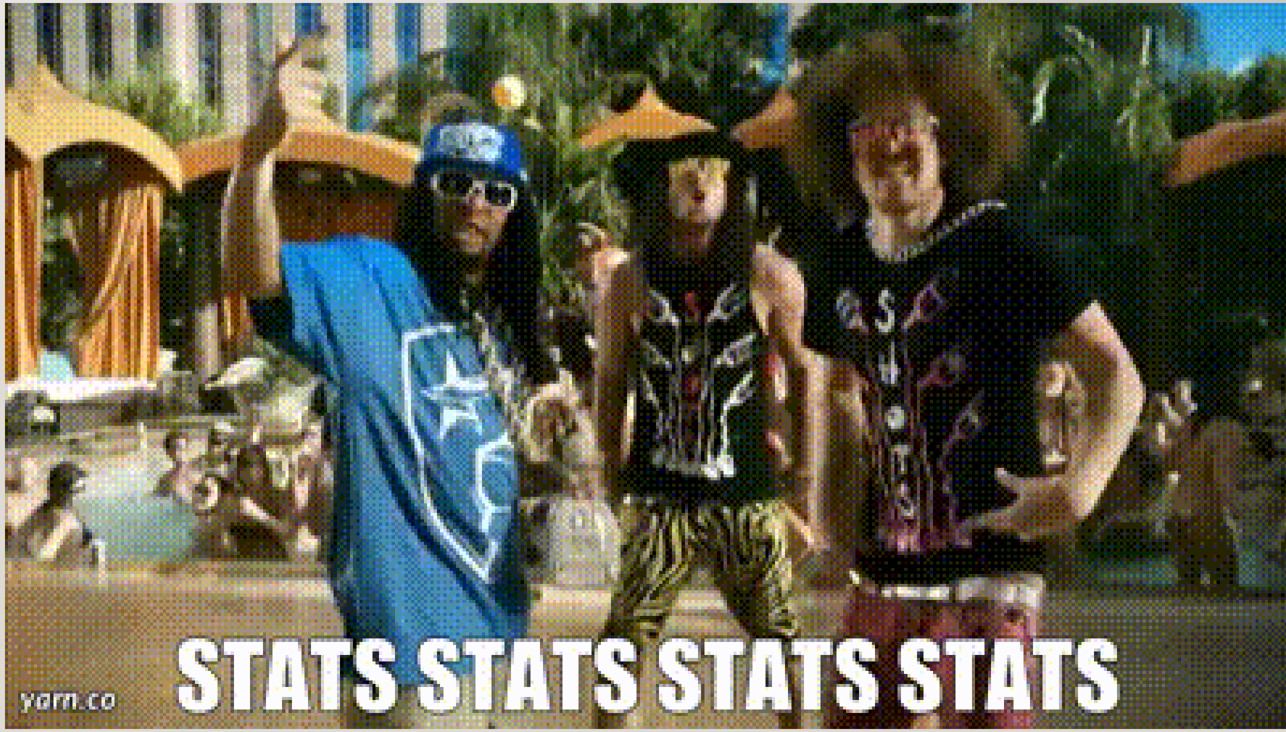
EXPLORING, CLEANING, AND BASIC DESCRIPTIVE STATS

PREPARING DATA

- To this point we should be able to load in some data into a dataframe.
- Before we can do some machine learning we need to prepare this data.
- The first step is to explore our data, or learn about it, which leads to…

STATS! STATS! STATS! STATS! STATS!

- The primary things that statistics gives us is a langue to describe data.
  - Descriptive statistics.
- There are a few basic statistics that we've likely seen/used before.
- These statistics allow us to describe one variable (feature) of data at a time.

