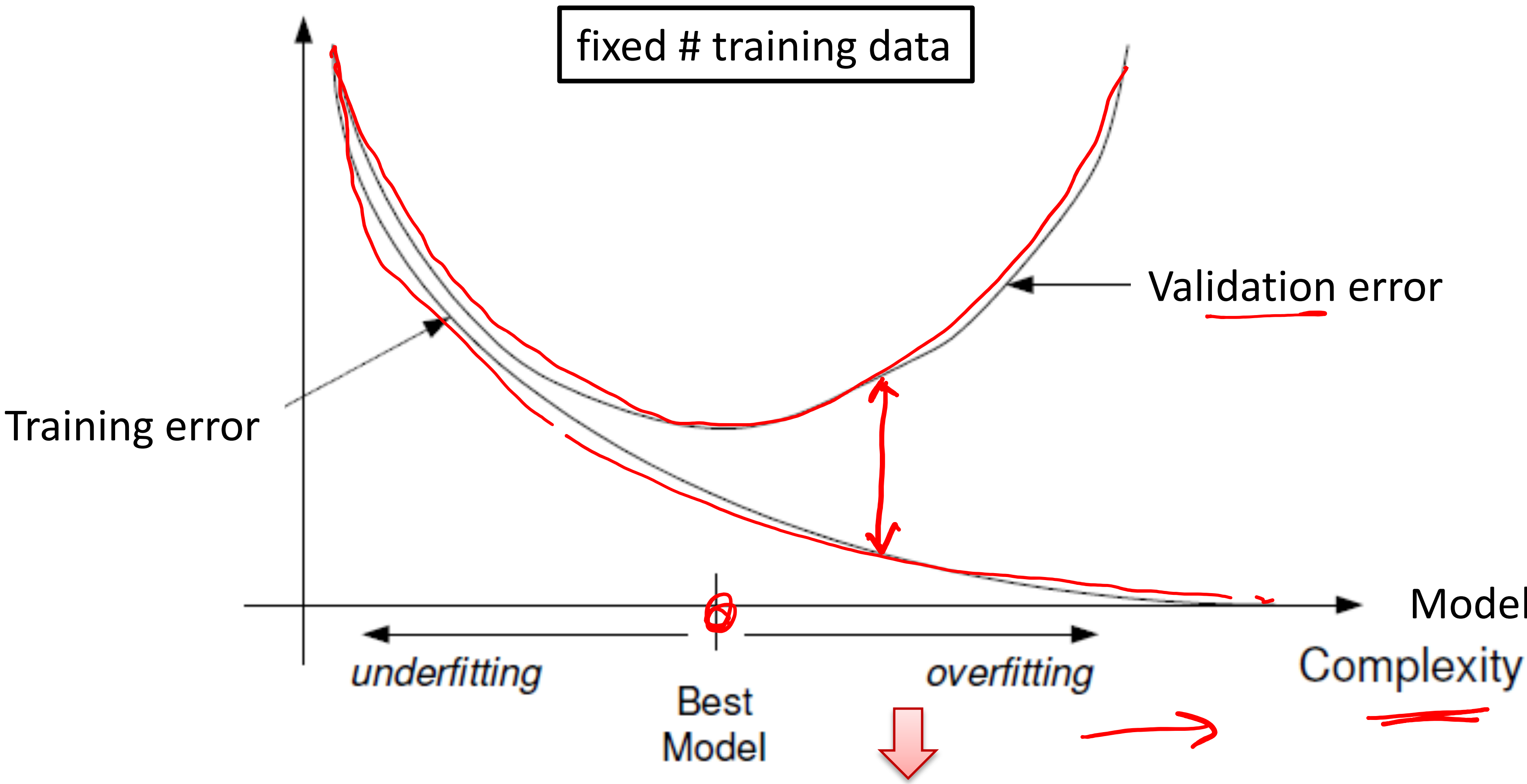
Model selection
Best practices

Aarti Singh & Geoff Gordon

Machine Learning 10-701
Mar 10, 2021

Training vs. Test Error



Training error is no longer a good indicator of test error

Examples of Model Spaces

Model Spaces with varying complexity:

