Outline

- Occam's Razor

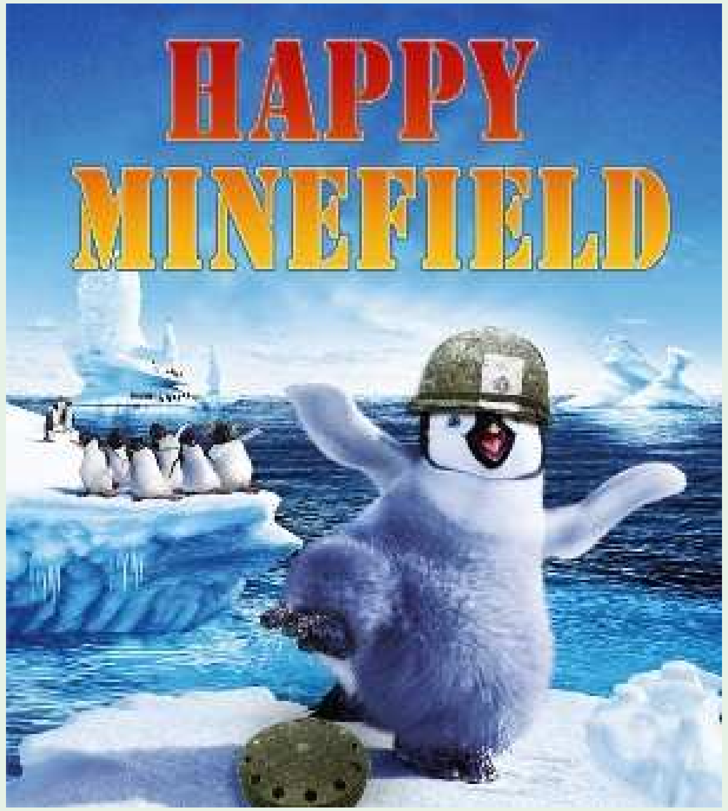
- Sampling Bias

- Data Snooping

The principle

If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.

Most common trap for practitioners - many ways to slip ☹



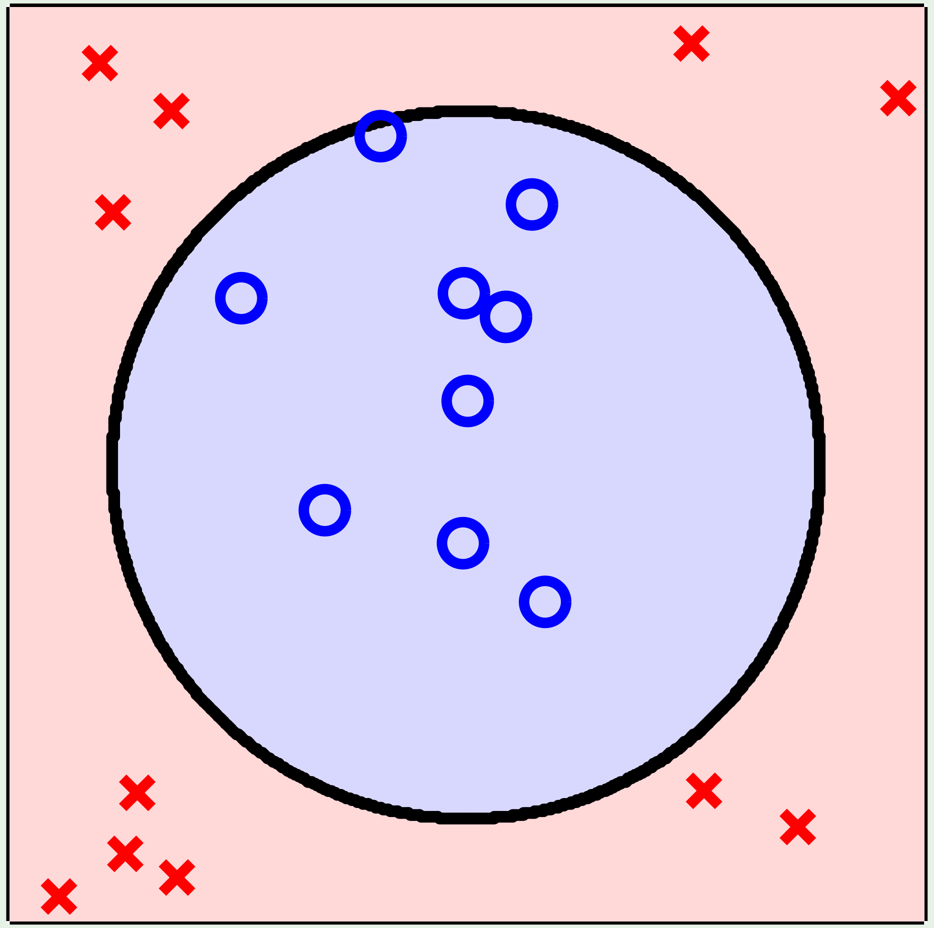
Looking at the data

Remember nonlinear transforms?

