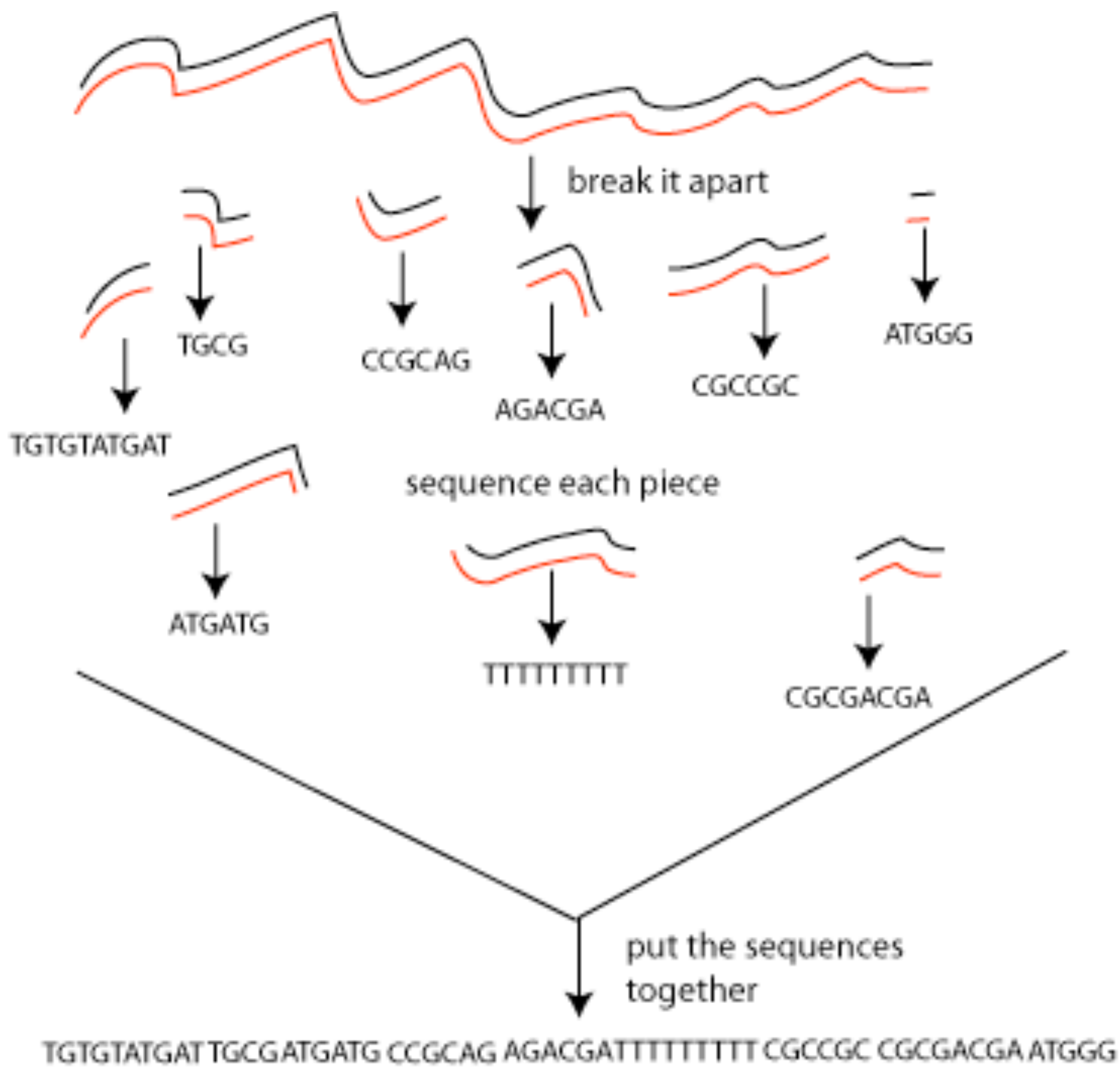
Shotgun genome sequencing

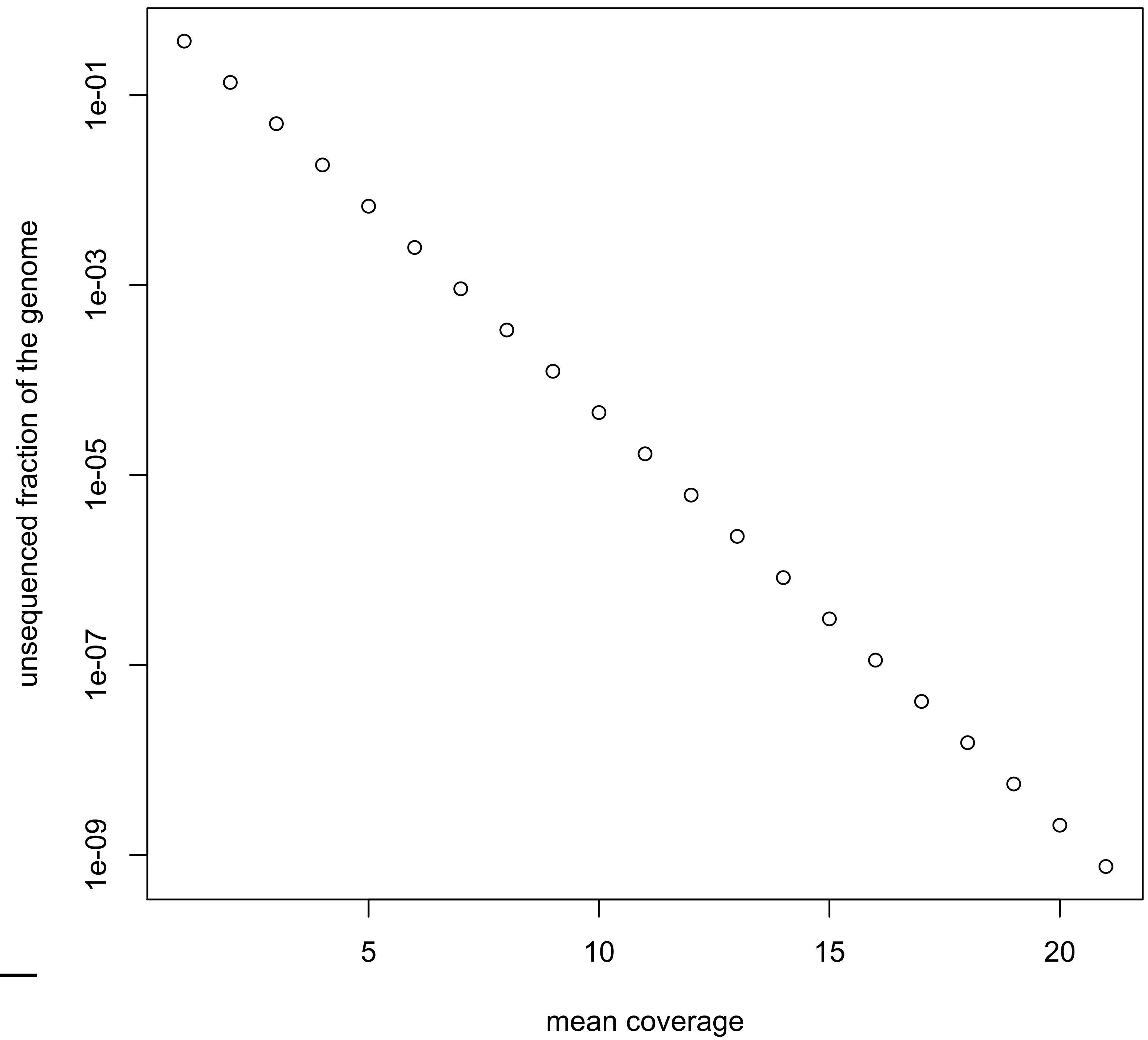


Sequencing coverage

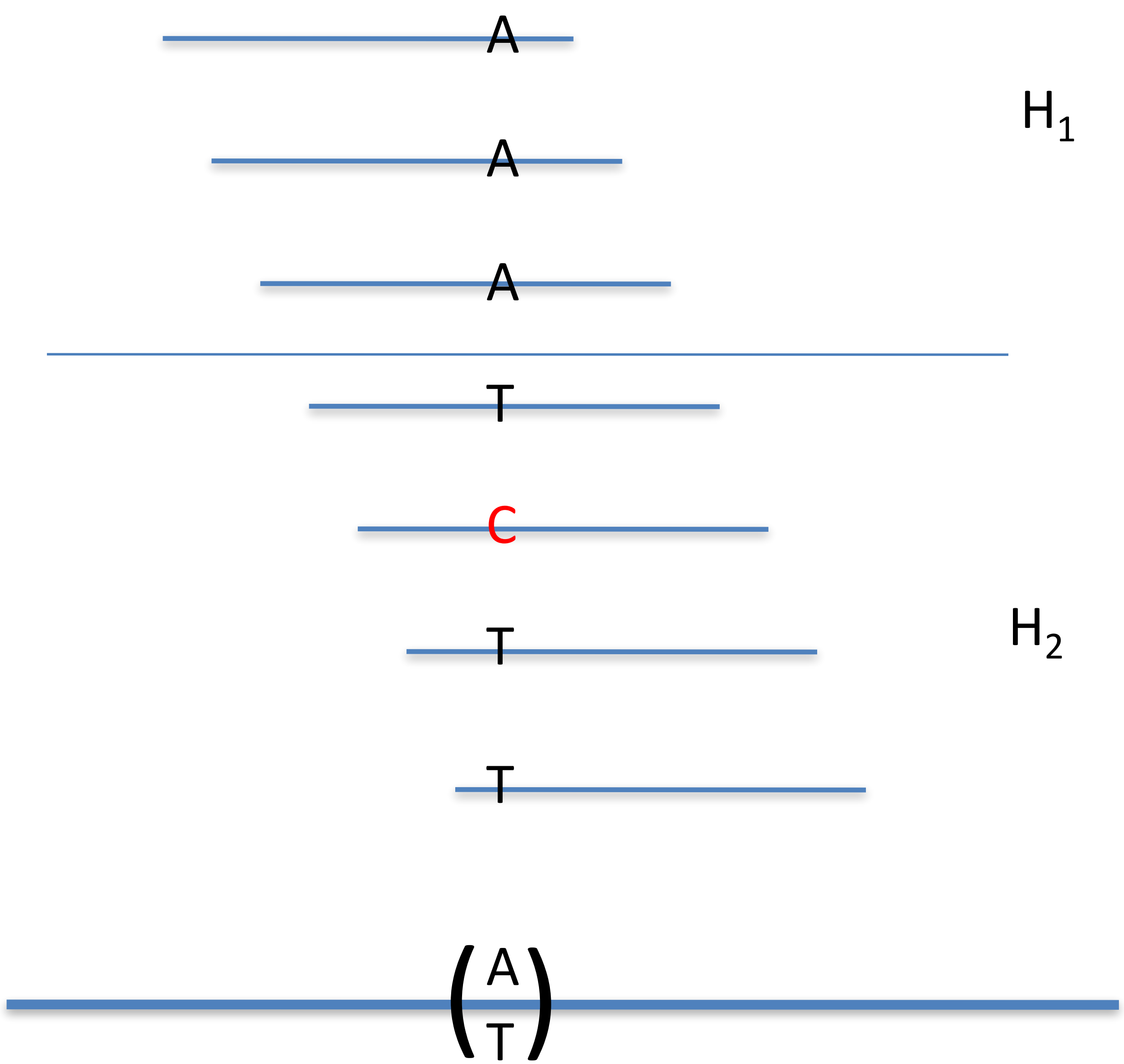
coverage	$P_{poisson}(X=0, coverage)$
1	$3.67\ 10^{-1}$
2	$1.35\ 10^{-1}$
3	$4.97\ 10^{-2}$
4	$1.83\ 10^{-2}$
5	$6.73\ 10^{-3}$
6	$2.47\ 10^{-3}$
7	$9.11\ 10^{-4}$
8	$3.35\ 10^{-4}$
9	$1.23\ 10^{-4}$
10	$4.53\ 10^{-5}$
11	$1.67\ 10^{-5}$
12	$6.14\ 10^{-6}$
13	$2.26\ 10^{-6}$
14	$8.31\ 10^{-7}$
15	$3.05\ 10^{-7}$
16	$1.12\ 10^{-7}$
17	$4.13\ 10^{-8}$
18	$1.52\ 10^{-8}$
19	$5.60\ 10^{-9}$
20	$2.06\ 10^{-9}$
21	$7.58\ 10^{-10}$

$$P_{poisson}(X = x \mid \lambda) = \lambda^x \, \frac{e^{-\lambda}}{x!}$$

$$P_{poisson}(X = 0 \mid \lambda) = \lambda^0 \, \frac{e^{-\lambda}}{0!} = e^{-\lambda}$$



