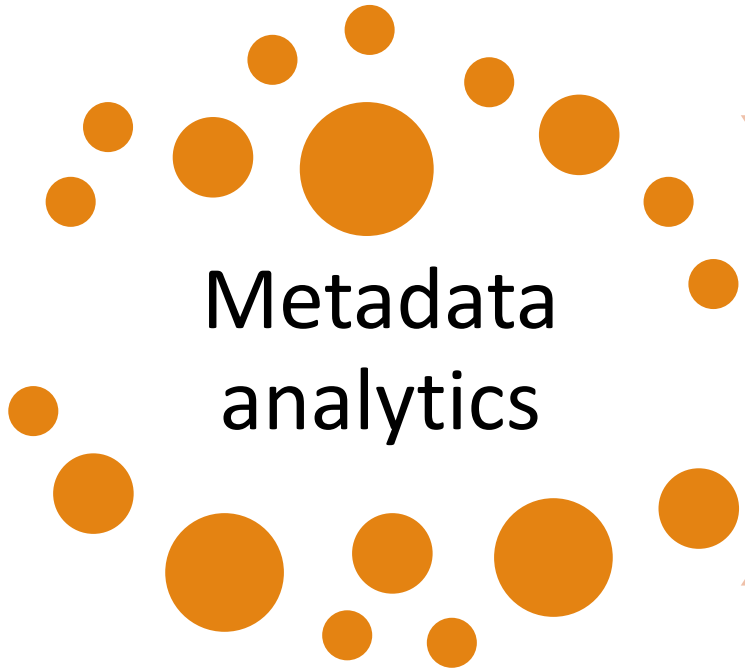
# Conceptual and Computational Workflows in Metadata Analytics

Jian Qin

School of Information Studies

Syracuse University

# Metadata analytics



## What is it?

- The (very brief) history
- The motivation
- The perspective

- Theories
- Methodologies
- Applications

# What is (big) metadata?

An earlier definition

"…the structured, semi-structured, or unstructured descriptions of scientific data stored in repositories" (Bratt et al., 2017)

An updated brief version

The structured or semi-structured descriptions of information and/or data objects.

An updated long version

The structured or semi-structured descriptions of information and/or data objects in the forms of library catalogs, indexing databases, and metadata repositories.

# The (very brief) history

Price, Derek J. de Solla, 1922-1983.

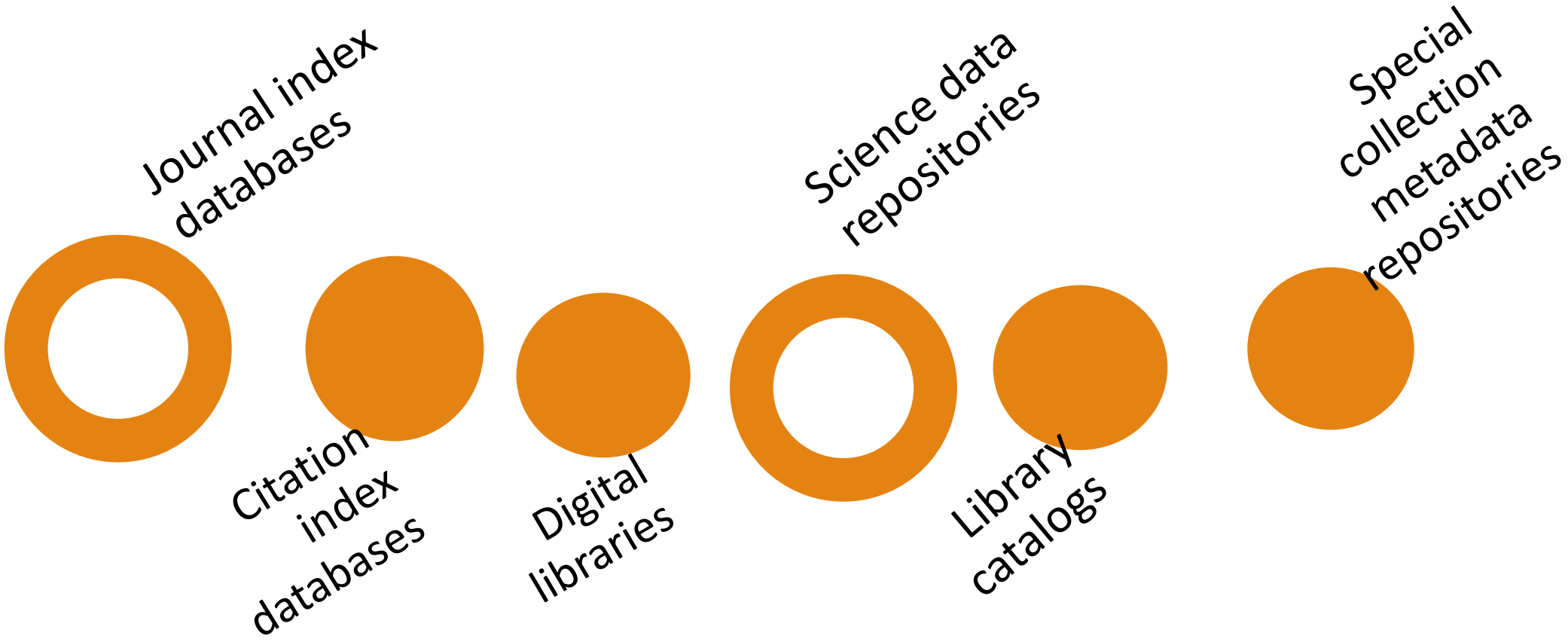
- Price's model: Preferential attachment process
- Power law distribution of citation network, first example of scale-free network
- Price's law: square root law for relationships between authors and publications
- Exponential growth of science and half-life of science literature

Eugene Eli Garfield, 1925-2017

- Science citation index (which inspired the PageRank algorithm by Google co-founders)
- Journal impact factor

Bibliometrics, Scientometrics:
Theories (laws) built on math
Quantitative methods
Macro- and micro-scale
Authors, publications, citations

# The changing landscape of metadata…



# Old issues, great challenges

- The perpetual ambiguous author names → Name disambiguation
- Metadata is never readily usable for analysis → Constant data cleaning and processing
- Difficult to reuse code and workflows → Reinventing the wheel
- Limitations of current metadata structures → Linked data

Very large volume of data and very complex structures require careful planning for metadata analytics to avoid reinventing the wheel and/or waste of time and efforts.

# Workflows are a method for ensuring effectiveness and quality of metadata analytics.

# What is a workflow?

- What is a workflow?
- “the activity of defining the sequence of tasks needed to manage a business or computational science or engineering process” with four broad aspects:
  - composition,
  - mapping,
  - execution, and
  - Provenance.” (Deelman et al., 2009, p. 529).

# An example of conceptual workflow in metadata analytics: Name disambiguation solutions (1)

**Goal**

---

1) improve accuracy of name-centric retrieval of information from GenBank

2) improve accuracy of data integration between GenBank and other sources

Documentation from Evernote

**Task**

---

1) Resolve each Author referenced in Genbank to a unique identifier (resolution).

In Genbank, authors are referenced by first initial and last name, giving rise to the following forms of ambiguity:

A) Multiple authors with the same last name and first initial (polysemy). Example: multiple authors named 'Smith, J.'

B) A single author with multiple name variants (synonymy). This can occur due to a name change, spelling variation (Anglicization of foreign names), or misspelling. Example: a single author referred to as 'Adams, E.' in one record and 'Adams-Hoffert, E.' in another.

