MAGic made easy

Jeffrey Demaine

McMaster University

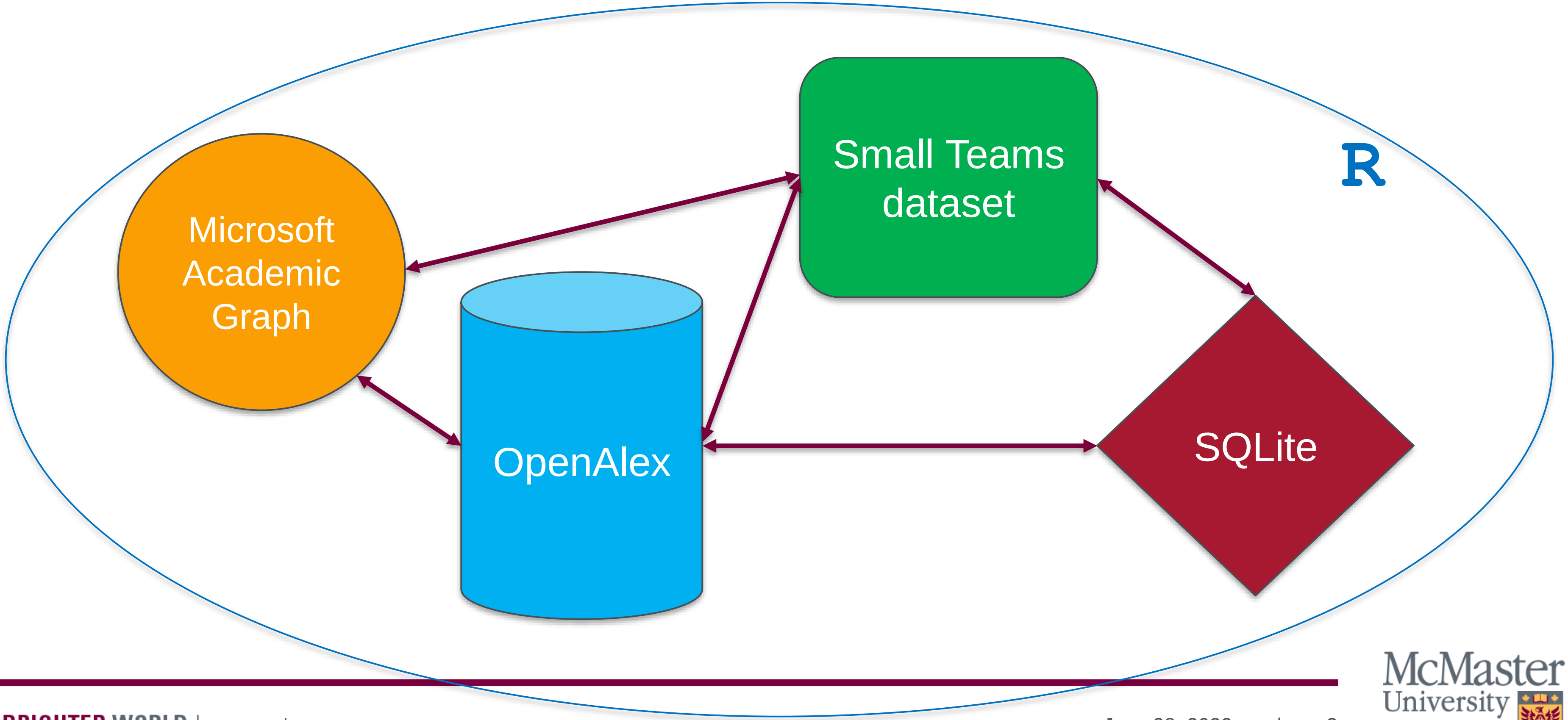
demainj@mcmaster.ca

BRIC

June 16, 2022

Challenge: connect the dots into a linear storyline

Leveraging 3rd-party datasets (new metadata, new patterns)



Panoramix

- Collects herbs
- Combines them into a ***Potion magique***

Similarly, today’s goal:

- How to collect a dataset via an API
- Combine with another datasets using SQL.
- All in a single R script (a “recipe”).

