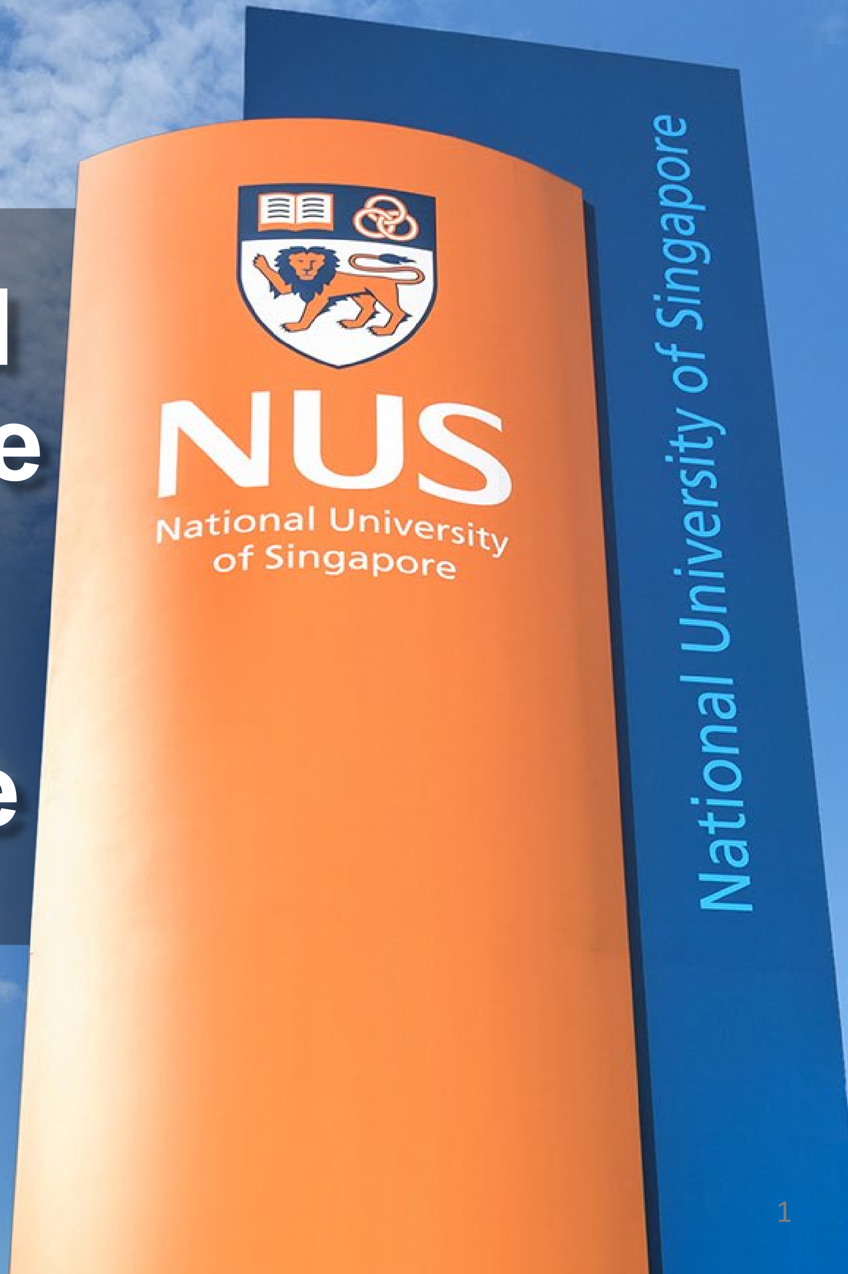
Some *opinion* and advice on machine learning in population-based genomic medicine

Wong Limsoon


NUS
National University of Singapore
National University of Singapore

A confession

I use machine learning in very limited ways these days

If you properly

- Resolve batch effects
- Control confounding factors
- Use informative features

Then any simple analysis methods (including machine learning methods) give equally good results

Machine learning currently has quite weak validation practices

A “black box” produced by a machine learning method may not be what you think it is

In the GWAS context

If you properly

- Resolve batch effects
- Control confounding factors
- Use informative features

Then any simple analysis methods (including machine learning methods) give equally good results

Resolving batch effects

Not an issue, as not much batch effects

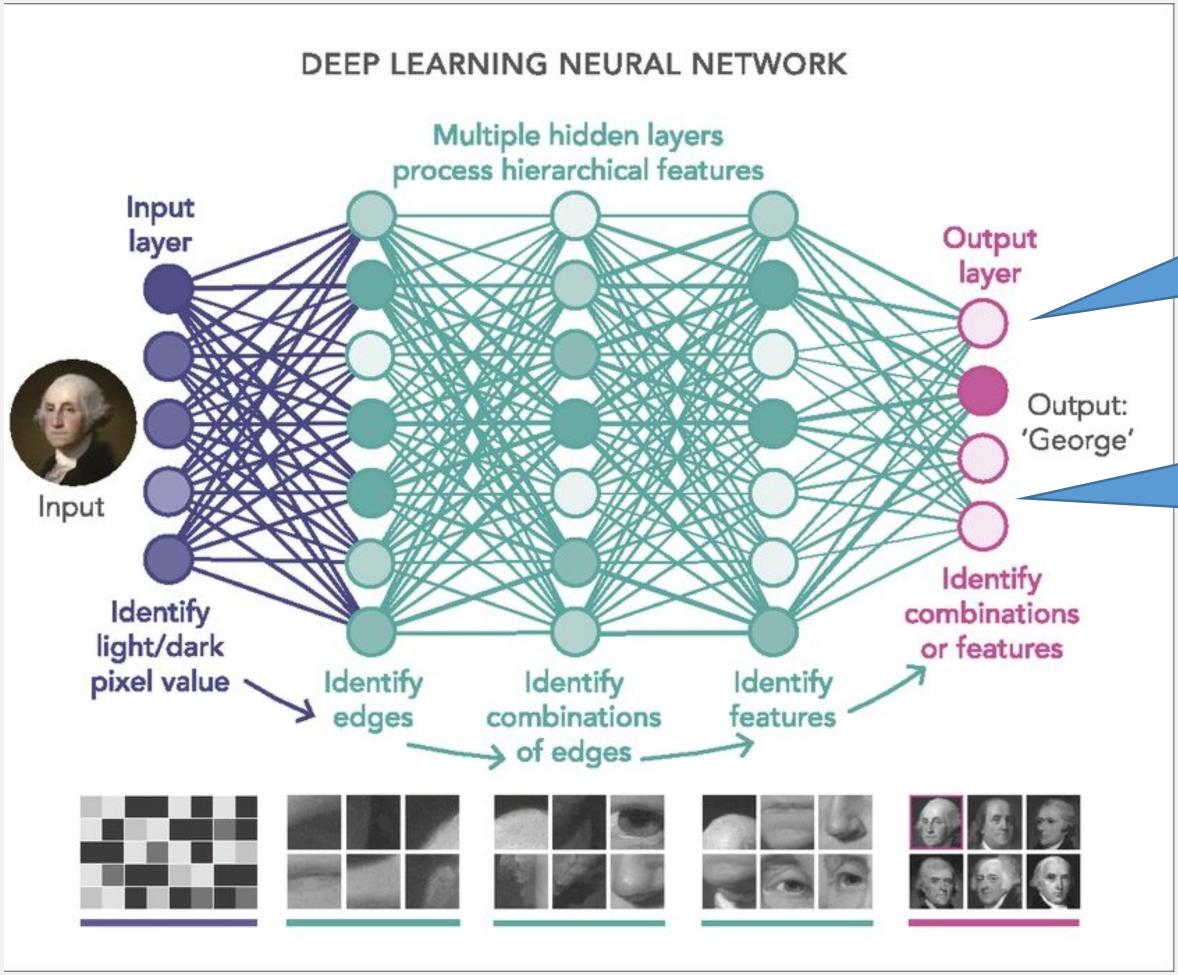
Controlling confounding factors, e.g. population structure

An issue, somewhat solved by stratification, sample selection, etc.

