Natural Language Processing

Lecture 16

Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

Bocconi

Examples: Sentiment

- Input: reviews
- Output: positive, negative, neutral
- Use: business intelligence, market analysis

Examples: Hate Speech

- Input: tweets, blog posts, etc.
- Output: presence (binary) or type of hate speech
- Use: platform health, public policy

Examples: Mental Health

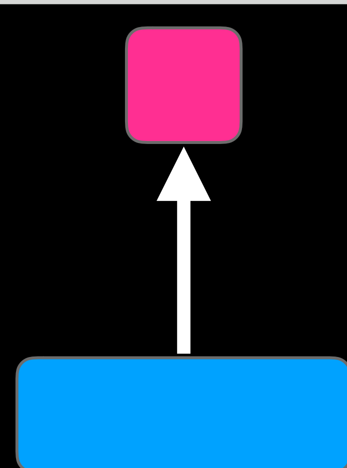
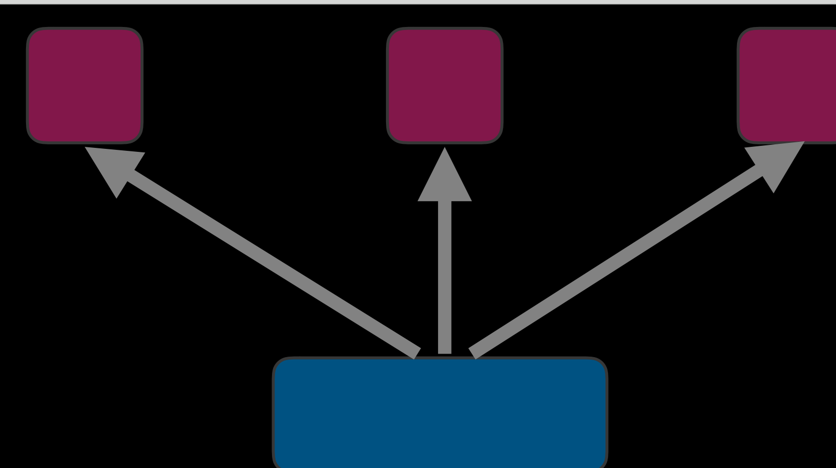
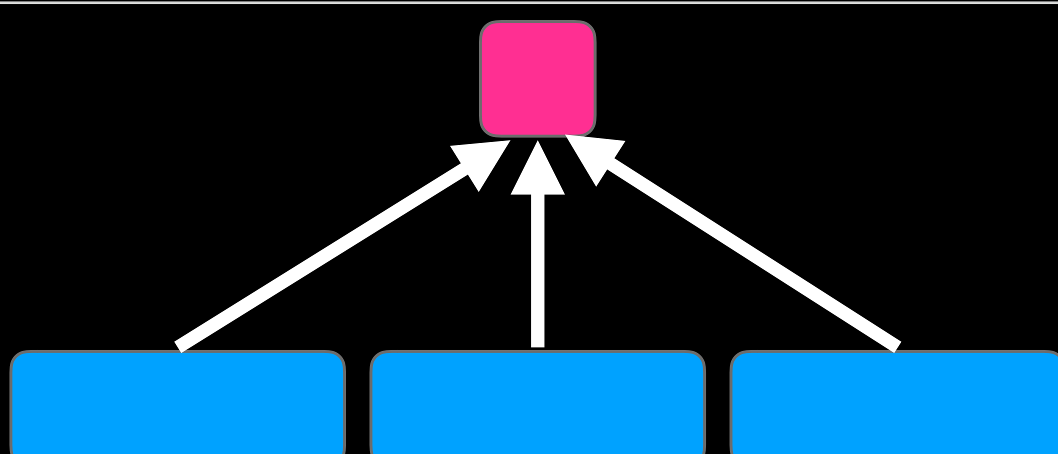
- Input: social media
- Output: presence of risk for mental health condition
- Use: psychologist support, risk screening

Examples: Geolocation

Author Attribute Prediction

- Input: tweet history
- Output: coordinates or predefined region
- Use: social media analysis, targeting

Types of Text Classification

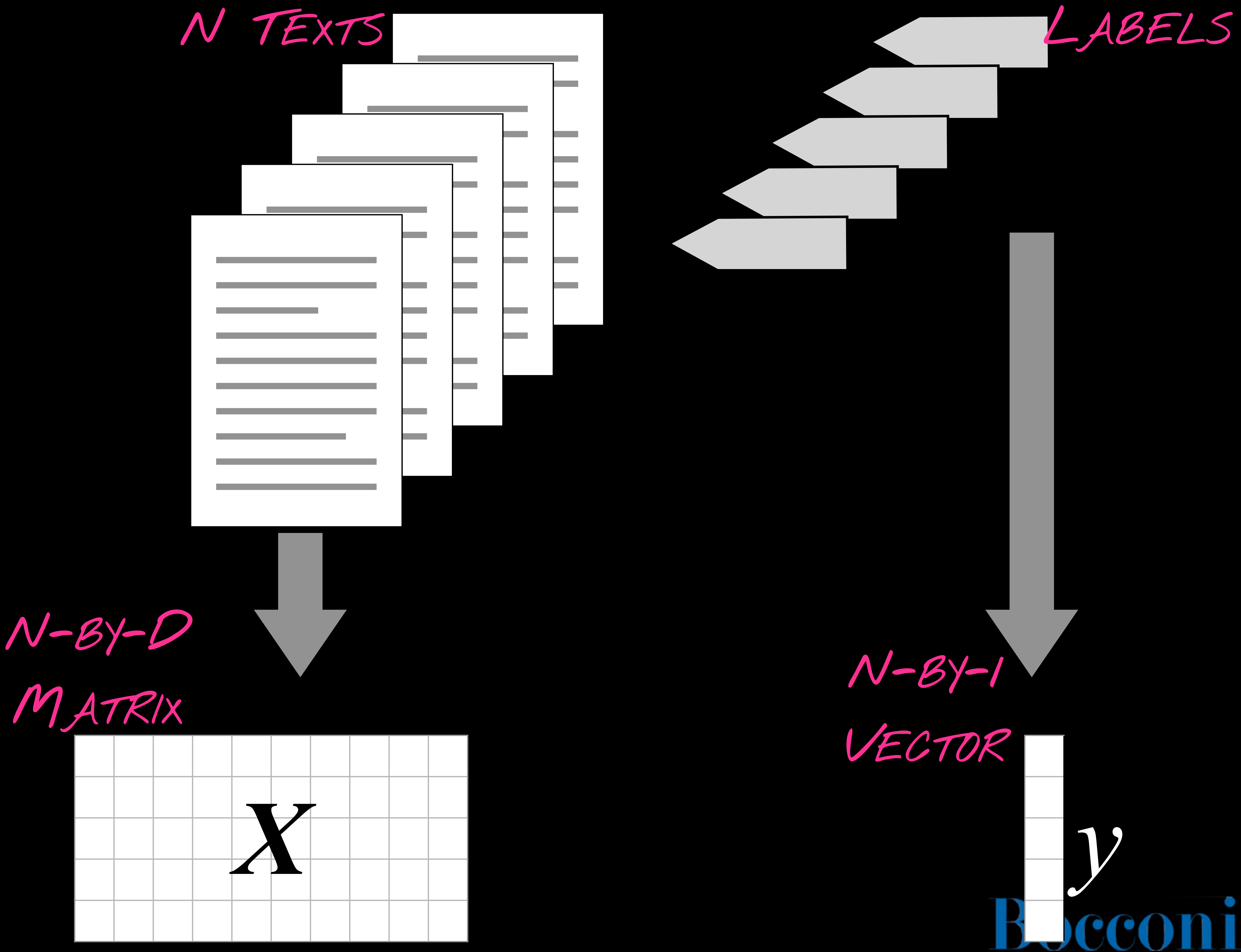
	Fixed length output	Variable length output
Fixed length	 Logistic Regression, Perceptron, Feed-Forward Network, Random Forest, Naive Bayes, SVM, …	 STRUCTURED Multitask Learning, Decoder
Variable length	 Convolutional Neural Networks (CNN)	PREDICTION  Recurrent Neural Networks (RNN), Hidden Markov Models (HMM), Conditional Random Fields

