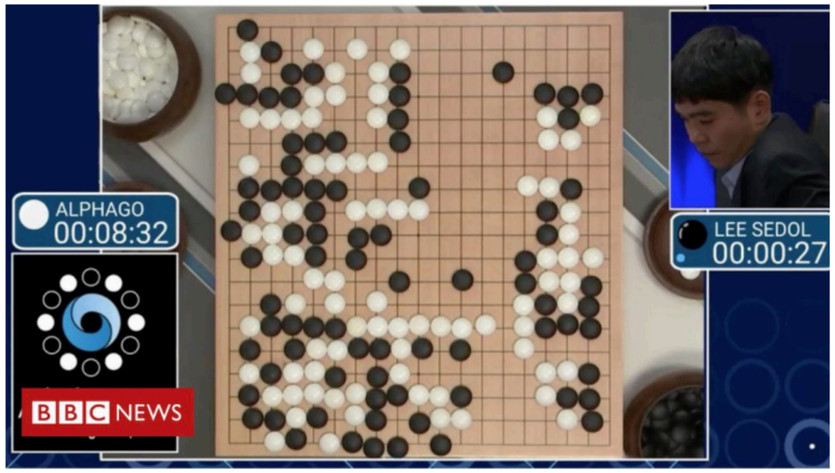
Introduction to Reinforcement Learning

- Meeting time is MW 4:30. Office hours TBA. I am Alex Olshevsky, http://sites.bu.edu/aolshevsky/
- Your TA is Andrew Huss, ahuss@bu.edu
- Lecture vs discussion.
- What is this class about?
- Machine learning is about learning rules from observing patterns.
- Old image recognition approach: try to build a model of what is in the image.
- Modern image recognition approach: create a data set of millions of labeled images, and try to learn a complicated rule that predicts labels from images.
- Old(er) robotics: design control strategies for robots to walk.
- New(ish) robotics: try to build robots that learn to walk from trial and error.
- The machine learning paradigm has achieved a number of recent successes.

(a) A robot hand solves a Rubik's cube.



(b) AlphaGo plays the (human) world champion.

(c) Spectators watch AlphaStar play Starcraft.

(d) A quadruped navigates a rocky terrain.

- Very roughly speaking, we can subdivide machine learning into supervised and unsupervised.
- Supervised: you need to create a data set with millions of images. You create the labels.
- Unsupervised: no labels.
- RL is unsupervised learning. It’s about learning from experience.
- Example: an agent that learns to play Pacman by playing many games (you will code this!).
- A car that learns to race from many crashes (you will code this too!).
- Starting point: what is a good algorithm for learning from trial and error?
  Easy: do I like mousse or croissants more?
  More difficult: think about how a baby learns to walk.

- This class: will teach you the basic algorithms of RL.
- Lecture fairly heavy on mathematical theory.
  We will derive and motivate the basic methods.
  Implementing RL methods using existing packages is not difficult...goal is to understand where the algorithms come from.
- Will ask you to think about the derivations done in class on the homeworks.
- You will have one midterm, which will be easier than the homework, and reflect a more basic level of understanding.
- The last part of the class will be dedicated to your final project.
  There will be no final exam.

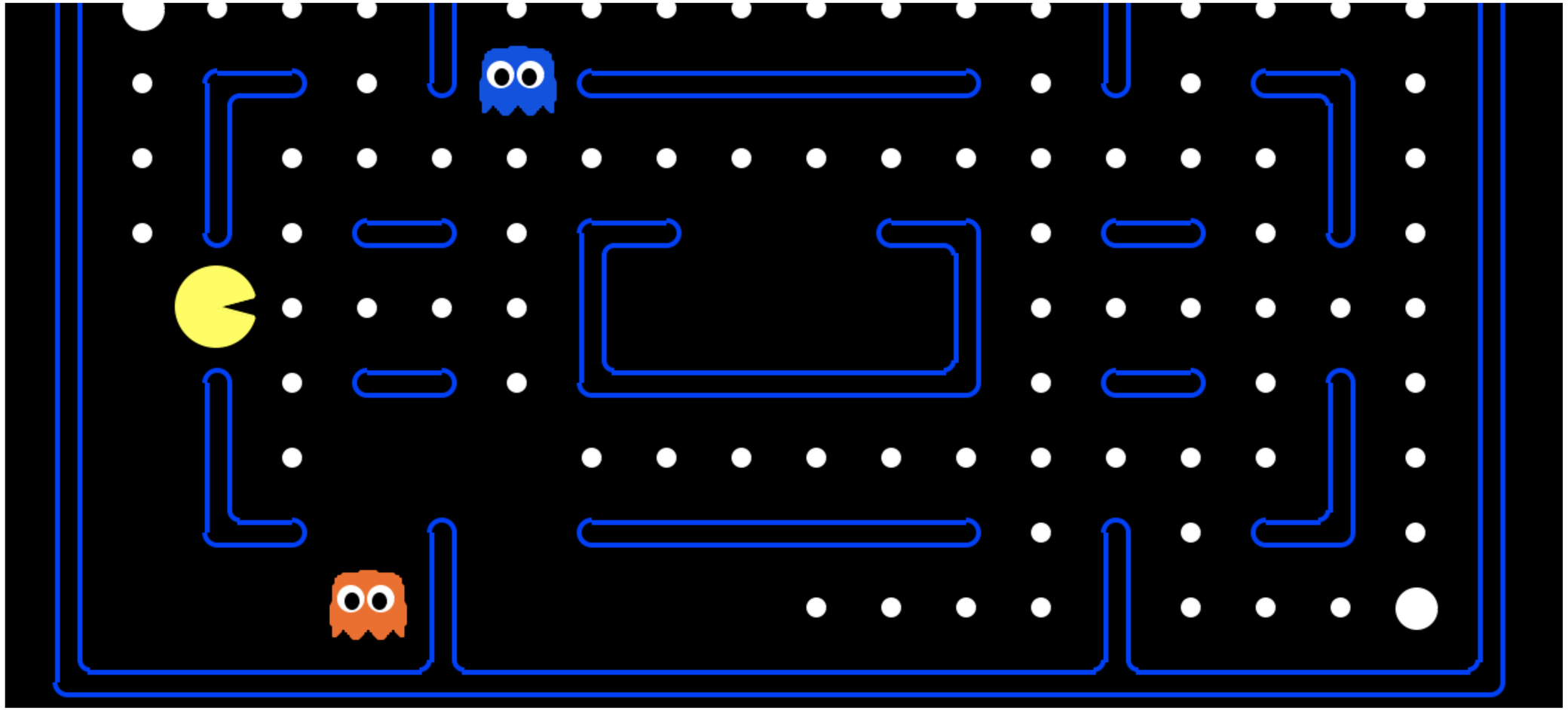
- Homework: 20%
- Coding exercises: 20%
- Midterm: 30%
- Final project: 30%
- The final project can be done in a group of up to X people.
  X is TBD but will probably be between 5-7.
- You will need to both write up a report and give a presentation.
- The last 1-2 weeks of class will be devoted to such presentations.
- [Go over syllabus]

