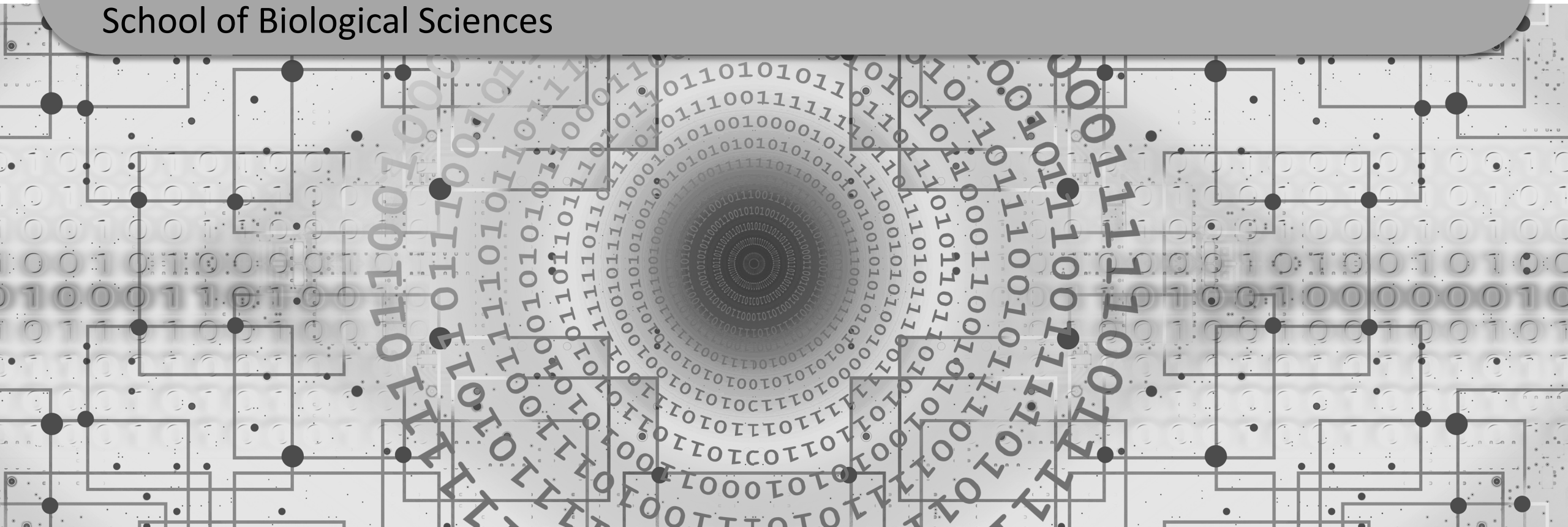
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

How Statistics go Wrong in Data Science

BS0004 Introduction to Data Science

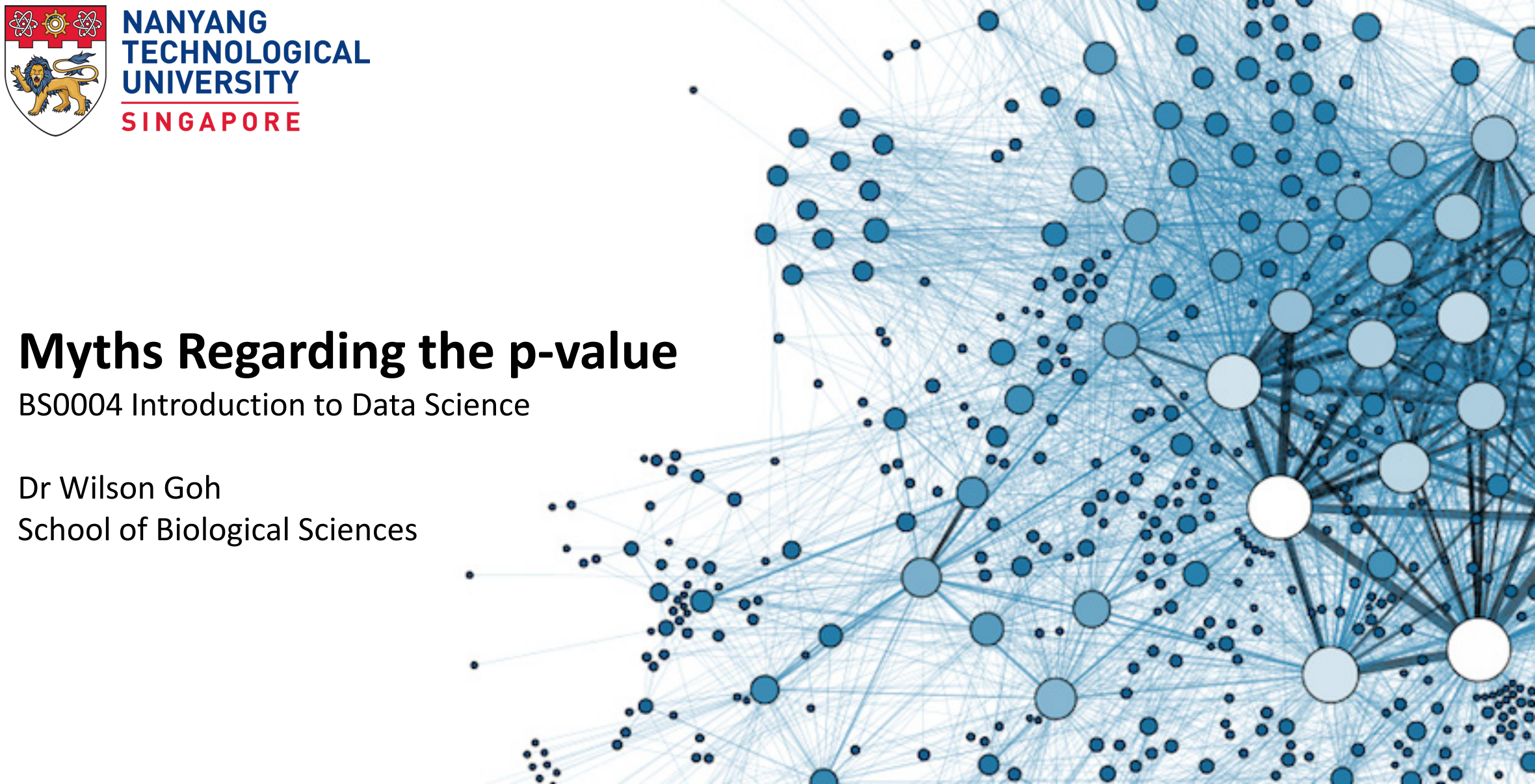
Dr Wilson Goh
School of Biological Sciences



Learning Objectives

By the end of this topic, you should be able to:

- Explain the myths regarding the p-value.
- Describe p-value instability and its implications.
- Describe the various approaches for checking reproducibility.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Myths Regarding the p-value

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Revisiting the p-value

The p-value is the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.

The p-value is widely used in statistical hypothesis testing, specifically in null hypothesis significance testing.

The p-value is never meant to be an absolute indicator of whether a hypothesis is correct or not.

Why do we love the p-value?

p-values are easy to calculate, even for complicated statistics. Many statistics do not lend themselves to easy analytic calculation; but using permutation and bootstrap procedures p-values can be calculated even for very complicated statistics.

p-values are relatively easy to understand. The p-value is a point metric bounded between 0 and 1 where 0 is more significant, 1 is not significant.

p-values have simple, universal properties. p-values come from the same background distirbution

Why do we love the p-value?

p-values are calibrated to error rates scientists care about Regardless of the underlying statistic, calling all P-values less than 0.05 significant leads to on average about 5% false positives even if the null hypothesis is always true.

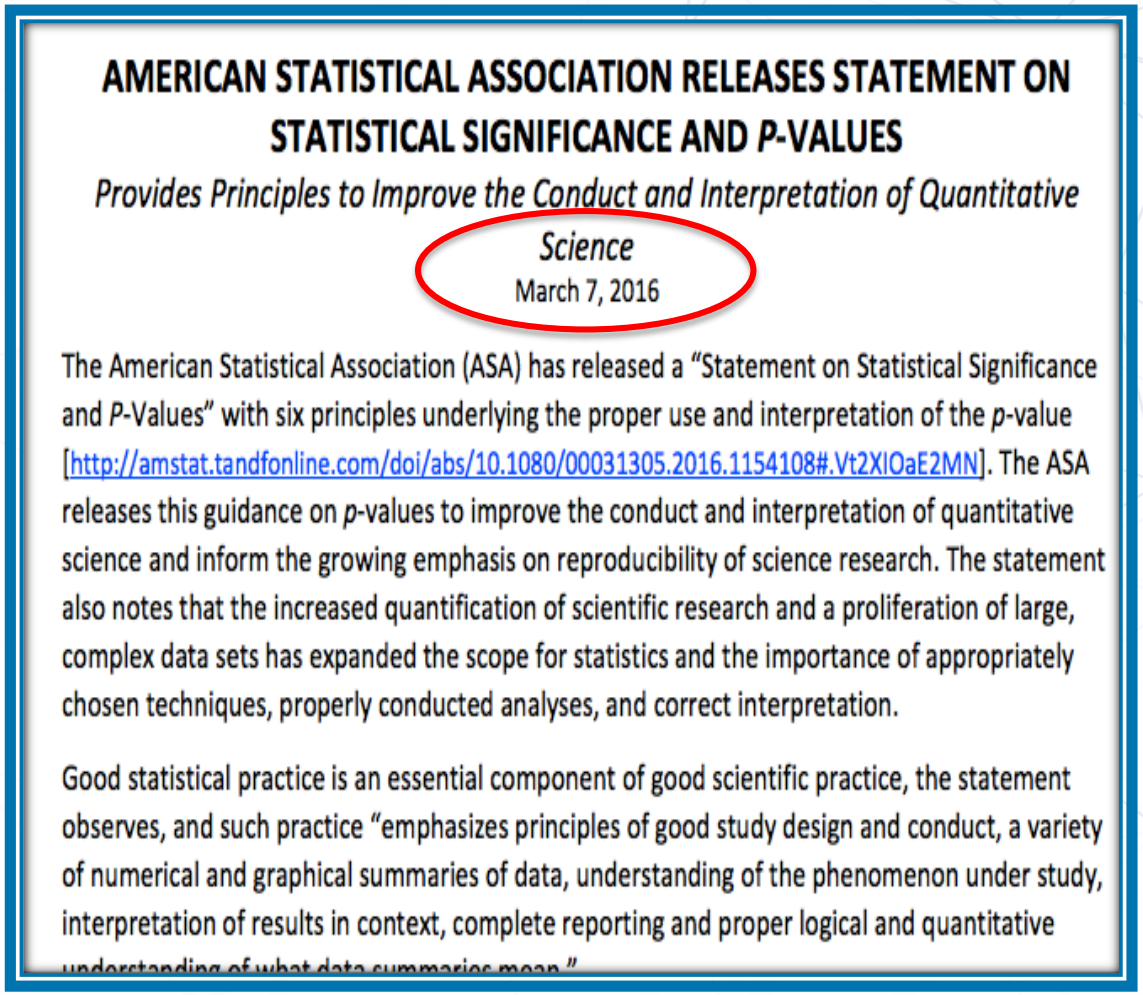
p-values are useful for multiple testing correction. The advent of new measurement technology has shifted much of science from hypothesis driven to discovery driven making the existing multiple testing machinery useful. Using the simple, universal properties of p-values it is possible to easily calculate estimates of quantities like the false discovery rate - the rate at which discovered associations are false.

p-values are "reproducible". All statistics are reproducible with enough information. Given the simplicity of calculating p-values, it is relatively easy to communicate sufficient information to reproduce them [note that in this case: it just means if you use the exact same data and same method, you will get the same p-value. It does not refer to statistical reproducibility given many resampling from population].

p-value is Misused and Misunderstood

P-value:

- Is not reliable.
- Is not reproducible.
- Does not relate directly with effect size.
- Is still an area of active research and philosophical debates.



AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

Source: Wasserstein and Lazer. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 2016

p-value is Misused and Misunderstood

"**The p-value was never intended to be a substitute for scientific reasoning.** Well-reasoned statistical arguments contain **much more** than the value of a single number and whether that number exceeds an arbitrary threshold."

---Ron Wasserstein, ASA Executive Director

The ASA statement is intended to steer research into a **post p<0.05 era**.

Source: Wasserstein and Lazer. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 2016

What the p-value does or is?

- p-values can indicate how **incompatible** the data are with a specified statistical model.
- p-values do not measure the probability that the studied hypothesis is true.
- p-value **does not** measure the size of an effect or the importance of a result.

- Scientific conclusions and business or policy decisions **should not** be based only on whether a p-value passes a specific threshold.
- Proper inference requires **full reporting** and transparency.
- **By itself**, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Source: Wasserstein and Lazer. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 2016

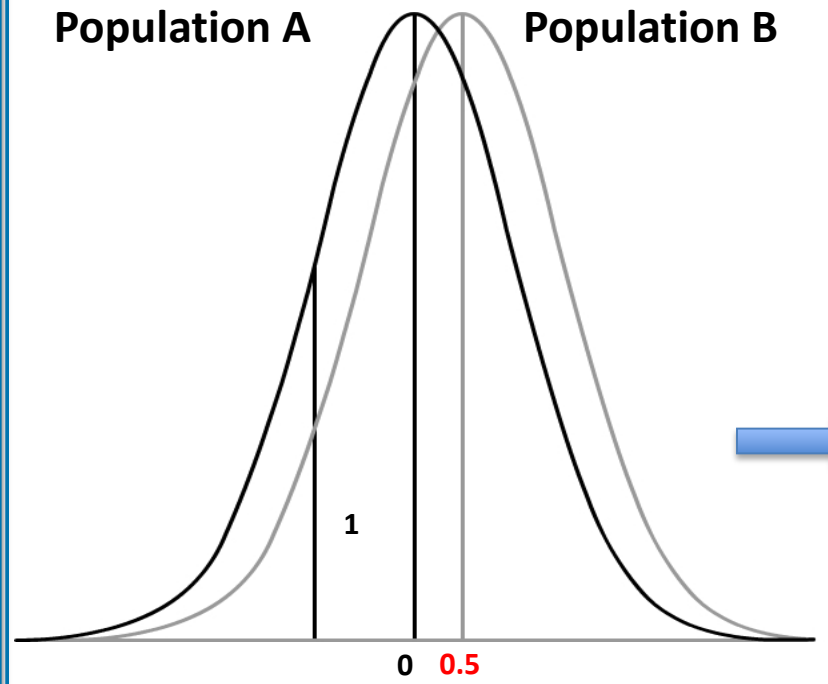
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

p-value Instability Issues and Implications

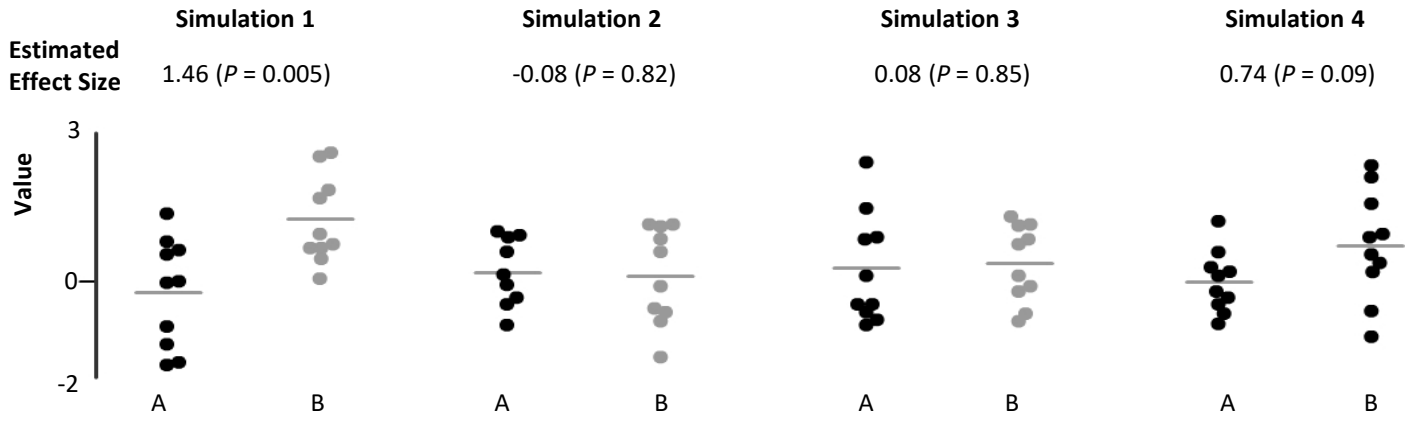
BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Demonstrating Instability



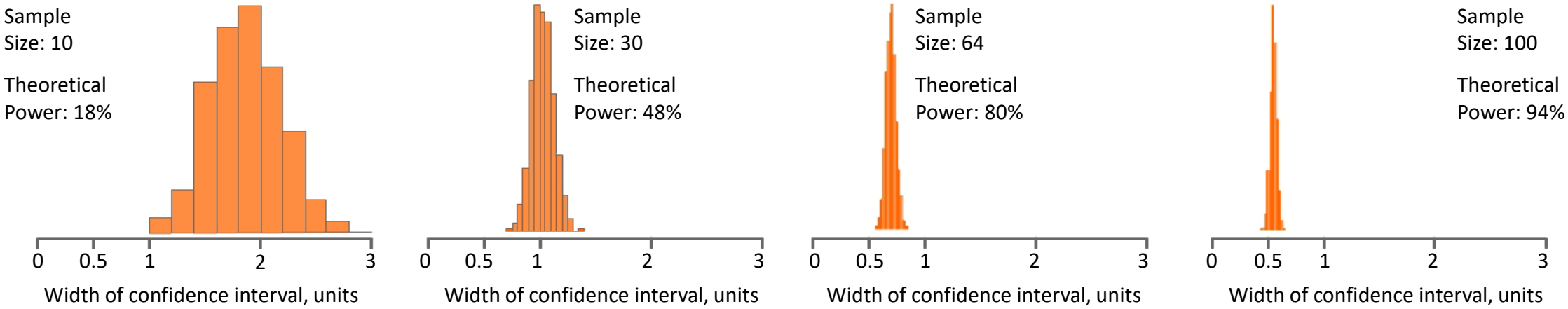
Simulated data distributions of two populations. The difference between the mean values is 0.5, which is the true (population) effect size. The standard deviation (the spread of values) of each population is 1.



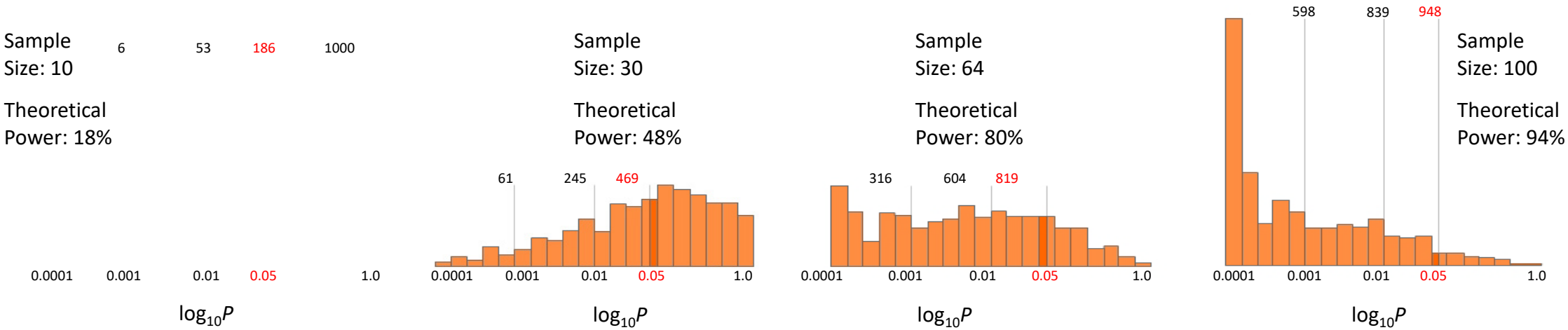
Small samples show substantial variation.

