Welcome to DATA 151

I’m so glad you’re here!

DATA 151: CLASS 2B
INTRODUCTION TO DATA SCIENCE (WITH R)

EXPERIMENTAL DESIGN

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

WELCOME!

Please sit down in your **previous locations (and groups)** from last class and get the **box of beads used on Tuesday**.

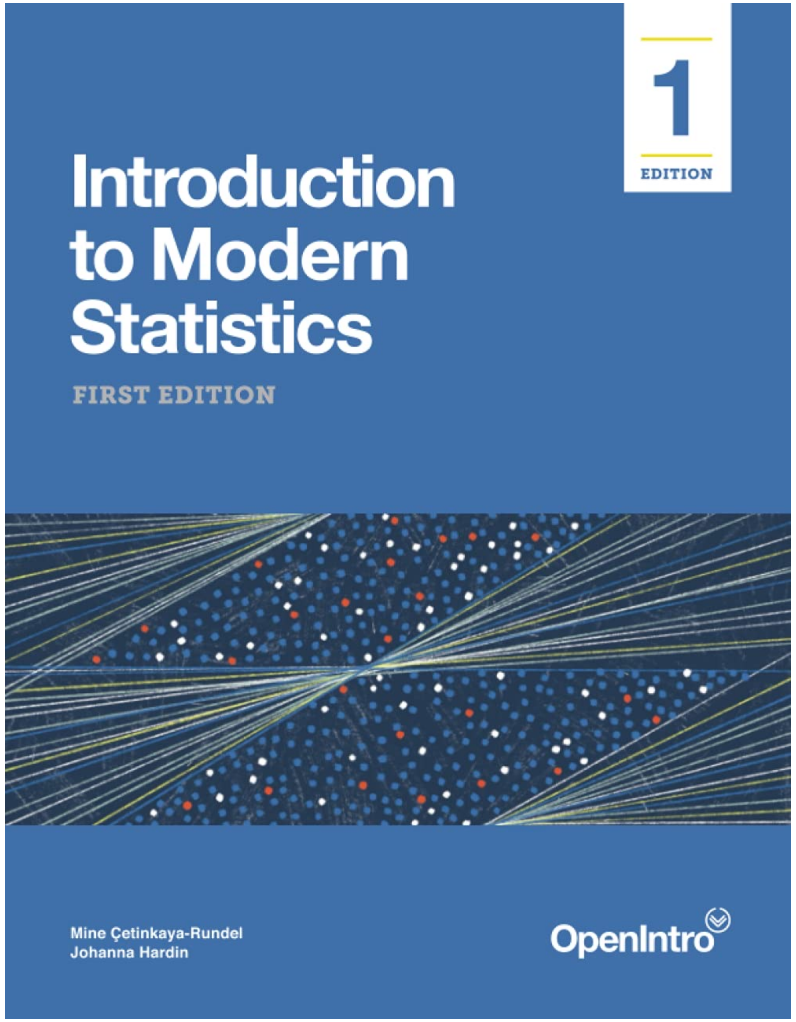
You will also need your **Worksheet from Class 2A**

AGENDA: DATA 151 – CLASS 2B

Time	Topics
2:30 – 3:00	Finish Sampling Activity from Class 2A Sampling Designs
3:00 – 3:15	Experiments vs Observational Studies Confounding Variables
3:15 – 3:20	*5 minute break*
3:20 – 3:45	Designing Experiments
3:45 – 4:00	Getting into project groups

ANNOUNCEMENTS

RELEVANT READING



Introduction to Modern Statistics:

- Tuesday:
  - iMStat - Ch 2: Study Design Sections:
    - 2.1: Sampling Principles and Strategies
- Thursday:
  - iMStat - Ch 2: Study Design Sections:
    - 2.2: Experiments
    - 2.3 Observational Studies

HOMEWORK REMINDER

Due next week:

- *HW #2: Practice Problems* **(due on WISE 9/15)**
- *Project Milestone #0: Communication Plan*
  - **Due on WISE 9/15**
  - *One submission per group*

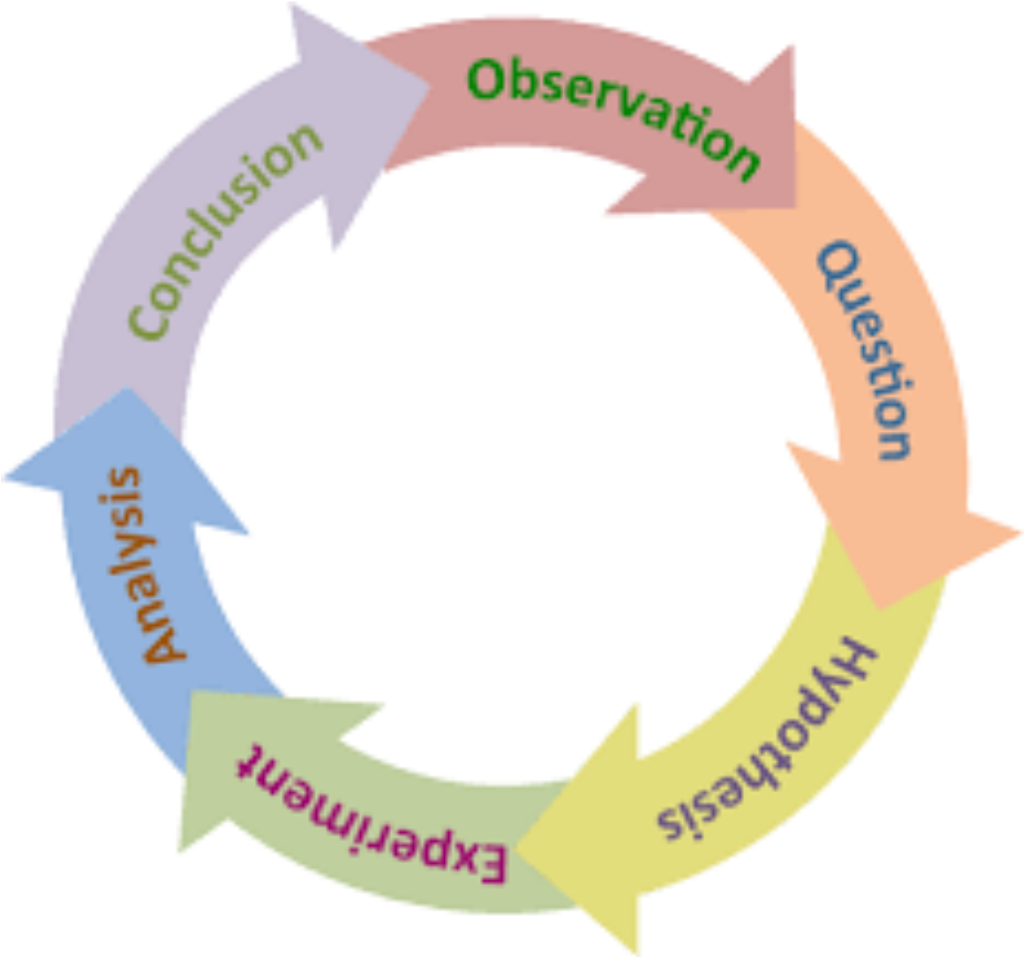
FRIENDLY REMINDERS

- Please **bring your laptop to class** on Tuesday next week (and all class days after that) for programming in R
- If you would like to download R onto your personal computer
  - Download **R** first (https://www.r-project.org/)
  - Then download **R Studio** (https://www.rstudio.com/products/rstudio/)
  - We will be using the **Willamette R Studio Server** in class

MAKING DECISIONS WITH DATA

THE STATISTICAL METHOD MIRRORS THE SCIENTIFIC METHOD

1. Questions / Purpose
2. Hypothesis
3. Experiment
4. Analysis
5. Conclusion / Report findings



EXPERIMENTAL DESIGN

RELATIONSHIPS BETWEEN VARIABLES

Many analyses are motivated by a researcher looking for a relationship between two variables.

