Bioinformatics in and out of Beijing: a Bibliometrical analysis

Leon Rauschning

Contents

At the China College of the German Academic Scholarship Foundation in the working group on the Barefoot Doctor Programme and health collaboration between China and Tanzania, the history of bioinformatics in China came up in a discussion.

As a follow-up, I did a bibliometrics analysis using the Clarivate Web of Science resource, the results of which are compiled in this Rmarkdown document.

If you have ideas for extending this analysis or using a different approach to explore the origins of this young field in China, I'd love to hear them & perhaps collaborate on it! Feel free to reach out!

1 Obtaining the dataset

1.1 Prepare university list

Download HTML of a search for »China University« in the institution search at Web of Science. From the HTML, extract a list of button fields corresponding to the selection

```
grep -oPe '(?<=" lang="en">).*?(?=\</span\>)' universities.html > universities.txt

sed 's/\&amp\;/\&/g' universities.txt | sed ':a; N; $!ba; s/\n/ OR /g'
# | xclip -selection CLIPBOARD
```

1.2 Merge downloaded records

Search Web of Science for non-retracted articles in »Mathematical & Computational Biology« with author affiliations at the list of universities. Download full records in TSV format in batches of 1k, merge locally:

```
cp recs/savedrecs.txt mergedrecs.tsv # could also do this with head -n 1 >, might be cleaner
for x in recs/savedrecs\(*.txt; do tail -n 1000 $x >> mergedrecs.tsv; done

# Michigan counts as china, apparently
# Something about spelling correction?
```

```
# remove it
grep -v "Michigan" mergedrecs.tsv > processedrecs.tsv
```

1.3 Read in Data

```
data <- read.table('./processedrecs.tsv', sep='\t', header=TRUE, quote=NULL, comment.char='')
```

2 Plots

```
library(ggplot2)
```

There are 3 articles in the data from the 1930s (and one in 1929), all published in Biometrika. These aren't on what is typically considered bioinformatics, and from the 1940s until 1973 there is no paper in the dataset.

Two papers are annotated as published in January 2025 (Search conducted: 2024-10-02). I think these might be conference publications that have already been accepted, but could not definitely confirm this.

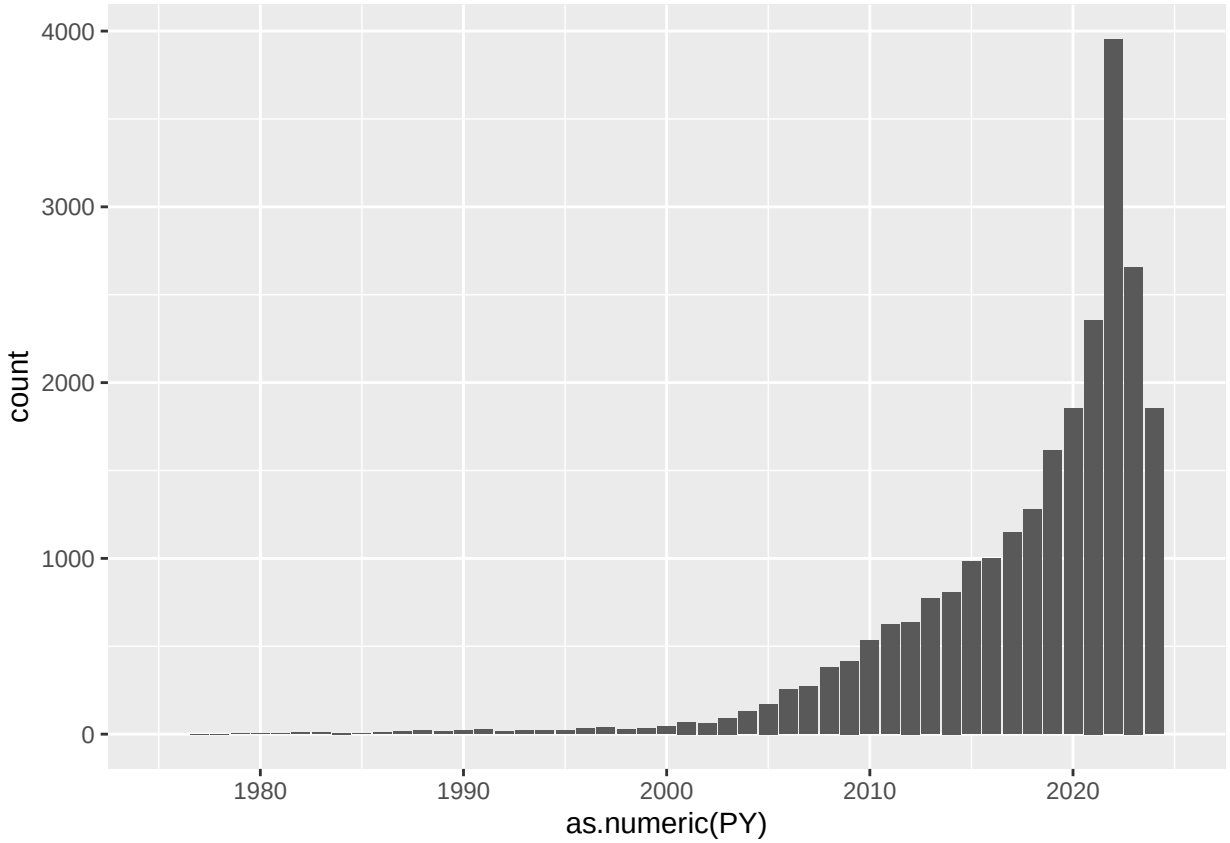
2.1 Basic Stats

```
ggplot(data=data) + geom_bar(aes(as.numeric(PY))) + xlim(1975, 2025)
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_count()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale
## range (`geom_bar()`).
```



