Lecture 1: Introduction to Simulation

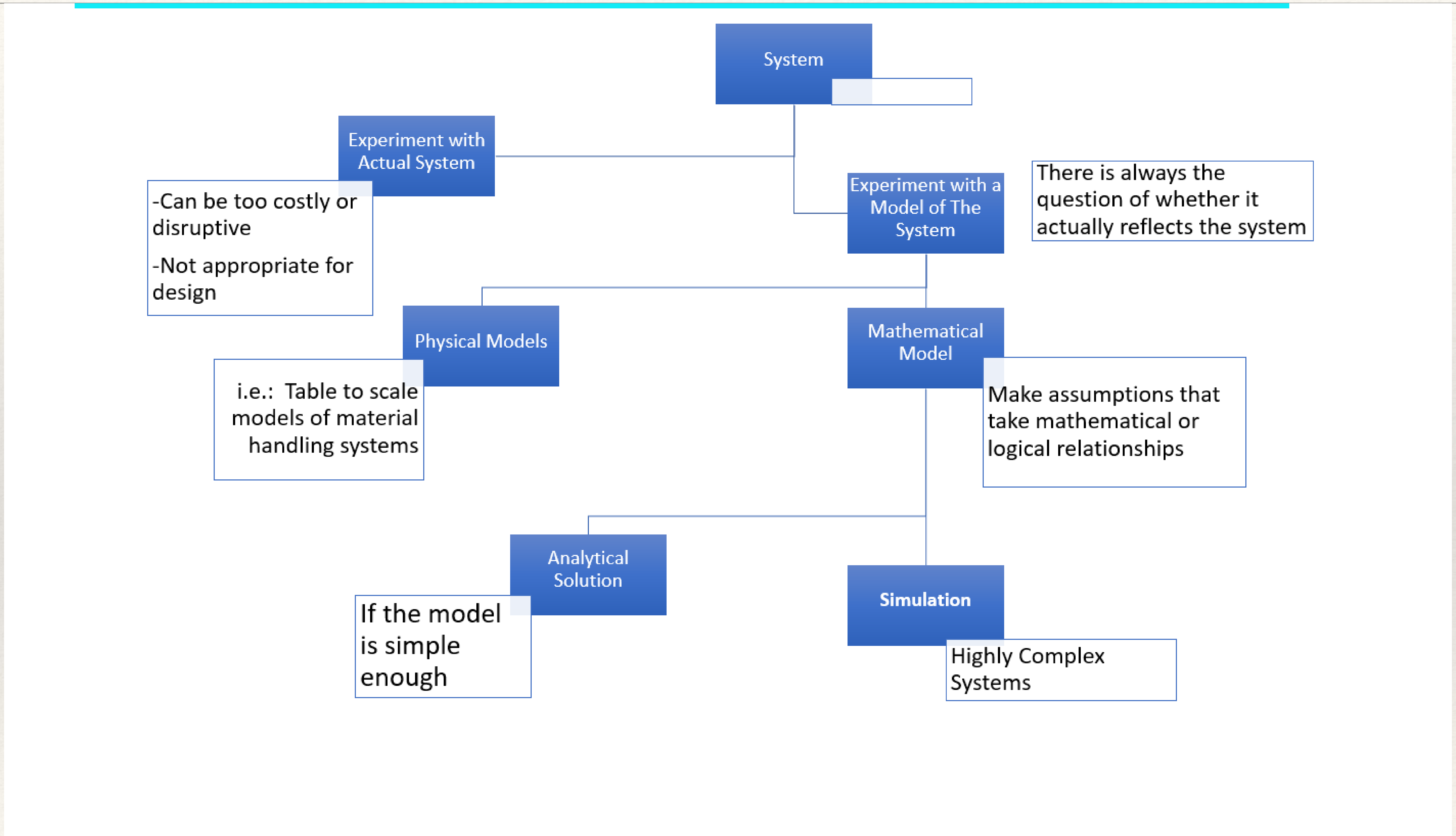
END 322E System Simulation

Mehmet Ali Ergün, Ph.D.

System

- A system is any set of related components that together work toward some purpose.
  - Manufacturing facility
  - Bank operations
  - Airport operations (passengers, security, planes, crews, baggage)
  - Transportation/logistics/distribution operation
  - Hospital facilities (emergency room, operating room, admissions)

Ways to Study a System



Model

- A model is the conceptualization of a real system
- Requires set of assumptions, hence a simplification
- Relations between entities of the system can be expressed in terms of
  - Mathematical
  - Logical
- Can be physical or mathematical

Modeling as an activity

- Why do we use models?
  - Show some aspects of the real-world systems
  - Ask what-if questions
  - They help us to understand what is going on so we can make decisions

- Famous Quotation:
  - *Essentially, all models are wrong, but some are useful.*

George E. P. Box (Professor Emeritus of Statistics at the University of Wisconsin)

- Good to remember: A model is an approximation

What is Simulation?

- *Simulation* – very broad term – simply: A **modeling** approach to imitate or mimic **complex** real systems, usually via computer. (A computer model of a real system)

Simulation is everywhere

- Entertainment
    - Many popular games are based on simulation.

- Aviation
    - Wind tunnel simulations help design more efficient planes.
    - Pilots trained on flight simulators mean fewer crashes.

- Manufacturing
    - Simulation is used to optimize production lines and schedule productions

- Military
    - Planning logistic operations
    - War scenario games

Some Application Areas

- Design/improve a processing center for shipping orders.
- Design/improve visitor flow at an amusement park.
- Design/improve an automobile assembly line.
- Design/improve the back office processes for a bank.
- Design/improve the baggage handling system at an airport
- Reduce waiting times in an emergency department.
- Process improvement at an ambulatory surgery center.
- Evaluate a disaster plan for emergency services.

When Simulation is Appropriate?

- To study and experiment on complex systems
- Informational, organizational, and environmental changes can be simulated to explore better alternatives
- Knowledge gained when building simulation model toward suggesting improvement
- To verify analytical solutions
- To prepare what-if scenarios
- Animations help visualization of the operations

Drawbacks of Simulation

- Results are not exact due to randomness, only approximations
  - May lead to misinterpretations
  - Statistical design, analysis of simulation experiments requires attention in detail
- Not easy to explore all possible alternatives, no optimality guarantee
- Sometimes very time-consuming/costly
  - One single run tells us little, often with high variance

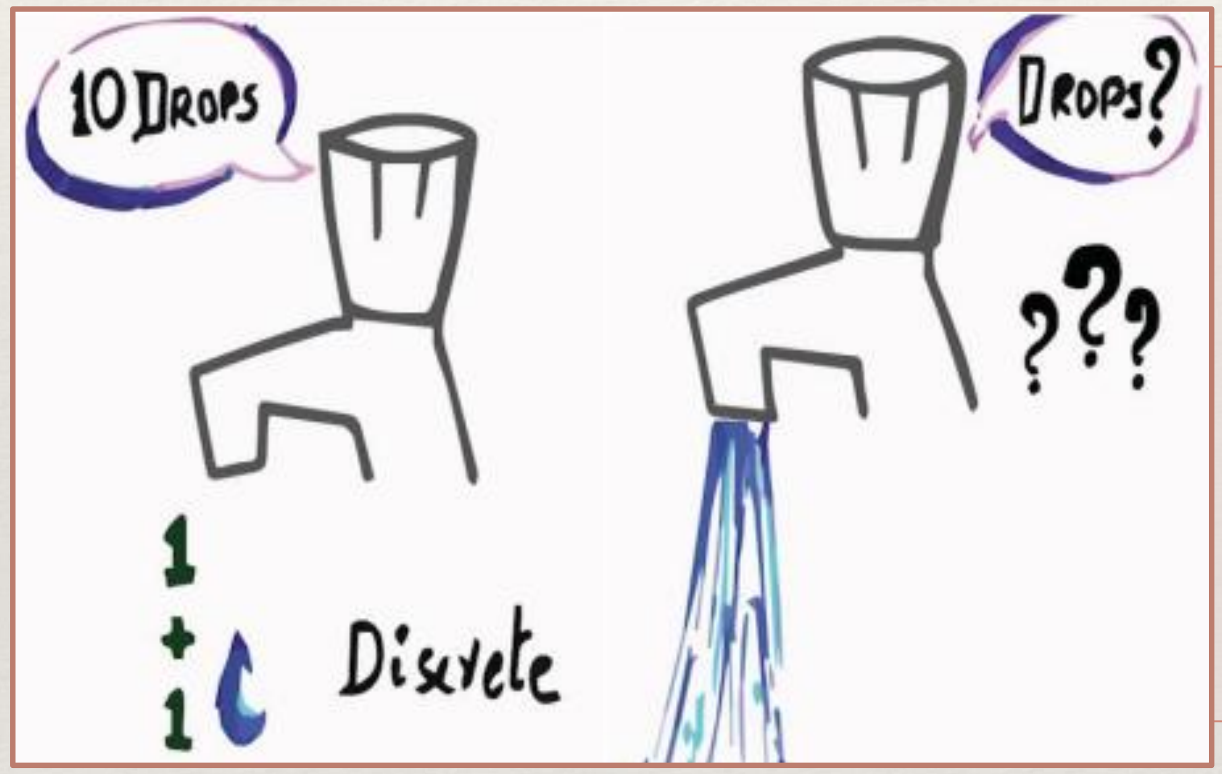
Don’t simulate when… (Banks & Gibson)

- A “common sense” solution exists
- The problem can be solved analytically*
  - i.e. simple queuing system
- If direct experiments are easier and cheaper to perform
- Simulation cost exceeds potential savings:
  - Remember to budget for: planning/problem definition, data collection, model validation, experimentation and analysis
- There are not enough resources (people, money, time) available

Types of Simulation

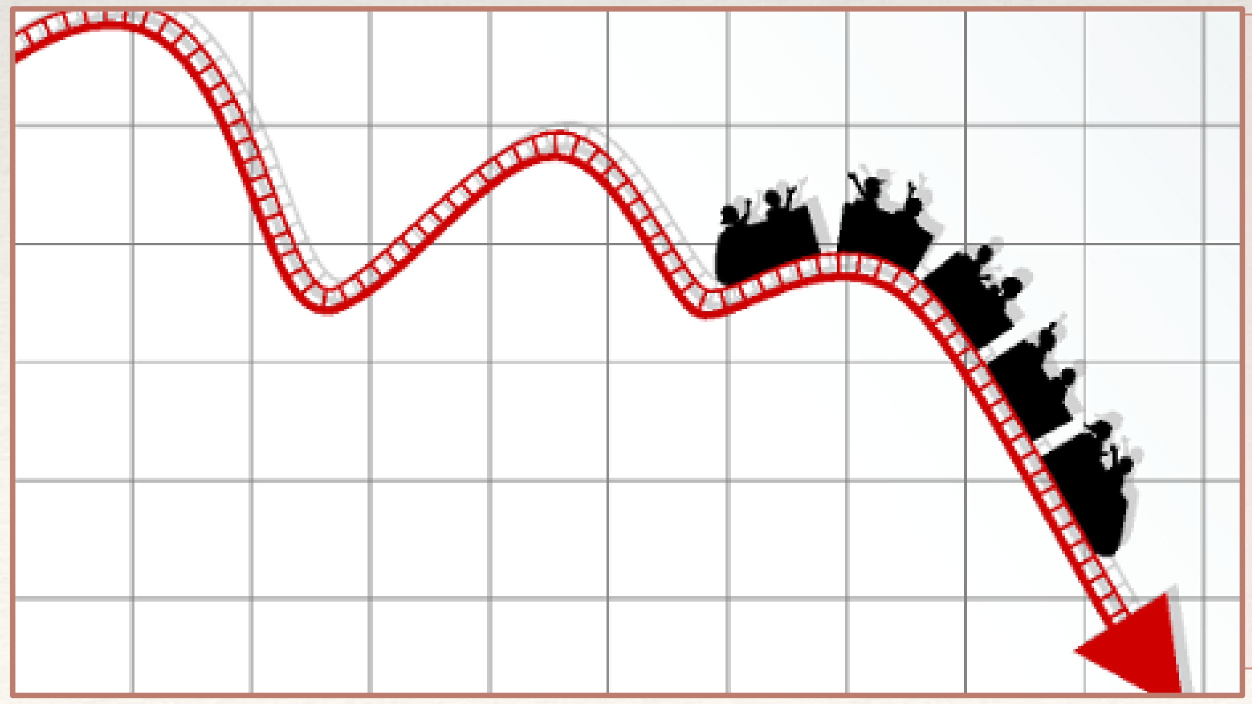
Static vs. *Dynamic*

- Does time have a role in model?



Continuous-change vs. *Discrete-change*

- States change continuously or at discrete points of time?



Deterministic vs. *Stochastic*

- Does randomness play a role?