actions

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

$$R_t = -1$$
on all transitions

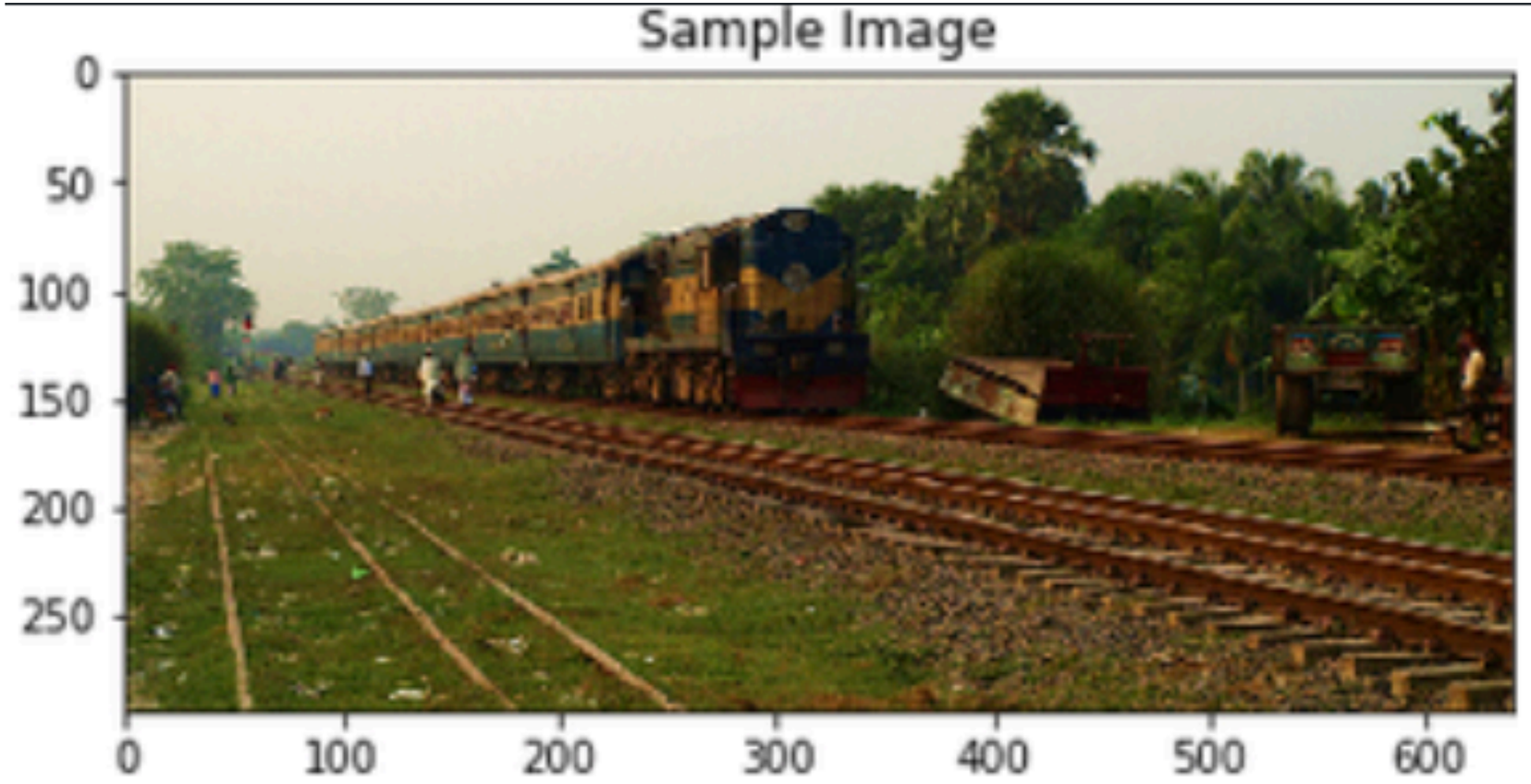
Your first two coding exercises: you'll basically play around with shortest path in this kind of graph
This is a toy example.
It's useful because it allows you to understand the ideas in this class in the simplest possible setting.



Coding exercise 3: Pacman

Coding exercise 4: PyTuxKart
Your final project will be improving the algorithms you designed in coding assignment 4 to race on many courses
There will be a competition among the teams in the class
[Videos]

- Class will be taught in Python. More or less any ML these days has to be in Python.
- I will have Andrew give you an introductory lecture on Python basics **next Wednesday**.
- Mostly, however, I'll expect you to work things out yourself.
  This is a useful like skill.
- Suggestion: use piazza!
  If you have a problem, others probably might as well.
- If something doesn't work, a useful trick is to google the error message.
- Start on the coding exercises early!
  It is **very hard** to predict how long things will take.
  Expect that making things work will take the vast majority of your time.
- You should expect to do well if:
  — you got a good grade (A, A-) in EK 381.
  — you either have lots of coding experience, or are confident of your ability to learn, or can spend a lot of time coding.
- Conversely....
- Keep in mind: this class is one level **harder** EK 381:
  — the math during lecture is more advanced
  — very little hand-holding on coding assignments
- Along the way, we will learn some pretty fun machine learning.




```
a train on a train track with trees in the background
```

Image-to-Caption

Instructions: There is a country in northern Asia that starts with "M". Go south from there, across a desert, and you will enter a very large country. Take the population of that larger country. Now divide that number by 10. Which country starts with "R" and has approximately that many people? That country used to be part of a much larger socialist state. One of the countries that was a part of that state starts with "G" and is also a female first name. Which American artist, born in 1887, has that first name? She died in which US state? Take the last of the two words of that state, which is the name of a country. What is the name of a peninsula in that country that starts with "Y"? At the tip of that peninsula is a popular vacation spot that starts with "C". If you were to sail from that location, what large island nation would you find? Who led that country until 2008? His first name is based on what Latin word, and what does it mean in English? What letter does that word start with? What profession starts with that letter and depends on the rain and sunshine? What is something from there that starts with "B" and is sold in strips? What is an American actor whos last name is that product?

