Worldwide Gender Differences in Public Code Contributions

and how they have been affected by the COVID-19 pandemic

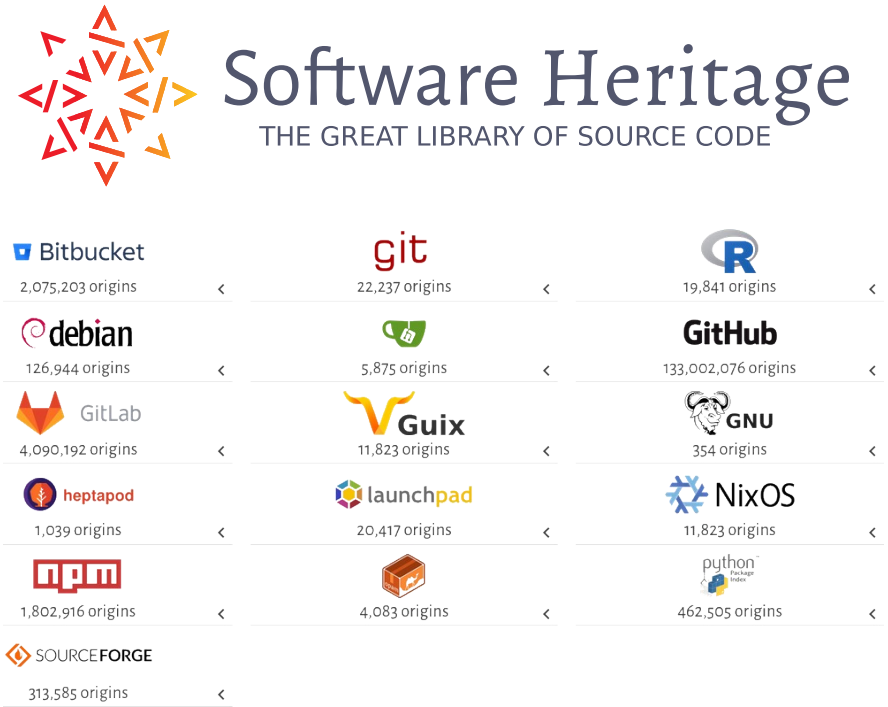
Davide Rossi - University of Bologna, Italy
Stefano Zacchiroli - Télécom Paris, Polytechnic Institute of Paris, France

Research Questions

Context: gender imbalance in Free/Open Source Software contributions

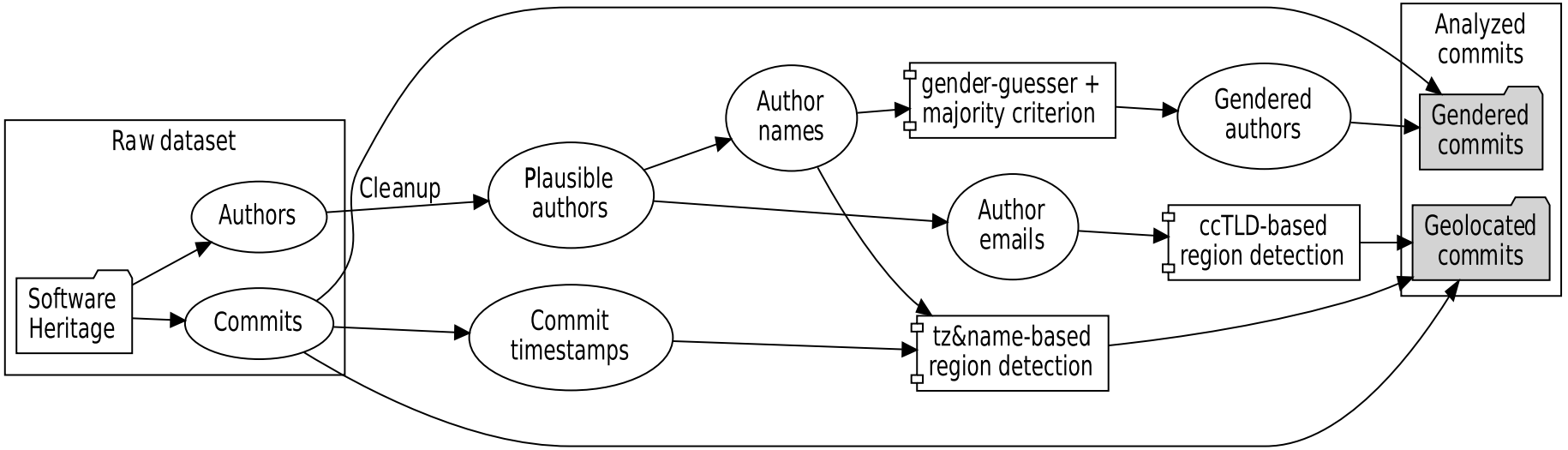
- **RQ1**. what is the overall breakdown *by gender and UTC offset* in contributions (and contributors) to public source code?
- **RQ2**. what is the overall breakdown *by gender and by world regions* in contributions (and contributors) to public source code?
- **RQ3**. Has the impact of the COVID-19 pandemic on contributions to public code been quantitatively different by gender?

