DSPT
DATA SCIENCE PORTUGAL

The “D” in Data Science

Saghir Bashir

{i} ilustat
www.ilustat.com

Objectives

My objective is to encourage you to:

> Understand the limits and consequences of your data

Motivation?

> To often I see data being used inappropriately
> To often I see inappropriate data being used

Outline

Google Flu Trends

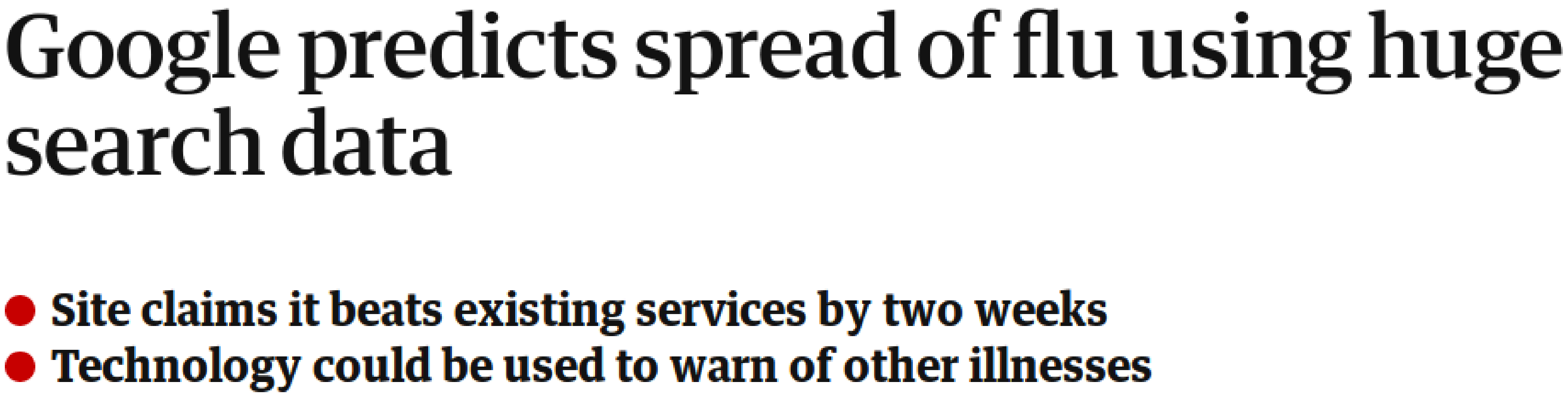
WHO Mortality Data

Translating Languages

Summary

The Guardian

13th November 2008



Google predicts spread of flu using huge search data

- Site claims it beats existing services by two weeks
- Technology could be used to warn of other illnesses

Source: https://www.theguardian.com/technology/2008/nov/13/google-internet

The Warning – Big News Headline

Big Data / Unicorn / Social Media / AI / …

saves humanity from

Disease / Dying / Fake News/ Bad stuff / …

The Warning – Big News Headline

Big Data / Unicorn / [illegible] a / AI / ...

saves [illegible] from

Disease / D [illegible] News/ Bad stuff / ...

