contributed by: jjunge@princeton.edu

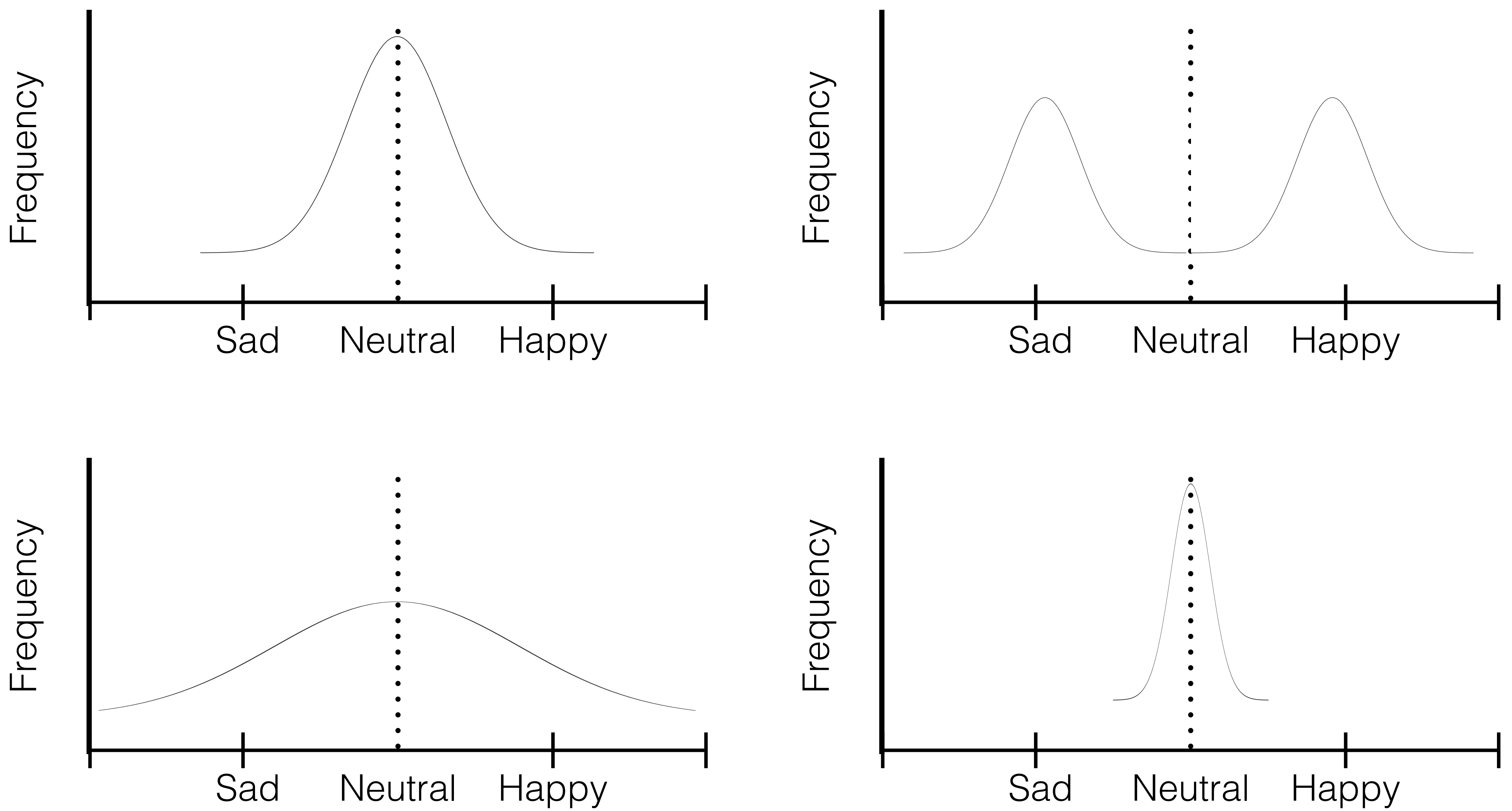
← continuous categorical →

← Scales →

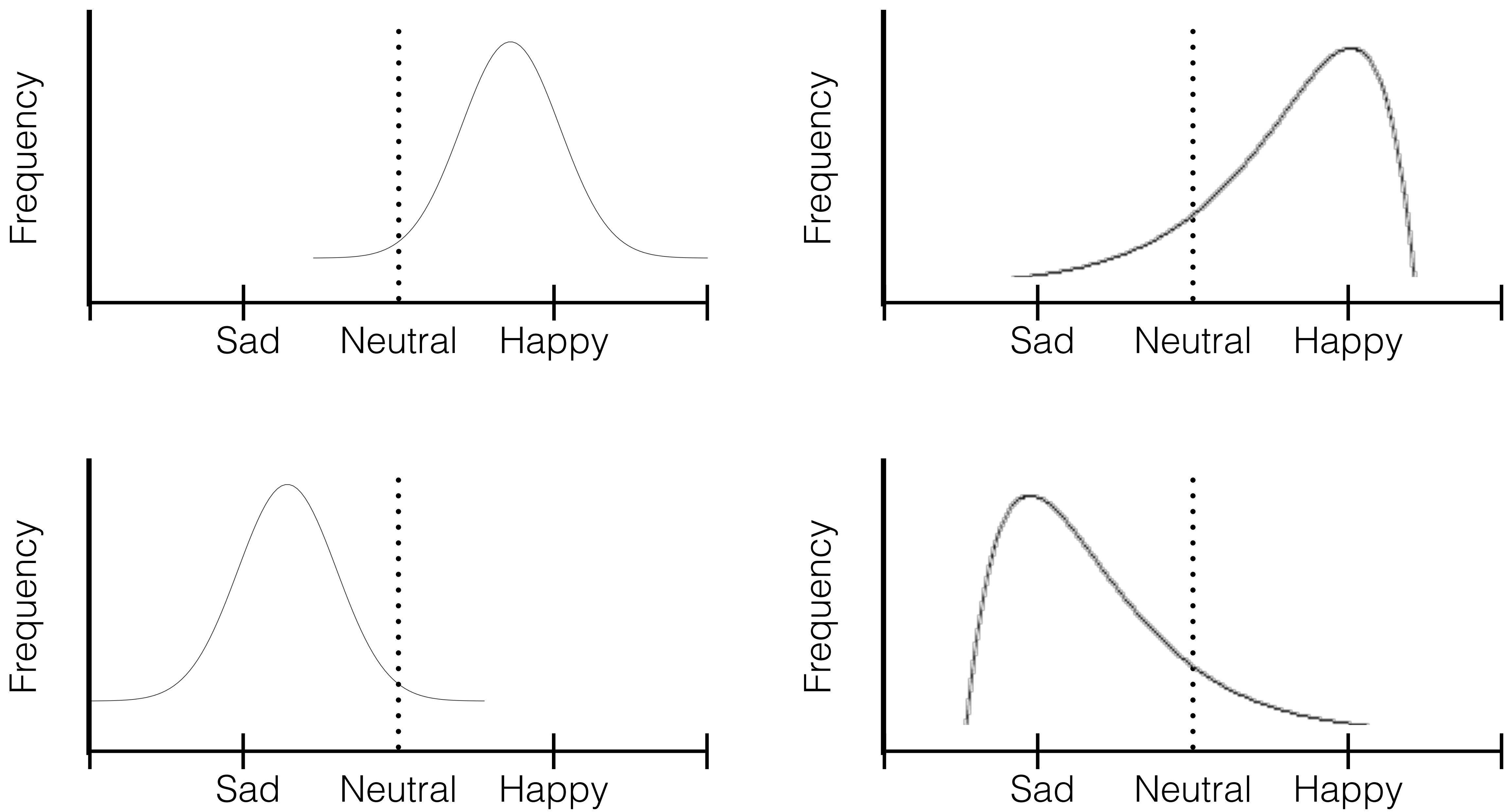
Ratio → Interval → Ordinal → Nominal

Ratio	Interval	Ordinal	Nominal
Interpret 0 **Magnitude**	**Quantity** **Mean (average)** **Comparison**	**Order**	**Frequency**
Kelvin	Celsius	Cold Cool Warm Hot	Comfortable/ Uncomfortable

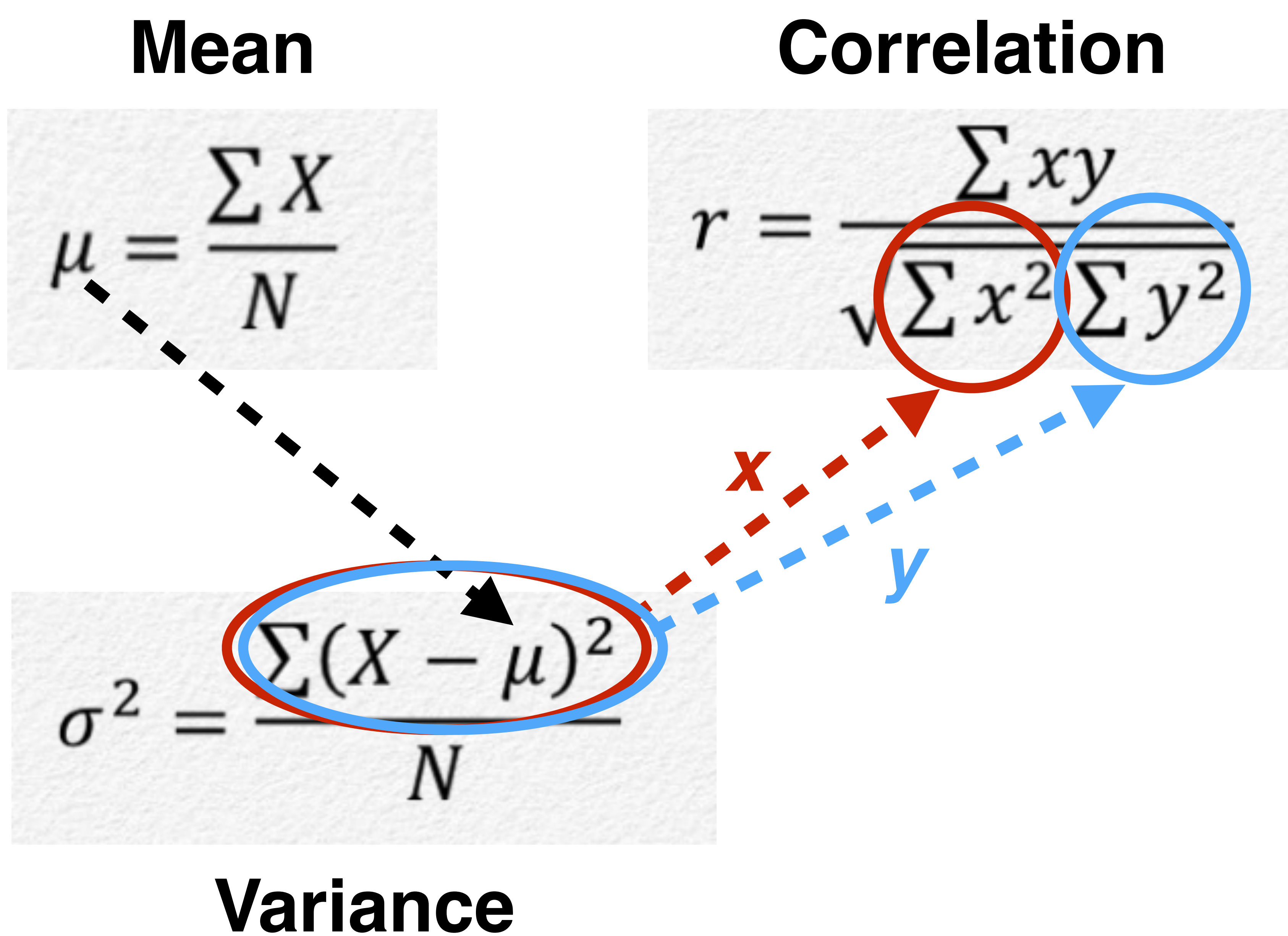
Distributions Personified



Distributions Personified



Equations


Mean
$\mu = \frac{\sum X}{N}$
Correlation
$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
x
y
$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
Variance

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

20

10 40

$$\frac{20}{\sqrt{400}}$$

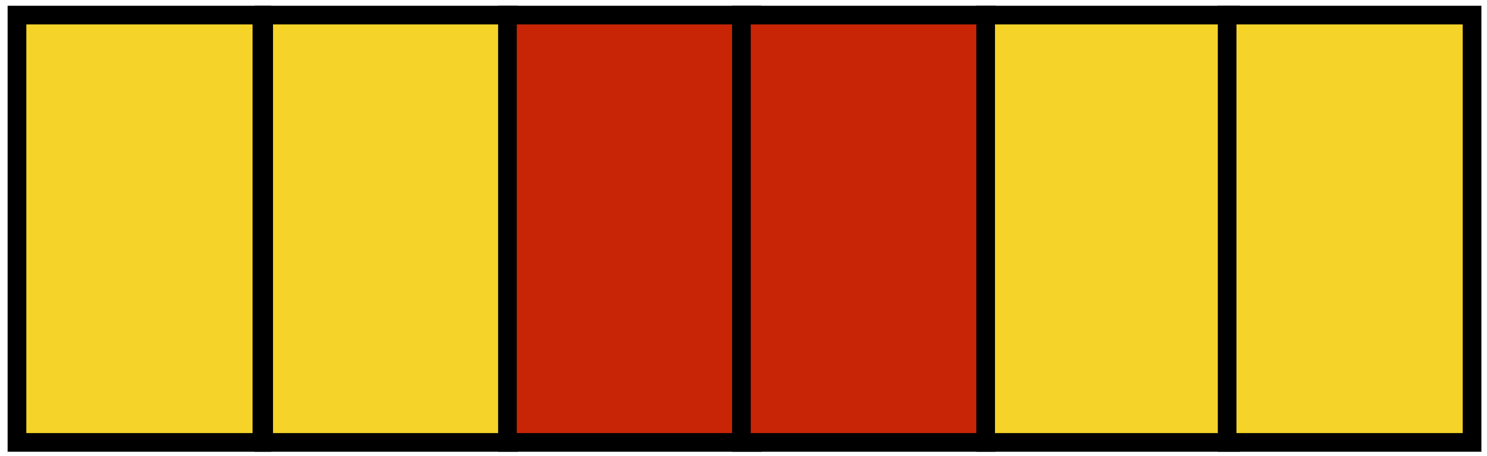
Vary Together

Total Variance

x^2	x		X	Y		y	y^2
4	-2	3	1	2	6	-4	16
1	-1	3	2	4	6	-2	4
0	0	3	3	6	6	0	0
1	1	3	4	8	6	2	4
4	2	3	5	10	6	4	16
10							40

x	y	xy
-2	-4	8
-1	-2	2
0	0	0
1	2	2
2	4	8
		20

One Event (Either)

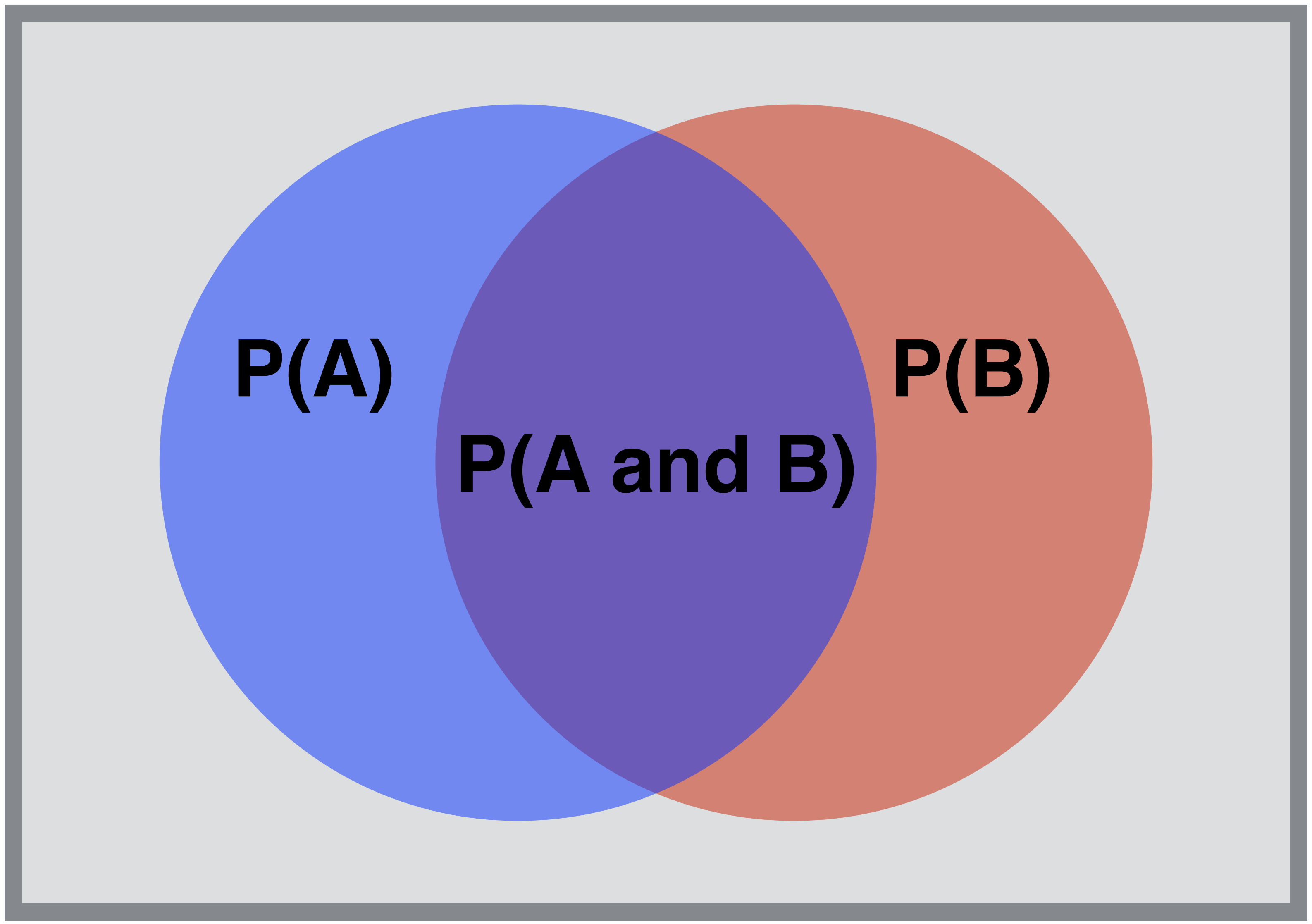


Either  or 

1/6 + 1/6 = **2/6** =1/3

Add

P(A or B)



P(A) = 1/3
P(B) = 1/3
P(A and B) = 1/9
P(notA and notB) = 4/9

P(A or B) = 1 - [P(notA) x P(notB)] = 5/9

2/3 2/3

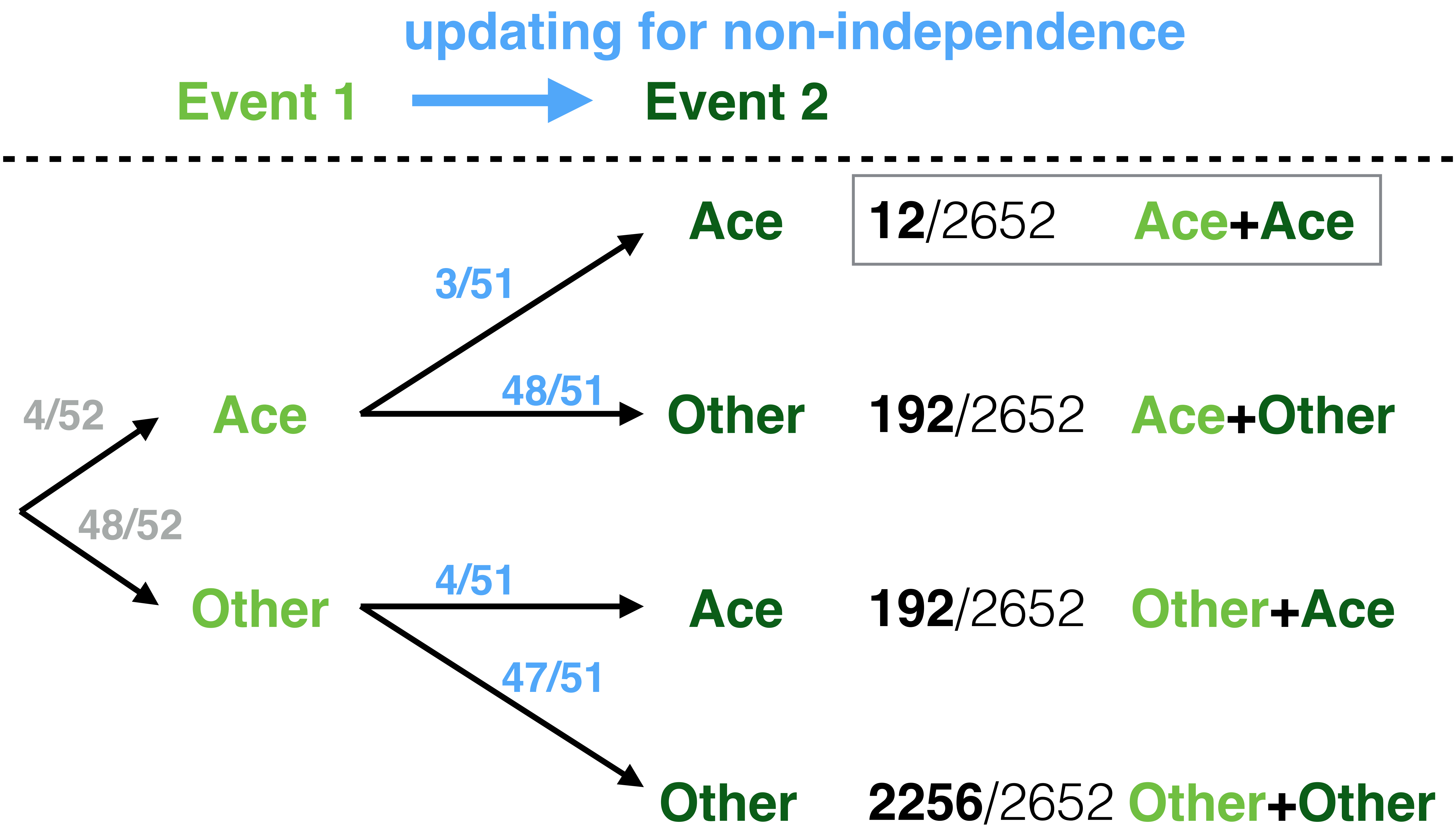
If	$0 \leq [a, b, c, d] \leq 1$	Probabilities
&	$a + b = 1$	Event 1
&	$c + d = 1$	Event 2
then	$\mathbf{ac + bc + ad + bd = 1}$	Event 1 & 2

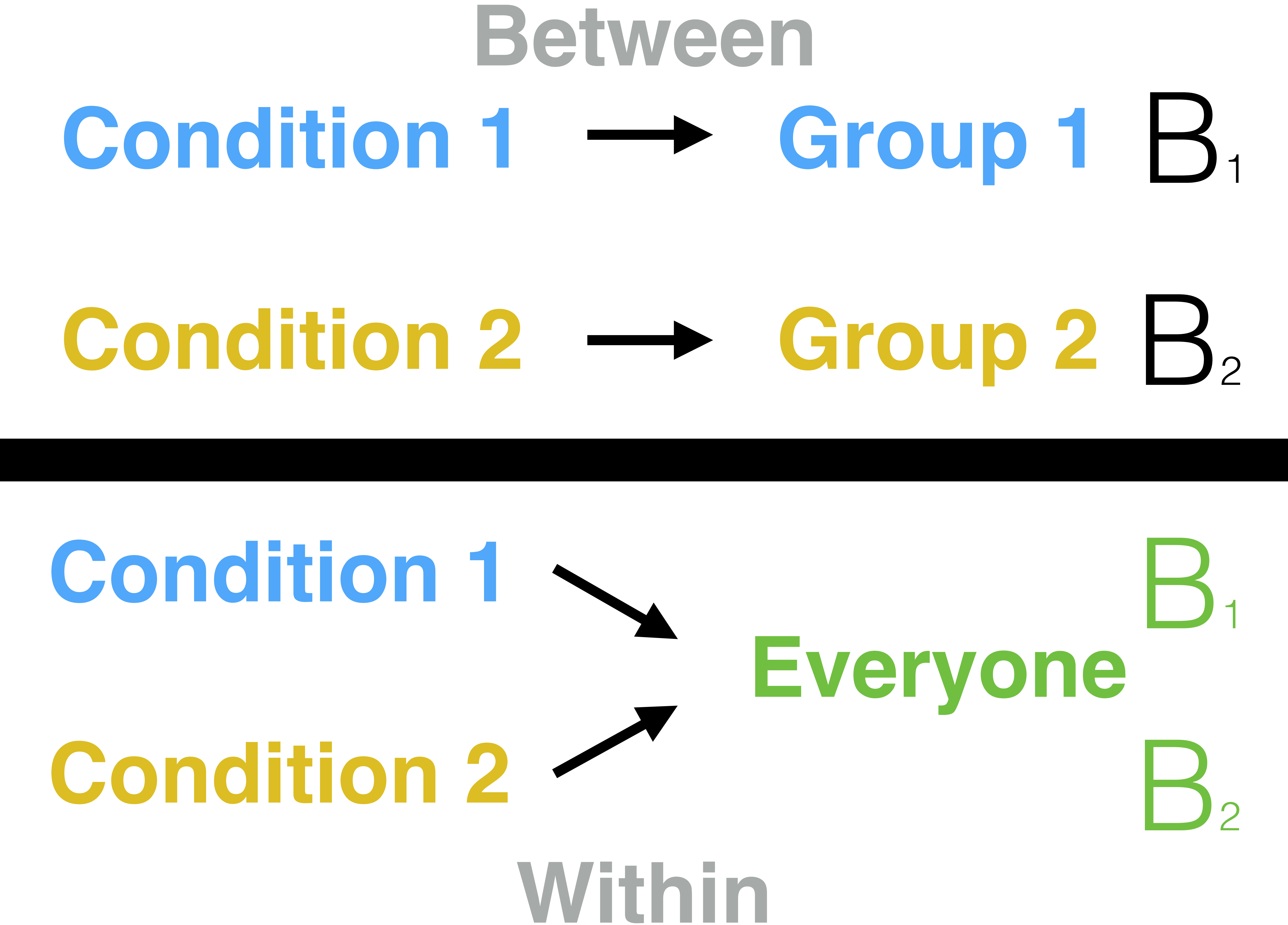
generalizes…

(a, b) & (c, d, e, f…)

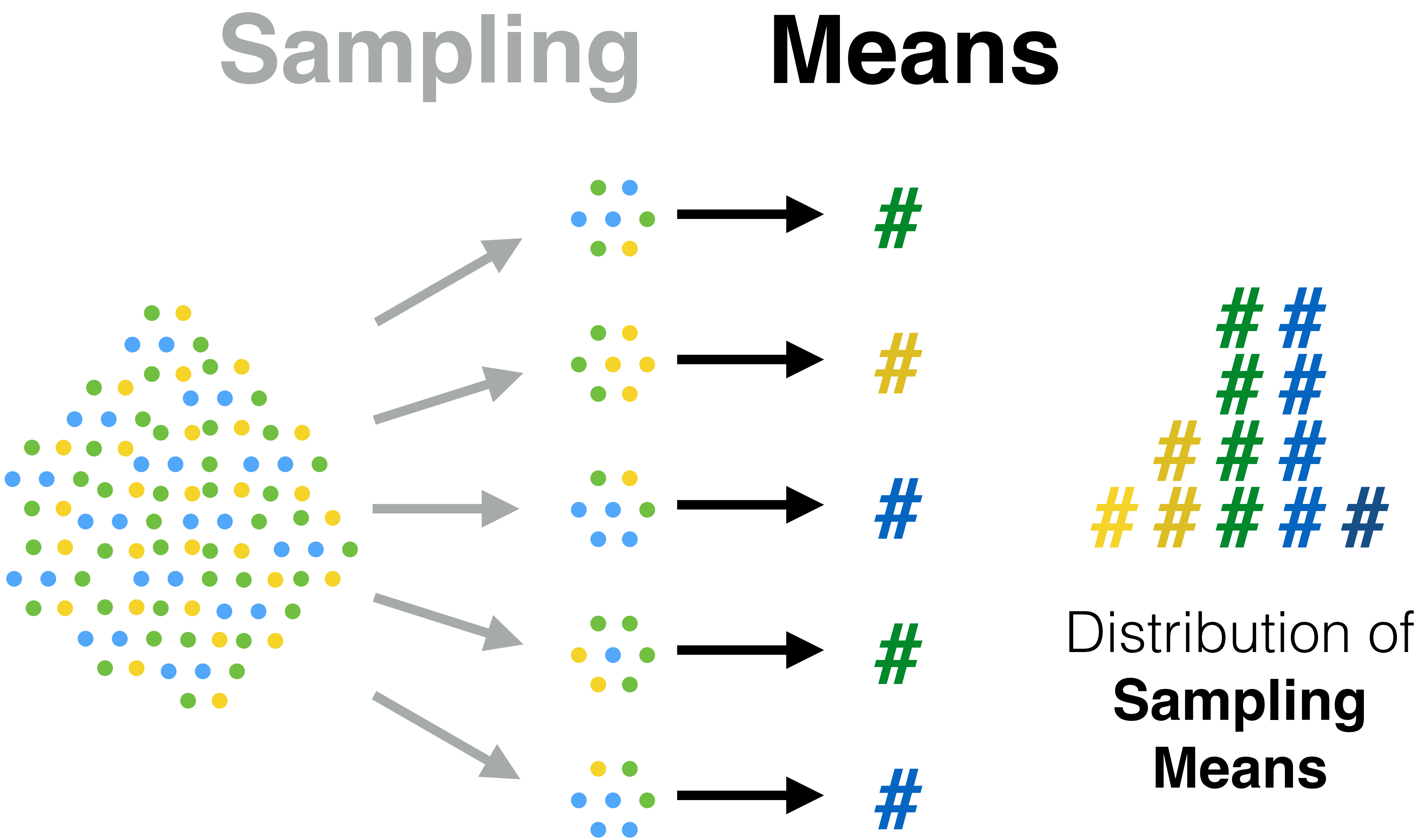
$ac + bc + ad + bd + ae + be + af + bf\ldots = 1$