Results: **Mongolia is a country in northern Asia that starts with "M". Going south from there, across a desert, you enter China, which is a very large country. China's population is approximately 1.3 billion. Dividing that by 10 gives 130 million. The country that starts with "R" and has approximately that many people is Russia. Russia was part of the Soviet Union, which was a socialist state. Georgia is a country that was a part of that state. The American artist with that first name is Georgia O'Keeffe. She died in New Mexico. The last of the two words of that state is "Mexico". The peninsula in Mexico that starts with "Y" is Yucatan. The popular vacation spot that starts with "C" is Cancun. If you were to sail from that location, you would find Cuba. Fidel Castro led Cuba until 2008. His first name is based on the Latin word "fidelis", which means "faithful". The letter that word starts with is "F". The profession that starts with that letter and depends on the rain and sunshine is farming. Something from there that starts with "B" and is sold in strips is bacon. An American actor whose last name is that product is Kevin Bacon.**

GPT-3 example

RL Basics

- Questions?
- RL is about learning from interactions with the world.
- Basic setup:

  — you are at a state

  — you take an action

  — you observe a reward

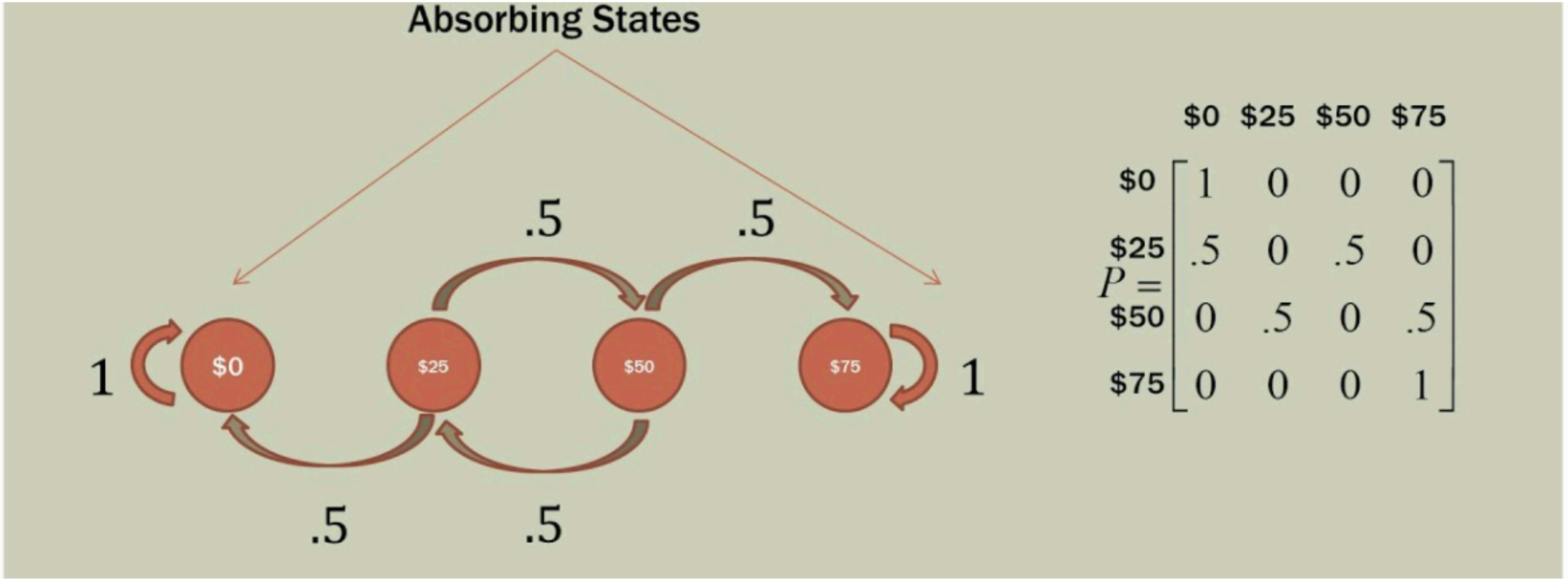
  — you transition to a new state

  Can you figure out what the best actions to take?
- Think about playing a computer game as an example.

- RL is useful because it **can** be model-free. Many systems are too complicated to model explicitly.
- Even when systems are not that complicated, RL seems to perform quite well.
- “Do you really want to code a system that has to learn that gravity pulls down?” Maybe…
- Applications go beyond merely game playing to optimization, control, robotics.

- Reinforcement learning uses the theory of Markov chains quite heavily. So we'll start by discussing this theory.
- The assumption here is that you've seen Markov chains before in EK 381. So we'll go very quickly.
- Refer back to your 381 lecture notes and lecture videos for a refresher.
- A discrete random variable takes values in some finite or countable set (as opposed to a continuous random variable, which can be real-valued).
- Good news (sort of): almost all the random variables in this class will take values in a finite set.
- We can consider an infinite sequence of random variables $X_1, X_2, X_3, \ldots$
- DEFINITION: a sequence if discrete random variables $X_1, X_2, \ldots$ is called a Markov Chain if $P(X_t = j \,|\, X_{t-1} = i, X_{t-2} = i', X_{t-3} = i'', \ldots) = P(X_t = j \,|\, X_{t-1} = i)$
- We will sometimes use the shorthand ``The Markov chain is in state i at time t'' to denote the event $\{X_t = i\}$.

