

# 1 - Introduction to Digital Demography

Diego Alburez-Gutierrez  
MPIDR

European Doctoral School of Demography 2019-20

30/03/2020



MAX-PLANCK-INSTITUT  
FÜR DEMOGRAFISCHE  
FORSCHUNG

MAX PLANCK INSTITUTE  
FOR DEMOGRAPHIC  
RESEARCH

# Agenda

1. Course setup and final assignment
2. Introduction to digital demography
3. Data and approaches
4. Current topics

# Hi from the Lab of Digital and Computational Demography!



Figure 1: DiCoDe

## Course setup

## Week's schedule

- ▶ Monday 11:30-13.00 - "Introduction to digital demography"
- ▶ Tuesday 11:30-13.00 - "Crowd-sourced online data"
- ▶ Wednesday 11:30-13.00 - "Digital trace data"
- ▶ Thursday 15:00-17.00 - "Computational approaches"
- ▶ Friday - No class

# Technical setup

- ▶ Online lectures, whenever possible at:
  - ▶ [https://meet.jit.si/EDSD20\\_digital\\_demography](https://meet.jit.si/EDSD20_digital_demography)
- ▶ Course materials (syllabus, presentations, readings, assignment)
  - ▶ [https://github.com/alburezg/EDSD20\\_digital\\_demography](https://github.com/alburezg/EDSD20_digital_demography)
- ▶ Requirements
  - ▶ Running installation of RStudio
  - ▶ packages: 'tidyverse', 'data.table'

# Final assignment

## Goal

- ▶ Hands-on experience using crowd-sourced digital data.
- ▶ Sample of Familinx database: user-generated genealogical records

## Exercises

1. Compute: historical lifespan in Sweden
2. Discuss: lifespan and life expectancy
3. Evaluate: bias in the online genealogies

*See full instructions in syllabus!*

# Evaluation

Assignments due Friday April 3 at midnight (CET):

1. A written report (Word document or pdf file).
2. The R scripts used to produce the empirical results.

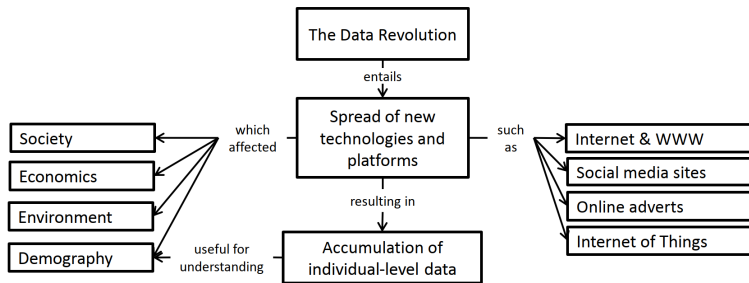
*See full instructions in syllabus!*



## Introduction to Digital Demography

# What is 'digital' demography?

- ▶ Digital vs analogue
- ▶ Online vs offline
- ▶ 'Big' vs 'small' data



**Fig. 1** The Data Revolution and new sources of data for demographic analysis.

# Digital data sources for demographic research

1. Crowd-sourced online data
2. Digital Trace Data (online and offline)
3. Simulations (made-up data)

## 'Big data' is not new data



Figure 2: Demographer collecting Big Data for the 1925 US census

## You are where you E-mail: Using E-mail Data to Estimate International Migration Rates\*

**Emilio Zagheni**

Max Planck Inst. for Demographic Research  
Konrad-Zuse-Str. 1, Rostock, Germany  
zagheni@demogr.mpg.de

**Ingmar Weber**

Yahoo! Research Barcelona  
Av. Diagonal 177, Barcelona, Spain  
ingmar@yahoo-inc.com

