Foundations of Data Equity
The worst equity problem facing data science is that people are making prejudiced choices but they don’t know it.
We want a “silver bullet” against bias, prejudice, human error, and injustice.
What is the average number of students across these three classrooms?

CLASSROOM A: 3

CLASSROOM B: 6

CLASSROOM C: 9

Average: (3 + 6 + 9) / 3 = 6 students
If you got 6, you’re right.
If you got 7, you’re right.

You just did the same math, but from different perspectives.
From the teachers' perspective:

$$3 + 6 + 9 = 18$$

$$18 ÷ 3 = 6$$
From the students’ perspective:

\[3 + 3 + 3 + 6 + 6 + 6 + 6 + 6 + 6 + 6 + 6 + 9 + 9 + 9 + 9 + 9 + 9 + 9 + 9 + 9 = 126\]

\[126 \div 18 = 7\]
Is our project a success?

Example: Does our project increase average monthly income?

<table>
<thead>
<tr>
<th></th>
<th>Before Project</th>
<th>After Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Monthly Income</td>
<td>$800</td>
<td>$1300</td>
</tr>
</tbody>
</table>
Is our project a success?

- **Zip Code #1**
  - Before Project: $700
  - After Project: $1800

- **Zip Code #2**
  - Before Project: $800
  - After Project: $1400

- **Zip Code #3**
  - Before Project: $900
  - After Project: $700
Is our project a success?

<table>
<thead>
<tr>
<th>Zip Code #1</th>
<th>Before Project</th>
<th>After Project</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$700</td>
<td>$1400</td>
</tr>
<tr>
<td>Zip Code #2</td>
<td>Income Gap: $200</td>
<td>Income Gap: $1100</td>
</tr>
<tr>
<td>Zip Code #3</td>
<td>$1800</td>
<td>$1400</td>
</tr>
</tbody>
</table>

Income Gap: $1100
This isn’t just a math problem. It’s a human problem.
We want to think of data science like this:

**CHOOSE:** What question should we ask?

→ **Objective Data**

→ **Objective Results**

→ **Objective Analysis**

→ **Objective Decision**
It’s a lot more like...

Choose source  Choose effect  Choose research question  Choose scale
Choose covariates  Choose definitions  Choose criteria  Choose methodology
Choose scope  Choose moderators  Choose cutoffs  Choose mechanism
Choose enumerators  Choose project design  Choose denominators
Choose timeframes  Choose mediators  Choose resolution
Choose metrics  Choose controls  Choose language  Choose relationships
Choose dependent variables  Choose data cleaning  Choose collection tool  Choose model
Choose sample  Choose identity categories
THIS TALK CAN'T BE UNSEEN...
There is no ‘right’ choice.

Equity is a process not a binary state.
“We were aware that we were asking them to give up some work time to come to the trainings and we didn’t want them to be penalized for that, so we controlled for it in the model.

We were patting ourselves on the back for being so equity-minded.”
“I can’t wait to get out of this program...”

“You think this was a success?!”

“I’ve barely slept in a month!”

Uh oh...
\[ \hat{Y} = b_0 + b_1 T_1 + b_2 \text{Time}_2 + \ldots + b_p X_p \]

(If you do twice as much work, you’ll get twice as much product.)
The first model wasn’t “wrong”.

But the new model was a better choice because it reflected the equity that we wanted it to.
We got two totally different answers even though we were using:

The same data.
The same research question.
The same methodology.
The same analysts.

Even controlling for the same variables.
Every choice reflects a worldview.

Even the smallest choices can have huge impacts.
Their results are not statistically significant.

Our results are not statistically significant.

Our results contain uncertainty.

Our results contain uncertainty because...
Choosing Data Equity...
The Data Equity Framework

- Funding
- Motivation
- Project Design
- Data Collection
- Analysis
- Interpretation
- Communication
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.
The order of steps reflects the typical data project process, but you can address them in any order.
Even if you can only address some of the steps in your project, it is still worth doing.
What you do in each step of the Data Equity Framework will have equity impacts and interactions in other steps and your project as a whole.

The Data Equity Framework works holistically and individually.
Our plan for this workshop:

1. Funding
2. Motivation
3. Project Design
4. Data Collection & Sourcing
5. Analysis
6. Interpretation
7. Communication & Distribution
Funding
Motivation
Step 3: Project Design
Project Design is the phase where the WHY becomes the HOW.

Critical step in data equity.
‘Geographical Sample Area’ based on whose geography?
RCTs: The Gold Standard ... of What?
Phase 4: Data Collection & Sourcing
Measuring social constructs (Demographics)

Who is constructing the categories?
Whose definitions are we using?
Measuring social constructs (Demographics)

And how are you going to use this?
Step 5: Analysis
Mental Health Risk

In our community, which children are most at risk for mental health issues?

<table>
<thead>
<tr>
<th></th>
<th>Percentage at Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Black Youth</strong></td>
<td></td>
</tr>
<tr>
<td><strong>White Youth</strong></td>
<td></td>
</tr>
</tbody>
</table>
Mental Health Risk

In our community, which children are most at risk for mental health issues?
Mental Health Risk

In our community, which children are most at risk for mental health issues?

- **YOUTH IN POVERTY**
- **YOUTH NOT IN POVERTY**

PERCENTAGE AT RISK

0 10 20 30 40 50
Mental Health Risk

We saw that the Black children were at risk, that male children were at risk, and children not living in poverty were at risk.
About Distribution of Key Predictor Variables

BLACK  WHITE
MALE     FEMALE
NOT POOR   POOR

It seems like you’re letting the data talk.
But really you’re telling the data what to say.
Example:

Project for a school district in Texas.

**Research question:** “What are the trends in Black, Indigenous and Hispanic male dropout rates compared to White male students?

Which of our two interventions are working best?”
Example:

Project for a school district in Texas.

Changed to:

“How much are our interventions effective in removing the barriers we put up to Black, Indigenous and Hispanic boys remaining in school in our district?”
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.

Interpretation Stage

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.

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

The numbers have meaning and a narrative to explain them.
“Our study shows that being black puts you at the highest risk for adult illiteracy.”
Communication & Distribution
Each audience will have a ‘sweet spot’ for how to best distribute your information.
Data Visualization

Ditch your ‘best’ practices for an adaptive, user-oriented system.
Data Viz “best practices” are not culturally universal.
Why does this matter for equity?
Our data reflects how we see the world, but that’s a good thing.

It means we can choose equity.
WE ALL COUNT
project for equity in data science

Thank you!

weallcount.com/dop
hello@weallcount.com
Thank you.

weallcount.com

Heather Krause, PStat

support@weallcount.com

@datassist