Reliance on science by inventors: Hybrid extraction of **in-text** patent-to-article citations

In-text citation extraction + scoring system

Data linking : PCS to MAG

- PCS = patent citations to science
- Data links PCS to MAG (Microsoft academic graph)
- MAG likes google scholar, but MAG is openly available for download by registering for an Azure account

Challenges in Linking NPLs to Scientific Articles

In text PCS extraction

https://miro.com/app/board/uXjVPI_JcBI=/?share_link_id=339918531539

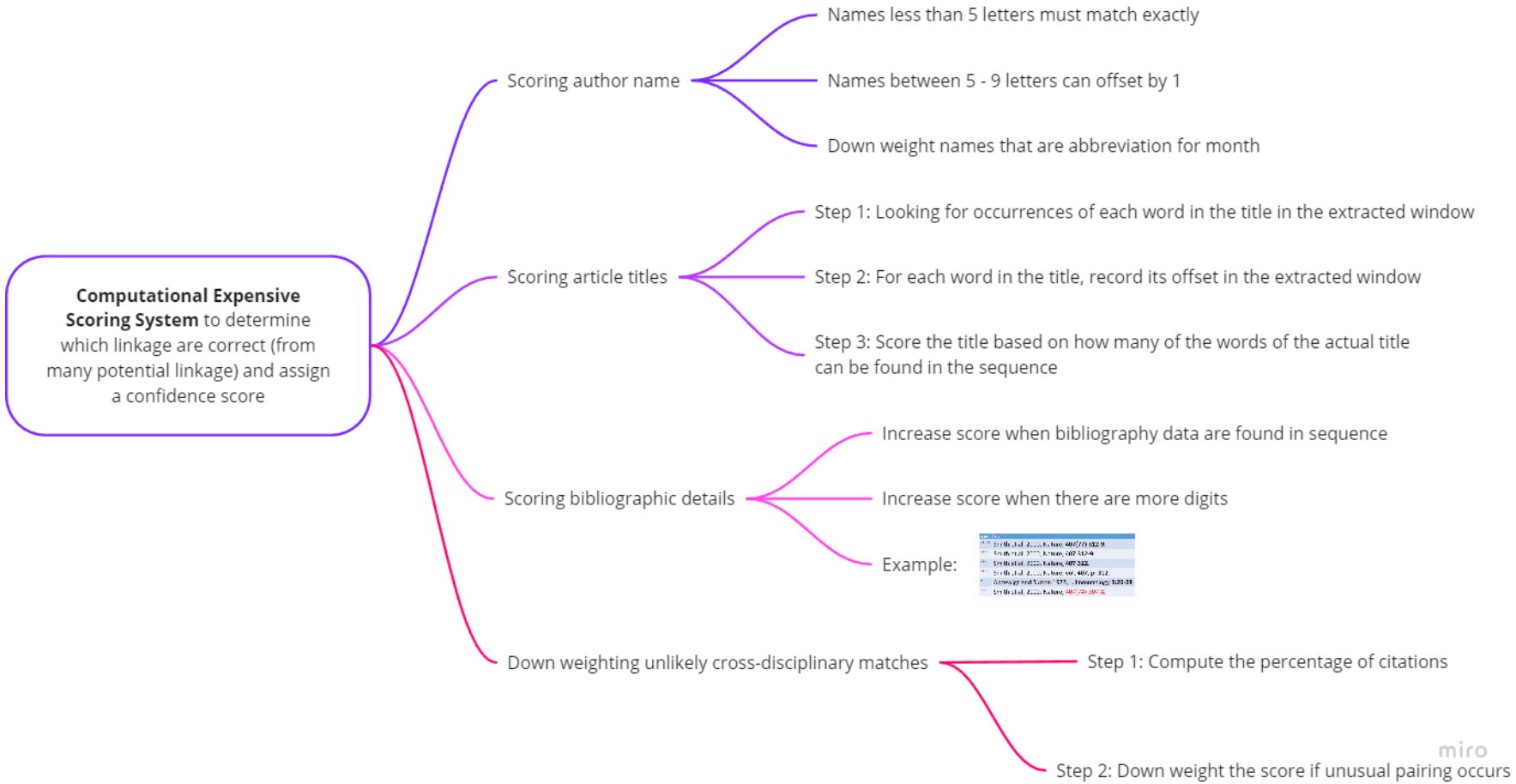
Steps in linking algorithm

segment the unstructured citation based on words (“biomedical”) , numbers, and journal names

execute computational-inexpensive matching to generate large potential PCS linkages – matching author, year, the longest word in the title of the article

