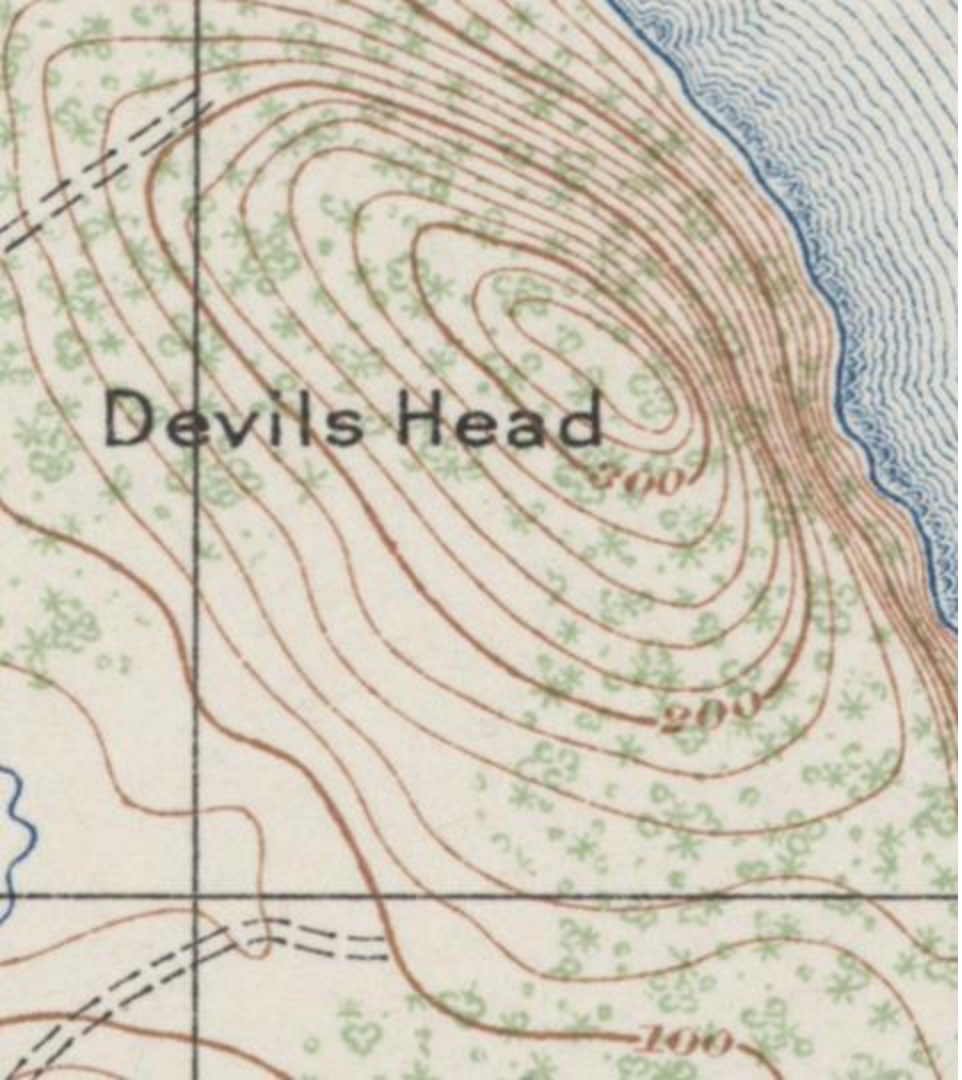


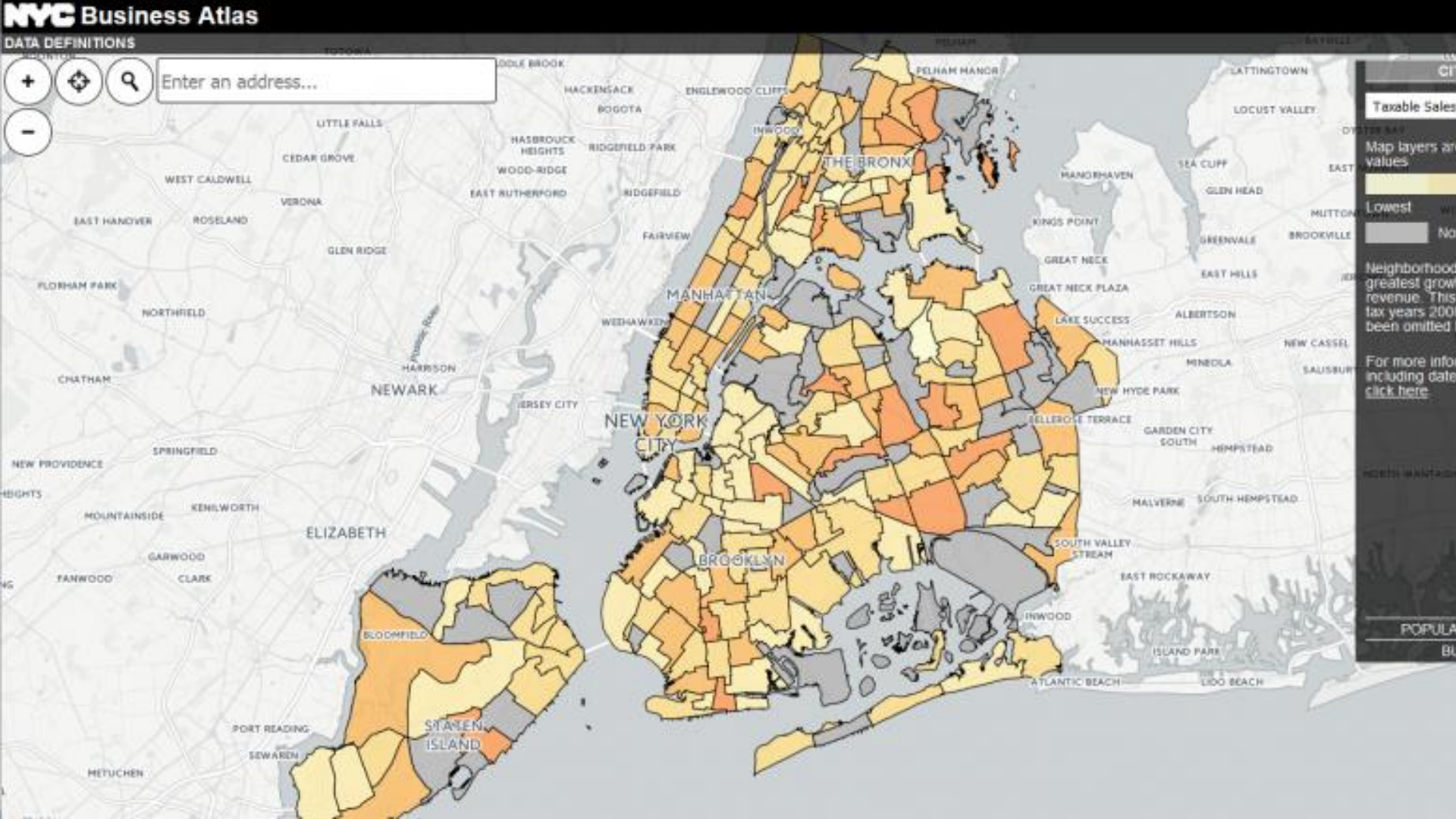
A Mapping Lens for Estimating (Public) Data Value

Abhishek Nagaraj

Data Innovation Lab

UC Berkeley and NBER

May 21st, 2024


NYC Business Atlas
DATA DEFINITIONS
Enter an address...
Taxable Sales
Map layers ar
values
Lowest
Neighborhood
greatest grow
revenue. This
tax years 200
been omitted
For more info
including date
click here
POPULA
BU
THE BRONX
MANHATTAN
BROOKLYN
NEW YORK CITY
NEWARK
ELIZABETH
STATEN ISLAND
JERSEY CITY
HACKENSACK
BOGOTA
ENGLEWOOD CLIFFS
PELHAM MANOR
LATTINGTOWN
LOCUST VALLEY
LITTLE FALLS
CEDAR GROVE
WEST CALDWELL
VERONA
EAST HANOVER
ROSELAND
GLEN RIDGE
FLORHAM PARK
NORTHFIELD
CHATHAM
SPRINGFIELD
NEW PROVIDENCE
MOUNTAINSIDE
KENILWORTH
GARWOOD
FANWOOD
CLARK
HASBROUCK HEIGHTS
WOOD-RIDGE
EAST RUTHERFORD
RIDGEFIELD PARK
RIDGEFIELD
FAIRVIEW
WEEHAWKEN
HARRISON
INWOOD
MANORHAVEN
SEA CUFF
GLEN HEAD
KINGS POINT
GREENVALE
BROOKVILLE
GREAT NECK
GREAT NECK PLAZA
EAST HILLS
LAKE SUCCESS
ALBERTSON
MANHASSET HILLS
NEW CASSEL
MINEOLA
NEW HYDE PARK
BELLEROSE TERRACE
GARDEN CITY SOUTH
HEMPSTEAD
MALVERNE
SOUTH HEMPSTEAD
SOUTH VALLEY STREAM
EAST ROCKAWAY
INWOOD
ISLAND PARK
ATLANTIC BEACH
LIDO BEACH
BLOOMFIELD
PORT READING
SEWAREN
METUCHEN

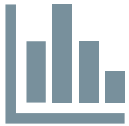
How would you value the New York Business Atlas?

Intended Beneficiaries

Entrepreneurs and Small Business Owners	Community that will make the most direct use of data housed in the Business Atlas. Improved decision-making capabilities engendered through free access to market research data and analytics that would normally come at great cost. Evidence of market opportunities provided by the Atlas can be useful for securing financing and investment for new businesses.

Agenda for today

Why do we need to value public data?



Challenges with valuing public data

A framework to value data

Case studies from my own research

Why should we value public data?