The number of papers is rising exponentially, though without normalization to the global bioinformatics and Chinese non-bioinformatics paper output, it is hard to interpret that trend. Anecdotally, bioinfo seems to have gotten off the ground a bit later in China than in the US and Europe (Late 90s/Early 2000s).

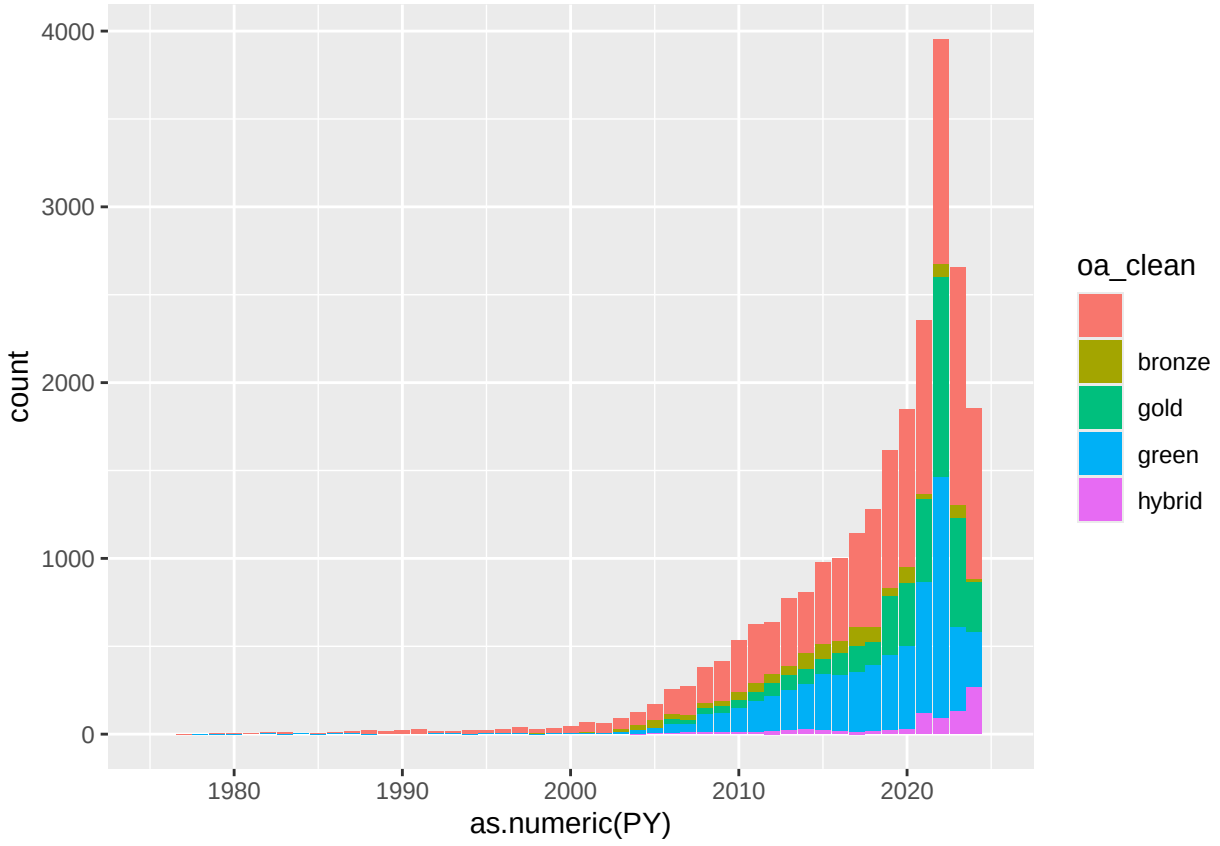
The spike in papers published in 2022 could correspond to an increased focus on manuscript writing during the lockdowns, or publications of COVID-19-related articles. It could also be explained by articles published in 2023/2024 not yet being indexed in the Web of Science database.

```
data$oa_clean <- tolower(sub("[ ,].*", "", data$OA))
ggplot(data=data) + geom_bar(aes(as.numeric(PY), fill=oa_clean)) + xlim(1975, 2025)
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_count()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale
## range (`geom_bar()`).
```



