The Data Economy

Isaac Baley
isaac.baley@upf.edu

Bojos per l'Economia

Who am I?

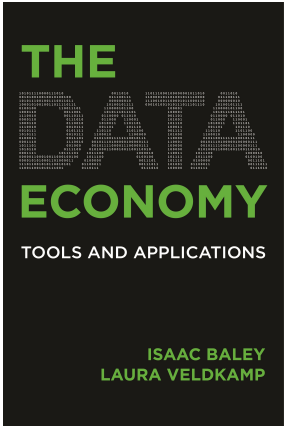
About me

- From **Mexico City**
- Studied at **ITAM**: Economics + Mathematics + Master's
- **Work experience**:
  - Professor (swimming, piano, math)
  - Research Assistant at ITAM
  - Ministry of Finance and Ministry of Economics
- **PhD in Economics**: New York University, 2009–2015
- In Barcelona for 10 years
  - **Associate Professor** at Universitat Pompeu Fabra
  - **Researcher** at CREi and Barcelona School of Economics
- Caps de setmana a Sadernes, **Alta Garrotxa**.

My research topics

Micro Frictions $\Longrightarrow$ Macro Dynamics

- **Search frictions:** finding the right job takes time and effort
- **Adjustment costs:** changing prices or resizing a factory is costly
- **Information frictions:** decisions are made with imperfect information
- **Macro:** these frictions shape business cycles and long-run growth



Princeton University Press, 2025

More information

What is Data?

What is Data?

Data is digitized information

1. It describes people, firms, or nature (numbers, text, images, GPS, clicks)
2. It is an input into **prediction** algorithms (AI), value comes from **reducing uncertainty**
3. It is a **by-product** of our digital footprint
4. It is often **non-rival**, many people can use the same data
5. It can be shared and **traded**

Our digital footprint

Our digital footprint

- List **three** data points you created *today* before 10:30.

  (Write them down. Then we will share a few examples.)

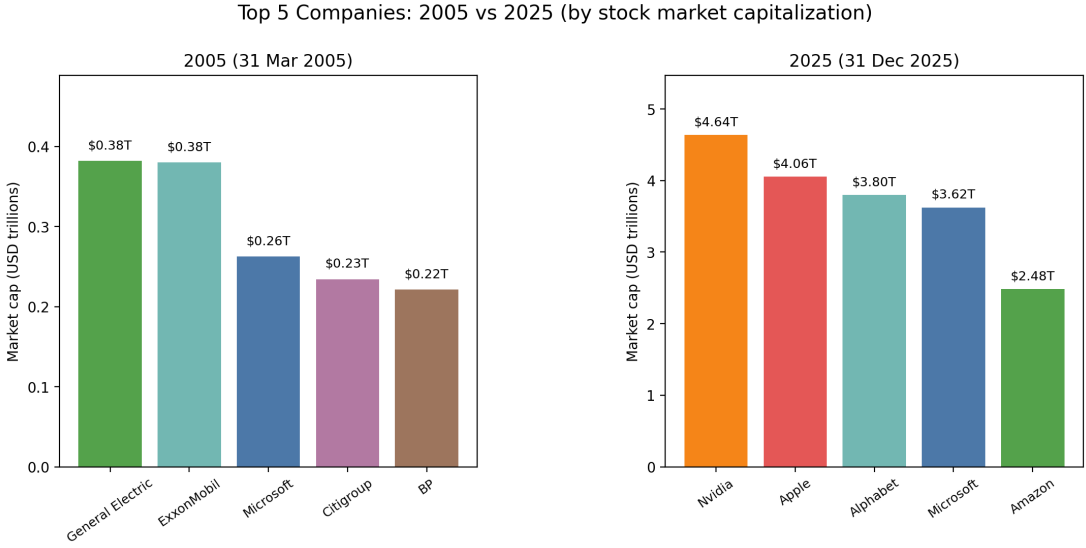
- Some examples:
  - Location trace (phone / maps)
  - A message / search / song played
  - A payment or purchase (card / Bizum)
  - Transport tap-in (metro / bus)

Key takeaway

Most data is not "collected in a lab". It is produced automatically when we live.

What is the Data Economy?

Who are the most valuable companies today?



Where does the rise in the data economy come from?

