Some bad practices in data analysis and machine learning

WONG Limsoon

Plan

PCA for dimension reduction

Indiscriminately using “high” PCs

Mindlessly discarding “low” PCs

Pearson correlation measure

Mechanically trusting high correlation scores

Not seeing association behind low correlation scores

Classification accuracy as indicator of model quality

Misinterpreting accuracy w/o consideration of prevalence

Unjustifiably treating all test instances as equal

Asinine propagation of bias

Irresponsible use of badly prepared test sets

A common advice on using PCA

PCA is the process of computing the principal components and using them to perform a change of basis on the data, …

It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible

Wikipedia

This assumes variations in the first few PCs are more meaningful / useful than the other PCs. Is this a sound assumption?

PCA, intuitively

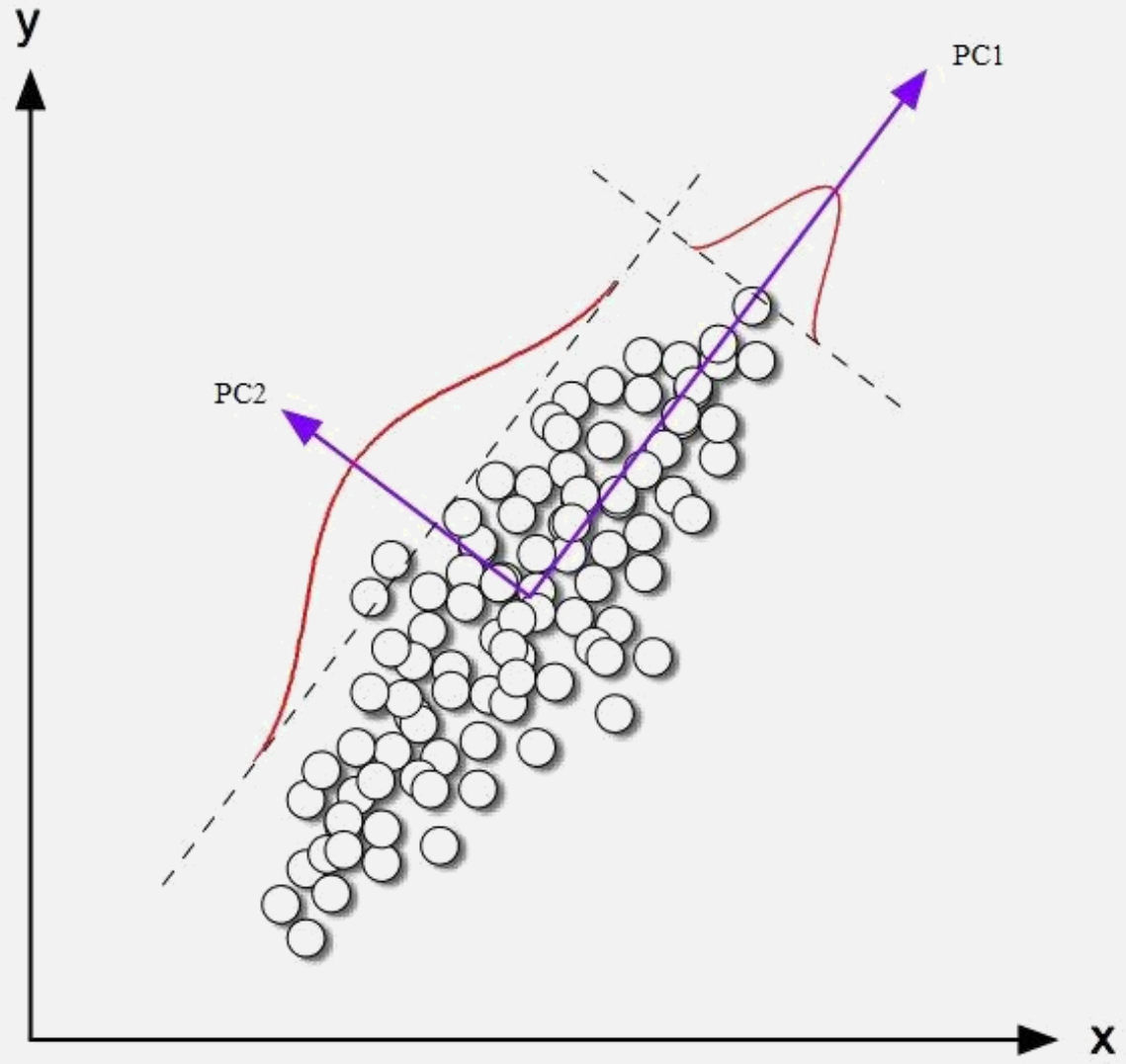


Image credit: Alessandro Giuliani

Exercise #1

Madrid and Warsaw are at almost the same distance to Italian cities

Are Madrid and Warsaw near each other?

Distances of European cities (km) from the main cities of Latium

	Rome	Latina	Frosinone	Viterbo	Rieti
Amsterdam	430	447	449	415	409
Athens	347	321	331	346	364
Barcelona	283	305	293	292	271
Beograd	227	222	236	220	238
Berlin	393	400	409	374	373
Bern	227	249	247	220	205
Bonn	353	370	372	339	330
Bruselles	388	406	406	371	365
Bucharest	364	355	368	359	378
Budapest	268	261	274	246	259
Calais	418	448	446	418	405
Copenhagen	510	522	527	492	491
Dublin	622	645	641	615	600
Edinburgh	637	655	655	625	615
Frankfurt	318	333	336	302	295
Hamburg	435	448	453	417	414
Helsinki	727	729	739	706	713
Istanbul	452	430	443	443	464
Lisbon	615	637	622	624	604
London	474	494	493	464	456
Luxembourg	325	346	346	315	307
Madrid	449	470	458	460	440
Marseille	200	223	213	202	183
Moscow	782	773	785	759	774
Munich	230	245	250	216	213
Oslo	664	675	682	646	645
Paris	365	386	383	357	343
Prague	305	313	320	286	290
Sofia	294	273	286	280	301
Stockholm	653	658	668	632	636
Warsaw	435	433	444	413	421
Vienna	255	254	265	233	240
Zurich	227	246	246	214	205

PCA of distance matrix of European cities to Italian cities

Factor loadings and proportions of explained variance

Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569	0.000043	0.000005

PC1 accounts for >99% of variance & correlates to distance of European cities to Latium cities

PC2, PC3, … account for < 1% of variance. Are they useless?

Variance is de-convoluted into real factors by PCA

PCs that don't correspond to real factors are thus Gaussian-like residual noise

"Pre-whitening" can therefore be used to check which PC's are informative:

Inject small Gaussian noise into data

Compare PC_i pre and post noise injection

High correlation $\Rightarrow$ PC_i carries info

Low correlation $\Rightarrow$ PC_i carries no info

Pre-whitening to distinguish informative vs uninformative PCs

So, PC4 and PC5 are residual noise

Whereas, PC1, PC2, and PC3 carry real information

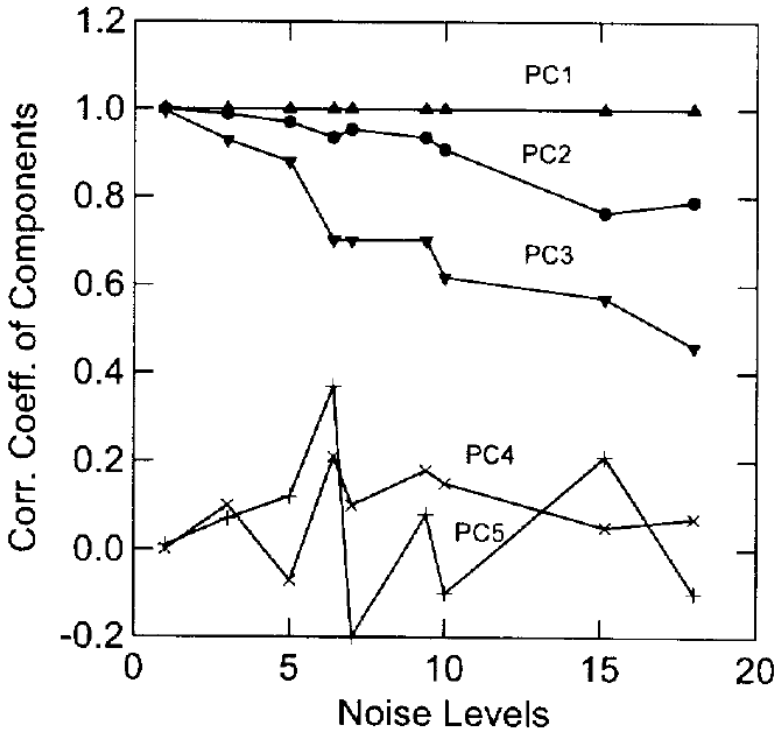
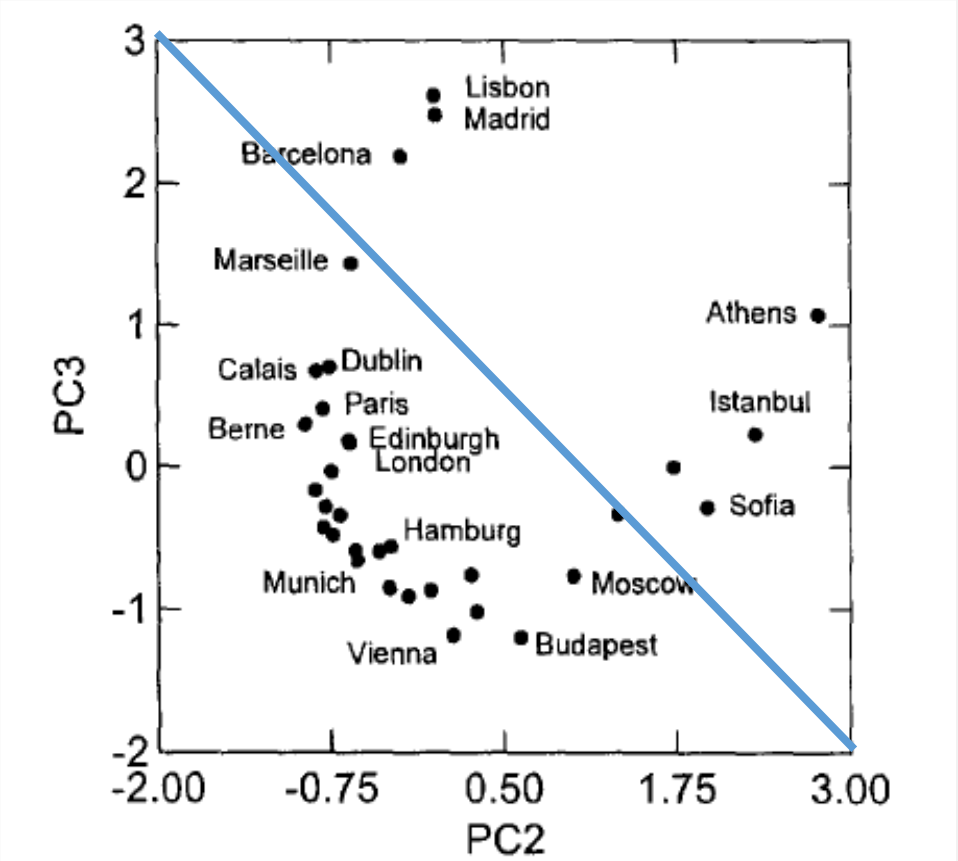


Fig. 3. The figure displays the degree of recognition (Pearson's correlation coefficient) between the noise-corrupted PCs and their original counterparts, for different amounts of noise. The abscissa in expressed in SD units (mm) (1 mm = 3 km).

