Retiring Adult: New Datasets for Fair Machine Learning

John Miller

Joint work with Frances Ding, Moritz Hardt, Ludwig Schmidt

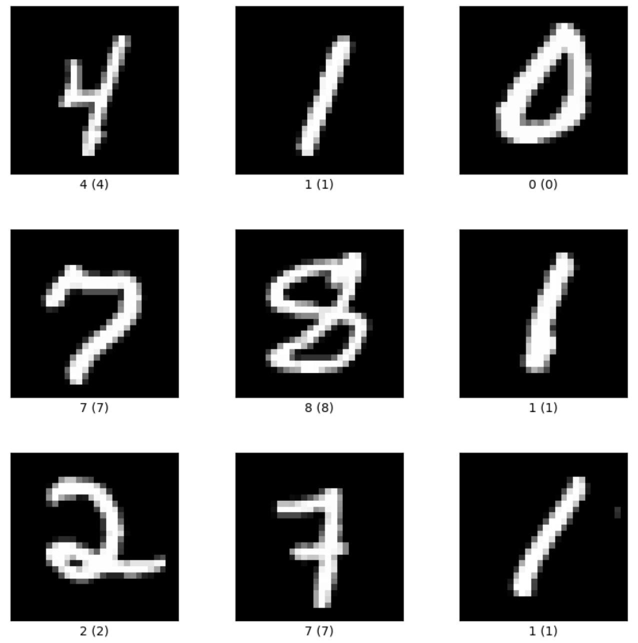
Berkeley
UNIVERSITY OF CALIFORNIA

https://folktables.org/

TOYOTA
RESEARCH INSTITUTE

Datasets as benchmarks

IMAGENET



UCI Machine Learning Repository

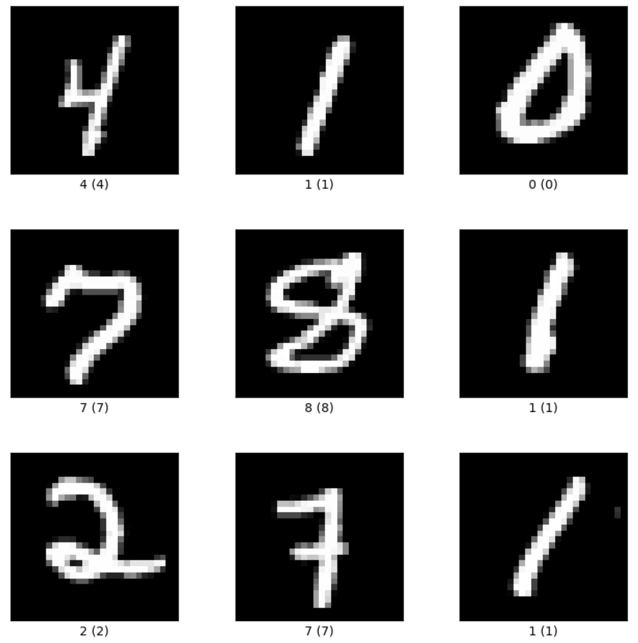
Center for Machine Learning and Intelligent Systems

kaggle

Datasets as benchmarks

- Common training and test sets for model builders

IMAGENET



UCI

Machine Learning Repository

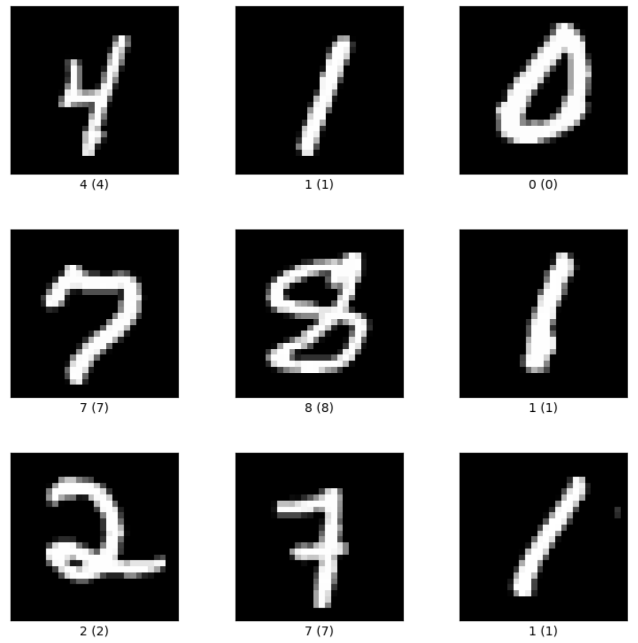
Center for Machine Learning and Intelligent Systems

kaggle

Datasets as benchmarks

- Common training and test sets for model builders
- Datasets also:

IMAGENET



UCI Machine Learning Repository

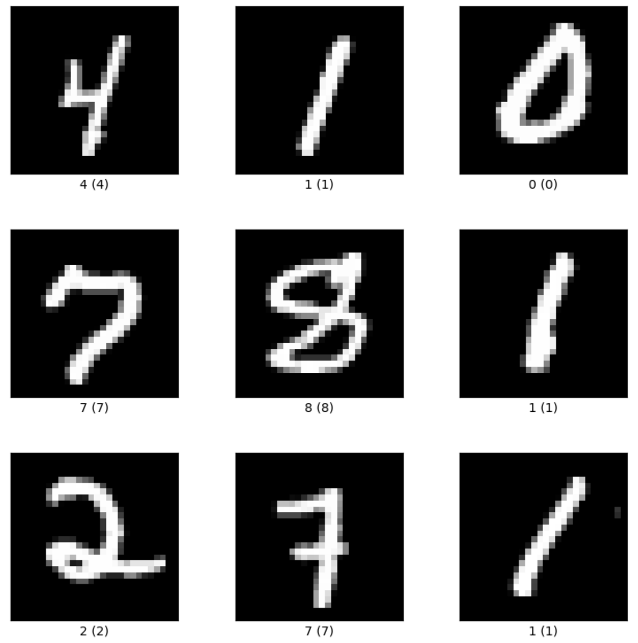
Center for Machine Learning and Intelligent Systems

kaggle

Datasets as benchmarks

- Common training and test sets for model builders
- Datasets also:
  - Formulate problems

IMAGENET



UCI Machine Learning Repository

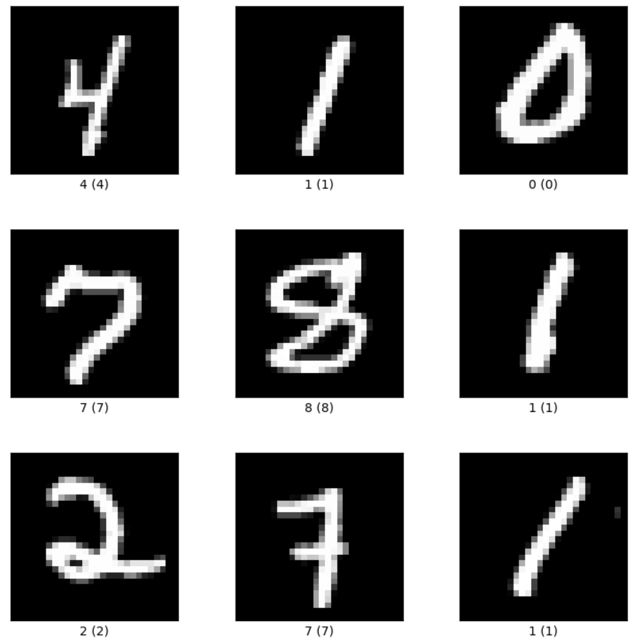
Center for Machine Learning and Intelligent Systems

kaggle

Datasets as benchmarks

- Common training and test sets for model builders
- Datasets also:
  - Formulate problems
  - Organize research communities

IMAGENET



UCI Machine Learning Repository

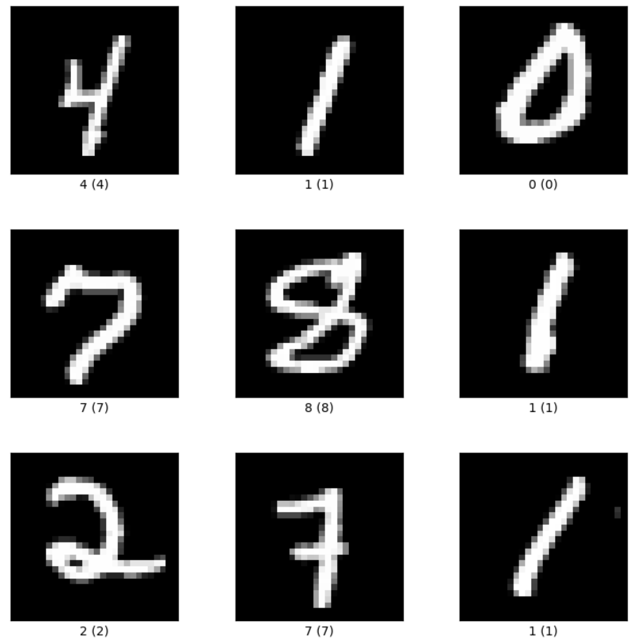
Center for Machine Learning and Intelligent Systems

kaggle

Datasets as benchmarks

- Common training and test sets for model builders
- Datasets also:
  - Formulate problems
  - Organize research communities
  - Serve as an interface between academia and industry