- Nearest-Neighbor classifiers with increasing neighborhood sizes k = 1,2,3,…

  Large neighborhood => lower complexity

- Decision Trees with increasing depth k or with k leaves

  Higher depth/ More # leaves => higher complexity

- Neural Networks with increasing layers or nodes per layer

  More layers/Nodes per layer => higher complexity

- MAP estimates with stronger priors (larger hyper-parameters β_H, β_T for Beta distribution or smaller variance for Gaussian prior)

How can we select the right complexity model ?

Judging Test error

- Training error of a classifier f

$$\frac{1}{n}\sum_{i=1}^{n} 1_{f(X_i)\neq Y_i}$$

Training Data

$$\{X_i, Y_i\}_{i=1}^{n}$$

- What about test error?

  Can't compute it.

- How can we know classifier is not overfitting?

  Hold-out or Cross-validation

Hold-out method

Can judge test error by using an independent sample of data.

Hold - out procedure:

n data points available $D \equiv \{X_i, Y_i\}_{i=1}^n$

1) Split into two sets (randomly and preserving label proportion):

Training dataset $D_T = \{X_i, Y_i\}_{i=1}^m$

Validation/Hold-out dataset $D_V = \{X_i, Y_i\}_{i=m+1}^n$

often m = n/2

2) Train classifier on D_T. Report error on validation dataset D_V.

Overfitting if validation error is much larger than training error

Hold-out method

Drawbacks:

- May not have enough data to afford setting one subset aside for getting a sense of generalization abilities
- Validation error may be misleading (bad estimate of test error) if we get an “unfortunate” split

Limitations of hold-out can be overcome by a family of sub-sampling methods at the expense of more computation.

