Leveraging the Leiden dataset

Jeffrey Demaine

McMaster University

BRIC

June 15, 2022

Theme

- Third-party datasets are a rich source of metadata.
- Unexpected discoveries can be made by looking for patterns.

1. **Leiden** dataset (*Done*)
   - Today part 1**: Gender of authors** at 30 Canadian universities.
   - Today part 2**: Fractional counting** of publications at uWaterloo.

2. **Small Teams** dataset (*Currently*)

   Lingfei Wu; Dashun Wang; James Evans, 2021, "Replication Data for: Large teams develop and small teams disrupt science and technology", https://doi.org/10.7910/DVN/JPWNNK, Harvard Dataverse, V1
   https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/JPWNNK

   - Tomorrow: "**MAGic made easy**" @ ~3:40PM

3. **Standardized author citation metrics** (*Maybe someday*)

   Ioannidis JPA, Baas J, Klavans R, Boyack KW (2019) A standardized citation metrics author database annotated for scientific field. *PLoS Biol* **17**(8): e3000384. https://doi.org/10.1371/journal.pbio.3000384

Overview

1. **Leiden** dataset

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C., Tijssen, R.J., van Eck, N.J., van Leeuwen, T.N., van Raan, A.F., Visser, M.S. and Wouters, P. (2012), The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. J Am Soc Inf Sci Tec, 63: 2419-2432. https://doi.org/10.1002/asi.22708

- Today part 1**: Gender of authors** at 30 Canadian universities.

Jeffrey Demaine, **Trends in authorship by women at Canadian universities 2006 to 2019.** *Canadian Journal of Information and Library Science*, **44**(2/3), 1-11: Dec. 2021 https://doi.org/10.5206/cjilsrcsib.v44i2.13687

- Today part 2**: Fractional counting** of publications at uWaterloo.

Jeffrey Demaine, **Fractionalization of research impact reveals global trends in university collaboration**, *Scientometrics*, 10.1007/s11192-021-04246-w, (2022).

Finding patterns in the Leiden dataset

- Produced by CWTS (*Ludo Waltman, Nees Jan van Eck, Paul Wouters…*)
- **“Leiden ranking”**
- Based on Web of Science data
- **161,700** rows
- **86** columns
- **1225** universities
- **11** years in 4-year slices (*I don’t know why…*)
- 2006-2009 to 2016-2019
- The Leiden Ranking for 2022 will be **released on June 22**! (**covering 2017-2020**)

Leiden dataset (1) - *Gender of Canadian faculty*

This study:

- 30 Canadian universities
- Remember: it's ***Leiden's*** gender-classification algorithm, not mine!

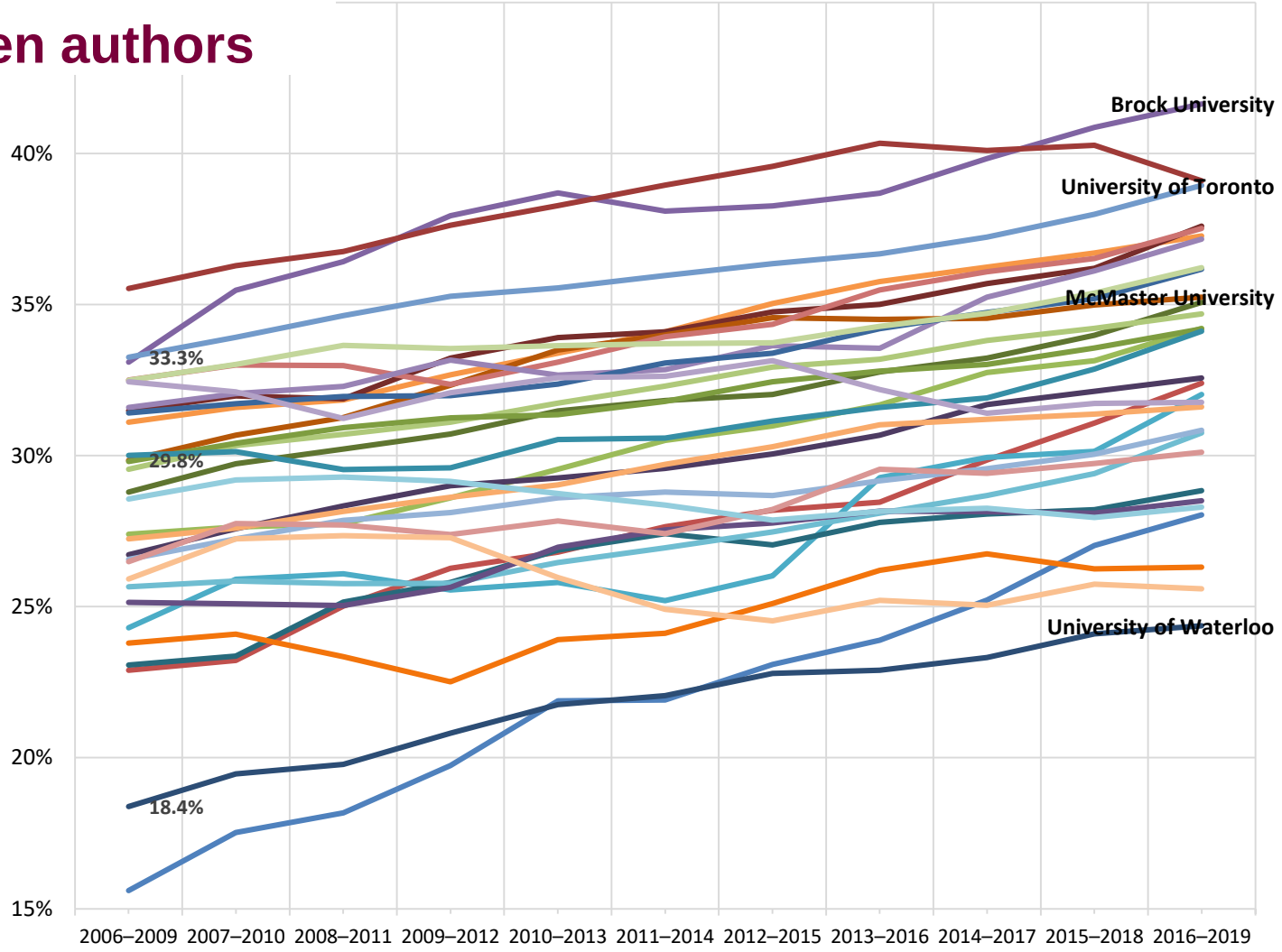
- `Gender_A`
- `Gender_A_MF`
- `A_gender_unknown`
- `A_M`
- `A_F`

