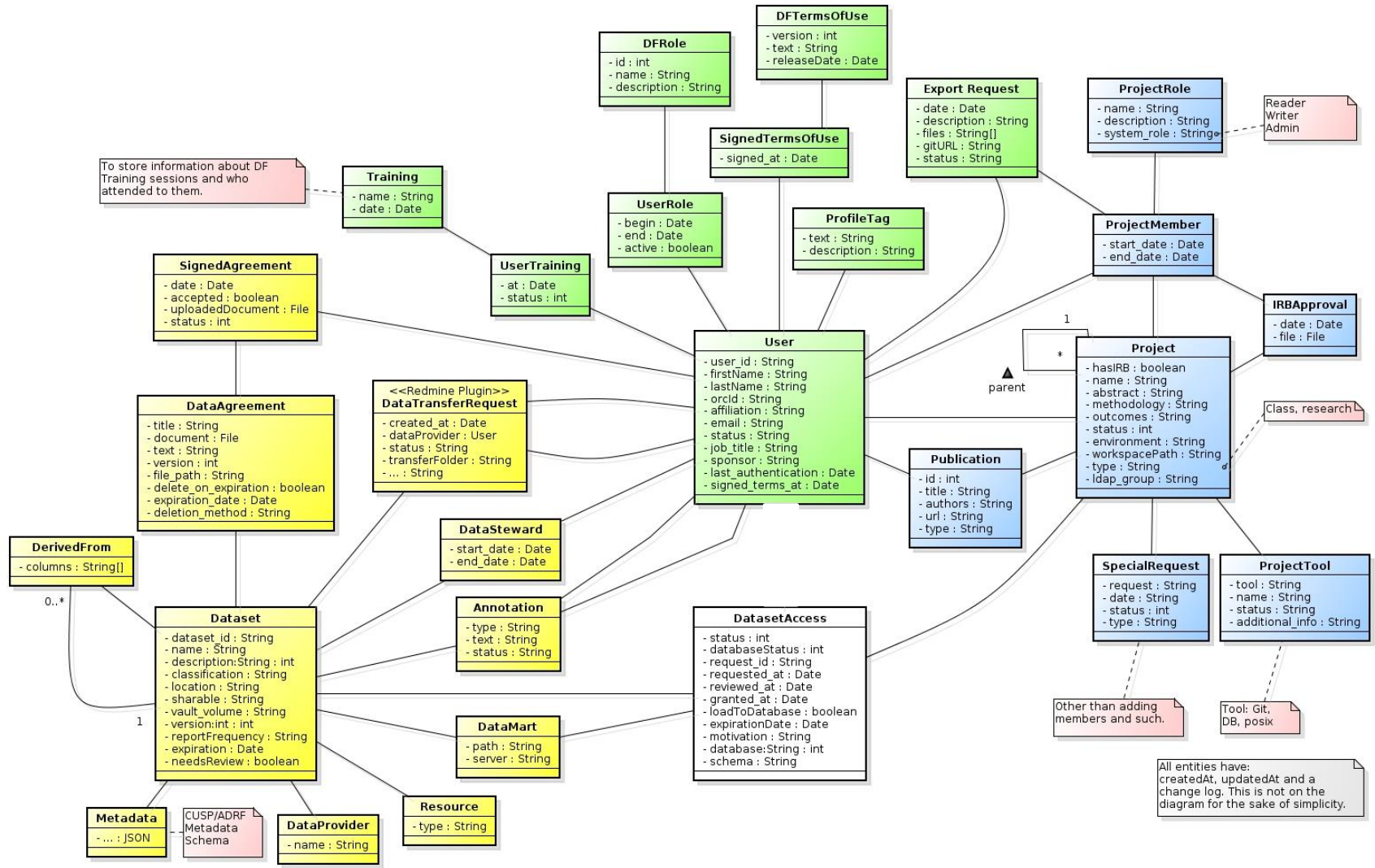

#class-3-fall17
97 | 0 | Add a topic
Search
Wednesday, January 3rd
Beau Anderson (CT) 4:27 PM
Using the idhs.hh_indcase_spells table, I'm able to select a sample for our project that has x unique records (based on ssn_hash) in it that meet the criteria, and store the results into a table on our team's schema. This is well and good, but we want more info about the heads of household in our sample, such as the information in the idhs_member_info table (sex, educational attainment, health, work experience)
When I
with 4x
1. Is there
2. If not,
Drew 5:2
@Beau A
exclusive
be able to
what you
what you
#class-3-fall17
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 rec
Susan H for posing question and Rick Hendra for a great response!) - this doc will also be linked on the class w
https://docs.google.com/document/d/1GTnuPAWxxtw3CUncX238cWwVbzx6FAdhl5O1pXsuNgg/edit?usp=s
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
#class-3-fall17
97 | 0 | Add a topic
Elena Semenova 11:49 AM
I asked that before and didn't get an a
between dates in fields: exit_date, cu
values between mentioned data? Also
violation (work release to community
clayton.hunter 11:54 AM
we may need to check with @Vivek Ananda or @Dana W
suspect those are cumulative values for each individual - s
clayton.hunter 11:56 AM
and ccvio_date is a helper column that combines all ccvi
properly (I believe that is the case for all columns that end
1 reply 6 days ago
Drew 12:05 PM
@Elena Semenova sorry about this, might have gotten los
in an e-mail: Jail time is calculated on how much time inm
jails prior coming to prison. Thought I had circulated, bu
2
Elena Semenova 9:09 PM
HI DOC data gurus! Do you know wha
incarcerations for lower offence class (
he/she was hiding from law enforceme
Vivek Ananda 11:27 PM
It mostly is bad data please email me t
American Com
BROWSE BY
Census.gov
Ame
About the Sur
Respond to th
News & Updat
Data
Guidance for D
Geography & A
Technical Documentatio
Methodology
Library
Operations an Administratior
Contact Us
Slide 8 of 65
8:09 AM
Sunday
10/14/2018

