Private Use of Public Information

Abhishek Nagaraj

UC Berkeley

December 5
National Academies Panel
“Advancing Commercialization from Federal Labs”
Washington DC

Dr John Snow, 1854

The Map of the Broad Street Pump

edureka!

BIG DATA

7 connected devices per person

COMMUNICATION

By 2020 each person will own an avg of 7 connected devices

71% of shoppers are multi channel

RETAIL

Retail companies are using IoT devices to manage their sales & customer acquisition

MEDICAL

Medical data disclosure is the second most breached source of data

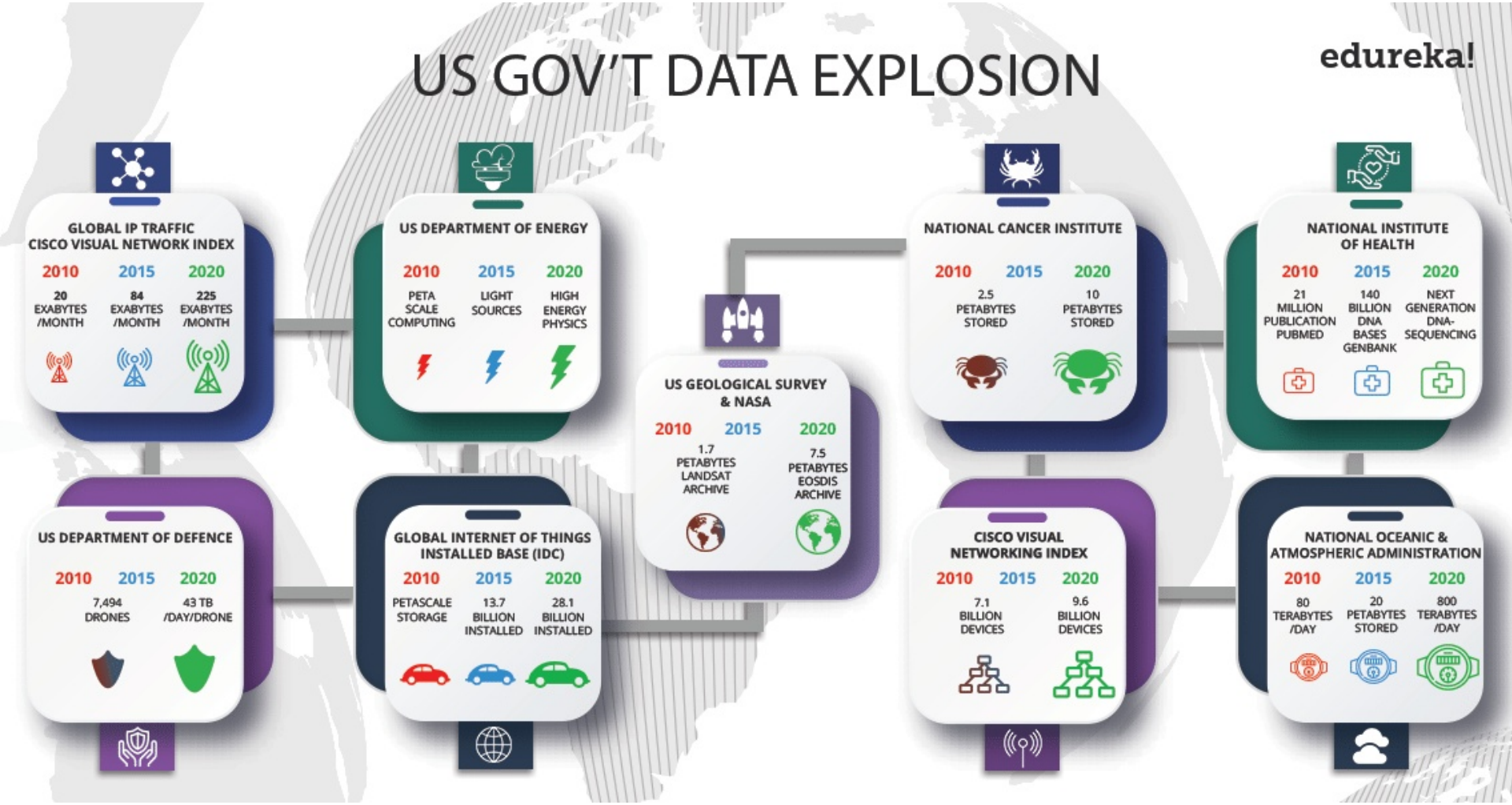
30% annual growth rate

INDUSTRIAL

Project will increase in connected machine - to - machine devices over the next 5 yrs

VEHICLES

23.6 million cars is having internet access by 2016, raising from 8.7 million in 2010


US GOV'T DATA EXPLOSION
edureka!
GLOBAL IP TRAFFIC CISCO VISUAL NETWORK INDEX
2010 20 EXABYTES /MONTH
2015 84 EXABYTES /MONTH
2020 225 EXABYTES /MONTH
US DEPARTMENT OF ENERGY
2010 PETA SCALE COMPUTING
2015 LIGHT SOURCES
2020 HIGH ENERGY PHYSICS
US GEOLOGICAL SURVEY & NASA
2010 1.7 PETABYTES LANDSAT ARCHIVE
2020 7.5 PETABYTES EOSDIS ARCHIVE
NATIONAL CANCER INSTITUTE
2010 2.5 PETABYTES STORED
2020 10 PETABYTES STORED
NATIONAL INSTITUTE OF HEALTH
2010 21 MILLION PUBLICATION PUBMED
2015 140 BILLION DNA BASES GENBANK
2020 NEXT GENERATION DNA-SEQUENCING
US DEPARTMENT OF DEFENCE
2010 7,494 DRONES
2015 43 TB /DAY/DRONE
GLOBAL INTERNET OF THINGS INSTALLED BASE (IDC)
2010 PETASCALE STORAGE
2015 13.7 BILLION INSTALLED
2020 28.1 BILLION INSTALLED
CISCO VISUAL NETWORKING INDEX
2010 7.1 BILLION DEVICES
2020 9.6 BILLION DEVICES
NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION
2010 80 TERABYTES /DAY
2015 20 PETABYTES STORED
2020 800 TERABYTES /DAY

On I.T.

Intel searches for the value in open data

By **Mohana Ravindranath**
May 2, 2014

For decades, Intel has generated the bulk of its revenue by manufacturing processors and other parts for personal computers, remaining comfortably among the top semiconductor vendors.

Today, it is exploring business opportunities in a new, less tangible area — the free exchange of information between the federal government and the public, often called "open data."

It is an ongoing research project. Last year, Intel sponsored the National Civic Day of Hacking, during which groups of entrepreneurs and developers across the country were asked to invent ways to use government data — making public transportation information easier to digest, for instance. And in

On I.T.

Intel searches for the value in open data

By **Mohana Ravindranath**
May 2, 2014

For decades, Intel has generated the bulk of its revenue by manufacturing processors and other parts for personal computers, remaining comfortably among the top semiconductor vendors.

Today, it is exploring business opportunities in a new, less tangible area — the free exchange of information between the federal government and the public, often called "open data."

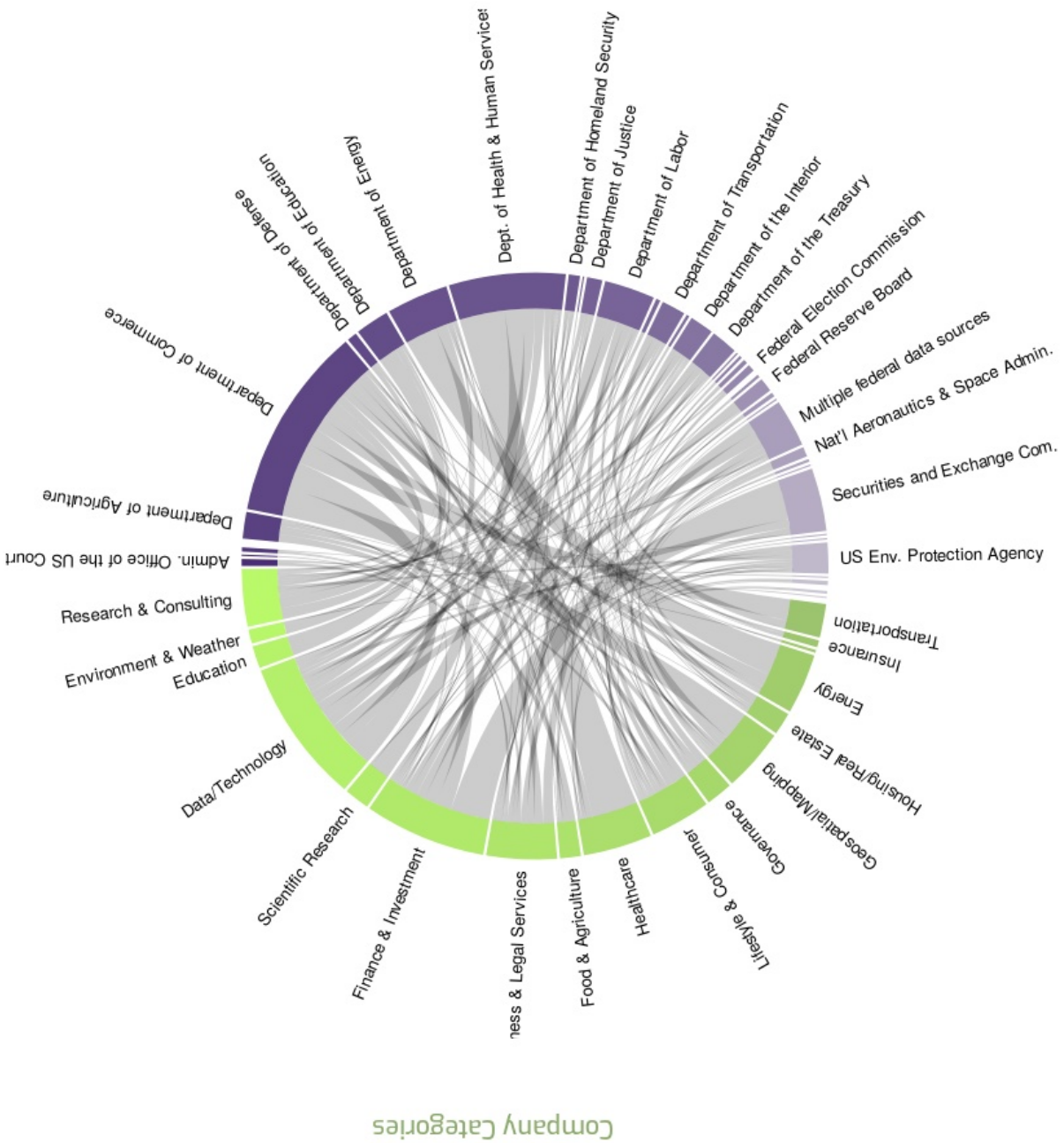
It is an ongoing research project. Last year, Intel sponsored the National Civic Day of Hacking, during which groups of entre
government data — m

Monsanto Buys Climate Corp For $930 Million

Bruce Upbin Former Contributor ⓘ

Tech

I manage our technology coverage.


