CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Hands-on Tutorial

Supported by Microsoft Research

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

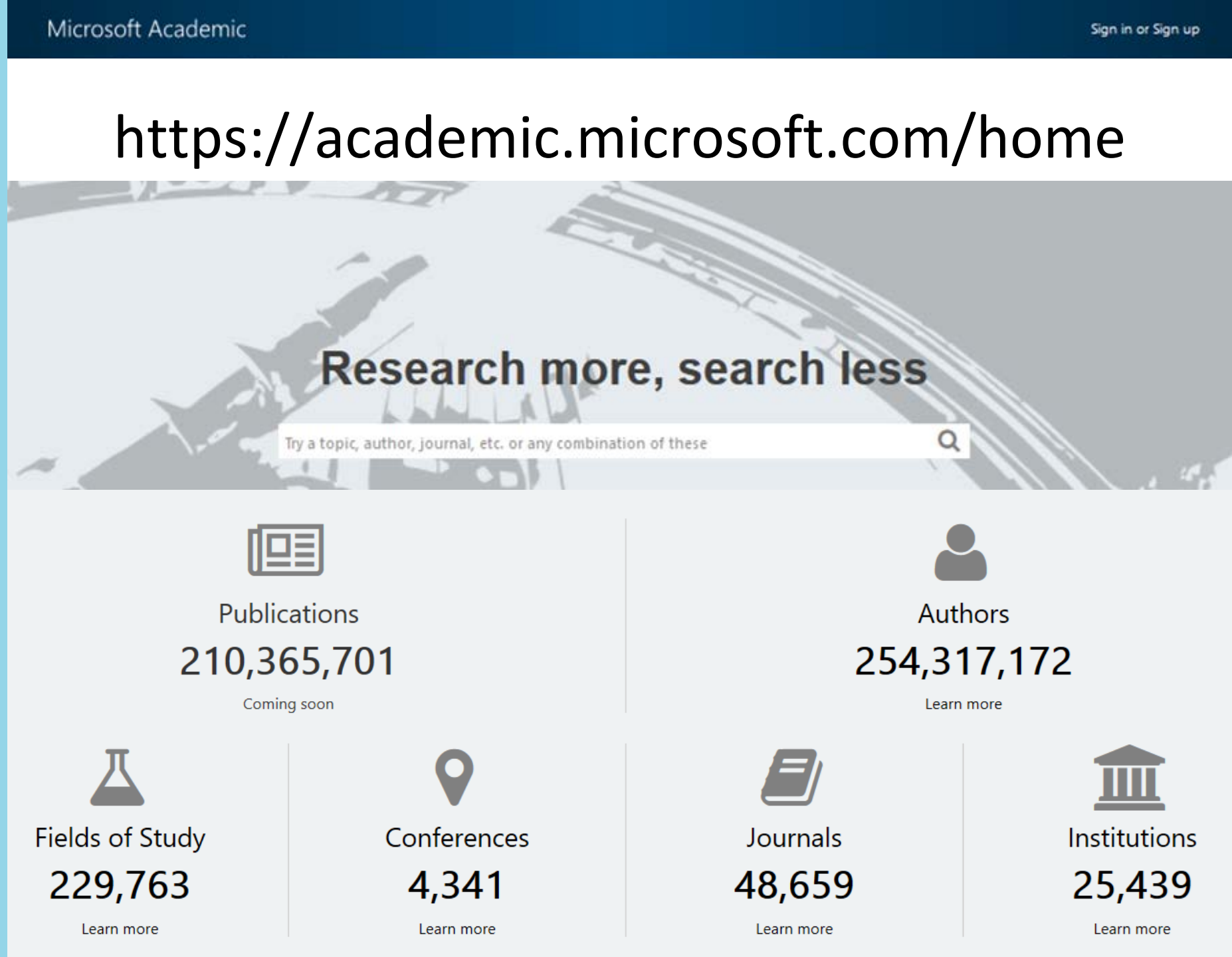
B1G
ACADEMIC ALLIANCE

Program overview

- The CADRE project (Val Pentchev)
- Hands on intro to CADRE (Mat Hutchinson)
- Interactive demo with packages and notebooks (Filipi Silva)
- CADRE fellow presentation (Yi Bu)
- Demo for scalability and Reproducibility (Xiaoran Yan)
- Q&A and conclusion

Microsoft Academic Graph

An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web* 2015.



Tutorial Resources

- https://cadre.iu.edu/
- https://cadre.iu.edu/news-and-events/events/rome
- https://github.com/iuni-cadre/ISSI-tutorial

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

The CADRE project

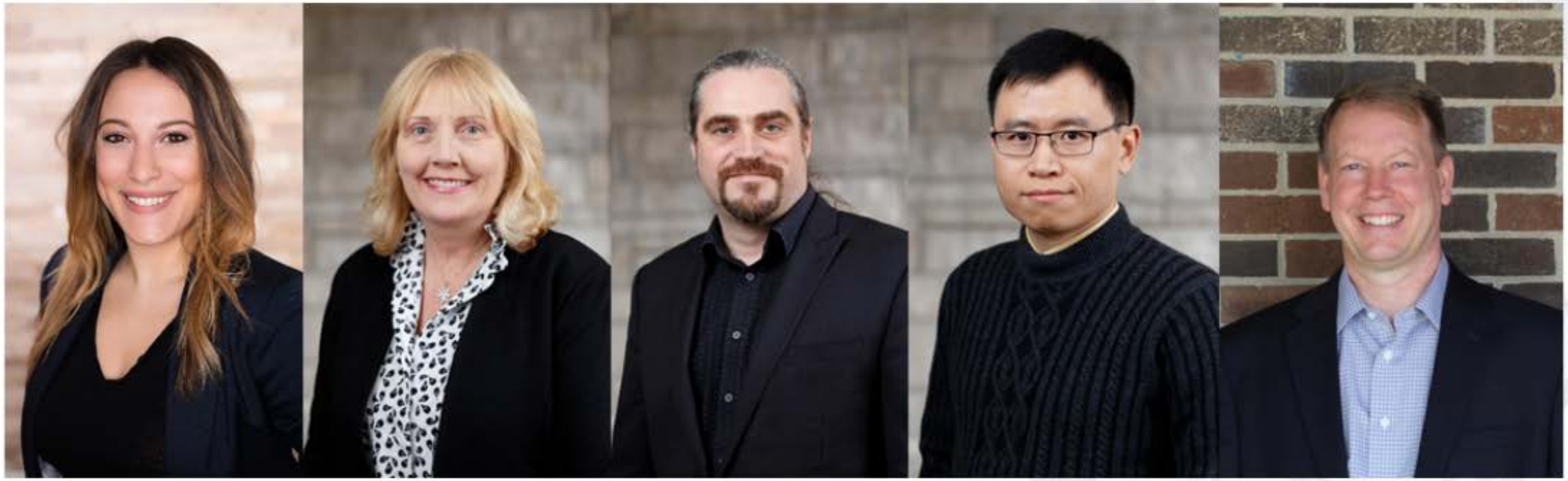
Val Pentchev

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

BIG
ACADEMIC ALLIANCE

CADRE Leadership



Partners

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

LIBRARIES

BIG
ACADEMIC ALLIANCE

- University of Iowa Libraries
- University of Michigan Libraries
- Michigan State University Libraries
- University of Minnesota Libraries
- Ohio State University Libraries
- Penn State University Libraries
- Purdue University Libraries
- Rutgers University Libraries

- Health Partners
- Pervasive Technology Institute
- Midwest Big Data Hub
- South Big Data Hub
- West Big Data Hub
- Microsoft Research
- Web of Science Group

This project was made possible in part by the Institute of Museum and Library Services LG-70-18-0202.

CADRE Project - The Dilemma

- Libraries cannot provide researchers with **sustainable, standardized access** to licensed datasets for text & data mining

- It is cost-prohibitive for most individual libraries to **develop and implement infrastructure** to provide access to licensed big data sets and large or unwieldy open data sets

- Many researchers who could benefit from text and data mining library-acquired resources, lack programming skills and would only be able to do so via a **graphical user interface**

CADRE Project - The Solution

- CADRE is a cloud-based platform that will provide **secure access to library-licensed datasets** and open, non-consumptive datasets

- By sharing the cost of this solution across a large number of academic libraries, we are able to provide a **superior solution at a lower cost to members**, as well as a free service tier for non-members

- CADRE will feature a graphical **user interface**; standardized , multiple data formats; shared and custom **computational resources**; and a **space to share and store** queries, algorithms, derived data, results of analyses, workflows, and visualizations.