Using informative features

An issue, we are still using features with diluted info

And is exacerbated when using machine learning in some cases


DEEP LEARNING NEURAL NETWORK
Multiple hidden layers process hierarchical features
Input layer
Output layer
Output: 'George'
Input
Identify light/dark pixel value
Identify edges
Identify combinations of edges
Identify features
Identify combinations or features

Because neural networks can learn high-level features from low-level inputs, we get lazy…

But no idea what configurations of nodes and edges in a neural network are needed to learn what features

Features with diluted information are often used in machine learning

In the context of GWAS

SNPs are de facto features

They have "structures" (in the same gene, pathway, etc.)

They have "interactions" (genetic linkage, epistasis, etc.)

Real explanations are often revealed at higher levels

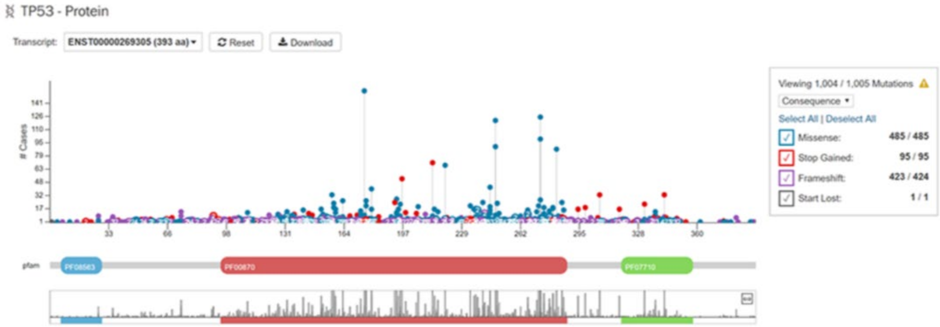
But such higher-level info is often insufficiently exploited, even totally ignored

Good explanations are often revealed at higher levels

TP53 are mutated in as many ways in as many cancer patients

But many patients have mutations in TP53

Mutational processes shape the landscape of TP53 mutations in human cancer.



TP53 - Protein

Transcript: ENST00000269305 (393 aa) ▾ Reset Download

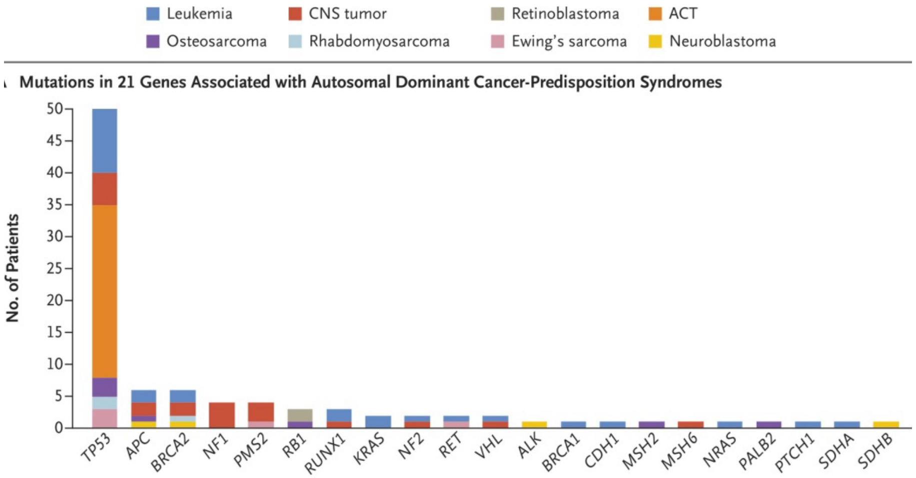
Viewing 1,004 / 1,005 Mutations

Consequence ▾

Select All | Deselect All

☑ Missense:	485 / 485
☑ Stop Gained:	95 / 95
☑ Frameshift:	423 / 424
☑ Start Lost:	1 / 1

NCI GDC Data Portal (https://portal.gdc.cancer.gov/genes/ENSG00000141510)



Mutations in 21 Genes Associated with Autosomal Dominant Cancer-Predisposition Syndromes

https://www.nejm.org/doi/full/10.1056/NEJMoa1508054

Provide / use higher-level info as much as possible

Machine learning methods have a hard time finding SNP-cancer associations, like the TP53 ones

Confused by noise from millions of SNPs

Diluted as each patient has his own mutations in TP53

Even when TP53 SNPs were found by machine learning methods, they couldn’t tell you these are TP53 ones

These methods see SNP-level (not gene-level) info, since this is what they are provided with

Another confession

I haven't done much work on GWAS these days

But I am thick-skinned

I am going to use this one as my example:

Sharlee Climer, Alan R. Templeton, Weixiong Zhang, "Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis GWAS", *PLoS Comput Biol*, 10(9):1003766, 2014

Missing heritability

Single genetic variations cannot account for much of the heritability of diseases, behaviours, and other phenotypes

Combinatorial interactions may account for a substantial portion of this “missing heritability”

But their discoveries have been difficult

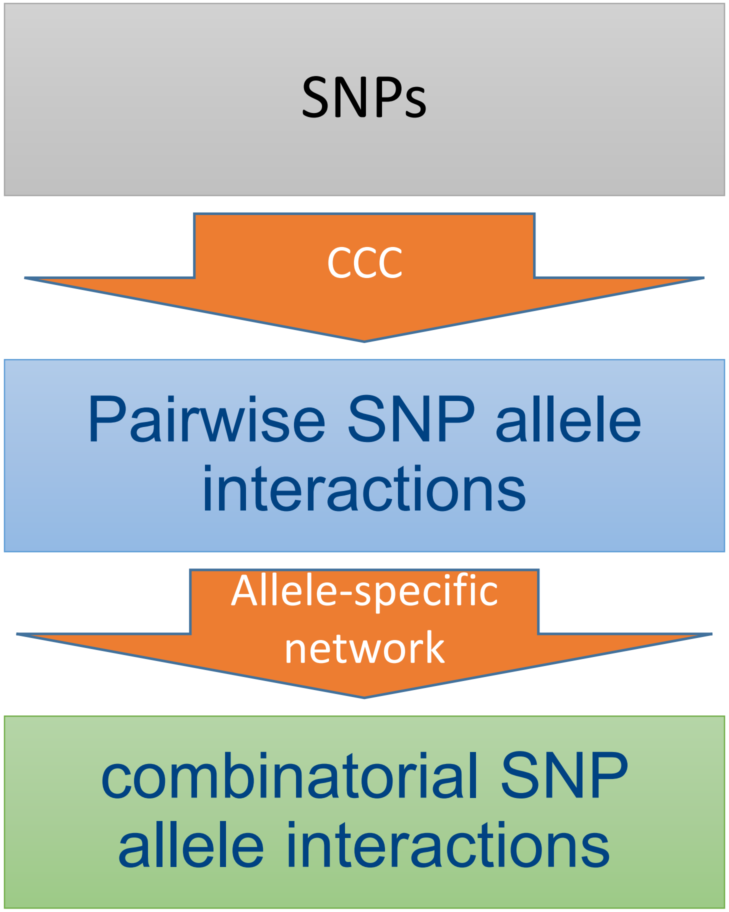
10^{12} pairwise SNP interactions, 10^{18} triplets, etc.