Federal Agencies
Dept. of Health & Human Service
Department of Homeland Security
Department of Justice
Department of Labor
Department of Transportation
Department of the Interior
Department of the Treasury
Federal Election Commission
Federal Reserve Board
Multiple federal data sources
Nat'l Aeronautics & Space Admin.
Securities and Exchange Com.
US Env. Protection Agency
Transportation
Insurance
Energy
Housing/Real Estate
Geospatial/Mapping
Government
Lifestyle & Consumer
Healthcare
Food & Agriculture
ness & Legal Services
Finance & Investment
Scientific Research
Data/Technology
Environment & Weather
Education
Research & Consulting
Admin. Office of the US Court
Department of Agriculture
Department of Commerce
Department of Defense
Department of Education
Department of Energy
Company Categories

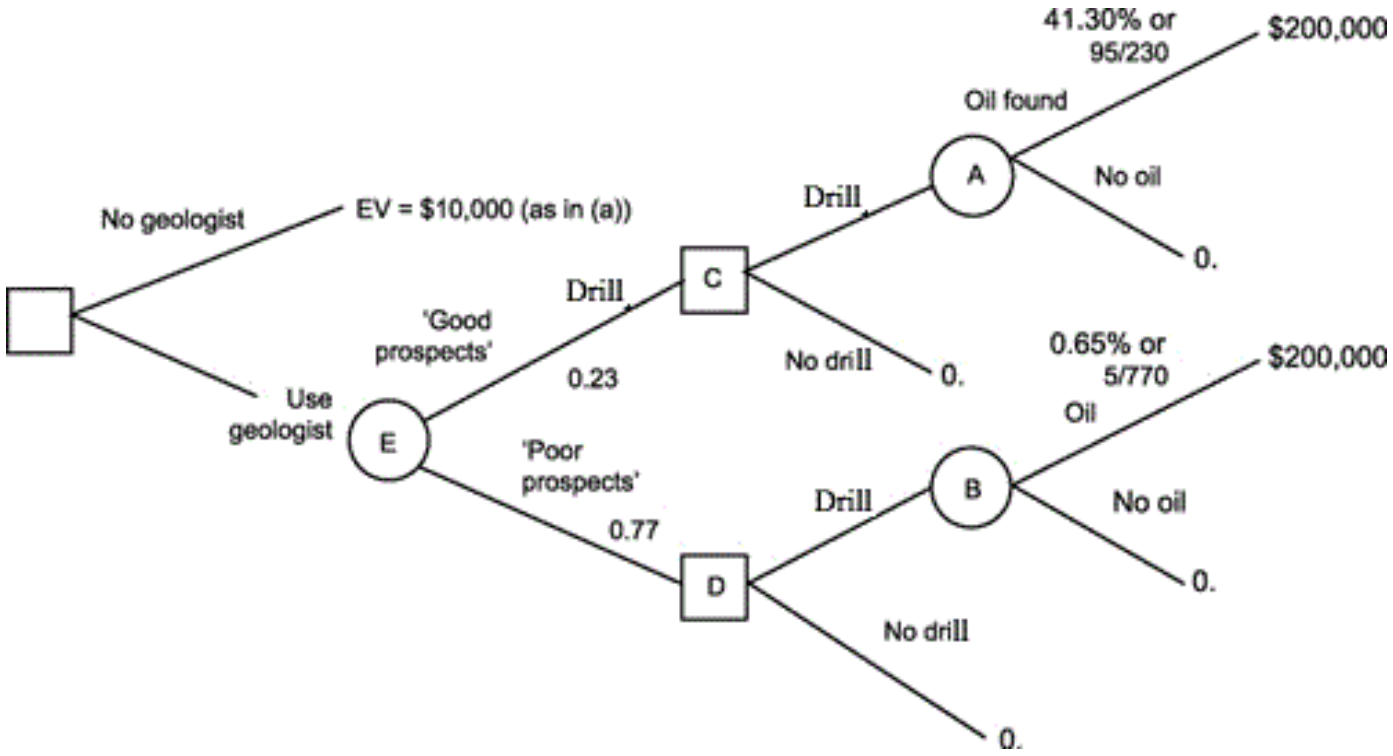
Academic Work on the Private Value of Government Investment

- Financial
  (SBIR Grants, Subsidies, R&D Tax Credits etc)
- Physical Infrastructure
  (Roads, Communication Networks, Airports etc)
- Technology and Science
  (University research, Public R&D, Patents etc)

Open Question:

How to Assess the Private Value of Public Data?

Theoretical Framework

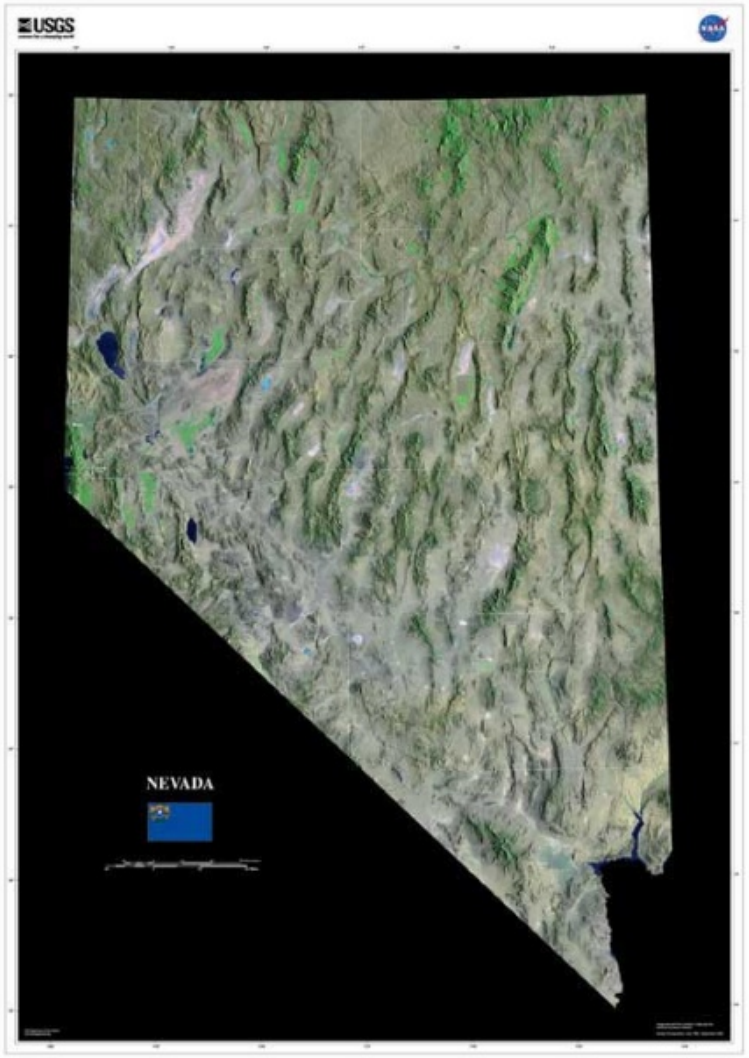


Today's Talk: Discuss 2 recent papers studying value of NASA satellite data for private-sector innovation

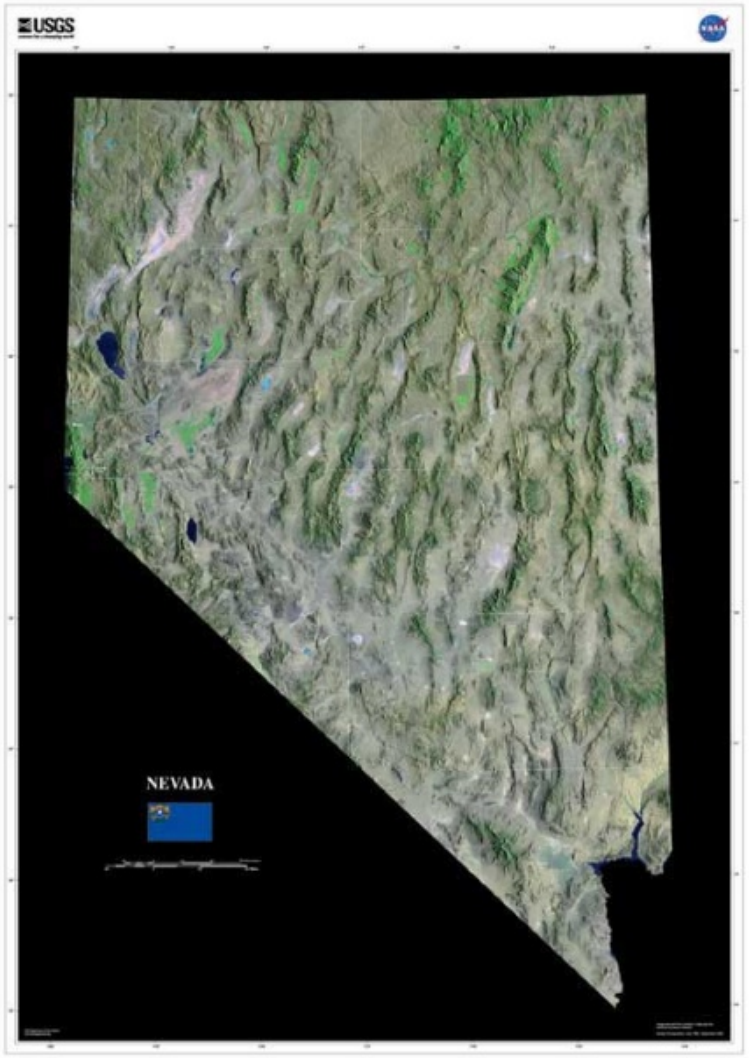
- Empirical framework for assessing the causal impact of public data
- Estimates along few different margins: firm performance, firm entry & scientific innovation

The Private Impact of Public Data:
Landsat Satellite Maps and Gold Exploration

Landsat and Gold Exploration



Landsat and Gold Exploration



(from: Rowan & Wetlaufer 1975)

Landsat and Gold Exploration

Ideal Experiment$\rightarrow$
Randomly assign mapping information to some regions and not others and collect discovery data

The Landsat “Experiment”

Historical Record of Landsat Global Coverage: Mission Operations, NSLRSDA, and International Cooperator Stations

Samuel Goward, Terry Arvidson, Darrel Williams, John Faundeen, James Irons, and Shannon Franks

The Landsat "Experiment"

Historical Record of Landsat Global Coverage: Mission Operations, NSLRSDA, and International Cooperator Stations

Samuel Goward, Terry Arvidson, Darrel Williams, John Faundeen, James Irons, and Shannon Franks

the changing global environment" (U.S. Congress, 1992). The advisory committee for NSLRSDA requested a detailed analysis of observation coverage within the U.S. Landsat holdings, as well as that acquired and held by International Cooperator (IC) stations. Our analyses, to date, have found gaps of varying magnitude in U.S. holdings of Landsat global coverage data, which appear to reflect technical or administrative variations in mission operations. In many cases it may be possible to partially fill these gaps in U.S. holdings through observations that were acquired and are

