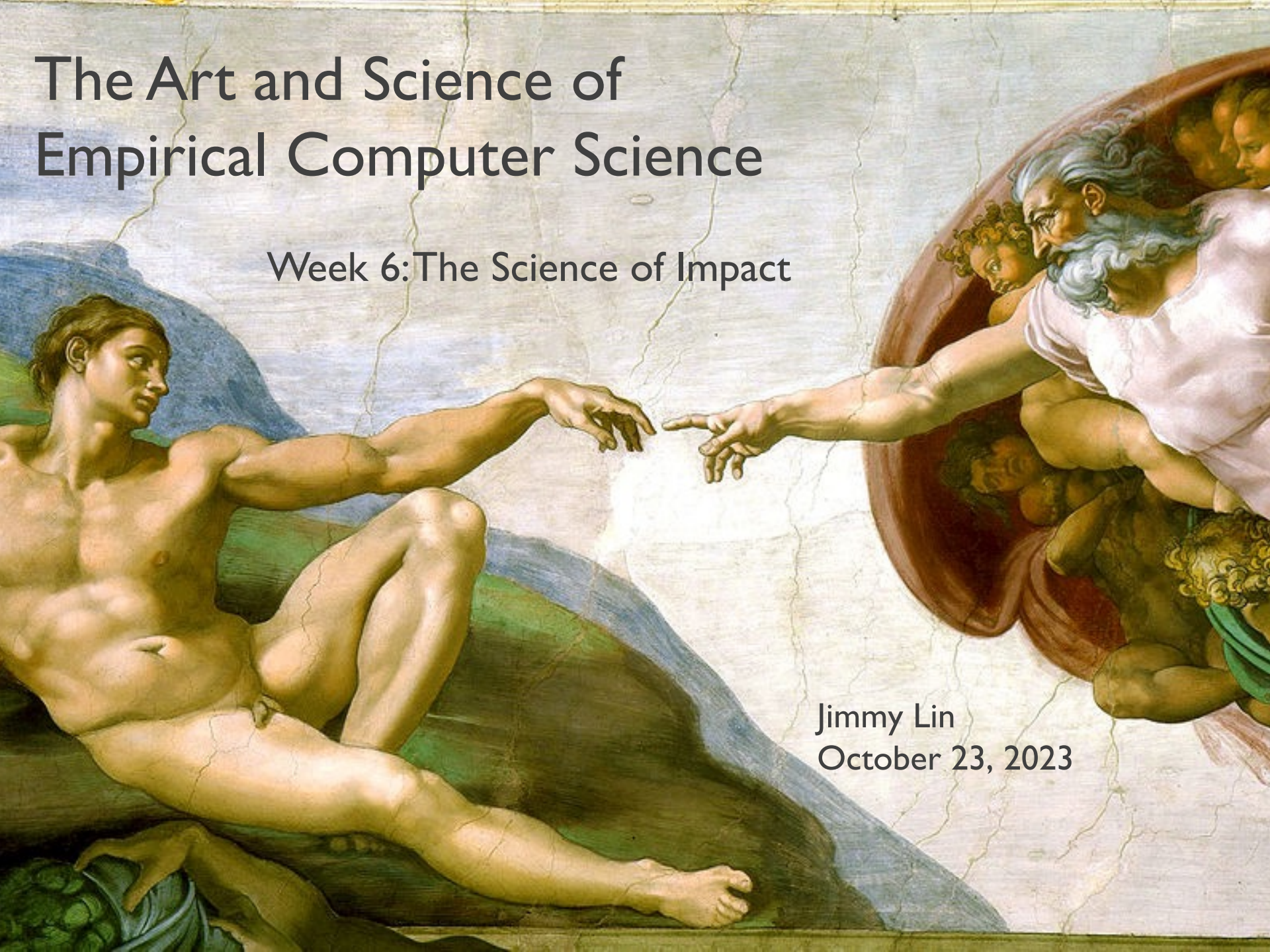# The Art and Science of Empirical Computer Science

Week 6: The Science of Impact

Jimmy Lin
October 23, 2023

Social Media: Waste of Time?

Position A: Early-stage researchers should actively incorporate social media use as a component of their career development. This means appropriate use of sites like Twitter, Facebook, and LinkedIn to build professional reputation, engage with the community, hear about recent work by others, etc.

Position B: Early-stage researchers should stay off social media. It's a complete waste of time.

Let's debate!

How to measure impact?

Let's measure it in terms of citations

Price (1965): Distribution of citation follows a power law
$$P(c) \sim c^{-\gamma}$$

a

Many papers with a few citations

p_c

Few papers with a many citations
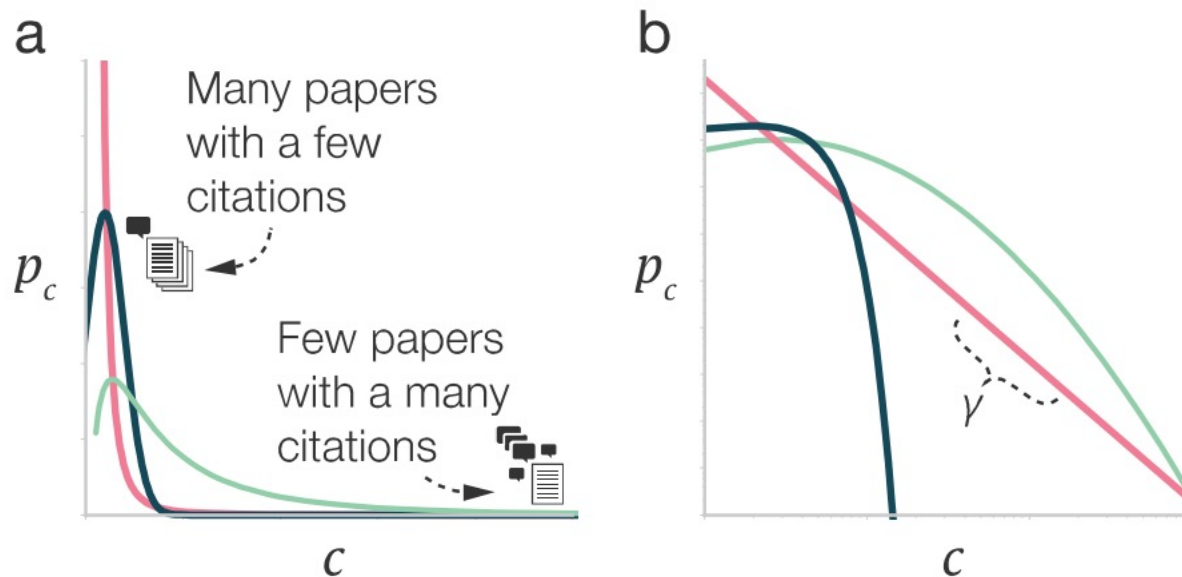
c

b

p_c

γ

c

Figure 3.2.2 **Illustrating normal, power-law, and lognormal distributions**. (a) Comparing a power law and a lognormal function to a normal distribution on a linear-linear plot. (b) The same comparison shown on a log-log plot, helping us see the fundamental difference between the normal and fat-tailed distributions in the high citation regime. A power law follows a straight line on a log-log plot, its slope indicating the citation exponent γ. Sometimes, it can be difficult to tell lognormal and power laws apart, as they appear similar on a log-log plot.
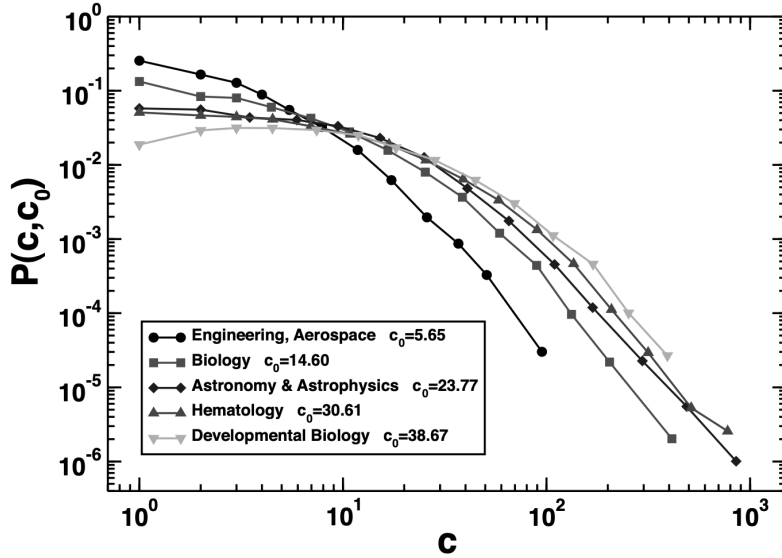
Fig. 1. Normalized histogram of the number of articles $P(c, c_0)$ published in 1999 and having received c citations. We plot $P(c, c_0)$ for several scientific disciplines with different average number c_0 of citations per article.
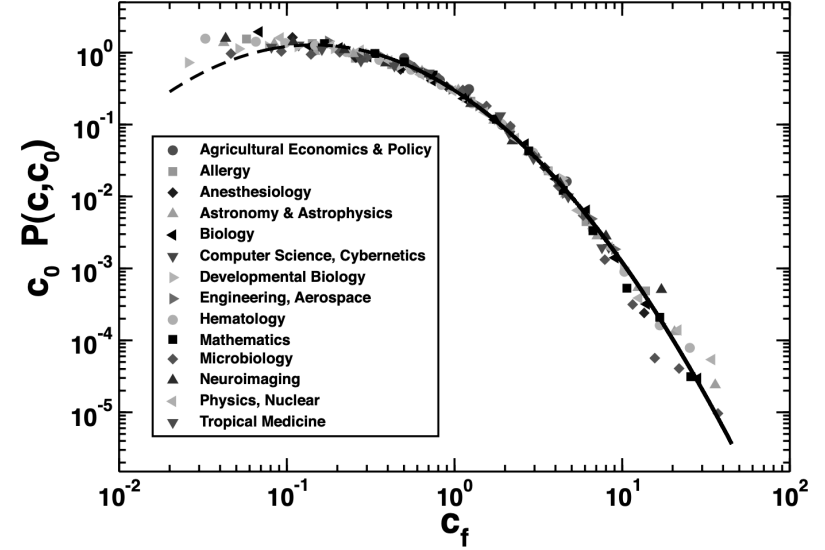