Too many to screen efficiently

Severe multiple testing

Also need to account for “diploid semantics” in the design of a screening metric

From SNPs to higher-level more informative features


SNPs
CCC
Pairwise SNP allele interactions
Allele-specific network
combinatorial SNP allele interactions

Genet Epidemiol, 38(7): 610–621, 2014

Custom correlation coefficient, CCC

CCC is allele specific

$$CCC_{ij} = R_{ij} * F_i * F_j * w$$

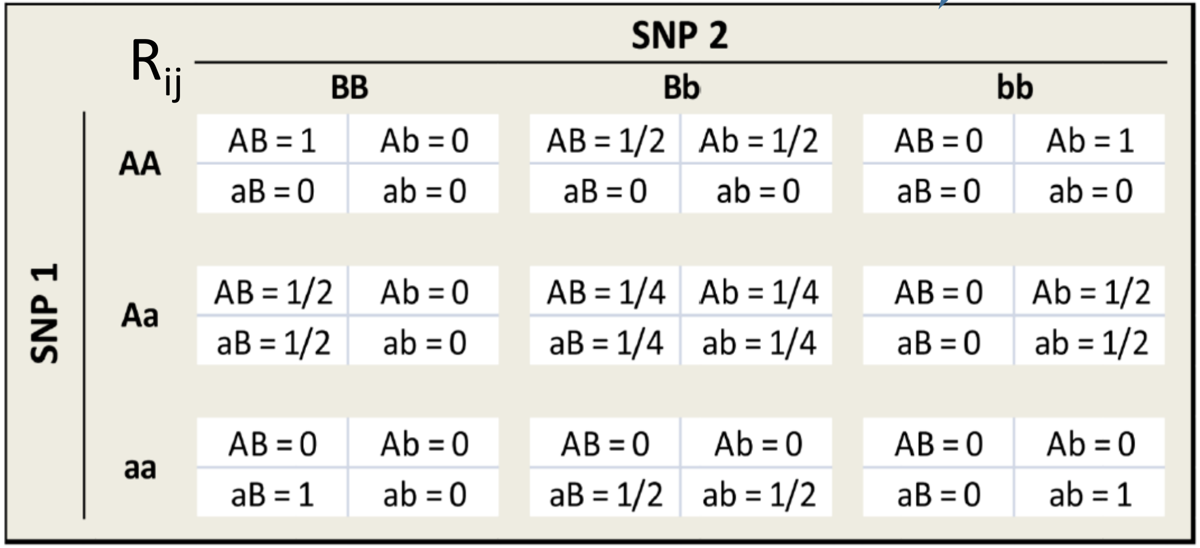
Rare alleles have more weight

"Diploid semantics"

for allele i of SNP1 and allele j of SNP2

F_i, F_j are 1 – frequencies of allele i and j

w is a scaling factor

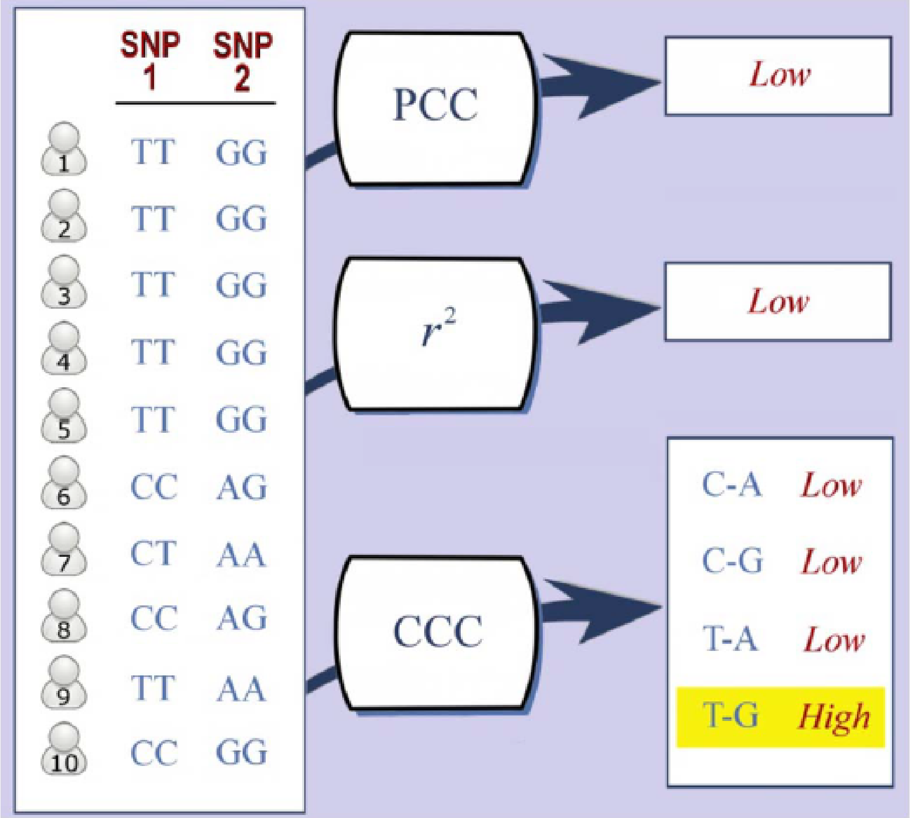


R_{ij}	SNP 2					
SNP 1	BB		Bb		bb	
AA	AB = 1	Ab = 0	AB = 1/2	Ab = 1/2	AB = 0	Ab = 1
	aB = 0	ab = 0	aB = 0	ab = 0	aB = 0	ab = 0
Aa	AB = 1/2	Ab = 0	AB = 1/4	Ab = 1/4	AB = 0	Ab = 1/2
	aB = 1/2	ab = 0	aB = 1/4	ab = 1/4	aB = 0	ab = 1/2
aa	AB = 0	Ab = 0	AB = 0	Ab = 0	AB = 0	Ab = 0
	aB = 1	ab = 0	aB = 1/2	ab = 1/2	aB = 0	ab = 1

CCC is more “sensitive” than PCC and r^2

$$\mathrm{PCC}_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

$$r = \frac{P_{AB}\,P_{ab} - P_{Ab\ PaB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$



	SNP 1	SNP 2
1	TT	GG
2	TT	GG
3	TT	GG
4	TT	GG
5	TT	GG
6	CC	AG
7	CT	AA
8	CC	AG
9	TT	AA
10	CC	GG

Figure 1. Genotypes for ten individuals for a pair of SNPs. The first five individuals are perfectly correlated, but the others are not correlated at all. The absolute value of PCC is 0.3 and r^2 returns 0.0, due to the uncorrelated individuals. CCC supplies four correlation values, each of which corresponds to a specific type of correlation. These values are low for three of the possible combinations, but a high value of 0.7 for the T-G combination was returned.
doi:10.1371/journal.pcbi.1003766.g001

CCC is more efficient than PCC and r^2

$$CCC_{ij} = R_{ij} * F_i * F_j * w$$

n: sample size, m: # of SNPs

F_i is computed once for each SNP allele i in O(n) time

R_{ij} is looked up in O(1) time

CCC_{ij} is computed in O(1) time

∴ CCC complexity = O(m^2 + n)

$$PCC_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

$$r = \frac{P_{AB}\ P_{ab} - P_{Ab\ PaB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

PCC complexity = O(m^2 * n)

r^2 complexity = O(m^2 * n)

∴ CCC is much faster

Sample size of 1,000; CCC is 1,000 times faster than PCC & r^2

Allele-specific psoriasis network analysis

Construct allele-specific network using 929 psoriasis cases and 681 controls in GAIN GRU genome-wide data: 443,020 autosomal SNPs