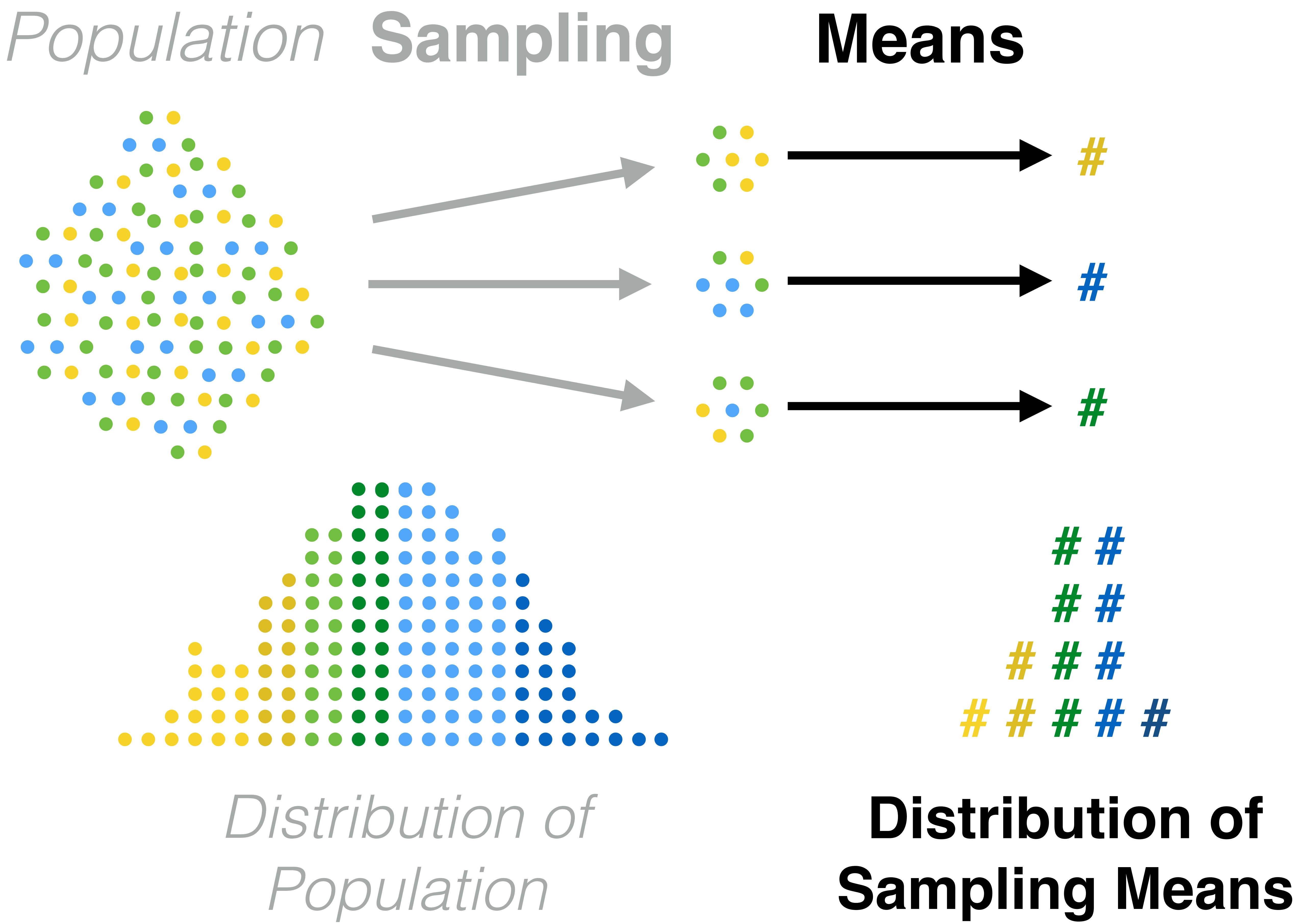
Sampling Without Replacement: If you pick two cards from a deck, what is the probability of drawing two Aces?



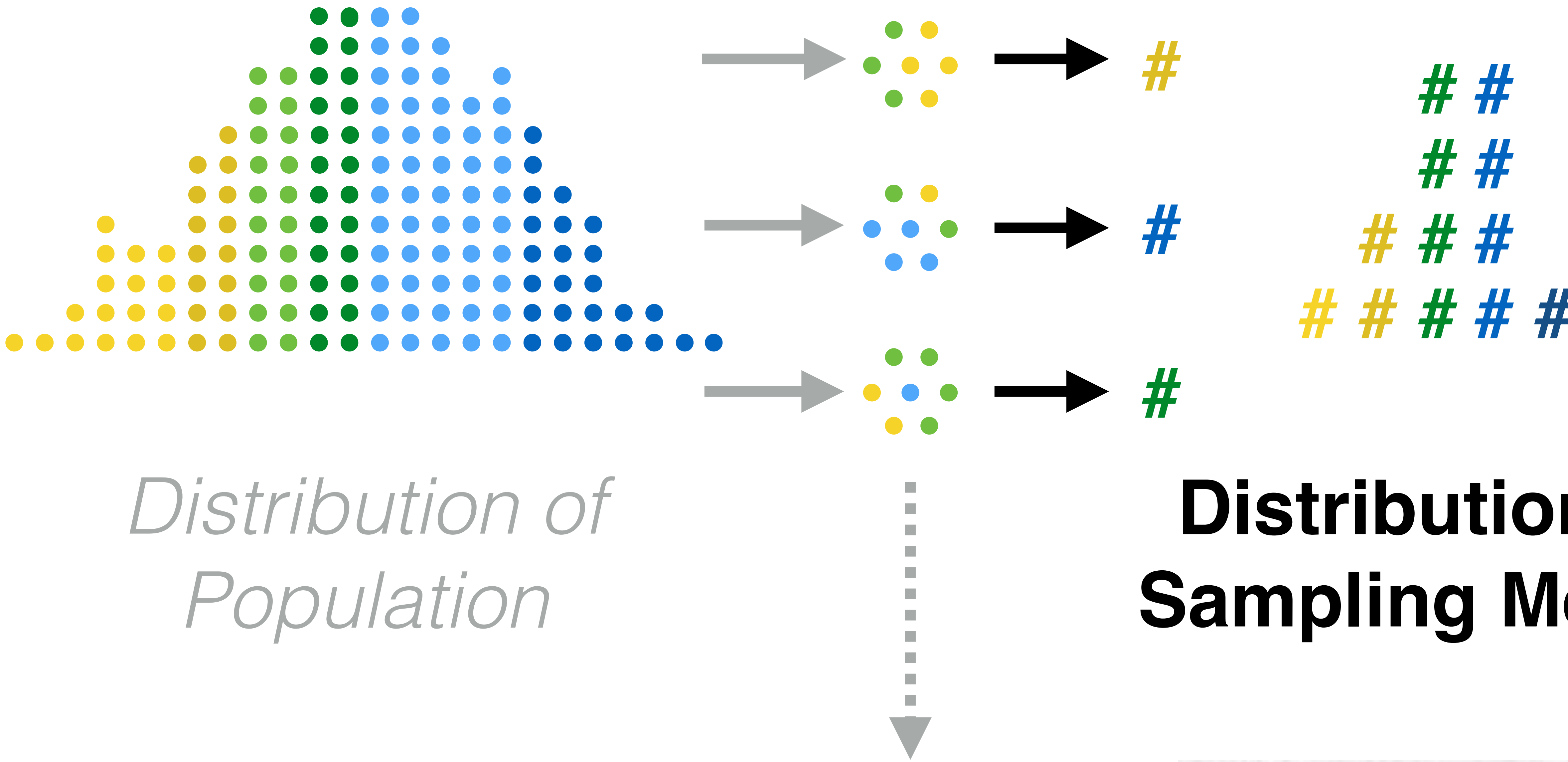

Between
Condition 1
Group 1
B1
Condition 2
Group 2
B2
Condition 1
Condition 2
Everyone
B1
B2
Within

Population




Population
Sampling
Means
#
#
#
# #
# #
# # #
# # # # #
Distribution of Population
Distribution of Sampling Means