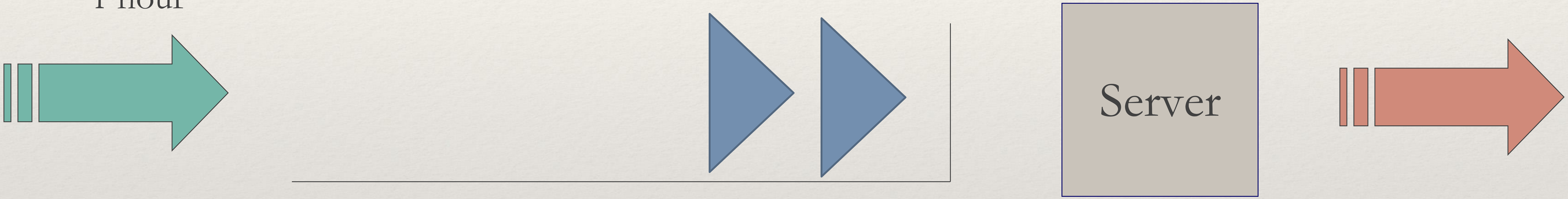
Randomness

- Simulation models that contain no random variables are called deterministic simulations
    - Scheduled arrivals to a dentist
- A stochastic simulation model has one or more random variables as inputs
    - Simulation of a bank involves random arrivals of customers and service times
- Random inputs lead to random outputs
- Output measures are averages, variances, and distributions, hence are statistical estimates

Randomness and System Performance

Patients arrive with an average time between arrivals of 1 hour

The average treatment time is 55 minutes.



How will this system perform?

Arrival/Service Process	**Result**
Constant/Constant	No waiting.
Random*/Constant	5 hour average wait.
Random*/Random*	10 hour average wait.

*** Exponential Distribution**

Randomness in a simulation

- Typical random components: processing time, service time, customer arrival times, transportation times, machine failures or repair times, etc.

- ***For example,*** if you head to the drive-through window at a local fast-food restaurant for a late-night snack, you cannot know exactly how long it will take you to get there, how many other customers may be in front of you when you arrive, or how long it will take to be served, to name just a few variables

Organizing a simulation study

- Most of the work is not the simulation itself

- Start with: what is the study about?
  - What will it be used for?
  - Who has to be involved?
  - What will be the measure of effectiveness?
  - How long will the study take?
  - What is the cost?

Conceptual Model

- Get the required information and the data
    - Not the same!
- Write an initial assumptions document
    - What is the model?
    - What will be ignored or simplified?
    - Where can you get additional data, if needed?
- Discuss the document with key people

Program the model

- An issue: code the model in a programming language, or use a simulation package?
  - Programming language may run faster, and marginal software cost usually low, but it may be buggy and take longer to develop
  - Simulation package may produce many fewer errors, lower total cost, but less flexible
- Pilot runs and debugging (verify and validate)
  - Compare results with real system's output, if real system exists
  - Even if it doesn't, get opinions from Subject Matter Experts (SME) on the model
  - Important not only to be correct (valid), but also to look correct to influential people (credible)

Analyze data and report results

- Design experiments that will answer the desired questions
  - Determine how these will run: replications, warm-up length, length of each run
- Statistical analysis is essential
  - Run the model as required within these policies
- Document, present, and use the model and its results
  - People often underestimate importance of credibility
    - Decision-makers often aren't technically competent
    - They won't act unless they feel that the model is reasonable
  - Important to keep them informed

Lecture 2: Probability Review

END 322E System Simulation

Mehmet Ali Ergün, Ph.D.

Sample Space

- Elementary events: elements of some set Ω
  - Think of it as set of possible outcomes of an experiment
  - E.g., $\{H, T\}$ for flipping a coin once
  - Elementary events are H and T
- This set Ω is called the sample space
- One more example:
  - Suppose we throw 3 dice in sequence
  - What are the possible outcomes? (This defines Ω)
  - How many possible outcomes are there?

Events

- Events are subsets of the sample space Ω of elementary events
- The set E of possible events must be a σ-field, i.e.:
  - Entire space Ω is an event
  - Complement of any event is an event
  - Union and intersection of two events are events
  - Union of a countable number of events is an event

- Example event: all outcomes in throwing of 3 dice in which the total shown is at least 17

Probability

- To each event $e \epsilon E$, we assign a real number called its probability, P(e)
- Axioms of probability:
  - For each e, $0 \leq P(e) \leq 1$
  - $P(\Omega) = 1, P(\emptyset) = 0$
  - For countable sequence of mutually exclusive events e_i , we have:
    - $P(\bigcup_i e_i) = \sum_i P(e_i)$
- Examples:
  - Cast of a single die: P(each face)=1/6
  - Same die, different events: $\emptyset, \{1,3,5\}, \{2,4,6\}, \{1,2,3,4,5,6\} \rightarrow$?

Conditional Probability

- Consider an experiment that consists of flipping a coin twice
$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$$
- Flipping a coin twice: $P\big((H,H)\big) = {}^{1}/_{4}$
- If the first flip H, then what is the probability of (H,H) ?
- Let A be the event of both flips are H and B the event first flip is H then the probability of A conditional on B is denoted by:
$$P(A|B) = {}^{1}/_{2}$$

Conditional Probability and Independence

- In general: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- In the previous example: $P(A) \neq P(A|B)$ because event A and B are not independent
  - In order for A to happen B must happen
- Events C and B are independent if following holds:

$$P(C \cap B) = P(C)P(B) \ \rightarrow P(C|B) = P(C)$$

Random Variables (RV)

- When we perform experiments, we are sometimes primarily concerned about a numerical quantity.
  - Essentially a function defined on sample space Ω
  - i.e. Sum of the 3 dice tosses
- We call such quantities random variables
- We need random variables for simulations, e.g.
  - Inter-arrival times for costumers/orders/vehicles
  - Service times for customers
  - Repair times for machines
- Assume their distributions, then generate them on demand

Discrete RVs

- A random variable that can take either a finite, or at most, a countable number of possible values is said to be a discrete random variable.
  - X = Number of customers arriving to a shop in a day
- Probability mass function, p(x)$\geq$ 0 is

$$p(x) = P(X = x)$$

- Since X must take on one value:

$$\sum_{0}^{\infty} p(x) = 1$$

- Cumulative density function

$$P(X \leq x) = F(x) = \sum_{0}^{x} p(x)$$

Discrete RVs (Example)

- The Industrial Engineering Department has a lab with six computers reserved for its students.
- Let X denote the number of these computers that are in use at a particular time of day.
- Suppose that the probability distribution of X is as given in the following table; the first row of the table lists the possible X values and the second row gives the probability of each such value.

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Discrete RVs (Example)

- We can now use elementary probability properties to calculate other probabilities of interest. For example, the probability that at most 2 computers are in use $P(X \leq 2)$

$$P(X = 0 \text{ or } 1 \text{ or } 2) = P(X \leq 2) = F(2)$$
$$= p(0) + p(1) + p(2)$$
$$= .05 + .10 + .15 = .30$$

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

Discrete RVs (Example)

- The probability that at *least 3 computers are in use?*

$$P(X \geq 3) = 1 - F(2) = 1 - P(X \leq 2)$$
$$= 1 - .30$$
$$= .70$$

- The probability that between 2 and 5 computers inclusive are in use?

$$P(2 \leq X \leq 5) = P(X = 2, 3, 4, or\ 5)$$
$$= .15 + .25 + .20 + .15 = 0.75$$

Continuous RVs

- Continuous Random Variables
  - X = Time waited at a bus station
  - Range of X = $[0,\infty)$
  - Probability density function, $f(x) \geq 0$ is

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

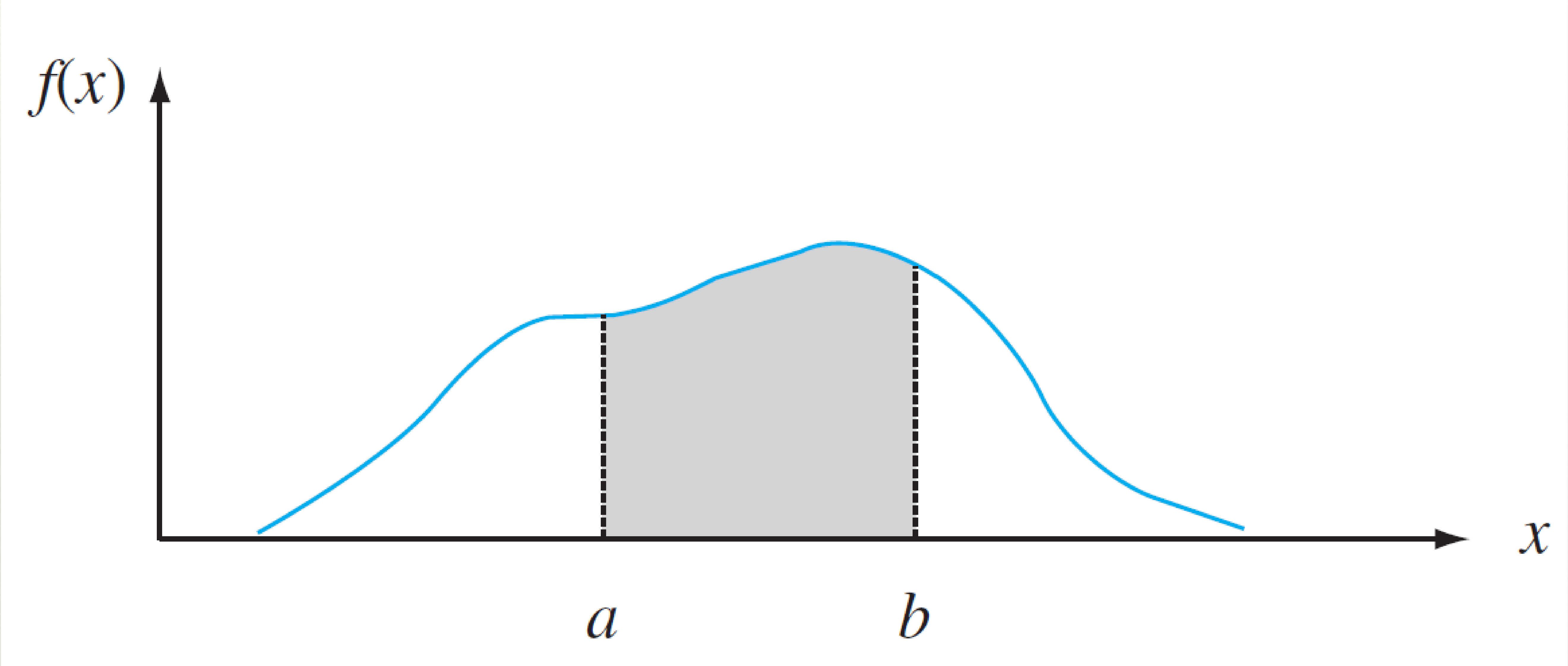
  - $\int_{-\infty}^{\infty} f(x)dx = 1$

  - Cumulative density function

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

Continuous RVs

- The probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function



- The graph of $f(x)$ is often referred to as the *density curve*.
- $P(X = a) = 0$ for any constant a

Random Variables (Example)

- The life of a laser-ray device used to inspect cracks in aircraft wings is a continuous random variable with a pdf given by

$$f(x) = \begin{cases} \frac{1}{2} e^{-x/2}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

- What is the probability that the life of the laser device is between 2 and 3 years?

$$P(2 \leq X \leq 3) = \frac{1}{2} \int_{2}^{3} e^{-x/2}\, dx = -e^{-\frac{3}{2}} + e^{-1} = 0.145$$

Expectation and Variance

- The expectation or the mean is a measure of the central tendency of a random variable. E(X) is given by if
  - Discrete rv

$$E(X) = \sum_{all\ x} x\, P(X = x)$$

  - Continuous rv

$$E(X) = \int_{-\infty}^{\infty} x\, f(x) dx$$

- The variance V(X) is a measure of spread of the random variable around the mean E(X) given by

$$\mathrm{V}(X) = \mathrm{E}[X^2] - E[X]^2$$

Expectation (Examples)

- The mean number of computers occupied in IE department at a given time:

$E(X) = \sum x\, P(X = x)$

$= 0 \times 0.05 + 1 \times 0.1 + 2 \times 0.15 + 3 \times 0.25 + 4 \times 0.20 + 5 \times 0.15 + 6 \times 0.1$

= 3.3 computers per day

- The mean life years of laser ray device is

$$E(X) = \frac{1}{2}\int_0^\infty x\, e^{-\frac{x}{2}} dx = xe^{-\frac{x}{2}}|_0^\infty + \int_0^\infty e^{-\frac{x}{2}} dx$$
$$= 0 + \frac{1}{1/2} e^{-\frac{x}{2}}|_0^\infty$$
$$= 2\ years$$