Landsat and Gold Exploration

Ideal Experiment→
Randomly assign mapping information to some regions and not others and collect discovery data

This project: Exploit variation in *timing* of mapping and compare discoveries in difference-in-difference framework with block and year fixed effects

Landsat and Gold Exploration

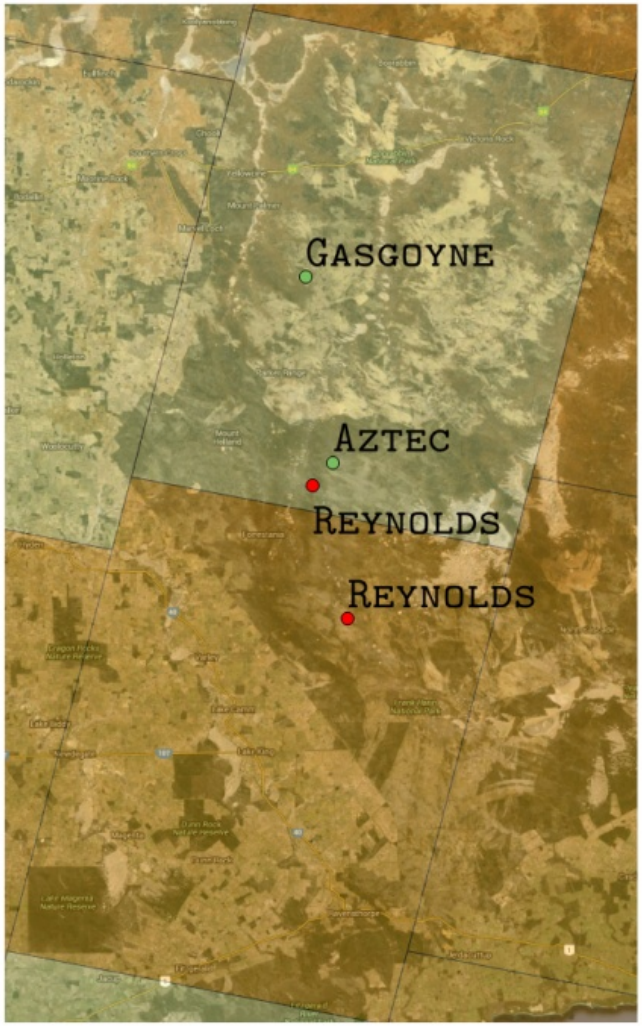
Ideal Experiment→
Randomly assign mapping information to some regions and not others and collect discovery data

This project: Exploit variation in *timing* of mapping and compare discoveries in difference-in-difference framework with block and year fixed effects

Landsat and Gold Exploration

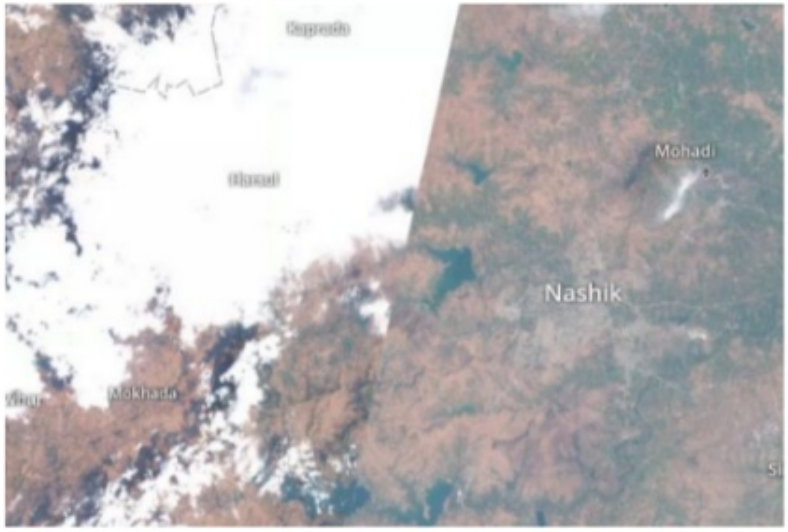
Ideal Experiment→
Randomly assign mapping information to some regions and not others and collect discovery data

This project: Exploit variation in *timing* of mapping and compare discoveries in difference-in-difference framework with block and year fixed effects



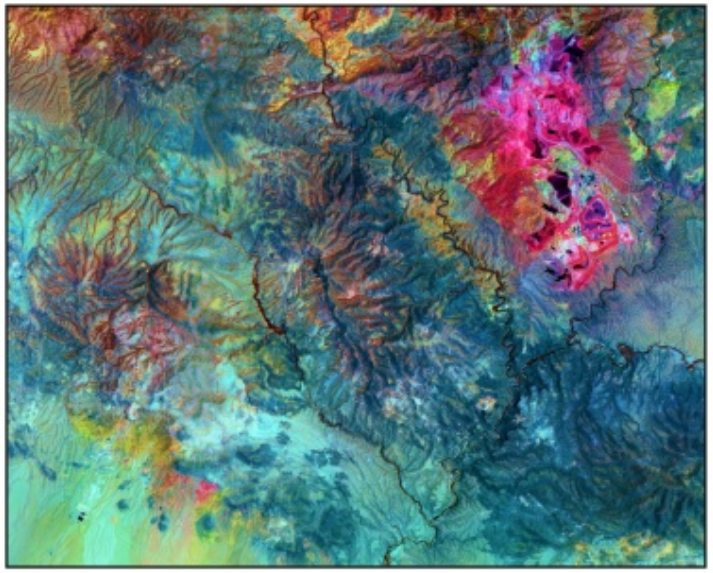
Causes Behind Variation

- **1. Technical Failures**→ Landsat program aimed for global coverage, but literature documents many coverage gaps linked to technical failures (Goward et. al, 2006)
- **2. Cloud-Cover in Imagery**→ Maps containing over 30% of cloud-cover practically unusuable for analysis

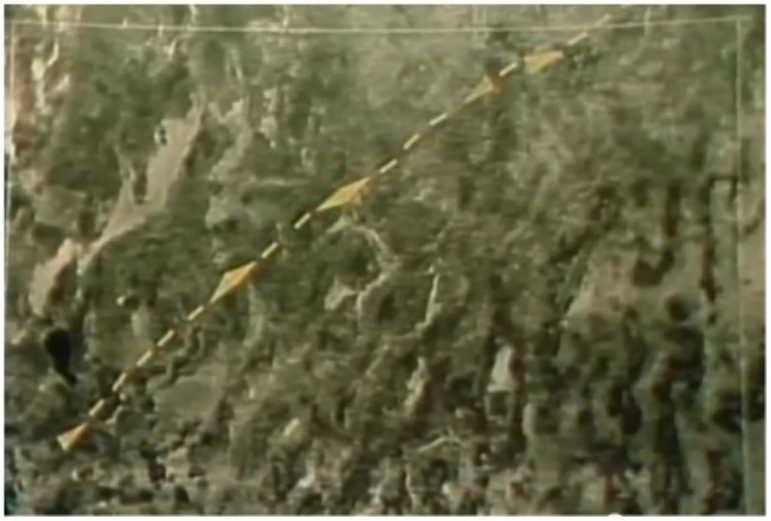


Landsat and Gold Exploration

- Economically important industry ($2.5 billion in exploration costs in 2014) with about 300 unique entities in my data
- Traditional Techniques: Existing data + aerial techniques + on-the-ground exploration
- Firms classified as juniors or seniors

