Quantifying the evolution of individual scientific impact

Zirui Yan

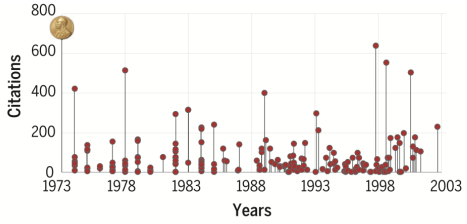
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute

November 13, 2023

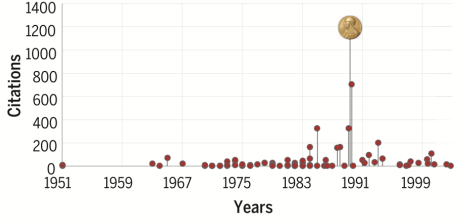
Contents

- R. Sinatra, et al. Quantifying the evolution of individual scientific impact. Science, 2016.
  1. How does impact involve in a career?
  2. Who is going to have an outstanding achievement?
  3. And when?
- Further experiments
  1. Impact of career time.
  2. Impact of publication field.
  3. Improving the accuracy.

Publication history of two Nobel laureates



Frank A. Wilczek
Physics Nobel,
2004



John B. Fenn
Chemistry Nobel,
2002

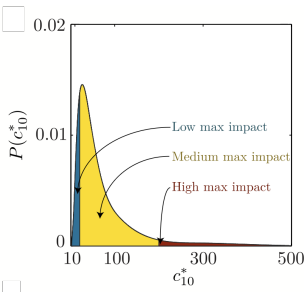
Problem

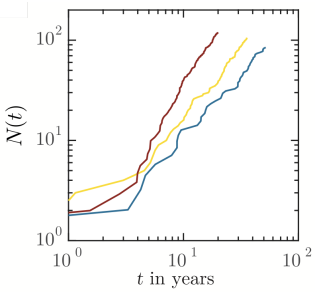
- How do impact and productivity change over a scientific career?
- Does impact follow predictable patterns?
- Can we predict the timing of a scientist's outstanding achievement?
- Can we model scientific careers in quantitative and predictive terms?

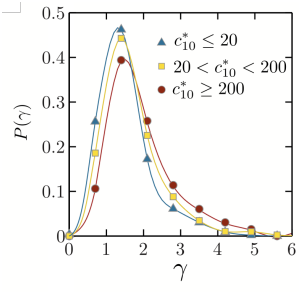
Dataset

- American Physical Society (APS) dataset
  - journal family *Physical Review*
  - 20 years of career + 10 papers + at least one paper every 5 years.
  - $500,000$ papers over 110 years
  - 3000 careers
- Impact of paper: Cumulative citations over 10 years c_{10}.

Distribution of max impact and productivity

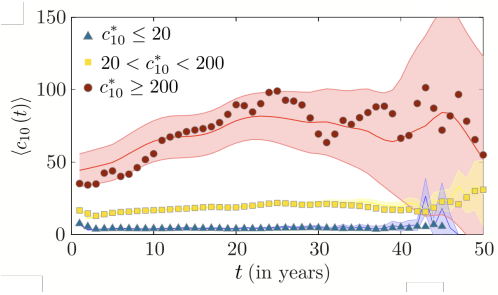


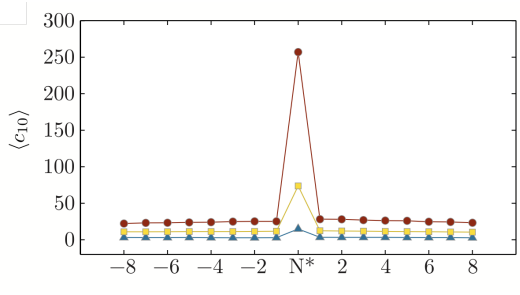




- Split into 3 groups: Low/Medium/High max impact.
- More products are expected from a high max impact group.
- Number of publication $N(t) = t^{\gamma}$.

