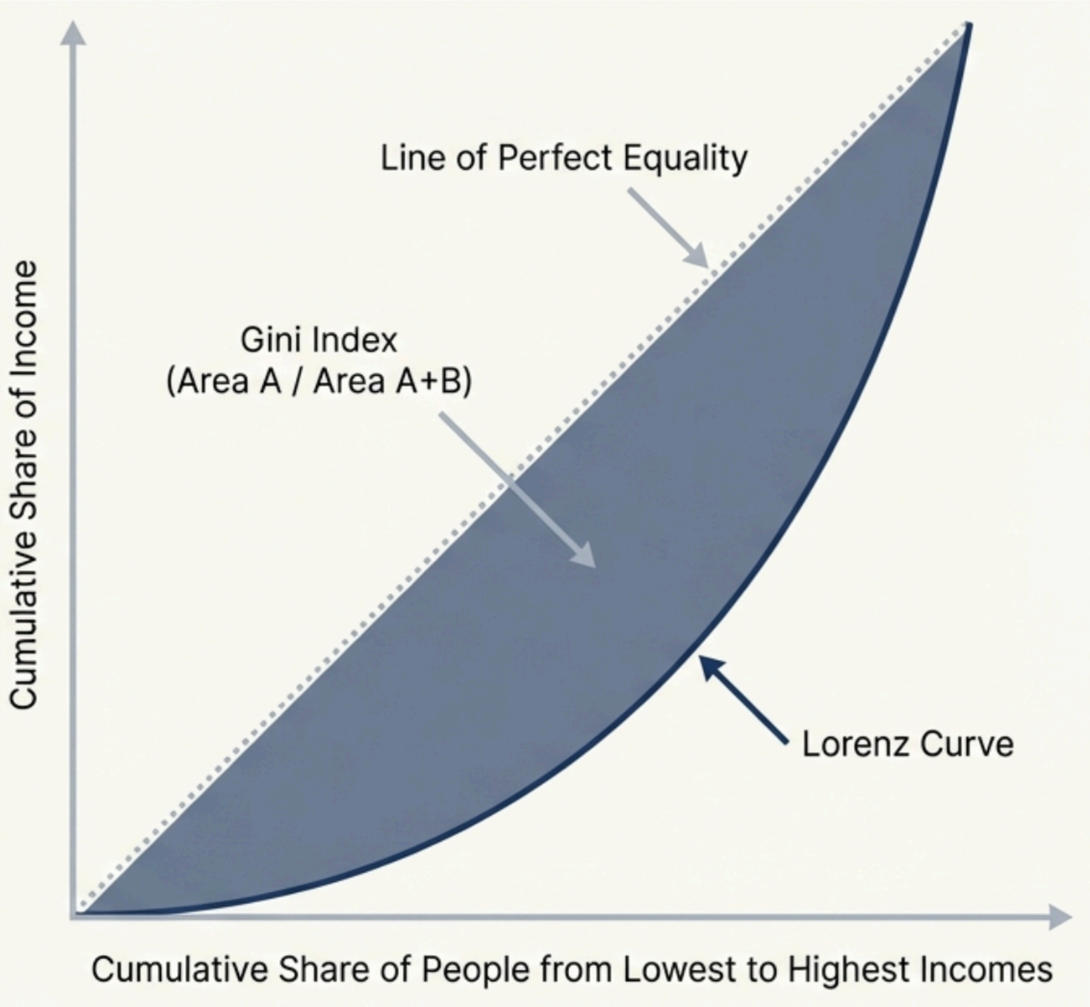
Gini Index Gain: Measuring US Income Inequality

Di Wu, Hsien Tseng Wang & Juan J. DelaCruz
School of Business, Lehman College, CUNY

Part 1: Introduction to Gini Index Gain
Part 2: A Simulated Demonstration
Part 3: US 2003 and 2023 Application

