

AI and Data in Social Sciences

Directions for Social and Behavioral Science Methodologies in the Next Decade.

An opinionated view, not a research presentation, or a review...

AI can improve data infrastructure for the social sciences through enabling innovations in measurement.

Frame **how** AI enables new forms of measurement and **why** these new forms of measurement are critical to building globally and nationally consistent measures of complex social, environmental and economic phenomena.

Data and AI: Three Ideas

AI systems are ***not databases*** but they can be used to reason over heterogenous inputs to produce data.

The ***schema is dead*** and this has important implications for social science.

The ***boundary between data and models is blurring***- new measurement methodologies, facilitated by AI, shift the epistemological position of "data" 🌶️ 🌶️ and we just have to get over it .

Me writing a “prompt”

Respond in Json form, using 1 for Texas and 0 for France. Your answer should repeat the question and then include your 1 or 0 label:

1. Are people happier in Paris Texas or Paris France?
2. Are there more cows in Texas or France?
3. Which has a higher per capita reate of pickup truck ownership?
4. Which has better bread?

REMEBER TO ONLY RESPOND IN JSON FORM

Machine response to my prompt

```
[
  {
    "question": "Are people happier in Paris Texas or Paris France?",
    "answer": 1
  },
  {
    "question": "Are there more cows in Texas or France?",
    "answer": 1
  },
  {
    "question": "Which has a higher per capita rate of pickup truck ownership?",
    "answer": 1
  },
  {
    "question": "Which has better bread?",
    "answer": 0
  }
]
```

Respond in Json form, using 1 for Texas and 0 for France. Your answer should repeat the question and then include your 1 or 0 label:

1. Are people happier in Paris Texas or Paris France?
2. Are there more cows in Texas or France?
3. Which has a higher per capita reate of pickup truck ownership?
4. Which has better bread?

REMEBER TO ONLY RESPOND IN JSON FORM

Me writing a "prompt"

AI systems are not databases!

Machine
response to my
prompt

Copy

```
{
  "question": "Are people happier in Paris Texas or Paris France?",
  "answer": 1
},
{
  "question": "Are there more cows in Texas or France?",
  "answer": 1
},
{
  "question": "Which has a higher per capita rate of pickup truck ownership?",
  "answer": 1
},
{
  "question": "Which has better bread?",
  "answer": 0
}
]
```

Can you repeat the above but include reasoning to justify your judgement.

The ability to reason is part of what makes these things exciting...

 Copy

```
[
  {
    "question": "Are people happier in Paris Texas or Paris France?",
    "answer": 1,
    "reasoning": "People in Paris, Texas, may experience a slower pace of li
  },
  {
    "question": "Are there more cows in Texas or France?",
    "answer": 1,
    "reasoning": "Texas has a massive cattle industry, leading to a higher r
  },
  {
    "question": "Which has a higher per capita rate of pickup truck ownershi
    "answer": 1,
    "reasoning": "Pickup trucks are more commonly associated with American l
  },
  {
    "question": "Which has better bread?",
    "answer": 0,
    "reasoning": "France is renowned for its baking traditions, including ba
  }
]
```

Can you repeat the above but be sure to ground your results in a web search. If you cannot find relevant information on the internet make your response "null". Add a field for "query" which includes the search terms you used. Add a field for "citation" which includes urls supportign your claim.

The ability to reason over heterogenous inputs, in this example websites, is the core of makes AI based measurement exciting for the social sciences.

```
Copy
    "question": "Are people happier in Paris Texas or Paris France?",
    "answer": "null",
    "reasoning": "null",
    "query": "happiness in Paris Texas",
    "citation": "null"
  },
  {
    "question": "Are there more cows in Texas or France?",
    "answer": 1,
    "reasoning": "Texas has a massive cattle industry, leading to a higher r",
    "query": "number of cows in Texas",
    "citation": "https://animal-world.com/how-many-cows-are-in-texas/"
  },
  {
    "question": "Which has a higher per capita rate of pickup truck ownershi",
    "answer": 1,
    "reasoning": "Pickup trucks are the most popular vehicle type in Texas,",
    "query": "pickup truck ownership rate in Texas",
    "citation": "https://motorandwheels.com/pickup-truck-demographics-segmer"
  },
  {
    "question": "Which has better bread?",
    "answer": 0,
    "reasoning": "France is renowned for its baking traditions, including ba",
    "query": "bread quality in France",
    "citation": "https://www.chefsresource.com/why-is-bread-such-a-common-fc"
  }
]
```

Is AI good at producing data?

