Practical 1: Firm size, profit persistence, and where firms are from

Practical session, Day 1 — data, code and regressions

Dr. Ernil Sabaj

Warwick Summer School
International Business and Finance

20 July 2026

This hour

Three empirical pillars under this morning's frameworks — estimated live.

- **A. Size.** The firm size distribution is a power law: one regression, one parameter, big consequences.
- **B. Persistence.** The AR(1) test of sustained competitive advantage — with its biases on the table.
- **C. Nationality.** 6% of US-listed firms are legally somewhere they are not.

Every claim this morning was a hypothesis. This hour we estimate three of them.

No statistics background is assumed — every concept and every line of code is explained when it first appears.

The plan: from this morning's slides to tests

Claim / diagram from the lecture	**Data**	**What we estimate**	**The claim is in trouble if...**
"An MNE has a distinct **home country**" (MNE definition slides)	Compustat `fic` vs `loc`, 38,954 firms	Share incorporated away from HQ; by country	the legal and real home usually coincide. (They don't: 6%, Cayman 3rd.)
Firms are heterogeneous; **the biggest go abroad** (HMY sorting slide)	Sales of all listed firms, 2024	the size-inequality number ζ (Part A); typical size by origin country	sizes are roughly equal (ζ far from 1), or cross-listers are ordinary-sized
VRIO $\Rightarrow$ **sustained** advantage (VRIO decision tree)	ROA panel, 1990–2024	the persistence number ρ (Part B); a 5-year "where did they end up" table incl. exit	$\rho \approx 0$ (no advantage) or $\rho \approx 1$ (no competition)
"MNEs differ from domestic firms" (misconceptions slides)	1,780 foreign-HQ vs 5,239 US-HQ listed firms	trait gaps; a valuation comparison, foreign vs US firms (Part D)	controls absorb the gaps entirely

Everything is pre-computed on the slides; every number re-runs from `replication/` in class.

The data this hour: one file

`ib_panel.dta` — built from Compustat, the standard database of company accounts.

- One row = **one firm in one year** (a "firm-year"). **441,198 rows**; 39,383 firms; 1980–2025.
- It contains **every firm ever listed on a US stock exchange — including the dead ones**. Firms that went bankrupt or were bought stay in the data up to their last year. (Why that matters becomes the heart of Part B.)
- Firms appear for different lengths of time (they list, they leave); the typical firm is observed for 8 years.
- Money amounts (sales, costs) are the *total over the firm's four quarters of that year*; balance-sheet amounts (assets) are the *year-end value*. The flag `full_year==1` marks years where all four quarters are present — we always filter on it.

Software. Slides show **Stata**; the identical analysis in **Python** ships in `replication/student_pack_day1/`. (A second dataset — World Bank firm surveys — joins us tomorrow, not today.)

The variables, in plain words

Name	What it means	Guard (filter)
`sale`	the firm's revenue that year, \$m	—
`cogs`	**cost of goods sold**: what it cost to make what was sold	—
`assets`	everything the firm owns, \$m (year-end)	—
ROA	**return on assets**: operating profit ÷ assets — profit per dollar of stuff	keep $[-1, 1]$; assets $>$ \$10m
Markup	`sale` ÷ `cogs`: revenue per dollar of production cost	keep $[0.5, 10]$
Tobin's q	what the stock market says the firm is worth ÷ its accounting value	keep $[0.3, 15]$
R&D intensity	research spending ÷ sales (missing $\to$ 0)	—

- A **guard** is simply a filter we apply before using a ratio. Why: a ratio *explodes* when the bottom of the fraction is near zero — one software firm with production costs of almost nothing can show a "markup" of 500 and single-handedly wreck an average.
- Guards are *choices we made*, so we state them on every slide — and you should always ask any researcher what theirs were.

Three actual rows of ib_panel

Before any regression: this is literally what the data looks like.

firm	fyear	sale ($m)	cogs ($m)	roa	markup	dlrsn*
APPLE INC	2024	391,035	198,907	+0.34	1.97	—
DOXIMITY INC	2024	570	49	+0.18	11.53	—
ENRON CORP	2000	100,638	97,820	+0.03	1.03	02

*`dlrsn` = Compustat's deletion reason — the code recording *why* a firm left the panel (acquired, bankrupt, delisted...). It becomes the star of Part D.

- One row = one firm-year. `sale` = revenue; `cogs` = cost of goods sold (what it cost to make what was sold); both in $m.
- Doximity's markup of 11.5 is why the guards exist: software COGS ≈ 0.
- Enron's final row is fiscal 2001 with `dlrsn` = 02 (bankruptcy): **exit is data, not absence**.

Four ways this data will mislead you if you let it

1. **Survivorship**: if you only study firms that are still around, you are studying the winners. (Ask "what do successful firms do?" of survivors only, and everything they do looks like success.) Our fix: firms that disappeared stay in our analysis, with their disappearance recorded.