As expected, the share of articles published as Open Access seems to have increased in the 2010s. Still, many articles are not published as open access, or otherwise are not indexed as Open Access by the Web of Science. Especially green OA may be under-indexed.

2.2 Citations

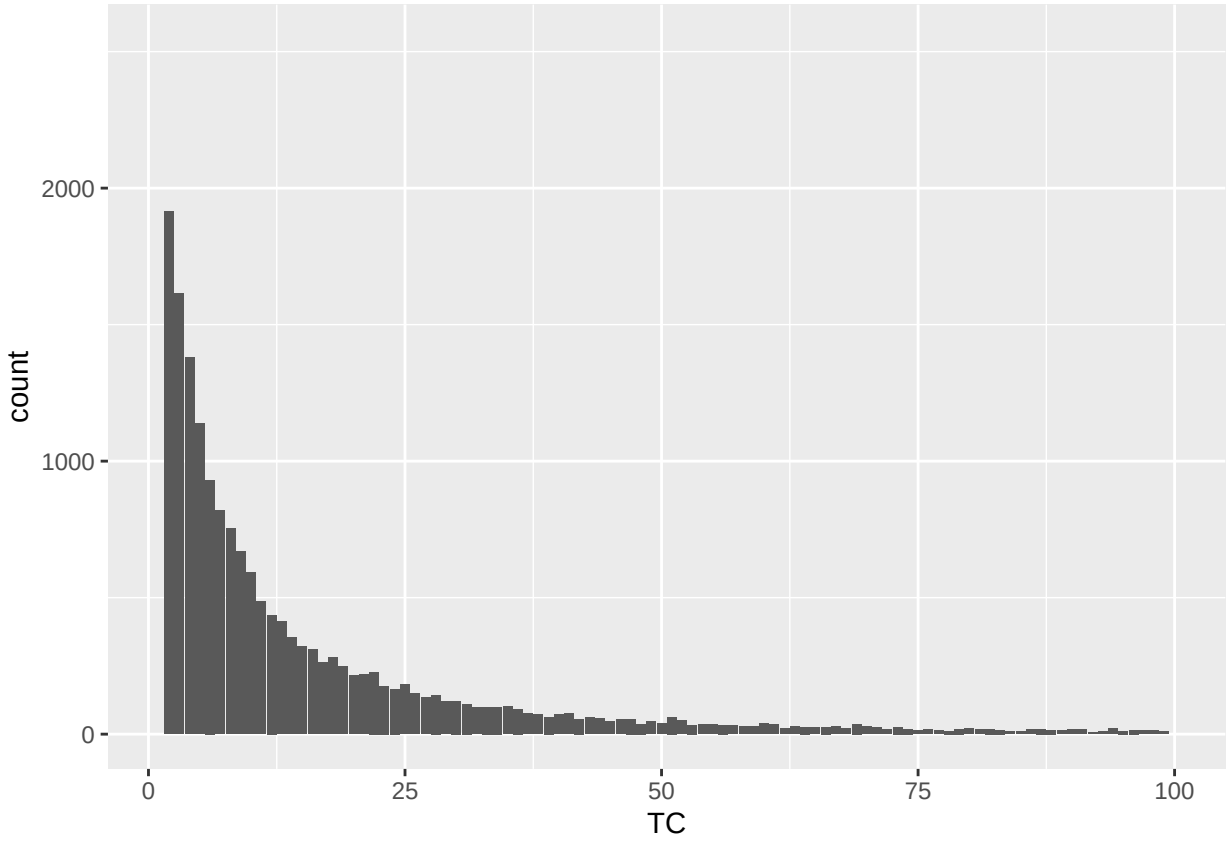
```
data$TC <- as.numeric(data$TC)
```

```
## Warning: NAs introduced by coercion
```

```
ggplot(data=data) + geom_bar(aes(TC)) + xlim(1, 100)
```

```
## Warning: Removed 4480 rows containing non-finite outside the scale range
## (`stat_count()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale
## range (`geom_bar()`).
```




```
mean(data$TC, na.rm=TRUE)
```

```
## [1] 19.7008
```

```
median(data$TC, na.rm=TRUE)
```

```
## [1] 5
```

