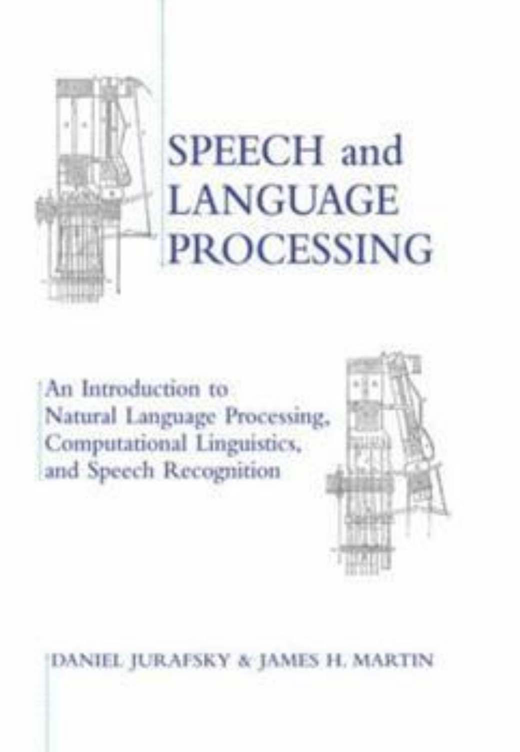
Ukraine International Airlines Flight 752

https://en.wikipedia.org/wiki/Ukraine_International_Airlines_Flight_752

www.youtube.com/watch?v=etlBZInTE-I



Naive Bayes and Sentiment Classification

CH04

Evaluate a Classifier

"We're addicted and annotated data is our heroine."

—?

Supervide

Gold Labels

aka. Golden Truth, Golden Standard

Human-defined classes/labels for each document

Human judgment

Manually labeled

Manually annotated

Gold Labels

to err is human!

Gold Labels

to err is human; to forgive, divine!

"Save This Word! All people commit sins and make mistakes. God forgives them, and people are acting in a godlike (divine) way when they forgive." - An Essay on Criticism, Alexander Pope.

Silver Labels

aka. Silver Truth, Silver Standard

Gold is very expensive!

Finding gold is needs a lot of effort!

automated-defined classes/labels for each document

Machine judgment

Machine labeled

Machine annotated

Transduction

Transductive Inference

Data has the labels already!

Language Models

Evaluation

$$(x, y) \rightarrow f(x) == y$$

Evaluation

Boolean (Binary) Classifier

Contingency Table

aka. Confusion Matrix

	Gold Positive	Gold Negative
Model Positive	True Positive	False Positive
Model Negative	False Negative	True Negative

Type I

Perfect Classifier

	Gold Positive	Gold Negative
Model Positive	N+	0
Model Negative	0	N-

2N

Other Metrics

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

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

Precision

High vs. Low

gold standard labels

		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** = $\frac{tp}{tp+fp}$
	system negative	**false negative**	**true negative**	
		recall = $\frac{tp}{tp+fn}$		**accuracy** = $\frac{tp+tn}{tp+fp+tn+fn}$

spam +

not spam

Figure 4.4 Contingency table

What scenarios require high precision?

Recall

High vs. Low

gold standard labels

Spam, not Spam

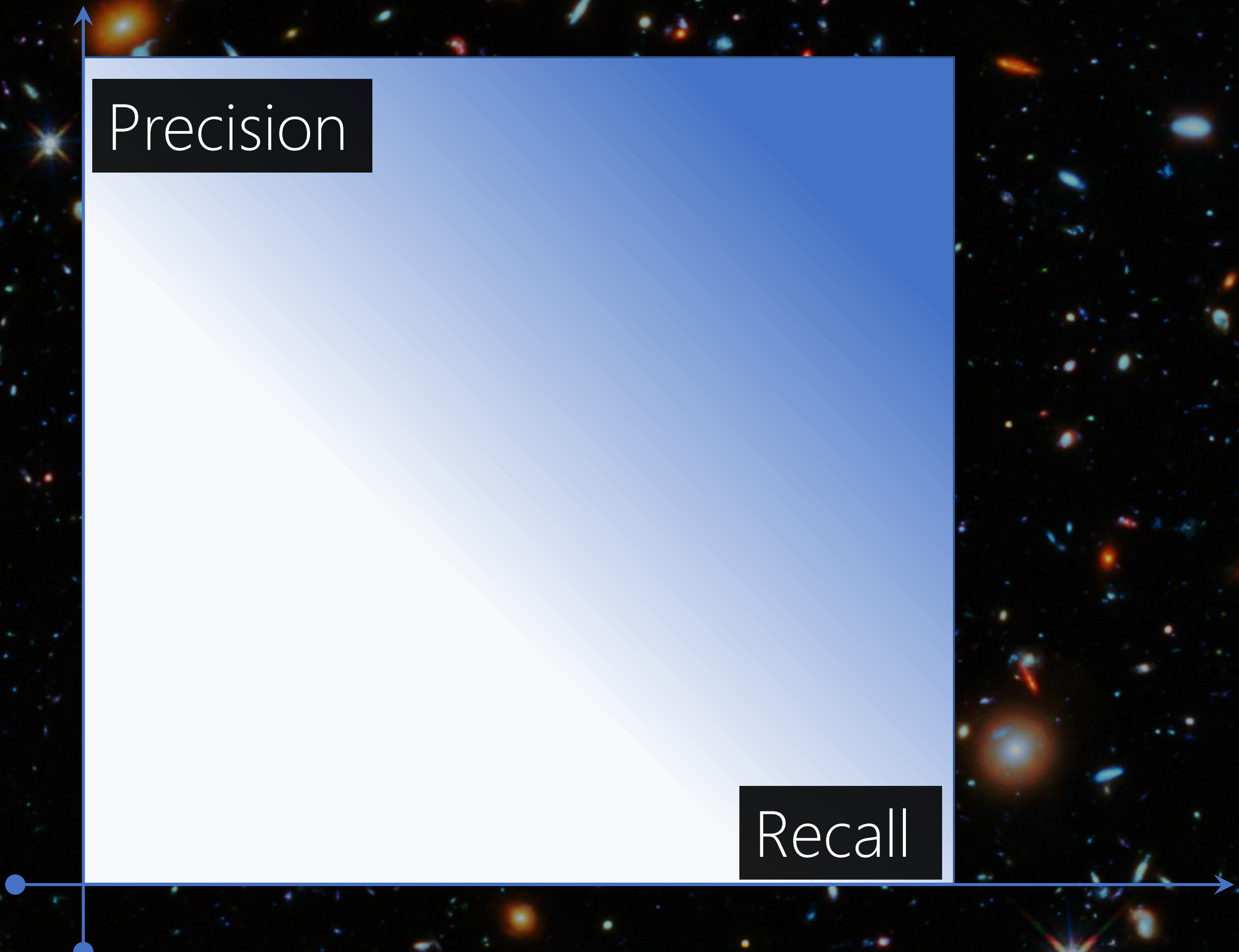
		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{tp}{tp+fp}$
	system negative	**false negative**	**true negative**	release, not
		recall $= \frac{tp}{tp+fn}$		**accuracy** $= \frac{tp+tn}{tp+fp+tn+fn}$

missile

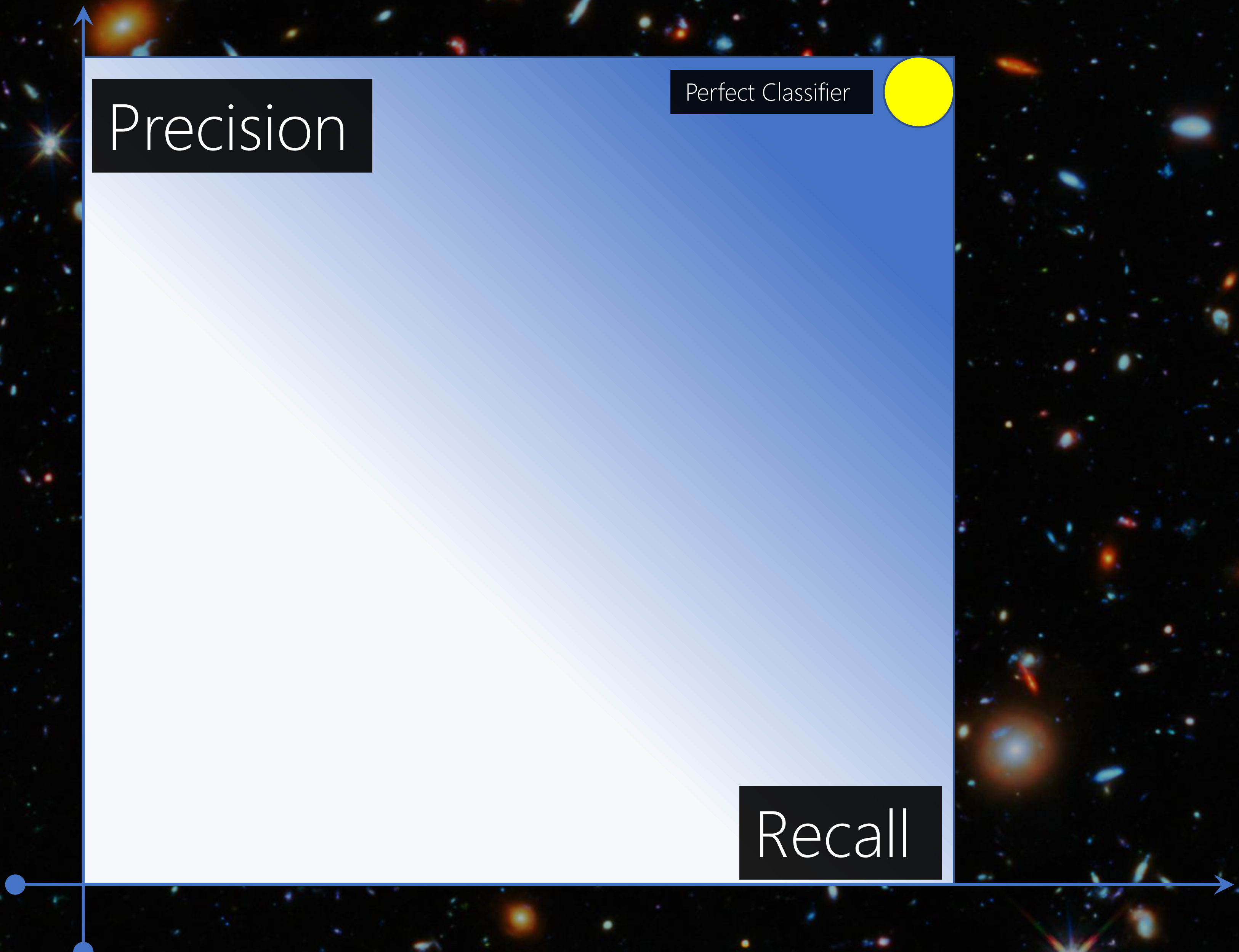
Figure 4.4 Contingency table

What scenarios require high recall?

