Impact of lexical and sentiment factors on the popularity of scientific papers

JULIAN SIENKIEWICZ* & **EDUARDO G. ALTMANN†**

Max Planck Institute for the Physics of Complex Systems, Dresden

∗ current address: Faculty of Physics, Warsaw University of Technology
† current address: School of Mathematics and Statistics, University of Sydney

9 November 2016

MOTIVATION

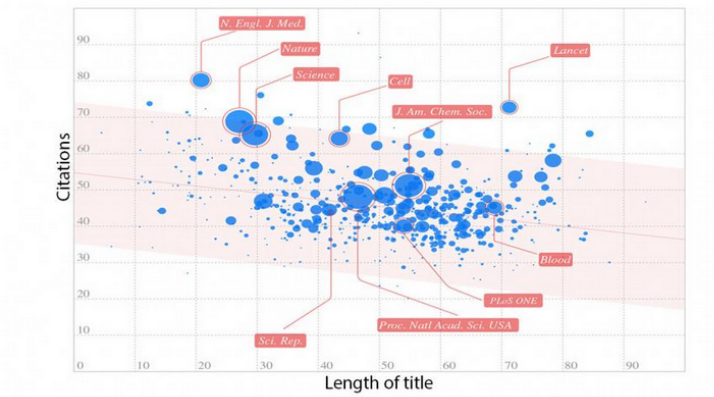
1. Citations for an article can be regarded as a **proxy for the attention** (popularity) the paper achieved in the scientific community.
2. Investigation how **textual properties** of scientific papers affect the adoption and spread of scientific results as quantified by the number of citations.
3. Last but not least: dealing with recent results (e.g., Letchford et al. R Soc Open Sci 2, 150266) showing that is the **negative correlation** between title length and citations (i.e., shorter the titles more citations).

Science AAAS

Home | News | Journals | Topics | Careers

Latest News | ScienceInsider | ScienceShots | Sifter | From the Magazine | About News | Quizzes

SHARE



With a few exceptions, studies published in journals that tend toward shorter paper titles get more citations annually than those published in journals with longer paper titles. A. Letchford et al., R.Soc. Open sci. 2 (2015)

In brief, papers with shorter titles get more citations, study suggests

By **Dalmeet Singh Chawla** | Aug. 25, 2015 , 7:00 PM

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment > News > 2016 > March > Article

NATURE | NEWS

Papers with shorter titles get more citations

Intriguing correlation mined from 140,000 papers.

Boer Deng

26 August 2015

Rights & Permissions

To William Shakespeare, brevity was the soul of wit. For scientists, it may be even more valuable, as conciseness seems to correlate with how frequently a research paper is cited.

Adrian Letchford and his colleagues at the University of Warwick in Coventry, UK, analysed the titles of 140,000 of the most highly cited peer-reviewed papers published between 2007 and 2013 as listed on Scopus, a research-paper database. They compared the lengths of the papers' titles with the number of times each paper was cited by other peer-reviewed papers— a statistic sometimes used as a crude measure of importance.

As they report in *Royal Society Open Science*[1], "journals which publish papers with shorter titles receive more citations per paper".

GOALS

KEY FACTORS

Systematic investigation of how different **textual properties** of scientific papers such as
- text length,
- text complexity,
- sentiment

affect the number of **citations** they acquire.
In this way we want identify **key factors** that influence scientific popularity.

GOALS

KEY FACTORS

Systematic investigation of how different **textual properties** of scientific papers such as
- text length,
- text complexity,
- sentiment

affect the number of **citations** they acquire.
In this way we want identify **key factors** that influence scientific popularity.

CITATION PATTERN DIFFERENCES

We want to show **differences** between **most cited** (top) and **typical** papers using **quantile regression** approach.