~~The top cited paper is Uni Michigan – seems the text search for China caught this as well.~~ Fixed, see above.

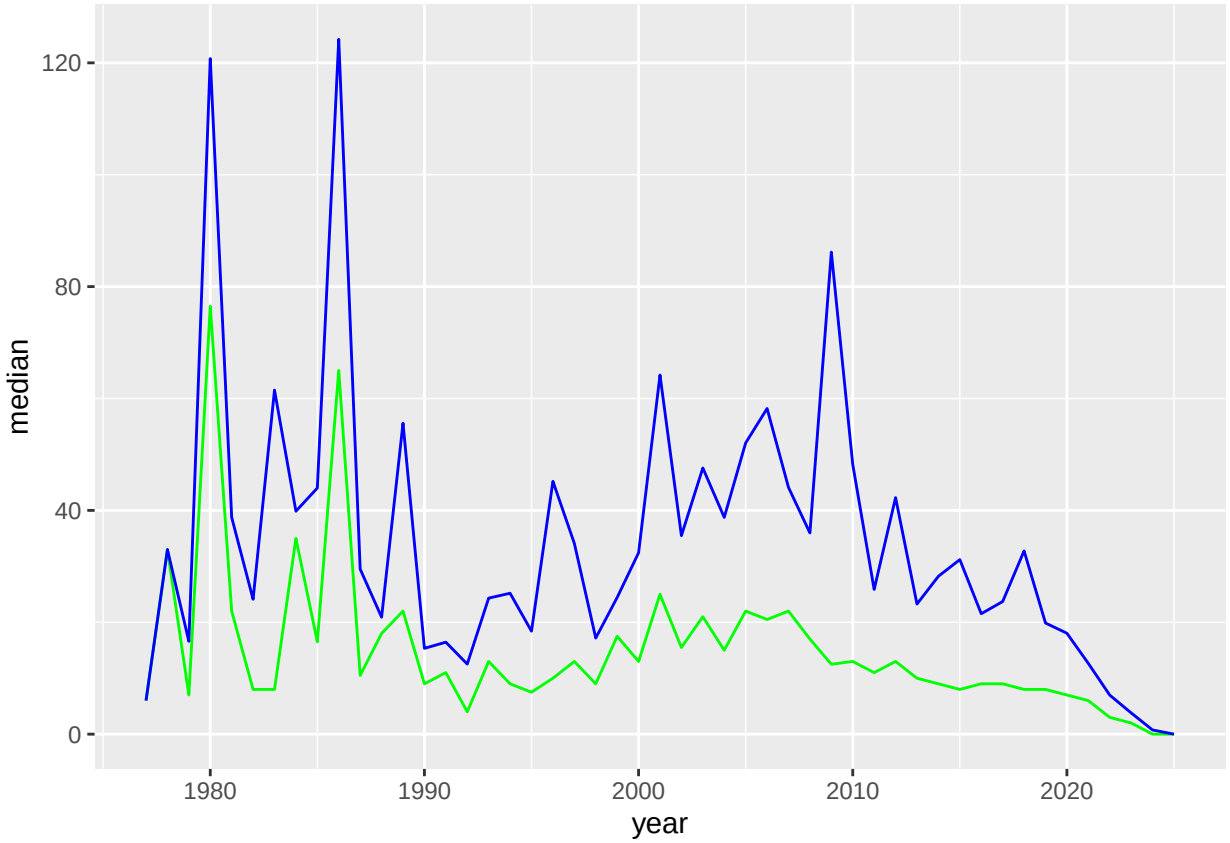
A few papers in the dataset are very highly cited (10k-20k citations), including some »canon« bioinformatics software and file format publications. These are excluded in the plots, but drive up the mean citation count (see also below). Other than the very long tail, the distribution seems to follow a power law?

```
citbyyear <- data.frame(1977:2025)
colnames(citbyyear) <- c('year')
citbyyear$median <- lapply(citbyyear$year, FUN=function(x) {median(data[data$PY == x, 'TC'], na.rm=TRUE
citbyyear$mean <- lapply(citbyyear$year, FUN=function(x) {mean(data[data$PY == x, 'TC'], na.rm=TRUE) })

citbyyear$year <- as.numeric(citbyyear$year)
citbyyear$mean <- as.numeric(citbyyear$mean)
citbyyear$median <- as.numeric(citbyyear$median)

citbyyear
```

```
##    year median        mean
## 1  1977    6.0   6.0000000
## 2  1978   33.0  33.0000000
## 3  1979    7.0  16.6000000
```

```
## 4  1980   76.5 120.7500000
## 5  1981   22.0  38.7500000
## 6  1982    8.0  24.1111111
## 7  1983    8.0  61.5000000
## 8  1984   35.0  39.8571429
## 9  1985   16.5  44.0000000
## 10 1986   65.0 124.2000000
## 11 1987   10.5  29.5000000