Data Discovery

- Step 1: Create the set of corpora and metadata (computer science technology) - Competition
- Step 2; Figure out how you learn from it and automate it (machine learning techniques) - Engagement
- Step 3: Gamification – recognize and emphasize patterns (with human curation) – Rinse and repeat

Operatic


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
python
jupyter
R
Data in cloud
Alternative: local servers
Security Module
FedRAMP security certified

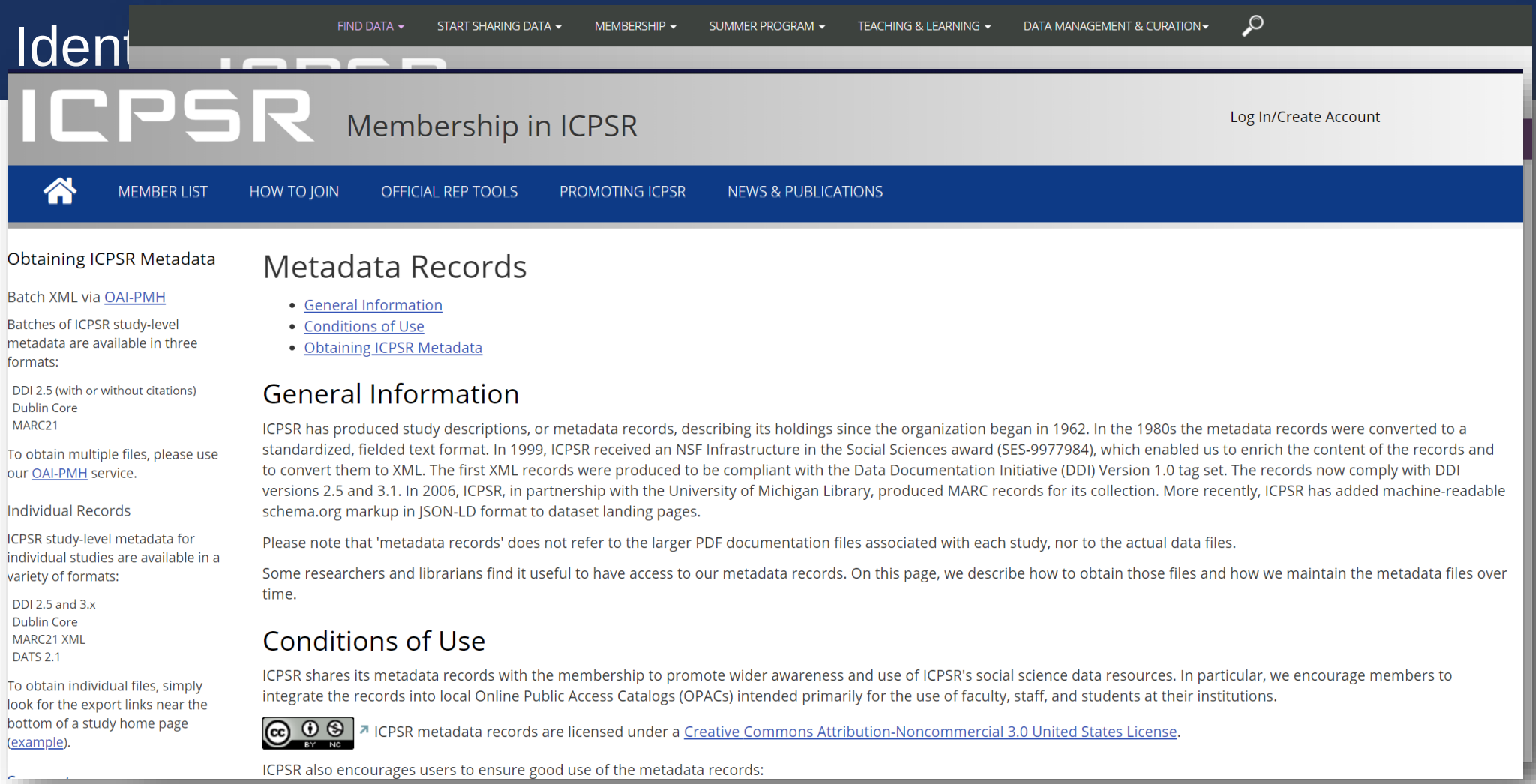

DFTermsOfUse
- version : int
- text : String
- releaseDate : Date
DFRole
- id : int
- name : String
- description : String
Export Request
- date : Date
- description : String
- files : String[]
- gitURL : String
- status : String
ProjectRole
- name : String
- description : String
- system_role : String
Reader
Writer
Admin
SignedTermsOfUse
- signed_at : Date
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
1
*
parent
Project
- hasIRB : boolean
- name : String
- abstract : String
- methodology : String
- outcomes : String
- status : int
- environment : String
- workspacePath : String
- type : String
- ldap_group : String
Class, research
<<Redmine Plugin>>
DataTransferRequest
- created_at : Date
- dataProvider : User
- status : String
- transferFolder : String
- ... : String
DataAgreement
- title : String
- document : File
- text : String
- version : int
- file_path : String
- delete_on_expiration : boolean
- expiration_date : Date
- deletion_method : String
Publication
- id : int
- title : String
- authors : String
- url : String
- type : String
DataSteward
- start_date : Date
- end_date : Date
DerivedFrom
- columns : String[]
0..*
1
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
SpecialRequest
- request : String
- date : String
- status : int
- type : String
ProjectTool
- tool : String
- name : String
- status : String
- additional_info : String
DataMart
- path : String
- server : String
Other than adding members and such.
Tool: Git, DB, posix
Metadata
- ... : JSON
CUSP/ADRF Metadata Schema
DataProvider
- name : String
Resource
- type : String
All entities have: createdAt, updatedAt and a change log. This is not on the diagram for the sake of simplicity.

Sch

		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

Competition

Ident


FIND DATA
START SHARING DATA
MEMBERSHIP
SUMMER PROGRAM
TEACHING & LEARNING
DATA MANAGEMENT & CURATION
ICPSR
Membership in ICPSR
Log In/Create Account
MEMBER LIST
HOW TO JOIN
OFFICIAL REP TOOLS
