Deep Mining Scholarly Big Data in the Large Language Model Era

DR. JIAN WU
ASSOCIATE PROFESSOR OF COMPUTER SCIENCE
OLD DOMINION UNIVERSITY, VIRGINIA, UNITED STATES

Self-Introduction

2004: B.S. in Physics and Astronomy

2011: Ph.D. in Astronomy and Astrophysics

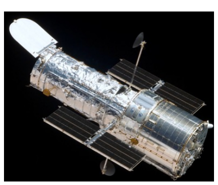
Big Data

2011-2017: Postdoctoral fellow, Information Sciences and Technology

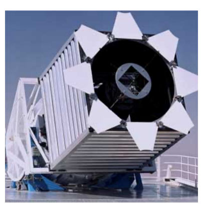
2017-2018: Assistant Teaching Professor, Information Sciences and Technology

2018-2025: Assistant Professor (tenure track), Computer Science

2025-: Associate Professor (effective July 25), Computer Science



Hubble Space Telescope



Sloan Digital Sky Survey

CiteSeer^{X}

Scholarly Big Data + AI

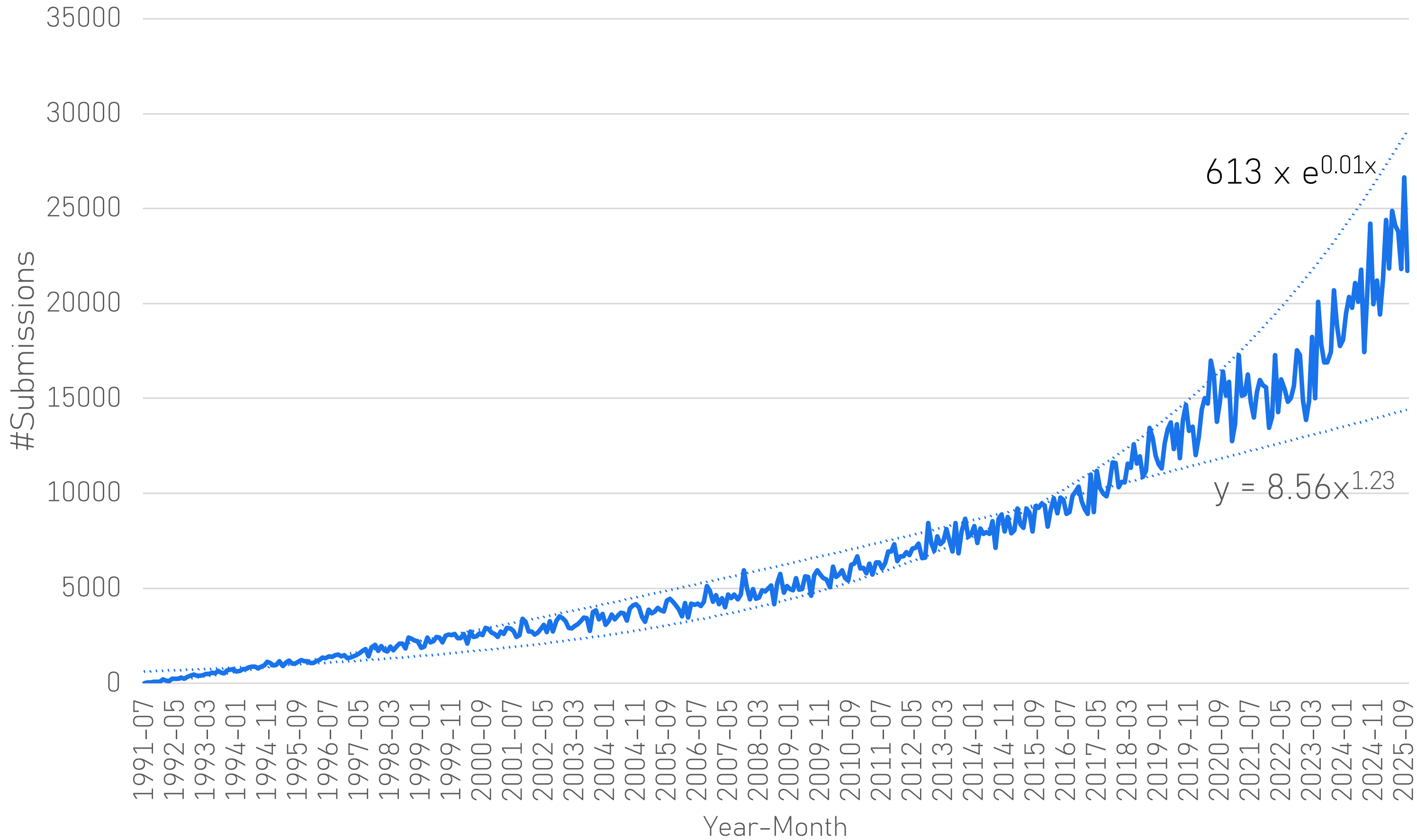
Conference Proceedings | ETDs | KGs | Tables

Journals | Technical Drawings | Keyphrases | Figures

Acknowledgments

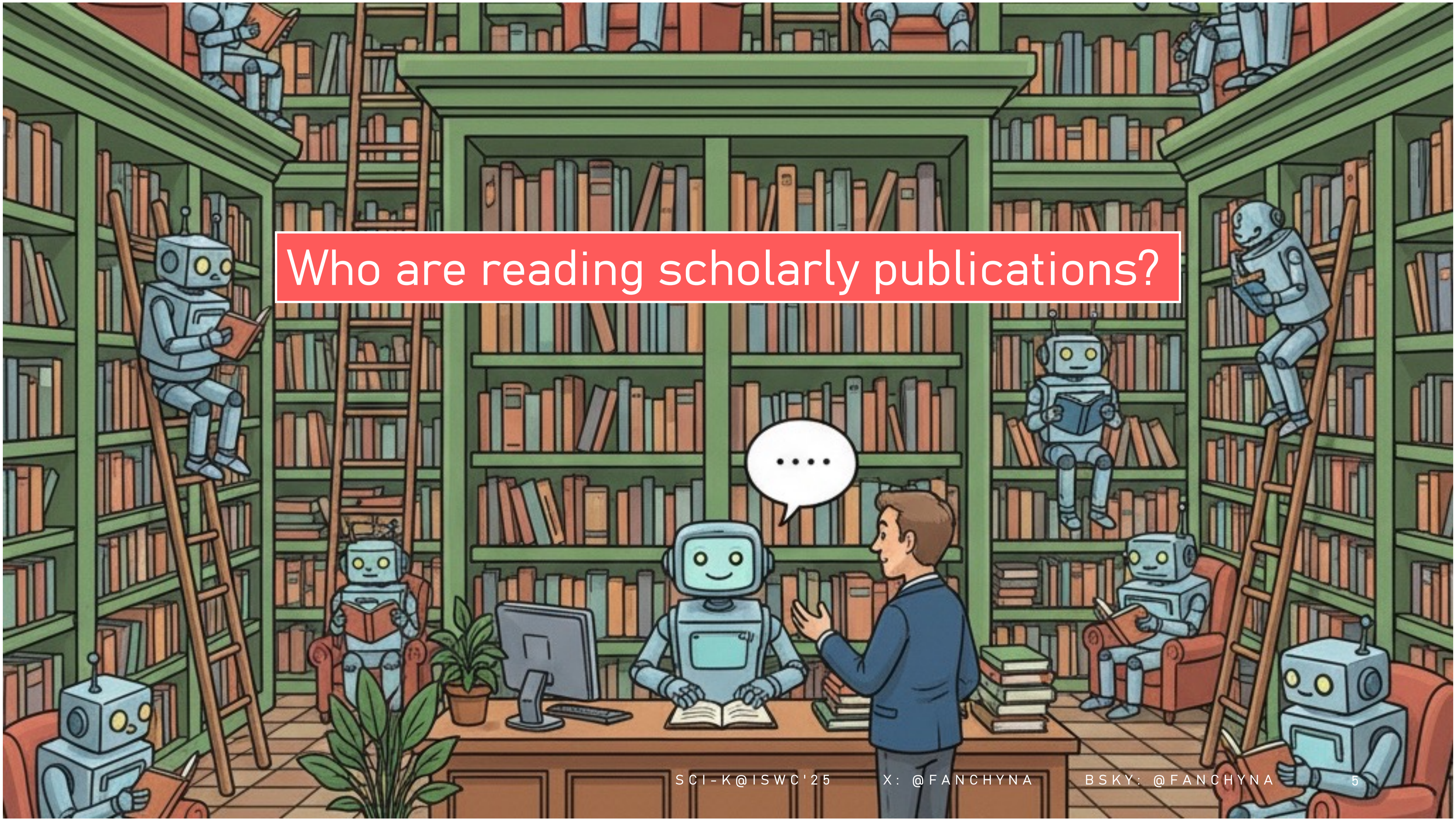
- Dr. C. Lee Giles (Pennsylvania State University)
  - PI of CiteSeerX
  - Eminent David Reese Professor of Information Sciences and Technology
  - A pioneer of AI and its applications on Scholarly Big Data
- Dr. Edward A. Fox (Virginia Tech)
  - Director of NDLTD
  - Eminent Professor of Computer Science
  - An advocate of collecting, mining, and improving electronic theses and dissertations (ETDs) and building communities

- This chart displays the number of new submissions received during each month since August 1991 (after 34.3 years).
- Total number of submissions as of October 27, 2025
= 2,867,652.



https://arxiv.org/stats/monthly_submissions

Who are reading scholarly publications?