CADRE Project - Indiana University Network Science Institute (IUNI)

- **IUNI** – http://iuni.iu.edu **a unique startup** in an established academic institution
  - A **cross-campus, transdisciplinary institute** that brings together faculty who engage in **network research** from various scientific fields
- **IUNI's mission**
  - To **strengthen the theories**, methods, analytic tools, and practice **of network science**, and to **foster collaborative, interdisciplinary network science approaches** to understanding and improving the complex challenges of our world
- **IUNI's Teams**
  - A team of IT professional
  - A team of research scientists

CADRE Project – Goals
Identify Constituents' Needs

- Understanding **users' needs** and expectations
  - User stories, Product Owner Council meetings, **Communication**
- **Informatics/computer science researchers** and labs
  - **APIs, Notebooks,** Access to **Raw Data** and **Cloud Compute** Resources
- **Science of Science** community
  - **Interface Access to Databases** and Cloud Native Technologies
- **Library and research** community outside of computer science
  - **Web Interface** guiding Query Building and **suggesting the most appropriate backend technology** on a case by case basis

CADRE Project – Goals
Research Asset Commons

- **Federated Login**
  - Access from **any affiliated institution** using single sign on (CILogon, inCommon, Shibboleth etc.)
  - **Restricted access to proprietary resources**, based on login credentials
- **Collaboration**
  - Ability to **save and share** with specific users, community or the public **metadata, queries, results, annotations, visualizations, algorithms, code, containers and virtual machines**
  - Community building and collaboration based around **same data access privileges** and goals
- **Reproducibility, Replicability, Provenance and Transparency**
  - Use of the same, **well documented original datasets**
  - **DOIs** identifying any and every data change or permutation
  - **Saved and shared** workflows a pipelines trough **Packages and Containers**
  - **Ability to publish using unique identifiers** leading back to Research Asset Commons

CADRE Project – Goals
Identify the Proper Technology

- **Raw data access**
  - Access to **XML, JSON, CSV** etc. files in their native form. **Containerized** tools and packages
  - Access to data using **cloud native technologies** like **U-SQL and Athena/Glue**
  - Access to cloud **distributed computing** using **Databricks** and **SPARK** on **HDInsight** and **EMR**
- **Database access**
  - Researching on currently available **cloud and serverless Relational Database** implementations for each dataset and query type
  - Researching on currently available **Graph Database** implementations for each dataset and query type. Currently comparing **Neo4j, Tiger Graph, AgensGraph**, **cloud native and in-memory** alternatives
- **Web interface**
  - **Guided Query Building**
  - Ability to **suggest the most appropriate technology** on a case by case basis
  - **User control** over execution and use of resources

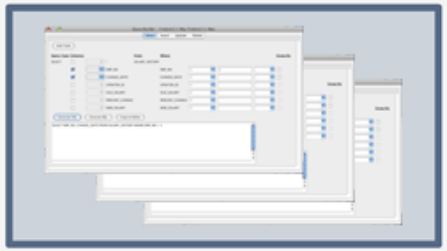
SHARED BIGDATA-GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)

AUTHENTICATION

FEDERATED SINGLE SIGN ON

- Custom
- InCommon
- Shibboleth
- Globus

WEB QUERY INTERFACE



RESEARCH ASSET COMMONS

Saved and Shared Results, Tools and Datasets

- Visualizations
- Annotations
- Data Derivatives
- Metadata

ACCESS BY PROGRAMMING AND VISUALIZATION TOOLS

Gephi, sigma, tableau, js, kibana, D3, S2, php, JavaScript

APPLICATION PROGRAMMING INTERFACE (API)

CLOUD COMPUTE GATEWAY

Machine Learning

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

Text Analytics

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

Database Analytics

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

Data Processing and Querying

elasticsearch, Spark, hadoop, Hive, SPARQL, presto, SQL, Cypher, R, Future processes

CLOUD AND LOCAL COMPUTE RESOURCES

Jetstream, Commercial Cloud Services, Big Red II

AUTHORIZATION

GRANULAR DATASET PERMISSIONS

CLOUD STORAGE

Indexed Storage

- Relational Databases
- Graph Databases
- Future Databases

Raw Format Data Lake Storage

- Web of Science
- MS Academic Graph
- USPTO Data
- Future Databases

LOCAL AND INSTITUTIONAL STORAGE

- Future Data Sources
- Future Datasets

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Hands on intro to CADRE

Mat Hutchinson

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

BIG
ACADEMIC ALLIANCE

Demo 1

https://github.com/iuni-cadre/ISSI-tutorial

Questions?

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Interactive demo

Filipi Silva

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

BIG
ACADEMIC ALLIANCE

Demo 2

https://github.com/iuni-cadre/ISSI-tutorial

Demo 3

https://github.com/iuni-cadre/ISSI-tutorial

Questions?

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

CADRE Fellows

Xiaoran Yan

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

B1G
ACADEMIC ALLIANCE

CADRE related events

Apr. 2019

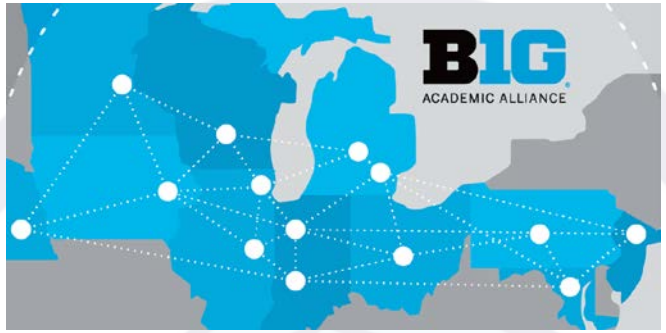
- 2019 CADRE meeting
- CADRE Fellowship open

Sep. 2019

- 1st Fellows announced
- ISSI workshop & tutorial

May. 2020

- 2020 CADRE meeting
- BTAA Library Conference 2020
- 2020 CADRE hack-a-thon



ISSI

SAPIENZA
UNIVERSITÀ DI ROMA

INDIANA UNIVERSITY
BLOOMINGTON

CADRE Fellowship program

- Gain access to the big bibliometric data sets
- Receive data and technical support for your project
- Join the CADRE community on Slack channels, GitHub repositories and other platforms
- Have early access to free cloud computing resources
- Receive travel scholarships

Utilizing Data Citation for Aggregating, Contextualizing, and Engaging with Research Data in STEM Education Research

Researchers: Michael Witt, Loran Carleton Parker, Ann Bessenbacher

Affiliation: Purdue University

MCAP: Mapping Collaborations and Partnerships in SDG Research

Researchers: Jane Payumo, Devin Higgins, Scout Calvert, Guangming He

Affiliation: Michigan State University

The global network of air links and scientific collaboration – a quasi-experimental analysis

Researchers: Katy Börner, Adam Ploszaj, Lisel Record, Bruce Herr II

Affiliation: Indiana University Bloomington and University of Warsaw

Measuring and Modeling the Dynamics of Science Using the CADRE Platform

Researchers: Russell Funk, Michael Park, Thomas Gebhart, Britta Glennon, Julia Lane, Raviv Murciano-Goroff, Matthew Ross, Jina Lee, Erin Leahey

Affiliation: University of Minnesota, University of Pennsylvania, New York University, Boston University, University of Arizona

Comparative analysis of legacy and emerging journals in mathematical biology

Researchers: Marisa Conte, Samuel Hansen, Scott Martin, Santiago Schnell

Affiliation: University of Michigan and University of Michigan Medical School

Systematic over-time study of the similarities and differences in research across mathematics and the sciences

Researcher: Samuel Hansen

Affiliation: University of Michigan

A user story from CADRE fellows

Understanding citation impact of scientific publications through ego-centered citation networks

Researchers: Yi Bu, Chao Min, Ying Ding

Affiliation: Indiana University Bloomington and Nanjing University

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Exploring ego-centered citation networks: A technical introduction

Yi Bu[1], Chao Min[2], and Ying Ding[1]

1: School of Informatics, Computing, and Engineering, Indiana University, U.S.A.
2: School of Information Management, Nanjing University, China

LIBRARIES

INDIANA UNIVERSITY NETWORK SCIENCE INSTITUTE

BIG ACADEMIC ALLIANCE

Understanding citation impact of scientific publications

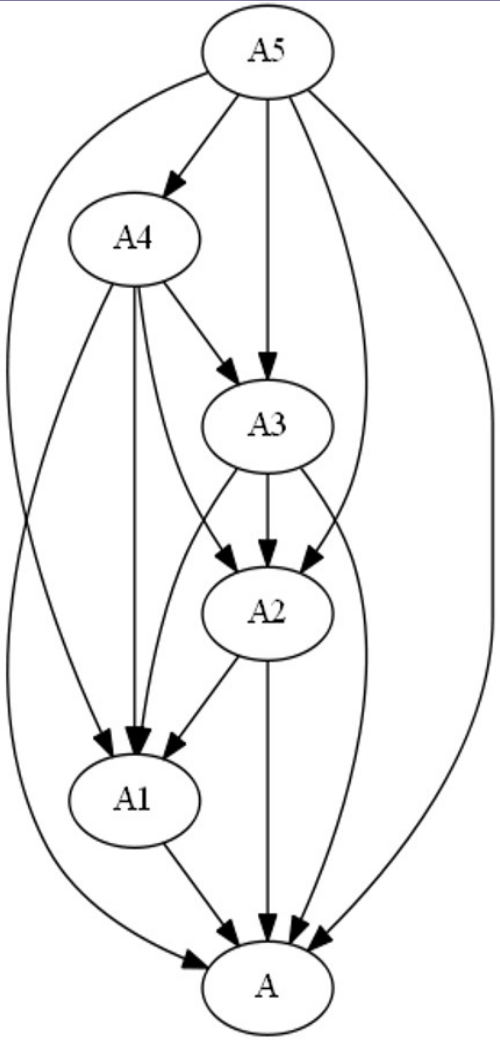
- Scientific impact as a type of impact
- Citation impact as a type of scientific impact
  - ✓ Citation impact among all types of impact
  - ✓ Citation impact of scientific publications
- Benefits from understanding citation impact
  - ✓ Indicator perspective: Measuring citation impact offers a useful way of examining the scientific impact of a publication.
  - ✓ More general perspective: Measuring citation impact can also assist in understanding knowledge diffusion and the use of information.

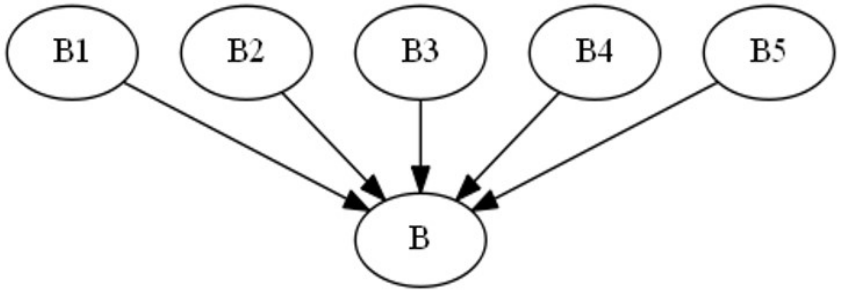
Understanding citation impact of scientific publications (cont.)

- Previous ways of understanding citation impact of scientific publications:
  - ✓ Count-based strategies: raw citation count, normalized citation measures…
  - ✓ Network-based strategies: PageRank, EigenFactor…

Understanding citation impact of scientific publications (cont.)