# An example of conceptual workflow in metadata analytics: Name disambiguation solutions (2)

2) Enhance metadata associated with each author referenced in GenBank (attribution).

To improve the ability to resolve a given GenBank author to an author referenced in another source, additional metadata will be associated with each uniquely identified author. This metadata could include:

A) Full name
B) Name variants
C) Organizational affiliations
D) Co-author affiliations
E) Subject matter expertise
F) etc.

Scope

---

The name disambiguation application will:

1) need to occasionally re-analyze the GenBank database when a significant amount of new information has been incorporated
2) need to handle updates to external resources such as author information in Pubmed
3) run as an offline (non-realtime) process

# An example of conceptual workflow in metadata analytics: Name disambiguation solutions (3)

Steps

---

Resolution:

1) Get access to data and set up dev / test environment

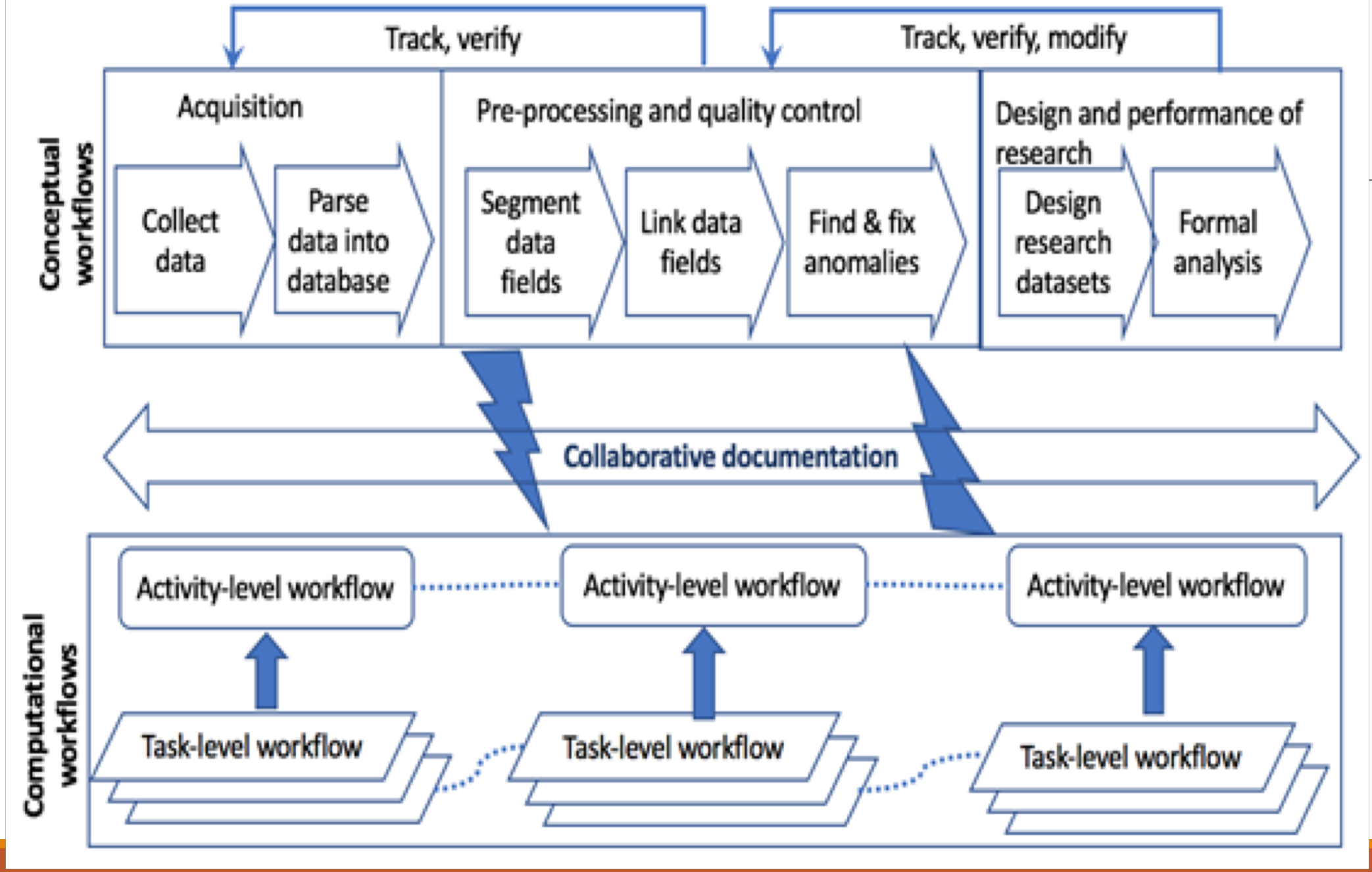
2) Develop ground-truth / test data for assessing accuracy

3) Develop algorithm for assigning similarity score between each pair of author refs in GenBank database based on available metadata

5) Determine similarity threshold for considering two refs as same individual

4) Apply clustering method to group all refs

5) Measure accuracy and modify similarity / clustering algorithms as necessary




*Workflows in (big) metadata analytics (Bratt, Hemsley, Qin, 2017)*

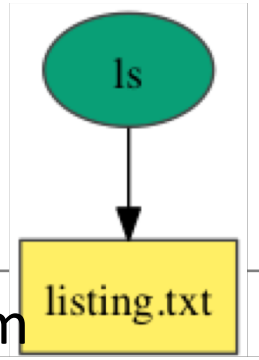
# Why do we need workflows in metadata analytics?

- Align your data collection, processing, and analysis with your research goals
  - make sure the data you collected are you needed for answering your research question
- Make a feasible plan step by step to keep your data and research stay on track
- Establish provenance for your research project to assure the reproducibility and replicability of your research