PROMOTING ICPSR
NEWS & PUBLICATIONS
Obtaining ICPSR Metadata
Batch XML via OAI-PMH
Batches of ICPSR study-level metadata are available in three formats:
DDI 2.5 (with or without citations)
Dublin Core
MARC21
To obtain multiple files, please use our OAI-PMH service.
Individual Records
ICPSR study-level metadata for individual studies are available in a variety of formats:
DDI 2.5 and 3.x
Dublin Core
MARC21 XML
DATS 2.1
To obtain individual files, simply look for the export links near the bottom of a study home page (example).
Metadata Records
General Information
Conditions of Use
Obtaining ICPSR Metadata
General Information
ICPSR has produced study descriptions, or metadata records, describing its holdings since the organization began in 1962. In the 1980s the metadata records were converted to a standardized, fielded text format. In 1999, ICPSR received an NSF Infrastructure in the Social Sciences award (SES-9977984), which enabled us to enrich the content of the records and to convert them to XML. The first XML records were produced to be compliant with the Data Documentation Initiative (DDI) Version 1.0 tag set. The records now comply with DDI versions 2.5 and 3.1. In 2006, ICPSR, in partnership with the University of Michigan Library, produced MARC records for its collection. More recently, ICPSR has added machine-readable schema.org markup in JSON-LD format to dataset landing pages.
Please note that 'metadata records' does not refer to the larger PDF documentation files associated with each study, nor to the actual data files.
Some researchers and librarians find it useful to have access to our metadata records. On this page, we describe how to obtain those files and how we maintain the metadata files over time.
Conditions of Use
ICPSR shares its metadata records with the membership to promote wider awareness and use of ICPSR's social science data resources. In particular, we encourage members to integrate the records into local Online Public Access Catalogs (OPACs) intended primarily for the use of faculty, staff, and students at their institutions.
ICPSR metadata records are licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
ICPSR also encourages users to ensure good use of the metadata records:

Creat

inputs > data_sets_

```
[
    {
        "dat
        "uni
        "tit
        "nam
        "des
surveys fiel
present data
social and p
opinions on
1952 Nationa
attitudes an
interview sc
collect data
before and a
```

