Sentiment analysis
Evaluation

Model Tuning

Find the best running settings of the mode

- #layers
- Activation functions
- Probs. assumption

Model Tuning

Find the best running settings of the mode

- Checking the performance of model on Train and Test
- For all different possibilities

Blind grid search! Brute-force

Model Tuning

Find the best running settings of the mode

- Learn the performance of model on Train and Test
- For all different possibilities

Guided grid search!

Model Tuning

Find the best running settings of the mode

- Learn the performance of model on Train and Test
- For all different possibilities

Guided grid search!

https://www.kaggle.com/c/petfinder-adoption-prediction/discussion/125436

Mongrel Jedi

PetFinder.my Contest: 1st Place Winner Disqualified

307

Posted in petfinder-adoption-prediction a year ago

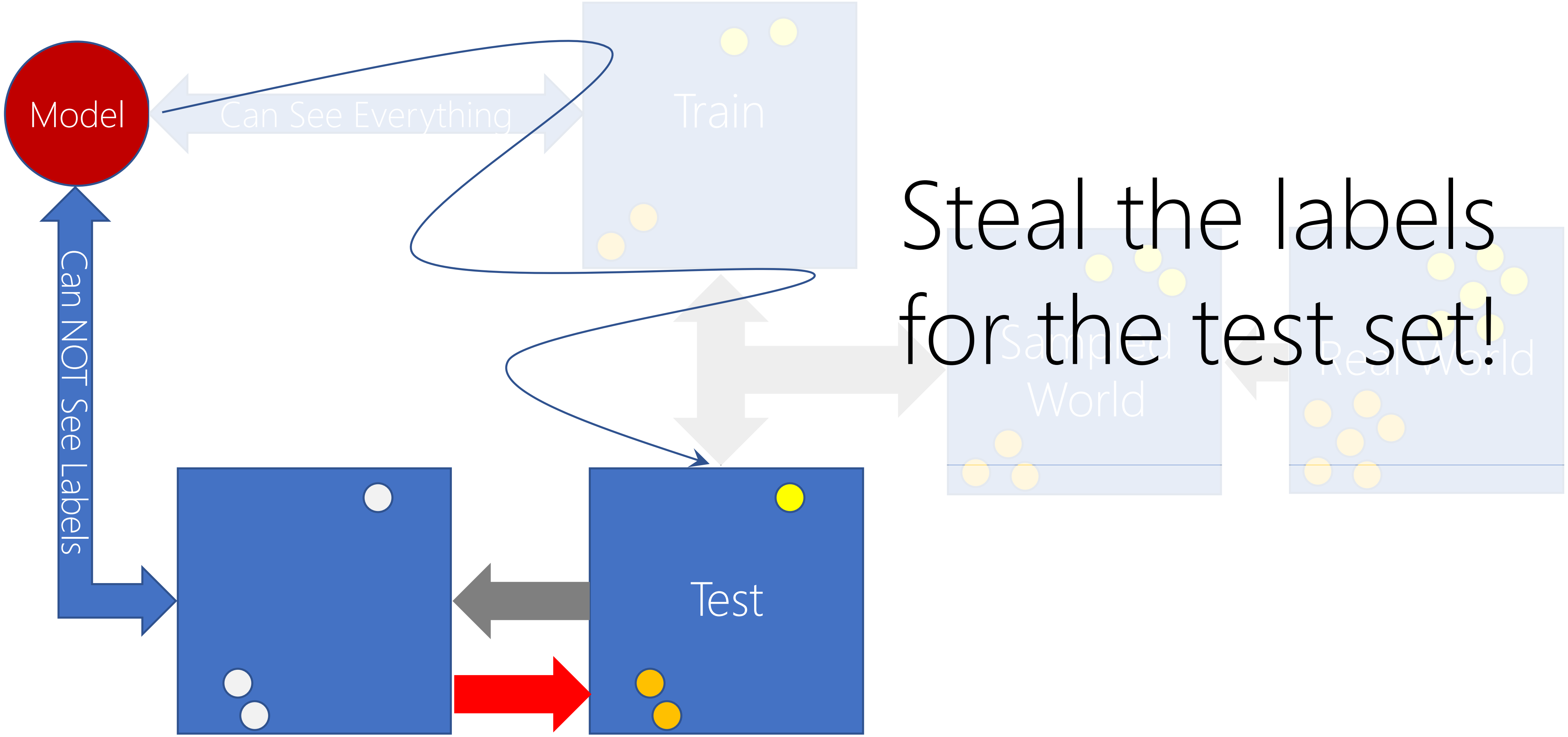
Dear Participants,

We would like to announce that the 1st Place Team, Bestpetting has been disqualified from the contest for cheating. The Kaggle Grandmaster cheater has also been permanently banned on this platform as the evidence points towards him being the key party behind this fraudulent activity.

Here is what the Bestpetting team did in the PetFinder.my contest:

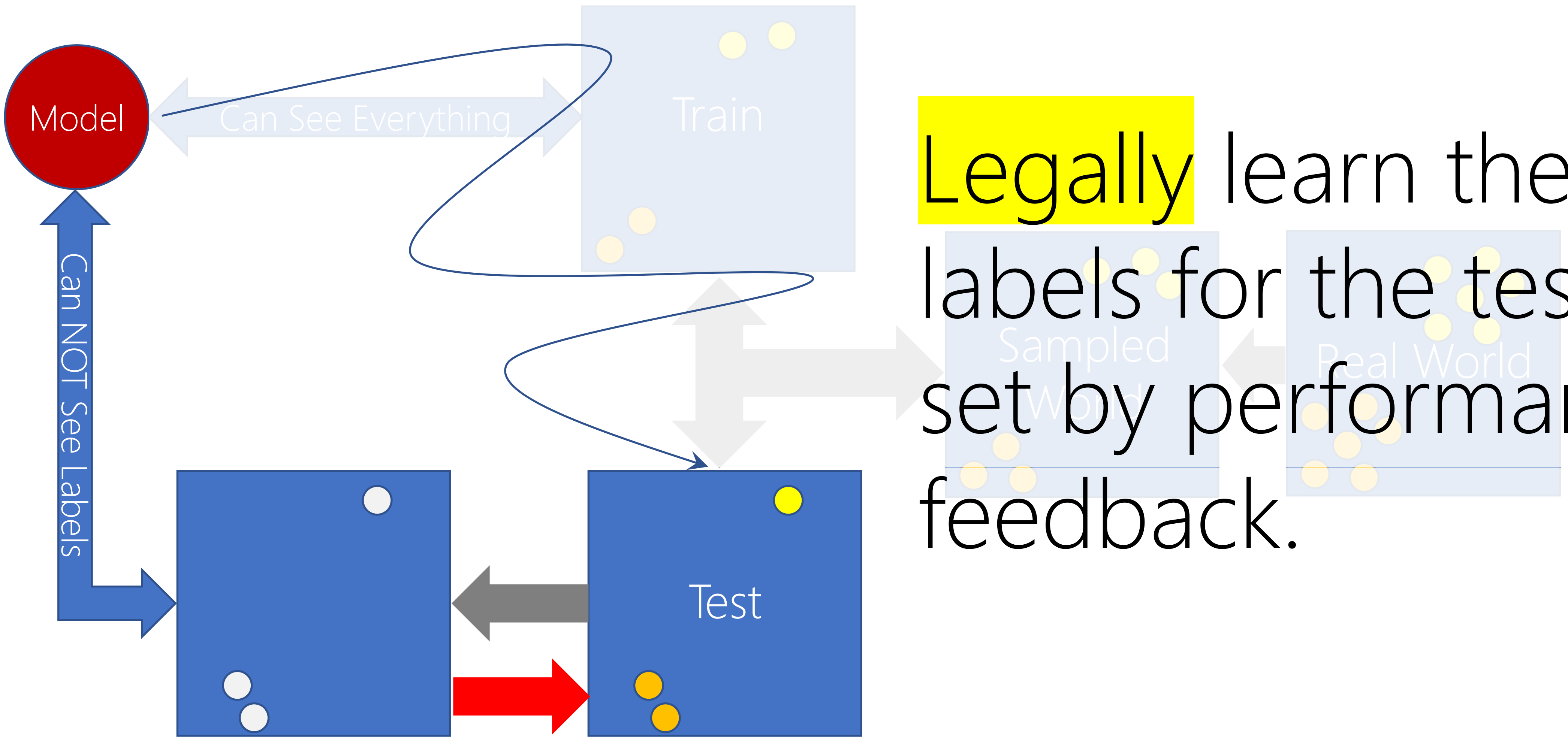
- They fraudulently obtained adoption speed answers for the private test data (possibly by scraping our website)

Labeled Data = {Train} U {Test}



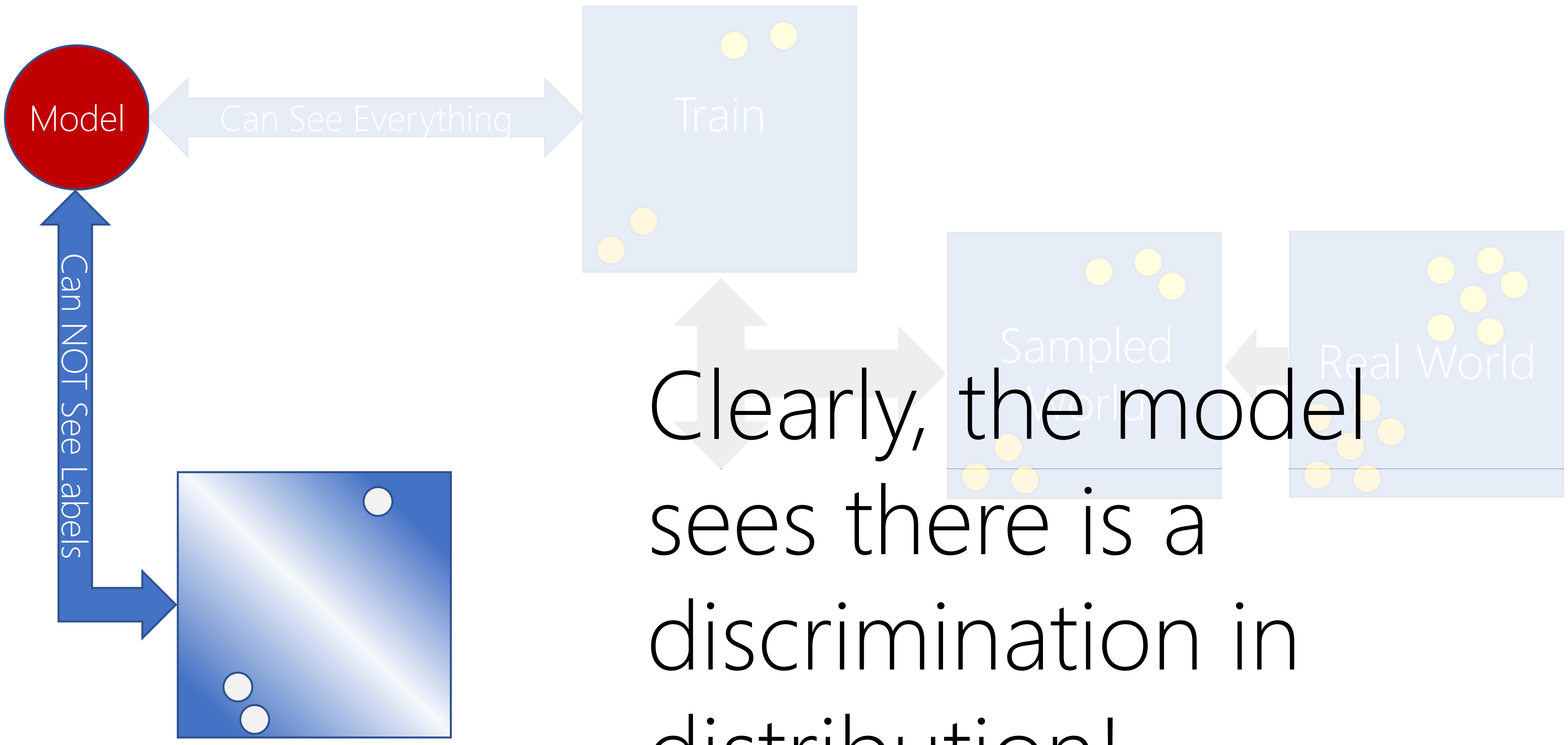
Steal the labels for the test set!

Labeled Data = {Train} U {Test}



Legally learn the labels for the test set by performance feedback.