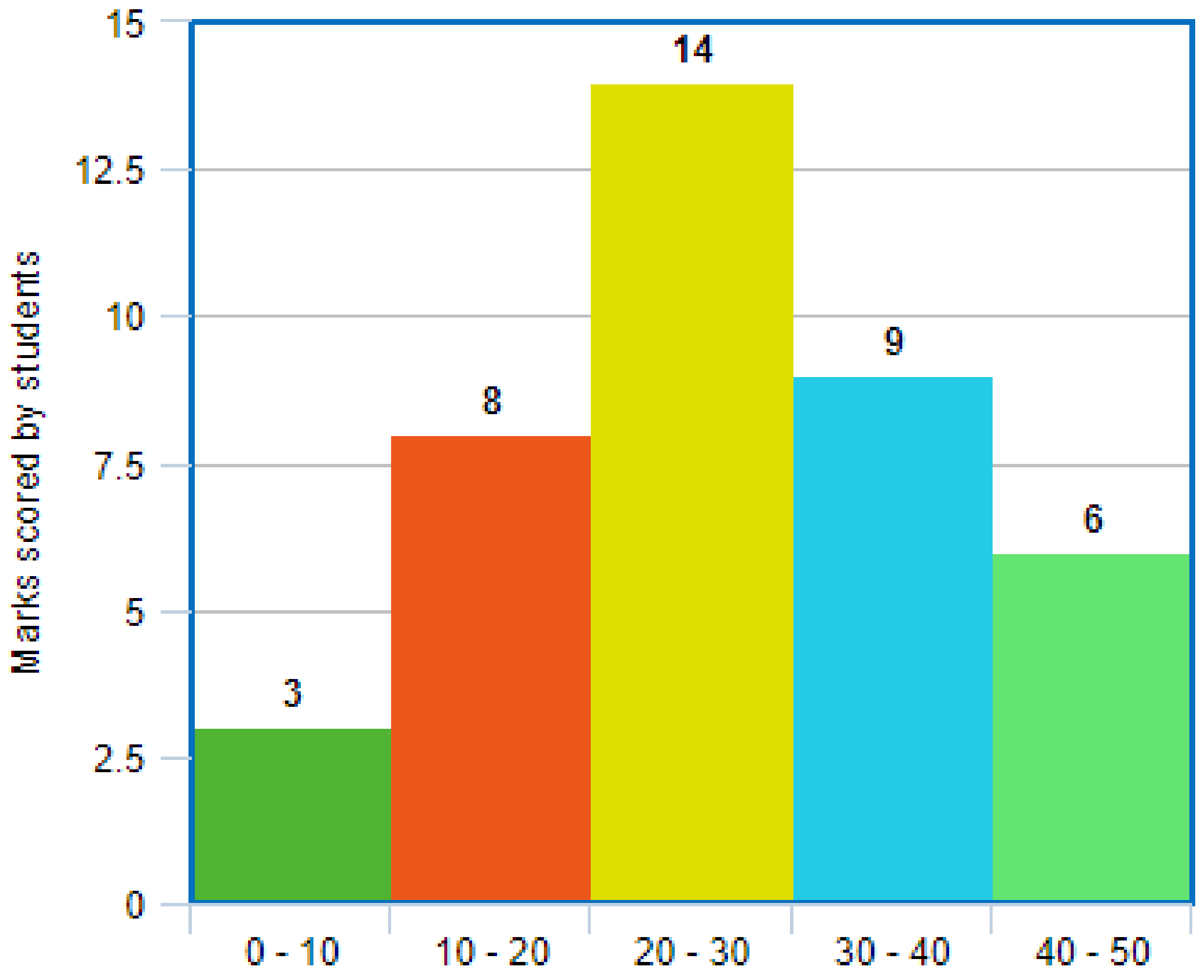


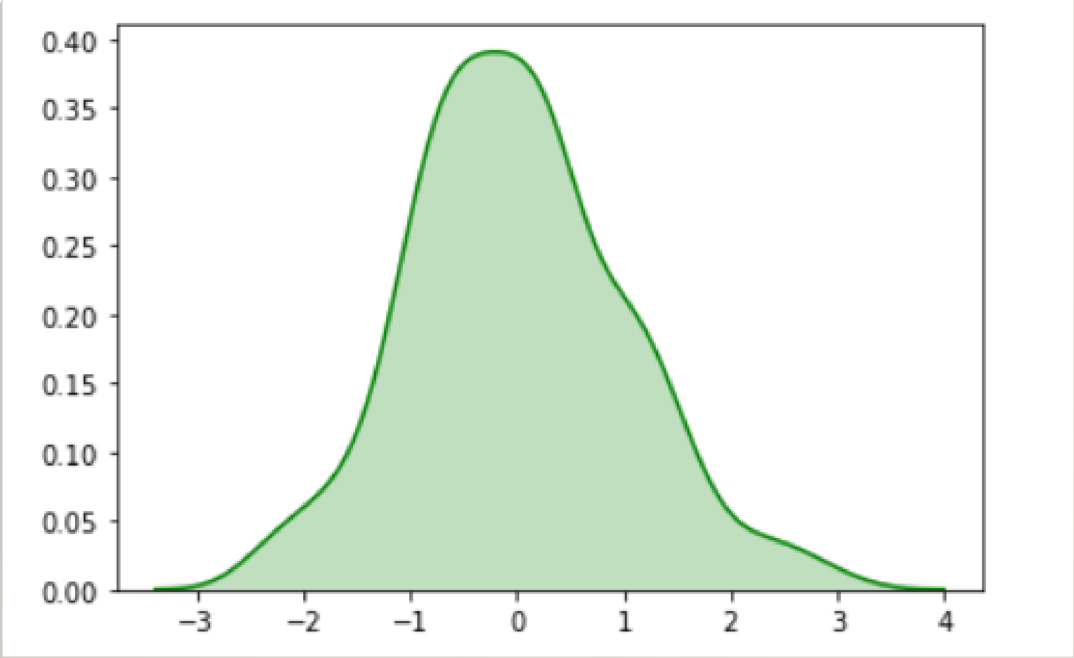
VISUALIZING STATISTICS

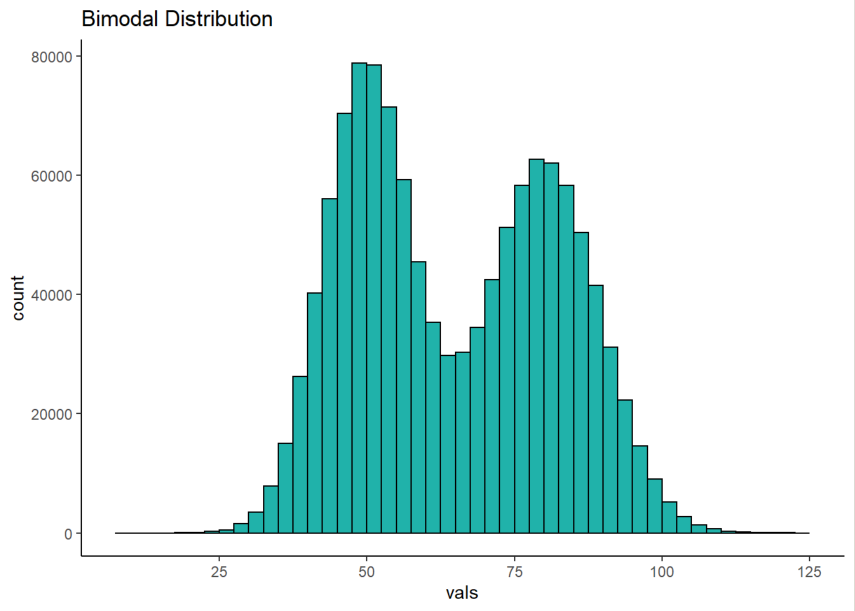
- We can calculate the statistics we need to look at.
- We can also visualize these statistics to help understand them.

HISTOGRAM

- A histogram is a plot to show us the stats of one variable.
- It is generated by breaking a variable into "bins", and counting the number of records in each bin.
