WE ALL COUNT

project for equity
in data science

Foundations of Data Equity

WE ALL COUNT

project for equity
in data science

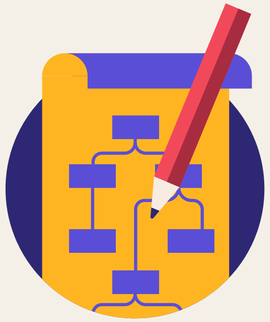
Data is Objective

Data is Not Objective

Data Equity Framework

Funding

Motivation



Project Design

Data Collection & Sourcing

Analysis

Interpretation

Communication & Distribution

The Data Equity Framework addresses equity issues systematically in each step of a data project. Some form of these steps is universal to all types of data projects.

Funding

Motivation

Project Design

Data Collection & Sourcing

Analysis

Interpretation

Communication & Distribution

Tools for Data Collection & Sourcing

Craft a plan for addressing Social Construct (Demographic) Data.

Create or locate a Data Biography for any data you use or generate.

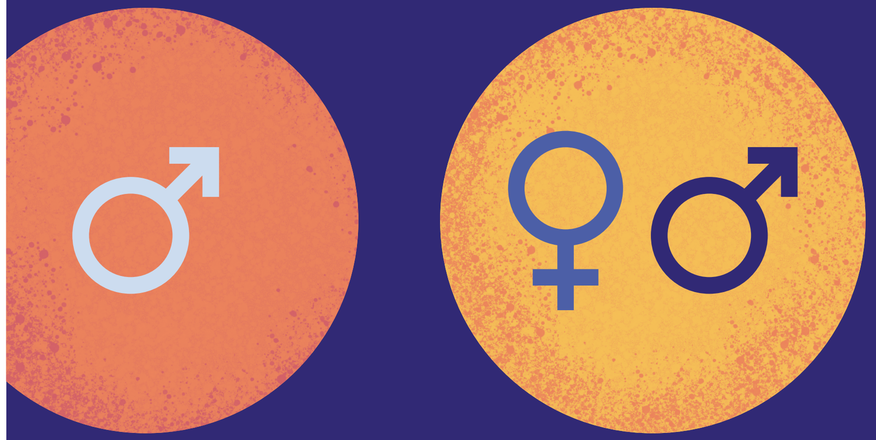
Measuring social constructs (Demographics)

And how are you going to use this?

Disaggregating data is essential....

.....but can lead to serious equity problems.

Measuring social constructs (Demographics)



Intersectionality

Proxies

Small Sample Sizes

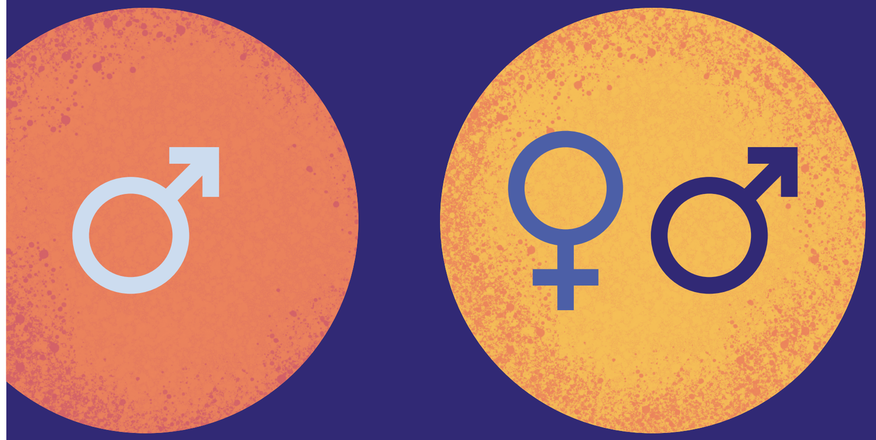
Percent having access to adequate academic support

Male Students 89%
Female Students 71%

Black Students 87%
White Students 76%

Black Male Students 72%
White Male Students 95%
Black Female Students 53%
White Female Students 85%

Measuring social constructs (Demographics)



Intersectionality

Proxies

Small Sample Sizes


PROXIES
WE ALL
COUNT

It’s the only data you have

You’re trying to understand the effect race is having an a trend or experience

You’re trying to talk about racism

You’re trying to show that race itself actually causes something

It’s the only data you have...