Source: Halsey et al, The fickle P value generates irreproducible results, Nature Methods, 2015

Demonstrating Instability



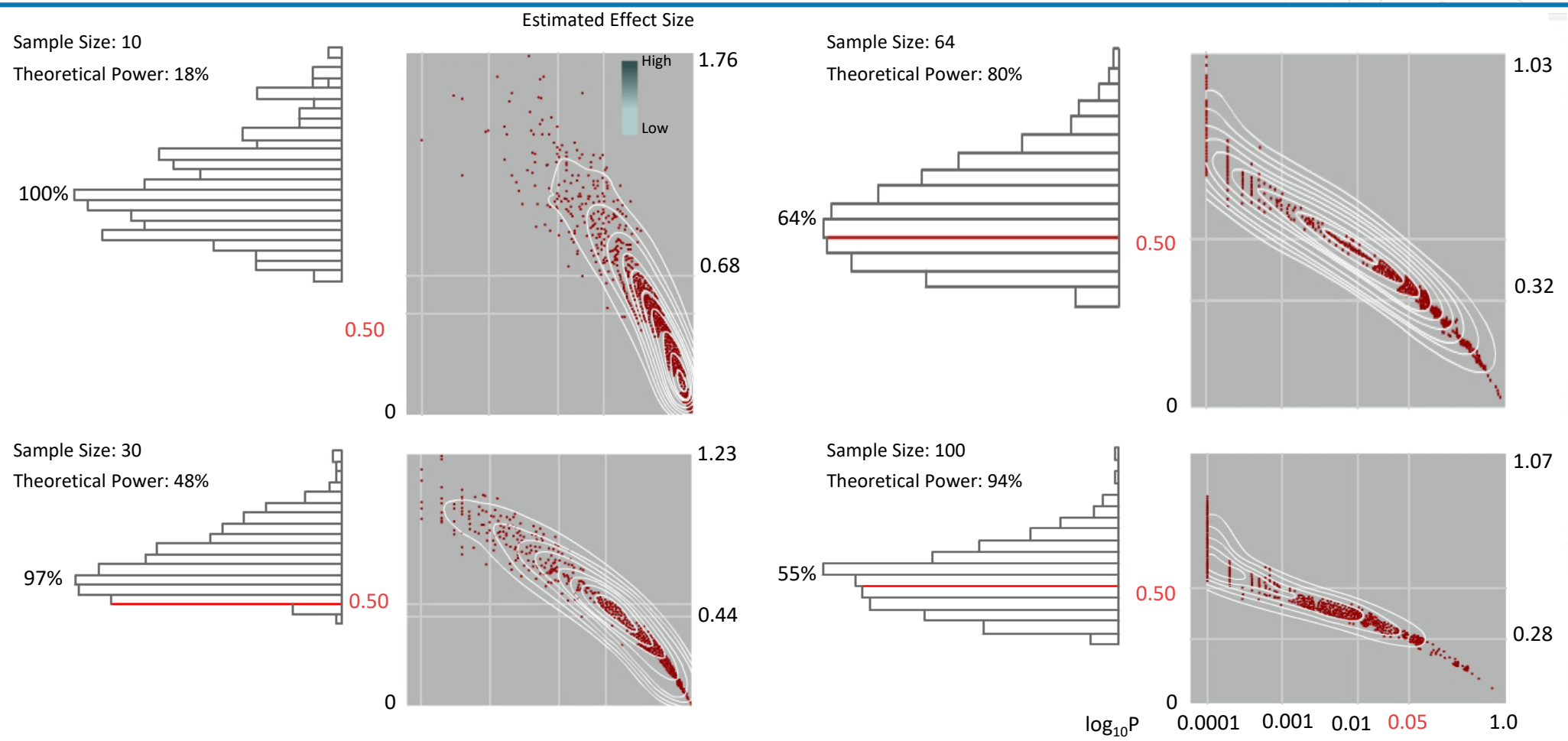
A larger sample size estimates effect size more precisely based on the CI but…



Individual p-values still vary widely even when sample size is high (pass the alpha but unstable).

Source: Halsey et al, The fickle P value generates irreproducible results, Nature Methods, 2015

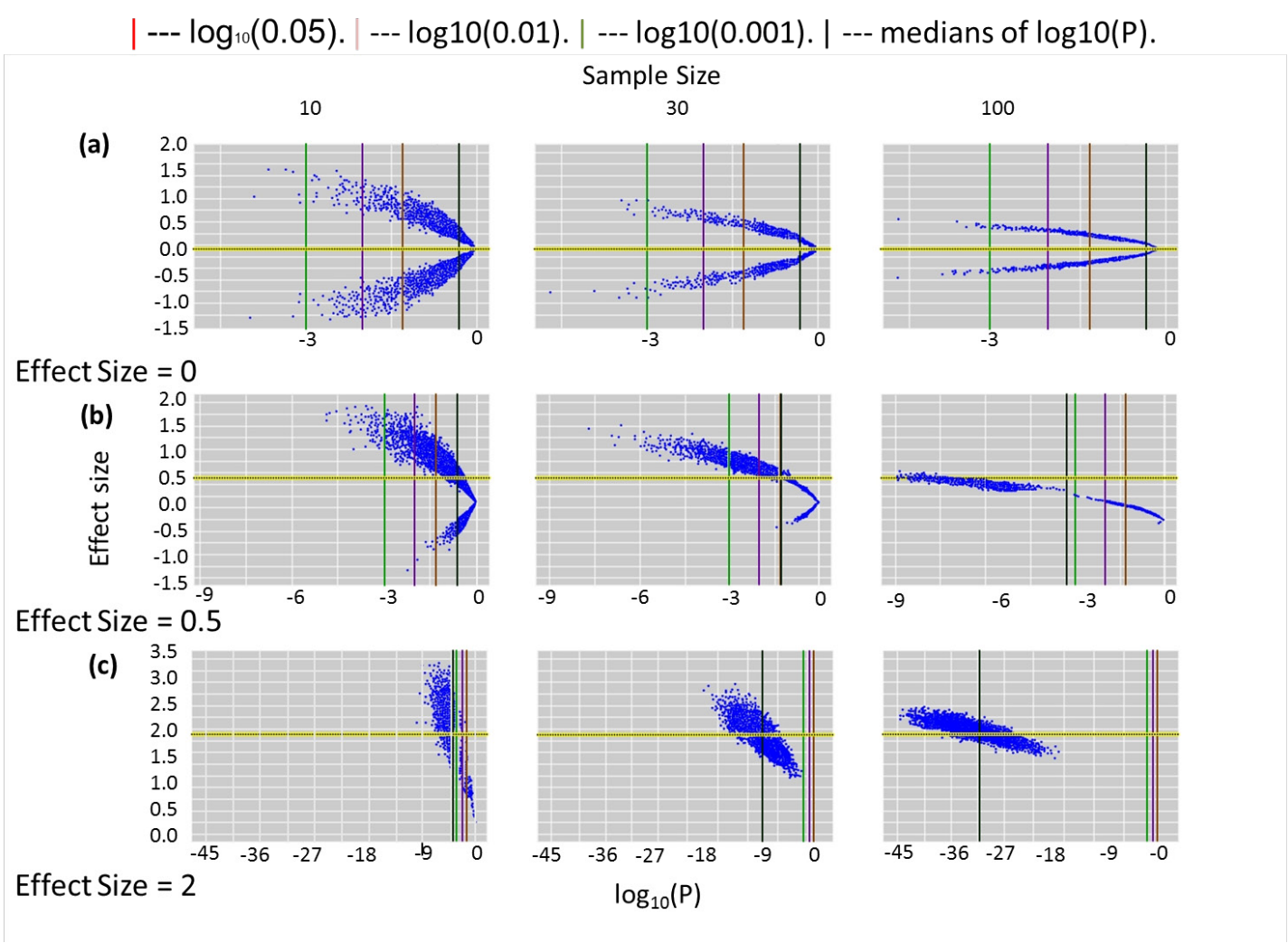
Demonstrating Instability



Wide p-value variability/instability gives rise also to gross misestimating of the true effect size. This is not resolvable by simply increasing sample size.

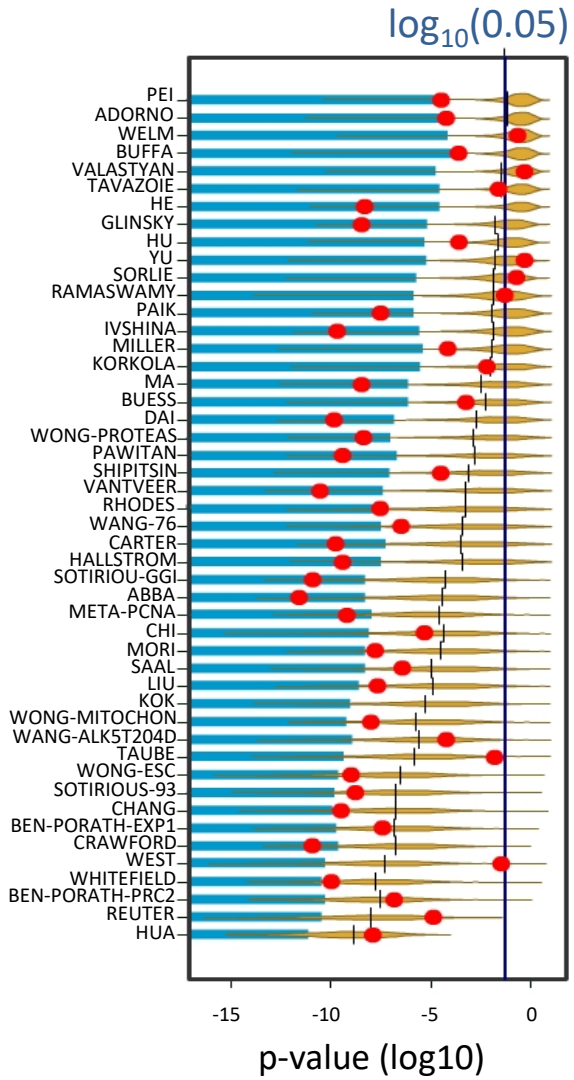
Source: Halsey et al, The fickle P value generates irreproducible results, Nature Methods, 2015

Demonstrating Instability



Source: Wang et al. Feature selection in clinical proteomics: with great power comes great reproducibility. Drug Discovery Today, 2017

