

2 - Crowd-sourced online data

Diego Alburez-Gutierrez
MPI IDR

European Doctoral School of Demography 2019-20

31/03/2020



MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

Agenda

1. Q&A
2. Crowd-sourced data
3. User-generated family trees
4. Limitations and bias

Q&A

- ▶ Download the course materials
- ▶ Knit an Rmarkdown document
- ▶ Review the assignment instructions
- ▶ Other?

Crowd-sourced data

What is crowd-sourced data?

- ▶ ‘Bottom-up’ user-generated content
- ▶ Usually large
- ▶ Available online - may have been produced offline
- ▶ By-product of decentralized activity

Examples in this session

1. Recruitment platforms
2. Online activism in the UK
3. Knowledge building in Wikipedia
4. Family history research

The issue of representation

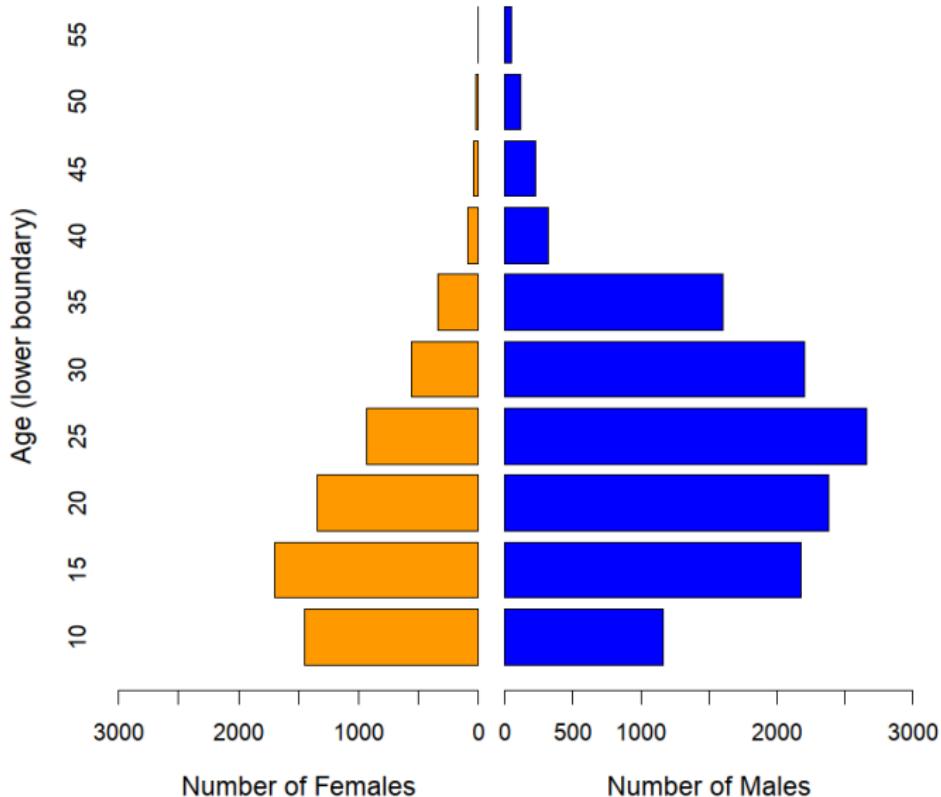


Figure 1: Why so many men?

Crowd-sourcing as recruitment

Aim: Strategies for recruiting iPhone users to participate in a web survey through Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk.

- ▶ Pull-method (cash-based platforms): recruits were more cost efficient and committed to the survey task
- ▶ Push method (ad-recruitment): recruits were more demographically diverse

Antoun, C., Zhang, C., Conrad, F.G., and Schober, M.F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* 28(3):231–246.

Question time!



1. Which challenges do you foresee when crowd-sourcing data?

Question time!



1. How can crowd-sourcing systematically bias the data collection?

Question time!



1. How can crowd-sourcing systematically bias the data collection?
2. What are the privacy considerations?

Some magic sampling...



```
n <- c("Anna", "Lara", "Ilgi", "Octavio", "Qi", "Luca"
      , "Miguel", "Madalina", "Margherita", "Niall"
      , "Momoko", "Alexander", "Rustam",
      "Serena", "Daniel", "Andres", "Heiner")

set.seed(42)
who <- sample(n, 3, replace = F)
n <- n[!n %in% who]
print(who)

## [1] "Heiner" "Qi"      "Anna"
```

Crowd-sourcing as recruitment: representativeness

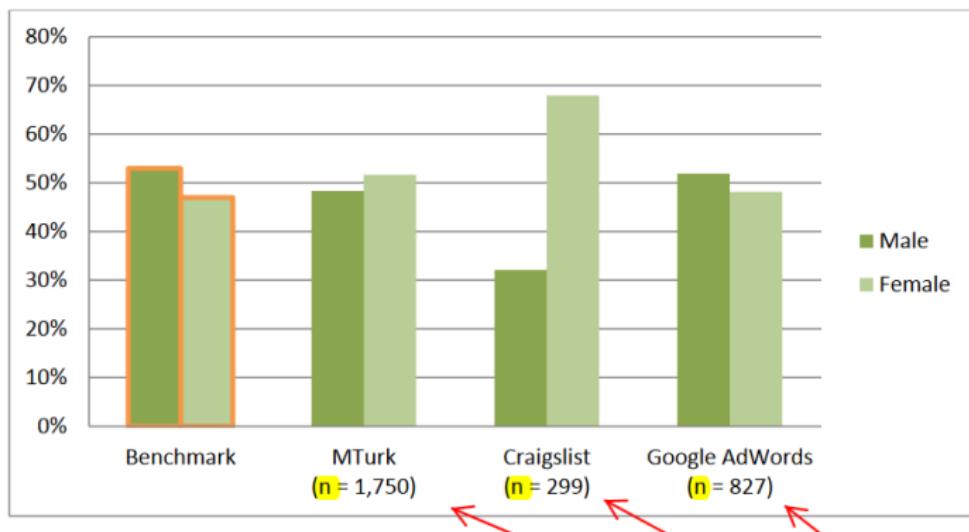


Figure 2: By sex

Antoun, C., Zhang, C., Conrad, F.G., and Schober, M.F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* 28(3):231–246.

