# Rich Context:
# support for cross-agency data stewardship, measuring dataset impact on public policy

Paco Nathan @pacoid

derwen.ai

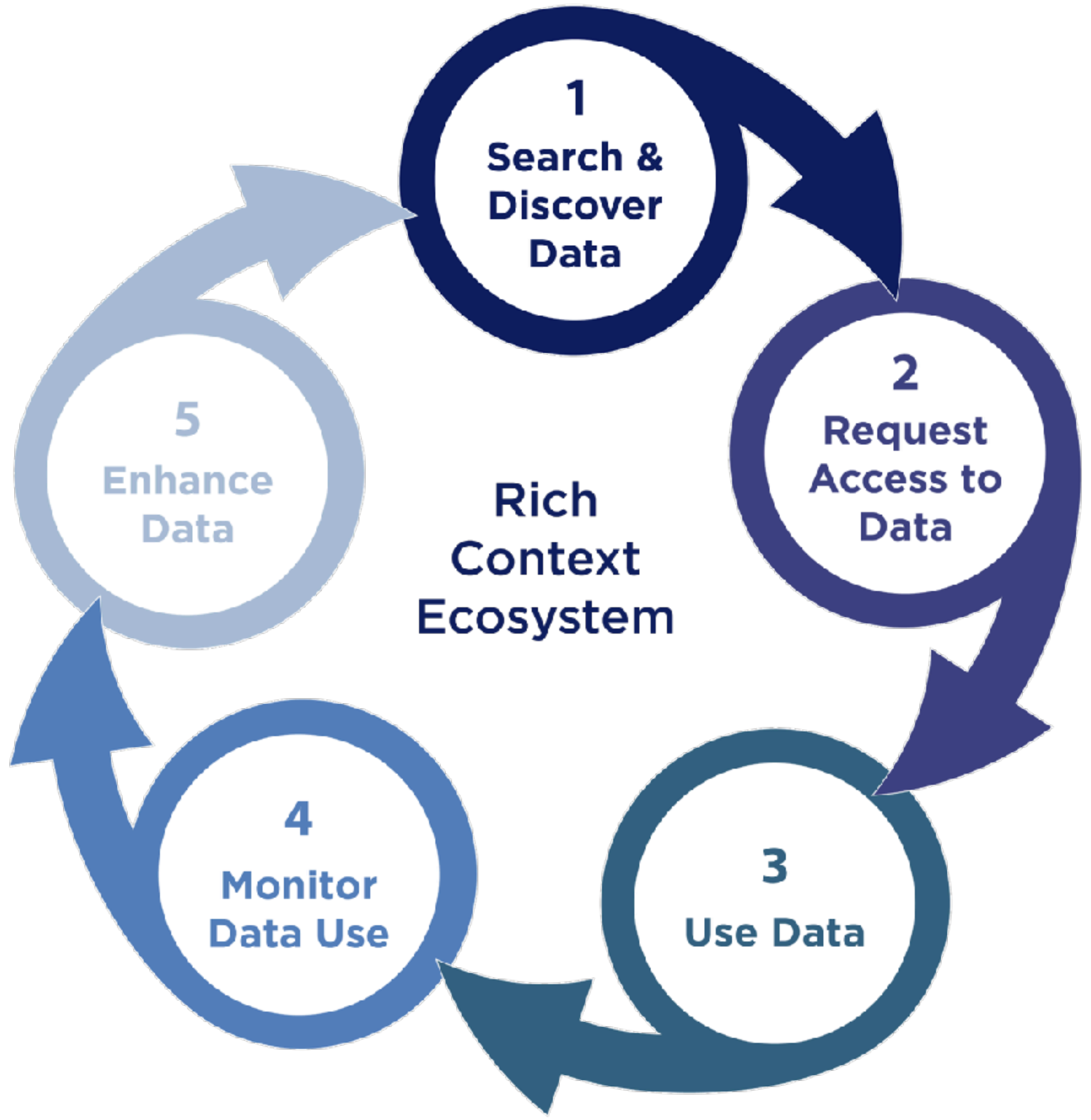
# part 1:
# public private partnership – a case study

# Administrative Data Research Facility

## Coleridge Initiative

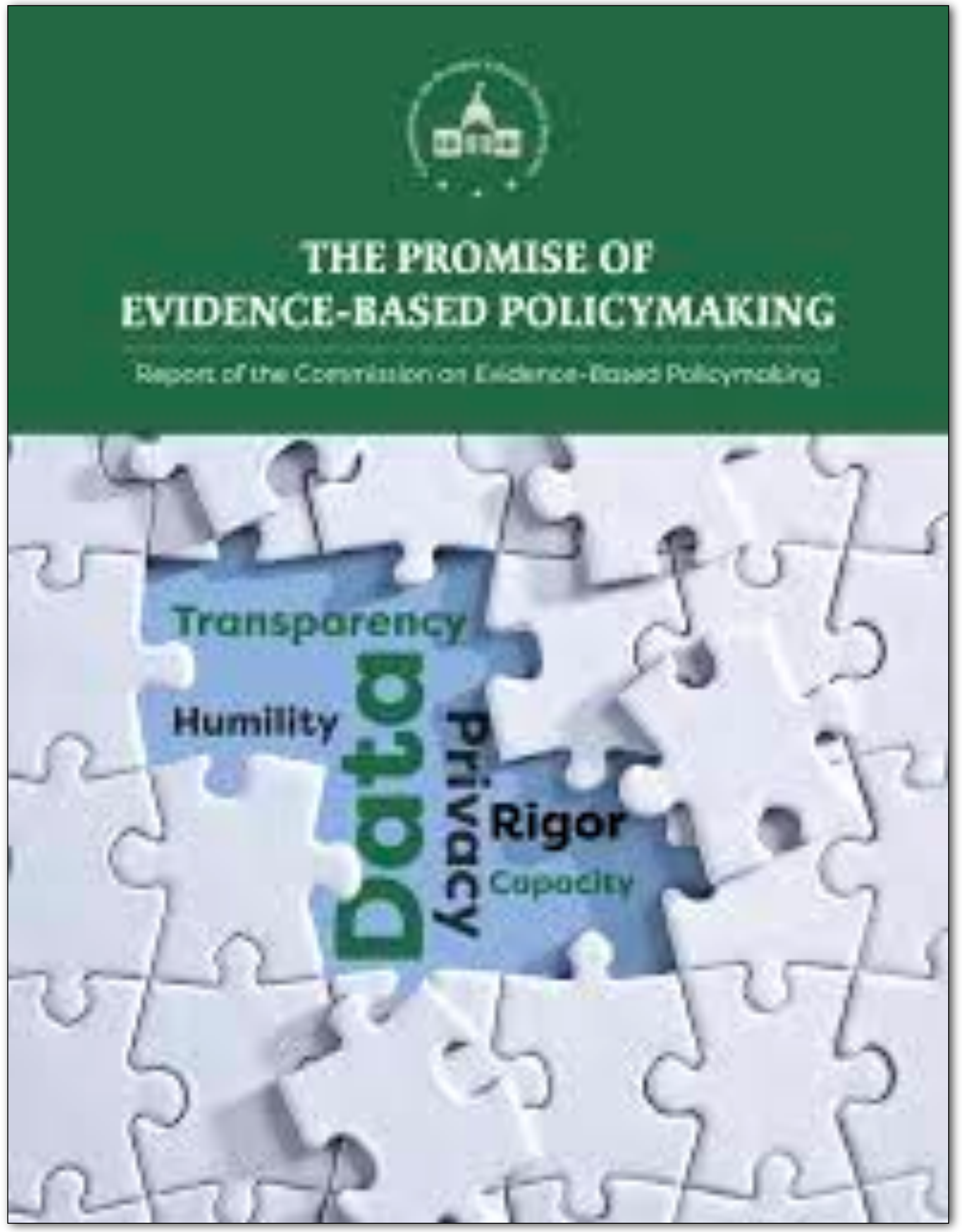
**Julia Lane**, et al.  NYU Wagner

- FedRAMP-compliant **ADRF framework** on AWS GovCloud: “public agency capacity to accelerate the effective use of new datasets”
- for research projects using cross-agency sensitive data, in US and EU – **now in use by 30+ agencies**
- cited as the first federal example of Secure Access to Confidential Data in the final report of the Commission on **Evidence-Based Policymaking**
- augments **Data Stewardship** practices; collaboration with Project Jupyter on the related **data gov features**



# Evidence-Based Policymaking

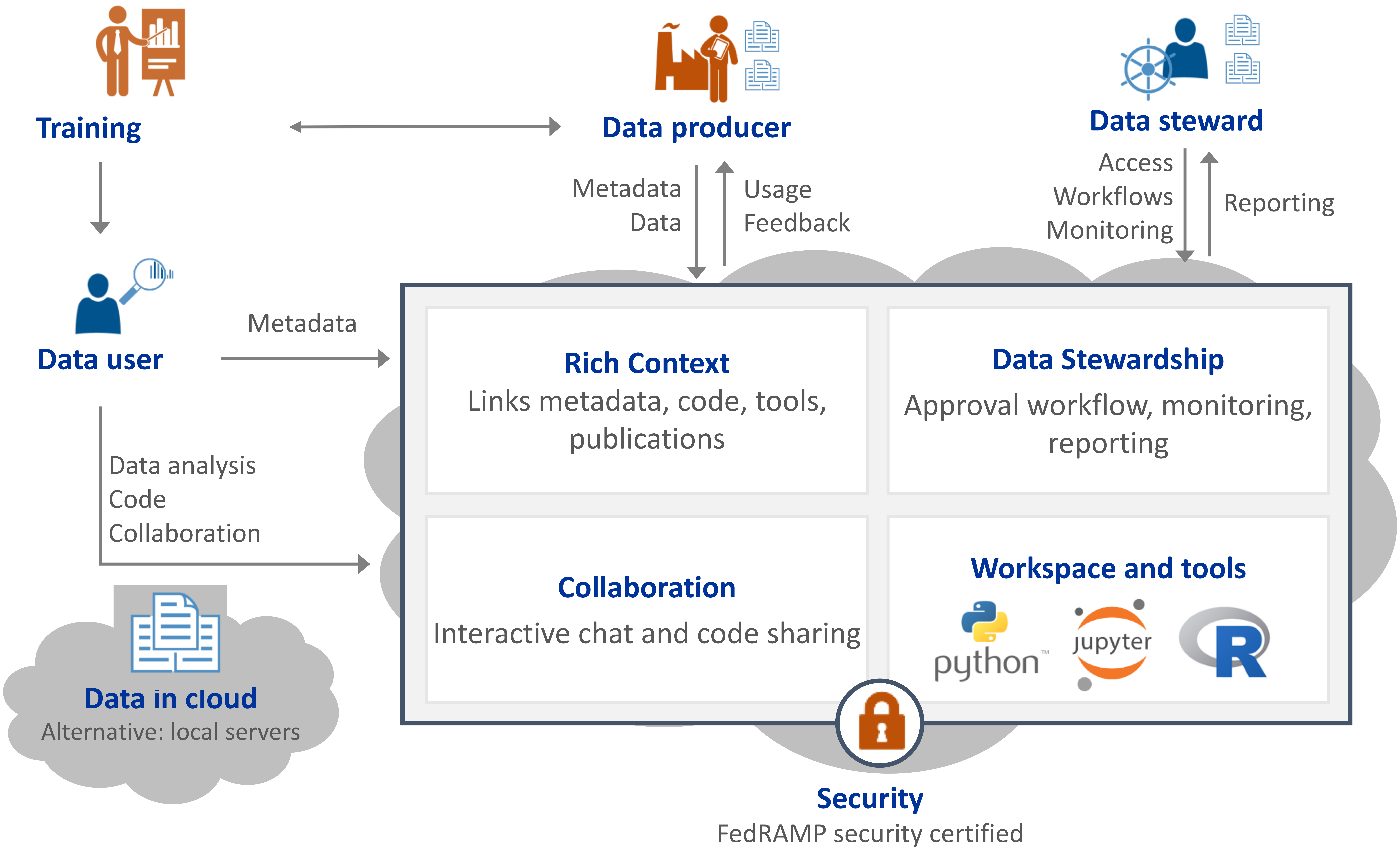
- **Foundations for Evidence-based Policymaking Act** (2018)
- **Information Quality Act** (2001)
- **NIH Strategic Plan for Data Science** (2018)
- **US federal data strategy**
- **Year-1 Action Plan** (2019)



THE PROMISE OF
EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

See also:
**“Evidence-based decision making: What DOE, USDA and others are learning”**
Wyatt Kash
*FedScoop* (2019-06-28)



**Training**

**Data producer**

**Data steward**

Metadata
Data

Usage
Feedback

Access
Workflows
Monitoring

Reporting

**Data user**

Metadata

Data analysis
Code
Collaboration

**Data in cloud**
Alternative: local servers

**Rich Context**
Links metadata, code, tools, publications

**Data Stewardship**
Approval workflow, monitoring, reporting

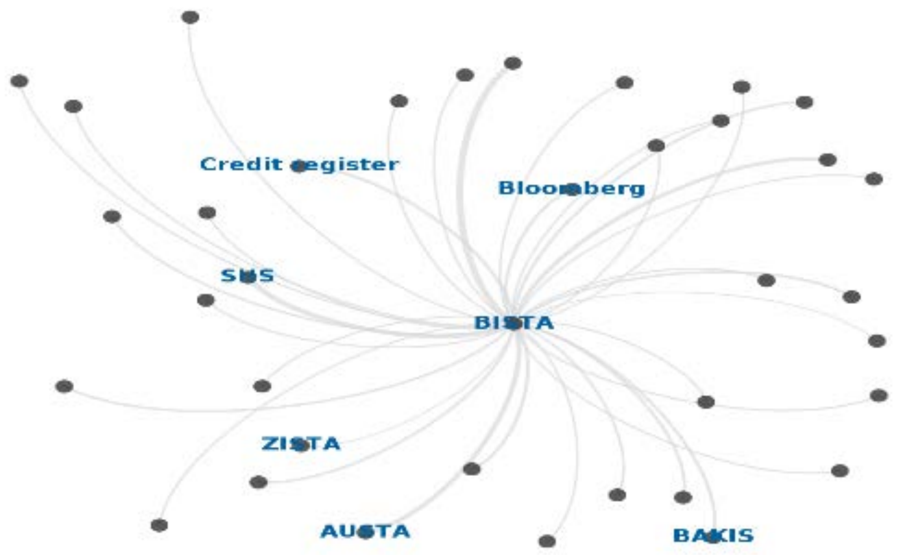
**Collaboration**
Interactive chat and code sharing

**Workspace and tools**

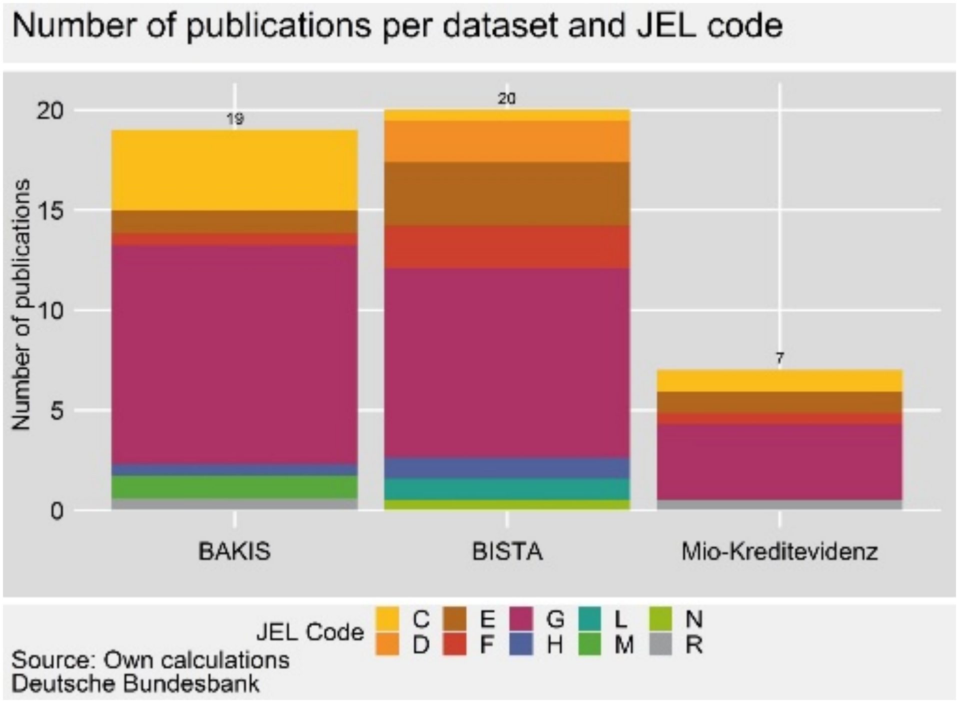
**Security**
FedRAMP security certified

# Research Data Centers

To date, we successfully launched a machine learning competition and obtained algorithms to extract used datasets from publications. Based on the results, we implemented two prototypes that will serve as building blocks for a unified system.



**Recommend data** to researchers (“*based on your interest, you might also like this data*”).



Source: Own calculations
Deutsche Bundesbank

**Data impact factor** (*“This dataset generates most research/ societal / policy value”)*

The two prototypes already create value by enabling more optimal data usage and by supporting effective resource allocation into valuable data that generates results.

*– Hendrik Doll, Deutsche Bundesbank*

**The bigger picture**

Our vision is a system, where researchers and analysts apply for confidential data access, get data recommendations, have a secure remote digital workspace, and provide feedback on data. Such a system relies on three non-traditional data sources: Data usage information, implicit knowledge in researchers’ heads (incentives for sharing), and structured administer metadata (*annodata*) to automatically govern data access.

# Research Data Centers

**Factsheet: Building an integrated data access system for empirical research**

Hendrik Doll, Stefan Bender, Jannick Blaschke, Christian Hirsch, Christian Resch[1]

*Federal Reserve Board, Washington, D.C., October 1-2, 2019*

**Our idea**

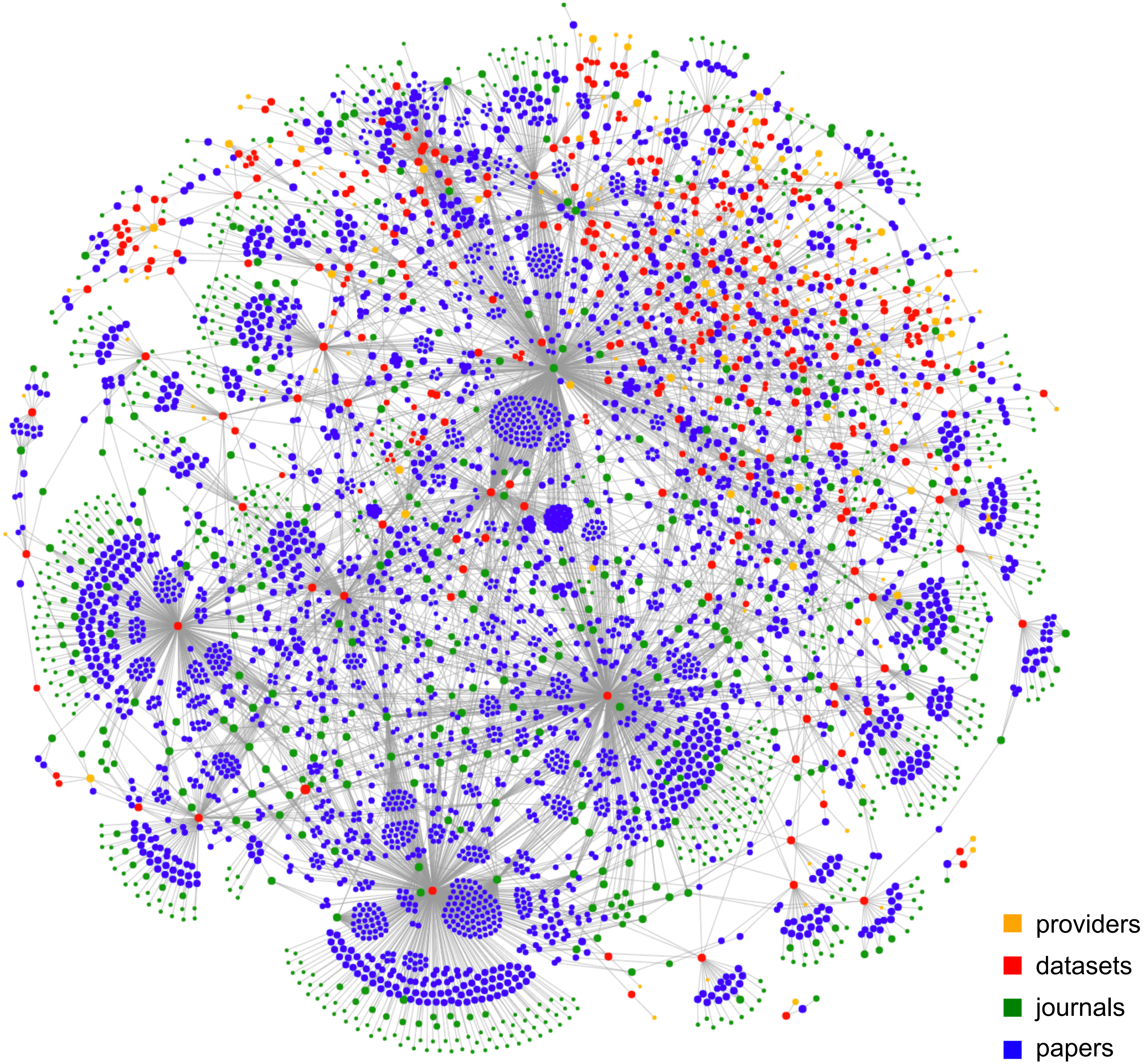
Empirical research and evidence-based policy decisions increasingly rely on microdata. Research publications are well referenced and findable, however structured information on data usage is less available. The current project aspires to change this by building a data-centric ecosystem with rich context and a community around microdata.

# Need: recommenders for researchers/analysts

- Objective: provide better means of *search and discovery* for social science researchers and agency analysts.
- Collect workflow telemetry and query logs to augment the graph.
- Currently developing recommender systems based on the graph.
- This accelerates research and also assists training (e.g., onboarding agency analysts).
- Near-term goal: identify people with specific expertise.
- Long-term goal: learn workflow configurations to support AutoML meta-learning.

# Rich Context

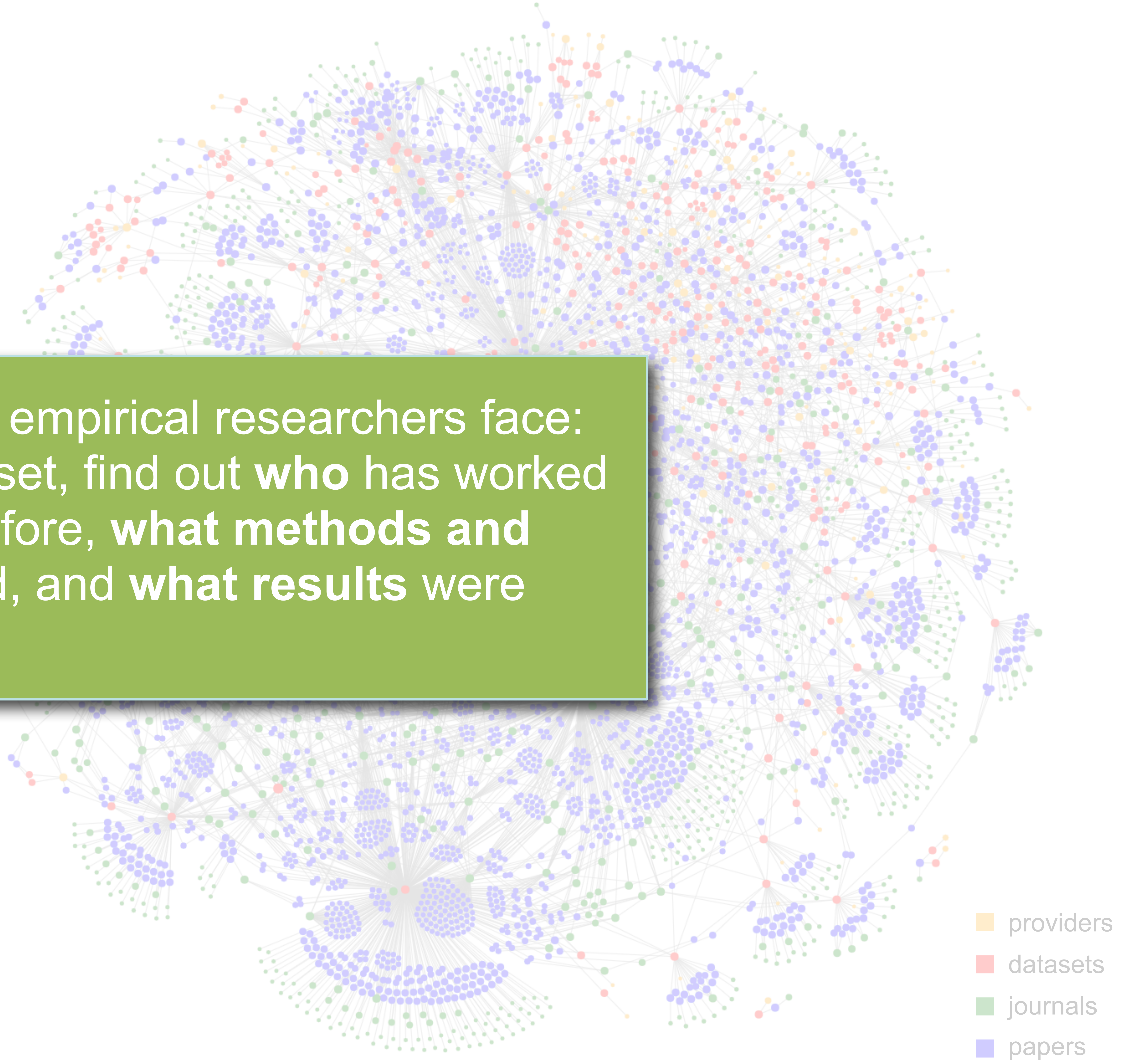
- Focus on *socioeconomic impact*
- Funded by Schmidt Futures, Sloan, Overdeck
- Partnering with Bundesbank, USDA, etc.
- Collaboration with SAGE Pub, RePEc, ResearchGate, Digital Science, etc.



# Rich Context

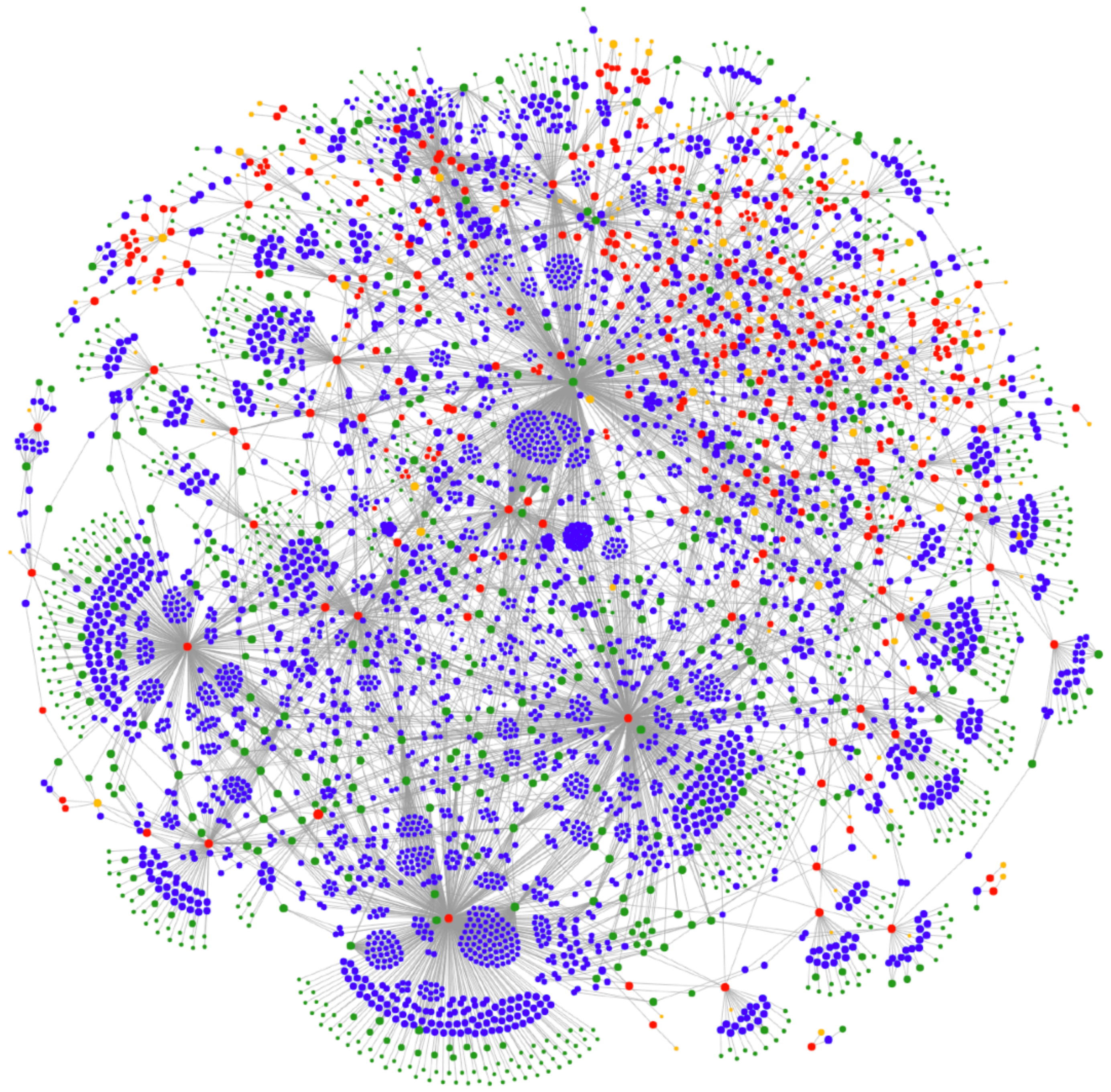
- Focus on *socioeconomic impact*
- Funded by Schmidt Futures, Sloan, Overdec
- Partnering with USDA, etc.
- Collaboration w Pub, RePEc, R Digital Science, etc.

Challenges that empirical researchers face: for a given dataset, find out **who** has worked with the data before, **what methods and code** were used, and **what results** were produced.



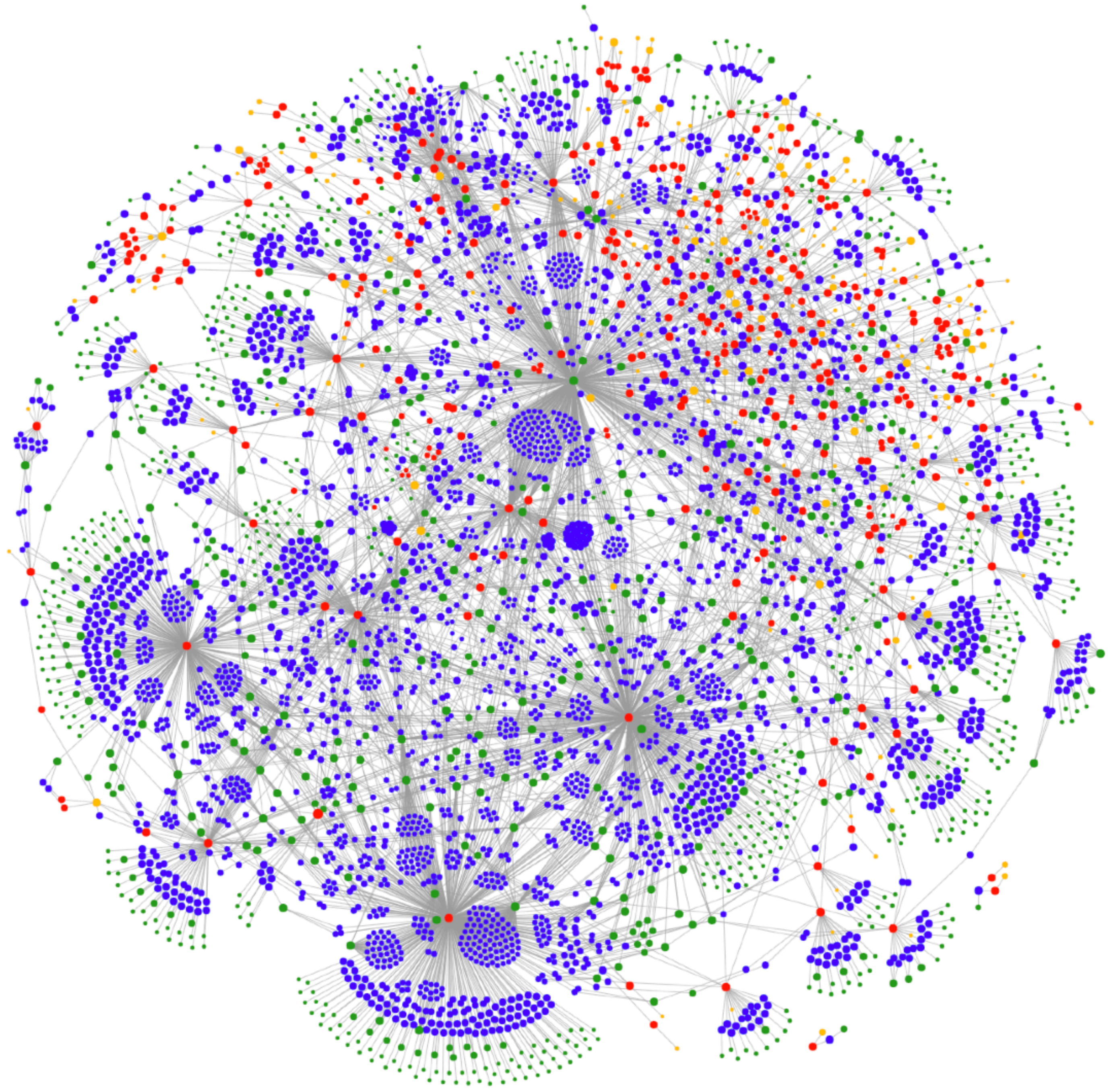
# Knowledge Graph – why?

- Allow flexibility for metadata representation
- Measure metadata quality
- Prepare features for ML models
- Build recommenders for *experts*, *topics*, *tools*, etc.
- Engage the public with automated data inventories
- Recommend configurations to new analysts
- Identify which datasets get used with others
- Quantify impact of datasets on policy



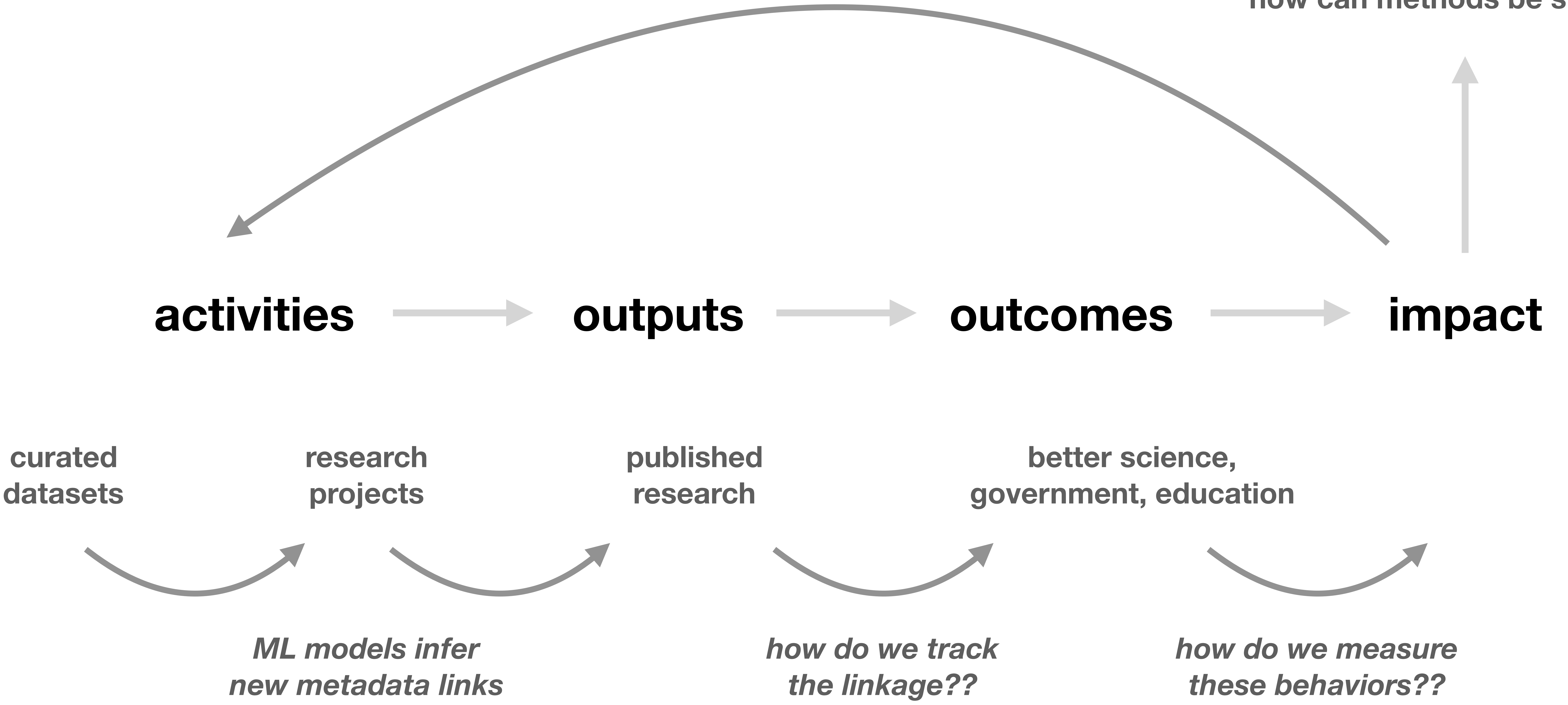
# Knowledge Graph – how?

- Manual data entry and curation of linked data
- Use persistent identifiers whenever possible:
  DOI, ISSN, ROR, ORCID, etc.
- Leverage ML models to infer missing metadata
- Federate queries of discovery services APIs
- Suggest corrections for metadata errors
- Use HITL to build feedback loops that engage experts, and provide convenient means for manual override
- Identify errors by using unit tests, ontology axioms, graph analytics, etc.
- **Collaborate with agency libraries!**



# KG process

- who are the expert people?
- which topics are emerging?
- how can methods be shared?



