

Using open competitions to drive innovation and collaboration.

Hi, I'm Ian, I work at SAGE where I help make processes and systems work better.



Ian Mulvany
Head of Transformation, Product Innovation
SAGE Publishing



Julia Lane

- Professor, Robert F. Wagner Graduate School of Public Service, NYU
- Professor, Center for Urban Science and Progress, NYU
- Provostial Fellow, Innovation Analytics



Paco Nathan

- Advisor on ML and AI
- Previously Director of learning O'Reilly Media

Most of the work presented today is that of Julia Lane and Paco Nathan, I did a little bit.



Semantic Web

Text Mining

Competitions

Social Science

Knowledge Graphs

Open Science

These are the key topics that we will concern ourselves with today.
They seem pretty relevant to the Force11 community!



Machine Learning

Persistent Identifiers

Very sadly we have a problem with persistent identifiers. Their very absence is at the heart of our story :(



“Our job is increasingly moving from writing code, to knowing what the right libraries are to use, and what the right problems are to collaborate on”

I threw this quote in from a friend of mine. I mean, it's true, right, nowadays finding the person who has a the solution is usually the fastest way to solve things!

– *Conor Masterson*



Community Building

And when we can structure our communities so that this kind of knowledge sharing can happen better, we all win!



Rich Context



The broad name of the project I'm talking about today is the "Rich Context" project, and its about connecting data in the social sciences to the papers where that data is used.

[Training](#)[Computing](#)[Connecting](#)[Rich Context ▾](#)[Resources](#)[Events](#)[About](#)

COLERIDGE INITIATIVE ABOUT

VISION

Our goal is to change the empirical foundation of social science, statistical and public agencies in the United States and transform understanding of how our society works. We are a fast growing university-based startup that has already created dozens of pilot projects, worked with over 100 agencies—federal, state and local.

Our team is led by world renowned data scientist Julia Lane—and we are building new data platform, the NYU Administrative Data Platform, to analyze sensitive and confidential microdata and provide training and consulting services to build agency capacity to serve society.

The project is being run by the Coleridge Initiative, a research group based out of NYU. They are looking at how to use data to improve social science outcomes.



TEAM

Directors



Rayid Ghani

- Director, Center for Data Science and Public Policy
- Senior Fellow, Harris School of Public Policy
- Senior Fellow, Computational Institute, The University of Chicago



Robert Goerge

- Senior Research Fellow, Chapin Hall
- Senior Fellow, Harris School of Public Policy
- Senior Fellow, Computational Institute, The University of Chicago



Frauke Kreuter

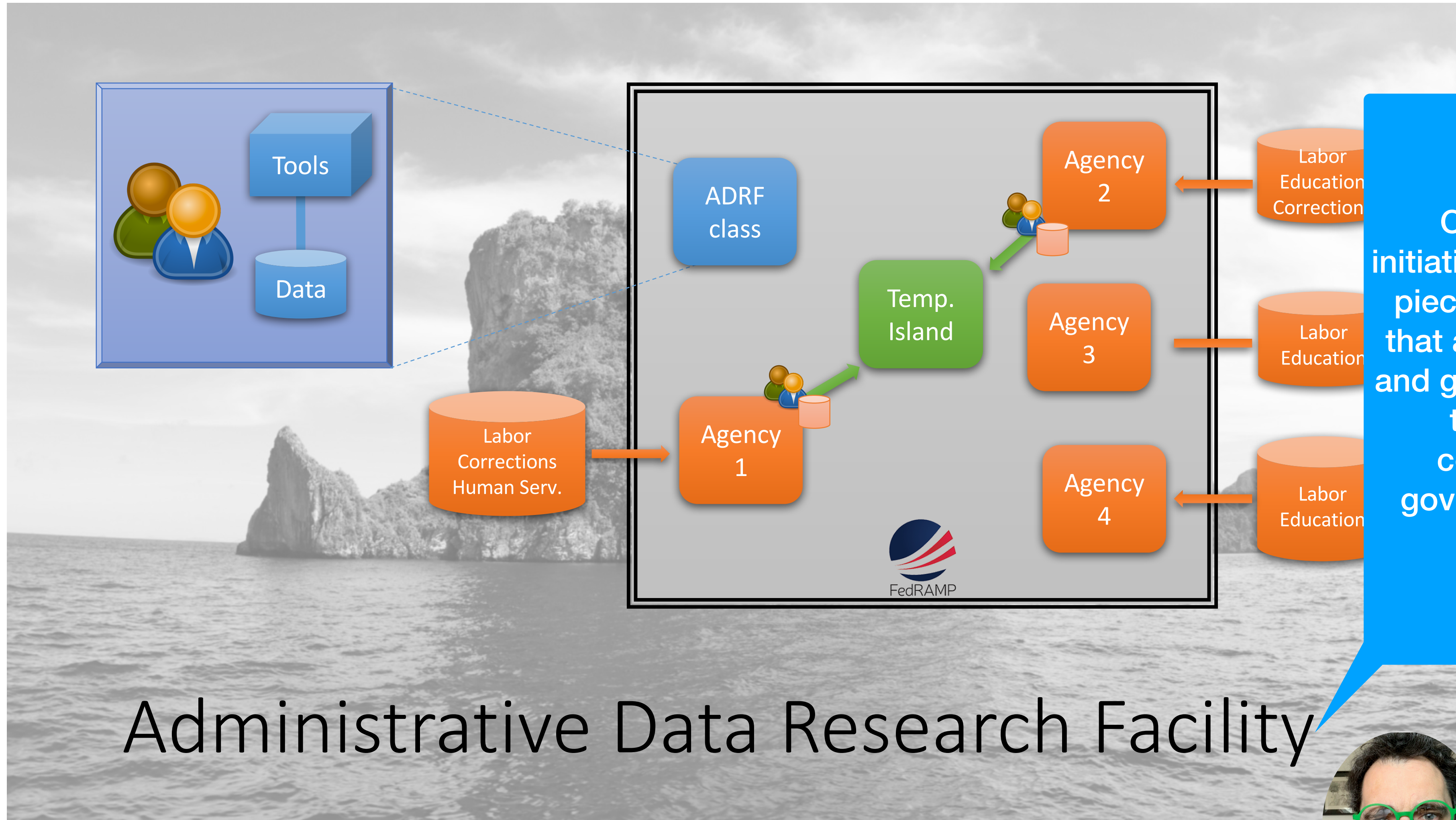
- Professor, Joint Program in Survey Methodology, University of Maryland
- Professor, Methods and Statistics, University of Mannheim
- Head, Statistical Methods Group, German Institute for Employment Research, Nuremberg



Julia Lane

- Professor, Robert F. Wagner Graduate School of Public Service, NYU
- Professor, Center for Urban Science and Progress, NYU
- Provostial Fellow, Innovation Analytics

Tiered Settings



One of their key initiatives is the ADRF - a piece of infrastructure that allows researchers and government officials to analyse and collaboration on government data in a secure way.

Administrative Data Research Facility



Typical Problems of working with Administrative data

- Sensitive data, e.g. unemployment insurance wage records, criminal records. Etc.
- Requires tiered access
- Crosswalking identifiers can be hard
- Skills within government are often low

Solutions that Coledrige / ADRF have worked on

- Secure cloud infrastructure
- Admin and reporting on access and usage
- Buy-in of a community of data experts that can help each other
- Providing hands on training to government.



I like to compare the social sciences today to where bioinformatics was say 10 or 15 years ago. There is an explosion of data, and emerging patterns for how to store, identify and collaborate on that data, it's an exciting time!



THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking



- **January 13, 2017: International and State Models for Managing Data**
 - o Charles Rothwell—National Center for Health Statistics, U.S. Department of Health and Human Services
 - o David Mancuso—Washington State Department of Social and Health Services
 - o Domenico Parisi—Mississippi State University
 - o Ivan Thaulow—Statistics Denmark
 - o Kenneth Dodge—Duke University
 - o Robert Goerge—University of Chicago
 - o Roxane Silberman—Secure Data Access Centre, France
 - o Shawna Webster—National Association for Public Health Statistics and Information Systems
 - o Stefan Bender—Deutsche Bundesbank

Public Law 115–435 115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) SHORT TITLE.—This Act may be cited as the “Foundations for Evidence-Based Policymaking Act of 2018”.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.

TITLE I—FEDERAL EVIDENCE-BUILDING ACTIVITIES

Sec. 101. Federal evidence-building activities.

TITLE II—OPEN GOVERNMENT DATA ACT

TITLE II—OPEN GOVERNMENT DATA ACT

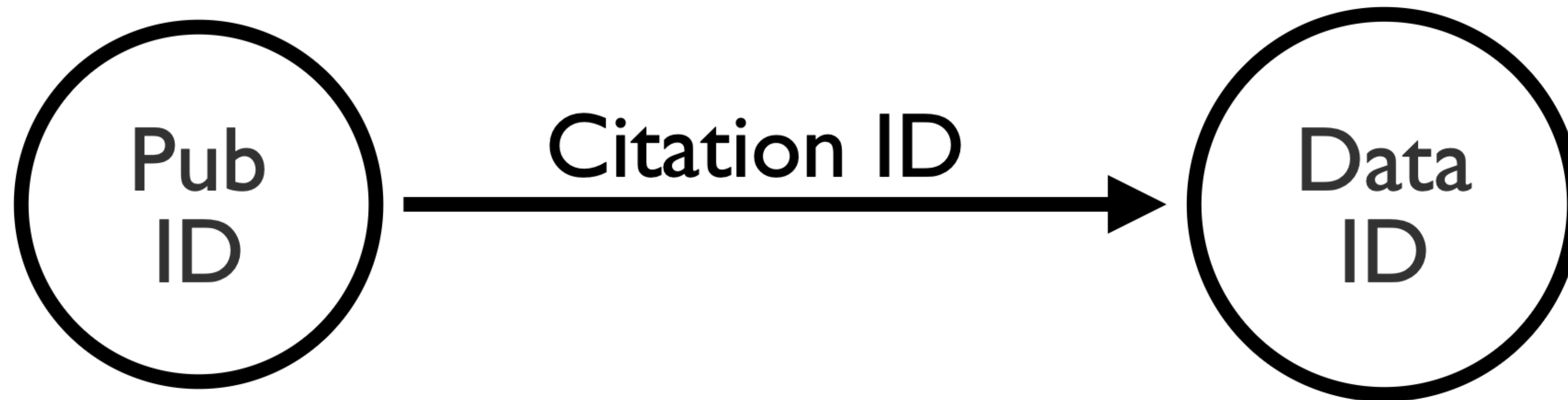
Sec. 101. Federal evidence-building activities.