Goals for Today

- Understand how to robustly **evaluate** results
- Learn how to **improve** performance

Text Classification

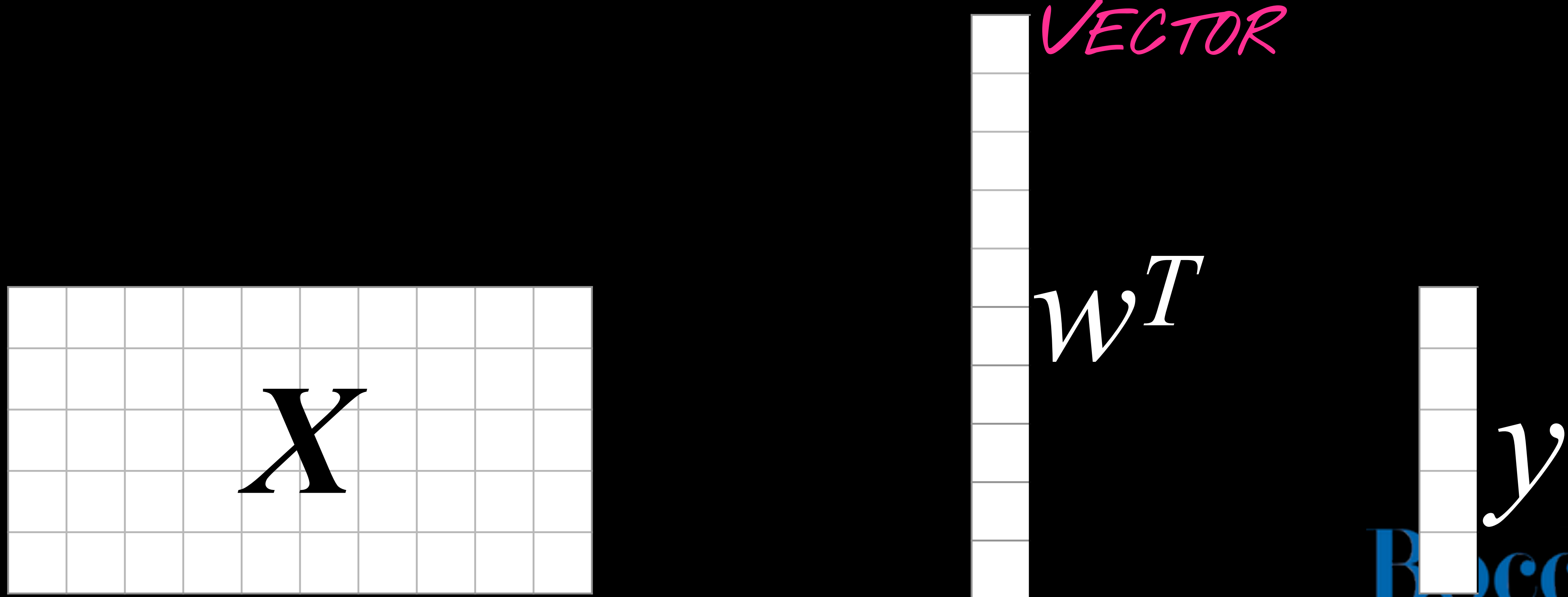
Bocconi


N TEXTS
LABELS
N-BY-D
MATRIX
X
N-BY-1
VECTOR
y
Bocconi

Fitting

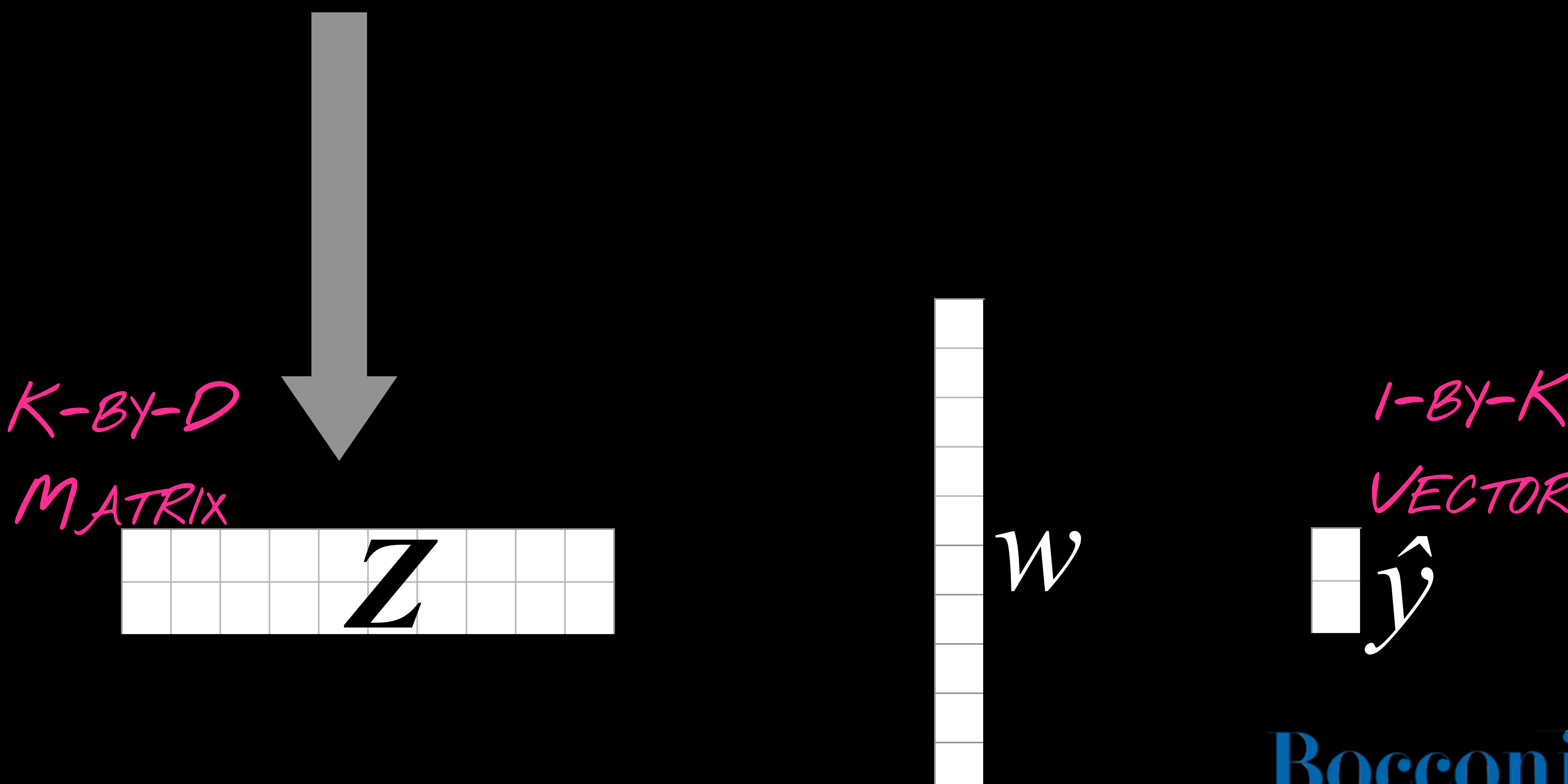
$$f(\boldsymbol{X}) = y$$

D-BY-1 VECTOR



Predicting

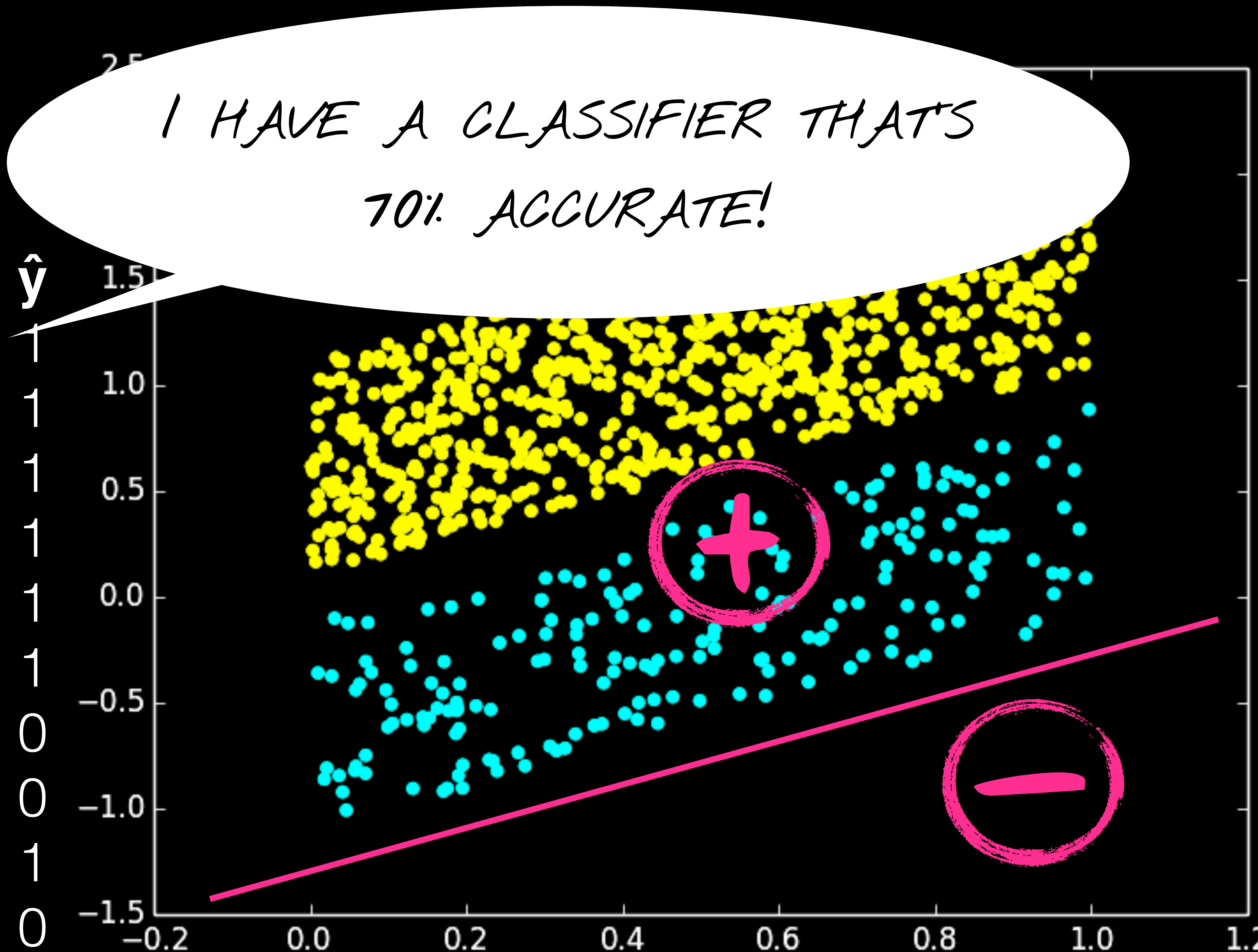
$$f(\mathbf{Z}) = \mathbf{Z}\, w^T = \hat{y}$$



Evaluating Performance

Performance Problems

x	y	ŷ
frog	1	1
deer	1	1
wolf	1	1
dog	1	1
bear	1	1
fish	1	1
bird	1	0
cat	1	0
stone	0	1
tree	0	0



A 70% ACCURATE CLASSIFIER

True and False

	predicted	
gold	1	0
1	TP	FN
0	FP	TN

TARGET = ANIMAL

x	**y**	**ŷ**	
frog	1	1	true positive
deer	1	1	
wolf	1	1	
dog	1	1	
bear	1	1	
fish	1	1	
bird	1	0	false negative
cat	1	0	
stone	0	1	false positive