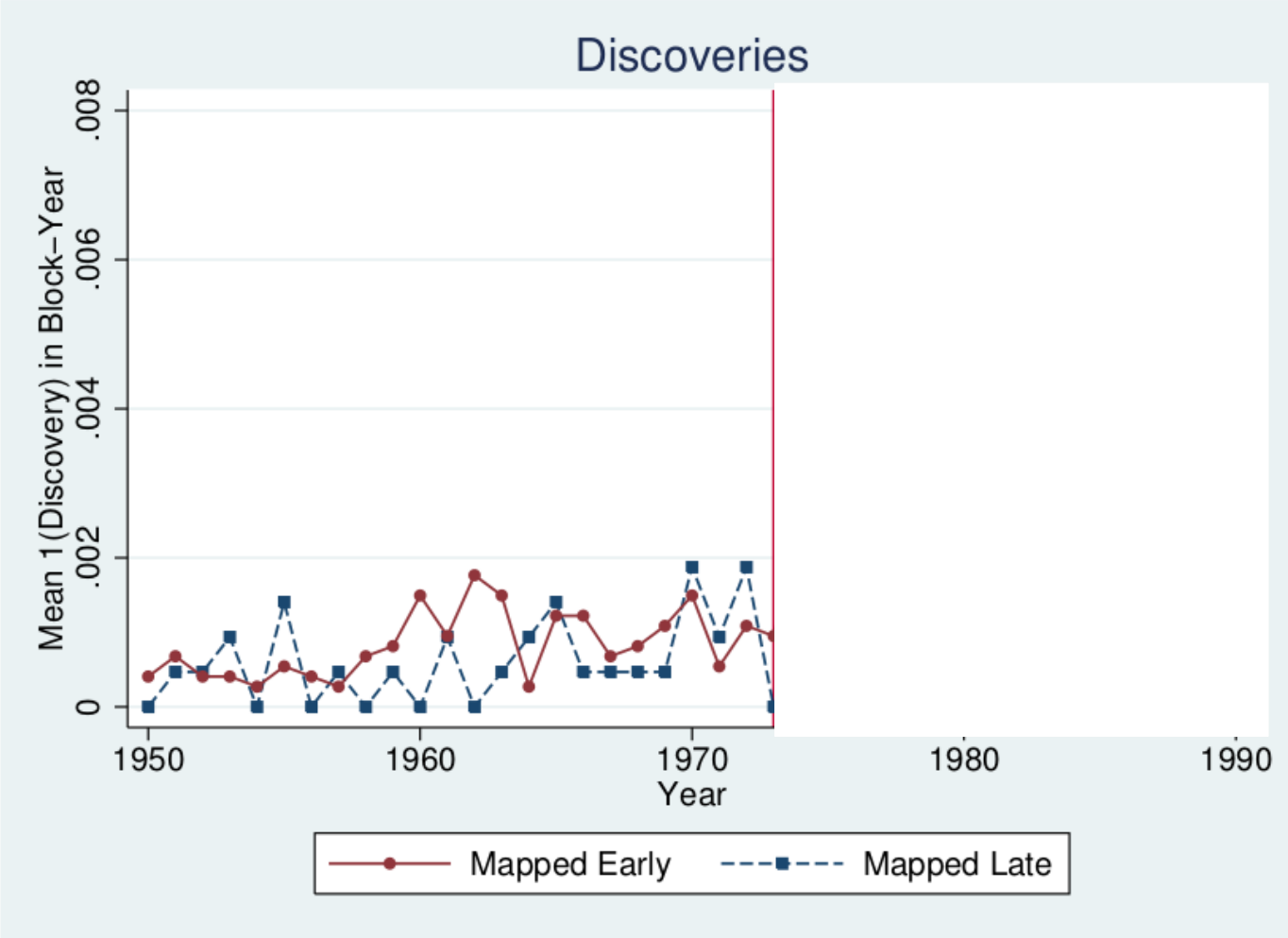


Landsat and Gold Exploration

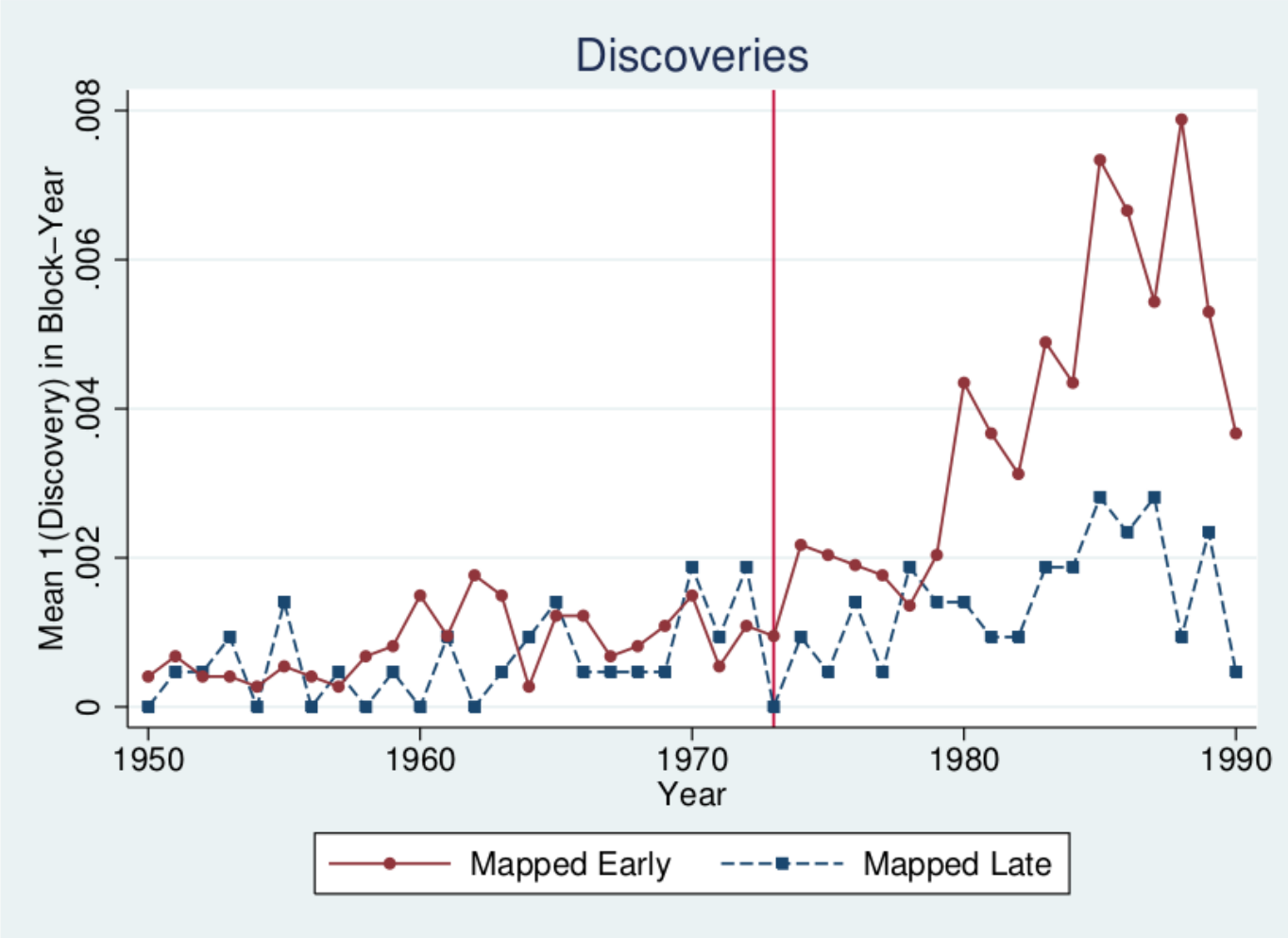


1. Identify "lineaments" or faults in the earth's surface
2. Detect specific minerals like iron, that are often markers of gold and other valuable minerals

Landsat and Gold Exploration



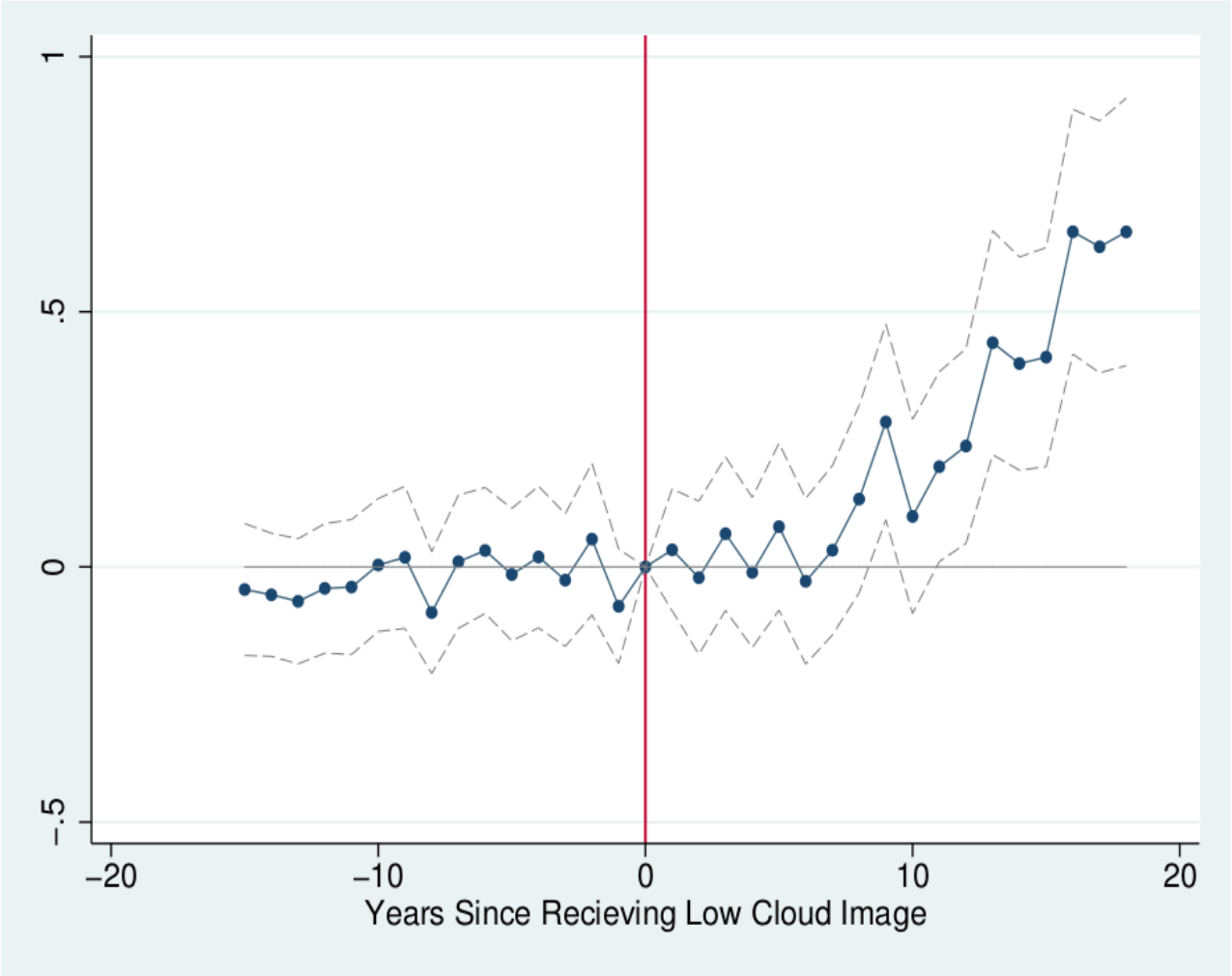
Landsat and Gold Exploration



Landsat and Gold Exploration

	Any Disc.	Any Disc.	Any Disc.	Any Disc.
Post Mapped	0.251***		0.152***	
	(0.0265)		(0.0294)	
Post Low-Cloud		0.267***		0.164***
		(0.0276)		(0.0274)
Block FE	No	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	389213	389213	389213	389213

Landsat and Gold Exploration



Junior vs. Senior Firms

- Pre-Landsat average discoveries:
  - $\overline{1(Junior)}$: 0.008
  - $\overline{1(Senior)}$: 0.069
- Baseline estimate: $\approx 0.16 = 0.016$ (junior) $+ 0.144$ (senior)

	1(Junior)	1(Junior)	1(Senior)	1(Senior)
Post Mapped	0.0288***		0.127***	
	(0.00563)		(0.0285)	
Post Low-Cloud		0.0472***		0.121***
		(0.00651)		(0.0260)
Percent Gain	**355.68%**	**583%**	**182.39%**	**174.95%**
Block FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	389213	389213	389213	389213

Landsat and Gold Exploration

- Landsat imagery had a causal role in doubling the rate of gold discovery in regions around the world
- Junior firms increased market share from about 8% to 25% in regions with freely available satellite imagery
- Back of the envelope calculations suggest additional $10B worth of gold reserves attributed to Landsat imagery in the US alone over a 15 year period
- First studies to illustrate the long-term and industry-wide implications of remotely sensed satellite data

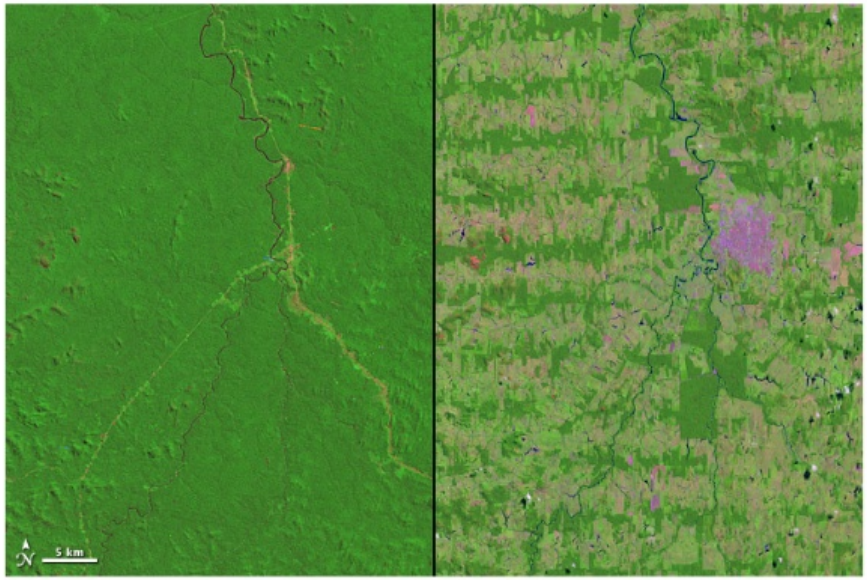
Does Data Access Democratize Science?

