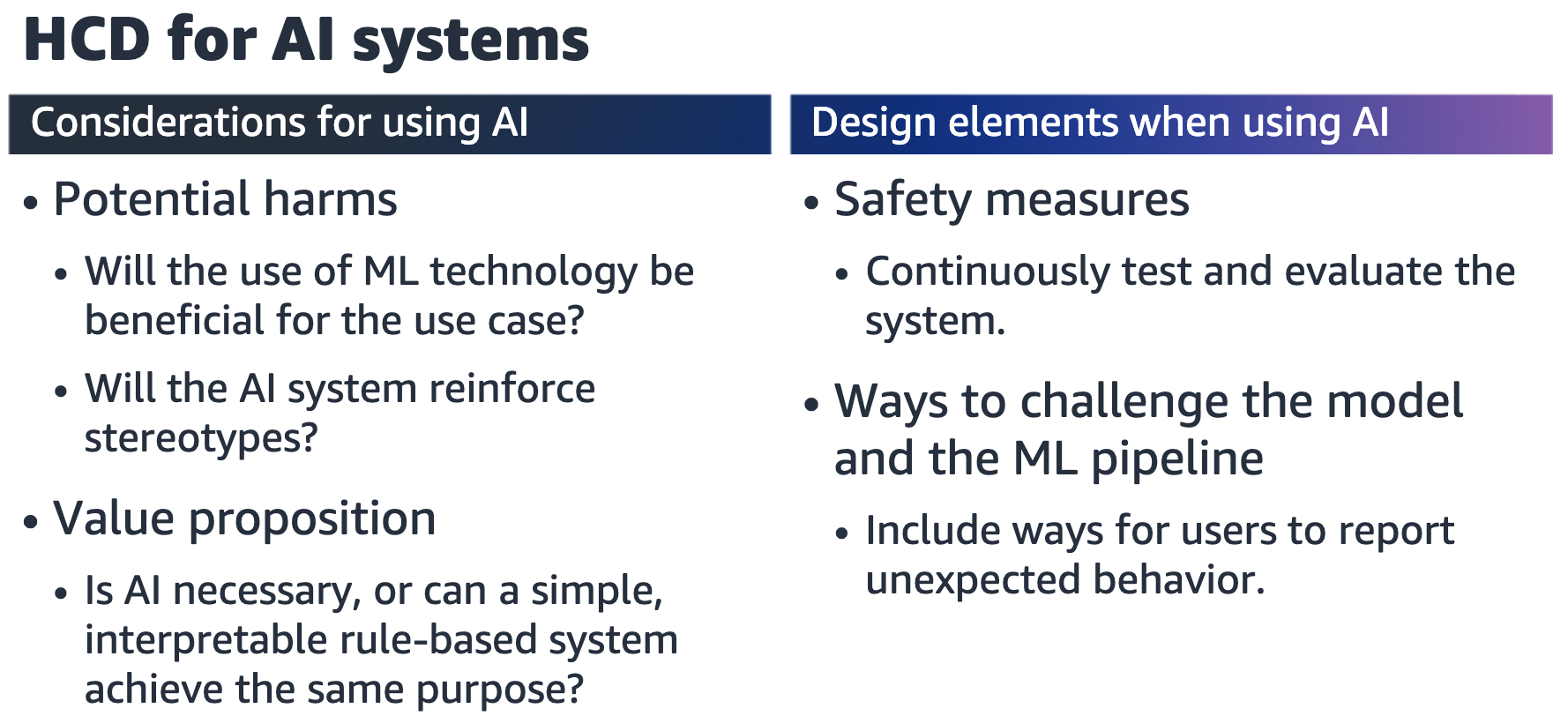


To help with the design of fair models, it is important to consider how the AI system

will impact humans. Human-centered design (HCD) is an approach that aims to make

(AI) systems relevant to the users, by asking questions to clarify their needs and

requirements. HCD also considers the risks AI can pose.



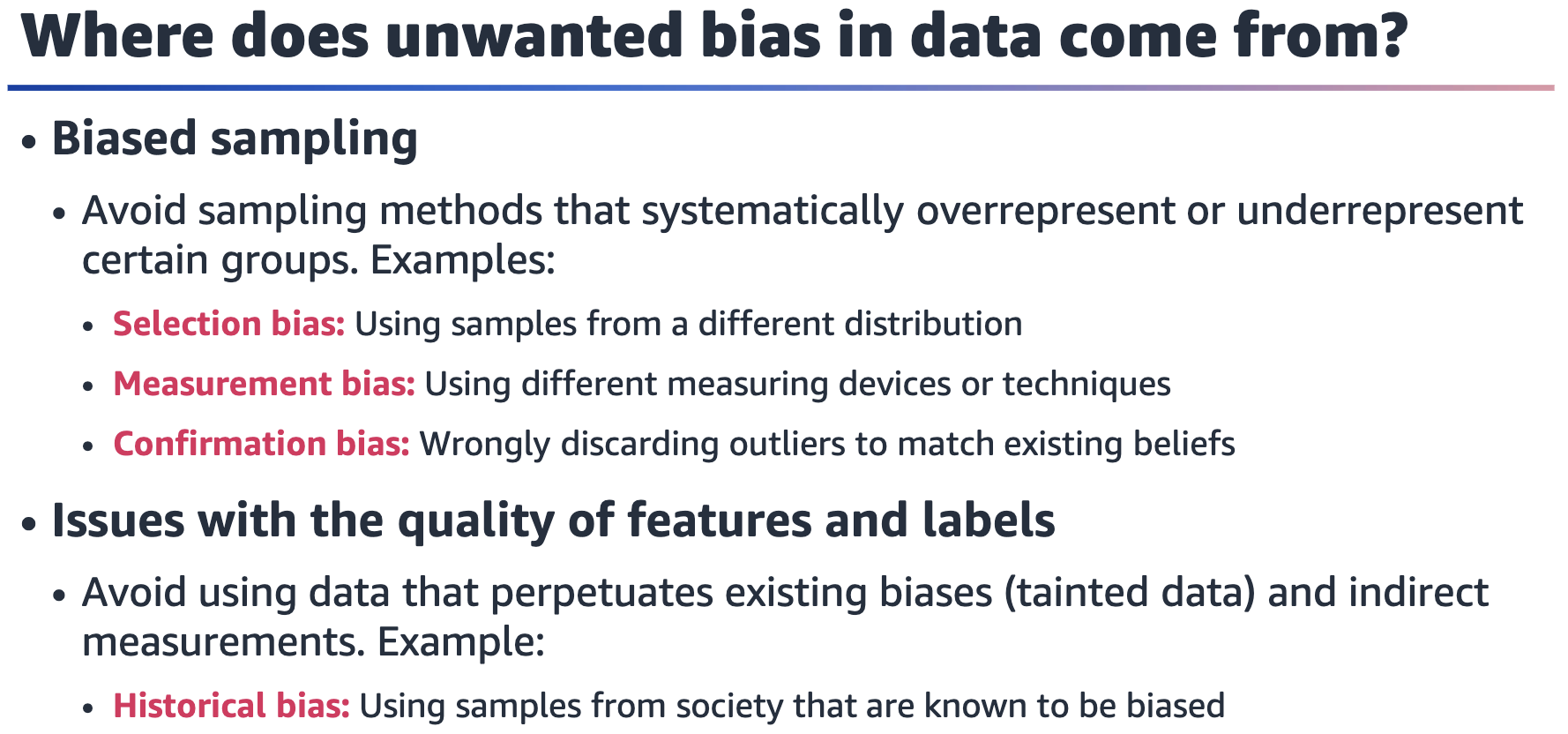
Human-centered design (HCD) is a big area of research. The main goal of HCD for AI is