tree	0	0	true negative

accuracy = (TP+TN) / (P + N)

precision = TP / (TP + FP)

recall = TP / (TP + FN)

F1 = 2 (prec x rec) / (prec + rec)

ACCURACY = 7/10 = 0.7

PRECISION = 6/7 = 0.86

RECALL = 6/8 = 0.75

F1 = 0.81

Bocconi

	predicted		
gold		1	0
	1	TP	FN
	0	FP	TN

Changing Target

accuracy = (TP+TN) / (P + N)

precision = TP / (TP + FP)

recall = TP / (TP + FN)

F1 = 2 (prec x rec) / (prec + rec)

TARGET = THING

x	y	ŷ	
frog	0	0	true negative
deer	0	0	
wolf	0	0	
dog	0	0	
bear	0	0	
fish	0	0	
bird	0	1	false positive
cat	0	1	
stone	1	0	false negative
tree	1	1	true positive

ACCURACY = 7/10 = 0.7

PRECISION = 1/3 = 0.33

RECALL = 1/2 = 0.5

F1 = 0.4

Bocconi

	predicted	
gold	1	0
1	TP	FN
0	FP	TN

MICRO Averaging

WEIGH BY CLASS SIZE

accuracy = (TP+TN) / (P + N)

precision = TP / (TP + FP)

recall = TP / (TP + FN)