University	Field	Period	Frac_counting	impact_P	gender_A	gender_A_MF	A_gender_unknown	A_M	A_F
Brock University	All sciences	2006–2009	0	1041	1502	1360	142	910	450
Brock University	All sciences	2007–2010	0	1124	1965	1790	175	1155	635
Brock University	All sciences	2008–2011	0	1214	2395	2169	226	1379	790
Brock University	All sciences	2009–2012	0	1280	2554	2325	229	1443	882
Brock University	All sciences	2010–2013	0	1342	2671	2432	239	1491	941
Brock University	All sciences	2011–2014	0	1431	2776	2510	266	1554	956
Brock University	All sciences	2012–2015	0	1538	2970	2692	278	1662	1030
Brock University	All sciences	2013–2016	0	1603	3087	2787	300	1709	1078
Brock University	All sciences	2014–2017	0	1666	3246	2930	316	1763	1167
Brock University	All sciences	2015–2018	0	1666	3334	3030	304	1792	1238
Brock University	All sciences	2016–2019	0	1696	3366	3060	306	1786	1274

Trend in women authors

Average:
2006-09: 28%
2016-19: 33%

Slope:
+ 0.5% per year



Female authors by field

Ratio as of 2016-2019:

Few women in*:*

- *Engineering*
- *Comp Sci*
- *Math*

Captain Obvious

	Rank	All sciences	Biomedical & health sciences	Life & earth sciences	Mathematics & computer science	Physical sciences & engineering	Social sciences & humanities
Brock University	1	**41.6%**	46.4%	38.1%	17.2%	24.4%	49.1%
York University	2	**39.1%**	50.9%	34.5%	23.0%	13.0%	53.6%
University of Toronto	3	**38.9%**	41.0%	37.2%	18.2%	20.5%	51.4%
University of Guelph	4	**37.6%**	45.0%	35.6%	19.3%	24.8%	48.3%
UQàM	5	**37.5%**	50.7%	30.7%	12.5%	23.8%	57.4%
Université de Montréal	6	**37.3%**	41.6%	36.0%	13.3%	24.9%	46.7%
Dalhousie University	7	**37.2%**	40.1%	35.2%	15.4%	23.2%	55.1%
McGill University	8	**36.2%**	40.1%	34.2%	15.7%	18.5%	49.3%
University of Ottawa	9	**36.2%**	39.4%	30.1%	18.6%	17.8%	50.2%
McMaster University	10	**35.2%**	38.4%	31.9%	15.6%	20.8%	49.2%
Queen's University	11	**35.1%**	42.2%	34.3%	11.5%	17.5%	49.9%
Univ of British Columbia	12	**34.7%**	39.3%	33.4%	14.9%	16.8%	46.5%
University of Calgary	13	**34.2%**	38.2%	31.1%	15.7%	15.9%	51.2%
Université Laval	14	**34.2%**	40.0%	30.7%	17.4%	18.0%	41.9%
University of Manitoba	15	**34.1%**	37.8%	29.2%	15.5%	17.5%	52.6%
Univ of Saskatchewan	16	**32.6%**	40.3%	30.5%	15.1%	16.8%	44.4%
Ryerson University	17	**32.4%**	55.0%	33.5%	11.0%	12.7%	53.5%
University of Regina	18	**32.0%**	48.1%	29.3%	13.0%	19.8%	48.6%
Memorial Univ of Nfld	19	**31.8%**	40.4%	31.2%	16.3%	18.4%	46.5%
Western University	20	**31.6%**	34.1%	29.7%	20.1%	17.8%	46.1%
University of Alberta	21	**30.8%**	38.4%	26.4%	13.7%	15.4%	47.6%
Simon Fraser Univ	22	**30.8%**	39.5%	33.3%	19.9%	12.9%	43.2%
Université de Sherbrooke	23	**30.1%**	39.8%	30.5%	11.6%	8.6%	57.9%
Concordia University	24	**28.8%**	43.9%	28.8%	15.8%	17.2%	46.3%
Carleton University	25	**28.5%**	36.0%	30.9%	6.7%	18.2%	48.0%
University of Victoria	26	**28.3%**	41.4%	31.3%	21.3%	13.8%	45.7%
INRS	27	**28.0%**	45.3%	31.7%	11.8%	22.5%	37.4%
Univ of New Brunswick	28	**26.3%**	44.1%	23.1%	16.9%	17.6%	50.8%
University of Windsor	29	**25.6%**	39.2%	22.0%	19.3%	16.2%	42.4%
University of Waterloo	30	**24.4%**	42.3%	27.6%	11.7%	14.2%	40.1%
	Mean	*33.0%*	*42.0%*	*31.4%*	*15.6%*	*18.0%*	*48.4%*

There is a secret door in the data…

The **error rate** of gender-classification algorithm tells us something about the names.

