What is the probability that at least 2 people in a group of size n have the same birthday?

minor assumptions:

- no Feb. 29
- $P(\text{person } i \text{ born on day } j) = \frac{1}{365}$

$n \leq 365$ otherwise by pigeonhole there must be a match

$P(\geq 2 \text{ of } n \text{ have same birthday})$

$P(\geq 1 \text{ of } \binom{n}{2} \text{ pairs is a birthday match})$

event $m(i,j)$: person i & person j match

$$= P \bigcup_{1 \leq i < j \leq n} m(i,j)$$

$$P\Big(m(1,2) \cup m(1,3) \cup m(1,4) \cup \dots \cup m(1,n) \quad \dots$$
$$\cup\ m(2,3) \cup m(2,4) \cup \dots \cup m(2,n) \quad \dots$$
$$\vdots$$
$$\cup\ m(n-3,n-1) \cup m(n-2,n-1) \dots$$
$$\cup\ m(n-1,n) \Big)$$

$$\leq \sum_{1 \leq i < j \leq n} P(m(i,j)) = \binom{n}{2} \cdot \frac{1}{365}$$

not useful, even as an upper bound

$$\overbrace{n<2 : P \leq 0}^{\text{accurate}}$$

$n=2$: $P \leq \frac{1}{365}$

$n=10$: $P \leq 12.4\%$

$n=28$: $P \leq 1.04$

$P(\geq 2 \text{ of } n \text{ have same birthday}) = 1 - P(\text{all } n \text{ birthdays are different})$

event $m(i,j)$: person i & person j match

$\overline{m(i,j)}$: person i & person j don't match

$$1 - P \underbrace{\bigcap}_{1 \leq i < j \leq n} \overline{m(i,j)} = 1 - P\Big(\overline{m(1,2)} \cap \overline{m(1,3)} \cap \overline{m(1,4)} \cap \ldots \cap \overline{m(1,n)} \quad \ldots$$

$$\overline{m(2,3)} \cap \overline{m(2,4)} \cap \ldots \cap \overline{m(2,n)} \quad \ldots$$

$$\vdots$$

$$\cap \ \overline{m(n-3,n-1)} \cap \overline{m(n-2,n-1)} \ldots$$

$$\cap \ \overline{m(n-1,n)} \Big)$$

Problem: these aren't independent

$P(\geq 2 \text{ of } n \text{ have same birthday}) = 1 - \underbrace{P(\text{all } n \text{ birthdays are different})}_{P_n}$

- consider people iteratively (in order)
- event H_i : person i doesn't match previous $i-1$ people

$P_1 = P(H_1) = 1$

$P_2 = P(H_1 \cap H_2) = P(H_1) \cdot P(H_2 | H_1) = 1 \cdot \frac{364}{365} = P(H_2)$

$P_3 = P(H_1 \cap H_2 \cap H_3) = P(H_1) \cdot P(H_2 | H_1) \cdot P(H_3 | (H_2 \cap H_1)) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365}$

$P_4 = P(H_1 \cap H_2 \cap H_3 \cap H_4) = P(H_1) \cdot P(H_2 | H_1) \cdot P(H_3 | (H_2 \cap H_1)) \cdot P(H_4 | (H_3 \cap H_2 \cap H_1))$

$= 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365}$

$$P(\geq 2 \text{ of } n \text{ have same birthday}) = 1 - \underbrace{P(\text{all } n \text{ birthdays are different})}_{P_n}$$

$$P_4 = P(H_1 \cap H_2 \cap H_3 \cap H_4) = P(H_1) \cdot P(H_2 | H_1) \cdot P(H_3 | (H_2 \cap H_1)) \cdot P(H_4 | (H_3 \cap H_2 \cap H_1))$$

person: 1 2 3 4 ... n

$$P_n = P(H_1 \cap \ldots \cap H_n) = P\left(\bigcap_{i=1}^{n} H_i\right) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdots \frac{365-(n-1)}{365}$$

$$= \frac{P(365, n)}{365^n} = \frac{\frac{365!}{(365-n)!}}{365^n}$$

{ ways to choose n from 365 (no repetition, order matters)