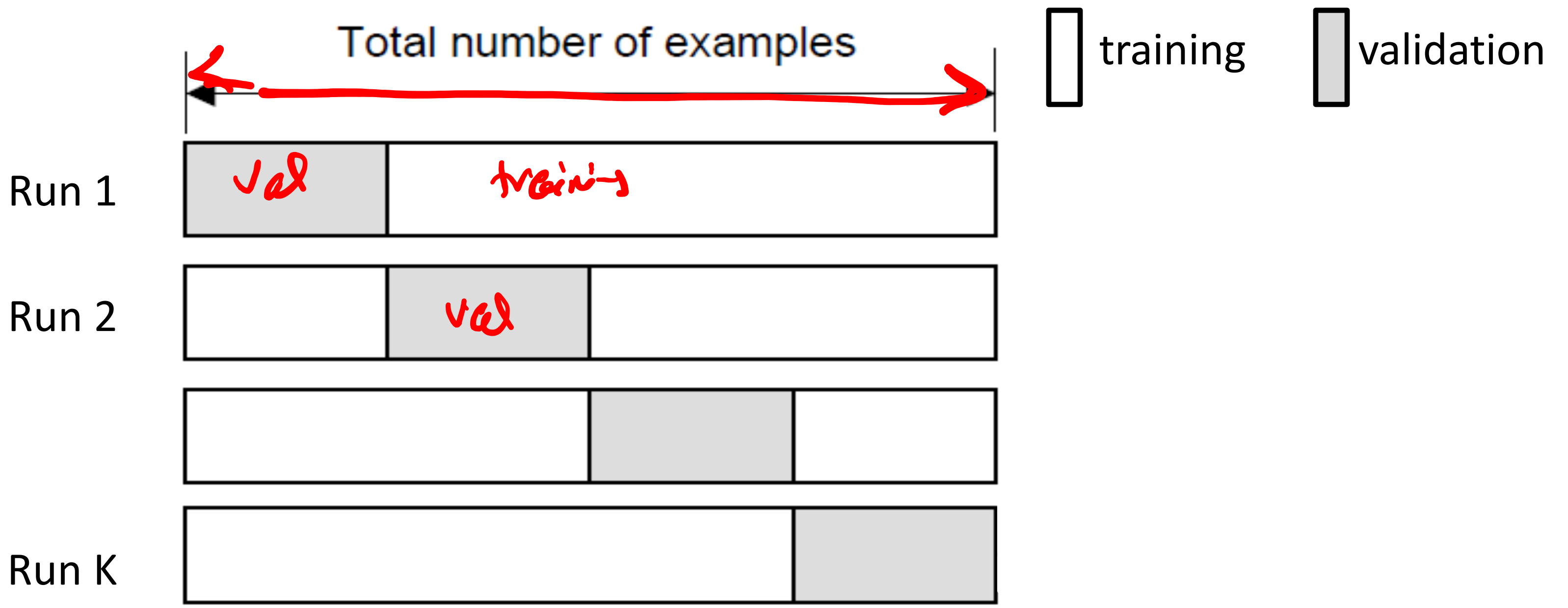
Cross-validation

K-fold cross-validation

Create K-fold partition of the dataset.

Do K runs: train using K-1 partitions and calculate validation error on remaining partition (rotating validation partition on each run).