Citation varying with time



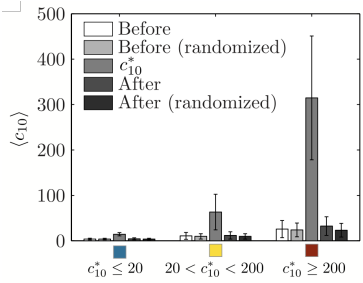


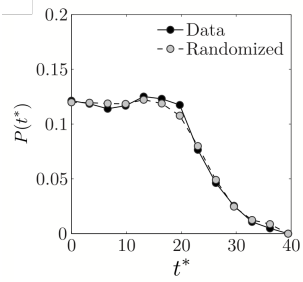
- High max impact group will constantly have a higher impact.
- High max impact happens randomly during the career.

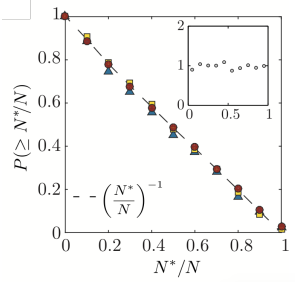
Random-impact rule

- Keep the publication time and citation number.

- Reshuffle the publication index.







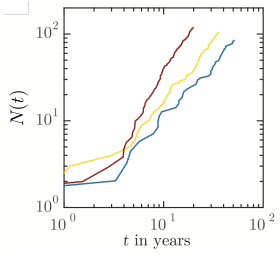
Random-impact rule

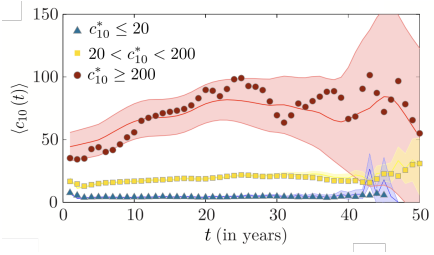
- Impact is random in a career.

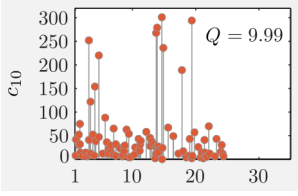
- There is always hope! If you keep publishing!

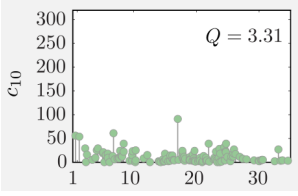
The role of scientist

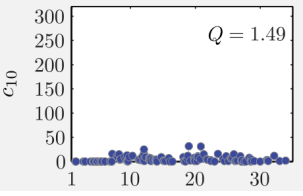
- There is systematic differences in impact between careers.











Models

- **Q-model:** Impact of a paper j by scientist i is

$$c_{10,ij} = p_j \, Q_i \ . \tag{1}$$

$$\text{Impact of j-th paper} = \text{lucky} * Q$$

- **Baseline: R-model:**

$$c_{10,j} = p_j \ , \tag{2}$$

where $p_j \sim P(c_{10})$.

  - The only factor differentiating two scientists is their overall productivity N.

Estimation in Q model

- The joint probability $P(p, Q, N)$ is verified to be log-normal.

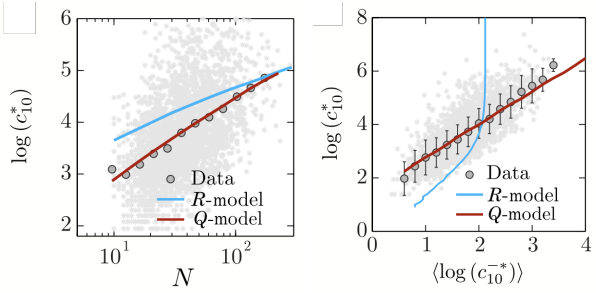
- Maximum-likelihood approach.

$$\mu = (\mu_p, \mu_Q, \mu_N) = (0.92, 0.93, 3.34) \tag{3}$$

$$\begin{aligned} \sum &= \begin{pmatrix} \sigma_p^2 & \sigma_{p,Q} & \sigma_{p,N} \\ \sigma_{p,Q} & \sigma_Q^2 & \sigma_{Q,N} \\ \sigma_{p,N} & \sigma_{Q,N} & \sigma_N^2 \end{pmatrix} \\ &= \begin{pmatrix} 0.93 & 0.00 & 0.00 \\ 0.00 & 0.21 & 0.09 \\ 0.00 & 0.09 & 0.33 \end{pmatrix} \end{aligned} \tag{4}$$

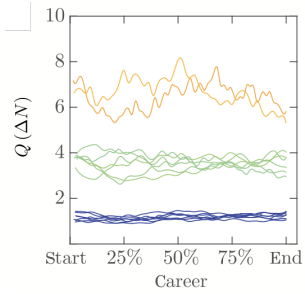
- $\sigma_{p,Q} = \sigma_{p,N} = 0$.