Labeled Data = {Train} U {Test}

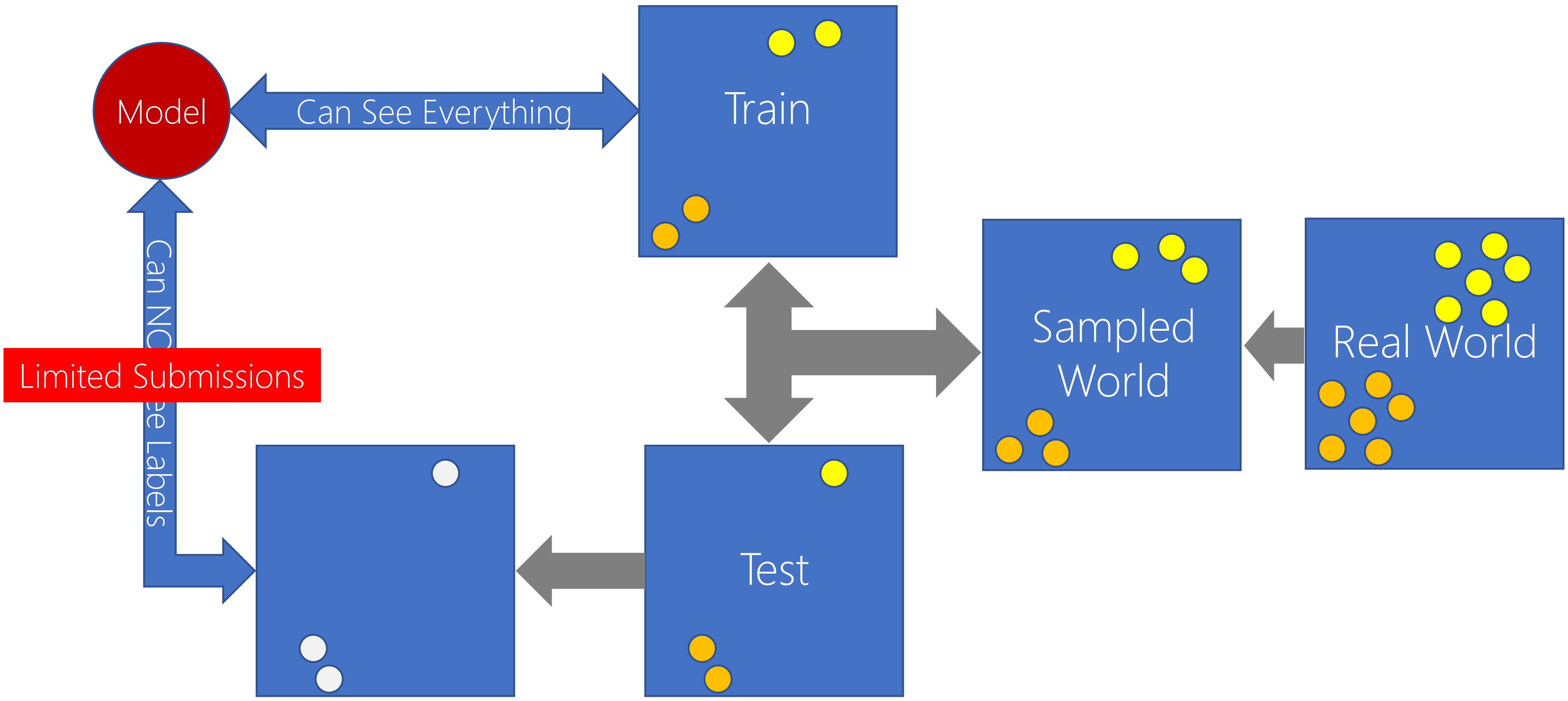


Clearly, the model sees there is a discrimination in distribution!

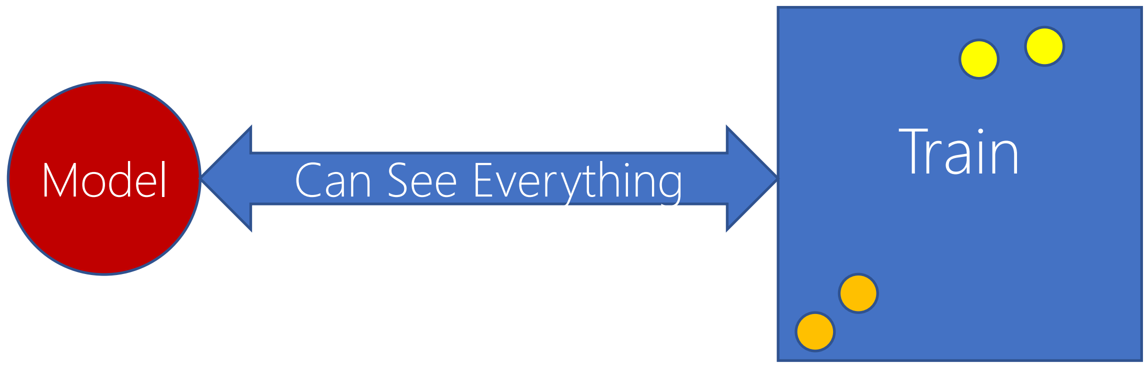
Labeled Data = {{Train} U {Valid}} U {Test}

The model intentionally ignores parts of his available knowledge and challenges itself to uncover those parts!

