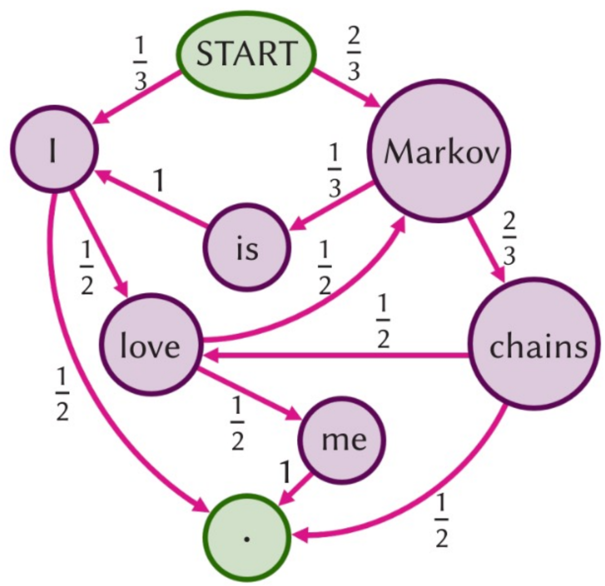
CS 4100: Introduction to AI

Wayne Snyder
Northeastern University

Lecture 12: Markov Chains and Linear Algebra Review



Andrey Markov and Markov Chains

Markov was part of the great tradition of mathematics in Russia.

Markov started out working in number theory but then got interested in probability. He enjoyed poetry and the great Russian poet Pushkin.

Markov studied the sequence of letters found in the text of *Eugene Onegin*, in particular the sequence of consonants and vowels. He sought a way to describe the patterns in the text. This eventually led to the idea of a system in which one transitions between states, and the probability of going to another state depends only on the current state.

Hence, Markov pioneered the study of systems in which the future state of the system depends only on the present state in a random fashion. Classic examples in modern life include the movement of stock prices and the dynamics of animal populations.

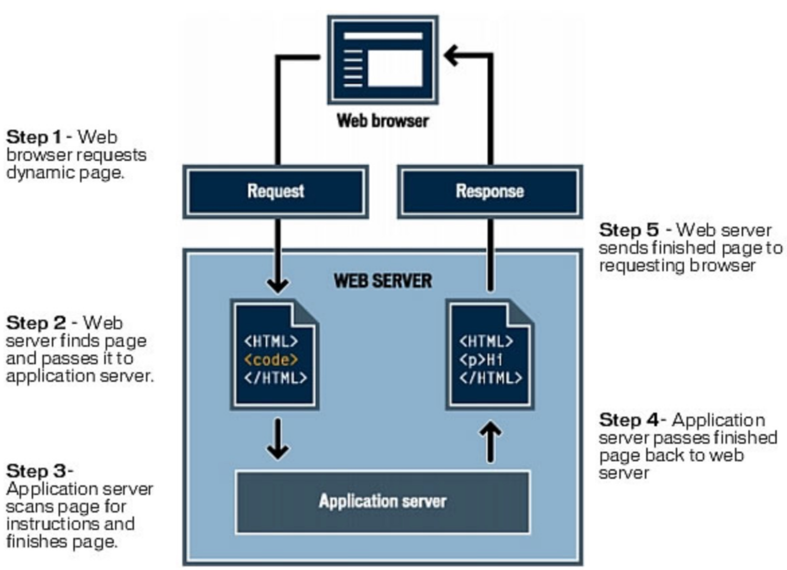
These have since been termed Markov Chains.

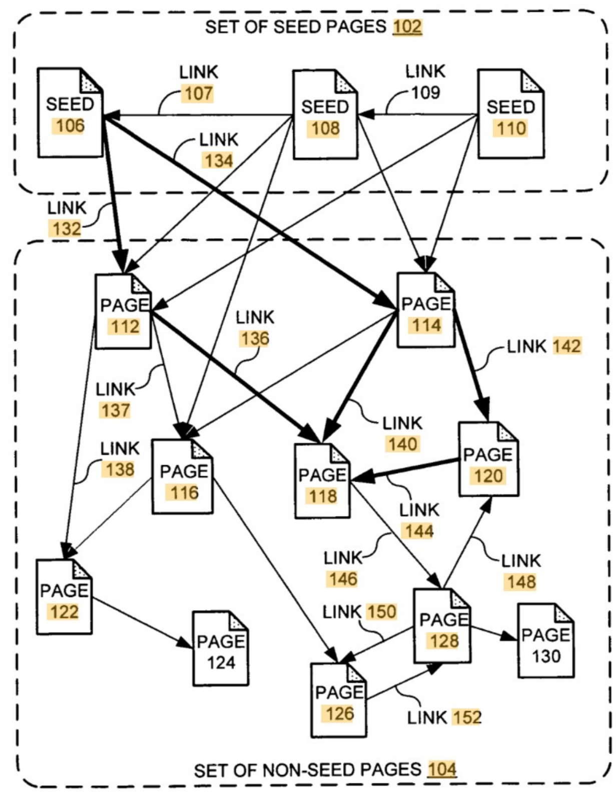
Markov chains are essential tools in understanding, explaining, and predicting phenomena in computer science, physics, biology, economics, and finance.

Markov Chains

Many applications in computing are concerned with how a system behaves over time.

Think of a Web server that is processing requests for Web pages, or the Page Rank algorithm Google uses to serve such requests:





Markov Chains

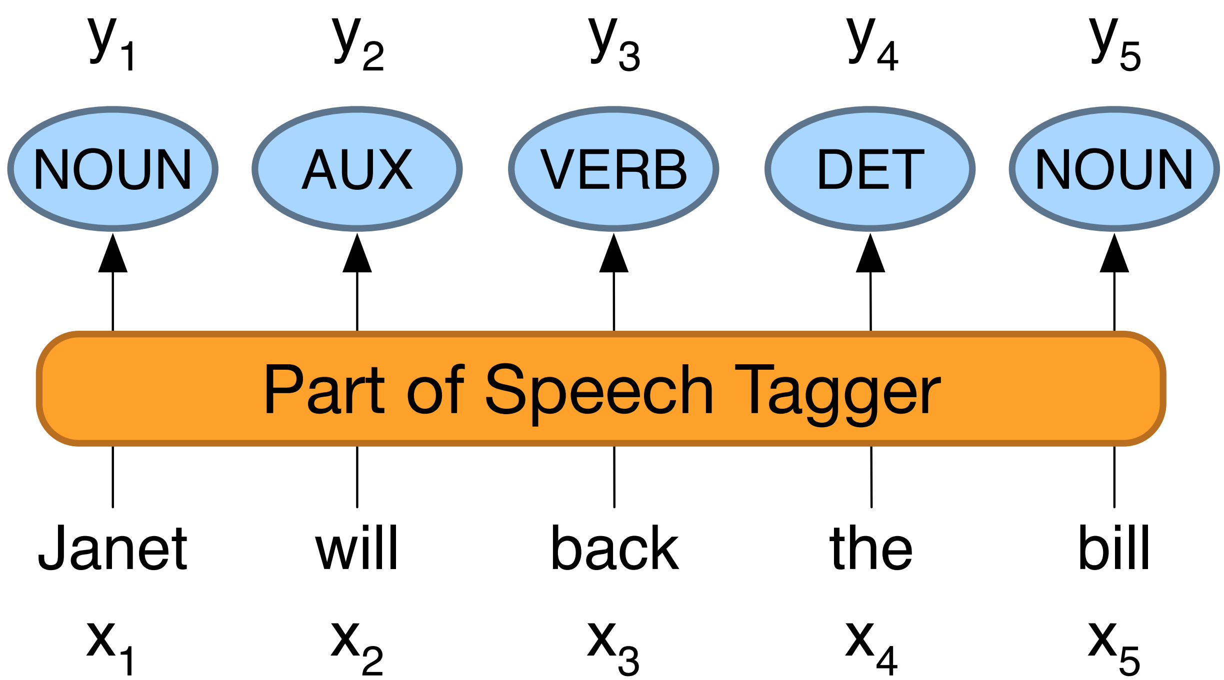
In Artificial Intelligence, we might want to understand how swarms of drones can learn to work together towards a common goal:

Britain's Royal Air Force chief says drone swarms ready to crack enemy defenses

By Sebastian Sprenger Jul 14, 2022

Markov Chains

Or – our goals are a bit more modest in this course – how to label text with annotations indicating parts of speech:



All of these various problems can be addressed successfully with Markov Chains.

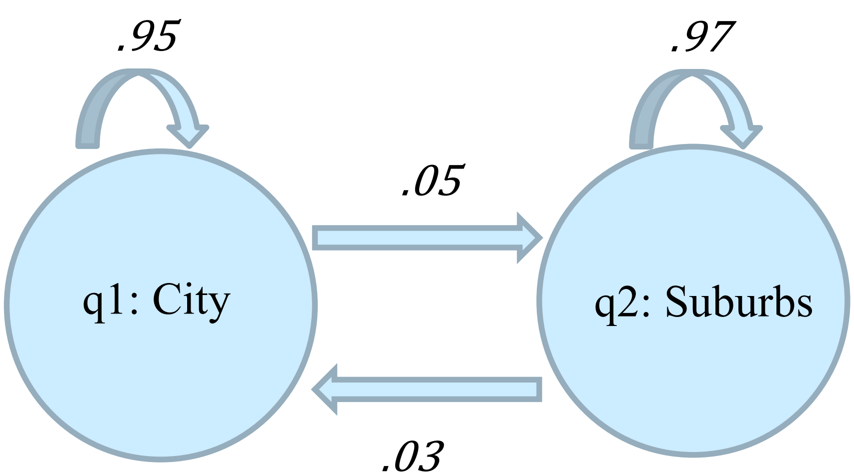
Markov Chains

Markov Chains are directed graphs defined by (Q, A, π):

$Q = q_1 q_2 \ldots q_N$	a set of N states
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Alternately:

π may be an initial population distribution of discrete individuals, and $\pi_1 \ldots \pi_n$ is a partition of this population among the states.



A

	From City	From Suburbs
To City	.95	.03
To Suburbs	.05	.97

$\pi = [\,.6, \quad .4]$

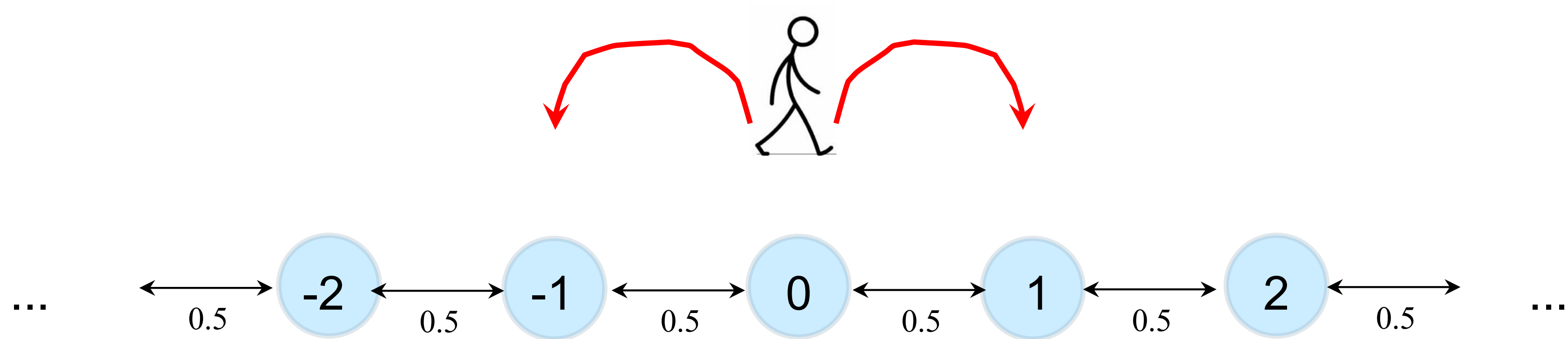
OR:

$\pi = [\,600000, \quad 400000]$

Digression: Random Walks

A good example of a Markov Chain is a random walk in n dimensions, in which the idividual is a point moving in N-space.

One-Dimensional Random Walk: Start at 0; at each step, flip a coin and go right if heads and left if tails:



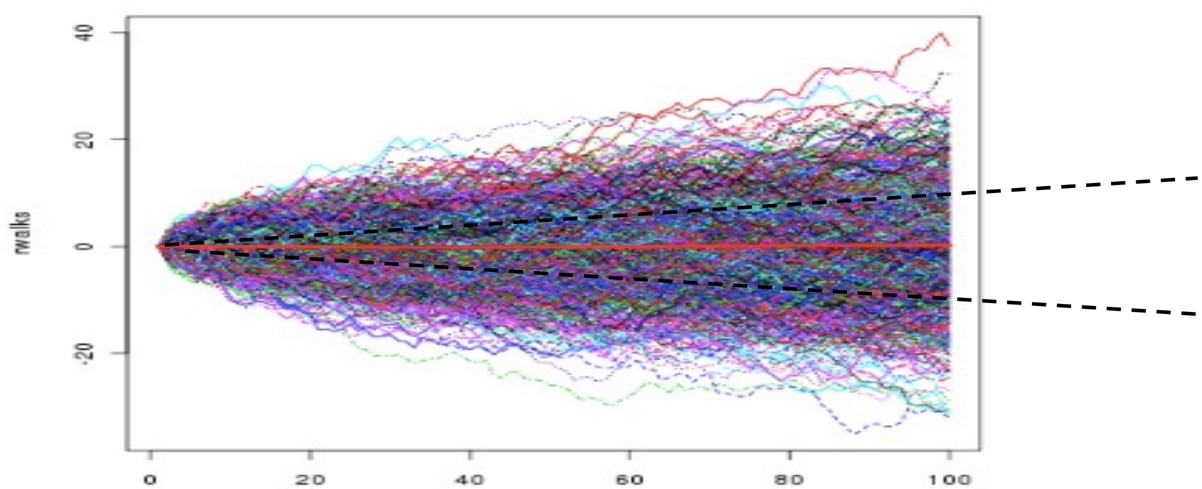
Markov Chains and Random Walks

Results about 1D Random Walks:

- As n -> infty, the probability that the marker is in a particular location approaches 0; but the probability that you return to this position later approaches 1.0; you will return to this position infinitely many times!