It’s not uncommon administrative data to collect race and gender and that’s pretty much it, for demographics.

It is possible to use this data to get a sense of who is using your programs, however, the potential for incorrect results and poor decisions around this are multitude.

Check out our video on Simpson’s Paradox. You’re really telling your data what to say rather than exploring what it actually says in a case like this. If we only have data on race, we tend to see things like ‘it’s a race thing!’.

You’re trying to understand the effect race is having an a trend or experience...

When you have data that shows that different races are experiencing something differently, it seems logical to say so. For example, school district data that shows that children of color are, on average, getting lower test scores than white children. It is not inherently incorrect to say this. What is important is how you say it. If we have data that shows that children of color are getting lower test scores, race here is a proxy. There is nothing about race that is directly causing the lower test scores. It’s important to make this clear and dig deeper into discovering exactly WHAT race is a proxy for in the data.

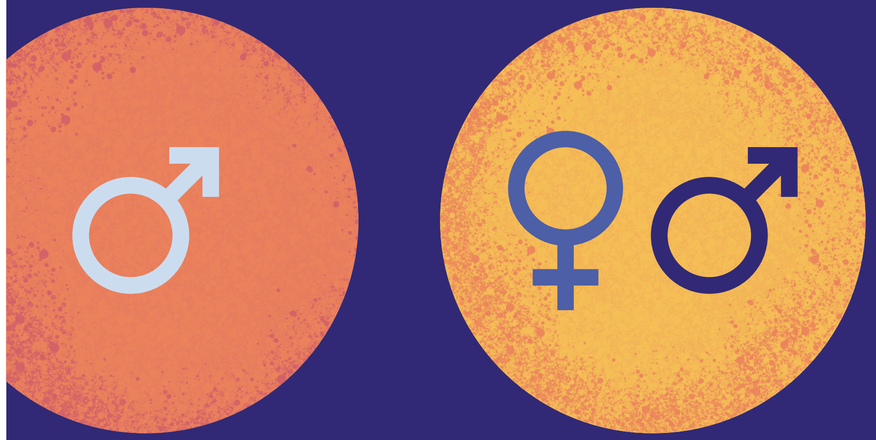
You’re trying to talk about racism...

Lots of times we want to include a race variable in our data project because we want to highlight the effects and impacts of racism. This is another case where things are tricky. By using race as a proxy for racism we can accidentally place the locus of power in the wrong place and encourage more racism. Also by using race as a proxy we mathematically homogenize the experience of an entire group of diverse people, which, of course, is part of racism.

You’re trying to show that race itself actually causes something...

When you’re doing causal data work and you would like to show that race itself is actually causing something, you might be no longer using race as a proxy. In certain medical studies, using race as a direct variable to test how a medication is being metabolized or how to treat certain diseases most successfully, you may want to use a direct variable of race. However, even then it’s good to check whether you could use a more directly measurable piece of data rather than the social construct of race.

Measuring social constructs (Demographics)



Intersectionality

Proxies

Small Sample Sizes

Example: Research on student experiences.

