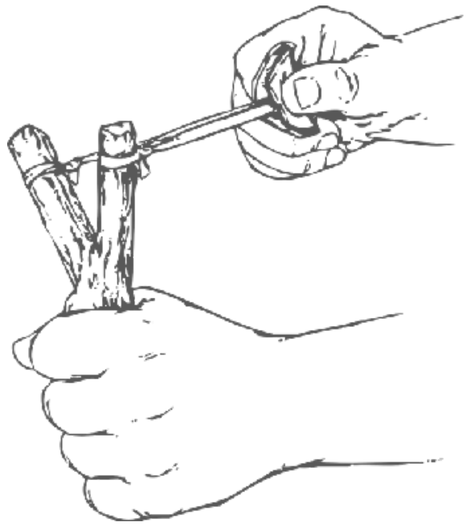
INTRODUCTION TO DATA ANALYSIS

WHAT'S DATA?

LEARNING GOALS



- appreciate the diversity of data
- distinguish different kinds of variables
  - dependent vs independent
  - nominal vs ordinal vs metric
- get familiar with basic aspects of experimental design
  - factorial designs, within- vs between subjects design
  - repeated measures, randomization, fillers and controls

WHAT DOES "DATA" MEAN?

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available
— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

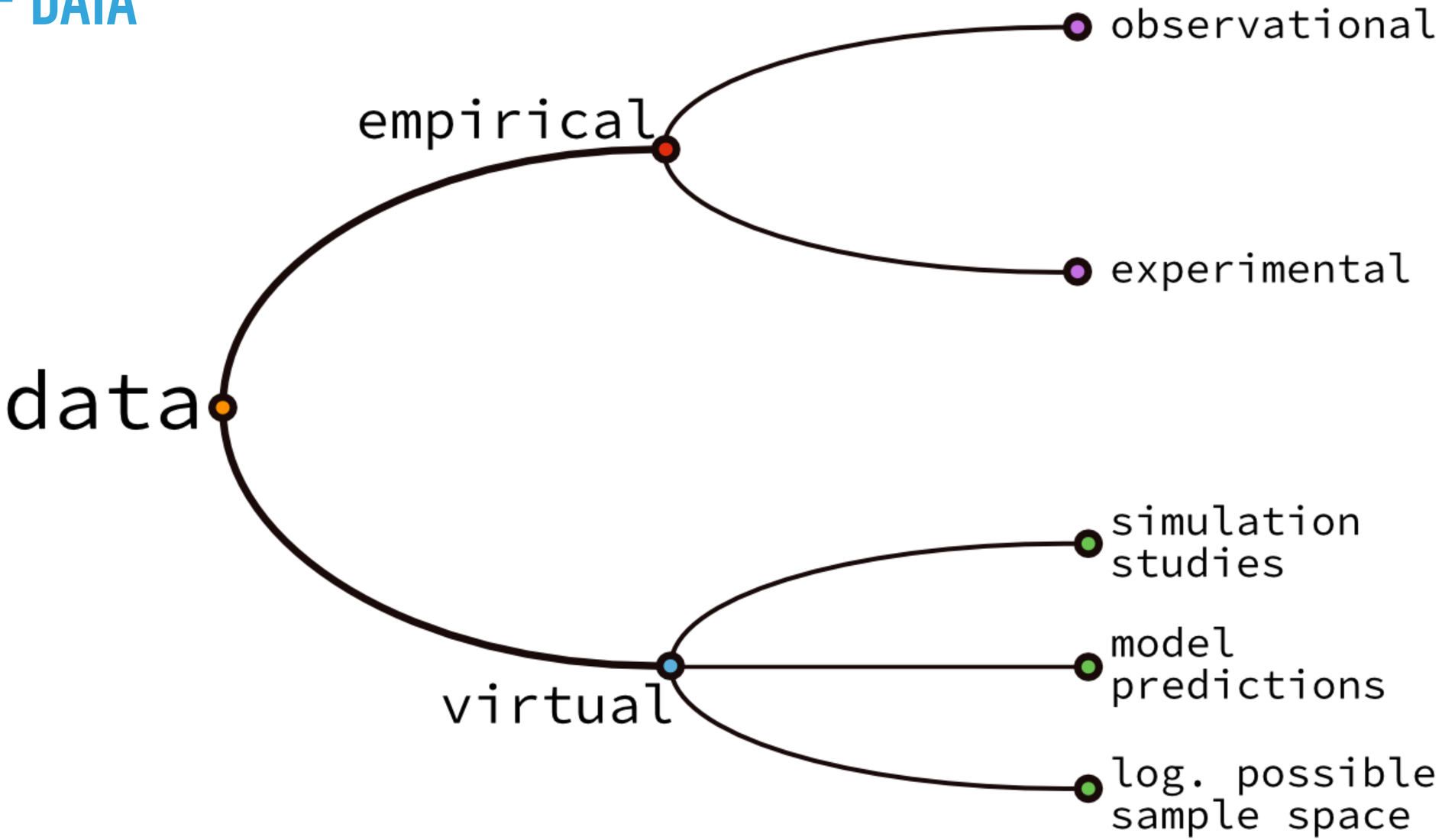
3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Merriam-Webster®