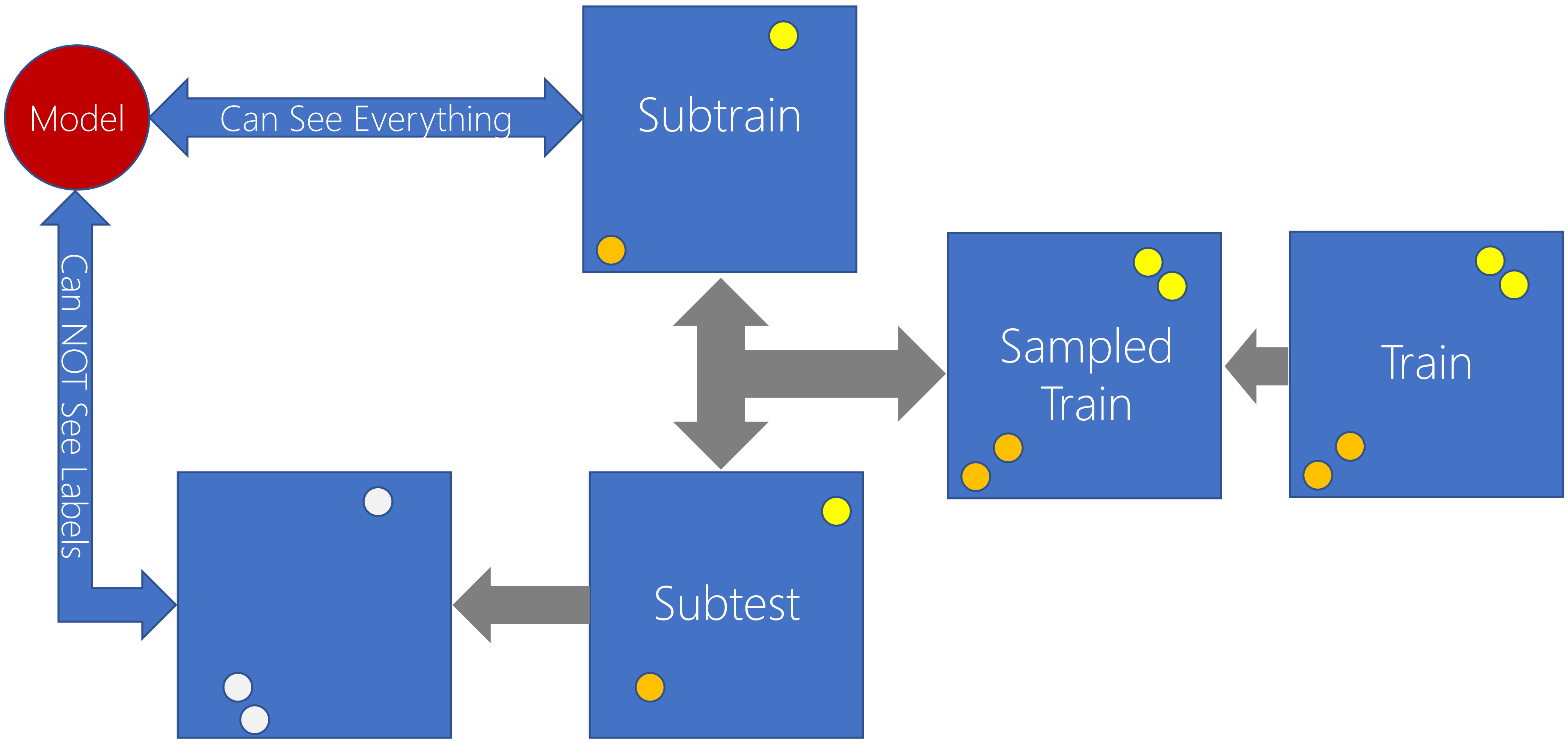
Labeled Data = {Train} U {Test}



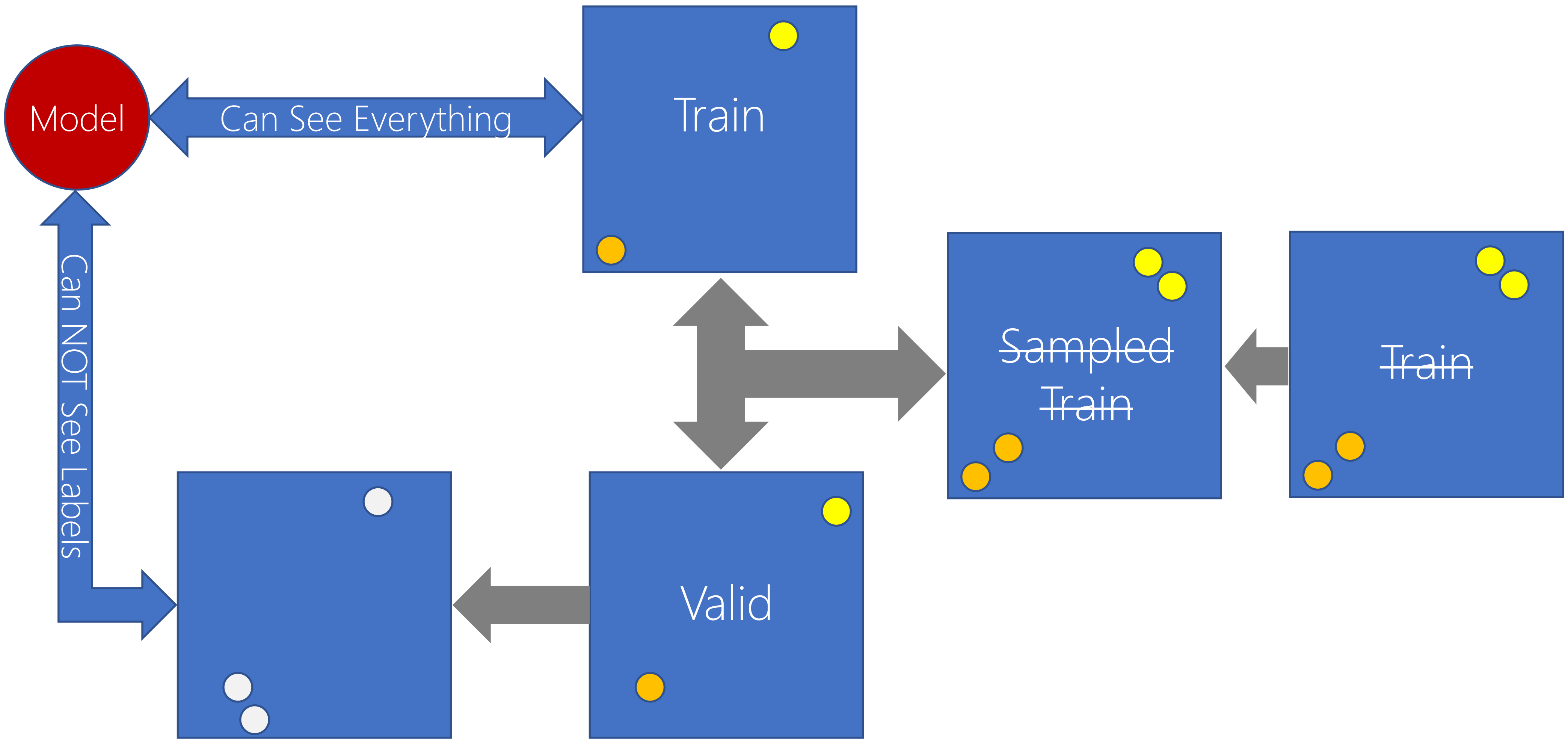
Labeled Data = {Train}



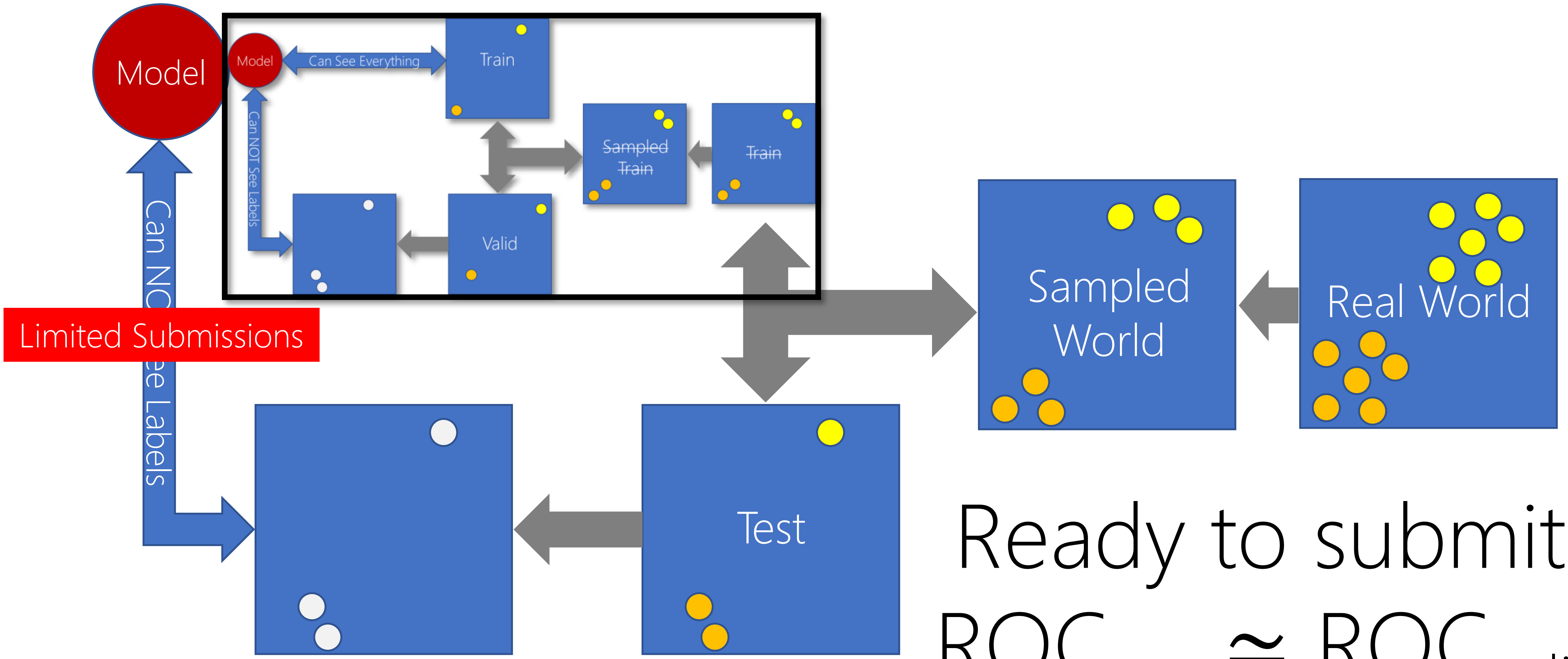
Labeled Data = {{Subtrain} U {Subtest}



Labeled Data = {{Train} U {Valid}}



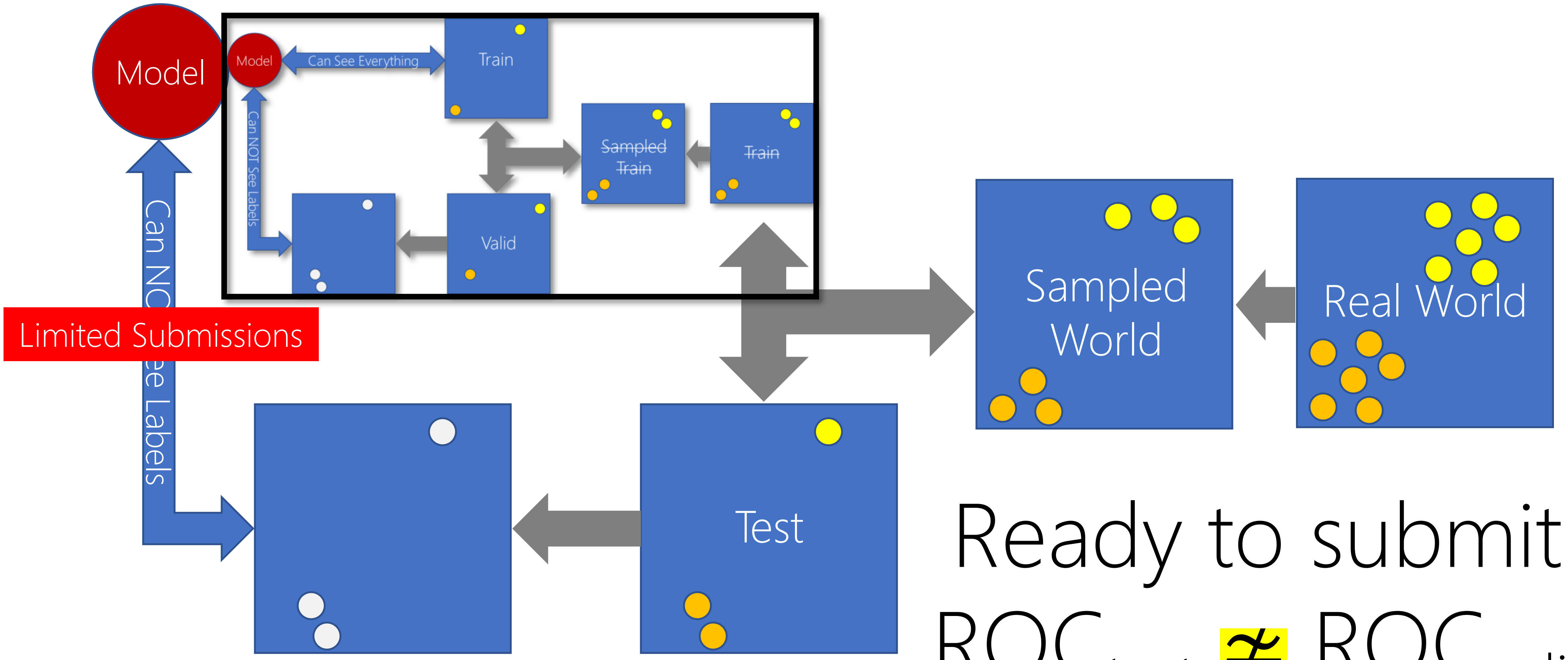
Labeled Data = {Train} U {Test}



Ready to submit.

$ROC_{test} \simeq ROC_{valid}$

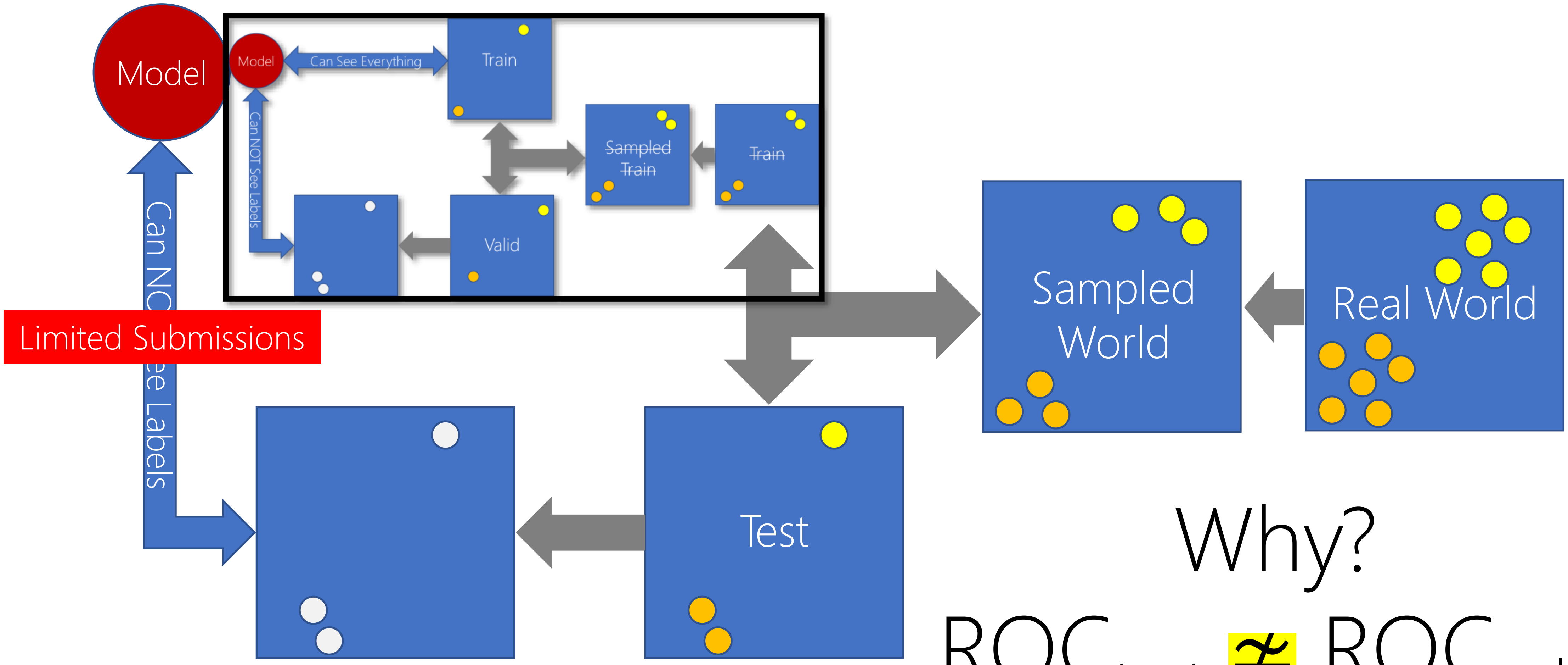
Labeled Data = {Train} U {Test}



Ready to submit.

$ROC_{test} \neq ROC_{valid}$

Labeled Data = {Train} U {Test}

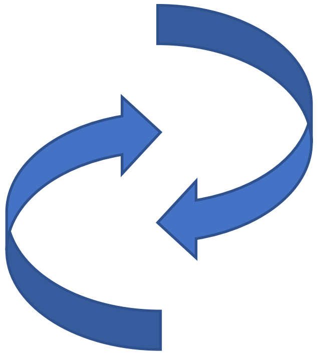


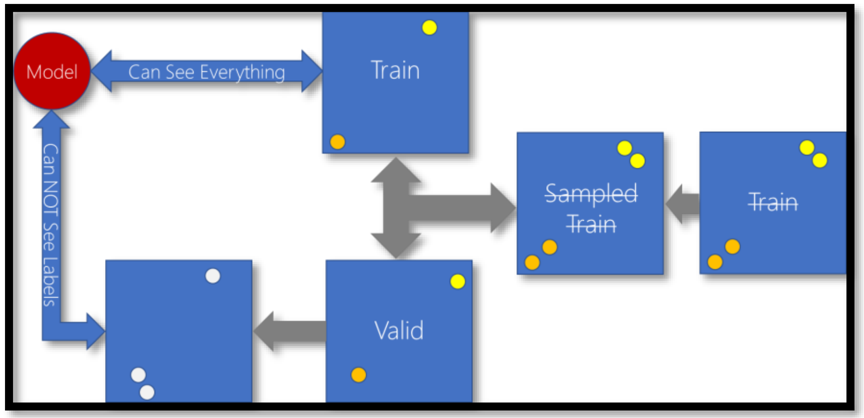
Why?

$ROC_{test} \neq ROC_{valid}$

Cross-Validation

1 practice vs. Multiple practice





Cross-Validation

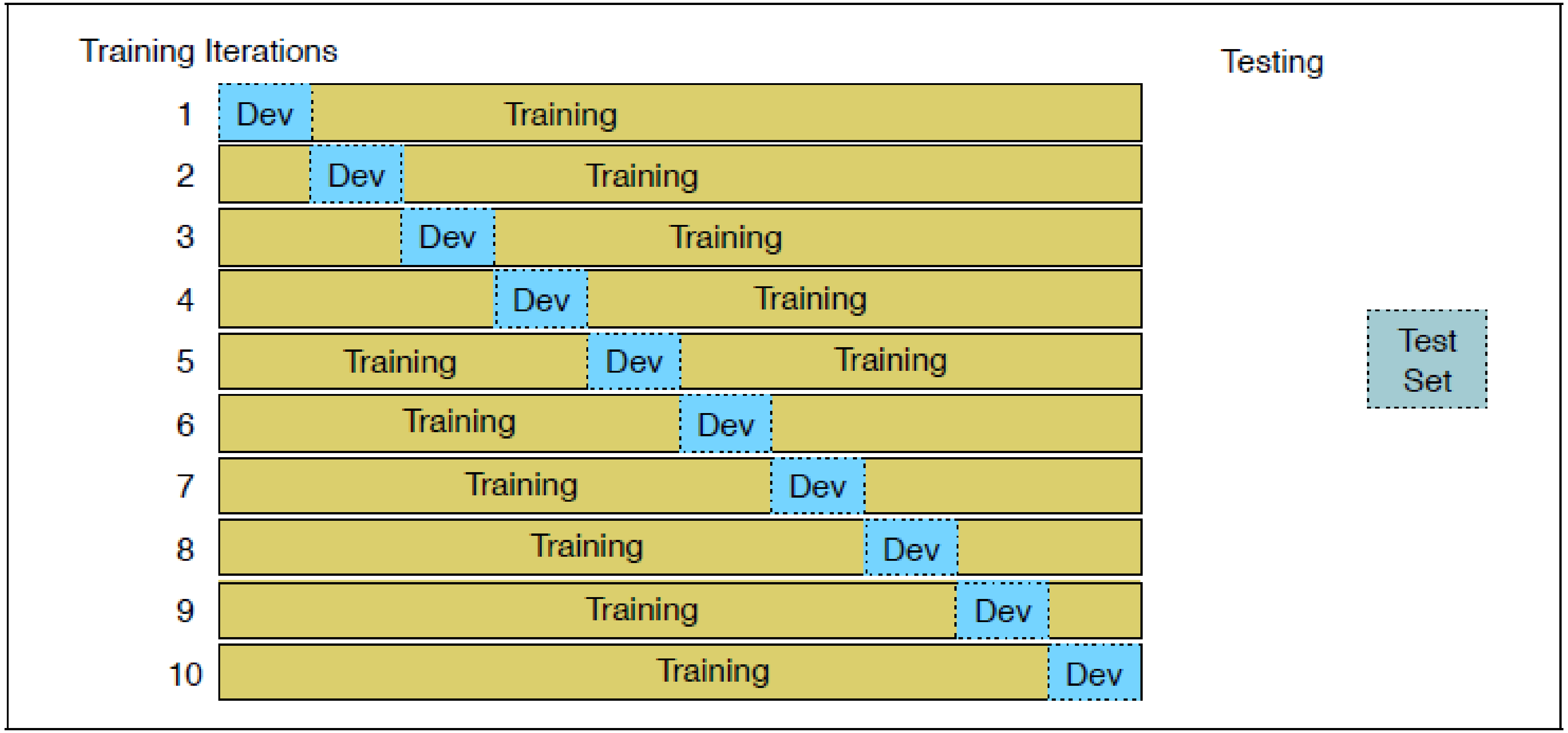
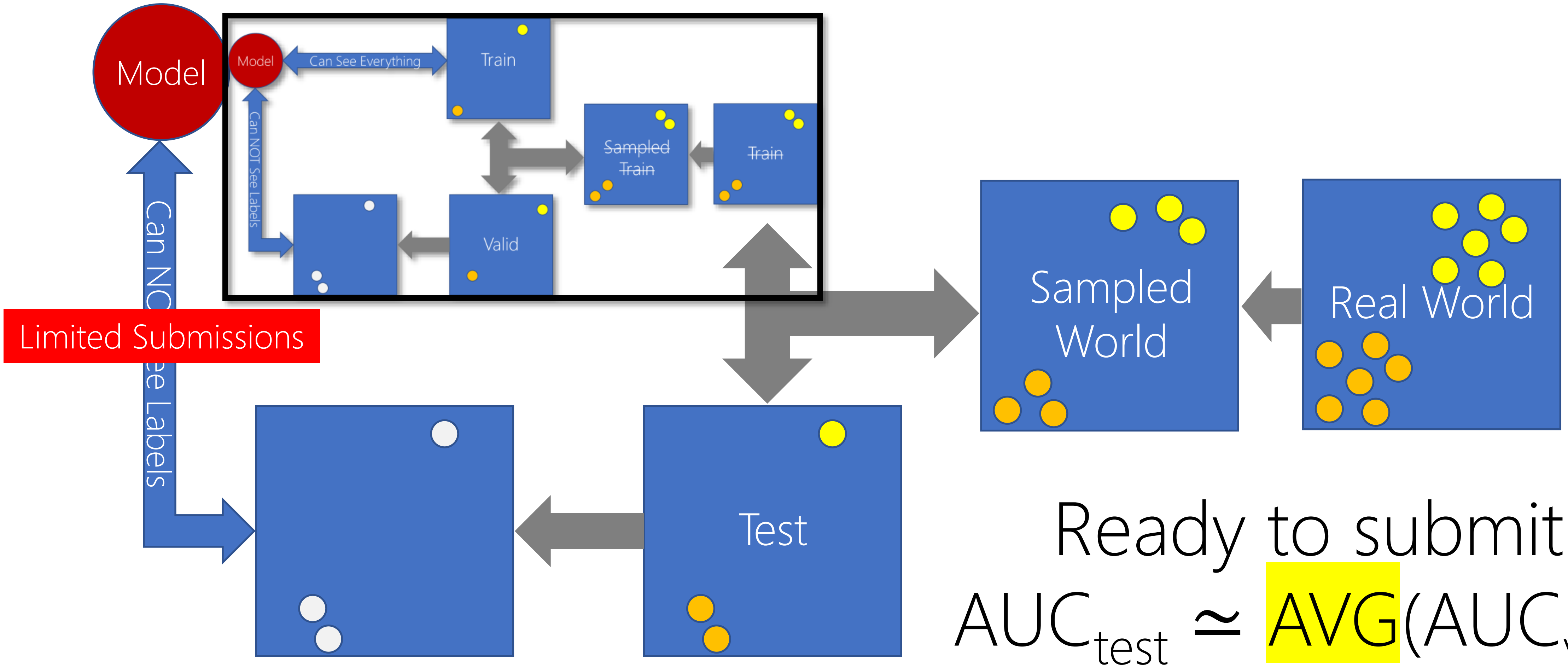


Figure 4.7 10-fold cross-validation

Labeled Data = {Train} U {Test}



Ready to submit.

