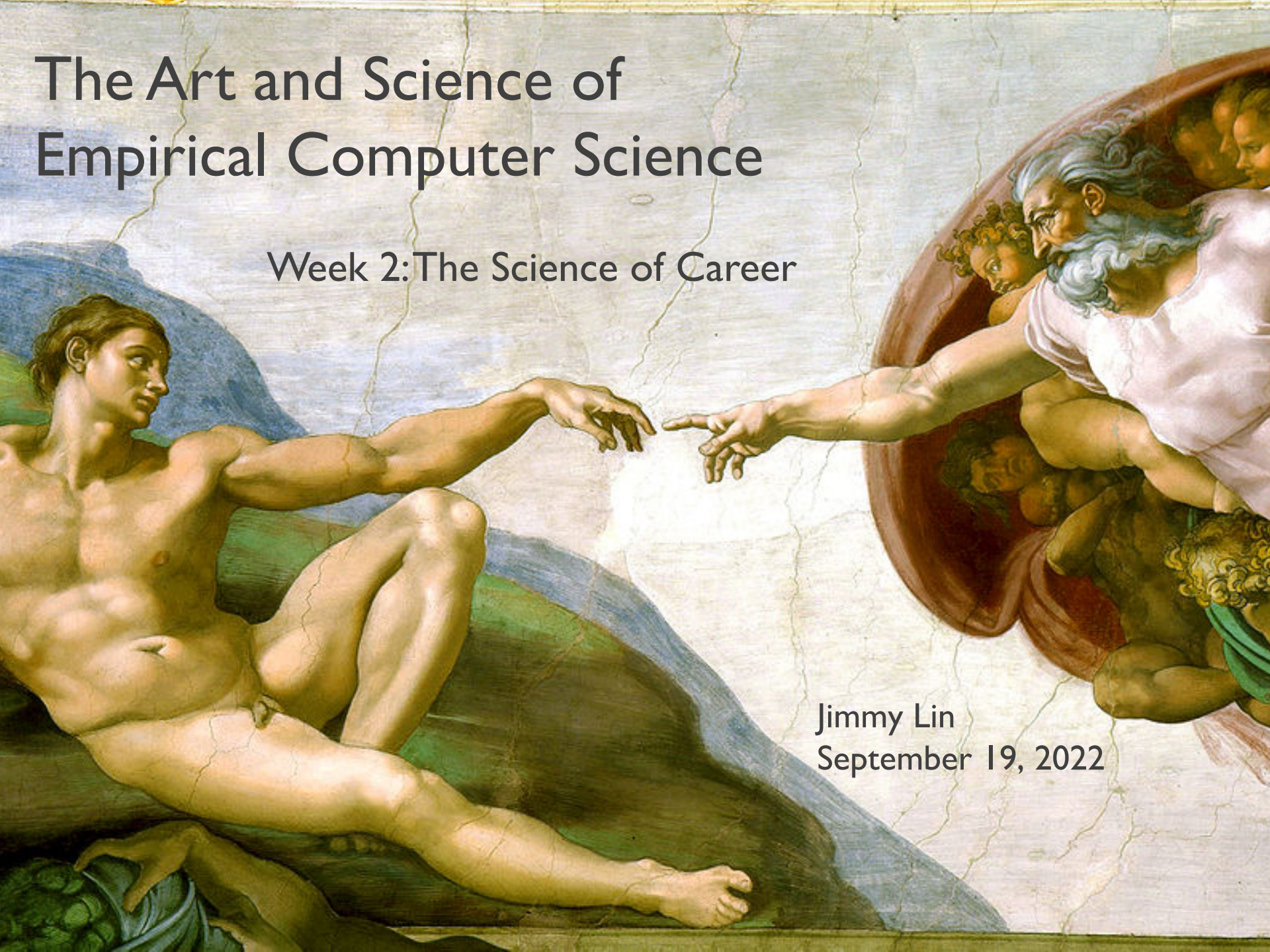
The Art and Science of Empirical Computer Science

Week 2: The Science of Career



Jimmy Lin
September 19, 2022

your *graduate student* career
your *academic* career
your *scientific* career
your career