Bernoulli Distribution

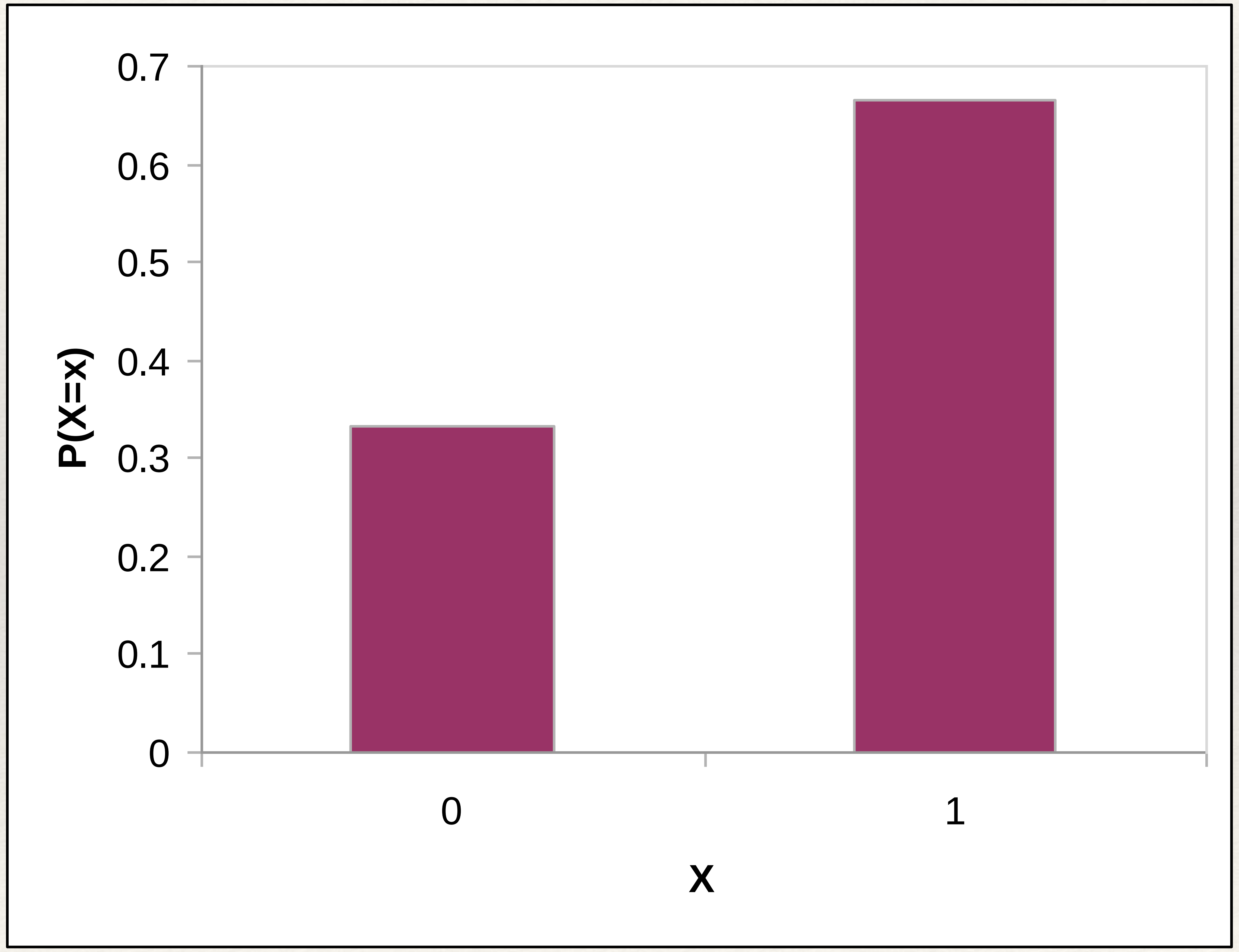
- The simplest form of random variable.
  - Success/Failure
  - Flip coin Heads/Tails

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$E[X] = p$$

$$Var(X) = p(1 - p)$$



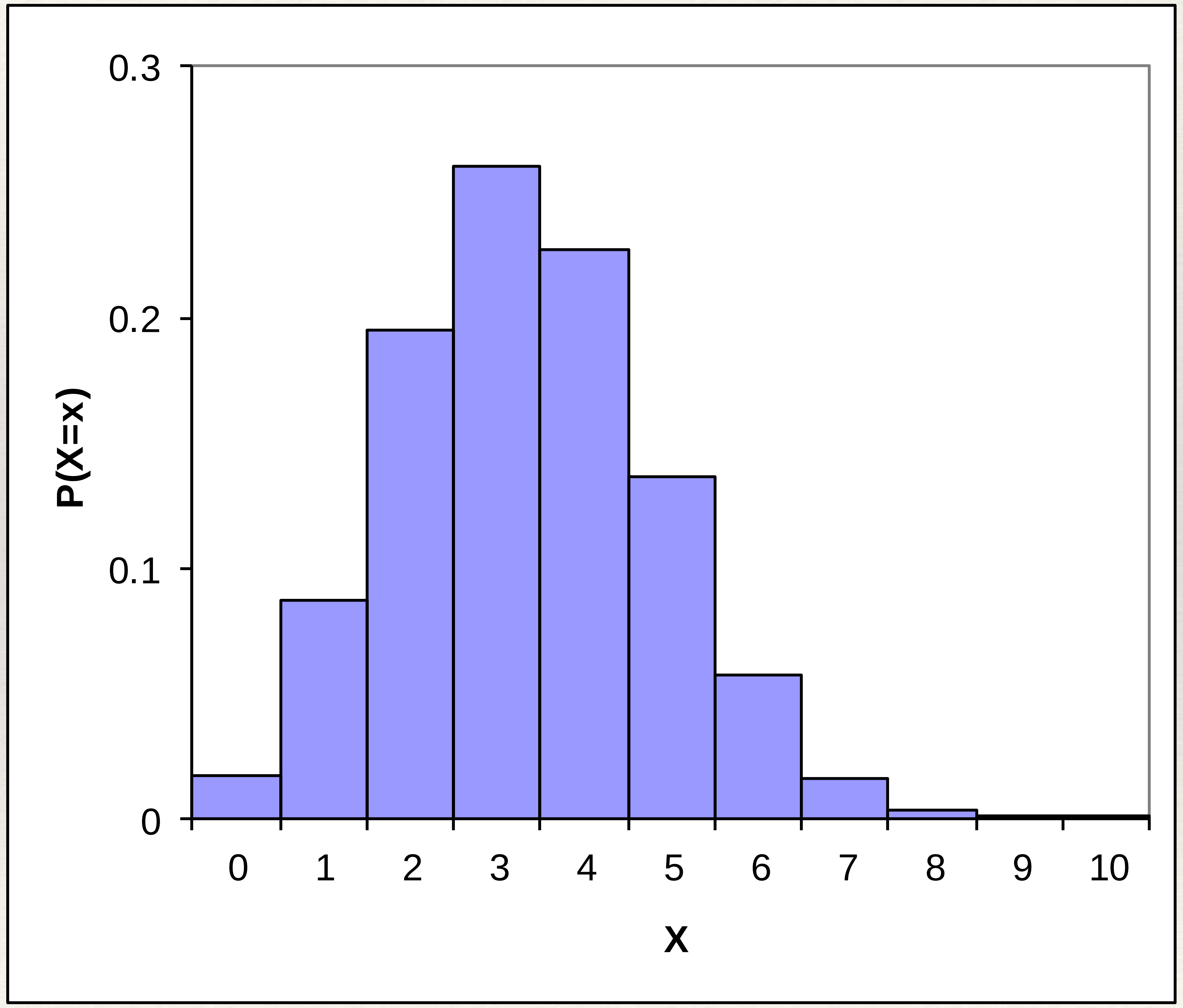
Binomial Distribution

- The number of successes in *n* Bernoulli trials.
  - Or the sum of n Bernoulli random variables.
  - Number of heads/tails out of n flips

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

$$E[X] = np$$

$$Var(X) = np(1-p)$$



Binomial Example

- An airline knows that 5 percent of the people making reservations on a certain flight will not show up. Consequently their policy is to sell 52 tickets for a flight that can hold only 50 passengers.
- Then $p = P$(a certain passenger to show up) = .95,
  X = The number of passengers show up for the flight,
  $X \sim \mathrm{Bin}(52,.95)$.
- What is the probability that there will be a seat for every passenger who shows up?

$$P(X \le 50) = P(X=0) + P(X=1) + \cdots + P(X=49) + P(X=50)$$
$$= \binom{52}{0} p^0 (1-p)^{52} + \binom{52}{1} p^1 (1-p)^{51} + \cdots$$

Binomial Example

- But it is too much calculation. Solution: use axioms of probability:

$$P(X \leq 50) = 1 - P(X > 50)$$

$$= 1 \; - (P(X = 51) + P(X = 52))$$

$$= 1 - \; \binom{52}{51} p^{51}(1-p)^{1} + \binom{52}{52} p^{52}(1-p)^{0}$$

$$= 1 \; - \; 52(0.95)^{51}(0.05)^{1} + (0.95)^{52}$$

$$= 0.74$$

Geometric Distribution

- The number of Bernoulli trials required to get the first success.

$$P(X = x) = p(1-p)^{x-1}$$

$$E[X] = \frac{1}{p}$$

$$Var(X) = \frac{(1-p)}{p^2}$$

Geometric Distribution - Example

- Starting at a fixed time, we observe the gender of each newborn child at a certain hospital until a boy (B) is born. Let , assume that successive births are independent, and define the RV X by of births observed until the first boy is born.

- Then also assume $p = P(\text{a newborn is Boy}) = .49$, so
  X = the number births until the first Boy,
  $X \sim \text{Geom}(.49)$.

- The probability that 3 births are observed until the first boy?

$$P(X = 3) = (.51)^2(.49)^1 = .127$$

- Expected number of births until a boy is born

$$E[X] = \frac{1}{0.49} = 2.04$$

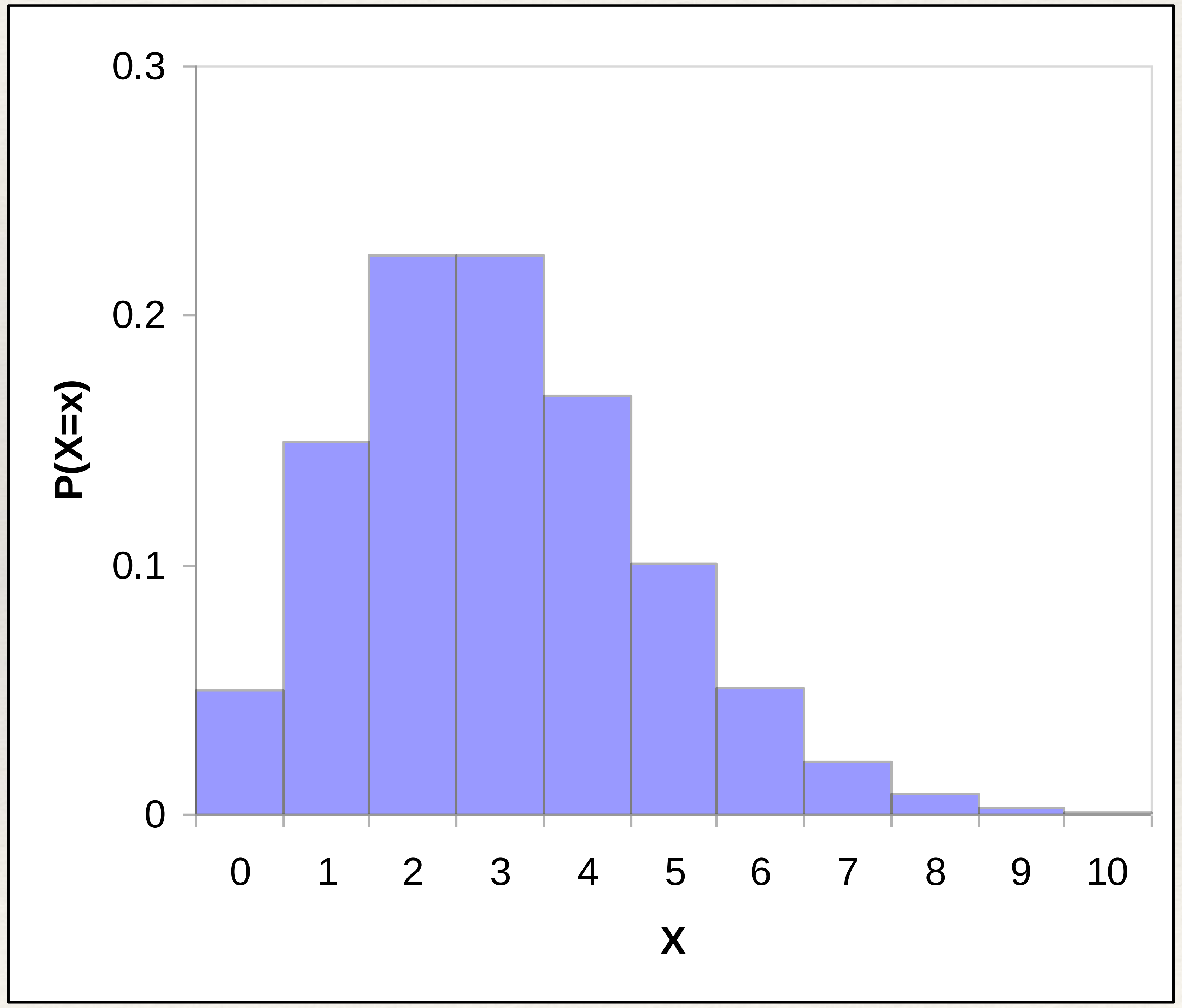
Poisson Distribution

- The number of random events occurring in a fixed interval of time
  - Random batch sizes
  - Number of defects on an area of material

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

$$E[X] = \mu$$

$$\mathrm{Var}[X] = \mu$$



Poisson Example

- Let X denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that X has a Poisson distribution with, so on average traps will contain 4.5 creatures.