Crowd-sourcing as recruitment: representativeness

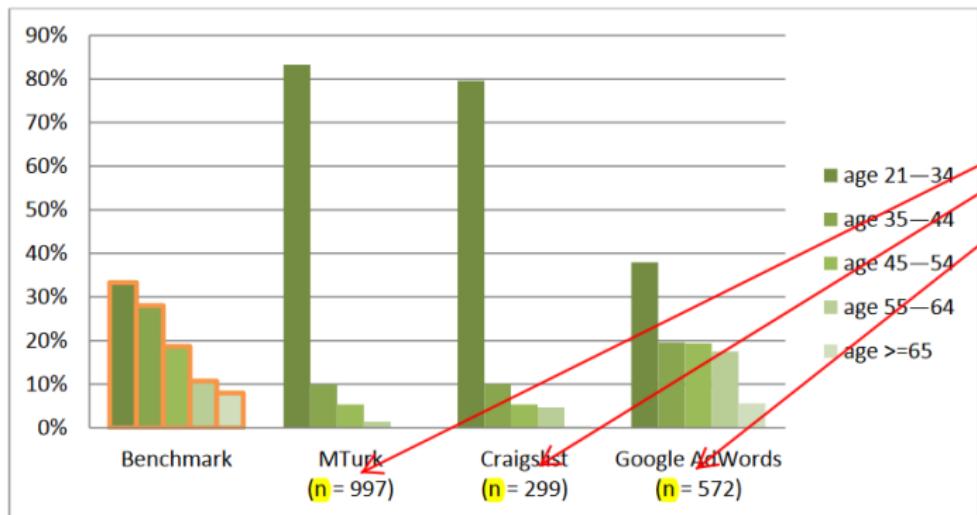


Figure 3: By age group

Antoun, C., Zhang, C., Conrad, F.G., and Schober, M.F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field Methods* 28(3):231–246.

An example closer to home

The screenshot shows a Facebook page for "Health Behavior Survey". The page header includes standard social media controls: Liked, Following, Share, and a three-dot menu. On the right, there's a "Send Message" button. The main content area has a sidebar on the left with links to Home, Posts, Reviews, Photos, About, and Community, and a "Create a Page" button. The main content area displays two posts:

- Health Behavior Survey** posted on 21 March at 17:43: "Here we are! Read this article to know more about our research." (with a link to an external article).
- Max Planck Institute for Demographic Research** posted on 20 March at 17:58: "A team of interdisciplinary scientists led by Daniela Perrotta and André Graw want to analyze people's health behavior during the coronavirus pandemic. They are currently conducting a survey via Facebook: https://www.demogr.mpg.de/.../coronavirus_survey_that_is_what..." (with a video thumbnail showing four people in a video call).

On the right side, there are sections for "Community" (with 287 likes and 302 followers) and "About" (listing the location as Konrad-Zuse-Str. 1, 18057 Rostock, Germany, and providing a phone number, email, and website). The "About" section also indicates the page is managed by "College & University - Scientist".

Figure 4: Learning more about the coronavirus

Visit and like: <https://www.facebook.com/Health-Behavior-Survey-106662670948465/>!

A different strategy: post-stratification

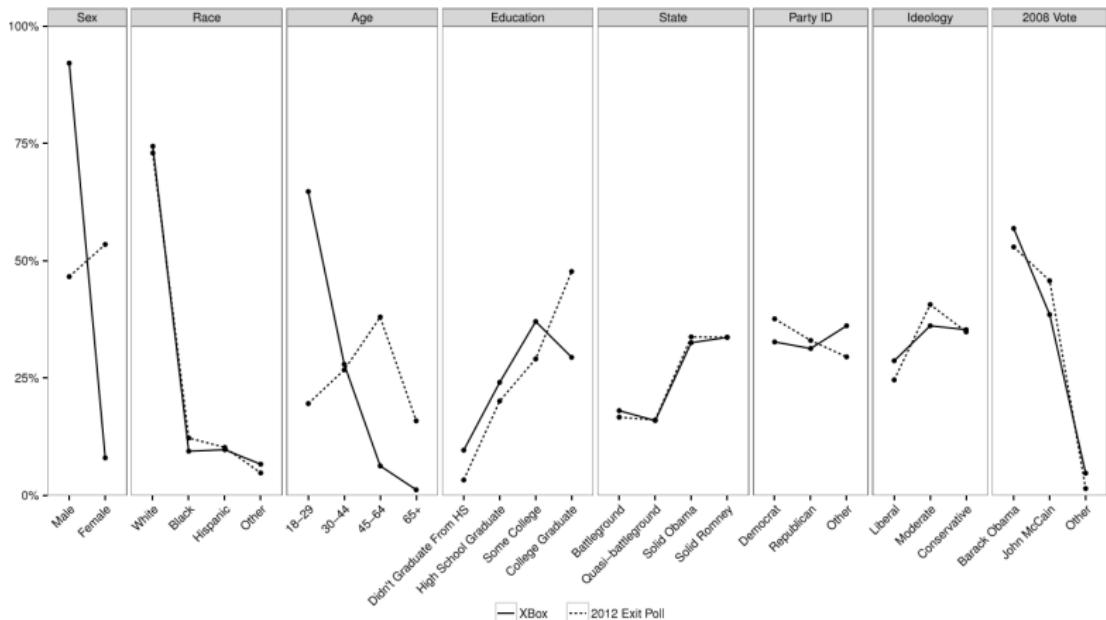


Figure 5: Different demographics

Wang, Rothschild, Goel, and Gelman 2015. Forecasting elections with non-representative polls. International Journal of Forecasting, 31 (3).

Forecasting with non-representative polls

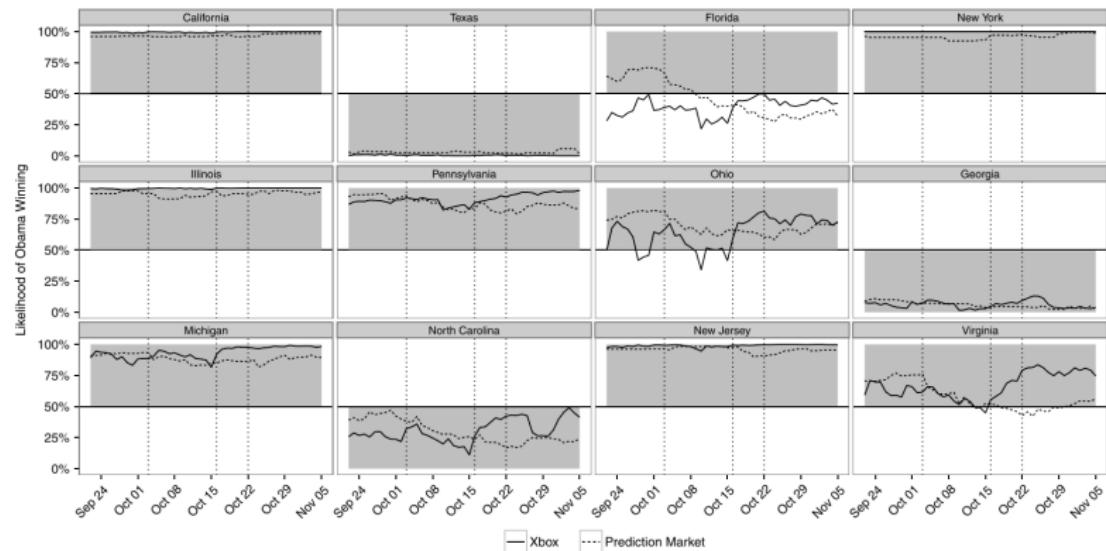


Figure 6: Likelihood of Obama victory

Wang, Rothschild, Goel, and Gelman 2015. Forecasting elections with non-representative polls. International Journal of Forecasting, 31 (3).

