

# Practical advice on using AI and machine learning on omics data

Wong Limsoon



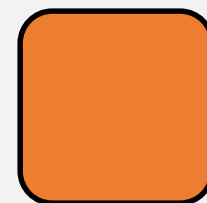
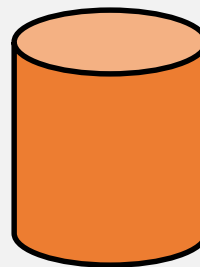
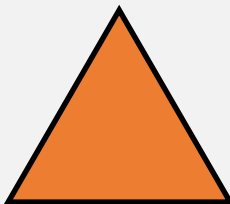
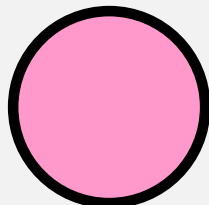
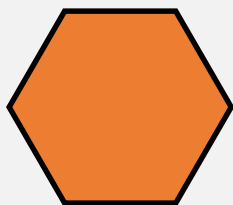
**NUS**  
National University  
of Singapore

National University of Singapore

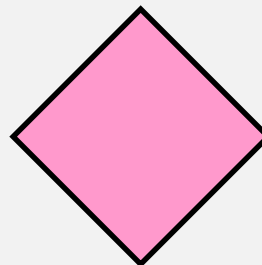
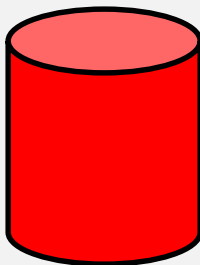
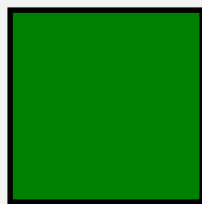
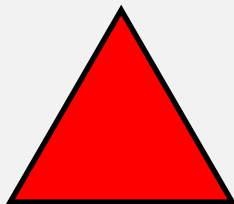
**We focus only on  
knowledge  
discovery &  
classifier learning  
here...**

# What is classifier learning?

Jonathan's blocks



Jessica's blocks

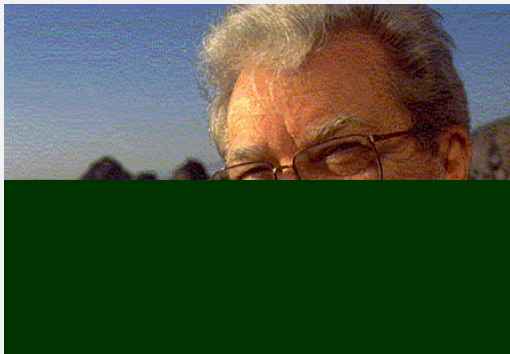


Whose block  
is this?

Jonathan's rules  
Jessica's rules

: Blue or Circle  
: All the rest

# | What is classifier learning?



Question: Can you explain how?

# Classifier learning

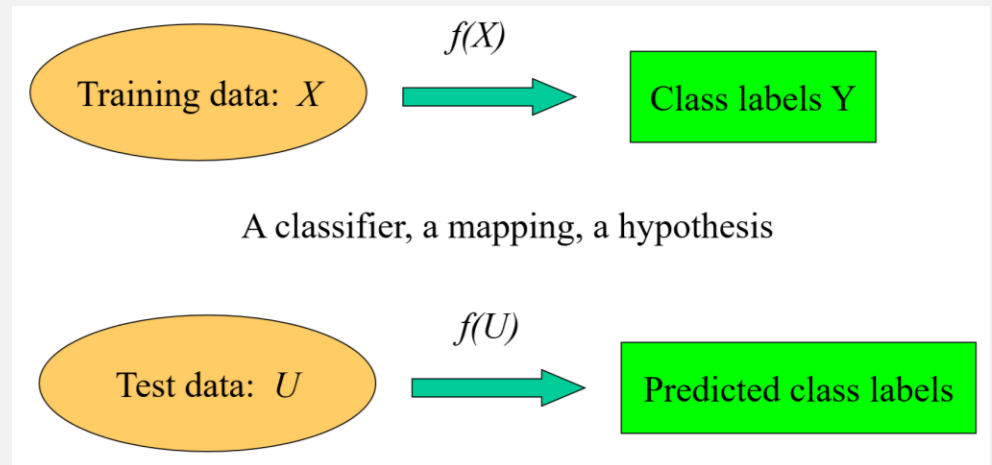
Learn from past experience, and use the learned knowledge to classify new data

Knowledge learned by intelligent algorithms

Examples

*Clinical diagnosis for patients*

*Cell type classification*



# Key steps of learning a classifier

Training data gathering

Feature generation

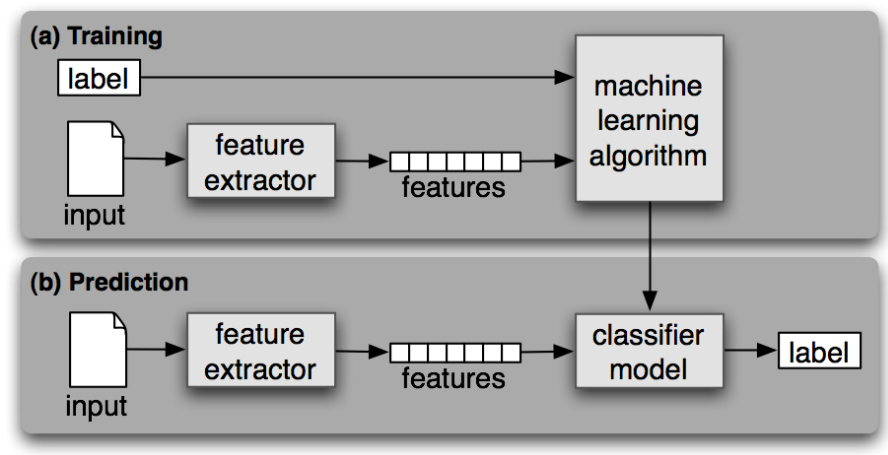
*k-grams, colour, texture, ...*

Feature selection

*Entropy,  $\chi^2$ , CFS, t-test, ...*

Feature integration by machine learning

*SVM, ANN, PCL, CART, C4.5, kNN, ...*



# Data

Classification application involves  $> 1$  class of data

*Normal vs disease cells for a diagnosis problem*

Training data is a set of instances (samples, points, etc.) with known class labels

Test data is a set of instances whose class labels are to be predicted

Outlook	Temp	Humidity	Windy	class
Sunny	75	70	true	Play
Sunny	80	90	true	Don't
Sunny	85	85	false	Don't
Sunny	72	95	true	Don't
Sunny	69	70	false	Play
Overcast	72	90	true	Play
Overcast	83	78	false	Play
Overcast	64	65	true	Play
Overcast	81	75	false	Play
Rain	71	80	true	Don't
Rain	65	70	true	Don't
Rain	75	80	false	Play
Rain	68	80	false	Play
Rain	70	96	false	Play

# Features (aka attributes)

Categorical features

*Outlook, windy*

Continuous or numerical features

*Temp, humidity*

Discretization

Outlook	Temp	Humidity	Windy	class
Sunny	75	70	true	Play
Sunny	80	90	true	Don't
Sunny	85	85	false	Don't
Sunny	72	95	true	Don't
Sunny	69	70	false	Play
Overcast	72	90	true	Play
Overcast	83	78	false	Play
Overcast	64	65	true	Play
Overcast	81	75	false	Play
Rain	71	80	true	Don't
Rain	65	70	true	Don't
Rain	75	80	false	Play
Rain	68	80	false	Play
Rain	70	96	false	Play



