Data-Intensive Innovation and the State: Evidence from AI Firms in China

Beraja, Yang and Yuchtman

Discussion by Matilde Bombardini

BerkeleyHaas

Creative and important paper

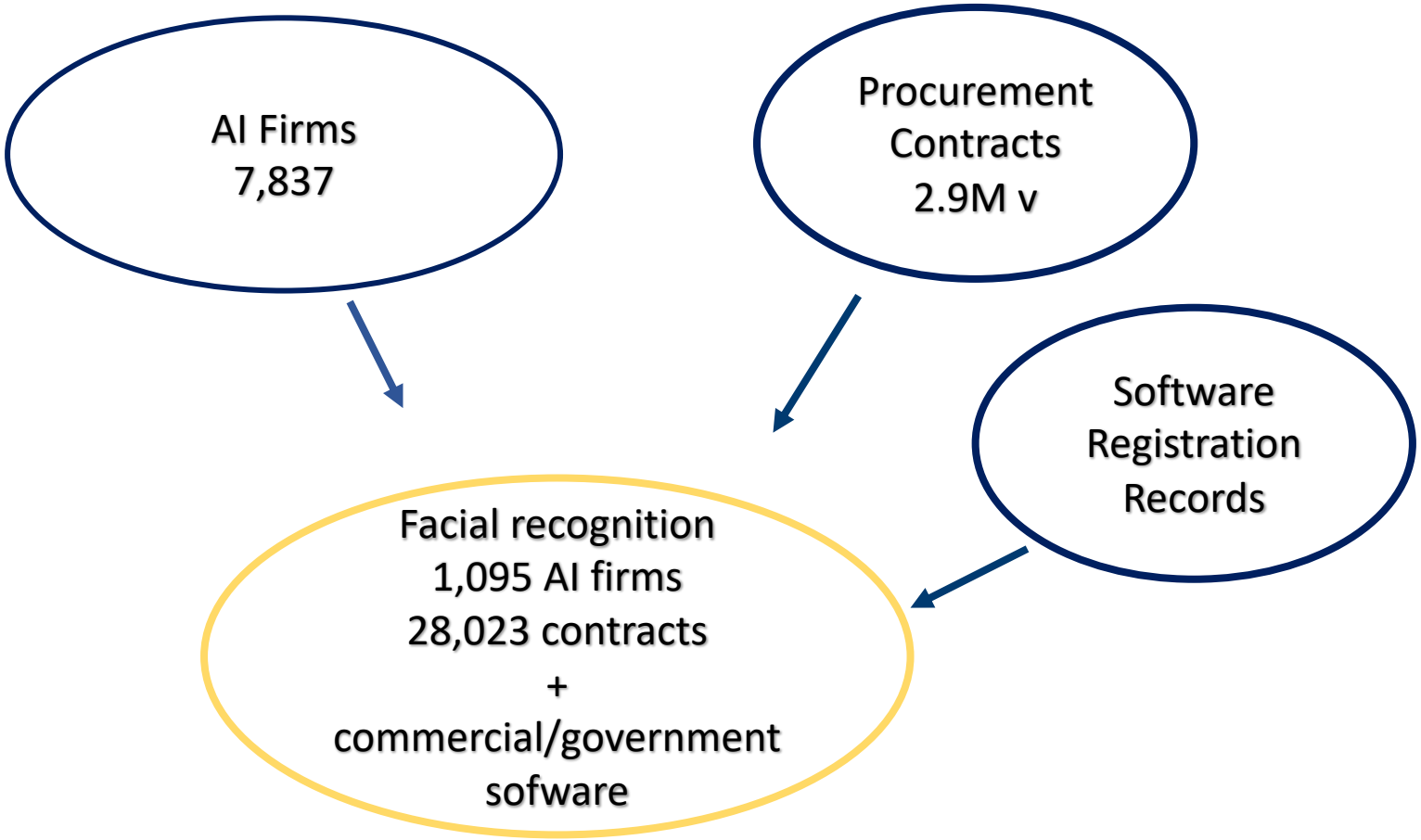
- Impressive data collection
- First paper to show that government AI software procurement has important spillovers on commercial AI development, together with Beraja, Kao, Yang and Yuchtman (2021)
- Contributes to literature on role of government in firm dynamics Moretti, Steinwender, van Reenen (2019)
  - In particular spillovers from procurement contracts: Ferraz, Finan and Szerman (2015)

Research Question

- Q: Does a firm that receive a data-intensive AI software government contract produce more or less commercial AI software?