Precision-Recall



Precision-Recall



Balance Classes

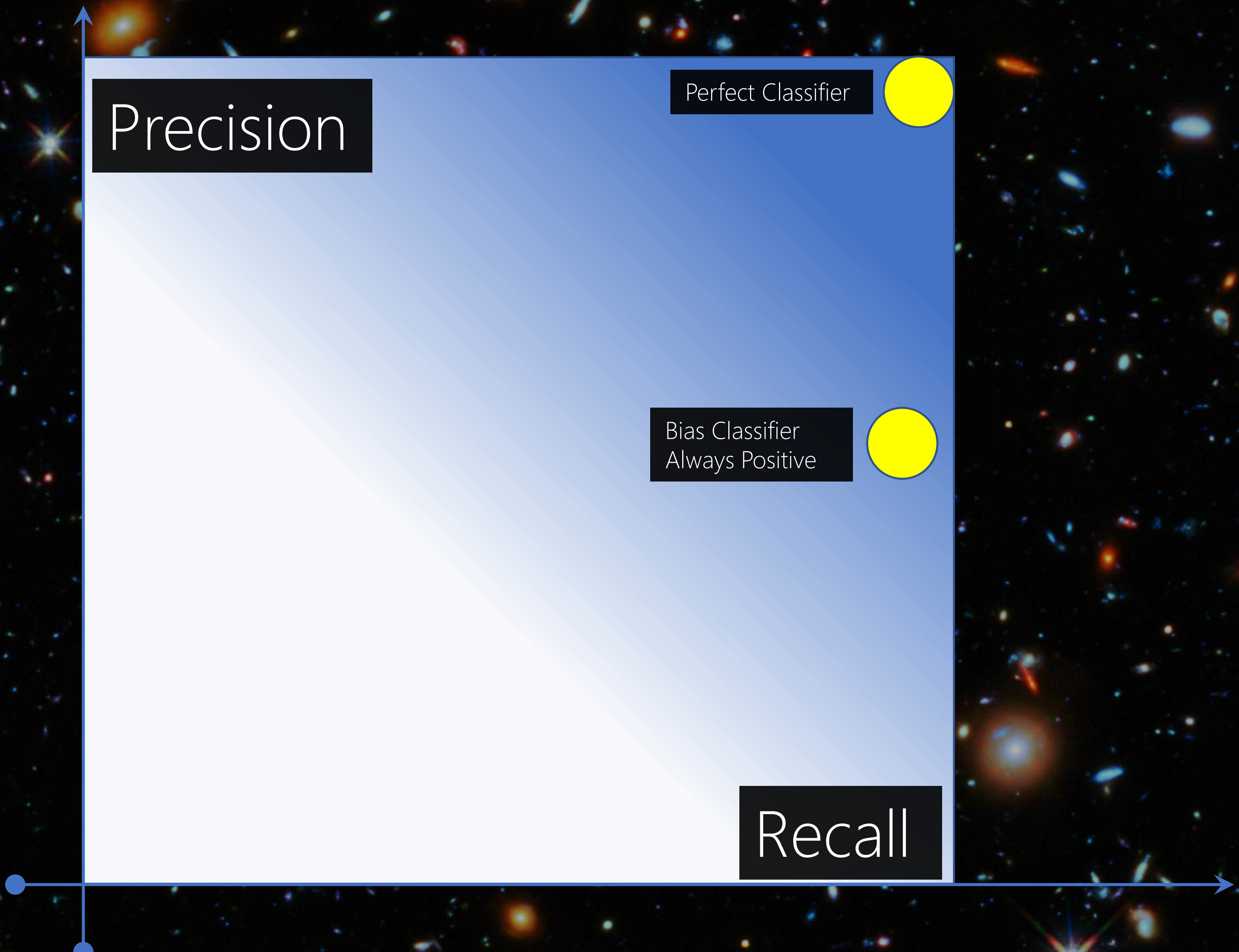
~50% Positive, ~50% Negative

	Gold Positive (50)	Gold Negative (50)
Model Positive	50	50
Model Negative	0	0

$$\text{Precision} = \frac{\mathbf{50}}{\mathbf{50+50}} = 0.5$$

$$\text{Recall} = \frac{\mathbf{50}}{\mathbf{50+0}} = 1.0$$

Precision-Recall



Balance Classes

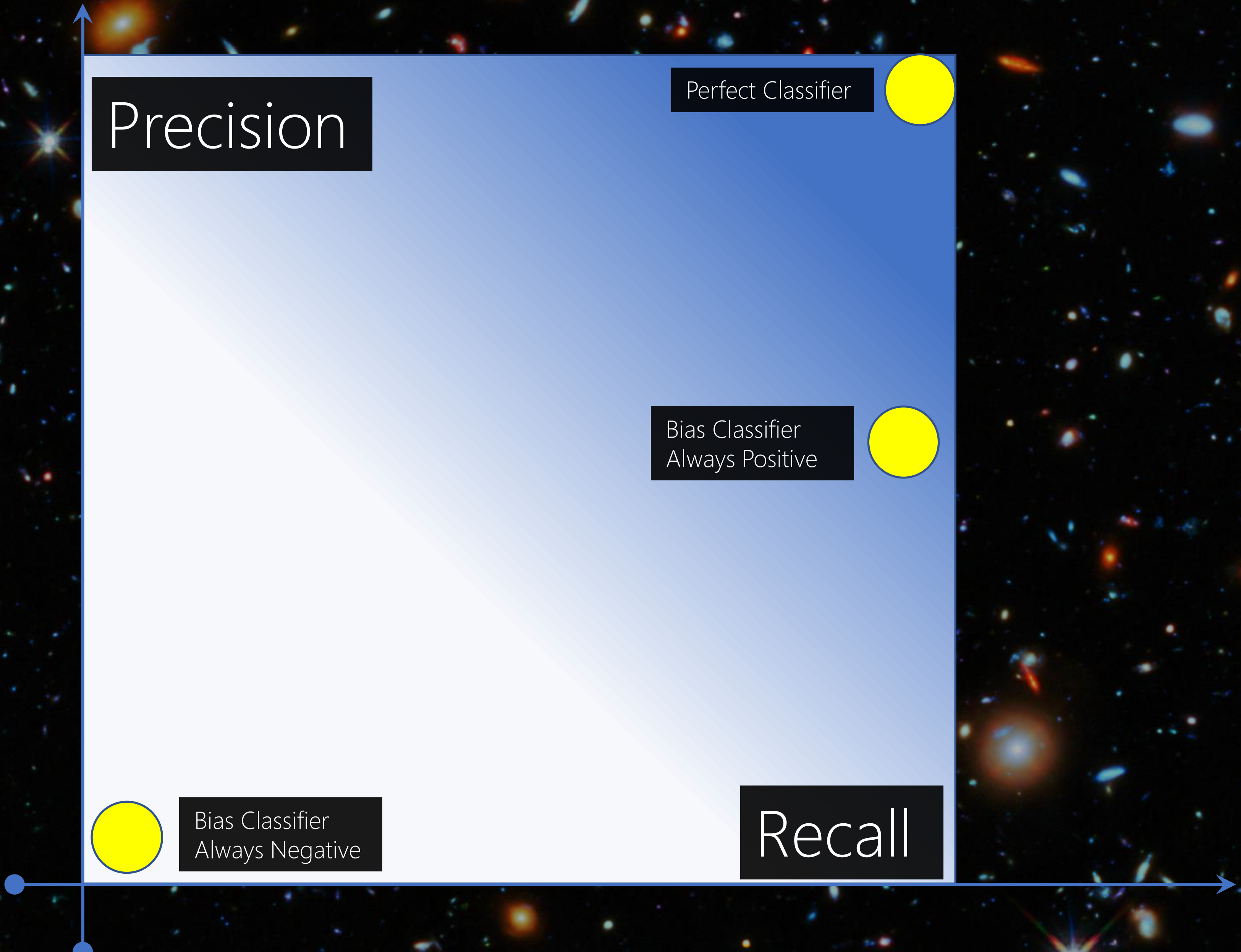
~50% Positive, ~50% Negative

	Gold Positive (50)	Gold Negative (50)
Model Positive	0	0
Model Negative	50	50

Precision $= \frac{0}{0+0} = 0.0$

Recall$=\frac{0}{50+0} = 0.0$

Precision-Recall



Balance Classes

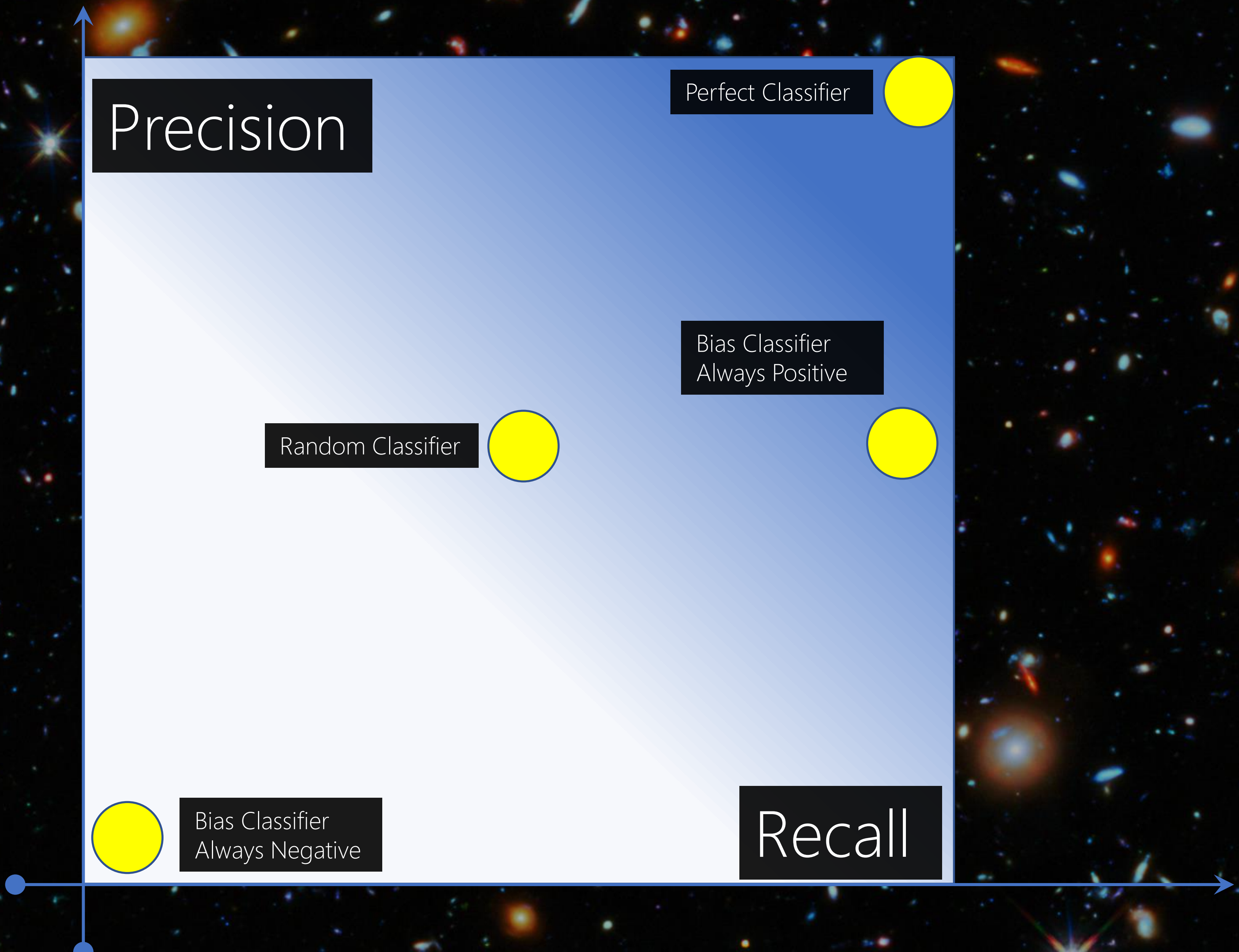
~50% Positive, ~50% Negative

	Gold Positive (50)	Gold Negative (50)
Model Positive	25	25
Model Negative	25	25

$$\text{Precision} = \frac{25}{25+25} = 0.5$$

$$\text{Recall} = \frac{25}{25+25} = 0.5$$

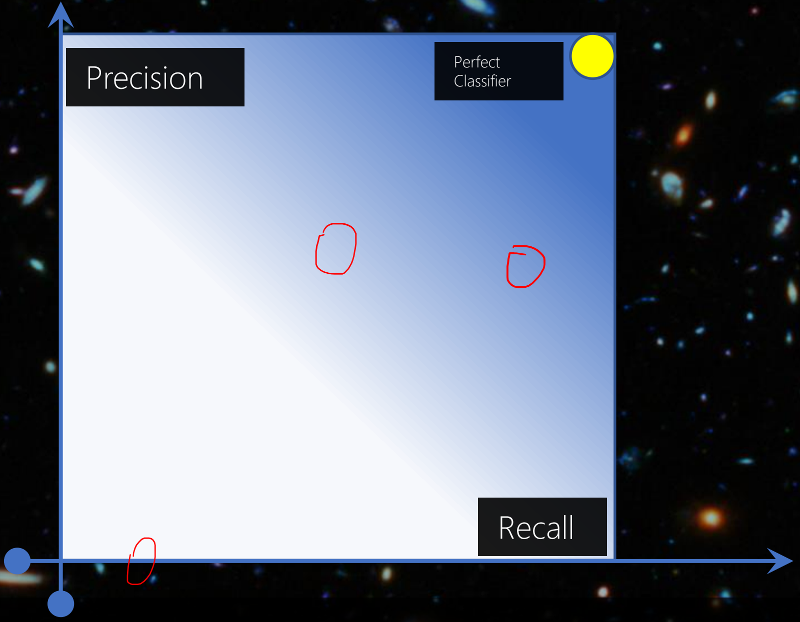
Precision-Recall



Imbalance (Unbalanced) Classes

~10% Positive, ~90% Negative

	Gold Positive (10)	Gold Negative (90)
Model Positive	?	?
Model Negative	?	?



Bias Positive Classifier?

Bias Negative Classifier?

Random Classifier?

Average of Precision and Recall: A Single Metric

$$\text{AVG-PR} = \frac{P+R}{2} = 0.5$$

Same weights to Precision and Recall

Not fair! high precision may discount low recall or vice versa

Average of Precision and Recall: A Single Metric

$$\text{Harmonic AVG-PR} = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2\left(\frac{P \times R}{P + R}\right)$$

Same weights to Precision and Recall

Conservative! More toward the lower number.

$$\text{HarmonicMean}(a_1, a_2, a_3, a_4, \ldots, a_n) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}$$

F_β-Measure

Weights to Precision and Recall Separately

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \text{or} \left(\text{with } \beta^2 = \frac{1-\alpha}{\alpha}\right) \quad F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \begin{cases} \beta > 1: & favors\ Recall \\ \beta = 1: & 2\left(\frac{P \times R}{P + R}\right) \\ 0 \leq \beta < 1: & favors\ Precision \end{cases}$$

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

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

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

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

$$R_{happy} = \frac{\mathbf{8}}{\mathbf{8+5+3}} = ?$$

$$R_{sad} = \frac{\mathbf{60}}{\mathbf{10+60+30}} = ?$$

$$R_{neutral} = \frac{\mathbf{200}}{\mathbf{1+50+200}} = ?$$

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} =?