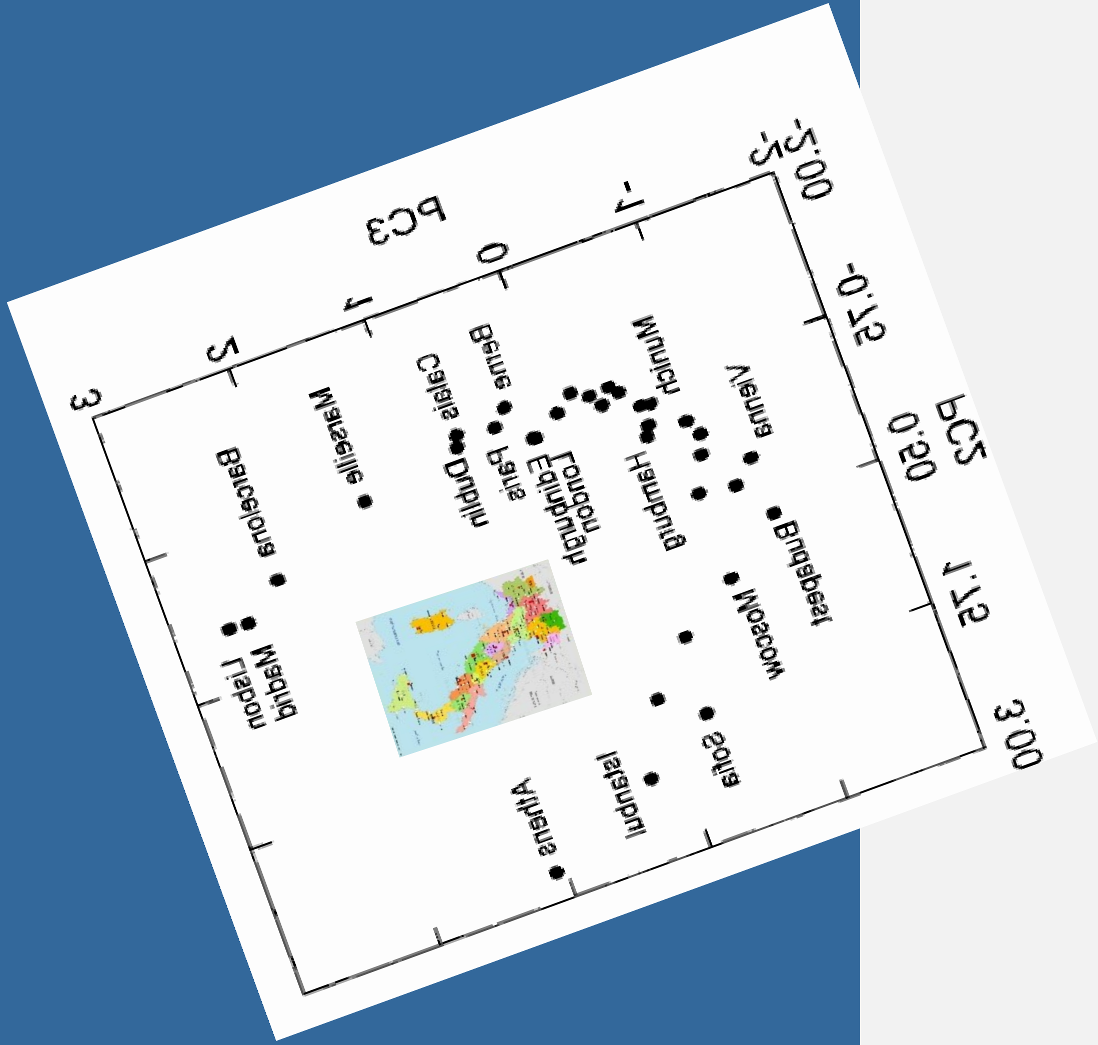
Giuliani et al., *Physics Letters A*, 247:47-52, 1998