IMAGENET



UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

kaggle

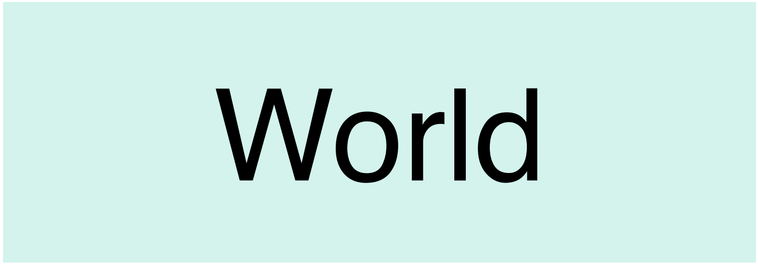
Fairness in machine learning

Fairness in machine learning

- Very active research area

Fairness in machine learning

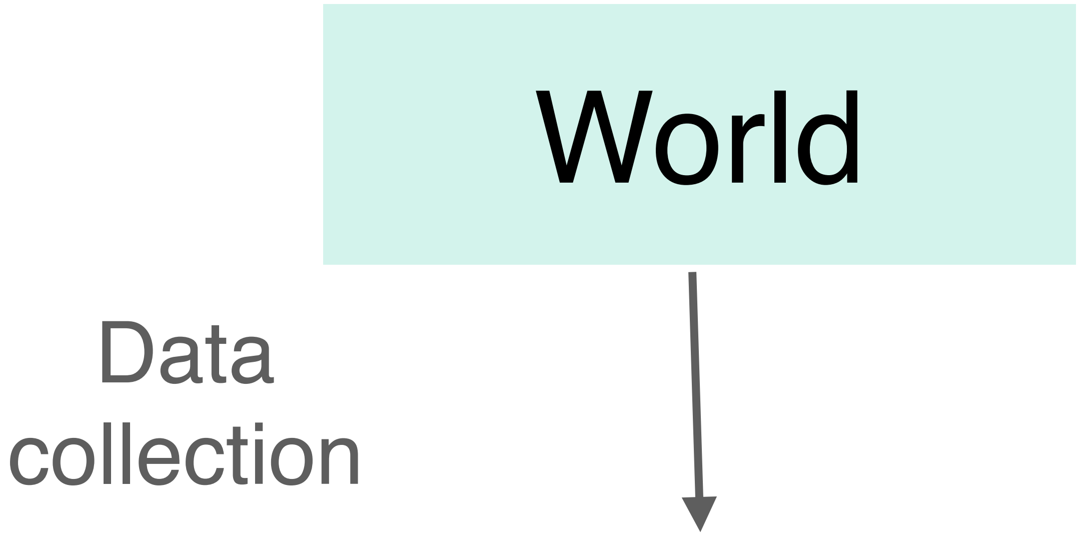
- Very active research area



World

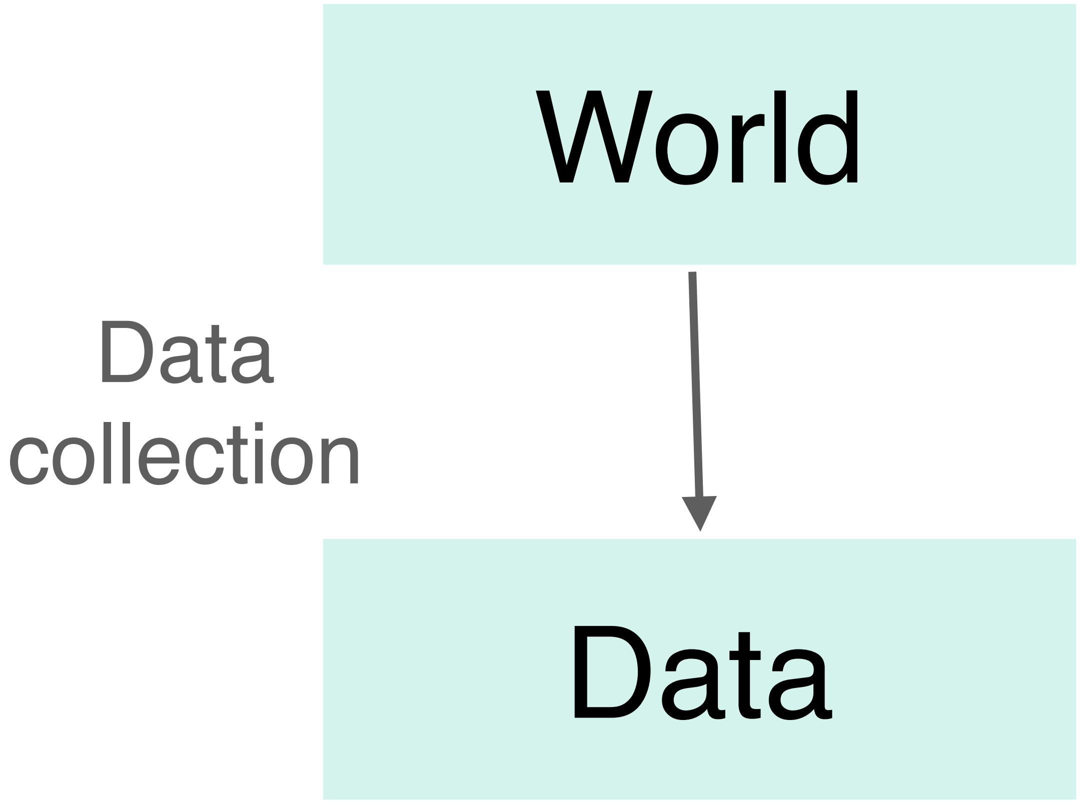
Fairness in machine learning

- Very active research area



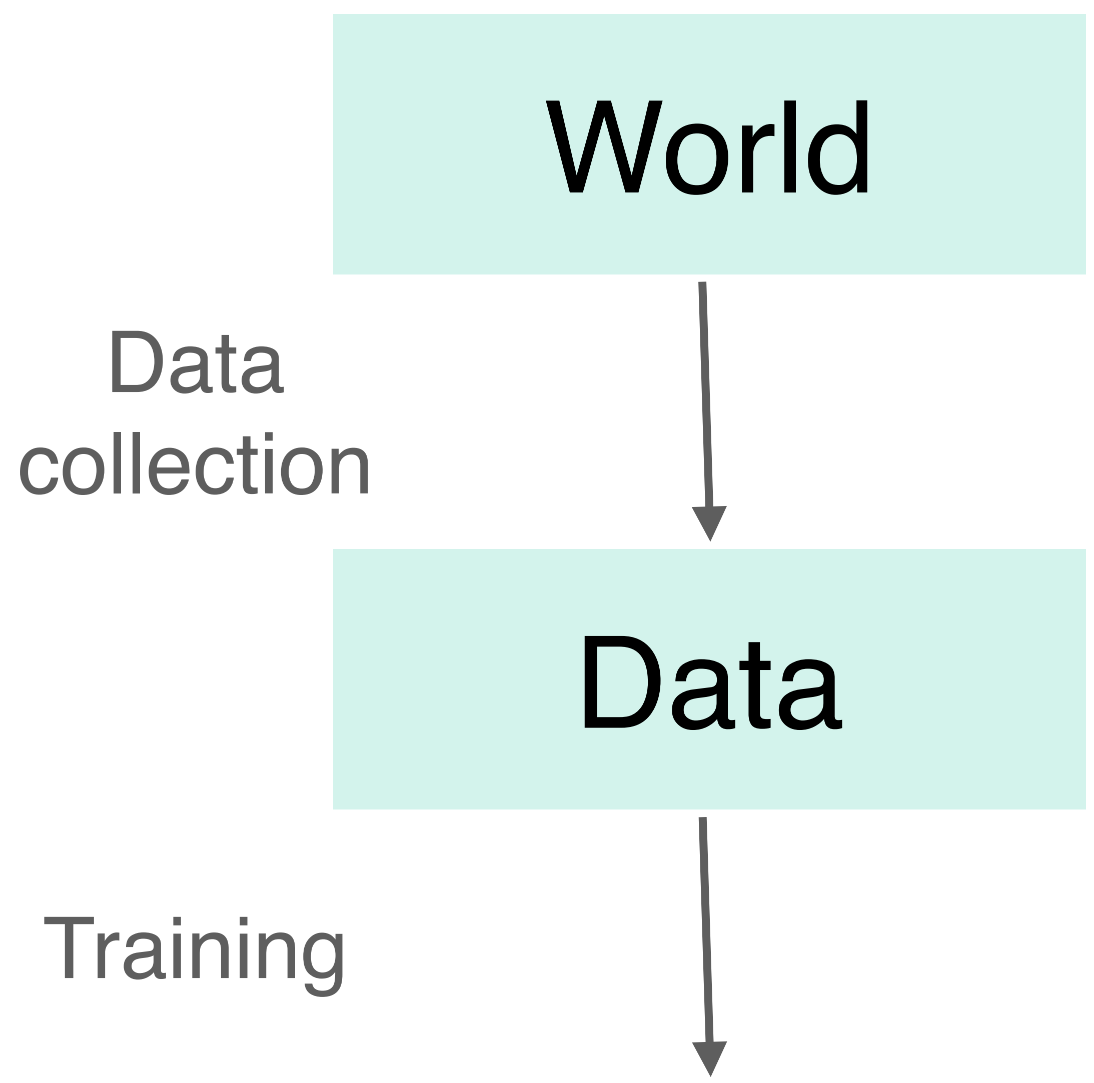
Fairness in machine learning

- Very active research area



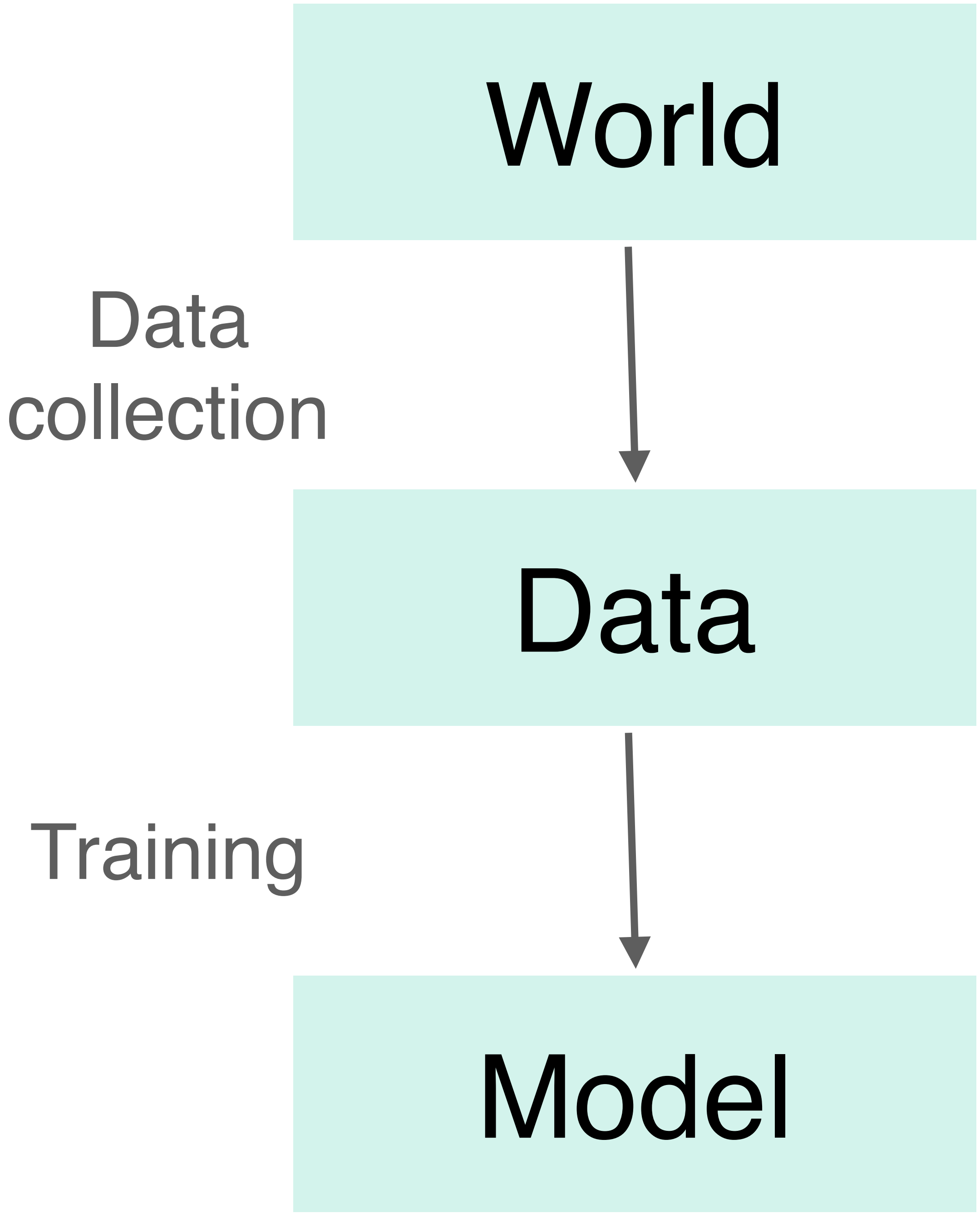
Fairness in machine learning

- Very active research area



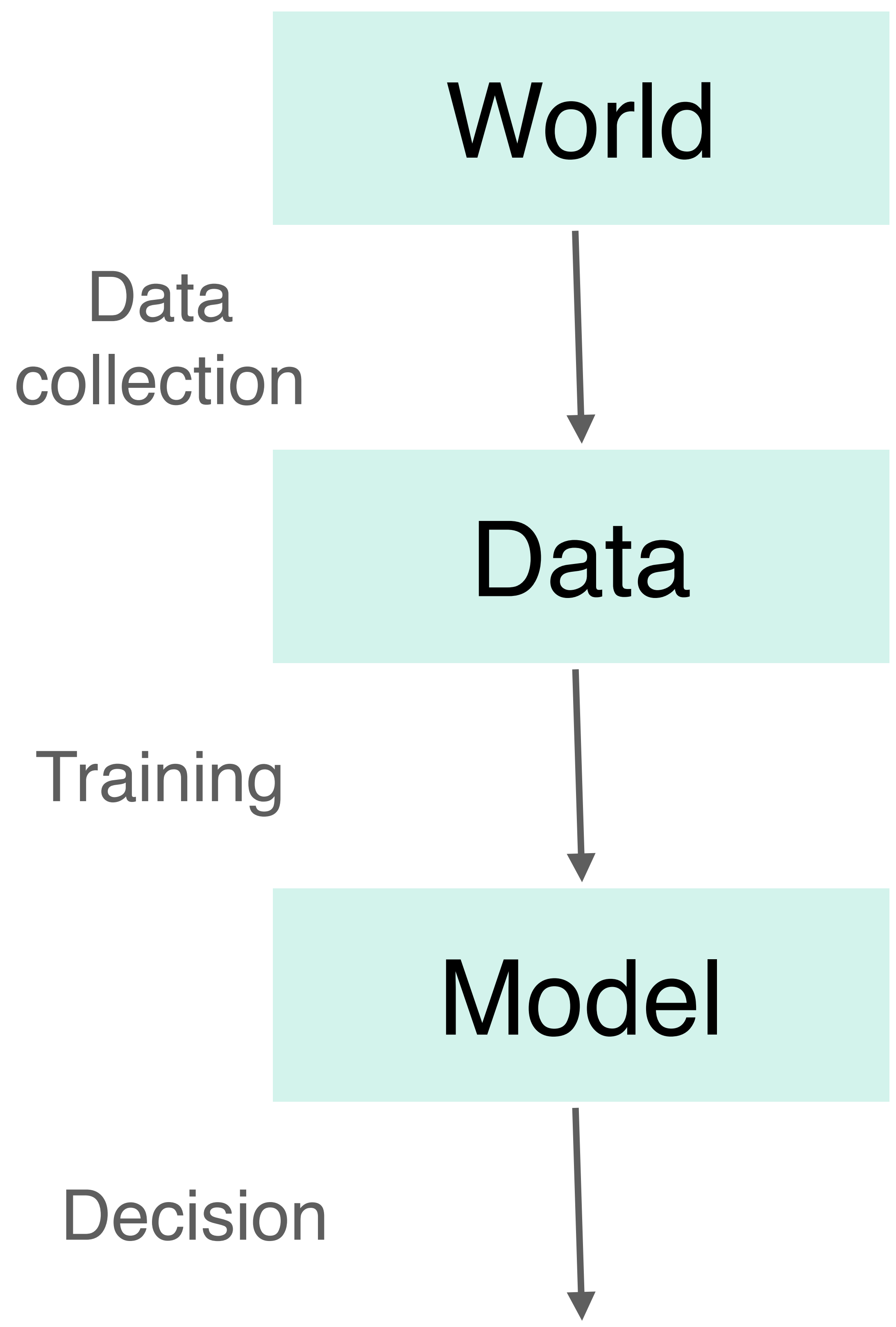
Fairness in machine learning

- Very active research area



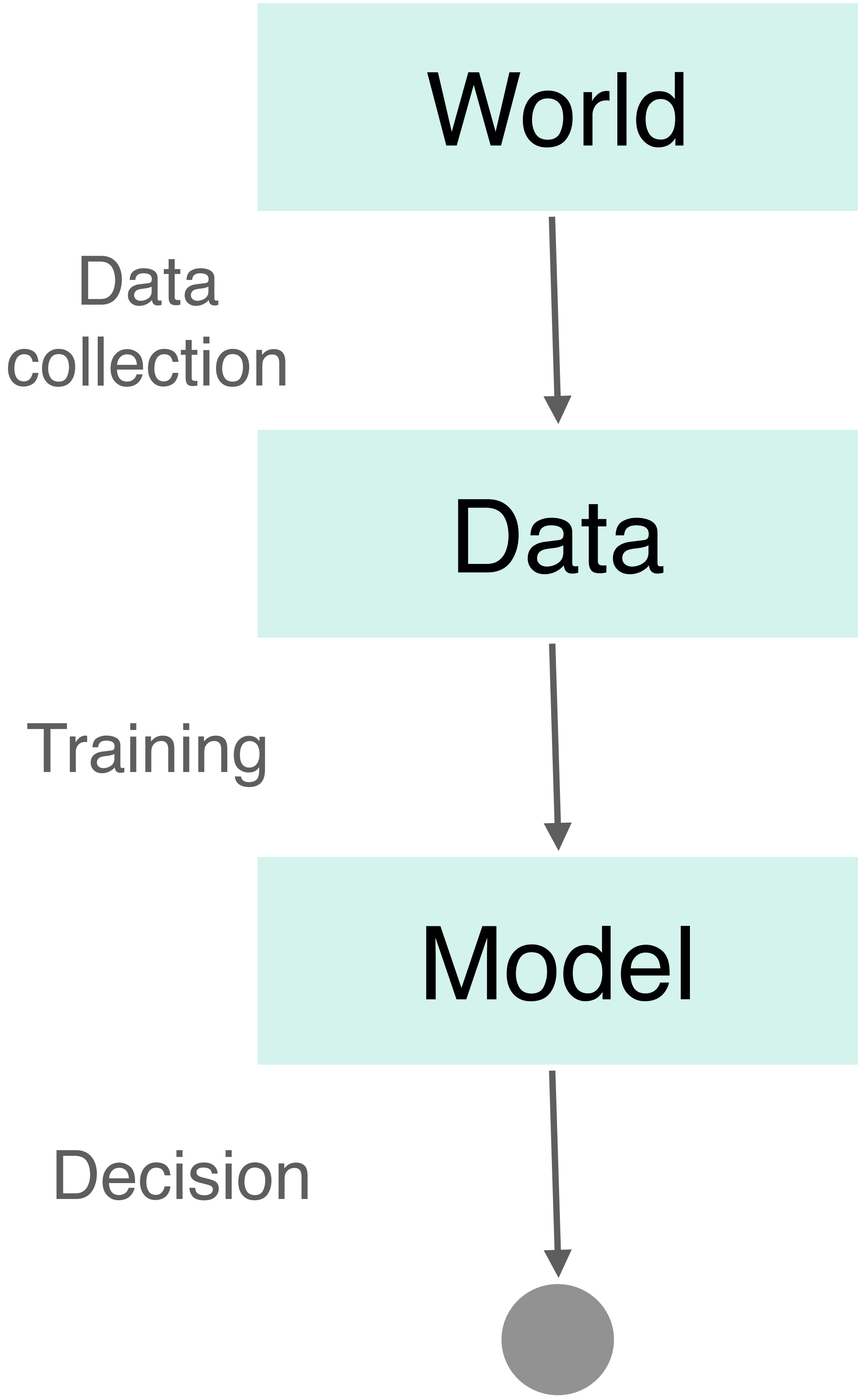
Fairness in machine learning

- Very active research area



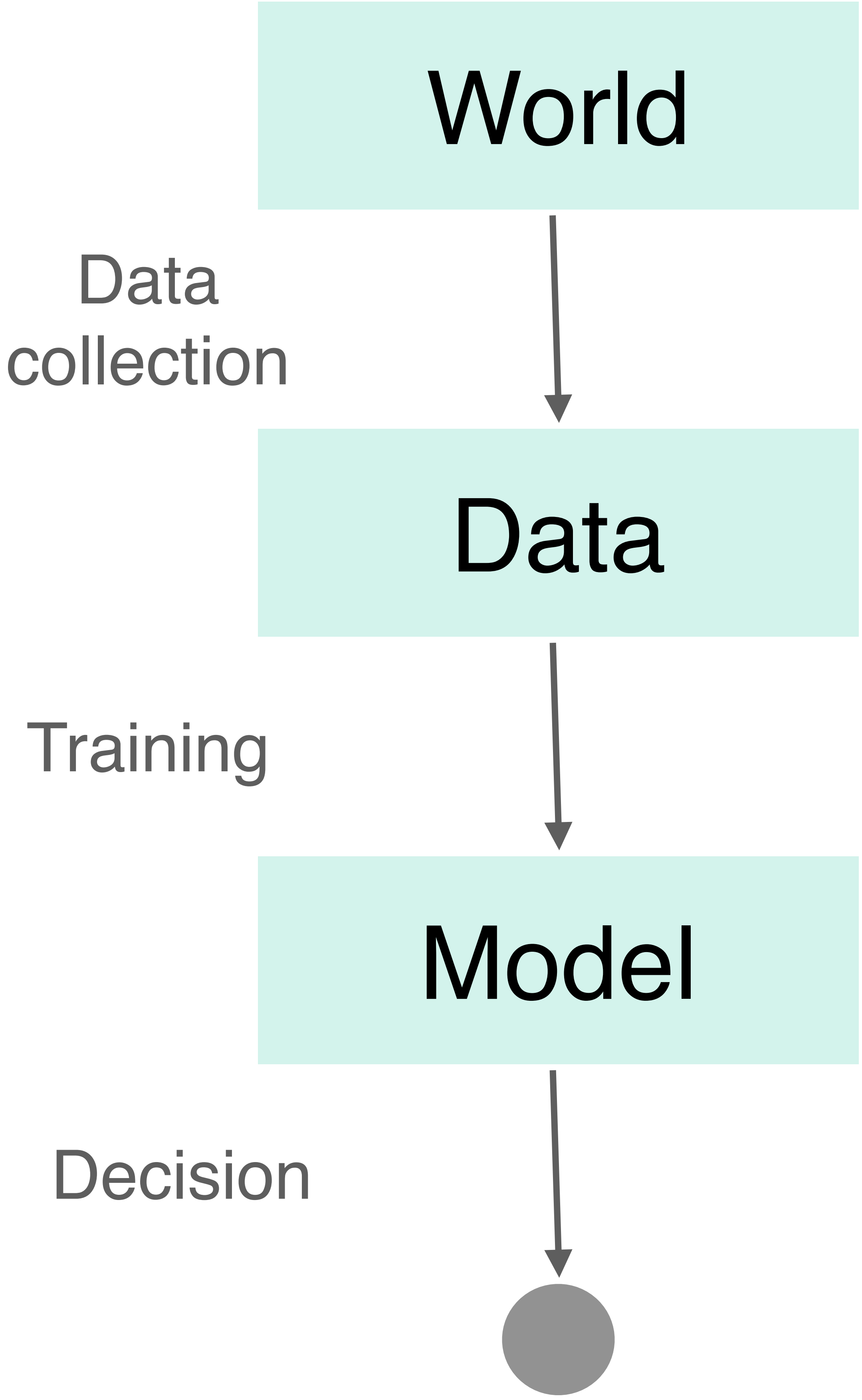
Fairness in machine learning

- Very active research area



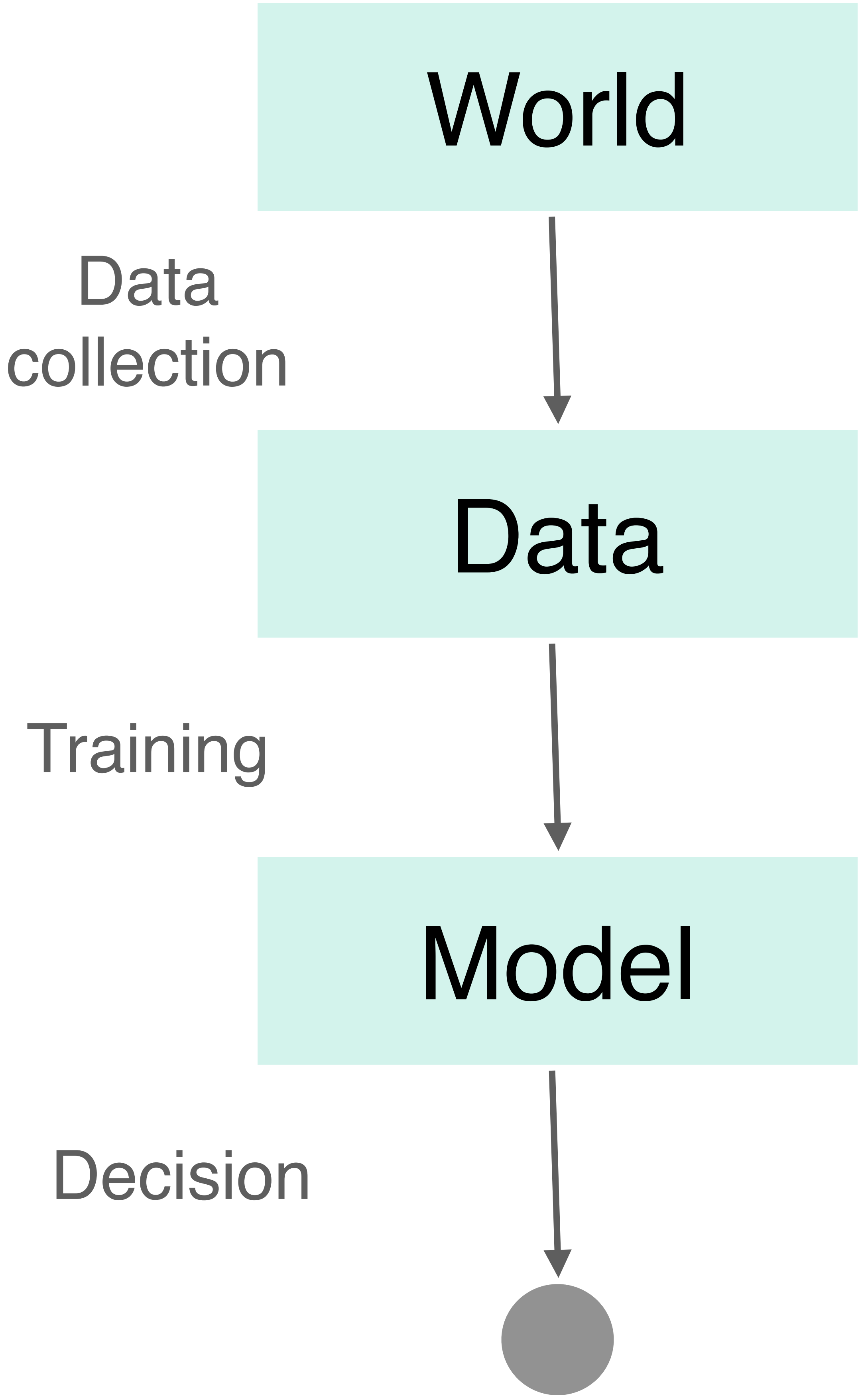
Fairness in machine learning

- Very active research area
- Often study *interventions* to satisfy “fairness criteria” intended to track normative goals



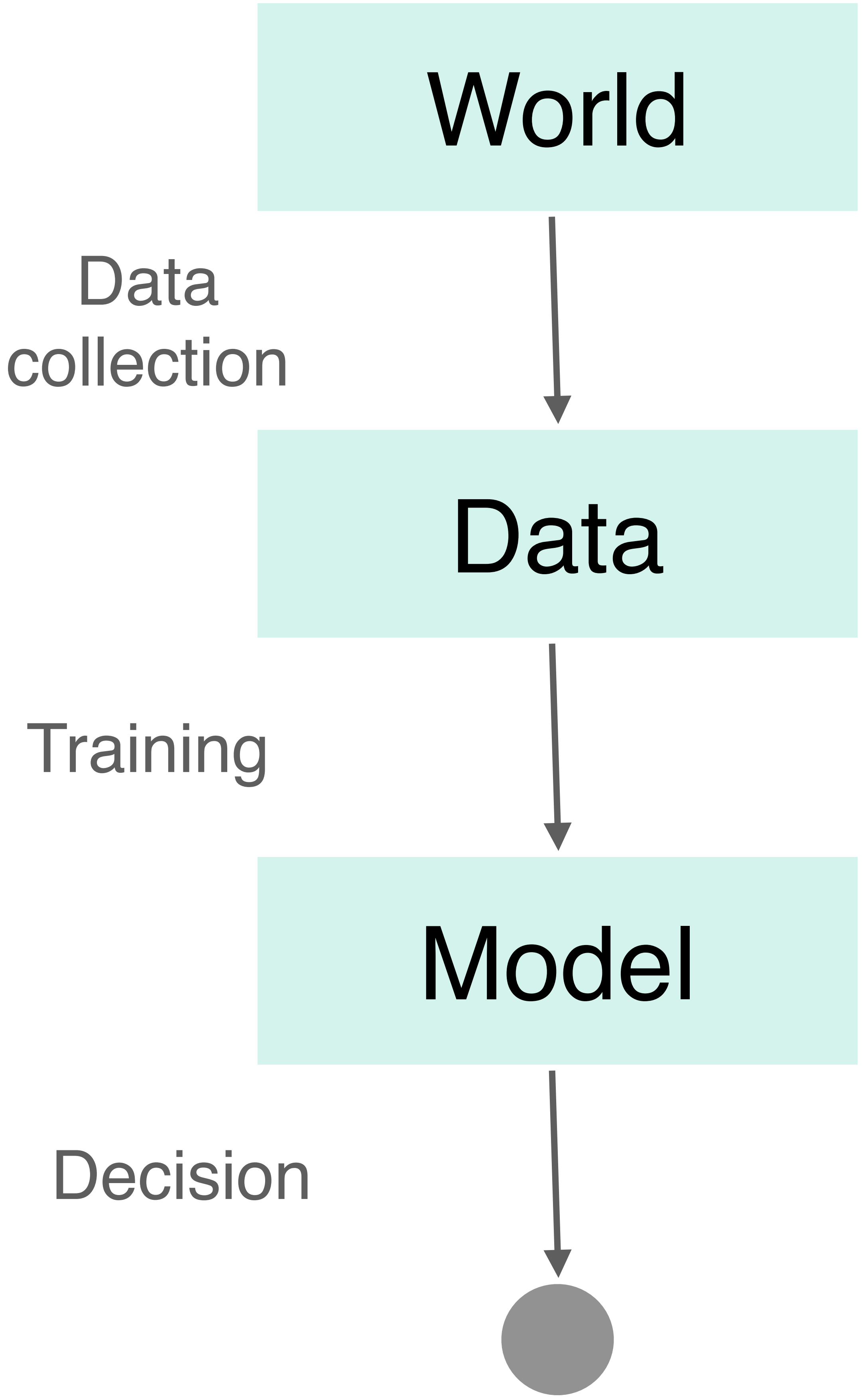
Fairness in machine learning

- Very active research area
- Often study *interventions* to satisfy "fairness criteria" intended to track normative goals
- Field has recognized the importance of datasets [Buolamwini, Gebru (2018); Gebru et al. (2018); Jo, Gebru (2019); Gray, Suri (2019); Prabhu, Birhana (2020)]



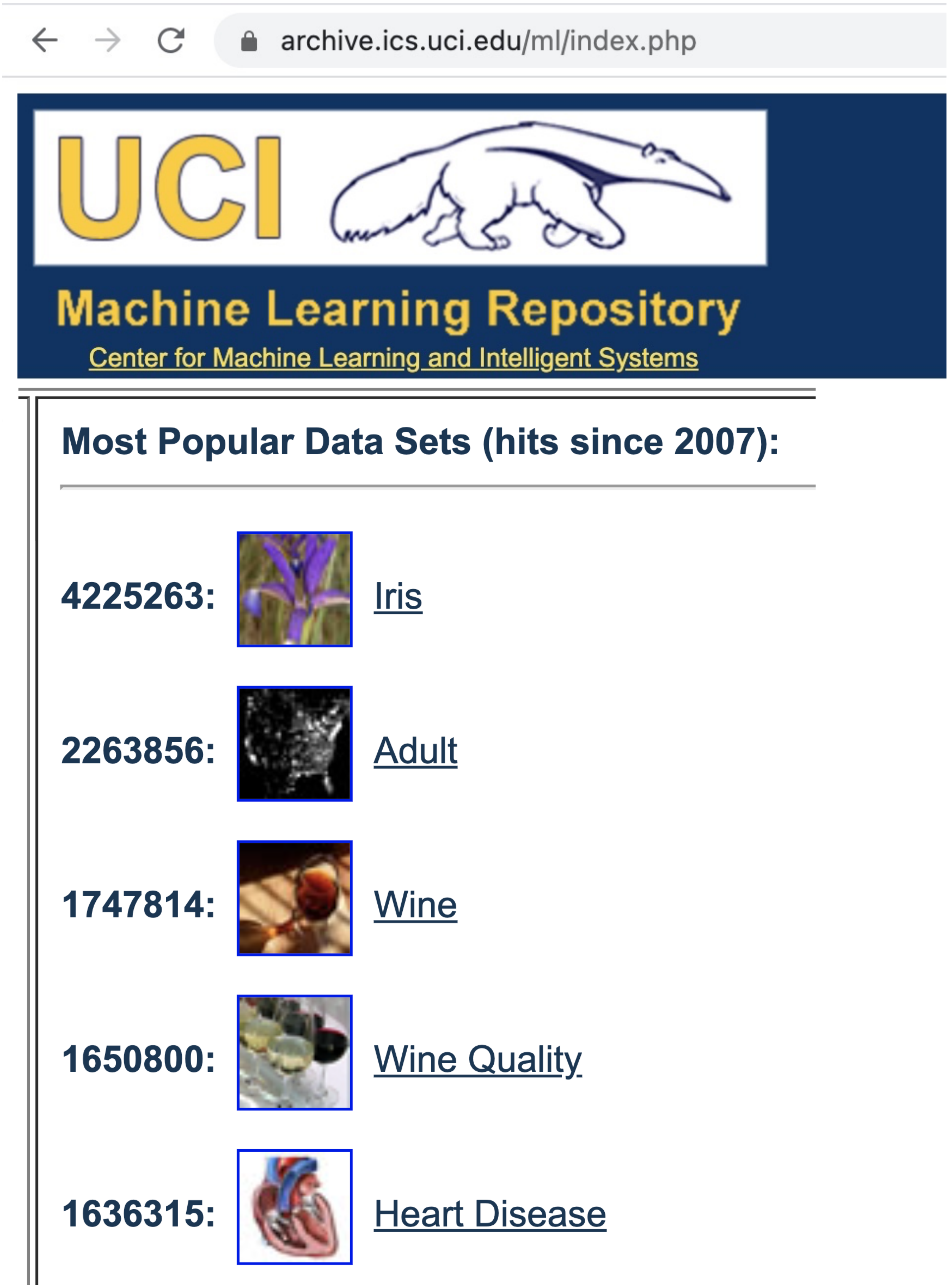
Fairness in machine learning

- Very active research area
- Often study *interventions* to satisfy "fairness criteria" intended to track normative goals
- Field has recognized the importance of datasets [Buolamwini, Gebru (2018); Gebru et al. (2018); Jo, Gebru (2019); Gray, Suri (2019); Prabhu, Birhana (2020)]
- When it comes to tabular data, most papers focus on **UCI Adult**