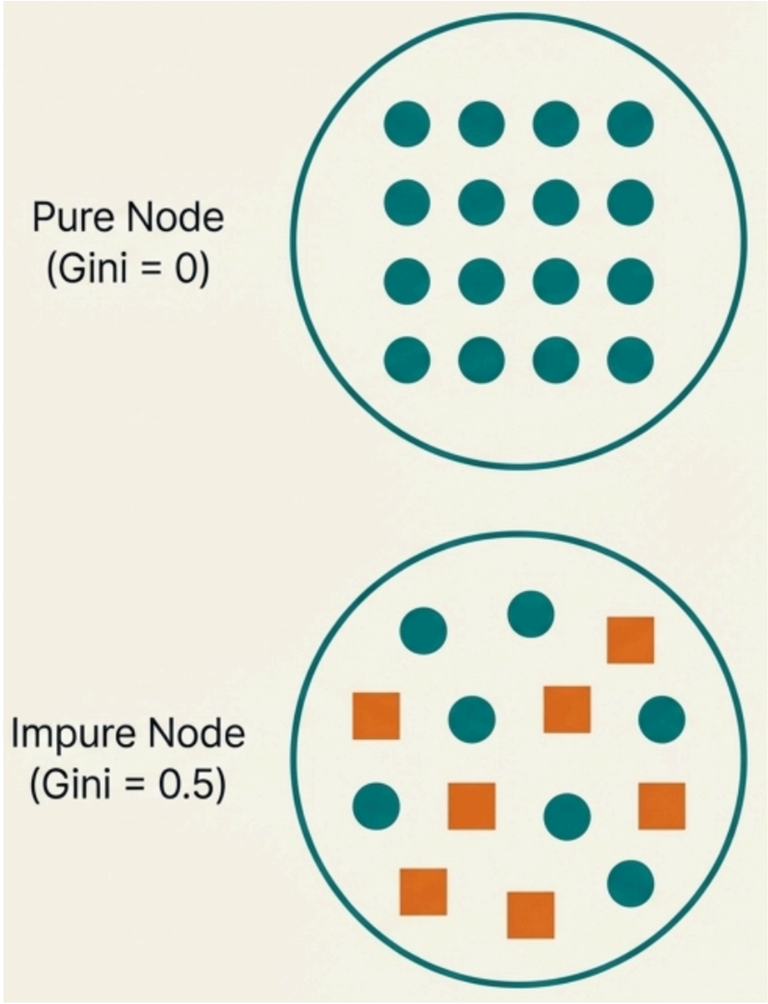
The Gini Index in Economics

- The Gini Index is a standard metric used by economists and policymakers to quantify economic inequality, with a scale from 0 to 1.
  - 0 represents perfect equality.
  - 1 represents maximum inequality.
- The Lorenz Curve visualize it.



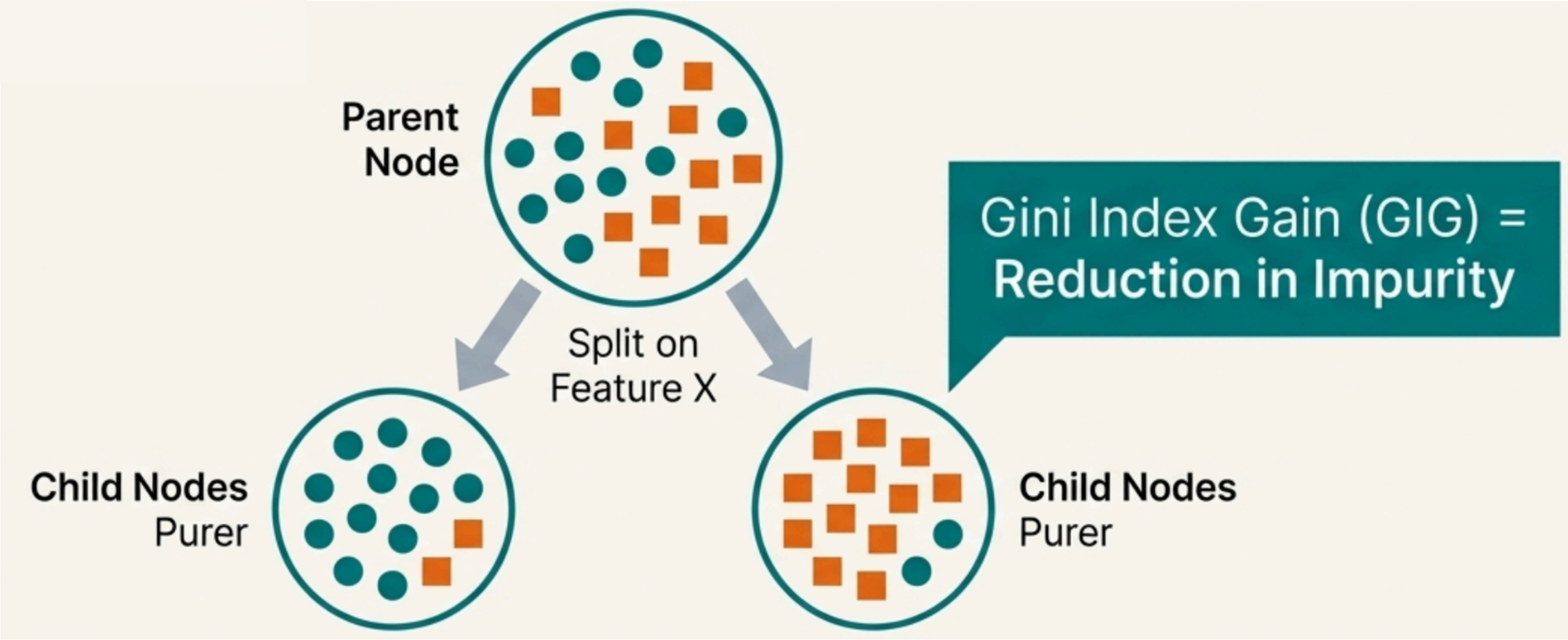
The Gini Index in Decision Tree

- In the field of machine learning, decision tree algorithms use Gini Index to measure impurity.
  - 0 represents a “pure” node.
  - 1 represents a “random” node.



The Gini Index Gain in Decision Trees

- Decision tree splits recursively top-down.