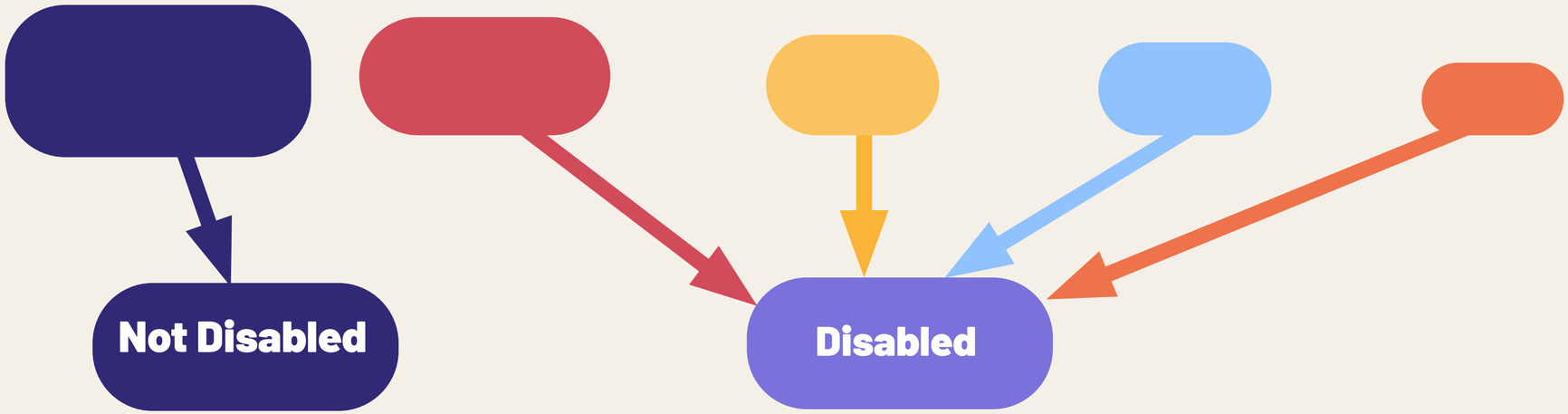
Let's say you are at an Educational Service Unit (ESU) and administering a survey to parents of high schools students and interested in the differences in parent perceptions for parents who have students with disabilities and parents who have students without disabilities.

You decide to include the terms the Nebraska Student and Staff Record System (NSSRS) uses to classify students with disabilities for a question about whether their student has disabilities or not instead of asking a yes or no type question. The survey includes options such as Emotional Disturbance, Deaf-Blindness, Hearing Impaired, Multiple Impairments, Orthopedic Impairment, Other Health Impairment, Specific Learning Disability, Speech Language Impairment, Visual Impairment, Autism, Traumatic Brain Injury, Intellectual Disability).

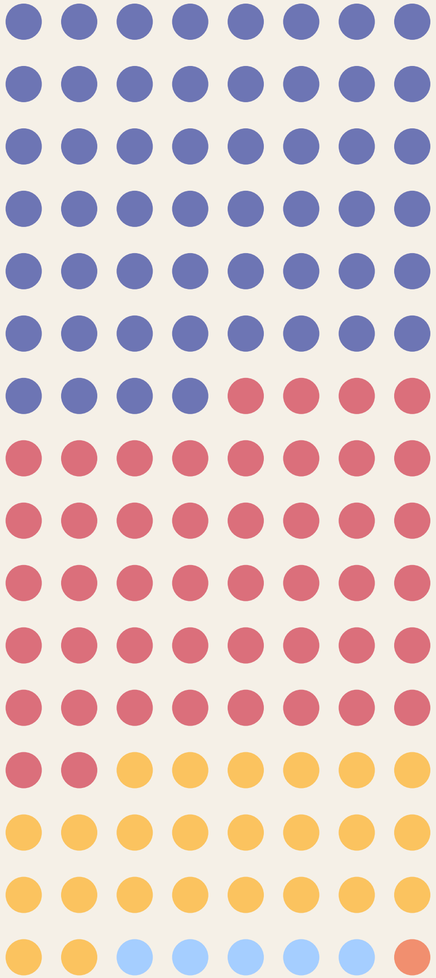
You get very few student disability responses in some categories.

To aggregate or not?

Option 1: Collapsing



- Puts people back in the “other” or “not normal” box.
- Diminished trust.
- Mathematically weaker.



Option 2: Not Collapsing

- Have to report very wide confidence intervals.
- Weakens the voice of the smallest groups.