Anatomy of a sequenced loci



P_{hetero}

$$P(H_1) = P(H_2) = 0.5$$

$$P_{error}$$

Some probabilities

Well known binomial distribution

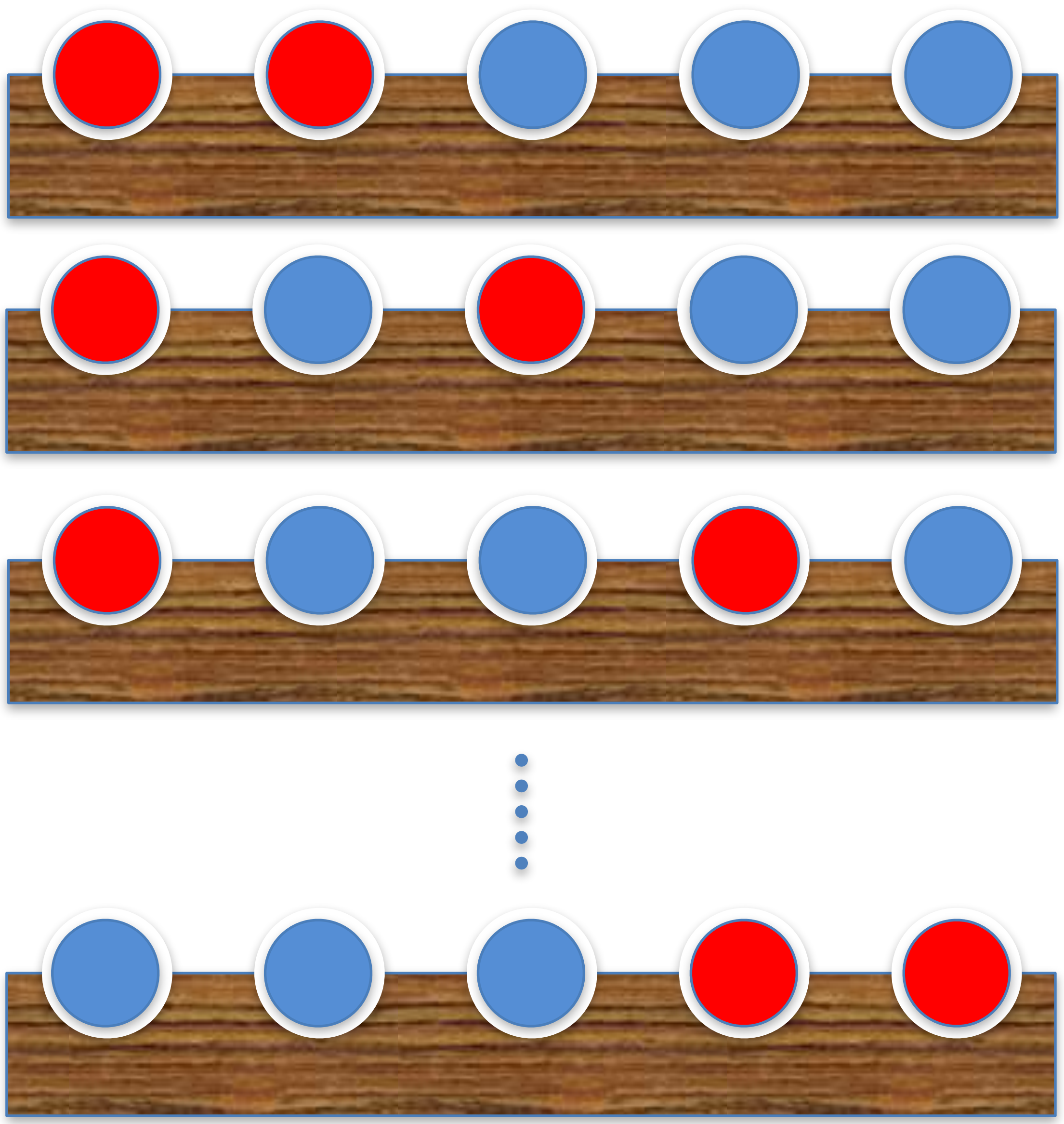
$$p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{p!(n-p)!}$$

Binomial coefficients

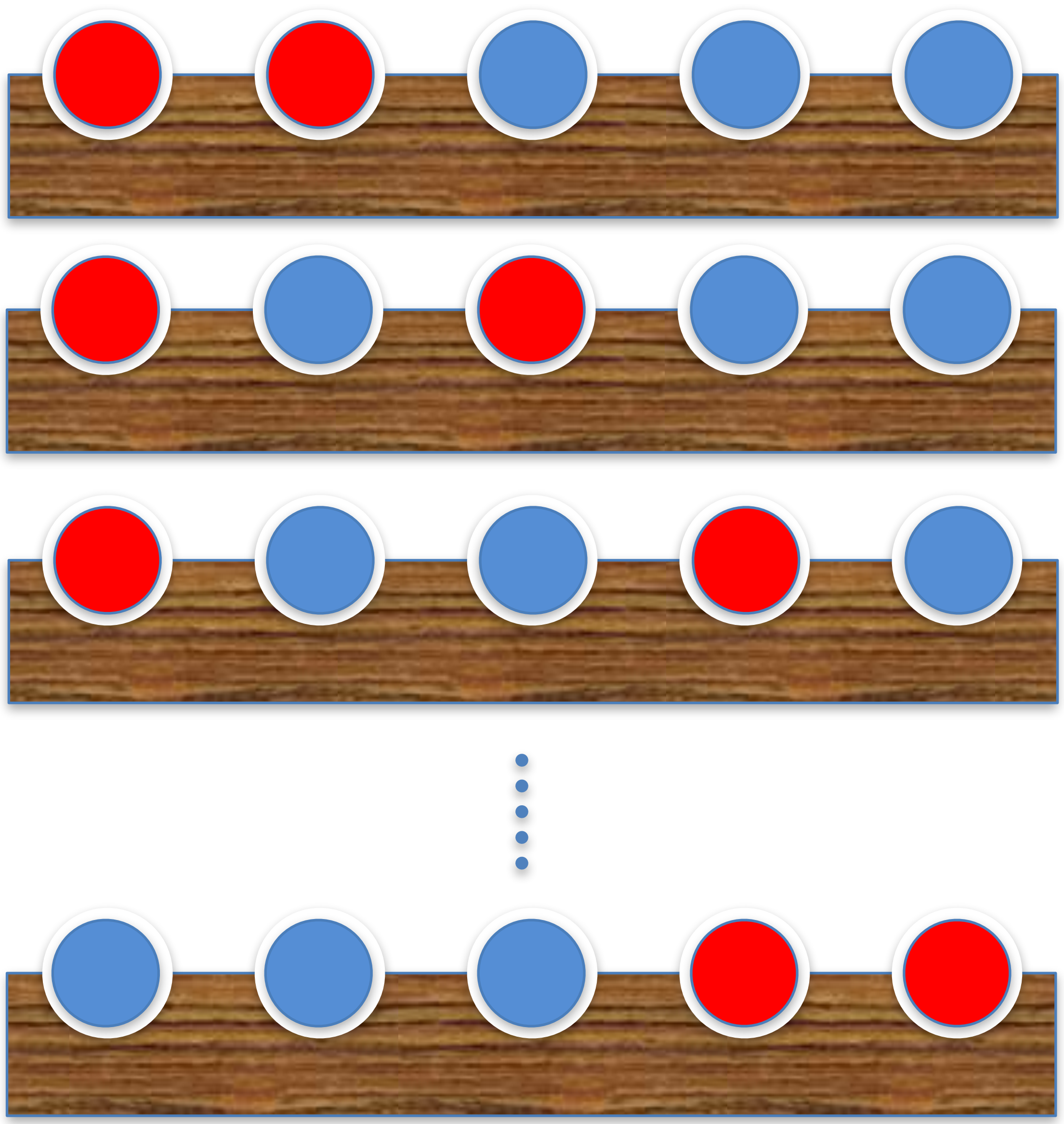
$$\binom{L_S}{l_S} = \frac{L_S!}{l_S!\,(L_S - l_S)!}$$

How many ways exist for ranging the beads ?

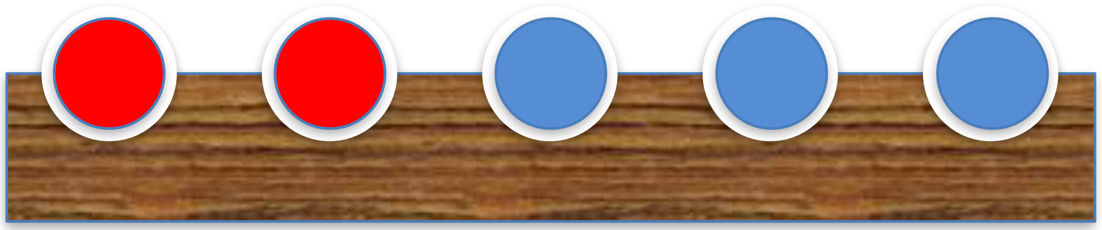


Complexe formula but simple explanation

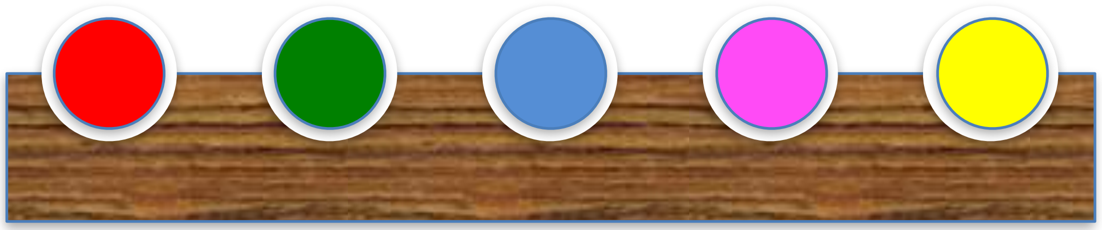
$$\binom{L_S}{l_S} = \frac{L_S!}{l_S!\,(L_S - l_S)!}$$



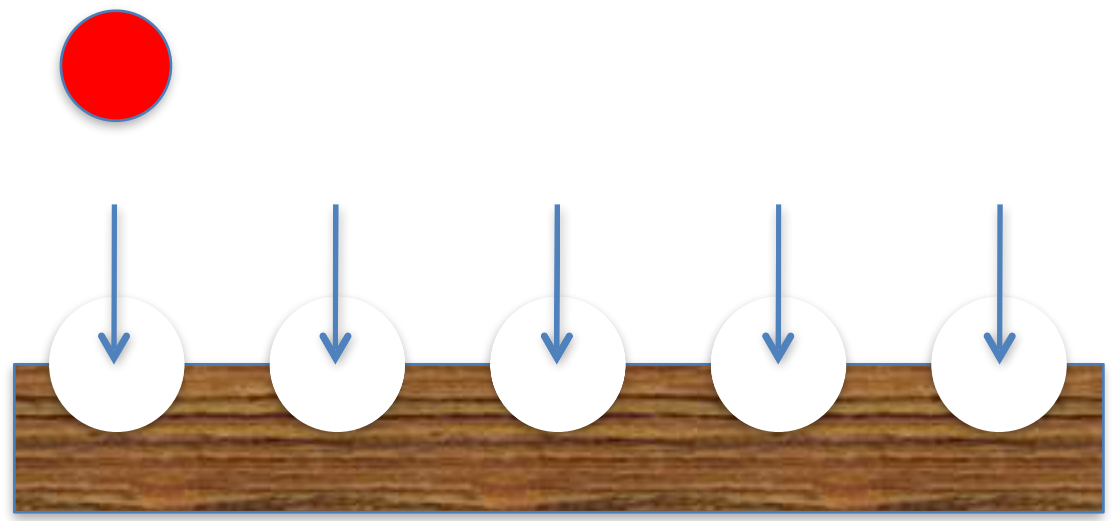
Going back to binomial coefficient



How many permutation of N beads ?