Abhishek Nagaraj
Esther Shears
Mathijs de Vaan

UC Berkeley

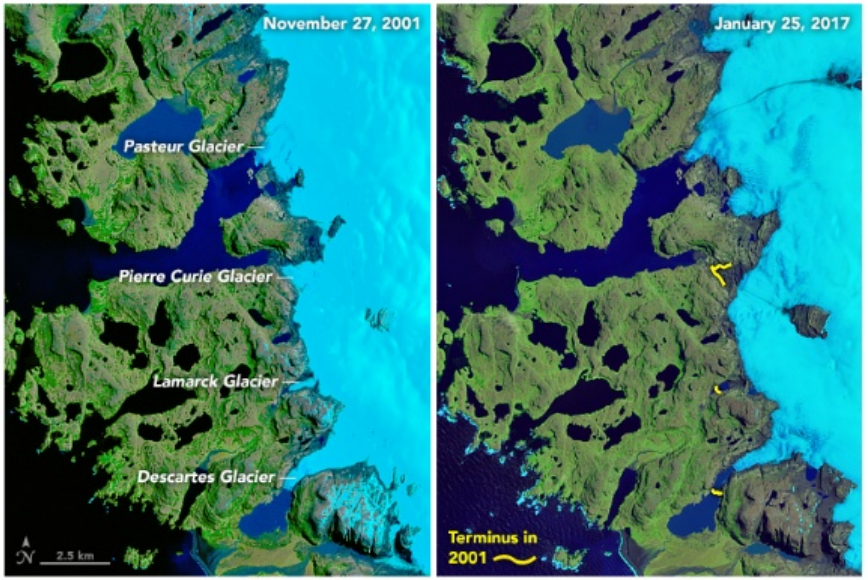
Landsat Applications

1. Deforestation in the Amazon Rainforest – Rondonia, Brazil in 1975 (left) & 2012 (right)[a]



[a]https://landsat.visibleearth.nasa.gov/view.php?id=78596

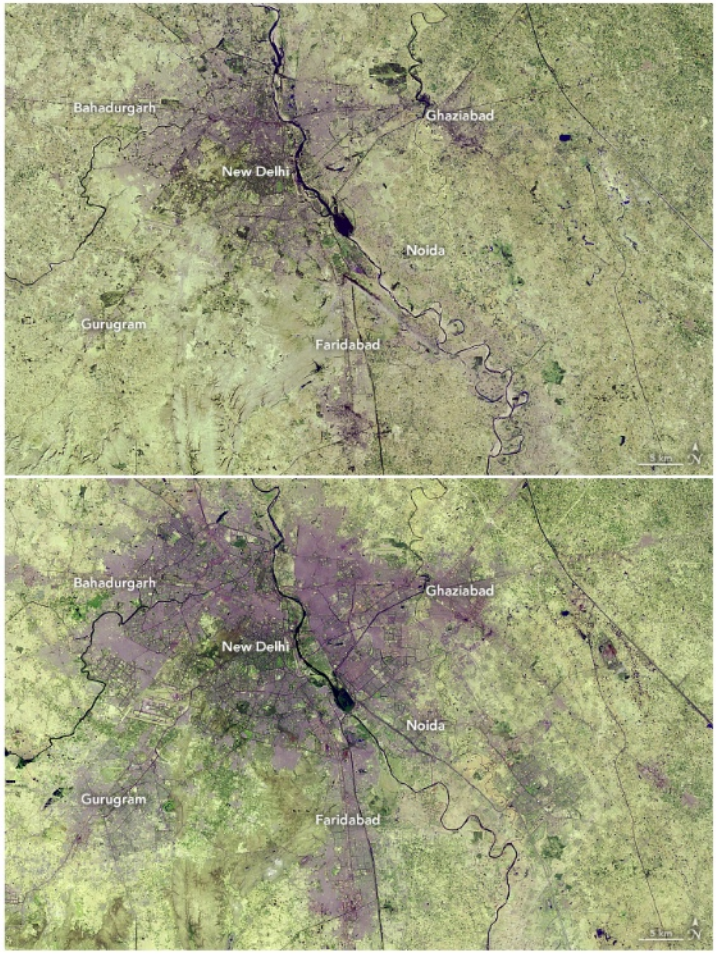
2. Glacier Melt in the Indian Ocean – Kerguelen Is. in 2001 (left) & 2017 (right)[a]



[a]https://landsat.visibleearth.nasa.gov/view.php?id=92059

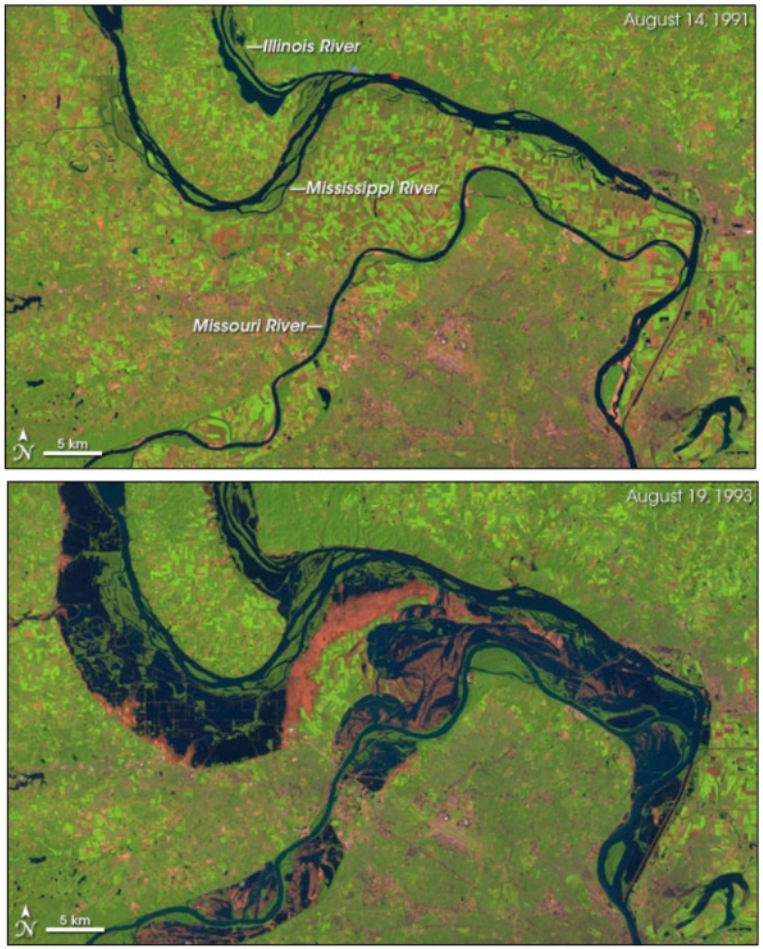
Landsat Applications

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



[a]https://landsat.visibleearth.nasa.gov/view.php?id=92813

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



[a]https://landsat.visibleearth.nasa.gov/view.php?id=5422

Our Focus:

How did the reductions in cost and data sharing restrictions after the Land Remote-Sensing Policy Act (effective 1995) affect ***scientific efforts*** that built on Landsat data?

Two Main Sources of Data

Publication Data:

- Build a dataset of about 24,000 publications by 34,000 authors from 1974 to 2005 that refer to Landsat in the title/abstract
- For each paper, we use ML algorithms that “geoparse” the text and match it to a specific block on earth
- Repeat the same process for author institution locations

Scopus Matching

REPORTS

A Mini-Surge on the Ryder Glacier, Greenland, Observed by Satellite Radar Interferometry

Ian Joughin, Slawek Tulaczyk, Mark Fahnestock, Ron Kwok
+ See all authors and affiliations

Science 11 Oct 1996:
Vol. 274, Issue 5285, pp. 228-230
DOI: 10.1126/science.274.5285.228

Abstract

Satellite radar interferometry reveals that the speed of the Ryder Glacier increased roughly threefold and then returned to normal (100 to 500 meters/year) over a 7-week period near the end of the 1995 melt season. The accelerated flow represents a substantial, though short-lived, change in ice discharge. During the period of rapid motion, meltwater-filled supraglacial lakes may have drained, which could have increased basal water pressure and caused the mini-surge. There are too few velocity measurements on other large outlet glaciers to determine whether this type of event is a widespread phenomenon in Greenland, but because

Scopus Matching

match to Landsat block for this place

REPORTS

A Mini-Surge on the Ryder Glacier, Greenland, Observed by Satellite Radar Interferometry

Ian Joughin, Slawek Tulaczyk, Mark Fahnestock, Ron Kwok
+ See all authors and affiliations

Science 11 Oct 1996:
Vol. 274, Issue 5285, pp. 228-230
DOI: 10.1126/science.274.5285.228

Abstract

Satellite radar interferometry reveals that the speed of the Ryder Glacier increased roughly threefold and then returned to normal (100 to 500 meters/year) over a 7-week period near the end of the 1995 melt season. The accelerated flow represents a substantial, though short-lived, change in ice discharge. During the period of rapid motion, meltwater-filled supraglacial lakes may have drained, which could have increased basal water pressure and caused the mini-surge. There are too few velocity measurements on other large outlet glaciers to determine whether this type of event is a widespread phenomenon in Greenland, but because

Two Main Sources of Data

Publication Data:

- Build a dataset of about 24,000 publications by 34,000 authors from 1974 to 2005 that refer to Landsat in the title/abstract
- For each paper, we use ML algorithms that “geoparse” the text and match it to a specific block on earth
- Repeat the same process for author institution locations

- **Landsat (meta) Data:**
  - For 12,577 scenes that cover earth’s landmasses we obtain Landsat coverage data
  - This includes image date, image quality (e.g. cloud cover) that documents large variation in coverage