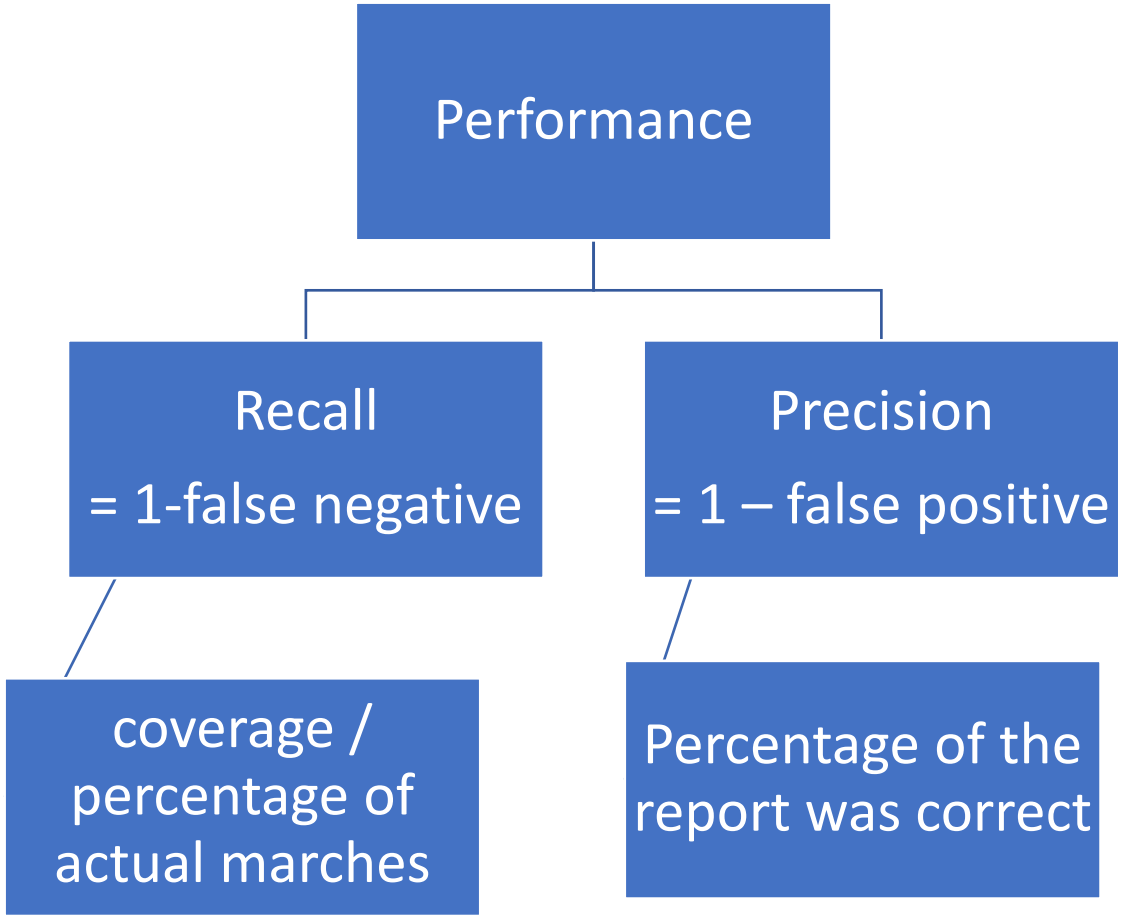
apply computational-expensive scoring to determine the likelihood of each potential PCS represents actual PCS, and assign a confidence score to the linkage

Computational Expensive Scoring System



https://miro.com/app/board/uXjVPI_JcBI=/?share_link_id=339918531539

Linkage algorithm performance



Confidence score explanation by precision

Panel A: Precision										
Conf. score >=	1	2	3	4	5	6	7	8	9	10
	93.53	95.98	97.60	99.37	99.47	99.54	99.64	99.67	99.71	100