Scholarly Big Data

- First mentioned in Dr. Lee Giles' keynote for CIKM in 2013
- Usually refers to the large-scale digital data generated throughout the scholarly communication and research activity.
  - Publications and citations, e.g., proceedings, references
  - Research artifacts, e.g., data, software
  - Scholarly communications, e.g., reviews
  - Scientometric, e.g., citation counts, h-index
  - Researcher and institutional metadata, e.g., author profiles

Mining Scholarly Big Data

- Information Extraction
  - Metadata extraction (title, author, keywords)
  - Table and Figure extraction
  - Citation extraction
  - Knowledge extraction (entities, relations)
  - Data extraction (table data, figure data)
  - Reasoning extraction (hypothesis, evidence)
- Information Classification
  - Document classification (subject category)
  - Scientific Claim Verification (true/false, stance)
  - Author name disambiguation
- Information Generation
  - Figure captioning
  - Hypothesis generation
- Applications
  - Reproducibility and replicability assessment
  - Data compilation
  - Building digital libraries and datasets

4 Retrospective Eras

- Metadata-centric Era (1990s-2010s)
- Semantic Enrichment Era (2010s-2018)
- Content-based Mining Era (2018-2022)
- Semantic Reasoning Era (2023 - present)

Metadata-centric Era (1990s–2010s)

- Driven by digital libraries and availability of network data
- Metadata: Titles, authors, publication year, venues, citations, etc.

Digital Libraries

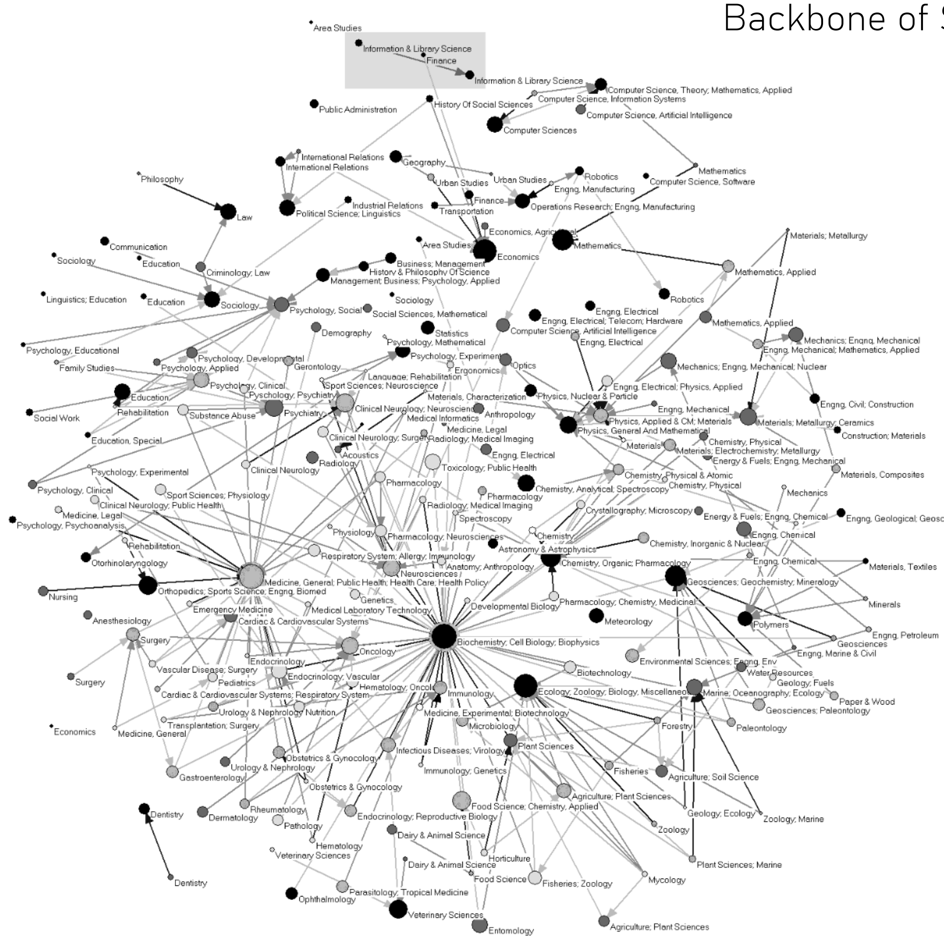
- Documents are organized in a connected manner, by inverse index, citation networks, co-author networks, or other structures so that they are more findable, navigable, and usually provide meta-level knowledge.
  - NDLTD (1996 – present)
  - Web of Science (1997 – present)
  - CiteSeer (1998 - present)
  - Google Scholar (2007 - present)
  - AMiner (2008 - present)

2003

Early Studies on Citation Networks

Boyack, Klavans, and Böner (2005) Mapping the Backbone of Science

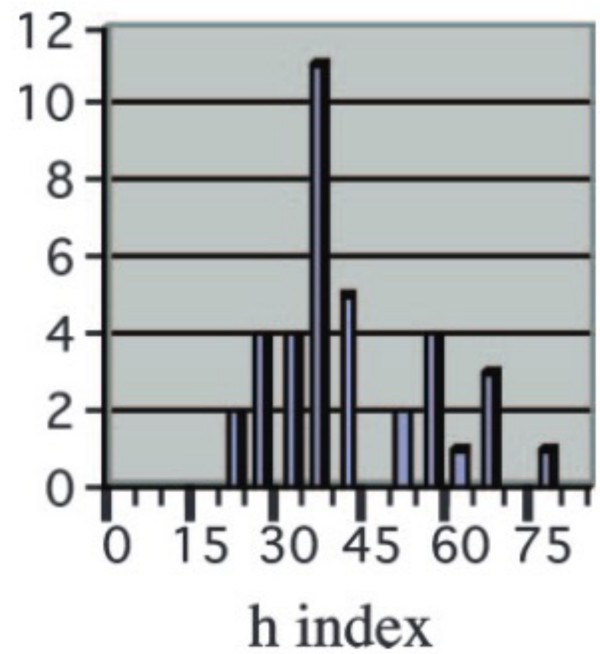
- Map of the backbone of science with 212 clusters comprising 7000 journals.
- Circle sizes (area) denote the number of journals in each cluster.
- Circle color depicts the independence of each cluster, with darker colors depicting greater independence.
- Dominant cluster-to-cluster citing patterns are indicated by arrows. Arrows show all relationships where the citing cluster gives more than 7.5% of its total citations to the cited cluster, with darker arrows indicating a greater fraction of citations given by the citing cluster.



WEB OF SCIENCE

H-index: an Impact Evaluation Metric

Histogram giving the number of Nobel prize recipients in physics in the last 20 years versus their h index.



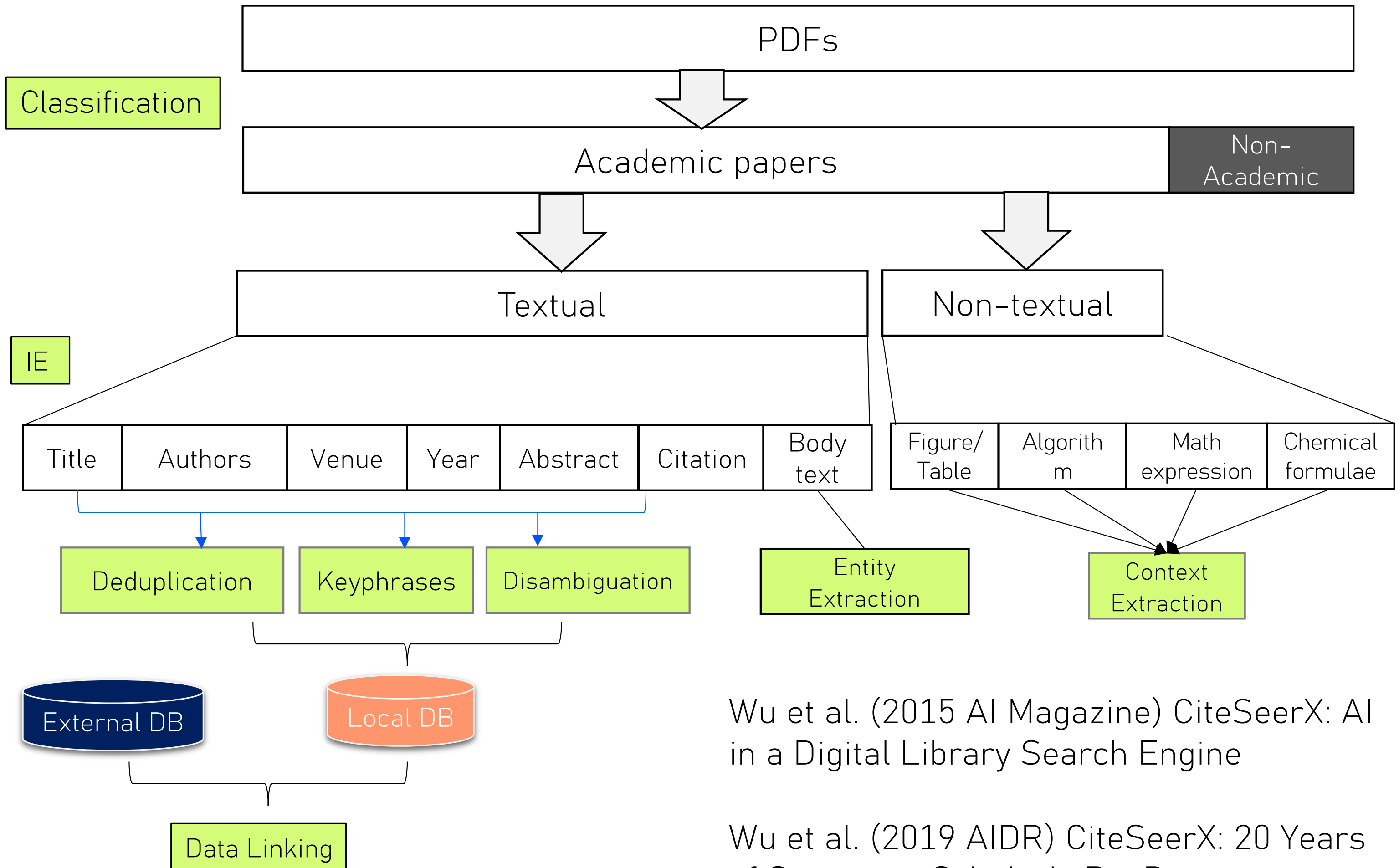
Hirsch (2005 PNAS) An index to quantify an individual's scientific research output

Semantic Enrichment Era (2010–2018)