$AUC_{test} \simeq$ AVG(AUC_{valid})

Multiclass Evaluation

Multiclass Evaluation

		Truth		
		Happy	Sad	Neutral
Model	Happy			
	Sad			
	Neutral			

		Truth for Happy	
		Happy	Not Happy
Model	Happy		
	Not Happy		

		Truth for Sad	
		Sad	Not Sad
Model	Sad		
	Not Sad		

		Truth for Happy	
		Neutral	Not Neutral
Model	Neutral		
	Not Neutral		

Multiclass Evaluation

		Truth		
		Happy	Sad	Neutral
Model	Happy	8	10	1
	Sad	5	60	50
	Neutral	3	30	200

		Truth for Happy	
		Happy	Not Happy
Model	Happy	8	11
	Not Happy	8	340

		Truth for Sad	
		Sad	Not Sad
Model	Sad	60	55
	Not Sad	40	212

		Truth for Happy	
		Neutral	Not Neutral
Model	Neutral	200	33
	Not Neutral	51	83

Multiclass Evaluation

		Truth		
		Happy	Sad	Neutral
Model	Happy	8	10	1
	Sad	5	60	50
	Neutral	3	30	200

		Truth for Happy	
		Happy	Not Happy
Model	Happy	8	11
	Not Happy	8	340

$$P_{happy} = \frac{8}{8+11} = 0.42$$

		Truth for Sad	
		Sad	Not Sad
Model	Sad	60	55
	Not Sad	40	212

$$P_{sad} = \frac{60}{60+55} = 0.52$$

		Truth for Happy	
		Neutral	Not Neutral
Model	Neutral	200	33
	Not Neutral	51	83

$$P_{neutral} = \frac{200}{200+33} = 0.85$$

Multiclass Evaluation

		Truth		
		Happy	Sad	Neutral
Model	Happy	8	10	1
	Sad	5	60	50
	Neutral	3	30	200

$$P_{happy} = \frac{8}{8+10+1} = 0.42$$

$$P_{sad} = \frac{60}{5+60+50} = 0.52$$

$$P_{neutral} = \frac{200}{3+30+200} = 0.85$$

$$R_{happy} = \frac{8}{8+5+3} = ? \qquad R_{sad} = \frac{60}{10+60+30} = ? \qquad R_{neutral} = \frac{200}{1+50+200} = ?$$

Multiclass Evaluation: Macro-Avg

$$Macroavg = \frac{1}{K}\sum_{i=1}^{K} Metric_K$$

$$Macroavg = \frac{1}{3}[P_{happy} + P_{sad} + P_{neutral}]$$

		Truth		
		Happy	Sad	Neutral
Model	Happy	8	10	1
	Sad	5	60	50
	Neutral	3	30	200

		Truth for Happy	
		Happy	Not Happy
Model	Happy	8	11
	Not Happy	8	340

$P_{happy} = \frac{8}{8+11} = 0.42$

		Truth for Sad	
		Sad	Not Sad
Model	Sad	60	55
	Not Sad	40	212

$P_{sad} = \frac{60}{60+55} = 0.52$

		Truth for Happy	
		Neutral	Not Neutral
Model	Neutral	200	33
	Not Neutral	51	83

$P_{neutral} = \frac{200}{200+33} = 0.85$

Multiclass Evaluation: Micro-Avg

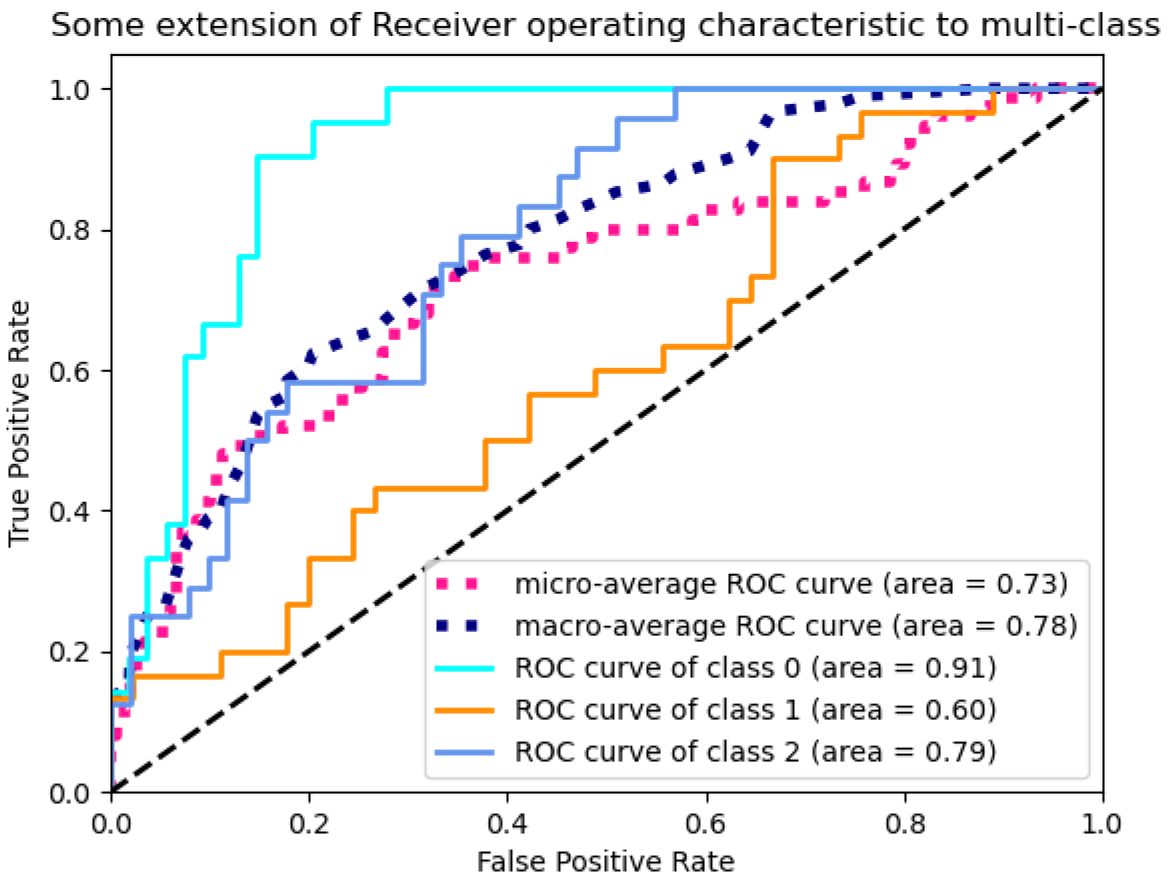
		Truth		
		Happy	Sad	Neutral
Model	Happy	8	10	1
	Sad	5	60	50
	Neutral	3	30	200

Pool

		True	Not True
Model	True	268	99
	Not True	99	?

P_{true} =?

R_{true} =?



https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

Macro vs. Micro Averaging

Macro vs. Micro Averaging

- Microavg is dominated by the more frequent class since the counts are pooled.
- Macroavg better reflects the statistics of the smaller classes, and so is more appropriate when performance on all the classes is equally important.

Transparent (Interpretable) Model

Qualitative Analysis

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

How (why) the model make “this prediction”?

"*** ** ****" → Happy

"*** ** ****" → Sad

"*** ** ****" → Neutral

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

How (why) the model make “this prediction”?
Manually inspect the test or valid sets.
Very time consuming!

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

How (why) the model make “this prediction”?

Look at the model’s parameters.

On what part of the input the model pay attention?

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

$$\text{Sigmoid}([x_{i:|V|}\ 1][w_{1:|V|}]) > 0.5 \rightarrow \text{Positive}$$

When we look at all w_i → we see only w_{34}= 0.8
Others are either 0 or negative.

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

$$\text{Sigmoid}([x_{i:|V|}\ 1][w_{1:|V|}]) > 0.5 \rightarrow \text{Positive}$$

Only w_{34} brings an input to the positive class.
Only $x_{34}w_{34}$ brings an input to the positive class. What is x_{34}?
The model learns to keep $x_{34}w_{34}$ to correctly classify positive instances.

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

[I am happy to be a student]

[0, 0, .., 1, 0, 0, …, 1, 0,0, 1, …, 1, 0, 0, …, 1, 0,0, 1, …]

x_{34} = [happy]

Transparent (Interpretable) Model

Qualitative (Descriptive) Analysis

Whenever “happy” is in the input, the model classify the input as positive!

Bagging Model for Product Title Quality with Noise

CIKM AnalyticCup 2017

Tam T.Nguyen
Ryerson University
nthanhtam@gmail.com

Hossein Fani
University of New Brunswick
hossein.fani@gmail.com

Ebrahim Bagheri
Ryerson University
ebrahim.bageri@gmail.com

Gilbero Titericz
Airbnb, Inc.
giba1978@gmail.com

P R O B L E M

is_clear

"hot sexy red clutch rug sack travel backpack unisex cheap with free gift"

LAZADA
Effortless Shopping

Is_concise

"Hot Sexy Tom Clovers Womens Mens Classy Look Cool Simple Style Casual Canvas Crossbody Messenger Bag Handbag Fashion Bag Tote Handbag Gray"

group	features	name
statistics	#char(length)	
	#term	xg_feat
	price	price_feat
information	color	color
	brand	brand
	category entropy	entropy_feat
n-gram term	(1,3)-gram	bow.3grams
n-gram char	(1,6)-gram	boc.6grams
	#upper char	
	#special char	
	html escape	
	#invisible char	
sparse feature	category one-hot-encoding	sp_feat
leave-one-out encode		
embedding	word2vec[2]	
part-of-speech	#adjective	
	#verb	
	#noun	
	#number	
multilingual characters	#non-english char	
	#chinese char	char_set_feat

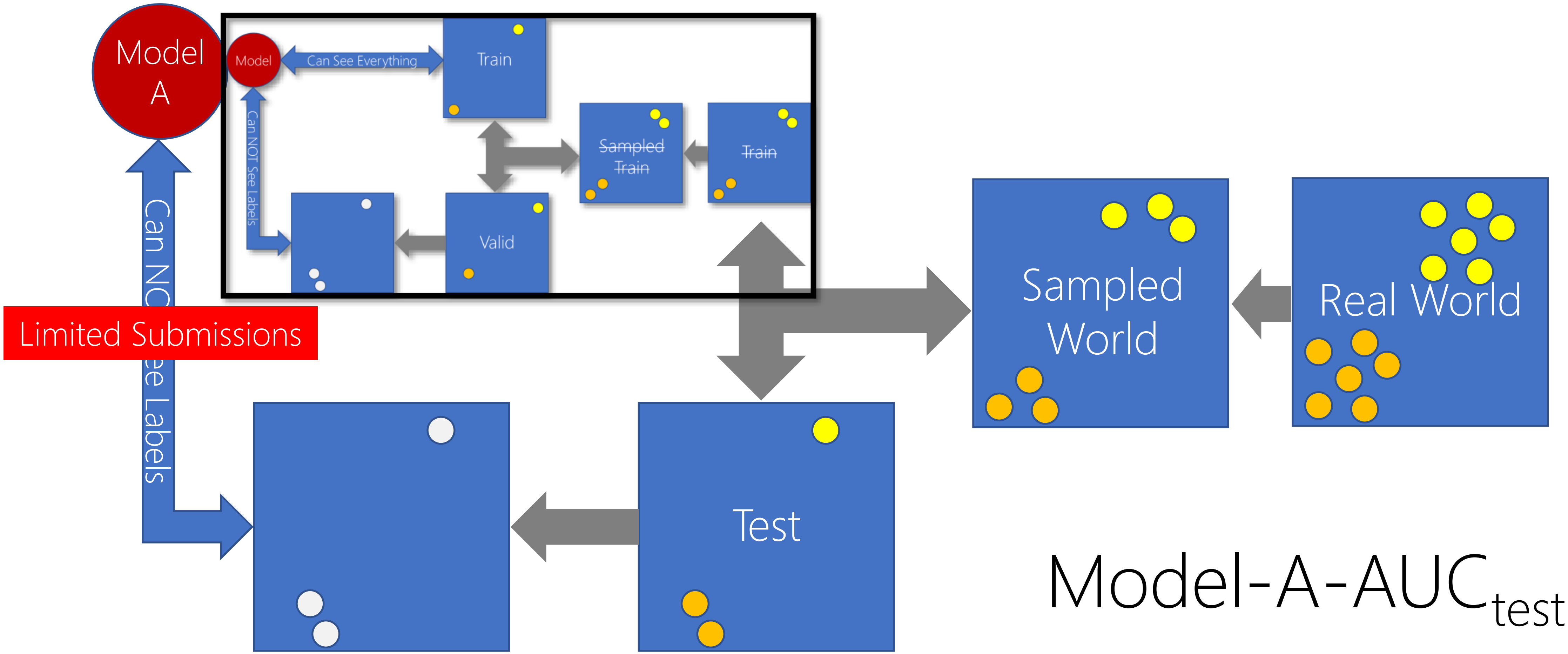
Table 3: Most important features based on linear SVM.