# Feature selection

# Basic idea of feature selection



Choose features that have high inter-class distance and low intra-class distance

E.g. t-statistic

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

# Food for thought



*Original photographer unknown/*

See also [www.cs.gmu.edu/~jessica/DimReducDanger.htm](http://www.cs.gmu.edu/~jessica/DimReducDanger.htm)

© Eamonn Keogh

# Machine learning methods

# Popular machine learning methods

K-nearest neighbor (kNN)

Support vector machines (SVM)

Naïve Bayes

Neural networks

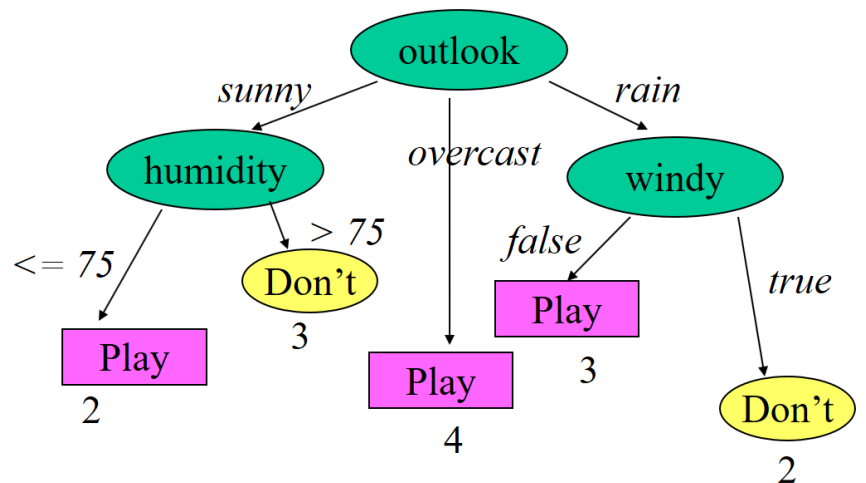
Decision trees & many more...

*WEKA is a nice free package for  
classifier learning*

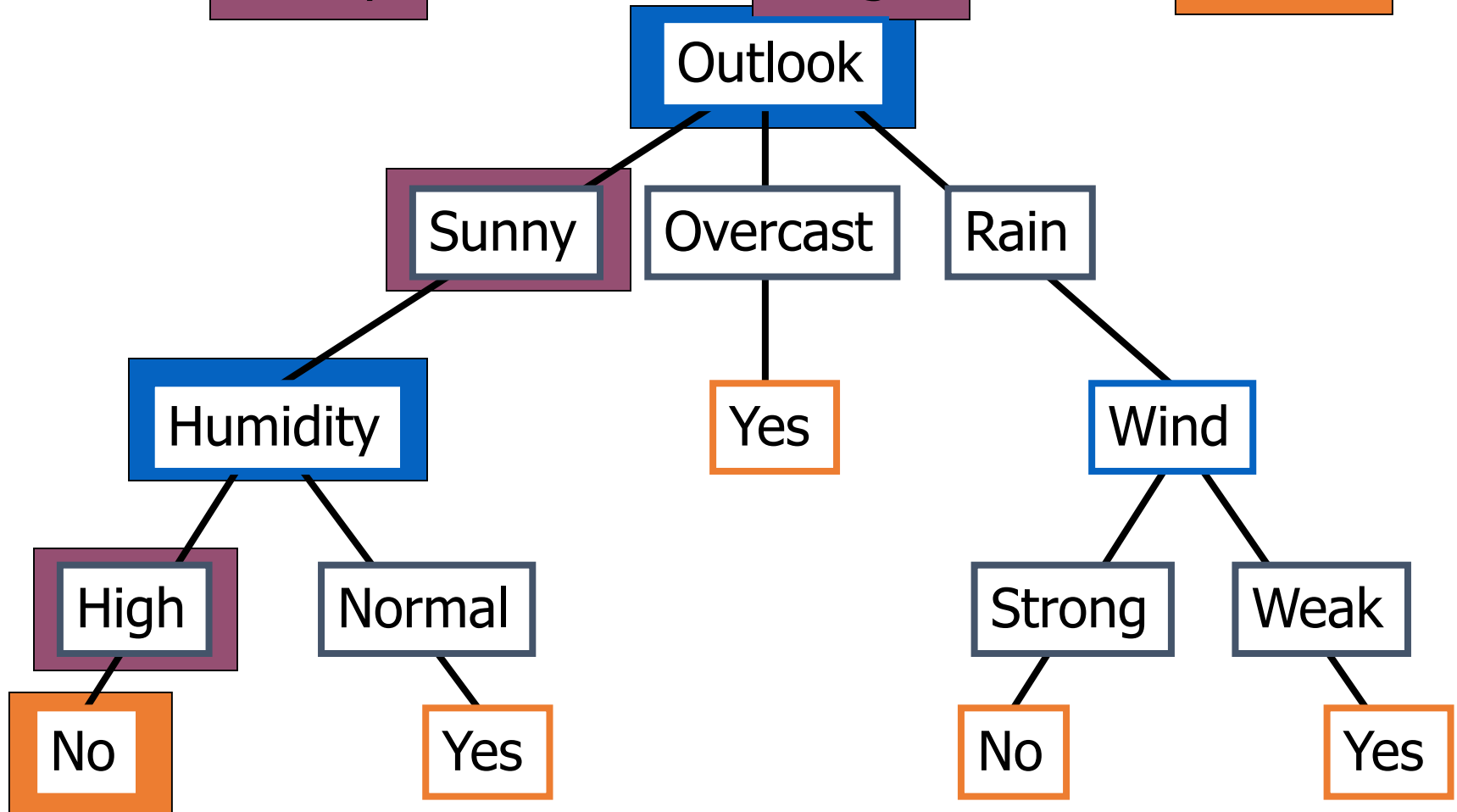


# Decision tree

Outlook	Temp	Humidity	Windy	class
Sunny	75	70	true	Play
Sunny	80	90	true	Don't
Sunny	85	85	false	Don't
Sunny	72	95	true	Don't
Sunny	69	70	false	Play
Overcast	72	90	true	Play
Overcast	83	78	false	Play
Overcast	64	65	true	Play
Overcast	81	75	false	Play
Rain	71	80	true	Don't
Rain	65	70	true	Don't
Rain	75	80	false	Play
Rain	68	80	false	Play
Rain	70	96	false	Play



Outlook Temperature Humidity Wind PlayTennis  
Sunny Hot High Weak No



# Decision tree construction

Select the “best” feature as root node of the whole tree

Partition dataset into subsets using this feature so that the subsets are as “pure” as possible

After partition by this feature, select the best feature (wrt the subset of training data) as root node of this sub-tree

Recursively, until the partitions become pure or almost pure



# Ensemble classifier

$h_1, h_2, h_3$  are indep classifiers w/ accuracy = 60%

$C_1, C_2$  are the only classes

$t$  is a test instance in  $C_1$

$h(t) = \operatorname{argmax}_{C \in \{C_1, C_2\}} |\{h_j \in \{h_1, h_2, h_3\} \mid h_j(t) = C\}|$