2. **Look-ahead**: using information that was not yet known at the time. Example: sorting firms by their 2024 profits "at the start of 2024" — nobody knew those profits yet. Fine for describing history; cheating if you claim you could have traded on it.

3. **Broken ratios**: mergers, giant write-offs and changed year-ends make a ratio jump for *accounting* reasons, not real ones. The `full_year` flag and the guards catch most of this.

4. **The cast changes**: the 1985 data is full of factories; the 2024 data is software and biotech. A "40-year trend" can just mean *different firms are in the data now* — always ask which it is.

None of these is exotic. Each one has ruined published research.

Before any regression: know your panel

Firm-years (`full_year==1`, sales > 0)	404,281
Firms	22,864
Years	1980–2025
Median years per firm	8
Firms present in all 45 years	268

- Firms **enter** (they list on the exchange) and **leave** (bought, failed, went private) — and both events are related to how well the firm was doing. Who is *in* the data is itself informative.
- Only 268 firms are present in all 45 years — so anything you compute "over time" mixes changing firms with changing behaviour (trap 4 from the last slide).
- The guards from the variables slide are applied throughout: ROA kept in $[-1, 1]$, markups in $[0.5, 10]$, assets $>$ \$10m — and we say so on every result.

The only statistics you need this hour

- A **regression** draws the *best-fitting straight line* through a cloud of points. That is all it is.
- The **coefficient** (or *slope*) says: when the thing on the right goes up by 1, the thing on the left moves by *this much* on average.
- The **standard error** (se) is the "$\pm$" on the slope — how much it would wobble if we had drawn a different sample. Small se = precisely measured.
- The t-**statistic** = slope $\div$ se. Rule of thumb: if $|t| > 2$, the slope is too big to be a fluke of sampling — we call it *statistically significant*.
- R^2 (0 to 1): how much of the variation in the left-hand thing the line accounts for. It measures *fit*, never *causation*.
- **Logs**: when both sides are in logarithms, the slope reads as *"a 1% change in x goes with a β% change in y"* — percentages, not units. We use this constantly.

Six ideas. Everything this week is built from them.

Part A. How unequal are firm sizes? A one-number answer

Heights follow a **bell curve**: almost everyone is near average, and a 3-metre person is impossible. Firm sizes do not — they are **fat-tailed**: *giants are common*. Our panel runs from $10m firms to a $600bn one, and every size in between is populated.

The pattern has a name — **Zipf's law**: *at any size, doubling the size roughly halves the number of firms that big*. The 2nd-largest firm is about half the largest; the 10th about a tenth. One number, ζ ("zeta"), measures it:

$$\Pr(\text{Size} > s) \;\approx\; \left(\frac{s}{s_{\min}}\right)^{-\zeta}, \qquad \zeta = 1 \;\Leftrightarrow\; \text{exactly the halving rule.}$$

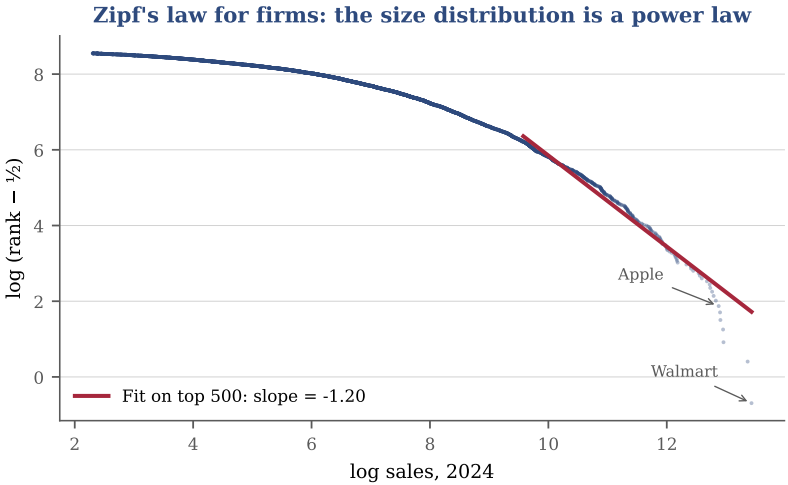
- **How we estimate** ζ: rank firms by size (1 = biggest); plot log(rank) against log(size); the slope of the best-fit line is $-\zeta$. (We use rank $-\frac{1}{2}$, a small technical correction — Gabaix & Ibragimov 2011.)
- **Why it matters**: measured on *all* US firms — millions of them, via Census data — Axtell (2001, *Science*) found $\zeta \approx 1$. With a tail that fat, a handful of giant firms *are* the aggregate economy: a bad year for three of them shows up in GDP (Gabaix 2011).

