CS5010 Artificial Intelligence Principles

Lecture 10 Uncertainty

Probability theory

Lei Fang

University of St Andrews

About me (the other lecturer)

Lei (pronounced as `Lay`) Fang

- Lecturer in School of Computer Science, St Andrews
    - Background in Computer Science (1st degree)
    - Ph.D. in statistical learning stuff
- Mostly working on statistical learning, Bayesian machine learning etc nowadays
- Office in Jack Cole Building, School of Computer Science
- Email: lf28@st-andrews.ac.uk
- Office hour by appointment, either in-person or on Teams, just email me

What to cover for the rest of the course

Uncertainty (coming two-three weeks)

Uncertainty (coming two-three weeks)

In a nutshell, use **probability theory** to

- equip machine with some uncertainty reasoning capability
- human intelligence unconciously does it all the time, e.g.
    - we judge how likely to win a lottery and invest accordingly
    - bring an umbrella or not by weighing the likelihood of raining

Uncertainty (coming two-three weeks)

In a nutshell, use **probability theory** to

- equip machine with some uncertainty reasoning capability
- human intelligence unconciously does it all the time, e.g.
    - we judge how likely to win a lottery and invest accordingly
    - bring an umbrella or not by weighing the likelihood of raining
- Human also known to be notoriously bad at uncertainty reasoning
    - we will see a few examples
- We will have a revision on probability theory today
- Practical 2 is about uncertainty and AI
    - shall we use AI to replace some uncertainty management in society?
    - more about this later ...

Searching (next next two weeks)

Searching (next next two weeks)

Equip machine with some problem solving skills via **searching**

Searching (next next two weeks)

Equip machine with some problem solving skills via **searching**

- It turns out a lot of problems are just searching problem
    - with some modelling/transformation process (abstraction)
    - e.g.
        - finding a route from St Andrews to Edinburgh
        - solving a sudoku puzzle
- Again, human uncontiously does it sometimes
    - machine is probably better than us at this for certain problems

Searching (next next two weeks)

Equip machine with some problem solving skills via **searching**

- It turns out a lot of problems are just searching problem
    - with some modelling/transformation process (abstraction)
    - e.g.
        - finding a route from St Andrews to Edinburgh
        - solving a sudoku puzzle
- Again, human uncontiously does it sometimes
    - machine is probably better than us at this for certain problems
- We will learn a range of searching algorithms
    - search is a field well studied in CS, maths and beyond
    - learning objective: be able to compare and constrast them
        - pick a suitable algorithm for your problem
- More on this in the last two weeks

This lecture

This lecture

- Notion of uncertainty
- Probability theory

This lecture

- Notion of uncertainty
- Probability theory

Relevant book chapters of AI-AMA

- Chapter 13 Quantifying uncertainty
- Chapter 14 Probabilistic reasoning

Why we need probability theory and uncertainty reasoning ?

Let's start with a sad but relevant story

The case of Sally Clark - "one of the great miscarriages of justice in modern British legal history"

- Sally Clark, a solictor in Cheshire, also a mother of two

Let’s start with a sad but relevant story

The case of Sally Clark - ”one of the great miscarriages of justice in modern British legal history”

- Sally Clark, a solictor in Cheshire, also a mother of two
- Sadly, both of Sally's children died out of sudden infant death syndrome (SIDS), a rare desease with a chance at
- In Nov 1999, following the death of her second child, Sally Clark was convicted of murder at Chester Crown Court

A key arugment in court: a peaedistrician professor, Sir Roy Meadow, testifying the chance of two children from an affluent family suffering cot death was 1 in 73 million $\frac{1}{8543} \times \frac{1}{8543}$

- the jurors swayed by the professor and gave a guilty verdict in 1999
- and the uncertainty reasoning (statistics) used was completely wrong

Sally was convicted of murdering both of her infant kids

Sally Clark was eventually exonerated and freed after serving 3 years in prison

She suffered a number of severe psychiatric problems and died sadly in 2007

Uncertain reasoning done by human intelligence

Uncertain reasoning done by human intelligence

- Clearly, there are some uncertainty reasoning in Sally Clark's case
- The chance of a SIDS happening in an affuent family is

    - it means 1 out of 8541 middle class families with a child death, and the death is attributed to SIDS
    - which is the level of uncertainty or *degree of belief*
- The professor believes the chance of two children died from both SIDS is

  1 in 73 million chance

- At a surface level, this uncertainty reasoning seems "reasonable"
- But **very wrong**, we will see why later and do the inference properly at a later lecture

Another example bad example: COVID vaccine protection rate

uncertainty inference by humans

Another example bad example: COVID vaccine protection rate

uncertainty inference by humans

- Deaths counts due to COVID in the UK in two age groups (I cited from government SAGE report between week 32 and week 35 2021)

age/death	Not Vaccinated	Double Vaccinated
70-79	129	428
80+	155	928

Another example bad example: COVID vaccine protection rate

uncertainty inference by humans

- Deaths counts due to COVID in the UK in two age groups (I cited from government SAGE report between week 32 and week 35 2021)

age/death	Not Vaccinated	Double Vaccinated
70-79	129	428
80+	155	928

- Anti-vac argue COVID vaccine are useless
- They claim: double vaccinated is more likely to die !
    - age 70-79, double vac death rate is
    - age 80+,
- Another uncertainty reasoning done by human and awfully **wrong**
    - which again looks reasonable at surface level
    - we will see how to do it properly next lecture

But what exactly done wrong ?

But what exactly done wrong ?

AND how to do it properly ?

But what exactly done wrong ?

AND how to do it properly ?

Hope the two examples have motivated you enough

But what exactly done wrong ?

AND how to do it properly ?

Hope the two examples have motivated you enough

We will see how to use **probability theory** to do uncertainty reasoning properly

- find the correct probability that Sally's guilty and the protection rate of COVID vaccine
- you can appreciate the importance of **probability theory** in proper reasoning

Before we start, something to reflect though ...

- how easy homo-sapiens can be misled and how un-intelligent we are
- the worst: we donot realise our igorance but hold it (mostly just predudice) dearly and firmly
    - even a well respected professor made such a mistake
- how to train a "humble" AI ?
    - admit one's ignorance is actually intelligent!
    - same applies to AI
    - "not very sure about this case" is better than a 100\% confident but wrong answer

Probability theory

Probability space (the axioms)

Probability space (the axioms)

A probability space consists of three elements , *a triple*

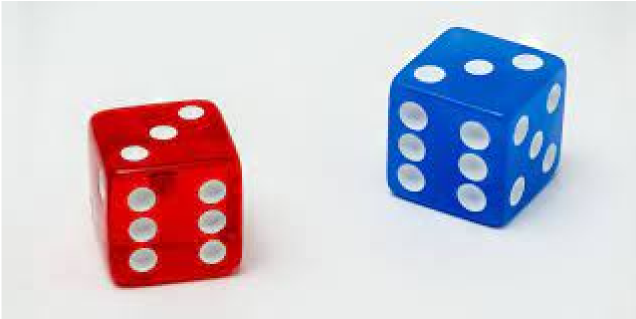
Sample space : the set of all possible worlds

- e.g. the result of rolling a 6 facet die

Probability space (the axioms)

A probability space consists of three elements , *a triple*

Sample space : the set of all possible worlds



- e.g. the result of rolling a 6 facet die
- e.g. the experiment of flipping a coin twice
- needs to be **exhaustive** and **mutually exclusive**
    - **exhaustive**: should contain all the possibilities
    - **mutually exclusive**: only one of the outcome is possible at a time

An **event** and its collection **event space**

- e.g. for the dice case, means the outcome is an even number
- e.g. for the coin case, we can define an as at least one head turning up:

An **event** and its collection **event space**

- e.g. for the dice case, means the outcome is an even number
- e.g. for the coin case, we can define an as at least one head turning up:

- is an event, called **certain event**
- is also an event, **null event**
- the collection of events we *care about* is called **event space**
    - for a discrete sample space, the event space can simply be
        - 6 facet die example,
        - coin tossing (one toss) example,

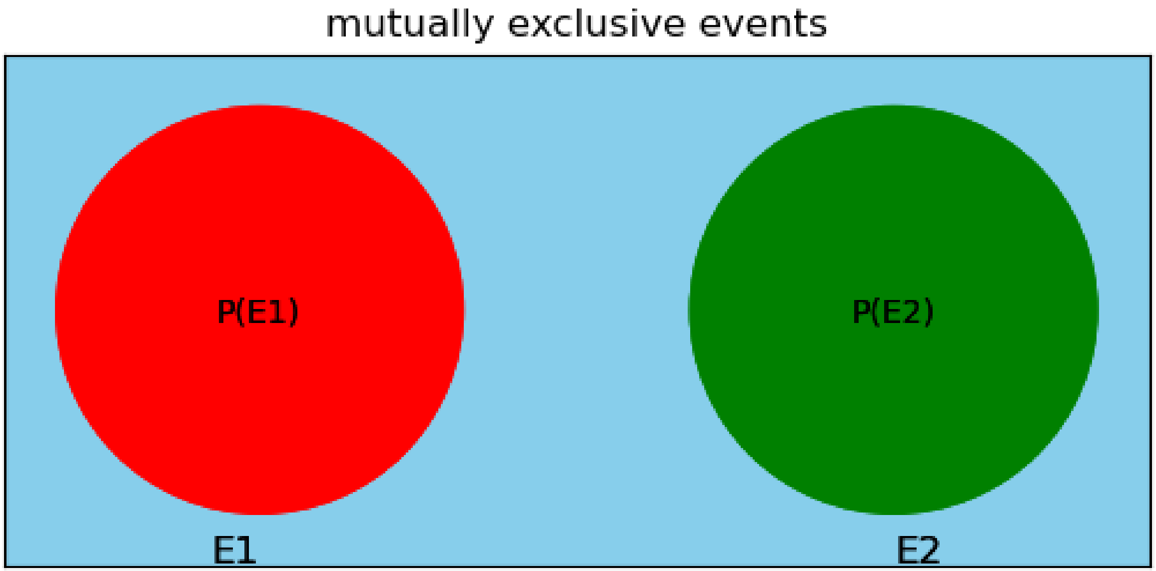
A **probability measure** : that assigns event a probability

- must satisfy:
  - (certain event) and
  - for any
  - $P(E_1 \cup E_2) = P(E_1)$
    $+ P(E_2)$
    if and are mutually exclusive