# Tools for (computational) workflow management

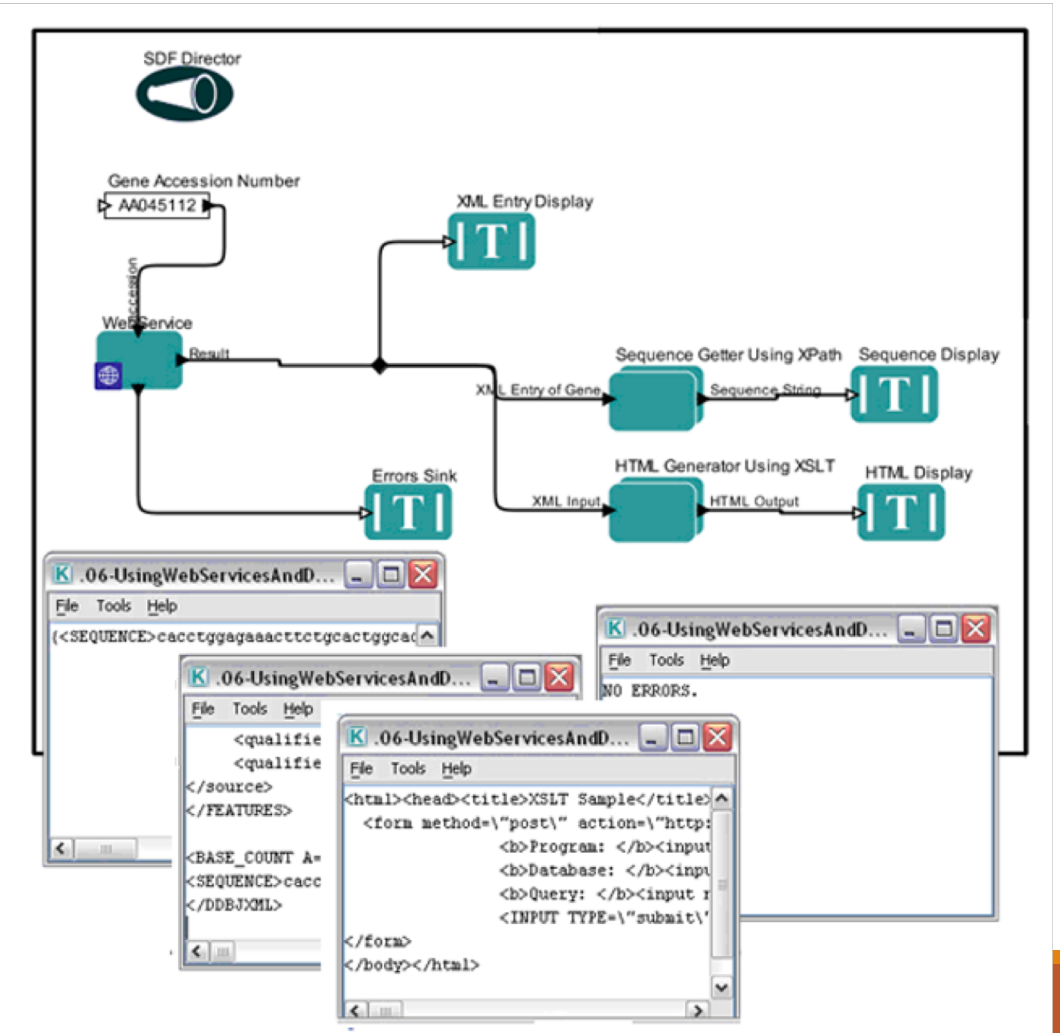
https://pegasus.isi.edu/

Pegasus Workflow Management System




Process

```
<!--  generated: 2013-05-18 11:47:33.892834  -->
<!--  generated by: gideon  -->
<!--  generator: python  -->
<adag xmlns="http://pegasus.isi.edu/schema/DAX"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://pegasus.isi.edu/schema/DAX
http://pegasus.isi.edu/schema/dax-3.4.xsd" version="3.4"
name="process">
  <job id="ID0000001" name="ls">
    <argument>-l /</argument>
    <stdout name="listing.txt" link="output"/>
    <uses name="listing.txt" link="output" register="false"
    transfer="true"/>
  </job>
</adag>
```




Kepler Workflow Management System

https://kepler-project.org/

# Now we have conceptual and computational workflows, what comes next?

# Linked data

- Available on the web
- Available as structured data readable by a machine
- Available in a non-proprietary format
- Expressed using open World Wide Web Consortium (W3C) standards
- Linked to other data on the web

What implications are there to the fields of bibliometrics and scientometrics?

# Name disambiguation solutions

- Traditional solution: use algorithms to automatically disambiguating author names
  - However, if database producers keep current practice in abbreviating names, the problem will remain unresolved.
- New solution: creating globally unique ID for researchers and authors

ORCID

ResearcherID

VIAF
Virtual International Authority File

**LC Linked Data Service**
**Authorities and Vocabularies**

**Union List of Artist Names®**
Getty Research Institute

# Identifying things with globally unique identifiers

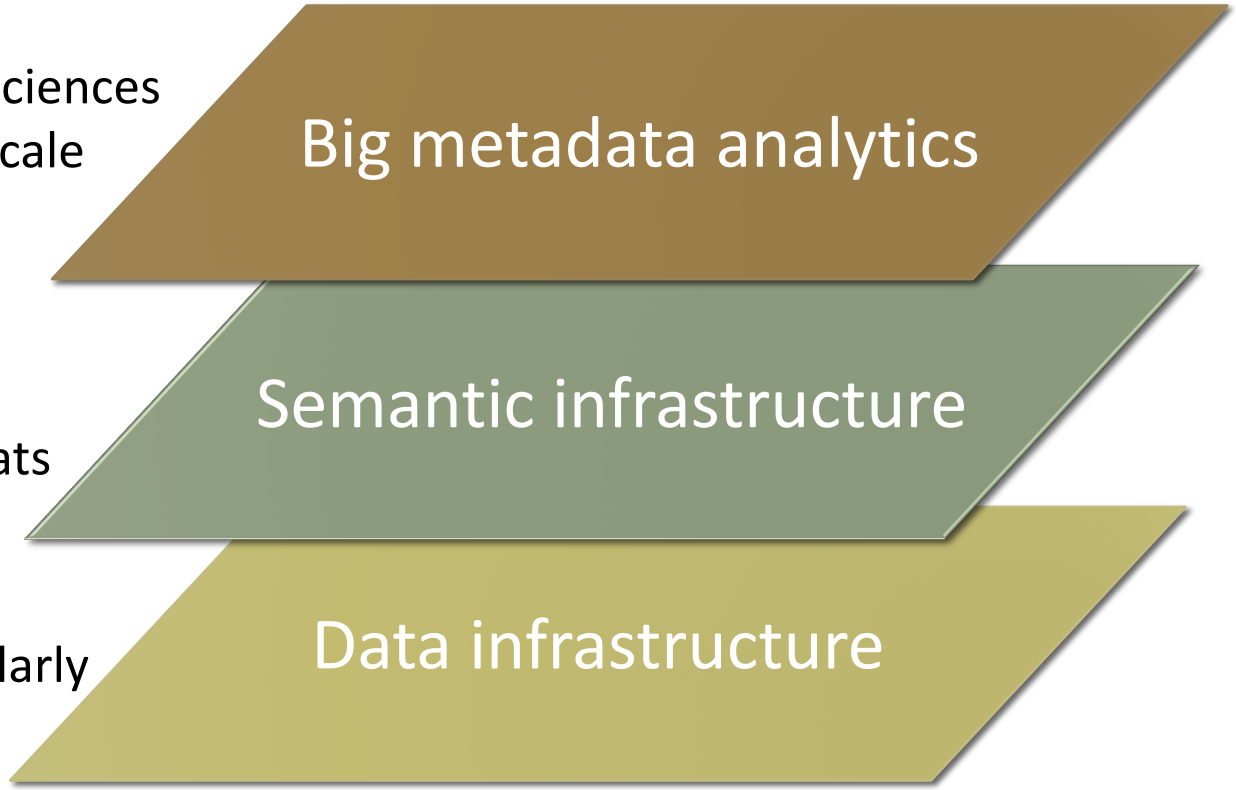
- **Subject terms** in controlled vocabularies
- **Events** (political, cultural, public health, social, …)
- **Publications** (papers, versions of a paper, journals, …)
- **Datasets** (research data, census data, observation data, sensor data…)
- **Cultural objects** (archives, museum objects, digital surrogates of physical objects…)

# Metadata is changing to broaden the research horizon

Bibliometrics and scientometrics
Knowledge discovery for humanities and social sciences
Data services to support interdisciplinary large-scale research

Uniquely identified authors, organizations, taxonomic classes, subject terms, datasets, publications, etc. in structured, linkable formats

Index databases; library catalogs; metadata repositories for datasets; digital libraries for scholarly pubs, special collections, and cultural objects