- The probability that a trap contains exactly five creatures is

$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = .1708$$

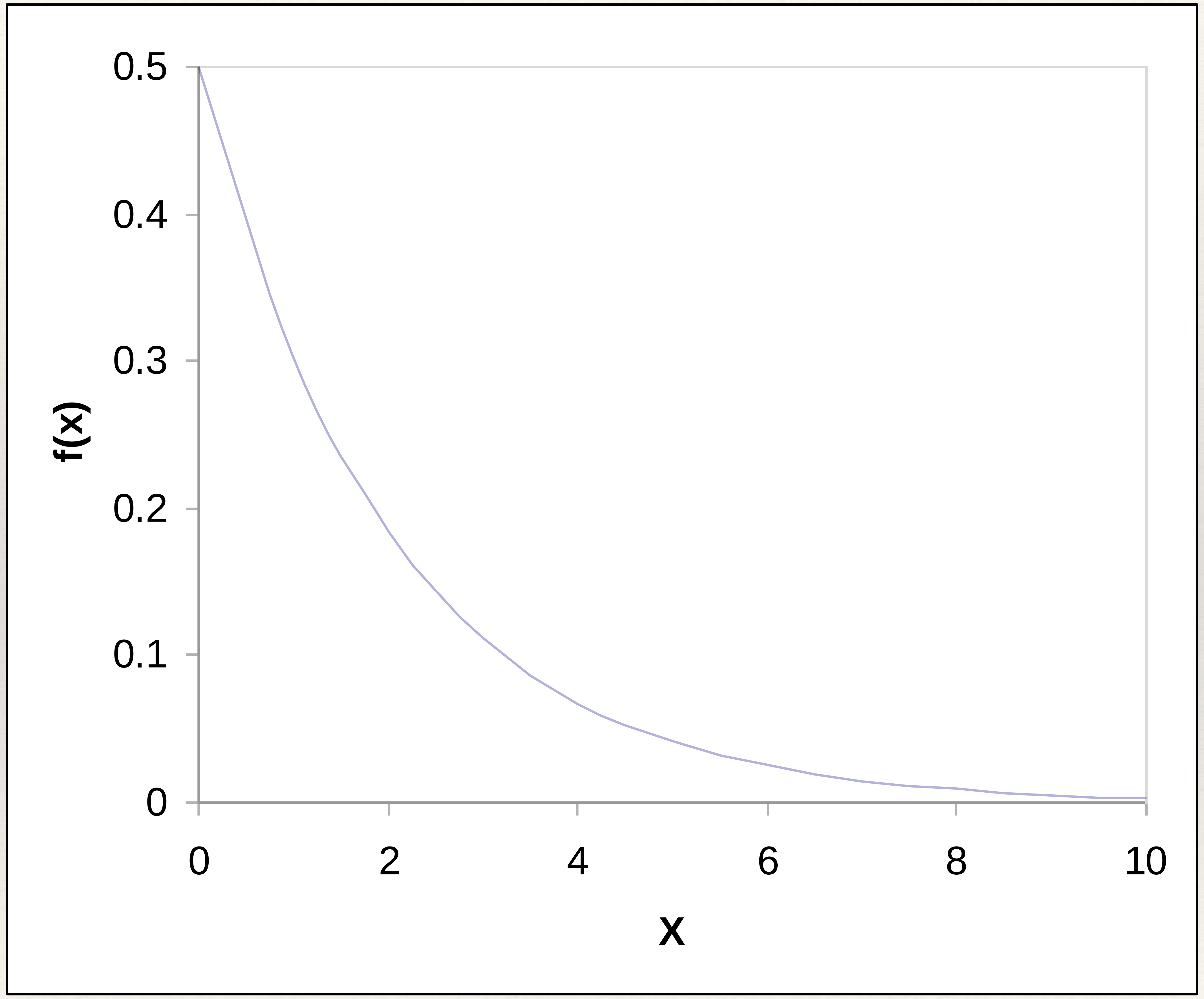
- The probability that a trap has at most five creatures is

$$P(X \leq 5) = \sum_{x=0}^{5} \frac{e^{-4.5}4.5^x}{x!} = 0.7029$$

Exponential Distribution

- Model times between events
  - Times between arrivals, failures
  - Times to repair, Service Times
  - Probability Density: $f(x) = \lambda e^{-\lambda x}$
  - Cumulative Density: $P(X \leq x) = F(x) = 1 - e^{-\lambda x}$
  - $E[X] = \frac{1}{\lambda}$ $\qquad Var[X] = \frac{1}{\lambda^2}$
- Memoryless

$$f(X > x + y | X > y) = F(x)$$



Exponential Example

- Suppose that calls are received at a 24-hour "suicide hotline" according to a Poisson process with rate $\lambda = 0.5$ call per day.
- Then the number of days X between successive calls has an exponential distribution with parameter value 0.5, so the probability that more than 2 days elapse between calls is

$$
\begin{aligned}
P(X > 2) &= 1 - P(X \le 2) \\
&= 1 - \left(1 - e^{-(.5)(2)}\right) \\
&= e^{-(.5)(2)} = 0.368
\end{aligned}
$$

- The expected time between calls is

$$E[X] = \frac{1}{0.5} = 2$$

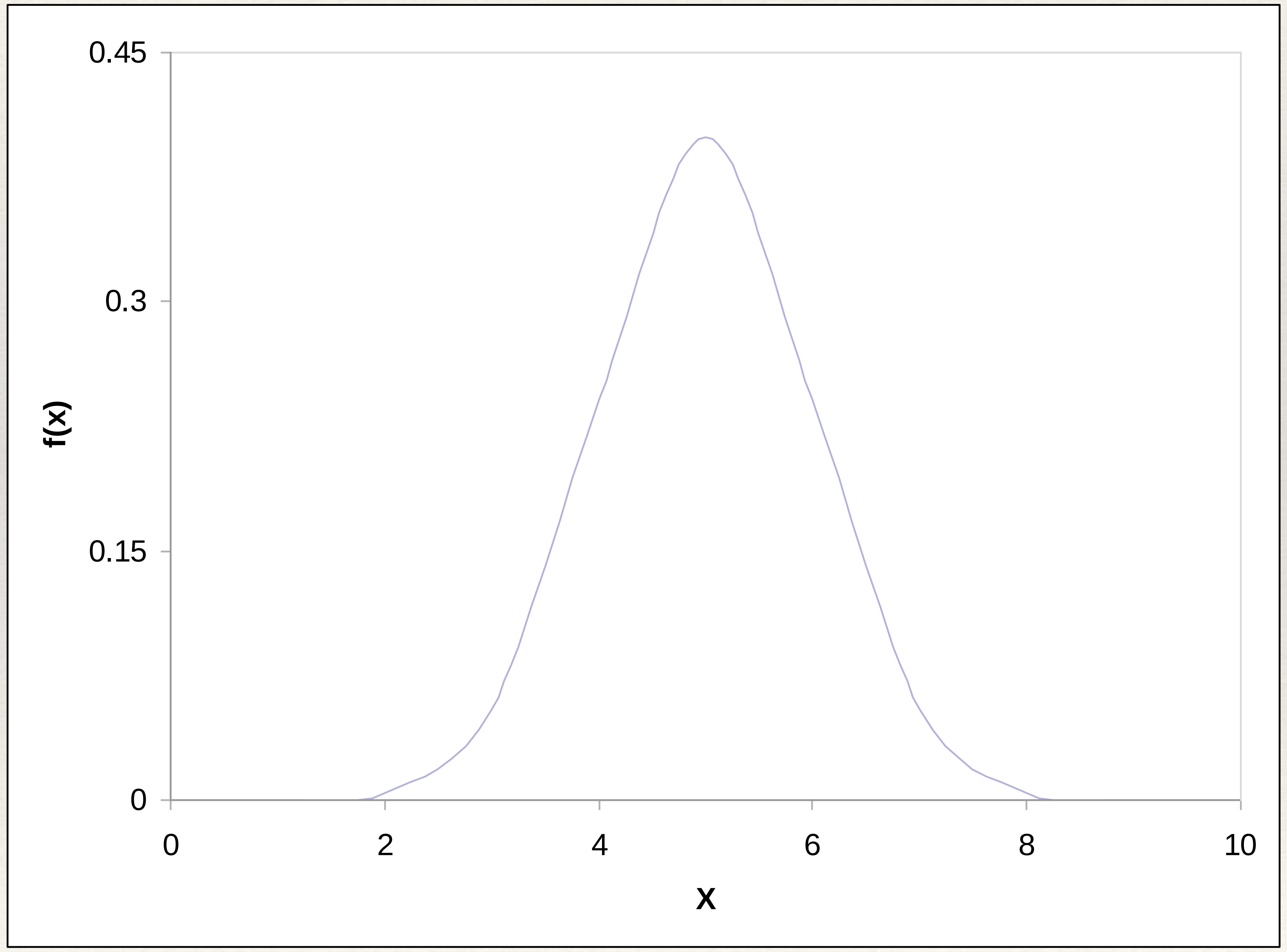
Normal Distribution

- The distribution of the average of iid random variables are eventually normal
  - Distribution of heights $X \sim N(\mu, \sigma)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$



- Left Tail distribution

$$P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \varphi(\frac{x-\mu}{\sigma})$$

Check φ from standard normal table

- Central Limit Theorem

$X_1, X_2, \ldots, X_n$independent and identically distributed random variables with mean μ and variance σ, and

$$\bar{X} = \frac{X_1 + X_2 + \ \ldots + X_n}{n}$$

Then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \ as\ n \to \infty$$

Normal Distribution Example

- The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions.
- The article “Fast-Rise Brake Lamp as a Collision-Prevention Device” (*Ergonomics*, 1993: 391–395) suggests
that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 sec and standard
deviation of .46 sec.

Normal Distribution Example

- What is the probability that reaction time is between 1.00 sec and 1.75 sec? If we let X denote reaction time, then standardizing gives

$$1.00 \leq X \leq 1.75$$

Thus normalizing

$$\frac{1.00 - 1.25}{.46} \leq \frac{X - 1.25}{.46} \leq \frac{1.75 - 1.25}{.46}$$

$$P(1.00 \leq X \leq 1.75) = P\left(\frac{1.00 - 1.25}{.46} \leq Z \leq \frac{1.75 - 1.25}{.46}\right)$$

$$= P(-.54 \leq Z \leq 1.09) = \varphi(1.09) - \varphi(-.54)$$
$$= .8621 - .2946 = .5675$$

Triangular Distribution

- Used in situations were there is little or no data.
  - Just requires the minimum (a), maximum(b) and most likely (m)s value.

