

## Nachrichten aus dem Datenraum

### Die Stars auf dem roten Teppich der Wissenschaft – Ein Blick auf *highly-cited papers*

In dieser Rubrik haben wir uns vor einiger Zeit bereits mit Publikationen befasst, die von sehr vielen Autor\*innen verfasst wurden. Diesmal soll es um Publikationen gehen, die sehr viel Aufmerksamkeit in der Wissenschaft auf sich gezogen haben. Zitationen in anderen Publikationen sind ein Maß für diese Aufmerksamkeit. Das am Fraunhofer INT entwickelte KATI-System bietet eine einfache Möglichkeit, alle Publikationen zu finden, die bspw. mindestens 10.000-mal zitiert wurden. Das sind derzeit knapp 300 Publikationen. Die ältesten Arbeiten stammen aus dem Jahr 1970, die jüngste ist erst 2018 publiziert worden.

Schaut man sich die Top 10 der am häufigsten zitierten Paper an, so fällt auf, dass in allen wichtige und grundlegende Methoden eingeführt wurden, auf die viele Wissenschaftler\*innen in ihrer täglichen Arbeit zurückgreifen. So wird eine Verbesserung von PCR-Tests in der Publikation mit den zweitmeisten Zitationen behandelt. Diese Tests spielen während der aktuellen Corona-Pandemie eine zentrale Rolle. Insgesamt können 6 Publikationen aus den Top 10 den Lebenswissenschaften und speziell der Biochemie zugeordnet werden. Das legt die Frage nahe, wie sich diese *highly-cited papers* über die verschiedenen Disziplinen verteilen und ob sie in einzelnen Disziplinen besonders häufig sind.

Um diese Frage zu beantworten, nutzen wir die sogenannten Subject Categories, die eine solche Einteilung in verschiedene Disziplinen darstellen. Im ersten Schritt berechnen wir die *erwartete Häufigkeit*. Dazu benötigt man die relative Häufigkeit einer Subject Category in Bezug auf die Gesamtzahl aller Publikationen in der Datenbank. Diese multipliziert man dann mit der Anzahl der Publikationen in unserem Sample und erhält daraus die gewünschte *erwartete Häufigkeit*. Diesen Wert vergleichen wir mit der tatsächlich *beobachteten Häufigkeit*. Das Ergebnis ist in der nachfolgenden Abbildung dargestellt.