Report average validation error

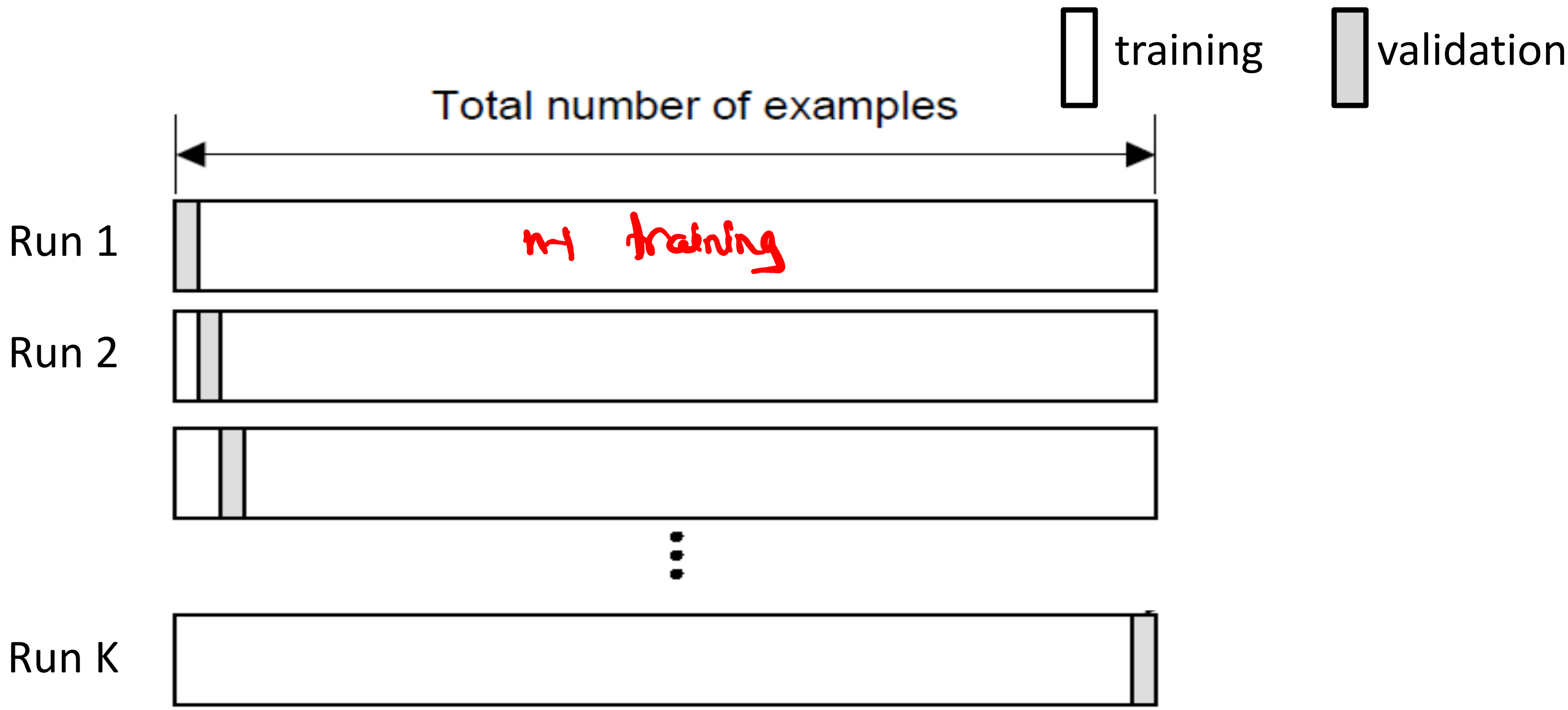


Cross-validation

Leave-one-out (LOO) cross-validation

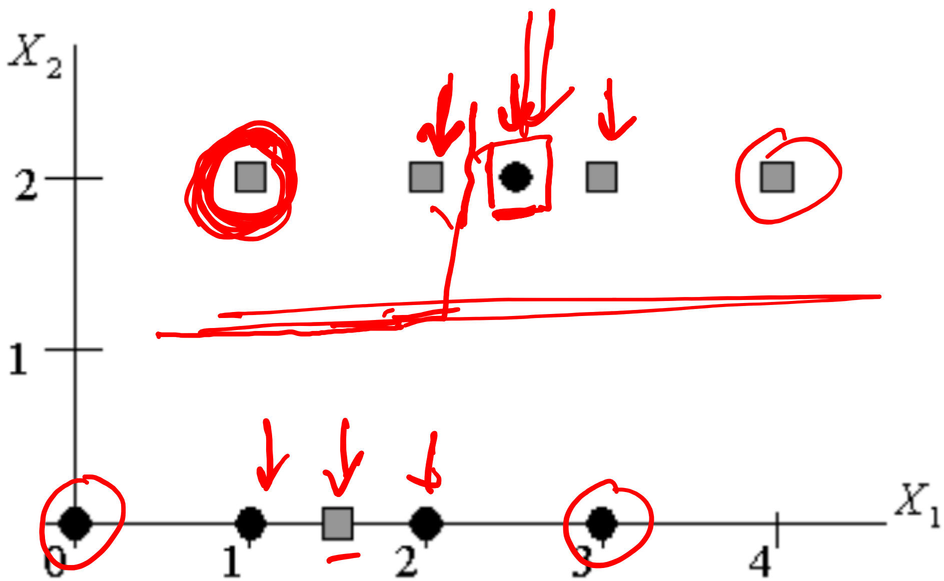
Special case of K-fold with K=n partitions

Equivalently, train on n-1 samples and validate on only one sample per run for n runs



Cross-validation

What is the leave-one-out cross-validation error of the given classifiers on the following dataset?