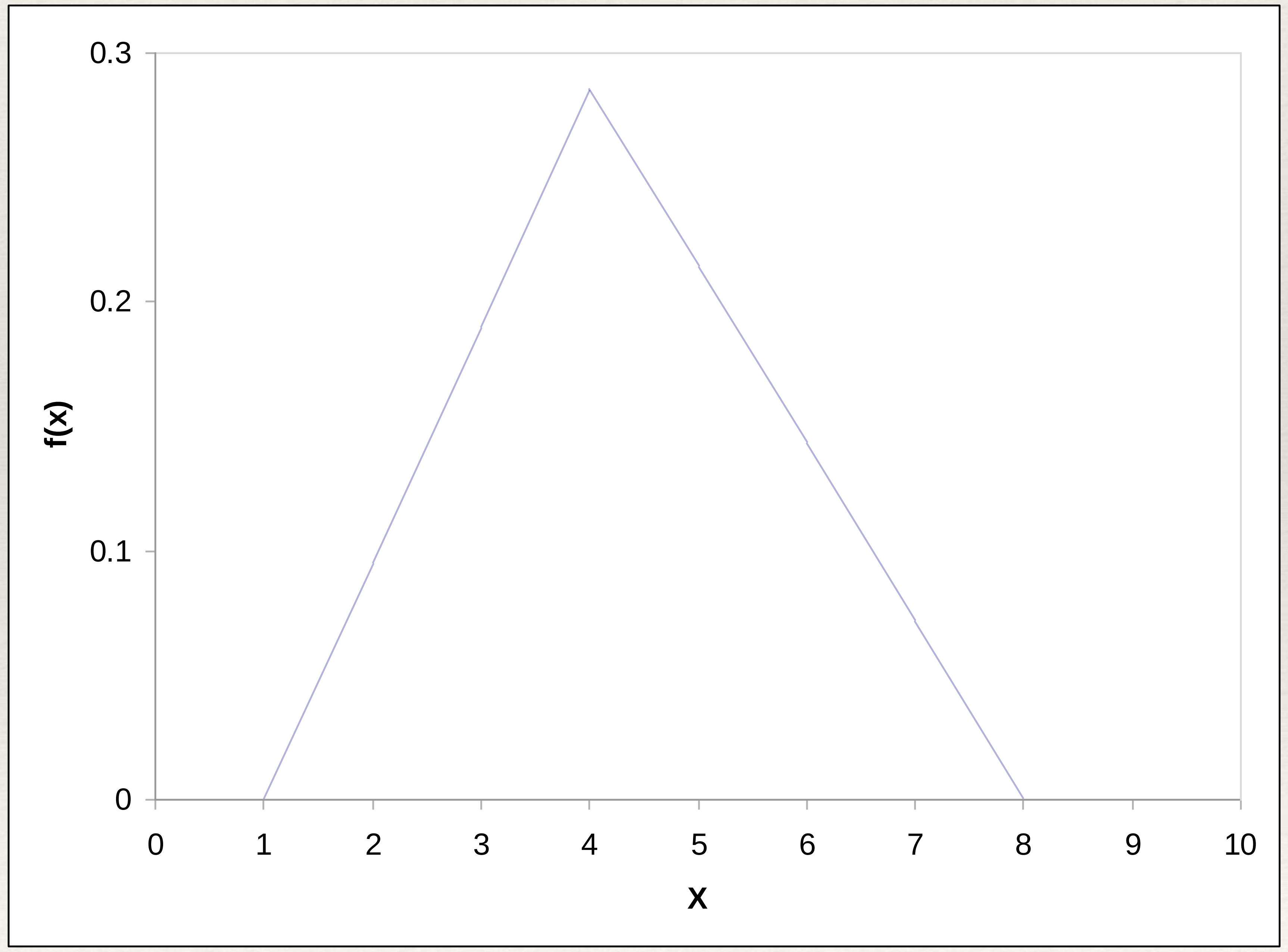
$$f(x) = \frac{2(x-a)}{(m-a)(b-a)}, \quad a \le x < m$$

$$= \frac{2(b-x)}{(b-m)(b-a)}, \quad m \le x < b$$

$$= 0, \quad \text{otherwise}$$

$$E[X] = (a+b)/2$$

$$Var(X) = (b-a)^2/12$$



Lecture 3: Queuing Theory and Poisson Processes

END 322E System Simulation

Mehmet Ali Ergün, Ph.D.

What is Queueing Theory

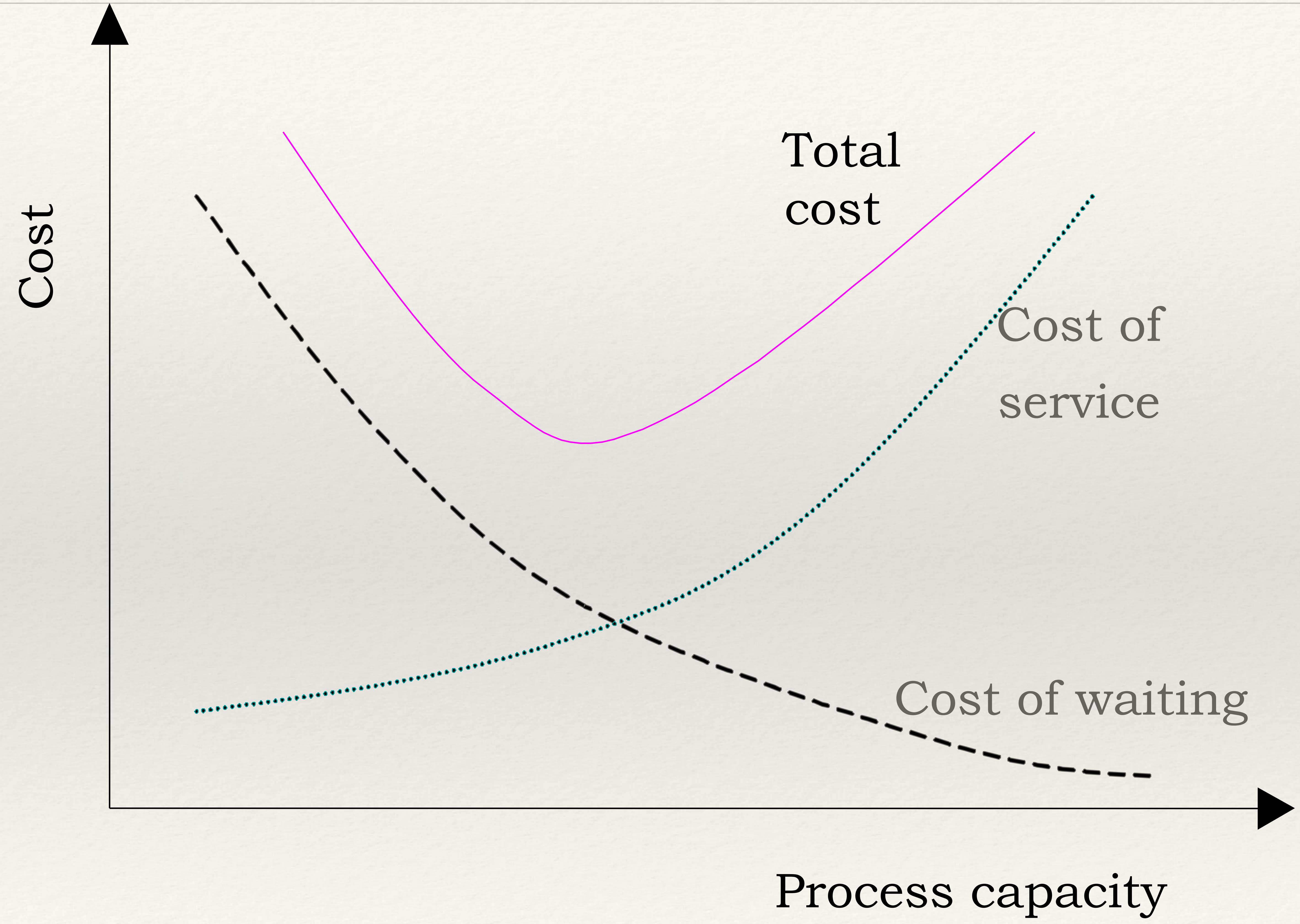
- Mathematical analysis of queues and waiting times in stochastic systems.
  - Used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times.
- Queues arise when the short term demand for service exceeds the capacity
  - Most often caused by random variation in service times and the times between customer arrivals.
  - If long term demand for service > capacity the queue will explode!

Why is Queuing Analysis Important?

- Capacity problems are very common in industry and one of the main drivers of process redesign
  - Need to balance the cost of increased capacity against the gains of increased productivity and service
- Queuing and waiting time analysis is particularly important in service systems
  - Large costs of waiting and of lost sales due to waiting

Prototype Example – ER at a Hospital
Patients arrive by ambulance or by their own accord
One doctor is always on duty
More and more patients seeks help ⇒ longer waiting times
➢ ***Question*: Should another MD position be instated?**

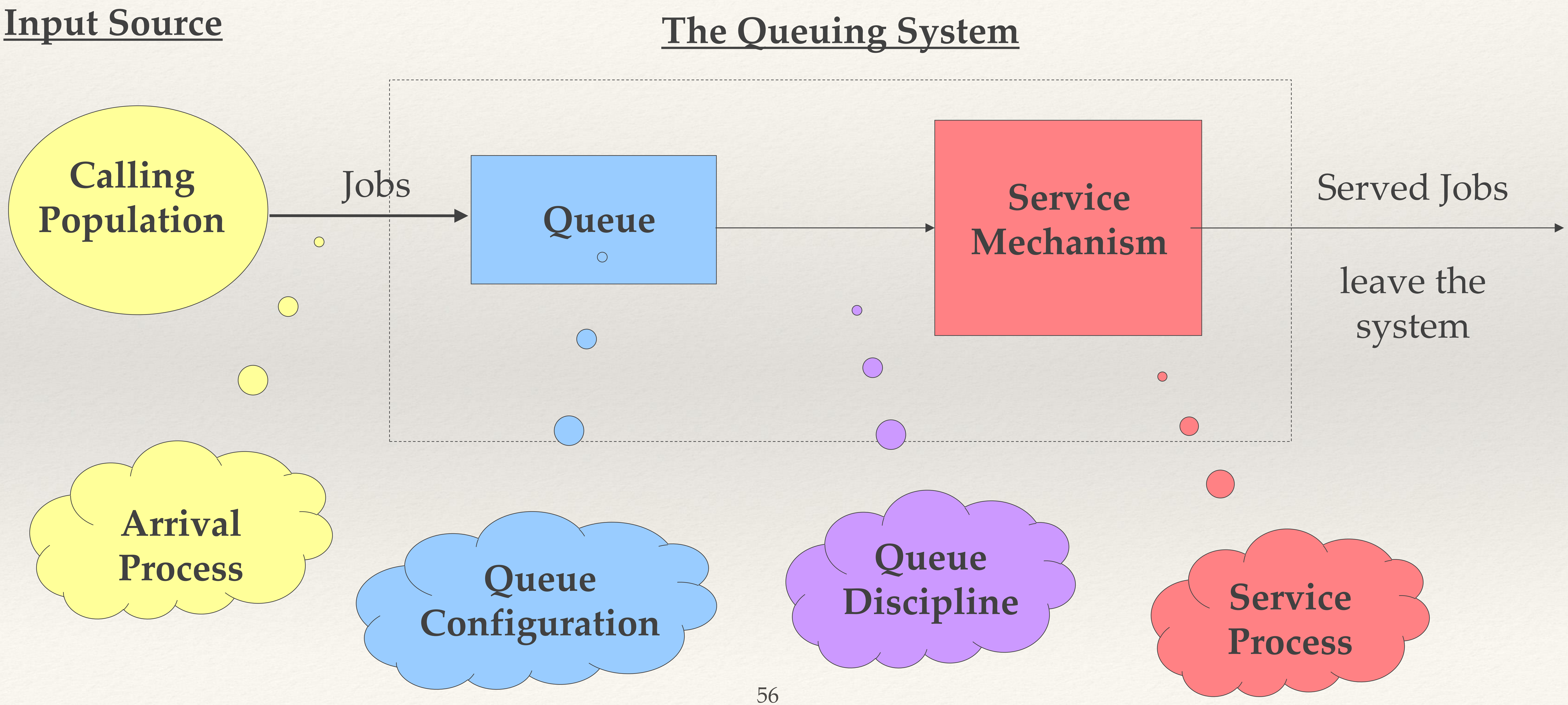
Cost/Capacity Trade-off



Examples of Queueing Systems

- Commercial Queueing Systems
  - Commercial organizations serving external customers
  - Ex. Dentist, bank, ATM, gas stations, plumber, garage …
- Transportation service systems
  - Vehicles are customers or servers
  - Ex. Vehicles waiting at toll stations and traffic lights, trucks or ships waiting to be loaded, taxi cabs, fire engines, elevators, buses …
- Business-internal service systems
  - Customers receiving service are internal to the organization providing the service
  - Ex. Inspection stations, conveyor belts, computer support …
- Social service systems
  - Ex. Judicial process, the ER at a hospital, waiting lists for organ transplants or student dorm rooms …

