Predicting the Future Evolution of Scientific Output

Yannis Manolopoulos
Open University of Cyprus

1st AIMinScience Workshop, Lyon, 25 August 2020

Thanks to

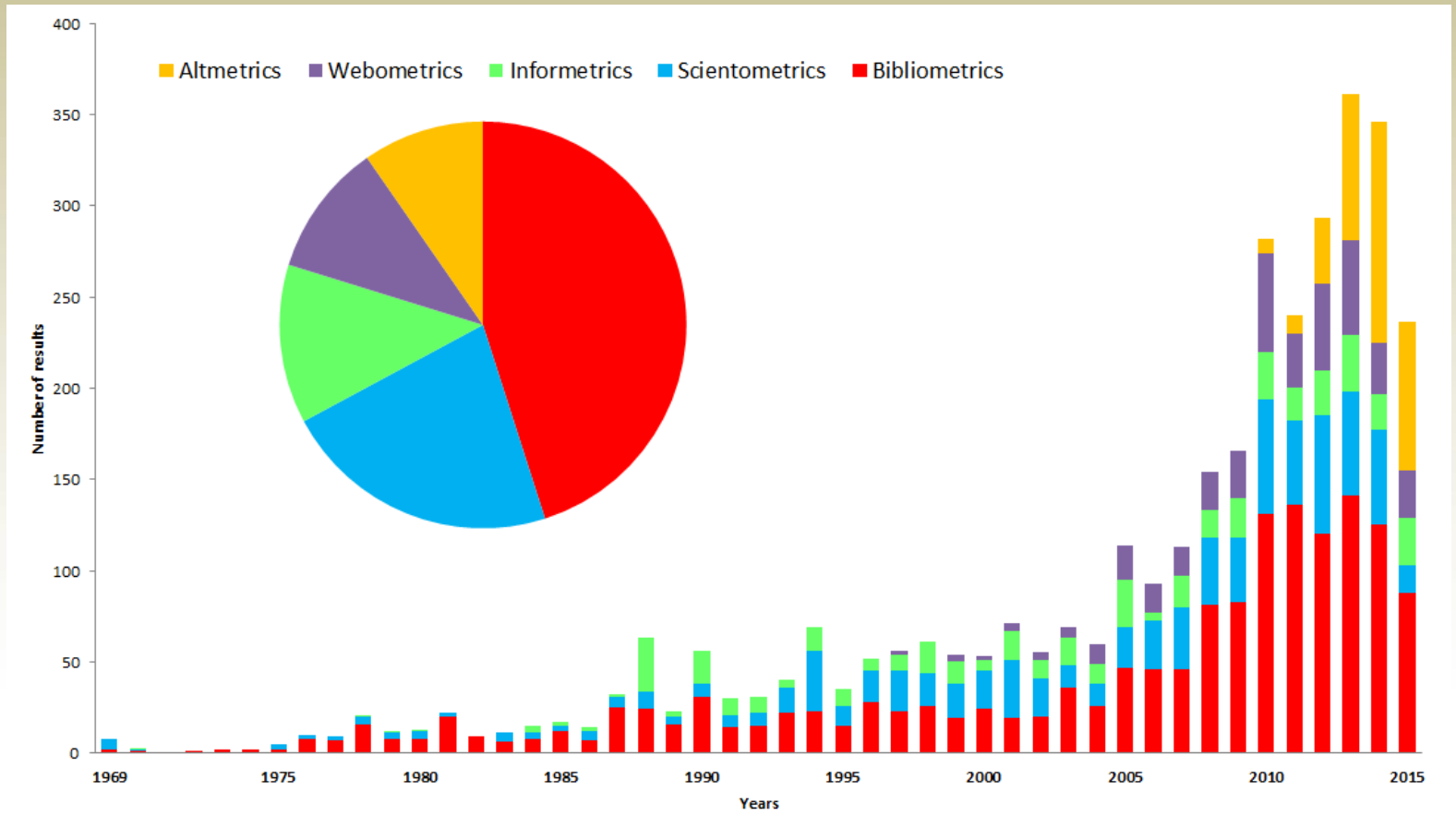
- Dimitris Katsaros
- Antonis Sidiropoulos
- Antonia Gogoglou
- Theodora Tsikrika
- George Stoupas

Outline

- Scientometrics
- Prediction of influence
- Taxonomy of approaches
- Factors affect prediction
- Challenges
- Own contribution
- Epilogue

What is Scientometrics ?

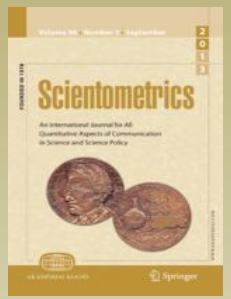
- Science of Science
- Measure and analyze science, technology and innovation



Number of results returned by Google Scholar for the terms Bibliometrics, Scientometrics, Informetrics, Webometrics, and Altmetrics, searching only within the titles of the documents, by year (1969-2015)

Major outlets in the field

JASIST,
1950,
Wiley



Scientometrics
1978, Springer

Informetrics
2007, Elsevier

Collnet Journal of Scientometrics & Info Management
2007, Taylor & Francis

IJ of Bibliometrics in Business & Management
2017, Interscience

Quantitative Science Studies
2019, MIT Press

ISSI, Leuven, 2021

WIS, Sri Lanka, 2020

Why is it important?

- Assist government, and society in general, make better R&D management decisions and assess the likely outcomes
- Policy makers and researchers need to assess the impacts of a nation's or institution's scientific enterprise
- Identify novelty and innovation in science portfolios
- Peer review is based on personal judgment, time-consuming and costly

What determines future influence?

- **Price (1965)**: current visibility, publishing venue and age highly influence a publication's future impact
- Are *institution affiliations*, *collaborations* and *interdisciplinarity* decisive factors in future outreach?
- How important is *current status* and *publishing patterns*?

Derek J. de Solla Price: Networks of Scientific Papers, Science, 30 July 1965.

Is it possible to predict future influence?

- Metrics for scientific output are cumulative but impact can decrease?
- Complicated underlying mechanism that determines future output - random effect too?
- Science is so diverse and dynamic… "**one-to-fit-all**" approaches good enough?

