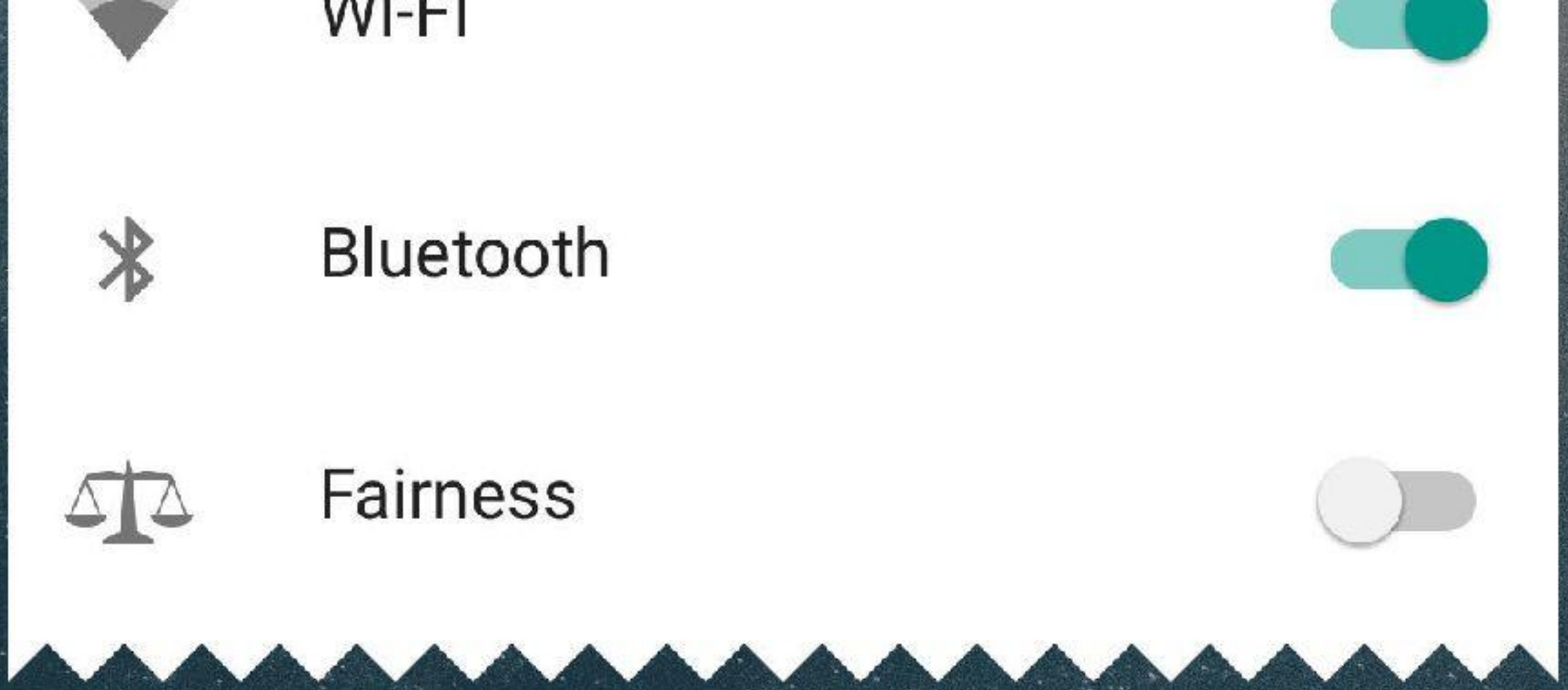


Fair is not the default

go/fair-not-default [redacted] Google Confidential


San Marco Square
Venice


Isn’t it curious how nearly every person’s head
falls along roughly the same horizon line?


It’s the result of the photographer being
average height. From an *outlier’s* perspective,
the world might look significantly less uniform.

This is a talk about the role of humans in machine learning.

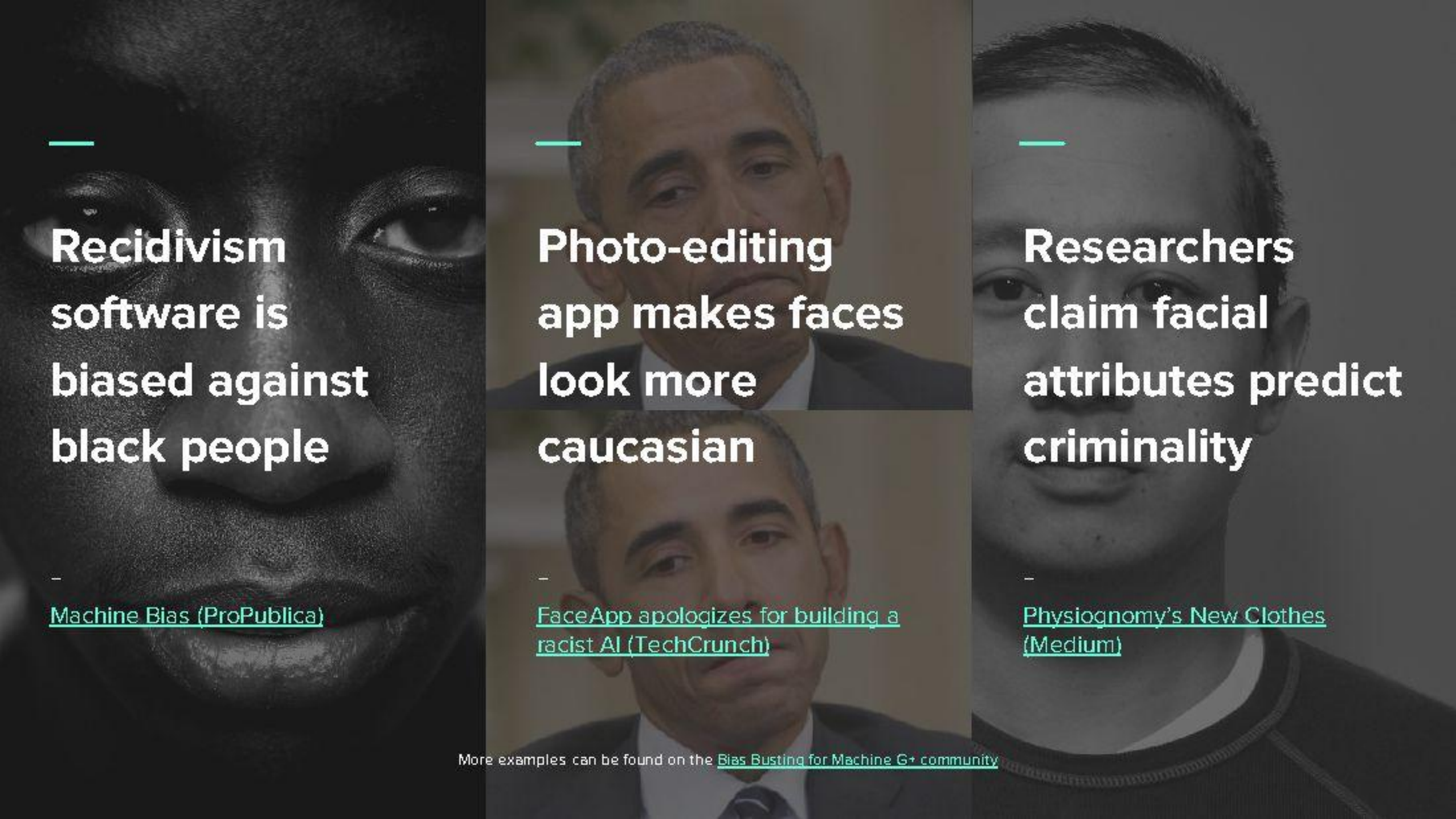
But really it’s a talk about the role of humans in decision making.

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. But these instructions don't come from nowhere. Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."

— The Guardian

"It's true that they can follow instructions at superhuman speed, with superhuman fidelity and over unimaginable quantities of data. But these instructions don't come from nowhere. Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."

— The Guardian



Recidivism software is biased against black people

Machine Bias (ProPublica)

Photo-editing app makes faces look more caucasian

FaceApp apologizes for building a racist AI (TechCrunch)

Researchers claim facial attributes predict criminality

Physiognomy's New Clothes (Medium)

More examples can be found on the Bias Busting for Machine G+ community

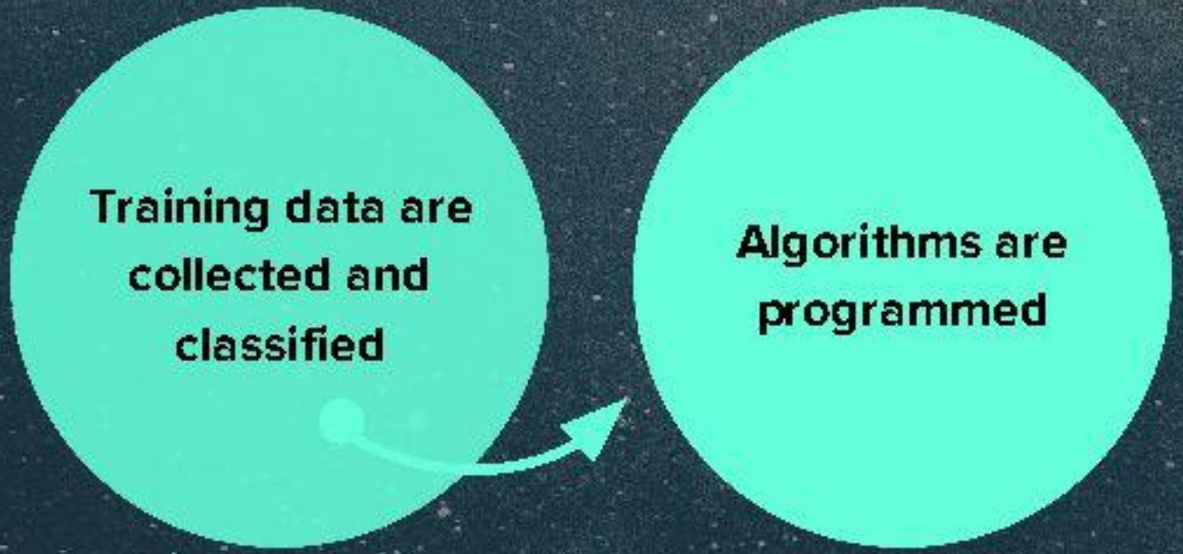
When we presume that human judgment can—or should—be removed from the loop, the result is an unconscious bias network effect.