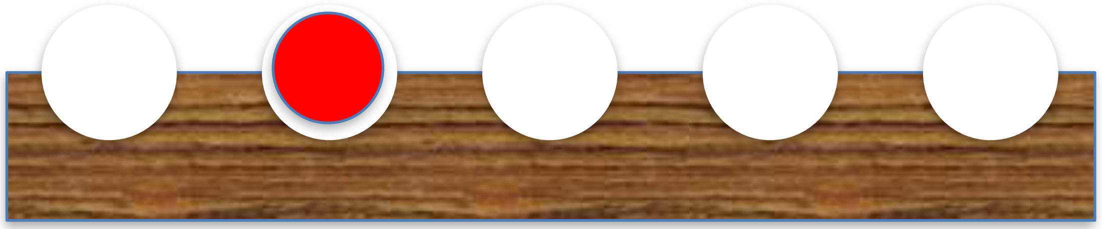


5 possibilities for the first bead

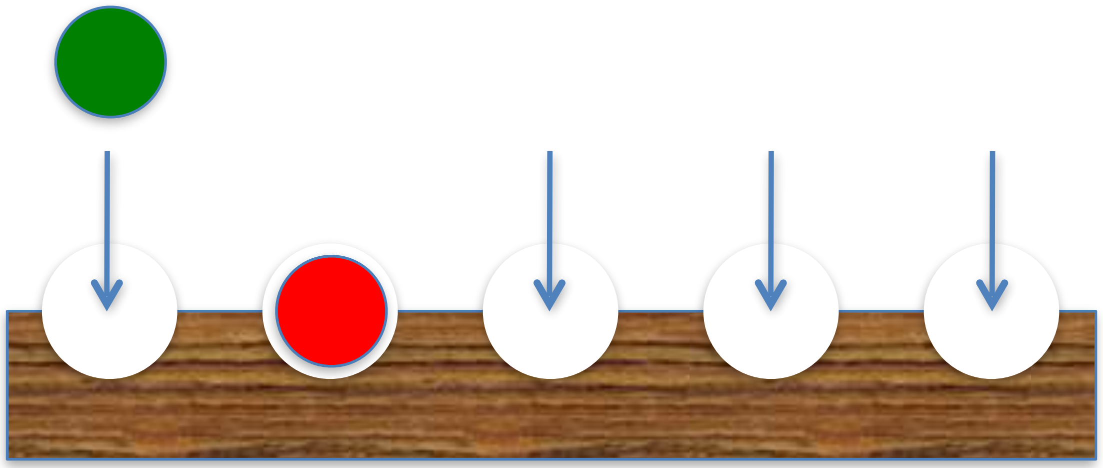


5

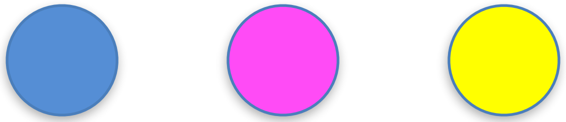
3 possibilities for the second bead



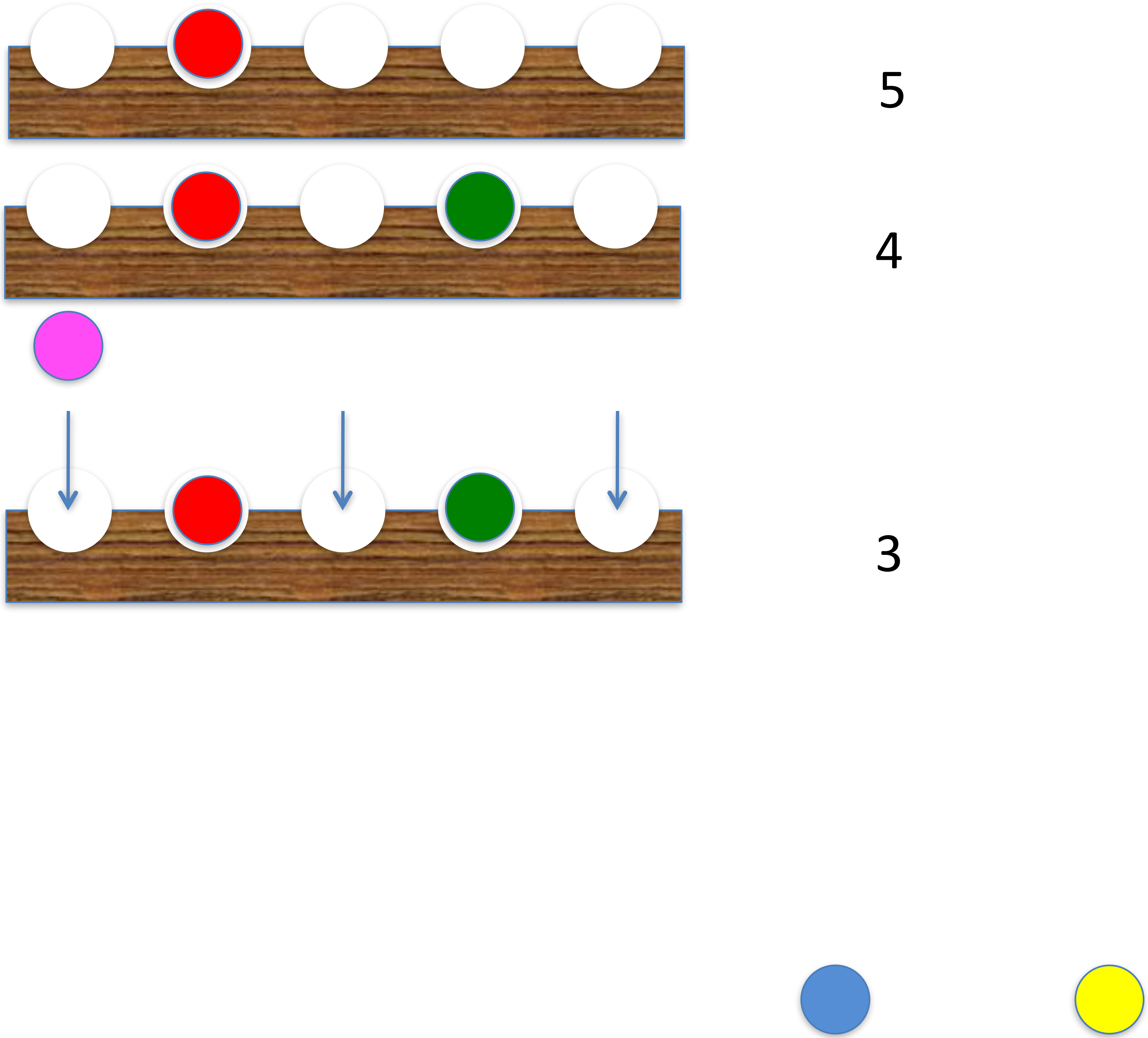
5



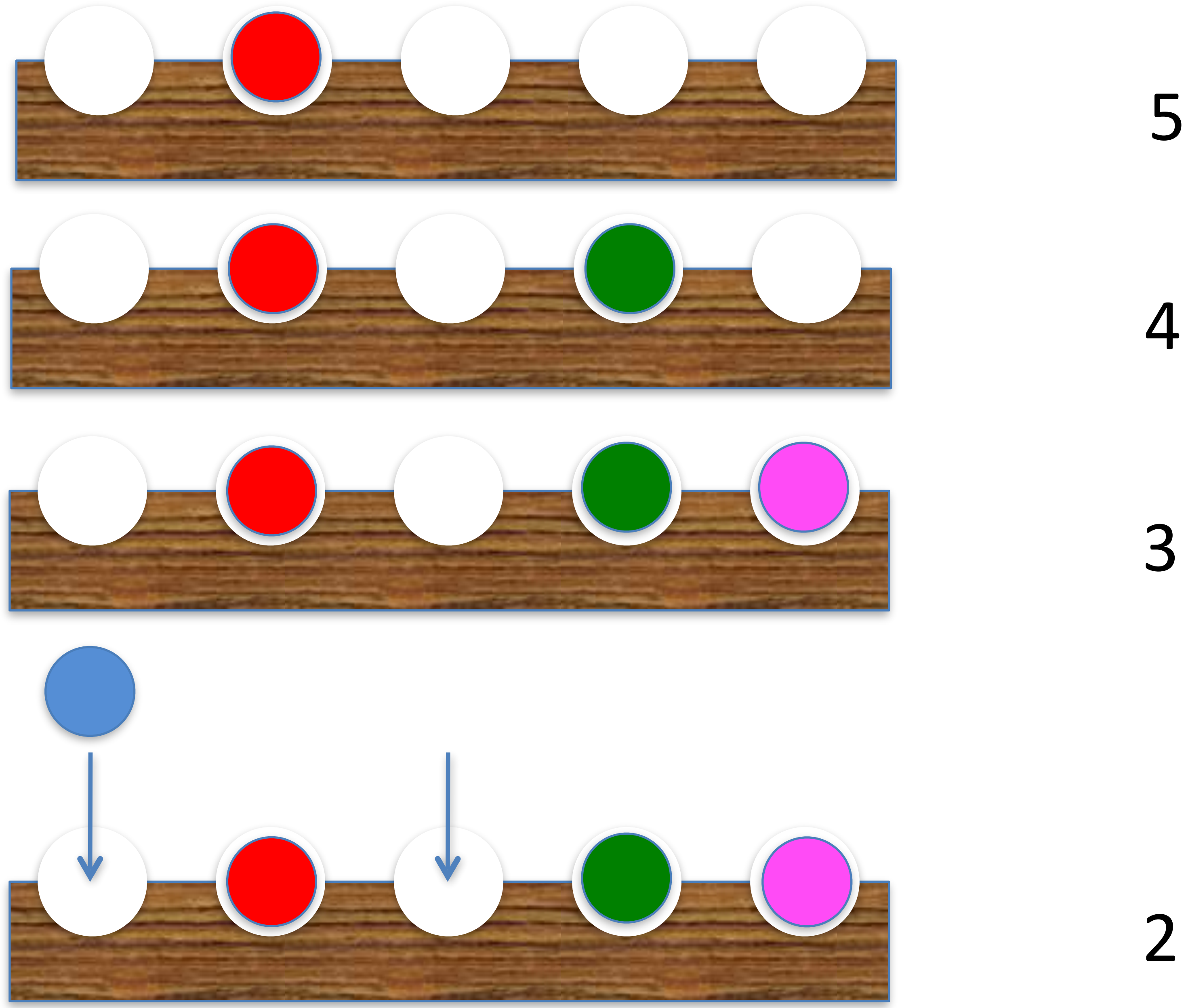
4



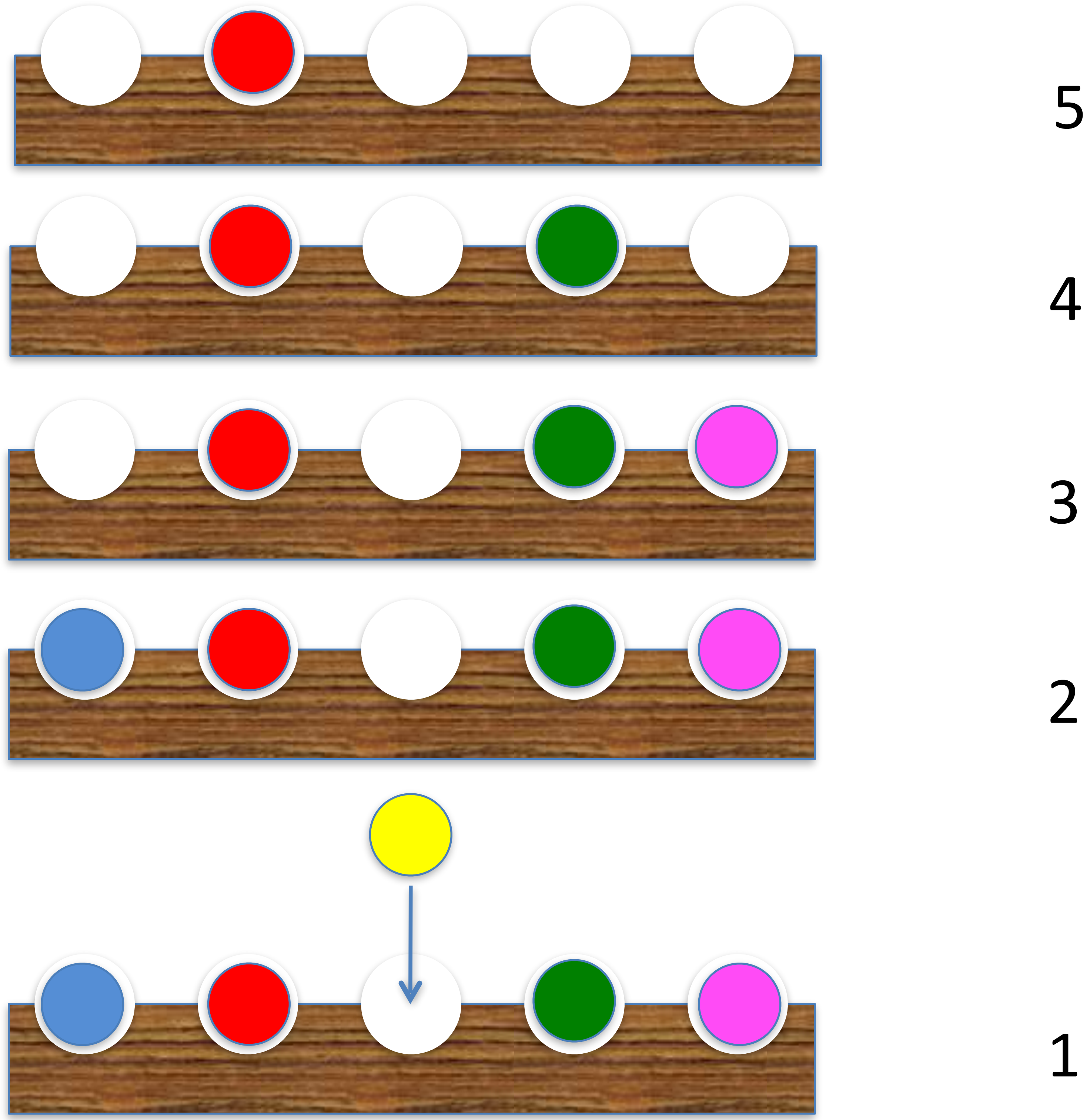
3 possibilities for the third bead



2 possibilities for the fourth bead



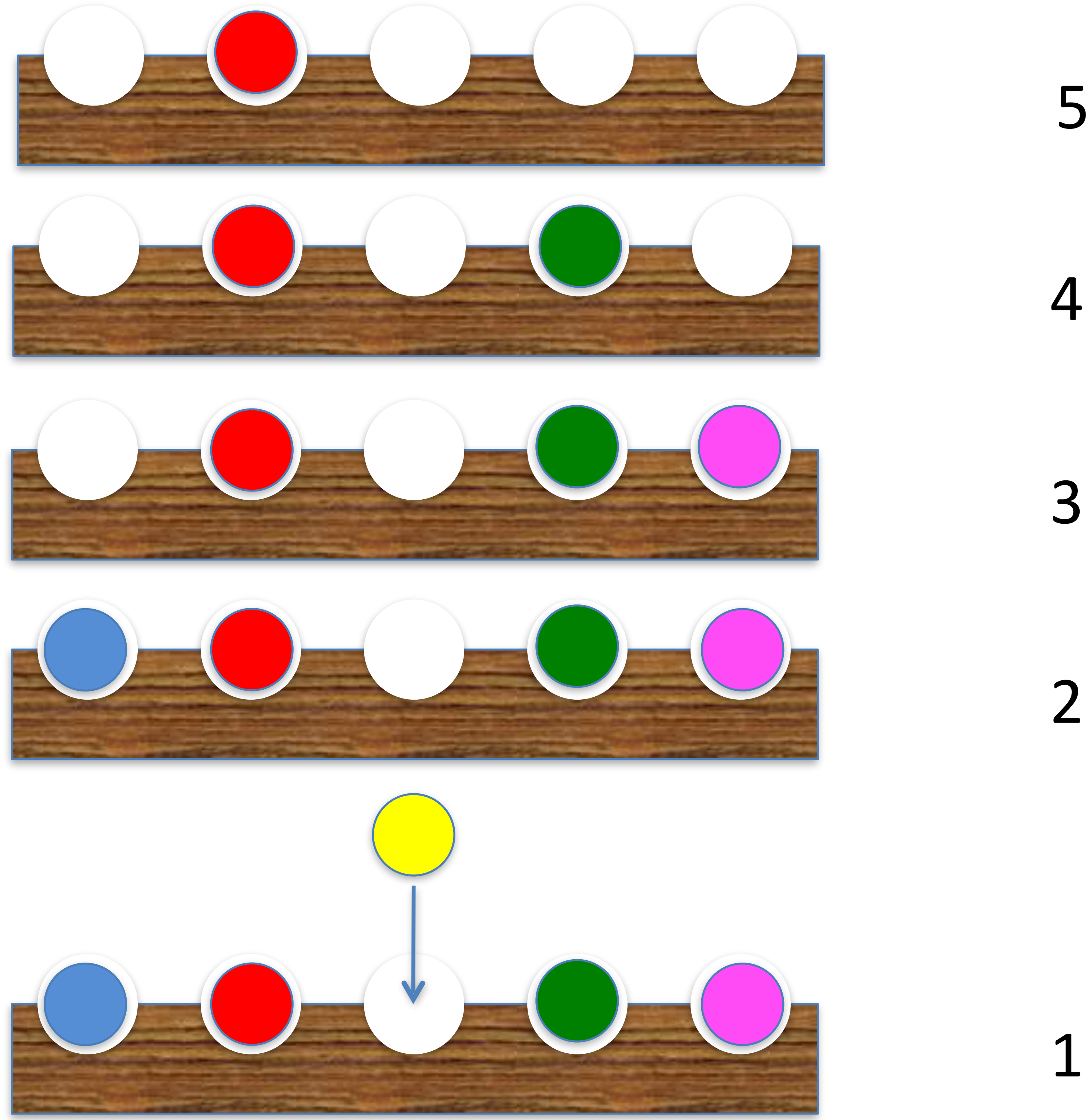
One for the last bead



$$5 \times 4 \times 3 \times 2 \times 1 = 5!$$