- Financial costs incurred when collecting and maintaining public data
- For instance, the Mayor's Office of Data and Analytics had to pay for salary, acquiring data assets, and providing training sessions

Why should we value public data?

- There are privacy costs incurred when publicly releasing data

- For instance, some of the underlying data powering Business Atlas contained personally identifiable information (e.g. sales tax data from the DOF), so it needed to be anonymized.

ACDEB
Advisory Committee on Data for Evidence Building

Easier to quantify costs, but value is harder …

United States™ Census Bureau

H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018

115th Congress (2017-2018)

Challenges associated with valuing data

Tracing

Spillovers

Diffuse

Counterfactual

1. Tracing

It is difficult to trace who uses public data, how, and for what purpose

a. Outside of academic contexts, users often do not leave behind a public record of their adoption
b. This is not unlike the challenge economists face when valuing internet-based technological innovations, such as social media

Challenges associated with valuing data

Individuals may not be aware that data is influencing their decisions

a. There are often positive spillovers to people that never directly use the data, let alone know it exists
b. Simply counting the number of data downloads and requests represents the ‘tip of the iceberg’

Challenges associated with valuing data

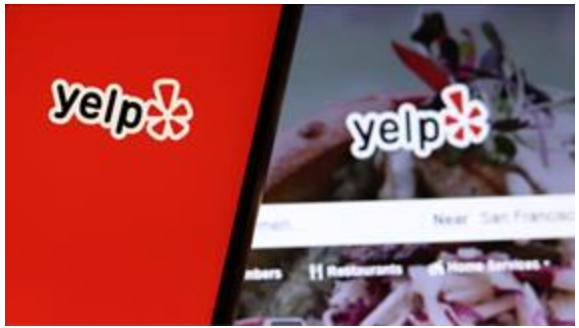
It can be prohibitively challenging to quantify the value of data empirically

a. Wide range of possible applications (predicting crop yields, forecasting weather investing, improving urban design)
b. Unclear *when* a dataset contributes to a decision, or *which subset* does
c. It can be difficult to capture dynamics beyond simple point-in-time estimates of value

Challenges associated with valuing data

The value of public depends on the counterfactual case: what *would* have happened had the data not been available?

a. Is there a reasonably close substitute?
b. Would the data actually change the decision outcomes?

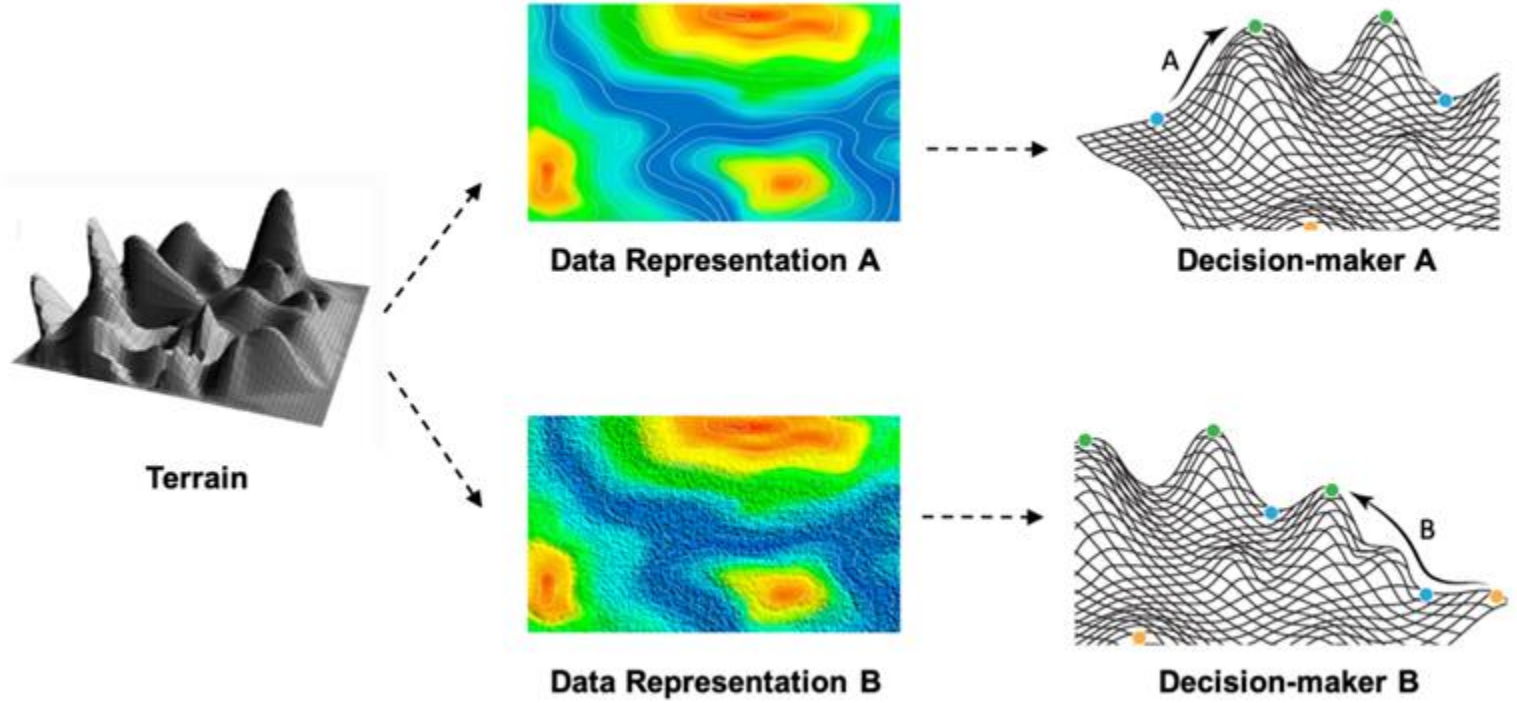


Solution: A Mapping “Lens” to Value Data

Data as a map

Each dataset essentially acts like a cartographic map: it provides an (imperfect) representation of the landscape of entities it describes

A mapping framework to estimate data value



The mapping framework in a nutshell

01

Define the relevant terrain and the relevant data-driven decisions

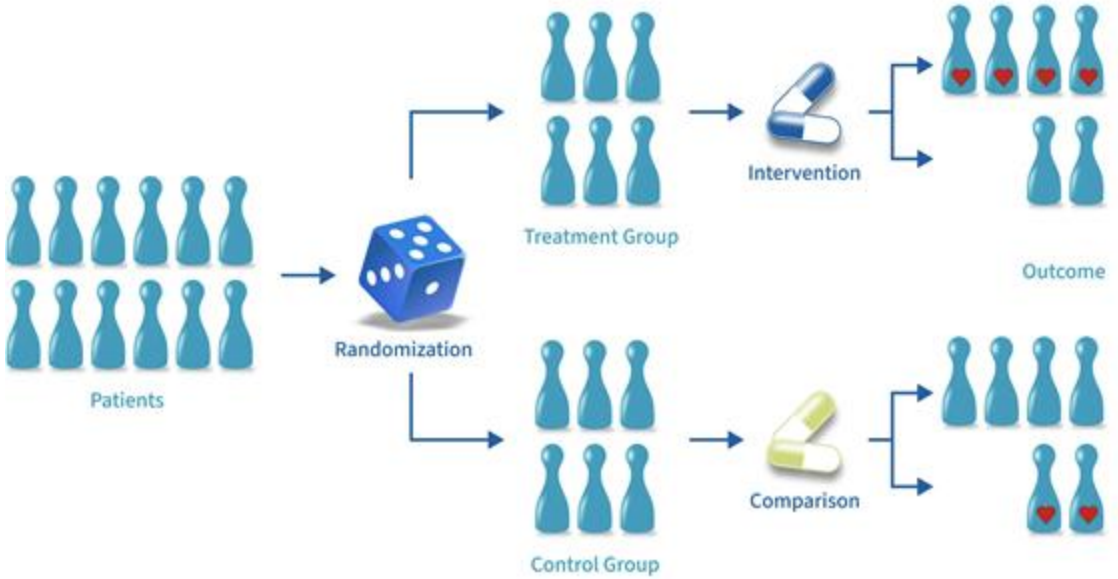
02

Specify alternate representations of the terrain; these alternate representations could reflect differences in coverage, access, cost, and so on.

03

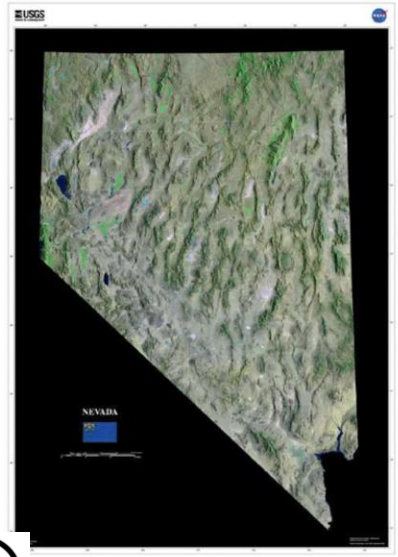
Map differences in decisions to differences in the data representation regime

The gold standard: randomized control trials (RCTs)