Possible Consequence of p-value Instability – Useless Signatures



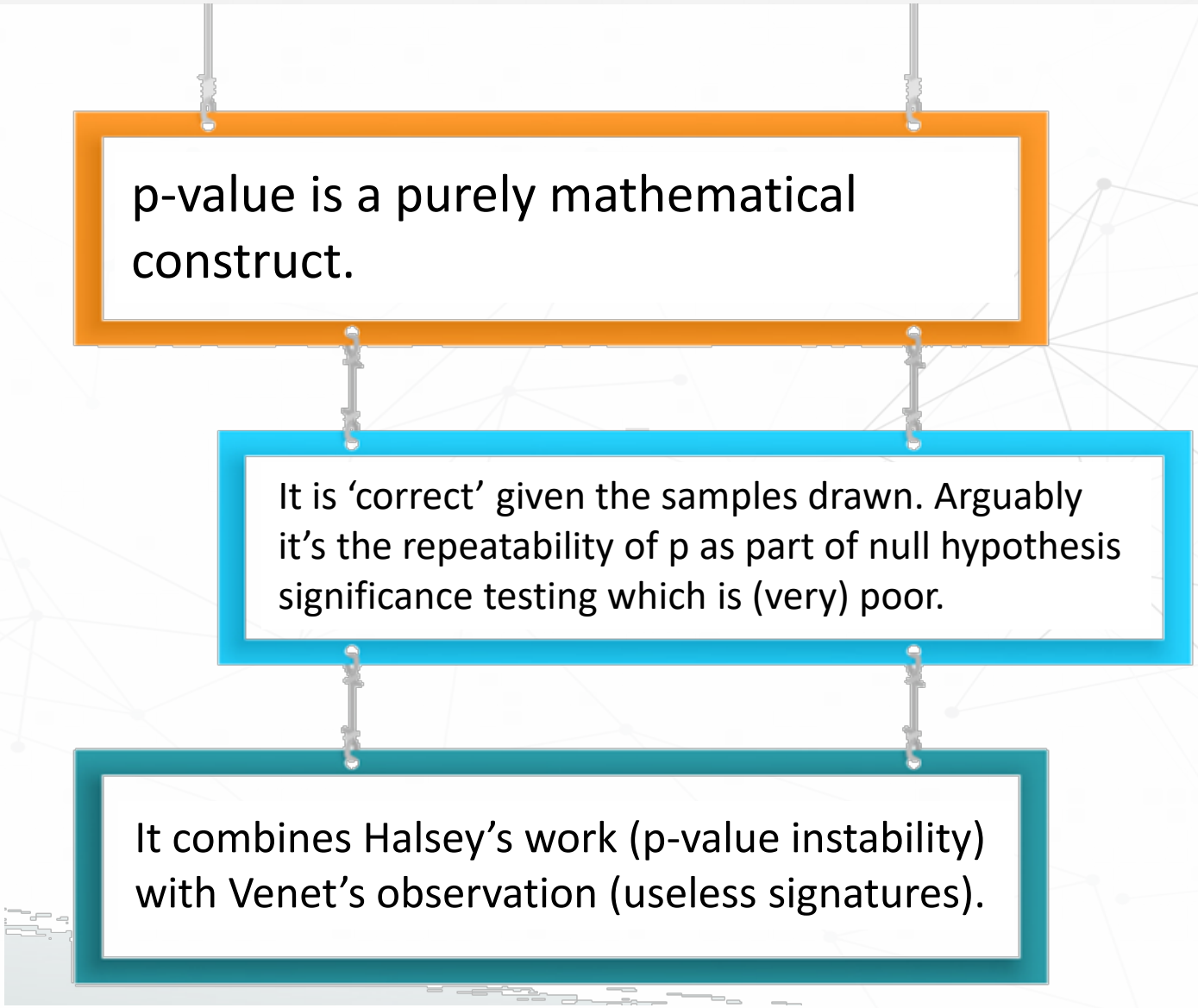
The x-axis denotes the p-value of association with overall survival. Red dots stand for published signatures, yellow shapes depict the distribution of p-values for 1000 random signatures of identical size, with the lower 5% quantiles shaded in green and the median shown as black line. Signatures are ordered by increasing sizes.

Although the signatures were shown to be significant In one study, their association with survival turns out to be non-significantly better than random signatures. *The red dots lies within the yellow shapes.

The suspected culprit? --- p-value instability

Source: Venet et al, Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome, PLOS 2011

Useless Signatures Arise from p-value Instability



Source: Wang and Goh, Sample-to-sample p-value variability and its implications in multivariate analysis, IJBRA, 2018

p-value Instability and its Implications for Multiple Testing

A toy multivariate scenario involving two groups 1 and 2 across 4 genes with arbitrarily defined means.

Genes	Group 1 (means)	Group 2 (means)
A	0	0.5
B	0	1
C	0	2
D	0	3

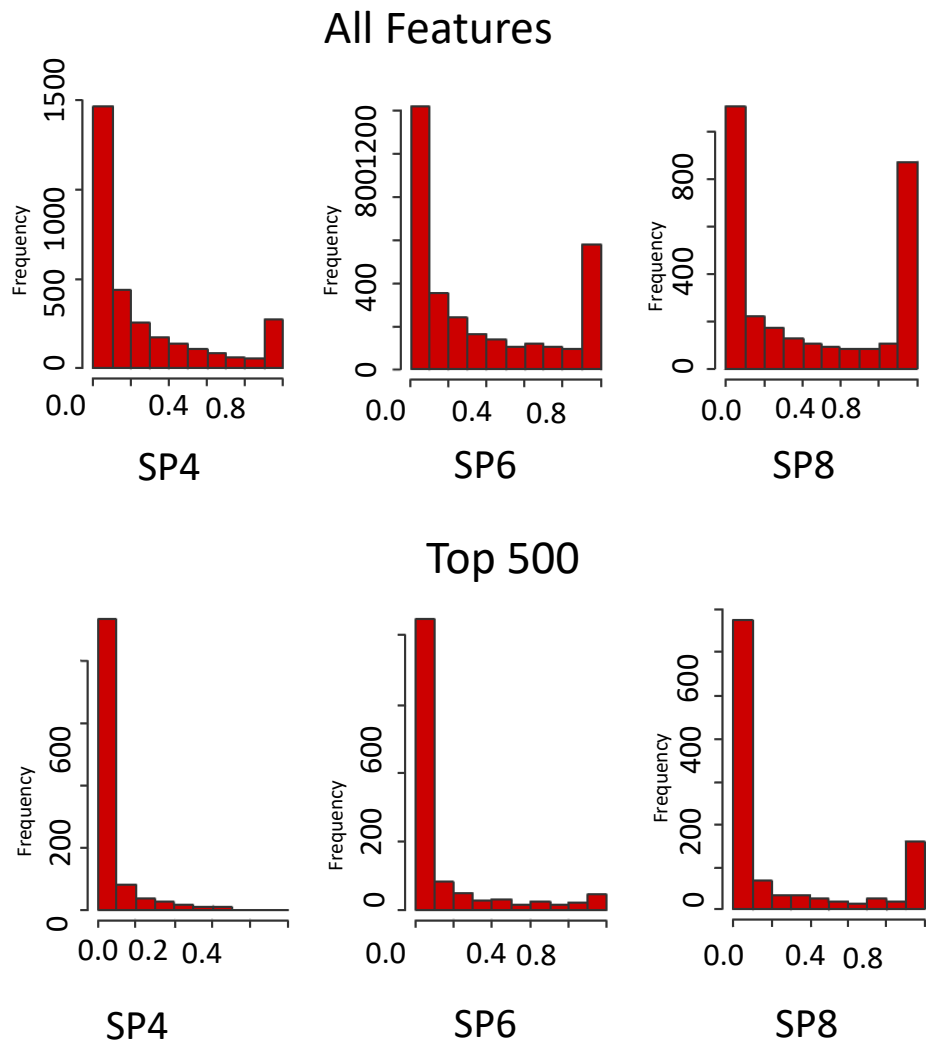
With small sample size, the rank orderings are wrong 40% of the time.

The proportion of correct orderings given sampling size increments.

Sampling Size	10	30	64	100
Correct Ordering/ Total Resamplings	0.57	0.90	0.96	0.99

Source: Wang and Goh, Sample-to-sample p-value variability and its implications in multivariate analysis, IJBRA, 2018

p-value Instability and its Implications for Multiple Testing



- On real data, take samples of size 4, 6 and 8, 1000 times. Do the t-test. Take all features with $p < 0.05$. Count the number of times each feature is significant.
- Repeat same experiment. But limit to only top 500 features each time (rank by p-value).
- Plot both results. Most of the top 500 features do not turn up consistently across resamplings.
- Therefore, the top features in real data are not stable and the signature is not stable.
- Now, think about what happens if we do Bonferroni correction. Will it be a useful procedure?

Source: Wang and Goh, Sample-to-sample p-value variability and its implications in multivariate analysis, IJBRA, 2018

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Evaluation of Performance

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

A Good Feature-selection Method is Reproducible

Different methods produce different p-values. It is imperative that the method gives stable p-value.

Good method must be able to make consistent and reproducible selections, **even at small sample size**.

Across different resamplings from the population, we expect a good method to return similar feature set.

It is important to check the sampling-to-sampling p-value distribution (spread of p-value) or alternatively, check the number of times a feature is significant over all simulations.

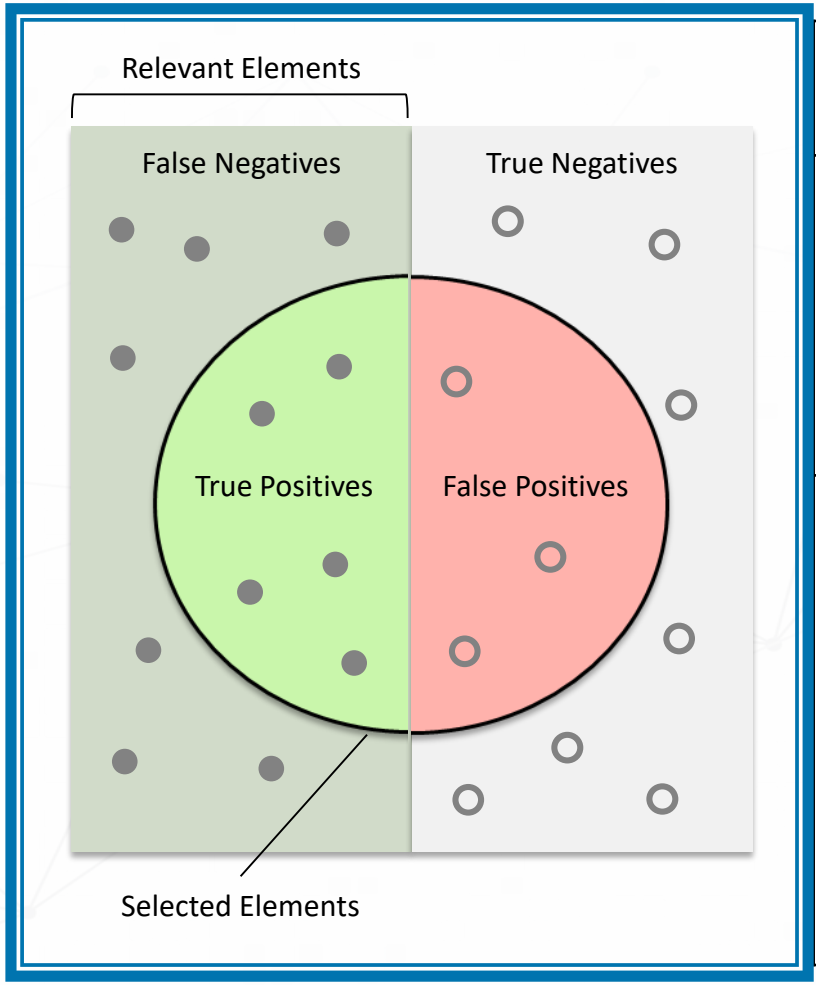
Source: Goh and Wong, Advancing Clinical Proteomics via Analysis Based on Biological Complexes: A Tale of Five Paradigms. JPR, 2016

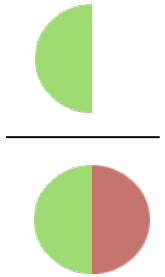
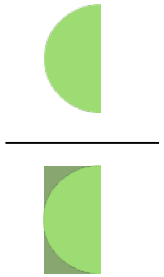
How to reduce naive reliance on the p-value?