Then  $\operatorname{prob}(h(t) = C_1)$

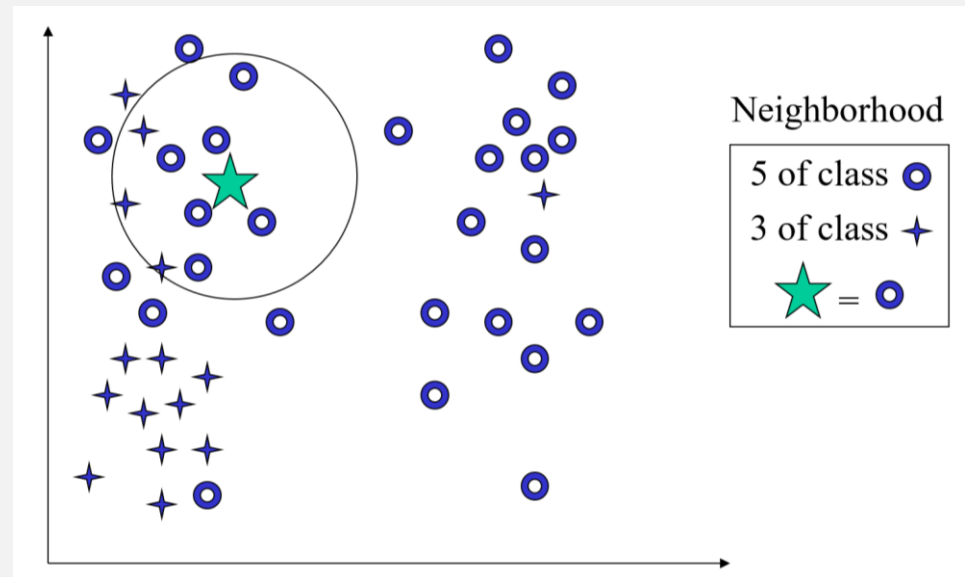
$$\begin{aligned} &= \operatorname{prob}(h_1(t)=C_1 \ \& \ h_2(t)=C_1 \ \& \ h_3(t)=C_1) + \\ &\quad \operatorname{prob}(h_1(t)=C_1 \ \& \ h_2(t)=C_1 \ \& \ h_3(t)=C_2) + \\ &\quad \operatorname{prob}(h_1(t)=C_1 \ \& \ h_2(t)=C_2 \ \& \ h_3(t)=C_1) + \\ &\quad \operatorname{prob}(h_1(t)=C_2 \ \& \ h_2(t)=C_1 \ \& \ h_3(t)=C_1) \\ &= 60\% * 60\% * 60\% + 60\% * 60\% * 40\% + \\ &\quad 60\% * 40\% * 60\% + 40\% * 60\% * 60\% = 64.8\% \end{aligned}$$

# kNN

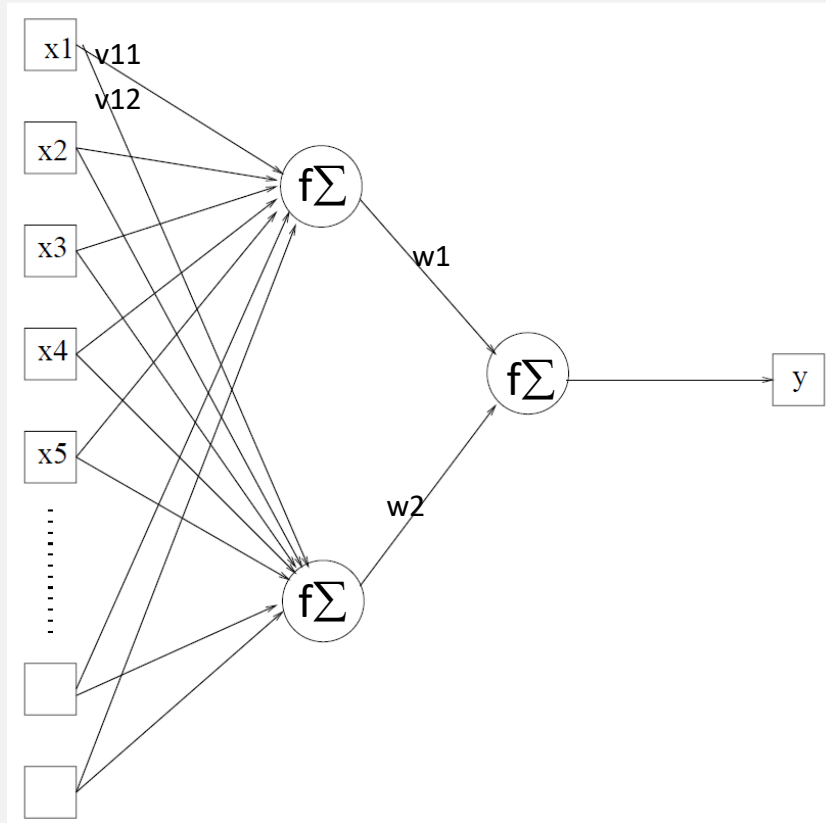
Given a new case

Find  $k$  “nearest” neighbours, i.e.,  $k$  most similar points in the training data set