And we (Googlers) are just as susceptible to this effect as our users.

Training data are collected and classified

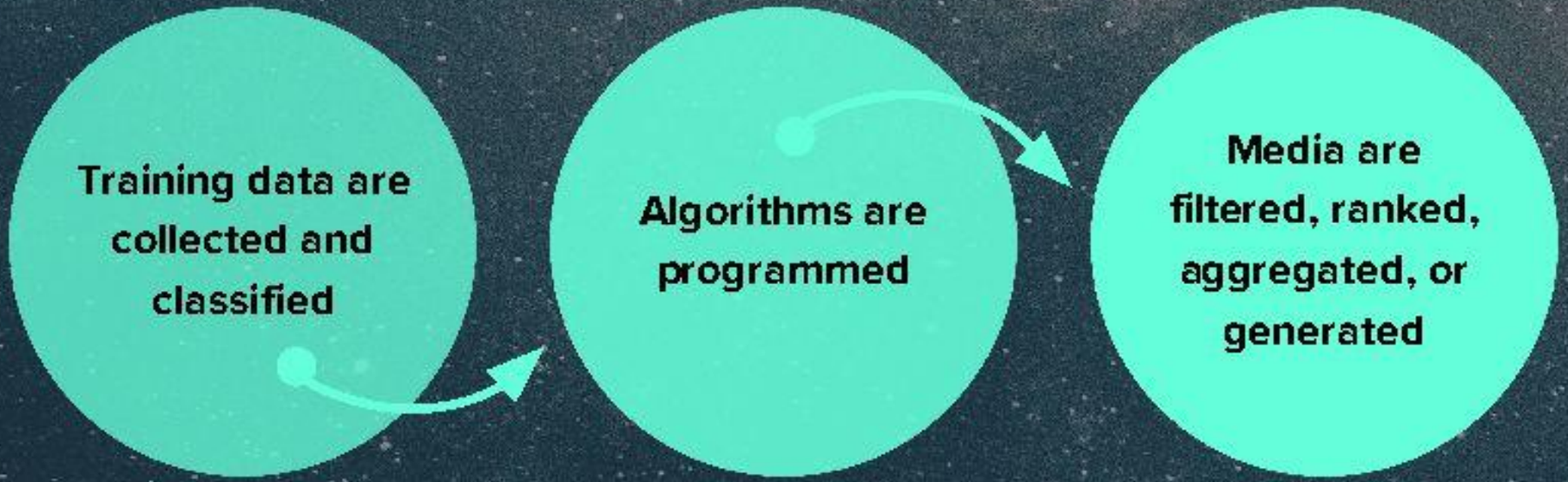
CREDIT

Latent Bias, Blaise Agüera y Arcas



CREDIT

Latent Bias, Blaise Agüera y Arcas



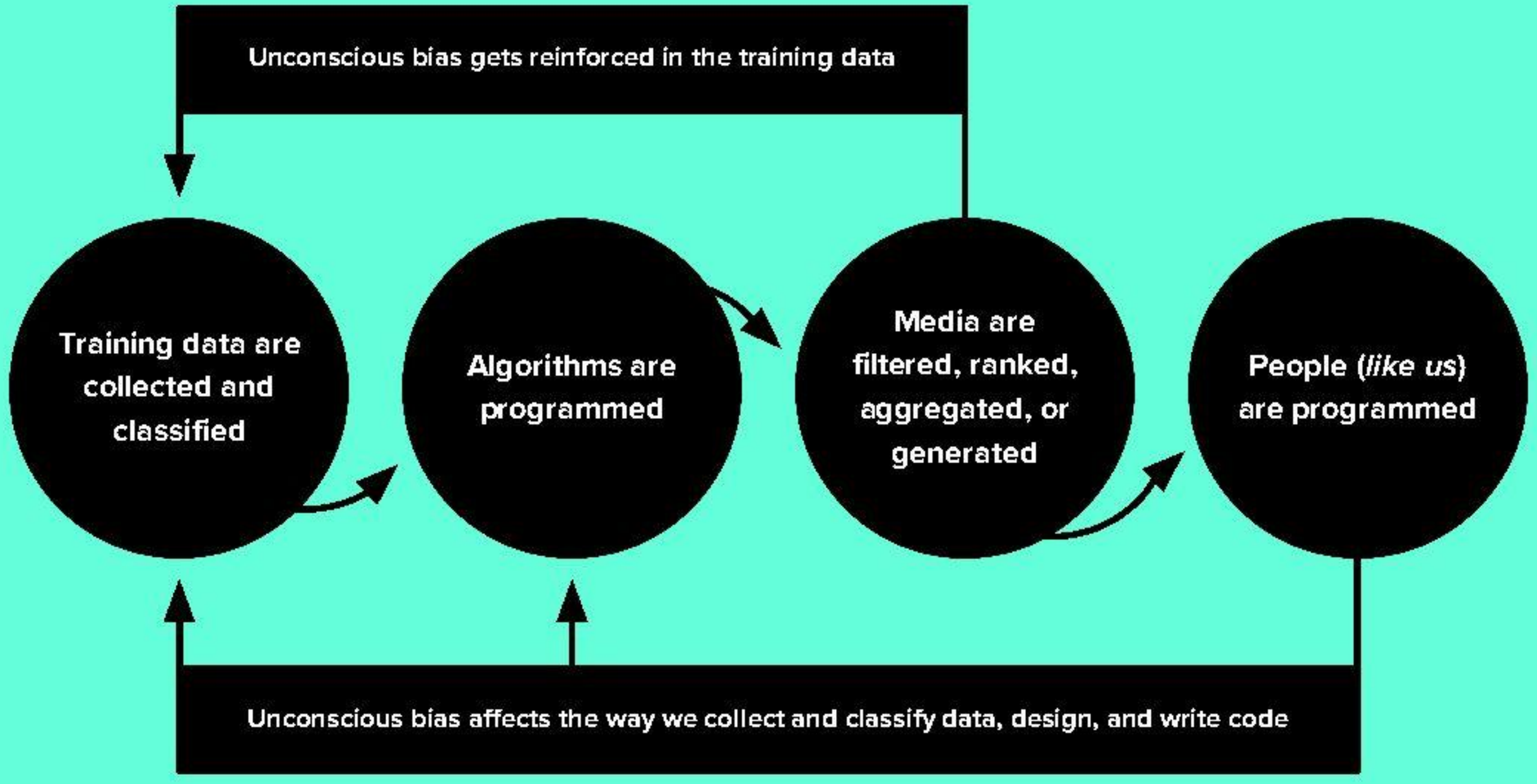
CREDIT

Latent Bias, Blaise Agüera y Arcas



CREDIT

Latent Bias, Blaise Agüera y Arcas

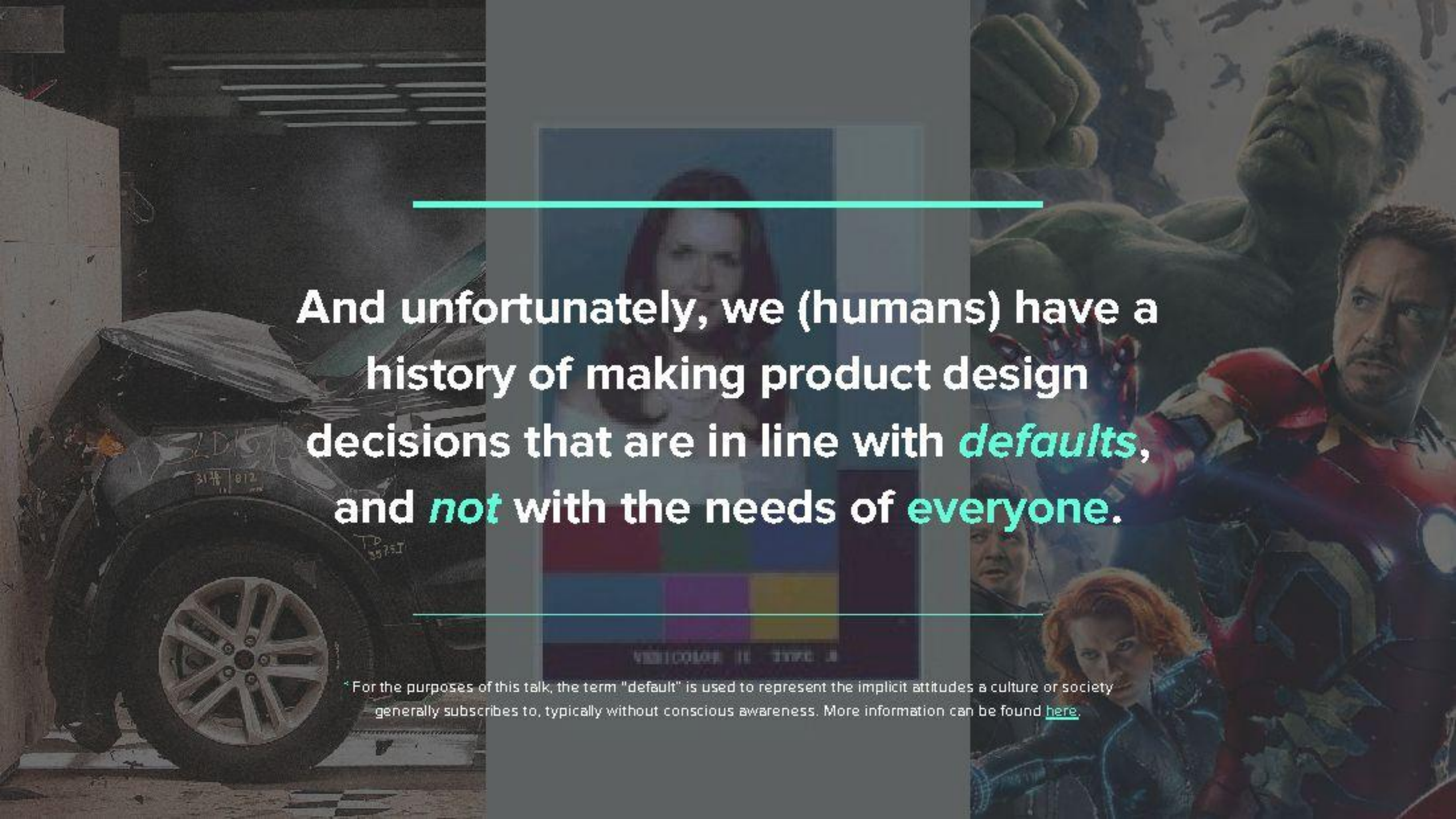

Unconscious bias gets reinforced in the training data
Training data are collected and classified
Algorithms are programmed
Media are filtered, ranked, aggregated, or generated
People (*like us*) are programmed
Unconscious bias affects the way we collect and classify data, design, and write code

If we are going to make the world's information universally accessible and useful, we must strive to make the products we are developing—apps, infrastructure, models, and more—work for everyone.

Learn more at go/ml-fairness

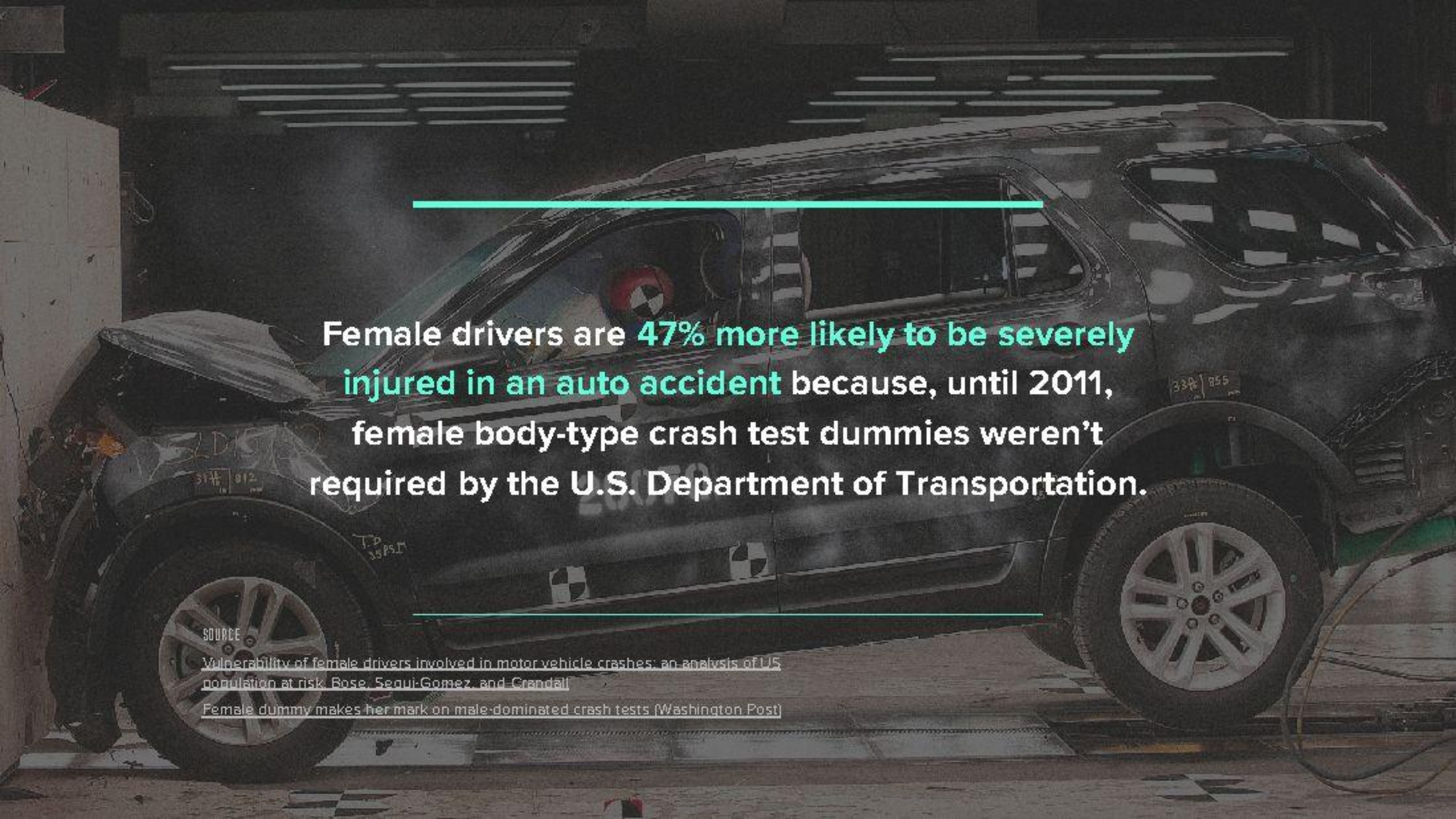
If we are going to make the world's information universally accessible and useful, we must strive to make the products we are developing—apps, infrastructure, models, and more—**work for everyone.**

Learn more at go/ml-fairness



And unfortunately, we (humans) have a history of making product design decisions that are in line with *defaults*, and *not* with the needs of everyone.