RED ALERT

The Guardian

th November 2008

Google predicts uge search data

- Site claims i
- Technolo

DATA RED ALERT

Source: https://www.theg technology/2008/nov/13/google-internet

Google Flu Trends

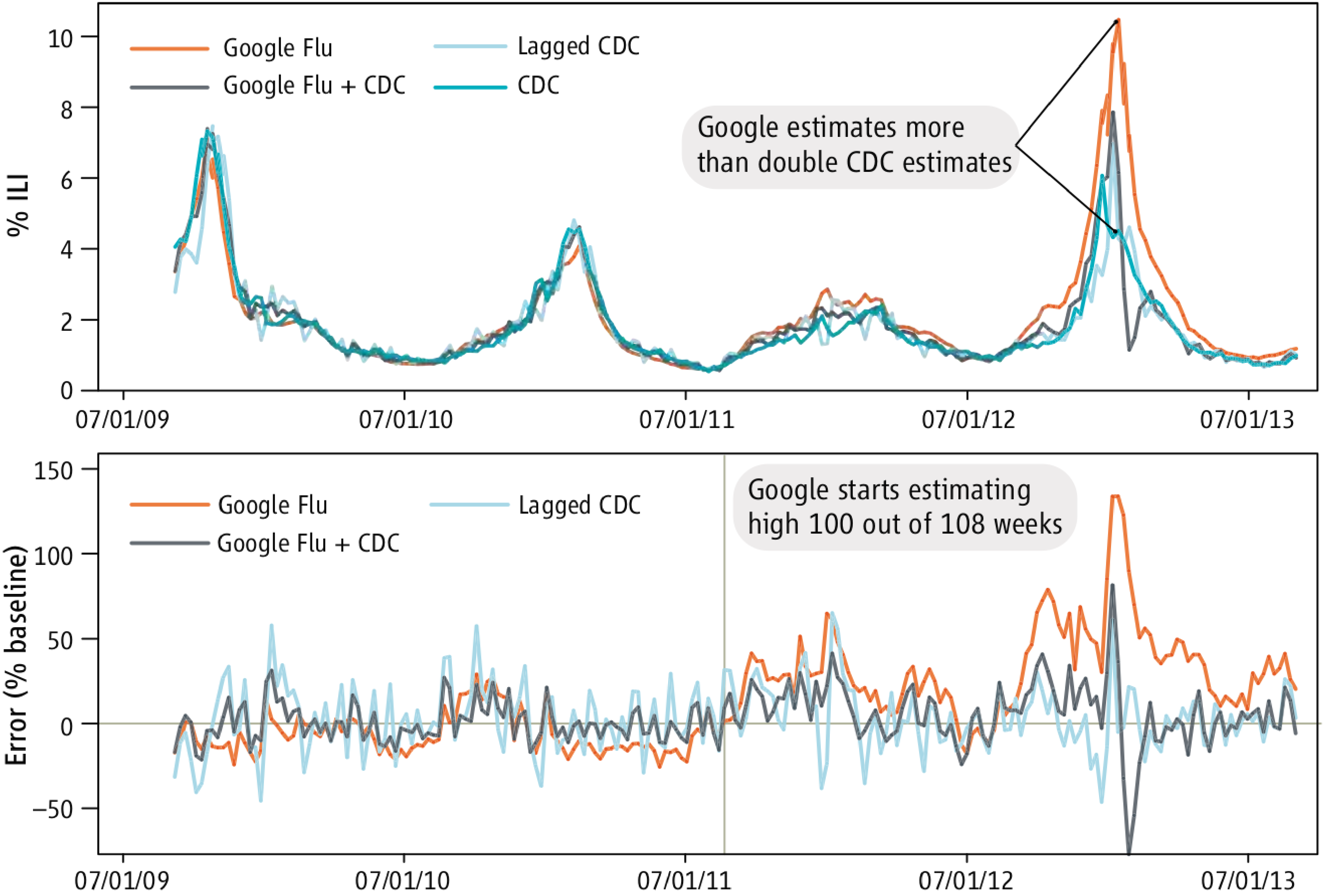
CDC
Centers for Disease Control and Prevention

> Best US predictions from Centre for Disease Control (CDC)

- Based on surveillance reports from labs across US
- By DESIGN - data and analyse give **reliable unbiased** predictions

> “Google searches” predict influenza like illness (ILI)

- Started with US and ended with 25 countries
- Found search terms **correlated with CDC data** (“training”)
- Then predicted using data from more recent searches
- Initially out performed CDC but then…



Source: http://science.sciencemag.org/content/343/6176/1203

My Main Issue…

- Essentially search terms were a “surrogate” for ILI
  - → Based on correlation with high CHANCE to find terms
- They are a bad surrogate
  - → Google tweaks algorithms (e.g. search box suggestions)
  - → People behaviour changes (e.g. news of bird flu epidemic)
  - → Correlation is not causation
- Surrogates have uses
  - → Blood pressure, cholesterol, … for cardiovascular events
  - → Well establish and widely recognised