Extrapolating from Wikipedia

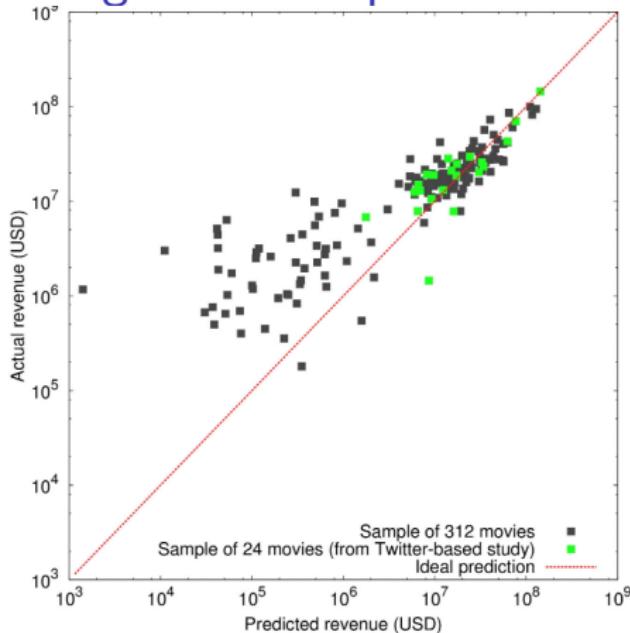


Figure 5. First weekend box office revenue in the U. S. against its predicted value by the Wikipedia model at $t = -30$ days. Green dots are representing the smaller sample of 24 movies common in Twitter and Wikipedia studies, and black dots are movies from the 2010 sample of 312 movies. Note that negative predicted revenues for some of the very unpopular movies could not be shown in the logarithmic scale.
doi:10.1371/journal.pone.0071226.g005

Figure 7: Predicting movie revenues by editor's activity

Mestyán, M., Yasseri, T., and Kertész, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS ONE 8(8):e71226.

A healthy dose of skepticism

It's Difficult to Make Predictions, Especially About the Future

Understand crowd-source directed behaviour

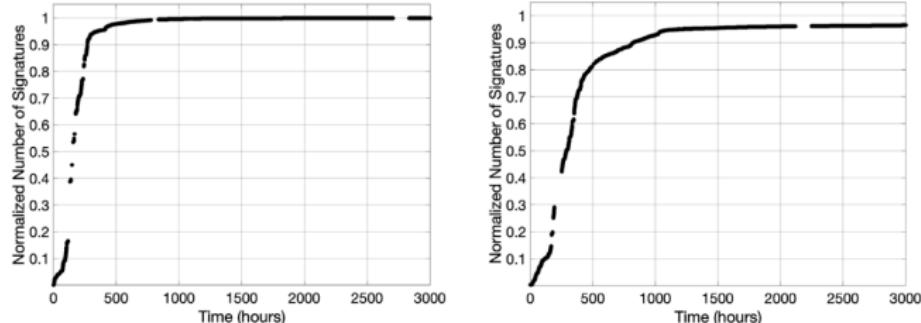


Figure 3 Growth of the number of signatures for 2 example petitions. Left panel: 'Stop the badger cull' and right panel: 'Reconsider West Coast Mainline franchise decision'.

Figure 8: Cumulative number of signatures in two online petitions in the UK

Yasseri, T., Hale, S.A., and Margetts, H.Z. (2017). Rapid rise and decay in petition signing. EPJ Data Science 6(1):20.

Petition to improve coverage of Covid-19 tests in LATAM

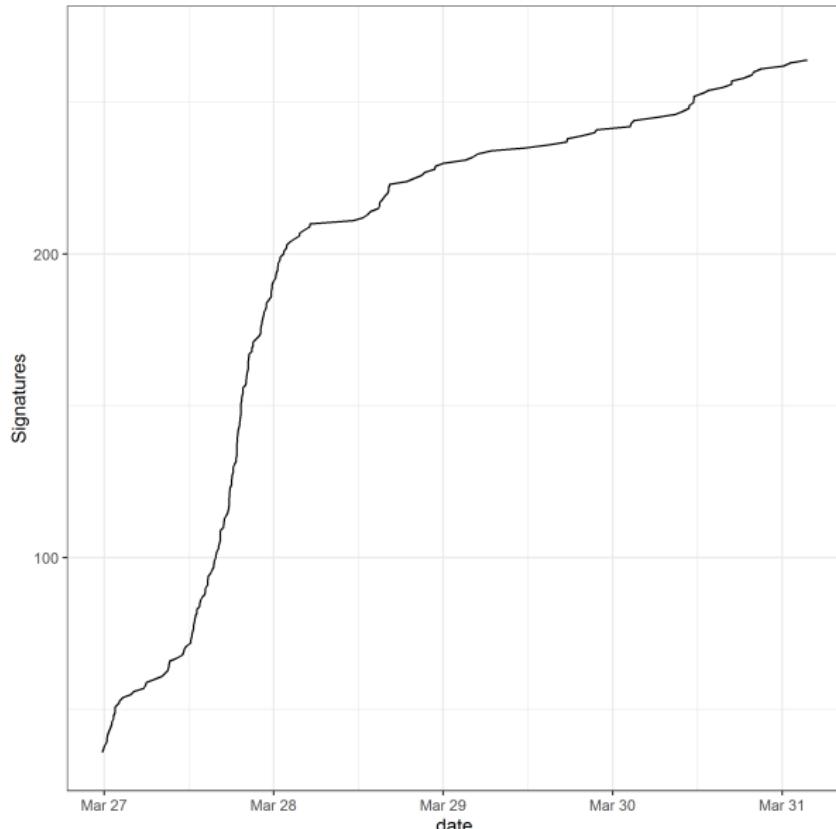


Figure 9: Please sign! <http://tinyurl.com/vp8ukpj>

Question time!



Think about a way in which you could use crowd-sourcing in your own research.