Nodes are SNP alleles

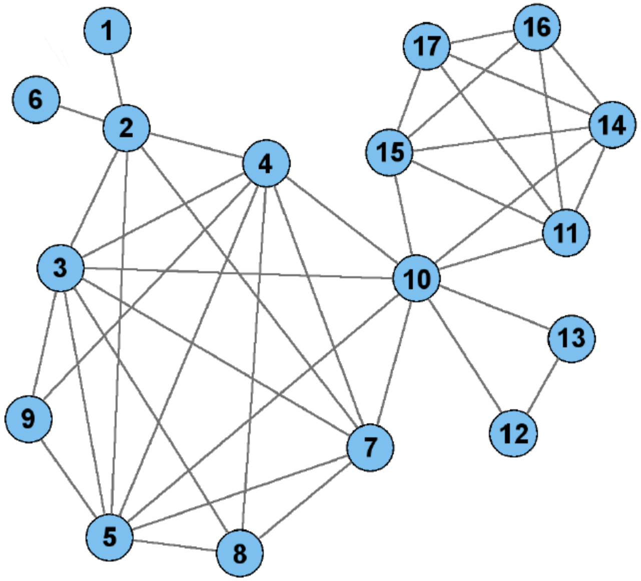
Edges link SNP alleles (i,j) with $CCC_{ij} > \theta$

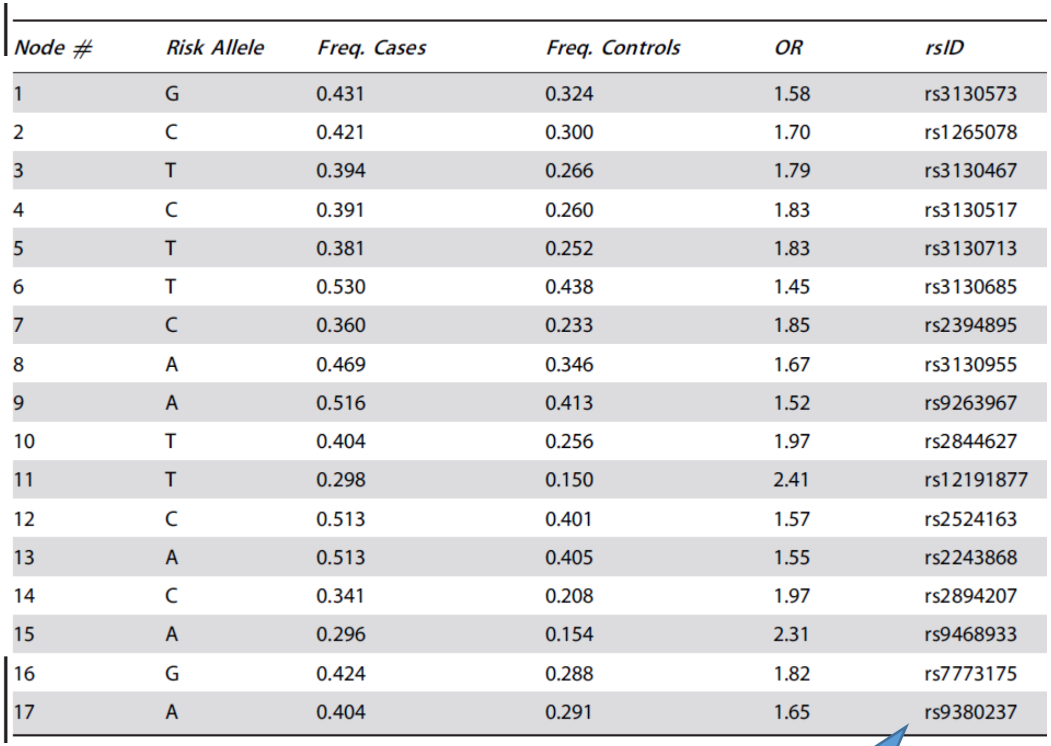
θ is set here so that # nodes = # edges

Each connected component is a combinatorial interaction of SNP alleles

Test it and its complement allele pattern for association with phenotype (psoriasis)

Top connected component, ps1





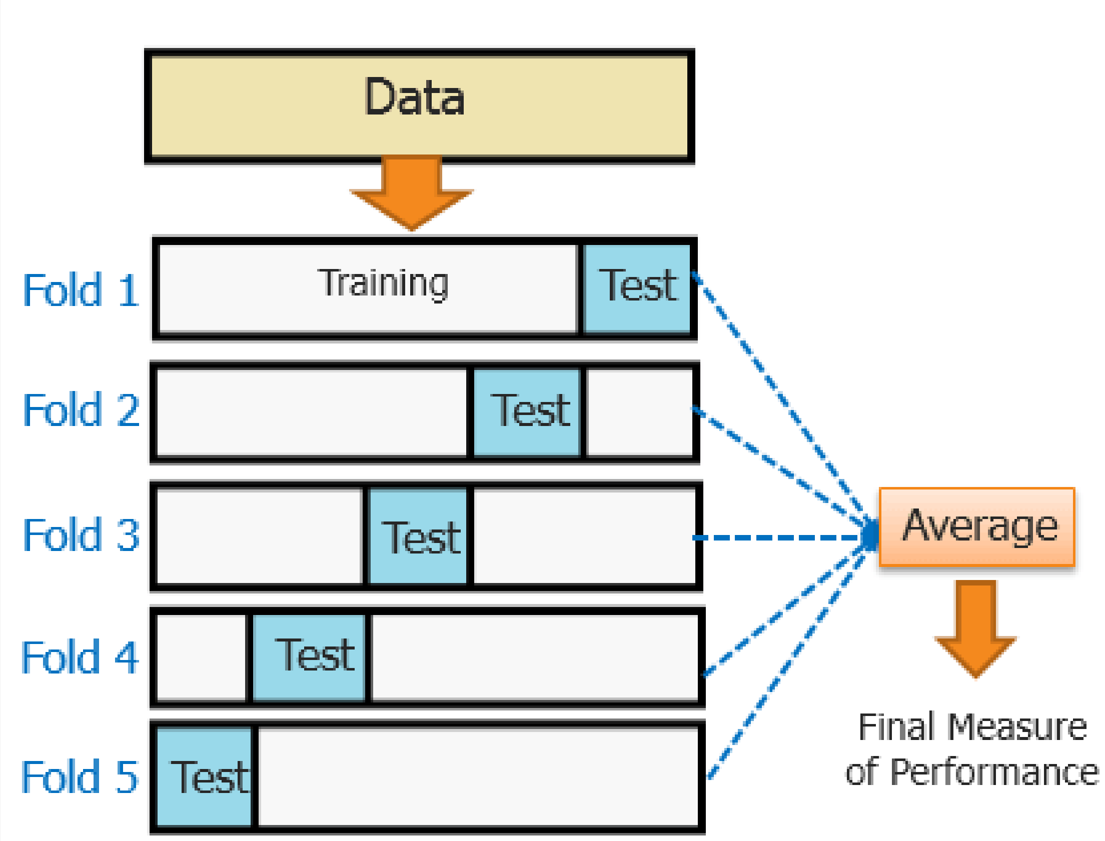
Node #	*Risk Allele*	*Freq. Cases*	*Freq. Controls*	*OR*	*rsID*
1	G	0.431	0.324	1.58	rs3130573
2	C	0.421	0.300	1.70	rs1265078
3	T	0.394	0.266	1.79	rs3130467
4	C	0.391	0.260	1.83	rs3130517
5	T	0.381	0.252	1.83	rs3130713
6	T	0.530	0.438	1.45	rs3130685
7	C	0.360	0.233	1.85	rs2394895
8	A	0.469	0.346	1.67	rs3130955
9	A	0.516	0.413	1.52	rs9263967
10	T	0.404	0.256	1.97	rs2844627
11	T	0.298	0.150	2.41	rs12191877
12	C	0.513	0.401	1.57	rs2524163
13	A	0.513	0.405	1.55	rs2243868
14	C	0.341	0.208	1.97	rs2894207
15	A	0.296	0.154	2.31	rs9468933
16	G	0.424	0.288	1.82	rs7773175
17	A	0.404	0.291	1.65	rs9380237

