What is Data?
Part 1: Definitions and Types

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

August 24, 2016

Prof. Michael Paul

Prof. William Aspray

Overview

This lecture will…

- first introduce some definitions,
- then show some examples of data types and how to describe them mathematically,
- and then preview how to do this in practice, using the MiniTab Express software.

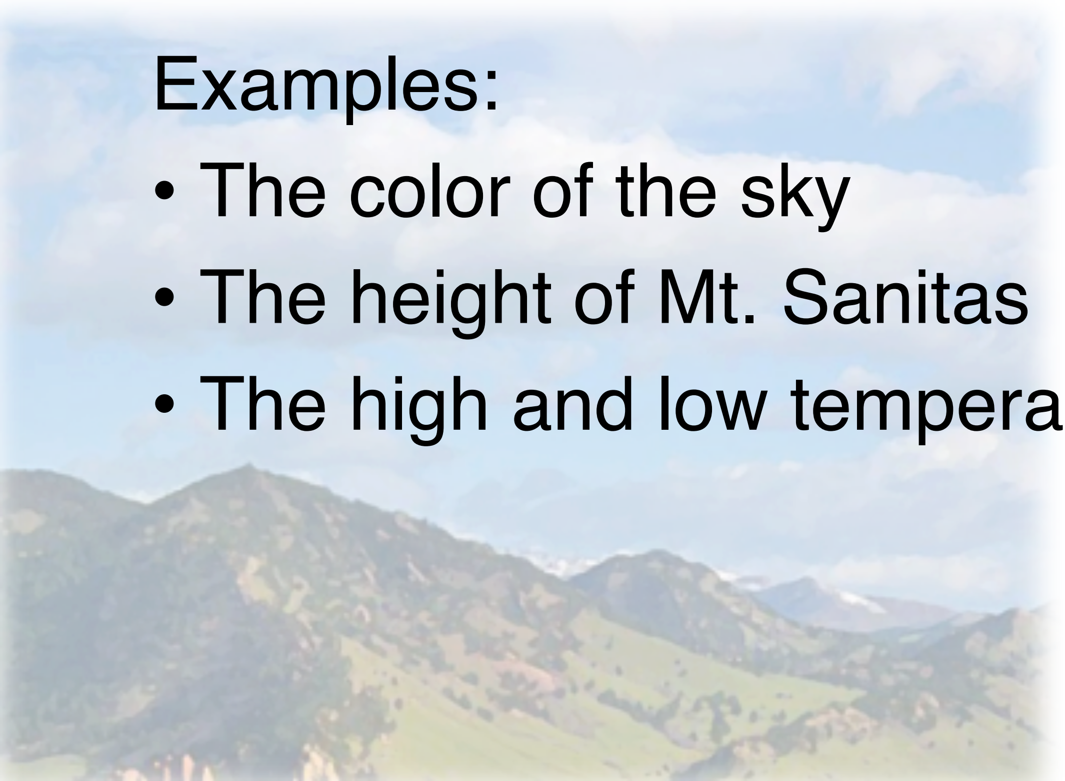
What is data?

Loosely:
Observation(s) about the world

Note on grammar:

Historically:
data = plural
datum = singular

Common today:
data = singular (and sometimes plural)



Examples:

- The color of the sky
- The height of Mt. Sanitas
- The high and low temperatures yesterday

What is a statistic?

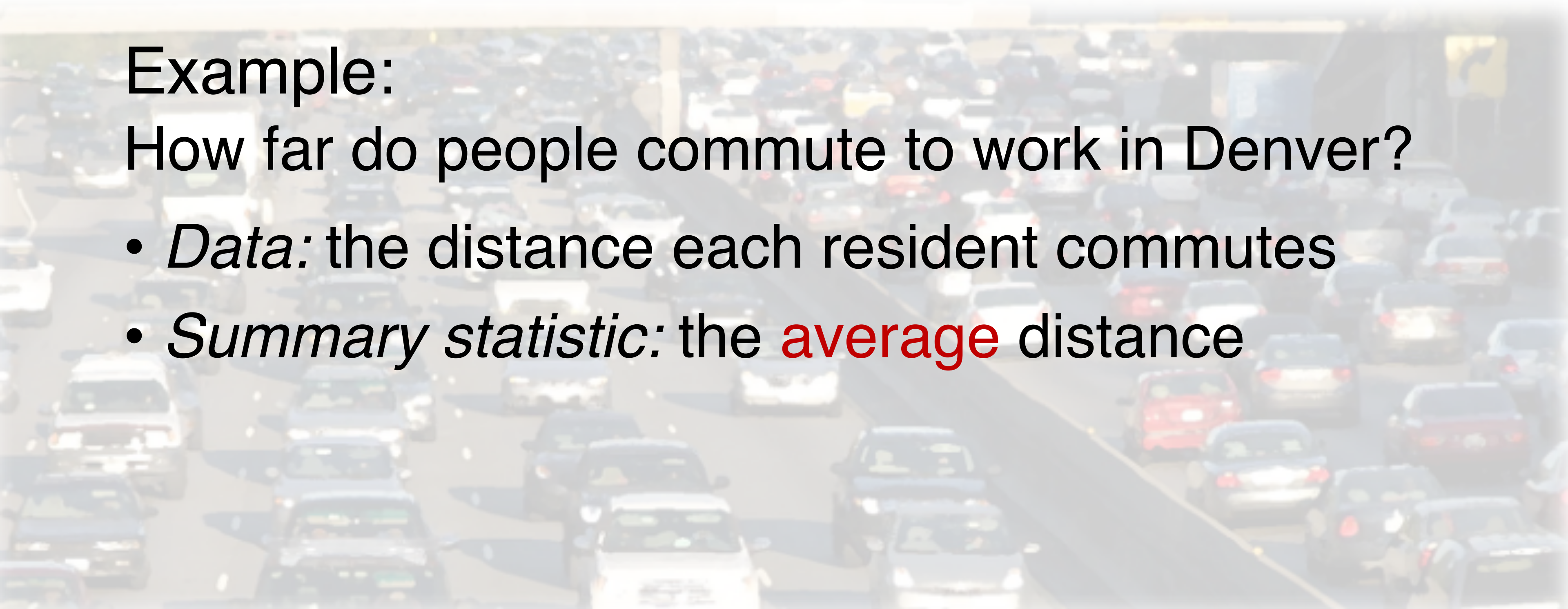
A statistic is a value computed from data

A **summary statistic** summarizes many pieces of data with a concise number

What is a statistic?

A statistic is a value computed from data

A **summary statistic** summarizes many pieces of data with a concise number



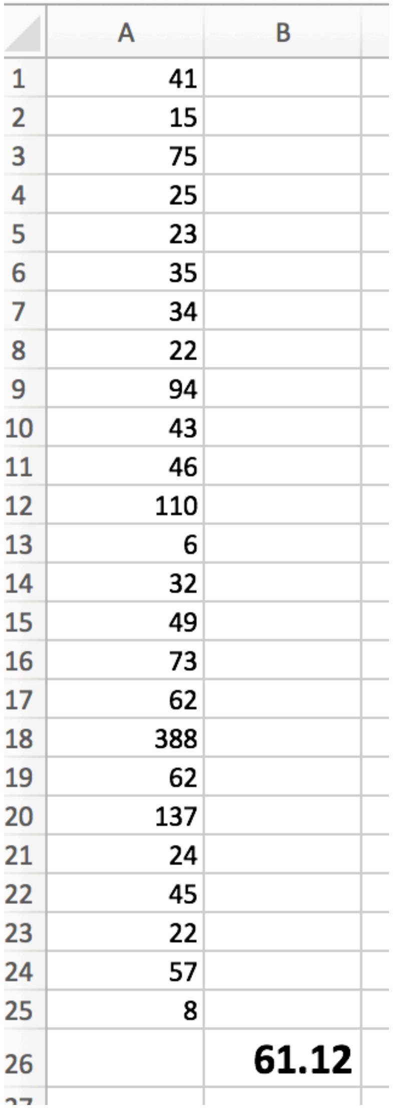
Example:

How far do people commute to work in Denver?

- *Data:* the distance each resident commutes
- *Summary statistic:* the average distance

What is a statistic?

Data values:



	A	B
1	41	
2	15	
3	75	
4	25	
5	23	
6	35	
7	34	
8	22	
9	94	
10	43	
11	46	
12	110	
13	6	
14	32	
15	49	
16	73	
17	62	
18	388	
19	62	
20	137	
21	24	
22	45	
23	22	
24	57	
25	8	
26		**61.12**

Average:

It can be hard to make sense of many different values

Summary statistics allow us to understand the general pattern

It's not practical to compute statistics by hand!

That's why we use software in this course.

