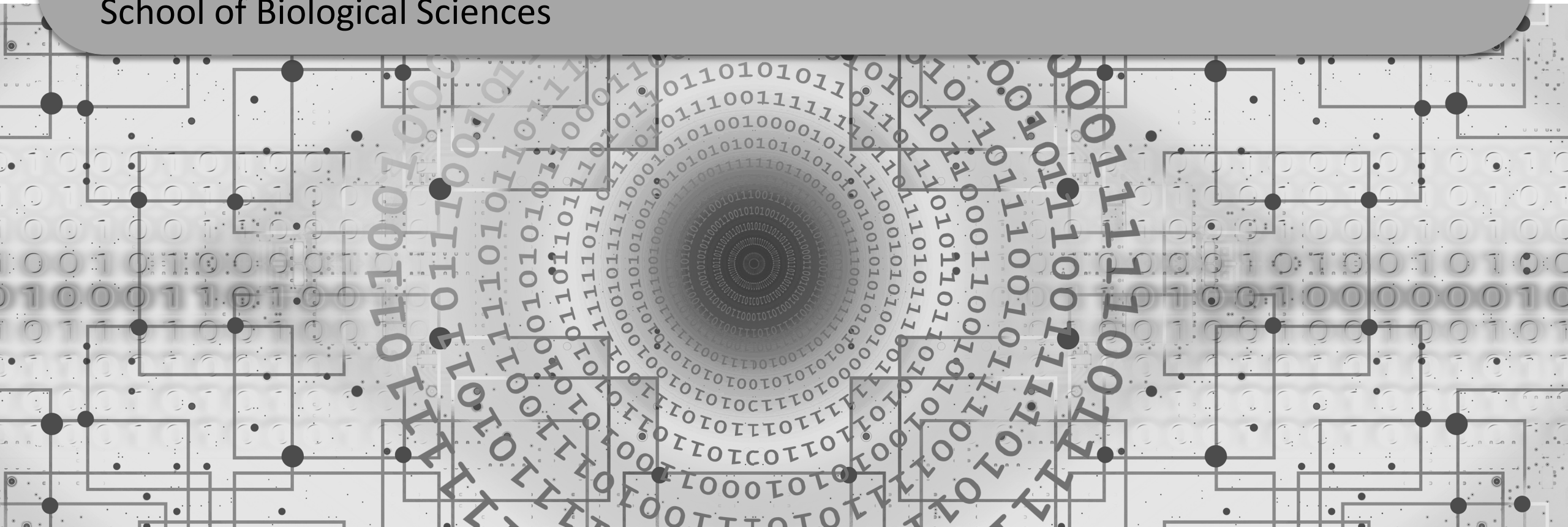


# Experimental Design and Confounders

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



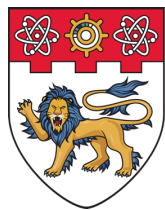
# Learning Objectives

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

- Describe the following terms in experimental design:
  - Bias and fallacies
  - Independent and Identically Distributed (IID)
  - Inclusion criteria
  - Confounders
  - Simpson's paradox
  - Batch effects
  - Domain-specific laws
  - Non-association
  - Context
  - Meta and mega analysis







**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

# **Bias and Fallacies**

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Bias

## Axiom:

- An unfair/tainted perspective.
- “The mind sees what it chooses to see.” --- *Robert Langdon, The da Vinci Code*

## Commonly encountered as follows:

- You see your favorite gene X turn up in a screen, you jump for joy.
- You believe gene X causes disease Y. You only look for evidence in support of your belief.

## How to avoid bias?

- Consider evidences objectively.
- Weigh-in/check your thinking with others to derive more fair-handed interpretations.



# Sampling/ Ascertainment Bias

Sample is collected such that it is non-representative of the actual population. Estimation of the population parameter from this sample is thus biased.

It can arise from :

- Self-selection
- Pre-screening (or advertising)

# Sampling/ Ascertainment Bias

In 1936 a postal survey was conducted to predict the next president of the USA.

The survey was comprised of readers of the American Literary Digest magazine, with additional responses from registered car and phone owners.

The survey predicted Alf Landon, the Republican candidate, would easily win. The actual election was an easy victory for Franklin Roosevelt.

What happened?



# Sampling/ Ascertainment Bias

The people surveyed were not randomly chosen and were not a statistically representative sample of the American population.

They were disproportionately rich, when compared to the average voter, and more likely to vote Republican.

# Cherry Picking

The act of only considering individual cases or data that confirms a particular position, while ignoring a significant portion of related cases or data that may contradict that position.

If I flipped a fair coin 100 times and I withheld half the data, I can convince you the coin has two heads.



# Publication Bias

A type of bias occurring in published academic research. Publication bias is of interest because literature reviews of claims about support for a hypothesis or values for a parameter will themselves be biased if the original literature is contaminated by publication bias.

# Publication Bias

In science, we only see the good stuff. But we never see what fails.

A positive study is 3x more likely to be published. So does this mean that scientists are smart people and always succeed in their projects? (you know this is not true!)

But what is more dangerous is that a commonly held but erroneous assertion is held to be truth, and only subsequent works that supports it are publishable, while works that do not support it are assumed to be due to be mistakes (or incompetence).



# Insensitivity to Sample Size

**Insensitivity to sample size** is a cognitive bias that occurs when the probability of obtaining a sample statistic is judged without respect to the sample size.

# Insensitivity to Sample Size

People tend to deploy “thinking shortcuts” or heuristics.

**Heuristics** are economical (reduce thinking effort) and pretty effective usually, but they can also lead to systematic and predictable errors.

Insensitivity to sample size stems from the “**representativeness heuristic**” where people compare an event to another which is largely similar in characteristics, but neglect consideration of other factors (e.g. sample size).



# Fallacies

## Axiom:

- An error in reasoning.
- “Having observed 99 heads, the next coin flip must be a tail.” ---  
*Compulsive Anonymous Gambler*

## Commonly encountered as follows:

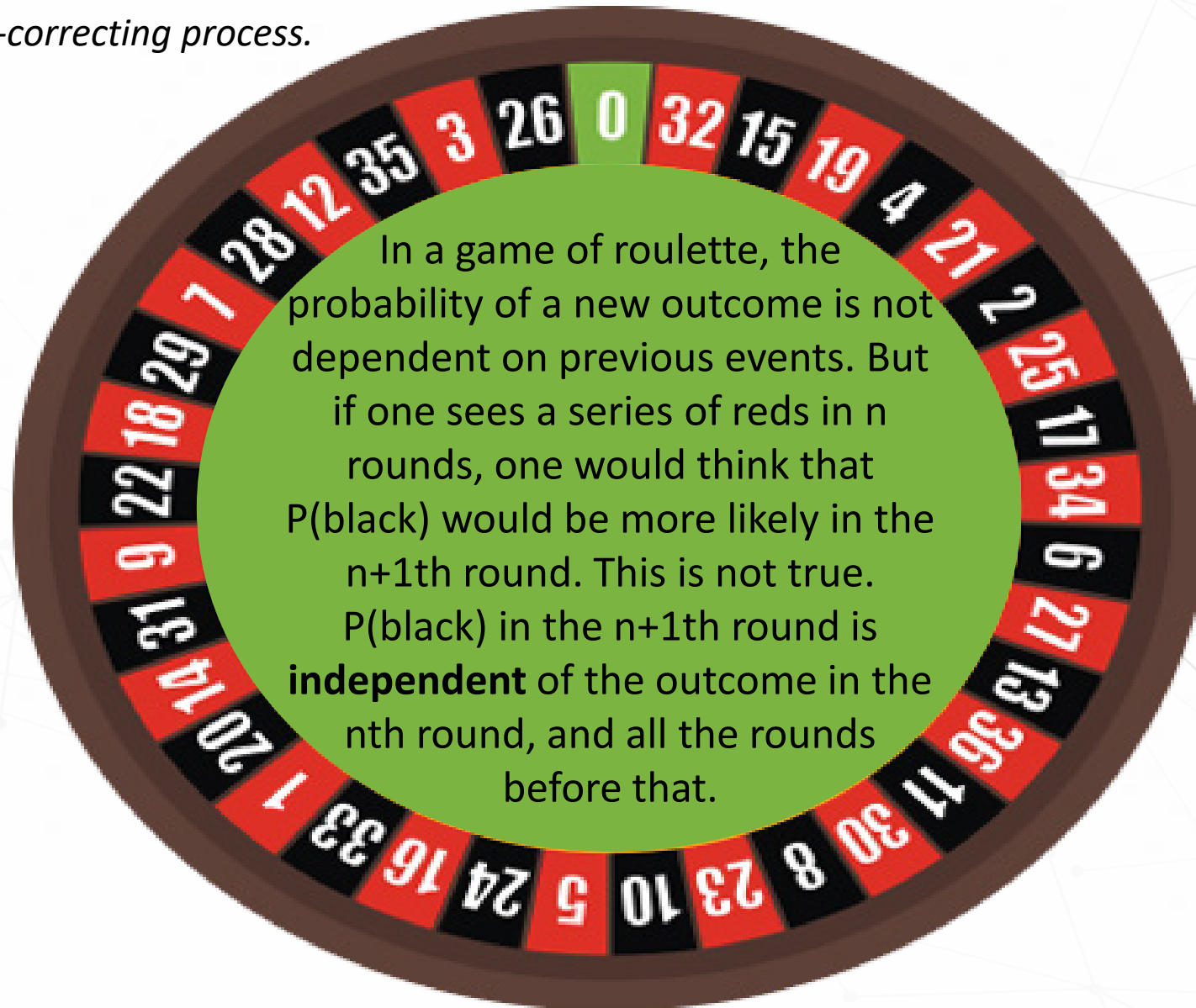
- Gene X is significantly up-regulated in Disease Y, you claim X causes Y.
- When predicting who will come out of the men’s bathroom next, you assume equal probabilities between men and women.

## How to avoid fallacies?

- Check your reasoning often.
- Write out your logic flow and look for gaps/flaws.
- Check with others and see if you can argue it through.

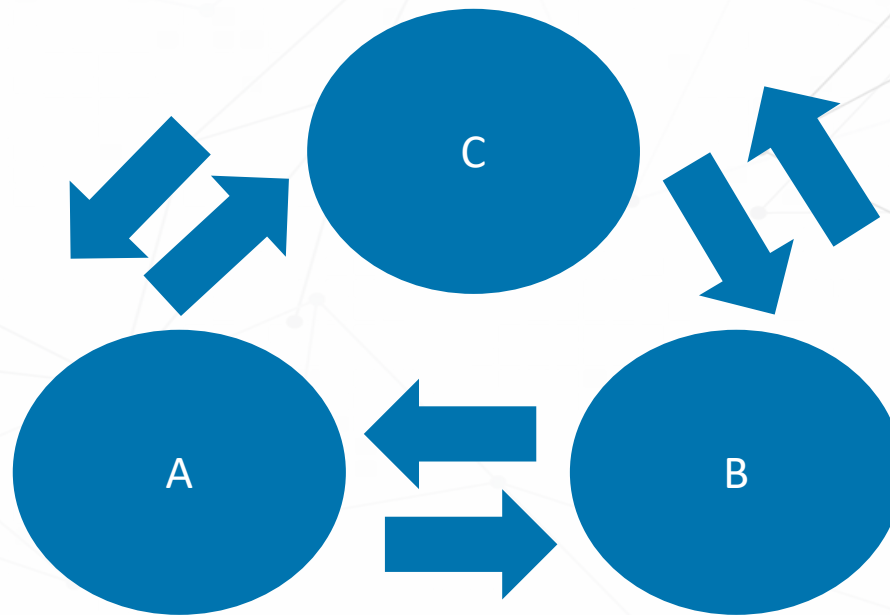
# Gambler's Fallacy

*Chance is not a self-correcting process.*



# Correlation-causation

When two variables A, B are correlated, there are at least 6 possibilities: A causes B, B causes A, A and B are controlled by C, A causes C which causes B, B causes C which causes A.