{ understanding, planning, making decisions about }

Course Description

Graduate students in computer science aspire to "do computer science" (research), but what exactly does that mean? It involves, among a multitude of activities, reading papers, learning the "state of the art", advancing knowledge, writing papers, and (hopefully) getting them published. Graduate students learn how to do these things under tutelage of professors, but rarely is there explicit or deliberate instruction on these myriad activities. With a focus on empirical computer science, this course covers elements that comprise the research enterprise, synthesizing both "art" — personal experiences I have accumulated over the years — as well as "science" — insights derived from quantitative analyses. The hope is that knowledge and actionable advice from this course will help graduate students better understand research, hopefully leading to more productive and fulfilling careers.

your *graduate student* career
your *academic* career
your *scientific* career
your career

understanding → characterizing

Characterizing scientific careers

… in terms of products?

… in terms of what else?

Quantity vs. Quality
Quantitative vs. Qualitative

How should we evaluate excellence?

Position A: Researchers should be evaluated *solely* on the quality of their publications. Quantity is irrelevant and we shouldn't even bother counting.

Position B: Researchers should be evaluated on *both* the quality and quantity of their publications. High-quality publications are of course important, but quantity is also an important component of excellence.

Interesting Data Points

Albert Einstein: 248 papers

Peter Higgs: 25 papers (by 84)

Paul Erdös: 1475 papers

Let’s debate!

“The Science of Science”, Part I
The Science of Career

What’s it about?

1. Quantitative characterizations.
2. Correlational analysis.
3. Proposed explanatory models.

Metrics should be used for everything.

Metrics convey *some* signal.

Metrics are useless and should be ignored.

"If you can not measure it, you can not improve it." – William Thompson (aka Lord Kelvin)

"All models are wrong, but some are useful."
– George Box

"When a measure becomes a target, it ceases to be a good measure." – Goodhart's Law

Metrics convey *some* signal

How much signal?
For what purpose?
Correlation vs. Causation
Normative vs. Positive

"The Science of Science", Part I

The Science of Career

Chapter 1: Productivity of a Scientist

Discussion Points

Definition of productivity
Venues
The growing number of scientists

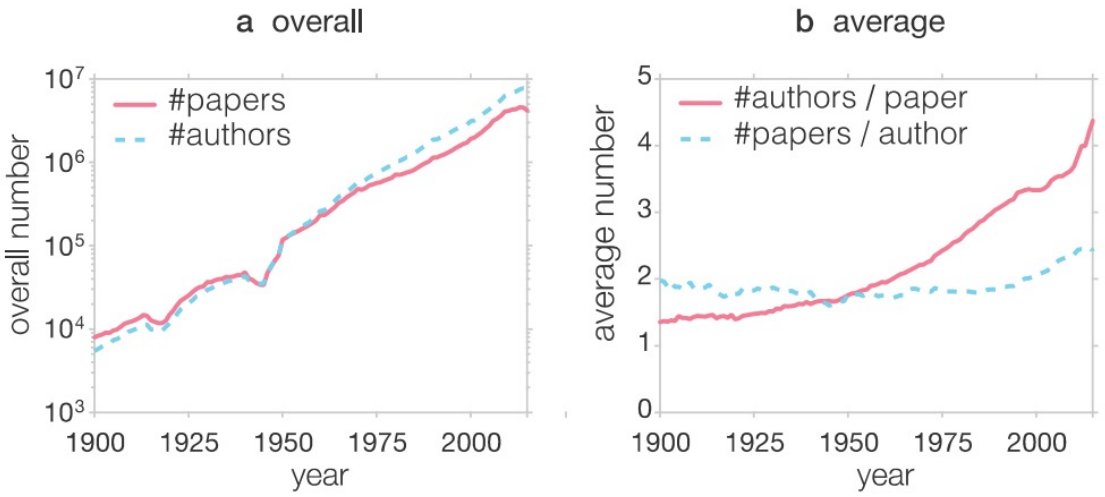


Figure 1.1.1 **The growing number of scientists**. (a) During the past century, both the number of scientists and the number of papers has increased at an exponential rate. (b) The number of papers co-authored by each scientist has been hovering around two during the past 100 years, and increased gradually in the past 15 years. This growth is a direct consequence of collaborative effects: Individual productivity is boosted as scientists end up on many more papers as co-authors. Similar trends were reported using data within a single field [5]. For physics, for example, the number of papers co-authored by each physicist has been less than one during the past 100 years, but increased sharply in the past 15 years. After Dong *et al*. [4] and Sinatra *et al*. [5].