- A: More (and almost exactly by the same amount as the government AI software)

Vast data collection effort



Data intensity measure (1)

- Current measure of prefecture data intensity:

$$\frac{\#\ surveillance\ cameras}{population}$$

- I am puzzled by this measure
- Assume
  - Harbin (pop 10.6M) has 100,000 cameras
  - Daqing (pop 2.9M) has 50,000 cameras
- Measure would imply that contracts in Daqing are more data intensive than in Harbin, but are they?

Data intensity measure (2)

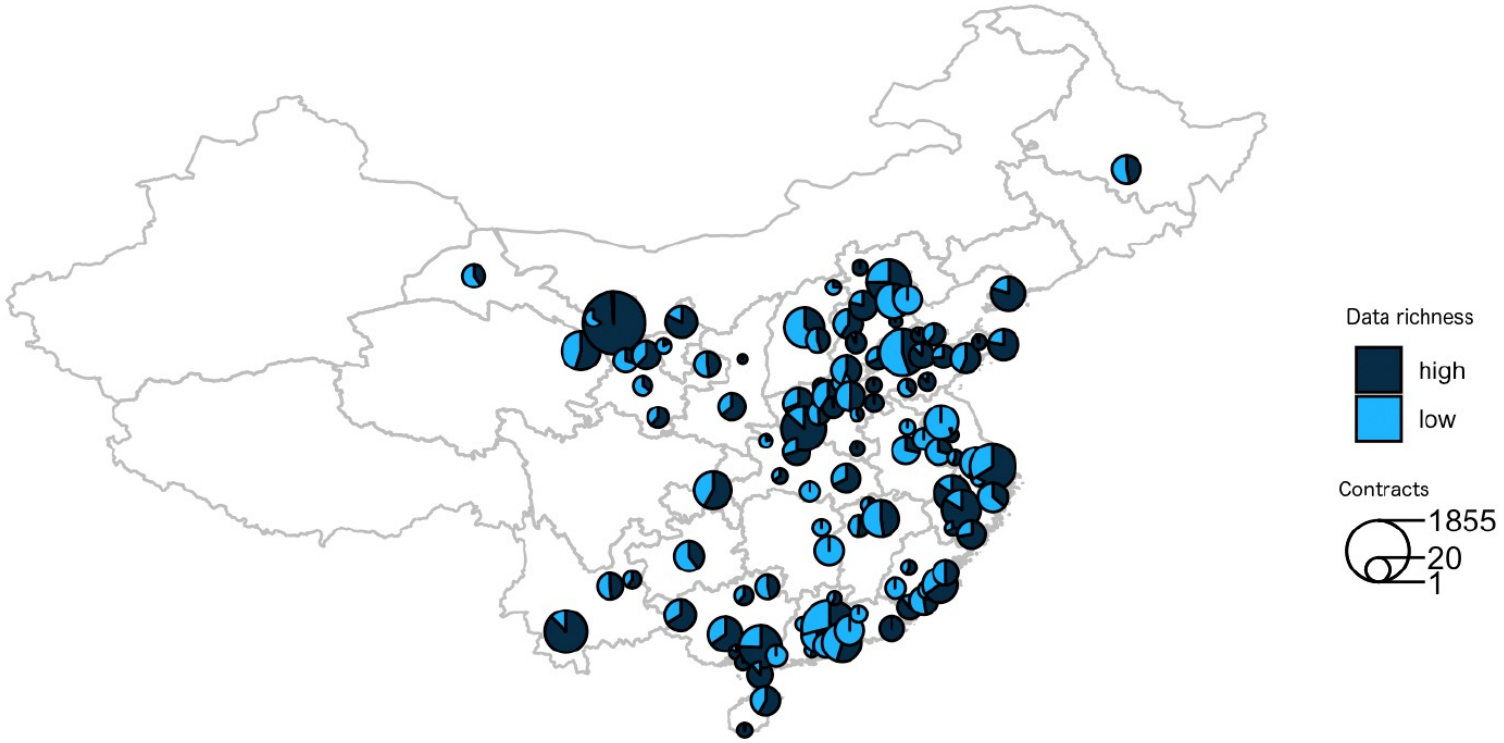
- My prior was that the measure would be a proxy for how many images are collected

$$\#\ surveillance\ cameras \times population$$

- Perhaps discounted by difficulty of actually capturing people's images in larger areas

$$\frac{\#\ surveillance\ cameras \times population}{(inhabitated)\ land}$$

Data intensity measure (3)



