Tidy Census + LLMs

Divij Sinha 01-30

Tidy Census

The Census Bureau

Data Collections

- Decennial Census: last round 2020
- American Community Survey: every year
- Other useful surveys - CPS, Economic Survey etc.

Decennial Census

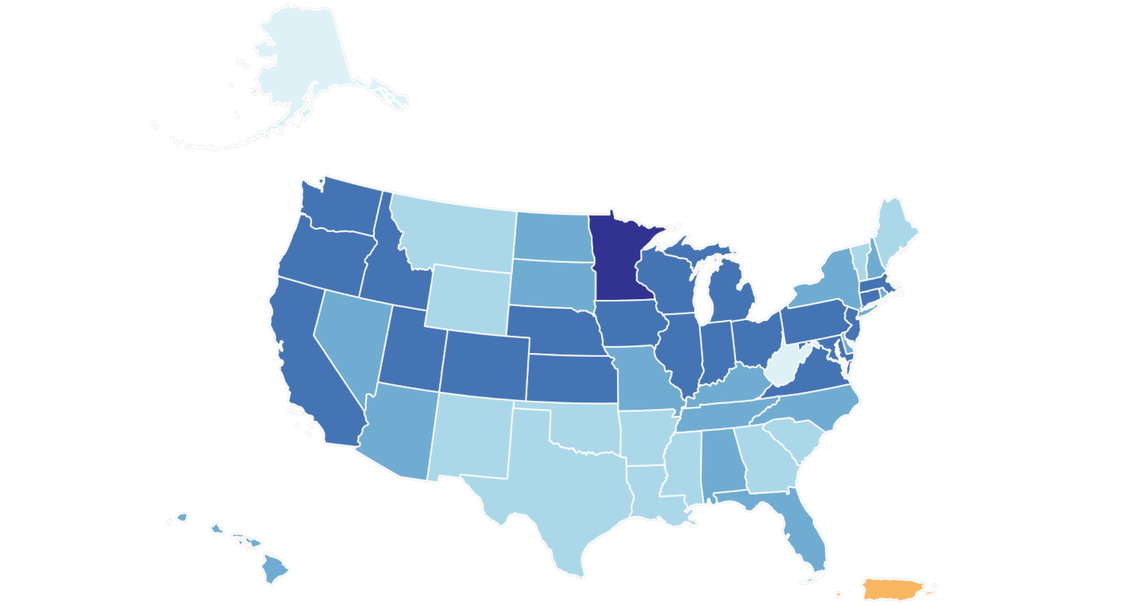
- Every 10 years
- Pros
  - No sampling errors, no estimations
  - All crosstabs present
  - Counts everybody (Does it?)
    - How is it conducted?
- Cons
  - Slow
  - Expensive

2020 Census Self-Response by State

This map features self-response rates from households that responded to the 2020 Census online, by mail, or by phone.

National Self-Response
67.0%

Alabama Self-Response
63.6%



Self-Response Rate (%)
0-15 16-30 31-40 41-50 51-56 57-62 63-68 69-74 75-85 86-100

Alabama Total Self-Response Rate

Select Date
1/28/2021

Select Mode
Total

Select State
(All)