← sample space (repetition, order matters)

$P(\geq 2 \text{ of } n \text{ have same birthday}) = 1 - P(\text{all } n \text{ birthdays are different})$

$$1 - \frac{\frac{365!}{(365-n)!}}{365^n}$$

$n = 2 \rightarrow 0.27\% = \frac{1}{365}$

$n = 4 \rightarrow 1.64\%$

$n = 10 \rightarrow 11.7\%$

$$\boxed{\sum_{1 \leq i < j \leq n} P(m(i,j)) = \binom{n}{2} \cdot \frac{1}{365} \qquad n = 10 : P \leq 12.4\%}$$

decent bound

$P(\geq 2 \text{ of } n \text{ have same birthday}) = 1 - P(\text{all } n \text{ birthdays are different})$

$$1 - \frac{\frac{365!}{(365-n)!}}{365^n}$$

$n = 2 \rightarrow 0.27\% = \frac{1}{365}$

$n = 4 \rightarrow 1.64\%$

$n = 10 \rightarrow 11.7\%$

$n = 23 \rightarrow 50.7\%$

$n = 30 \rightarrow 70.6\%$

$n = 70 \rightarrow 99.9\%$

$n \sim 116 \rightarrow 1 - \frac{1}{10^9}$

$n \sim 127 \rightarrow 1 - \frac{1}{10^{11}}$ { estimated number of stars in our galaxy

$n \sim 300 \rightarrow 1 - \frac{1}{10^{80}}$ { estimated number of atoms in the universe

$n > 365 \rightarrow 100\%$

TESTING FOR A DISEASE

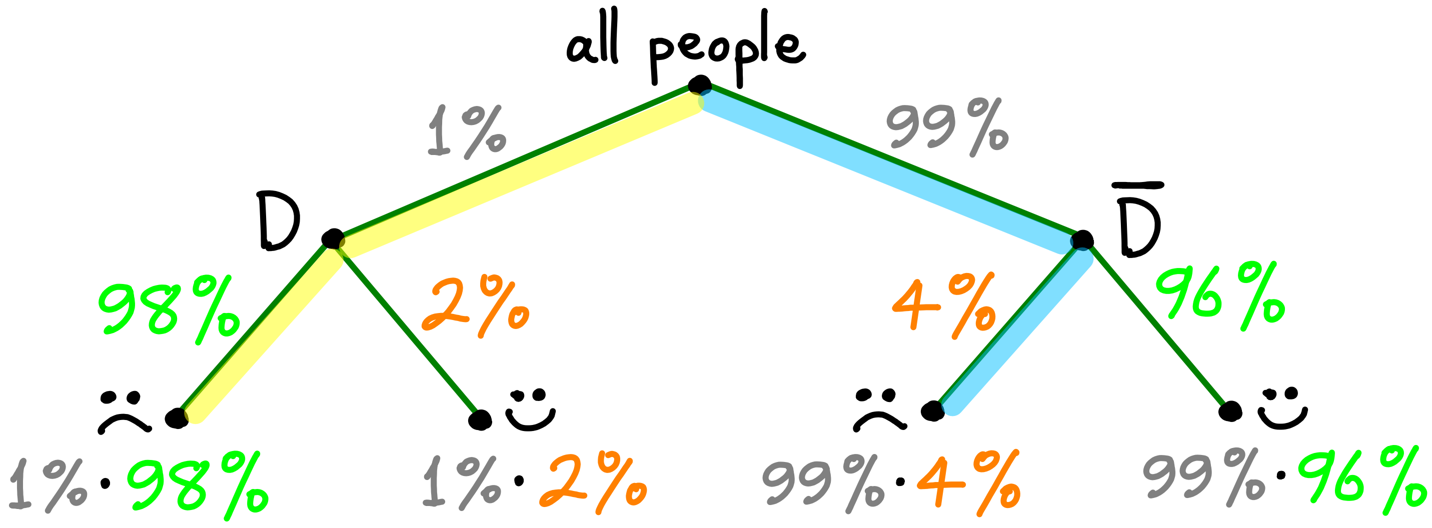
- suppose 1% of the population has a disease.
- There is a diagnostic test with the following performance:
  - if a person with the disease is tested,
    - ↳ the test confirms this 98% of the time. : true positive
    - → 2% : the test says the person is ok. : false negative (BAD)
  - if a person without the disease is tested,
    - ↳ the test confirms this 96% of the time. : true negative
    - → 4% : the test says the person has the disease. : false positive (also bad)
- What is the probability someone has the disease if they test positive?