What You Can Do:

- Decide on how to deal with this before distributing your data collection tool and before analysis.
- Report your results in more than one way, including collapsed, uncollapsed, and hybrid perspectives.
- Be transparent about the dilemmas, compromises and choices you face with your data team, your survey respondents, and your audience.

What You Should Stop Doing:

- Using the word “other”. There are better terms for combined or catch-all categories that don’t have dehumanizing connotations.
- Discounting small sample sizes with the words “not statistically significant”. Just because you have wide confidence intervals doesn’t indicate the data doesn’t have meaning.

Data collection is a very social and relational activity.

It is deeply embedded with subjectivity.

Two focuses today:
1. Social Constructs
2. Data Biographies

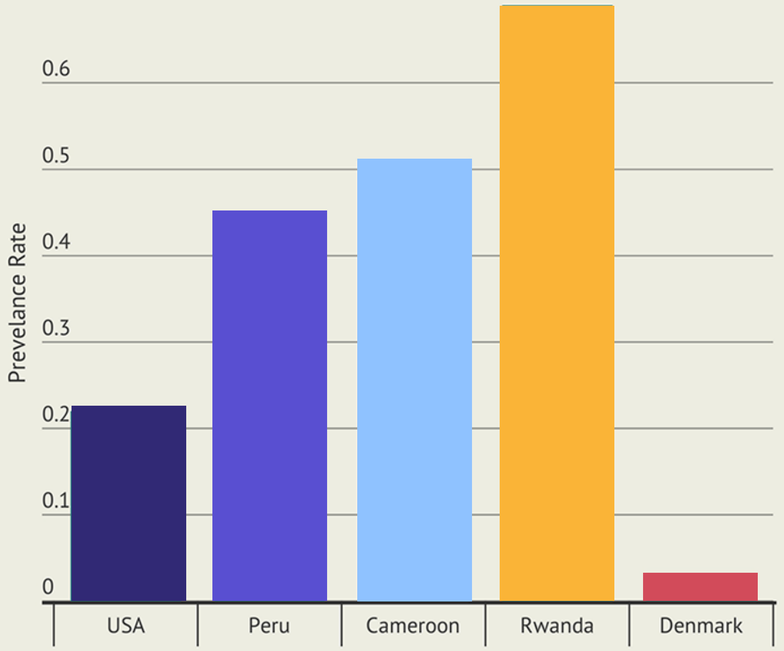
Data Biographies must accompany each dataset you are using and include at minimum:

When What Who Why How Where

Original Graph

Here, each country's data is from a different survey with different questions, criteria and definitions.

"Intimate Partner Violence Rates"