Feature-selection stability	Cross-technical replicate reproducibility	False positive analysis via class-label reshuffling (resampling statistics)	Cross-validation
Some terminologies:	If the p-value does not change much, then we say that the p-value is stable.	If the feature set is always the same across resamplings, we say that the feature set is reproducible.	p-value stability implies feature-selection reproducibility.

Some Evaluation Metrics

Elements = Features



How many selected items are relevant?	How many relevant items are selected?
Precision = 	Recall = 

Precision: Of the selected feature, how many are correct?

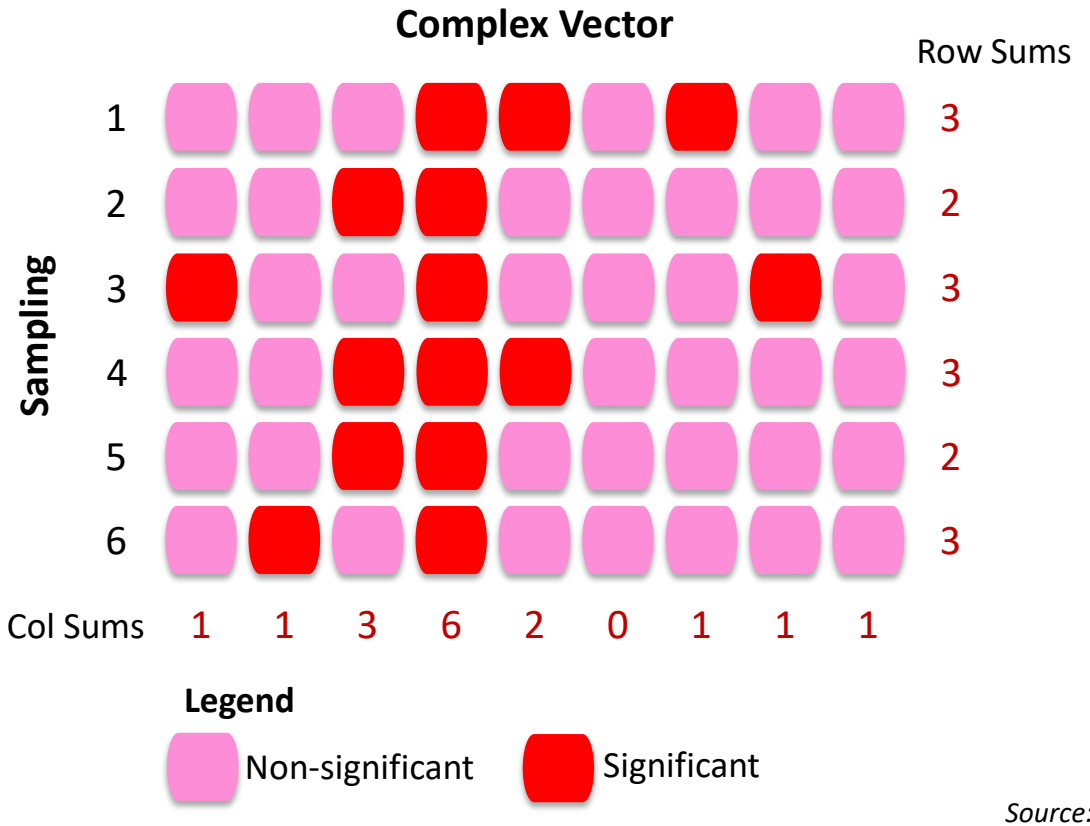
Recall: Of the selected feature, what is the proportion of all the correct ones we got?

Precision and recall can be combined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Feature-selection Stability

The binary matrix is useful for comparing stability and consistency of significant features produced by some feature-selection method.

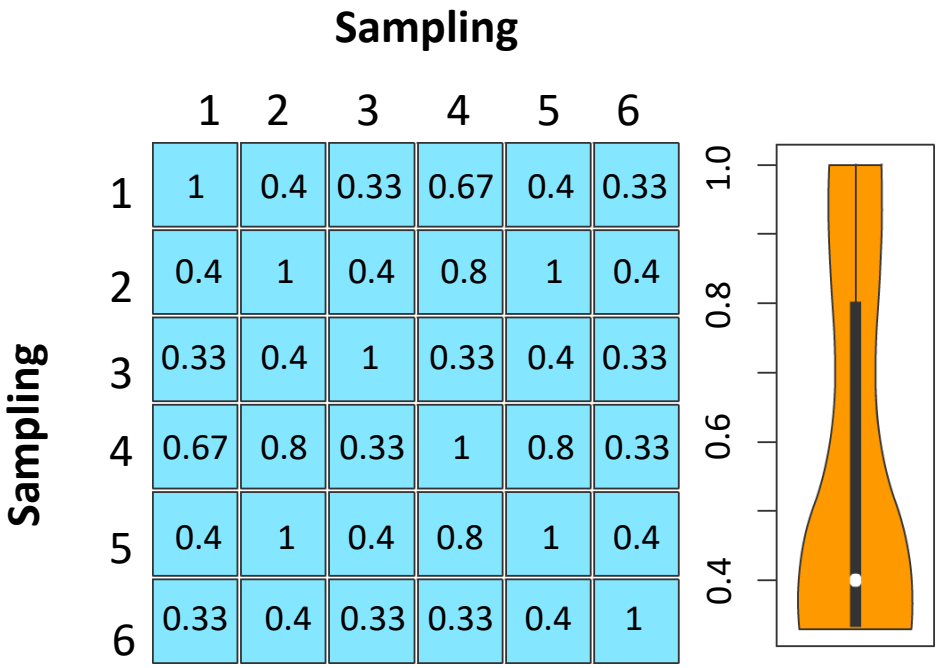


The rows represent each simulation. The columns are a nominal feature vector. Red represents features reported as significant while pink are non-significant. The row sums provides information on the number of significant features while the column sums provide information on the relative stability of each feature (i.e., out of n simulations, how many times is the feature reported as significant).

Source: Goh & Wong, Design principles for clinical network-based proteomics. Drug Discovery Today, 2016

Feature-selection Stability

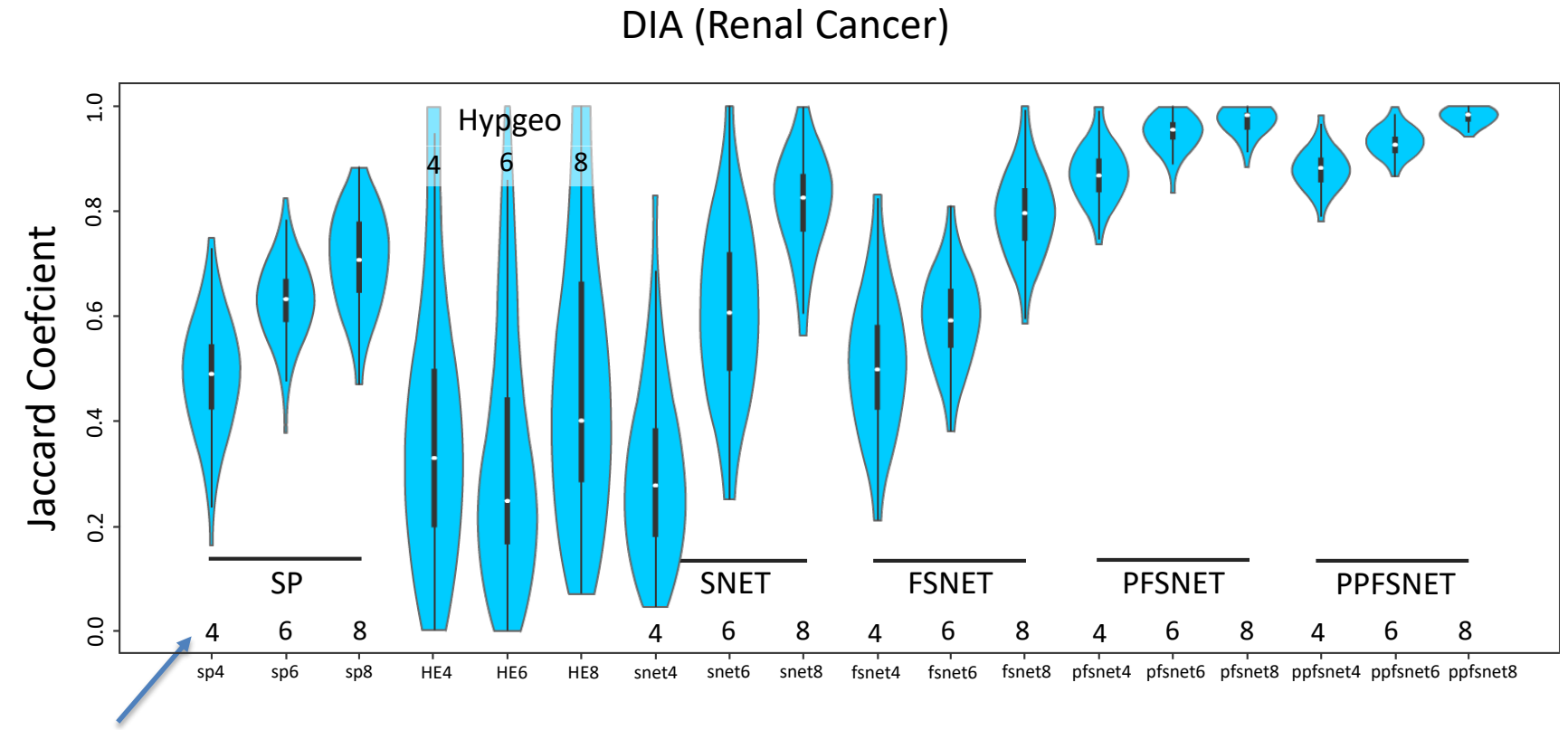
Inter-sample agreement rates. The violin shows that the inter-sampling similarity has a very wide IQR. So what can we say about the stability of the method? Is this a global or local estimate?



Similarity can be examined between simulations. Here the Jaccard score (intersection/ union) is used to calculate pairwise similarity. The distribution of pairwise similarities can be summarised diagrammatically using the violin plot, which provides centrality information such as the median and inter-quartile range, and distribution information based on the kernal density, making for easier analysis.