TITLE I—FEDERAL EVIDENCE-BUILDING ACTIVITIES

The ADRF project was cited as an example of best practice in a report that led to the US Open Data Act, so the ADRF team have a great track record here!



Evidence Based



A key to making policy decisions evidence based is being able to see how the research ties to the data.

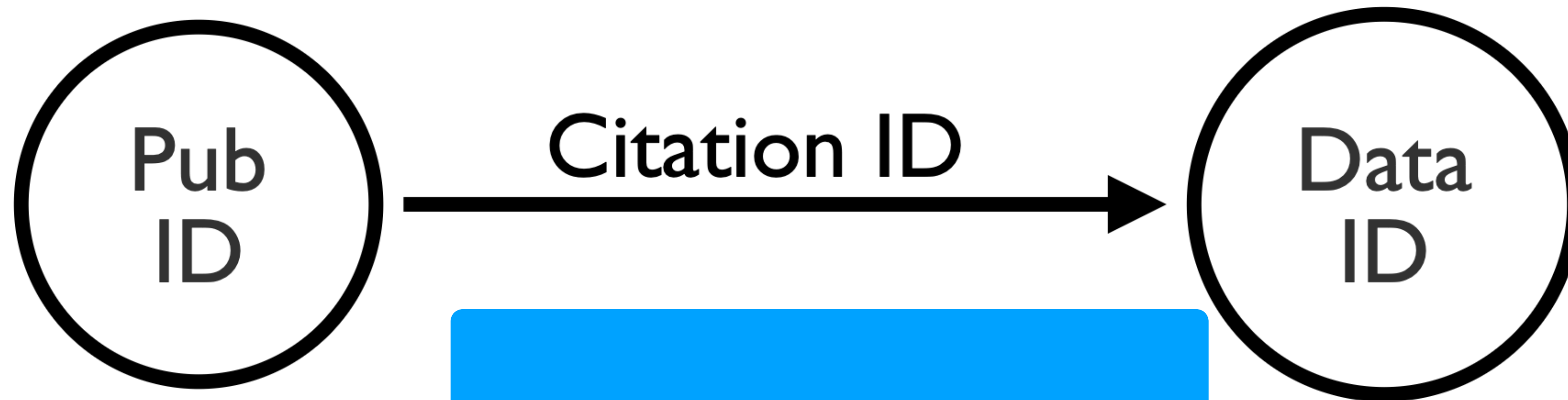
Larger Vision



If we could make these links, and link to the users of the data and papers, we could get to a virtuous circle of connections.

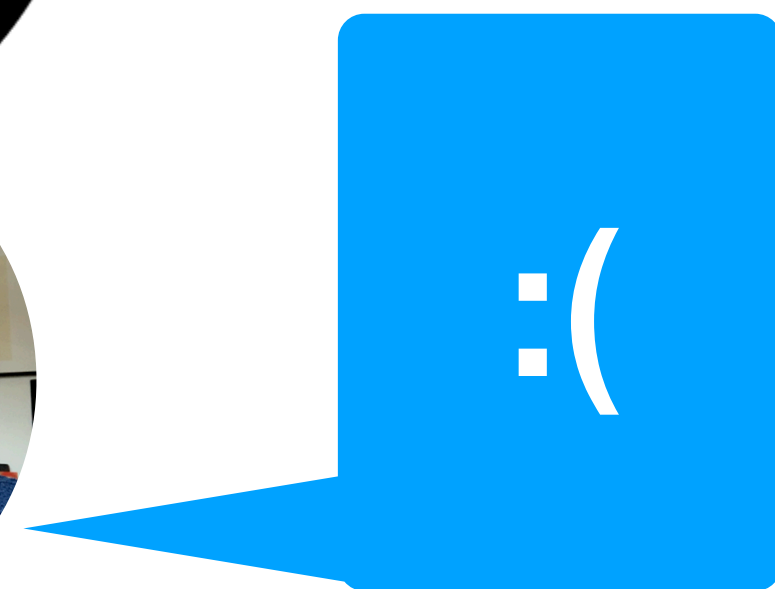
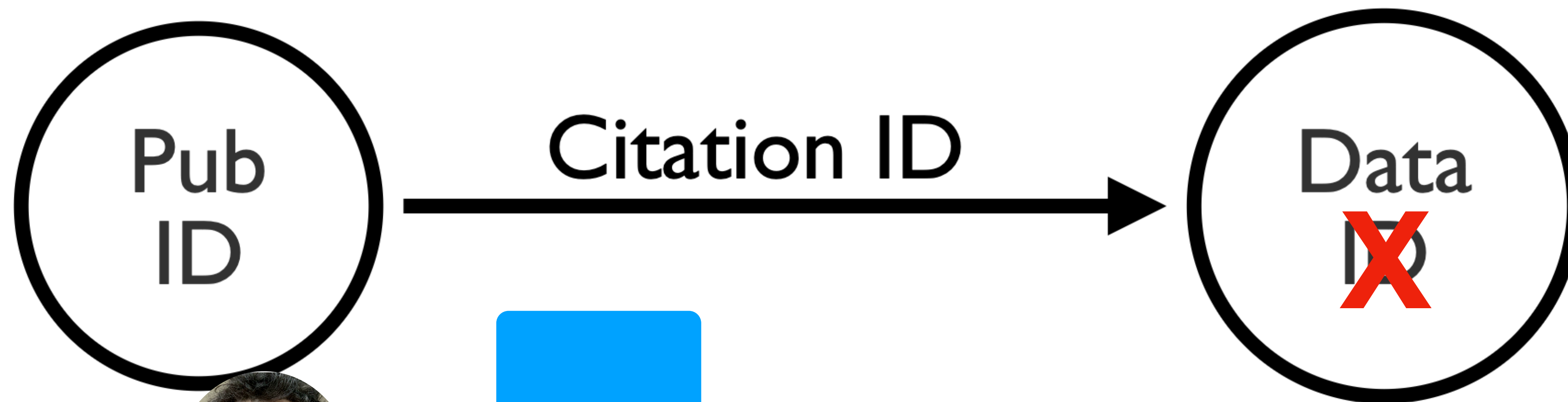


Evidence Based



But in the social sciences, there is one big nasty elephant in the room!

Evidence Based



No Persistent Identifiers :(

AllenNLP framework, neural networks

NER, Topic Modelling, AllenNLP/Biome Framework

LSTM, CNN

bi-LSTM neural network, CRF model

TensorFlow, Perceptron, RNN, Keras, Pyspark, SpaCy

Machine

Logistic regression & SVM

CNN-LSTM-CRF, Bi-LSTM Coreference Model

CRF

Learning

LightGBM model

Bi-Directional Attention Flow, CNN

CNN, RNN/RCNN

CRF, RNN/LSTM



But that's OK, because we can draw on the work, CRF model
magic of AI, right, right?

bi-LSTM neural network, tagging

Rich Context Competition

September 2018 — February 2019

\$2000 to each team to pass phase 1

\$20,000 to the winner of phase 2

All outputs had to be made openly available

How do you get up to speed with machine learning from a standing start?

The Rich context team decided to run a competition and incentivise ML experts to help solve their problem for them.

And you know what, it mostly worked! They got 20 teams to participate and some partial solutions to the problem.

More importantly, this approach to drive collaboration really worked, and is being carried on later this year!

The rest of this deck is about that comp!



Goal:

Build a machine learning model that can identify **datasets** referenced in a paper, along with the paper's **research fields**, and **methods** used in the paper.



This is what we wanted the teams to create machine learning models to do for us.

Phase 1

Training Corpus

2500 papers **with references** to data

2500 papers **with no references** to data

Test Corpus

2500 papers **with references** to data

2500 papers **with no references** to data

In the first phase teams had 5k papers to train on, and we compared their models against 5k papers that were held back



2500 papers **with references** to data

Bundesbank
ICPSR catalog

Data annotations manually created using NYU built tool

Full text & methods labels, provided by SAGE



This is one of the bits where SAGE helped, we provided training data, along with Bundesbank and ICPSR.

Status Messages

- Article data successfully saved!
- Loaded article 5 coding for user agordon

- [Return to DataSetCitation mention coding List](#)

Code DataSetCitation mentions

DataSetCitation ID:

Article 5 - Explaining the recent decline in cocaine use among young adults: Further evidence that perceived risks and disapproval lead to reduced drug use

- Publication Date: Jan. 1, 1990
- Original PDF: [2137171.pdf](#)

Explaining the Recent Decline in Cocaine Use among Young Adults: Further Evidence That

Perceived Risks and Disapproval Lead to Reduced Drug Use

Author(s): Jerald G. Bachman, Lloyd D. Johnston and Patrick M. O'Malley

Source: Journal of Health and Social Behavior, Vol. 31, No. 2 (Jun., 1990), pp. 173-184

Published by: American Sociological Association

Stable URL: <https://www.jstor.org/stable/2137171>

Accessed: 06-08-2018 16:15 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide

range of content in a trusted digital archive. We use information technology and tools to increase productivity and

facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

Record Data Set Mentions

Mention ==> :

|

Mention List

Monitoring the Future: Questionnaire Responses	3017	<input type="button" value="Remove"/>
"Monitoring the Future" project	3016	<input type="button" value="Remove"/>
"Monitoring the Future" series	3015	<input type="button" value="Remove"/>
Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth	3014	<input type="button" value="Remove"/>

Find in Article Text

Search for:

Data Set Info

Data Set 54 - Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1980

• Title: Monitoring the Future: A Continuing

The ICPRS data only listed papers that mentioned their data. To make things more useful for training we had some brave people go in and tag exactly where in the papers the data sets were mentioned. They did this using a tool that the team at NYU built. (One of the interesting things to observe from the outside is that the whole process involved building a lot of tools!).



