Probability and Statistics for Data Science

STATISTICS & PROBABILITY

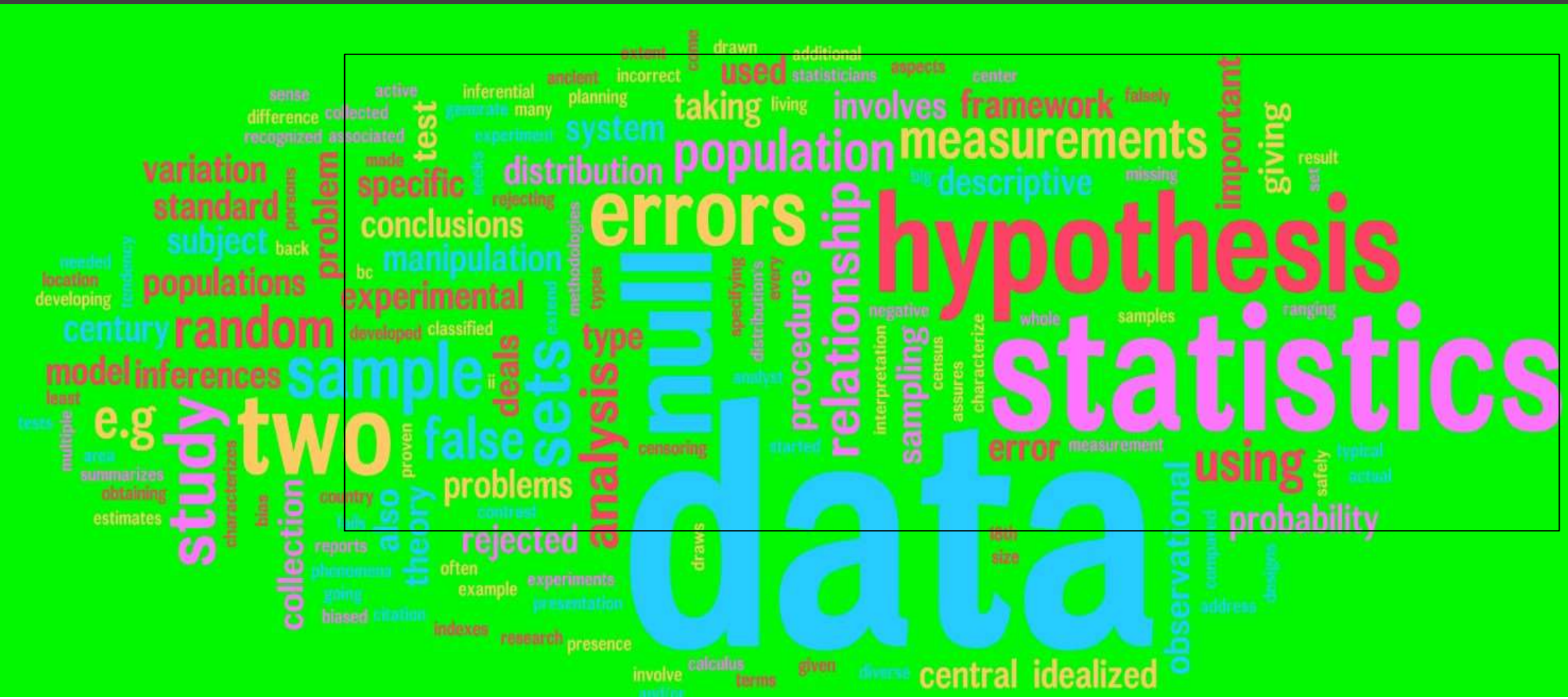
Science ***Data***

Variation

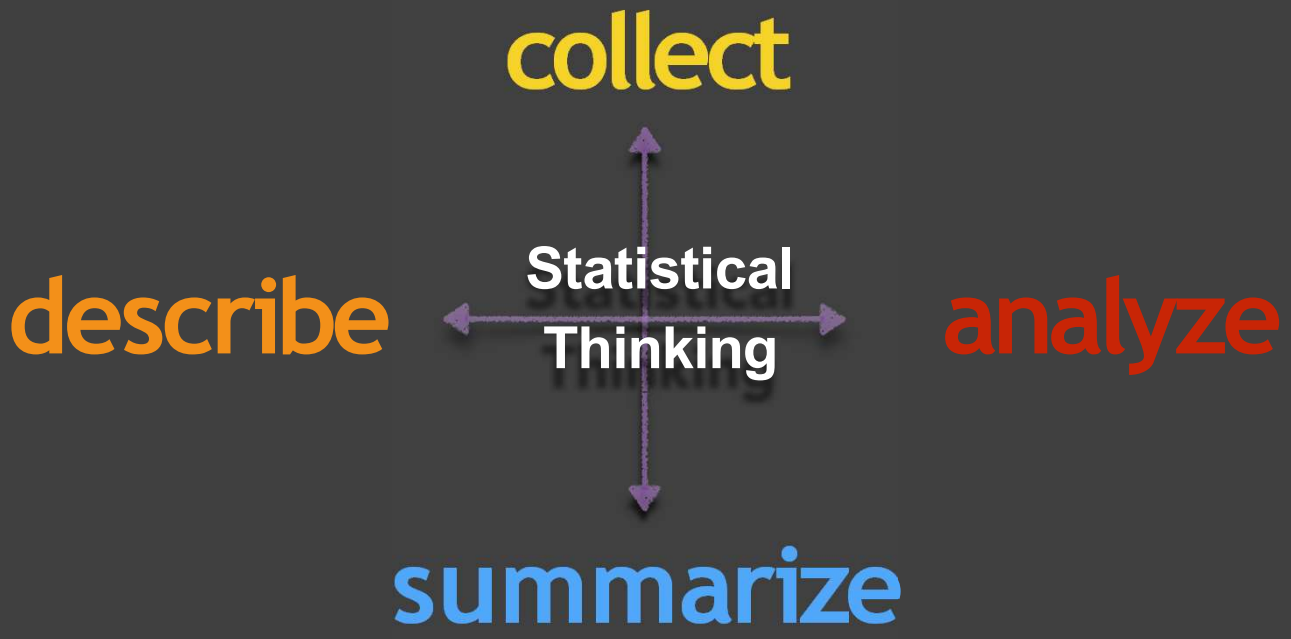
Chance

Randomness

STATISTICS & DATA



PROBABILITY THEORY & STATISTICAL METHODS



Statistical Thinking

Two Examples from Today's News

Example 1
A sample survey

GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market

1025 randomly sampled **Americans**

TELEPHONE SURVEY

GALLUP POLL, SEPTEMBER 2015

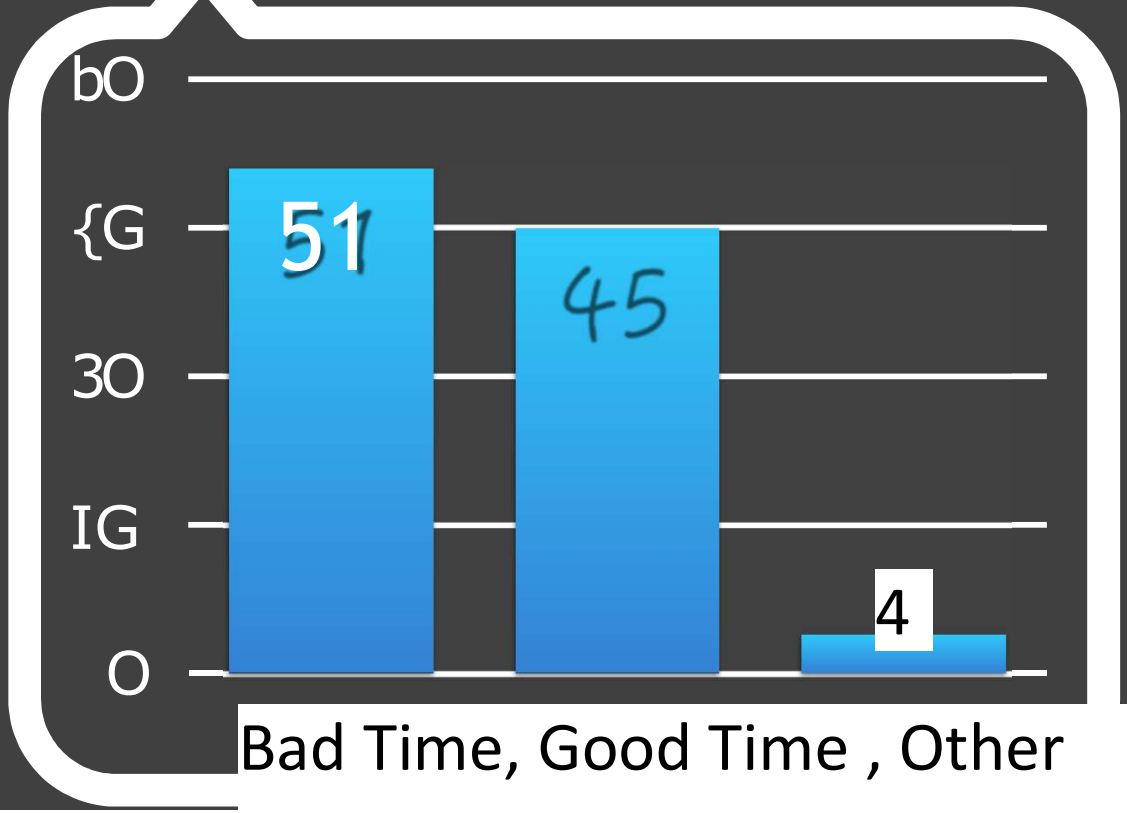
Americans' perception
on the job market

"Is it a GOOD TIME to FIND A JOB?"