*For the purposes of this talk, the term "default" is used to represent the implicit attitudes a culture or society generally subscribes to, typically without conscious awareness. More information can be found here.



Female drivers are **47% more likely to be severely injured in an auto accident** because, until 2011, female body-type crash test dummies weren't required by the U.S. Department of Transportation.

SOURCE

Vulnerability of female drivers involved in motor vehicle crashes: an analysis of US population at risk, Bose, Segui-Gomez, and Crandall

Female dummy makes her mark on male-dominated crash tests (Washington Post)



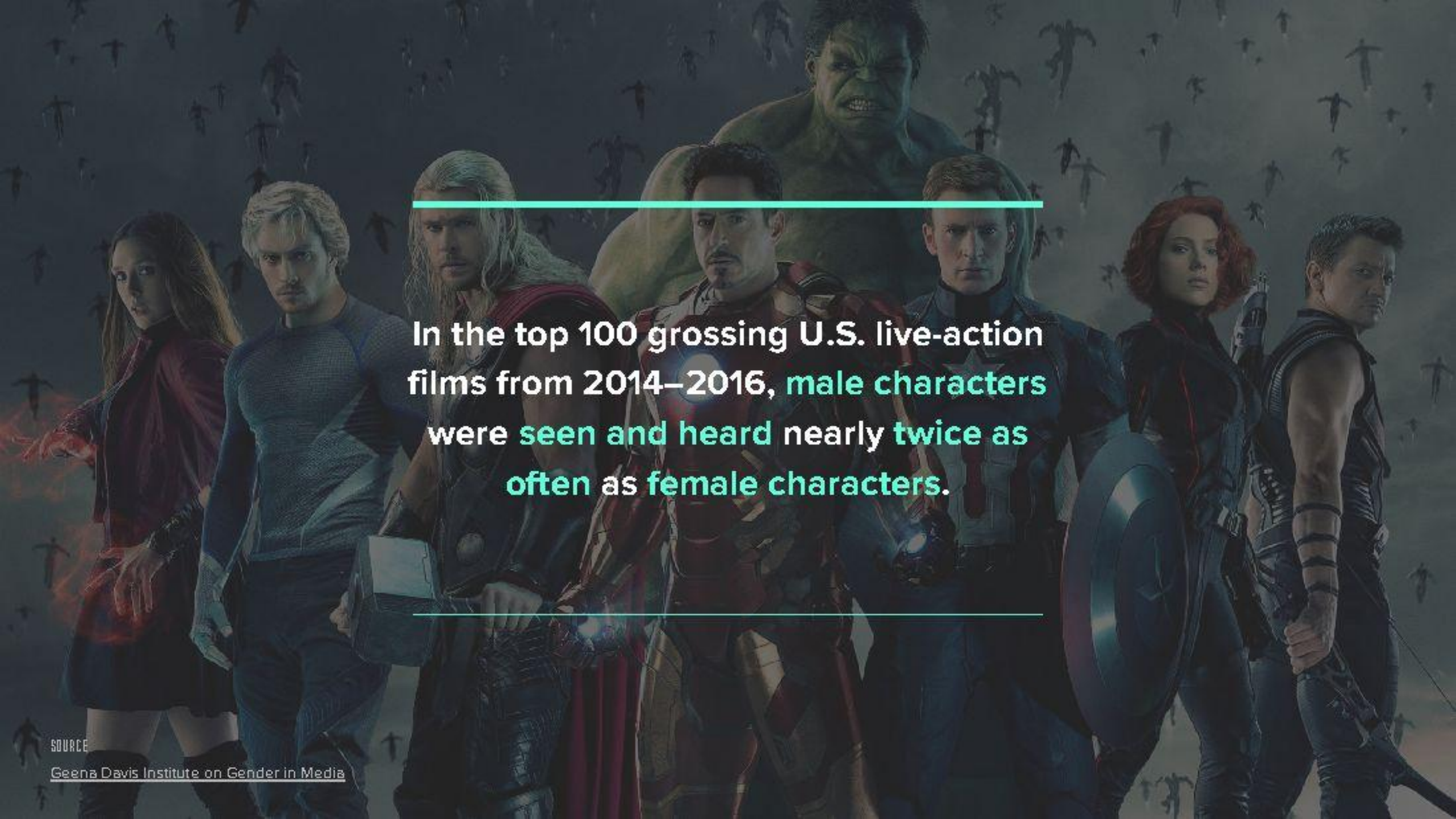
Until the 1990's, virtually all color calibration for color film was based on caucasian skin tones. It wasn't until Kodak started receiving complaints from chocolatiers and wood furniture manufacturers that they began to invest research into supporting a broader range of brown tones.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

Colour Balance, Image Technologies, and Cognitive Equity. Roth

How Photography Was Optimized for White Skin Color (Priceonomics)



In the top 100 grossing U.S. live-action films from 2014–2016, male characters were seen and heard nearly twice as often as female characters.

