

# Using the Forbes Billionaires List as a Research Dataset

## Methods, Challenges, and Insights on the Global Super-Rich

Lidia Ceriani

Alba di Canazei, January 7th 2026



UNIVERSITÀ  
di **VERONA**  
Dipartimento  
di **SCIENZE ECONOMICHE**

CES **ifo**



**LISER**

Luxembourg Institute of  
Socio-Economic Research

# Outline

- 1 Motivation
- 2 Data Availability
- 3 Main Issues with using Forbes Data
- 4 Some Hands-on Solutions
  - Unique Record Identifiers
  - Country Names
  - Anchoring Strategies
- 5 AI-Assisted Completion of Missing Attributes

# Forbes List of Billionaires: Overview

- An **annual ranking** of the world's billionaires by estimated net worth, published by the business magazine *Forbes*. It includes individuals and families whose total assets are judged to be  $\geq$  US\$1 billion
- The first *World's Billionaires* list was published in March 1987, and it has appeared each year since then, most recently in 2025 with over 3,000 billionaires
- The list is assembled by *Forbes* analysts who estimate net worth using:
  - ▶ publicly available financial data
  - ▶ stock holdings and valuations
  - ▶ company ownership and private assets
  - ▶ interviews and research into individual wealth structures
  - ▶ For private companies or assets without transparent market values, additional indices (e.g., market indices) and expert judgments are often used
- In addition to the annual print list, *Forbes* maintains a *Real-Time Billionaires* tracker that updates estimated net worth continuously based on stock market movements and other valuation inputs
- It is not a random or administrative dataset; editorial judgment and evolving methodology shape the list over time, and wealth estimates are inherently approximations

# Why Use the Forbes Billionaires List?

## ► Direct observation of the extreme tail

- ★ Forbes provides one of the very few global data sources that directly observe individuals at the very top of the wealth distribution, where surveys and most administrative data fail

## ► Global coverage and long time span

- ★ Annual world wide billionaire rankings since the late 1990s allow cross-country and long-run analyses of extreme wealth concentration

## ► Rich individual-level metadata

- ★ Information on net worth, industry, country, age, gender, and self-made status enables research on wealth accumulation mechanisms and heterogeneity at the top (Vermeulen, 2018)

## ► Unique insight into top-wealth dynamics

- ★ Entry, persistence, and exit from the billionaire threshold can be studied, shedding light on mobility and volatility among the ultra-wealthy

## ► Complementarity with survey and tax data

- ★ Forbes data can be combined with household surveys and tax records to improve the measurement of top wealth shares and inequality (Saez e Zucman, 2016; Bach, Thiemann e Zucco, 2019)

## ► High relevance for policy debates

- ★ The concentration of billionaire wealth is central to discussions on taxation, inheritance, market power, and political influence

# Data Availability

- Historical Forbes Data have been cleaned and published
  - ▶ Freund e Oliver (2016) (1996 to 2009)
- Currently, Forbes charges a fee to get a clean dataset. But it is still possible to retrieve information by web-scraping
  - ▶ Kaggle (1997 to 2024)
  - ▶ Real Time Billionaires (2020-2025)
  - ▶ Additional information from singularly sourced years from web scraping
- In this Laboratory, we will work on selected years, not the full time series

# Forbes data are not a ready-made panel

- Forbes data are invaluable for studying extreme wealth, but they are not a ready-made panel
- Treating them as such requires some deliberate actions
  - ▶ Identifier construction
  - ▶ Anchoring of stable attributes
  - ▶ Transparency about measurement changes

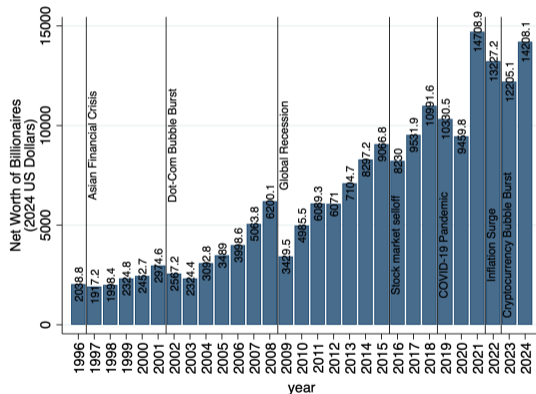
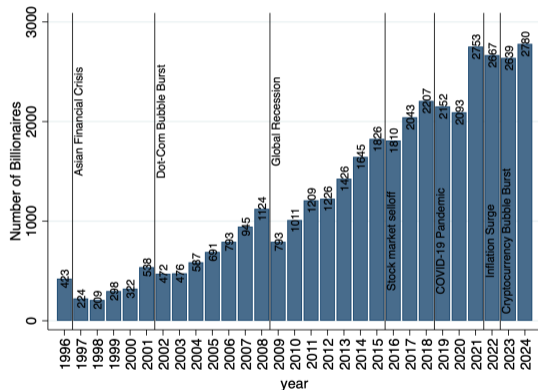
# Forbes data are not a ready-made panel