Big Data and Data Intelligence

- Vast and varied ecosystem of recorded bibliometric data is growing in volume, velocity and variety (Big Data Era)

- Human knowledge and understanding is limited → need for Data Intelligence
- Data Intelligence has been utilized in various disciplines (marketing, business, security, etc.)

Taxonomy of approaches

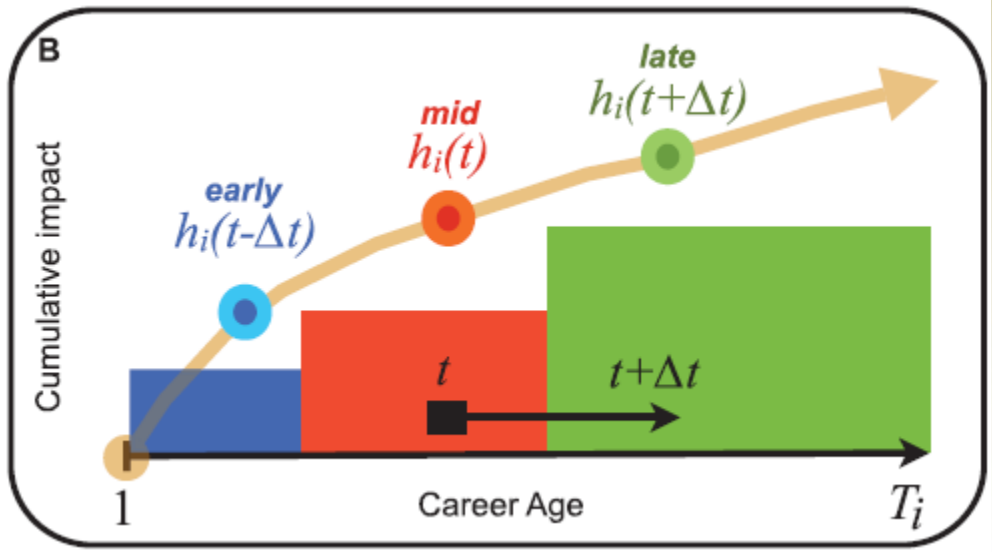
Categorization attribute	Examples of each category	Related work
Scientific entity	Publication	[3, 5–8, 10–12, 16, 21, 25, 37, 40]
	Author	[1, 17, 20, 23, 28, 29, 31, 34, 36, 37, 41]
	Venue	[6]
	Institution	[36]
Target variable	h-index	[1, 31]
	Citation count	[7, 12, 16, 20, 23, 28, 40]
	Increase in h-index/citations	[11]
	Shift in impact group	[6, 34]
	Relative ranking	[3, 5, 8, 17, 21, 31, 34]
	Rank position in a network	[21, 25, 37, 41]
	Award or promotion	[29, 33, 36]
Model	Classification	[5, 6, 8, 11, 23]
	Regression	[1, 3, 12, 16, 17, 20, 21, 36]
	Statistical modeling	[29, 34]
	Time series	[28]
	Citation networks	[25, 37, 41]
	Combination of the above	[7, 10, 40]

Scientific entities

- Scientific entity: most approaches focus on **publications,**
- Higher availability of complete records for publications
- The evaluation of other 3 categories results from aggregation of respective portfolios
  - For authors, etc. increased complexity of calculations
  - Difficulties in disambiguation across online sources for the other 3 categories

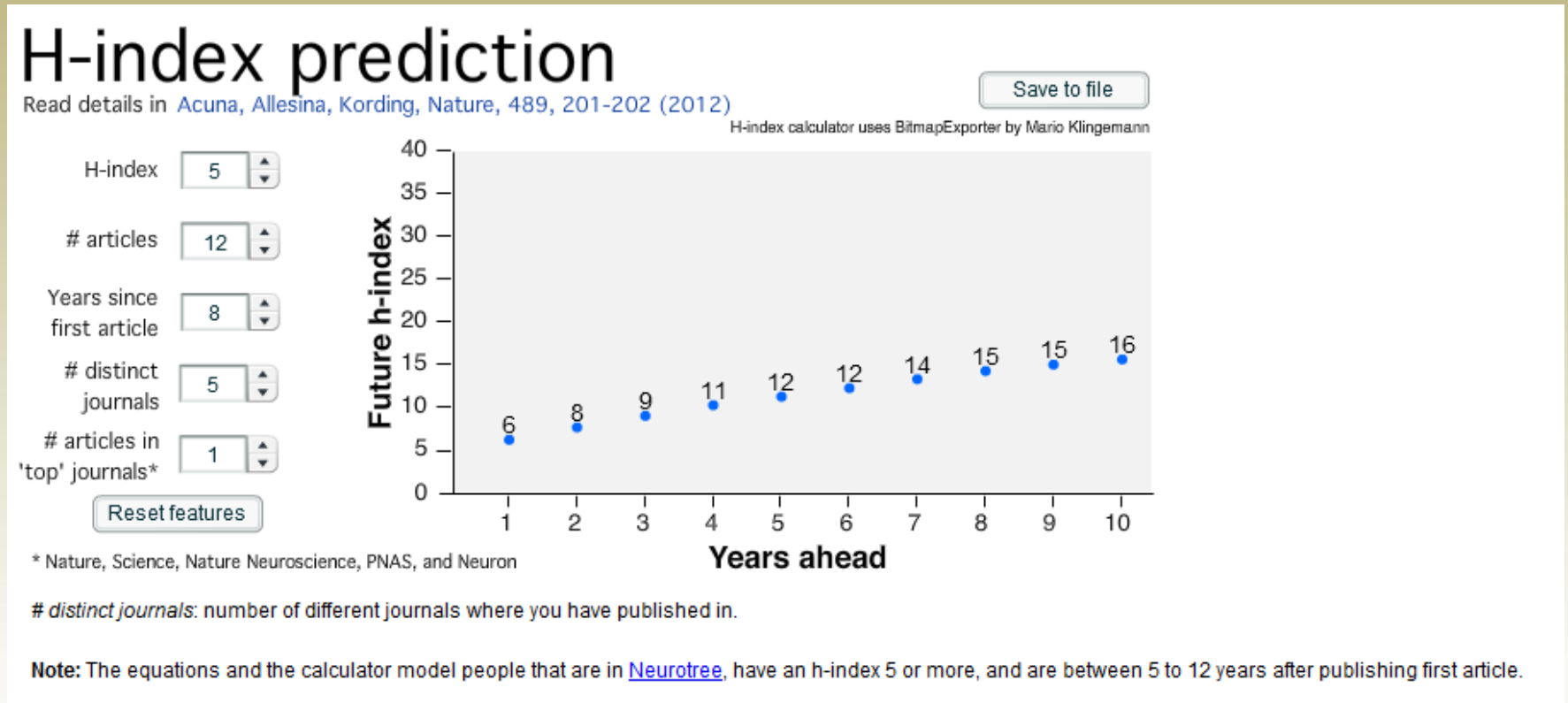
Target variable (1)