SOURCE

Geena Davis Institute on Gender in Media







Where are these disconnects coming from?



Are automakers **intentionally** trying to harm women?



Was the decision to color balance for light skin **overtly** racist?



Are creatives in the film industry marginalizing women **on purpose**?

These are often the behaviors of rational actors making what *seemed* like 'obvious' choices, *without* malice or ill-will.

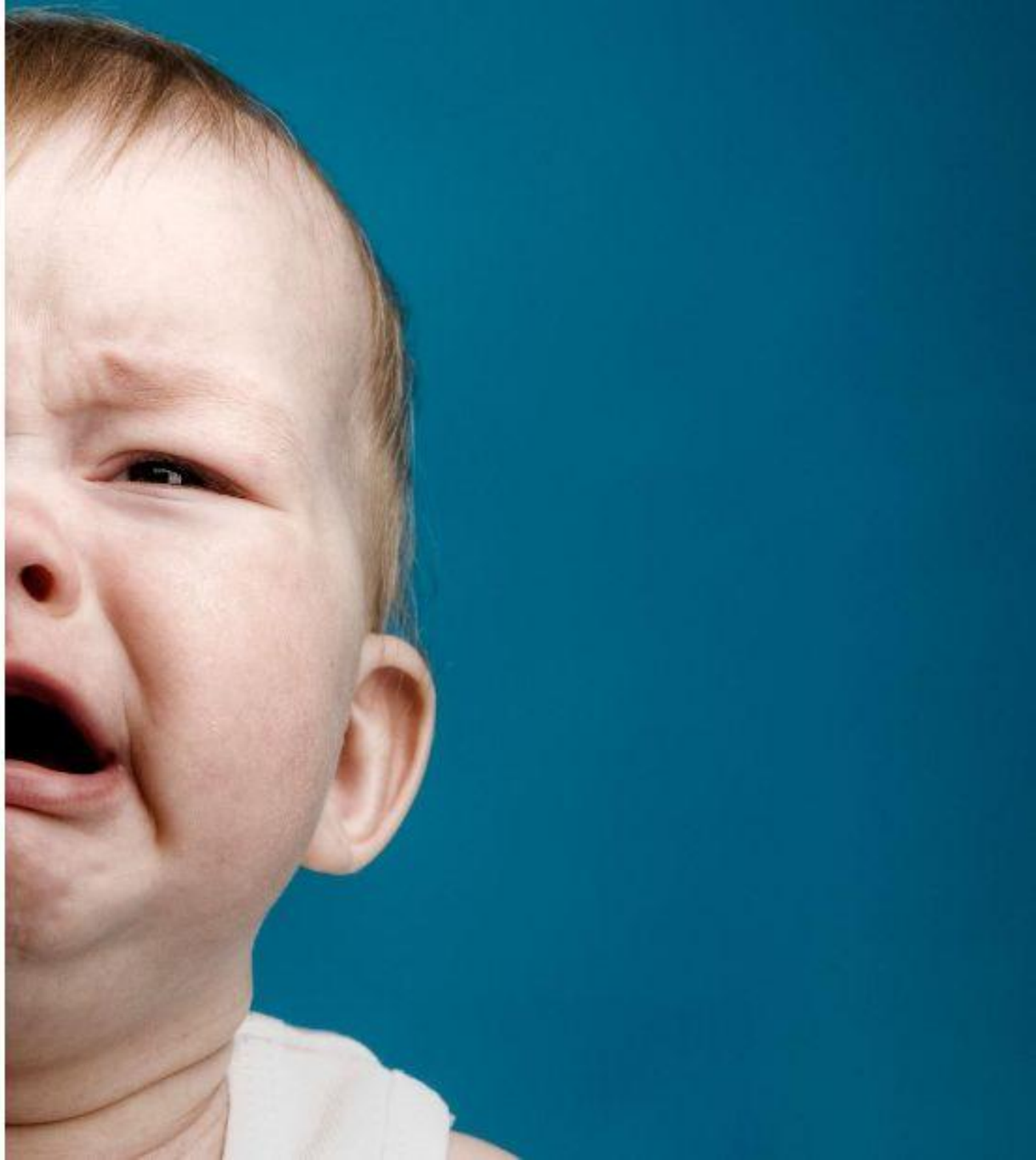
SOURCES

Why Heuristics Work, Gigerenzer

Rationality, Blume and Easley

Recommendations Implicit in Policy Defaults, McKenzie, Liersch, and Finkelstein

Take for example the case of “David”





Or was it “Diana”?



For the purposes of this study, it was both.

Participants watched a series of videos of babies.
For each video, **half of the participants were told a boy's name, the other half a girl's name.**

SOURCES

Sex Differences: A Study of the Eye of the Beholder, Condry and Condry
Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

When the babies did something unambiguous, reports were not affected by the perceived gender.

If the baby clearly smiled, for example, everyone said the baby was smiling or happy.



SOURCES

Sex Differences: A Study of the Eye of the Beholder, Condry and Condry

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

Then the babies played with a jack-in-the-box toy. When it suddenly popped up, the child was startled and jumped backward.



SOURCES

Sex Differences: A Study of the Eye of the Beholder, Condry and Condry

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science



SOURCES

Sex Differences: A Study of the Eye of the Beholder | Condry and Condry

Steven Pinker & Elizabeth Spelke | The Scien[illegible] Gender & Science


And most of us would agree that
it would be *irrational* behavior to
treat a fearful child the same way
we'd treat an angry child.



"If knowledge of a child's gender affects adults' perception of that child, then male and female children are going to elicit different reactions from the world, different patterns of encouragement. These perceptions matter, even in parents who are committed to treating sons and daughters alike."

— Elizabeth Spelke

SOURCES

Steven Pinker & Elizabeth Spelke | The Science of Gender & Science

Same child, same reaction, different perception.

REPLICATION AND RELATED STUDIES

The Gender Stereotyping of Emotions, Plant, Hyde, Keltner, Devine

A comparison of observed and reported adult-infant interactions: Effects of perceived sex, Culp, Cook, Housley

Adult perceptions of the infant as a function of gender labeling and observer gender, Delk, Madden, Livingston, Ryan

Surprising Smiles and Unanticipated Frowns: How Emotion and Status Influence Gender Categorization, Smith, Lafrance, Knol, Moes

Human perception drives virtually every facet of machine learning.

I propose we* make machine-learning intentionally human-centered and intervene for fairness.

*We = Humans. This isn't something any one company should be doing in isolation, but we're in a good position to start.

Tenets

Designing for fairness

01

Be Accountable

We can’t take our hands off the steering wheel.

In rejecting the myth of neutral data, we are committing to be more conscious and conscientious.

01

Be Accountable

Present day
↳ ~~Robots~~
↳ ~~Yada yada yada~~
↳ ~~Future~~

Present day
↳ **Humans**
↳ **Still humans**

02

Be Skeptical

Challenge assumptions at every turn.

We can’t blindly rely on the systems that underpin conventional wisdom.

02

Be Skeptical

Journalism

Fake news is indistinguishable

Customer reviews

Male reviews skew average scores

Standardized tests

SAT scores don't predict grades

Medical science

Experiments over-recruit whites

Crime statistics

Racial profiling is real

03

Be Humane

Success metrics should bring out the best in human nature.

Standard engagement metrics confine people to whatever they've done *before*, rather than empowering what they're capable of doing *next*.

03

Be Humane

Time well spent

timewellspent.io

Learning and expression

Exploration and connection

User-defined goals

04

Be Humble

We don't *always* know better.

The tech industry is in love with "disruption", but frequently that means imposing a vision of the future *onto* users and expecting them to adapt.