Emily Wiegand, Neil Miller and Jenna Chapman from Chapin Hall at the University of Chicago, Mengxuan Zhao, Marcos Ynoa and Ekaterina Levitskaya from the CUNY Graduate Center, Computational Linguistics program

Successfully Saved article (note green box Status Messages)



Here you can see how one data set is referred in lots of different ways within one paper.

```
{  
  "citation_id": 1959,  
  "publication_id": 109,  
  "data_set_id": 322,  
  "mention_list": [  
    "Deutsche Bundesbank's Securities Holdings Statistics",  
    "Microdatabase Securities Holdings Statistics",  
    "Securities Holdings Statistics",  
    "Securities holdings statistics"  
  ],  
  "score": 1.0  
},
```


Phase 2

5000 unlabelled publications

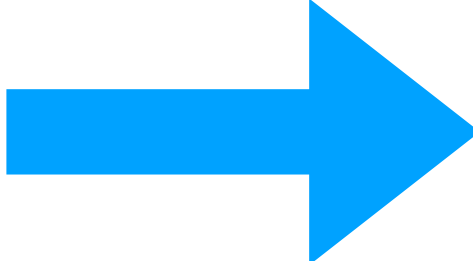
Teams had to discover datasets from the first phase's data catalog

research methods and fields

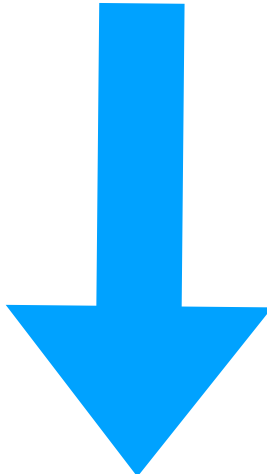
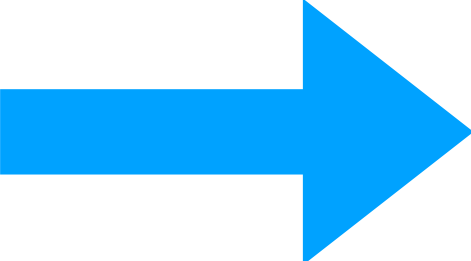


In the second phase they just had to run over a corpus of 5K new papers.

Submissions



config.sh



The team in NY helped entrants take their code and dockerise it, and those docker containers were uploaded and run by the NY team.

Judging

Confusion matrix was generated comparing team predictions with our data on our hold-outs.

10 random papers per set of disciplines were manually evaluated for accuracy of methods and field suggestions



This is how we judged the teams.

ADR | Data Stewardship

rc.adrf.cloud.s3-website-us-east-1.amazonaws.com/rc/qualitative-review

ADR | Data Stewardship

Navigation

- Rich Context
- Submission Review
- Qualitative Review **NEW**
- Projects
- Explore Data
- User Directory
- Data Steward **NEW**
- Bookmarks

Submissions Qualitative Review


RichContext / Qualitative Review

Information

Judge: jmorgan Status: TODO

Mentions 60 Fields 60 Methods 60

Publication	Team 1	Team 2	Team 3	Team 4
Banking on the Margin in Canada PDF	-1 0 +1 Empty	-1 0 +1 1999 Statistics Canada Survey of Financial Security	-1 0 +1 Empty	-1 0 +1 Economic Development Quarterly Volume 22 Number Social Sciences and Humanities Research Council
Estimating Household Water Use: A Comparison of Diary, Prompted Recall, and Free Recall Methods PDF	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 DATA COLLECTION
Houses of Healing PDF	-1 0 +1 HITS	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 quantitative scales prison
Higher, Faster, Louder: Representations of the International Music Competition PDF	-1 0 +1 SAGE	-1 0 +1 Washington Post	-1 0 +1 the Cold War	-1 0 +1 prestige organizing committees Van Cliburn International Piano Competition
Providing reviews of evidence to COPD patients: controlled prospective 12-month trial PDF	-1 0 +1 Empty	-1 0 +1 Medical Interview Satisfaction Scale	-1 0 +1 Empty	-1 0 +1 TQEH Research Foundation
Commission Versus Receipt of Violence During Pregnancy: Associations With Substance Abuse Variables PDF	-1 0 +1 Pacific Islands Families Study CTS	-1 0 +1 Empty	-1 0 +1 Violence	-1 0 +1 three violence subtypes b Pearson s chi square anal National Family Violence S
The Punitive Turn in Social Policies: Critical Race Feminist Reflections on the USA, Great Britain, and Beyond PDF	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty
Economic Vulnerability among Low-Educated Europeans PDF	-1 0 +1 IALS ESS registers European Social Survey The European Social Survey ACTA	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 IALS European Social Surveys ESS ESS



Judging interface

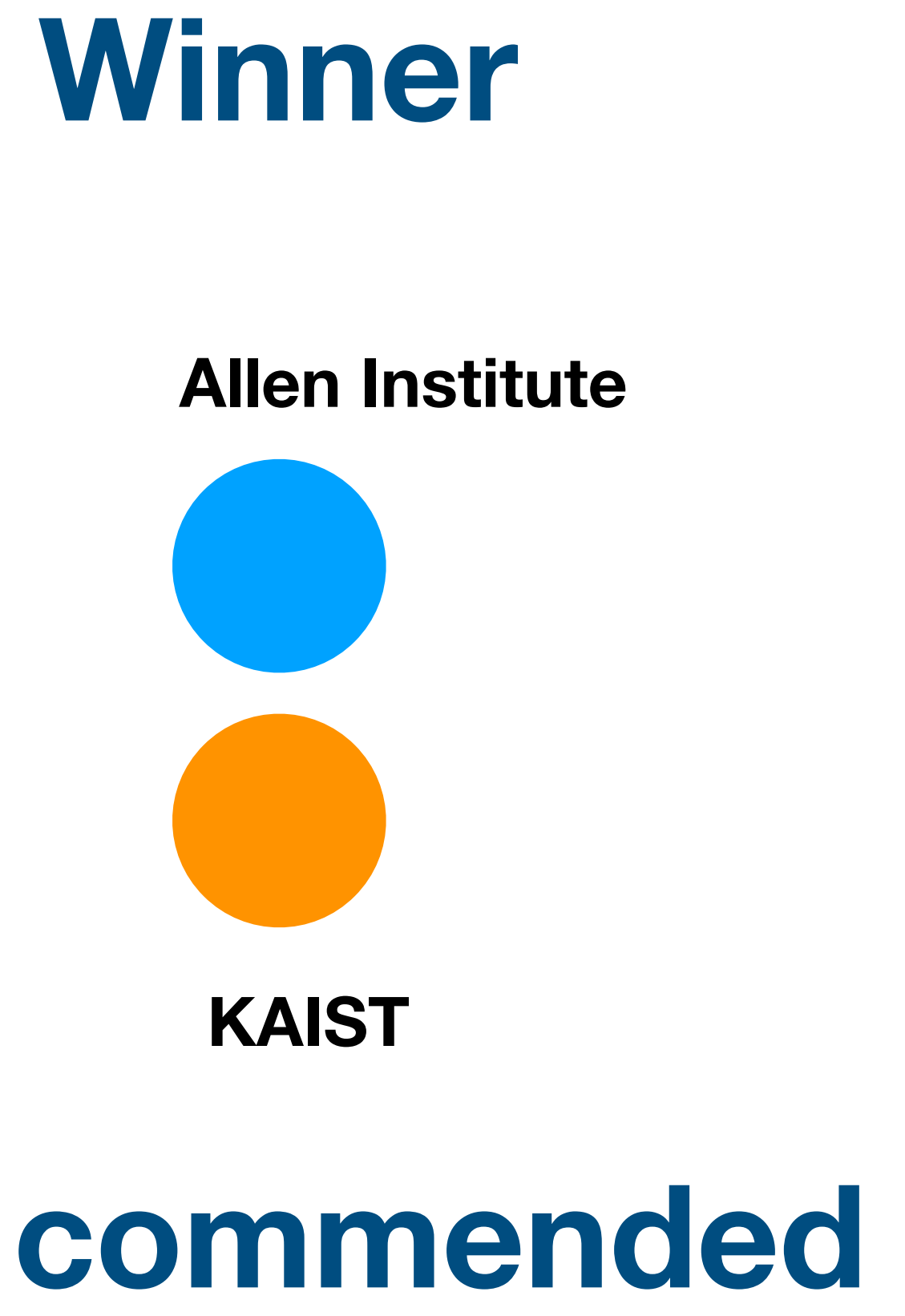
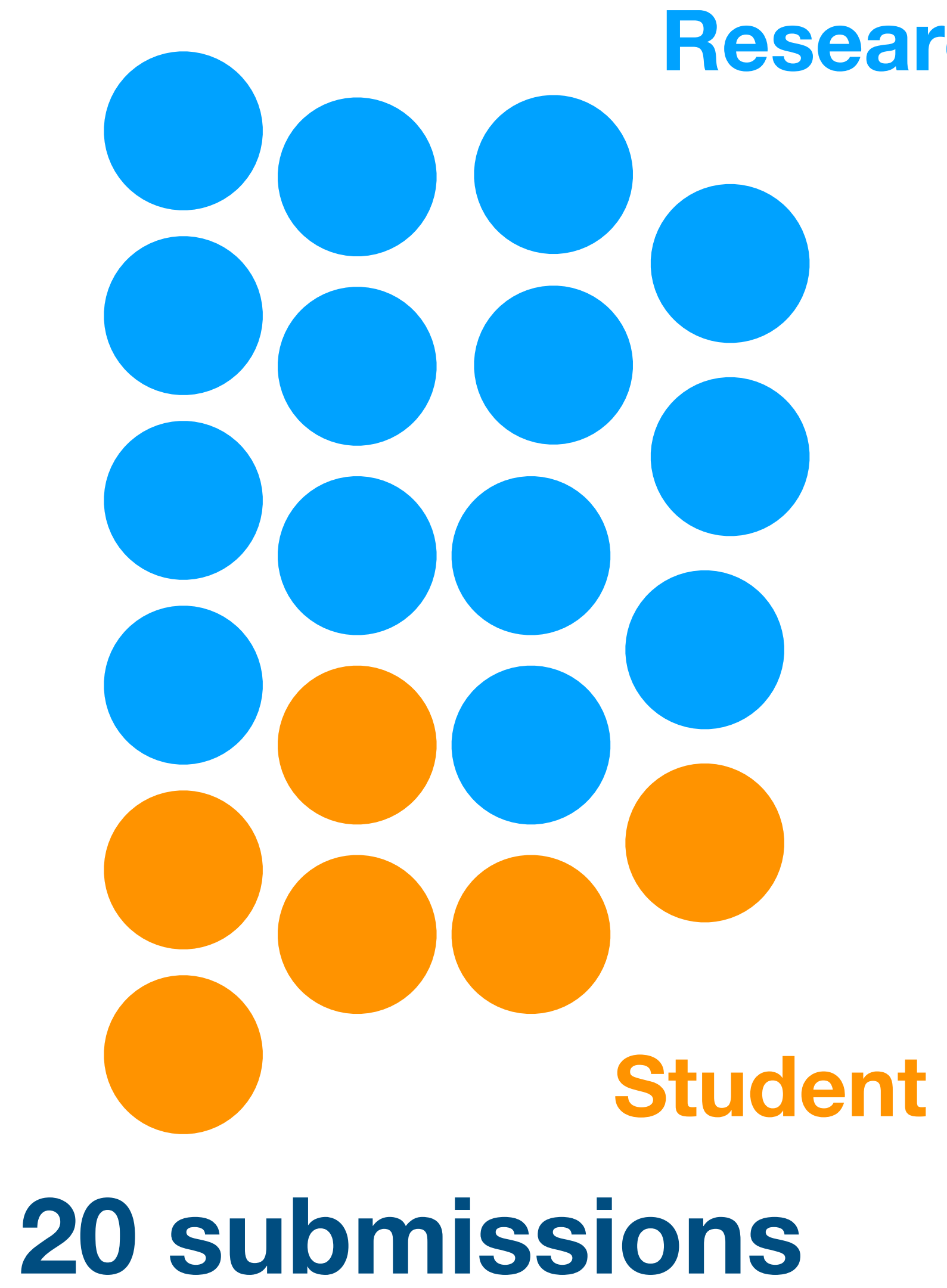
Another tool was built to help the expert judges get to consistent voting across all the entries.