DATA

Web of Science service

FILTERING

- papers marked as articles published in the period of **1995—2004**
- papers needed to fulfil two conditions:
  1. journal active in **all** mentioned years (e.g., `PLOS` journals absent)
  2. there had to be at least **1.000 articles** in the given period (e.g., `Rev Mod Phys` absent)

DATA

Web of Science service

FILTERING

- papers marked as articles published in the period of **1995—2004**
- papers needed to fulfil two conditions:
  1. journal active in **all** mentioned years (e.g., `PLOS` journals absent)
  2. there had to be at least **1.000 articles** in the given period (e.g., `Rev Mod Phys` absent)

OUTCOME

- over **4.300.000** articles from over **1.500** different journals,
- information about the title of the paper, the number of its authors, full abstract contents and OECD category it had been classified to,
- the **number of citations** it acquired between being published and 31st December 2014

MEASURES

property	title	abstract
length	number of characters	number of words
complexity	—	Gunning fog index F
	z-index	z-index
	Herdan's C	Herdan's C
sentiment	valence	valence
	arousal	arousal
number of authors		

1. Fog index: $F = \left(\frac{\#words}{\#sentences} + 100 \frac{\#complexwords}{\#words} \right)$
2. Herdan's C: $C = \frac{\log N}{\log M} \left[\begin{array}{c} \mathrm{M} - \text{text length} \\ \mathrm{N} - \text{vocabulary size} \end{array} \right]$
3. z-index: $z_{M,N} = \frac{N - \mu(M)}{\sigma(M)}$
4. Valence — emotional sign of the text (positive - 9, neutral - 5, negative - 1)
5. Arousal — level of emotional activation (low - 1, medium - 5, high - 9)

QUANTILE REGRESSION (QR)

IDEA

Find coefficients α and β of the relation

$$Y = \alpha(\tau) + \beta(\tau)X$$

dividing the dataset so that τ points lay below the line and $(1 - \tau)$ are above it.

ADVANTAGES

- we can examine different regimes (ranges) of Y,
- the log of p-th quantile is equal to the p-th quantile log-transformed Y

QUANTILE REGRESSION (QR)

IDEA

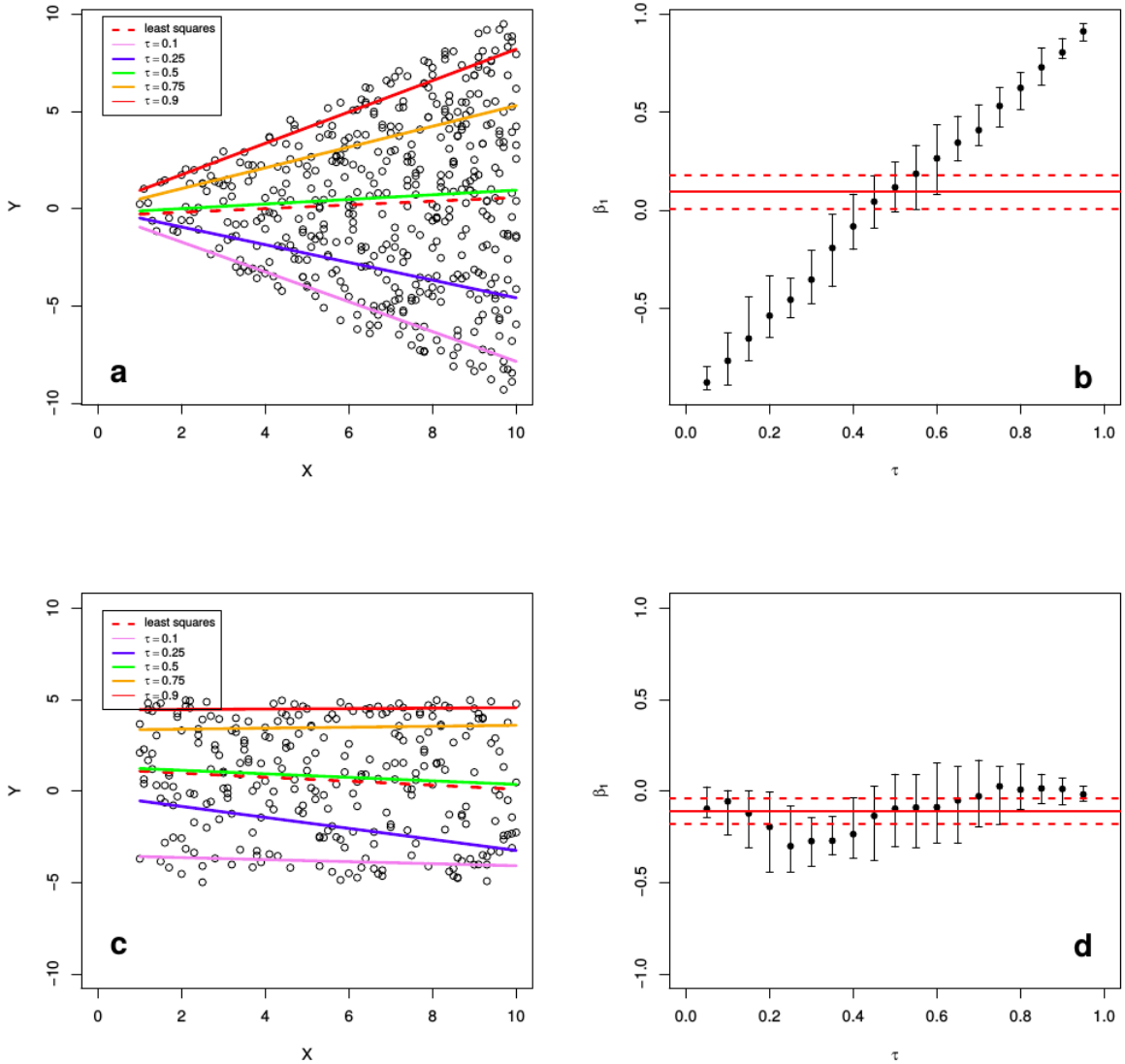
Find coefficients α and β of the relation