inputs > ICPSR_citations_metadata_enriched_full_combined_coded_status.csv

2015-11-10	10.3886/ICPSR07355	ANES 1974 Time Series St	1978-01-01	10.1017/S000712340000	doi	Journal Article	literature	Generational replacemer
2015-11-10	10.3886/ICPSR07381	ANES 1976 Time Series St	1978-01-01	10.1017/S000712340000	doi	Journal Article	literature	Generational replacemer
1992-02-16	10.3886/ICPSR07233	Political Change in Britain,	1978-01-01	10.1017/S000712340000	doi	Journal Article	literature	Generational replacemer
1992-02-16	10.3886/ICPSR07234	Political Change in Britain,	1978-01-01	10.1017/S000712340000	doi	Journal Article	literature	Generational replacemer
1992-02-16	10.3886/ICPSR07004	Political Change in Britain,	1978-01-01	10.1017/S000712340000	doi	Journal Article	literature	Generational replacemer
2016-09-20	10.3886/ICPSR07213	ANES 1952 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2016-09-22	10.3886/ICPSR07214	ANES 1956 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2016-09-22	10.3886/ICPSR07215	ANES 1958 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07216	ANES 1960 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2016-12-01	10.3886/ICPSR07217	ANES 1962 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07235	ANES 1964 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07259	ANES 1966 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07281	ANES 1968 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07298	ANES 1970 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2016-09-20	10.3886/ICPSR07010	ANES 1972 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07355	ANES 1974 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07381	ANES 1976 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2015-11-10	10.3886/ICPSR07655	ANES 1978 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
2016-02-26	10.3886/ICPSR07763	ANES 1980 Time Series St	1982-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	The decline of electoral
1999-10-07	10.3886/ICPSR09093	American National Electio	1992-01-01	http://www.jstor.org/stabl	url	Journal Article	literature	'Sophisticated' voting in

Point t

ARTICLE

Reported Consumption of Low
Foods by American Children a

Nutritional and Health Correlates, NHANES III

Ashima K. Kant, PhD

Objective: To examine the contribution of foods of modest nutritional value to the diets of American children and adolescents.

Methods: The data were from the third National Health and Nutrition Examination Survey, 1988 to 1994, and included 4852 children and adolescents, aged 8 to 18 years. Foods reported in the 24-hour dietary recall were grouped into the following low-nutrient-density (LND) food categories: visible fat; table sweeteners, candy, and sweetened beverages; baked and dairy desserts; salty snacks; and miscellaneous. The independent association of the number of LND foods mentioned in the recall with intake of food groups, macronutrients, micronutrients, and body mass index was examined by means of regression procedures to adjust for multiple covariates.

Results: The LND foods contributed more than 30% of daily energy, with sweeteners and desserts jointly

MANY AMERICAN
dren and adoles
consume diets
provide marg
amounts of sever
trients, including vitamins A, E, B_6, a
late, and the minerals calcium, m
sium, iron, and zinc.[1,2] These ess
nutrients have well-known metabolic
tions, and many have been linked to
promotion and disease prevention.
cent survey data also suggest an alarming
trend in increasing prevalence of adipos-
ity in US children and adolescents.[7,8] Es-
timates from the Continuing Survey of
Food Intakes by Individuals, 1994 to 1996,
suggest a dramatic increase in consump-
amination of the contribution of added
sugar or carbonated beverages to the di-
ets of children.[11-18] There is some evi-
dence that food selection patterns fa-
vored in childhood may track through
adult years[19]; therefore, a better under-

METHODS

The NHANES III is a multistage, stratified, probability sample of the noninstitutionalized, civilian US population, aged 2 months and older.[20] The survey was conducted by the National Center for Health Statistics and included administration of a questionnaire at home and a full medical examination along with a battery of tests in a special mobile examination center.[20] Demographic and medical history information was obtained during the household interview. The examination at the mobile examination center included physical and dental examinations, dietary interview, body measurements, and collection of blood and urine samples. Body weight, height, and circumference at various body sites were measured by standardized procedures in the mobile examination center.[20]

DIETARY ASSESSMENT METHOD

A 24-hour dietary recall was collected by a trained dietary interviewer in a mobile examination center interview with the use of an automated, microcomputer-based interview and coding system.[20] The type and amount of foods consumed were recalled with recall aids such as abstract food models, special charts, measuring cups, and rulers to help in quantifying the amounts consumed. Special probes were used to help the recall of commonly forgotten items such as condiments, accompaniments, fast foods, and alcoholic beverages.

ANALYTIC SAMPLE

All NHANES III respondents aged 8 to 18 years with a 24-hour recall considered complete and reliable by the National Center for Health Statistics were eligible for inclusion in this study (n=4889). Recalls of pregnant (n=34) or nursing (n=3) respondents were excluded, to result in a final analytic sample size of 4852 (2383 boys and 2469 girls) representing a population estimate of 38738796 children and adolescents.

ages in the recall, (2) amount (in grams) of LND foods and beverages, and (3) percentage of total energy from LND foods and beverages.