- Alternate to last: For any boundary (say, the position k), you will cross this boundary with probability 1.0.

- After n steps, your expected distance from the start is $\sqrt{\frac{2n}{\pi}}$ locations, 0.5 probability in negative region, 0.5 in positive.



For coin flips, after 100 flips, expected distance from original is sqrt(200/pi) = 7.98

Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

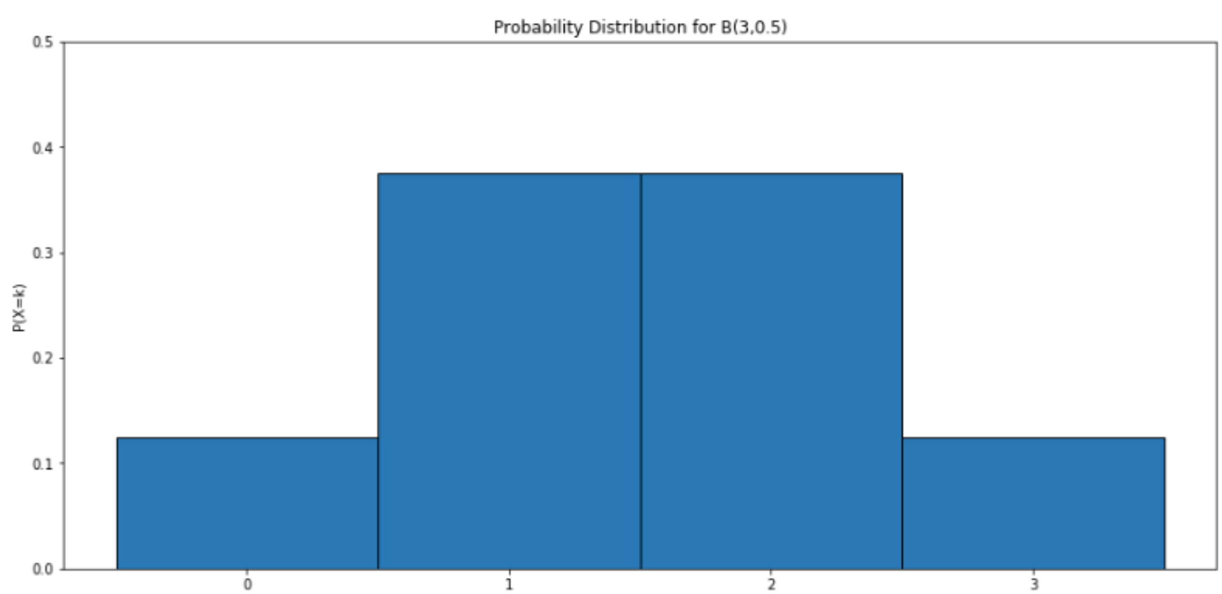
Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

```
[0.125, 0.375, 0.375, 0.125]
```