1. Data storage/processing capabilities
   - Global data storage to exceed 200 zettabytes (one trillion GB) in 2025.
2. Data science and AI breakthroughs
   - Global AI software market to reach $1.81 trillion by 2030.
3. Data abundance from our digital footprint
   - Global data generation to reach 181 zettabytes in 2025.

Data as input into predictions

Statistical tool: Bayes' Law

Conditional probability

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{when } \mathbb{P}(B) > 0.$$

- Example: Draw 1 card from a 52-card deck. Let $A = \{\text{Ace}\}$ and $B = \{\text{Spade}\}$.
- Then $\mathbb{P}(B) = 13/52 = 1/4$ and $\mathbb{P}(A \cap B) = 1/52$, so

$$\mathbb{P}(A \mid B) = \frac{1/52}{1/4} = \frac{1}{13}.$$

Bayes' Law

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

- $\mathbb{P}(A)$ prior (before data)
- $\mathbb{P}(B \mid A)$ likelihood (data quality)
- $\mathbb{P}(A \mid B)$ posterior (after data)
- $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B \mid A) = 1/4$, $\mathbb{P}(B) = 1/4$,

$$\mathbb{P}(A \mid B) = \frac{(1/4)(1/13)}{1/4} = \frac{1}{13}.$$

Example 1: a COVID-style test

Setup

- Prior prevalence: $\mathbb{P}(\text{infected}) = 1\%$.
- Test sensitivity: $\mathbb{P}(+ \mid \text{infected}) = 95\%$.
- Test specificity: $\mathbb{P}(- \mid \text{not infected}) = 98\%$ or $\mathbb{P}(+ \mid \text{not infected}) = 2\%$.

- **Question:** What is the probability that I have COVID if I get one positive test (data)?
- We need to compute $\mathbb{P}(\text{infected} \mid +)$.

Example 1 (solution)

- We want $\mathbb{P}(I \mid +)$ where I means infected.

$$\mathbb{P}(I \mid +) = \frac{\mathbb{P}(+ \mid I)\mathbb{P}(I)}{\mathbb{P}(+ \mid I)\mathbb{P}(I) + \mathbb{P}(+ \mid \textit{not } I)\mathbb{P}(\textit{not } I)}.$$

Plug numbers:

$$\mathbb{P}(I \mid +) = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.02 \cdot 0.99} = \frac{0.0095}{0.0095 + 0.0198} \approx 0.324.$$

- One positive test raises the probability from 1% to about **32%**. That is a big update.
- What if we take a second test? The posterior becomes the prior!

Key takeaway

Bayes Law allows us to incorporate data to change our beliefs. More data, more precise beliefs.

Why Bayes is the core of the data economy

- A company observes data (clicks, sales, sensor readings).
- Data updates beliefs about unknowns (demand, risk, consumer tastes).
- Updated beliefs change decisions (prices, inventory, hiring, ads).
- **Better decisions $\Rightarrow$ better performance $\Rightarrow$ more activity $\Rightarrow$ more data**

Key takeaway

Data is valuable because it changes decisions towards the most profitable.

Example 2: Predict demand and hire workers

Ice-cream sales

A manager at a classic ice-cream shop in Barcelona must decide how many workers to schedule.

- State: demand is **High** (H) or **Low** (L).
- Prior: $\mathbb{P}(H) = 0.4$ (based on season).
- Data: early-morning foot traffic $S \in \{\text{Busy}, \text{Quiet}\}$.
  - $\mathbb{P}(\text{Busy} \mid H) = 0.8$
  - $\mathbb{P}(\text{Busy} \mid L) = 0.3$

Question: If early-morning foot traffic is Busy, what is the probability that ice-cream demand is high?

Example 2 (solution)

- We need to compute $\mathbb{P}(H \mid \text{Busy})$:

$$\mathbb{P}(H \mid B) = \frac{\mathbb{P}(B \mid H)\mathbb{P}(H)}{\mathbb{P}(B \mid H)\mathbb{P}(H) + \mathbb{P}(B \mid L)\mathbb{P}(L)} = \frac{0.8 \cdot 0.4}{0.8 \cdot 0.4 + 0.3 \cdot 0.6} = \frac{0.32}{0.32 + 0.18} = \frac{0.32}{0.50} = 0.64.$$

- Data (traffic) increases the belief that demand will be high (from 40% to 64%).
- Which situation makes data *more* valuable?
  1. A very uncertain situation (say $\mathbb{P}(H) = 0.5$)
  2. A very certain situation (say $\mathbb{P}(H) = 0.9$)

Key takeaway

Data is more valuable when we are more uncertain; there is more "updating" of our beliefs. The more informative the data, the stronger the update.