  - Binning is just grouping records into a group for each range.
- The y axis is just the count of each bin.
- Allows us to visualize range, estimate mean/median, and see the distribution.





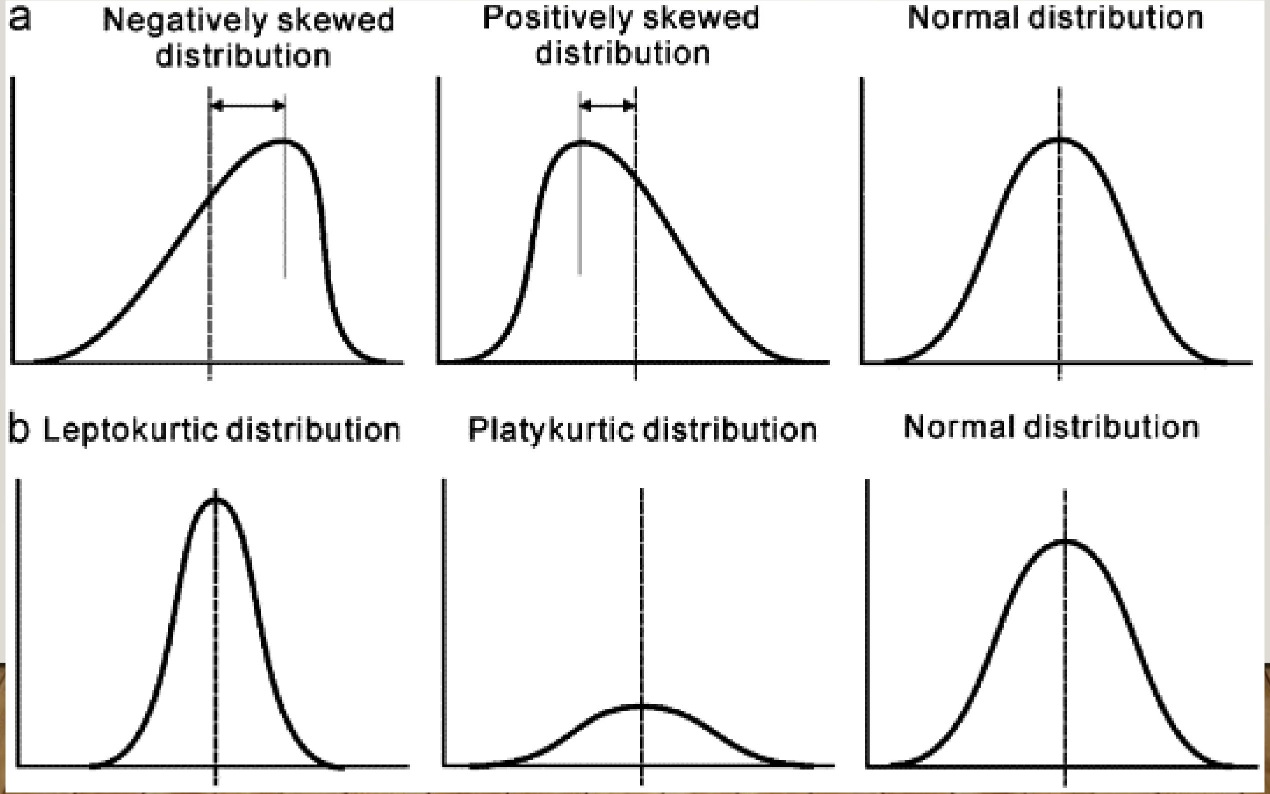


VIEWING DISTRIBUTIONS

- We normally look at the "shape" when discussing one numeric variable.
- This type of density plot shows us the same data as a histogram, but smoother.
- Note – the normal (bell) distribution is common, but not universal!