### ABSTRACT

International migration is one of the major determinants of demographic change. Although efforts to produce comparable statistics are underway, estimates of demographic flows are inexistent, outdated, or largely inconsistent, for most countries. We estimate age and gender-specific migration rates using data extracted from a large sample of Yahoo! e-mail messages. Self-reported age and gender of anonymized e-mail users were linked to the geographic locations (mapped from IP addresses) from where users sent e-mail messages over time (2009-2011). The users' country of residence over time was inferred as the one from where most e-mail messages were sent. Our estimates of age profiles of migration are qualitatively consistent with existing administrative data sources. Selection bias generates uncertainty for estimates at one point in time, especially for developing countries. However, our approach allows us to compare in a reliable way migration trends of females and males. We document the recent increase in human mobility

### Author Keywords

Demographics, Migration, Mobility, E-mail data

### ACM Classification Keywords

J.4 Social and Behavioral Sciences; H.4.3 Communications Applications

### General Terms

Experimentation, Human Factors

### INTRODUCTION

International migration is an important driver of demographic growth in many countries [13], and a major source of uncertainty in demographic projections carried out by the United Nations [14]. Migrations have relevant social, economic, and environmental consequences that are felt for decades in both sending and receiving countries. Although there is growing interest in quantifying international

## Figure 3: Using email data

# International migration



Figure 4: Conditional probabilities of migration using IP data

Bogdan State, Ingmar Weber, and Emilio Zagheni. 2013. Studying inter-national mobility through IP geolocation. In Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13).

# Current topics in digital demography

1. Methodological developments
2. Understand internet users and online use
3. Migration (internal and external)
4. Mortality and morbidity
5. Online and offline fertility dynamics

# 1. Methodological developments

1. Inference from non-representative samples
2. Understand and address online bias
3. Nowcast demographic processes



# 1. Big Data, Big Bias?

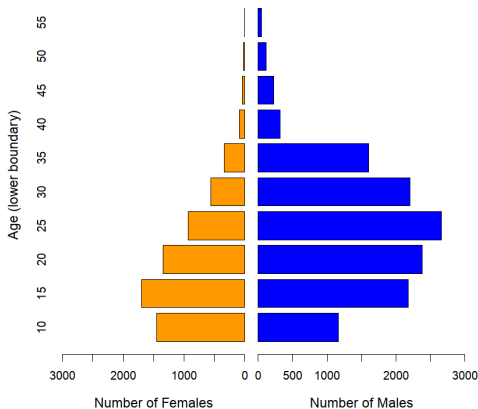


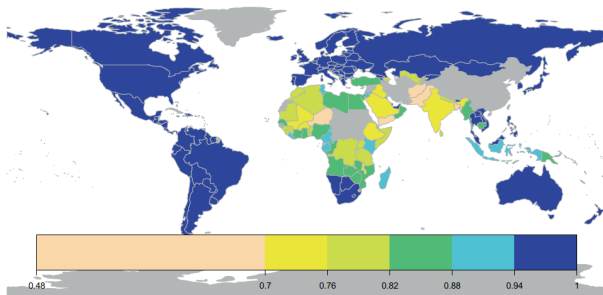
Figure 5: Looks a bit odd?

Zagheni, E., Garimella, V.R.K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from Twitter data. Paper presented at the 23rd International Conference

## 2. Understand internet users and online use

1. Infer demographics (age, sex, location, SE status, etc) from image and text
2. Track inequalities in online access
3. Consequences of platform use for users

## 2. The digital gender gap



(b) Mobile Phone Gender Gap Index predicted from Online model using the Facebook age 25-29 user Gender Gap Index

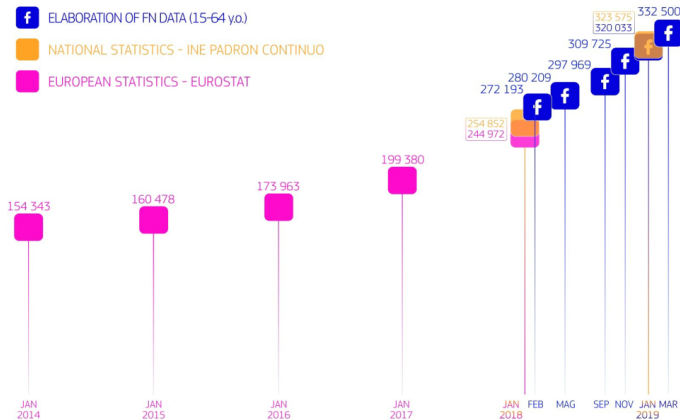
Figure 6: Inequalities in digital access

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development* 107:189–209.

