The Tragedy of Lonely Data

Elias Albagli

October 2017

Outline

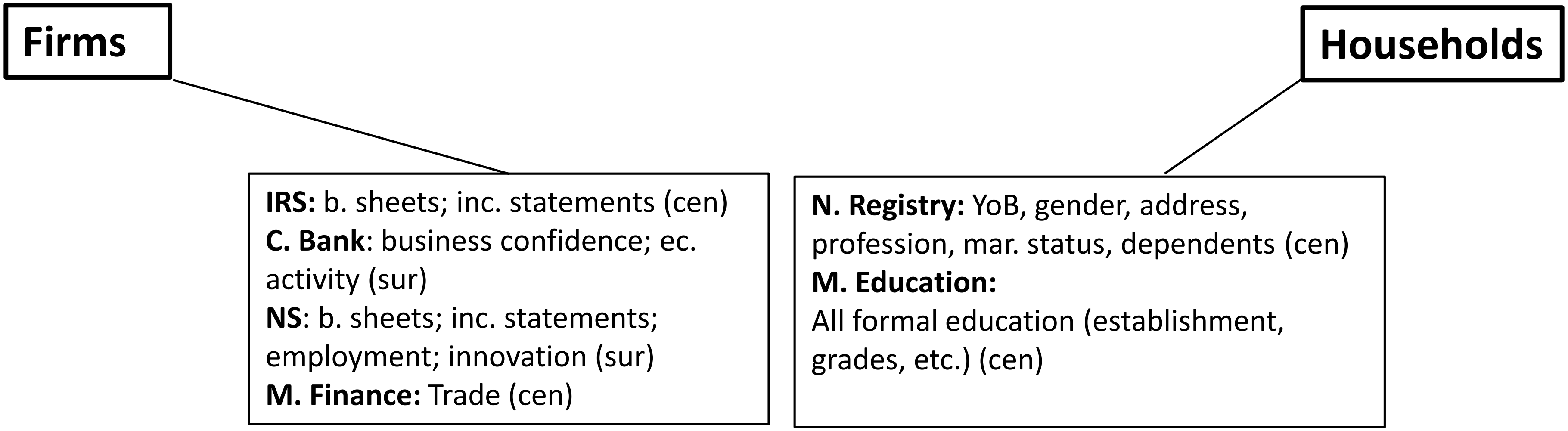
1. Chilean data is pretty good… but lonely

2. The potential: examples from academic research and applied public policy

3. Why we fail: some hypotheses

4. Where do we go from here: the Danish example

When it comes to research in economics, Chilean data is pretty good… even under advanced countries‘ standards



