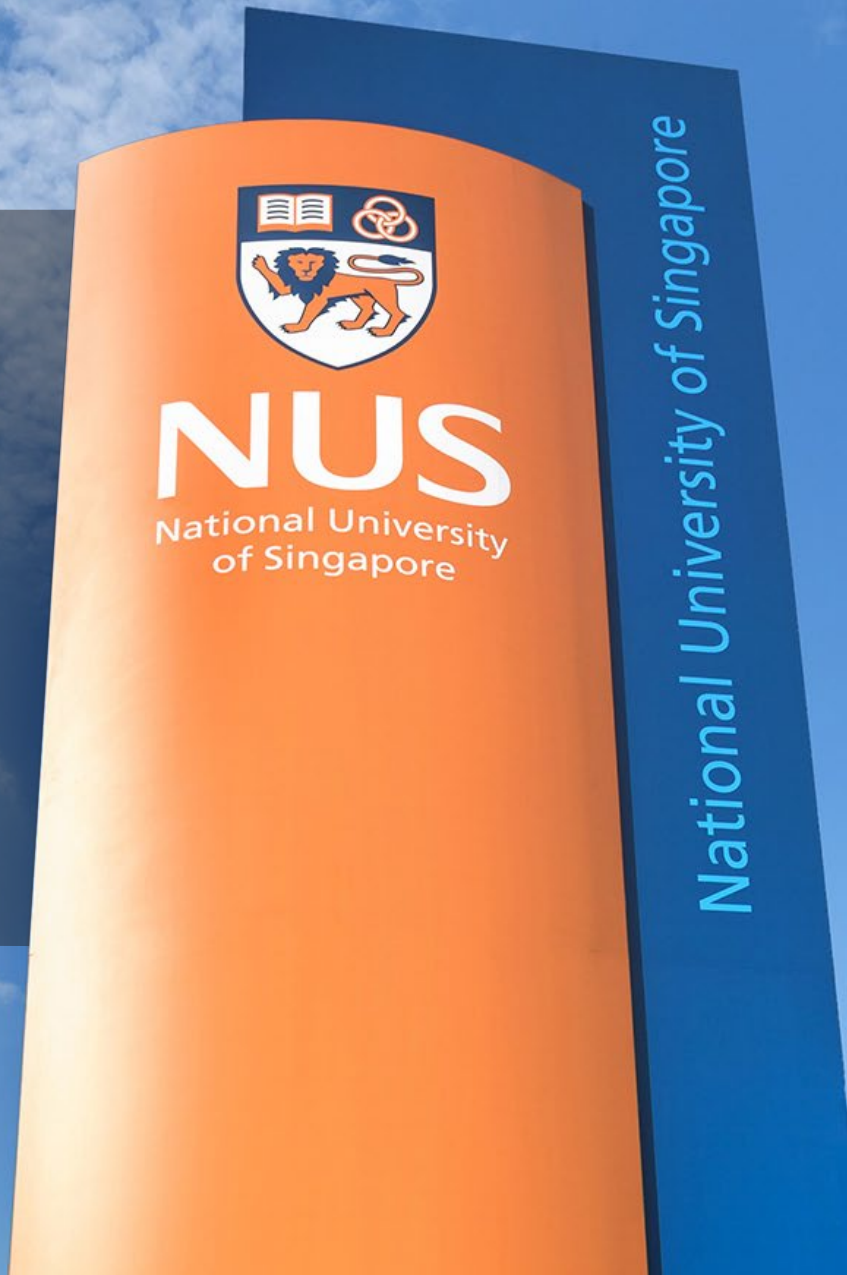
Practical advice for bewildered lay analysts

Wong Limsoon



About Limsoon

Position

Kwan-Im-Thong-Hood-Cho-Temple Chair Professor, Dept of Computer Science, NUS

Research

Database systems & theory, knowledge discovery, bioinformatics & computational biology

Honours

ACM Fellow

FEER Asian Innovation Gold Award 2003

ICDT Test of Time Award 2014

Lecture plan

Testing a hypothesis

Test sample fidel to population?

Right null hypothesis? Right null distribution?

Finding a better hypothesis & explaining why it is better

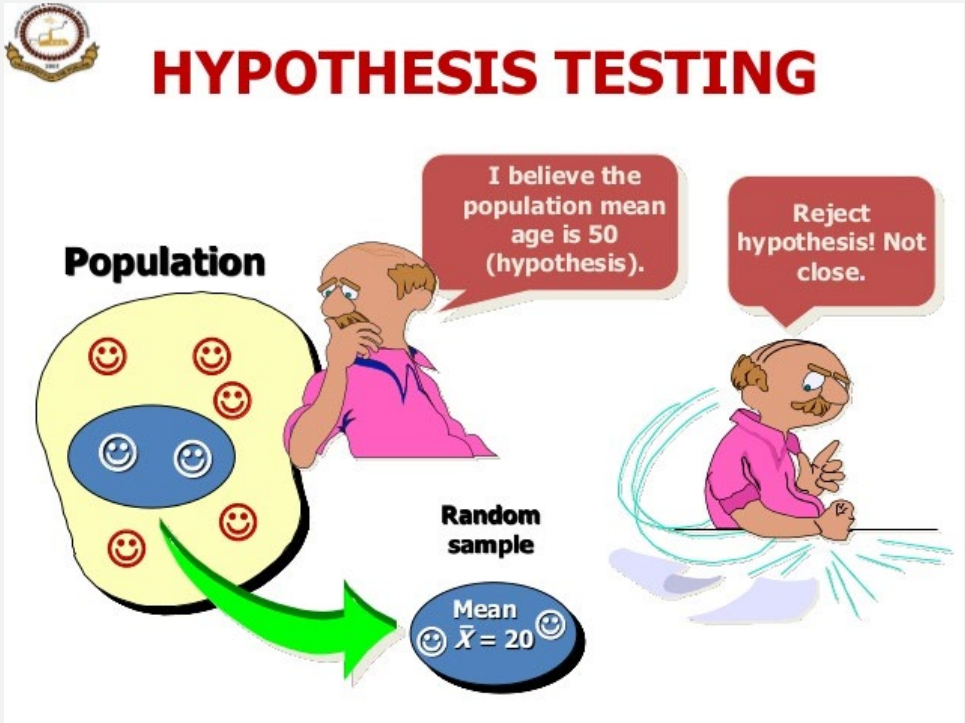
Exceptions? Trend reversals? Trend enhancements?

Data may be telling more than what you think

Assessing a prediction model

Reproducible? Meaningful?

Am I testing this hypothesis correctly?



A seemingly obvious conclusion

SNP	Genotypes	Group: Controls [n(%)]		Group: Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

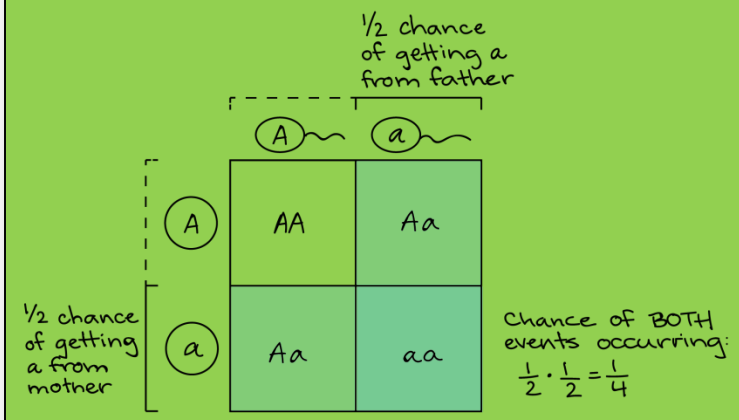
A scientist claims the SNP rs123 is a great biomarker for a disease

If rs123 is AA or GG, unlikely to get the disease

If rs123 is AG, a 3:1 odd of getting the disease

A straightforward $\chi 2$ test. Anything wrong?

Sample may not be fidel to real-world population



Basic rule of human genetics

SNP	Genotypes	Group: Controls [n(%)]		Cases [n(%)]		χ^2	*P* value
rs123	AA	1	0.9%	0	0.0%		4.78E-21[b]
	AG	38	35.2%	79	97.5%		
	GG	69	63.9%	2	2.5%		

Abbreviation: SNP, single nucleotide polymorphism.

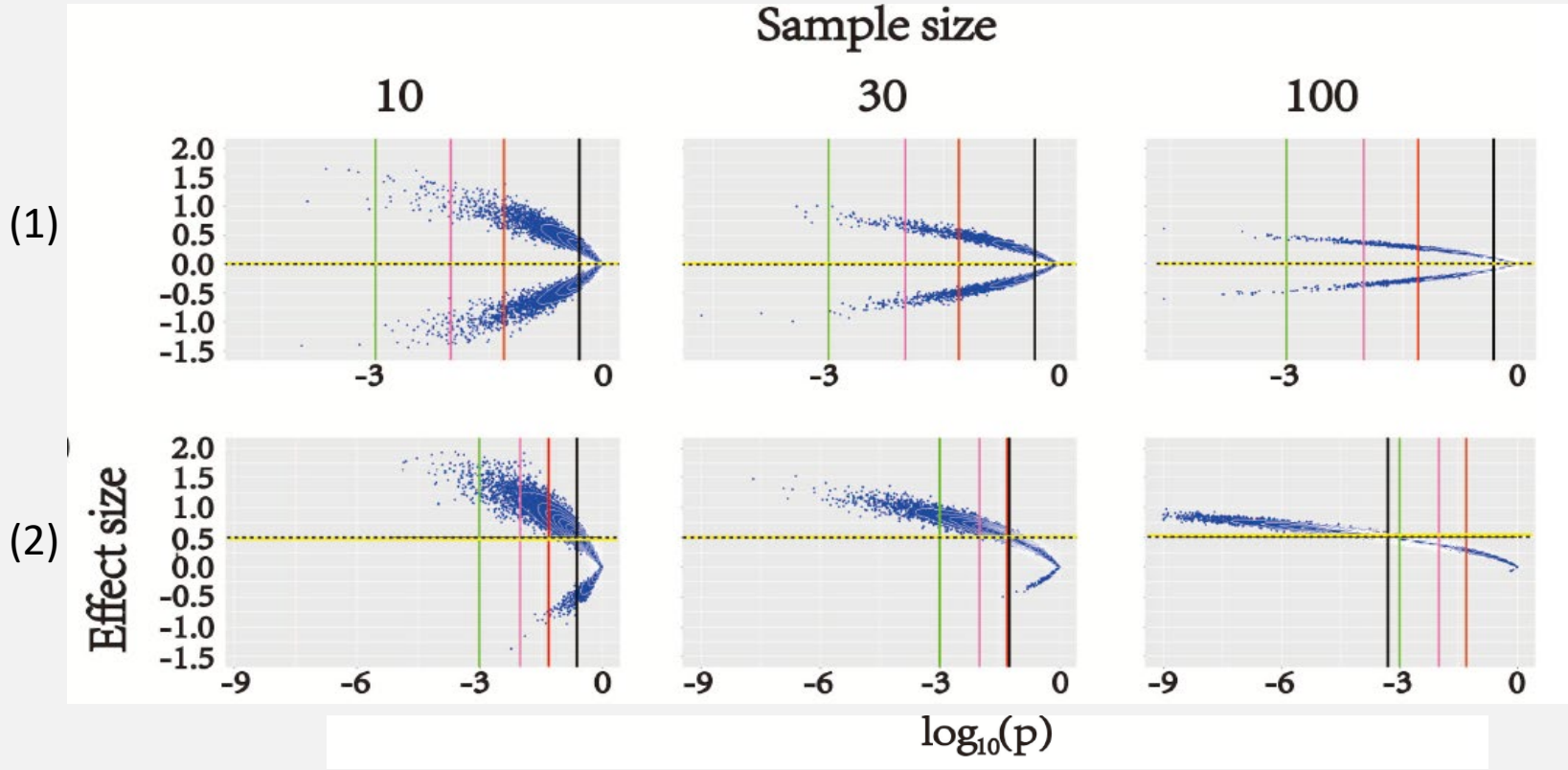
AG = 38 + 79 = 117,
controls + cases = 189