Shockley's Model

$$N \sim p_1\, p_2\, p_3\, p_4\, p_5\, p_6\, p_7\, p_8$$

F_1. Identify a good problem
F_2. Make progress with it
F_3. Recognize a worthwhile result
F_4. Decide when to stop the research and start writing up the results
F_5. Write adequately
F_6. Profit constructively from criticism
F_7. Show determination to submit the paper for publication
F_8. Make changes if required by the journal or the referees

"The Science of Science", Part I

The Science of Career

Chapter 2: The *h*-index

Discussion Points

Kinda like democracy?

Correlation vs. Causation

"The Science of Science", Part I

The Science of Career

Chapter 3: The Matthew Effect

Discussion Points

It's real.

Normative vs. Positive

"The Science of Science", Part I

The Science of Career

Chapter 4: Age and Scientific Achievement

Discussion Points

Burden of knowledge

Conceptual vs. Experimental

"The Science of Science", Part I

The Science of Career

Chapter 5: Random Impact Rule

Discussion Points

Definition of impact

Implications?

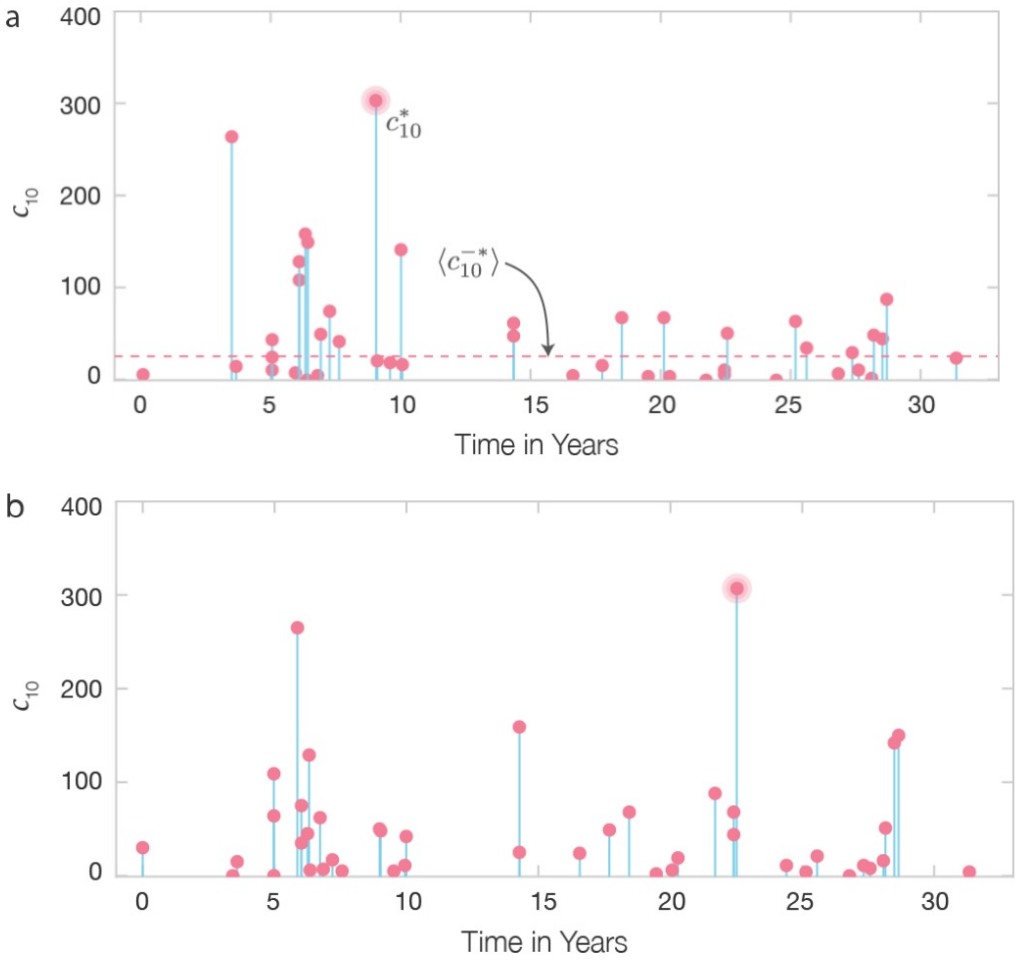


Figure 1.5.1 **Publication history of Kenneth G. Wilson**. (a) The horizontal axis indicates the number of years after Wilson's first publication and each vertical line corresponds to a publication. The height of each line corresponds to c_{10}, *i.e.* the number of citations the paper received after 10 years. Wilson's highest impact paper was published in 1974, 9 years after his first publication; it is the 17th of his 48 papers, hence t^{*}=9, N^{*}=17, N = 48. (b) Shuffled career of Wilson, where we keep the locations of the pins, but swap the impact of each paper with another one, thereby breaking the temporal ordering of when best work occurs within a career. After Sinatra *et al.* [116].