- Poll 1: Depth 1 Decision tree using best feature
- Poll 2: 1-NN classifier

Cross-validation

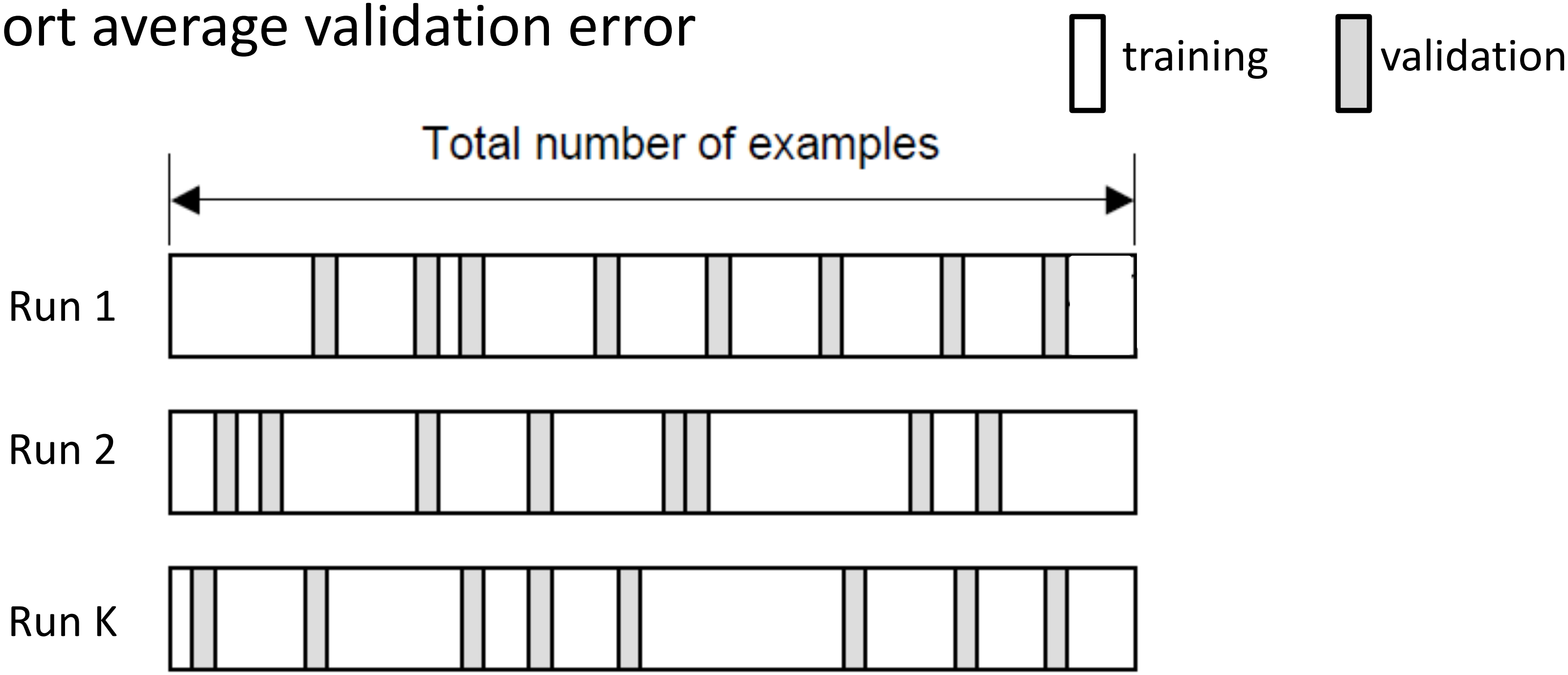
Random subsampling

Randomly subsample a fixed fraction αn ($0 < \alpha < 1$) of the dataset for validation.

Compute validation error with remaining data as training data.

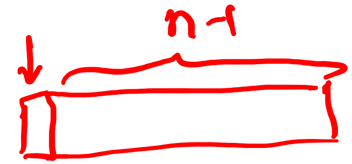
Repeat K times

Report average validation error



Practical Issues in Cross-validation

How to decide the values for *K* and *α*?



- Large K (more data for training)

  + Validation error can approximate test error well

  - Observed validation error will be unstable (few validation pts)

  - The computational time will be very large as well (several runs)

- Small K (less data for training)

  + The #runs and, therefore, computation time are reduced

  + Observed validation error will be stable (many validation pts)

  - Validation error cannot approximate test error well

Common choice: K = 10, $\alpha = 0.1$ ☺

Model selection using Hold-out/Cross-validation

- Train models of different complexities and evaluate their validation error using hold-out or cross-validation

- Pick model with smallest validation error (averaged over different runs for cross-validation)

➢ When using hold-out or cross-validation for model selection, test error should be reported using independent data

ML best practices

- Training vs Validation vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose
- Reproducibility
- Interpreting results

ML best practices

- ➢ Training vs Validation vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose
- Reproducibility
- Interpreting results

ML best practices

- Training vs Validation vs Testing accuracy
- ➢ Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose
- Reproducibility
- Interpreting results

Baselines are extremely important: biased classes

Accuracy of classifier

➢ Are these good classifiers?