D : have disease $\qquad \bar{D}$: don't have disease

- What is the probability someone has the disease if they test positive?

Reality :	1%	$P(D)$	99%	$P(\bar{D})$
Test ☹	98%	$P(☹ \mid D)$	4%	$P(☹ \mid \bar{D})$
Test ☺	2%	$P(☺ \mid D)$	96%	$P(☺ \mid \bar{D})$

$P(D \mid ☹) = ?$



$1\% \cdot 98\% = P(☹ \cap D)$

$99\% \cdot 4\% = P(☹ \cap \bar{D})$

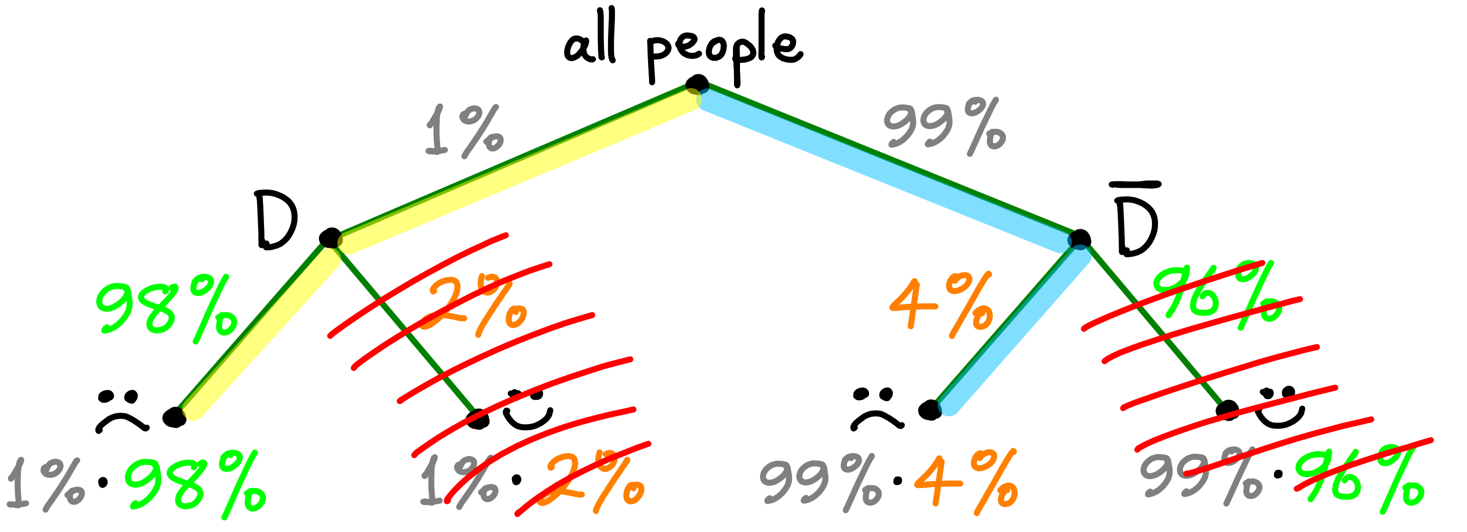
D : have disease $\qquad$ $\bar{D}$: don't have disease

- What is the probability someone has the disease if they test positive?