label	name	coef.	label	name	coef.
clarity	t	0.442989	conciseness	my	1.087967
	sexy	0.398171		ph	0.968527
	exy	0.398026		c	0.957356
	sex	0.384535		ocal	0.931618
	urse	0.368735		local	0.925727
	purse	0.341463		r	0.920576
	purs	0.341463		loca	0.912073
	rse	0.338007		sg	0.909163
	xy	0.334105		cal	0.888805
	purse	0.326108		loc	0.882074

Statistical Significance Testing

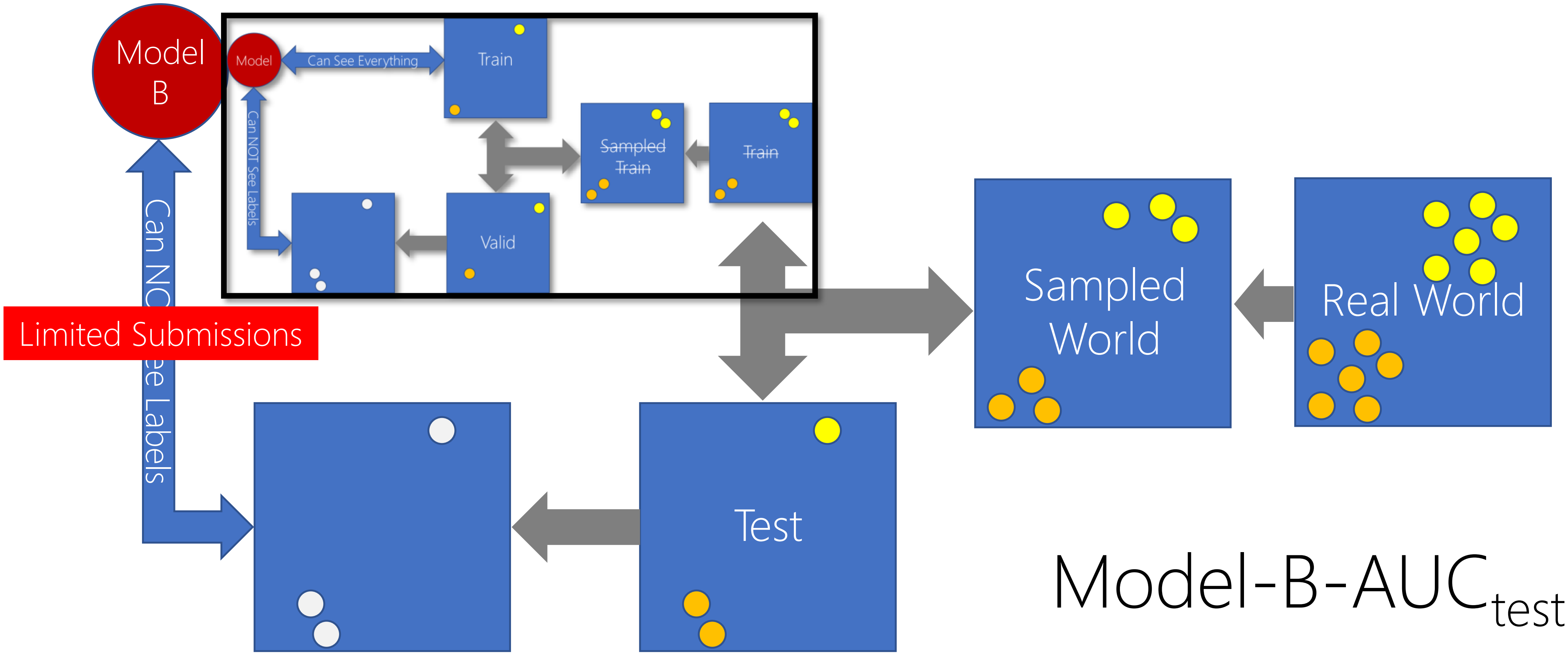
Model Comparison

Statistical Significance Testing



$\text{Model-A-AUC}_{\text{test}}$

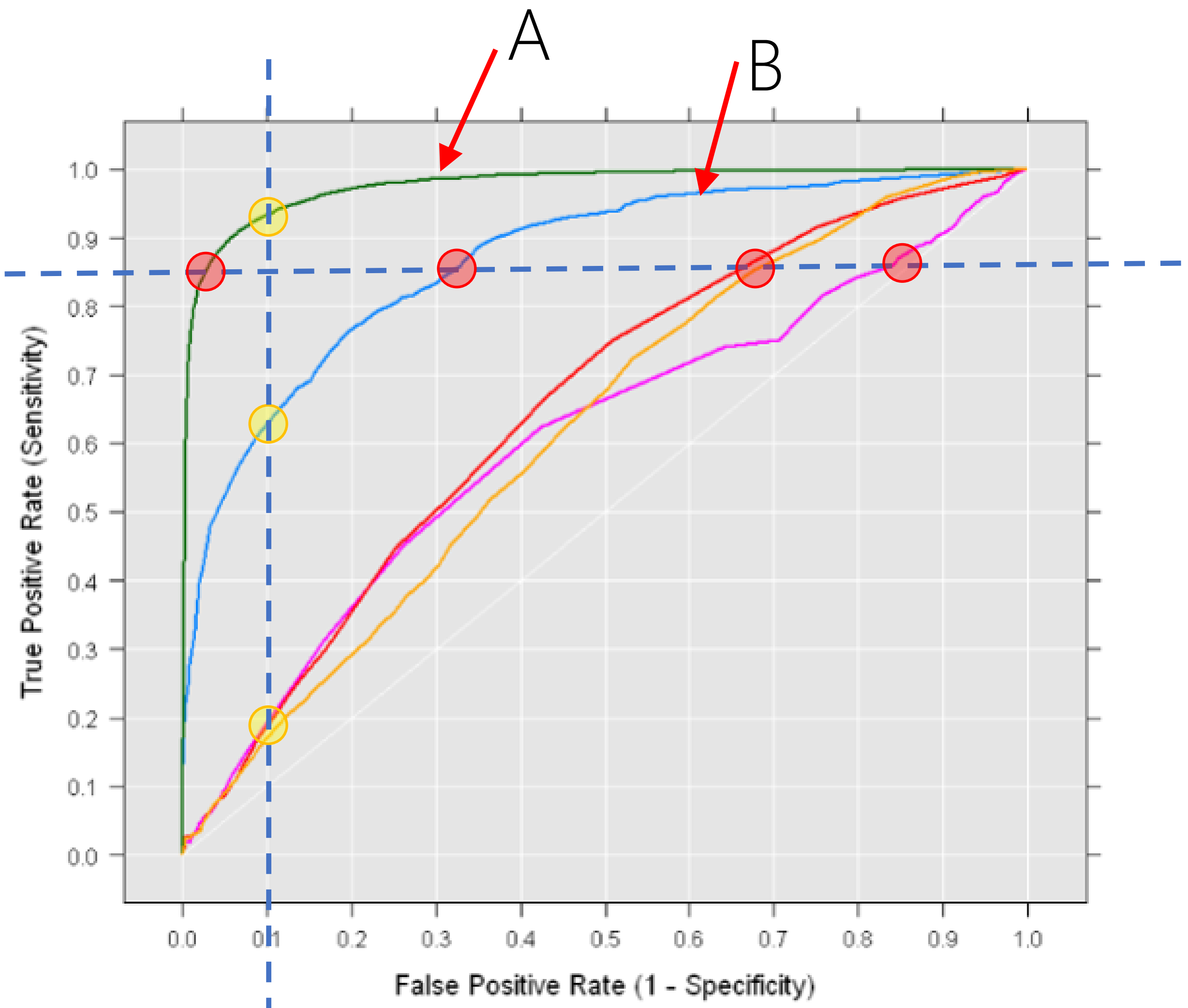
Statistical Significance Testing



$$\text{Model-B-AUC}_{test}$$

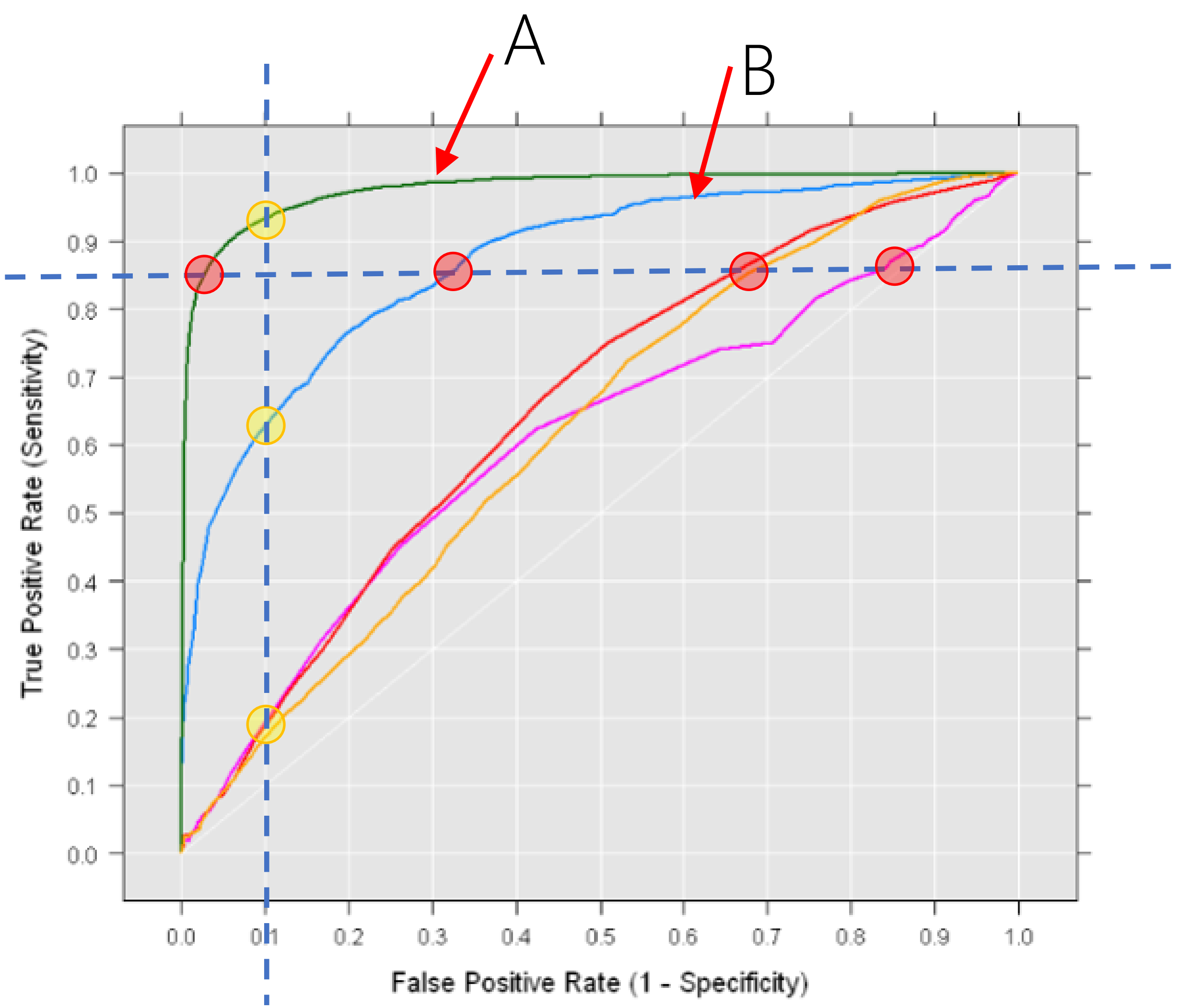
Statistical Significance Testing

Which one? A or B? A-AUC_{test} > B-AUC_{test} → A



Statistical Significance Testing

Which one? A or B? A-AUC_{test} > B-AUC_{test} → A



I am pessimist! Although, I trust you, but still there is a possibility that A is not a better model. Why? Where?