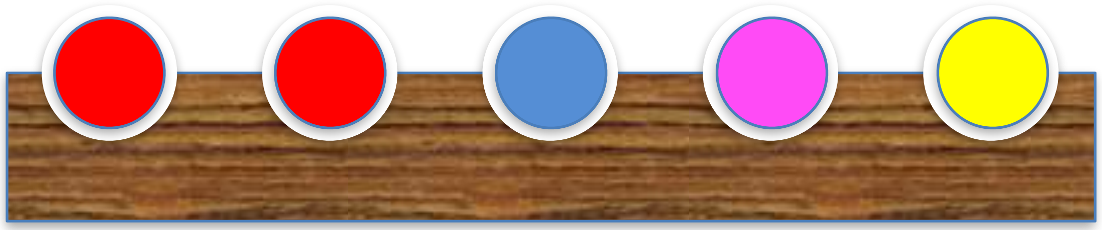
You understand now the first term of the formula

$$\binom{N}{a} = \frac{N!}{a!\,(N-a)!}$$



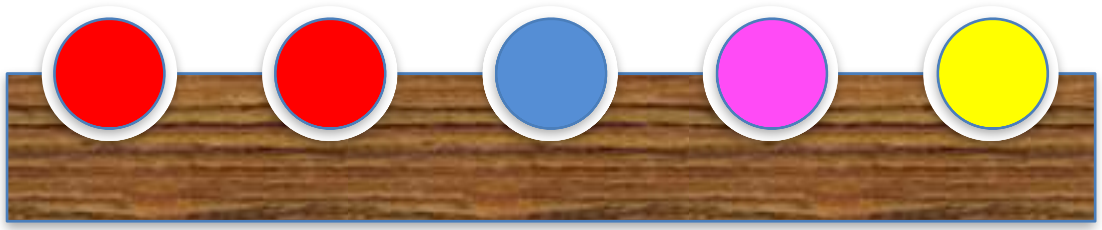
5 x 4 x 3 x 2 x 1 = 5!

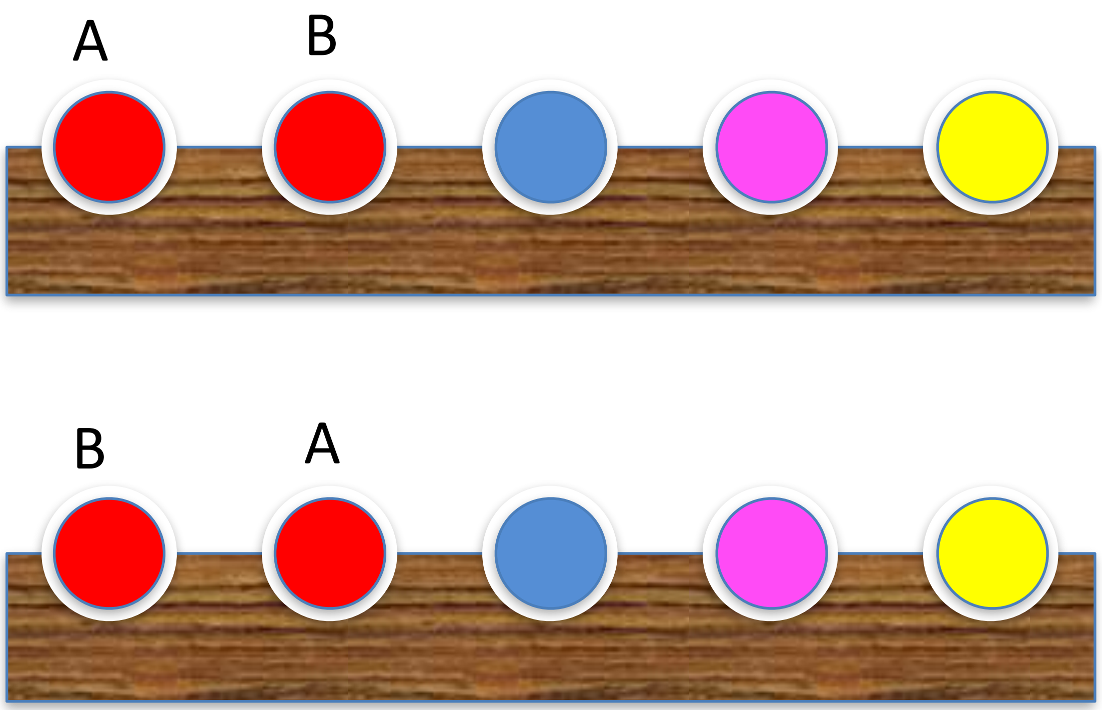
Two beads of the same color



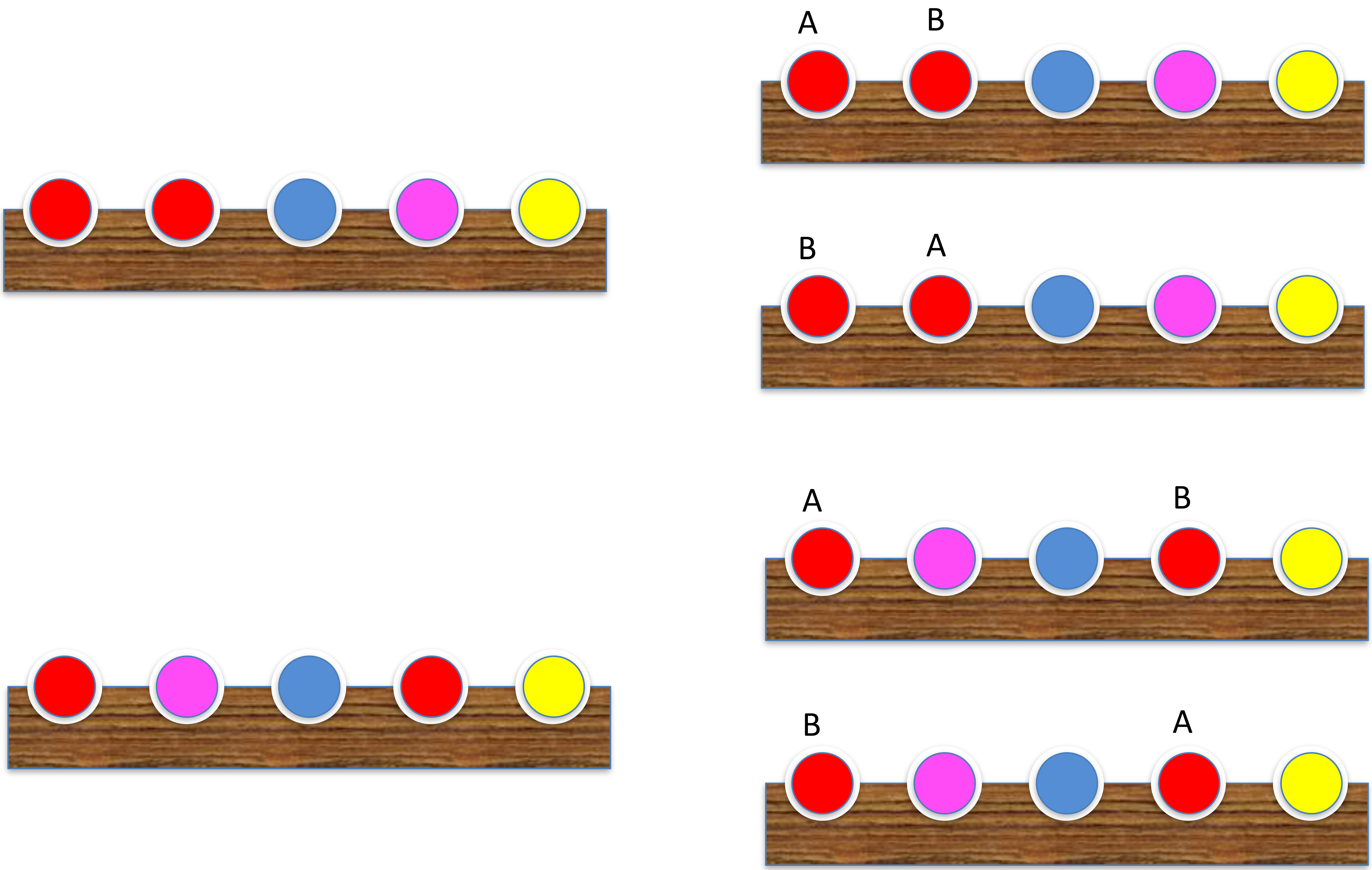
Two beads of the same color

Two hidden configurations

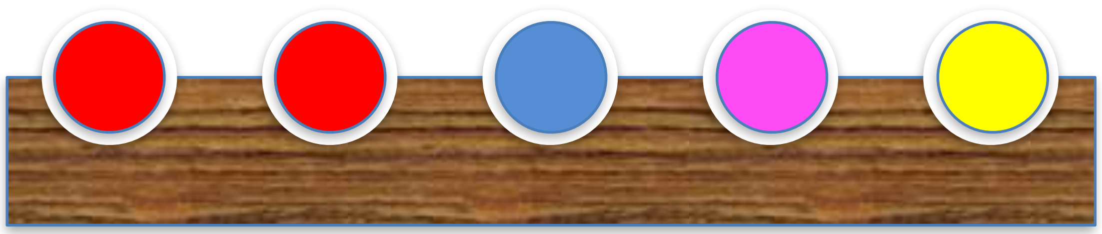




You must divide by the count of red bead permutation



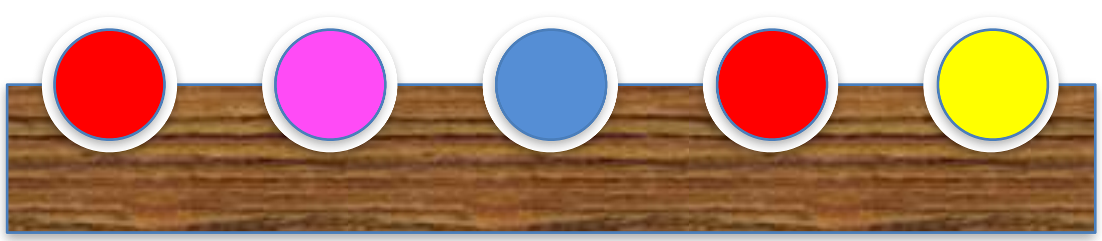
You must divide by the count of red bead permutations