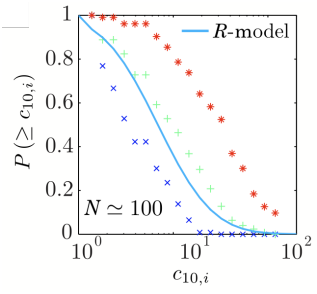
Goodness of model

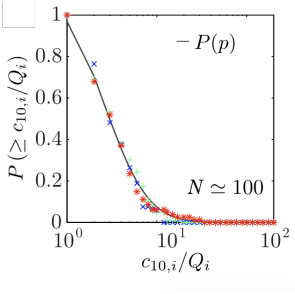


- R-model can not capture the correlation between c_{10}^* and N.
- R-model can not capture the correlation between c_{10}^* and c_{10}^{-*}, average citation exclude the most cited paper.
- Q-model is a good fit.

Number of highest citation



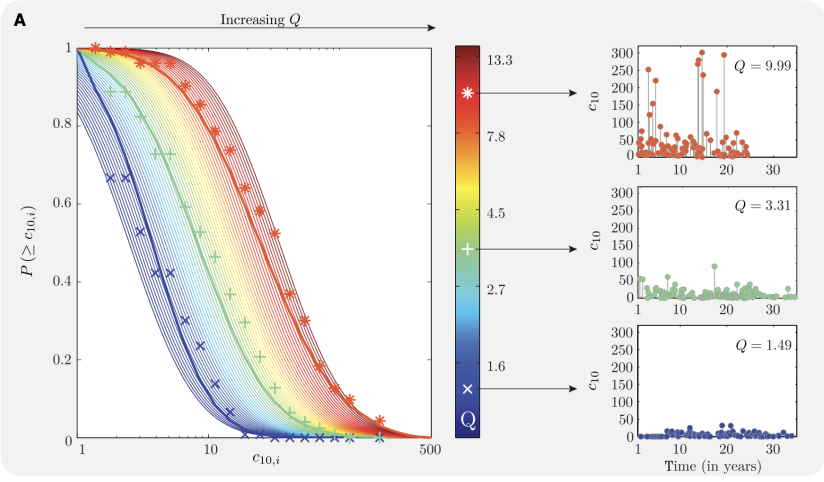




- Sliding widow: Q factor is a "constant" within career.
- Q-model capture the difference between different group.
- p is pure lucky!

Predicting individual Q factor

- calculate Q by maximizing the individual likelihood



Results

- Who is going to have an outstanding achievement?

  Lucky scientists with high Q value.

- And when?

  Randomly within their career.

Further experiment

National Institutes of Health Open Citation Collection (NIH-OCC) dataset[1]

- MedLine, PubMed Central (PMC), and CrossRef.
- 20 years of career + 10 papers
- 551274 careers with Publication since 1800
- Impact of paper: Average citations.

Longer history, larger dataset!

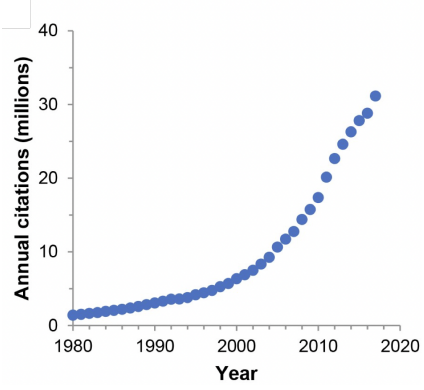
[1] Hutchins, B. Ian, et al. "The NIH Open Citation Collection: A public access, broad coverage resource." PLoS Biology, 2019.

