

Data Analytics and Mathematical Statistics
Day 1

Dr Edmund Low
National University of Singapore
edmundlow@nus.edu.sg

Source: Towards Data Science, Spring Board, The Software Report

Agenda

Lecture 1 and Tutorial 1

- How to reason with data
  - The quantitative approach
  - The issue of measurement
  - Data collection
  - Introduction of group project requirements

Lecture 2 and Tutorial 2

- Cleaning and exploring our data
  - How to clean datasets
  - Descriptive statistics
  - Data visualization
  - Hands-on data cleaning and exploration

Agenda

Lecture 3 and Tutorial 3

- Using our sample and assessing the findings
  - The role of probability in reasoning with data
  - Hypothesis testing
  - Hands-on assessing sample findings

Lecture 4 and Tutorial 4

- Investigating trends and relationships
  - Using models
  - Checking model assumptions
  - Hands-on model building

Agenda

Lecture 5 and Tutorial 5

- Reviewing our quantitative analyses
  - Data misrepresentations
  - Communicating our insights
  - Quiz

Lecture 6

- Final project presentation

Why data?

Leaders

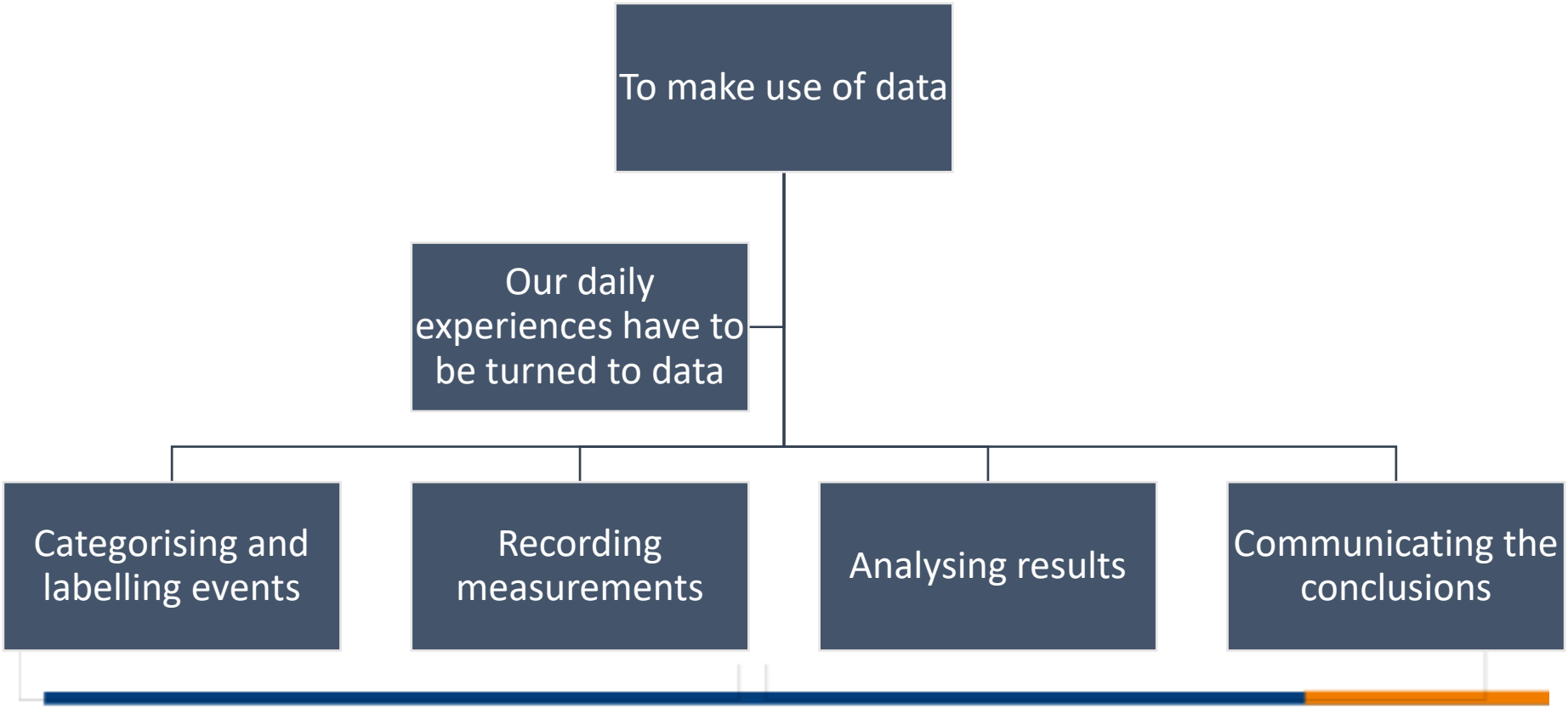
May 6th 2017 edition ›

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

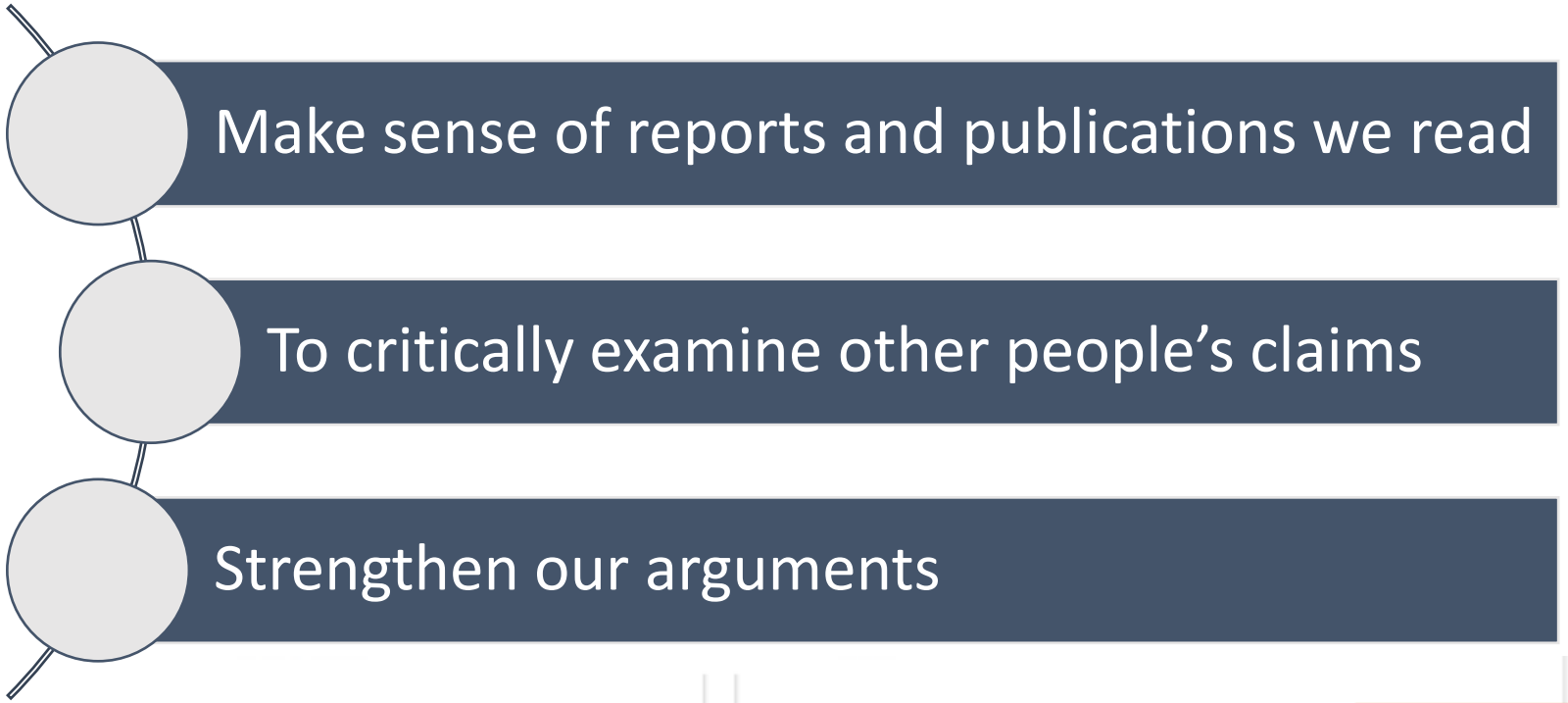
The Economist, 6 May 2017

Converting…well, everything…to data



Reasoning with data, quantitatively…

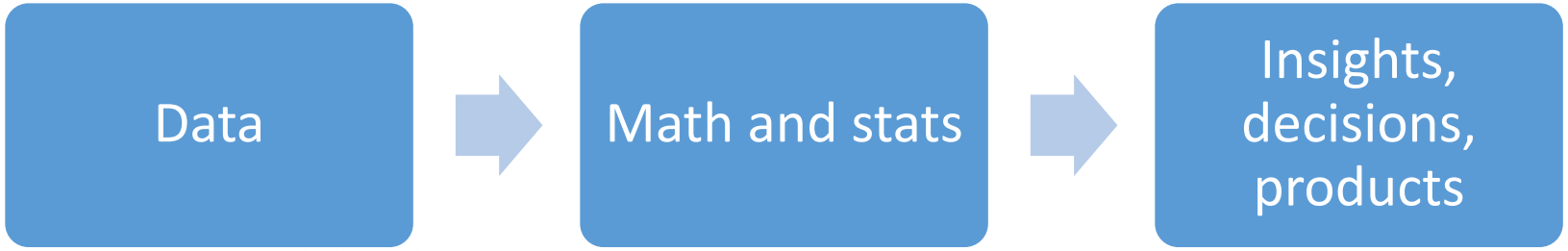
- Understanding quantitative tools such as significance will allow us to:



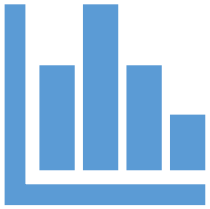
The quantitative toolkit

"Data science (and analytics) is the transformation of data using mathematics and statistics into valuable insights, decisions, and products"

- John Foreman, Chief Data Scientist, Mailchimp.com



The quantitative toolkit



While the course will cover some of these data analytics tools founded on math and statistics

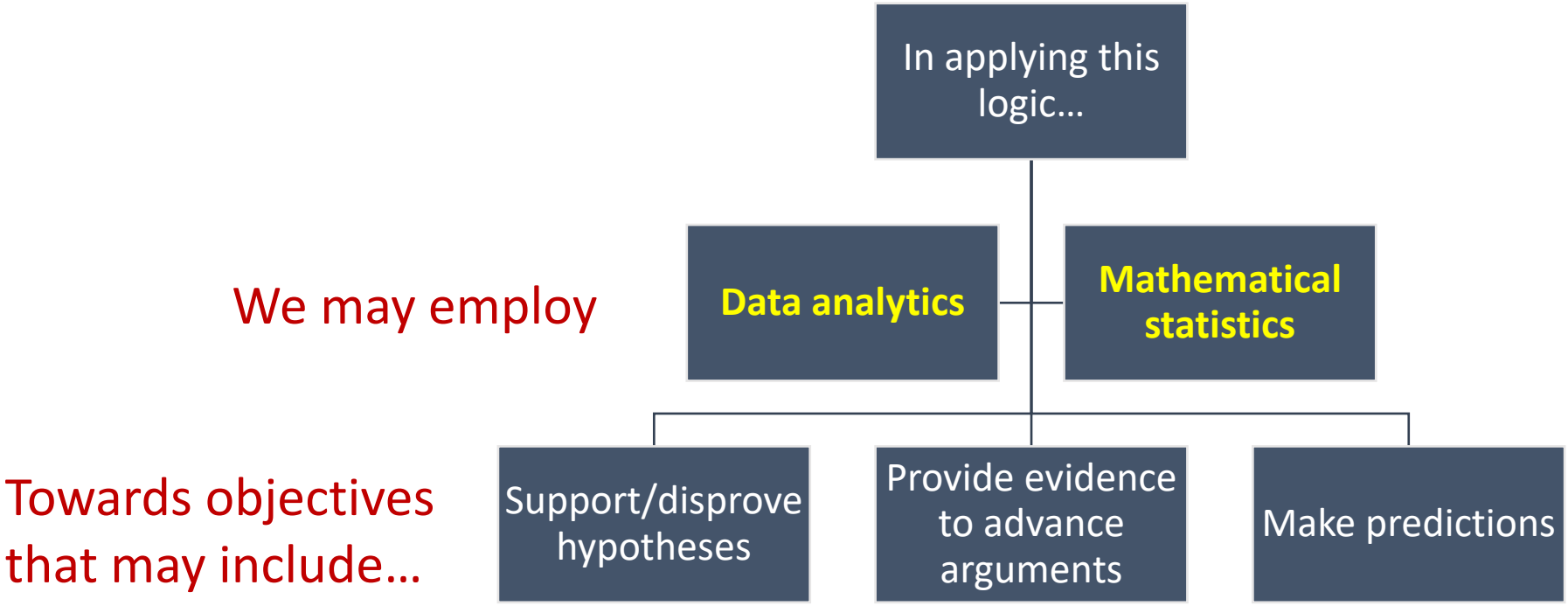
Another key part of the equation here is the underlying reasoning process in employing these tools to answer questions→ focus of the course

Motivation

Delve into the thinking process that underpins the use of data – quantitative reasoning

We can think of quantitative reasoning as applying a certain set of logic that allows us to work with data

Reasoning quantitatively…



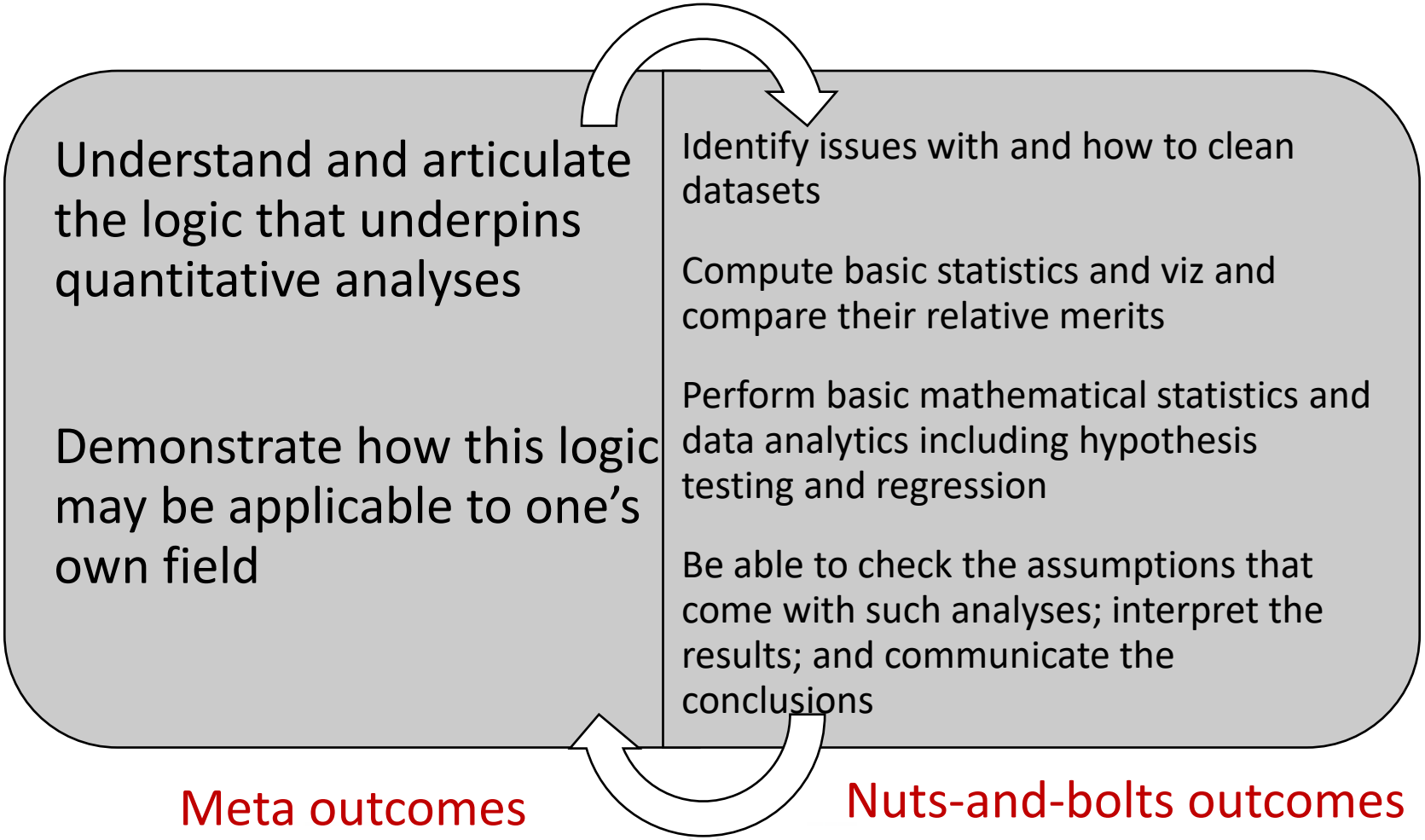
Course objective

- Develop the relevant skills that will allow us to apply this logic, so that we may effectively utilise data as a resource, and become more critical consumers of quantitative information

Active, effective "engagers" of data

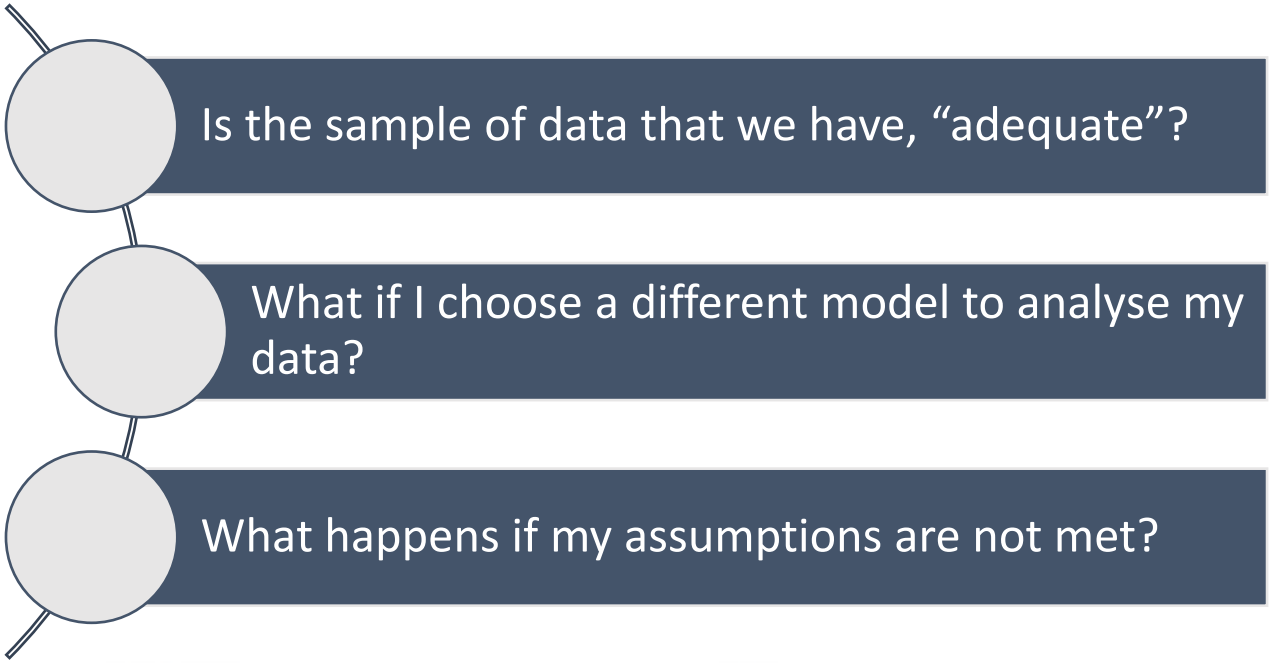
Informed consumers of data that we encounter