04

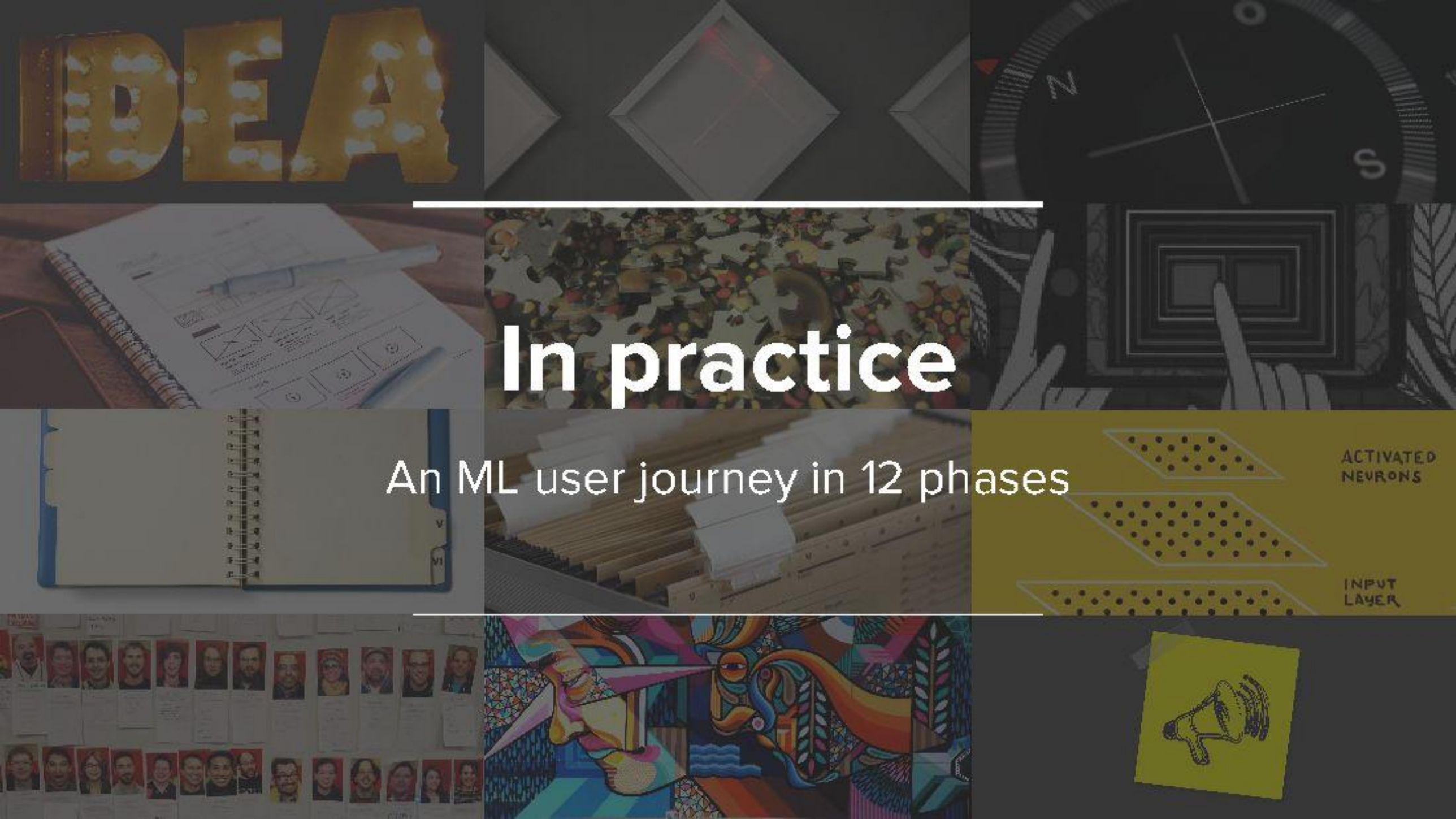
Be Humble

"There are two ways to get people used to automation. The soft and fuzzy way, ... just keep reassuring people until they're comfortable. And then there's the second way: let the humans ***take control when they need*** *to."*

— NPR

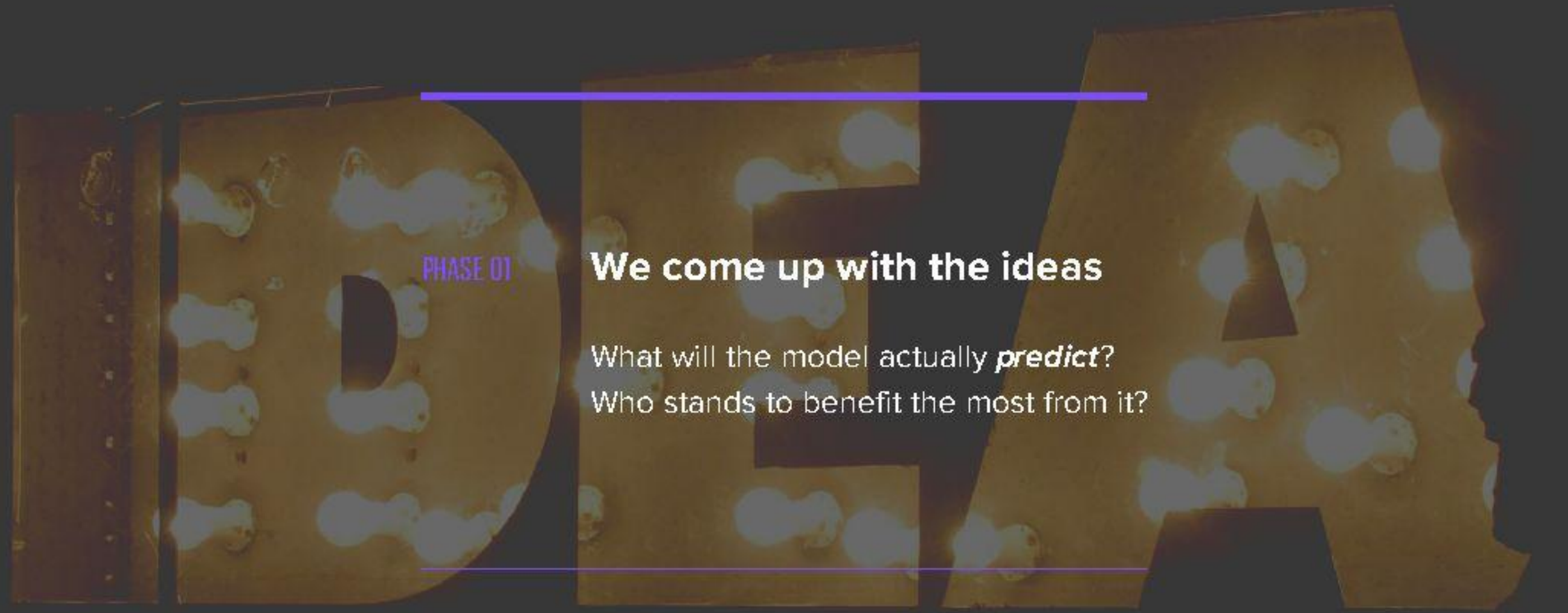
Augmenting → NOT → Automating

Supporting → NOT → Replacing



In practice

An ML user journey in 12 phases



PHASE 01

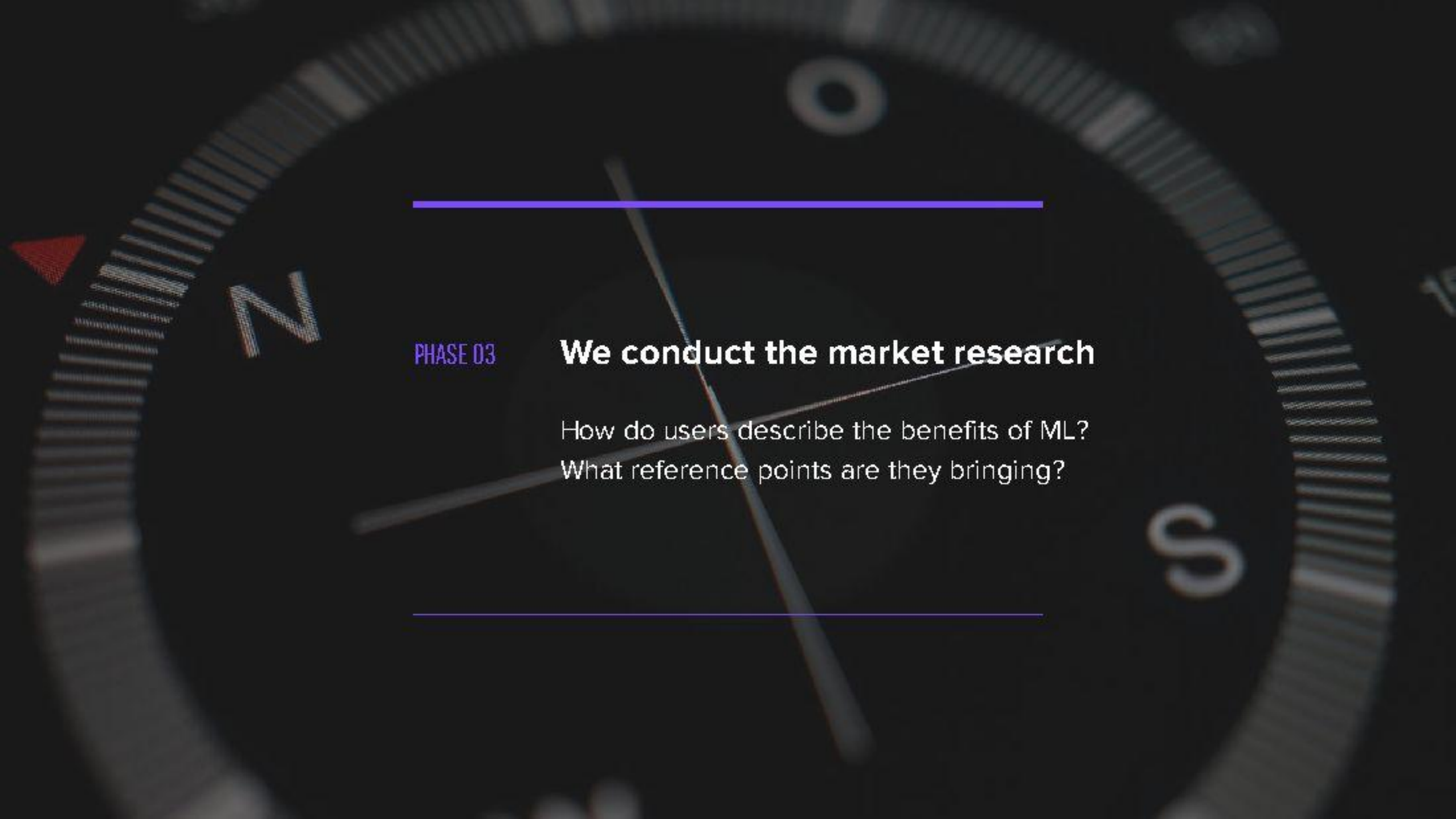
We come up with the ideas

What will the model actually ***predict***?
Who stands to benefit the most from it?