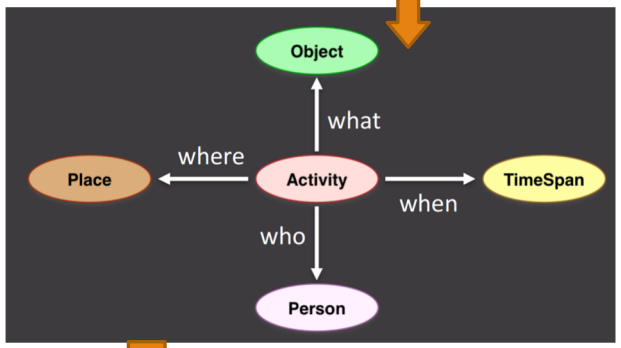
# Example: Mining large (meta)datasets for the humanities

| date_issued | description | coverage | series |
|---|---|---|---|
| 1989-04-03 | <p>"Perestroika, glasnost: they don't even talk about them in Castro's Cuba. Violent insurgencies? The Cubans support them, the Soviets say they prefer political solutions. What are Gorbachev and Castro really talki hopeful? Joining us live from Havana spokesman for the Soviet foreign mi commercials.</p> | | |
| 1989-04-04 | <p>"Imagine a space con Pollard gave the Israelis. What Pollar there to haunt US-Israeli relations. ... Jonathan Pollard's father and the aut Israel." Includes commercials.& | | |
| 1989-04-05 | <p>"It's been a horror-st Massachusetts where the criminally insane and those committed for civil reasons are held together." Includes commercials.</p> | Massachusetts | Nightline |
| 1989-04-06 | <p>"... that's the cost [$100-200 million] of cleaning-up [the Exxon Valdez oil spill], but are Americans willing to pay the much higher price of not having it happen again?" Includes commercials.</p> | Alaska | Nightline |

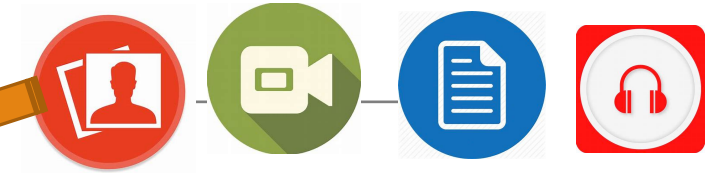
Text mining in metadata to discover trends, patterns, and phenomena for humanities and social sciences scholars

Leonard, Peter. (2014). Mining large datasets for the humanities. IFLA WLIC . http://library.ifla.org/930/1/119-leonard-en.pdf

# Example: Remodeling and transforming special collection metadata to linked archives




modeling

Archival collections

```
"@context": "https://linked.art/ns/v1/linked-art.json",
"id": "https://data.getty.edu/museum/11733",
"type": "ManMadeObject",
"label": "Attic Black-Figure Neck Amphora",
"classified_as":
[{"id": "aat:300148696", "type": "Type", "label": "Amphora"}]
"made_of":
[{"id": "aat:300015045", "type": "Material", "label": "Terracotta"}]
"produced_by":
{"id": ".../create", "type": "Production", "label": "Creation event"}
// ...
}
```

Ready for publishing and sharing as well as consuming by machines

Model and linked data examples from:
Sanderson, R. & Newbury, D. (2018). Linked.Art: Our Linked Open Usable Data Model. https://www.slideshare.net/azaroth42/**linkedart-data-model-walkthrough**

# Case study: GenBank Metadata mining

"From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months."

-- NCBI. (2017). Growth of GenBank and WGS, http://www.ncbi.nlm.nih.gov/genbank/statistics.