1. Which platform would you use?
2. Which challenges do you foresee?
3. Special measures to protect privacy?

Some magic sampling...



```
who <- sample(n, 3, replace = F)
n <- n[!n %in% who]
print(who)

## [1] "Momoko"    "Alexander"  "Luca"
```

Online genealogies

A genealogy is the history of a population

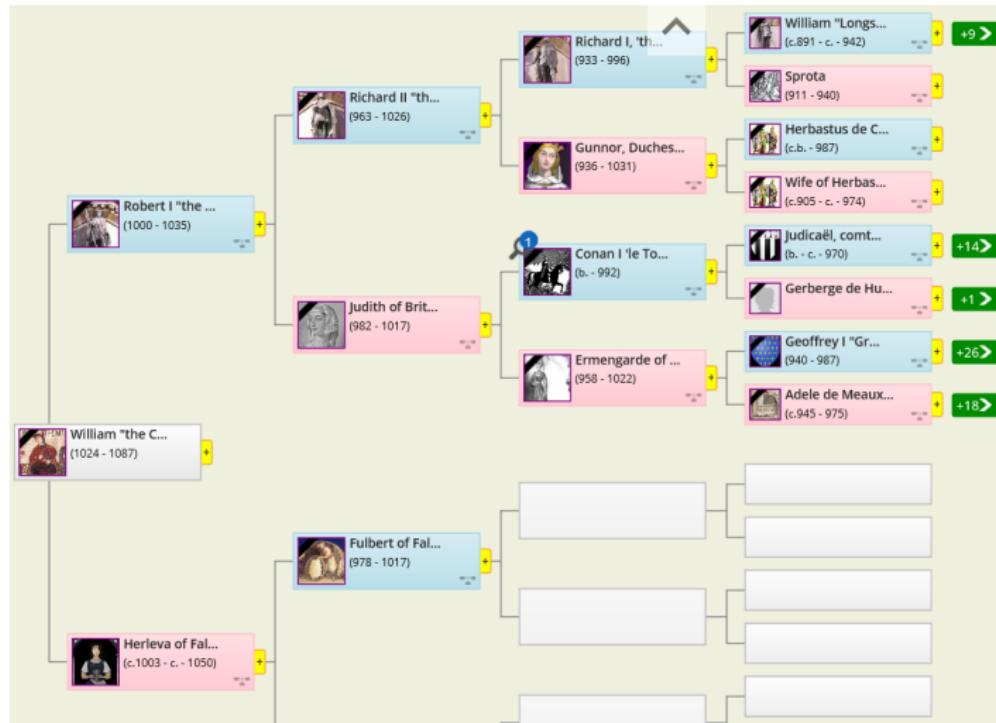


Figure 10: A Geni.com family tree

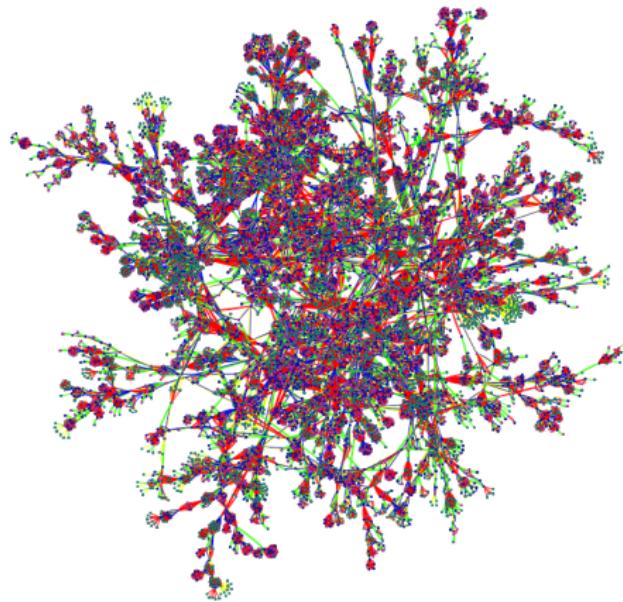


Figure 11: Cool dataviz, but what do we learn from the data?

Fire, M. and Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. ACM Trans. Intell. Syst. Technol. 6(2):28:1–28:22.

Demographic measures from online genealogies

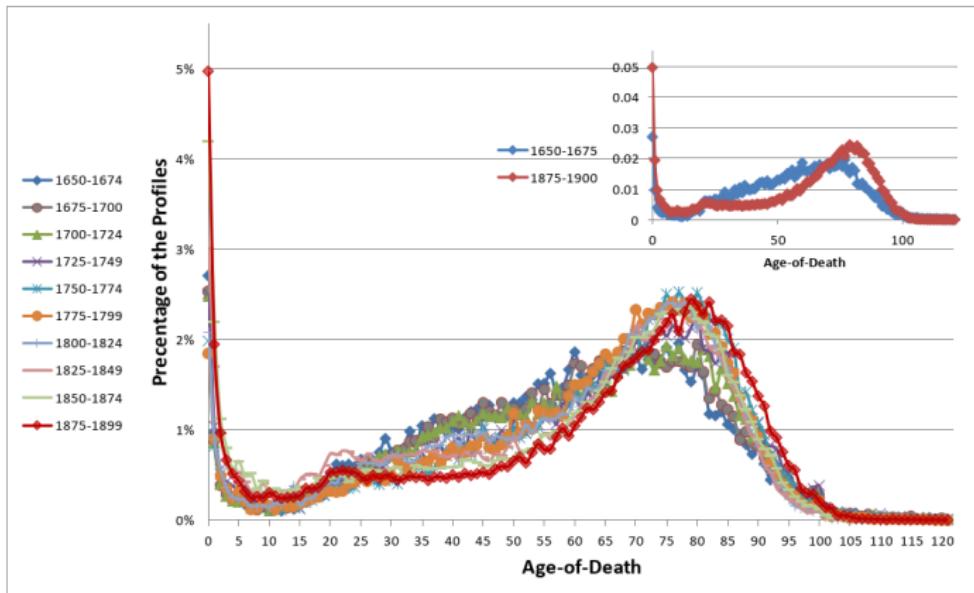


Figure 12: Lifespan Variation ('world')

Fire, M. and Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. ACM Trans. Intell. Syst. Technol. 6(2):28:1–28:22.

Demographic measures from online genealogies

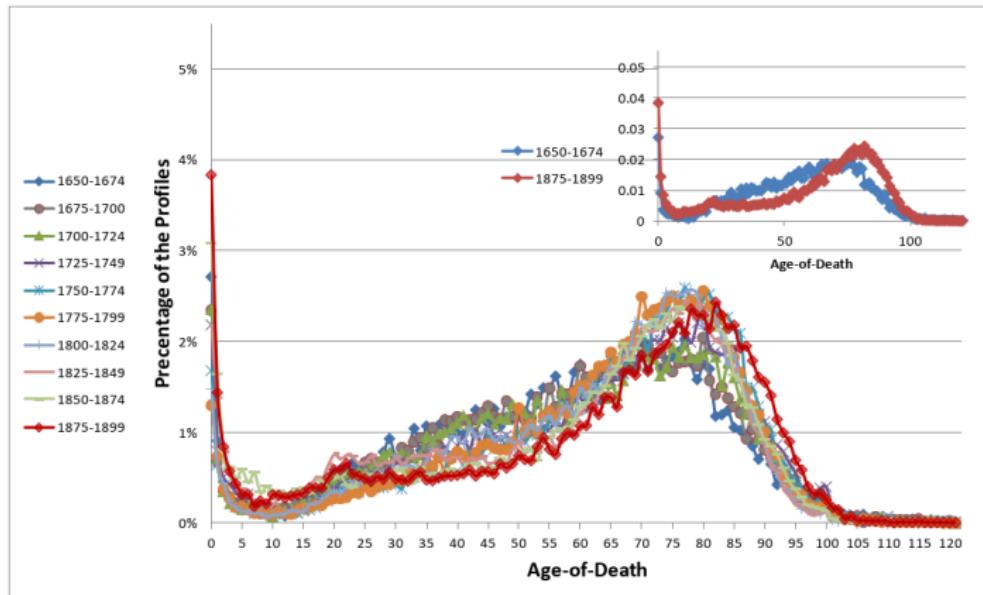


Figure 13: Lifespan Variation ('USA') - what's the difference??