- OK, suppose the sequence is a Markov Chain. We will usually write $p_{ij}(t) = P(X_t = j \,|\, X_{t-1} = i)$.
- The quantities $p_{ij}(t)$ are a way to describe the process.
- In the case where $p_{ij}(t)$ does not depend on t, we will say that the Markov chain is homogeneous.
- In that case, we'll just write p_{ij}. Good news: almost all Markov chains in this class will be homogeneous.
- Suppose that each random variable X_i has range in $\{1, \ldots, n\}$. Then $\sum_{j=1}^{n} p_{ij} = 1$ for all $i = 1, \ldots, n$.
- It is standard to stack these up into the matrix $P = [p_{ij}]$. This is called the transition matrix.
- The matrix P is an $n \times n$ matrix which is nonnegative and its rows sum up to one. Such matrices are called stochastic.

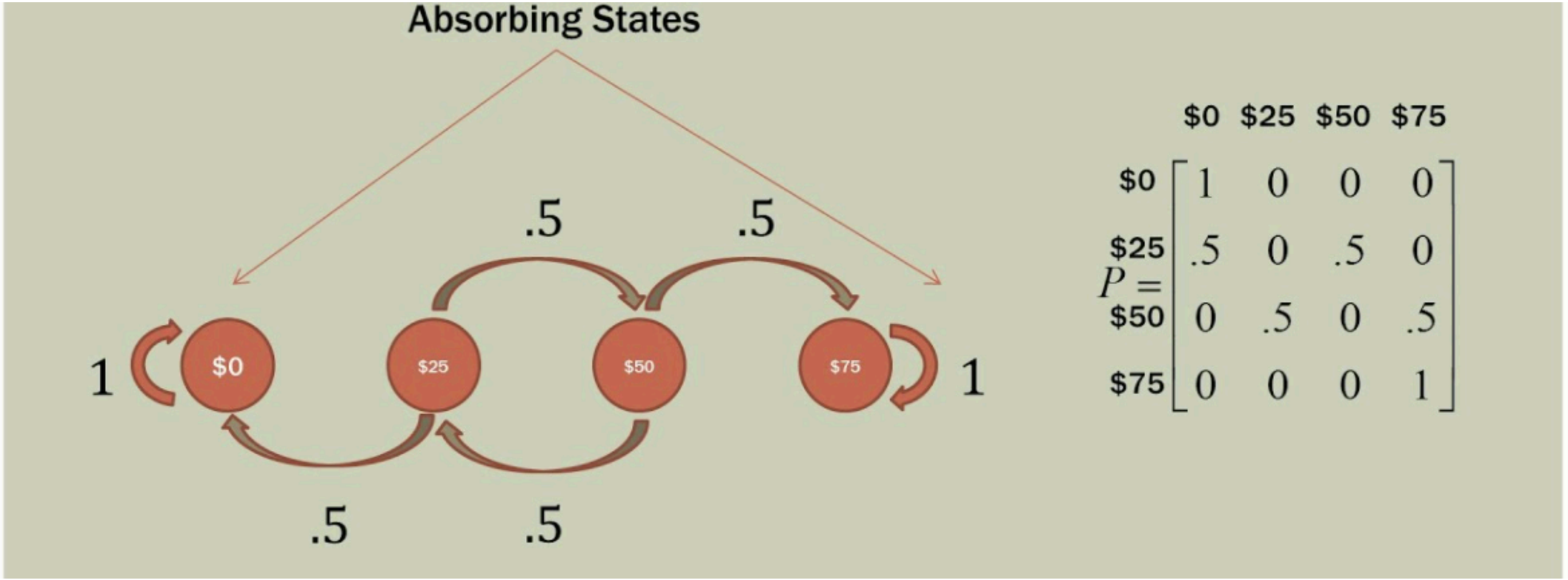


Credit to Brandon Folz. This is a common "Gambler's Ruin" Markov chain.
Observe the transition matrix is nonnegative with rows adding up to one.
The columns, however, need not add up to one. The matrix need not be symmetric.

- Suppose we are given an homogeneous Markov chain $X_0, X_1, X_2, \ldots$. Let's talk about two-step transition probabilities.
- Suppose the Markov chain takes values in $\{1, \ldots, n\}$ and suppose i_0, i_1, i_2 are all in $\{1, \ldots, n\}$ (informally, we will say that i_0, i_1, i_2 are states). Can we write $P(X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0)$ in terms of the quantities p_{ij}?
- Can write:

$$P(X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0) = P(X_2 = i_2 \,|\, X_1 = i_1, X_0 = i_0) P(X_1 = i_1, X_0 = i_0 \,|\, X_0 = i_0)$$

- Indeed, the LHS is $P(X_2 = i_2, X_1 = i_1, X_0 = i_0)/P(X_0 = i_0)$.
- Whereas the RHS is $\dfrac{P(X_2 = i_2, X_1 = i_1, X_0 = i_0)}{P(X_1 = i_1, X_0 = i_0)} \dfrac{P(X_1 = i_1, X_0 = i_0)}{P(X_0 = i_0)}$
- Thus $P(X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0) = P(X_2 = i_2 \,|\, X_1 = i_1) P(X_1 = i_1 \,|\, X_0 = i_0) = p_{i_1 i_2} p_{i_0 i_1}$
- Punchline: given the matrix P, we can compute the probability of any two-step trajectory.



$$P = \begin{array}{c} \\ \$0 \\ \$25 \\ \$50 \\ \$75 \end{array} \begin{array}{c} \begin{array}{cccc} \$0 & \$25 & \$50 & \$75 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

P(X_2 = 50, X_1 = 25 | X_0 = 50) = (1/2) x (1/2) = 1/4

What about P(X_2 = 50 | X_0 = 50)?