Assign new case to the same class to which most of these neighbours belong



# Neural networks



$$y = f \left( \sum_j w_j f \left( \sum_i x_i v_{ij} \right) \right)$$

# Classifier assessment

# Common assessment measures

Accuracy

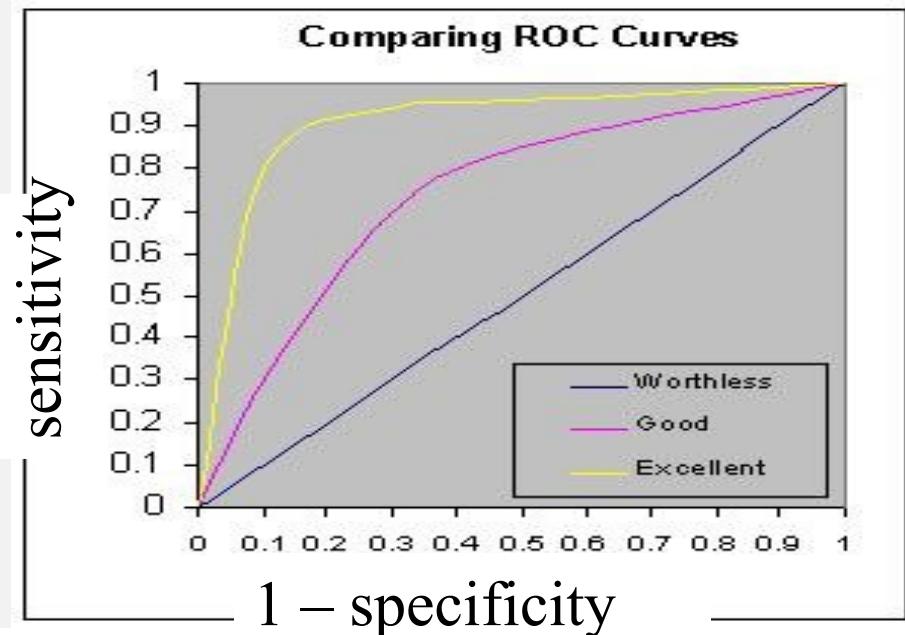
Sensitivity

Specificity

Precision

ROC curves

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN



# | Food for thought

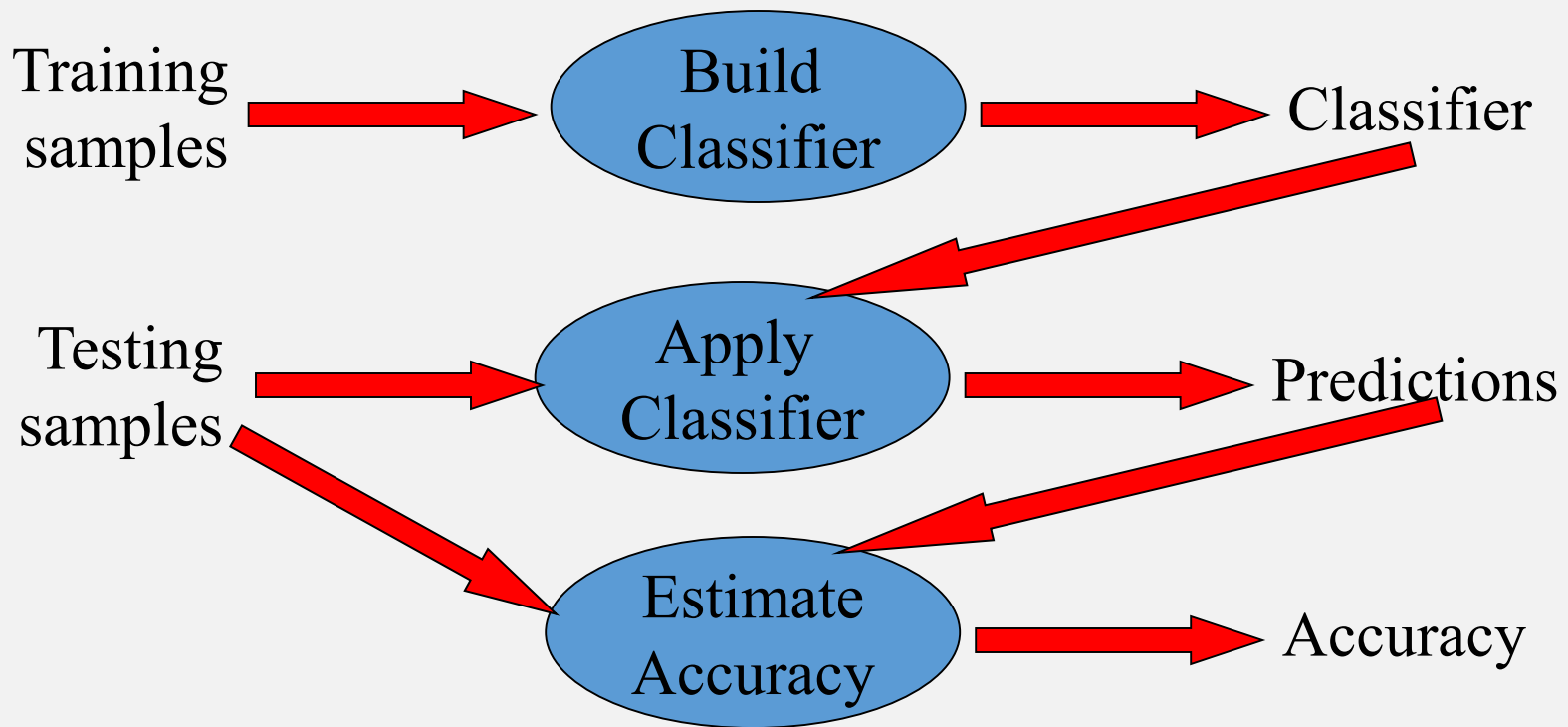
You have a classifier. On a test set having 20% +ve and 80% -ve cases, the classifier's recall and precision are both 80%

Suppose you test it on a new test set having 80% +ve and 20% -ve cases. What do you expect its accuracy to be?

You may assume that the +ve (resp. -ve) cases in both test sets are equally sufficiently representative of the +ve (resp. -ve) real-world population

What lesson have you learned?

# Estimating accuracy of classifier learning



**Testing samples are NOT to be used during “Build Classifier”**

# Cross validation

1.Test	2.Train	3.Train	4.Train	5.Train
--------	---------	---------	---------	---------

1.Train	2.Test	3.Train	4.Train	5.Train
---------	--------	---------	---------	---------

1.Train	2.Train	3.Test	4.Train	5.Train
---------	---------	--------	---------	---------

1.Train	2.Train	3.Train	4.Test	5.Train
---------	---------	---------	--------	---------

1.Train	2.Train	3.Train	4.Train	5.Test
---------	---------	---------	---------	--------

Divide samples into  
k roughly equal parts

Each part has  
similar proportion of  
samples from  
different classes

Use each part to test  
other parts

Total up accuracy



# | Food for thought

What is the logical basis of cross validation?

*Hint: Central limit theorem*

What / whose accuracy does it really estimate?

# An old example

# Childhood ALL

Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid >50

Diff subtypes respond differently to same treatment

∴ Match treatment to subtype

Conventional diagnosis

*Immunophenotyping*

*Cytogenetics*

*Molecular diagnostics*

Can gene expression profiling be used to replace all these?

# Childhood ALL subtype diagnosis by classifier learning

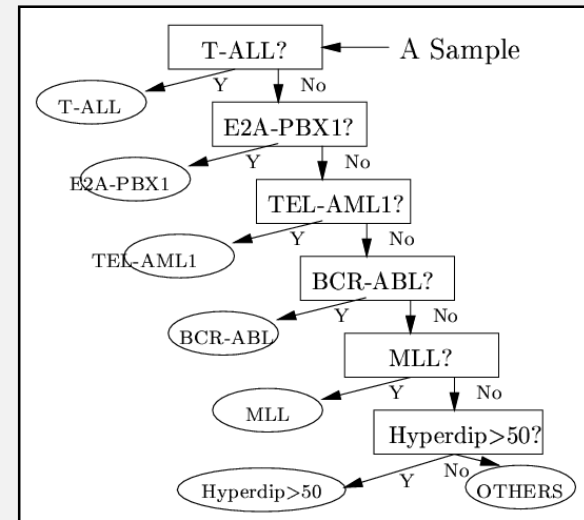
Gene expression data  
collection

Gene selection by  $\chi^2$

Classifier training

Apply classifier for  
diagnosis of future cases

# Childhood ALL subtype diagnosis workflow



Paired datasets	Ingredients	Training	Testing
T-ALL vs OTHERS1	OTHERS1 = {E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL, Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

# Accuracy of various classifiers

Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS <sup>1</sup>	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS <sup>2</sup>	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS <sup>3</sup>	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS <sup>4</sup>	2:0	3:0	1:4	2:0
MLL vs OTHERS <sup>5</sup>	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

The classifiers are all applied to the 20 genes selected by  $\chi^2$  at each level of the tree