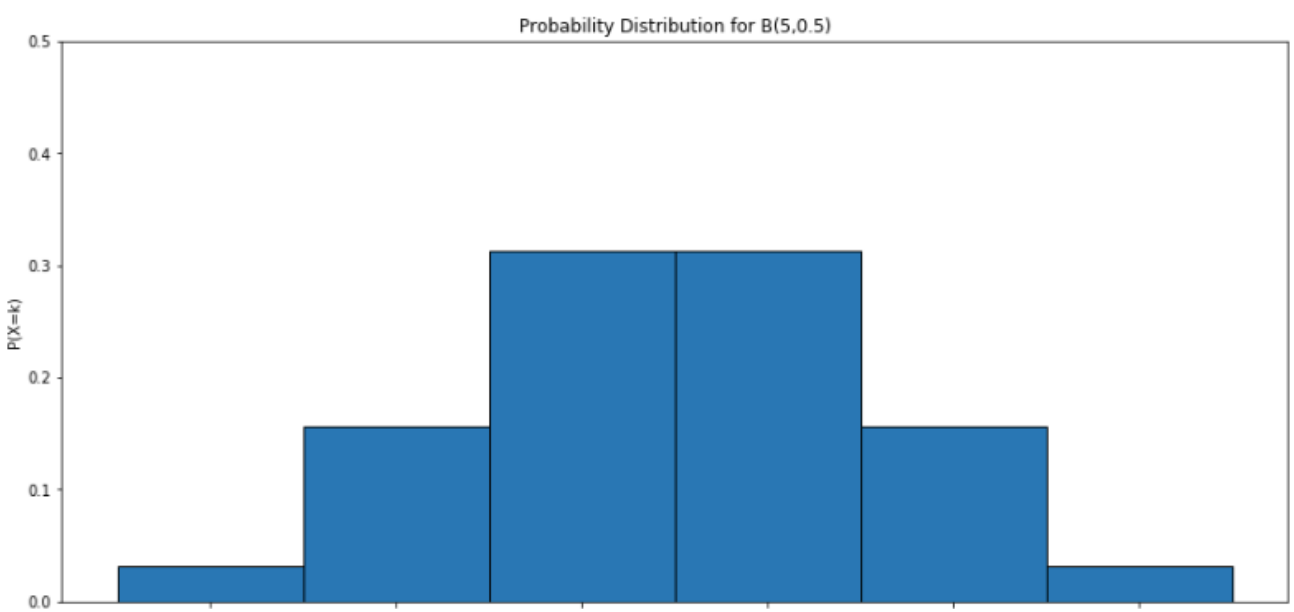


Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

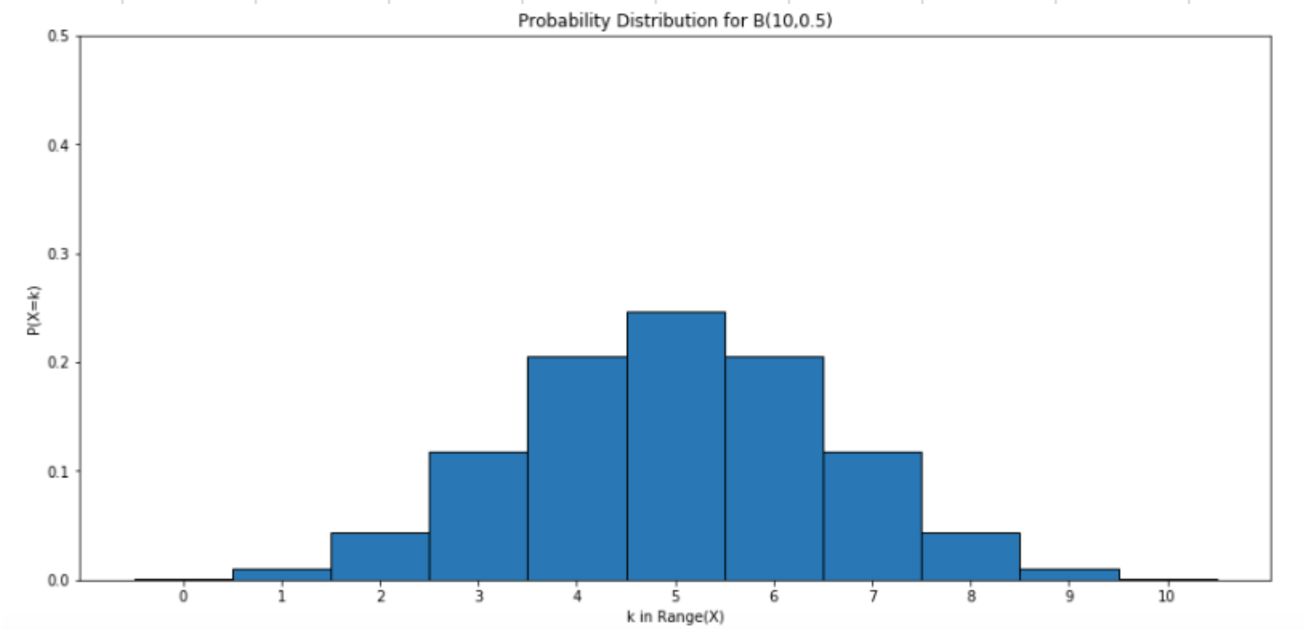


Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

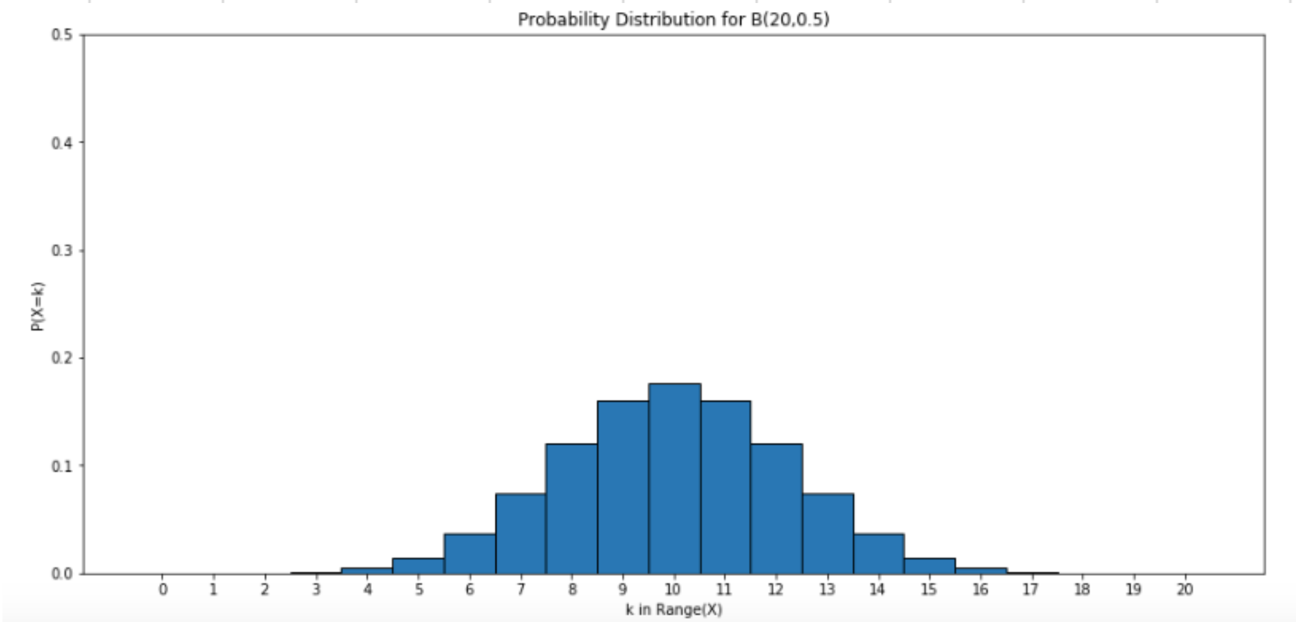


Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														



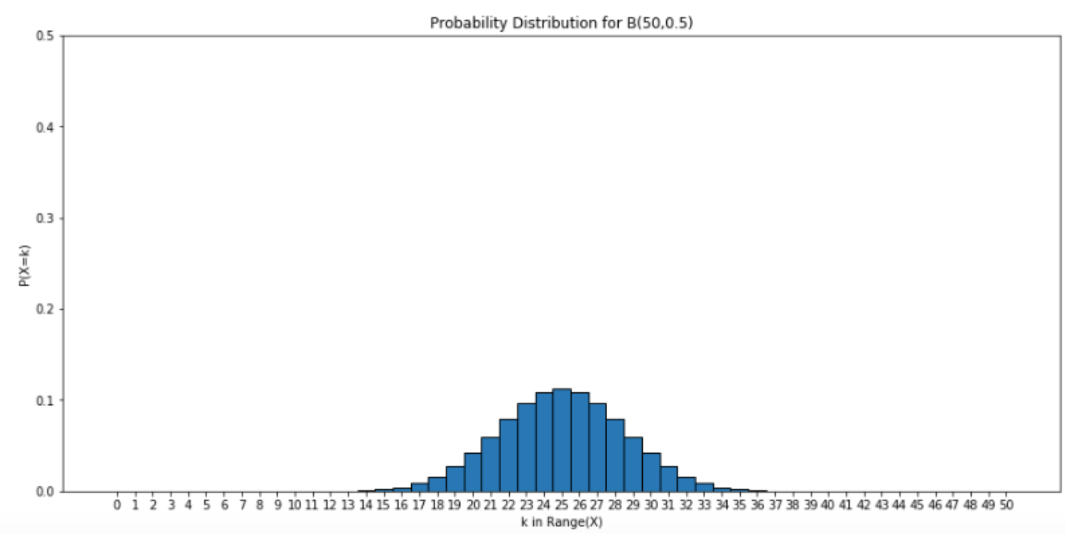
Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

```
[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0001, 0.0003, 0.0008, 0.002, 0.0044, 0.0087, 0.016, 0.
027, 0.0419, 0.0598, 0.0788, 0.096, 0.108, 0.1123, 0.108, 0.096, 0.0788, 0.0598, 0.0419, 0.027, 0.016, 0.0087, 0.004
4, 0.002, 0.0008, 0.0003, 0.0001, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
```