OR = 3.64 (CI: 2.75--4.80)

P < 5.01 x 10^{-16} (Bonferroni corrected)

Freq in cases: 22%, in control: 7%

3 SNPs in known psoriasis-associated genes (SEEK1, SPR1, HCR)



Machine learning has quite weak validation practices

Computational validations

Phenotype permutations, i.e. null distribution for OR

Genotype permutations, i.e. null distribution for CCC

Boot-strap trials

Independent validation

Phenotype permutations

P-values based on phenotype permutations agree with Bonferroni-corrected p-values

Genotype permutations

Edges unlikely to be false positives

Max CCC in permuted networks = 0.6515

Min CCC in unpermuted network = 0.6949

Boot-strap trials

Ps1 robustly reproduced in 1,000 boot-strap rounds using random 50% of cases and controls

Ave OR = 3.66 (CI: 3.64—3.69)

Ave P < 2.91 x 10^{-11}

Independent validation

Ps1 replicated using GAIN ADO dataset (439 psoriasis cases, 728 controls)

OR = 3.86 (CI: 2.98—5.01)

P < 1.81 x 10^{-25}

Freq in cases: 26%, controls: 8%

Brief comparison w/ PCC

A network constructed using PCC to link SNPs, same # of nodes and edges as CCC network

PCC network is more dispersed ⇒ fewer "believable modules"

Genotype-permuted PCC networks have higher PCC values than the unpermuted network ⇒ more false positives

PCC network took much longer to build

Some caveats

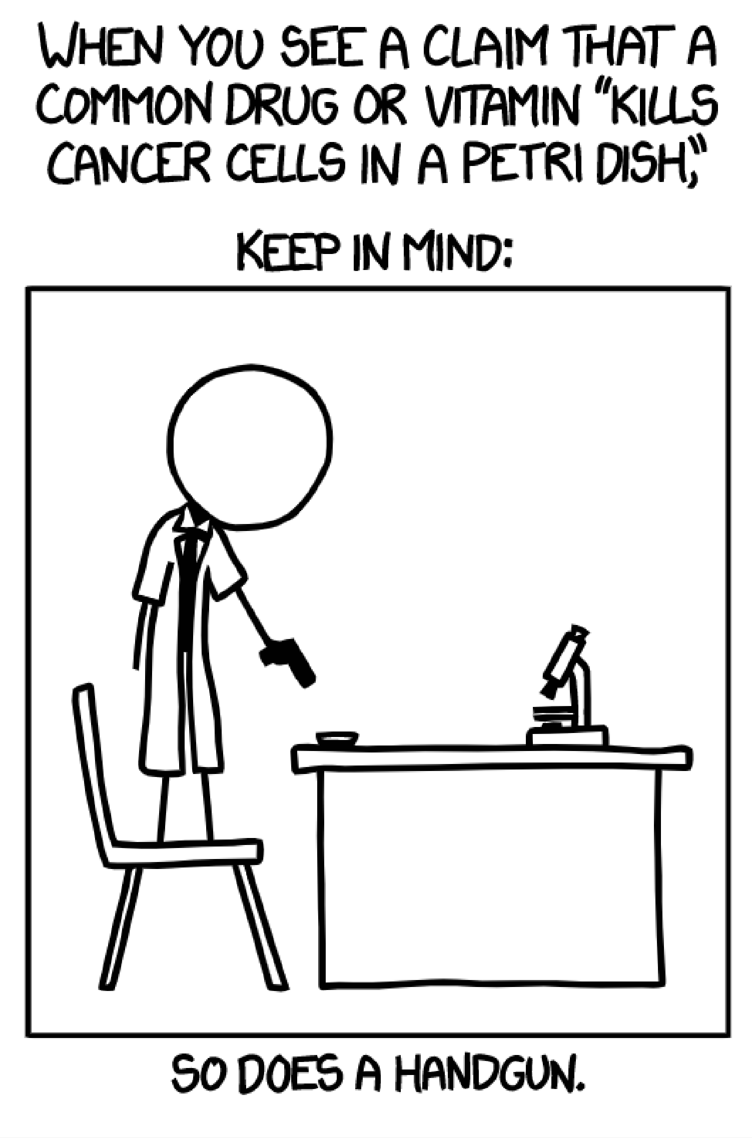
Though CCC is much more efficient to compute that PCC and r^2, it still took ~50 “desktop” days to compute the allele-specific psoriasis network

But parallelizes easily; ran in 1 day on 45 desktops

Didn’t take care of linkage disequilibrium, population structure, etc.

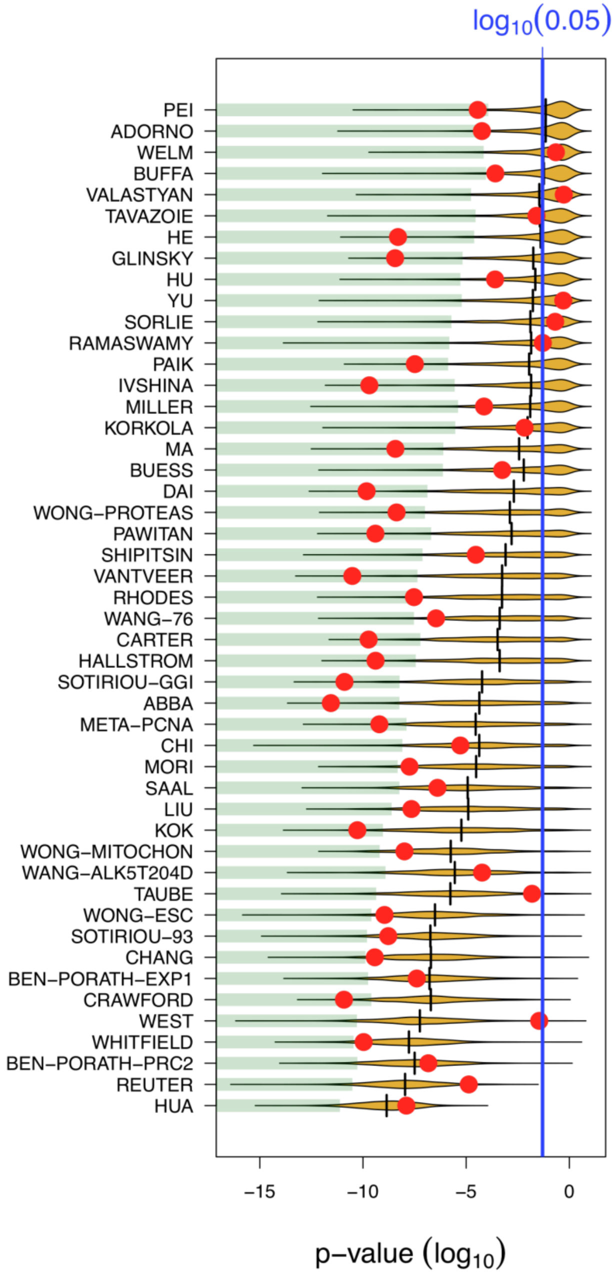
Can do these easily at post-processing

An unrelated story about validation


WHEN YOU SEE A CLAIM THAT A COMMON DRUG OR VITAMIN "KILLS CANCER CELLS IN A PETRI DISH,"
KEEP IN MIND:
SO DOES A HANDGUN.