Fire, M. and Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. ACM Trans. Intell. Syst. Technol. 6(2):28:1–28:22.

Geni.com: a social network for genealogists



William "the Conqueror" FitzRobert, Duke of Normandy, King of England MP

French: Roi d'Angleterre Guillaume FitzRobert, le Conquérant

Gender: Male
Birth: October 14, 1024
Château de Bayeux, Falaise, Calvados, Normandie, France

Death: September 09, 1087 (62)
Prieuré de Saint-Gervais, Rouen, Seine-Maritime, Haute-Normandie, France (Wounds suffered at the siege of Mantes)

Place of Burial: Abbaye Saint-Étienne, Abbaye aux Hommes, Caen, Calvados, Basse-Normandie, France

Immediate Family:

- Son of Robert I "the Magnificent", Duke of Normandy and Herleva of Falaise
- Husband of Matilda of Flanders
- Father of Robert II "Curthose", Duke of Normandy; Adeliza de Normandie, Princess of England; William II "Rufus", King of England; Cecilia, Abbess of Holy Trinity, Richard and 5 others
- Brother of Adelaide of Normandy, Countess Of Aumale
- Half brother of Robert de Mortagne, Earl of Cornwall; Odo, Bishop of Bayeux; Jeanne de Conteville; Rohesia deConteville; Muriel de Conteville and 2 others

Matches ?

[Research this Person](#)

[Contact Profile Managers](#)

[View Tree](#)

[Edit Profile](#) ?

Figure 14: Everyone's relative

Built on top of (private) genomic data

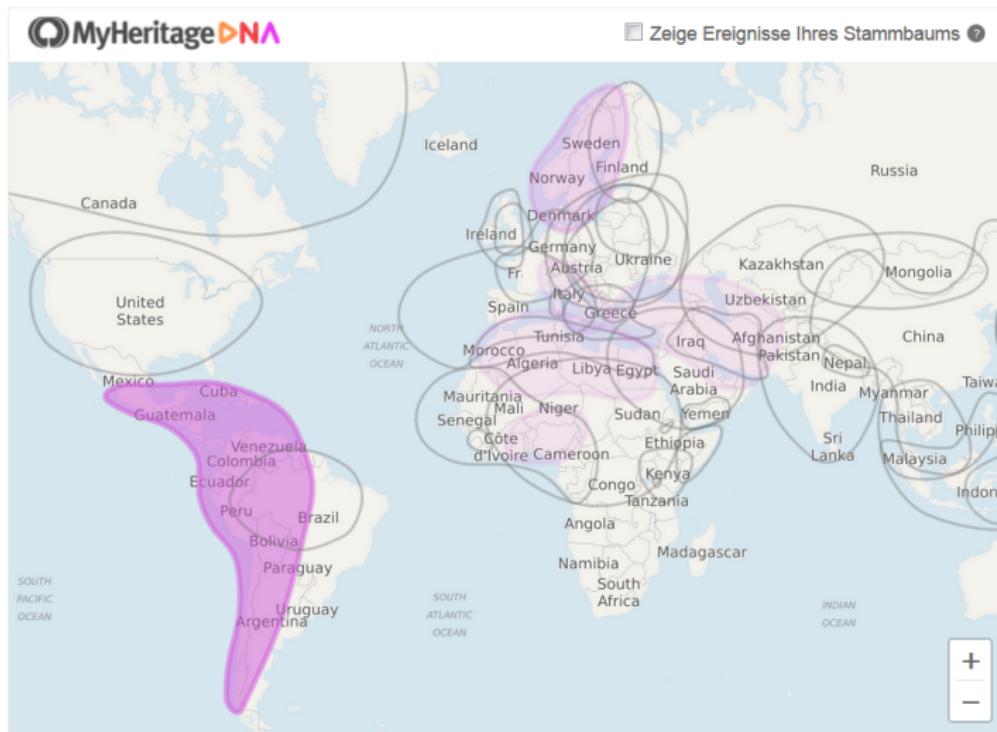


Figure 15: 'Ethnicity' estimates

Our example: Familinx data

1. Genealogy-driven social media data
2. Goal: register entire population of the world
3. 86M unique profiles over last 400 years
4. Curated, with quality checks
5. Geo-coded events - 55% Europe; 30% North America

Kaplanis, J., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360(6385):171–175.

Geographic distribution in Familinx

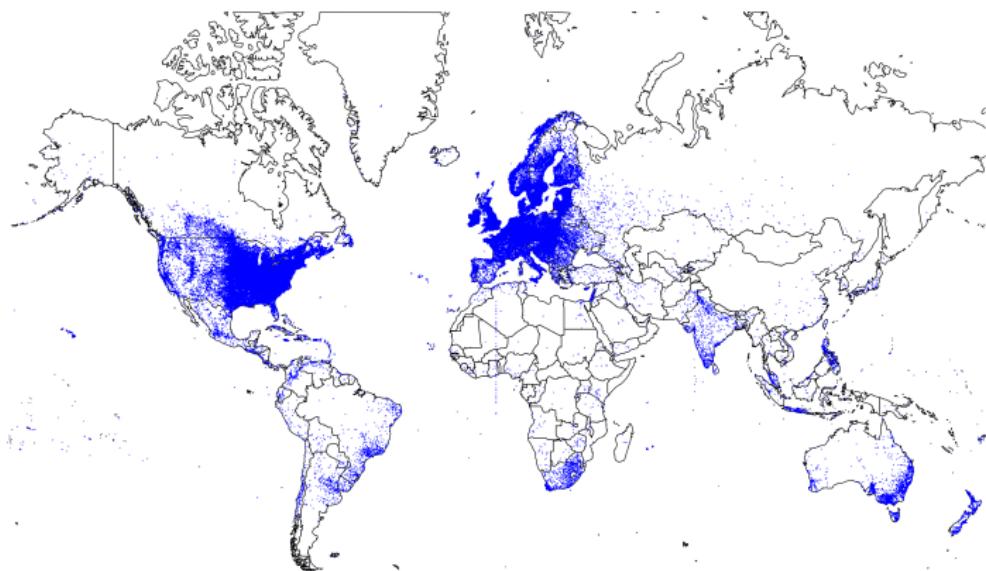


Figure 16: Birth events worldwide

Population alive by age group (all countries)