We Could Use Official Health Data

CDC
Centers for Disease Control and Prevention

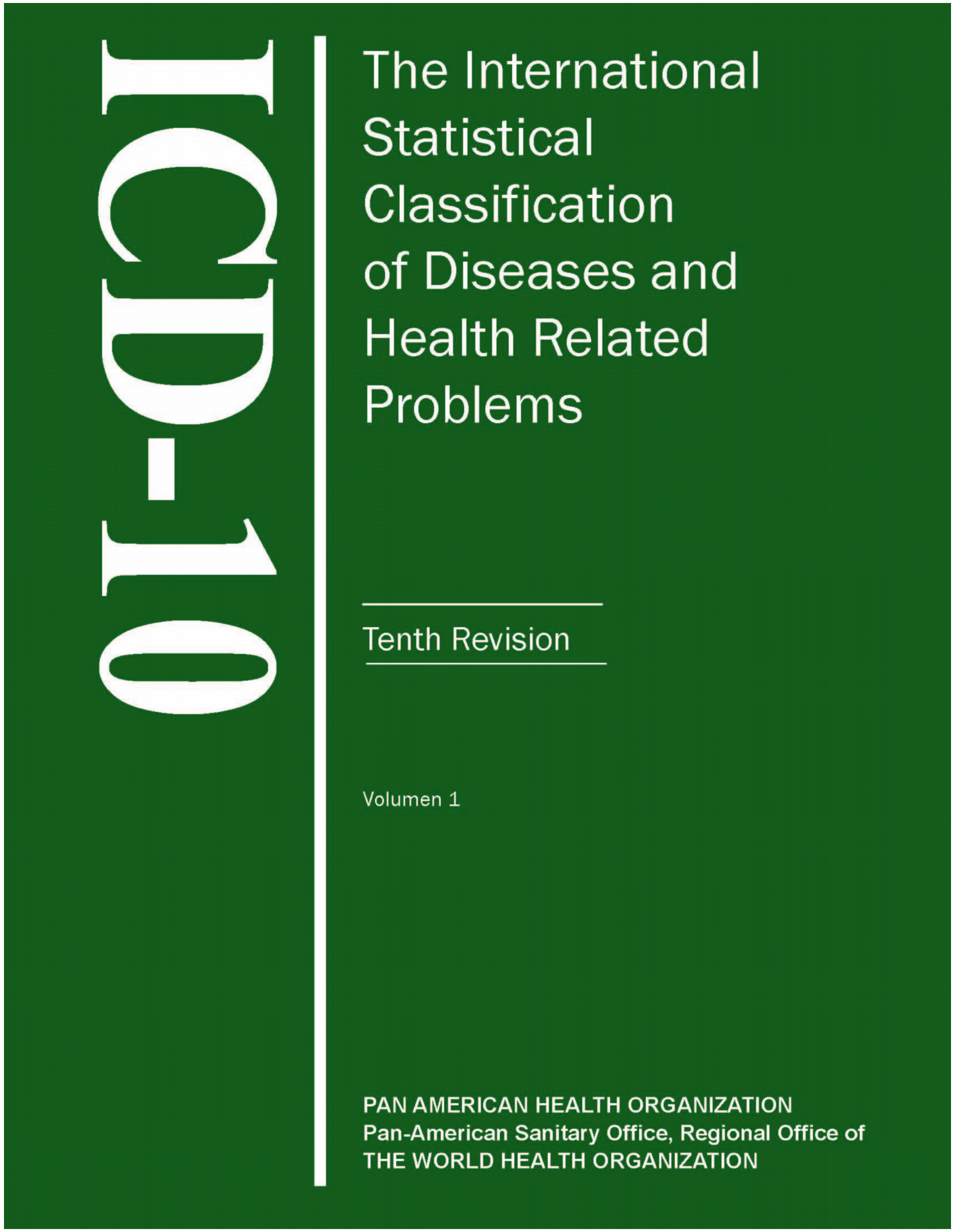
World Health Organization

Good Idea but...

There is always a story and data challenges...

- > **WHO Mortality Database**
- > **Data reported by country registration systems**
- > **Compilation of mortality data by:**
  - → **Age, sex, year and cause of death**
  - → **International Classification of Diseases (ICD)**