GALLUP POLL, SEPTEMBER 2015

Americans' perception

on the job market



GALLUP POLL, SEPTEMBER 2015

Americans' perception
on the job market

Margin of Error
is 4% for 95%
confidence
interval

Margin of Error?

4%? 95%?

Confidence Interval?

STATISTICAL THINKING I

What was this study trying to find out?

Population of interest:
Americans

Information (variable) of interest:
Their perception on the job market

STATISTICAL THINKING I

STATISTICAL THINKING I

Why should we care about the opinion of the 1025 Americans who participated in this survey?

1025 Survey Participants

Representative?

300 Million Americans

STATISTICAL THINKING I

Statistics ***derives***

Knowledge

Sample ➡ ***Population***

Learning activity I:
understand mathematical notations

Notations

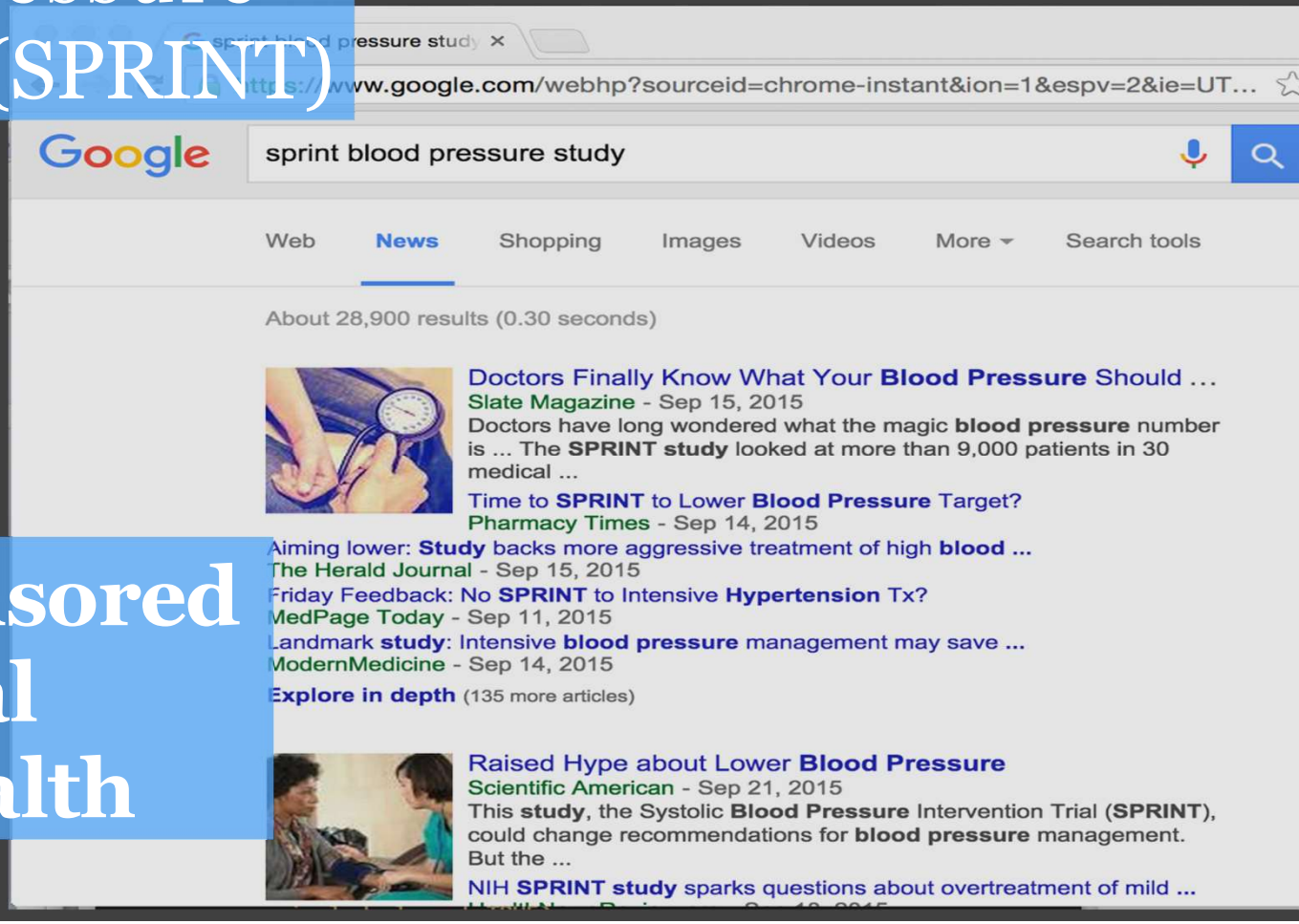
- **Statistics rely on computation of numerical summaries of data.**
- **Mathematical notation and equation formally describe such computation.**
- **Data are organized by individuals and variables.**
- **Variables, denoted by letters close to the end of the English alphabet such as X, Y.**
- **X with a subscript i is X's value for individual i.**
- **Summation sign.**

