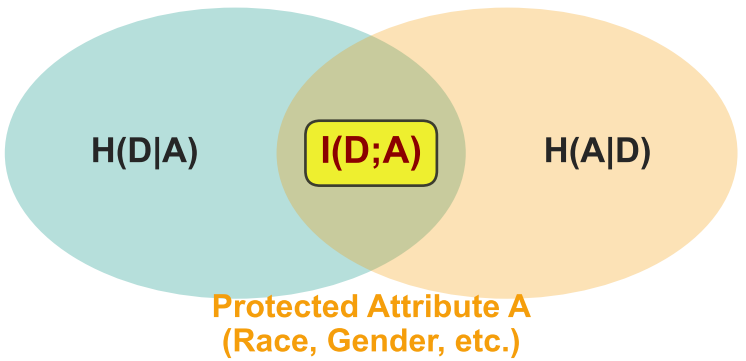
Information Theory of Bias: I(D; A) and Shannon Entropy

Mutual Information: I(D; A)

$$I(D; A) = H(D) - H(D|A) = H(A) - H(A|D)$$

(Approve/Deny)



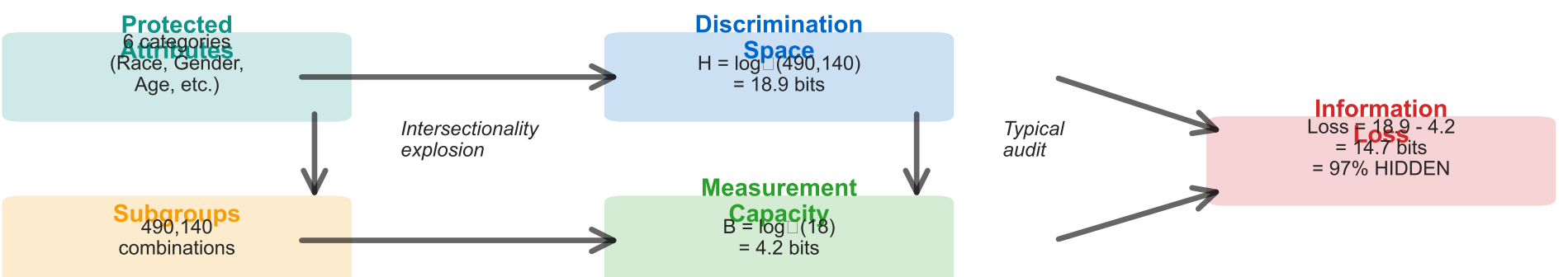
Bias Interpretation

No Bias
$I(D; A) = 0$
Decisions independent of group

Bias Present
$I(D; A) > 0$
Decisions leak group information

Example: Loan approval with I(D; A) = 0.21 bits
→ Knowing group reduces decision uncertainty by 21%

Shannon Entropy: Quantifying Unmeasurable Discrimination



Result: 97% of discrimination patterns are INVISIBLE
Only 4.2 bits measured out of 18.9 bits total discrimination space
→ Measurement bottleneck makes bias undetectable at scale