- Machine learning and deep learning
- Entity extraction and linking (knowledge graph)
- Subject category classification
- Keyphrase extraction
- Digital Libraries and Datasets:
  - OpenAIRE (2011 -- present)
  - CORE (2012 -- present)
  - Semantic Scholar (2015 -- present)
  - Microsoft Academic Graph (2015 – 2021)

CiteSeerX

AI in CiteSeerX



Wu et al. (2015 AI Magazine) CiteSeerX: AI in a Digital Library Search Engine

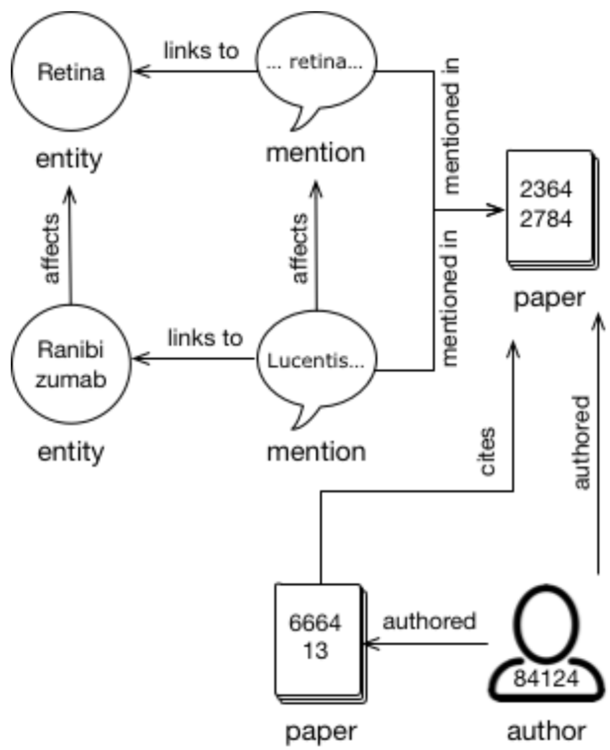
Wu et al. (2019 AIDR) CiteSeerX: 20 Years of Service to Scholarly Big Data

Semantic Scholar

- Basic functionalities
  - Search
  - Browse
  - Download

- AI-powered functionalities
  - TLDR summarization
  - Citation intent and influence classifications
  - Field of study classification
  - Paper recommendation
  - Metrics (most influential citations, etc.)

Building Digital Library Knowledge Graphs



Ammar et al. (2018 NAACL-HLT) Construction of the Literature Graph in Semantic Scholar

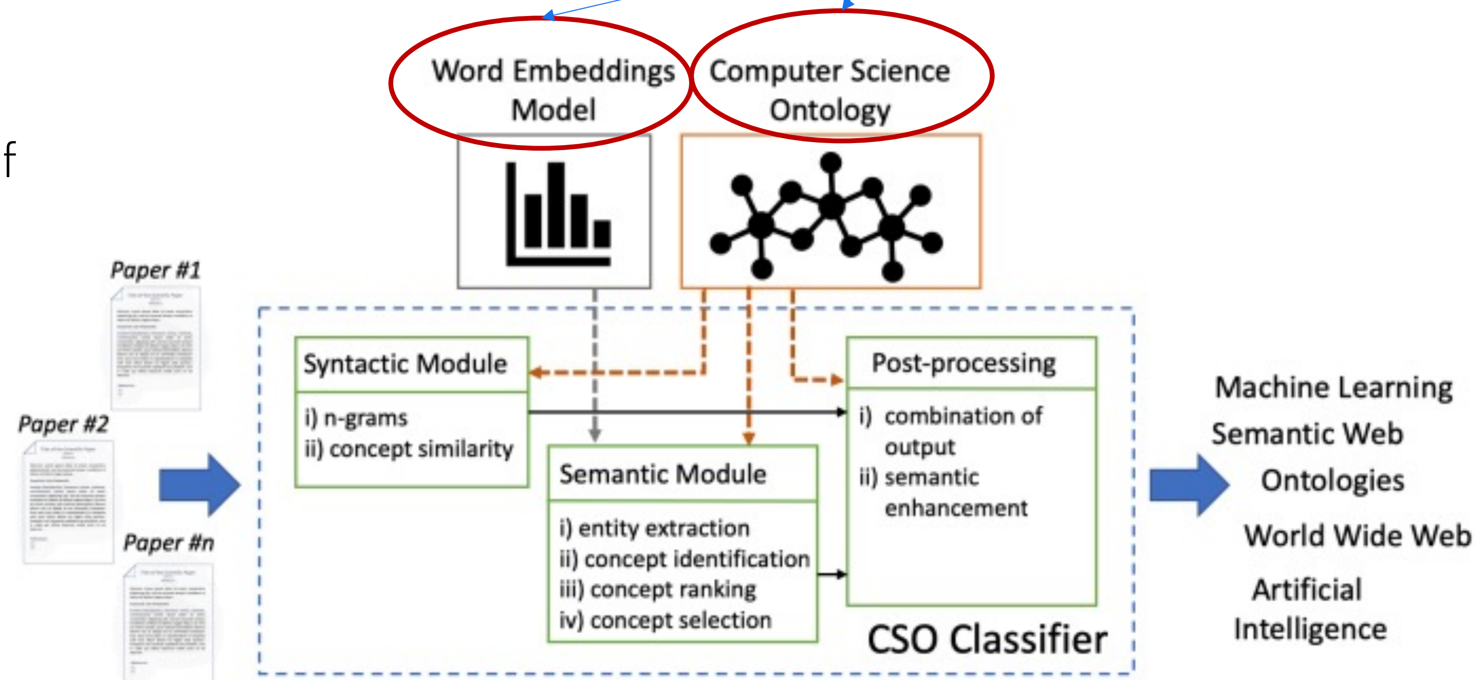
Part of the literature graph.

- Publisher provided metadata is often noisy and incomplete, it is often necessary to directly extract metadata from the PDFs.
- Machine Learning is heavily used for metadata extraction, entity extraction, and linking.
- GROBID (Lopez et al. 2009): CRF
- ScienceParse (AI2) BiLSTM

Topic Classification

In addition to ontology, word embedding was used.

The architecture of the workflow of the Computer Science Ontology (CSO) classifier.



Salatino et al. (TPDL 2019) The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles

Content-based Mining Era (2018-2022)

- Mining components constituting scholarly publications
- Project 1: SCORE: Automatically assign explainable “confidence scores” to Social and Behavioral Science (SBS) research results and claims
  - Theory/model extraction
  - Open Access Datasets and Software (OADS) URL extraction
  - Application of mined features: reproducibility and replicability assessment
- Project 2: Uncertainty-aware data extraction from complex scientific tables
  - Conformal prediction for quantifying the uncertainty of table data extraction

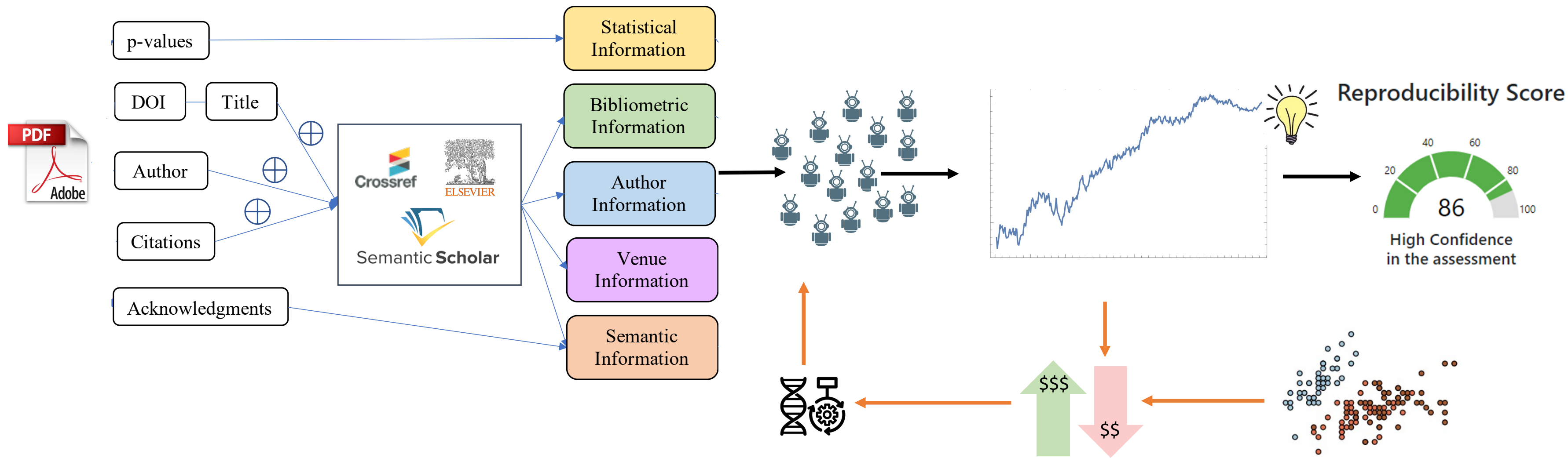
Why Assessing Reproducibility and Replicability?