⇒population is ~62% AG

⇒population is >9% AA, unless AA is lethal

Sampling bias happens often

	Distribution		Mean		Standard deviation		Sample size		
Scenario	A	B	A	B	A	B			
(1)	Normal	Normal	0	0	1	1	10	30	100
(2)	Normal	Normal	0	0.5	1	1	10	30	100



Wang, Sue, & Goh. *Drug Discovery Today*, 22(6):912-918, 2017

An old story

"Dewey Defeats Truman" was a famously incorrect banner headline on the front page of the *Chicago Tribune* on November 3, 1948, the day after incumbent United States President Harry S. Truman won an upset victory over Republican challenger and Governor of New York Thomas E. Dewey in the 1948 presidential election.

President-elect Truman holding the infamous issue of the *Chicago Tribune*, telling the press, "That ain't the way I heard it!"

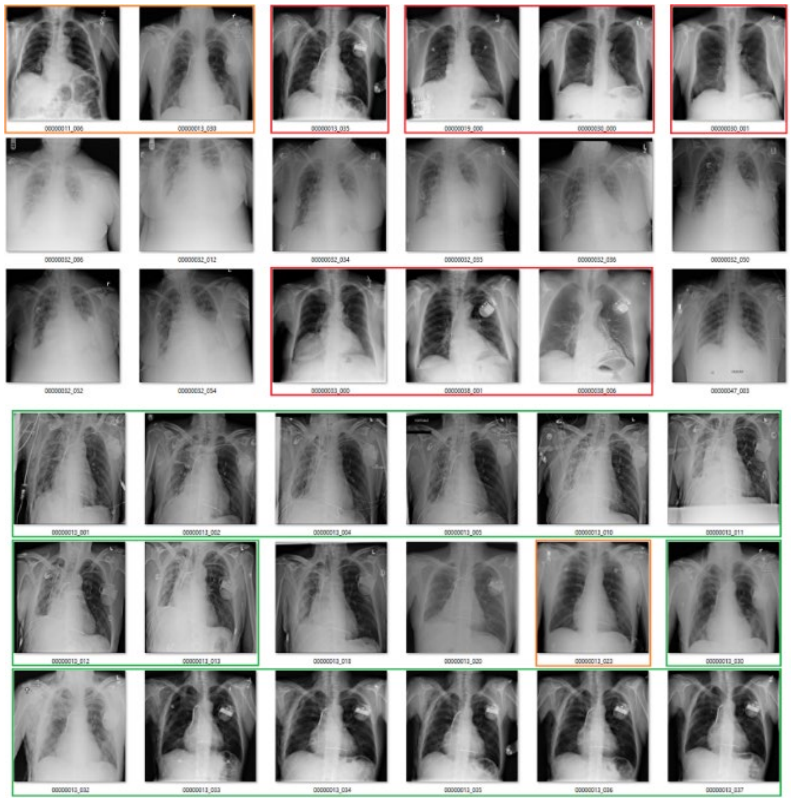
The reason the Tribune was mistaken is that their editor trusted the results of a phone survey… Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses.

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

A very recent story

Disease	MetaMap Precision / Recall / F1-score			Our Method 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 /	[illegible]	83.3
[illegible]osis	100.0 /	100.0 /	100.0	100.0 /	100.0 /	[illegible]
[illegible]T	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
[illegible]rdiomegaly	94.1 /	95.7 /	94.9	96.7 /	95.7 /	[illegible]
Mass	87.7 /	99.6 /	93.3	94.8 /	99.2 /	[illegible]
Nodule	69.7 /	90.0 /	78.6	95.[illegible] /	[illegible] /	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

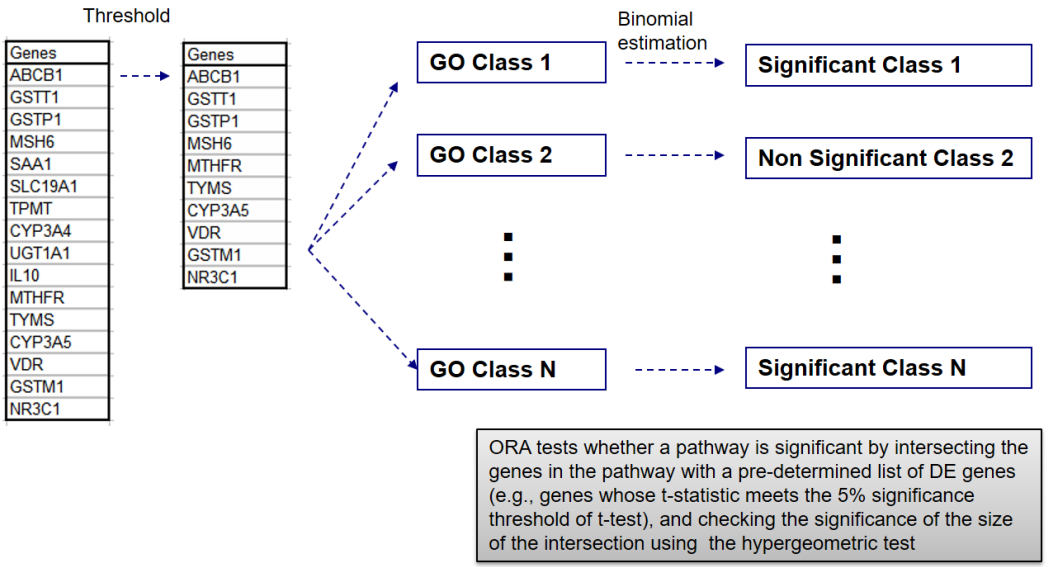
Really good results from a study published in CVPR 2017



Actually the dataset contained many mis-labeled data

Biased data – many pneumo-thorax cases were patients treated with chest drain

A seemingly obvious conclusion

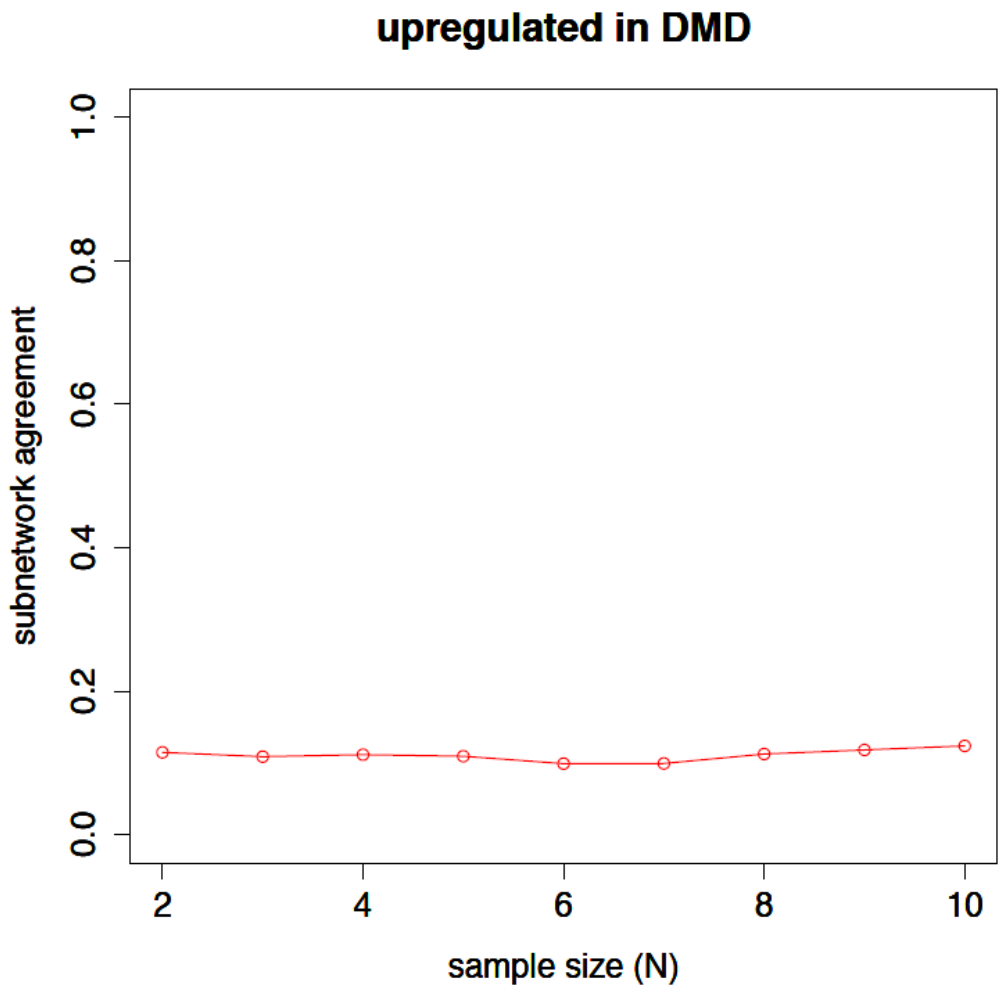


A biological pathway is claimed as an explanation for a disease phenotype as it is enriched with differentially expressed genes

ORA p-value << 0.05

A straightforward hypergeometric test. Anything wrong?