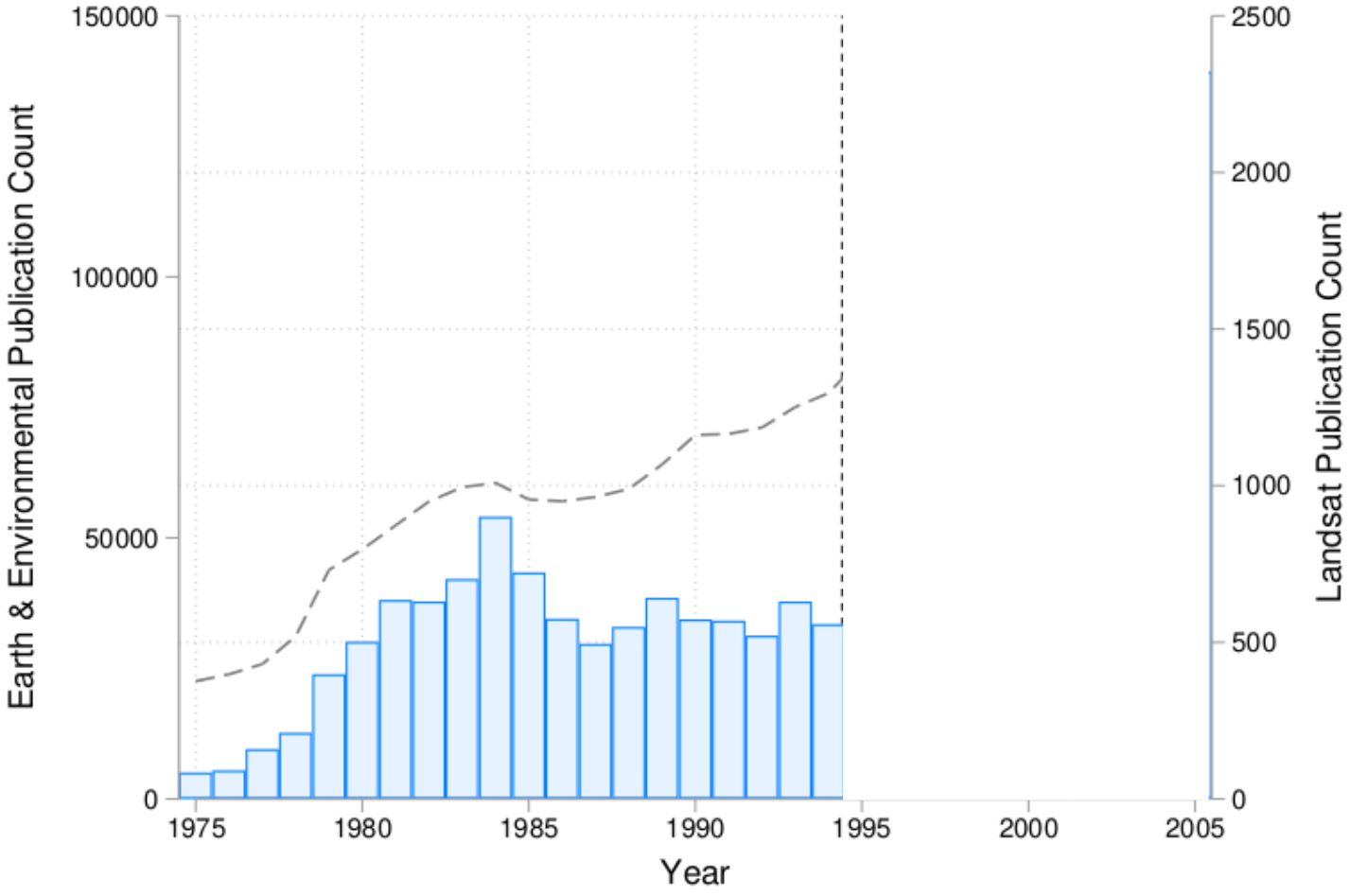
Research Design

- **Descriptive Statistics:**
  Explore how publication output changes in response to reduction in data costs and sharing restrictions

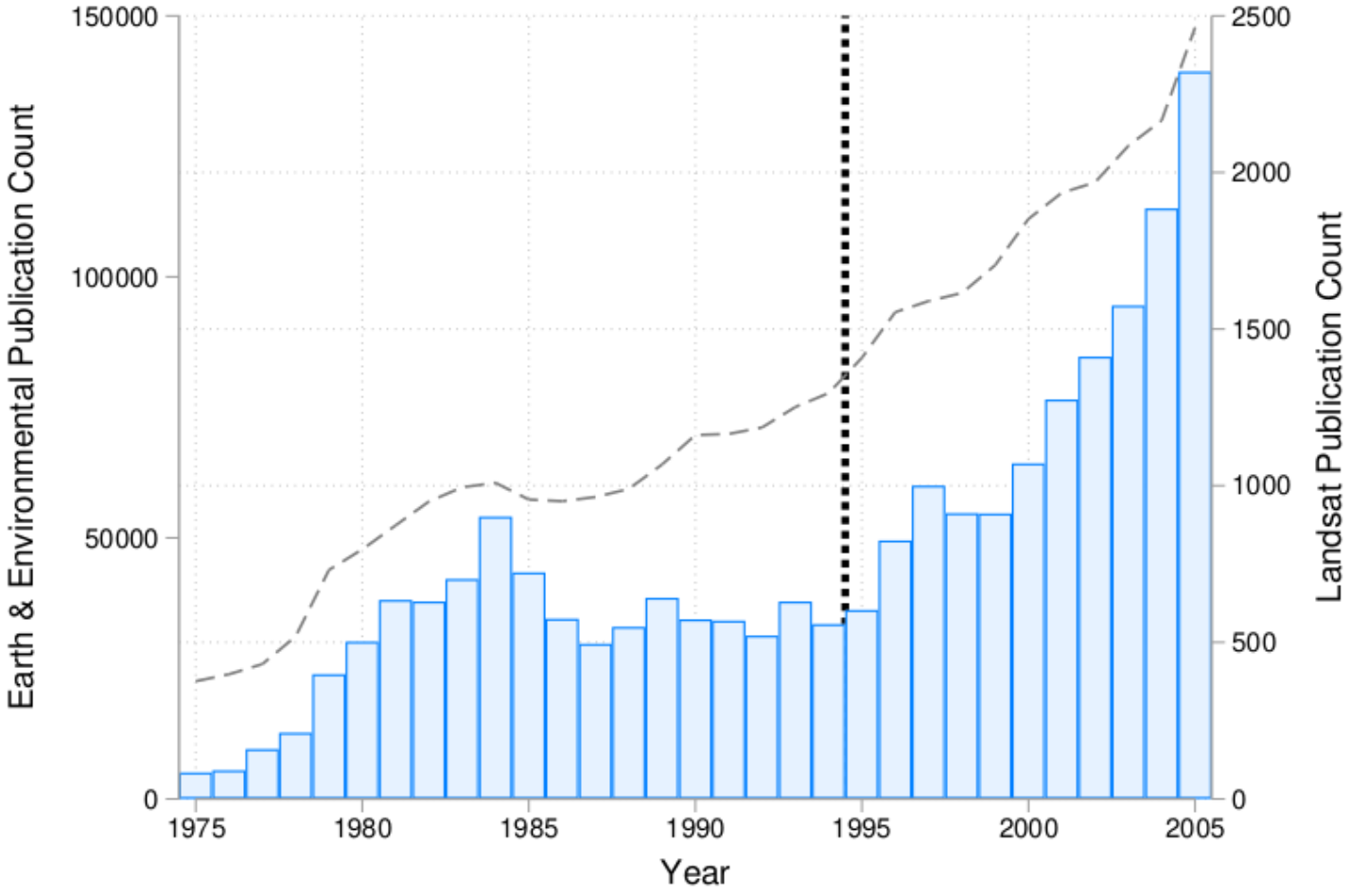
- **DD Regression:**

  Compare high-coverage and low-coverage blocks in terms of their response to reduction in costs (i.e. post 1994) in a difference-in-difference framework with year and block fixed effects

Impact on Rate of Science (Total Pubs)



Impact on Rate of Science (Total Pubs)