$$Y = \alpha(\tau) + \beta(\tau)X$$

dividing the dataset so that τ points lay below the line and $(1 - \tau)$ are above it.

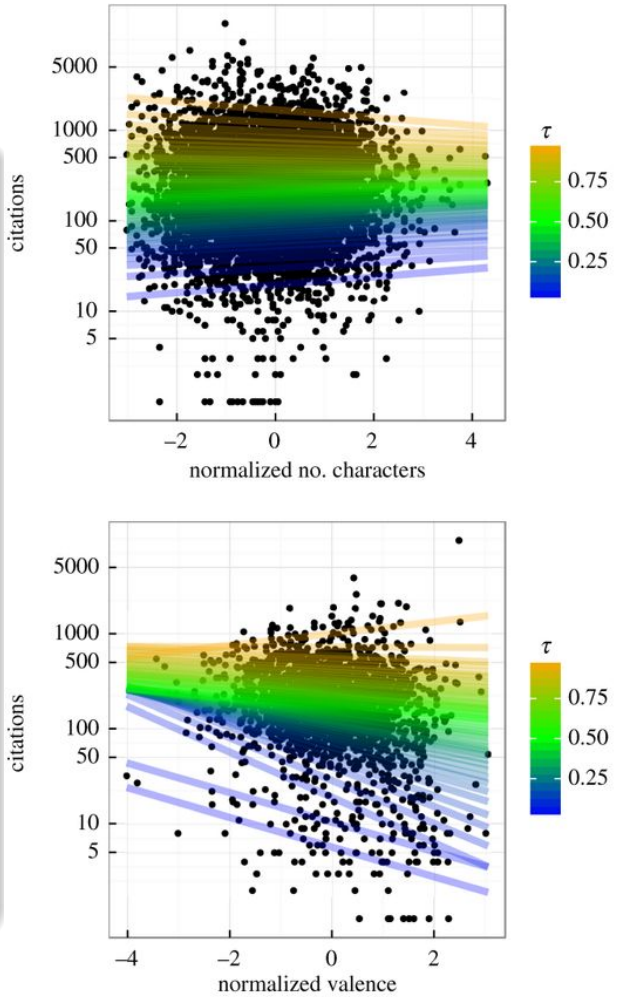
ADVANTAGES

- we can examine different regimes (ranges) of *Y*,
- the log of *p*-th quantile is equal to the *p*-th quantile log-transformed *Y*