Confidence score explanation by Recall

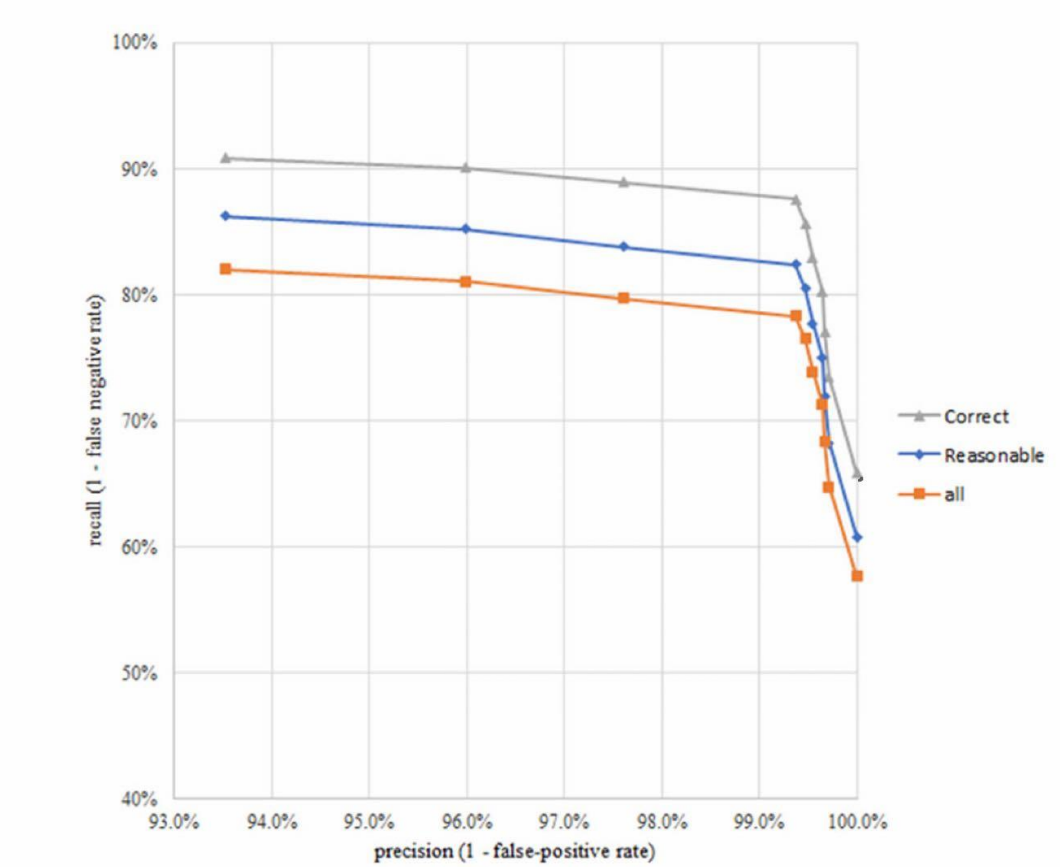
Panel B: Recall

Conf. score >=	1	2	3	4	5	6	7	8	9	10
Correct	90.86	90.11	88.95	87.57	85.69	82.91	80.20	77.09	73.44	65.89
Reasonable	86.25	85.20	83.82	82.38	80.49	77.74	74.99	71.91	68.25	60.85
All	82.05	81.04	79.73	78.31	76.49	73.87	71.26	68.33	64.79	57.70

Note: $N = 1000$, 100 matches per confidence level. Correct: citation has the correct year and first author's surname (i.e., without misspellings) and some other identifying information including one or more of journal, title, volume, issue, or page (i.e., year and author alone is not Correct). A Correct citation may not have the journal specified but cannot have the *wrong* journal. Reasonable: citation may have the year off by one and may have a letter misspelled of the first author's surname, and has some other identifying information. It may not have the journal specified, or it may have the wrong journal.

- Correct: citations have the correct year and first author’s name (without misspellings)
- Reasonable: allowed to have the year off by one and may have misspelled a letter for author’s name
- All: Correct + Reasonable

Precision-Recall Tradeoffs



Reliance on Science: Worldwide **front-page** patent citations to scientific articles

Data linkage + confidence score

Confidence score explanation by precision

TABLE 4 Precision (1—False positives) in a random sample of 100 PCS linkages per confidence level (USPTO only)

(1) Actual matches	(2)	(3) Sample of 100	(4)	(5) Precision
Confidence level	**Non-patent references linked**	**Manually marked incorrect**	**% Correct**	**Estimated cumulative % correct**
10	14,632,844	0	100%	100.00%
9	653,258	1	99%	99.96%
8	404,045	3	97%	99.88%
7	292,615	7	93%	99.76%
6	155,446	11	89%	99.65%
5	172,955	12	88%	99.53%
4	112,732	9	91%	99.47%
3	291,628	41	59%	98.76%
2	379,671	79	21%	97.04%
1	589,531	96	4%	93.93%

Abbreviation: PCS, patent citations to science.