- *h*-index is a popular quantity to be predicted
  - More stable than citation counts, limited range and non decreasing values



Orion Penner, Raj Pan, Alexander Petersen, Kimmo Kaski and Santo Fortunato: "On the Predictability of Future Impact in Science", Scientific Reports, 2013.

Target variable (2)



- http://klab.smpp.northwestern.edu/h-index.html

Target variable (3)

- **Predicting next year** ($R^2 = 0.92$):

$$h_{+1} = 0.76 + 0.37\sqrt{n} + 0.97h - 0.07y + 0.02j + 0.03q$$

- **Predicting 5 years into the future** ($R^2 = 0.67$):

$$h_{+5} = 4 + 1.58\sqrt{n} + 0.86h - 0.35y + 0.06j + 0.2q$$

- **Predicting 10 years into the future** ($R^2 = 0.48$):

$$h_{+10} = 8.73 + 1.33\sqrt{n} + 0.48h - 0.41y + 0.52j + 0.82q$$

Key: n, number of articles written; h, current h-index; y, years since publishing first article; j, number of distinct journals published in; q, number of articles in *Nature, Science, Nature Neuroscience, Proceedings of the National Academy of Sciences* and *Neuron*.

Daniel Acuna, Stefano Alesina and Konrad Kording: "Predicting Scientific Success", Nature, 489, 201-202, 2012.

Target variable (4)

- Exponential distribution of bibliometric quantities and their crude nature led to alternative formulation of prediction
  - Will this paper increase your h-index?
  - When this paper will get a first citation?
  - What will be the yearly increase of the target variable?
  - Will you reach the top ranking / group of a venue, institution or research field?

Modeling approaches (1)

- **Classification:** a set of predefined categories is based on the current and past state → future approximated by the category behavior
- **Issues:**
  - limiting and oversimplified
  - behavior often deviates from one's cohort
  - predicting only certain aspects of future state not total output

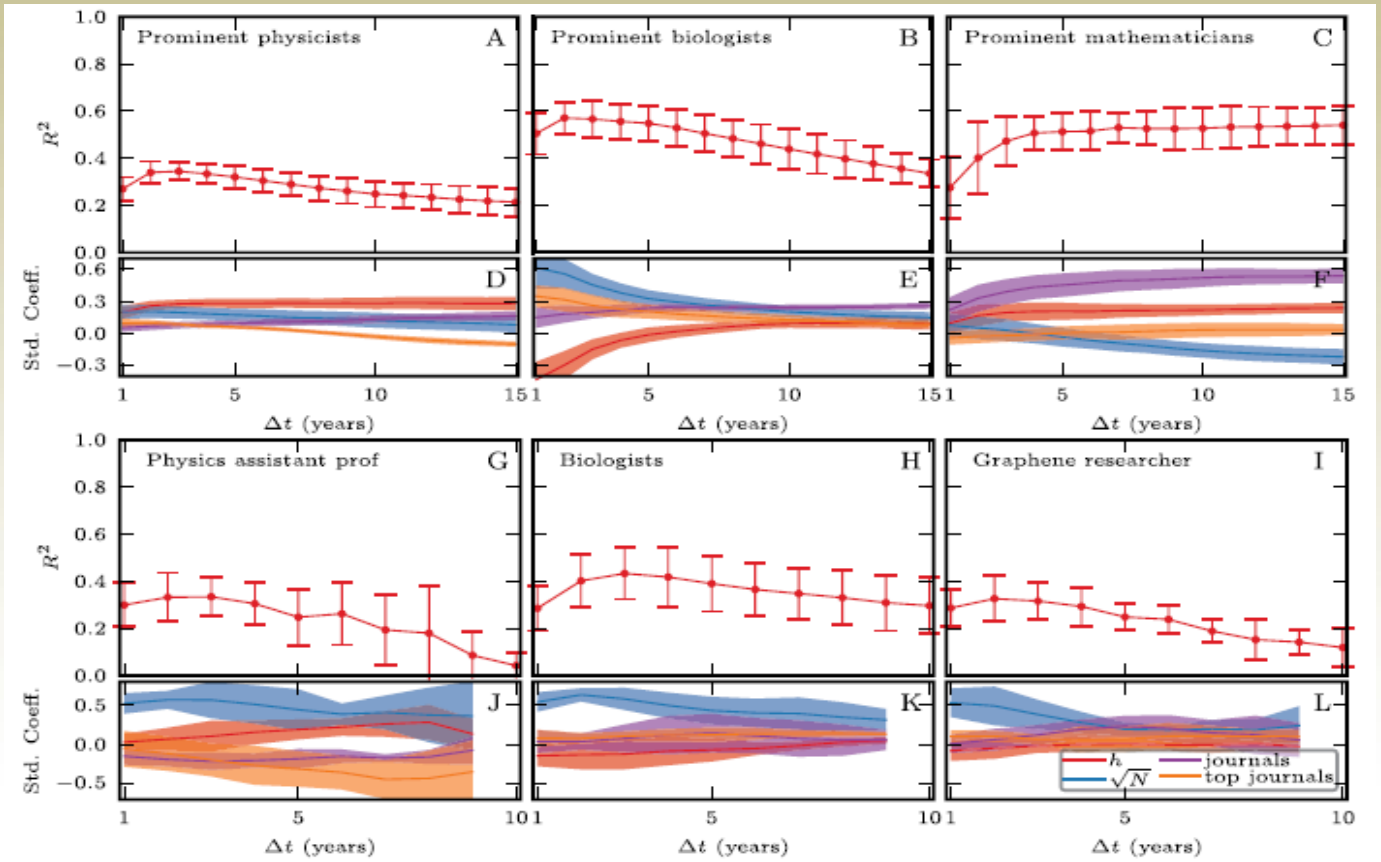
Modeling approaches (2)

- **Regression:** predict a numerical quantity, e.g. *h*-index or citation counts. Seminal work by Acuna [1]
- **Issues:**
  - Difficulty predicting highly skewed distributions
  - "One-to-fit-model" cannot be adapted to various publishing patterns
  - Need for fine-tuned, age and field adjusted models

Daniel Acuna, Stefano Alesina and Konrad Kording: "Predicting Scientific Success", Nature, 489, 201-202, 2012.