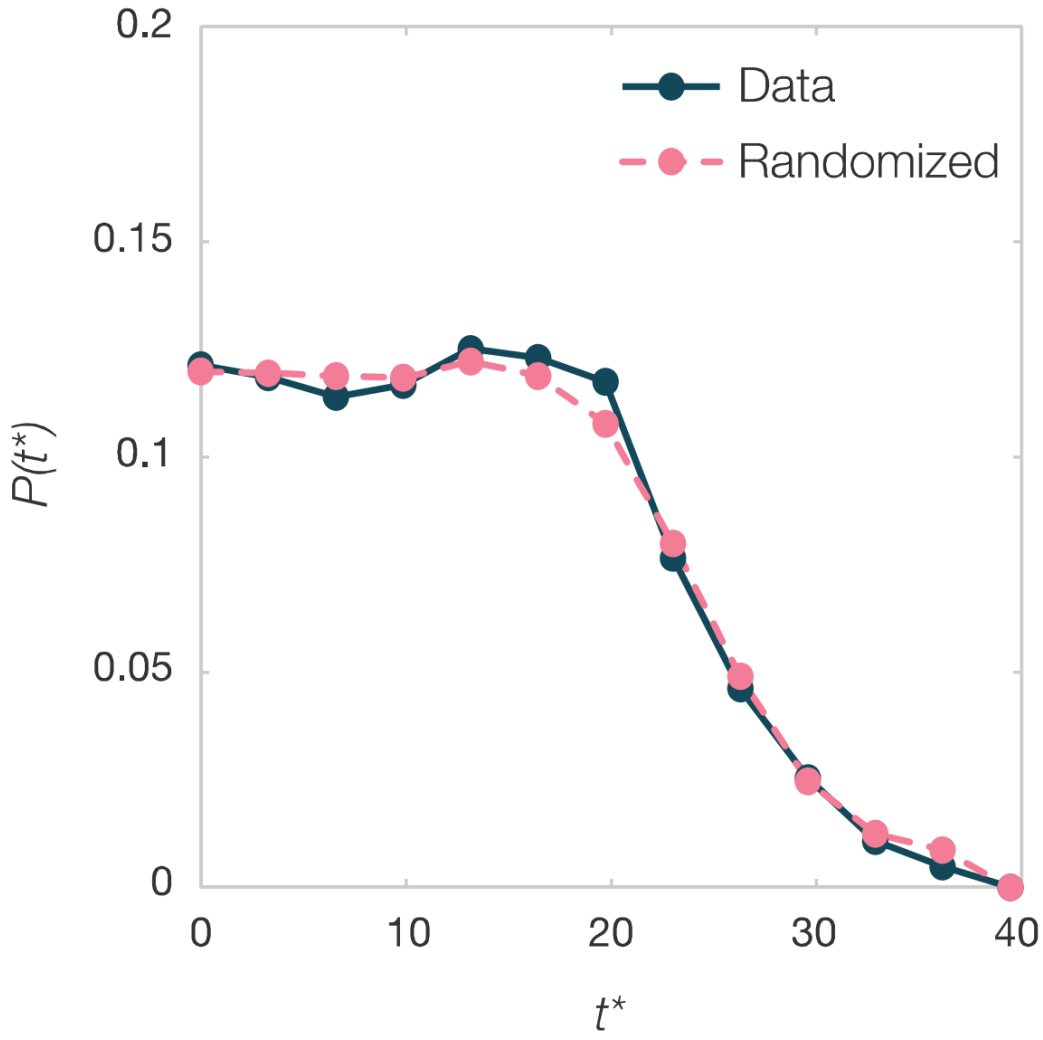


Figure 1.5.2 **The Random Impact Rule.** Distribution of the publication time t^* of the highest impact paper in scientists' career (black circles) and for randomized impact careers (black circles). The lack of differences between the two curves indicates that impact is random within a scientist's sequence of publication. After Sinatra *et al.* [116].

“The Science of Science”, Part I
The Science of Career
Chapter 6: The *Q*-Factor

Discussion Points

R-model vs. *Q*-model
Q-factor stability 🙁