$\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$

or $\mathbf{z} = (1, x_1^2, x_2^2)$ or $\mathbf{z} = (1, x_1^2 + x_2^2)$

Snooping involves $\mathcal{D}$, not other information

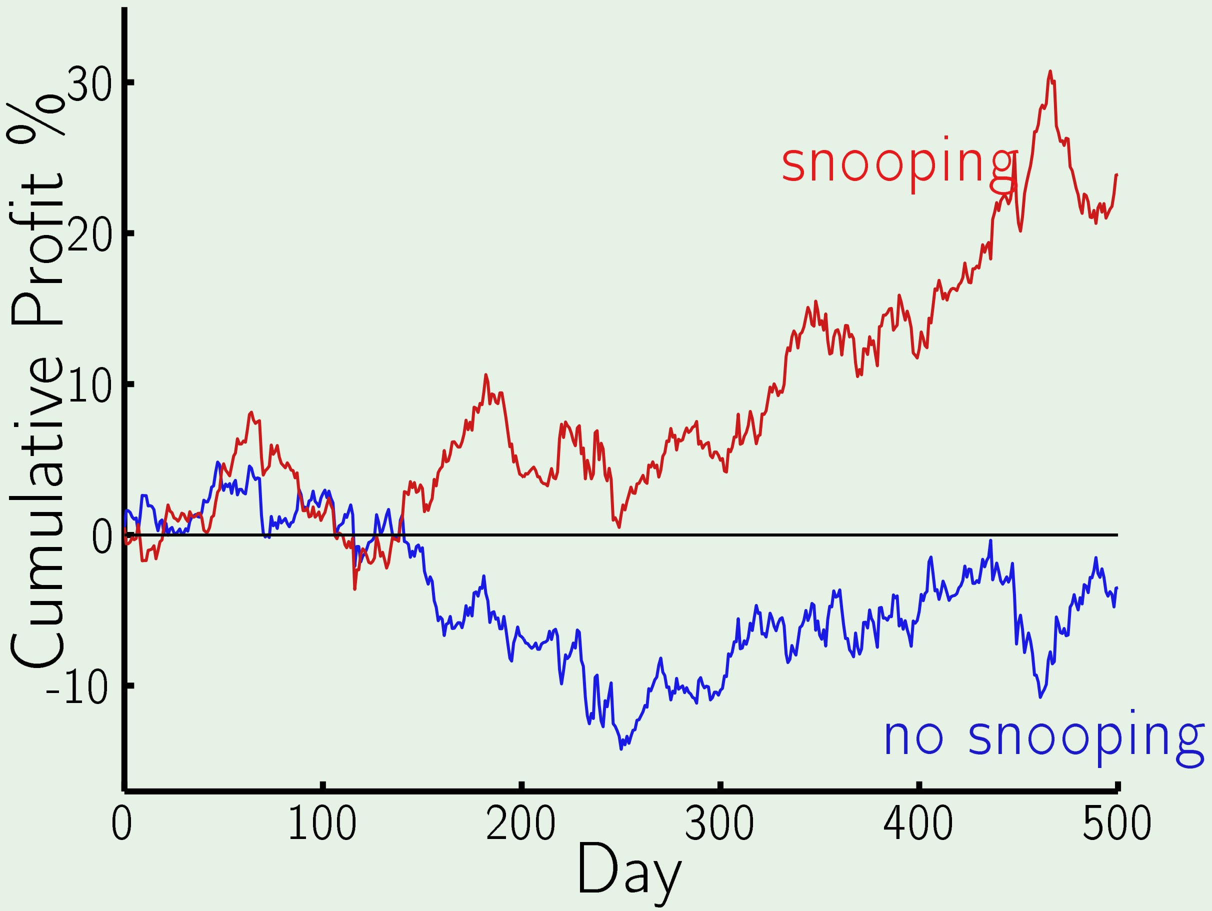


Puzzle 4: Financial forecasting

Predict US Dollar versus British Pound

Normalize data, split randomly: $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$

Train on $\mathcal{D}_{\text{train}}$ only, test g on $\mathcal{D}_{\text{test}}$



$$\Delta r_{-20}, \Delta r_{-19}, \cdots, \Delta r_{-1} \rightarrow \Delta r_0$$

Reuse of a data set

Trying one model after the other **on the same data set**, you will eventually 'succeed'

If you torture the data long enough, it will confess

VC dimension of the **total** learning model

May include what **others** tried!

Key problem: matching a *particular* data set

Two remedies

1. **Avoid** data snooping

   strict discipline

2. **Account for** data snooping

   how much data contamination

Puzzle 5: Bias via snooping

Testing long-term performance of "buy and hold" in stocks. Use **50 years** worth of data

- All currently traded companies in S&P500
- Assume you strictly followed buy and hold
- Would have made great profit!

Sampling bias caused by 'snooping'