Learning outcomes – takeaways



Takeaways

- With such understanding of the process in reasoning with data, we can also ask questions of it



Takeaways

- Even though this reasoning may already be subconscious, it is important to make explicit and be aware of this process so that we can be more rigorous in its application
  - Being able to adapt it to suit our needs
  - Be cognizant of the strengths and limitations in the use of data, and the conclusions we can draw
  - Identify when data has been misrepresented, misused, misinterpreted etc

Before and after

Taking the course before or after other data analytics courses

- **Before:** Can provide the foundational perspective and mindset to better prepare for the acquisition of analytics and visualization skillsets

- **After:** Or an opportunity to reflect upon and gain better understanding, in hindsight, of the logical process that underpins the utilization of the tools that we have learned earlier

Before and after

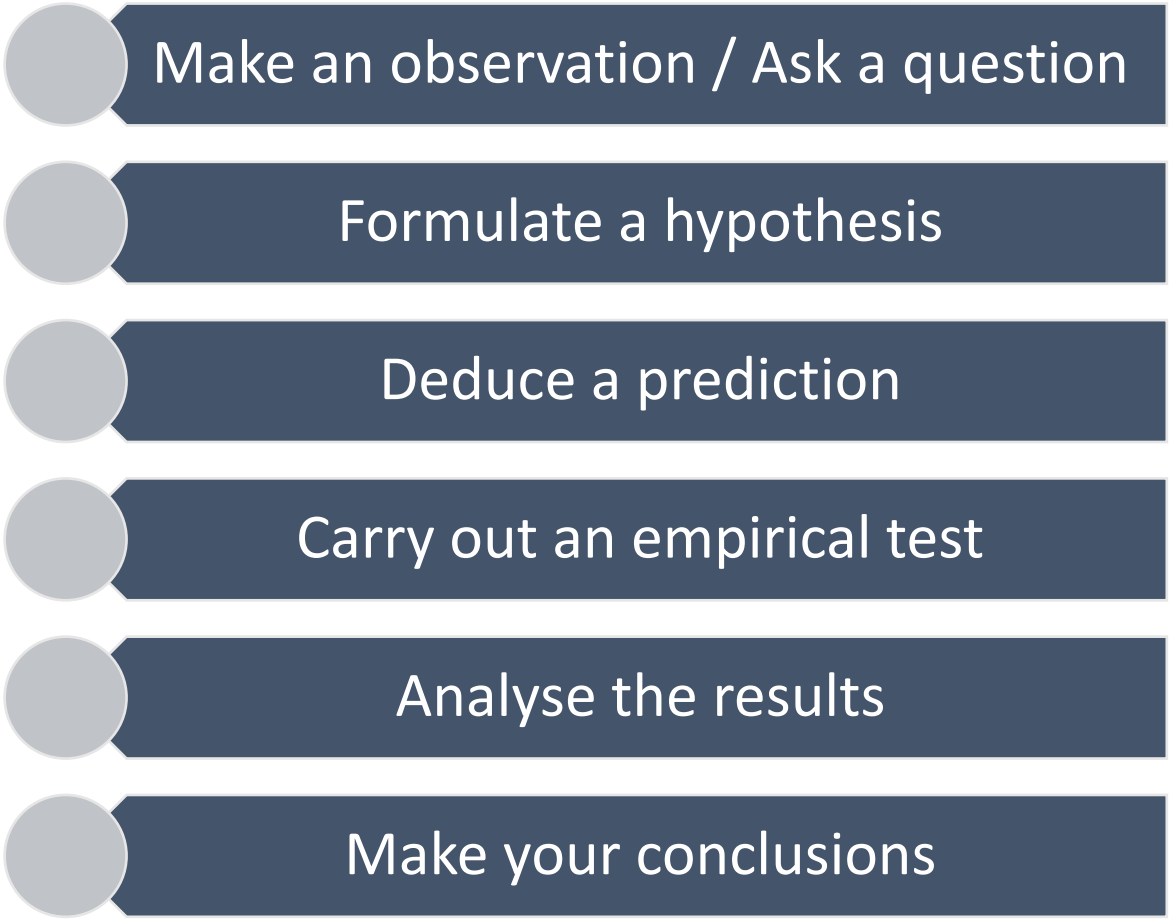
Taking the course as a standalone

- Data is ubiquitous in this day and age

- The course can help prepare us to be an informed “data citizen/employee” who can both “read” data critically and actively use data as a resource to add value in our daily lives/at work

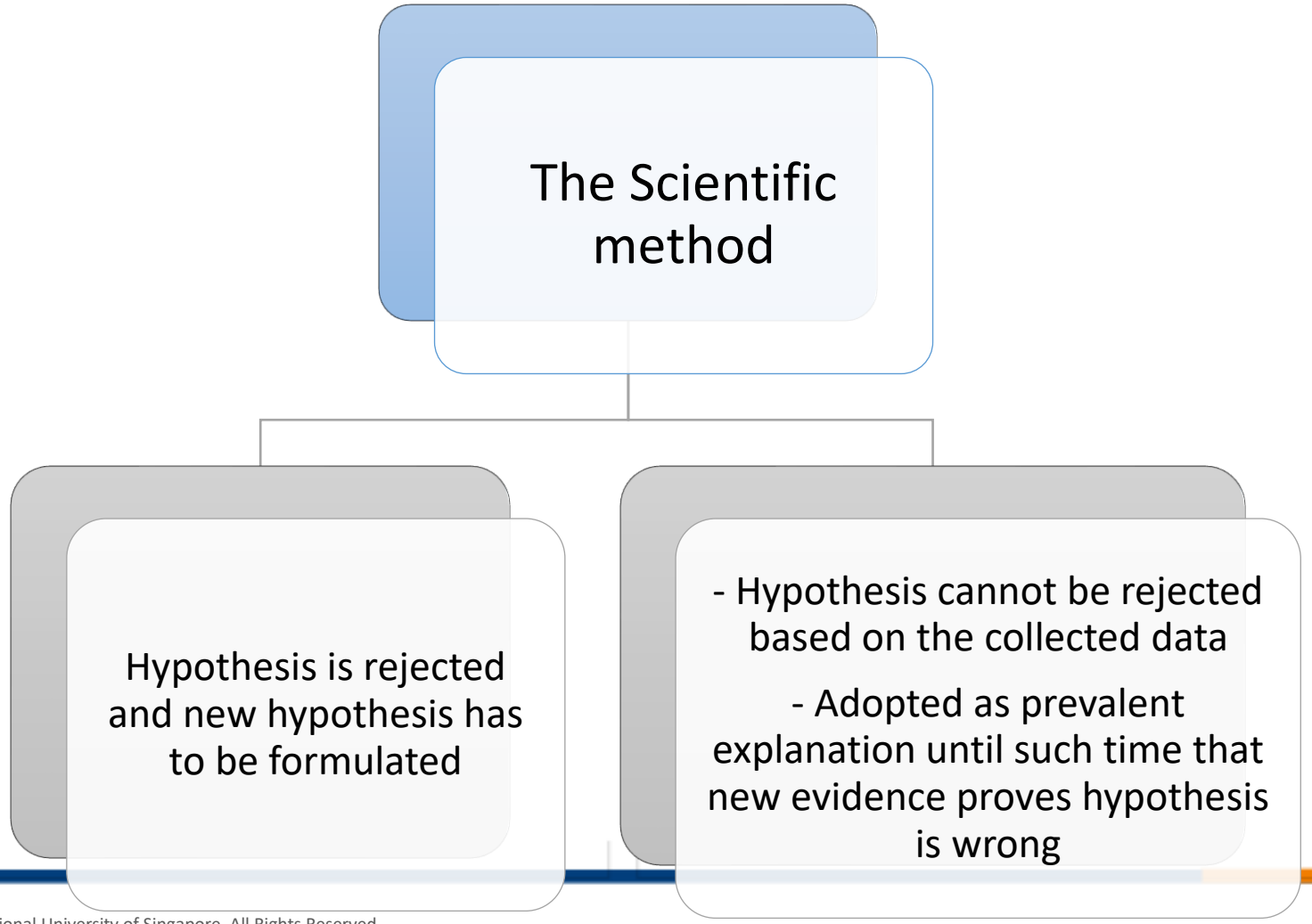
Reasoning with data

The Scientific Method



Method of investigation by which objective knowledge regarding phenomenon of interest is obtained

The Scientific Method



Let's take a look at each of these steps…

Making an obs, asking a qn

- What type of beer has higher alcohol content, the one brewed with ale yeast, or the one with lager yeast?