Reality :	1%	$P(D)$	99%	$P(\bar{D})$
Test ☹	98%	$P(☹ \mid D)$	4%	$P(☹ \mid \bar{D})$
Test ☺	2%	$P(☺ \mid D)$	96%	$P(☺ \mid \bar{D})$

$P(D \mid ☹) = ?$

universe shrank



$1\% \cdot 98\% = P(☹ \cap D)$

$99\% \cdot 4\% = P(☹ \cap \bar{D})$

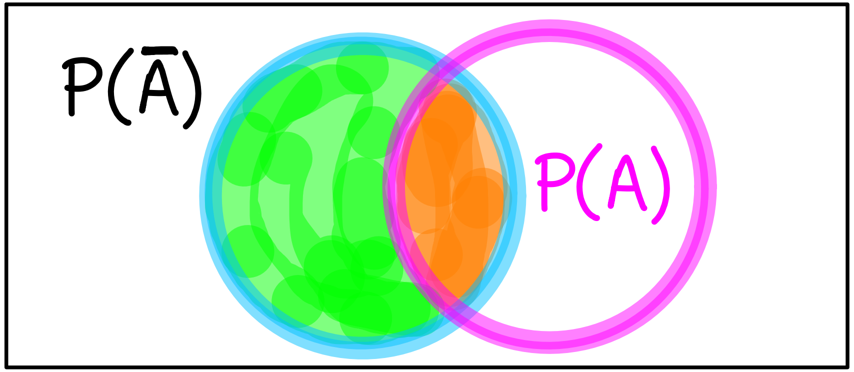
new valid outcomes

$$\boxed{\frac{1\% \cdot 98\%}{1\% \cdot 98\% + 99\% \cdot 4\%}}$$

new sample space

$$P(A \cap B) = P(A|B) \cdot P(B)$$
$$= P(B|A) \cdot P(A)$$

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



$$P(B) = \begin{cases} P(B|A) \cdot P(A) = P(A \cap B) \\ + \\ P(B|\bar{A}) \cdot P(\bar{A}) = P(\bar{A} \cap B) \end{cases}$$

BAYES' THEOREM

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B|A)\cdot P(A) + P(B|\bar{A})\cdot P(\bar{A})}$$

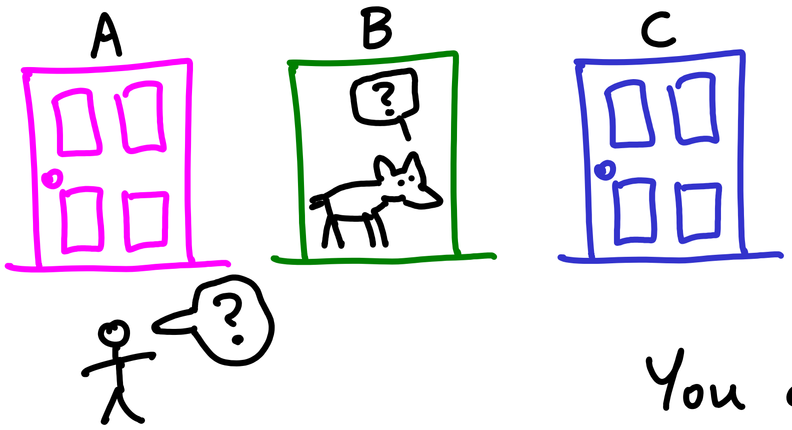
D : have disease

Reality :	1%	P(D)	99%	P($\bar{D}$)
Test ☹	98%	P(☹ | D)	4%	P(☹ | $\bar{D}$)
Test ☺	2%	P(☺ | D)	96%	P(☺ | $\bar{D}$)

P(D | ☹) = ?

$$P(D|\text{☹}) = \frac{P(\text{☹}|D)\cdot P(D)}{P(\text{☹}|D)\cdot P(D) + P(\text{☹}|\bar{D})\cdot P(\bar{D})} = \frac{98\%\cdot 1\%}{98\%\cdot 1\% + 4\%\cdot 99\%} \sim 19.8\%$$

The Monty Hall problem



1 of these 3 doors hides a car.
The other 2 hide goats.