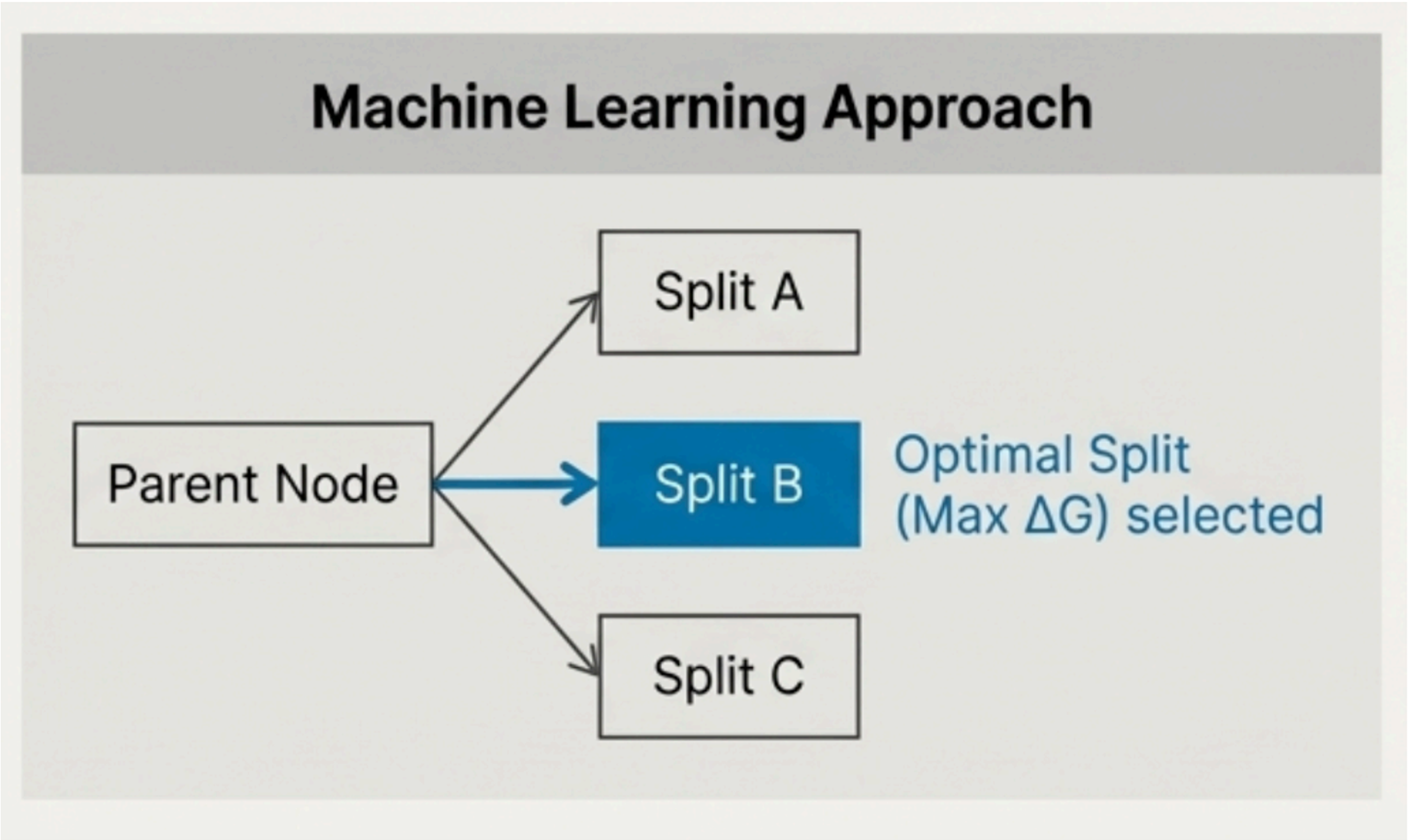


The Gini Index Gain Calculation

- Let $Gini(D)$ be the Gini Index of a dataset *D*.
- $|D|$ denotes the number of elements in dataset *D*.
- Let $D_i(i = 1,2,...,n)$ be the n subsets splitting the dataset *D* by attribute *A*.
- $Gini_A(D) = \Sigma_{i=1}^{n} \frac{|D_i|}{|D|} Gini(D_i)$
- $\Delta G_A = Gini(D) - Gini_A(D)$