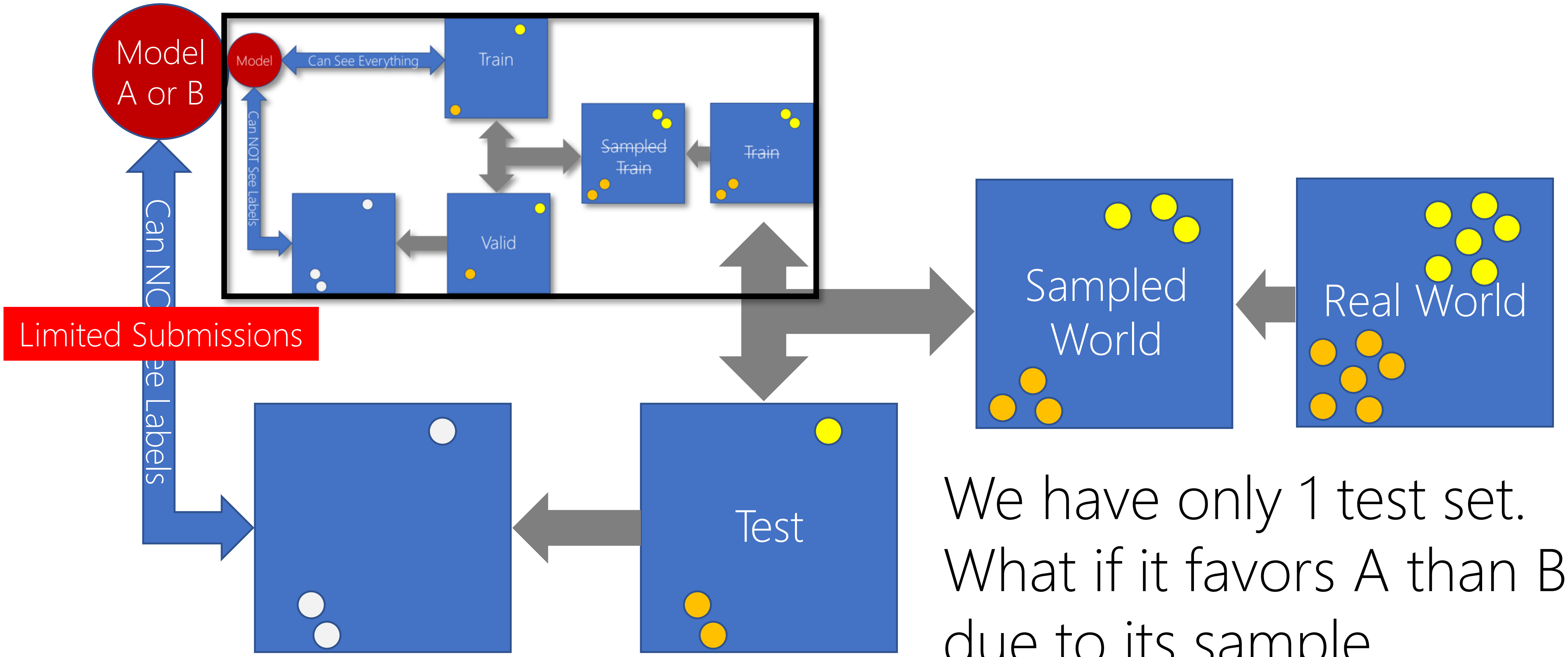
Statistical Significance Testing

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6

A is better than B.

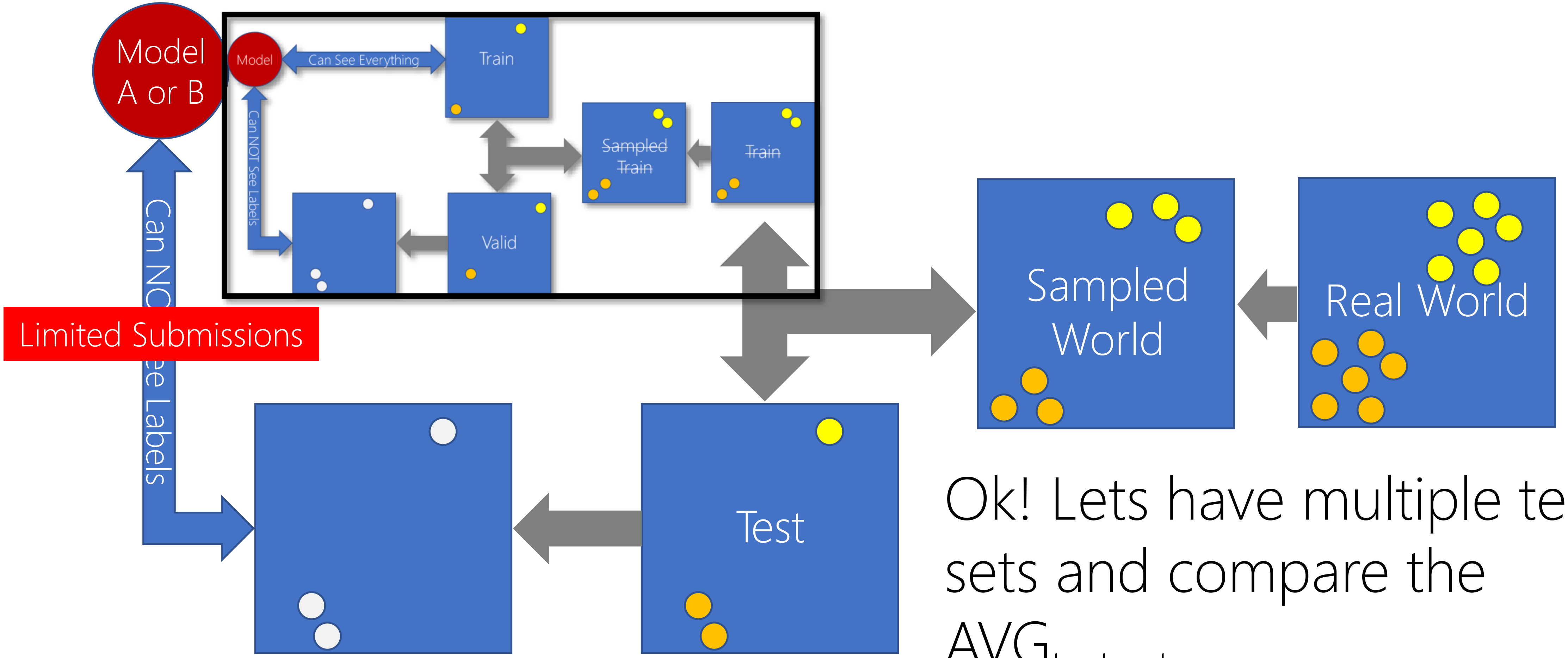
A is significantly better than B → 0.99 >> 0.6

Statistical Significance Testing



We have only 1 test set. What if it favors A than B due to its sample distribution!

Statistical Significance Testing



Ok! Lets have multiple test sets and compare the $\text{AVG}_{\text{test sets}}$

Statistical Significance Testing

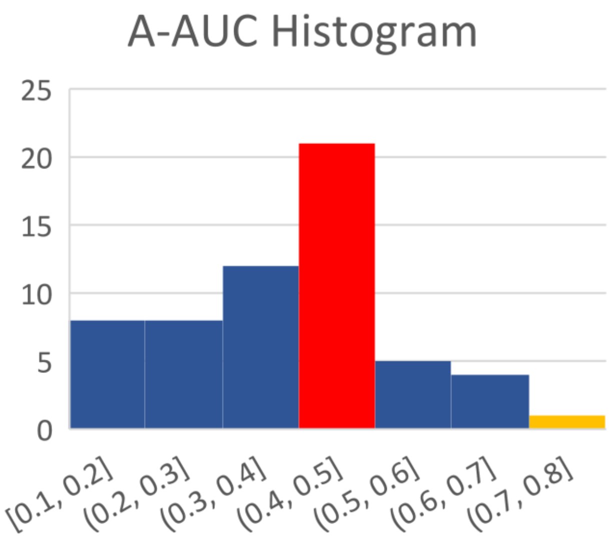
Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
AVG		0.6225		0.6

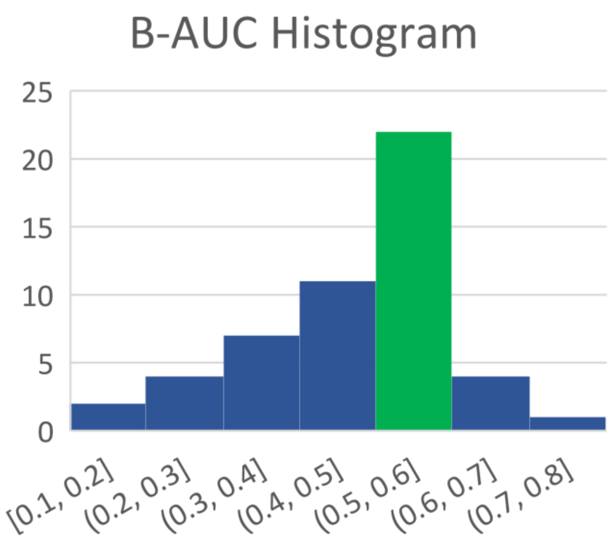
1) By average, A is better than B. However, clearly B is better than A.
2) By average, A is better than B but only slightly NOT significantly!

What is the problem here?

Statistical Significance Testing

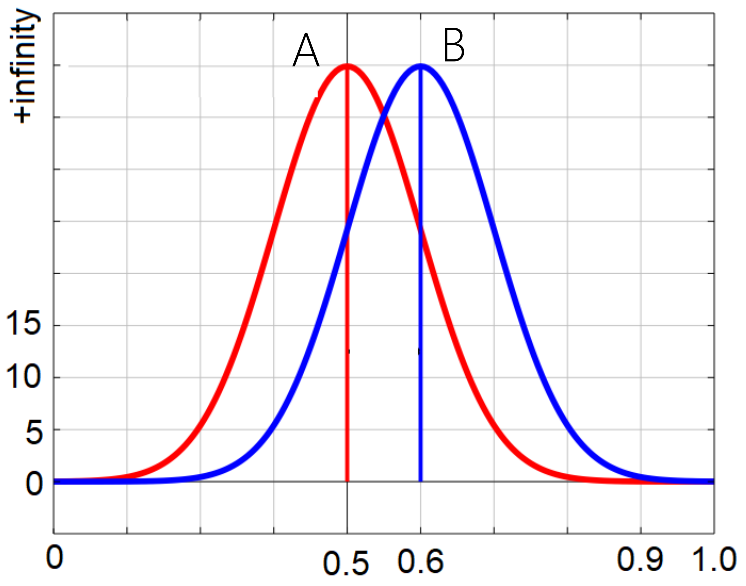
Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
Expectation		E[X]		E[Y]





Statistical Significance Testing

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC_{test-1}	0.99	B-AUC_{test-1}	0.6
Test-2	A-AUC_{test-2}	0.5	B-AUC_{test-2}	0.6
Test-3	A-AUC_{test-3}	0.5	B-AUC_{test-3}	0.6
Test-4	A-AUC_{test-4}	0.5	B-AUC_{test-4}	0.6
Expectation		E[X]		E[Y]



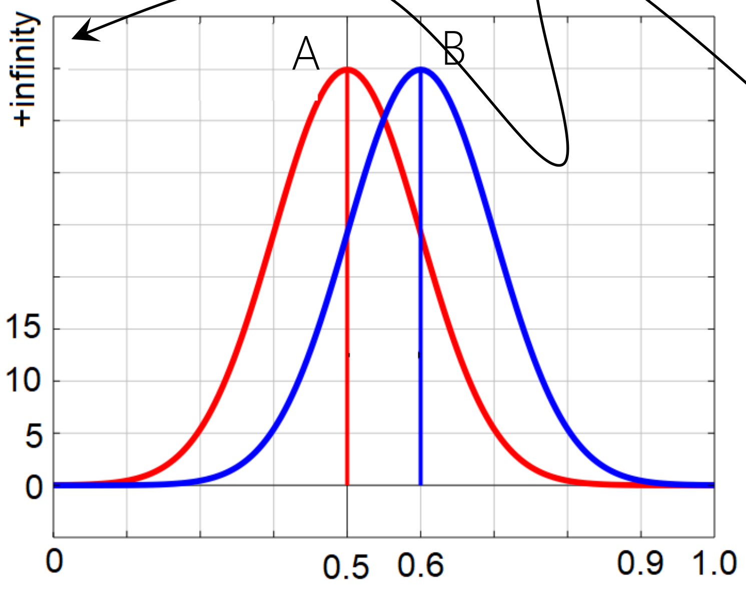
$$P(a<X<b) = \int_a^b f(x)dx.$$

$$P(0.2<X<0.4) > P(0.7<X)$$

$$E(X) < E(Y)$$

Statistical Significance Testing

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
Expectation		E[X]		E[Y]



1) Labeled data is already expensive.
2) Sometimes testing is slow.

Reporting for a lot of runs on different test sets is very challenging!

Statistical Significance Testing: *t*-test

Assuming $X_1, X_2, \ldots X_n$ are n random variable $X_i \sim N(\mu, \delta)$, and iid

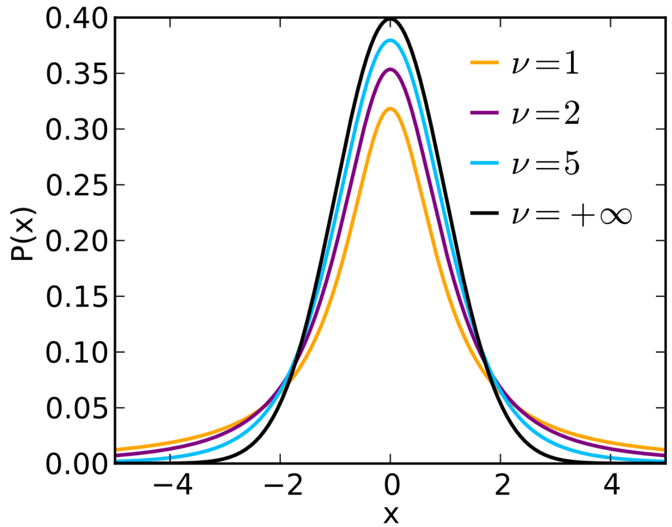
then the [Student's] *t*-distribution with ν=n-1 *degrees of freedom* can be defined as the distribution of the location of the sample mean (AVG=$\bar{x}$) relative to the *true* mean (μ), divided by the sample standard deviation, after multiplying by the standardizing term $\sqrt{n}$.