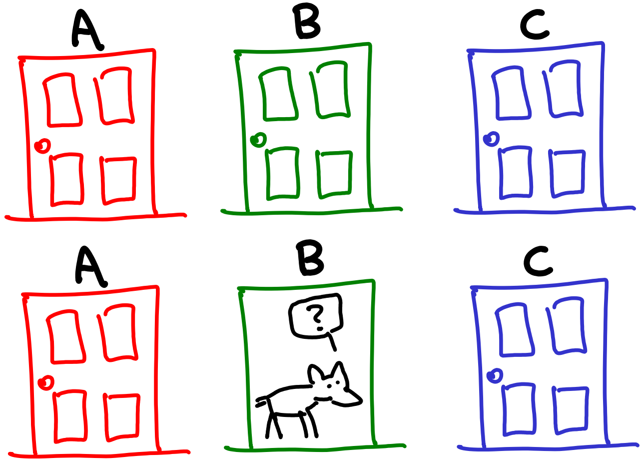
You get to choose a door and keep what's behind.

Without loss of generality → 1) you choose door A.

just some door with a goat → 2) door B is opened, revealing a goat.

You're given a choice: → keep A OR switch to C ?

An incorrect analysis:



$\} \; \bar{B}$

- event i: car is at door i

$P(A) = P(B) = P(C) = \frac{1}{3} \qquad P(\bar{A}) = P(\bar{B}) = P(\bar{C}) = \frac{2}{3}$

Bayes' theorem $\rightarrow$ $P(A|\bar{B}) = \underbrace{\frac{P(\bar{B}|A)\cdot P(A)}{P(\bar{B})}}_{}$

$\downarrow$

$$P(C|\bar{B}) = \frac{P(\bar{B}|C)\cdot P(C)}{P(\bar{B})} = \frac{1\cdot \frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

$\hookrightarrow$ conclusion: it makes no difference

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

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

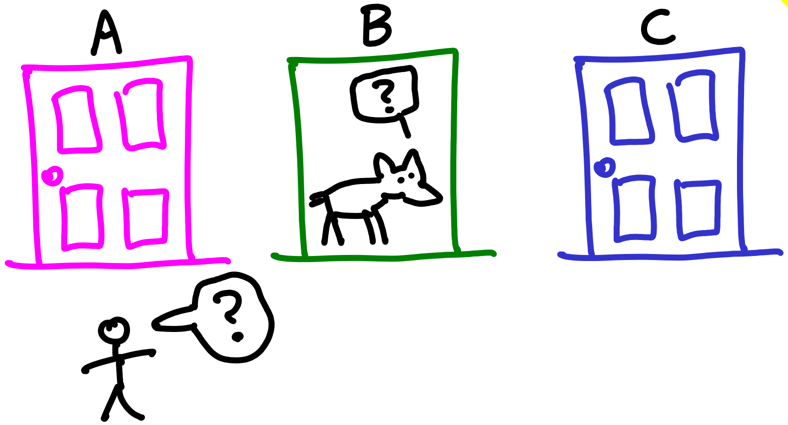
- What is wrong? $\rightarrow$ You are not in the picture.
- if you hadn't selected A, this analysis would be correct.

Without loss of generality → 1) you choose door A.