Data intensity measure (4)

- Yunnan, Guangxi have high intensity measures
- Should we assume that firms are located close to government units that request services?
  - Later exercise suggests that we shouldn't, but BKYY paper sees to imply that city A government contracts with city A firms
    - otherwise what does the rain instrument for protests do in that paper?
  - Important to describe the geography of AI firms and their government clients (does gravity hold?)

- To exclude measuring convergence/catch-up should include interaction

$$Province_i \times T_{it}$$

  - This only uses variation within province

Event-study specification

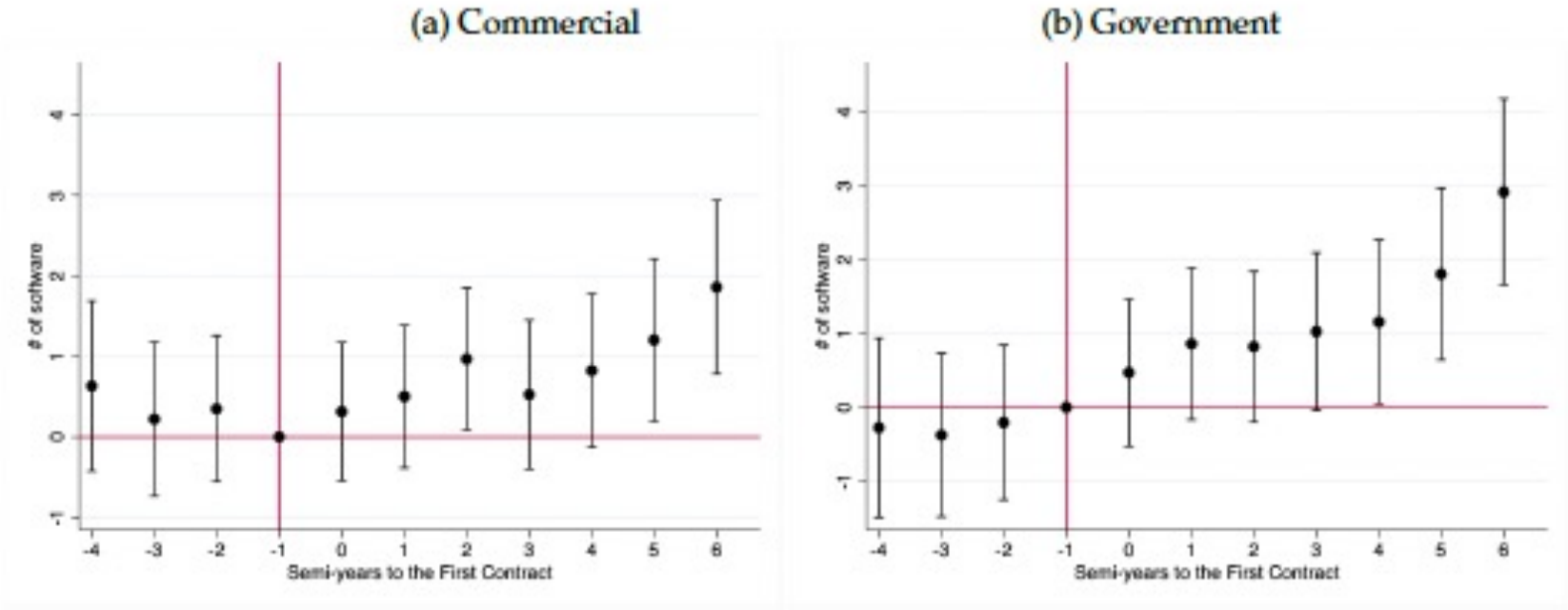
of software releases
(government or commercial)

$$y_{it} = \sum_T \beta_{1T} T_{it} Data_i + \sum_T \beta_{2T} T_{it} + \alpha_t + \gamma_i + \epsilon_{it}.$$