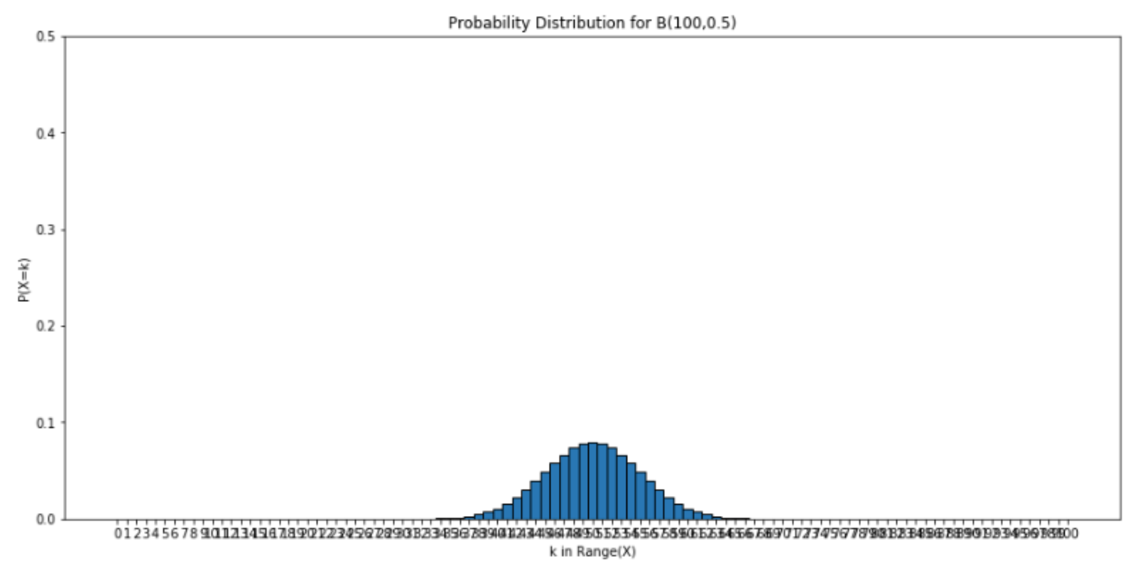


Markov Chains and Random Walks

Results about 1D Random Walks:

- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

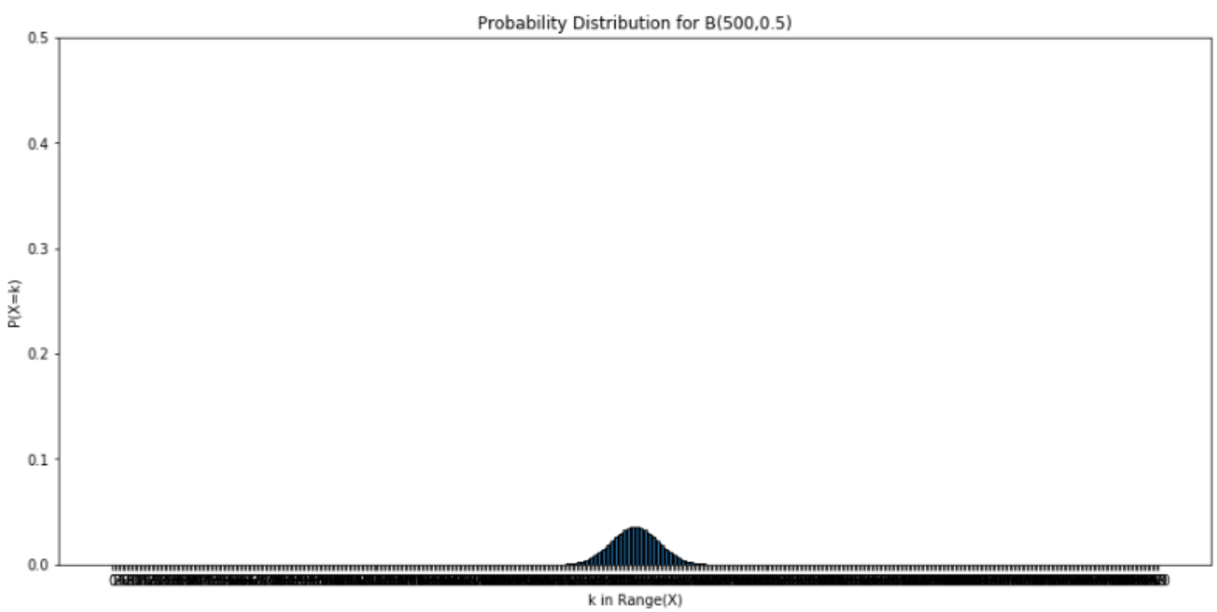


Markov Chains and Random Walks

Results about 1D Random Walks:

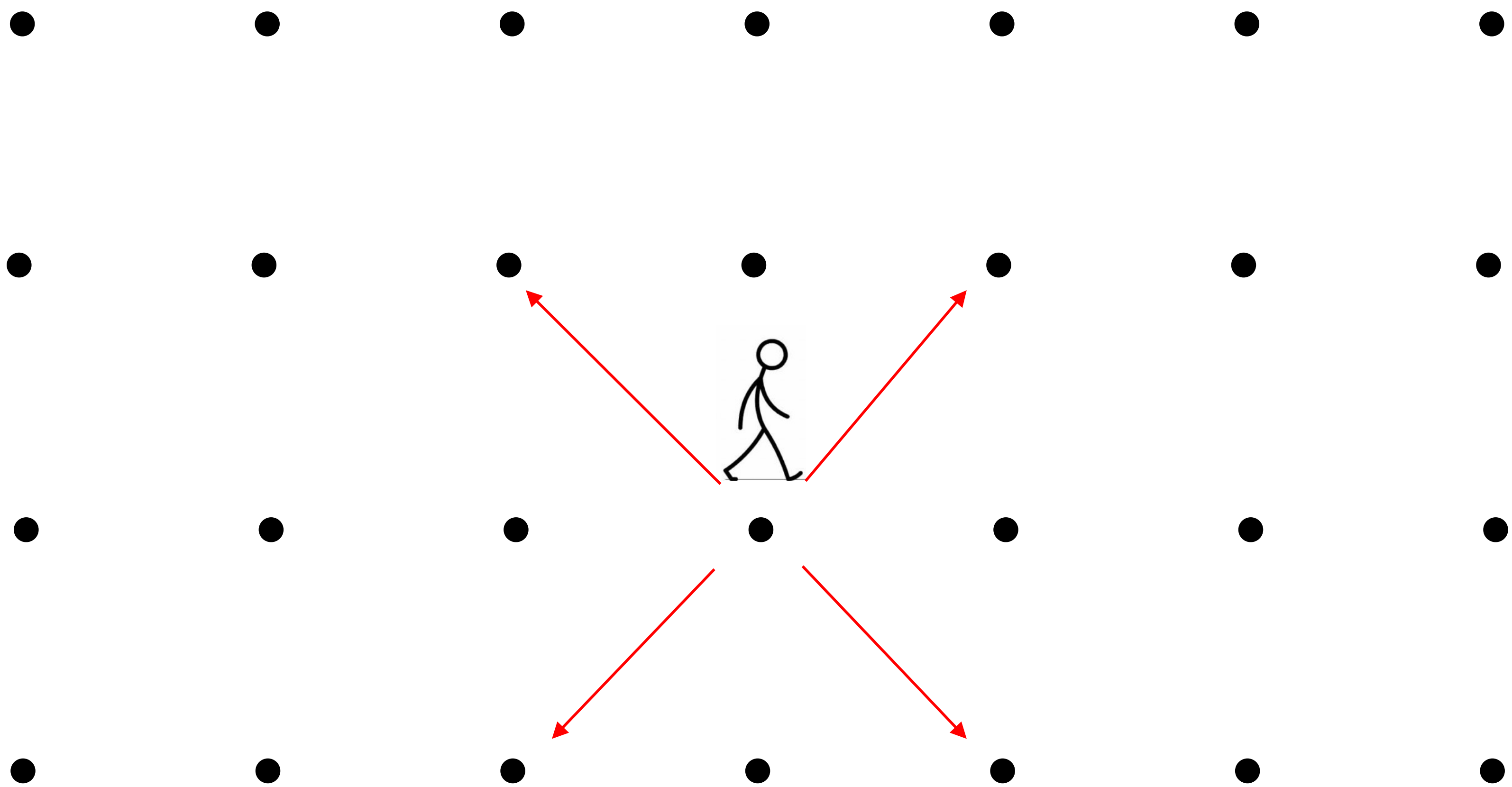
- As we saw in that problem, the probability that you are in local positions after a small number n of rounds, can be calculated using the binomial:

Time Step	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	-7
1							1.000							
2						0.500		0.500						
3					0.250		0.500		0.250					
4				0.125		0.375		0.375		0.125				
5			0.063		0.250		0.375		0.250		0.063			
6		0.031		0.156		0.313		0.313		0.156		0.031		
7	0.016		0.094		0.234		0.313		0.234		0.094		0.016	
8														
9														

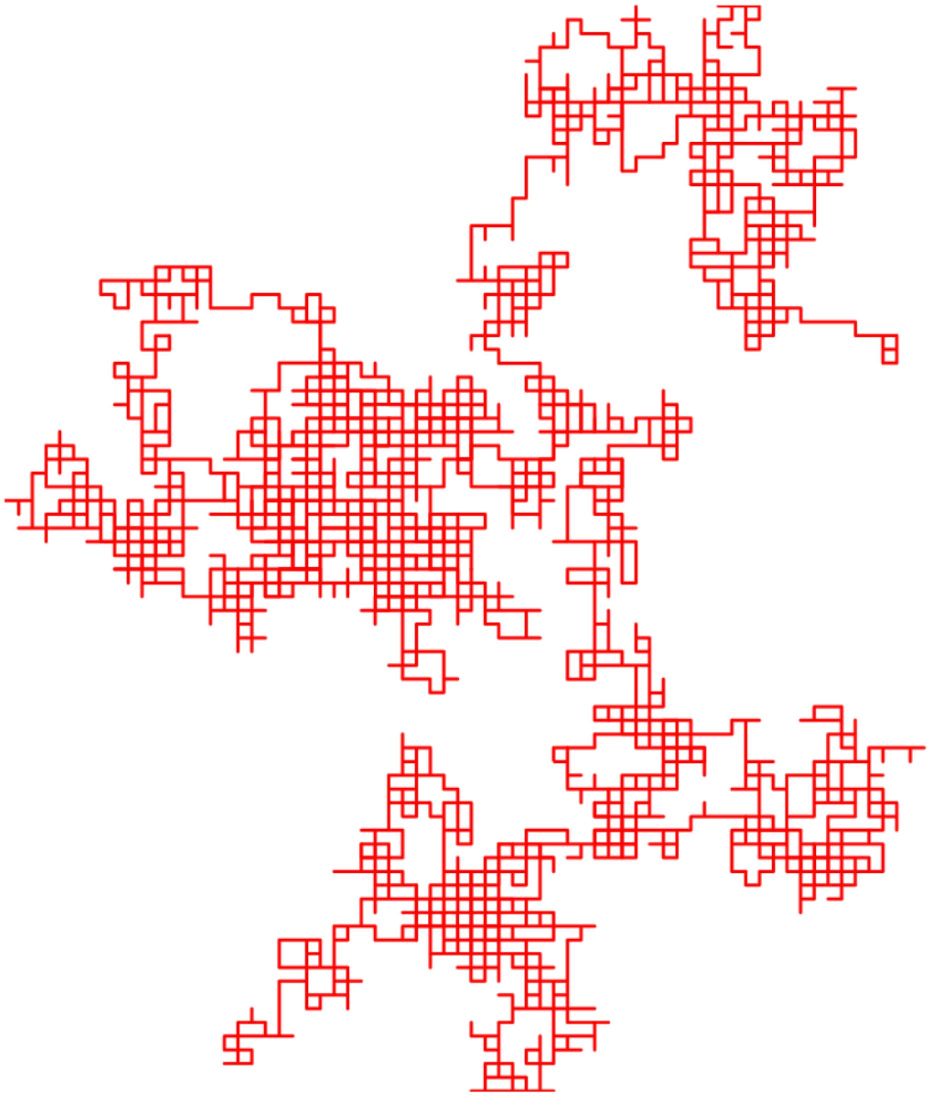


Markov Chains and Random Walks

Two-Dimensional Random Walk: Start at (0,0); at each step, flip two coins, one to determine if you should go left or right, and one to determine if you should go up or down. You go in one of four directions with ¼ probabiliy:



Markov Chains and Random Walks



Markov Chains and Random Walks

Antony Gormley's *Quantum Cloud* sculpture in London was designed by a computer using a random walk algorithm.

End of digression!

Markov Chains: Essential Properties

There are two equivalent ways of thinking about the dynamics of such a system, depending on whether you are interested in:

- One individual moving among the states – thus you start with a probability distribution describing whether this individual is likely to start his journey; at each time step he chooses the next state with the probabilities labelling the edges.