Source: Euronews

Formulating hypotheses

- Made on the basis of limited evidence, survey of existing literature, or exploratory studies
- Suggested explanation of the observation we made
  - Generally provide a causal explanation or propose some correlation
- A hypothesis must be falsifiable:
  - Objective methods are available for the hypothesis to be proven incorrect

Falsifiability – Example

- "No human is immortal" is not falsifiable, as we would have to observe a human living forever to falsify this claim

- "All humans are immortal" is falsifiable, as one dead human would disprove this claim

Formulating hypotheses

- Causation:

- Falsifiability

Exploratory vs confirmatory studies

- We saw that exploratory studies are one way in which hypotheses can be generated
- We may have a question that we are invested in, e.g. what motivates employees to be satisfied in our company
- But we don’t have any hypothesis in advance as to what the pertinent factors are

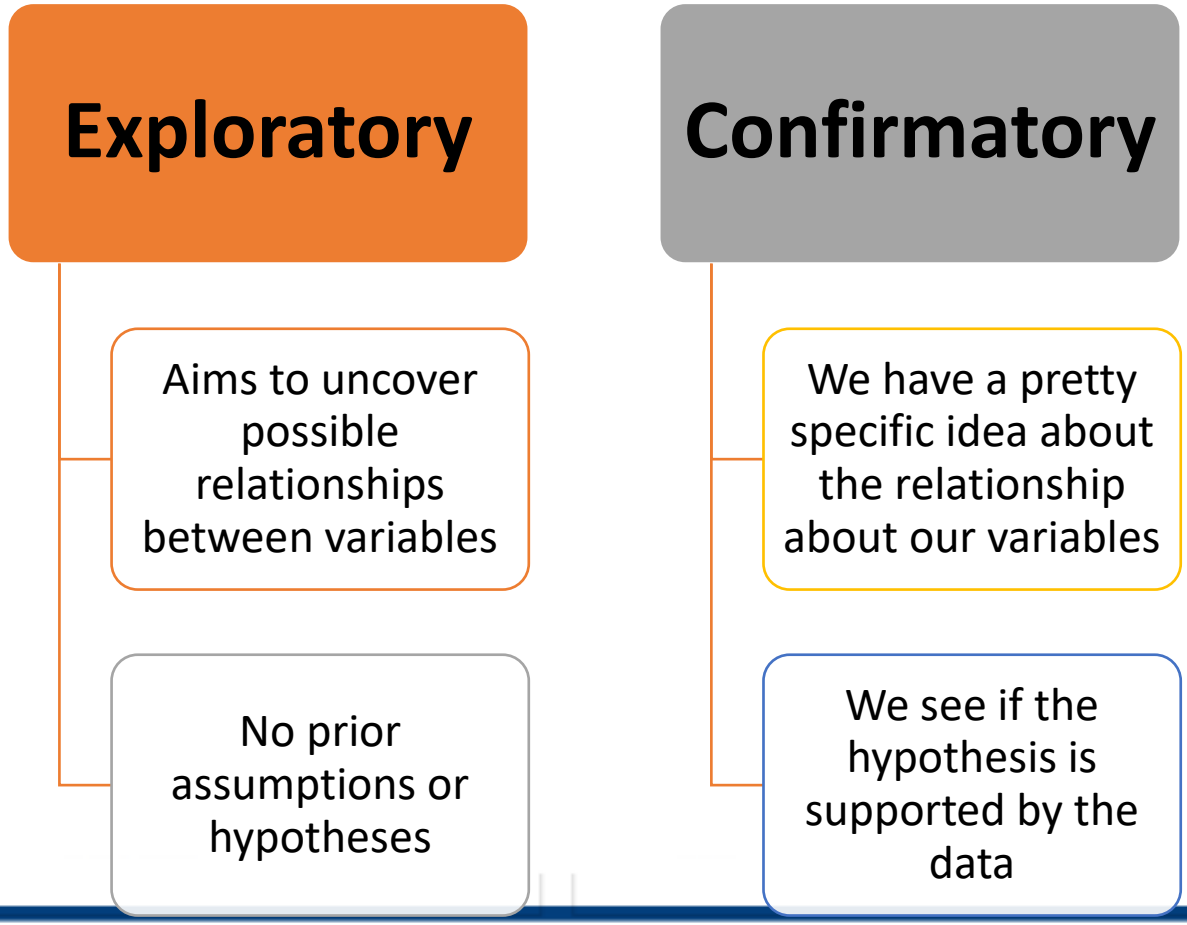
Exploratory vs confirmatory studies

- So we could collect some data through surveys with employees in your own department

- Based on the data collected, we might then identify e.g. salary and no. of promotions as factors that crop up quite often

- From here, we'll then formulate the hypothesis that these two factors increases satisfaction, and see if our confirmatory study can then reproduce the findings

Exploratory vs confirmatory studies



Deducing a prediction

- Well, now we have a possible answer, how to we go about assessing our claim?

- More importantly, how to go about assessing our claim, quantitatively?

Deduction of prediction

- We have our hypothesis
  - Need a way to test it and predict what will happen in the test
  - I.e. we deduce a specific prediction based on our hypothesis (general statement)
- IF-THEN statements, i.e. “if my hypothesis is true, then I expect to see X result in an experiment
- Predictions are usually test-specific

Operationalise

- Hypothesis may be formulated in terms of theoretical concepts

- In order for us to test the hypothesis, we need to move from the theoretical concepts to actual measures of these concepts
- → Operationalisation

- "What are measurable quantities related to the theoretical concept?"

Operationalise

- It is quite easy in our example since alcohol content is already measurable

- But it may not be that straightforward in other problems that we're interested in, e.g.
  - "What are the factors that increases employee satisfaction?"

- How do we measure satisfaction in this case?

Operationalisation

Turnover rate per month

Satisfaction survey scores

Average length of tenure of employees

- Thinking about how we measure also makes us think about how we might design a sampling study to collect all these data

Operationalise

- Through operationalisation, theoretical concepts can now take on “values”
  - Measurable quantities
  - Can be tested

- Through measurement, we can quantify these variables

Variables

- A random variable is a quantity that can take on one or more values, each of them associated with a probability, e.g.
  - Alcohol content in a beer brew
  - The number of people who have left the company each month

Types of Variable

- Numerical variables
  - –Continuous variables
  - –Discrete variables
- Categorical variables
  - –Ordinal variable
  - –Nominal variable

Name	Sex	Age	Marital status	No of children	Income	Smoking
John Smith	male	24	single	0	Low	never smoked
Mary Brown	female	35	married	1 to 3	High	current smoker
Adam Jones	male	42	divorced	4 to 6	Medium	former smoker
Jane Robertson	female	29	divorced	Above 6	High	never smoked

http://www-ist.massey.ac.nz/dstirlin/CAST/CAST/Hstructures/structures_c2.html

Likert Scale

- What is the variable type?

Very Interested 5	Somewhat Interested 4	Neutral 3	Not Very Interested 2	Not at All Interested 1
Very Much 5	Somewhat 4	Undecided 3	Not Really 2	Not at All 1
Very Much Like Me 5	Somewhat Like Me 4	Neutral 3	Not Much Like Me 2	Not at All Like Me 1
Very Happy 5	Somewhat Happy 4	Neutral 3	Not Very Happy 2	Not at All Happy 1
Almost Always 5	Sometimes 4	Every Once In a While 3	Rarely 2	Never 1

Dependent and Independent Variables

- Variables: labeled either as “dependent” or “independent” variables

- Dependent variable
  - Corresponds to the outcome we are interested in explaining
  - Fully or partially determined by other variables

- Independent variable
  - Varies independently of any other variable
  - Fully or partially responsible for our DV

Dependent and Independent Variables

After operationalisation, we can deduce our prediction

- If our hypothesis is true, then I expect to observe:

- An increase in satisfaction survey scores with an increase in monthly salary, number of promotions attained, and amount of bonuses received

- Keep our prediction strictly to what we expect to observe, since our test may not be able to prove causality (more on that later)

Operationalisation issues

- Construct validity: does the variable that we have operationalised actually measure what it is supposed to measure?

- E.g. is average tenure length really measuring satisfaction?

Empirical test

- Empirical test of prediction through the collection of data

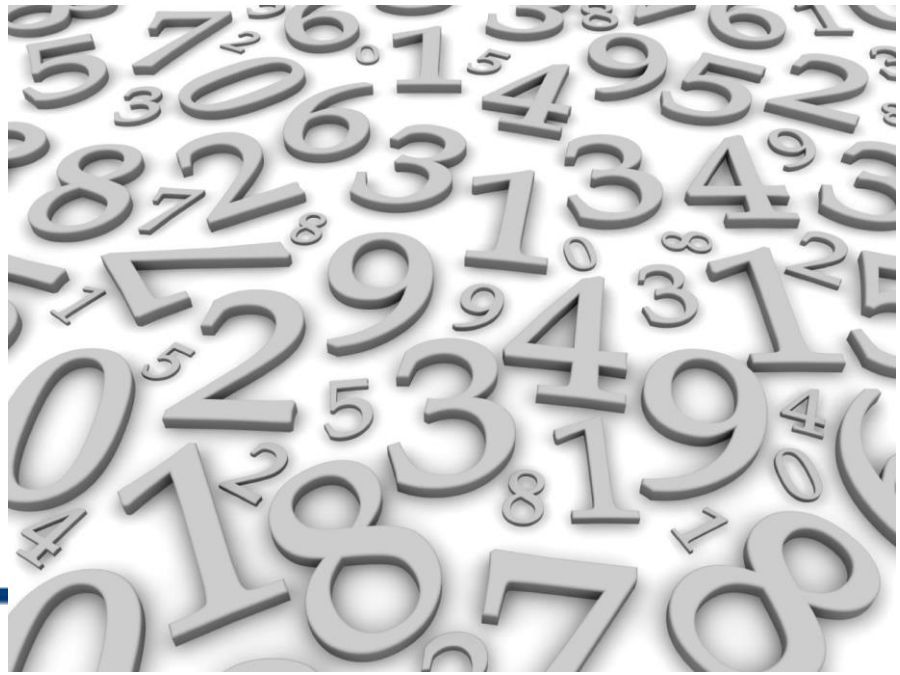
Analyse the data

Draw conclusions

Data

- Numerical / categorical values of physical quantities
  - E.g. alcohol content, satisfaction survey score, school

- Be focusing exclusively on quantitative data in the form of numbers, categories



Empirical test

- How do we go about collecting data?