There are also other possibilities: A and B are simply correlated by chance alone.

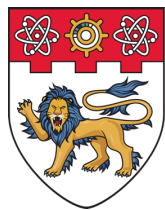
# Ludic

Use of inappropriate model to represent real life. Assuming flawless statistical models apply to situations where they actually don't. Consider the following conversation/example:

*Jason: Since about half the people in the world are female, the chances of the next person to walk out that door being female is about 50/50.*

*Sarah: Do you realize that is the door to Dr. Chao, the gynecologist?*





**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
SINGAPORE

# **Independent and Identically Distributed (IID)**

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences






# Independent and Identically Distributed (IID)

The condition of IID states that every sample has equal chance of being selected (**identically distributed**). The selection of one sample does not influence the chance of another being selected (**independent**). This is a common assumption used in many statistical models but...

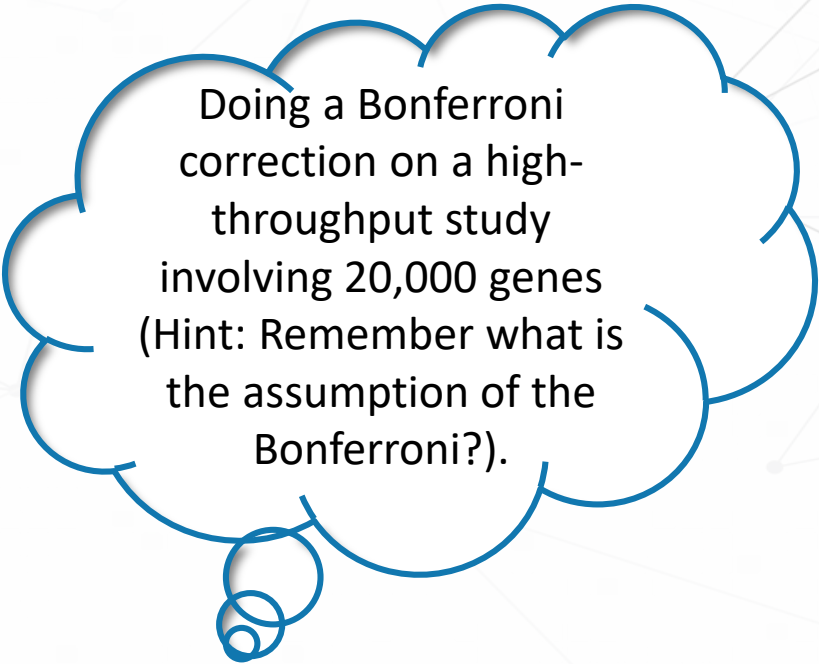


# Does IID reflect reality?

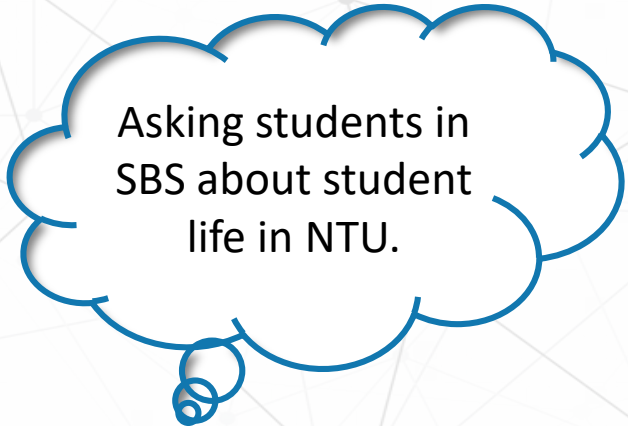
Consider the following scenarios.  
Which of the following violate IID and why?



Bringing your friends and family with you to a poll.



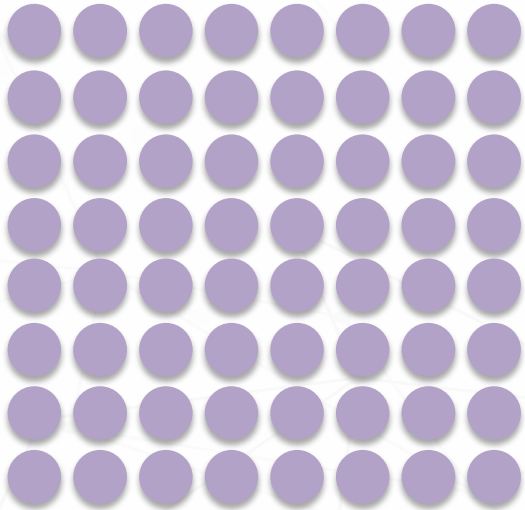
Doing a Bonferroni correction on a high-throughput study involving 20,000 genes (Hint: Remember what is the assumption of the Bonferroni?).



Asking students in SBS about student life in NTU.

# Does IID reflect reality?

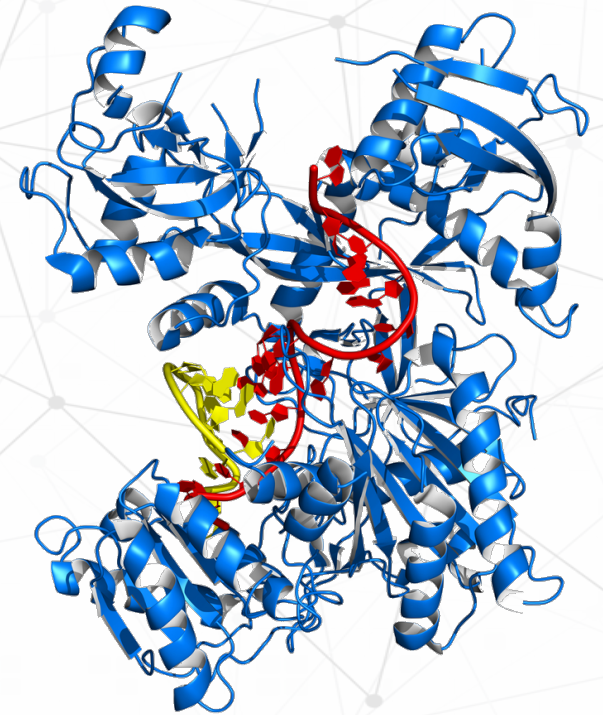
## Assumption



Statistical assumptions  
often do not reflect  
biological reality.

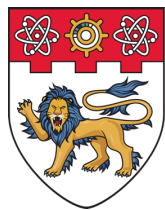
All genes behave independently. All genes have equal probability of being sampled/detected.

## Reality



Genes do not behave independently. High abundance genes are easier to detect.





**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

## **Inclusion Criteria**

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Inclusion Criteria

In clinical testing, we carefully choose the sample to ensure the test is valid.

- Independent: Patients are not related
- Identical: Similar # of male/female, young/old, in cases and controls (apples to apples)



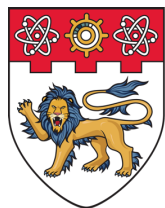


# Inclusion Criteria

In big data analysis, and in many datamining works, people sometimes do not set inclusion criteria.

This is not sound as it leads to the generation of hidden confounders.

However, setting very stringent inclusion criteria may limit our ability to generalise (limited scope).



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
SINGAPORE

# What are Confounders?

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# A confounder is...

- A confounder is a variable that can create spurious associations.
  - Let's say you want to study whether eating sweets causes lung cancer. You only have smokers in your "eating sweets" category, and non-smokers in your "non-sweet eating" category, whether or not someone smokes then becomes a confounding variable (**your study is likely to identify a spurious link between sweet eating and lung cancer**).
- A confounder is also referred to as a lurking variable in statistics.
- Let's examine this concept using the following scenario.



≠

# WHAT ARE CONFOUNDERS?

≠





Data is mostly gathered to investigate a certain situation, clear doubts or research a certain objective to reach a conclusion.



This superstition was set up in medieval times so that people were more visible at night and avoid getting into accidents.



Some questions have a straightforward answer which could be based on facts...

'Can monkeys be trained to use money and behave like humans to use it to buy things from the market?

We can probably give them tokens which could behave as money and check if they will barter that with people for food.'



...and is probably enough to base one's judgement based on a single set of data.

However, some questions need a more in-depth complex analysis due to their conflicting and comparative nature.



'If my dog could talk, I wonder what it would call me!'

Don't walk under a ladder, it will bring you years of bad luck.



However, it simply means that one should avoid walking under a ladder since you never know when it might fall on you and end up hurting you.







If you walk of your house while eating a banana ,it will cause somthing unfortunate to happen.

Such analysis which are captured with the help of statistics involves a hidden relationship between the variables.



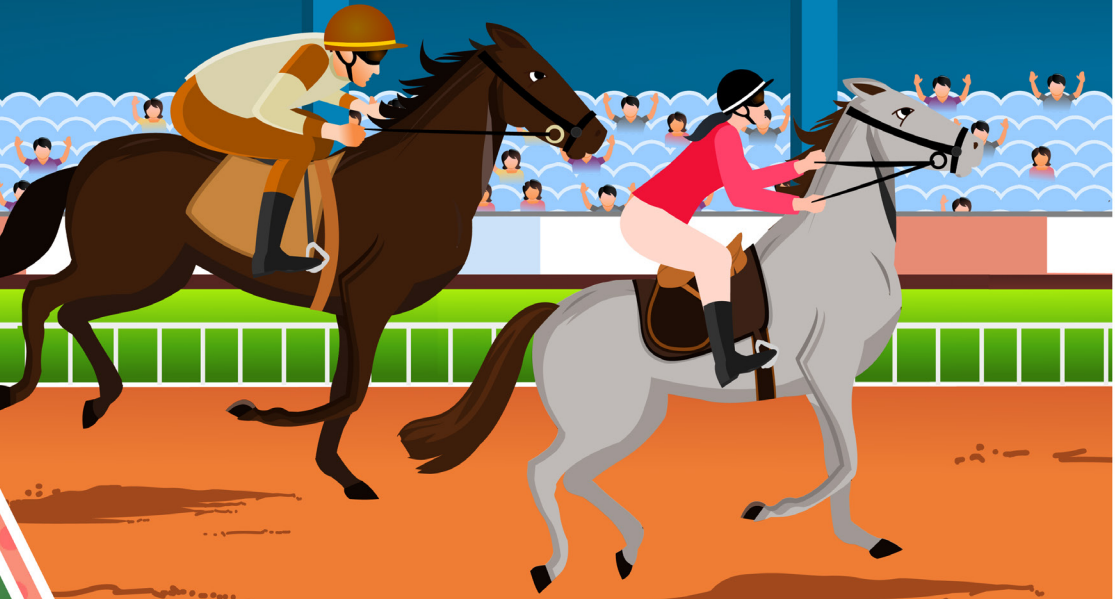
This myth was established so that people would be more careful about disposing off banana peels and avoid other people from slipping on them.