**activities** → **outputs** → **outcomes** → **impact**

curated datasets

research projects

published research

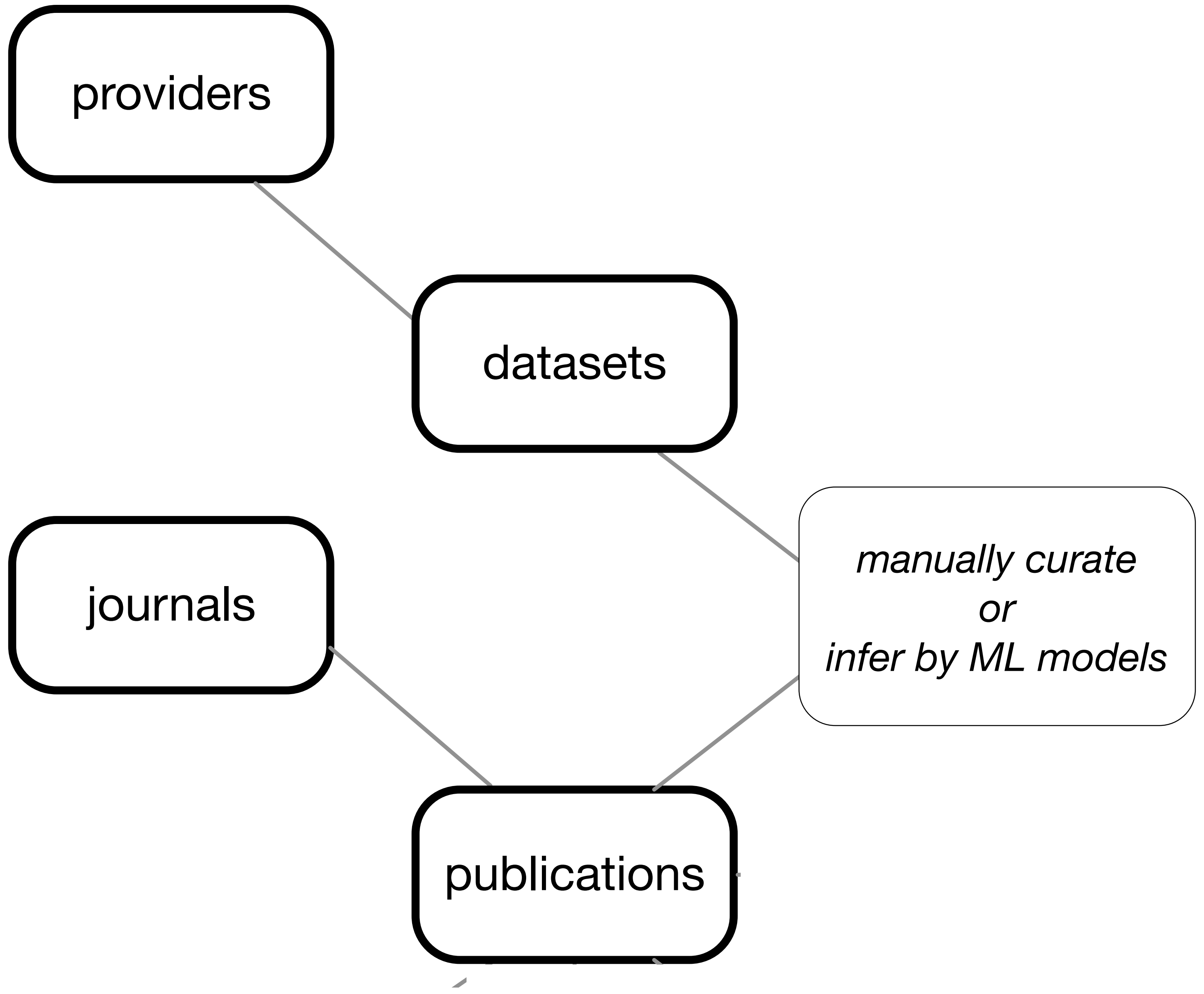
better science, government, education

*ML models infer new metadata links*

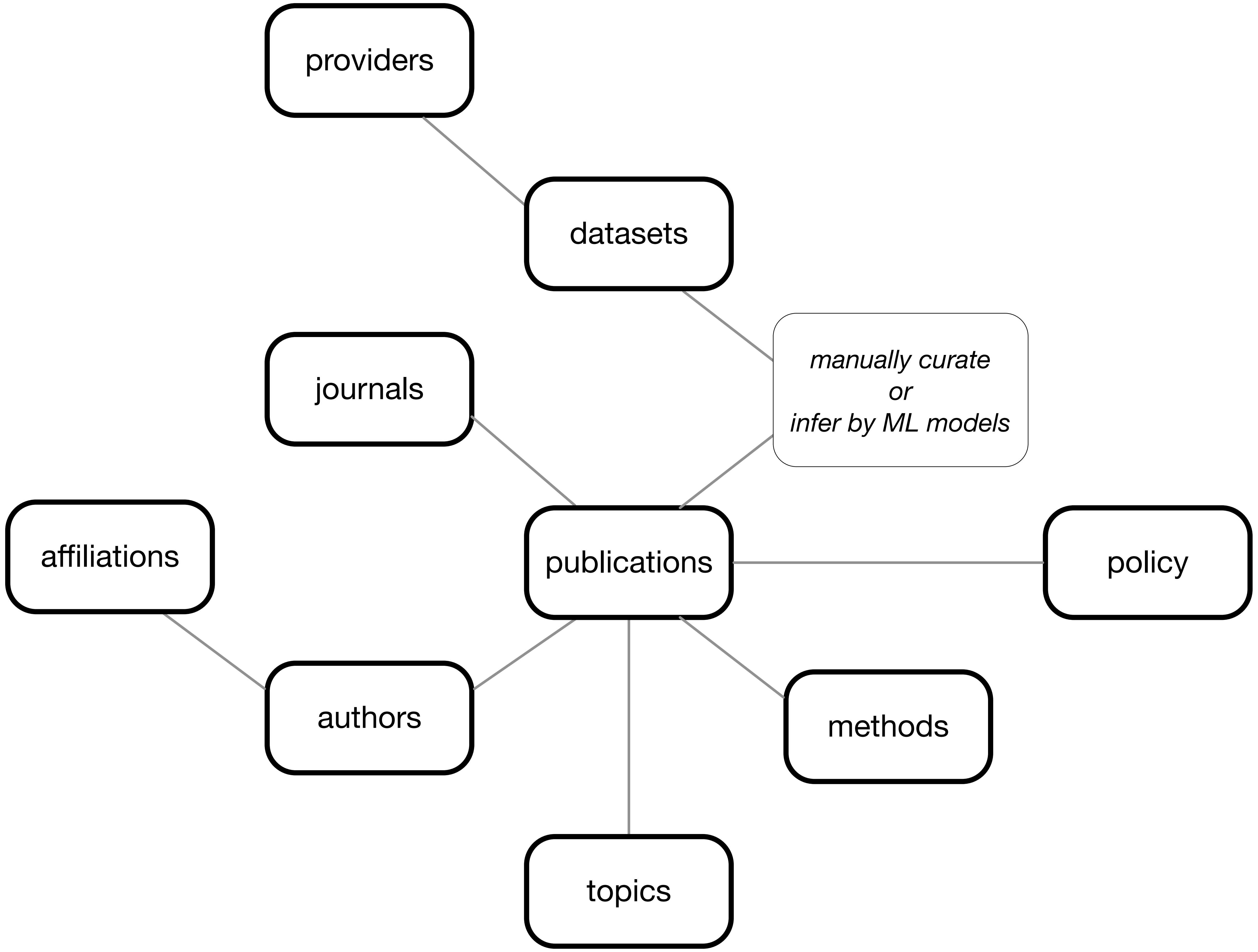
*how do we track the linkage??*

*how do we measure these behaviors??*

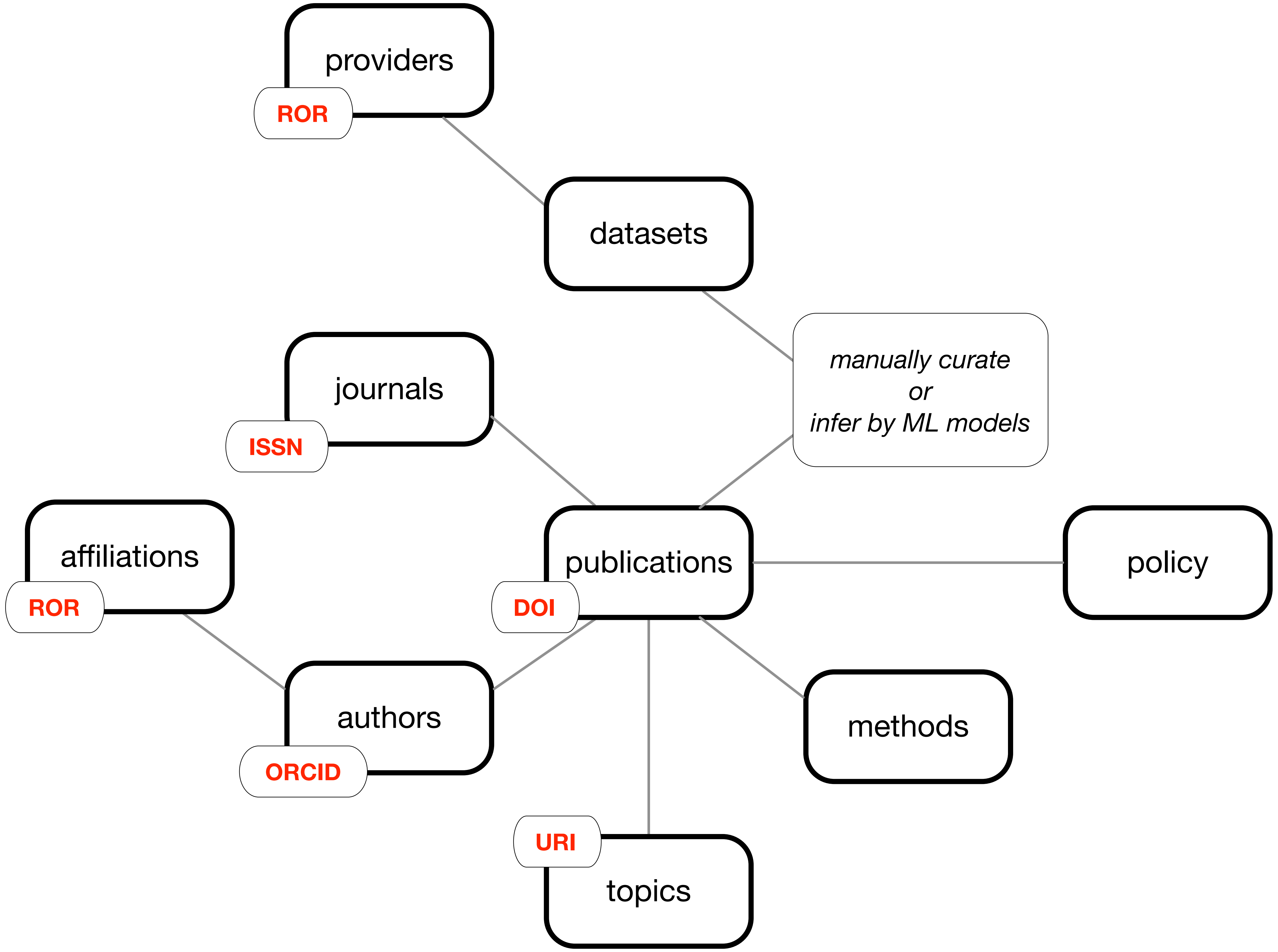
# KG construction and representation



# KG construction and representation

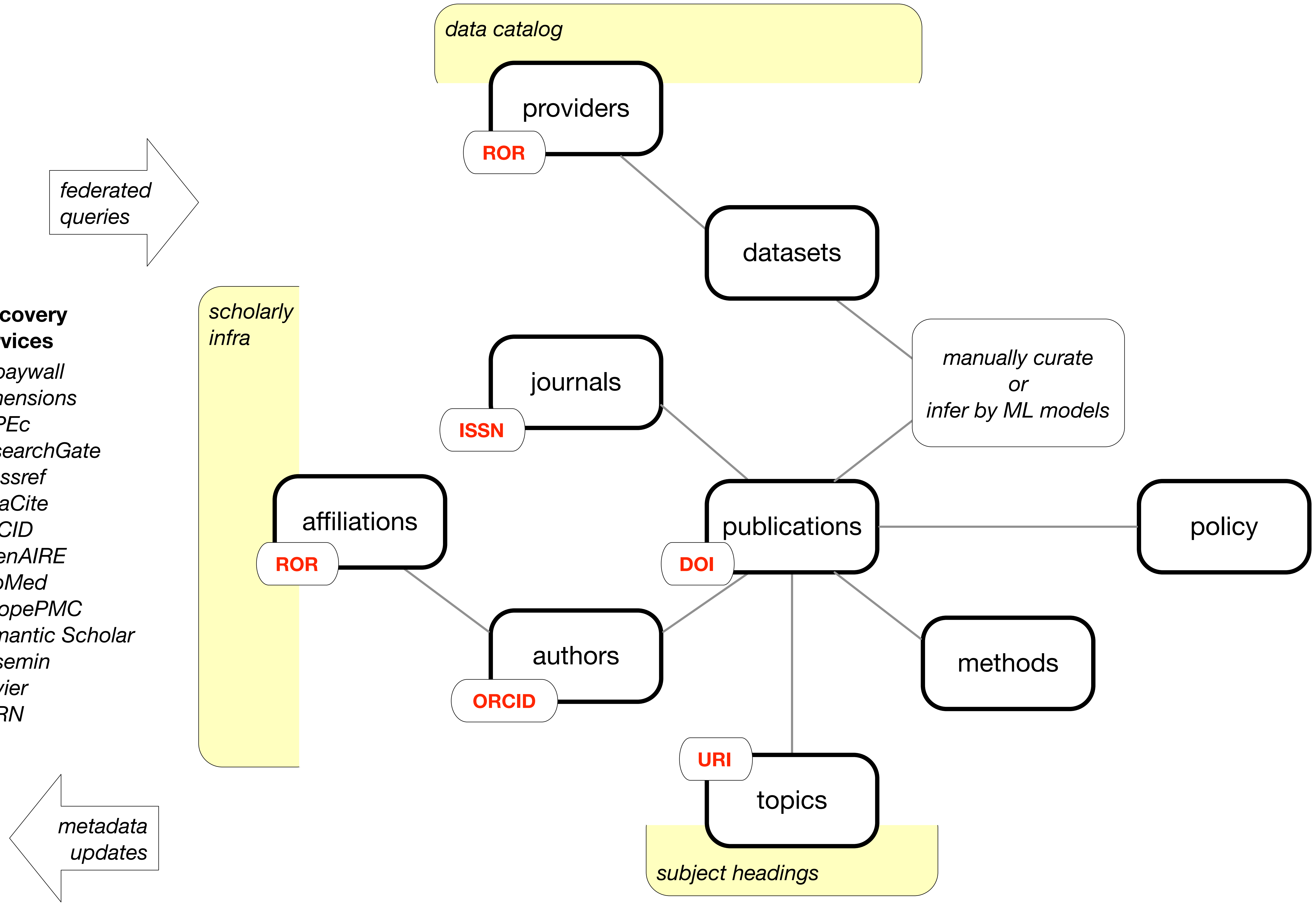


# KG construction and representation



See github.com/Coleridge-Initiative/rclc/wiki/Corpus-Description

# KG construction and representation

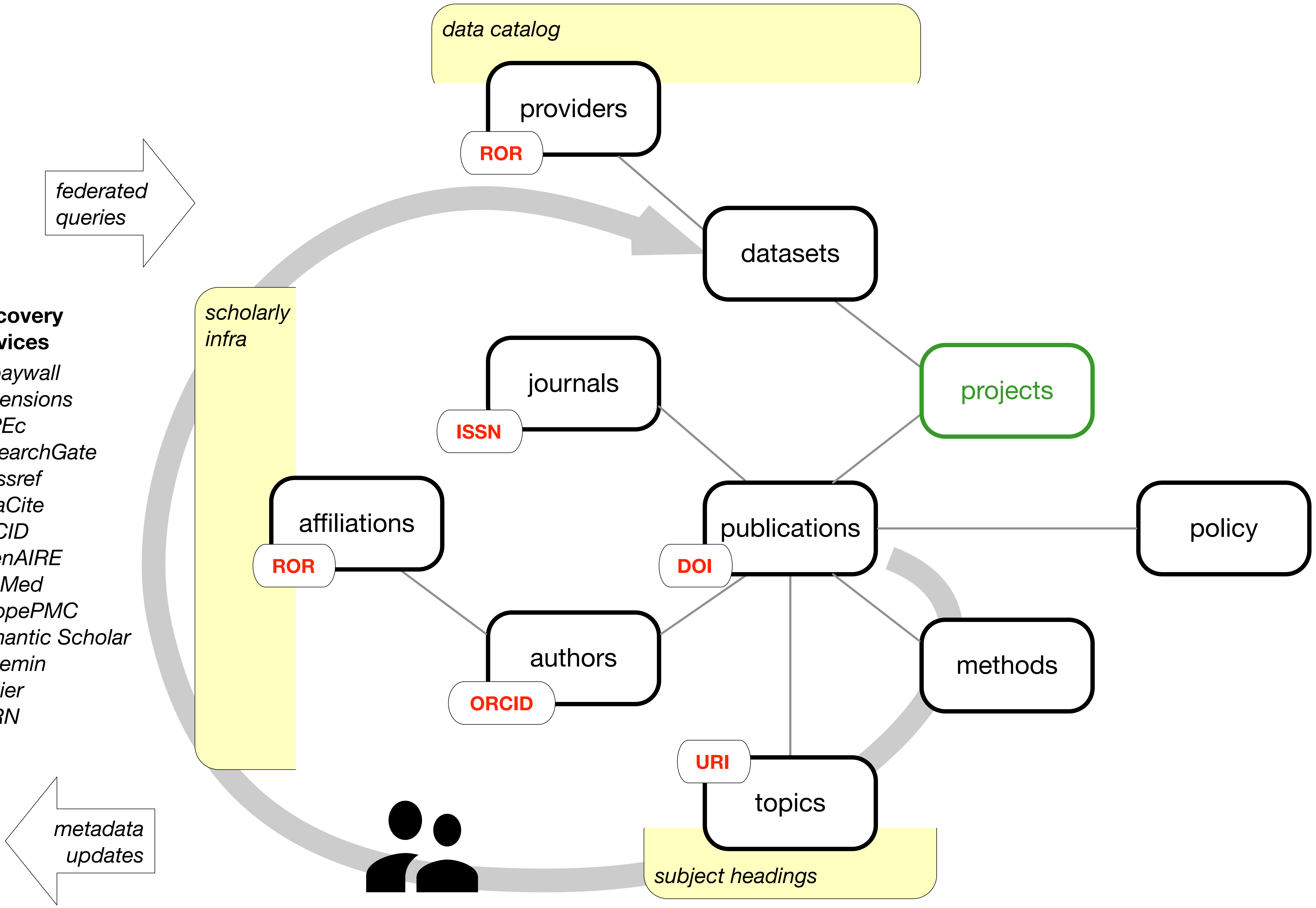


**Discovery Services**

*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# KG construction and representation



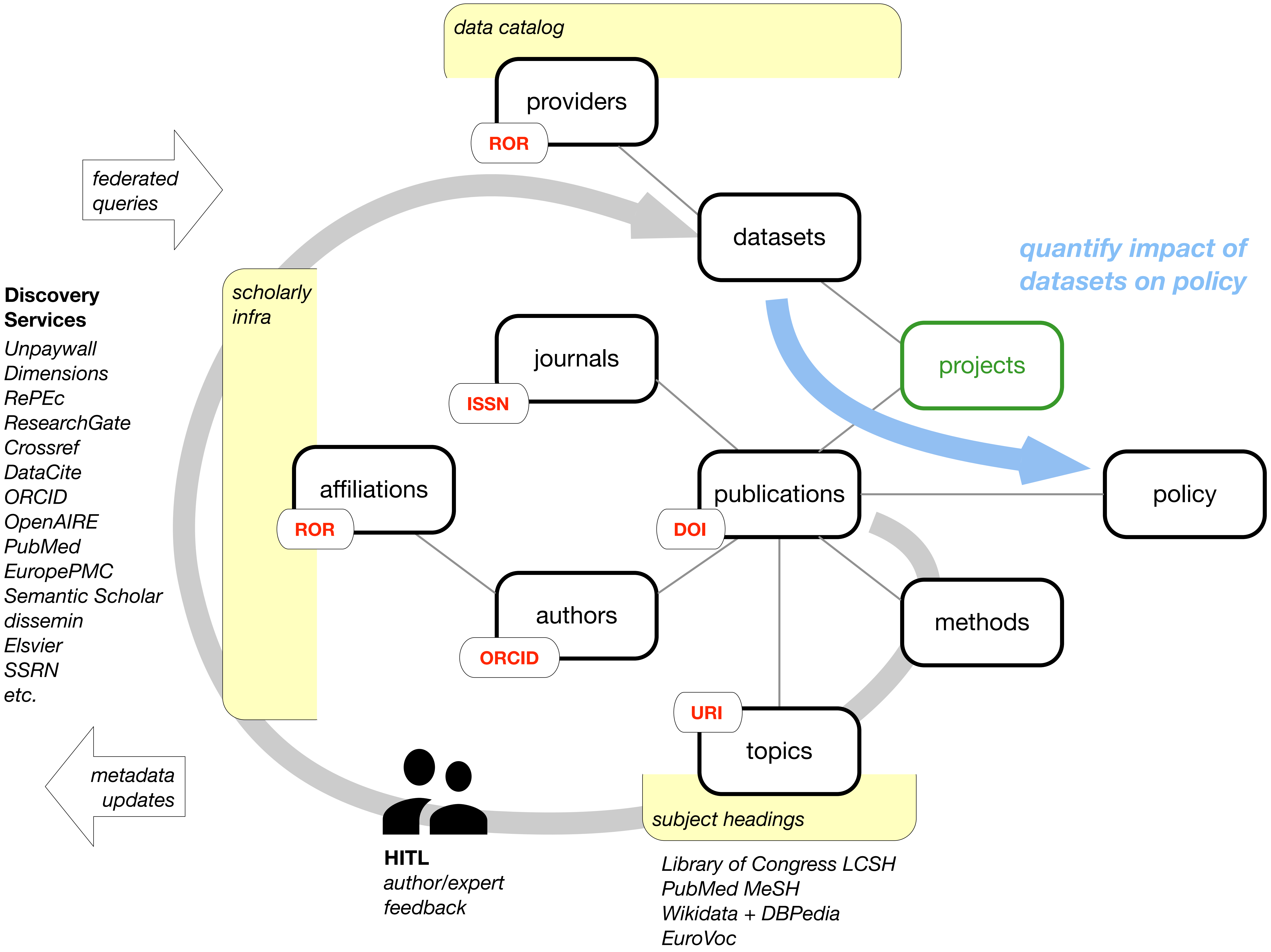
**Discovery Services**

*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*

**HITL**
*author/expert feedback*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# KG construction and representation



*data catalog*

providers

ROR

*federated queries*

datasets

*quantify impact of datasets on policy*

**Discovery Services**

*Unpaywall*
*Dimensions*
*RePEc*
*ResearchGate*
*Crossref*
*DataCite*
*ORCID*
*OpenAIRE*
*PubMed*
*EuropePMC*
*Semantic Scholar*
*dissemin*
*Elsvier*
*SSRN*
*etc.*

*scholarly infra*

journals

ISSN

projects

affiliations

ROR

publications

DOI

policy

authors

ORCID

methods

URI

topics

*metadata updates*

**HITL**
*author/expert feedback*

*subject headings*

*Library of Congress LCSH*
*PubMed MeSH*
*Wikidata + DBPedia*
*EuroVoc*

# Open Source Projects

- **RCGraph** – Rich Context knowledge graph management
  **github.com/Coleridge-Initiative/RCGraph**
- **richcontext.scholapi** – federated discovery services and metadata exchange across scholarly infrastructure APIs
  **pypi.org/project/richcontext-scholapi**
- **adrf-onto** – controlled vocabulary for ADRF and Rich Context using OWL, SKOS, DCAT, PAV, CITO, FaBiO, etc.
  **github.com/Coleridge-Initiative/adrf-onto**
- **RCLC** – ML leaderboard competition
  **github.com/Coleridge-Initiative/rclc**

See also:
**“Machine Learning Highlights for Rich Context”**

# Funded additions to Project Jupyter

Make datasets and projects top-level constructs, support metadata exchange and privacy-preserving telemetry from notebook usage:

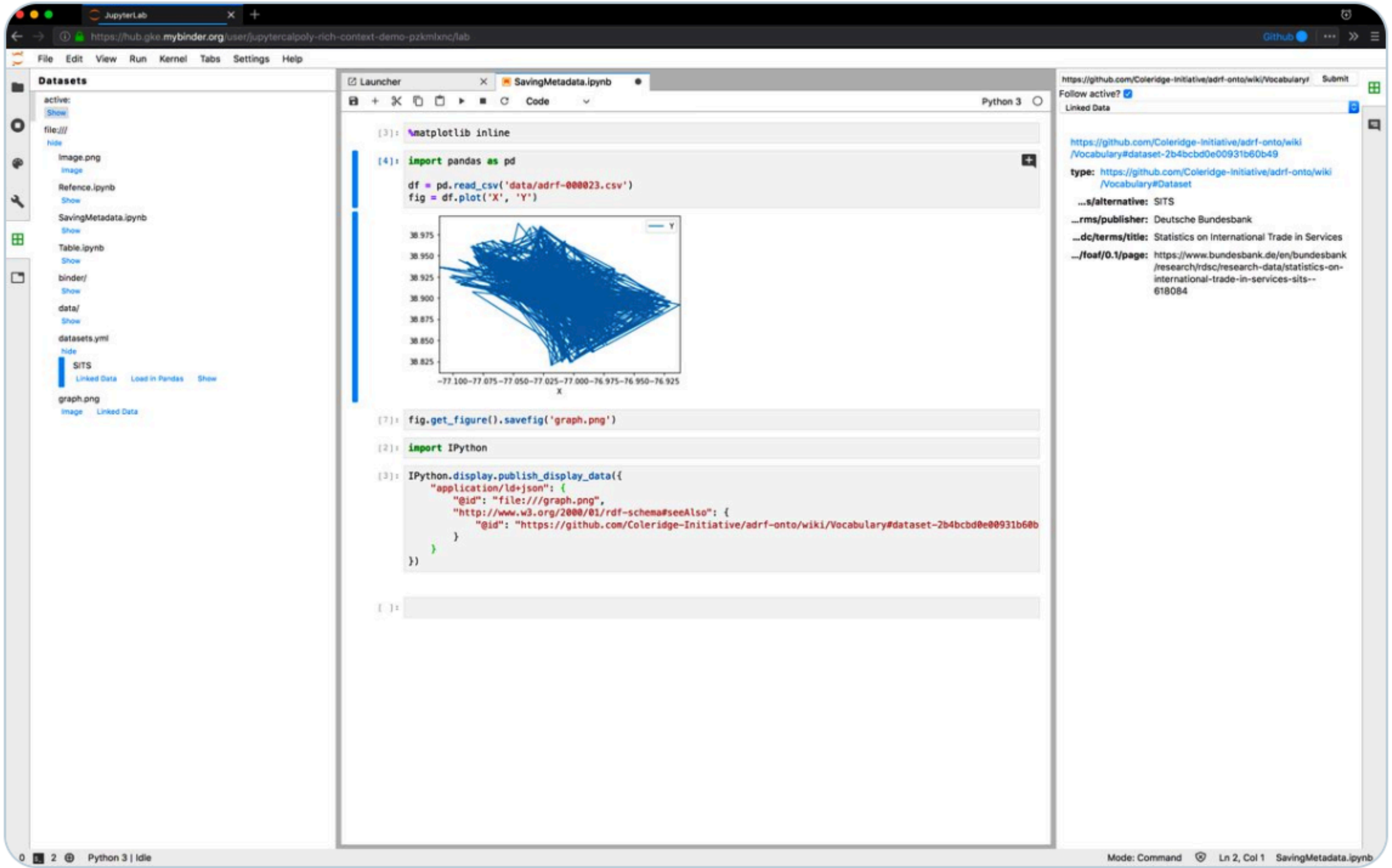
- JupyterLab **Commenting** and real-time collab similar to Google Docs
- JupyterLab **Data Explorer**: register datasets within research projects
- JupyterLab **Metadata Explorer**: browse metadata descriptions, get recommendations through knowledge graph inference (via extension)
- **Data Registry** (original proposal)
- **Telemetry** (privacy-preserving, reports usage)

**Saul Shanabrook**
@SShanabrook

twitter.com/SShanabrook/status/1182442214980501505

Replying to @choldgraf

Data Catalog Vocabulary and other related vocabularies are useful here. w3.org/TR/vocab-dcat/ We are building a way to explore metadata defined in JSON LD that uses these in JupyterLab github.com/jupyterlab/jup… cc @pacoid



Datasets

active:
Show

file:///
hide

Image.png
Image

Refence.ipynb
Show

SavingMetadata.ipynb
Show

Table.ipynb
Show

binder/
Show

data/
Show

datasets.yml
hide

SITS
Linked Data Load in Pandas Show

graph.png
Image Linked Data

```
[3]: %matplotlib inline

[4]: import pandas as pd

     df = pd.read_csv('data/adrf-000023.csv')
     fig = df.plot('X', 'Y')
```

```
[7]: fig.get_figure().savefig('graph.png')

[2]: import IPython

[3]: IPython.display.publish_display_data({
         "application/ld+json": {
             "@id": "file:///graph.png",
             "http://www.w3.org/2000/01/rdf-schema#seeAlso": {
                 "@id": "https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#dataset-2b4bcbd0e00931b60b
             }
         }
     })

[ ]:
```

https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#

Follow active? ☑

Linked Data

https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#dataset-2b4bcbd0e00931b60b49

type: https://github.com/Coleridge-Initiative/adrf-onto/wiki/Vocabulary#Dataset

...s/alternative: SITS

...rms/publisher: Deutsche Bundesbank

...dc/terms/title: Statistics on International Trade in Services

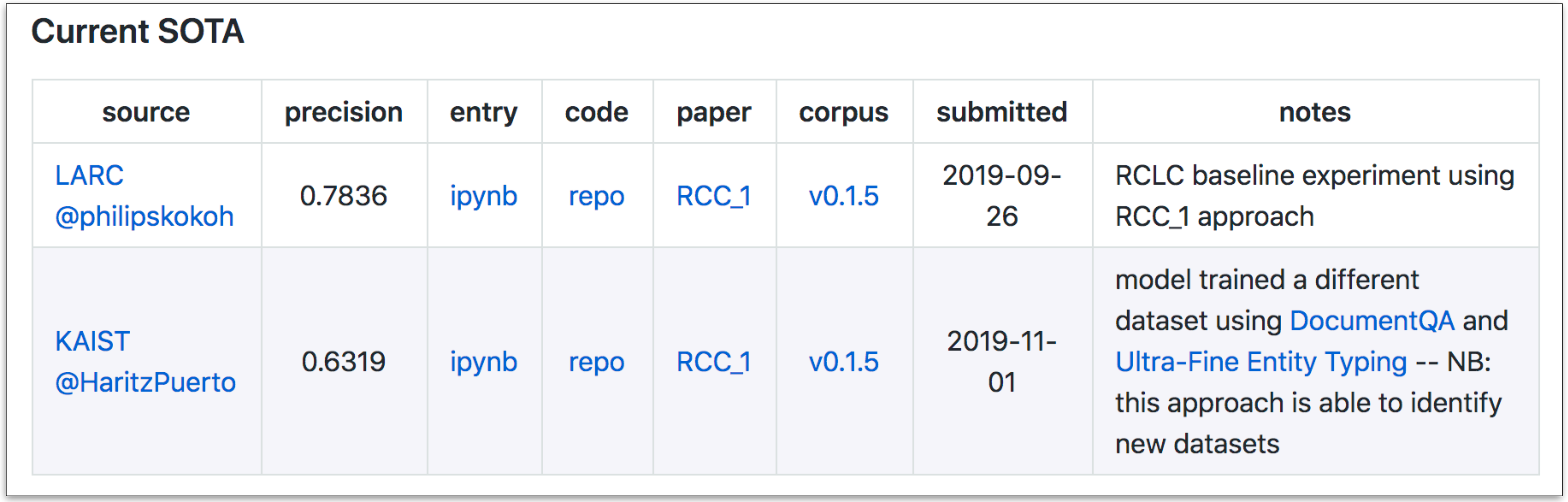
.../foaf/0.1/page: https://www.bundesbank.de/en/bundesbank/research/rdsc/research-data/statistics-on-international-trade-in-services-sits--618084

12:46 AM · Oct 11, 2019 · Twitter Web App

# ML Leaderboard Competition

## github.com/Coleridge-Initiative/rclc

- update from RCC competition in 2018-2019
- ongoing ML leaderboards (similar to **NLP-progress**)
- open source, hosted on GitHub
- highly curated test sets, all open-access publications
- teams collaborate via GH issues on corpus data quality, etc.
- focus on *precision* for ML model evaluation