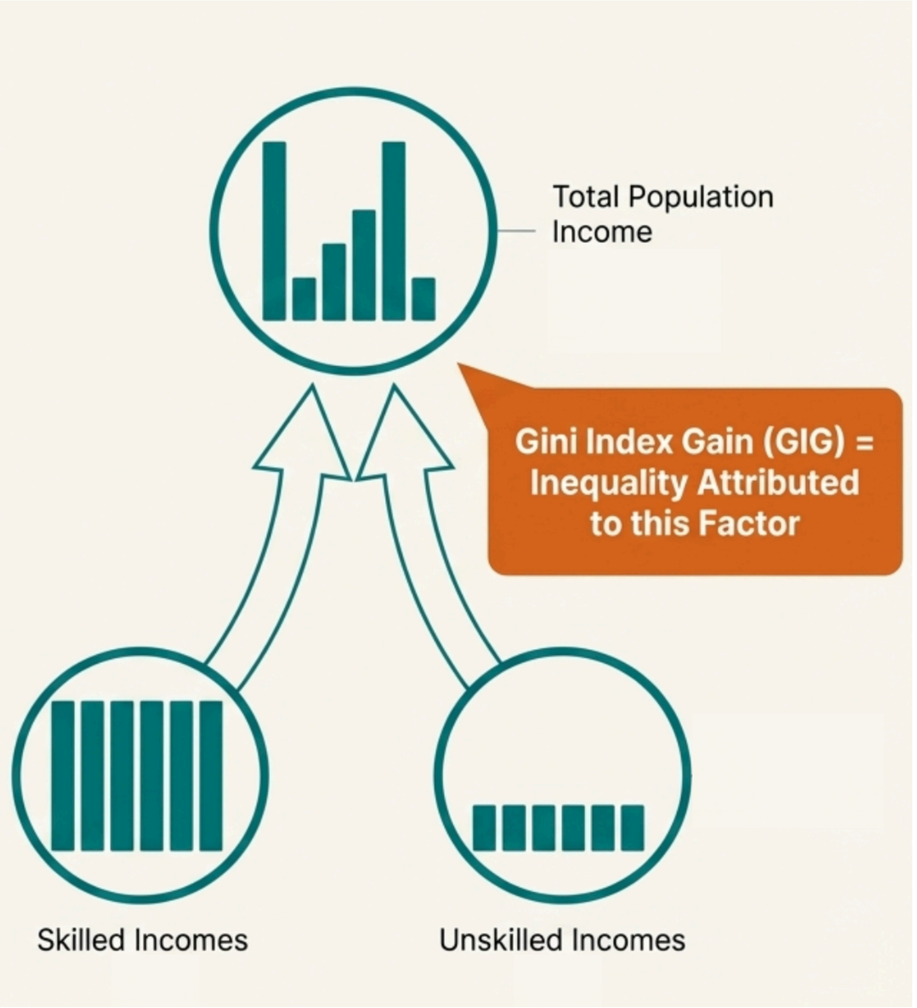
The Gini Index Gain

- The goal of decision tree is to construct a tree from root to leaf that gets purer and purer.
- The algorithm chooses the feature split that results in the max reduction in impurity.
- This reduction is called the Gini Index Gain.



The Gini Index Gain in Economics

- We repurpose the Gini Index Gain to measure the inequality contributed by merging subgroups.
  - For example, if we split a population by education (Skilled or Unskilled), the resulting Gini Index Gain reveals how much the skill-based divide contributes to overall income inequality.

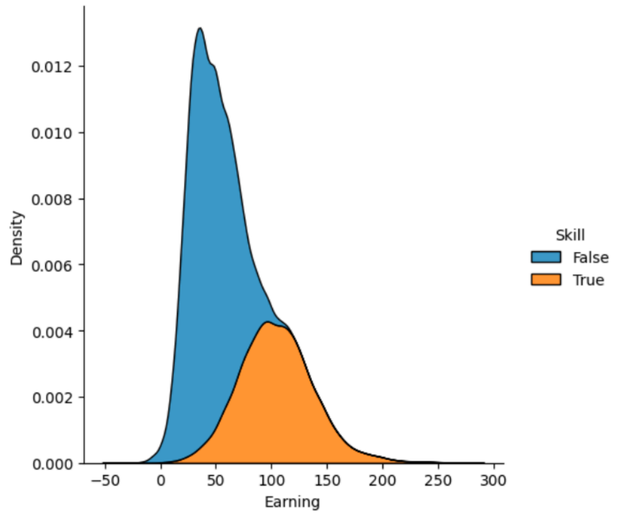


Part 1: Introduction to Gini Index Gain
Part 2: A Simulated Demonstration
Part 3: US 2003 and 2023 Application

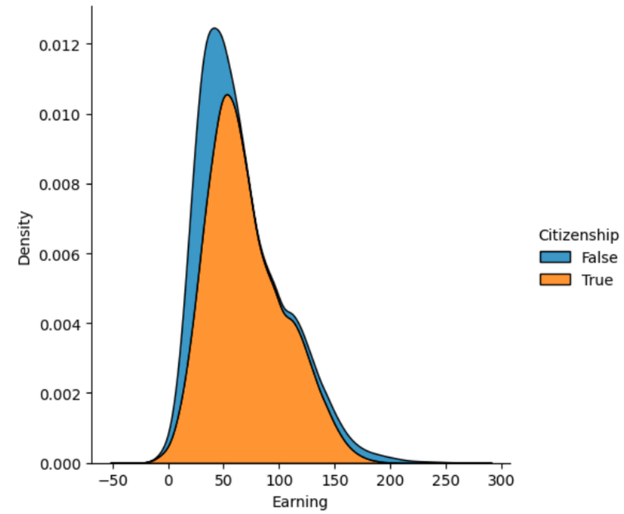
A Simulated Population

- We simulated four distinct social groups:

(U, sigma, n)	Skilled	Unskilled
US-Born	100, 30, 10000	50, 20, 15000
NonUS-Born	150,40,1000	30,10,5000