This allows us to measure the **ethnic diversity** across these universities & fields.

There is a secret door in the data…

The **error rate** of gender-classification algorithm tells us something about the names.

This allows us to measure the **ethnic diversity** across these universities & fields.

Error rate by field

3 universities are not like the others

What do they have in common?

Now the error isn't a bug –
it's a ***feature***!

Column: `PA_gender_unknown`
Snapshot of final period: 2016-2019

University	All sciences	Biomedical & health sciences	Life & earth sciences	Mathematics & computer science	Physical sciences & engineering	Social sciences & humanities
University of Victoria	57.4%	7.1%	11.8%	27.2%	78.8%	8.4%
Carleton University	49.4%	9.7%	8.2%	15.0%	86.2%	7.4%
Simon Fraser University	48.8%	12.9%	6.7%	31.3%	76.6%	7.5%
University of Regina	33.6%	6.5%	42.6%	18.9%	49.4%	10.5%
York University	24.2%	11.9%	14.1%	26.4%	47.9%	10.6%
University of Alberta	22.0%	12.7%	16.4%	27.6%	42.8%	9.5%
Univ of British Columbia	21.8%	10.5%	11.1%	18.9%	54.9%	9.4%
University of New Brunswick	18.5%	7.6%	11.0%	28.6%	28.8%	9.3%
University of Waterloo	18.5%	9.0%	11.8%	23.7%	24.6%	10.6%
Ryerson University	18.4%	10.0%	16.0%	24.7%	24.5%	12.9%
McGill University	17.9%	8.5%	12.8%	18.9%	48.0%	7.8%
Univ of Saskatchewan	17.8%	14.0%	17.0%	31.7%	26.2%	13.2%
University of Toronto	17.1%	10.8%	12.7%	16.2%	49.8%	8.4%
Concordia University	16.3%	7.4%	12.3%	23.3%	21.8%	8.3%
University of Manitoba	15.4%	12.5%	15.4%	28.1%	26.8%	10.7%
University of Windsor	15.1%	10.3%	9.4%	13.7%	21.9%	17.3%
Memorial Univ of Nfld	15.0%	11.5%	12.8%	28.5%	19.5%	6.5%
Université de Montréal	14.0%	5.1%	5.0%	9.3%	44.3%	3.2%
Queen's University	13.9%	9.3%	8.6%	16.1%	27.4%	8.1%
INRS	13.5%	4.6%	8.7%	8.4%	19.2%	3.5%
University of Calgary	13.2%	11.1%	14.2%	20.1%	21.4%	8.3%
Western University	12.6%	9.3%	11.3%	19.7%	24.4%	7.3%
McMaster University	12.5%	10.4%	11.5%	18.5%	21.6%	9.4%
University of Guelph	12.4%	9.2%	10.3%	10.4%	29.5%	8.7%
Dalhousie University	11.5%	8.3%	11.4%	15.9%	23.9%	6.9%
University of Ottawa	10.0%	8.9%	10.1%	16.8%	15.6%	7.3%
Brock University	9.1%	6.4%	10.2%	30.0%	15.9%	4.8%
Univ de Sherbrooke	6.7%	2.9%	3.3%	8.3%	14.8%	4.3%
UQàM	5.3%	4.3%	4.7%	7.5%	7.5%	4.5%
Université Laval	4.9%	3.5%	4.8%	8.4%	9.6%	3.6%

Error rate by field

3 universities are not like the others

What do they have in common?

Now the error isn't a bug –
it's a ***feature***!