Participation and results



There was global participation in the comp!

Participation and results



Nicely there were both senior and undergraduate teams taking place.

Sheet1

team	phase 1 model			phase 2 model					
	holdout 1 precision	holdout 1 recall	holdout 1 F1	holdout 1 precision	holdout 1 recall	holdout 1 F1	holdout 2 precision	holdout 2 recall	holdout 2 F1
rcc-03	0.54399	0.1336	0.21451	0.35726	0.19635	0.25342	0.28623	0.1875	0.25455
rcc-05	0.21003	0.09564	0.11365	0.1054	0.08944	0.09677	0.52272	0.20535	0.29487
rcc-14	0.21051	0.1475	0.17346	0.29609	0.1195	0.17028	0.28623	0.07143	0.11765
rcc-17	0.17093	0.16197	0.16633	N/A	N/A	N/A	0.46988	0.34821	0.4

Best models identified about 1/2 of the data sets

Pitfalls



There were a few issues that we identified with how the training data was created, and the entries run

Annotators of the training data were not able to find all mentions in papers in the ICPRS data set.

Teams identified data set references that were not tagged in the training data

Running competition entries for judging required a lot of effort and back and forth with the team

Outputs from this phase

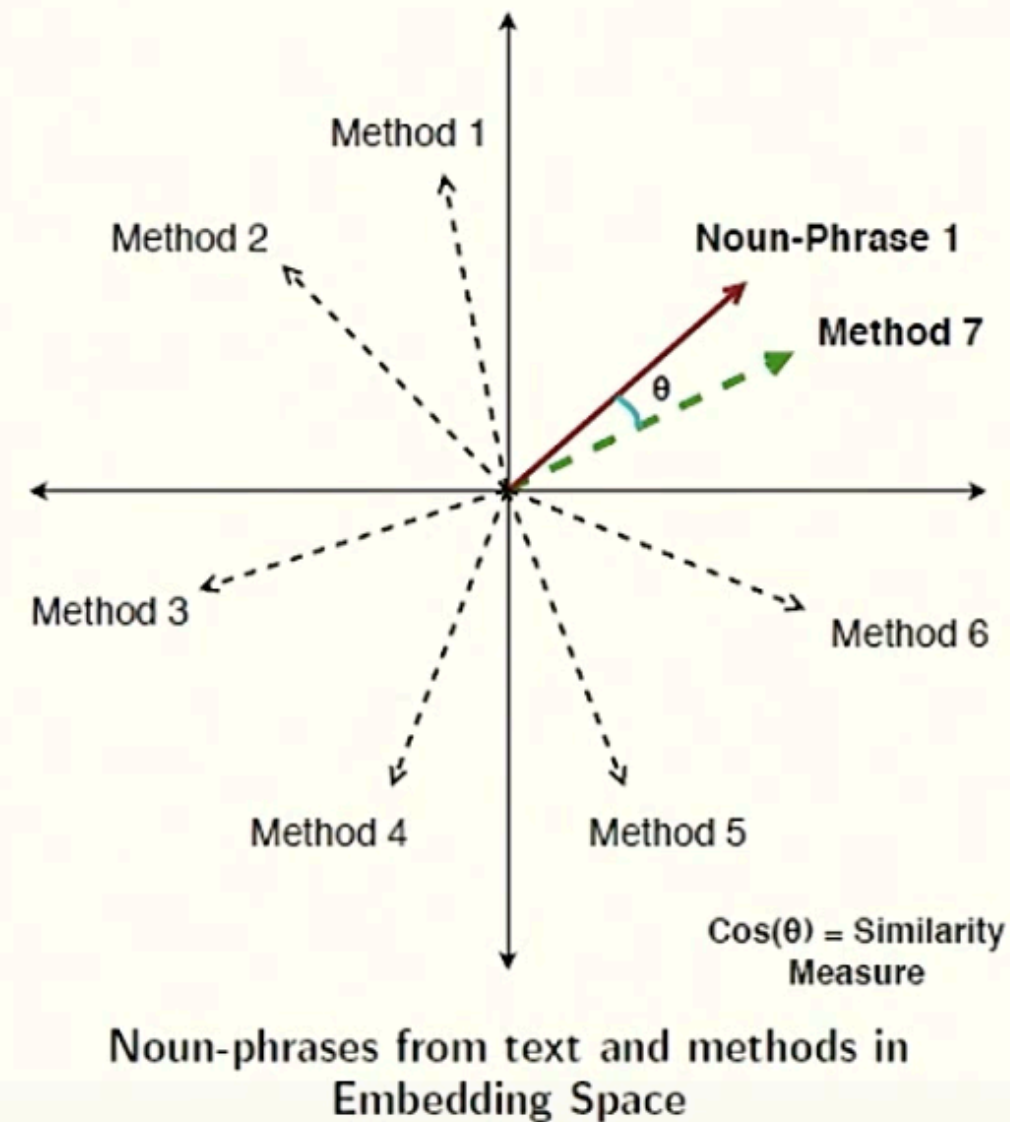


YouTube^{GB}

Search

Identification of Research Methods and Fields

Methods Identification - Word2Vec



- Closest research method vector found by measuring cosine similarity between noun phrase vectors and method vectors

Click on Tools, Sign, and Comment to access additional features.

February workshop

1:42:08 / 3:19:52

Data Science Group

DICE @ Rich Context Competition

February



<https://www.youtube.com/watch?v=PE3nFrEkwoU>

Outputs from this phase

Rich Search and Discovery for Research Datasets

Navigation

Contents:

Chapter 1: Introduction
Chapter 2: Who's Waldo: Conceptual issues when characterizing data in empirical research
Chapter 3: Enriching

Rich Search and Discovery for Research Datasets: *Building the next generation of scholarly infrastructure*

Julia Lane, Ian Mulvany, Paco Nathan (eds.)

Copyright 2019 NYU.

Contents:

- Chapter 1: Introduction
- Chapter 2: Who's Waldo: Conceptual issues when characterizing data in empirical research
- Chapter 3: Enriching context and enhancing engagement around datasets
- Chapter 4: Metadata for Administrative and Social Science Data
- Chapter 5: Competition Design
- Chapter 6: Allen Institute for Artificial Intelligence (AI2)
- Chapter 7: KAIST
- Chapter 8: GESIS
- Chapter 9: DICE