How Do Firms Profit from Data?

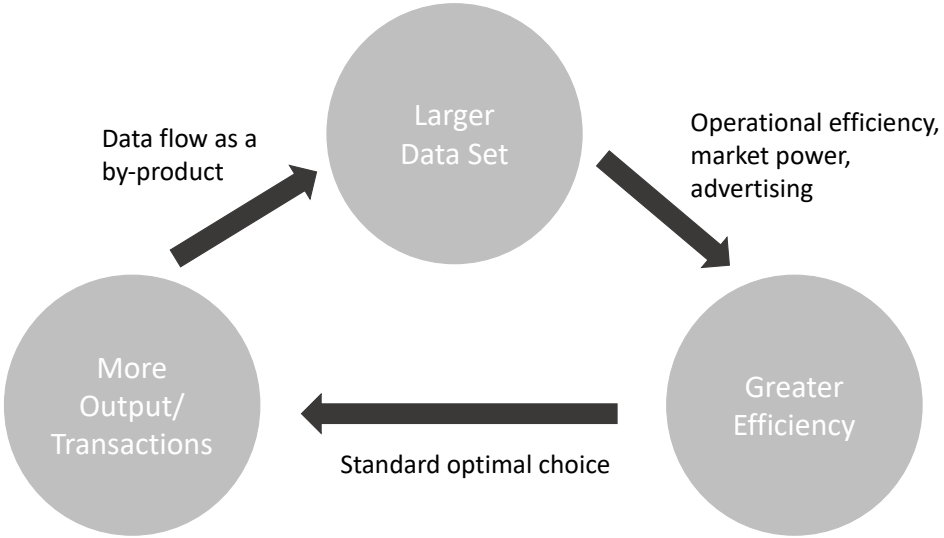
How Do Firms Profit from Data?

1. **Better prediction and data-driven decisions**
   - Anticipate shocks to costs and revenues
2. **Improve resource allocation in uncertain environments**
   - Better inventory management
   - Targeted advertising
3. **Develop new business models**
   - Data harvest, data trade, data platforms
4. **Create market power**
   - Customer capital, barriers to entry and entrepreneurship

Break Time! (30 mins)

Data Feedback Loop

The Data Feedback Loop



Example: Platforms around us

Pick one platform (T-mobilitat, Wallapop, Glovo, Bizum, Spotify)

1. What transactions generate data?
2. How does the platform use it to improve predictions?
3. Where do you see the feedback loop?
4. Is service better with more users/transactions?
5. What are the effects on competition?

Data Accumulation

Data stock and data flows

- Key objects
  - **Data stock** D_t: the history of data, summarized by how *precise* are our predictions
  - **Data flow** δ_t: new information added today (more precision).
  - **Data depreciation** ζ: If the world changes, old data predicts less well.
- Data accumulation:

$$D_{t+1} = (1 - \zeta)\, D_t + \delta_t.$$

- Very similar to physical capital (machinery, equipment, intangibles...)

$$K_{t+1} = (1 - \zeta)\, K_t + I_t.$$

Key takeaway

Data behaves like physical capital: it accumulates, but it also **depreciates**.

Where does the new data flow δ_t come from?

- **Three main sources:**
  1. First-party: collected directly from own activity
     - consumers, users, employees, suppliers, customers ...
  2. Second-party: observed, shared or bought
     - observing competitors or market trends
  3. Third-party data: aggregated by market prices or intermediaries
     - platforms, data brokers
- **Who also observes the data matters**
  - (i) Public data is shared by everyone $\Longrightarrow$ More useful if we want to coordinate
  - (ii) Private data is owned by firm i $\Longrightarrow$ More useful if we want to differentiate

Divergence and Convergence

Consequences of the data feedback

Data Feedback $\Rightarrow$ Divergence and Inequality

- **Superstar firms:** big firms get bigger, a few firms dominate
- **Market power:** large firms set rules (fees, ranking, access to customers).
- **Data poverty trap:** small firms learn slowly and may stay small

Quick debate: Are superstar firms good or bad for consumers? For workers?