- Local details are missing!
  - ✓ “Deep” or “wide” impact?





Understanding citation impact of scientific publications (cont.)

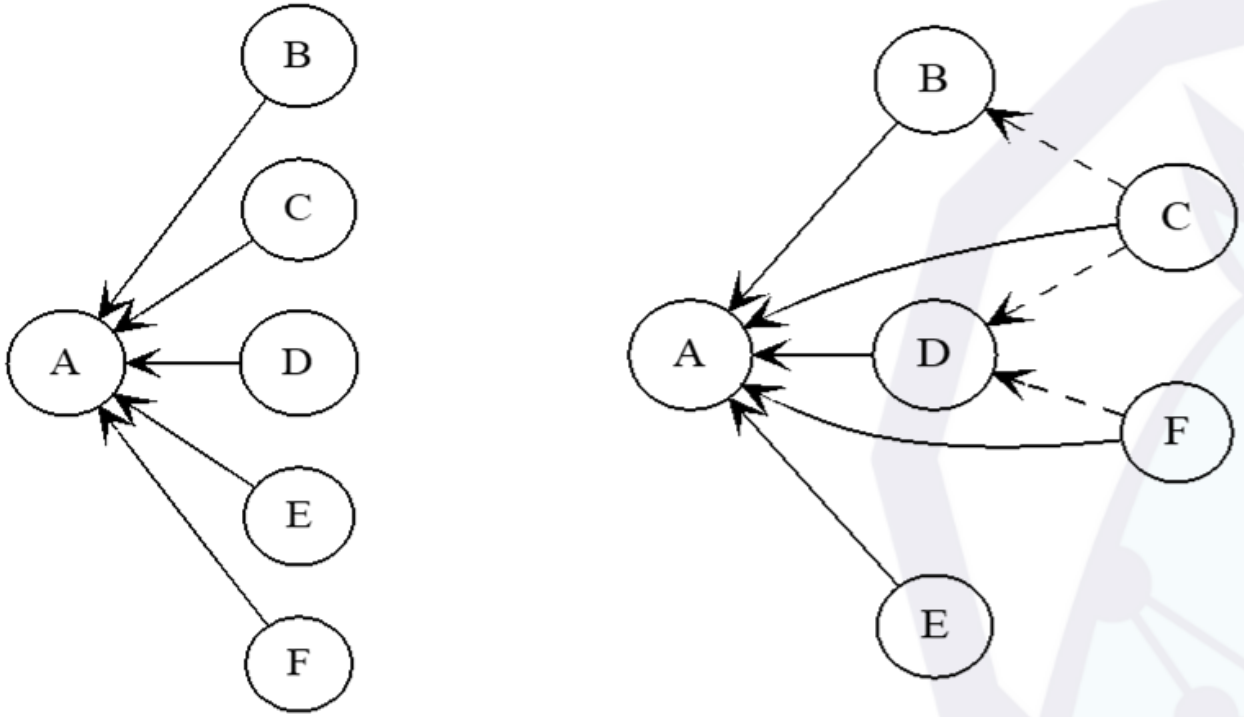
- Local details are missing!
  - ✓ How does an article impact other research, and what are the patterns? The direct citations between citing publications (DCCPs) offer a good way to mine how a publication impacts other research.

		citing publication					
		SSH	BHS	PSE	LES	MCS	subtotal
Cited publication	SSH	11138	224	16	5	37	11420
	BHS	440	1254	2	11	1	1708
	PSE	137	1	19	3	18	178
	LES	57	13	3	11	0	84
	MCS	194	0	17	0	26	237
	subtotal	11966	1492	57	30	82	13627

year	SSH	BHS	PSE	LES	MCS
2006	13	0	0	0	0
2007	111	0	0	0	0
2008	455	0	2	2	4
2009	753	9	3	0	0
2010	1155	19	0	1	0
2011	1310	80	2	1	12
2012	1092	39	3	1	9
2013	1440	187	19	3	41
2014	1110	449	30	2	31
2015	1161	361	12	12	13
2016	1491	290	44	57	60
2017	1329	274	63	5	67

Published year and discipline distributions of citing publications of *h*-index article's DCCPs

Ego-centered citation networks as a tool to understand citation impact

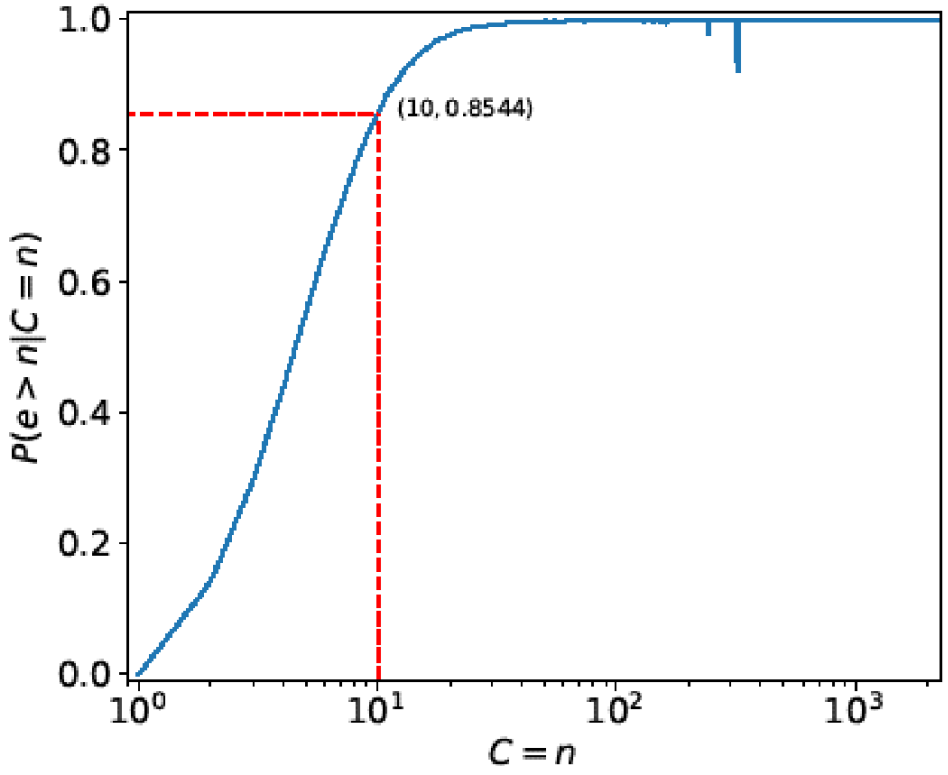


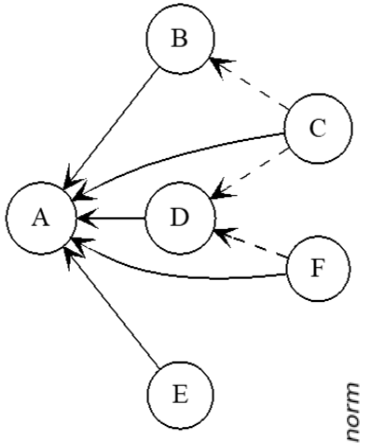
(a) (b)

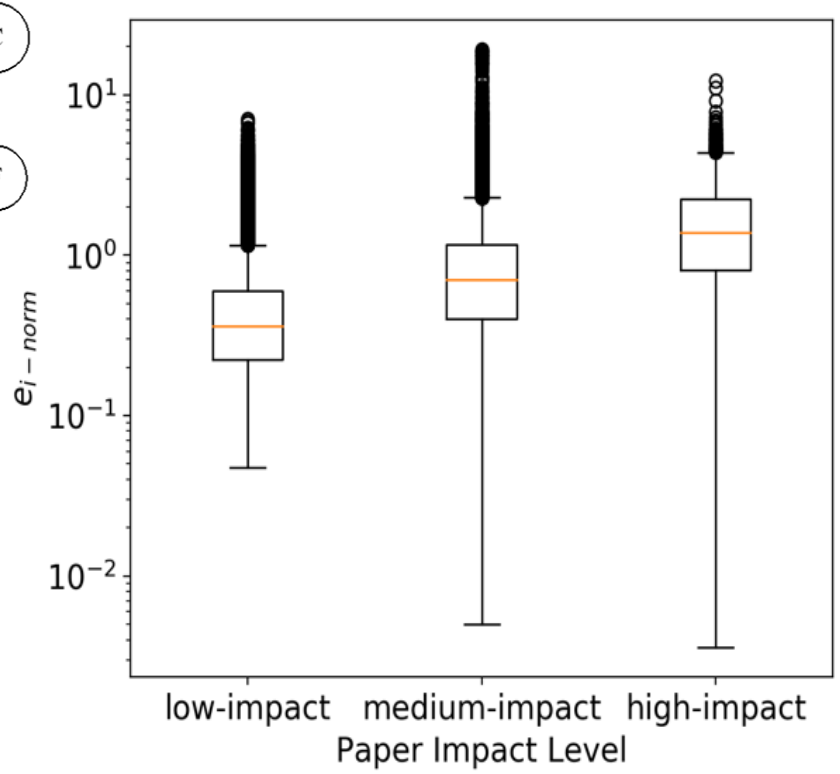
Preliminary research questions

- Do DCCPs occur frequently?
- How does DCCPs differ in papers with different citation impacts and in different years?