Data vs information

Data is usually considered the smallest "piece"

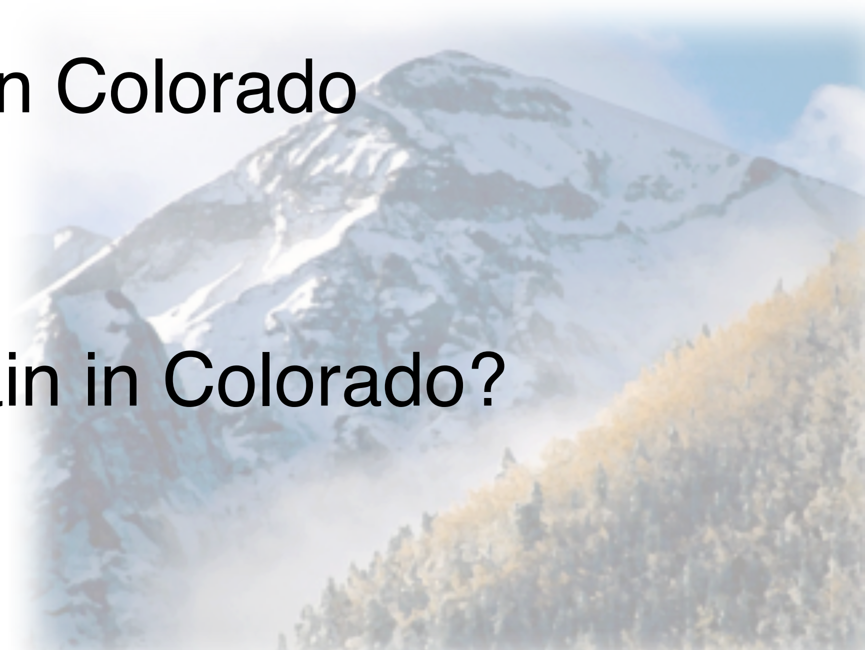
Pieces of data can combine to form information

Example of **data:**

- Height of each mountain in Colorado

Example of **information:**

- What is the tallest mountain in Colorado?



Other phrases to know

Big data

Data mining

Data science

Other phrases to know

Big data

- Very large amounts of data (usually more than can fit on one computer)

Newer technology makes it easier to use big data, so more companies are taking advantage of it

Other phrases to know

Big data

Examples of big data:

- Amazon has billions of transaction records
- Google has trillions of search query logs

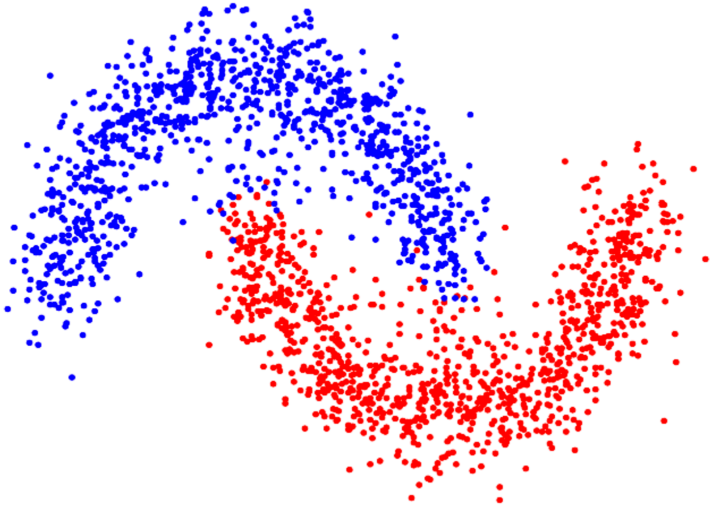
These companies can find interesting patterns in their data to improve their products

Other phrases to know

Data mining

The science and process of discovering patterns in data

- Related to data science, but has its own history within computer science



Other phrases to know

Data science

The science and process of extracting information, knowledge, and insights from data

This field includes:

- Data analysis
- Statistics
- Visualization

This course (along with INFO-2301) will teach the foundations of data science

Other phrases to know

Data science

How is data science different from information science?

- **Data science is part of information science,** but information science is broader and includes the study of how information is and should be used

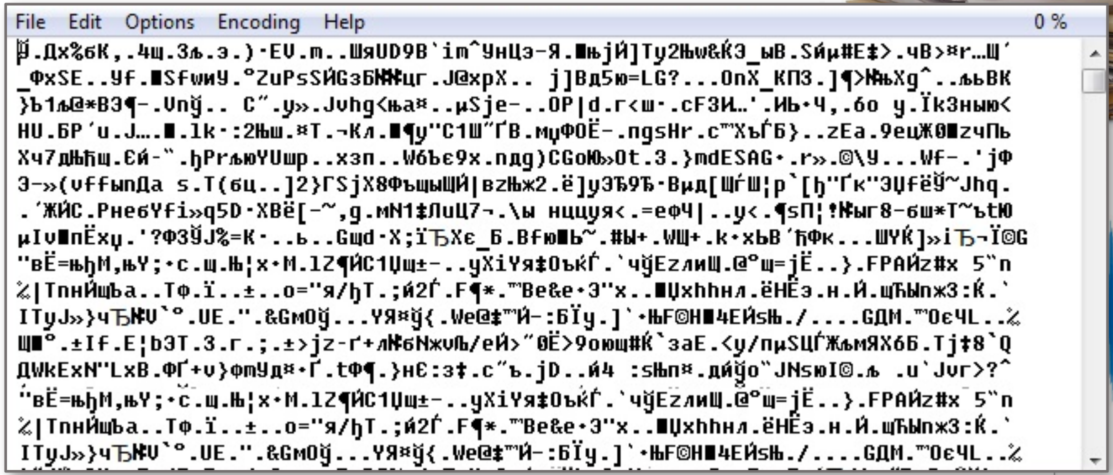
Pause

Questions at this point?

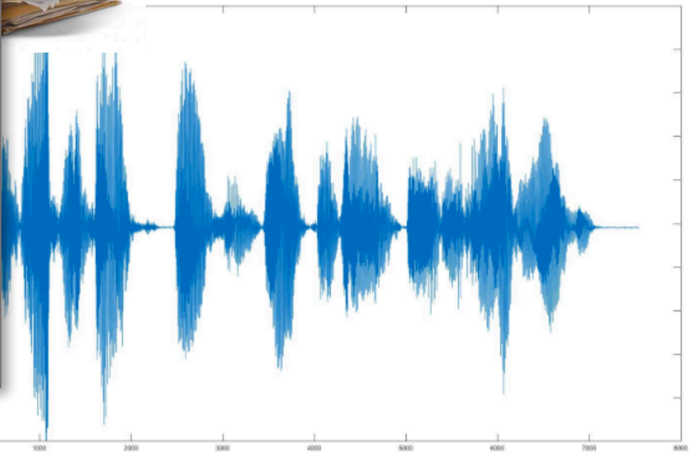
What does data look like?

Data comes in many forms

- Some forms are more useful than others



```
ᵽ.Дх%бК,.4щ.3ѧ.э.)·EU.m..ШяUD9B`im^УнЦэ-Я.■њjЙ]Ty2Њw&Ќ3_ыB.Sйμ#E‡>.чB>¤r…Щ'
_ФxSE..Уf.■SfwиУ.°ZuPsSЙGзБ№№цr.J@xpX.. j]Bд5ю=LG?...OnX_КПЗ.]¶>№љXg^..љьBK
}Ъ1љ@*B3¶-.Unў.. C".y».Jvhg<њa¤..μSje-..OP|d.r<ш·.cF3И…'.Иь·Ч,.бo y.ÏkЗныю<
HU.БP'u.J….■.lk·:2Њш.¤T.¬Кл.■¶y"C1Ш"ЃB.мyФOЁ-.ngsHr.c™ХъЃБ}..zEa.9eцЖ0■zчПь
Хч7дЊћщ.Єй-".ђPrљюYUшр..хзп..Wбъє9x.nдg)CGoЮ»Ot.3.}mdESAG·.r».©\У...Wf-.'jФ
Э-»(vffыnДa s.T(бц..]2}ЃSjX8ФъщыЩЙ|вzЊж2.ё]yЭЂ9Ђ·Bμд[ЩѓШ¦p`[ђ"Ѓк"ЭЏfёЎ~Jhq.
.'ЖЙC.PнeбYfi»q5D·XBё[-~,g.мN1‡ЛuЦ7¬.\ы нццyя<.=eфЧ|..y<.¶sП¦‼Ѓыr8-бш*T~ъtЮ
μIv■пЁxy.'?Ф3ЎJ%=К·..ь..Gшd·X;ïЂXє_Б.Bfю■ь~.#Ы+.WЩ+.k·xьB'ћФк...ШYЌ]»iЂ¬Ï©G
"вЁ=њђM,њY;·c.щ.Њ¦x·M.lZ¶ЙC1Џщ±-..yXiYя‡0ъкЃ.`чўEzлиЩ.@°щ=jЁ..}.FPAЙz#x 5"n
%|TпнЙщЪa..Tф.ï..±..o="я/ђT.;й2Ѓ.F¶*.™Be&e·Э"x..■Џxhhнл.ёHЁэ.н.Й.щЋЫnж3:Ќ.`
ITyJ»}чЂ№U`°.UE.".&GмOў...YЯ¤ў{.We@‡™Й-:БЇy.]`·ЊF©H■4EЙsЊ./....GДM.™0єЧL..%
Щ■°.±If.E¦bЭT.3.r.;.±>jz-r+л№бNжvЊ/eЙ>"0Ё>9oющ#Ќ`заE.<y/пμSЦЃЖљмЯХбБ.Tj‡8`Q
ДWkExN"LxB.ФЃ+v}фmУд¤·Ѓ.tФ¶.}нЄ:з‡.c"ъ.jD..й4 :sЊп¤.дйўo"JNsюI©.ѧ .u`Jvr>?^
"вЁ=њђM,њY;·c.щ.Њ¦x·M.lZ¶ЙC1Џщ±-..yXiYя‡0ъкЃ.`чўEzлиЩ.@°щ=jЁ..}.FPAЙz#x 5"n
%|TпнЙщЪa..Tф.ï..±..o="я/ђT.;й2Ѓ.F¶*.™Be&e·Э"x..■Џxhhнл.ёHЁэ.н.Й.щЋЫnж3:Ќ.`
ITyJ»}чЂ№U`°.UE.".&GмOў...YЯ¤ў{.We@‡™Й-:БЇy.]`·ЊF©H■4EЙsЊ./....GДM.™0єЧL..%
```