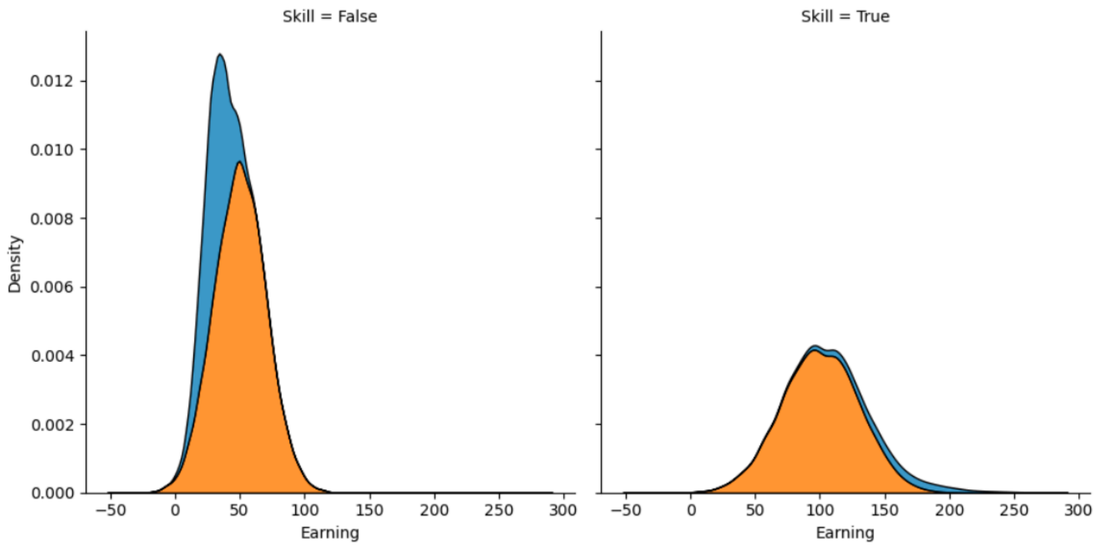
(b) Skill (T/F)



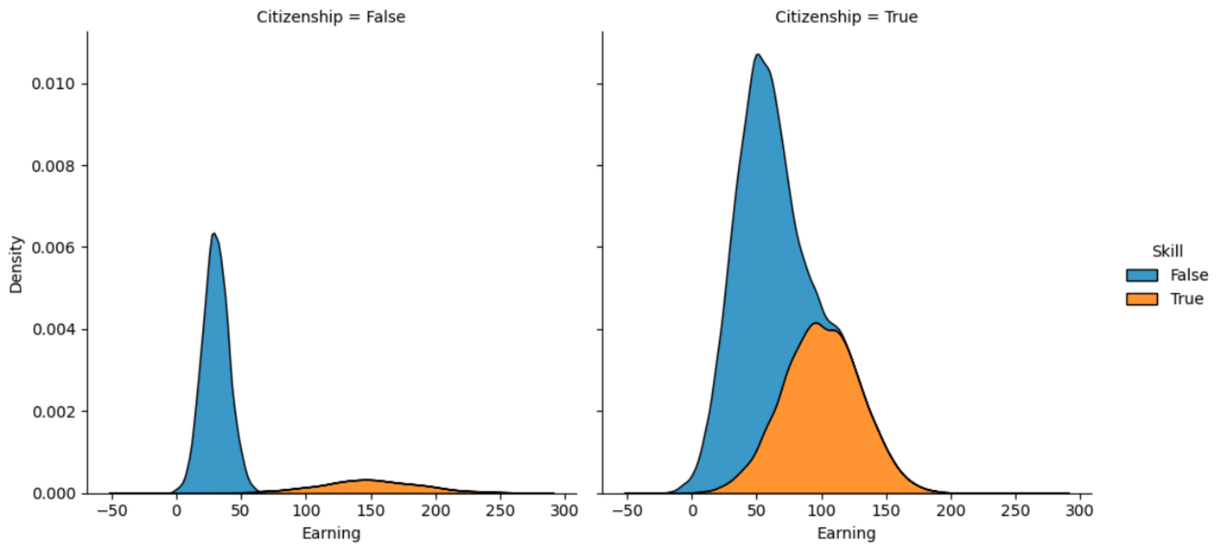
(c) Citizen (T/F)

A Simulated Population

- Splitting by Skill and By Birthplace leads to different results.



(a)



(b)

Applying Gini Index Gain

- Calculating the Gini Index Gain reveals Skill drives more inequality.

Table 1—ΔG for the *Skill*.

Dataset	N	Gini
$Gini(D)$	31000	0.3204
$Gini(Skilled)$	11000	0.1808
$Gini(UnSkilled)$	20000	0.2516
$Gini_{Skill}(D)$		0.2265
ΔG_{Skill}		**0.0940**

Table 2—ΔG for the *Citizenship*.

Dataset	N	Gini
$Gini(D)$	31000	0.3204
$Gini(Citizen)$	25000	0.2780
$Gini(Foreigner)$	6000	0.4237
$Gini_{Citizenship}(D)$		0.3062
$\Delta G_{Citizenship}$		**0.0142**

Validation

- We compared our splits using four established statistical metrics.