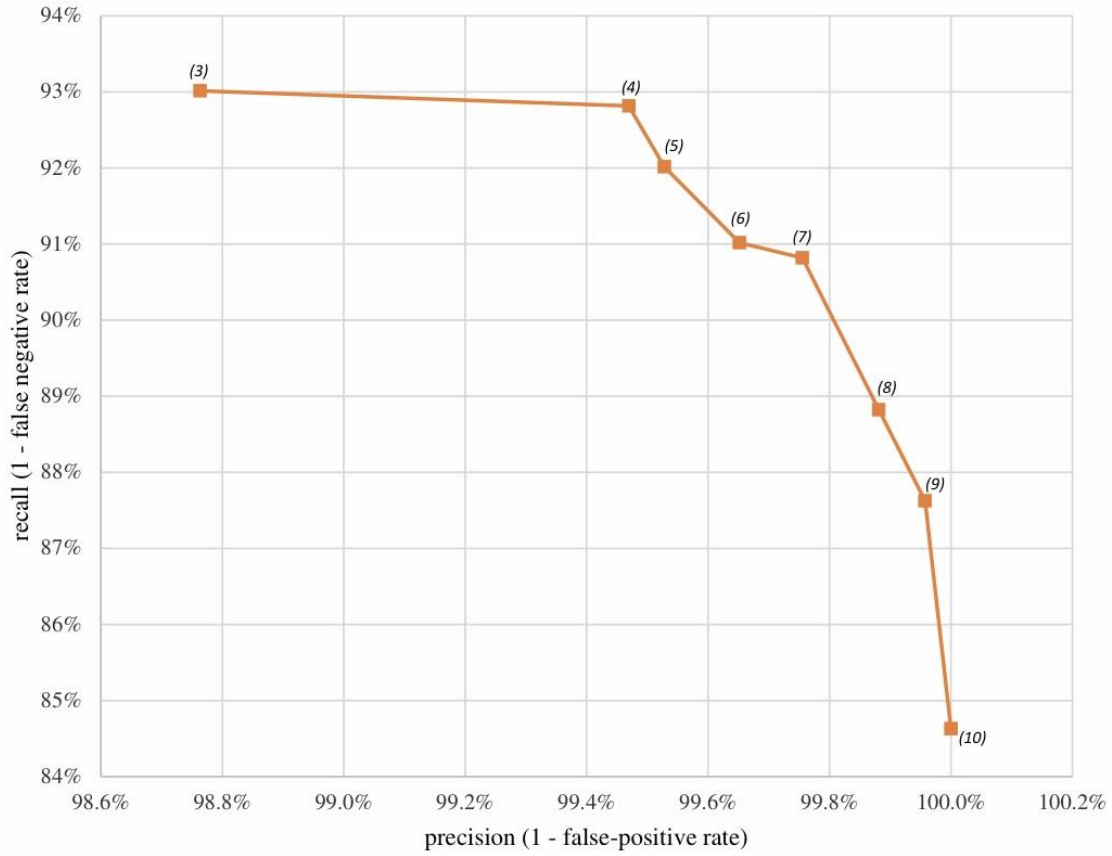
Confidence score explanation by the recall

TABLE 5 Recall (1—False negatives) as measured against 501 known-good USPTO references

Confidence level	Non-patent references linked	# Found (of 501 known)	Recall
10	14,632,844	424	84.63%
9	653,258	439	87.62%
8	404,045	445	88.82%
7	292,615	455	90.82%
6	155,446	456	91.02%
5	172,955	461	92.02%
4	112,732	465	92.81%
3	291,628	466	93.01%
2	379,671	467	93.21%
1	589,531	468	93.41%