Borrowing from the identification revolution

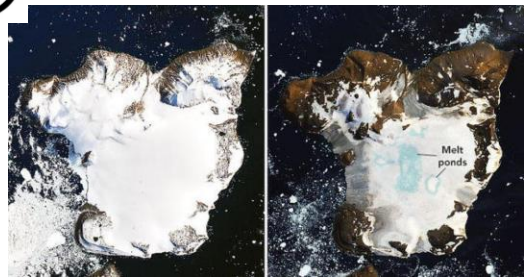
4 case studies from my own research



1

2

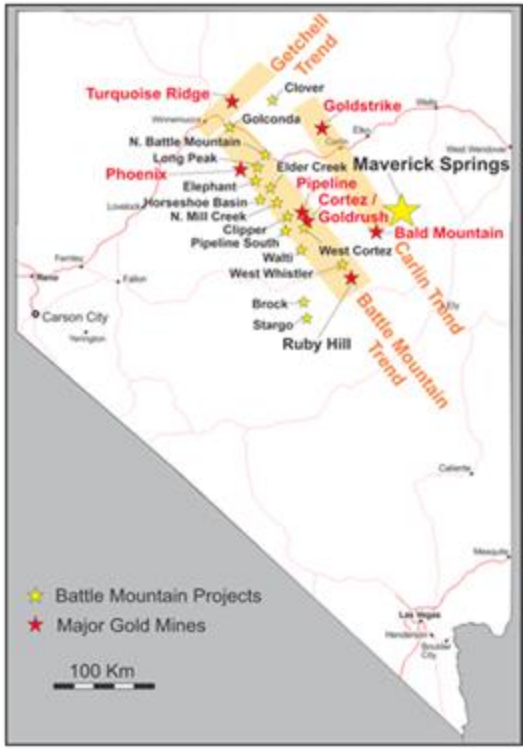
3



4

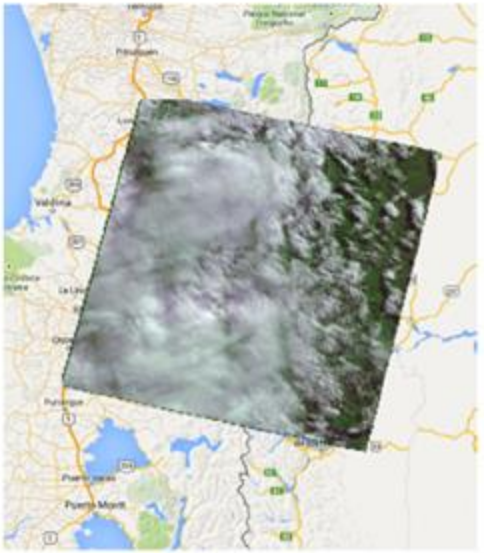
Example 1: the private impact of public (Landsat) data

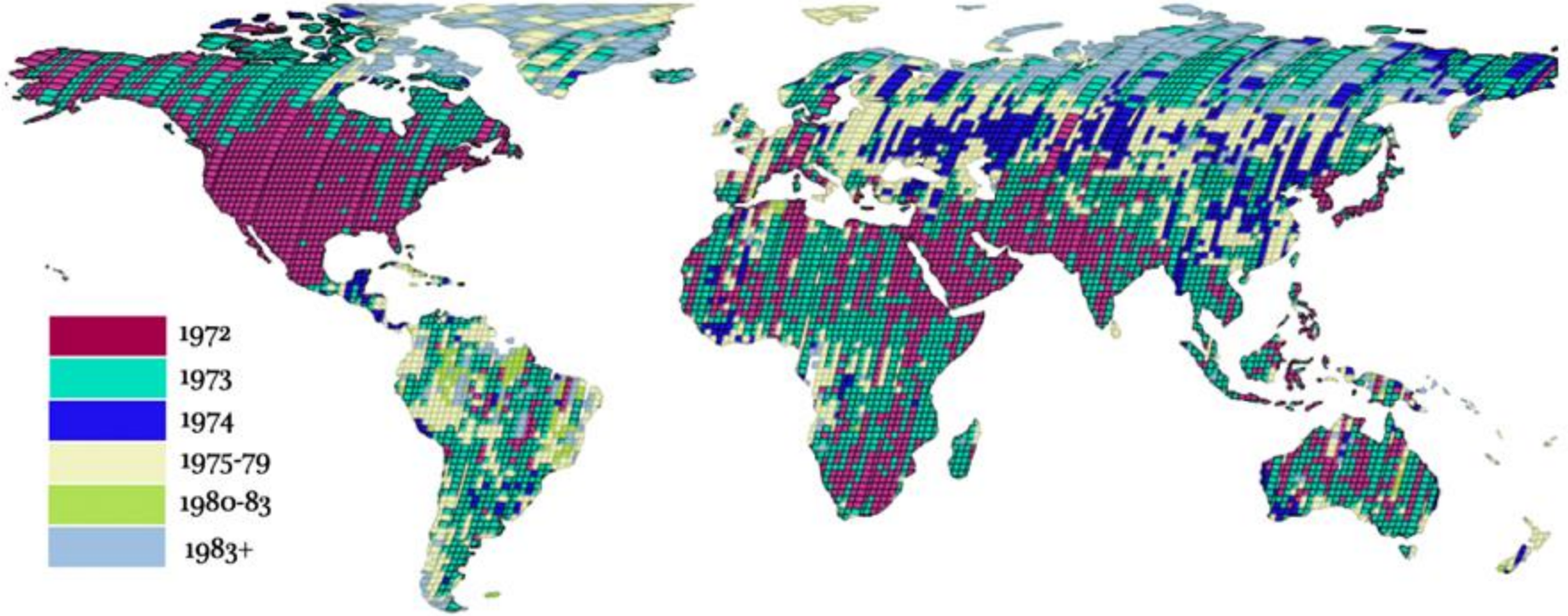
(from: Rowan & Wetlaufer 1975)



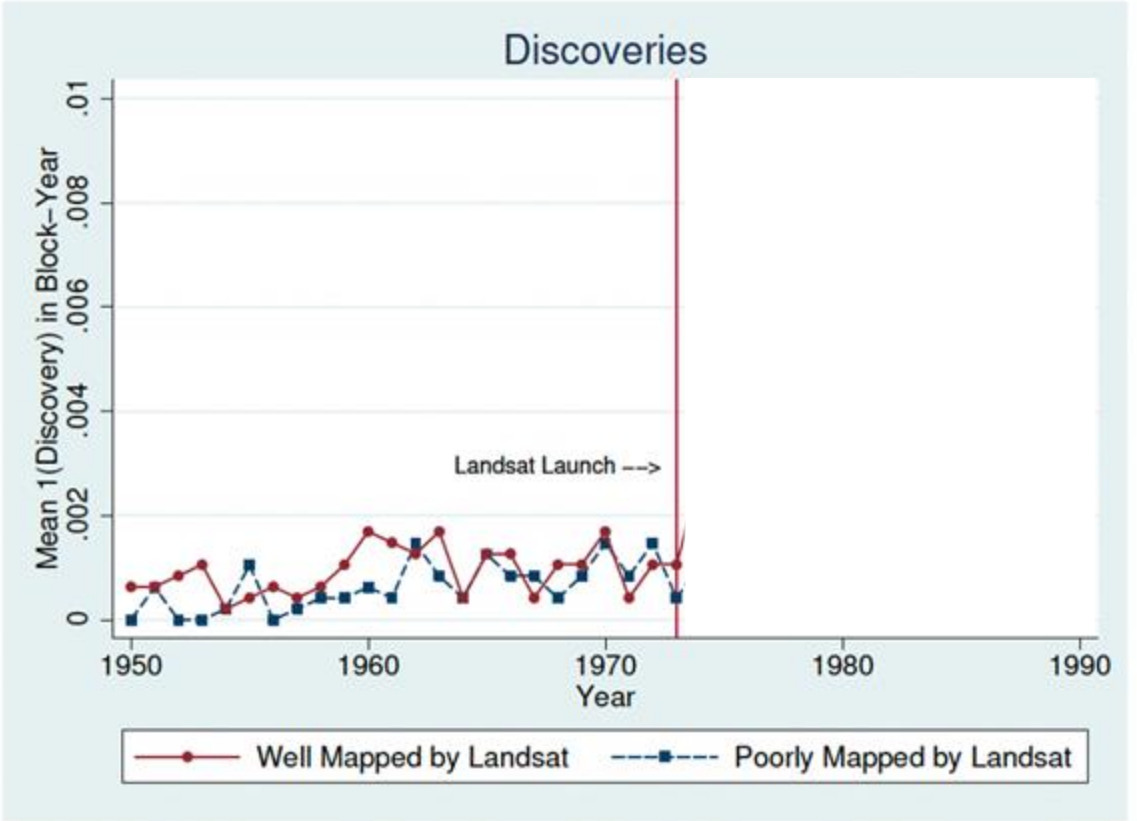
Example 1: the private impact of public (Landsat) data