Population **Sampling** **Means**

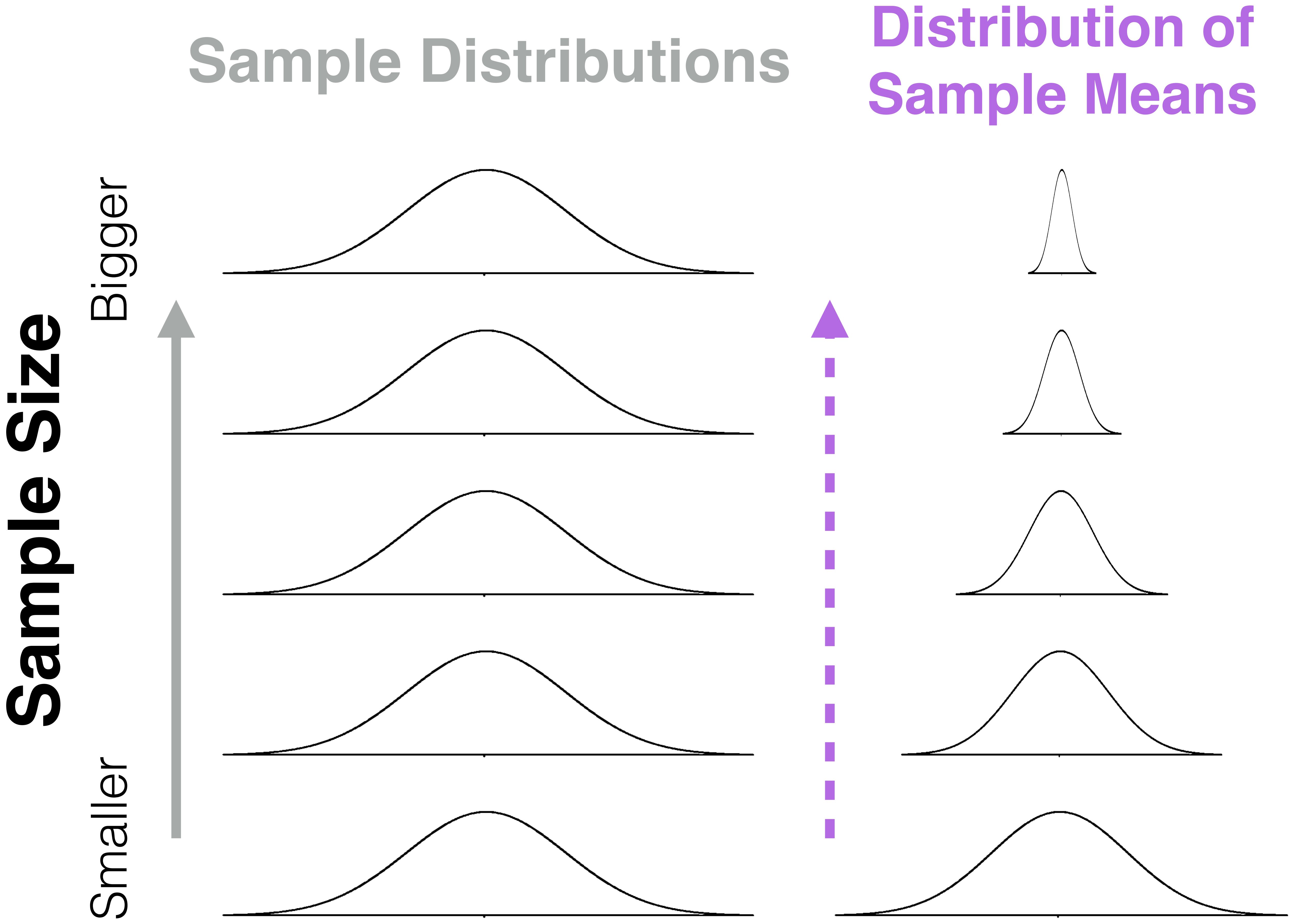


Distribution of Population

Distribution of Sampling Means

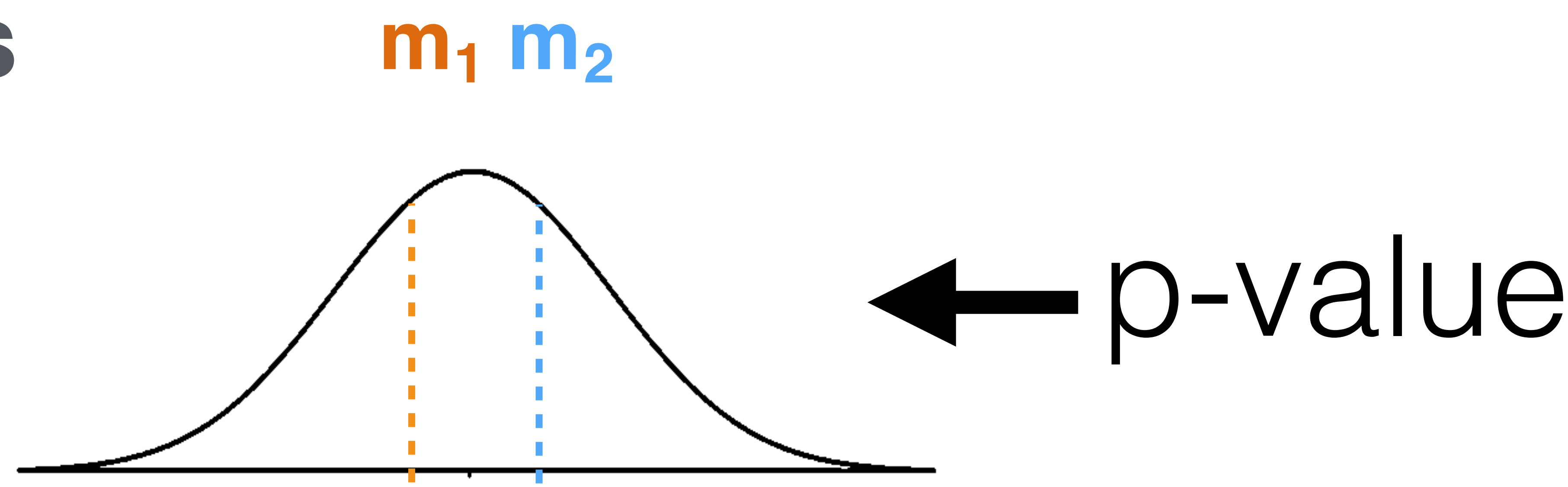
N = 7

$$\sigma_M = \frac{\sigma}{\sqrt{N}}.$$

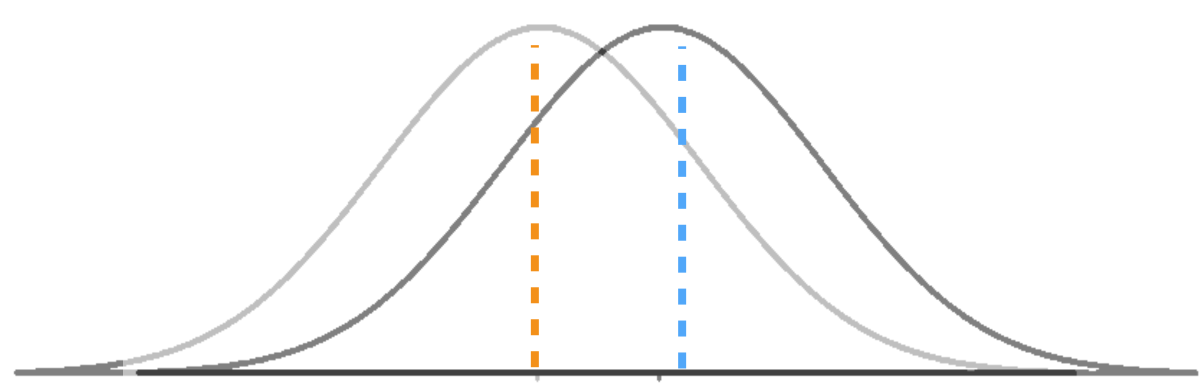


Sample Distributions
Distribution of Sample Means
Sample Size
Bigger
Smaller

Two Samples

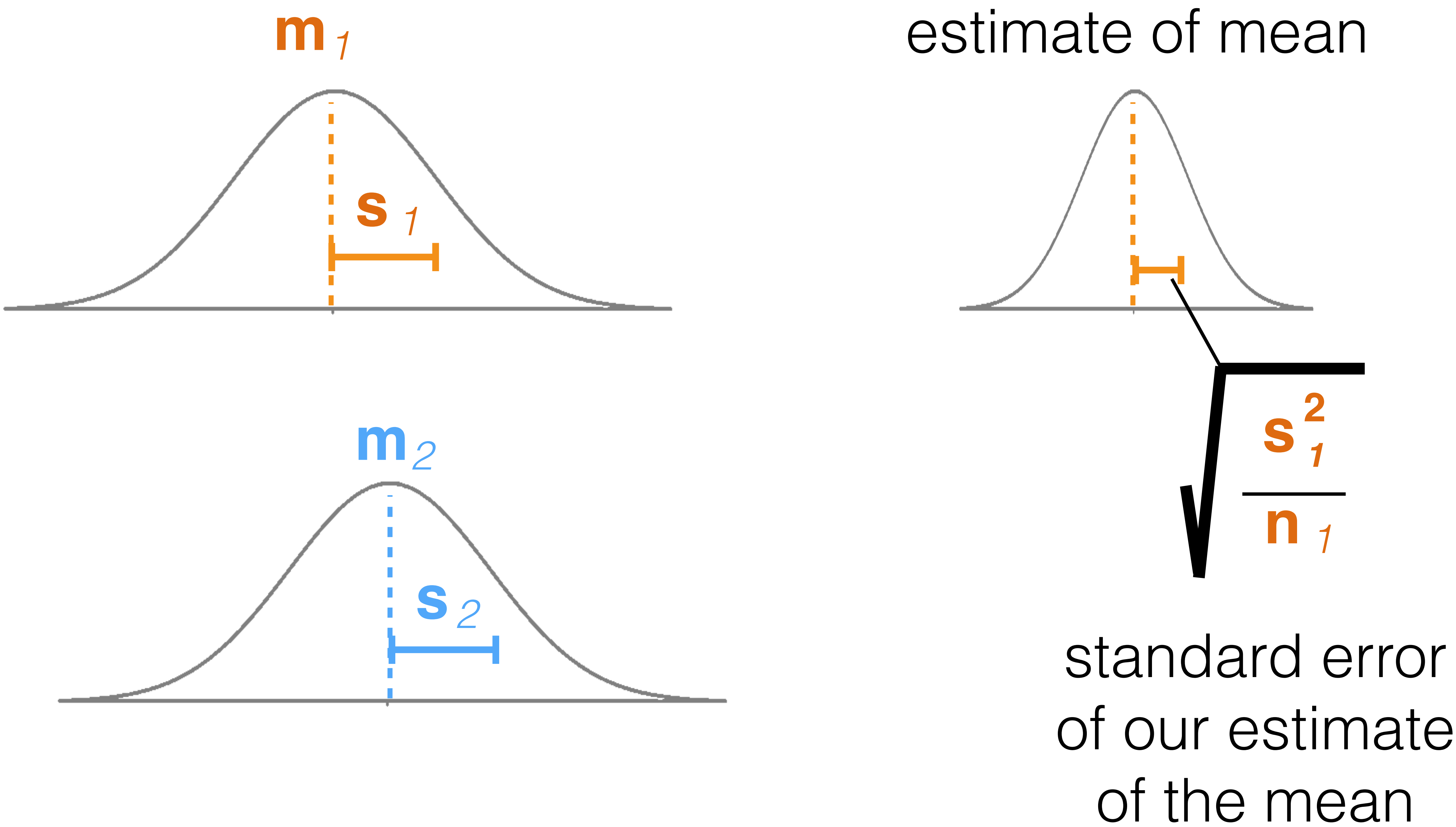


Random

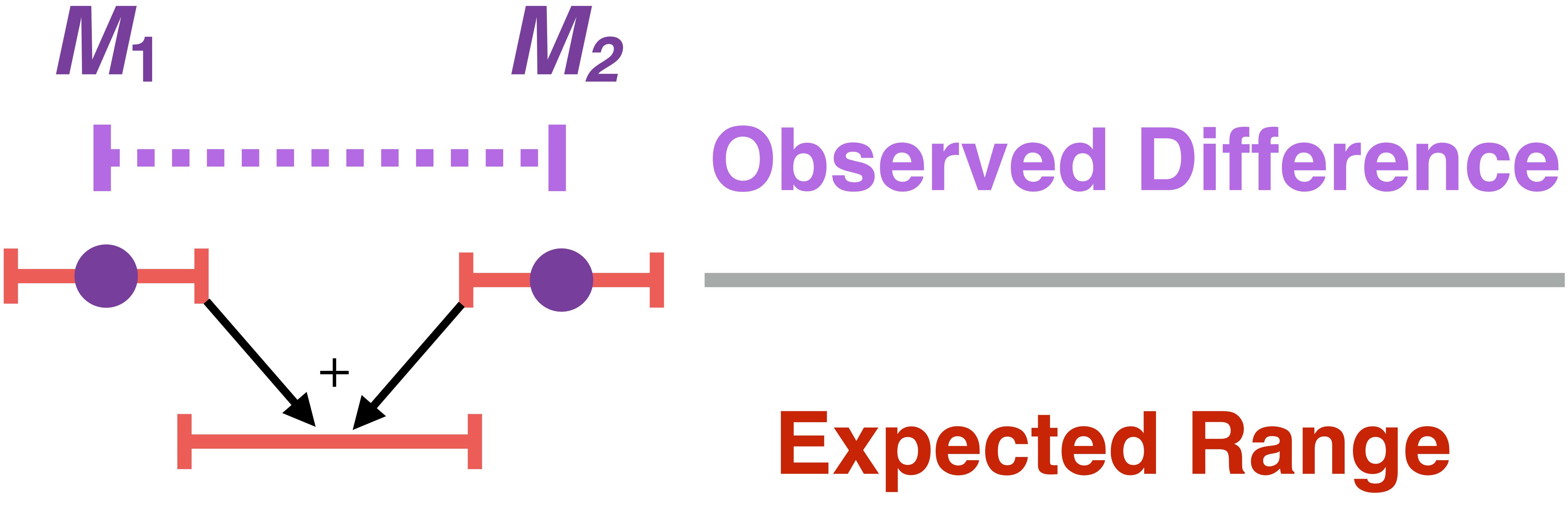


Real

Two Samples



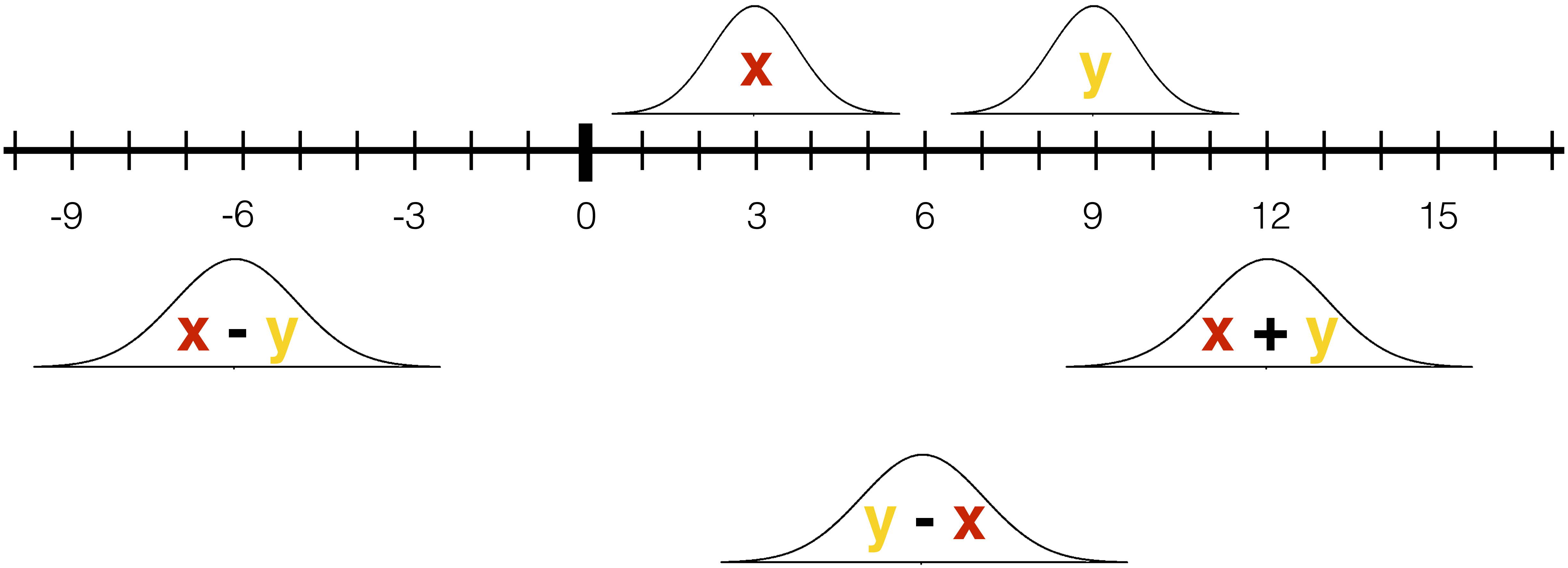
m_1
s_1
m_2
s_2
estimate of mean
$\sqrt{\frac{s_1^2}{n_1}}$
standard error
of our estimate
of the mean