Disappointing performance



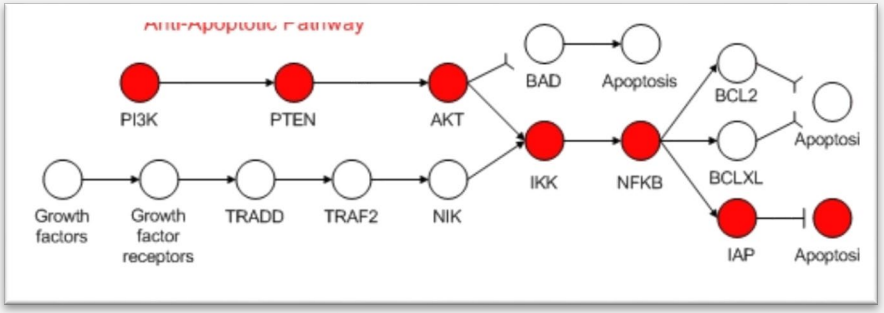
DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010

Null hypothesis may be inappropriate

The null hypothesis underlying ORA basically says “Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size”

This null hypothesis is obviously false



A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behavour of genes in a pathway is more coordinated than random ones

ORA-Paired: New null hypothesis

Let g_i be genes in a given pathway P

Let p_j be a patient

Let q_k be a normal

Let $\Delta_{i,j,k} = Expr(g_i,p_j) - Expr(g_i,q_k)$

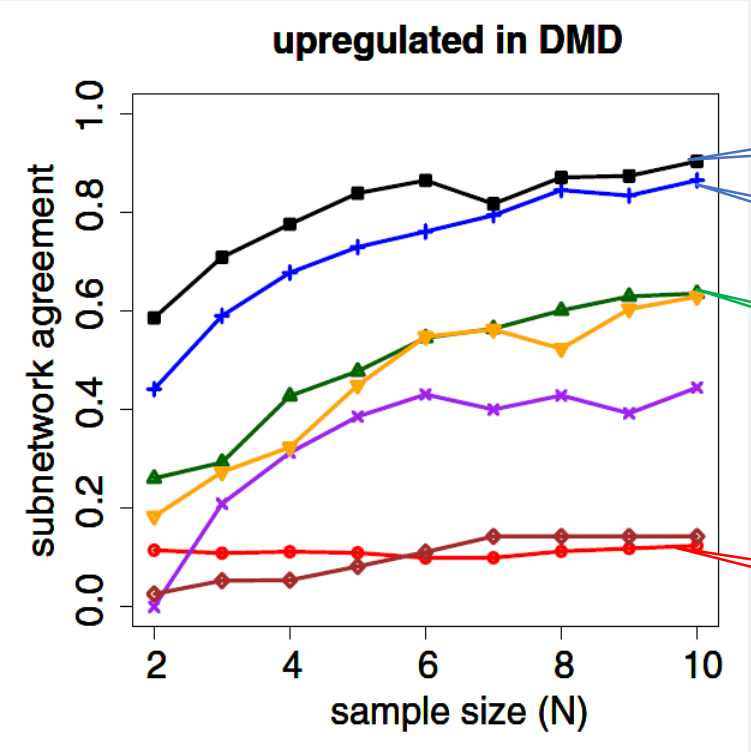
Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Null hypothesis is now much more reasonable…

"Pathway P is irrelevant to the difference between patients and normals, &

Thus genes in P behave similarly in patients and normals"

After fixing the null hypothesis and one other issue…



ESSNet: Subnetwork issue fixed more cleverly

NEA-paired: Null hypothesis & subnetwork issues fixed

ORA-paired: Null hypothesis issue fixed

ORA: Null hypothesis and subnetwork issues

Lim, et al. JBCB, 13(4):1550018, 2015

Not so fast…

How to test $\Delta_{i,j,k}$ = ~0?

ORA-Paired: New null hypothesis

Let g_i be genes in a given pathway P

Let p_j be a patient

Let q_k be a normal

Let $\Delta_{i,j,k} = Expr(g_i,p_j) - Expr(g_i,q_k)$

Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

Null hypothesis is now much more reasonable…

"Pathway P is irrelevant to the difference between patients and normals, &

Thus genes in P behave similarly in patients and normals"

Test statistic is t-statistic, $t = \mu_\Delta / (\sigma_\Delta / \text{sqrt}(n))$

Null distribution is t-distribution

Degrees of freedom is |patients| * |normal| * |P|

What do you think?

t-statistic is test statistic ≠ t-distribution is null distribution

Testing the null hypothesis

"Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals"

- **Method #1**
  - T-test w/ a conservative degree of freedom
    - **E.g., | normal | + | patients |**
- **Method #2**
  - By the null hypothesis, a dataset and any of its class-label permutations are **exchangeable**

  ⇒ Get null distribution by class-label permutations
    - **Only for large-size sample**
- **Method #3**
  - Modified null hypothesis
    - **"Pathway P induces gene-gene correlations, and genes in P behave according to these gene-gene correlations;**
    - **P is irrelevant to the diff betw patients and normals and so, genes in P behave similarly in patients and normals"**

  ⇒ Get null distribution using datasets that conserve gene-gene correlations in the original dataset
    - **E.g., array rotation**

A little more biology background for the next example …

Synthetic lethal pairs

Fact/postulate

When a pair of genes is synthetic lethal, their mutations avoid each other

Observation

Mutations in genes (A,B) are seldom observed in the same subjects

Conclusion by abduction

Genes (A,B) are synthetic lethal

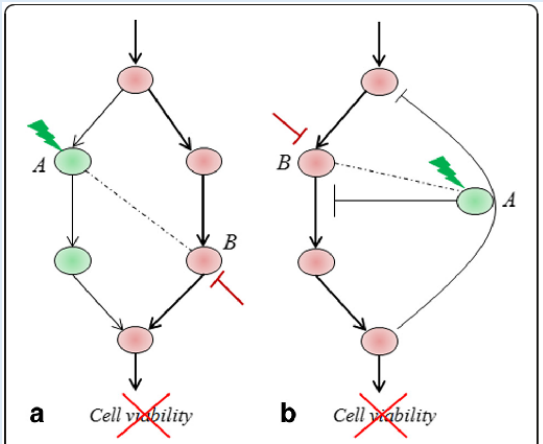
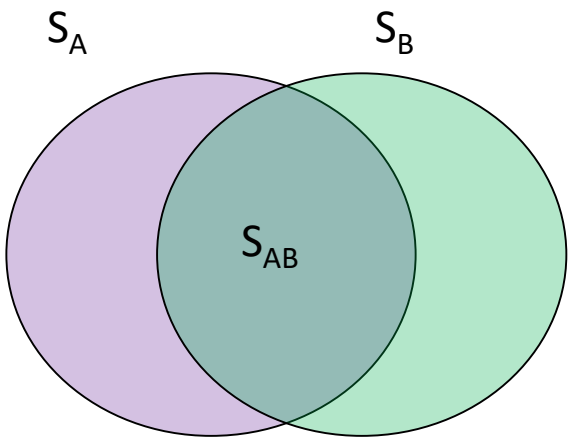


Fig. 7 Two models for pathway-based targeting of synthetic lethal genes *B* in conjunction with deleted/downregulated genes *A*: **a** parallel pathways model where targeting *B* results in disruption of both survival pathways, and **b** negative feedback-loop model where targeting *B* shunts of (forward) signals for cell survival

Why interested in synthetic lethality?

They are good cancer treatment targets

A seemingly obvious approach based on hypergeometric test



$$P[X \leq |S_{AB}|] = 1 - P[X > |S_{AB}|], \quad (1)$$

where $P\ [X > |S_{AB}|]$ is computed using the hypergeometric probability mass function for $X = k > |S_{AB}|$:

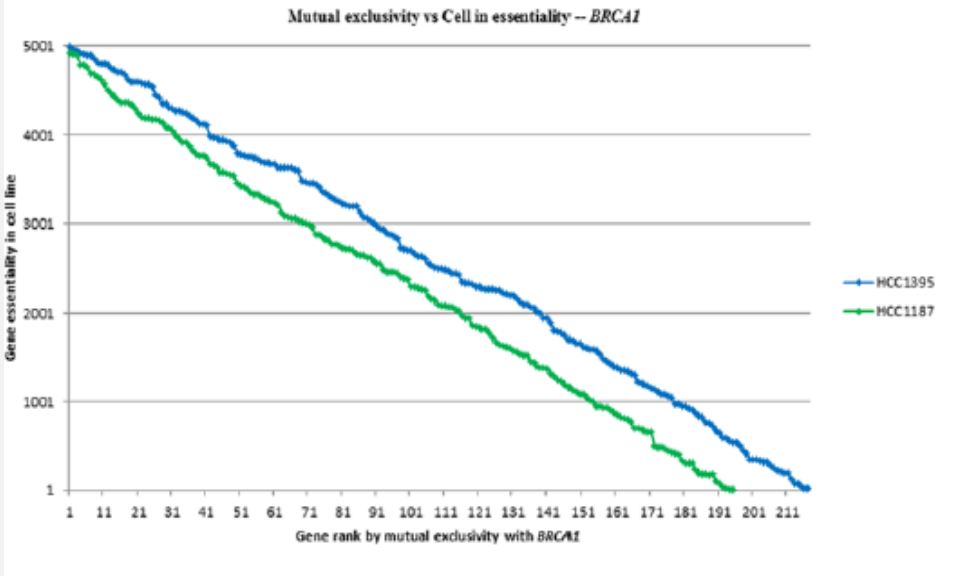
$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

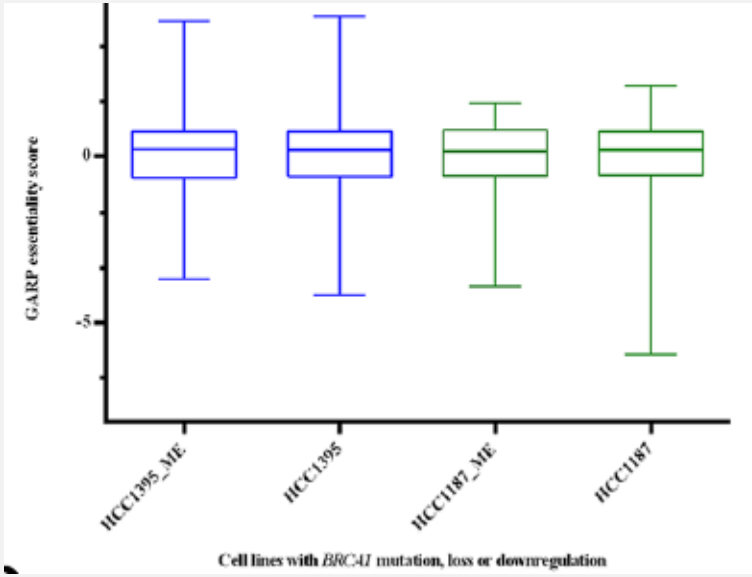
Mutations of genes (A,B) avoid each other if $P[X \leq S_{AB}] \leq 0.05$

Anything wrong with this?

Srihari et al. *Biology Direct*, 10:57, 2015.

What is happening?