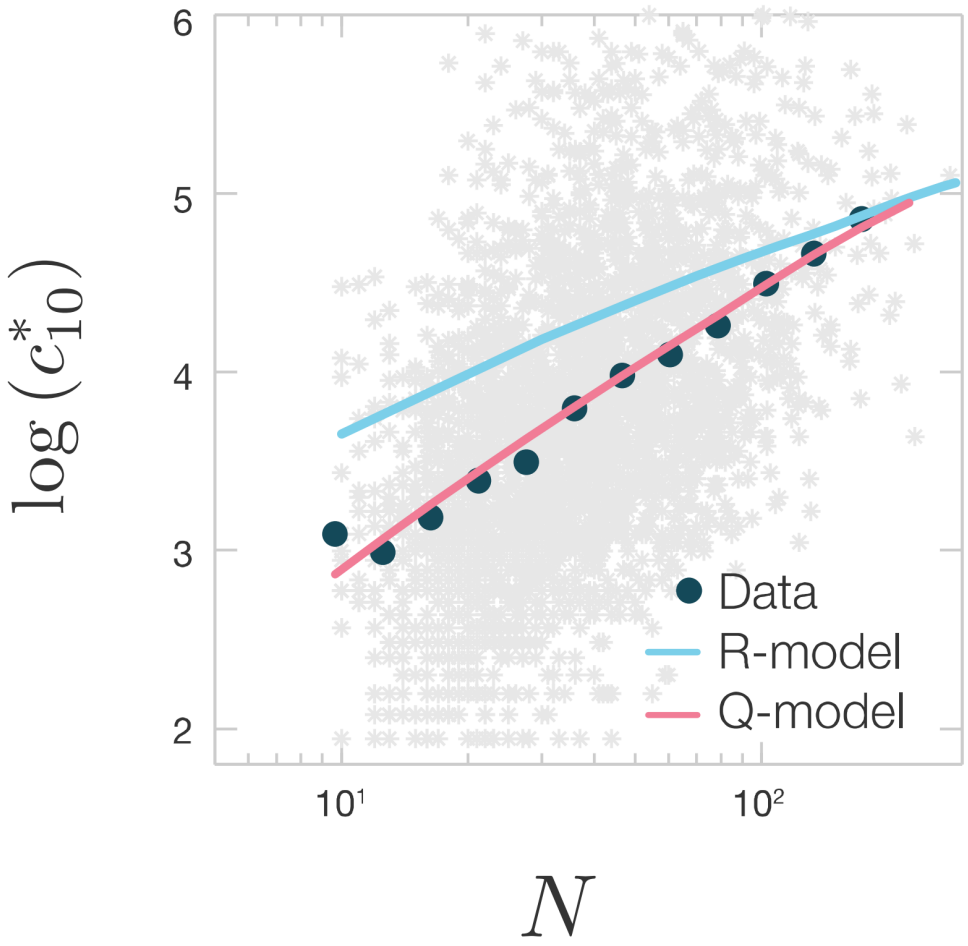


Figure 1.6.1 **Scientific careers are not random**. Scatter plot shows the citation of the highest impact paper, c_{10}^* vs the number of publications N during a scientist's career. Each grey dot corresponds to a scientist. The circles are the logarithmic binning of the scattered data. The cyan curve represents the prediction of the R-model, which shows systematic deviation from data. The red curve corresponds to the analytical prediction of the Q model. After Sinatra *et al.* [116].