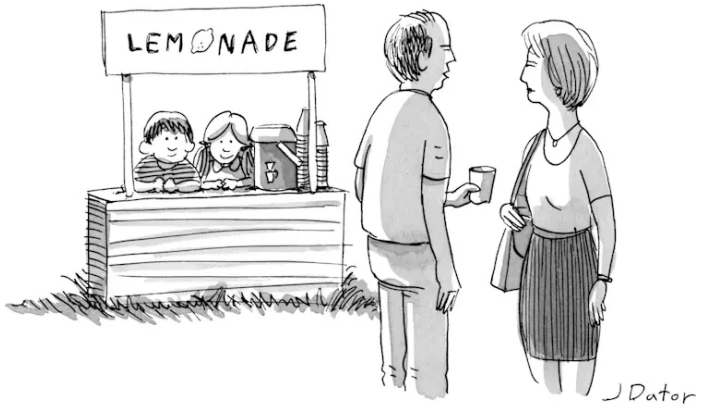
What could push the other way?

Convergence Forces

- **Data sharing**: open data, open source
- **Data portability and interoperability**: users can move their data
- **Data markets**: small firms can buy analytics
- **Competition policy and privacy rules**: GDPR and Digital Service Act in Europe

Data Barter

Data price is zero. Data value is positive.

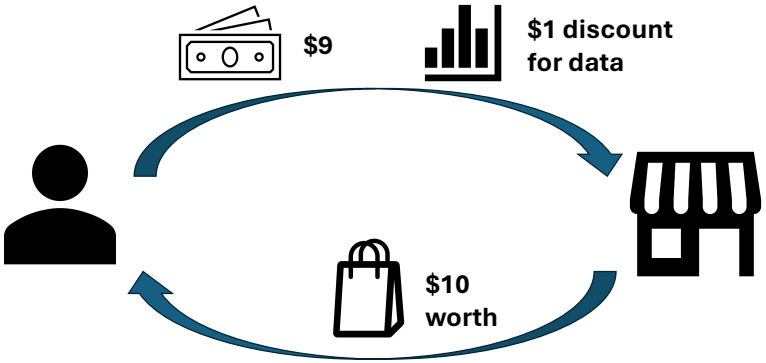


"It's free, but they sell your information."

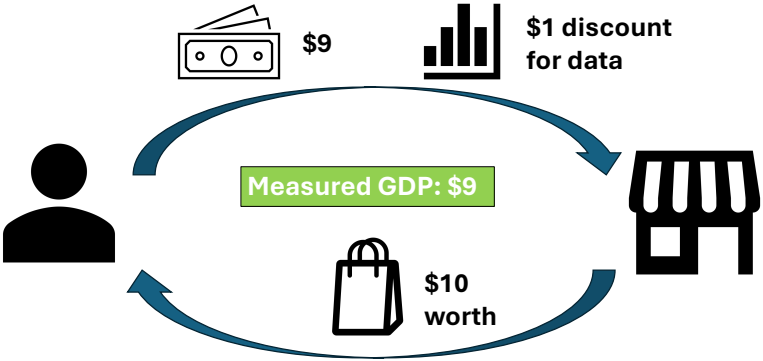
CartoonStock.com

Data barter (partial barter)

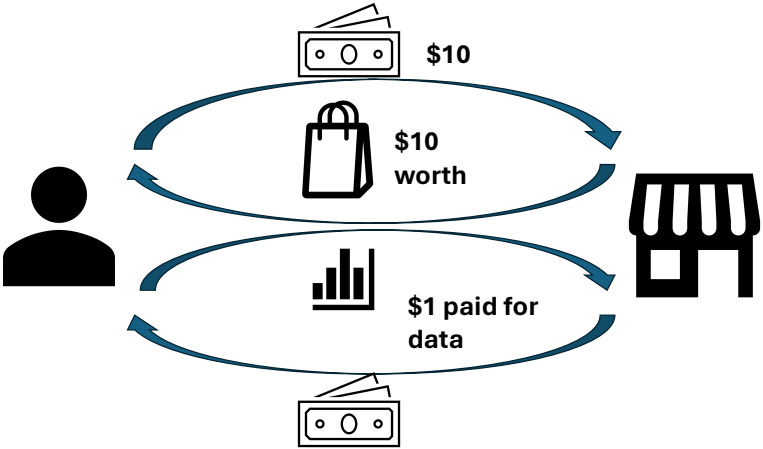
- Many services feel "free" (zero monetary price) but are paid with data or attention.
- Example: Supermarkets give discounts in exchange for purchase data.