Outputs from this phase

November 2019 workshop



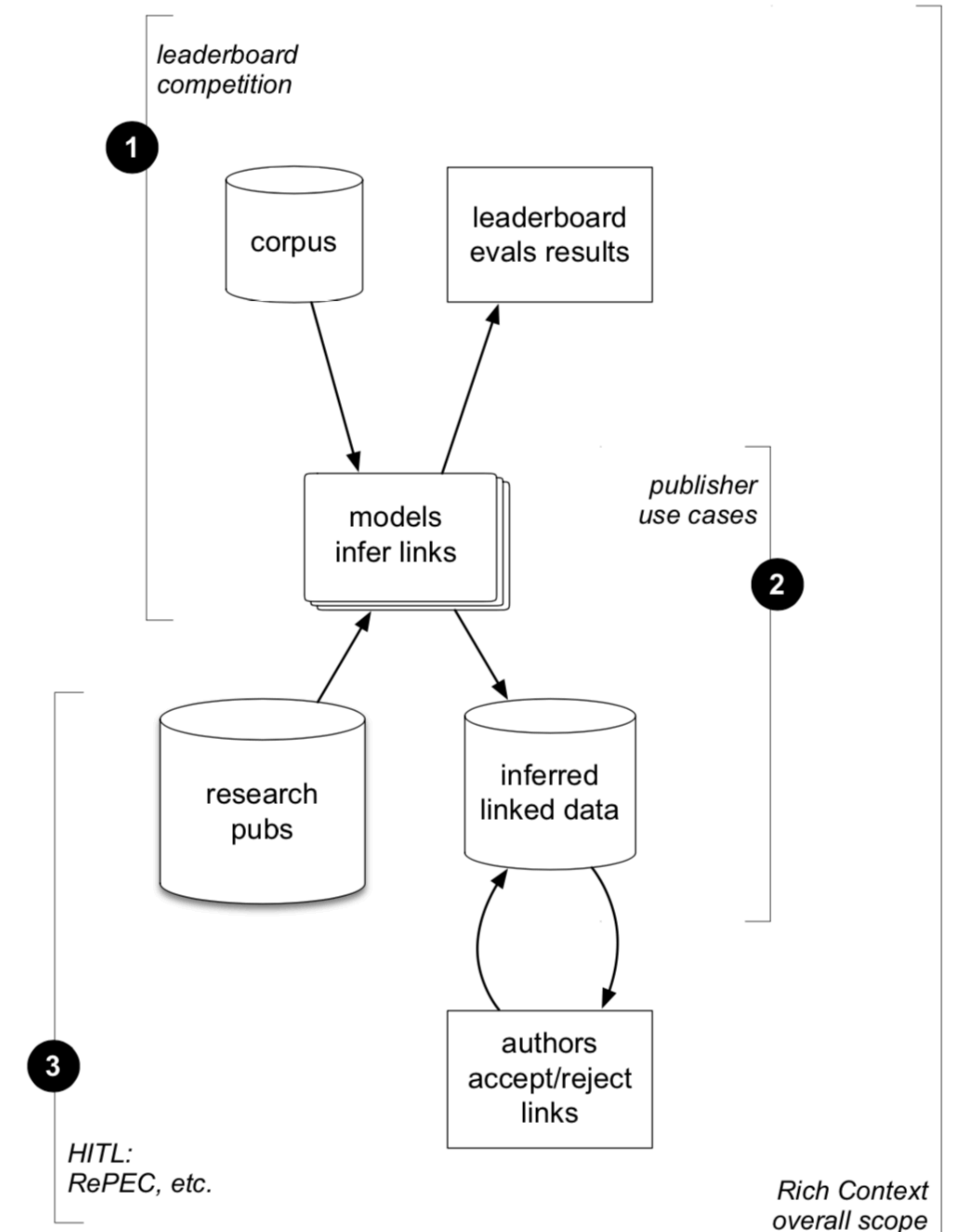
The key thing, though, is that we got to a minimally viable research output, with enough momentum to keep going, and in Nov 2019 there is going to be a workshop to help set the future direction of the initiative.

- **goal #1.** Identify compelling use cases that would be transformed by access to dataset search and discovery tools (starting from Evidence-Based Policymaking)
- **goal #2.** Take stock of existing practices
- **goal #3.** Catalyze a community that works together to integrate open source projects for common needs in data/metadata infrastructure (JupyterLab, spaCy, PyTorch rdflib, Egeria, W3C standards for metadata, Amundsen and its emerging category, etc.)
- **goal #4.** Identify where we need centralized services (e.g., a global repository of datasets, having persistent identifiers) to complete the knowledge lifecycle
- **goal #5.** Define a platform (akin to Amazon, Etsy, LinkedIn) for the initial use cases, which can be broadly adopted:
- **goal #6.** Generate business model(s) that can be seeded with initial research-funding support and subsequently become self-sustaining.



Comp Phase 2

- collaboration with SAGE Pub, Digital Science, RePEc, etc.; partnering with Bundesbank (EU)
- knowledge graph **vocabulary** integrates W3C metadata standards: DCAT, PAV, DCMI, CITO, FaBiO, FOAF, etc.
- **data as a strategic asset:** knowledge graph produces an open corpus for the **leaderboard competition**
- **human-in-the-loop AI** used to infer metadata then confirm with authors via **RePEC**, etc.
- adjacent work: graph embedding, meta-learning, persistent identifiers, reproducible research



Leaderboard examples



We are also going to run the competition again, and this time will take lessons from the machine learning community and run it as an open leaderboard.

[View on GitHub](#)

NLP-progress

Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

Tracking Progress in Natural Language Processing

Table of contents

English

- [Automatic speech recognition](#)
- [CCG](#)
- [Common sense](#)
- [Constituency parsing](#)
- [Coreference resolution](#)
- [Dependency parsing](#)
- [Dialogue](#)
- [Domain adaptation](#)

NLP Progress



Search for papers, code and tasks

[Browse state-of-the-art](#)

[Follow](#)

Browse state-of-the-art

1442 leaderboards • 1307 tasks • 1307 datasets • 16375 papers with code

[Follow on Twitter for updates](#)

Computer Vision

Semantic Segmentation 32 leaderboards 614 papers with code	Image Classification 51 leaderboards 517 papers with code	Object Detection 52 leaderboards 430 papers with code	Image Generation 48 leaderboards 209 papers with code
---	--	--	--

[See all 694 tasks](#)

Natural Language Processing

Machine Translation 43 leaderboards 475 papers with code	Language Modelling 8 leaderboards 396 papers with code	Question Answering 42 leaderboards 386 papers with code	Sentiment Analysis 21 leaderboards 311 papers with code
---	---	--	--

Papers with code

New submission process



We are also going to use a different infrastructure for running the competition entries, one we hope will lead to more success in being able to run entries.



GitHub

open by
default



Jupyter

reproducible
by default

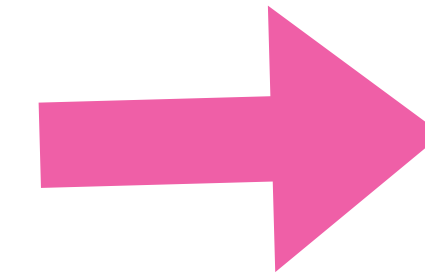


Binder

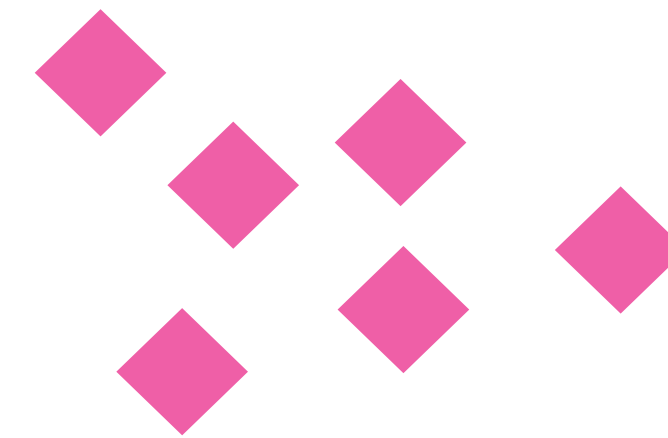
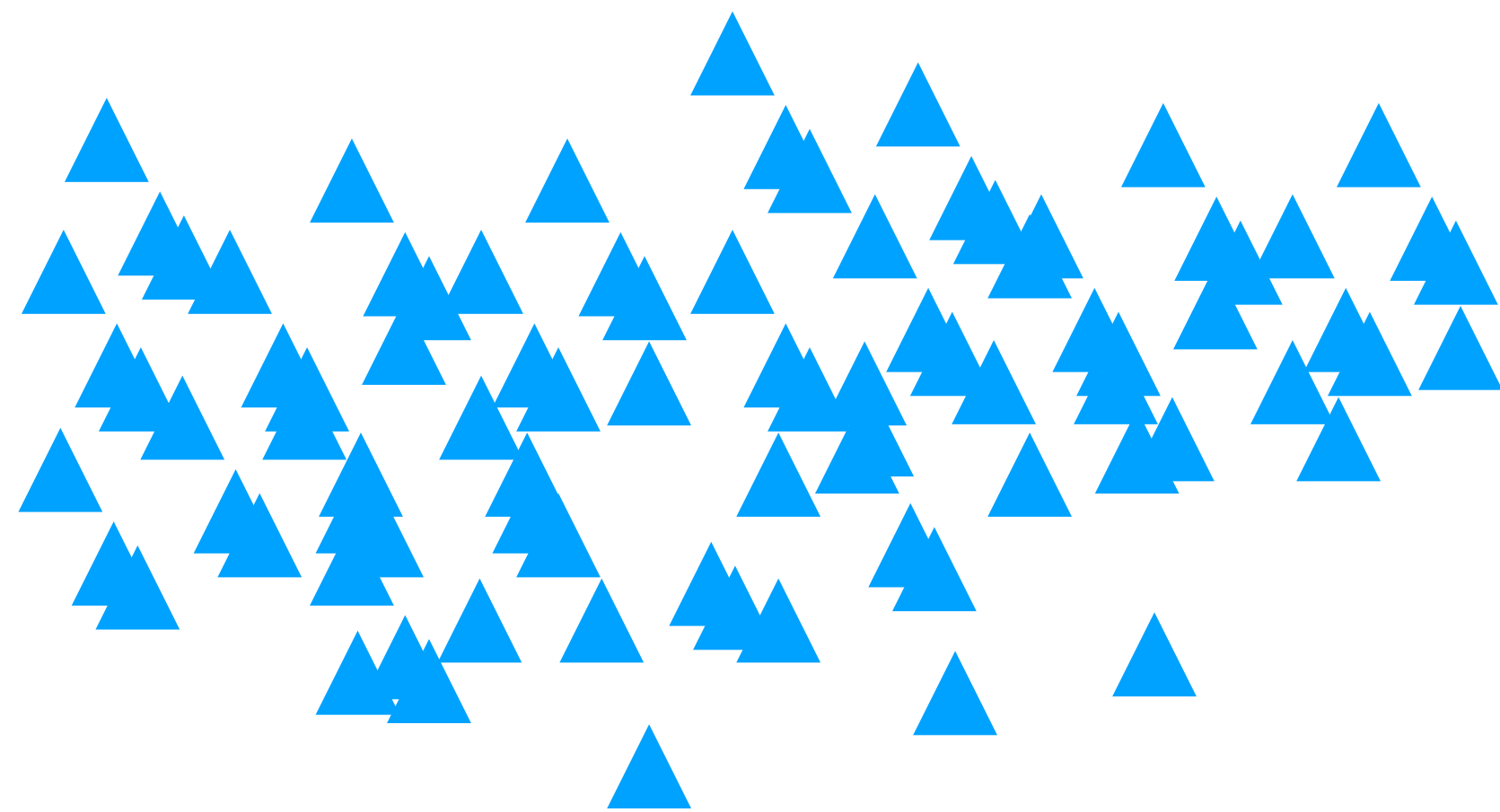
Scalable by
default

New training protocol

Model generated



Lots of Training data



Small data set

We hope that adoption of transfer learning can help get teams to better models more quickly



Normal Machine Learning

Transfer learning

New training protocol

Move from 5000 training papers to 500

=> Easier to robustly tag

While using fewer training examples to get there.