Random variation: technical errors or cloud coverage limited data quality

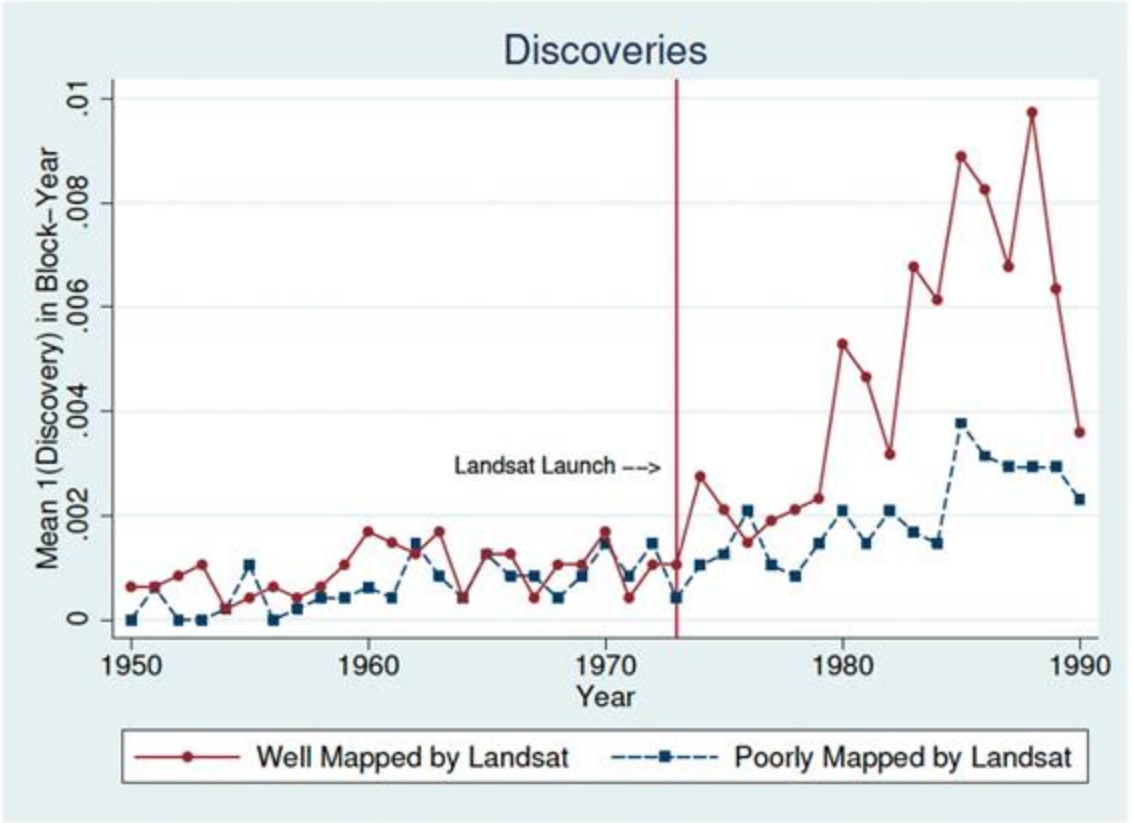



1972
1973
1974
1975-79
1980-83
1983+

Result 1: effect on discovery of new gold deposits



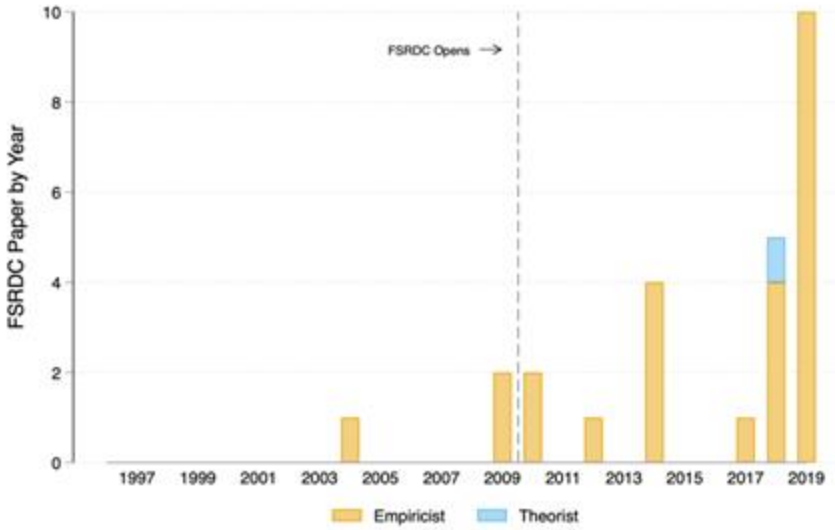
Result 1: effect on discovery of new gold deposits



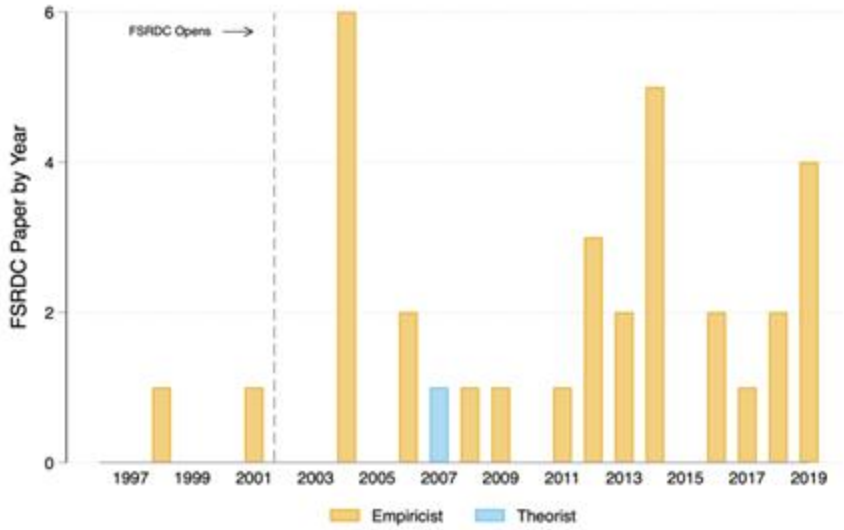
Example 2: Census Data and academic research