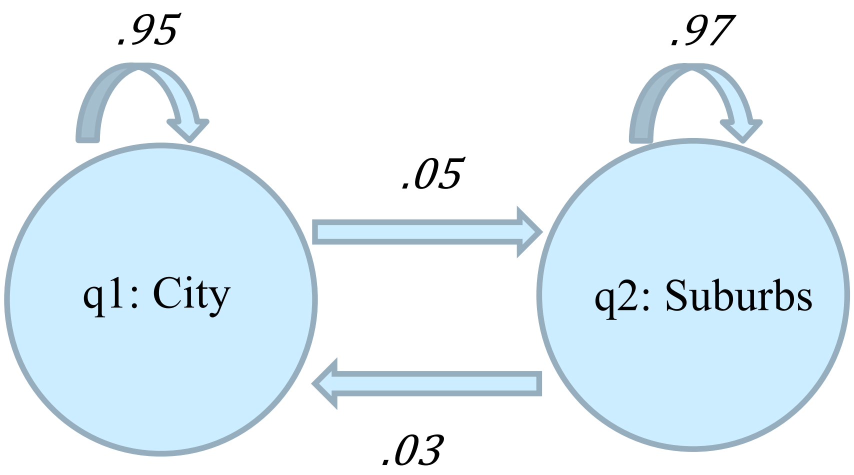
- A population of many individuals moving en mass among the states – thus you start with a percentage of the total population in each state, and track how they move over time (again, following the probabilities on the edges).

The most important property of a Markov Chain is that it is a stochastic process (depending on random events) which is memory-less: the behavior of an individual is independent of past states he may have been in.

Most important questions to ask about a Markov Chain:

1. How does it evolve over time?
2. Does it converge to a steady state (where there is no change in the population dynamics)?
3. What happens if some of the information (e.g., transition probabilities) is unknown: can it be inferred from the observation of its behavior?
4. What extensions of the basic model are useful (e.g., Hidden Markov Models)?

Markov Chains; How do they evolve?



A	From City	From Suburbs
To City	.95	.03
To Suburbs	.05	.97

$$\pi = [\ .6, \quad .4]$$

OR:

$$\pi = [\ 600000, \quad 400000]$$

	City	Suburbs			From City	From Suburbs
2000	600000	400000		To City:	0.95	0.03
2001	582000	418000		To Suburbs:	0.05	0.97
2002	565440	434560				
2003	550204.8	449795.2				
2004	536188.416	463811.584				
2005	523293.343	476706.657				
2006	511429.875	488570.125				
2007	500515.485	499484.515				
2008	490474.246	509525.754				

Etc., until...

2226	375000.001	624999.999	
2227	375000.001	624999.999	
2228	375000.001	624999.999	
2229	375000.001	624999.999	
2230	375000.001	624999.999	
2231	375000.001	624999.999	
2232	375000.001	624999.999	
2233	375000.001	624999.999	
2234	375000.001	624999.999	
2235	375000.001	624999.999	
2236	375000.001	624999.999	
2237	375000.001	624999.999	
2238	375000.001	624999.999	
2239	375000	625000	
2240	375000	625000	
2241	375000	625000	
2242	375000	625000	
2243	375000	625000	
2244	375000	625000	
2245	375000	625000	

Markov Chains and Linear Algebra

We note that the overall state of the system can be represented by a vector, either of the population or the probabilities:

$$\frac{1}{1,000,000}\begin{bmatrix} 600,000 \\ 400,000 \end{bmatrix} = \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}$$

When a vector contains real numbers which sum to 1.0, we call it a probability vector.

Note that A also has a similar property: each of its columns sums to 1.0 as well. A square matrix with this property is called a stochastic matrix.

$$A = \begin{bmatrix} .95 & .03 \\ .05 & .97 \end{bmatrix}$$

Markov Chains and Linear Algebra

Taking this Markov Chain through its successive states can be accomplished through matrix multiplication:

Definition. A *Markov chain* is a dynamical system whose state is a probability vector and which evolves according to a stochastic matrix.

That is, it is a probability vector $\mathbf{x_0}$ and a stochastic matrix $A \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{x_{k+1}} = A\mathbf{x_k} \quad \text{for } k = 0, 1, 2, \dots$$

$$\mathbf{x_1} = A\mathbf{x_0} = \begin{bmatrix} .95 & .03 \\ .05 & .97 \end{bmatrix} \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix} = \begin{bmatrix} 0.582 \\ 0.418 \end{bmatrix}$$

$$\mathbf{x_2} = A\mathbf{x_1} = \begin{bmatrix} .95 & .03 \\ .05 & .97 \end{bmatrix} \begin{bmatrix} 0.582 \\ 0.418 \end{bmatrix} = \begin{bmatrix} 0.565 \\ 0.435 \end{bmatrix}$$

To answer the question for 2020, i.e., $k = 20$, we note that

$$\mathbf{x_{20}} = \overbrace{A \cdots A}^{20} \mathbf{x_0} = A^{20}\mathbf{x_0}.$$

Markov Chains and Linear Algebra

In [49]:

```
import numpy as np

# stochastic matrix A
A = np.array(
    [[0.95,0.03],
     [0.05,0.97]])
#
# initial state vector x_0
x_0 = np.array([0.60,0.40])
print('x_0:',x_0)
#
# compute A times x_0
x_1 = A @ x_0
print('x_1:',x_1)
#
# compute A times x_0
x_2 = A @ x_1                    #  maxtrix multiplication is @
print('x_2:',x_2)
#
#
# compute A^20
A_20 = np.linalg.matrix_power(A, 20)
#
# compute x_20
x_20 = A_20 @ x_0
print('x_20',x_20)

#
# compute A^240
A_240 = np.linalg.matrix_power(A, 240)
#
# compute x_240
x_240 = A_240 @ x_0
print('x_240',x_240)
```

```
x_0: [0.6 0.4]
x_1: [0.582 0.418]
x_2: [0.56544 0.43456]
x_20 [0.417456 0.582544]
x_240 [0.375 0.625]
```

Markov Chains and Linear Algebra

What can we say about the distant future? Will a given Markov Chain eventually stabilize to a steady state?

We can use Linear Algebra to answer this question:

Generalizing MCs: Hidden Markov Models

Hidden Markov Models are Markov Chains with observations and emission probabilities. The states are considered to be unobservable or “hidden.”