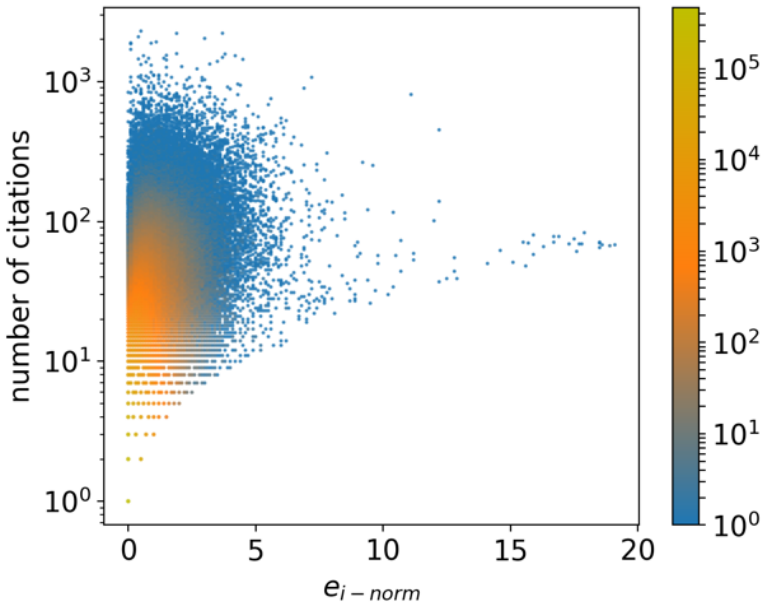
Preliminary results: The universality of DCCPs

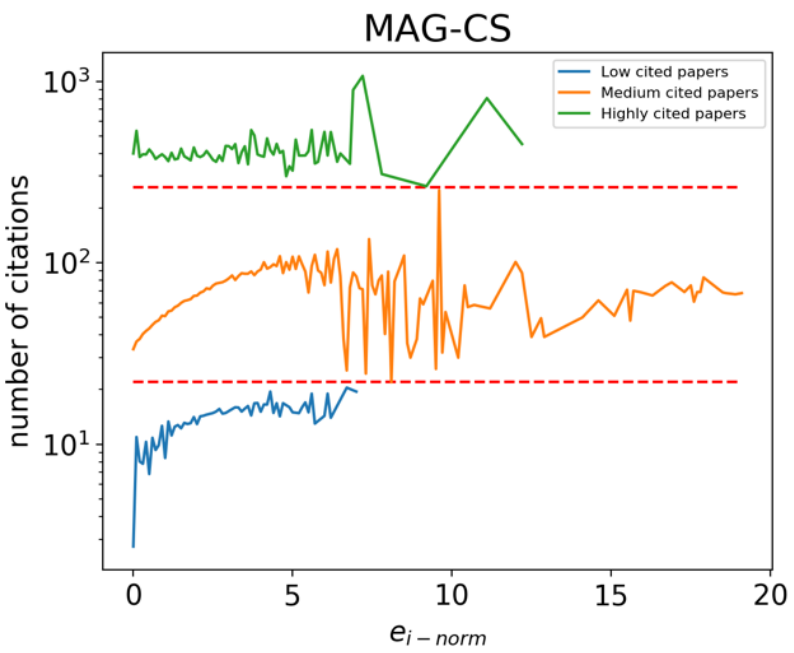


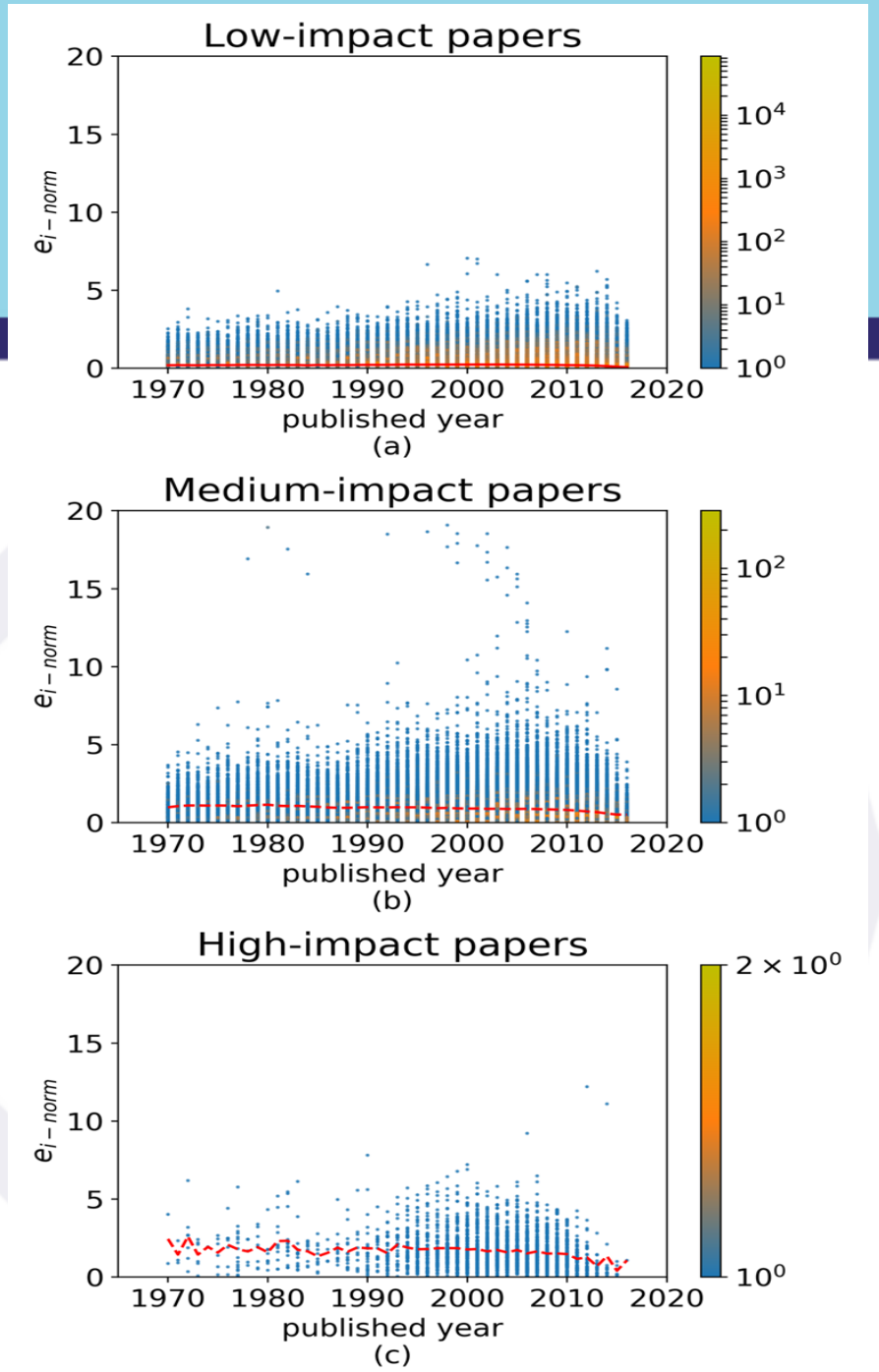




Preliminary results (cont.)







Technical details: Extracting citing relationships from the raw MAG tables

- SQL extraction as a .txt file:

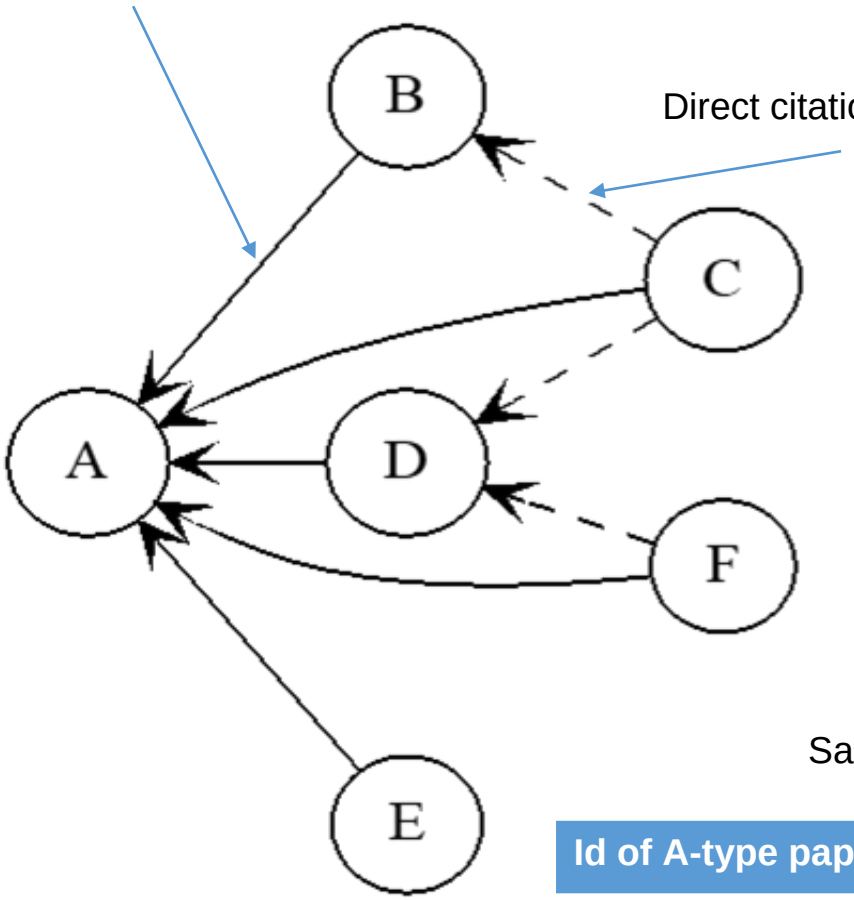
```
import psycopg2
conn = psycopg2.connect(database = 'core_data', user = 'buyi', password = [redacted])
cur = conn.cursor()
cur.execute("SELECT paper_id, paper_reference_id FROM mag_core.paper_references;")
outFile = open("mag_citing.txt", "w+")
lines = ['citing id======cited id']
for row in cur:
if str(row[0]) in paper_id_set and str(row[1]) in paper_id_set:
lines.append('{:}======{:}'.format(str(row[0]), str(row[1])))
if len(lines) % 100000 == 0:
outFile.write('\n'.join(lines) + '\n')
lines = []

outFile.write('\n'.join(lines) + '\n')
cur.close()
```

- .txt file to a Python dictionary:
  - ✓ If paper in paper_citing.keys()

Difficulty 1: How to extract DCCPs?