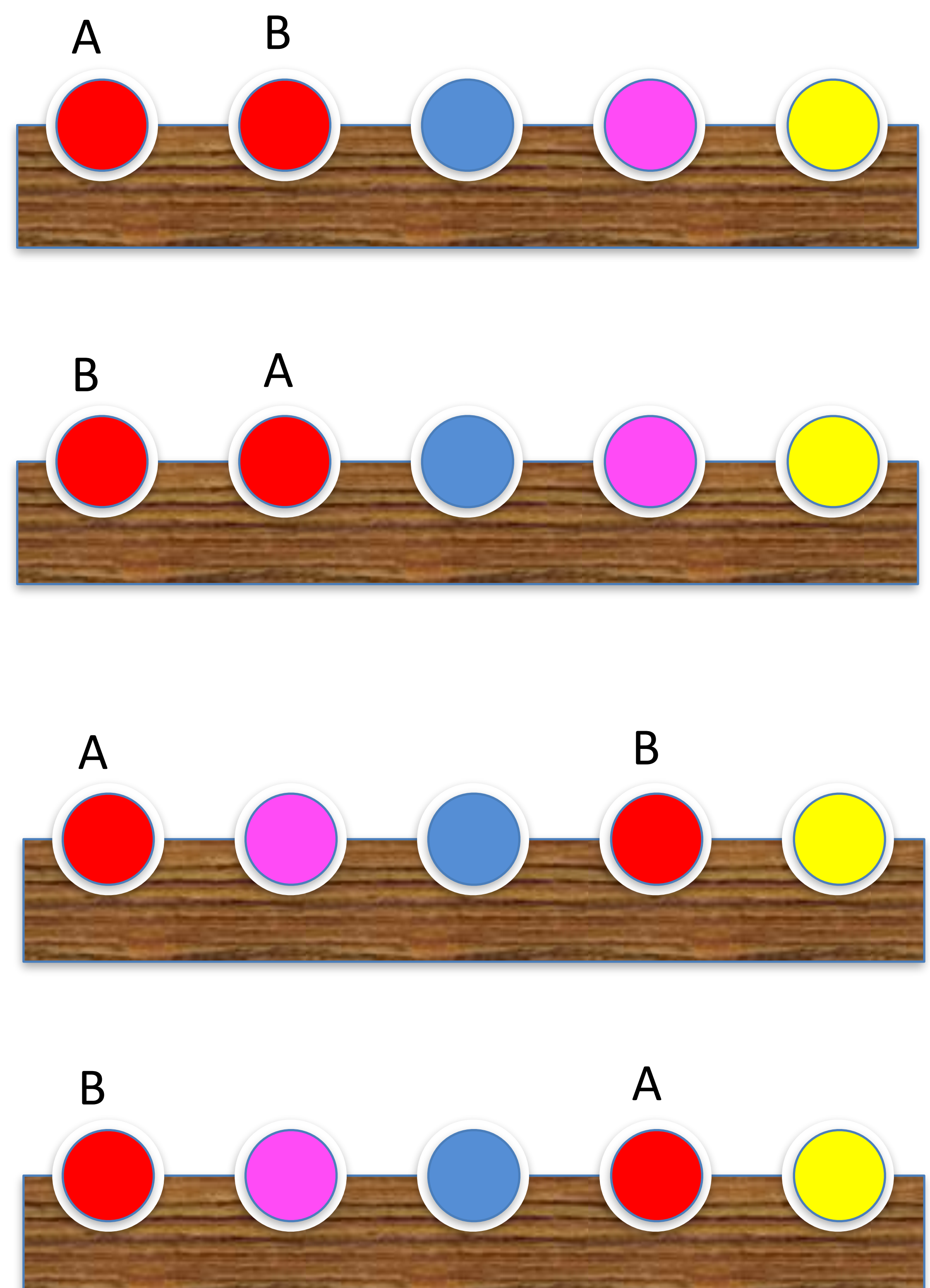


If you have *a red b beads you have*

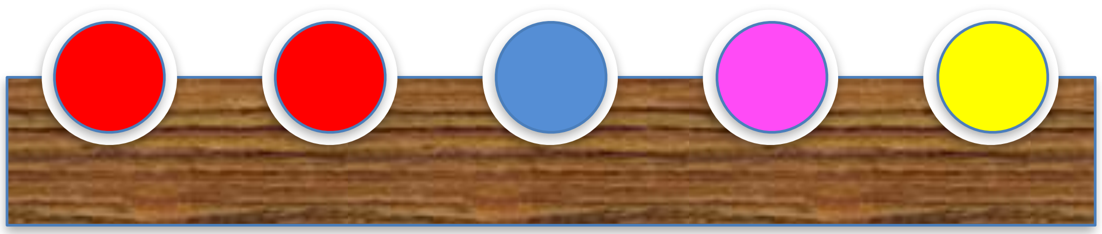
$$a!$$

Ways to range them

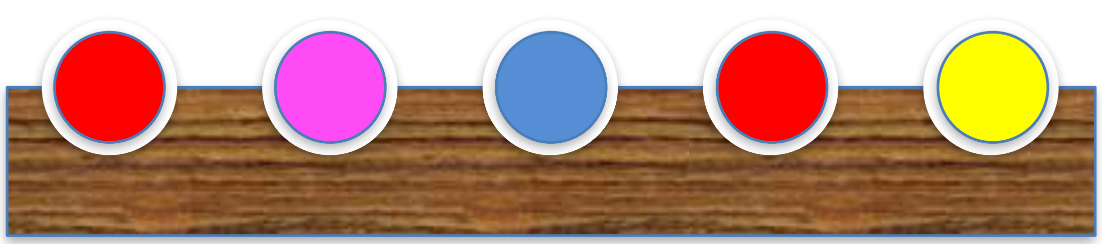




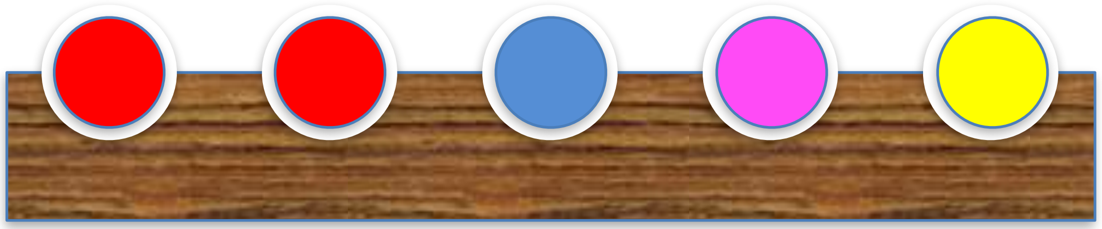
You understand now the second term

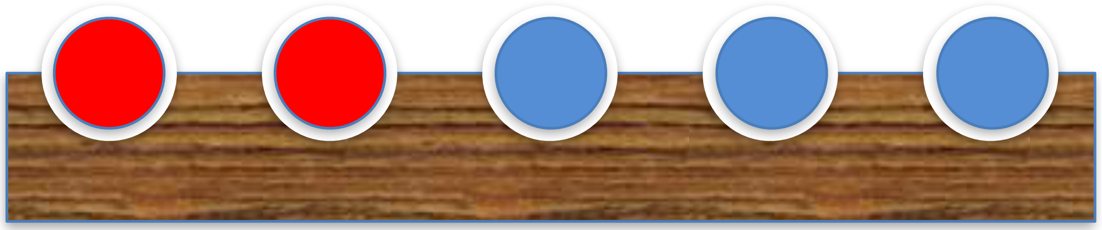


$$\binom{N}{a} = \frac{N!}{a!\ .\ (N-a)!}$$



If you paint not red bean in blue...





For *b* blue beads you have : $b!$ ways to arrange them

But $$b = N - a$$

$$\binom{N}{a} = \frac{N!}{a!\,.\,(N-a)!}$$

Some probabilities

Well known binomial distribution

$$p(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$$

$$\binom{n}{x}=\frac{n!}{p!(n-p)!}$$

Less well known multinomial distribution

$$p(N_1,N_2,N_3,...)=\frac{N!}{N_1!N_2!N_3!...}P_1^{N_1}P_2^{N_2}P_3^{N_3}...$$

And its application to DNA sequences

$$p(N_a, N_c, N_c, N_t) = \frac{N!}{\prod_{i \in \{a,c,g,t\}} N_i!} \prod_{i \in \{a,c,g,t\}} P_i^{N_i}$$

$$\sum_{i \in \{a,c,g,t\}} N_i = N \ , \sum_{i \in \{a,c,g,t\}} P_i = 1$$

Homozygote loci

	a	c	g	t
a	aa	ac	ag.	at
c		cc	cg	ct
g			gg	gt
t				tt

Probability to read a base at an homozygote loci XX

All errors are equiprobable

$$P(x = X) = 1 - P_{error}$$

$$P(x = Z, \forall\, Z \neq X) = \frac{P_{error}}{3}$$

Heterozygote loci

	a	c	g	t
a	aa	ac	ag.	at
c		cc	cg	ct
g			gg	gt
t				tt

Probability to read a base at an heterozygote loci XY

All errors are equiprobable

$$P(x = Z \,\forall\, Z \in \{X,Y\}) = \left(\frac{1}{2} - P_{error}\right) + \frac{1}{2}\frac{P_{error}}{3}$$
$$= \frac{1}{2} - \frac{5}{6}P_{error}$$

$$P(x = Z \,\forall\, Z \notin \{X,Y\}) = \frac{1}{2}\left(1 - P(x = X) - P(x = Y)\right)$$
$$= \frac{5}{6}P_{error}$$

Bayesian model

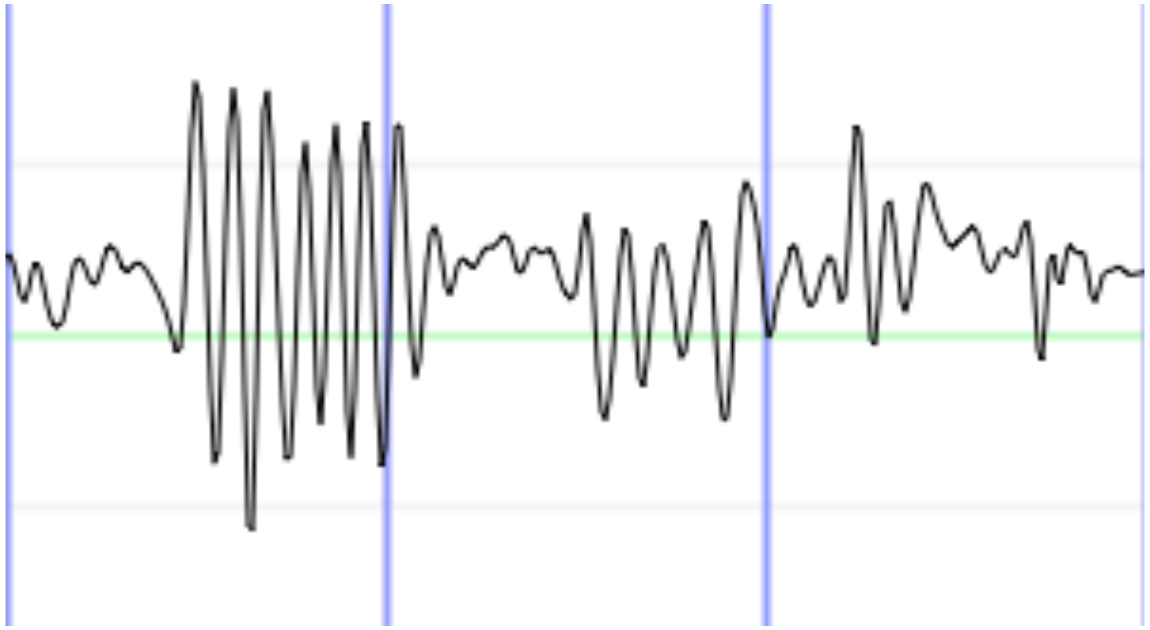
$$p(N_a, N_c, N_c, N_t|XY) = \frac{N!}{\prod_{i \in \{a,c,g,t\}} N_i!} \prod_{i \in \{a,c,g,t\}} P_i^{N_i}$$

$$P(N_a, N_c, N_c, N_t \cap XY) = P(N_a, N_c, N_c, N_t|XY)\; P(XY)$$
$$P(N_a, N_c, N_c, N_t \cap XY) = P(XY|N_a, N_c, N_c, N_t)\; P(N_a, N_c, N_c, N_t)$$

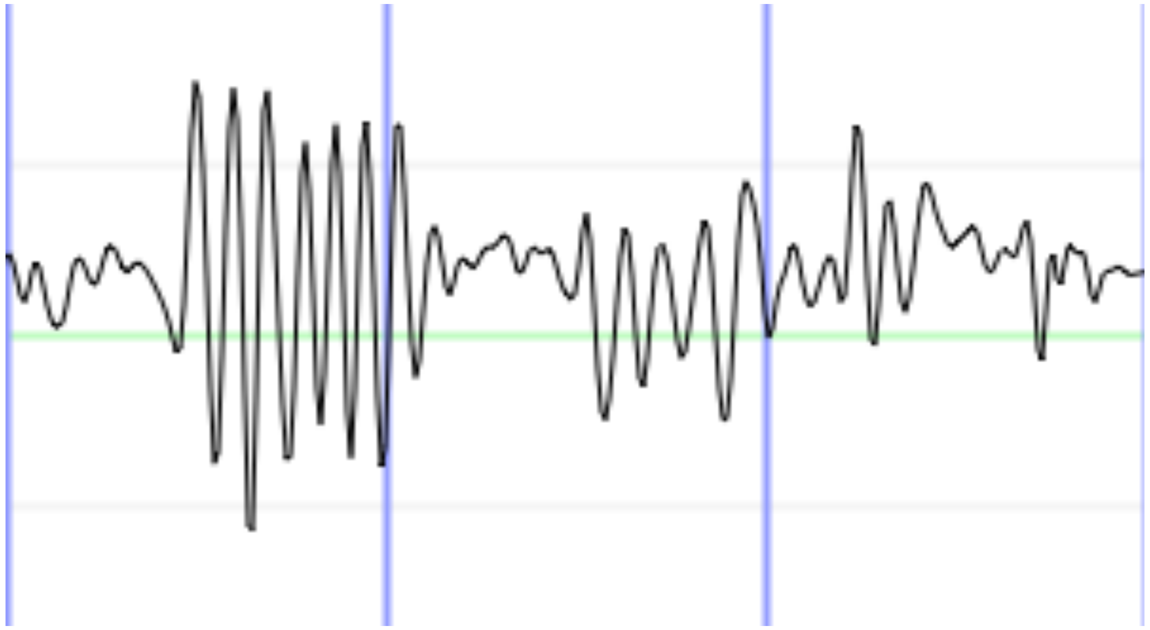
$$P(XY|N_a, N_c, N_c, N_t) = \frac{P(N_a, N_c, N_c, N_t|XY)\; P(XY)}{P(N_a, N_c, N_c, N_t)}$$