Dataset

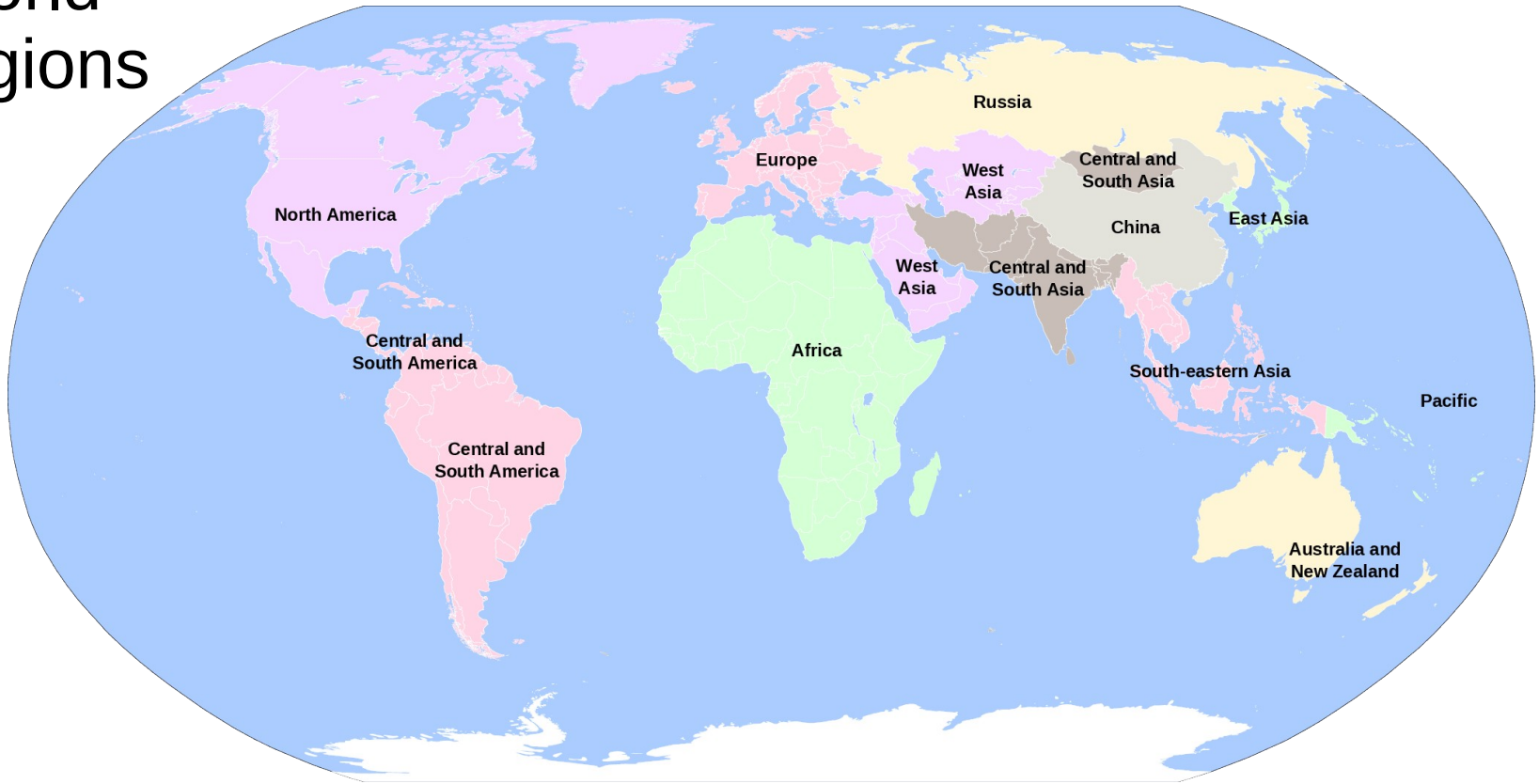


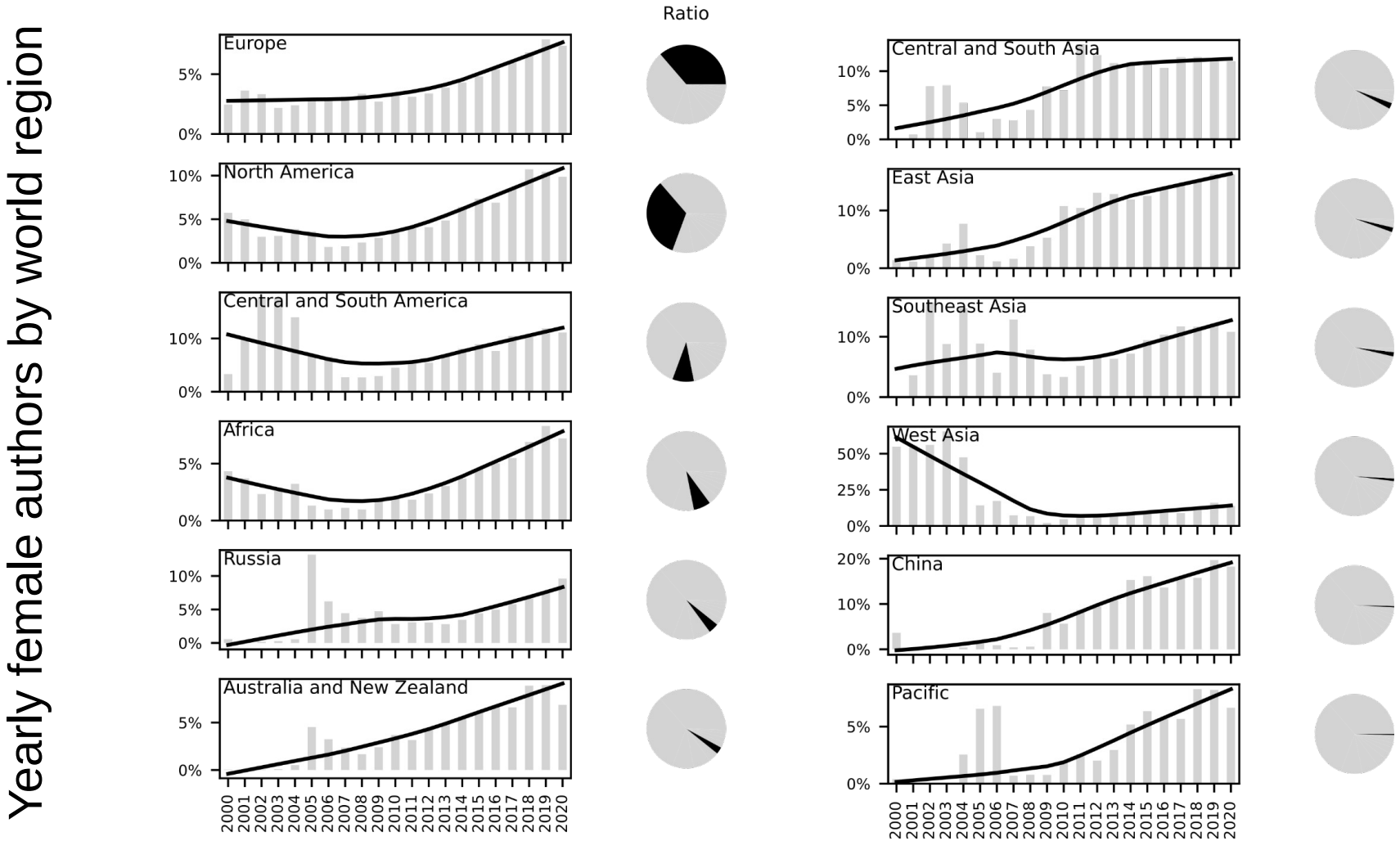
- 160 million public projects
- Revisions (commits) and their authors from 1970-01-01 to 2021-07-07 (50 years)
- 2.2 billion commits in total
- 43 million (unique) authors
- Sourced from different VCS systems ⇒ only common metadata are usable
  - Commit timestamp (with UTC offset)
  - Author’s name
  - Author’s email