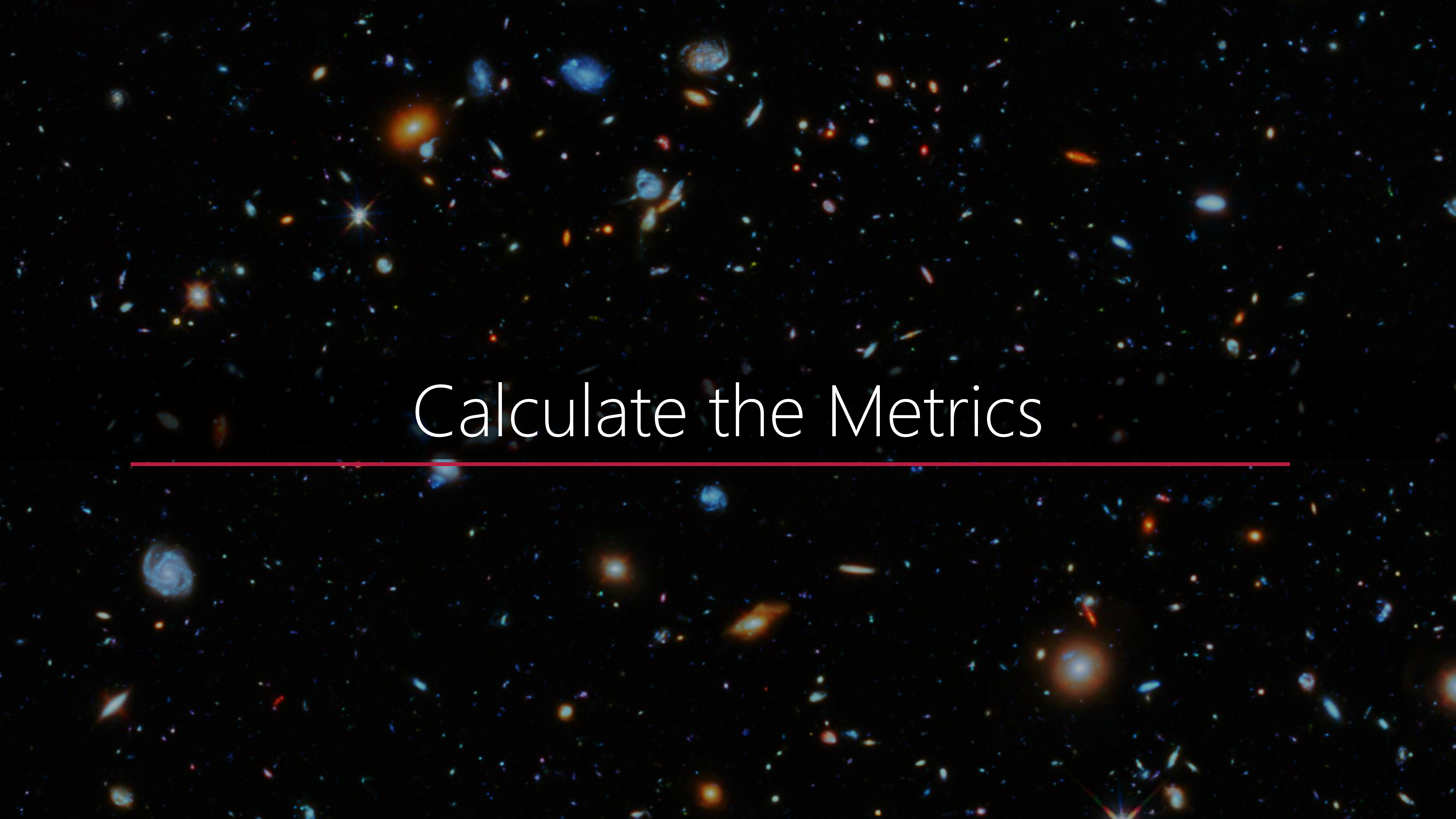
Macro vs. Micro Averaging

Macro vs. Micro

M[illegible]cro is dominated by the more frequent class.

M[illegible]cro better reflects the statistics of the smaller classes.

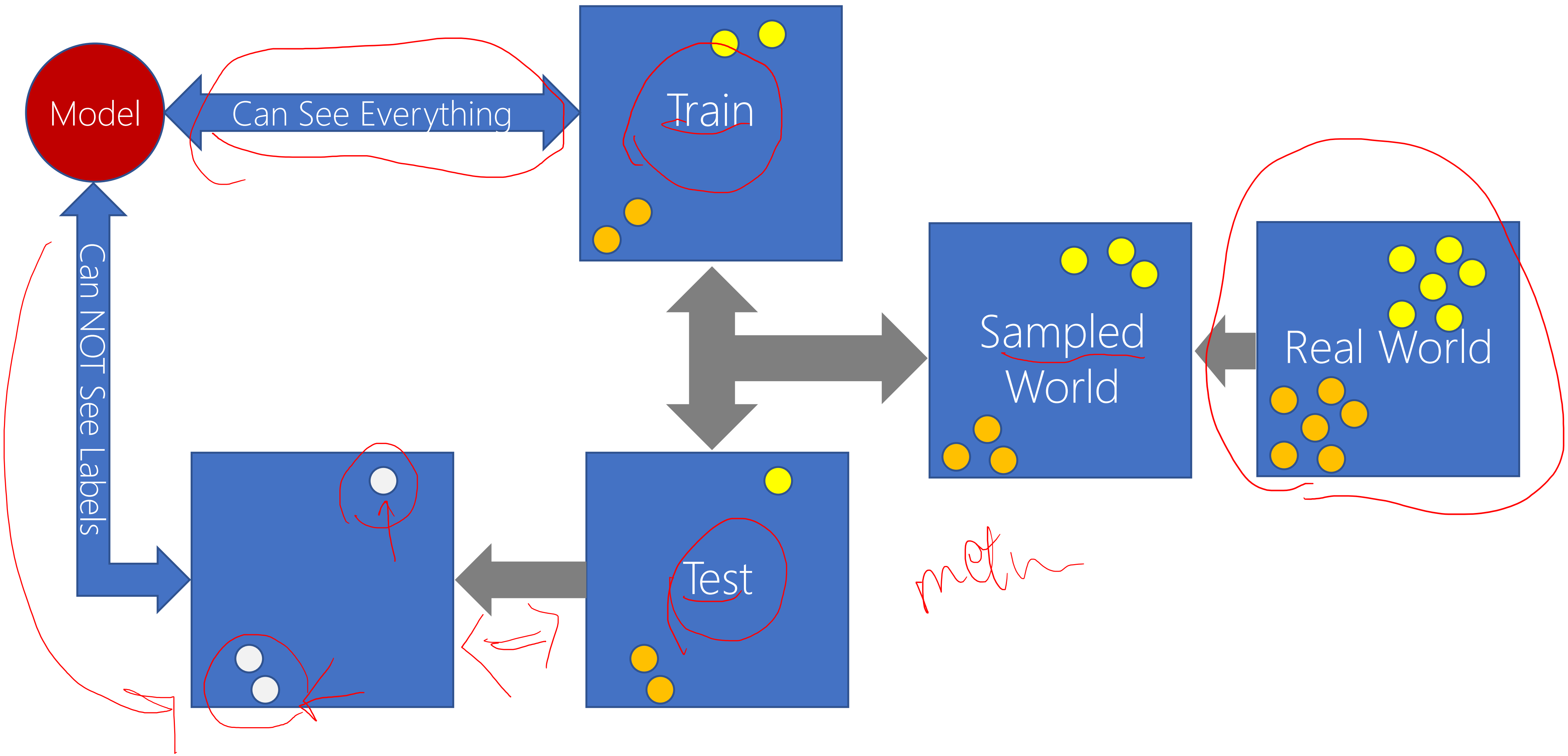
M[illegible]cro more appropriate when performance on all the classes is equally important.



Calculate the Metrics

Labeled Data = {Train} U {Test}

Labeled Data = {Train} U {Test}



Labeled Data = {Train} U {Test}

$$\{\text{Train}\} \cap \{\text{Test}\} \stackrel{?}{=} \emptyset$$

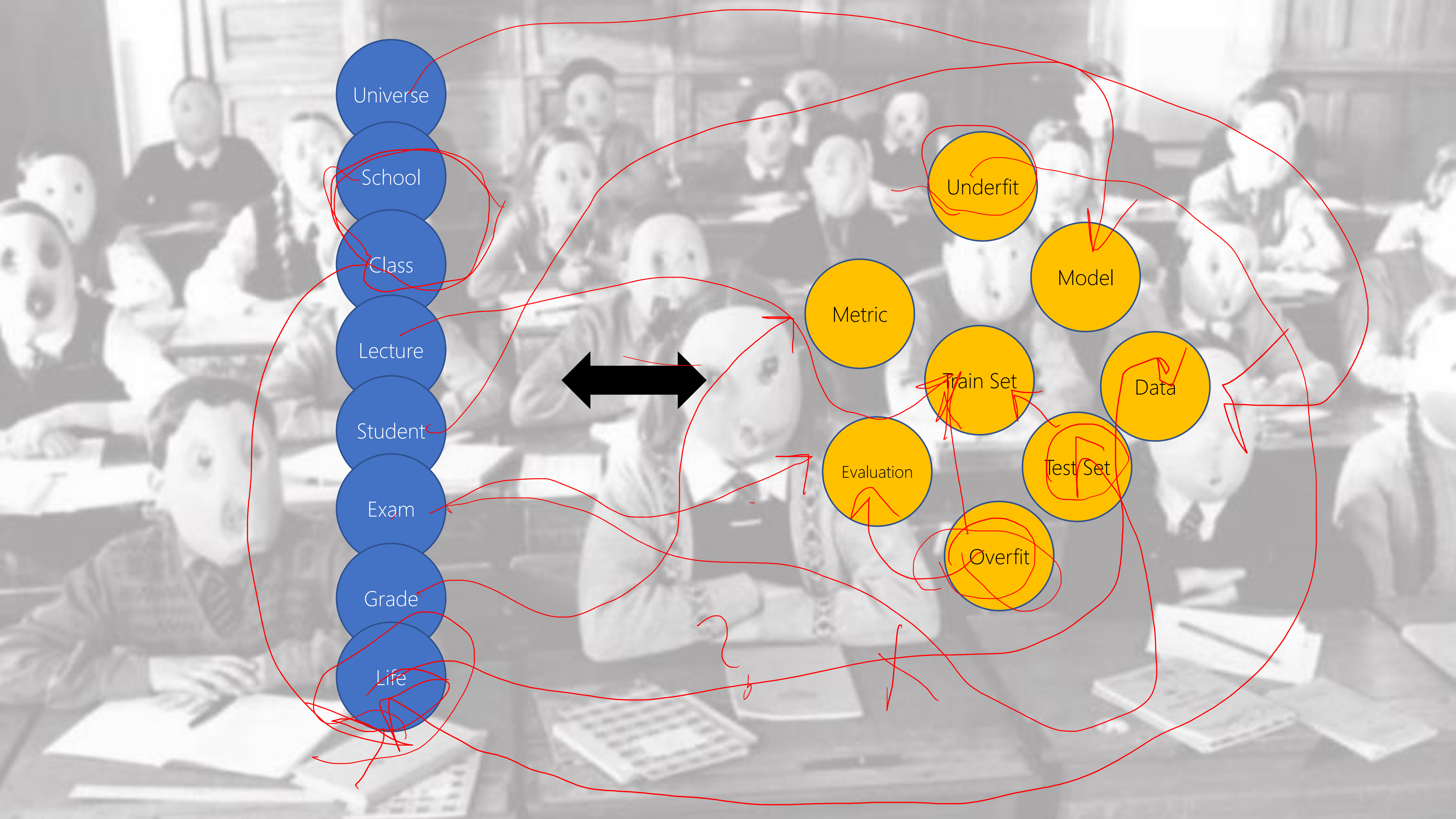
Imbalance Labeled Data = {Train} U {Test}

$$\{\text{Train}\} \cap \{\text{Test}\} \stackrel{?}{=} \emptyset$$

Train and test sets presumably follow same distribution!

Underfit → Balance fit ← Overfit


Another Brick in the Wall
- Pink Floyd
- The Wall

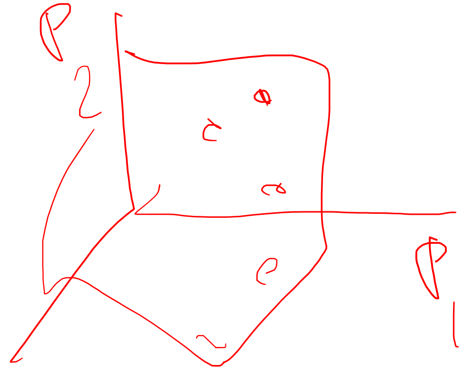


Universe
School
Class
Lecture
Student
Exam
Grade
Life
Underfit
Model
Metric
Train Set
Data
Evaluation
Test Set
Overfit

Model Tuning

Find the best running settings of the mode

- n-Gram LM: what is the best n?
- Probs. assumptions
- Neural Model: #layers, Activation functions



Model Tuning

Find the best running settings of the mode

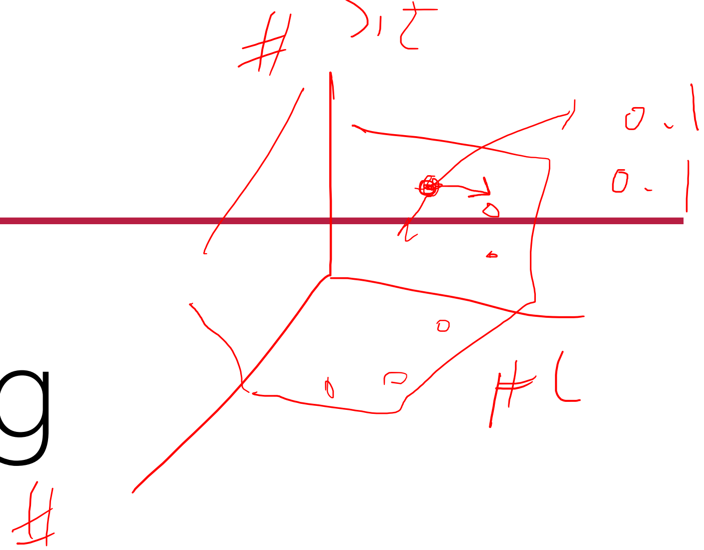
- Checking the performance of model on Train and Test
- For all different possibilities