- Reproducibility: same data, same method
- Replicability: different data, same method
- Reproducibility and replicability crisis in
  - Social and Behavioral Science (SBS) (Camerer 2016 Nature; Camerer 2018 Nature)
  - Computer Science (Moraila et al. 2014 PloS; Collberg et al. 2016)
  - Artificial Intelligence (Raff et al. 2019 NeurIPS; Gundersen et al. 2018 AAAI; Haibe-Kains et al. 2020 Nature; Ajayi et al. 2023 ICDAR)
  - Biomedical Science (Gentleman et a. 2005)

Manual Reproduction and Replication are Not Scalable

- Average time to reproduce the main results in one paper
  - **Reproduce**: Table Structure Recognition (an AI task): 8 hours (using code and data provided by the original authors; Ajayi et al. 2023 ICDAR)
  - **Reproduce**: General AI tasks: 53.5 days (using re-implemented codes and data provided by the original authors; Raff 2023 AAAI)
  - **Replicate**: Social and Behavioral Science: months – up to 1 year (using the same methods and new data collected from new user studies)

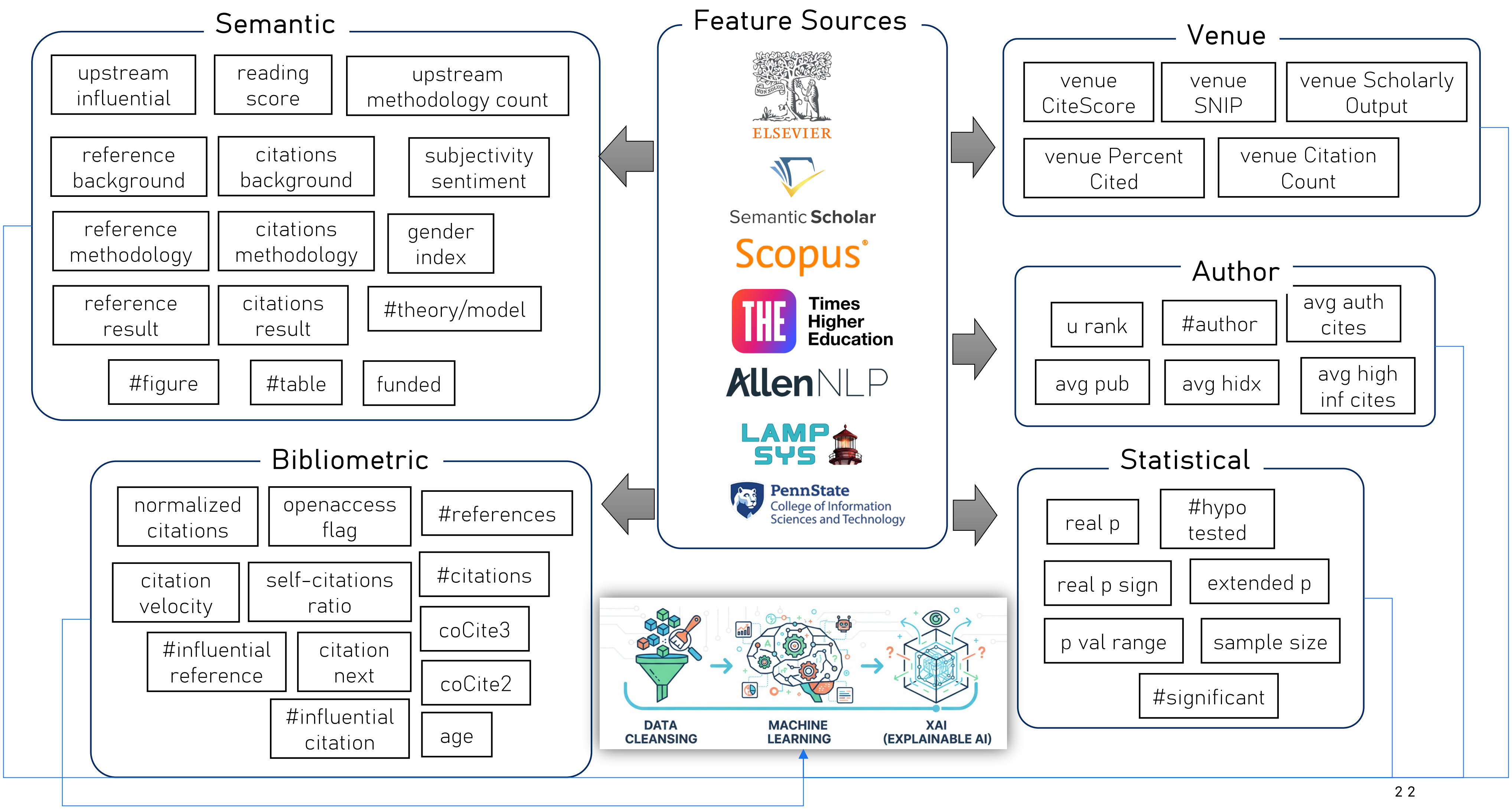
SCORE: A Synthetic Prediction Market for Estimating Confidence in Published Work (Rajtmajer et al. 2022 AAAI)



Synthetic prediction markets— *Prediction markets populated by artificial agents (trader-bots), trained and updated within human-expert prediction markets, but deployable "offline".*

- Trader-bots will represent atomic (human-interpretable) properties of relevant signals, including features extracted fromm the full text, metadata, and evaluation metrics after the paper is published.
- Bots will learn trading patterns from subject matter experts engaged in prediction markets, but unlike their human counterparts, will have comprehensive, unbiased view of the existing literature and metadata.

Feature extraction is the prerequisite!



Semantic
upstream influential
reading score
upstream methodology count
reference background
citations background
subjectivity sentiment
reference methodology
citations methodology
gender index
reference result
citations result
#theory/model
#figure
#table
funded
Feature Sources
ELSEVIER
Semantic Scholar
Scopus
THE
Times Higher Education
AllenNLP
LAMP SYS
PennState
College of Information Sciences and Technology
Venue
venue CiteScore
venue SNIP
venue Scholarly Output
venue Percent Cited
venue Citation Count
Author
u rank
#author
avg auth cites
avg pub
avg hidx
avg high inf cites
Bibliometric
normalized citations
openaccess flag
#references
citation velocity
self-citations ratio
#citations
coCite3
#influential reference
citation next
coCite2
#influential citation
age
Statistical
real p
#hypo tested
real p sign
extended p
p val range
sample size
#significant
DATA CLEANSING
MACHINE LEARNING
XAI (EXPLAINABLE AI)

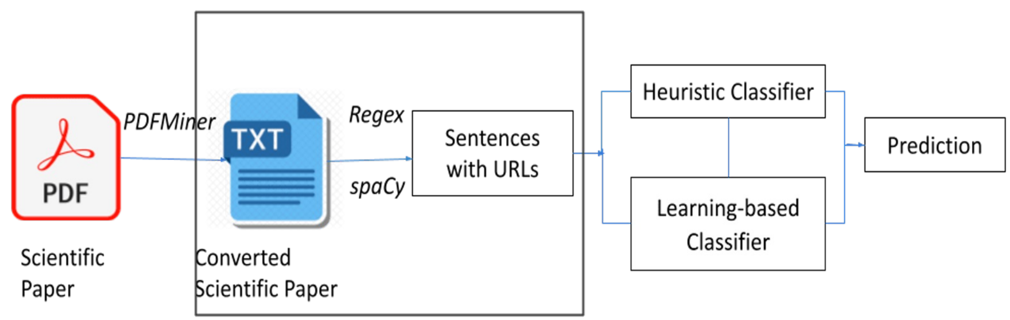
Open-Access Datasets and Software (OADS) URL Extraction

The dataset is publicly available and searchable online at http://www.nrel.gov/lci/database/.

Dataset URL

All of the presented structural figures were produced using pymol (http://pymol.sourceforge.net).

Software URL

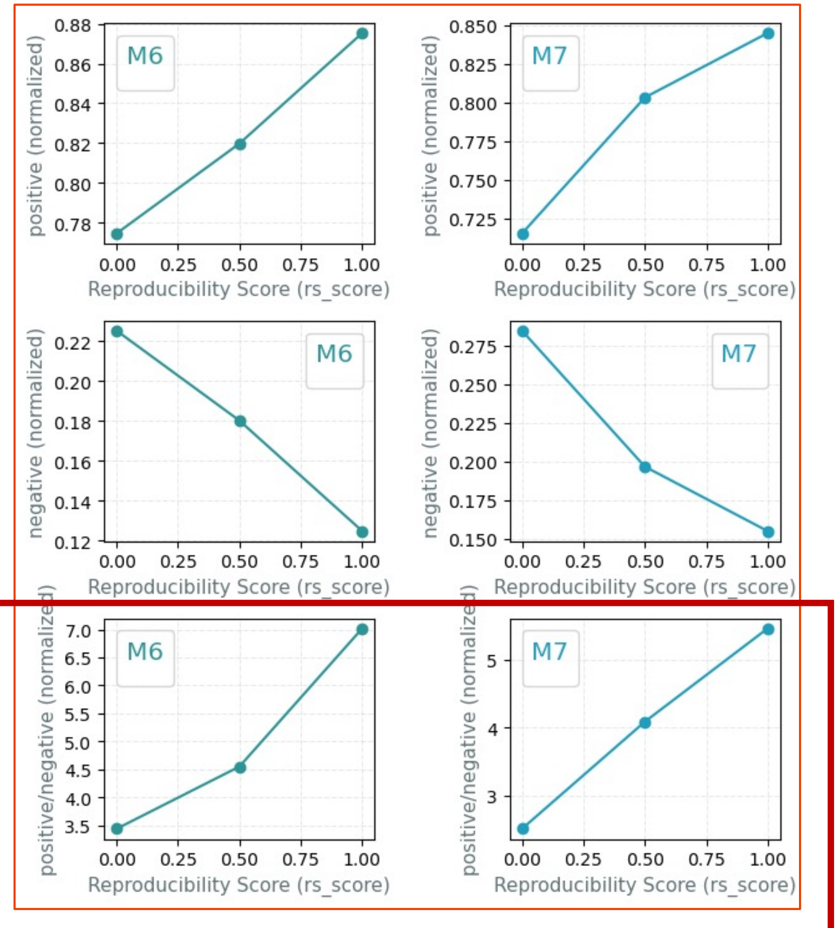


The Architecture of the hybrid OADS URL classifier.