Fig. 2. Rescaled probability distribution $c_0 P(c, c_0)$ of the relative indicator $c_f = c/c_0$, showing that the universal scaling holds for all scientific disciplines considered (see Table 1). The dashed line is a lognormal fit with $\sigma^2 = 1.3$.

Radicchi et al. Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, 105(45):17268-17272, 2008.

How to measure impact?

Let's measure it in terms of citations

Okay, let's model it

Price Model (1976)

1. **The growth of the scientific literature.** New papers are continuously published, each of which cite a certain number of previous papers.
2. **Preferential attachment.** The probability that an author chooses a particular paper to cite is not uniform, but proportional to how many citations the paper already has

Bianconi–Barabási Model (2001)

1. **Growth:** In each time step, a new paper i with a certain number of references and fitness η_i is published, where η_i is a random number chosen from a distribution $p(\eta)$. Once assigned, the paper's fitness does not change over time.

2. **Preferential Attachment:** The probability that the new paper cites an existing paper is proportional to the product of paper i's previous citations and its fitness η_i.

Wang–Song–Barabási Model (2013)

1. **The exponential growth of science.** In order for papers to gain new citations, new papers must be published; hence, the rate at which these new papers are published affects how existing papers will accumulate citations.
2. **Preferential attachment** captures the fact that highly cited papers are more visible and thus more likely than their less-cited counterparts to be cited again.
3. **Fitness** captures the inherent differences between papers, accounting for the perceived novelty and importance of a discovery.
4. **Aging** captures how new ideas are integrated into subsequent work: Every paper's propensity for citations eventually fades, in a fashion best described by a log-normal survival probability.