### 3. Migration (internal and external)

1. Estimate flows and stocks
2. Mobility by subgroup (eg. undocumented, highly-skilled)
3. Cultural assimilation of immigrants

### 3. Quantifying international migration



**Fig 8. Stocks of international migrants and FN migrants from Venezuela in Spain.** The national and Eurostat statistics refer to all ages, while the FN-derived estimates refer to ages 15–64.

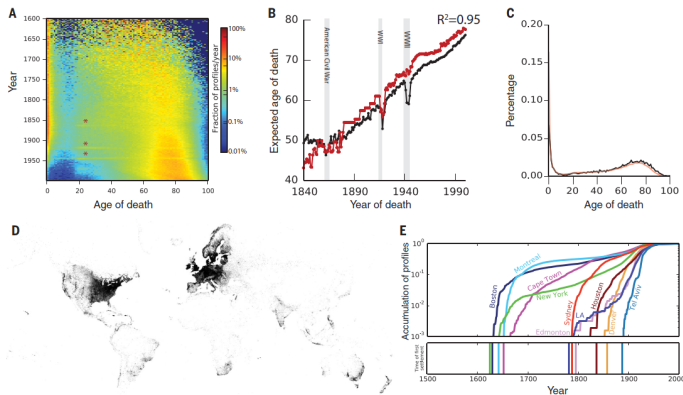
<https://doi.org/10.1371/journal.pone.0224134.g008>

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. PLOS ONE 14(10):e0224134. doi:10.1371/journal.pone.0224134.

## 4. Mortality and morbidity

1. “Historical” mortality estimates (obituaries, genealogies)
2. Enhanced data collection with IOT and online surveys
3. Morbidity from online behaviour (digital traces)

## 4. Long-term mortality patterns



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

## 5. Online and offline fertility dynamics

1. Estimate fertility from friendship networks
2. Online discourse around reproduction
3. Partnership formation and assortative mating (online dating)



# Estimating male fertility from FB

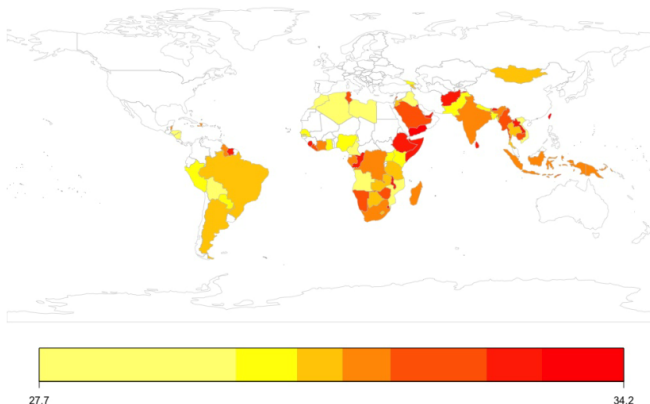


Figure 1: Prediction of Male MAC 2017 for countries without UN ground truth data.

# Homework

1. Download Familinx sample data for Sweden:  
[https://github.com/alburezg/EDSD20\\_digital\\_demography/tree/master/Assignment](https://github.com/alburezg/EDSD20_digital_demography/tree/master/Assignment)
2. Use the script in the `Assignment/R` directory to load it in your R global environment
3. Review the assignment instructions in the syllabus