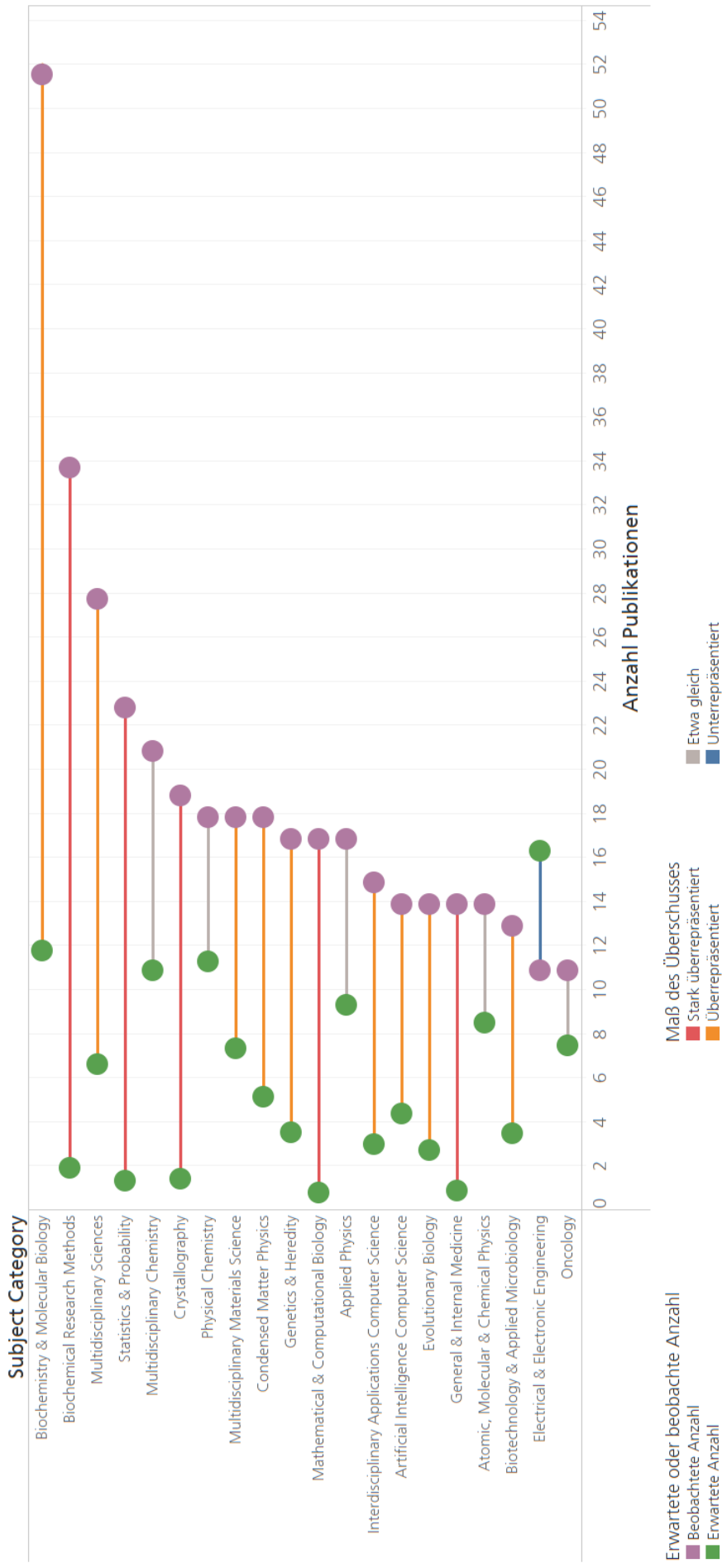
Es ist bspw. zu erwarten, dass etwa 12 Publikationen aus der Disziplin *Biochemistry & Molecular Biology* in unserem Datensatz zu finden sind. Tatsächlich sind es aber 52, also etwas mehr als 4-mal so viele. Diese Disziplin ist also überrepräsentiert. Noch größer ist die Diskrepanz für die Disziplinen *Biochemical Research Methods* sowie *Statistics & Probability*, von denen man gerade mal 2 Publikationen erwartet hätte. Es sind aber 34 bzw. 23, also jeweils deutlich mehr als 10-mal so viele Publikationen. Hingegen ist die Disziplin *Electrical & Electronic Engineering* unterrepräsentiert: Statt der erwarteten knapp 17 Publikationen sind nur 11 in unserem Datensatz enthalten.

In welchen Disziplinen nun besonders viele *highly-cited papers* publiziert werden, hängt von verschiedenen Faktoren ab. Dazu zählt zum einen natürlich die Publikations- und Zitationskultur, die sich von Fach zu Fach unterscheidet. So wird bspw. in den Lebenswissenschaften mehr publiziert und zitiert, als in den Sozialwissenschaften. Auf der anderen Seite scheinen manche Disziplinen prädestiniert zu sein, *highly-cited papers* hervorzubringen, da in diesen Methoden von grundlegender Bedeutung für den Fortschritt der Wissenschaft entwickelt werden. Dazu zählen sicherlich *Biochemical Research Methods* sowie *Statistics & Probability*. Manchmal ist es

aber vielleicht auch schlicht Zufall. So wurde die am häufigsten zitierte Arbeit – mehr als 90.000 Zitationen! - im Jahr 1996 in der Zeitschrift *Physical Review Letters* publiziert. Diese ist der sehr allgemein gehaltenen Disziplin *Multidisciplinary Physics* zugeordnet. In dieser Publikation geht es um die sogenannte Dichtefunktionaltheorie, das Arbeitspferd, wenn es um die Simulation von Werkstoffen und Biomolekülen geht.

## Vergleich der erwarteten und der beobachteten Anzahl an Publikationen in einer Subject Category

Dargestellt werden die Top-20 der Subject Categories in unserem Datensatz. Eine Subject Category ist **stark überrepräsentiert**, wenn sie mehr als 10-mal häufiger ist als erwartet. Sie ist **überrepräsentiert**, wenn sie mehr als doppelt so häufig ist, bzw. **unterrepräsentiert** wenn sie seltener ist, als erwartet. Ist sie zwischen weniger als 2-mal so häufig, so wird sie als **etwa gleich häufig** bezeichnet. Dies wird über die Farbe der Linie zwischen den Datenpunkten dargestellt.



Erwartete oder beobachtete Anzahl  
 Beobachtete Anzahl  
 Erwartete Anzahl

Maß des Überschusses  
 Stark überrepräsentiert  
 Überrepräsentiert

Etwa gleich  
 Unterrepräsentiert