Salsabil et al. (2022 Sci-K) A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software

Citation Context Sentiments vs. Reproducibility Scores

- Positive correlation between
  - ratio of #citation contexts (positive sentiments) over #citation context (negative sentiment)
  - reproducibility score



In this paper, we make a first attempt towards the second question, by studying a family of algorithms named DirectSet, in which the DirectPred algorithm proposed by **Tian et al. (2021)** is a special case with *positive*

Although we tried to train split network with the same training data we used, we failed to reproduce their results and used the model trained by the authors **[32]**. *negative*

Obadage et al. (2024 ACM REP) Can Citations Tell Us About a Paper's Reproducibility? A Case Study of Machine Learning Papers

Our Ongoing Research

- Goal: building a new benchmark for LLM agents to replicate claims in Social and Behavioral Science papers
- Data : 200+ papers, pre-registrations, and human replication study reports from the SCORE project (Nosek et al. 2021 ARP)
- Replicability (new data), multi-difficulty level, multi-stage, no-human involved

CENTER FOR OPEN SCIENCE

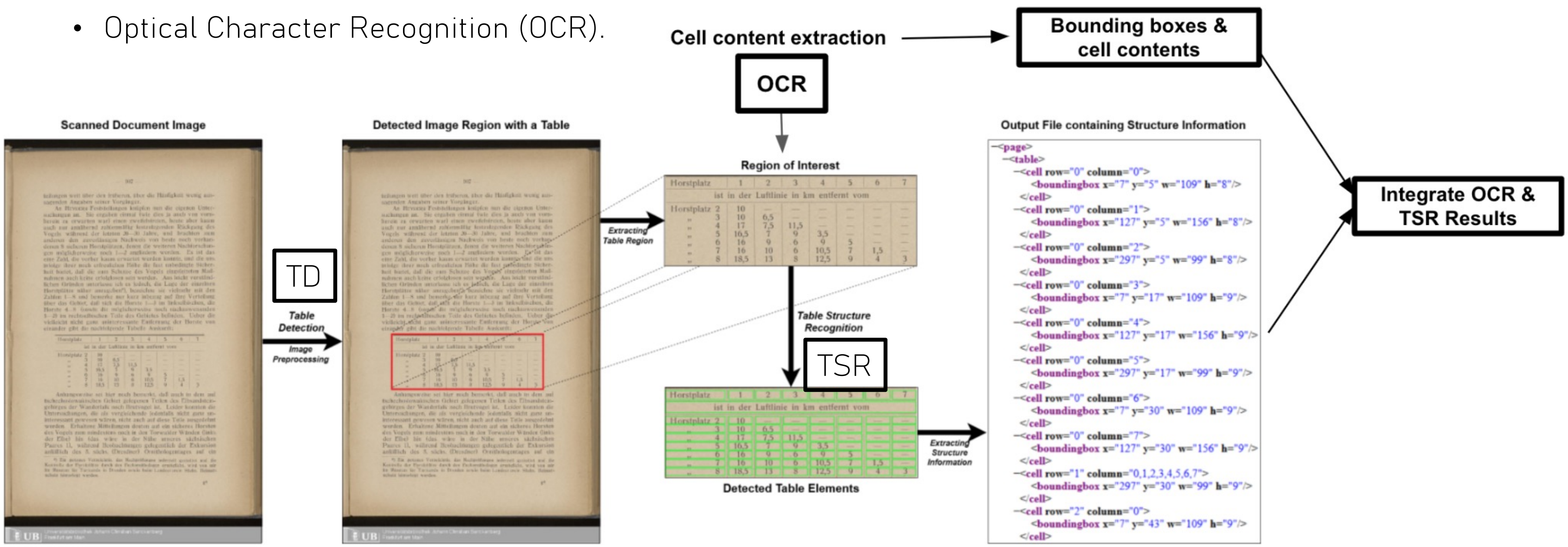
PennState

UNIVERSITY OF NOTRE DAME

OLD DOMINION UNIVERSITY

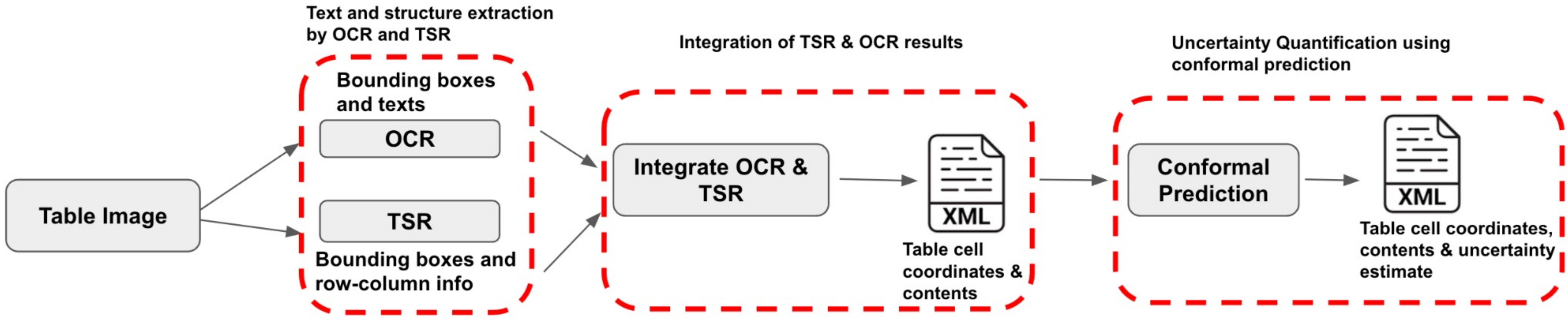
Project 2: Table Data Extraction

- Identifying and extracting structured data from tables embedded in PDFs and scanned images.
  - Table detection (TD)
  - Table structure recognition (TSR), and
  - Optical Character Recognition (OCR).



Uncertainty-Aware Table Data Extraction

- Why doing this? Existing data extraction methods usually report an overall performance of precision, recall, and F1 and do not estimate the uncertainty of extracted data at the cell level.
- Method 2: Conformal Prediction

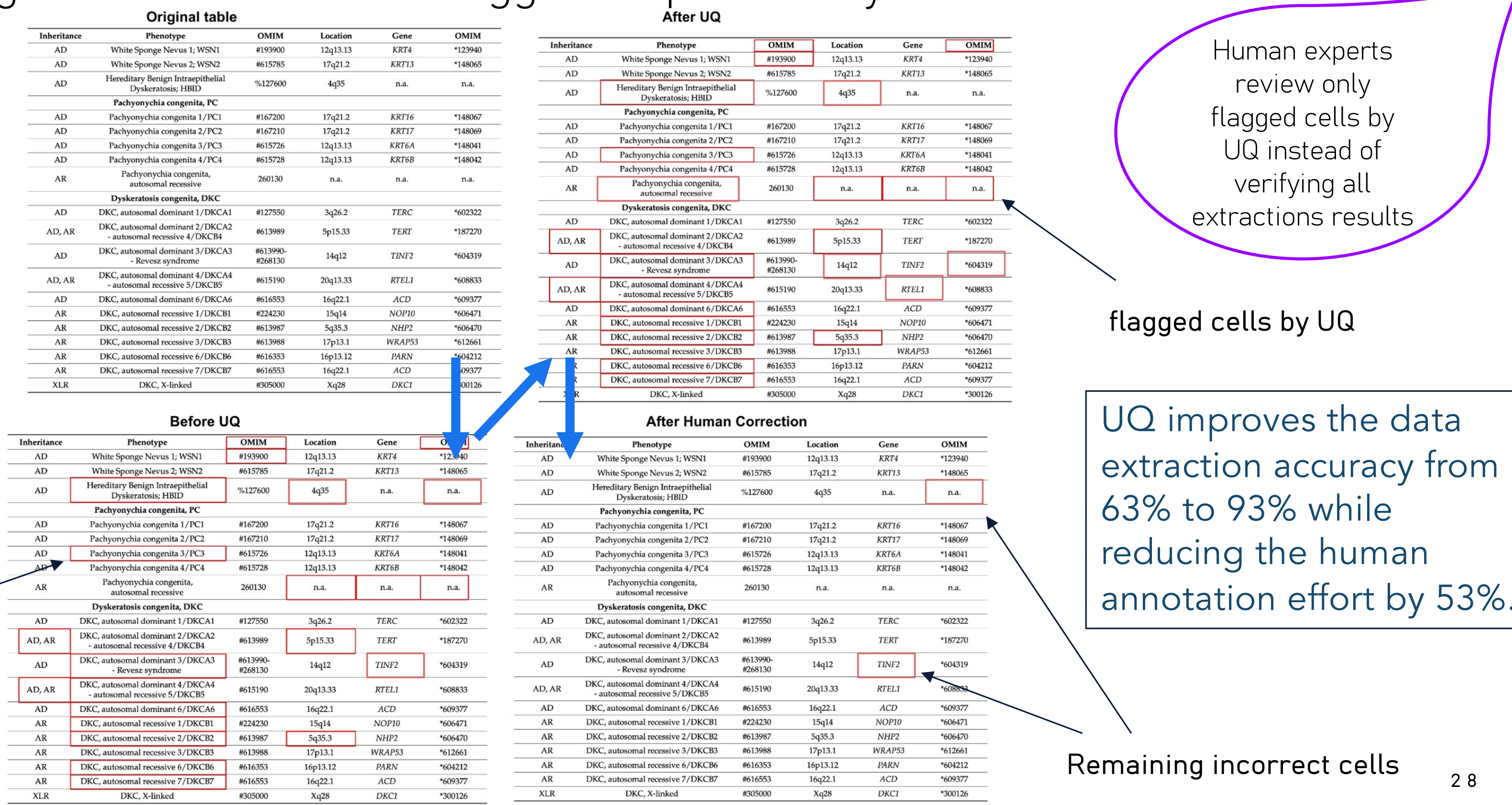


Ajayi et al. (2025 ICDAR) Uncertainty-Aware Complex Scientific Table Data Extraction

Uncertainty-Aware Table Data Extraction

Ajayi et al. (2025 ICDAR)
Uncertainty-Aware Complex Scientific Table Data Extraction

- Cells with high uncertainties are flagged as potentially incorrect.