Components of a Basic Queueing System



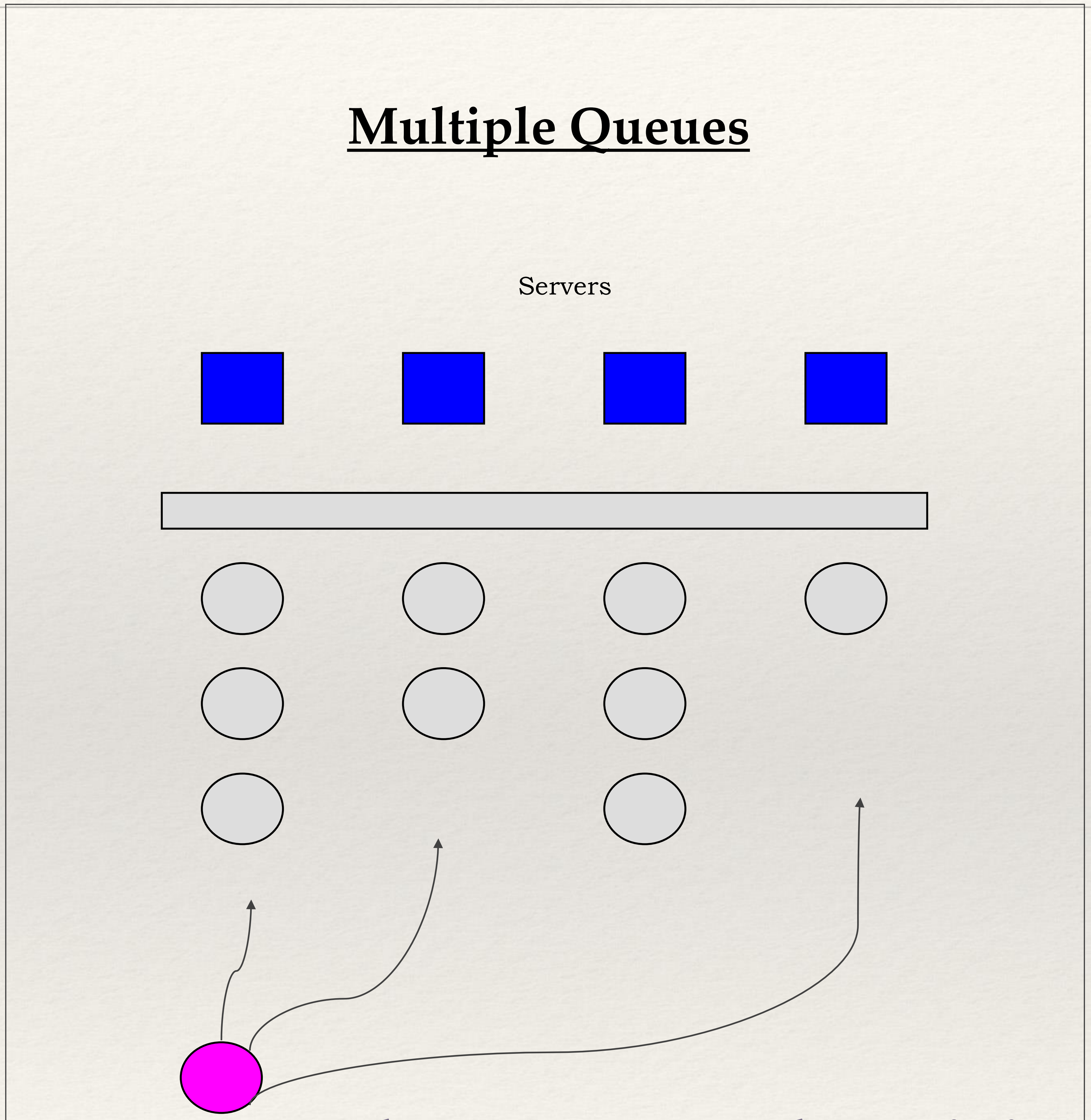
Components of a Basic Queueing System

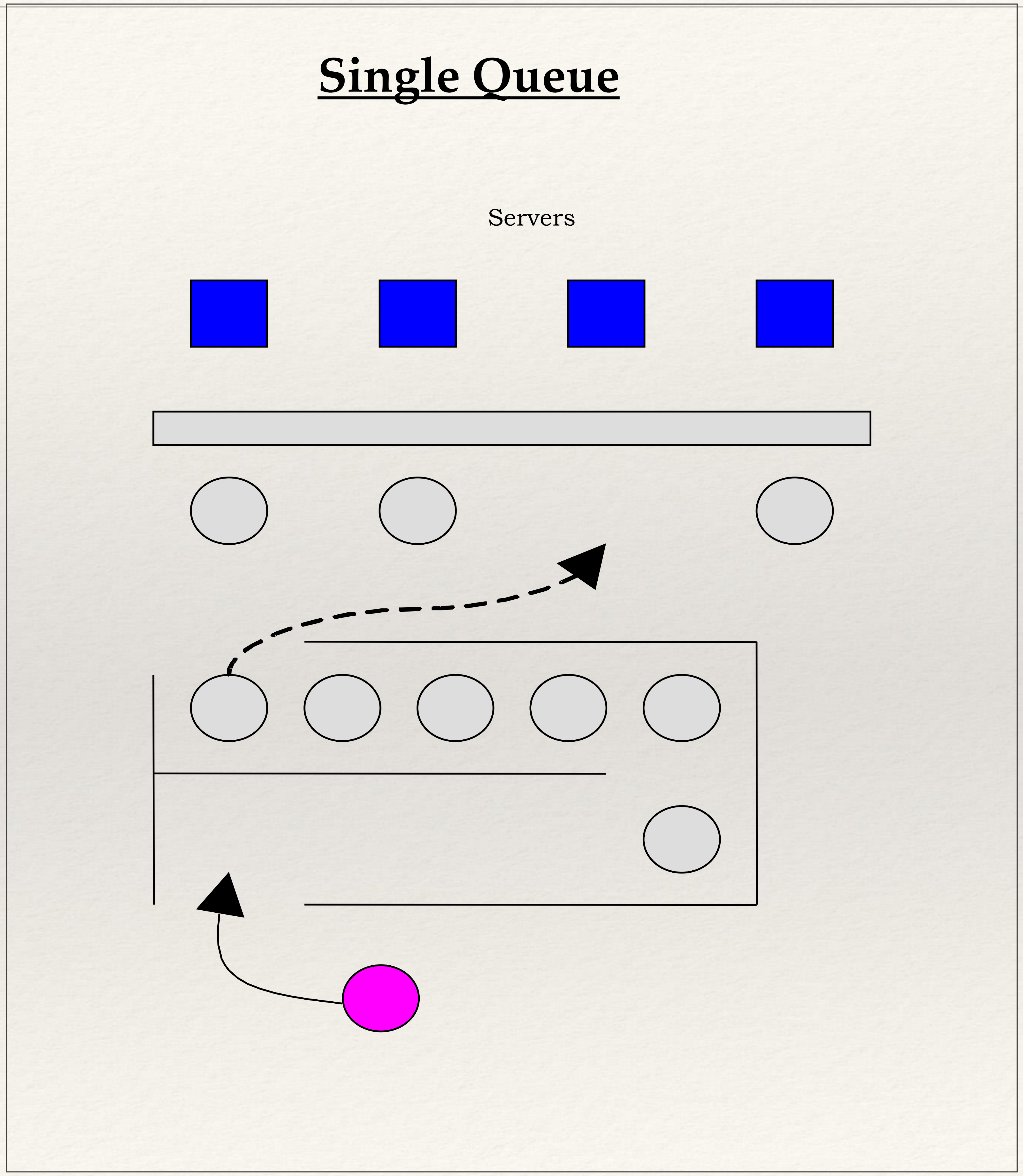
- The calling population
  - The population from which customers/jobs originate
  - The size can be finite or infinite (the latter is most common)
  - Can be homogeneous (only one type of customers/ jobs) or heterogeneous (several different kinds of customers/jobs)
- The Arrival Process
  - Determines how, when and where customer/jobs arrive to the system
  - Important characteristic is the customers'/jobs' inter-arrival times
  - To correctly specify the arrival process requires data collection of interarrival times and statistical analysis.

Components of a Basic Queueing System

- The queue configuration
  - Specifies the number of queues
    - Single or multiple lines to a number of service stations
  - Their location
  - Their effect on customer behavior
    - Balking and reneging
  - Their maximum size (# of jobs the queue can hold)
    - Distinction between infinite and finite capacity

Two different Queue Configurations





The queue you have joined is always the slowest ☺

Multiple v.s. Single Customer Queue Configuration

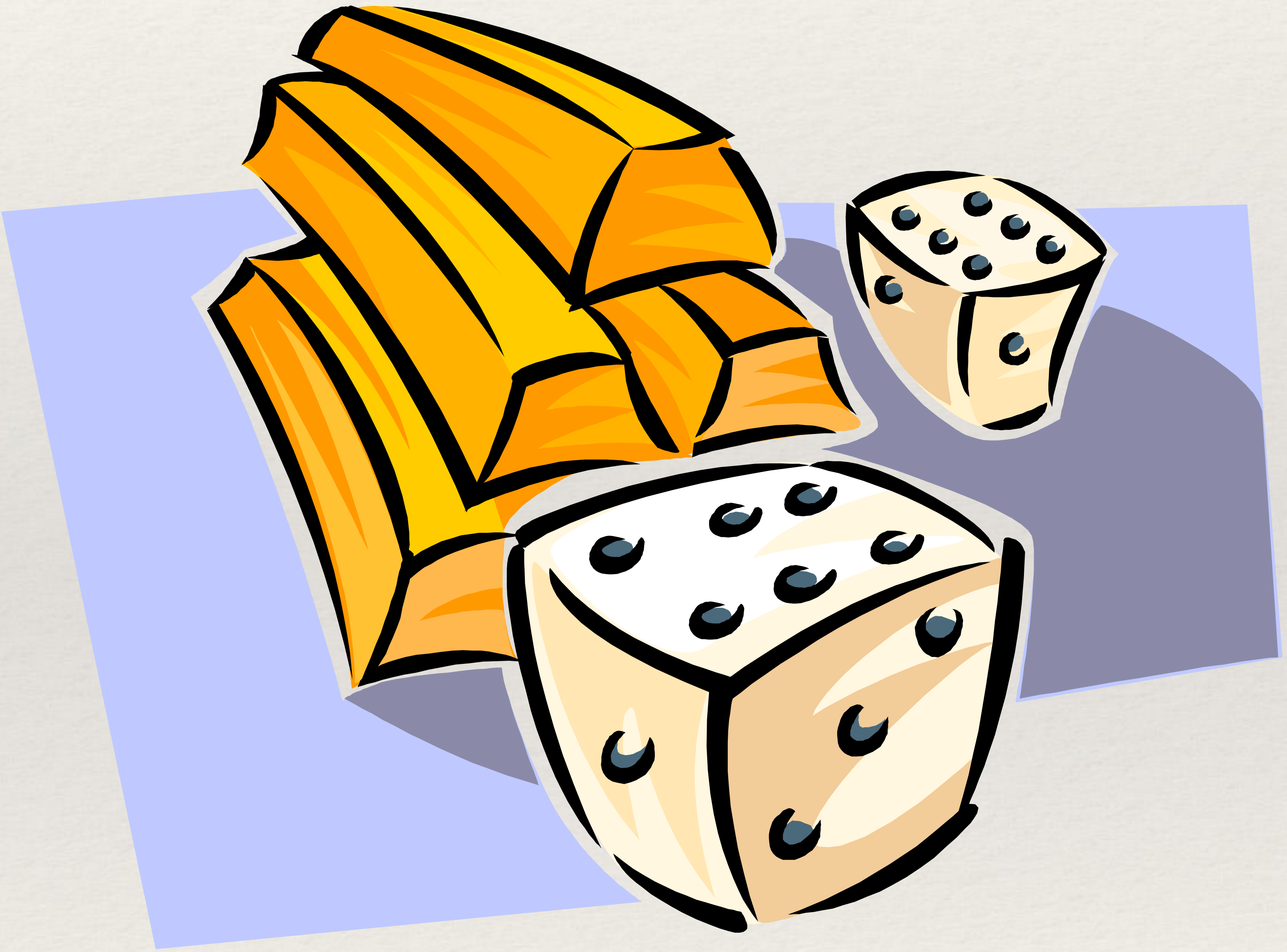
Multiple Line Advantages	Single Line Advantages
❖ **The service provided can be differentiated** ❖ Ex. Supermarket express lanes ❖ **Labor specialization possible** ❖ **Customer has more flexibility** ❖ **Balking behavior may be deterred** ❖ Several medium-length lines are less intimidating than one very long line	❖ **Guarantees fairness** ❖ FIFO applied to all arrivals ❖ **No customer anxiety regarding choice of queue** ❖ **Avoids "cutting in" problems** ❖ **The most efficient set up for minimizing time in the queue** ❖ **Jockeying (line switching) is avoided**

Components of a Basic Queueing System

- The Service Mechanism
  - Can involve one or several service facilities with one or several parallel service channels (**servers**) - Specification is required
  - The service provided by a server is characterized by its service time
    - Specification is required and typically involves data gathering and statistical analysis.
    - Most analytical queuing models are based on the assumption of exponentially distributed service times, with some generalizations.
- The queue discipline
  - Specifies the order by which jobs in the queue are being served.
  - Most commonly used principle is FIFO.
  - Other rules are, for example, LIFO, SPT, EDD…
  - Can entail prioritization based on customer type.