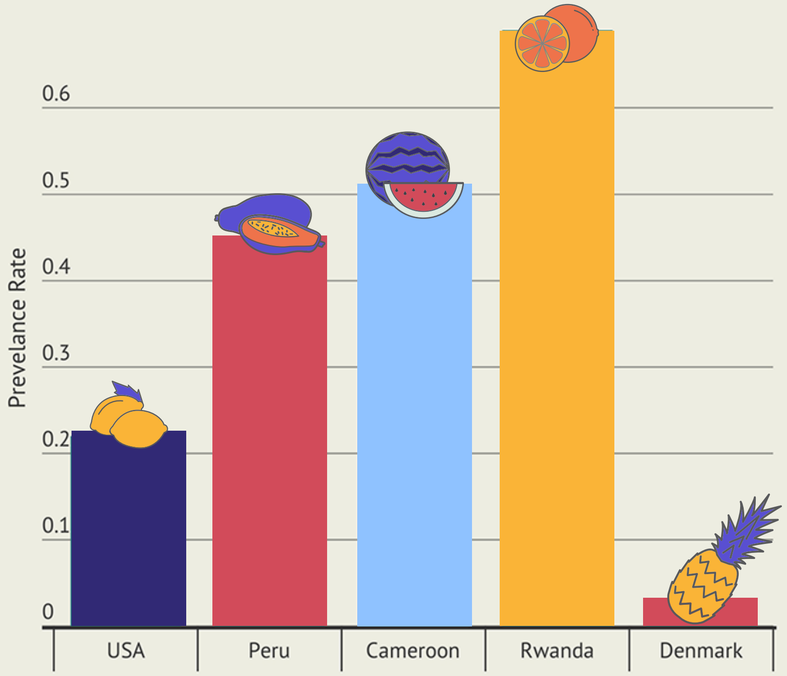


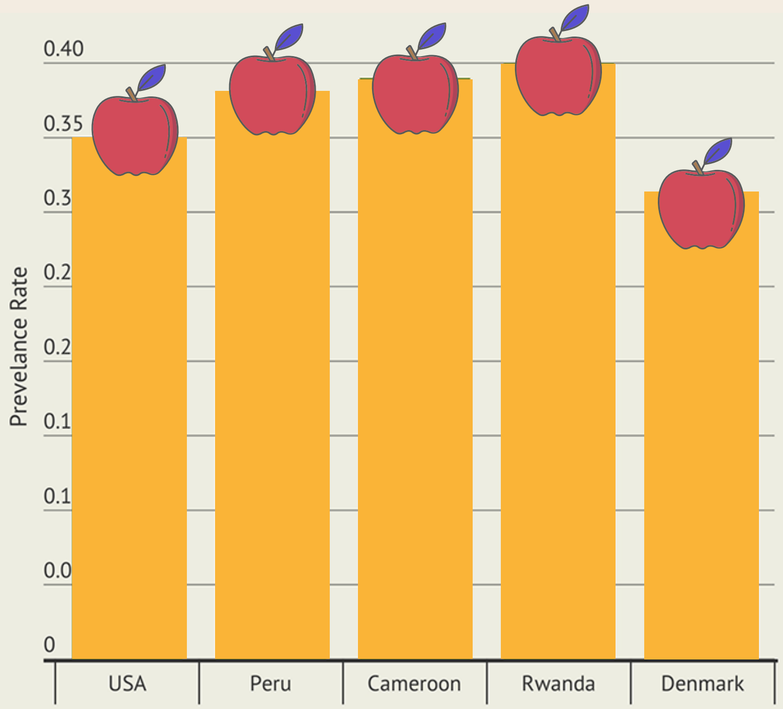
Some datasets counted women aged 18-60, while some counted ages 14-80.

Some datasets were self-reported incidents, some were asked by survey takers, and others were based solely on police reports.