Singing on the dinner table while eating summons the devil.

However, when we use statistics we can never completely prove any of the conclusions drawn.

Let's explore the relationship between two different variables...



Would being a woman make me faster?



When female jockeys started riding, they were convinced that they were faster than male jockeys.



In order to test if gender has any effect on speed, an equestrian magazine gathered some data...



They randomly selected 35 male jockeys...



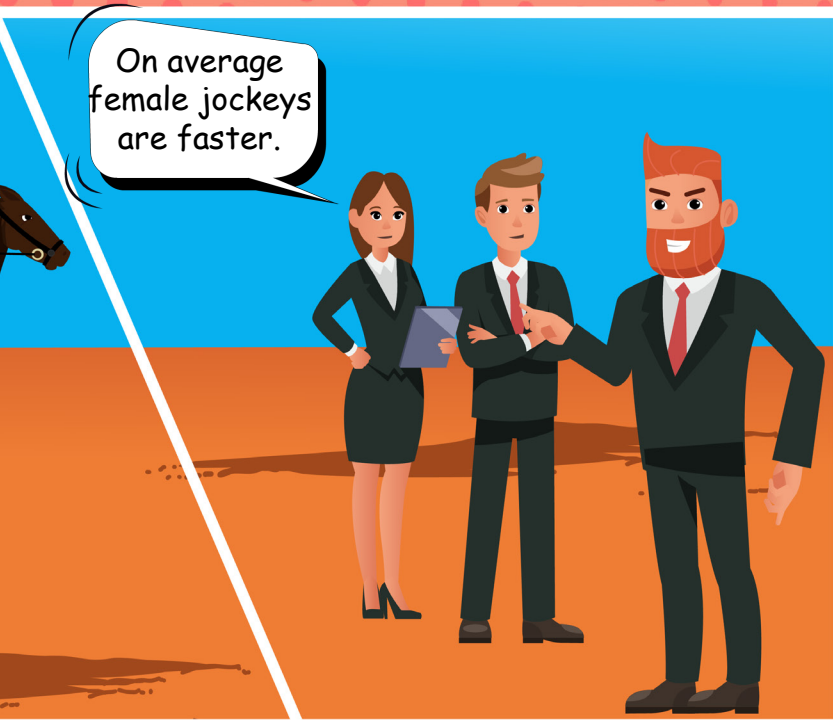
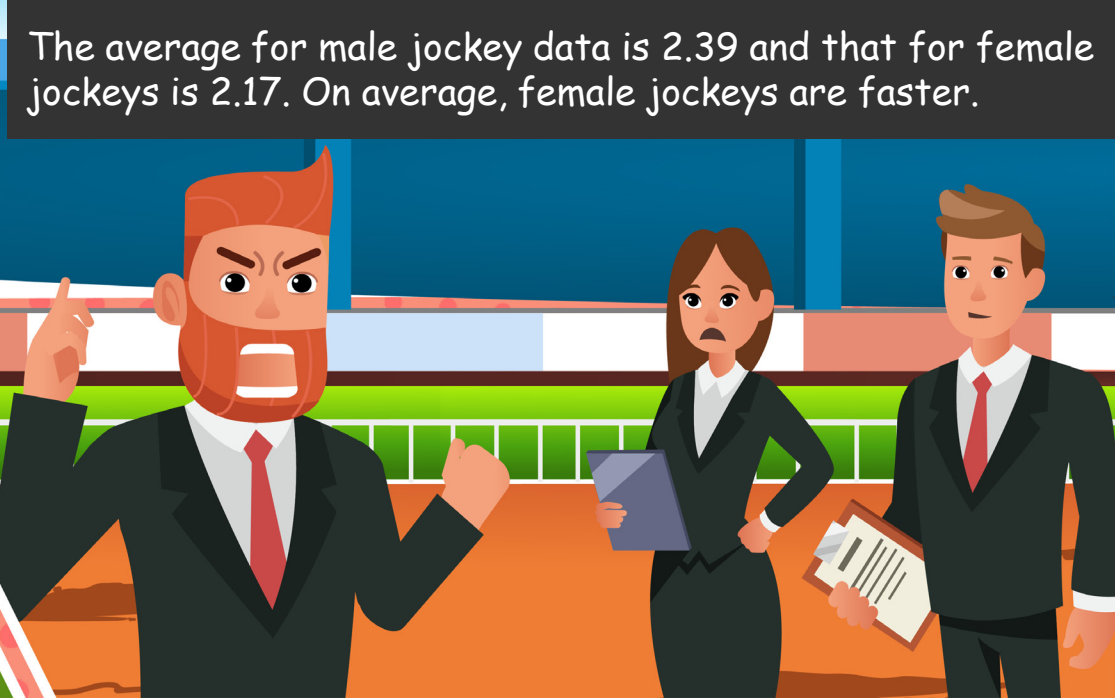
...and 35 female jockeys...



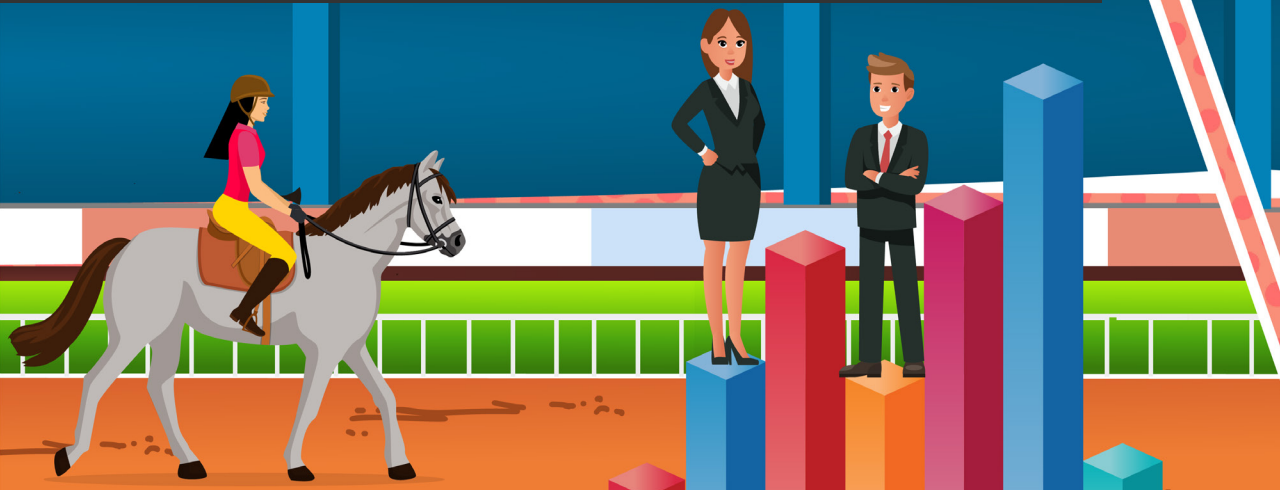
...and timed them for a 1 mile run.







But when we compare a Histogram of female jockeys' data....



Female jockeys data had one big hump on the fast side and one small hump on the slow side!

...with the Histogram for male jockeys' data....



Male jockeys data had one small hump on the fast side and one big hump on the slow side!

A man with a red beard and a black suit is pointing upwards with his right hand. To his left, a woman in a black business suit and a man in a black suit are looking at him. The woman is holding a blue folder. In the background, a grey horse and a brown horse are standing on a racetrack. The scene is set against a blue sky and a green fence.

Why should both groups be skewed in different directions?

I think we are missing something. Why does the data look like this?

This means that the **relationship** between the two variables may **not be as simple** as we thought.



There are different breeds of horses...

I am a nimble,  
sly, speedy  
runt...

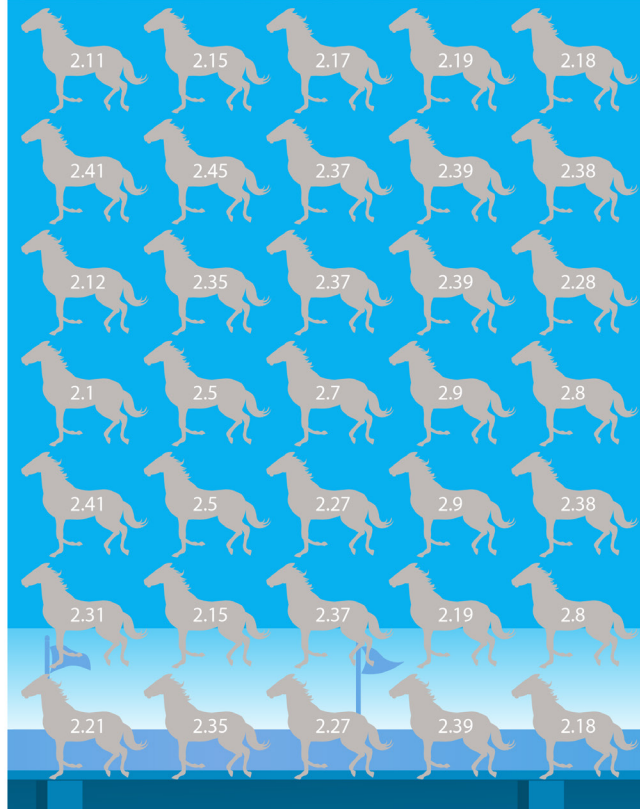
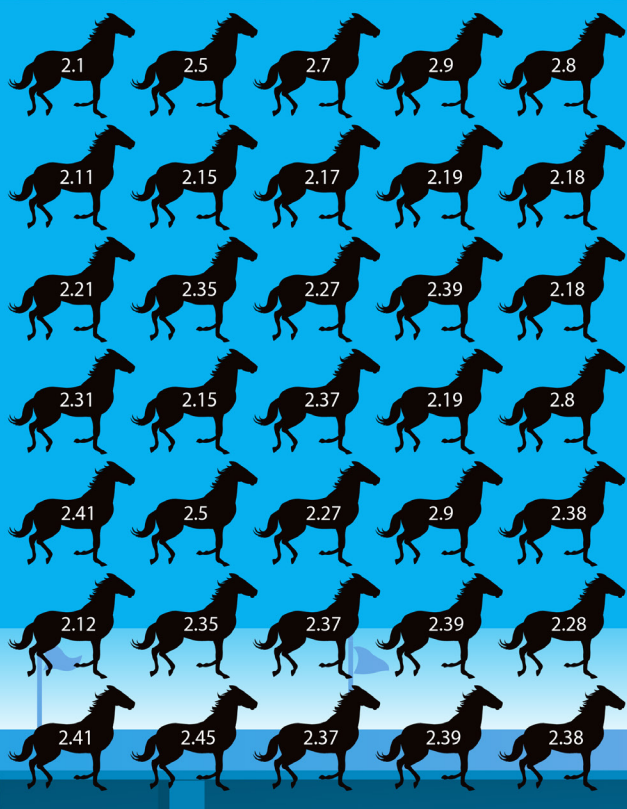
I am a big, mean,  
burly grunt..

