# Why should we care about data linkages?

## or Confessions of a Data Hoarder

Tim Beatty

University of California, Davis

# My name is Tim and I am a data hoarder.

Off the top of my head, over the course of my career, I have used :

- Survey data – governmental and have run my own surveys.
- Private data:
  - Scanner (multiple countries / IRI and Kilts / Household and Retail)
  - Private price data (Food and Gas)
  - Firm data (D&B / TD Linx)
  - Internal firm data
  - Location based data
- Administrative data (multiple countries):
  - Patient level data (UK / CA)
  - SNAP (MN/CA) / WIC (CA)
  - Workers Compensation (CA)
  - Food Safety (FSIS)

I rarely use the same data twice. This is a (very) bad strategy.

# Novel linked data is only important if it advances knowledge

**Novel linked data has no intrinsic value**

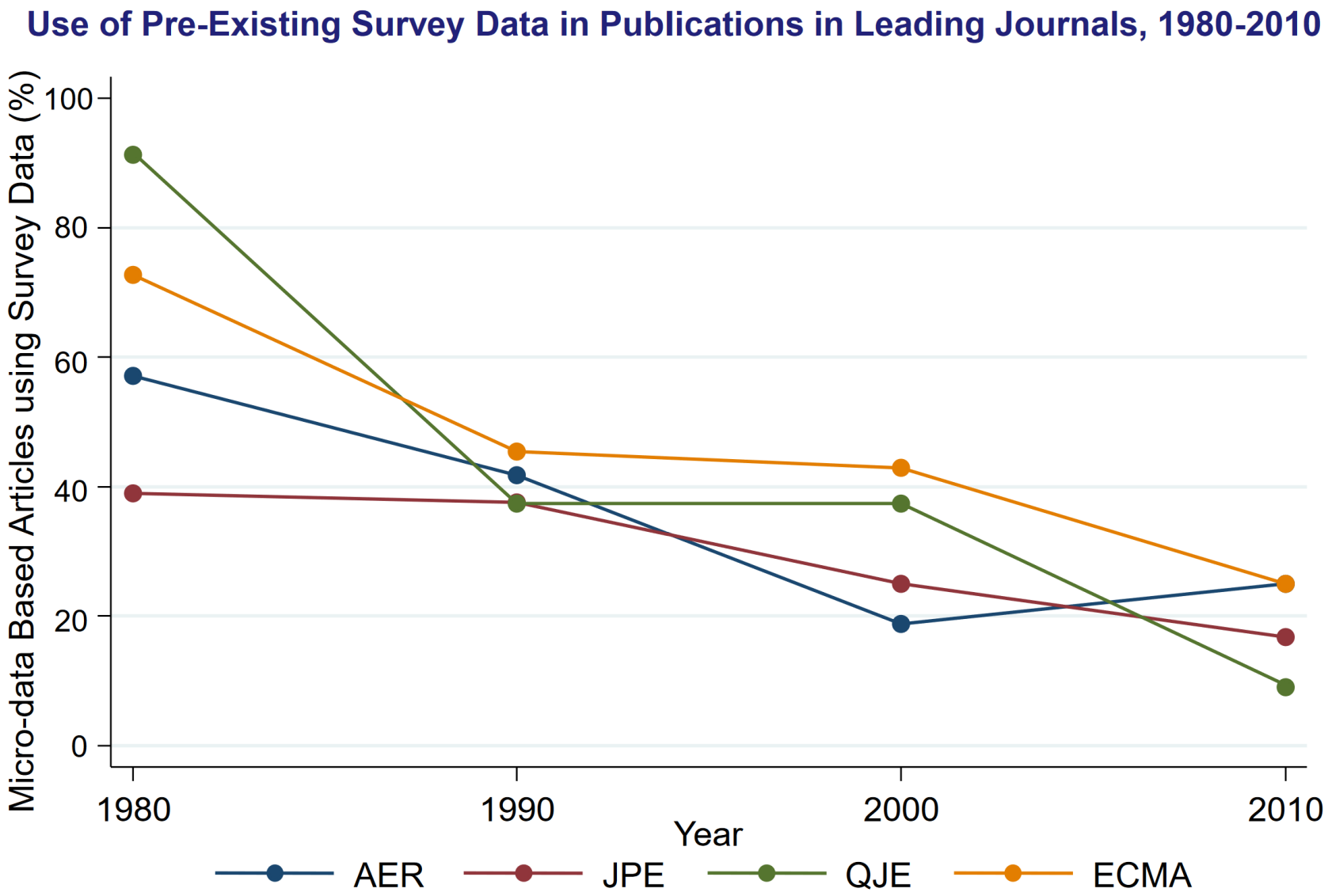
- I try to remind myself of this often.