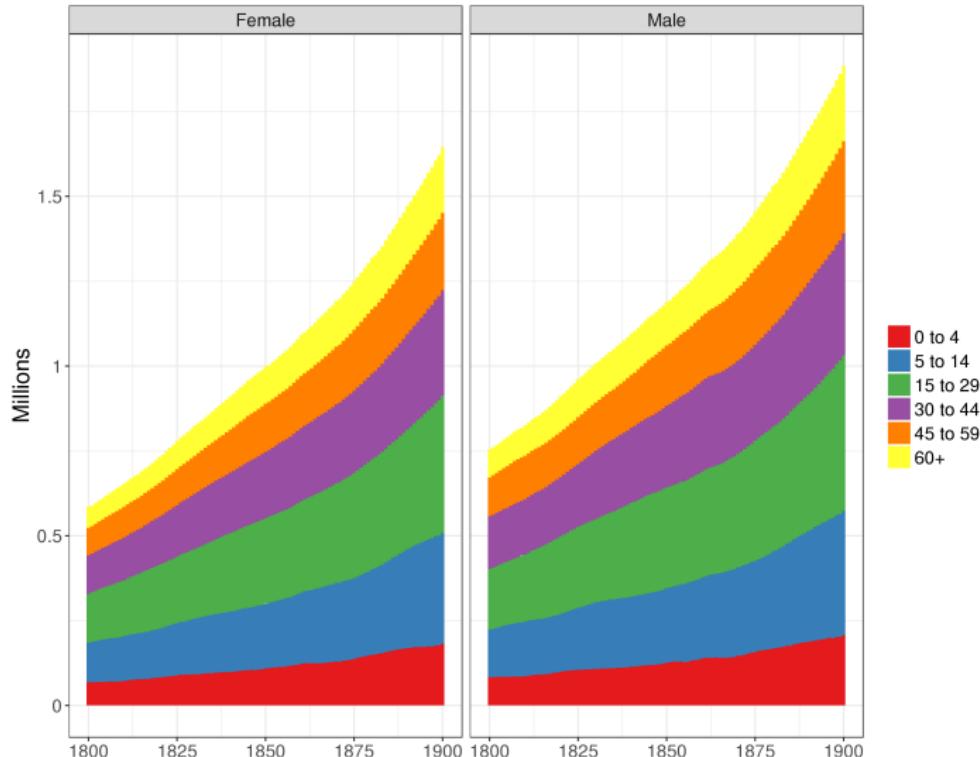
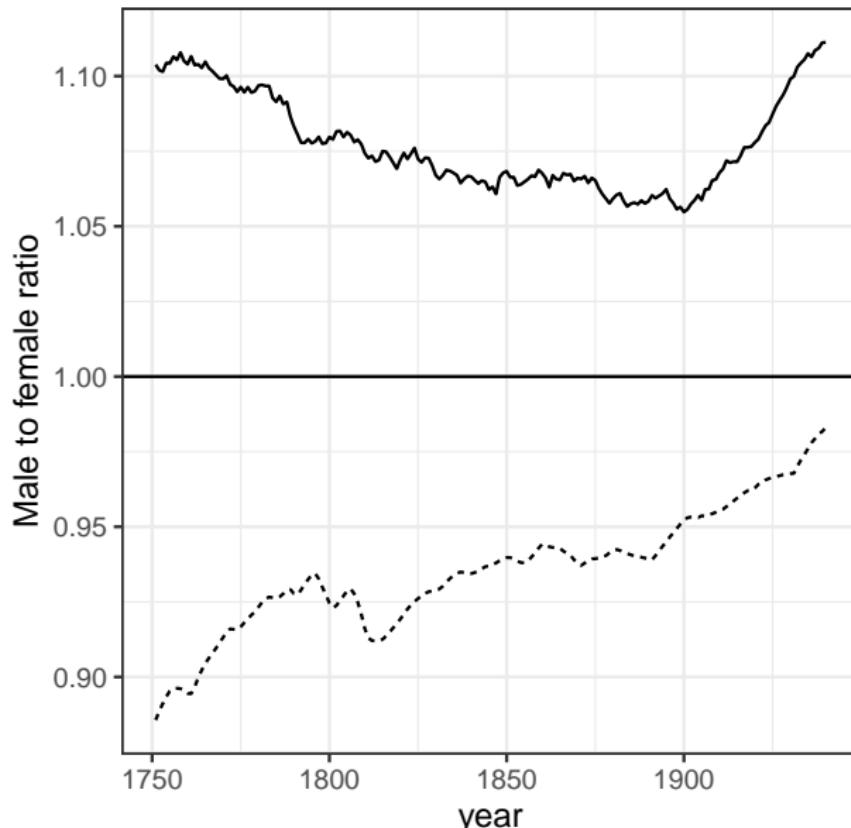


Figure 17: Yearly censuses from Familinx data

Male bias in online genealogies (Familinx)



source — Familinx --- HMD

Question time!



1. Why are there so many men?
2. Can we fix this?
3. What data would we need?

Some magic sampling...



```
who <- sample(n, 3, replace = F)
n <- n[!n %in% who]
print(who)
```

```
## [1] "Ilgi"    "Daniel"   "Lara"
```