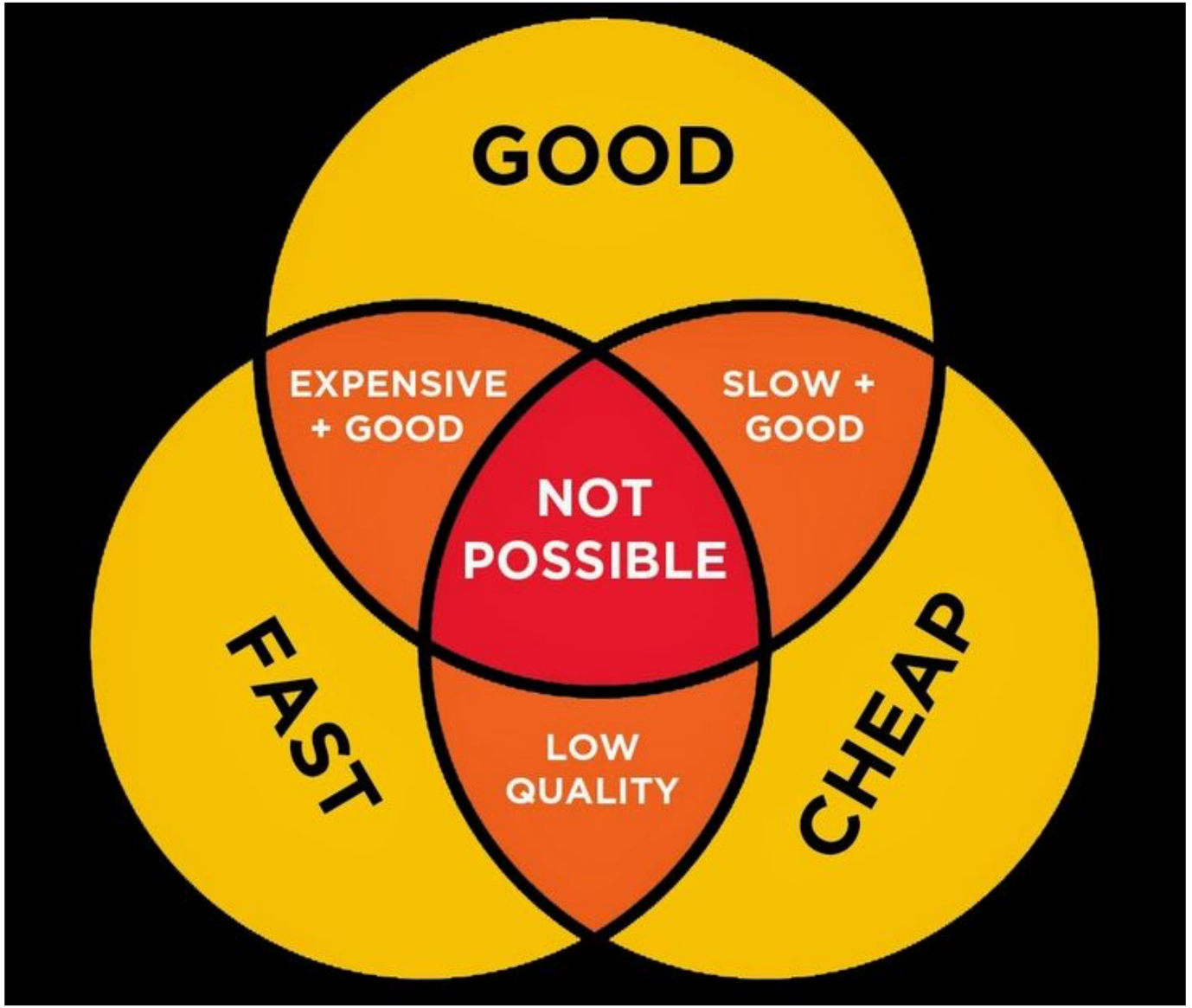
Geographies

State
County
City
Congressional District
Town and Township
Tribal Area

Top

American Community Survey
Alternative?