P
$+$

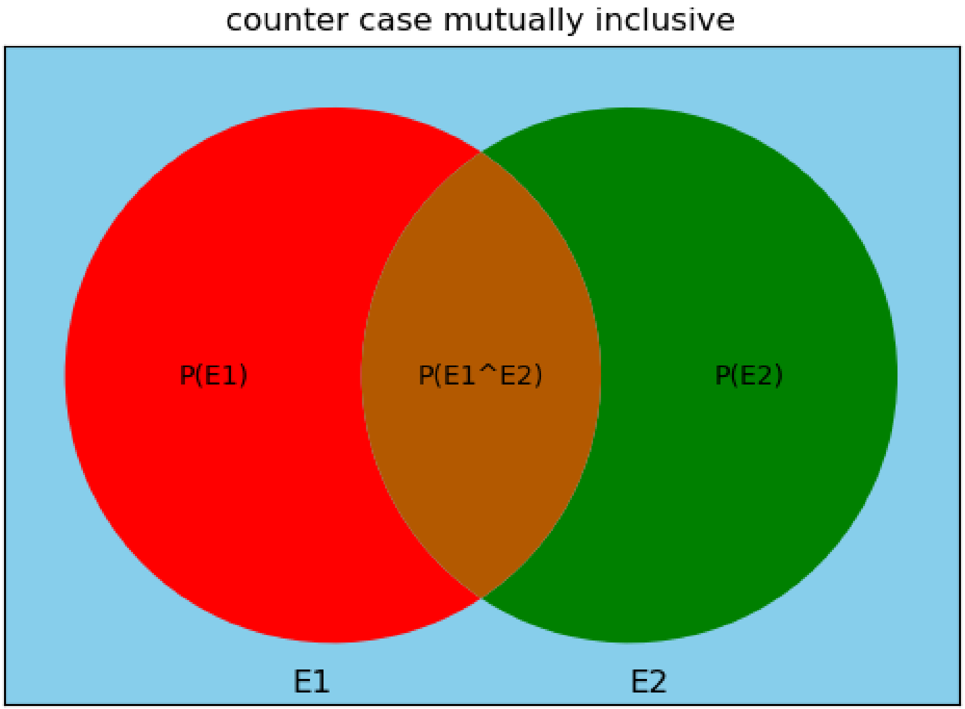
A **probability measure** : that assigns event a probability

- must satisfy:
  - (certain event) and
  - for any
  - if and are mutually exclusive



If and are not mutually exclusive,

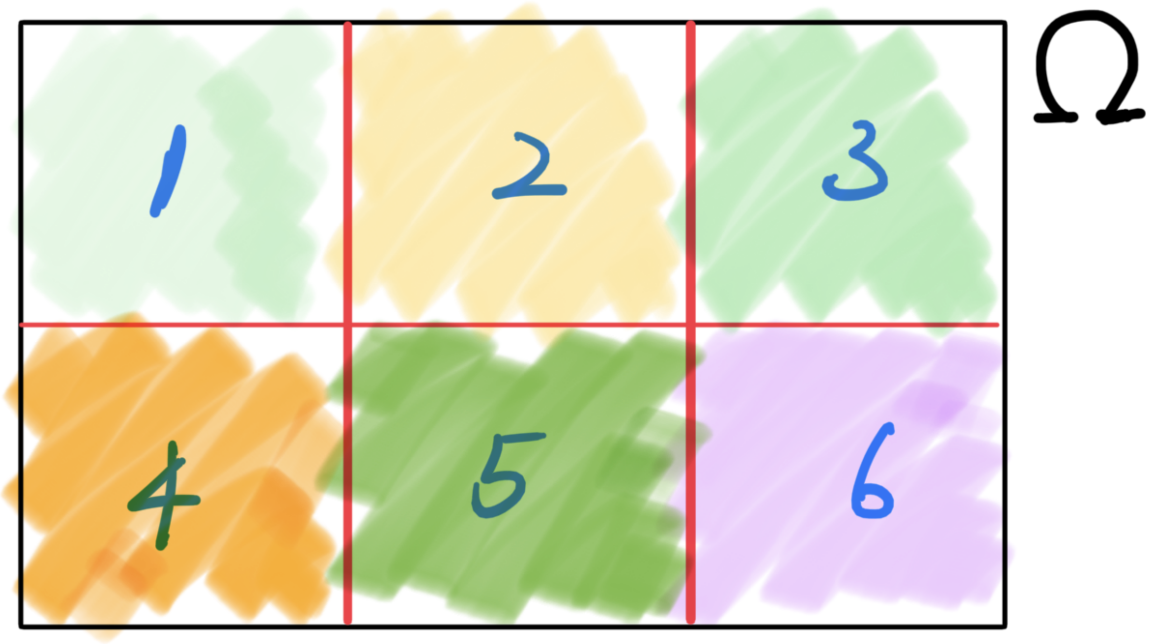
If and are not mutually exclusive,



Caculate event’s probability based on the triple

For the experiment of die throwing

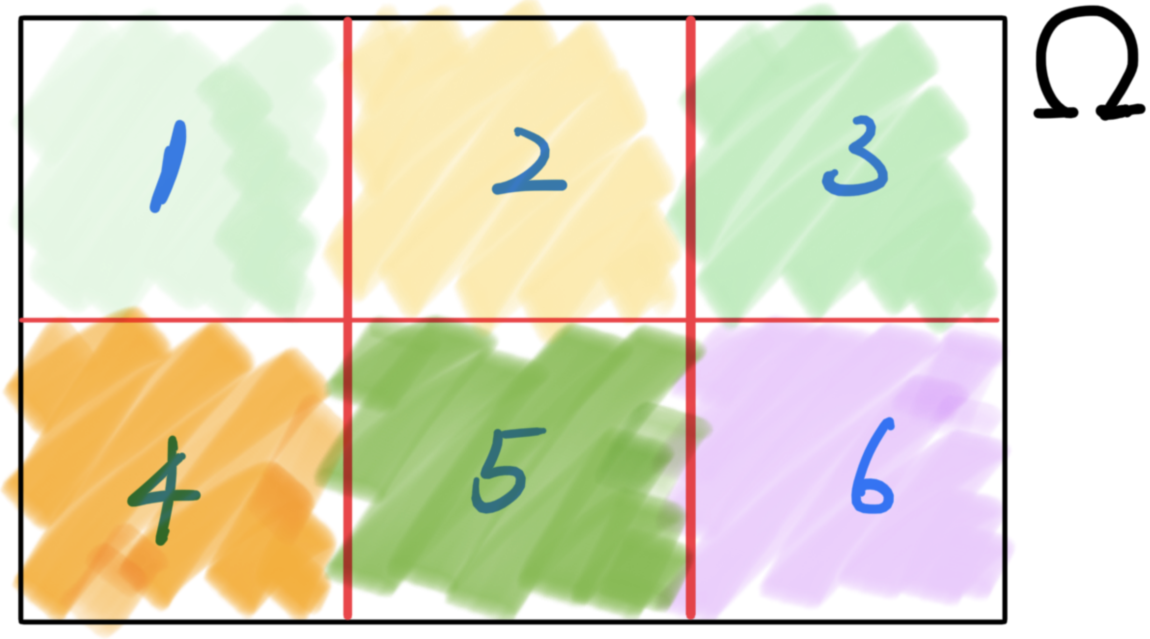
-
- Events: elements of the power set of : (elements!)
- Probability measure: if the die is fair, the singleton events' probabilities are



Caculate event’s probability based on the triple

For the experiment of die throwing

-
- Events: elements of the power set of : (elements!)
- Probability measure: if the die is fair, the singleton events' probabilities are

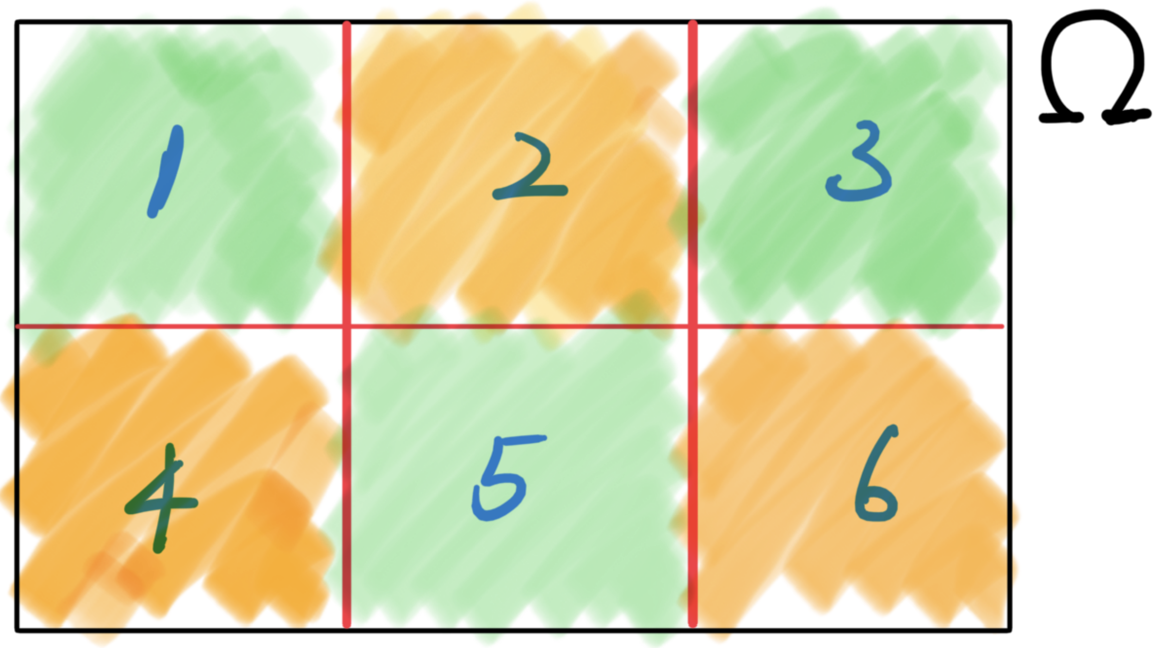


- What's the probability that an even number showing up:
    - ; and the singleton events are disjoint

$$P(\text{Odd number}) = P(\{1, 3, 5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = 1/6 + 1/6 + 1/6 = 1/2$$

$P(\text{Even number} \cup \text{Odd number}) = P(\text{Even number}) + P(\text{Odd number}) = 1/2$
$+ 1/2 = 1$

- the two events, $\{2, 4, 6\}$ (orange cells) and $\{1, 3, 5\}$ (green cells) are disjoint



Calculate event's probability based on probability triple

The axioms, or probability triple, are not the most convenient tool to use in practice

- for example, tossing a coin 10,000 times and calculate the probability or
- the sample space has elements, from to

Calculate event's probability based on probability triple

The axioms, or probability triple, are not the most convenient tool to use in practice

- for example, tossing a coin 10,000 times and calculate the probability or
- the sample space has elements, from to

We need more convenient tool: namely *random variables*

- if we are interested in the total count of 10000 tosses, assume and and

- is the *random variable* we want to work with
- more on this next ...