© ALLNUMIS.COM

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE ECOLOGY OF RESEARCH
Small teams disrupt and big teams consolidate to drive progress in science and technology PAGES 330 & 378
EPIDEMIOLOGY
DISEASE CONTROL
Meet Nigeria's prime protector of public health
PAGE 310
PLANETARY SCIENCE
NEPTUNE'S NEW MOON
Discovery of Hippocamp hints at history of impacts
PAGES 320 & 350
NEUROSCIENCE
LEARNING NOT TO FEAR
Neural pathway sheds light on therapy for stress disorder
PAGES 335 & 339
NATURE.COM
21 February 2019 £10
Vol. 566, No. 7744

Large teams develop and small teams disrupt science and technology

We analyzed teamwork from more than 65 million papers, patents and software products over 100 years.

Nature Article

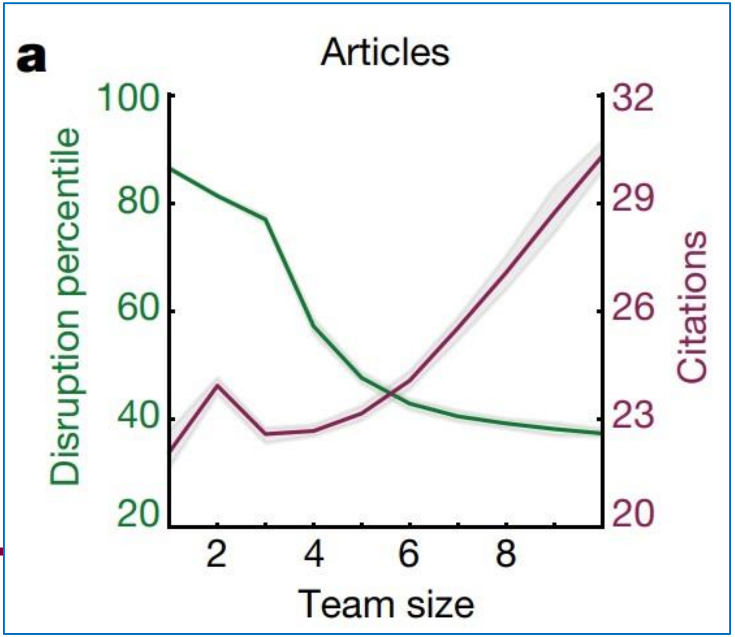
Coverage of the “Small Teams” research in the news

Can Big Science Be Too Big? - *New York Times*
https://www.nytimes.com/2019/02/13/science/science-research-psychology.html

Small Teams of Scientists Have Fresher Ideas - *The Atlantic*
https://www.theatlantic.com/science/archive/2019/02/why-small-science-still-matters/582685/