## Multiple Sources, No Unified Data-Generating Process

- Forbes combines:
  - ▶ public filings
  - ▶ private company estimates
  - ▶ journalistic investigations
  - ▶ self-reported information
- Methods and coverage evolve over time, so changes may reflect measurement rather than real dynamics
  - ▶ Example: In 2017, Wilbur Ross's reported net worth was highly disputed, with Forbes initially listing him at \$2.9 billion but later revising it to under \$700 million after financial disclosures revealed fewer assets, a change Ross disputed by claiming over \$2 billion was held in undisclosed family trusts

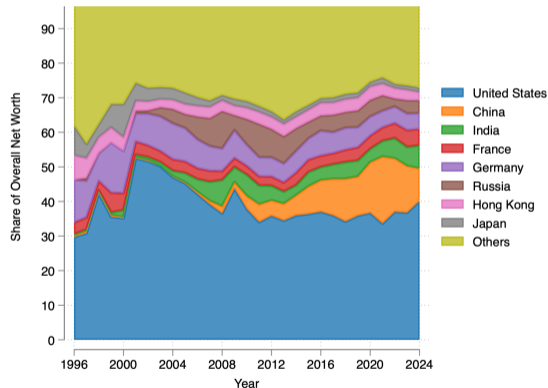
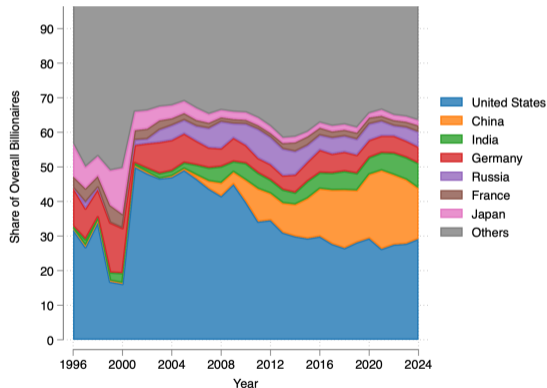
# Forbes data are not a ready-made panel

- Not a big issue is you just want to check and display the number of billionaires, and their net worth across the years



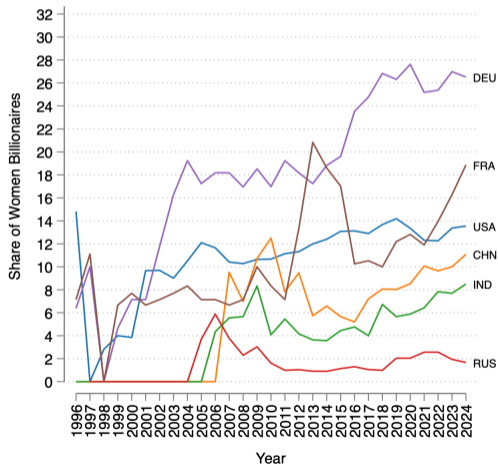
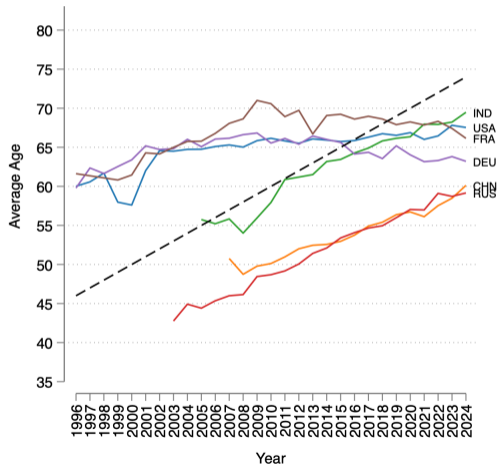
# Forbes data are not a ready-made panel

- But it becomes problematic when you want to study also other billionaires' characteristics across the years
  - ▶ There graphs need a harmonized label for the country of citizenship of billionaires



# Forbes data are not a ready-made panel

- But it becomes problematic when you want to study also other billionaires' characteristics across the years
  - ▶ There graphs need a complete series of ages and gender for each years under analysis