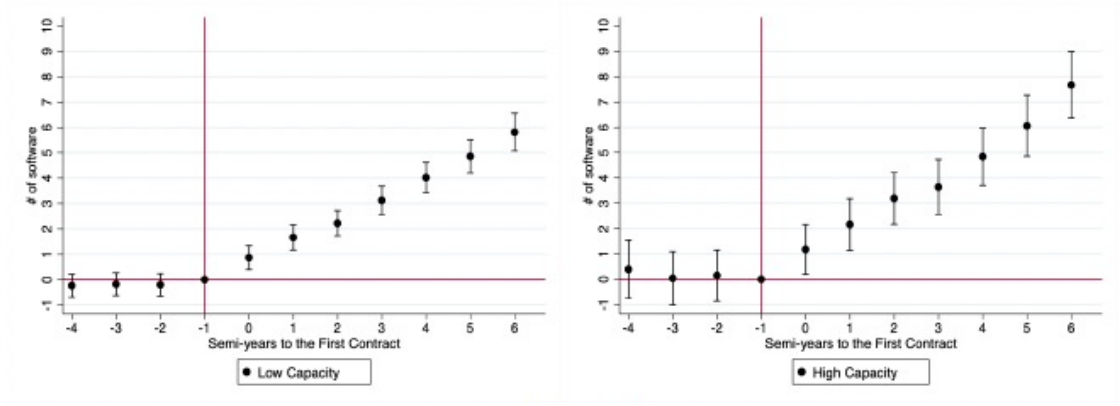
Notation I prefer: $\sum_{T=-4}^{6} \beta_{1T} \mathbb{I}(t - FirstContractPeriod_i = T) Data_i$

- Staggered treatment
- Heterogeneous treatment/Triple D-in-D: $Data_i$=1 if first contract data intensive

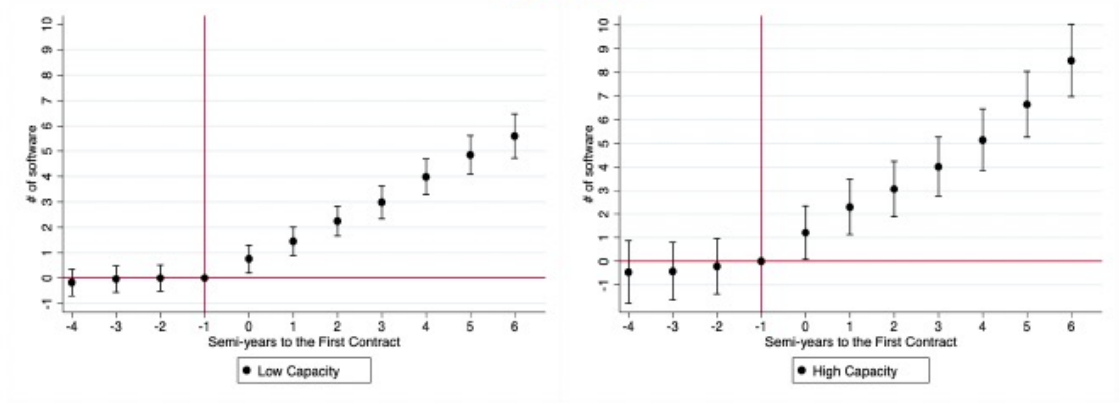
Main result


(a) Commercial
(b) Government
# of software
Semi-years to the First Contract
-4
-3
-2
-1
0
1
2
3
4
5
6

Main result broken down



(a) Commercial



(b) Government

Main result: comments (1)

- The coefficient estimates are incredibly similar for commercial-use and government-use software
  - Roughly 1 extra software release for high data intensity
  - This is out of an average of 10 (for both commercial and government use), so similar semi-elasticity as well
  - Timing is also remarkably similar (no delay in commercial)
- This seems a very large effect: 10% increase in software releases (with CRS this implies the data is equivalent to 10% subsidy)

Main result: comments (2)

- Government software increases by more for data-intensive contracts
  - Are these contracts larger? i.e. do they require more distinct pieces of software?
  - Are you implying that this is all indirect (i.e. excluding the first contract itself?)