We don't ride no  
wimpy horses!

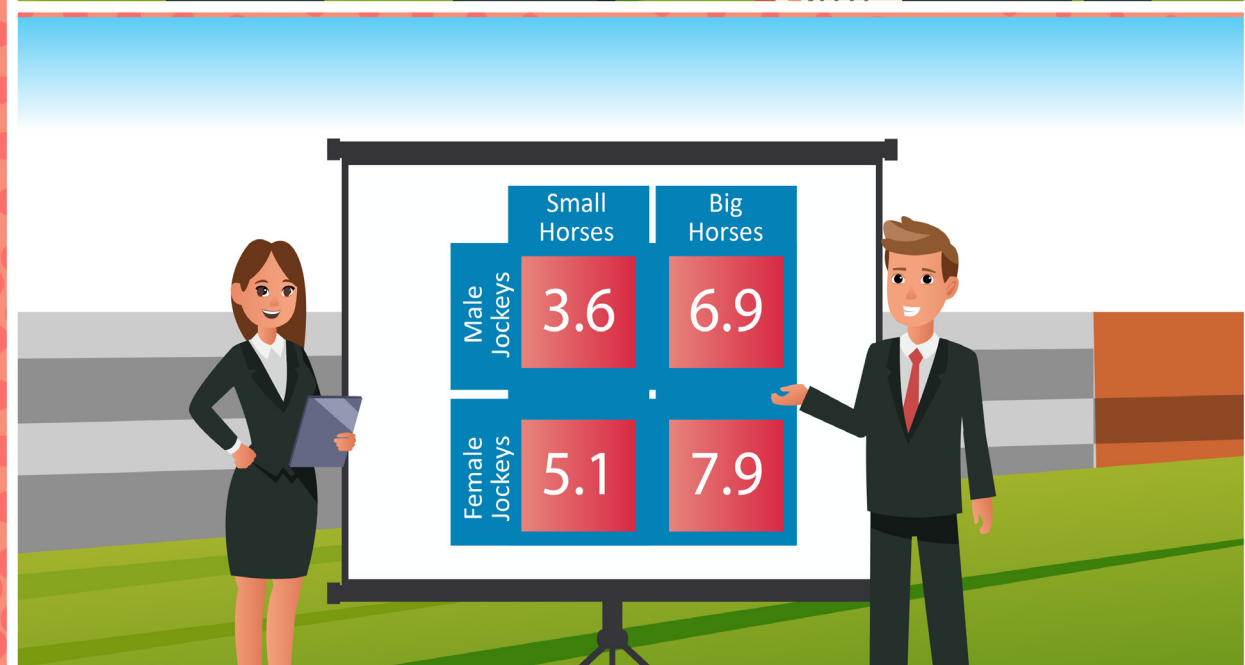
Duh, they are  
faster!

...and male jockeys tend to prefer the larger, slower horses...

...while female jockeys prefer smaller, faster horses!




So it's no wonder the female jockeys seemed faster overall....



When we account for the choice of horse by calculating average times by both jockey gender and horse type, the results are surprising...







While we were busy looking for relationship between two variables, a hidden, lurking variable was sneaking around.

Wreaking havoc with our conclusions.

Unfortunately, lurking variables can damage all kinds of statistical analyses.

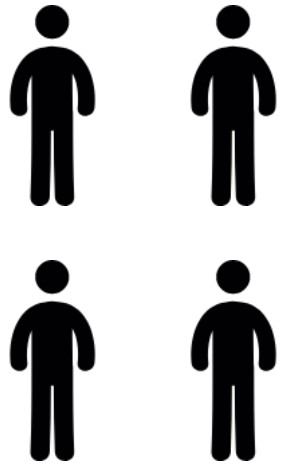
It's a part of a statistician's job to dig around for lurking variables.

# Why should we care about confounders in bio data science?

- Clinical/Biology samples are complex, and are different in many unexpected ways.
- Imbalance in any of these ways (variables) may lead towards unexpected correlations (Anna Karenina Effect).
- They may also lead towards non-detection of true signal (Loss of power).
- Poor experimental design can lead towards horrifying situations where the variable of interest is completely entangled with a confounder (and cannot be disambiguated) --- this scenario is called perfect confounding.

# The Idea of Perfect Confounding

Let's say we want to know if drinking Yakult makes you taller. We design the following experiment:



Yakult Drinkers



Non-Yakult Drinkers

What is wrong here?

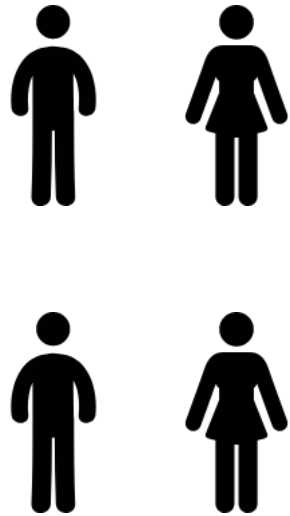


# The Idea of Perfect Confounding

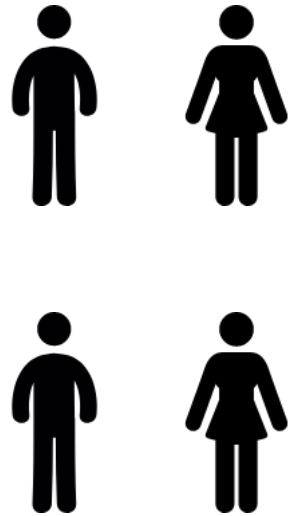
- Yakult drinkers being tested are all male.
- Males are in general, taller than females.
- Therefore, we are likely to observe that Yakult drinkers are taller and therefore conclude that a correlation exists.
  - However, we cannot say for sure (**do not forget that Yakult is full of milk calcium and nutrition and can in fact, aid growth**).
  - In this case, because the Yakult and gender variables are completely mixed together in the same category, we cannot disentangle them. This situation is known as “perfect confounding”.
- Note that not all unbalanced variables are confounders. The unbalanced variable must contribute towards the outcome of interest. For example, if the Yakult drinking group is composed entirely of biology students on one hand, and engineering students on the other, you may not care as much.

# Remedies

## Balanced Design

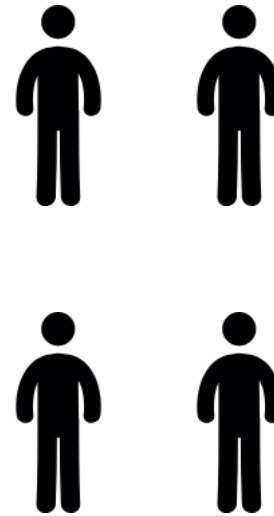


Yakult Drinkers

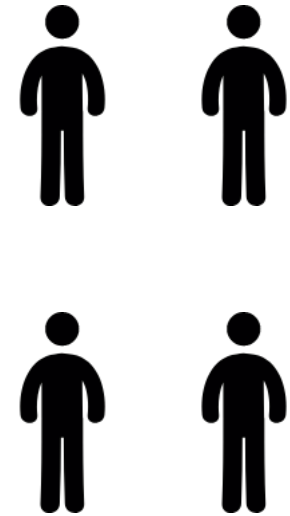


Non-Yakult Drinkers

## Remove the Confounder



Yakult Drinkers



Non-Yakult Drinkers

In your opinion, which approach is better? Which approach is more feasible in real world practice?

# Hunting for Confounders

## Use the Contingency Table

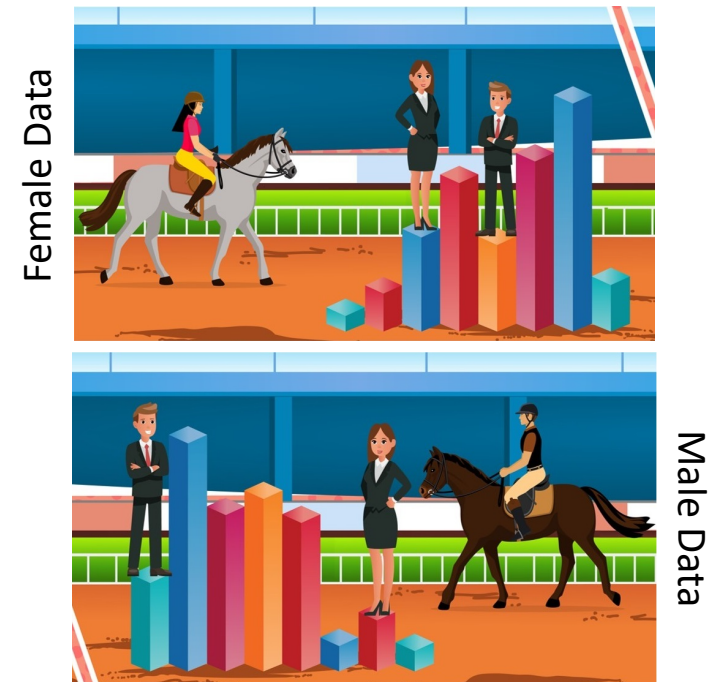
Suspected confounder

Variable of Interest	Suspected confounder		Averages
		Small Horses	Big Horses
	Male Jockeys	3.6	6.9
	Female Jockeys	5.1	7.9

Check that your results are consistent given any split of a 'suspected' confounder.

## Graphs and Plots

Check for multi-modality of your data



If you see multiple peaks, then there may be subpopulations (splittable by the confounder) in your data.



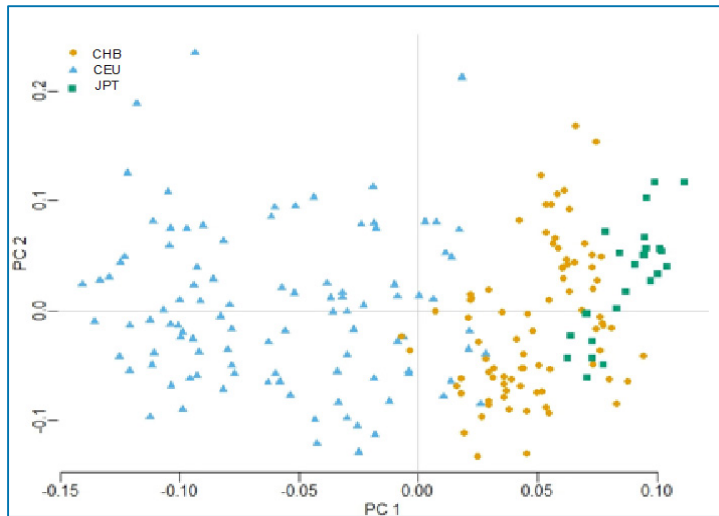
# Thinking Time

What do you make of the 1,000 differential genes identified in this study---Are they reliable?