Among top ME-genes, GARP score ranks seems to correlate with mutual exclusion ranks

GARP scores of ME-genes (viz. significantly mutually exclusive mutations to BRCA1) are similar to other genes

Hypergeometric distribution does not reflect real-world mutations

$$P[X \le |S_{AB}|] = 1 - P[X > |S_{AB}|], \quad (1)$$

where $P\ [X > |S_{AB}|]$ is computed using the hypergeometric probability mass function for $X = k > |S_{AB}|$:

$$P[X > |S_{AB}|] = \sum_{k=|S_{AB}|+1}^{|S_B|} \frac{\binom{|S_A|}{k}\binom{|S|-|S_A|}{|S_B|-k}}{\binom{|S|}{|S_B|}}$$

Hypergeometric distribution assumes

- Mutations are independent
- Mutations have equal chance to appear in a subject

Real-life mutations

Inherited in blocks; those close to each other are correlated

Some subjects have more mutations than others, e.g. those with defective DNA-repair genes

Null distribution is not hypergeometric, binomial, etc.

An engineer's solution

Group genes into genomic clusters

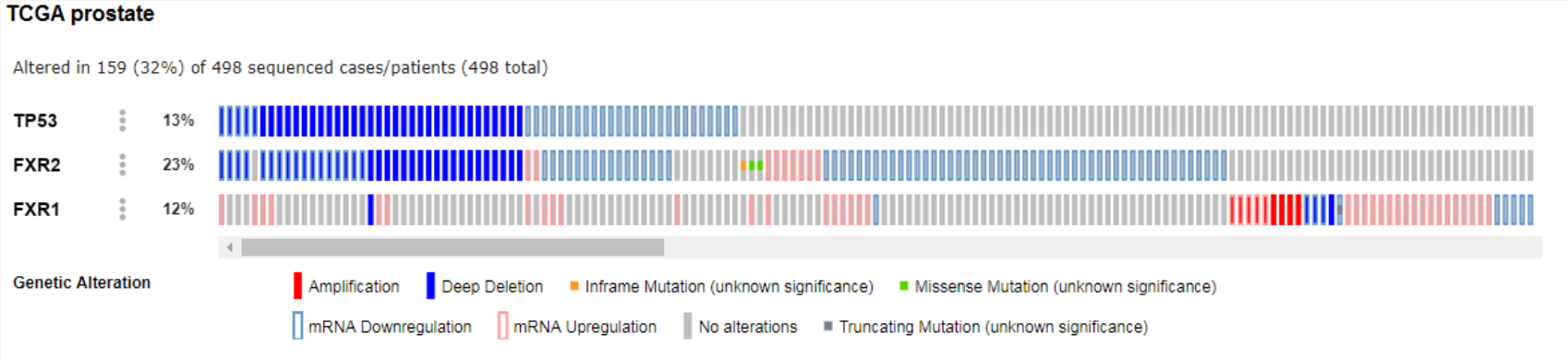
Test genes in far-apart genomic clusters for mutually exclusive mutations

Mutually exclusive clusters should contain synthetic-lethal & *collateral-lethal* gene pairs

Illustrative example

FXR2 is located near TP53

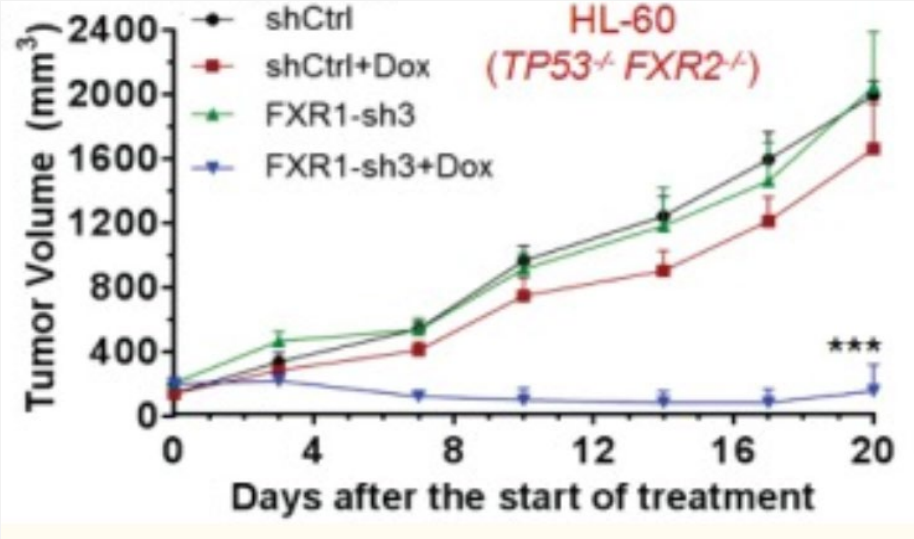
FXR1 & 2 are paralogs that buffer each other's function



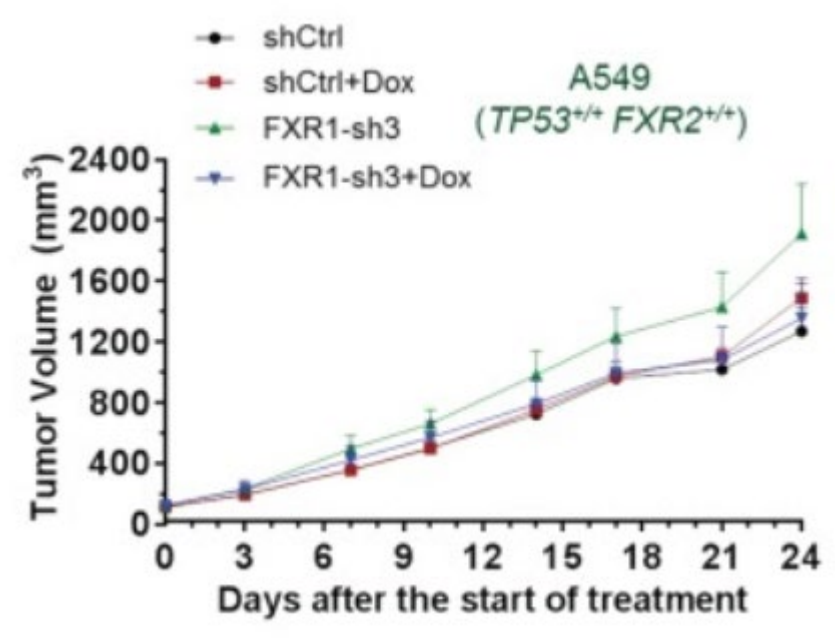
Is FXR1 synthetic lethal to TP53?

Does inhibiting FXR1 lead to cell death for TP53-deleted cell lines?

Collateral lethality



Fan et al., eLife, 6:e26129, 2017



Tumour bearing homozygous TP53/FXR2 co-deletion shrinks upon doxycycline-induced FXR1 knock down

Learning points

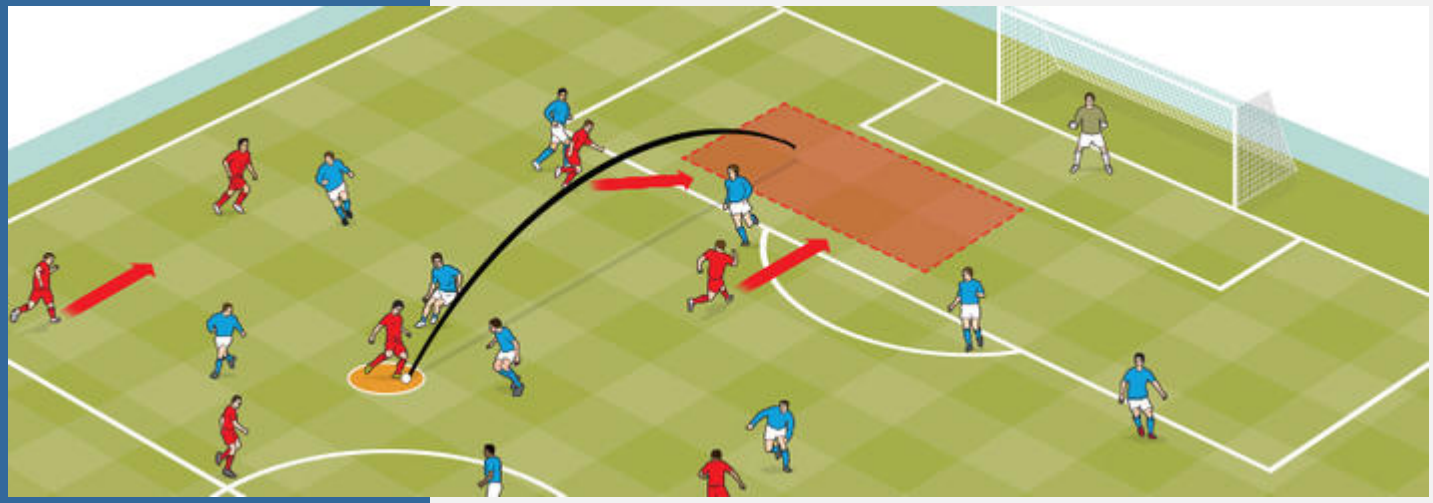
Sample fidelity to population

Right null hypothesis

Right null distribution

Note that using a test statistic does not mean you must use its nominal null distribution

How do I find deeper insight from data?



Getting lost in data

The Australian adult dataset from UCI machine learning repository contains demographic data of 32k adults

If a freq-pattern mining method is run on this dataset, thousand of patterns like these are produced

- {Race = White, Occupation = Adm-clerical, Income>50K}: 439,
- {Race = White, Occupation = Adm-clerical, Income<50K}: 2,645,
- {Race = White, Occupation = Craft-repair, Income>50K}: 844, and
- {Race = White, Occupation = Craft-repair, Income<50K}: 2,850,

A lay analyst will be quite lost…

Think in terms of a contingency table helps

Context
Race = White

Occupation	Income>50K	Income<50K
Adm-clerical	439 (14%)	2,645 (86%)
Craft-repair	844 (23%)	2,850 (77%)

Related patterns can be put into the form of contingency tables

These tables may be more palatable for compare-and-contrast analysis

A seemingly obvious conclusion

Context
Race = White

Occupation	Income>50K	Income<50K
Adm-clerical	439 (14%)	2,645 (86%)
Craft-repair	844 (23%)	2,850 (77%)

The data shows that, in Australia, craft repairers tend to earn more than administrative clerks

23% of the former vs 14% of the latter has high income

A straightforward $\chi 2$ test. Anything wrong?

Contradictions as deeper insight

Context
Race = White, Workclass = Self-emp-not-inc

Occupation	Income>50K	Income<50K
Adm-clerical	16 (35%)	30 (65%)
Craft-repair	90 (18%)	409 (82%)

The “unincorporated self-employed” work class is a contradiction to the conclusion that “craft repairers tend to earn more than administrative clerks”

Exceptions as deeper insight

Context	Occupation	Income>50K	Income<50K
Race = White, Sex = Male	Adm-clerical	251 (24%)	787 (76%)
	Craft-repair	829 (24%)	2,695 (76%)

Context	Occupation	Income>50K	Income<50K
Race = White, Sex = Female	Adm-clerical	188 (9%)	1,858 (91%)
	Craft-repair	15 (9%)	155 (91%)

The conclusion “craft repairers tend to earn more than administrative clerks” holds for neither male nor female

The conclusion is an artefact of male earning more than female

A seemingly obvious conclusion

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
III	25	245	270
IV	48	212	260
V	57	233	290
Total	225	1125	1350

Vaccines I-V are not equal in efficacy

0.001 < $\chi 2$ test p-value < 0.01 is significant

A straightforward $\chi 2$ test. Anything wrong?