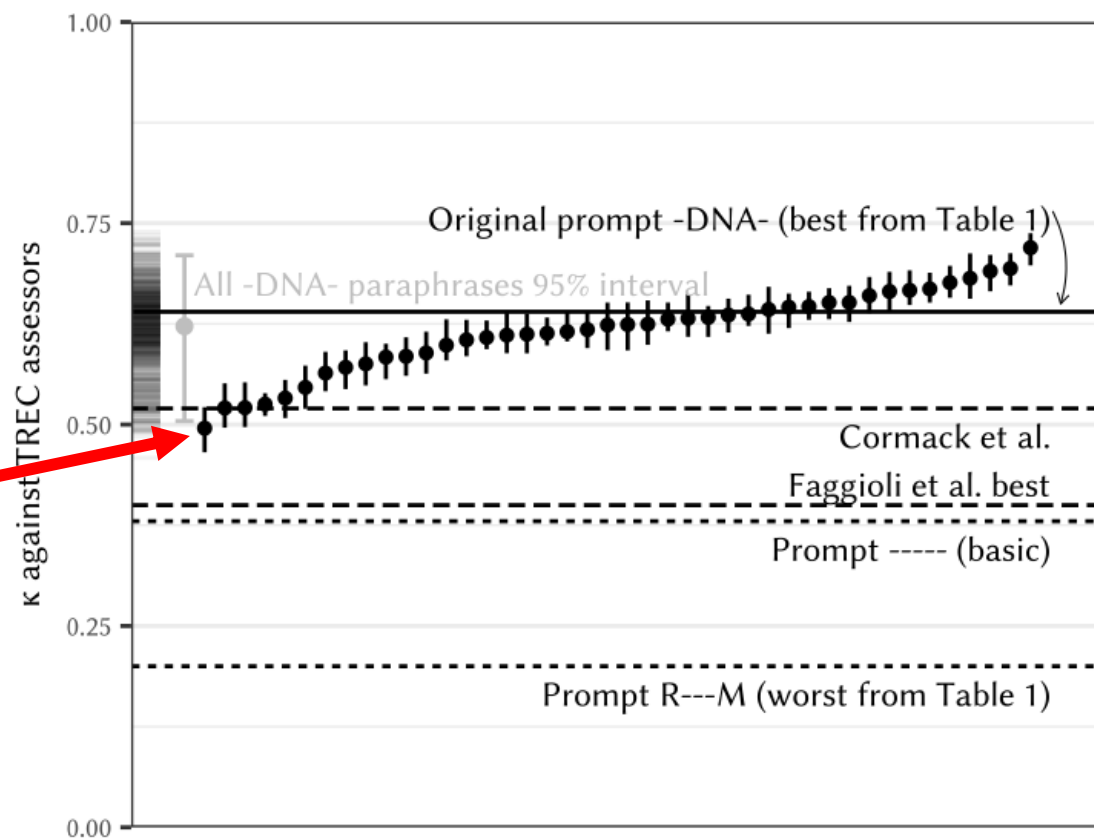
To evaluate AI produced data one needs “benchmarks”, data sets containing the correct answer and all relevant inputs.

For example, NIST for 20+ years has published a benchmark called “TREC” designed for measure the relevance of a document to a question.

There are large within/between prompt variations in quality.

Using TREC we found GPT-4 was able to quantitatively measure relevance of a document better than humans (crowd workers and experts in web search (employees)).

AI based measurement is a moving target, what is possible is evolving rapidly.



	Relative accuracy	Latency	Relative throughput	Relative cost
Employees	+24%	hours–days	$\times^1/_{100}$	$\times 8$
Best crowd	+19%	hours–days	$\times^1/_{15}$	$\times 5$
Typical crowd	—	hours	$\times 1$	$\times 1$
LLM (GPT-4)	+28%	minutes–hours	$\times 10$	$\times^1/_{20}$

Data and AI

- Asking an AI model to regurgitate facts is not interesting or useful- ***AI systems are not databases.***
- Using a model's **ability to reason**, especially over **heterogenous inputs** like text, tabular data, images, maps/satellite, voice, ... alone or in combination is exciting.
- A bit of evidence that it works.
- Why does this matter???