DISTRIBUTIONS

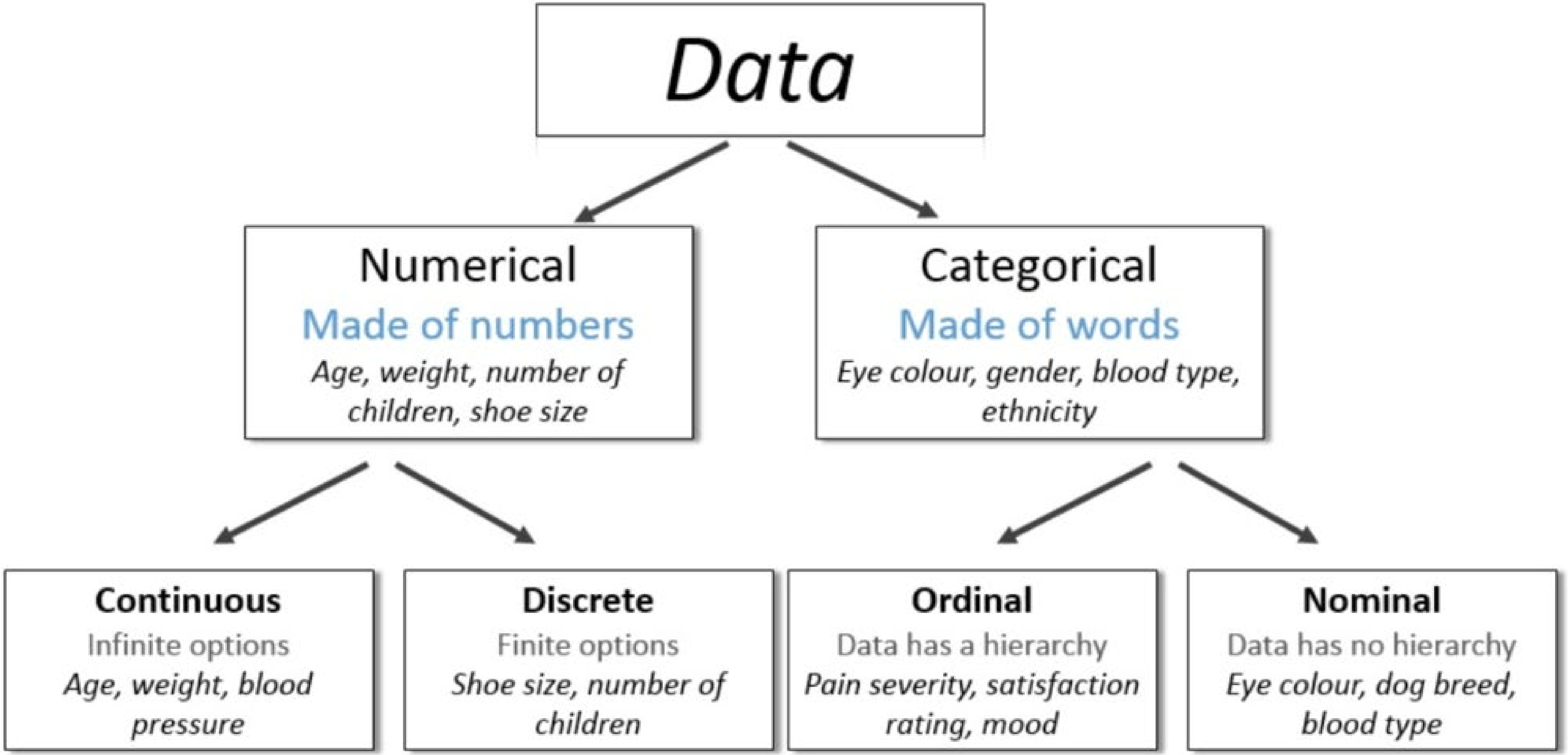
- These individual statistics are helping to build us up to looking at the distribution – a visual representation of the "shape" of the data's distribution.



VALUES IN A DATASET

- Each of these values is a certain type of value.
- Some values are descriptors, like name or hair color.
- Some values are measurements, like height or bank account balance.

- We can break datatypes into a few divisions…


Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

VARIABLE TYPES

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread: *categorical, ordinal - could also be used as numerical*

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

	Categorical	Quantitative
Definition	*Take on names or labels*	*Take on numeric values*
Examples	Marital Status	Height
	Smoking Status	Population Size
	Eye Color	Square Footage
	Level of Education	Class Size