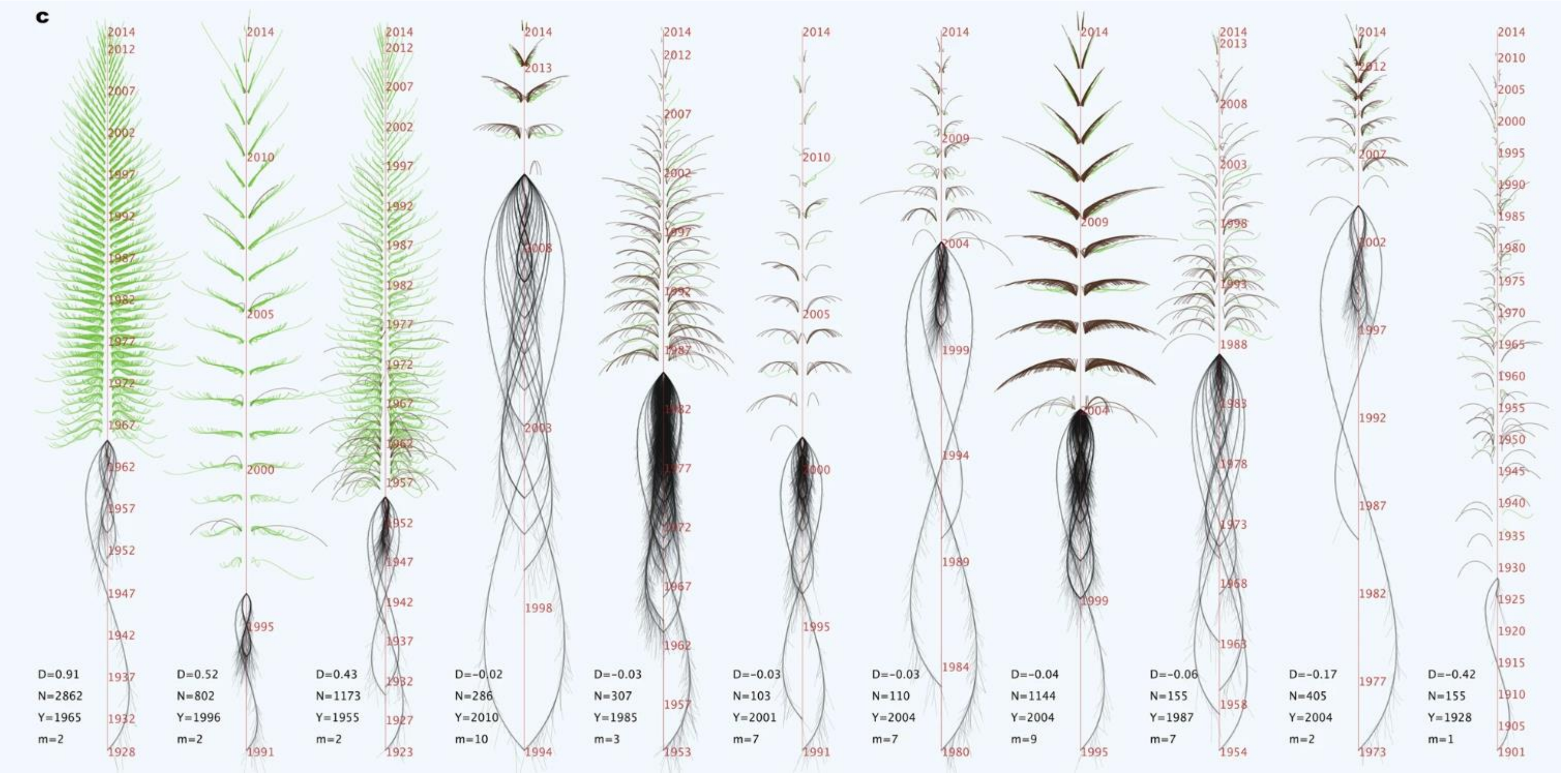
Bigger teams aren't always better in science and tech - *Phys.org*
https://phys.org/news/2019-02-bigger-teams-science-tech.html

- **Large teams produce mainstream research**
  - Accepted by the "big journals”
  - Quickly cited
- **Small teams produce disruptive research**
  - Quirky, innovative
  - Citations take some time