GOALS OF DATA ANALYSIS

- explanation: understand / find the true relation between variables of interest
  - e.g., causal mechanism or correlation
- prediction: accurately predict hitherto unobserved (e.g., future) data points
  - e.g., for medical image classification (tumor recognition)

KINDS OF DATA



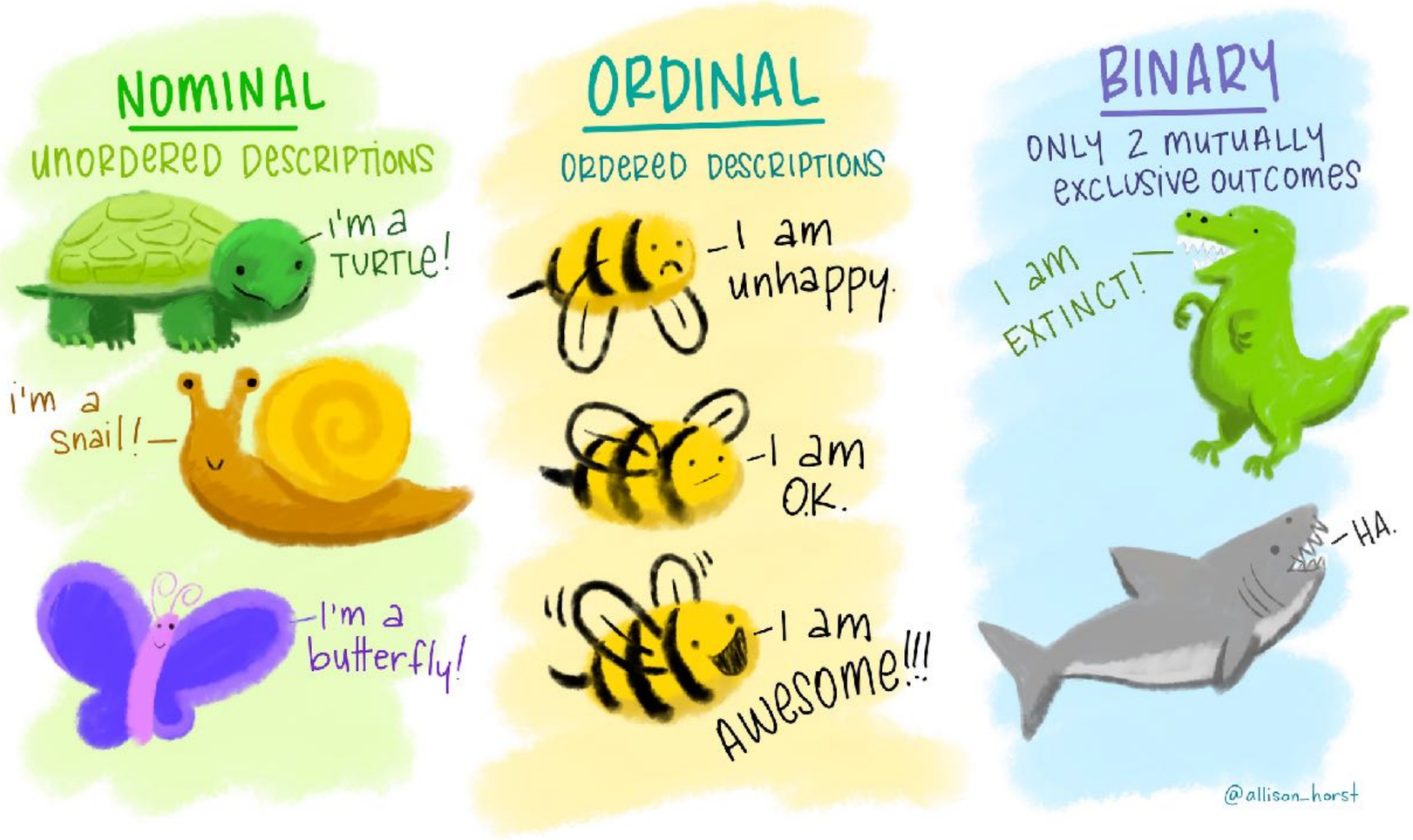
RECTANGULAR DATA

- columns represent variables
- rows are associated observations

```
# proportion of tutorials attended and exam pass/fail
exam_results <-
  tribble(
    ~student,   ~tutorial_proportion,   ~pass,
    "Jax",      0.0,                    TRUE,
    "Jason",    0.78,                   FALSE,
    "Jamie",    0.39,                   TRUE
  )
exam_results
```

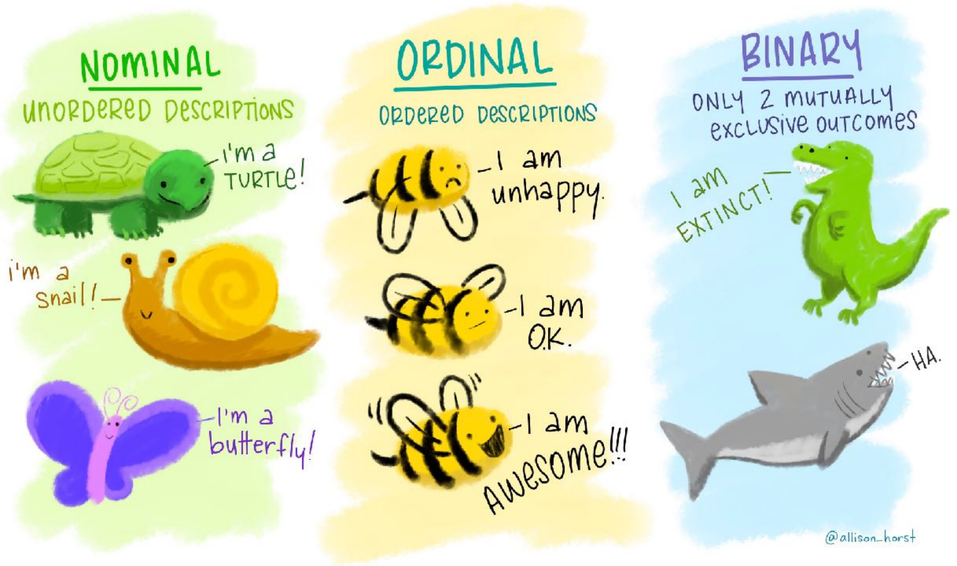
```
## # A tibble: 3 x 3
##   student tutorial_proportion pass
##   <chr>                 <dbl> <lgl>
## 1 Jax                     0    TRUE
## 2 Jason                   0.78 FALSE
## 3 Jamie                   0.39 TRUE
```

KINDS OF VARIABLES



KINDS OF VARIABLES

variable type	representation in R
nominal / binary	unordered factor
Boolean	logical vector
ordinal	ordered factor
metric	numeric vector



DEPENDENT VS INDEPENDENT VARIABLES

- dependent variables represent data we want to explain / predict
  - ♢ dep. variable ≠ what's measured
- independent variables represent data we want to use as explanans / conditional information based on which to make predictions
- distinction is entirely purpose-driven

```
# A tibble: 3 x 3
  maker  price consumption
  <chr>  <dbl>       <dbl>
1 Audi   43900         7.2
2 Volvo  61350         6.8
3 Toyota 34290         5.3
```

It's not possible to say which of these variables has to be (for logical reasons) a dependent or independent variable. That depends on the goal of explanation/prediction.