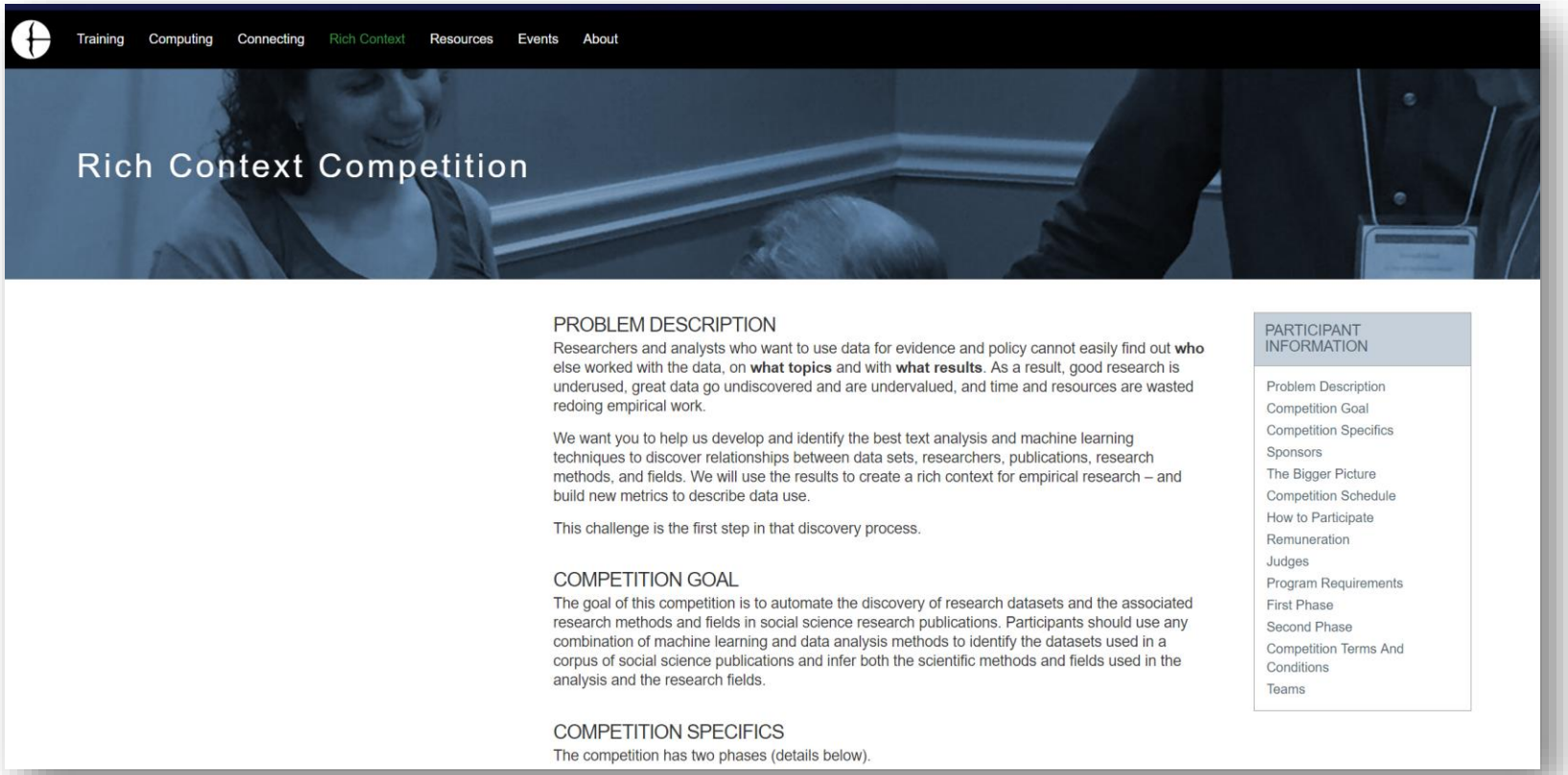
The NHANES III nutrient database for individual foods, which is derived from the US Department of Agriculture's Survey Nutrient Database, was used for determining energy and nutrient content of all foods.[23] We examined the mean intake of selected micronutrients (vitamins A, E, C, B_6, and folate, and the minerals calcium, magnesium, iron, and zinc) by tertiles of the reported number of LND foods. The reported intake of each nutrient was also examined with reference to the most recent age-sex–specific standard available, the estimated average requirement.[3-6] As an estimate of dietary misreporting, the ratio of reported energy intake to energy expenditure for basal needs was computed. Energy expenditure for basal needs was estimated by prediction equations developed by the Dietary Reference Intakes Committee for normal and overweight or obese 3- to 18-year-olds.[24]

Data on serum concentrations of vitamin C, folate, and carotenoids were obtained from the National Center for Health Statistics public release compact disks.[25,26] The methods used for measurement of these serum analytes and their associated errors have been described.[25,26] Serum folate, ascorbate, and the carotenoids—α-carotene, β-carotene, β-cryptoxanthin, lutein-zeaxanthin, and lycopene—were chosen because the dietary intake of each nutrient is believed to be a major determinant of serum concentration of the respective nutrient and can thus serve as a biomarker of dietary exposure.[4,5] Serum homocysteine has been reported as an independent predictor of coronary heart disease risk and may have dietary determinants.[27,28]

STATISTICAL ANALYSES

The number, amount, and proportion of daily energy contributed by all LND foods and subgroups were computed, and each LND food variable was examined as weighted tertiles. The mean daily energy, percentage of energy from macronutrients, and

Run competition (https://coleridgeinitiative.org/richcontextcompetition)



Hold breath…

- Twenty teams
  - Half research labs
  - Half students
- Eight countries
  - US
  - Korea
  - Germany
  - Russia
  - Singapore
  - Indonesia
  - Spain
  - India