Estimate it yourself (Stata)

```
use ib_panel, clear              // open the data file
keep if fyear == 2024 & full_year == 1 & sale > 10
                                 // keep: year 2024, clean years,
                                 //   real firms (sales over $10m)
gsort -sale                      // sort rows, biggest sales first
gen rank  = _n                   // rank = row number: 1 = biggest
gen lrank = ln(rank - 0.5)       // take logs (-0.5 = the small
gen lsale = ln(sale)             //   technical correction)
reg lrank lsale if rank <= 500   // best-fit line, top 500 firms
```

`reg` y x draws the best-fit line of y on x. Its slope here is $-\zeta$, so flip the sign. One special rule for *ranked* data: the usual standard-error formula does not apply — the honest uncertainty is $\hat{\zeta}\sqrt{2/N}$, reported on the results slide as the "GI se".

The picture: a straight line across three orders of magnitude



How to read it: each dot is **one firm**; the axes are in **logs** (each step $= \times 10$). Down-and-to-the-right: bigger firms have lower ranks. A **straight line** is exactly what Zipf's law predicts — and it holds from \$100m firms to \$600bn ones. The slope of the fitted line is $-\hat{\zeta} \approx -1.2$.

Source: own calculations, Compustat, fiscal 2024, firms with sales $>$ \$10m. Fit on the largest 500 firms.

Reading the estimate — and the trap in the lower tail

Sample	$\hat{\zeta}$	GI se	R^2
Top 500 firms	**1.20**	0.08	0.96
All 5,165 listed firms >$10m	0.41	—	0.83

- ("GI se" = the corrected standard error from the last slide — the $\pm$ on our estimate.) $\hat{\zeta} = 1.20 \pm 0.08$: close to the halving rule, and close to the 1.06 that Axtell (2001) found using Census data on *all* US firms, small ones included.
- So why does the **full sample** give 0.41? Not because the law fails — because our data is **truncated**: it contains only firms big enough to be listed on a stock exchange. The millions of corner shops that would fill in the bottom of the picture simply are not in the file, and their absence flattens the fitted line.
- Rule: estimate a *tail* pattern *on the tail* — hence the top-500 fit.
- Lesson 1 of the week: before interpreting any number, ask what the data's coverage did to it.

Concentration of activity: the top 10 firms are **12%** of all US-listed sales; the top 100 are **45%**. "Multinationals dominate world trade" is not a conspiracy — it is what $\zeta \approx 1$ looks like.

The test: sort firms by profitability, then wait ten years

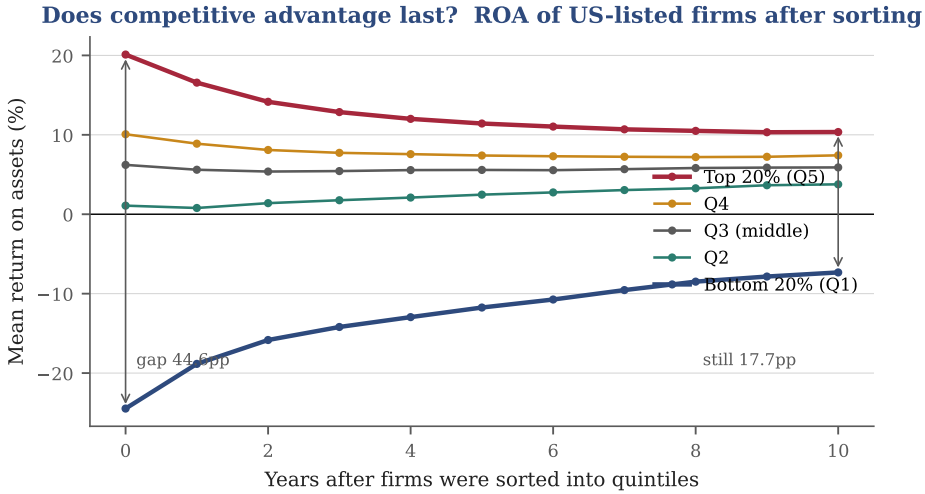
Design (this is Mueller's 1986 persistence-of-profits design):

1. Every year, rank all US-listed non-financial firms by return on assets.
2. Split them into five equal groups (quintiles). Q5 = most profitable 20%.
3. Follow each group for the next ten years and take the average ROA.

What should we see?

If markets are competitive	the lines converge fast. Rivals imitate; abnormal profit is competed away.
If VRIO is right	the gap persists — resources that are valuable, rare, inimitable and organised keep the top firms on top.

The answer: both, a bit



Source: own calculations, Compustat quarterly. US-listed non-financial firms, sorting years 1990–2019, 4-quarter fiscal years, assets > **$**10m. ROA = operating income after depreciation / total assets.