Example 2: Census Data and academic research

(i) University of Michigan-Ann Arbor

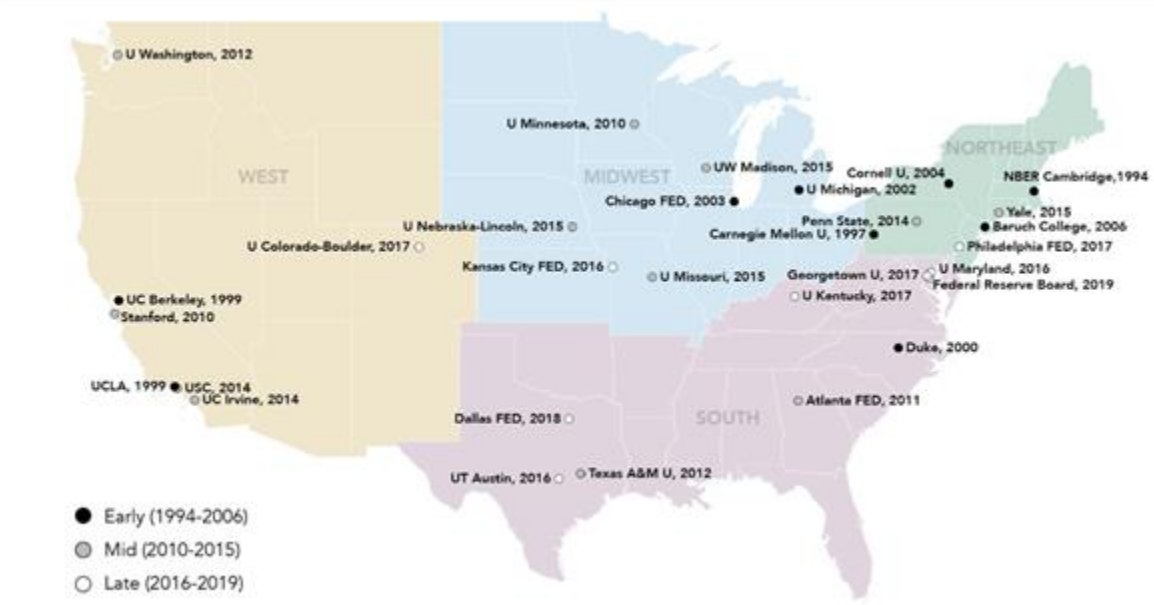


(ii) Stanford University

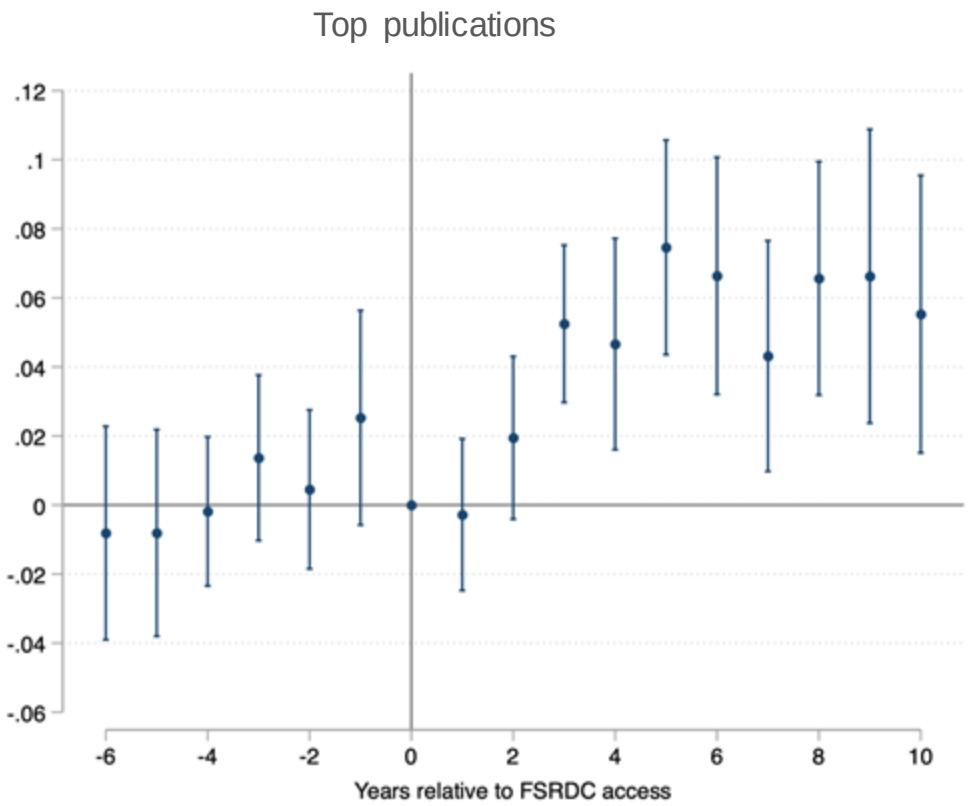


Example 2: Census Data and academic research

Random variation: timing variation across regions based on equity considerations



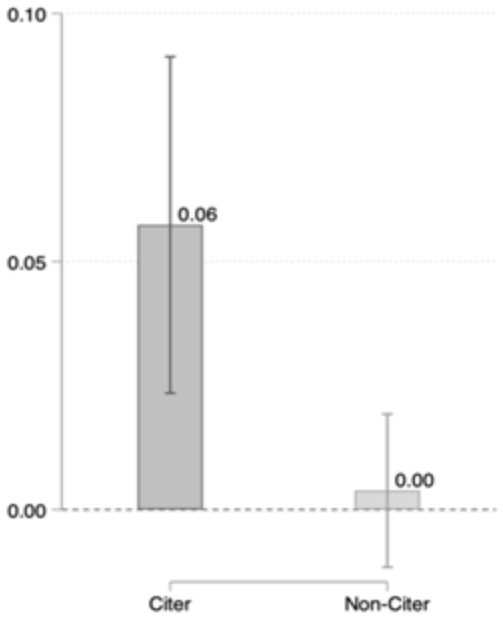
Result: Effect on publication outcomes



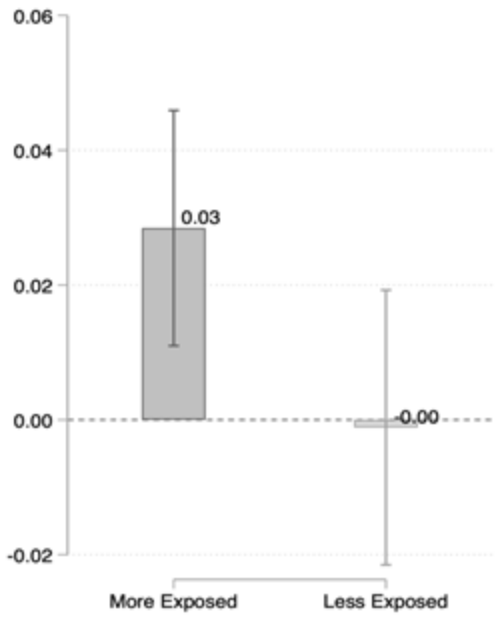
Result: Spillover Effects to Non-Users

Increase in Top Publications

i) By citation status

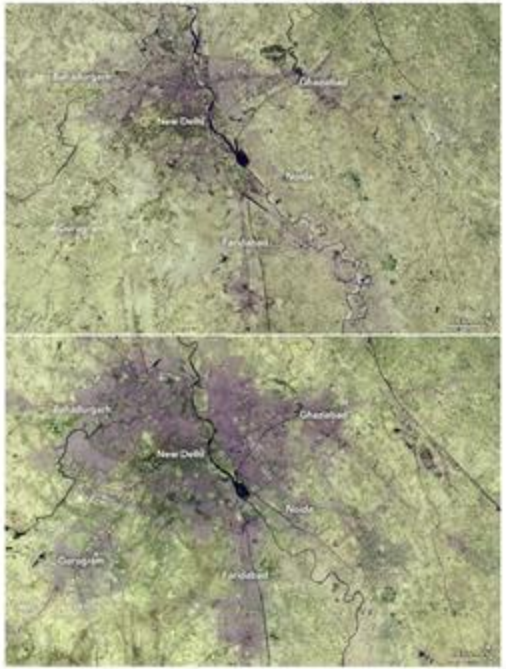


ii) By exposure status

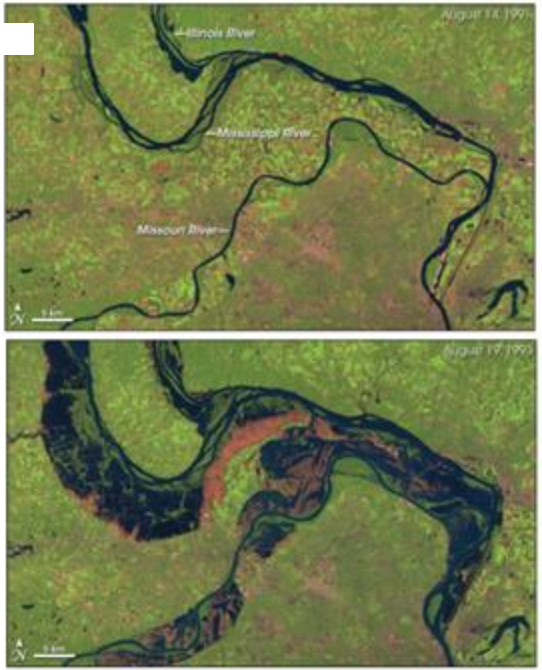


Example 3: Cost of data and environmental research

3. Urbanization in India – New Delhi in 1989 (top) & 2018 (bottom)[a]



4. Flooding of the Mississippi River – St. Louis, Missouri in 1991 (top) & 1993 (bottom)[a]



Example 3: Cost of data and environmental research

Public Law 98-365
98th Congress

An Act

To establish a system to promote the use of land remote-sensing satellite data, and for other purposes.

July 17, 1984
[H.R. 5155]

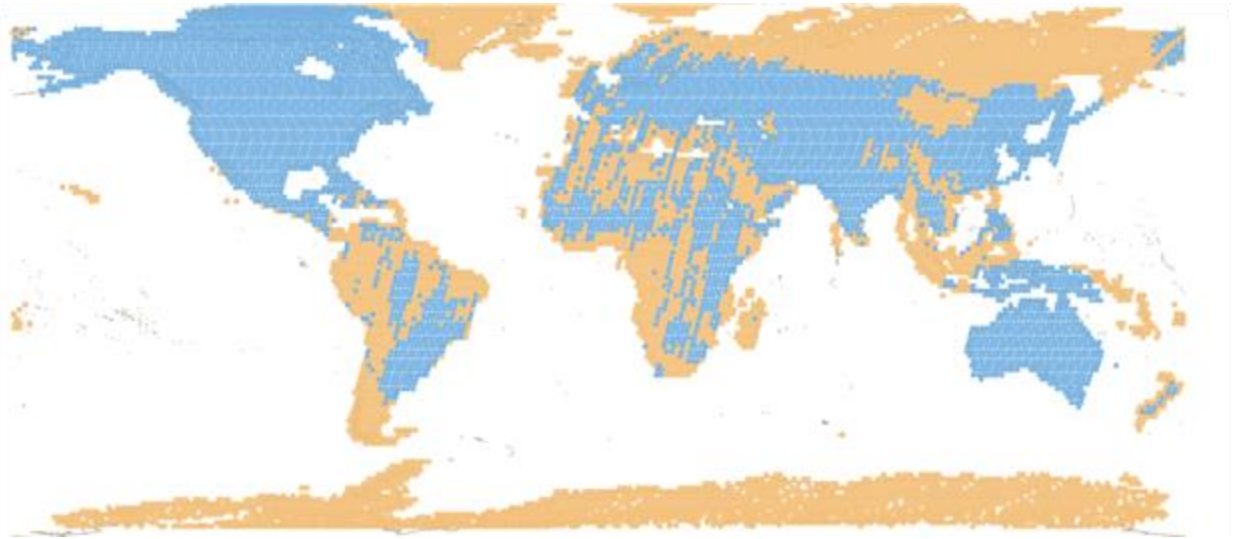
Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That this Act may be cited as the "Land Remote-Sensing Commercialization Act of 1984".

Land Remote-Sensing Commercialization Act of 1984.
Communications and telecommunications.
15 USC 4201 note.

TITLE I—DECLARATION OF FINDINGS, PURPOSES, AND POLICIES

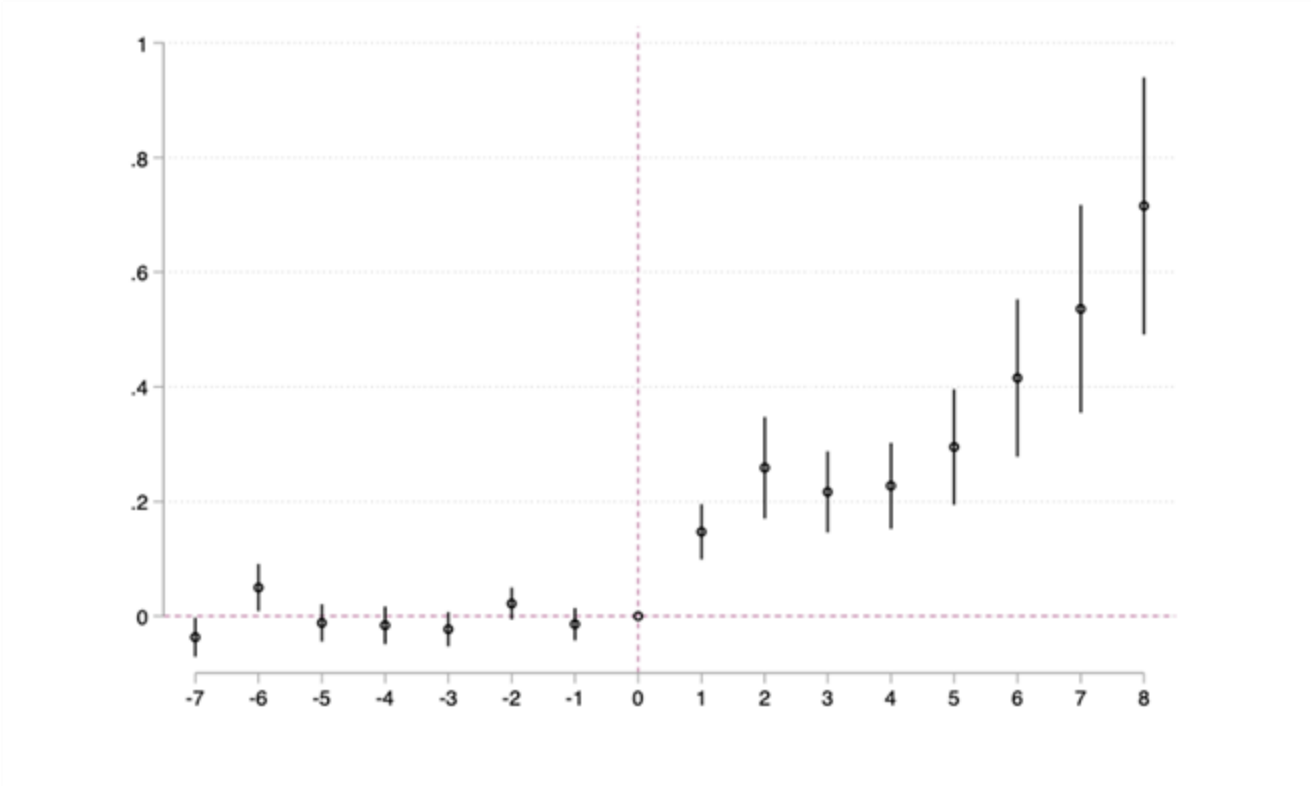
Example 3: Cost of data and environmental research