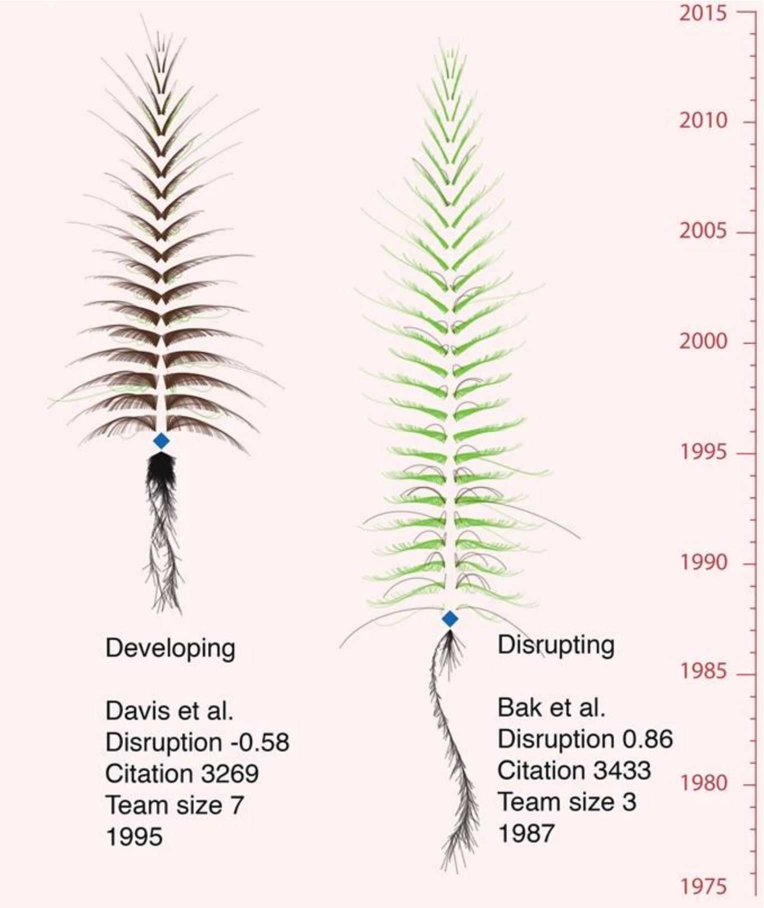
Large teams develop and small teams disrupt science and technology

Due to the increasing speed & size of mainstream science, the "Top 1%" (i.e. most cited) is attracting all the attention. Research that is less immediately impactful is being overlooked. This is leading to a **lack of innovation**.



Motivation

- Leverage the *Small Teams* dataset to identify McMaster's ***most innovative*** (~~cited~~) research.
- Strategic planning: Can we be more *disruptive* in order to *differentiate* ourselves?



Small Teams dataset (*Wu, Wang, and Evans*: **19.4MB**)

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

Includes a “Disruption Score"

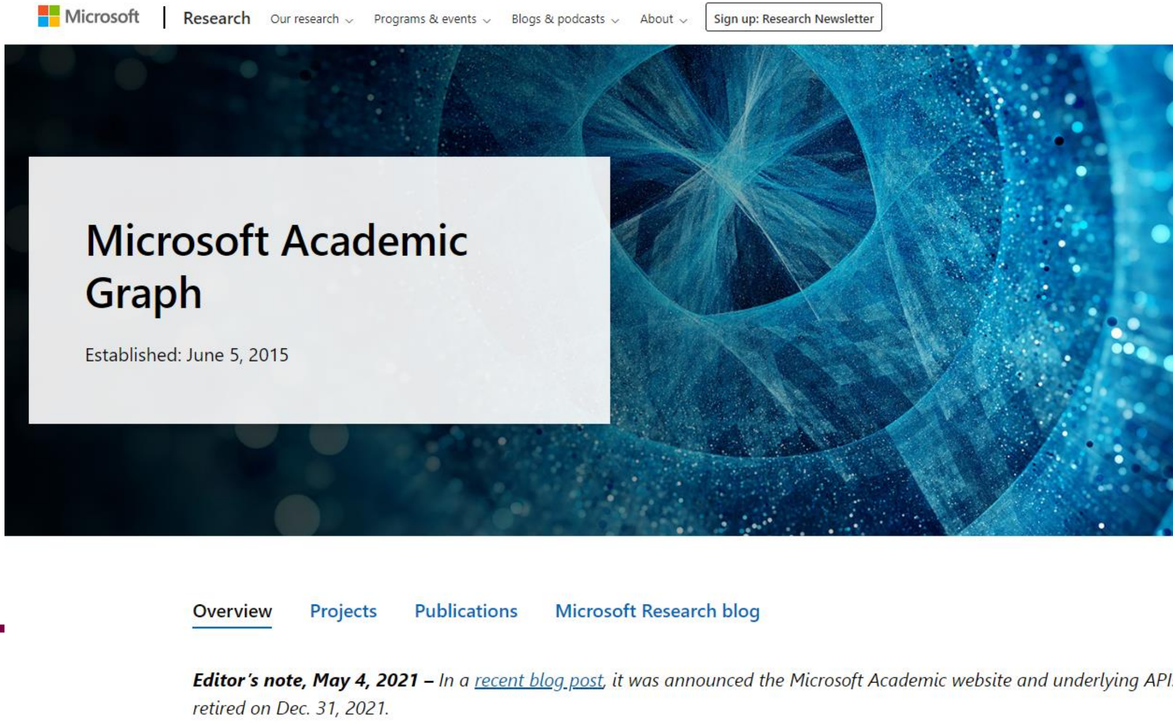
MAGPaperId, Year, Field, Team size, Collab?, Citations, Disruption

```
1970392578 1830  10 1  0  3  0.75
2108276706 1842  5  1  0  1  0.3333333333333333
2022566795 1846  7  1  0  12 0.2
2065789632 1850  9  1  0  3  0.15
219463075  1851  5  1  0  3  0.21428571428571427
```