Example 2 A clinical trial

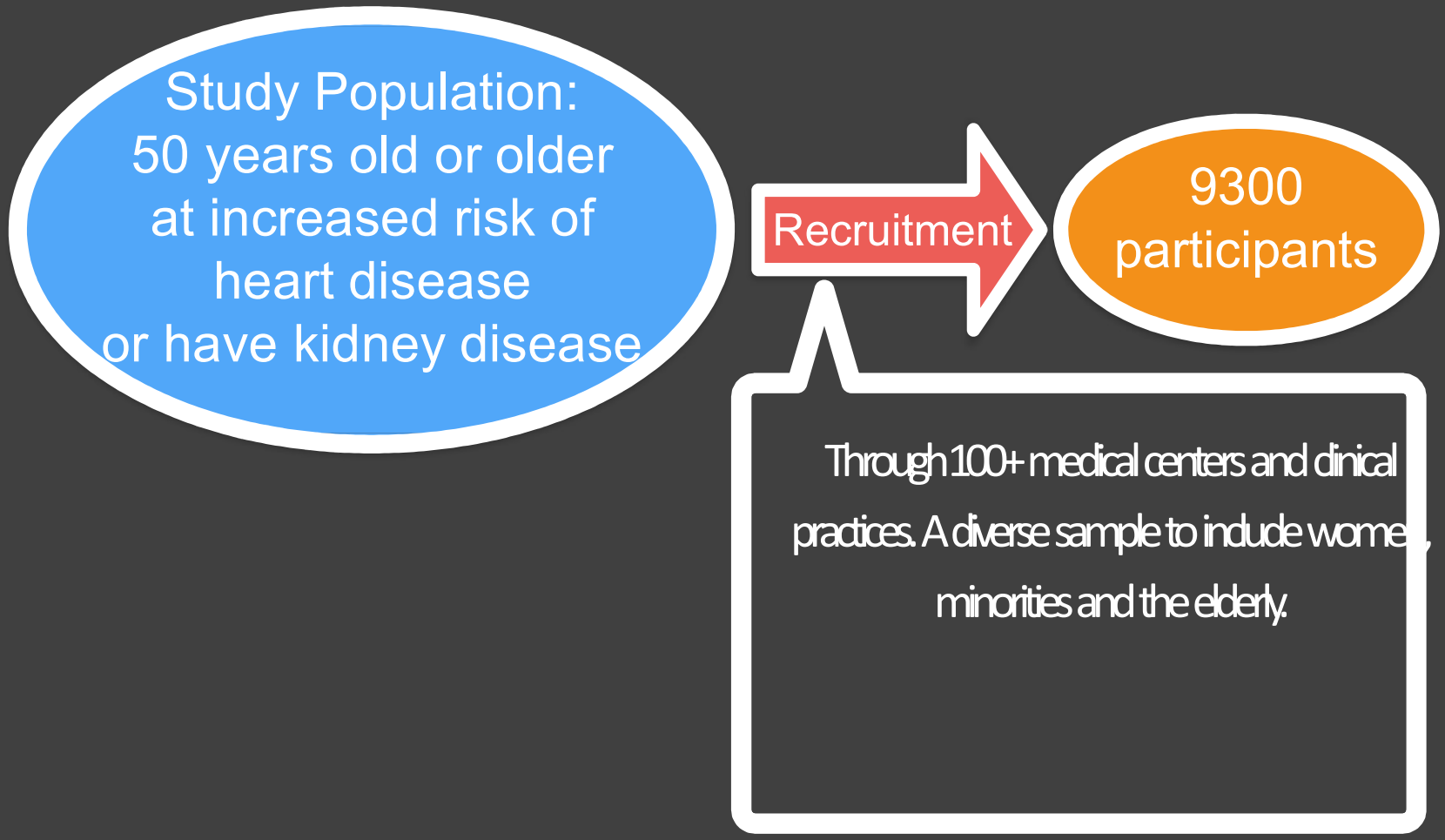
Example 2 - NIH News Release

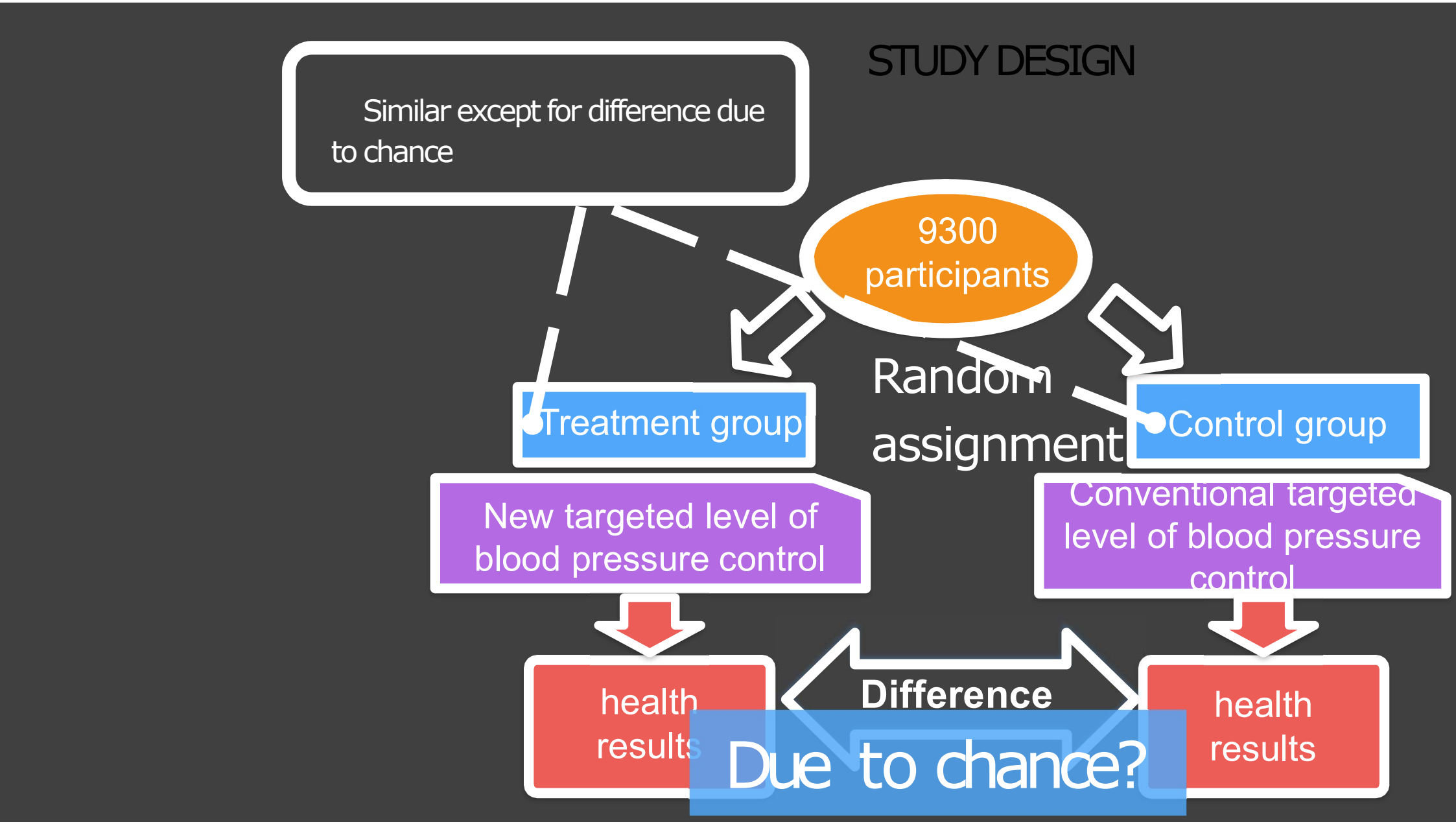
Systolic Blood Pressure Intervention Trial (SPRINT)

A clinical trial sponsored by the National Institutes of Health



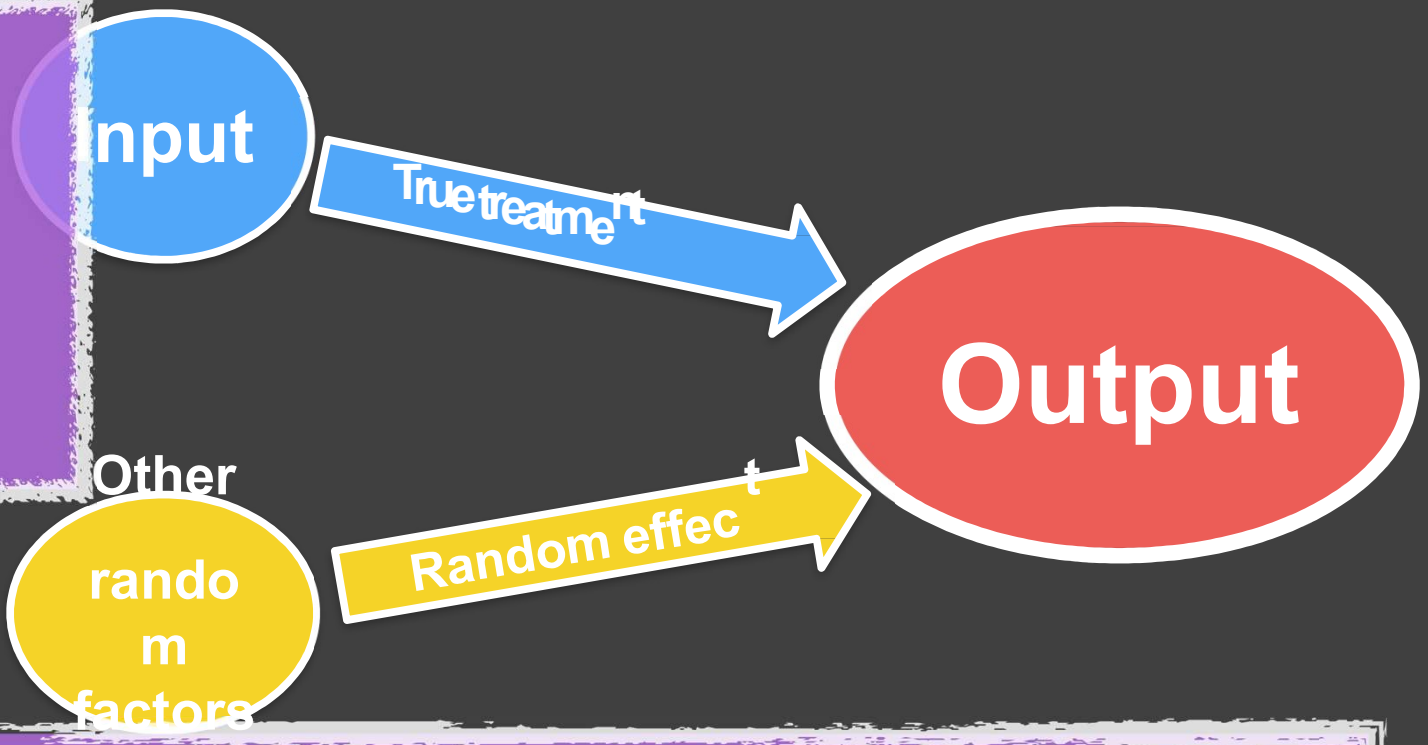
Study participants




STUDY DESIGN
Similar except for difference due to chance
9300 participants
Random assignment
Treatment group
Control group
New targeted level of blood pressure control
Conventional targeted level of blood pressure control
health results
Difference
health results
Due to chance?