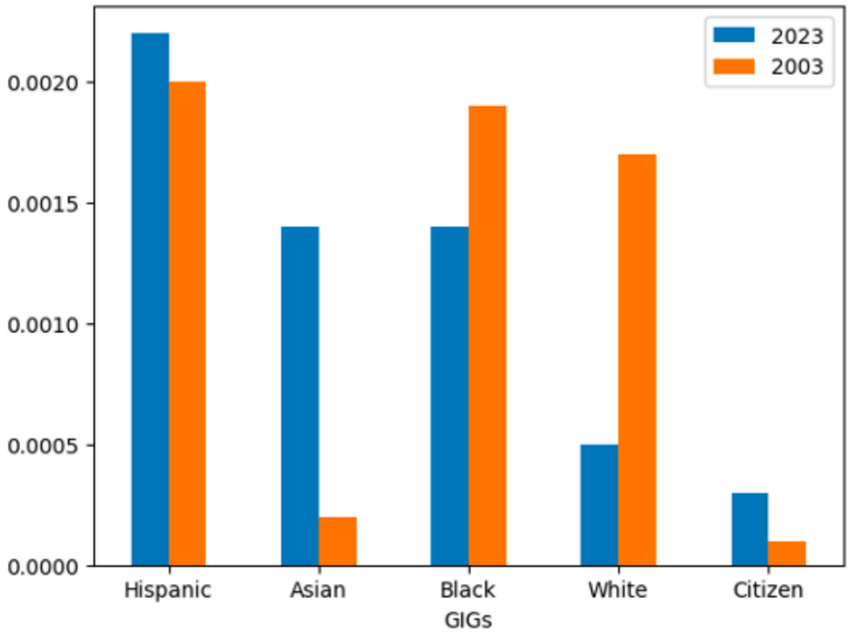
Table 3—Validation Results.

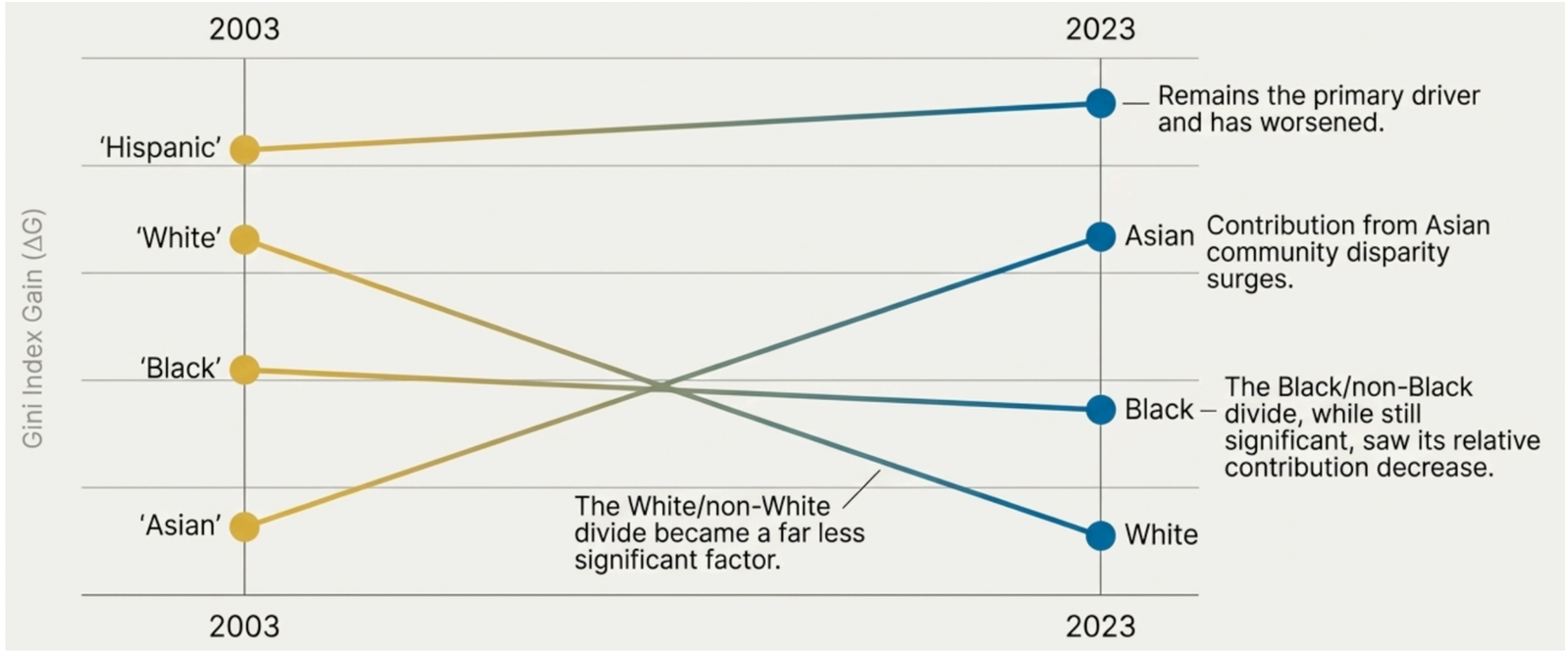
Metric	By Skills	By Citizenship	Passing Threshold
Premium Ratio	2.31 (pass)	1.41 (pass)	> 1.2
Distribution Overlap	0.27 (pass)	0.46 (fail)	< 0.35
Mann-Whitney U Test p-value	0.00 (pass)	0.00 (pass)	< 0.05
Population Balance Ratio	0.55 (pass)	0.24 (fail)	$[0.3, 0.7]$