**Don't worry about  
small differences in  
accuracy etc.**

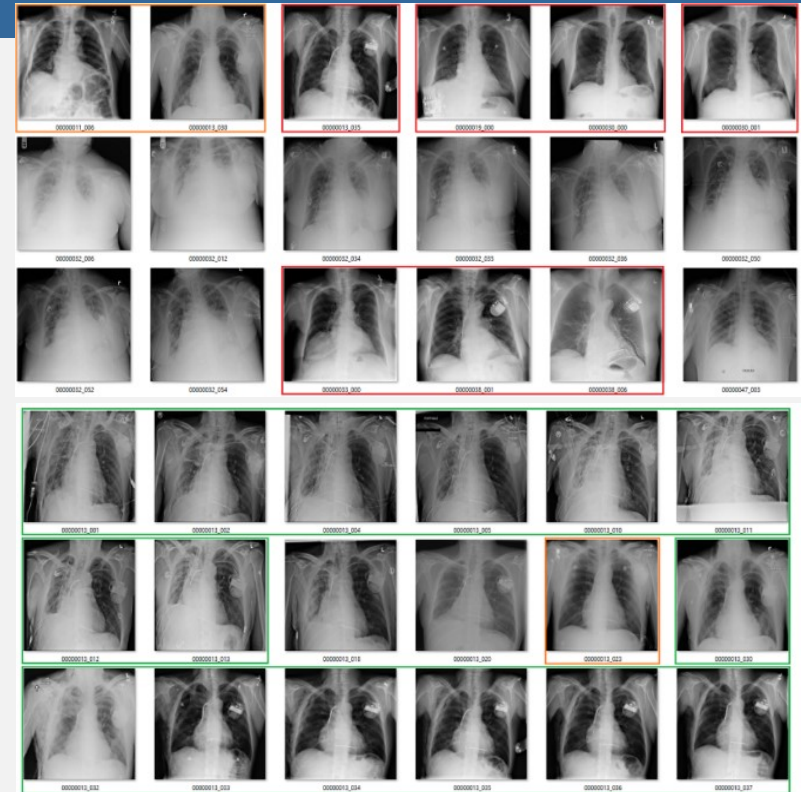
**These differences  
are meaningless**

**More important to figure out whether you have learned a meaningful classifier that is robust**



# A recent story

Disease	MetaMap			Our Method		
	Precision / Recall / F1-score			Precision / Recall / F1-score		
	OpenI					
Atelectasis	87.3 /	96.5 /	91.7	88.7 /	96.5 /	92.4
Cardiomegaly	100.0 /	85.5 /	92.2	100.0 /	85.5 /	92.2
Effusion	90.3 /	87.5 /	88.9	96.6 /	87.5 /	91.8
Infiltration	68.0 /	100.0 /	81.0	81.0 /	100.0 /	89.5
Mass	100.0 /	66.7 /	80.0	100.0 /	66.7 /	80.0
Nodule	86.7 /	65.0 /	74.3	82.4 /	70.0 /	75.7
Pneumonia	40.0 /	80.0 /	53.3	44.4 /	80.0 /	57.1
Pneumothorax	66.7 /	90.9 /	76.9	76.9 /	57.1 /	66.7
Consolidation	94.1 /	64.0 /	76.2	94.1 /	64.0 /	82.4
Fibrosis	100.0 /	100.0 /	100.0	100.0 /	100.0 /	100.0
PT	100.0 /	75.0 /	85.7	100.0 /	75.0 /	85.7
Hernia	100.0 /	100.0 /	100.0	100.0 /	100.0 /	100.0
Total	77.2 /	84.6 /	80.7	89.8 /	85.0 /	87.3
ChestX-ray14						
Atelectasis	88.6 /	98.1 /	93.1	96.6 /	97.3 /	96.9
Cardiomegaly	94.1 /	95.7 /	94.9	96.7 /	95.7 /	96.2
Infiltration	87.7 /	99.6 /	93.3	94.8 /	99.2 /	94.0
Pneumonia	69.7 /	90.0 /	78.6	95.9 /	85.4 /	88.1
Pneumothorax	87.4 /	100.0 /	93.7	94.3 /	98.8 /	96.5
Consolidation	72.8 /	98.3 /	83.7	95.2 /	98.3 /	96.7
Edema	72.1 /	93.9 /	81.6	96.9 /	93.9 /	95.43
Emphysema	97.6 /	93.2 /	95.3	100.0 /	90.9 /	95.2
Fibrosis	84.6 /	100.0 /	91.7	91.7 /	100.0 /	95.7
PT	85.1 /	97.6 /	90.9	97.6 /	97.6 /	97.6
Hernia	66.7 /	100.0 /	80.0	100.0 /	100.0 /	100.0
Total	82.8 /	95.5 /	88.7	94.4 /	94.4 /	94.4



Really good results from a study published in CVPR 2017

Actually the dataset contained many mis-labeled or biased data  
Biased data – many pneumo-thorax cases were patients treated with chest drain

Will a classifier learned from yesterday's samples work well on tomorrow's samples?

Samples from diff batches are grouped together, regardless of subtypes and treatment response

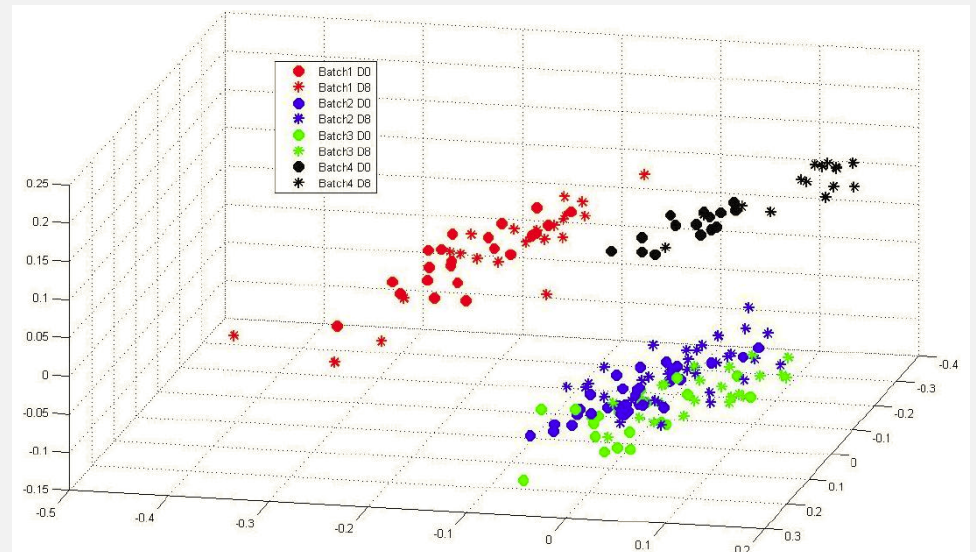
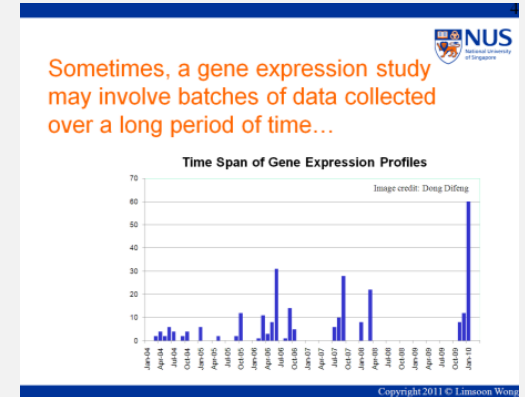