Data processing

The process of modifying and organizing data for analysis is called **data processing**

Data before processing is called **raw data**

Representing data

Note on grammar:
The plural of matrix is **matrices**

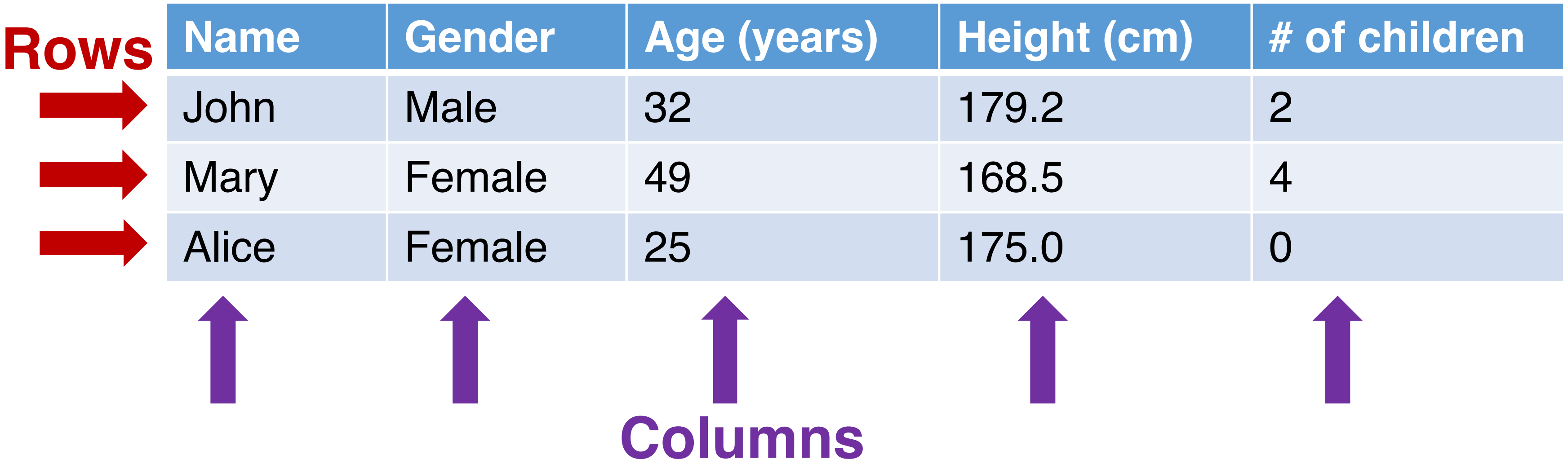
A common way of representing and organizing data is with a **data matrix:**

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

- Also called a **table**
- Equivalent to a **spreadsheet** (e.g., Microsoft Excel)

Representing data

A common way of representing and organizing data is with a **data matrix:**



Rows

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Columns

Representing data

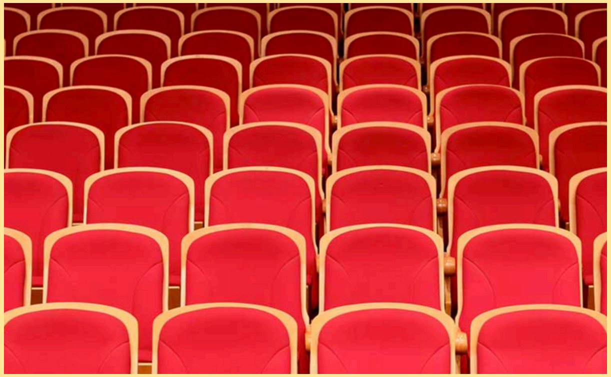
A common way of representing and organizing data is with a **data matrix:**

Rows

Name	Gender	Age (years)	Height (cm)	# of children
Joh				
Ma				
Alic				

How to remember which is which:

Rows:



Columns:

Representing data

A common way of representing and organizing data is with a **data matrix:**

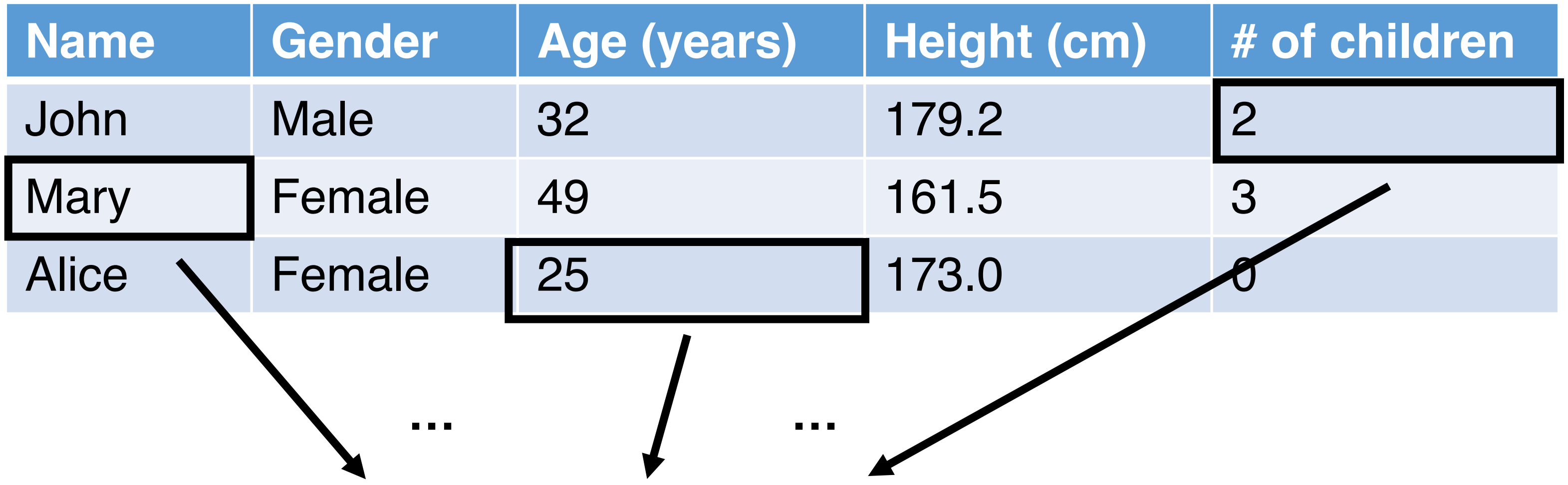
Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

This top row is the **header** row, which describes the columns

- We don’t count this as part of the data

Representing data

A common way of representing and organizing data is with a **data matrix:**



Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

Each box is called a **cell**

Representing data: variables

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Each column is a **variable**

- Also called an **attribute**

Representing data: variables

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Example: The 2nd column is the gender **variable**

Representing data: variables

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Example: The 2nd column is the gender variable

- The cell in the header row is the **name** of the variable

Representing data: variables

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Example: The 2nd column is the gender variable

- The cell in the header row is the **name** of the variable
- The cells in the 3 data row are the variable **values**

Representing data: observations

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

Each row is an **observation** (or **observational unit**)

- Also called a **case**
- Also called an **instance**

Representing data: observations

How do we interpret the matrix?

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	168.5	4
Alice	Female	25	175.0	0

The 1st row is an observation of a person named John

- Every observation has values for the 5 variables

Where does data come from?

Data tables don’t simply exist in the universe waiting to be discovered.

People have to create data!

People have to make choices about:

- What variables to include and how to define them
- What values the variables can take and how to measure them

Be aware that these choices can affect how the data is interpreted! (we’ll discuss this next week)

Pause

Questions at this point?