Empirical test

- Empirical test of prediction through the collection of data

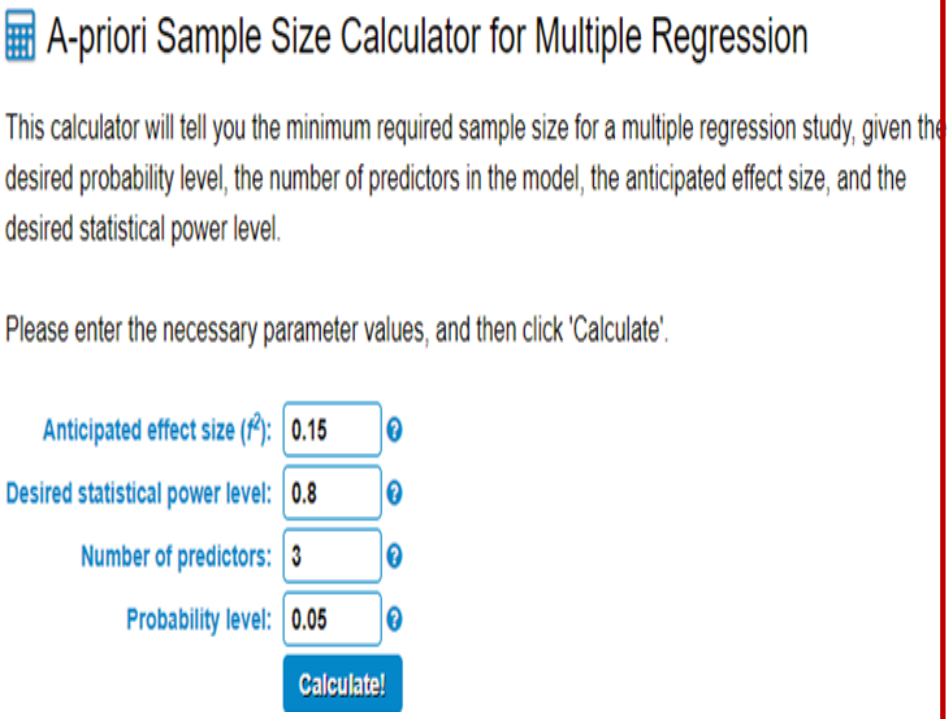
- Data Collection

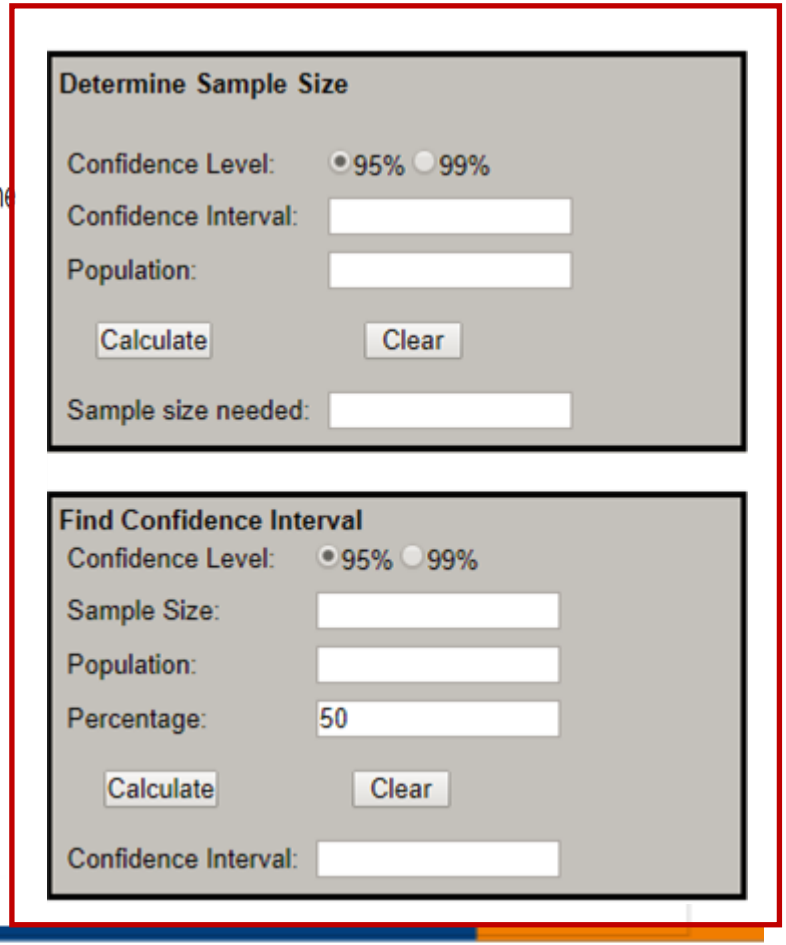
Designed experiments

Observational studies

Retrospective studies

Sample size?

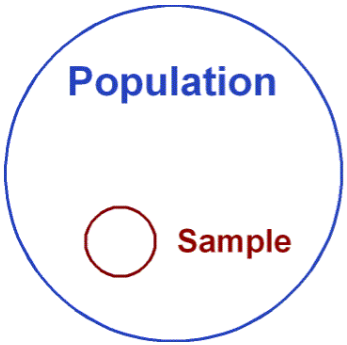




http://www.surveysystem.com/sscalc.htm

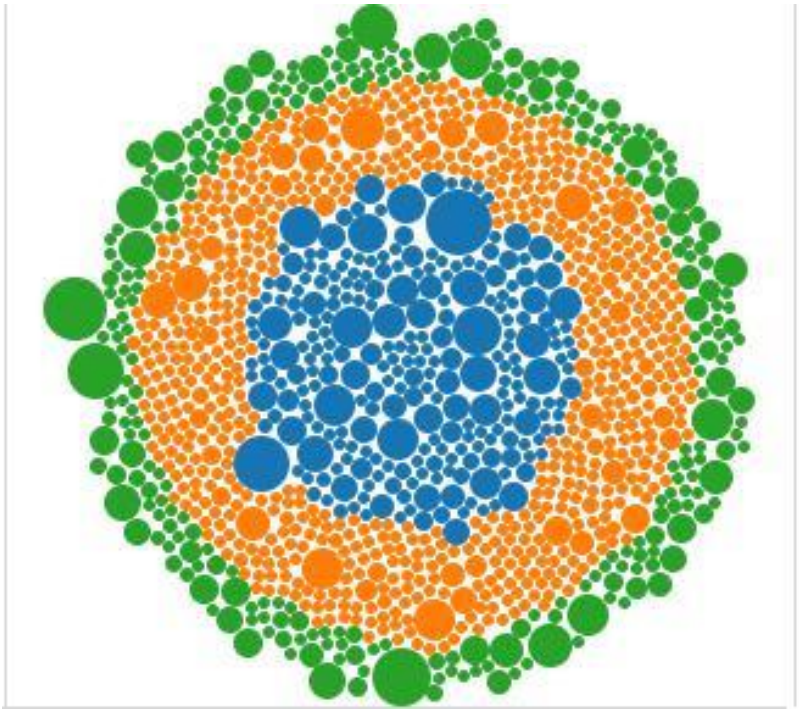
Random Sampling

- What is random sampling?
  - –Random sampling, also known as probability sampling, is a probability based method of data collection that ensures each portion of the population undergoing studying has a chance to be selected at random.
- Why random sampling?
  - –One of the best ways to achieve unbiased results in a study is through random sampling
  - –Everyone has an equal chance of being selected.
- But it is difficult and expensive.