The earthquake observer in Victoria national park

A geologist, his bow tie and his seismograph

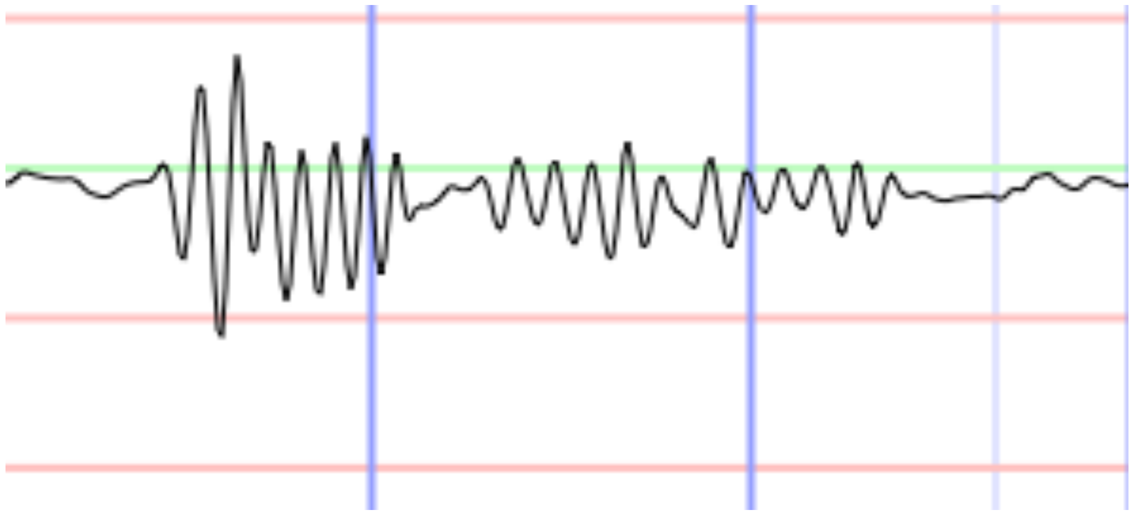


The earthquake observer in Victoria national park

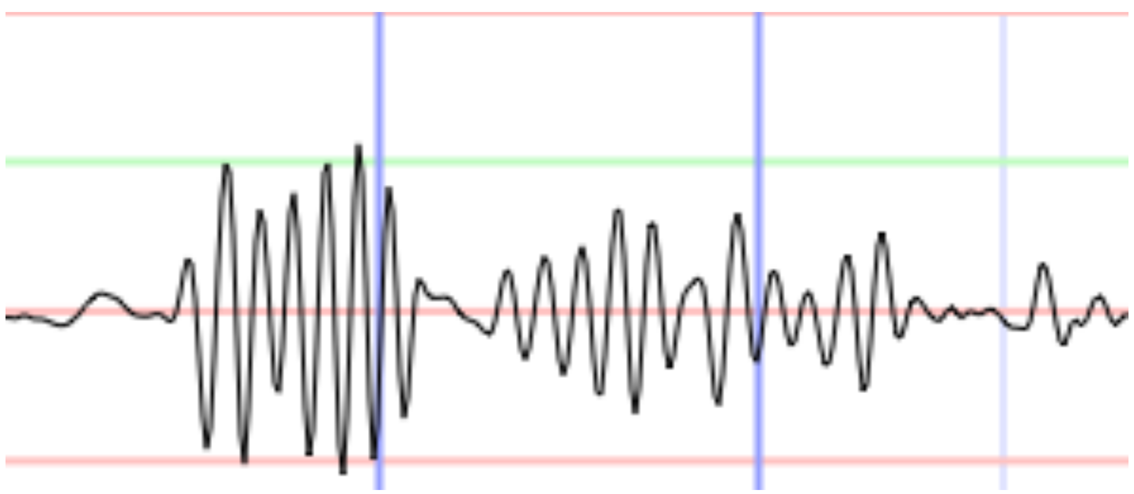


The both models

The elephant model

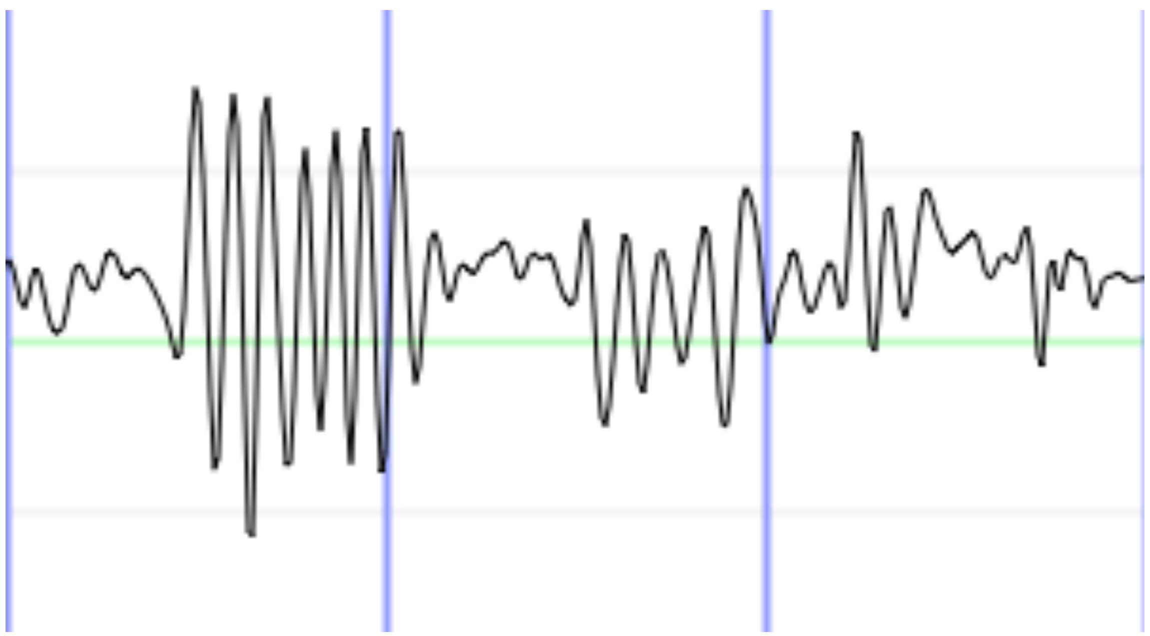


The earthquake model



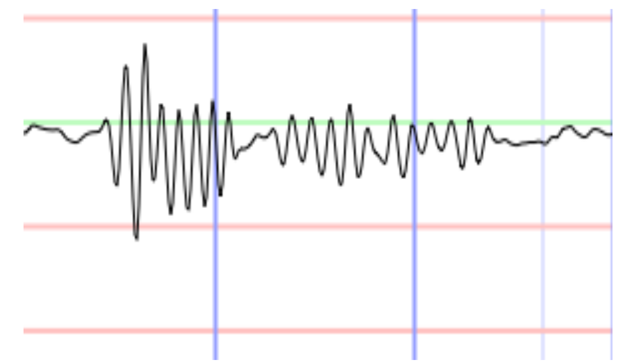
Recognize between both

A real recording : R



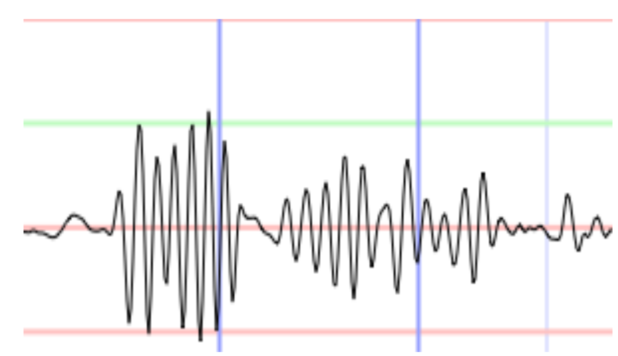
?

The elephant : E



$$P(R \mid E) = 0.1$$

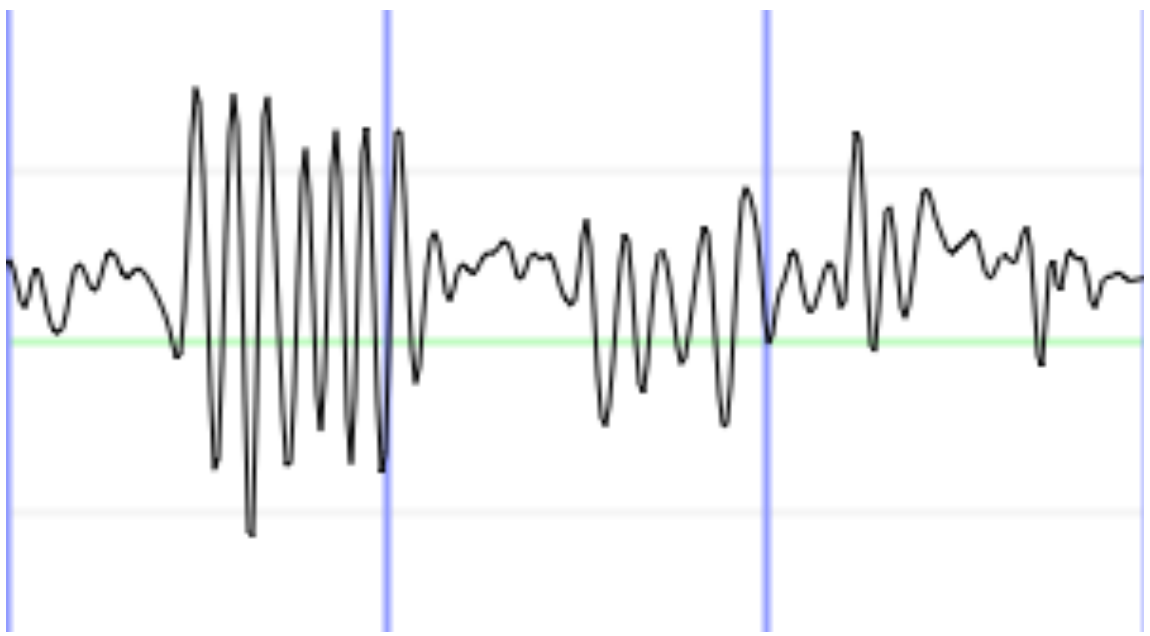
The earthquake : Q



$$P(R \mid Q) = 0.4$$

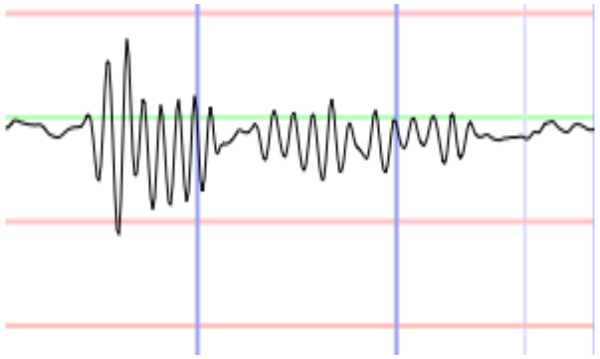
A priori knowledges

A real recording



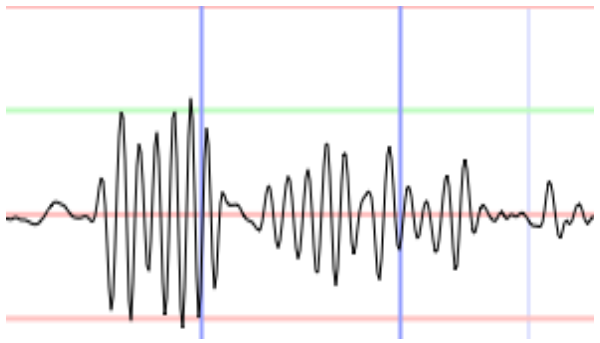
?