Types of variables

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

Categorical variables

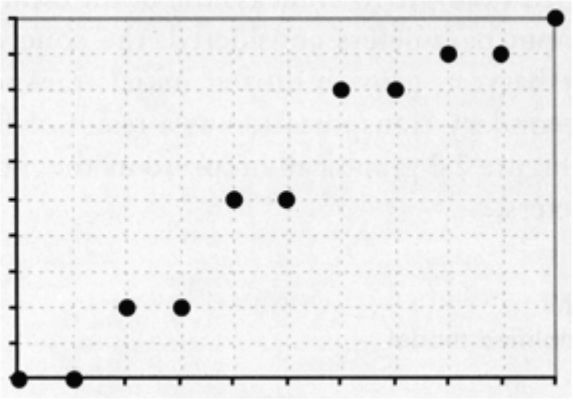
Numerical variables

Types of variables: numerical

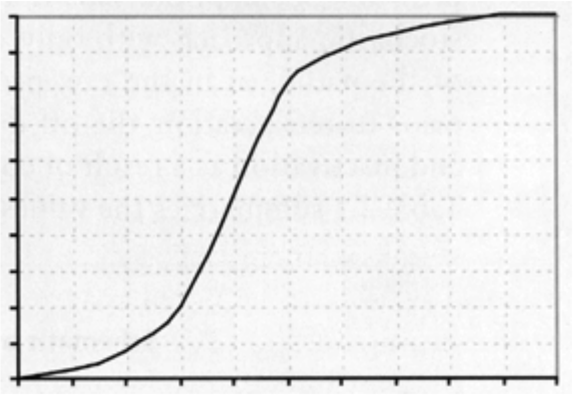
Numerical variables have a range or set of numbers as possible values

- Numerical variables can either be **discrete** or **continuous**

Discrete values have separation between them; they can be counted

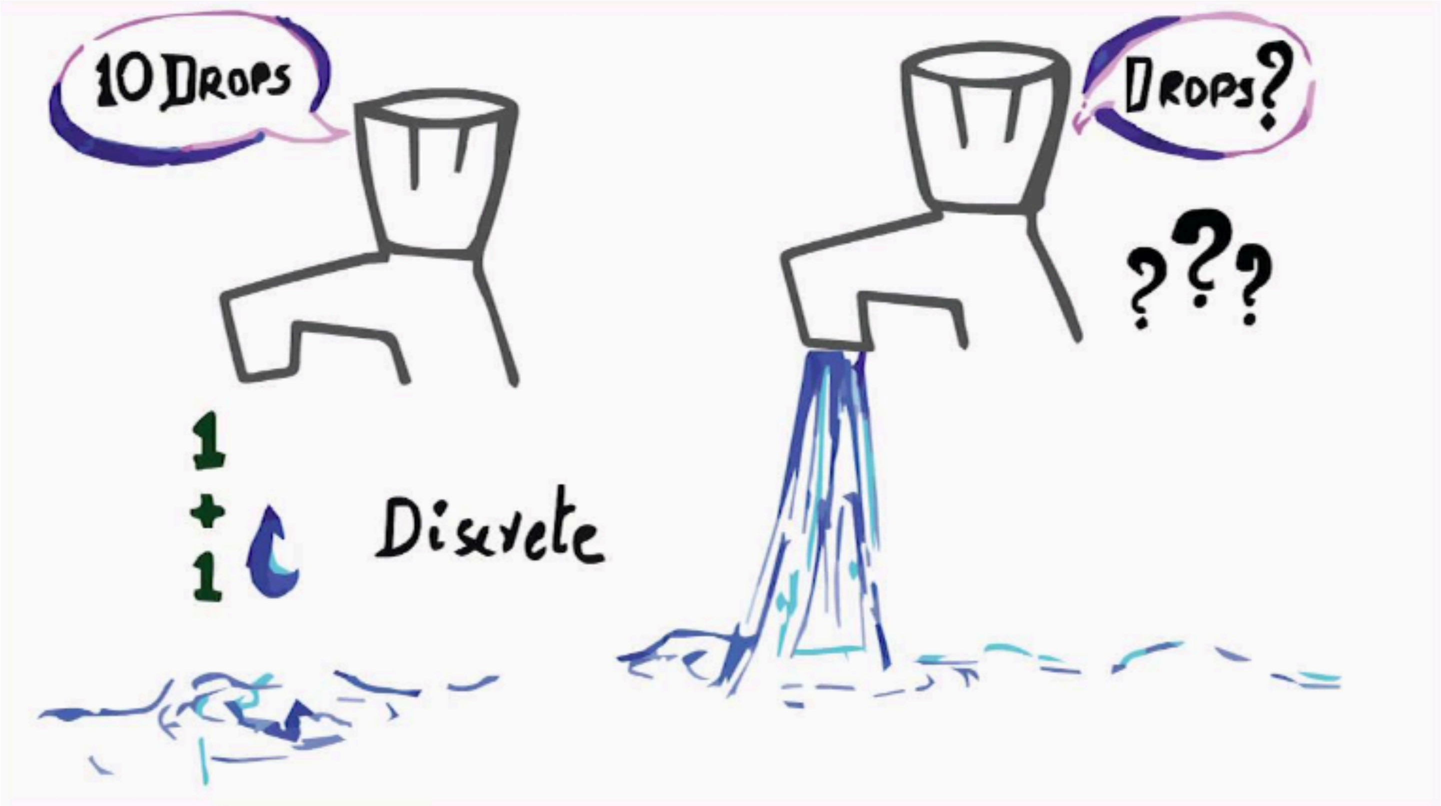


Continuous values can be plotted as a smooth line without gaps; a spectrum



Types of variables: numerical

Discrete vs continuous: can it be counted?



From: TAPtheTECH, https://www.youtube.com/watch?v=WX0hnuniLpI

Types of variables: numerical

Discrete examples:

- The number of people in this room
- The number of hairs on your head

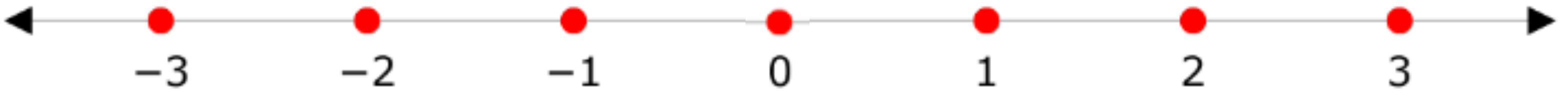
Continuous examples:

- The loudness of sound
- The brightness of light
- The passage of time

Types of variables: numerical

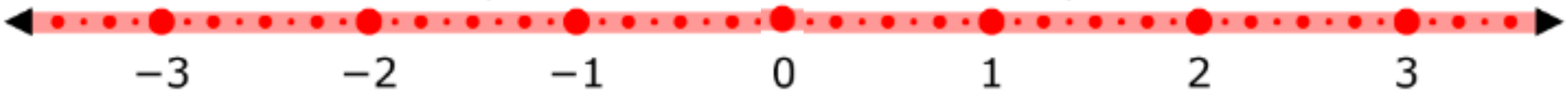
Discrete examples:

- **Integers** (also called **whole numbers,** but can be negative too)



Continuous examples:

- **Real numbers**



Types of variables: numerical

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

Continuous **Discrete**

Both

- Time passed since birth is continuous
- Number of years since birth is discrete

Types of variables: categorical

Categorical variables have a set of categories they can take as values

- Names instead of numbers

Examples of categorical values:

- Colors of paint
- Brands of cola
- Breeds of dogs

All categorical values are also discrete

Types of variables: categorical

Categorical variables can also be divided as **ordinal** and **nominal** variables

Ordinal categories have some type of ordering

- Example:
  small → medium → large

Note:
Numerical values are also ordinal

Nominal categories include everything else

- We mostly won't make the distinction between ordinal and nominal categories, but it can be useful to be aware of

Types of variables: categorical

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

Categorical variables

- Name and gender are both **nominal** (not ordered)

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

There are usually additional rules for what values a variable can have beyond numbers vs categories

The set of values a variable can take is called the **domain** of the variable

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

⬆

What is the **domain** of the *name* variable?

- Any text

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

⬆

What is the **domain** of the *gender* variable?

- A set of valid options:
  - Agender
  - Cis Female
  - Cis Male
  - Transgender Female
  - …

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

What is the **domain** of the *age* variable?

- Any positive number (greater than zero)
  - Or any positive integer if we define it as whole years

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

What is the **domain** of the *height* variable?

- Any positive number (greater than zero)

Types of variables: domain

Pay attention to what values the variables can have:

Name	Gender	Age (years)	Height (cm)	# of children
John	Male	32	179.2	2
Mary	Female	49	161.5	3
Alice	Female	25	173.0	0

What is the **domain** of the *children* variable?