	λ	μ	σ
(solid blue)	1	7	1
(dashed green)	2	7	1
(dashed red)	1	9	1
(dashed cyan)	1	7	0.5
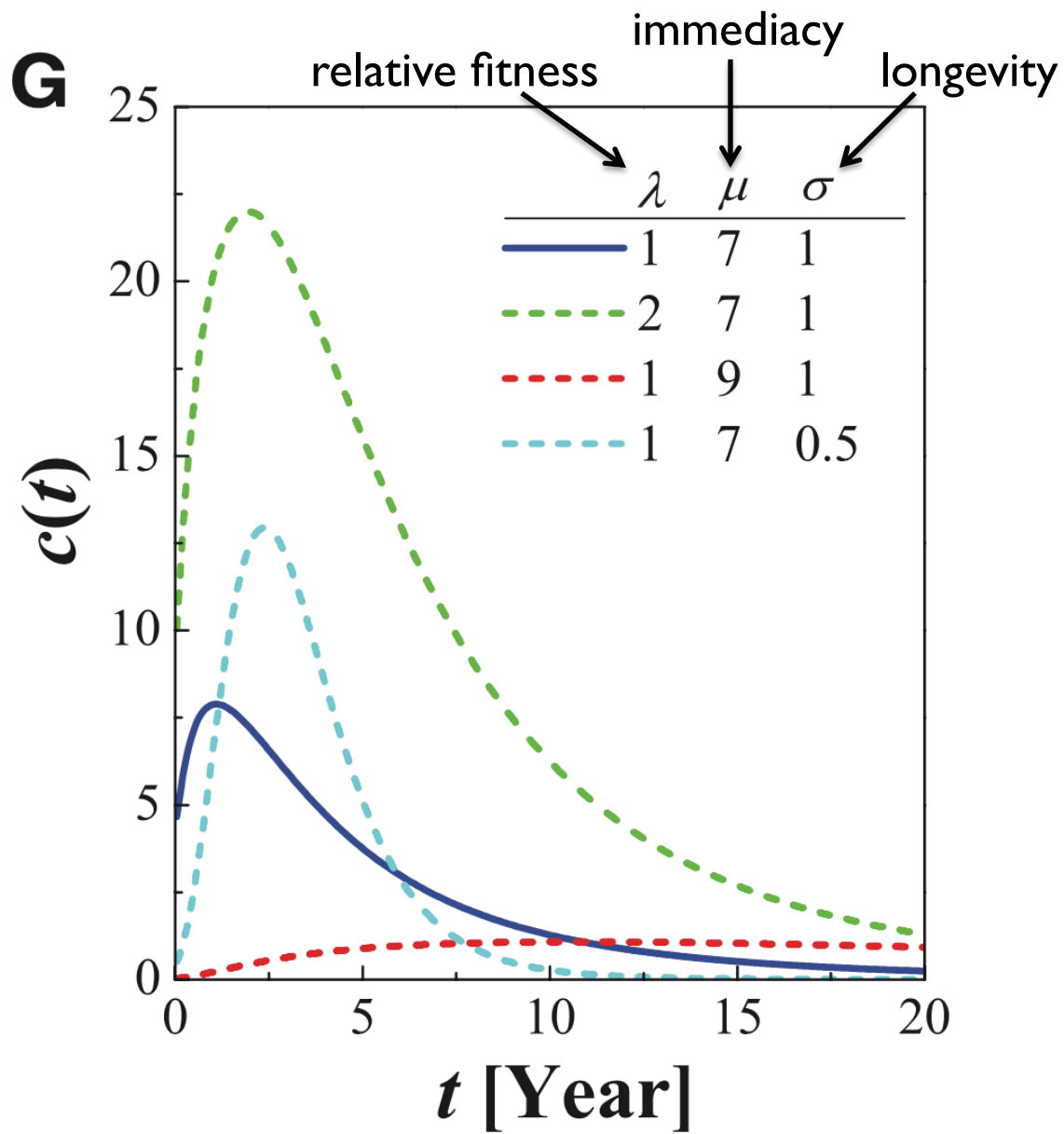
How to measure impact?