The elephant : E



$P(R \mid E) = 0.1$
$P(E) = 0.8$

The earthquake : Q



$P(R \mid Q) = 0.4$
$P(Q) = 0.2$

Probability of a model knowing an event

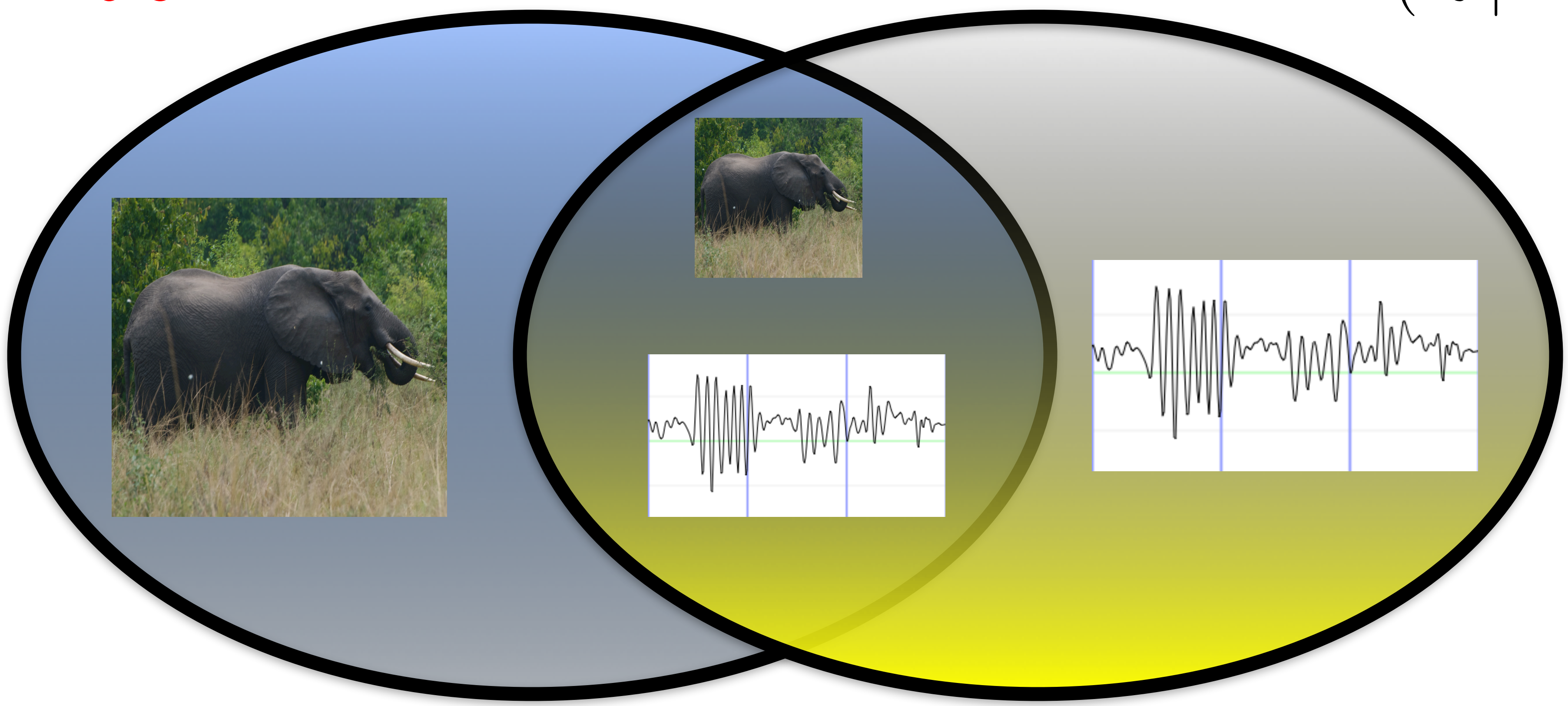
$$P(R \mid E) = 0.1$$

Probability of a model knowing an event

$$P(R \mid E) = 0.1$$

$P(E) = 0.8$

$P(R \mid E) = 0.1$



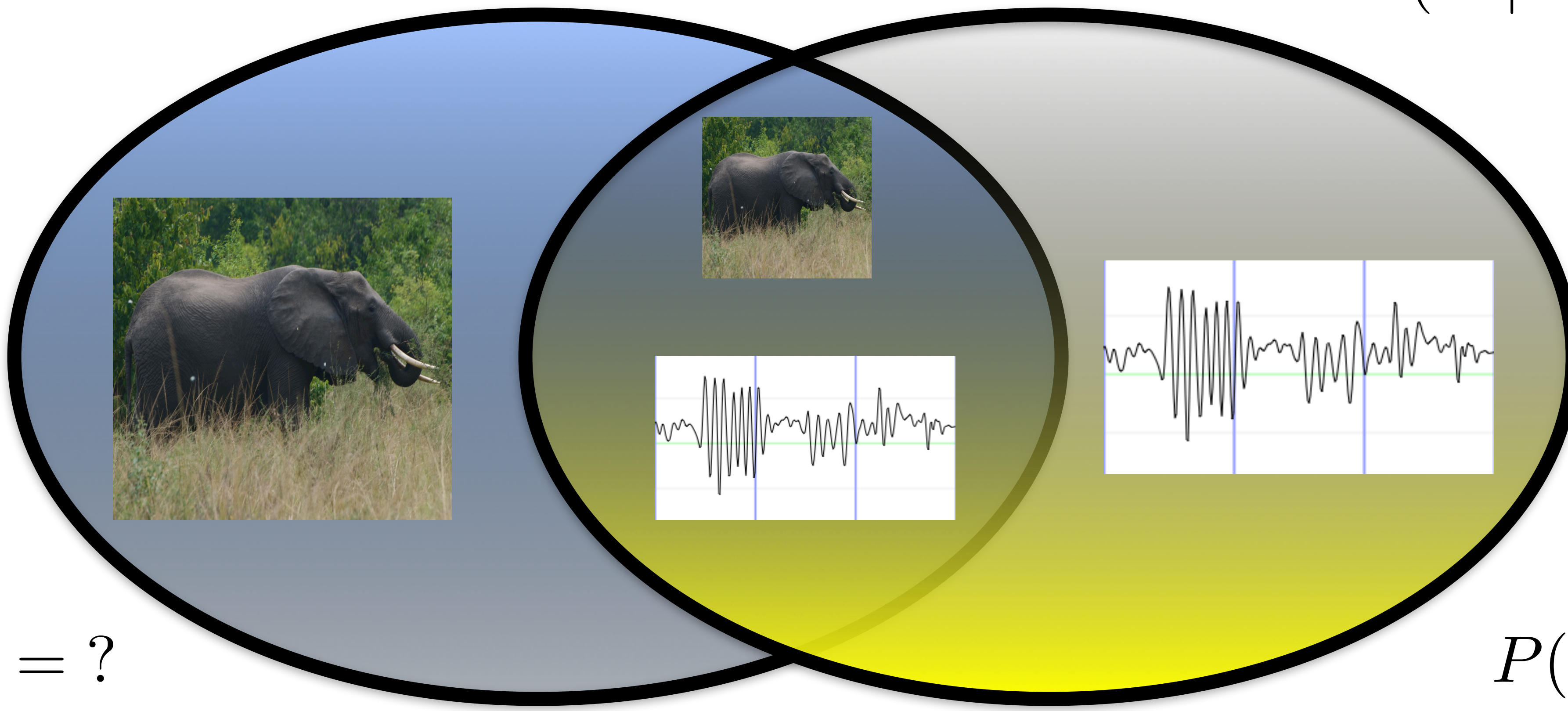
$$P(R \,\&\, E) = P(E) \,.\, P(R \mid E) = 0.08$$

Probability of a model knowing an event

$$P(R \mid E) = 0.1$$

$P(E) = 0.8$

$P(R \mid E) = 0.1$



$P(E \mid R) = ?$

$P(R) = ?$

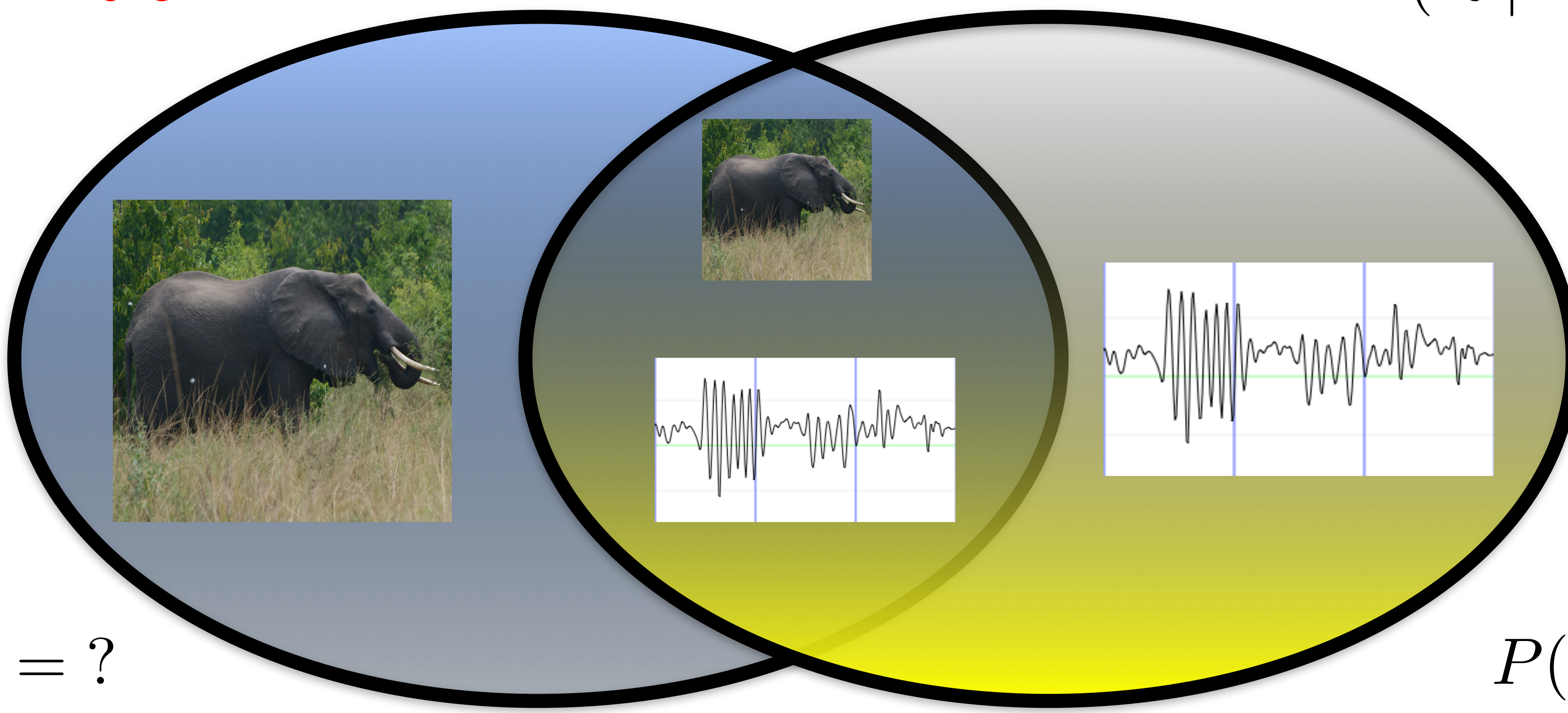
$$P(R \,\&\, E) = P(E) \,.\, P(R \mid E) = 0.08$$
$$= P(R) \,.\, P(E \mid R)$$

Probability of a model knowing an event

$$P(R \mid E) = 0.1$$

$P(E) = 0.8$

$P(R \mid E) = 0.1$



$P(E \mid R) = ?$

$P(R) = ?$

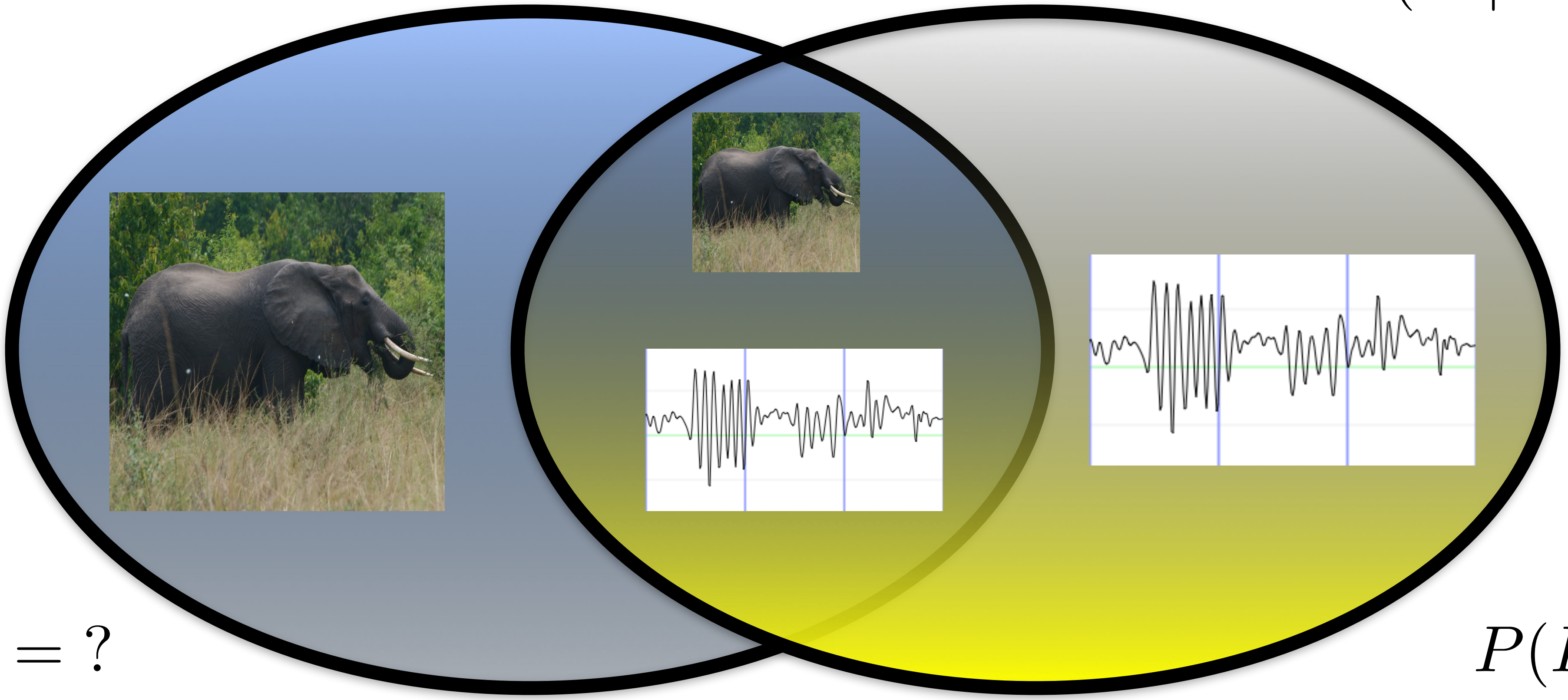
$$P(E)\ .\ P(R \mid E) = P(R)\ .\ P(E \mid R)$$