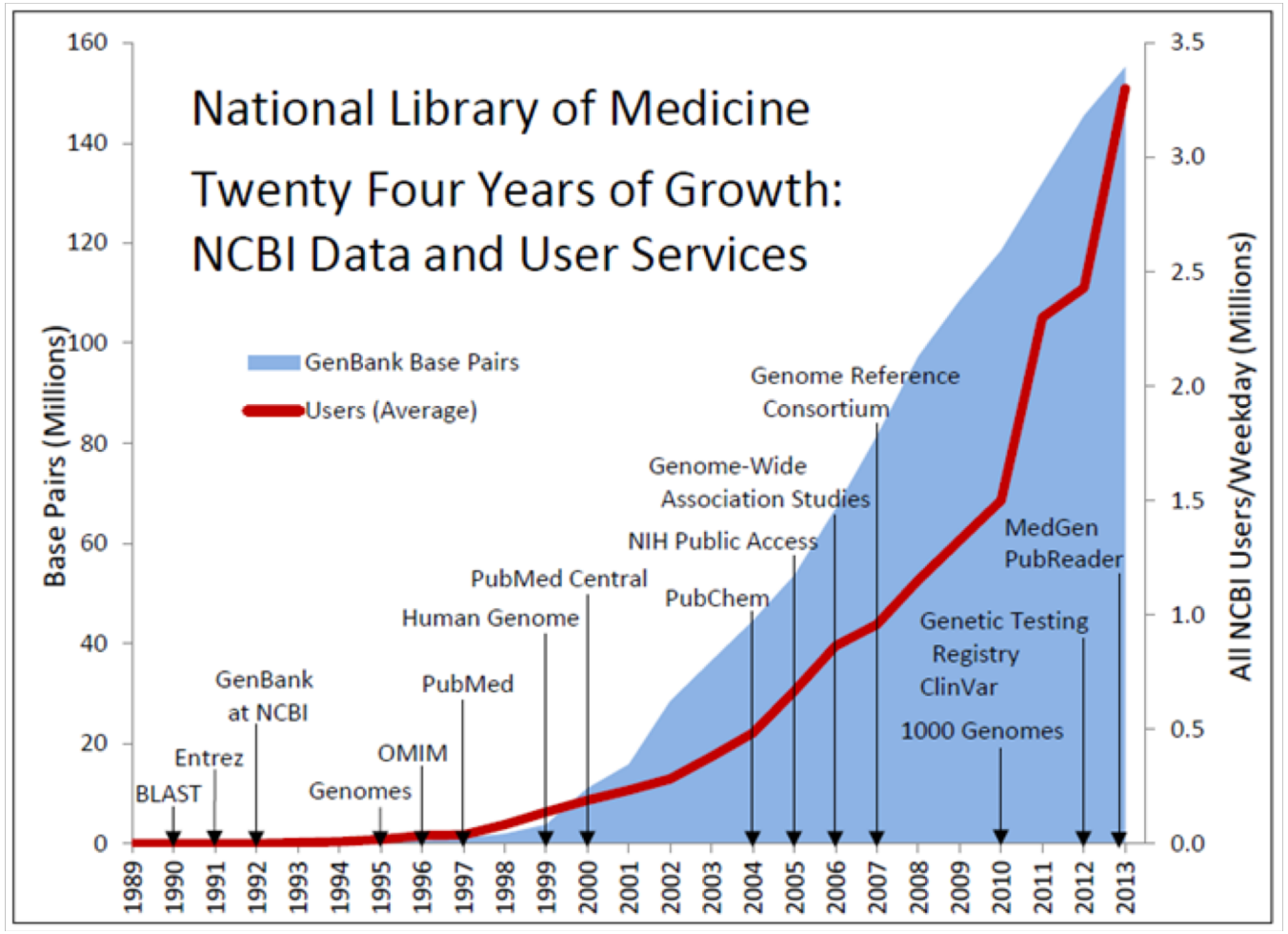


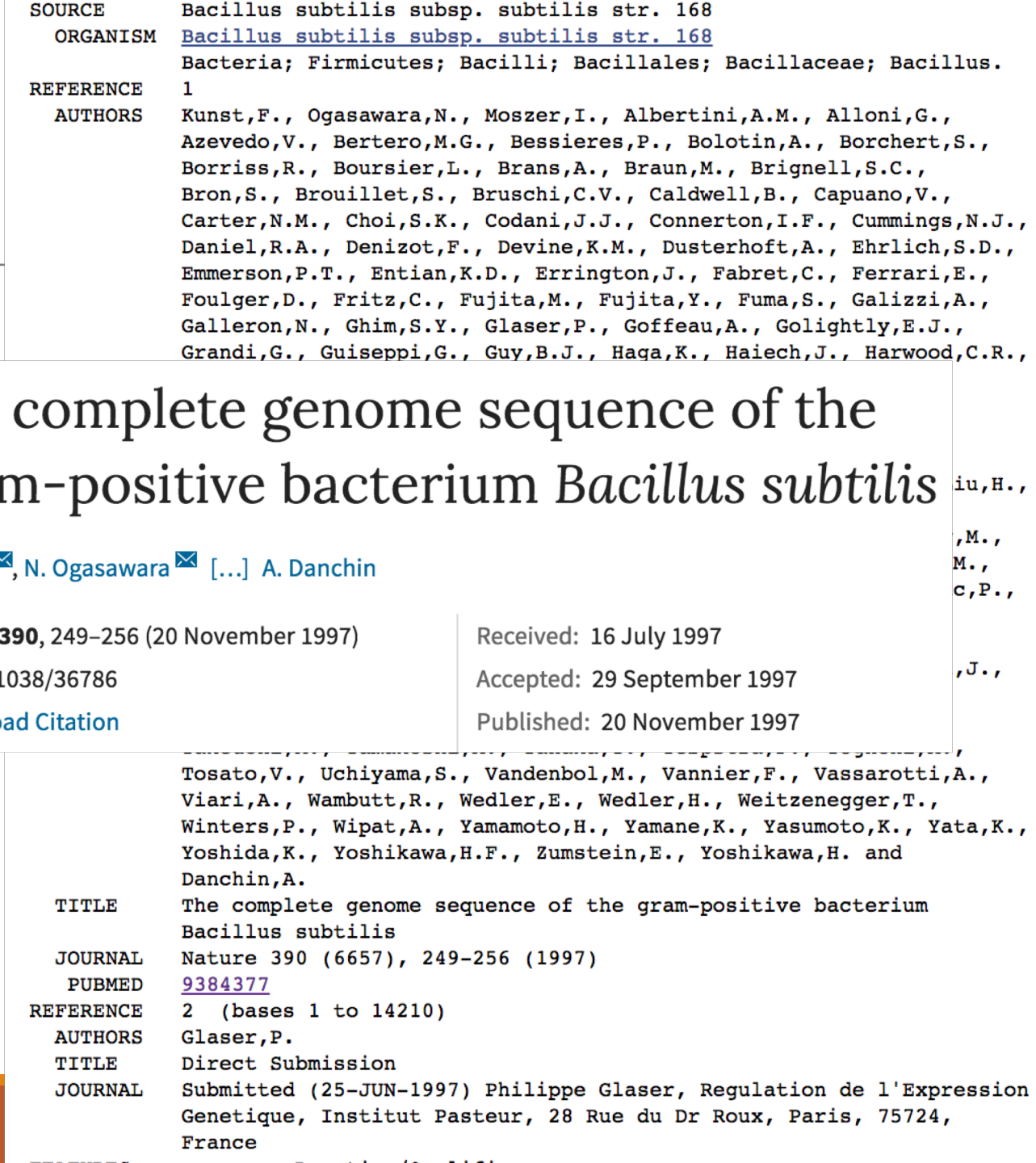

Image credit: https://www.nlm.nih.gov/about/2015CJ.html

# GenBank's big metadata as a source for quantitative studies of team science

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

# Collaboration across countries, labs, and fields

- Big problems, big data (and big metadata), and big teams
- Relations between data production and paper publication
- Large scale studies of collaboration networks to find patterns, structures, and empirical evidence for in-depth exploration



SOURCE Bacillus subtilis subsp. subtilis str. 168
ORGANISM Bacillus subtilis subsp. subtilis str. 168
Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus.
REFERENCE 1
AUTHORS Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S., Borriss,R., Boursier,L., Brans,A., Braun,M., Brignell,S.C., Bron,S., Brouillet,S., Bruschi,C.V., Caldwell,B., Capuano,V., Carter,N.M., Choi,S.K., Codani,J.J., Connerton,I.F., Cummings,N.J., Daniel,R.A., Denizot,F., Devine,K.M., Dusterhoft,A., Ehrlich,S.D., Emmerson,P.T., Entian,K.D., Errington,J., Fabret,C., Ferrari,E., Foulger,D., Fritz,C., Fujita,M., Fujita,Y., Fuma,S., Galizzi,A., Galleron,N., Ghim,S.Y., Glaser,P., Goffeau,A., Golightly,E.J., Grandi,G., Guiseppi,G., Guy,B.J., Haga,K., Haiech,J., Harwood,C.R.,