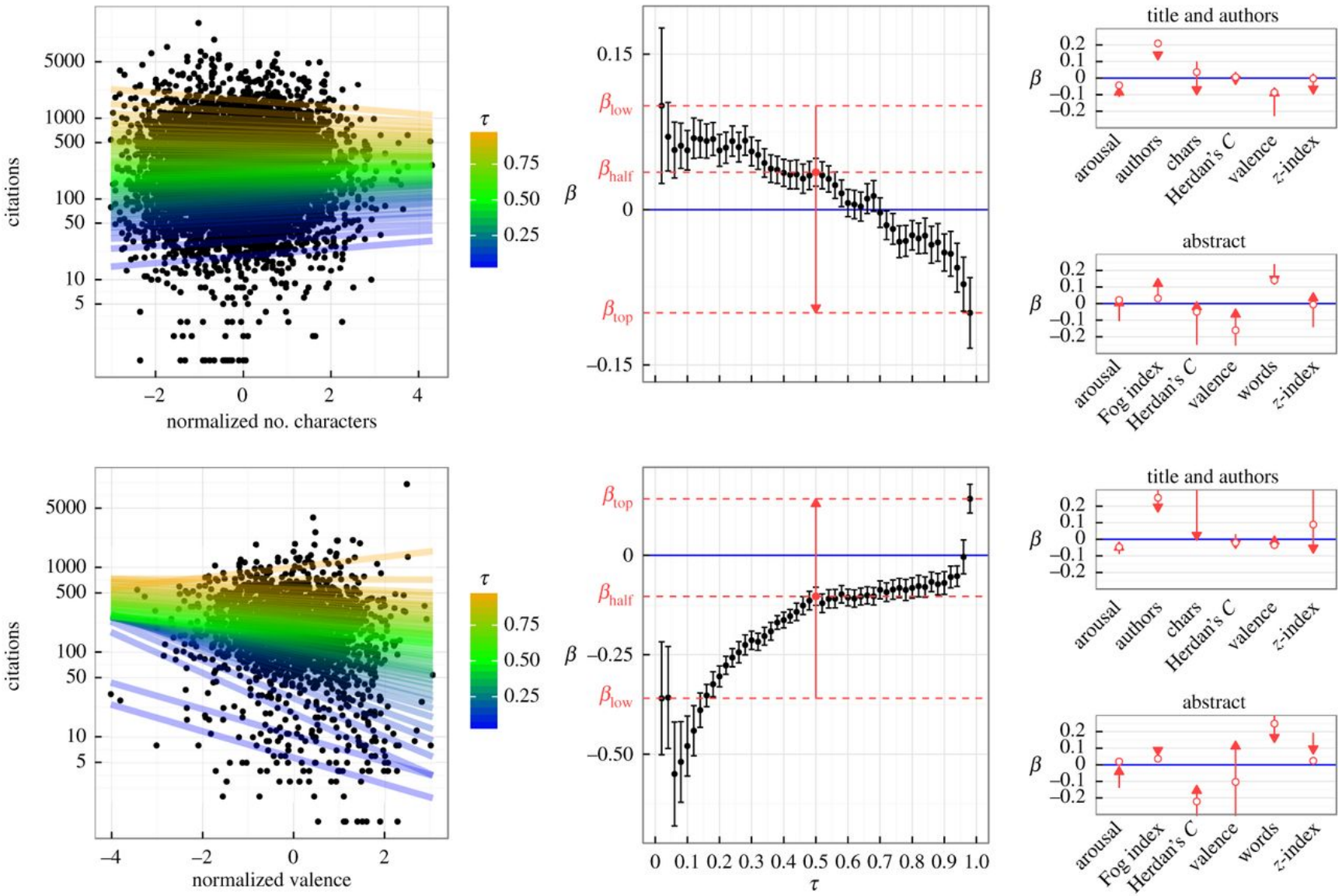


RESULTS - QR - DISCUSSION

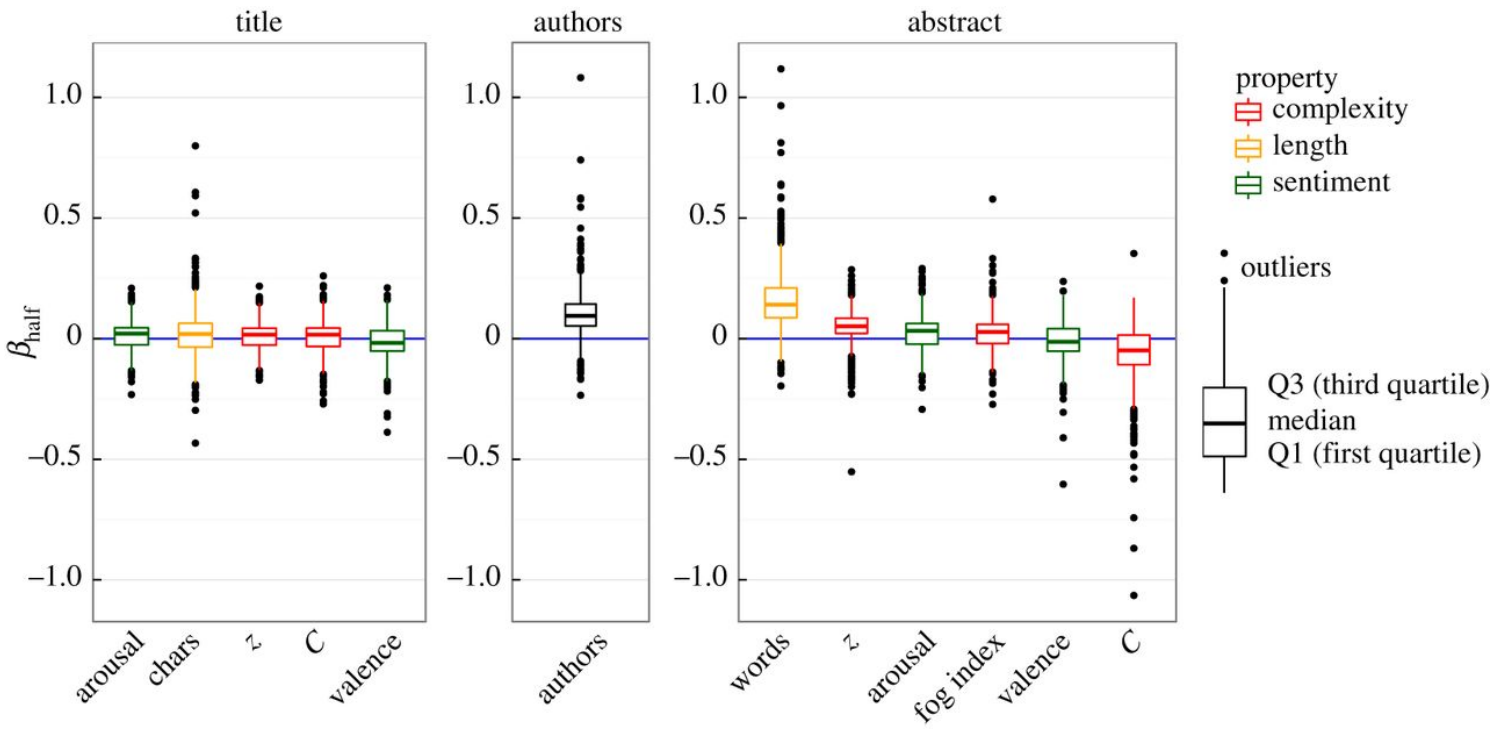
- broad scattering of the points — visual inspection fails even to detect whether the relation between X and Y is positive or negative,
- Pearson correlation coefficient r yields: $r = 0.02 \pm 0.01$ for title length (Science) and $r = -0.21 \pm 0.03$ for valence (Nature Genetics),
- such difference motivates us to go beyond linear correlations, which rely on the (homoscedasticity) assumption of uniform errors in the whole dataset.



RESULTS - QR


citations
normalized no. characters
τ
0.75
0.50
0.25
β
β_low
β_half
β_top
normalized valence
τ
title and authors
abstract
arousal
authors
chars
Herdan's C
valence
z-index
Fog index
words