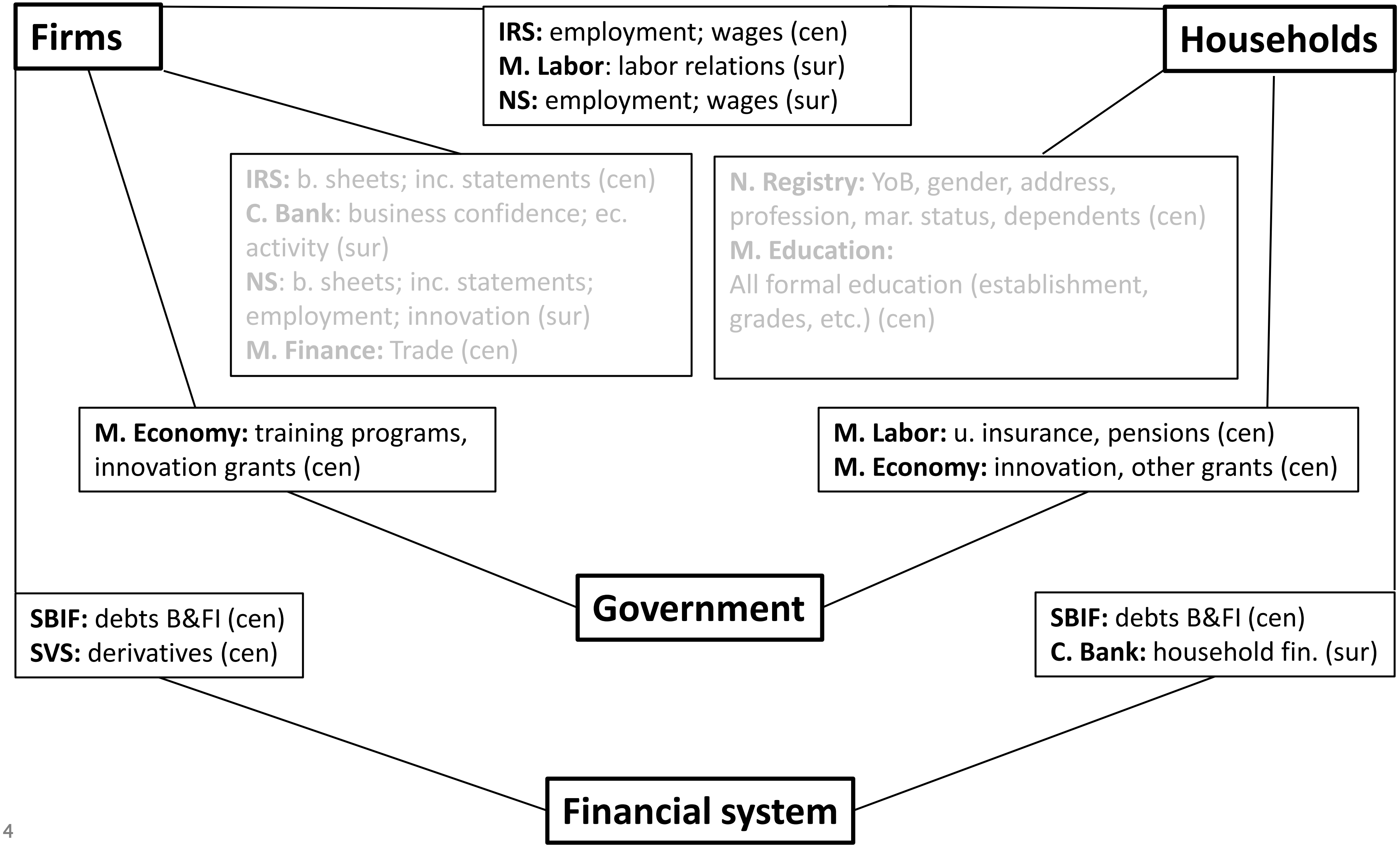
Firms

IRS: b. sheets; inc. statements (cen)
C. Bank: business confidence; ec. activity (sur)
NS: b. sheets; inc. statements; employment; innovation (sur)
M. Finance: Trade (cen)

Households

N. Registry: YoB, gender, address, profession, mar. status, dependents (cen)
M. Education:
All formal education (establishment, grades, etc.) (cen)

When it comes to research in economics, Chilean data is pretty good… even under advanced countries‘ standards



Firms

IRS: employment; wages (cen)
M. Labor: labor relations (sur)
NS: employment; wages (sur)

Households

IRS: b. sheets; inc. statements (cen)
C. Bank: business confidence; ec. activity (sur)
NS: b. sheets; inc. statements; employment; innovation (sur)
M. Finance: Trade (cen)

N. Registry: YoB, gender, address, profession, mar. status, dependents (cen)
M. Education:
All formal education (establishment, grades, etc.) (cen)