The data generating process

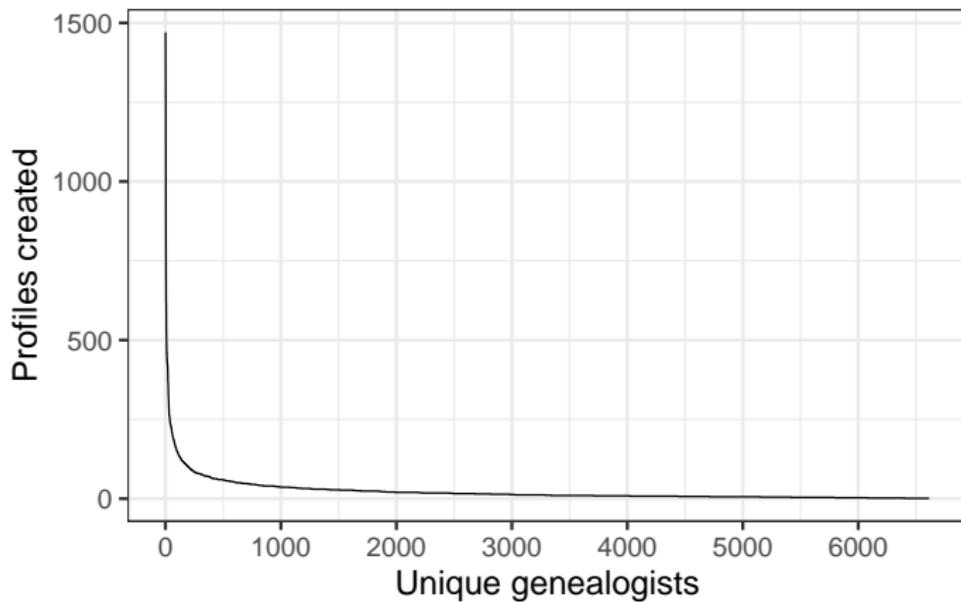
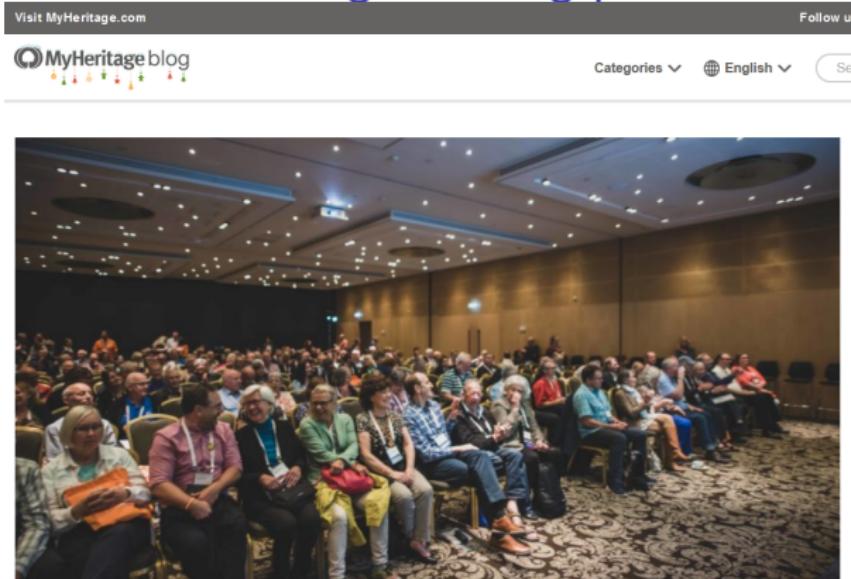


Figure 19: Power law function of genealogists

A closer look at the data generating process



MyHeritage LIVE 2019 Recap

By Esther · September 12, 2019 · Events And Webinars

Like 106

Comments 3

f Share

t Tweet

e Email

Share

Figure 20: The crowd-sourcers, median age: 65

<https://blog.myheritage.com/2019/09/myheritage-live-2019-recap/>

The Swedish sample

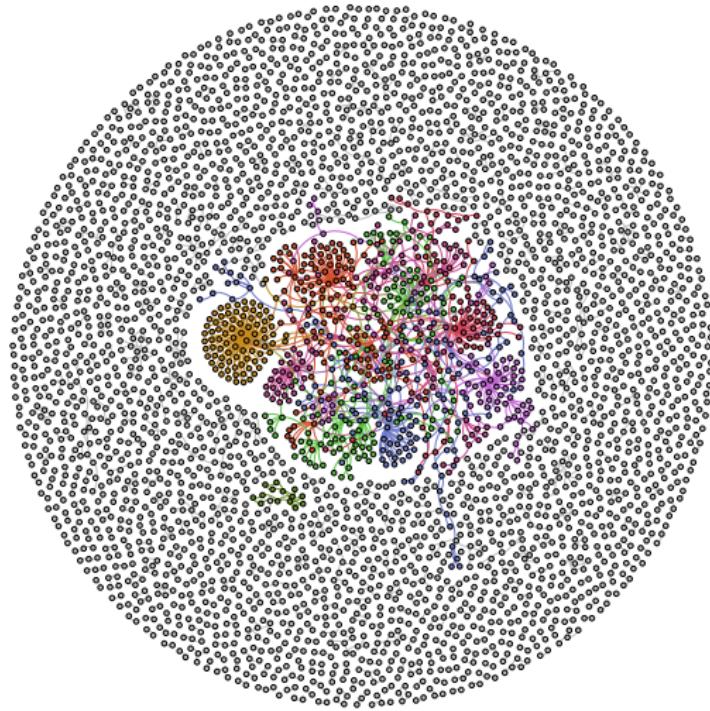


Figure 21: Subsetting a genealogical network

Exploring FamiliNX

```
library(data.table); library(tidyverse); library(knitr)

# Read sample FamiliNX data using data.table
fread("../Assignment/Data/sweden_genealogy.csv") %>%
  # To get a sense of the structure of the data:
  select(profileid, father, mother, birth_year, death_year)
  slice(1:4) %>%
  kable()
```

profileid	father	mother	birth_year	death_year
136	NA	NA	NA	NA
264	57536639	69836161	1875	NA
708	83768131	48140261	NA	NA
722	NA	NA	1694	1767

Geographic distribution in our sample

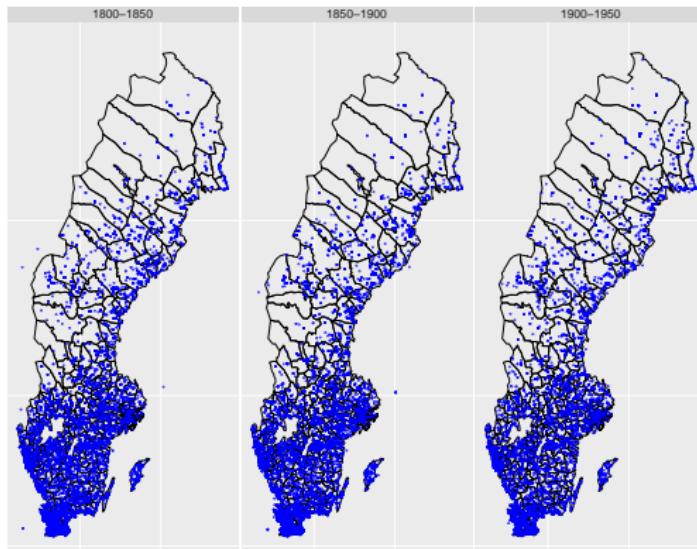


Figure 22: Birth events in Sweden over time

Diego Alburez-Gutierrez (2020)

Correcting bias: an example using post-stratification weights

Starting point: Online genealogies are a non-representative sample of real-world genealogies.