RESULTS - COMPARISON OF FACTORS

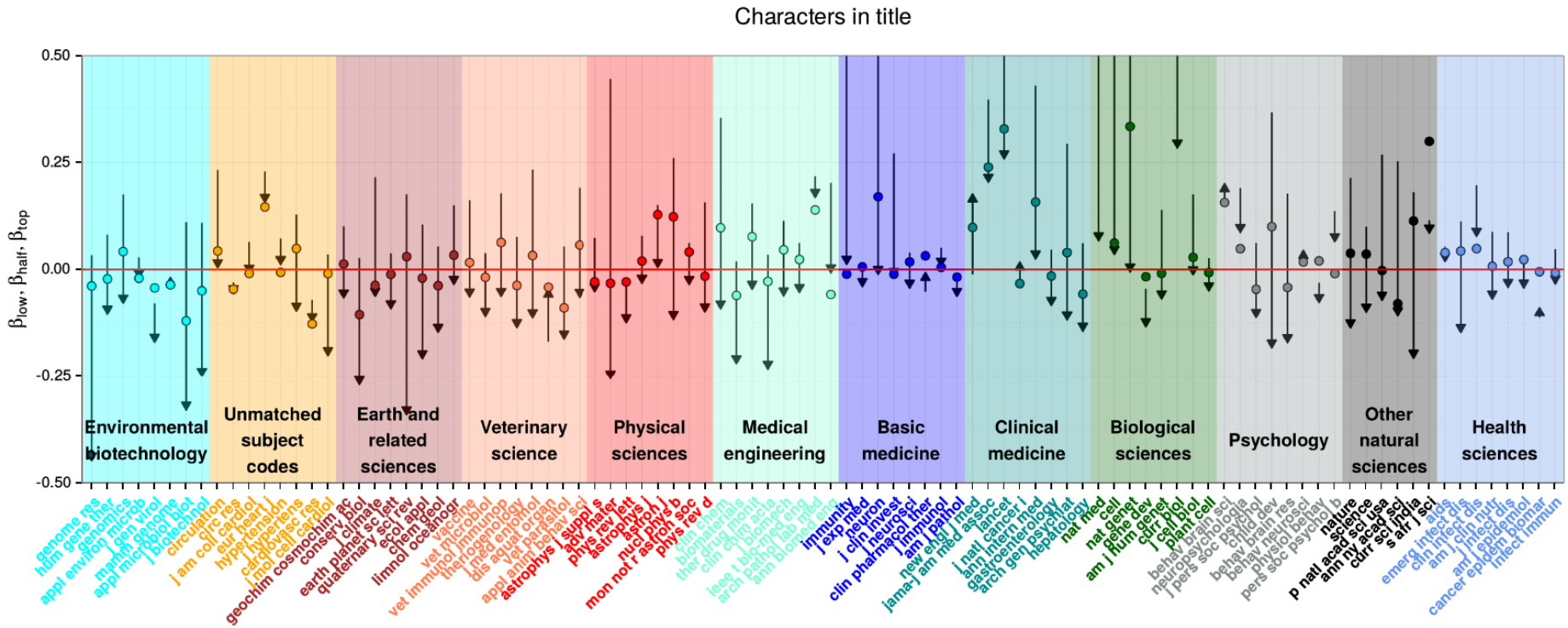


- The influence of factors is overall rather weak - $|\beta| < 0.5$ ($\beta = \ln 2$ means that the number of citations Y doubles by moving 1 standard deviation in X).
- the strongest factors are (i) the number of words in the abstract, (ii) the number of authors, and (iii) z-index in the abstract (over 75% of journals — equivalently, the whole box, are placed above zero).
- factors in the abstract are more visible than in the title

RESULTS - DIFFERENCE BETWEEN TYPICAL AND TOP PAPERS

property	factor	$\beta_{\text{top}} > \beta_{\text{half}}$	$\beta_{\text{top}} < \beta_{\text{half}}$	$\beta_{\text{top}} \neq \beta_{\text{half}}$
length	no. of characters (title)	2.6%	44.4%	47.0%
	no. of words (abstract)	8.3%	29.4%	36.7%
			mean	41.9%
complexity	Herdan's *C* (title)	18.7%	8.5%	27.2%
	Herdan's *C* (abstract)	34.9%	6.5%	41.4%
	z-index (title)	8.3%	16.7%	25.0%
	z-index (abstract)	24.6%	7.7%	32.3%
	fog index (abstract)	26.4%	8.0%	34.4%
			mean	32.0%
sentiment	arousal (title)	11.0%	13.5%	24.5%
	arousal (abstract)	15.7%	13.7%	29.4%
	valence (title)	16.1%	11.3%	27.4%
	valence (abstract)	29.2%	5.7%	34.9%
			mean	29.1%
	no. of authors	4.0%	39.6%	43.6%
			overall mean	33.7%