Source: http://www.who.int/healthinfo/mortality_data/en/



ICD-10

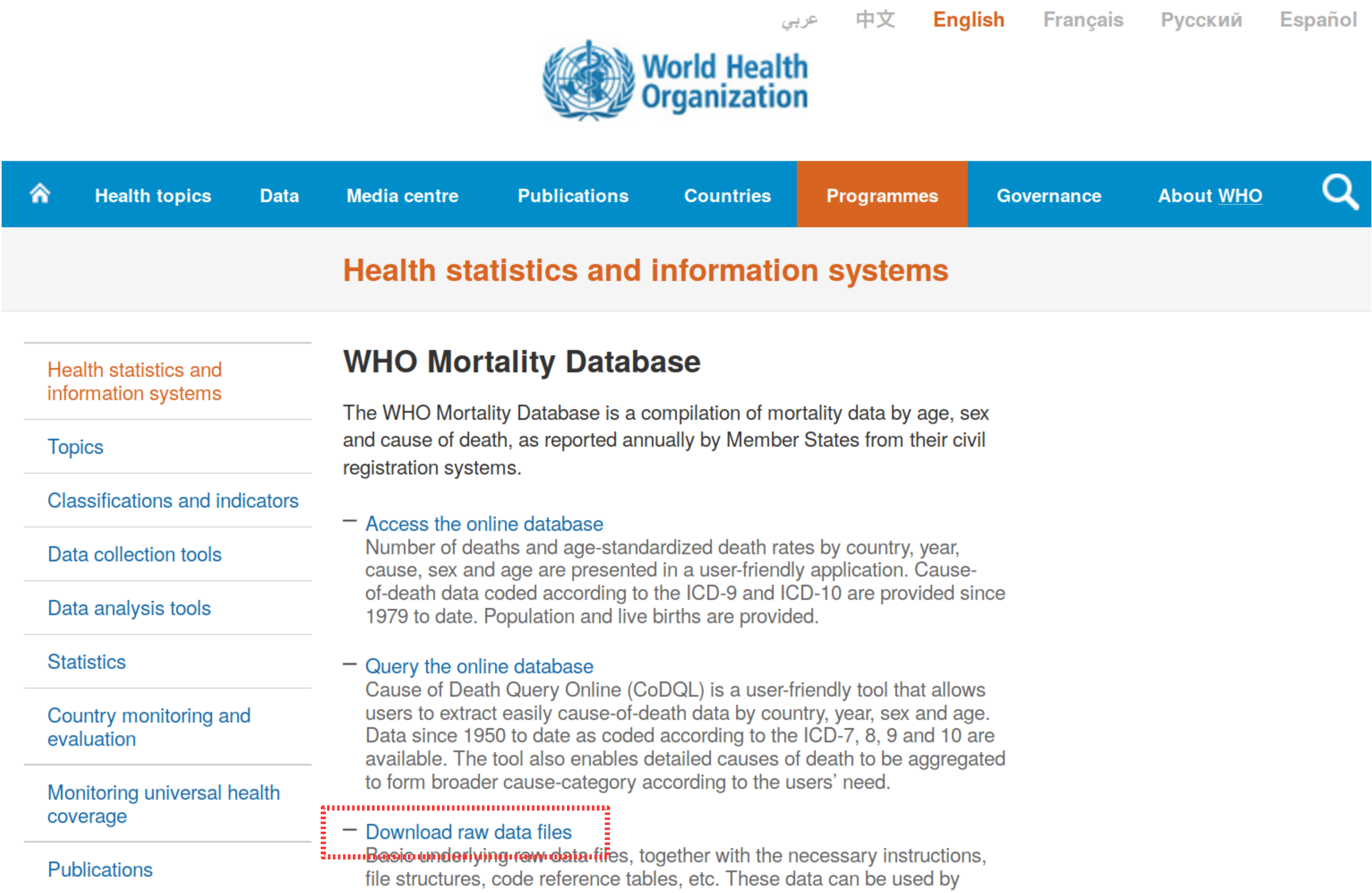
The International Statistical Classification of Diseases and Health Related Problems

Tenth Revision

Volumen 1

PAN AMERICAN HEALTH ORGANIZATION
Pan-American Sanitary Office, Regional Office of
THE WORLD HEALTH ORGANIZATION

ICD	Revised	Used
7	1955	1958 – 1967
8	1965	1968 – 1978
9	1975	1979 – 1994
10	1989	1995 –



Health statistics and information systems

- Health statistics and information systems
- Topics
- Classifications and indicators
- Data collection tools
- Data analysis tools
- Statistics
- Country monitoring and evaluation
- Monitoring universal health coverage
- Publications