Source: Goh & Wong, Design principles for clinical network-based proteomics. Drug Discovery Today, 2016

Inter-sample Agreement Rates



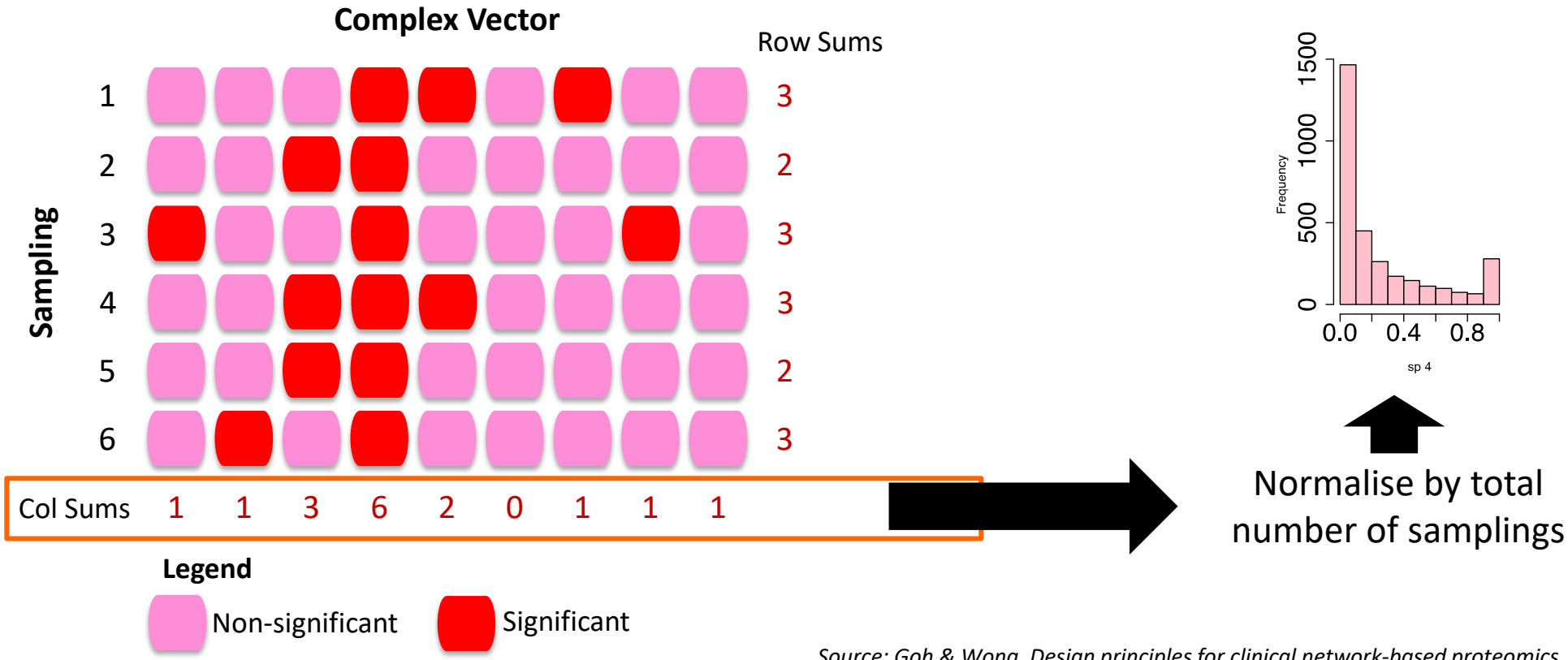
Sample Size

Jaccard = Intersection(A,B)/ Union(A,B)

Which method here has the highest inter-sample agreement rate?

Feature-selection Stability

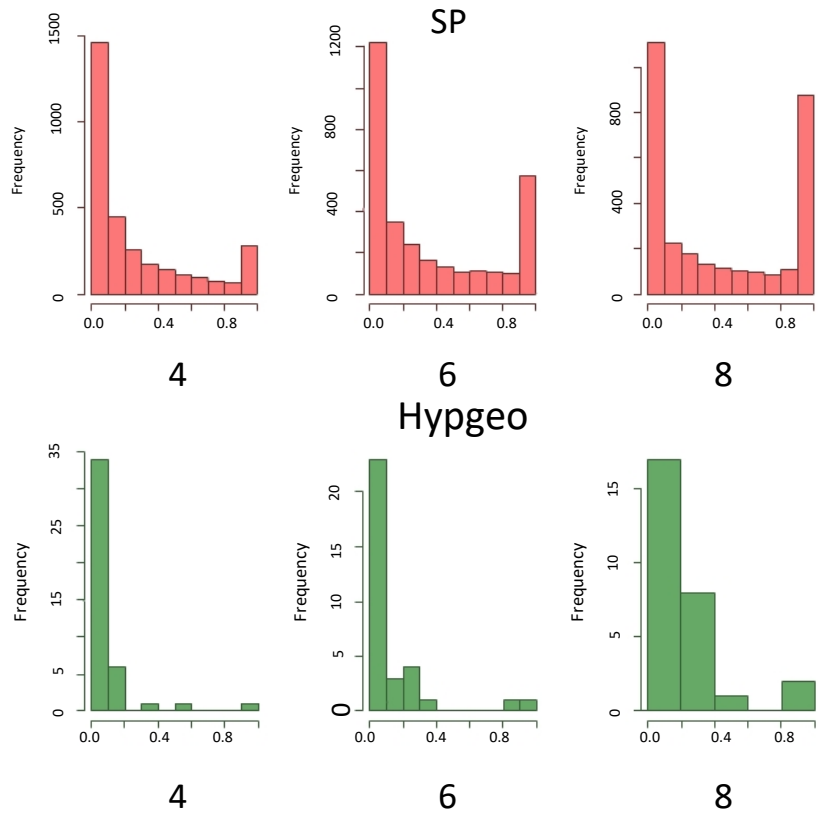
Individual feature reproducibility

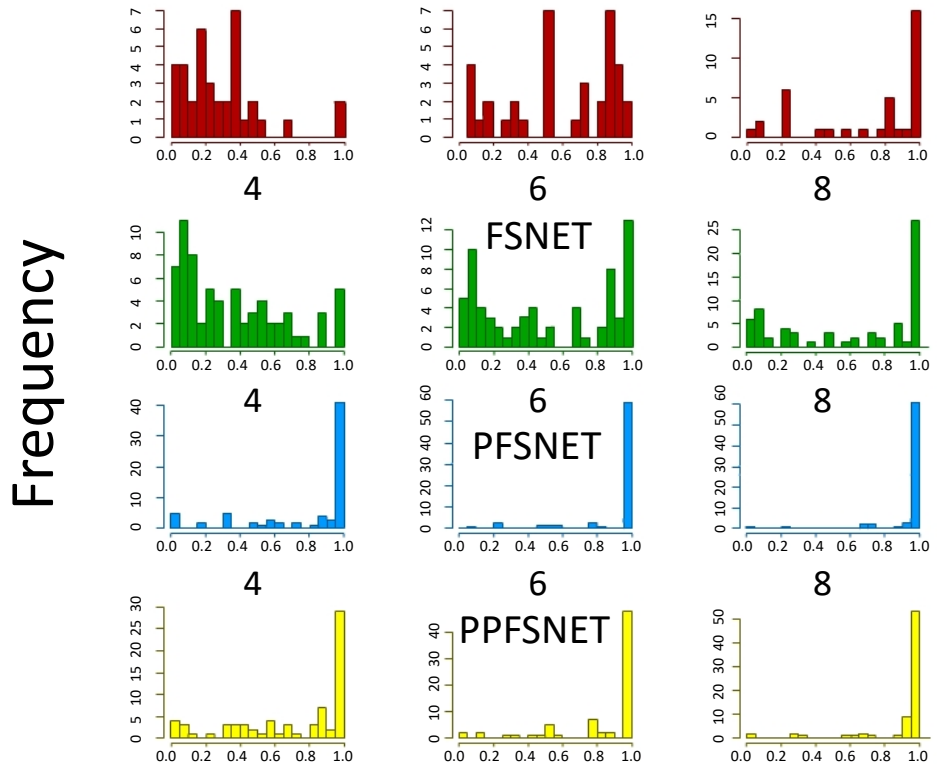


Normalise by total number of samplings

Source: Goh & Wong, Design principles for clinical network-based proteomics. Drug Discovery Today, 2016