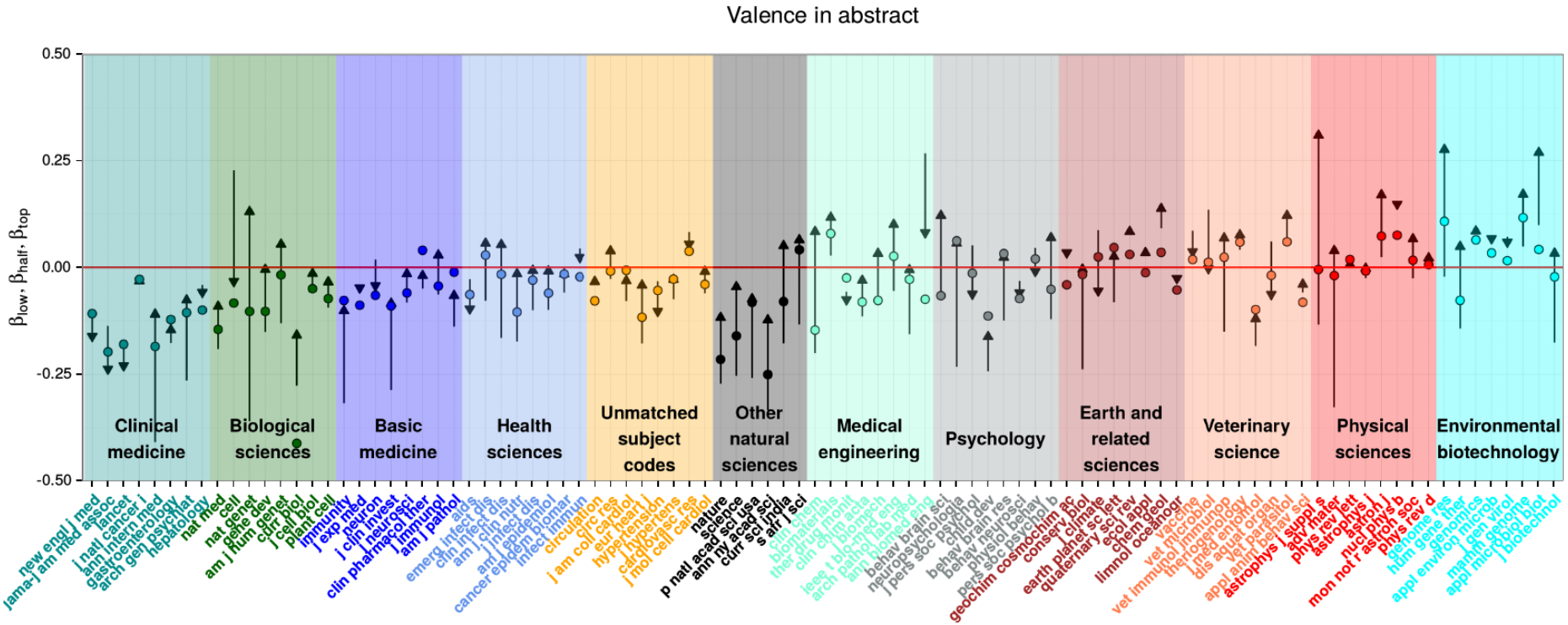
RESULTS - COMPARISON ACROSS JOURNALS (CHARS IN TITLE)



Easy comparison of the factor strength (calculating $\exp(\beta \Delta X)$, one can directly estimate how much gain in citations is obtained on average by a move in ΔX standard deviations in the variable X):

- for *Lancet* $\beta_{\text{half}} = 0.33$ and thus extending the length of the title by 1 standard deviation gives almost 40% gain in citations
- for *Nature* $\beta_{\text{half}} = 0.038$ and thus one obtains less than 4% gain.

RESULTS - COMPARISON ACROSS JOURNALS (VALENCE IN ABSTRACT)



Variation across journals is partially explained by disciplines, e.g. for *clinical medicine* all values of β in the case of valence in abstract are below zero, whereas for *physical sciences*, the majority is positive.

SUMMARY

1. Investigation on how textual properties of scientific papers relate to the number of citations they receive,
2. Main finding: correlations are **non-linear** and affect differently **most-cited** and **typical** papers,
3. In most journals **short titles** correlate **positively** with citations only for the **most** cited papers, for typical papers the correlation is in most cases **negative**,
4. Statistically significant effect present for **most factors**, but it is typically **weak** ($|\beta| < 0.5$),
5. large **variability** across journals

details & some data: **R Soc Open Sci 3, 160140 (2016)**