WHO Mortality Database

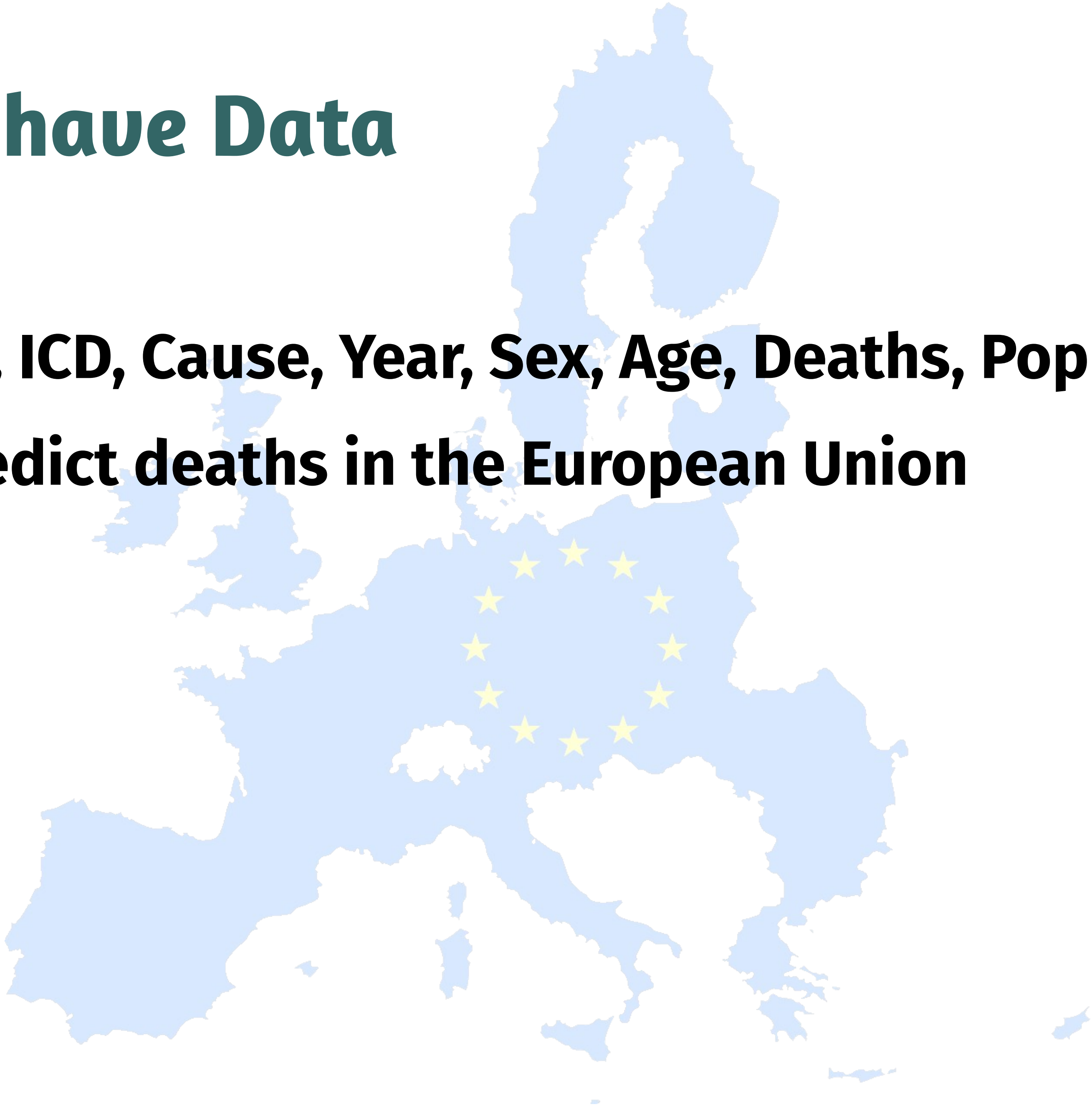
The WHO Mortality Database is a compilation of mortality data by age, sex and cause of death, as reported annually by Member States from their civil registration systems.

– Access the online database
Number of deaths and age-standardized death rates by country, year, cause, sex and age are presented in a user-friendly application. Cause-of-death data coded according to the ICD-9 and ICD-10 are provided since 1979 to date. Population and live births are provided.

