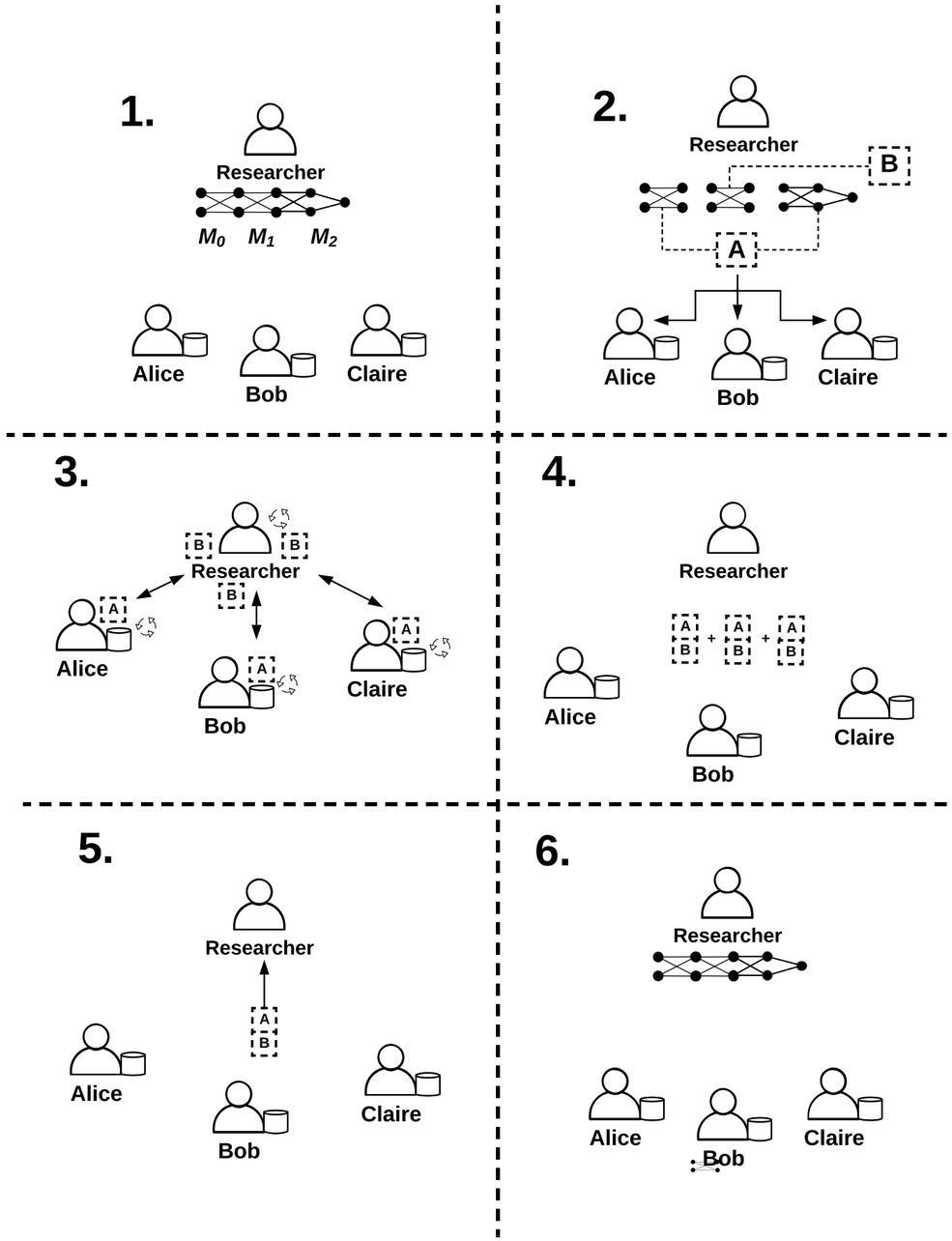
# SplitNN / Federated Learning




# Vulnerability A

## Researcher can reverse engineer X values if;

*- Has access to the **current weights and biases** of the model*
*- Has access to the **output** of that was produced by the model from those X values*

## The Case for Reversing X as Researcher

*- The Researcher starts with knowledge of the complete model, **M**.*
*- Only Alice knows **X**, her input data, and **Y**, her output data.*
*- The Researcher distributes the beginning segment, $M_0$, and the end segments, $M_2$, of the model to Alice to train.*
*- When Alice feeds her data through, the Researcher will recieve activation signals, **A,** to pass through his segment, $M_1$.*

*With knowledge of **A** and the version of $M_0$ that was combined with **X** to produce **A**, the Researcher can reverse engineer the value of Alice's **X.***

*Privately, if the Researcher cycles through all possible permutations of **X** as an input to $M_0$, they will eventually produce a result which matches the one they recieved from Alice, **A**.*

*Any **X^** where $M_0(X^)==A$ is potentially identical to the **X** data held by Alice. There is a chance that two different values of **X** will create the same **A** value. However this is made less likely as dimensionality is added to **A**. The more neurons which exist in the joining layer, the greater the likelihood of a 1:1 mapping of X:A.*

*Malicious Researchers could **generate rainbow tables** mapping **A** values produced by all plausible $M_0(X^)$. This would allow them to resolve **X** in all cases of **A**.*

## Researcher can approximate labels if;

*- I have access to the **current weights and biases** of the model*
*- I have access to the **gradients fed backward** through that model*

*In addition, if the Researcher has knowledge of the gradient signals being sent back to Alice. They can use these to update the original version of $M_0$ in parrallel to and independently of Alice. This allows the Researcher to maintain knowledge of $M_0$ over each epoch.*

*Matching an Alice's **A** to a particular $M_0(X^)$ over a large enough number epochs proves **X^** to be identical to Alice's **X** beyond all reasonable doubt.*

*In some case the dimensionality of **X** may make it unfeasble to explore all possible permutations with **X^**. However, permutations can be visited in order of a factor chosen by some heuristic, for example; conformity to expected statistical norm.*

## Key Takeaway:

***The Model should be processed entirely independently of any parties who may have prior knowledge of the model.***