Statistical thinking II

Statistical inference estimates the possible extent of the random effect for establishing the size of the true effect



The observed effect is called statistical significance if it is unlikely to occur purely by chance

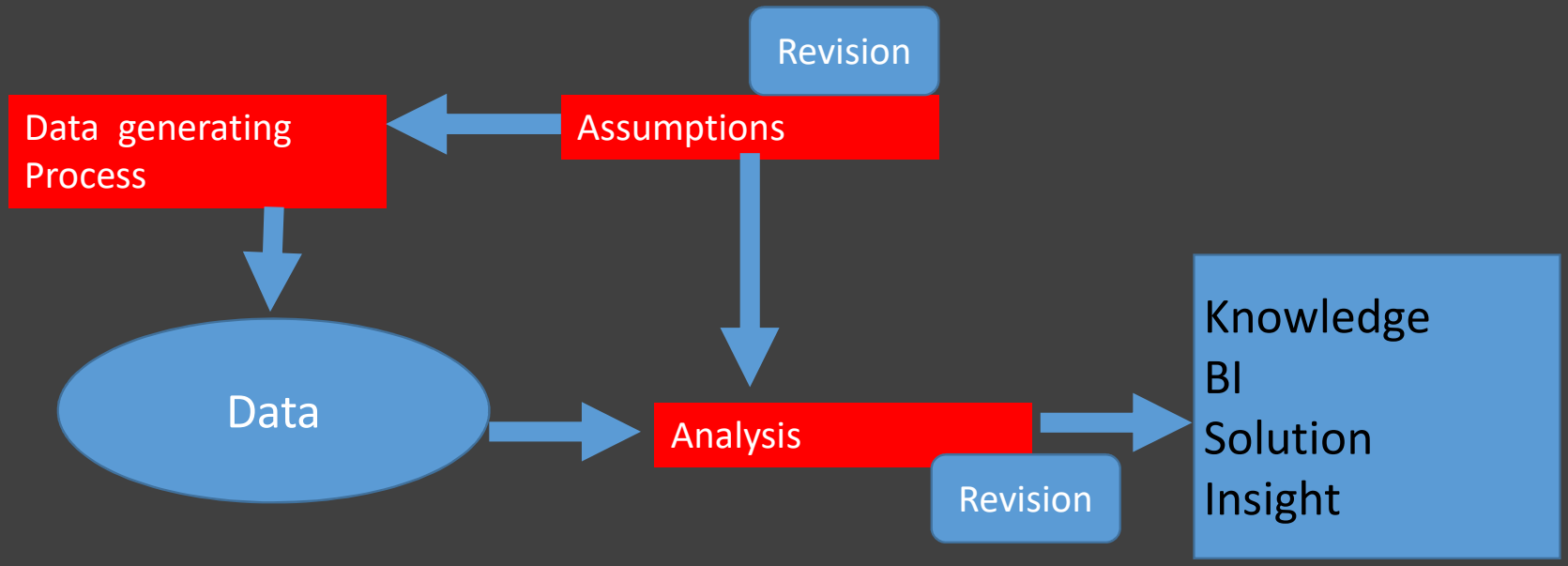
STATISTICAL THINKING 2

Statistics ***establishes***

Statistical Significance of observed signal

by studying randomness

Data to Solution



Garbage in, Garbage Out?

- Validity of the results depends on Validity of assumptions on the data generating process
- Regarding the sampling , randomization , measurements, independence , etc
- They are often violated big data
- Data Scientist investigate these assumptions and propose solution

Derive good answers
from data

Data: numbers with context

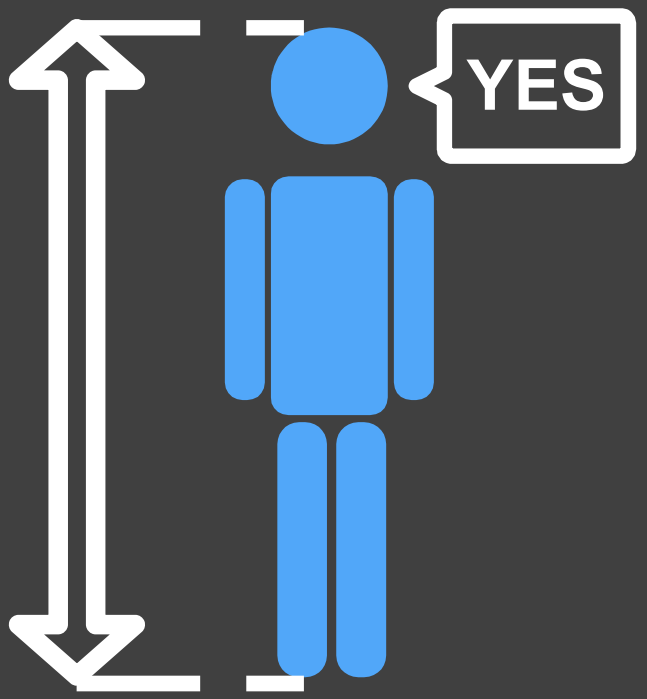
- degrees — today's highest temperature
- years old — age of a cancer patient when the cancer was diagnosed
- pounds — weight of a 10 years old boy
- seconds — how long a study participant can hold the plank position

NO

From Individuals to Statistics

Individuals: units of observation in a data set.

Also called study participants, study units, subjects, etc.



An individual in a study

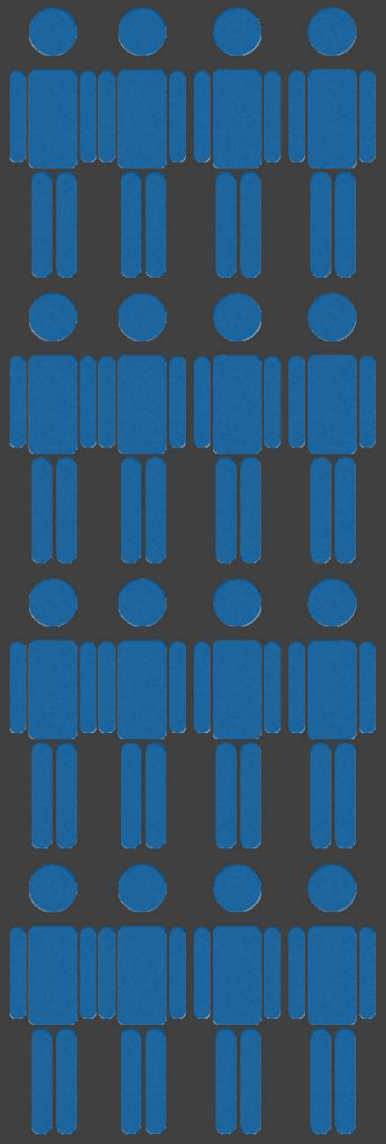
"Yes"

- ***Gender:*** *Male*
- ***"Q:Is it a good time to find a job?":***
- ***Height:*** *70 inches*
- ***Education, income, social behaviors***

From Individuals to Statis

Individuals:
units of observation in a da

Also called study
participants, study subjects



Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School

From Individuals to Statistics

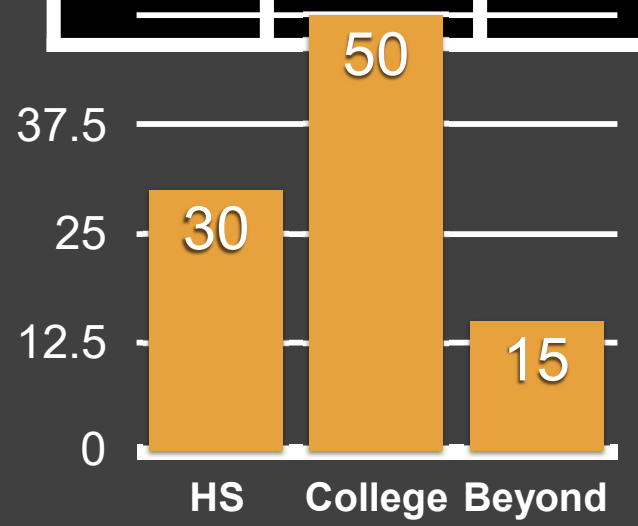
- **Categorical:** the values represent different categories for the individuals; do not have arithmetical meaning.
- Quantitative: the values represent numerical quantities that can be ordered and averaged.
- Ordinal: the values represent ordered categories; such as “how often do you exercise?”—*Everyday, frequently, sometimes, rarely, never.*

From Individuals to Statistics

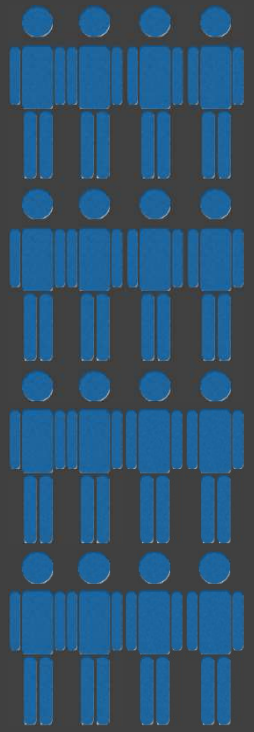
Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School
…			

et.