## 12 1988   18.0  20.9047619
## 13 1989   22.0  55.5882353
## 14 1990    9.0  15.3333333
## 15 1991   11.0  16.4400000
## 16 1992    4.0  12.5294118
## 17 1993   13.0  24.3157895
## 18 1994    9.0  25.1904762
## 19 1995    7.5  18.4090909
## 20 1996   10.0  45.1935484
## 21 1997   13.0  34.1025641
## 22 1998    9.0  17.1785714
## 23 1999   17.5  24.4375000
## 24 2000   13.0  32.4186047
## 25 2001   25.0  64.1884058
## 26 2002   15.5  35.4843750
## 27 2003   21.0  47.5760870
## 28 2004   15.0  38.7480315
## 29 2005   22.0  52.0523256
## 30 2006   20.5  58.2086614
## 31 2007   22.0  44.0996310
## 32 2008   17.0  35.9658793
## 33 2009   12.5  86.1634615
## 34 2010   13.0  48.2218045
## 35 2011   11.0  25.8683788
## 36 2012   13.0  42.2805643
## 37 2013   10.0  23.2597403
## 38 2014    9.0  28.2202970
## 39 2015    8.0  31.2008155
## 40 2016    9.0  21.5284715
## 41 2017    9.0  23.7050611
## 42 2018    8.0  32.7537138
## 43 2019    8.0  19.8680297
## 44 2020    7.0  18.0372570
## 45 2021    6.0  12.6515280
## 46 2022    3.0   6.9820390
## 47 2023    2.0   3.8009782
## 48 2024    0.0   0.7486516
## 49 2025    0.0   0.0000000
```

```
ggplot(citbyyear) +
    geom_line(aes(x=year, y=median), colour="green") +
    geom_line(aes(x=year, y=mean), colour="blue")
```