Definitions:

- **Response/Dependent variable (Y):** the variable one suspects is affected by the explanatory variable(s).
  - Variable that is of interest to study
- **Explanatory/Independent variable (X):** the variable whose effect one wants to study
  - Is thought to explain or influence the response variable

OBSERVATIONAL STUDIES VS EXPERIMENTS

Definitions:

- **Observational Study:** observes individuals and measures variables of interest but does not attempt to influence the responses

- **Experiment:** deliberately imposes some treatment on individuals to measure their response
  - Used when the goal is to understand **cause and effect**
  - Random experiments are the only source of fully convincing data

CONFOUNDING VARIABLE

Observational studies that try to look at cause-and-effect relationships fail because of confounding variables

Definitions:

- **Confounding variable (noun):** is a variable that is not among the explanatory or response variables in a study but that may influence the response variable
- **Confounding (verb):** occurs when two variables are associate in such a way that their effects on a response variable cannot be distinguished from each other

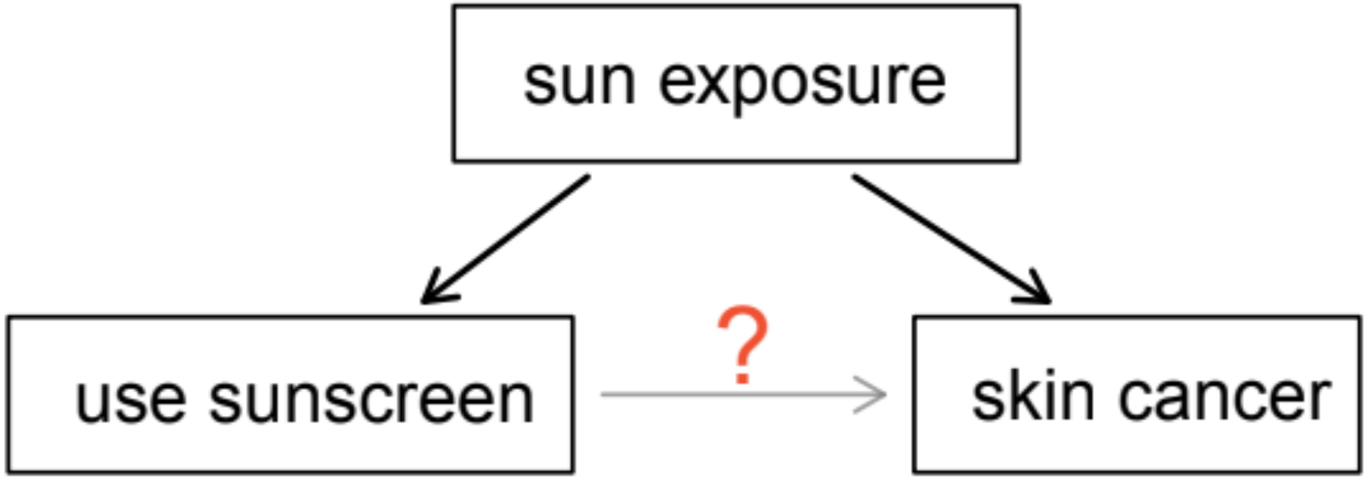
CONFOUNDING VARIABLE

Example: Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer.

Does this mean sunscreen causes skin cancer?

CONFOUNDING VARIABLE

Sun exposure is a confounding variable, which is a variable that is correlated with both the explanatory and response variables.



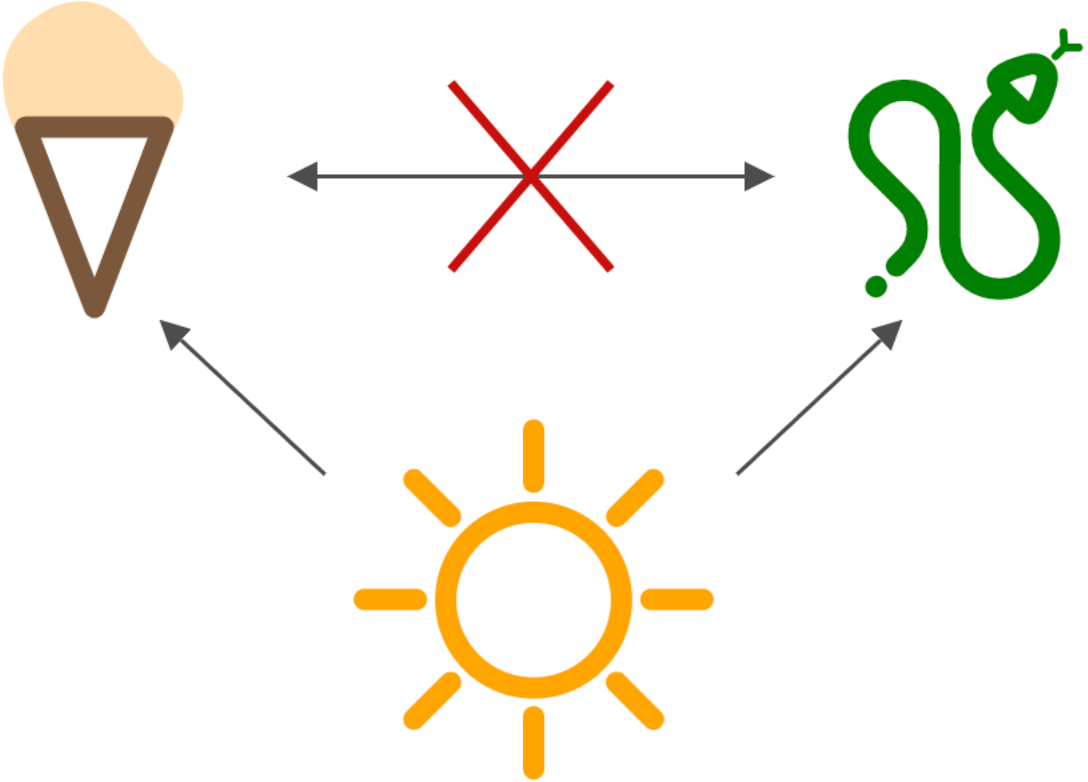
EXAMPLE

Example: In the 2011 NBA finals, the Dallas Mavericks defeated the Miami Heat. One headline read, “Miami’s real problem this series: Not enough high fives,” citing an observational study[1] that found that teams exhibiting more touching, such as high fives, early in the season had better performance later in the season. The article then went on to imply that giving high fives improves basketball performance.

a. What is wrong with using the language “giving high fives improves basketball performance” as an interpretation of the findings from the study?

b. Name a few confounding variables that could be explaining why team who give more high fives are associated with better performance later on in the season.