UCI Adult dataset

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

UCI

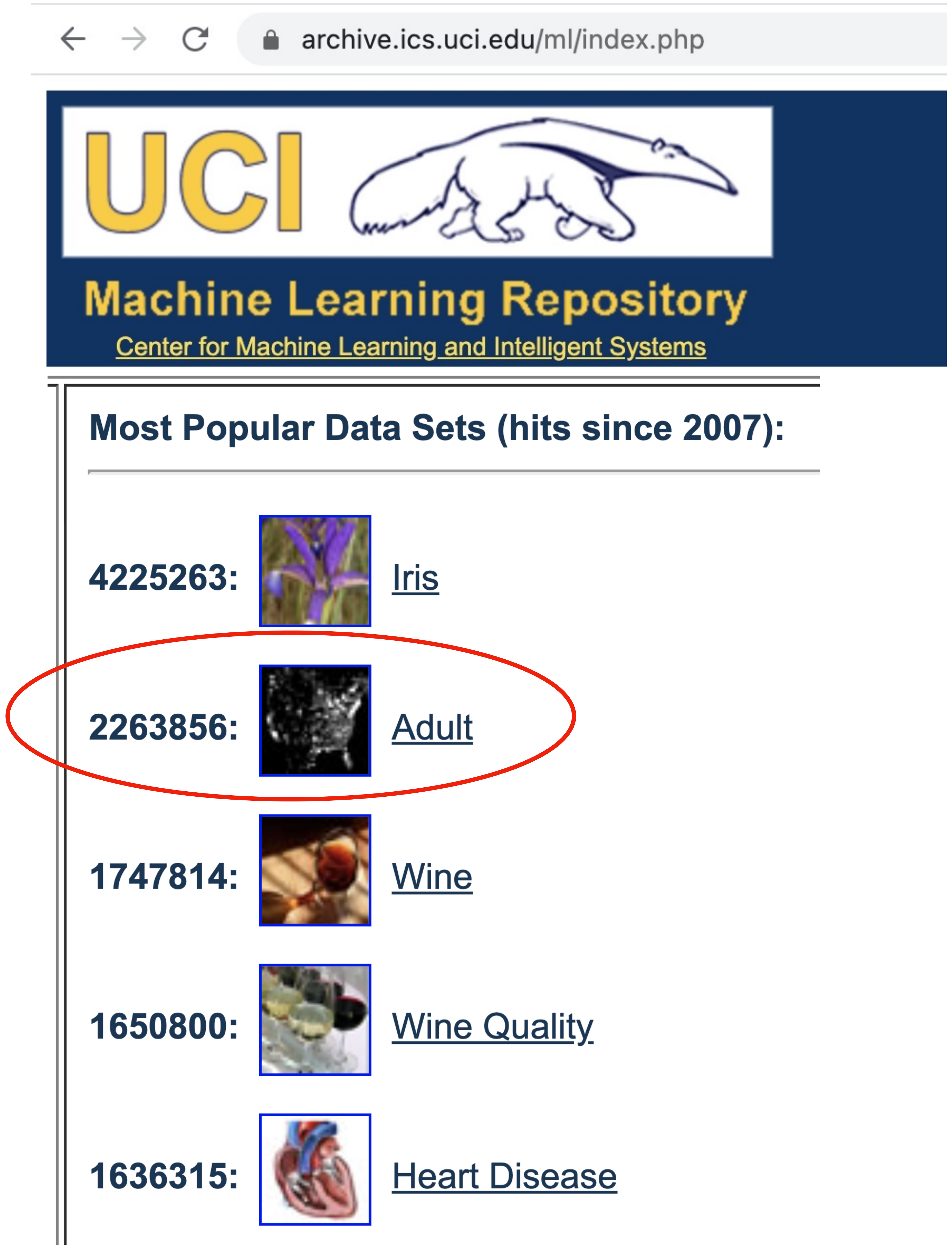
Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263:	Iris
2263856:	Adult
1747814:	Wine
1650800:	Wine Quality
1636315:	Heart Disease

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

UCI

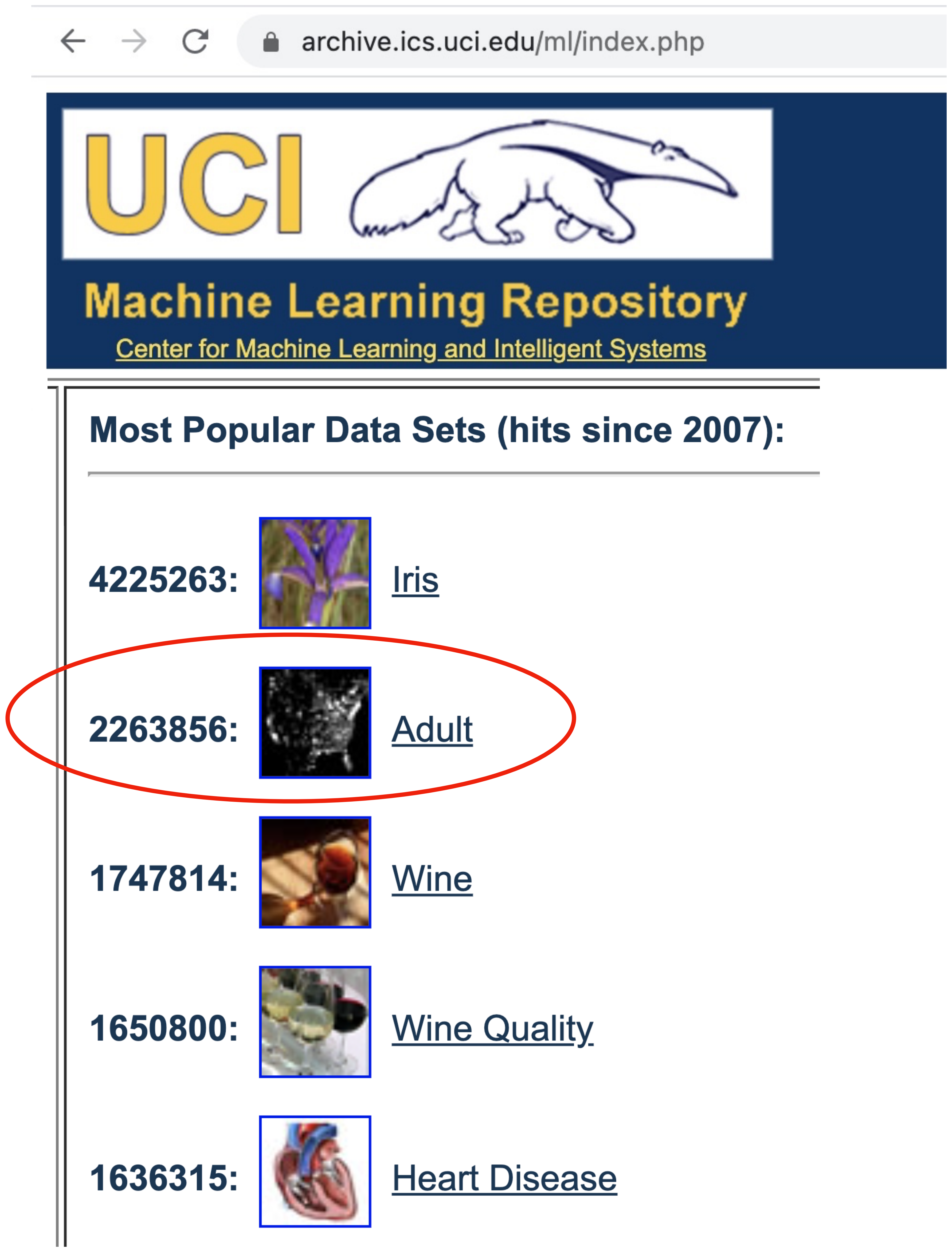
Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263:	Iris
2263856:	Adult
1747814:	Wine
1650800:	Wine Quality
1636315:	Heart Disease

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

UCI

Machine Learning Repository

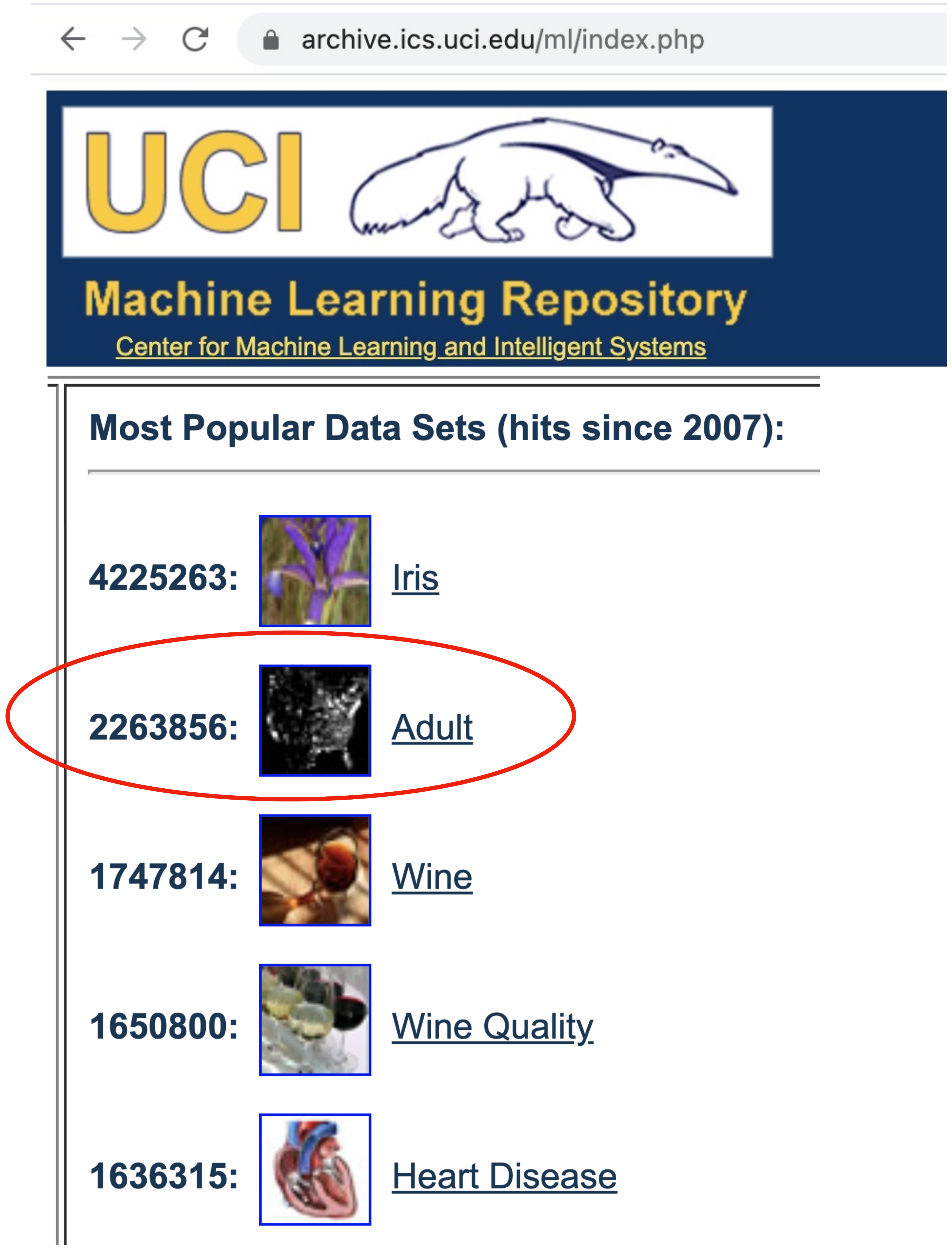
Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263:	Iris
2263856:	Adult
1747814:	Wine
1650800:	Wine Quality
1636315:	Heart Disease

- Extracted by Barry Becker and Ronny Kohavi from “1994 Census database”

UCI Adult dataset



UCI Machine Learning Repository

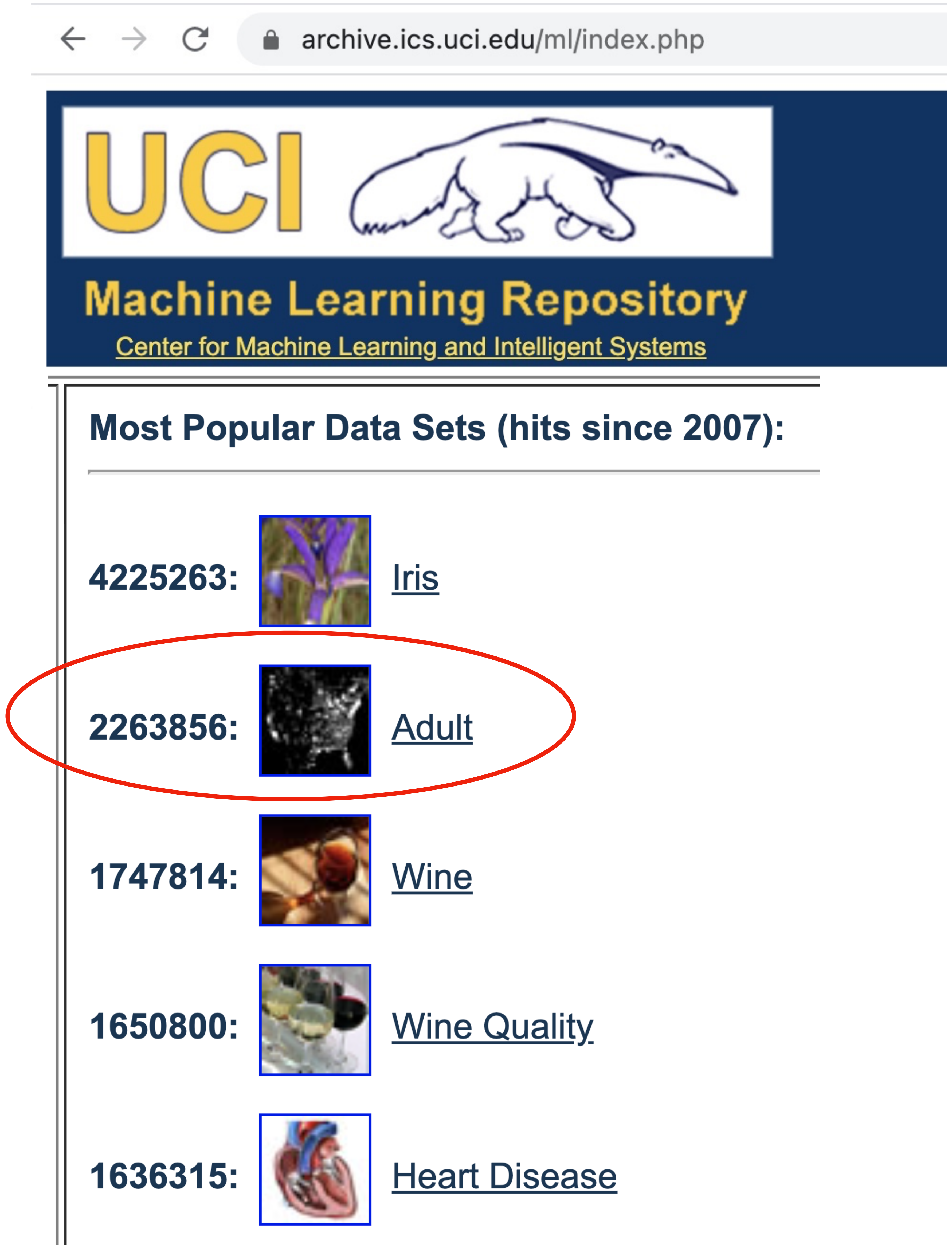
Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263:	Iris
2263856:	Adult
1747814:	Wine
1650800:	Wine Quality
1636315:	Heart Disease

- Extracted by Barry Becker and Ronny Kohavi from “1994 Census database”
- “Prediction task is to determine whether a person makes over $50k a year”

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263: Iris

2263856: Adult

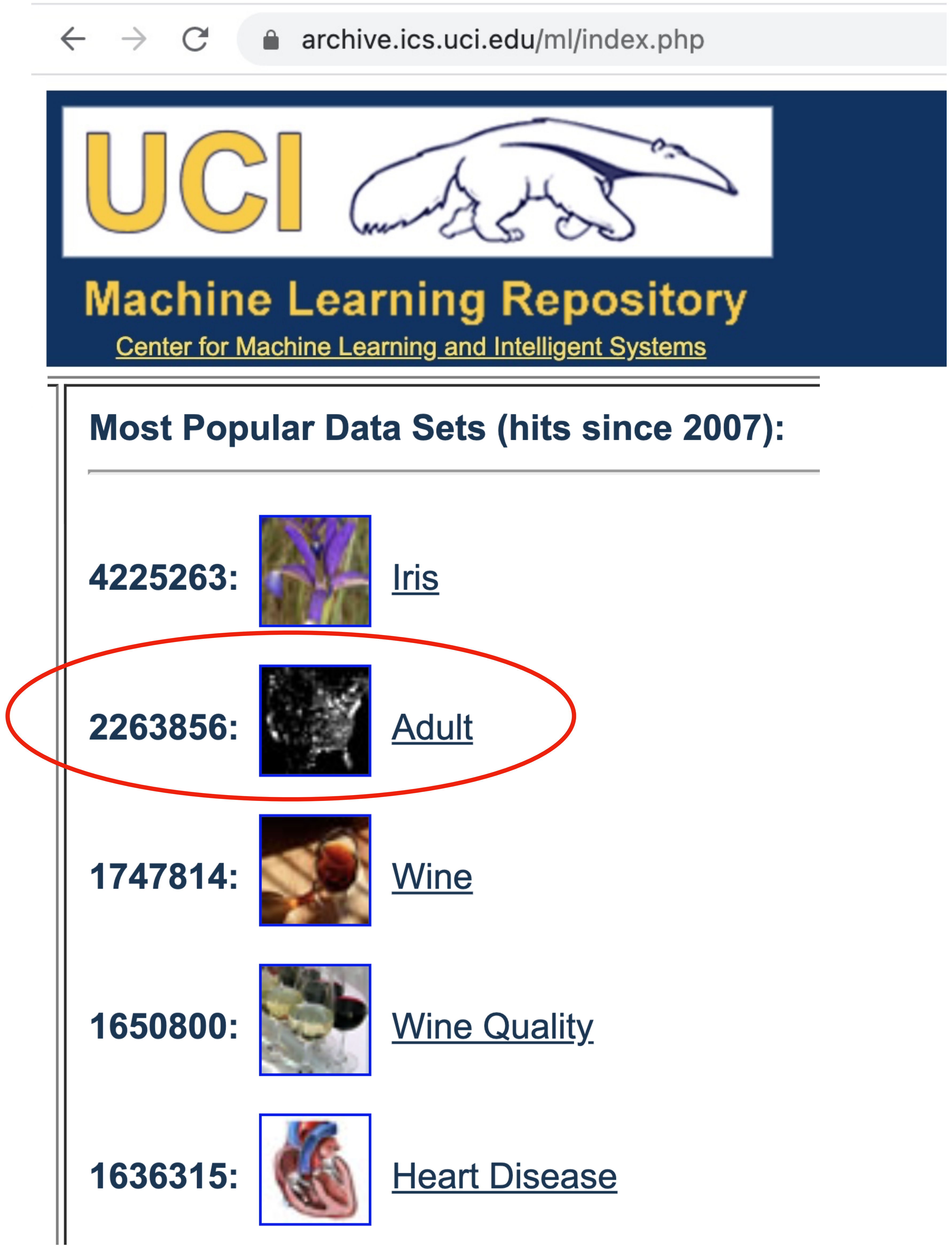
1747814: Wine

1650800: Wine Quality

1636315: Heart Disease

- Extracted by Barry Becker and Ronny Kohavi from “1994 Census database”
- “Prediction task is to determine whether a person makes over $50k a year”
- 14 features available:

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Most Popular Data Sets (hits since 2007):

4225263: Iris

2263856: Adult

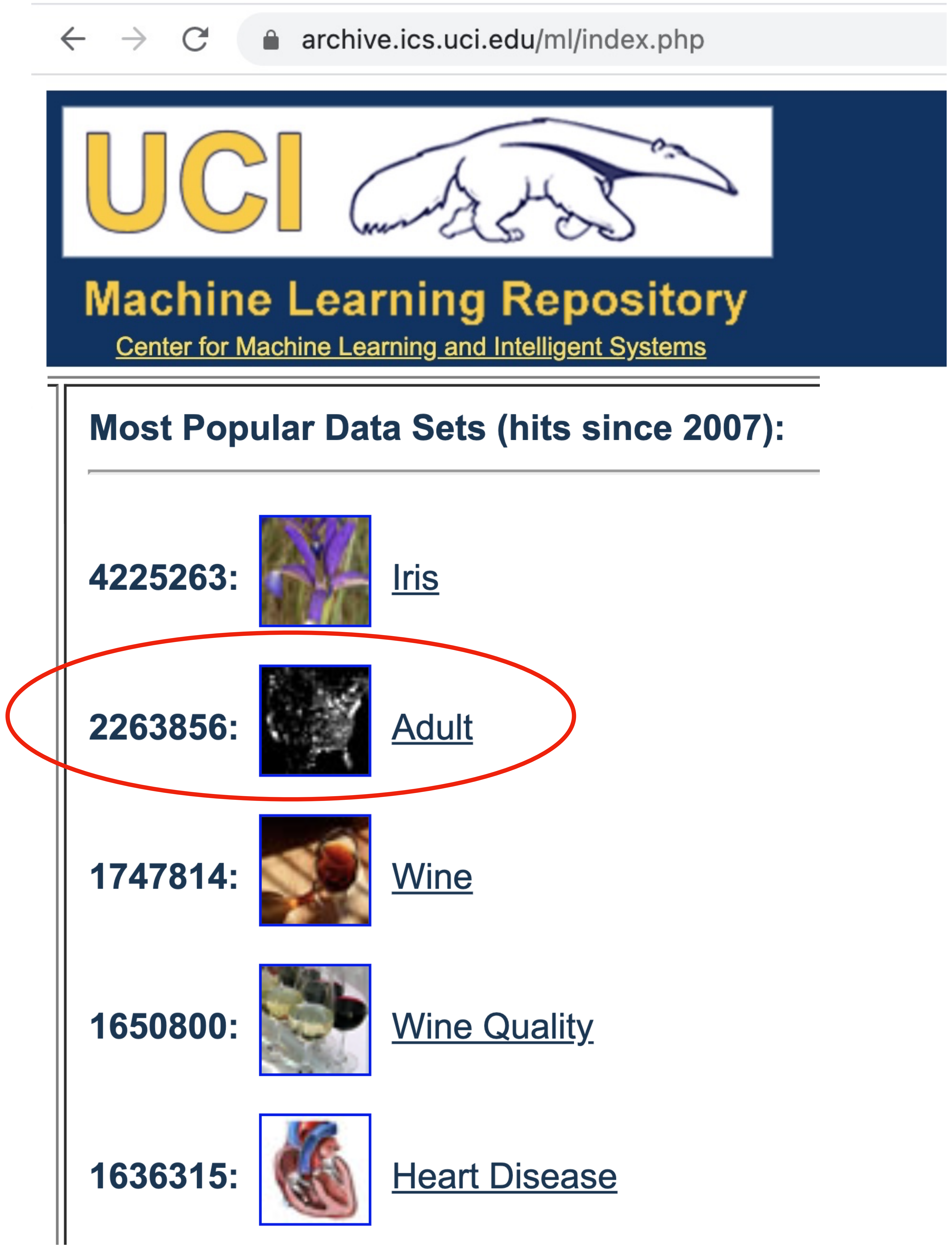
1747814: Wine

1650800: Wine Quality

1636315: Heart Disease

- Extracted by Barry Becker and Ronny Kohavi from “1994 Census database”
- “Prediction task is to determine whether a person makes over $50k a year”
- 14 features available:
  - Age, occupation, education, marital-status, race, sex**,** capital-gain, capital-loss, hours-per-week, native-country, etc

UCI Adult dataset



archive.ics.uci.edu/ml/index.php

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

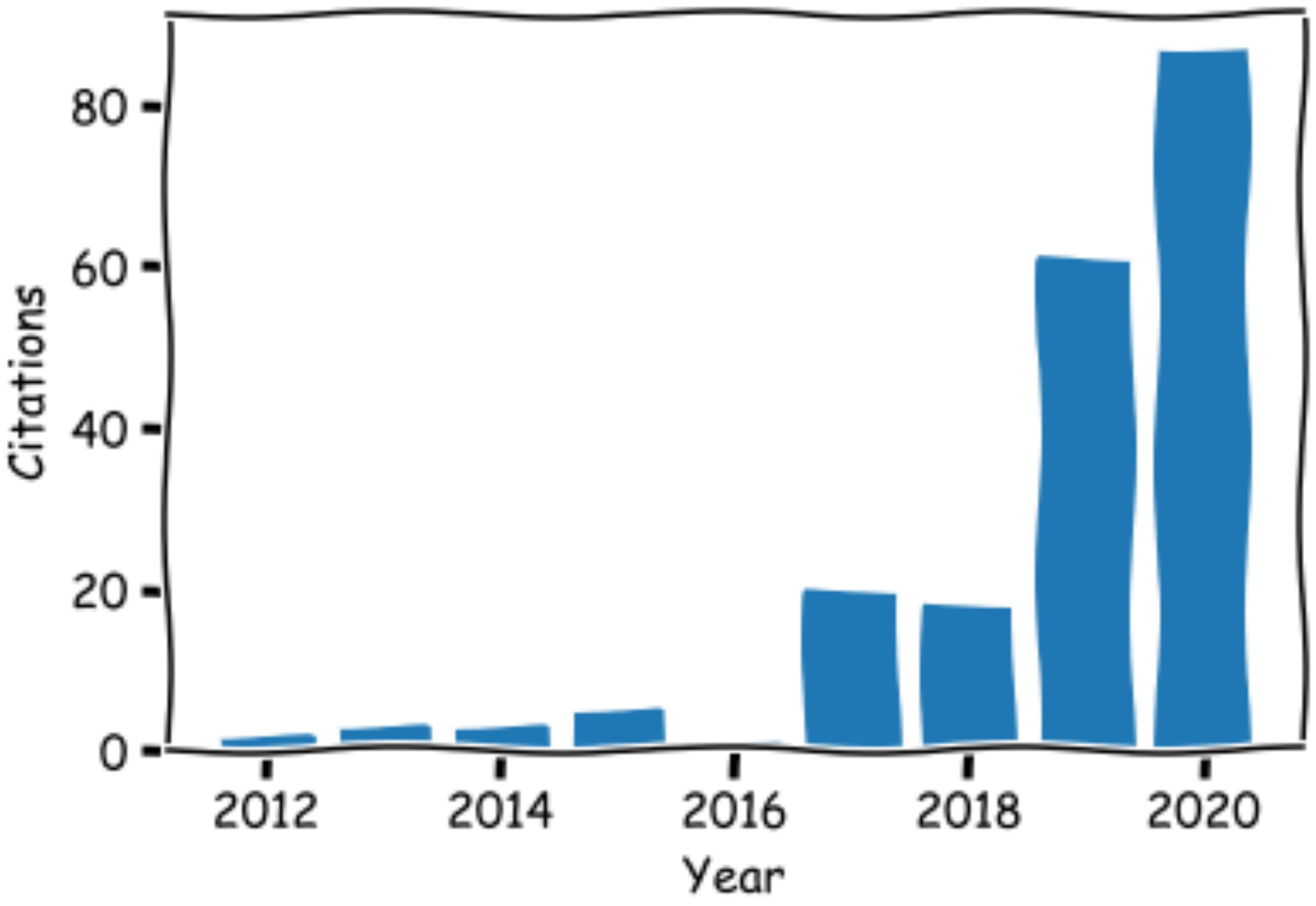
Most Popular Data Sets (hits since 2007):

4225263:	Iris
2263856:	Adult
1747814:	Wine
1650800:	Wine Quality
1636315:	Heart Disease

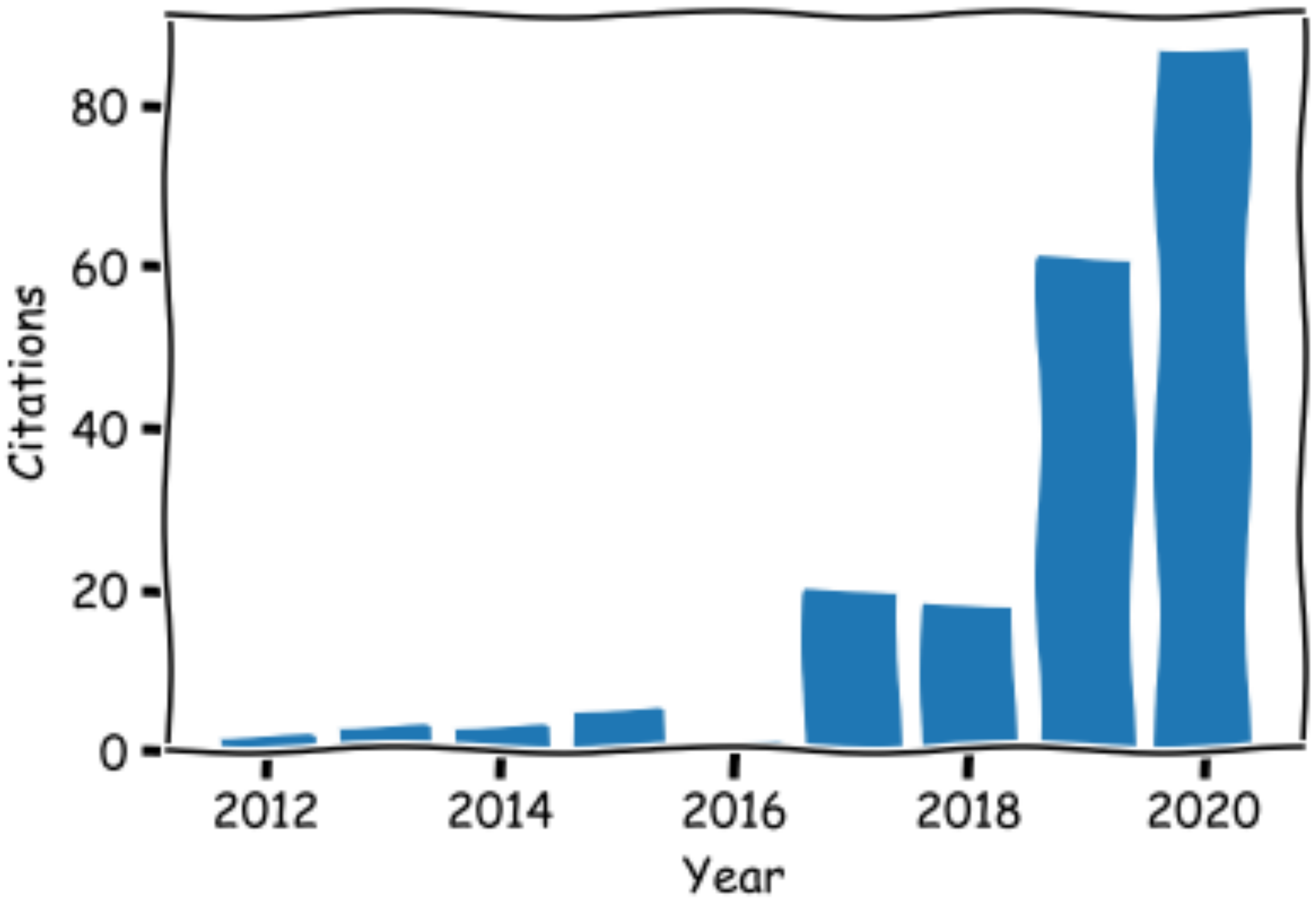
- Extracted by Barry Becker and Ronny Kohavi from "1994 Census database"
- "Prediction task is to determine whether a person makes over $50k a year"
- 14 features available:
  - Age, occupation, education, marital-status, **race, sex,** capital-gain, capital-loss, hours-per-week, native-country, etc

Fairness papers using UCI Adult

Fairness papers using UCI Adult



Fairness papers using UCI Adult



Lower bound: 300+ fairness papers cite UCI Adult

Where does UCI Adult come from?
Is UCI Adult a good empirical foundation for fair ML?

Reconstruction of UCI Adult

United States™ Census Bureau

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data

United States™
Census
Bureau

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data

- **Source**: 1994 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)

United States™ Census Bureau

IPUMS CPS

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data
- **Source**: 1994 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)
    - Attribute names and encodings have changed from UCI Adult description

United States™ Census Bureau

IPUMS CPS

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data
- **Source**: 1994 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)
  - Attribute names and encodings have changed from UCI Adult description
  - Some “native countries” in CPS ASEC are coded as “unknown” in UCI Adult

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data

- **Source**: 1994 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)

  - Attribute names and encodings have changed from UCI Adult description

  - Some “native countries” in CPS ASEC are coded as “unknown” in UCI Adult

- Succeeded in finding all but one column: *fnlwgt*

United States™ Census Bureau

IPUMS CPS

Reconstruction of UCI Adult

- We matched all 48,842 records in UCI Adult to available Census data
- **Source**: 1994 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)
    - Attribute names and encodings have changed from UCI Adult description
    - Some “native countries” in CPS ASEC are coded as “unknown” in UCI Adult
- Succeeded in finding all but one column: *fnlwgt*
- We recover the **raw income variable**

Good aspects of UCI Adult

Good aspects of UCI Adult

- Established survey and data collection standards by the Census Bureau

Good aspects of UCI Adult

- Established survey and data collection standards by the Census Bureau

- Census data arguably avoids some known issues raised with other ML datasets

Good aspects of UCI Adult

- Established survey and data collection standards by the Census Bureau

- Census data arguably avoids some known issues raised with other ML datasets

    - Quality sampling frame capturing the US population

Good aspects of UCI Adult

- Established survey and data collection standards by the Census Bureau

- Census data arguably avoids some known issues raised with other ML datasets

    - Quality sampling frame capturing the US population

    - Skilled and compensated labor (in contrast with MTurk data collection)

Good aspects of UCI Adult

- Established survey and data collection standards by the Census Bureau
- Census data arguably avoids some known issues raised with other ML datasets
  - Quality sampling frame capturing the US population
  - Skilled and compensated labor (in contrast with MTurk data collection)
  - Thorough documentation

Issues with UCI Adult: external validity

Issues with UCI Adult: external validity

- 25+ years old

Issues with UCI Adult: external validity

- 25+ years old
    - Census encodings for *race* have changed substantially

Issues with UCI Adult: external validity

- 25+ years old
    - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources

CENSUS ’90

Issues with UCI Adult: external validity

- 25+ years old
    - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
    - State information, many other features *not* included

CENSUS ’90

Issues with UCI Adult: external validity

- 25+ years old
  - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
  - State information, many other features *not* included
- $50k income cut-off is way above the median income

CENSUS ’90

Issues with UCI Adult: external validity

- 25+ years old
  - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
  - State information, many other features *not* included
- $50k income cut-off is way above the median income
  - 76th percentile of income in 1994

CENSUS '90

Issues with UCI Adult: external validity

- 25+ years old
  - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
  - State information, many other features *not* included
- $50k income cut-off is way above the median income
  - 76th percentile of income in 1994
  - 88th percentile in the Black population

CENSUS '90

Issues with UCI Adult: external validity

- 25+ years old
  - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
  - State information, many other features *not* included
- $50k income cut-off is way above the median income
  - 76th percentile of income in 1994
  - 88th percentile in the Black population
  - 89th percentile among women

CENSUS ’90

Issues with UCI Adult: external validity

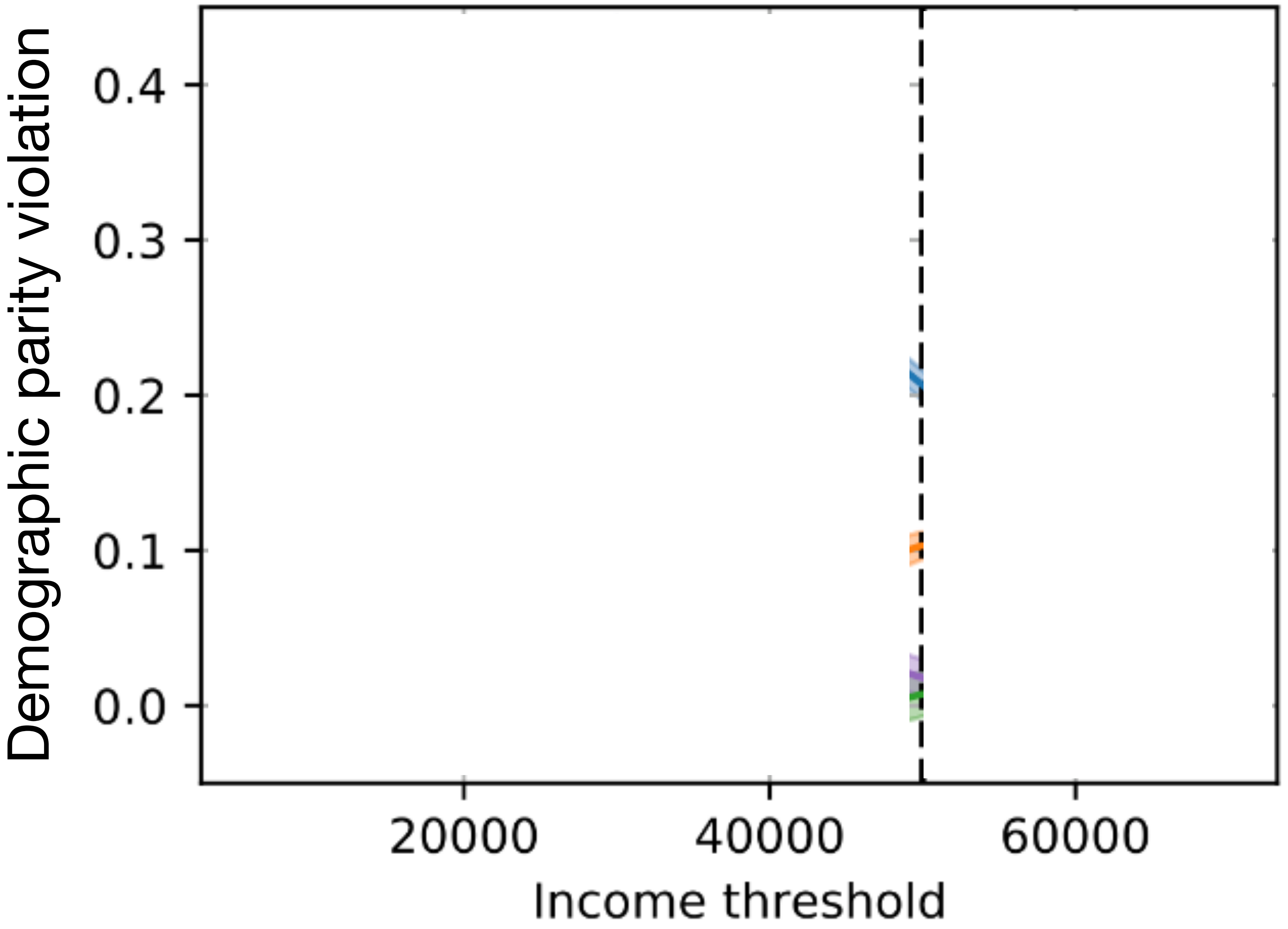
- 25+ years old
  - Census encodings for *race* have changed substantially
- Under-utilizes the wealth of census data sources
  - State information, many other features *not* included
- $50k income cut-off is way above the median income
  - 76th percentile of income in 1994
  - 88th percentile in the Black population
  - 89th percentile among women

CENSUS ’90

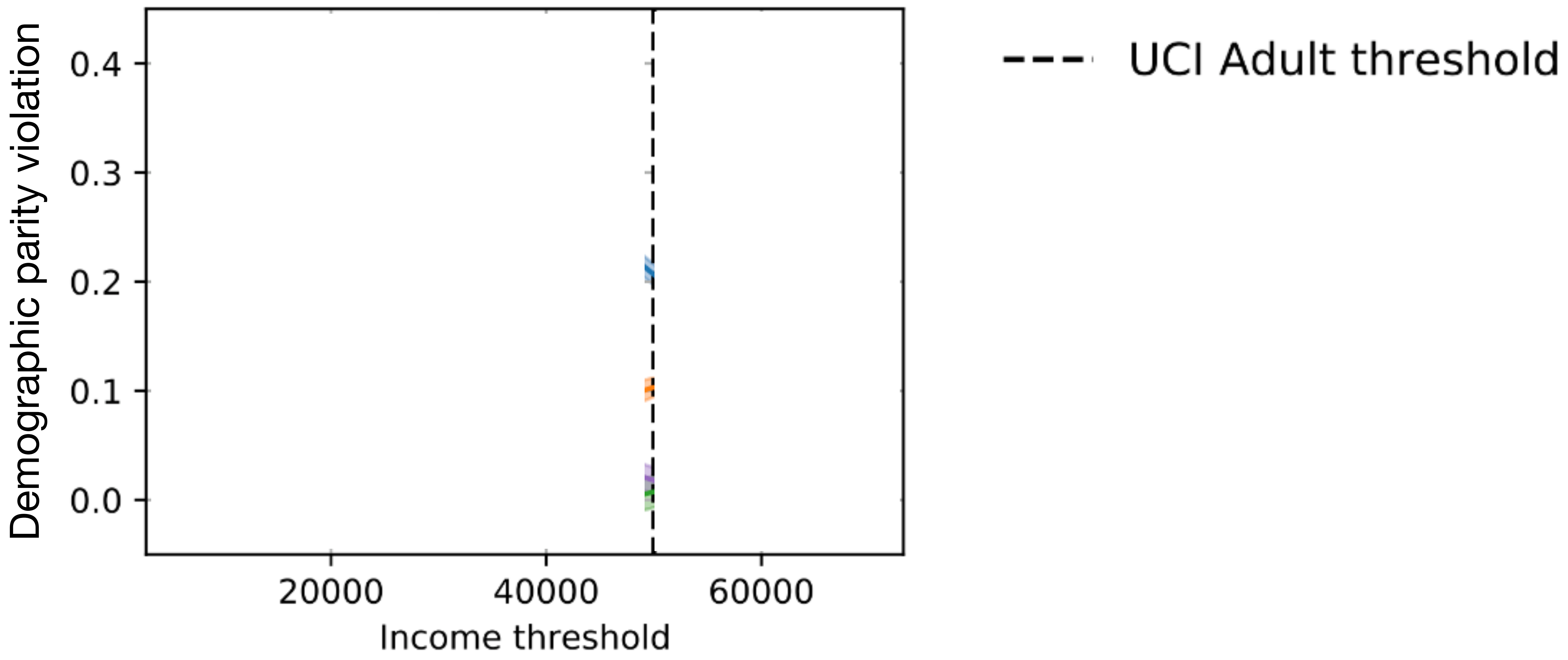
Classifier accuracy is **higher** for individuals in minority groups, the **opposite** setting of much of fair ML