Random variation: technical errors and cloud cover + changes in access regime

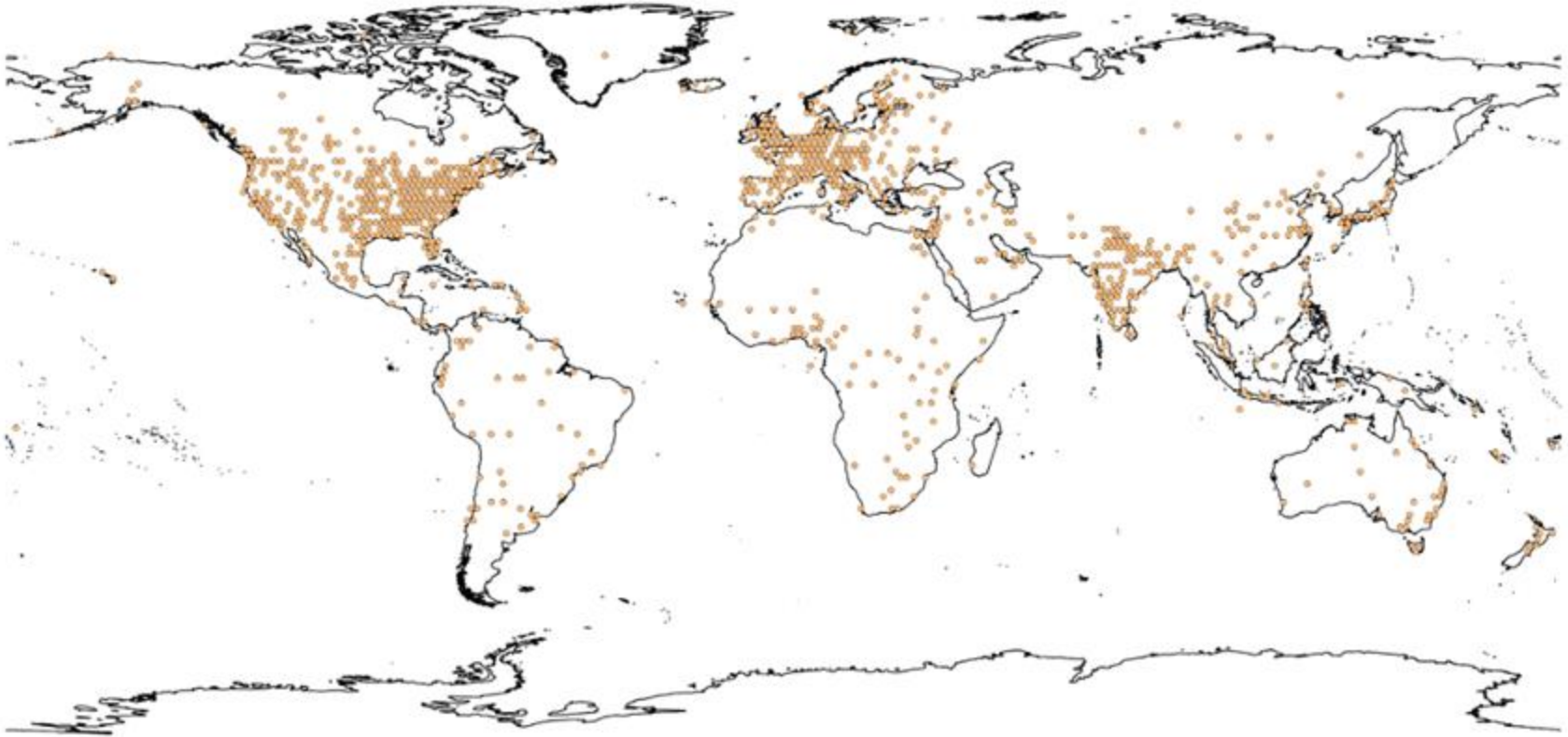


Blocks With Above or Below Median Number (18) of Landsat Images by 1985

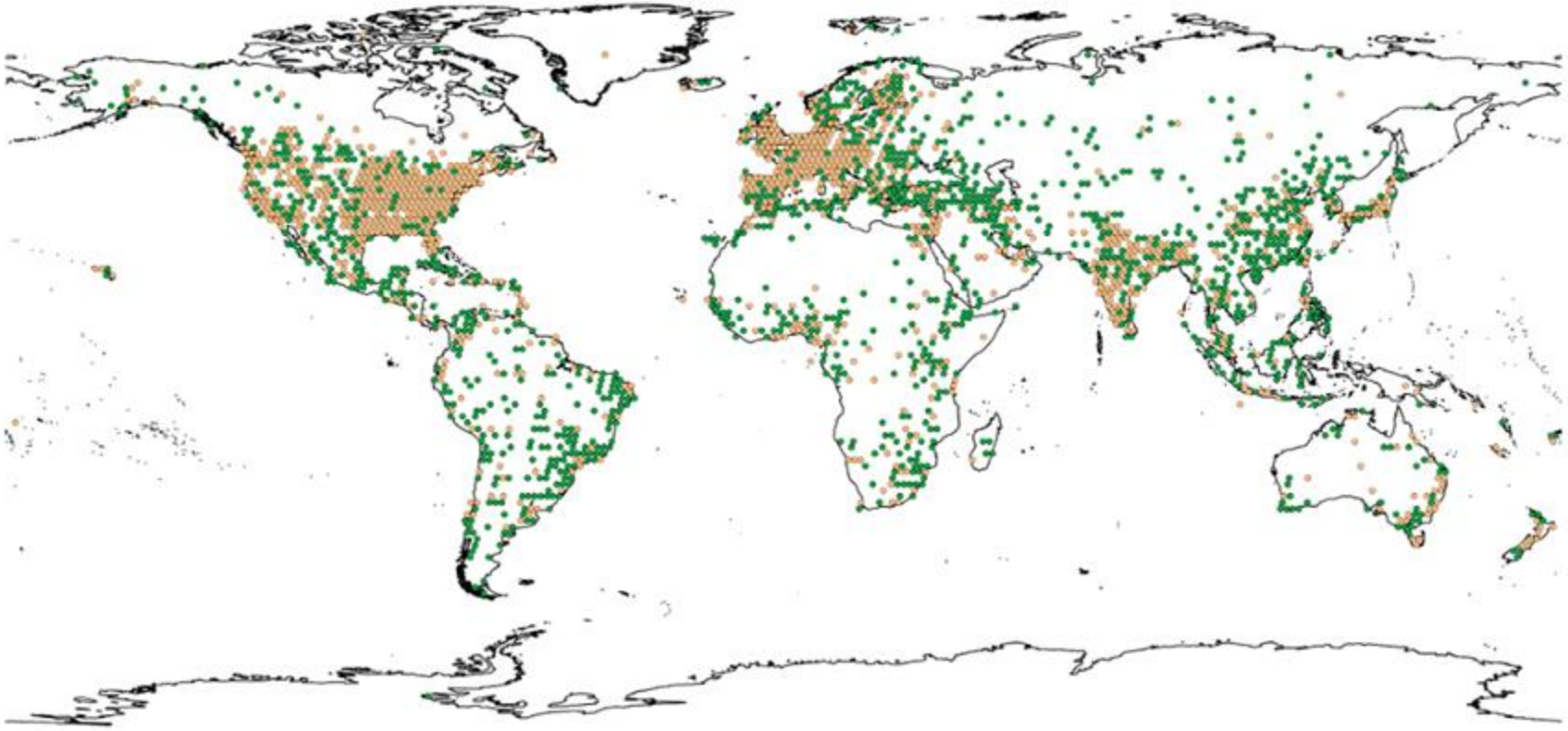
Result: Effect of cost reductions on follow-on publications