	Test accuracy
• Classifier 1	92%
• Classifier 2	87%

Test dataset had 9300 normal patients and 700 patients with cancer

Baselines are extremely important: multiple classes

Accuracy of classifier

➢ Are these good classifiers?

	Test accuracy
• Classifier 1	52%
• Classifier 2	44%

Test dataset 10000 images: 2 classes, 5000 images each

Test dataset 10000 images: 10 classes, 1000 images each

Baselines are extremely important: regression

Accuracy of regressor

➢ Are these good predictors?

	Test Mean Squared Error	
• Regressor 1	25	✓
• Regressor 2	100	✗

Standard deviation of test data ~7

$(x - E[x])^2 \approx 49$

MSE vs $R^2 := 1 - \text{MSE}/\text{Variance}$

(Fraction of variance explained by predictor)

ML best practices

- Training vs Validation vs Testing accuracy
- Baselines
- ➢ Mean vs Best accuracy
- ➢ Standard deviation (Std)
- Underlying goal/purpose
- Reproducibility
- Interpreting results

Best run test accuracy doesn't make a classifier better

Test Accuracy of classifier

	Mean	Best run
• Classifier 1	92%	97%
• Classifier 2	87%	100%

High mean test accuracy doesn't make a classifier better

Test Accuracy of classifier

	Mean
• Classifier 1	92%
• Classifier 2	87%

High mean test accuracy doesn't make a classifier better

Test Accuracy of classifier

	Mean	Std
• Classifier 1	92%	15%
• Classifier 2	87%	5%

High mean test accuracy doesn't make a classifier better

Test Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

ML best practices

- Training vs Validation vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- ➢ Underlying goal/purpose
- Reproducibility
- Interpreting results

Purpose often dictates validity of classifier

Test Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

- ➢ Which classifier would you choose when recommending movies?
- ➢ Which classifier would you choose when diagnosing serious illness?

Purpose often dictates validity of regressor

Accuracy of regressor

- Are these good predictors?

	MSE
• Regressor 1	25
• Regressor 2	0.0001

Purpose often dictates validity of regressor

Test Accuracy of regressor

➢ Are these good predictors?

	MSE	Task
• Regressor 1	25	Predict age of a person
• Regressor 2	0.0001	Predict proportion of lead in water

MS(quared)E vs. MA(bsolute)E

Units important

ML best practices

- Training vs Validation vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose
- ➢ Reproducibility
- Interpreting results

Reproducibility

- All model choices mentioned?
  - Model order, Step-size, batch-size, initialization, order of cross-validation, training/validation/test/hold-out set size, …
- Experimental platform details?
  - Which GPUs, CPUs, memory, …
- Data and code availability?
- Proof details?

ML best practices

- Training vs Validation vs Testing accuracy
- Baselines
- Mean vs Best accuracy
- Standard deviation
- Underlying goal/purpose
- Reproducibility
- ➢ Interpreting results

Interpreting 'correct' results correctly is important too

- Correlation vs Causation

(Some measure of) Structure of brain is correlated with (some measure of) Function of brain, hence structure shapes function.

- Higher-order dependence

Expression of gene A is uncorrelated (or has statistically insignificant correlation) with gene B, hence gene A has no influence on expression of gene B; OR hence genes A and B function independently.

Interpreting ‘correct’ results correctly is important too

- Confounding variables

  Given data from a surveillance camera, an ML algorithm could predict with high accuracy when a subway is busy. Hence, it has learnt to detect crowd.

  Given images of US and Russian tanks, an ML algorithm could classify them with high accuracy. Hence, it learnt to distinguish between their salient capabilities.