## Series GSE5859

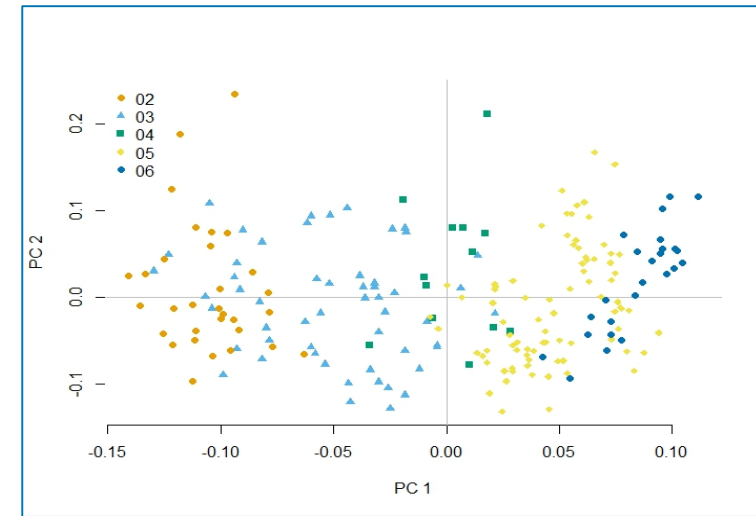
[Query DataSets for GSE5859](#)

Status Public on Jan 07, 2007  
Title Allelic Differences Account for Gene Expression Differences Among Population.  
Organism [Homo sapiens](#)  
Experiment type Expression profiling by array  
Summary Expression level of genes in lymphoblasts from individuals in three HapMap populations (CEU, CHB, JPT) were compared. More than 1,000 genes were found to be significantly different ( $P < 0.05$ ) in mean expression level between the CEU and CHB+JPT samples.  
Keywords: Comparison of Gene Expression Profiles from Lymphoblastoid cells



2D PCA scatterplot (grouped by ethnicity)

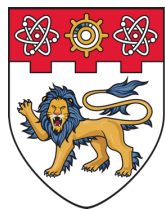
(Note: PCA is a summary of total variation in data and is useful for summarizing the visualization of samples with thousands of variables into 2D plots).



2D PCA scatterplot (grouped by data collection year  
--- 02 refers to years 2002, 03 to 2003 and so on)

# Thinking Time

- Common genetic variants account for differences in gene expression among ethnic groups.
- Take some time to think through the problem and data.
- Check out the original paper here:
  - <https://www.ncbi.nlm.nih.gov/pubmed/17206142>
- Check out the rebuttal and criticism here:
  - <https://www.ncbi.nlm.nih.gov/pubmed/17597765>
- Is this a case of perfect confounding? What are you able to conclude/not conclude based on the study? Have the authors dug themselves into an unrescuable position?



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

# **Simpson's Paradox**

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences





# Simpson's Paradox

Watch:  
[https://ed.ted.com/lessons/  
how-statistics-can-be-  
misleading-mark-liddell](https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell)

The presence of lurking variables leads to a reversal of findings once the data has been split by the lurking variable (e.g. male and female).

Best Practice:  
Beware anytime data is aggregated. Try to keep dataset balanced across any split by sub-variables (very hard to do).  
Check that the findings are consistent despite splitting by each potential variable.

# Simpson's Paradox

Looks like A is better		
Overall		
	A	B
Lived	60	65
Died	100	165

Looks like B is better		
Women		
	A	B
Lived	40	15
Died	20	5

Men		
	A	B
Lived	20	50
Died	80	160

Looks like A is better		
History of heart disease		
	A	B
Lived	10	55
Died	70	50

No history of heart disease		
	A	B
Lived	10	45
Died	10	110

# Simpson's Paradox

Looks like A is better		
Overall		
	A	B
Lived	60	65
Died	100	165

Taking A:

- Men = 100 (63%)
- Women = 60 (37%)

Looks like B is better		
Women		
	A	B
Lived	40	15
Died	20	5

Men		
	A	B
Lived	20	50
Died	80	160

Taking B:

- Men = 210 (91%)
- Women = 20 (9%)

Looks like A is better		
History of heart disease		
	A	B
Lived	10	55
Died	70	50

No history of heart disease		
	A	B
Lived	10	45
Died	10	110

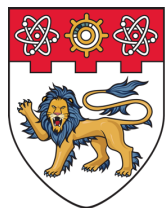
Men taking A:

- History = 80 (80%)
- No history = 20 (20%)

Men taking B:

- History = 55 (26%)
- No history = 155 (74%)





**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
SINGAPORE

# Batch Effects

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Batch Effects



**Batch effects are sub-groups of measurements that have exhibit different behavior across conditions and are unrelated to the biological or scientific variables in a study.**



If not properly dealt with, these effects can have a particularly strong and pervasive impact. This can lead to selection of wrong variables from data.

# Some Simple Examples

Oven A tends to overheat. Oven B has uneven heating issues. You bake 5 cookies in each oven set to the same temperature. They turn out differently.



Two people split 10 samples equally between them on a western blot. Person A tends to press down harder on average. Person B tends to press lighter. Blots by person A turn out darker generally.



Pipetting



# A More Complex Example

## Transcriptomics

You have 2 phenotypes, A and B, with 2 samples each. You split these into 2 runs, 1 and 2 and analyse their gene profiles (A1 B1 and A2 B2). You find that samples tends to cluster by run rather than phenotype.



## Question

If you run the samples as A1 A1 and B2 B2, what will happen?



# Two Ways of Dealing with Batch Effects

## Batch Correction Algorithms

### Advantages

- Maintains the “scale” of the data while removing batch-correlated variation.

### Disadvantages

- Difficult to use.
- Many different types (need to know how the algorithm works).
- Can affect data integrity (create false positives).

## Re-normalise the Data

### Advantages

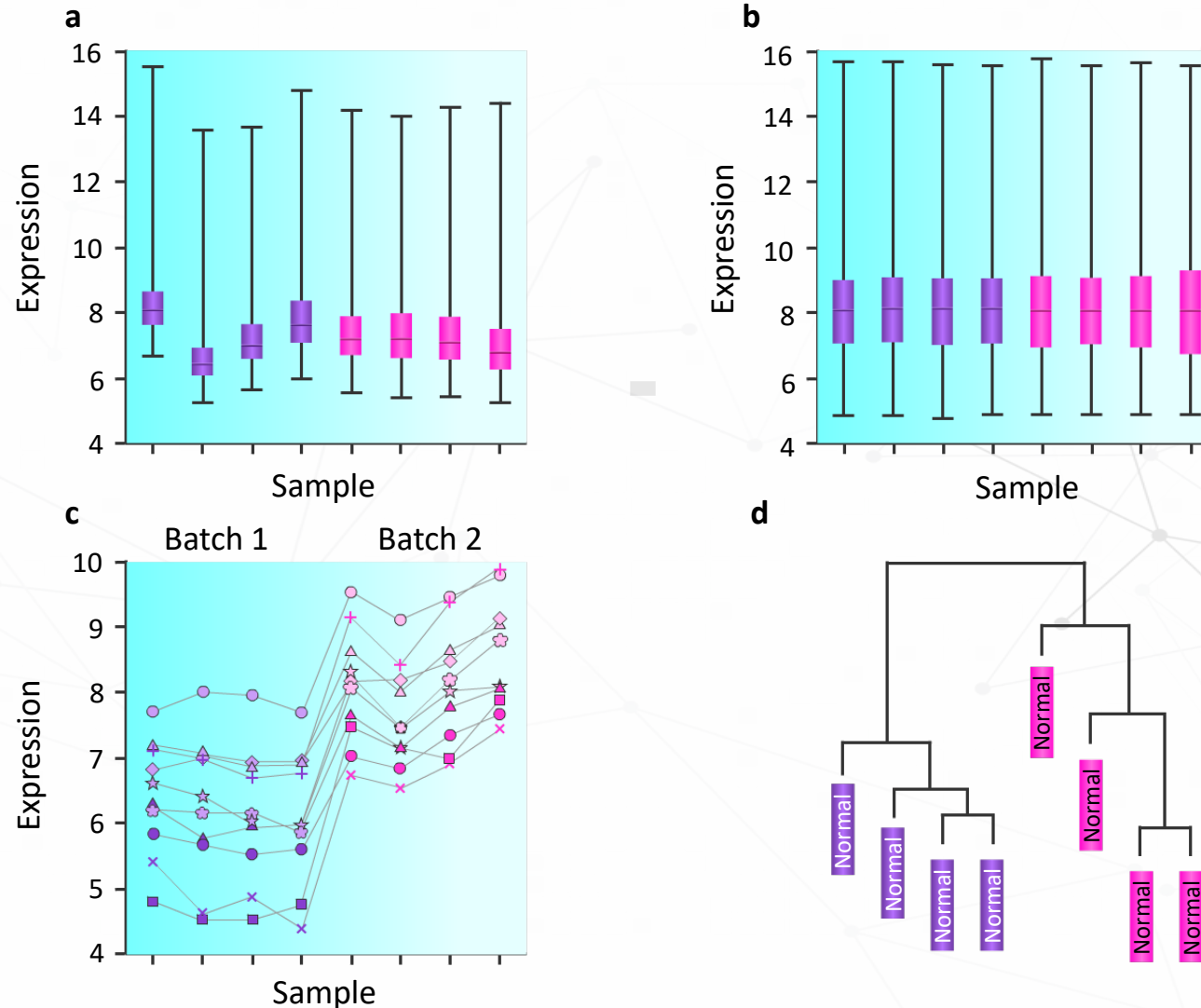
- Simple to use and understand.
- Does not adversely affect data integrity.
- Does not require prior knowledge of batch factors.

### Disadvantages

- Changes the “scale” of the data e.g. in z-norm, you lose information on actual data magnitude.
- Limited efficacy.

# Two Ways of Dealing with Batch Effects

Simple normalisation does not guarantee batch effect removal.



Source: Leek et al, Nature Reviews Genetics, 2010



# Two Ways of Dealing with Batch Effects

## Exploratory Analyses

Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time).

Plot individual features versus biological variables and batch surrogates.

Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates.

## Downstream Analyses

Do you believe that measured batch surrogates (processing time, Laboratory, etc.) represent the only potential artefacts in the data?

Yes

Use measured technical variables as surrogates for batch and other technical artefacts.

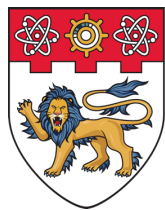
No

Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA).

Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/ surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat.

## Diagnostic Analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects.



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
SINGAPORE

# Domain-specific Laws

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Why do you think the data on the right looks suspicious?

