Introduction to Data Analysis

HSE University and University of London Undergraduate Program in International Relations

Instructors: Georgy Tarasenko, Ioann Dovgopoliy

Logistics

- 16 seminars in the spring semester
  - 1st seminar - introduction
  - 9th seminar - midterm
  - 16th seminar - final project’s presentation

Logistics

- 6 **Home Assignments** (week 2, 4, 6, 10, 12, 14) - 25%
  - Grade for the HA can be changed if different compared to the grade for the ***Problem Set (checks whether the related HA was done independently)***
  - Bonus - completing tasks on DataCamp
- **Quizzes** (each week except 9 and 16) - 15%.
  - 5 min test at the beginning of a seminar. Problem Set can be randomly assigned instead of a quiz.

Logistics

- **Midterm** (week 9) - 25%
  - In-class lab - 7-10 problems sets (including bonus tasks), 120 min (content of the weeks 1-8).
- **Final project** - 35%
  - Group research project on the topic within IR (or social sciences) with application of the data analysis tools

Logistics

- Plagiarism
- Late Assignment Policies
- Communication:
  - Telegram chat: https://t.me/+_34nUYYmK35kNDcy
  - DM via e-mail: gtarasenko@hse.ru or Telegram: @georgy_tarasenko
  - All the materials can be accessed via GitHub (the link will be sent in the chat)

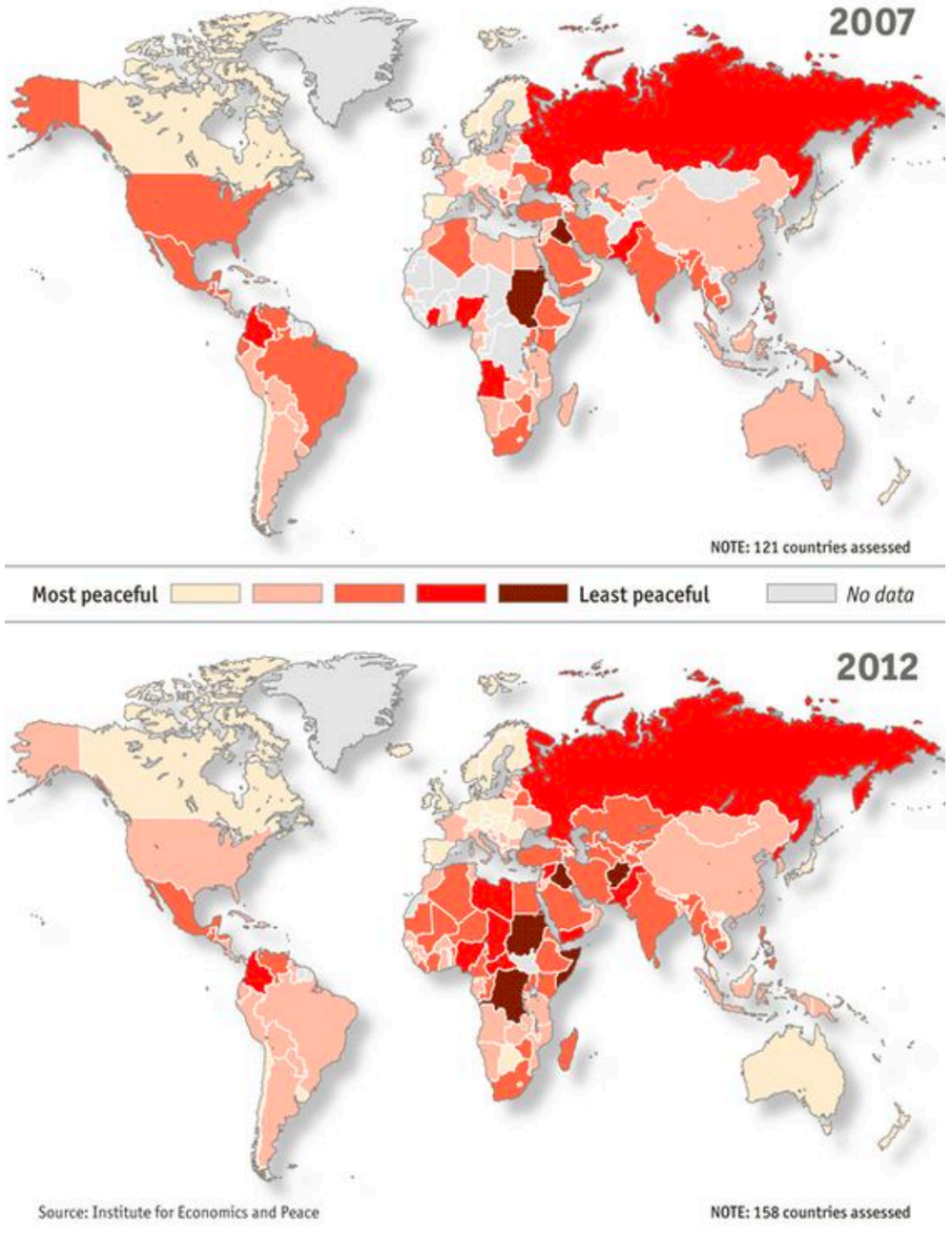
Computational International Relations

- a subfield within computational social science
- relies on the mining and processing of vast quantities of digital social footprint to study, model and explain world events