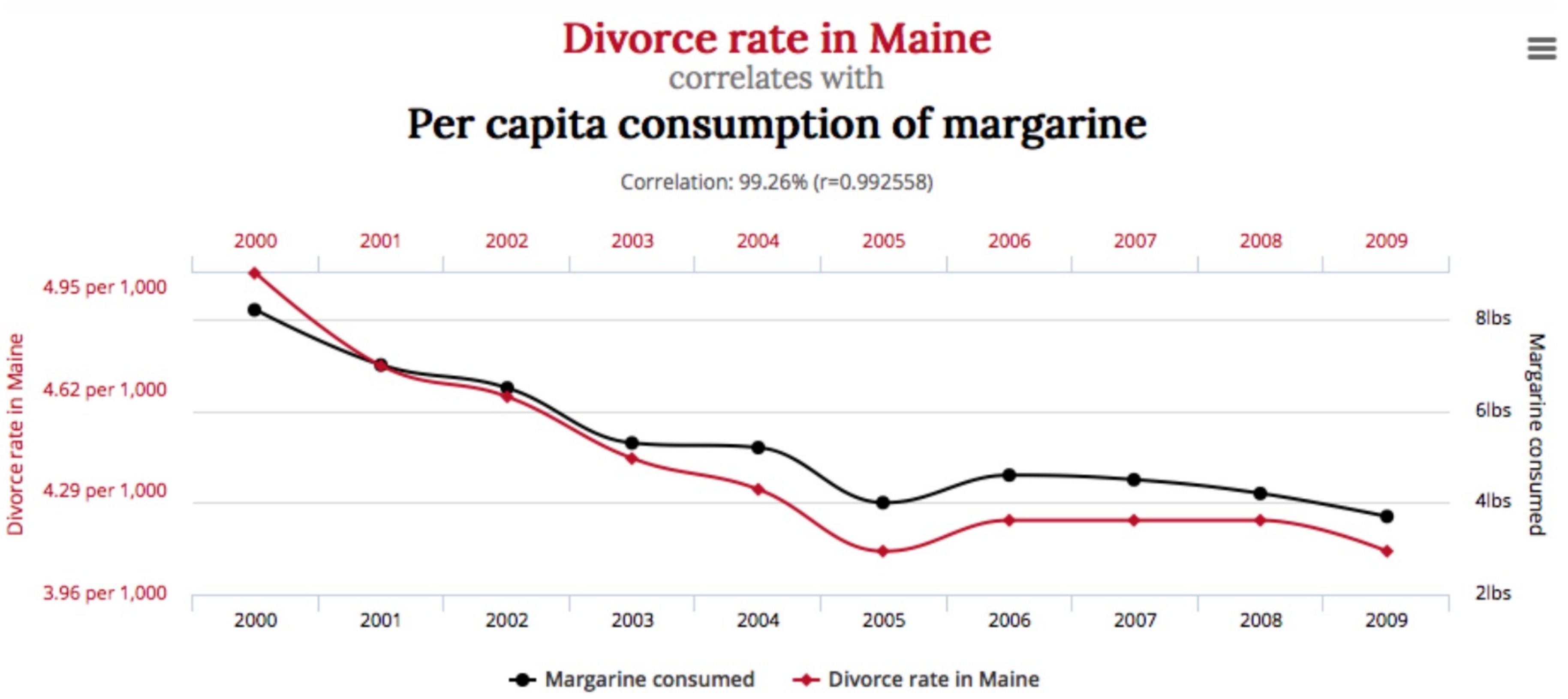
CORRELATION ≠ CAUSATION



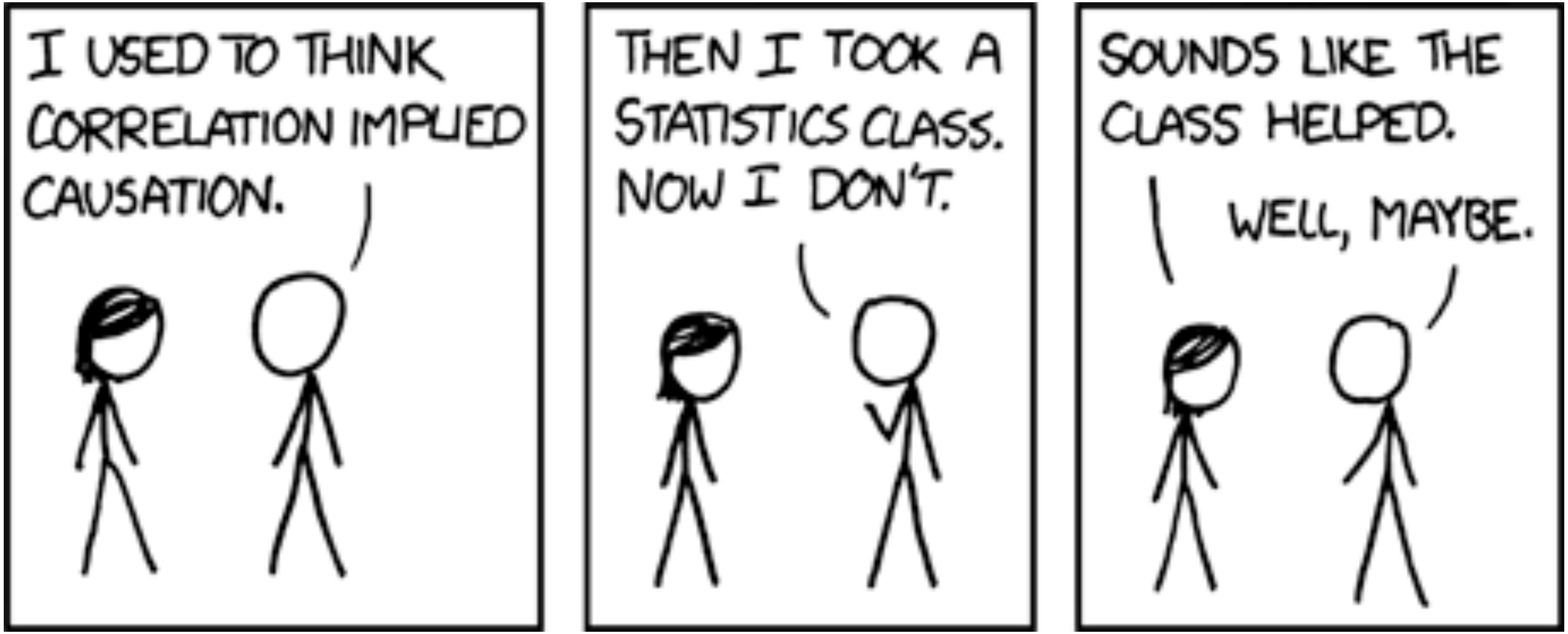
- **Common response:** An apparent relationship illusion due to a third variable

- **Example:**
  - Ice cream sales
  - Snake bites
  - Sun shine

SPURIOUS CORRELATION



CORRELATION ≠ CAUSATION


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

KEY TERMS FOR EXPERIMENTS

- **Experimental units**
  - The smallest collection of individuals to which treatments are applied
    - Ex: Humans (called subjects), plants, plots of land, animals, etc

- **Factor**
  - The explanatory variables of an experiment
  - Factors have levels
  - *Note: Think categorical variables*

KEY TERMS FOR EXPERIMENTS

- **Treatment**
  - A specific condition applied to the individuals in an experiment which correspond to the levels of the factor

- **Control Group**
  - A treatment group that receives an inactive treatment or an existing baseline treatment

RANDOMIZED COMPARATIVE EXPERIMENTS

One remedy to help control for possible confounding is to perform a ***comparative (controlled) experiment***:

- Experimental units are randomly assignment to each level of the treatment
- Most well-designed experiments compare two or more treatments

PRINCIPLES OF EXPERIMENTAL DESIGN

Randomized experiments are build on four principles:

- **1) Control**
  - (verb) Control for lurking variables that might affect the response, most simply by comparing two or more treatments
  - (noun) May also be referred to as the "non-treatment"

- **2) Randomization**
  - Use chance to assign experimental units to treatments

PRINCIPLES OF EXPERIMENTAL DESIGN

- **3) Replication**
  - Use enough experimental units in each group to reduce chance of variation in the results

- **4) Blocking**
  - The arranging of experimental units in groups (blocks) that are known to be similar to one another
  - Blocking factors is typically a source of variability but not the primary interest
  - <u>Common Examples</u>: Space and time…

CAUSAL CONCLUSIONS

Why can causal conclusions be made from a well-designed randomized experiment?

Part 1: By using a chance mechanism (random number generator) to decide which experimental units get each treatment, a researcher can (hopefully) ensure that all possible confounding variables are thoroughly mixed between the two treatment groups.