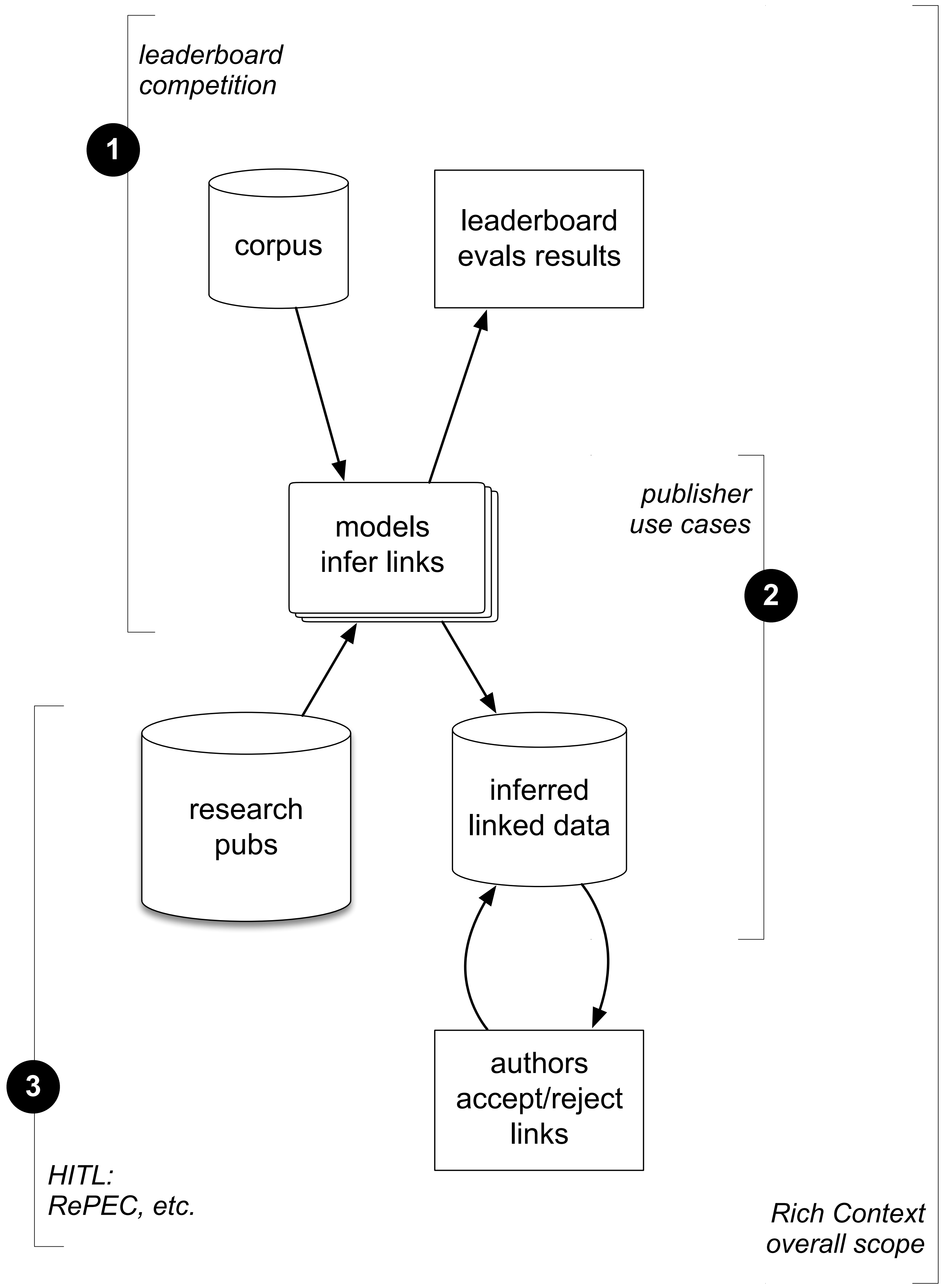


**Current SOTA**

| source | precision | entry | code | paper | corpus | submitted | notes |
|---|---|---|---|---|---|---|---|
| LARC @philipskokoh | 0.7836 | ipynb | repo | RCC_1 | v0.1.5 | 2019-09-26 | RCLC baseline experiment using RCC_1 approach |
| KAIST @HaritzPuerto | 0.6319 | ipynb | repo | RCC_1 | v0.1.5 | 2019-11-01 | model trained a different dataset using DocumentQA and Ultra-Fine Entity Typing -- NB: this approach is able to identify new datasets |

# Human-in-the-loop

- semi-supervised learning, aka "human-in-the-loop" – in progress via RePEc
- interact with authors to confirm metadata inferred by ML models
- feedback from experts improves the corpus metadata and the ML modeling



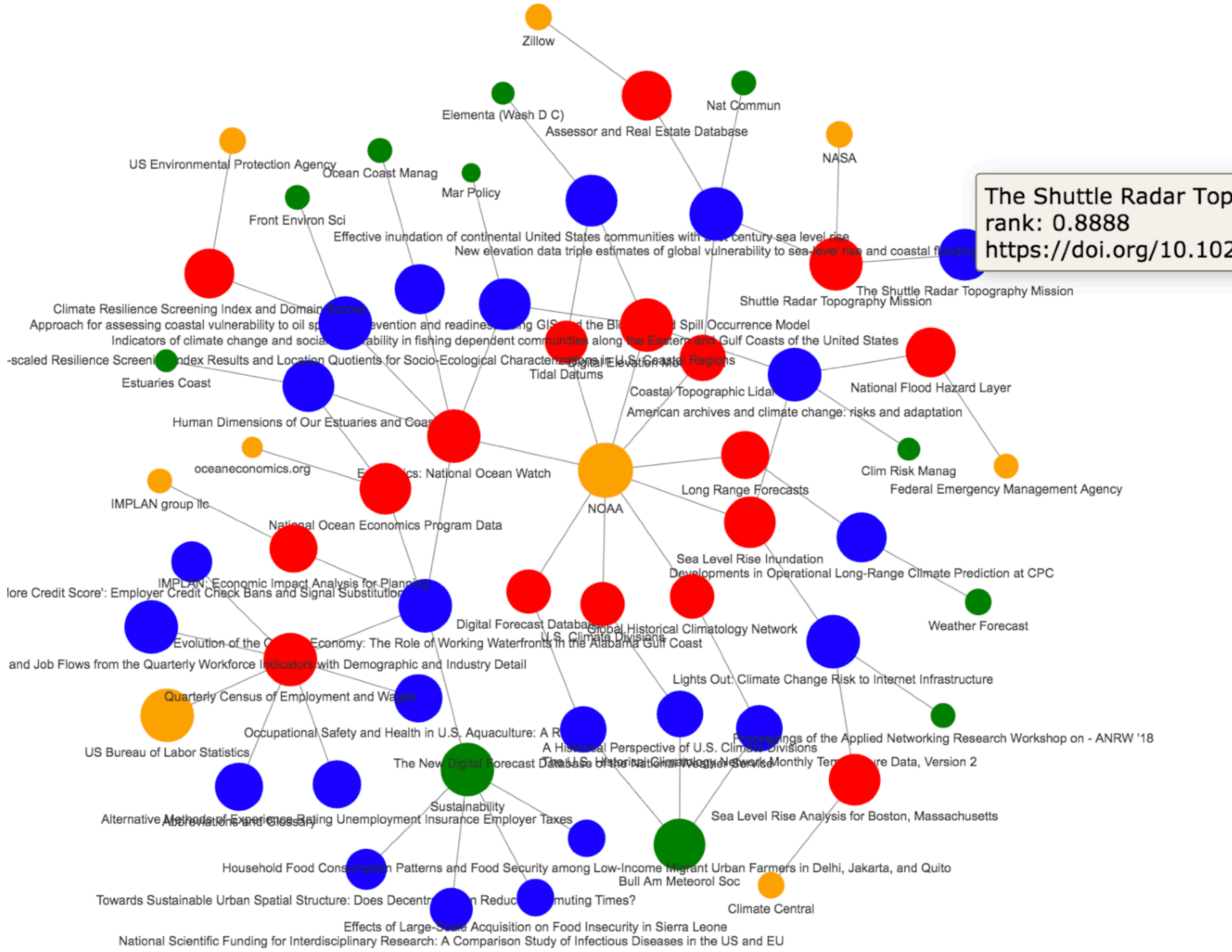
See also:

**"Human-in-the-loop AI for scholarly infrastructure"**

**"New initiative to help with discovery of dataset use in scholarly work"**, Christian Zimmerman

# Collaboration with NOAA

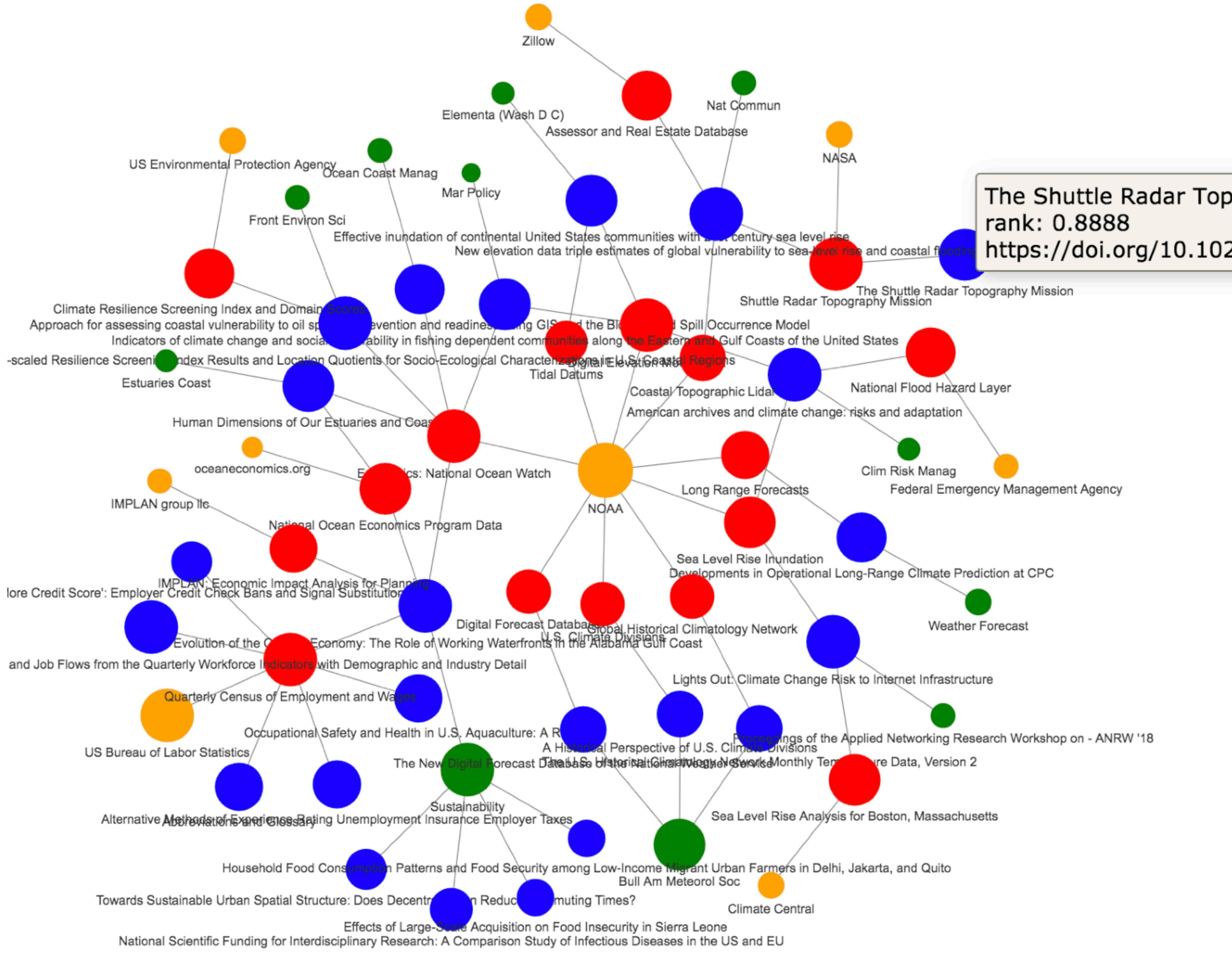
- Initial focus on coastal inundation and community resilience, working with NOS
- Develop reusable dataset discovery services, so that the public and researchers can find trustworthy, high-impact data
- Identify experts who have used the data and the associated research topics, associated analytical methods and tools, and related datasets (e.g., Zillow, EPA, NASA, FEMA, etc.)
- Generalize for other federal agencies, such as USDA and NSF, as well as to international organizations, such as Deutsche Bundesbank
- Bring in AI expertise from industry and academia: KAIST, LARC, Recognai, DLA, Primer AI, GESIS, AllenAI, etc.



Zillow
Elementa (Wash D C)
Nat Commun
Assessor and Real Estate Database
US Environmental Protection Agency
Ocean Coast Manag
Mar Policy
NASA
Front Environ Sci
The Shuttle Radar Top
rank: 0.8888
https://doi.org/10.102
Effective inundation of continental United States communities with 21st century sea level rise
New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding
The Shuttle Radar Topography Mission
Shuttle Radar Topography Mission
Climate Resilience Screening Index and Domain Scores
Approach for assessing coastal vulnerability to oil spill prevention and readiness using GIS and the Blowout and Spill Occurrence Model
Indicators of climate change and social vulnerability in fishing dependent communities along the Eastern and Gulf Coasts of the United States
-scaled Resilience Screening Index Results and Location Quotients for Socio-Ecological Characterizations in U.S. Coastal Regions
Digital Elevation Model
Estuaries Coast
Tidal Datums
Coastal Topographic Lidar
National Flood Hazard Layer
American archives and climate change: risks and adaptation
Human Dimensions of Our Estuaries and Coasts
oceaneconomics.org
Economics: National Ocean Watch
NOAA
Long Range Forecasts
Clim Risk Manag
Federal Emergency Management Agency
IMPLAN group llc
National Ocean Economics Program Data
Sea Level Rise Inundation
Developments in Operational Long-Range Climate Prediction at CPC
IMPLAN: Economic Impact Analysis for Planning
lore Credit Score': Employer Credit Check Bans and Signal Substitution
Weather Forecast
Digital Forecast Database
Global Historical Climatology Network
U.S. Climate Divisions
Evolution of the Ocean Economy: The Role of Working Waterfronts in the Alabama Gulf Coast
and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail
Lights Out: Climate Change Risk to Internet Infrastructure
Quarterly Census of Employment and Wages
Occupational Safety and Health in U.S. Aquaculture: A R
Proceedings of the Applied Networking Research Workshop on - ANRW '18
US Bureau of Labor Statistics
A Historical Perspective of U.S. Climate Divisions
The New Digital Forecast Database of the National Weather Service
The U.S. Historical Climatology Network Monthly Temperature Data, Version 2
Sustainability
Alternative Methods of Experience Rating Unemployment Insurance Employer Taxes
Abbreviations and Glossary
Sea Level Rise Analysis for Boston, Massachusetts
Household Food Consumption Patterns and Food Security among Low-Income Migrant Urban Farmers in Delhi, Jakarta, and Quito
Bull Am Meteorol Soc
Towards Sustainable Urban Spatial Structure: Does Decentralization Reduce Commuting Times?
Climate Central
Effects of Large-Scale Acquisition on Food Insecurity in Sierra Leone
National Scientific Funding for Interdisciplinary Research: A Comparison Study of Infectious Diseases in the US and EU

# Agency Benefits

- Understand more about NOAA’s user community, to help outreach and get feedback, especially for researchers or commercial entities using NOAA data in novel ways.
- Help quantify the value and impact of data and research, especially in the context of NOAA’s Blue Economy initiative.
- Relevance given the Federal Data Strategy and the OPEN Government Data Act: comprehensive data inventory, tools to help data users find trustworthy and relevant data.
- Explore a novel solution to dataset search.


Zillow
Elementa (Wash D C)
Nat Commun
Assessor and Real Estate Database
NASA
US Environmental Protection Agency
Ocean Coast Manag
Mar Policy
Front Environ Sci
Effective inundation of continental United States communities with century sea level rise
New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding
The Shuttle Radar Top
rank: 0.8888
https://doi.org/10.102
The Shuttle Radar Topography Mission
Shuttle Radar Topography Mission
Climate Resilience Screening Index and Domain Scores
Approach for assessing coastal vulnerability to oil sp
Spill Occurrence Model
Indicators of climate change and socia
ability in fishing dependent communities along the Eastern and Gulf Coasts of the United States
-scaled Resilience Screeni
ndex Results and Location Quotients for Socio-Ecological Characterizations in U.S. Coastal Regions
Digital Elevation Mo
Tidal Datums
Estuaries Coast
Coastal Topographic Lidar
National Flood Hazard Layer
American archives and climate change: risks and adaptation
Human Dimensions of Our Estuaries and Coas
oceaneconomics.org
cs: National Ocean Watch
NOAA
Long Range Forecasts
Clim Risk Manag
Federal Emergency Management Agency
IMPLAN group llc
National Ocean Economics Program Data
Sea Level Rise Inundation
Developments in Operational Long-Range Climate Prediction at CPC
IMPLAN: Economic Impact Analysis for Planning
lore Credit Score': Employer Credit Check Bans and Signal Substitution
Digital Forecast Database
Global Historical Climatology Network
U.S. Climate Divisions
Weather Forecast
Evolution of the
Economy: The Role of Working Waterfronts in the Alabama Gulf Coast
and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail
Lights Out: Climate Change Risk to Internet Infrastructure
Quarterly Census of Employment and Wages
Occupational Safety and Health in U.S. Aquaculture: A R
A Historical Perspective of U.S. Climate Divisions
Proceedings of the Applied Networking Research Workshop on - ANRW '18
US Bureau of Labor Statistics
The New Digital Forecast Database of the National Weather Service
The U.S. Historical Climatology Network Monthly Temperature Data, Version 2
Sustainability
Alternative Methods of Experience Rating Unemployment Insurance Employer Taxes
Abbreviations and Glossary
Sea Level Rise Analysis for Boston, Massachusetts
Household Food Consumption Patterns and Food Security among Low-Income Migrant Urban Farmers in Delhi, Jakarta, and Quito
Bull Am Meteorol Soc
Towards Sustainable Urban Spatial Structure: Does Decentr
n Reduc
muting Times?
Climate Central
Effects of Large-Scale Acquisition on Food Insecurity in Sierra Leone
National Scientific Funding for Interdisciplinary Research: A Comparison Study of Infectious Diseases in the US and EU

# Additional Information

Rich Context @ NYU Coleridge Initiative
**coleridgeinitiative.org/richcontext**

- **white paper**
- **upcoming book** (Jan 2020)
- **feedback/propose collaboration**

EDITED BY
JULIA I. LANE · IAN MULVANY · PACO NATHAN
RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS
BUILDING THE NEXT GENERATION OF SCHOLARLY INFRASTRUCTURE

**“Empty rhetoric over data sharing slows science”**
*Nature* (2017-06-12)

**“Experiences of the Deutsche Bundesbank”**
Stefan Bender
*CEMLA* (2019-05-28)

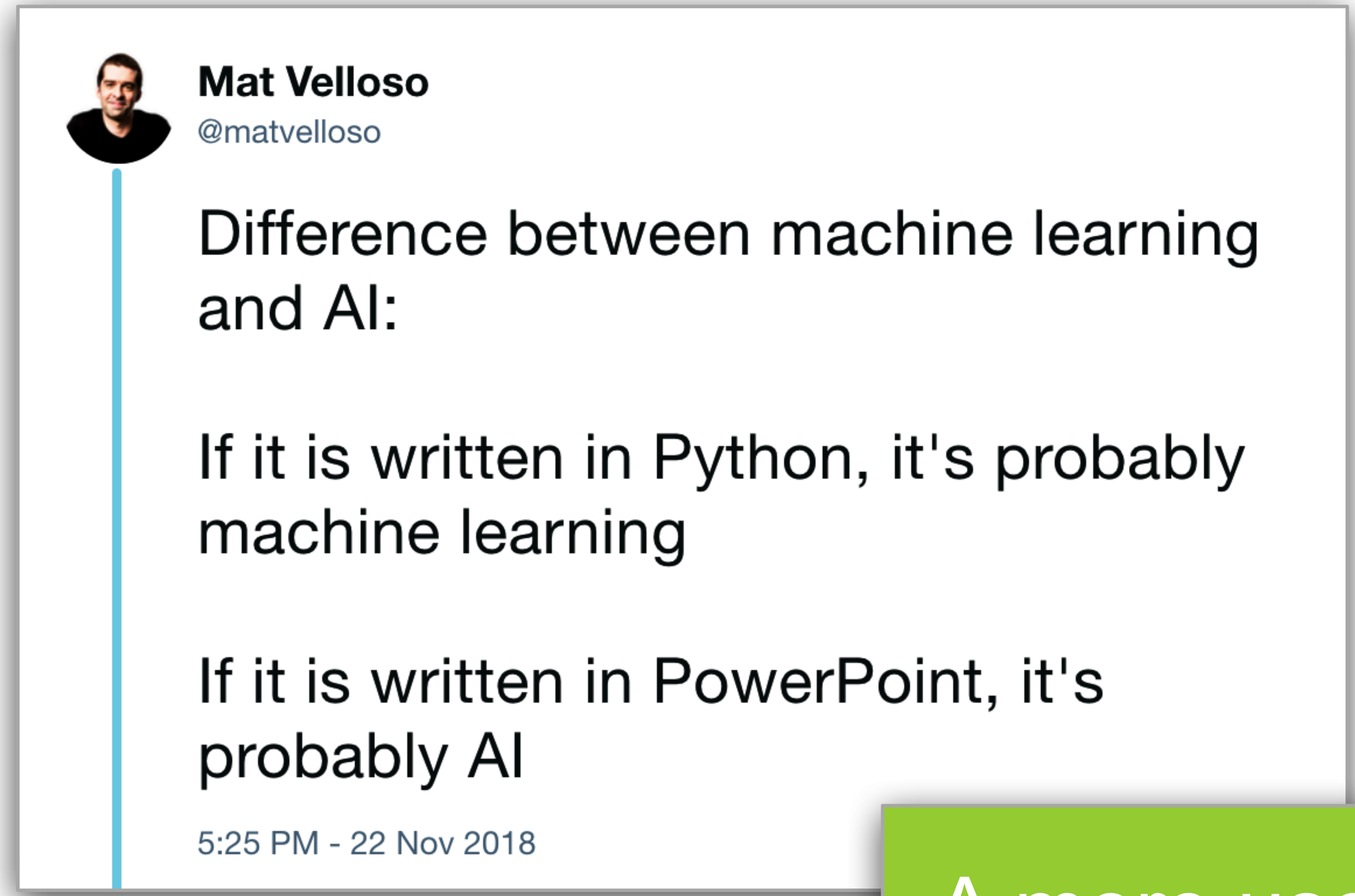
**“Where’s Waldo: Finding datasets in empirical research publications”**
Julia Lane
*AKBC* (2019-05-22)

**“Google data set search”**
Ian Mulvany
*ScholCommsBlog* (2019-11-19)

**“Impact for social science researchers”**
Ian Mulvany
*FORCE11* (2019-11-17)

# part 2:
# AI practices circa 2020 – perspectives from industry

# “Two Cultures” for AI



**Mat Velloso**
@matvelloso

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

5:25 PM - 22 Nov 2018

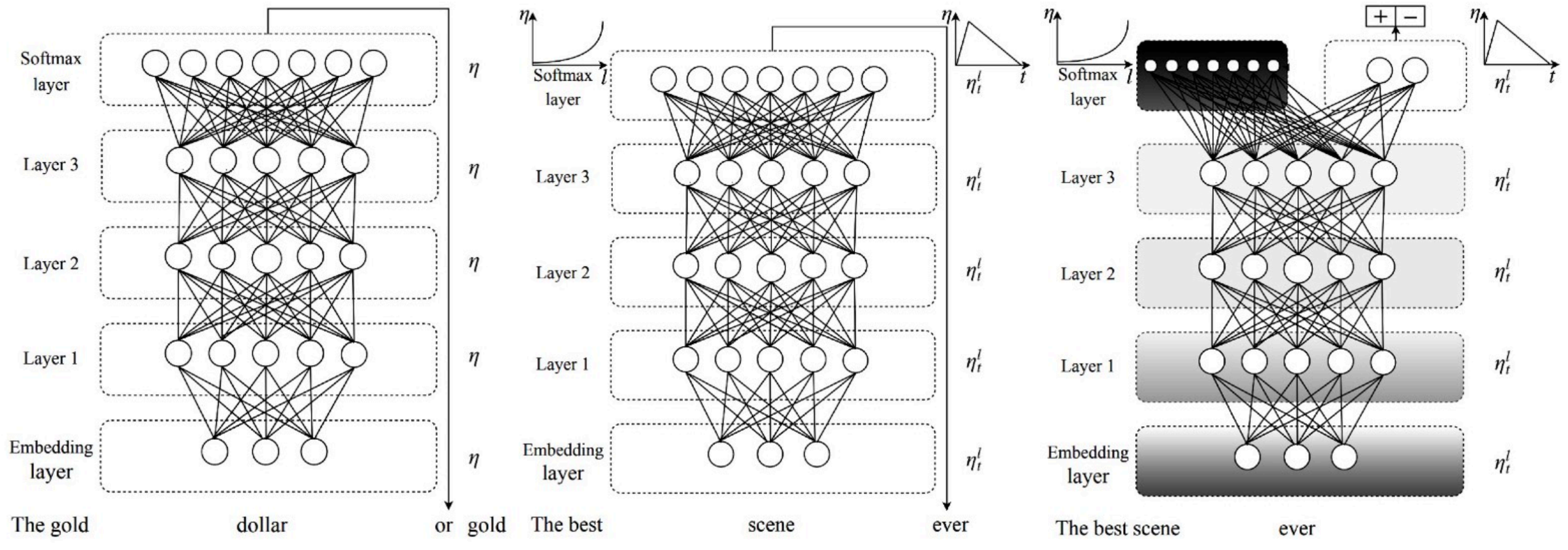
A more useful distinction:

- **ML** is about the tools and technologies
- **AI** is about use case impact on social systems

# So many ways to use optimized gradients…

AI finds pervasive use throughout industry: surpassing human benchmarks – although it’s probably better to think in terms of **how to augment social systems**:

- *deep learning*; **Krizhevsky 2012**
- *reinforcement learning*; **Silver 2017**
- *transformers* (NLP); **Radford 2019**
- *weak supervision*; **Ratner 2017**

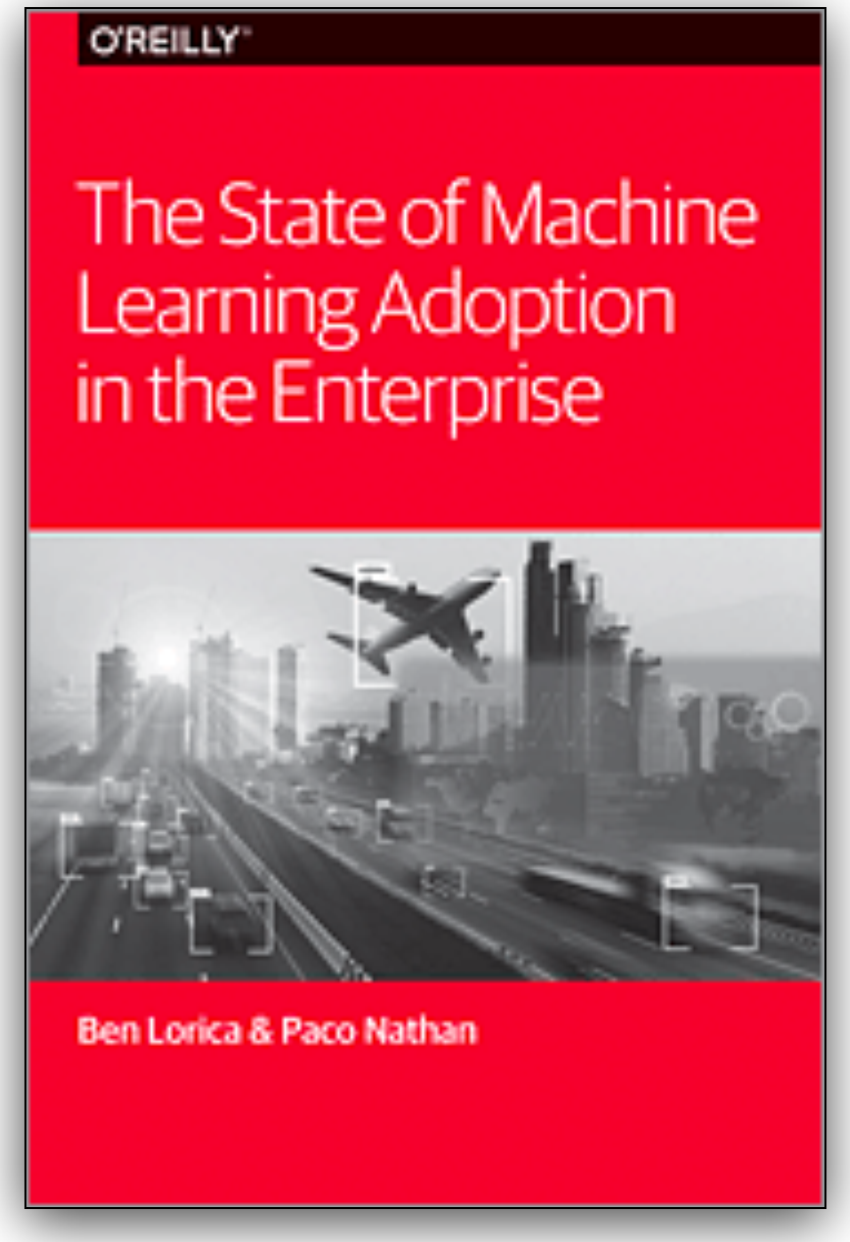


(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

As well as: *self-supervision*, *knowledge graph, transfer learning*, *active learning* (“human-in-the-loop”), and so on…

# Industry surveys for AI and Cloud adoption

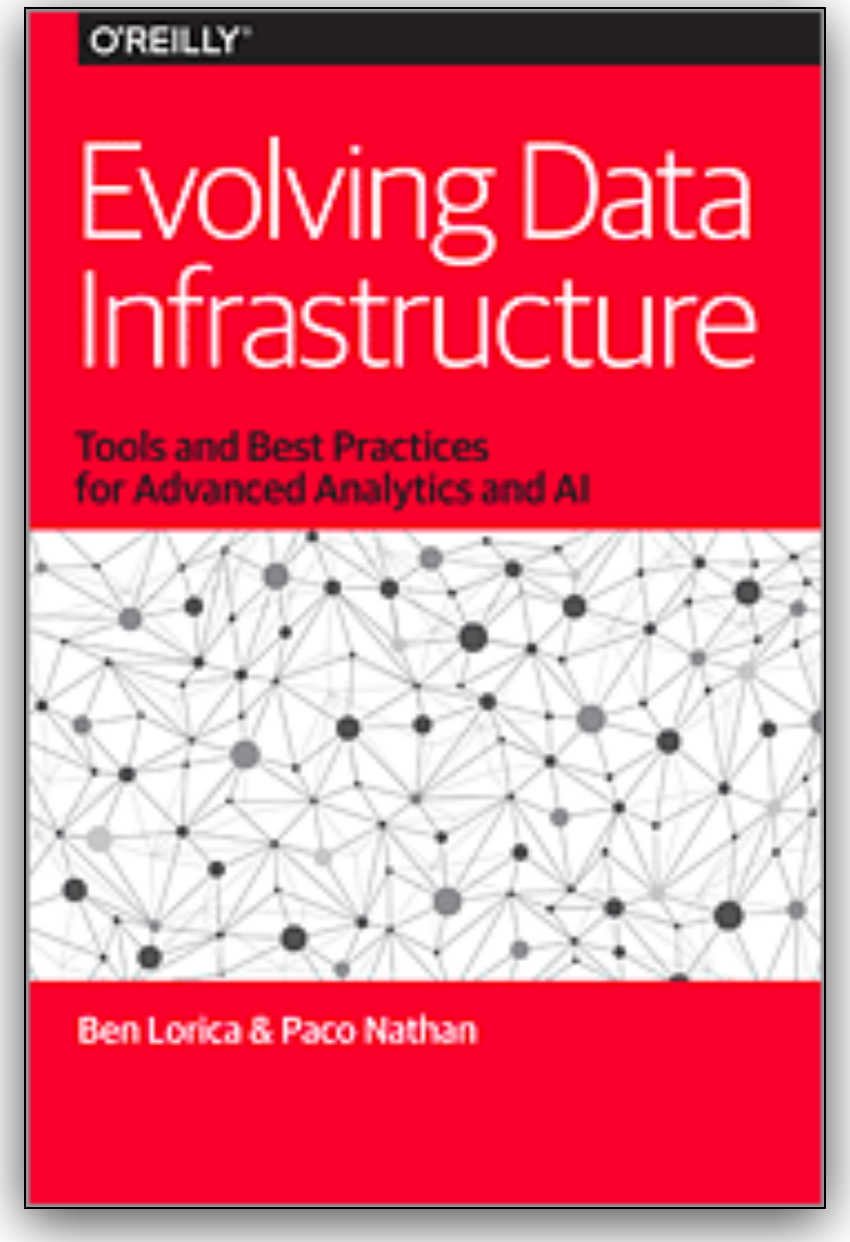
- “Three surveys of AI adoption reveal key advice from more mature practices”
  **Ben Lorica**, **Paco Nathan** *O’Reilly Media* (2019-02-20)
- Episode 7, *Domino*: surveying “ABC” adoption in enterprise (2019-03-03)



O'REILLY

The State of Machine Learning Adoption in the Enterprise

Ben Lorica & Paco Nathan

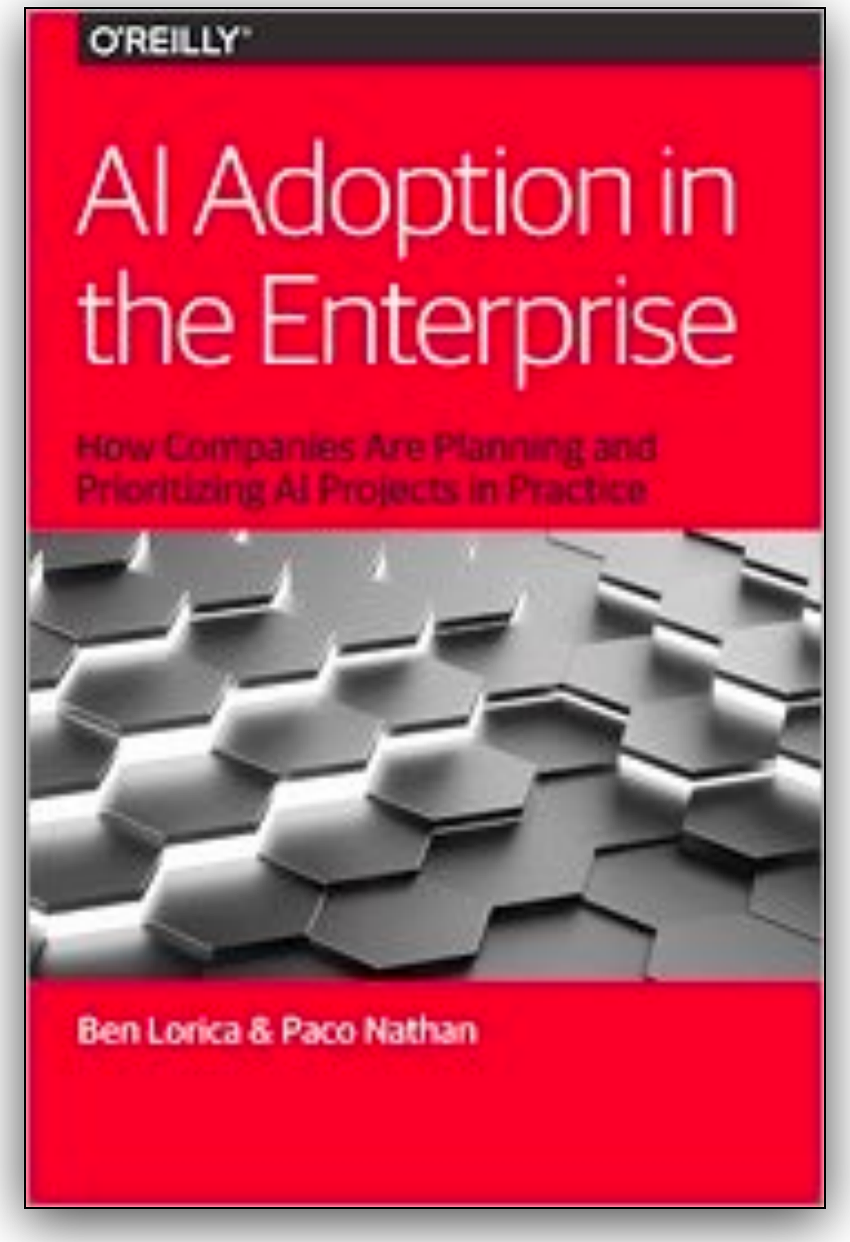


O'REILLY

Evolving Data Infrastructure

Tools and Best Practices for Advanced Analytics and AI

Ben Lorica & Paco Nathan



O'REILLY

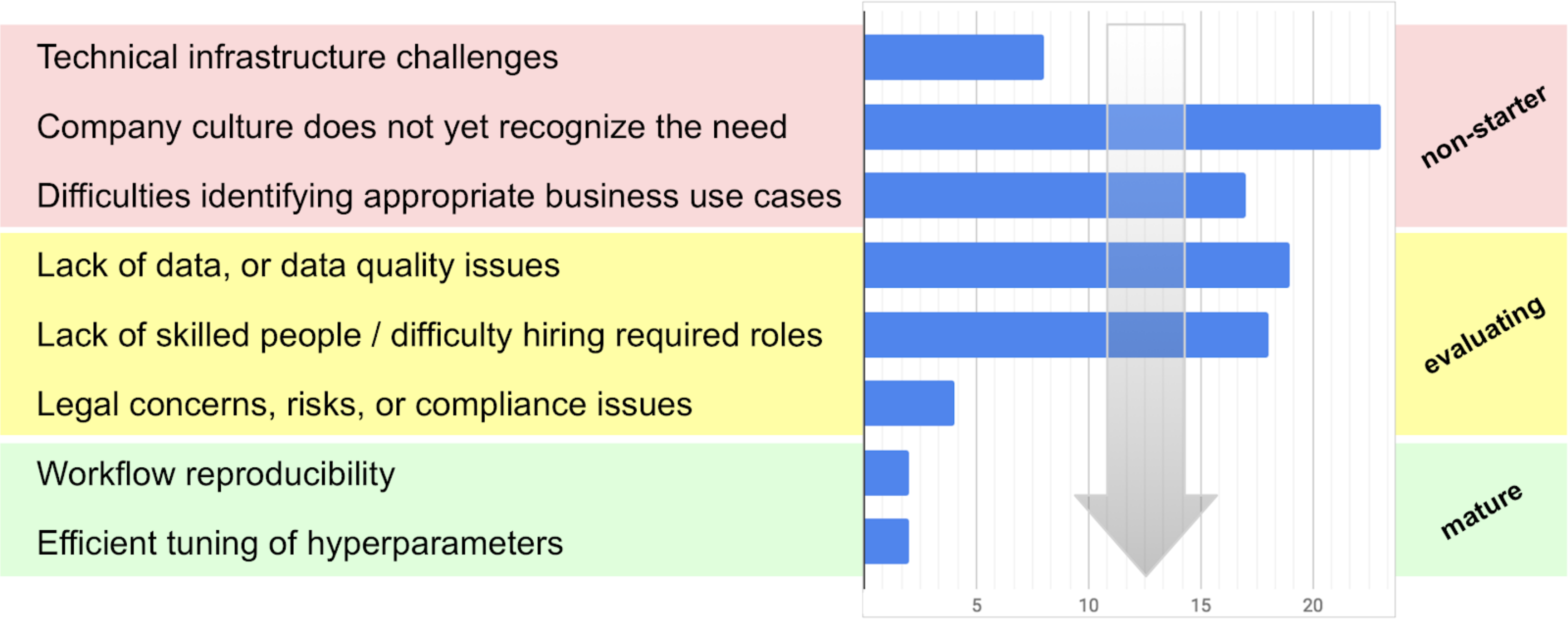
AI Adoption in the Enterprise

How Companies Are Planning and Prioritizing AI Projects in Practice

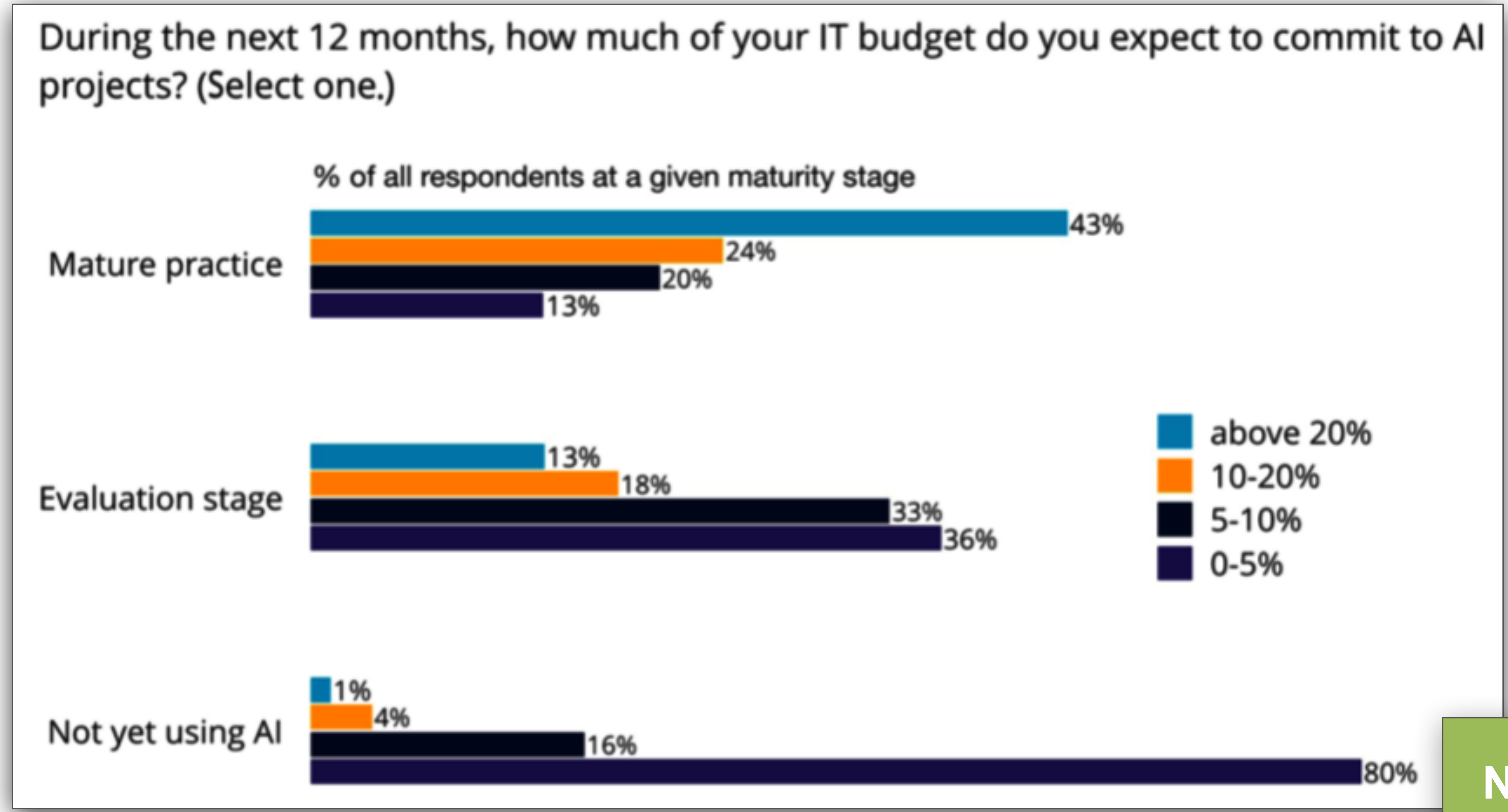
Ben Lorica & Paco Nathan

# Challenges in “ABC” adoption in enterprise

Leading risks for “ABC” (AI, Big Data, Cloud) adoption in enterprise, based on ORM surveys – viewed as a “survival analysis”:



# An accelerating gap in AI funding



During the next 12 months, how much of your IT budget do you expect to commit to AI projects? (Select one.)

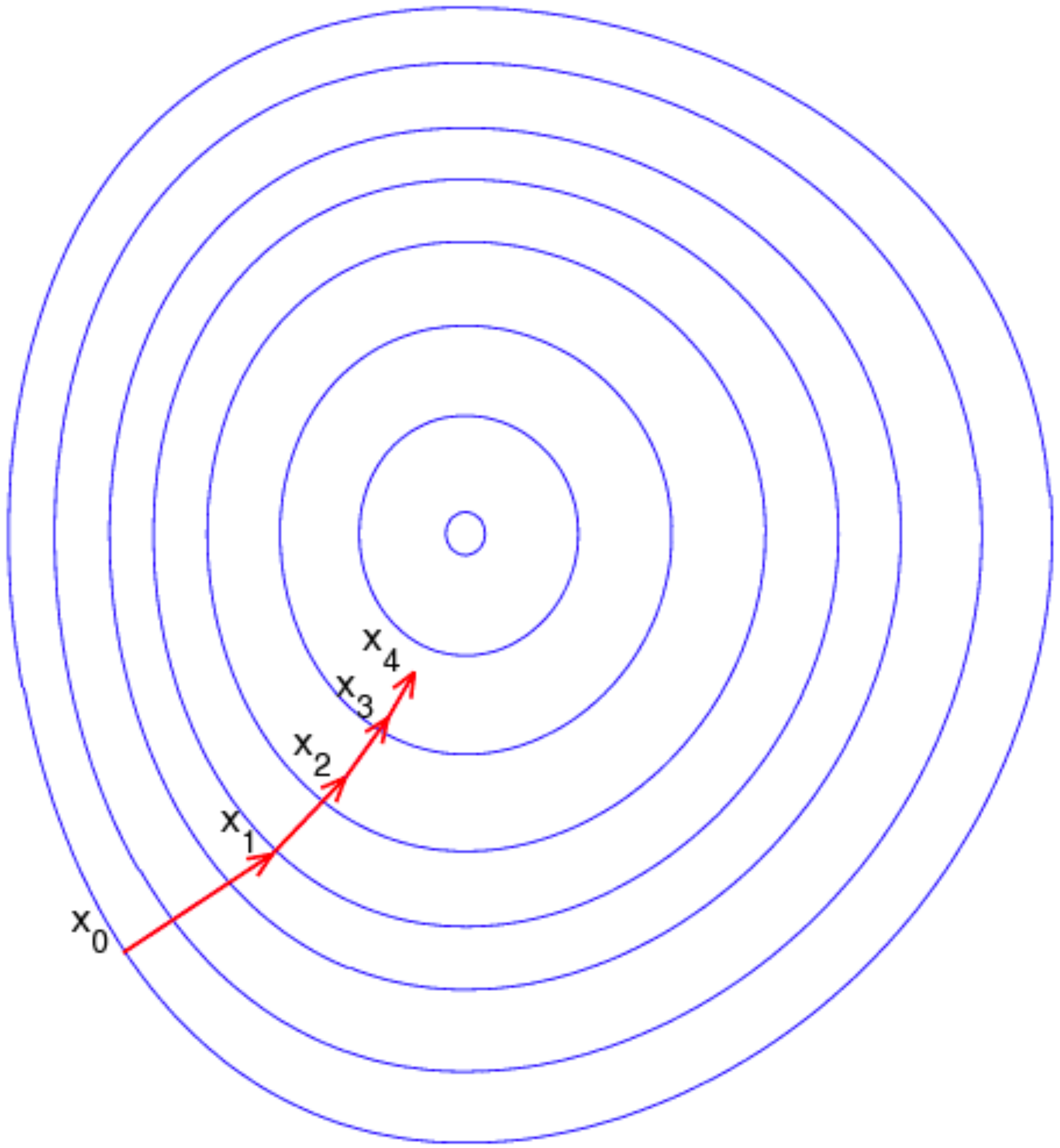
**2020 spending expected to increase 31%** – *WSJ*

**Note: firms with early advantage are investing more, moving still further away from the pack.**

# Concerns…

Large investments by **hyperscalers** which own public clouds, have access to enormous amounts of training data, and hired large teams of AI talent … are now sold back through a range of AutoML services.