Blind grid search! Brute-force

1 2 3 4 5 6 10

3

Model Tuning

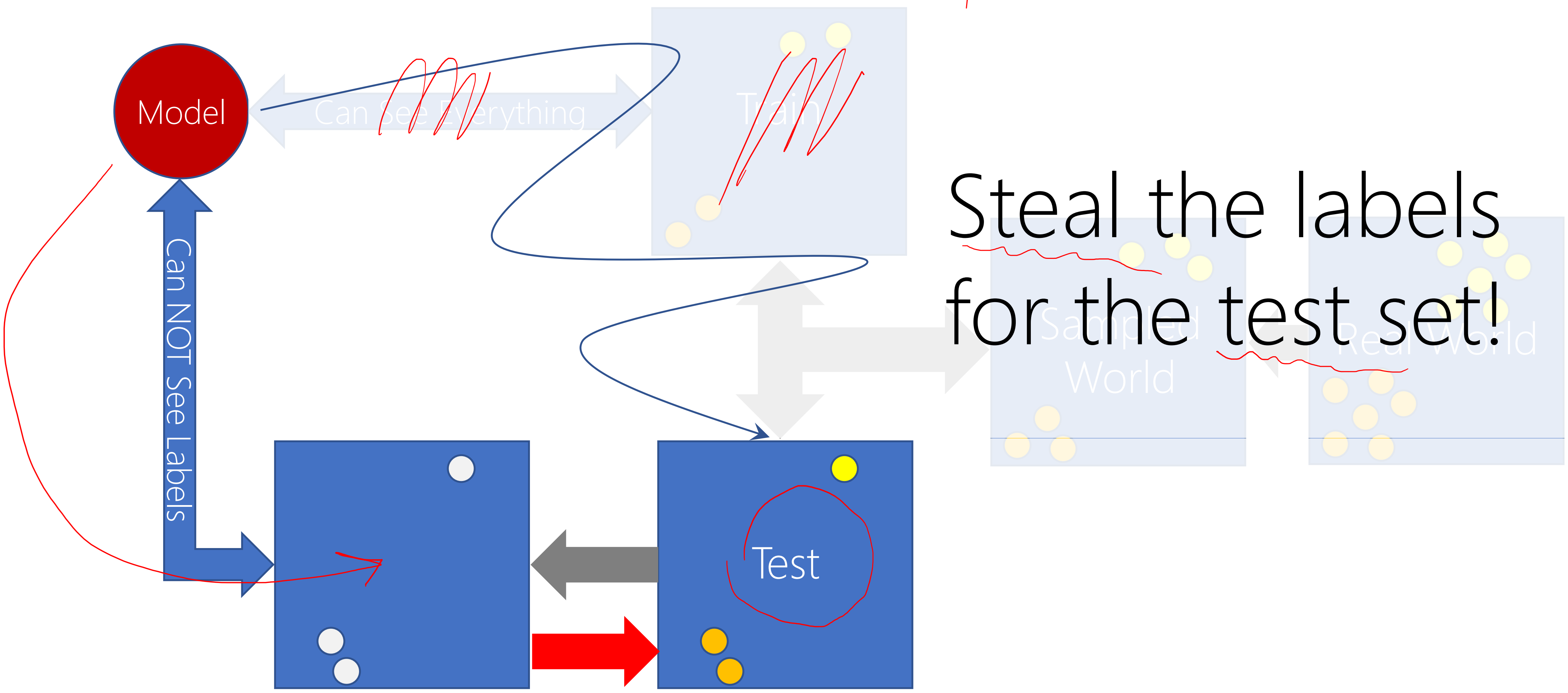


Find the best running settings of the mode

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

Guided grid search!

Labeled Data = {Train} U {Test}



Steal the labels for the test set!

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

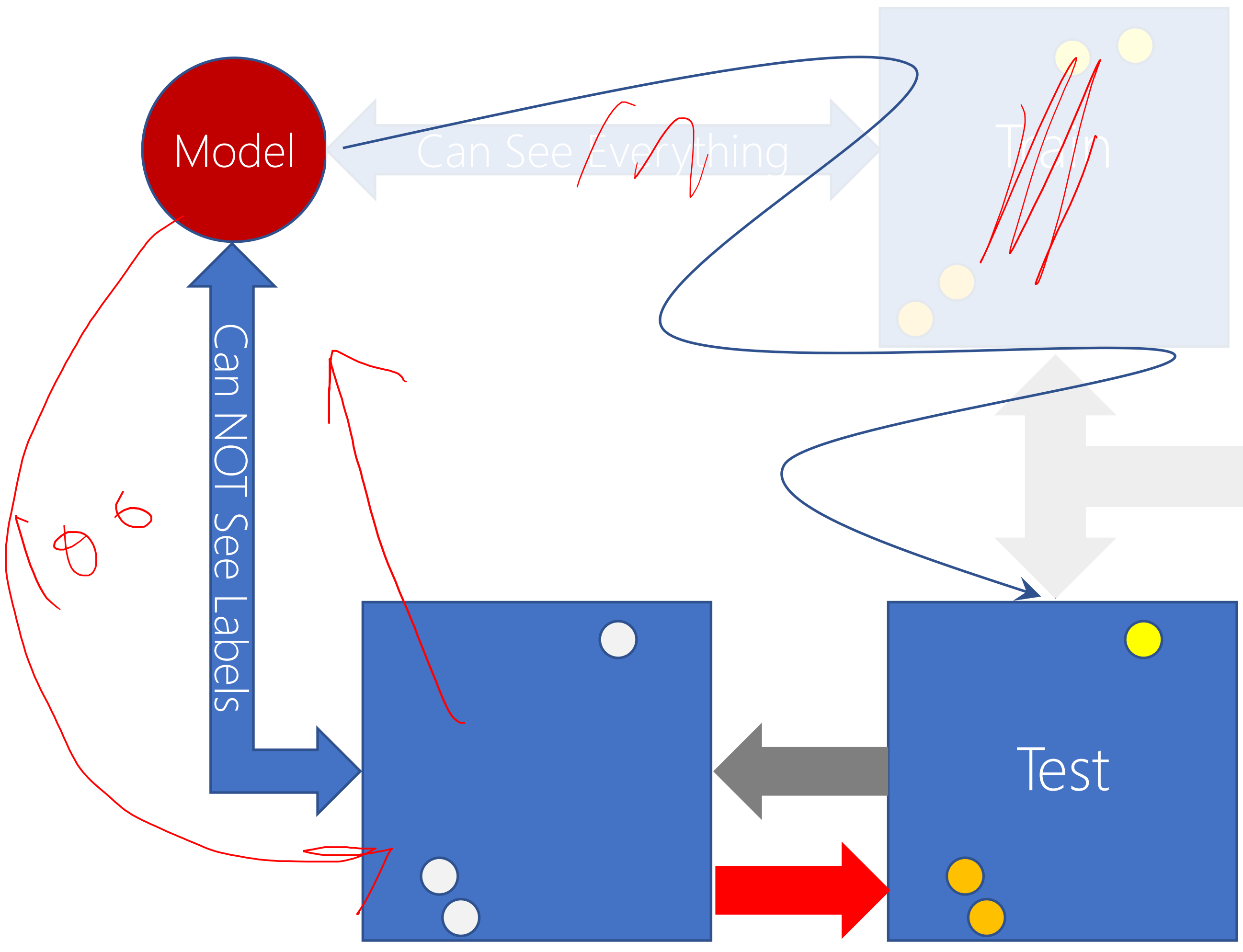
Dear Participants,

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}



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!

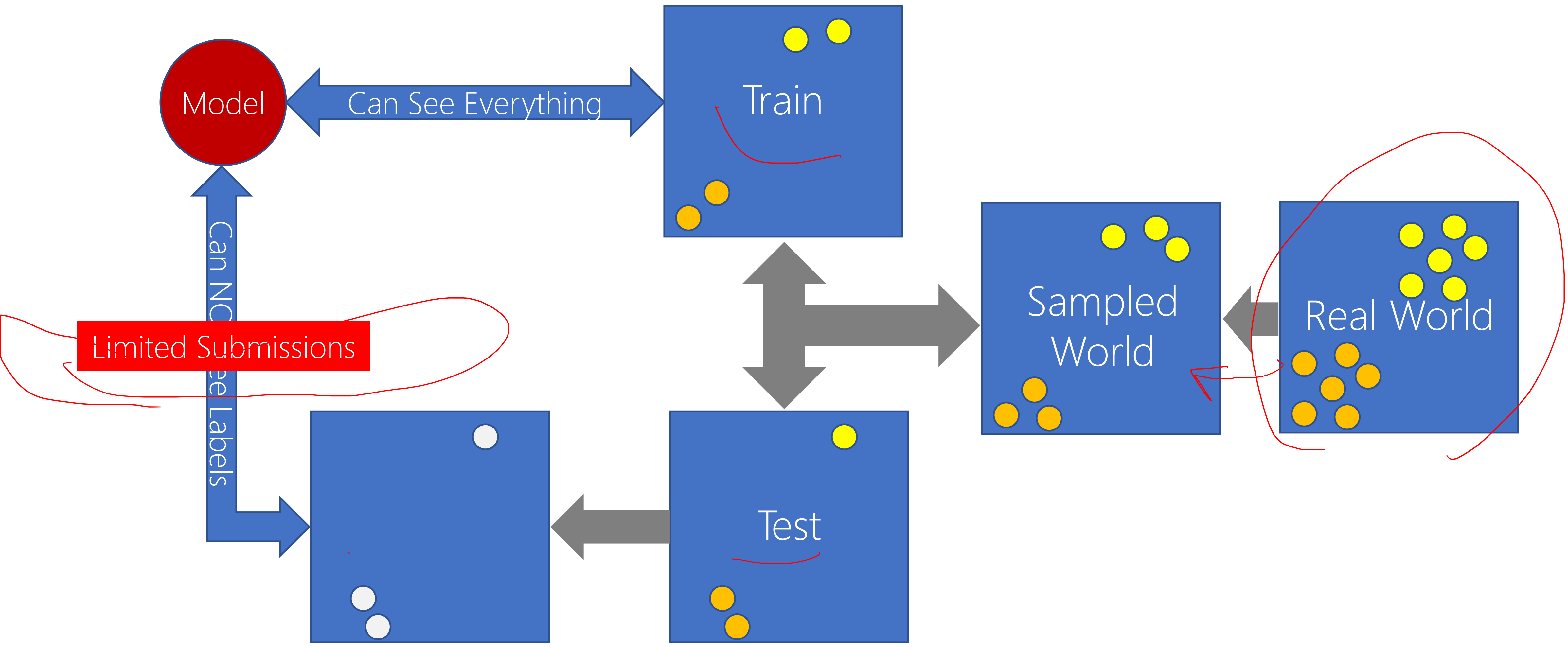
Labeled Data = {{Train} U {Valid}} U {Test}