Varying the income threshold

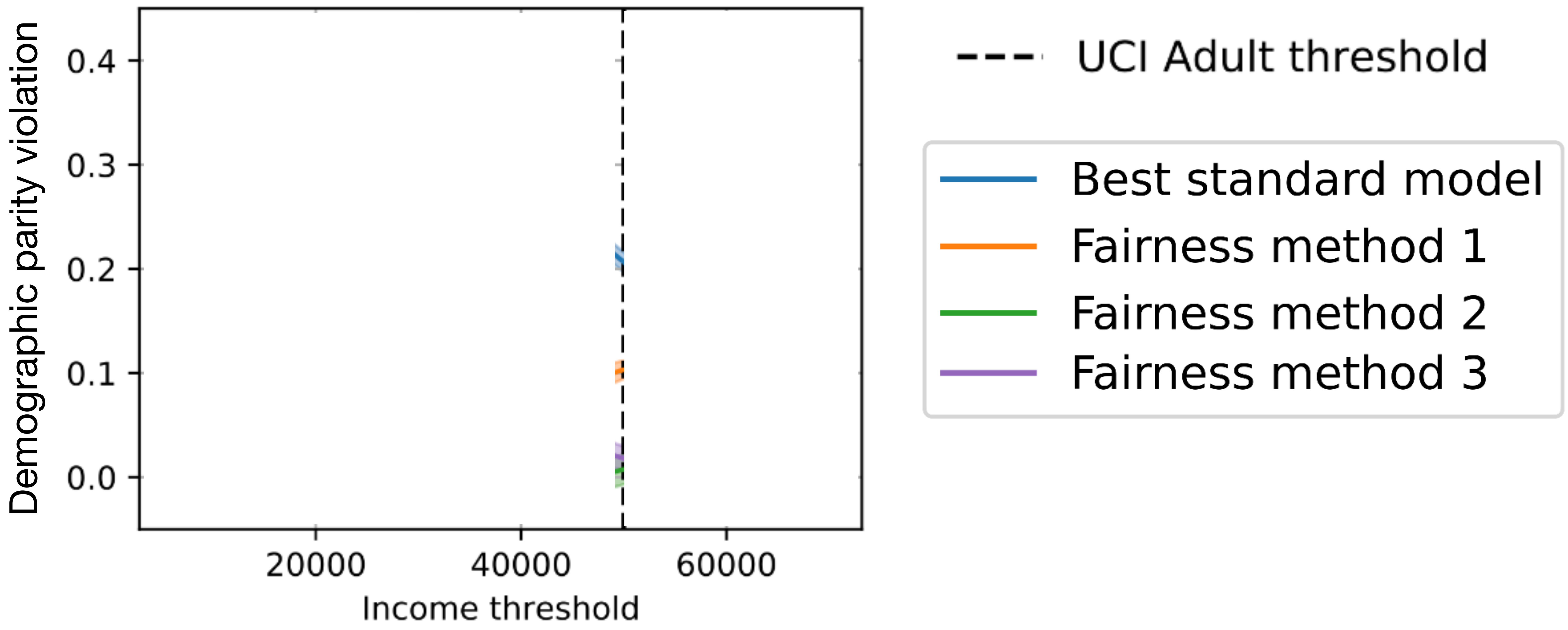
Varying the income threshold



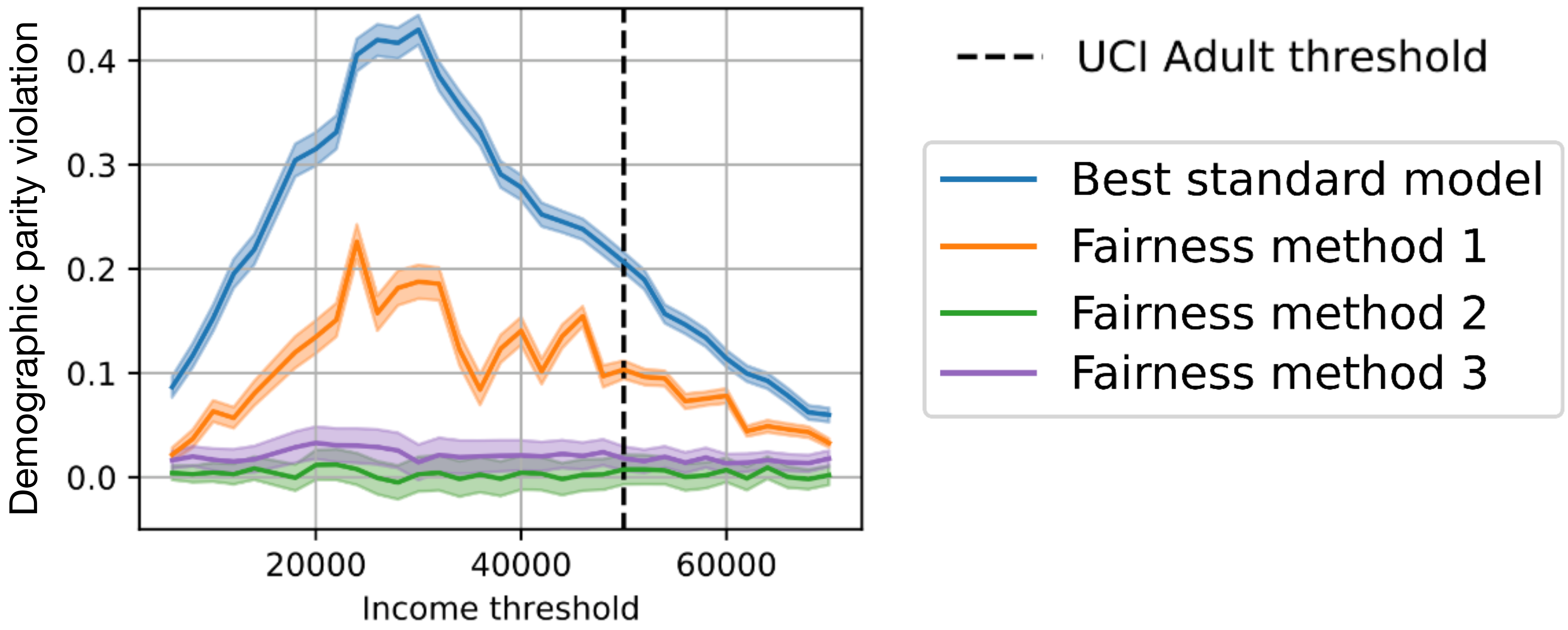
Varying the income threshold



Varying the income threshold



Varying the income threshold



Can we get the best of Census data while avoiding issues with UCI Adult?

Opportunity for new datasets

United States™
Census
Bureau

Opportunity for new datasets

- What do we want:

United States™
Census
Bureau

Opportunity for new datasets

- What do we want:
  - More, and more diverse datasets

United States™ Census Bureau

Opportunity for new datasets

- What do we want:
  - More, and more diverse datasets
  - Test beds for whether methods transfer well between different settings

United States™ Census Bureau

Opportunity for new datasets

- What do we want:
  - More, and more diverse datasets
  - Test beds for whether methods transfer well between different settings
  - Enough data to study intersection groups

United States™ Census Bureau

Opportunity for new datasets

- What do we want:
    - More, and more diverse datasets
    - Test beds for whether methods transfer well between different settings
    - Enough data to study intersection groups
- US Census Bureau provides many different data products

Opportunity for new datasets

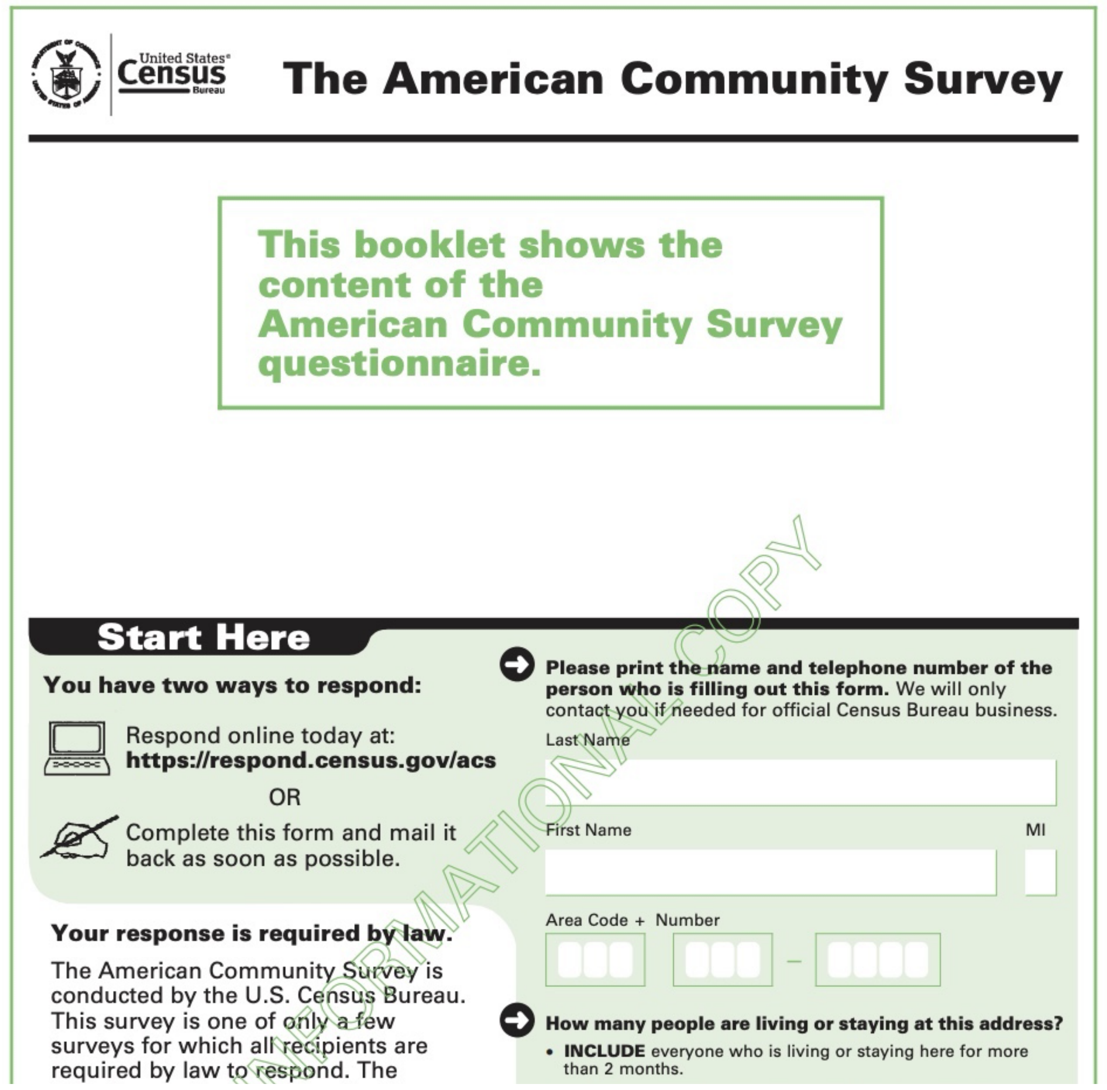
- What do we want:
  - More, and more diverse datasets
  - Test beds for whether methods transfer well between different settings
  - Enough data to study intersection groups
- US Census Bureau provides many different data products
- **Goal**: Expose more of these data products for as benchmark datasets for ML

United States™
Census
Bureau

Opportunity for new datasets

- What do we want:
  - More, and more diverse datasets
  - Test beds for whether methods transfer well between different settings
  - Enough data to study intersection groups
- US Census Bureau provides many different data products
- **Goal**: Expose more of these data products for as benchmark datasets for ML

United States™ Census Bureau



United States® Census Bureau

The American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

Start Here

You have two ways to respond:

Respond online today at:
https://respond.census.gov/acs

OR

Complete this form and mail it back as soon as possible.

Your response is required by law.

The American Community Survey is conducted by the U.S. Census Bureau. This survey is one of only a few surveys for which all recipients are required by law to respond. The

➔ **Please print the name and telephone number of the person who is filling out this form.** We will only contact you if needed for official Census Bureau business.

Last Name

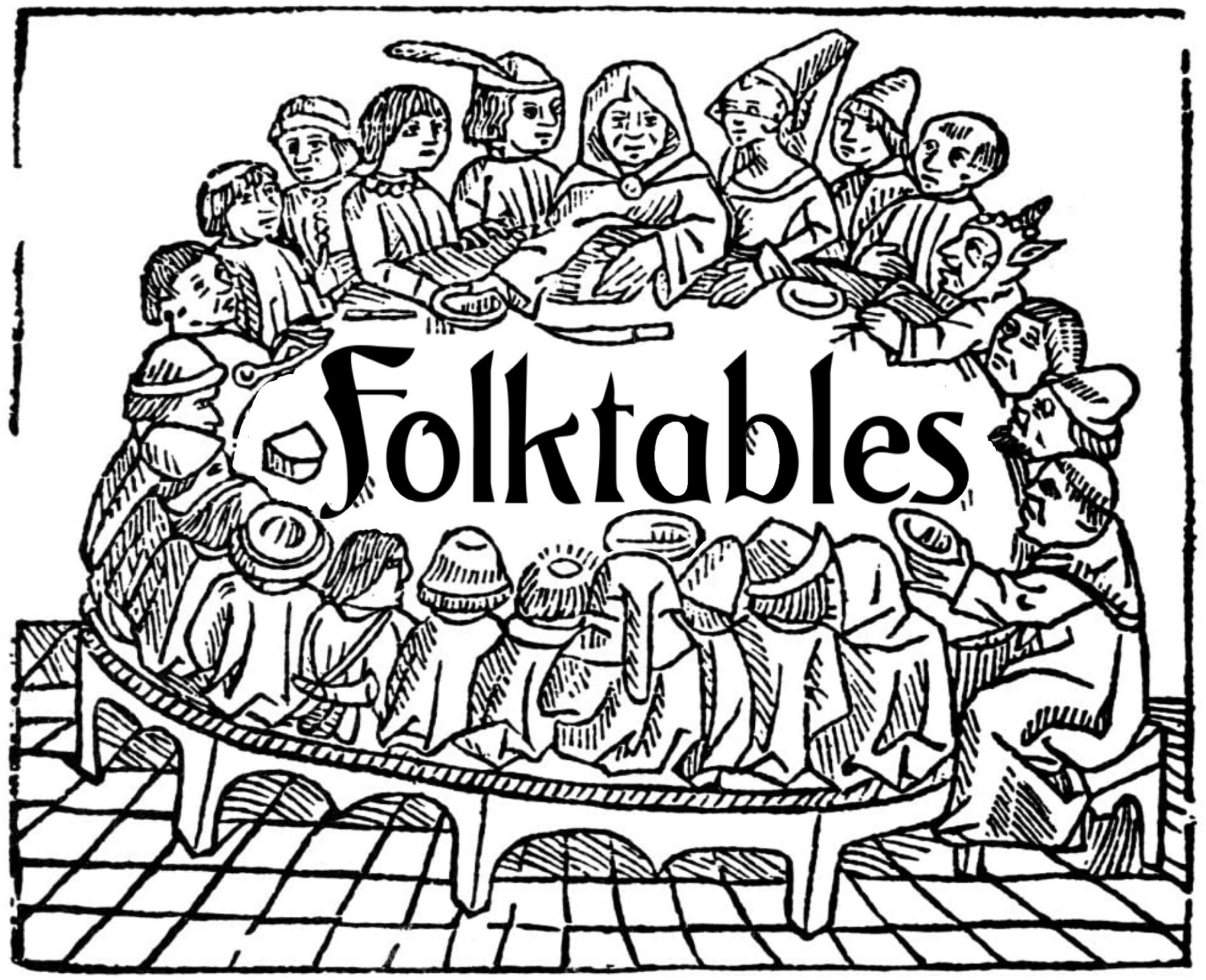
First Name MI

Area Code + Number

➔ **How many people are living or staying at this address?**

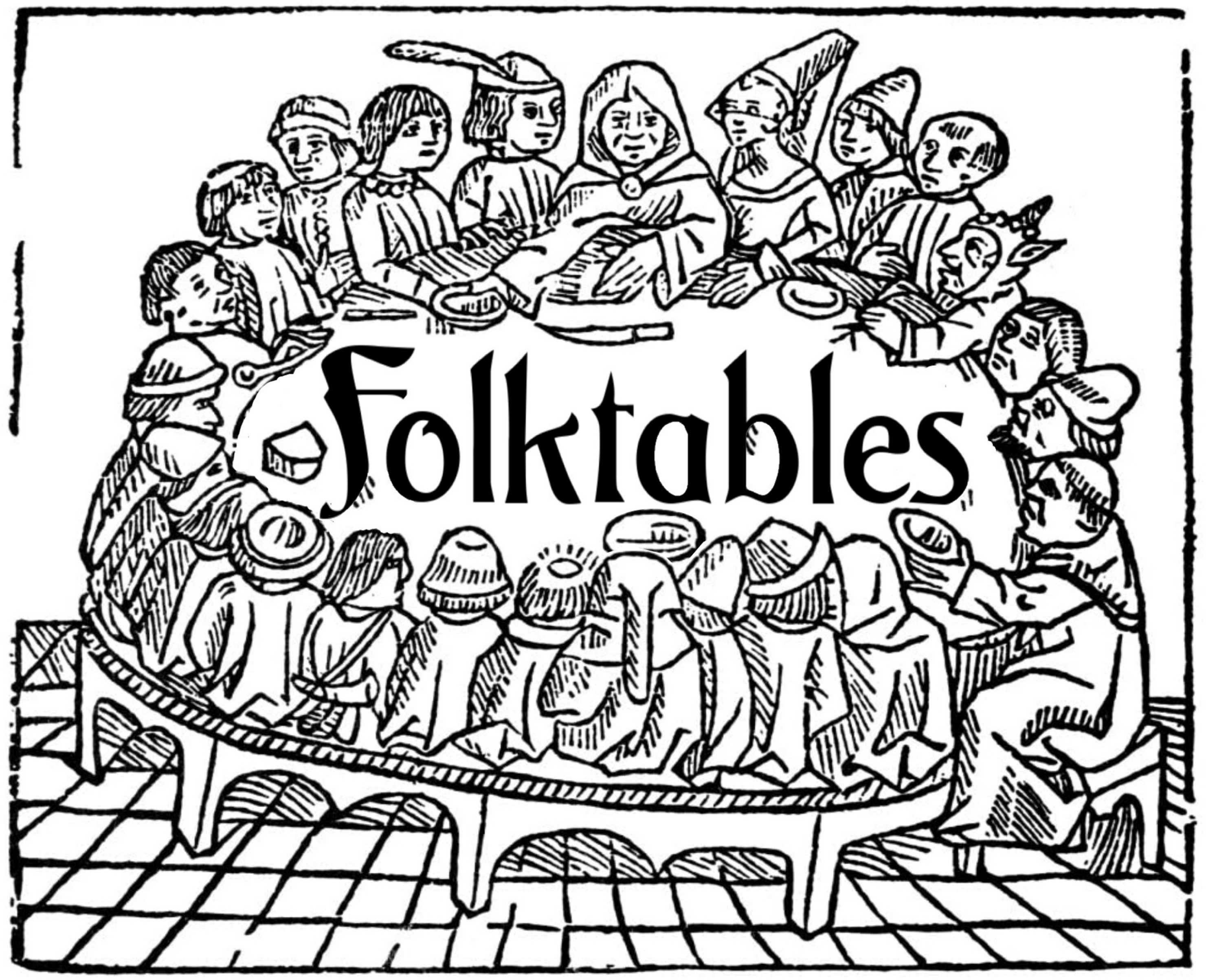
- **INCLUDE** everyone who is living or staying here for more than 2 months.

Folktables: New datasets from Census microdata



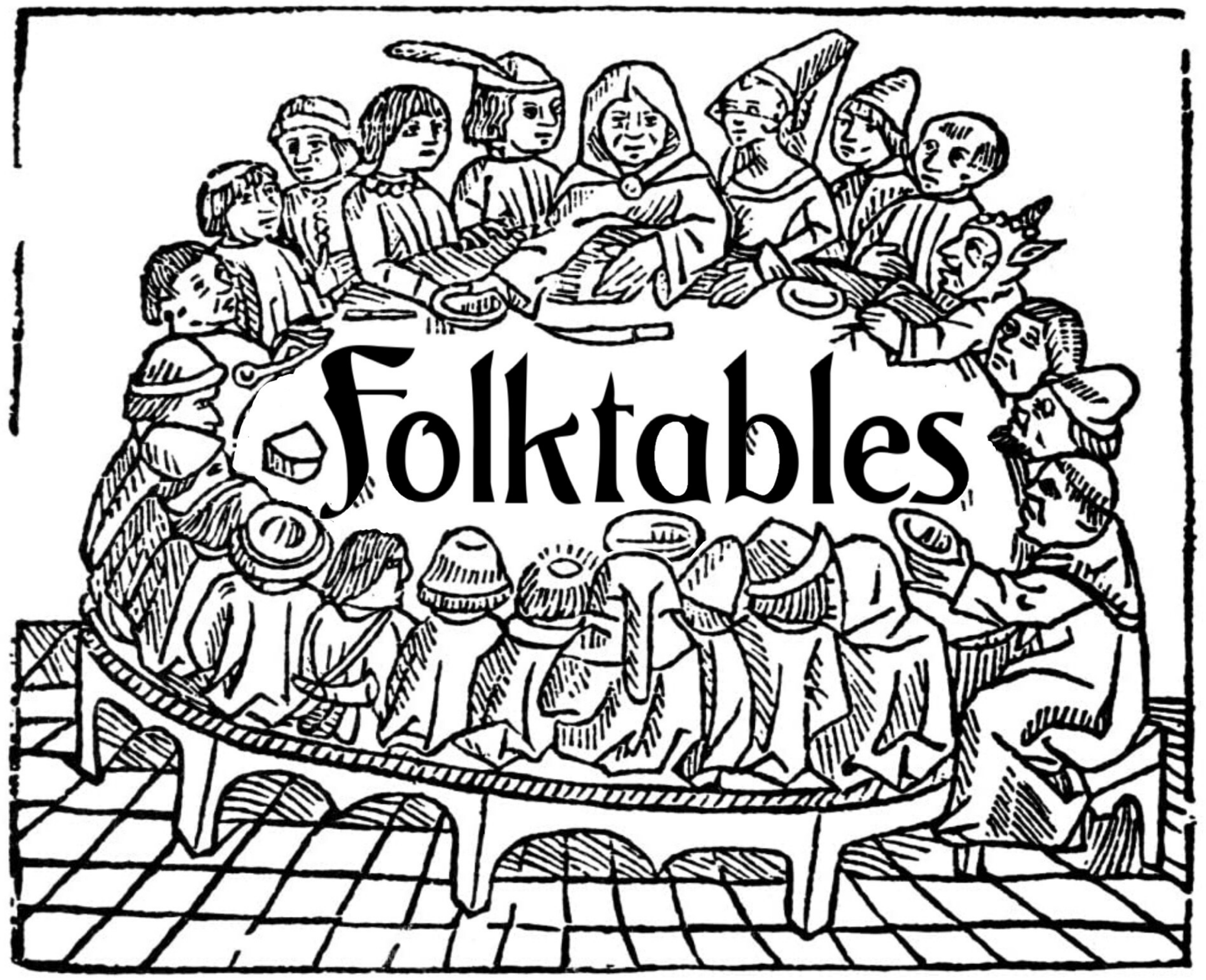
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:



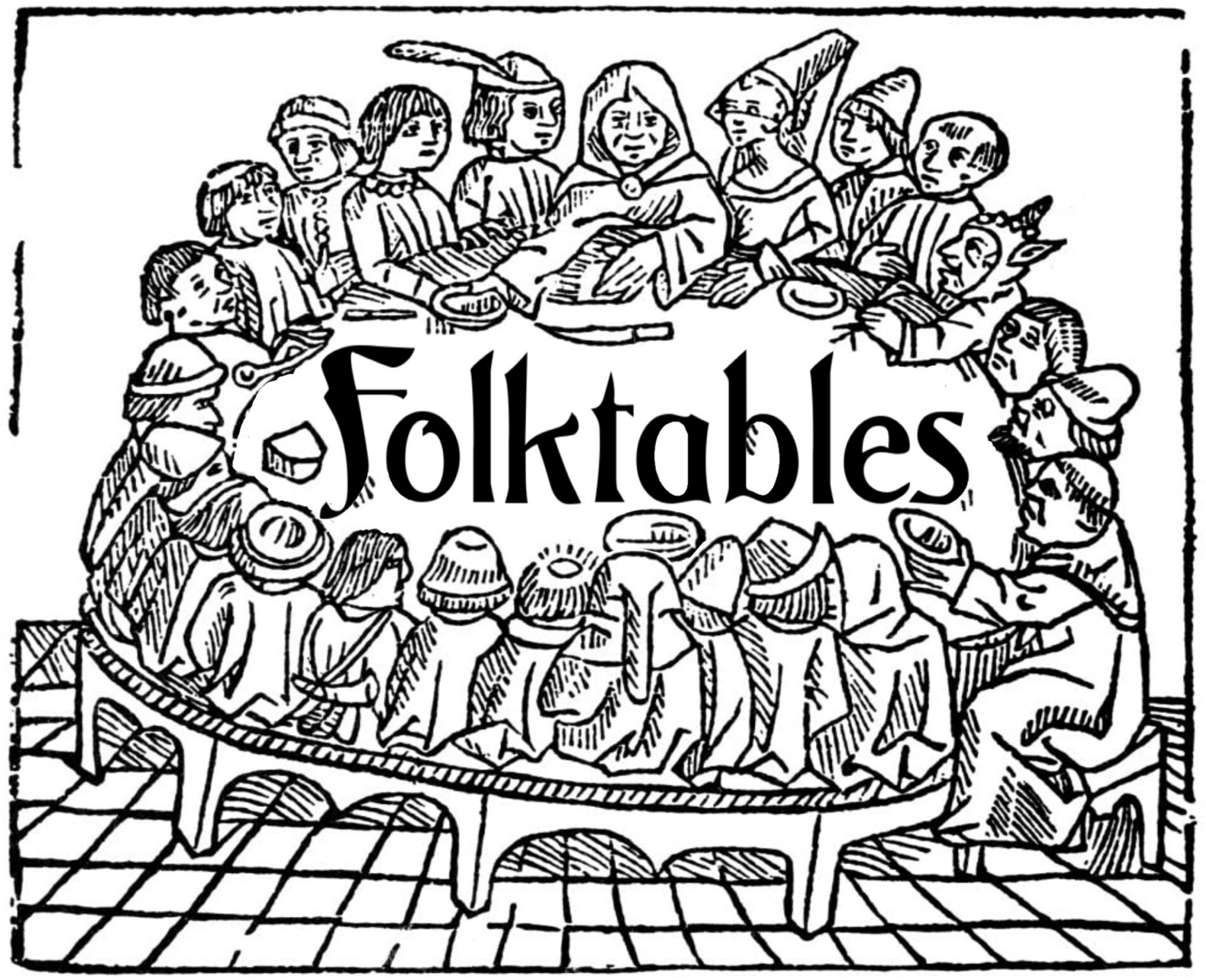
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:
  - Income, employment, health coverage, travel time to work, mobility



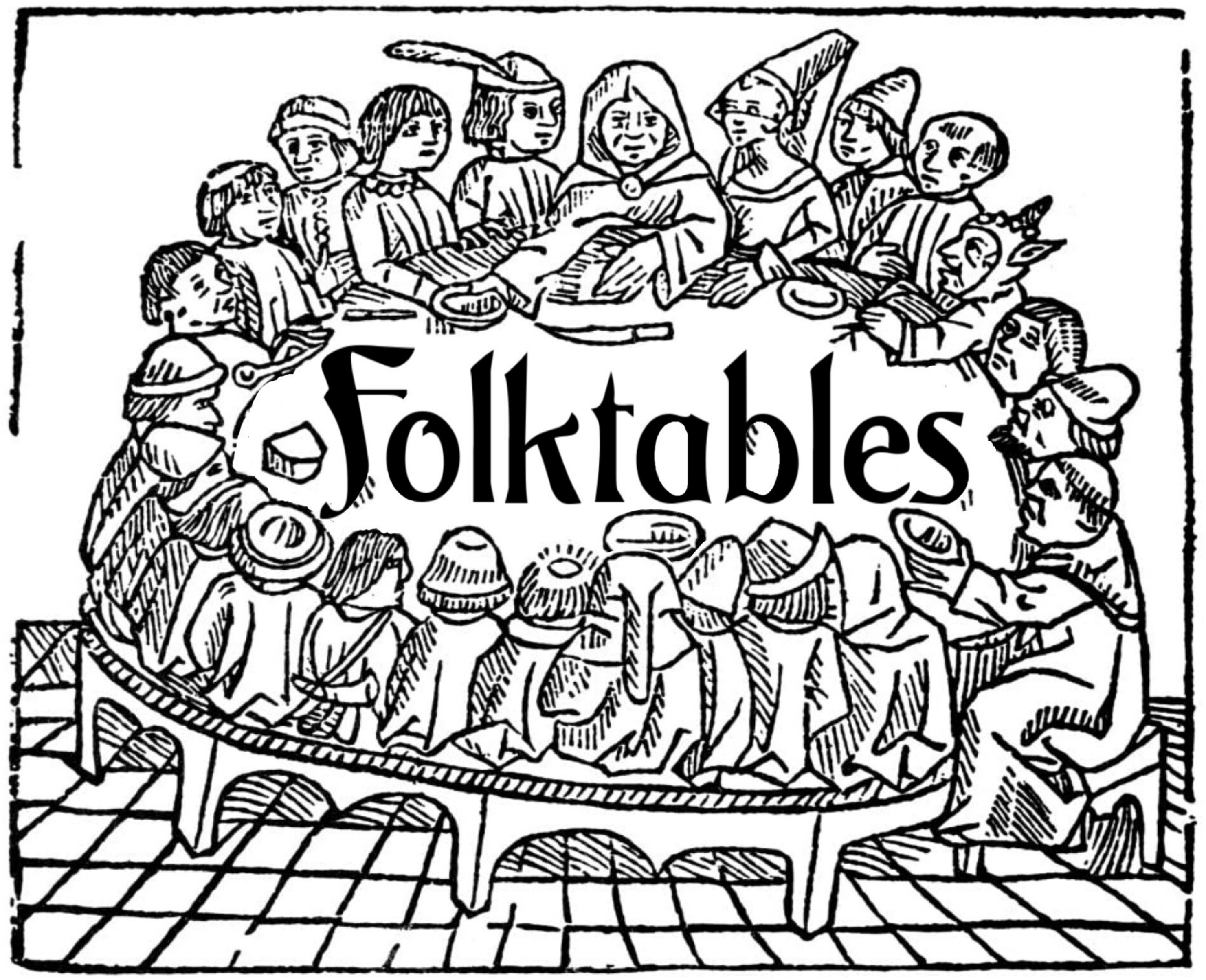
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:
    - Income, employment, health coverage, travel time to work, mobility
    - New features, larger sample size, updated Census encodings



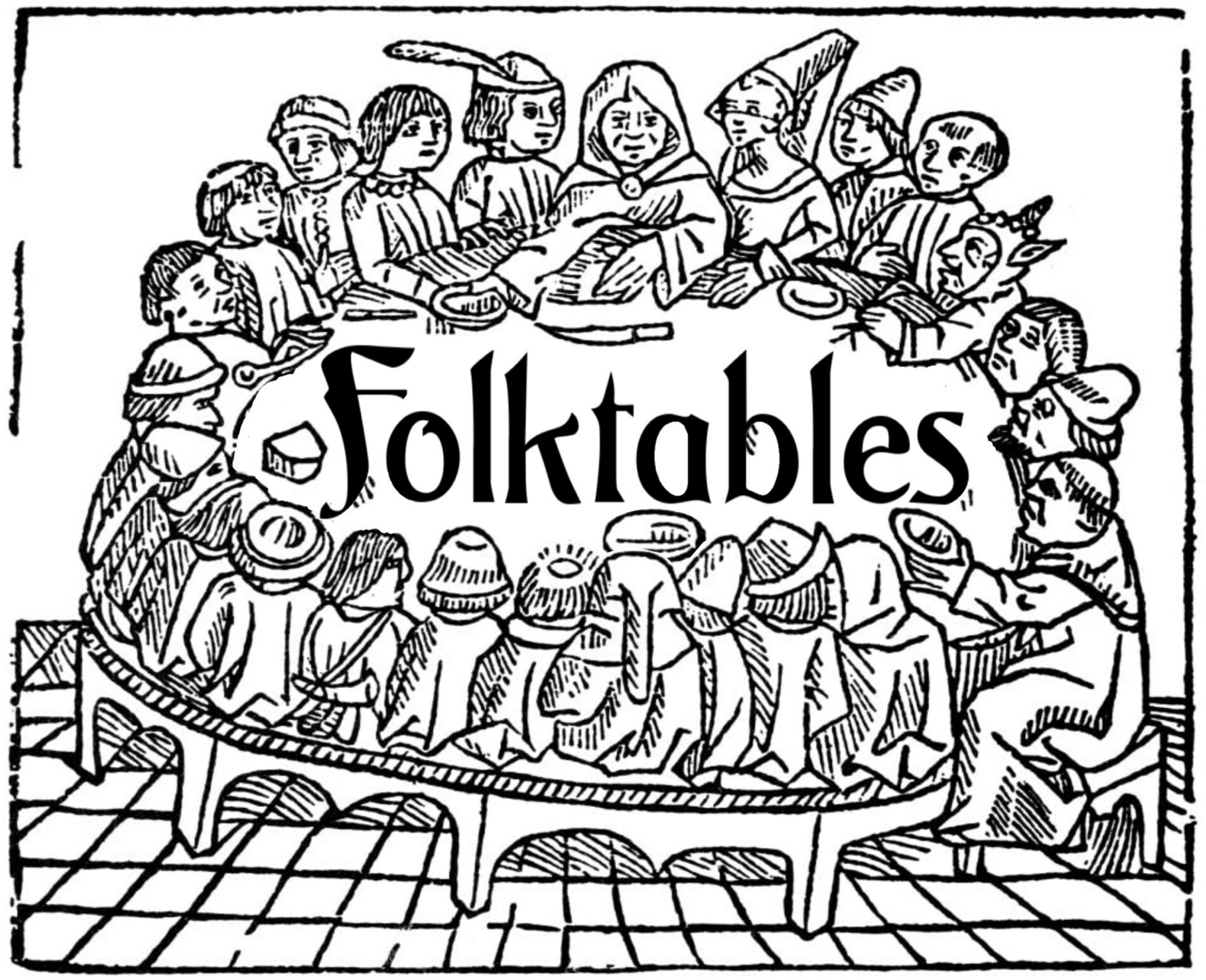
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:
  - Income, employment, health coverage, travel time to work, mobility
  - New features, larger sample size, updated Census encodings
- Easy to create other predictions problems



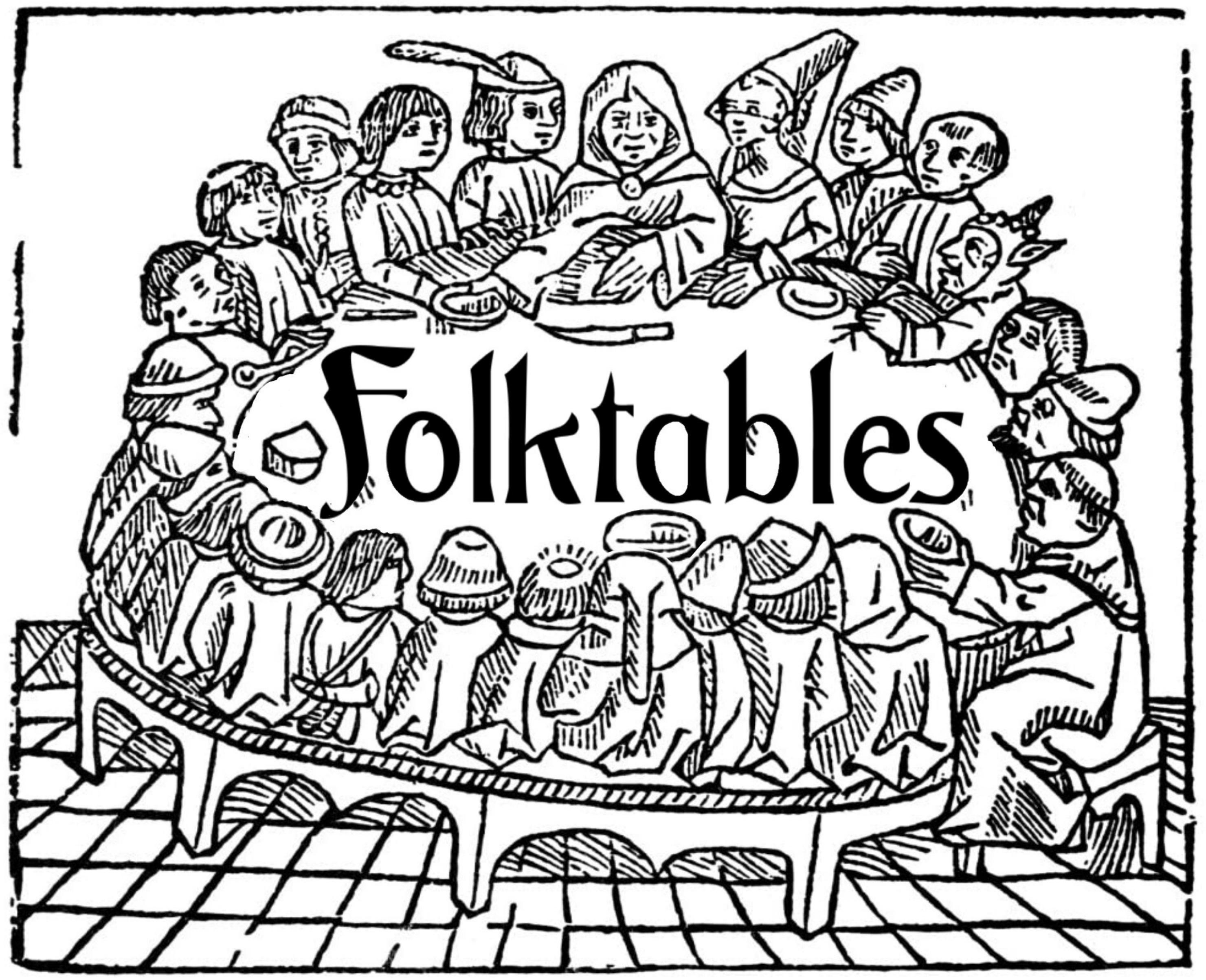
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:
    - Income, employment, health coverage, travel time to work, mobility
    - New features, larger sample size, updated Census encodings
- Easy to create other predictions problems
- 50 states, multiple years



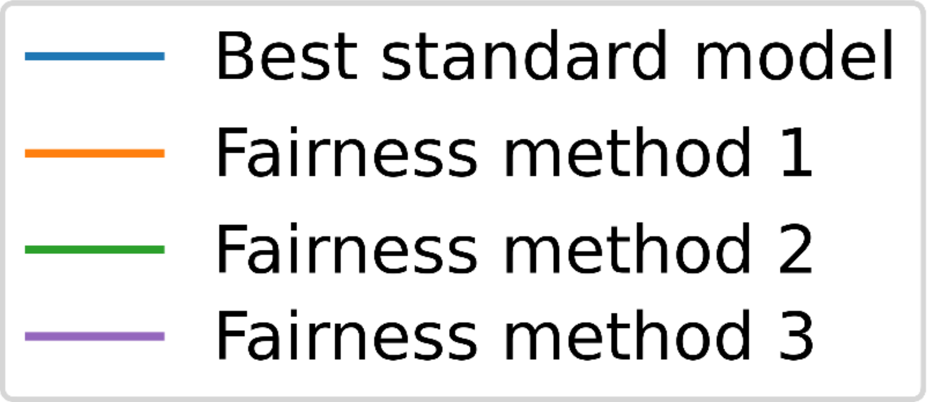
Folktables: New datasets from Census microdata

- New prediction problems from ACS data:
    - Income, employment, health coverage, travel time to work, mobility
    - New features, larger sample size, updated Census encodings
- Easy to create other predictions problems
- 50 states, multiple years
    - Ideal for studying performance variation and distribution shift



Stability over time

Stability over time



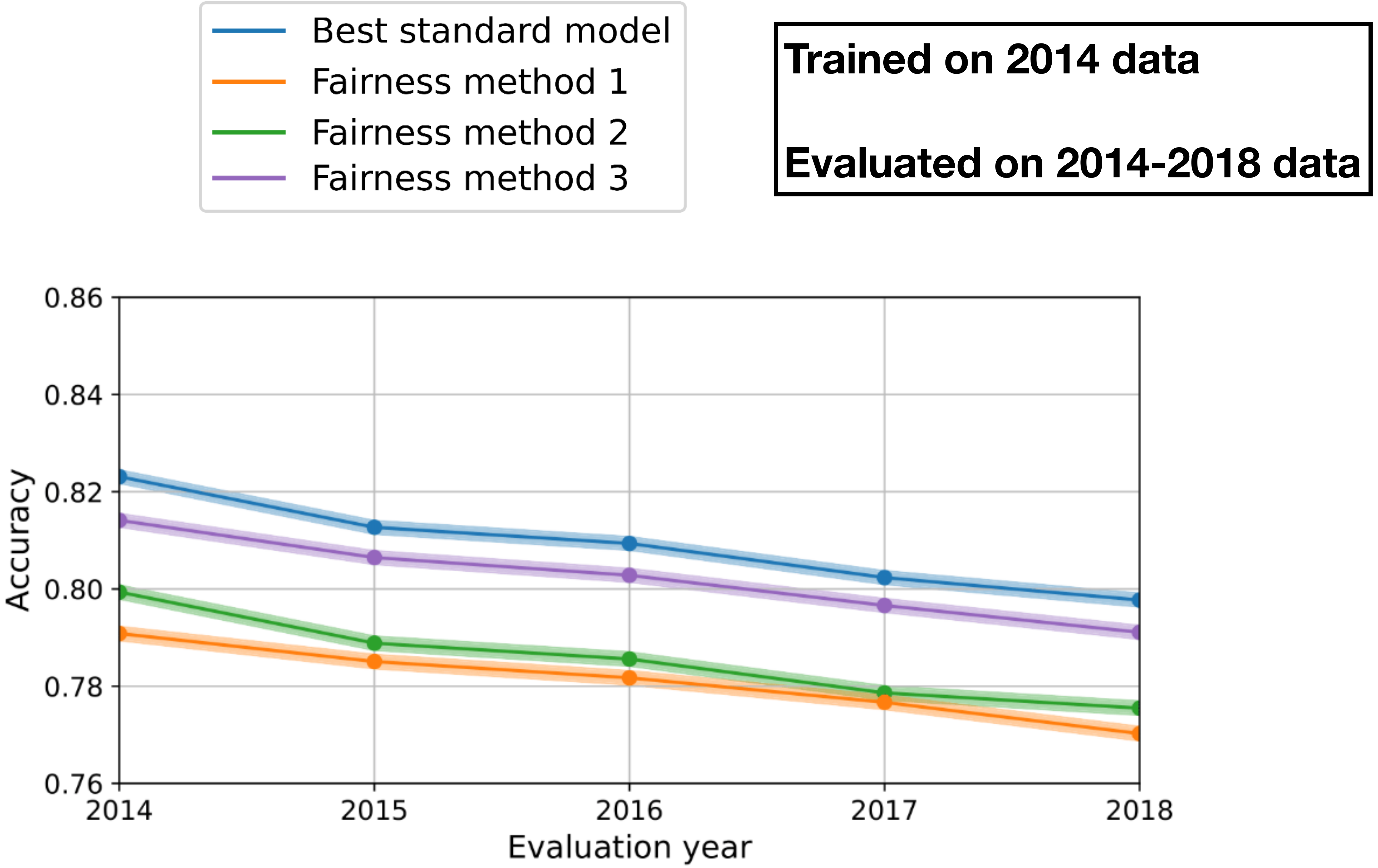
Stability over time

- 🟦 Best standard model
- 🟧 Fairness method 1
- 🟩 Fairness method 2
- 🟪 Fairness method 3