- ▶ Online genealogies \neq offline genealogies
- ▶ Unknown 'weights' - derive from comparison to trusted sources
- ▶ Understand data-generating process

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

Enumerator - death events

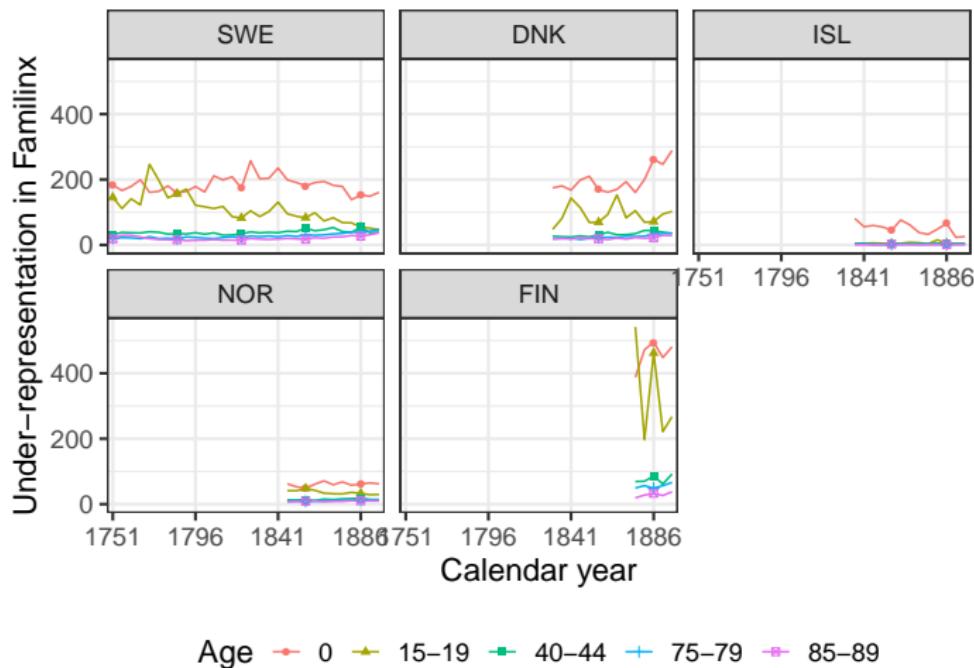


Figure 23: Observed deaths in four countries

Denominator or exposure

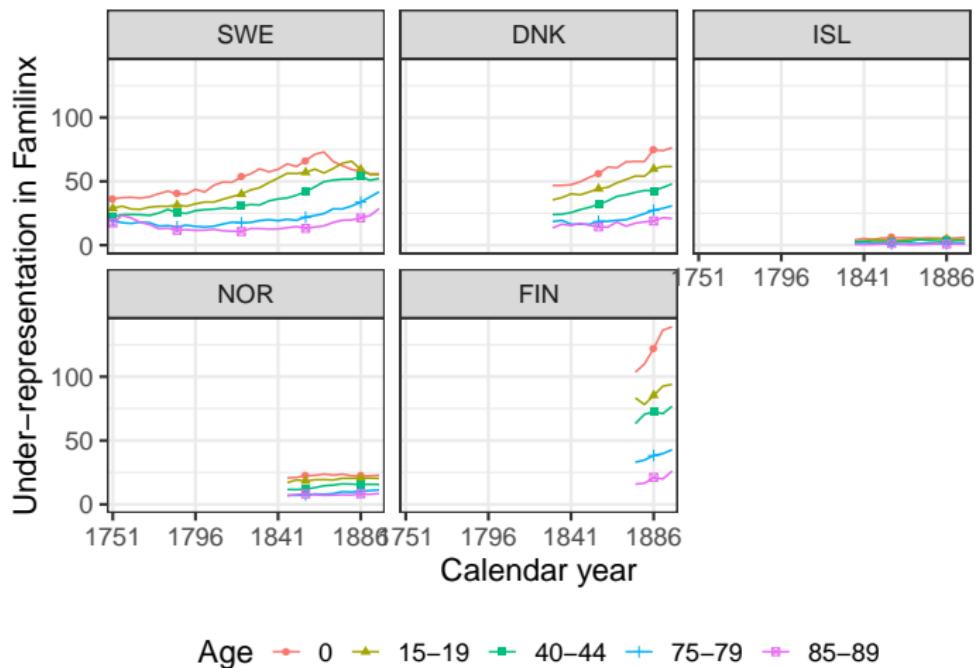


Figure 24: Population alive by age and sex in four countries

'Corrected' demographic rates (Sweden)

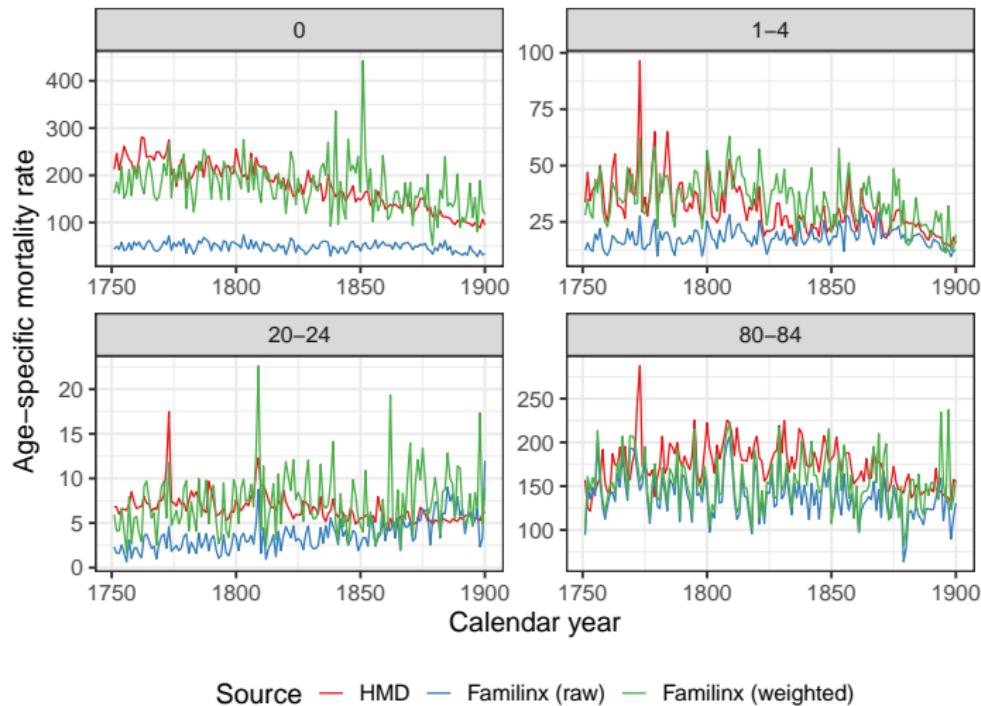


Figure 25: Weighted mortality rates

Suggested homework

- ▶ Read Sofia Gil's tutorial on using the FB Marketing API:
https://github.com/SofiaG11/Using_Facebook_API
- ▶ Explore website:
<https://www.digitalgendergaps.org/data/?report=2020-03-02>
- ▶ Start working on Exercise 1 from the final assignment