Original table

Inheritance	Phenotype	OMIM	Location	Gene	OMIM
AD	White Sponge Nevus 1; WSN1	#193900	12q13.13	*KRT4*	*123940
AD	White Sponge Nevus 2; WSN2	#615785	17q21.2	*KRT13*	*148065
AD	Hereditary Benign Intraepithelial Dyskeratosis; HBID	%127600	4q35	n.a.	n.a.
	Pachyonychia congenita, PC				
AD	Pachyonychia congenita 1/PC1	#167200	17q21.2	*KRT16*	*148067
AD	Pachyonychia congenita 2/PC2	#167210	17q21.2	*KRT17*	*148069
AD	Pachyonychia congenita 3/PC3	#615726	12q13.13	*KRT6A*	*148041
AD	Pachyonychia congenita 4/PC4	#615728	12q13.13	*KRT6B*	*148042
AR	Pachyonychia congenita, autosomal recessive	260130	n.a.	n.a.	n.a.
	Dyskeratosis congenita, DKC				
AD	DKC, autosomal dominant 1/DKCA1	#127550	3q26.2	*TERC*	*602322
AD, AR	DKC, autosomal dominant 2/DKCA2 - autosomal recessive 4/DKCB4	#613989	5p15.33	*TERT*	*187270
AD	DKC, autosomal dominant 3/DKCA3 - Revesz syndrome	#613990-#268130	14q12	*TINF2*	*604319
AD, AR	DKC, autosomal dominant 4/DKCA4 - autosomal recessive 5/DKCB5	#615190	20q13.33	*RTEL1*	*608833
AD	DKC, autosomal dominant 6/DKCA6	#616553	16q22.1	*ACD*	*609377
AR	DKC, autosomal recessive 1/DKCB1	#224230	15q14	*NOP10*	*606471
AR	DKC, autosomal recessive 2/DKCB2	#613987	5q35.3	*NHP2*	*606470
AR	DKC, autosomal recessive 3/DKCB3	#613988	17p13.1	*WRAP53*	*612661
AR	DKC, autosomal recessive 6/DKCB6	#616353	16p13.12	*PARN*	*604212
AR	DKC, autosomal recessive 7/DKCB7	#616553	16q22.1	*ACD*	09377
XLR	DKC, X-linked	#305000	Xq28	*DKC1*	00126

After UQ

Inheritance	Phenotype	OMIM	Location	Gene	OMIM
AD	White Sponge Nevus 1; WSN1	#193900	12q13.13	*KRT4*	*123940
AD	White Sponge Nevus 2; WSN2	#615785	17q21.2	*KRT13*	*148065
AD	Hereditary Benign Intraepithelial Dyskeratosis; HBID	%127600	4q35	n.a.	n.a.
	Pachyonychia congenita, PC				
AD	Pachyonychia congenita 1/PC1	#167200	17q21.2	*KRT16*	*148067
AD	Pachyonychia congenita 2/PC2	#167210	17q21.2	*KRT17*	*148069
AD	Pachyonychia congenita 3/PC3	#615726	12q13.13	*KRT6A*	*148041
AD	Pachyonychia congenita 4/PC4	#615728	12q13.13	*KRT6B*	*148042
AR	Pachyonychia congenita, autosomal recessive	260130	n.a.	n.a.	n.a.
	Dyskeratosis congenita, DKC				
AD	DKC, autosomal dominant 1/DKCA1	#127550	3q26.2	*TERC*	*602322
AD, AR	DKC, autosomal dominant 2/DKCA2 - autosomal recessive 4/DKCB4	#613989	5p15.33	*TERT*	*187270
AD	DKC, autosomal dominant 3/DKCA3 - Revesz syndrome	#613990-#268130	14q12	*TINF2*	*604319
AD, AR	DKC, autosomal dominant 4/DKCA4 - autosomal recessive 5/DKCB5	#615190	20q13.33	*RTEL1*	*608833
AD	DKC, autosomal dominant 6/DKCA6	#616553	16q22.1	*ACD*	*609377
AR	DKC, autosomal recessive 1/DKCB1	#224230	15q14	*NOP10*	*606471
AR	DKC, autosomal recessive 2/DKCB2	#613987	5q35.3	*NHP2*	*606470
AR	DKC, autosomal recessive 3/DKCB3	#613988	17p13.1	*WRAP53*	*612661
R	DKC, autosomal recessive 6/DKCB6	#616353	16p13.12	*PARN*	*604212
R	DKC, autosomal recessive 7/DKCB7	#616553	16q22.1	*ACD*	*609377
R	DKC, X-linked	#305000	Xq28	*DKC1*	*300126

Before UQ

Inheritance	Phenotype	OMIM	Location	Gene	OMIM
AD	White Sponge Nevus 1; WSN1	#193900	12q13.13	*KRT4*	*12 40
AD	White Sponge Nevus 2; WSN2	#615785	17q21.2	*KRT13*	*148065
AD	Hereditary Benign Intraepithelial Dyskeratosis; HBID	%127600	4q35	n.a.	n.a.
	Pachyonychia congenita, PC				
AD	Pachyonychia congenita 1/PC1	#167200	17q21.2	*KRT16*	*148067
AD	Pachyonychia congenita 2/PC2	#167210	17q21.2	*KRT17*	*148069
AD	Pachyonychia congenita 3/PC3	#615726	12q13.13	*KRT6A*	*148041
AD	Pachyonychia congenita 4/PC4	#615728	12q13.13	*KRT6B*	*148042
AR	Pachyonychia congenita, autosomal recessive	260130	n.a.	n.a.	n.a.
	Dyskeratosis congenita, DKC				
AD	DKC, autosomal dominant 1/DKCA1	#127550	3q26.2	*TERC*	*602322
AD, AR	DKC, autosomal dominant 2/DKCA2 - autosomal recessive 4/DKCB4	#613989	5p15.33	*TERT*	*187270
AD	DKC, autosomal dominant 3/DKCA3 - Revesz syndrome	#613990-#268130	14q12	*TINF2*	*604319
AD, AR	DKC, autosomal dominant 4/DKCA4 - autosomal recessive 5/DKCB5	#615190	20q13.33	*RTEL1*	*608833
AD	DKC, autosomal dominant 6/DKCA6	#616553	16q22.1	*ACD*	*609377
AR	DKC, autosomal recessive 1/DKCB1	#224230	15q14	*NOP10*	*606471
AR	DKC, autosomal recessive 2/DKCB2	#613987	5q35.3	*NHP2*	*606470
AR	DKC, autosomal recessive 3/DKCB3	#613988	17p13.1	*WRAP53*	*612661
AR	DKC, autosomal recessive 6/DKCB6	#616353	16p13.12	*PARN*	*604212
AR	DKC, autosomal recessive 7/DKCB7	#616553	16q22.1	*ACD*	*609377
XLR	DKC, X-linked	#305000	Xq28	*DKC1*	*300126

After Human Correction

Inheritance	Phenotype	OMIM	Location	Gene	OMIM
AD	White Sponge Nevus 1; WSN1	#193900	12q13.13	*KRT4*	*123940
AD	White Sponge Nevus 2; WSN2	#615785	17q21.2	*KRT13*	*148065
AD	Hereditary Benign Intraepithelial Dyskeratosis; HBID	%127600	4q35	n.a.	n.a.
	Pachyonychia congenita, PC				
AD	Pachyonychia congenita 1/PC1	#167200	17q21.2	*KRT16*	*148067
AD	Pachyonychia congenita 2/PC2	#167210	17q21.2	*KRT17*	*148069
AD	Pachyonychia congenita 3/PC3	#615726	12q13.13	*KRT6A*	*148041
AD	Pachyonychia congenita 4/PC4	#615728	12q13.13	*KRT6B*	*148042
AR	Pachyonychia congenita, autosomal recessive	260130	n.a.	n.a.	n.a.
	Dyskeratosis congenita, DKC				
AD	DKC, autosomal dominant 1/DKCA1	#127550	3q26.2	*TERC*	*602322
AD, AR	DKC, autosomal dominant 2/DKCA2 - autosomal recessive 4/DKCB4	#613989	5p15.33	*TERT*	*187270
AD	DKC, autosomal dominant 3/DKCA3 - Revesz syndrome	#613990-#268130	14q12	*TINF2*	*604319
AD, AR	DKC, autosomal dominant 4/DKCA4 - autosomal recessive 5/DKCB5	#615190	20q13.33	*RTEL1*	*608833
AD	DKC, autosomal dominant 6/DKCA6	#616553	16q22.1	*ACD*	*609377
AR	DKC, autosomal recessive 1/DKCB1	#224230	15q14	*NOP10*	*606471
AR	DKC, autosomal recessive 2/DKCB2	#613987	5q35.3	*NHP2*	*606470
AR	DKC, autosomal recessive 3/DKCB3	#613988	17p13.1	*WRAP53*	*612661
AR	DKC, autosomal recessive 6/DKCB6	#616353	16p13.12	*PARN*	*604212
AR	DKC, autosomal recessive 7/DKCB7	#616553	16q22.1	*ACD*	*609377
XLR	DKC, X-linked	#305000	Xq28	*DKC1*	*300126

UQ improves the data extraction accuracy from 63% to 93% while reducing the human annotation effort by 53%.

Semantic Reasoning Era (2023 - present)

- Mining the reasoning knowledge and processes in scholarly publications using large language models (LLMs) and vision language models (VLMs)
  - Paper QA and SciTableQA
  - Scientific Claim Verification (Scientific Hypothesis Evidencing)
  - Hypothesis generation (ongoing)
  - Data Compilation (ongoing)
  - Hypothesis-evidence extraction (ongoing)

Semantic Reasoning has Gained Interests in Industry

- **AllSci**: hypothesis-centric, AI-powered
  - More than 12 million scientific hypotheses
  - Using AI-guided tools to help researchers formulate better hypotheses
- **Scite**: using citation context for QA and Table search
- **Consensus**: search engine + QA

https://guides.pnw.edu/AISearch

Paper QA – Offloading Reading to Bots

- Single document QA
  - Answer questions after reading a single scholarly document provided by a user
- Multi-document QA
  - Answer questions after reading multiple documents provided by a user