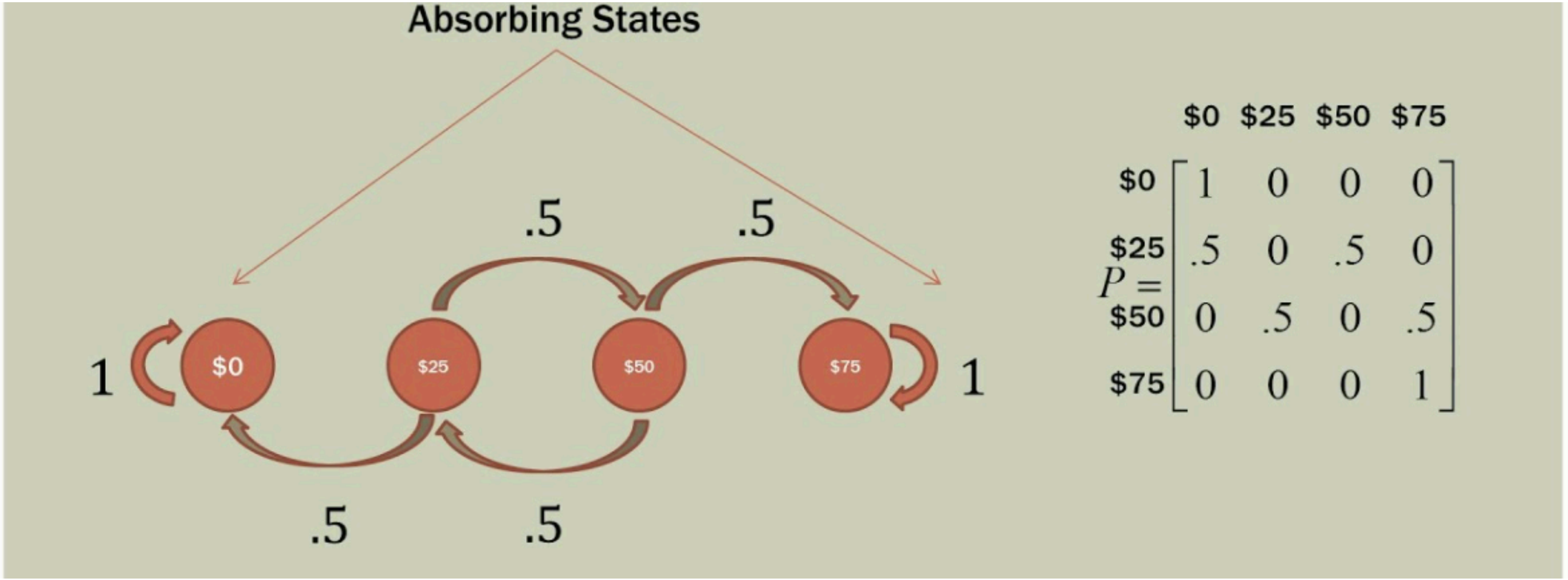
- Analogously:

$$P(X_3 = i_3, X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0) = p_{i_2 i_3} p_{i_1 i_2} p_{i_0 i_1}$$

- More generally:

$$P(X_n = i_n, \ldots, X_1 = i_1 \,|\, X_0 = i_0) = p_{i_{n-1} i_n} p_{i_{n-2} i_{n-1}} \cdots p_{i_0 i_1}$$

- Punchline: given the matrix P, we can compute the probability of any trajectory.



$$P = \begin{array}{c} \\ \$0 \\ \$25 \\ \$50 \\ \$75 \end{array} \begin{array}{c} \begin{array}{cccc} \$0 & \$25 & \$50 & \$75 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

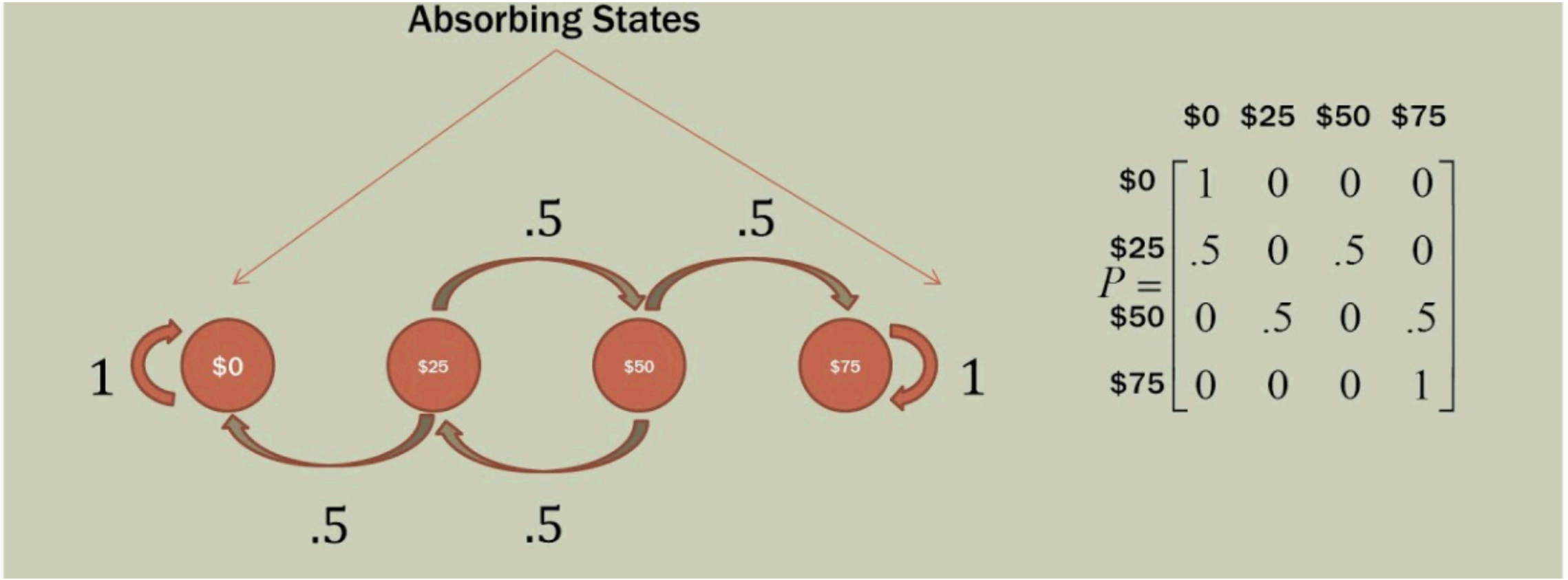
$$P(X_3 = 75, X_2 = 75, X_1 = 75 \mid X_0 = 50) = (1/2) \times 1 \times 1 \times 1 = 1/2$$

