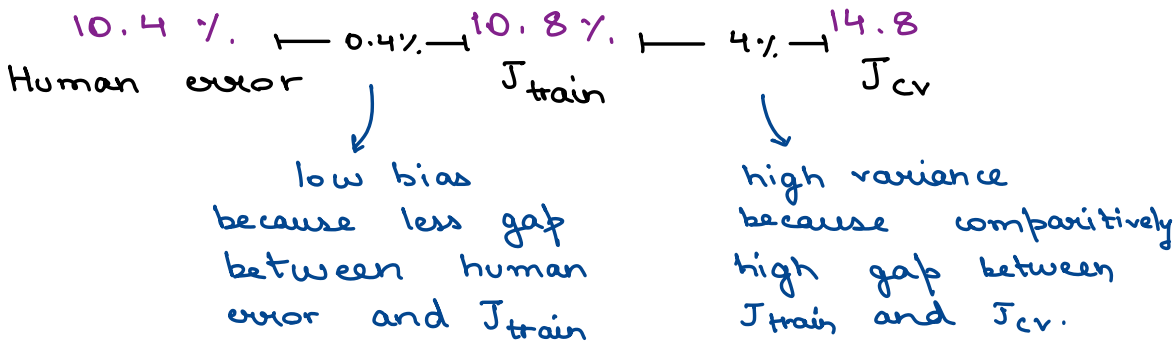The need to have a baseline for bias and variance

Example, take speech recognition, it has J_{train} error: 10.8% and J_{cv}: 14.8%.

We might say it has high bias and high variance (i.e. underfit), but we have to be realistic with our data.

Some audio on the internet is inelligible, even humans make 10.4% error while converting speech to text.

If even humans make that much error then it's safe to say the algorithm performs very good.

10.4%   ⊢— 0.4% —⊣ 10.8%   ⊢— 4% —⊣ 14.8

Human error J_{train} J_{cv}

low bias
because less gap
between human
error and J_{train}

high variance
because comparitively
high gap between
J_{train} and J_{cv}.

Basically,

1. Human error ← less gap → J_{train} ← high gap → J_{cv}

   eg.   10.4%,   ⟷ J_{train} : 10.8%.   ⟷ J_{cv} : 14.8%.
       ⟹ High variance

2. Human error ← high gap → J_{train} ← less gap → J_{cv}

   eg. 10.4%. ⟷ J_{train} : 14.8%. ⟷ J_{cv} : 15%.
   ⟹ High bias

3. Human error ← high gap → J_{train} ← high gap → J_{cv}

   eg. 10.4% ⟷ J_{train} : 14.8%. ⟷ J_{cv} : 18%.
       ⟹ High variance and high bias

Establishing a baseline level of performance

How to obtain the level of error we can possibly hope to get to?

→ Comparing to human level performance

→ Competing other algorithms' performance

→ Take a guess based on prior experience