– Query the online database
Cause of Death Query Online (CoDQL) is a user-friendly tool that allows users to extract easily cause-of-death data by country, year, sex and age. Data since 1950 to date as coded according to the ICD-7, 8, 9 and 10 are available. The tool also enables detailed causes of death to be aggregated to form broader cause-category according to the users' need.

– Download raw data files
Basic underlying raw data files, together with the necessary instructions, file structures, code reference tables, etc. These data can be used by

Yay we have Data



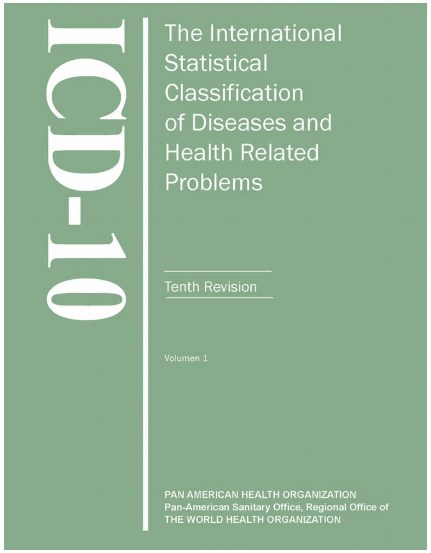
- Country, ICD, Cause, Year, Sex, Age, Deaths, Population
- Let’s predict deaths in the European Union
- But ...

Map source: https://commons.wikimedia.org/wiki/File:Flag_map_of_the_European_Union.png

Data Related Challenges (1)

- How do you define the European Union?
  - Start (1951) 6 countries – Now (2018) 28 countries
  - The UK has voted to leave (2019)
  - What is a fair comparison with the “EU average”?
- How do you define a country?
  - East and West Germany – Reunified in 1990
  - Czech Republic & Slovakia were formerly Czechoslovakia

Data Related Challenges (2)



ICD-10

The International Statistical Classification of Diseases and Health Related Problems

Tenth Revision

Volumen 1

PAN AMERICAN HEALTH ORGANIZATION
Pan-American Sanitary Office, Regional Office of
THE WORLD HEALTH ORGANIZATION

- How do you handle:
  - Partial coverage (e.g. cities only not rural)
  - ICD – Causes could be split or joined
  - Countries used ICD revisions at different times

- These issues have to be addressed by experts
  - Modelling (including ML & AI) **CANNOT** do this

“So what? I work with NLP!”

There is always a story and data challenges…

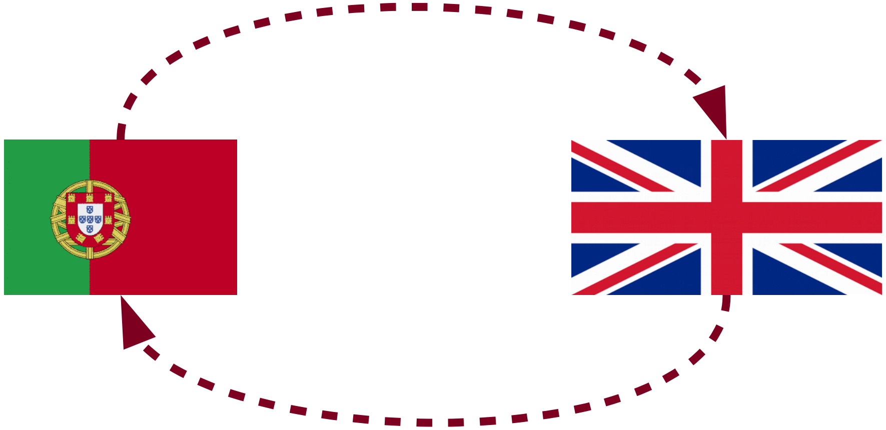
- **Natural Language Processing**
- **Sentiments Analysis**
- **Translation Engines**
- **…**

Languages & Translations

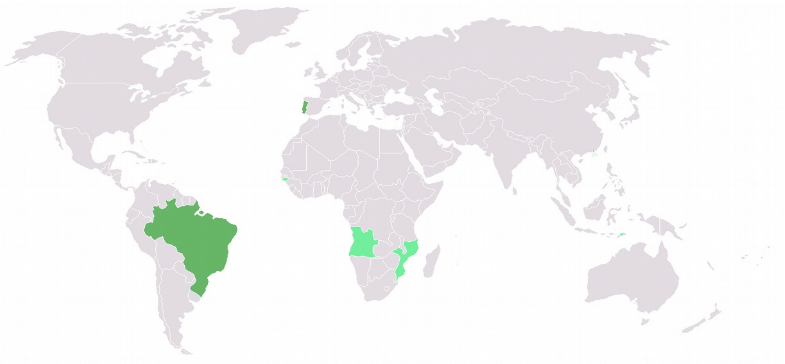
Imagine that we have **1 million** articles, books, regulations, *etc.* available in both Portuguese and English

> We plan to develop a translation system

> What potential data issues can you foresee?