Random variable and probability distribution

- Formally, a random variable is a mapping from to some possible value range
    - if is discrete, is called a *discrete random variable*, e.g. result of one coin tossing 0 (tail),1 (head)
    - if is continuous, is called a *continuous random variable*, e.g. Gaussian
- A random variable is also associated with a probability distribution , which satisfies

Random variable and probability distribution

- Formally, a random variable is a mapping from to some possible value range
    - if is discrete, is called a *discrete random variable*, e.g. result of one coin tossing 0 (tail),1 (head)
    - if is continuous, is called a *continuous random variable*, e.g. Gaussian
- A random variable is also associated with a probability distribution , which satisfies

- Notation: is a shorthand notation for
- Capital letter are random variables;
- smaller letters are particular values r.v.s can take

Example 1

- For a coin tossing example, , let random variable in other words, and are mappings
    - : a *mapping* from to
    - the associated probability distrbution is and , for some
    - e.g. , a bent coin

Example 1

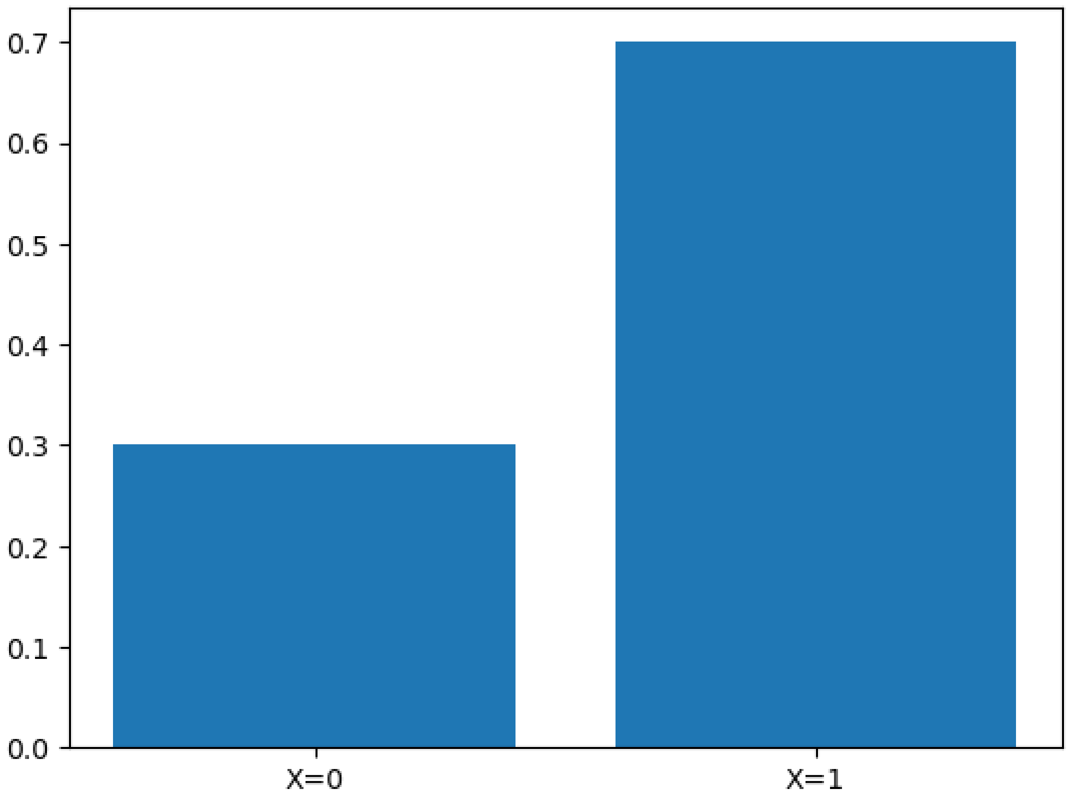
- For a coin tossing example, , let random variable

  in other words, and are mappings

  - : a *mapping* from to
  - the associated probability distrbution is and , for some
  - e.g. , a bent coin

In [8]:

```
# hide_code_in_slideshow()
names = ['X=0', 'X=1']; p =0.7;
values = [1-p, p]
plt.bar(names, values); plt.show();
```


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
X=0
X=1

Example 2

Toss a coin 3 times,

Define a r.v. as the number of times head turns up

- then , ,
- the possible value for :

Example 2

Toss a coin 3 times,

Define a r.v. as the number of times head turns up

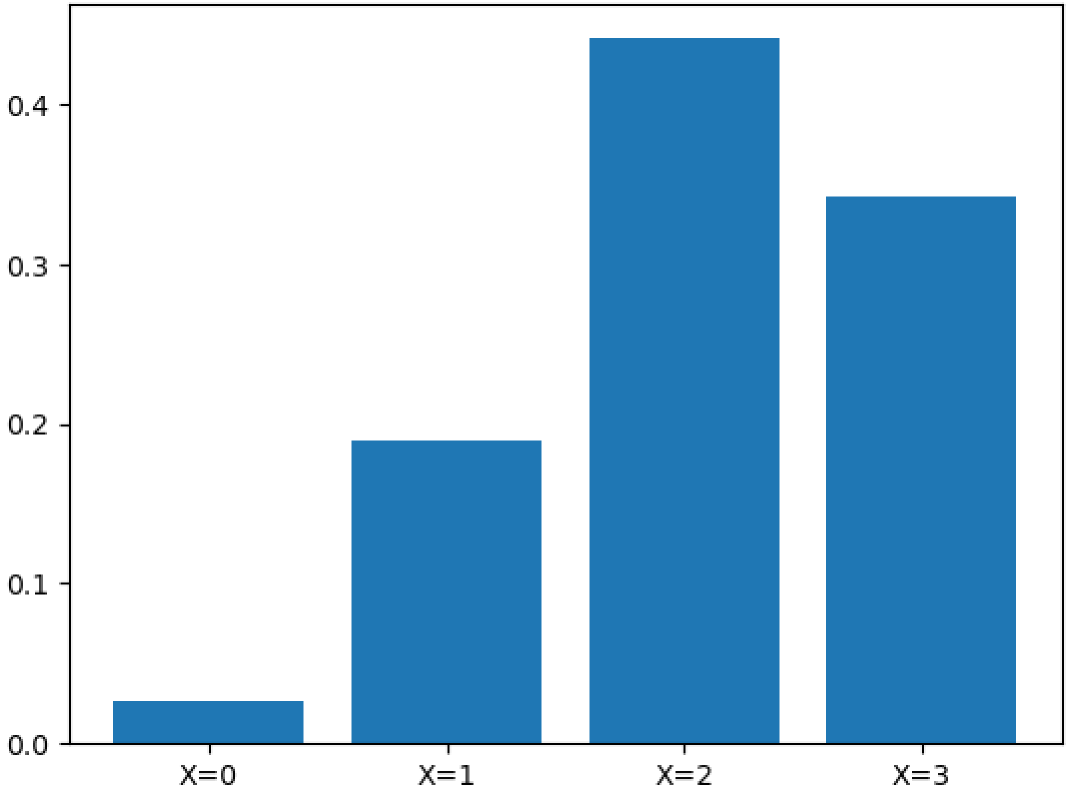
- then , ,
- the possible value for :
- the probability distribution is ; ; ;
- note that actually defines an event, e.g. is which is an event
- note that

- for a bent coin with , the distribution looks like below

- for a bent coin with , the distribution looks like below

In [19]:

```
p = 0.7; n=3; names = ["X=" + str(number) for number in range(n+1)];
values = binom.pmf(np.arange(n+1), n, p).tolist()
plt.bar(names, values); plt.show();
```



In general, the number of heads showing up for tosses of a coin with a head probabilty is a **Binomial distribution**

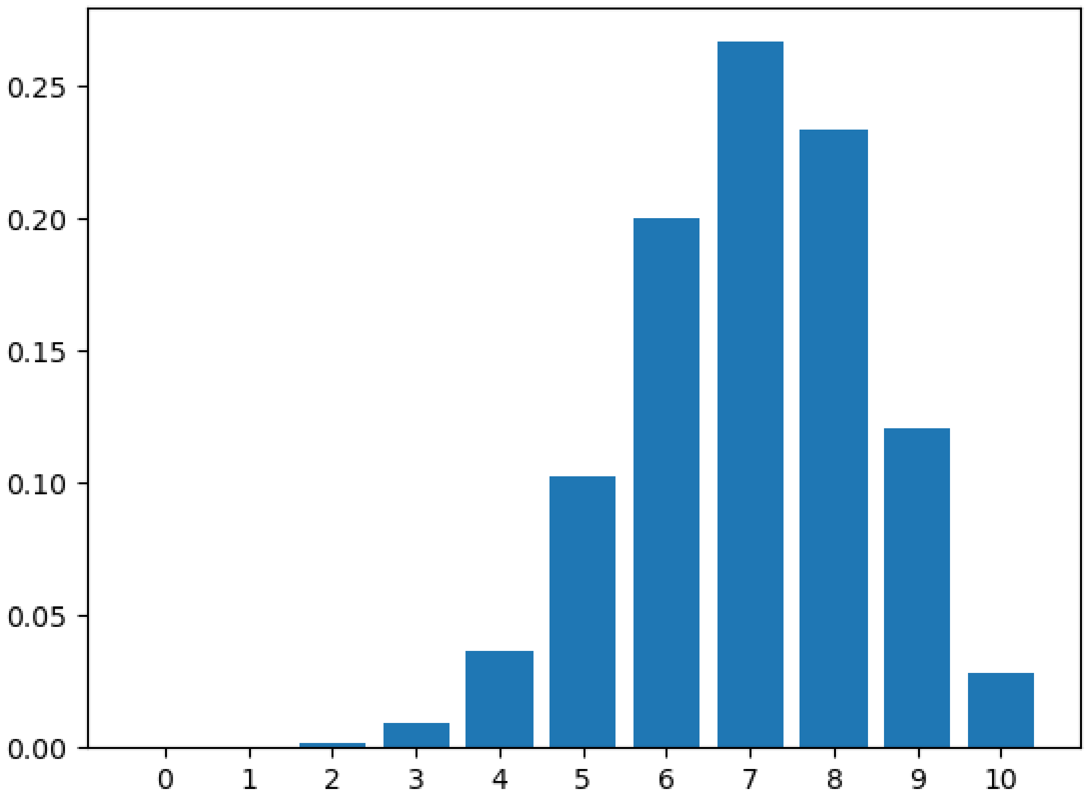
- is binomial coefficient, e.g. : i.e. out of the three tosses, how many ways to see head twice: HHT, HTH, THH
- the plot below is tossing times with each success probability
- key intuition: tells how likely you are going to see a result of
  - in this example, the most likely result, called *mode* is 7
  - you are almost impossible to observe , i.e. all 10 tosses are tail, the probability is

In general, the number of heads showing up for tosses of a coin with a head probabilty is a **Binomial distribution**

- is binomial coefficient, e.g. : i.e. out of the three tosses, how many ways to see head twice: HHT, HTH, THH
- the plot below is tossing times with each success probability
- key intuition: tells how likely you are going to see a result of
  - in this example, the most likely result, called *mode* is 7
  - you are almost impossible to observe , i.e. all 10 tosses are tail, the probability is

In [18]:

```
p = 0.7; n=10; names = [str(number) for number in range(n+1)];
values = binom.pmf(np.arange(n+1), n, p).tolist()
plt.bar(names, values); plt.show();
```



0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
9
10

For the 10000 tossing case,

- what's probability of more head than tail ?

$$P(\text{more heads than tail}) = P(X > 5000) = \sum_{x>5000} P(X = x)$$

- what's the probability of even number of head shows up ?

$$P(\text{even toss}) = P(X = \{2, 4, 6, \ldots, 10000\}) = \sum_{x \in \{2,4,\ldots,10000\}} P(X = x)$$

Example 3 (Continuous random variable, Gaussian)

Remember if is continuous, the random variable is called a continuous r.v.