Column: `PA_gender_unknown`
Snapshot of final period: 2016-2019

University	All sciences	Biomedical & health sciences	Life & earth sciences	Mathematics & computer science	Physical sciences & engineering	Social sciences & humanities
University of Victoria	57.4%	7.1%	11.8%	27.2%	78.8%	8.4%
Carleton University	49.4%	9.7%	8.2%	15.0%	86.2%	7.4%
Simon Fraser University	48.8%	12.9%	6.7%	31.3%	76.6%	7.5%
University of Regina	33.6%	6.5%	42.6%	18.9%	49.4%	10.5%
York University	24.2%	11.9%	14.1%	26.4%	47.9%	10.6%
University of Alberta	22.0%	12.7%	16.4%	27.6%	42.8%	9.5%
Univ of British Columbia	21.8%	10.5%	11.1%	18.9%	54.9%	9.4%
University of New Brunswick	18.5%	7.6%	11.0%	28.6%	28.8%	9.3%
University of Waterloo	18.5%	9.0%	11.8%	23.7%	24.6%	10.6%
Ryerson University	18.4%	10.0%	16.0%	24.7%	24.5%	12.9%
McGill University	17.9%	8.5%	12.8%	18.9%	48.0%	7.8%
Univ of Saskatchewan	17.8%	14.0%	17.0%	31.7%	26.2%	13.2%
University of Toronto	17.1%	10.8%	12.7%	16.2%	49.8%	8.4%
Concordia University	16.3%	7.4%	12.3%	23.3%	21.8%	8.3%
University of Manitoba	15.4%	12.5%	15.4%	28.1%	26.8%	10.7%
University of Windsor	15.1%	10.3%	9.4%	13.7%	21.9%	17.3%
Memorial Univ of Nfld	15.0%	11.5%	12.8%	28.5%	19.5%	6.5%
Université de Montréal	14.0%	5.1%	5.0%	9.3%	44.3%	3.2%
Queen's University	13.9%	9.3%	8.6%	16.1%	27.4%	8.1%
INRS	13.5%	4.6%	8.7%	8.4%	19.2%	3.5%
University of Calgary	13.2%	11.1%	14.2%	20.1%	21.4%	8.3%
Western University	12.6%	9.3%	11.3%	19.7%	24.4%	7.3%
McMaster University	12.5%	10.4%	11.5%	18.5%	21.6%	9.4%
University of Guelph	12.4%	9.2%	10.3%	10.4%	29.5%	8.7%
Dalhousie University	11.5%	8.3%	11.4%	15.9%	23.9%	6.9%
University of Ottawa	10.0%	8.9%	10.1%	16.8%	15.6%	7.3%
Brock University	9.1%	6.4%	10.2%	30.0%	15.9%	4.8%
Univ de Sherbrooke	6.7%	2.9%	3.3%	8.3%	14.8%	4.3%
UQàM	5.3%	4.3%	4.7%	7.5%	7.5%	4.5%
Université Laval	4.9%	3.5%	4.8%	8.4%	9.6%	3.6%

Leiden dataset (2) - *Fractionalization of Impact*

`Frac counting: 0` = Whole counted

`Impact_P` = total # publications

`P_top1` = # of pubs in Top 1% most cited.

`P_top50to90` = # of pubs between the average (i.e. 50%) and Top 90% most cited.

`P_bottomHalf` = # of pubs from 0% to 50% most cited.

There are **three trends** here:

1. Time
2. Level of impact
3. Divergence between levels (*i.e. Trend 1 x Trend 2*)

Field	Period	Frac counting	impact_P	P_top1	P_top90to99	P_top50to90	P_bottomHalf
All sciences	2006–2009	0	6954	87	708	2988	3171
All sciences	2007–2010	0	7405	90	752	3200	3363
All sciences	2008–2011	0	7925	96	831	3412	3586
All sciences	2009–2012	0	8355	97	908	3601	3749
All sciences	2010–2013	0	8916	98	980	3796	4042
All sciences	2011–2014	0	9512	106	1048	4111	4247
All sciences	2012–2015	0	10121	127	1131	4377	4486
All sciences	2013–2016	0	10618	136	1169	4637	4676
All sciences	2014–2017	0	11078	183	1280	4751	4864
All sciences	2015–2018	0	11452	198	1357	4834	5063
All sciences	2016–2019	0	12156	240	1429	5171	5316
All sciences	2006–2009	1	4149	49	383	1764	1953
All sciences	2007–2010	1	4395	45	414	1882	2054
All sciences	2008–2011	1	4719	44	467	2013	2195
All sciences	2009–2012	1	4947	42	494	2123	2288
All sciences	2010–2013	1	5241	40	545	2201	2455
All sciences	2011–2014	1	5485	44	567	2336	2538
All sciences	2012–2015	1	5733	57	589	2451	2636
All sciences	2013–2016	1	5871	54	587	2513	2717
All sciences	2014–2017	1	5958	81	617	2496	2764
All sciences	2015–2018	1	6025	82	644	2492	2807
All sciences	2016–2019	1	6134	86	624	2568	2856

Disappearing impact

Over 10 yrs, output increased by

- 4426 **whole** papers (up **64%**)
- 1837 **fractional** papers (up **44%**)

What happened to the other 20%?