2

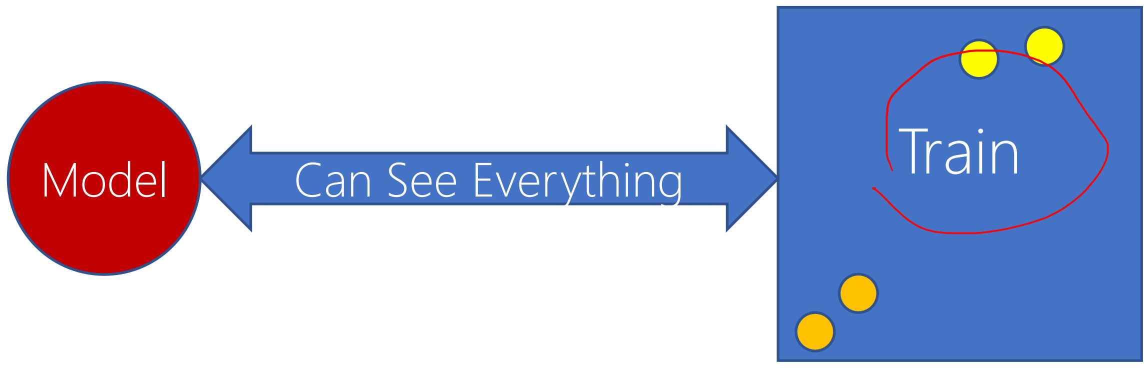
3

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

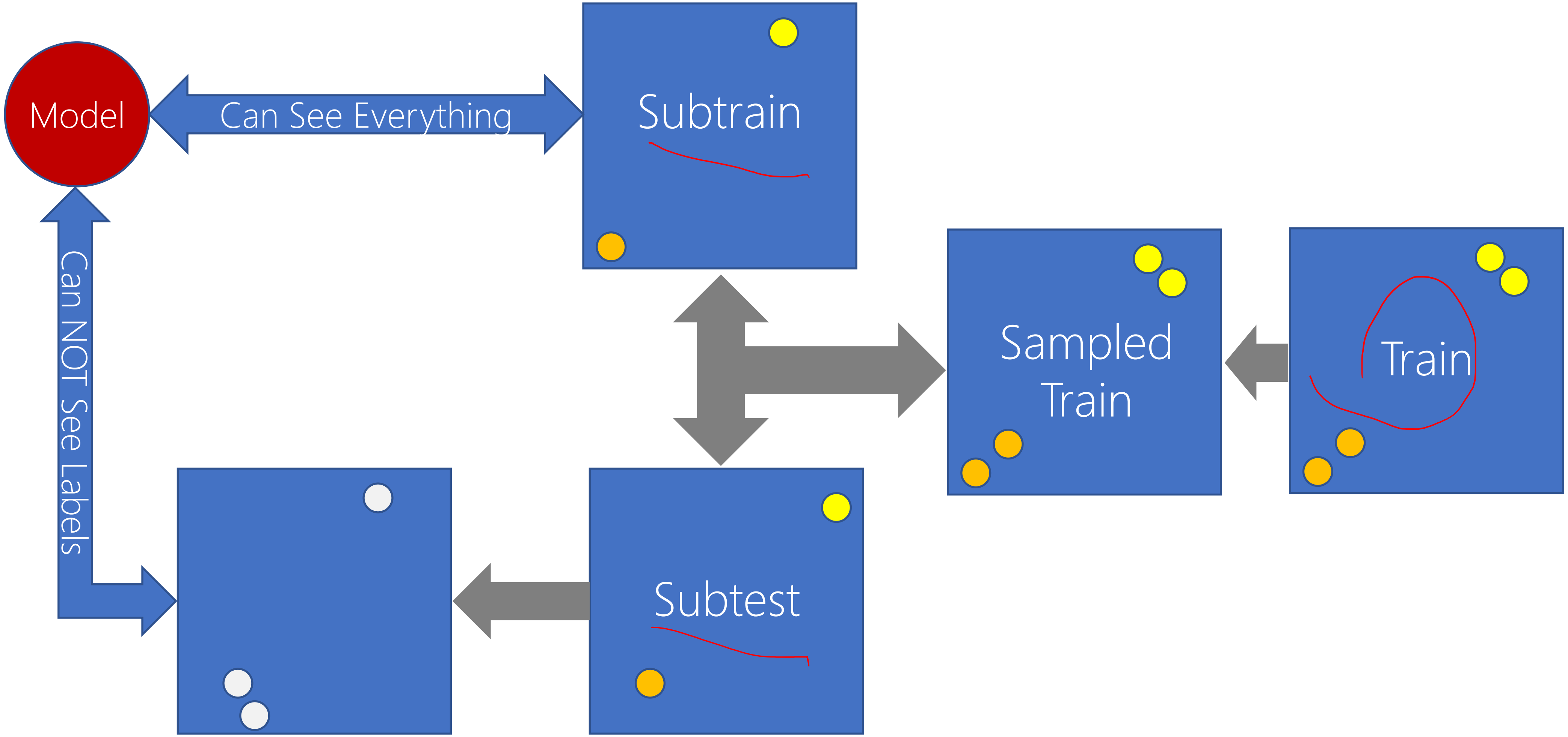
Labeled Data = {Train} U {Test}



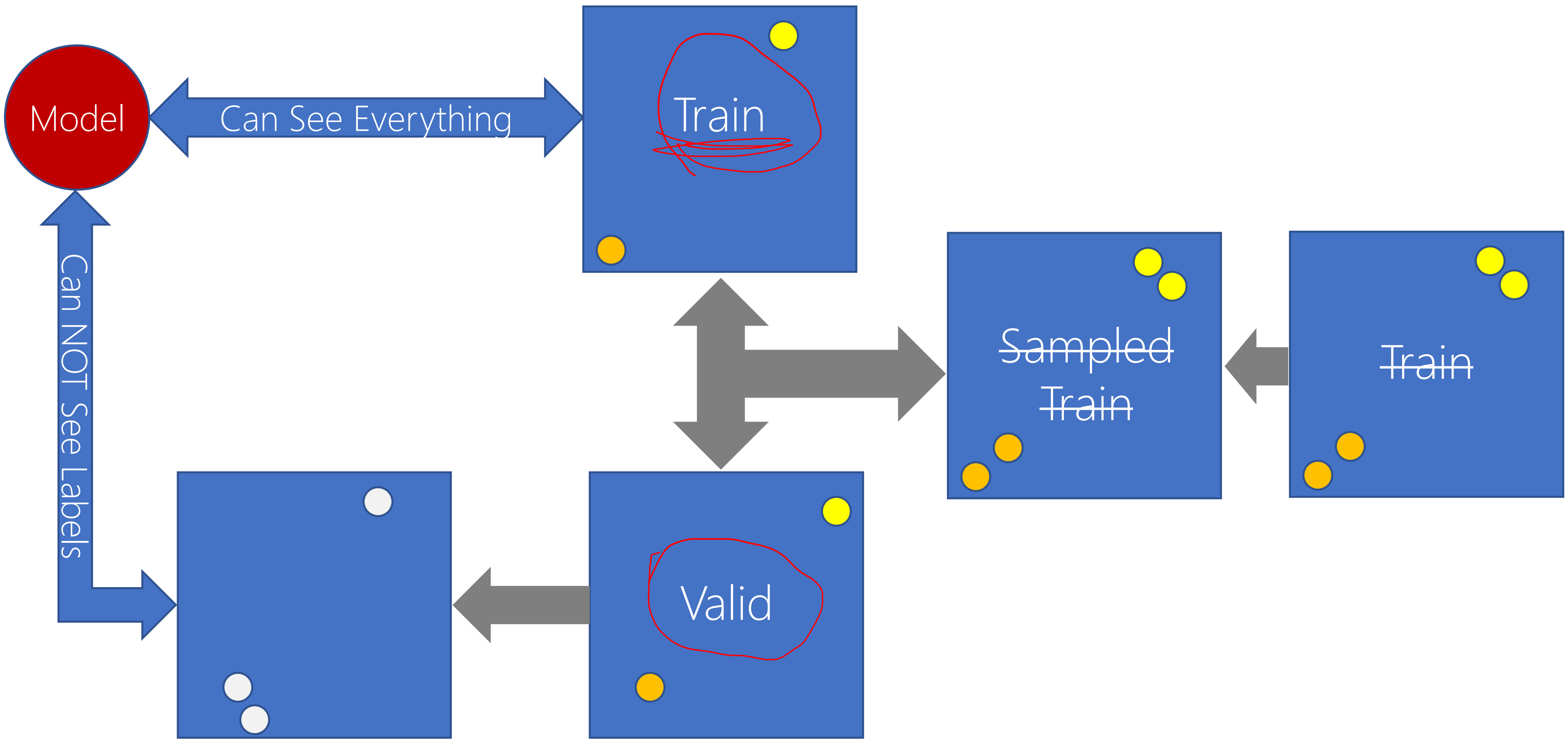
Labeled Data = {Train}



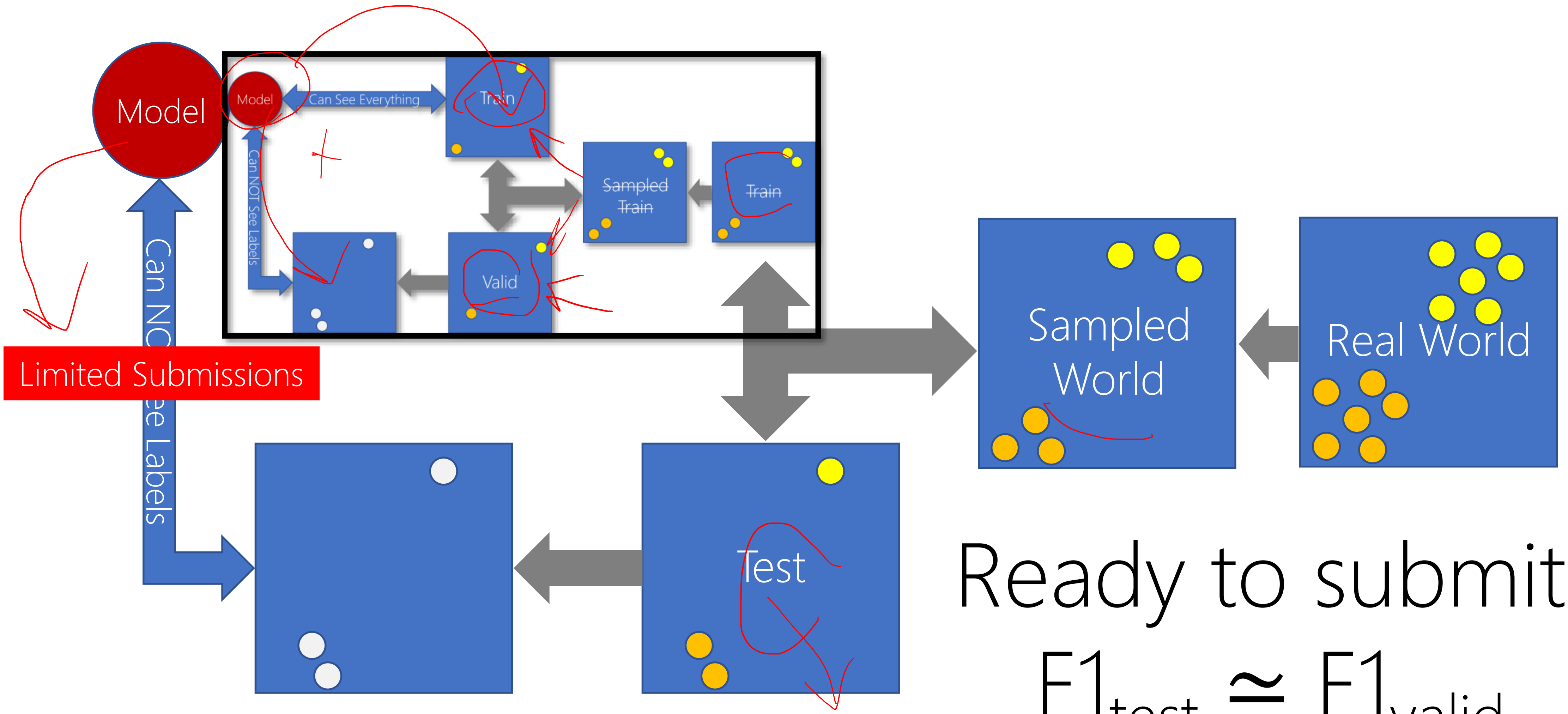
Labeled Data = {{Subtrain} U {Subtest}}



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



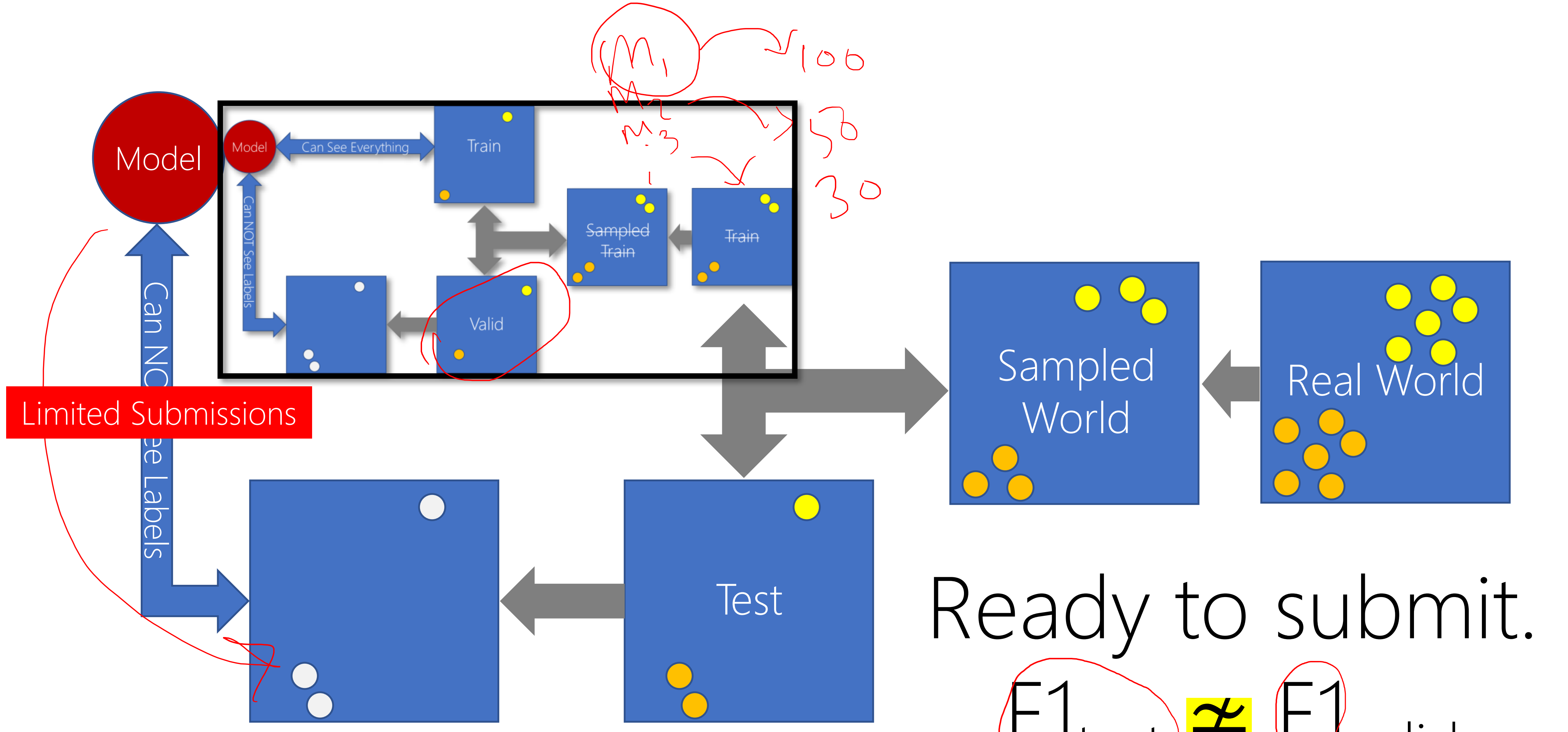
Labeled Data = {Train} U {Test}



Ready to submit.