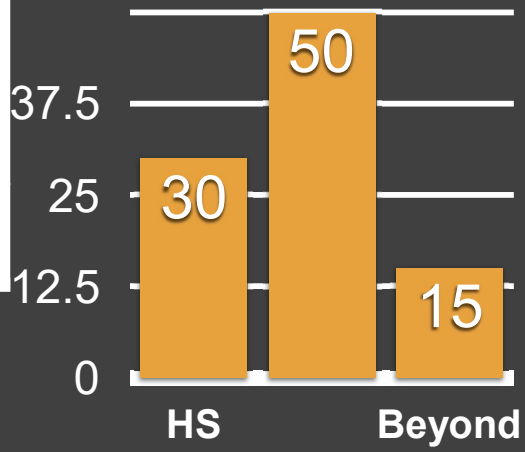
High School	College	Beyond college
20	50	15



From Individuals to Statistics



Individual	Gender	Height	Education
1	Male	70	College
2	Female	68	College
3	Male	69	High School
...			

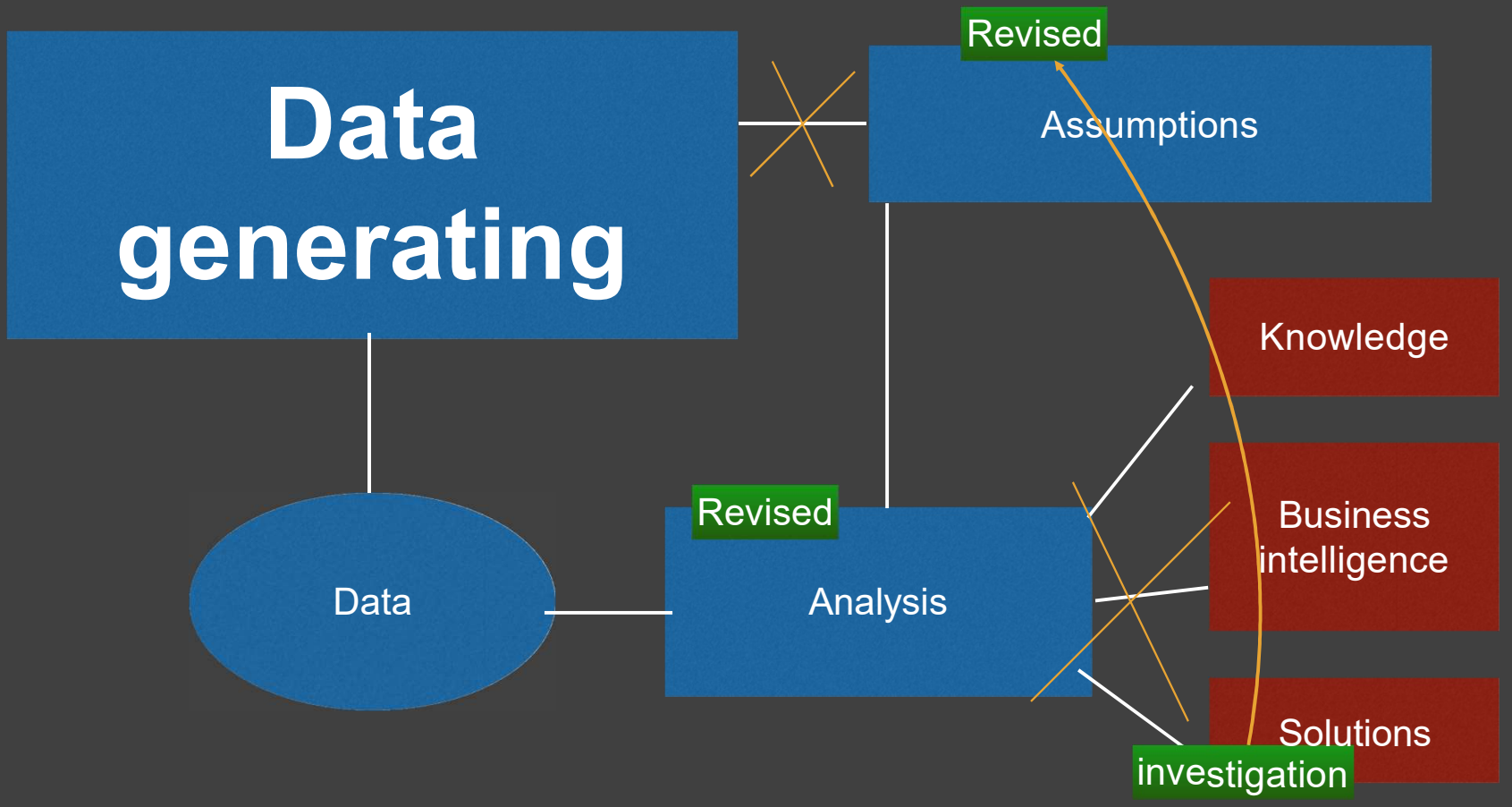


From Individuals to Statistics

Statistics are

Summaries of Numerical Data that don't tell the whole story, but are useful and meaningful

From Data to answers



Garbage in, Garbage out?

- Validity of results depend on the validity of assumptions on the data generating process.
  - regarding the sampling, randomization, measurements, independence, etc
- They are often violated for big data.
- Data scientists investigate these assumptions and propose solutions.

Display numerical data

Displaying categorial variable

- For categorical variable, we summarize the data using the **counts** of observed occurrences of each value.
- Alternatively, we can use ***percentage*** or *proportion*.

Pie chart

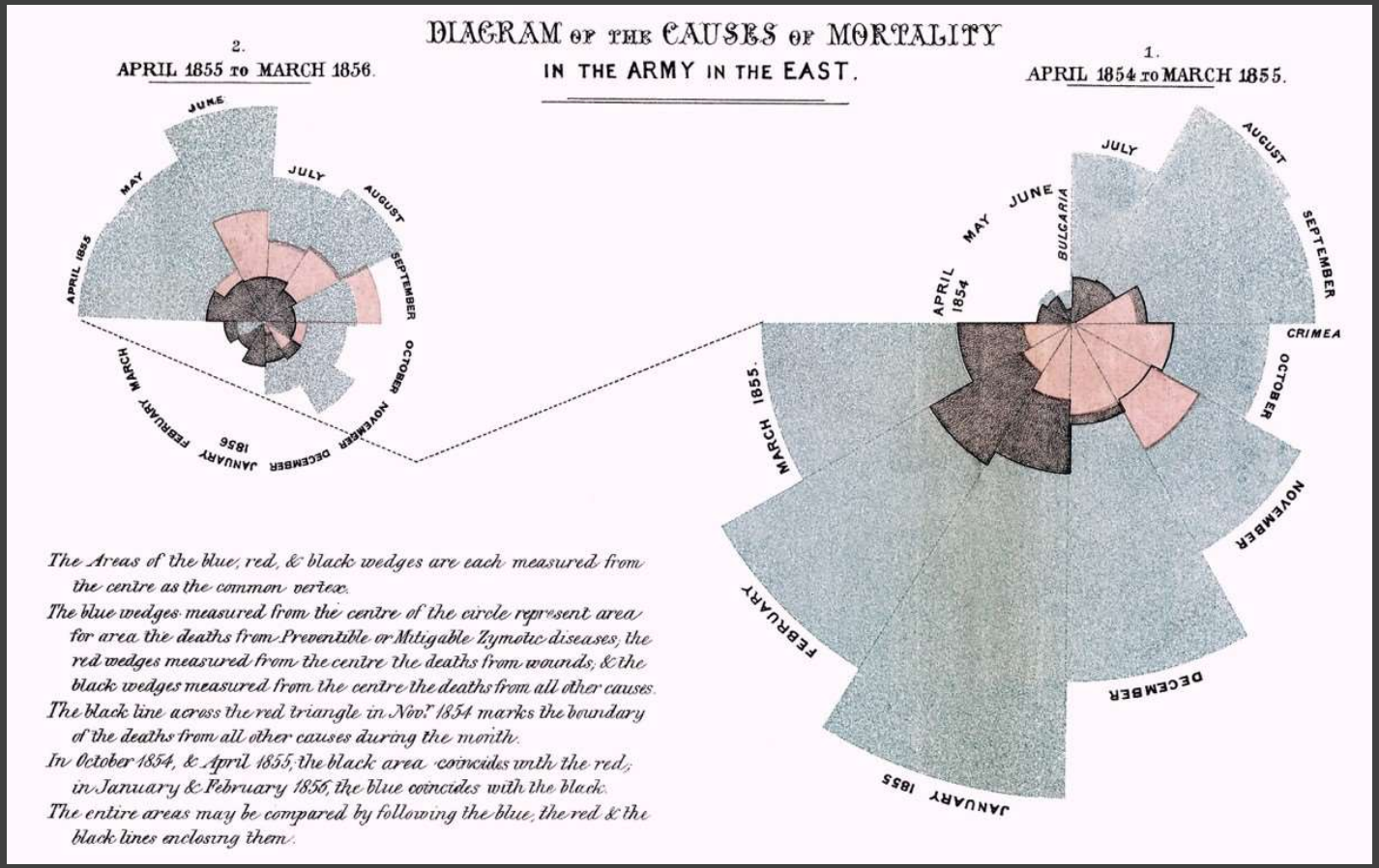
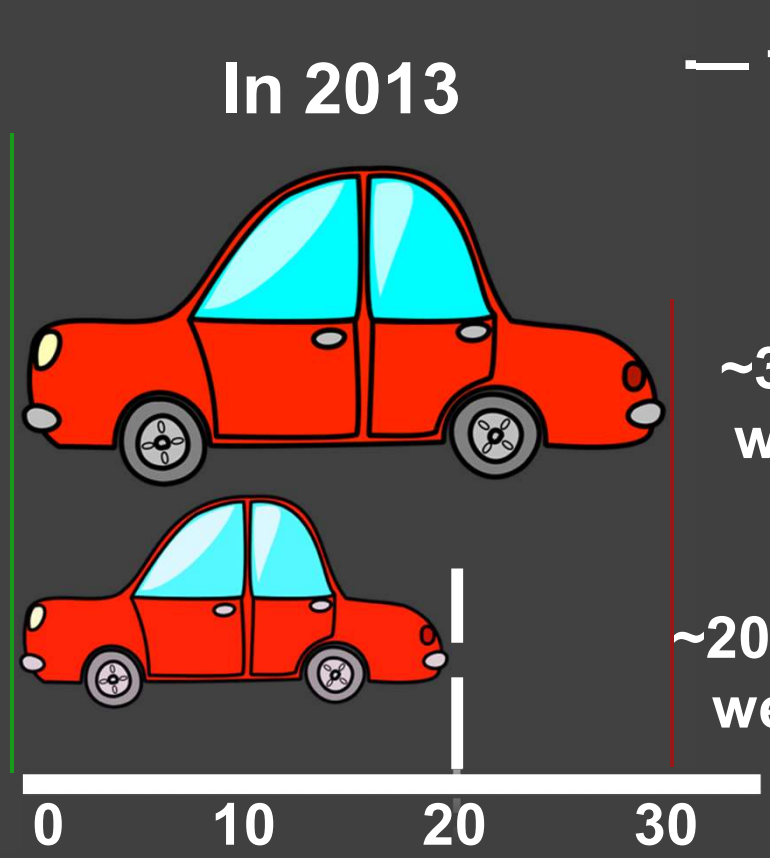