Engagement

NYU Game Center

NYU | TISCH SCHOOL OF THE ARTS GAME CENTER

ABOUT ACADEMICS EVENTS PEOPLE GAMES BLOG

rker@nyu.edu

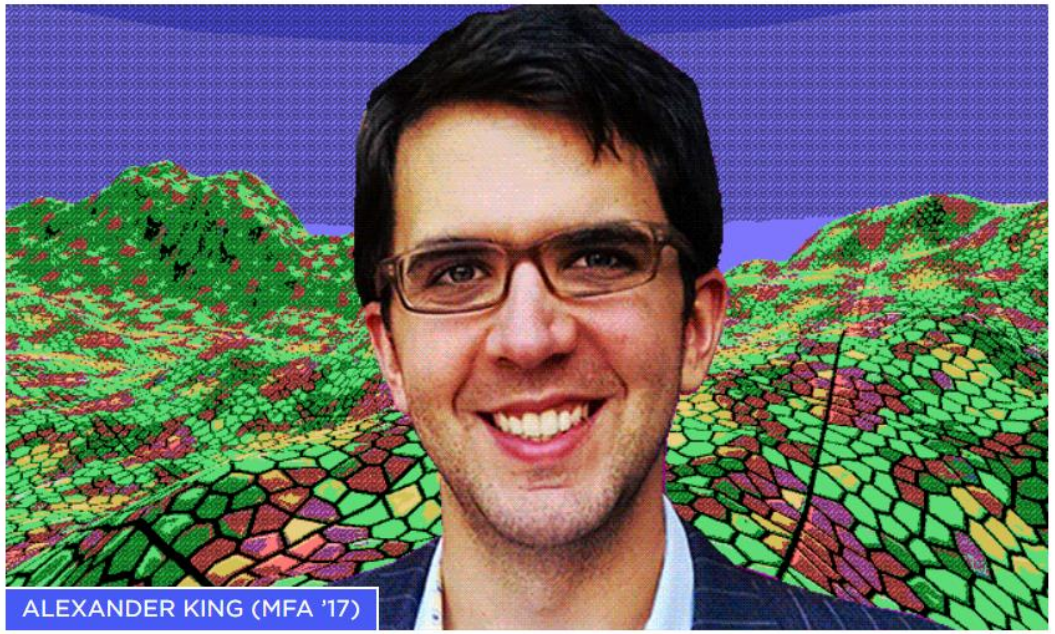
Matt Parker

MATT PARKER

COURSES TAUGHT:
- DESIGNING FOR IMPACT: ICIVICS
- DESIGNING FOR MUSEUMS: AMNH
- CODE LAB 0
- CODE LAB 2
- CODE LAB 1
- THESIS 1
- GAMES 101

Matt Parker is a game designer, teacher, and new media artist. His work has been displayed at the American Museum of Natural History, Brooklyn Academy of Music, SIGGRAPH Asia, the NY Hall of

Alexander King (MFA '17)



ALEXANDER KING (MFA '17)

Alexander King is an independent developer, freelance game designer and consultant. His work centers on data-driven design and simulation, and his games have been featured in festivals like ALT.CTRL.GDC and IndieCade. Alexander has an MFA in Game Design from the NYU Game Center. Before working in games, Alexander was an analytics consultant working in finance and ecommerce. Now the economies he models are largely fictional ones.

A Game Designer's View of Engagement

BACKGROUND

What exists

- A large database for researchers and lawmakers to access.

The problem:

- Data is difficult to parse and utilize effectively.
- Users are not encouraged to interact with or contribute to data.

Project goals

- Have researchers establish context and meaning to datasets.
- Enable users to easily find related datasets and research.
- Create an effective system which encourages researchers to add metadata and organizational system to datasets.