Reading the picture

- The gap between best and worst **halves within four years** — 44.6 percentage points falls to 25.0. Competition is real and it is fast.
- But it **never closes**. Ten years on, the top quintile still earns 10.4% and the bottom still loses 7.3% — a gap of **17.7 points**.
- About **40% of the initial advantage survives a decade**. Neither "competition eliminates profit" nor "good firms stay good" wins outright.

This is the honest version of VRIO:
advantage is real, decays quickly, and leaves a durable residue.

Part B. Persistence, done properly: the AR(1)

The fan chart shows convergence visually. The workhorse estimate of *how fast* is one number:

$$\text{ROA}_{it} \;=\; \alpha \;+\; \rho\,\text{ROA}_{i,t-1} \;+\; \gamma_t \;+\; \varepsilon_{it}$$

- In words: *next year's profitability* = *a constant* + $\rho\times$ *this year's* + *noise*. So ρ ("rho") is **how much of your lead you keep each year**. $\rho = 0.8$: a firm 10 points above average today is, on average, 8 points above next year. $\rho = 0$: advantage gone instantly (perfect competition). ρ near 1: advantage never dies (VRIO's dream).
- The **half-life** — years until half the lead is gone — is $\ln(0.5)/\ln(\rho)$. It turns ρ into something a manager can feel.
- Two choices we make, in plain words: (i) **year effects** γ_t: allow every year its own average, so booms and recessions do not masquerade as firm persistence; (ii) **cluster the standard errors by firm**: tell the software that *the same firm's years move together* — they are not fresh independent observations — so it does not overstate how certain we are. (Pretending 30 years of Apple are 30 independent firms would fake precision.)

This is Mueller's (1986) *Persistence of Profits* design, run on 45 years of Compustat.

The regression, in Stata

```
use ib_panel, clear                          // open the data
keep if full_year==1 & assets>10 & inrange(roa,-1,1)   // the guards
drop if inlist(real(gsector),40,60)         // drop financials & RE:
                                             //   their ROA isn't comparable
egen id = group(gvkey)                       // give each firm a number
xtset id fyear                               // tell Stata: panel data --
                                             //   this firm, this year
* (1) pooled       (2) + year effects       (3) firm fixed effects
reg  roa L.roa,                cluster(id)
reg  roa L.roa i.fyear,        cluster(id)
xtreg roa L.roa i.fyear, fe    cluster(id)
```

`L.roa` means "*last* year's roa" (that is why `xtset` is needed); `i.fyear` adds the year effects; `cluster(id)` is the same-firm-years-move-together correction; `fe` in column (3) compares each firm *with its own average* — next slide shows what that means.

What fixed effects actually do (in one picture's worth of algebra)

The firm-FE estimator demeans everything within each firm:

$$(\text{ROA}_{it} - \overline{\text{ROA}}_i) \;=\; \rho\,(\text{ROA}_{i,t-1} - \overline{\text{ROA}}_i) + \ldots$$

- **What dies**: everything constant about a firm — its industry, its founding era, "being Apple". Column (3) cannot be driven by some firms being permanently great, because permanent greatness has been subtracted.
- **What survives**: only *within-firm movement over time*. The FE question is different, not better: "when *this* firm is above its own average, how long does the excursion last?"
- **The price**: this method has a known technical quirk — with short firm histories it *understates* persistence somewhat (the next slide uses this).
- Habit to keep: when two methods disagree, **say which question each answers** before asking which is right.

Results: how fast does advantage die?

Dep. var.: ROA_{it}	(1) Pooled	(2) + year FE	(3) Firm FE
$ROA_{i,t-1}$ (ρ)	**0.807**	**0.802**	**0.443**
	(0.003)	(0.004)	(0.006)
Implied half-life	3.2 yrs	3.1 yrs	0.9 yrs
N (firm-years)	165,792	165,792	165,792
Firms	17,092	17,092	17,092

SEs clustered by firm. Sample: US-listed non-financials, 1990–2024, consecutive-year pairs.

- $\rho \approx 0.81$: a firm 10pp above the mean today is, on average, **8.1pp above next year** — and half the gap is gone in about three years.
- Advantage is real and it decays. Both halves of that sentence carry equal weight.

How to read a regression table (once, properly)

Take column (1) of the last slide: $\hat{\rho} = 0.807$, (0.003), $N = 165{,}792$.

- **The coefficient** (0.807): the slope. Meaning: a firm 1 point more profitable today is expected to be 0.81 points more profitable next year. First question to ask of any coefficient: *is that big?* Here yes — it means half an advantage survives three years.
- **The number in parentheses** (0.003): the standard error — the $\pm$ on the slope. $t = 0.807/0.003 \approx 269$, far above 2: not a fluke. (With 165,000 observations almost everything clears the bar — which is why significance is the *least* interesting number here.)
- **"SEs clustered by firm"** (the table footnote): the same-firm correction from earlier. Without it the table would claim far more precision than it has.
- N **and Firms**: how many rows and how many distinct firms went in. Always check the sample matches the claim.
- R^2: the share of the variation the line explains — a measure of *fit*, never of *causation*.