Rich Context Leaderboard

Leaderboard 1

Entity Linking for Datasets in Publications

The first challenge is to identify the datasets used in research publications, initially focused on the problem of *entity linking*. Research papers will generally mention the datasets they've used, although there are no formal means to describe that metadata in a machine-readable way.

Identifying dataset mentions typically requires:

- extracting text from an open access PDF
- some NLP parsing of the text
- feature engineering (e.g., paying attention to where the text is located in the paper)
- modeling to identify up to 5 datasets per publication

See *Evaluating Models for Entity Linking with Datasets* for details about how the Top5uptoD leaderboard metric is calculated.

Current SOTA

source	precision	repo	version	date	contact
LARC	78.36	link	v0.1.5	2019-09-26	@philipskokoh

**First entry
has
identified
about 3/4s of
the data sets**

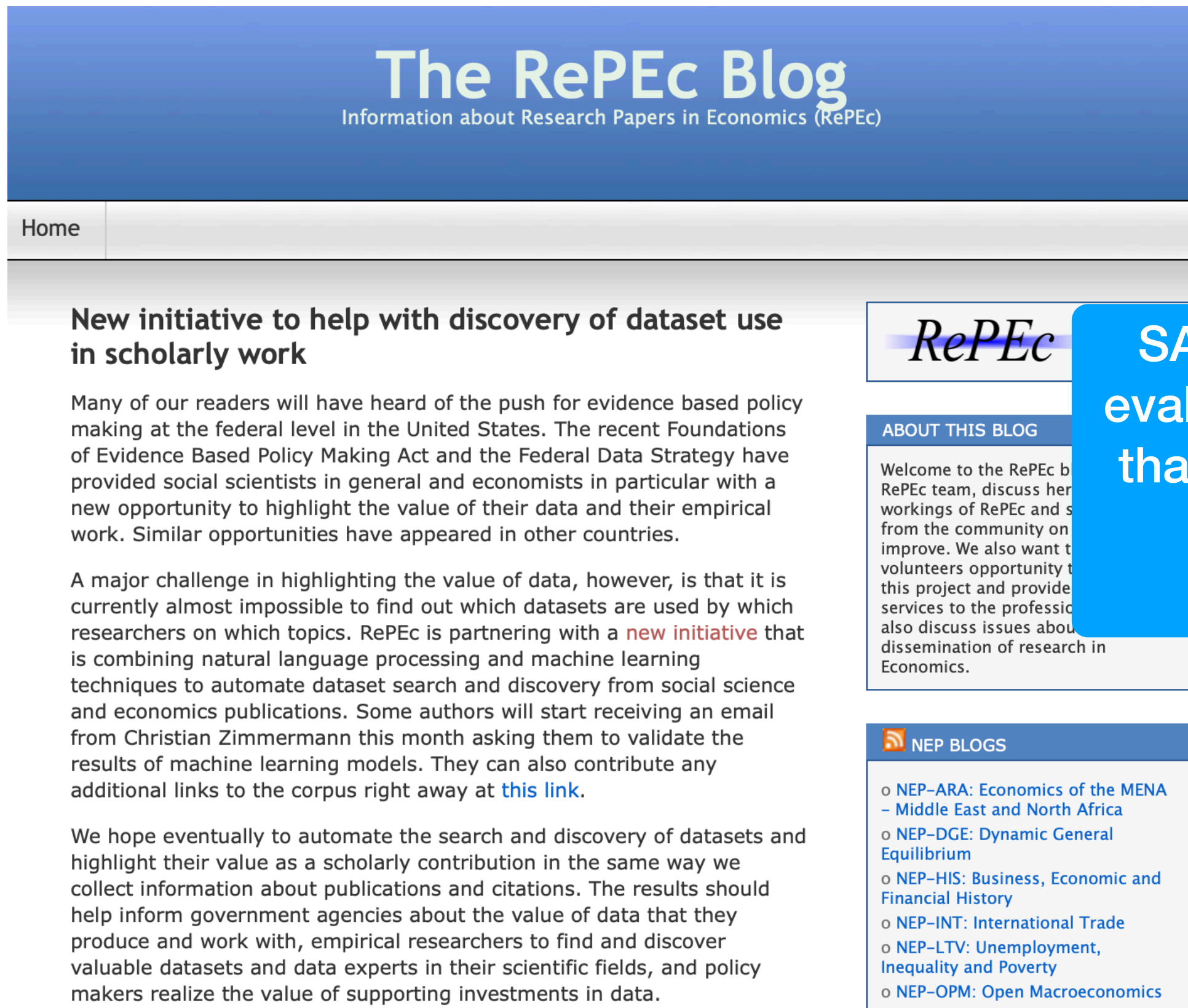
Initial trials of this have been promising.



HITL pattern for refining the data

RePeC

SAGE



The RePEc Blog
Information about Research Papers in Economics (RePEc)

Home

New initiative to help with discovery of dataset use in scholarly work

Many of our readers will have heard of the push for evidence based policy making at the federal level in the United States. The recent Foundations of Evidence Based Policy Making Act and the Federal Data Strategy have provided social scientists in general and economists in particular with a new opportunity to highlight the value of their data and their empirical work. Similar opportunities have appeared in other countries.

A major challenge in highlighting the value of data, however, is that it is currently almost impossible to find out which datasets are used by which researchers on which topics. RePEc is partnering with a **new initiative** that is combining natural language processing and machine learning techniques to automate dataset search and discovery from social science and economics publications. Some authors will start receiving an email from Christian Zimmermann this month asking them to validate the results of machine learning models. They can also contribute any additional links to the corpus right away at [this link](#).

We hope eventually to automate the search and discovery of datasets and highlight their value as a scholarly contribution in the same way we collect information about publications and citations. The results should help inform government agencies about the value of data that they produce and work with, empirical researchers to find and discover valuable datasets and data experts in their scientific fields, and policy makers realize the value of supporting investments in data.

RePEc

ABOUT THIS BLOG

Welcome to the RePEc blog. The RePEc team, discuss here the workings of RePEc and share with you from the community on how to improve. We also want to provide you with a volunteer's opportunity to help this project and provide services to the professional community. We also discuss issues about the dissemination of research in Economics.

NEP BLOGS

- o NEP-ARA: Economics of the MENA - Middle East and North Africa
- o NEP-DGE: Dynamic General Equilibrium
- o NEP-HIS: Business, Economic and Financial History
- o NEP-INT: International Trade
- o NEP-LTV: Unemployment, Inequality and Poverty
- o NEP-OPM: Open Macroeconomics

SAGE and RePeC are going to help evaluate the outputs of models, and in that way allow models to be updated quickly.



Economic Development Quarterly
Education and Urban Society
Education, Citizenship and Social Justice
Educational and Psychological Measurement
Educational Policy
Accounting, Auditing & Finance
Adult and Continuing Education
Journal of Educational Computing Research
Journal of Educational Technology Systems
Public Finance Review
Research in Education
Urban Education

Creating the data infrastructure

Work has started on defining the schema for the data link repository



Semantic underpinning of vocabulary - CITO!



ceteri TTL to use as an example for JupyterLab metadata service

1 contributor

89 lines (76 sloc) | 3.27 KB

Raw

Blame

His

```
1 @base <https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary> .
2
3 @prefix cito: <http://purl.org/spar/cito/> .
4 @prefix dbo: <http://dbpedia.org/ontology/> .
5 @prefix dbr: <http://dbpedia.org/resource/> .
6 @prefix dcat: <http://www.w3.org/ns/dcat/> .
7 @prefix dct: <http://purl.org/dc/terms/> .
8 @prefix dcterms: <http://purl.org/dc/terms/> .
9 @prefix dctypes: <http://purl.org/dc/dctypes/> .
10 @prefix fabio: <http://purl.org/spar/fabio/> .
11 @prefix foaf: <http://xmlns.com/foaf/0.1/> .
12 @prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
13 @prefix owl: <http://www.w3.org/2002/07/owl#> .
14 @prefix pav: <http://purl.org/pav/> .
15 @prefix prism: <http://prismstandard.org/namespaces/c/2.0/> .
16 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
17 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
18 @prefix schema: <http://schema.org/> .
```

And for the Force11 community I'm happy to say that CITO is going to play a role here!