DATA TYPES

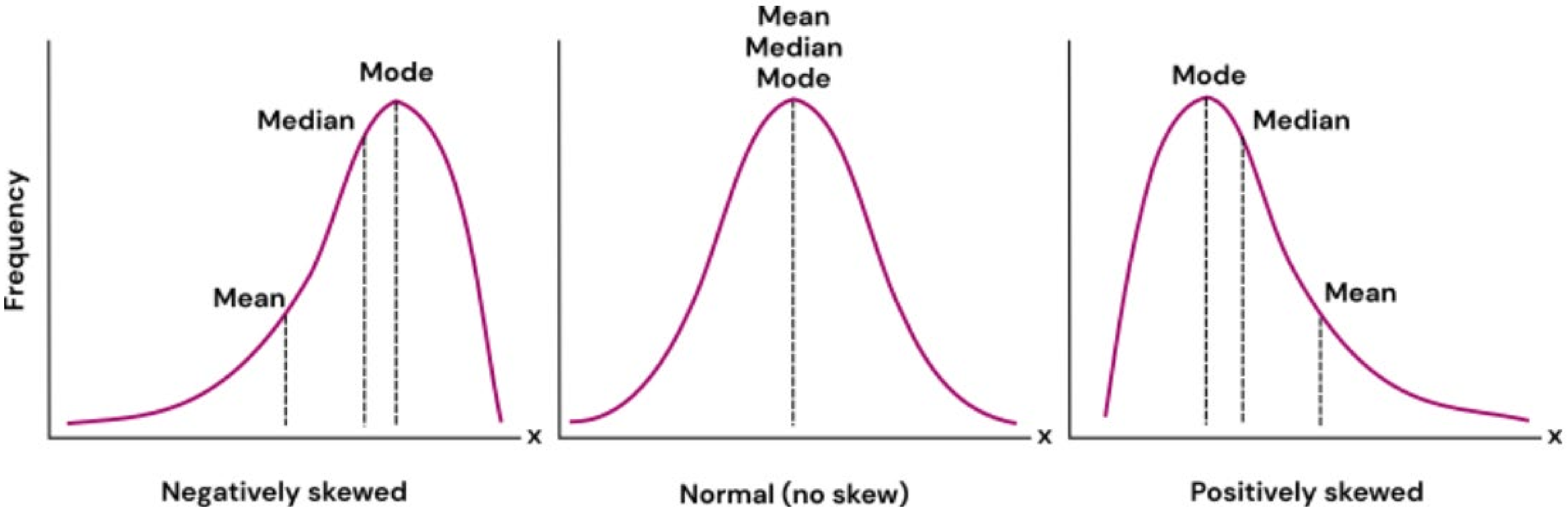
- The split between numerical and categorical is critical.
  - Predicting categorical vs numerical values when we get to ML is different.
  - How we analyze and process categorical vs numerical values is different.
- We generally group things by categorical variables.
  - E.g. group all students who took IB courses in HS when looking at earnings.
- We generally calculate things for numerical variables.
  - E.g. calculate the median of income for the IB group, compared to others.

RANGE

- Minimum – smallest value.
- Maximum – largest value.
- Range – distance between the minimum and the maximum values.
- Count (N) – number of records in dataset.

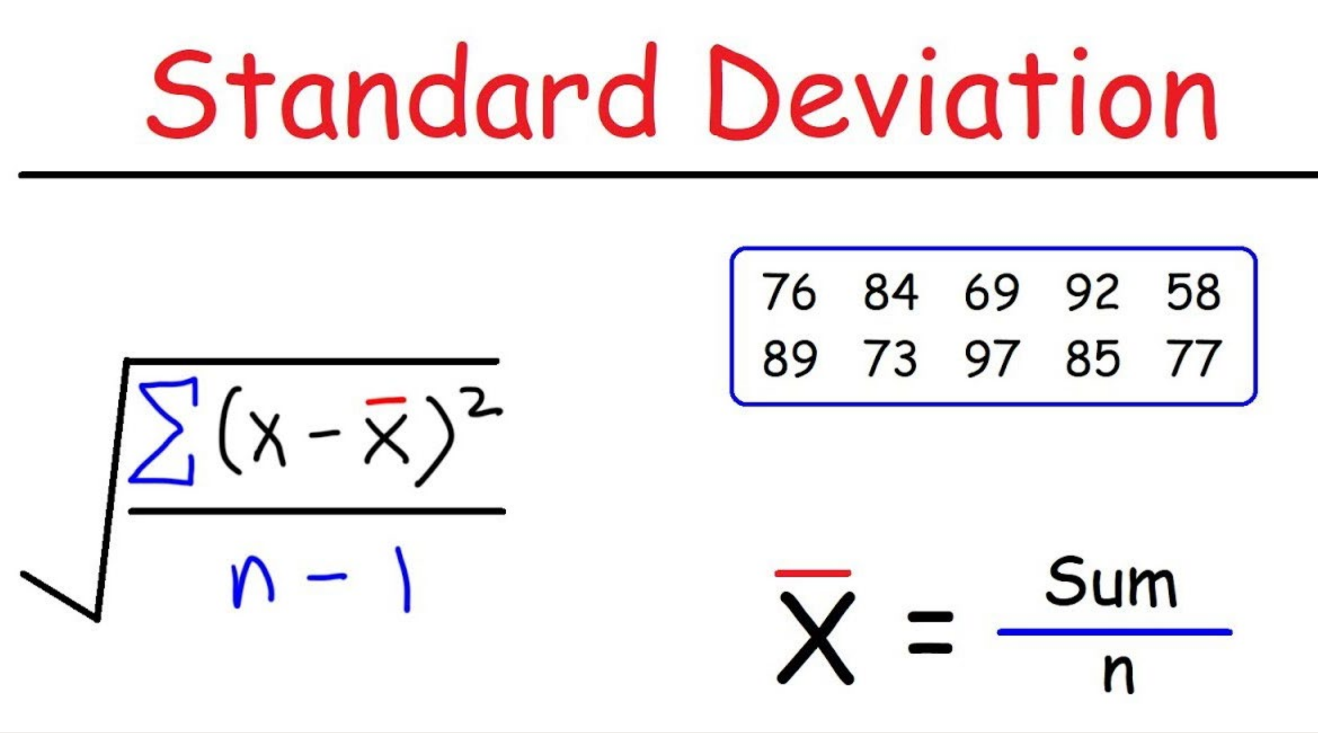
AVERAGE(S) – MEASURES OF CENTRAL TENDENCY