to ask questions about potential harms and risk, as well as the value proposition of using AI. From a practical design point, you should ensure that safety measures are in

place for AI systems. You should also include mechanisms to allow end users to

challenge unexpected model behavior and outcomes.



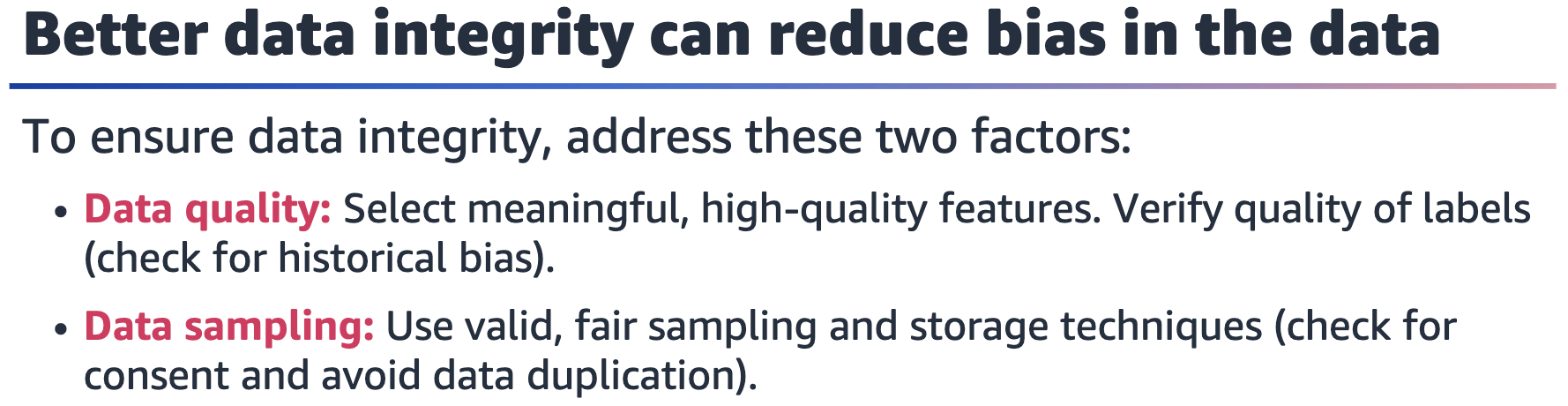


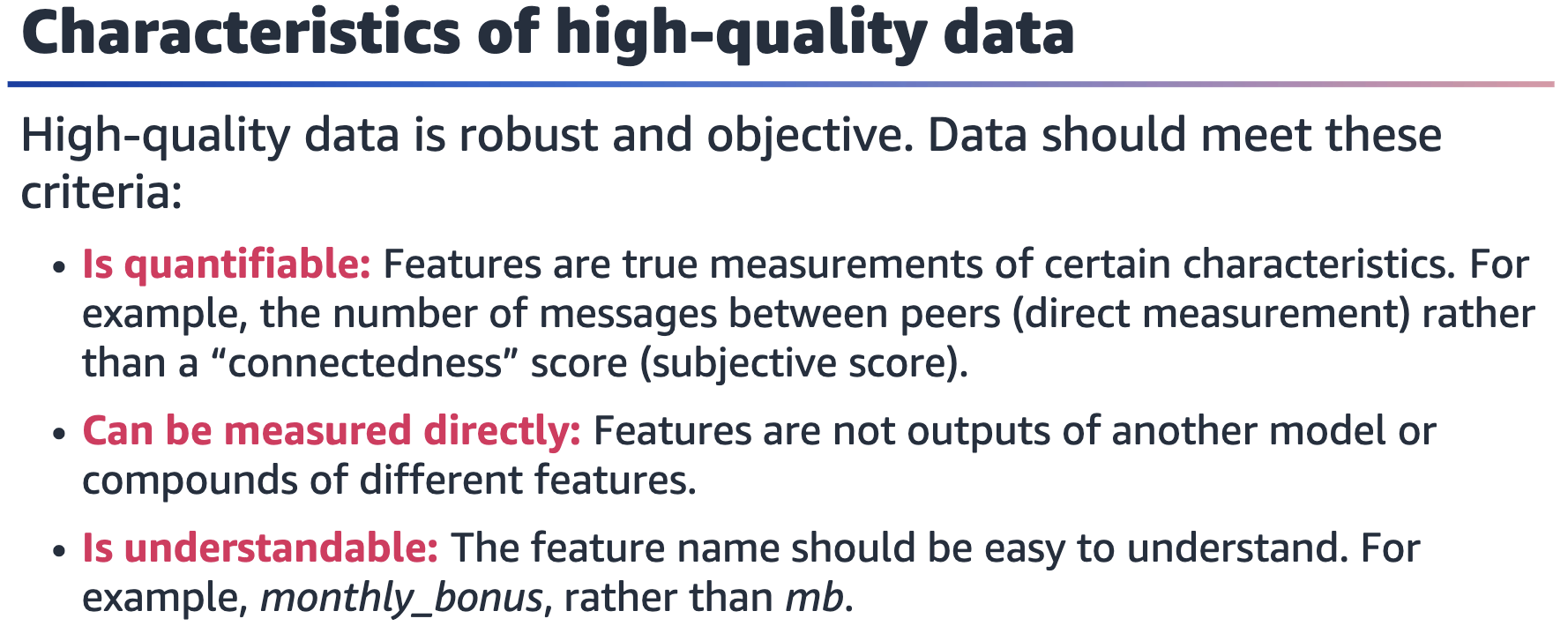
Bias in the outputs of ML models can stem from bias in the training data that