Anna Karenina effect

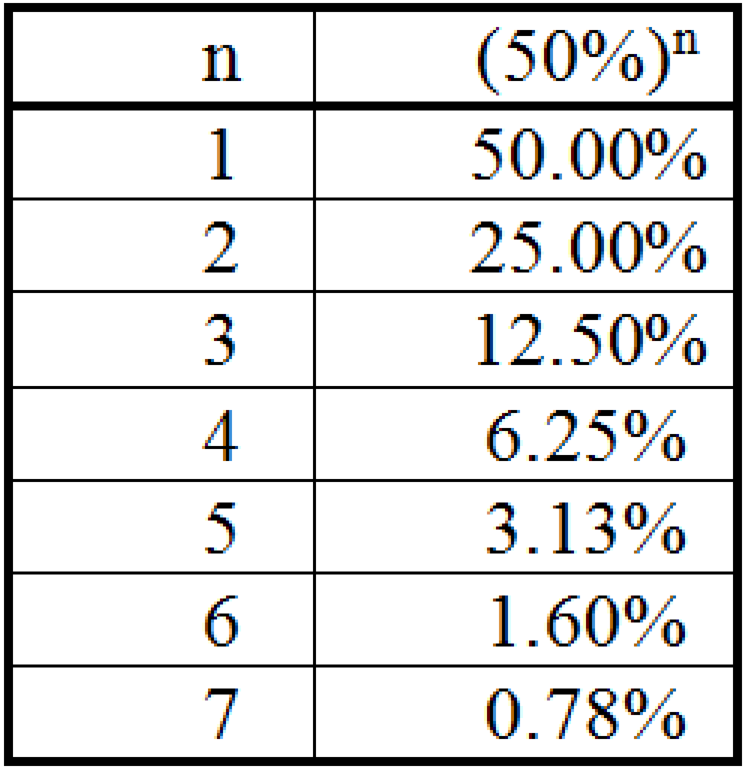
40-50% of random signatures also have p-value << 0.05 on breast cancer datasets



An engineer’s solution to eliminate random signatures

For any independent 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?

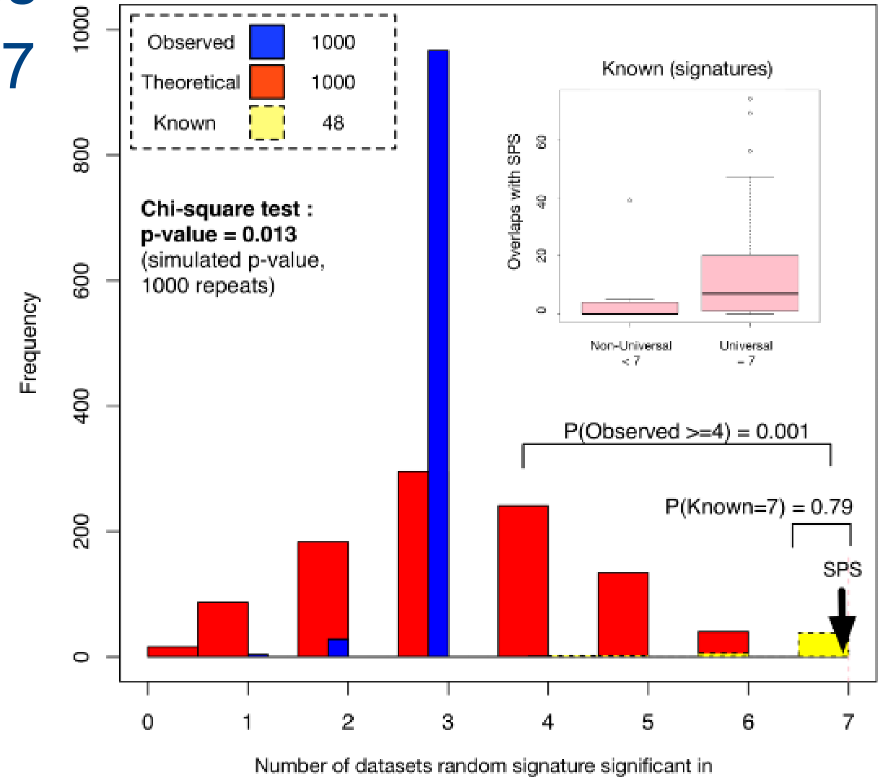


n	$(50\%)^n$
1	50.00%
2	25.00%
3	12.50%
4	6.25%
5	3.13%
6	1.60%
7	0.78%

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



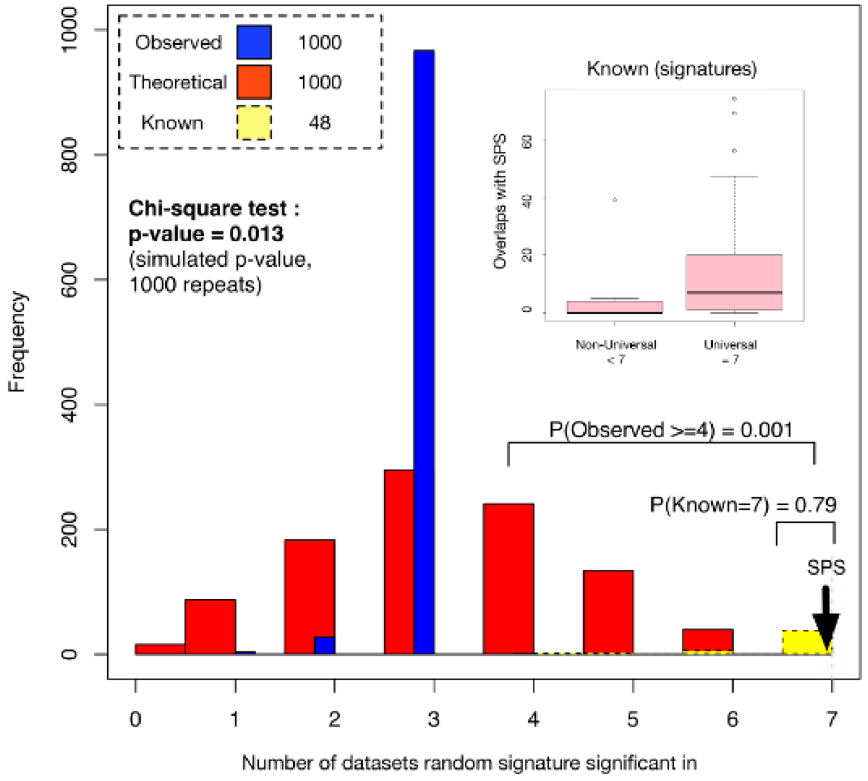
Goh & Wong. Drug Discovery Today, 24(1):31--36, 2019

A theory-practice gap

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

Blue histogram is observed distribution

The independent datasets are less independent than you think!



A “black box” produced by a machine learning method may not be what you think it is

Neural networks: A popular machine learning approach

Do you know what a neural network has learned?

When two neural networks trained on the same training datasets have the same high performance on the same test datasets, have they learned the same thing?