- For example Gaussian random variable has a distribution
- Note that and (need to do some integration here)
- One can calculate the probability by integration

Example 3 (Continuous random variable, Gaussian)

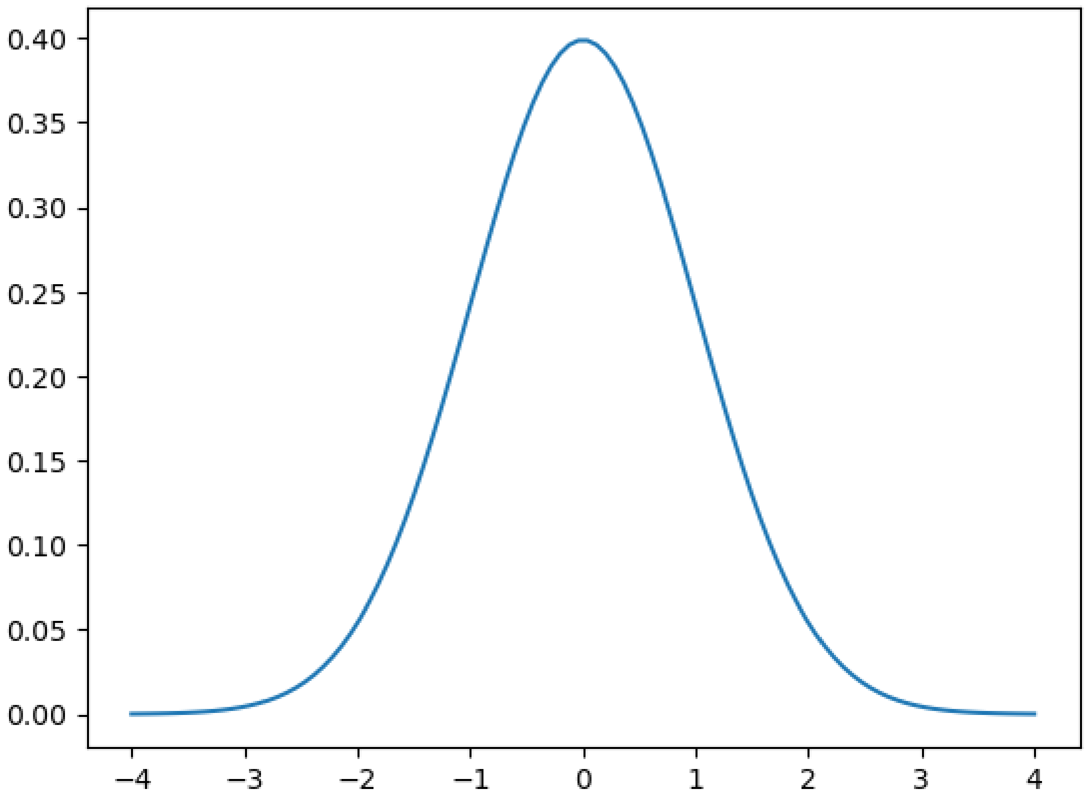
Remember if is continuous, the random variable is called a continuous r.v.

- For example Gaussian random variable has a distribution

- Note that and (need to do some integration here)
- One can calculate the probability by integration

In [17]:

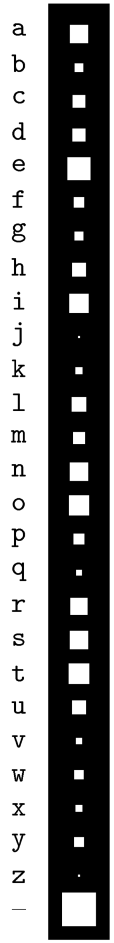
```
mu = 0; variance = 1; sigma = math.sqrt(variance)
x = np.linspace(mu - 4*sigma, mu + 4*sigma, 100)
plt.plot(x, stats.norm.pdf(x, mu, sigma)); plt.show();
```


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4

In practice, we don’t need to worry about the probability triple

- We care about the r.v. and its distribution (rather than the implied probability triple)
- Previous Gaussian r.v. is an example, we didn't specify its $(\Omega, P_\omega, \mathcal{F})$
- Here is another example of categorical random variable, $X \in \mathcal{A}_x$ where $\mathcal{A}_x = \{a, b, c \ldots, z, \_\}$
    - the English alphabet plus empty space "_"
    - the following is the probability distribution of alphabet in an English text
    - it basically tells you _, e, i, n, o are more likely to be used than e.g. letter z

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



- as expected, letters `a, e, i...` are among the most popular letters used in English text

Joint probability and random variables

- It is common to work multiple random variables at the same time
  - e.g. , are the random variables of two die tossing
  - or bigrams, is the first letter is the following letter
    - e.g. "student" has following bigrams: st, tu, de, en, nt
    - for bigram "", and

Joint probability and random variables

- It is common to work multiple random variables at the same time
    - e.g. , are the random variables of two die tossing
    - or bigrams, is the first letter is the following letter
        - e.g. "student" has following bigrams: st, tu, de, en, nt
        - for bigram "", and
- A joint event: denotd as : seperated by ",", means when X, Y are jointly true
    - some write :
    - dice example: , : the first toss is 3 and second toss is 6
- The probability distribution of the joint random variable then is
    - it gives the probability of the joint event is true
    - dice example,
- Joint distribution is a valid probability distribution: satisfies

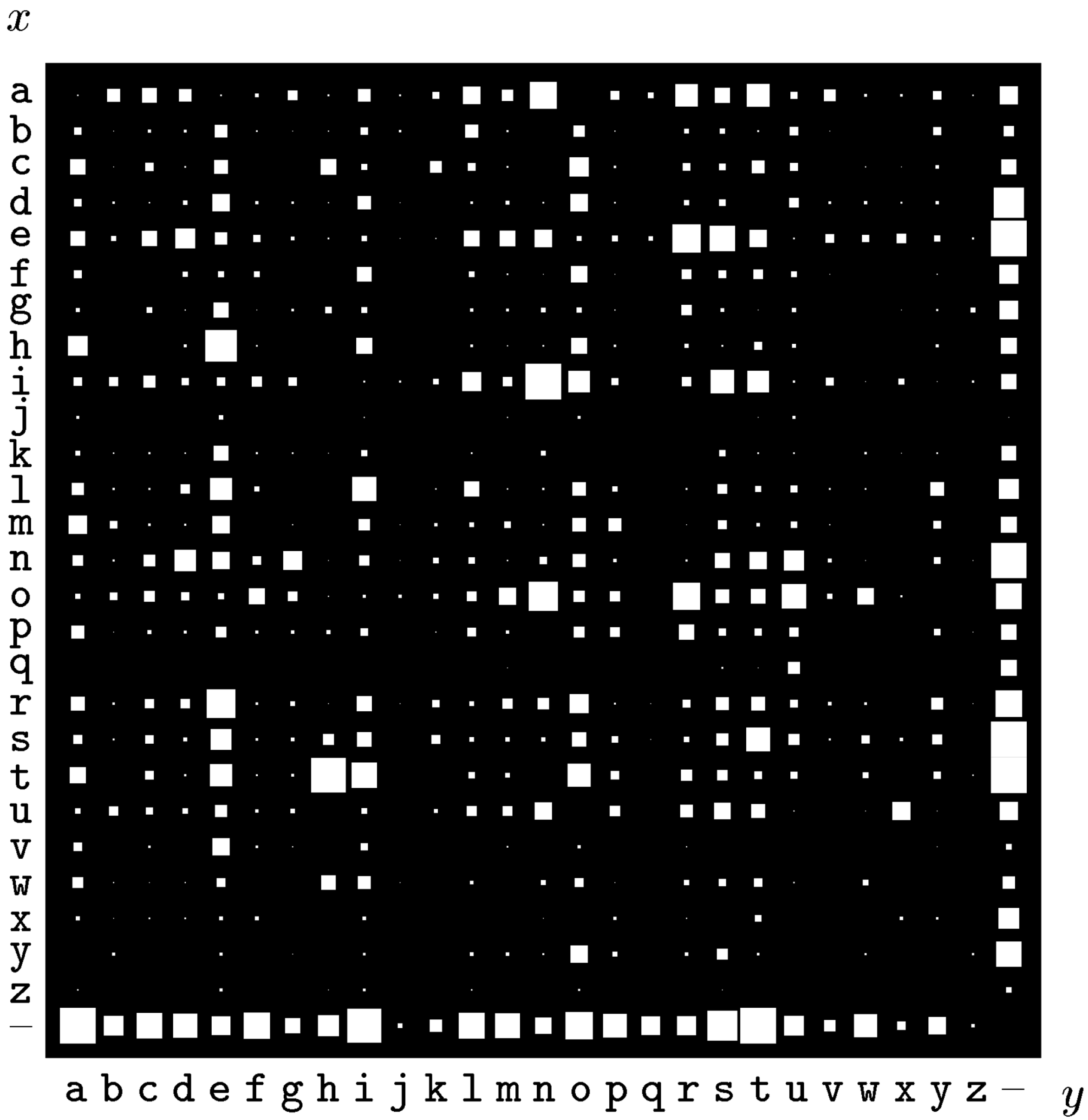
- 6×6 entries for $P(X, Y)$ of two dice tossing

X, Y	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

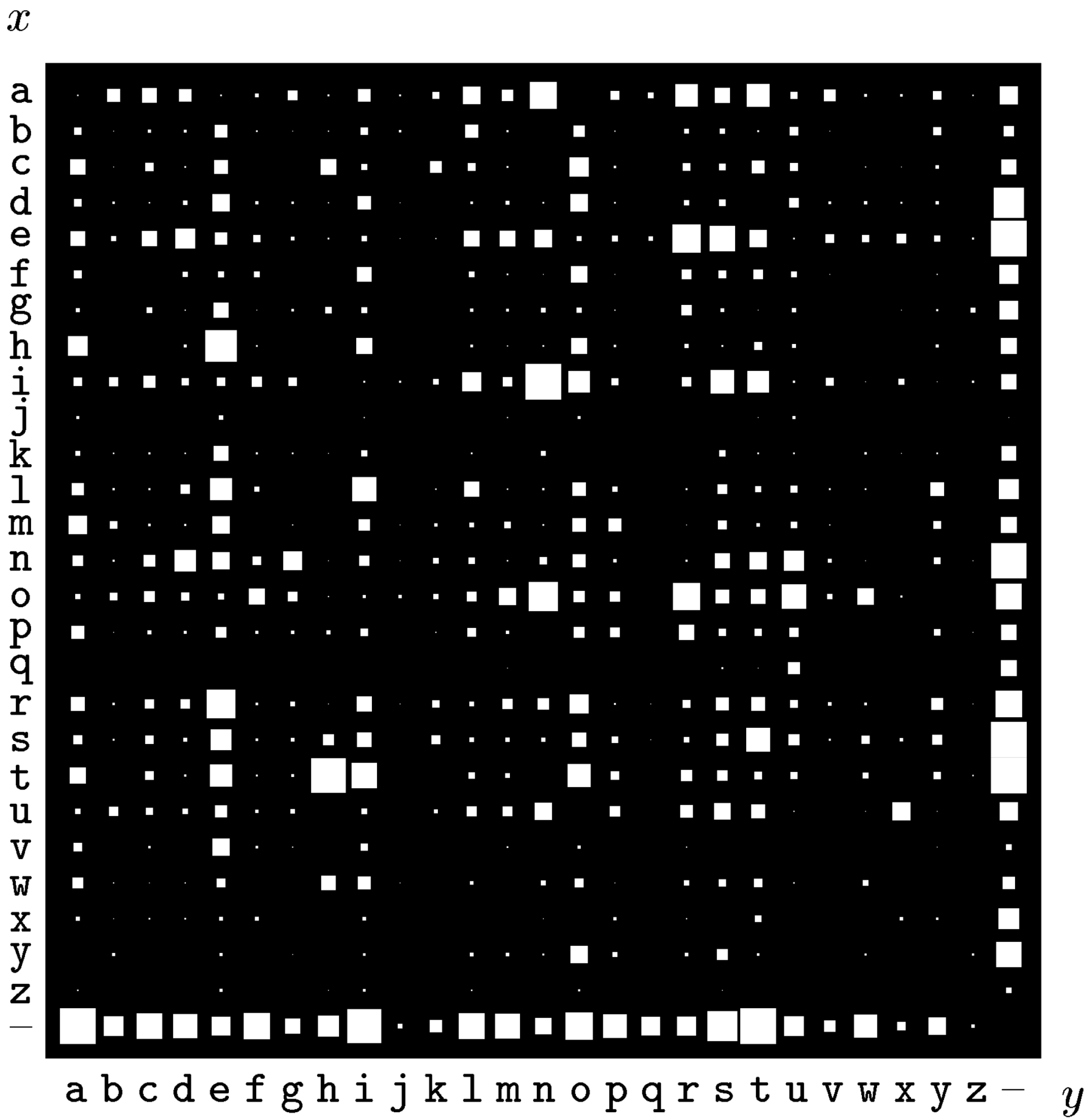
- it is a valid probability distribution
  - all positive $P(X, Y) \geq 0$
  - and $\sum_{x,y \in \{1,...,6\}} P(X = x, Y = y) = 36 \times \frac{1}{36} = 1$