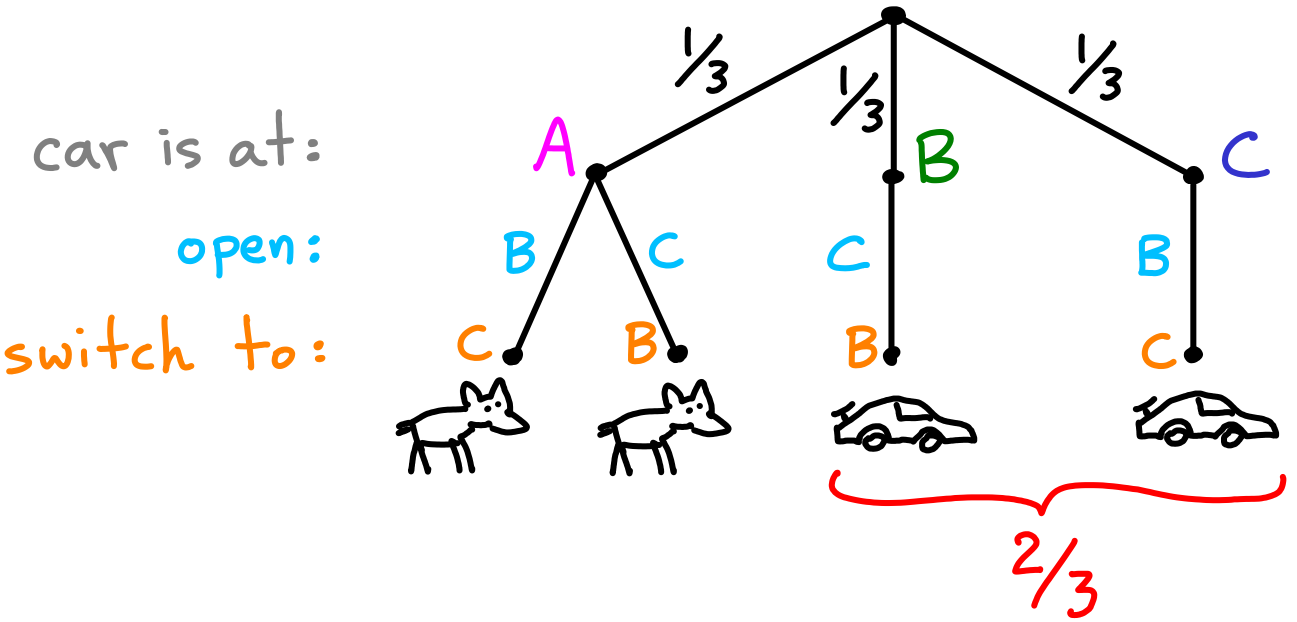
→ 2) door B is opened, revealing a goat.