CIR: Main Tasks

- **Language and Text**
  - e.g. Natural Language Processing (NLP)
  - all forms of the information retrieval (e.g. web-scraping)
- **Geospatial analysis**
  - Geographic Information Systems (GIS)
    - Mapping

Global peace index



CIR: Main Questions

- **Explicit mathematic modelling** (econometrics, “IR-metrics”)
  - Correlation analysis
  - Regression analysis
  - Structural Equation Modelling etc.
- **Network Analysis**
- **Exploratory data analysis, visual analysis**

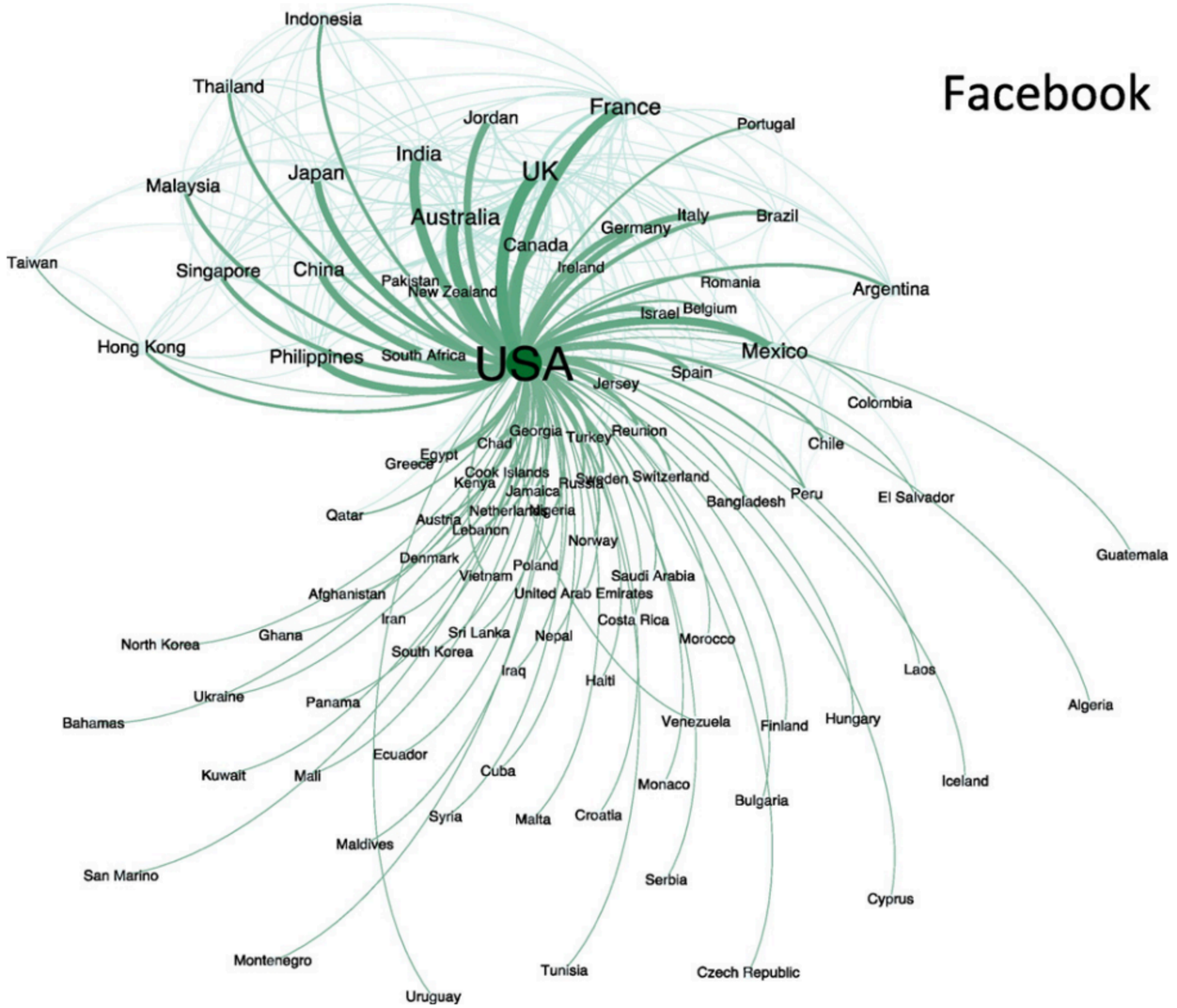


Fig. 1. Co-occurrence network based on Facebook data. *Note:* Node size varies by centrality. Line width indicates the frequency of co-occurrence.