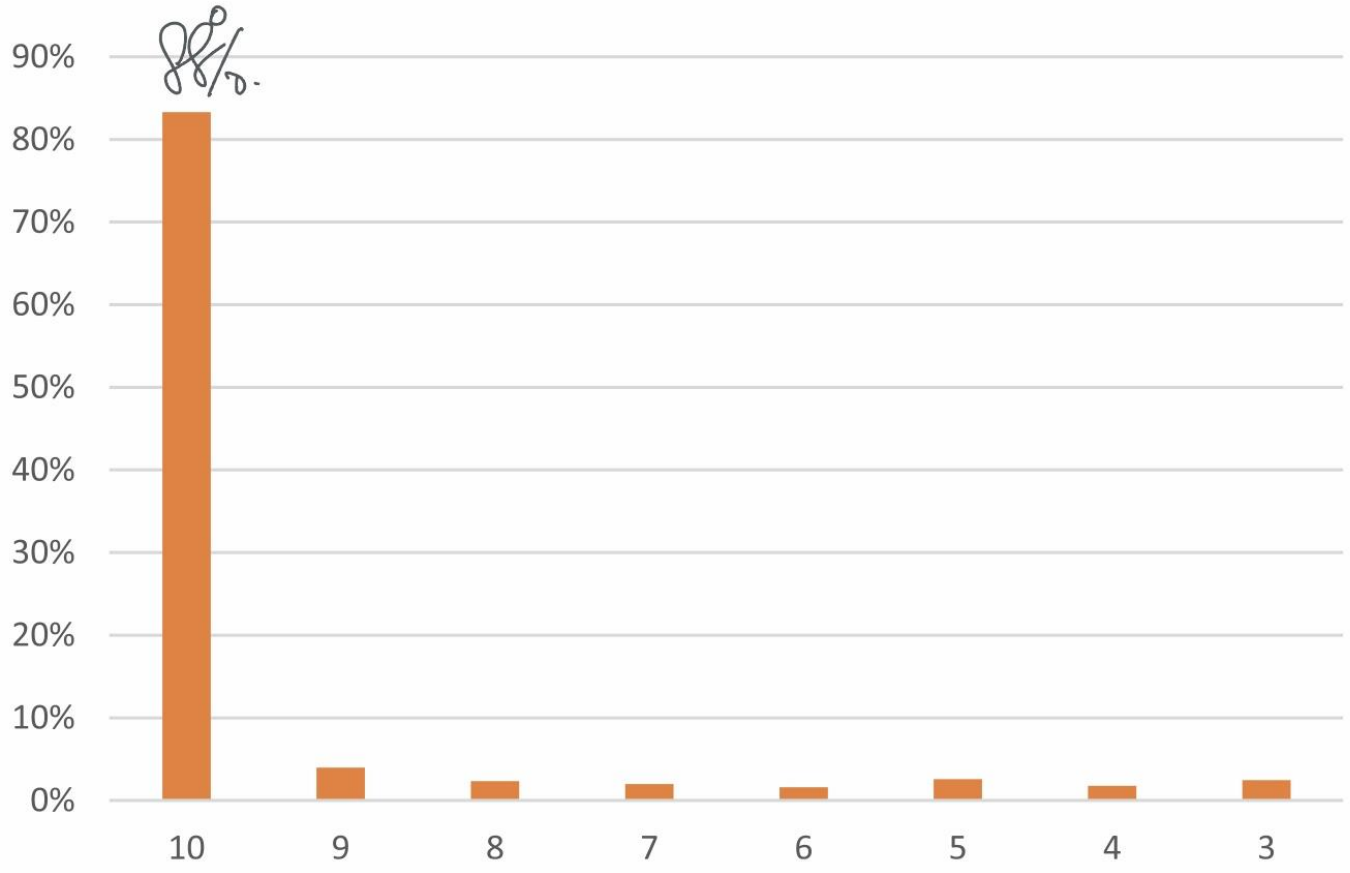
Precision-Recall Tradeoffs

FIGURE 5 USPTO linkage algorithm performance, recall versus precision [Color figure can be viewed at wileyonlinelibrary.com]



Confidence score distribution

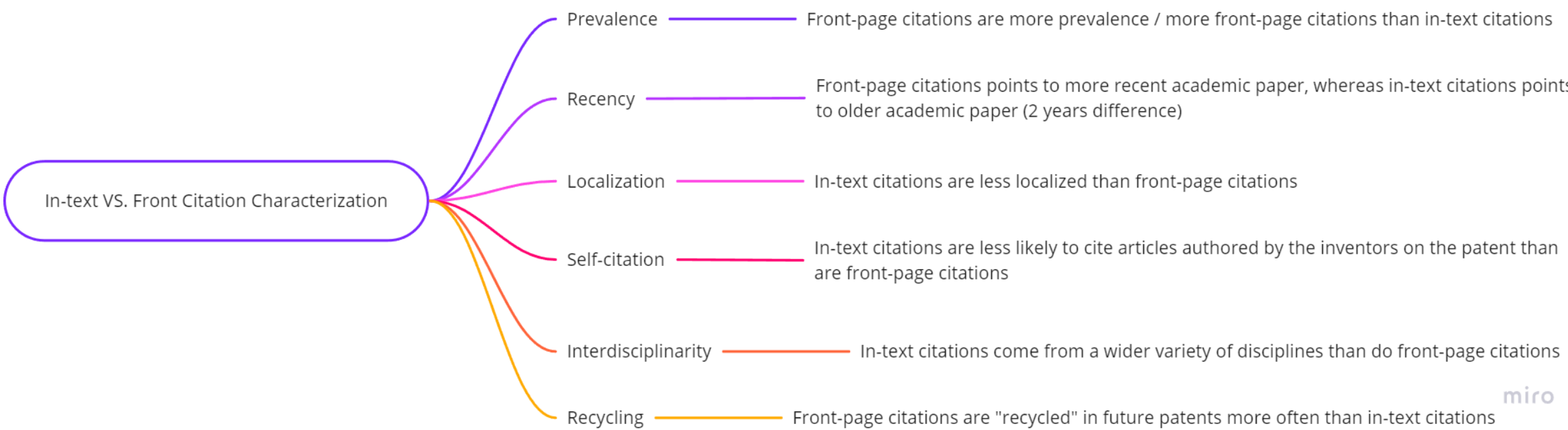
FIGURE 4 Distribution of confidence scores for linkages [Color figure can be viewed at wileyonlinelibrary.com]



In-text VS. Front Citations to Scientific Articles

https://miro.com/app/board/uXjVPI_JcBI=/?share_link_id=339918531539

First one to provide a descriptive characterization of In-text and front-page citation



Ignoring in-text citations can lead to understating the connection between academic science and commercial inventions