**Good news is that the potential to advance the research frontier is enormous**

- Ability to follow units over much longer periods.
- Limited attrition.
- Very large sample sizes.
- Allows application of modern causal research designs.

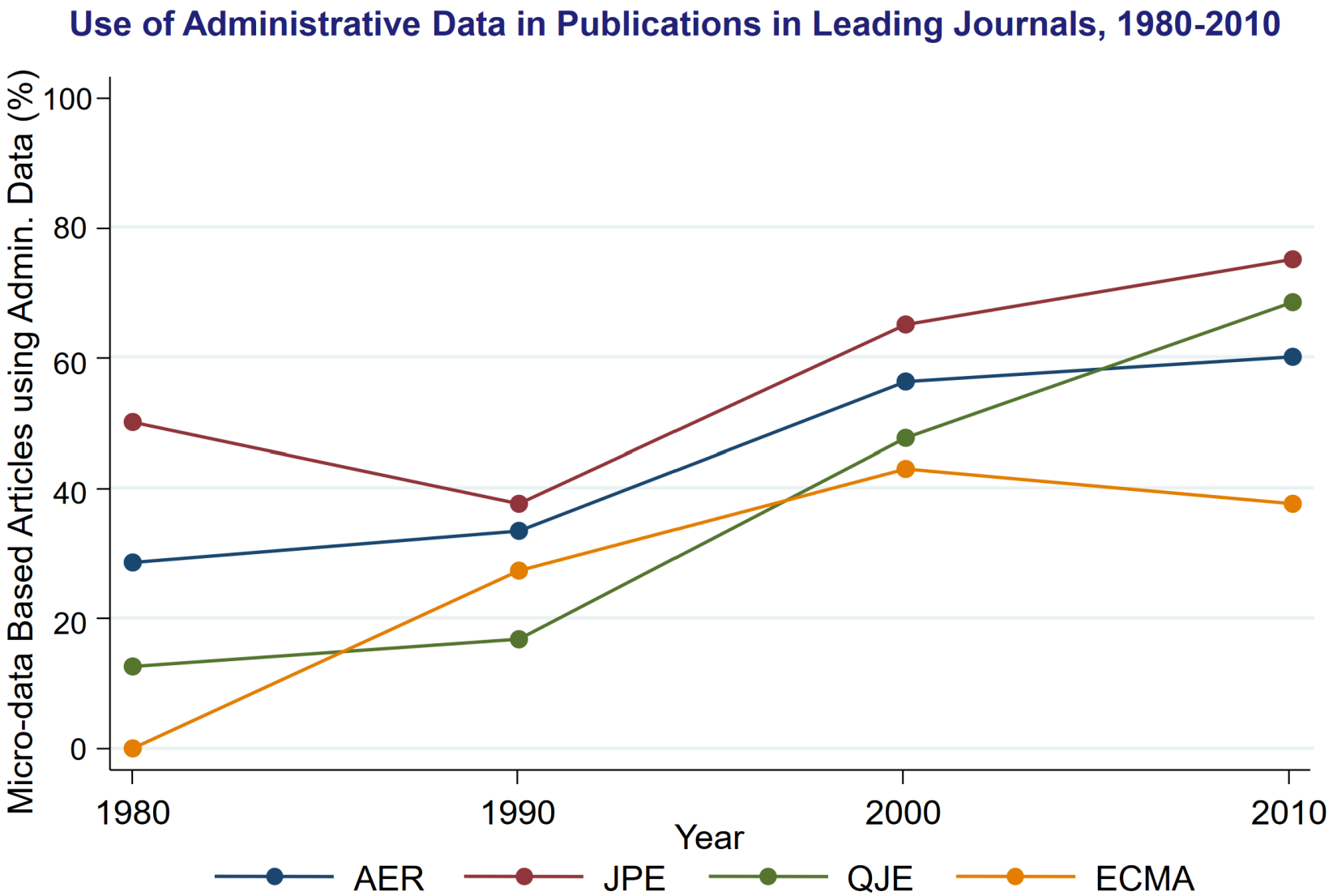
This bad news is this is well understood by our peers in Econ departments.