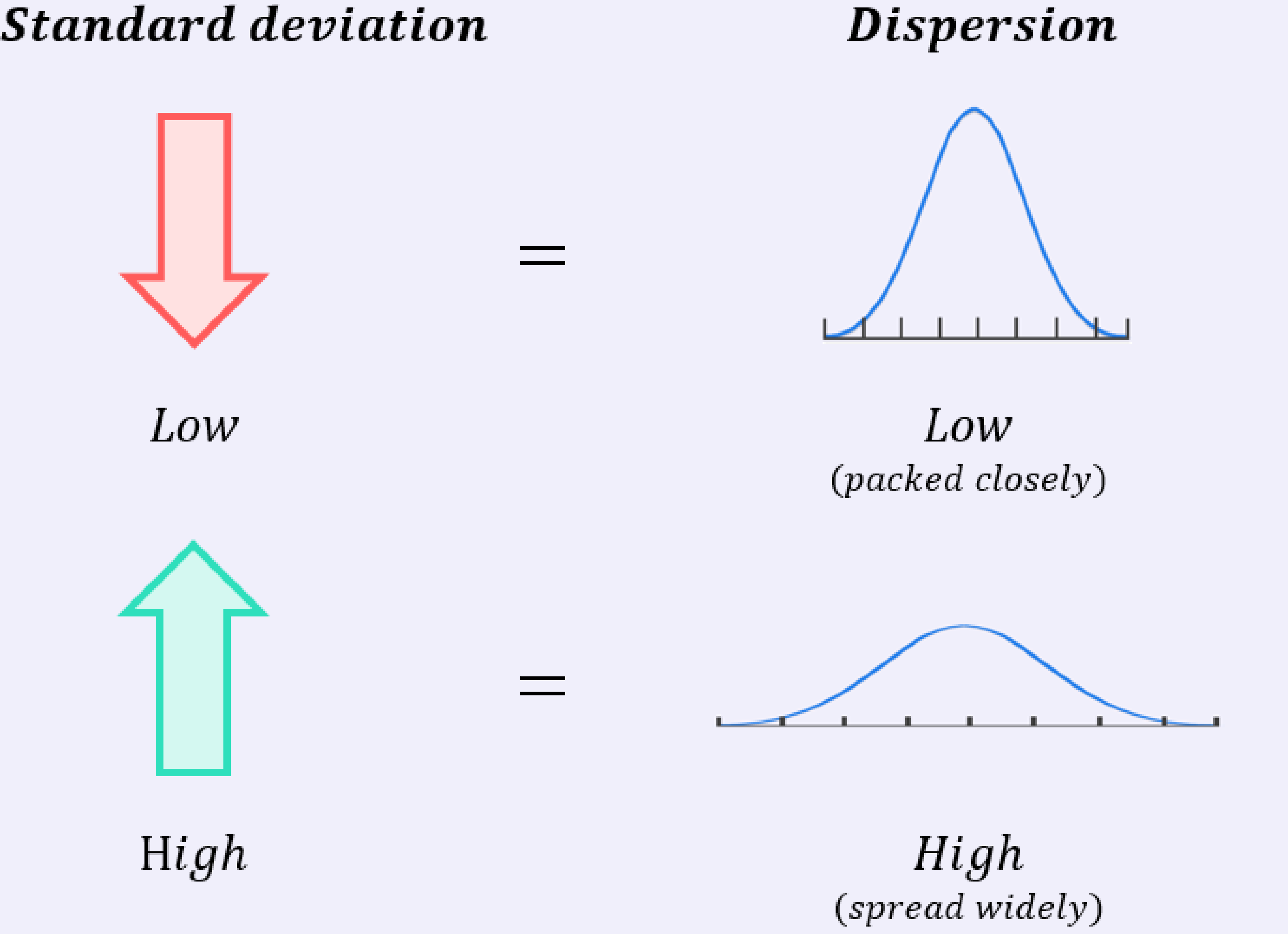
- We have 3 measures of average:
  - Mean – Add all values and divide by N.
  - Median – The value with 50% of other values above, and %50 below.
  - Mode – The most frequently occurring value.
- "Average" normally means the mean, but we should be specific.
- Median is very common is scenarios where there are outliers.
  - Why?
- Mode isn't usually all that useful with decimal numbers.