The Schema is Dead

The ability to reason over heterogeneous inputs and output well reasoned descriptors of a place, person, or other entity **at scale** is fundamentally new.

AI generated data holds particular promise for complex latent constructs (e.g. “vulnerability”, “resilience” or “well-being”). These are difficult to measure, require diverse inputs, and the information available varies between entities.

The schema is dead, we can integrate different forms of data by reasoning over them, rather than trying to shoehorn data into a single table.

Robust validation of AI generated data is essential “ground truth” for complex latent variables often does not exist.

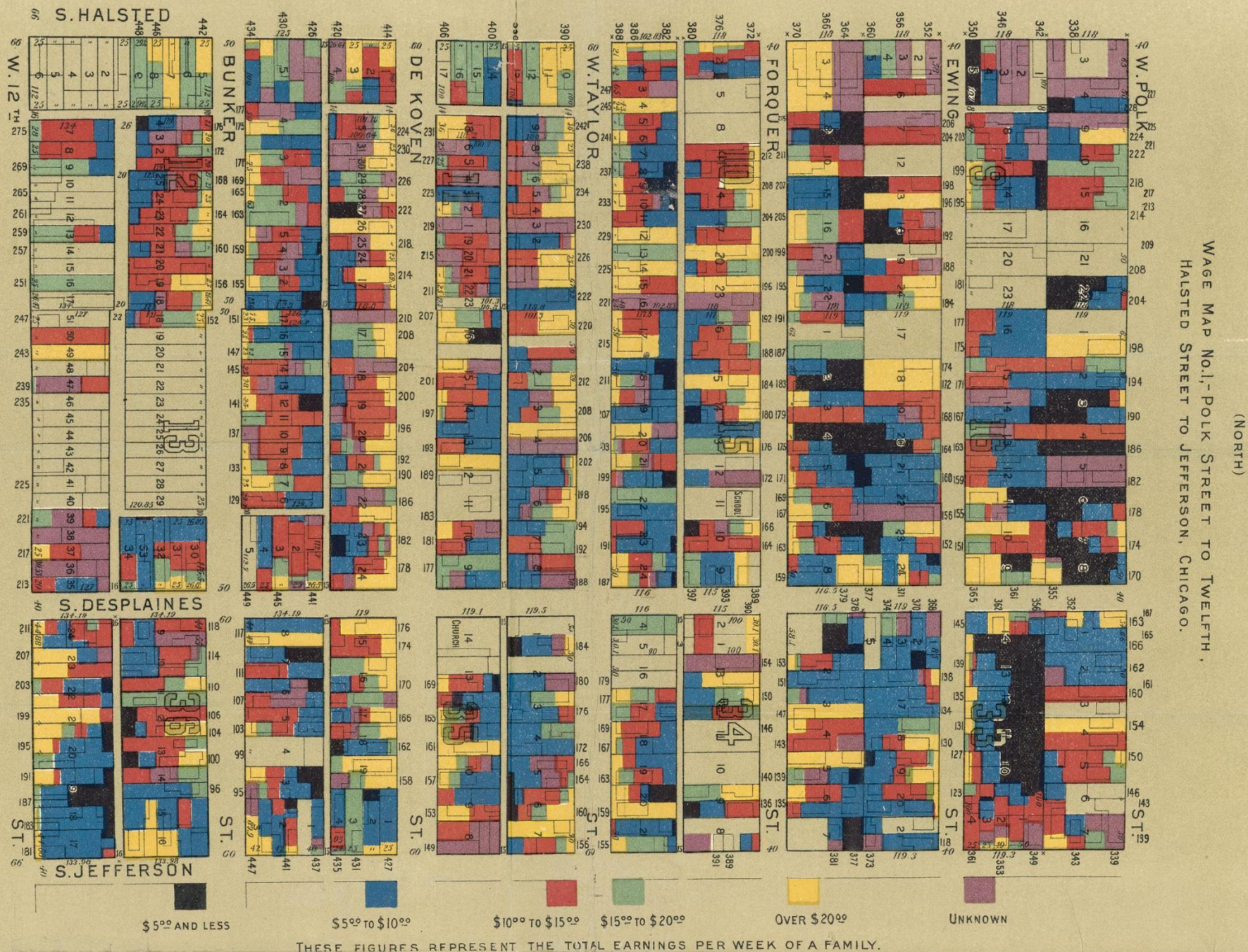
Is this data?