Image credit: Difeng Dong's PhD dissertation, 2011

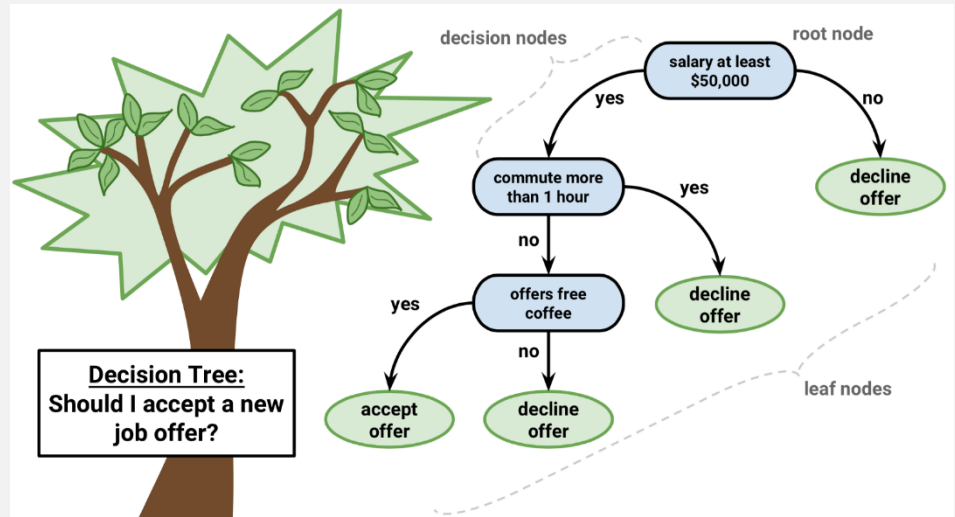
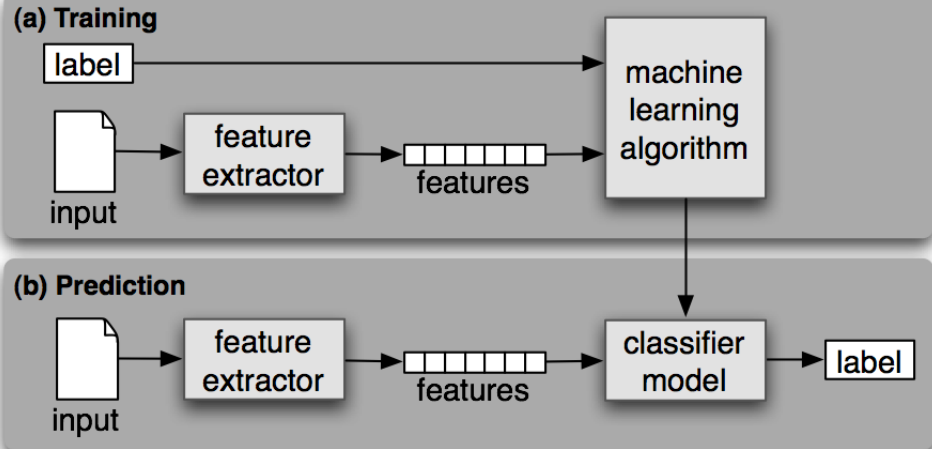
# Accuracy does not correlate with classifier similarity

NN	NN Acc. (%)	Acc. $t_1$ -sparse (%)	Acc. $t_2$ -sparse (%)	NPAQ r for $t_1$ -sparse (%)	NPAQ r for $t_2$ -sparse (%)
ARCH <sub>1</sub>	74.00	78.00	81.00	20.31	62.50
ARCH <sub>2</sub>	62.00	73.00	78.00	12.50	65.62
ARCH <sub>3</sub>	76.00	82.00	83.00	4.17	65.62
ARCH <sub>4</sub>	50.00	64.00	72.00	1.56	65.62
ARCH <sub>5</sub>	78.00	82.00	83.00	7.29	65.62
ARCH <sub>6</sub>	80.00	11.00	87.00	37.50	55.47
ARCH <sub>7</sub>	87.00	89.00	89.00	6.25	79.69

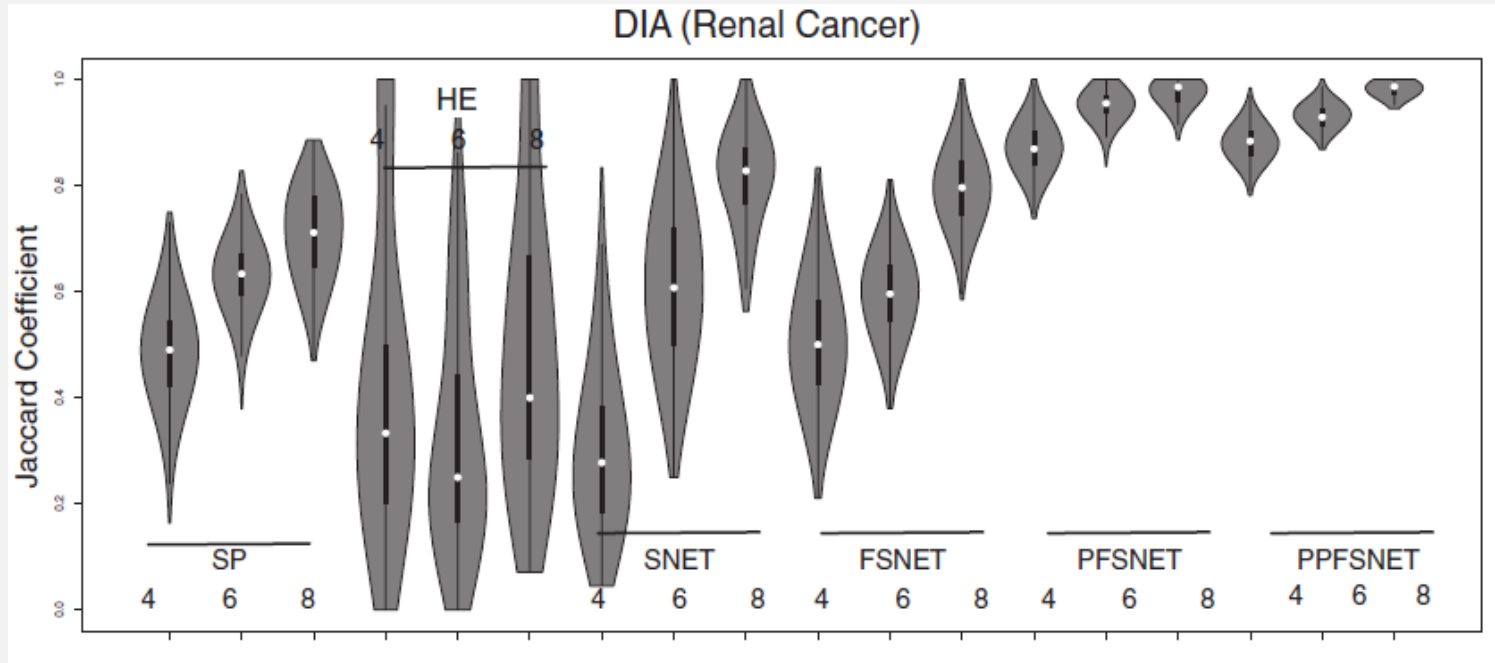
Although  $t_2$ -sparse and ARCH<sub>7</sub> are both ~90% accurate on the test set, they will disagree on ~80% of future cases

Table 2: First and second column refer to the baseline model where we use BNNs with 7 different architectures. The third and fourth represent the accuracies of sparsified models with  $t_1 = 0.03, t_2 = 0.05$  sparsification thresholds. The last 2 columns show NPAQ estimates for the difference between each sparsified model and the original model.