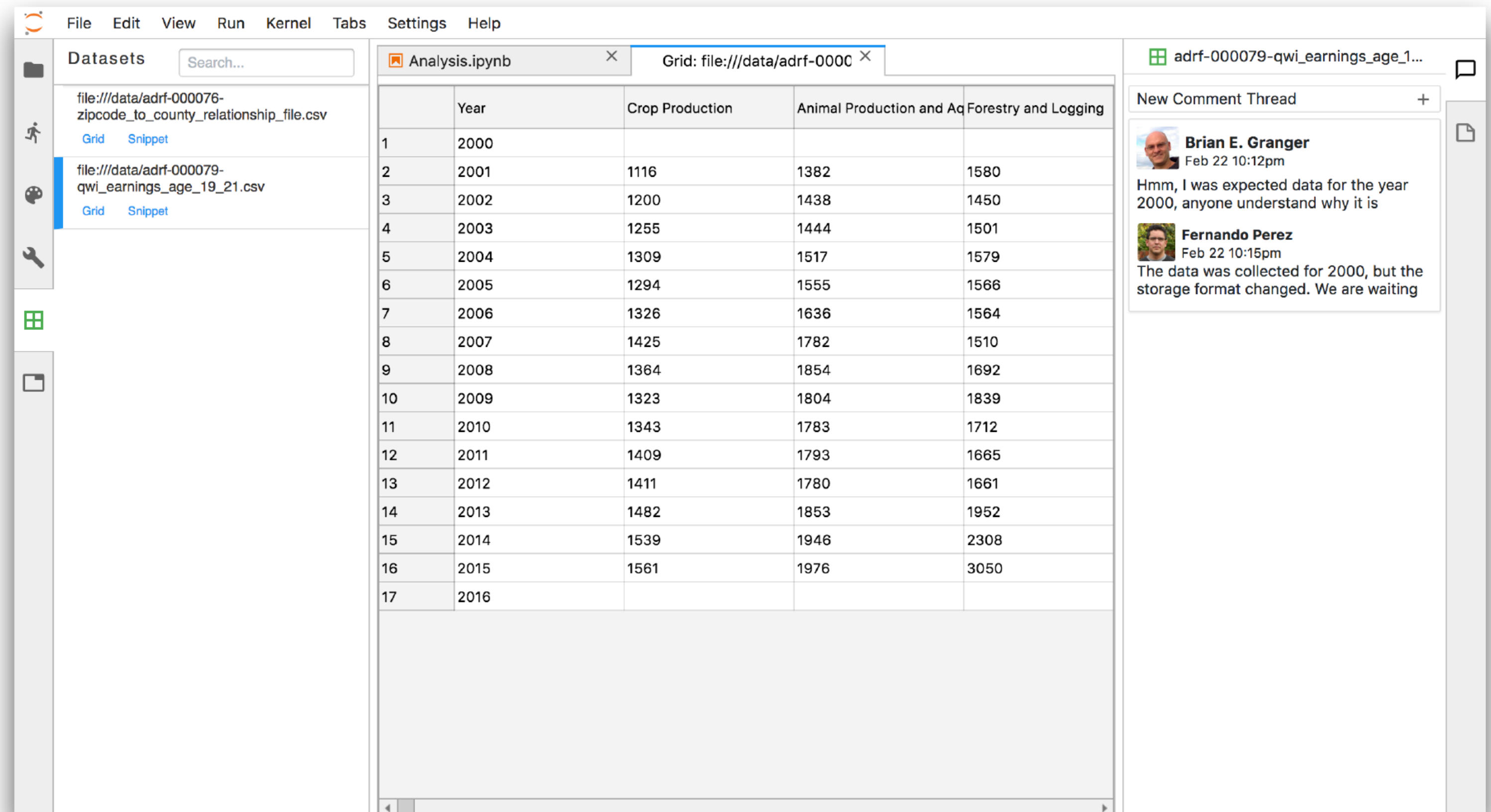
Integrating directly with Jupyter

Commenting and Annotation in JupyterLab

The Jupiter project has a grant as part of this project to make their data explorer connect with this piece of infrastructure.



Data browser



The screenshot displays the JupyterLab interface. On the left is the 'Datasets' panel with a search bar and two files listed: 'file:///data/adrf-000076-zipcode_to_county_relationship_file.csv' and 'file:///data/adrf-000079-qwi_earnings_age_19_21.csv'. The central notebook, 'Analysis.ipynb', shows a data grid with the following table:

	Year	Crop Production	Animal Production and Aq	Forestry and Logging
1	2000			
2	2001	1116	1382	1580
3	2002	1200	1438	1450
4	2003	1255	1444	1501
5	2004	1309	1517	1579
6	2005	1294	1555	1566
7	2006	1326	1636	1564
8	2007	1425	1782	1510
9	2008	1364	1854	1692
10	2009	1323	1804	1839
11	2010	1343	1783	1712
12	2011	1409	1793	1665
13	2012	1411	1780	1661
14	2013	1482	1853	1952
15	2014	1539	1946	2308
16	2015	1561	1976	3050
17	2016			

On the right, a comment thread is visible for the dataset 'adrf-000079-qwi_earnings_age_1...'. It includes a 'New Comment Thread' button and two comments:

- Brian E. Granger** (Feb 22 10:12pm): Hmm, I was expected data for the year 2000, anyone understand why it is
- Fernando Perez** (Feb 22 10:15pm): The data was collected for 2000, but the storage format changed. We are waiting

* Early prototype

Summary

- Comp format engaged ML community
- Growing this community: like the book, second comp, workshop, partnerships like SAGE, RePeC
- The bet is that is full workflow integration will be crucial
- Aim is to streamline second version of the comp
- Focus on Nov workshop is to find the right direction, and build on existing work and knowledge

How to participate

<https://github.com/Coleridge-Initiative/rclc/wiki/How-To-Participate>



If you would like to get involved, we would love to hear from you!

Thank you!

A postscript - tools and behaviours

While I was putting these slides together for the Force11 conference I was mulling over the kinds of capabilities machine learning allows us



NEED

TOOL



CAPABILITY



**THIS PART
FITS THE
PROBLEM**



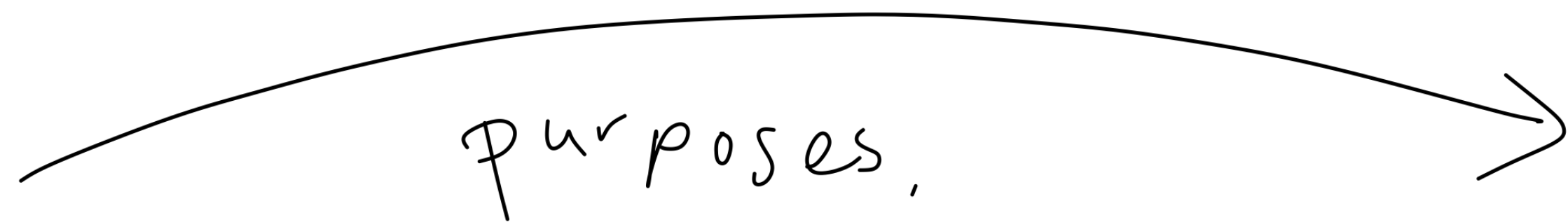
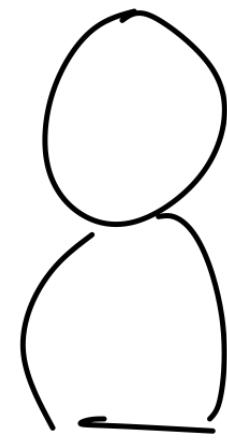
**THIS PART
FITS THE
PERSON**



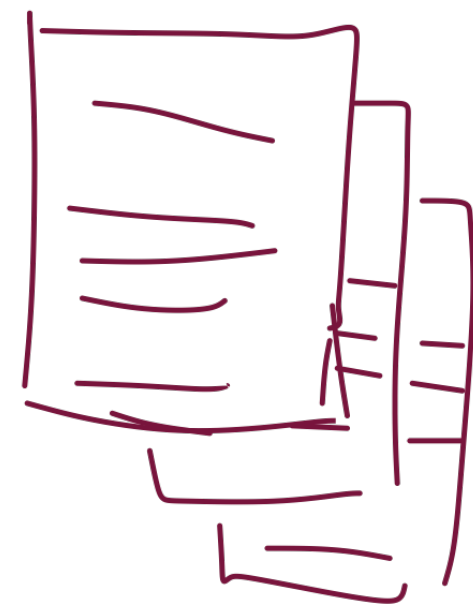
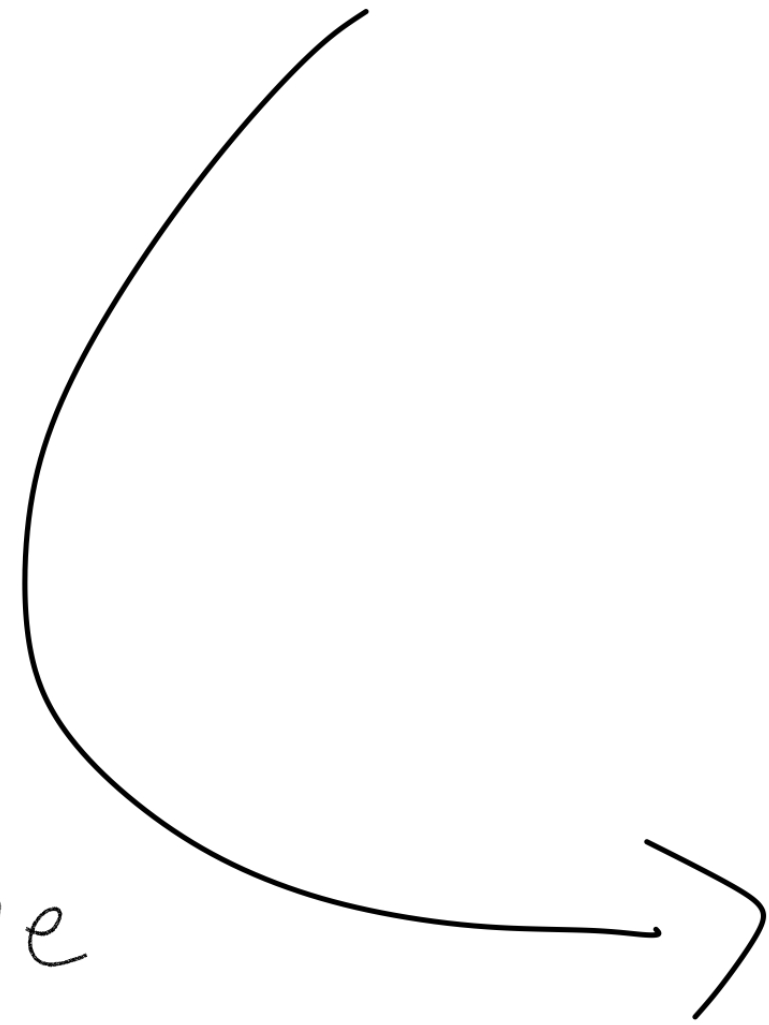
Brett Victor demonstrates brilliantly one way of thinking about tooling.