Dialects & Styles



- > What is meant by "Portuguese" & "English"?
  - → Angolan, Brazilian, Mozambican, Portuguese...
  - → American, Australian, British, Caribbean, Indian, ...
  - → Even within each "language" there are differences
- > Does it make sense to mix articles, books, regulations, ...?
  - → Writing styles differ
  - → Legalese, technical, scientific, business, journalistic, ...

Map source: https://commons.wikimedia.org/wiki/File:Map-Lusophone_World.png

The Data?

> Where did the data come from and how?
  - → Randomly scraped from the web? Quality?

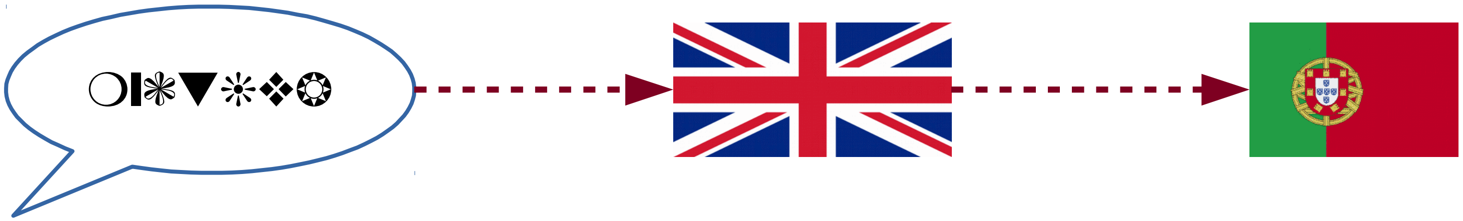
> Which periods are the translations from?
  - → Languages change over time
  - → How do you handle new words and phrases?

> How do you define “translation”?
  - → Word for word
  - → The author’s intention

Image source: https://commons.wikimedia.org/wiki/File:PessoaChapeu.jpg

Compromises can be made…

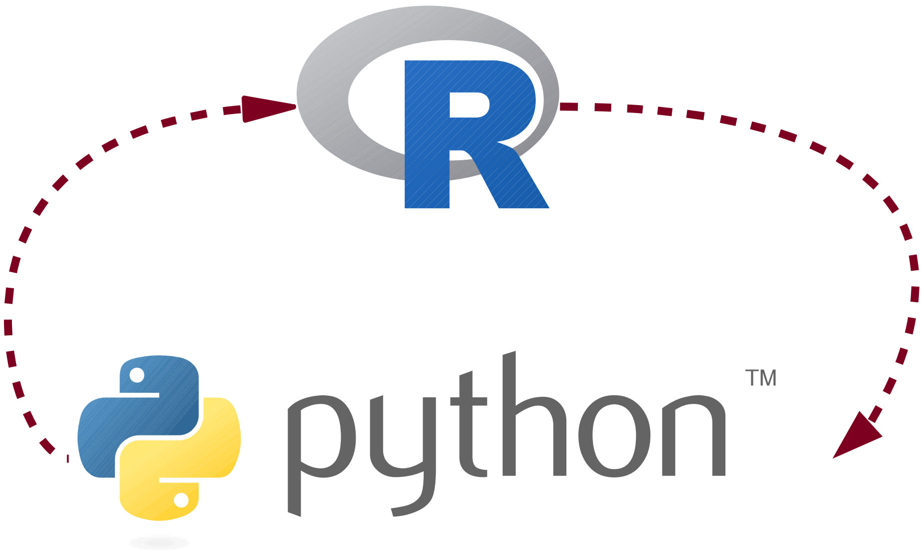
- Translating an “endangered” language
  - That is only translated into English but not Portuguese
- Translates “endangered” to Portuguese via English?
  - A rudimentary translation might be better than none
  - However users must be aware of the compromises