# Chetty 2012.




Source: Chetty 2012

# Chetty 2012




Source: Chetty 2012

# Whither Agricultural Economics?

I'm unaware of an equivalent figure on trends of administrative data use in Agricultural Economics.

Anecdotally, I think we lag 20 years behind Economics.

**Partially due to topic.**
- Firm/Farm administrative data can be harder to access.
- On the consumer side, a tradition of survey/experimental research.
- In development, a tradition of collecting own data.

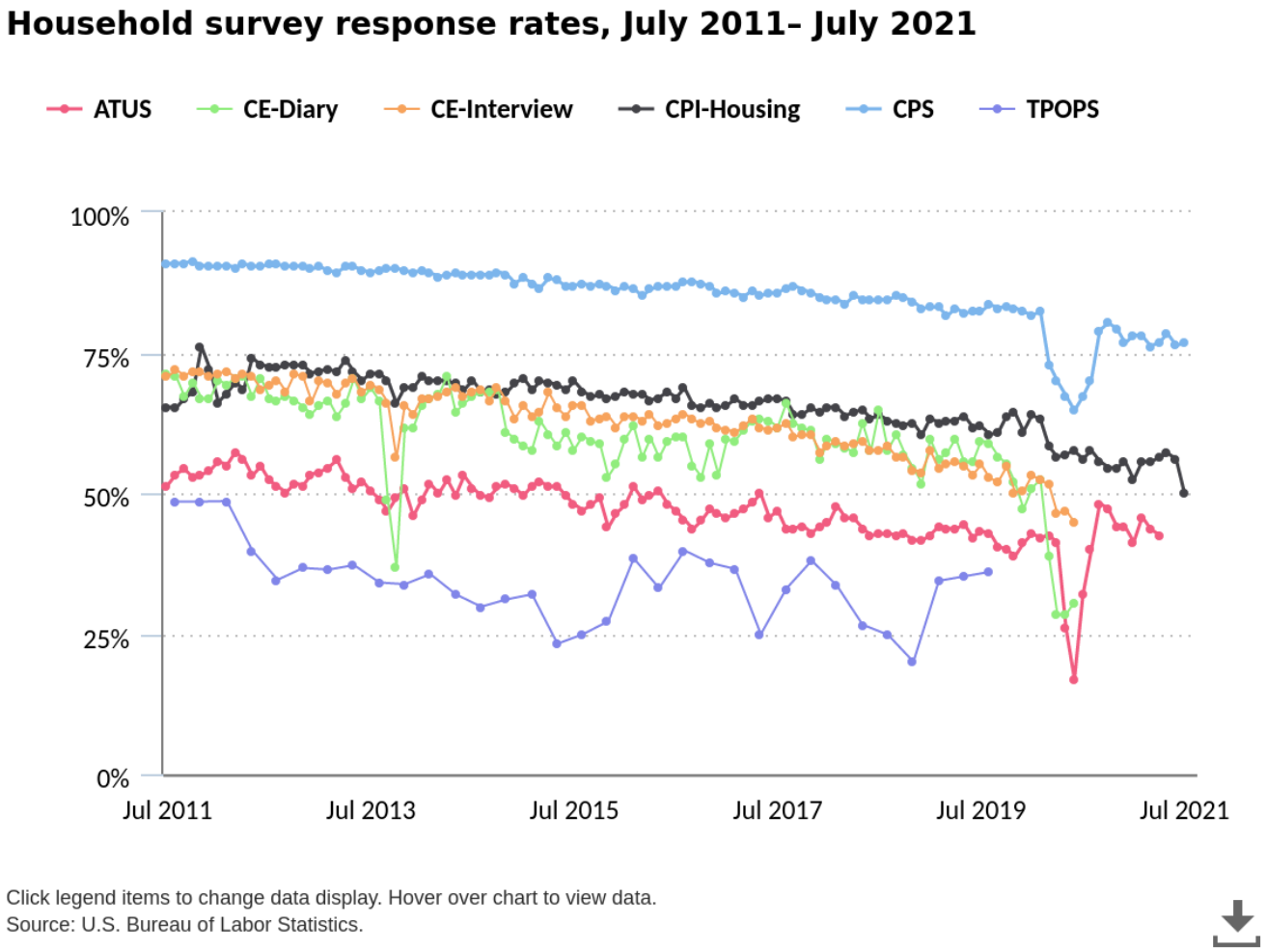
**Partially due to resources** – few ARE departments have ready access to an RDC.