Modeling approaches (3)

The predictive power of h-index increments for different disciplines averaged out over multiple age cohorts and performance levels



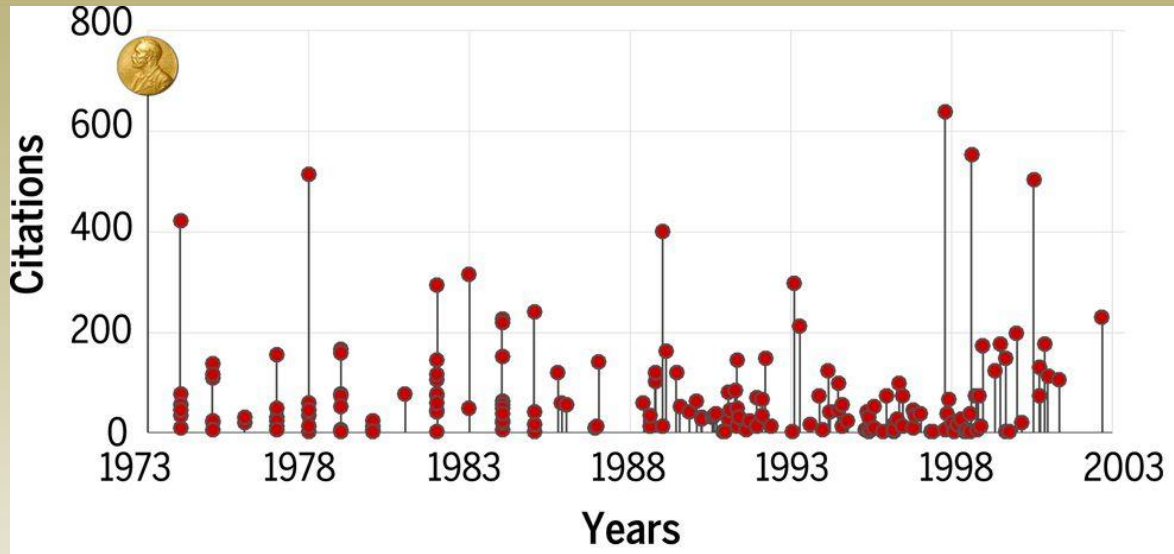
Orion Penner, Raj Pan, Alexander Petersen, Kimmo Kaski and Santo Fortunato: "On the Predictability of Future Impact in Science", Scientific Reports, 2013.

Modeling approaches (4)

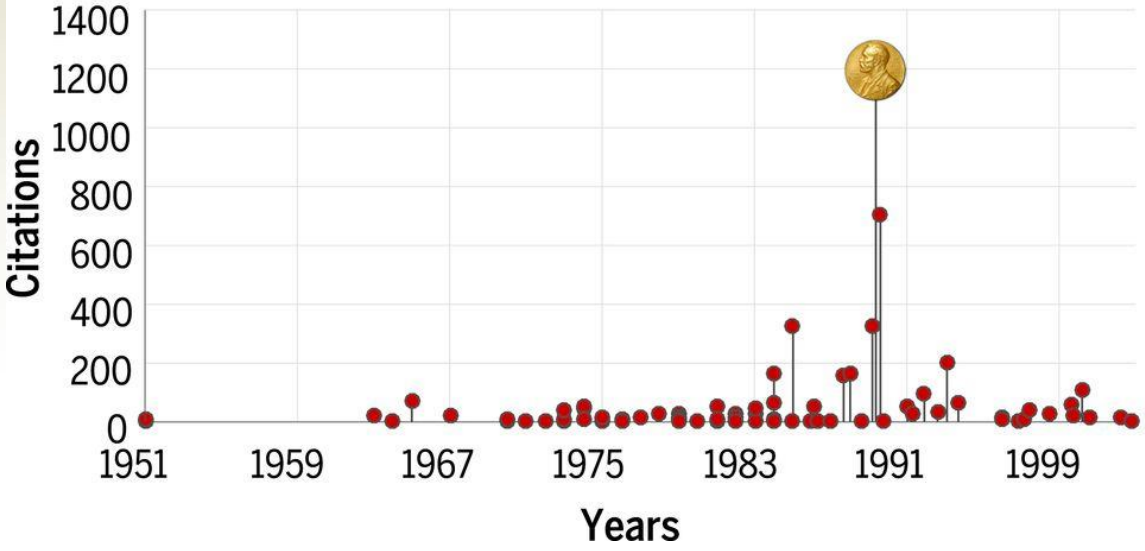
- **Statistical Modeling:** fit bibliometric quantities to complex distributions to approximate their evolution mechanism. Random Impact rule (Sinatra)
- **Issues:**
  - Abundance of data required to calculate statistical parameters (e.g. exponent)
  - Highly complex models when moving from publication level to author or institutional level

Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, Albert-László Barabási: "Quantifying the evolution of individual scientific impact", Science, 04 Nov 2016.