Reading order for life: sign $\rightarrow$ magnitude $\rightarrow$ SE/design $\rightarrow$ only then the stars.

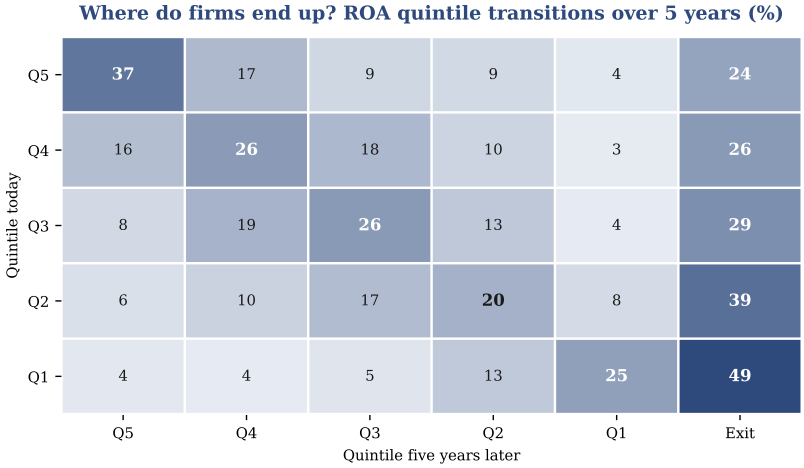
Why columns (1) and (3) disagree — and why we report both

Column (1) is the left column of the results table (plain regression); column (3) the right one (firm fixed effects). They give 0.79 and 0.46 — both are honest, and both are a little wrong, in opposite directions:

- **Column (1) is too high.** Some firms are simply *always* good (a great brand, forever). The plain regression cannot tell "being permanently great" apart from "keeping last year's lead", so both end up inside its 0.79.
- **Column (3) is too low.** Comparing each firm with its own average has a known technical quirk: when firm histories are short (our typical firm: 8 years), it *understates* persistence. (For reference, the quirk is called the Nickell bias — the name is all you need.)
- So the honest statement is a **bracket**: the truth lies *between* 0.46 and 0.79 — an advantage's half-life is somewhere between one and three years.

When two defensible methods disagree, report the bracket and explain why.
That sentence is worth more than either number.

The same fact as a transition matrix — now with exit



Source: own calculations, Compustat. Row = ROA quintile in year t (sorting years 1990–2018); column = position in $t + 5$. Rows sum to 100.

Three numbers to take from that matrix

1. **37%** of top-quintile firms are still top-quintile five years on — almost double the 20% a random shuffle would give. Persistence is real.
2. **49%** of bottom-quintile firms are simply *gone* in five years, versus 24% of top-quintile firms. Exit is the main event in the bottom of the distribution, and any analysis that ignores it is describing survivors.
3. Direct jumps Q1→Q5 (4.0%) and Q5→Q1 (4.3%) are rare. Fortunes change by **drift or death**, not by teleport.

This is the quantitative content of "sustained competitive advantage": real, partial, and always competing against the exit margin.

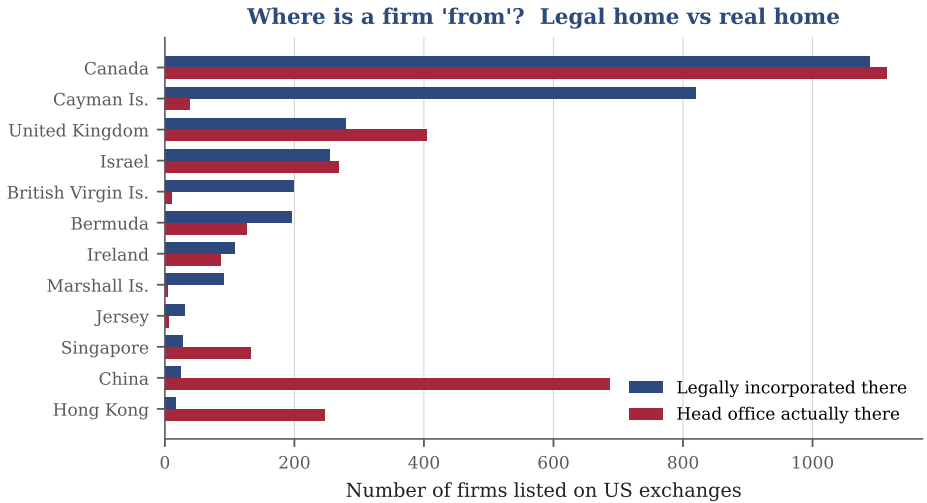
Now the trap — and this is the part that matters