PHASE 02

We define the ML strategy

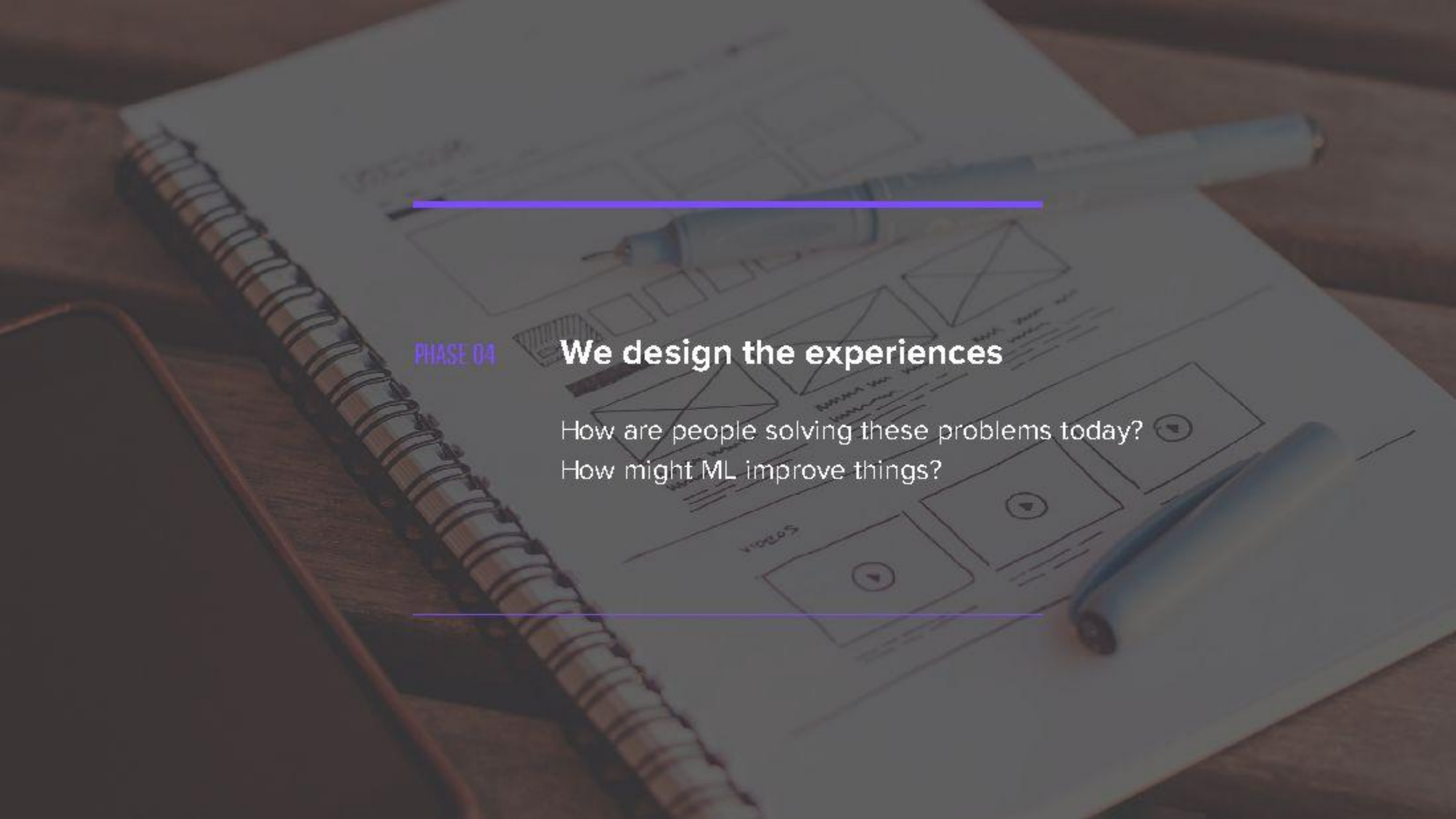
Why would heuristics be less effective than a machine-learned model?



PHASE 03

We conduct the market research

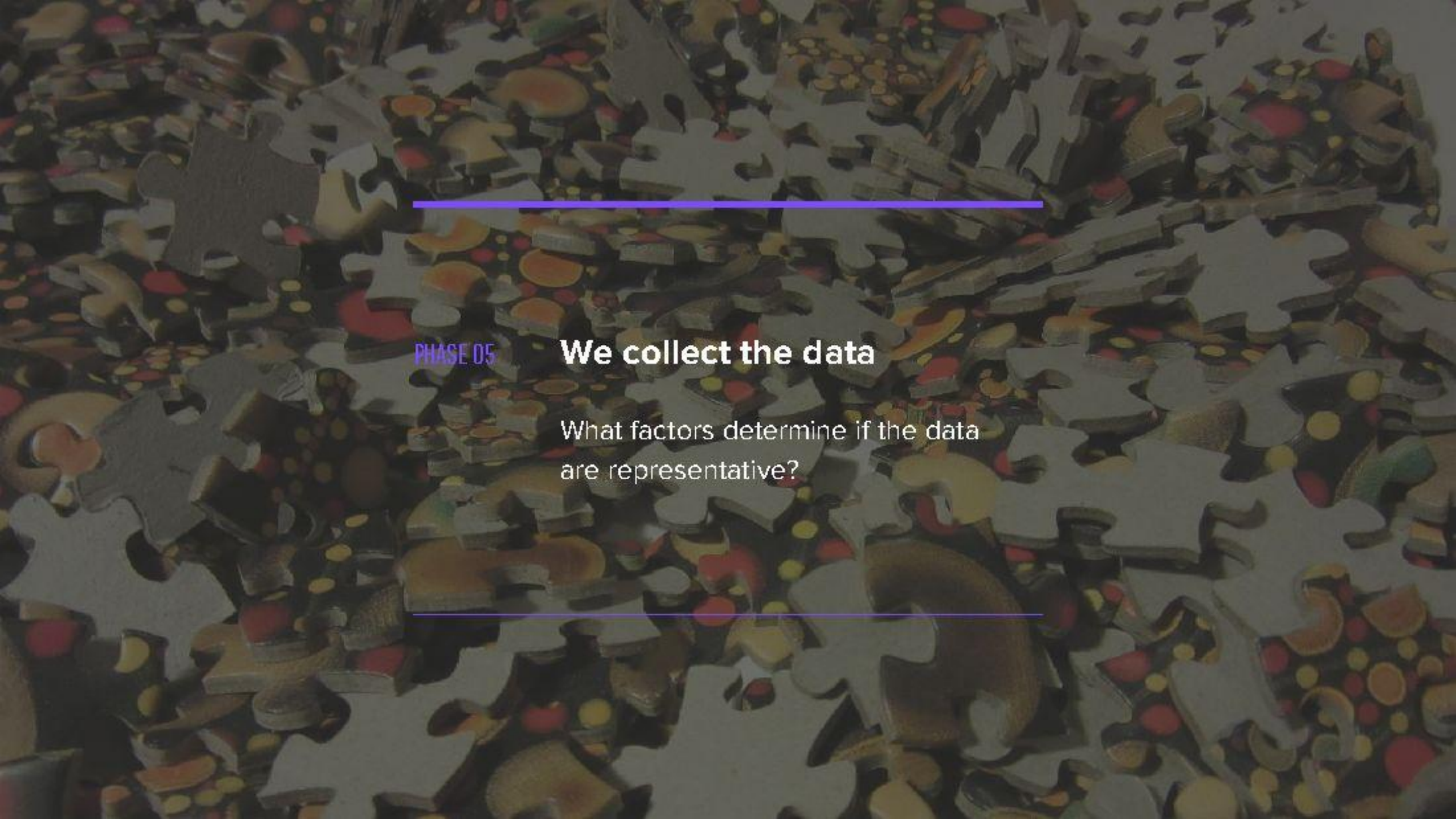
How do users describe the benefits of ML?
What reference points are they bringing?