CAUSAL CONCLUSIONS

Why can causal conclusions be made from a well-designed randomized experiment?

Part 2: Since the confounding variables ”cancel” each other out between the two groups, researchers can say that what is causing the observed differences in the groups is the treatment being studied

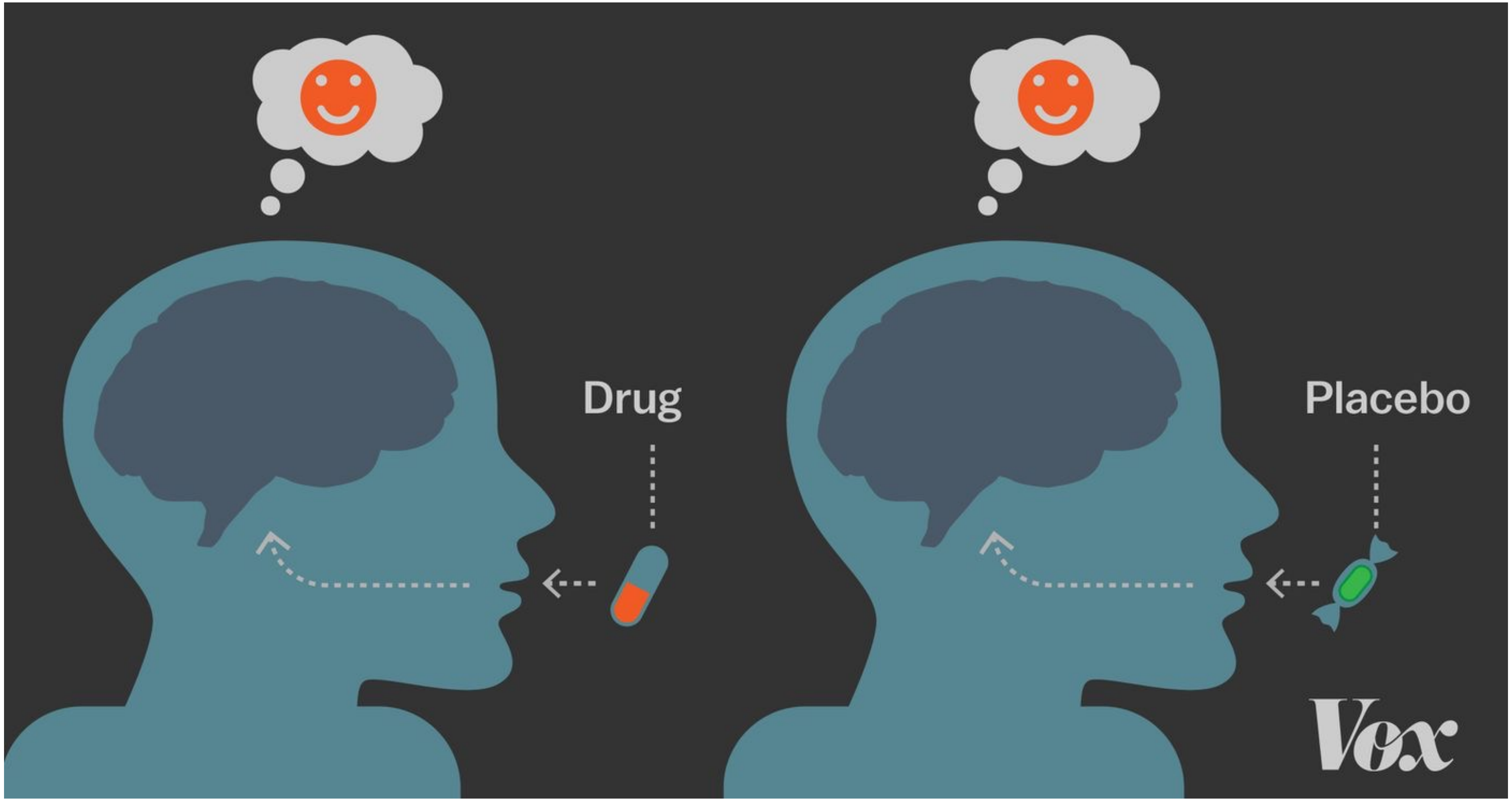
CAUTIONS ABOUT EXPERIMENTATION

- The logic of a randomized comparative experiment depends on out ability to treat all the subjects the same in every way except for the actual treatments being compared

- When any randomized experiment is done on humans, one must be watchful of the ***placebo effect***.

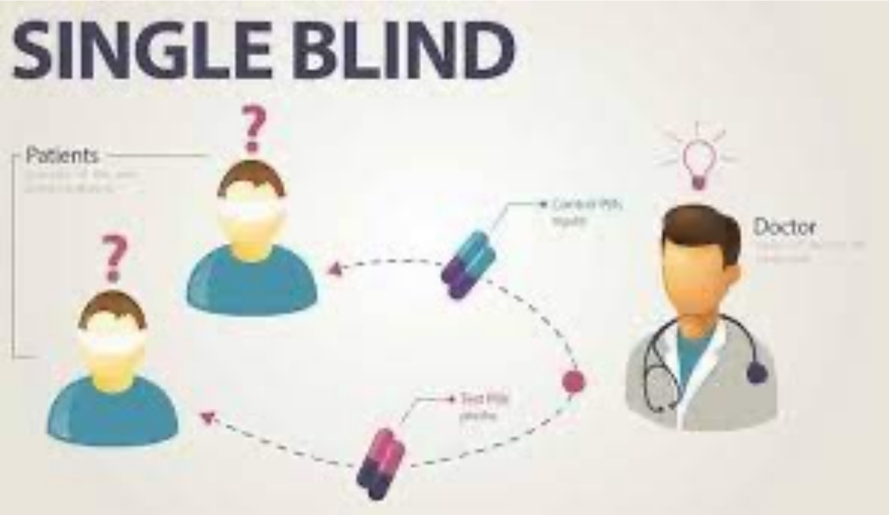
- Sometimes placebo's are called ***sugar pills*** because they do not contain the active drug that is being tested.

CAUTIONS ABOUT EXPERIMENTATION



CAUTIONS ABOUT EXPERIMENTATION

- ***What is the Placebo Effect?***
  - The subject's perceived positive/beneficial effect of the active treatment while being assigned to the control group



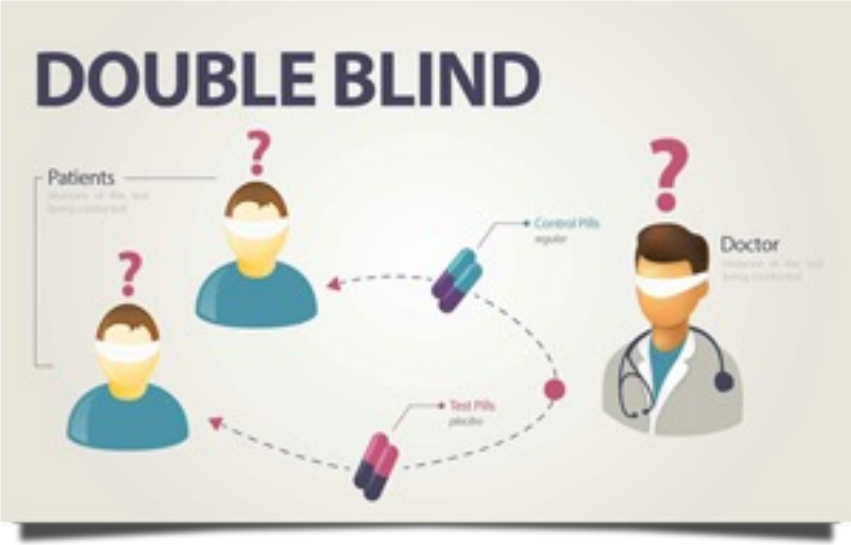
How can you control for the Placebo Effect?

