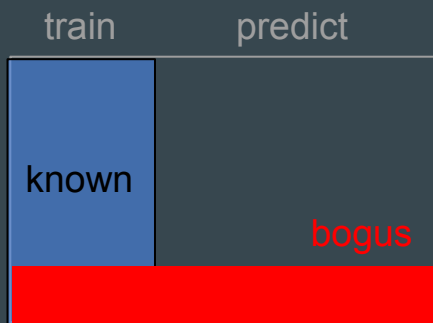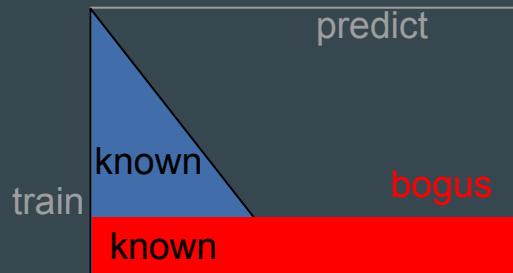# One-sided labels

Ideal Machine Learning scenario
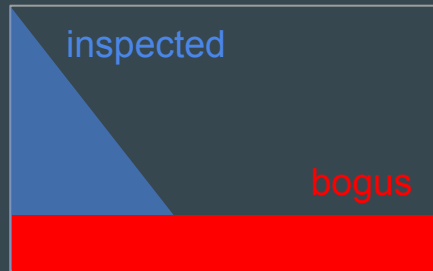    Labeled set representative of population

Real-life Machine Learning scenario
    Labeled set non-representative,
    but still fully labeled

Our Machine Learning scenario
    One sided labels for non-representative set,
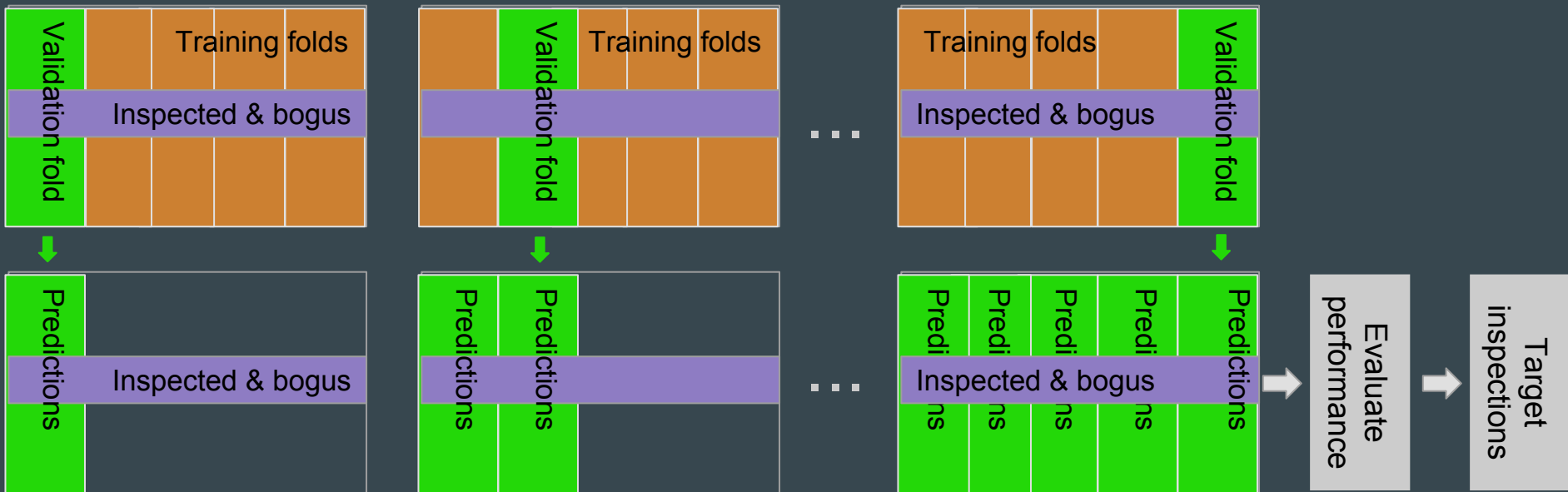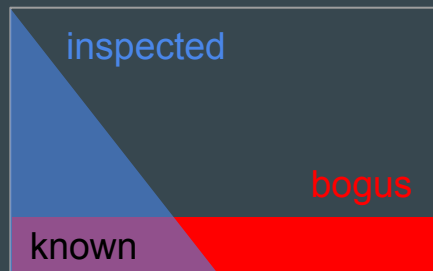    but labeled class rare in the population
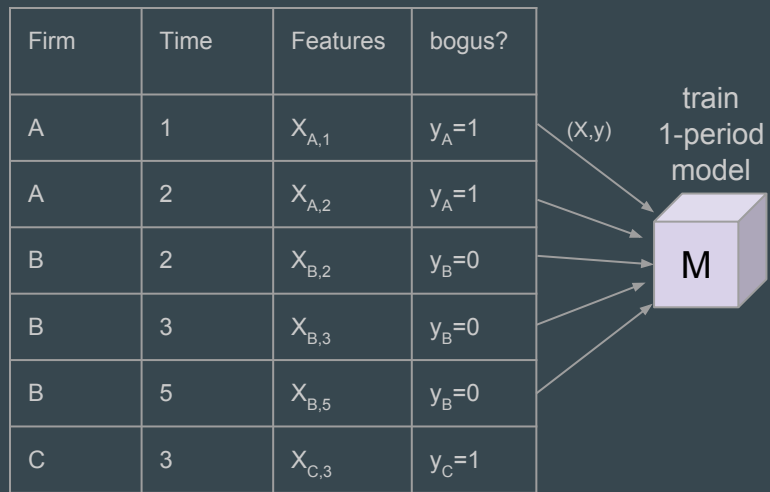
# One-sided labels - solution