algorithms learn from. Bias in the training data can be because of different effects.

For example, biased sampling methods might lead to a skewed representation, or

features could be of low quality or not reproducible.

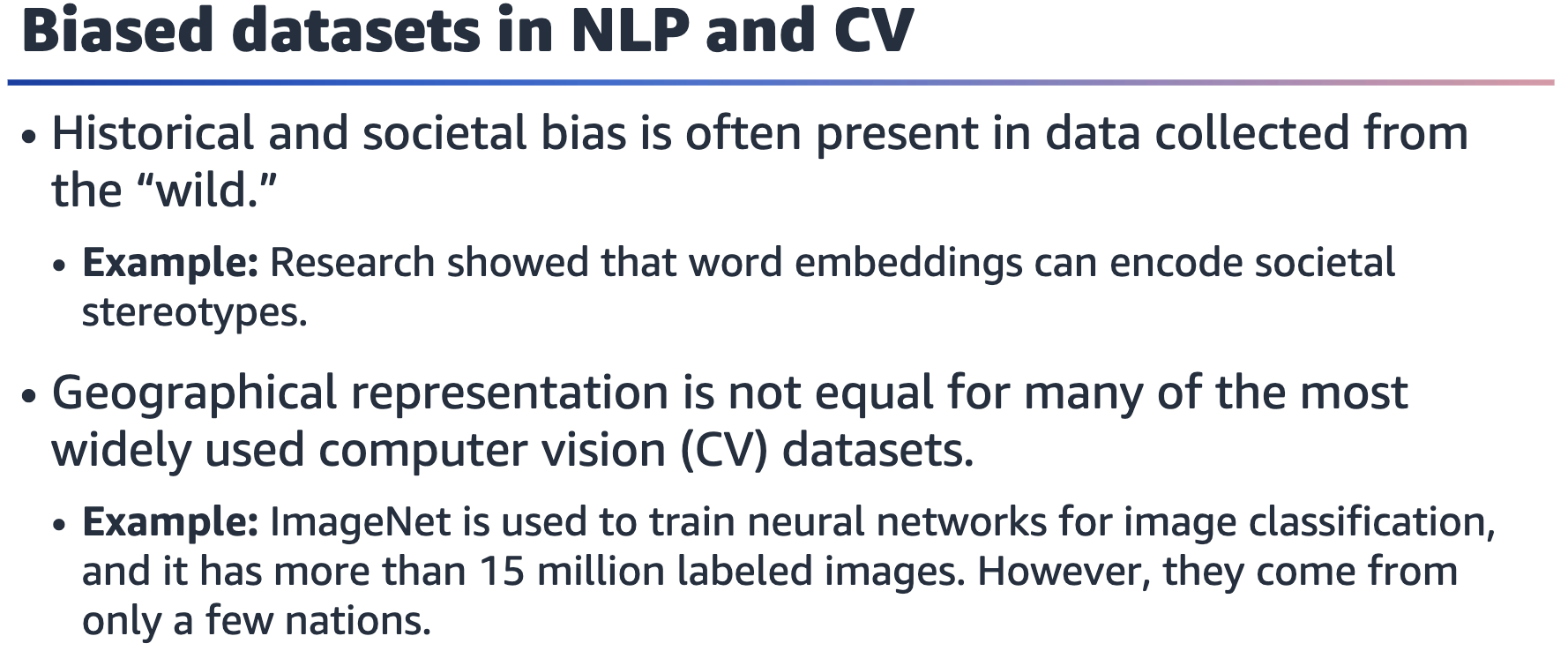




Data should be reproducible and not depend on who recorded it. If the data quality is

poor, then the model may end up perpetuating existing biases or even amplify the

bias. It is therefore very important to ensure that the data quality is as high as possible.