PC2 & PC3 are the angular orientation of European cities centered on Latium

So, you can tell Madrid is not near Warsaw

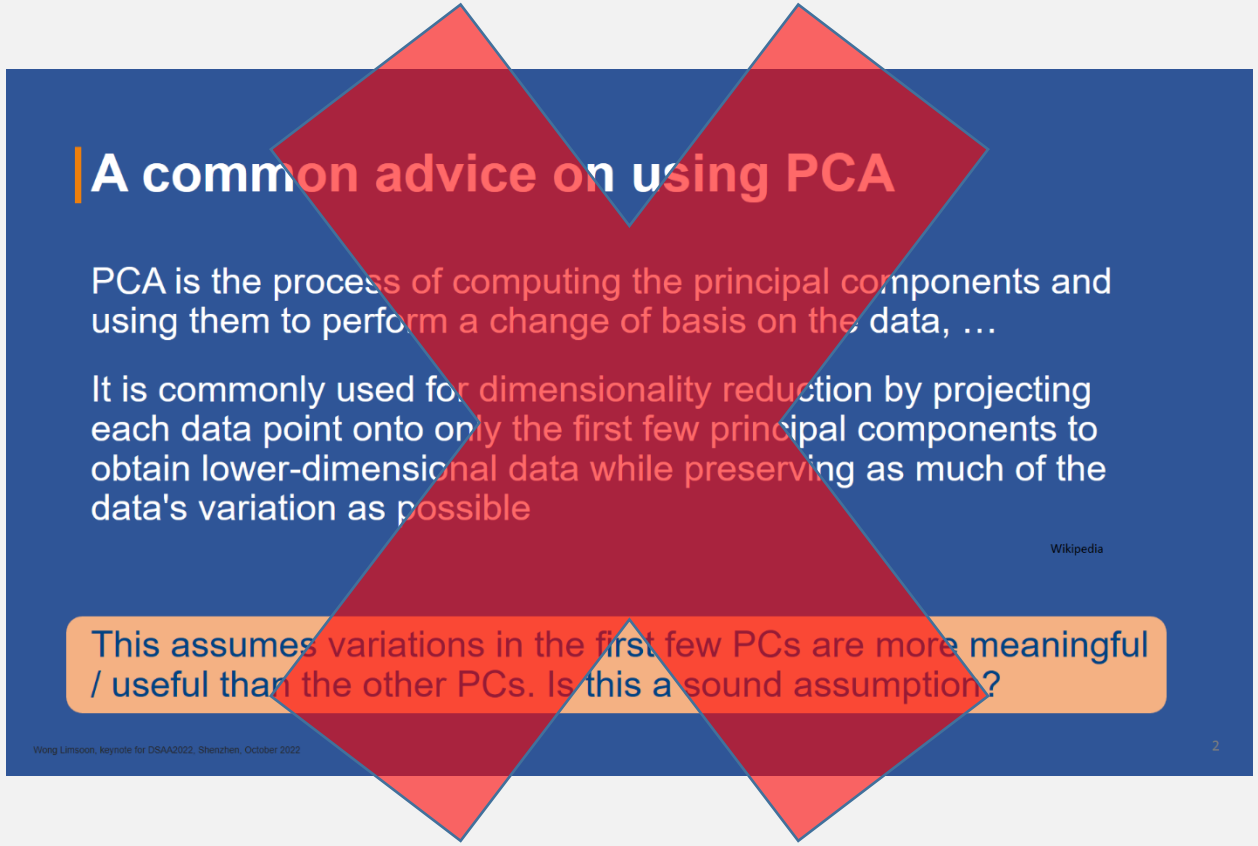


Giuliani et al., *Physics Letters A*, 247:47-52, 1998

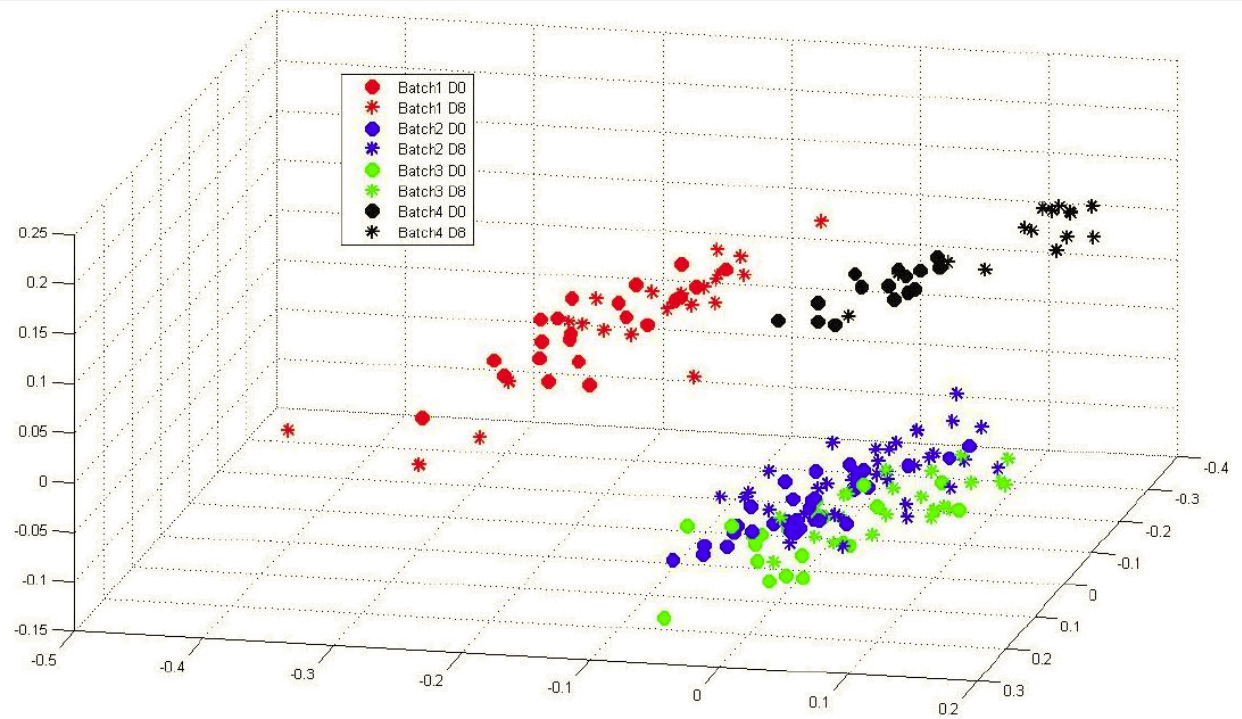


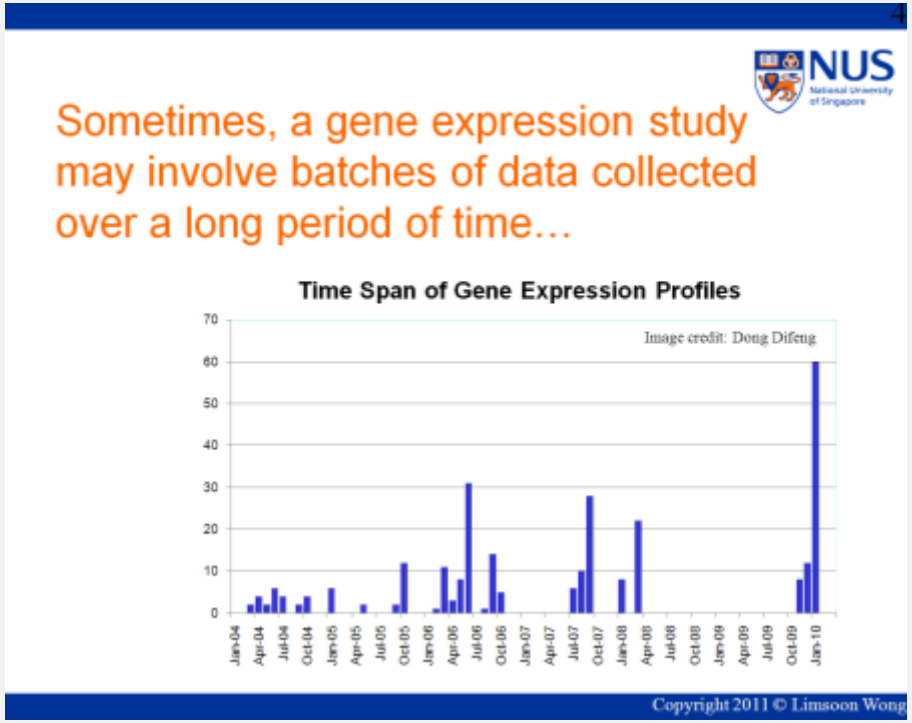

GREENLAND (DENMARK)
GREENLAND SEA
BARENTS SEA
WHITE SEA
NORWEGIAN SEA
Reykjavik
ICELAND
FAROE ISLANDS (DENMARK)
FINLAND
Helsinki
RUSSIA
Moscow
NORWAY
Oslo
SWEDEN
Stockholm
Tallinn
ESTONIA
LATVIA
LITHUANIA
Vilnyus
Minsk
BELARUS
NORTH SEA
DENMARK
Copenhagen
BALTIC SEA
UNITED KINGDOM
Dublin
IRELAND
London
Amsterdam
NETHERLANDS
Berlin
GERMANY
Warsaw
POLAND
Kiev
UKRAINE
Brussels
BELGIUM
Prague
CZECH REPUBLIC
ATLANTIC OCEAN
Paris
SLOVAKIA
MOLDAVIA
Vienna
AUSTRIA
HUNGARY
ROMANIA
Bucharest
Bern
SWITZERLAND
FRANCE
SERBIA
Belgrade
BULGARIA
BLACK SEA
ITALY
Rome
Madrid
SPAIN
PORTUGAL
Lisbon
TURKEY
ALBANIA
GREECE
Athens
TYRRHENIAN SEA
IONIAN SEA
MEDITERRANEAN SEA
Algiers
Tunis
Rabat
MAROCCO
ALGERIA
TUNISIA
VectorStock®
VectorStock.com/30039770

PCs corresponding < 1% of variation can be informative



PCA scatter plot





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

Image credit: Difeng Dong's PhD dissertation, 2011

Exercise #2

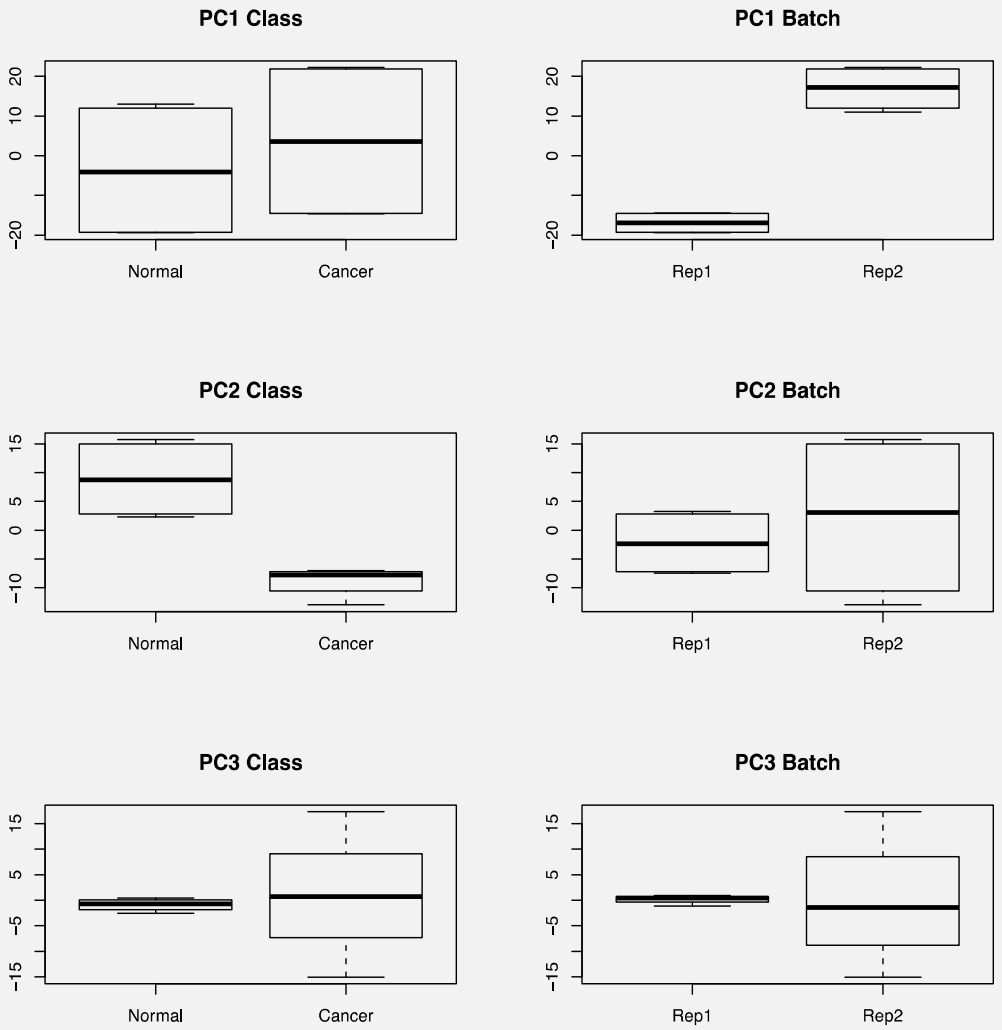
Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment, etc.

Batch effects is a big challenge faced in biological research, especially towards translational research and precision medicine

How do you know which PCs are dominated by batch effects?

Paired boxplots of PCs

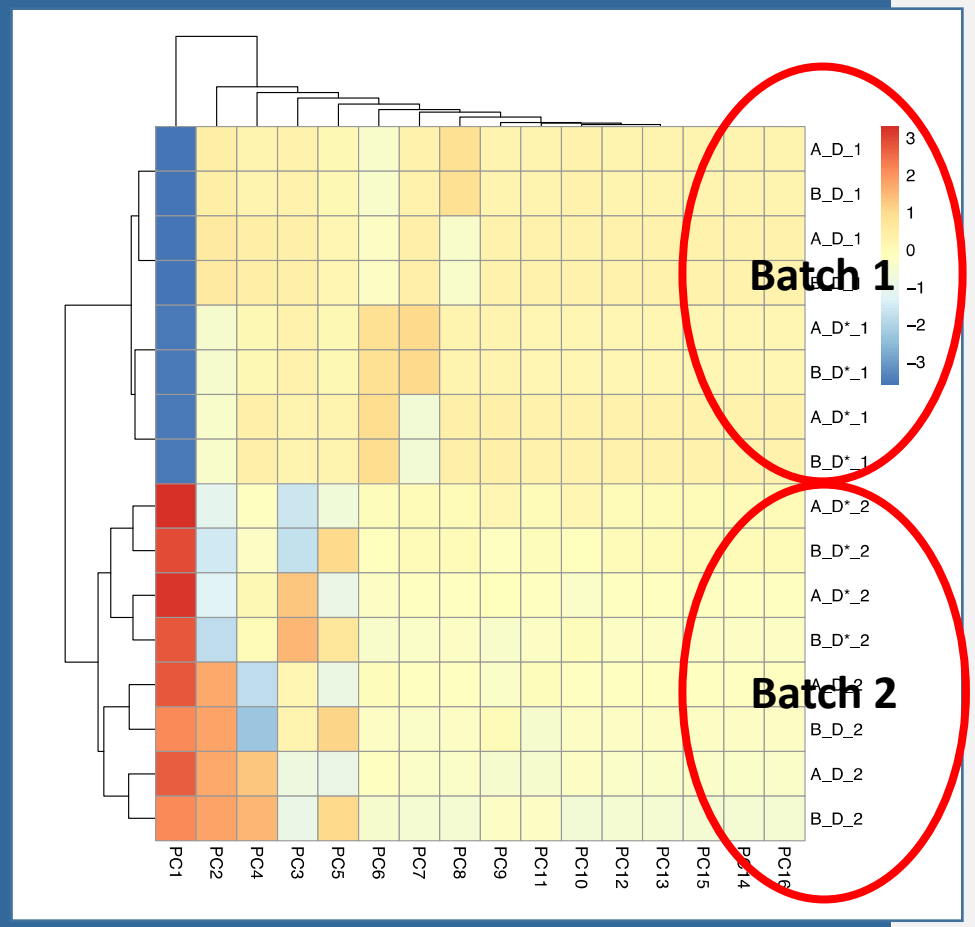
See which PC is enriched in batch effects by showing, side by side, distribution of values of each PC stratified by class and suspected batch variables



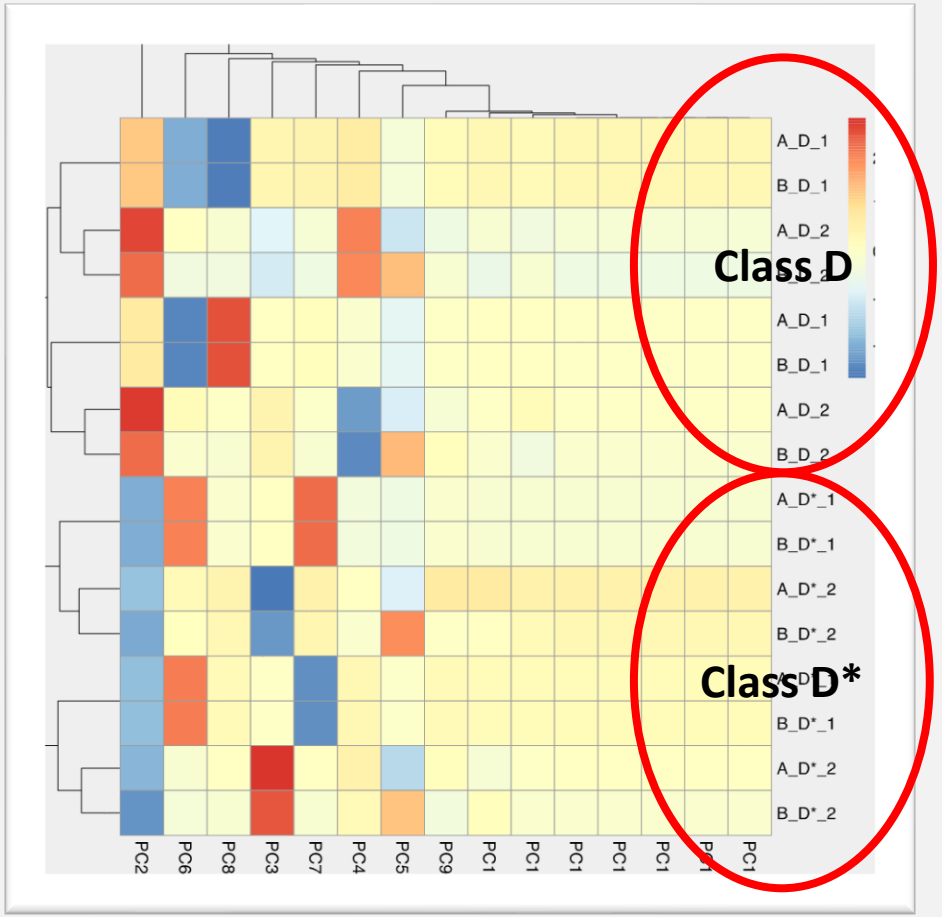
Goh & Wong, *BMC Genomics*18(Suppl2):142, 2017

