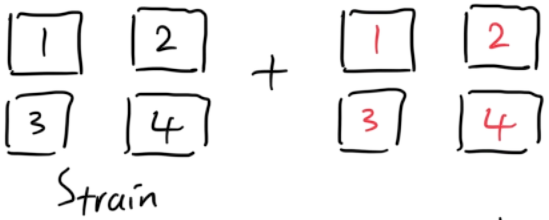
12) More on Data

Imbalanced data

- Proportion of data label is not evenly distributed (ex. fraud detection)
- Issue: model is learning to give a biased response
- Solution: ① Get more data (best way but not always possible)
  ② Undersample (drop common data)
  ③ Oversample (duplicate rare data) → for small sample sizes
  ④ Data augmentation (create transformation of rare data)
  ⑤ SMOTE (generates rare data by linear interpolation of feature space)

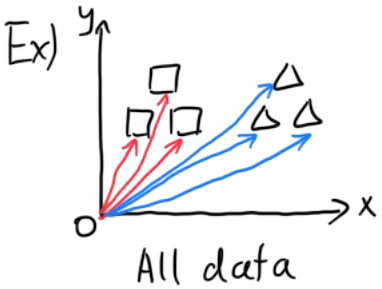
Data Augmentation

① Noise Augmentation: make a noisy copy (not redundant nor uncorrelated)

Ex) 

- Unlike oversampling, reduce overfitting risks
- particularly feasible in images

② Feature Augmentation: make new feature to data as nonlinear combination of existing ones

Ex) 

- Adding extra dimension may help separate data
- 3D model outperforms 2D model statistically if $P < 0.05$ for T-test
- particularly useful in signal processing

Save/Load Trained Model

- Pros: Skips the long training process of model
- Workflow: