Machine Learning Basics

Zhijun Yin, PhD, MS

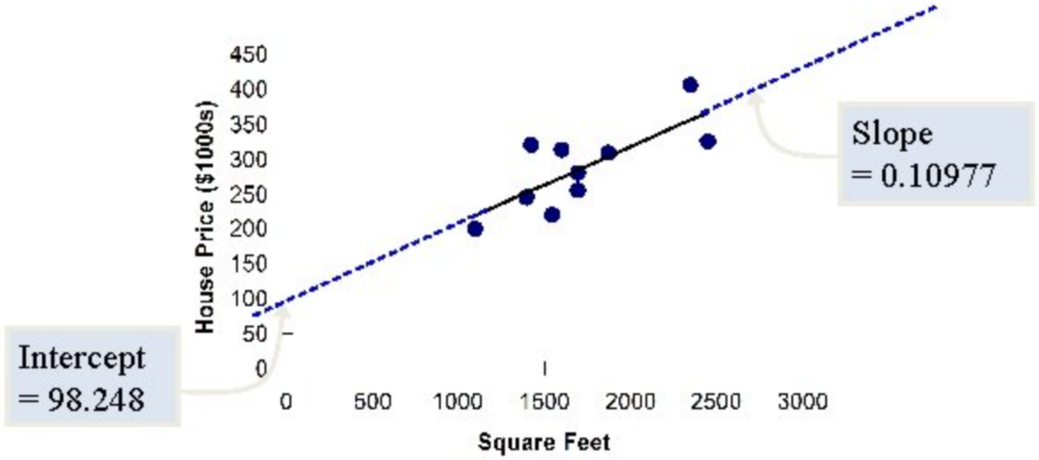
Assistant Professor of Biomedical Informatics and Computer Science

Vanderbilt University Medical Center

Wednesday, June 14, 2023

A Simple Linear Regression

- Suppose that we want to predict the house price based on the square feet, what should we do?
- We collect 10 records from Zillow to build our knowledge about the relationship between house price and square feet
- While we can easily process the 10 data points within our mind, we plan to build a function to learn about the relationship.



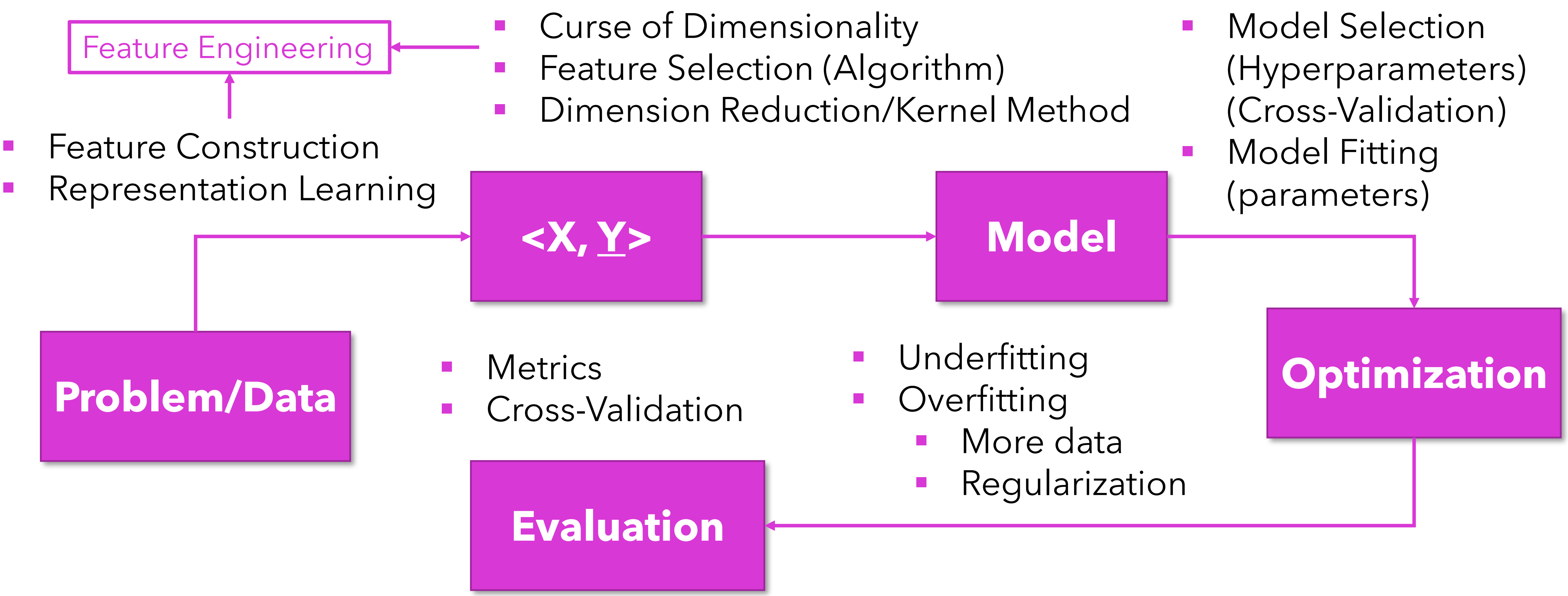
$$\text{house price} = 98.24833 + 0.10977\,(\text{square feet})$$

Machine Learning

Linear algebra and **probability/statistics** and **optimization** are the mathematical pillars of machine learning

To construct a function that classifies the training data correctly, so it can generalize to unseen test data

Machine Learning Pipeline



Feature Engineering

Represent each data point, x, in a dataset / domain, X, with a vector of features/predictors/independent variables

$$x = (x_1, x_2, \ldots, x_p)$$

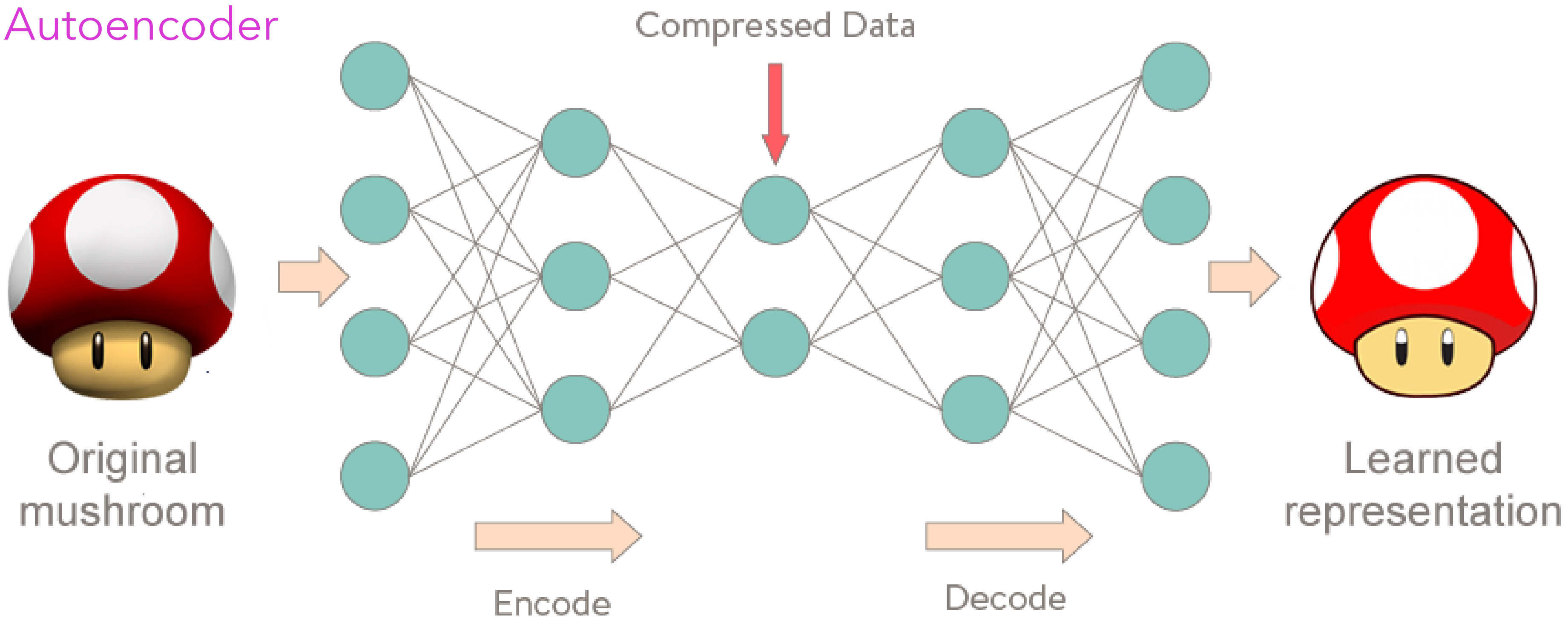
Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.

– Andrew Ng, Machine Learning and AI via Brain simulations

Representation Learning

- Unsupervised learning
- Learn dense representation
- Feed into deep learning models

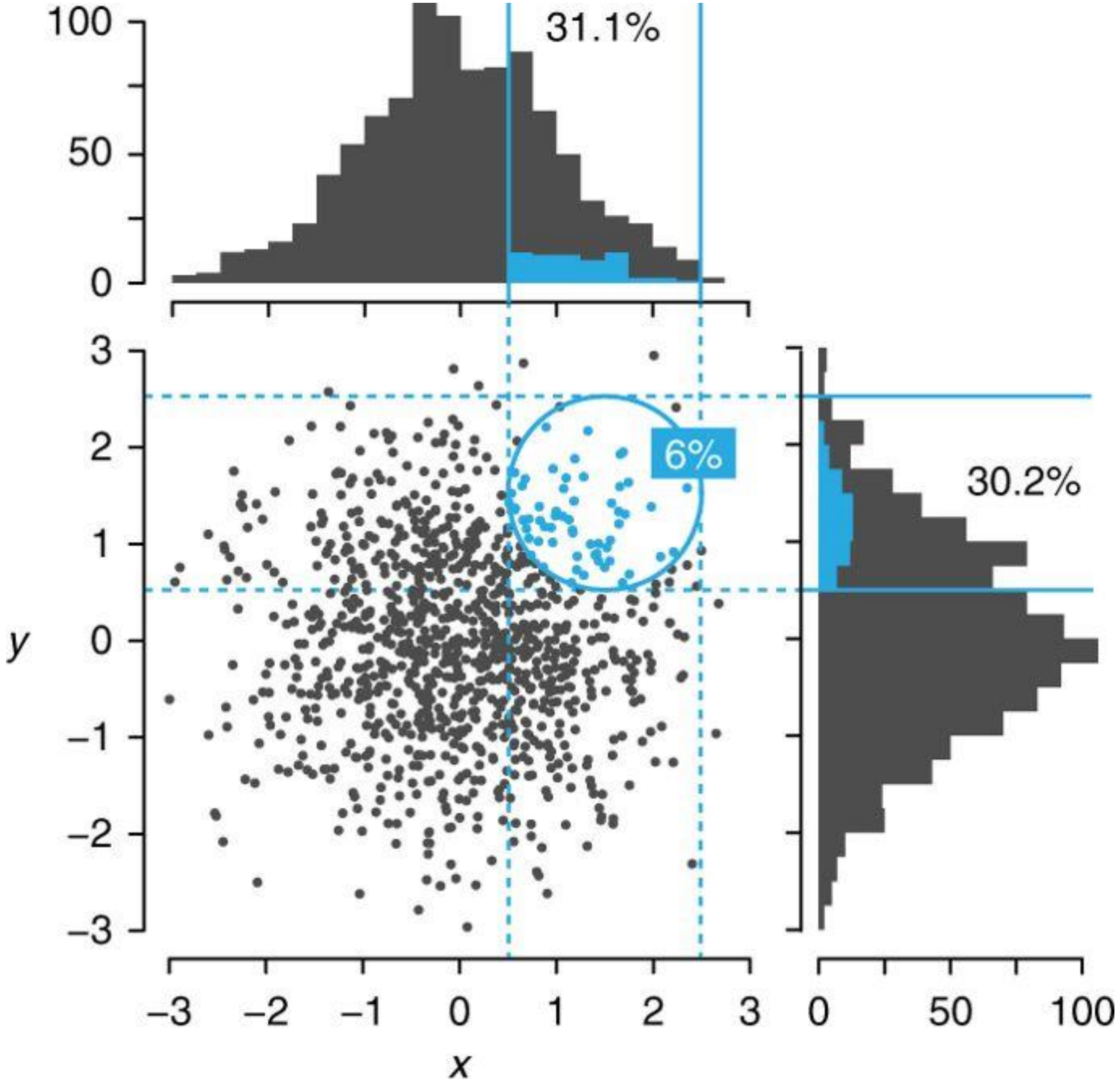
Example: Autoencoder



https://blog.myyellowroad.com/unsupervised-sentence-representation-with-deep-learning-104b90079a93

Curse of Dimensionality

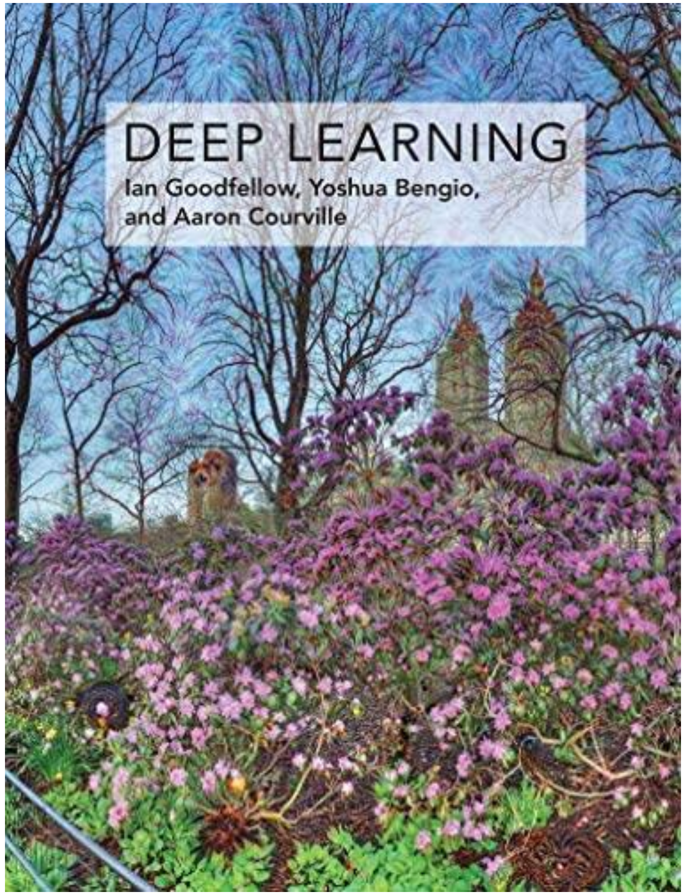
Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. σ = 1, only 6% fall within σ of (x, y) = (1.5, 1.5) (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins within blue solid lines) are within σ of 1.5. Blue bins in histograms correspond to the blue points.



Altman, Naomi, and Martin Krzywinski. "The curse (s) of dimensionality." *Nat Methods* 15 (2018): 399-400.

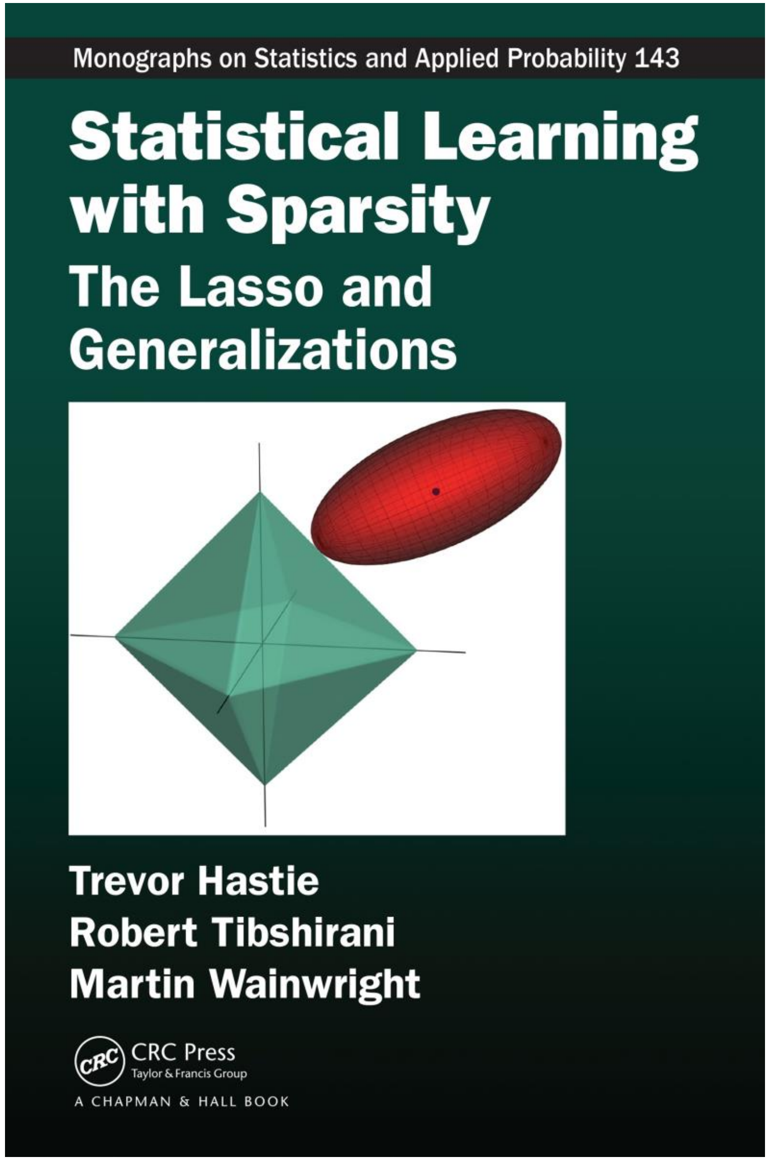
Manifold Learning

Many machine learning problems seem hopeless if we expect the machine learning algorithm to learn functions with interesting variations across all of $\mathbb{R}^n$. **Manifold learning** algorithms surmount this obstacle by assuming that most of $\mathbb{R}^n$ consists of invalid inputs, and that interesting inputs occur only along a collection of manifolds containing a small subset of points, with interesting variations in the output of the learned function occurring only along directions that lie on the manifold, or with interesting variations happening only when we move from one manifold to another. Manifold learning was introduced in the case of continuous-valued data and in the unsupervised learning setting, although this probability concentration idea can be generalized to both discrete data and the supervised learning setting: the key assumption remains that probability mass is highly concentrated.



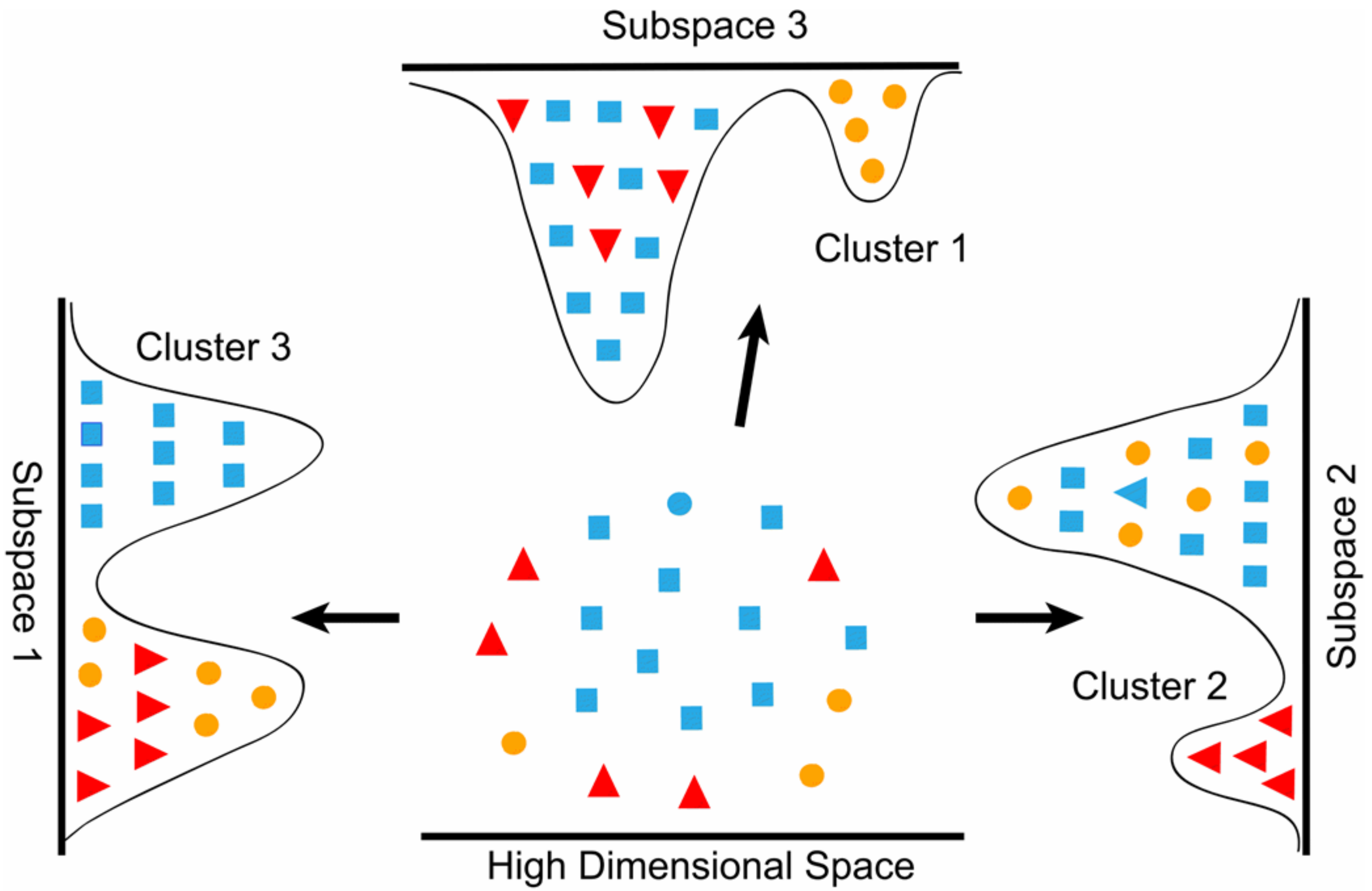
Sparse Models

- "We are drowning in information and starving for knowledge." -Rutherford D. Roger
- A sparse statistical model is one in which only a relatively **small number of parameters** (or predictors) play an important role.
- The advantages of sparsity are **interpretation** of the fitted model and **computational convenience**



Dimension Reduction

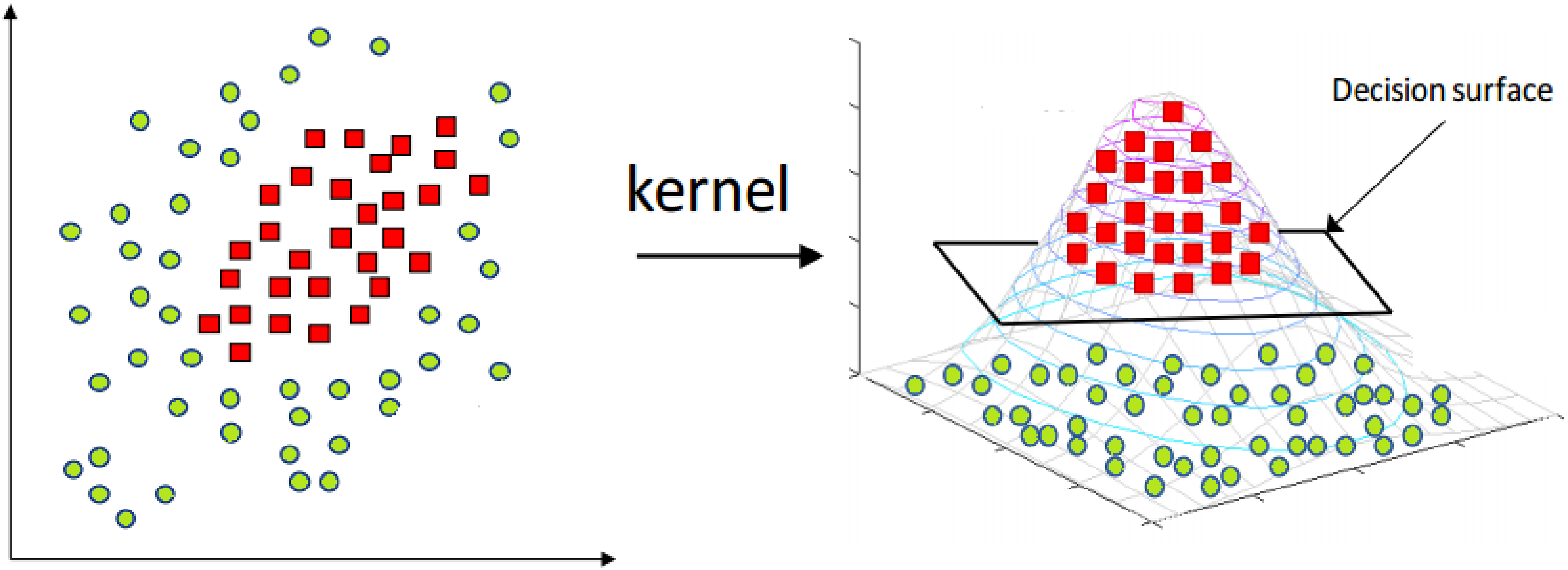
- Feature selection
- Principal Component Analysis
- t-Distributed Stochastic Neighbor Embedding (t-SNE)



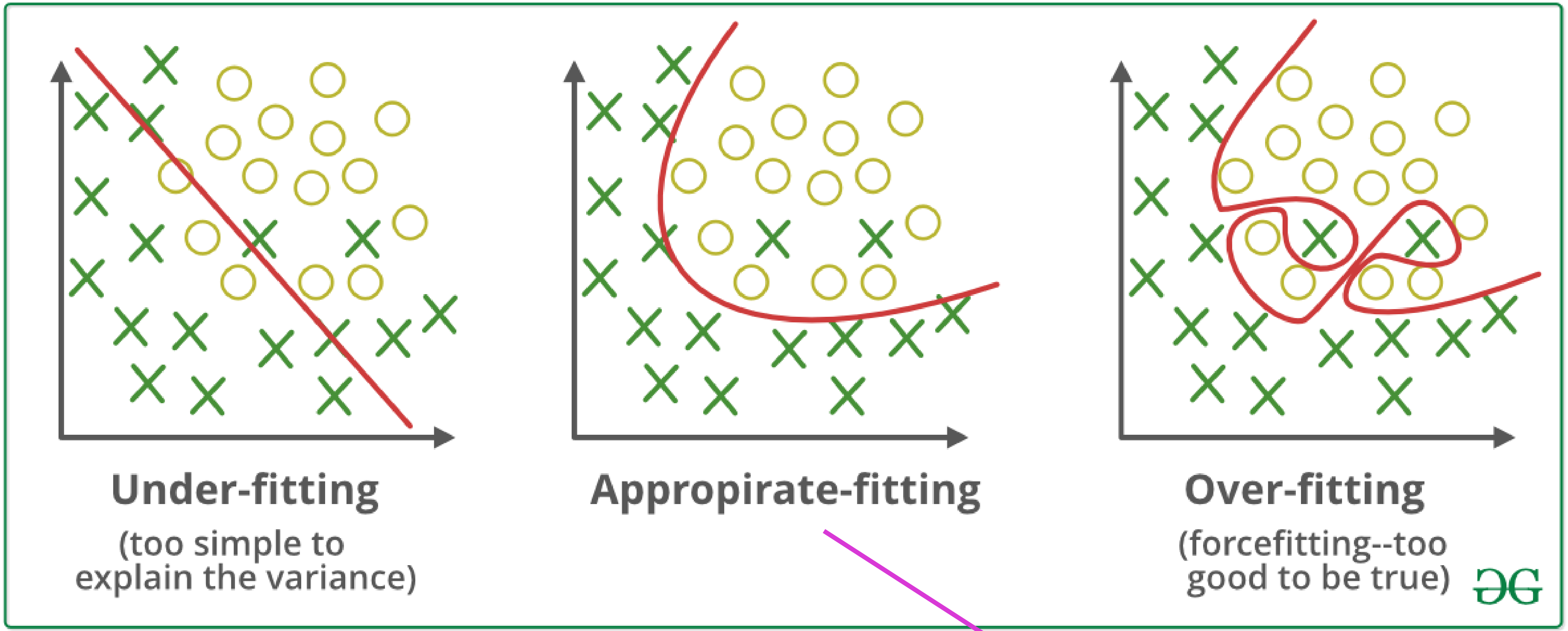
https://i.stack.imgur.com/Klvel.png

Separating data in higher dimension space might be much easier, efficient

Kernel Method (trick)



https://maelfabien.github.io/machinelearning/largescale/



Underfitting vs. Overfitting

Generalizability

https://www.geeksforgeeks.org/regularization-in-machine-learning/

Regularization

Lasso (L1):

Hyperparameter (model complexity)
(model selection)

$$J(\beta) = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=0}^{P} |\beta_j|$$

Parameter (model fitting)

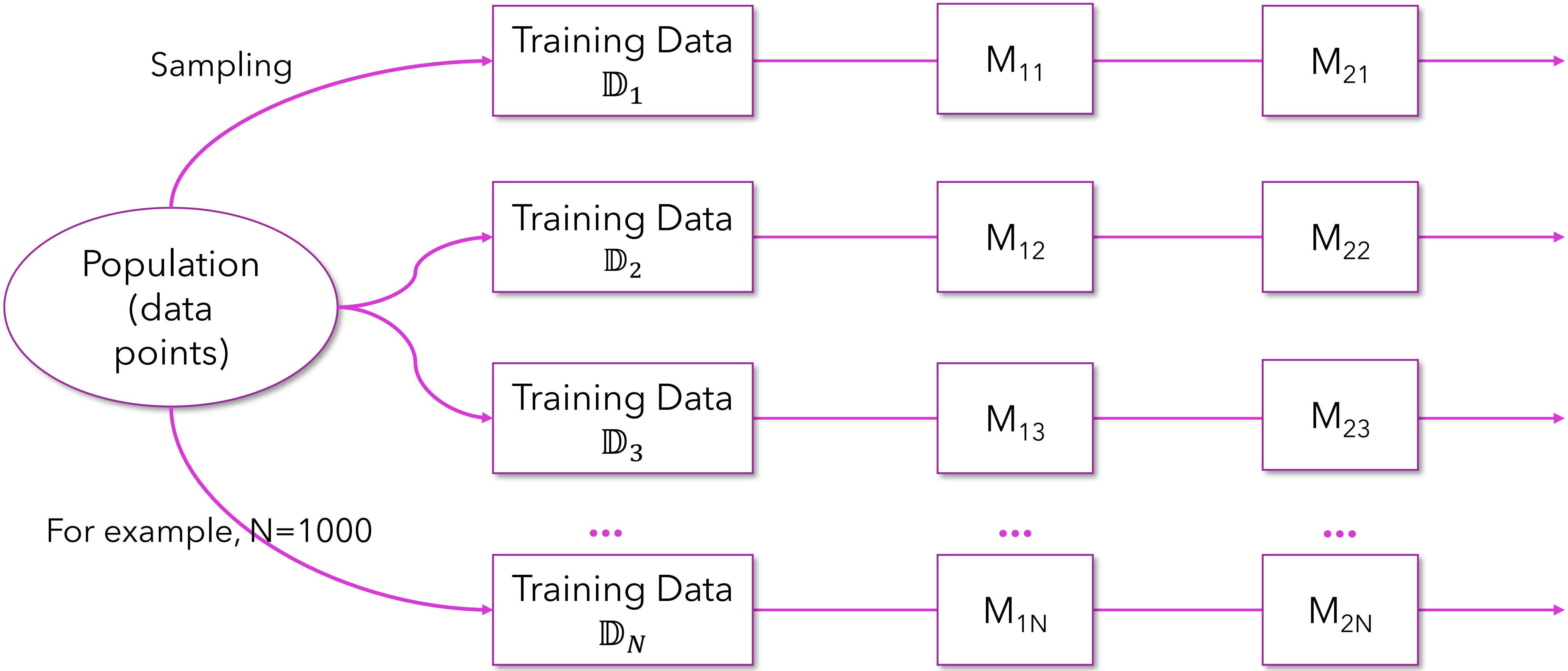
Ridge (L2):

$$J(\beta) = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{P} x_{ij}\beta_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{P} |\beta_j|^2}_{\text{Regularization Item}}$$

Model Comparison

“Utopian”

Compare M_1 and M_2 in a Perfect Situation

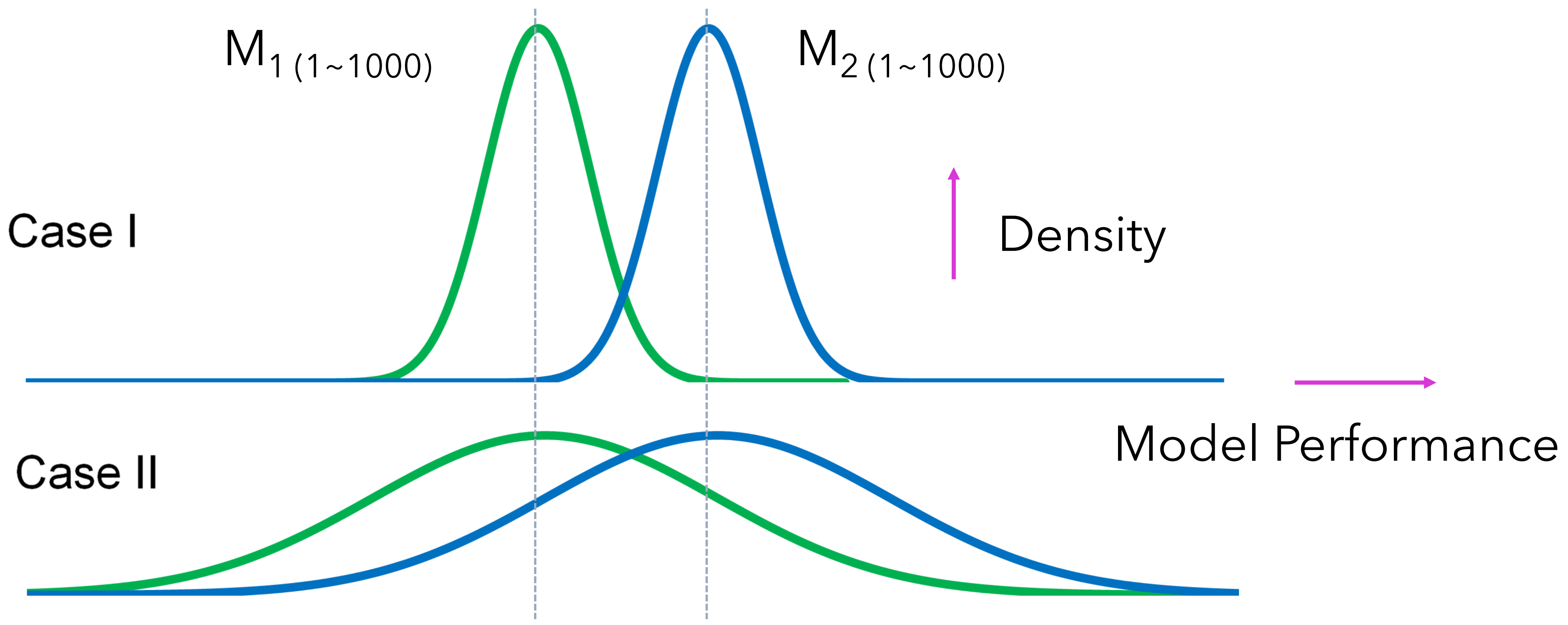


Compare M_1 and M_2

Can we reply on any single sampled training dataset, say, $\mathbb{D}_{17}$, to determine which model performs better?

By keeping sampling training dataset $\mathbb{D}_i$,

- we can always find $\mathbb{D}_m$, where M_1 outperforms M_2
- we can always find $\mathbb{D}_n$, where M_2 outperforms M_1



Which model performs better?

Null Hypothesis (H_0)

- All statistical significance tests start with a null hypothesis
- A statistical significance test measures the strength of evidence that the data sample supplies for or against some proposition of interest
- This proposition is known as a 'null hypothesis'
- It usually relates to there being 'no difference' between groups' or 'no effect' of a treatment

https://www.ucl.ac.uk/child-health/short-courses-events/about-statistical-courses/research-methods-and-statistics/chapter-6

Alternative Hypothesis (H_a)

A statement of what a statistical hypothesis test is set up to establish. For example,

- In clinical trial of new drug, H_a might be the drug has effect, on average, compared to current drug
- In machine learning, H_a might be M_1 outperforms M_2
- $\underbrace{H_0: \mu = k,\ H_a: \mu > k;\ H_0: \mu = k,\ H_a: \mu < k;}_{\textit{One-sided test}}\ \underbrace{H_0: \mu = k,\ H_a: \mu \neq k}_{\textit{Two-sided test}}$

P-value

- The probability of obtaining the observed data sample if the null hypothesis were true
- Smaller p-values suggest that the null hypothesis is less likely to be true
- We have NOT disproved the null hypothesis; the sample is unlikely BUT NOT IMPOSSIBLE

https://www.ucl.ac.uk/child-health/short-courses-events/about-statistical-courses/research-methods-and-statistics/chapter-6

Level of Significance (α)

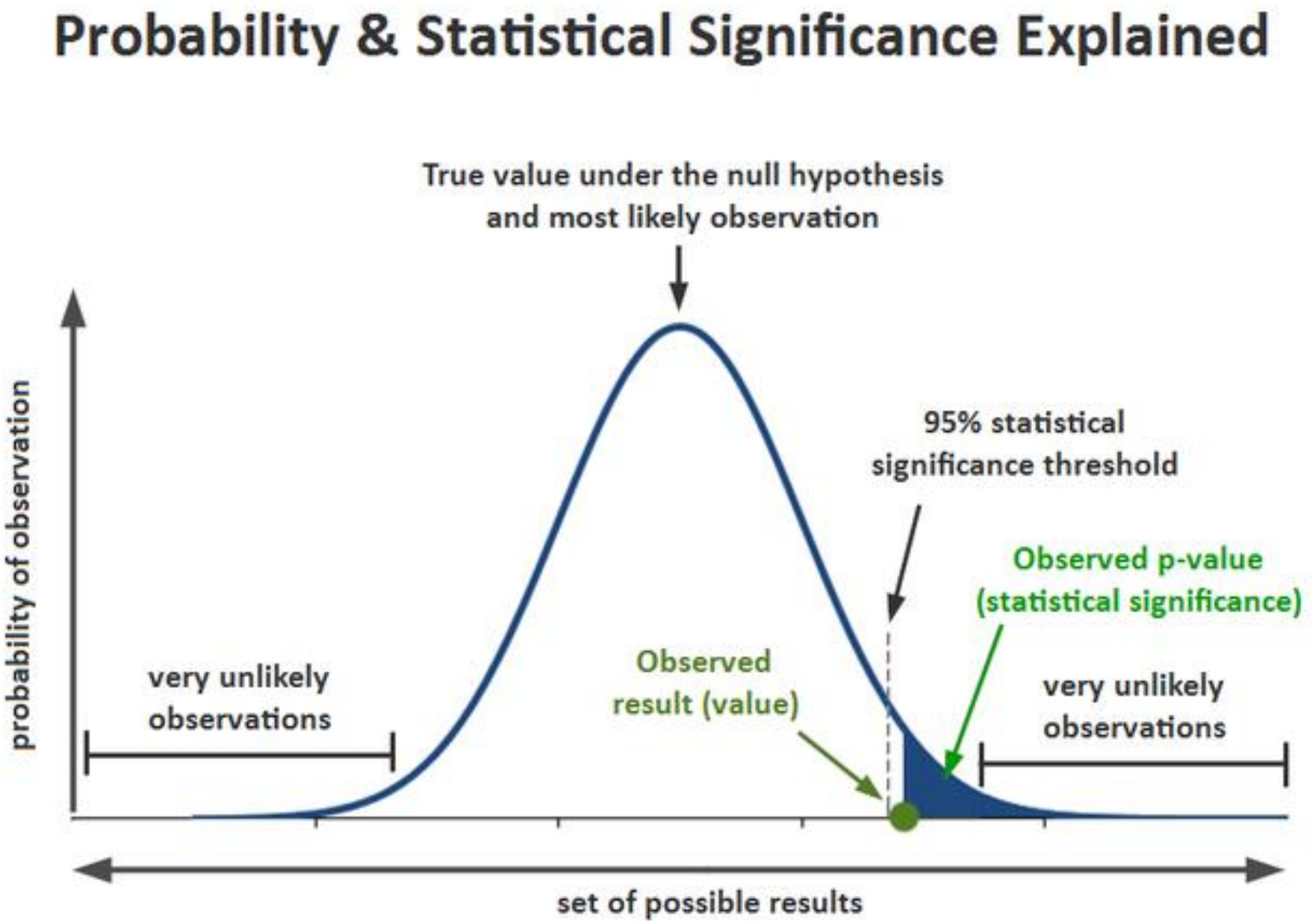
It acts like a threshold, and purely for report being significant purpose.

- In (traditional) practice, α is set to be 0.05. A p-value of less than 0.05 is called as significant
- p-value is a probability, there is no sudden changeover from being unlikely to being likely
- It is always best to report the actual p-values

https://www.ucl.ac.uk/child-health/short-courses-events/about-statistical-courses/research-methods-and-statistics/chapter-6

Hypothesis Testing and P-value

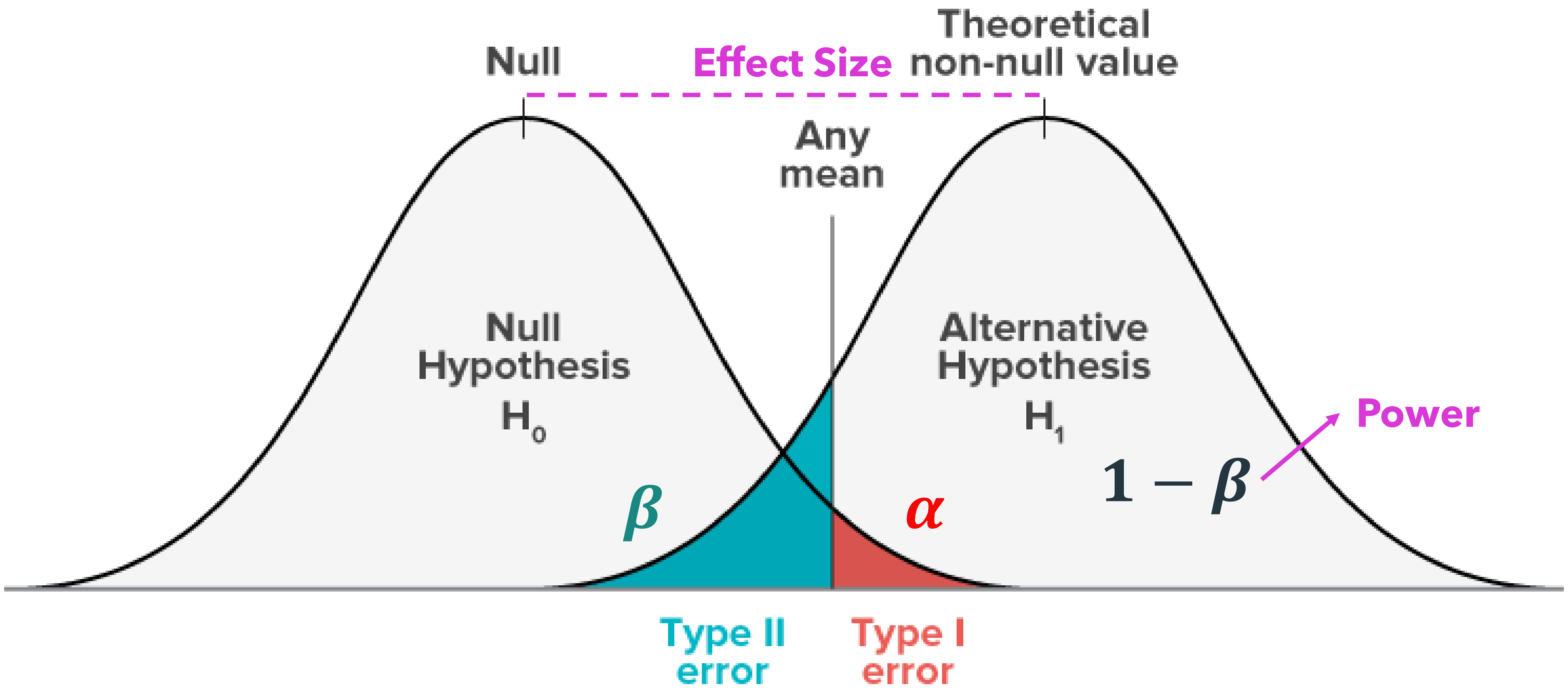
The maximal turning point is located at the expected value. Until the grey line for the 95% significance threshold is reached, all values verifying the null hypothesis are between the very unlikely observations



https://steemit.com/steemstem/@aximot/statistical-hypothesis-testing-the-lottery-tickets

Type I, II Error

Demo



Type I, II Error and Metrics

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Precision ↑

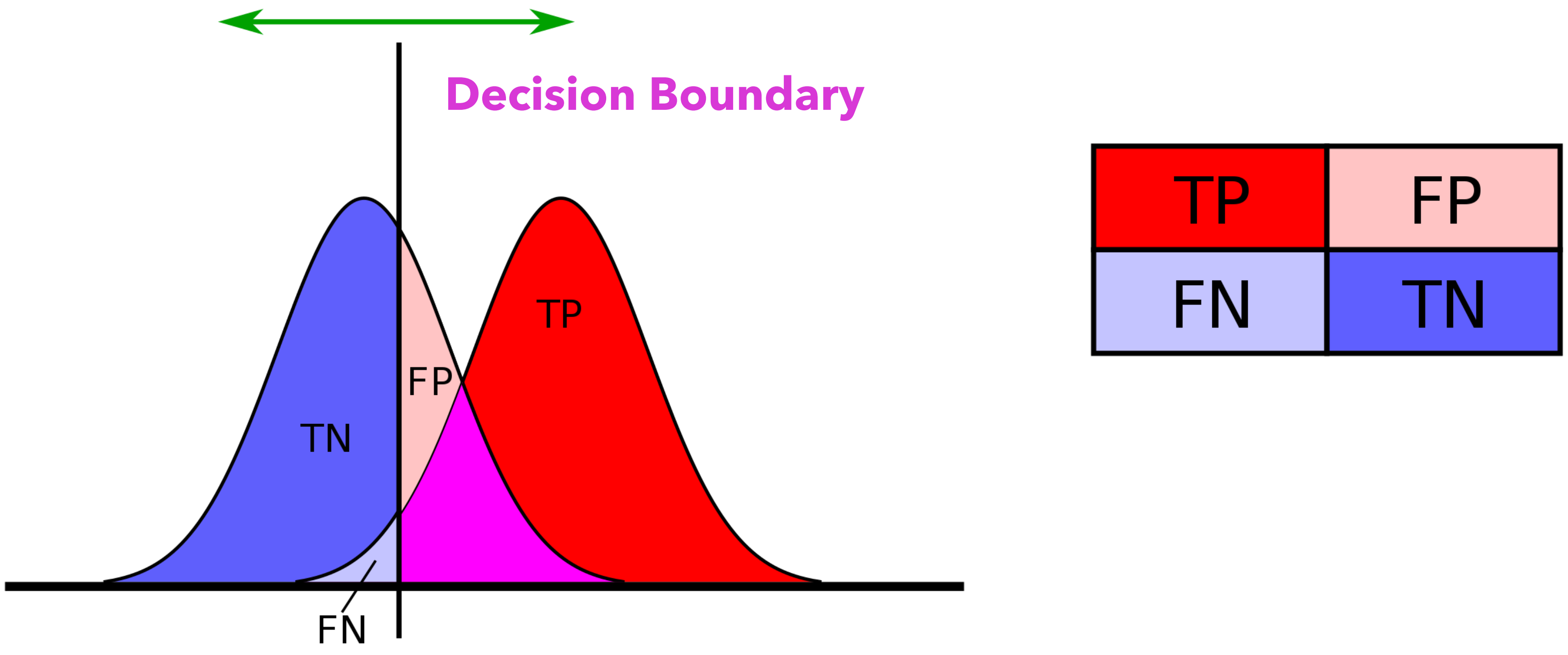
Recall ←

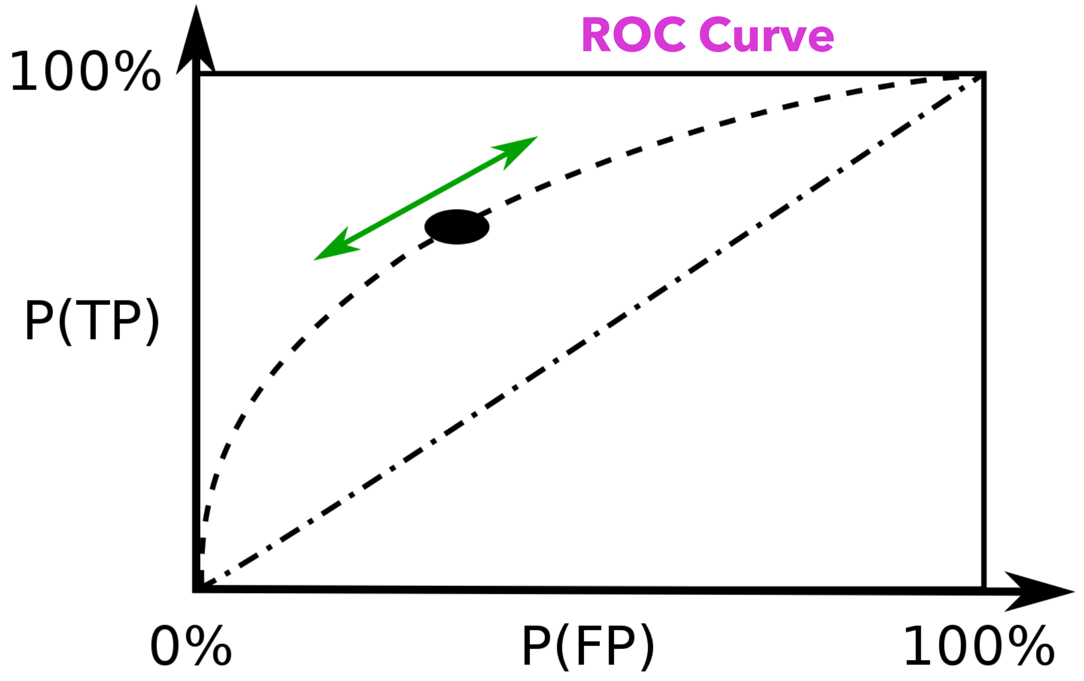
https://tex.stackexchange.com/questions/449762/want-to-write-a-confusion-matrix-in-latex

		CONDITION determined by "Gold Standard"			
	TOTAL POPULATION	CONDITION POS	CONDITION NEG	PREVALENCE $\frac{\text{CONDITION POS}}{\text{TOTAL POPULATION}}$	
TEST OUT-COME	TEST POS	True Pos TP	*Type I Error* False Pos FP	*Precision* Pos Predictive Value PPV = $\frac{\text{TP}}{\text{TEST P}}$	False Discovery Rate FDR = $\frac{\text{FP}}{\text{TEST P}}$
	TEST NEG	*Type II Error* False Neg FN	True Neg TN	False Omission Rate FOR = $\frac{\text{FN}}{\text{TEST N}}$	Neg Predictive Value NPV = $\frac{\text{TN}}{\text{TEST N}}$
	ACCURACY ACC ACC = $\frac{\text{TP + TN}}{\text{TOT POP}}$	*Sensitivity (SN), Recall* Total Pos Rate TPR TPR = $\frac{\text{TP}}{\text{CONDITION POS}}$	*Fall-Out* False Pos Rate FPR FPR = $\frac{\text{FP}}{\text{CONDITION NEG}}$	Pos Likelihood Ratio LR + LR + = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio DOR DOR = $\frac{\text{LR +}}{\text{LR -}}$
		Miss Rate False Neg Rate FNR FNR = $\frac{\text{FN}}{\text{CONDITION POS}}$	*Specificity (SPC)* True Neg Rate TNR TNR = $\frac{\text{TN}}{\text{CONDITION NEG}}$	Neg Likelihood Ratio LR - LR - = $\frac{\text{TNR}}{\text{FNR}}$	

Type I, II Error and AUC

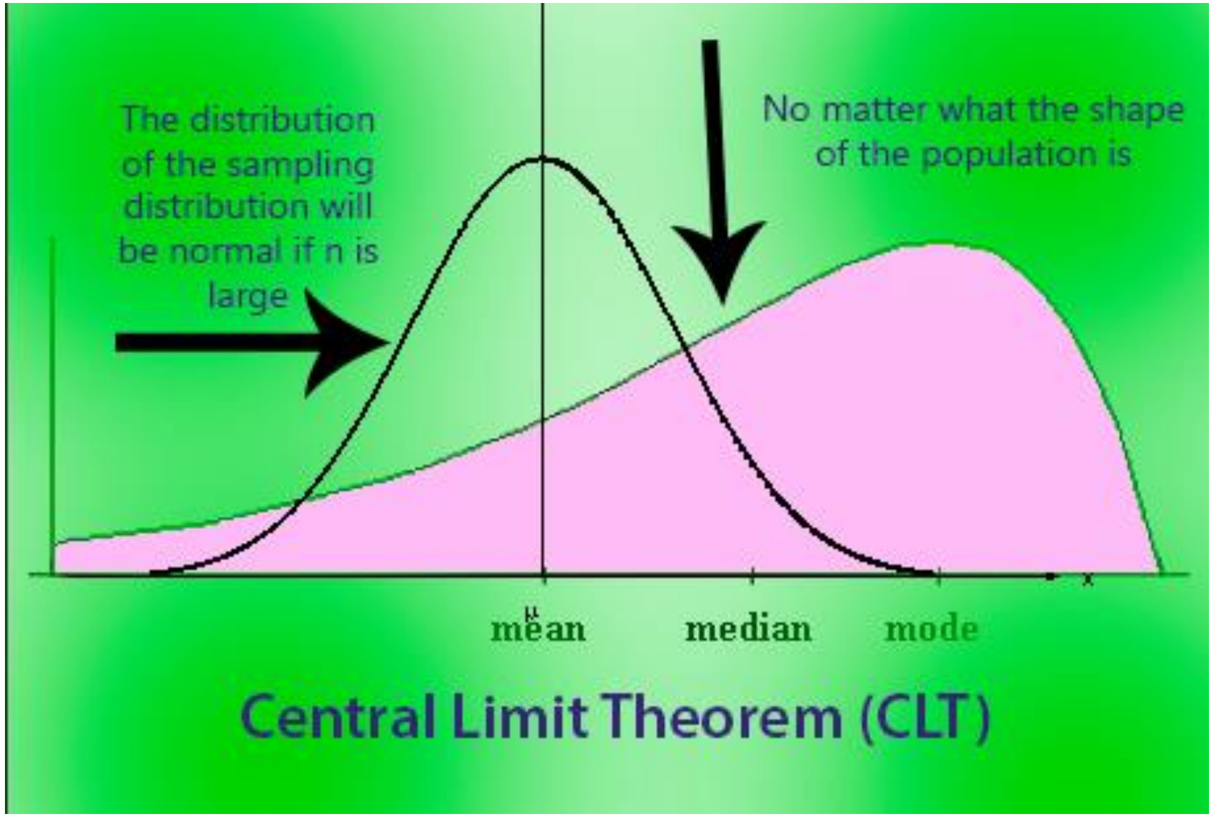
Area Under the Receiver Operating Characteristics (ROC) Curve





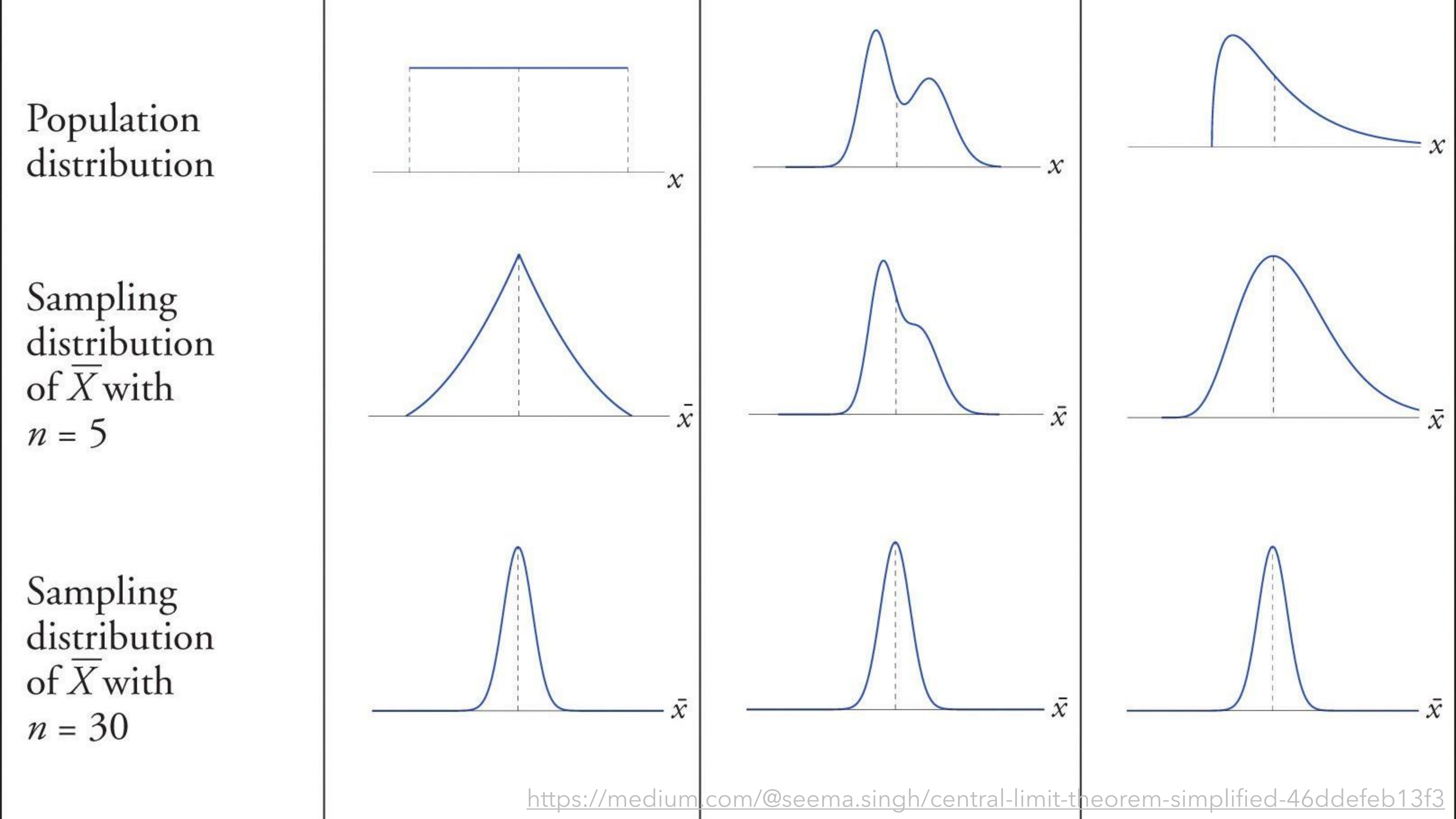
https://en.wikipedia.org/wiki/Type_I_and_type_II_errors

Central Limit Theorem



For large sample sizes, the sampling distribution of means will approximate to normal distribution even if the population distribution is not normal.

https://medium.com/@seema.singh/central-limit-theorem-simplified-46ddefeb13f3


Population distribution
Sampling distribution of $\overline{X}$ with $n = 5$
Sampling distribution of $\overline{X}$ with $n = 30$
x
x
x
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$
$\bar{x}$

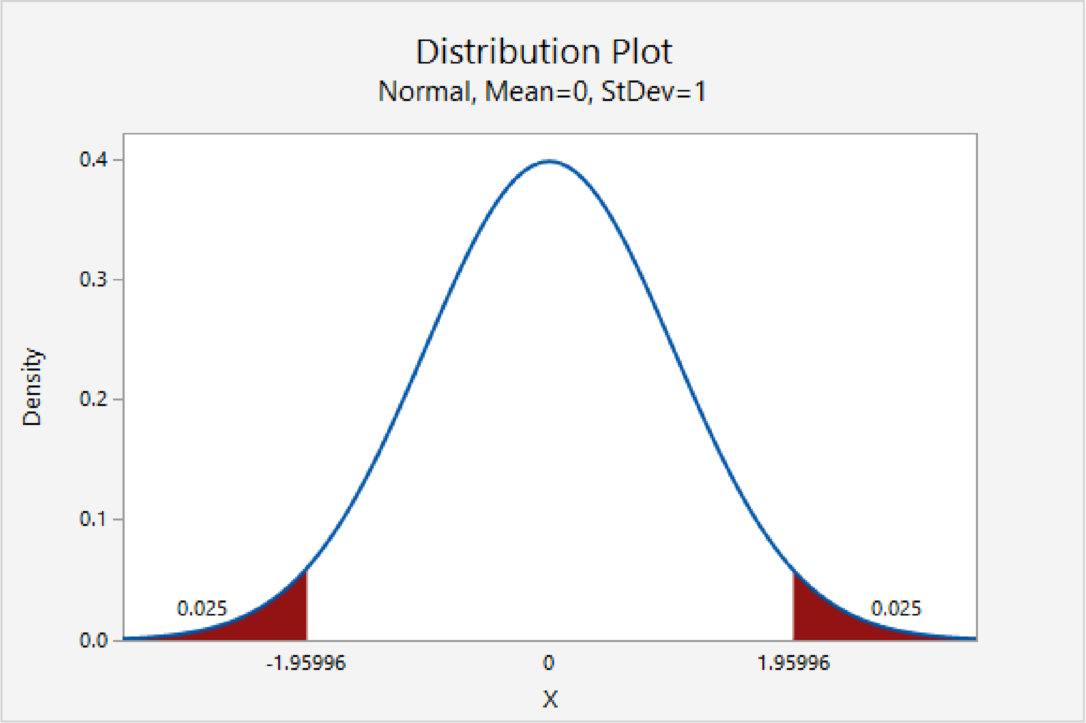
One Sample z-test

Note: the population is normally distributed, and the population variance, σ^2, is known.

Test Statistic: $z = \frac{\bar{x} - M}{\sigma / \sqrt{n}}$

M: a specified value to be tested

n: the size of the sample



https://online.stat.psu.edu/stat200/book/export/html/146

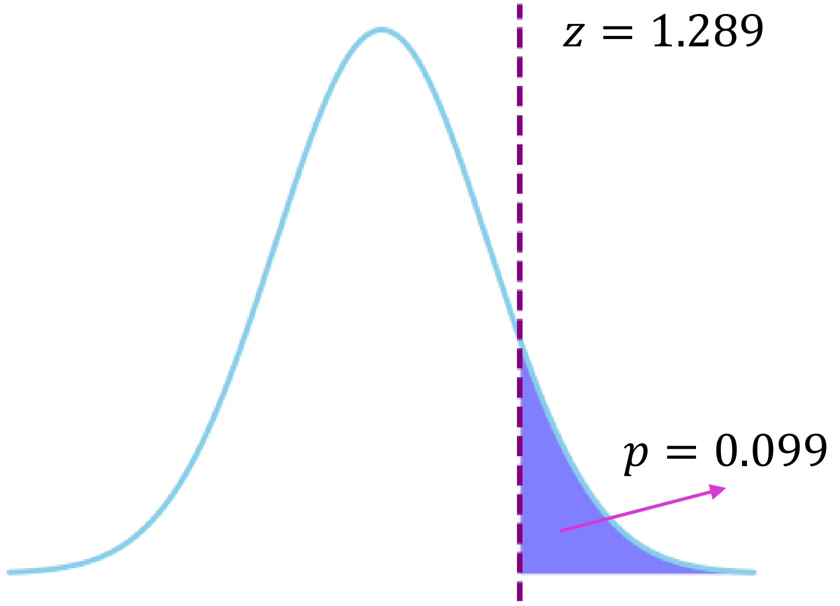
One Sample z-test

Example: A herd of 1,500 steer was fed a special high-protein grain for a month.

A random sample of 29 (n) were weighed and had gained an average of 6.7 ($\bar{x}$) pounds. If the standard deviation of weight gain for the entire herd is 7.1 (σ), test the hypothesis that the average weight gain per steer for the month was more than (or not equal to) 5 (M) pounds.

One Sample z-test

- Null hypothesis: H_0: $M = 5$
- Alternative hypothesis: H_a: $M > 5$
- $z = \frac{6.7-5}{\frac{7.1}{\sqrt{29}}} = \frac{1.7}{1.318} = 1.289$
- from scipy.stats import norm
- p = 1 - norm.cdf(1.289) = 0.099



One Sample t-test

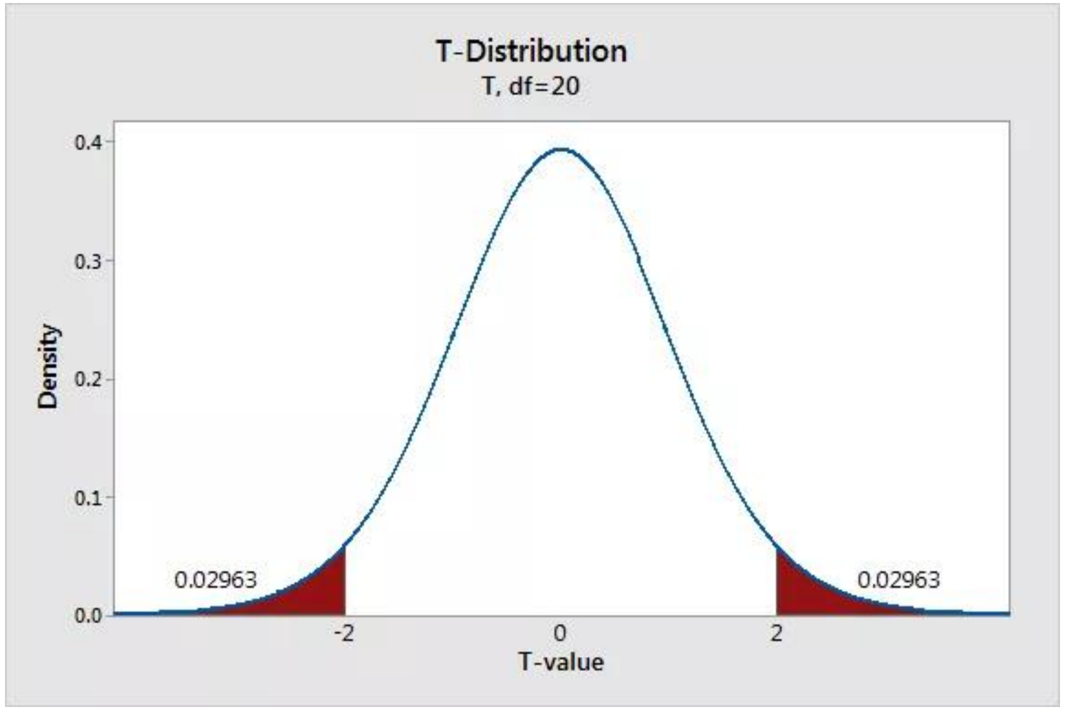
Note: the population is normally distributed, and the population variance, σ^2, is unknown.

Test Statistic: $t = \frac{\bar{x} - M}{s/\sqrt{n}}$

M: a specified value to be tested

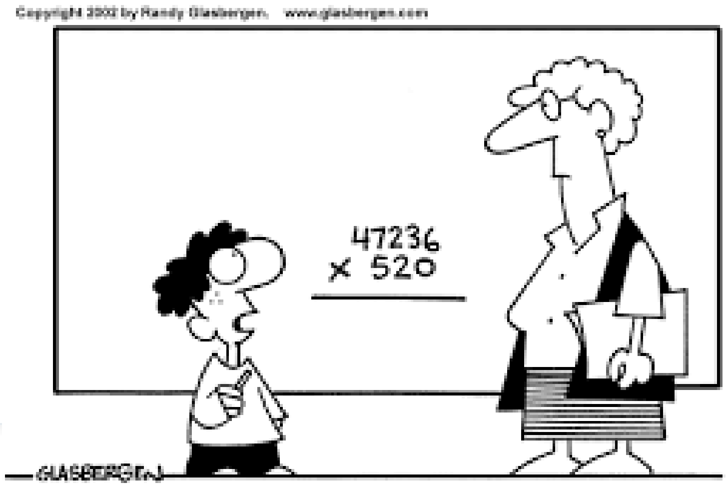
n: the size of the sample

s: the standard deviation of the sample



https://online.stat.psu.edu/stat200/book/export/html/146

One Sample t-test



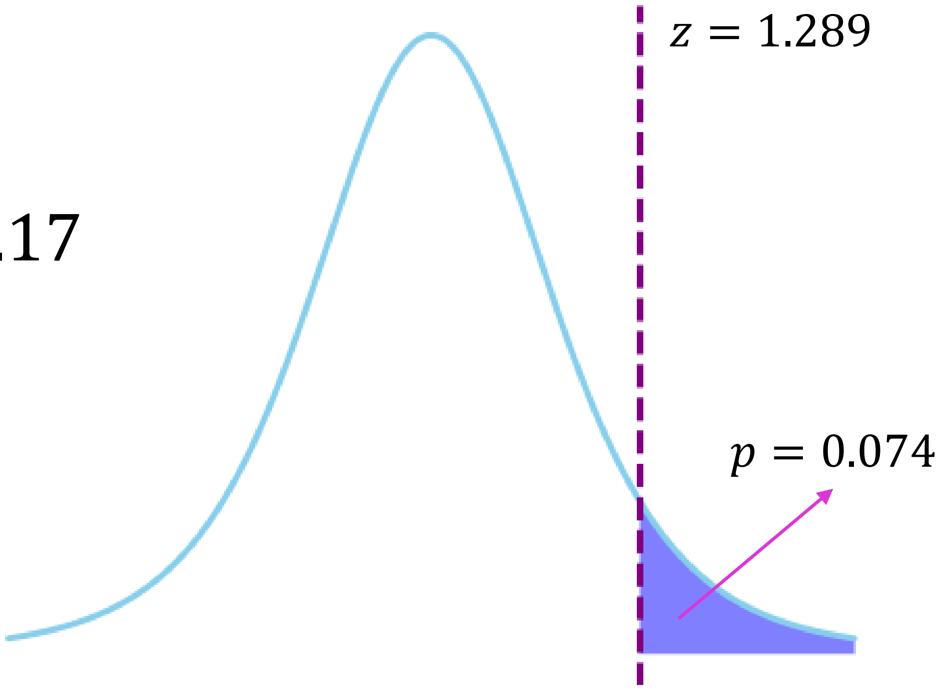
"Aren't there enough problems in the world already?"