- Any positive integer (including zero)

Types of variables: domain

A domain is defined by a **set**

A set is a collection of values

- We'll define sets mathematically next week

Examples:

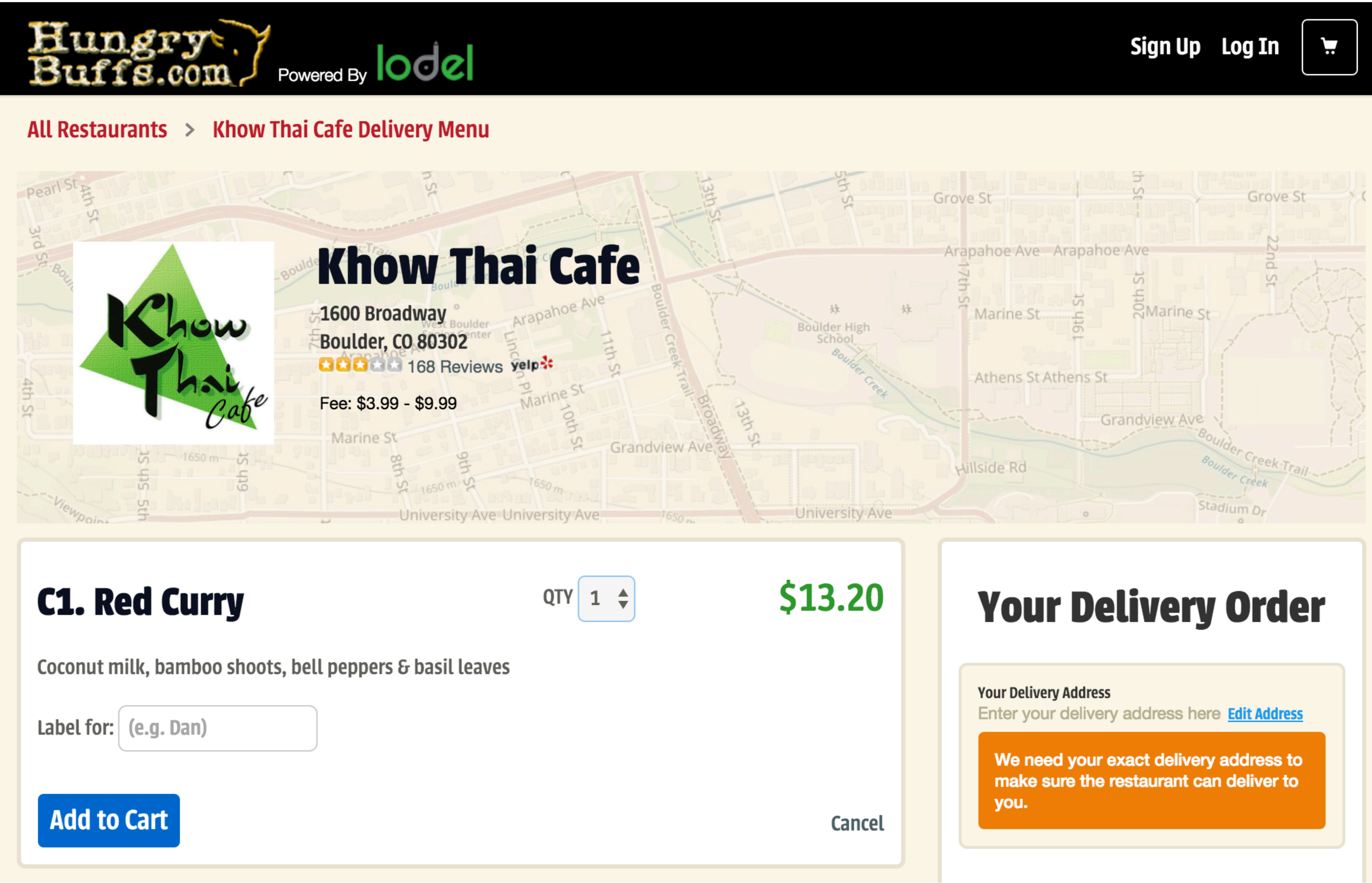
- Set of genders
- Set of dog breeds
- Set of integers
- Set of real numbers
- Set of positive real numbers

Other terminology

The textbook calls the possible values **levels,** but note that this term only applies to categorical values.

Data is everywhere: a silly example

- What are the attributes of Thai curry?



All Restaurants > Khow Thai Cafe Delivery Menu

Khow Thai Cafe

1600 Broadway
Boulder, CO 80302

★★★☆☆ 168 Reviews yelp

Fee: $3.99 - $9.99

C1. Red Curry

QTY 1

$13.20

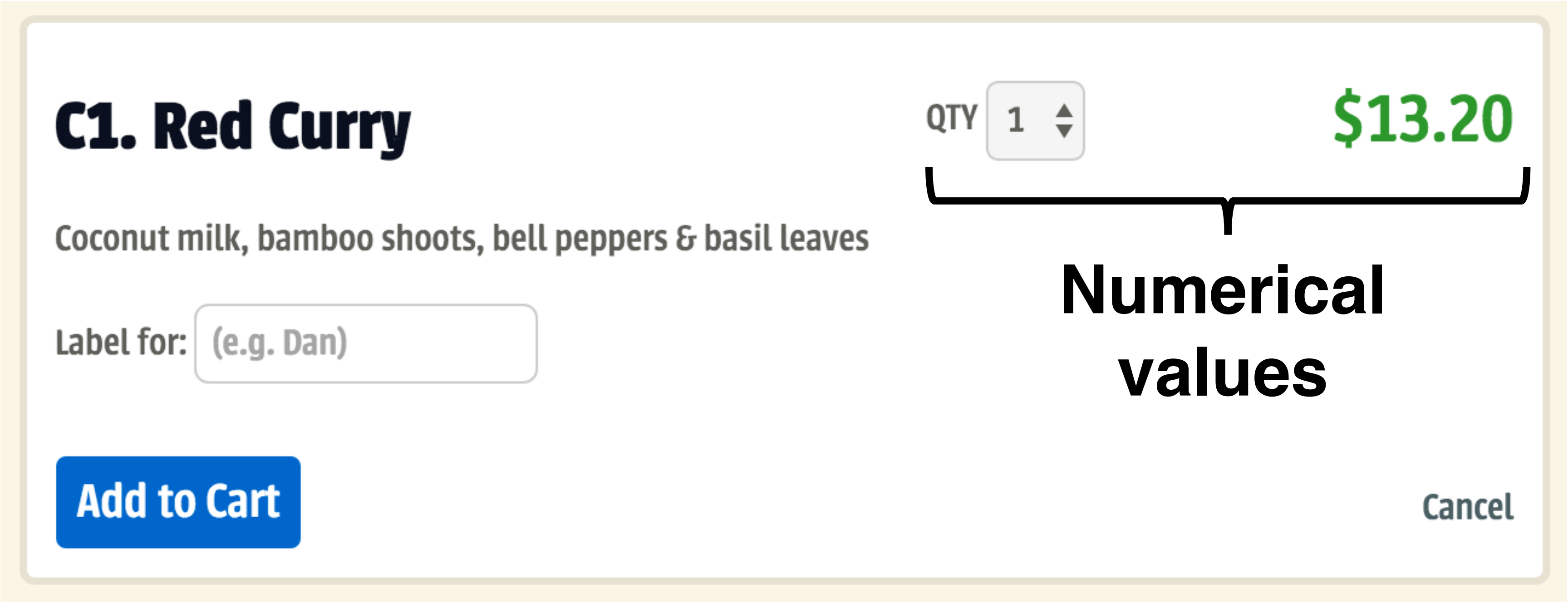
Coconut milk, bamboo shoots, bell peppers & basil leaves

Label for: (e.g. Dan)

Your Delivery Order

Your Delivery Address
Enter your delivery address here Edit Address

We need your exact delivery address to make sure the restaurant can deliver to you.