EXPERIMENTAL DATA

- experimental data typically has:
  - at least one dependent variable
  - at least one independent variable
  - some association of observations between variables

```
tribble(
  ~subj_id,     ~group,        ~systolic,
  1,            "treatment",   118,
  2,            "control",     132,
  3,            "control",     116,
  4,            "treatment",   127,
  5,            "treatment",   122
)
```

```
## # A tibble: 5 x 3
##   subj_id group     systolic
##     <dbl> <chr>        <dbl>
## 1       1 treatment      118
## 2       2 control        132
## 3       3 control        116
## 4       4 treatment      127
## 5       5 treatment      122
```

FACTORIAL DESIGN

- if all independent variables are at most ordinal in nature, we have a factorial design
- a 2x3 factorial design has:
  - two factors
  - one with two levels
  - another one with three levels
- a 2x3 factorial design has 6=2*3 experimental conditions (= design cells)

```
tribble(
  ~subj_id,     ~group,        ~systolic,
  1,            "treatment",   118,
  2,            "control",     132,
  3,            "control",     116,
  4,            "treatment",   127,
  5,            "treatment",   122
)
```

```
## # A tibble: 5 x 3
##   subj_id group     systolic
##     <dbl> <chr>        <dbl>
## 1       1 treatment      118
## 2       2 control        132
## 3       3 control        116
## 4       4 treatment      127
## 5       5 treatment      122
```

WITHIN- & BETWEEN-SUBJECTS DESIGNS

- within-subjects design: every participant contributes at least one observation to each experimental condition
- between-subjects design: not every participant contributes data to each experimental condition

```
tribble(
  ~subj_id,     ~group,        ~systolic,
  1,            "treatment",   118,
  2,            "control",     132,
  3,            "control",     116,
  4,            "treatment",   127,
  5,            "treatment",   122
)
```

```
## # A tibble: 5 x 3
##   subj_id group     systolic
##     <dbl> <chr>        <dbl>
## 1       1 treatment      118
## 2       2 control        132
## 3       3 control        116
## 4       4 treatment      127
## 5       5 treatment      122
```

WITHIN- & BETWEEN-SUBJECTS DESIGNS

- within-subjects design: every participant contributes at least one observation to each experimental condition
- between-subjects design: not every participant contributes data to each experimental condition

```
tribble(
  ~subj_id,     ~group,        ~systolic,
  1,            "treatment",   118,
  2,            "control",     132,
  3,            "control",     116,
  4,            "treatment",   127,
  5,            "treatment",   122
)
```

Example of a between-subject design.

Different designs have different pro's and cons's

between-subjects	within-subjects
no confound between conditions	possible cross-contamination between conditions
more participants needed	fewer participants needed
less associated information for analysis	more associated data for analysis

REPEATED MEASURES

- single-shot experiment: every participant contributes exactly one data point to exactly one experimental condition
- repeated measures: every participant contributes more than one observation to at least one experimental condition
  - repetition can lead to data contamination
  - calls for fillers, randomization and item-variability

```
tribble(
  ~subj_id,     ~group,         ~systolic,
  1,            "treatment",    118,
  2,            "control",      132,
  3,            "control",      116,
  4,            "treatment",    127,
  5,            "treatment",    122
)
```

This is a single-shot experiment.

TYPES OF TRIALS

- critical: belongs to an experimental condition
- filler: used to introduce variance, disguise experimental purpose, avoid repetition etc.
- control: used to check whether participants paid attention, understood the task, etc.

SAMPLE SIZE

- how many observations does a study need for each experimental condition?
- answer depends on goals of statistical analysis
  - power-calculation, error control, etc.

HOMEWORK

- read Chapter 3
- work on HW1
  - to be submitted next Friday before noon
  - released later today
  - see course website & email announcement