Start with predicting (inspected & bogus) vs. the world

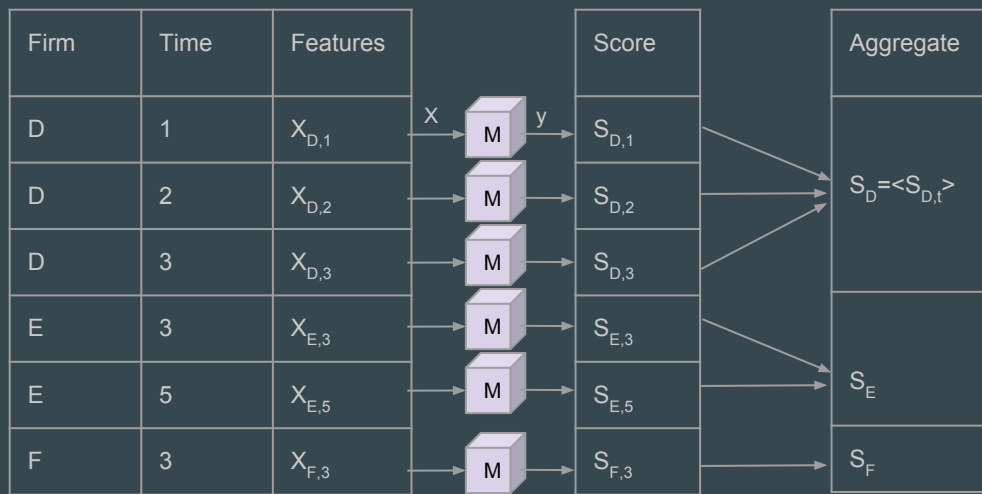Use cross-validation to predict for each unknown firm (when it is in the validation fold)
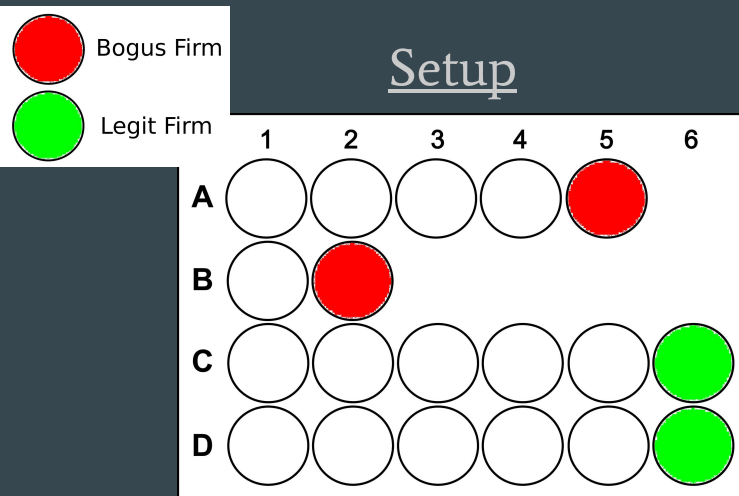
Evaluate performance

# Train

| Firm | Time | Features | bogus? |
|------|------|----------|--------|
| A | 1 | $X_{A,1}$ | $y_A=1$ |
| A | 2 | $X_{A,2}$ | $y_A=1$ |
| B | 2 | $X_{B,2}$ | $y_B=0$ |
| B | 3 | $X_{B,3}$ | $y_B=0$ |
| B | 5 | $X_{B,5}$ | $y_B=0$ |
| C | 3 | $X_{C,3}$ | $y_C=1$ |

(X,y)

train
1-period
model

M

# Predict

| Firm | Time | Features | | | Score | Aggregate |
|------|------|----------|---|---|-------|-----------|
| D | 1 | $X_{D,1}$ | X M y | | $S_{D,1}$ | |
| D | 2 | $X_{D,2}$ | M | | $S_{D,2}$ | $S_D=<S_{D,t}>$ |
| D | 3 | $X_{D,3}$ | M | | $S_{D,3}$ | |
| E | 3 | $X_{E,3}$ | M | | $S_{E,3}$ | |
| E | 5 | $X_{E,5}$ | M | | $S_{E,5}$ | $S_E$ |
| F | 3 | $X_{F,3}$ | M | | $S_{F,3}$ | $S_F$ |

# Point-in-time simulation



Train & Predict

T = 6

Setup

Train | Predict

T = 3

Money saved

Roll back observations to the state of knowledge at time period T

Bogus Firm

Legit Firm