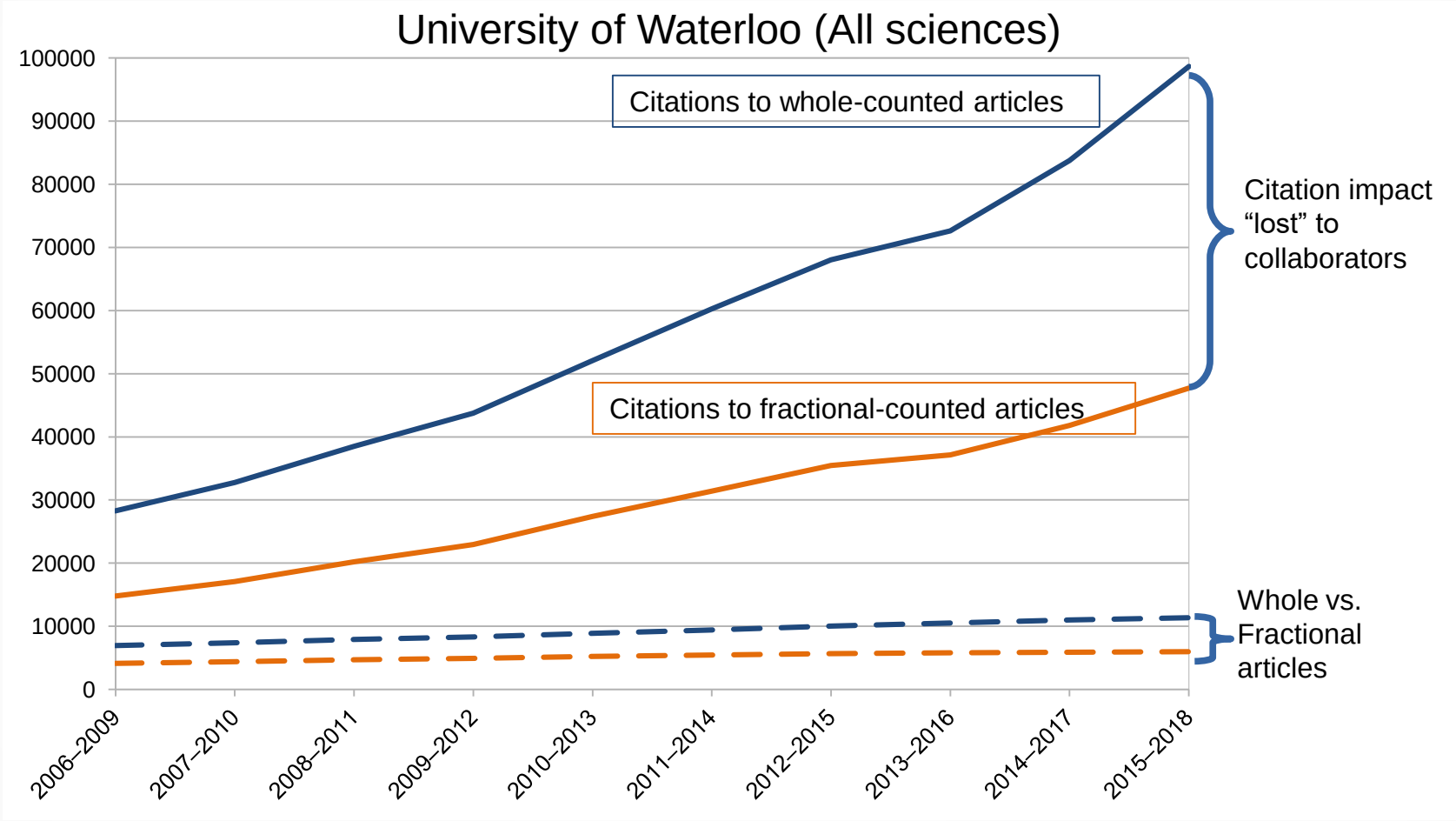
- Collaborators got it

Yes, authors *wrote* 64% more articles
But…
productivity (*of the university*) increased by 44%

University	Period	Fractional Counting	Publications	Top 1% most cited
University of Waterloo	2006–2009	0	6923	76
University of Waterloo	2007–2010	0	7378	82
University of Waterloo	2008–2011	0	7894	84
University of Waterloo	2009–2012	0	8321	93
University of Waterloo	2010–2013	0	8864	
University of Waterloo	2011–2014	0	9434	
University of Waterloo	2012–2015	0	10030	131
University of Waterloo	2013–2016	0	10514	138
University of Waterloo	2014–2017	0	10981	187
University of Waterloo	2015–2018	0	11349	197
University of Waterloo	2006–2009	1	4131	41
University of Waterloo	2007–2010	1	4380	41
University of Waterloo	2008–2011	1	4699	38
University of Waterloo	2009–2012	1	4924	43
University of Waterloo	2010–2013	1	5210	
University of Waterloo	2011–2014	1	5437	
University of Waterloo	2012–2015	1	5679	57
University of Waterloo	2013–2016	1	5810	53
University of Waterloo	2014–2017	1	5898	80
University of Waterloo	2015–2018	1	5968	82

+64%

+44 %


University of Waterloo (All sciences)
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Citations to whole-counted articles
Citations to fractional-counted articles
Citation impact "lost" to collaborators
Whole vs. Fractional articles
2006–2009
2007–2010
2008–2011
2009–2012
2010–2013
2011–2014
2012–2015
2013–2016
2014–2017
2015–2018

The *Fractionalization* of impact

- As collaboration *increases*, fractional impact ***decreases***

- ***"Fractionalization ratio":***
  - In 2006-2009:
    4131 ÷ 6923 = 0.5967
  - By 2015-2018:
    5968 ÷ 11349 = 0.5258
  - A ***decrease*** of 0.071 (**~12%**)

- *Collaboration 'tax' on impact*

University	Period	Fractional Counting	Publications	Top 1% most cited
University of Waterloo	2006–2009	0	6923	76
University of Waterloo	2007–2010	0	7378	82
University of Waterloo	2008–2011	0	7894	84
University of Waterloo	2009–2012	0	8321	93
University of Waterloo	2010–2013	0	8864	101
University of Waterloo	2011–2014	0	9434	116
University of Waterloo	2012–2015	0	10030	131
University of Waterloo	2013–2016	0	10514	138
University of Waterloo	2014–2017	0	10981	187
University of Waterloo	2015–2018	0	11349	197
University of Waterloo	2006–2009	1	4131	41
University of Waterloo	2007–2010	1	4380	41
University of Waterloo	2008–2011	1	4699	38
University of Waterloo	2009–2012	1	4924	43
University of Waterloo	2010–2013	1	5210	40
University of Waterloo	2011–2014	1	5437	46
University of Waterloo	2012–2015	1	5679	57
University of Waterloo	2013–2016	1	5810	53
University of Waterloo	2014–2017	1	5898	80
University of Waterloo	2015–2018	1	5968	82