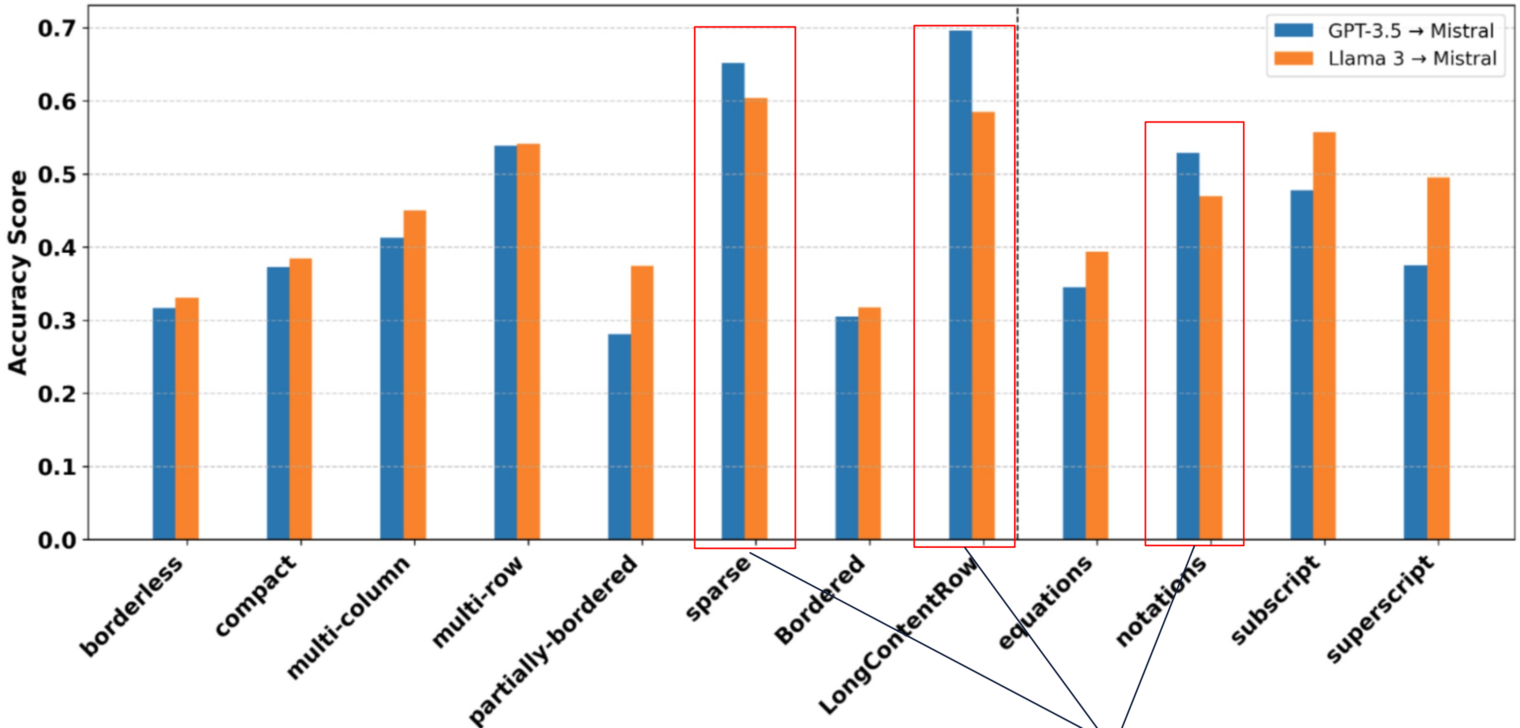
SciTableQA

- **Question**: What is the total reduction for Unary, NP Int, and Clause Att error types?
- **LLM Answer**: 337
- **LLM Explanation**: Add the reduction values for Unary (143), NP Int (129), and Clause Att (65) error types to get the total reduction.
- **GT: 337**
- **Reasoning validity**: score = 3 (explanation aligns with ground truth & leads to correct answer)

error type	bottom-up	top-down	reduction
PP Attach	850	668	182 (21%)
1-word Span	687	489	198 (29%)
Unary	555	412	143 (25%)
NP Int	464	335	129 (28%)
Clause Att	376	311	65 (17%)
Different label	374	269	105 (28%)
Mod Attach	317	264	53 (17%)
Co-ordination	379	225	154 (41%)
UNSET add	291	216	75 (26%)
NP Attachment	216	161	55 (25%)
Other	413	304	109 (26%)

Ajayi et al. (2025 TPDL) SciTableQA: A Question-Answering Benchmark for Complex Scientific Tables

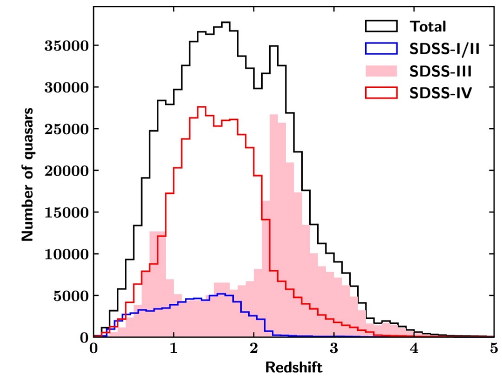
How Well Can LLM Do? Cross-LLM QA Evaluation



- **Llama-3 outperforms** GPT-3.5 for tables containing 9/12 complex features.
- **GPT-3.5 outperforms** Llama-3 for tables with 3 other features.

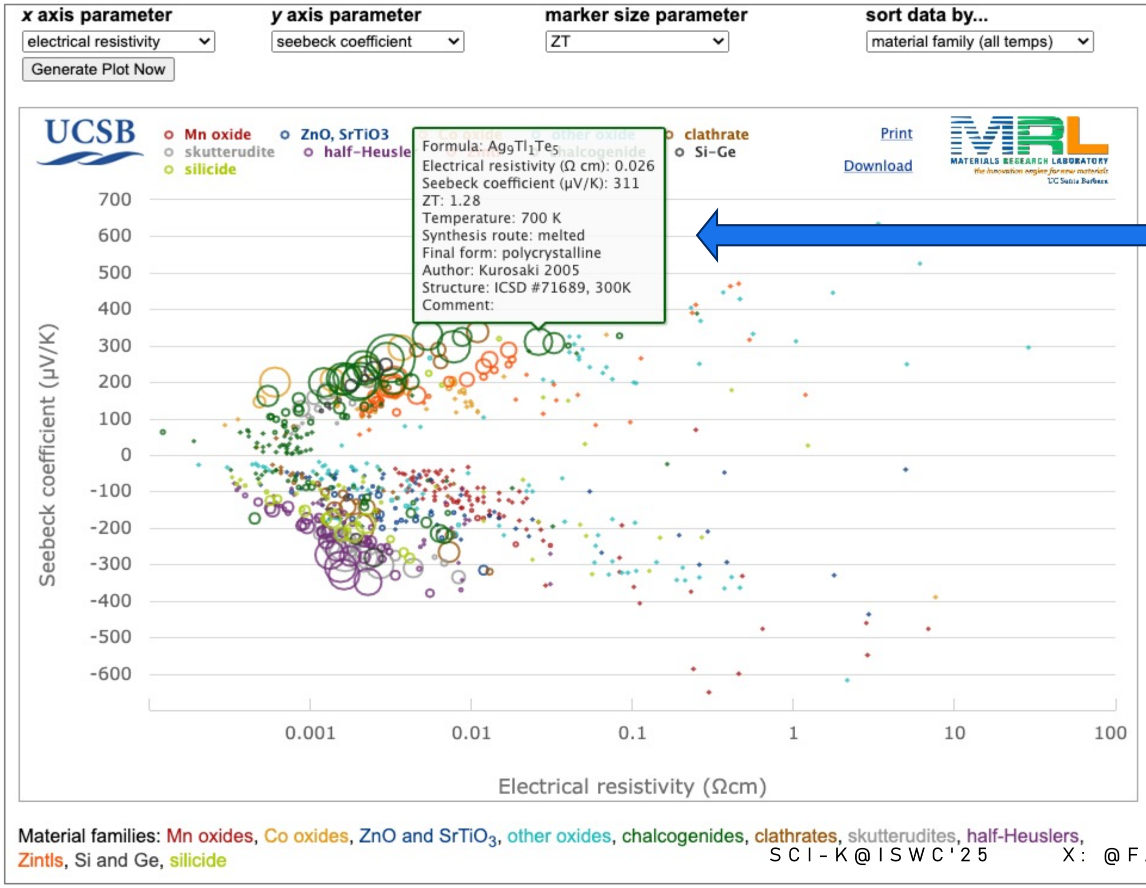
Data Compilation

- Automatic gathering data from multiple papers into a database
- Why important?
  - Tons of data are published in PDFs
  - Manually collecting data is very time-consuming
  - Need to compile data to get a sense of the answer to important questions in the literature (vs in just one paper at a time)
- Why is it hard? -- Usually needs reasoning and hard to generalize!



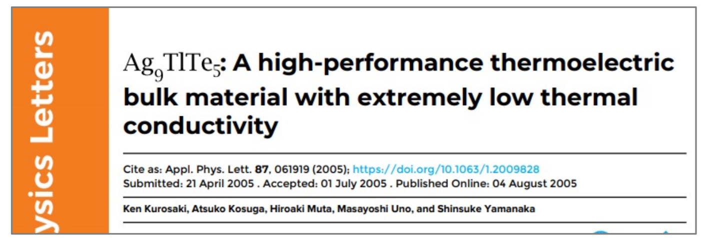
(Manually) Compiled Data in Materials Science

http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp



18,000 data points from 100+ papers

source paper


Physics Letters

Ag_9TlTe_5: A high-performance thermoelectric bulk material with extremely low thermal conductivity

Cite as: Appl. Phys. Lett. 87, 061919 (2005); https://doi.org/10.1063/1.2009828
Submitted: 21 April 2005 . Accepted: 01 July 2005 . Published Online: 04 August 2005

Ken Kurosaki, Atsuko Kosuga, Hiroaki Muta, Masayoshi Uno, and Shinsuke Yamanaka

https://aip.scitation.org/doi/10.1063/1.2009828

Gaultois et al. 2013

Our Work on Table Data Compilation

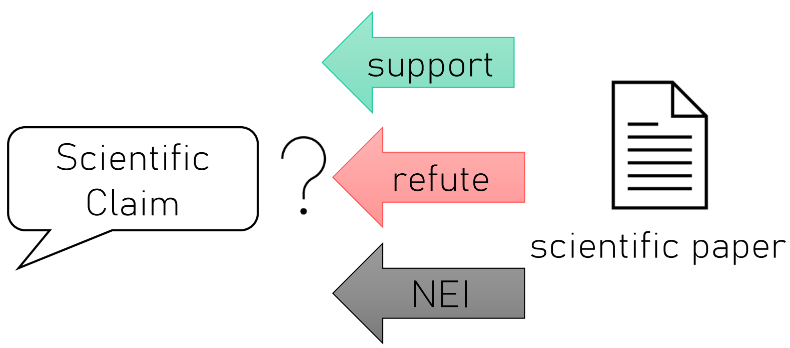
1. Data compilation usually involves two steps: extracting data from a table and populating the data into a a bigger table or a database
2. Neural models can achieve 60+% F1-scores on data Extraction (Ajayi et al. 2025 ICDAR), not including data population
3. VLMs can achieve anywhere between 50-90% F1-score on data Compilation using a pretty standard prompt (few-shot, no prompt tuning) (Domminage et al. ongoing project)

Challenge: Papers do not always follow a standard way to show data, and even measure data!