Data after all is supposed to represent the real world, not output from a robot.

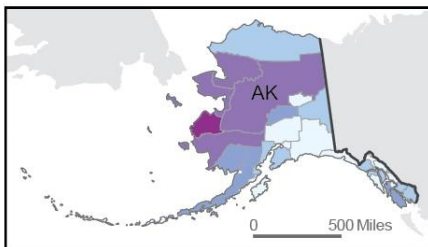
This idea that the output from an AI model would hold the same epistemological position as data rooted in observation may seem unsettling/objectionable.

I'd argue that we're already there. That much of what we consider first class data is from a model, even the official statistics we use to make large national resource allocations (albeit those models are easier to interrogate than an AI system with billions of parameters.)

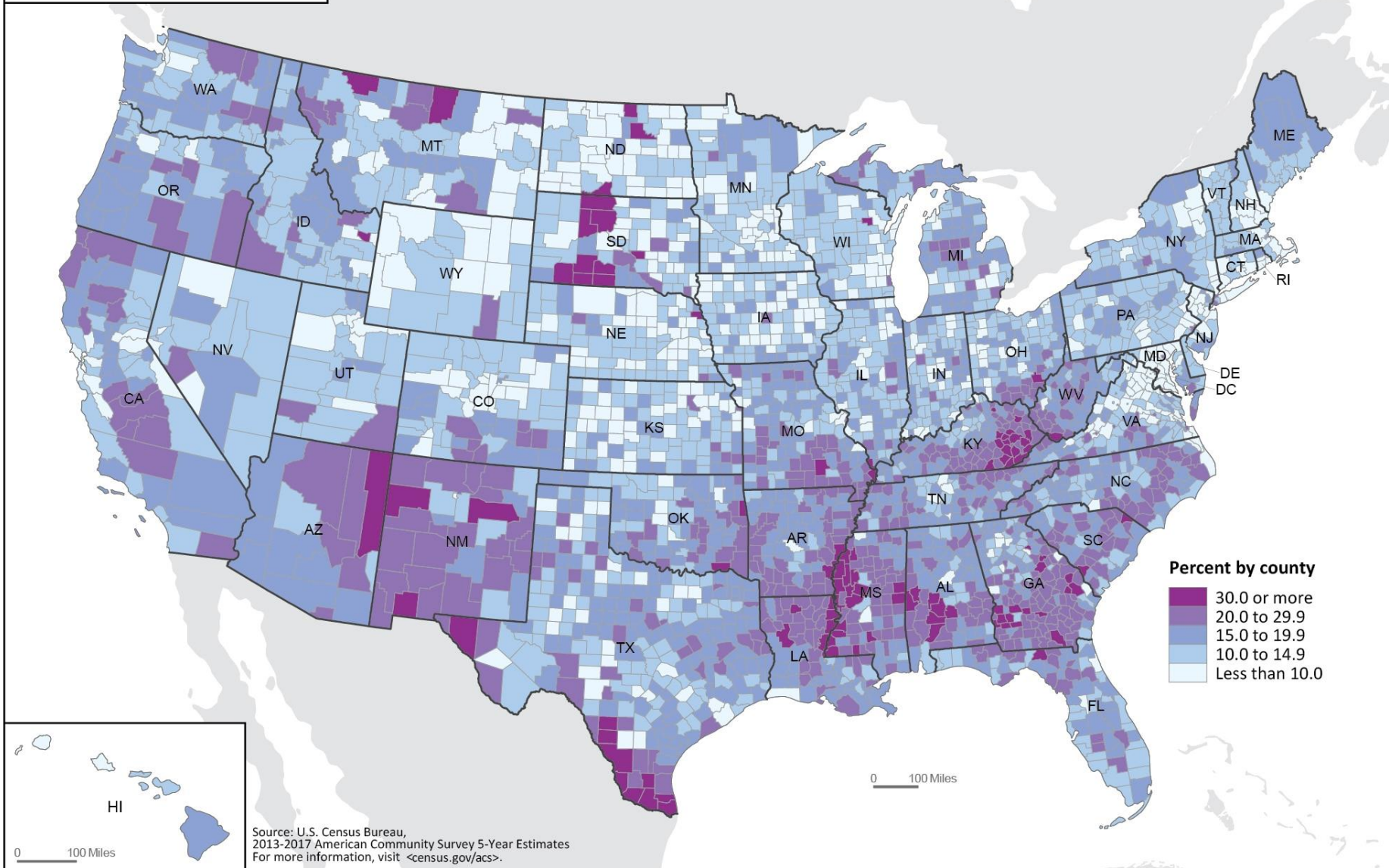
Hull House 1895 Map of Income in Chicago