Trend-strengthening subpopulation as deeper insight

Computation of the $\chi 2$

Type of vaccines	Had flu	$(O-E)^2/E$	Avoided flu	$(O-E)^2/E$
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	**25 (45.0)**	**8.889**	**245 (225.0)**	**1.778**
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- **Vaccine III contributes to the overall $\chi 2$= (8.889+1.778)/16.564 = 64.4%**

Vaccine III vs. rest

Type of vaccines	Had flu	Avoided flu	total
III	25	245	270
I, II, IV, V	200	880	1080
Total	225	1125	1350

- **$\chi 2$ =12.7 with 1 d.f.**
- **P < 0.001**

$\chi 2$ with Vaccine III removed

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
IV	48	212	260
V	57	233	290

- **$\chi 2$ =2.983 with 3 d.f.**
- **0.1<p<0.5, not statistically significant**

Vaccine III is different from / better than the rest

Can these tactics be automated?

Formulation of a hypothesis

"For Chinese, is drug A better than drug B?"

Three components of a hypothesis

Context (under which the hypothesis is tested), e.g. Race = Chinese

Comparing attribute, e.g. Drug = A or B

Target attribute/target value, e.g. Response = positive

$\langle \{\text{Race=Chinese}\},\ \text{Drug=A|B},\ \text{Response=positive} \rangle$

Algo for rough hypothesis analysis

Given a hypothesis H

Add values of an extra attribute A to context of H

Re-calculate test statistic

Test statistic is reversed ➔ Contradiction?

Test statistic becomes insignificant ➔ Exception?

Test statistic is strengthened ➔ Better explanation?

Brute-force on small datasets

Freq-pattern mining on big datasets & immediate superset search on freq patterns

A frequent pattern ≈ a population

A superset of a frequent pattern ≈ a subpopulation

Excelah! - Hypothesis

Target Attribute	
Target Attribute:	Income
Preliminary Analysis	
Delete Preliminary Analysis	
Initial Hypothesis Test Parameters	
Number of context variables:	2
Choose statistic:	Odds ratio
Perform Initial Hypothesis Test	
Follow-up Analysis Parameters	
Min. Population:	20
Min. % Change:	1
Significance Level (α):	0.05
Follow-up Analysis	
Delete Follow-up Analysis	

Constraints
Race = White
Sex = Male

Observed Values			
Education	>50K	<=50K	Total
Bachelors	1727 (51.6%)	1622 (48.4%)	3349
HS-grad	1335 (21.4%)	4893 (78.6%)	6228
Total	3062	6515	9577

Expected Values			
Education	>50K	<=50K	Total
Bachelors	1070.8	2278.2	3349
HS-grad	1991.2	4236.8	6228
Total	3062	6515	9577

Chi-Square Values			
Education	>50K	<=50K	Total
Bachelors	402.13	189.01	909.02
HS-grad	216.25	101.63	

Statistics Summary	
Chi-square Statistic	909.02
P-Value	1.074E-199
Odds Ratio (Bachelors / HS-grad)	3.902432841

Implication Analysis Report

In the context of {Race: White, Sex: Male}, {Education: Bachelors} is 3.902 times more likely than {Education: HS-grad} to be {Income: >50K}

Follow-up Analysis Report

Trend Enhancements

There are subpopulations for which the trend is enhanced; specifically: {WorkClass: Private}, {Marital-Status: Never-married}, {Marital-Status: Married-civ-spouse}, {Marital-Status: Divorced}, {Marital-Status: Married-spouse-absent}, {Marital-Status: Separated}, {Marital-Status: Widowed}, {Occupation: Protective-serv}, {Relationship: Not-in-family}, {Relationship: Husband}, {Relationship:

Trend Supporters

There are subpopulations which support this trend; specifically: {WorkClass: State-gov}, {WorkClass: Self-emp-not-inc}, {WorkClass: Federal-gov}, {WorkClass: Local-gov}, {WorkClass: Self-emp-inc}, {Occupation: Adm-clerical}, {Occupation: Exec-managerial}, {Occupation: Handlers-cleaners}, {Occupation: Prof-specialty}, {Occupation: Other-service}, {Occupation: Sales}, {Occupation: Craft-repair},

Trend Reversals

There are no subpopulations for which this trend is reversed.

No trend-reversing subpopulation. Hypothesis is very likely true

Trend Exceptions

There are subpopulations which are exceptions to this trend; specifically: {Occupation: Transport-moving}, {Occupation: Tech-support}, {Native-Country: England}, {Native-Country: Canada}, {Native-Country: Italy}, {Native-Country: Poland}

Interesting subpopulations for further investigation. E.g. the hypothesis is insignificant for the European immigrant subpopulations; perhaps they all have degrees?

Chi-square | T-Test | Wilcoxon_Rank_Sum | Query_Results | Analysis_Chi2-WorkClass | Analysis_Chi2-Marital-Status | Analysis_Chi2-Occupation | Analysis_Chi2-Rel

Download this excel plug-in at https://github.com/dblux/excelah

Learning points

Exceptions

Trend reversals

Trend enhancements

Sometimes the little bits (that you want to discard) are more informative than what you think

We tend to ignore non-associations

Many technologies for association and correlation mining

Frequent patterns, association rules, …

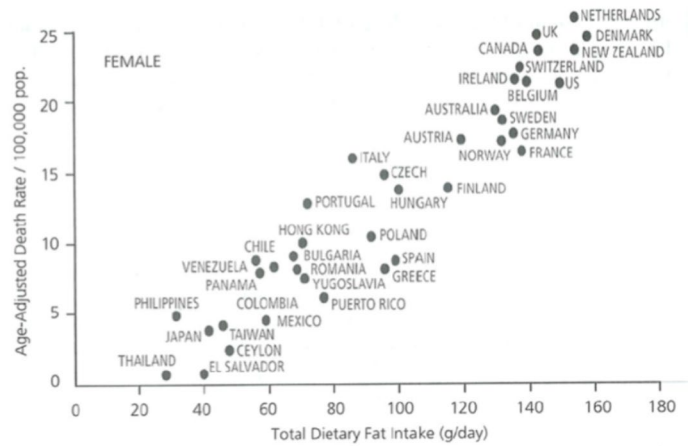
But ignore non-associations

Not interesting

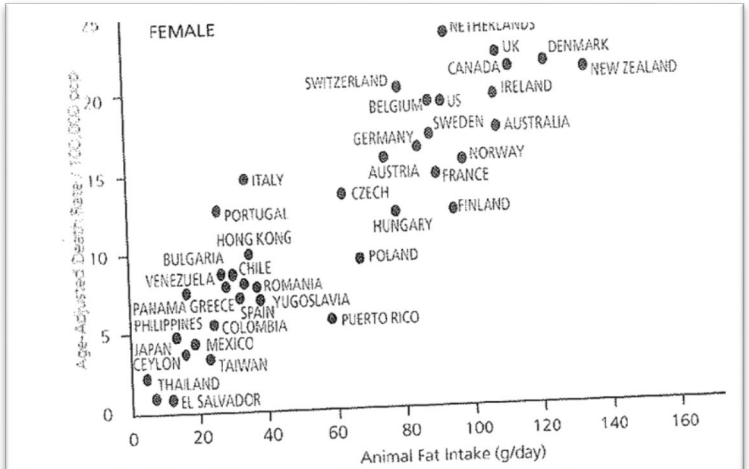
Too many of them

Is this a good thing?

We love to find correlations like these…



Dietarv fat intake correlates with breast cancer



Animal fat intake correlates with breast cancer

But not non-correlations like this…

Plant fat intake doesn't correlate with breast cancer

There is much to be gained when we take both into our analysis

A: Dietary fat intake correlates with breast cancer

B: Animal fat intake correlates with breast cancer

C: Plant fat intake doesn't correlate with breast cancer

Given C, we can eliminate A from consideration, and focus on B!

We tend to ignore context!

We have many technologies to look for associations and correlations

Frequent patterns, association rules, …

We tend to assume the same context for all patterns and set the same global threshold

This works for a focused dataset

But for big data where you union many things, this spells trouble

The right context is crucial

⟨{Race=Chinese}, Drug=A|B, Response=positive⟩

Context	Comparing attribute	response= positive	response= negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{pos}$
	Drug=B	N^B_{pos}	$N^B - N^B_{pos}$

If A/B treat the same single disease, it is ok

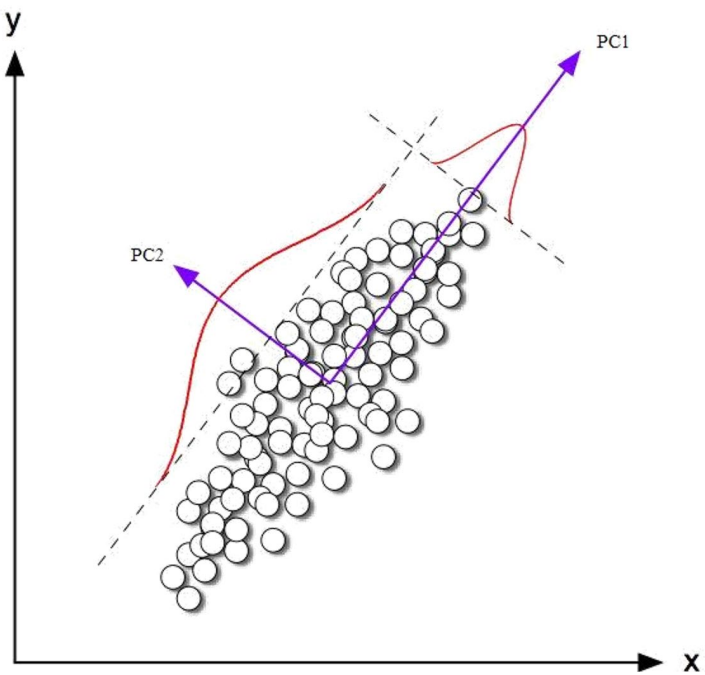
If B treats two diseases, but A one, it is not sensible

The disease has to go into the context

In PCA, lower PCs account for minute amounts of variance; these PCs are often ignored. Should they?

A quick reminder about PCA

PCA, in modern English ☺



Introduction

- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a *rotation* of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \text{Var}(Y_p)$. Idea: Y_1 is most important, then Y_2, etc.

9 / 33

Definition of PCA

- Given $X = (X_1, \ldots, X_p)'$
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$
- Find the SLC $a'_{(1)} = (a_{11}, \ldots, a_{p1})$ so that $Y_1 = a'_{(1)}X$ has maximal variance
- Find the SLC $a'_{(2)} = (a_{12}, \ldots, a_{p2})$ so that $Y_2 = a'_{(2)}X$ has maximal variance, subject to the constraint that Y_2 is uncorrelated to Y_1.
- Find the SLC $a'_{(3)} = (a_{13}, \ldots, a_{p3})$ so that $Y_3 = a'_{(3)}X$ has maximal variance, subject to the constraint that Y_3 is uncorrelated to Y_1 and Y_2
- Etc...