Part 1: Introduction to Gini Index Gain
Part 2: A Simulated Demonstration
Part 3: US 2003 and 2023 Application

US Census Data 2003 VS 2023

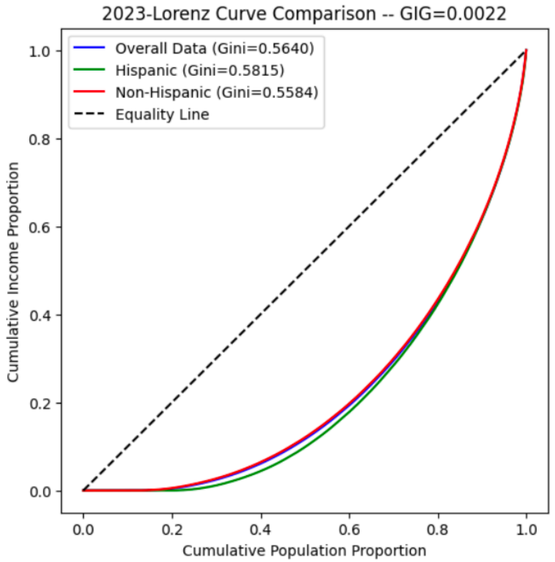
- We applied the ΔG to 2003 and 2023 U.S. census data to compare inequality drivers across race/ethnicity, citizenship, and education.



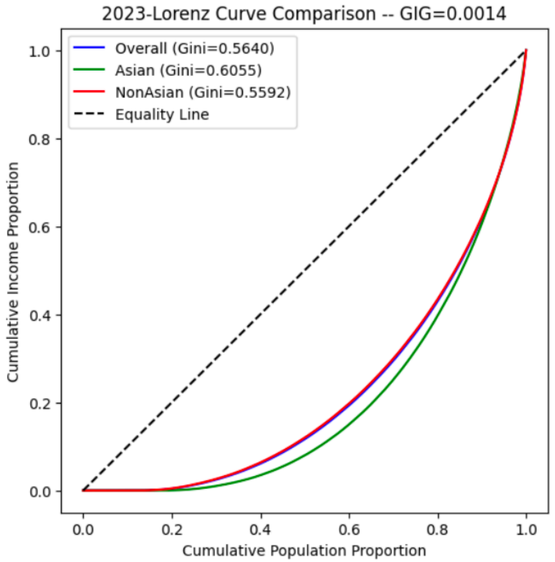


Gini Index Gains by Ethnicity

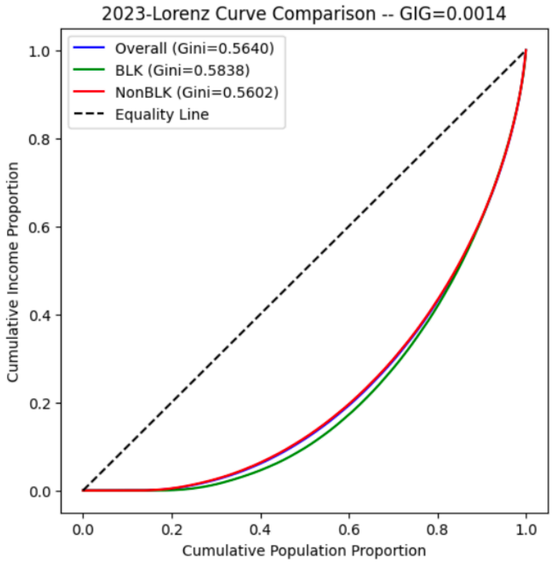
- We compared the Gini Index Gains of splitting by ethnicity.