**rs123** chi-square p-value = 4.78E-21

Genotypes	Controls[n(%)]	Disease[n(%)]
AA	1(0.9%)	0(0%)
AG	38(35.2%)	79(97.5%)
GG	69(63.9%)	2(2.5%)



# Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Why do you think the data on the right looks suspicious?

**rs123** chi-square p-value = 4.78E-21

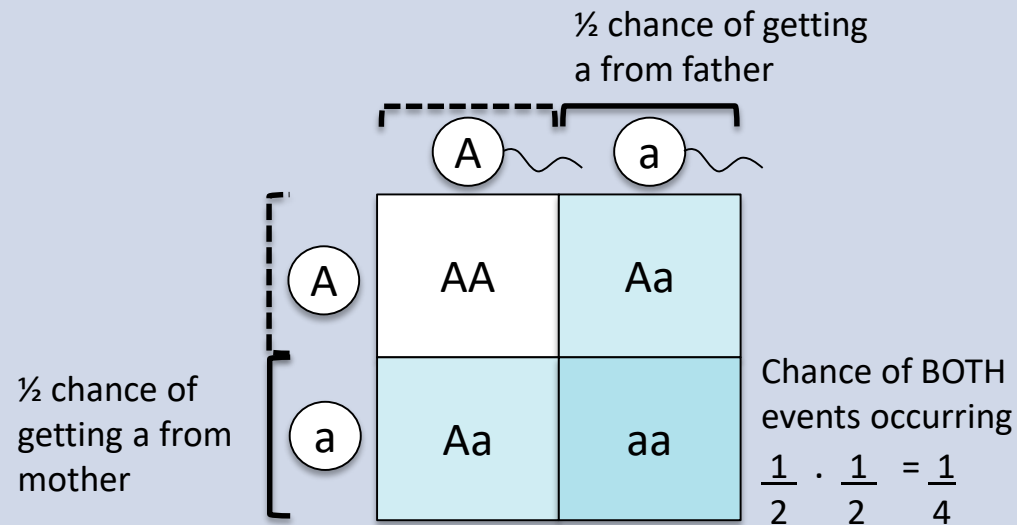
Genotypes	Controls[n(%)]	Disease[n(%)]	N= 189
AA	1(0.9%)	0(0%)	1/189 (<1%)
AG	38(35.2%)	79(97.5%)	117/189 (62%)
GG	69(63.9%)	2(2.5%)	71/189 (37.9%)

# Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Let's use what we know about simple human genetics.

Let's calculate backwards.



- 62% of our samples are AG.
- So let's say, the probability of a mother and a father both being AG is  $0.62 * 0.62 = 0.38$ .
- And the probability of them having a child that is AA is  $0.25 * 0.62 * 0.62 = 0.09$  (9%).

# Domain-specific Laws

Laws of genetics gives us an expectation on genotype distribution frequencies.

Let's look at our table again.

We expect 9%. But our data says AA is only < 1%. So unless AA is lethal, our samples do not reflect expectation.

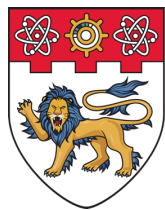
Therefore, via the use of domain-specific laws (in this, mendelian segregation proportion) we infer that our samples could be biased.

**rs123** chi-square p-value = 4.78E-21

	Genotypes	Controls[n(%)]	Disease[n(%)]	
<1% AA	AA	1(0.9%)	0(0%)	1/189
62% AG	AG	38(35.2%)	79(97.5%)	117/189
38% GG	GG	69(63.9%)	2(2.5%)	71/189

N= 189





**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
SINGAPORE

# Non-association

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



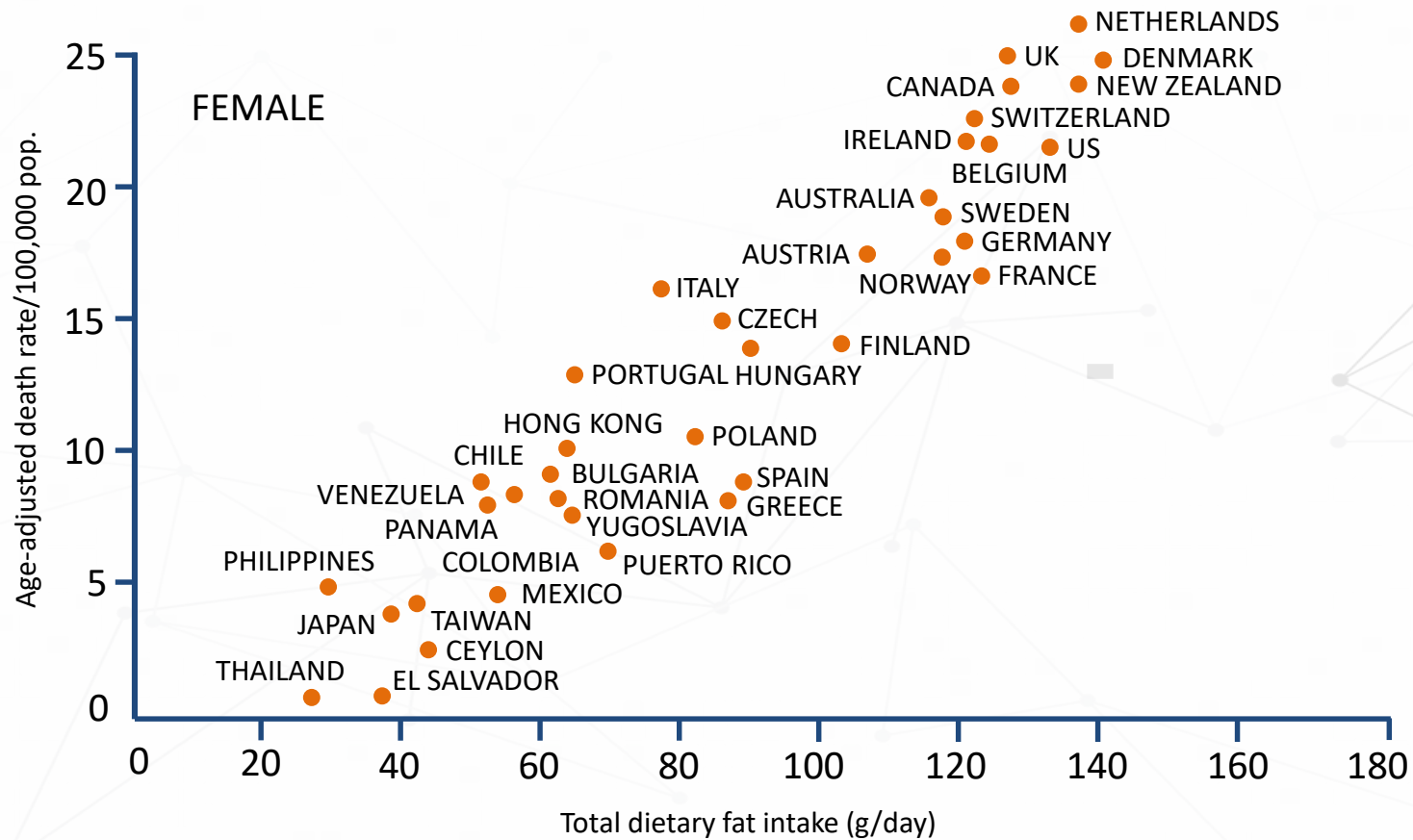
# Positive vs. Negative Space



- What is positive to you?
- What is negative to you?
- In the image here, which one do you think is more important?

- We have many methods to look for associations and correlations (positive space), for example statistical test.
- We tend to ignore non-associations (negative space).
  - We think they are not interesting/ informative.
  - There are too many of them.
- We also tend to ignore relationship between associations (aka multi-collinearity).

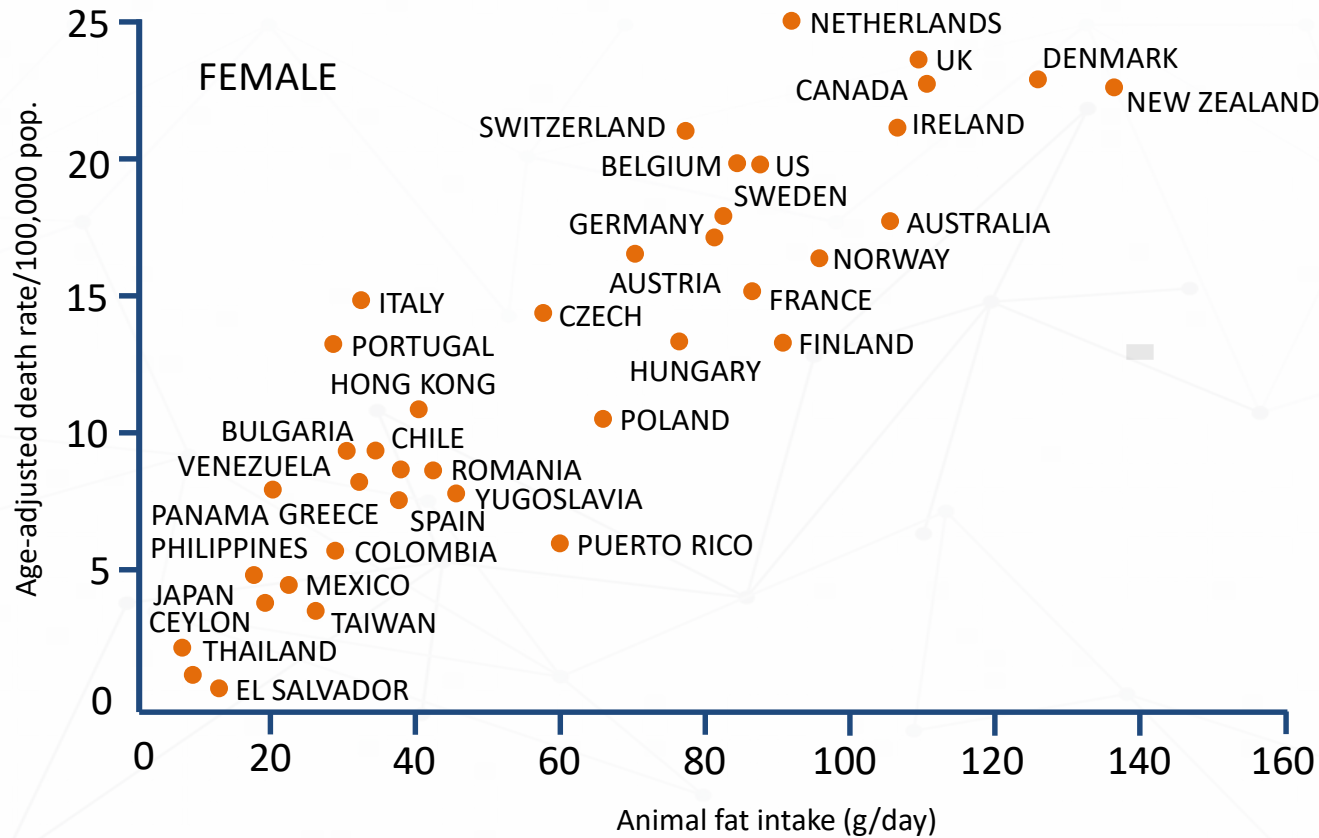
# We love to find correlations like this...