10 / 33

Madrid and Warsaw are at almost the same distance to Latium 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

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

PCA of distance matrix of European cities to Latium cities

Factor loadings and proportions of explained variance

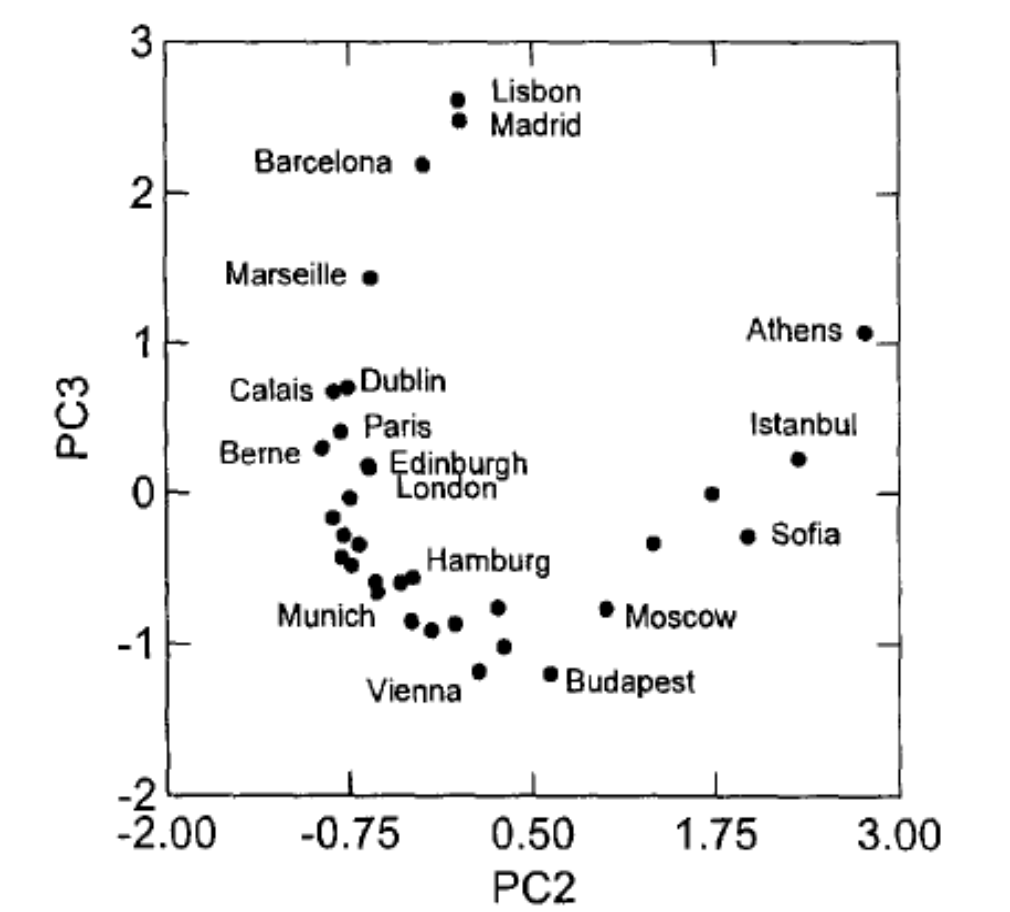
Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569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

PC1 correlates with distance of European cities to Latium cities

PC2, PC3, … account for < 1% of variance

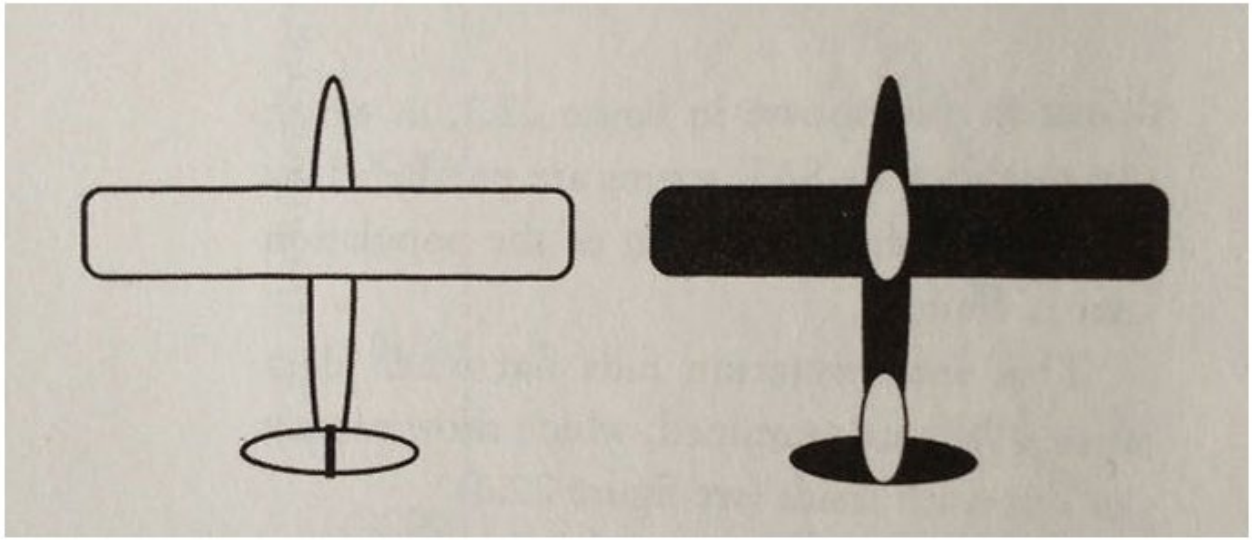
Are PC2, PC3, … useless / non-informative?

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



Another old story

Abraham Wald's analysis of survivability of bombers in WWII



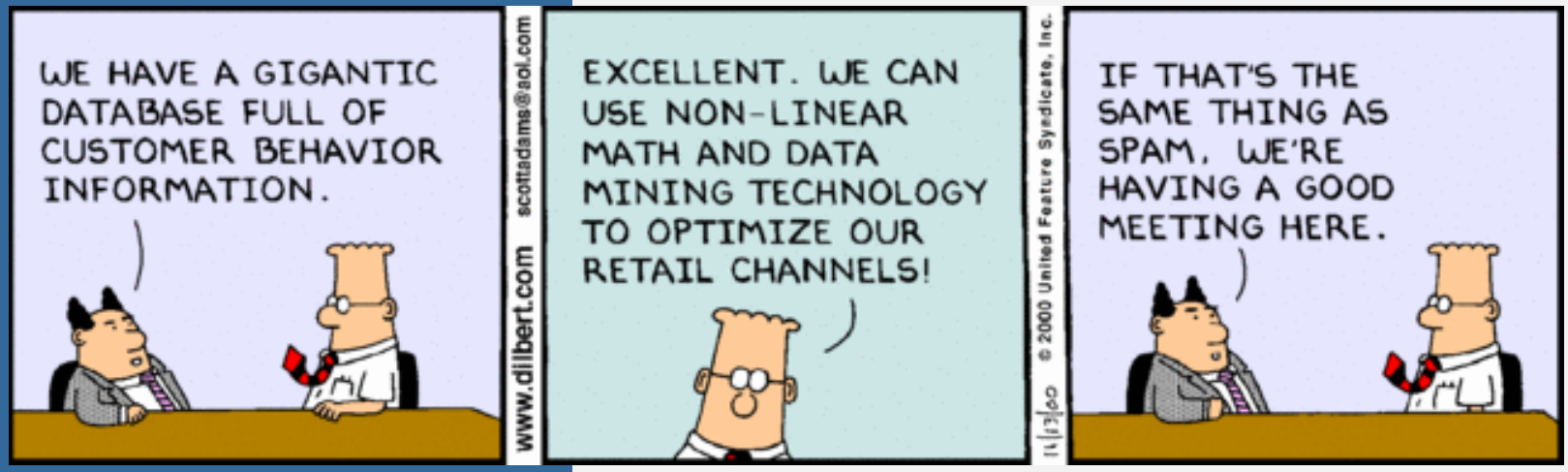
Undamaged plane (left). A plane shaded everywhere bullets struck returning aircraft (right).

Learning points

Mechanically applying data mining, statistical testing, etc. can only take you so far

“It is so easy to make bad inference with data… there’s a creative part of understanding quantitative data that requires a sort of artistic or creative approach to research.” ---Nate Bolt

Have I constructed a “meaningful” model?



Prediction models are often evaluated for accuracy etc. on some test sets

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

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{No. of predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Is this too simple minded?

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Precision = TP / (TP + FP)

Accuracy does not correlate with classifier similarity

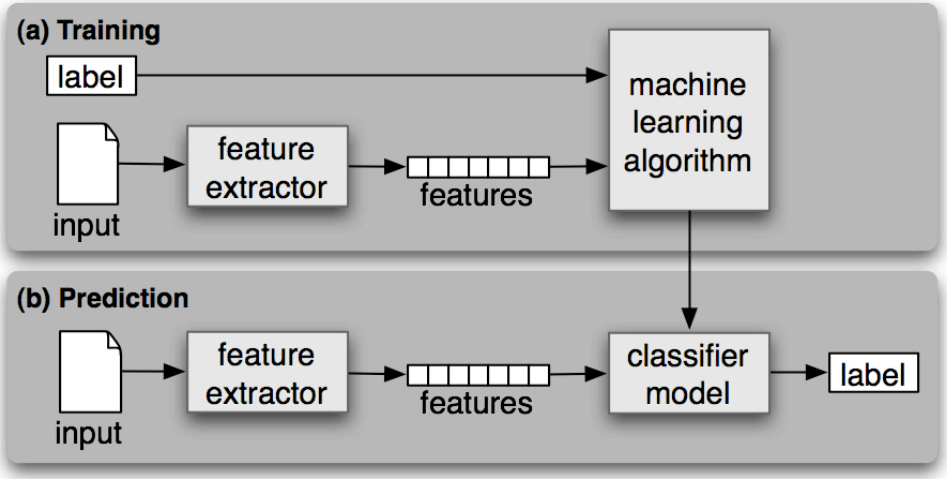
NN	NN Acc. (%)	Acc. t_1-sparse (%)	Acc. t_2-sparse (%)	NPAQ r for t_1-sparse (%)	NPAQ r for t_2-sparse (%)
$ARCH_1$	74.00	78.00	81.00	20.31	62.50
$ARCH_2$	62.00	73.00	78.00	12.50	65.62
$ARCH_3$	76.00	82.00	83.00	4	
$ARCH_4$	50.00	64.00	72.00	1	
$ARCH_5$	78.00	82.00	83.00	7	
$ARCH_6$	80.00	11.00	87.00	37.50	55.4
$ARCH_7$	87.00	89.00	89.00	6.25	79.69