References:

• Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai,

“Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word

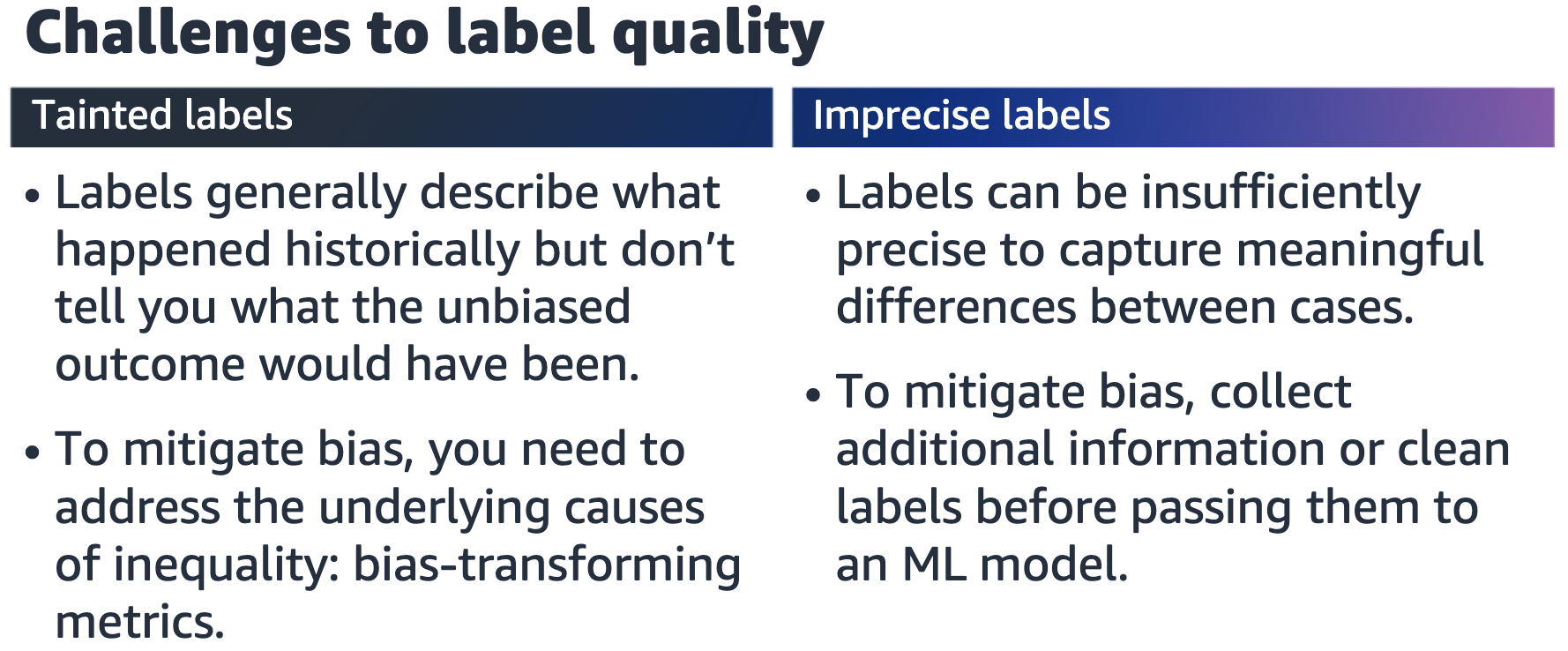
Embeddings,” arXiv (July 2016). <https://doi.org/10.48550/arXiv.1607.06520>.

• Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, D. Sculley,

“No Classification without Representation: Assessing Geodiversity Issues in Open

Data Sets for the Developing World,” arXiv (November 2017).

<https://doi.org/10.48550/arXiv.1711.08536>.



History is full of examples of disparate outcomes for different groups. If you train on

past biased decision-making, you might encode the historical, societal, or

geographical bias. Therefore, it’s important to think about the patterns that led to

biased outcomes to begin with.

References:

• Solon Barocas and Andrew D. Selbst, “Big Data’s Disparate Impact,” California Law

Review 104, no. 3. (June 2016): 671–732, <http://dx.doi.org/10.15779/Z38BG31>.

• Sandra Wachter, Brent Mittelstadt, and Chris Russell, “Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-discrimination Law,” West Virginia Law Review 123, no. 3. (2021): 2–51, <https://researchrepository.wvu.edu/wvlr/vol123/iss3/4>.

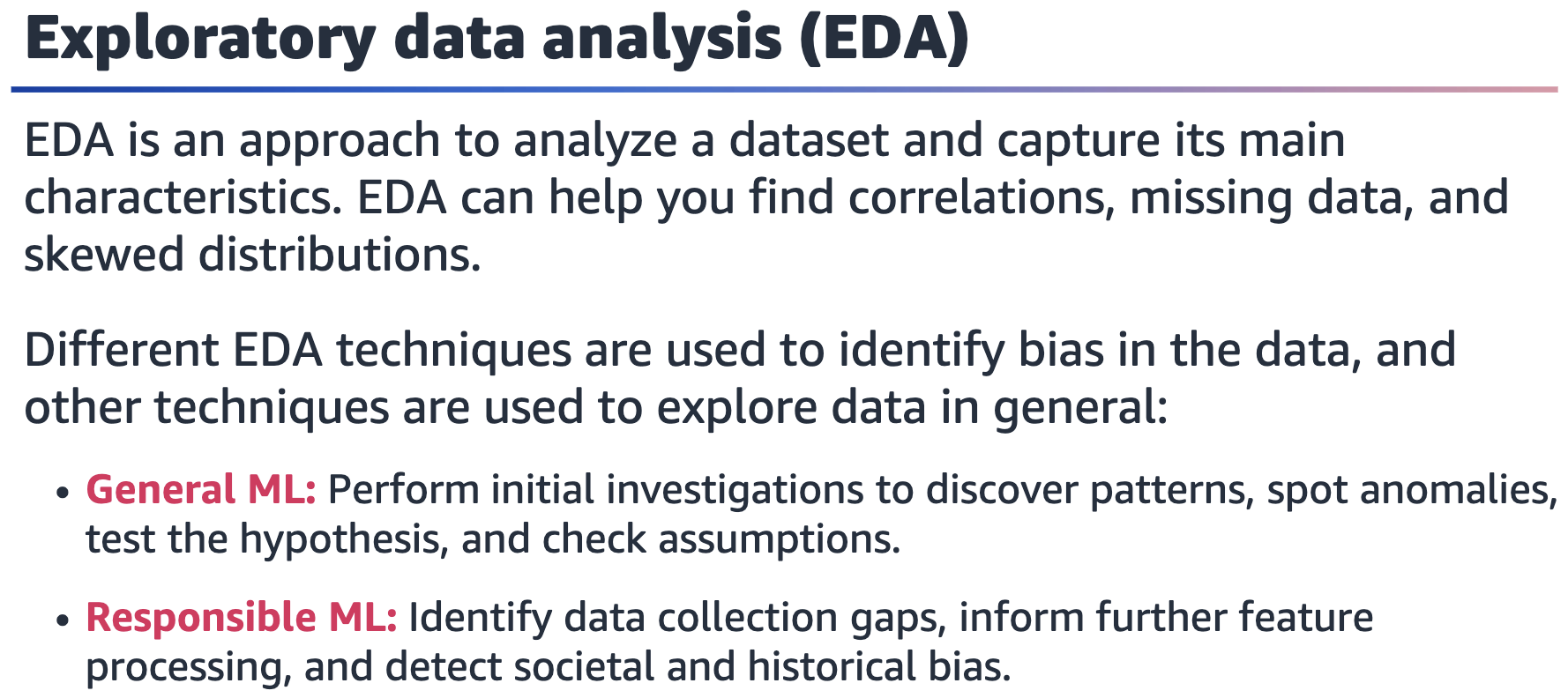


A datasheet tries to capture all the important information about a dataset that is

used to train a model. To complete a datasheet, you must answer several important

questions — many of which can help to identify bias before the model is even trained.





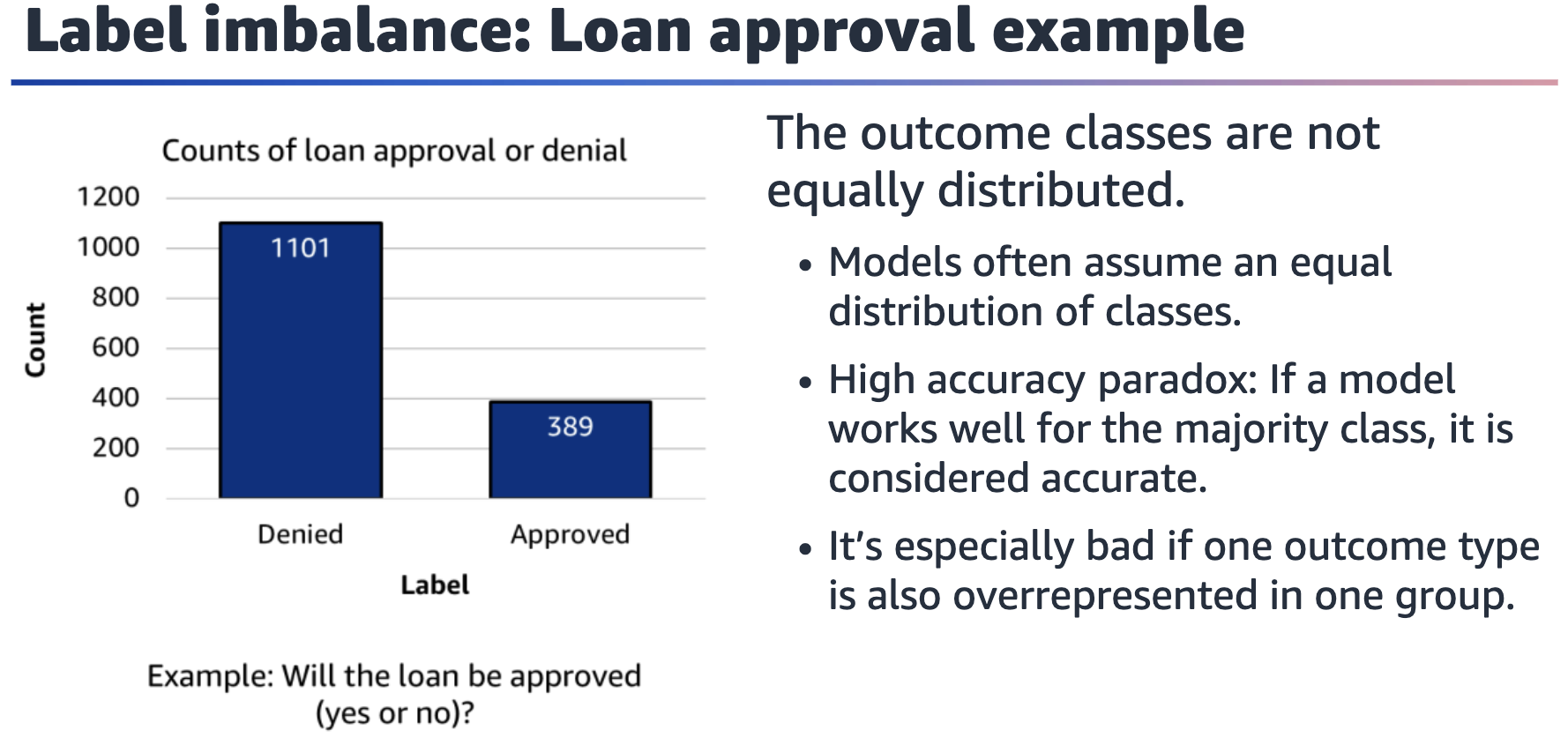
Exploratory data analysis (EDA) is important in ML in general, as it can help you

understand what kind of algorithm would be required to solve the ML problem. EDA

is also relevant to identify bias in the dataset by looking for correlations and data

skews. EDA includes the creation of summary statistics and graphical representations

of data (such as histograms and plots).