F1 = 2 (prec x rec) / (prec + rec)

ANIMAL			THING		
x	**y**	**ŷ**	**x**	**y**	**ŷ**
frog	1	1	frog	0	0
deer	1	1	deer	0	0
wolf	1	1	wolf	0	0
dog	1	1	dog	0	0
bear	1	1	bear	0	0
fish	1	1	fish	0	0
bird	1	1	bird	0	0
cat	1	0	cat	0	1
stone	0	1	stone	1	0
tree	0	0	tree	1	1

ACC = (7+7)/(10+10) = 14/20 = 0.7

PREC = (6+1)/(7+3) = 7/10 = 0.7

REC = (6+1)/(8+2) = 7/10 = 0.7

F1 = 0.7

MACRO Averaging

WEIGH ALL CLASSES EQUALLY

	predicted	
gold	1	0
1	TP	FN
0	FP	TN

accuracy = (TP+TN) / (P + N)

precision = TP / (TP + FP)

recall = TP / (TP + FN)

F1 = 2 (prec x rec) / (prec + rec)

ANIMAL

x	**y**	**ŷ**
frog	1	1
deer	1	1
wolf	1	1
dog	1	1
bear	1	1
fish	1	1
bird	1	1
cat	1	0
stone	0	1
tree	0	0

THING

x	**y**	**ŷ**
frog	0	0
deer	0	0
wolf	0	0
dog	0	0
bear	0	0
fish	0	0
bird	0	0
cat	0	1
stone	1	0
tree	1	1

ACC = (0.7 + 0.7) / 2 = 0.7

PREC = (0.86 + 0.33) / 2 = 0.6

REC = (0.5 + 0.75) / 2 = 0.63

F1 = 0.61

Bocconi

Metrics Overview

- **accuracy** can be too general
- **precision** and **recall** are per-class measures
- **precision** = how many of instances labeled as target class are actually *in* target class?
- **recall** = how many of *all* target class instances in data identified correctly?
- **F1** = symmetric mean of precision and recall

Bocconi

Baselines

Baseline: Total Recall

	predicted	
gold	1	0
1	TP	FN
0	FP	TN

PREDICT MAJORITY CLASS FOR ALL

accuracy = (TP+TN) / (P + N)

precision = TP / (TP + FP)

recall = TP / (TP + FN)

F1 = 2 (prec x rec) / (prec + rec)

TARGET = ANIMAL

x	**y**	**ŷ**
frog	1	1
deer	1	1
wolf	1	1
dog	1	1
bear	1	1
fish	1	1
bird	1	1
cat	1	1
stone	0	1
tree	0	1

true positive

false positive

ACCURACY = 8/10 = 0.8

PRECISION = 8/10 = 0.8

RECALL = 8/8 = 1.0

F1 = 0.9

Bocconi

The Hulk

(dumb but powerful)

- Character 2–6 grams
- TFIDF weights
- L2-regularized Logistic Regression with balanced classes
- Can be further improved with dimensionality reduction