PHASE 04

We design the experiences

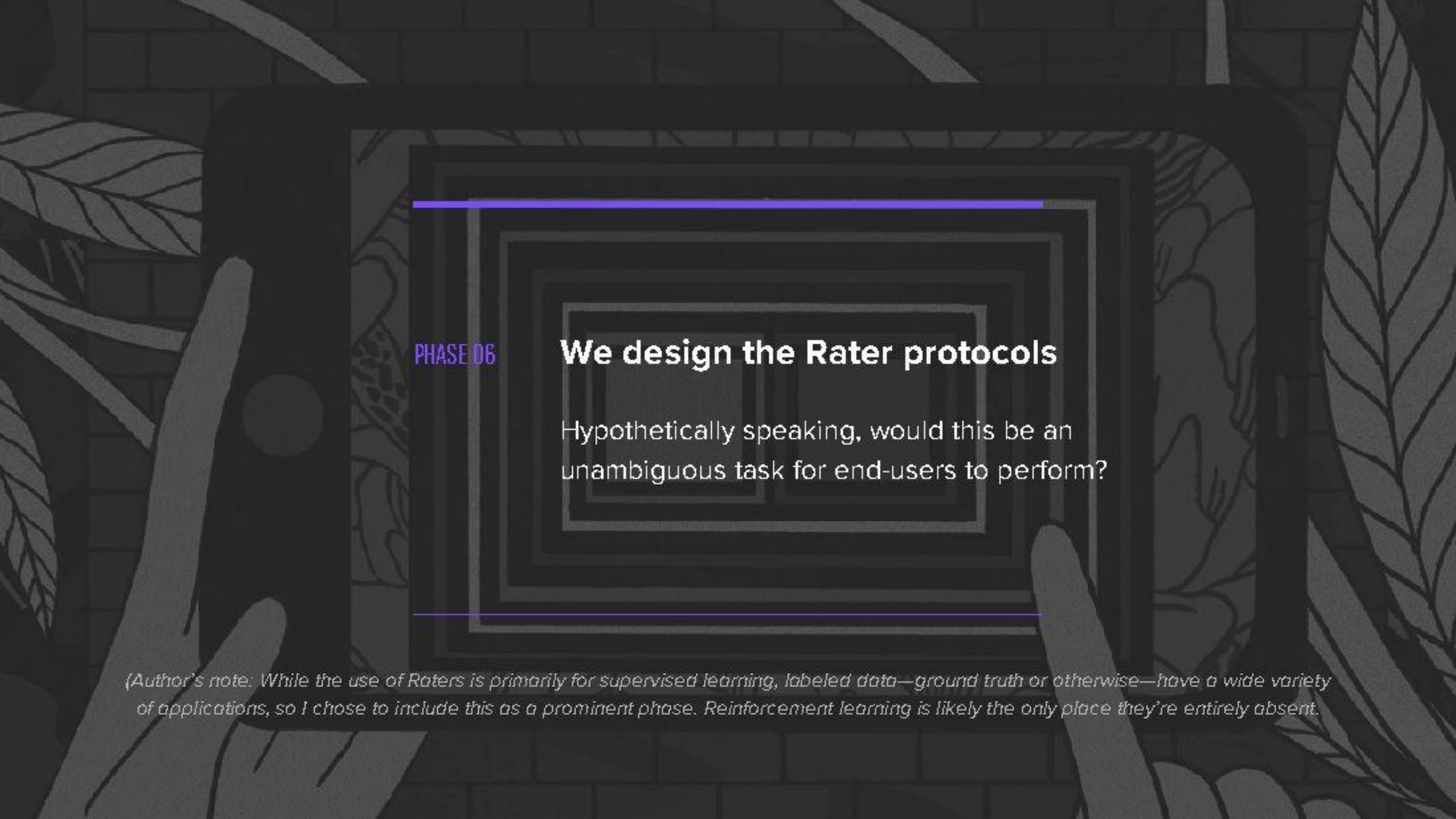
How are people solving these problems today?
How might ML improve things?



PHASE 05

We collect the data

What factors determine if the data are representative?



PHASE 06

We design the Rater protocols

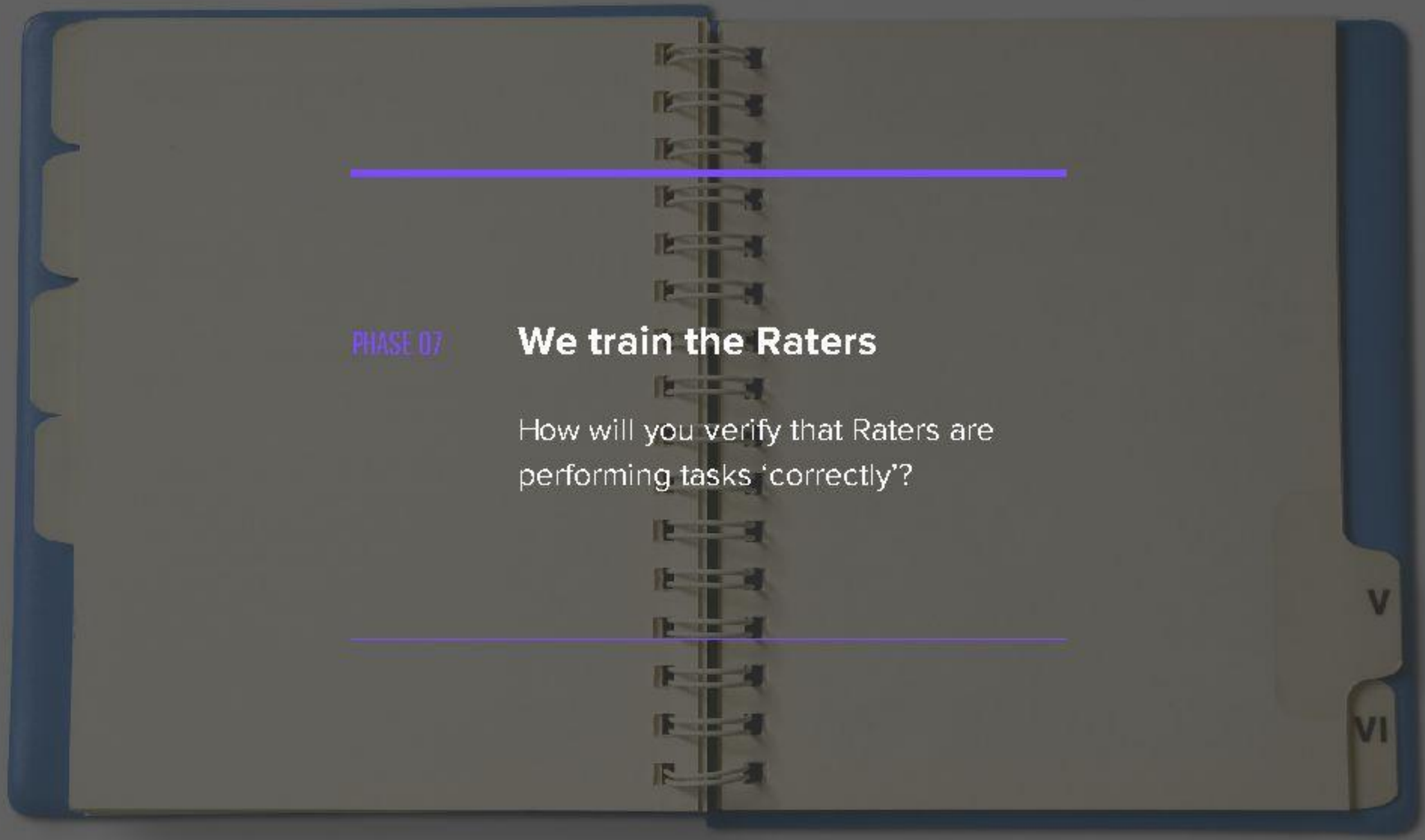
Hypothetically speaking, would this be an unambiguous task for end-users to perform?

(Author's note: While the use of Raters is primarily for supervised learning, labeled data—ground truth or otherwise—have a wide variety of applications, so I chose to include this as a prominent phase. Reinforcement learning is likely the only place they're entirely absent.

PHASE 06

Sidebar on Raters

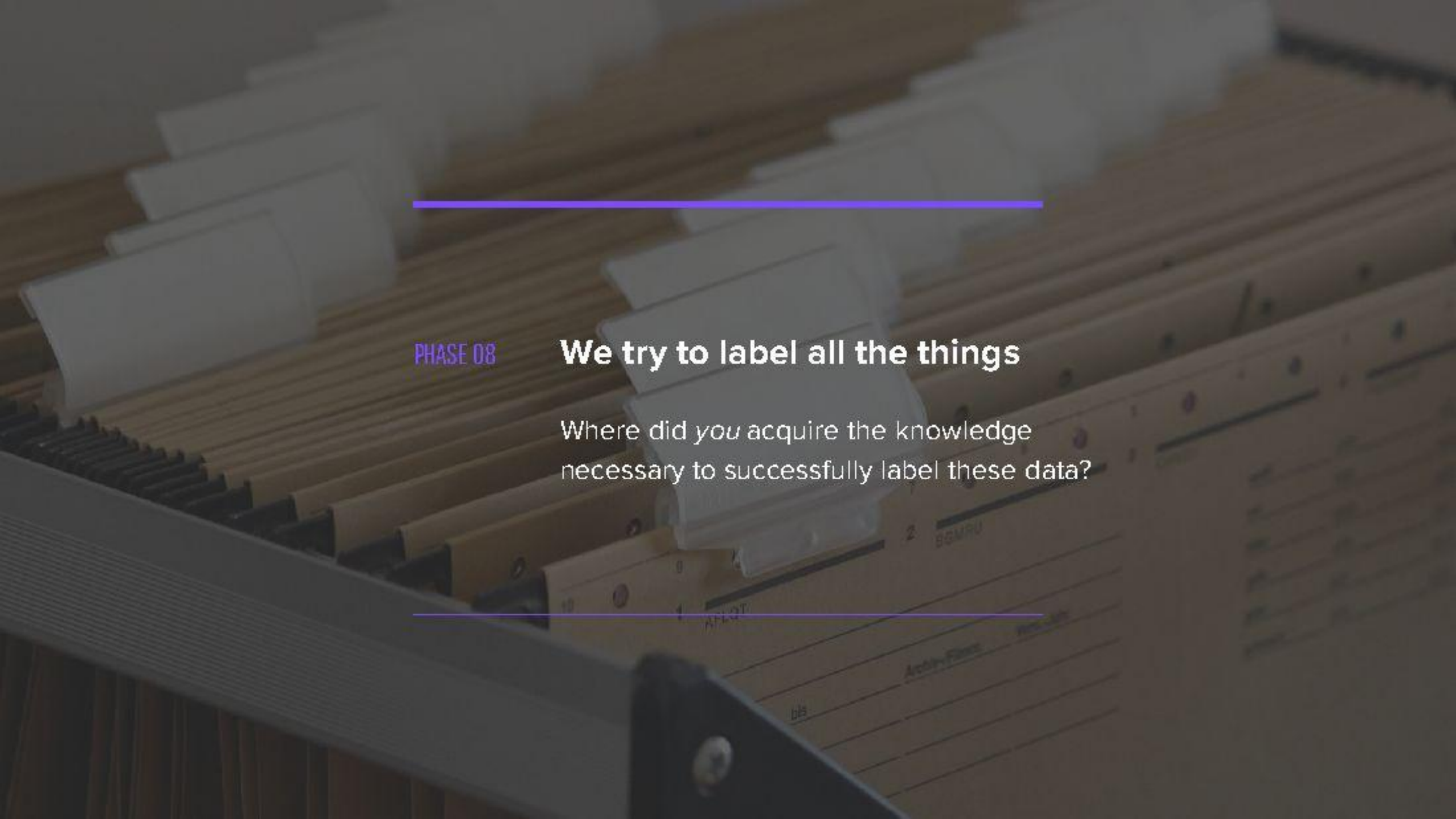
Speed and Agreement are the bedrock measures of "click-workers". The general idea is that if a lot of people can perform many quick tasks, the sheer volume of consensus will balance their lack of individual expertise. But without careful consideration for the diversity of Raters, click-work turns into exponential groupthink; baking cultural biases directly into training data.