As part of EDA, you can analyze the distribution of the labels. Samples per class might

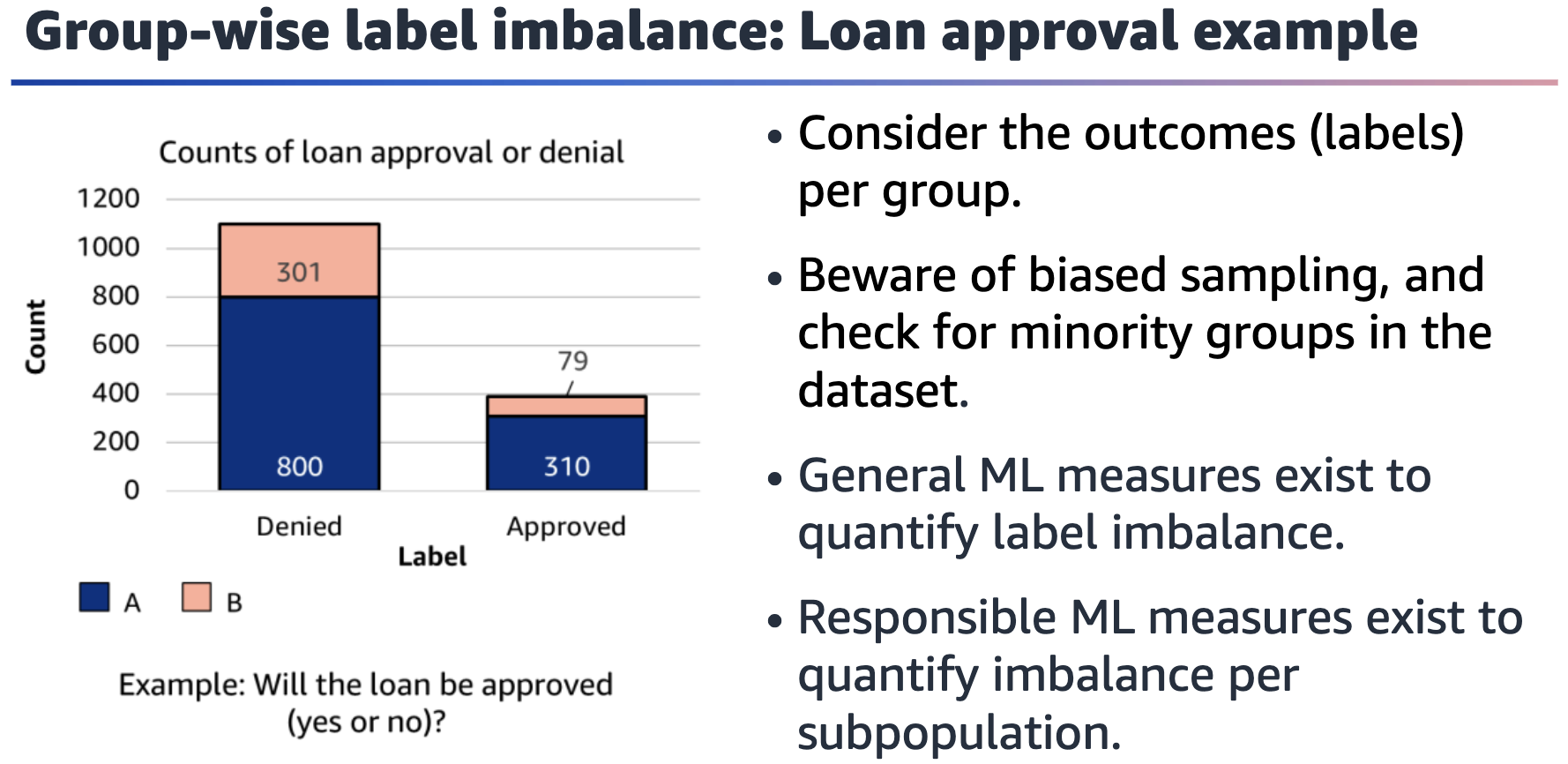
not be equally distributed, and the ML model might not work well for the infrequent

labels. In the chart on the slide, more data points are available for the “denied” label.

Models learn from data, so by seeing more negative outcomes, the model might

predict more negative outcomes in general. This can be especially bad if one group of

people is overrepresented in a certain outcome class.

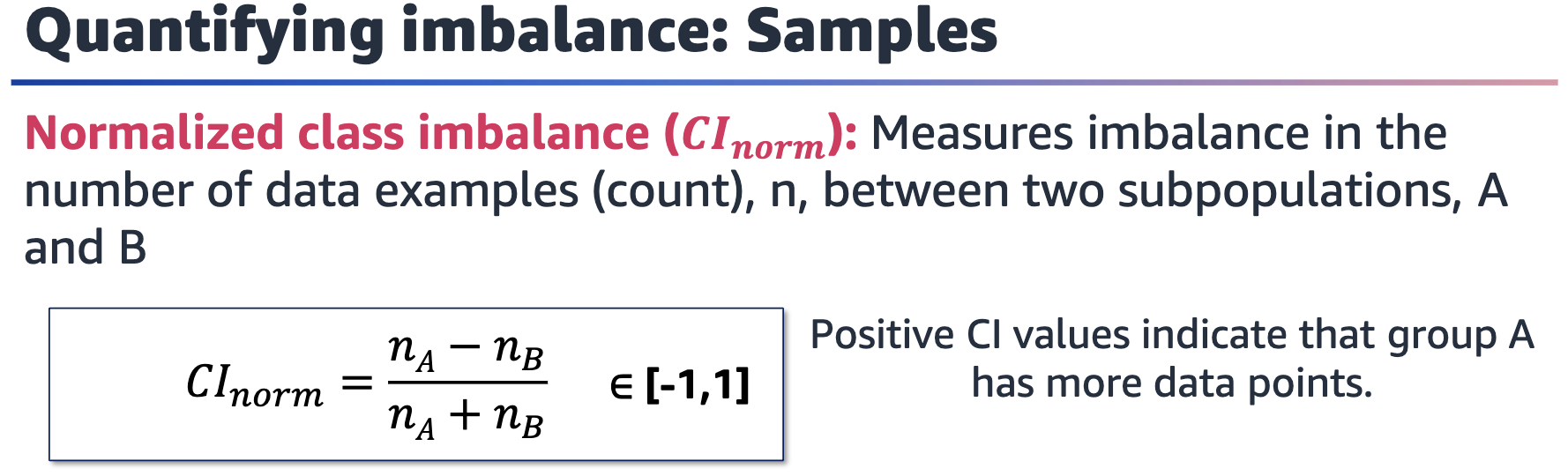


Bar chart showing denied and approved outcome for loan applications split by two groups, A and B. Group B has a worse ratio of approved outcomes than group A.