Remove batch effects-laden PCs

(Notation: A/B_D/D*_1/2 refers to the dataset, class and batches respectively)

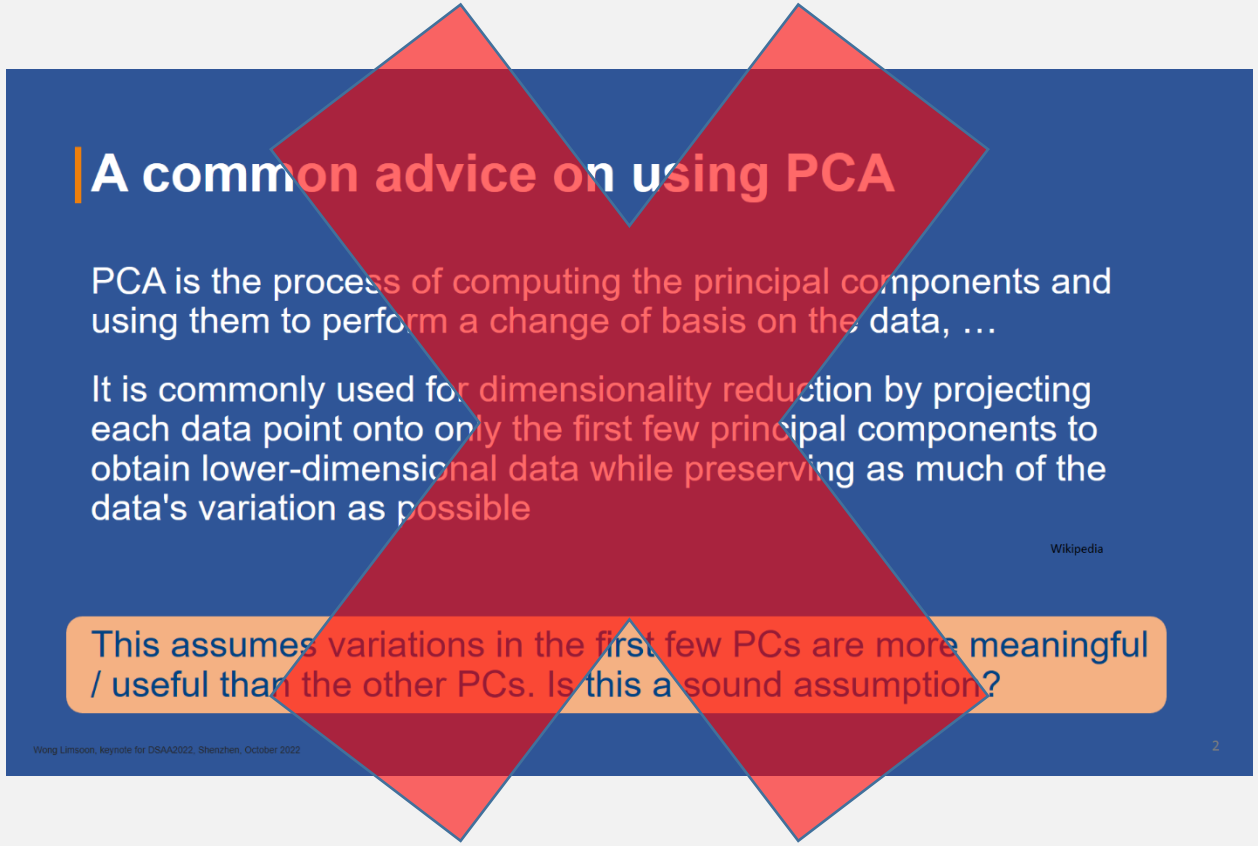


Batch effects dominate



Class-effect discrimination recovered

Top PCs can correspond to irrelevant or confounding information



A common advice, paraphrased from a business analytics class

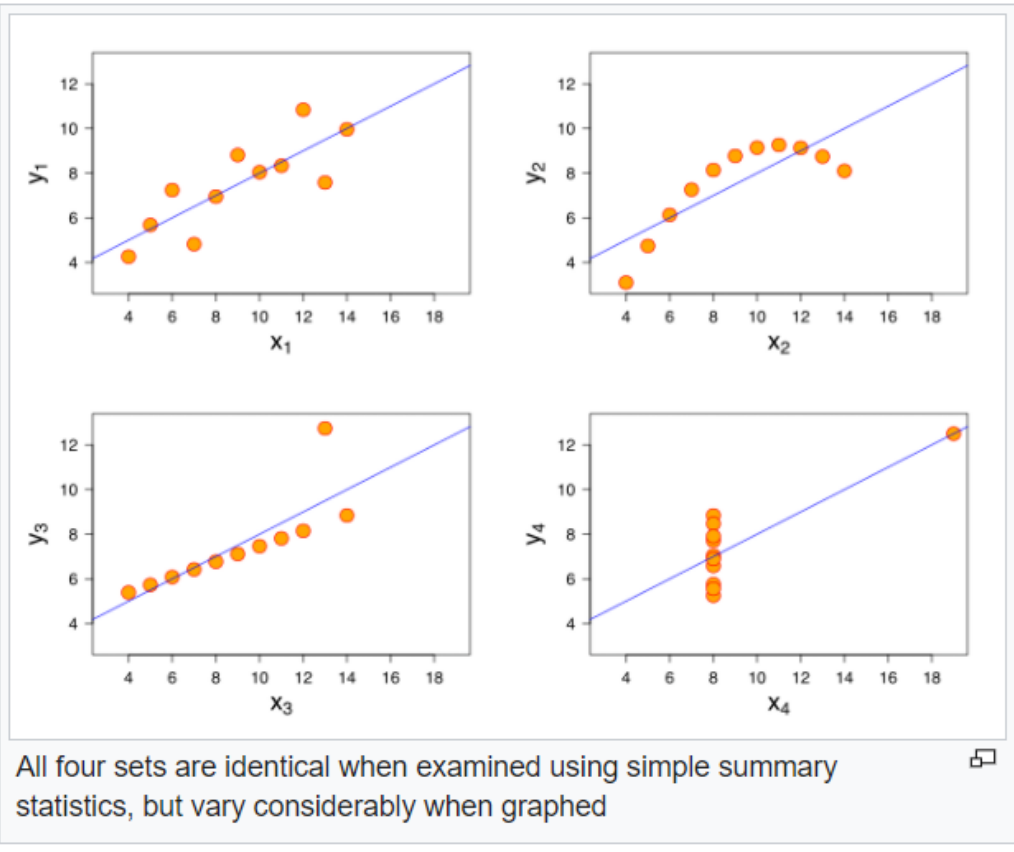
A correlation value of 0.8 between X and Y tells that an increase in X will lead to an increase in Y

And this insidious converse: No or low correlation between X and Y implies no relationship between X and Y

Are these sound advices? Note that correlation value is valid only when Y varies tightly and linearly wrt X

Anscombe quartet

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Wikipedia

High correlation does not imply good association

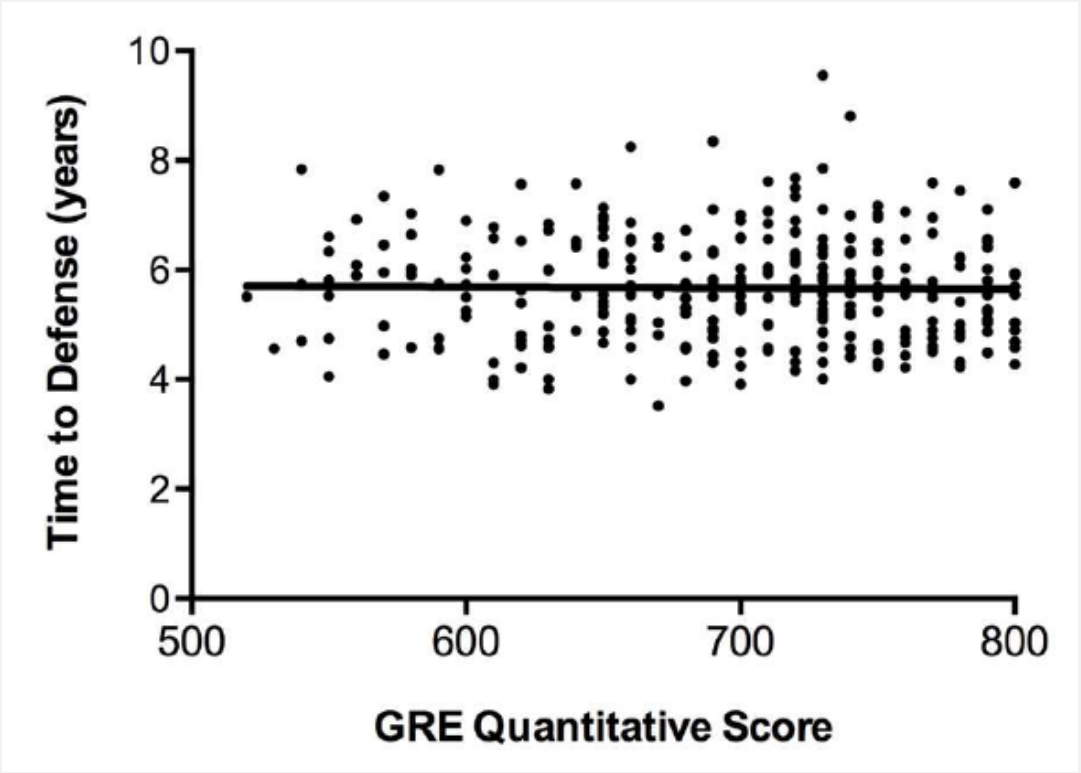
Vanderbilt GRE Study

Correlate GRE scores to things like year to thesis defense, publication count, etc.

Found no or low correlation

Conclude GRE is useless for PhD student admission purpose

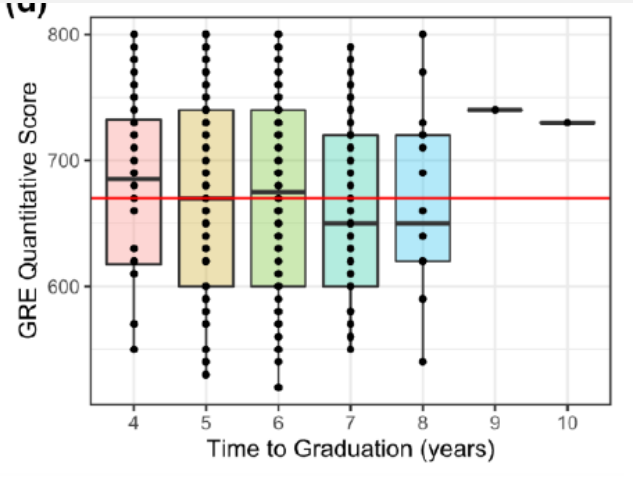
Really?



Moneta-Koehler et al., *PLOS ONE*, 12(1):e0166742, 2017

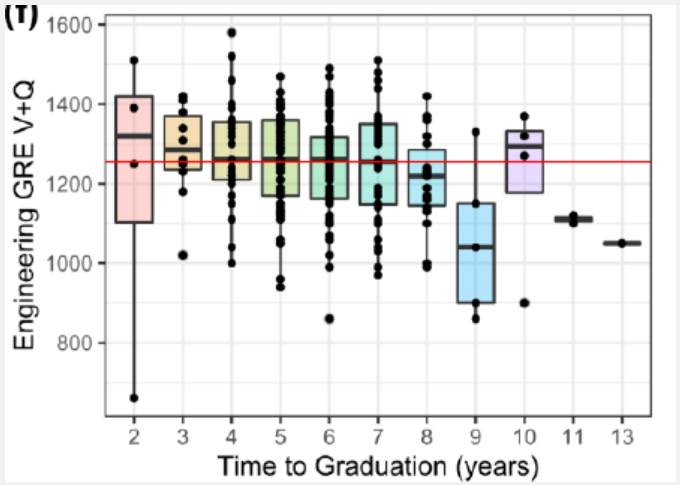
A different visualization

A different trend is easily grasp if we swap the axes and bin the points by yr



Alternative view of data of Moneta-Koehler et al., *PLOS ONE*, 12(1):e0166742, 2017

This trend is reproduced in a multi-institutional STEM PhD study



Alternative view of data of Petersen et al., *PLOS ONE*, 13(10):e0206570, 2018

This trend is reproduced in my own university's data!