Trained on 2014 data

Evaluated on 2014-2018 data

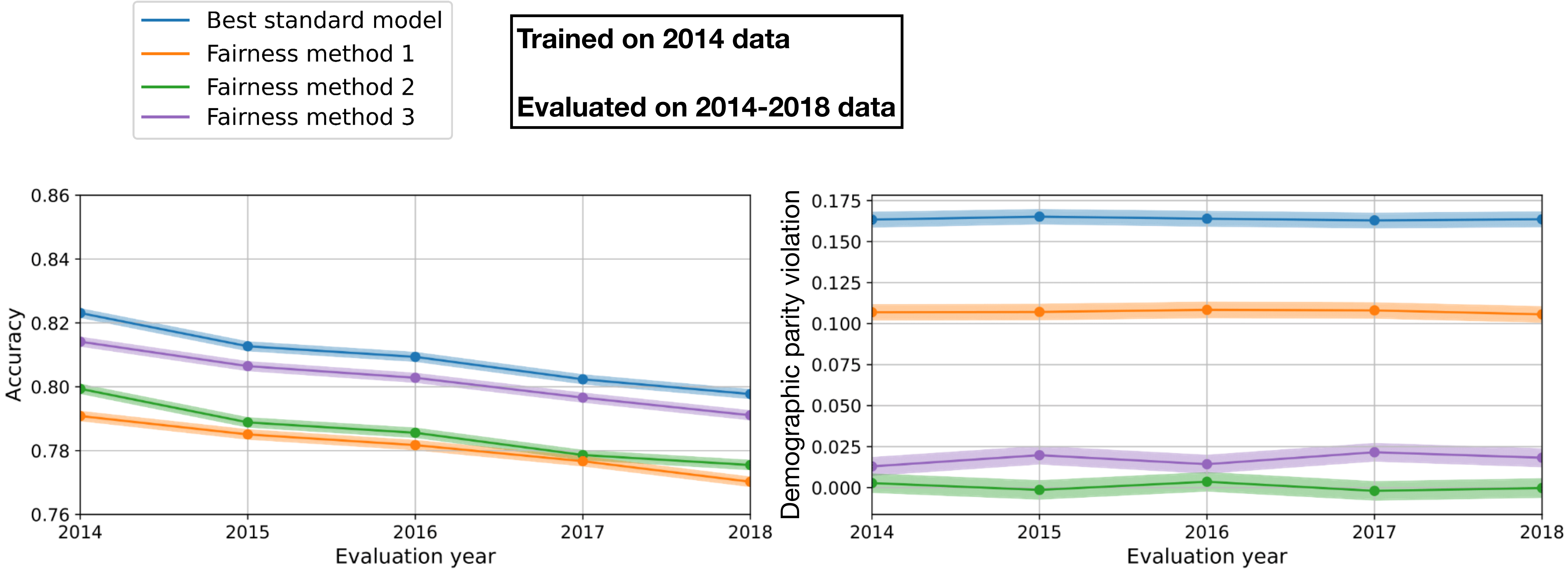
Stability over time



Trained on 2014 data

Evaluated on 2014-2018 data

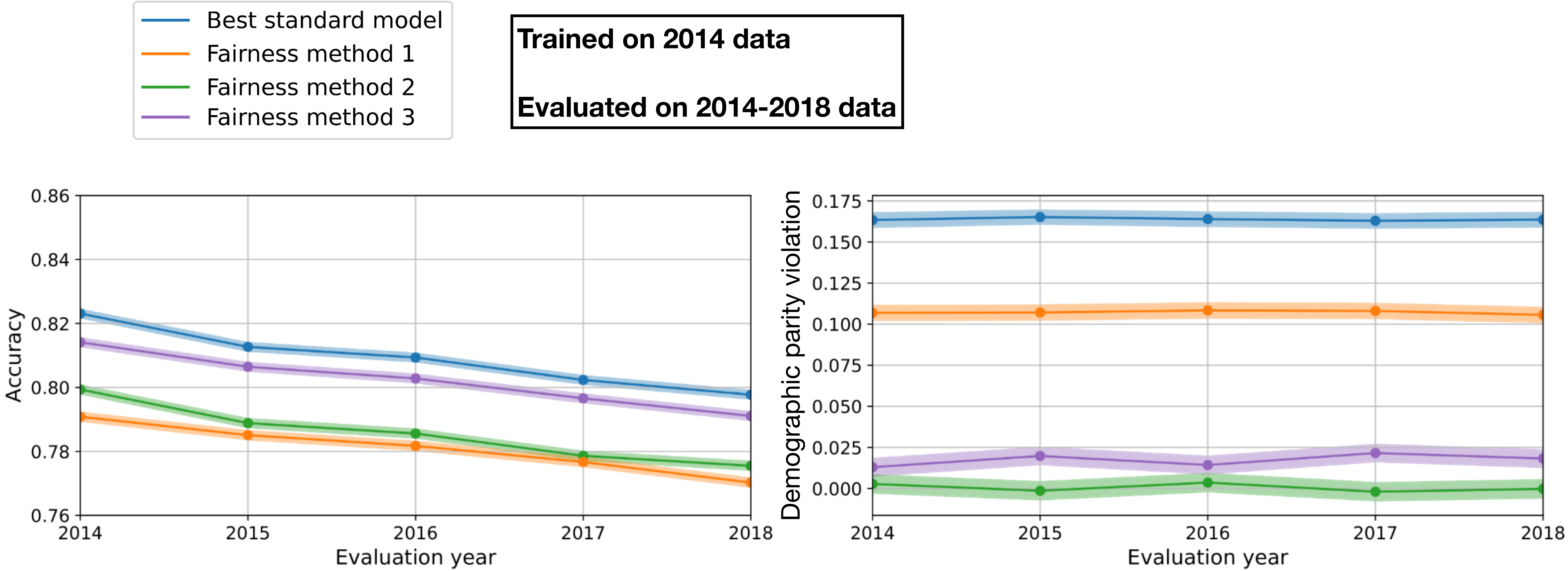
Stability over time



Trained on 2014 data

Evaluated on 2014-2018 data

Stability over time



Trained on 2014 data

Evaluated on 2014-2018 data

Fairness metrics are more stable over time than predictive accuracy

Transferring interventions between states

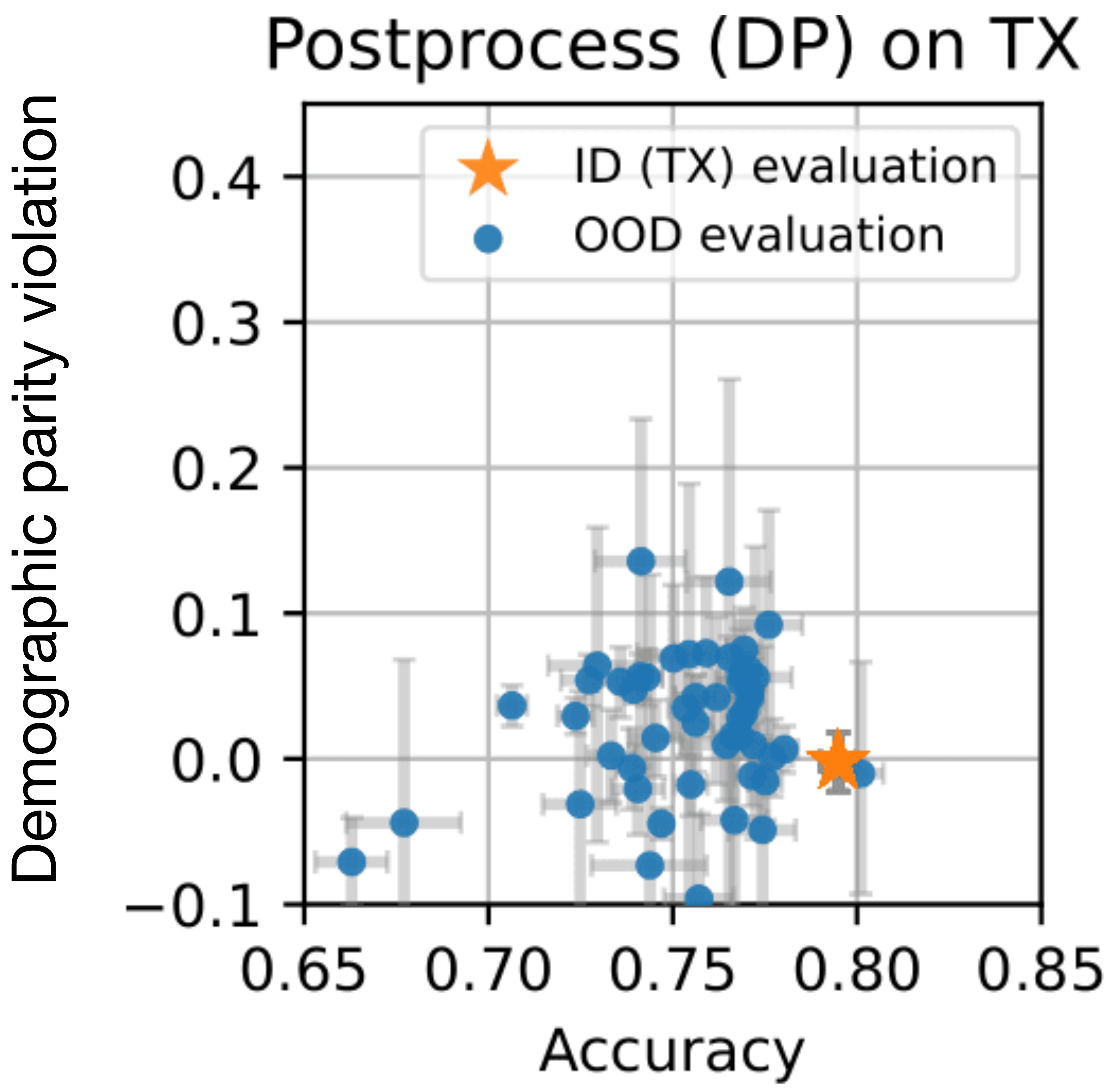
Transferring interventions between states

- Trained classifier on Texas data to enforce demographic parity

Transferring interventions between states

- Trained classifier on Texas data to enforce demographic parity
- Evaluated classifier on all 50 states (each dot corresponds to a state)

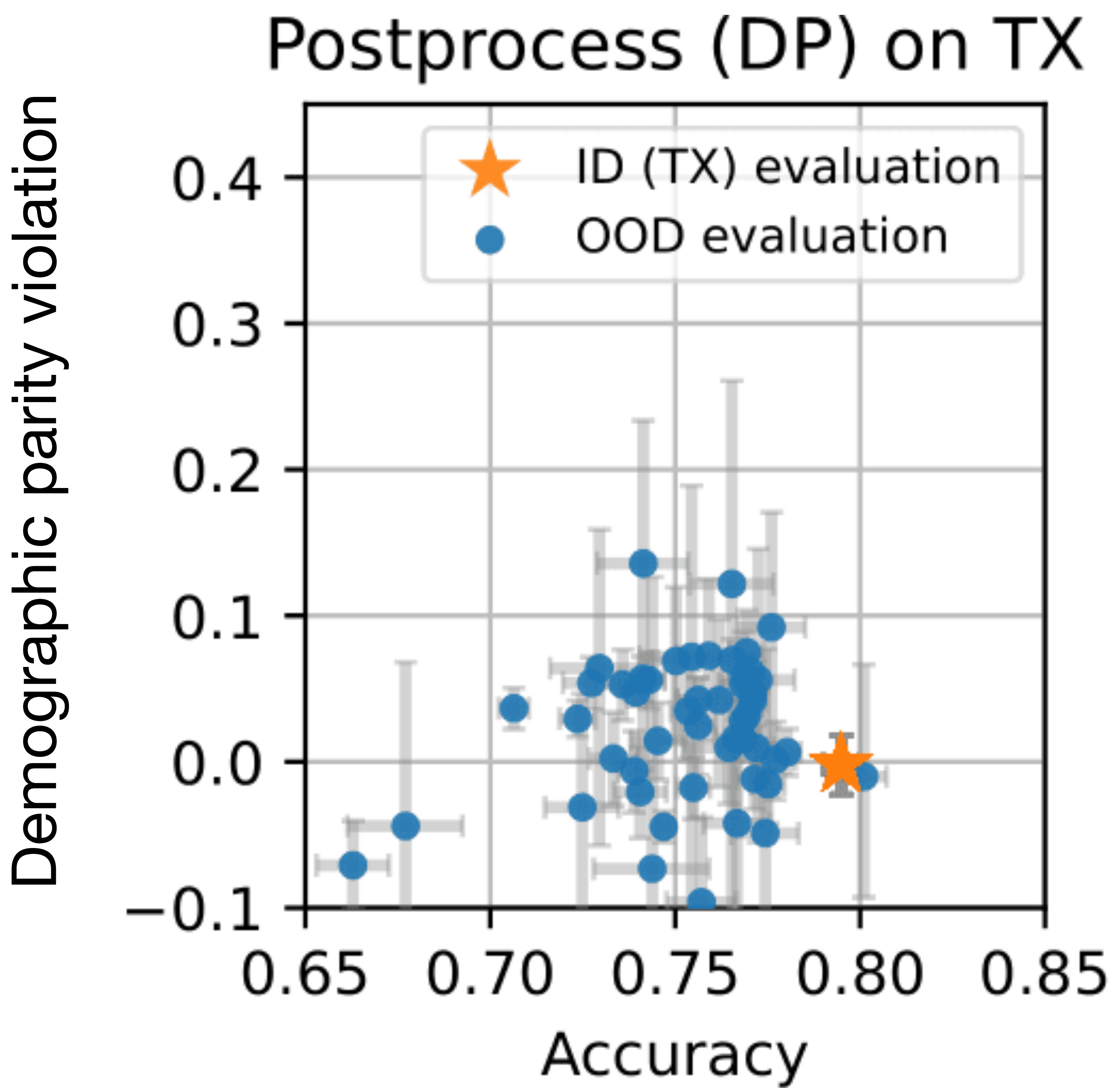
Transferring interventions between states



- Trained classifier on Texas data to enforce demographic parity

- Evaluated classifier on all 50 states (each dot corresponds to a state)

Transferring interventions between states



- Trained classifier on Texas data to enforce demographic parity
- Evaluated classifier on all 50 states (each dot corresponds to a state)
- Demographic parity violations vary substantially across states

Scope and limitations

Scope and limitations

- Census data used here for benchmark datasets, **not** substantiative studies about income, employment, housing, healthcare, etc.

Scope and limitations

- Census data used here for benchmark datasets, **not** substantiative studies about income, employment, housing, healthcare, etc.
- US-centric data; still a need for datasets in international contexts

Scope and limitations

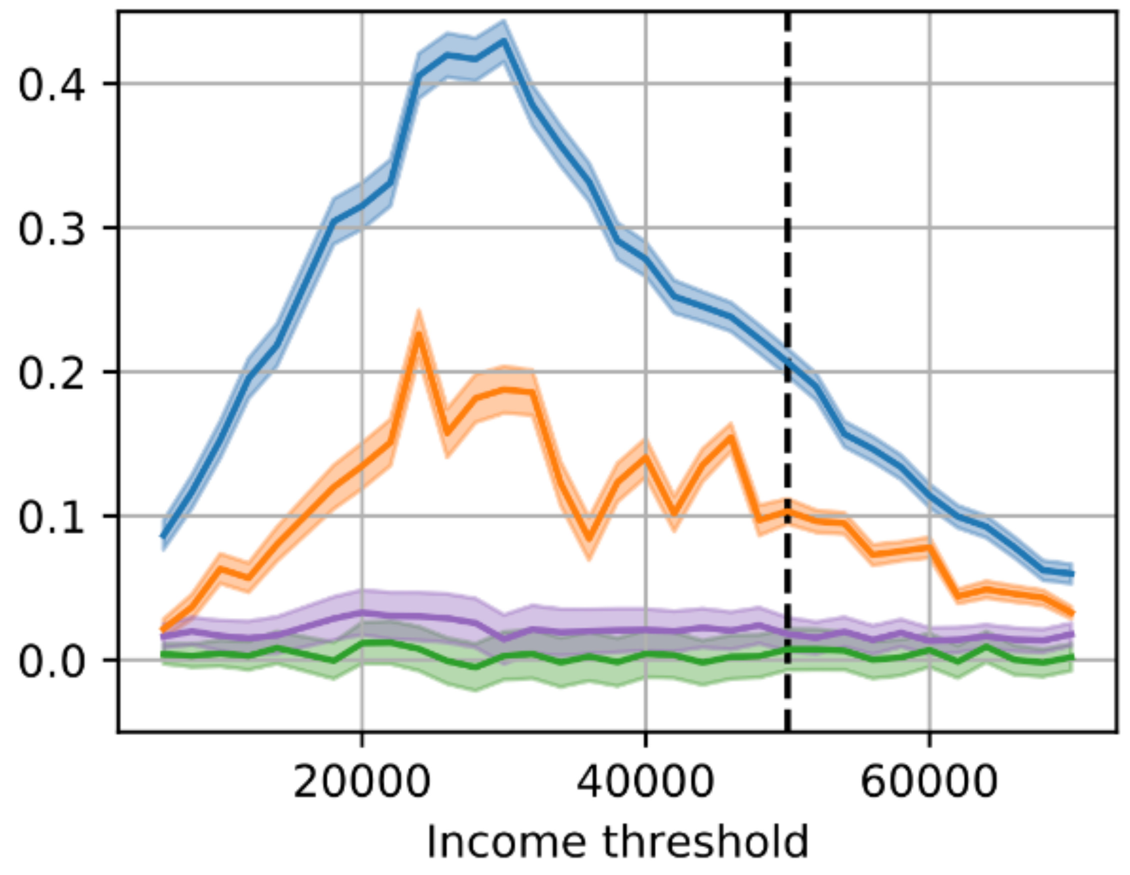
- Census data used here for benchmark datasets, **not** substantiative studies about income, employment, housing, healthcare, etc.
- US-centric data; still a need for datasets in international contexts
    - Contributions welcome!

Scope and limitations

- Census data used here for benchmark datasets, **not** substantiative studies about income, employment, housing, healthcare, etc.
- US-centric data; still a need for datasets in international contexts
  - Contributions welcome!
  - **Idea**: Leverage the international census data from the 102 countries in IPUMS international

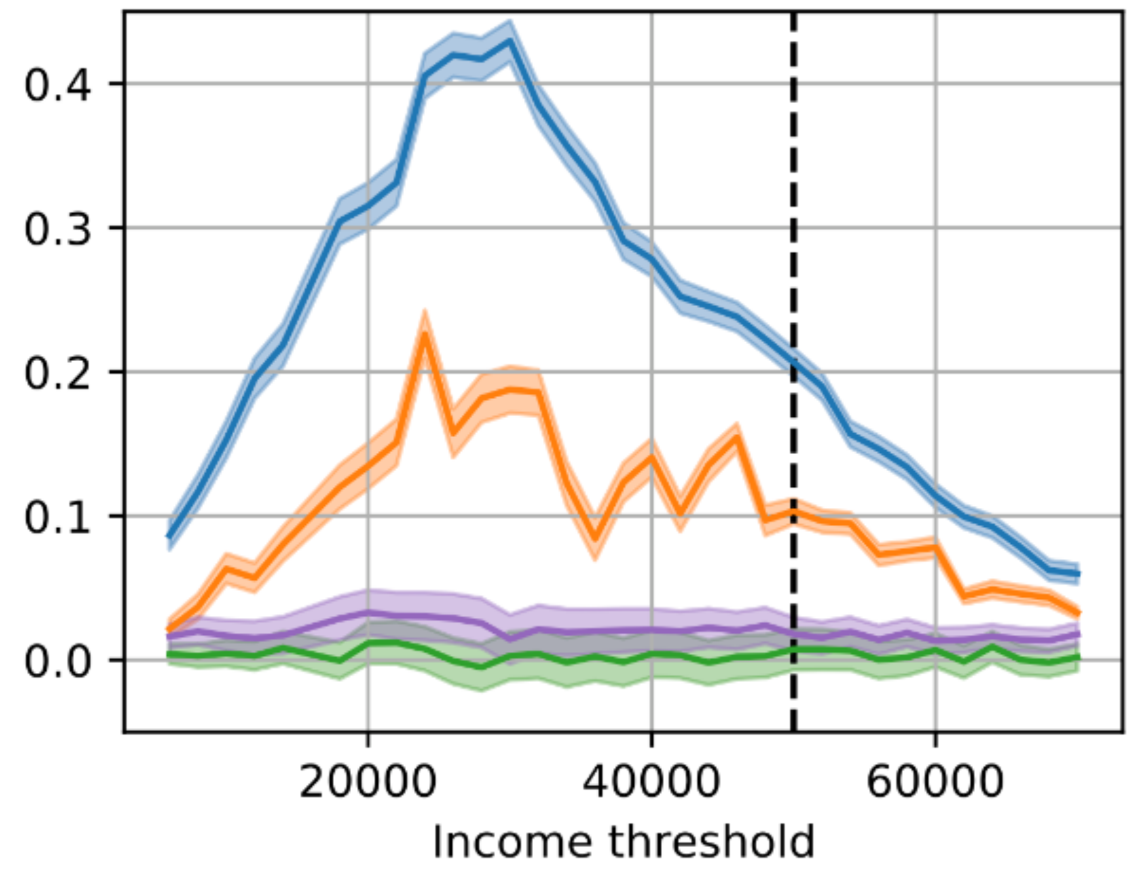
IPUMS
INTERNATIONAL

Takeaways



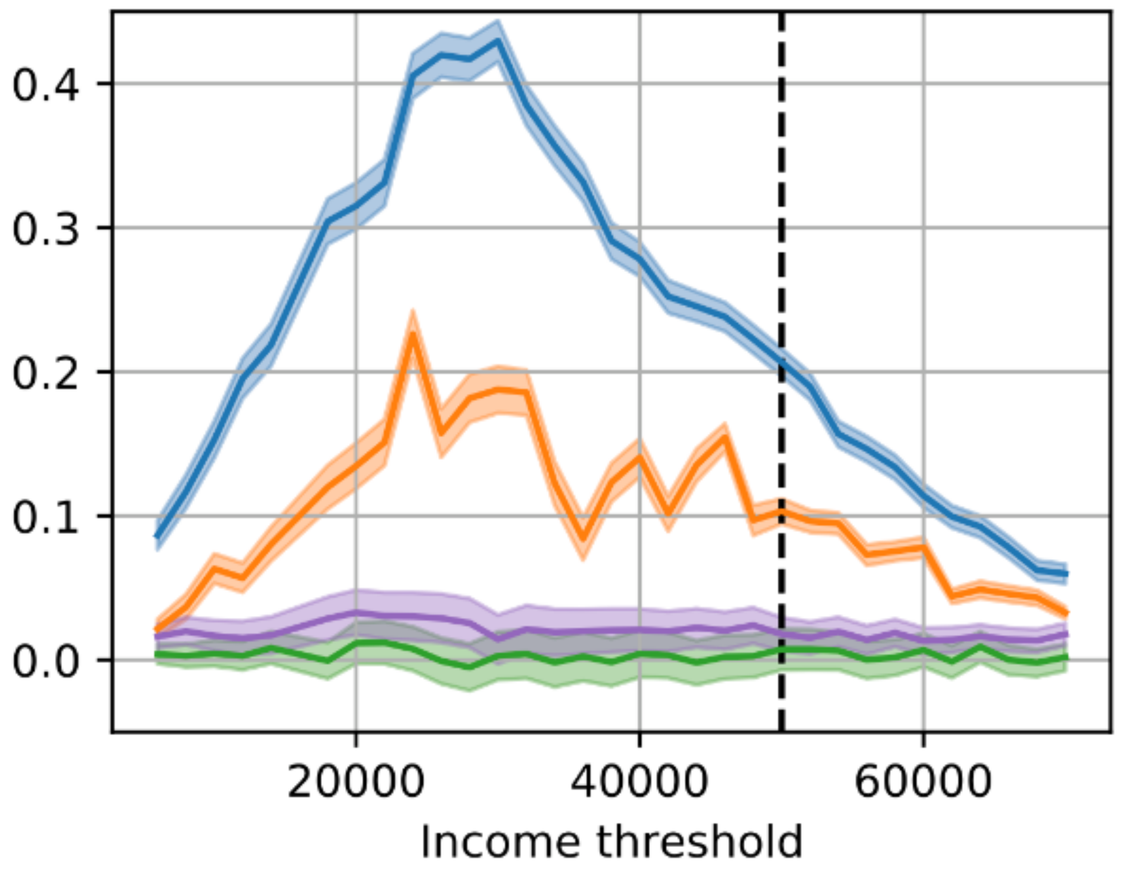
Takeaways

- Fairness community has recognized the importance of datasets

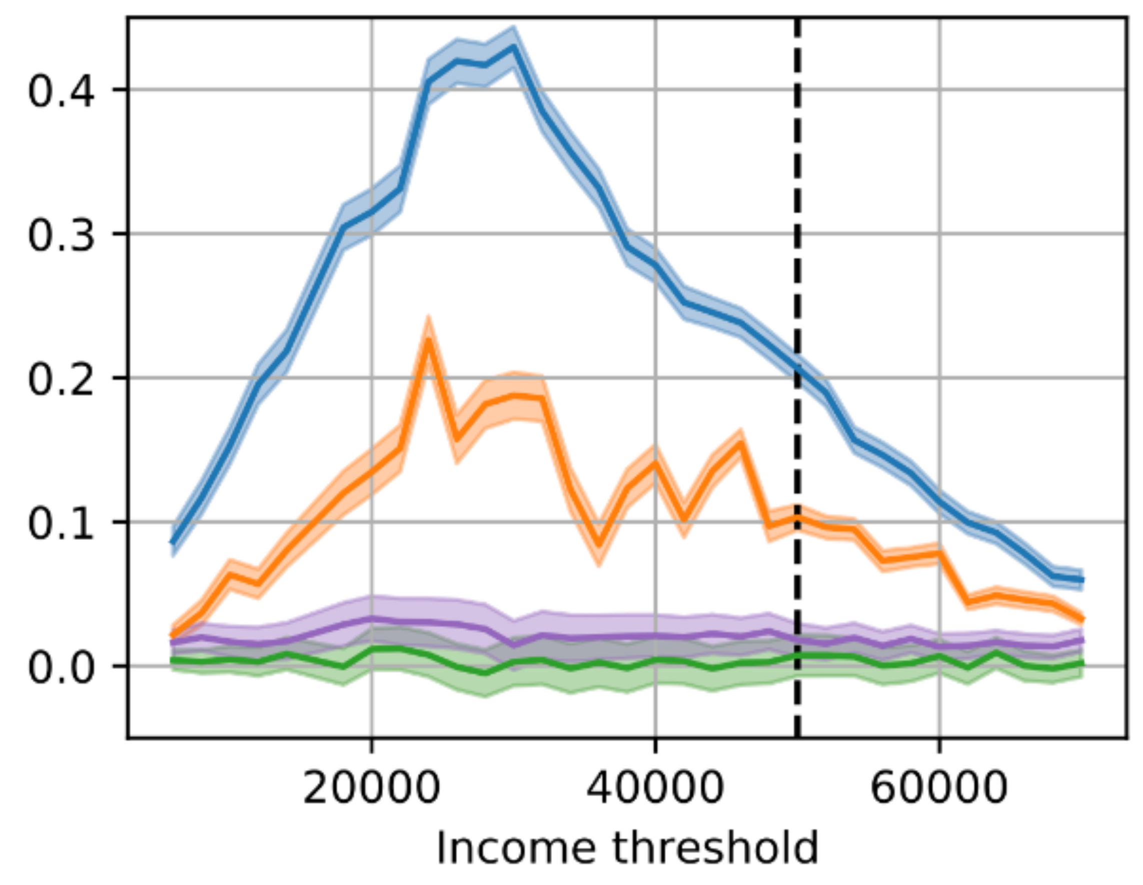


Takeaways

- Fairness community has recognized the importance of datasets
    - Datasets the community relies on still lacking