Jupyter

Making Computational Research with Sensitive Data Possible and Valuable

Brian E. Granger
Associate Professor
Cal Poly

Julia Lane
Professor
NYU

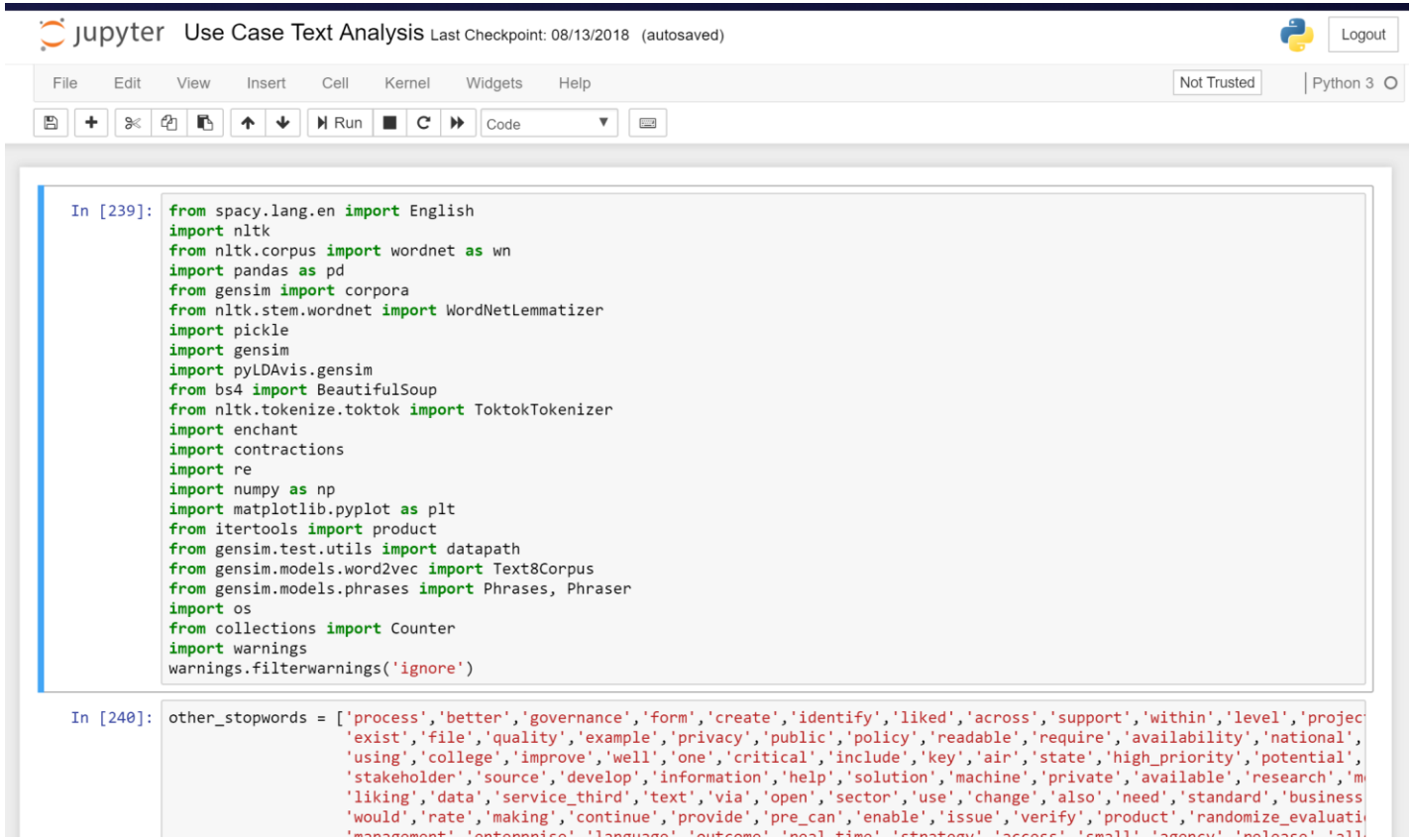
Fernando Perez
Assistant Professor
UC Berkeley

Alfred P. Sloan FOUNDATION

SCHMIDT FUTURES

Overdeck Family Foundation

The role of Jupyter Notebooks



A computer scientist's view of engagement

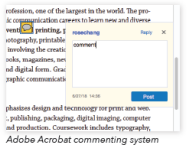
Designing Comments

By Rose Chang, Meredith Granger, Alena Mueller, and Taka Shimokobe
Thanks to Tim George, Brian Granger, and Ana Ruvalcaba

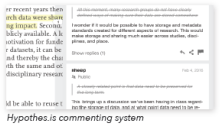
JupyterLab is adding real-time collaboration features to enable multiple users to view and edit documents at the same time. In the context of real-time collaboration, a commenting and annotation system becomes important to enable effective remote teamwork.

1 Background Research

The first step of any design is researching similar existing systems. We read W3C's Web Annotation Working Group documents to help us build a conceptual model and vocabulary. In particular, we examined users' feedback about existing systems to discover the advantages and issues of each.



Adobe Acrobat commenting system

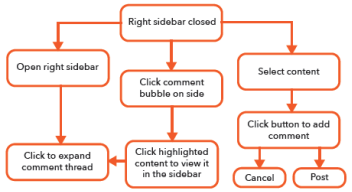


Hypothes.is commenting system

Figma commenting system

DropBox commenting system

We planned by making an **interaction map**: What are all the things a user might want to do?



Why it's needed

Users of JupyterLab will soon be able to edit documents in real-time. In these situations, users often need to leave contextual comments and annotations for themselves and others.

What it should be

Our commenting system needs to work in notebooks, datasets, text files, and other formats.

2 Defining Why and What

3 Answering Big Questions

As we planned interactions, big questions started to come up. What does it mean to resolve a comment? How is that different from deleting? What happens on a past version of the document? Here are some of the rules we came up with:

Time and History

-The document changes in real-time and only comments on current content show.

-If a previous version of the document is viewed, only the comments from that version display.

-When content is deleted, comments on it are only visible in past versions.

-In previous versions of the document, comments are shown in the state (resolved/open) that they were.

Delete and Resolve

-Delete removes a comment from all versions.

-Resolving a comment hides it unless viewing resolved comments.

-Resolved comments only show if their content exists.

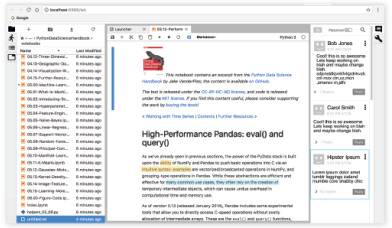
-Anyone can resolve or reopen a comment thread.