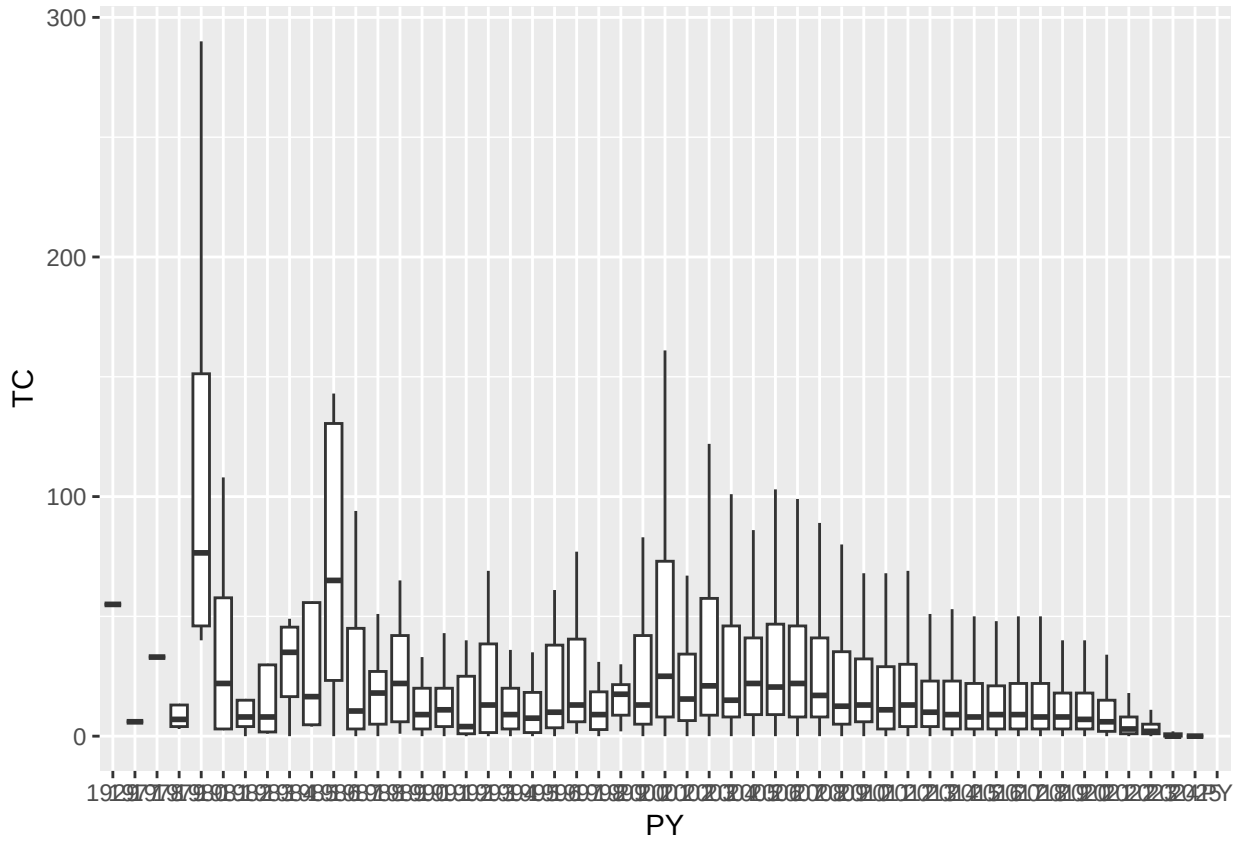
Analysis of the mean/median (blue/green) citations of bioinformatics papers by year.

In the years until 1990, low numbers of papers overall leads to a very high year-to-year variance. Even after that, individual extremely highly cited publications drive up the average of entire years by a factor of 1.5-2x (about what would be expected for a 10-20k citation paper in a year of ~1k publications with typically ~15-20 citations).

Citations per paper seem to have dropped over the years, which likely is a consequence of more recent papers not having the chance to have been cited by other papers yet. Unfortunately, Web of Science doesn't seem to provide an age-matched measure of citations out of the box.