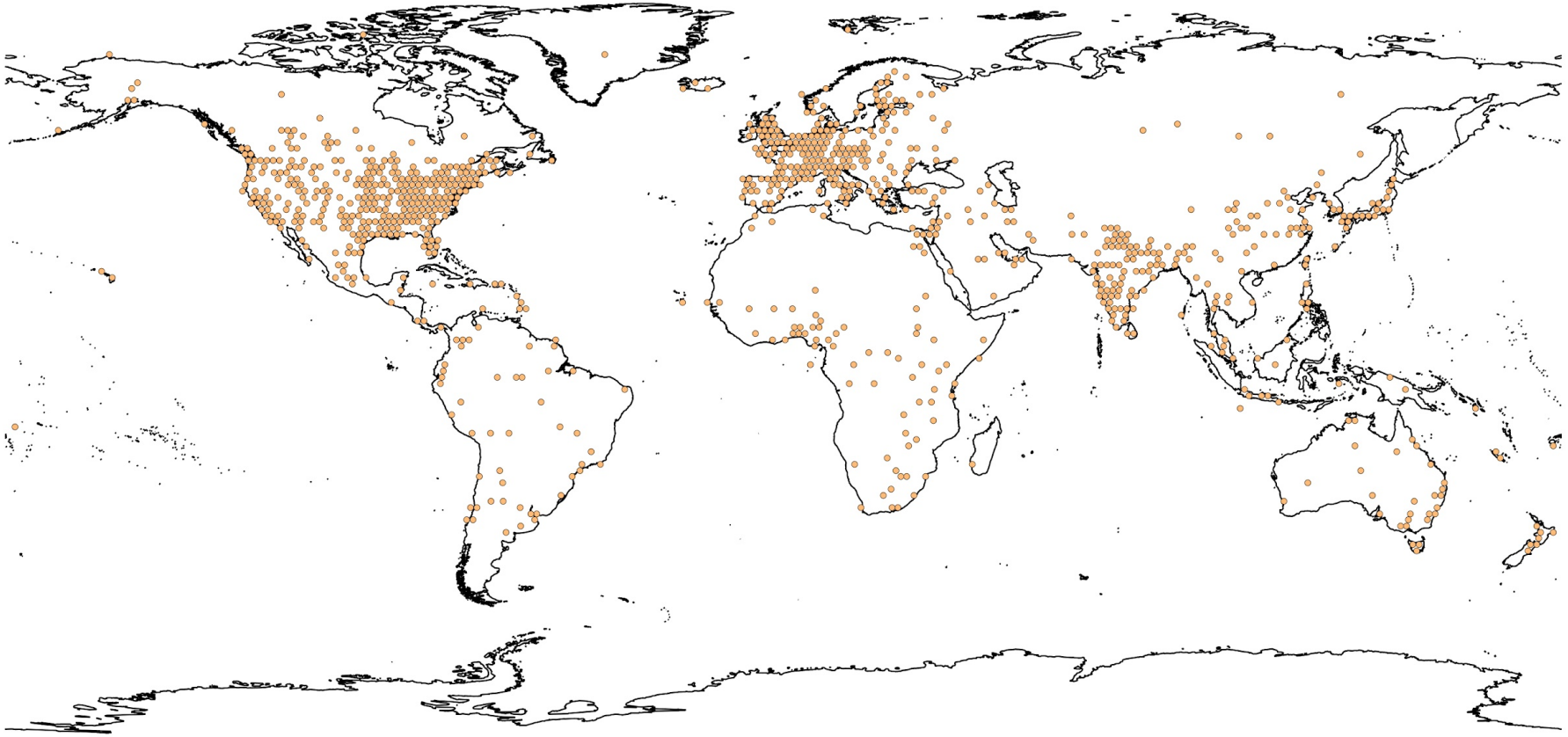


Direction of Science

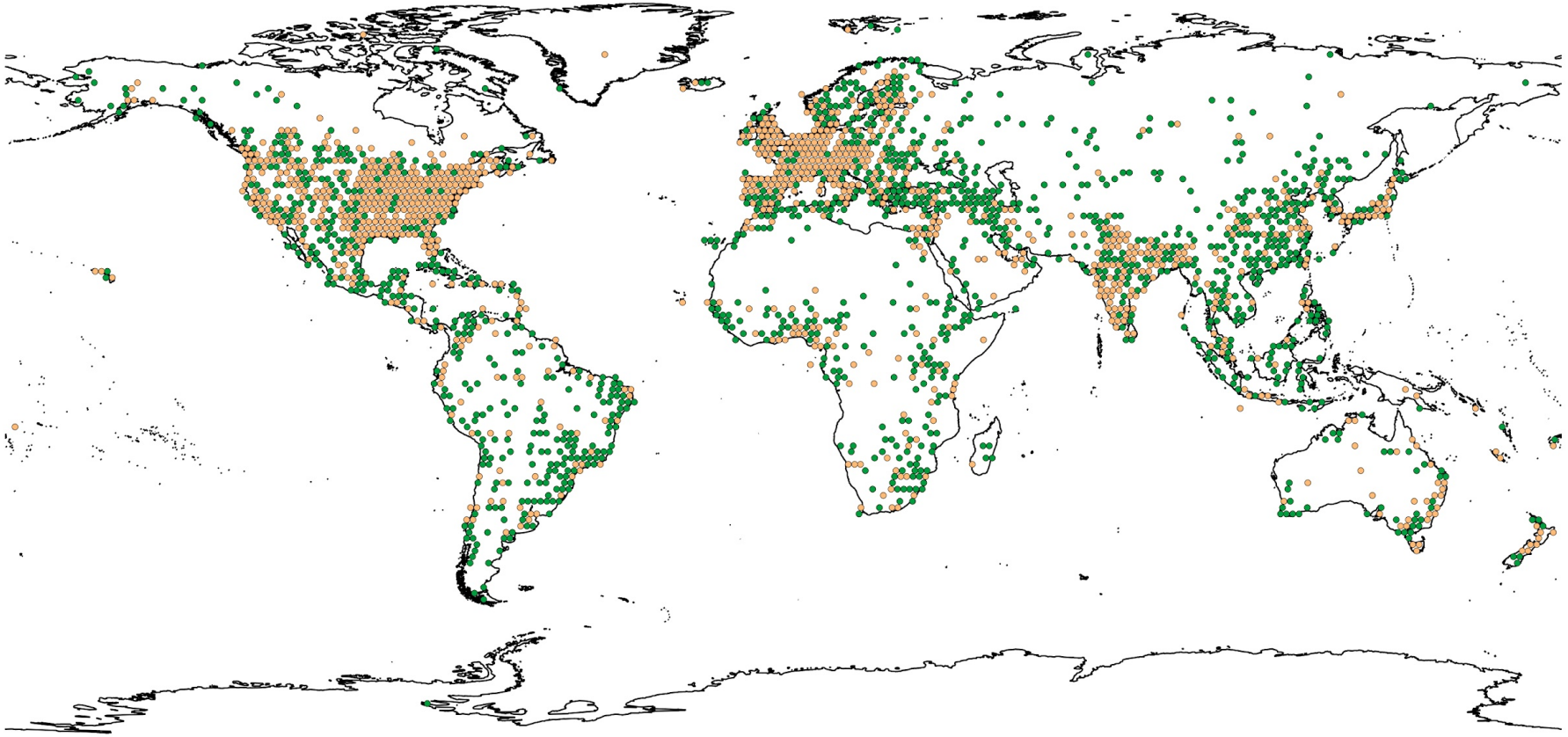
- Improved data access should increase the diversity of researchers
  - Lower-ranked researchers
  - Developing Country researchers

- These researchers are more likely to study “local” topics typically ignored in science
  - More research “about” the developing world and ignored regions

Author Locations (pre 1995)

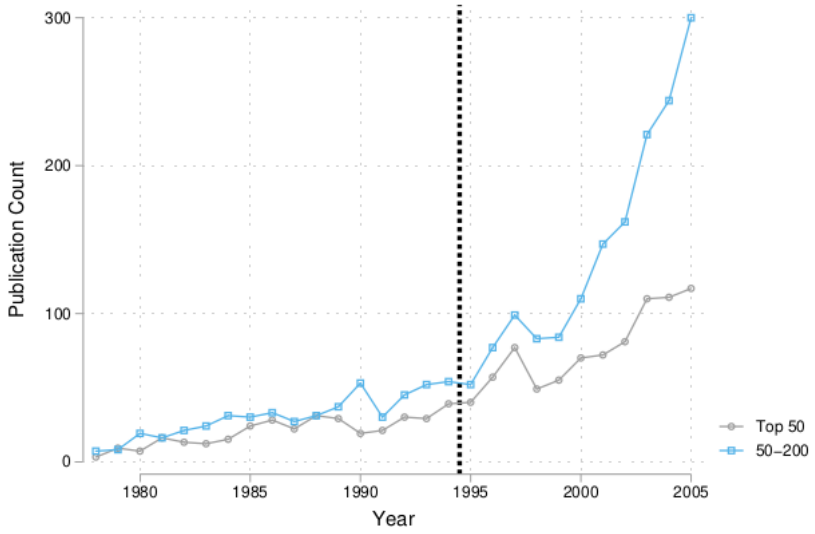


Author Locations (pre+post 1995)



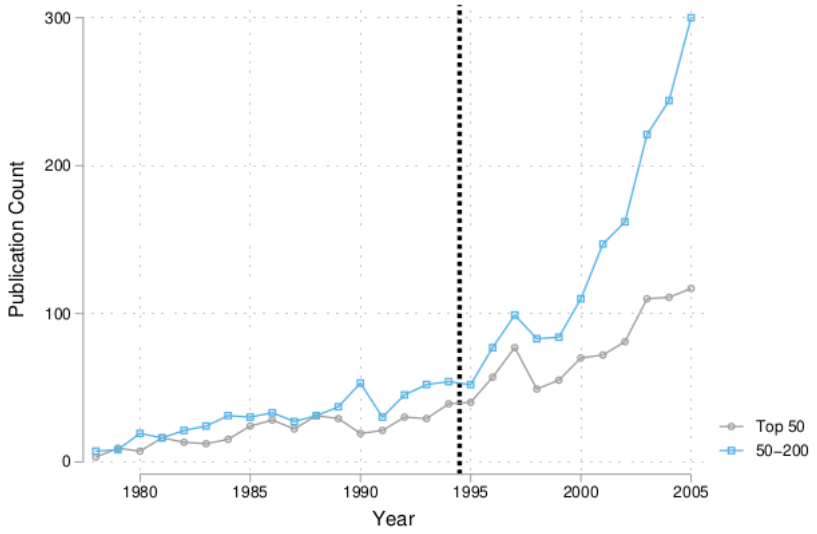
Reductions in Author Inequality

(b) Publications by Authors' Institutional Rank

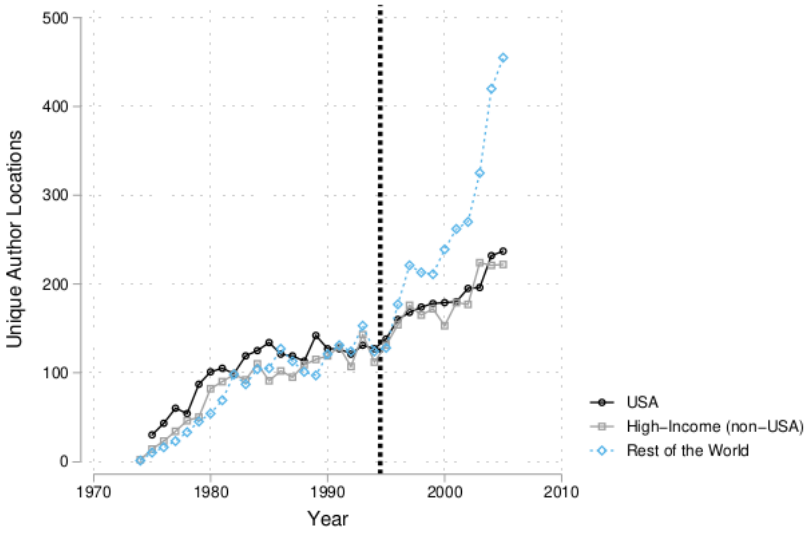


Reductions in Author Inequality

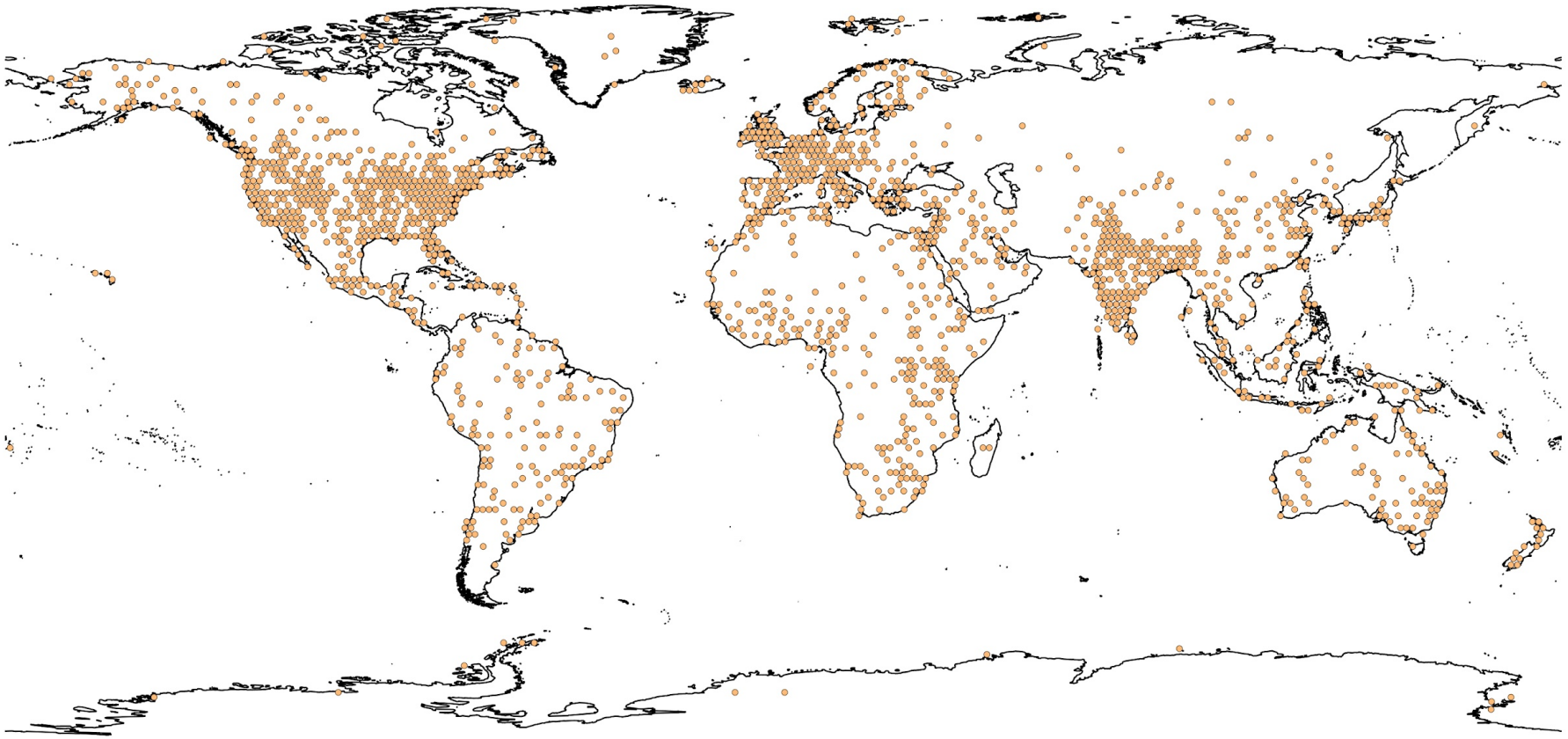
(b) Publications by Authors' Institutional Rank