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n},$$

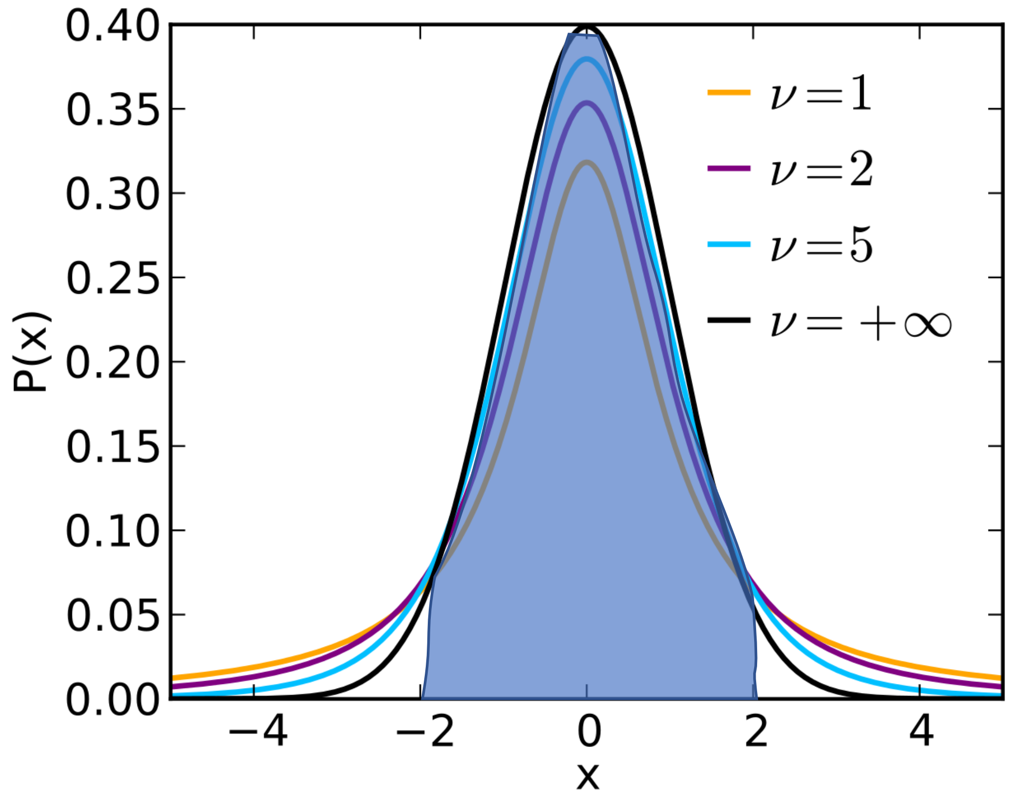
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

t-value

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$



Statistical Significance Testing: *t*-test



$$\Pr(-A < T < A) = 0.9,$$

$$\Pr\left(-A < \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} < A\right) = 0.9,$$

$$\Pr\left(\overline{X}_n - A\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + A\frac{S_n}{\sqrt{n}}\right) = 0.9.$$

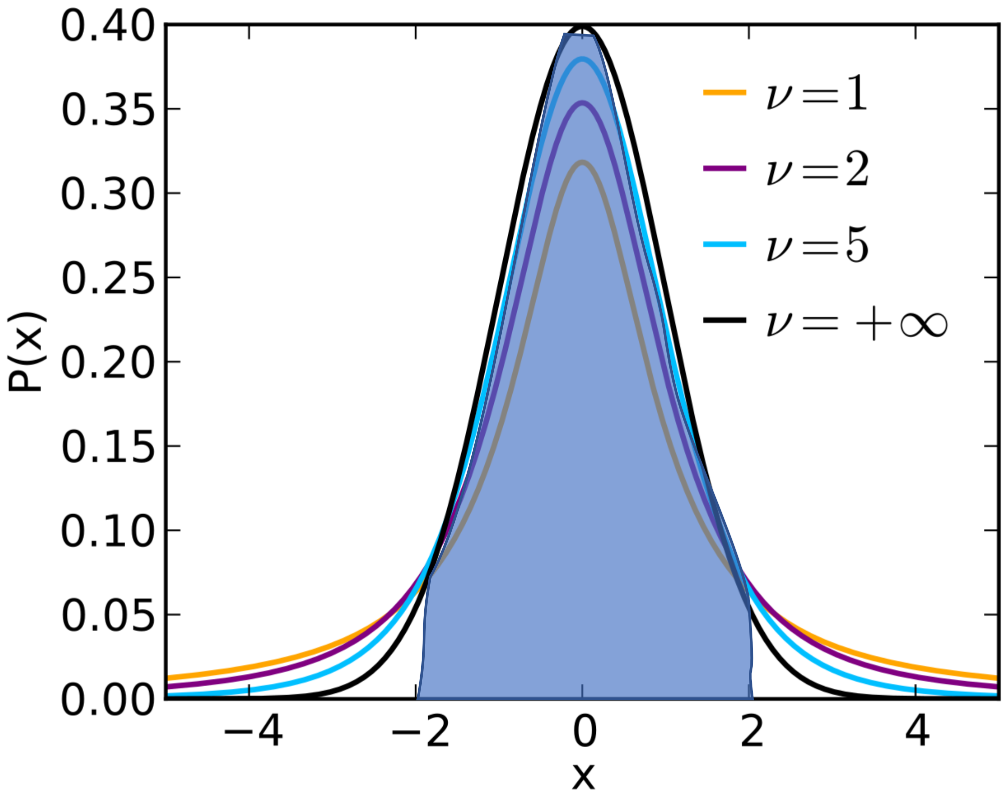
2-tailed (2-sided)

The true mean μ lies between $\overline{X}_n \pm A\frac{S_n}{\sqrt{n}}$
with 90% probability
with 90% confidence
90% of the time

p-value = 1 – Pr(…) = 10% = 0.1

The less p-value, the more confidence.

Statistical Significance Testing: t-test



$$\Pr(-A < T < A) = 0.9,$$

$$\Pr\left(-A < \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} < A\right) = 0.9,$$

$$\Pr\left(\overline{X}_n - A\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + A\frac{S_n}{\sqrt{n}}\right) = 0.9.$$

	90%	95%	97.5%	99%	99.5%	99.95%	**1-Tail Confidence Level**
	80%	90%	95%	98%	99%	99.9%	**2-Tail Confidence Level**
	0.100	0.050	0.025	0.010	0.005	0.0005	**1-Tail Alpha**
df	0.20	0.10	0.05	0.02	0.01	0.001	**2-Tail Alpha**
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	

Statistical Significance Testing: Paired t-test

2-tailed (2-sided), independent two-sample (paired) t-test
Given two groups and assuming equal variances:

$$\Pr\left(-A < \frac{(\overline{X_A} - \overline{X_B}) - (\mu_A - \mu_B)}{s_p\sqrt{\frac{2}{n}}} < A\right) = 0.9,$$

$$s_p = \sqrt{\frac{s_{X_A}^2 + s_{X_B}^2}{2}}.$$

Statistical Significance Testing: Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

Claim (**Alternative** Hypothesis H1):
A is significantly better than B on average ($\mu_A > \mu_B$).
Subclaim: the difference of their averages ($\mu_A - \mu_B$) is significant.
Subsubclaim: $\mu_A \neq \mu_B$

Statistical Significance Testing: Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

Null Hypothesis H0:

A is NOT significantly better than B on average. ($\mu_A \le \mu_B$).

Subnull: The difference of their averages ($\mu_A - \mu_B$) is NOT significant.

Subsubnull: $\mu_A = \mu_B$ is significant.

Statistical Significance Testing: Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

1) We try to reject H0 in favor of H1.
2) Rejecting H0 does not prove H1. Only more confidence about H1.
3) "Not rejecting" H0 does not prove H0.

Statistical Significance Testing: Paired *t*-test

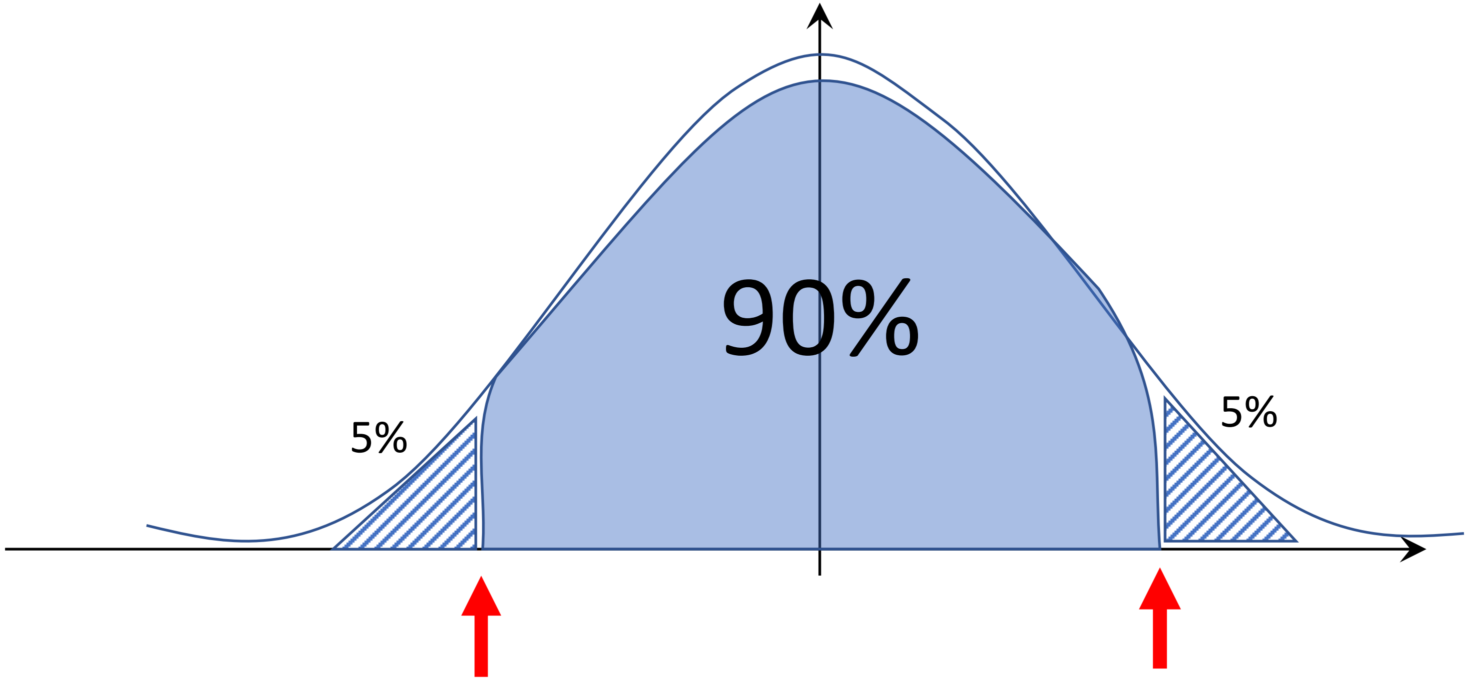
Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

Degree of freedom = 2*(4-1) = 6

Confidence about the H0 = 0.9 = 90% => p-value = 0.1

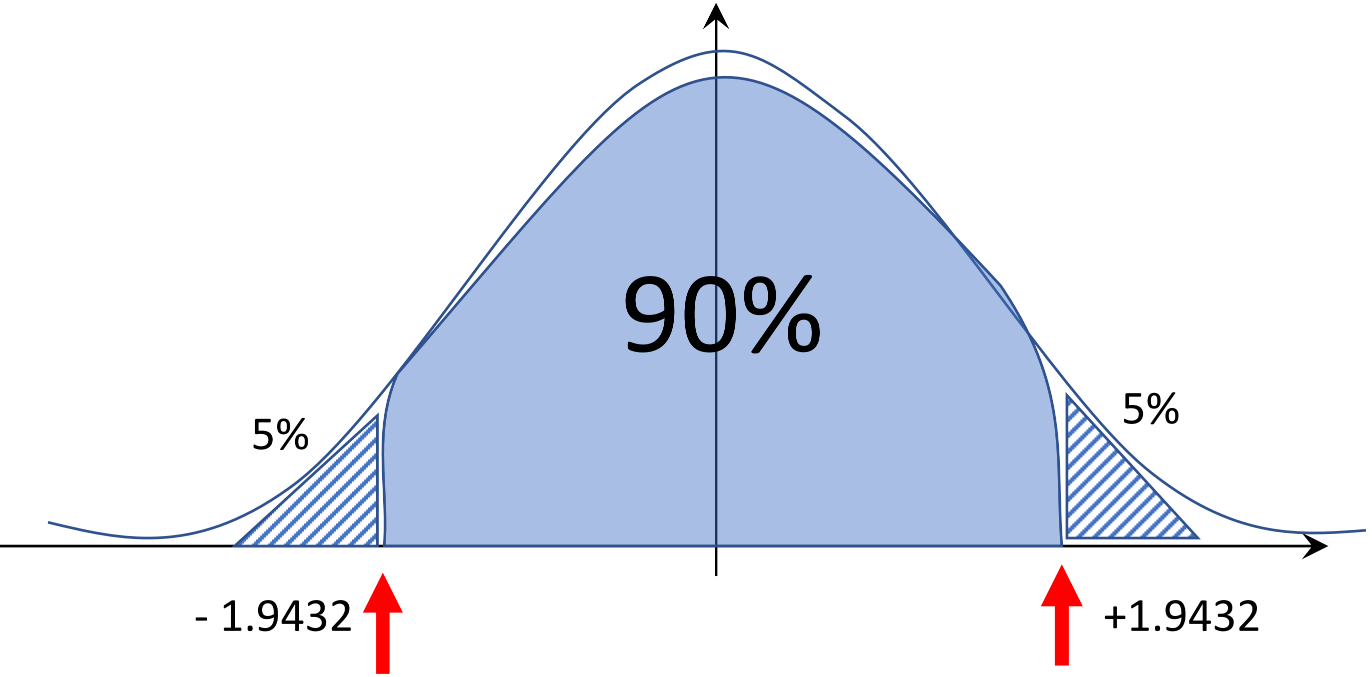
Statistical Significance Testing : Paired t-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	



Statistical Significance Testing : Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test\text{-}1}$	0.99	B-AUC$_{test\text{-}1}$	0.6
Test-2	A-AUC$_{test\text{-}2}$	0.5	B-AUC$_{test\text{-}2}$	0.6
Test-3	A-AUC$_{test\text{-}3}$	0.5	B-AUC$_{test\text{-}3}$	0.6
Test-4	A-AUC$_{test\text{-}4}$	0.5	B-AUC$_{test\text{-}4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	