$F1_{test} \simeq F1_{valid}$

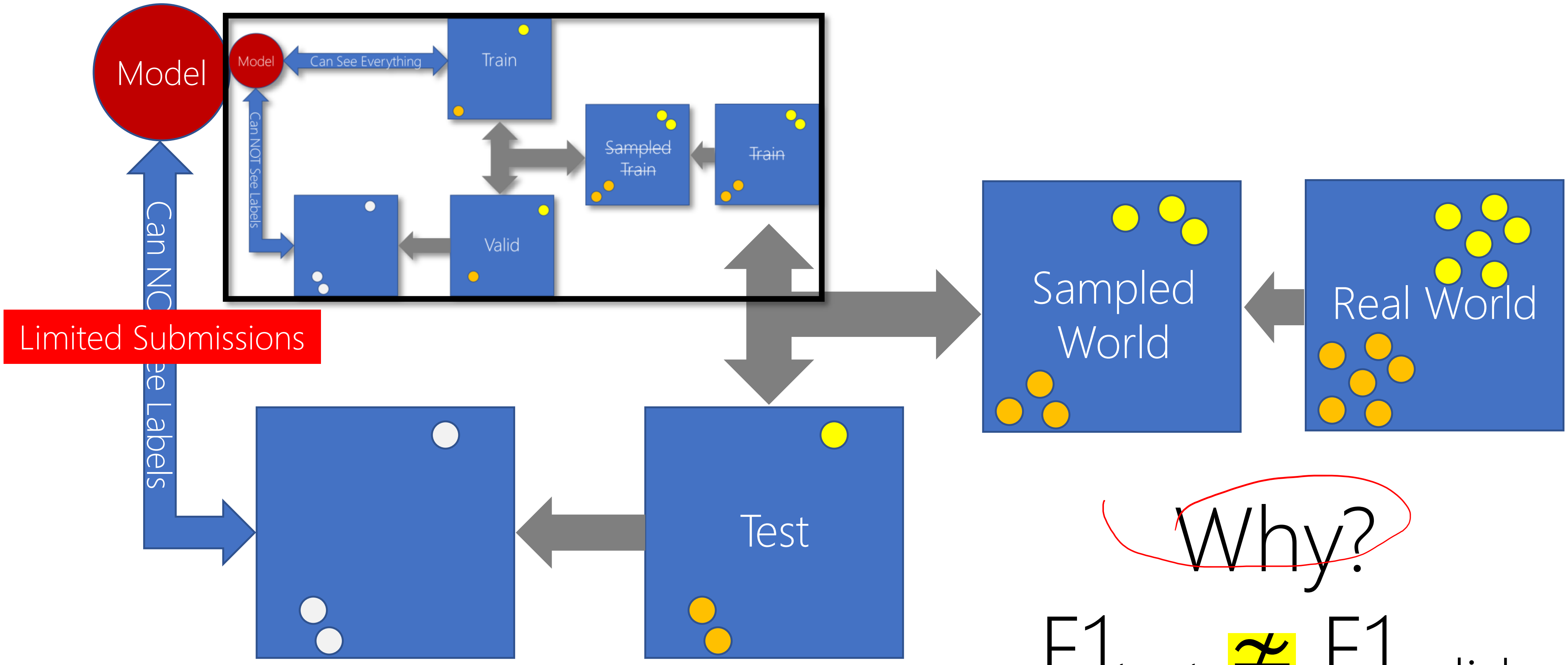
Labeled Data = {Train} U {Test}



Ready to submit.

$F1_{test} \neq F1_{valid}$

Labeled Data = {Train} U {Test}

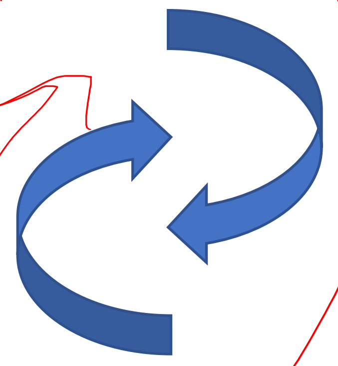


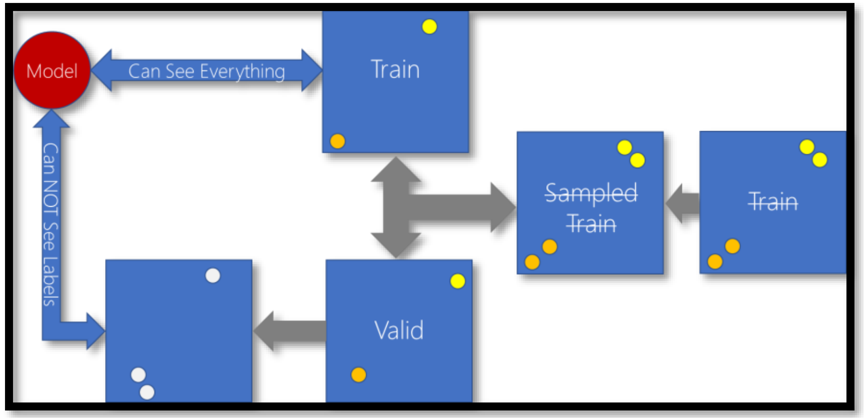
Why?