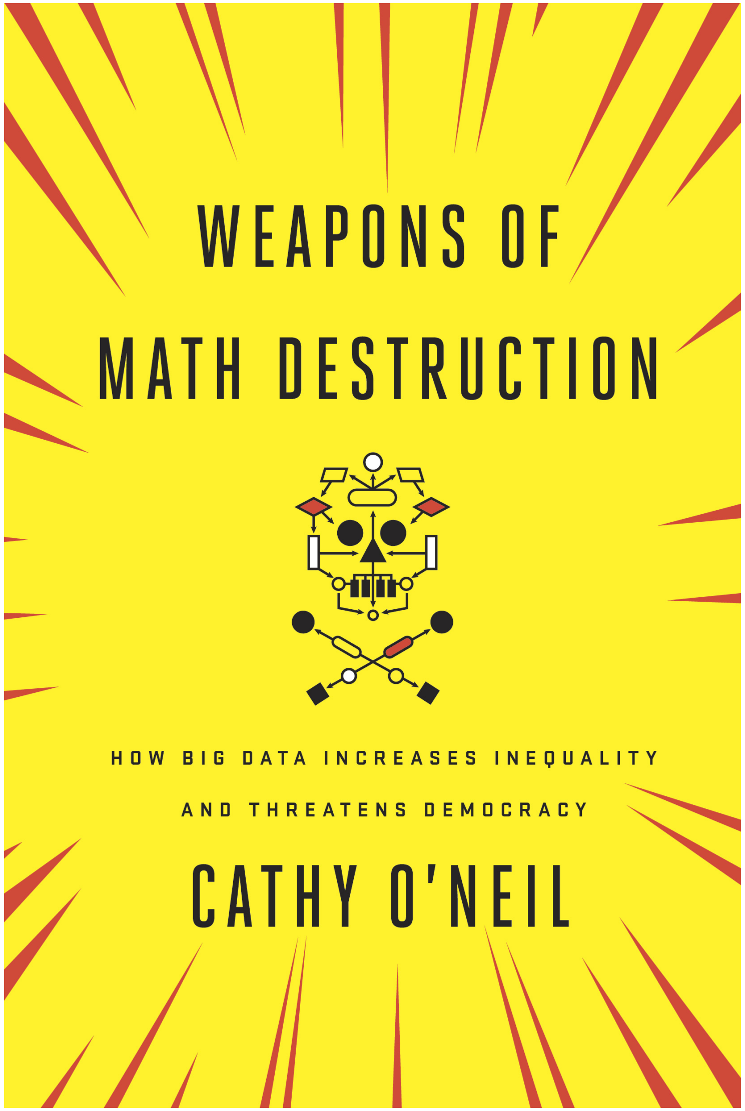
Vote

How confident would you be in an “A.I.” system that translates between R & Python?

- Very
- 50 – 50
- Erm sort of…
- Are you crazy?



Recommendation



Cathy O'Neil's website:

- https://mathbabe.org/

Ted talk:

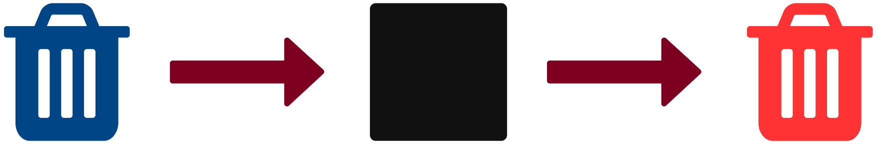
- https://youtu.be/_2u_eHHzRto

Google talk:

- https://youtu.be/TQHs8SA1qpk

Summary

- Data is often seen as a technical challenge
  - Cleaning & preparing it to summarise, visualise & analyse
- Do you really know and understand your data?
  - Are the data reliable and usable?
- Data have limits
  - Is your data appropriate? valid? biased?
- Analyses cannot save bad or inappropriate data
  - Garbage in, Garbage out



DSPT
DATA SCIENCE PORTUGAL

Thank you

Saghir Bashir

{i} ilustat
www.ilustat.com

Notice: All product names, logos, and brands are property of their respective owners. All company, product and service names used in this website are for identification purposes only. Use of these names, logos, and brands does not imply endorsement.