Regularization

Regularization

$$y = \boldsymbol{X}\, w^T + e$$

D-BY-1

VECTOR

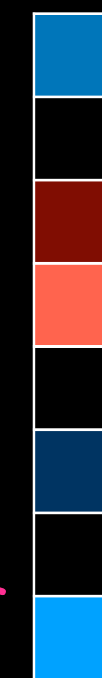
w^T

$$||w||$$

Regularization Norms

L1 NORM

$$||W||_1 = \sum_{i=1}^{\boldsymbol{N}} |w_i|$$



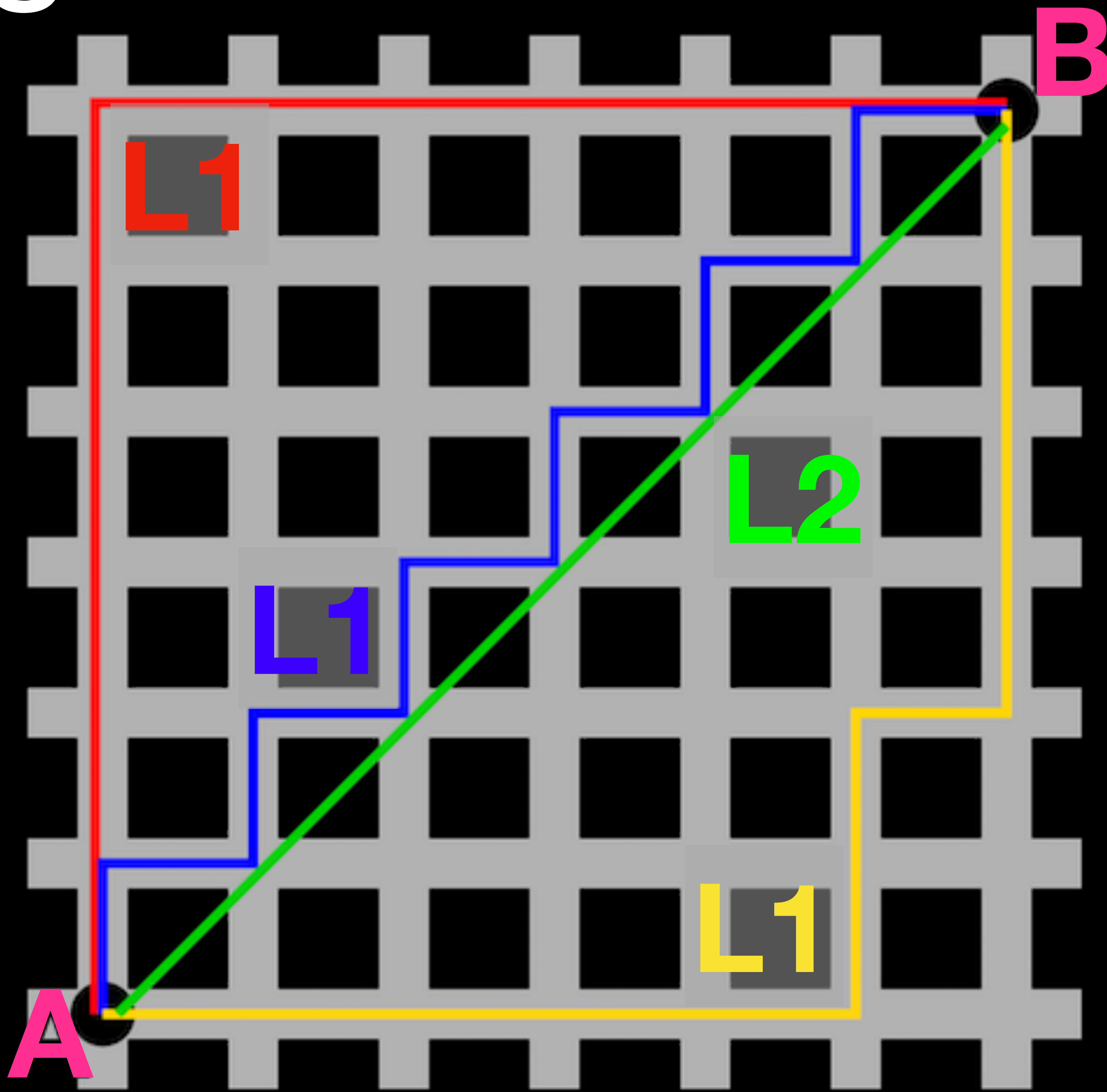
SPARSE

L2 NORM

$$||W||_2 = \sqrt{\sum_{i=1}^{\boldsymbol{N}} w_i^{\boldsymbol{2}}}$$