The data pipeline

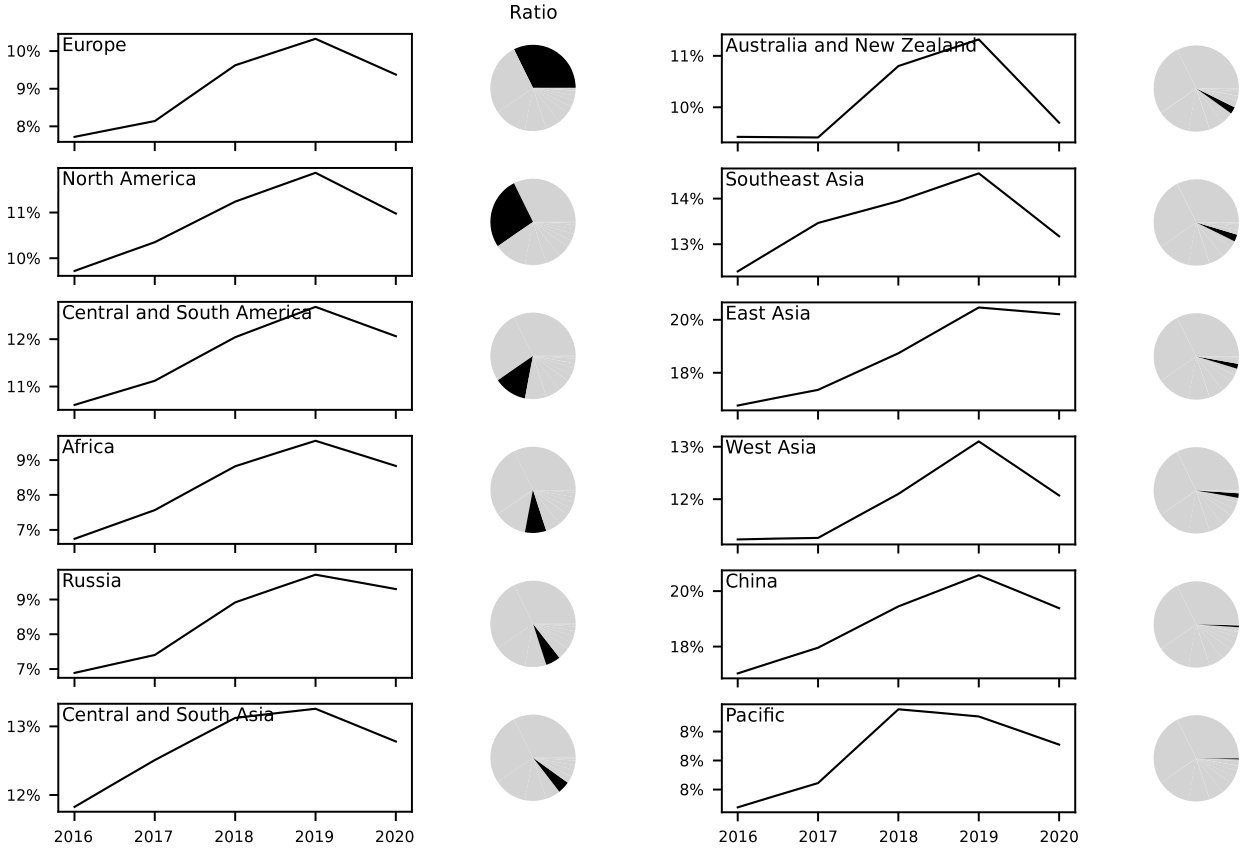


World regions




Yearly female authors by world region
Ratio
Europe
North America
Central and South America
Africa
Russia
Australia and New Zealand
Central and South Asia
East Asia
Southeast Asia
West Asia
China
Pacific
0%
5%
10%
25%
50%
20%
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Effects of the pandemic (yearly female authors, 2016—2020)