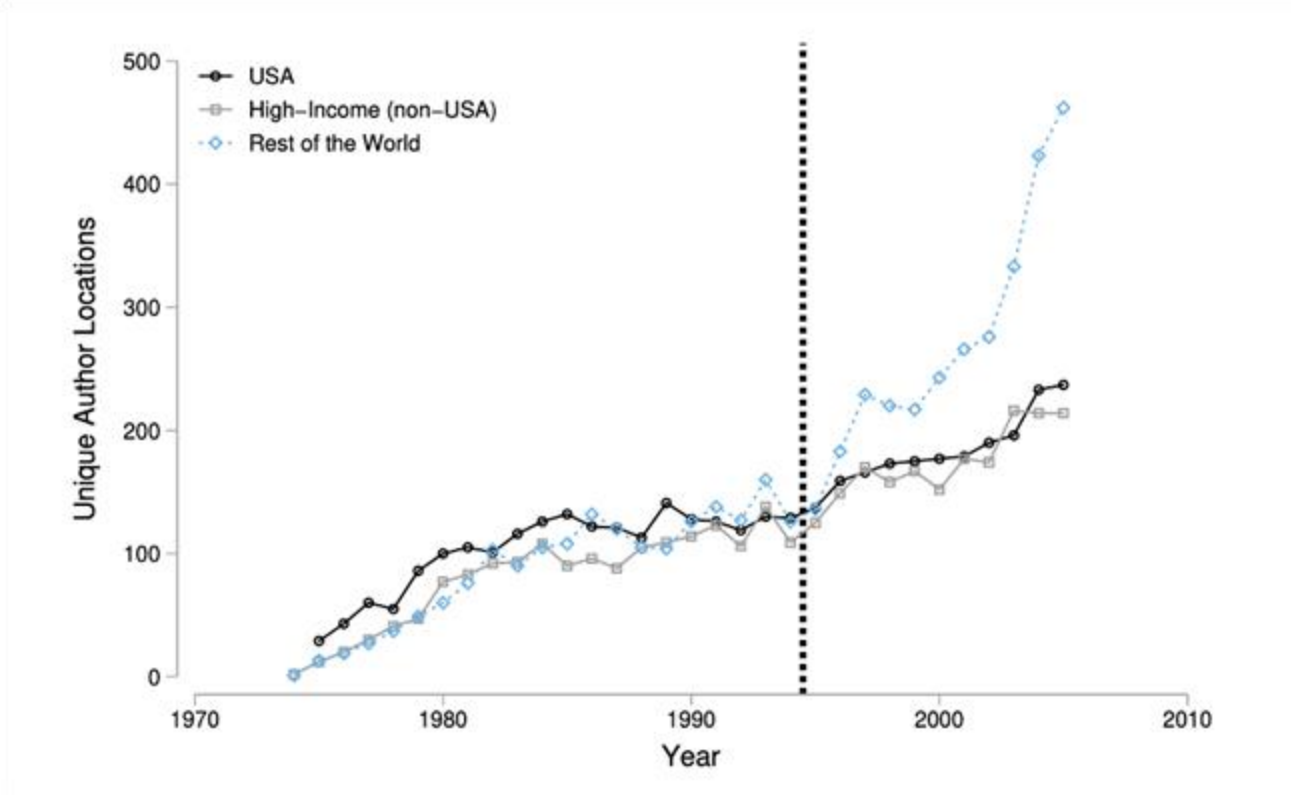
Author locations



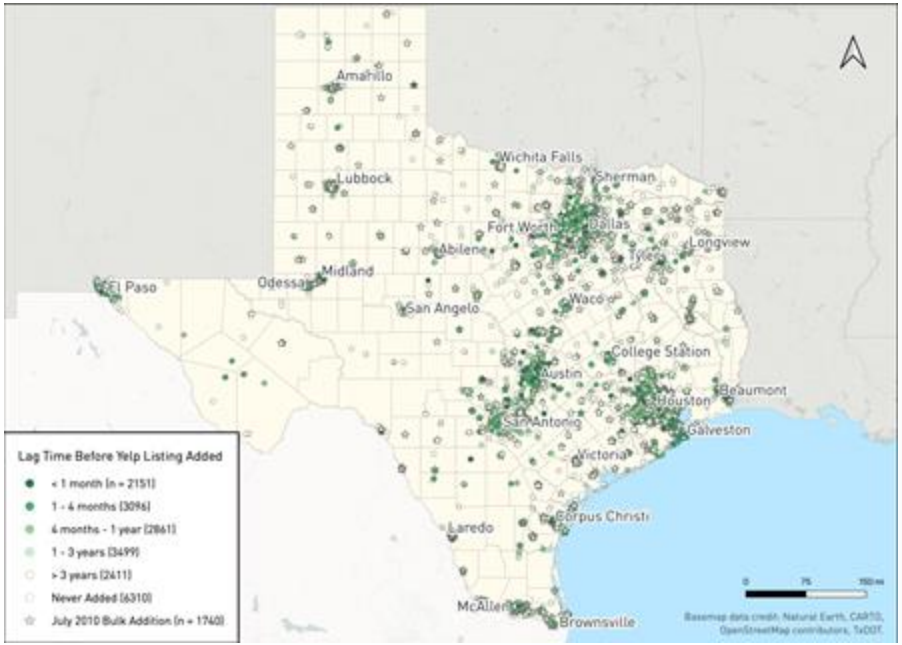
Author locations

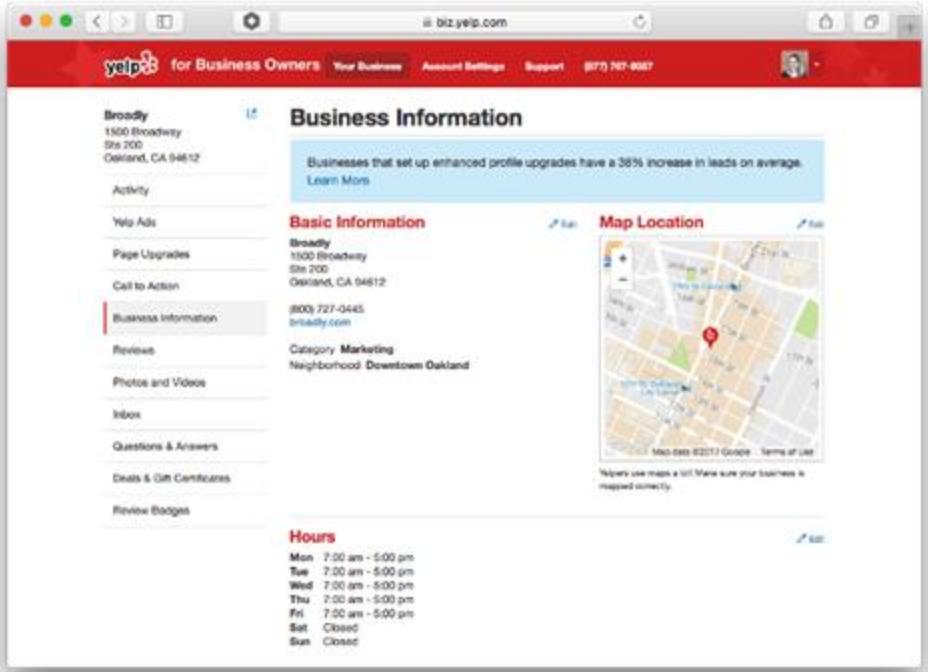


Result: Effect of cost reductions on *author* locations



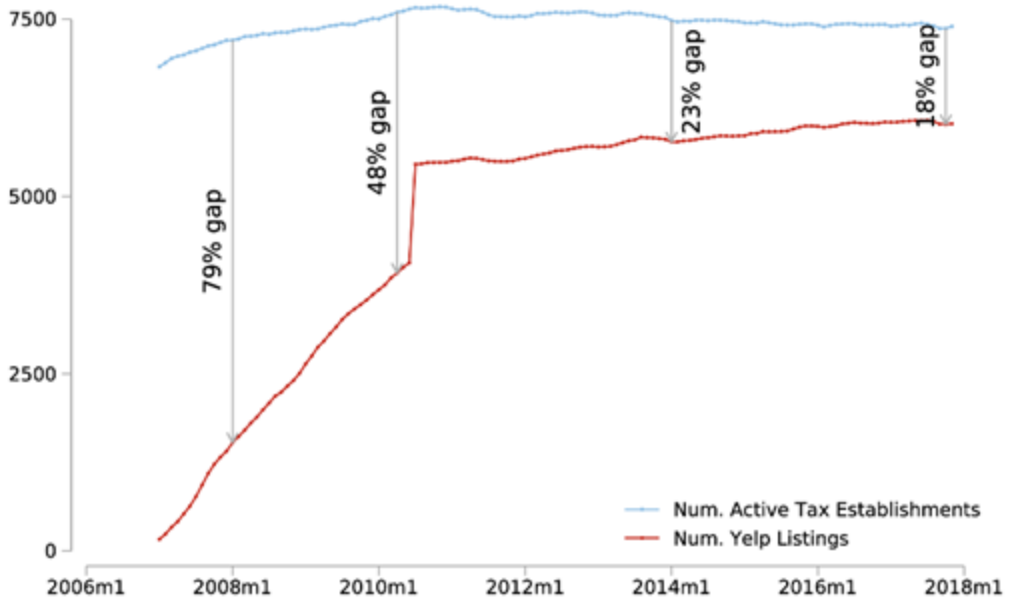
Example 4: Online listing data and business performance





Example 4: Online listing data and business performance

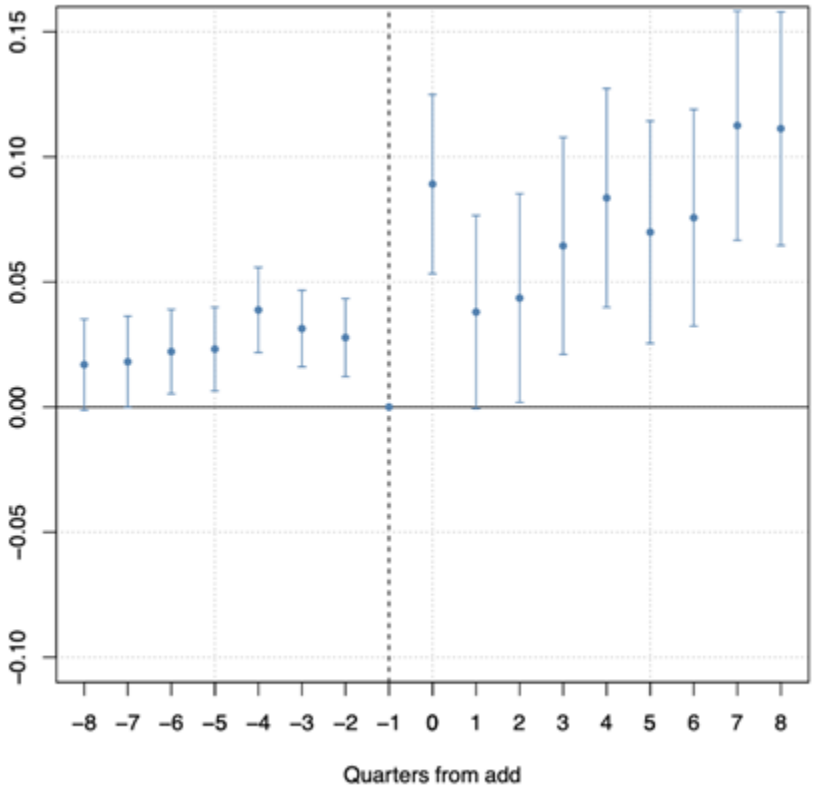
Random variation: Yelp worked with a local business data aggregator to add over a thousand local restaurants and bars in their database in bulk at a single point in time