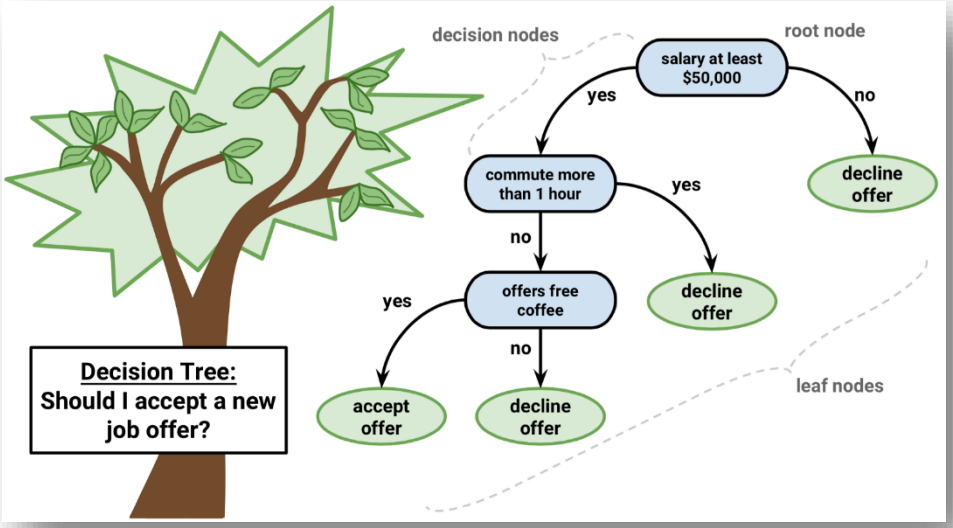
Although t2-sparse and ARCH7 are both ~90% accurate on the test set, they will disagree on ~80% of future cases

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 orignal model.

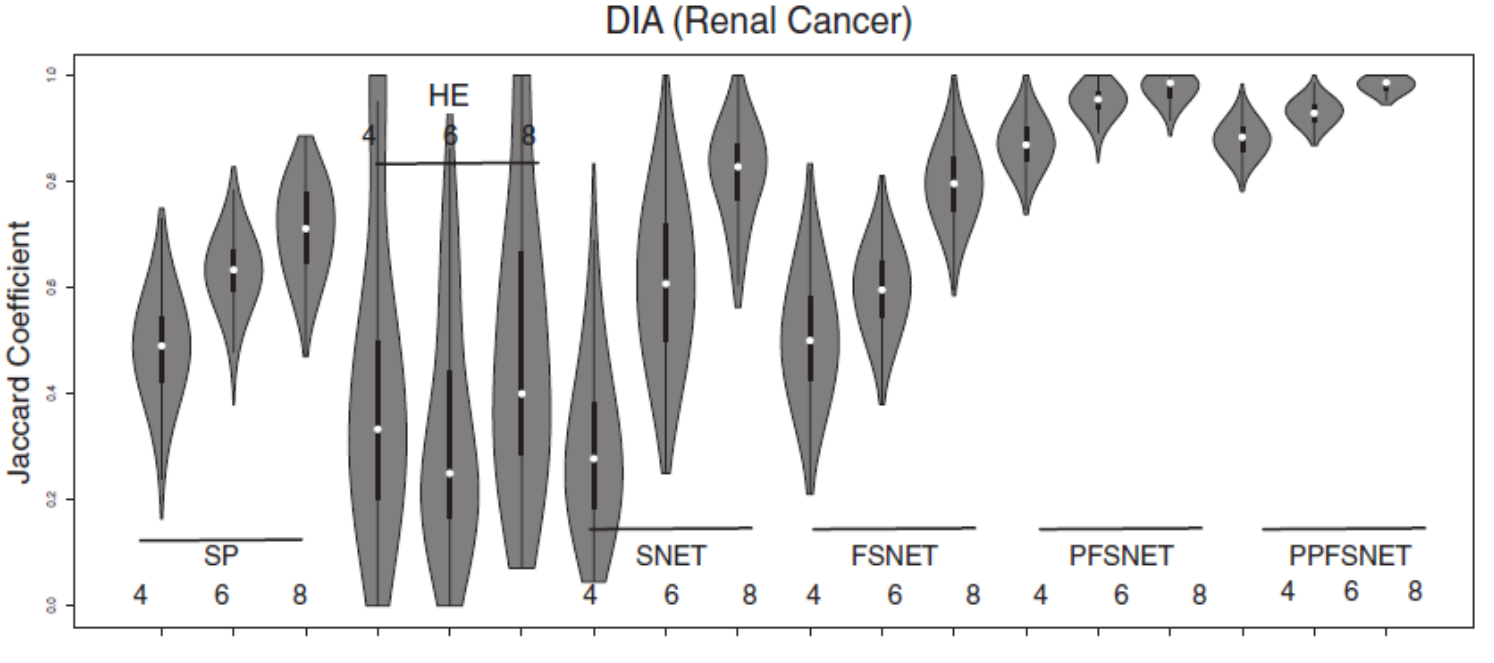
Credit: Teodora Baluta

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





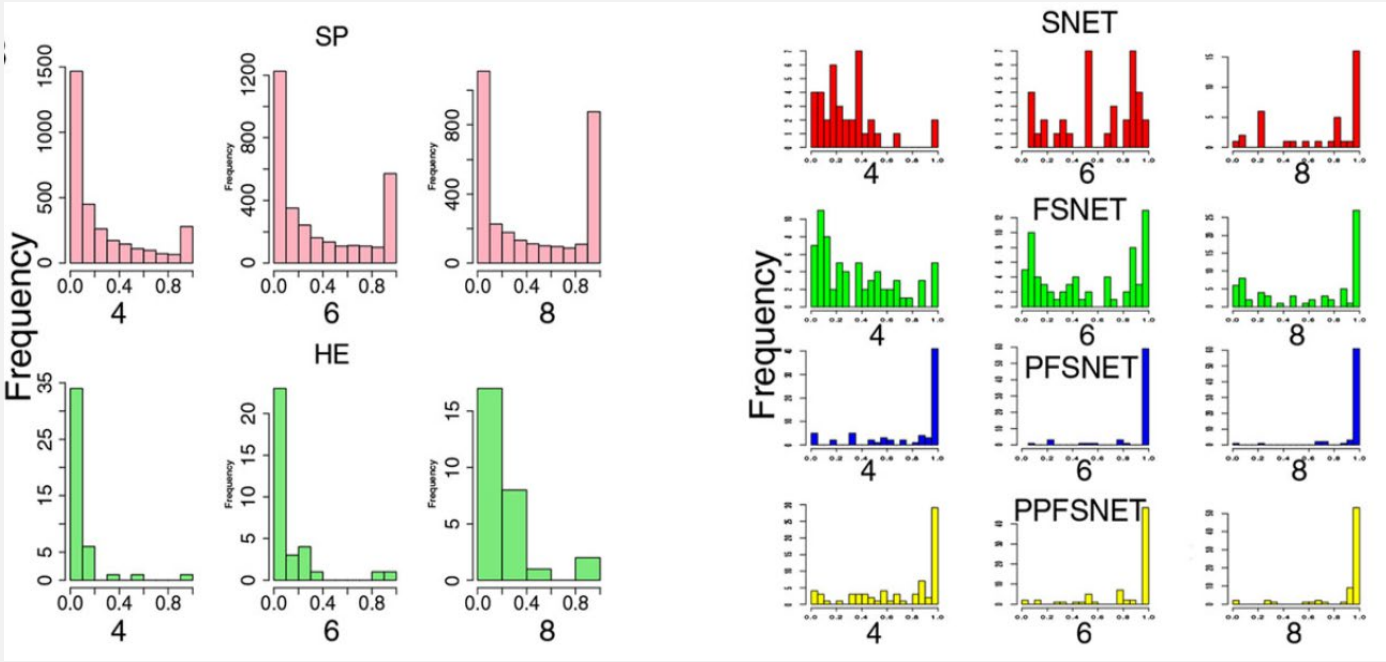
High accuracy does not imply features used are reproducible



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

HE uses high-level features (complexes) but incc

SNET/FSNET/PFSNET/PPFSNET uses high-level features (complexes) but correct null hypothesis/distribution

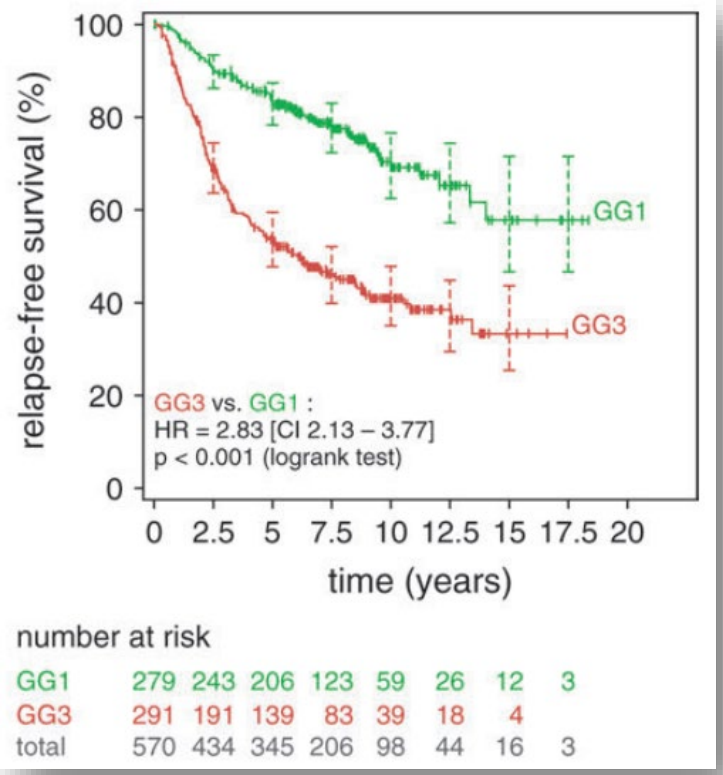
Me		accuracy	CV p-val	CV accuracy/pval
SP			0.91	1.08
HE	162	0.98	0.91	1.08
SNET	21	0.84	0.06	14.00
FSNET	36	0.96	0.06	16.00
PFSNET	65	0.92	0.06	15.33
PPFSNET	66	0.96	0.06	16.00