- Incorporating blinding into the experiment
- In a (single) blind experiment, only the researcher (the person measuring the response variable) is aware of which group each participant belongs to but not the subject.

CAUTIONS ABOUT EXPERIMENTATION

- The researcher can potentially include his/her own bias when interacting with the subject. One can account for this by performing a double-blind experiment.

- ***What is a double-blinded experiment?***
  - Neither the researcher not the subject is aware of which group each participant belongs to.



TYPES OF EXPERIMENTAL DESIGNS

1. Completely Randomized Design

2. Randomized Block Design

3. Matched Pairs Design

COMPLETELY RANDOMIZED DESIGN

- Also known as **CRD**

- The simplest experimental design, in terms of analysis and convenience

- Subjects are randomly assigned to treatments

- Typically done by listing treatment levels and randomly assigning random numbers to each

COMPLETELY RANDOMIZED DESIGN

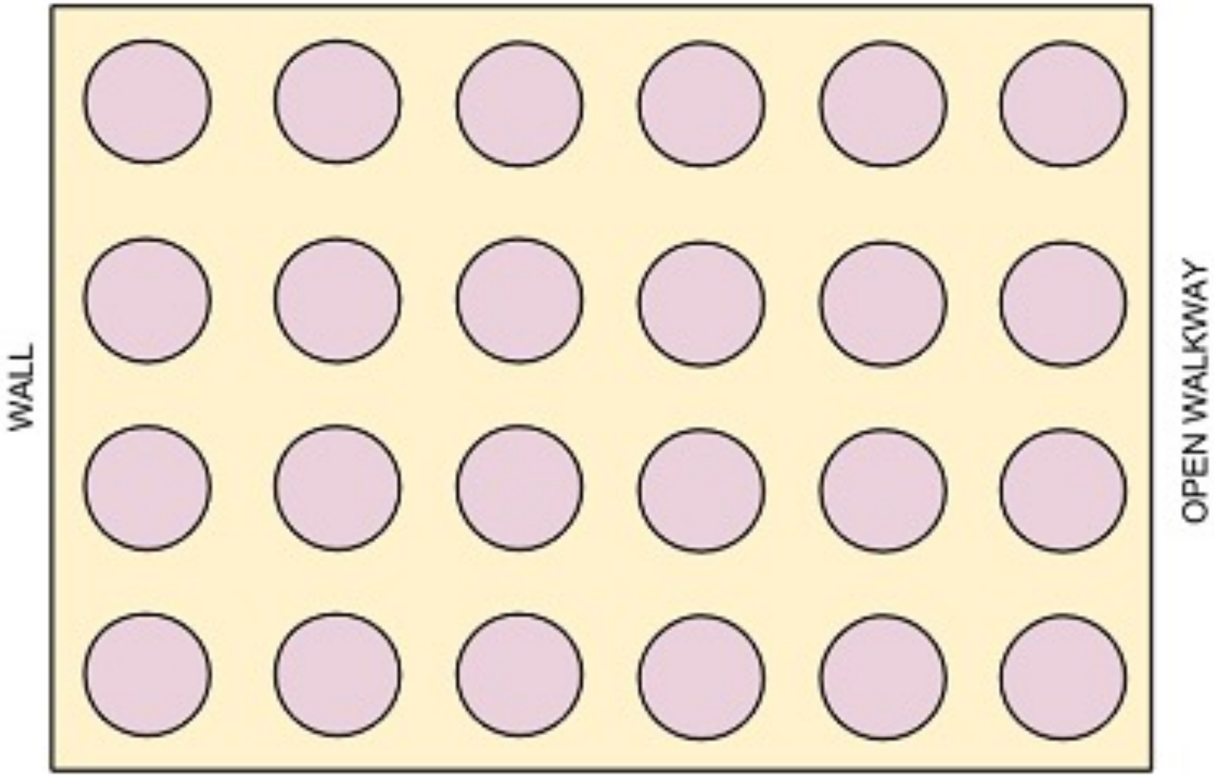
Consider the set up:

- In a greenhouse experiment we want to study a single factors (fertilizer) with 4 levels
- We have enough space for 24 experimental units (a potted plant)
- To maintain balance in the experiment, we will have 6 replications of each treatment

Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

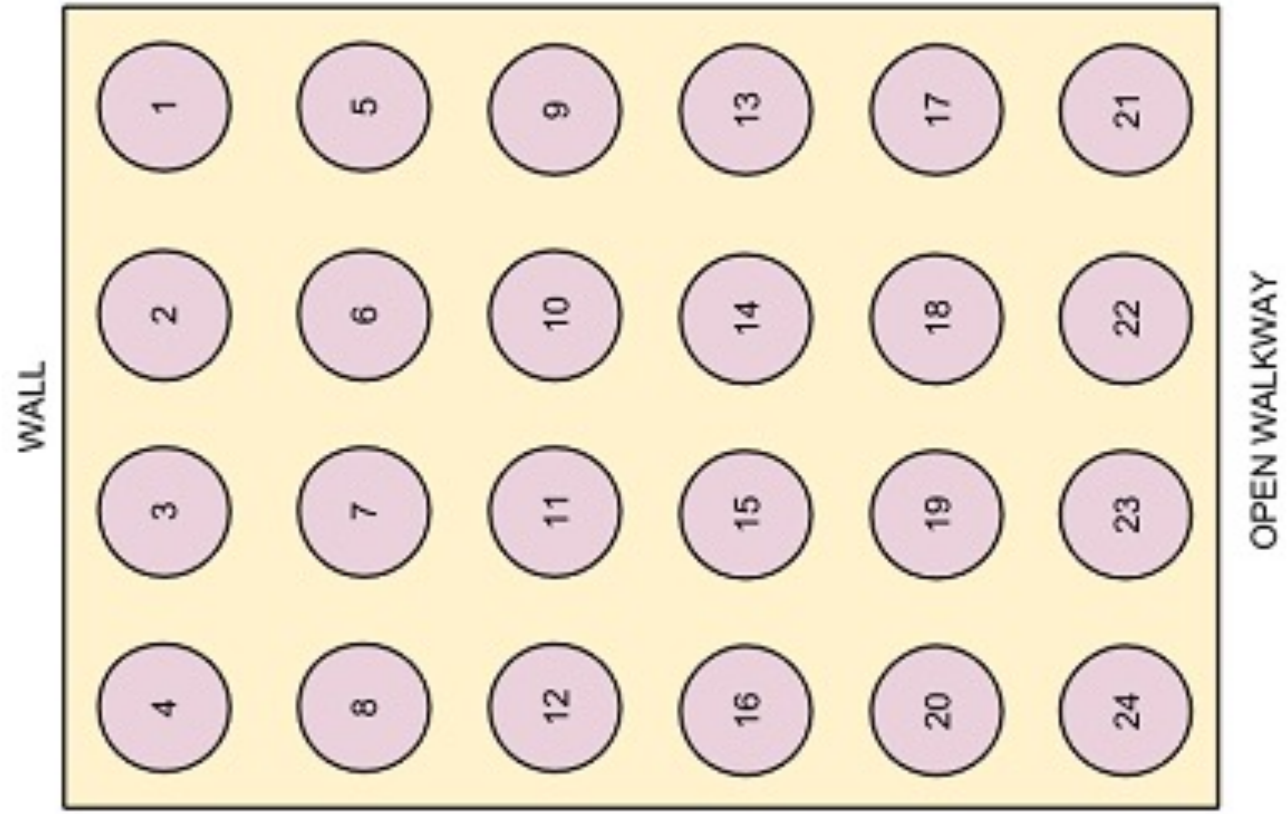
Greenhouse Diagram and bench used for the experiment (viewed from above):



Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

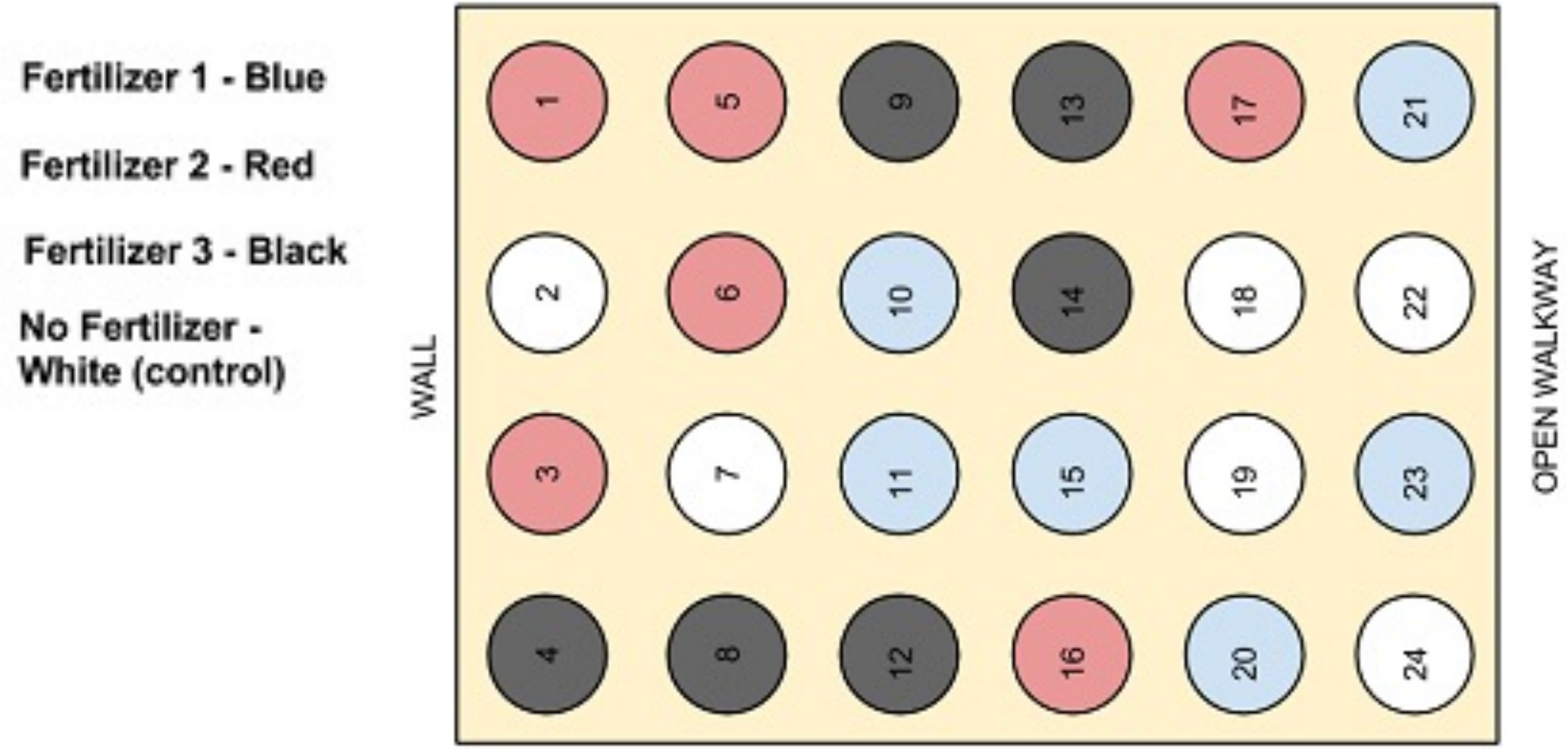
Step 1: Assign it experimental unit a unique id



Source: https://newonlinecourses.science.psu.edu/stat502/node/175/

COMPLETELY RANDOMIZED DESIGN

Step 2: Randomly assign each experimental units to treatments



COMPLETELY RANDOMIZED DESIGN

… ***but what if there are known nutrient gradients across the bench?***

A vanilla CRD will not control for this!

RANDOMIZED (COMPLETE) BLOCK DESIGN

- Also known as **RCBD**
- Variation between blocks is accounted for assigning at least one of each treatment to each block
- Effects of blocks not of interest
- Standard design for agricultural experiments

RANDOMIZED (COMPLETE) BLOCK DESIGN

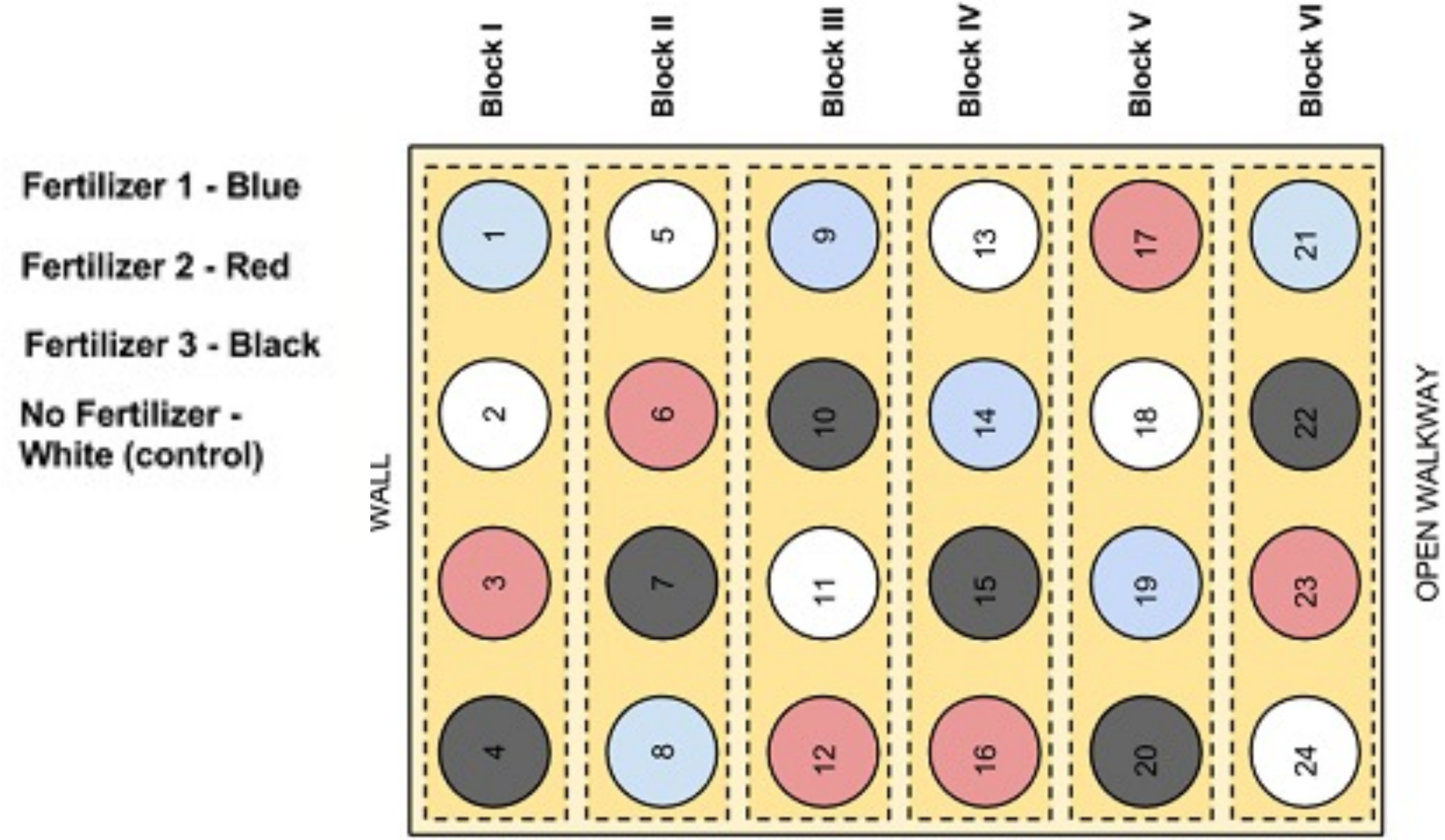
In a block design, the random assignment of experimental units to treatments is carried out within each block

What are the steps in performing a blocked experiment?