Individual Feature Reproducibility

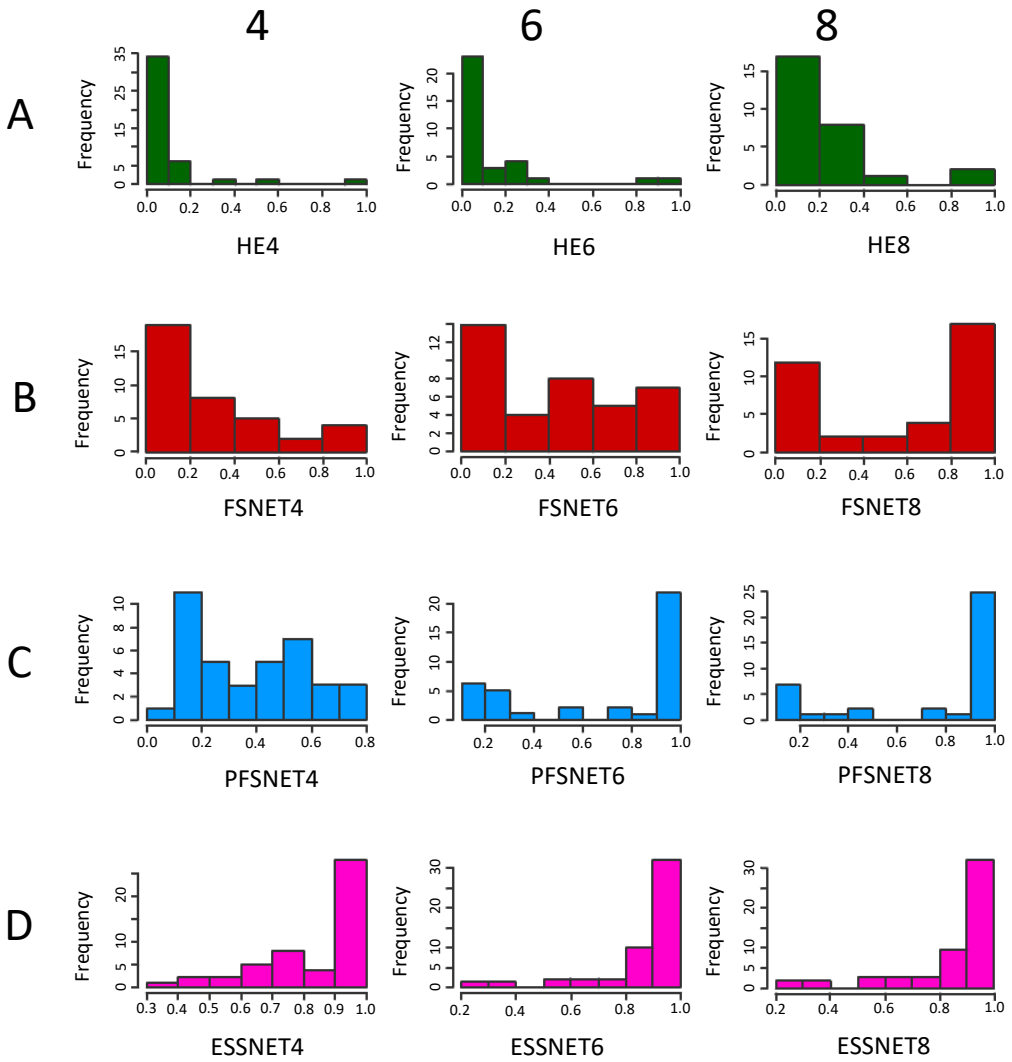




0 --- feature only observed once
1 --- feature observed all the time

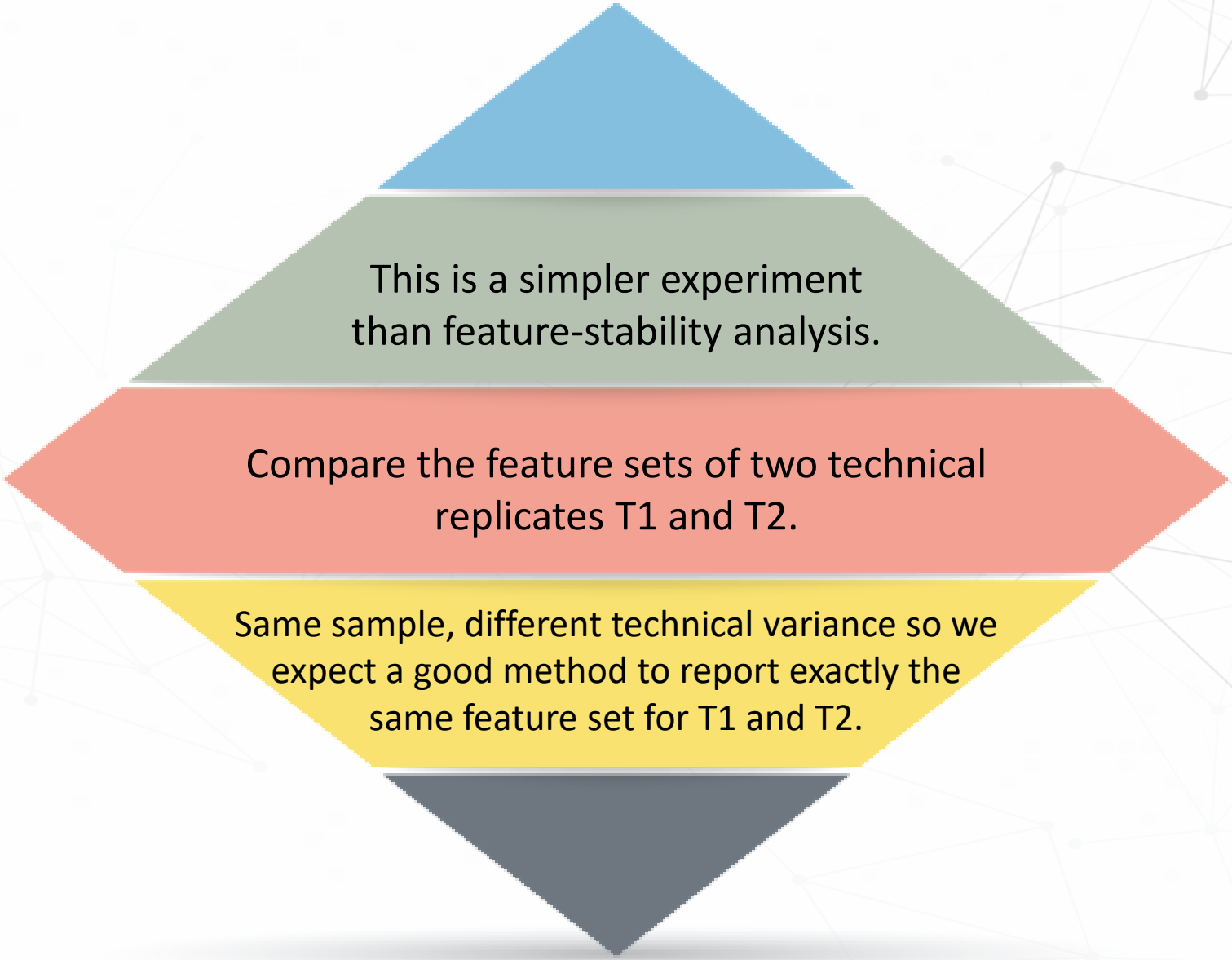
Which feature-selection method shows that its predictions are always stable?

Power at Small Sample Sizes



- Resampling analysis at various sample sizes also tells about the sensitivity of a method.
- Which method (A-D) works well with small sample sizes?
- With large sample sizes, we are more likely to get strong signal from all the relevant features.
- This is much harder with small sample sizes, and the method needs to be able to be sensitive to weak signal. But at the same time, have controlled false positive rates (non-hypersensitivity).

Cross-technical Replicate Reproducibility



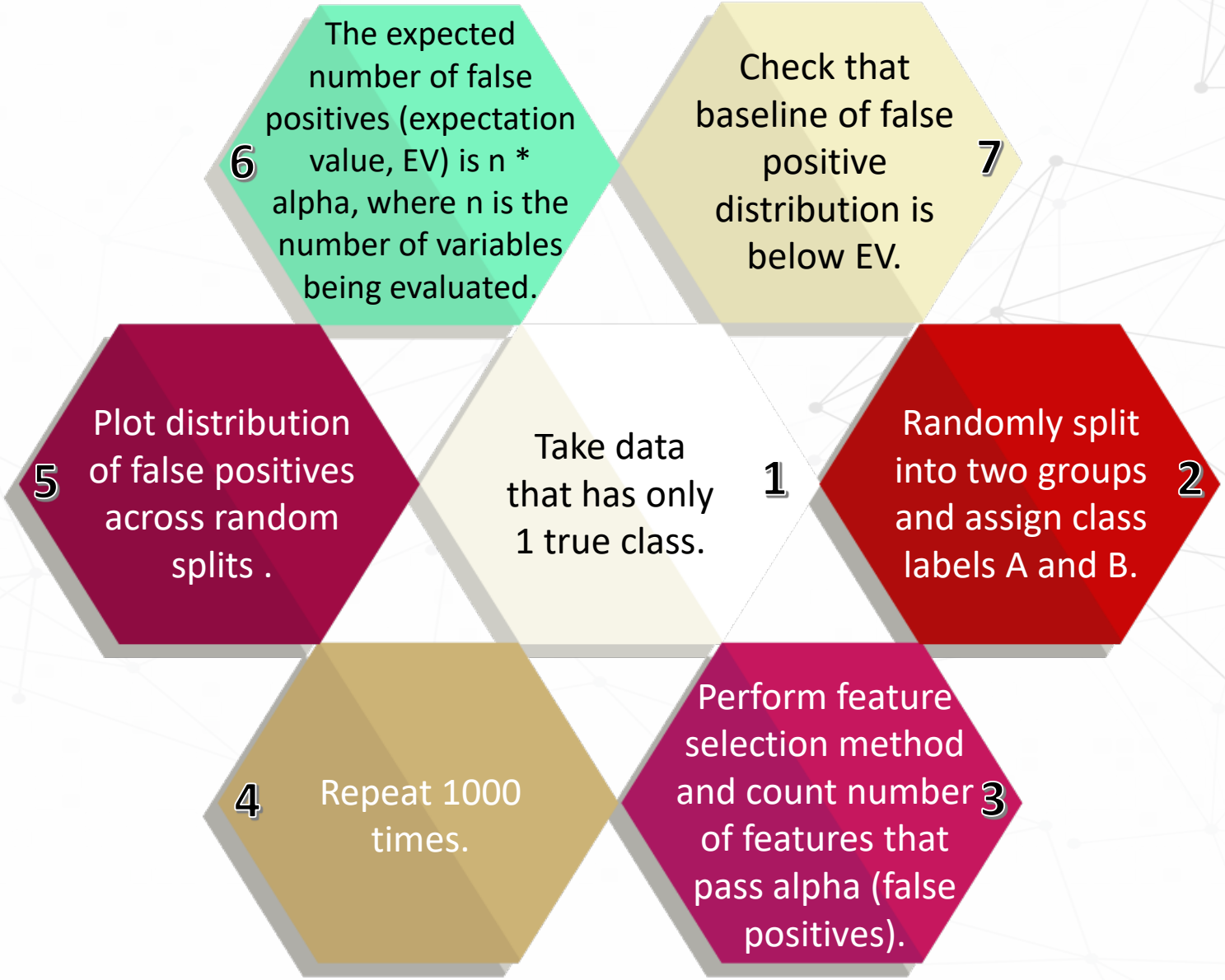
Cross-technical Replicate Reproducibility

Direct cross-replicate reproducibility per method:

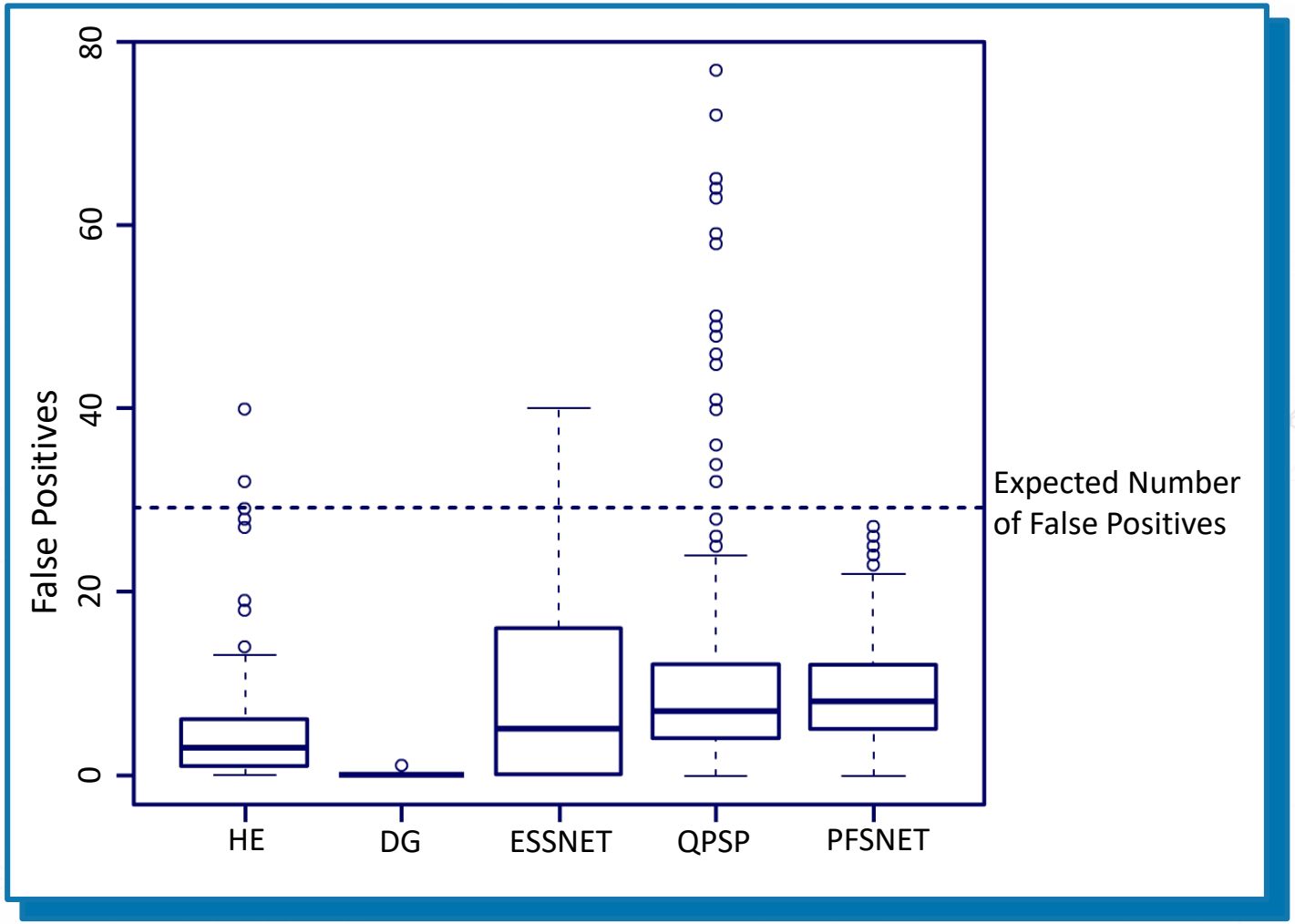
Number of Terms	HE	DG	ESSNET	QPSP	PFSNET
Replicate 1	4	1	35	86	45
Replicate 2	6	2	29	75	46
Overlaps	0.25	0.5	0.83	0.66	0.94