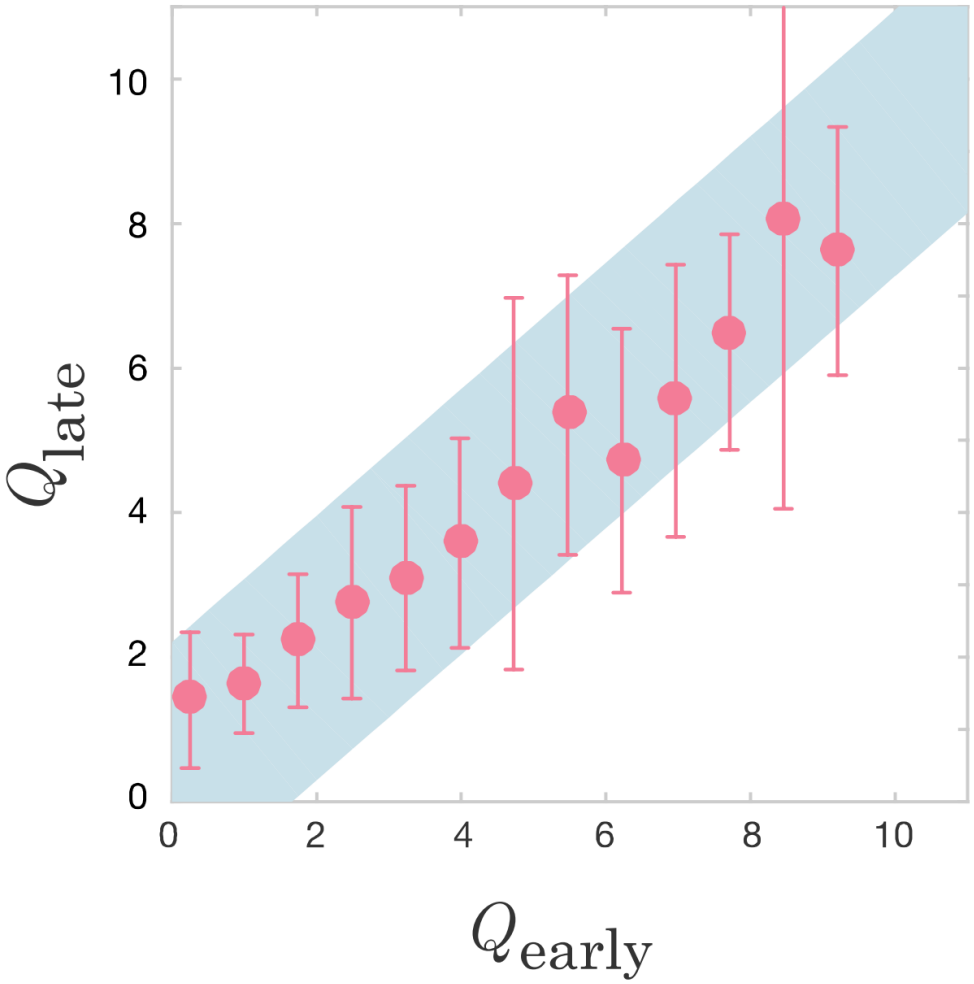


Figure 1.6.3 **The *Q*-factor appears relatively stable during a career**. We compare the Q parameter at early-career (Q_{early}) and late-career (Q_{late}) stage of 823 scientists with at least 50 papers. We measured the two values of the Q parameters using only the first and second half of published papers, respectively. We perform these measurements on the real data (circles) and on randomized careers, where the order of papers is shuffled (gray shaded areas). For most of the careers (95.1%), the changes between early- and late-career stages fall within the fluctuations predicted by the shuffled careers, suggesting that the Q parameter is relatively stable throughout a career.