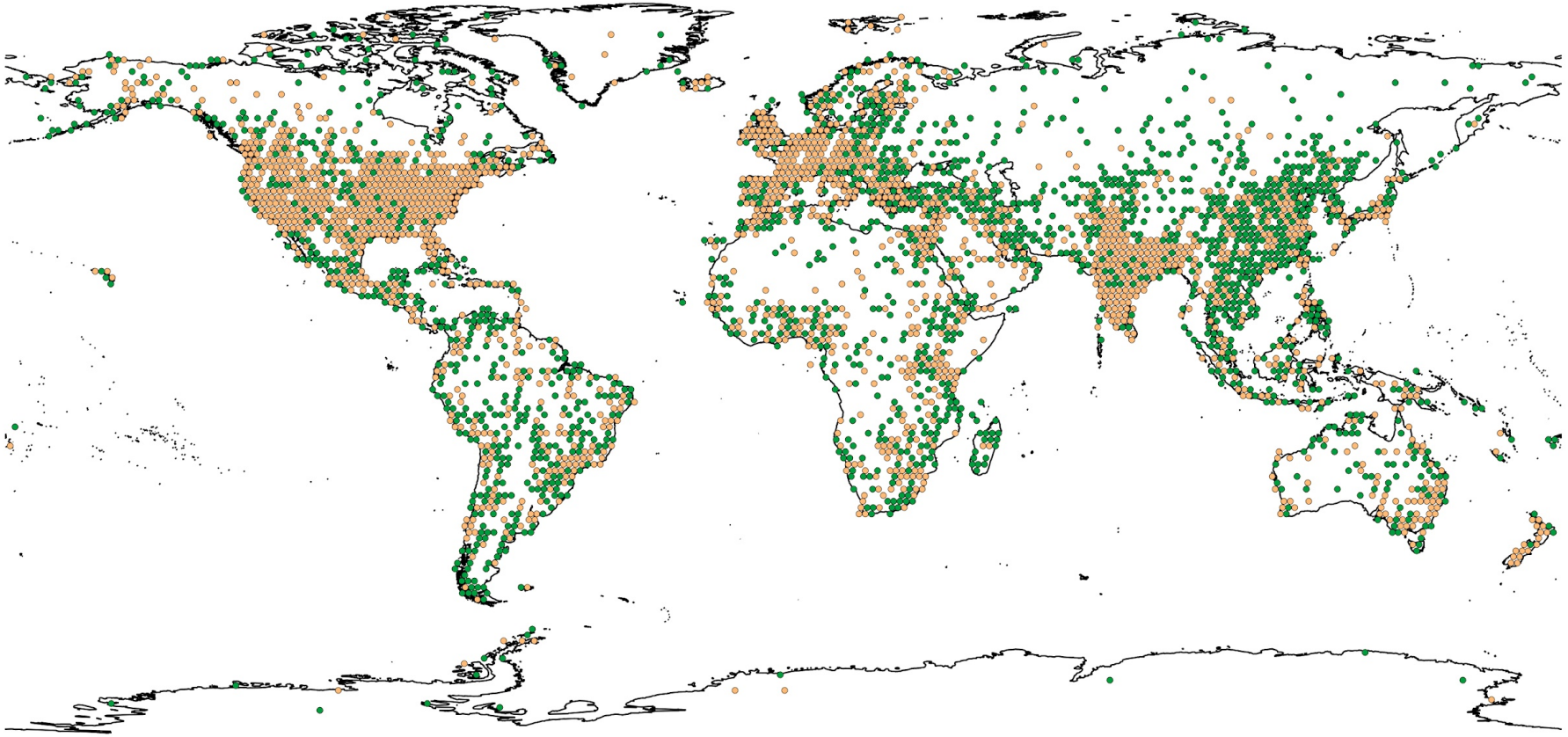
(c) Author Locations by Country Income



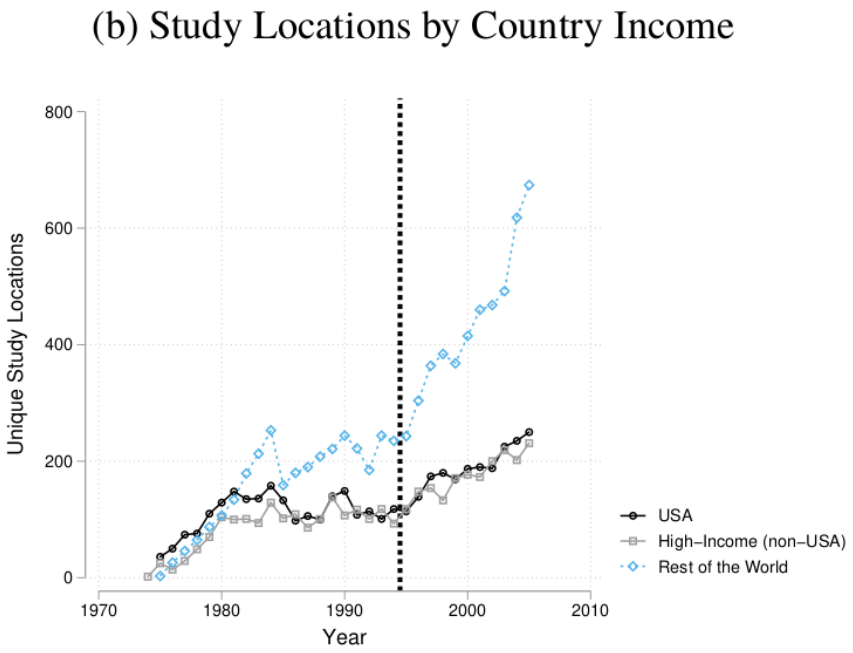
Study Locations (pre 1995)



Study Locations (pre+post 1995)

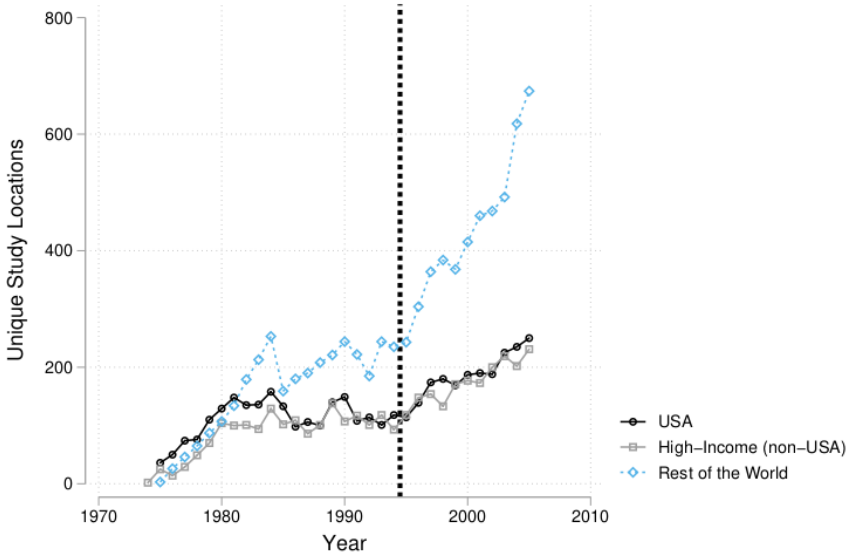


Reductions in Study Inequality

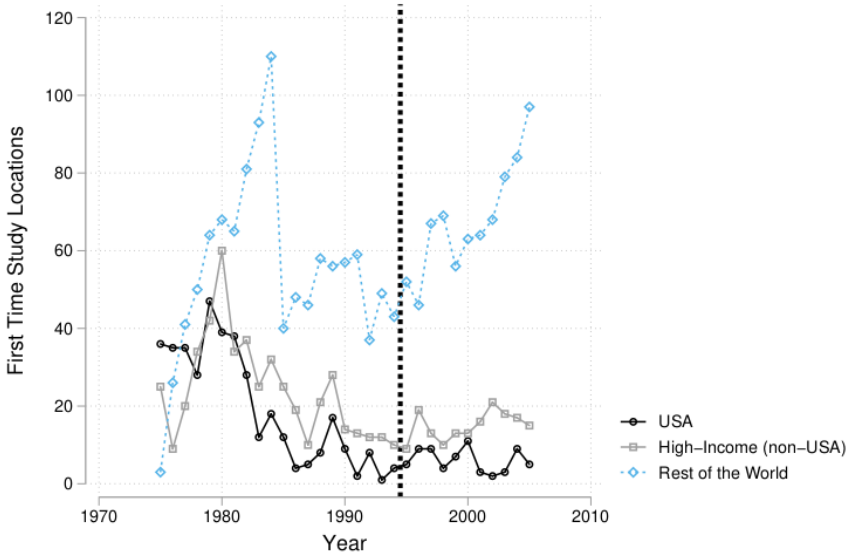


Reductions in Study Inequality

(b) Study Locations by Country Income



(c) First-Time Study Locations by Country Income



Implications

- Even though remote sensing data is lifeblood of modern environmental science, we understand little about the role of data access in shaping science

- Reductions in the cost and sharing restrictions on data have important implications for science
  - ⬆quantity and quality of scientific output
  - ⬆diversity of authors in scientific field
  - ⬆diversity of topics in the field

Concluding Thoughts

- Across two studies, we showed how variation in Landsat coverage and cloud cover can be exploited to evaluate its long term impact

Concluding Thoughts

- Across two studies, we showed how variation in Landsat coverage and cloud cover can be exploited to evaluate its long term impact
- A key enabling factor was granular outcome data matched to Landsat scenes (gold discoveries and scientific papers)

Concluding Thoughts

- Across two studies, we showed how variation in Landsat coverage and cloud cover can be exploited to evaluate its long term impact
- A key enabling factor was granular outcome data matched to Landsat scenes (gold discoveries and scientific papers)
- This method enables long term evaluation of the value of public data at the industry level and accounts for indirect and direct effects of information

Concluding Thoughts

- Across two studies, we showed how variation in Landsat coverage and cloud cover can be exploited to evaluate its long term impact
- A key enabling factor was granular outcome data matched to Landsat scenes (gold discoveries and scientific papers)
- This method enables long term evaluation of the value of remote sensing data at the industry level and accounts for indirect and direct effects of information
- Promise in similar methods (e.g. randomized control trials) to evaluate the impact of public information

One Last Thing ...

(a) Schematic Tube Map
(Source: London Underground)

The Economics of Maps

(w/ Scott Stern)

(a) Schematic Tube Map
(Source: London Underground)

(b) Geographical Map
(Source: Simon Clarke)

twitter: @abhishekn

email: nagaraj@berkeley.edu