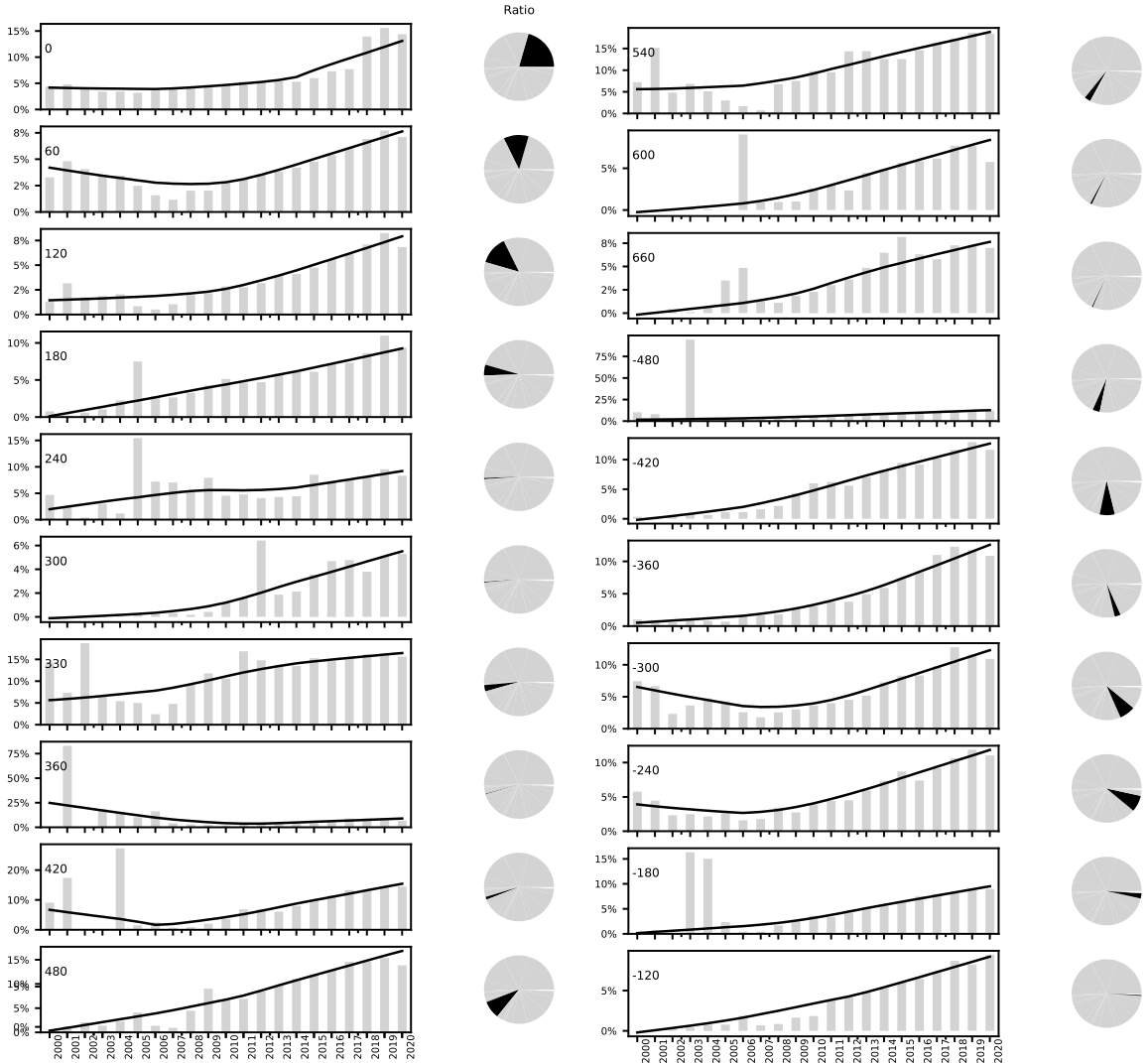


Conclusions, threats and future work

- Takeaways
  - **Gender gap is shrinking**, with women participation having increased steadily over the past 12 years with little differences among world zones
  - The ratio of **women participation during the COVID-19 pandemic has decreased**
- Threats
  - Construct validity: Accuracy/scale trade-off - Gender/zone detection
  - External validity: The study dataset is the largest *approximation* of contributions to public code that is readily and publicly available for analysis
- Future work
  - Apply multi-method techniques to dig into relevant outliers, e.g. why COVID-19 contraction in women participation appears to have impacted Asiatic regions less than others