1. Form groups (blocks)
    - All individuals within each block should be similar in regard to the lurking variable

2. Within each block, randomly assign experimental units to each treatment.

RANDOMIZED (COMPLETE) BLOCK DESIGN



Fertilizer 1 - Blue

Fertilizer 2 - Red

Fertilizer 3 - Black

No Fertilizer - White (control)

MATCHED PAIRS (AN EXTENSION OF BLOCKING)

- Big Idea: Create blocks by matching pairs of similar experimental units

- Chance is used to determine which unit in each pair gets each treatment

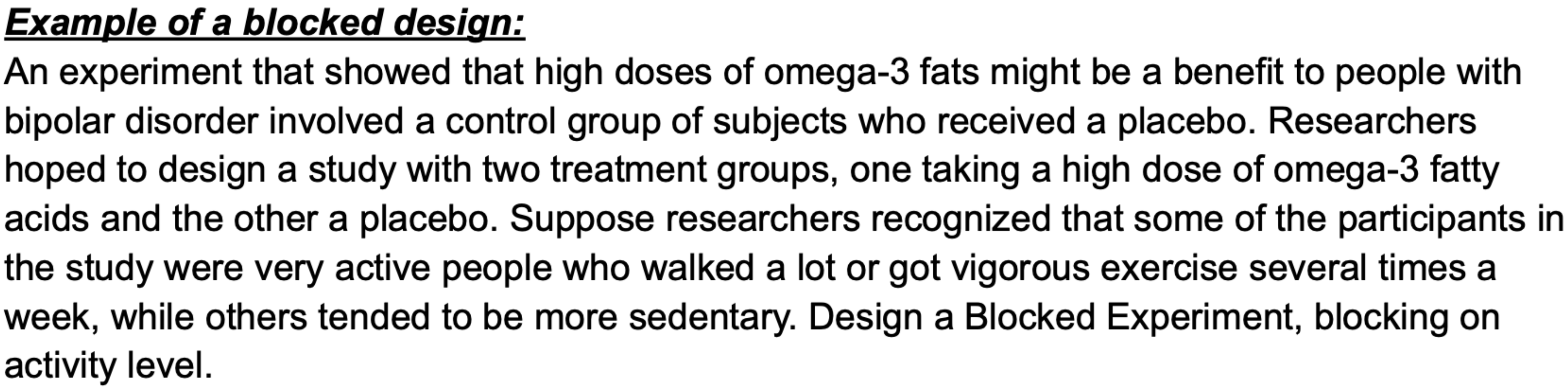
Ex: **Pre-Post (Before After) Studies**

Data from the same individual is related (treat like a block)

1. Assess baseline
2. Assign treatment
3. Find difference after

WORKSHEET EXAMPLES

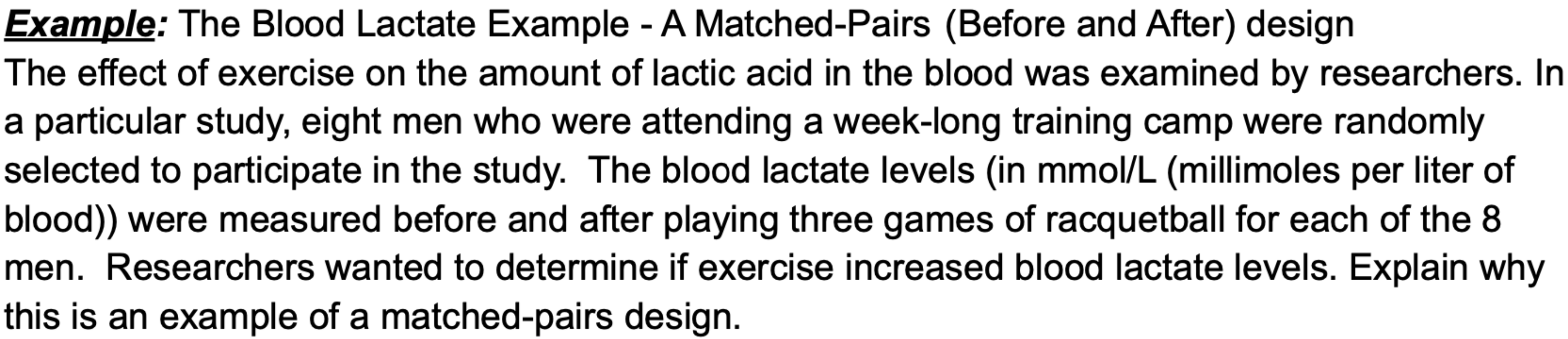
WORKSHEET EXAMPLES



Example of a blocked design:

An experiment that showed that high doses of omega-3 fats might be a benefit to people with bipolar disorder involved a control group of subjects who received a placebo. Researchers hoped to design a study with two treatment groups, one taking a high dose of omega-3 fatty acids and the other a placebo. Suppose researchers recognized that some of the participants in the study were very active people who walked a lot or got vigorous exercise several times a week, while others tended to be more sedentary. Design a Blocked Experiment, blocking on activity level.

WORKSHEET EXAMPLES



Example: The Blood Lactate Example - A Matched-Pairs (Before and After) design
The effect of exercise on the amount of lactic acid in the blood was examined by researchers. In a particular study, eight men who were attending a week-long training camp were randomly selected to participate in the study. The blood lactate levels (in mmol/L (millimoles per liter of blood)) were measured before and after playing three games of racquetball for each of the 8 men. Researchers wanted to determine if exercise increased blood lactate levels. Explain why this is an example of a matched-pairs design.