M. Economy: training programs, innovation grants (cen)

M. Labor: u. insurance, pensions (cen)
M. Economy: innovation, other grants (cen)

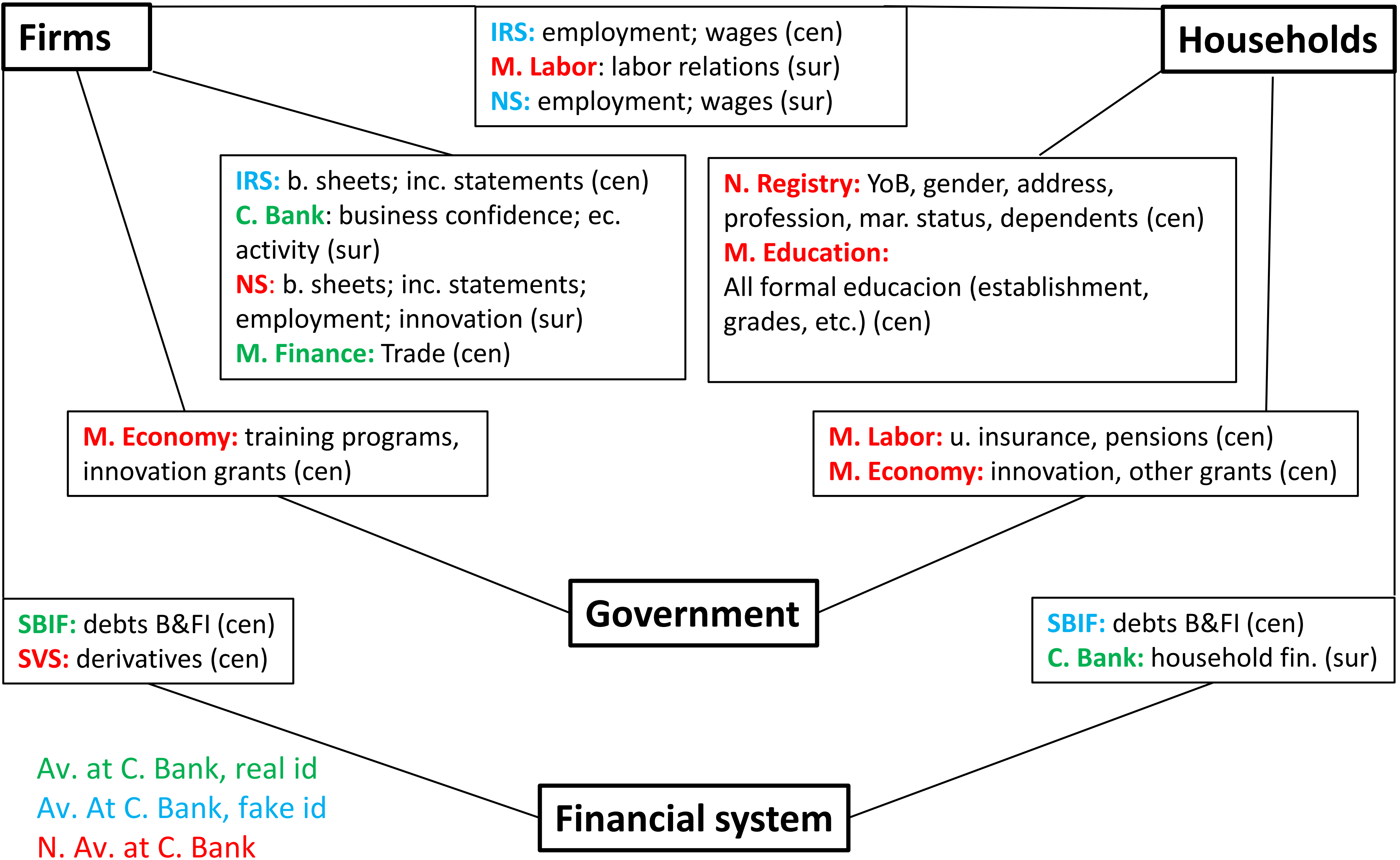
SBIF: debts B&FI (cen)
SVS: derivatives (cen)

Government

SBIF: debts B&FI (cen)
C. Bank: household fin. (sur)

Financial system

But existence and availability of merged information are two different things: Chilean data is, for the most part, lonely…



Firms

Households

Government

Financial system

IRS: employment; wages (cen)
M. Labor: labor relations (sur)
NS: employment; wages (sur)

IRS: b. sheets; inc. statements (cen)
C. Bank: business confidence; ec. activity (sur)
NS: b. sheets; inc. statements; employment; innovation (sur)
M. Finance: Trade (cen)

N. Registry: YoB, gender, address, profession, mar. status, dependents (cen)
M. Education:
All formal educacion (establishment, grades, etc.) (cen)

M. Economy: training programs, innovation grants (cen)

M. Labor: u. insurance, pensions (cen)
M. Economy: innovation, other grants (cen)

SBIF: debts B&FI (cen)
SVS: derivatives (cen)

SBIF: debts B&FI (cen)
C. Bank: household fin. (sur)

Av. at C. Bank, real id
Av. At C. Bank, fake id
N. Av. at C. Bank

Outline

1. Chilean data is pretty good… but lonely
2. **The potential: examples from academic research and applied public policy**
3. Why we fail: some hypotheses
4. Where do we go from here: the Danish example

Economics is full of interesting and important questions…

Is trade bad for you?