The *Frax Tax* paid by Waterloo

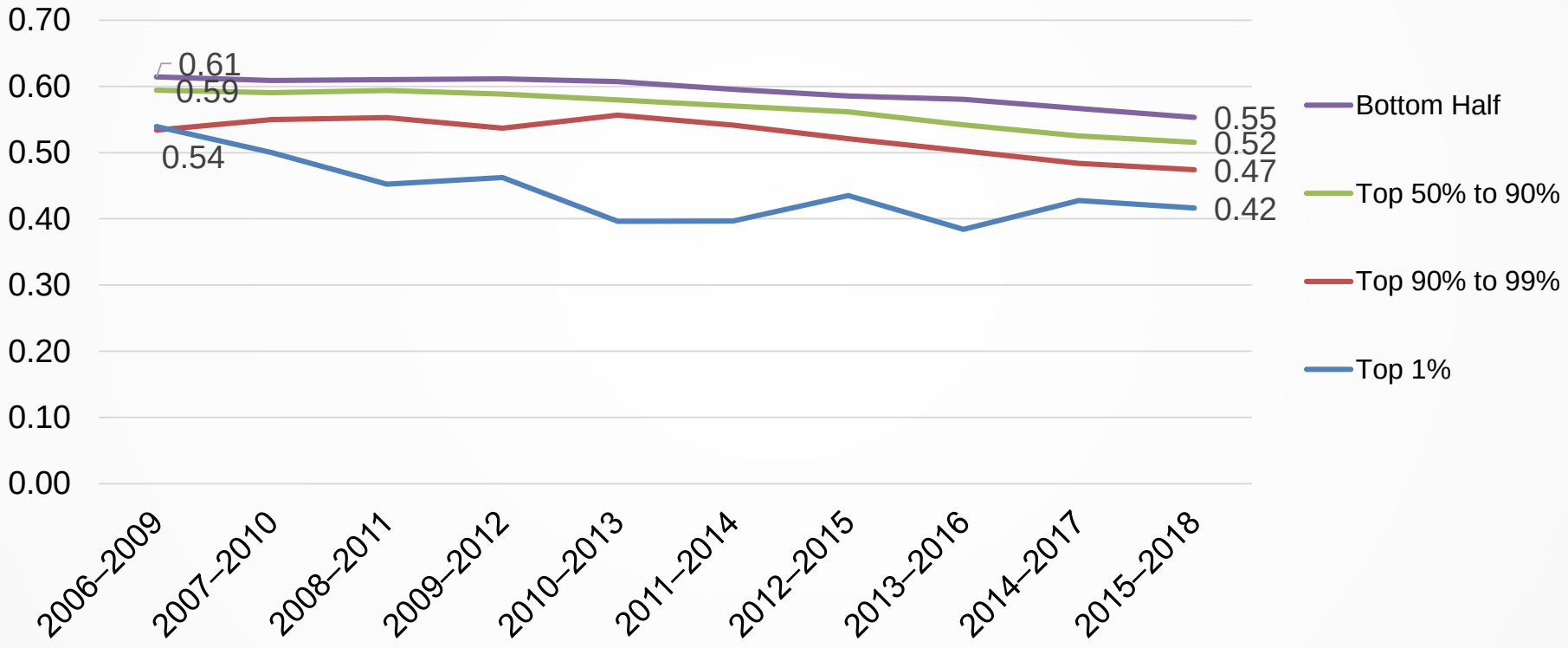
- As collaboration *increases*, fractional impact ***decreases***
- Collaboration 'tax':
  - In 2006-2009:
    4131 ÷ 6923 = 0.5967
  - By 2015-2018:
    5968 ÷ 11349 = 0.5258
  - A ***decrease*** of 0.071 (**~12%**)
- Top 1% has sharper decline
  - Frax = 0.5395 → 0.4162
  - A ***decrease*** of 0.123 (**~23%**)

University	Period	Fractional Counting	Publications	Top 1% most cited
University of Waterloo	2006–2009	0	6923	76
University of Waterloo	2007–2010	0	7378	82
University of Waterloo	2008–2011	0	7894	84
University of Waterloo	2009–2012	0	8321	93
University of Waterloo	2010–2013	0	8864	101
University of Waterloo	2011–2014	0	9434	116
University of Waterloo	2012–2015	0	10030	131
University of Waterloo	2013–2016	0	10514	138
University of Waterloo	2014–2017	0	10981	187
University of Waterloo	2015–2018	0	11349	197
University of Waterloo	2006–2009	1	4131	41
University of Waterloo	2007–2010	1	4380	41
University of Waterloo	2008–2011	1	4699	38
University of Waterloo	2009–2012	1	4924	43
University of Waterloo	2010–2013	1	5210	40
University of Waterloo	2011–2014	1	5437	46
University of Waterloo	2012–2015	1	5679	57
University of Waterloo	2013–2016	1	5810	53
University of Waterloo	2014–2017	1	5898	80
University of Waterloo	2015–2018	1	5968	82