Group imbalance can pose an additional complication — especially when negative

outcomes are disproportionally concentrated in one group. Therefore, you should

visualize and quantify outcomes (labels) per group.



Class imbalance (CI) works to assess the imbalance in group sizes.

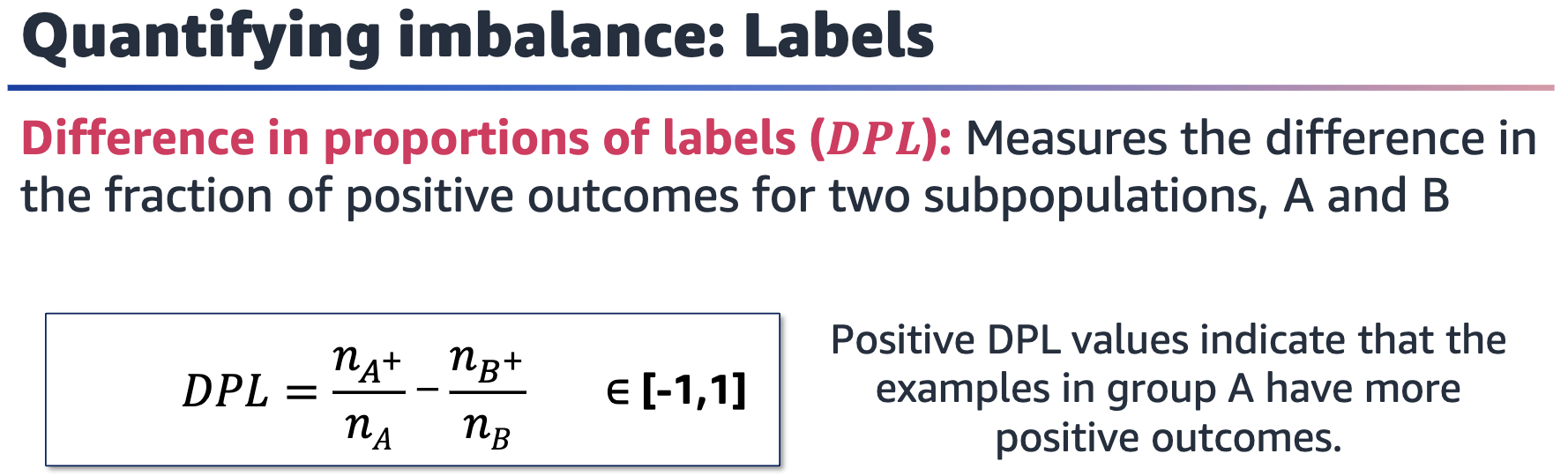
Positive CI values indicate that group A has more training samples in the dataset, and

a value of 1 indicates that the data only contains members of group A.

Values of CI that are near zero indicate a more equal distribution of members between groups. A value of zero indicates a perfectly equal partition between groups and represents a balanced distribution of samples in the training data.

Negative CI values indicate that group B has more training samples in the dataset, and

a value of -1 indicates that the data only contains members of group B.



For more metrics, see “Measure Pre-training Bias” in the Amazon SageMaker Developer Guide at

<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html>.

Difference in proportion of labels (DPL) is specifically used to assess label distribution.

Another metric is difference in predicted proportion of labels (DPPL). The equation is

the same as for DPL, but the numbers come from model predictions rather than

training data. You can often observe a shift for the worse in predicted labels.

Positive DPL values indicate that group A has a higher proportion of positive

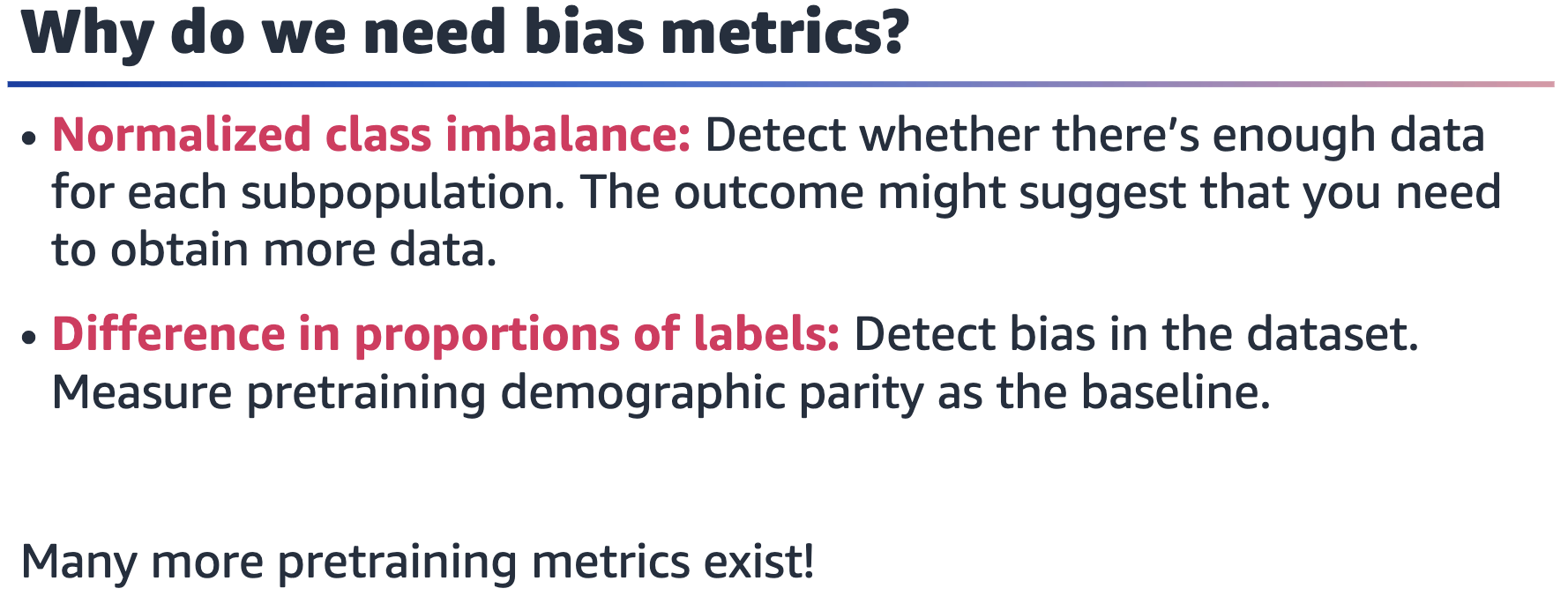
outcomes when compared with group B.