- OK, so we know how to compute $P(X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0)$.
- What about $P(X_2 = i_2 \,|\, X_0 = i_0)$?
- Just do:

$$P(X_2 = i_2 \,|\, X_0 = i_0) = \sum_{i_1=1}^{n} P(X_2 = i_2, X_1 = i_1 \,|\, X_0 = i_0)$$

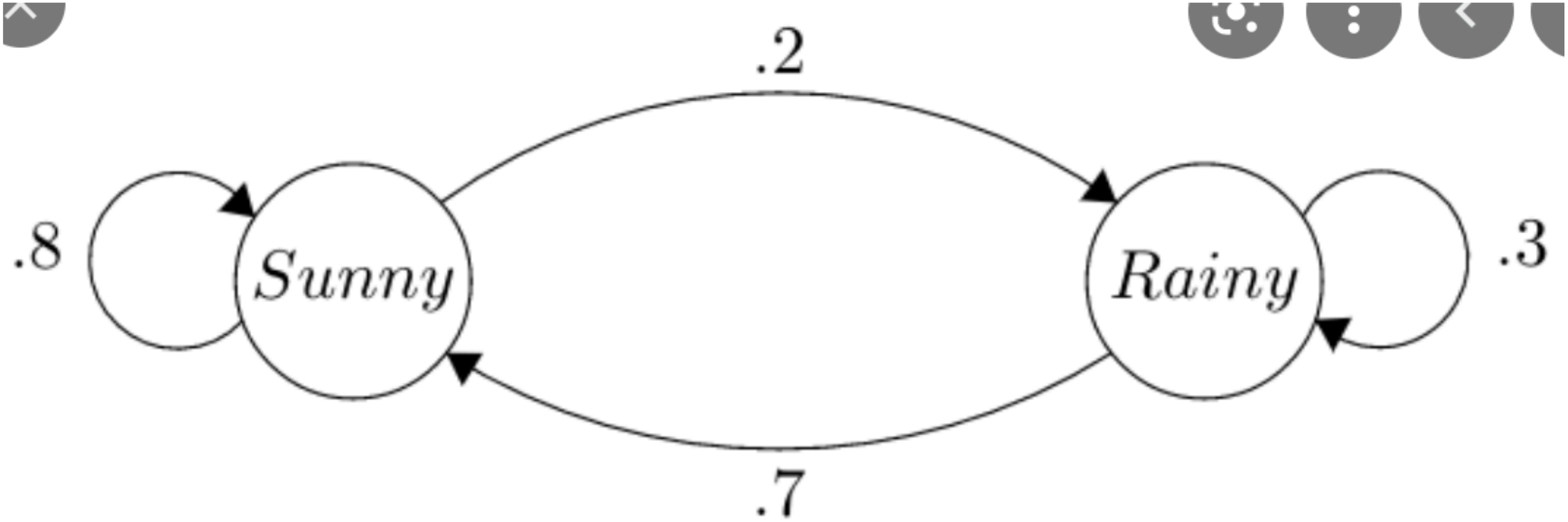
  recall here the Markov chain takes values in $\{1, \ldots, n\}$.

- So $P(X_2 = i_2 \,|\, X_0 = i_0) = \sum_{i_1=1}^{n} p_{i_0 i_1} p_{i_1 i_2}$



$$P = \begin{array}{c} \\ \$0 \\ \$25 \\ \$50 \\ \$75 \end{array} \begin{array}{c} \begin{array}{cccc} \$0 & \$25 & \$50 & \$75 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$P(X_2 = 50 \mid X_0 = 50) = P(0|50)\,P(50|0) + P(25|50)\,P(50|25) + P(50|50)\,P(50|50) + P(75|50)\,P(50|75) = P(25|50)\,P(50|25) = 1/4$$



P(X_2 = Sunny| X_0 = Rainy) = P(Sunny | Rainy) P(Sunny | Sunny) + P(Rainy | Rainy) P(Sunny | Rainy) = 0.7 x 0.8 + 0.3 x 0.7

- Let's take this equation $P(X_2 = i_2 \,|\, X_0 = i_0) = \sum_{i_1=1}^{n} p_{i_0 i_1} p_{i_1 i_2}$

  and examine it. Does ti remind you for something?
- Let us define the matrix $T_{ij} = P(X_2 = j \,|\, X_0 = i)$.
- So $T_{ij} = \sum_{k=1}^{n} p_{ik} p_{kj}$. What is this in matrix form?
- $T = P^2$.
- Very natural:

  — P is the matrix of "one step" transition probabilities.

  — To get the ``two step'' transition probabilities, just square P.

- OK, now let's consider $P(X_3 = i_3 \,|\, X_0 = i_0)$.
- Let's adopt the notation $[M]_{ij}$ to denote the i,j'th entry of the matrix M.
- Simple trick:

$$\begin{aligned} P(X_3 = i_3 \,|\, X_0 = i_0) &= \sum_{i_2=1}^{n} P(X_3 = i_3, X_2 = i_2 \,|\, X_0 = i_0) \\ &= \sum_{i_2=1}^{n} P(X_3 = i_3 \,|\, X_2 = i_2, X_0 = i_0) P(X_2 = i_2, X_0 = i_0 \,|\, X_0 = i_0) \\ &= \sum_{i_2=1}^{n} [P]_{i_2 i_3} [P^2]_{i_0 i_2} \end{aligned}$$

- Conclusion: $P(X_3 = i_3 \,|\, X_0 = i_0) = [P^3]_{i_0 i_3}$.
- Makes sense.
- Most general form: $P(X_n = b \,|\, X_0 = a) = [P^n]_{ab}$

- OK, all this is usually stated in the following way.
- Let's switch back to the indices i,j.
- We can talk about t-step transition probabilities:

  $r_{ij}(t) = P(X_t = j \,|\, X_0 = i).$

- We can stack $r_{ij}(t)$ into the matrix $R(t)$.
- What we showed: $R(t) = P^t$.
- Makes sense: P tells you everything you want to know about the chain.
- Let's talk about the rows of these matrices. Consider, for example, the first row of P:

  $[p_{11}, p_{12}, p_{13}, \ldots, p_{1n}]$

  or alternatively

  $[P(X_1 = 1 \,|\, X_0 = 1), P(X_1 = 2 \,|\, X_0 = 1), \ldots, P(X_1 = n \,|\, X_0 = 1)]$

- This is the distribution of the state after one step given that the initial state is one.