$Q = q_1 q_2 \ldots q_N$	a set of N states
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

$$Q = \{1\,(Healthy),\ 2\,(Fever)\}$$

$$\pi = [0.6,\ 0.4]$$

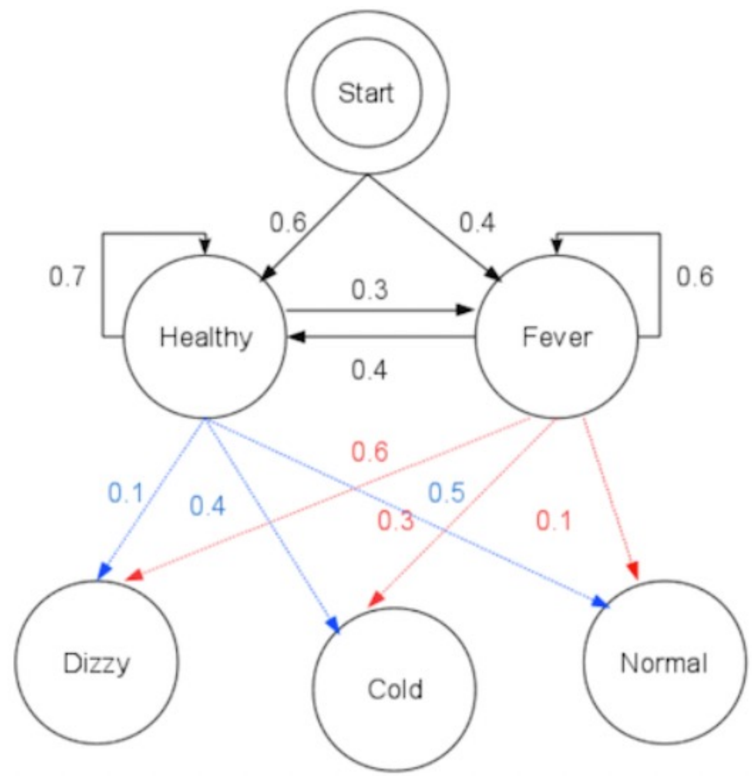
A =

	1	2
1	0.7	0.3
2	0.6	0.4

$$O = \{1\,(Dizzy),\ 2\,(Cold),\ 3\,(Normal)\}$$

B =

	1	2	3
1	0.1	0.4	0.5
2	0.6	0.3	0.1



Hidden Markov Models

$Q = \{\,1\,(Healthy),\ 2\,(Fever)\}$

$\pi = [\,0.6,\ 0.4\,]$

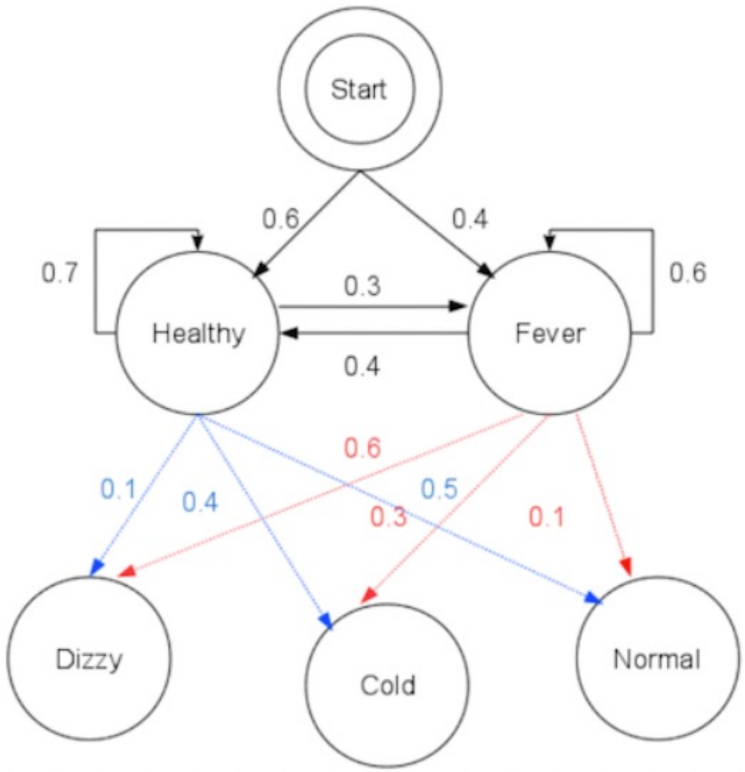
A =

	1	2
1	0.7	0.3
2	0.6	0.4

$O = \{\,1\,(Dizzy),\ 2\,(Cold),\ 3\,(Normal)\}$

B =

	1	2	3
1	0.1	0.4	0.5
2	0.6	0.3	0.1



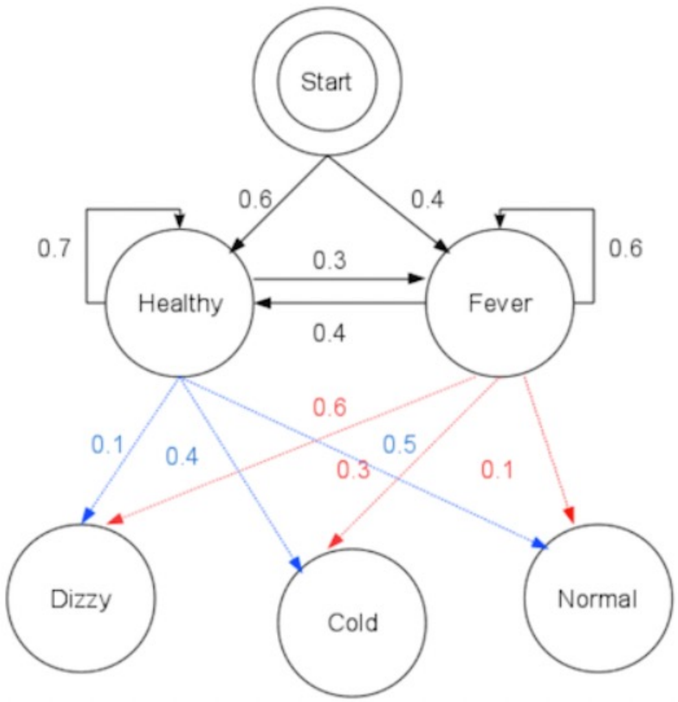
Hidden Markov Models

There are three main problems that are studied with HMMs:

Evaluation problem. Given the HMM M=(A, B, π,O,B) and the observation sequence $o_1 o_2 \ldots o_K$, calculate the probability that model M has generated the sequence.

Decoding problem. Given the HMM M=(A, B, π,O,B) and the observation sequence $o_1 o_2 \ldots o_K$, calculate the most likely sequence of hidden states $s_1, s_2, \ldots s_K$, that produced this observation sequence.

Learning problem. Given some training observation sequences $o_1 o_2 \ldots o_K$ and general structure of HMM (numbers of hidden and visible states), determine HMM parameters M=(A, B, π,O,B) that best fit the training data (alternately, determine some subset of the parameters, the others being given).



Hidden Markov Models for POS Tagging

Map from sequence $x_1,\ldots,x_n$ of words to $y_1,\ldots,y_n$ of POS tags