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

A seemingly obvious conclusion



A multi-gene signature is claimed as a good biomarker for breast cancer survival

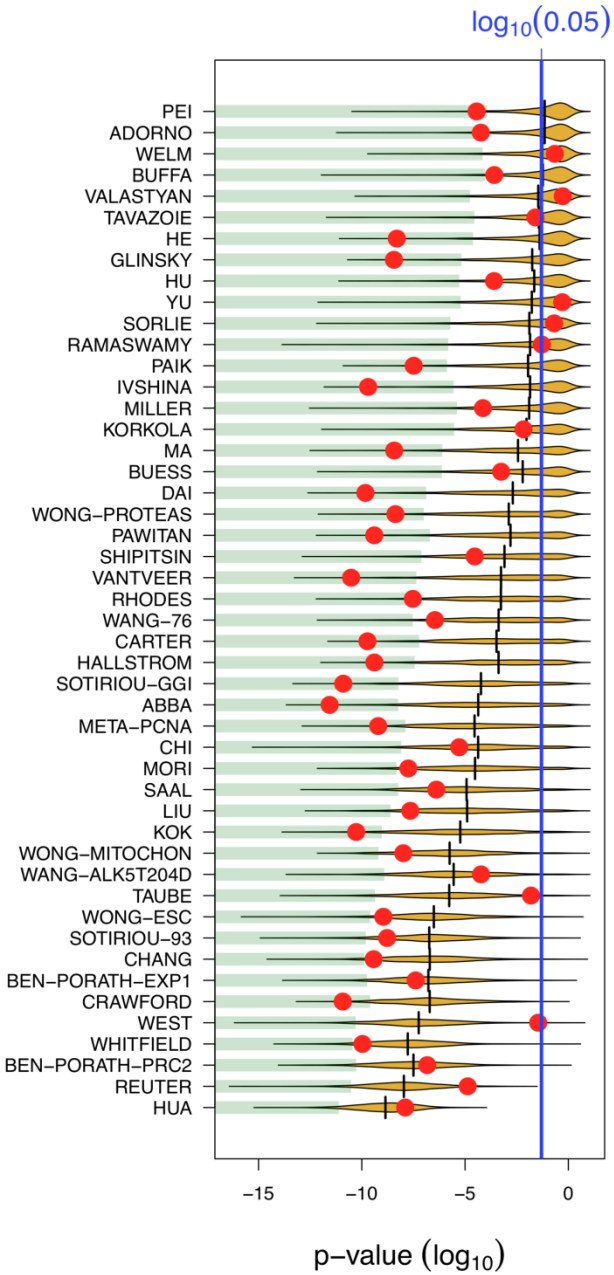
Cox's survival p-value << 0.05

A straightforward Cox's proportional hazard analysis. Anything wrong?

Are significant signatures meaningful?

40-50% of random signatures also have p-value << 0.05

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



An engineer's solution

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

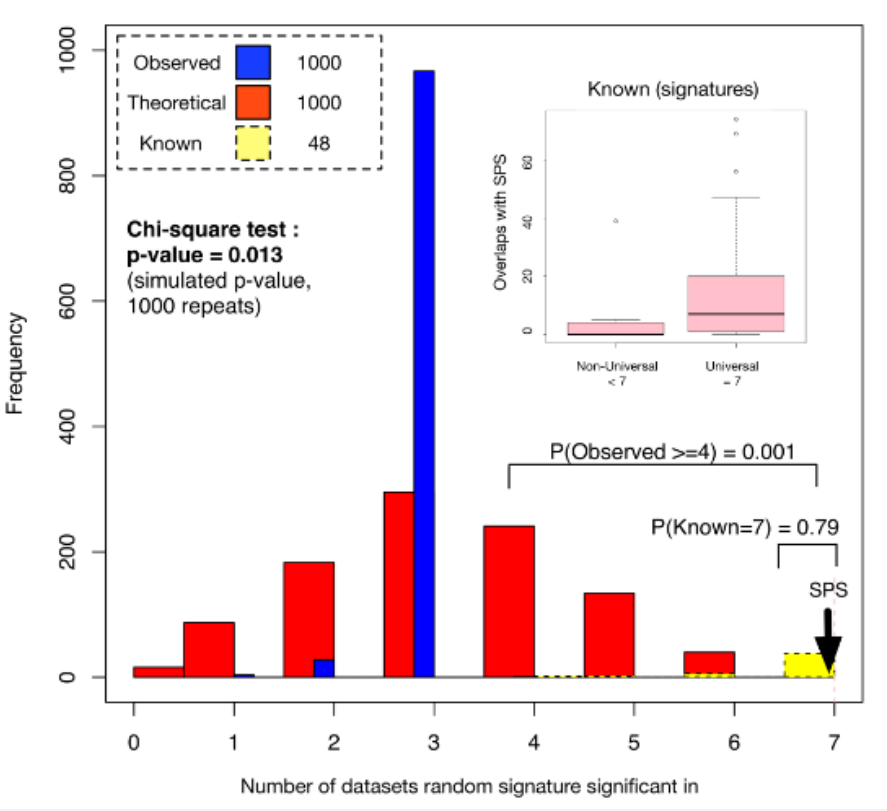
How many independent datasets are needed to avoid reporting random signatures as significant?

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



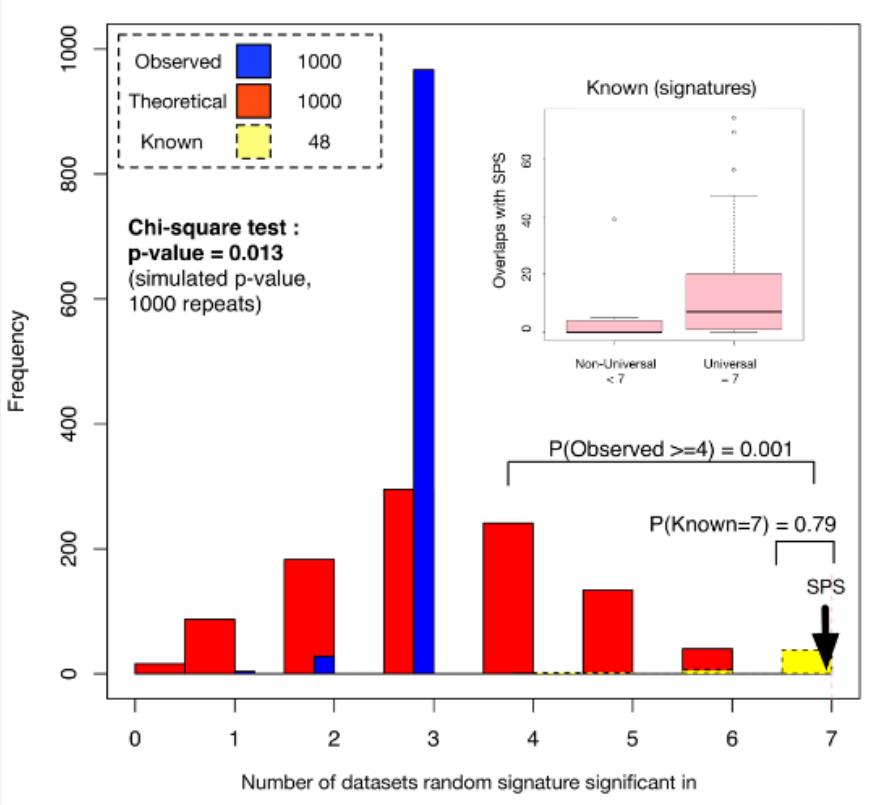
Goh & Wong. Drug Discovery Today, 24(1):31--36, 2019

A theory-practice gap

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



Learning points

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

Reproducibility of features selected

Consideration of confounding factors

Validate on many datasets

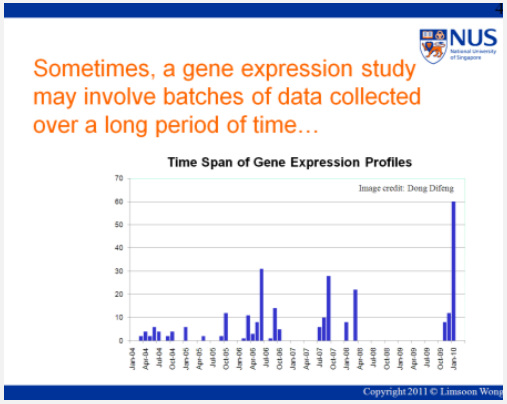
Some independent datasets are not as independent as you think

Batch effects

This is an important issue in analyzing clinical and many other types of real-world data

Discuss another time….

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



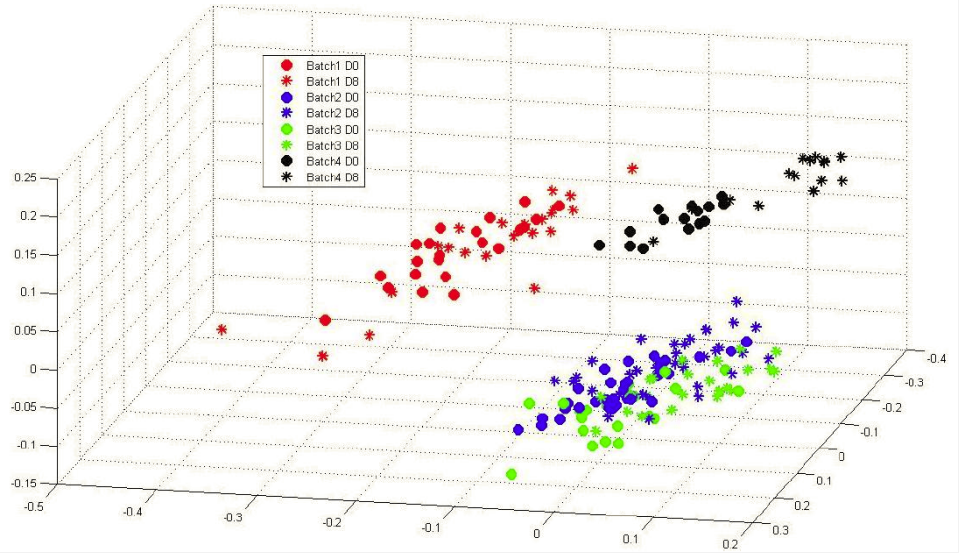
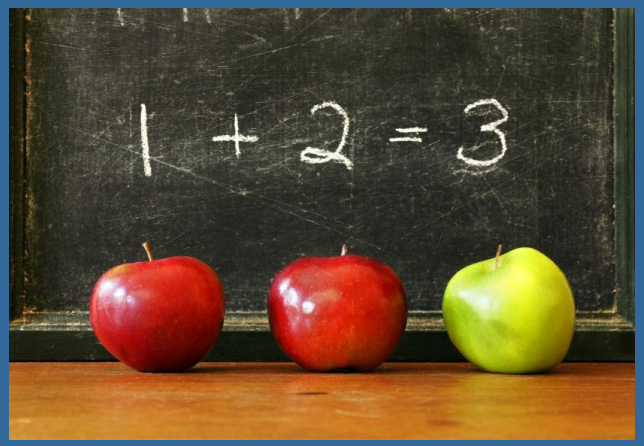


Image credit: Difeng Dong's PhD dissertation, 2011

Summary



Wong, **Big data and a bewildered lay analyst**, *Statistics & Probability Letters*, 136:73-77, 2018

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.

It is easy to make mistakes when analyzing data

Think in terms of contingency tables; i.e. compare & contrast

Look for subpopulations causing exception, contradiction, & trend strengthening

Mechanical use of data mining, statistical test, etc. can only take you so far