just some door with a goat →

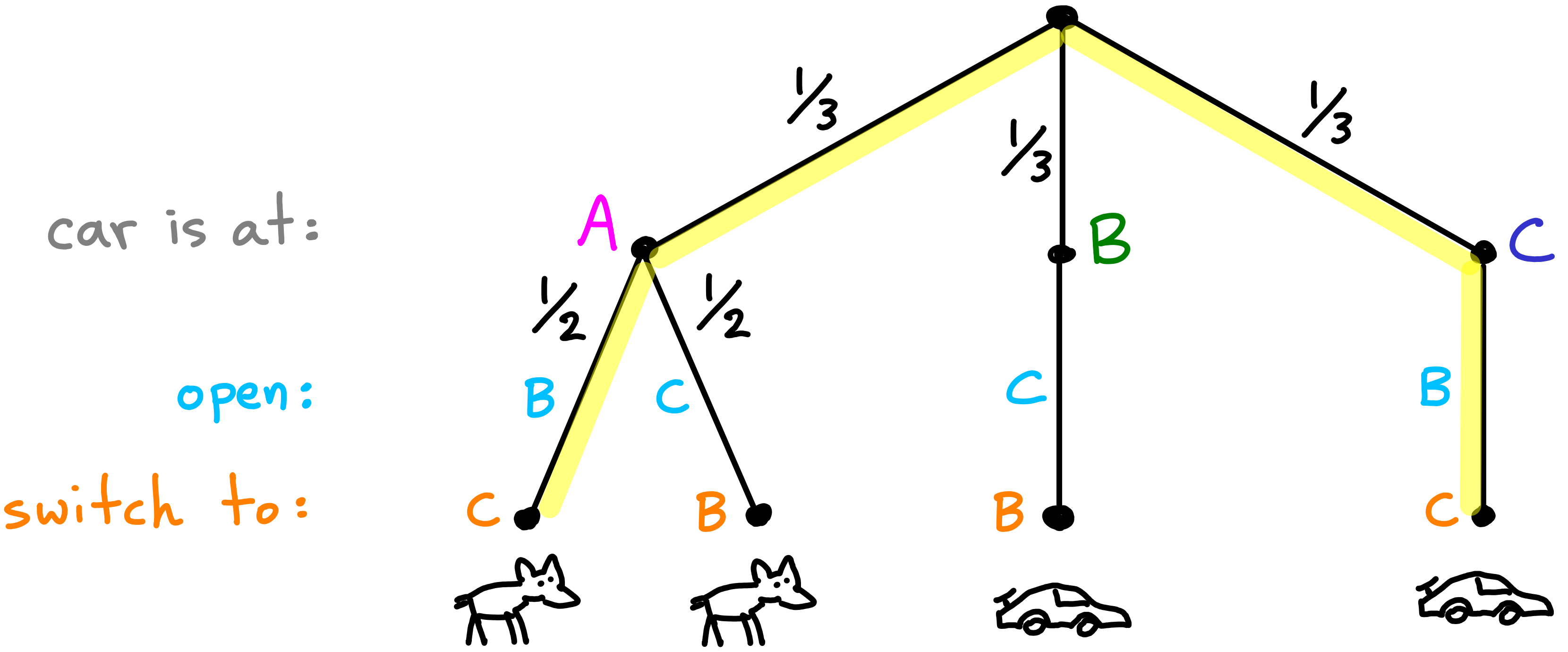
if the car had been at B,
then C would have opened.
(doors could have been renamed)



strategy:
switch



2 scenarios consistent with B opened

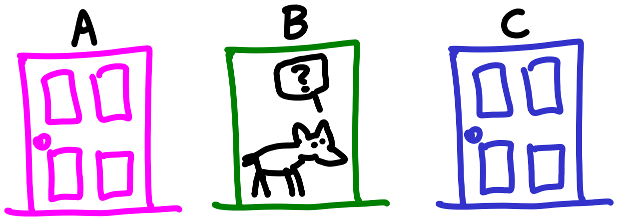


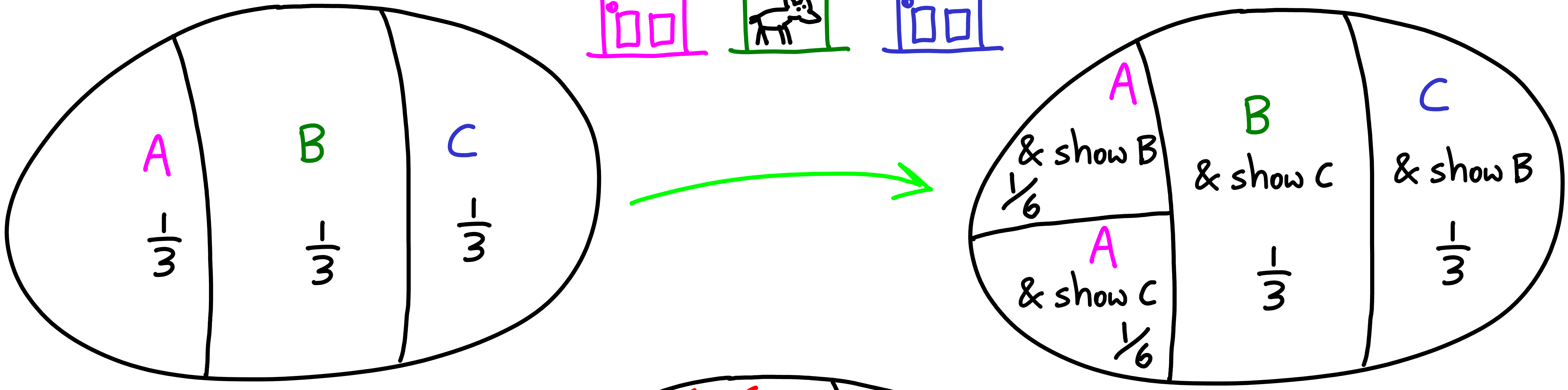
$$\frac{1}{2}\cdot\frac{1}{3}=\frac{1}{6}$$