**TensorFlow Dev Summit 2018**
Jeff Dean: “Can we replace scarce ML expertise with 100x computation?”



# Concerns...

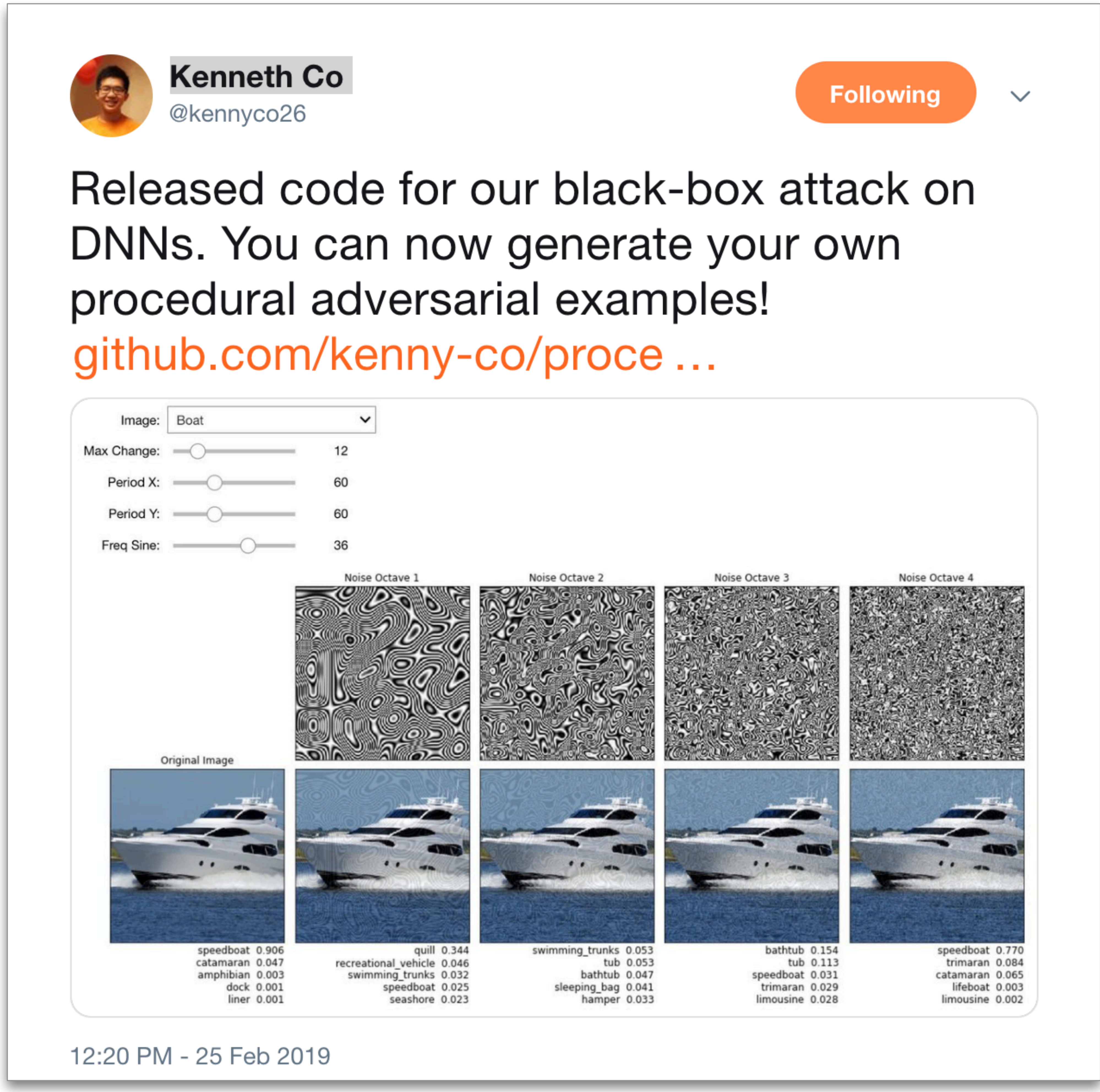
Potential *attack surfaces* are getting exposed much more rapidly than they're becoming understood.

Contemporary work in data privacy technologies **injects noise** into data streams – mirrored by generative AI efforts creating *security exploits* that inject carefully designed noise into data streams:

**github.com/kenny-co/procedural-advml**

The math of DL is not clearly understood, e.g., why don't we encounter local minima more frequently? Advanced work in numerical analysis is exploring these phenomena:

**"Understanding deep neural networks"**
**Michael Mahoney**



Kenneth Co
@kennyco26

Released code for our black-box attack on DNNs. You can now generate your own procedural adversarial examples!
github.com/kenny-co/proce ...

Image: Boat
Max Change: 12
Period X: 60
Period Y: 60
Freq Sine: 36

```
speedboat 0.906          quill 0.344     swimming_trunks 0.053   bathtub 0.154     speedboat 0.770
catamaran 0.047          recreational_vehicle 0.046   tub 0.053    tub 0.113     trimaran 0.084
amphibian 0.003          swimming_trunks 0.032   bathtub 0.047   speedboat 0.031   catamaran 0.065
dock 0.001               speedboat 0.025   sleeping_bag 0.041    trimaran 0.029    lifeboat 0.003
liner 0.001              seashore 0.023    hamper 0.033          limousine 0.028   limousine 0.002
```

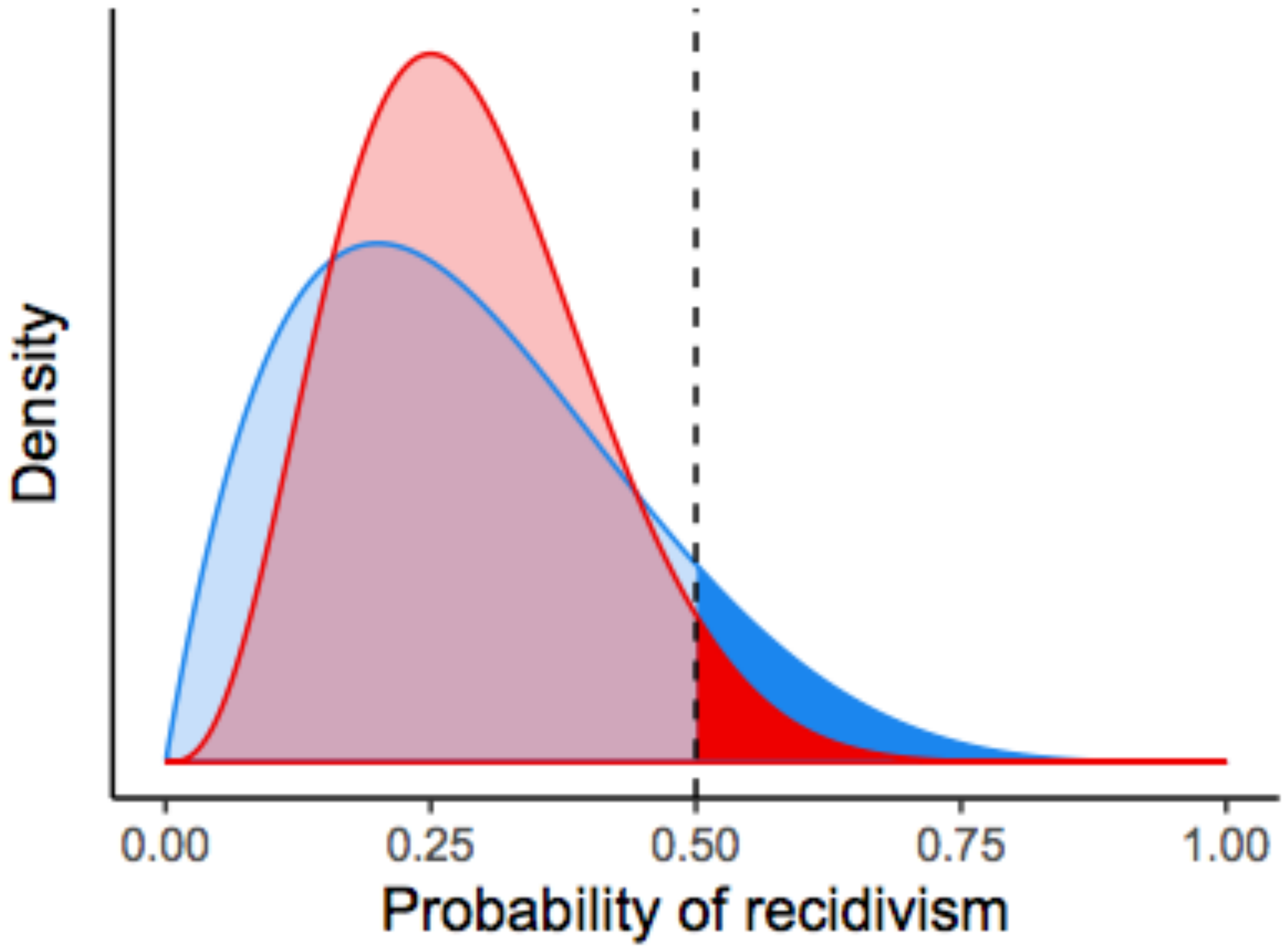
12:20 PM - 25 Feb 2019

# Concerns…

Understanding fairness and bias in use of data is a provably hard problem:

**“Why it’s hard to design fair machine learning models”**
**Sharad Goel**, **Sam Corbett-Davies**

recommended: open source **AIF360** toolkit
**http://aif360.mybluemix.net/**



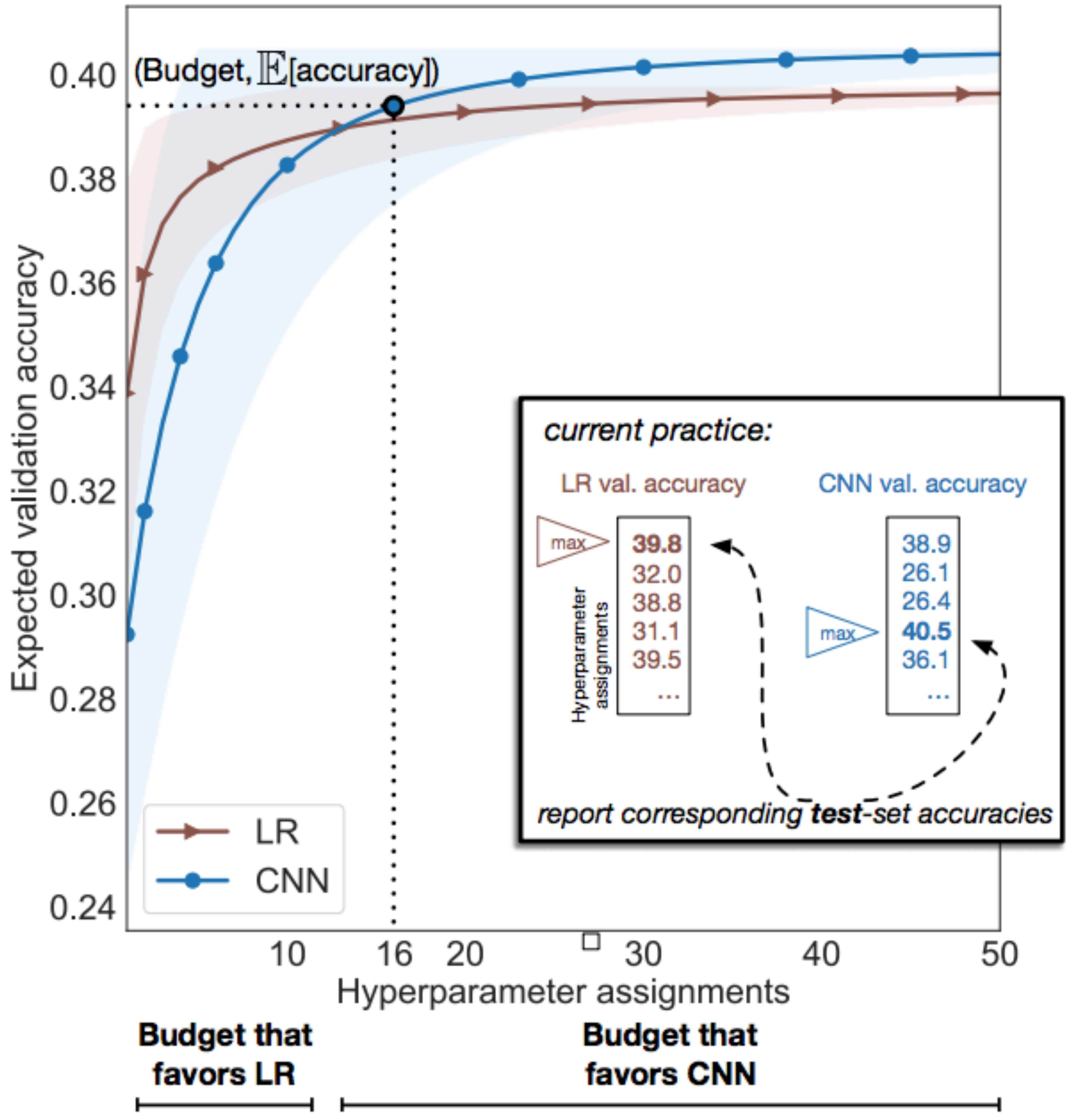
# Concerns…

Results in widely cited papers are too expensive for most organizations (except **hyperscalers)** to reproduce…

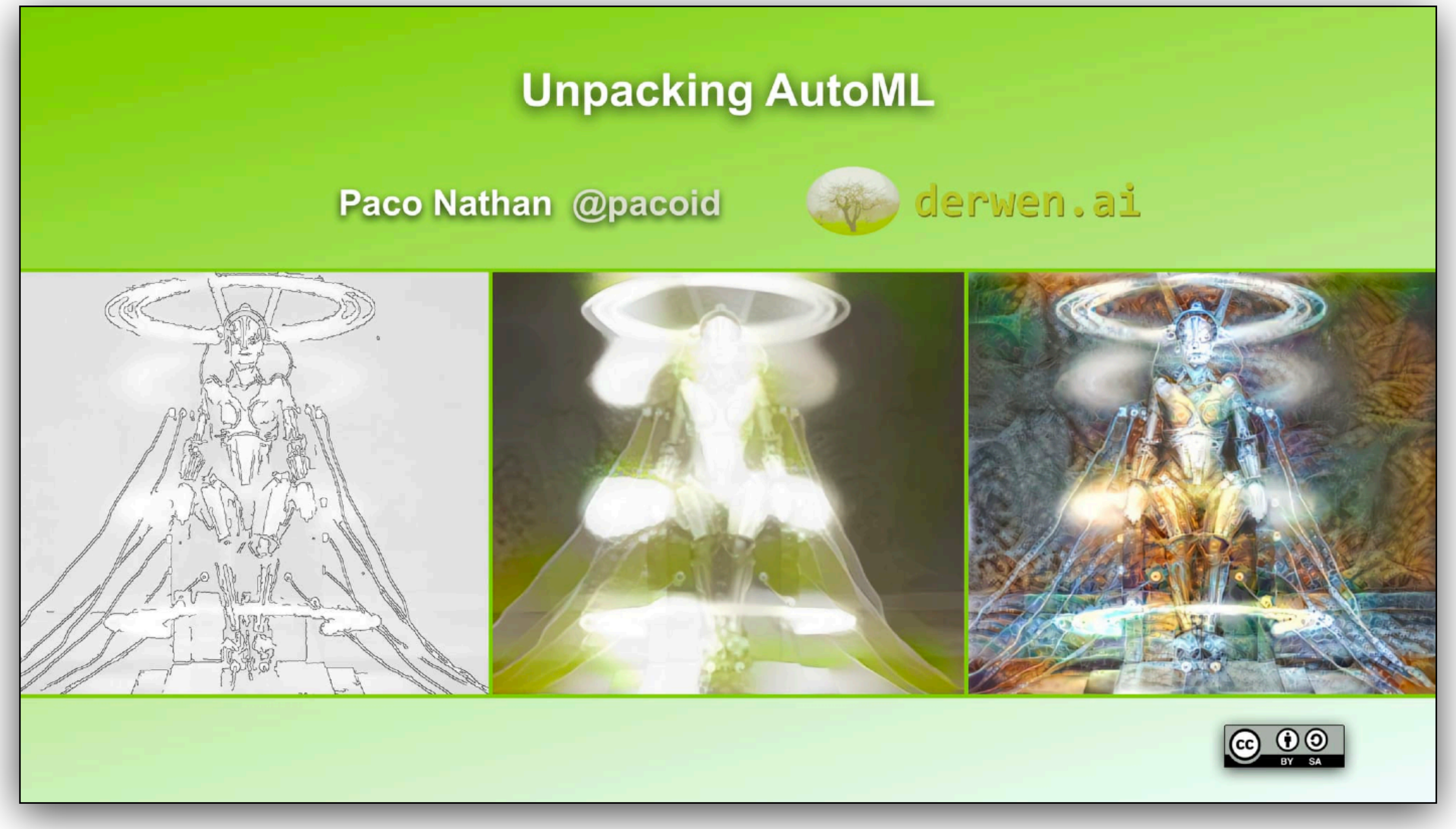
**“Show Your Work: Improved Reporting of Experimental Results”**
**Jesse Dodge**, **Suchin Gururangan**, **Dallas Card**, **Roy Schwartz**, **Noah A. Smith**

Policy: training a *transformer* model (GPT-2 w/ NAS) requires 5x the carbon footprint of operating an automobile over its lifetime…

**“Energy and Policy Considerations for Deep Learning in NLP”**
**Emma Strubell**, **Ananya Ganesh**,
**Andrew McCallum**



# Unpacking AutoML

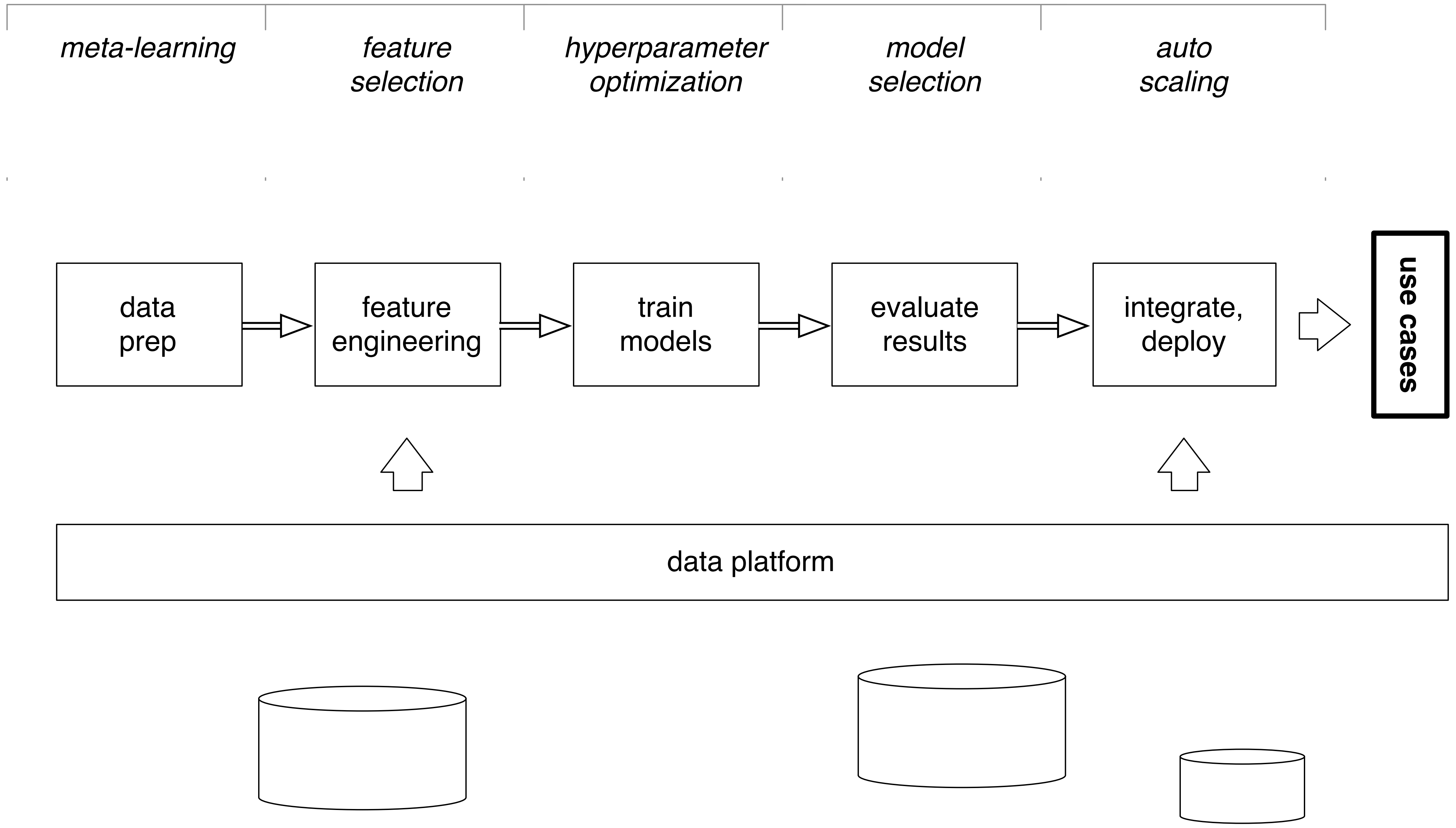


Unpacking AutoML

Paco Nathan @pacoid

**derwen.ai/s/yvkg**

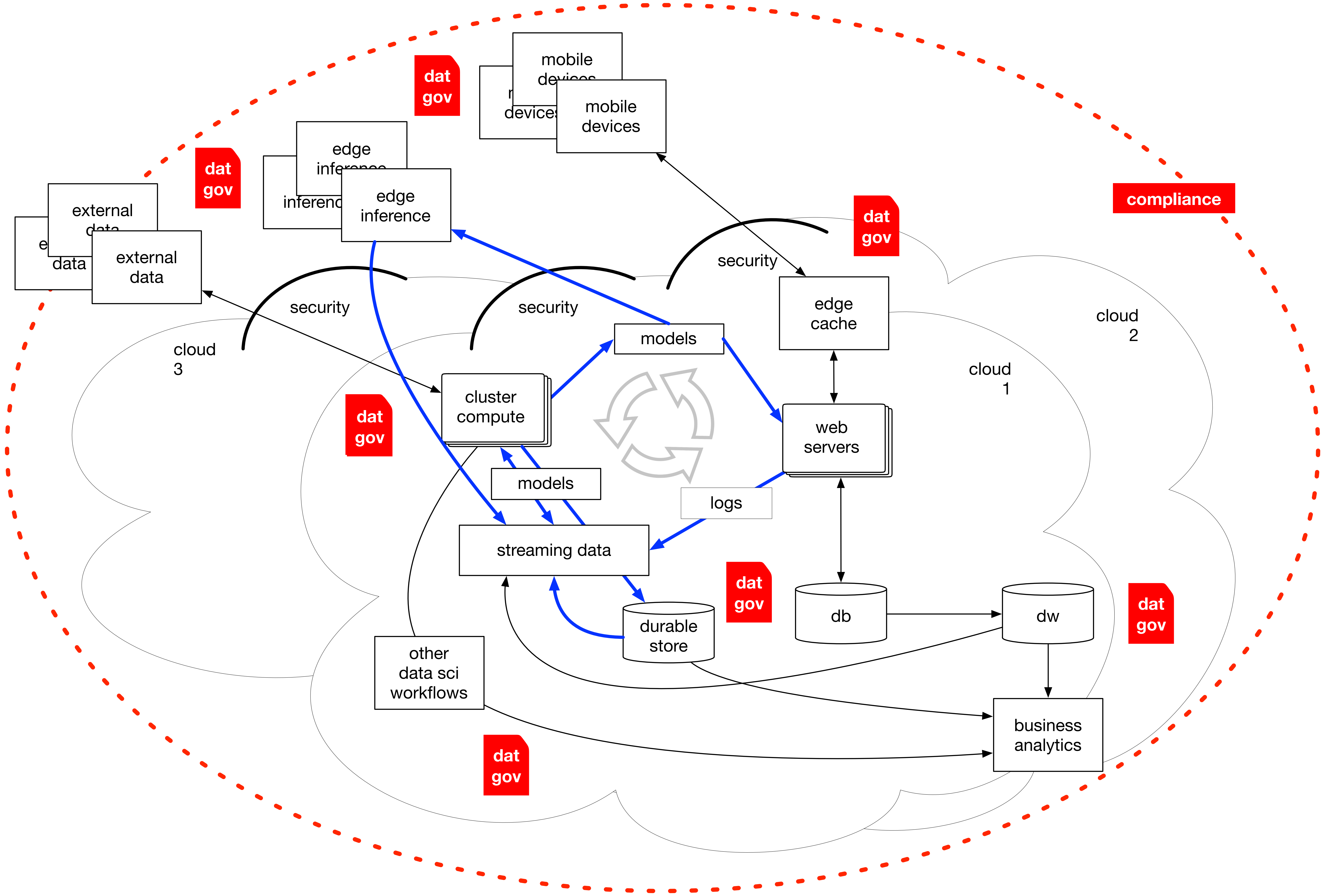
*we noted an uptick in adoption for a third aspect, co-evolving along with DG and* ***MLOps***



# Overview of Data Governance

Overview of Data Governance

Paco Nathan @pacoid derwen.ai



**derwen.ai/s/6fqt**

*we noted a resurgence in data governance – this report examines key themes, vendors, issues, etc.*