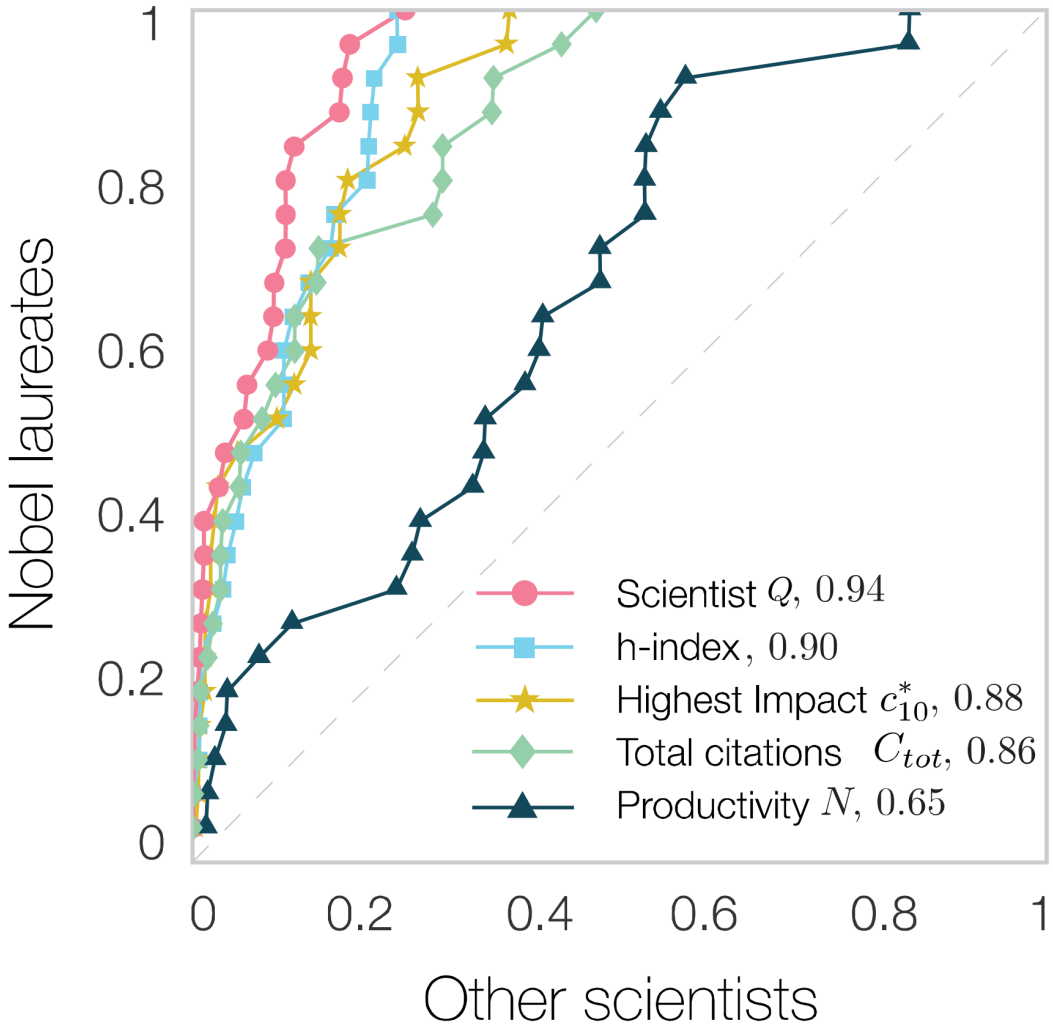


Figure 1.6.4 **Predicting Nobel Laureates**. ROC plot captures the ranking of scientists based on Q, productivity N, total number of citations C, citations of the highest-impact paper c_{10}^{*} and h-index. Each curve represents the fraction of Nobel laureates versus the fraction of other scientists for a given rank threshold. The diagonal (no-discrimination line) corresponds to random ranking; the area under each curve measures the accuracy to rank Nobel laureates (reported in the legend, with 1 being the maximum). After Sinatra *et al.* [116].

“The Science of Science”, Part I

The Science of Career

Chapter 7: Hot Streaks

Discussion Points

Contradiction?

tl;dr – keep trying?

That’s all for this week!