Direct citations to A



Direct citations between citing publications (from the perspective of A)

Sample output:

Id of A-type paper (focal)	Id of B-type paper	Id of C-type paper

Difficulty 1: How to extract DCCPs? (cont.)

- This task is computationally expensive:
  - ✓ In MAG, we have ~0.1 billion papers. The below Python script will perhaps take forever…

```
indirect_citation = defaultdict(list)
for paper in paper_year.keys(): # for papers that have pub_year information
    for citing_paper_1 in paper_citing[paper]:
        for citing_paper_2 in paper_citing[paper]:
            if citing_paper_1 in paper_citing[citing_paper_2]:
                temp = []
                temp.append(citing_paper_1)
                temp.append(citing_paper_2)
                indirect_citation[paper].append(temp)
```

Difficulty 2: Self-citations in ego-centered citation networks?

- If two papers (A and B) share at least one co-author and B cites A, such citation is called a self-citation (first-order self-citation).
- How about these circumstances, when B cites A?
  - ✓ A and B don't share co-authors, but A and C do, and B and C do. [second-order self-citations]
  - ✓ A and B don't share co-authors, but A and C do, B and D do, and C and D do. [third-order self-citations]
  - ✓ This indicates how researchers' social distance impacts on their self-citation patterns.
- **How to technically achieve these?**

Difficulty 2: Self-citations in ego-centered citation networks?

- Completing this task is also QUITE computationally expensive:
  - ✓ Deriving n-order self-citations need to know **the shortest paths and their lengths** in the co-authorship and citation networks
  - ✓ Such networks are quite huge (hundreds of millions of nodes in the citation network, and millions of nodes in the co-authorship network)

Questions?

Presenter: Yi Bu, Indiana University

Email: buyi@iu.edu
Website: https://buyi08.wixsite.com/yi-bu

CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Scalability & Reproducibility

Xiaoran Yan

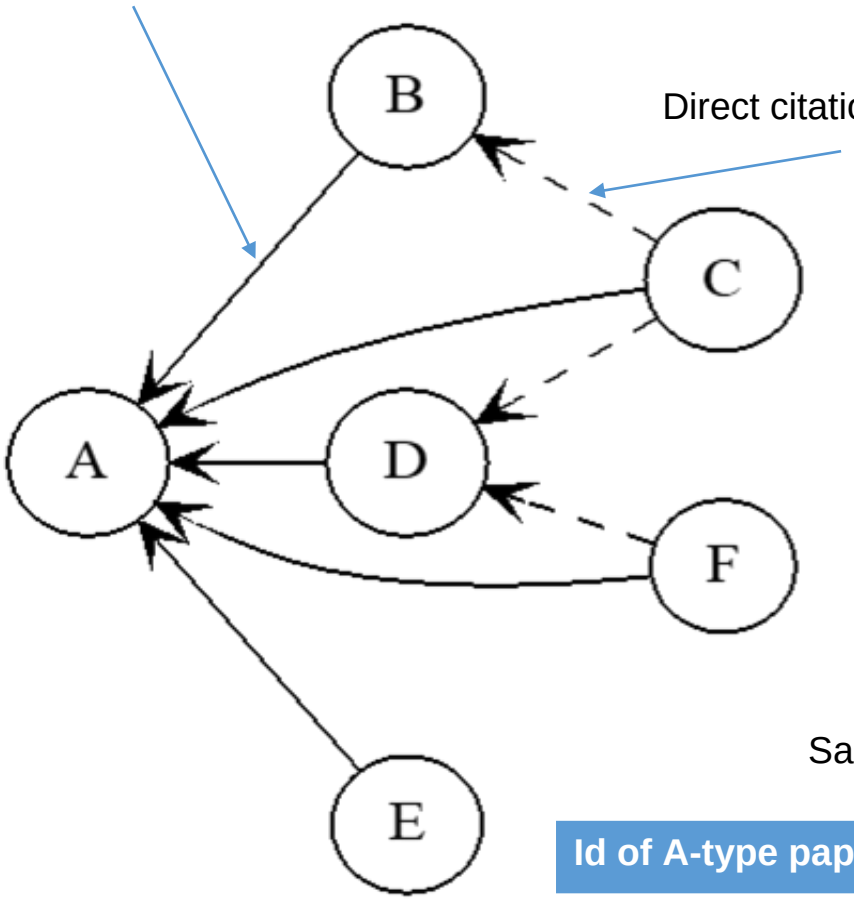
LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

B1G
ACADEMIC ALLIANCE

Difficulty 1: How to extract DCCPs?

Direct citations to A



Direct citations between citing publications (from the perspective of A)

Sample output:

Id of A-type paper (focal)	Id of B-type paper	Id of C-type paper

Difficulty 1: How to extract DCCPs? (cont.)

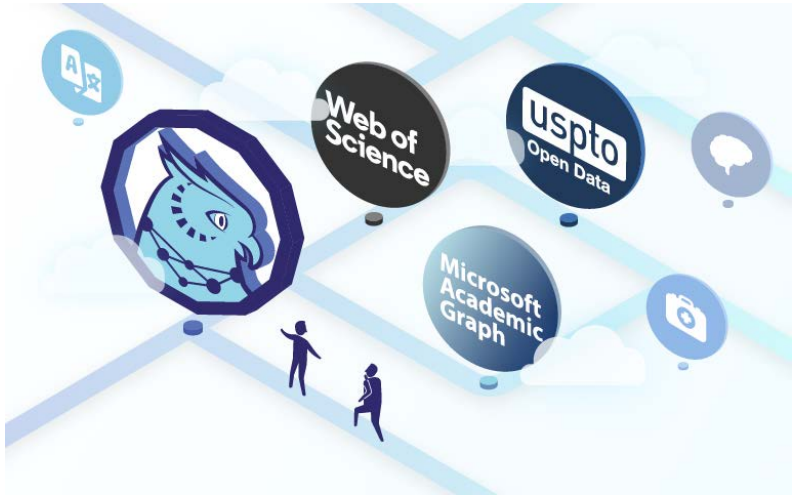
- This task is computationally expensive:
  - ✓ In MAG, we have ~0.1 billion papers. The below Python script will perhaps take forever…

```
indirect_citation = defaultdict(list)
for paper in paper_year.keys(): # for papers that have pub_year information
    for citing_paper_1 in paper_citing[paper]:
        for citing_paper_2 in paper_citing[paper]:
            if citing_paper_1 in paper_citing[citing_paper_2]:
                temp = []
                temp.append(citing_paper_1)
                temp.append(citing_paper_2)
                indirect_citation[paper].append(temp)
```

CADRE’s solution

- An easy to use graphical interface of a query builder with preview functionality
- A unified engine with optimized combinations of solutions based on relational/graph/document databases
- For users who want intuitive and quick access of data, no programing skills required
- In development: APIs for power users