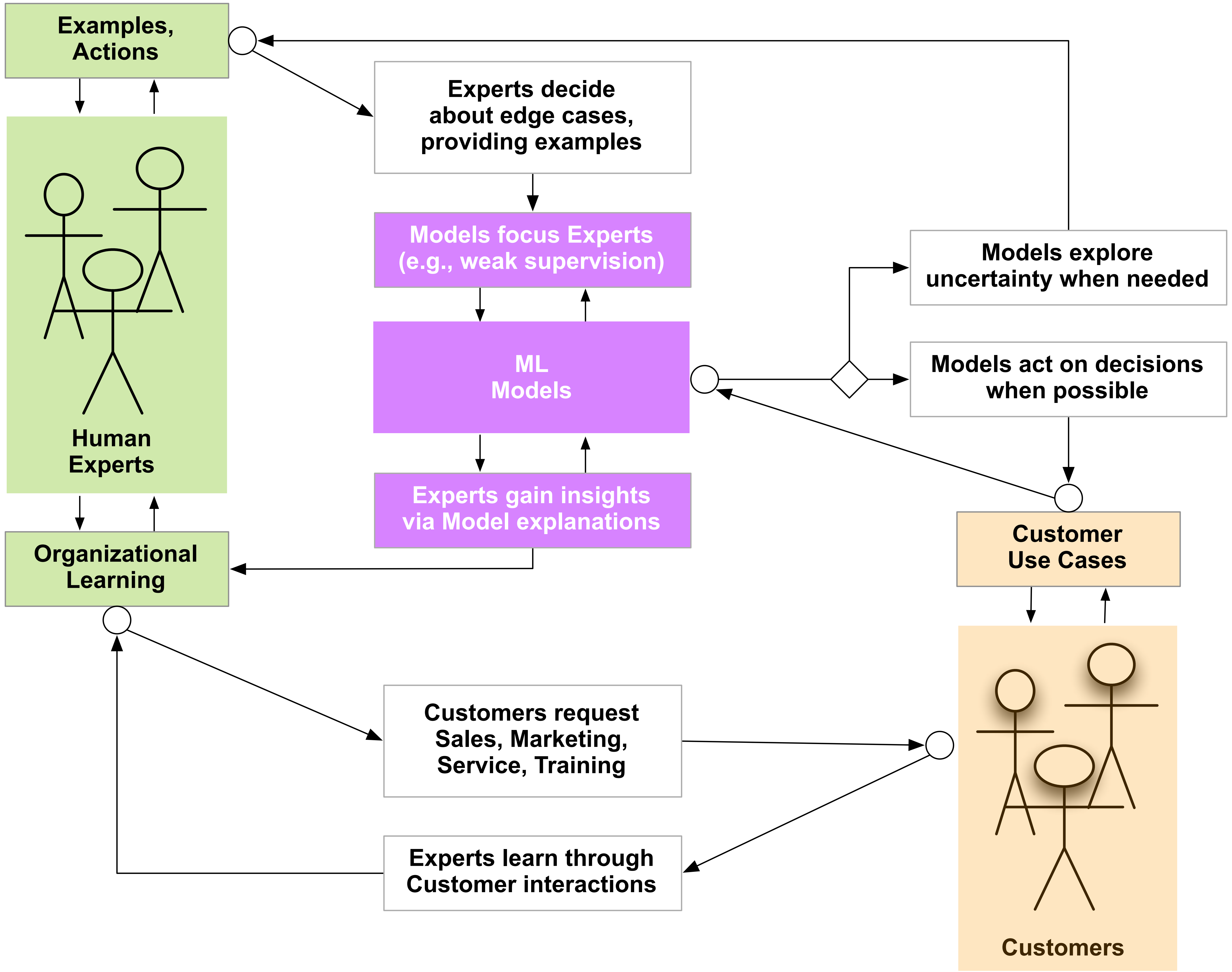
# Active Learning as a Data Strategy

Active Learning Use Cases

Paco Nathan @pacoid derwen.ai

**derwen.ai/s/d8b7**

*teams of people + machines, leveraging the complementary strengths of both*



Examples, Actions

Human Experts

Organizational Learning

Experts decide about edge cases, providing examples

Models focus Experts (e.g., weak supervision)

ML Models

Experts gain insights via Model explanations

Models explore uncertainty when needed

Models act on decisions when possible

Customer Use Cases

Customers

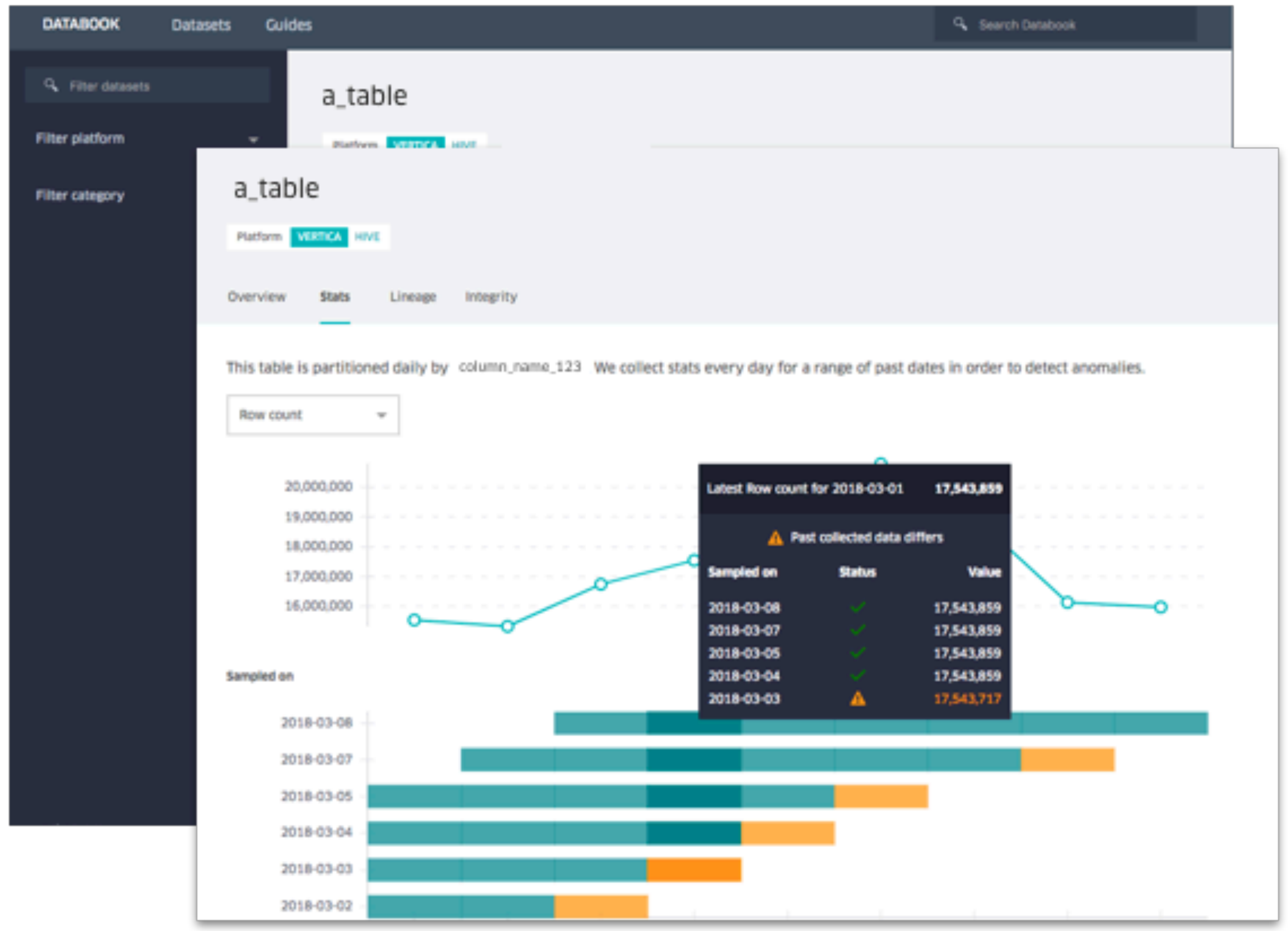
Customers request Sales, Marketing, Service, Training

Experts learn through Customer interactions

# Emerging category: watch the “AI Natives”

Projects (mostly OSS) that leverage **knowledge graph** of metadata about datasets and their usage:

- **Amundsen** @ Lyft
  *data discovery and metadata*
- **Data Hub** @ LinkedIn
  *data discovery and lineage*
- **Marquez** @ Stitch Fix
  *collect, aggregate, visualize metadata*
- **UMS** @ Uber
  *manage metadata about datasets*
- **Metcat** @ Netflix
  *data discovery, metadata service*
- **Dataportal** @ Airbnb
  *integrated data-space (not OSS)*



This table is partitioned daily by column_name_123 We collect stats every day for a range of past dates in order to detect anomalies.

Latest Row count for 2018-03-01 **17,543,859**

⚠ Past collected data differs

| Sampled on | Status | Value |
|---|---|---|
| 2018-03-08 | ✓ | 17,543,859 |
| 2018-03-07 | ✓ | 17,543,859 |
| 2018-03-05 | ✓ | 17,543,859 |
| 2018-03-04 | ✓ | 17,543,859 |
| 2018-03-03 | ⚠ | 17,543,717 |

# part 3:
# watch this closely –
# rapid evolution of hardware

# Hardware in perspective

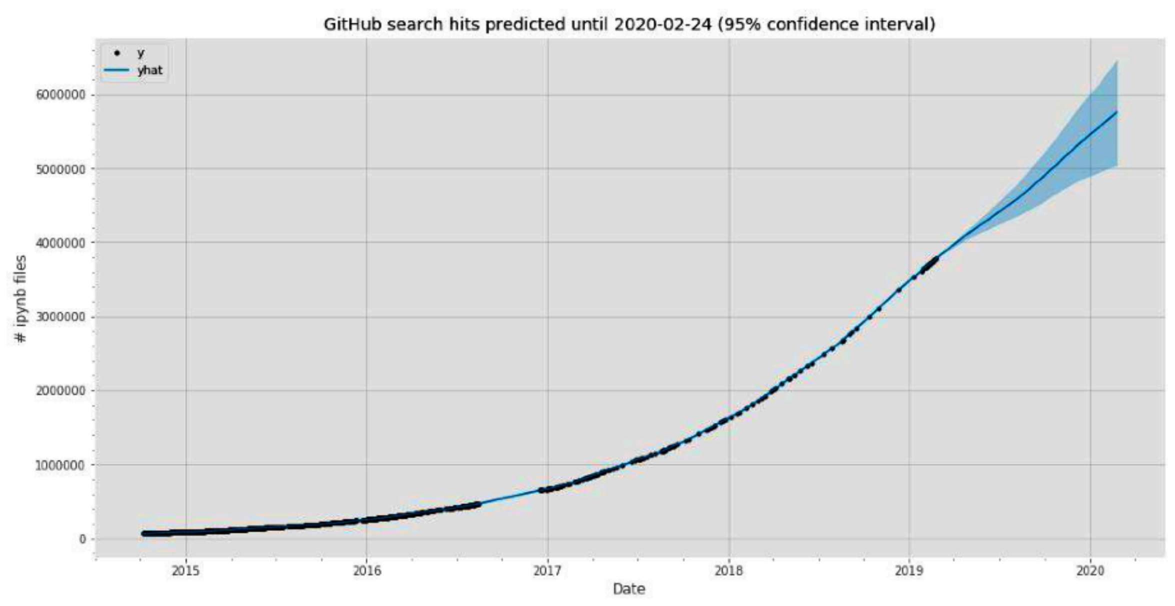
An **emerging trend** disrupts the past 15-20 years of software engineering practice:

**hardware > software > process**

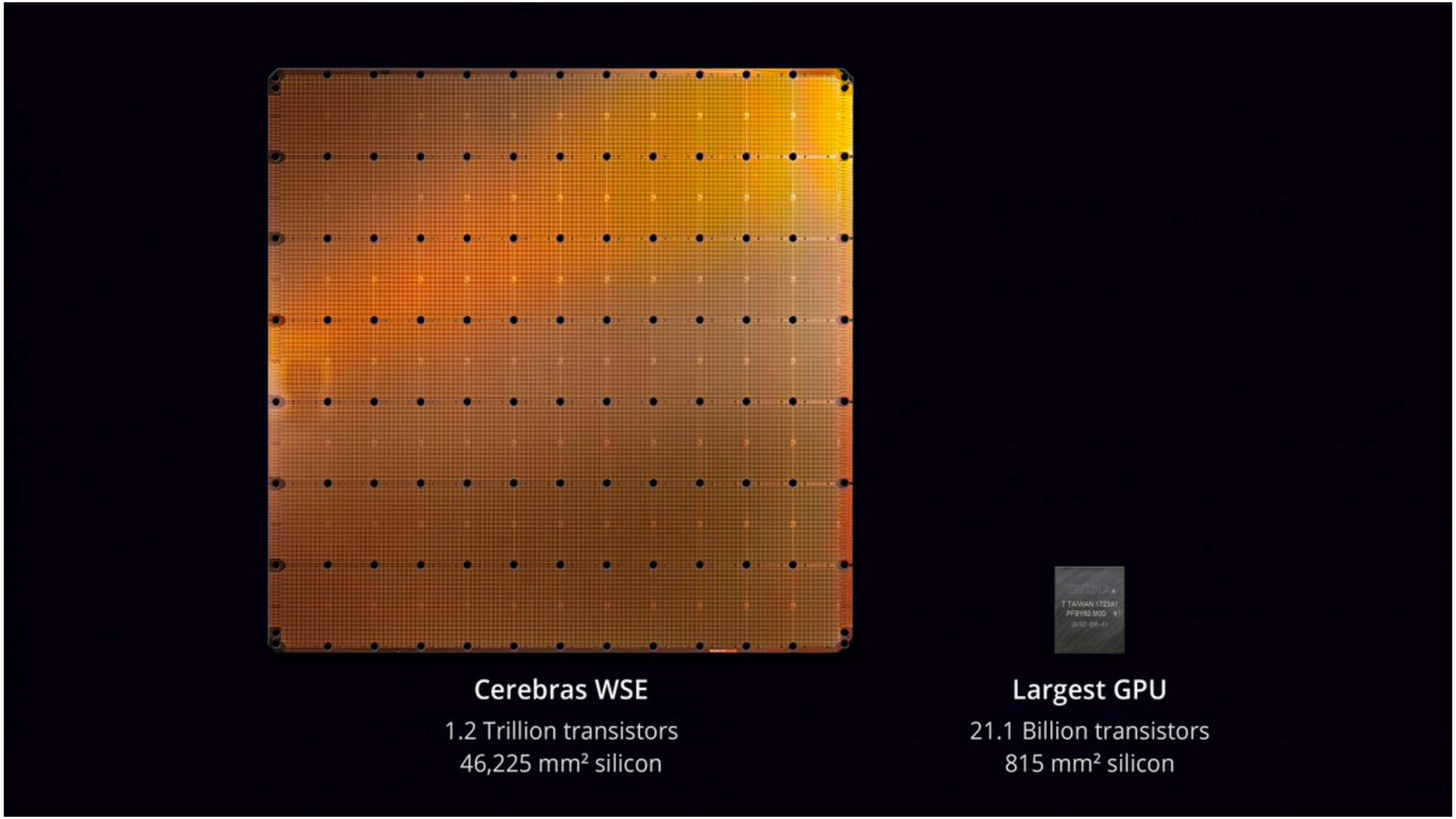
Hardware is now evolving more rapidly than software, which is evolving more rapidly than effective process

**Moore's Law** is all but dead, although ironically many inefficiencies grew to be based on it

**Project Jupyter**, **Apache Arrow**, **NumPyWren** and the related **Ray** are emblematic for data infrastructure transformation in enterprise



# ASICs the size of iPads, for ML



**Cerebras WSE**

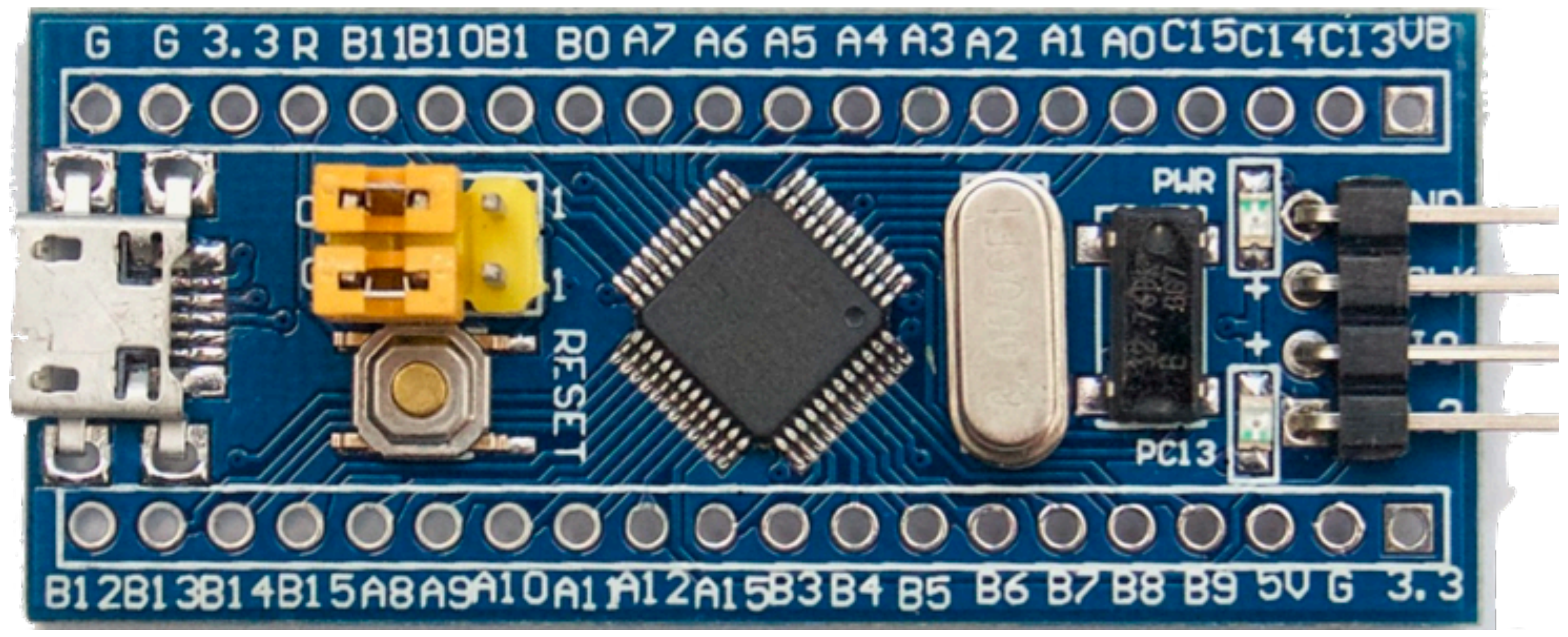
1.2 Trillion transistors
46,225 mm² silicon

**Largest GPU**

21.1 Billion transistors
815 mm² silicon

# Impact of edge devices and low-power ML

- “EIE: Efficient Inference Engine on Compressed Deep Neural Network”
  **Song Han**, et al., (2016-05-03)
- **Alasdair Allan**, Babilim Light Industries:
  The intelligent edge and the demise of big data?
  (2019-05-02)
- **Pete Warden**, Google:
  TensorFlow.js on low-power devices
  (2018-10-10)
- **tinyML Summit**, 2020 Feb 12-13, San Jose
  https://www.tinymlsummit.org/


G G 3.3 R B11 B10 B1 B0 A7 A6 A5 A4 A3 A2 A1 A0 C15 C14 C13 VB
PWR
RESET
PC13
B12 B13 B14 B15 A8 A9 A10 A11 A12 A15 B3 B4 B5 B6 B7 B8 B9 5V G 3.3

# Implications of rapid hardware evolution

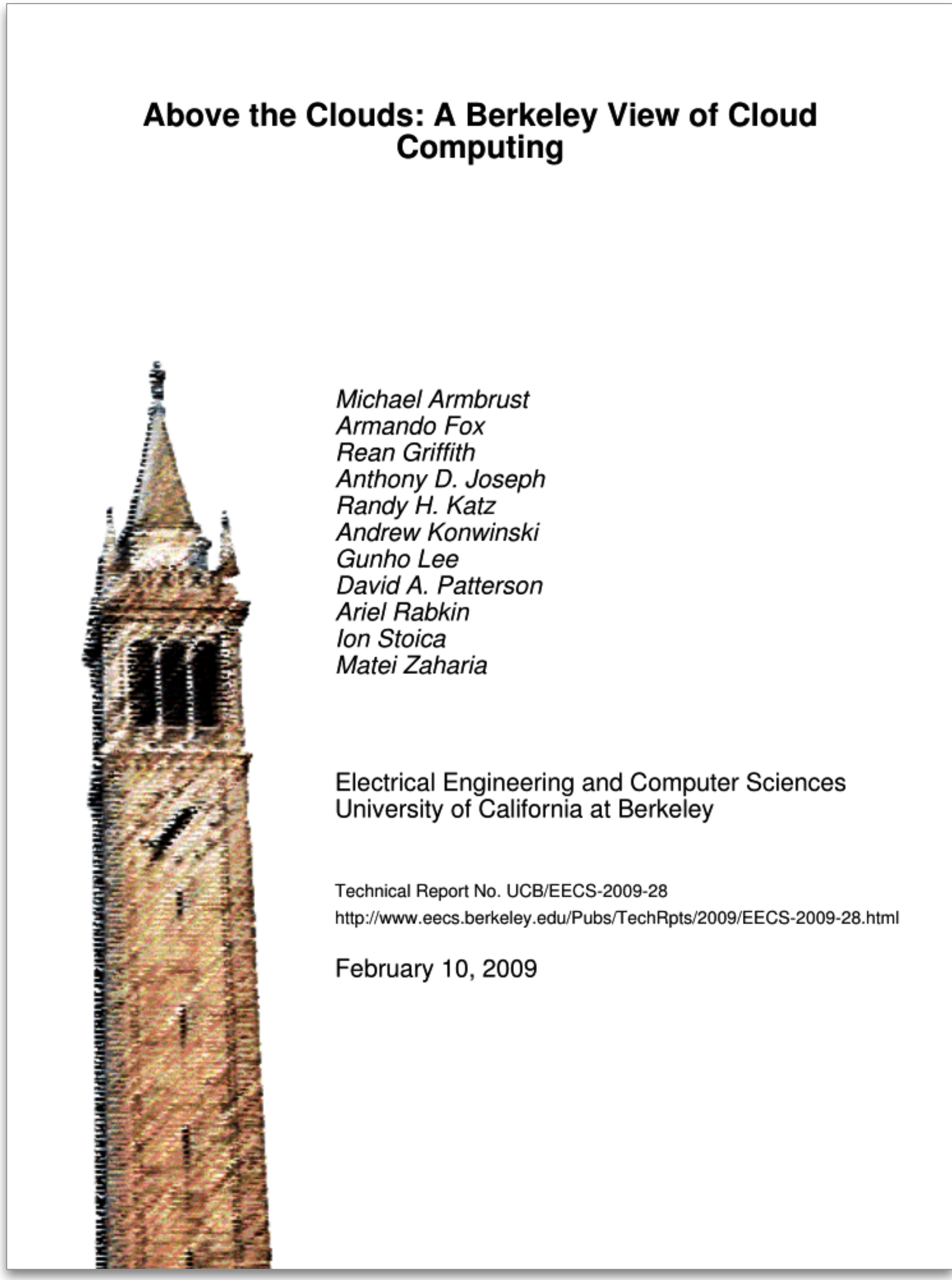
- ~6 firms (hyperscaler cartel) now drive the **Intel** roadmap … and extract most of the value?
- innovation cost curves have gone exponential, with side-effects becoming difficult to anticipate
- $4-10B per fab now, with potentially < 2 yrs utility?
- geopolitically orchestrated IP theft + cyberattacks:
  - undercutting R&D investments in the US
  - competitive pressure on US public cloud vendors

# Evolution of cloud patterns

UC Berkeley published a **2009 report** about early use cases for cloud computing, which foresaw the shape of industry deployments over much of the next decade, and led directly to **Apache Mesos** and **Apache Spark**

It's fascinating to study the **contrasts** between that 2009 report and its 2019 follow-up.