Frequency
Mode
Median
Mean
x
Negatively skewed
Mean
Median
Mode
x
Normal (no skew)
Mode
Median
Mean
x
Positively skewed

MEASURES OF DISPERSION

- Measures of dispersion tell us how "spread out" the values are?
  - Are values tightly clustered or scattered over a wide area?
  - Variance – a measure of how "varied" the values are, i.e. are they clustered over a small range or distributed broadly.
- Standard Deviation – the square root of variance. More commonly used for most analysis.
  - Roughly, "how far from the mean is a typical value?"





DISPERSION

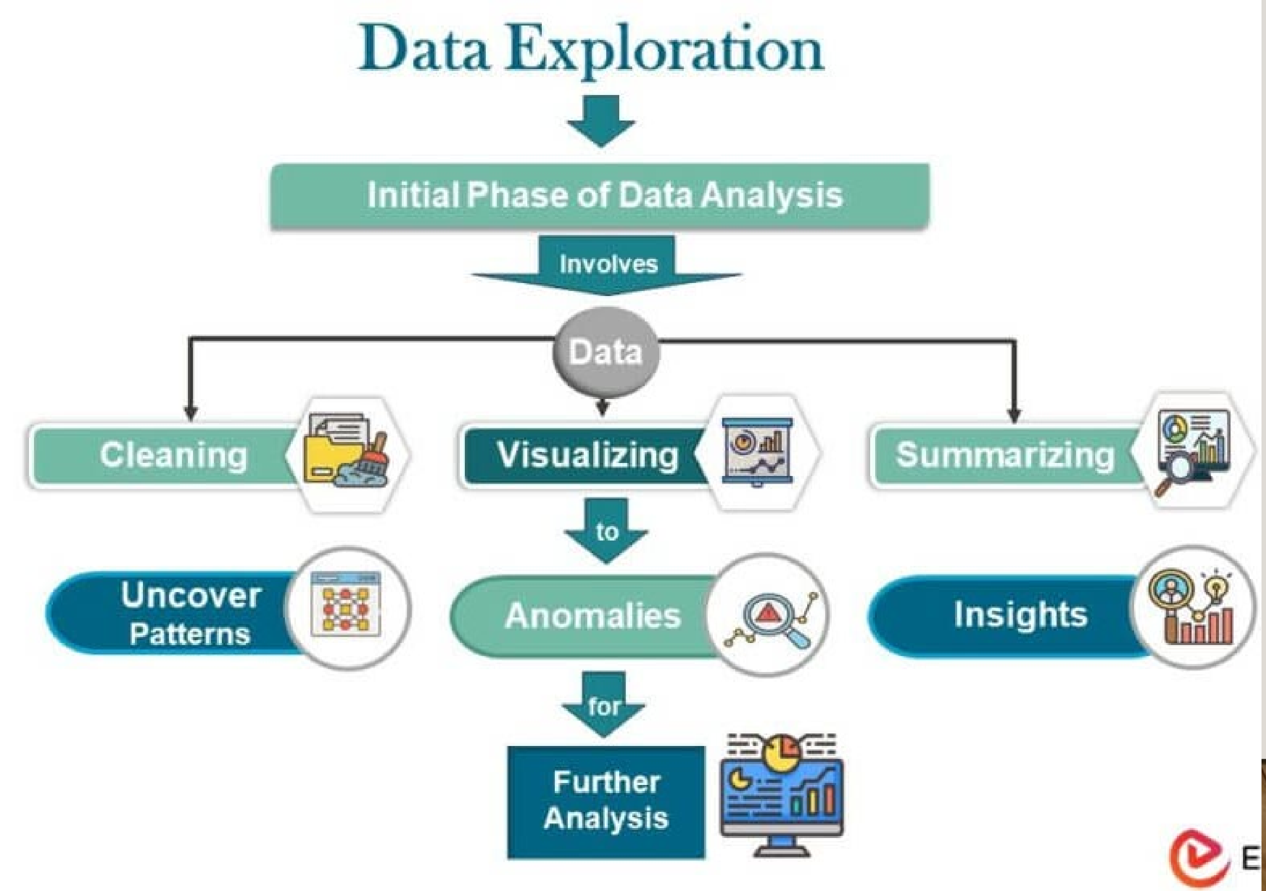
- Dispersion metrics tell us if our data is tightly grouped, or spread out.

SINGLE VARIABLE STATISTICS

- These simple stats help us describe data that we are dealing with.
- When we look at distributions soon, knowing a distribution pattern and these basic statistics can allow us to describe our data very accurately with a small amount of info.
- These are fundamental building blocks, we should be comfortable with each and what it means.
- Note: each of these stats looks at one variable at a time, we haven't looked at all at the relationships between them.
- You'll need to know mean, median, range, std and be comfortable with them.

WHY?

- We explore data to understand it and to know what needs to be done to get it ready.