M1
M2
+
Observed Difference
Expected Range

Variance Sum Law I

$$s^2_{X \pm Y} = s^2_X + s^2_Y$$



Two Independent Samples, Equal Variance

$M_1 - M_2$

0 ← H_0 Null

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\text{standard error of the statistic}}$$

$$\sqrt{\frac{2s^2}{n}}$$

$$s_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}}$$

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

$$s_{X \pm Y}^2 = s_X^2 + s_Y^2$$

Variance Sum Law I

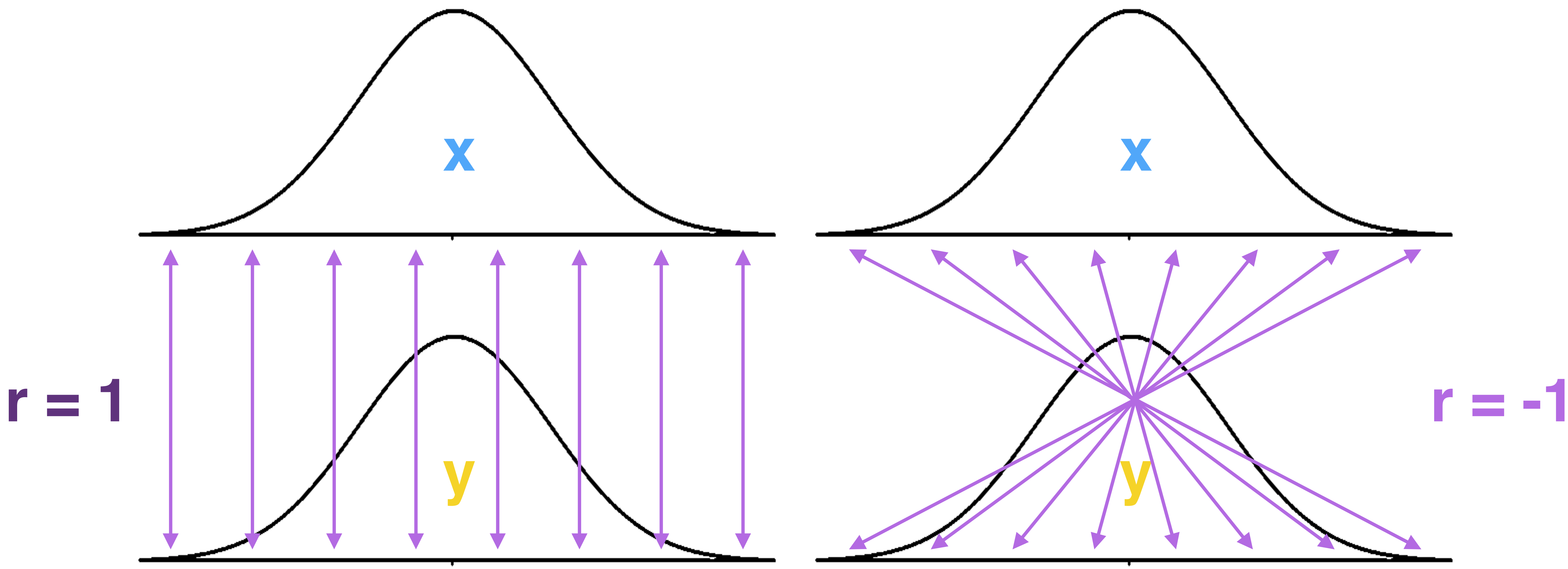
Variance Sum Law II

$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2 r s_X s_Y$$

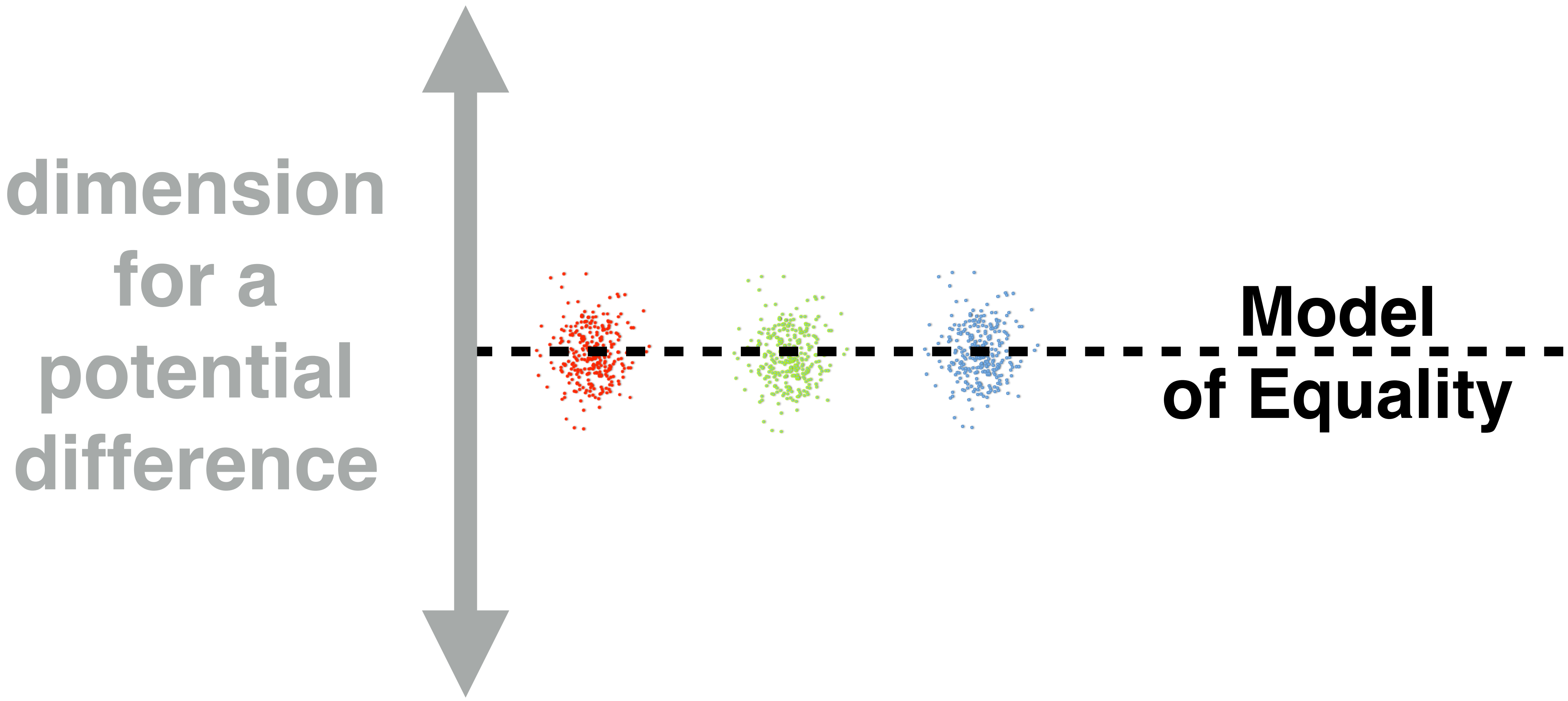
r = 1

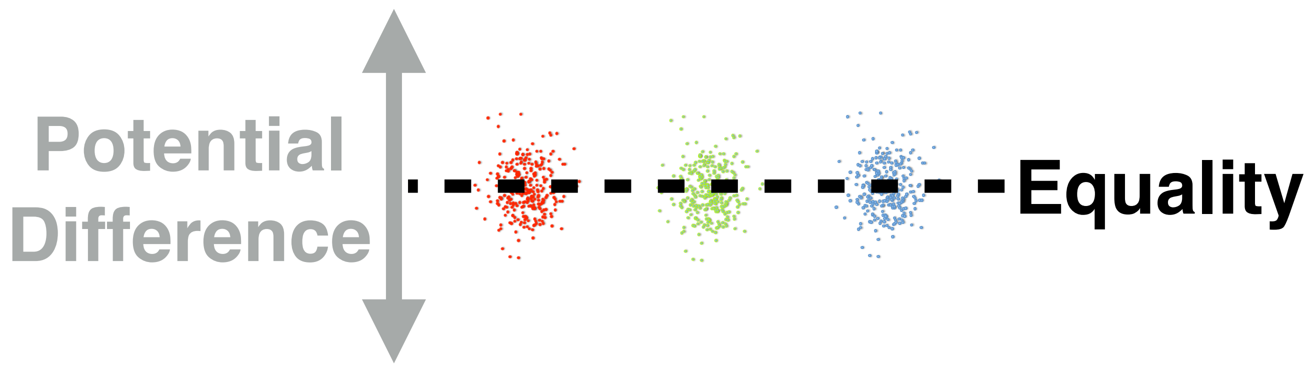
r = 0

r = -1

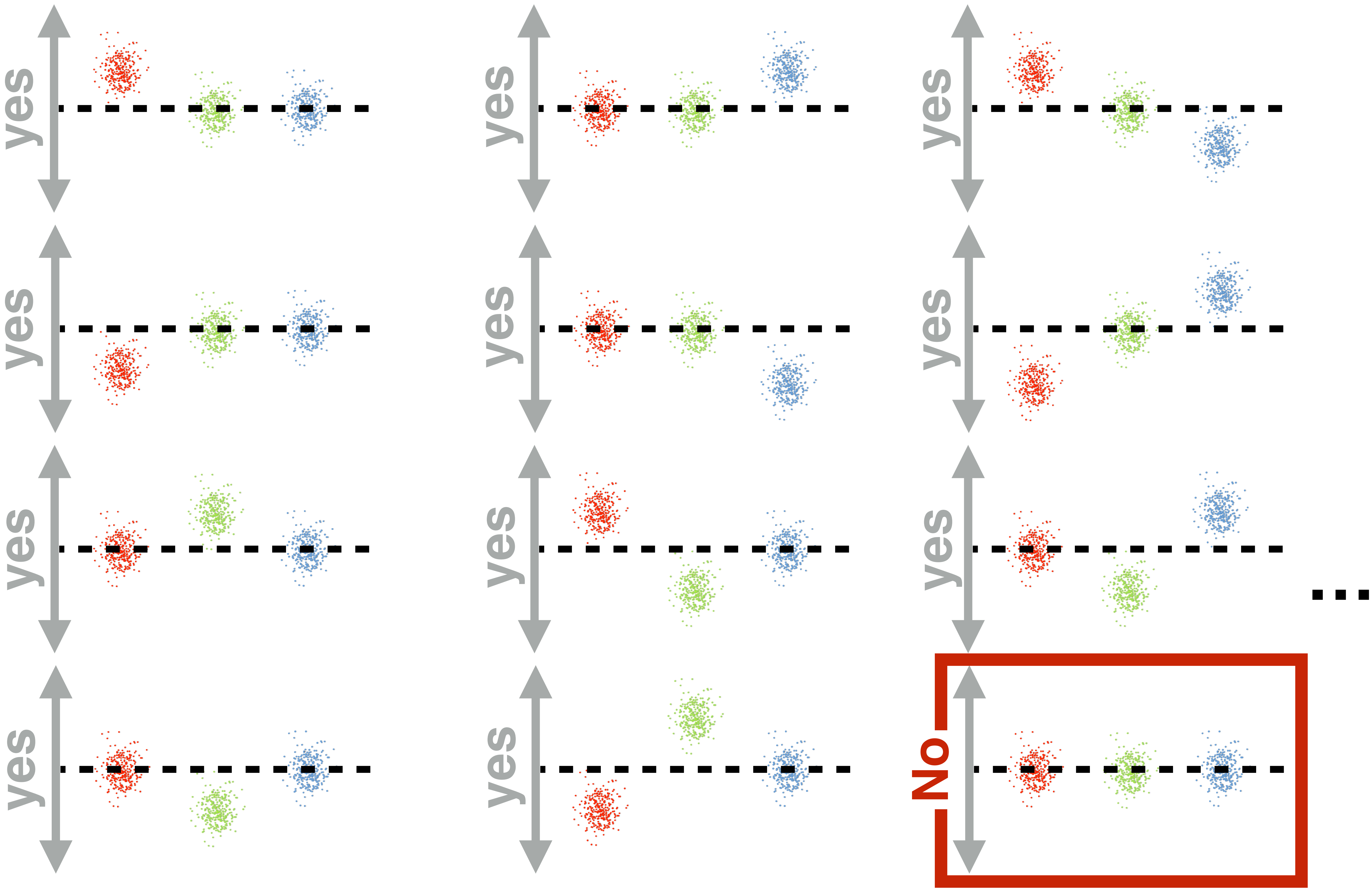


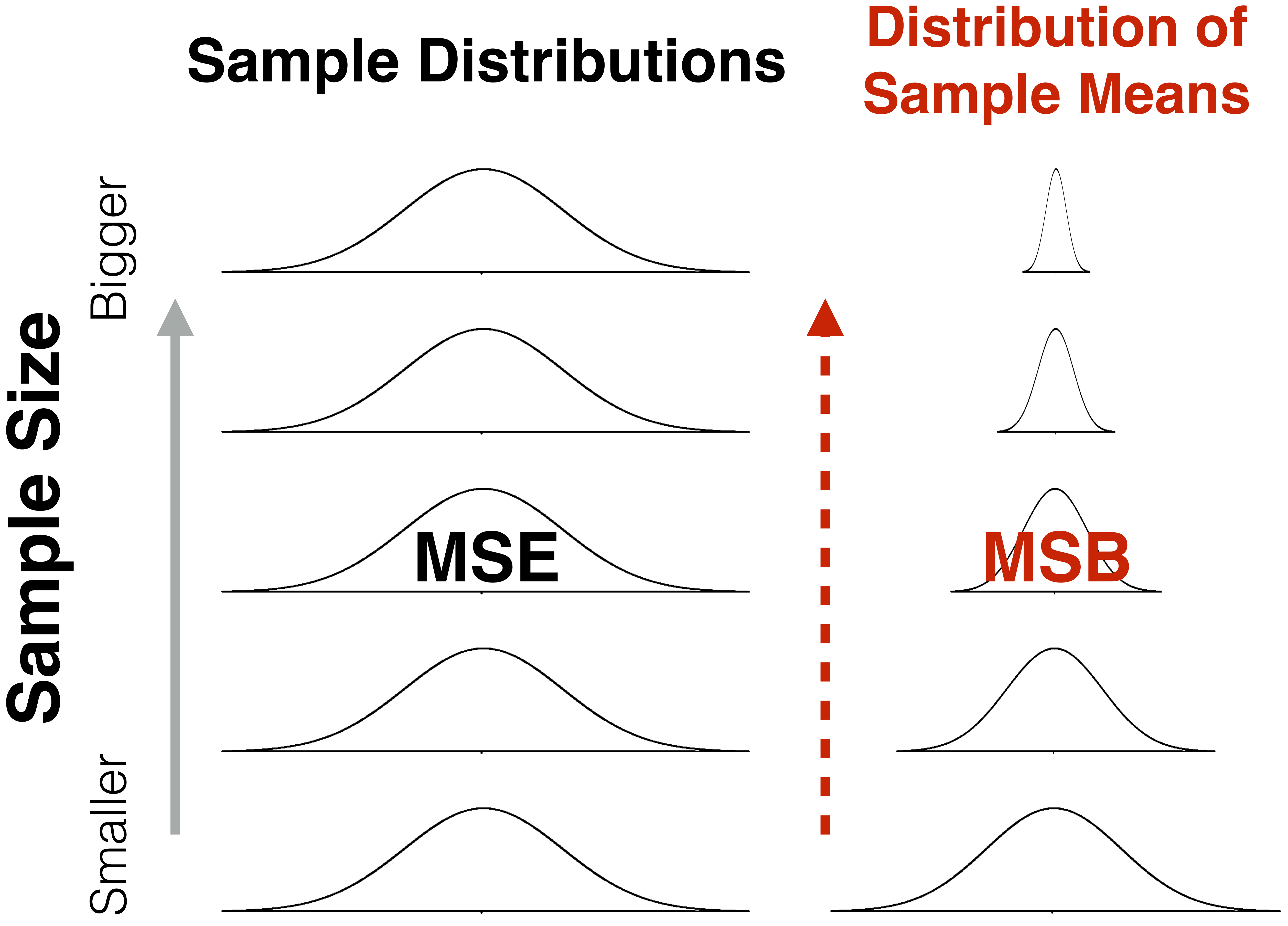
ANOVA asks:
Is Any One (or More) Different?