Goal

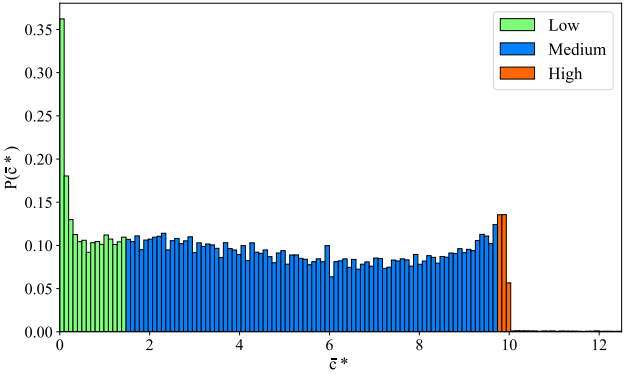
- Impact of career time.
- Impact of publication field.
- Improving the accuracy.

Influence of career time

- The citation number is exploding.

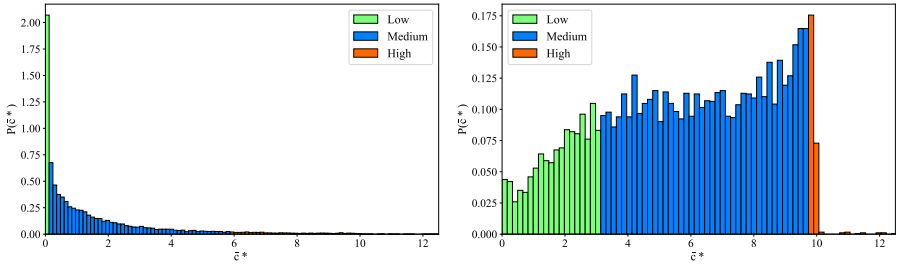


Influence of career time



Influence of career time

- Early: Last publication < 1990 v.s. Late: First publication > 1990.

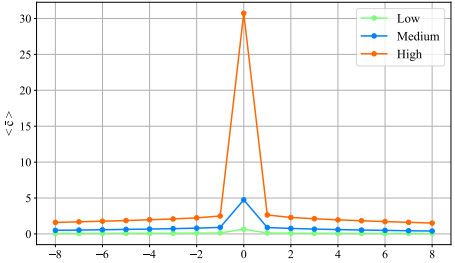


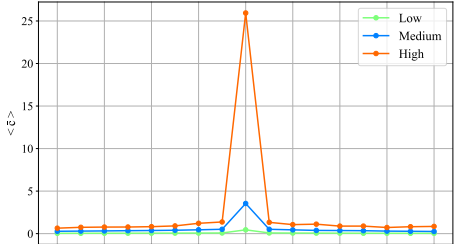
Percentage	20%	95%
all	1.57	9.74
early	0.07	5.98
late	3.15	9.82

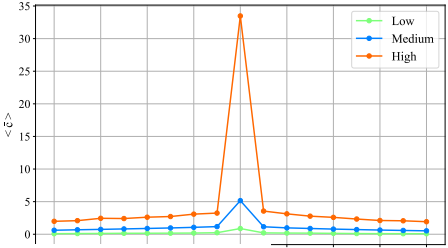
- Career time does impact the average citation.

Influence of career time

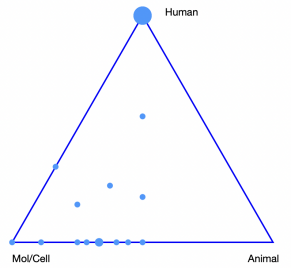
- Random rule still holds.