DIAGRAM of the CAUSES of MORTALITY
IN THE ARMY IN THE EAST.
2.
APRIL 1855 to MARCH 1856.
1.
APRIL 1854 to MARCH 1855.
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY 1856
FEBRUARY
MARCH
APRIL 1855
MAY
APRIL 1854
MAY
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY 1855
FEBRUARY
MARCH 1855
BULGARIA
CRIMEA
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Area principle

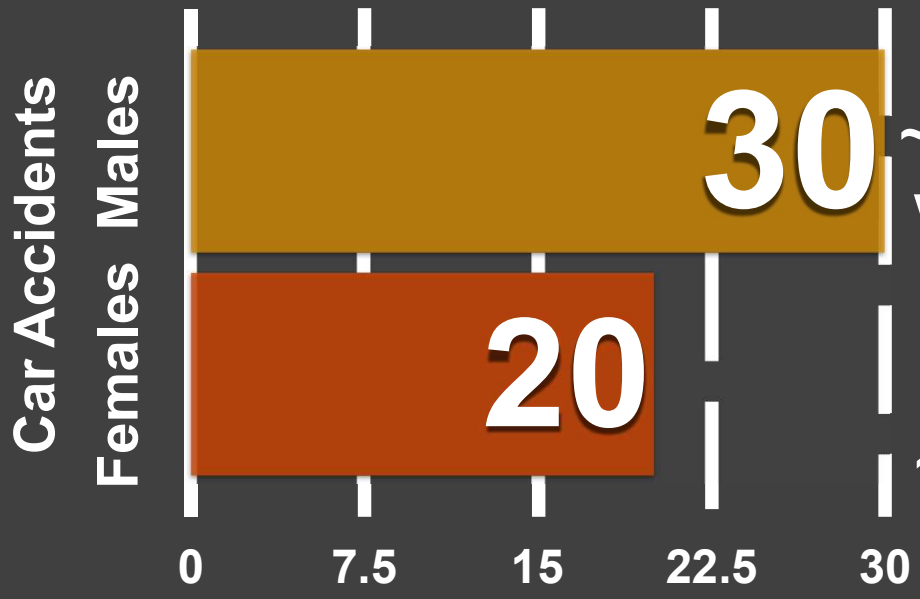


The area principle — the size of the area correlates with the data summaries.

~30% of accidental deaths of males were due to automobile accidents.

~20% of accidental deaths of females were due to automobile accidents.

Area principle

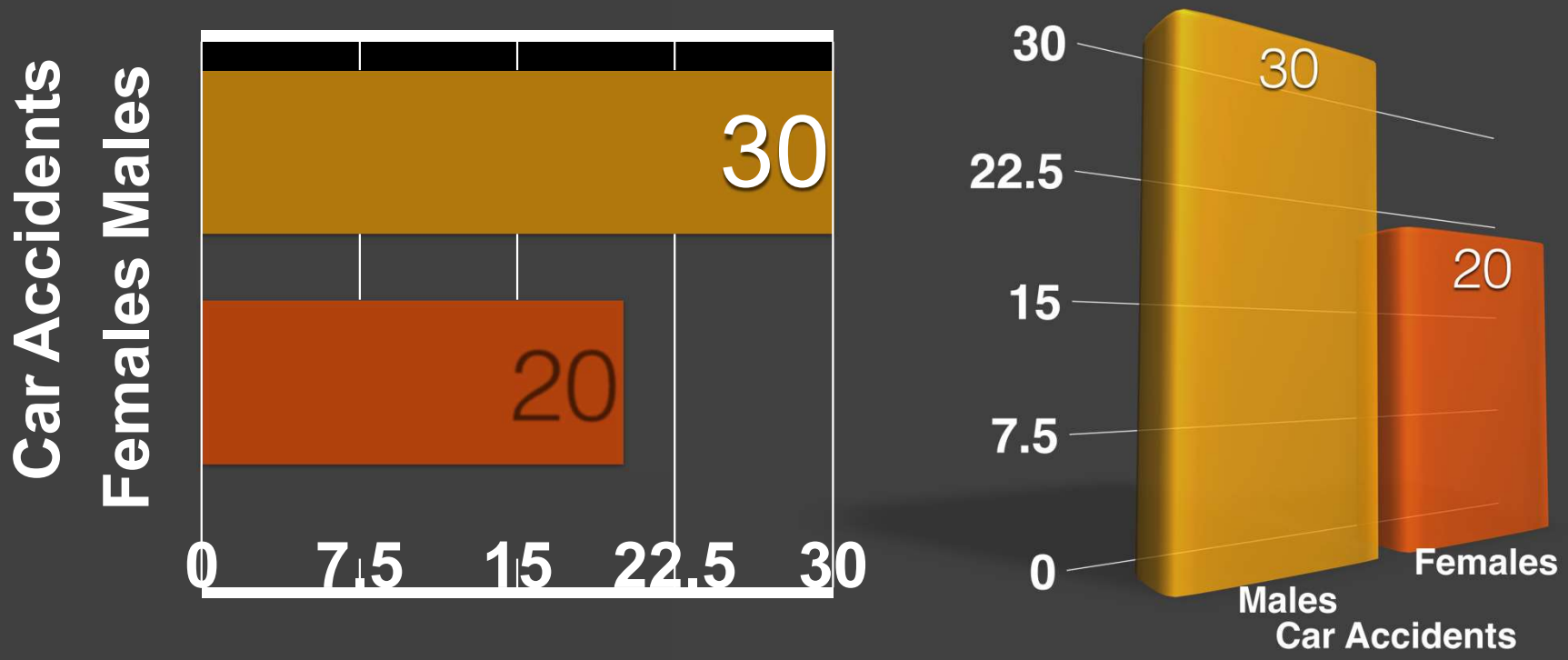


The area principle — the size of the area correlates with the data summaries.

~30% of accidental deaths of males were due to automobile accidents.