Random Impact Rule

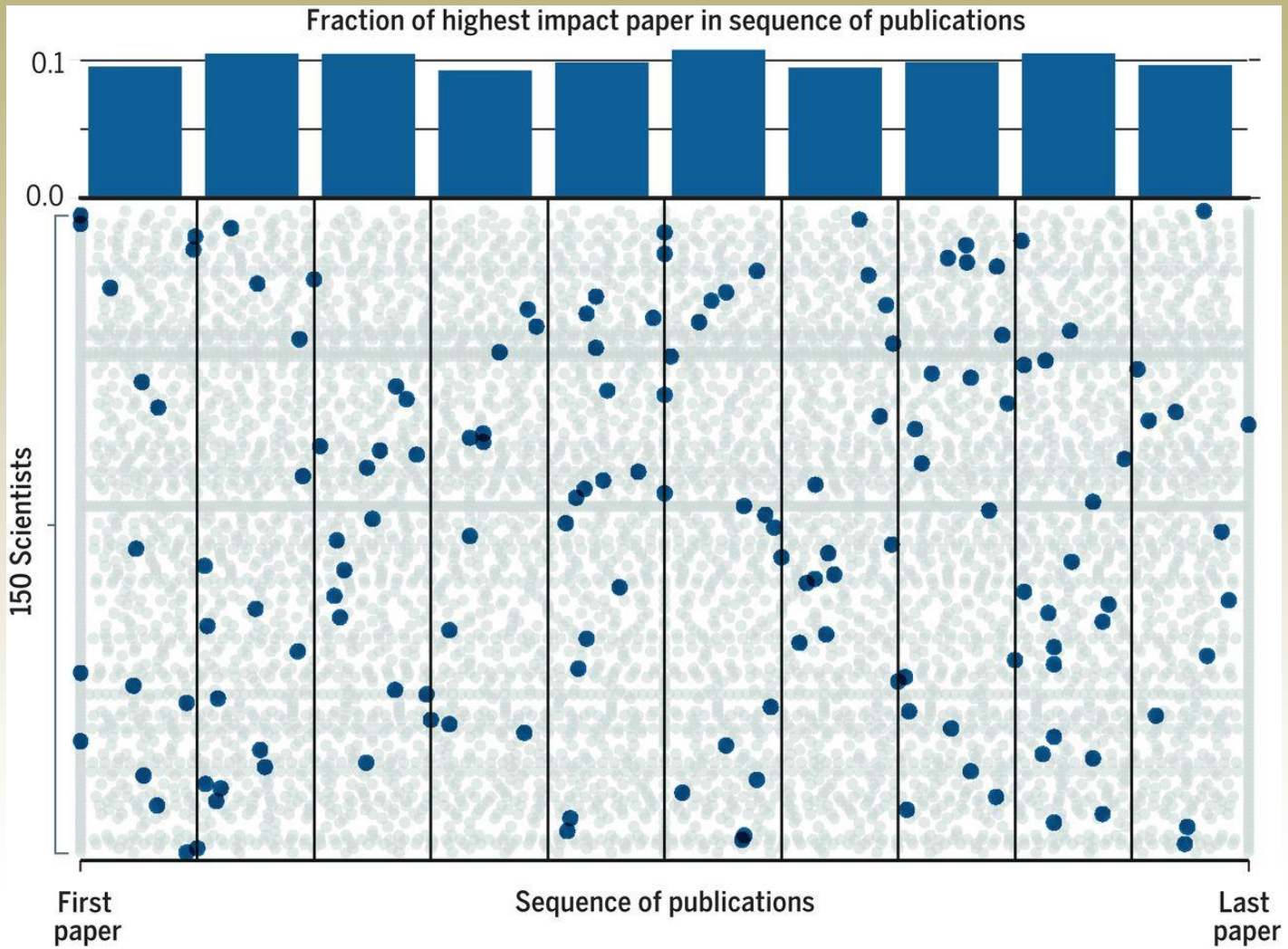


Frank A. Wilczek
Physics Nobel,
2004



John B. Fenn
Chemistry Nobel,
2002

Major discoveries occur at any point



Aaron Clauset, Daniel Larremore, Roberta Sinatra: "Data-driven predictions in the science of science", Science, Vol.355, pp.477-480, Feb 2017.

Love is so short, forgetting is so long

- Previous statistical attempts use *preferential attachment* combined with *randomness*
- Recent proposal* combines preferential attachment with *time decay* inspired by:

- Neruda's "Poema 20" "…love was short and intense, whereas forgetting lingered"

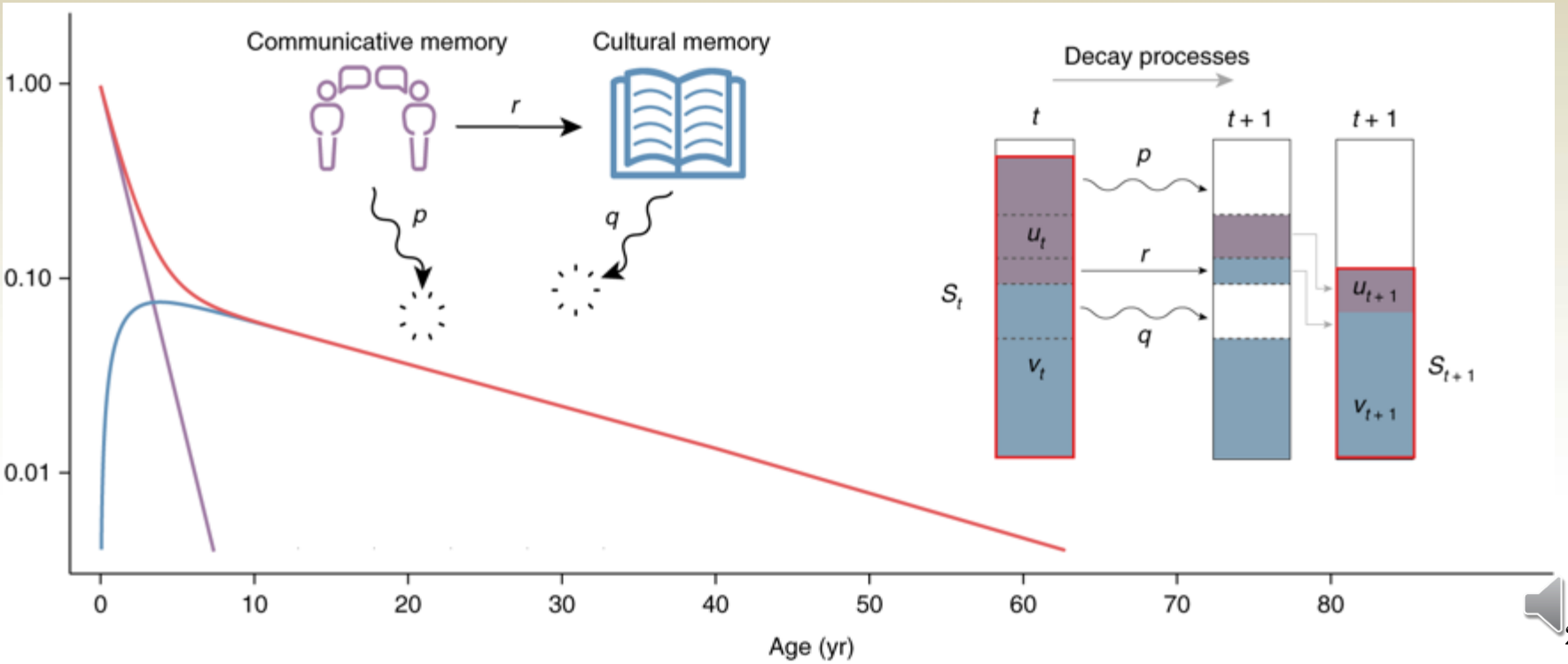
which leads to the exploitation of:

- ***Communicative*** and ***Cultural*** memory

Cristian Candia, C. Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási & César A. Hidalgo, "The universal decay of collective memory and attention", Natural Human Behavior, 3, 82–91 2019.

Communicative & Cultural memory

- The attention follows a decay:
  - a short-lived and fast-decaying phase connected to communicative memory, and
  - a longer-lived and slower-decaying phase connected to cultural memory

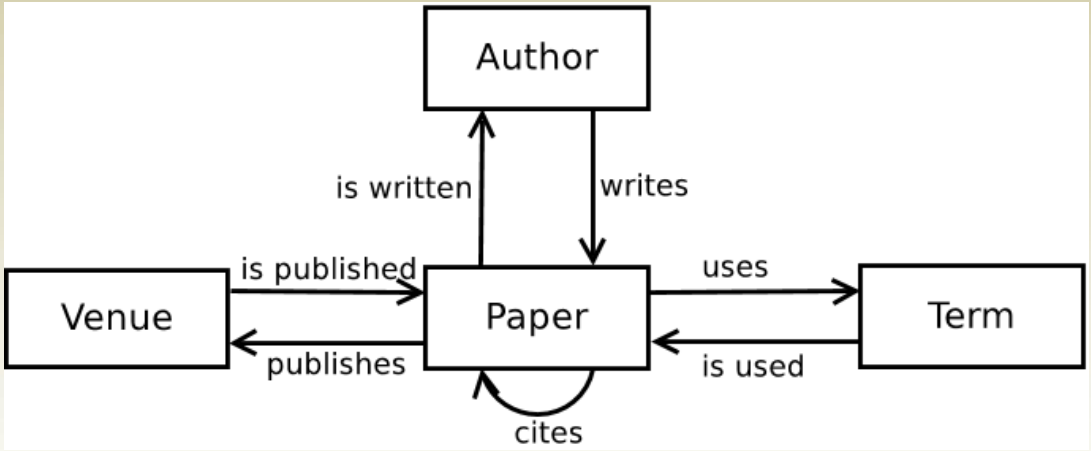


Modeling approaches (6)

- **Network approaches:** citation, co-citation, co-authorship, like link prediction
- **Time Series prediction:** ignore the multiple factors defining future output
- **Deep learning approaches**: Recurrent NN (RNN), Long Short-Term Memory (LSTM)
- **Combination of approaches**: gaining popularity recently

Factors of impact prediction (1)

- Research impact is a result of a **complex interplay** of various factors formulating interconnected networks



- For each case (age, field, level of maturity, country, etc.) the weight assigned to each factor may differ

Factors of impact prediction (2)

Categorization of factors characterizing scientific impact and its evolution modeled as features

Feature origin	Features
author-centric	popularity and productivity
	academic age and gender
	affiliation and academic position
	rank based on bibliometric indices
	disciplines/domains
publication-centric	popularity and age
	number of references and keywords
	topic allocation and number of authors
	relative ranking in portfolio or field
content-related	popularity of topic and novelty
	positioning of references and author ordering
	diversity of topic and MeSH terms
	abstract content
venue-centric	Journal Impact Factor (JIF)
	popularity of publishing house
	ranking in international lists (JCR, Scimago, etc.)
	number of issues/volumes
	open access
social	collaboration network
	citation network
	co-citation/bibliographic coupling
temporal	differences from past state
	normalization over time frame
	rate of activity within a time frame

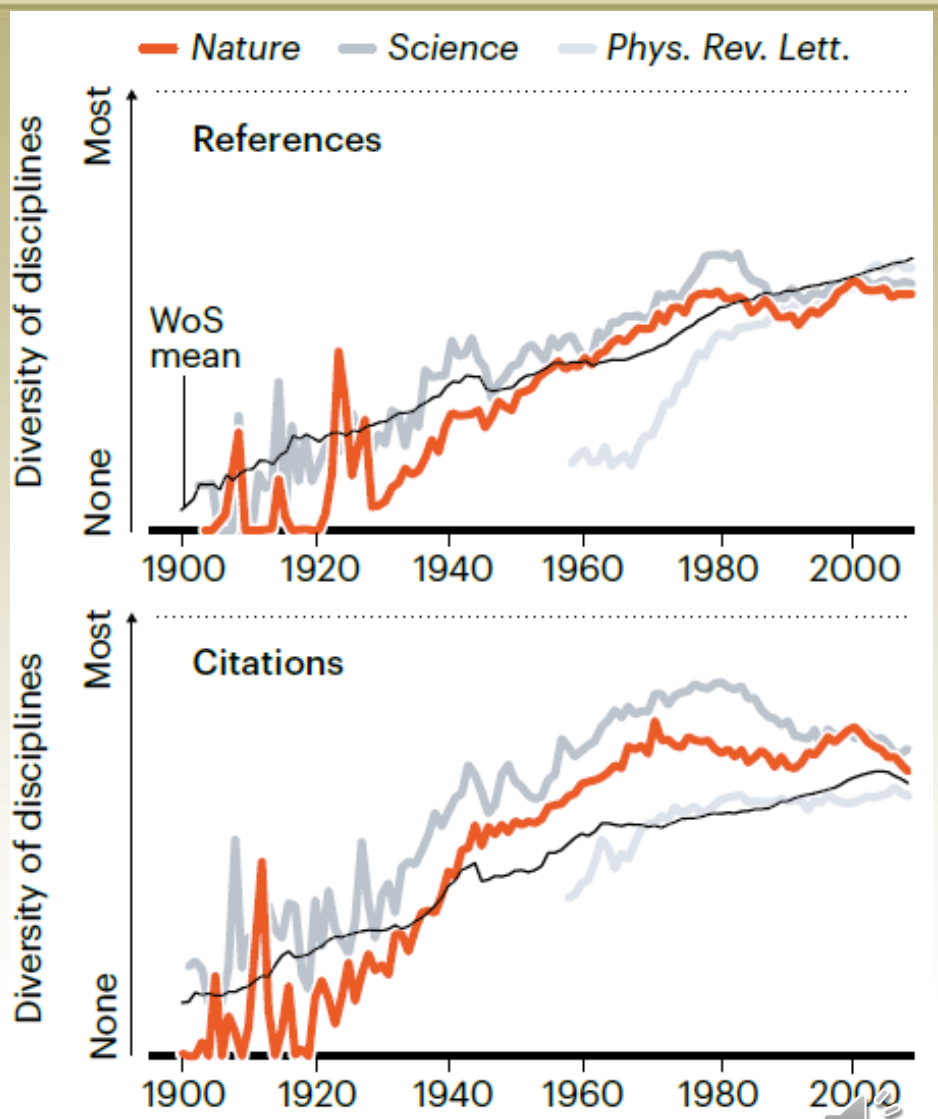
Factors of impact prediction (3)