UW's best research is returning ever less impact than its more average publications

Fractionalization ratio by impact percentile

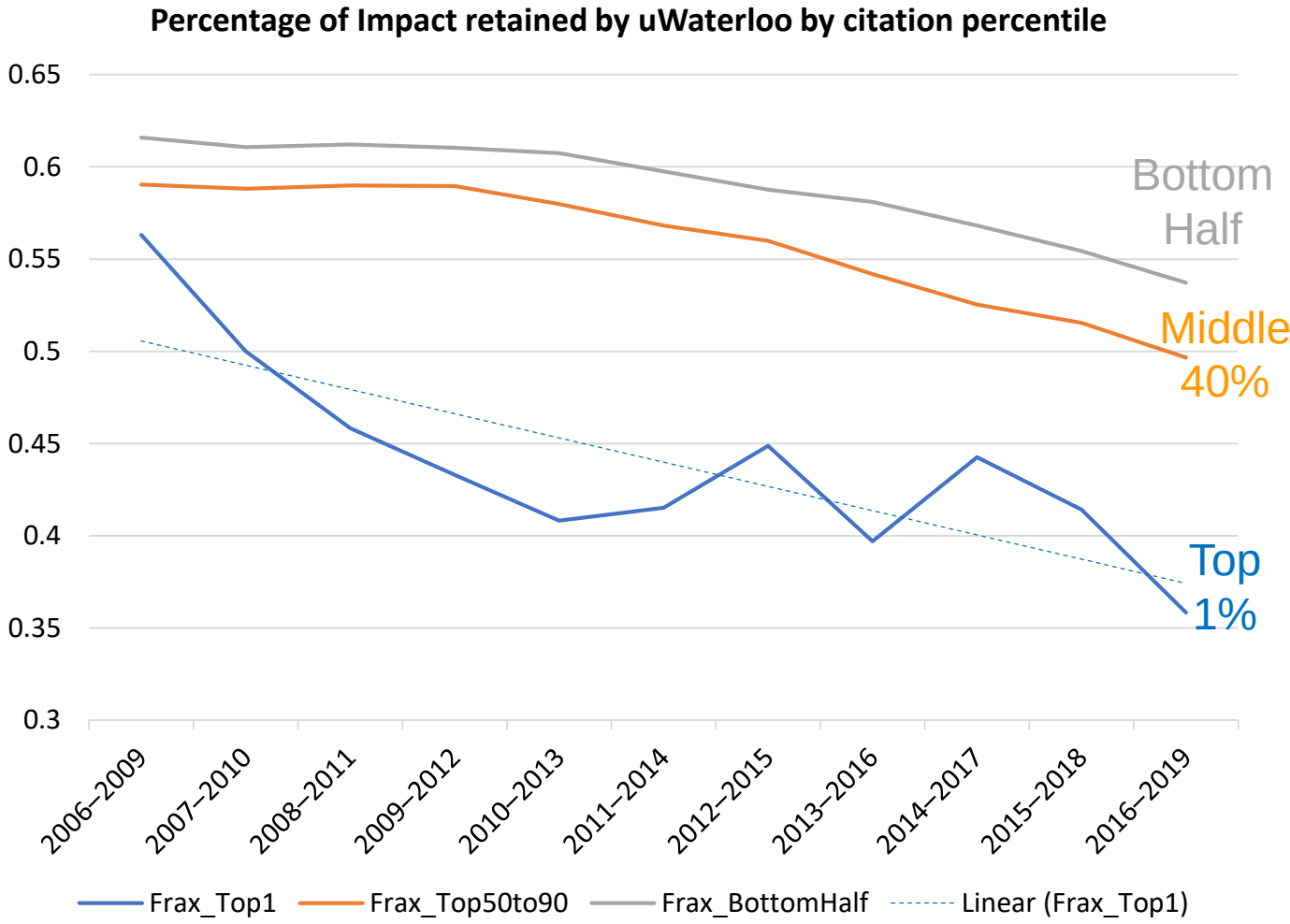
University of Waterloo – All sciences



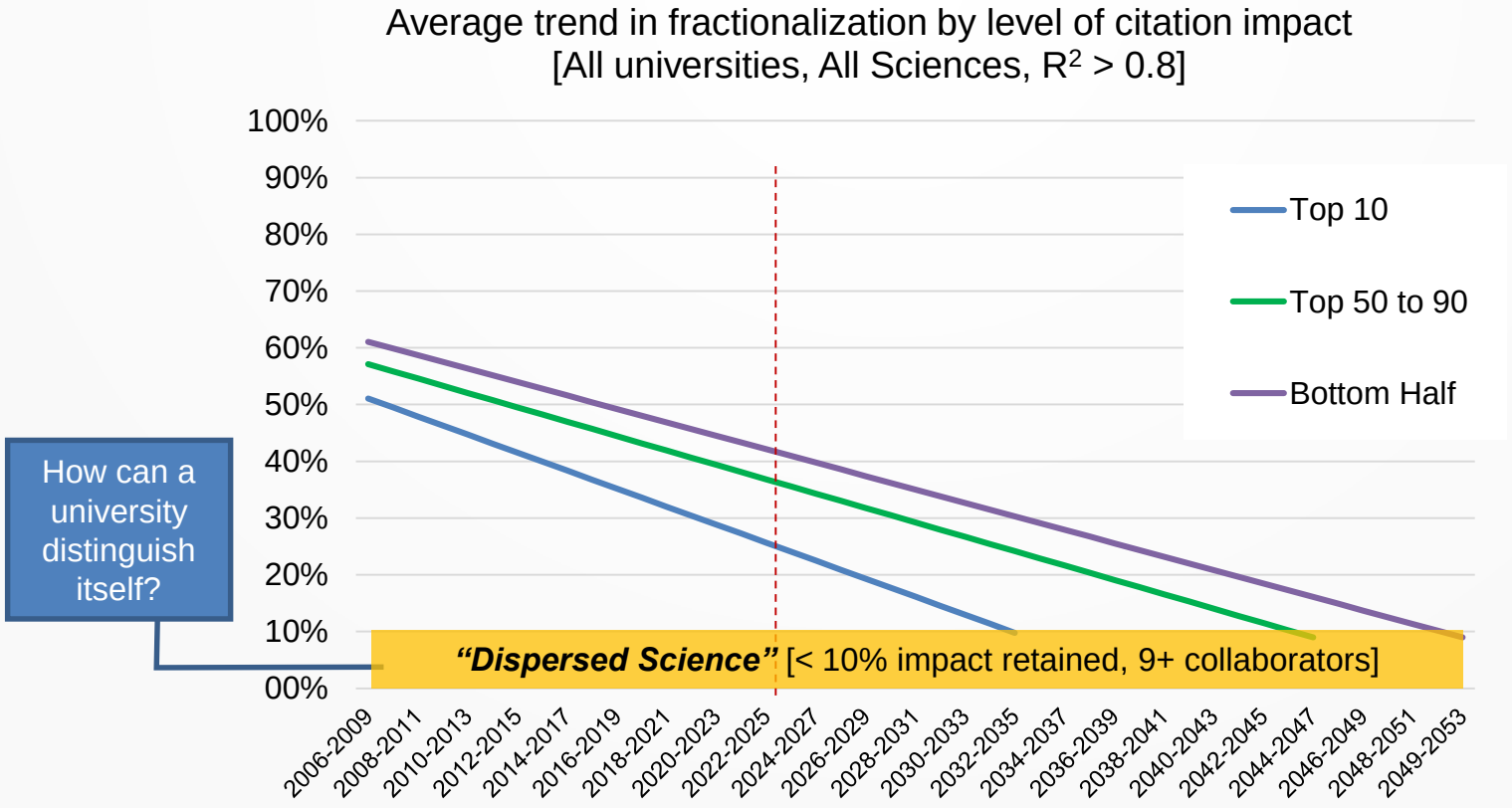
Trend in Impact

Average change:

- - 0.75% per year
- - 0.97% per year
- - 1.31% per year



Dispersed Science



Trend in the Impact of uWaterloo's research – *per year*

Field	Bottom Half	Middle 40%	Top 1%
Biomedical and health sciences	-0.88%	-1.27%	-1.93%
Life and earth sciences	-0.88%	-1.06%	-1.22%
Mathematics and computer science	**-0.88%**	**-1.13%**	**-2.34%**
Physical sciences and engineering	-0.57%	-0.66%	-0.17%
Social sciences and humanities	-0.52%	-0.80%	-0.46%
All sciences	***-0.75%***	***-0.97%***	***-1.31%***

- By 2029, the best research in Math & CS will **lose another 23%** of its impact!
- In contrast, the "worst" M&CS research will only lose **9%** of its impact.
- So which research returns the most net citations to Waterloo?
- Does more collaboration really lead to more impact if citations must be shared?
- As this trend is seen across all universities, the **distinction** between them **disappears** (*We're all working on the same stuff…together.*)
- How can uWaterloo remain a leader? *Is leadership becoming centrality?*