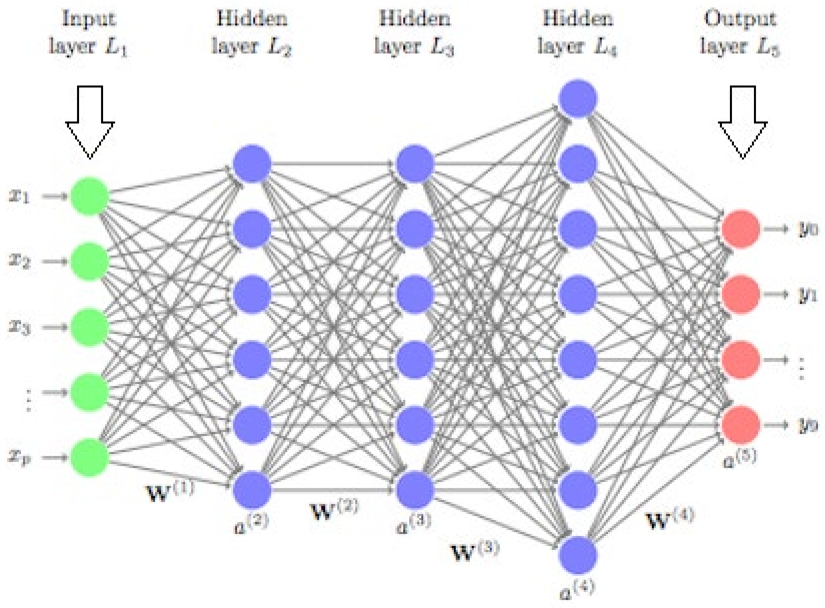
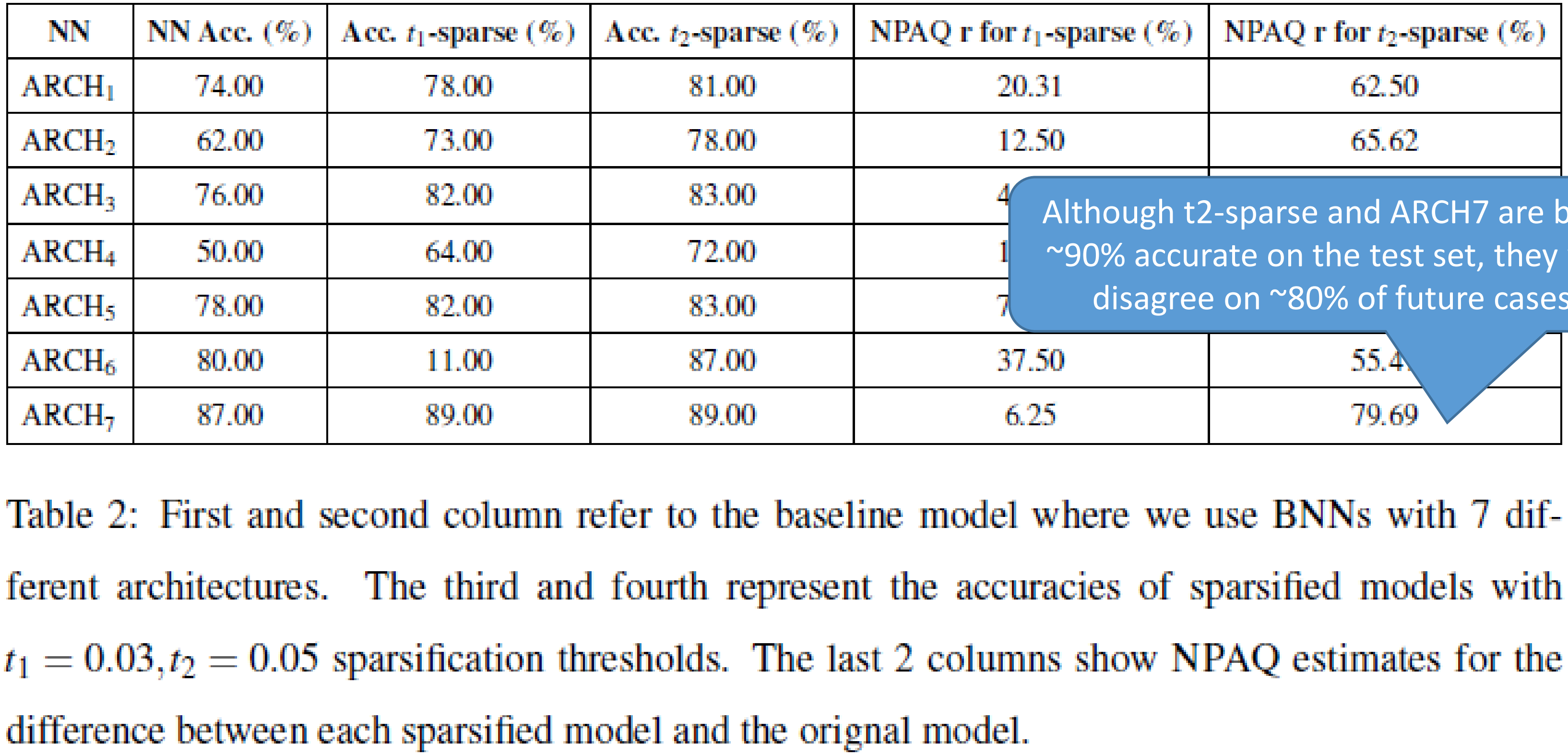


Image credit: University of Cincinnati

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 (%)
$ARCH_1$	74.00	78.00	81.00	20.31	62.50
$ARCH_2$	62.00	73.00	78.00	12.50	65.62
$ARCH_3$	76.00	82.00	83.00	4[illegible]	[illegible]
$ARCH_4$	50.00	64.00	72.00	1[illegible]	[illegible]
$ARCH_5$	78.00	82.00	83.00	7[illegible]	[illegible]
$ARCH_6$	80.00	11.00	87.00	37.50	55.4[illegible]
$ARCH_7$	87.00	89.00	89.00	6.25	79.69

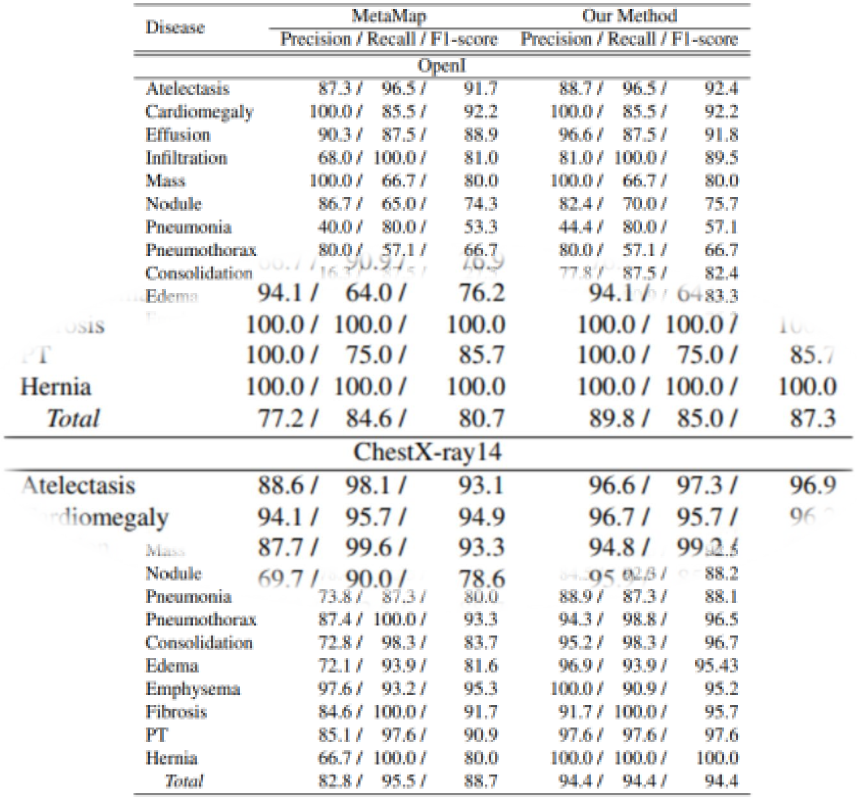
Table 2: First and second column refer to the baseline model where we use BNNs with 7 different architectures. The third and fourth represent the accuracies of sparsified models with $t_1 = 0.03, t_2 = 0.05$ sparsification thresholds. The last 2 columns show NPAQ estimates for the difference between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

Although t2-sparse and ARCH7 are both ~90% accurate on the test set, they will disagree on ~80% of future cases