Source: Goh and Wong, Advancing Clinical Proteomics via Analysis Based on Biological Complexes: A Tale of Five Paradigms. JPR, 2016

False Positive Analysis via Class-label Reshuffling



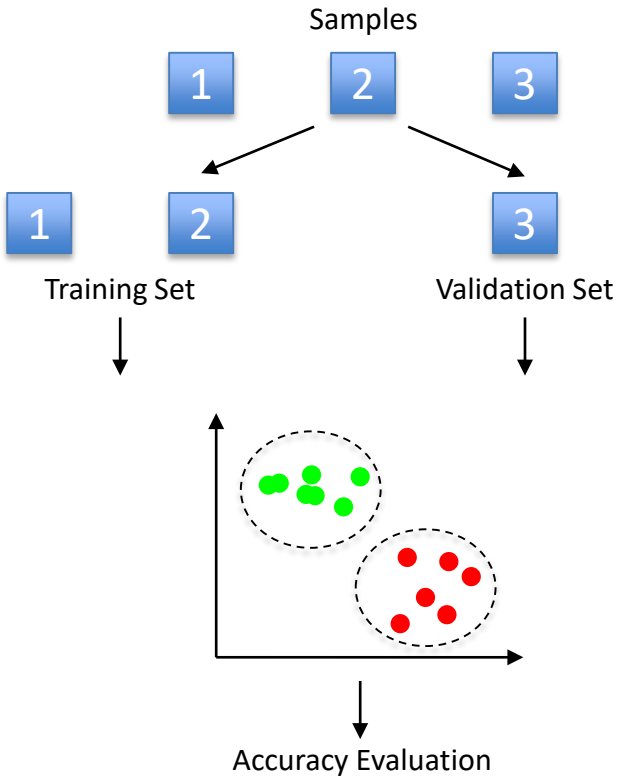
False Positive Analysis via Class-label Reshuffling



Source: Goh and Wong, Advancing Clinical Proteomics via Analysis Based on Biological Complexes: A Tale of Five Paradigms. JPR, 2016

Cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalise to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.



Cross-validation Accuracy is Not Reliable

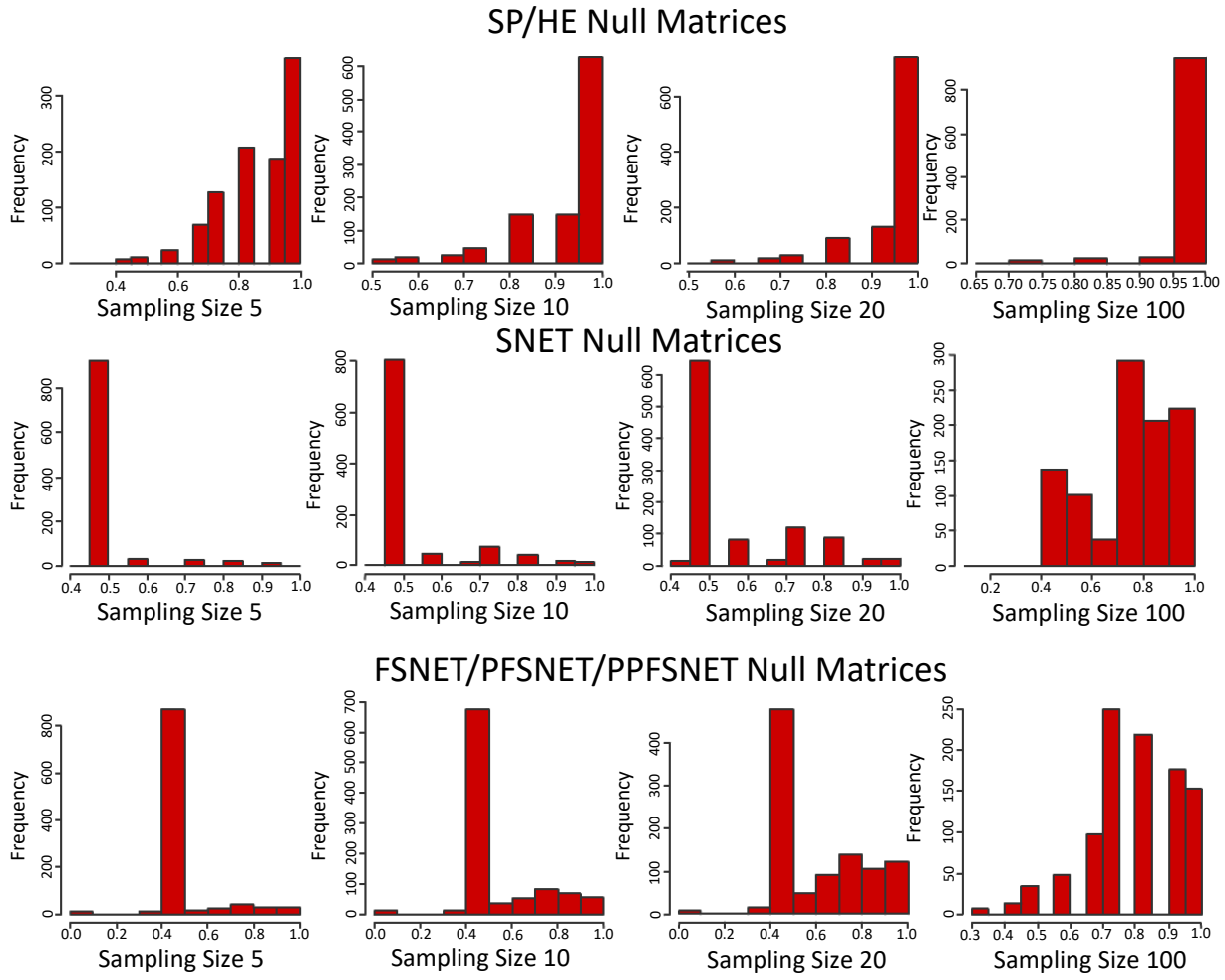
The numbers lie. Cross-validation by itself is not enough.

Method	Number of Features	CV of Accuracy
SP	1124	0.98
HE	162	0.98
SNET	0.25	0.5
FSNET	36	0.96
PFSNET	65	0.92
PPFSNET	66	0.96

All too high
To be true

Source: Goh and Wong. Evaluating feature-selection stability in next-generation proteomics, JBCB, 2016

Cross-validation Accuracy p-value



Source: Goh and Wong, JBCB, 2016

- Any random resampling can produce accuracies equal or better than the observed feature-set. We may define a p-value for the cross-validation accuracy based on resampling statistics where CV p-value = CV_accuracy|rand > obs|/|rand|
- The sampling size is equal to number of selected features in observed data.
- For certain methods, any random selection of genes will give high CV accuracy. So this means that the CV accuracy on its own is useless. SP is the two-sample t-test. What does this result tell you?

Normalised Cross-validation Accuracy

We normalise the CV accuracies by the CV p-value. Now we can see which method gives real meaningful information.

Method	Number 7 Features	CV7 Accuracy	CV7 p=val	CV7 Accuracy/pval
SP	1124	0.98	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

Now we know whether a high CV accuracy is meaningful or not.

Source: Goh and Wong, JBCB, 2016

Putting Together a Feature-selection Method Evaluation

Precision and Recall

Method/ Sampling Size				

	eature-stability scores			
Method/ Sampling Size				

Feature-stability

C

Normalised CV Accuracy

D

Combined Ranks

+ False Positive Checks

Source: Goh and Wong, JBCB, 2016

But we are not yet done…

	F-scores (0.05)					
	HIV	Diabetes	D1.2.301	Breast Cancer	Renal Cancer	Colon Cancer
t-test	0.27	0.10	0.58	0.66	0.54	0.62
Wilcoxon Rank Sum Test	NaN	NaN	0.08	0.70	0.52	0.62
Limma	0.38	0.26	0.61	0.34	0.32	0.56
Rank Products	0.25	0.20	0.30	0.39	0.47	0.56
Kolmogorov Smimov Test	NaN	NaN	NaN	0.71	0.53	0.68

- Meaningless methods and meaningless data.
- Different data gives different results.
- Simulated data do not fit real data.
- Maybe no Feature-selection method is better than any other if performance is averaged across all possible situations – No free lunch theorem (Wolpert and Macready).

- Current evaluation of feature-selection methods based on real data are not using meaningful benchmarks. Real data are different from one another, so performance ranks will always change depending on which datasets are being used.
- Upstream and down-stream data processing will also affect performance ranks.
- Use of meaningful benchmarks will stabilise feature-selection evaluations.
- Current benchmarks should include surveys on the p-value stability of individual methods.

Summary

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. The p-value is unstable.
2. Only rank by effect size, not by p-values.
3. There are several ways of checking for statistical reproducibility.
4. Use other metrics such as the Cohen's D and confidence interval to help augment the p-values.