The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*

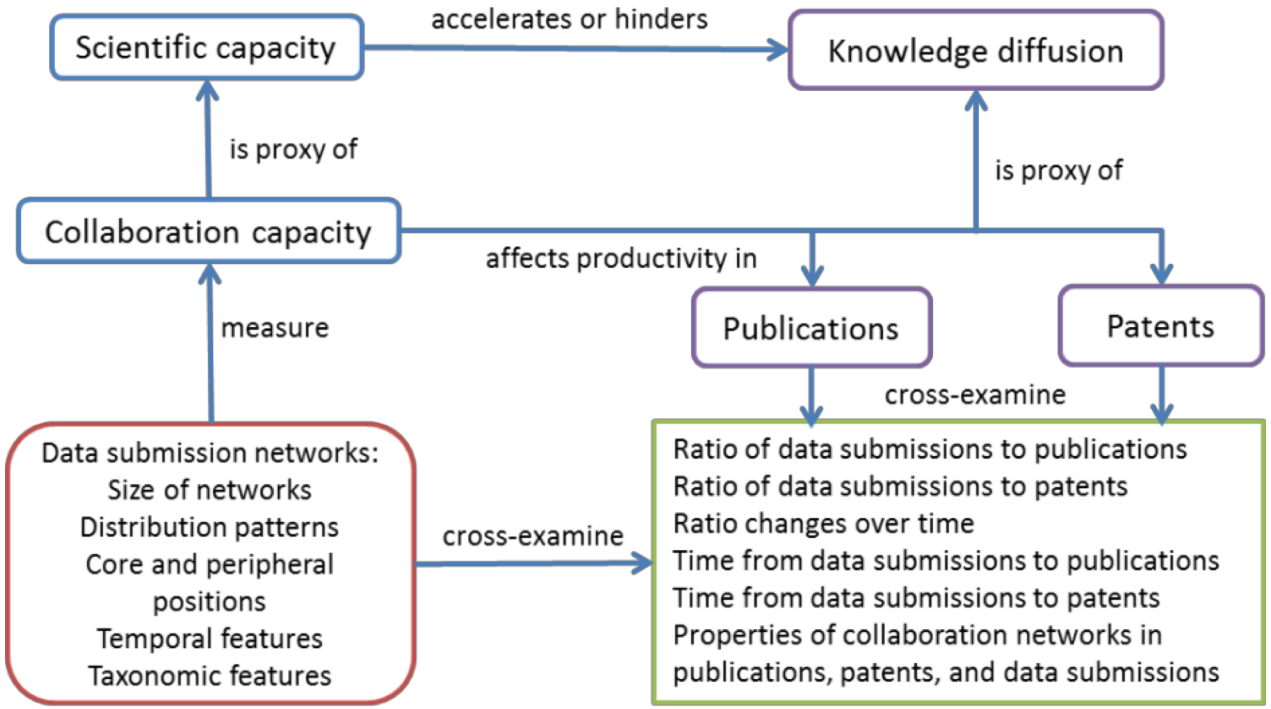
F. Kunst, N. Ogasawara [...] A. Danchin

*Nature* **390**, 249–256 (20 November 1997)
doi:10.1038/36786
Download Citation

Received: 16 July 1997
Accepted: 29 September 1997
Published: 20 November 1997

Tosato,V., Uchiyama,S., Vandenbol,M., Vannier,F., Vassarotti,A., Viari,A., Wambutt,R., Wedler,E., Wedler,H., Weitzenegger,T., Winters,P., Wipat,A., Yamamoto,H., Yamane,K., Yasumoto,K., Yata,K., Yoshida,K., Yoshikawa,H.F., Zumstein,E., Yoshikawa,H. and Danchin,A.
TITLE The complete genome sequence of the gram-positive bacterium Bacillus subtilis
JOURNAL Nature 390 (6657), 249-256 (1997)
PUBMED 9384377
REFERENCE 2 (bases 1 to 14210)
AUTHORS Glaser,P.
TITLE Direct Submission
JOURNAL Submitted (25-JUN-1997) Philippe Glaser, Regulation de l'Expression Genetique, Institut Pasteur, 28 Rue du Dr Roux, Paris, 75724, France

# The collaboration capacity framework




(Qin et al., 2018)

**Collaboration capacity:** the ability of an individual researcher or a team of researchers to collaborate throughout the data production and publication lifecycle and sustain a network of collaborators over time.

Assumptions:

- Collaboration capacity is a proxy for studying scientific capacity
- Data, publication, and patent together can be used as a proxy for studying knowledge diffusion
- Collaboration capacity significantly affects the level of research productivity and extent of knowledge diffusion

# Methods

- Source: metadata describing molecular sequences in GenBank

- Exploratory data analysis (EDA)
- Social Network Analysis (SNA)

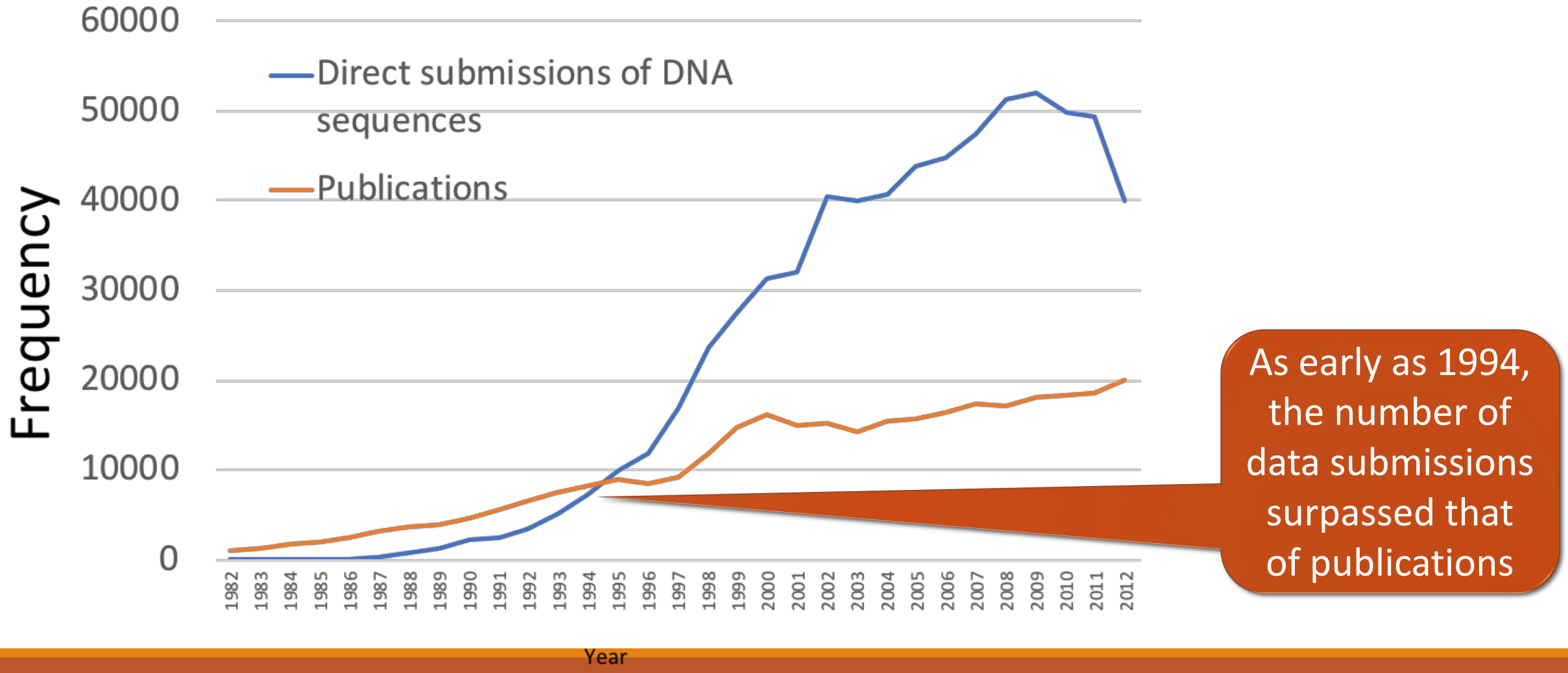
Purpose: using descriptive stats and visualization techniques to look for patterns, structures, and problems

- Based on our framework, datasets generated include:
  - Size of collaboration networks for data submission
  - Extent of knowledge diffusion
  - Rate of knowledge diffusion