- For the bigram case:
    - remember X, Y represents the first and second letter,
    - so $X = s, Y = t$, denotes a bigram "st"

- For the bigram case:
  - remember X, Y represents the first and second letter,
  - so $X = s, Y = t$, denotes a bigram "st"


x
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
–
a b c d e f g h i j k l m n o p q r s t u v w x y z –
y

- For the bigram case:
    - remember X, Y represents the first and second letter,
    - so $X = s, Y = t$, denotes a bigram "st"


x
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
–
a b c d e f g h i j k l m n o p q r s t u v w x y z –
y

Probability rule 1: marginal probability

There are only two rules

- summation rule or marginalisation

$$P(X) = \sum_y P(X, Y = y); \quad P(Y) = \sum_x P(X = x, Y),$$

  - $P(X), P(Y)$ are called marginal probability

X, Y	1	2	3	4	5	6	P(X)
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
P(Y)	1/6	1/6	1/6	1/6	1/6	1/6	

- the marginal distribution of the first throw is $P(X)$

$$P(X=1) = \sum_{y=\{1,2,...,6\}} P(X=1, Y=y) = 6 \times \frac{1}{36} = \frac{1}{6}$$

  - sum over the first row

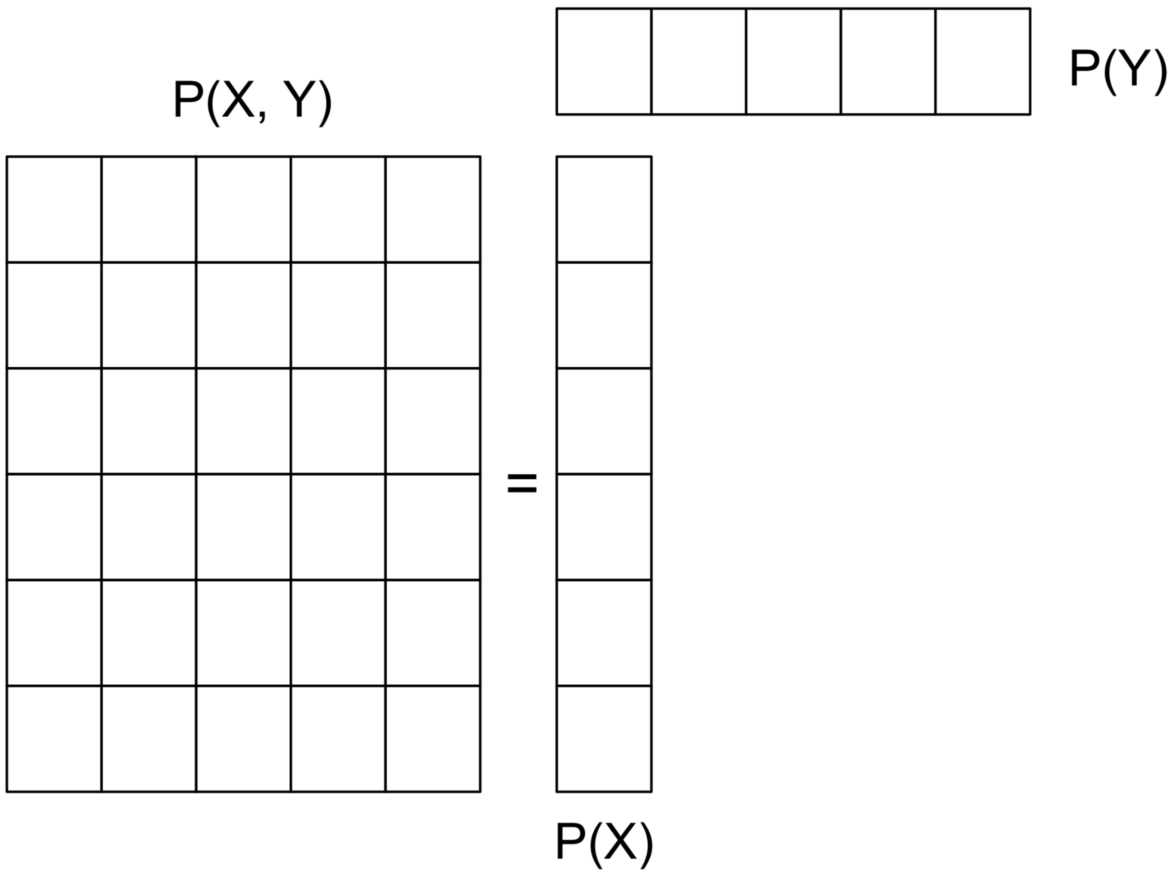
- the marginalisation is also called: summing over
  - collapsing the other dimension

X, Y	1	2	3	4	5	6	P(X)
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
P(Y)	1/6	1/6	1/6	1/6	1/6	1/6	

- the marginal distribution of the second throw is $P(Y)$

$$P(Y = 1) = \sum_{x=\{1,2,...,6\}} P(X = x, Y = 1) = 6 \times \frac{1}{36} = \frac{1}{6}$$

- sum over the first column


P(X, Y)
=
P(X)
P(Y)

Conditional probability

Before we introduce the second rule, we need a concept called conditional probability, denoted as

$$P(X|Y=y):$$

- the probability of event X occurring given that we know event Y=y to have occurred
- in short, the probability of X given Y

Conditional probability is a very useful concept for probabilistic inference

- inference: $P(\texttt{Burglar}|\texttt{Alarm})$, $P(\text{COVID}|\texttt{Test}=\textit{positive})$,
- prediction (inference over a future event): $P(\texttt{Rain\_today}|\texttt{Rain\_yesterday})$

Conditional probability of X given Y can be calculated:

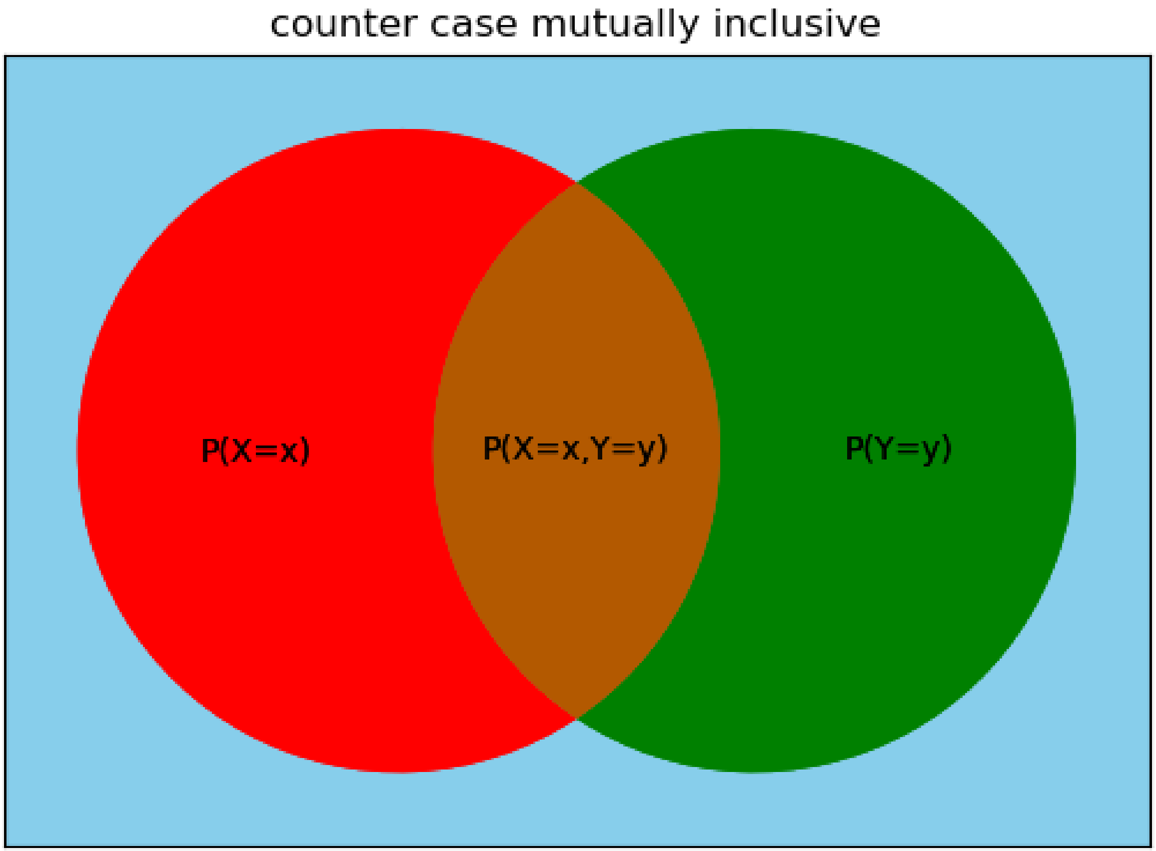
$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x, Y = y)}{\sum_x P(X = x, Y = y)}$$

- the ratio between joint and marginal probability

Conditional probability of X given Y can be calculated:

$$P(X=x|Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{P(X=x, Y=y)}{\sum_x P(X=x, Y=y)}$$