Thanks to bayes

$$P(R \mid E) = 0.1$$

$P(E) = 0.8$

$P(R \mid E) = 0.1$



$P(E \mid R) = ?$

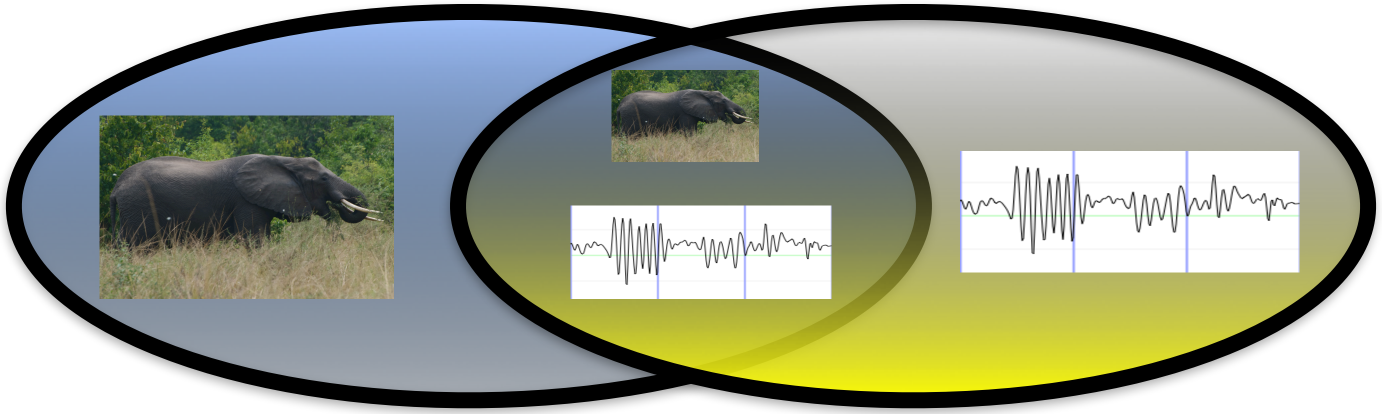
$P(R) = ?$

$$P(E) \,.\, P(R \mid E) = P(R) \,.\, P(E \mid R)$$

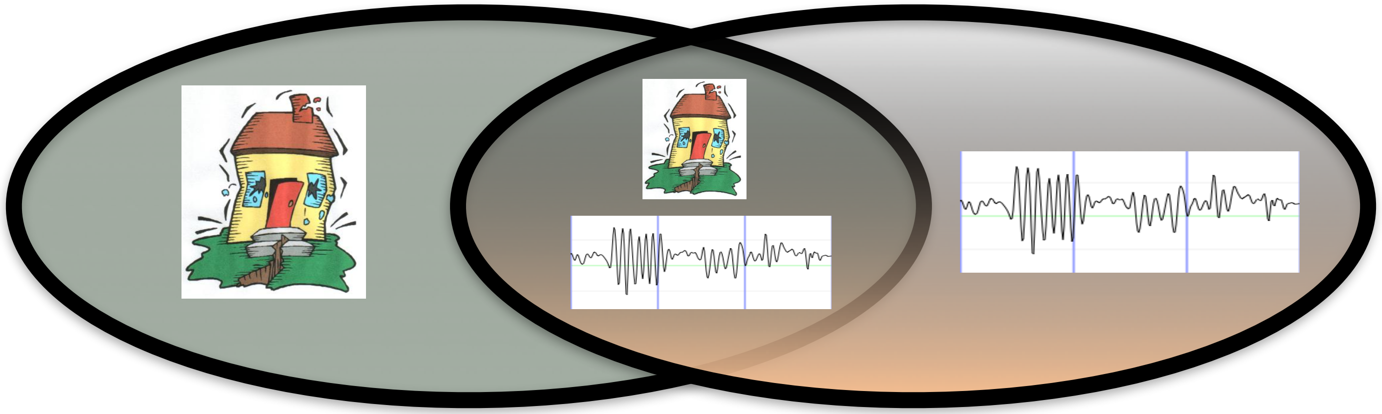
$$P(E \mid R) = \frac{P(R \mid E) \,.\, P(E)}{P(R)}$$

Bayes formula

Absolute probability of the record : P(R)

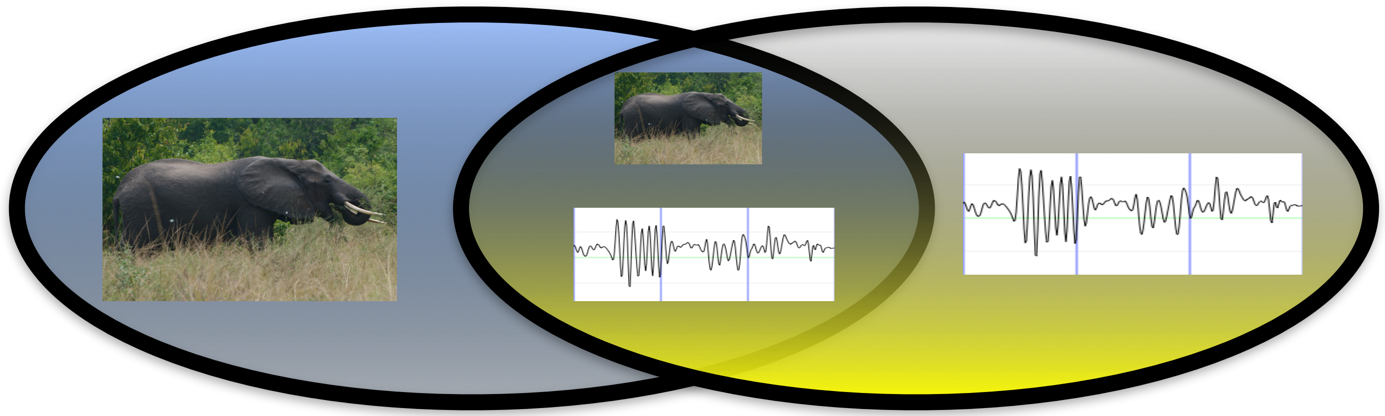


$$P(R\ \&\ E) = P(E)\ .\ P(R\,|\,E) = 0.08$$

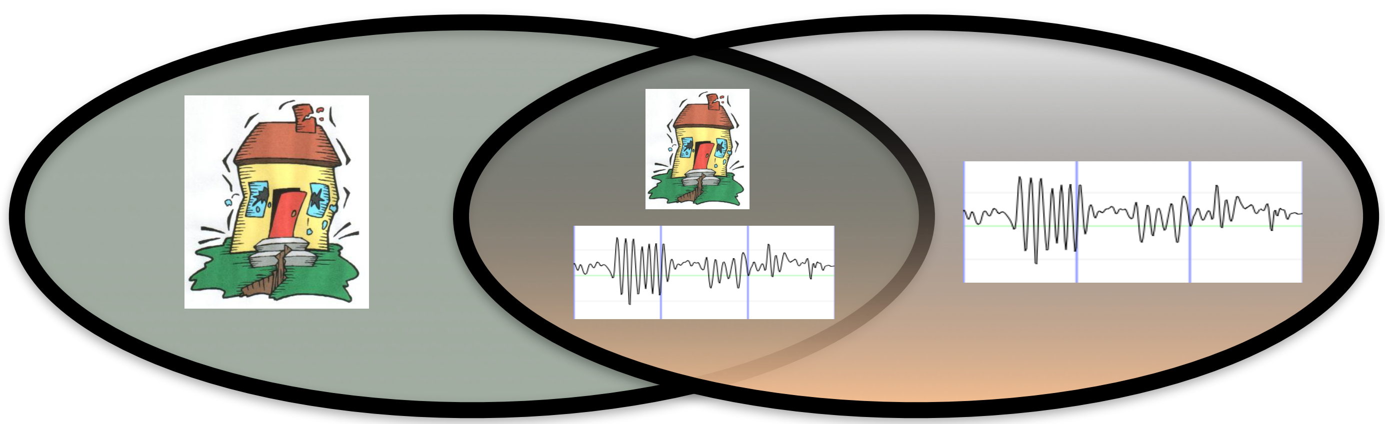


$$P(R\ \&\ Q) = P(Q)\ .\ P(R\,|\,Q)$$

Absolute probability of the record : P(R)



$$P(R \,\&\, E) = \color{red}{P(E)} \;.\; P(R \mid E) = 0.08$$



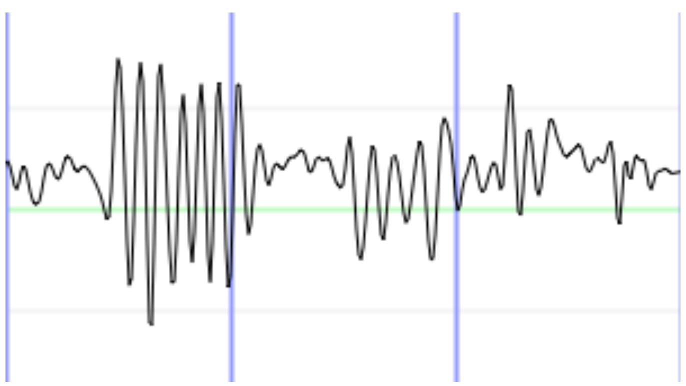
$$P(R \,\&\, Q) = \color{green}{P(Q)} \;.\; P(R \mid Q)$$

$$\begin{aligned} P(R) &= P(R \,\&\, Q) + P(R \,\&\, E) \\ &= \color{green}{P(Q)} \;.\; P(R \mid Q) + \color{red}{P(E)} \;.\; P(R \mid E) \end{aligned}$$

A priori knowledge can change conclusion

$$P(E \mid R) = \frac{P(R \mid E)\ .\ P(E)}{P(Q)\ .\ P(R \mid Q) + P(E)\ .\ P(R \mid E)}$$

$$= \frac{0.8\ \times\ 0.1}{(0.2\ \times\ 0.4) + (0.8\ \times\ 0.1)} = 0.5$$



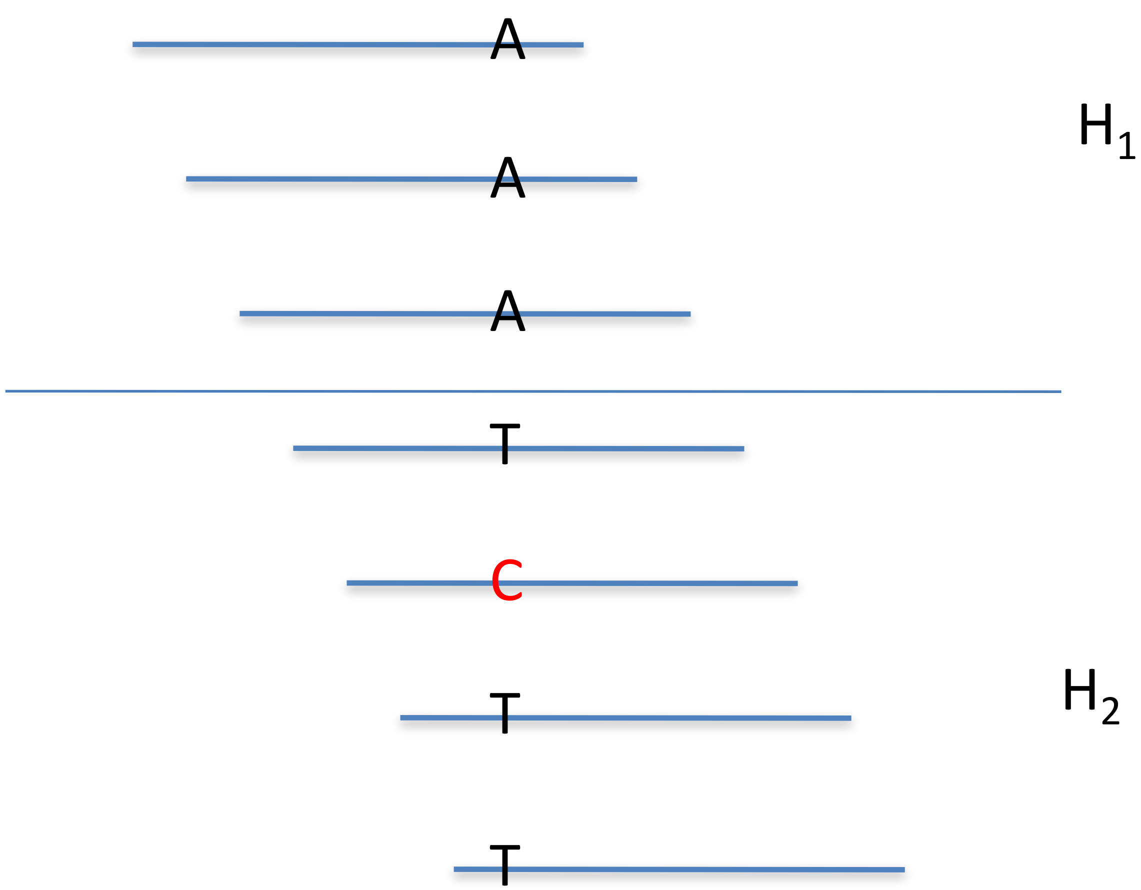
$$P(Q \mid R) = \frac{P(R \mid Q)\ .\ P(Q)}{P(Q)\ .\ P(R \mid Q) + P(E)\ .\ P(R \mid E)}$$

$$= \frac{0.2\ \times\ 0.4}{(0.2\ \times\ 0.4) + (0.8\ \times\ 0.1)} = 0.5$$

Genotype inference

$$P(XY|N_a, N_c, N_c, N_t) = \frac{P(N_a, N_c, N_c, N_t|XY)\ P(XY)}{P(N_a, N_c, N_c, N_t)}$$

$$P(N_a, N_c, N_c, N_t) = \sum_{i \in G} P(N_a, N_c, N_c, N_t|M_i)\ P(M_i)$$



	probability	
aa	1.5	10^{-3}
ac	2.9	10^{-4}
cc	1.7	10^{-8}
ag	4.9	10^{-6}
cg	1.4	10^{-9}
gg	5.8	10^{-11}
at	9.9	10^{-1}
ct	2.9	10^{-4}
gt	4.9	10^{-6}
tt	1.5	10^{-3}