- This is changing: Berkeley, Cornell, Florida (2021), Kentucky, Illinois (2021), Maryland, Minnesota, Missouri, Nebraska, Penn State, TAMU, Wisconsin.

**Partially due to training**

# Background: Crisis in Social Surveys

Many of the large social surveys have low and declining response rates.



# Background: Differential Privacy

Census Bureau adoption of a differential privacy standard may make public data less useful to researchers.

- Census and ACS will introduce additional privacy protection features.

Important work on this topic by Steve Ruggles suggests this may make these data less suitable for social science research.

- Not clear if this will extend to other federal gov't data.

Makes administrative data / linked data even more important.

May make secure access to public data necessary.

# No Free Lunch

There is no free lunch in Admin data.

Working with admin/linked data often requires increased institutional knowledge:

- Charlotte Ambrozek will talk a little about this later today.

Data was not designed for research purposes.

- The data is messier than you realize, even realizing the data is messy.
- It is often poorly documented.

Institutional memory on data is limited.

Time horizons are long.

# Concerns about equity

Administrative data can exacerbate existing hierarchies in the profession.

- Access to novel administrative data requires increased financial resources.
  - Research teams rather than sole investigators.
  - Access to well trained graduate students.
- Access to novel administrative data requires use of social and peer networks.
  - Data is not randomly assigned.

Credit to USDA-ERS for recognizing this and funding grant competitions.

Structural barriers may be difficult to overcome in the long run.

Economics towards a more lab based model.

- Rise of the pre-doc / more post-docs

# Concerns about reproducibility

The use of confidential and non-public data raises issues about reproducibility.

Reproducibility is sometimes possible, at increased cost, and so there will be less of it.

- AEA data editor blog describes the process in detail.

Important work by the AEA on this front.

- Create a social norm.

But challenging for smaller associations like AAEA to follow.

- May also exacerbate inequities.

# Administrative Data as Public Good

Should Administrative data / data linkages be a public good?

- Ideally, linked administrative data would be a public good.
- Public goods are underprovided...

Scarcity breeds novelty.

This suggests an important role for the state.

- This is why we know a lot about Scandinavia these days.
- USDA's Longitudinal Data for Research linking state SNAP data to federal surveys is promising.

Some concern federal support may be subject to change every four years.

# High Fixed Costs / Low Marginal Costs

**The first thing we do, let's kill all the lawyers" – noted admin data user W. Shakespeare.**

Benefits to centralized arrangements.

- Consortia can arrange master agreements that benefit multiple researchers.
- Knowledge of data structures can be institutionalized.
- Clean once / use lots.
- Data security can be enforced.

Examples:

- Kilts Center.
- California Policy Lab.

Headed towards gated data communities / proliferation of data platforms.

# Returns to investment in data / data linkages

Novel data is a input, not an output.

- As academic agricultural economists, our unit of output is journal articles.
- Need to find ways to increase the returns to the provision of a public good, e.g. increased use of data citation.
- Public good provision is best saved for life after tenure.
- Again, suggests the role for specialists where output is aligned with incentives.

Novel data may help get a paper published. But returns attach to the paper, not the data.

# Careful what you wish for...





**SUPREME COURT OF THE UNITED STATES**

No. 18–481

FOOD MARKETING INSTITUTE, PETITIONER *v.*
ARGUS LEADER MEDIA, DBA ARGUS LEADER

ON WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE EIGHTH CIRCUIT

[June 24, 2019]

JUSTICE GORSUCH delivered the opinion of the Court.

Congress has instructed that the disclosure requirements of the Freedom of Information Act do "not apply" to "confidential" private-sector "commercial or financial information" in the government's possession. But when does information provided to a federal agency

# The Frontier is moving

The data frontier is moving / has moved to the commercial sector.

- Location based data such as Safegraph became widely available during the pandemic.
- Credit data (e.g. UC-CCP)
- J.P. Morgan Chase Institute.
- Partnerships with private firms (e.g. work by Alexandra Hill on Strawberry farmers / Sarah Smith on Tomato growers).

# Wrapping Up

Administrative data and data linkages are not the future, they're the present.

- My sense is the recent crop of Assistant Professors gets this.

AgEcon is well behind – it will take investment to catch up.

- Commend ERS for pushing the profession forward on this.

I think equitable access needs to be at the center of any move forward.

Recognize, we will always be fighting the last war.