A Commonly Seen Queueing Model

- Service times as well as interarrival times are assumed independent and identically distributed
  - If not otherwise specified
- Commonly used notation principle: A/B/C/Capacity/Pop.size
  - A = The interarrival time distribution
  - B = The service time distribution
  - C = The number of parallel servers
- Commonly used distributions
  - M = Markovian (exponential) - *Memoryless*
  - D = Deterministic distribution
  - G = General distribution
- Example: M/M/c
  - Queuing system with exponentially distributed service and inter-arrival times and c servers



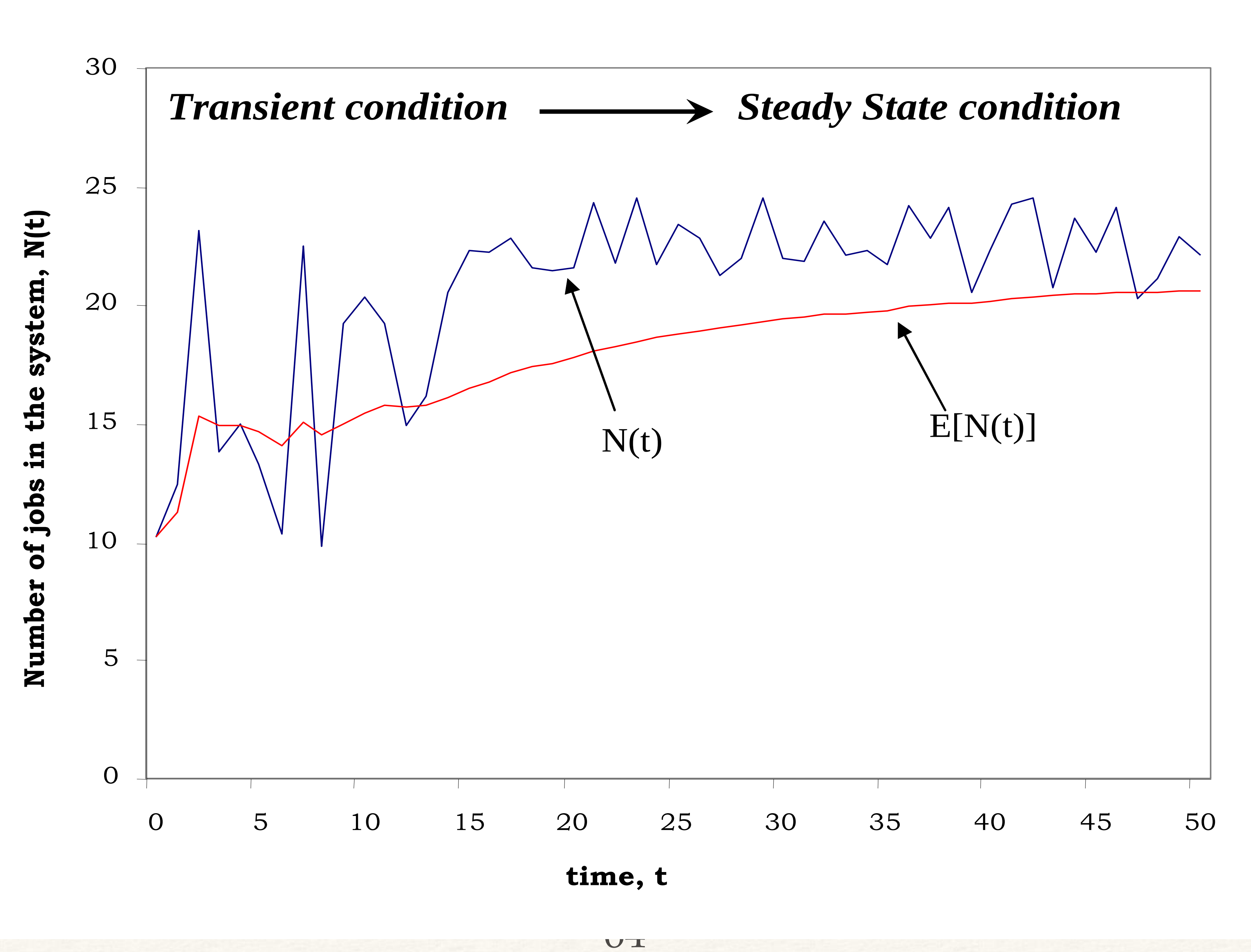
Steady State Analysis of Queues

- Steady State condition
  - Enough time has passed for the system state to be independent of the initial state as well as the elapsed time
  - The probability distribution of the state of the system remains the same over time (is stationary).
- Transient condition
  - Prevalent when a queuing system has recently begun operations
  - The state of the system is greatly affected by the initial state and by the time elapsed since operations started
  - The probability distribution of the state of the system changes with time

With few exceptions Queuing Theory has focused on analyzing steady state behavior

Transient and Steady State Conditions

- Illustration of transient and steady-state conditions
  - N(t) = number of customers in the system at time t,
  - E[N(t)] = represents the expected number of customers in the system.



Notation for Steady State Analysis

P_n= The probability that there are exactly n customers/jobs in the system (in steady state, i.e., when $t \to \infty$)

L= Expected number of customers in the system (in steady state)

L_q= Expected number of customers in the queue (in steady state)

W= Expected time a job spends in the system

W_q = Expected time a job spends in the queue

ρ = The utilization factor for the service facility. (= The expected fraction of the time that the service facility is being used)

Little's Formula

- Assume that $\lambda_n = \lambda$ and $\mu_n = \mu$ for all n

$$L = \lambda W$$

$$L_q = \lambda W_q$$

- Holds for all queueing systems regardless of the number of servers, queue discipline, or arrival and service time distributions

The M/M/1 - Model

Assumptions - the Basic Queuing Process

- ✓ Infinite Calling Populations
  - Independence between arrivals
- ✓ The arrival process is Poisson with an expected arrival rate λ
  - Independent of the number of customers currently in the system
- ✓ The queue configuration is a single queue with possibly infinite length
  - No reneging or balking
- ✓ The queue discipline is FIFO
- ✓ The service mechanism consists of a single server with exponentially distributed service times
  - μ = expected service rate when the server is busy

The M/M/1 - Model

- Steady State condition: $\rho = (\lambda/\mu) < 1$

Needs to be satisfied, otherwise the queue length goes to infinity

$P_0 = 1-\rho$

$P_n = \rho^n(1- \rho)$

$P(n \geq k) = \rho^k$

$L=\rho/(1- \rho)$

$W=L/\lambda=1/(\mu- \lambda)$

$L_q= \rho^2/(1- \rho) = L-\rho$

$W_q=L_q/\lambda= \lambda /(\mu(\mu- \lambda))$

Queuing Example 1

- Customers arrive at a sales counter manned by **a single person** according to a Poisson process with a mean rate of 20 per hour. The time required to serve a customer has an exponential distribution with a mean of 100 seconds. Find the average waiting time of a customer.

M/M/1 Example 1 - Solution

Solution

Arrival rate = λ = 20 customers per hour

Service rate = $\mu = \dfrac{3600}{100}$ = 36 customers per hour

The average waiting time of a customer in the queue $= \dfrac{\lambda}{\mu(\mu - \lambda)} = \dfrac{20}{36(36-20)}$

$$= \frac{5}{36\times 4}\text{hours} = \frac{5}{36\times 4}\times 60\times 60 = 125\text{ seconds}$$

The average waiting time of a customer in the system

$$= \frac{1}{(\mu - \lambda)} = \frac{1}{(36-20)}$$

$$= \frac{1}{16}\text{hours} = \frac{1}{16}\times 60\times 60 = 225\text{ seconds}$$

Queueing Example 2

- A university is about to lease a super computer
- There are two alternatives available
  - The M computer which is more expensive to lease but also faster
  - The C computer which is cheaper but slower
- Processing times and times between job arrivals are exponential $\Rightarrow$ M/M/1 model
  - λ = 20 jobs per day
  - μ_M = 30 jobs per day
  - μ_C = 25 jobs per day
- The leasing and waiting costs:
  - Leasing price: C_M = \$500 per day, C_C = \$350 per day
  - The waiting cost per job and time unit job is estimated to \$50 per job and day
- Question:
  - Which computer should the university choose in order to minimize the expected costs?

Example – Computer Procurement

- Compute expected time in system for each computer type

$$\mathbf{W}= \frac{\mathbf{1}}{(\boldsymbol{\mu}-\boldsymbol{\lambda})} = \frac{\mathbf{1}}{(\mathbf{30-20})} = \mathbf{1/10} \text{ days in Machine M}$$

$$\mathbf{W}= \frac{\mathbf{1}}{(\boldsymbol{\mu}-\boldsymbol{\lambda})} = \frac{\mathbf{1}}{(\mathbf{25-20})} = \mathbf{1/5} \text{ days in Machine C}$$

- Expected Cost is
  - *For Machine M, 1/10*20*50+500= $600 per day*
  - *For Machine C, 1/5*20*50+ 350=$550 per day (Slow machine is better)*

- *What if Machine M has the capability to process 50 jobs per day?*
  - *1/30*20*50+500= $533 per day (choose machine M)*

Counting Processes

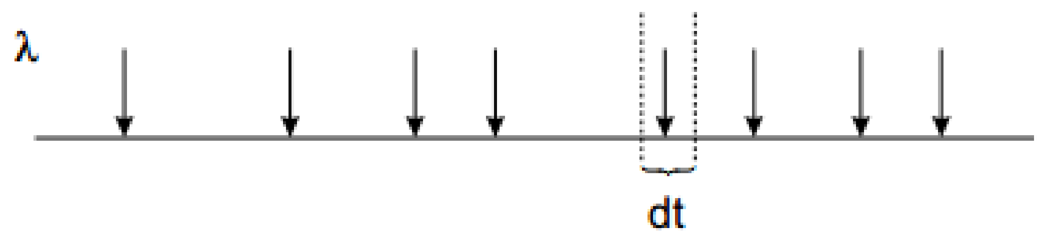
- A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if $N(t)$ represents the total number of "events" that occur by time t. Some examples of counting processes are the following:
  - If we let $N(t)$ equal the number of persons who enter a particular store at or prior to time t, then $\{N(t), t \geq 0\}$ is a counting process in which an event corresponds to a person entering the store. Note that if we had let $N(t)$ equal the number of persons in the store at time t, then $\{N(t), t \geq 0\}$ would *not* be a counting process (why not?).
  - If we say that an event occurs whenever a child is born, then $\{N(t), t \geq 0\}$ is a counting process when $N(t)$ equals the total number of people who were born by time t. (Does $N(t)$ include persons who have died by time t? Explain why it must.)
  - If $N(t)$ equals the number of goals that a given soccer player scores by time t, then $\{N(t), t \geq 0\}$ is a counting process. An event of this process will occur whenever the soccer player scores a goal.

Poisson Processes

Definition

The Poisson process can be defined in three different (but equivalent) ways:

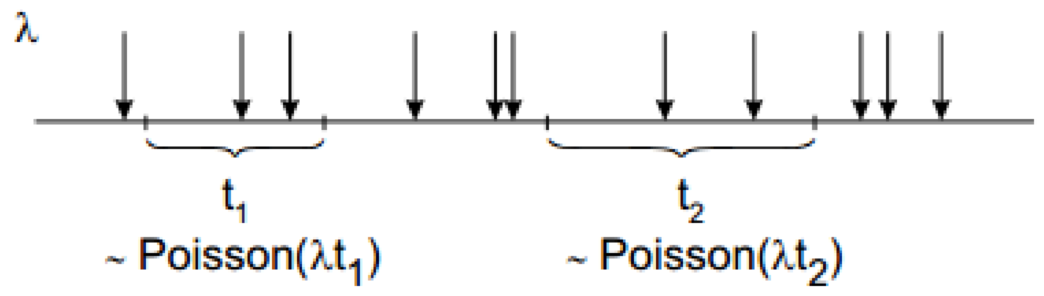
1. Poisson process is a pure birth process:
   In an infinitesimal time interval dt there may occur only one arrival. This happens with the probability λdt independent of arrivals outside the interval.



2. The number of arrivals $N(t)$ in a finite interval of length t obeys the Poisson(λt) distribution,

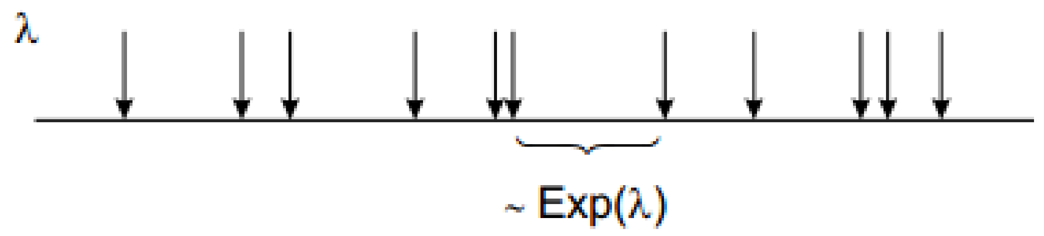
   $$\mathrm{P}\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

   Moreover, the number of arrivals $N(t_1, t_2)$ and $N(t_3, t_4)$ in non-overlapping intervals $(t_1 \le t_2 \le t_3 \le t_4)$ are independent.