- Autor et al. (QJE, 2014): employment & wage effects from exposure to trade (China)
    - Individuals who worked in industries experiencing high import growth face lower earnings.. Loses are larger for low skill, low initial tenure, and low attachment to labor force. High-wage workers are better able to move across employers with minimal earnings losses and are more likely to move out of manufacturing: **import shocks impose substantial labor adjustment costs, unevenly distributed across workers**.
- What about benefits from exporting? Here´s a nice paper to be written with Chilean data:
    - Use firm-level data on imports and exports from M. Finance (census), merge with data on employment and wages from IRS (census), demographic info from N. registry (census), and educational data from M. education (census).
    - Compute earnings and employment effects in both import substitution and exporting sectors. Compare effects by individual characteristics.
- Good luck merging the data!

Economics is full of interesting and important questions…

- Do credit constraints limit the expansion of young firms? Does it matter for aggregate productivity?
- Do firms respond to macroeconomic expectations? Do they adjust employment, capital, or both? How does it depend on credit constraints and labor market conditions?
- What are the main predictors of firms and consumers loan defaults?
- How can we detect firms misreporting wages (i.e., pension contributions avoidance)?
- Is the exchange rate pass through (to consumer prices) dependent on the cycle? Can firms manage margins to smooth it?
- Do grants given for innovation & training make a difference? How can we improve their impact?
- Are too many people studying journalism? (given unemployment and wages of journalists)
- What is the impact of unionization on wages? On labor productivity? On labor relations within the firm? What can we learn from good firm-level examples?
- Can *millennials* hold on to a job for more than 6 months? How many students are the first generation in their family to attend college? Are they discriminated (controlling on college grades, etc.)
- What is the real gender wage gap (controlling for college degree and grades, firm tenure, etc.)

…..few of which can be answered with isolated data

Outline

1. Chilean data is pretty good… but lonely

2. The potential: examples from academic research and applied public policy

3. **Why we fail: some hypotheses**

4. Where do we go from here: the Danish example

Incentive and coordination problems abound

- Separate mandates/separate incentives: some examples
    - IRS gathers data with tax-collecting purposes only.
    - INE conducts surveys to update aggregate series (demographics; labor market; activity).
    - Public adm. branches gather info related to specific objectives (M. labor: compliance with labor code; M. education: monitoring schooling outcomes; N. Registry: update demographic info).
        - **--> Little interest in merging info to understand deeper economic interactions.**
- Legal constraints (and their "interpretations")
    - Data protection laws impede data merging (or create enough ambiguity to justify inaction).
        - Statistical secrecy; tax-info secrecy; banking-info secrecy.
- In fact, institutions most concerned with understanding economic interactions through research have little or NO legal statue to obtain data
    - Central Bank: excluded from numerous data access agreements.
    - Universities/academics: limited access to limited data under unstructured (and unaccountable) institution-specific protocols.
    - A few notable exceptions rely on the perseverance (and luck) of individual researchers.

Why this is “a tragedy”

- Many developing countries´ government take suboptimal decisions due to lack of information. Indeed, gathering information is costly
  - We cannot give this excuse. We already spend the money collecting the data!
- If institutional agreements could be reached, a central repository of merged micro data could be available for government agencies, policymakers, and academics, at very low costs.
- **Our failure to do this has very tangible costs on our current state of knowledge**
  - Money is spent in public programs without a proper assessment of their actual impact.
  - Policy decisions are made with limited knowledge of possible effects.
  - Highly trained Chilean researchers would rather invest their time studying economic and social issues in other countries.
  - Highly trained international researchers unlikely to be interested in studying economic and social issues using Chilean data (this is not true for other countries).