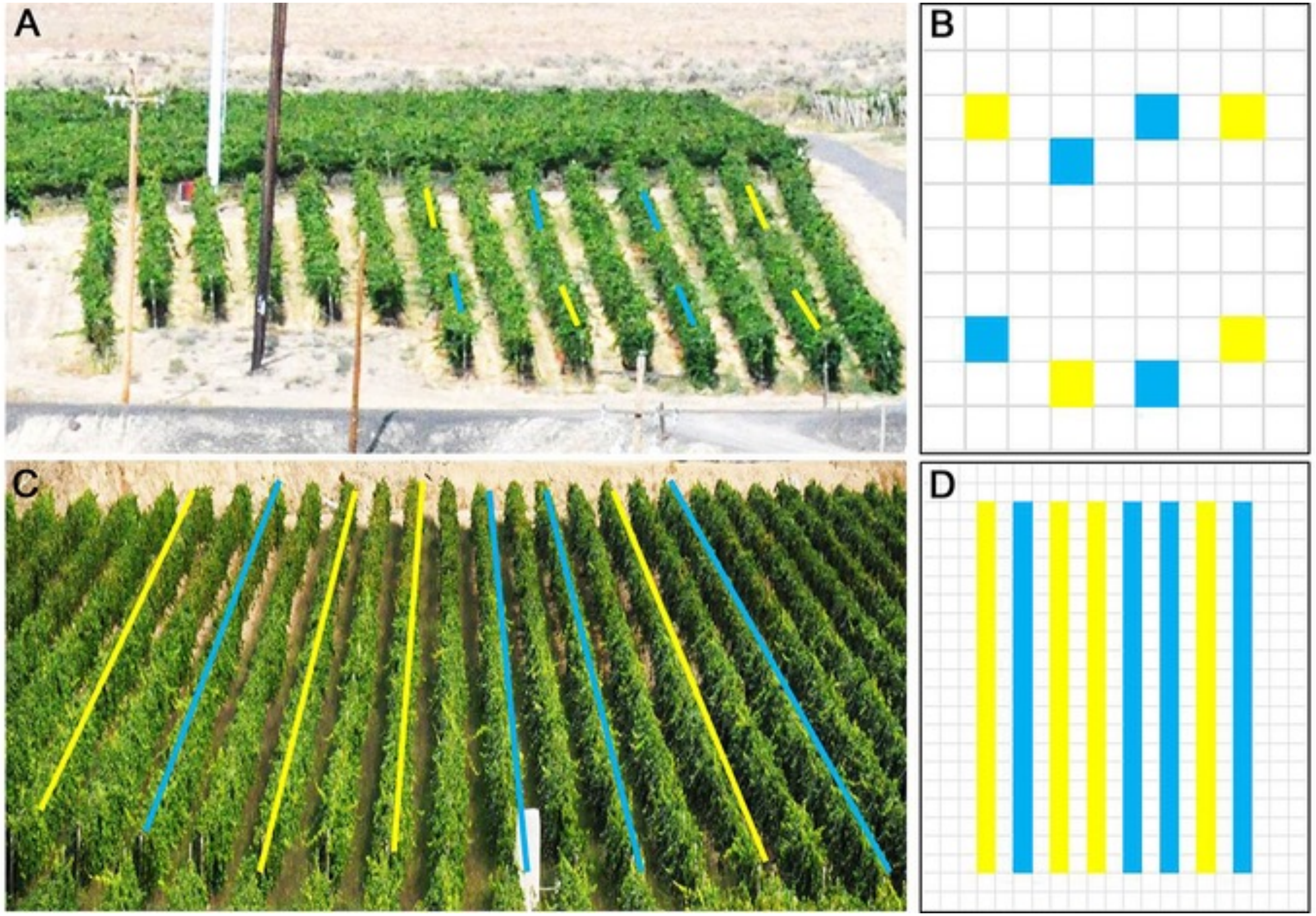
MORE EXAMPLES

EXAMPLE: QUESTION

A study is design to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are represented equally under different conditions. Which of the below is correct?

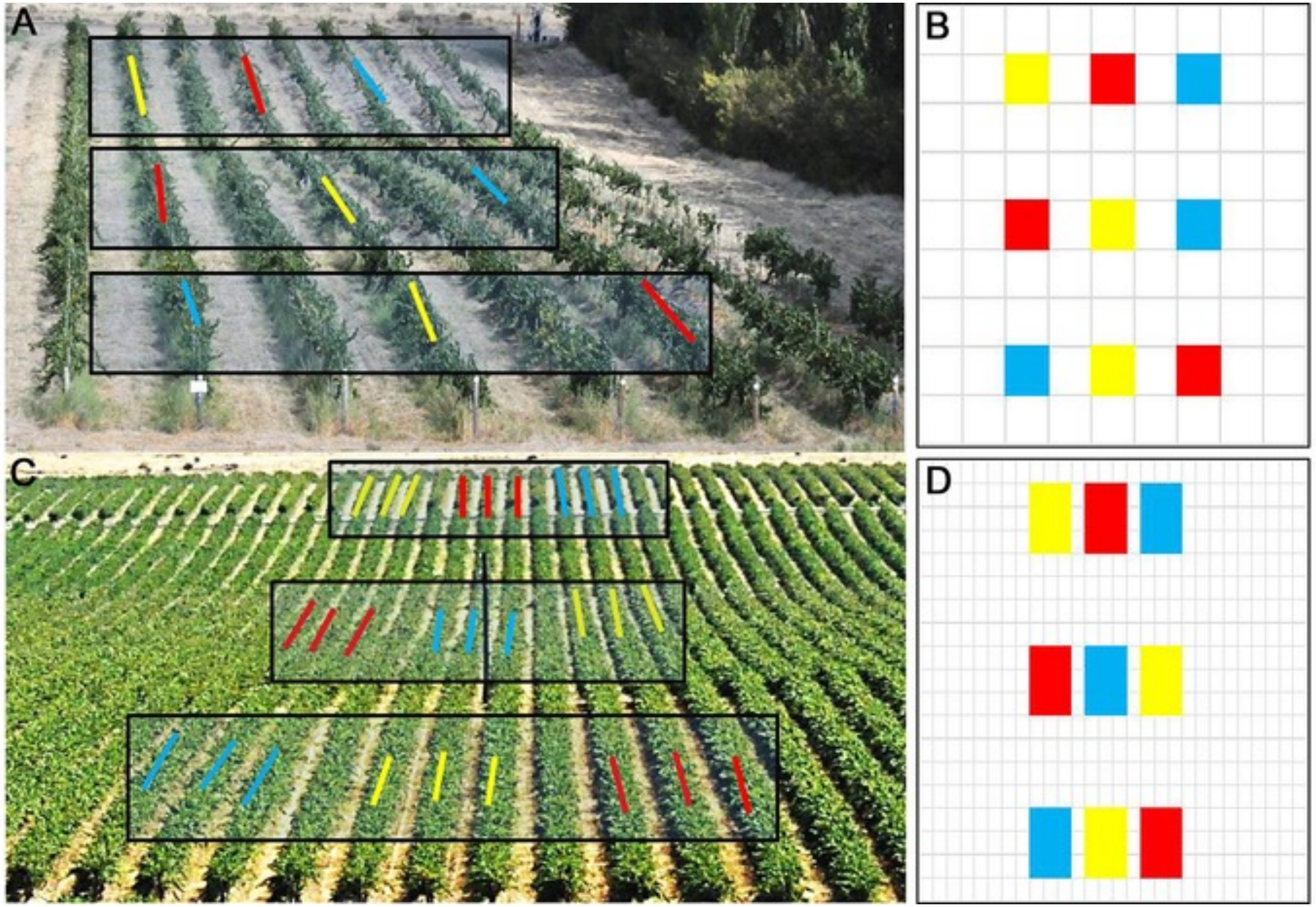
1. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
2. There are 2 explanatory vars (light and noise), 1 blocking var (gender), and 1 response var (exam performace)
3. There is 1 explanatory var (gender) and 3 response vars (light, noise, exam performance)
4. There are 2 blocking vars (light and noise), 1 explanatory var (gender), and 1 response var (exam performance)

VINEYARD STUDIES: CRD



Example of a Completely Randomized Design (CRD) with four replicates of a single treatment and a control in (A) a small and (C) a mid-size vineyard block. Yellow lines and blue lines indicate the control and treatment, respectively. Note: Both examples show buffer rows between treatments. In a CRD, sometimes the same treatment may end up in adjacent rows. Schematic diagrams of the trials shown in A and C are shown in B and D, respectively. Photos and illustrations by Hemant Gohil.

VINEYARD STUDIES: RCB



Example of a Randomized Complete Block (RCB) design with three replicates of two treatments and one control. The black frames in A and C represent the grouping of different zones that was done to account for variability due to slope. In this design, the treatments and control (red, blue, and yellow lines) are randomized within each zone. In the smaller field design (A), two to three vines in a single row might be the replicate unit. In a larger design (C), three to five vines across multiple rows might serve as the replicate unit. Schematic diagrams of the trials shown in A and C are shown in B and D, respectively, where individual cells represent a single vine. Photos and illustrations by Hemant Gohil.