Dietary fat intake correlates with breast cancer.

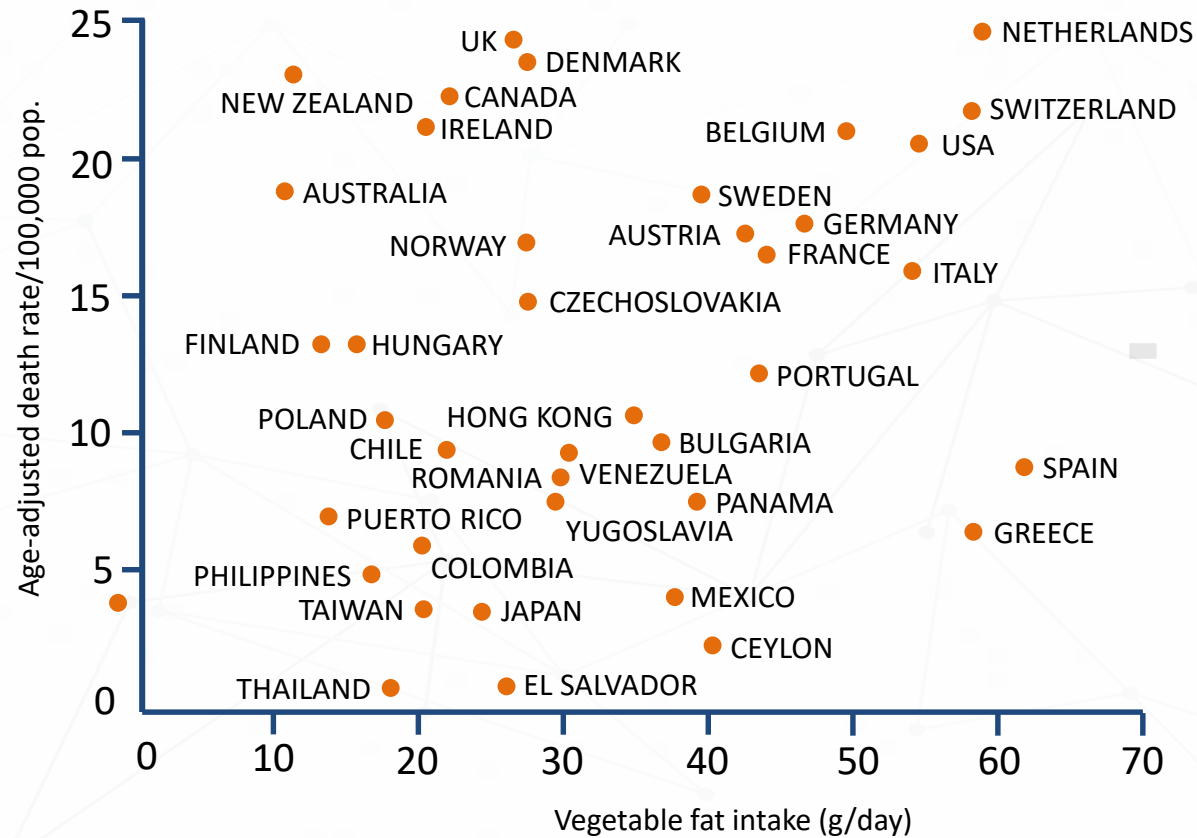


# And like this...(positive)



Animal fat intake correlates with breast cancer.

# But not this...(negative)



Plant fat intake doesn't correlate with breast cancer.

# But there is much to be gained...



A: Dietary fat intake correlates with breast cancer.

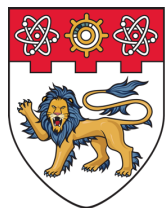
B: Animal fat intake correlates with breast cancer.



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

- Given C, we can eliminate A from consideration, and focus on B!
- **You may also conclude that not all fats are bad, and that you may quite liberally eat plant fat.**





**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

# Context

BS0004 Introduction to Data Science

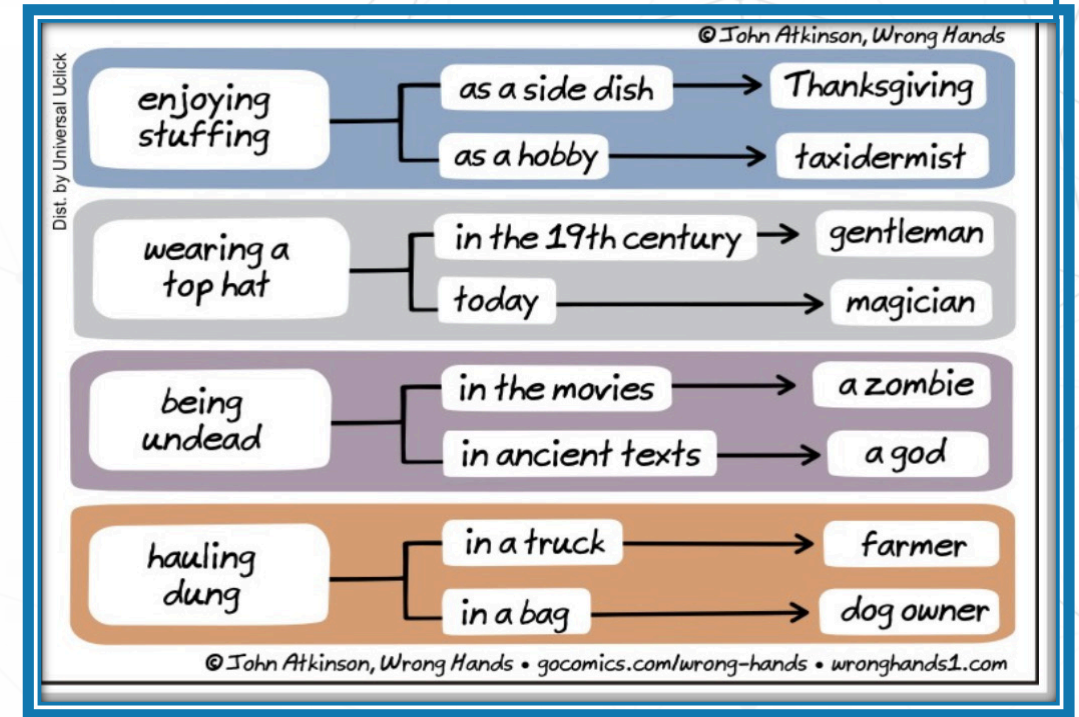
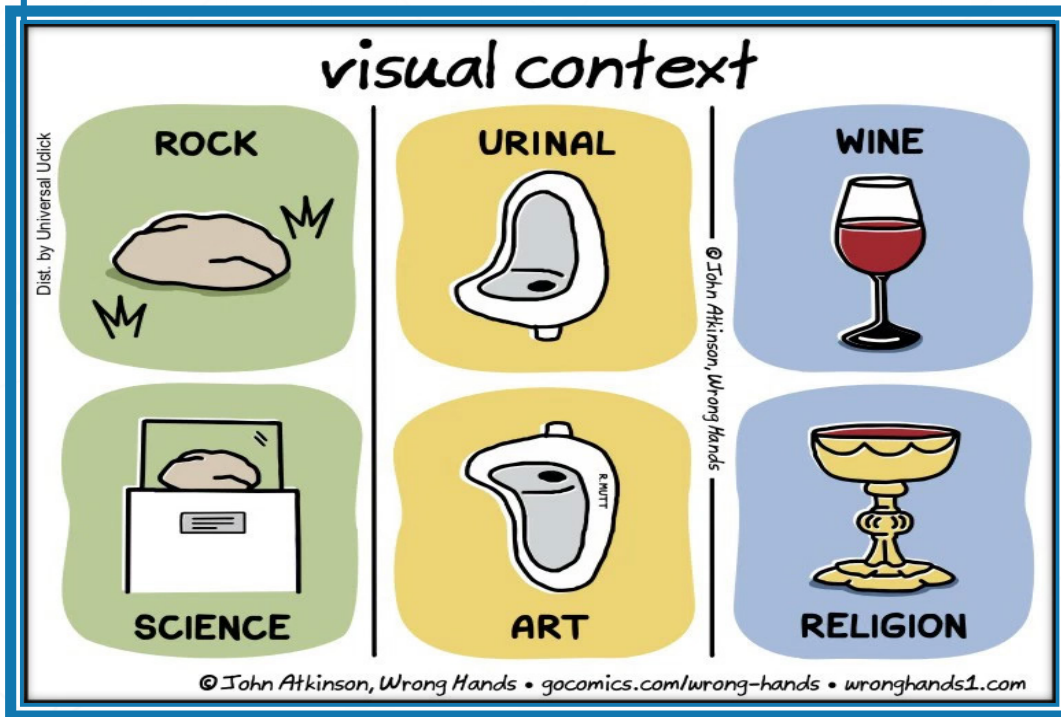
Dr Wilson Goh

School of Biological Sciences



# Context

The term 'context' is a noun. It is the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood.



Source: Creative Common License

<https://wronghands1.files.wordpress.com/2017/07/visual-context.jpg>

Source: Creative Common License

<https://wronghands1.files.wordpress.com/2017/02/contextual.jpg>



# Why is context important in biology?

## Gene isoform switching:

- Same gene, but produces different isoforms (splice variants) in different tissues, i.e. a gene functions differently in different parts of the body.
- Refer to lecture notes for link to website for further information.

## Human behavior:

- In our typical environment, we are generally well-behaved, well-adjusted individuals.
- In an alternative environment with new rules (e.g. Stanford Prison Experiment), people can behave in extreme ways.



## Context

## Gene networks:

- Genes do not function independently of each other but rather in pathways and networks.
- When several components of a single pathway are affected, we can generally deduce that this pathway (including the unobserved components) as a whole is important to the phenotype.

## Evolution:

- Interplay between genetics and environment (via natural selection).
- In Galapagos, finches varied from island to island (their beaks adapted to the type of food they ate; filling different niches on the Galapagos Islands).
- Refer to lecture notes for link to website for further information.



# Context (Biological Complexes)


**Postulate:** The chance of a protein complex being present in a sample is proportional to the fraction of its constituent proteins being correctly reported in the sample. Suppose proteomics screen has 75% reliability; a complex comprises proteins A, B, C, D, E; and screen reports A, B, C, D only but not E.



Complex has 60% ( $= 0.75 * 4 / 5$ ) chance to be present.



The unreported protein E also has  $\geq 60\%$  chance to be present, as presence of the complex implies presence of all its constituents (**improving coverage and recover missing proteins**).



Each of the reported proteins (A, B, C, and D) individually has 90% ( $= 100\% * 0.6 + 75\% * 0.4$ ) chance of being true positive, whereas a reported protein that is isolated has a lower 75% chance of being true positive (**removing noise**).



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

# Meta and Mega Analyses

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Big and Small Data

Data science isn't necessarily concerned only with big data. Small data is also important. But what's the difference?