CADRE’s solution

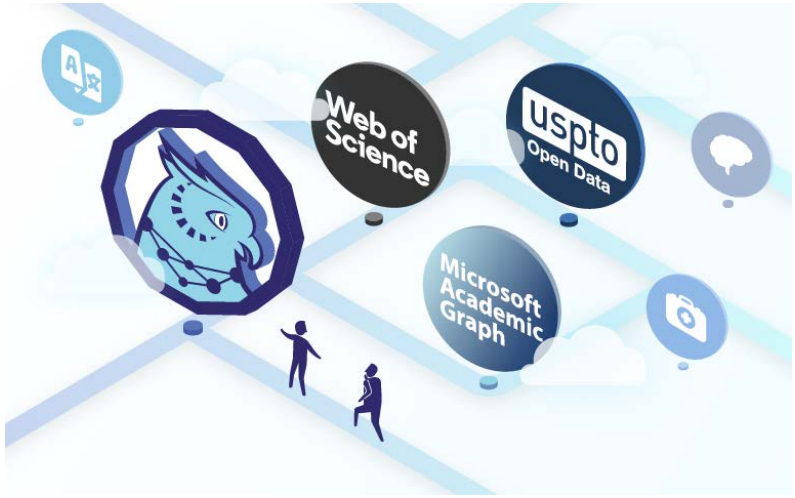


Access over 220 million scientific publications

Effortlessly query data and analyze results

Reproduce research & leverage tools

CADRE’s solution



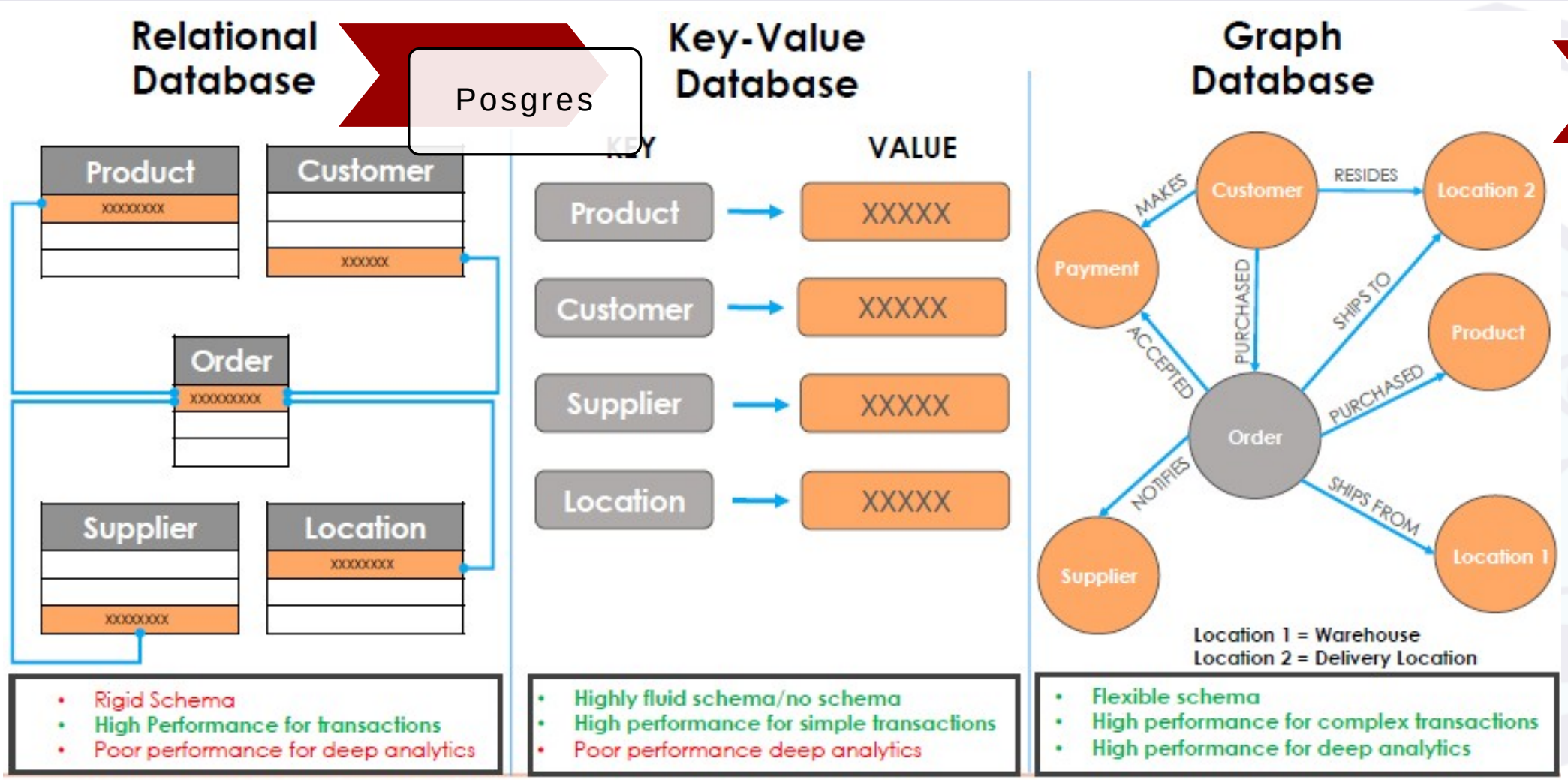
Databases

GUI-query

Notebooks

RAC

Databases behind the query interface



Gaurav Desphpande, vice president of marketing for TigerGraph

Databases behind the query interface

Database	Database Type	Language support	Distributed
Neo4j	Native graph database	Cypher	No
RedisGraph	In-memory data structure store (Graphs represented as sparse adjacency matrices)	Subset of Cypher	No
TigerGraph	Native graph database	GSQL	Yes
JanusGraph	Supports various storage backends (Apache Cassandra, Apache HBase, Google Cloud Bigtable, Oracle BerkeleyDB)	Native integration with the Apache TinkerPop (Gremlin stack).	Yes

Demo 4

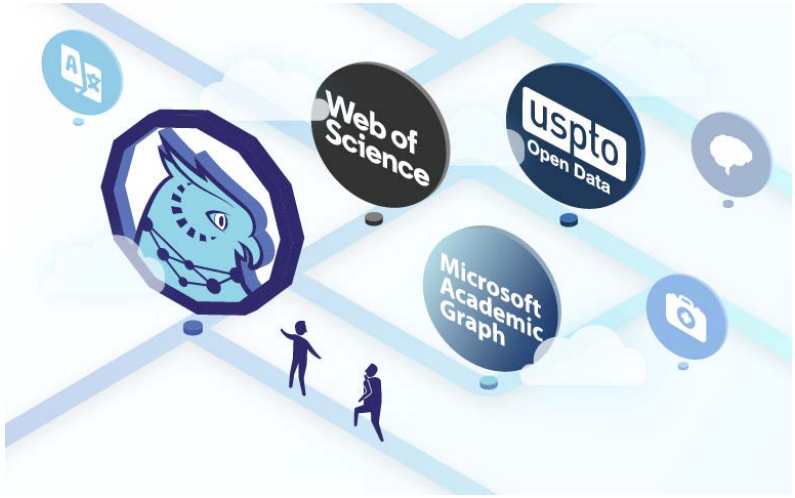
https://github.com/iuni-cadre/ISSI-tutorial

Questions?

Presenter: Xiaoran Yan, Indiana University

Email: yan30@iu.edu

CADRE’s solution



Access over 220 million scientific publications

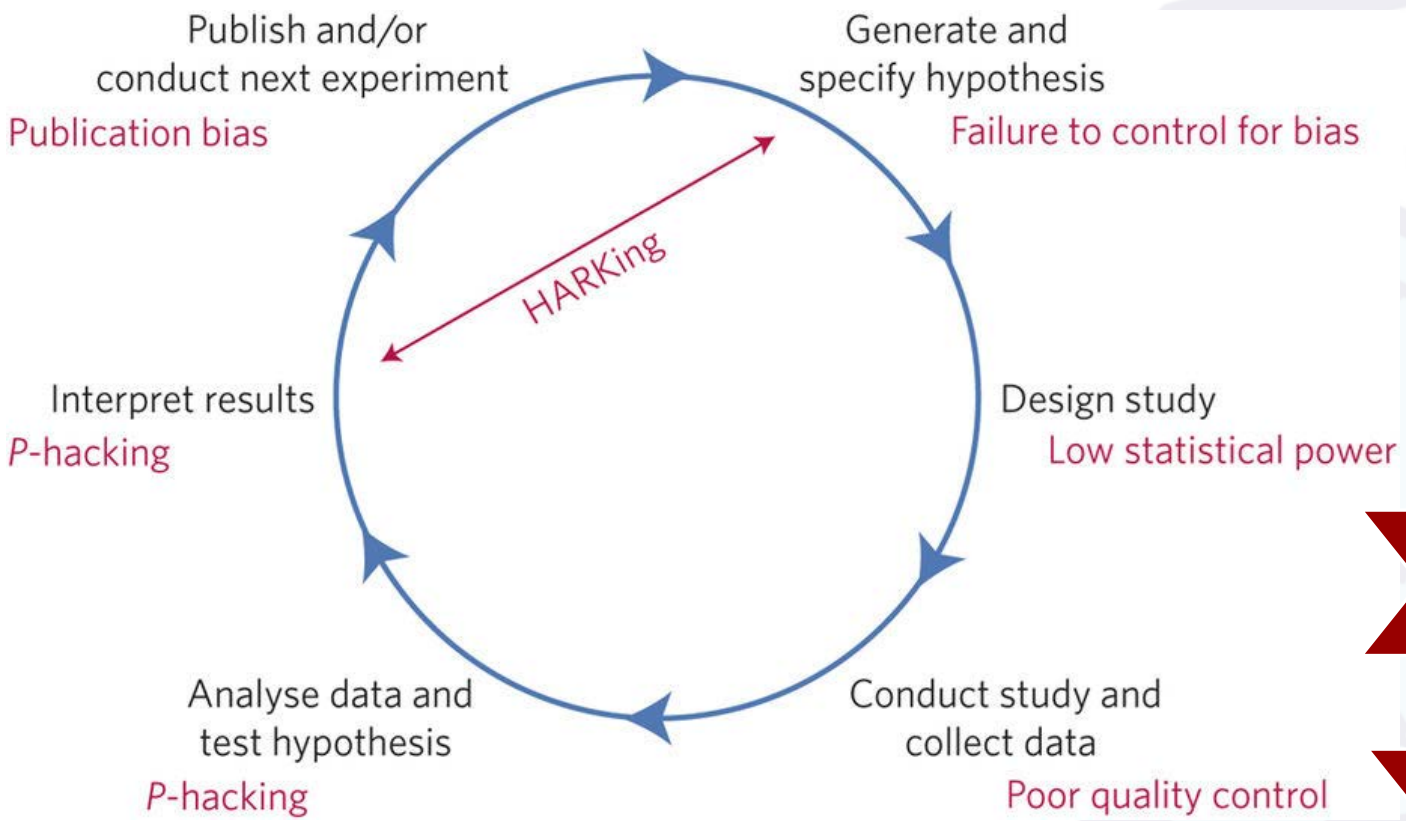
Effortlessly query data and analyze results

Reproduce research & leverage tools

The reproducibility “Crisis”

RAC

Notebooks



GUI-query

Databases

Marcus R. Munafò, et al. “A manifesto for reproducible science” (2017)

Spectrum of Reproducibility

Computational → Statistical → Empirical

Stodden, Victoria. “Resolving Irreproducibility in Empirical and Computational Research” (2013)

Current solutions

IU-AMBITION / MASS

Code | Issues 0 | Pull requests 0 | Projects 0

Matlab code of minimum absolute spectral similarity

6 commits | 1 branch

Branch: master | New pull request

everyxs Update README.md

LFR.mat	Add files via upload	
README.md	Update README.md	
absSpecSim.m	Add files via upload	
main.m	Add files via upload	7 months ago
sparsify.py	Added a Python implementation	7 months ago