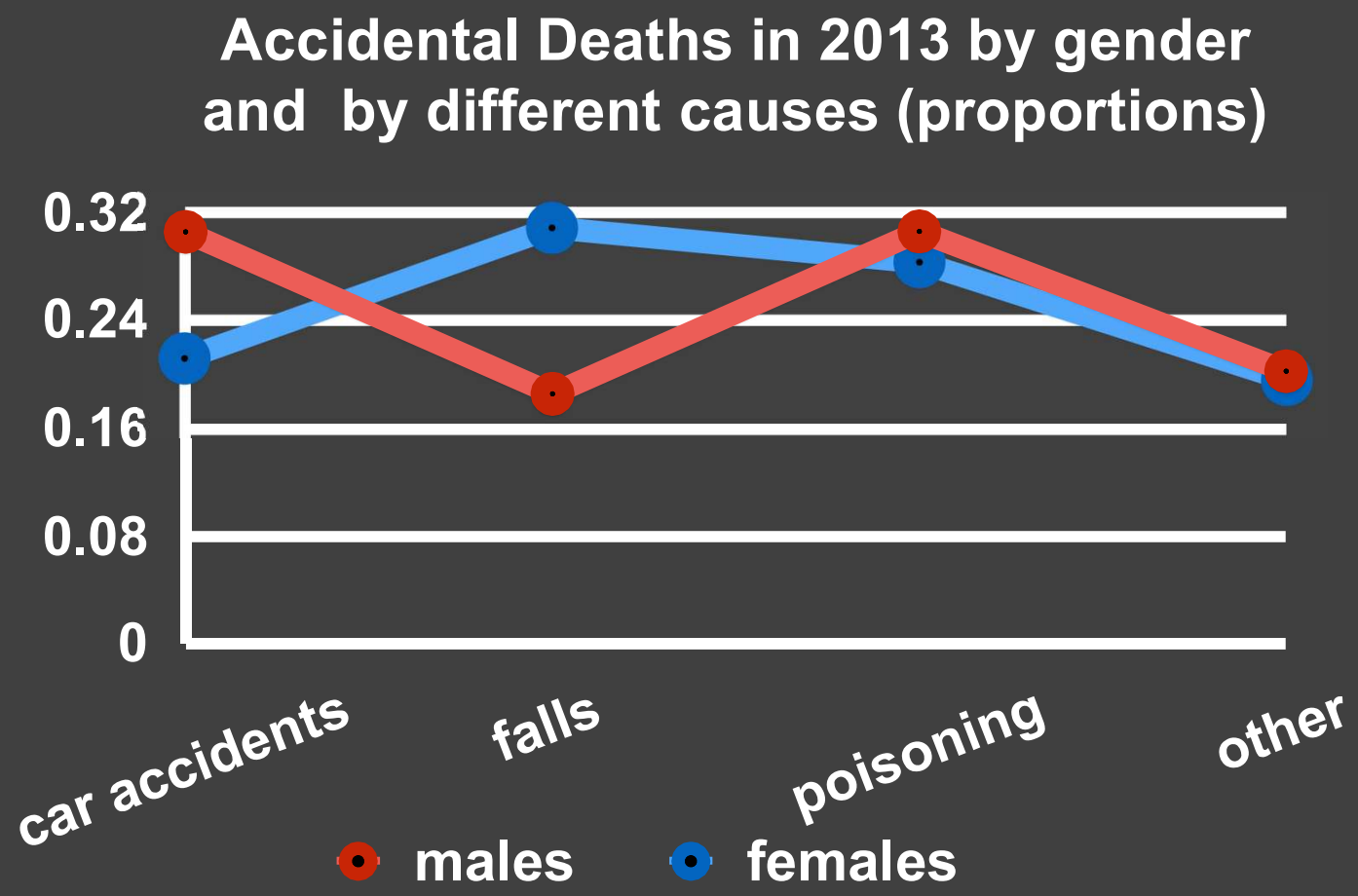
~20% of accidental deaths of females were due to automobile accidents.

3D effects?

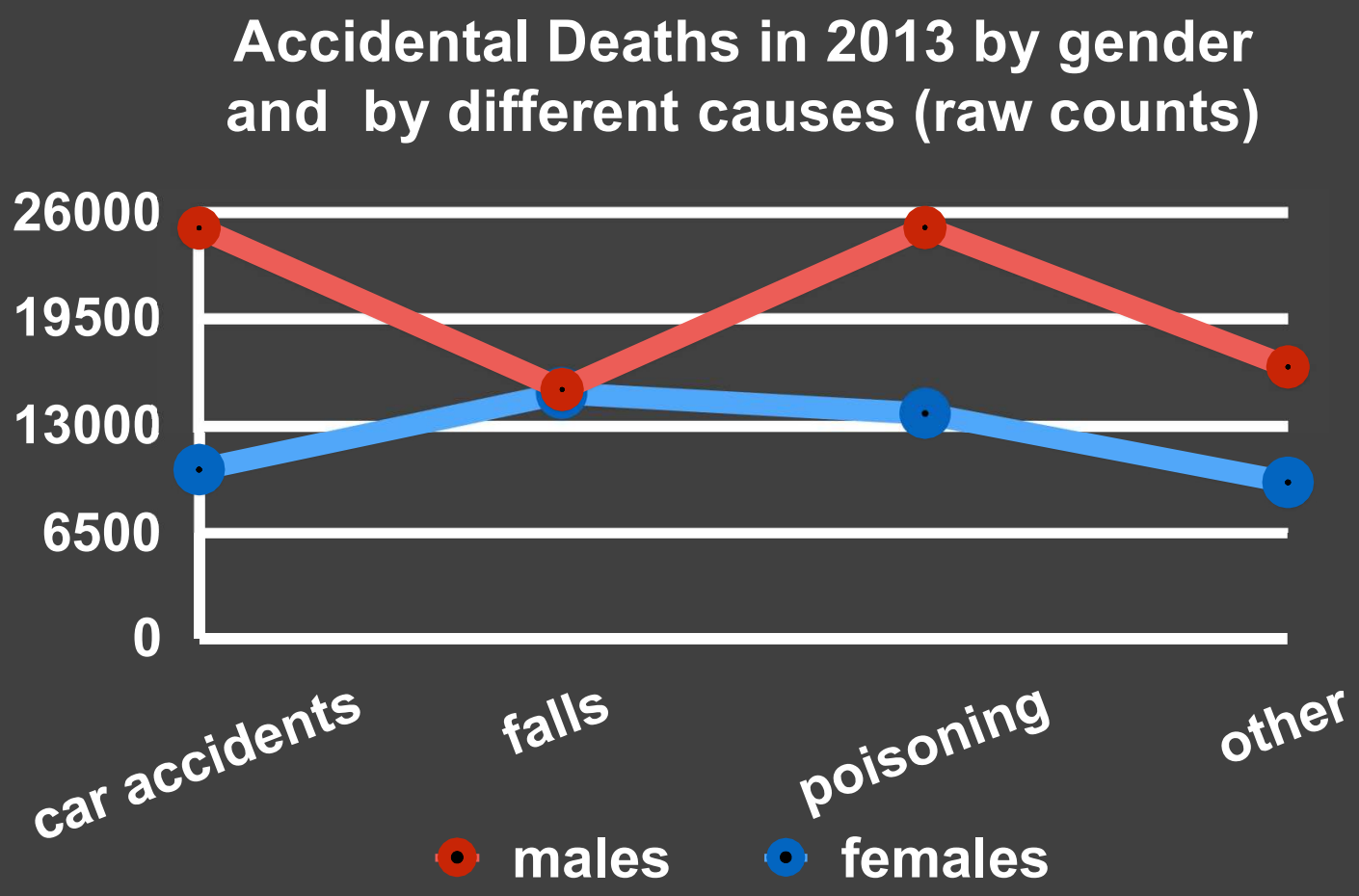


3D effects distort the visualization and often violate the area principle.