Features used by a prediction model are crucial for understanding the model and assessing its soundness



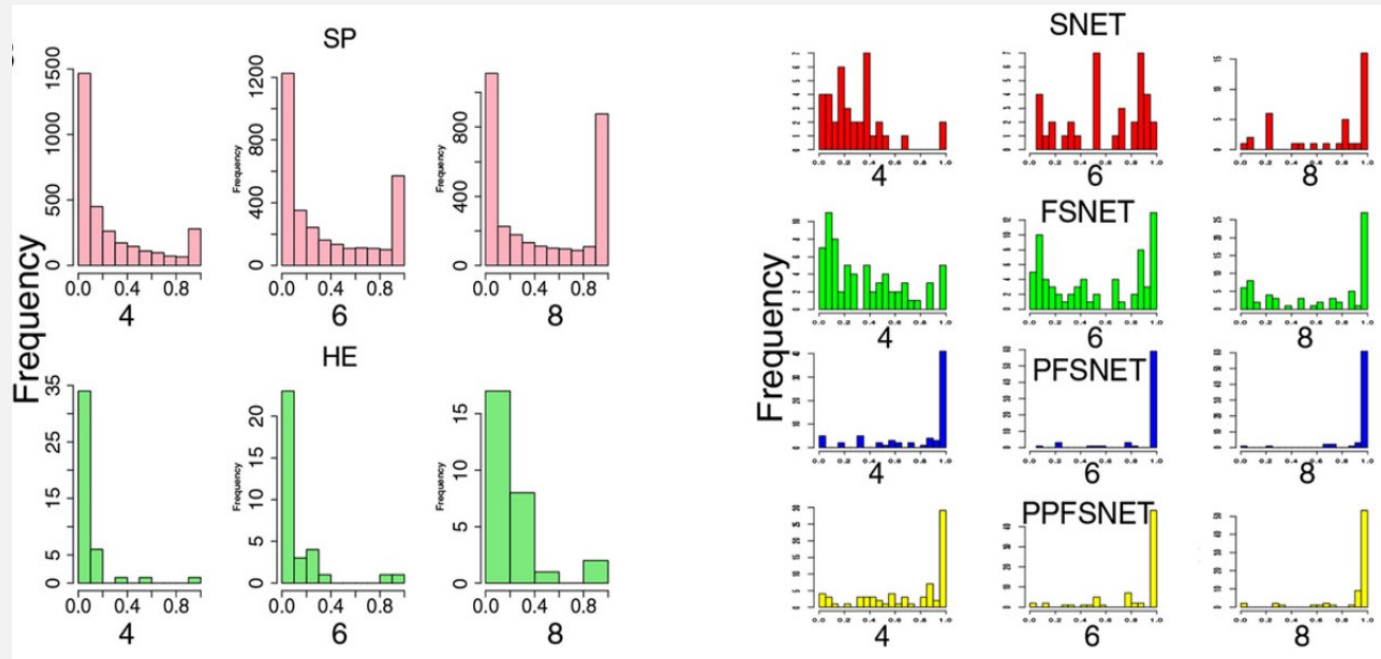
# High accuracy does not imply features used are reproducible



Agreement of feature sets selected from different samples of the same population is much poorer for methods that use no or “wrong” domain knowledge (SP, HE)

Goh & Wong. JBCB, 14(5):1650029, 2016.

# High accuracy does not imply features used are “meaningful”

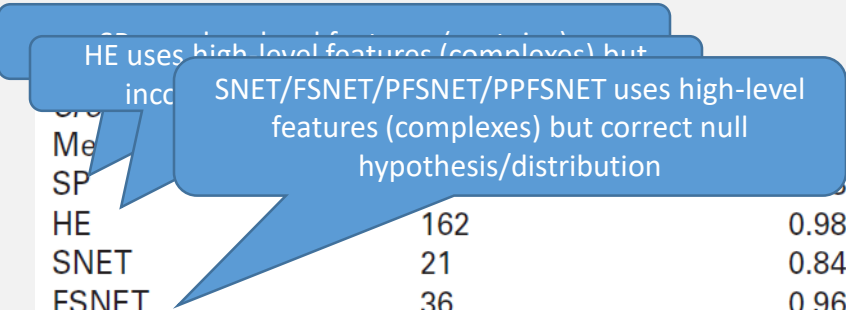


Features selected from different samples of the same population is much more unstable for methods that use no or “wrong” domain knowledge (SP, HE)

Goh & Wong. JBCB, 14(5):1650029, 2016.



# High accuracy does not imply features used are better than random



		accuracy	CV p-val	CV accuracy/pval
SP		0.91	0.91	1.08
HE	162	0.98	0.91	1.08
SNET	21	0.84	0.06	14.00
FSNET	36	0.96	0.06	16.00
PFSNET	65	0.92	0.06	15.33
PPFSNET	66	0.96	0.06	16.00

Classifiers trained on feature sets selected by SP, HE, etc. all have high accuracy

But they (SP/HE) may be confounded and result in classifiers not better than classifiers trained on comparable random feature sets of the same size

Goh, & Wong. Proteomics, 17(10):1700093, 2017

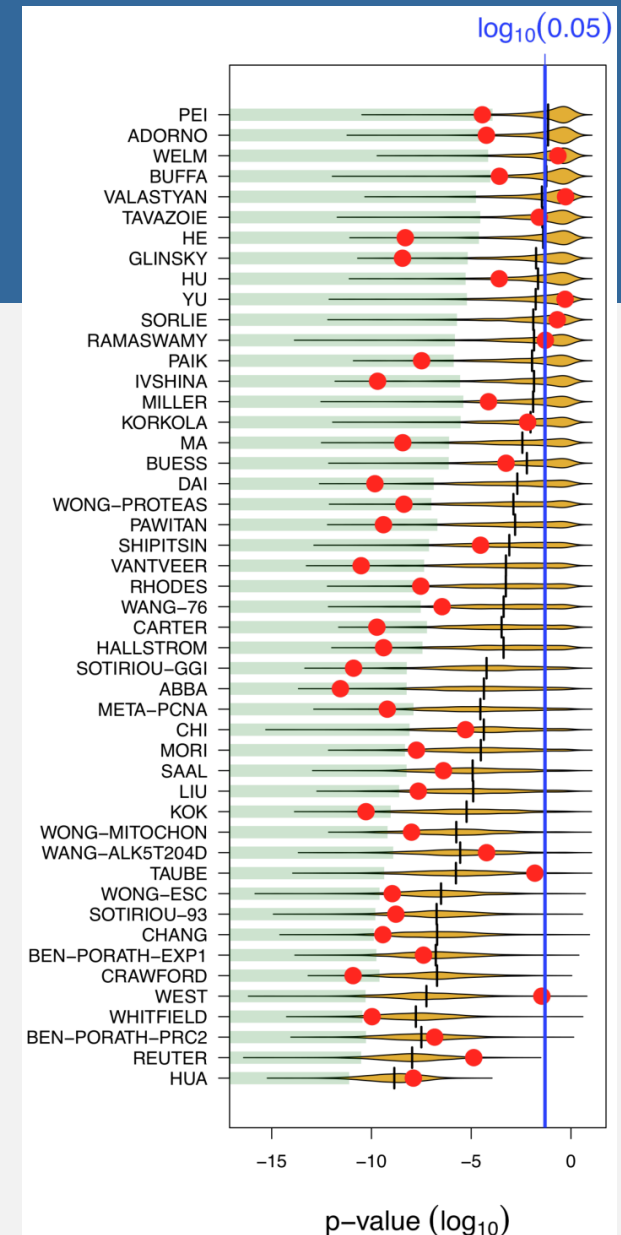
# **Some lessons from breast cancer survival signatures**