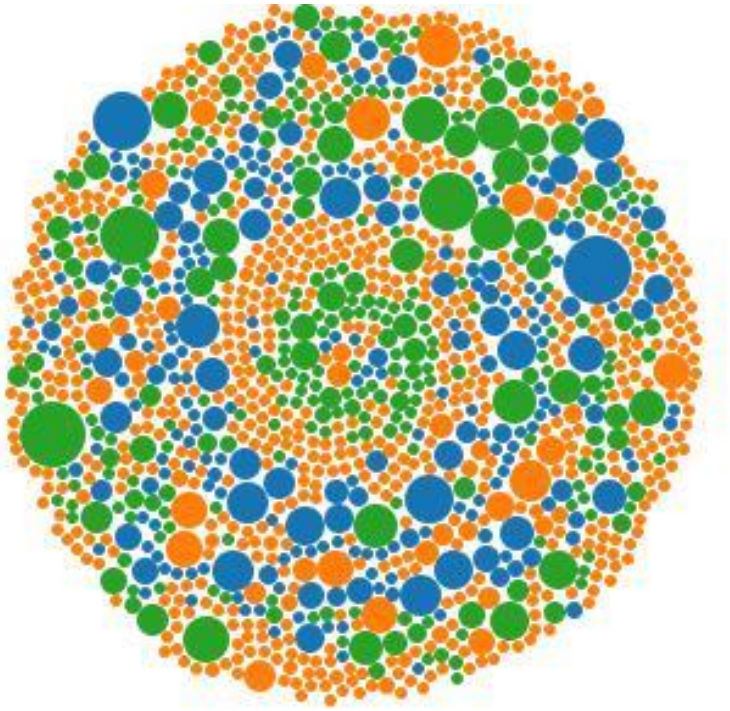


Random sampling

- If blue = those I sample



Convenient



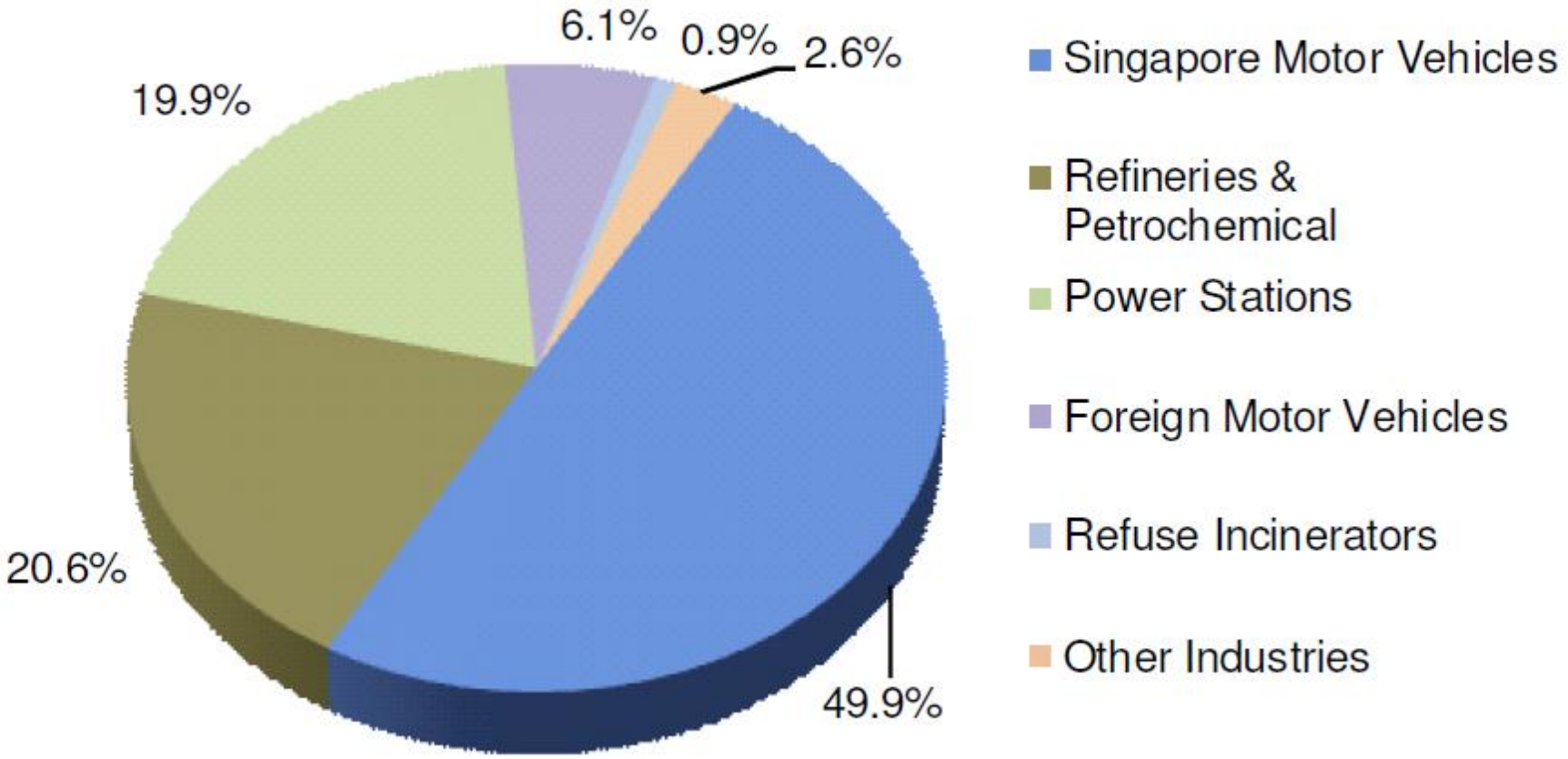
Random

Source: Paint by Numbers

Empirical test

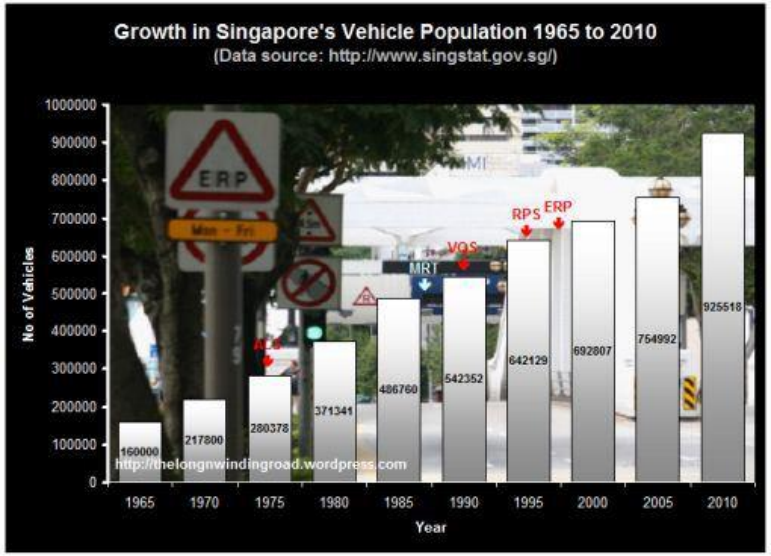
- Retrospective studies
  - Uses all or a sample of historical data archived over some period of time
  - E.g. relationship between diesel vehicle population and PM2.5 levels in Singapore

2012 Percentage Contribution of $PM_{2.5}$ Emissions



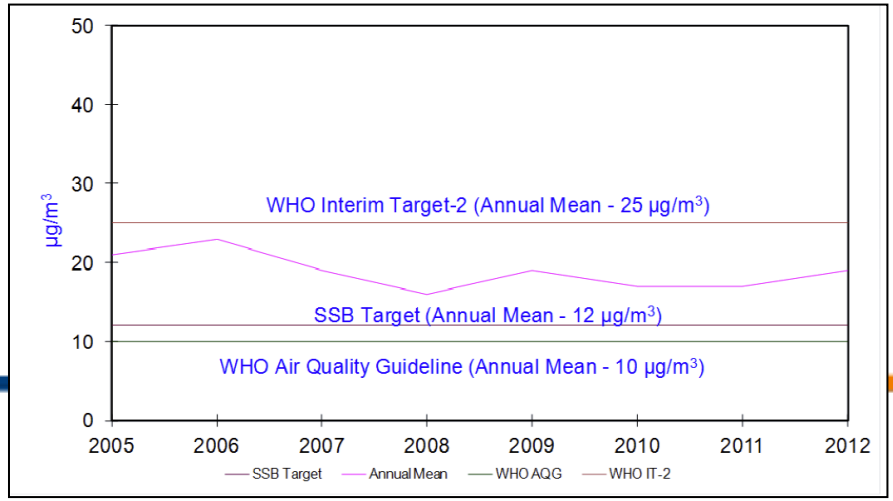
Source: NEA

Empirical test



Annual diesel vehicle population over the years

Daily PM2.5 concentrations



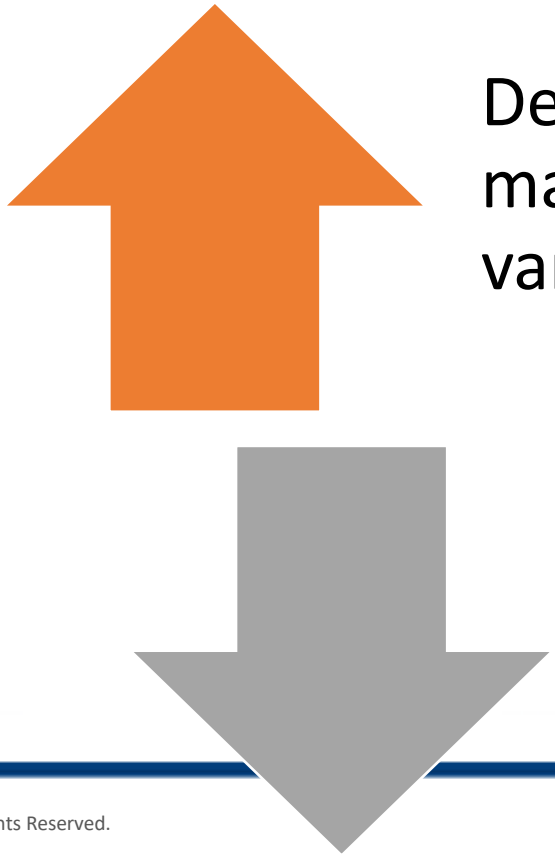
Empirical test

- Observational study
  - Observes population of interest with minimal disturbance
  - Usually carried out over a short period of time; allows variables not usually measured to be collected

- E.g. measurement of PM2.5 concentrations, observation of diesel vehicle counts, at selected locations in Singapore