C1. Red Curry

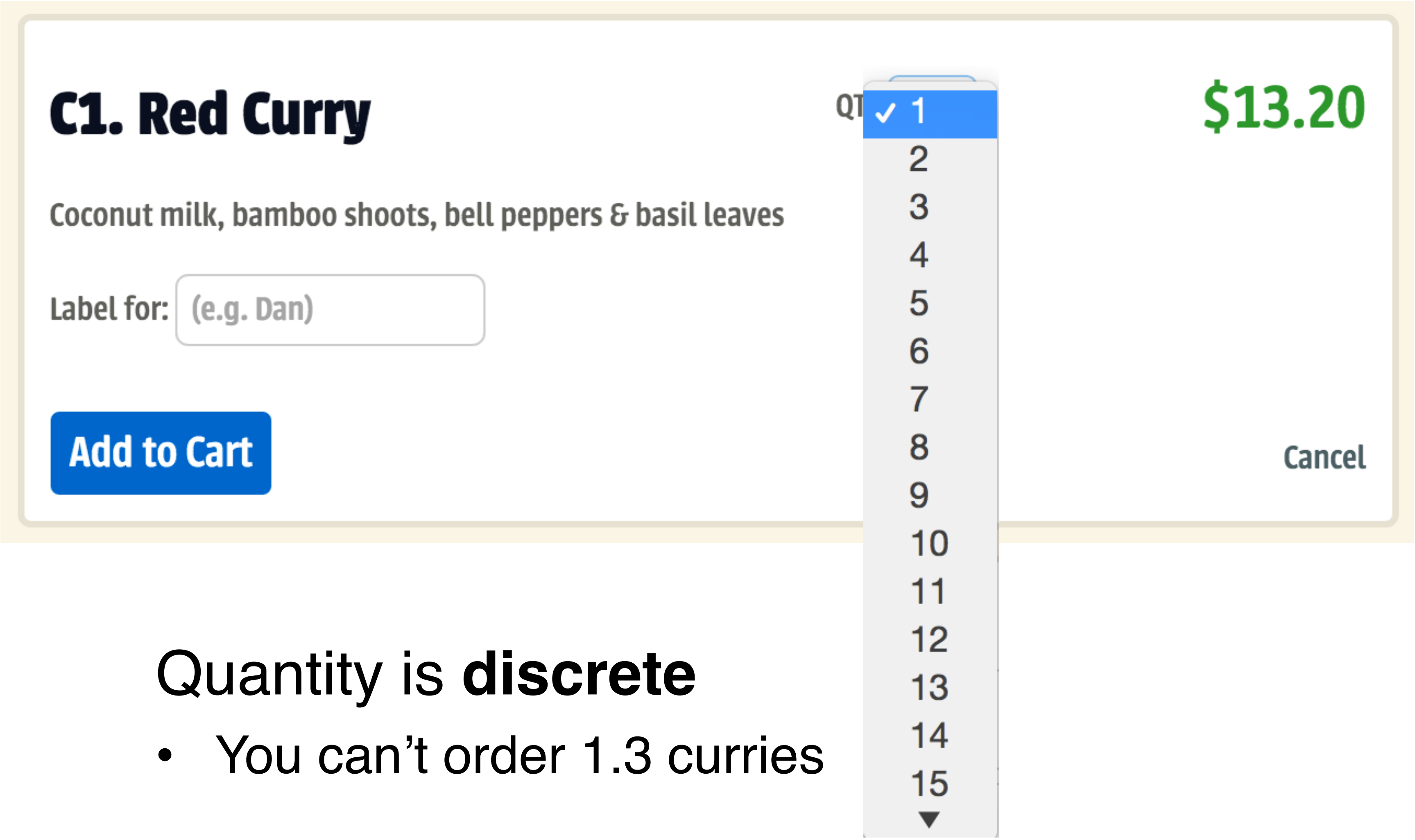
Coconut milk, bamboo shoots, bell peppers & basil leaves

Label for: (e.g. Dan)

QTY 1

$13.20

Numerical values



C1. Red Curry

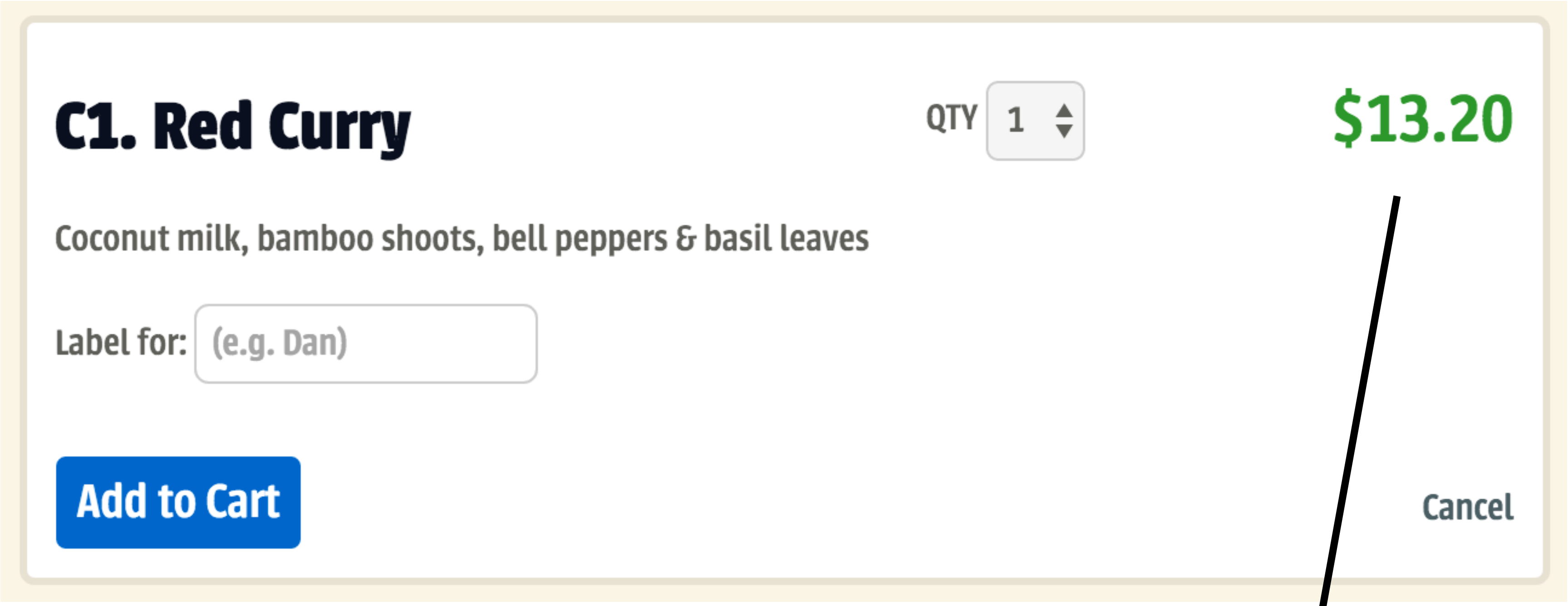
$13.20

Coconut milk, bamboo shoots, bell peppers & basil leaves

Label for: (e.g. Dan)

Quantity is **discrete**

- You can’t order 1.3 curries



C1. Red Curry

QTY 1

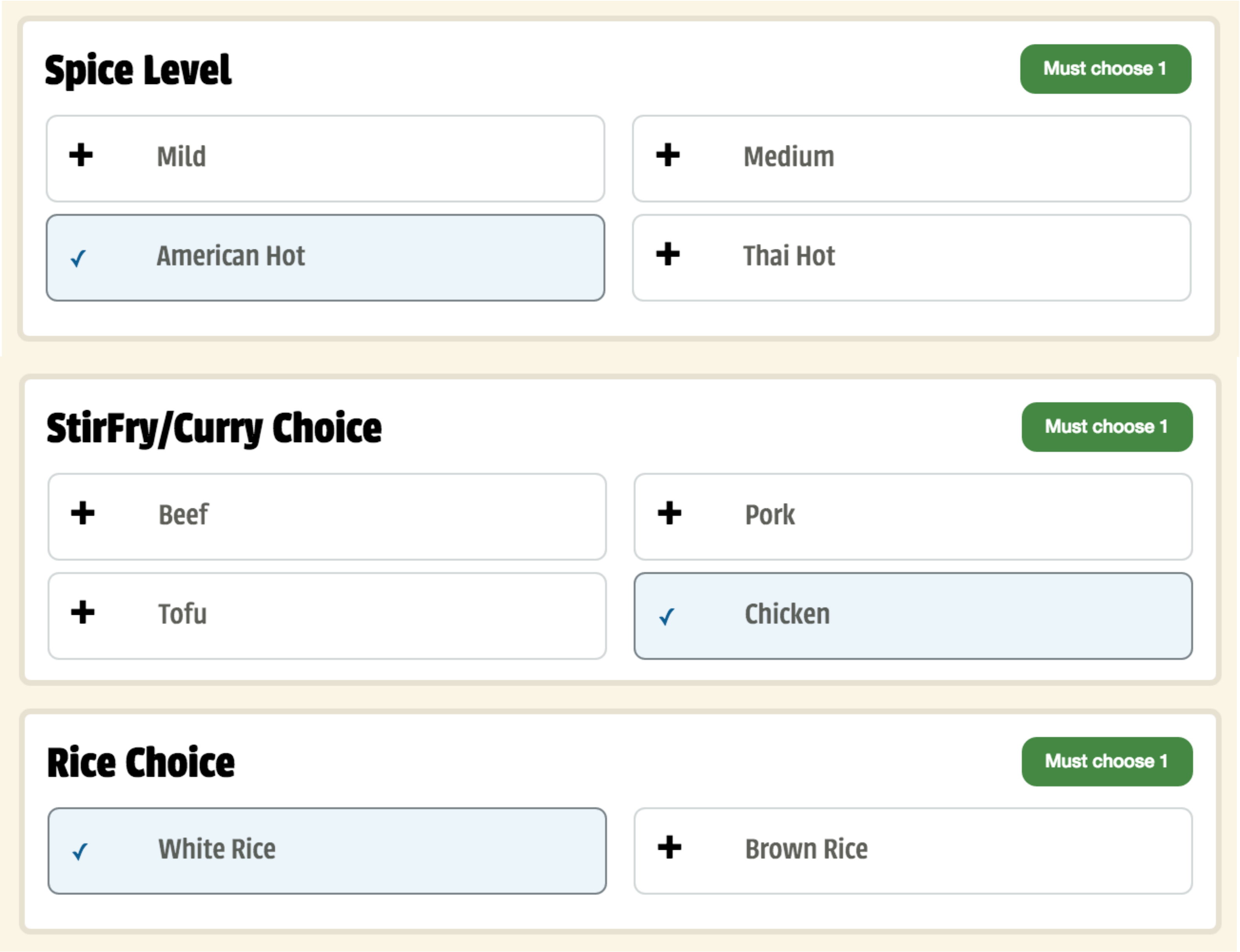
$13.20

Coconut milk, bamboo shoots, bell peppers & basil leaves

Label for: (e.g. Dan)