Collaboration / Fractionalization / Team size are all facets of the same issue:

Institute for Scientific Information:

- Potter, Ross W.K., Martin Szomszor, and Jonathan Adams. 2022. “Comparing Standard, Collaboration and Fractional CNCI at the Institutional Level: Consequences for Performance Evaluation.” *Scientometrics*, February, 1–14. https://doi.org/10.1007/S11192-022-04303-Y/FIGURES/3.
- Potter, Ross W.K., Martin Szomszor, and Jonathan Adams. 2020. “Interpreting CNCIs on a Country-Scale: The Effect of Domestic and International Collaboration Type.” *Journal of Informetrics* 14 (4). https://doi.org/10.1016/j.joi.2020.101075.
- Adams, Jonathan, David Pendelbury, and Ross Potter. 2022. “Making It Count: Research Credit Management in a Collaborative World.” **[NEW METRIC = “*Collaborative CNCI*”]**

- Thelwall, Mike. 2020. “Large Publishing Consortia Produce Higher Citation Impact Research but Coauthor Contributions Are Hard to Evaluate.” *Quantitative Science Studies* 1 (1): 290–302. https://doi.org/10.1162/qss_a_00003.

- Chu, Johan S.G., and **James A. Evans**. 2021. “Slowed Canonical Progress in Large Fields of Science.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (41): 2021636118. https://doi.org/10.1073/PNAS.2021636118/-/DCSUPPLEMENTAL. [***“Disruptiveness”***]

Jeff Demaine
Bibliomagician
demainj@mcmaster.ca