# Findings

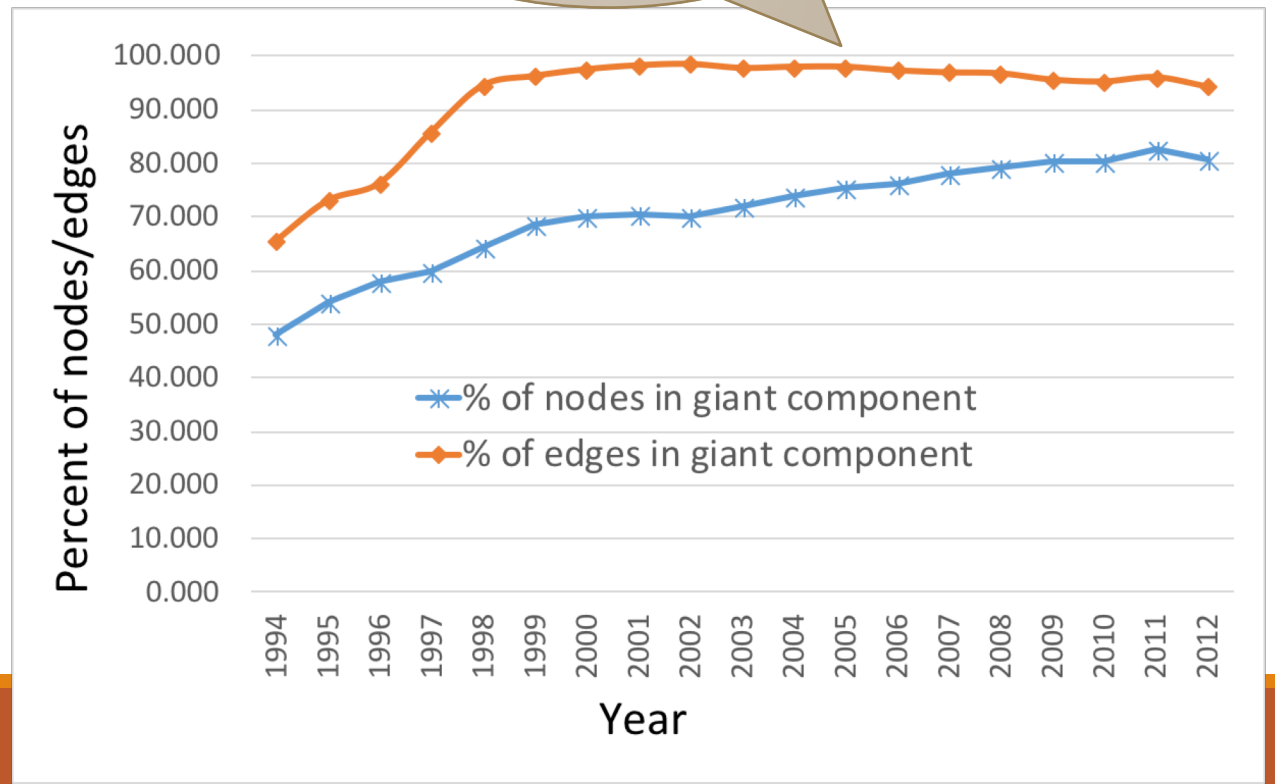
- Connectedness of collaboration networks
- Ratio of data submissions to publications