Microsoft Academic Graph

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15 Companion). ACM, New York, NY, USA, 243-246. http://dx.doi.org/10.1145/2740908.2742839

- 2015 to 2021
- "Graph" in the sense of a social network of metadata.



Microsoft Academic Graph

Established: June 5, 2015

Editor's note, May 4, 2021 – *In a recent blog post, it was announced the Microsoft Academic website and underlying API retired on Dec. 31, 2021.*

OpenAlex.org

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. ArXiv. https://arxiv.org/abs/2205.01833

OpenAlex indexes about **209 million** works, with about 50,000 added daily.
The canonical PID for works is DOI.

New works are collected from many sources:

- Crossref
- PubMed,
- Repositories [institutional and discipline-specific (e.g. ***arXiv***)

Many older works come from the now-defunct Microsoft Academic Graph.

MAG is no more…what to do?

The Microsoft Academic Graph Paper ID lives on as the accession number in OpenAlex

For example, the first row of the dataset has a MAGPaperId of `1970392578`. When preceded by a "W", this MAGPaperId can be used in the OpenAlex API to retrieve the article:

LIKE SO: `https://explore.openalex.org/works/W1970392578`

RESULT = "Baden Powell (1830) *Researches towards Establishing a Theory of the Dispersion of Light*."

We have a pathway:

Small Teams data
-> OpenAlex
-> article metadata
-> [filter]
-> Disruptivity of McMaster's research

“MAGic” is possible via OpenAlex

There is a new R package for this:

openalexR (*Massimo Aria* – Univ of Naples Federico II) https://github.com/massimoaria/openalexR

```
query_inst <- oaQueryBuild(
  entity = "works",
  filter = "institutions.id:I98251732",
  date_from = "2020-01-01", date_to = "2020-12-31"
)
```

Matching records based on a shared ID is easy with SQL.

RSQLite (*SQLite is a self-contained, ‘light’ database – no server required*)

```
SELECT SmallTeams.DisruptionScore, OpenAlexRecords.*
FROM SmallTeams INNER JOIN OpenAlexRecords
ON SmallTeams.MAGid = OpenAlexRecords.MAGid
```

MAGic made easy - Recipe

Ingredients:

- `openalexR` package
- ***Small Teams*** dataset
- `RSQLite` package

Steps:

1. Query **openalexR** for all records from university X for year Y. Load to a dataframe.
2. Load **Small Teams** dataset into a 2nd dataframe.
3. With **RSQLite**, write the dataframes to tables.
4. Use an SQL query to find intersection ("JOIN") of the two tables based on MAG ID.
5. *RESULT = The disruption score of research from X*

Jeff Demaine
Bibliomagician
demainj@mcmaster.ca