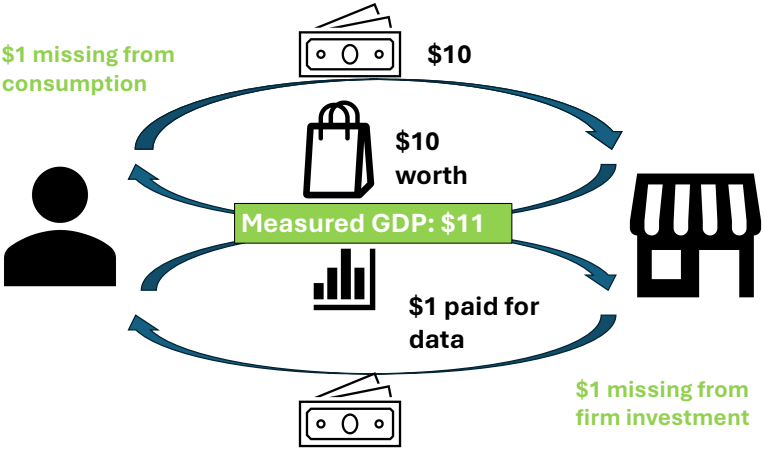
Bundling: two transactions with one price



Unbundling: revealing the hidden data price



Unbundling: revealing the hidden data price



GDP and missing data investment

- **GDP = Gross Domestic Product**
- Standard expenditure approach

$$\text{GDP}_{\text{exp}} = C + I + G + NX.$$

- Corrected approacth (data discount $= D$)

$$\text{GDP}_{\text{exp}}^{\text{corrected}} = (C + D) + (I + D) + G + NX.$$

Key takeaway

A data discount can imply **missing consumption and missing investment**. If we could observe a **privacy-protected price** and a **data-sharing price**, the difference reveals the value of data.

A “loyalty card” GDP correction

Exercise

Suppose a supermarket gives you **€1 off** each basket if you use a loyalty card.

- There are **2 million** loyalty baskets per month.

1. What is the monthly value of the discount D?
2. If the correction adds D to consumption and D to investment, what is the total monthly GDP correction?

- Discount value: $D = €1 \times 2{,}000{,}000 = €2{,}000{,}000$ per month.
- Total correction: $2D = €4{,}000{,}000$ per month.

Data Valuation

Valuation methodologies

1. **Cost-based:** How much does it cost to collect, clean, store, and protect the data?
   - Think: servers, engineers, labeling, cybersecurity.
2. **Revenue-based:** How much extra revenue or profit does the data generate?
   - Example: better targeting $\Rightarrow$ higher sales
3. **Market-based:** What price do similar datasets (or data services) trade for?
   - If data is bought/sold, the price contains information about value.
4. **Complementary inputs:** Data is useful only with other inputs that turn it into knowledge.
   - Labor (data scientists), cloud computing, software, and algorithms.
   - Proxy: how much firms spend on these complements to make data productive.

Example of revenue valuation

Setup

A small online shop improves its recommendation system using better data.

- Monthly visitors: 200,000
- Conversion rate increases from 2.0% to 2.3%
- Profit per sale: €12

What is the monthly profit gain from the data improvement?

- Sales before data: $200{,}000 \times 0.020 = 4{,}000$.
- Sales after data: $200{,}000 \times 0.023 = 4{,}600$.
- Extra sales: 600.
- Profit gain: $600 \times €12 = €7{,}200$ per month.

Wrap-up

Wrap-up

Key takeaway

1. **Bayes' Law** is the logic of how data changes beliefs and improves prediction.
2. The **data feedback loop** links activity $\rightarrow$ data precision $\rightarrow$ decisions $\rightarrow$ more activity.
3. This can increase **inequality** (superstar firms, market power).
4. **Data barter** hides a price; GDP can miss part of the data economy.
5. Valuing data requires a method: costs, revenues, or market prices.

Final question: Should data have an explicit price?

Exercise

Imagine your phone offers two plans:

1. **Free**: you share app-usage data
2. **€5/month**: no tracking

Questions:

- Which plan do you choose? Why?
- What does your choice imply about the value of your data?

Moltes gràcies!

isaac.baley@upf.edu