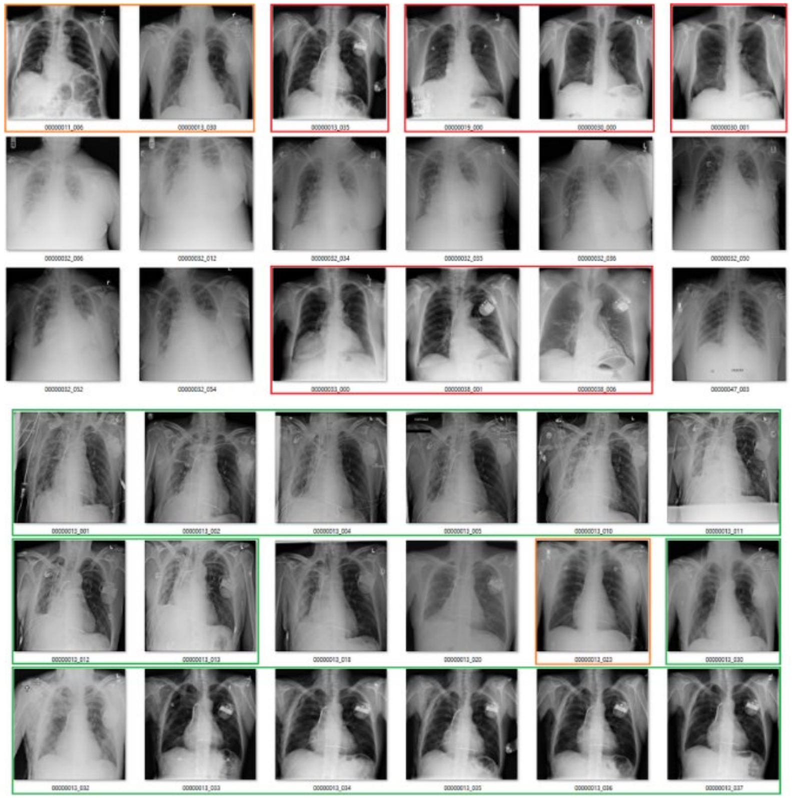
Credit: Teodora Baluta

https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/

A very recent story

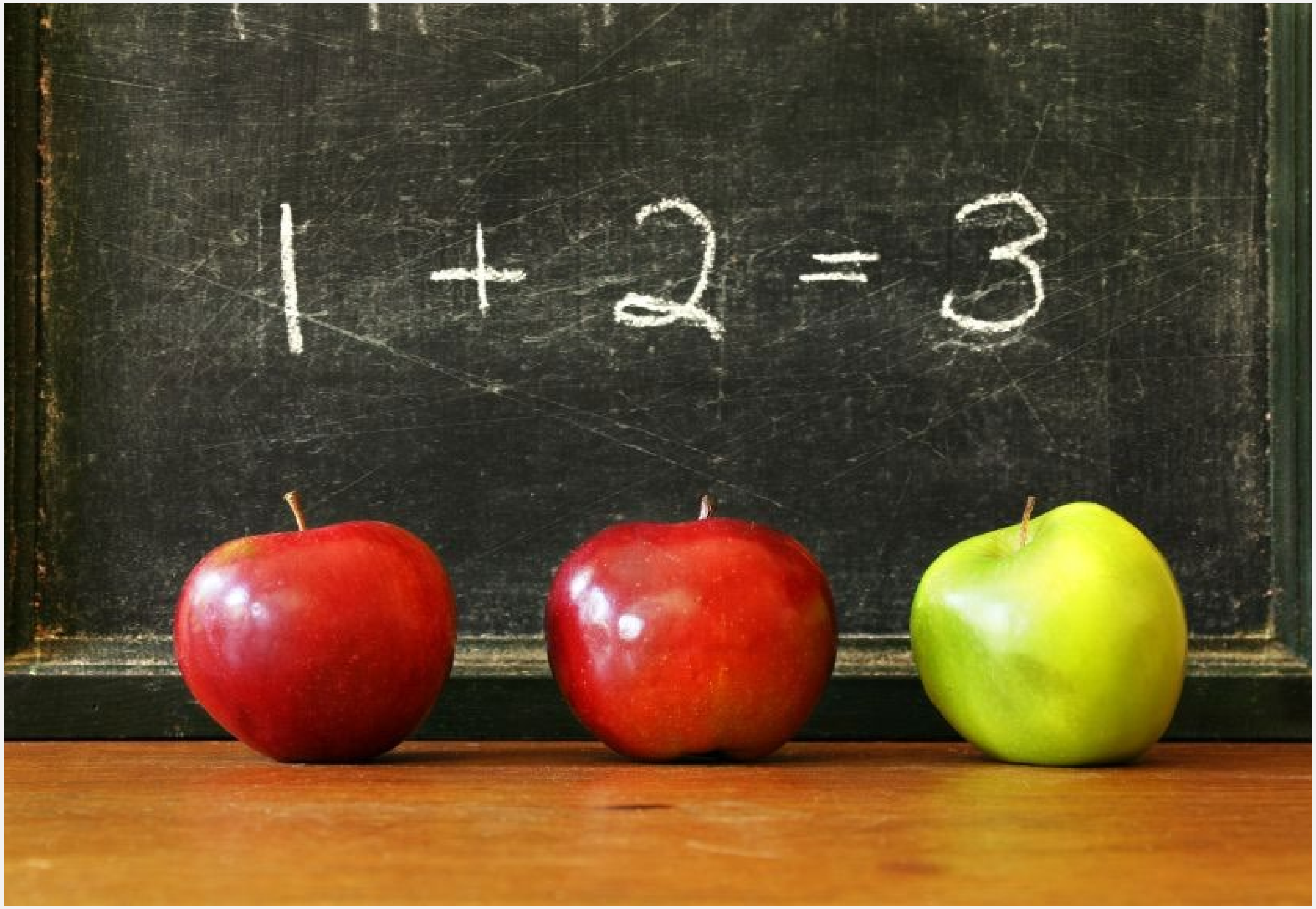


Disease	MetaMap			Our Method		
	Precision /	Recall /	F1-score	Precision /	Recall /	F1-score
			OpenI			
Atelectasis	87.3 /	96.5 /	91.7	88.7 /	96.5 /	92.4
Cardiomegaly	100.0 /	85.5 /	92.2	100.0 /	85.5 /	92.2
Effusion	90.3 /	87.5 /	88.9	96.6 /	87.5 /	91.8
Infiltration	68.0 /	100.0 /	81.0	81.0 /	100.0 /	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	80.0 /	57.1 /	66.7	80.0 /	57.1 /	66.7
Consolidation	[illegible]	[illegible]	[illegible]	77.8 /	87.5 /	82.4
Edema	94.1 /	64.0 /	76.2	94.1 /	64[illegible]	83.3
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.[illegible]
Effusion	87.7 /	99.6 /	93.3	94.8 /	99.2 /	[illegible]
Infiltration	69.7 /	90.0 /	78.6	95.[illegible]	[illegible]	[illegible]
Mass	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	92.5
Nodule	[illegible]	[illegible]	[illegible]	[illegible]	82.3 /	88.2
Pneumonia	73.8 /	87.3 /	80.0	88.9 /	87.3 /	88.1
Pneumothorax	87.4 /	100.0 /	93.3	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

Dataset bias - many pneumothorax cases were patients treated with chest drain



Closing remarks

Closing remarks

Resolve batch effects + control confounding factors + use informative features ⇒ simple analysis methods can give good results

But it takes some understanding to design good features

Current validation practices are quite weak

Put more thoughts into here; test and test again

A “black box” produced by a machine learning method may not be what you think it is

Use w/ caution; avoid unless no choice