Number of citations to science per patent area 1

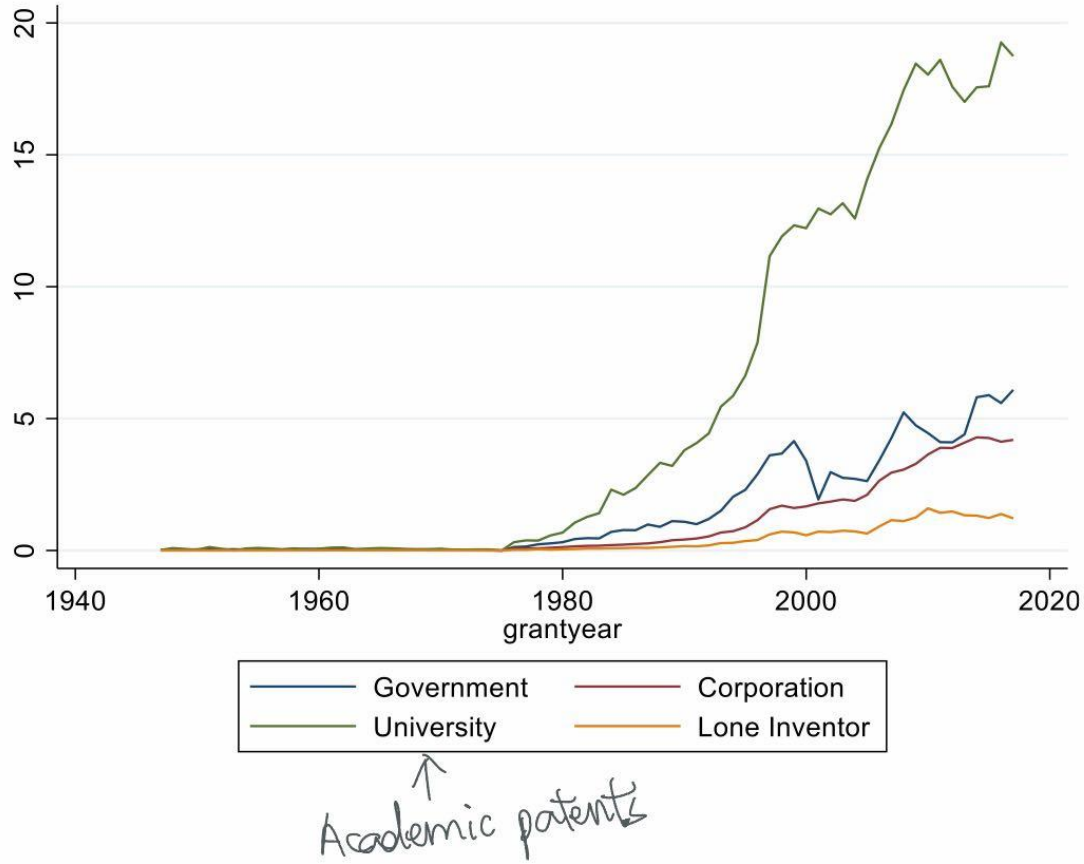


FIGURE 2 Average number of citations to science per patent, 1947–2018, by assignee type (USPTO only) [Color figure can be viewed at wileyonlinelibrary.com]

Number of citations to science per patent area 2

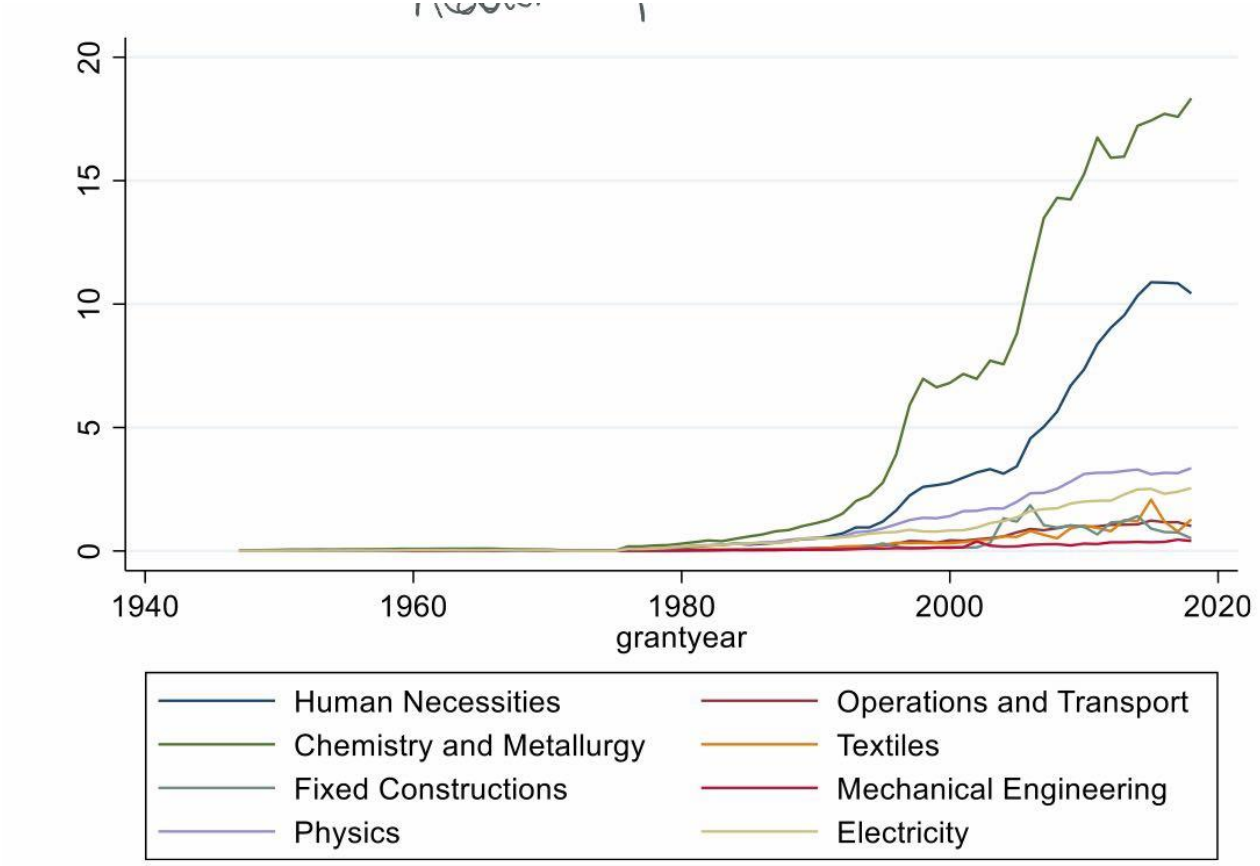


FIGURE 3 Citations to science per patent, by patent grant year and technical classification [Color figure can be viewed at wileyonlinelibrary.com]