$F1_{test} \neq F1_{valid}$

Cross-Validation

1 practice vs. Multiple practice





Cross-Validation

n folds

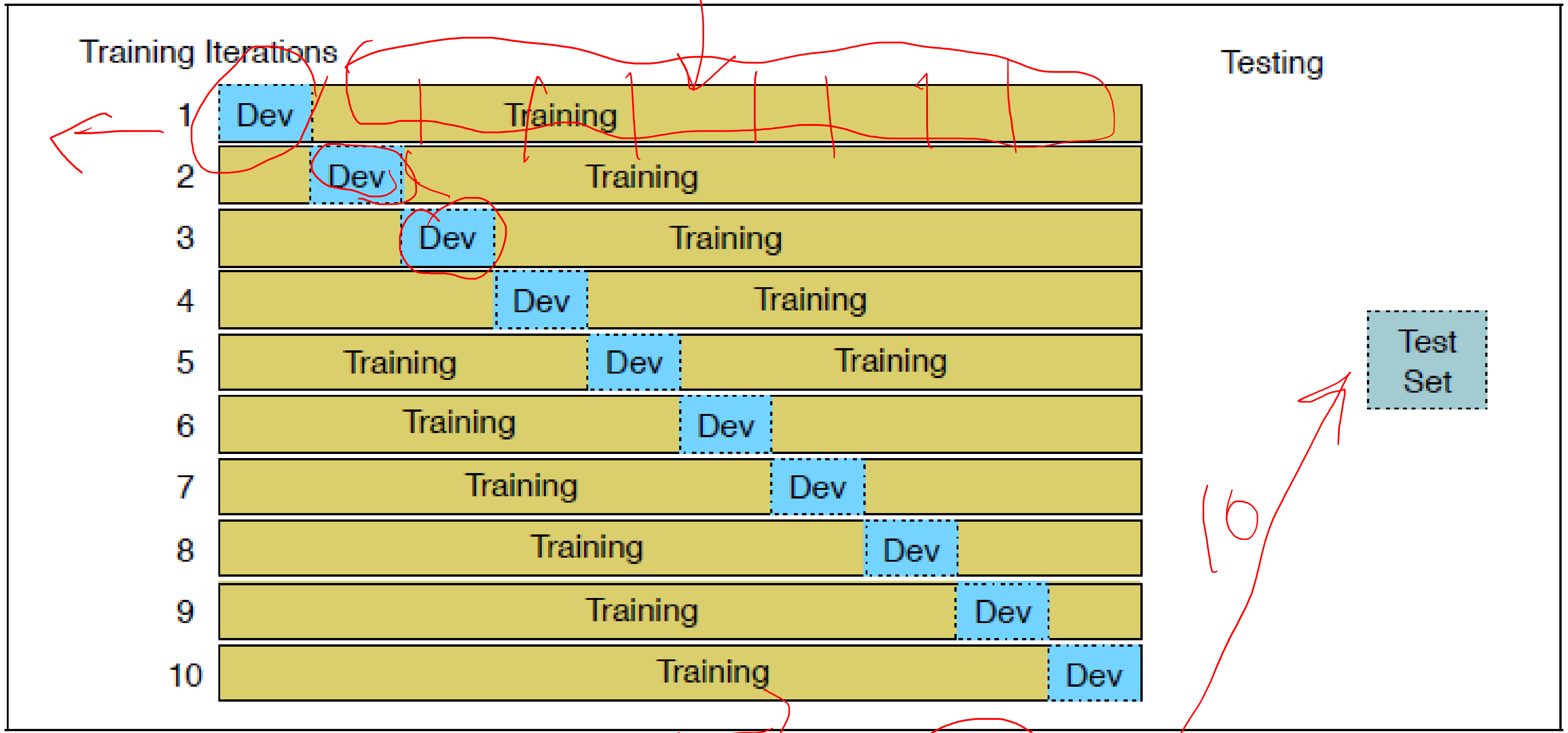


Figure 4.7 10-fold cross-validation

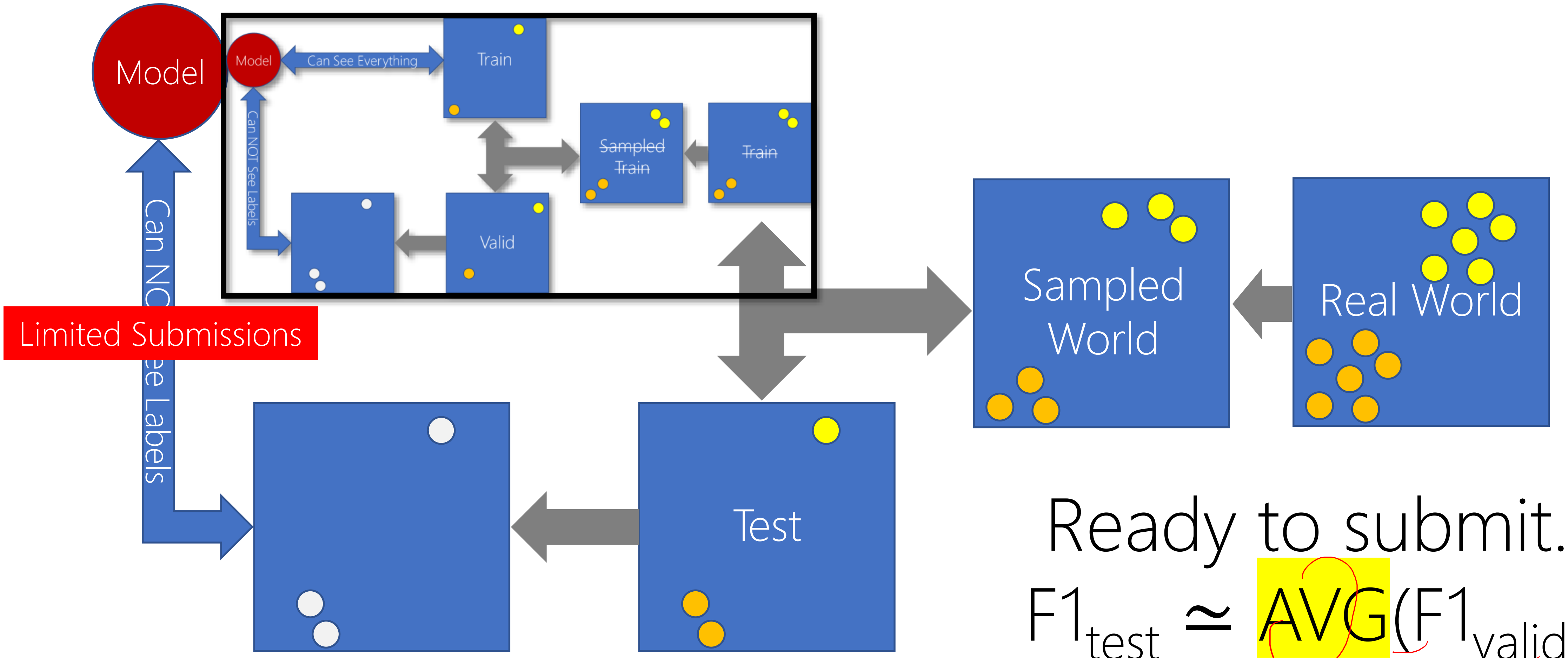
$m_1 f_1$
$m_1 f_2$
$m_1 f_3$

$m_2 f_1$
$m_2 f_2$

10

m_1 m_2

Labeled Data = {Train} U {Test}



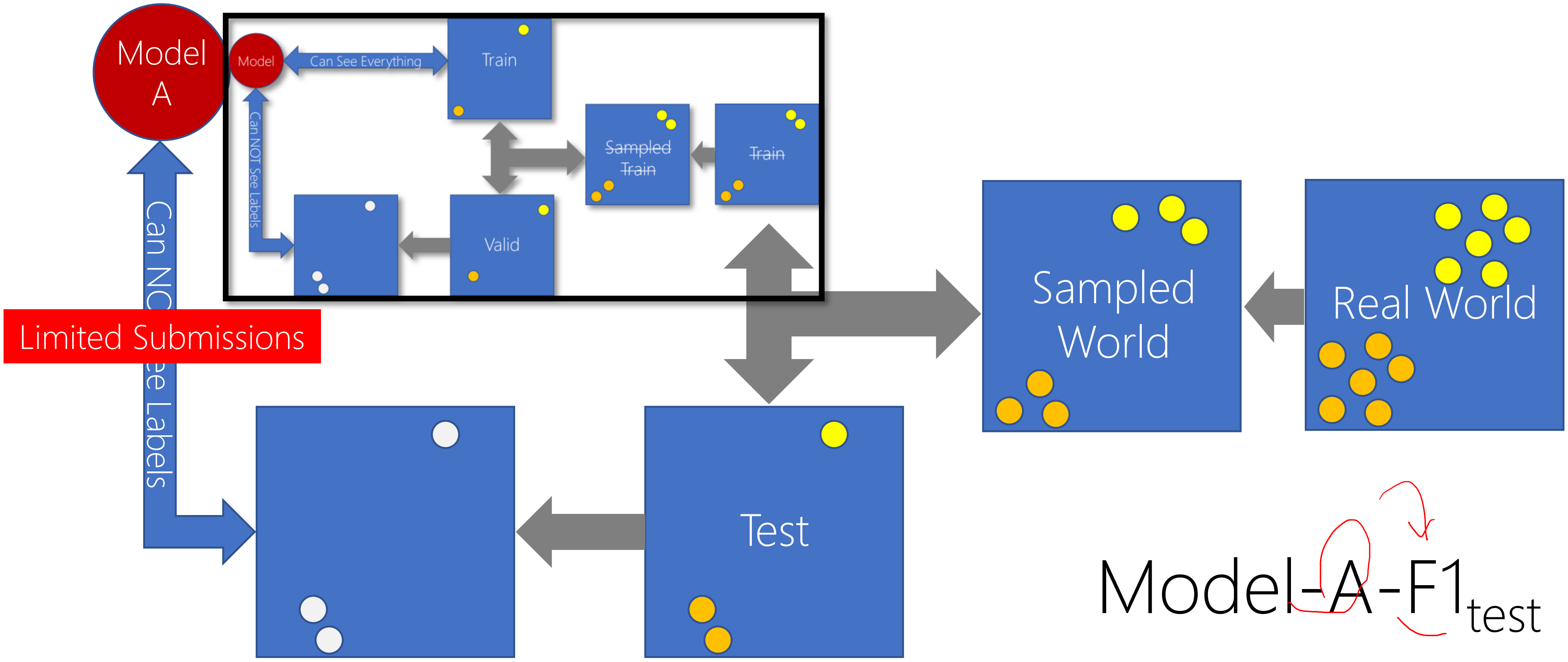
Ready to submit.

$F1_{test} \simeq AVG(F1_{valid})$

Statistical Significance Testing

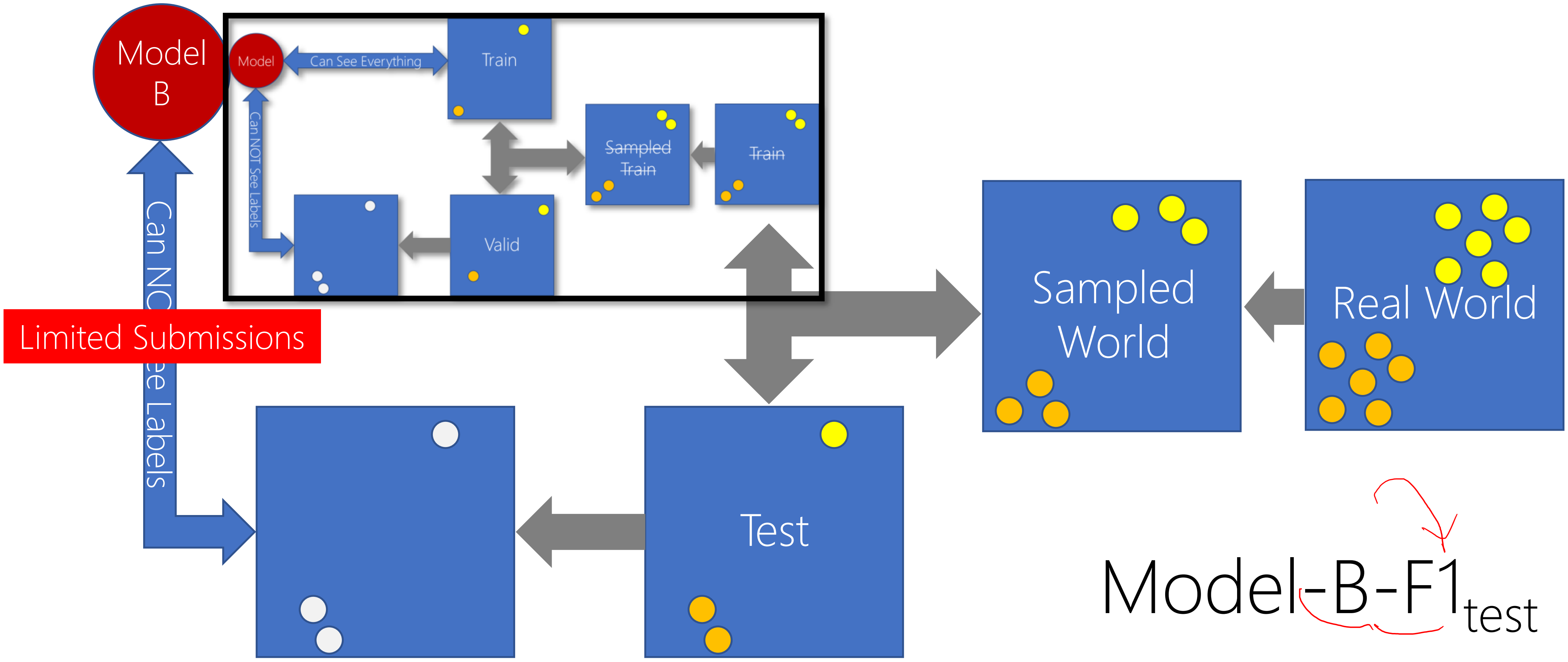
Model Comparison

Statistical Significance Testing



Model-A-F1$_{test}$

Statistical Significance Testing



$\text{Model-B-F1}_{\text{test}}$

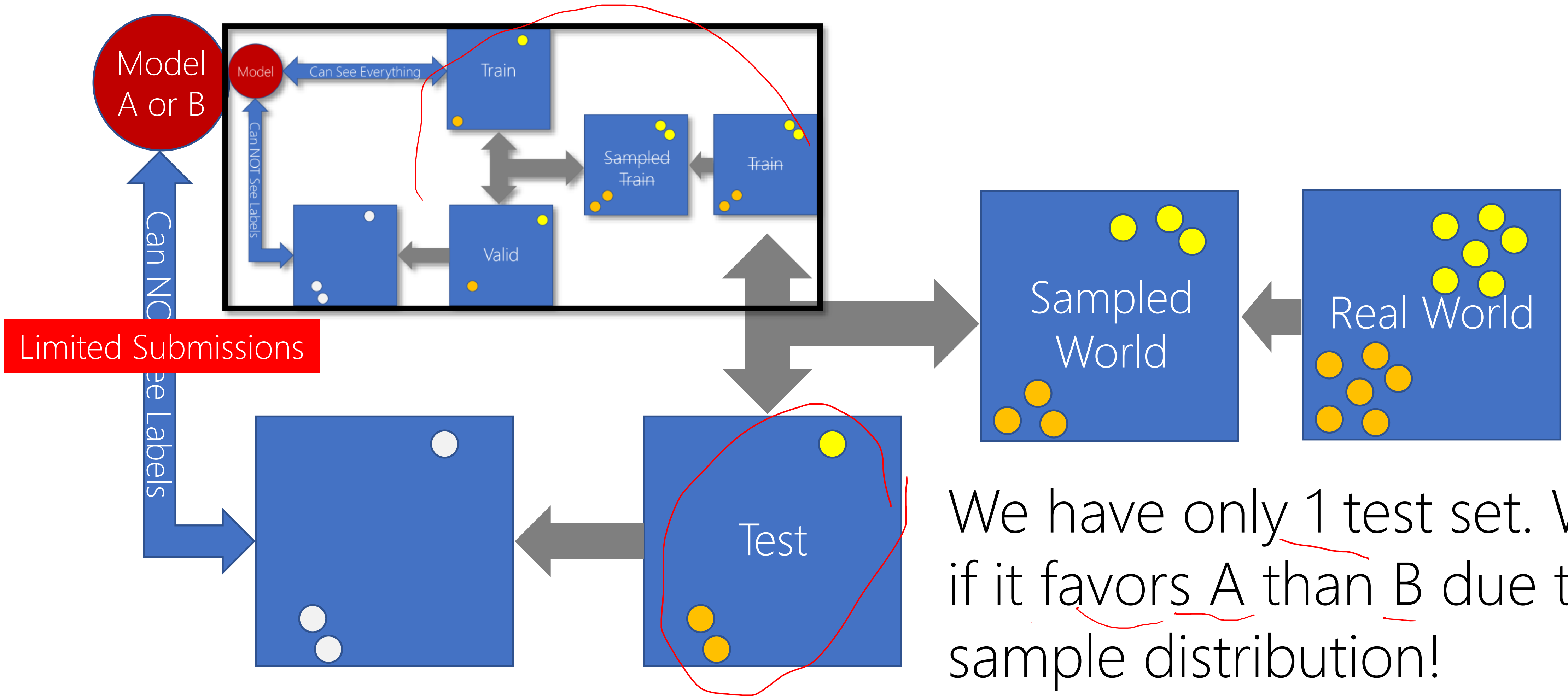
Statistical Significance Testing

Test Sets	A-F1	E.g.,	B-F1	E.g.,
Test-1	A-F1$_{test-1}$	0.99	B-F1$_{test-1}$	0.6

A is better than B.

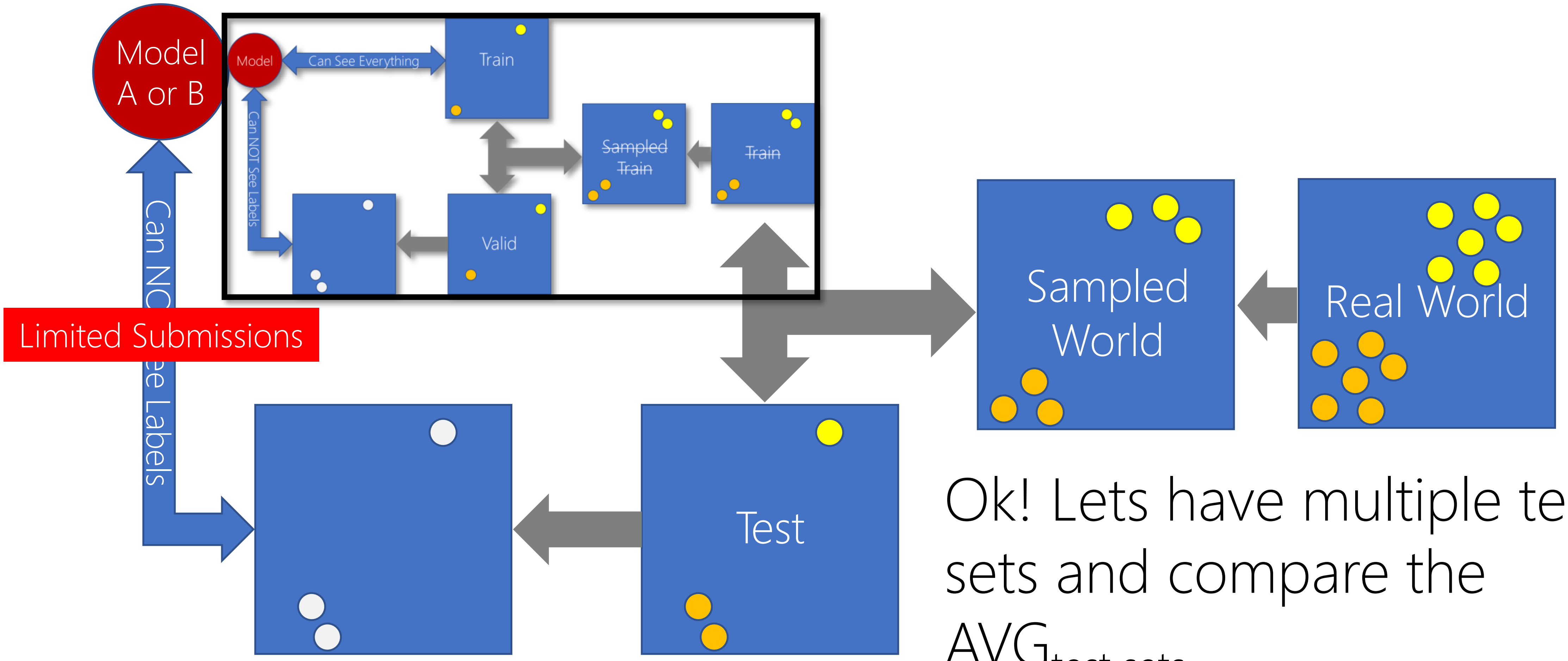
A is significantly better than B → 0.99 >> 0.6