The chart only follows firms that **still exist**. Here is who is left:

Starting quintile	After 3 yrs	After 5 yrs	After 10 yrs
Q5 (most profitable)	84.7%	76.1%	**51.5%**
Q4	83.1%	73.9%	49.2%
Q3	81.6%	71.0%	46.5%
Q2	75.0%	62.4%	36.5%
Q1 (least profitable)	63.1%	50.1%	**27.1%**

- Only **27% of the worst firms** are still listed after ten years. The rest were acquired, went private, or failed.
- So the Q1 line is drawn using the **survivors** — the unlucky ones that recovered enough to stay listed.
- **The true penalty for being a bad firm is worse than the chart shows.** Survivorship bias always flatters the losers.

The legal home and the real home are not the same place



Source: own calculations, Compustat quarterly. Every firm ever listed on a US exchange (n = 38,954), latest observation. `fic` = country of incorporation, `loc` = country of headquarters.

Three facts that should bother us

- The **Cayman Islands** is the **third-largest country of incorporation** for US-listed firms — 819 firms.
  Thirty-nine of them actually have an office there.

- **6.2%** of US-listed firms are incorporated somewhere other than where they are run.

- **687** firms run from China are listed in New York. **24** of them are incorporated in China.

So which country is the "home country"? Which government regulates them?

Passports of convenience: the biggest examples, 2024

Firm	Incorporated	Sales ($m)	Run from
Bunge Global SA	Switzerland	53,108	St. Louis / Chesterfield, USA
SLB (Schlumberger)	Curaçao	36,289	Houston, USA
Flex Ltd	Singapore	25,813	Austin, USA
Carnival Corp	Bermuda	25,021	Miami, USA
Royal Caribbean	**Liberia**	16,484	Miami, USA
Flutter Entertainment	Ireland	14,048	New York, USA
Norwegian Cruise Line	Bermuda	9,480	Miami, USA
GlobalFoundries	Cayman Islands	6,750	Malta, New York, USA

- Marshall Islands: **91** firms incorporated, **4** headquartered — almost all shipping.

Source: own calculations, Compustat quarterly, fiscal 2024.

Why? Four reasons, and only one is tax

1. **Tax.** Historically the big one. US inversions peaked before the 2017 Tax Cuts and Jobs Act cut the headline rate from 35% to 21% and moved the US toward territorial taxation. The 2021 OECD/G20 global minimum tax (Pillar Two, 15%) narrows the gap further.

2. **Liability and legal regime.** Cruise lines register ships in Liberia, Panama and the Bahamas — labour law, safety inspection and liability rules follow the flag, not the head office.

3. **Getting listed at all.** A Cayman shell lets a Chinese operating company sell shares in New York when Chinese law forbids foreign ownership of the operating business directly.

4. **Merger history.** Flutter is Paddy Power (Ireland) + FanDuel (US) + Stars Group (Canada). Nobody chose Ireland — it is a residue.

Nationality is a *choice variable*, not a fact about a firm.

What this does to the rest of the course

When we say "**FDI from country A to country B**", "**home country**", or "**emerging-market MNE**", we are relying on a legal address that the firm chose.

- Damgaard, Elkjaer & Johannesen estimate that **phantom FDI** — investment into corporate shells with no substance and no real link to the local economy — is about **40% of global FDI**, roughly $15 trillion in 2017.
- Luxembourg (population $\approx$ 600,000) hosts about as much inward FDI as the United States: $4 trillion, or $6.6 million *per inhabitant*.
- Ten economies — Luxembourg, the Netherlands, Hong Kong, BVI, Bermuda, Singapore, Cayman, Switzerland, Ireland, Mauritius — host over 85% of it.

Hold onto this for the FDI statistics in the next section — and for the "rise of the emerging-market MNE" debate.

Damgaard, Elkjaer & Johannesen (2024), "What is real and what is not in the global FDI network?", *Journal of International Money and Finance* 140; earlier as IMF WP/19/274.

Part D. The comparison the course is named after

Multinationals versus domestic firms — how many, from where, with what traits, worth how much.

First, an honest limitation — stated, not hidden

(Economists call what follows an "identification problem": the comparison we *can* make is not quite the comparison we *want*.)

What this file cannot see. Compustat North America has no geographic segments, no foreign pre-tax income (`pifo`), no foreign taxes (`txfo`). So for a **US-headquartered** firm, multinationality is *invisible here*: Apple and a Kansas utility look identical on the dimension we care about.

What it can see. 1,780 firms (2024) **headquartered outside the US but listed on US exchanges**. These are multinationals by construction — they crossed a border to raise capital, file with the SEC, and answer to foreign investors.

