# Regression

Supervised learning →

→ We're trying to predict some value given some data.

→ These systems are build using prior labeled data

→ for each $X_t$, there is an associated $Y_t$ :→ $\{x_t, y_t\}_{t=1}^{N}$

## *) Setting up data

$\frac{2}{3}$ training  $\frac{1}{3}$ testing

a) build system using training data,
b) test using testing data.
c) Compare system against others
d) find upper limit of error using this model

} training/testing data are from same distribution

? what IF we don't have this much data??

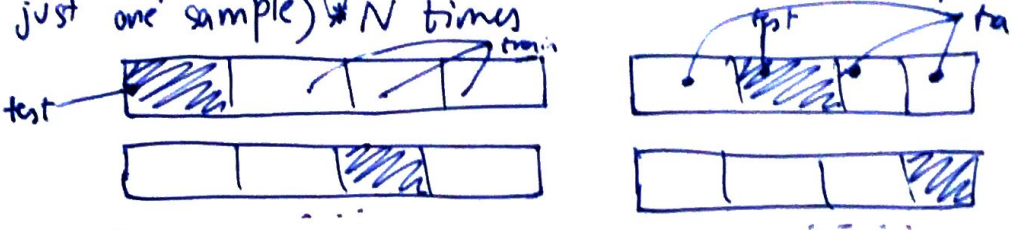
↳ we do Cross Validation

we do several training/testing runs while keeping track of errors.
Classifier ↓ statistics

## *) S-fold Cross Validation:

→ (divide data into S-parts, train on S-4 & test remaining part) S times

→ (If sample is really small, we build system on N-1 samples & test on just one sample) * N times




Learning function:

no noise? data = Z  $y = f(z)$

but usually, $x \approx z + \epsilon$

∴ we get, $g(x) \approx f(z)$
model  ideal learn fn

## OVer/UNDER FITTING

→ our model may give too specific or too broad results initially based on the choice of training data.