- **Multidisciplinarity:**
  - Breadth of inspiration (references) correlates to breadth of impact (citations)
    - True for journals such as Cell and Physical Review Letters
    - A typical journal today publishes articles inspired by and impacting about six disciplines
    - Nature and Science both have a greater breadth of impact (citations) and inspiration (references) than 99.7% of other journals

Alexander J. Gates, Qing Ke, Onur Varol and Albert-László Barabási, "Nature's reach: narrow work has broad impact", Nature. 2019 Nov; 575(7781):32-34.

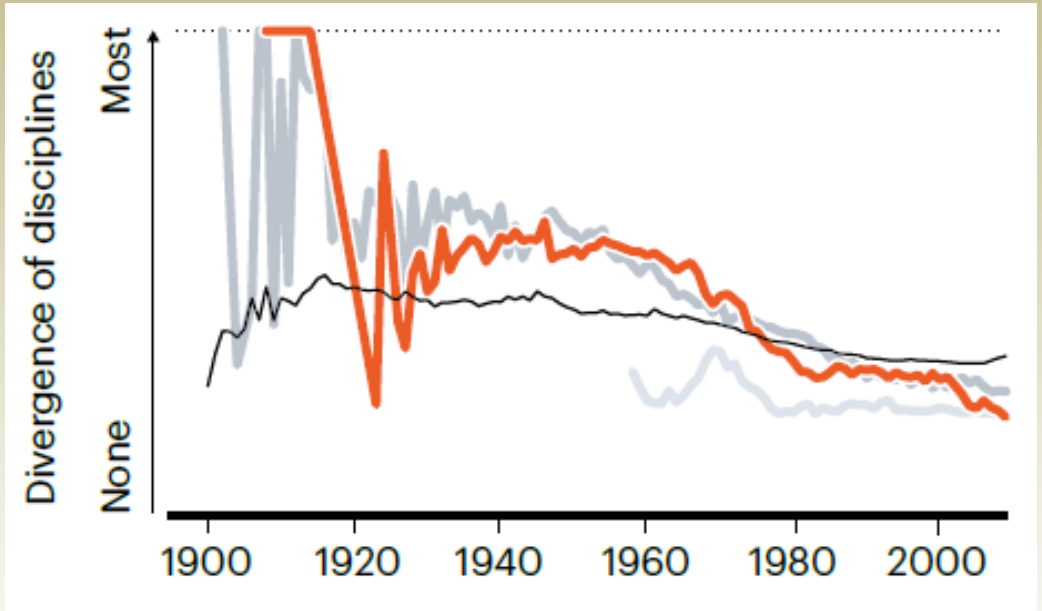
Inspiration and Impact: Interdisciplinarity

- How many, how diverse and how balanced disciplines are across an article's references and citations
- <u>This is growing across all of science</u>



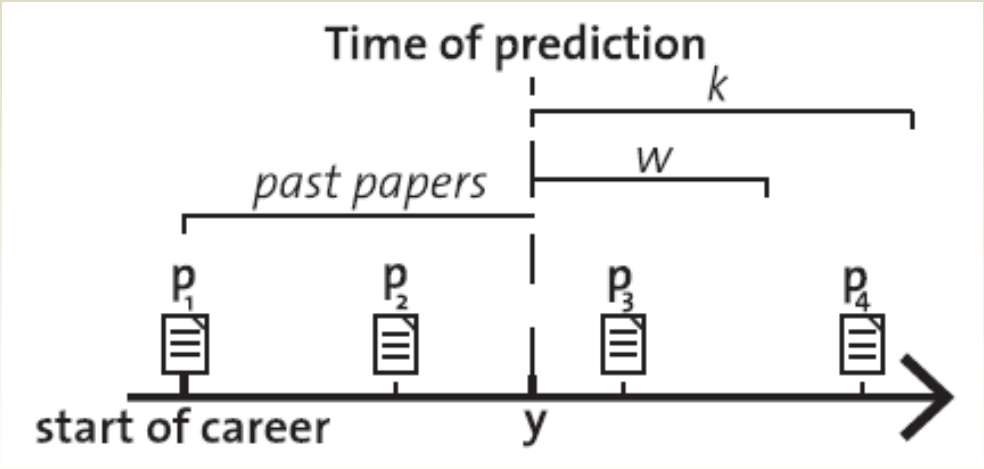
Inspiration and Impact: Cross-disciplinarity

- How much the disciplines in articles' references vary from those in their citations
- <u>The decline here is probably due to rising interdisciplinarity</u>



Challenges

- Focusing on measurable quantities, e.g. citations to past papers, timing of discoveries or promotions, awards, etc.
  - ➢Is it fair and meaningful?



Difference between predicting citations to existing papers and predicting occurrence of new publications

"Rich getting richer"

- **Preferential attachment:** the majority achieves low scores in these metrics, with a selected few attracting significant attention
- Inert property of citation based metrics to be non-decreasing *creates false self-fulfilling predictive models*
- Focus on other performance metrics

Timing is everything (?)

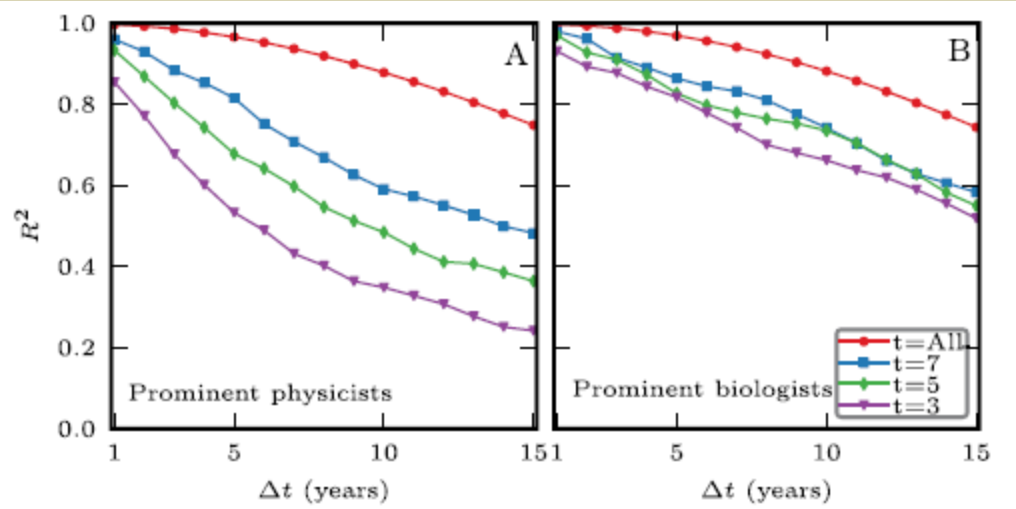
- Timing vs. magnitude of one's impact
  - Varying heterogeneous patterns, e.g. "*sleeping beauties*"
  - How to interpret an *abrupt boost in citations and how long will it last*?
  - *What obsolescence means in different fields*?
  - Younger researchers are found to be more productive but mature ones have a broader outreach

Are we really predicting impact?

- Predicting highly cited scholars or publications is often a different problem from identifying the truly **innovative** ones
- *Publications that conform to the mainstream within a field get cited more often than novel original works*

Time window?

- Long-term vs. short-term prediction
  - ➢ *Is short-term decisive of the whole career*?
  - ➢ Same factors at play in different time windows?



The "predictive power" of a regression model of the h-index across disciplines and career age cohorts (years since first publication t=3,5,7)

Future research directions ?

- Incorporation of **context-specific** data
- Integration with online presence and social media dissemination (***Altmetrics***)
- Cross-referencing of data from various online sources to mitigate **database bias**
- ***Unified*** framework across all levels and patterns with combination of approaches and varying input

Epilogue - Morals

- Avoid oversimplification which makes generalization harder.
- Hard to create an accepted ground truth dataset for model evaluation
- Do not encourage cheating statistics over the progress of science
- Do not focus on individual metrics; aim for overall fair adjustable models
- Combine data intelligence with human insight to decipher science dynamics

Science is a social dynamic versatile process

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, Ismail Rafols, "The Leiden Manifesto for Research Metrics", Nature, 22 April 2015.

Lab contribution (1)

- Sidiropoulos, Katsaros, Manolopoulos: “Generalized h-index for Disclosing Latent Facts in Citation Networks”, *Scientometrics*, Vol.72, No.2, 253-280, 2007.
- Sidiropoulos, Manolopoulos: “Generalized Comparison of Graph-based Ranking Algorithms for Publications and Authors”, *Journal of Systems & Software*, Vol.79, No.12, pp. 1679-1700, 2006.
- Sidiropoulos, Manolopoulos Y.: “A Citation-based System to Assist Prize Awarding”, *ACM SIGMOD Record*, Vol.34, No.4, pp.54-60, 2005.
- Sidiropoulos, Manolopoulos: “A New Perspective to Automatically Rank Scientific Conferences using Digital Libraries”, *Information Processing & Management*, Vol.41, No.2, pp.289-312, 2005.

Lab contribution (2)

- Gogoglou, Manolopoulos: “A Data-driven Unified Framework for Predicting the Evolution of Citation Dynamics”, *IEEE Transactions on Big Data*, accepted.
- Stoupas, Sidiropoulos, Gogoglou, Katsaros, Manolopoulos: “Rainbow Ranking: An Adaptable, Multidimensional Ranking Method for Publication Sets”, *Scientometrics*, Vol.116, No.1, pp.147-160, 2018.
- Gogoglou, Sidiropoulos, Katsaros, Manolopoulos: “The Fractal Dimension of a Citation Curve: Quantifying an Individual's Scientific Output Using the Geometry of the Entire Curve”, *Scientometrics*, Vol.111, No.3, pp.1751-1774, 2017.
- Sidiropoulos, Gogoglou, Katsaros, Manolopoulos: “Gazing at the Skyline for Star Scientists”, *Informetrics*, Vol.10, No.3, pp.789-813, 2016.
- Sidiropoulos, Katsaros, Manolopoulos: “Ranking and Identifying Influential Scientists vs. Mass Producers by the Perfectionism Index”, *Scientometrics*, Vol.103, No.1, pp.1-31, 2015.

Lab contribution (3)

- **A book to be published by Springer**
- **Editors: Katsaros D. and Manolopoulos Y.**
- **Title "Predicting the Dynamics of Research Impact"**
- **Topics**
  - Part I. Impact prediction: Citation curve modelling, Citation (of papers/authors) prediction, Impact indicators evolution prediction, Rising star scientists prediction
  - Part II. Citation network topology prediction: Publication number prediction, Missing link prediction, Models of citation network growth and their use in prediction, Multi-layer (heterogeneous) networks and impact prediction
  - Part III. Case studies to scientific disciplines: Award-winning researchers' prediction, Recommendations for co-authorship, citation, papers, Systems for (big) scholarly data
- **Email: manolopo@csd.auth.gr if interested in contributing**

Thank you!