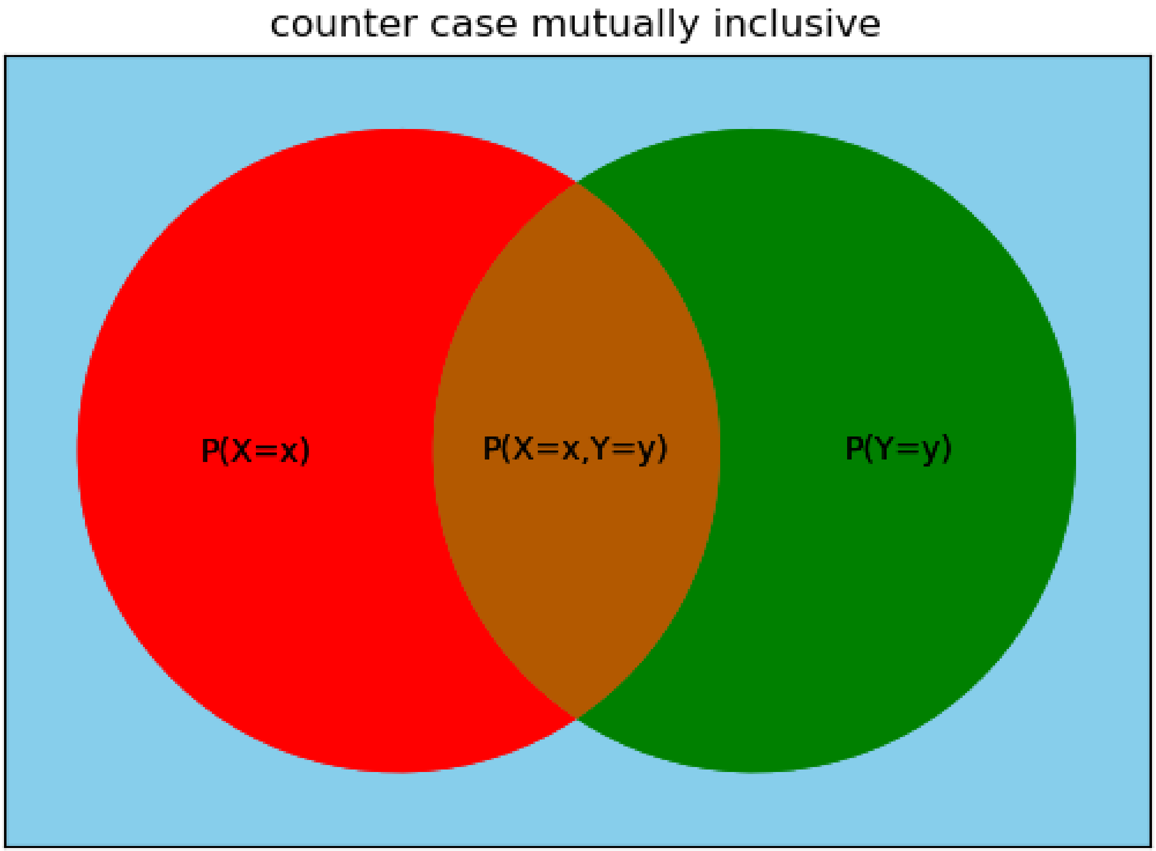
- the ratio between joint and marginal probability



Conditional probability of X given Y can be calculated:

$$P(X=x|Y=y) = \frac{P(X=x,Y=y)}{P(Y=y)} = \frac{P(X=x,Y=y)}{\sum_x P(X=x,Y=y)}$$

- the ratio between joint and marginal probability



- note that $P(X|Y=y)$ is a valid distribution of X:

$$P(X|Y=y) > 0 \ \text{ and } \ \sum_x P(X=x|Y=y) = 1$$

- given a joint distribution, the conditional is just the normalised row/column!

Conditional probability example COVID vaccine

age/death	Not Vaccinated	Double Vaccinated
80+	155	928

Define

- `Vac` $\in \{true, false\}$: random variable whether a 80+ is double vaccinated or not
- `Death` $\in \{true, false\}$: r.v. a 80+ died due to COVID

The true protection rate is actually:

$$P(\texttt{Death} = true | \texttt{Vac} = true) = \frac{P(\texttt{Death} = true, \texttt{Vac} = true)}{P(\texttt{Vac} = true)}$$

Conditional probability example COVID vaccine

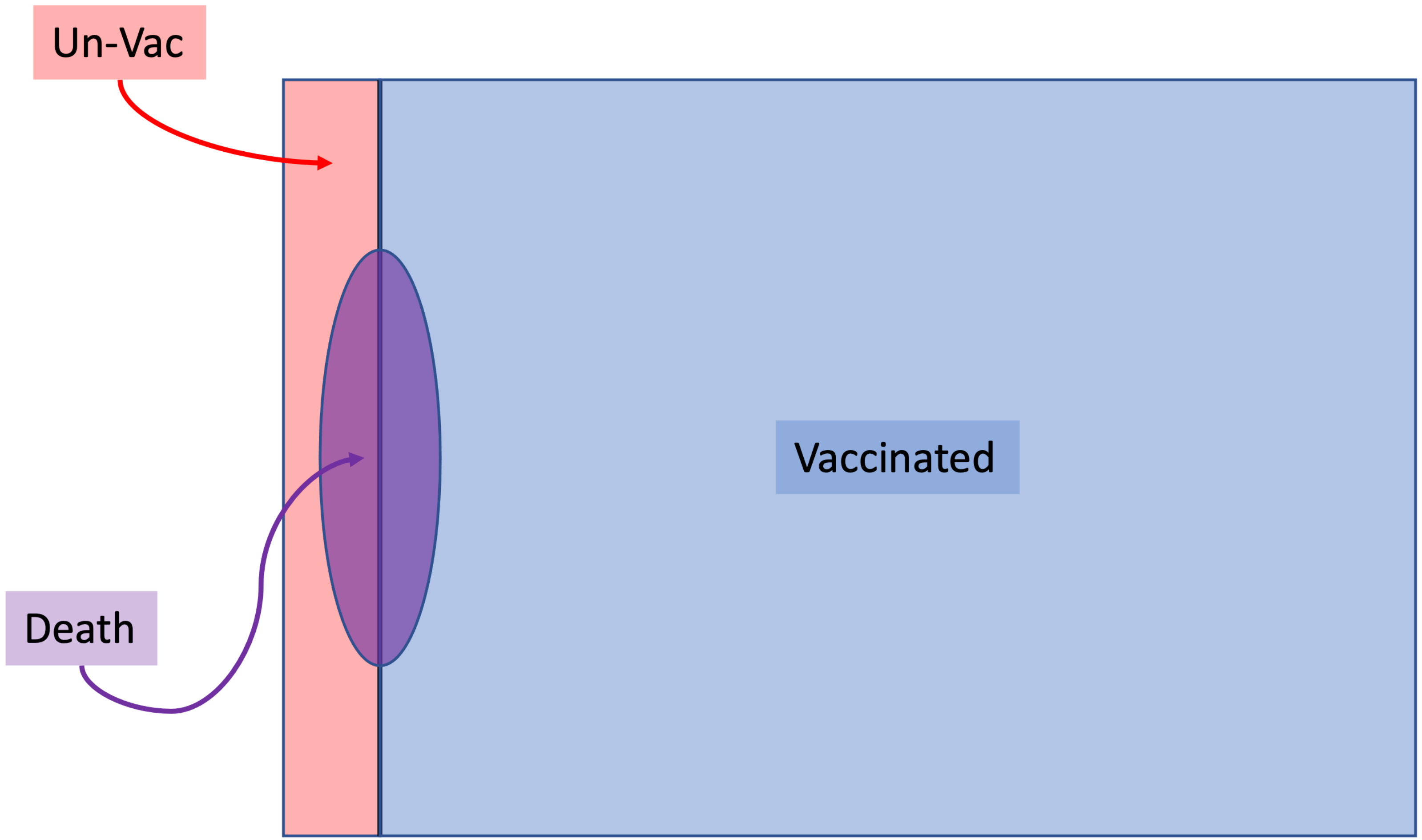
age/death	Not Vaccinated	Double Vaccinated
80+	155	928

Define

- `Vac` $\in \{true, false\}$: random variable whether a 80+ is double vaccinated or not
- `Death` $\in \{true, false\}$: r.v. a 80+ died due to COVID

The true protection rate is actually:

$$P(\texttt{Death} = true | \texttt{Vac} = true) = \frac{P(\texttt{Death} = true, \texttt{Vac} = true)}{P(\texttt{Vac} = true)}$$


Un-Vac
Vaccinated
Death

Conditional probability example COVID vaccine

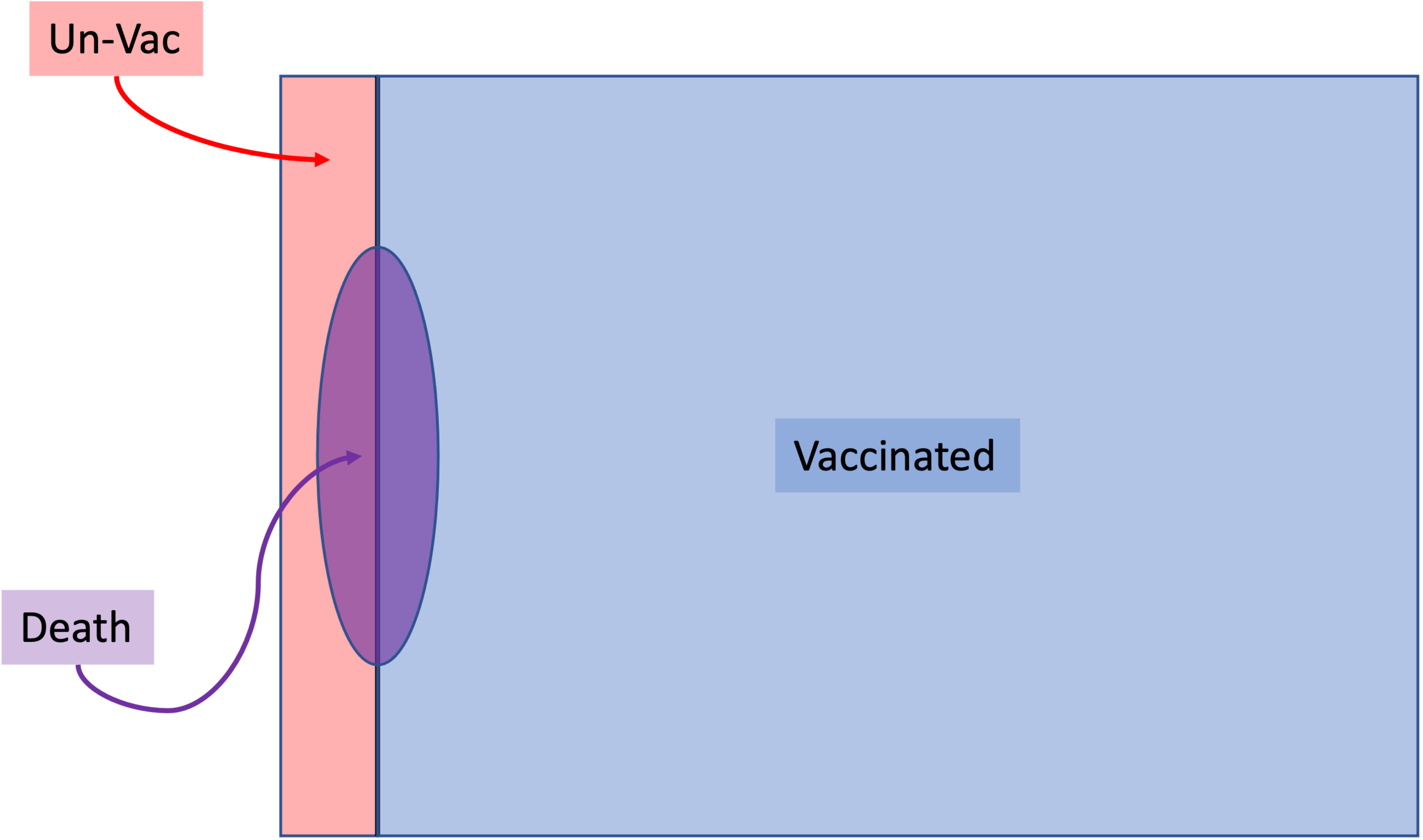
age/death	Not Vaccinated	Double Vaccinated
80+	155	928