Appendix

Yearly female authors by UTC offset



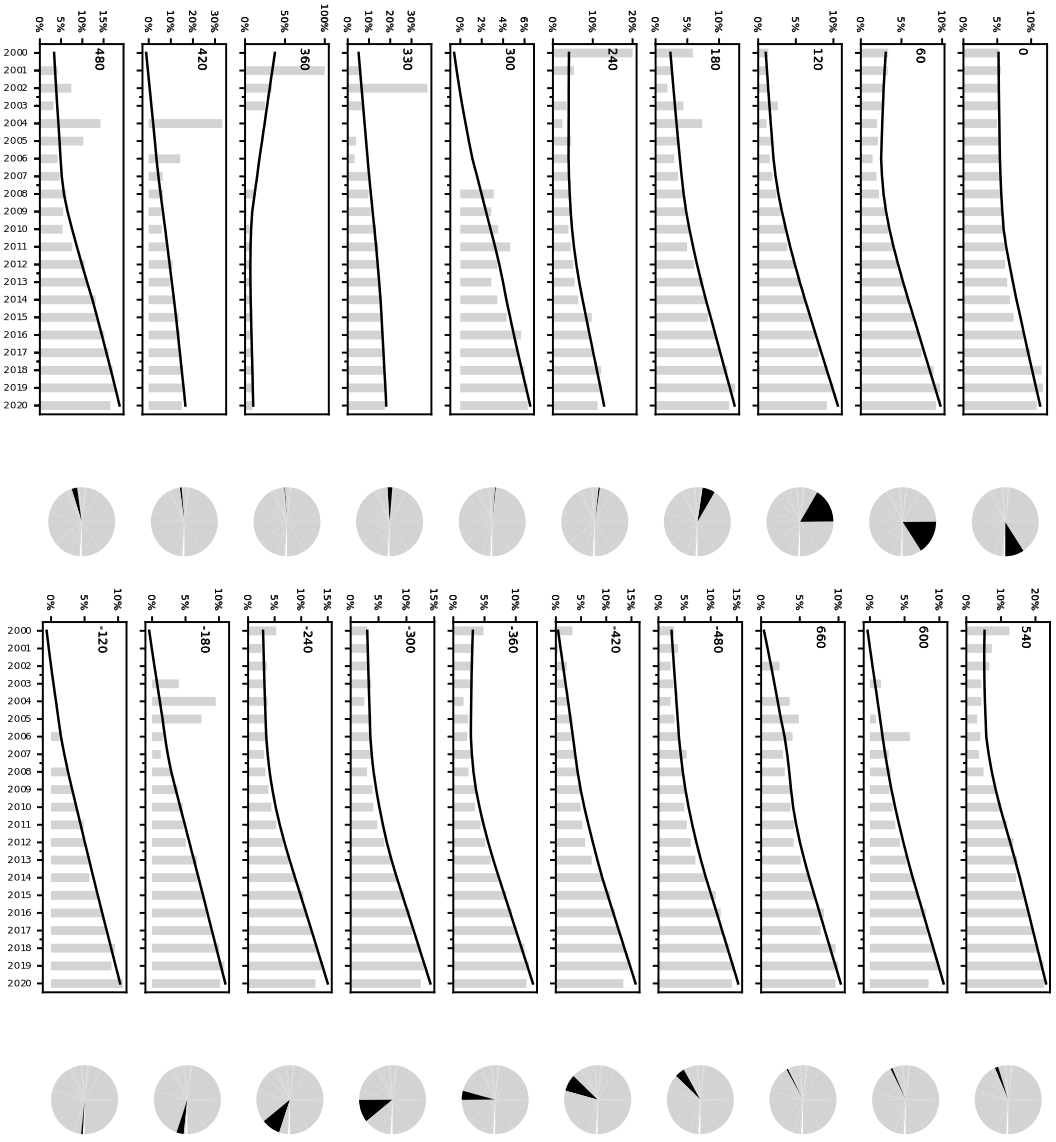
Gender detection

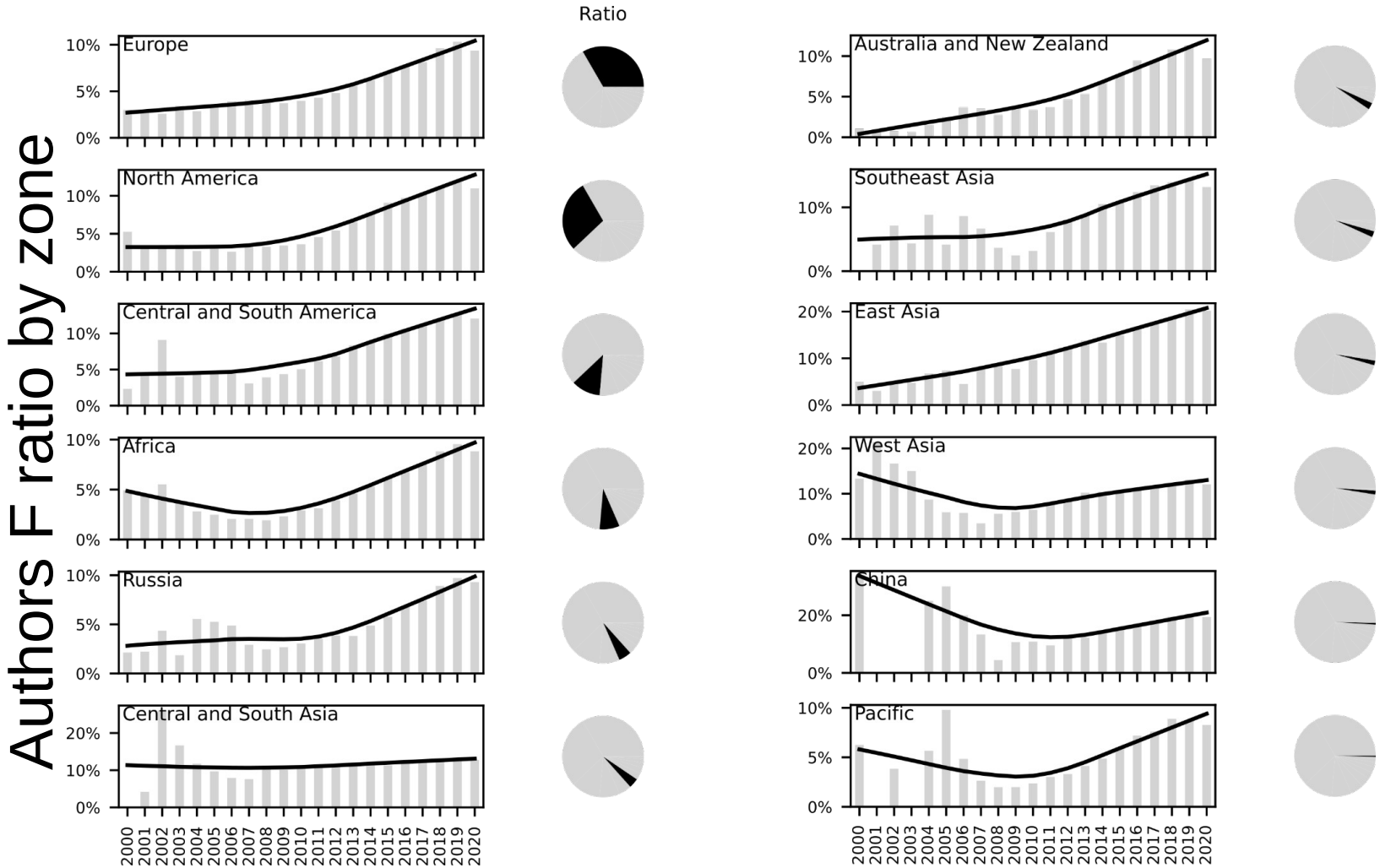
- Names are split in tokens
- Iff a strict majority of tokens for a given author is detected by gender-guesser as belonging to one gender we associate the majority gender to the author; otherwise their gender will remain unknown

Zone detection

- ccTLD-based detection
  - Straightforward
  - Problem: the adoption of national domains varies across zones
- tz&name-based detection
  - Likelihood measure of an author’s name to belong to a country
  - Limited to nations/territories in the commit’s time offset (at the time the commit was performed)
  - A score which is a sum of products name_frequency * population in that given offset for all candidate nations, aggregated by zone
  - Problem: time offset are not always reliable
- For this study: ccTLD if offset is zero, tz&name otherwise

Authors F ratio by UTC offset




Authors F ratio by zone
Ratio
Europe
North America
Central and South America
Africa
Russia
Central and South Asia
Australia and New Zealand
Southeast Asia
East Asia
West Asia
China
Pacific
0%
5%
10%
20%
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020