Empirical test

- Designed experiments

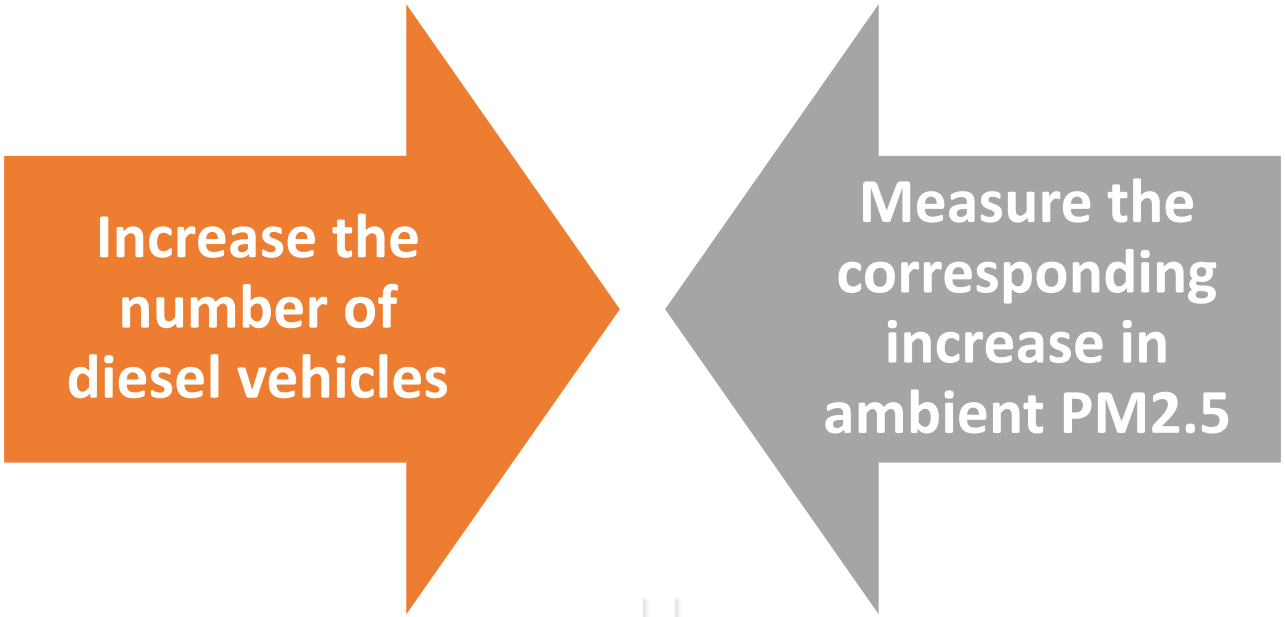


Deliberate changes made to controlling variables

Resulting change in dependent variable observed

Empirical study

- Designed experiments



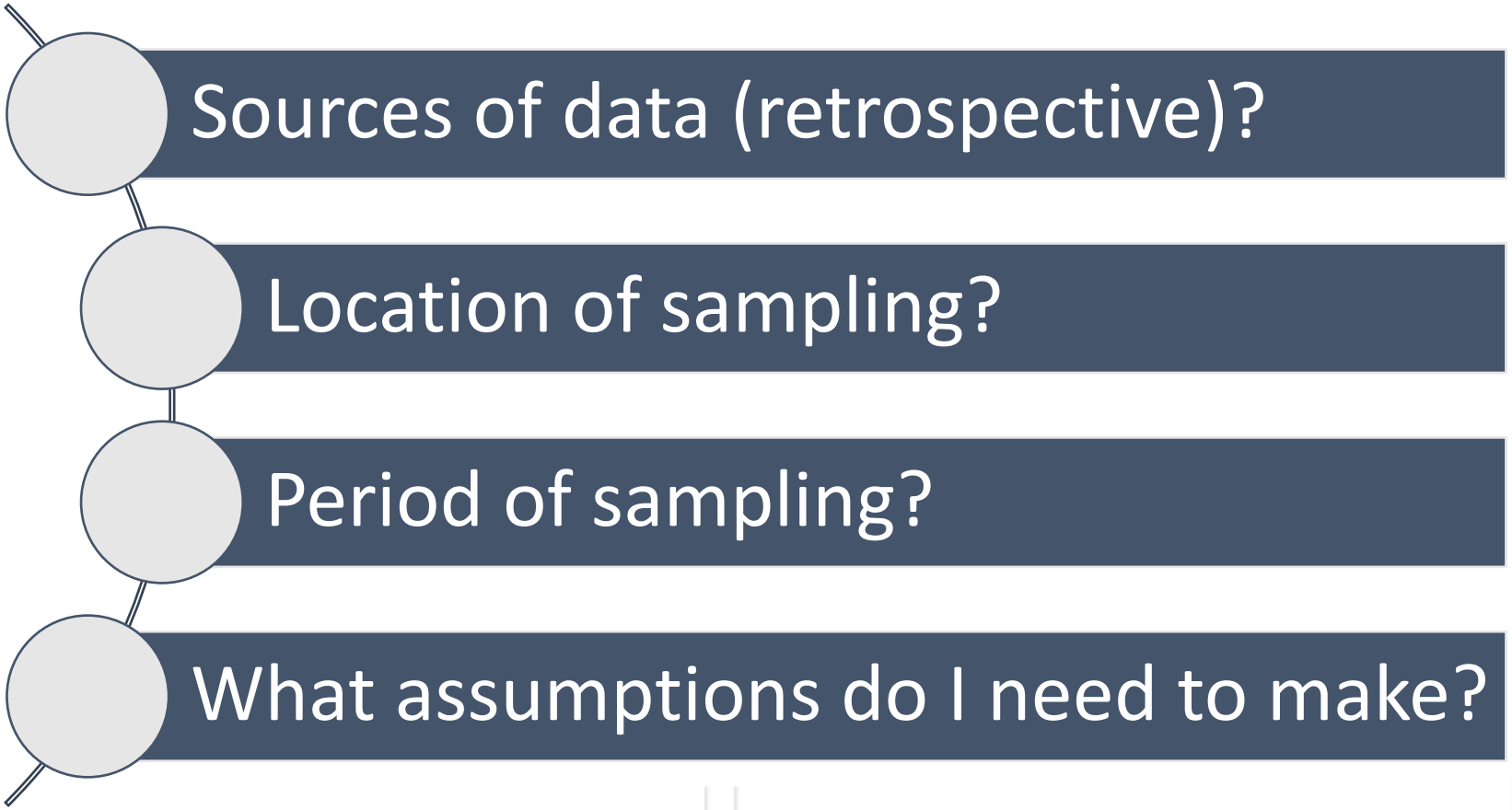
What is the method of collection?

- If we review past records of alcohol content in beer batches?
- If we take measurements during the brewing of the beer batches under normal operational circumstances?
- If we make deliberate changes to the amount of yeast that is placed across differing beer brews?

What is the method of collection?

- If we review past records of monthly employee turnover?
- If we administer a survey asking employees information on their satisfaction levels
- If we change the amount of bonuses given to a sample of 30 employees over time, and compare the survey scores