EVENLY DISTRIBUTED

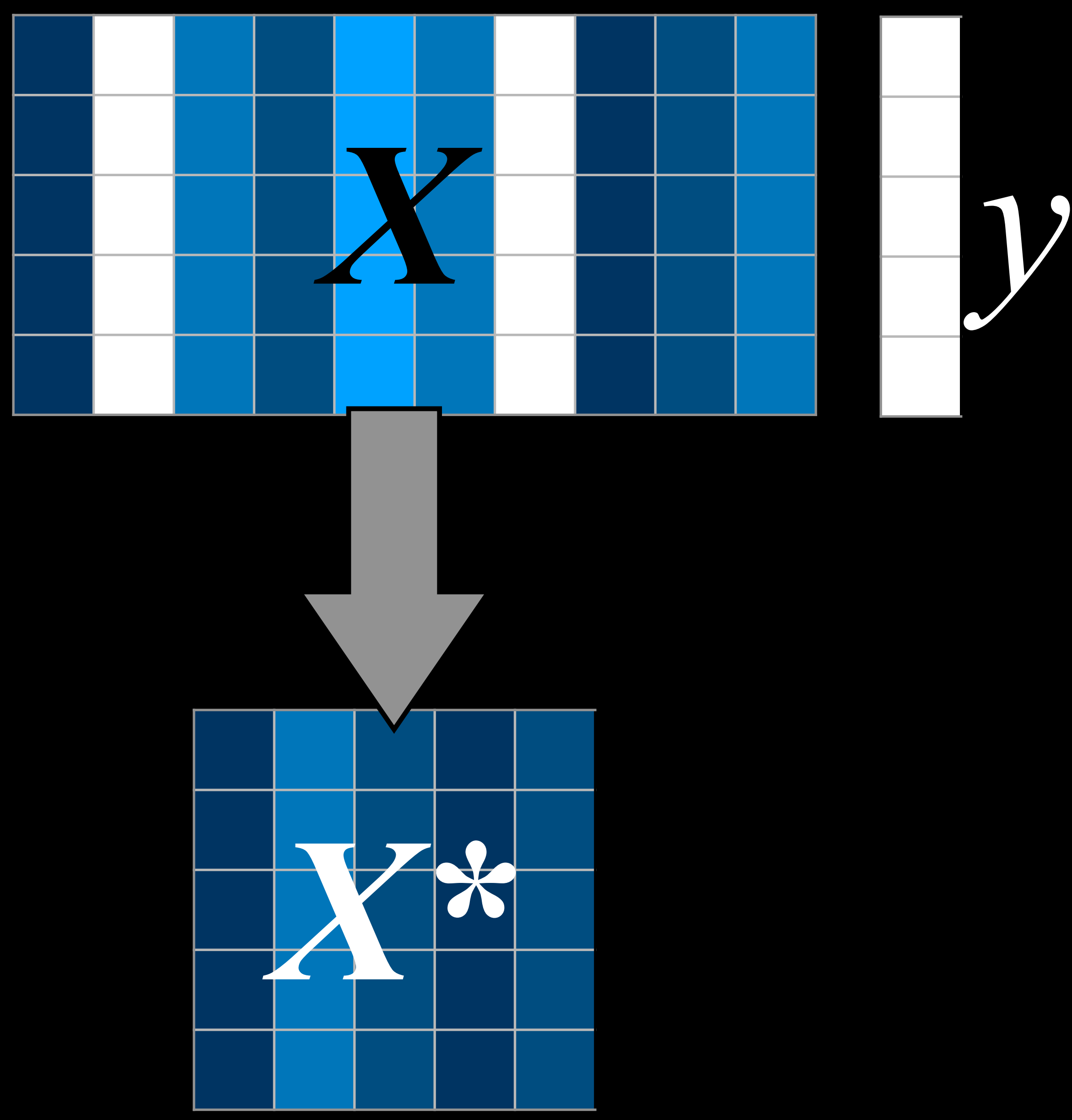
Regularization Norms



Bocconi

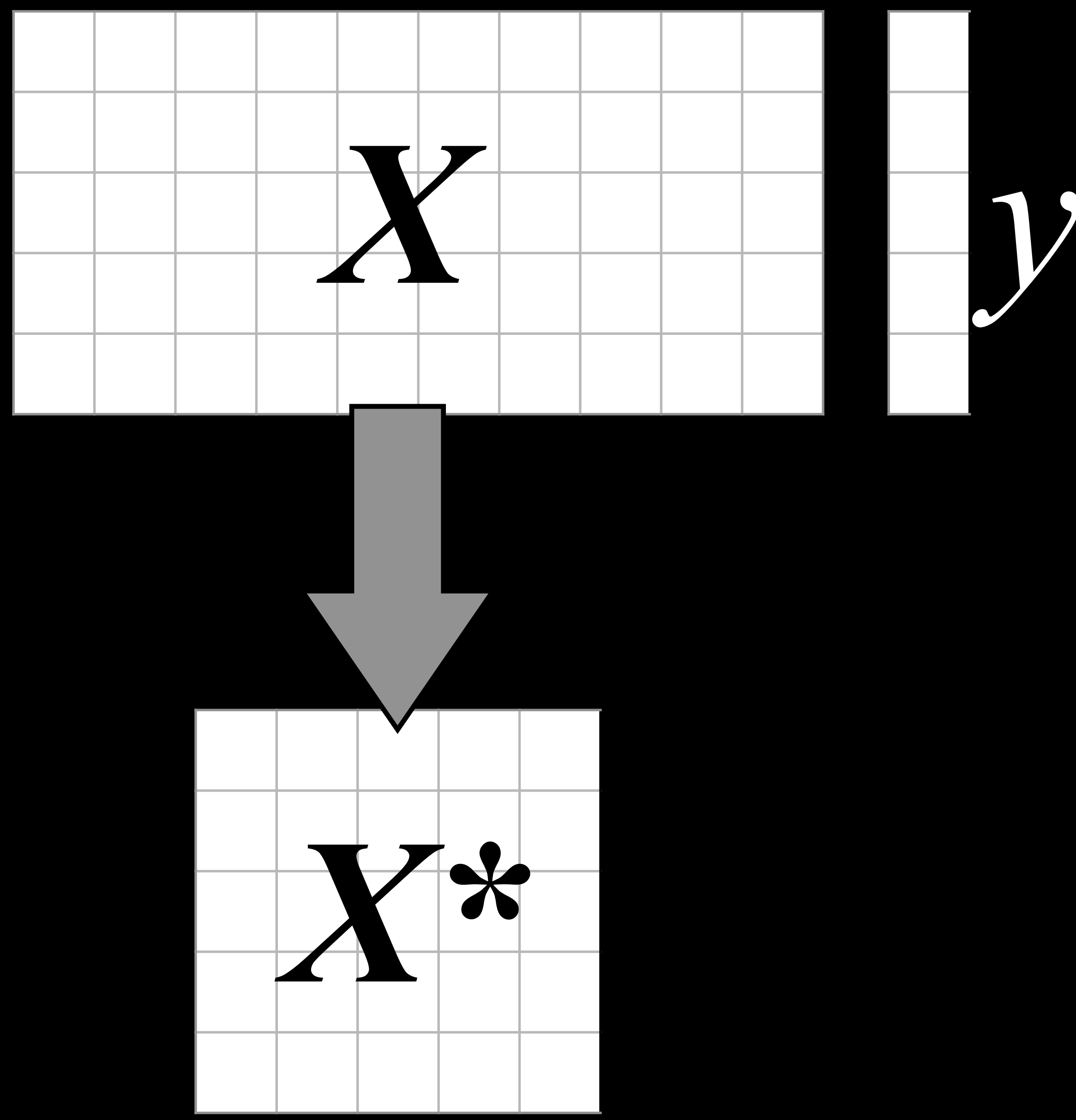
Feature Selection

Chi-Squared Selection



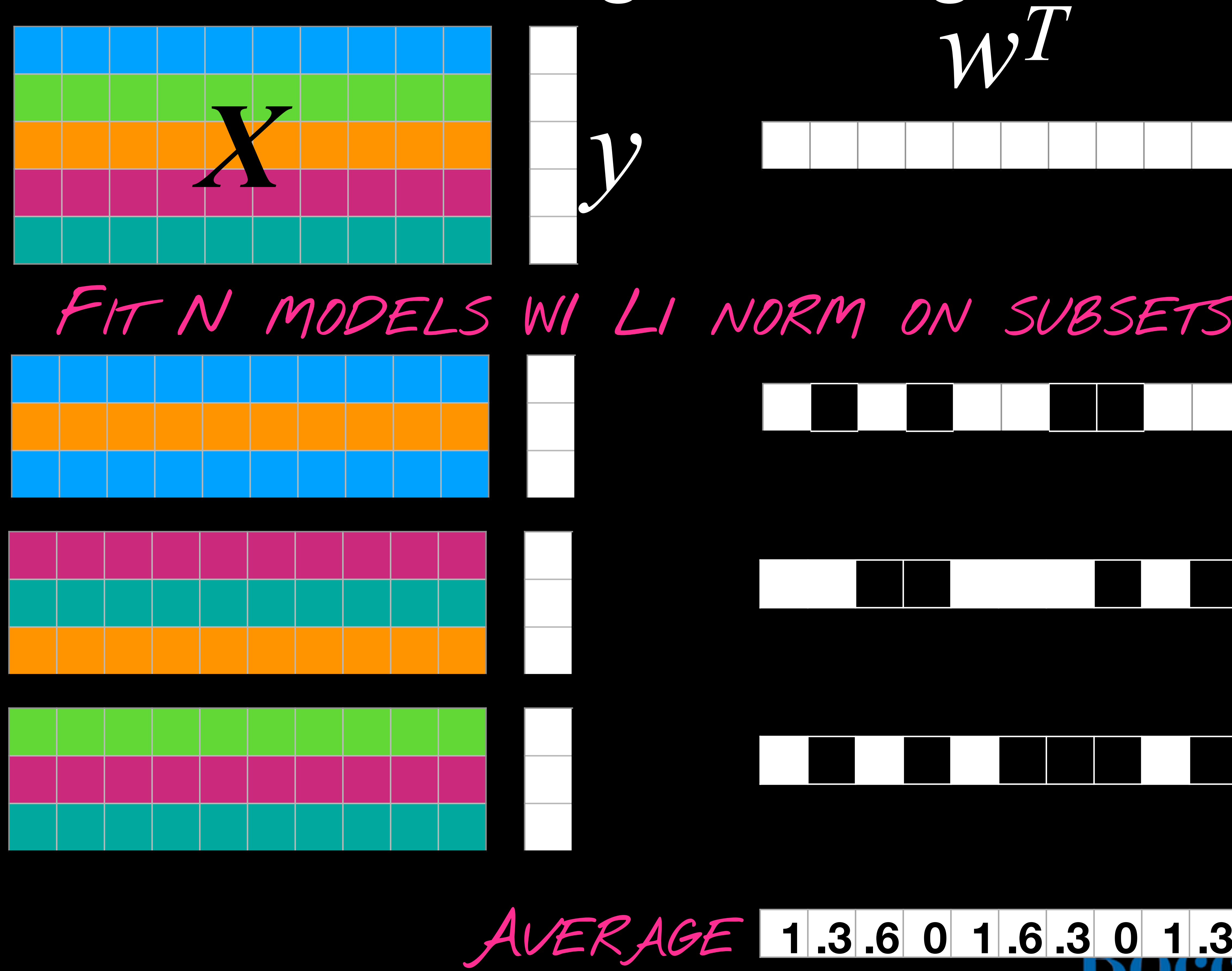
MEASURE CHI2 VALUE (CORRELATION) FOR EACH FEATURE WITH TARGET, SELECT TOP K BY CUTOFF