- Likewise, consider the second row

  $[p_{21}, p_{22}, p_{23}, \ldots, p_{2n}]$

  or alternatively

  $[P(X_1 = 1 \,|\, X_0 = 2), P(X_1 = 2 \,|\, X_0 = 2), \ldots, P(X_1 = n \,|\, X_0 = 2)]$
- This is the distribution of the state after one step given that the initial state is 2.
- Similarly, the k'th row of the matrix P is the distribution of the state after one step, given that the initial state is k.
- There is a way this is written in linear algebra. You might have seen the notation

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

- Then we $\mathbf{e}_k^T P$ is just the k'th row of P.
- We thus have that the distribution after one step starting from node k is $\mathbf{e}_k^T P$.

- Similarly, if we consider the e.g., third row of P^2, the entries are $[P(X_2 = 1 \,|\, X_0 = 3), P(X_2 = 2 \,|\, X_0 = 3), \ldots, P(X_2 = n \,|\, X_0 = 3)]$
- So the distribution after two steps of the chain, starting at node 3, is $\mathbf{e}_3^T P^2$
- More generally, the distribution after k steps provided you start at node i is $\mathbf{e}_i^T P^k$.
- If you initial distribution is $\mathbf{p}$, then after k steps your distribution is $\mathbf{p}^T P^k$.
- If your distribution at time t is $\mathbf{p}$, then your distribution at time $t + k$ is $\mathbf{p}^T P^k$.
- Summary: “moving forward” in the Markov chain is a matter of matrix multiplication!

- **Summary**: suppose $X_0, X_1, X_2, X_3, \ldots$ is a homogeneous Markov chain and every random variable X_i takes values in the finite set $\{1,\ldots,n\}$.
- This random process is defined by the numbers $P_{ij} = P(X_{t+1} = j \,|\, X_t = j)$.
- We stack these numbers up into the $n \times n$stochastic matrix P, called the probability transition matrix .
- $\mathbf{e}_i$ is a column vector with a 1 in the i'th entry and zeroes in all other entries.
- $\mathbf{e}_i^T P$ is a vector of numbers adding up to one. These numbers represent the distribution of the **next** state after state i.

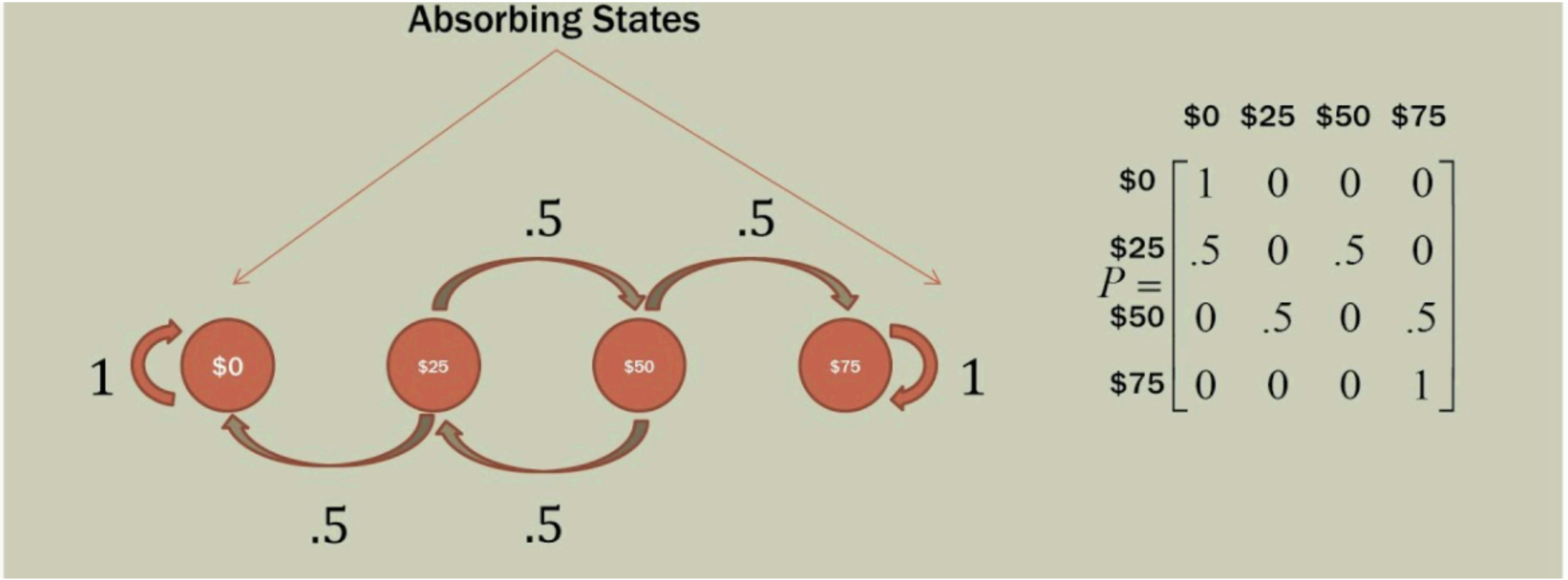
  This is a row matrix times a matrix.
- Aside: you can multiply vectors and matrices like that:

  $$\begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_1 b_{11} + a_2 b_{21} & a_1 b_{12} + a_2 b_{22} \end{pmatrix}$$