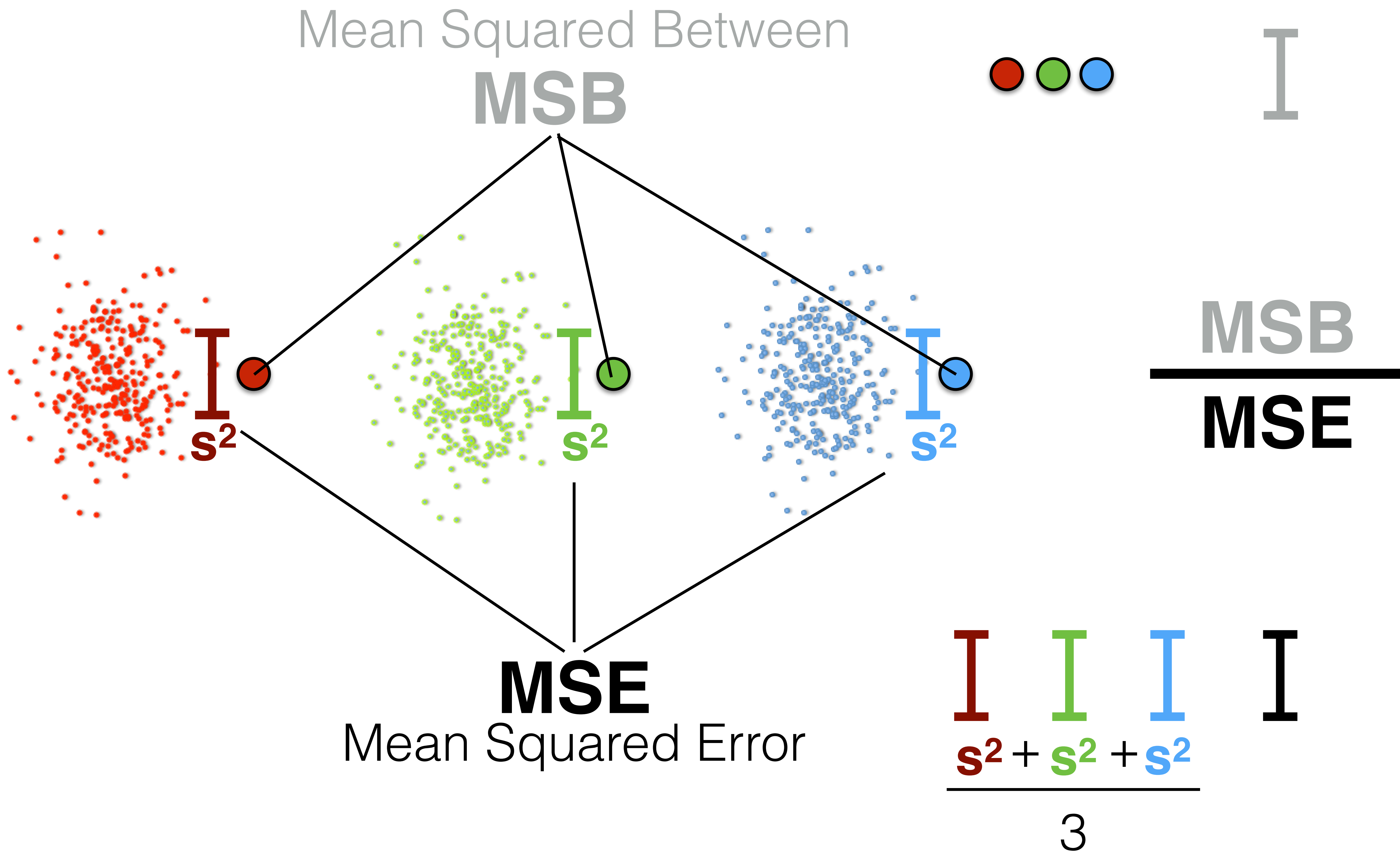




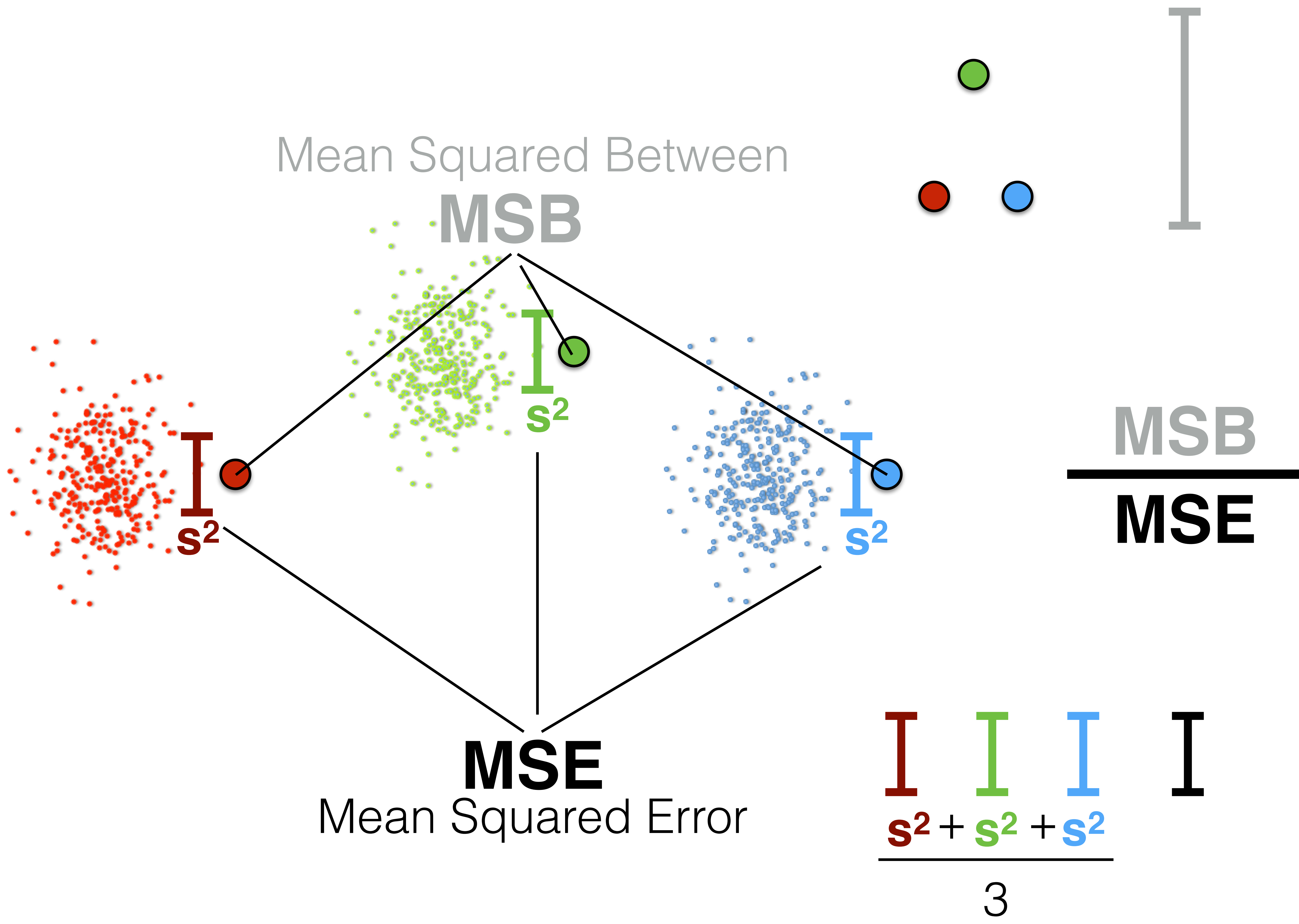
Is Any One (or More) Different?



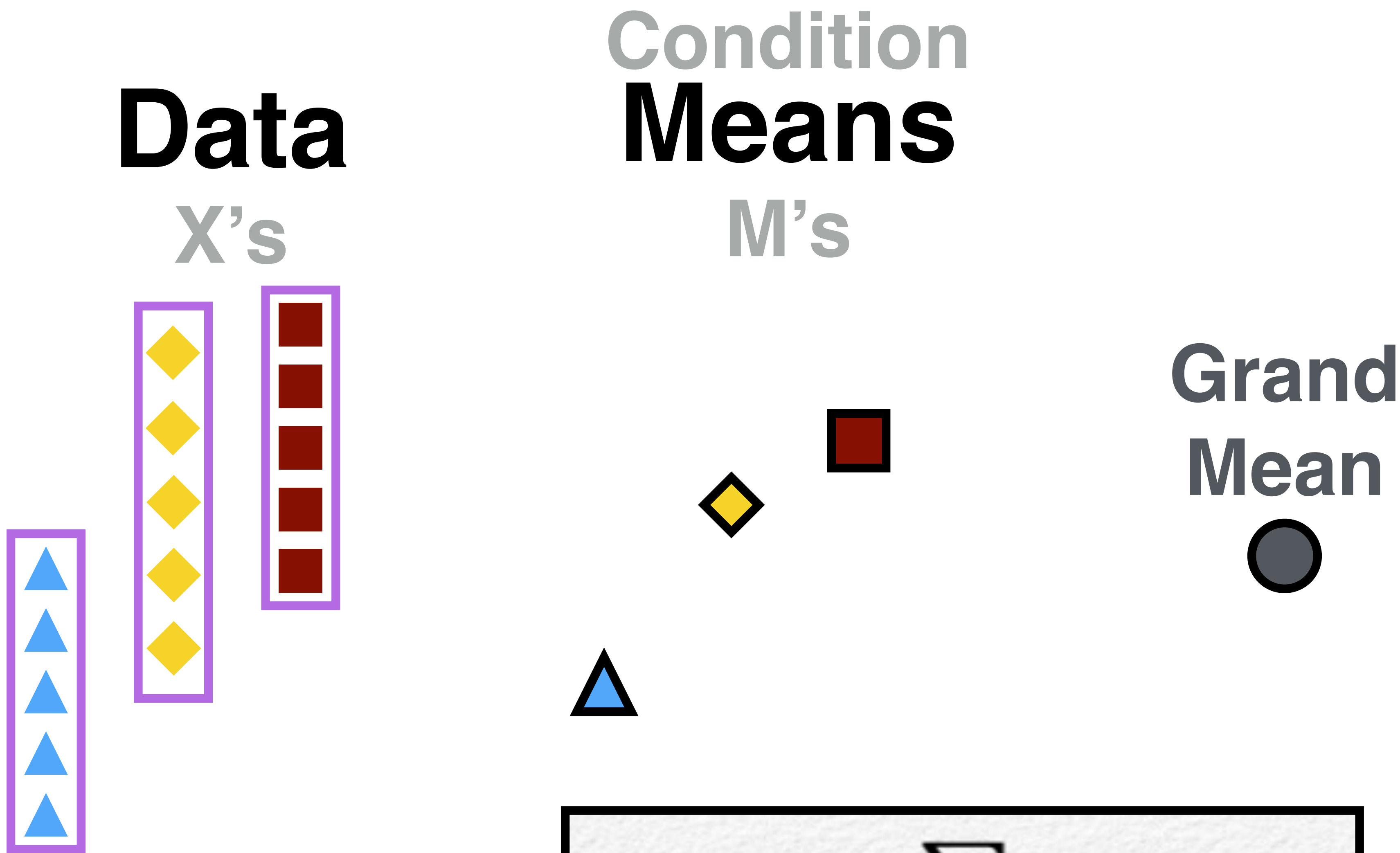

Sample Distributions
Distribution of Sample Means
Sample Size
Bigger
Smaller
MSE
MSB



Mean Squared Between
MSB
s2
s2
s2
MSB
MSE
MSE
Mean Squared Error
s2 + s2 + s2
3



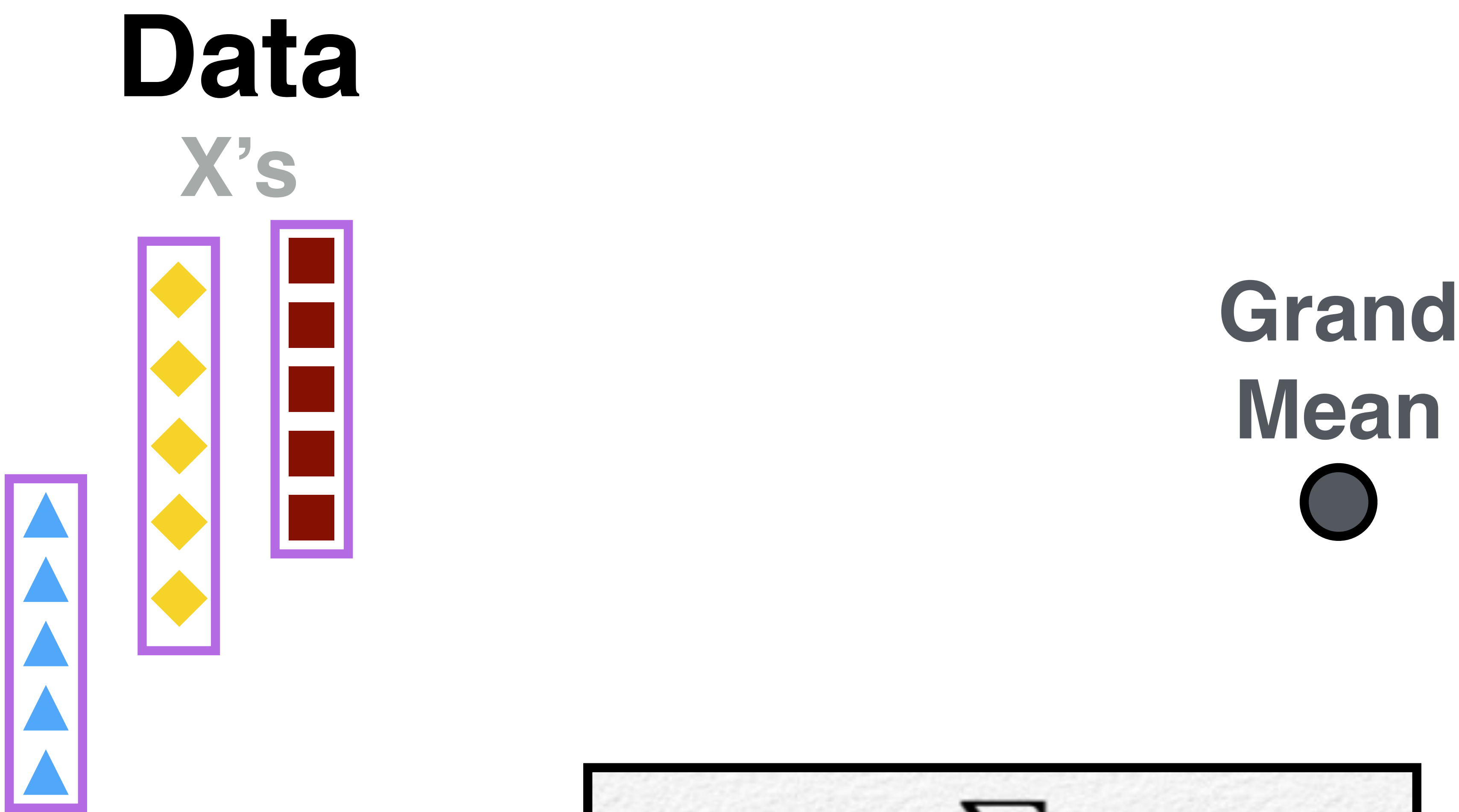
Mean Squared Between
MSB
s²
s²
s²
MSB
MSE
MSE
Mean Squared Error
s² + s² + s²
3