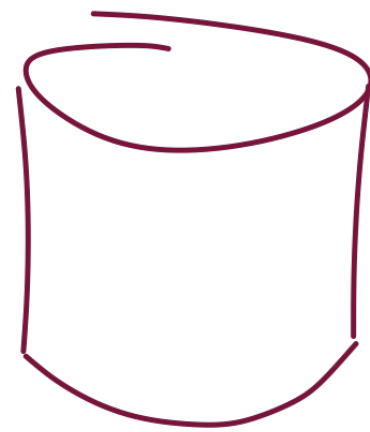
via Brett Victor



Some tools



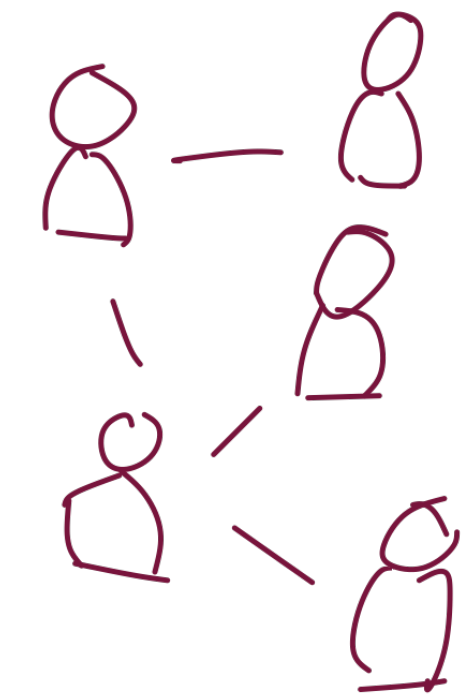
papers



data

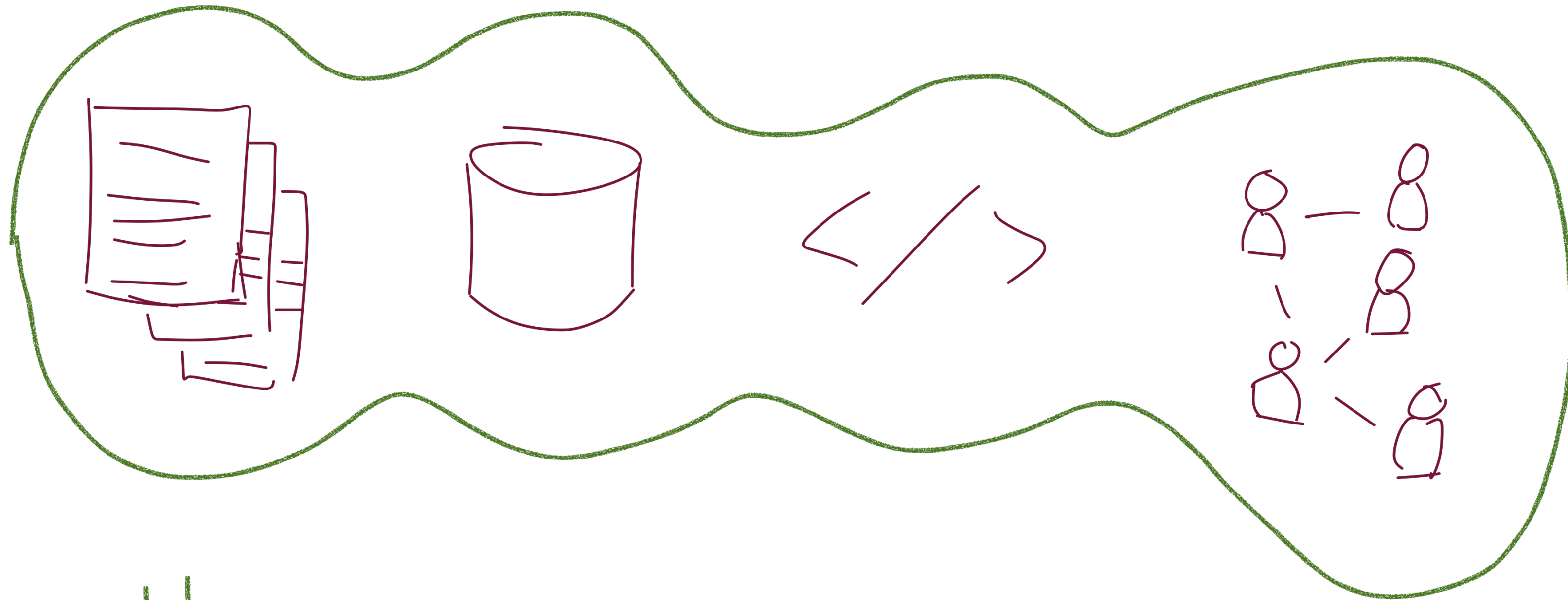


code



communities

The scholarly comms space has a "typical" set of tools that we have been discussing for donkeys years now.



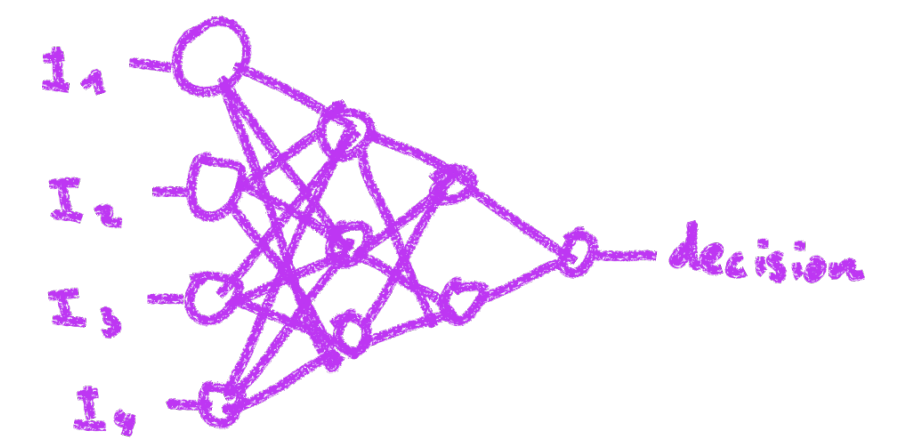
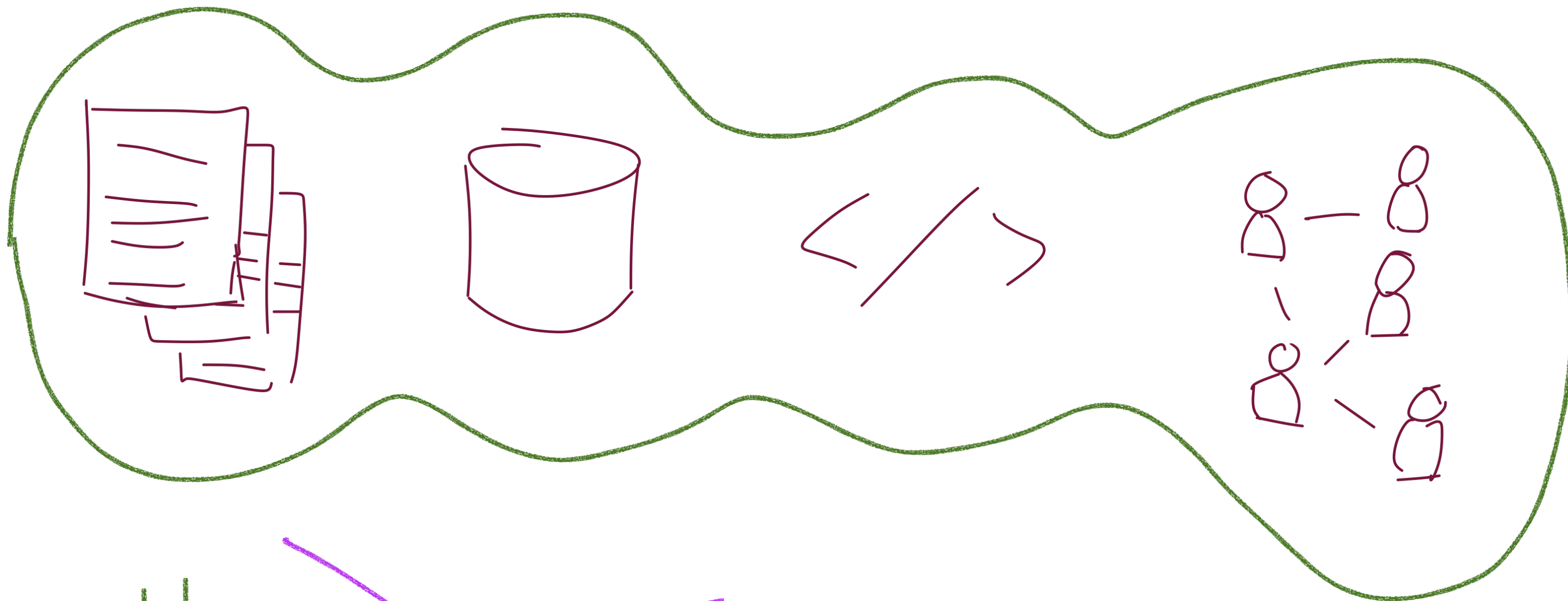
How to use

How to govern

How they respond to use



We kind of understand how these tools work



ML Models

~~How to use~~

~~How to govern~~

~~How they respond to use~~



ML Models can vastly expand the kinds of volumes of data that we can ask question of, in a way vastly expanding our capabilities.

To say this is hardly controversial, but as I observe that we are still talking about how to deal with data, let alone code, I just want to us to raise our eyes towards this approaching future.

We will need to agree norms and practices as a community for how to work with this new class of tool