Example: A professor wants to know if her introductory statistics class has a good grasp of basic math.

Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 (M) on the test. The 6 (n) students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence ($\alpha = 0.10$) that the mean score for the class on the test would be above 70?

One Sample t-test

- Null hypothesis: H_0: $M = 70$
- Alternative hypothesis: H_a: $M > 70$
- $\bar{x} = \frac{62+92+75+68+83+95}{6} = 79.17; s = 13.17$
- $z = \frac{79.17-70}{\frac{13.17}{\sqrt{6}}} = \frac{9.17}{5.38} = 1.71$
- from scipy.stats import t
- $p = 1$ - t.cdf(1.71) $= 0.074$



Two-Sample z-test for Comparing Two Means

Note: two populations is normally distributed, and the population variance, σ_1^2 and σ_2^2, are known.

Test Statistic: $z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Δ: the hypothesized difference between population means (0 if testing for equal means)

Two-Sample t-test for Comparing Two Means

Note: two populations is normally distributed, and the population variances, σ_1^2 and σ_2^2, are unknown.

Test Statistic: $t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Δ: the hypothesized difference between population means (0 if testing for equal means).

The degrees of freedom parameter for looking up the t-value is the smaller of $n_1 - 1$ and $n_2 - 1$.

https://www.cliffsnotes.com/study-guides/statistics

Paired Difference t-test

Note: a set of paired observation from a normal distribution. Two populations are normally distributed, and the population variances, σ_1^2 and σ_2^2, are unknown.

Test Statistic: $t = \frac{\bar{x}-\Delta}{\frac{s}{\sqrt{n}}}$

Δ: the hypothesized difference (0 if testing for equal means)

The degrees of freedom parameter for looking up the t-value is the smaller of $n-1$.

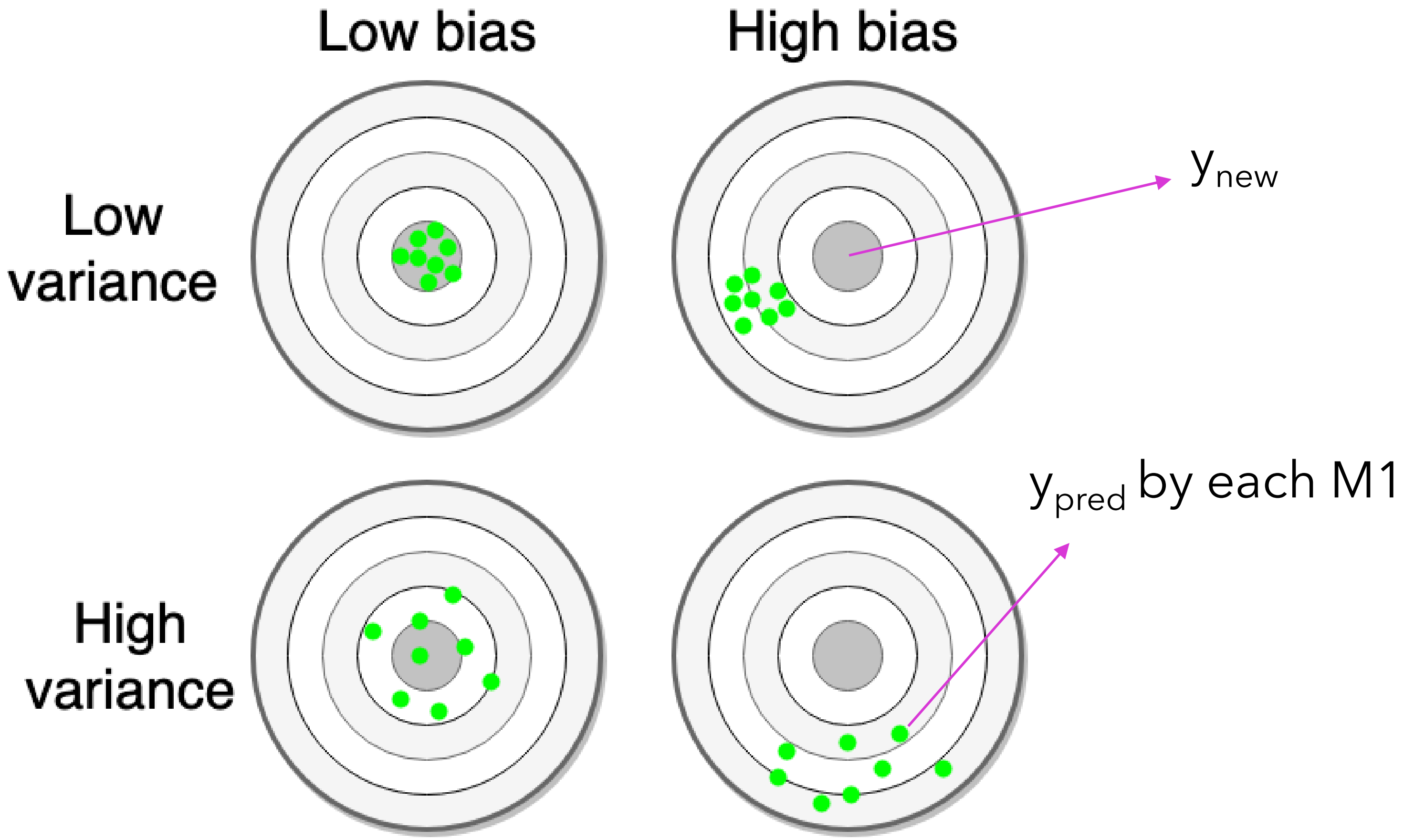
https://www.cliffsnotes.com/study-guides/statistics

Bias and Variance

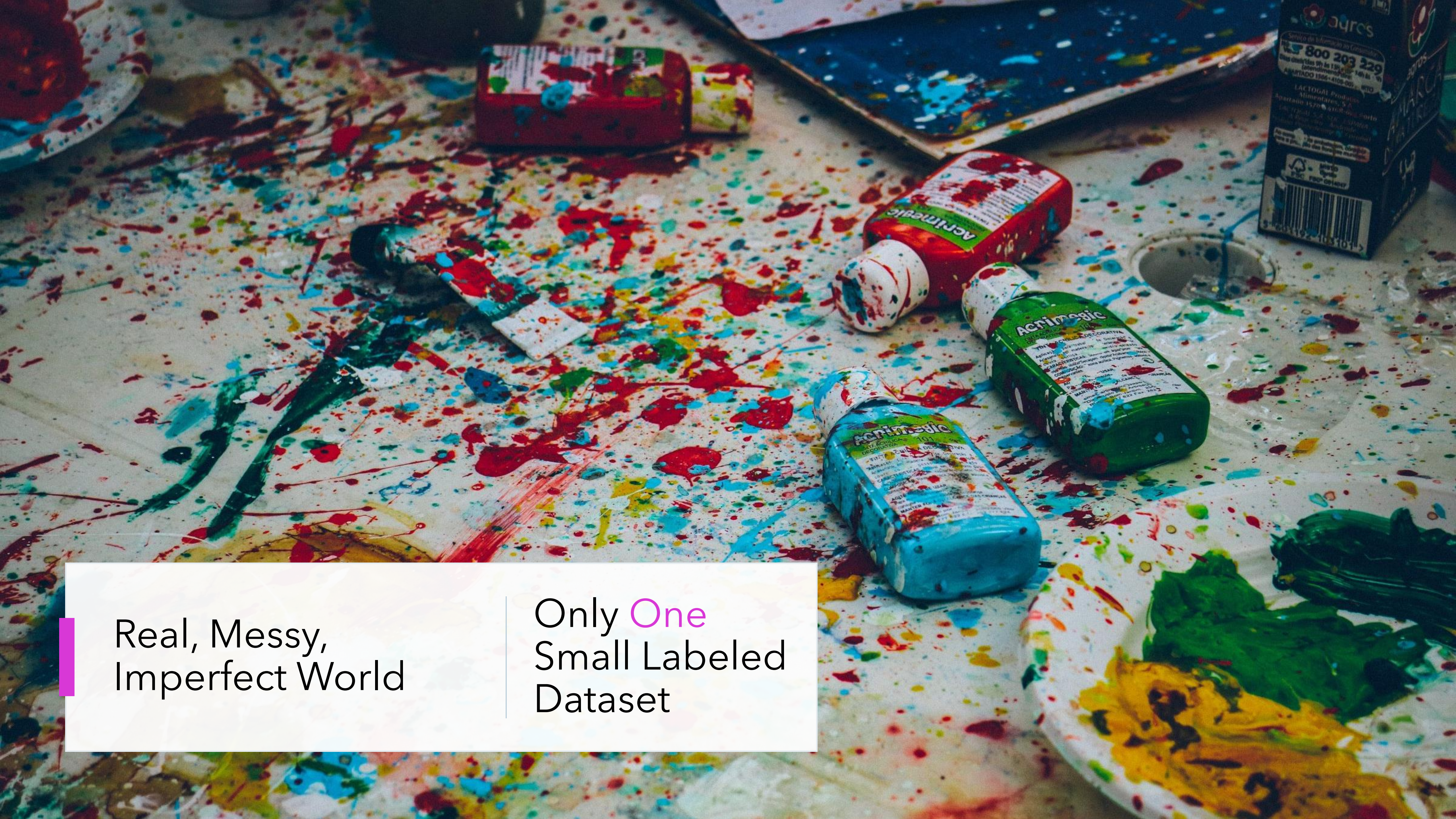
After fitting N(=1000) M_1 models,

given a new data point (x_{new}, y_{new}) that never shows up in training dataset,