	Big Data	Small Data
<b>Data Condition</b>	Usually unstructured, not ready for analysis	Usually structured, ready for analysis
<b>Location</b>	Cloud, Offshore, SQLServer, etc.	Database, Local PC
<b>Size</b>	Over 50k variables, over 50k individuals, random samples, unstructured.	File that is in a spreadsheet, that can be viewed on a few sheets of paper.
<b>Purpose</b>	No intended purpose.	Intended purpose for data collection.



# Meta-analysis

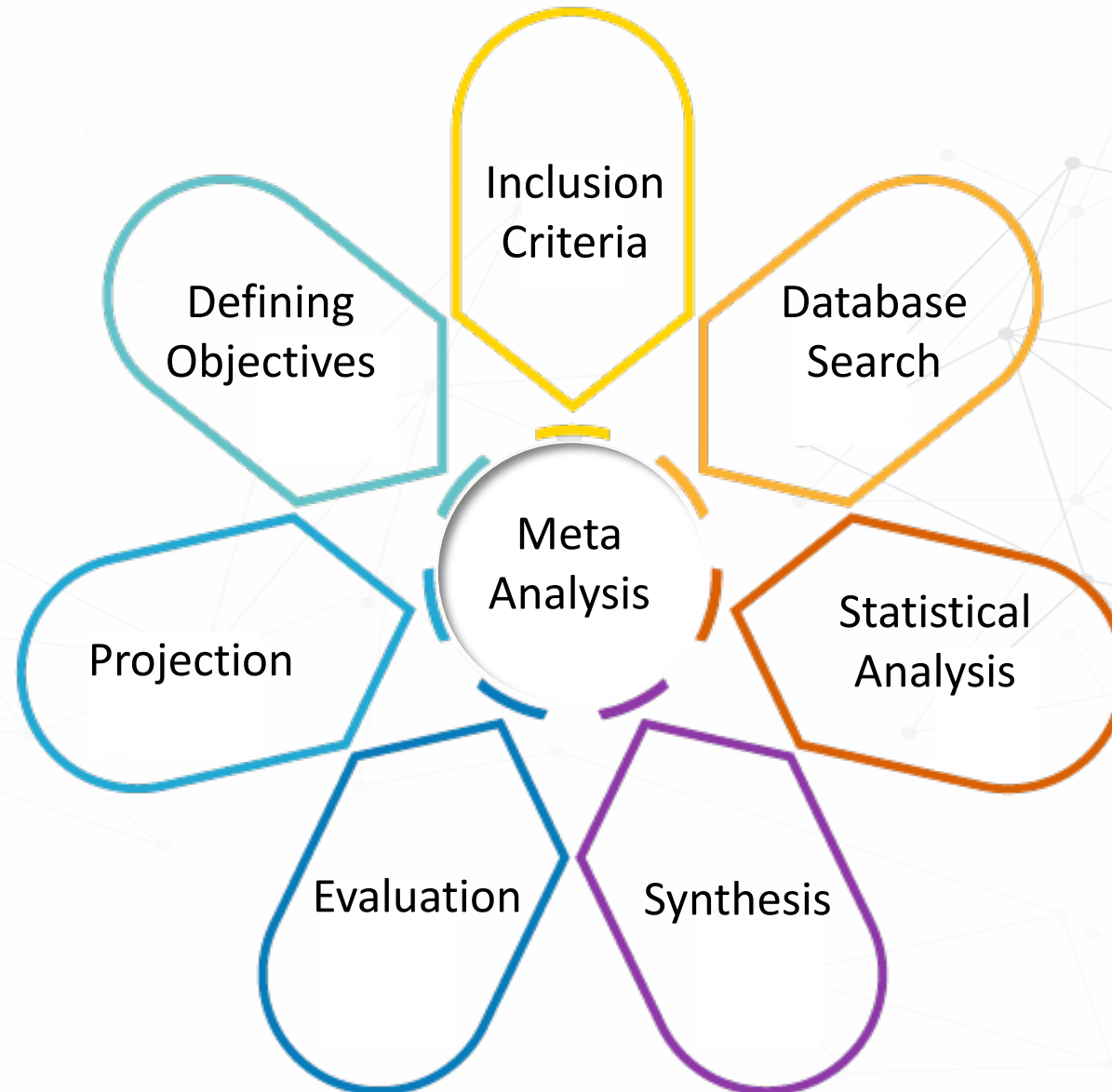
Meta-analysis is a statistical procedure that integrates the results of several independent studies.

It can be a very useful method to summarise data across many studies, but requires careful thought, planning and implementation.

A meta-analysis goes beyond a literature review.

**Is this equivalent to big data?**

# Considerations for Meta Analysis



# Between Small and Large Data

D1

→ Standard/Small Data Analysis

In series:

D1 D2 D3

→ Big Data (Mega) Analysis

In parallel:

D1

→ Small Data Analysis

D2

→ Small Data Analysis

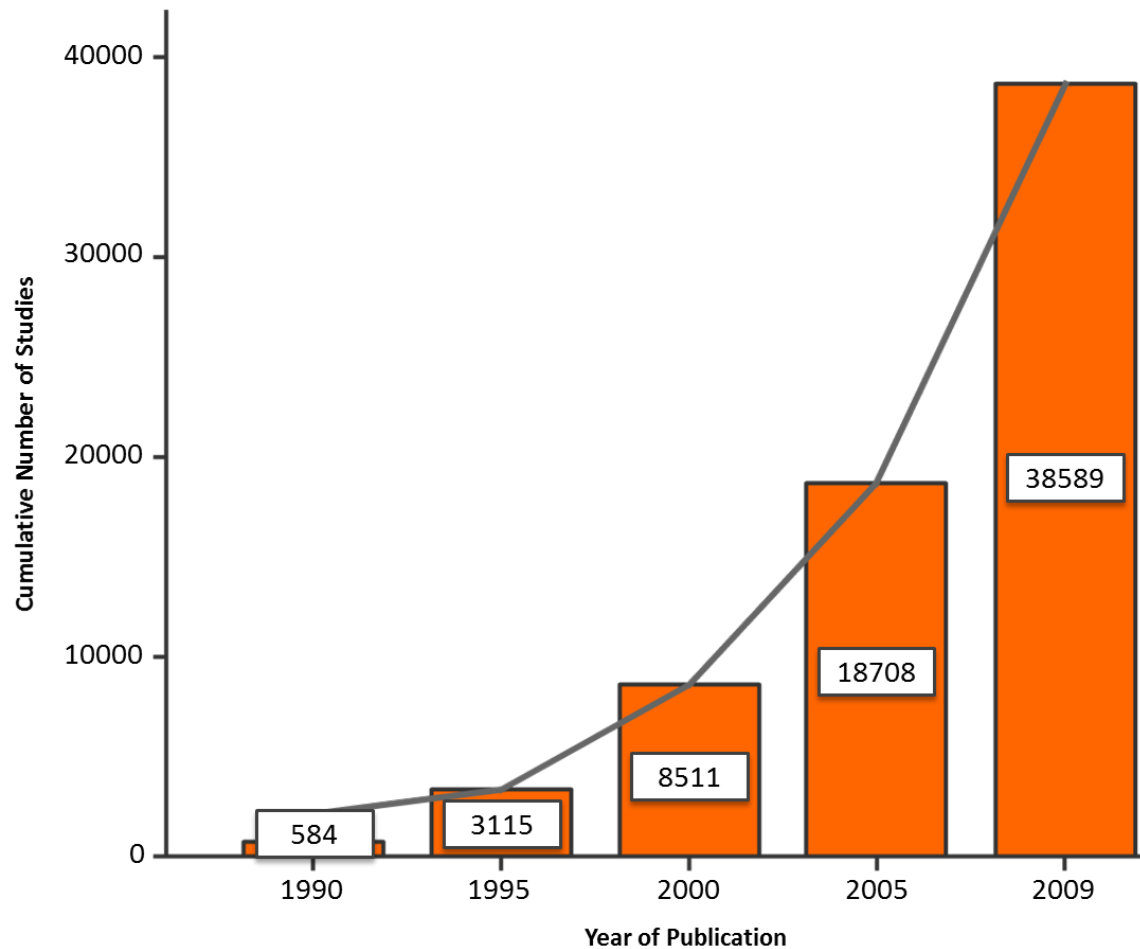
D3

→ Small Data Analysis

Integration  
“Meta-analysis”



# Meta-analysis is Increasingly Common



Cumulative number of publications about meta-analysis over time, until 17 December 2009 (results from Medline search using text "meta-analysis").

This upward trend is also partly because of the larger amount of existing data available to us. And not simply because meta is necessarily seen as more important.

# Papers Discussing Meta-analysis

Papers for discussion (feel free to add more):

- Berman and Parker, Meta-analysis: Neither quick nor easy, BMC Medical Research Meth, 2002.
- Haidich, Meta-analysis in medical research, Hippokratia, 2010.
- Nakagawa et al, Meta-evaluation of meta-analysis: ten appraisal questions for biologists, BMC Biology, 2017.

Questions for thought:

- Do the various papers agree with each other?
- What are some simple examples of finding consensus amongst the individual datasets?
- “Meta-analysis” is less powerful. Do you agree?

# Example of Mega-analysis (aka big data analysis or data pooling)

## Papers for discussion (feel free to add more):

- Hess et al. Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia, Schizophr Res, 2016.
- This paper puts together 9-11 datasets to generate pooled data for deriving markers for schizophrenia.

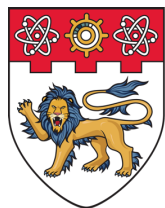
## Questions for thought:

- Do you foresee any problems? Comment on their methodology and critique their findings.
- You may also relate what Hess et al did and whether they should also have performed a meta-analysis as well. What should they expect to see?
- How would you have designed the analysis?



# Relating Meta-analysis and Big Data Analysis

	Meta-analysis	Big Data
<b>Addresses</b>	Heterogeneity	Power
<b>What it is</b>	Systematic review with synthesis of findings	Integration-based knowledge discovery
<b>How to do it?</b>	No set protocol	No set protocol
<b>Relies on</b>	Consistency	Strength of larger sample size (pooling)
<b>Uses</b>	Many datasets (in parallel)	Many datasets (in series)
<b>Achilles heel</b>	Data selection bias; not being “expansive” enough; many conflicting results; false negatives	Not addressing dataset; heterogeneity issues; false positives



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  
**SINGAPORE**

# Summary


BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences



# Key Takeaways from this Topic

- 
1. Bias and Fallacies, IID, Inclusion Criteria, Simpson's Paradox, Batch Effects and Domain-specific laws are the common forgotten assumptions in research design.
  2. Non-associations and Context are the commonly overlooked information in research design.
  3. A confounder is a variable that can create spurious associations. It is also referred to as a lurking variable in statistics.
  4. Meta analysis is a statistical method of combining the results of independent studies. It uses summary data from groups of people rather than data from individual subjects. In contrast, mega analysis refers to a technique of summarising the results of independent studies using data from the individual subjects.