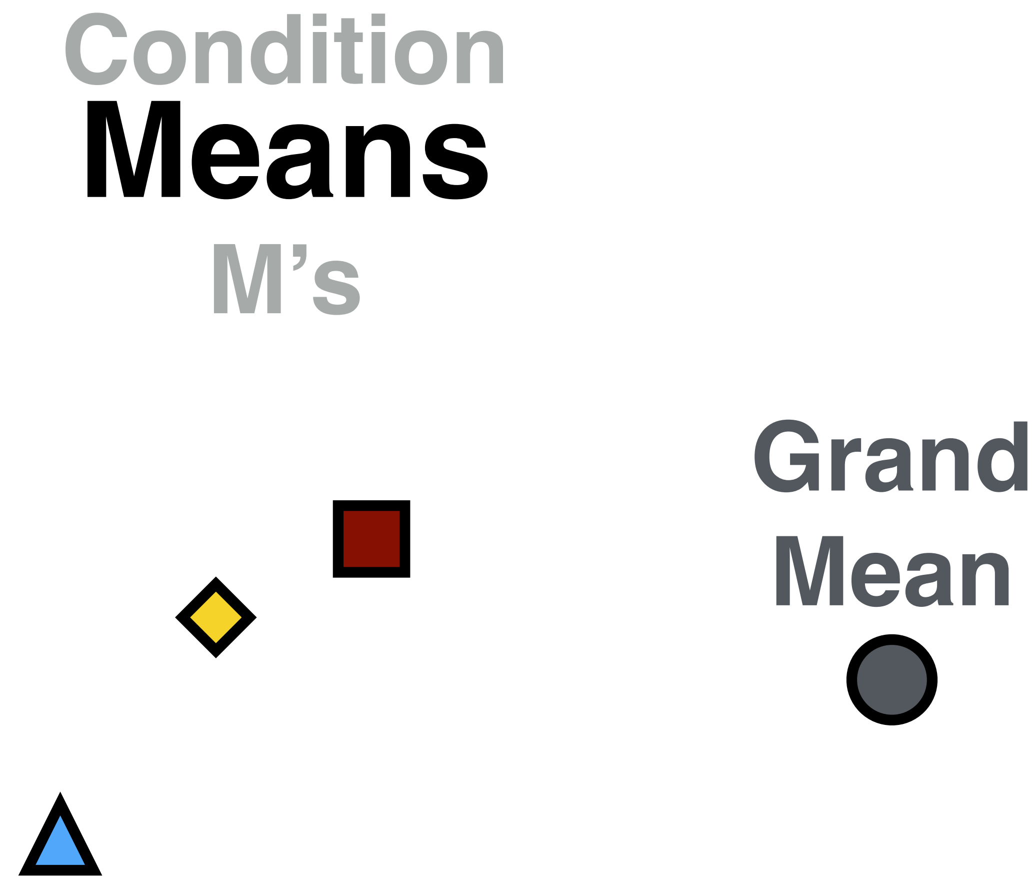
$$SSQ_{total} = \sum (X - GM)^2$$

$$SSQ_{condition} = n \sum (M_1 - GM)^2 + (M_2 - GM)^2 + \cdots + (M_k - GM)^2$$

$$SSQ_{error} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \cdots + \sum (X_{ik} - M_k)^2$$



$$SSQ_{total} = \sum (X - GM)^2$$



MSB

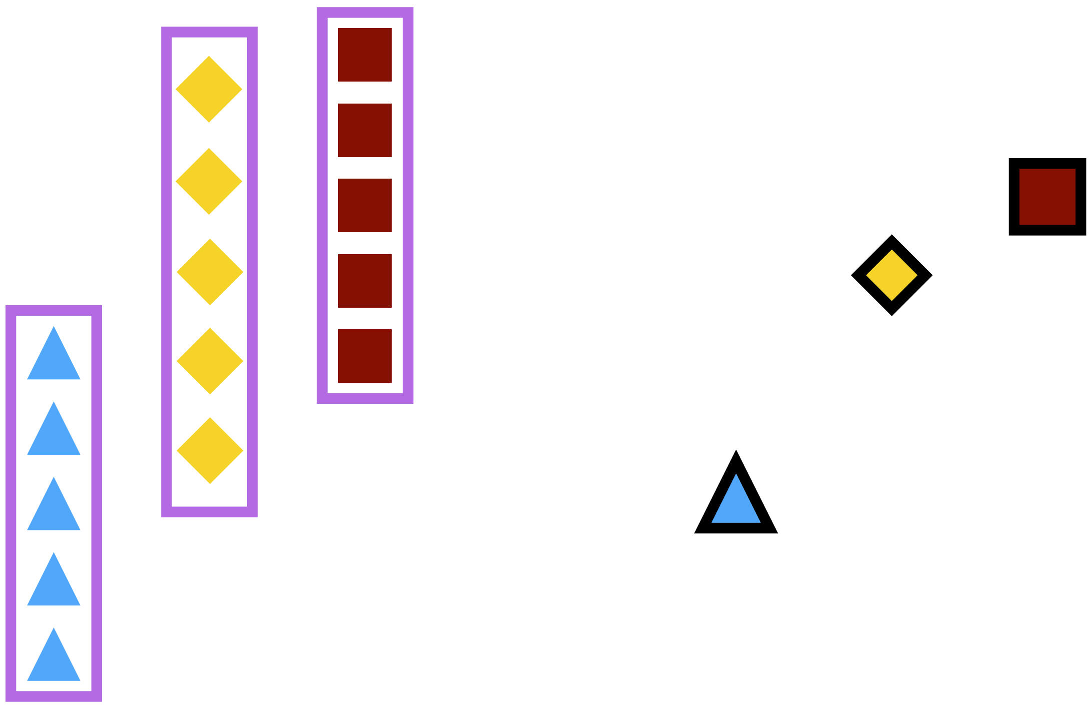
$$SSQ_{condition} = n \sum (M_1 - GM)^2 + (M_2 - GM)^2 + \cdots + (M_k - GM)^2$$

Data

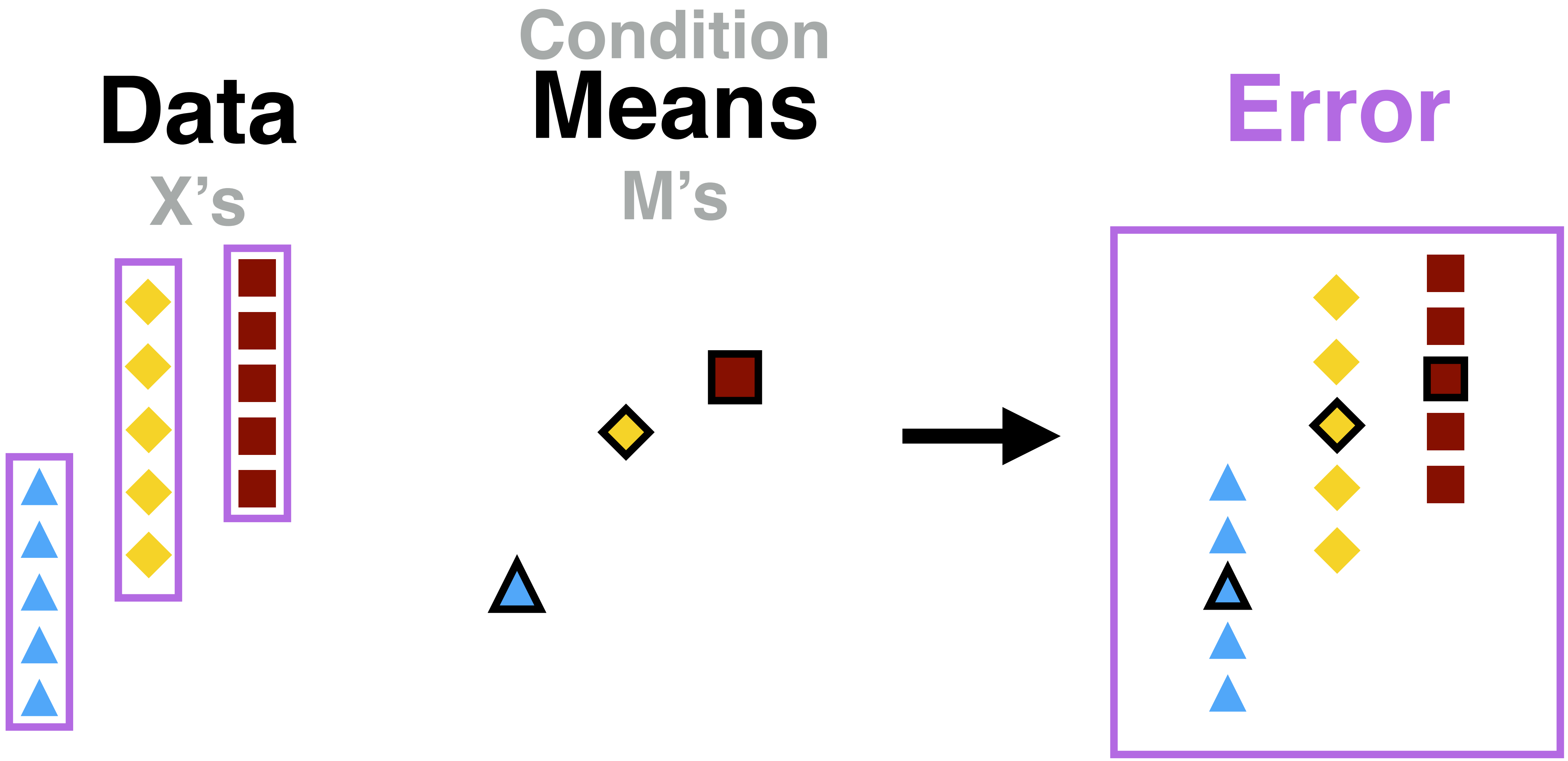
X's

Condition **Means**

M's



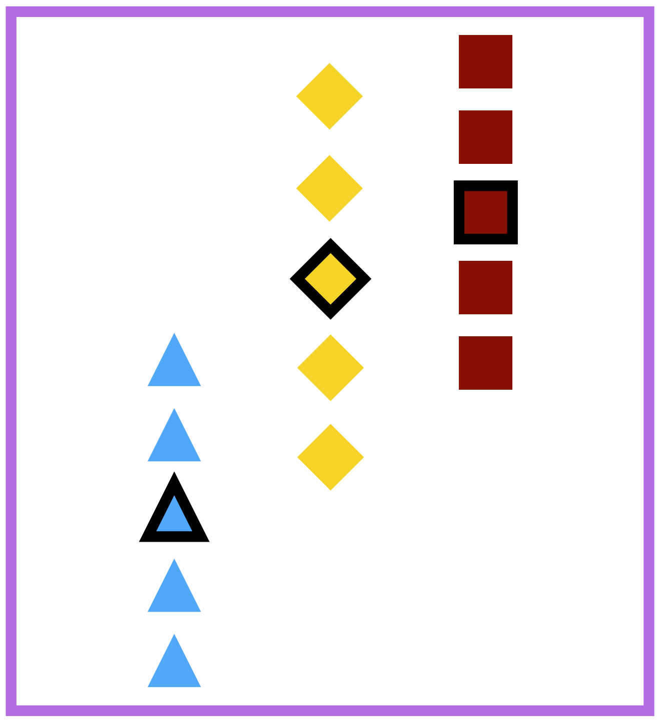
$$SSQ_{error} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \cdots + \sum (X_{ik} - M_k)^2$$