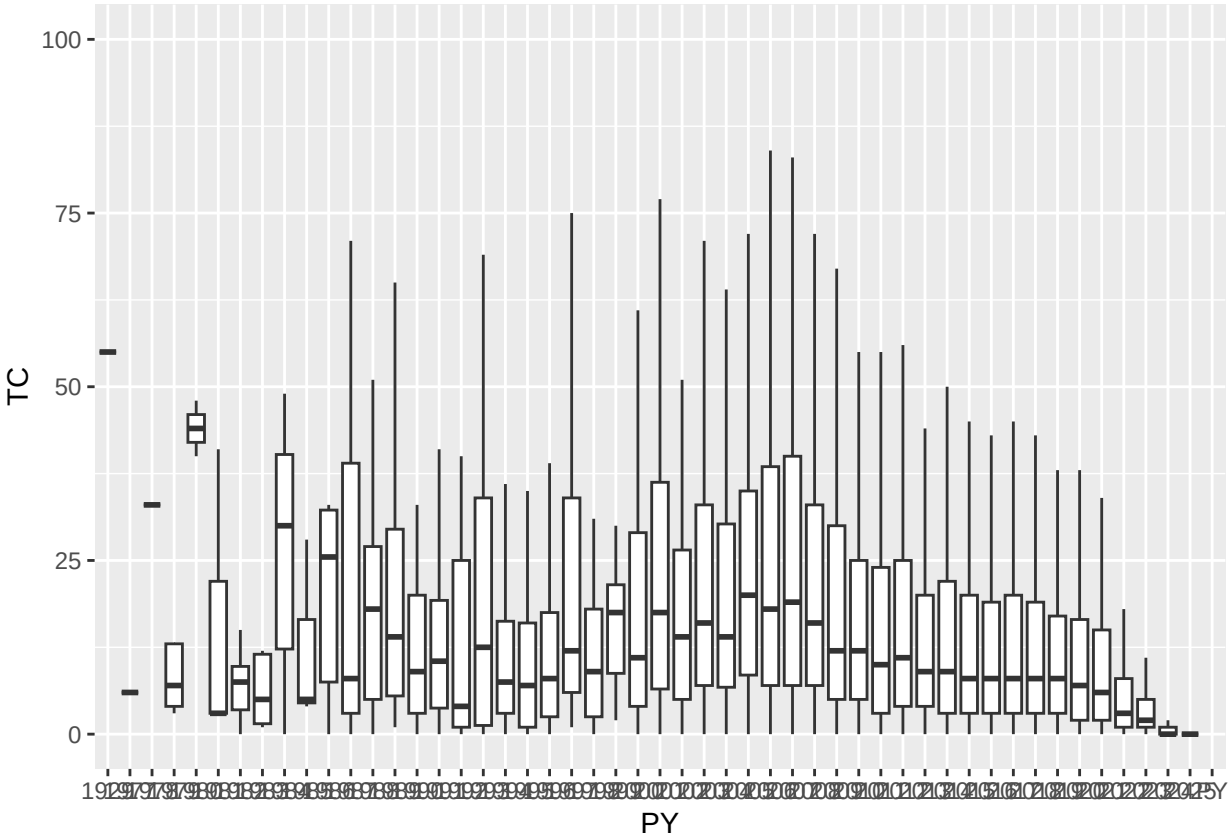
```
ggplot(data) +
    geom_boxplot(aes(PY, TC), outliers=FALSE)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
# filter to papers <100 citations
ggplot(data) +
    geom_boxplot(aes(PY, TC), outliers=FALSE) +
    ylim(0, 100)
```

```
## Warning: Removed 671 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Because of the high within- and between-year-variance in citations and the strong effect of very few extremely highly cited papers, the distribution of citations by year are shown here also as tukey plots. This reveals that the decreasing mean citations in the 2010s is driven by fewer very highly cited papers (the median and quartiles change only slightly, with unclear significance), consistent with the hypothesis of high-impact publications not having had the chance to be widely cited at the time of analysis. A comparison at a future timepoint would be an interesting avenue to further investigation.

3 Conclusions

Due to the fragmented nature of these data, not many conclusions can be drawn from it without further in-depth analysis and comparison to other discipline's output and global trends in bioinformatics publications.

What stands out is the strong exponential increase in bioinformatics publications at Chinese universities, as well as the impact of few very highly cited papers. Analysing other measures of research activity, like datasets deposited in repositories or code commits published in zenodo or GitHub may provide a more complete picture of how bioinformatics in China got to where it is now, as well as investigations of the history of bioinformatics departments at individual institutions.