predict y_{pred} by applying each M_1 on x_{new}



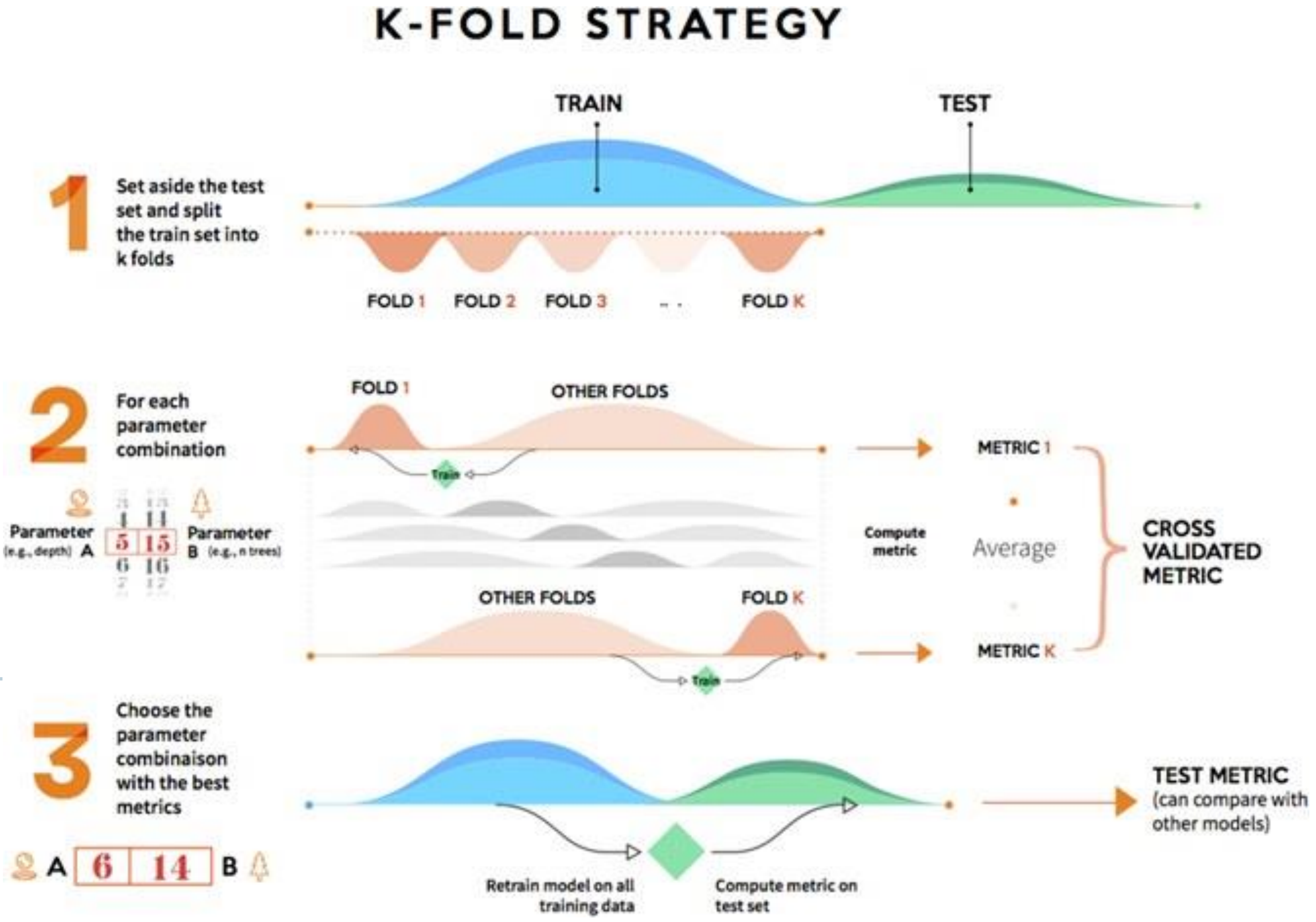
https://www.machinelearningtutorial.net/2017/01/26/the-bias-variance-tradeoff/



Real, Messy, Imperfect World

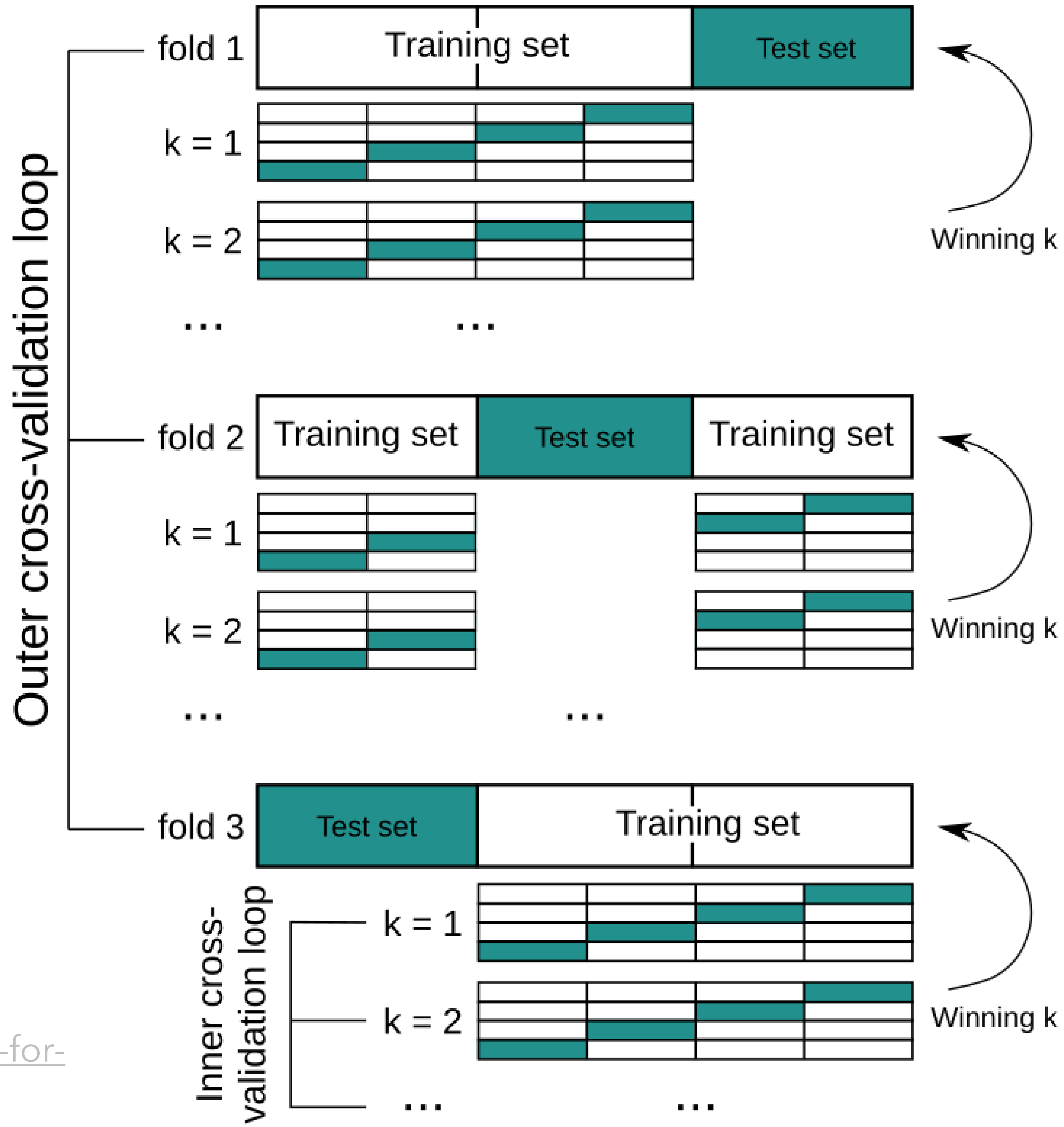
Only One Small Labeled Dataset

Cross-Validation



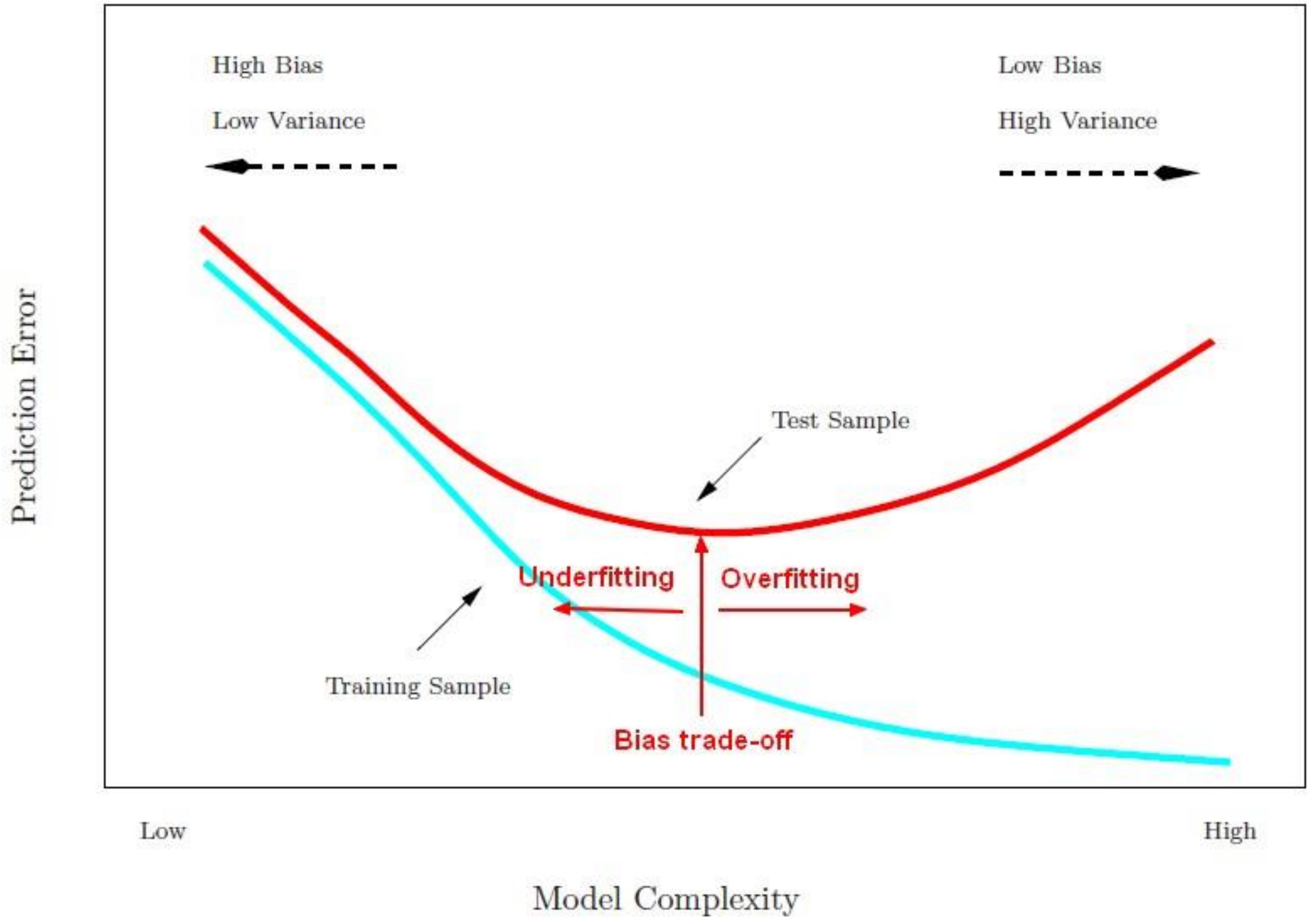
https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html

Cross-Validation



https://livebook.manning.com/book/machine-learning-for-mortals-mere-and-otherwise/chapter-3/v-4/287

Model Complexity



https://gerardnico.com/data_mining/overfitting

Reading Materials

- Ian Goodfellow, Yoshua Bengio and Aaron Courville, ***Deep Learning***. MIT Press, 2016. Deeplearningbook.org.
  Part I: Applied Math and Machine Learning Basics

- Bishop, Christopher M. ***Pattern Recognition and Machine Learning***. Springer, 2006.
  Chapter 1 & 2