Let's measure it in terms of citations
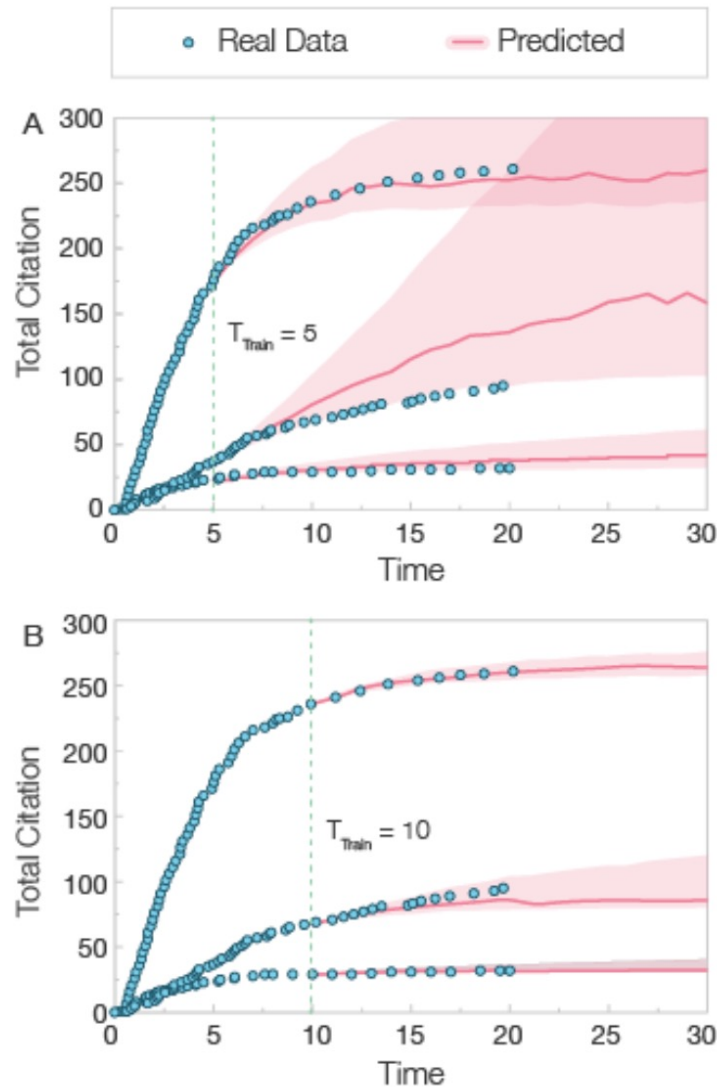Okay, let's model it
Who cares?

Figure 3.6.7 **Predicting future citations**. We can adapt the citation model to predict future citations of any given paper, by learning the model parameters from its existing citation histories. The uncertainty in estimating these parameters then translate into a prediction envelop with its most likely trajectory, similar to what we saw in the hurricane's example. As we observe more of its citation records, the envelop shrinks, and more accurately encapsulates a paper's citation history. After Wang *et al.* [66].

What do citation counts not capture?

Unfairly benefits surveys
"Negative citations"
"Perfunctory"

Category	Description	Example
Criticizing	Criticism can be positive or negative. A citing sentence is classified as "criticizing" when it mentions the weakness/strengths of the cited approach, negatively/positively criticizes the cited approach, negatively/positively evaluates the cited source.	Chiang (2005) introduced a constituent feature to reward phrases that match a syntactic tree but did not yield significant improvement.
Comparison	A citing sentence is classified as "comparison" when it compares or contrasts the work in the cited paper to the author's work. It overlaps with the first category when the citing sentence says one approach is not as good as the other approach. In this case we use the first category.	Our approach permits an alternative to minimum error-rate training (MERT; Och, 2003);
Use	A citing sentence is classified as "use" when the citing paper uses the method, idea or tool of the cited paper.	We perform the MERT training (Och, 2003) to tune the optimal feature weights on the development set.
Substantiating	A citing sentence is classified as "substantiating" when the results, claims of the citing work substantiate, verify the cited paper and support each other.	It was found to produce automated scores, which strongly correlate with human judgements about translation fluency (Papineni et al. , 2002).
Basis	A citing sentence is classified as "basis" when the author uses the cited work as starting point or motivation and extends on the cited work.	Our model is derived from the hidden-markov model for word alignment (Vogel et al., 1996; Och and Ney, 2000).
Neutral (Other)	A citing sentence is classified as "neutral" when it is a neutral description of the cited work or if it doesn't come under any of the above categories.	The solutions of these problems depend heavily on the quality of the word alignment (Och and Ney, 2000).

Table 2: Annotation scheme for citation purpose. Motivated by the work of (Spiegel-Rösing, 1977) and (Teufel et al., 2006)

Intent cateogry	Definition	Example
Background information	The citation states, mentions, or points to the background information giving more context about a problem, concept, approach, topic, or importance of the problem in the field.	Recent evidence suggests that co-occurring alexithymia may explain deficits [12]. Locally high-temperature melting regions can act as permanent termination sites [6-9]. One line of work is focused on changing the objective function (Mao et al., 2016).
Method	Making use of a method, tool, approach or dataset	Fold differences were calculated by a mathematical model described in [4]. We use Orthogonal Initialization (Saxe et al., 2014)
Result comparison	Comparison of the paper's results/findings with the results/findings of other work	Weighted measurements were superior to T2-weighted contrast imaging which was in accordance with former studies [25-27] Similar results to our study were reported in the study of Lee et al (2010).