Result: Effect on quarterly revenue

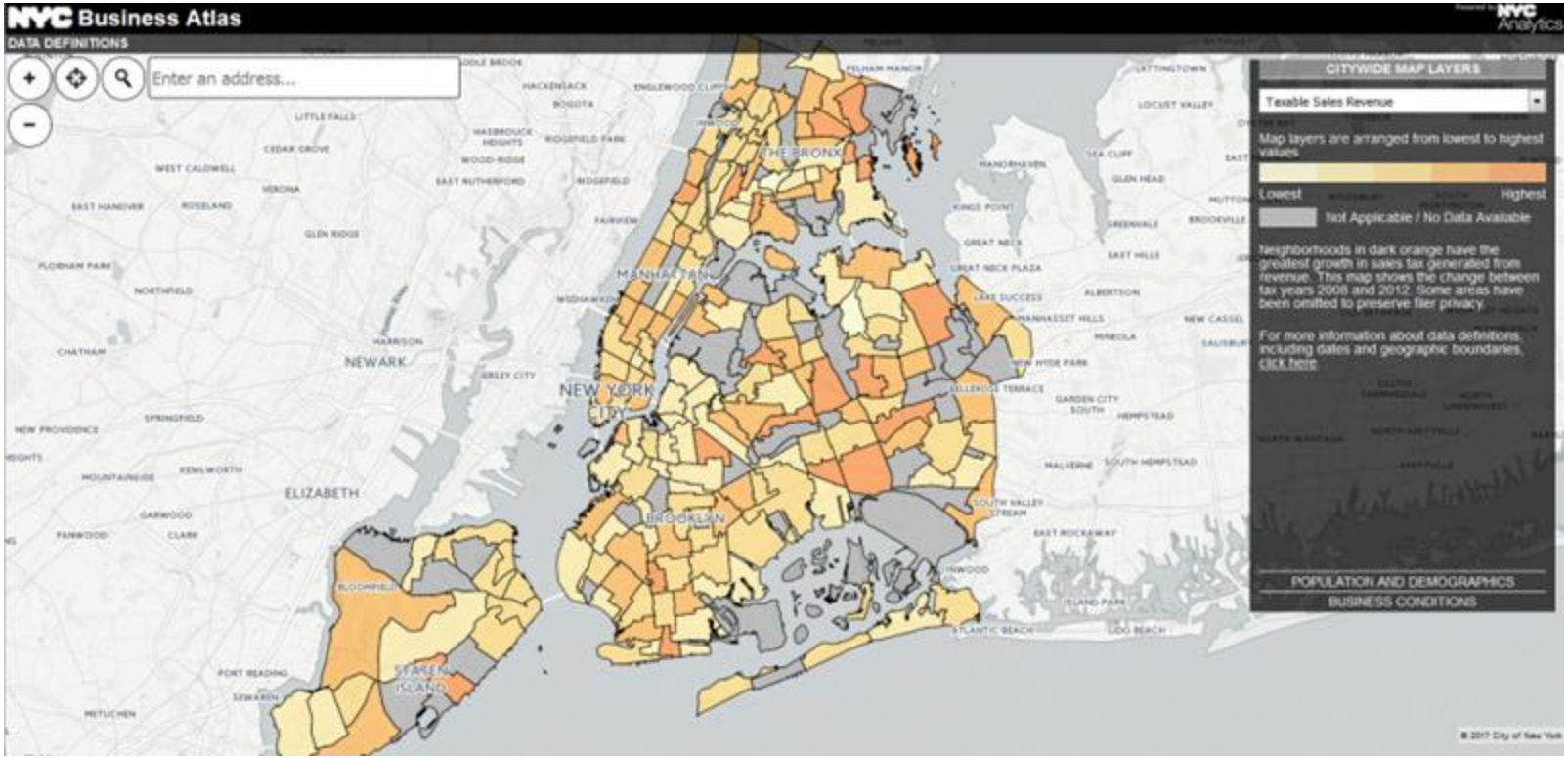
~$1500/Month

~$18M for 1k establishments



Question	Quasi-Experimental Variation	Outcome
Satellite data in the gold exploration industry	Patchy coverage due to technical difficulties and cloud cover during satellite data collection	Satellite data doubles discoveries and encourages the entry of new gold exploration firms
Census administrative data and economics research	Changing access at the university level given the need to access data via physical enclaves	Access to administrative data leads to more top publications by empirical researchers, and even among those who do not use the data directly
Satellite data cost and environmental science	Changing cost in access to data given changing government regulation	Lowering the cost of data access leads to more diversity in the set of regions and scientists that benefit
Restaurant listings and revenue on Yelp	Increased coverage of restaurant listings owing to a data purchase by Yelp	Improved coverage increases quarterly revenue by 5–10%

How would you value NYC Business Atlas?



How to apply this in practice

01

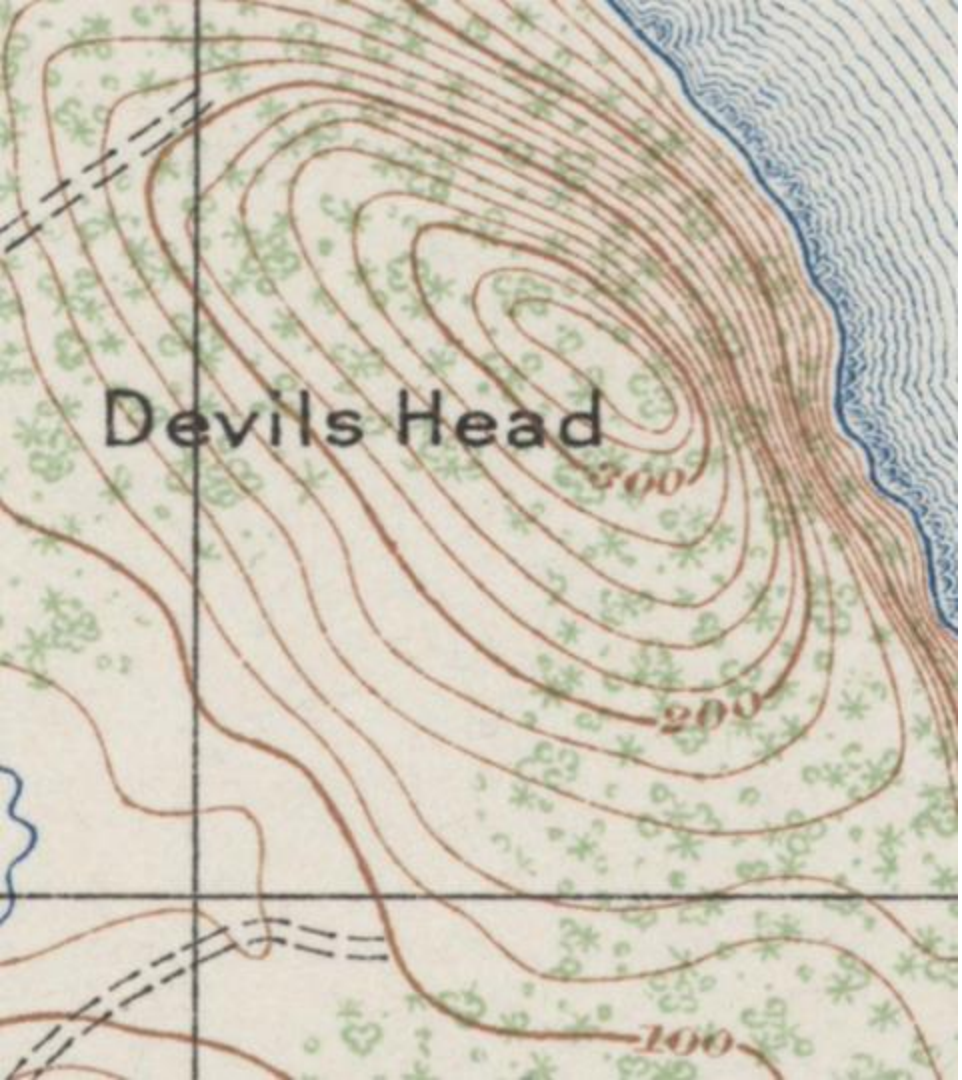
Define the relevant terrain and the relevant data-driven decisions

02

Specify alternate representations of the terrain; these alternate representations could reflect differences in coverage, access, cost, and so on.

03

Map differences in decisions to differences in the data representation regime



Thank you!

Abhishek Nagaraj
UC Berkeley and NBER
nagaraj@berkeley.edu

X: @abhishekn

References

Luca, Michael, Abhishek Nagaraj, and Gauri Subramani. *Getting on the Map: The Impact of Online Listings on Business Performance*. No. w30810. National Bureau of Economic Research, 2023.

Nagaraj, Abhishek. "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry." *Management Science* 68, no. 1 (2022): 564-582.

Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan. "Improving data access democratizes and diversifies science." *Proceedings of the National Academy of Sciences* 117, no. 38 (2020): 23490-23498.

Nagaraj, Abhishek, and Scott Stern. "The economics of maps." *Journal of Economic Perspectives* 34, no. 1 (2020): 196-221.

Nagaraj, Abhishek, and Matteo Tranchero. *How does data access shape science? Evidence from the impact of US census's research data centers on economics research*. No. w31372. National Bureau of Economic Research, 2023.