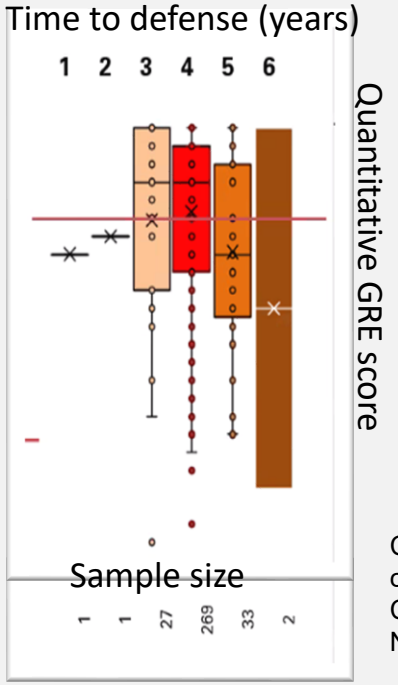


Chart communicated by Caroline Lee @ NUS

The top 20% GRE scorers are different from everyone else

Comparison	Risk Ratio	Chi-square p-value
top 20% vs bottom 20%	1.19	0.0500
top 20% vs bottom 20-40%	1.26	0.0160
top 20% vs bottom 40-60%	1.24	0.0205
top 20% vs bottom 60-80%	1.20	0.0484

In fact, the top 20% vs bottom 20% comparison tells us more…

The bottom 20% was likely accepted to Vanderbilt based on undergrad grades, reference letters, research statements, interviews, etc.

So, these other considerations are less informative than GRE

A simple model that explains GRE scores distribution

"A" students master 7-10 topics

"B" students master 6-9 topics

"C" students master 5-8 topics

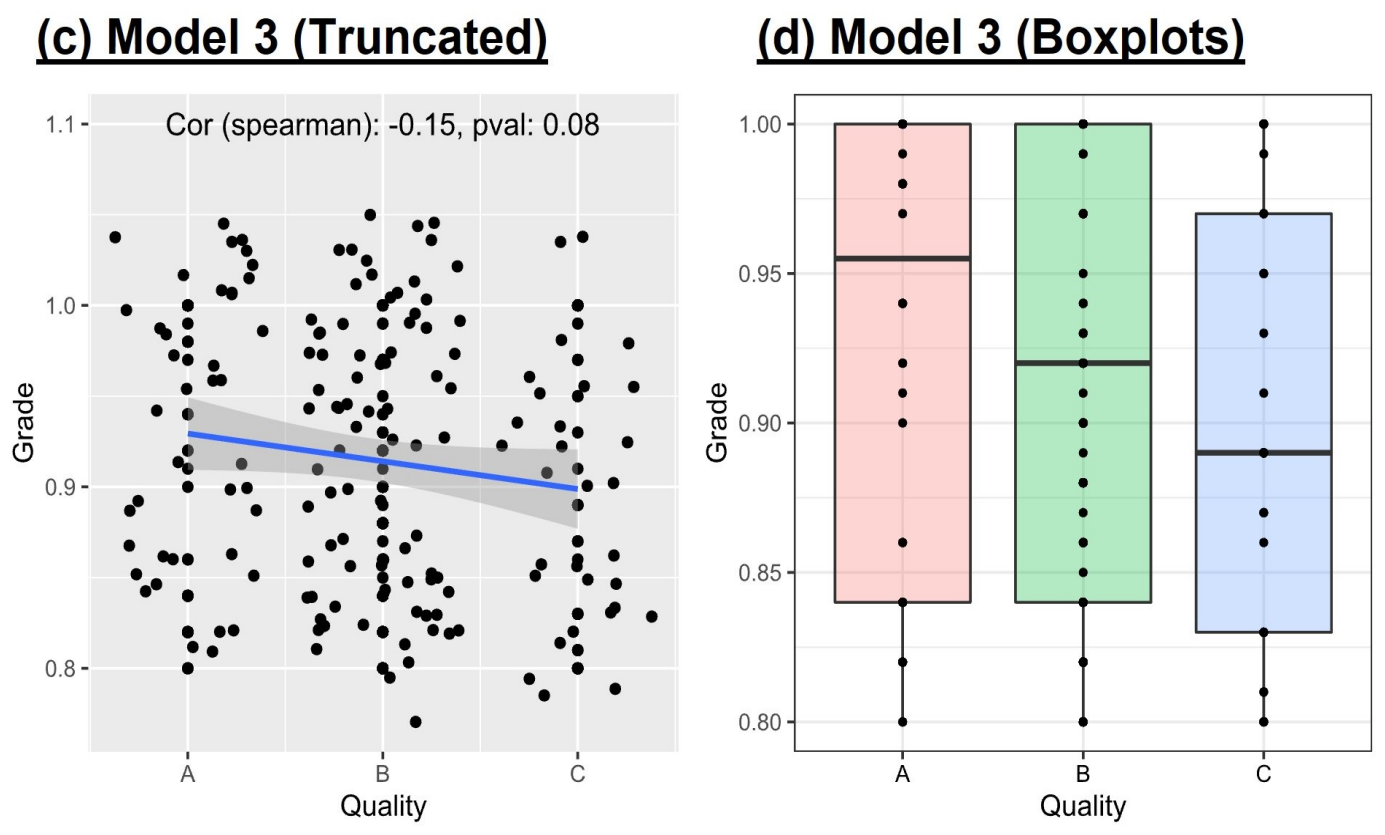
Exam covers 5 out of 10 topics, each topic is worth 20 marks. Then,

"A" students get 40-100 marks, P(80+|A) = 81%

"B" students get 20-100 marks, P(80+|B) = 63%

"C" students get 0-100 marks, P(80+|C) = 39%

Exercise: Derive P(A|80+), P(B|80+), P(C|80+)



50 "A" students, 150 "B" students, 150 "C" students

No correlation does not imply no association

A common practice in machine learning

Optimize and evaluate based on cross-validation accuracy

Is this a sound advice? Note that a vast majority of AI/machine learning projects fail on deployment

Exercise #3

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

Class proportion of test set is not always fidel to real life; accuracy determined from test set may not give the right picture of real-life performance

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

Test set:

20% +ve, 80% -ve

recall = 80%, precision = 80%

∴specificity = 95%, accuracy = 92%

New test set "real life":

80% +ve, 20% -ve

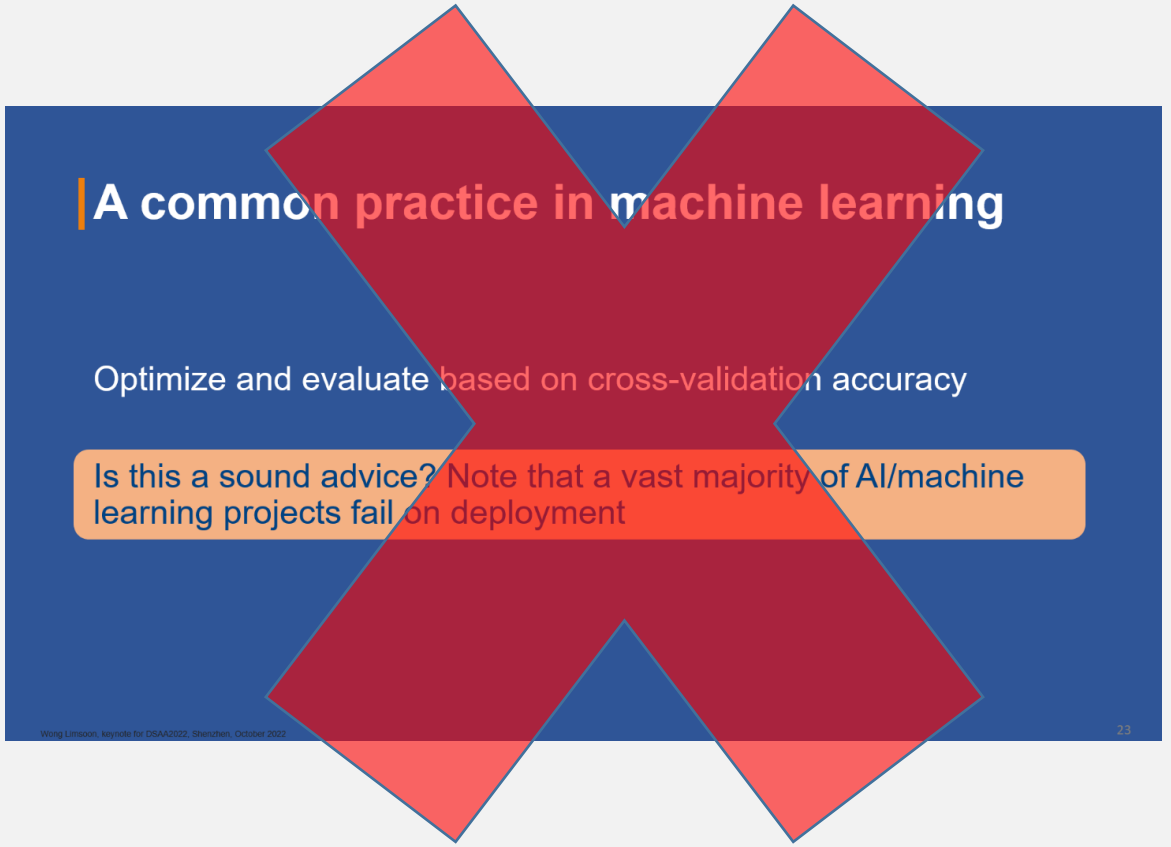
By "representativeness",

recall = 80%, specificity = 95%

∴ accuracy = 83%, precision = 98%

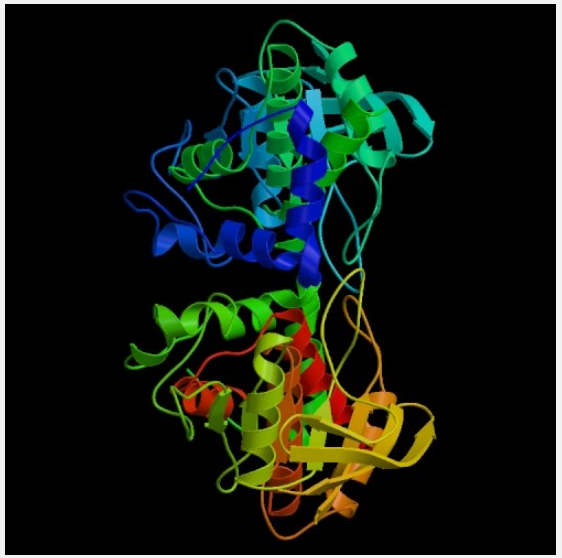
Accuracy measured from a test set must be calibrated for interpretability

Probably better to optimize wrt recall & specificity, as these are independent of class proportion



Protein function assignment

A protein is a large complex molecule made up of one or more chains of amino acids



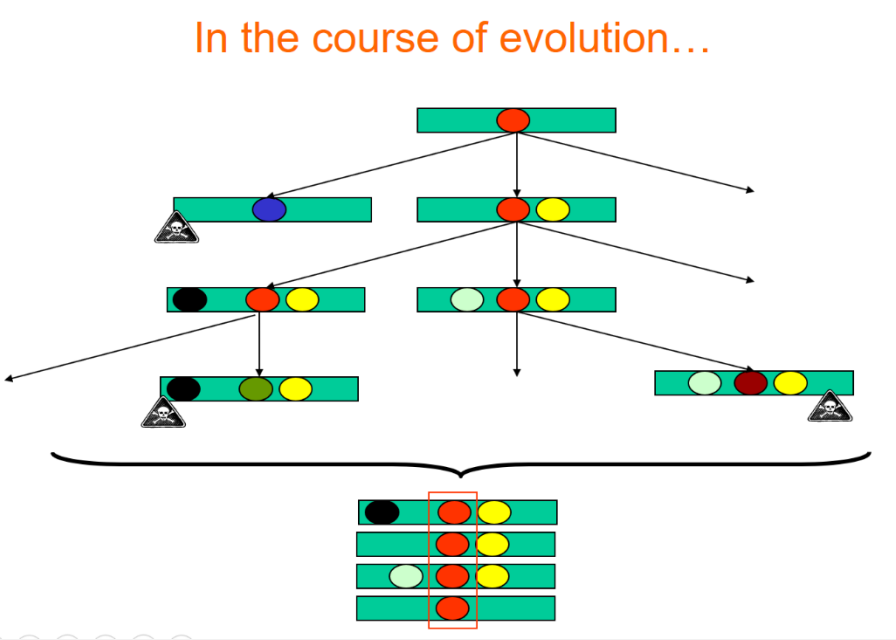
Usually, only the sequence of amino acid is known

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

Proteins perform a wide variety of activities in the cell

How do we predict the function of a protein?