ETC.

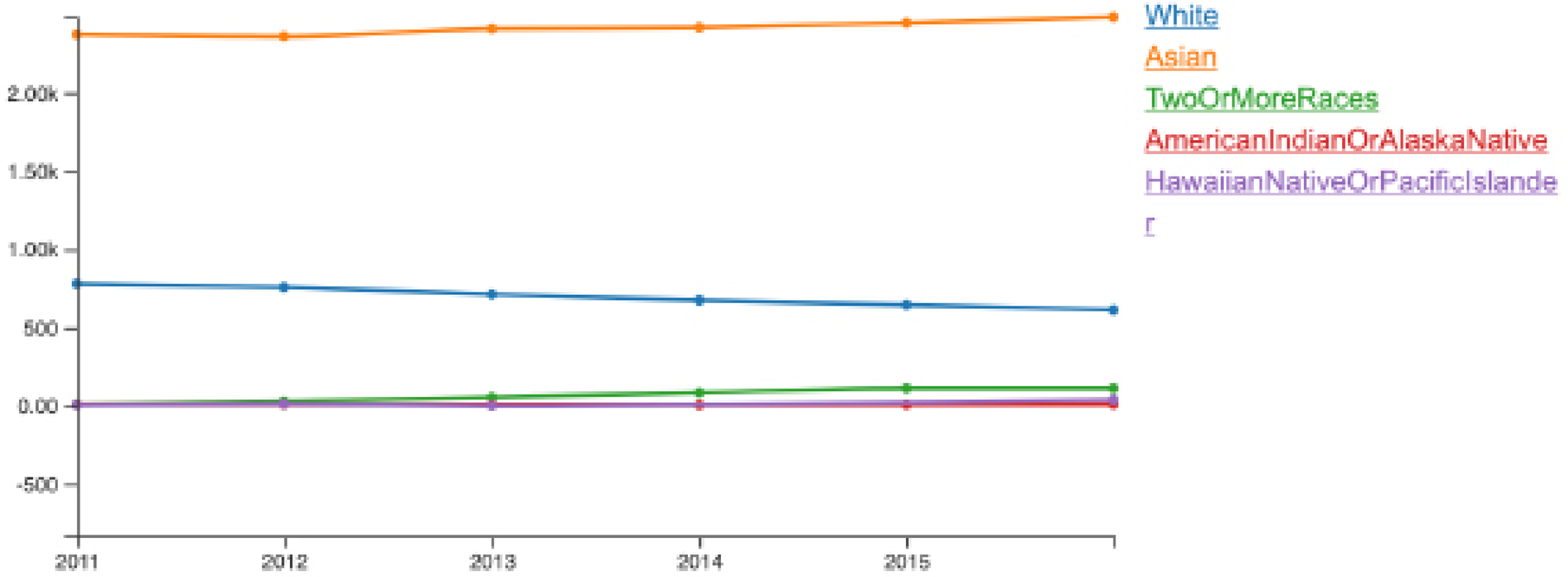
A Data Biography is a tool to help identify equity issues when apples are compared to oranges, and they are a tool that can help fix issues when sourcing from multiple datasets or trying to supplement your dataset with others.