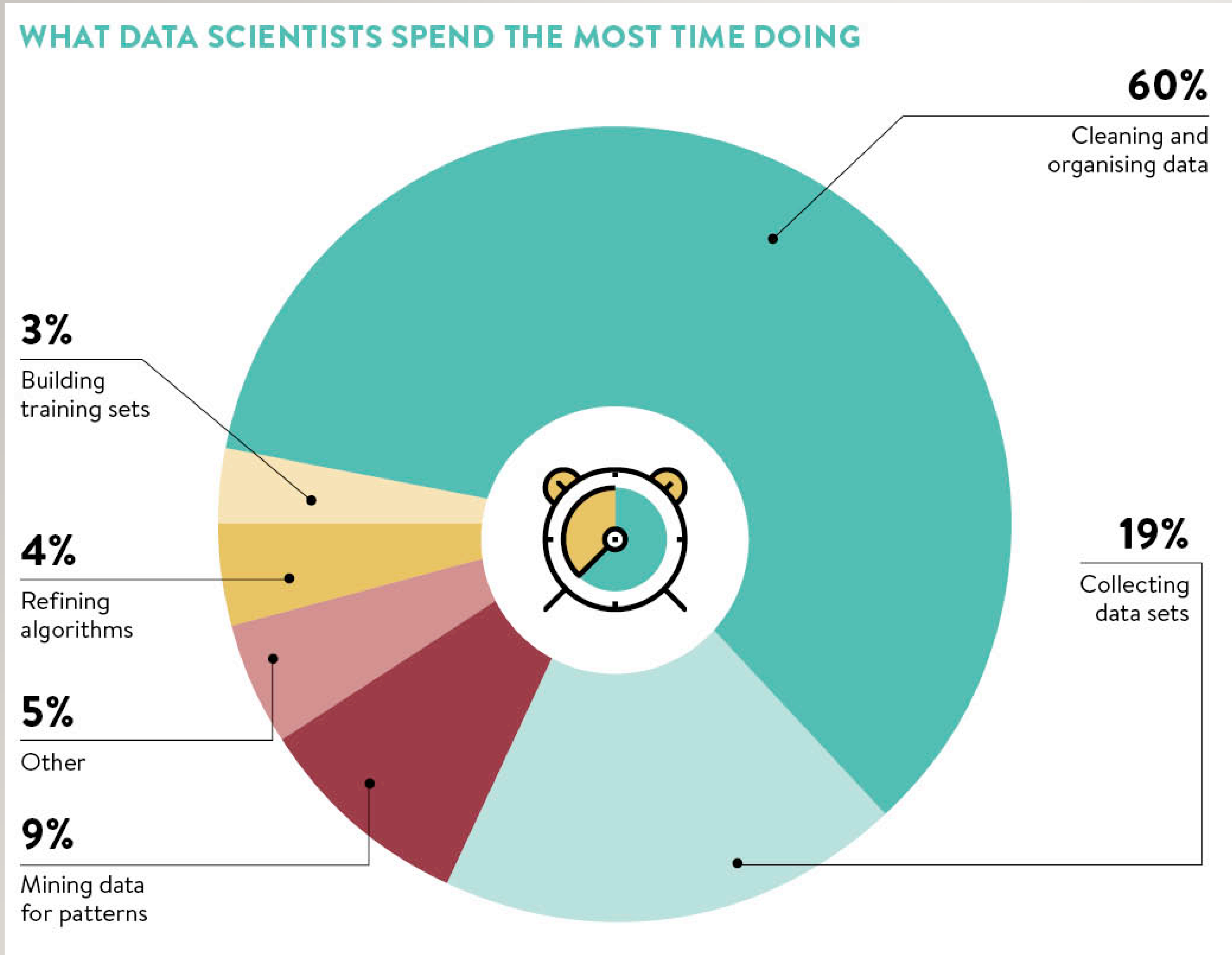


UNDERSTANDING THE DATA

- One reason for data exploration is to “understand” the data better prior to analysis.
- Understanding is a very vague goal.
- We don’t really *need* a deep understanding of stats to do ML.
- As you do more DS work (likely beyond this class) the understanding matters more:
  - Judging which variables are important/useful and which aren’t.
  - Choosing different options to improve accuracy of predictions.
  - Deciding on different transformations (modifying the data) to help make better predictions.
  - This stuff isn’t necessary for it to work, but comes up when making things good.
- For now, we want tools to explore the data, the why comes as we learn.

DATA CLEANUP

- One big reason to explore data is to know what we need to do to clean it.
- Cleaning data is needed, but also open ended.
  - Remove large outliers, or at least check that we should keep them.
  - Fix any errors – stray values, mistakes, etc…
  - Convert and correct types – we want numbers to be numbers, dates as dates, etc…
  - More analytics-focused cleanup – relating to values and distributions.
- Before we do analysis, we need to cleanup the data.
  - For now – outliers, errors, data type mistakes primarily.
- This cleanup depends on the data, our goal, and our understanding.



CLEANING DATA IS BIG!

- Most of the work in data science is getting data ready!
- In "normal" programs we tell the computer what to do, in ML we give it examples (data) and it figures it out – our work is on prepping the data, and setting the "rule" of learning.

OK, TIME TO PROGRAM…