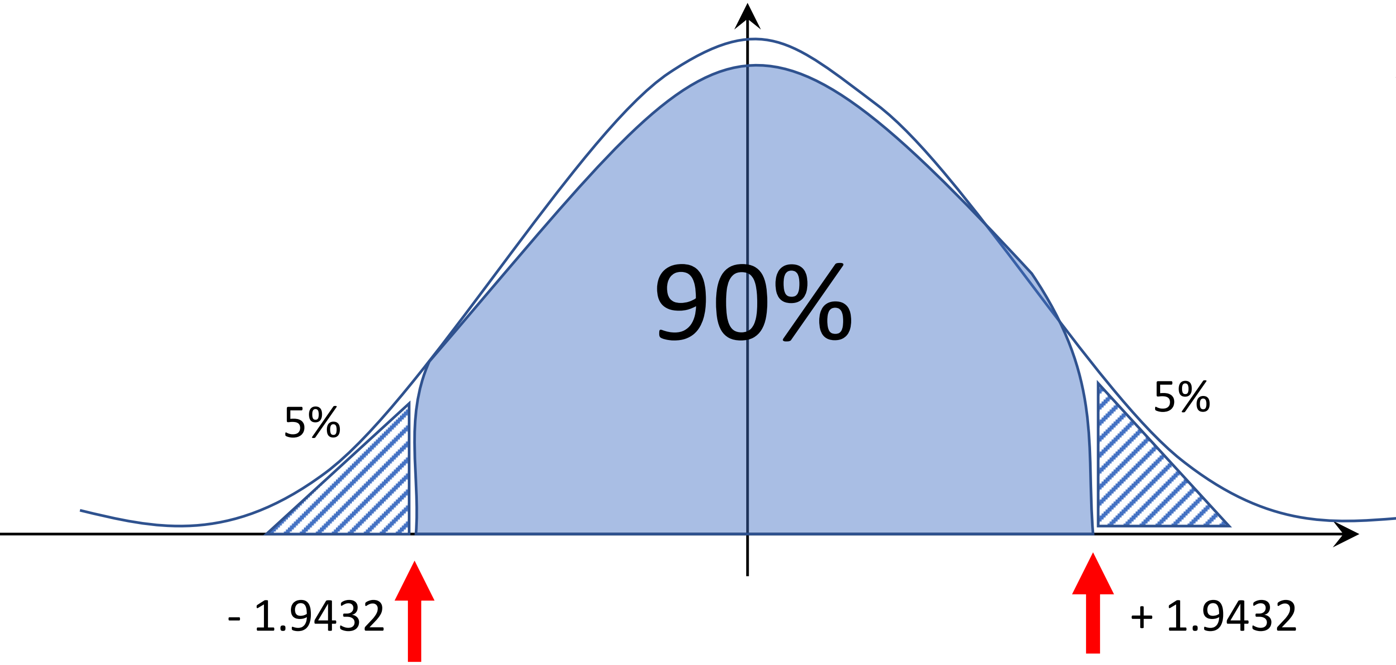
Statistical Significance Testing : Paired t-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$\text{A-AUC}_{\text{test-1}}$	0.99	$\text{B-AUC}_{\text{test-1}}$	0.6
Test-2	$\text{A-AUC}_{\text{test-2}}$	0.5	$\text{B-AUC}_{\text{test-2}}$	0.6
Test-3	$\text{A-AUC}_{\text{test-3}}$	0.5	$\text{B-AUC}_{\text{test-3}}$	0.6
Test-4	$\text{A-AUC}_{\text{test-4}}$	0.5	$\text{B-AUC}_{\text{test-4}}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

$$\Pr\left(-A < \frac{(\overline{X_A} - \overline{X_B}) - (\mu_A - \mu_B)}{s_p\sqrt{\frac{2}{n}}} < A\right) = 0.9,$$

- 1.9432

+ 1.9432



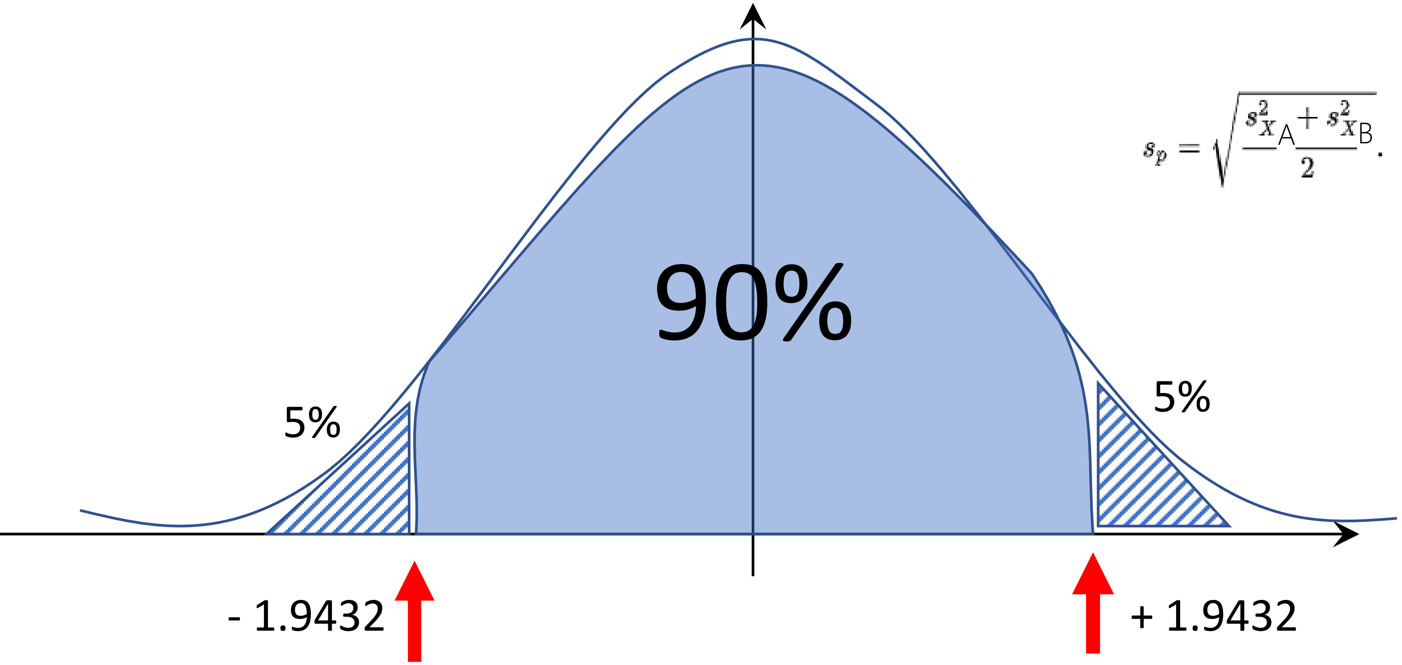
We assumed H0 ($\mu_A = \mu_B$), we should see t lies in between [-A, +A] with 0.9 probability.

Otherwise, what?

Statistical Significance Testing : Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

$$\Pr\left(\underbrace{-A}_{-1.9432} < t = \frac{\bar{X}_A - \bar{X}_B}{s_p\sqrt{\frac{2}{n}}} < \underbrace{A}_{+1.9432}\right) = 0.9,$$



$$s_p = \sqrt{\frac{s^2_{X_A} + s^2_{X_B}}{2}}.$$

$$t = \frac{\overline{X_A} - \overline{X_B}}{S_p\sqrt{\frac{2}{n}}} = \frac{\mathbf{0.6225 - 0.6}}{\sqrt{\frac{\mathbf{0.21+0}}{2}}\sqrt{\frac{2}{4}}} =$$

0.0981980506

Statistical Significance Testing : Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6
Test-2	A-AUC$_{test-2}$	0.5	B-AUC$_{test-2}$	0.6
Test-3	A-AUC$_{test-3}$	0.5	B-AUC$_{test-3}$	0.6
Test-4	A-AUC$_{test-4}$	0.5	B-AUC$_{test-4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

We could not reject H0.
H0 is probable 90% of times!
H0's probability is 90%!

Does not mean H0 is true! Only H0 could not be rejected.
Does not mean H1 is false. Only H1 could not have further support!

Statistical Significance Testing: Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

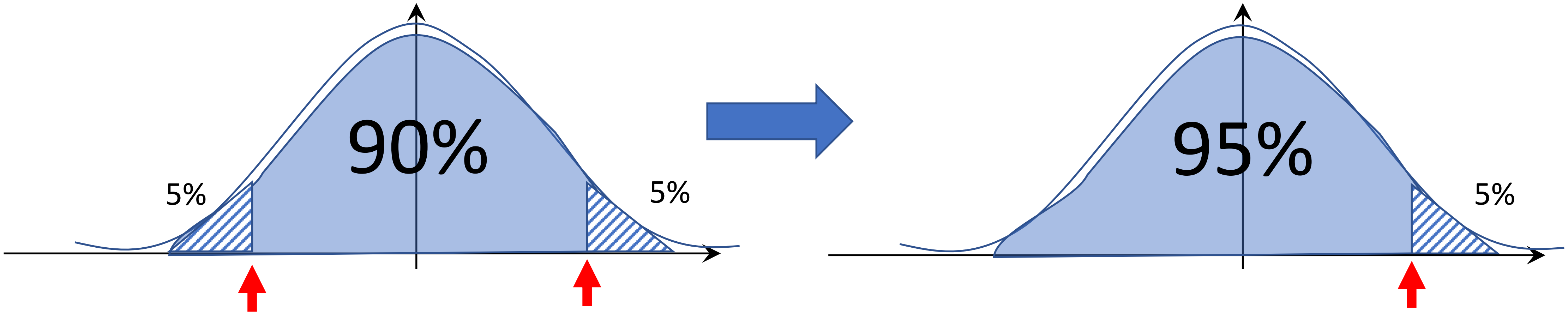
Claim (Alternative Hypothesis H1):
A is significantly better than B on average.
($\mu_A > \mu_B$) should be significant.

Statistical Significance Testing: Paired *t*-test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,
Test-1	$A\text{-}AUC_{test\text{-}1}$	0.99	$B\text{-}AUC_{test\text{-}1}$	0.6
Test-2	$A\text{-}AUC_{test\text{-}2}$	0.5	$B\text{-}AUC_{test\text{-}2}$	0.6
Test-3	$A\text{-}AUC_{test\text{-}3}$	0.5	$B\text{-}AUC_{test\text{-}3}$	0.6
Test-4	$A\text{-}AUC_{test\text{-}4}$	0.5	$B\text{-}AUC_{test\text{-}4}$	0.6
AVG		$\overline{X_A}$	$\overline{X_B}$	

Null Hypothesis H0:
A is NOT significantly better than B on average.
($\mu_A > \mu_B$) is NOT significant.
Indeed, $\mu_A \leq \mu_B$ is significant.

Statistical Significance Testing: Paired t-test



$$\Pr\left(-A < \frac{(\overline{X_A} - \overline{X_B}) - (\mu_A - \mu_B)}{s_p\sqrt{\frac{2}{n}}} < A\right) = 0.9 \Rightarrow \Pr\left(\frac{(\overline{X_A} - \overline{X_B})}{s_p\sqrt{\frac{2}{n}}} > A\right) = 0.05$$

$$0.0981980506 < 1.9432$$

Null Hypothesis H0:
The data is confirming H0.

df							
	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.[illegible]060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	

Statistical Significance Testing: Paired *t*-test

Notes:

- Don't forget about the assumptions:
  - Normal distribution of random variables
  - i.i.d of the random variables
    - K-fold on test set is not valid.
  - Same standard deviation
    - (t-test has a version for different standard deviation)

In NLP, we don't generally use paired *t*-tests

Most metrics are not normally distributed

Non-parametric Tests

Non-parametric Tests: bootstrap test

Test Sets	A-AUC	E.g.,	B-AUC	E.g.,	δ(A, B)
Test-1	A-AUC$_{test-1}$	0.99	B-AUC$_{test-1}$	0.6	0.39

1) A is better than B by $\boldsymbol{\delta}$

a) That was by luck

2) H1: The chance of luck is very low

b) H0: The chance of luck is high.

3) Assuming A NOT better than B, there should be a good chance for A better than B then!

$P(\boldsymbol{\delta}_{A,B}(\text{other test sets}) > 0.39 \mid H0) \sim 0.30$ or 0.40

Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. CRC press.

Non-parametric Tests: bootstrap test

	1	2	3	4	5	6	7	8	9	10	A%	B%	$\delta()$
x	AB	A~~B~~	AB	~~A~~B	A~~B~~	~~A~~B	A~~B~~	AB	~~A~~~~B~~	A~~B~~	.70	.50	.20
$x^{*(1)}$	A~~B~~	AB	A~~B~~	~~A~~B	~~A~~B	A~~B~~	~~A~~B	AB	~~A~~~~B~~	AB	.60	.60	.00
$x^{*(2)}$	A~~B~~	AB	~~A~~~~B~~	~~A~~B	~~A~~B	AB	~~A~~B	A~~B~~	AB	AB	.60	.70	-.10
...													
$x^{*(b)}$													

Figure 4.8 The bootstrap: Examples of b pseudo test sets being created from an initial true test set x. Each pseudo test set is created by sampling $n = 10$ times with replacement; thus an individual sample is a single cell, a document with its gold label and the correct or incorrect performance of classifiers A and B.

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In EMNLP 2012, 995–1005.

Non-parametric Tests: bootstrap test

function BOOTSTRAP(test set x, num of samples b) **returns** $p\text{-}value(x)$

Calculate $\delta(x)$ # how much better does algorithm A do than B on x
for $i = 1$ **to** b **do**
  for $j = 1$ **to** n **do** # Draw a bootstrap sample $x^{*(i)}$ of size n
    Select a member of x at random and add it to $x^{*(i)}$
  Calculate $\delta(x^{*(i)})$ # how much better does algorithm A do than B on $x^{*(i)}$
for each $x^{*(i)}$
  $s \leftarrow s + 1$ **if** $\delta(x^{*(i)}) > 2\delta(x)$
p-value(x) $\approx \frac{s}{b}$ # on what % of the b samples did algorithm A beat expectations?
return p-value(x)

Figure 4.9 A version of the bootstrap algorithm after Berg-Kirkpatrick et al. (2012).

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In EMNLP 2012, 995–1005.