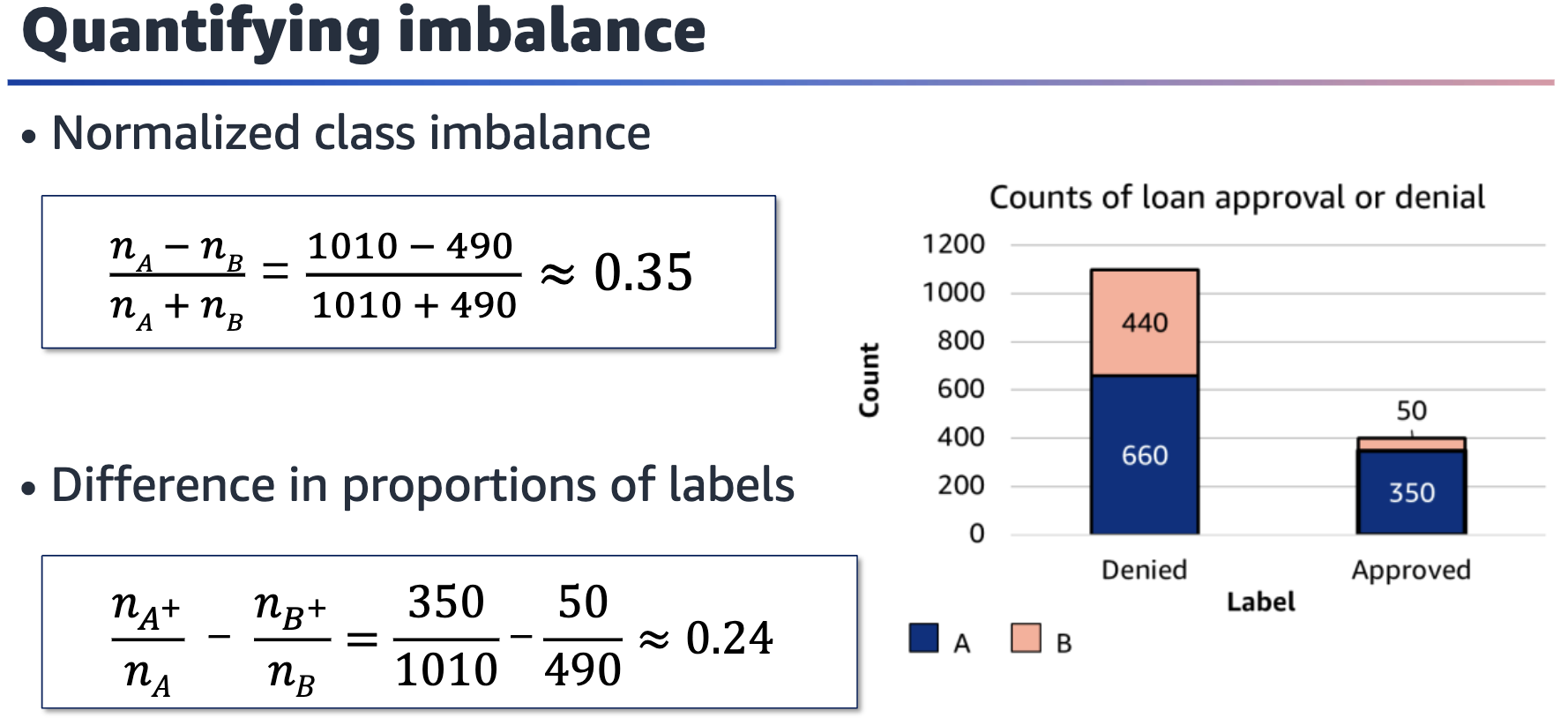
Values of DPL that are near zero indicate a more equal proportion of positive

outcomes between groups. A value of zero indicates perfect demographic parity.

Negative DPL values indicate that group B has a higher proportion of positive

outcomes when compared with group A.





Bar chart with counts of loan approvals and denials. There are 660 denials for group A and 440 denials for group B, for a total of 1,100 denials. There are 350 approvals for group A and 50 approvals for group B, for a total of 400 approvals.

This slide provides an example of calculating CI and DPL metrics on a sample dataset.

Note that the groups are represented almost equally for denied outcomes; however,

group B makes up a small portion of the total number of approvals.

For information about additional metrics to measure bias, see “Measure Pretraining

Bias” in the Amazon SageMaker Developer Guide at

<https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html>.