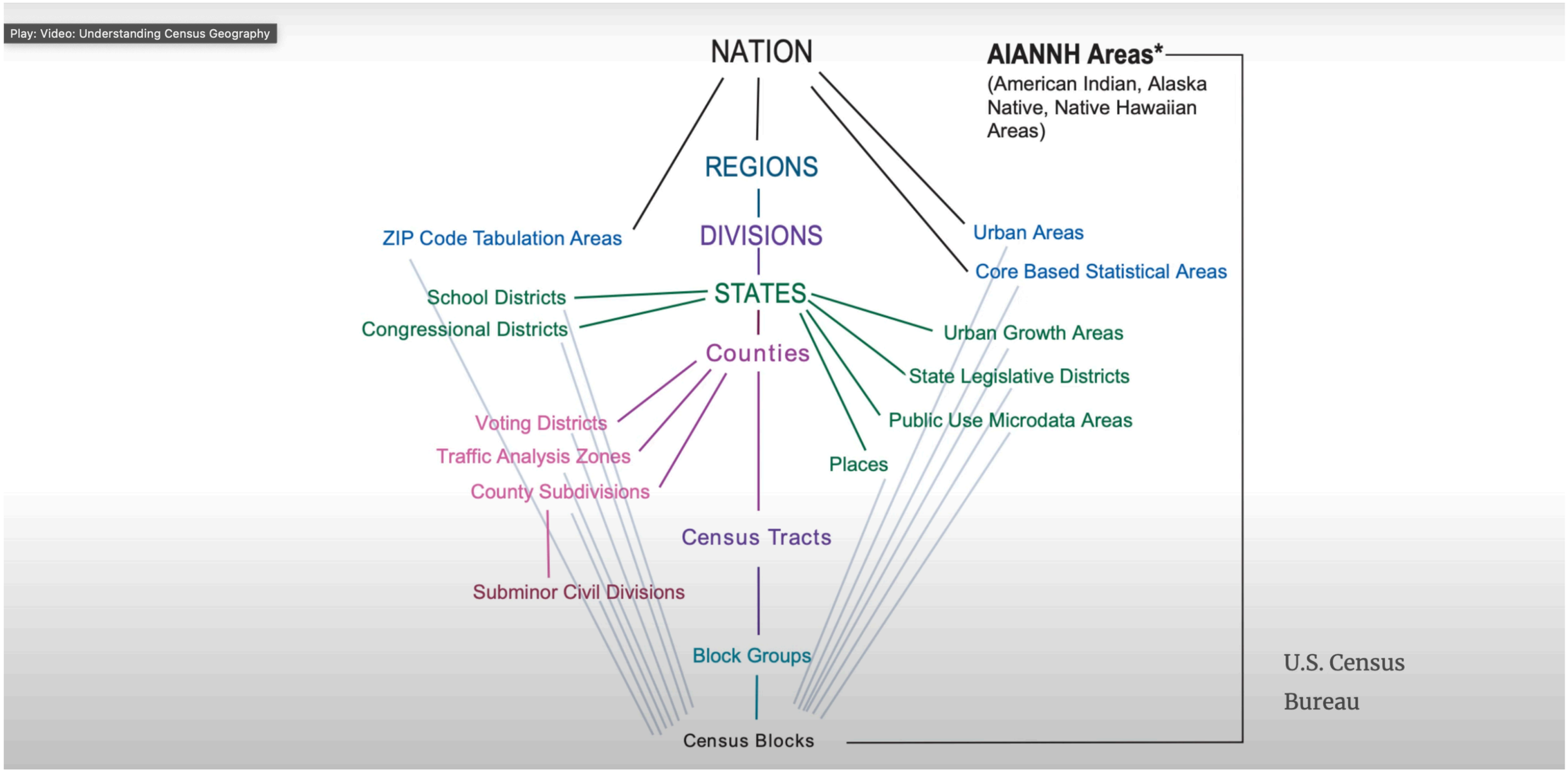
- Pros
  - Every year - fast
  - Cheap
  - Easy


GOOD
EXPENSIVE + GOOD
SLOW + GOOD
NOT POSSIBLE
FAST
LOW QUALITY
CHEAP

Solution

- Break into 2
  - ACS 1
    - Released every year - based on surveys conducted that year
    - Smaller sample - does not cover all geographies (only ~65k+)
  - ACS 5
    - Blended sample of last 5 years - Not timely
    - Covers all geographies - complete

MARGIN OF ERROR



Census geographies (source:censusreporter.org)

`tidycensus` + `tigris`

- Two packages that allow us to get a lot of acs and census data (NOT all data! See `censusapi` package https://www.hrecht.com/censusapi/)
- How to start use?
  - Get API Key - https://api.census.gov/data/key_signup.html
  - Figure out what variables you want
  - Figure out what geographies you want

Can AI solve a problem?

- Started here — https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker
- Now here — https://resobscura.substack.com/p/the-leading-ai-models-are-now-very
- Fast moving field - deepseek?

What does it mean for AI to "solve" a problem?

What problems is it good at solving?

- Narrow vs. General AI
- Disease
- Drug Discovery - alphafold
- Image recognition
- Personalized Recommendations
- Algorithmic Trading

NOT THE SAME AXIS AS HUMANS

Deep and narrow vs shallow and broad

Examples

- Counting - tokenization
- Image description - units of novelty
- Facial Recognition - Sensitivity vs specificity
- Image generation - example
- Chatbots - human like?

Should AI solve a problem?

What should be the goal of trying to get AI to solve problems?

Important questions

- Efficiency/productivity
- Creativity
- External validity
- Black box
- Who is ultimately accountable?
  - Is it AI? Programmers? The user who created the initial data?
  - Eg. Death penalty

Important questions

- Who is ultimately accountable?
- Is it AI? Programmers? The user who created the initial data?
- Why does this matter?