Dimensionality Reduction



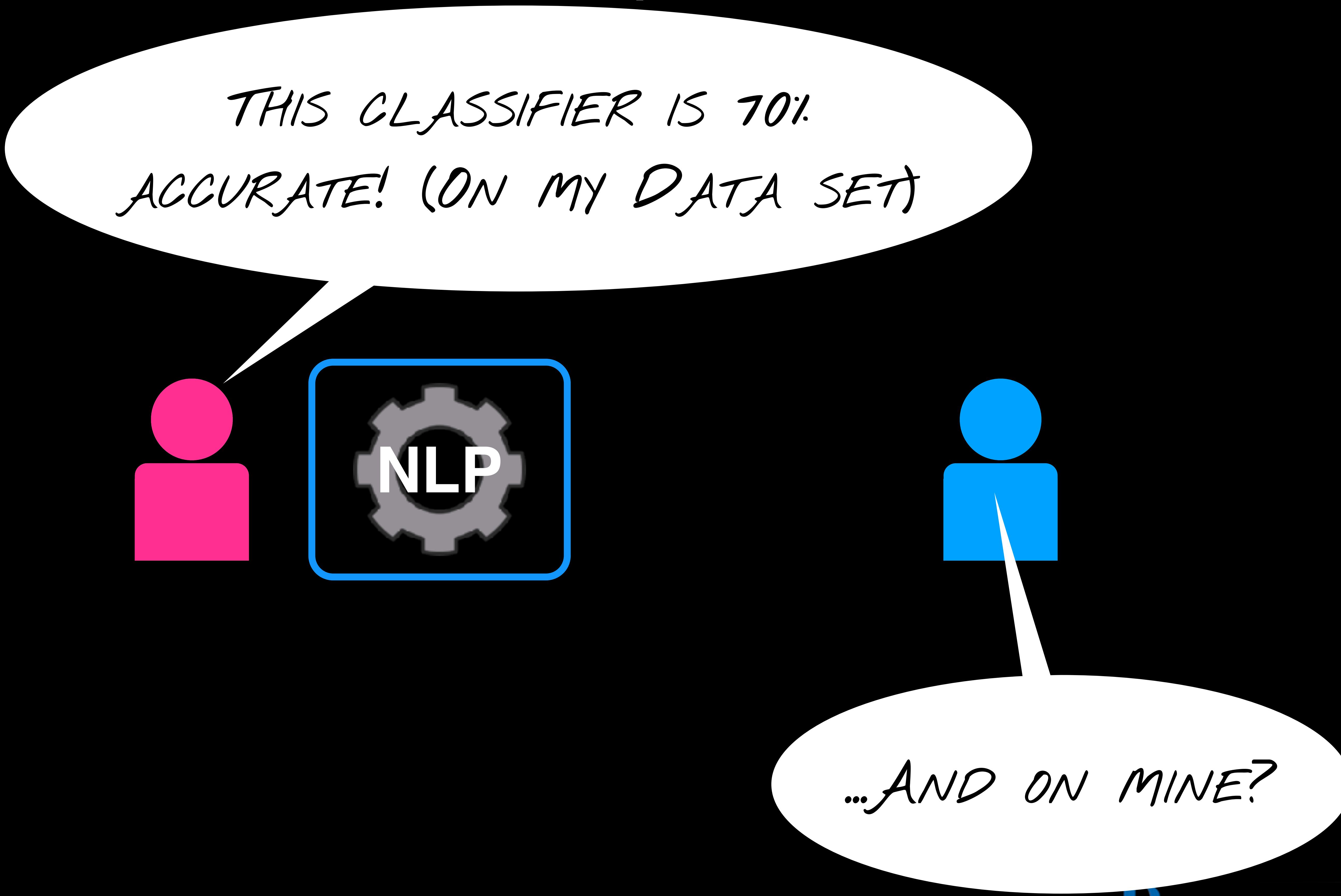
REDUCE DIMENSIONALITY TO PREVENT SPURIOUS CORRELATIONS WITH TARGET, BRING OUT LATENT DIMENSIONS

Randomized Logistic Regression



Significance Testing

What does a *p*-Value Tell Us?



Bootstrap Sampling

y	$\hat{y}1$		$\hat{y}2$	
1	1		1	
1	0	3/5	0	1/5
1	1		0	
0	1		1	
0	0		1	

COMPARE ON SUBSETS

1	1		1	
1	0	1/3	0	1/3
0	1		1	

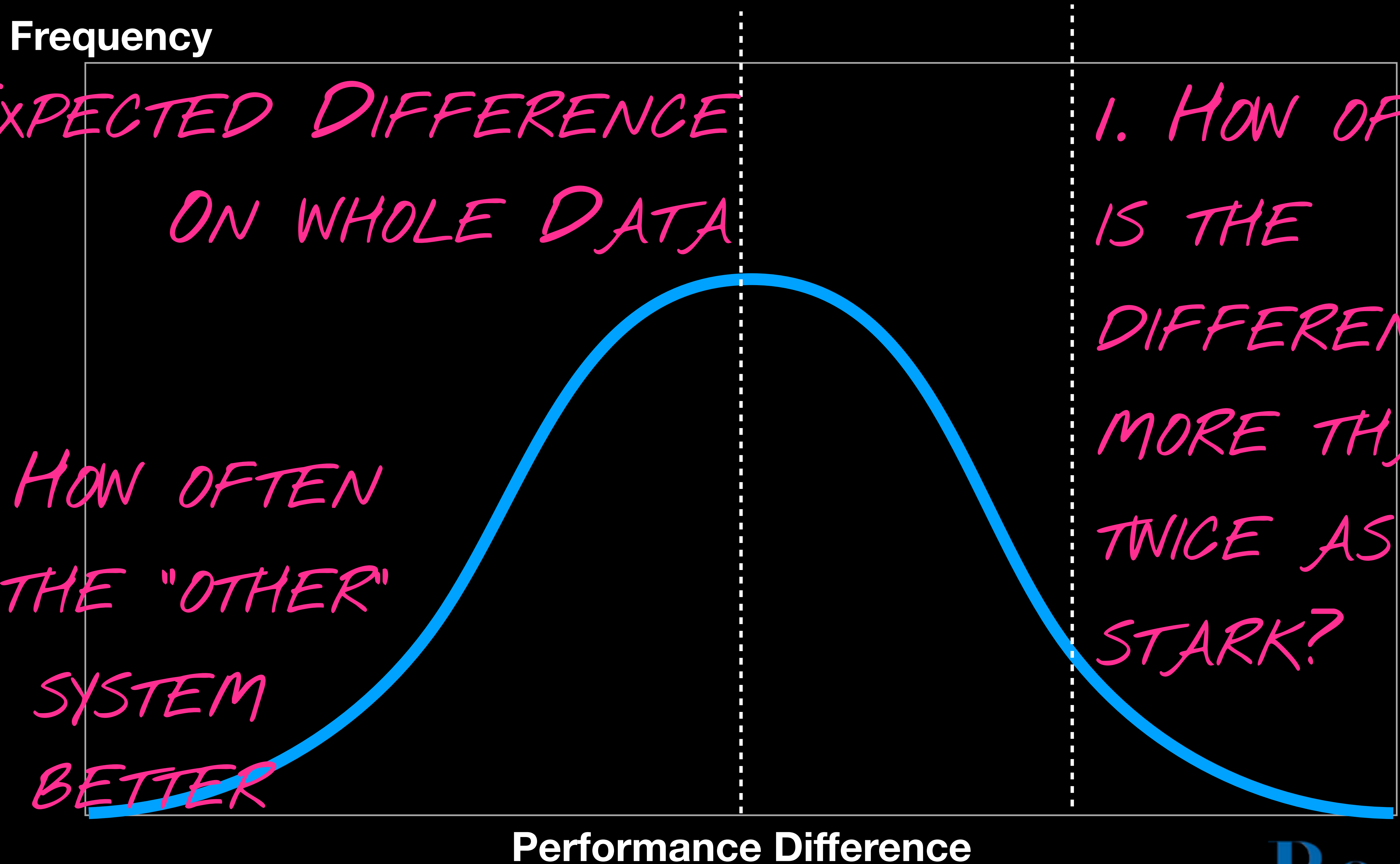
1	0		0	
1	1	1/3	0	0/3
0	0		1	

1	1		1	
1	1	2/3	0	1/3
0	1		1	

Bootstrap Sampling

SAMPLED DIFFERENCES FOLLOW NORMAL DISTRO:

CENTRAL LIMIT THEOREM



EXPECTED DIFFERENCE ON WHOLE DATA

1. HOW OFTEN IS THE DIFFERENCE MORE THAN TWICE AS STARK?

2. HOW OFTEN IS THE "OTHER" SYSTEM BETTER

Bootstrap Sampling

	System 1	System 2	Difference(1-2)
full	**82.13**	**81.89**	**0.24**
1	81.96	82.03	-0.07
2	81.86	82.61	-0.75
3	81.70	81.44	0.26
4	82.42	82.77	-0.35
5	81.89	81.06	0.83
6	81.39	81.24	0.15
7	81.96	81.58	0.37
8	82.57	81.65	0.92
9	82.50	82.67	-0.17
10	83.07	81.84	1.23
		p-value	**0.3**

Note: Significance is Binary!