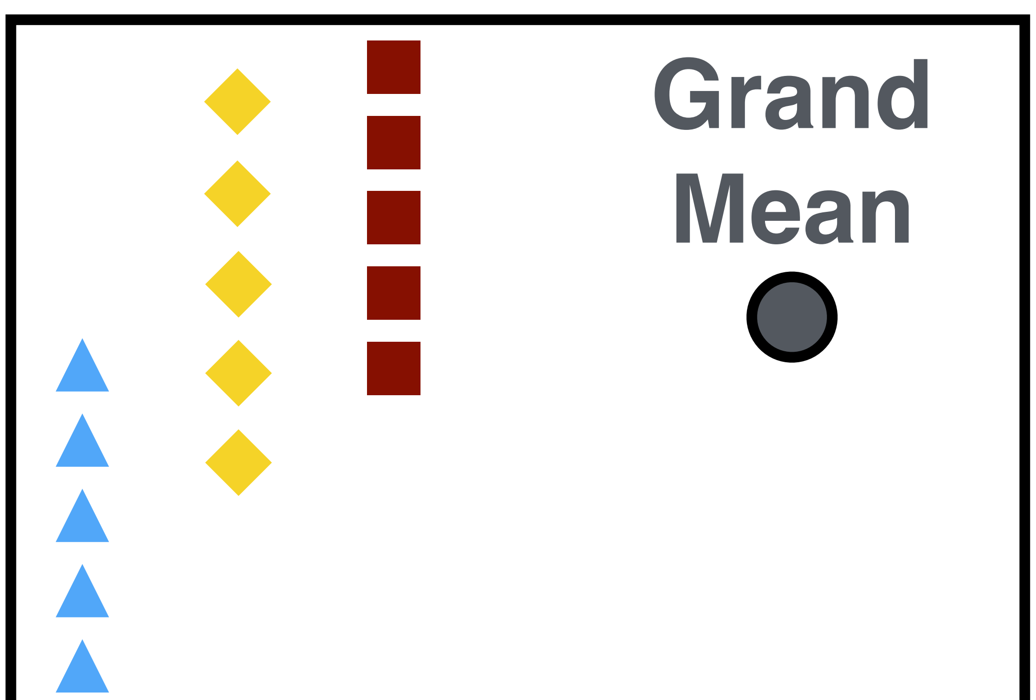
MSE

$$SSQ_{error} = \sum (X_{i1} - M_1)^2 + \sum (X_{i2} - M_2)^2 + \cdots + \sum (X_{ik} - M_k)^2$$

$$SSQ_{condition} = n\sum(M_1 - GM)^2 +$$

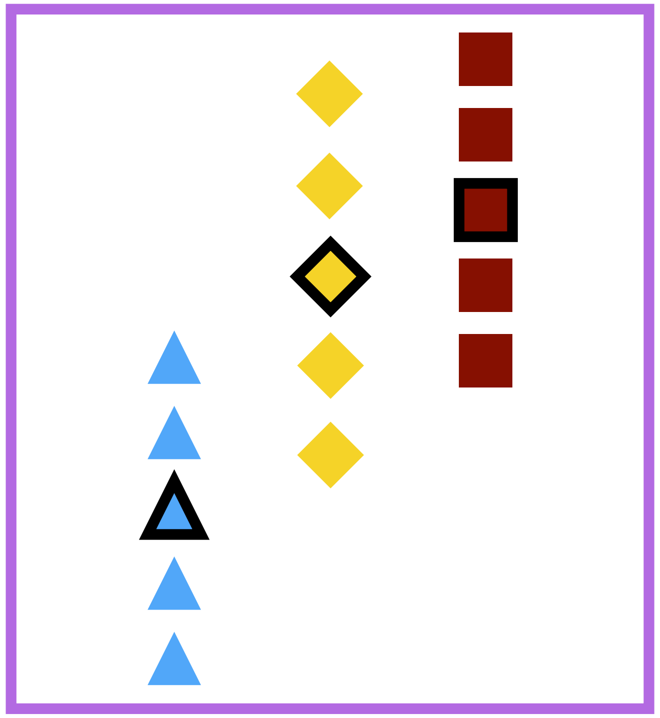


$$SSQ_{error} = \sum(X_{i1} - M_1)^2 +$$



$$SSQ_{total} = \sum(X - GM)^2$$

$$SSQ_{condition} = n\sum(M_1 - GM)^2 +$$



$$SSQ_{error} = \sum(X_{i1} - M_1)^2 +$$

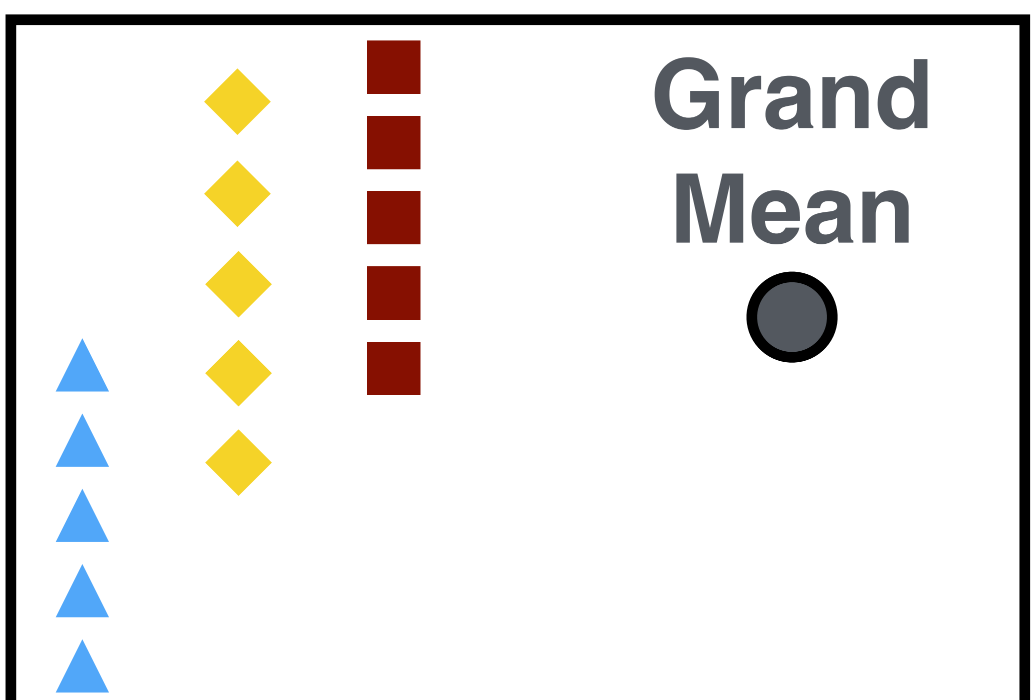
$$\text{MSB} = \text{SSQ}_{\text{condition}}/\text{dfn}$$

$$F = \frac{\text{MSB}}{\text{MSE}}$$

$$\text{MSE} = \text{SSQ}_{\text{error}}/\text{dfn}$$

$$SSQ_{condition} = n\sum(M_1 - GM)^2 +$$

$$\text{Effect Size} = \frac{SSQ_{condition}}{SSQ_{total}}$$



$$SSQ_{total} = \sum(X - GM)^2$$

P(A|B)

is the **likelihood** that **A is true**, **given B**.

New **evidence** can **update** our **beliefs**.

Making us more certain, or less certain, that a hypothesis or theory is true.

P(B|A)

is the **probability** of **observing B,**
if A is true

True theories are consistent with all observations.

Prior → Posterior → New Prior'

$$P(A) \rightarrow P(A|B) \rightarrow P(A')$$

A is a **hypothesis**. **B** is **evidence**.

P(A) is the **likelihood** that **A is true**.

P(B) is the **probability** of **observing B.**

P(A|B) is the **likelihood** that **A is true**, **given** **B**.

P(B|A) is the **probability** of **observing B**, **if** **A is true**

the **probability** of **observing B, if A is true**

the **likelihood** that **A is true**.

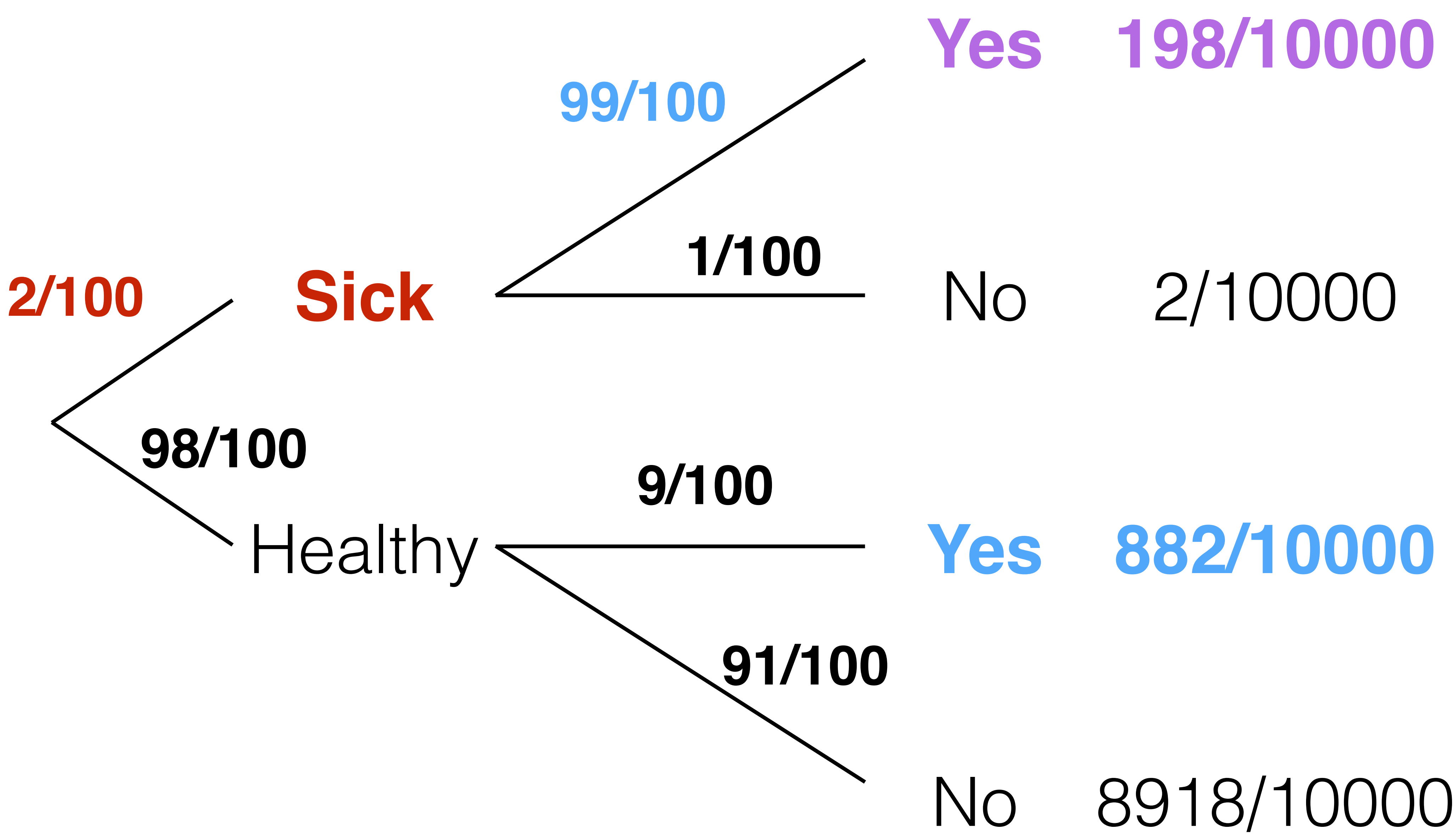
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

the **likelihood** that **A is true, given B**.

the **probability** of **observing B.**

2/100 99/100

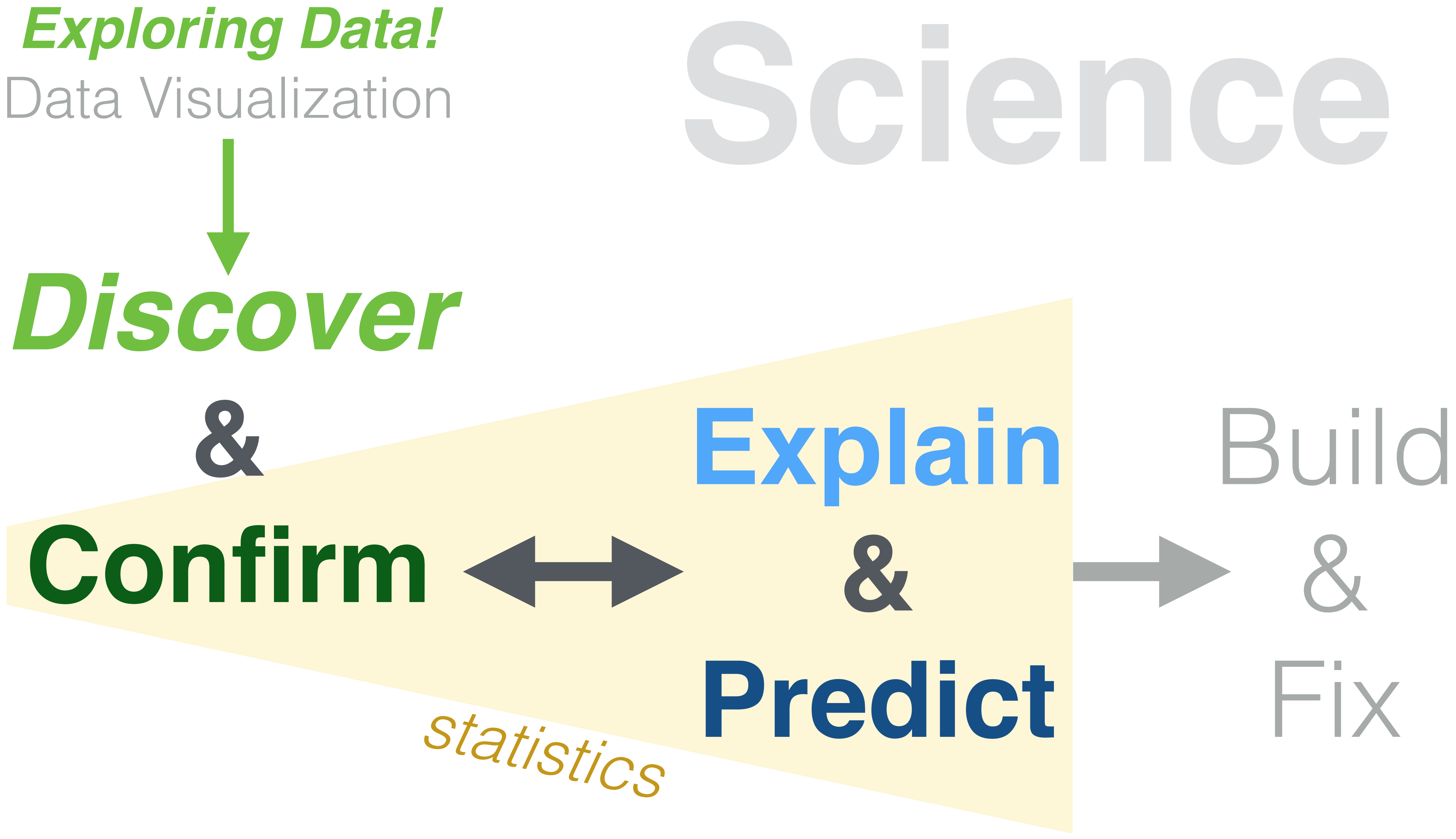
$$P(A)\ P(B|A)$$



$$P(A|B) = \frac{198}{198+882}$$

$$P(B)$$

198/10000 + 882/10000


Exploring Data!
Data Visualization
Science
Discover
&
Confirm
Explain
&
Predict
Build
&
Fix
statistics