Ajayi et al. (2025 ICDAR) Uncertainty-Aware Complex Scientific Table Data Extraction

Scientific Claim Verification

- Problem definition: Given a claim and a scientific paper, can AI tell us if the paper supports or refutes claim (or does not provide enough information)?



Verifying Claims (hypotheses) in Scientific Papers

Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences

Sai Koneru[1], **Jian Wu**[2], **Sarah Rajtmajer**[1]
[1] Pennsylvania State University, State College, PA
[2] Old Dominion University, Norfolk, VA
{sdk96, smr48}@psu.edu, j1wu@odu.edu

LREC COLING 2024

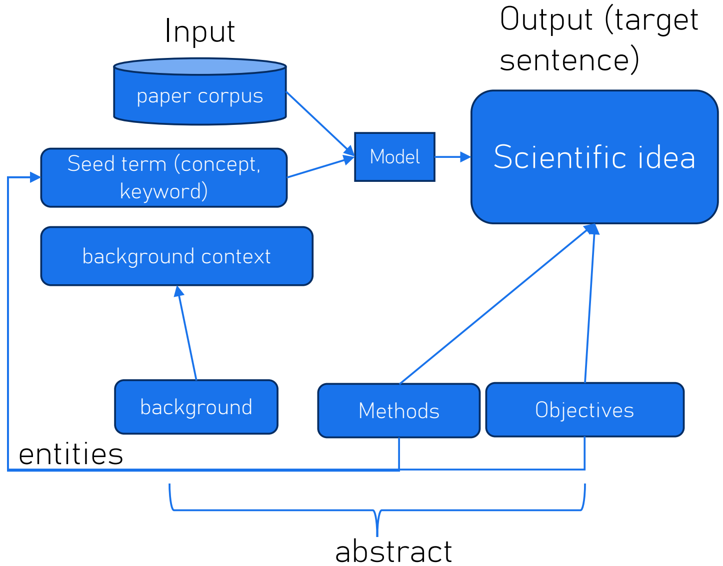
AI underperforms domain experts on verifying scientific hypotheses in social and behavioral sciences.

Best model of each type	Accuracy	Macro F1
embedding + supervised classification	70.31%	0.615
Transfer learning	67.97%	0.523
GPT3.5 few-shot	66.57%	0.576
PaLM2	62.87%	0.536

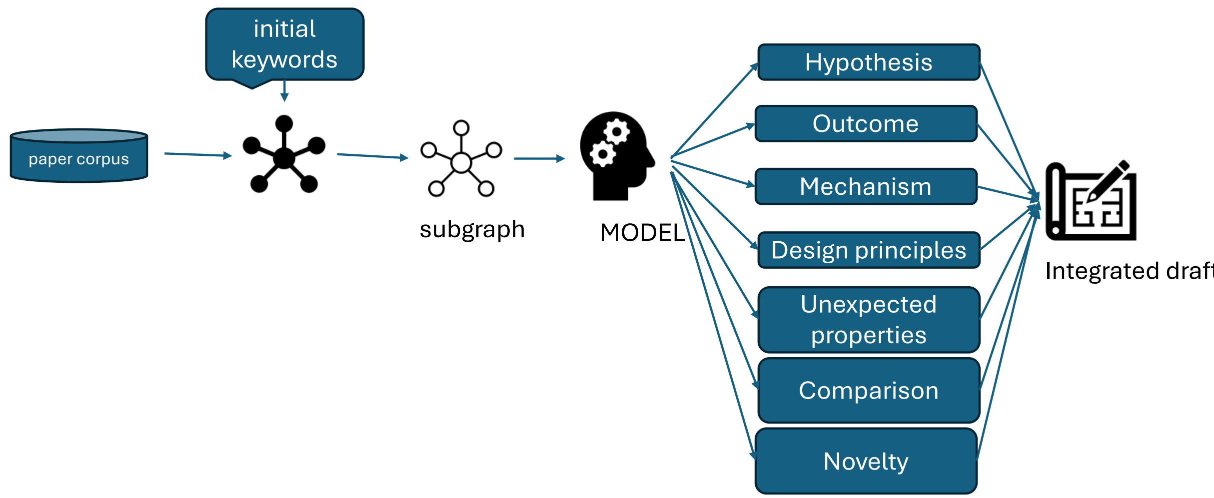
Data: 69 distinct hypotheses and 637 documents.

Hypothesis Generation

- Can AI help scientists propose novel and feasible scientific hypotheses and then plan experiments to verify them?



SciMon (Wang et al. ACL 2024)



Ghafarollahi et al., (2025 Advanced Materials)

Hypothesis Generation

- Evaluation is a major challenge!
  - Models have different input and output.
  - A lack of expert evaluation standard.

Stay for our talk about this topic!

From Philosophy to NLU: Evolving Definitions of Research Hypotheses

Jian Wu[1,*], Sarah Rajtmajer[2,*]

[1] *Old Dominion University*
[2] *The Pennsylvania State University*

Sci-K 2025

Ongoing Work

- We propose generating highly novel and interdisciplinary citation-enriched hypothesis proposals through iterative interactions between expert LLMs.
- We explore new methods to automatically evaluate the novelty of the generated hypotheses.

OLD DOMINION UNIVERSITY®

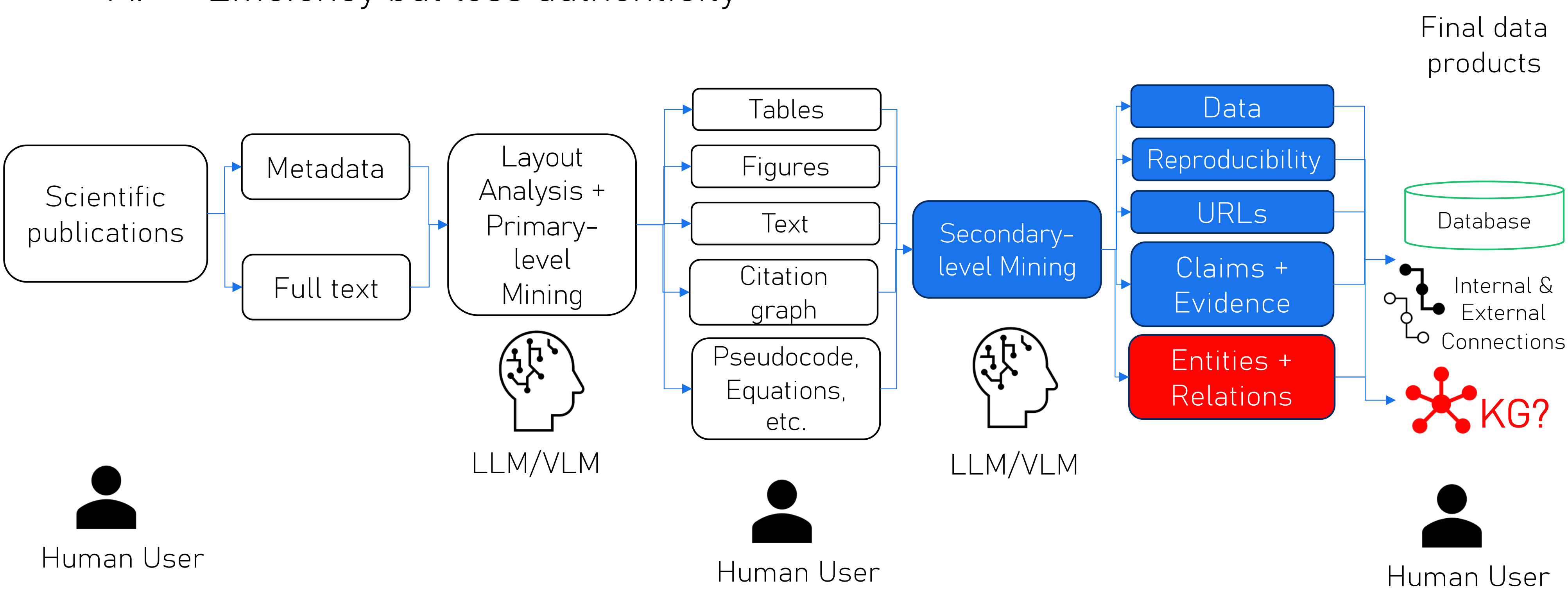
OAK RIDGE National Laboratory

PennState®

How Could Mining Scholarly Data Impact Scientific Information Access?

A Trade-off between efficiency and authenticity.

AI → Efficiency but loss authenticity



Do We still Need to Read Papers or Do We Rely on AI to Read Papers?

- Both should!
- We are still not sure how to train AI to read scientific papers and automatically mine the information the users need
- Humans should always read original papers to acquire the most genuine scientific knowledge

Summary

- AI plays an important role to read and digest the ever-growing scholarly big data to improve the efficiency of information access
- LLMs and VLMs allow us to mine nuanced knowledge from low-level content in scholarly papers but still far from being used as a human-expert level assistant
- Future research may consider two possible directions
  - Scale-out: extensively train smaller AI agents to become experts of a narrow field and then form a team to work collectively
  - Scale-up: train a huge AI to become an erudite

Many “successful” AIs are obtained by extensively training a weak AI (like a young kid) on a particular task in a narrow domain.