Alternative stories addressed (1)

- Authors are very open about differences between high-data-intensive contracts and low-data-intensive contracts

	Any contract		Public security contract		Public security contract by surveillance capacity	
	Yes	No	Yes	No	High	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Firm characteristics						
Year firm established	2009.3	2013.8	2008.9	2011.4	2007.5	2010.0
	(6.4)	(4.2)	(6.4)	(6.1)	(7.0)	(5.7)
Capitalization (millions USD)	22.8	5.1	26.4	4.1	35.3	19.9
	(210.3)	(42.8)	(229.1)	(14.4)	(295.0)	(165.4)
Rounds of investment funding	0.9	0.5	1.0	0.3	1.0	1.0
	(1.7)	(1.9)	(1.8)	(0.8)	(1.8)	(1.7)
Observations	1,093	6,041	919	174	387	532
Panel B: Software production before first contract receipt						
Total amount of software	22.7	14.6	23.8	14.8	27.4	21.2
	(37.9)	(24.5)	(39.9)	(16.4)	(45.0)	(35.8)
Commercial	9.0	6.3	9.4	6.7	10.1	8.8
	(17.1)	(12.5)	(17.9)	(9.6)	(20.1)	(16.1)
Government	7.3	4.0	7.8	4.1	10.0	6.3
	(16.3)	(8.2)	(17.2)	(7.0)	(17.7)	(16.6)
AI (video)	1.6	1.0	1.6	1.4	2.0	1.3
	(3.8)	(2.8)	(3.9)	(3.2)	(4.9)	(3.0)
Data-complementary	9.2	5.6	9.7	5.9	11.3	8.6
	(16.7)	(10.8)	(17.5)	(8.4)	(19.4)	(16.0)
Observations	956	6,042	835	121	345	490

Alternative stories addressed (2)

Concern	Solution
Sorting into public security	Look only *within* public security firms
Selection of better firms	No pre-trends
Time-invarying characteristics	Firm fixed effects
Signalling of high quality	Use only subsidiaries of past suppliers
Time-varying effects of time-invarying characteristics (e.g. firm productivity)	$\Sigma_T \beta_{3T} T_{it} X_i$ (not time fixed effects)
Time-varying effects of contract characteristics (e.g. richer contract)	$\Sigma_T \beta_{4T} T_{it} C_i$

Alternative story not addressed

- I could only come up with one
- What if prefectures that use intensively surveillance cameras pick firms with better growth potential?

Mechanisms

- Paper distinguishes between two channels:
  - Direct: shareable data from government used directly in commercial software
  - Indirect: may ↑ or ↓ other inputs (non-data software)
- Very neat that they can observe data-complementary non-AI (DCNA) software

1. DCNA software ↑ (some indirect effect)
2. Control for pre-contract DCNA software importance (not just indirect effect)

Mechanisms: suggestions (1)

- Government seems to be giving a very useful input to private companies:
  - The usefulness should be correlated to current stock of data the firm has or the price the firm is currently paying for data (any proxies available?): can you show that for less constrained firms the effect is smaller?
  - The procurement cost (how much the government pays for the contract) should be taking into account the usefulness of the data for the firm: does the government agency with high data intensity get any discount on similar contracts compared to low intensity data agencies?

Mechanisms: suggestions (2)

- At the moment, results suggest 100% immediate shareability b/w commercial and government
  - Any feature that could split sample into more or less shareable would strengthen the result substantially
- E.g. we should not see smaller commercial ↑ for firms that also produce types of AI software for medical use

Additional thoughts/questions

- Given the immense benefit for AI firms, why is Chinese government so selective in "subsidizing" only a few lucky firms?
- Does Chinese government place any explicit restrictions on use of contract-related data?
- You mention government can shape direction of innovation: implies China should specialize in facial recognition AI (compare to other AI technology). Is it true?
- Elephant in the (Zoom) room: this productivity benefit for a few AI firms comes at a huge cost to personal freedom and privacy

Exciting research agenda

- This is going to be a very influential paper
- Thank you for inviting me to discuss it!