Price is generally considered **continuous**

- Although at some level, it is discrete because the price can’t have fractions of pennies, e.g. $13.204



Spice Level

Must choose 1

+ Mild

+ Medium

✓ American Hot

+ Thai Hot

StirFry/Curry Choice

Must choose 1

+ Beef

+ Pork

+ Tofu

✓ Chicken

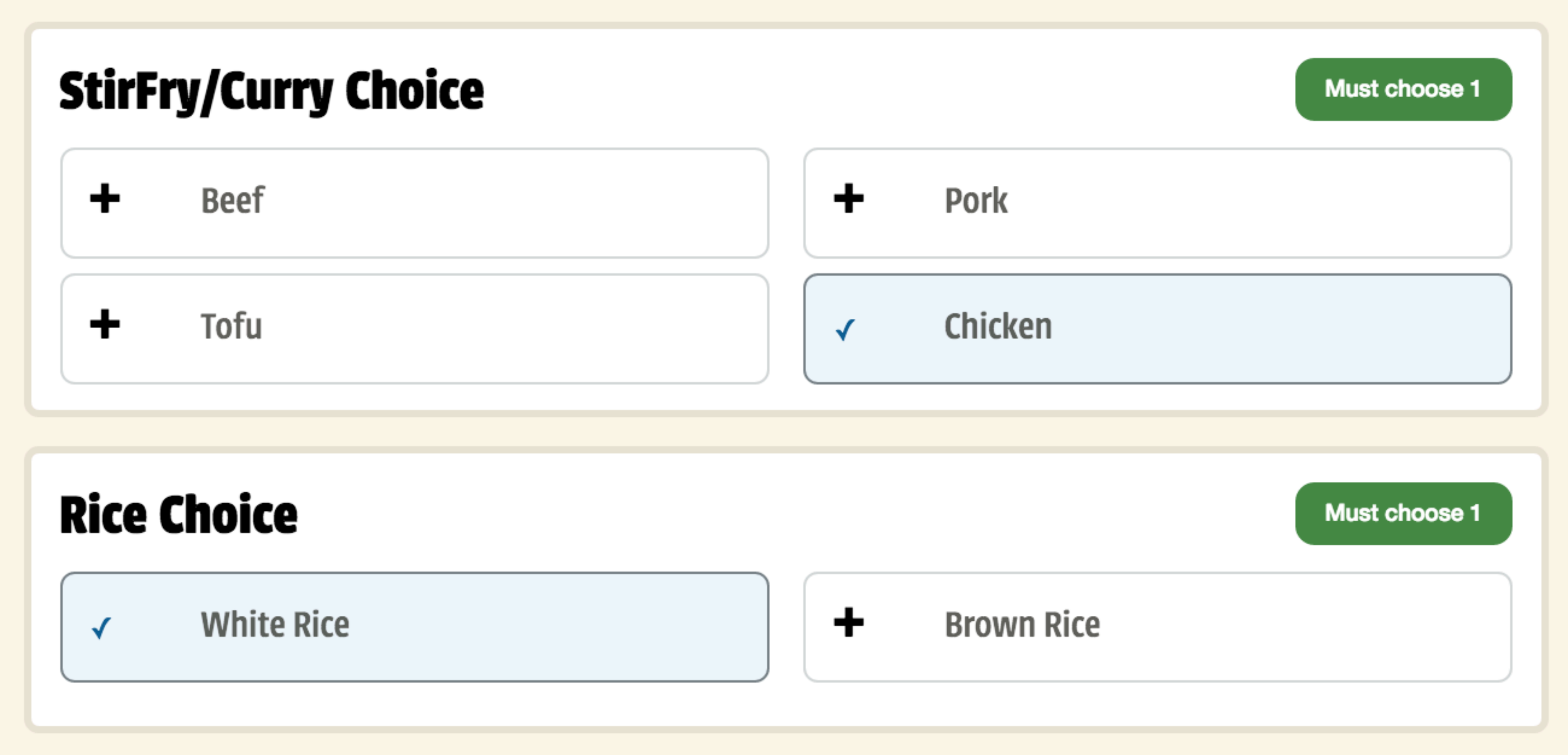
Rice Choice

Must choose 1

✓ White Rice

+ Brown Rice

Categorical values



StirFry/Curry Choice

Must choose 1

+ Beef

+ Pork

+ Tofu

✓ Chicken

Rice Choice

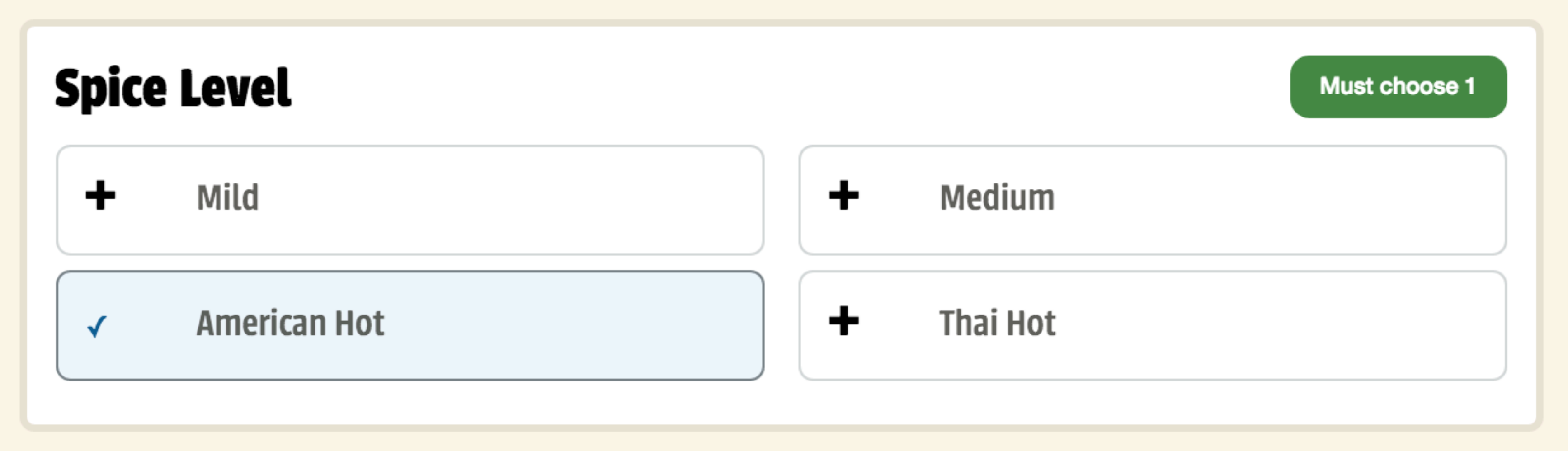
Must choose 1

✓ White Rice

+ Brown Rice

Types of protein and rice are **nominal** categories

The category values describe a characteristic of the dish



Spice Level

Must choose 1

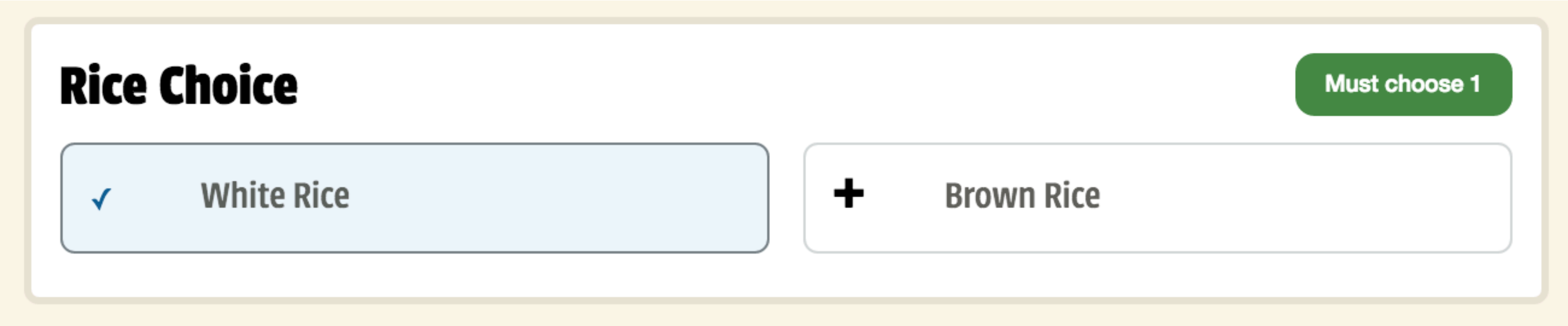
+ Mild	+ Medium
✓ American Hot	+ Thai Hot