Outline

1. Chilean data is pretty good… but lonely

2. The potential: examples from academic research and applied public policy

3. Why we fail: some hypothesis…

4. **Where do we go from here: the Danish example**

Some fun facts about Denmark

- You can safely park your baby´s stroller outside any restaurant (with the baby inside)
- They probably have the world´s best merged micro data system, housed at **Statistics Denmark (SD)**
  - Demography, education, elections, culture, religion; Employment, wages, wealth, consumption; Tax records, business sector, financial markets, foreign trade; Prices, national accounts, public finance, Geography, environment, energy. Oldest data from 1787. Most data from 1960´s.
- **Central registry model: intensive use of administrative data**
  - Individuals are characterized by ID numbers (personal; dwelling; workplace). All data collected by govt. agencies (ex: taxes; educational background) are reported to SD under ID numbers.
  - By law, SD must gather and make available for public sector branches merged micro data.
    - Surveys only complementary (also and with ID). Indeed, last census was 1980!
  - Communication strategy: SD grants researchers access to merged micro data (Statbank).
- Requisites and protocols
  - Research requests (sponsored by institutions) must specify objectives variables If approved, SD merges and anonymizes data; grants remote access to SD server.
  - No individual information extracted. Output (tables, regressions) monitored and sent via email.
- **→ Centralized databank with tested safety protocols probably safer than agency-specific, discretionary protocols**.

The medium-term consequence of getting your act together…

- Here are some titles illustrating the potential of having this data (all published in economic journals)
  - Labor markets:
    - Are skills firm-specific?; Returns to tenure, firm-specific human capital and worker heterogeneity; The Impact of Worker and Establishment-level Characteristics on Male–Female Wage Differentials; Adverse workplace conditions, high-involvement work practices and labor turnover.
  - Productivity and firm dynamics:
    - Firm patterns of entry and exit; The impact of R&D on productivity; Pay inequality and firm performance; The role of families in succession decisions and firm performance; Industrial clusters, firm location and productivity; Productivity growth and worker reallocation.
  - Trade:
    - The wage effects of offshoring; Human capital and wages in exporting firms; Exports, firm size, and firm dynamics; Offshoring, transition, and training; Micro-level wage effects of international outsourcing; Do multinational enterprises relocate employment to low-wage regions?
- I would risk saying: Denmark is the most overly-studied economy in the world (per capita terms)

How can we copy this model?

- A recent law proposal appears to have some elements that would facilitate the construction of a central repository of merged data for Chile: SEN law (National statistics system).
- Two important innovations:
  - Governance of INE (national statistics institute): favors independence and quality of statistics
  - Attributions of INE: can request information from any government branch (including SII)
    - In principle, this would allow merging all datasets mentioned previously
- However, there is no guarantee that INE will do so, much less make it available for researchers/institutions ("databases not regarded as confidential will be publicly available to researchers…")

Final Remarks:
Huge gap between existence & availability of merged micro data

- **Severely affects capacity to answer important questions and design good public policy.**
- This is not a money problem. Technological advances allow handling big data safely at low costs.
- Not really a problem of data confidentiality –data protection protocols have been established.
  - Indeed, centralized databank with tested safety protocols probably safer than alternative.
- It is rather a problem of **incentives and coordination**:
  - Govt. agencies use data for narrow purposes: no incentive to understand deeper economic interactions ("quality economic research" is not in anyone´s mandate). Ambiguous legal attributions provide perfect excuse to avoid inter-agency cooperation.
- Good news: we can learn from some remarkable examples
  - Denmark: institutional framework not only allows, but indeed mandates, creation of central data repository. SD makes (anonymized) micro-data available to researchers.
  - Indeed, there are plenty of international scholars waiting for their turn to access Danish data. This makes the Danish economy one of the most studied (certainly in per capita terms).
- New statistical regulatory framework in Chile (to be approved):
  - Enough attributions to request & merge data, but enough discretion not to do so (or make it available, for that matter). Important to push in this direction going forward!