A standard postulate based on evolution



Two proteins (not) inheriting their function from a common ancestor (do not) have similar amino acid sequences

Guilt by association

Compare *T* with seqs of known function in a db

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                    60        70        80        90       100
Amicyanin           MPHNVHFVAGVLGEAALKGPMMKKEQAYSLFFTEAGTYDYHCTPHPFMRGKVVV
                                                :.:  . ::. ::
Ascorbate Oxidase   ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                      70        80        90       100       110
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|  unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|  unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPWAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSVLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

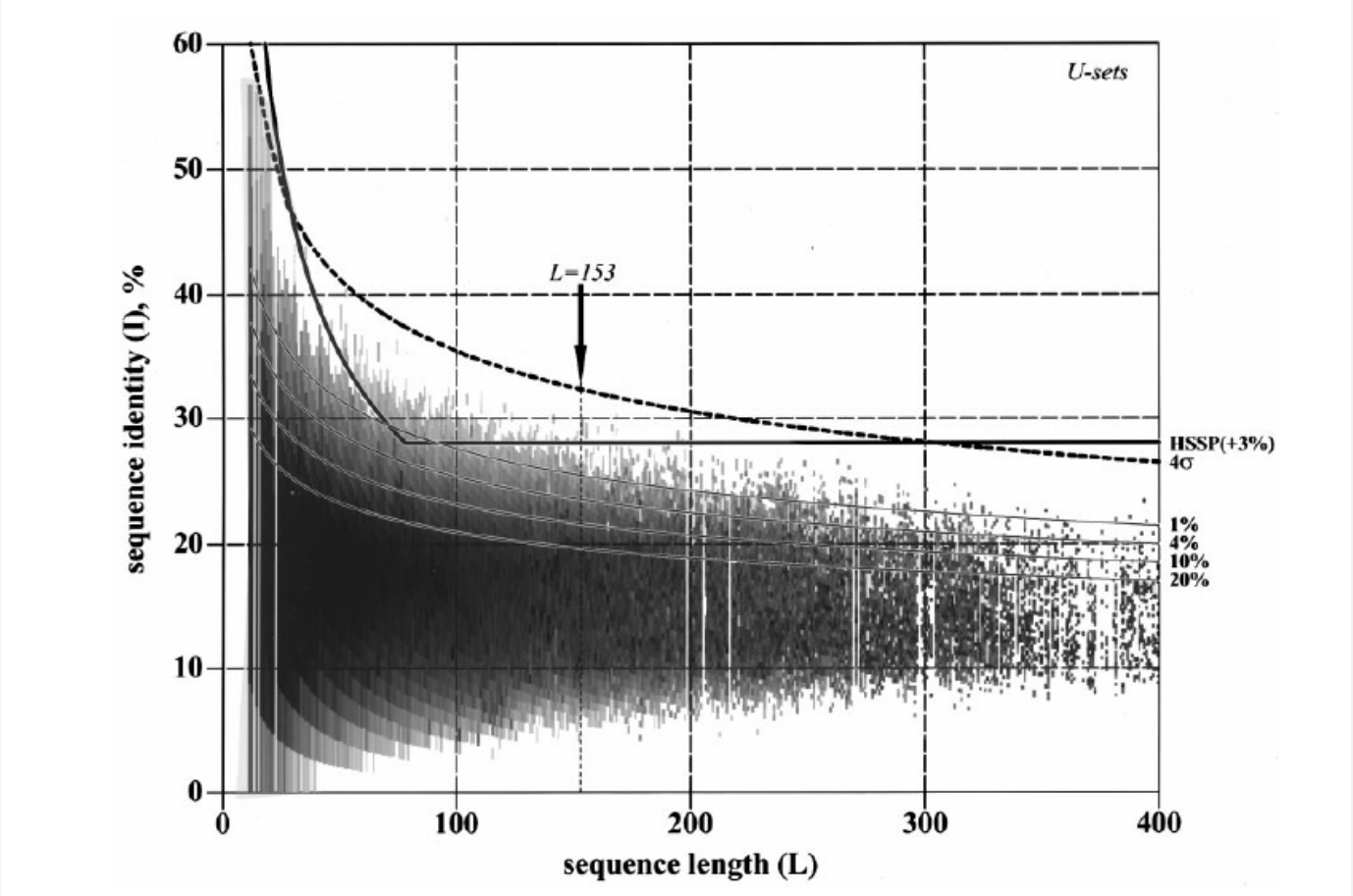
good match between
Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Twilight zone: Limit of sequence similarity-based protein function assignment

So, need clever methods for the twilight zone



Abagyan RA, Batalov S. *J Mol Biol.,* 273(1):355-68, 1997

DeepFam, deep learning for protein family prediction

This looks good

Really?

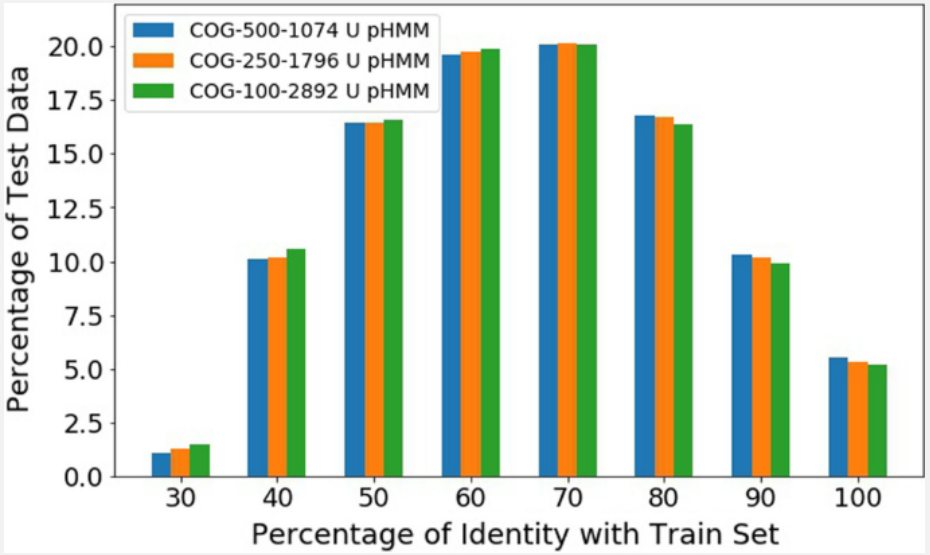
Table 2. Prediction accuracy (%) comparison of COG dataset

Dataset	COG-500-1074	COG-250-1796	COG-100-2892
DeepFam	**95.40**	**94.08**	91.40
pHMM	91.75	91.78	**91.67**
3-mer LR	85.59	81.15	75.44
Protvec LR	47.34	41.76	37.05

Bold indicates the best performance for each dataset.

Seo et al., *Bioinformatics,* 34(13):i254-i262, 2018

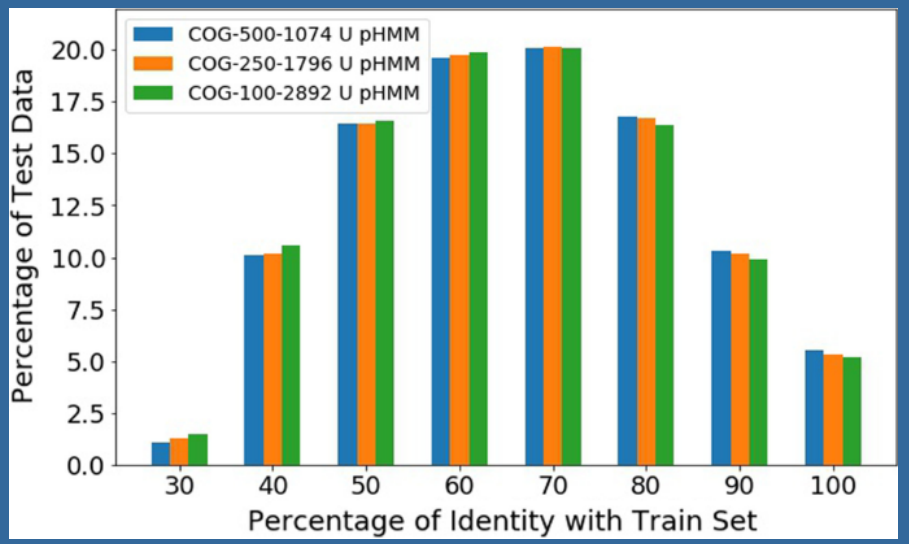
DeepFam's good accuracy is largely due to "easy" proteins… it doesn't advance the field



Dataset	Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
Identity: $0 < x \leq 30$							
COG-500-1074	EnsembleFam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**
	pHMM	69.54	73.75	55.51	70.62	70.85	73.55
	DeepFam	57.14	54.52	49.90	46.92	43.64	35.94
COG-250-1796	EnsembleFam	72.84	**77.07**	**81.02**	**82.14**	**84.66**	**86.45**
	pHMM	**75.39**	73.82	73.84	71.02	67.44	72.43
	DeepFam	32.44	32.54	30.24	29.53	30.02	28.68
COG-100-2892	EnsembleFam	**75.24**	**79.55**	**81.21**	**80.63**	**82.05**	**88.95**
	pHMM	63.44	59.69	53.45	48.16	47.42	57.57
	DeepFam	27.30	26.13	25.54	27.62	24.83	25.36

Neamul Kabir & Wong, *BMC Bioinformatics,* 23:90, 2022

If there are few twilight zone proteins in real life, maybe DeepFam's poor twilight zone performance is ok?