Spice levels are **ordinal** categories

The categories imply an ordering of increased spiciness

- Mild → Medium → American Hot → Thai Hot

Returning to our representation…



Rice Choice

Must choose 1

✓ White Rice

+ Brown Rice

Remember our terminology:

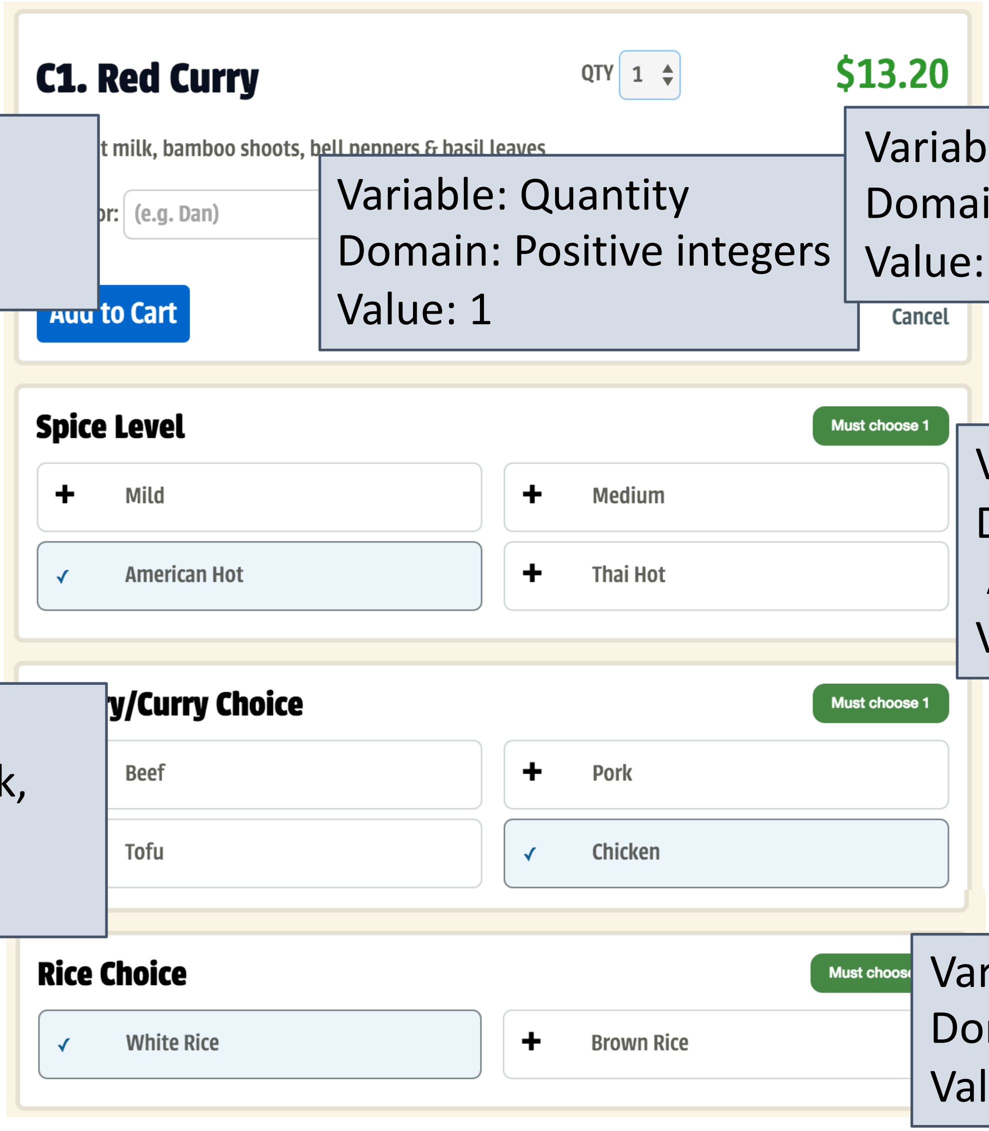
Variable: Rice Choice

Domain (“Levels”): {White, Brown}

Value: White

Notation:
Curly braces **{ }** are used to show that this is a **set**

Returning to our representation…



C1. Red Curry QTY 1 **$13.20**

t milk, bamboo shoots, bell peppers & basil leaves

or: (e.g. Dan)

Add to Cart

Cancel

Spice Level Must choose 1

+ Mild
+ Medium
✓ American Hot
+ Thai Hot

y/Curry Choice Must choose 1

Beef
+ Pork
Tofu
✓ Chicken

Rice Choice Must choos

✓ White Rice
+ Brown Rice

Variable: Dish
Domain: Any text
Value: "Red Curry"

Variable: Quantity
Domain: Positive integers
Value: 1

Variable: Price
Domain: Positive real numbers
Value: 13.20

Variable: Spice Level
Domain: {Mild,Medium, American Hot,Thai Hot}
Value: American Hot

Variable: Protein
Domain: {Beef,Pork, Tofu,Chicken}
Value: Chicken

Variable: Rice
Domain: {White,Brown}
Value: White

Returning to our representation…

We can organize all of this as a data matrix:

Dish	Qty	Price	Protein	Spice Level	Rice
Red Curry	1	13.20	Chicken	American Hot	White
…	…	…	…	…	…

Terminology reminder:
Each row is called a **observation** or **case** or **instance**

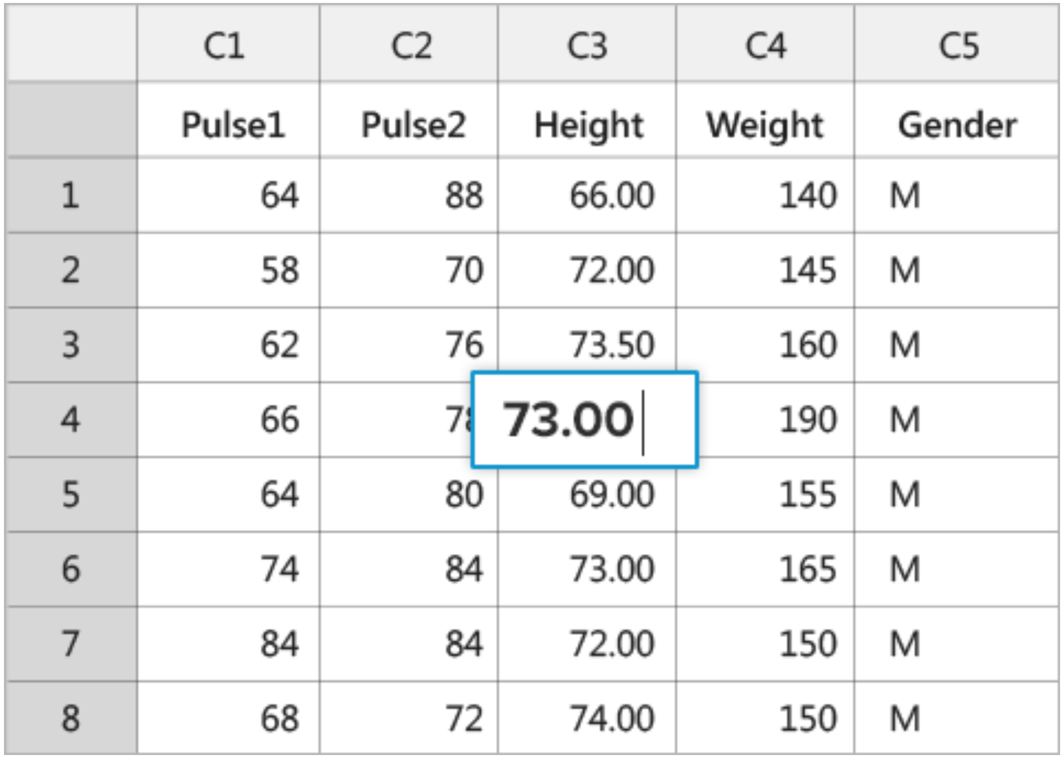
Raw data that you observe "in the wild" is not conveniently organized as variables, but you can conceptualize it this way

Pause

Questions at this point?

Representing data in practice

Most data analysis software uses a row/column representation



	C1	C2	C3	C4	C5
	Pulse1	Pulse2	Height	Weight	Gender
1	64	88	66.00	140	M
2	58	70	72.00	145	M
3	62	76	73.50	160	M
4	66	78	73.00	190	M
5	64	80	69.00	155	M
6	74	84	73.00	165	M
7	84	84	72.00	150	M
8	68	72	74.00	150	M

Representing data in practice

Software:

Getting started video:

http://support.minitab.com/en-us/minitab-express/1/getting-started/

Representing data in practice

We’ll practice on Friday – bring your laptops!