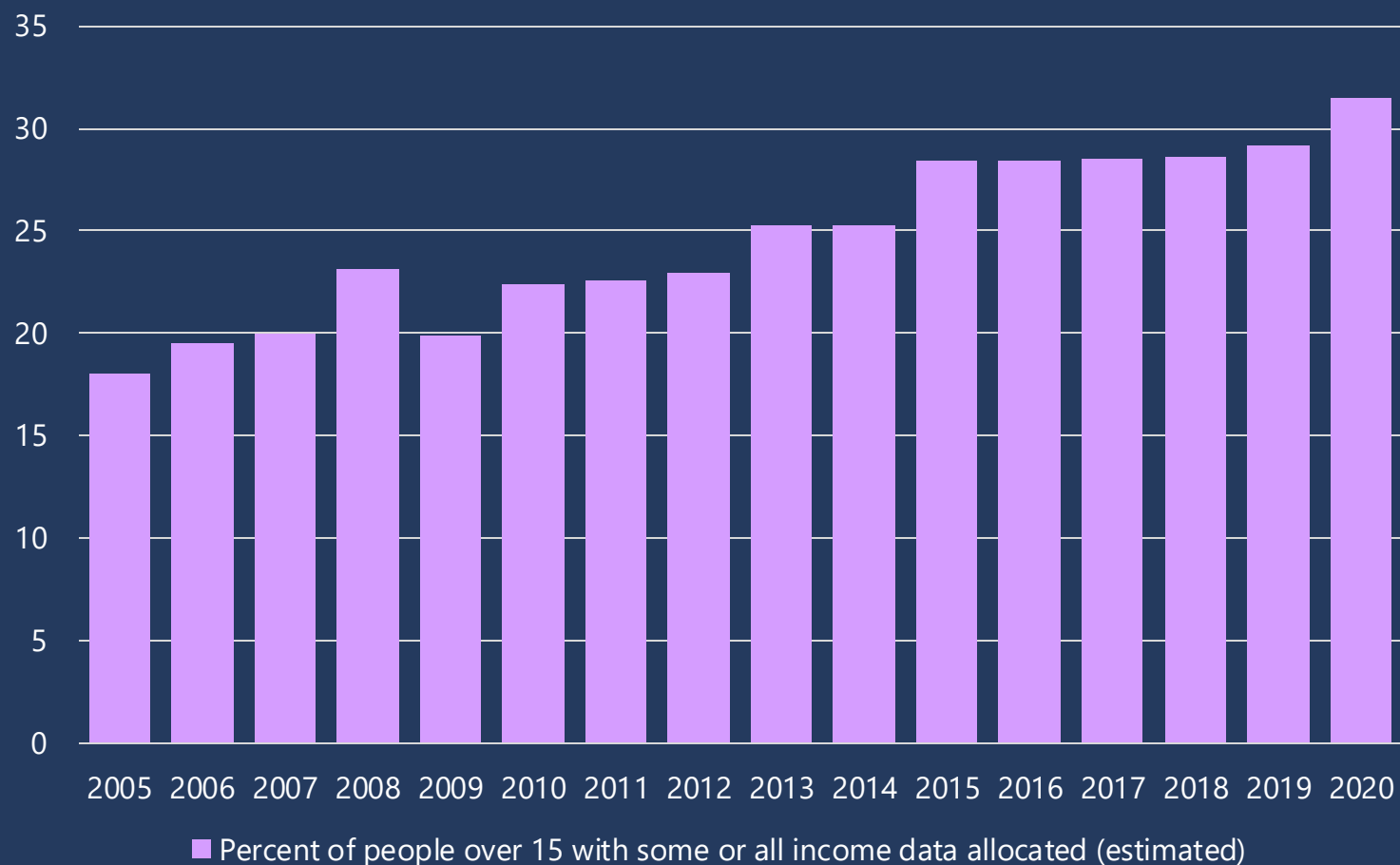
THESE FIGURES REPRESENT THE TOTAL EARNINGS PER WEEK OF A FAMILY.