The selection is the theory. Cross-listing has a large fixed cost (US GAAP/IFRS reconciliation, SOX compliance, litigation exposure). Only firms expecting to clear it come — **exactly the HMY sorting logic from this morning**, applied to the capital-market border instead of the product-market one. So: our "MNE" group is well-defined, but

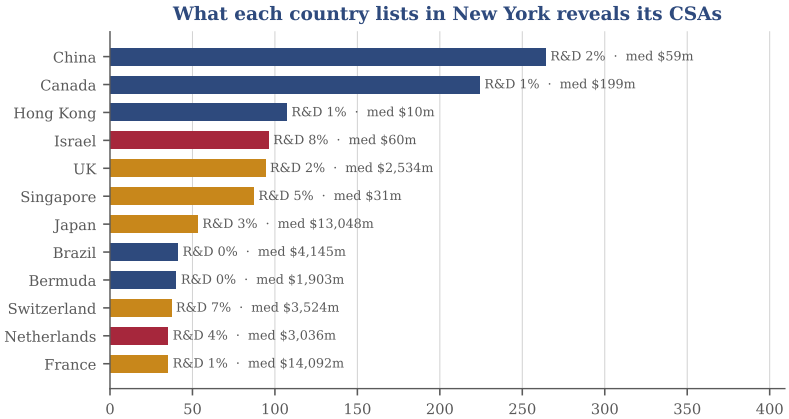
selected. Every comparison that follows
is a comparison of *firms that chose to come* with firms already home.

How many came, and when

From **4.3%** of US listings in 1980 to **15.8%** in 2024. One in six "US-listed firms" is run from abroad.

Source: own calculations, Compustat quarterly; firms with a fiscal-year record, loc $\neq$ USA.

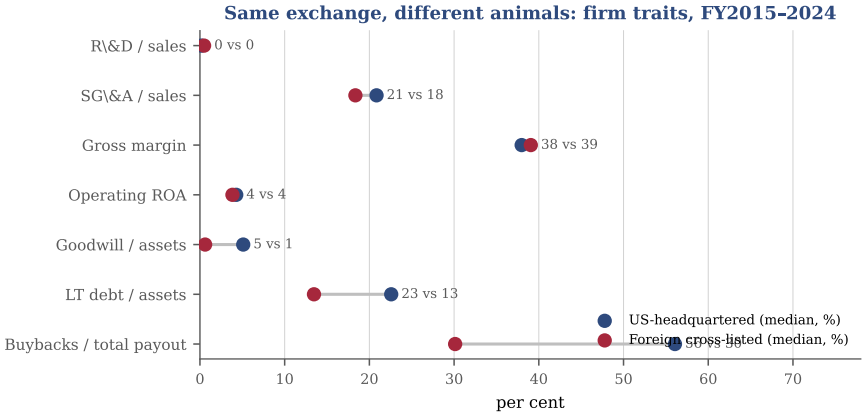
From where, and with what: origin profiles are CSAs made visible



- **Israel**: 96 firms, **70% Tech+Health**, R&D **8.1%** of sales — the "start-up nation" CSA in a bar.
- **Canada**: modal sector **Materials** (the resource CSA). **Japan/France**: median sales **$13–14bn** — only giants clear the fixed cost of cross-listing: HMY at the capital-market border.
- **China**: 264 firms, median sales **$59m** — the reverse-merger and VIE tail, not national champions.

Active firms, fiscal 2024, full four-quarter records.

Same exchange, different animals



- **Goodwill 0.6% vs 5.1% of assets**: US growth is **acquisition-led**; cross-listers arrive grown. (Goodwill is the accounting shadow of the M&A entry mode from this morning.)
- **Buybacks 30% vs 56% of payout**: dividend cultures travel with the passport. **Leverage 13% vs 23%**: foreign firms come conservatively financed.

Medians, FY2015–2024 pooled, non-financials, assets > $10m. $n = 1{,}730$ foreign vs 5,239 US firms.

What is a multinational worth? The cross-listing q regression

$$\ln q_{it} \;=\; \beta\,\text{Foreign}_i \;+\; \gamma' X_{it} \;+\; \delta_s + \delta_t + \varepsilon_{it}, \qquad q = \frac{\text{assets} + \text{mkt. cap} - \text{book equity}}{\text{assets}}$$

	(1) raw	(2) full controls	(3) excl. China/HK
Foreign	-0.188	$\mathbf{-0.145}$	-0.129
	(t $= -10.9$)	(t $= -9.0$)	(t $= -7.5$)
Implied q gap	-17%	$\mathbf{-13\%}$	-12%

$N = 38{,}735$ firm-years, FY2015–24; SEs clustered by firm. Controls in (2)–(3): ln sales, R&D intensity, leverage, sector and year FE.