# Data submissions vs. publications

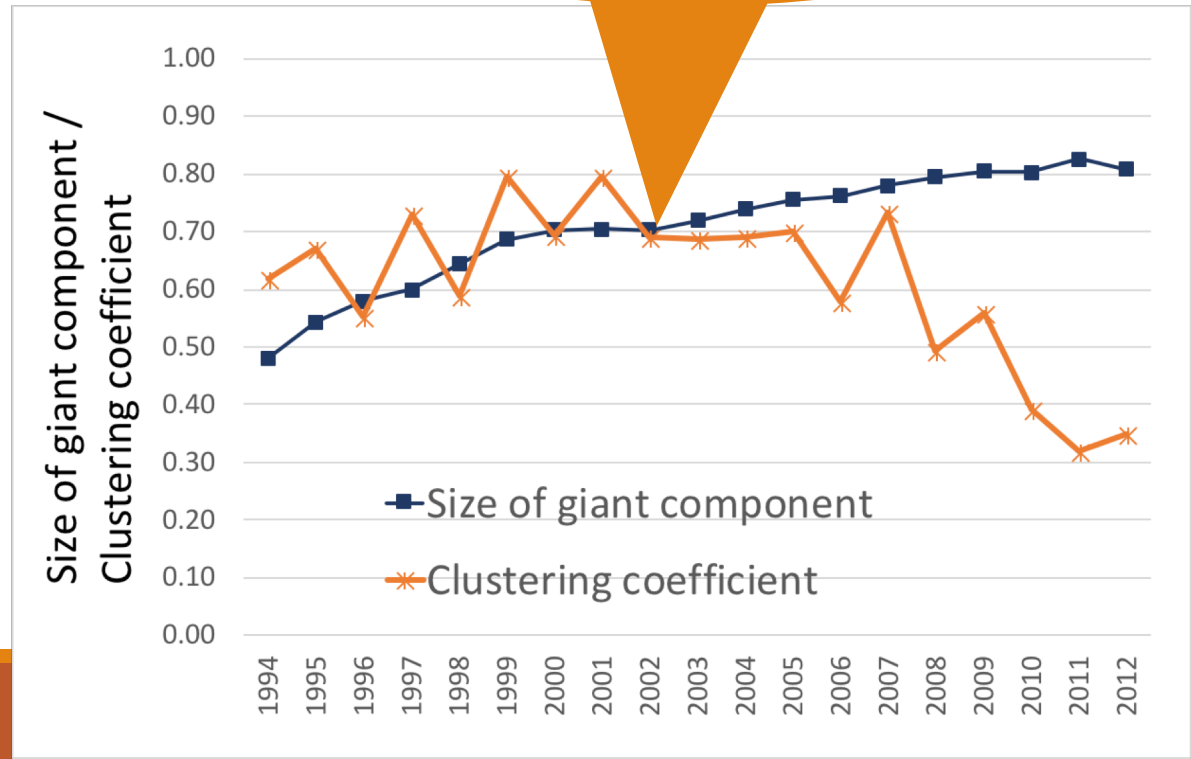


# Connectedness vs. distributedness

Authors remain well connected over time

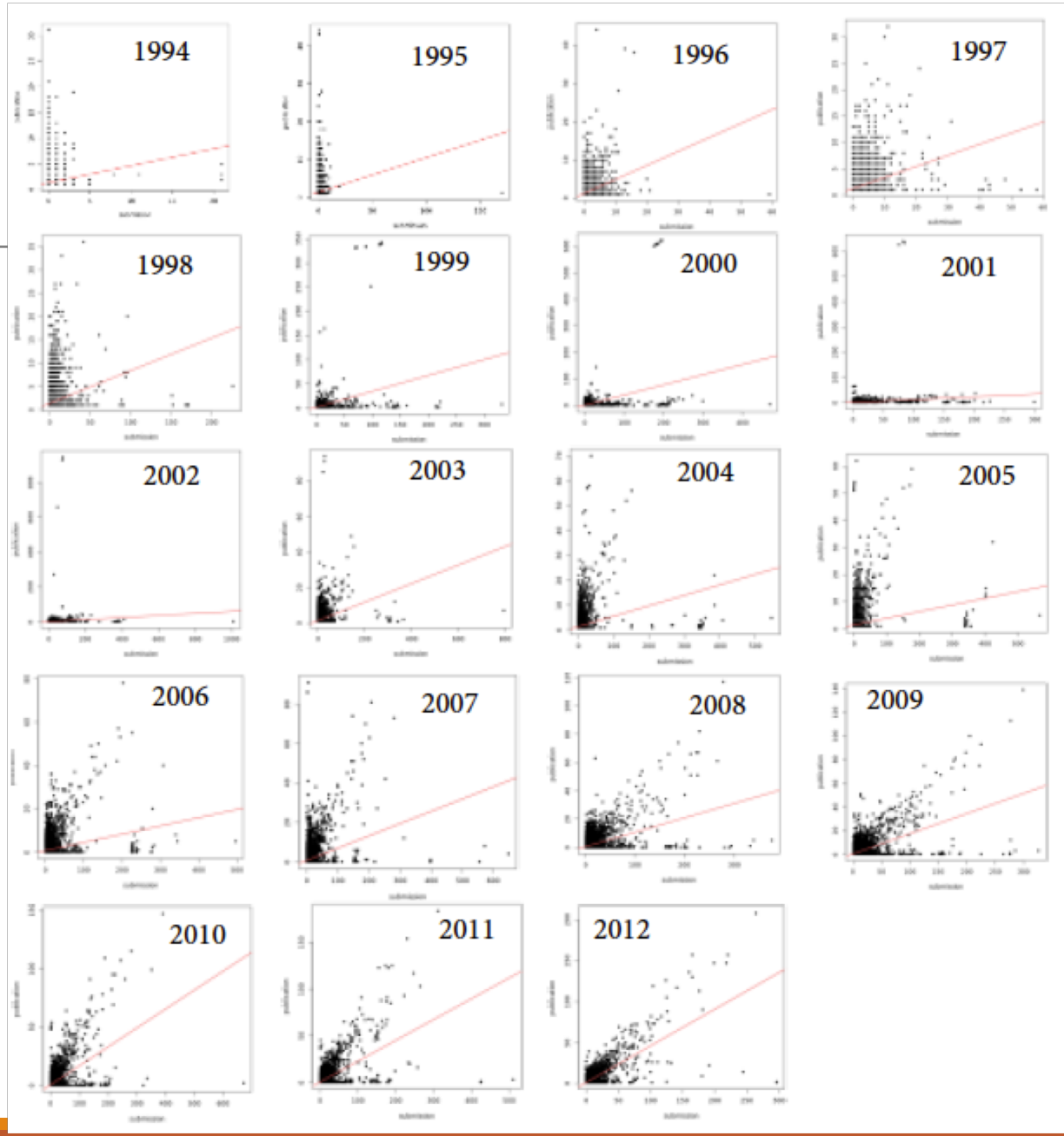


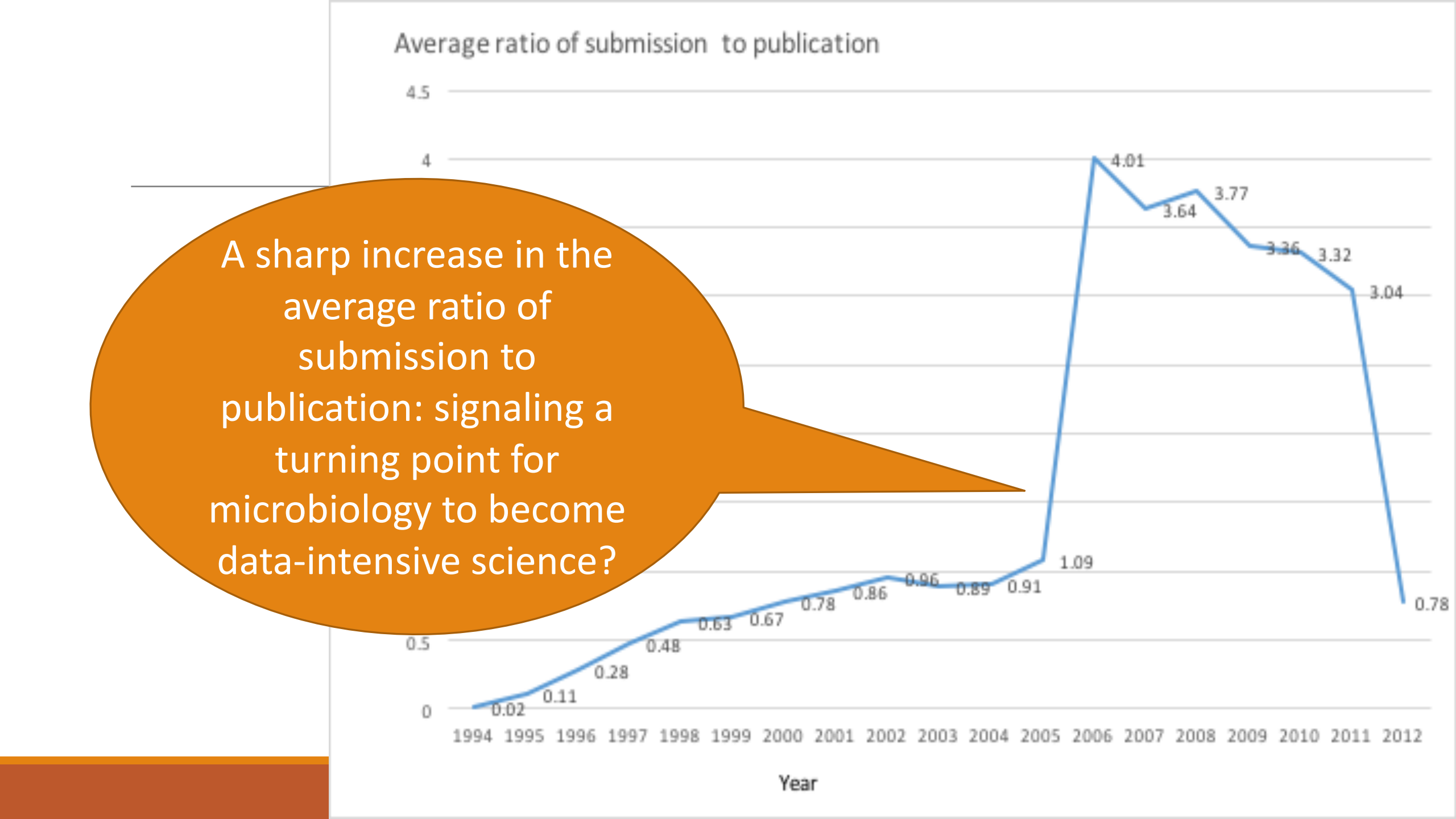

While more clusters of smaller communities emerged



# Ratio of submissions to publications

- x axis: # of authors who submitted sequence data
- y axis: # of authors who published a paper associated with the data submissions
- After 1998, more authors were involved in data production than those in paper publications
- Significant increment in productivity:
  - Before 1998, majority had a range between 20 publications and 50 data submissions
  - Since 2008, a sizable # of authors had a high productivity in the range of 50~100 publications and 100~300 data submissions




Average ratio of submission to publication
4.5
4
0.5
0
0.02
0.11
0.28
0.48
0.63
0.67
0.78
0.86
0.96
0.89
0.91
1.09
4.01
3.64
3.77
3.36
3.32
3.04
0.78
1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
Year
A sharp increase in the average ratio of submission to publication: signaling a turning point for microbiology to become data-intensive science?

# Conclusion

- (Big) metadata analytics uses metadata as the data source to:
  - Study phenomena, trends, behaviors, and relations
  - Produce semantically precise, linked data for better discovery, access, and management of information resources and datasets
- As an emerging research field, it faces great challenges in
  - Methodologies: workflows, tools, and practices that reduce reinventing the wheel and enhance research reproducibility
  - Data: scattered, in different formats, messy, and over 80% of time spent in getting data ready for analysis

# References


- Bratt, S., Hemsley, J., Qin, J. & Costa, M. (2017), Big data, big metadata and quantitative study of science: A workflow model for big scientometrics. *Proc. Assoc. Info. Sci. Tech.,* 54: 36–45. doi:10.1002/pra2.2017.14505401005
- Qin, J., J. Hemsley, & S. Bratt. (2018). Collaboration capacity: Measuring the impact of cyberinfrastructure-enabled collaboration networks. Science of Team Science (SCITS) 2018 Conference, Galveston, Texas, May 21-24, 2018.