The reference database comprises proteins with known function

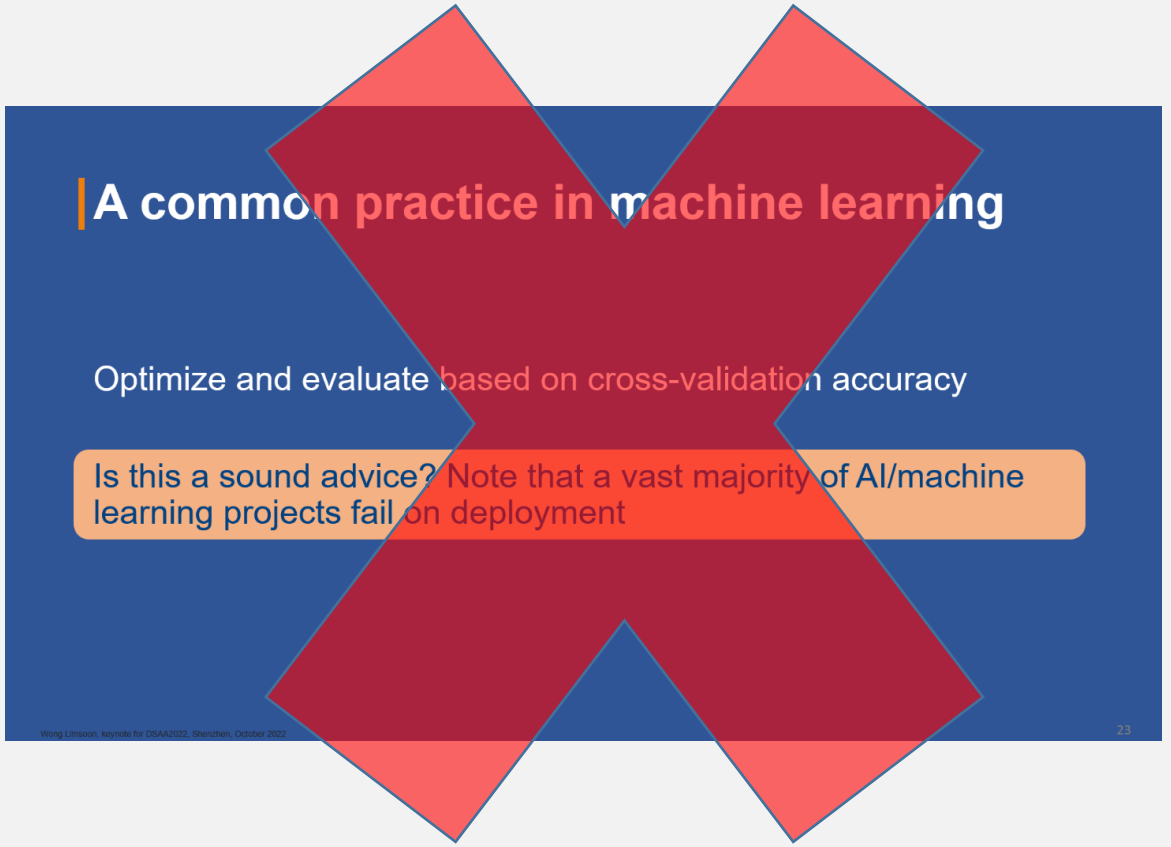
If no function is predicted for a protein, or a wrong function is predicted, there won't be any validated result for the protein

∴ Few twilight zone proteins can get into the reference database

Catch-22 !

Don’t be fooled by high accuracy on easy test sets

Need to stratify accuracy wrt easy and hard test instances



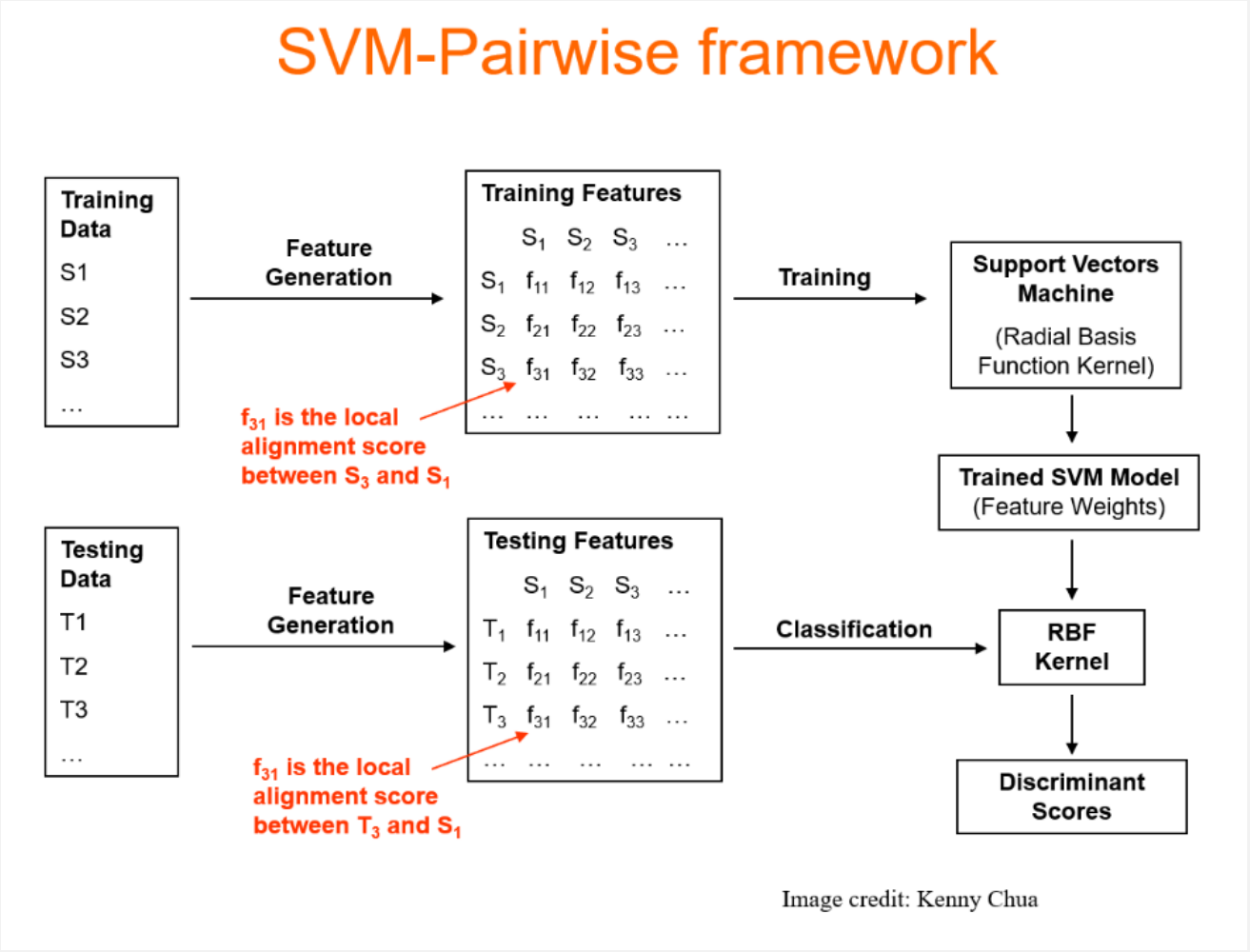
Exercise #4

How did EnsembleFam achieve its superior performance in the twilight zone?

Dataset	Method	predCount = 1	predCount = 2	predCount = 3	predCount = 4	predCount = 5	predCount > 5
Identity: $0 < x \leq 30$							
COG-500-1074	EnsembleFam	**72.07**	**81.00**	**82.82**	**84.96**	**85.33**	**85.27**
	pHMM	69.54	73.75	55.51	70.62	70.85	73.55
	DeepFam	57.14	54.52	49.90	46.92	43.64	35.94
COG-250-1796	EnsembleFam	72.84	**77.07**	**81.02**	**82.14**	**84.66**	**86.45**
	pHMM	**75.39**	73.82	73.84	71.02	67.44	72.43
	DeepFam	32.44	32.54	30.24	29.53	30.02	28.68
COG-100-2892	EnsembleFam	**75.24**	**79.55**	**81.21**	**80.63**	**82.05**	**88.95**
	pHMM	63.44	59.69	53.45	48.16	47.42	57.57
	DeepFam	27.30	26.13	25.54	27.62	24.83	25.36

EnsembleFam uses low-/dis-similarity information discarded by other methods!

Inspired by SVM-pairwise



Neamul Kabir & Wong, *BMC Bioinformatics,* 23:90, 2022

Exercise #5

What qualities should a test set have?

Representativeness

Absence of doppelgangers

Etc.

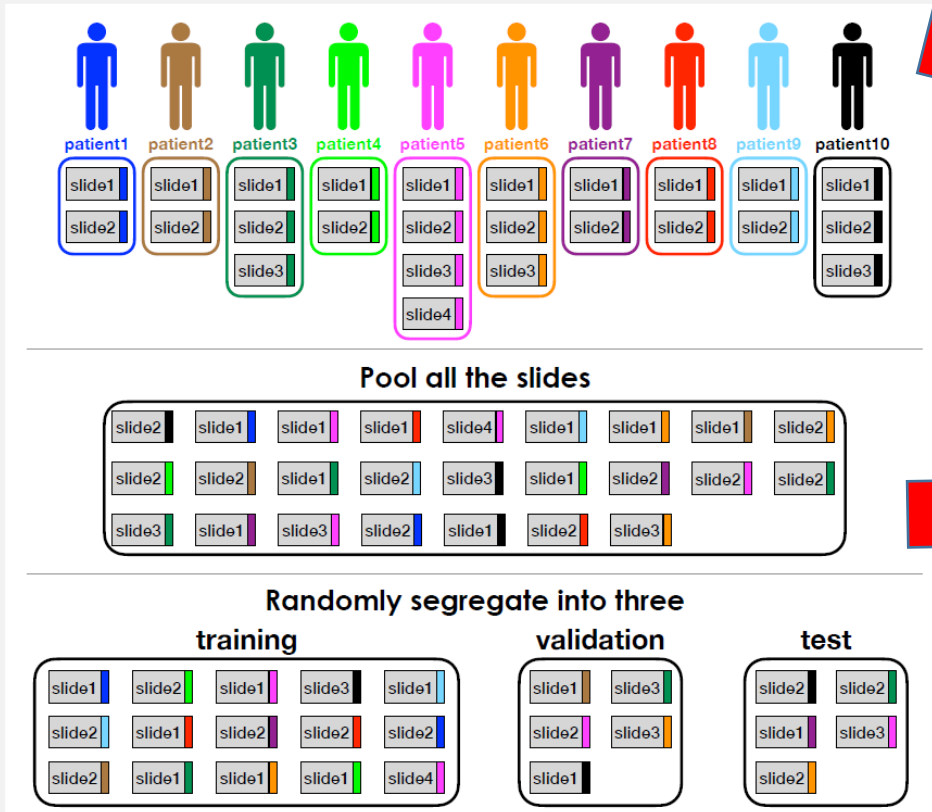
Do you / your students / your professors ever check test sets for these qualities?