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 $AVG_{\text{test sets}}$

Statistical Significance Testing

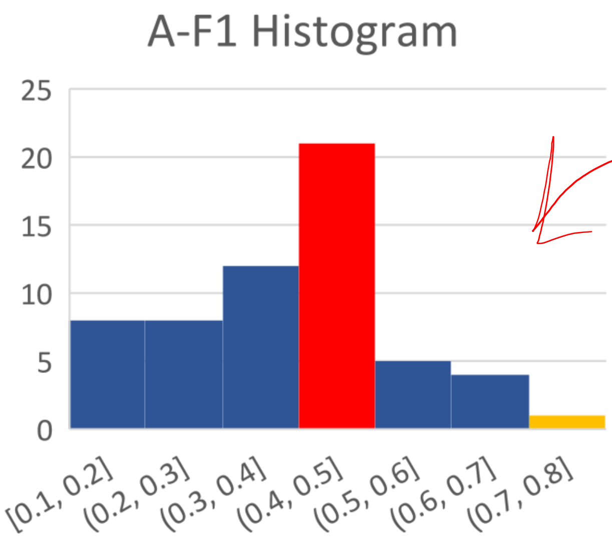
Test Sets	A-F1	E.g.,	B-F1	E.g.,
Test-1	A-F1$_{test-1}$	0.99	B-F1$_{test-1}$	0.6
Test-2	A-F1$_{test-2}$	0.5	B-F1$_{test-2}$	0.6
Test-3	A-F1$_{test-3}$	0.5	B-F1$_{test-3}$	0.6
Test-4	A-F1$_{test-4}$	0.5	B-F1$_{test-4}$	0.6
AVG		0.6225		0.6

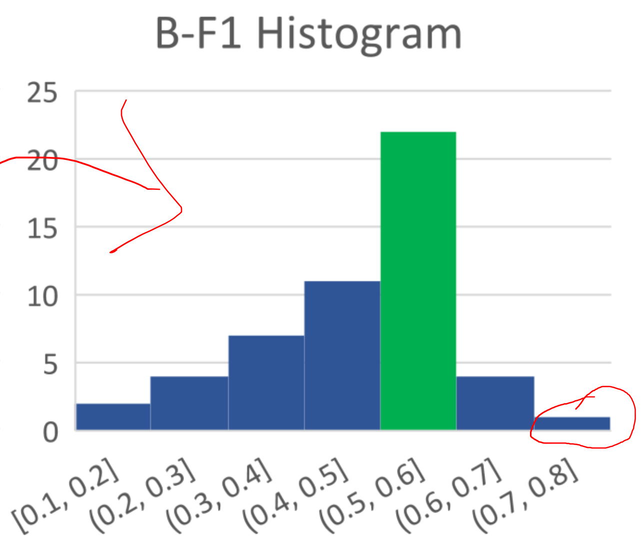
1) By average, A is better than B. However, clearly B is better than A.
2) By average, A is better than B but only slightly NOT significantly!

What is the problem here?

Statistical Significance Testing

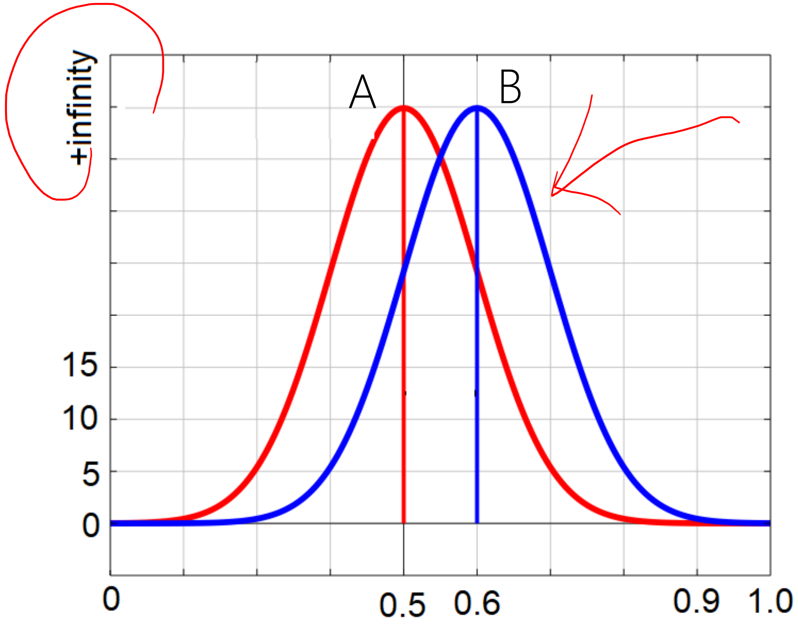
Test Sets	A-F1	E.g.,	B-F1	E.g.,
Test-1	A-F1$_{test-1}$	0.99	B-F1$_{test-1}$	0.6
Test-2	A-F1$_{test-2}$	0.5	B-F1$_{test-2}$	0.6
Test-3	A-F1$_{test-3}$	0.5	B-F1$_{test-3}$	0.6
Test-4	A-F1$_{test-4}$	0.5	B-F1$_{test-4}$	0.6
AVG		0.6225		0.6





Statistical Significance Testing

Test Sets	A-F1	E.g.,	B-F1	E.g.,
Test-1	$\text{A-F1}_{\text{test-1}}$	0.99	$\text{B-F1}_{\text{test-1}}$	0.6
Test-2	$\text{A-F1}_{\text{test-2}}$	0.5	$\text{B-F1}_{\text{test-2}}$	0.6
Test-3	$\text{A-F1}_{\text{test-3}}$	0.5	$\text{B-F1}_{\text{test-3}}$	0.6
Test-4	$\text{A-F1}_{\text{test-4}}$	0.5	$\text{B-F1}_{\text{test-4}}$	0.6
AVG		0.6225		0.6



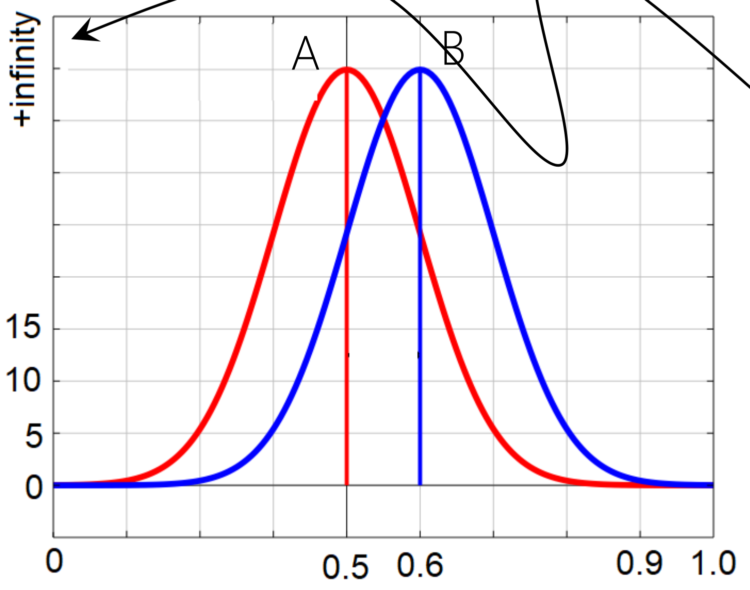
$$P(a<X<b) = \int_a^b f(x)dx.$$

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

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

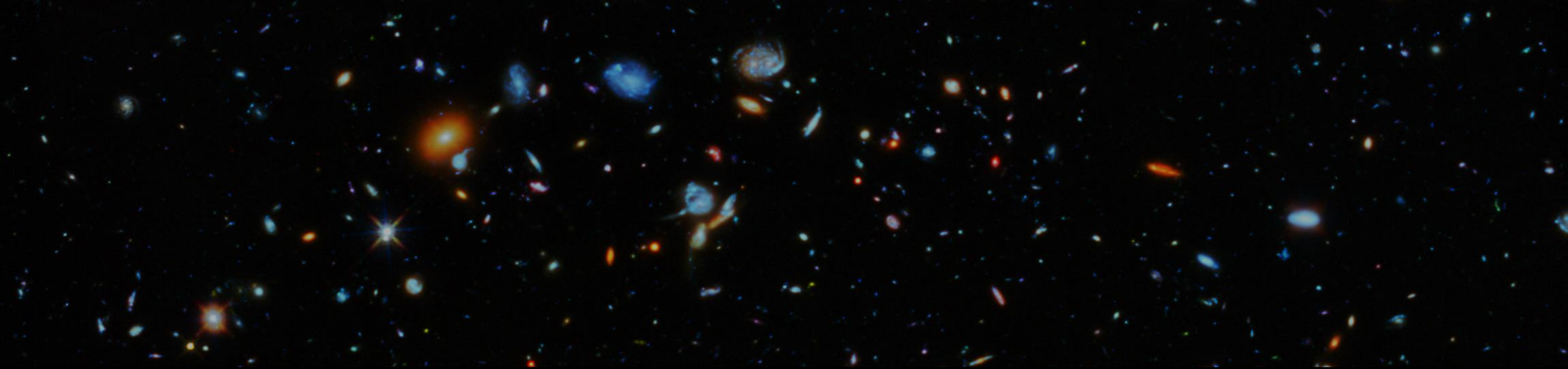
Statistical Significance Testing

Test Sets	A-F1	E.g.,	B-F1	E.g.,
Test-1	A-$F1_{test-1}$	0.99	B-$F1_{test-1}$	0.6
Test-2	A-$F1_{test-2}$	0.5	B-$F1_{test-2}$	0.6
Test-3	A-$F1_{test-3}$	0.5	B-$F1_{test-3}$	0.6
Test-4	A-$F1_{test-4}$	0.5	B-$F1_{test-4}$	0.6
AVG		0.6225		0.6



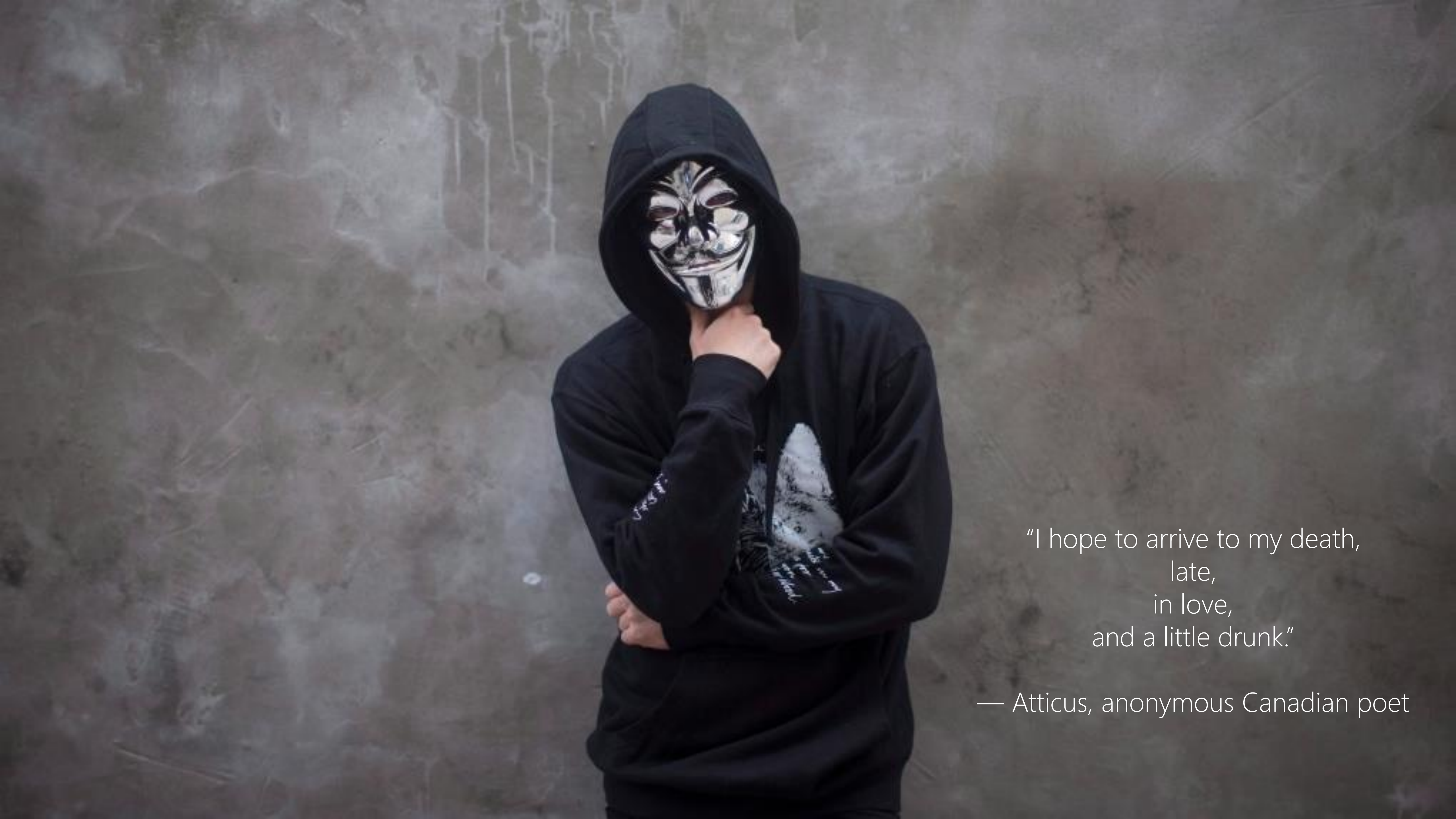
1) Labeled data is already expensive.
2) Sometimes testing is slow.

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



Statistical Significance Test

t-test


"I hope to arrive to my death,
late,
in love,
and a little drunk."
— Atticus, anonymous Canadian poet