# Forbes data are not a ready-made panel

## Entity Identification Is Non-Trivial

- Names change
  - ▶ Spelling
  - ▶ Transliteration
  - ▶ Accents
  - ▶ Use of titles (e.g. Prince), Suffixes (e.g. Sr, Jr, I, II), Denotations (e.g. Brothers)
- Individuals may appear as
  - ▶ Persons
  - ▶ Families
  - ▶ Couples
  - ▶ Holding entities
- The same individual may:
  - ▶ enter/exit the list due to valuation thresholds
  - ▶ Appear under different identifiers across years

**Panel IDs must be constructed, not assumed**

# Forbes data are not a ready-made panel

## Time-Invariant Characteristics Are Incompletely Observed

- Time-invariant individual characteristics (gender, birth year, self-made status)
  - ▶ missing in many years
  - ▶ inconsistently reported
  - ▶ sometimes updated retroactively
- Time-variant (country of citizenship, residency, source of wealth)
  - ▶ missing in many years
  - ▶ inconsistently reported

**Requires anchoring strategies, not year-by-year imputation**

# Forbes data are not a ready-made panel

## Family and Group Entries Complicate Longitudinal Analysis

- Assigning individual characteristics is complicated
  - ▶ Age
  - ▶ Gender
- Groups may
  - ▶ Split into individuals
  - ▶ Merge
  - ▶ Disappear
  - ▶ Composition can change over time

**Researchers must define who the unit of analysis is and document choices (e.g. “older individual retained”)**

# Forbes data are not a ready-made panel

## No Natural Panel Structure

- Forbes data are
  - ▶ Annual snapshots
  - ▶ Not designed for longitudinal inference
  - ▶ Mechanical year-to-year comparisons risk
    - ★ Spurious dynamics
    - ★ False persistence or volatility

**Panel use requires explicit modeling and careful documentation**

# Unique Record Identifiers I

Transform a raw name into a stable identifier (`uri`)

"François Pinault Jr. and Family"

- ▶ **Standardize case**

- ★ `gen uri = strlower(full_name) → "françois pinault jr. and family"`

- ▶ **Homogenize conjunctions**

- ★ `substr(uri, " and ", " & ", ".) → "françois pinault jr. & family"`

- ▶ **Remove punctuation**

- ★ `ustrregexra(uri, "[[:punct:]]", "") → "françois pinault jr family"`

- ▶ **Decompose accents**

- ★ `ustrnormalize(uri, "nfd") → "francois pinault jr family"`

- ▶ **Remove accent marks**

- ★ `ustrregexra(uri, "\\p{Mark}", "") → "francois pinault jr family"`

- ▶ **Remove non-identifying token**

- ★ `substr(uri, "family", "", ".) → "francois pinault jr "`

# Unique Record Identifiers II

Transform a raw name into a stable identifier (`uri`)

- ▶ **Remove suffixes**

- ★ `regexpr(uri, " jr$| sr$", "") → "francois pinault "`

- ▶ **Collapse internal spaces**

- ★ `stritrim(uri) → "francois pinault "`

- ▶ **Trim leading/trailing spaces**

- ★ `strtrim(uri) → "francois pinault"`

- ▶ **Create URI-style key**

- ★ `→ "francois-pinault" substr(uri, " ", "- ", .)`

- After visual inspection (`duplicates report uri year`), we find some duplicates

- ▶ B. Wayne Hughes; B. Wayne Hughes, Jr.; H. Ross Perot, Jr.; H. Ross Perot, Sr.; Henry Ross Perot, Jr. Henry Ross Perot, Sr.; Jim Davis; Jim Davis family; Li Li; Oleg Deripaska; Robert Miller; Wang Wei; Zhou Yifeng & family

- ★ `replace uri = "b-wayne-huges-jr" if strpos(full_name, "B. Wayne Hughes")>0 & strpos(full_name, "Jr")>0`

- ★ `replace uri = "li-li-shenzhen" if strpos(uri, "li-li")>0 strpos(city_of_residence, "Shenzhen")>0`

# Country Names

- The same concept (country of citizenship/residence) appears in different datasets with different codings
  - ▶ **Country names** (strings): "Italy", "Korea, Republic of", "Hong Kong SAR"
  - ▶ **ISO2 codes**: IT, KR, HK
  - ▶ **ISO3 codes**: ITA, KOR, HKG
- Choose one canonical representation (typically **ISO3**) and convert everything into it
  - ▶ `ssc install isocodes`
    - ★ Convert everything to the same case  
`replace citizenship = strupper(citizenship)`
    - ★ Generate ISO 3 codes from ISO2 codes  
`isocodes citizenship, gen(iso3c)`
    - ★ Generate countrynames from ISO2 codes  
`isocodes citizenship, gen(cntryname)`
- Political entities and special regions (e.g., Guernsey, Hong Kong, Taiwan, Kosovo) may require explicit manual treatment; always document recodes and check stability across years when constructing a panel