Barnett, G. A., Xu, W. W., Chu, J., Jiang, K., Huh, C., Park, J. Y., & Park, H. W. (2017). Measuring international relations in social media conversations. *Government Information Quarterly*, *34*(1), 37-44.

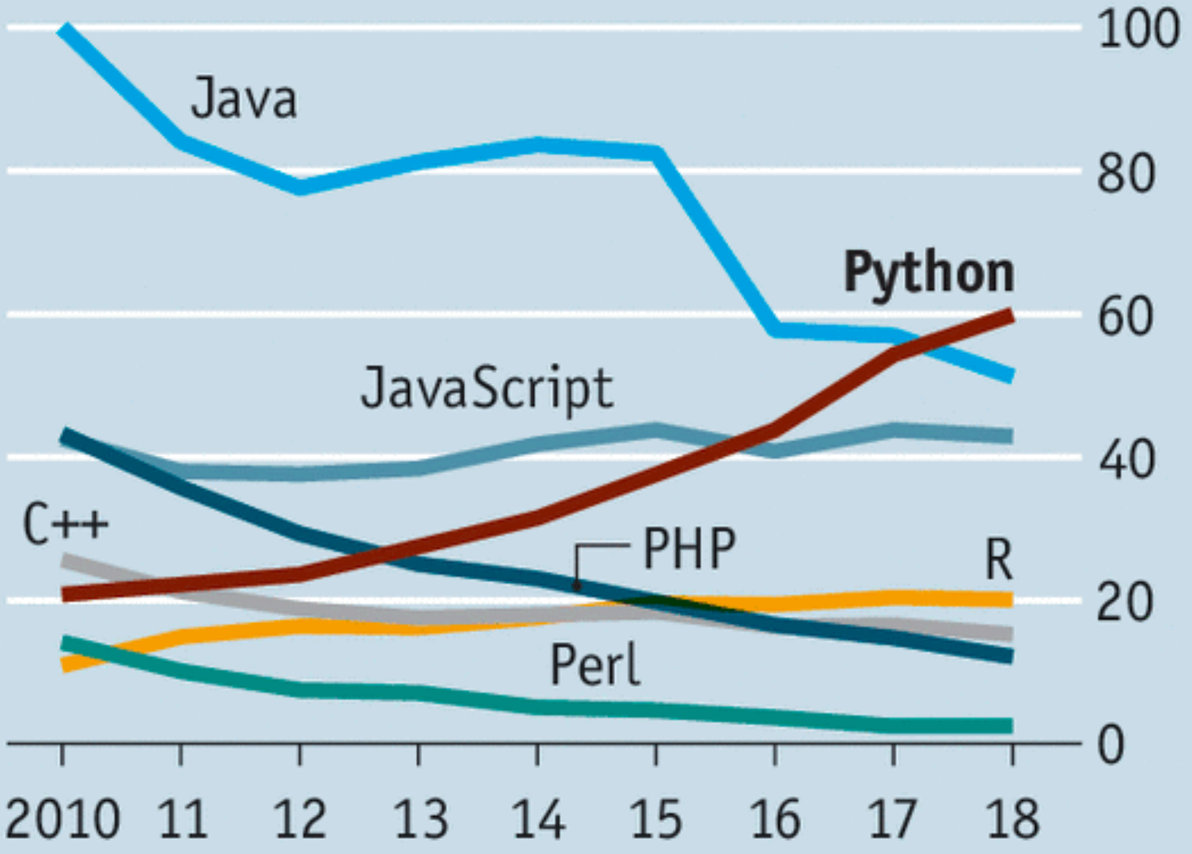
What is Python?

- Interpreted **high-level general-purpose** programming language invented in 1991 by a Dutch programmer Guido van Rossum
- Its embedded principles allowed Python to become **one of the most popular programming languages** in the world, highly preferable to use among social scientists
- The language's core **philosophy** includes aphorisms such as:
  - Beautiful is better than ugly
  - Explicit is better than implicit
  - Simple is better than complex
  - Complex is better than complicated
  - Readability counts.

Biggus uptickus

US, Google searches for coding languages

100=highest annual traffic for any language



Source: Google Trends

To-Do List for the next week

- Join the Telegram Chat: https://t.me/+_34nUYYmK35kNDcy
- **Install Anaconda Navigator (and launch Jupiter Notebook)**
  - Windows: https://www.youtube.com/watch?v=uOwCiZKj2rg
  - Mac OS: https://www.youtube.com/watch?v=iPsOCj_wKvY
- Read the article “Computational International Relations: What Can Programming, Coding and Internet Research Do for the Discipline?”: https://arxiv.org/pdf/1803.00105.pdf