Table 1: The definition and examples of citation intent categories in our SciCite.
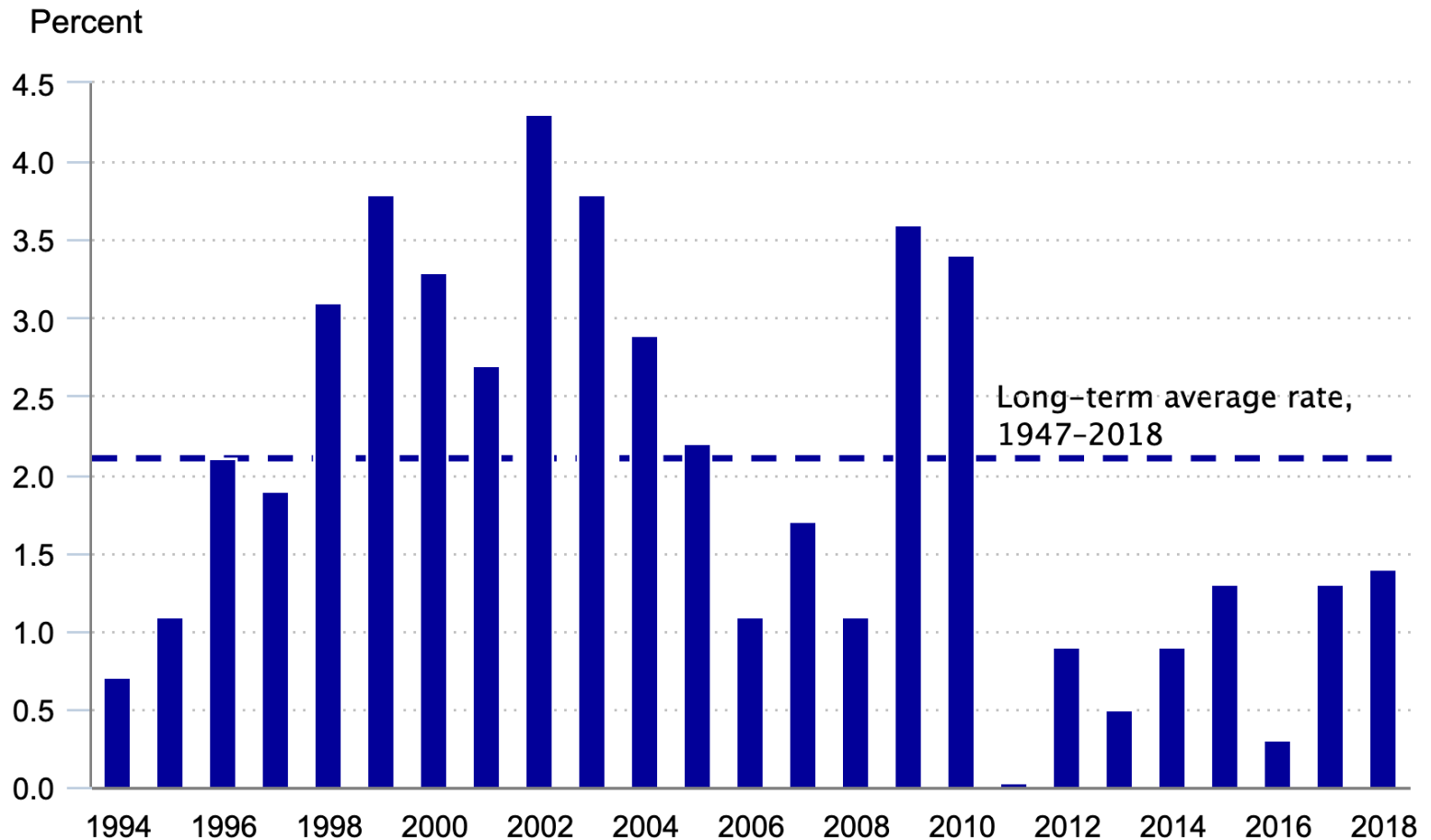
How to measure impact?

Let's measure it in terms of citations

No, let's not.

Then what?

Figure 1. Labor productivity growth: annual percent changes, nonfarm business sector, 1994–2018

Percent

4.5

4.0

3.5

3.0

2.5

Long–term average rate, 1947–2018

2.0

1.5

1.0

0.5

0.0

1994   1996   1998   2000   2002   2004   2006   2008   2010   2012   2014   2016   2018

That's all for this week!