- A robust **13% valuation discount** — Zaheer's **liability of foreignness**, priced. Not a China artifact.
- But mind the comparison group: Doidge–Karolyi–Stulz's famous cross-listing *premium* compares these firms to their **home-market peers**. Same firms, premium against home, discount against US hosts. **The sign of the "premium" is a choice of counterfactual.**

Issues facing the multinational, read from how they exit

% of all delistings, 1980–2025	Foreign cross-listed	US-headquartered
Acquired / merged	43.1	45.6
Stopped filing / moved to OTC	**39.7**	28.0
Went private	6.2	7.6
Bankruptcy	1.7	**3.8**
Liquidation	2.5	**9.3**

- **Far less likely to exit through financial distress**: bankruptcy 1.7% vs 3.8%, liquidation 2.5% vs 9.3% — together **4.2% vs 13.1%**. Weak firms rarely pay the cross-listing fixed cost, and the ones that do arrive established.
- **More likely to simply stop filing or move to OTC** (39.7% vs 28.0%): a delisting, not a death. Acquisition is only modestly rarer (43.1% vs 45.6%) — the market for corporate control reaches across the border more readily than the bankruptcy courts do.
- Five-year survival is nearly identical (72% vs 70%): once here, they stay at the same rate — they just *leave differently*.

Read the comparison like an economist

1. **Selection, everywhere.** Every gap mixes "being multinational" with "being the kind of firm that cross-lists". Nothing on the last five slides proves that *being* multinational *causes* the gaps.

2. **Composition.** The foreign group is 15% Chinese small-caps and 3% Japanese giants. Always re-run by origin before generalising — exercise 4 does exactly this.

3. **Accounting regimes travel too.** Part of the goodwill gap is IFRS vs US GAAP practice, not strategy. Part of the R&D gap is disclosure habit.

4. **"MNE" here means capital-market multinationality.** A German firm selling in 80 countries but listed only in Frankfurt is invisible; a Cayman shell is not. The clean fix is Compustat **Segments** — the one dataset this course is still missing.

New exercise 4: drop China and Hong Kong, recompute the traits dumbbell.
Which gaps survive? Those are the ones worth theorising about.

Your turn (15 min, laptops)

Open `replication/stata/`, run `02_persistence_and_nationality.do` to the AR(1) block, then:

1. Re-estimate ρ using **gross margin** instead of ROA. Is product-market advantage more or less persistent than asset productivity?
2. Estimate ρ by decade (1990s, 2000s, 2010s). Has competition sped up?
3. Re-run the Zipf regression for **2000** and **2024**. Did the tail get fatter?

```
gen gm = (sale - cogs)/sale
* ... then re-use the AR(1) block with gm in place of roa
```

Expected findings to discuss: $\rho_{\text{gm}} > \rho_{\text{roa}}$ (margins outlast returns); ρ falls modestly across decades; $\hat{\zeta}$ drifts down (fatter tail) — connect each to this morning's frameworks.

References I

- Axtell, R. (2001), 'Zipf Distribution of U.S. Firm Sizes', *Science* 293(5536), 1818–1820.
- Gabaix, X. (2011), 'The Granular Origins of Aggregate Fluctuations', *Econometrica* 79(3), 733–772.
- Gabaix, X. & Ibragimov, R. (2011), 'Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents', *JBES* 29(1), 24–39.
- Mueller, D. (1986), *Profits in the Long Run*, Cambridge University Press.
- Nickell, S. (1981), 'Biases in Dynamic Models with Fixed Effects', *Econometrica* 49(6), 1417–1426.
- Zaheer, S. (1995), 'Overcoming the Liability of Foreignness', *Academy of Management Journal* 38(2), 341–363.
- Doidge, C., Karolyi, G.A. & Stulz, R. (2004), 'Why are foreign firms listed in the U.S. worth more?', *Journal of Financial Economics* 71(2), 205–238.
- Coffee, J. (1999), 'The Future as History: Prospects for Global Convergence in Corporate Governance', *Northwestern University Law Review* 93(3), 641–707.
- Damgaard, J., Elkjaer, T. & Johannesen, N. (2024), 'What is real and what is not in the global FDI network?', *Journal of International Money and Finance* 140.
- Helpman, Melitz & Yeaple (2004), *AER* 94(1), 300–316.
- Cameron, A.C. & Miller, D. (2015), 'A Practitioner's Guide to Cluster-Robust Inference', *Journal of Human Resources* 50(2), 317–372.

Take this with you

- Every framework this morning was a **hypothesis**. Today you saw which parts survive estimation, which fail, and which the data cannot even address.
- A regression is an association until you have argued identification. We wrote that argument out loud each time — copy the habit.
- The dataset determines the question. Knowing what Compustat **cannot** answer (private firms, input prices, geographic segments) is half of empirical literacy.

Code and data: `replication/`. Do-files 01–04 reproduce every number shown today.