# Breast cancer survival signatures

40-50% of random signatures also have p-value  $\ll 0.05$

Significant signatures may be confounded; they are no better than random ones!



# | Food for thought

For any independent breast cancer survival dataset, a random signature has ~50% chance to be significant in it

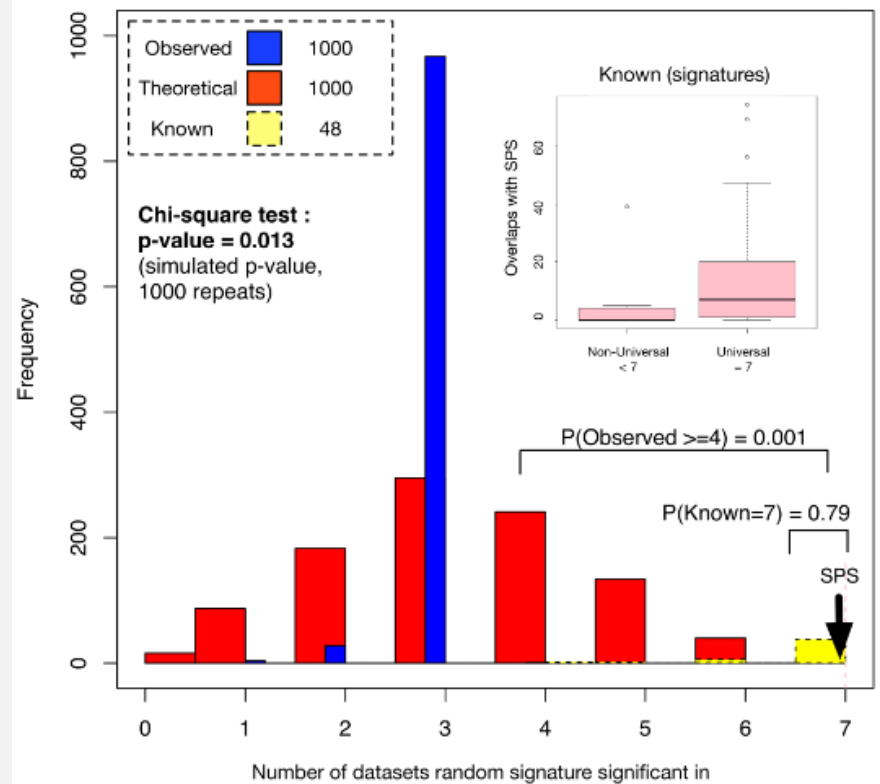
How many independent datasets are needed to avoid reporting random signatures as significant?

$n$	$(50\%)^n$
1	50.00%
2	25.00%
3	12.50%
4	6.25%
5	3.13%
6	1.60%
7	0.78%

# Test on 7 datasets

SPS & most known signatures are universally significant on 7 breast cancer datasets

Random signatures (same size as SPS) are hardly universal, even though they get better p-values than known signatures on some datasets



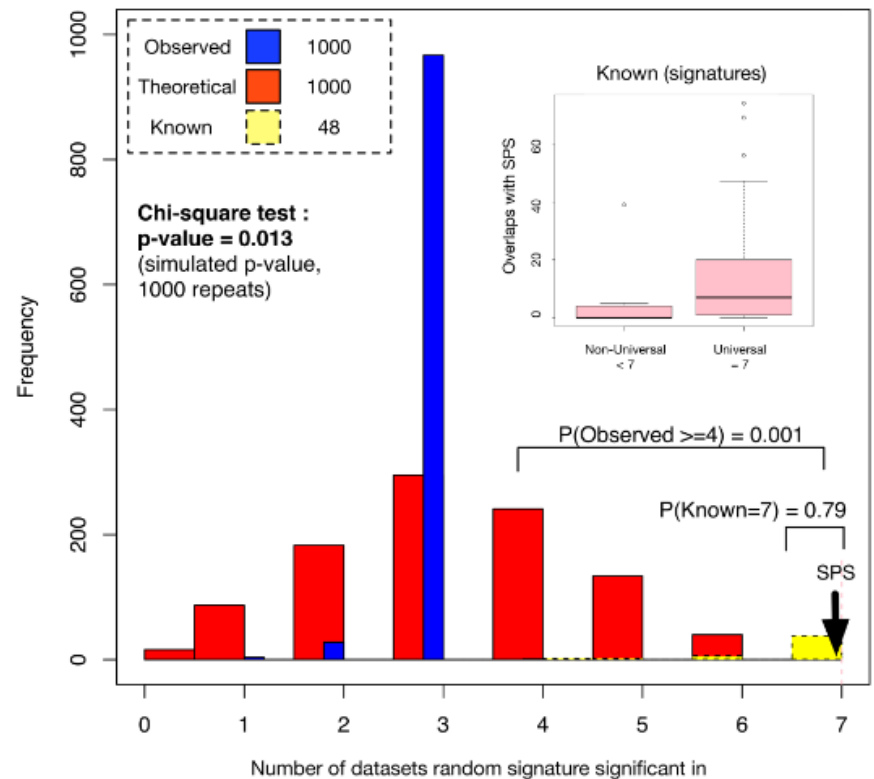
# Food for thought

40-50% of random signatures are significant in 1 dataset

Red histogram is expected # of random signatures significant in 1 to 7 independent dataset

Blue histogram is observed distribution

What does this figure tell?



# | Learning points

Many machine learning methods

Most methods give similar good results, when provided the same *informative* features

Accuracy etc. are too simple minded for assessing whether a prediction model is good

Validate on many datasets

Some independent datasets are *not* as independent as you think

# References

Li, et al. **Data mining techniques for the practical bioinformatician.** *The Practical Bioinformatician*, Chapter 3, pages 35-70, World Scientific, 2004

Li, et al. **Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients.** *Bioinformatics*. 19:71-78, 2003

Goh & Wong. **Dealing with confounders in -omics analysis.** *Trends in Biotechnology*, 36(5):488-498, 2018

Goh & Wong. **Why breast cancer signatures are no better than random signatures explained.** *Drug Discovery Today*, 23(11):1818-1823, 2018

Goh & Wong. **Turning straw into gold: Building robustness into gene signature inference.** *Drug Discovery Today*, 24(1):31-36, 2019