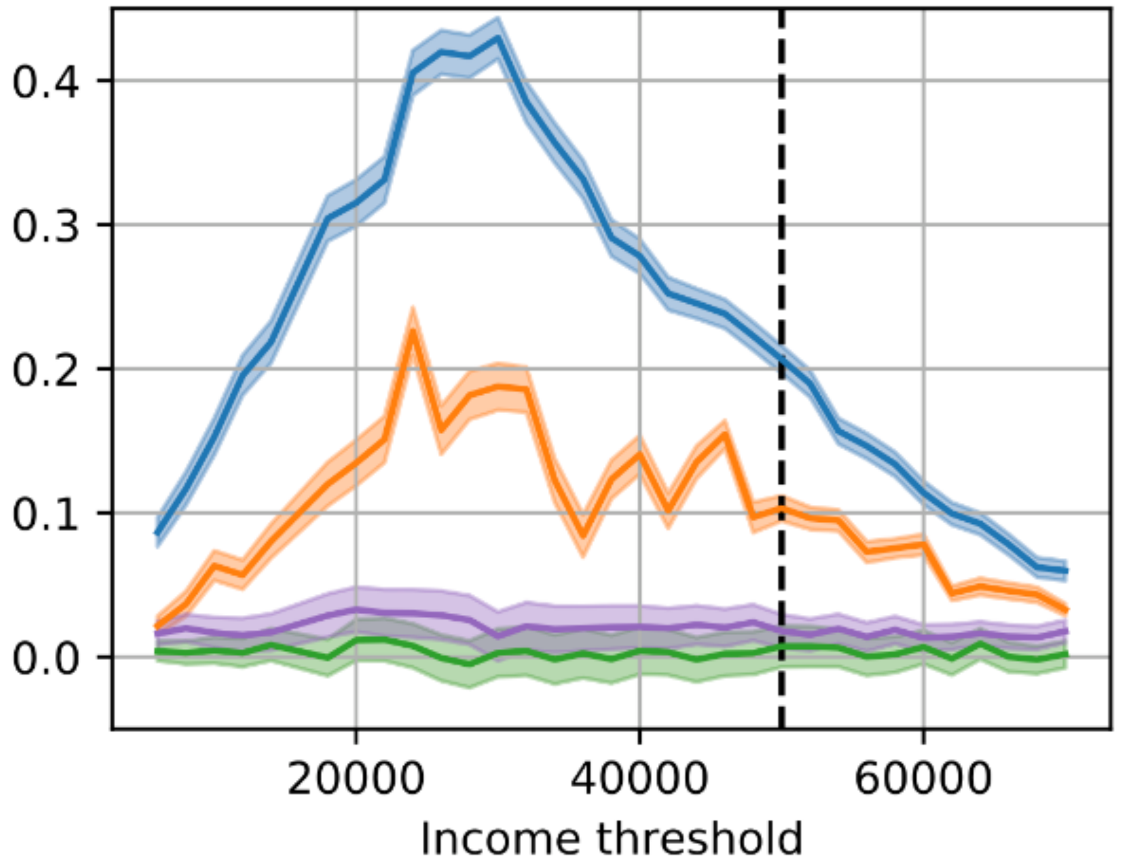
Takeaways

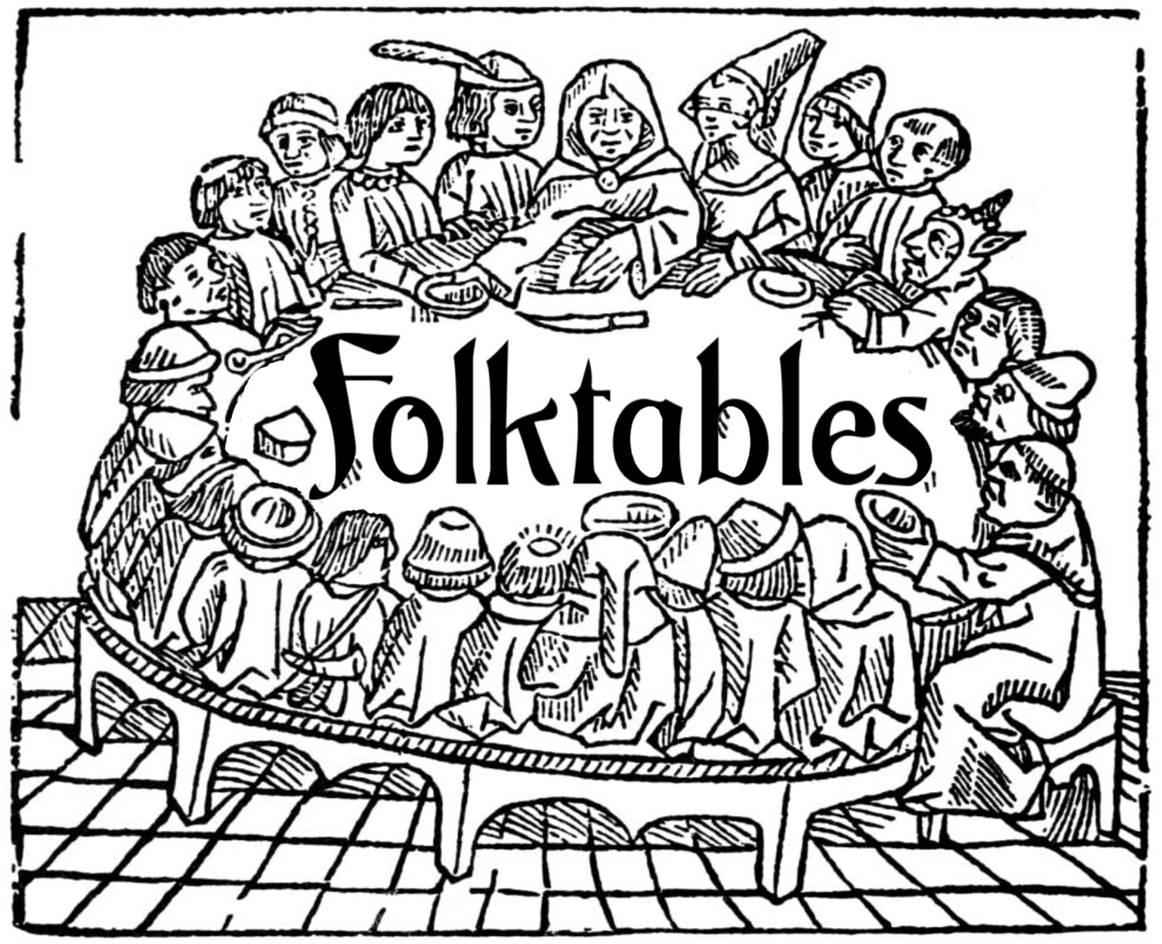


- Fairness community has recognized the importance of datasets
    - Datasets the community relies on still lacking
- Census data provides a rich resource for new high quality datasets

Takeaways

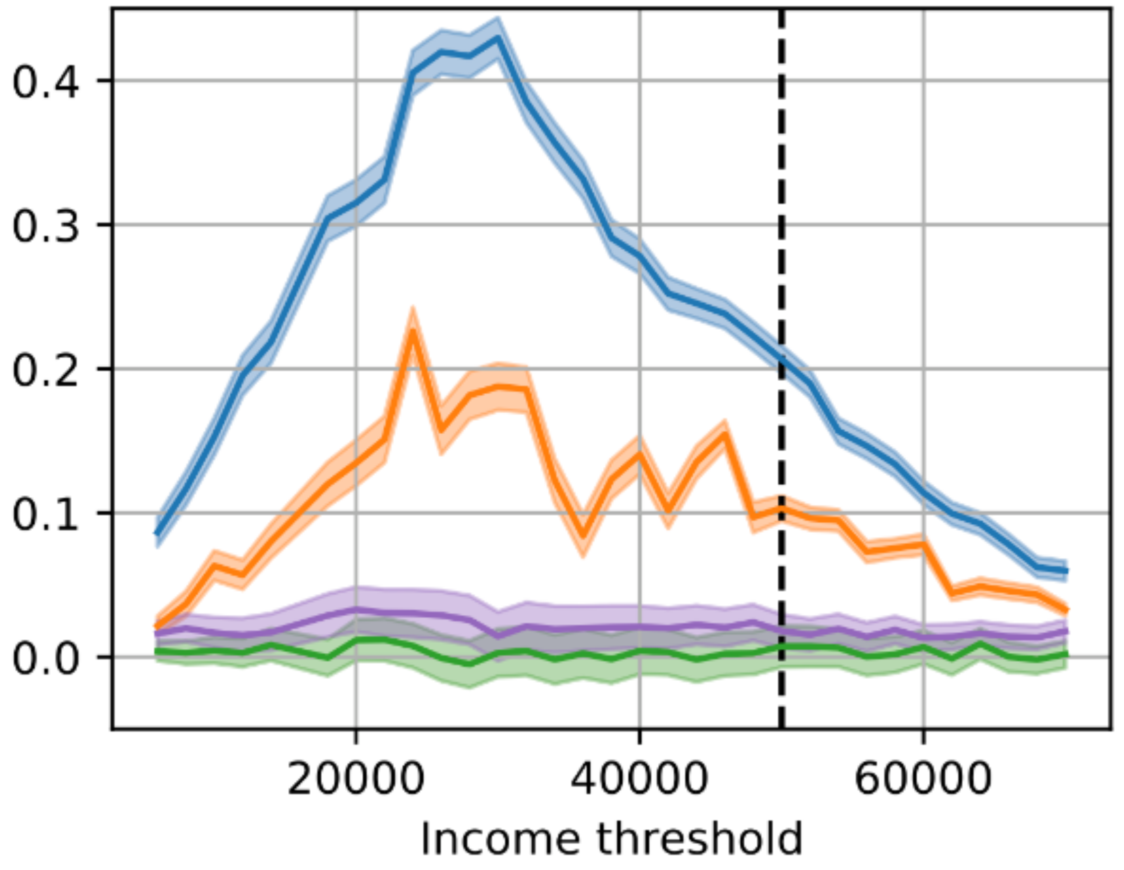
- Fairness community has recognized the importance of datasets
  - Datasets the community relies on still lacking
- Census data provides a rich resource for new high quality datasets
- Folktables provides access to Census data and several ready-to-use new datasets

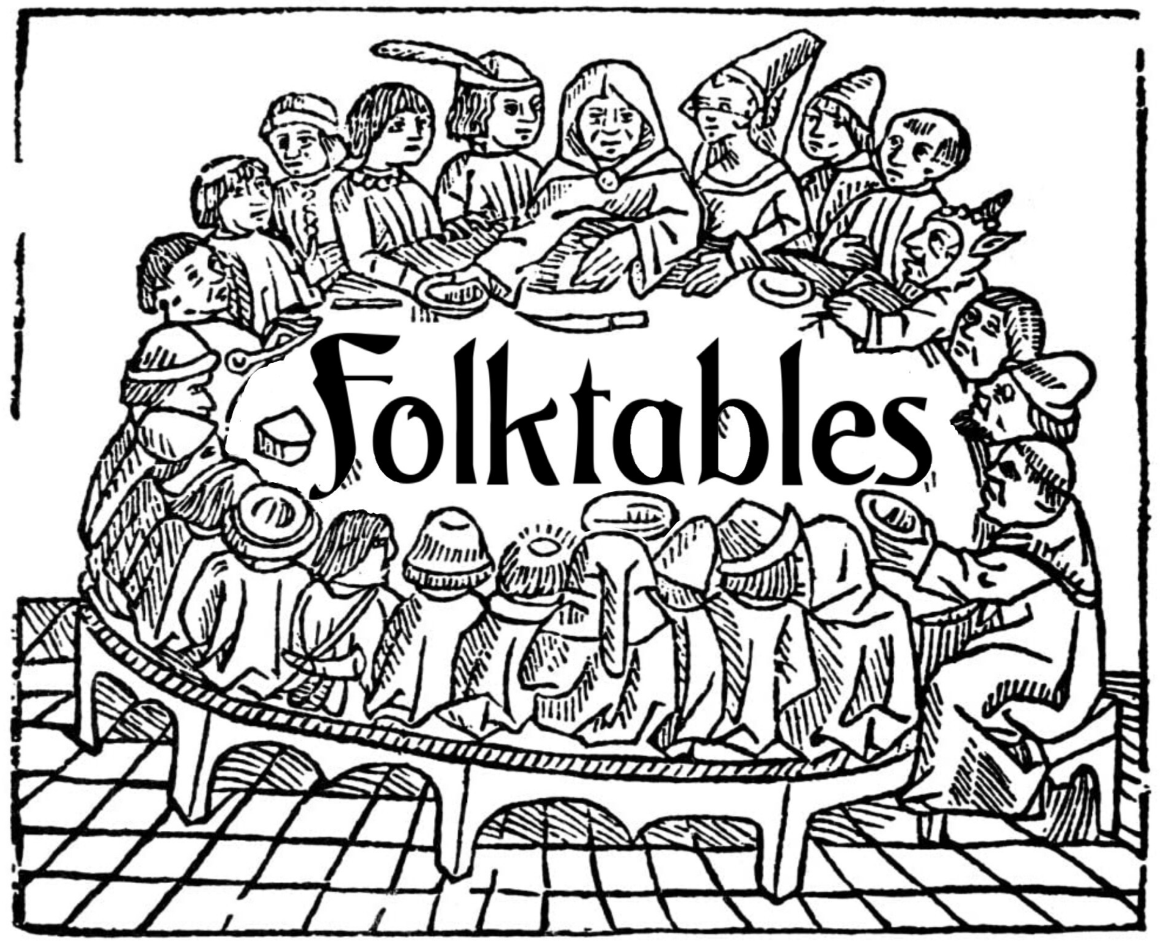




Takeaways

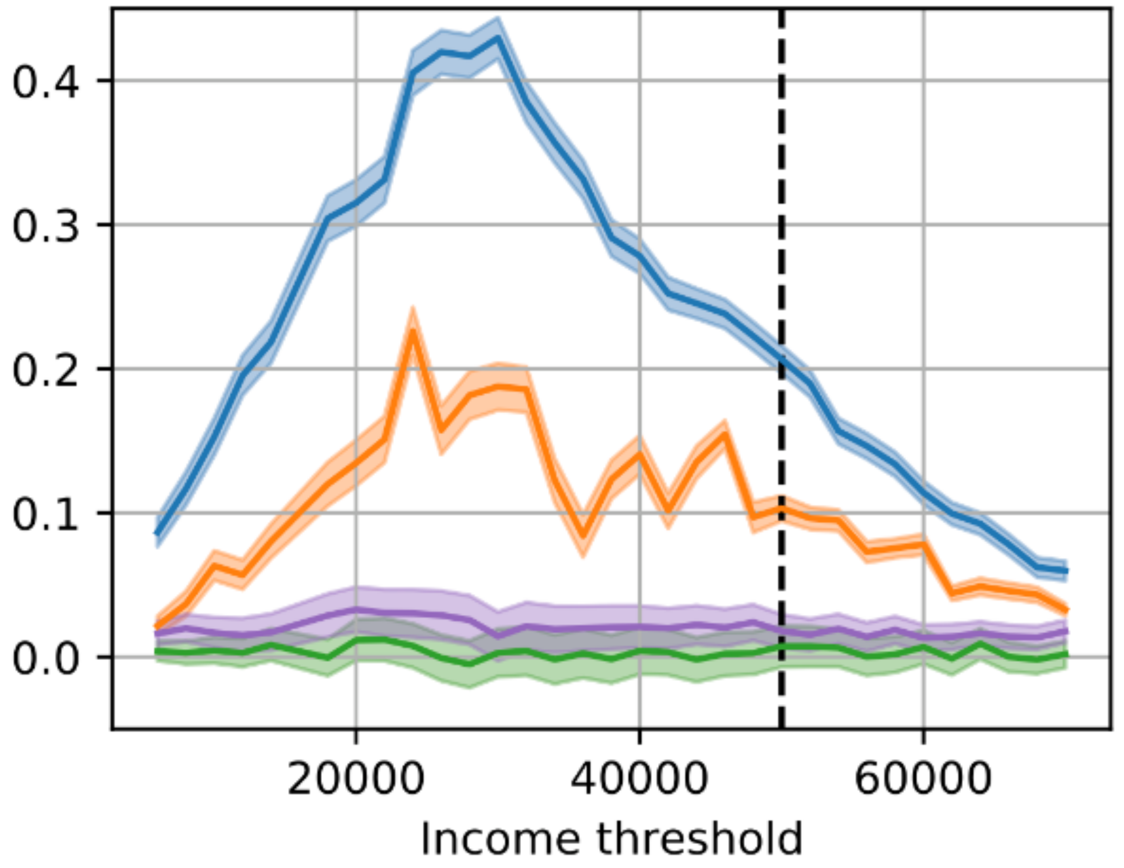
- Fairness community has recognized the importance of datasets
  - Datasets the community relies on still lacking
- Census data provides a rich resource for new high quality datasets
- Folktables provides access to Census data and several ready-to-use new datasets
- More empirical exploration of these datasets is needed

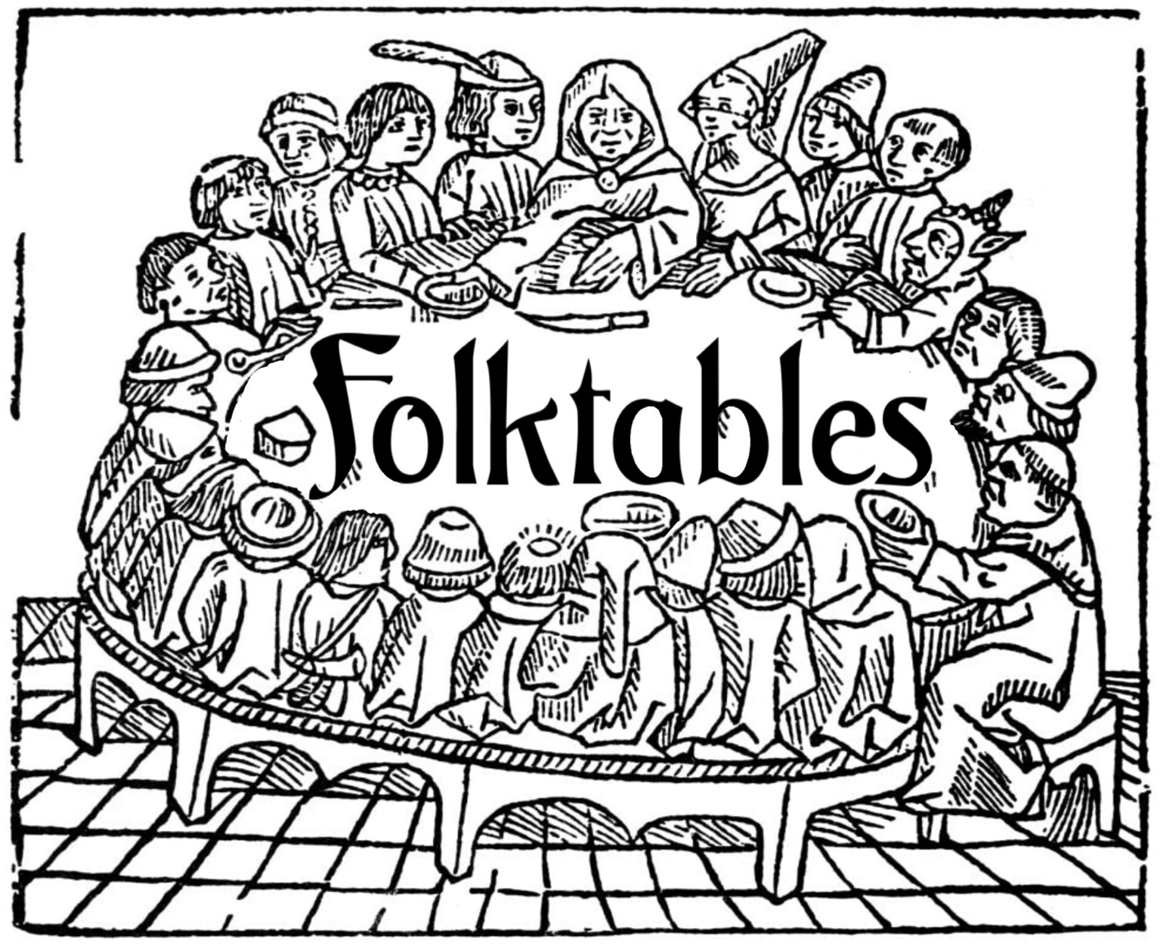




Takeaways

- Fairness community has recognized the importance of datasets
  - Datasets the community relies on still lacking
- Census data provides a rich resource for new high quality datasets
- Folktables provides access to Census data and several ready-to-use new datasets
- More empirical exploration of these datasets is needed
  - Distribution shift, variations between states, etc





New datasets: https://folktables.org