Define

- `Vac` $\in \{true, false\}$: random variable whether a 80+ is double vaccinated or not
- `Death` $\in \{true, false\}$: r.v. a 80+ died due to COVID

The true protection rate is actually:

$$P(\texttt{Death} = true | \texttt{Vac} = true) = \frac{P(\texttt{Death} = true, \texttt{Vac} = true)}{P(\texttt{Vac} = true)}$$



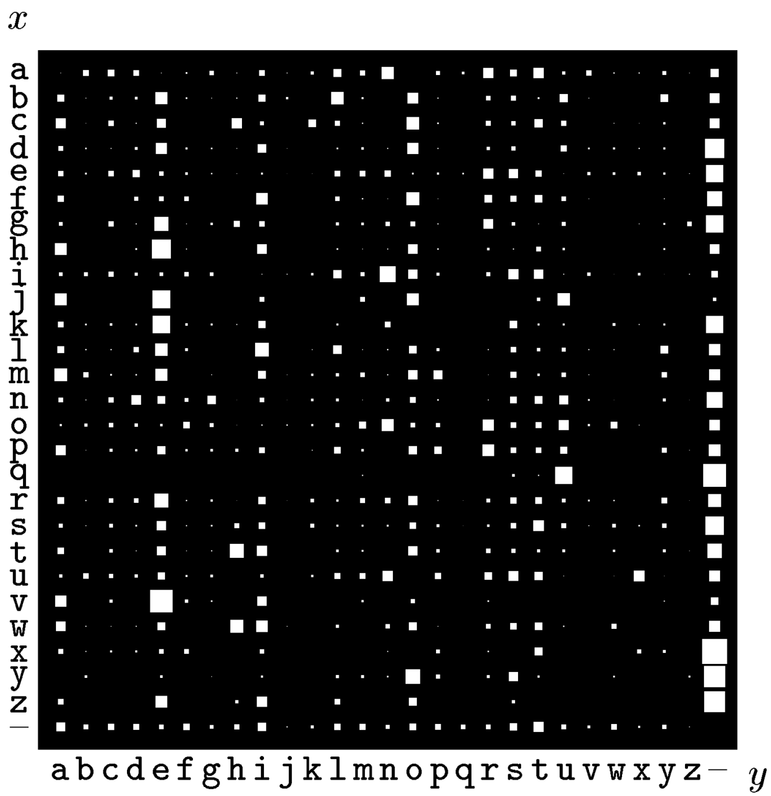
$$P(\texttt{Death} = true | \texttt{Vac} = true) = \frac{P(\texttt{Death} = true, \texttt{Vac} = true)}{P(\texttt{Vac} = true)} = \frac{\frac{928}{3 \times 10^6}}{\frac{2.94 \times 10^6}{3 \times 10^6}}$$

$$= 0.032\%$$

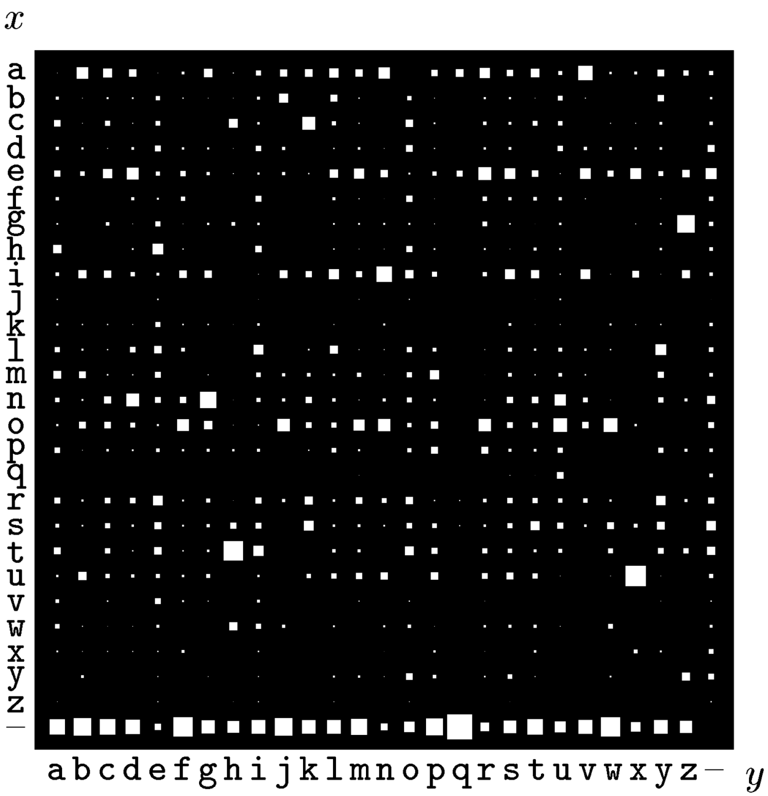
Conditional probability example

Bigram example

- left: $P(y|x)$: conditional probability of the second letter given the first letter
  - each row sum to one
  - e.g. check $P(y|x = \mathtt{q})$: the 17th row
    - it tells us it is very likely to see **`u`** following **`q`**: i.e. **`qu`**
    - also very likely to see **`q_`**: i.e. bigrams end with **`q`**, say Iraq, BBQ?



(a) $P(y \mid x)$

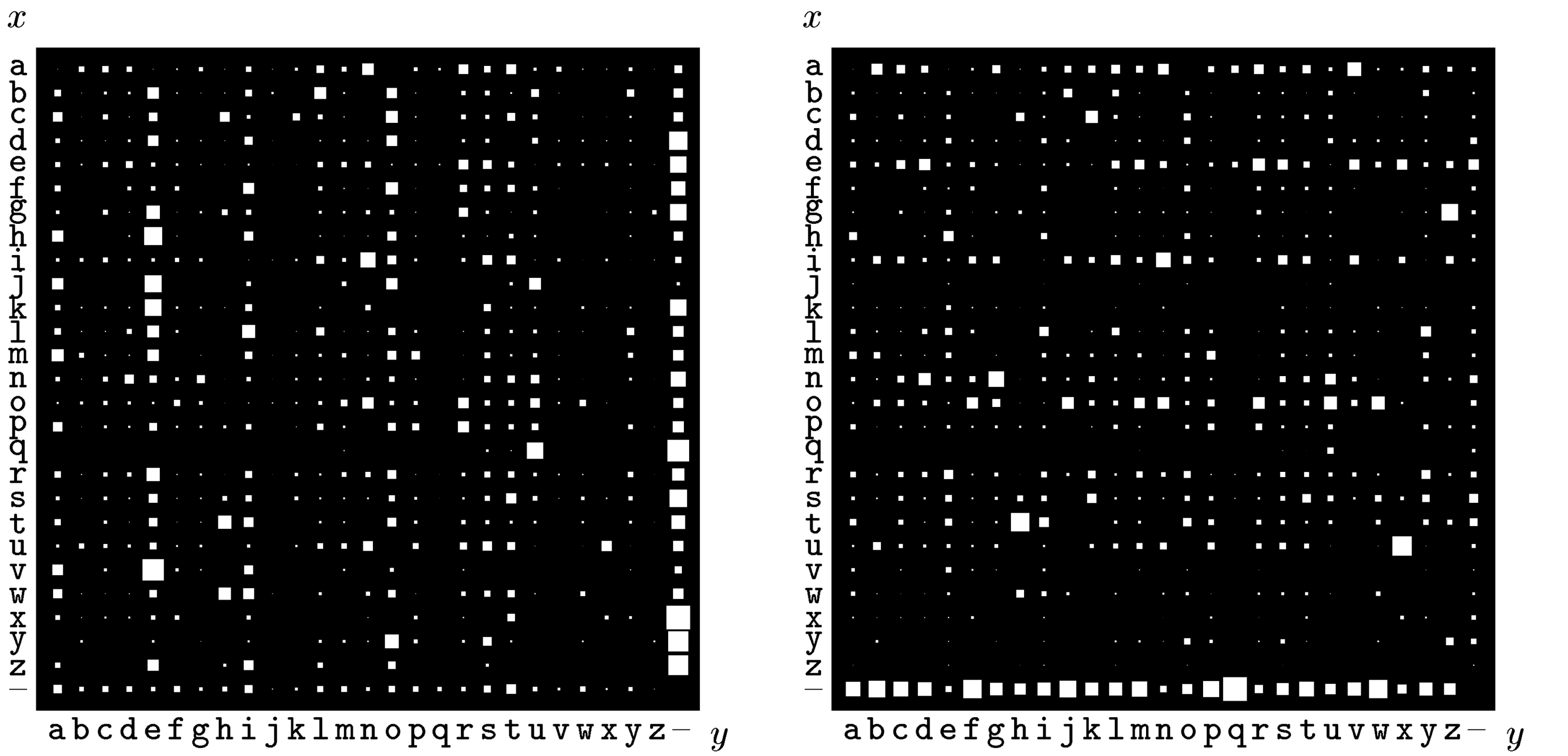


(b) $P(x \mid y)$

Conditional probability example

Bigram example

- left: $P(y|x)$: conditional probability of the second letter given the first letter
    - each row sum to one
    - e.g. check $P(y|x = \mathtt{q})$: the 17th row
        - it tells us it is very likely to see $\mathtt{u}$ following $\mathtt{q}$: i.e. $\mathtt{qu}$
        - also very likely to see $\mathtt{q\_}$: i.e. bigrams end with $\mathtt{q}$, say Iraq, BBQ?



(a) $P(y\,|\,x)$ (b) $P(x\,|\,y)$

- right $P(x|y)$: conditional probability of the first letter given second
  - each column sum to one
  - e.g. check $P(x|y=\texttt{q})$: the 17th column
    - conditional probability of the first letter given the second letter is `q`
    - the top three entries are $x = $ `a`, `e`, `_`: i.e. `aq`, `eq` and bigrams start with `q`

Probability rule 2: product rule

Product rule or chain rule

$$P(X,Y) = P(X)P(Y|X); \;\; P(X,Y) = P(Y)P(X|Y)$$

- the chain order doesn't matter
- joint distribution factorised as a product

Probability rule 2: product rule

Product rule or chain rule

$$P(X,Y) = P(X)P(Y|X); \;\; P(X,Y) = P(Y)P(X|Y)$$

- the chain order doesn't matter
- joint distribution factorised as a product

For joint distribution for more than two r.v.s

$$P(X,Y,Z) = P(X)P(Y|X)P(Z|X,Y)$$

Independence and independent random variables

If X, Y are independent, then

$$P(X,Y) = P(X)P(Y)$$

One can show that if X, Y are independent, then

$$P(X|Y) = P(X)$$

- intuition: knowing (conditional on Y) does not change the probability
- can you prove it by using the probability probability rules ?