American Community Survey 5-Year Estimates
**County Poverty Rate for the
United States: 2013-2017**



Allocation Rates for Income Data (American Comm. Survey)



About a third of the respondents to the American Community Survey leave some or all income questions blank. These blanks are corrected via “allocation.”

Our measures of income and poverty in the United States are, to an increasing degree, due to models.

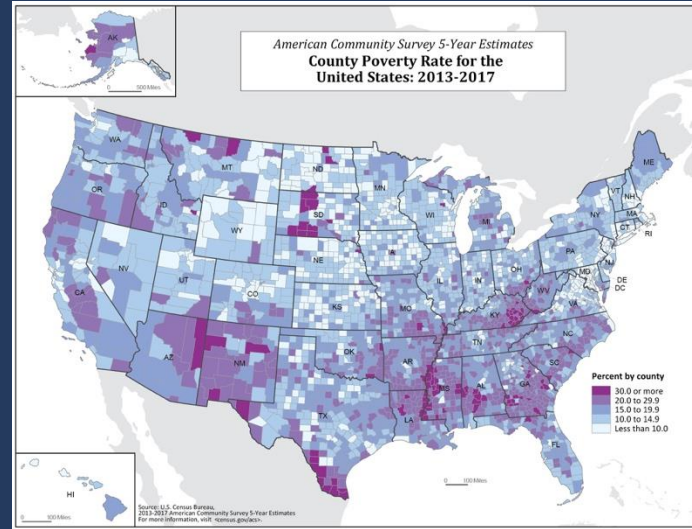
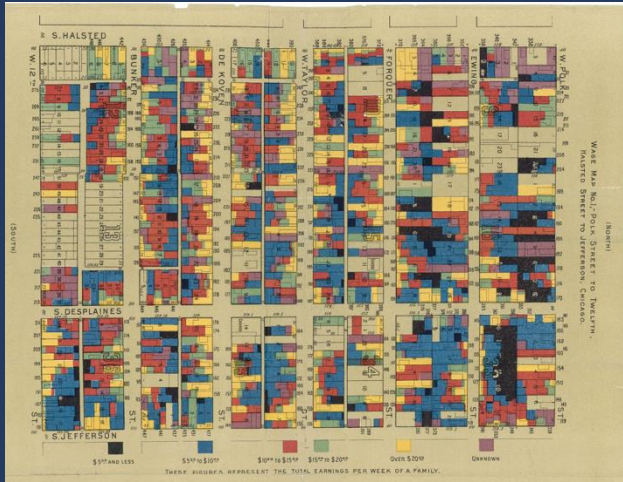
[US Census 202 p. 77](#)

Allocation is “the use of statistical procedures to impute the values for these missing data. Item allocation rates—the proportions of responses allocated for an item in a given geographic area—are determined by the content edits performed on the individual raw responses and closely correspond to item nonresponse rates.”

Data as a single realization of a model....

Due to the complexity of the sampling design and the weighting adjustments performed on the ACS sample, unbiased design-based variance estimators do not exist. As a consequence, the direct variance estimates are computed using a replication method that repeats the estimation procedures independently several times. The variance of the full sample is then estimated by using the variability across the resulting replicate estimates.

[acs_design_methodology_ch12_2014.pdf \(census.gov\)](#)



There's been a long-term trend toward more model assisted data.

**DATA AS
DIRECT OBSERVATION**

MODEL-BASED DATA
Fully Synthetic data.

AI Generated Data:

AI generated data is exciting and potentially transformational. There are multiple flavors:

- **Fusion:** Combining inputs to produce new data sets. This might refine/enrich existing data sources, reasons across heterogeneous supports for interdisciplinary problems (people/pixels), measure complex latent constructs.
- **Synthesis:** Generating new data, imputing missing data. Satellite Foundation Models estimate what's under a cloud occluded pixel in a satellite image. Estimating how a change in A relates to change in B (e.g. where might people go after a flood).

Conclusions

- Innovations in AI-based measurement have the potential to new forms of social scientific data infrastructure.
 - The ability to handle heterogenous inputs opens the potential for globally/nationally integrated data sets – this is critical for a changing planet
- In the social sciences measurement is often seen as a means to an end, not an end itself.
 - ... but to observe change we need consistent measurement (not one-off papers). NSF and other research funding agencies have a role.
 - Robust benchmarks are critical, again this is a potential place NSF to spur activity.
- **3 Ideas:** Not a database, schema is dead, boundary between models and data is increasingly blurry.