- $\mathbf{e}_i^T B$ is just the i'th row of B
- For all $i = 1,\ldots,n$,

  $$[aB]_i = \sum_{j=1}^{n} a_j B_{ji}$$

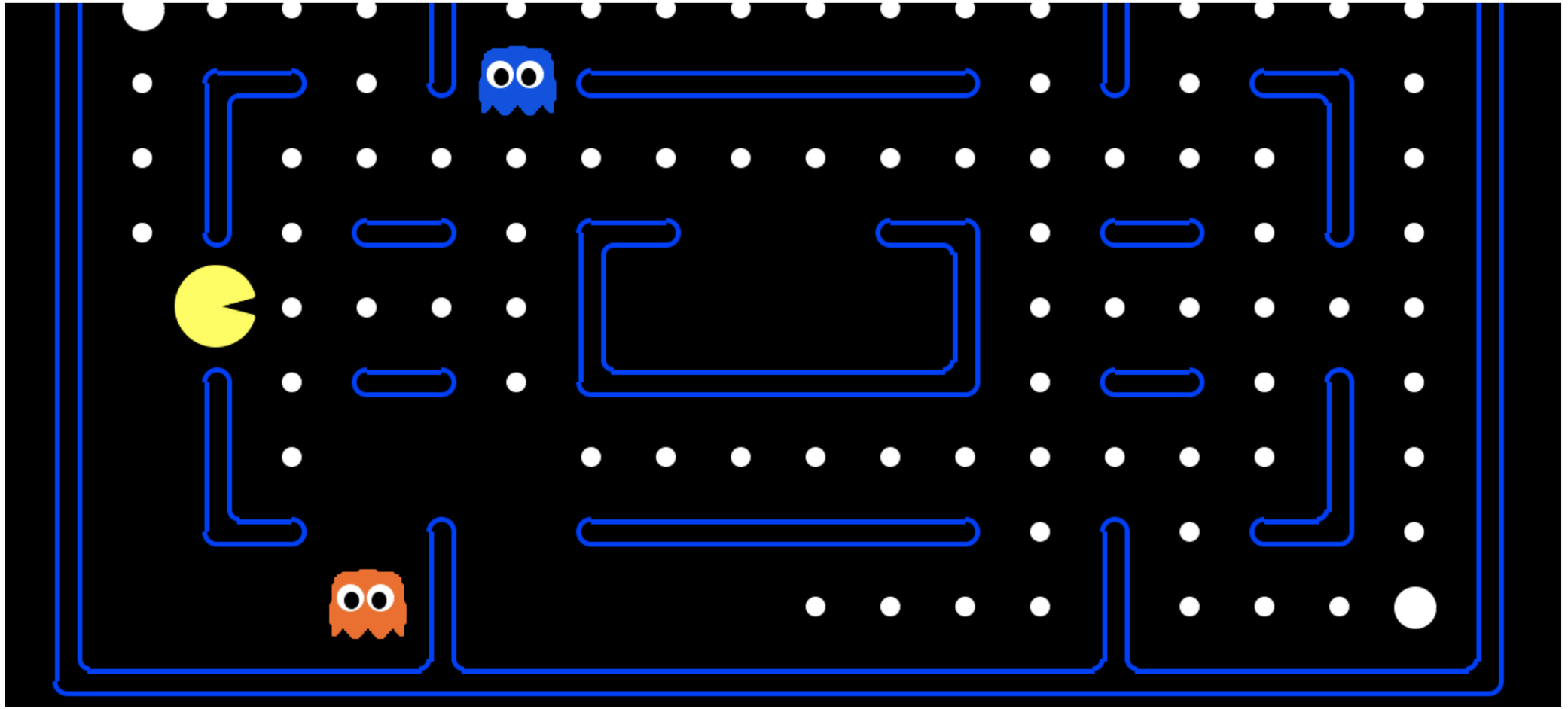
- The k-step chain is what happens if you look at the Markov chain at multiples of k, for example the 2-step chain might be $X_0, X_2, X_4, \ldots$
- The probability transition matrix of the k-step chain is P^k.
- The distribution of the next state in the k-step chain after state i is $\mathbf{e}_i^T P^k$.
- For example, the distribution of the next state after state 3 in the 2-step chain is $\mathbf{e}_3^T P^2$.
- Now suppose we initialize the Markov chain in a random location.
- Suppose the distribution of the initial location (i.e., of X_0) is in a vector $\mathbf{p}$.
- What is the distribution of X_k?
- Aside: don't get confused between P_{ij}, P, and $\mathbf{p}$, whose it's element is $\mathbf{p}_i$



$$P = \begin{array}{c} \\ \$0 \\ \$25 \\ \$50 \\ \$75 \end{array} \begin{array}{c} \begin{array}{cccc} \$0 & \$25 & \$50 & \$75 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Initial distribution: [0, 2/3, 1/3, 1]

- For all $i = 1, \ldots, n, P(X_k = i) = \sum_{l=1}^{n} P(X_k = i \,|\, X_0 = l) P(X_0 = l)$
- For all $i = 1, \ldots, n$,
  $P(X_k = i) = \sum_{l=1}^{n} [P^k]_{li} \mathbf{p}_l$
- $[P(X_k = 1), P(X_k = 2), \ldots, P(X_k = n)] = \mathbf{p}P^k$ if $\mathbf{p}$ is a row vector.
- If $\mathbf{p}$ is a column vector, then $[P(X_k = 1), P(X_k = 2), \ldots, P(X_k = n)] = \mathbf{p}^T P^k$
- In words: Markov chains propagate distributions forward by matrix multiplication.



Suppose the player moves towards the closest food
If there are multiple sources equally close, the player chooses randomly
Suppose the ghosts make random choices at each intersection
Is pacman a Markov chain?

- Consider a sequence of random variables $X_1, X_2, \ldots$ in which the distribution of X_t depends on X_{t-1}, X_{t-2}....but if you know, X_{t-1}, X_{t-2}, knowledge of other past samples is useless.
- More formally:

$$P(X_t = i_t \mid X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, X_{t-3} = i_{t-3}, X_{t-4} = i_{t-3}, \ldots, X_0 = i_0) = P(X_t = i_t \mid X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2})$$

- Simple trick: define $Y_t = \begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix}$. So Y_t is a random vector. If $X_t \in \{1, \ldots, n\}$, then Y_t can take one of n^2 possible values.
- Then the sequence $Y_0, Y_1, \ldots$ is a Markov chain!