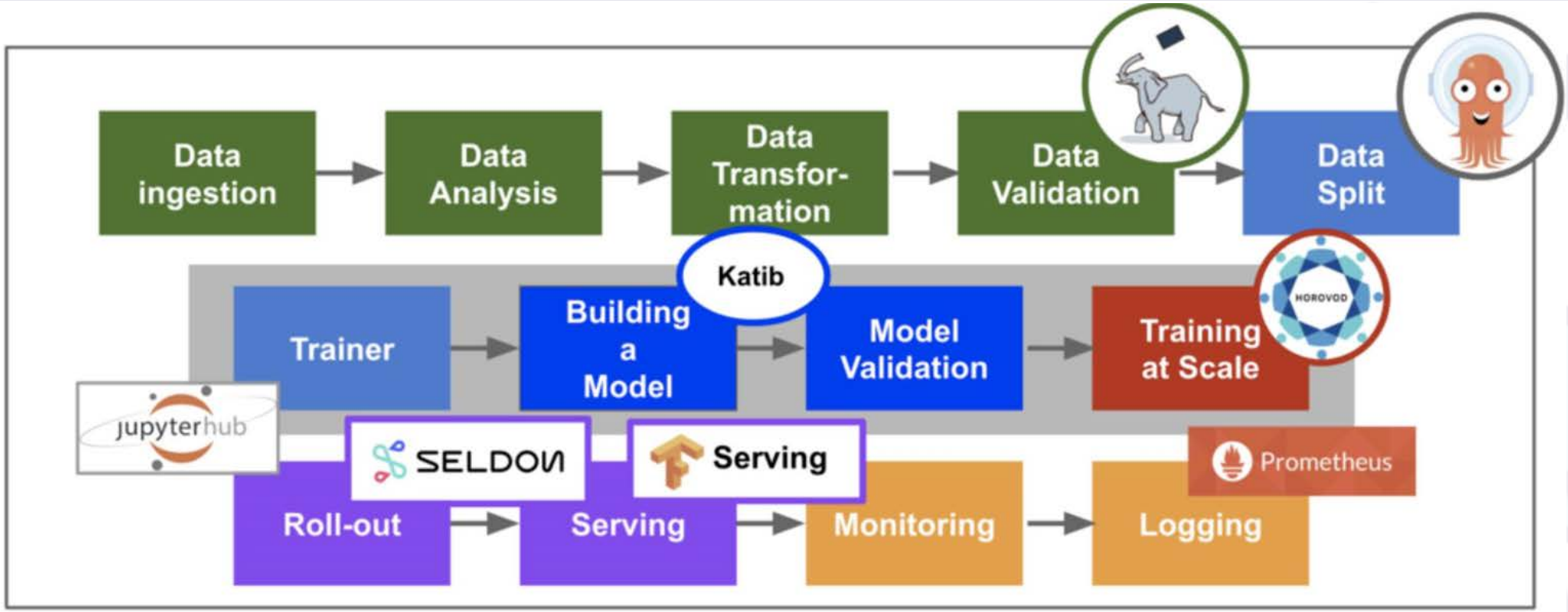
README.md

Matlab function for the Minimum absolute spectral similarity (MASS)

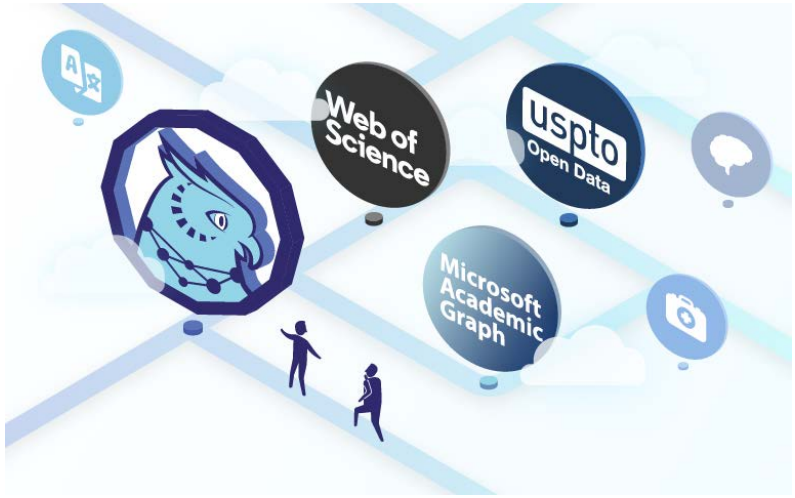
Publications

15. Yan, X., Jeub, L., Flammini, A., Radicchi, F., and Fortunato, S. Weight Thresholding on Complex Networks. *Accepted by Physical Review E, Issue/Batch: 10_05_18 (2018)*
    https://arxiv.org/abs/1806.07479
    Source code: https://github.com/IU-AMBITION/MASS
14. Faskowitz, J., Yan, X., Zuo, X.-N., and Sporns, O. Scientific reports, 8(1):12997
    https://www.nature.com/articles/s41598-018-31202-1