Comment Threads

-If multiple comments are in the same area of content, one click opens all threads.

-In the comment panel, users click a thread to expand it.

-Only the first comment is show in unfocused threads.



We used Figma, a art application, to make our prototypes. Figma is great for collaboration and allowed us to link one screen to another interactively. We created a functional prototype to test users without any coding.

4 Building Prototypes

5 Evaluating Our Work

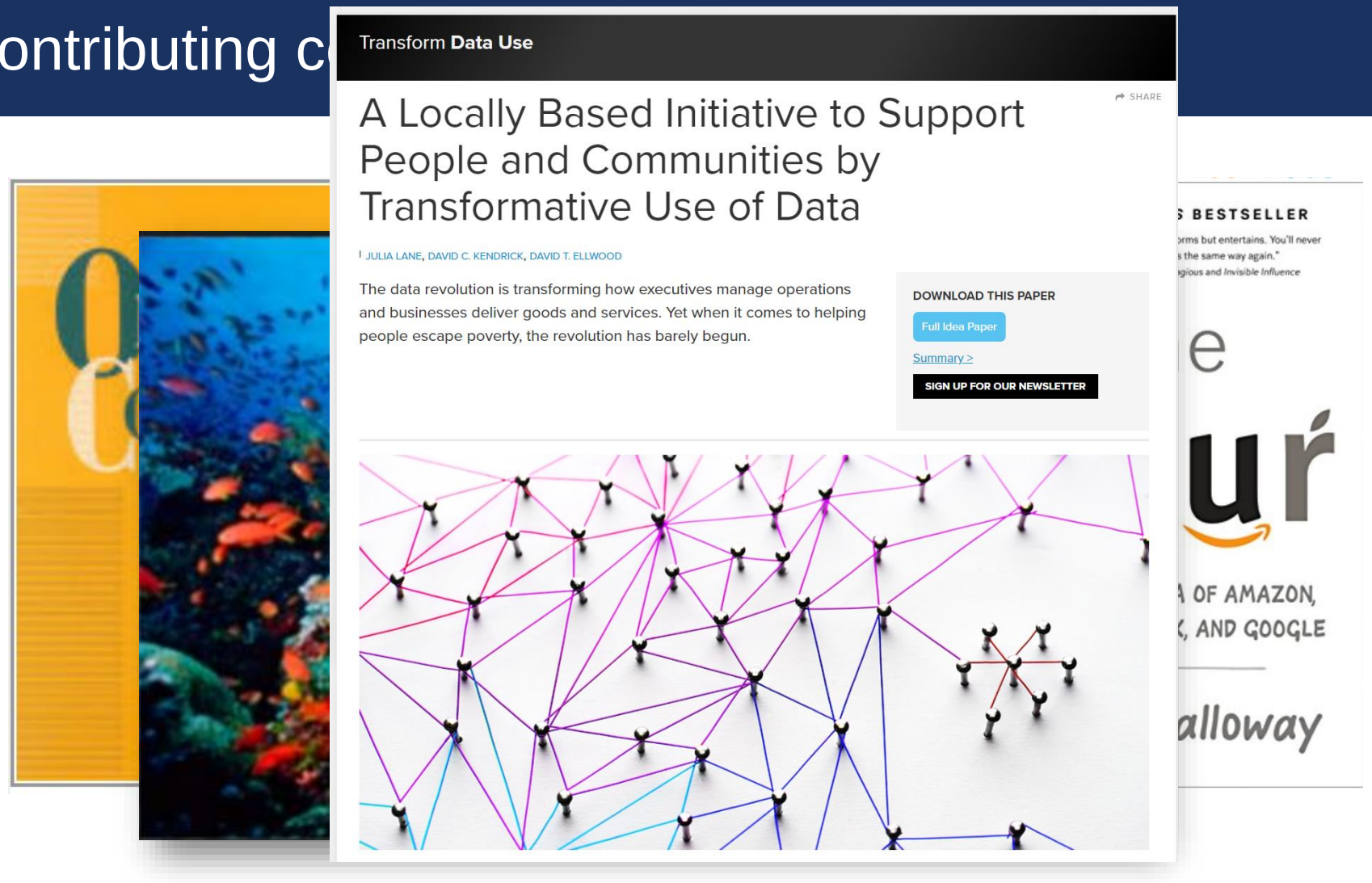
A major idea in early iterations was having three views of the comment panel. Users can view all threads, a group of threads, or an individual thread.

But typical users are going to have much simpler mental models.

The cycle of evaluating our work and building prototypes that meet JupyterLab's standards for usability is going to continue for much longer. We've started brainstorming prototypes to match that.

The Future

Build a contributing c



Takeaways

- Massive supply of new granular data
- New demand - local/city/regions
- Data science offers engagement and feedback tools
- Statistical skills critical
- Need to rethink data infrastructure

Help us and be a beta tester! Email me at julia.lane@nyu.edu

Comments welcome

- Julia.lane@nyu.edu
- https://coleridgeinitiative.org