(minor footnote: **vimeo.com/3616394**)



**Above the Clouds: A Berkeley View of Cloud Computing**

*Michael Armbrust*
*Armando Fox*
*Rean Griffith*
*Anthony D. Joseph*
*Randy H. Katz*
*Andrew Konwinski*
*Gunho Lee*
*David A. Patterson*
*Ariel Rabkin*
*Ion Stoica*
*Matei Zaharia*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-28
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

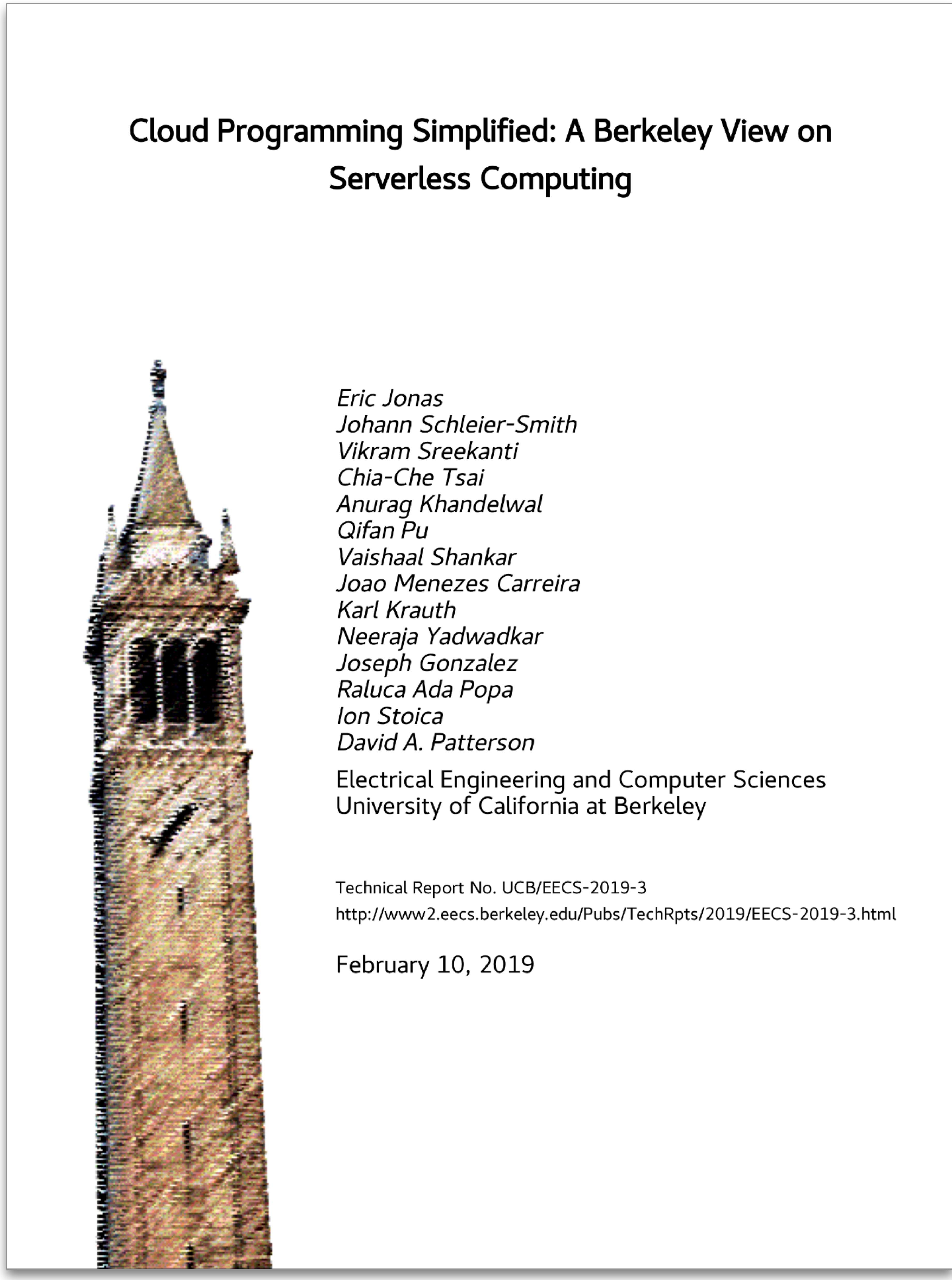
February 10, 2009

**2009**

# Evolution of cloud patterns

Early patterns of cloud use mirrored the **virtualization** familiar to enterprise firms in their existing on-prem architectures, mostly as a convenience – *for migration*

More **contemporary patterns** force restructuring – *for efficiency and security* – **decoupling computation and storage**



Cloud Programming Simplified: A Berkeley View on Serverless Computing

*Eric Jonas*
*Johann Schleier-Smith*
*Vikram Sreekanti*
*Chia-Che Tsai*
*Anurag Khandelwal*
*Qifan Pu*
*Vaishaal Shankar*
*Joao Menezes Carreira*
*Karl Krauth*
*Neeraja Yadwadkar*
*Joseph Gonzalez*
*Raluca Ada Popa*
*Ion Stoica*
*David A. Patterson*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-3
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-3.html

February 10, 2019

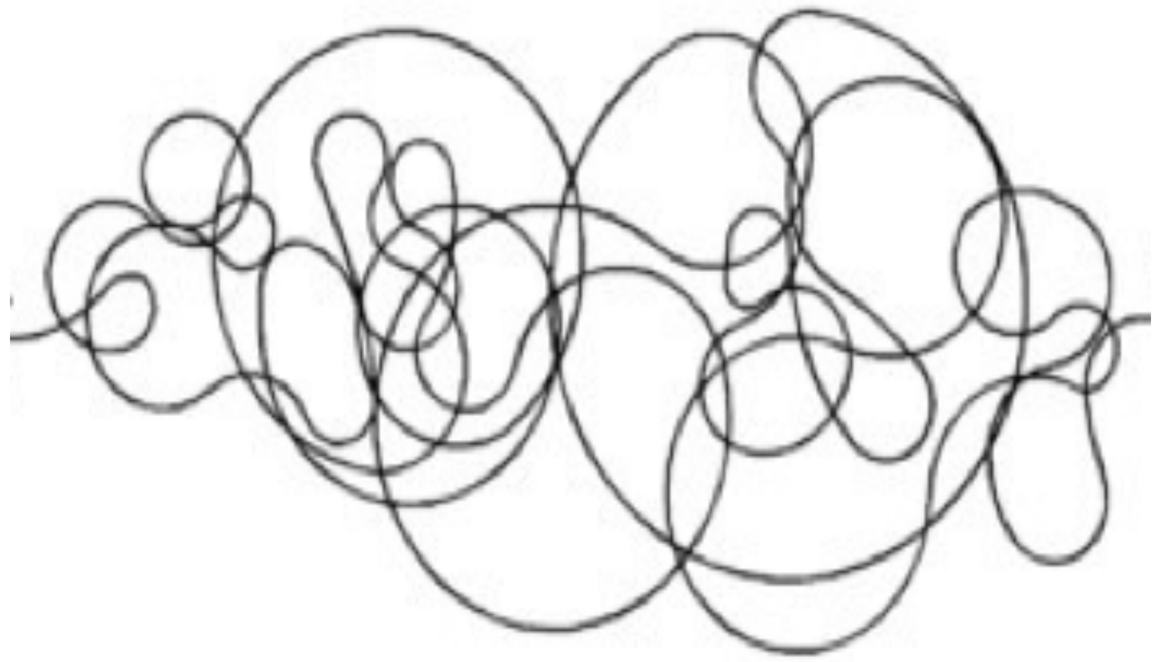
**2019**

# Key takeaways from “A Berkeley View”

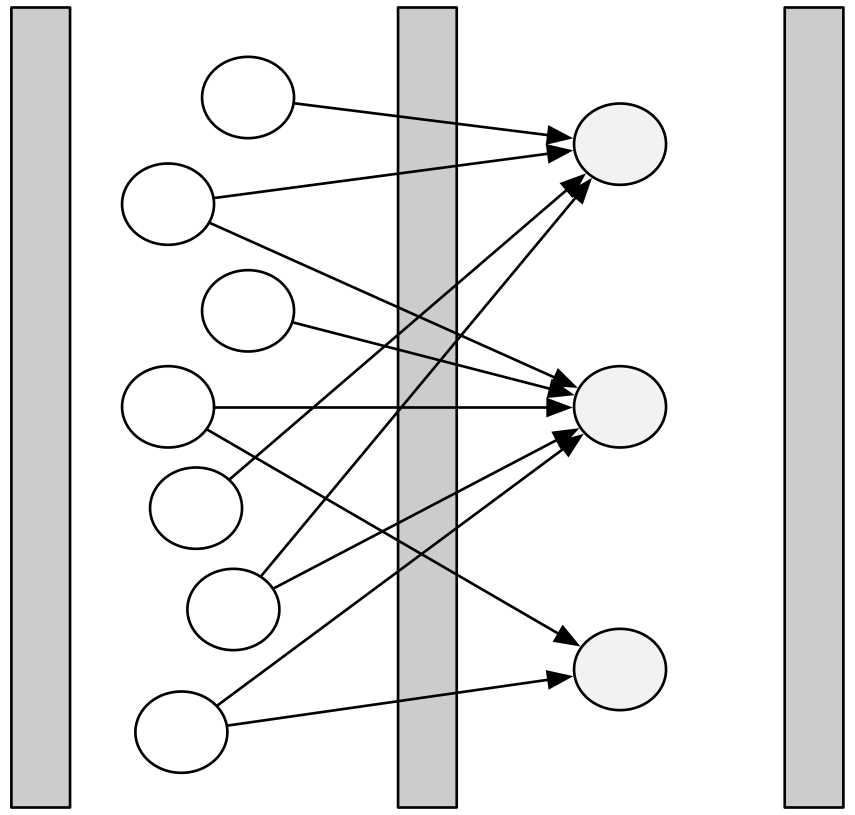
- physics + cloud economics imply less “framework” layers
- service offerings **migrate up** the tech stack
- issues with **data gravity**, increased **vendor lock-in**

- also, we cannot overstate the importance of **data integrity**
  - the risks tend to **undermine science**
  - questions of **global competitiveness** for training ML models; do we prioritize having volumes of training data or providing carefully **curated test sets**? (compare: China, US, EU

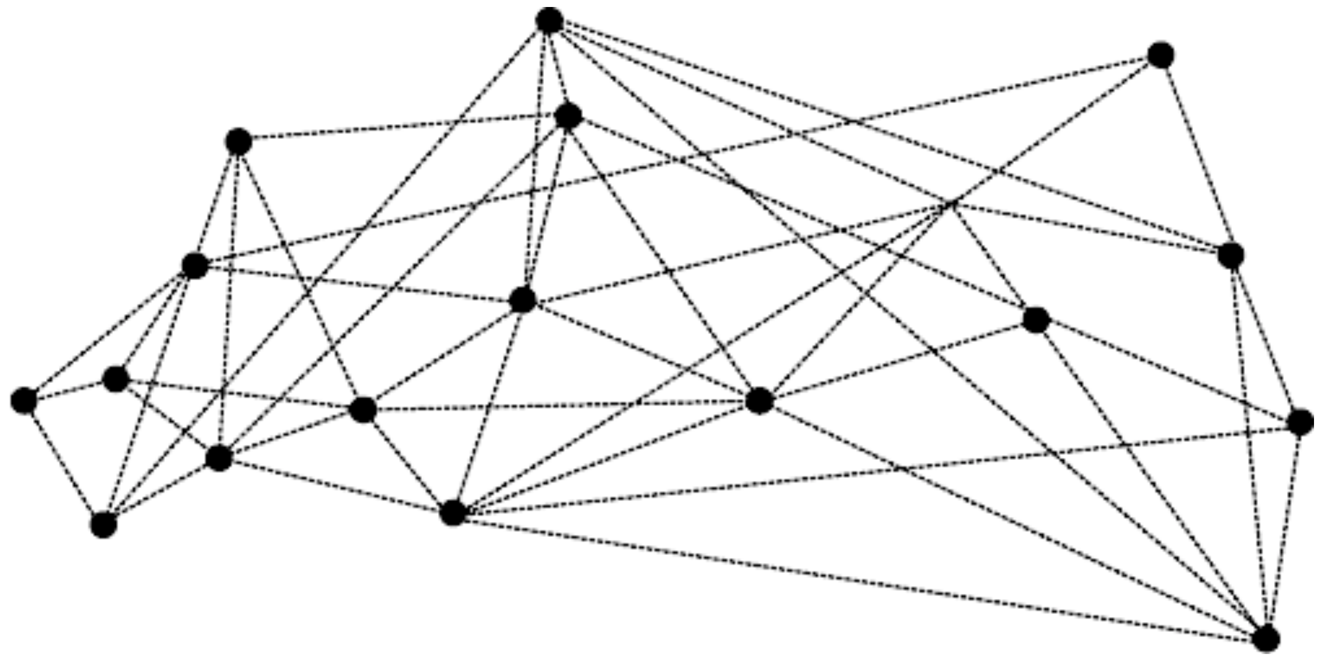
# Cluster topologies, by generation



1990s



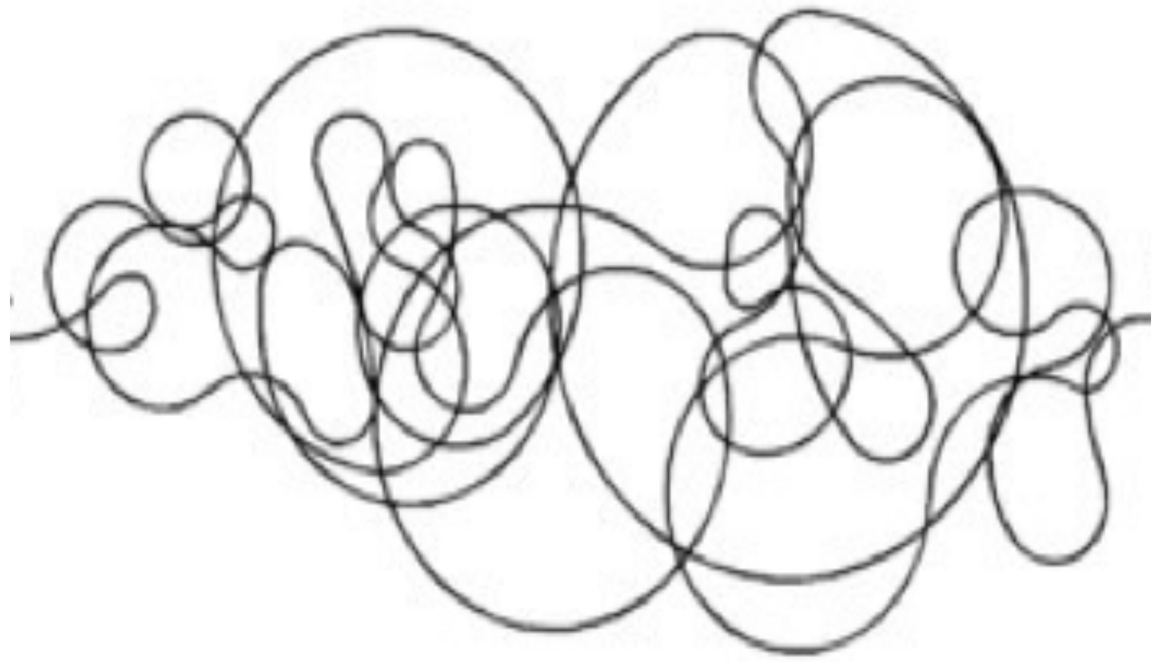
mid-2000s



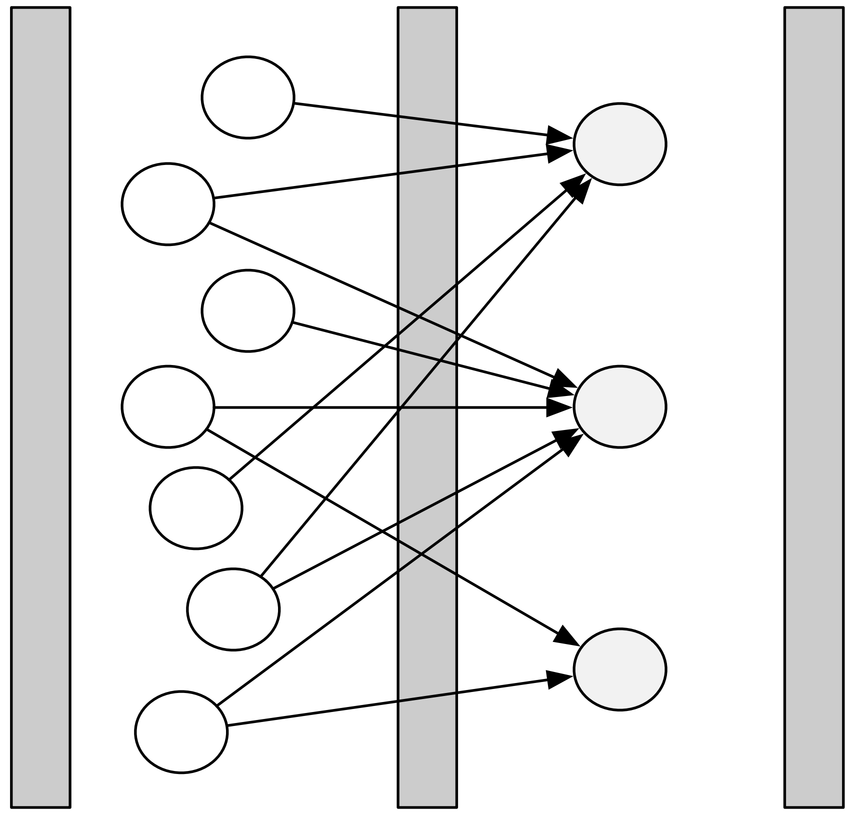
current

Opinion: one problem with software/hardware interface for distributed systems is that it's taken *decades* to prioritize the need for handling **graphs/tensors** directly within popular, accessible open source libraries, without having some **commercial database vendor** intermediate.

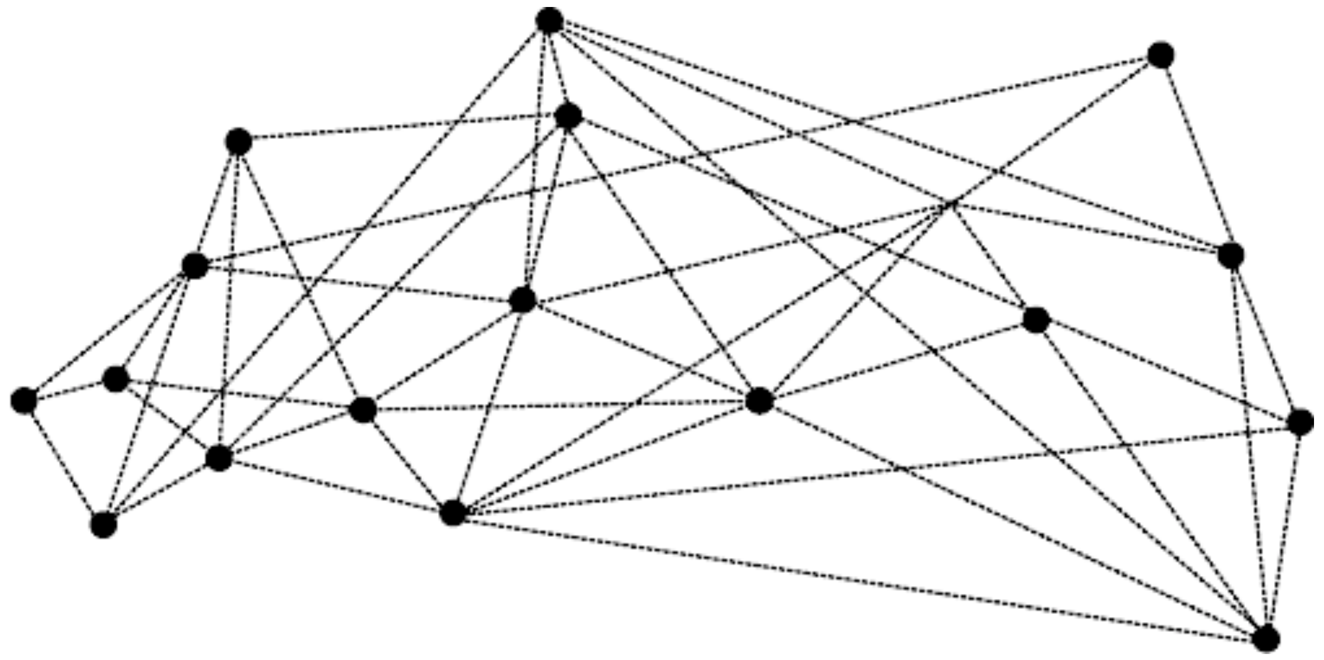
# Cluster topologies, by generation



**1990s**



**mid-2000s**



**current**

MPI

APACHE hadoop

APACHE Spark™

RAY

see also: **Jeff Dean** (2013)
**youtu.be/S9twUcX1Zp0**

# Next Generation Frameworks

Watch carefully about **Ray**, the next generation of distributed systems, i.e., what comes after Apache Spark:

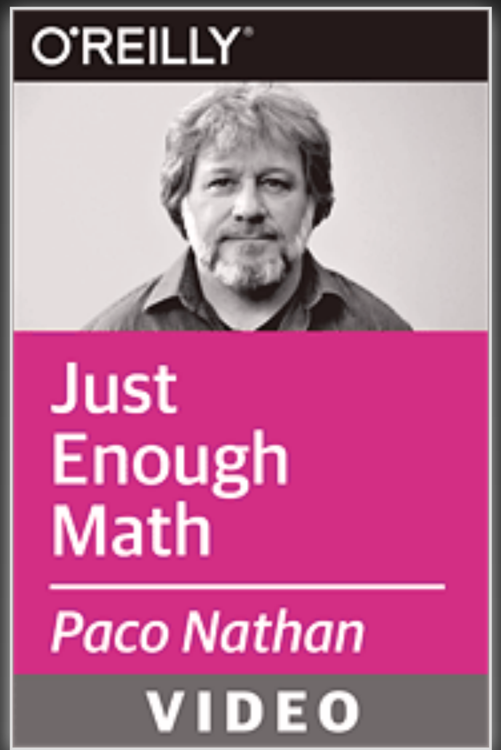
- focus on open source machine learning in Python
- accelerates work on laptops, scales to cloud without changing code; ~4 lines of code needed to parallelize
- supports ad-hoc queries, deep learning (TensorFlow, PyTorch), and reinforcement learning
- developed in response to contemporary ML research, cloud economics (after +10 years), taking advantage of rapid hardware evolution
- unifies *actor abstraction* with the *task-parallel abstraction*

See: **“Ray for the Curious”** by **Dean Wampler**

# publications, interviews, conference summaries…

https://derwen.ai/paco

@pacoid

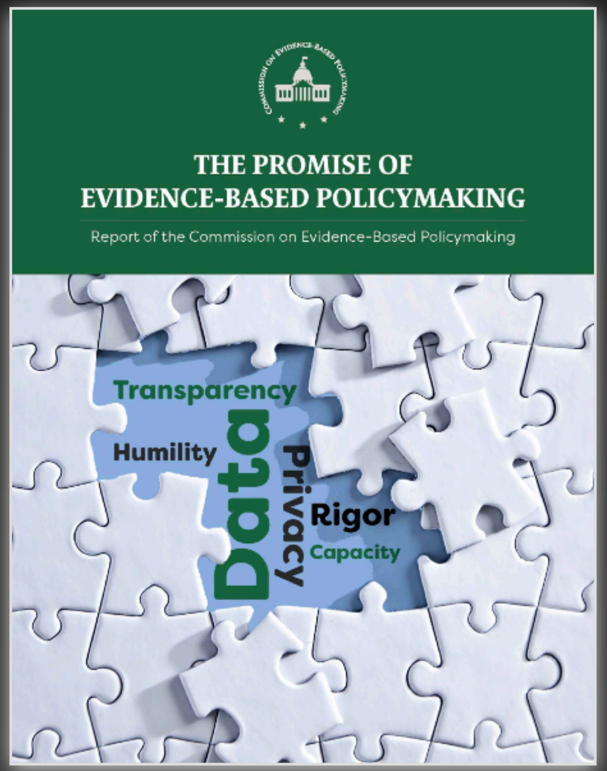

O'REILLY

Just Enough Math

Paco Nathan

VIDEO

Just Enough Math


THE PROMISE OF EVIDENCE-BASED POLICYMAKING

Report of the Commission on Evidence-Based Policymaking

Rich Context

Hylbert-Speys

Rev conf

Themes + Confs
per Pacoid

derwen.ai