Cut-offs: 0.1 (meh), 0.05 (standard), 0.01 (strict)

- (barely) not statistically significant (p=0.052)
- a barely detectable statistically significant difference (p=0.073)
- a borderline significant trend (p=0.09)
- a certain trend toward significance (p=0.08)
- a clear tendency to significance (p=0.052)
- a clear trend (p<0.09)
- a clear, strong trend (p=0.09)
- a considerable trend toward significance (p=0.069)
- a decreasing trend (p=0.09)
- a definite trend (p=0.08)
- a distinct trend toward significance (p=0.07)
- \borderline conventional significance (p=0.051)
- borderline level of statistical significance (p=0.053)
- borderline significant (p=0.09)
- did not quite reach conventional levels of statistical significance (p=0.079)
- did not quite reach statistical significance (p=0.063)
- did not reach the traditional level of significance (p=0.10)
- did not reach the usually accepted level of clinical significance (p=0.07)
- difference was apparent (p=0.07)
- direction heading towards significance (p=0.10)
- does not appear to be sufficiently significant (p>0.05)
- does not narrowly reach statistical significance (p=0.06)
- does not reach the conventional significance level (p=0.098)
- effectively significant (p=0.051)
- equivocal significance (p=0.06)
- essentially significant (p=0.10)
- extremely close to significance (p=0.07)
- failed to reach significance on this occasion (p=0.09)
- failed to reach statistical significance (p=0.06)
- fairly close to significance (p=0.065)
- fairly significant (p=0.09)
- falls just short of standard levels of statistical significance (p=0.06)
- fell (just) short of significance (p=0.08)
- fell barely short of significance (p=0.08)
- scarcely significant (0.05<p>0.1)
- significant at the .07 level
- significant tendency (p=0.09)
- significant to some degree (0<p>1)
- significant, or close to significant effects (p=0.08, p=0.05)
- significantly better overall (p=0.051)
- significantly significant (p=0.065)
- similar but not nonsignificant trends (p>0.05)
- slight evidence of significance (0.1>p>0.05)
- slight non-significance (p=0.06)
- slight significance (p=0.128)

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

Evaluation Don'ts

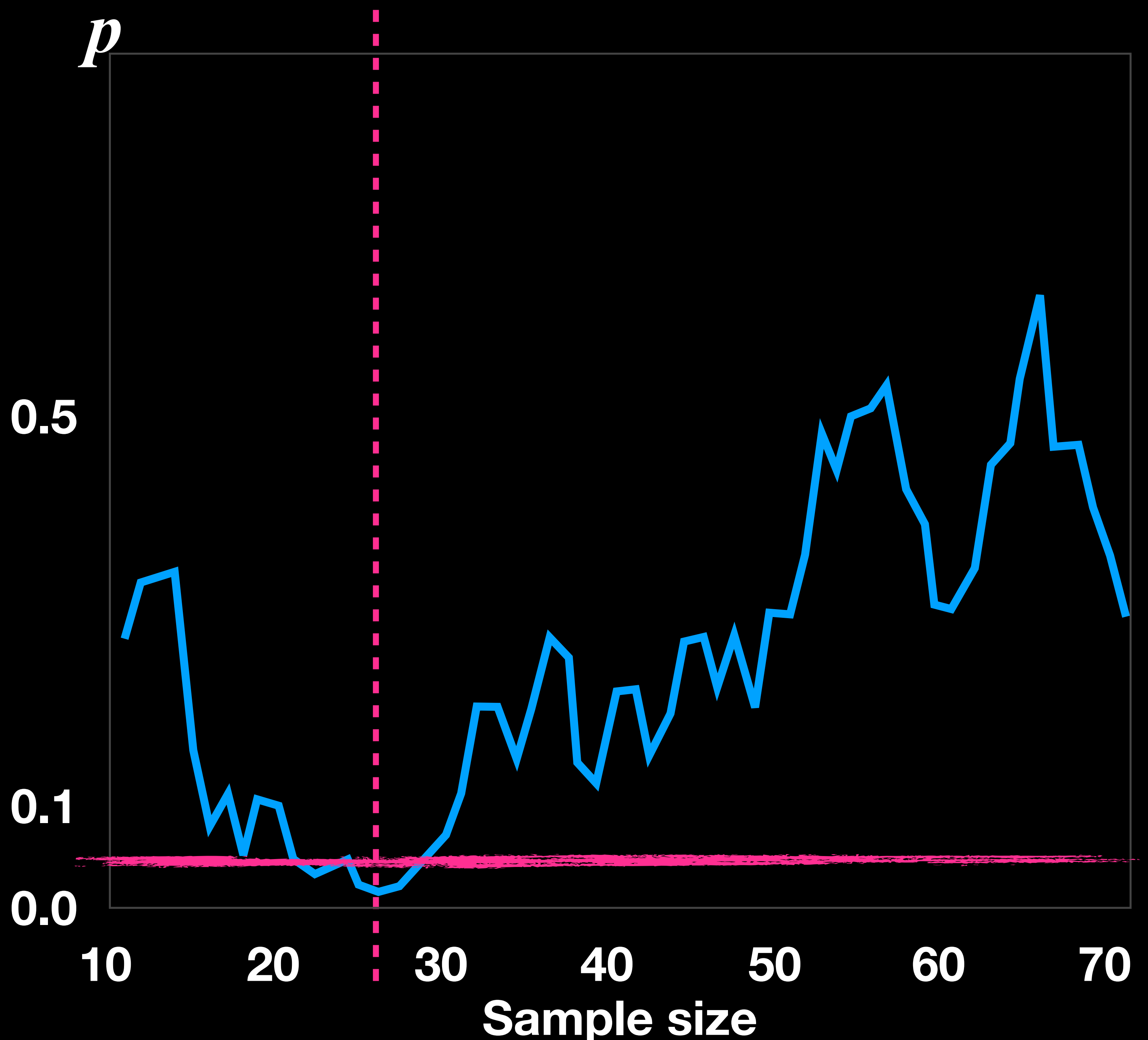
Don’t choose among metrics

metric	p
~~F1~~	0,0899
~~precision~~	0,062
~~recall~~	0,179
accuracy	0,0014

REPORT!

Bocconi

Don’t choose sample sizes



“We observed significant results at a sample size of 26”

…but not with smaller or larger samples!

Bocconi

Don't Choose Subsets

“Young, left-handed, vegetarian atheists are significantly less likely to get X”

…but the population as a whole isn't!

Wrapping Up

Take-home points

- Choose the **appropriate performance metric**
- Choose an **informative baseline**
- **Regularize, regularize, regularize**
- **Feature selection** can improve performance and provide insights
- Measure **significance** of improvement