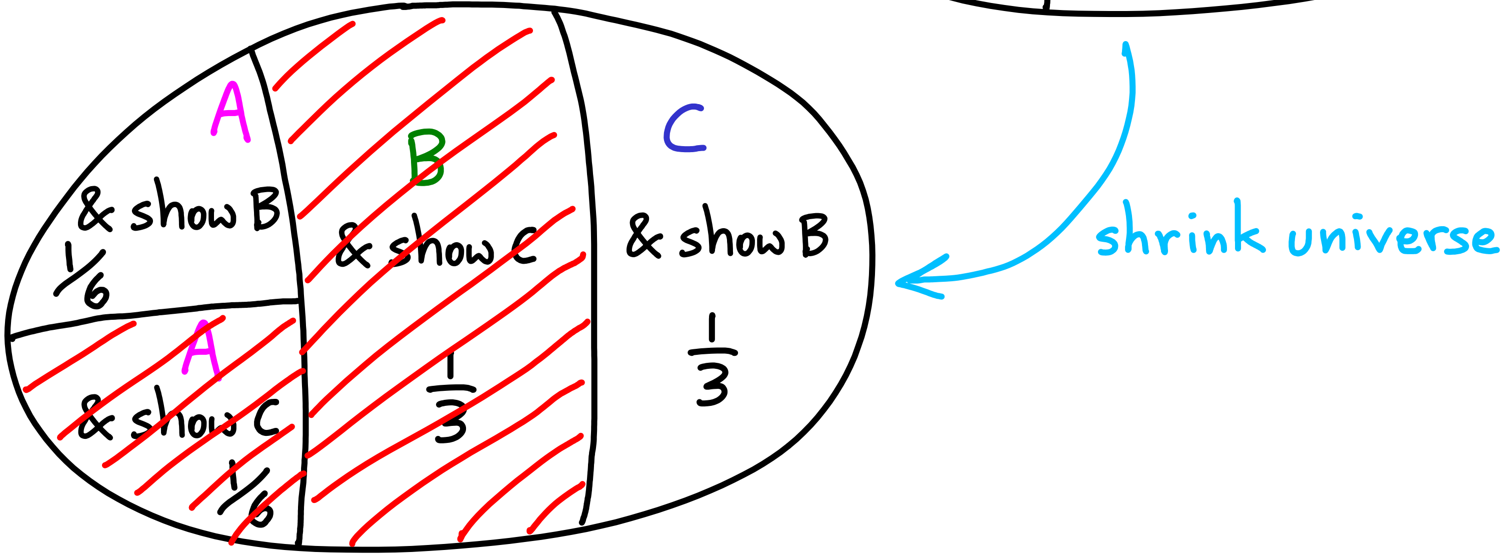
$$\underbrace{\frac{1}{3}}$$

$$\frac{\frac{1}{3}}{\frac{1}{2}}=\frac{2}{3}$$

new sample space $=\frac{1}{6}+\frac{1}{3}=\frac{1}{2}$







$$P(C \cap \text{show}B) = \frac{\frac{1}{3}}{\frac{1}{6}+\frac{1}{3}} = \frac{2}{3}$$