Big data pipelines in the industry



CADRE’s solution



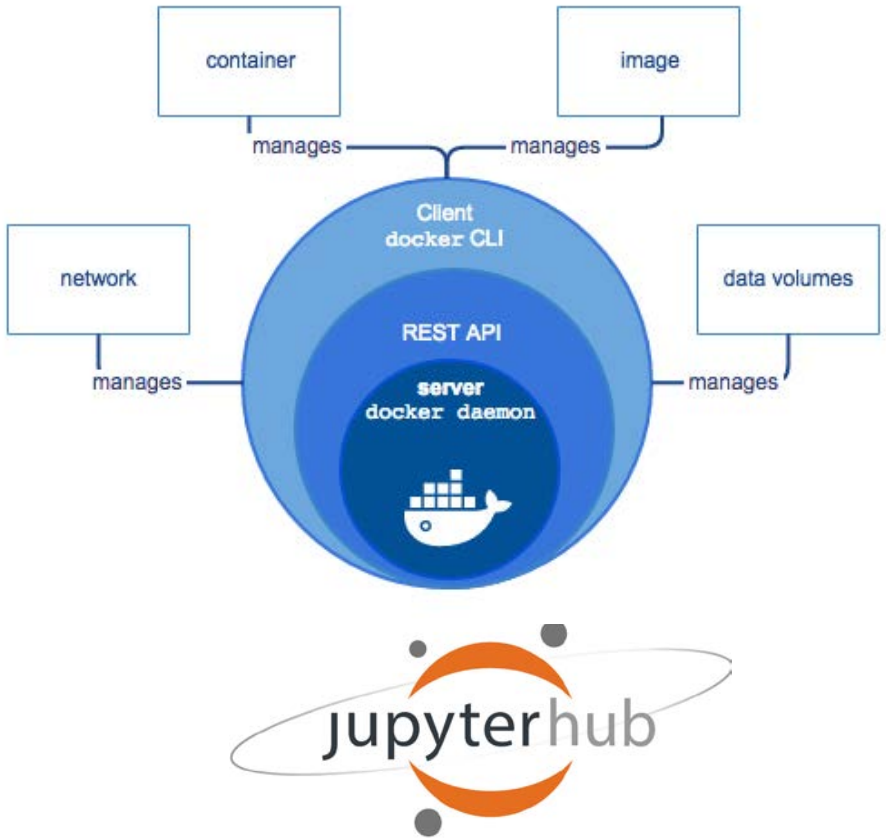
Databases

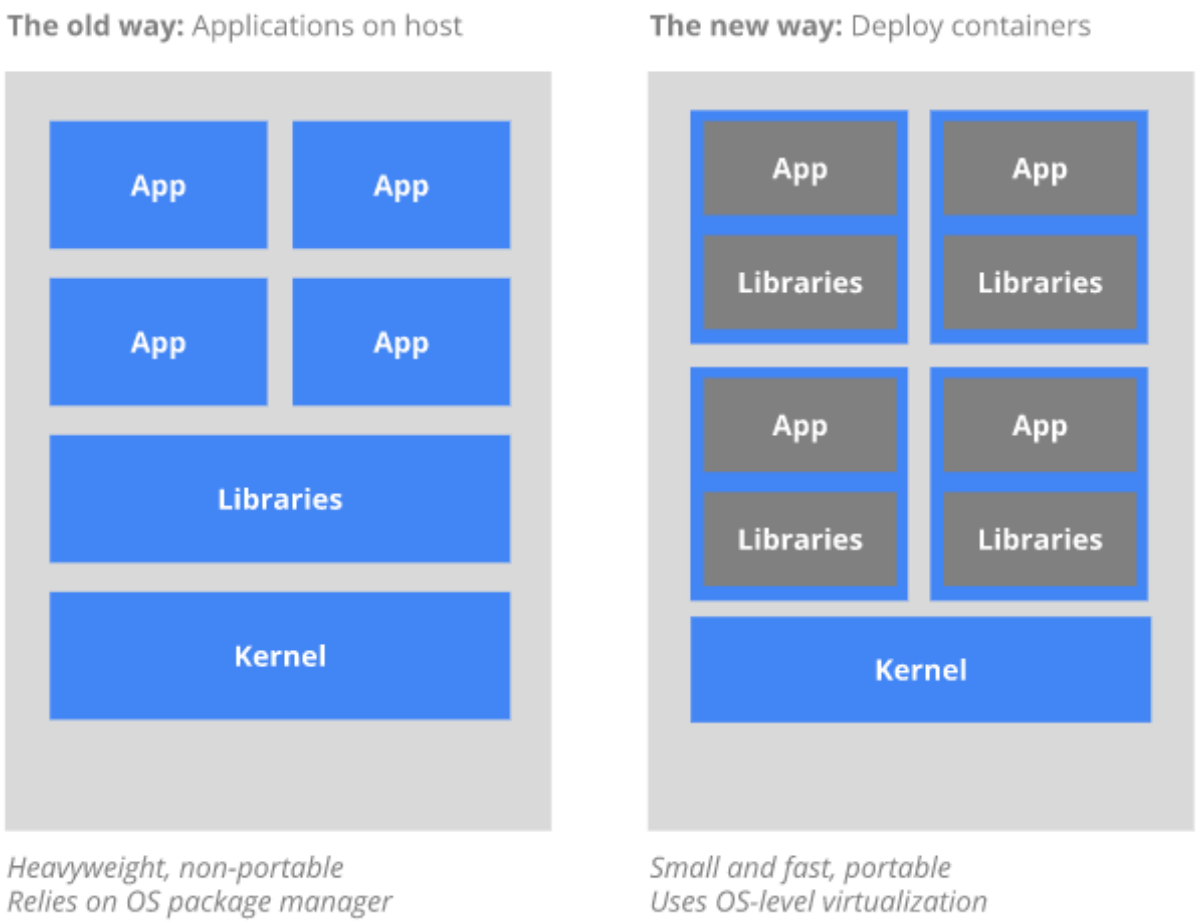
GUI-query

Notebooks

RAC

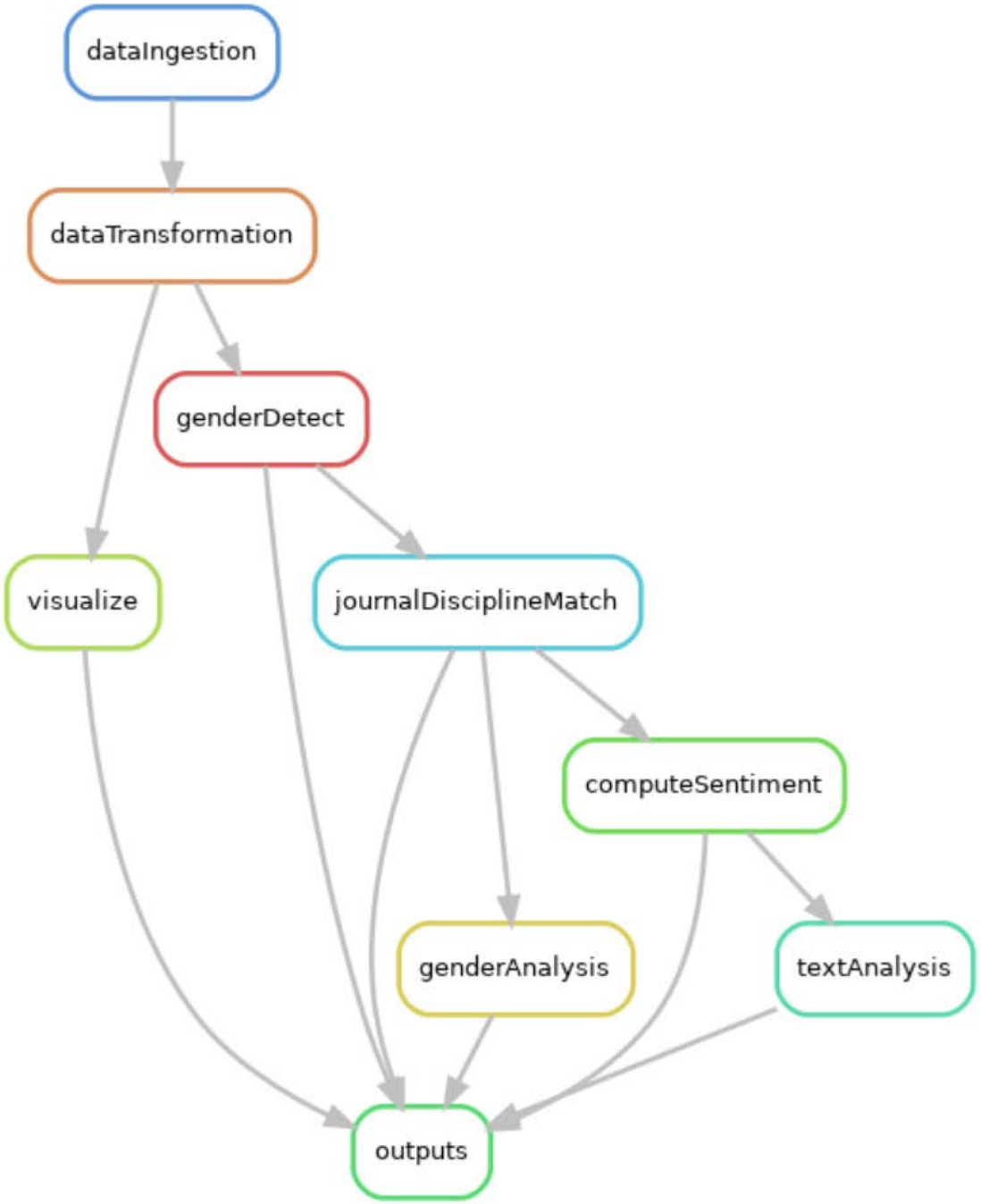
Empowered by the open-source ecosystem





Reproducible notebooks on Kubernetes

https://github.com/iuni-cadre/ReproducibilityDemo/wiki/A-demo-of-reproducibility



Demo 5

https://github.com/iuni-cadre/ISSI-tutorial

Reproducible notebooks on Kubernetes

binder

(beta)

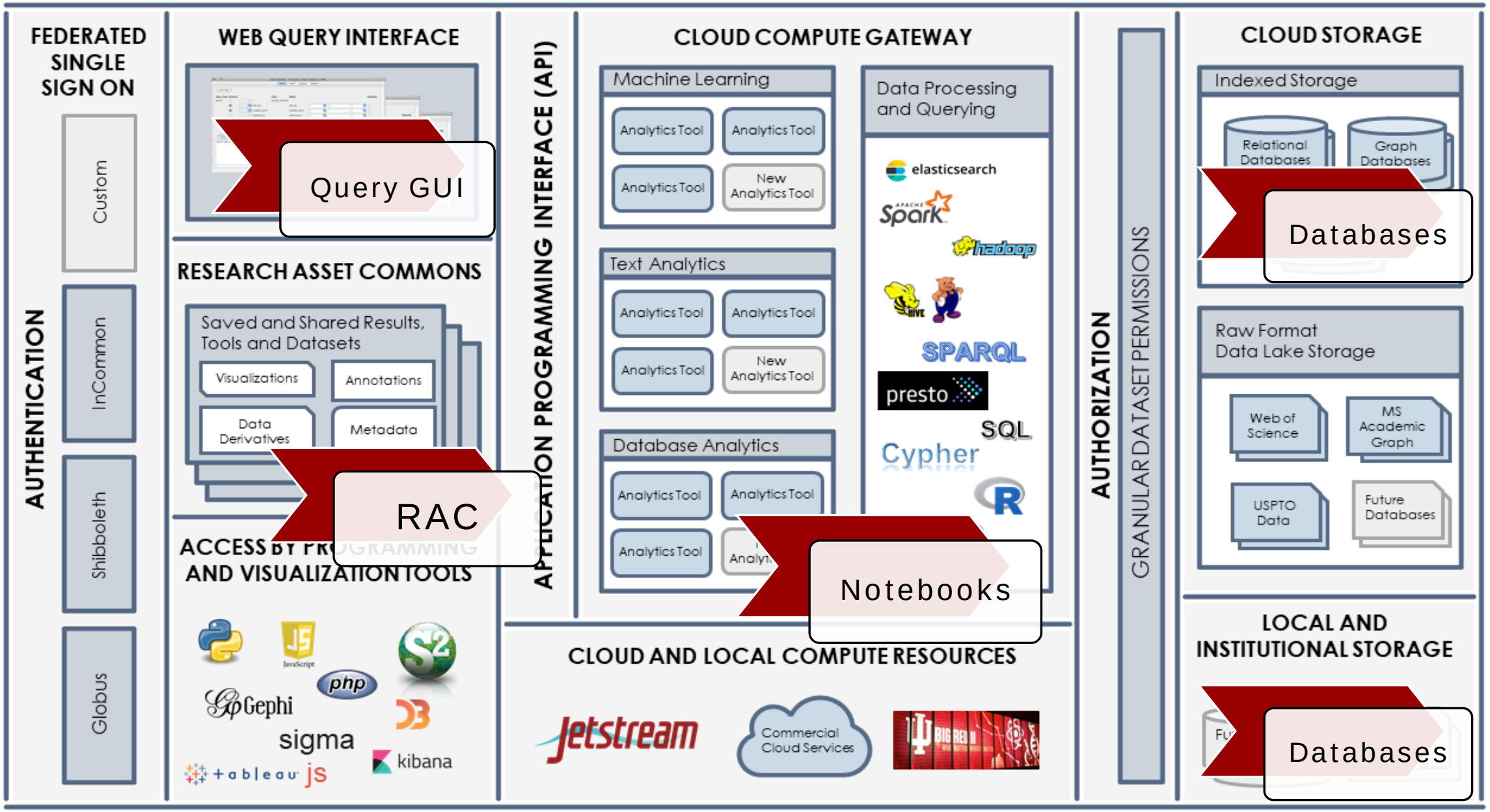
Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

The CADRE ecosystem

3rd party	•Plugins and extensions •Computing resources •Other data sets
RAC	•Package marketplace •Derivatives data •Pipeline builder
CADRE core	•Centralized databases •Data API •Coding environment

SHARED BIGDATA-GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)



CADRE
Collaborative Archive & Data Research Environment

Microsoft Research

ISSI

Q&A

The CADRE TEAM

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

B1G
ACADEMIC ALLIANCE

CADRE related events

Apr. 2019

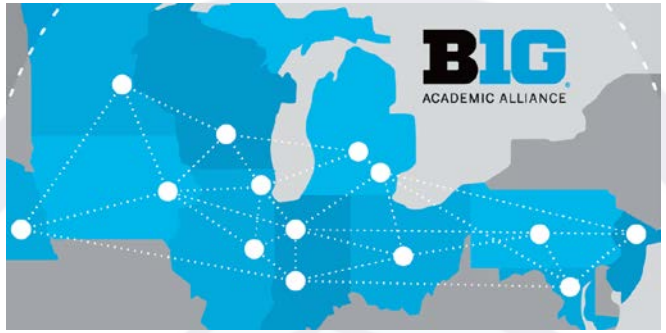
- 2019 CADRE meeting
- CADRE Fellowship open

Sep. 2019

- 1st Fellows announced
- ISSI workshop & tutorial

May. 2020

- 2020 CADRE meeting
- BTAA Library Conference 2020
- 2020 CADRE hackathon



ISSI

SAPIENZA
UNIVERSITÀ DI ROMA

INDIANA UNIVERSITY
BLOOMINGTON

Contact Us

https://cadre.iu.edu

cadre@iu.edu

@CADRE_Project

Tutorial Resources

- https://cadre.iu.edu/
- https://cadre.iu.edu/news-and-events/events/rome
- https://github.com/iuni-cadre/ISSI-tutorial