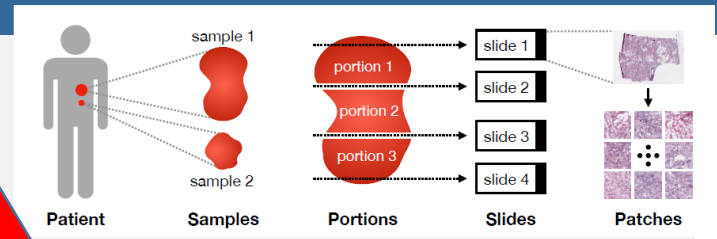
Exercise #6: Deep learning from histopath images for lung cancer diagnosis

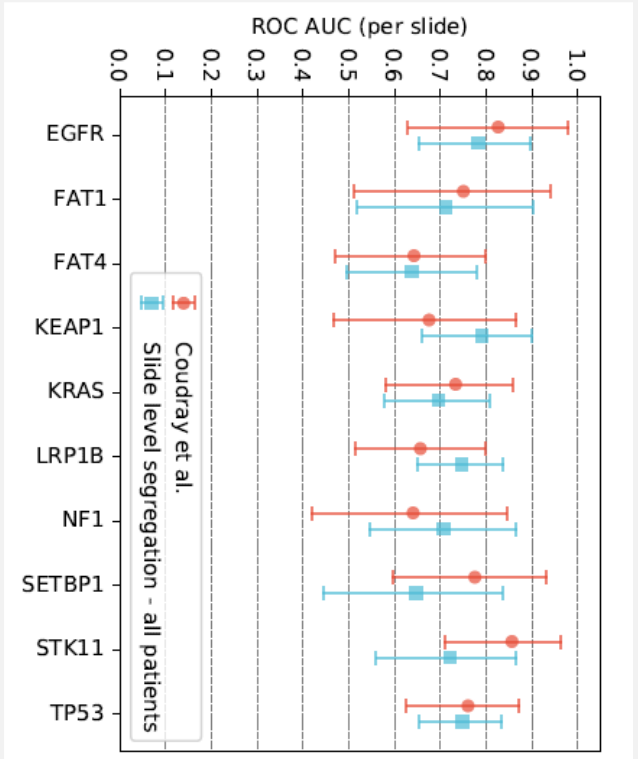
Coudray et al. report that common mutations in lung cancers can be predicted from histopath images using deep learning

Is this claim sound based purely on their results?

Coudrayet al, *Nat Med* 24:1559-1567, 2018



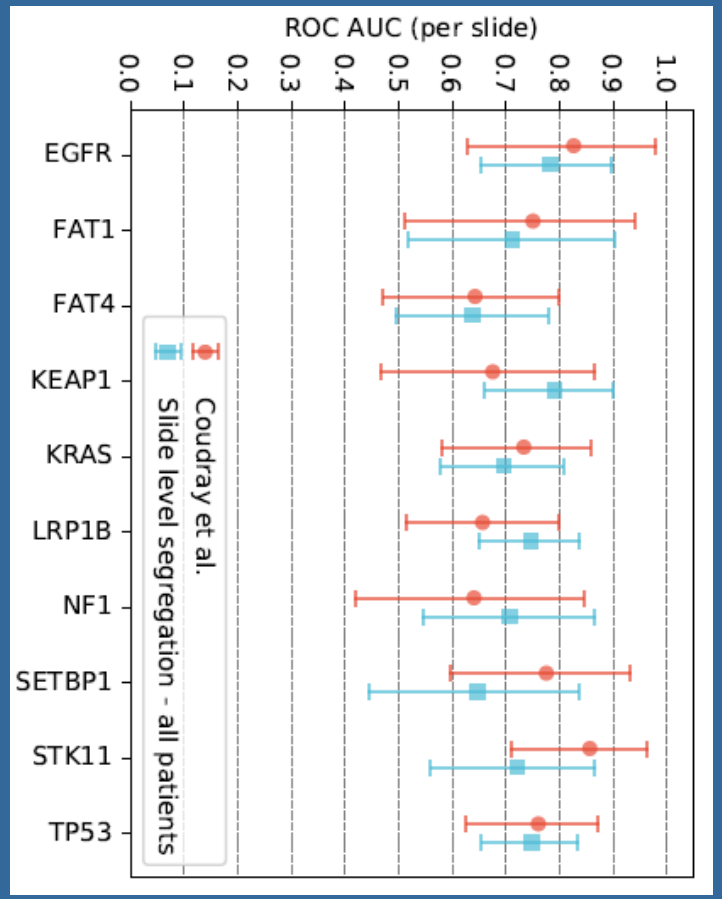


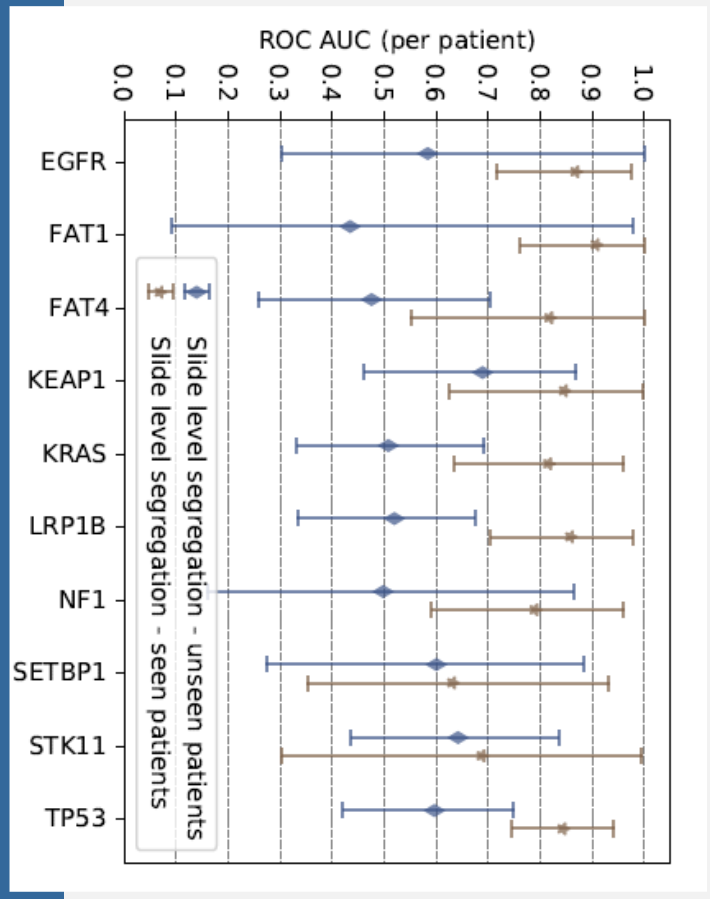


Wong Limsoon, keynote for DSAA2022, Shenzhen, October 2022

Image credit: Mustafa Umit Oner

Doppelgangers





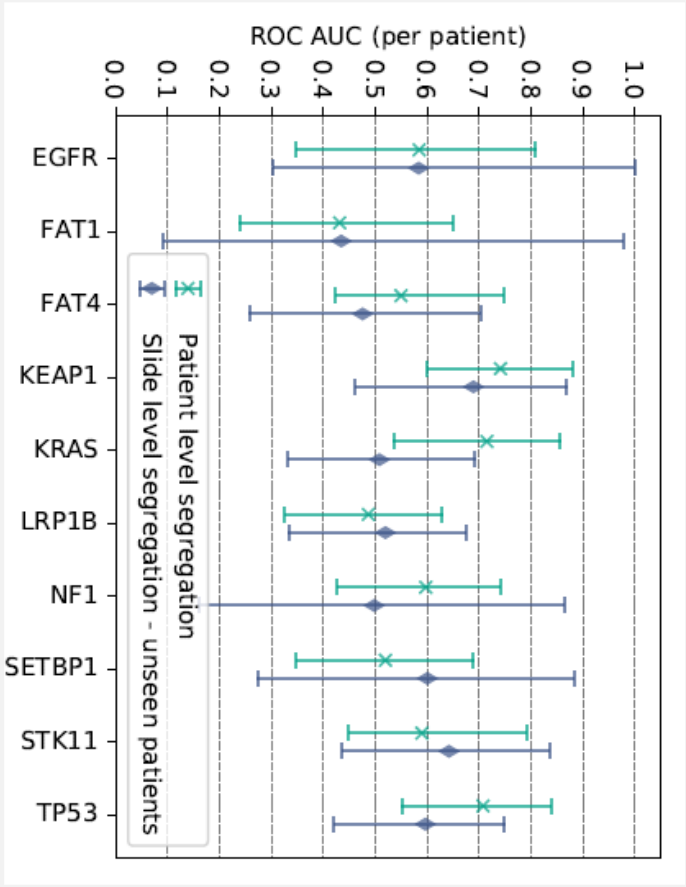
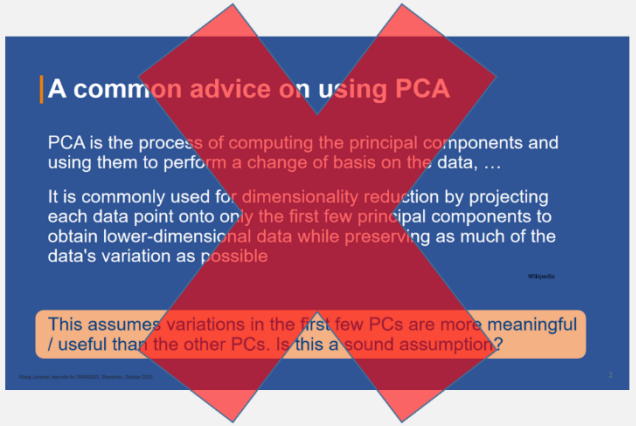
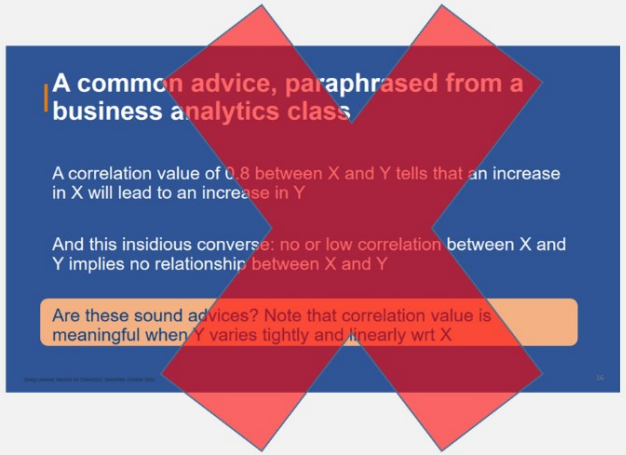


Image credit: Mustafa Umit Oner

Summary





PCA is not about dimensionality reduction

It concerns deconvolution of variations

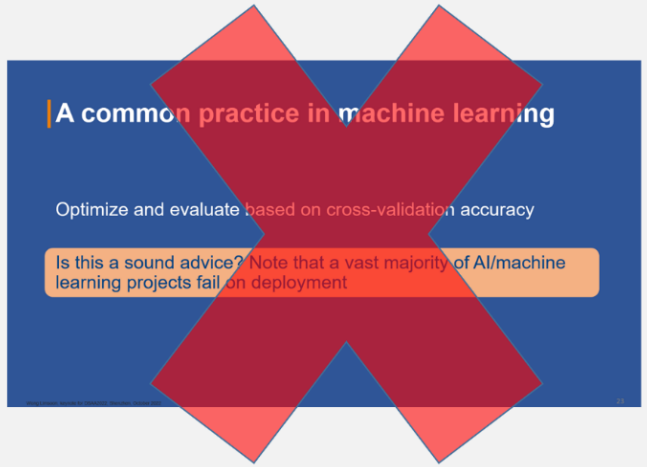
Top PCs may be irrelevant or confounding

Bottom PCs can be relevant and informative

High correlation may not imply association

No correlation does not imply no association

Summary



Accuracy measured from a test set must be calibrated wrt prevalence for interpretability

Accuracy should be “stratified” wrt easy and hard test cases

Discarded information can be very useful; cf. the informativeness of low PCs

Beware that test sets may not meet quality requirements or may not be used properly

References

Giuliani et al., “On the constructive role of noise in spatial systems”, *Physics Letters A*, 247:47-52, 1998

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics”, *BMC Genomics*, 18(Suppl 2):142, 2017

Neamul Kabir & Wong, “EnsembleFam: Towards more accurate protein family prediction in the twilight zone”, *BMC Bioinformatics*, 23:90, 2022

Goh et al., “What can scatterplots teach us about doing data science better?”, *International Journal of Data Science and Analysis,* in press