Let's go back to Sally Clark's case

- The professor believes the chance of two children died from both SIDS is

$$\frac{1}{8541} \times \frac{1}{8541}$$

  1 in 73 million chance

- He has assumed the two events are independent, let $C_1 = \text{SIDs}$ (and C_2) be the first (second) child's cause of death is SIDS, $C_1 = murder$ means dies from murder, the expert has assumed

$$P(C_1 = \text{SIDs}, C_2 = \text{SIDs}) = P(C_1 = \text{SIDs})P(C_2 = \text{SIDs})$$

  - do you agree ?

Let’s go back to Sally Clark’s case

- The professor believes the chance of two children died from both SIDS is

$$\frac{1}{8541} \times \frac{1}{8541}$$

  1 in 73 million chance

- He has assumed the two events are independent, let $C_1 = \text{SIDs}$ (and C_2) be the first (second) child's cause of death is SIDS, $C_1 = murder$ means dies from murder, the expert has assumed

$$P(C_1 = \text{SIDs}, C_2 = \text{SIDs}) = P(C_1 = \text{SIDs})P(C_2 = \text{SIDs})$$

    - do you agree ?

Not really!

- genetic runs in family
- environment factors also have a play (they breathe the same air, eat the same food etc.)

If the first child dies from one disease, the second's risk is increased!

- same as parents have diabetes, you are more likely to have it

Verify independence from joint distribution

Independence :

$$P(X, Y) = P(X)P(Y)$$

X,Y	1	2	3	4	5	6	P(X)
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
P(Y)	1/6	1/6	1/6	1/6	1/6	1/6	

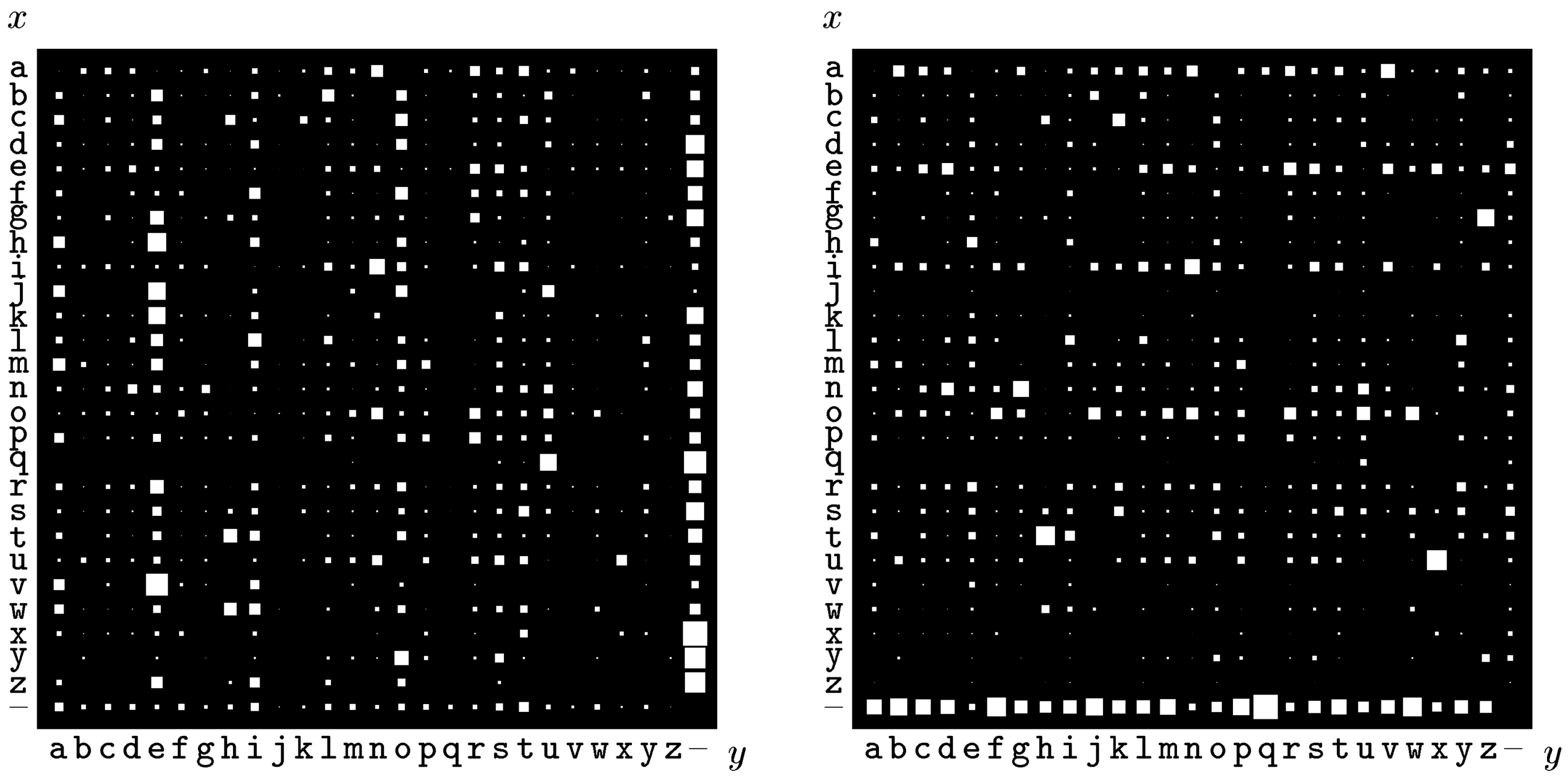
- marginalisation: find $P(X)$ and $P(Y)$
- check $P(X = x, Y = y) = P(X = x)P(Y = y)$

Verify indepedence from conditional distribution

We can also check indepedence from the conditional distribution

Note that if X, Y are independent, then $P(X|Y) = P(X)$

Question, is the bigram model independent ?



(a) $P(y \mid x)$

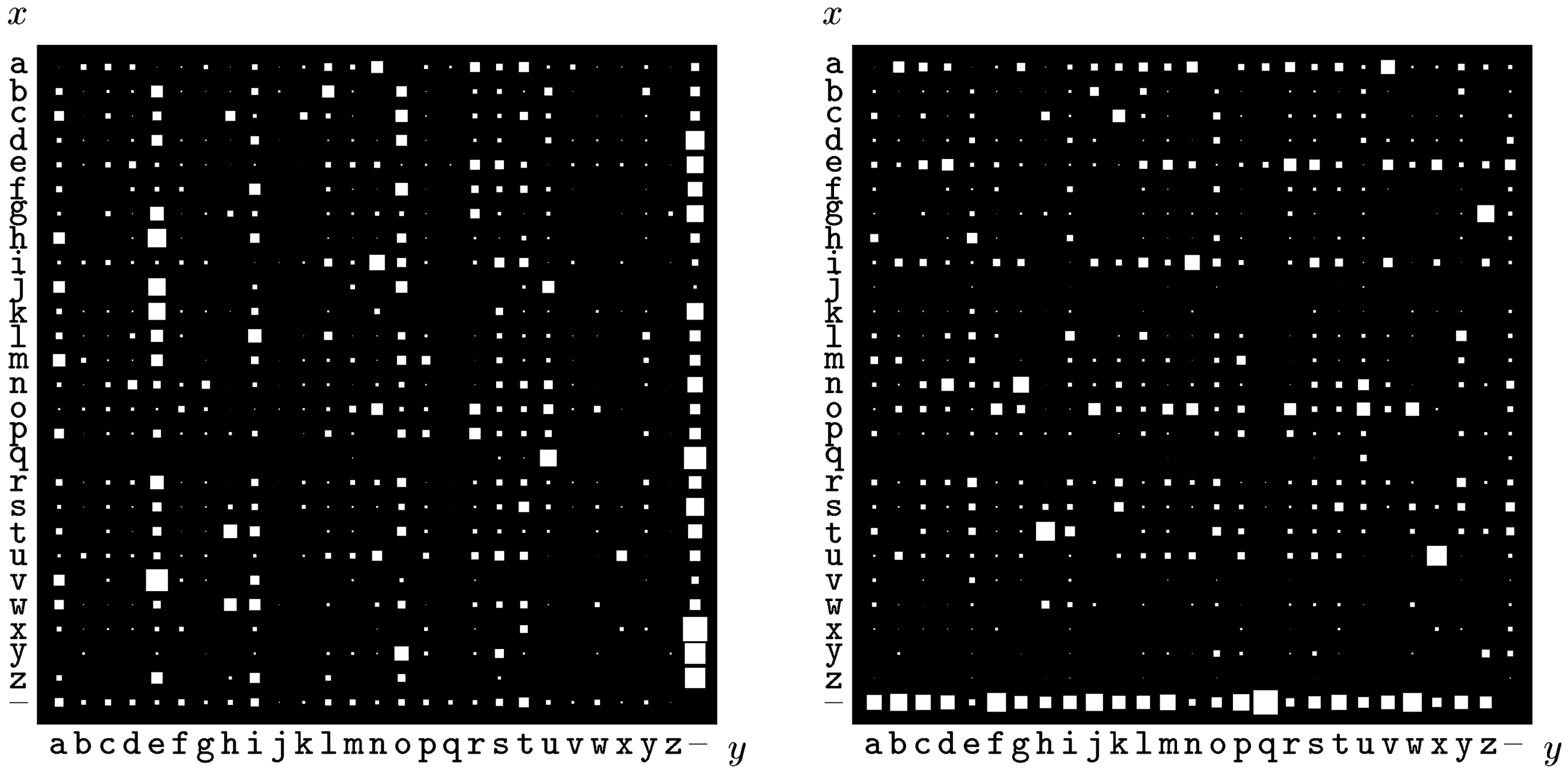
(b) $P(x \mid y)$

Verify indepedence from conditional distribution

We can also check indepedence from the conditional distribution

Note that if X, Y are independent, then $P(X|Y) = P(X)$

Question, is the bigram model independent ?



(a) $P(y \mid x)$ (b) $P(x \mid y)$

- No! the column of $P(x|y)$ are all drastically different!
  - knowing the second letter changes the distribution of the first $\Rightarrow$ NOT Independent
  - if independent, we'd expect all columns are the same $P(x|y) = P(x)$
  - the rows of $P(x|y)$ are also drastically different

Summary

- Probability theory
- Random variables and their distributions
- Two rules
    - sum rule
    - product rule

Next time

- Baye's rule
  - just conditional probability
  - tackle the two problems: Sally Clark and COVID case
- Probabilistic inference
  - as an inverse problem
- Some simple uncertainty based machine intelligence examples
  - Naive Baye's classifier: classify spam emails
  - Concept learning machine: can machine mimic your prediction ?