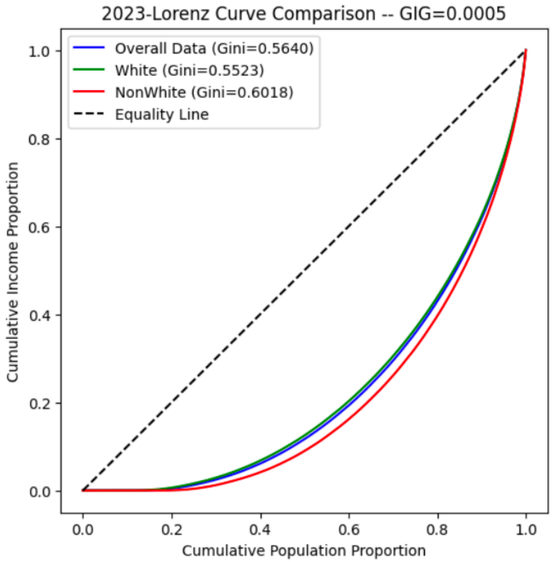
(a) Hispanic



(b) Asian



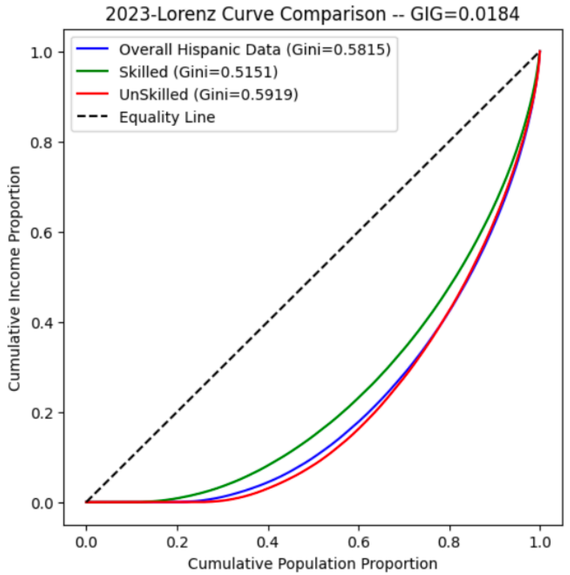
(c) Black



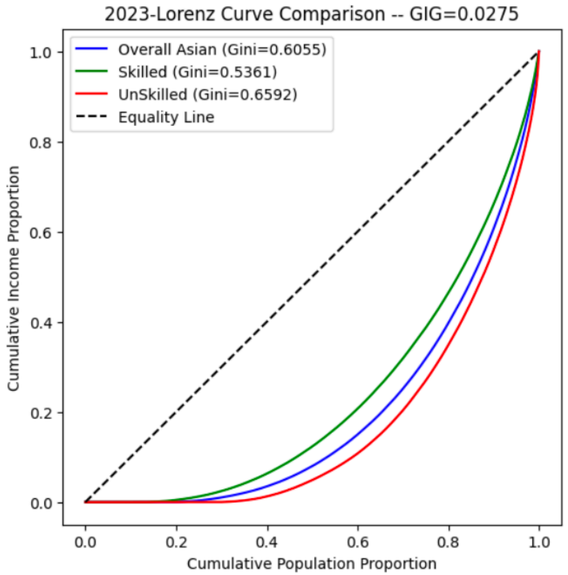
(d) White

Gini Index Gains by Education

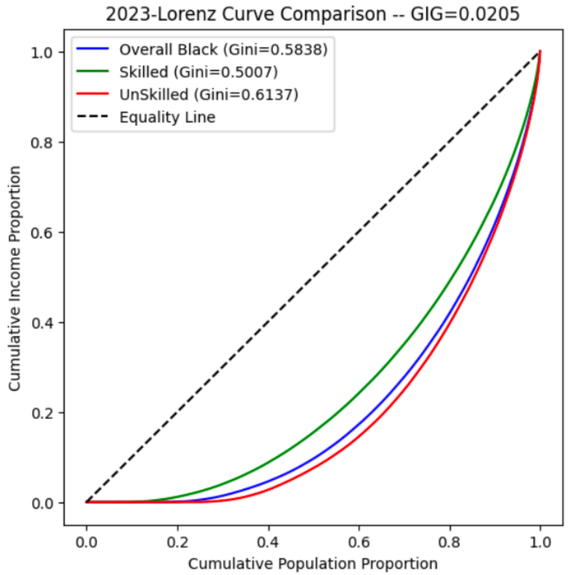
- We further compared the Gini Index Gains of splitting by education within each ethnicity group.



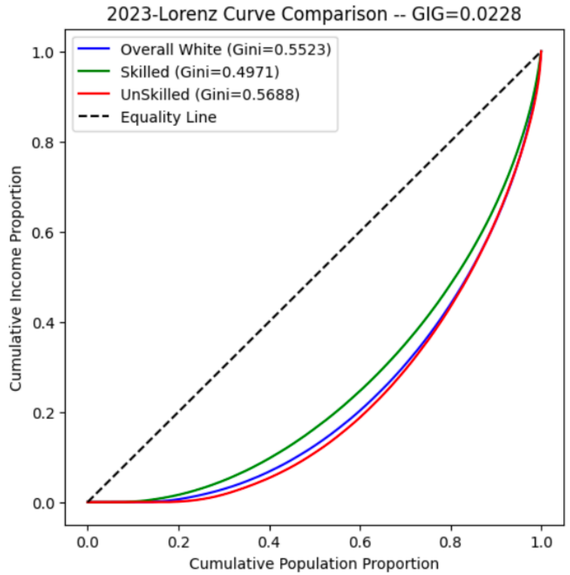
(a) Hispanic



(b) Asian



(c) Black



(d) White

Conclusion

- The Gini Index Gain is a valid tool to measure the inequality introduced by certain attributes within a group.
- The comparison of the Gini Index gains of ethnic groups in 2003 and 2023 shows the dynamics of population.
- We can further improve the tool by optimizing the threshold of splitting, other metics of splitting, etc.
- Applying the Gini Index Gain to more years may result a detailed trend analysis.

Question?

Thank you!