Data

reftype	confscore	magid	patent	uspto	wherefound	doi	pmid	diff_month	selfciteconf_avg	selfciteconf_avgno0	selfciteconf_max
app	6	979	US-8386899	1	bodyonly			50	0	0	0
app	10	3066	US-10494607	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	109	0	0	0
app	10	3066	US-7311905	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	49	0	0	0
app	10	3066	US-7700090	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	95	0	0	0
app	10	3066	US-8057789	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	105	0	0	0
app	10	3066	US-8293223	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	152	0	0	0
app	10	3066	US-8455250	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	95	0	0	0
app	10	3066	US-8586360	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	138	0	0	0
app	10	3066	US-8617535	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	94	0	0	0
app	10	3066	US-9200253	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	115	0	0	0
app	10	3066	US-9255248	1	frontonly	10.1016/S0301-472X(98)00008-3	9923457	177	0	0	0
exm	8	3603	KR-101889120-B1	0	frontonly			132	0	0	0
exm	8	3603	KR-20140014155-A	0	frontonly			132	0	0	0
app	1	7400	US-7473706	1	bodyonly			50	0	0	0
app	1	7400	US-7524885	1	bodyonly			37	0	0	0
app	1	7400	US-8202901	1	bodyonly			86	0	0	0
exm	10	9715	RU-2657196-C2	0	frontonly	10.1097/TA.0B013E31824EF9EC	22673256	44	0	0	0
app	3	10101	EP-239504-A1	0	bodyonly				0	0	0
app	3	10101	EP-239504-B1	0	bodyonly				0	0	0
app	9	11779	US-6848099	1	frontonly			129	0	0	0
app	10	11959	US-5229073	1	bodyonly	10.1016/0076-6879(86)29068-0	2941672	73	0	0	0
app	10	11959	US-5783400	1	both	10.1016/0076-6879(86)29068-0	2941672	113	0	0	0
app	10	11959	US-6210906	1	both	10.1016/0076-6879(86)29068-0	2941672	156	0	0	0
app	4	12687	US-8263069	1	bodyonly		10710805	117	0	0	0
app	10	12687	US-8871199	1	bodyonly		10710805	152	0	0	0
app	9	12687	US-9486468	1	both		10710805	149	0	0	0
app	10	12687	US-9968623	1	both		10710805	161	0	0	0

Data Variable Description 1

Variable	Type	Notes
reftype	string	**App** = from applicant **Exm**=from examiner (Note: non-USPTO refs are examiner unless otherwise indicated in the reference.) **Unk** = if unspecified in the unstructured reference (Note: most pre-2006 USPTO references are unkown.)
confscore	numeric	Assigned confidence score to the match.
magid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
doi	string	Digital Object Identifier as provided by Microsoft
pmid	numeric	PubMed ID as provided by Microsoft
patent	string	Only patents for which our algorithm established a PCS linkage are included. The format is as follows. The first two characters represent the country of the patent office, e.g. US for USPTO. Next is a hyphen (-), followed by the patent number. Non-USPTO patent numbers often include another hyphen followed by an alphanumeric suffix. The DOCDB data also includes these suffixes for USPTO patents, but we remove them as many of our academic users merge against USPTO data from patentsview.org or similar USPTO-based sources, where the suffixes are not included. If you are using DOCDB-based sources, such as PATSTAT, you will want to chop off the suffixes (i.e. end of the patent number starting with the final hyphen) for USPTO patents *only*. Leading zeroes are removed from all patent numbers.
wherefound	string	**frontonly**, **bodyonly**, or **both** (i.e., both on the front page of the patent, and also in the body text)
uspto	binary	Indicates whether the patent is from the USPTO. These can be matched up by patent family using *intlpatfamily.tsv*.

Data Variable Description 2

Variable	Type	Notes
diff_month	numeric	Temporal difference between paper publication and patent application in month
selfciteconf_avg	numeric	Average of confidence score of all authors in matching with an inventor of the citing patent based on first, middle, and last name
selfciteconf_avgnoo	numeric	Average of confidence score of only authors who are matched to an inventor of the citing patent based on first, middle, and last name
selfciteconf_max	numeric	Max confidence score of an author matched to an inventor of the citing patent based on first, middle, and last name