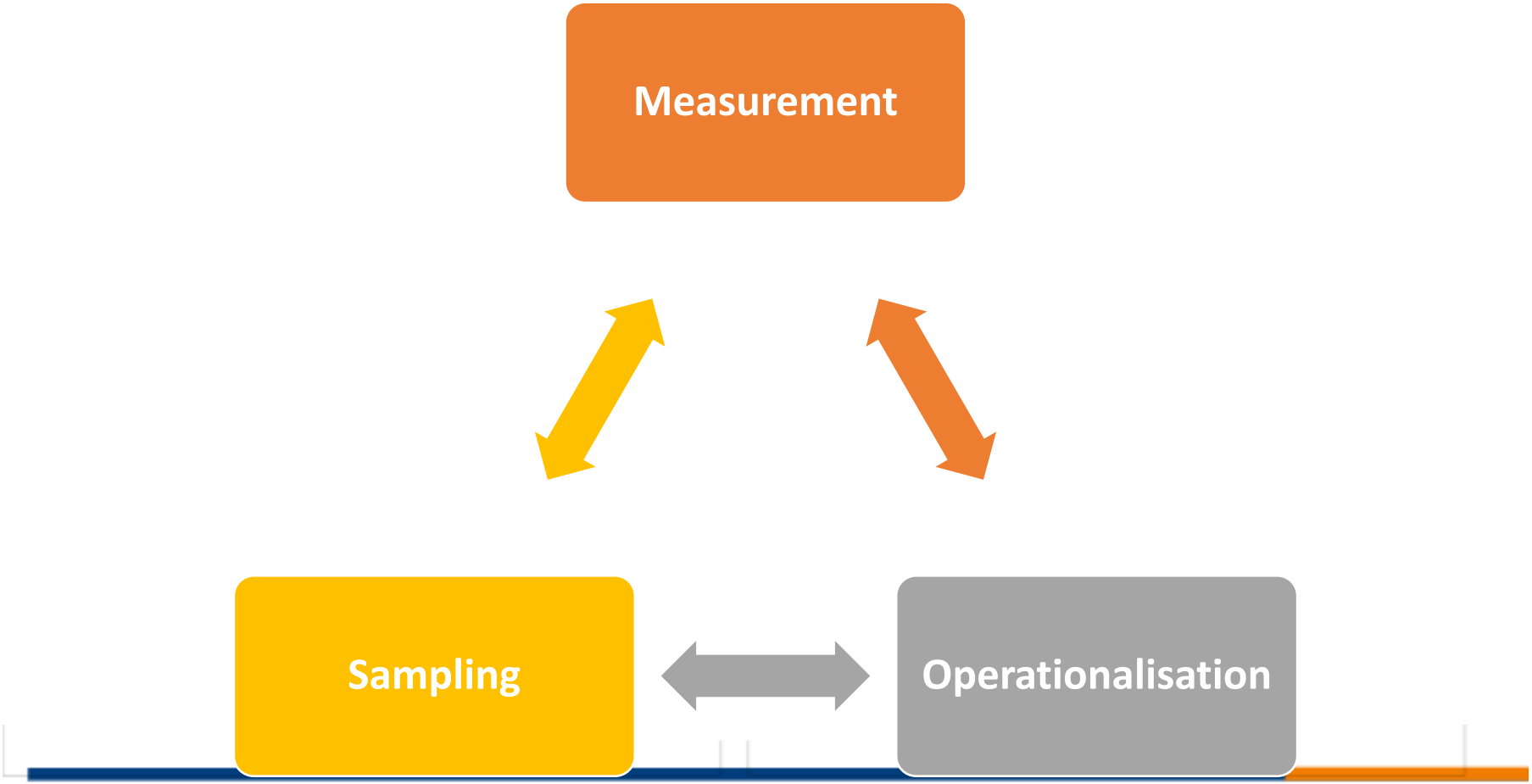
Asking questions when collecting data



Asking questions when collecting data

- For example, for our motor vehicle-PM2.5 study
  - How long do we measure, if we carry out an observational study? One month, one year?
  - Where do we measure? In just the west of Singapore, at major traffic junctions?
  - What assumptions do we need to make? E.g. do I measure PM2.5 at the same time instant as the number of motor vehicles, or some time after?

What could affect the validity of outcomes in our study

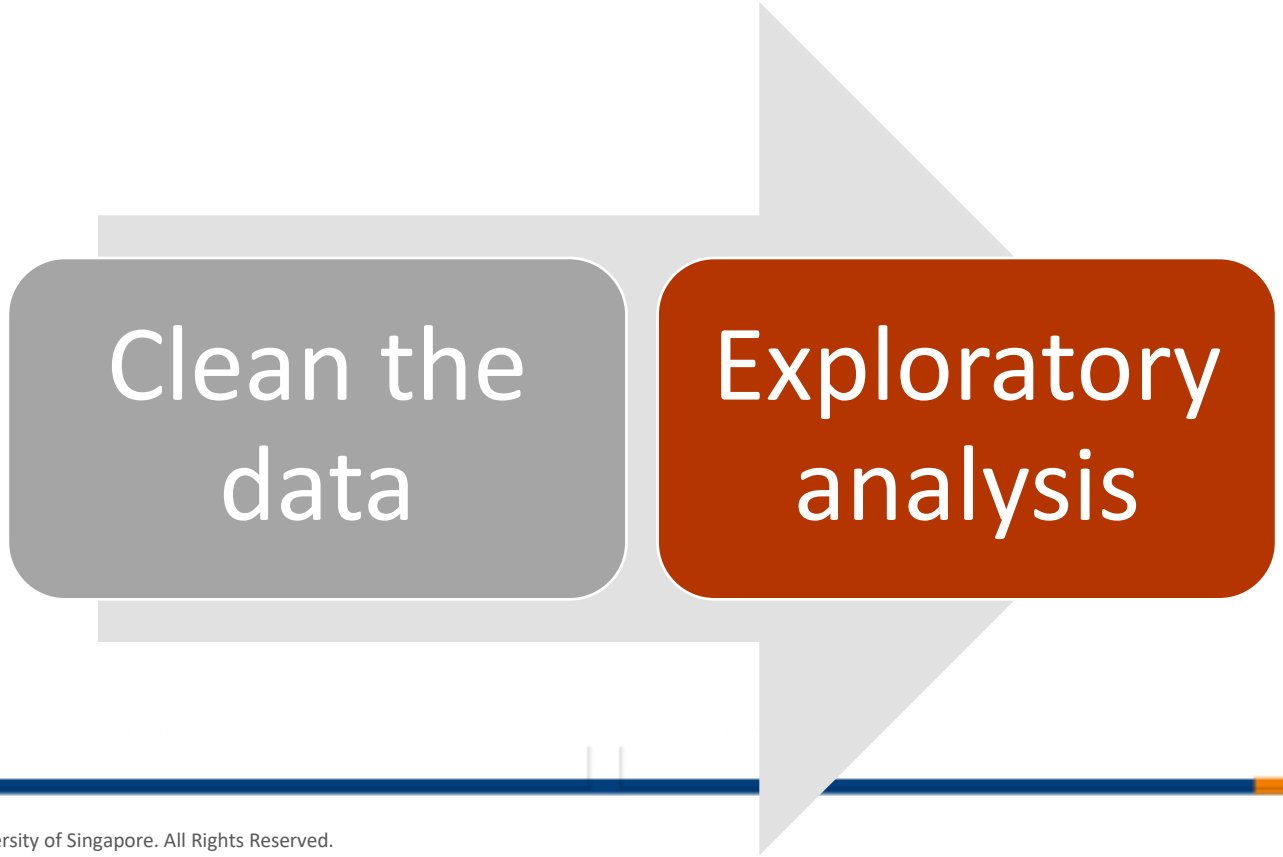


What could affect the validity of outcomes in our study

- Measurement
  - Alcohol content recorded incorrectly
  - Errors in transcribing employee satisfaction survey results to computer
- Operationalisation
  - Construct validity issues
- Sampling
  - We only use one pair of beer brews
  - We only survey senior management

Empirical study

- After collecting the data, prior to analysis, we need to



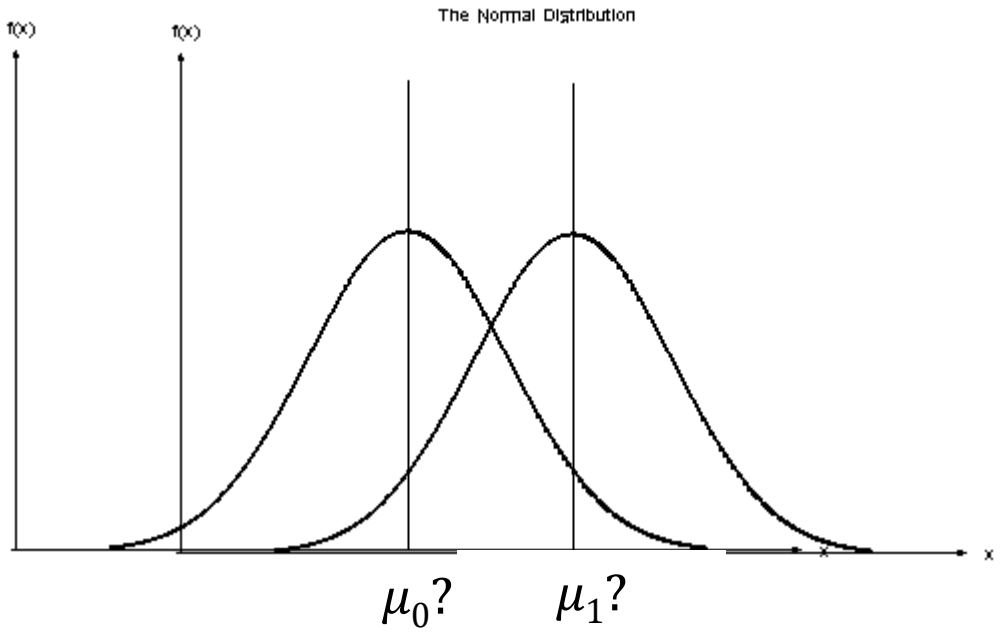
Analysis

- We have the data from our empirical test
  - Data for independent variable, dependent variable
  - Now what do we do?

- We deduced relationship between IV and DV
  - Represent this relationship in mathematical form, i.e. we model this relationship
  - E.g. $y = \beta_0 + \beta_1 x$

E.g. we predict that employee satisfaction would increase with monthly salary…how do we model this mathematically?

Analysis

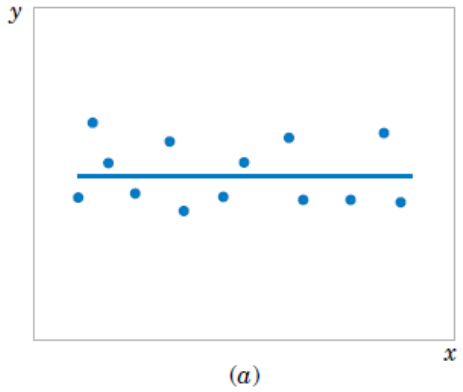


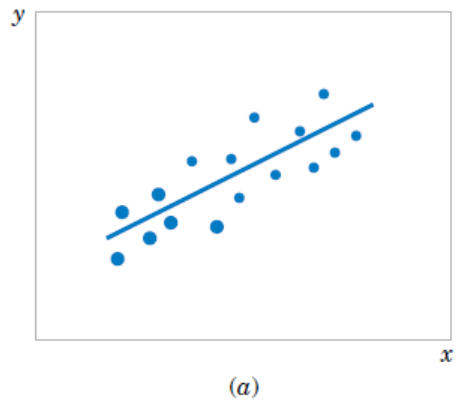
Analysis

- $y = \beta_0 + \beta_1 x$

- How do we use the data in our model?

- Do we have evidence to support our hypothesis?





Conclusion

Interpreting results	Making conclusions	New ideas	Communication
Does the evidence support or refute our hypothesis? Do all our assumptions hold?	How do we use this quantitative evidence to build a sound argument?	What follow up hypothesis can we generate to troubleshoot / add to what we know?	Communica-ting results to our audience

End of Day 1