Side by side comparison



Create Meaningful visualization



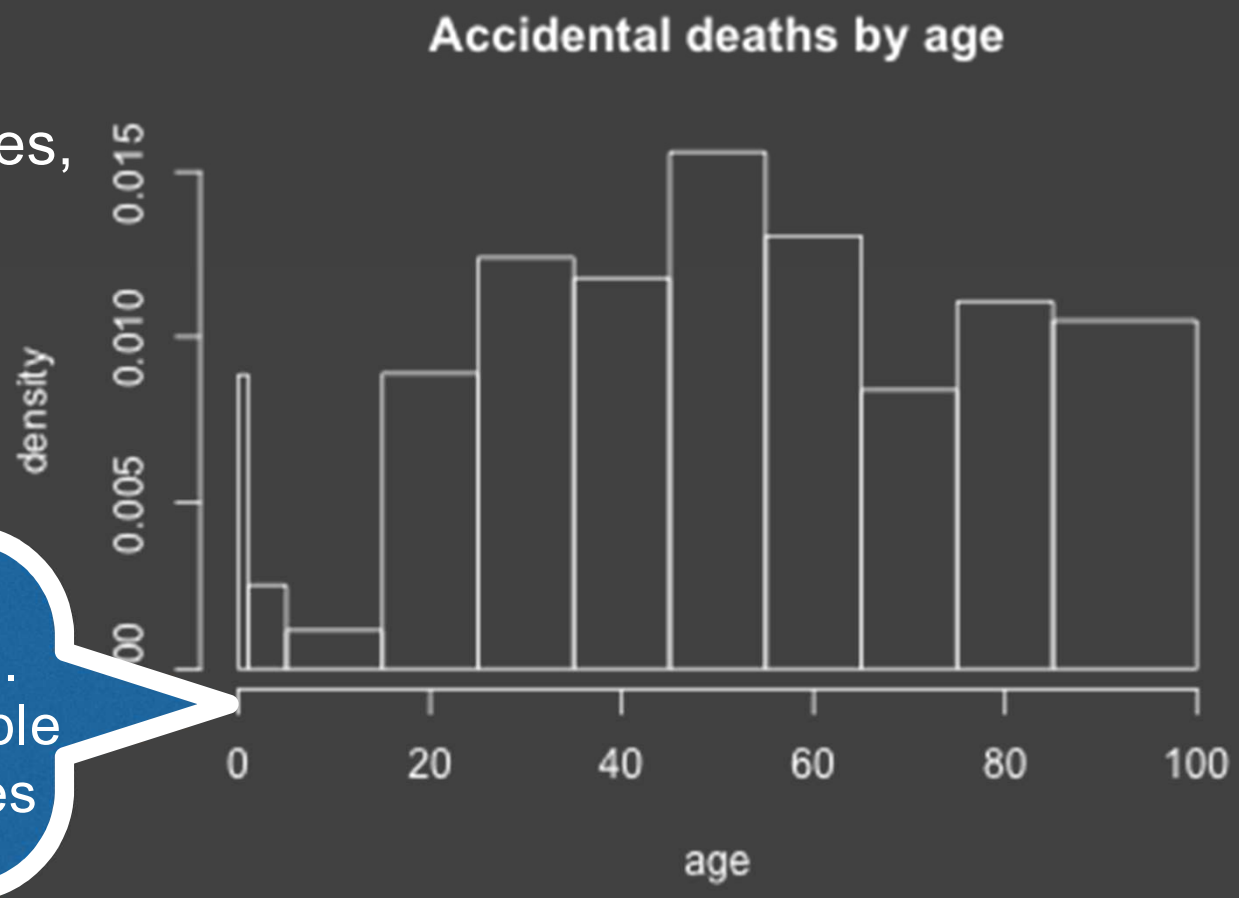
Displaying quantitative variable

- For quantitative variables, we also summarize the data using the counts of observed occurrences of values.
- Different from categorical variables, we may count occurrences **within intervals** rather than individual values.
- We also use percentage or proportion.

Displaying quantitative variable

Area of each bar
(The height of the bar multiplied by the width of the interval = the propotion of the death in the corresponding age interval.

Different from categorical variables, we may count occurrences within intervals rather than individual values.

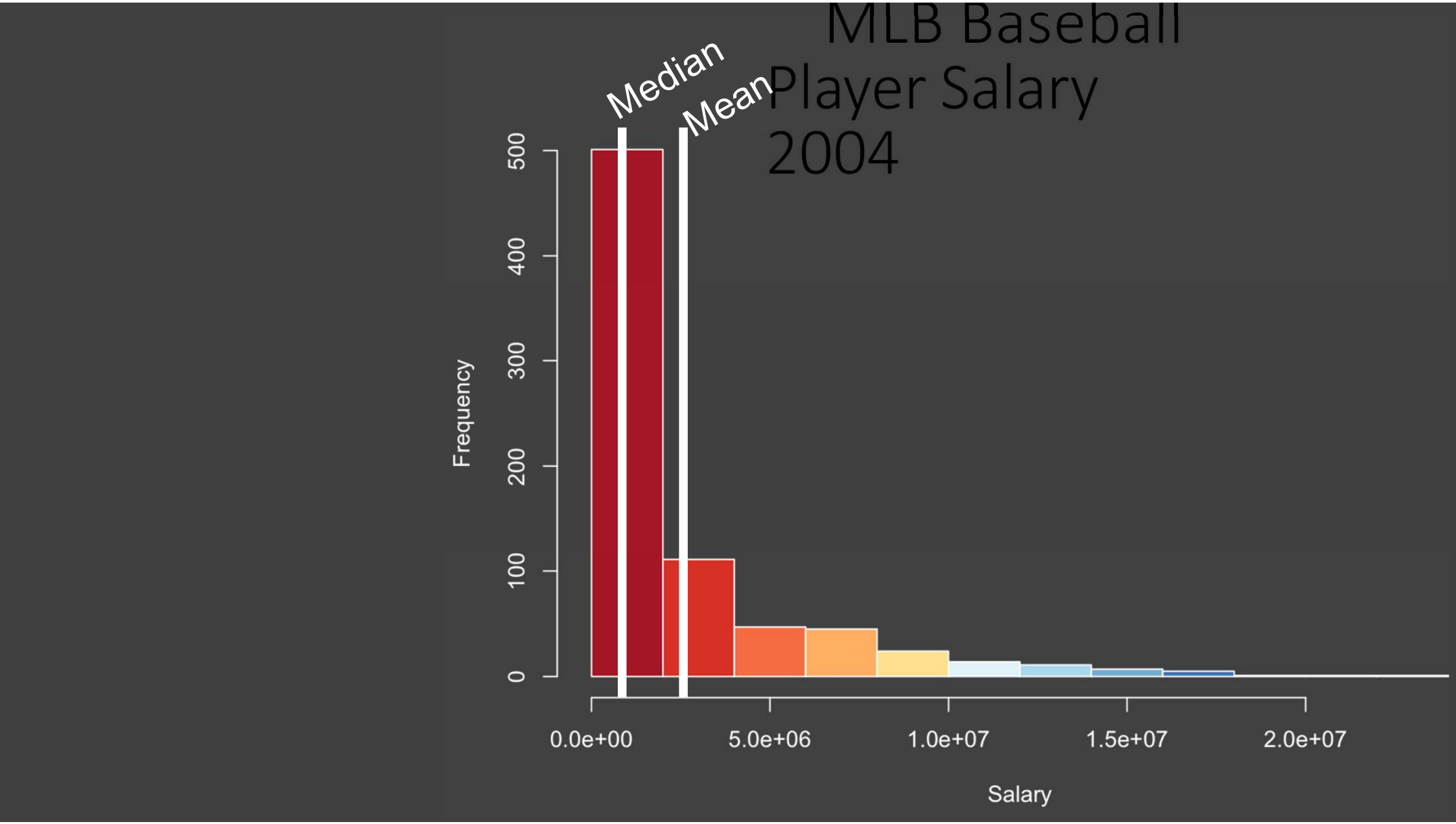


This is a histogram. Area principle also applies

Summarize numerical data

Center of Variation

- Summarizing center of variation:
  - **mean** (numerical average)
  - **median** (mid-point)
- When the data come with a few extremely large values
  - mean is more affected by them than median.
  - Sensitive to outliers.


MLB Baseball
Player Salary
2004
Median
Mean
Frequency
0
100
200
300
400
500
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Salary

Summarizing variation Standard Deviation

- For multiple observed values, **variation** is quantified by their **deviation from their center**.
- **Standard deviation**
  - deviation from the **sample mean**
  - the square root of variance—the average squared deviation.

Standard deviation is a parameter for normal distributions.

- It is used as a "**yard stick**" for variation.

- variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

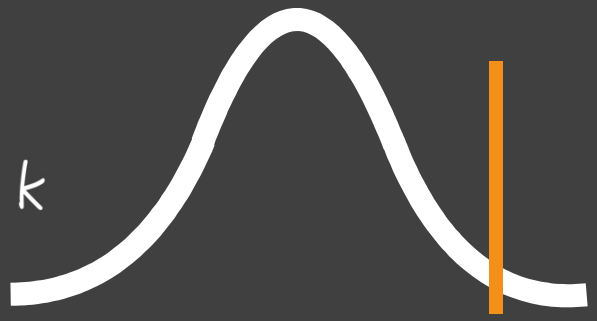
standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

- It ***standardizes*** variation to make
- random values from different variables comparable.

Standardization using mean and standard deviation

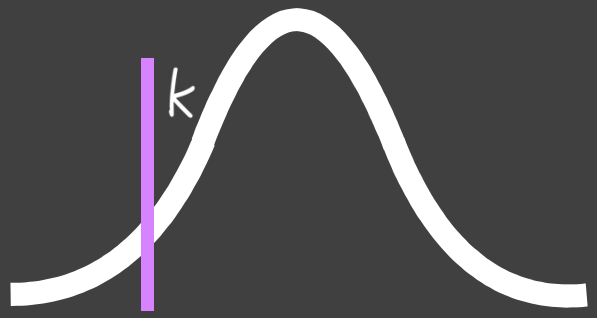
- X: a value observed
- We calculate **how many "standard deviations"** X is above/below from the mean

Standard deviation as a yard stick



Kim is making 25K a year, the income in his city has a mean of 20K and a standard deviation of 4K

Kim is 1.25 SD more than mean in his city



Lee is also making 25K a year, the income in his city has a mean of 30K and a standard deviation of 5K

Lee is 1 SD below the mean of his city.

Summarizing variation Quantiles

- Quantiles (or percentile): a value threshold of a variable that is defined to have a percent of data below it.
  - SAT critical reading, a score 600 is the 79th percentile.
- A set of special percentiles are called **quartiles**, which corresponds to 25%, 50% and 75% percentiles.
  - Quartiles divide data into quarters