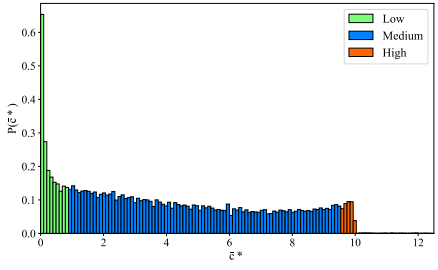


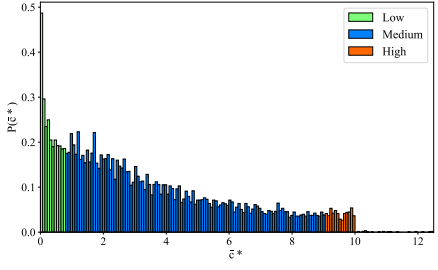


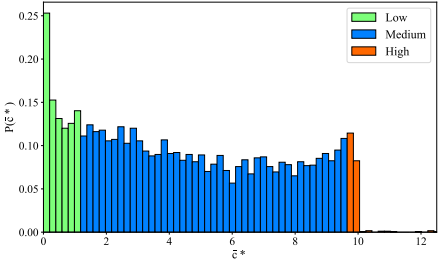


Influence of field









Influence of field

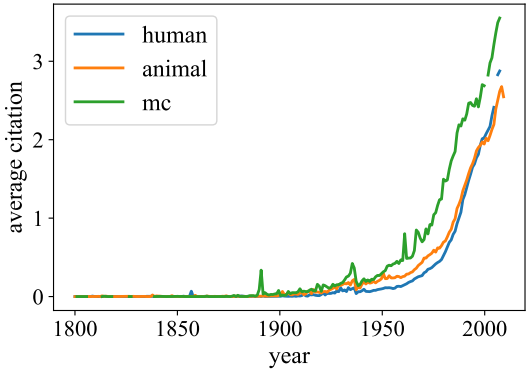
Percentage	20%	95%
all	1.57	9.74
human	0.92	9.62
animal	0.85	9.07
molecular cellular	1.34	9.72

Table: Percentage of average citation

- Field do impact the average citation.

- Better to publish cross-field papers.

Correlation between career time and filed



- Career time and filed are correlated.

Failure of Q-model

- We now have 4 factors: lucky (p), career year (y), filed (f) and scientist power Q.
- The dataset is too large which is computational inefficient
- Correlation between career time (y) and filed (f) such that

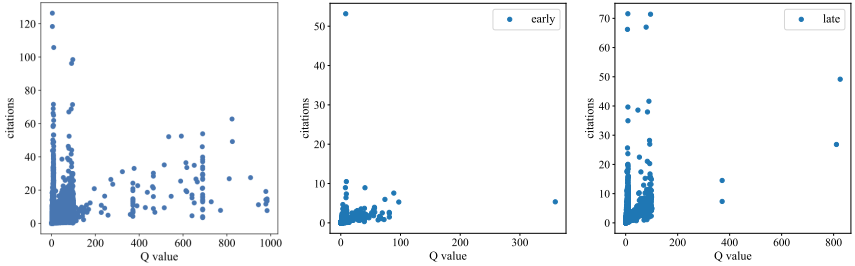
$$\bar{c} \neq p\mu_y fQ \tag{5}$$

- Use neural network to predict the mean of average citation

$$\mu_{\bar{c}} = \mathrm{NN}(y, f) \ . \tag{6}$$

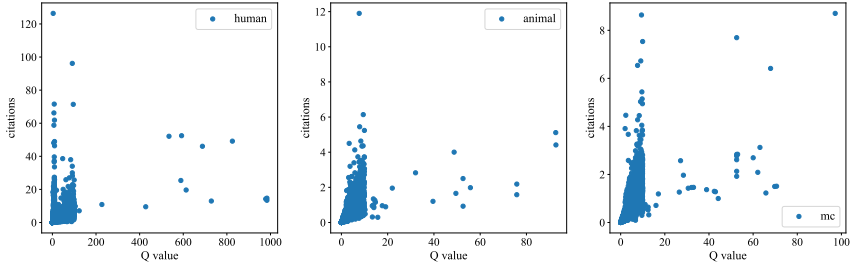
Neural network (preliminary)

- Trained three layer neural network with ReLU activation
- Q_i is average of $\bar{c}_j/\mathrm{NN}(y_j, f_j)$
- the influence of career time



Neural network (preliminary)

- influence of field



Conclusion of additional experiment

- Career time and publication field do impact the average citation.
- Q-model is not capable for this setting.
- Neural network can help enhance the prediction.

Future work

- Improve the prediction results of neural network.
- Rescale number of citations.
- Author name disambiguation.