# Anchoring Strategies

## Why Anchor Panel Information? (Example: Age)

- In Forbes-based panels, time-invariant or slow-moving attributes (e.g., age, birth year, gender) are often **missing** or **inconsistent** across years due to different sources, updates, and editorial changes
  - ▶ **Naive year-by-year use** can generate implausible dynamics (age not increasing by 1, jumps, or reversals)
  - ▶ **Anchoring**: Choose a **reliable reference observation** within each individual (e.g., the last non-missing age), then reconstruct the full time path in a consistent way
    - ★ For instance, choose the most recent information as the most reliable
    - ★ Back-calculate and forward-calculate consistent age for every year
    - ★ Always run some diagnostics, and keep original information

# AI-Assisted Completion of Missing Attributes

## Methods

- Key individual attributes (gender, year of birth, industry, self-made status) are missing or inconsistently reported across years in the Forbes data
- Some useful rules
  - ▶ **Observed data always dominate**
    - ★ AI is used *only* when no authoritative information is available in the original dataset or related structured sources
  - ▶ **Targeted retrieval, not prediction**
    - ★ AI is prompted to retrieve information from publicly available, authoritative sources (Forbes profiles, official biographies, reputable media), not to infer attributes algorithmically
  - ▶ **Anchoring and propagation**
    - ★ When an attribute is identified for an individual, it is anchored to a single, well-defined observation (e.g. last reliable year) and propagated across all panel years
  - ▶ **Structured outputs**
    - ★ AI responses are constrained to machine-readable formats (CSV) with explicit fields for values, sources, confidence, notes
- This enables reproducibility and automated validation

# AI-Assisted Completion of Missing Attributes

## Ethics and Transparency in AI-Assisted Data Construction

- AI augments human data collection; it does not replace transparent measurement or careful documentation
  - ▶ **No guessing, no silent inference**
    - ★ AI is explicitly instructed to return “unknown” when reliable sources do not exist; name-based or probabilistic guessing is avoided
  - ▶ **Provenance and auditability**
    - ★ Every AI-filled observation is flagged with a source indicator (e.g. `forbes / ai / missing`) and accompanied by cited references
  - ▶ **Human oversight and validation**
    - ★ AI outputs are spot-checked against original sources; ambiguous or conflicting cases are documented rather than forced into a single value
  - ▶ **Bias awareness**
    - ★ The approach recognizes cultural and linguistic bias in name-based inference, especially for gender and industry classification, and avoids automated classification when reliability is low
  - ▶ **Research integrity**
    - ★ AI-assisted enrichment is treated as a data construction step, fully disclosed in the methods and appendices, not as ground truth

# Bad Prompt vs. Good Prompt: Why Prompt Design Matters

## Bad Prompt

- ▶ “Guess the gender and age of these billionaires from their names.”
- ▶ “Fill in missing information as best as you can.”
- ▶ No restriction on sources
- ▶ No treatment of uncertainty
- ▶ No structured output

## Typical outcome

- ▶ Name-based inference (culturally biased)
- ▶ Silent hallucination
- ▶ Inconsistent or unverifiable answers
- ▶ No way to audit or reproduce results





## Good Prompt

- ▶ “Retrieve publicly available information from authoritative sources.”
- ▶ “Do *not* guess; return *unknown* if unavailable.”
- ▶ Explicit source requirements (Forbes, biographies, news)
- ▶ Mandatory uncertainty and ambiguity flags
- ▶ Strict CSV output with provenance

## Typical outcome

- ▶ Transparent data enrichment
- ▶ Reproducible, machine-readable results
- ▶ Explicit handling of missingness
- ▶ Ethical and defensible use of AI

# References

-  Bach, Stefan, Andreas Thiemann e Aline Zucco (2019). “Looking for the missing rich: tracing the top tail of the wealth distribution”. In: *International Tax and Public Finance* 26.6, pp. 1234–1258.
-  Freund, Caroline e Sarah Oliver (2016). *The Origins of the Superrich: The Billionaire Characteristics Database*. Working Paper No. 16-1.  
<http://dx.doi.org/10.2139/ssrn.2731353>: Peterson Institute for International Economics.
-  Saez, Emmanuel e Gabriel Zucman (2016). “Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data”. In: *The Quarterly Journal of Economics* 131.2, pp. 519–578.
-  Vermeulen, Philip (2018). “How Fat Is the Top Tail of the Wealth Distribution?” In: *Review of Income and Wealth* 64.2, pp. 357–387.