PHASE 07

We train the Raters

How will you verify that Raters are performing tasks ‘correctly’?



PHASE 08

We try to label all the things

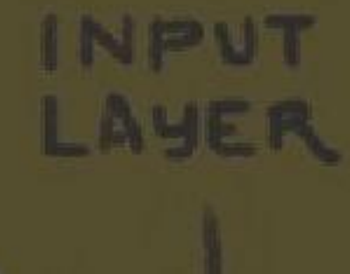
Where did *you* acquire the knowledge necessary to successfully label these data?

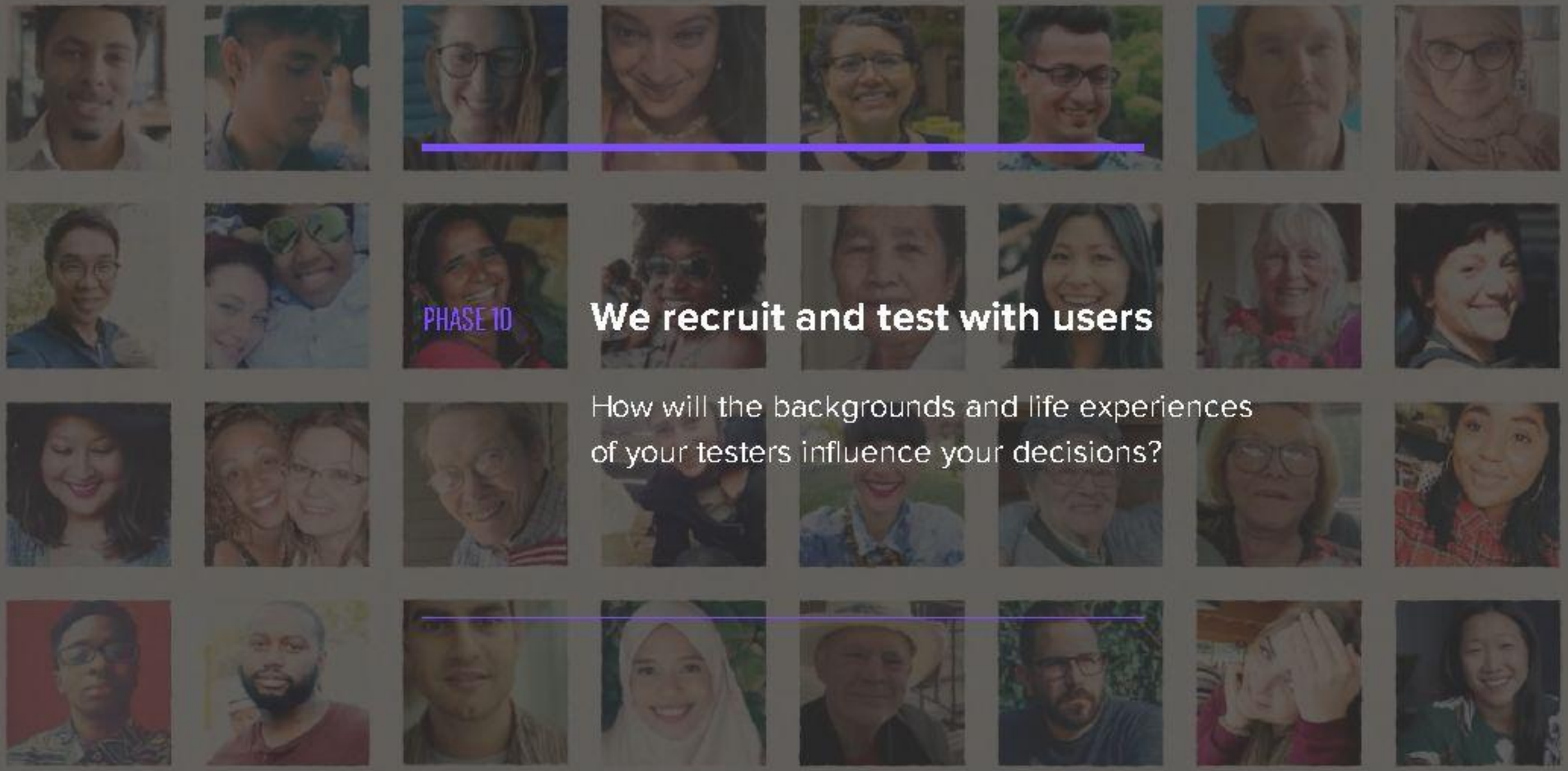
PHASE 09

We train the models

How will the model be debugged? What does ‘wrong’ look like?



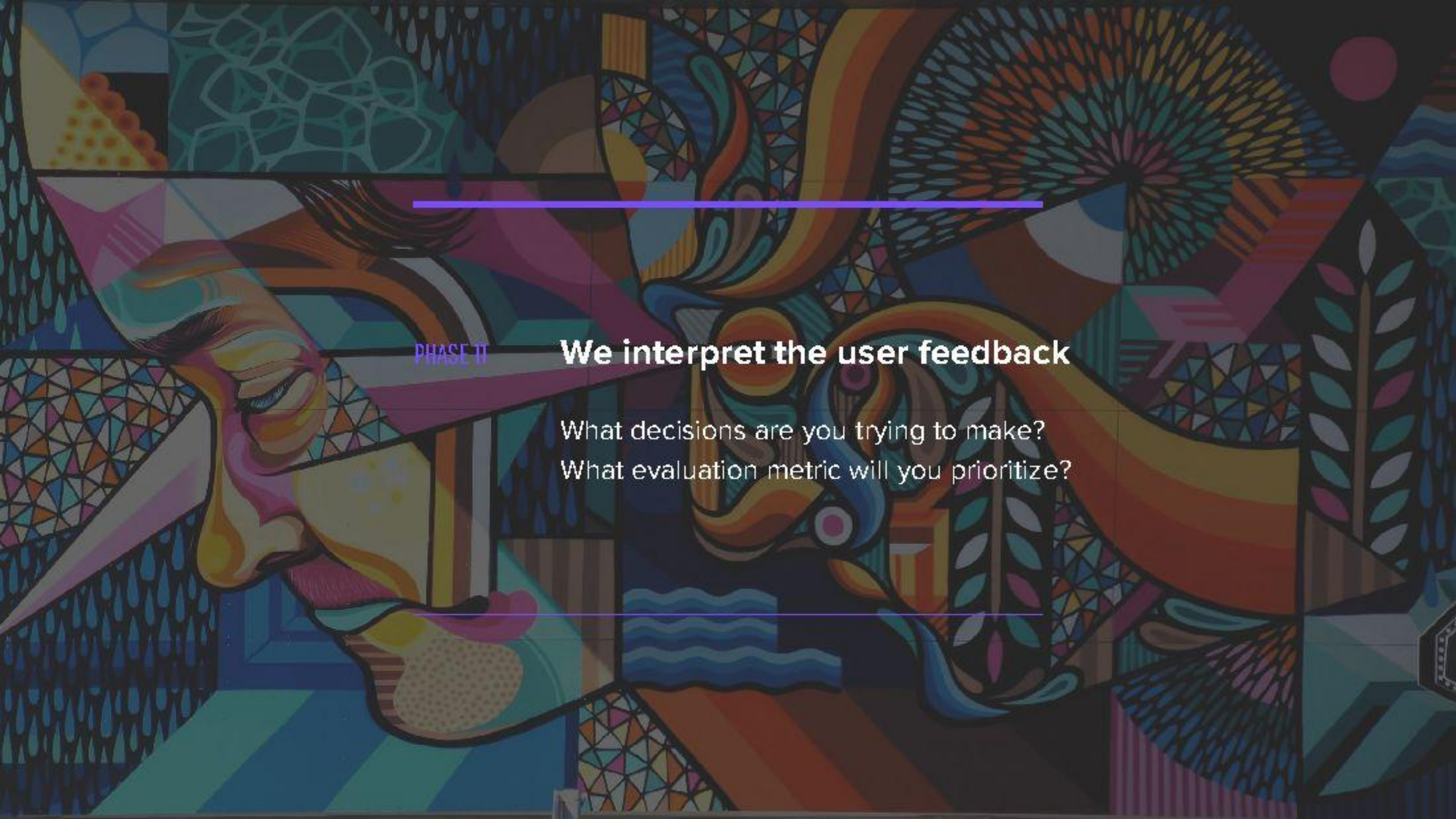




PHASE 10

We recruit and test with users

How will the backgrounds and life experiences of your testers influence your decisions?



PHASE II

We interpret the user feedback

What decisions are you trying to make?

What evaluation metric will you prioritize?

PHASE 12

We craft the PR

How will users see themselves reflected in marketing materials and demos?



Finally, I'd like to offer **3 human-centered diagnostics** when designing with ML.

If you're finding it tricky to answer these, it might be a signal to slow down and take a closer look.