count of Student by race

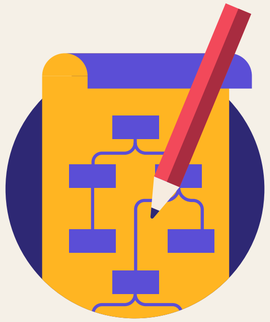
40 observations



Sources of bias can be identified in each step of the data life cycle.

Funding

Motivation



Project Design

Data Collection & Sourcing

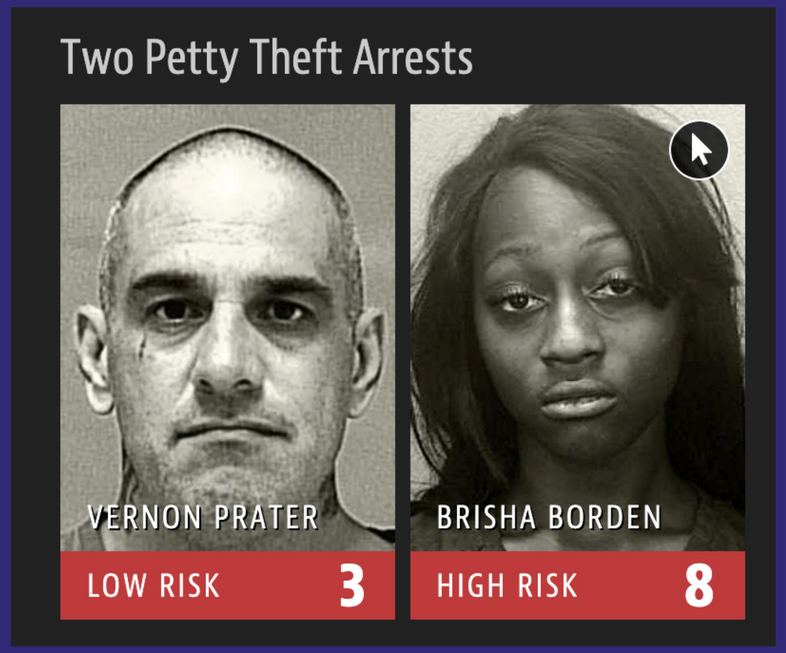
Analysis

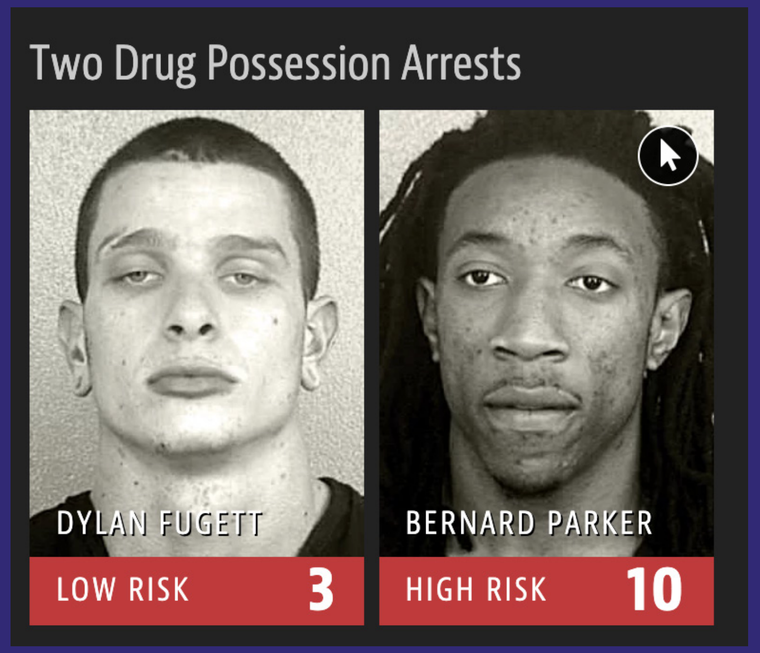
Interpretation

Communication & Distribution

Interpretation

COMPAS gives a score that predicts how likely it is that this person will reoffend.





~~COMPAS gives a score that predicts how likely it is that this person will reoffend.~~

COMPAS gives a score that predicts how likely it is that this person might be in contact with the police again, be arrested by those police and not have the money for immediate bail release.

Results Stage

This is the end of the Analysis Step (5).

Results are represented as numbers.

This is where we apply meaning by seeing what our methodology can tell us and how we can interpret the analysis based on our perspective.

Interpretation Stage

This is the end of the Interpretation Step (6).

The numbers have meaning and a narrative to explain them.

“85% of our respondents reported feeling happier on the days when they had eaten breakfast.”

“85% of our respondents reported feeling happier on the days when they had eaten breakfast.”

This is at the result stage. It only states the exact result from a specific question; it doesn’t try to say why breakfast made them happier, and it only addressed the people that were directly asked.

“Eating breakfast will make you happier.”

“Eating breakfast will make you happier.”

This is an interpretation of the results from the previous question. It’s very broad. It has taken the results and moved them into the stage of interpretation.

“The study shows that drinking red wine three times a week lowers the risk of heart disease.”

“The study shows that drinking red wine three times a week lowers the risk of heart disease.”

This is an interpretation. It has taken the results of whatever method they used to test the effects of red wine on heart disease and applied a causal ‘this causes that’ meaning. If you are curious about this interpretation, that’s good; it puts the emphasis on how they arrived at this conclusion and whether or not they have a good scientific basis for such a strong statement. It doesn’t mean it’s wrong.

“Our study shows that being black puts you at the highest risk for adult illiteracy.”

“Our study shows that being black puts you at the highest risk for adult illiteracy.”

This is an interpretation. The result of the count of black people with literacy challenges might be accurate, but the meaning inferred from the result is presented here with a specific perspective. This tips us off to the fact that it’s in the ‘interpretation stage’.

Our whole message is that THIS IS A STEP. BE AWARE THAT YOU ARE INTERPRETING.

Do it intentionally rather than subconsciously, and support your interpretation.

A strongly supported interpretation is much more powerful than hiding behind an unsupported claim that ‘science’ makes this ‘objectively true’.

What narrative are you choosing?

After deciding what the results *mean*, in order to communicate them, we often present them from a *perspective*.

A narrative is the combination of meaning and perspective.

Applying a narrative is saying “this is what this **means** from **this perspective**“.

Thank you.

weallcount.com

Heather Krause, PStat

support@weallcount.com

@datassist