3. The interarrival times are independent and obey the Exp(λ) distribution:

   $$\mathrm{P}\{\text{interarrival time} > t\} = e^{-\lambda t}$$



Poisson Process Example

- We model the arrivals of email messages at a server as a Poisson process. Suppose that on average 330 messages arrive per minute. What would you choose for the intensity λ in messages per second? What is the expectation of the interarrival time?
  - Because there are 60 seconds in a minute, we have
  - λ = (number of events) / (time unit) = 330 / 60 = 5.5
  - Since the interarrival times have an *Exp*(λ) *distribution, the*
  - expected time between messages is $1/\lambda = 0.18$ second, i.e.,
  - E(T) =1/λ = t / μ = (time unite) / (number of events) = 60/330=0.18

Properties of the Poisson Process

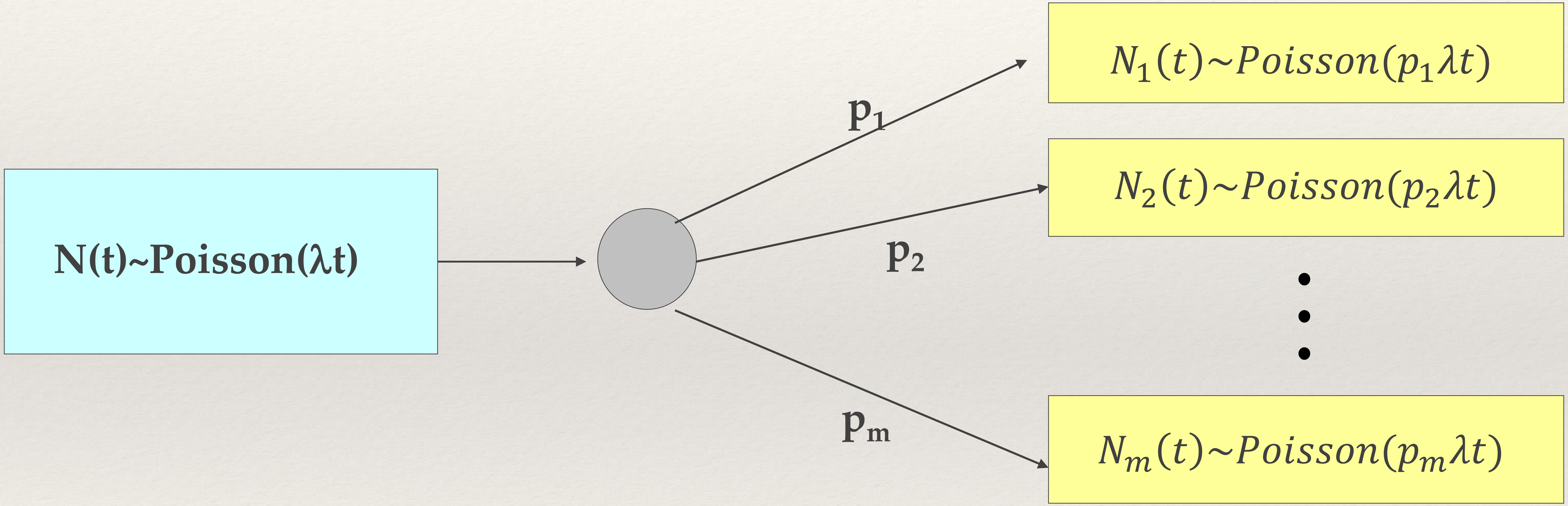
- Poisson processes can be aggregated or disaggregated and the resulting processes are also Poisson processes

a) Aggregation of m Poisson processes with intensities $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ renders a new Poisson process with intensity $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_m$.

b) Disaggregating a Poisson process $N(t) \sim \text{Poisson}(\lambda t)$ into m sub-processes $\{N_1(t), N_2(t), , \ldots, N_m(t)\}$ (for example m customer types) where $N_i(t) \sim \text{Poisson}(\lambda_i t)$ can be done if

- For every arrival we know the probability of belonging to sub-process i is p_i
- $p_1 + p_2 + \ldots + p_N = 1$, and $\lambda_i = p_i \lambda$

Disaggregating a Poisson Process



Lecture 4: Introduction to Discrete Event Simulation

END 322E System Simulation

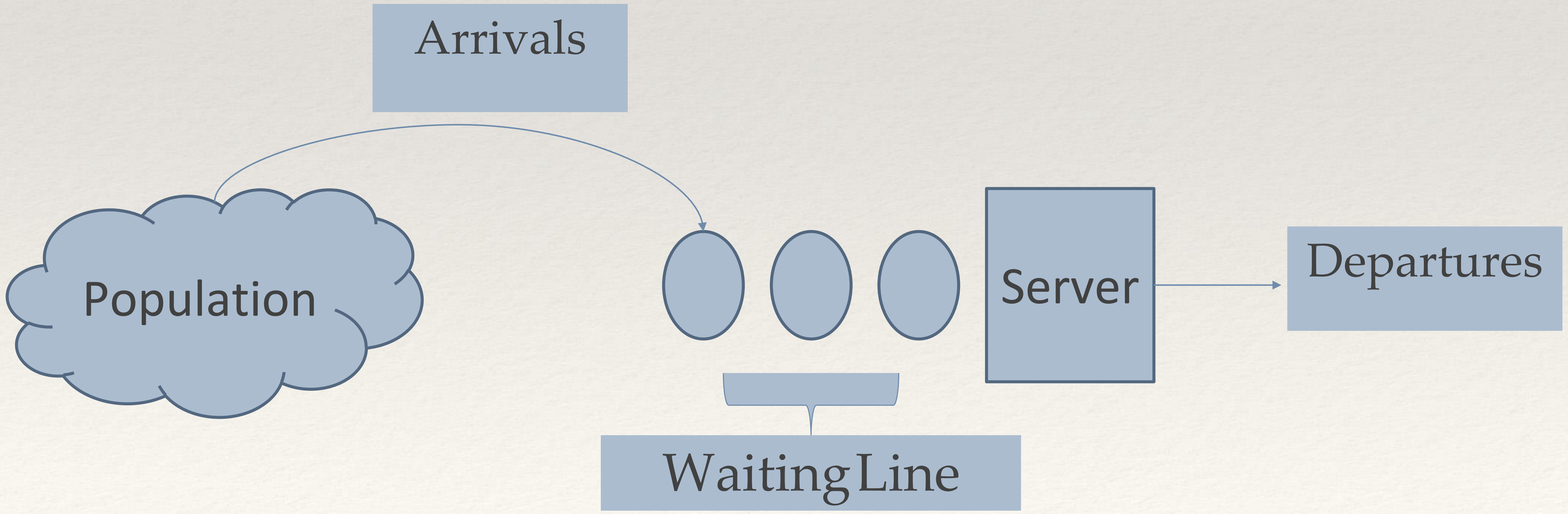
Mehmet Ali Ergün, Ph.D.

Introduction

- We will see some simulation examples that can be performed via using a simulation table
  - By hand
  - Spreadsheet
- The simulation table provides a systematic method for tracking system state over time
- Provide insights into the methodology of discrete event system simulation and descriptive statistics

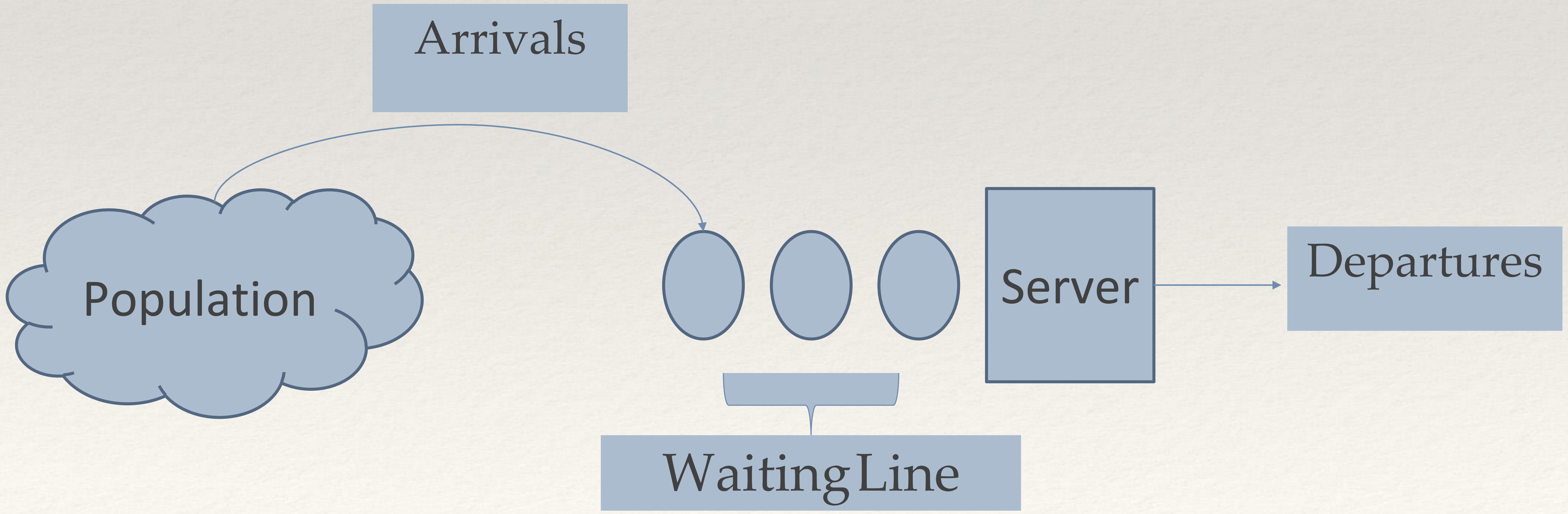
Simulation of a Queueing System

- Consider a single server queue (ATM)
  - Infinite calling population and infinite queue capacity
  - Single server (Only one customer served at a time)
  - Customers served in First in First out (FIFO) fashion
  - No departures from the queue (patient customers ☺)



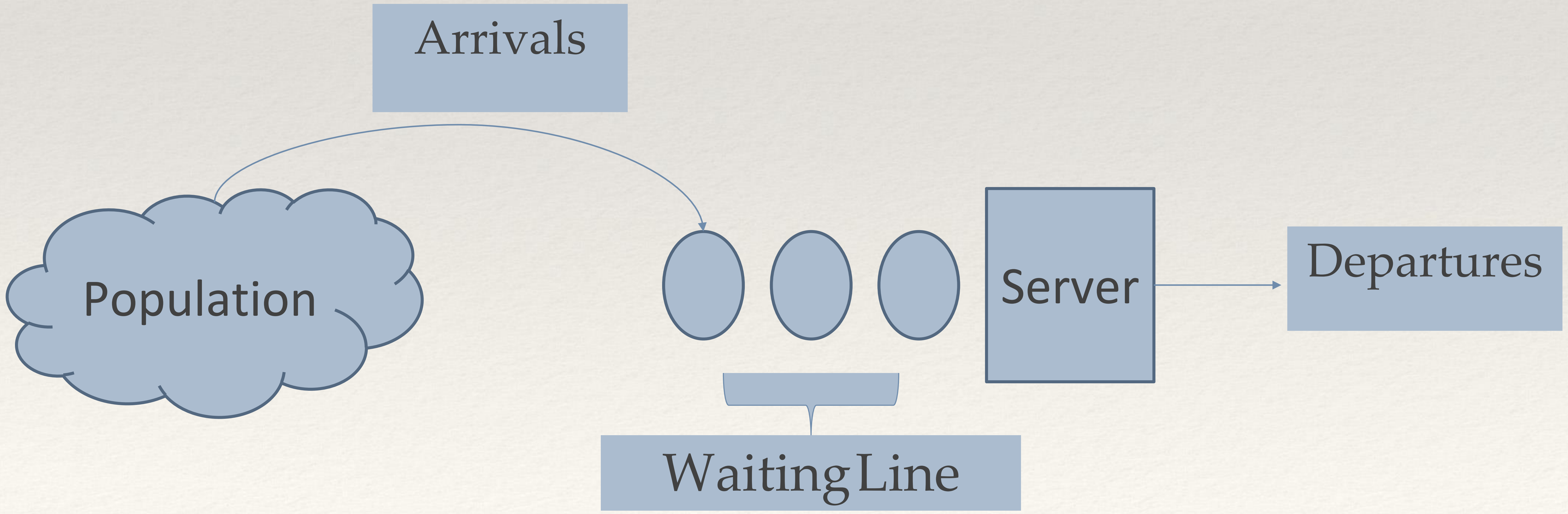
Simulation of a Queueing System

- Arrivals occur one at a time in a random fashion
- Once they join the waiting line they are eventually served
- Service times are of some random length according to a probability distribution



Simulation of a Queueing System

- State: Number of units in the system and the status of the server, busy or idle
- Event: A circumstance that cause a change in system state
  - Arrival of a customer
  - Departure of a customer after service completion
- Simulation Clock

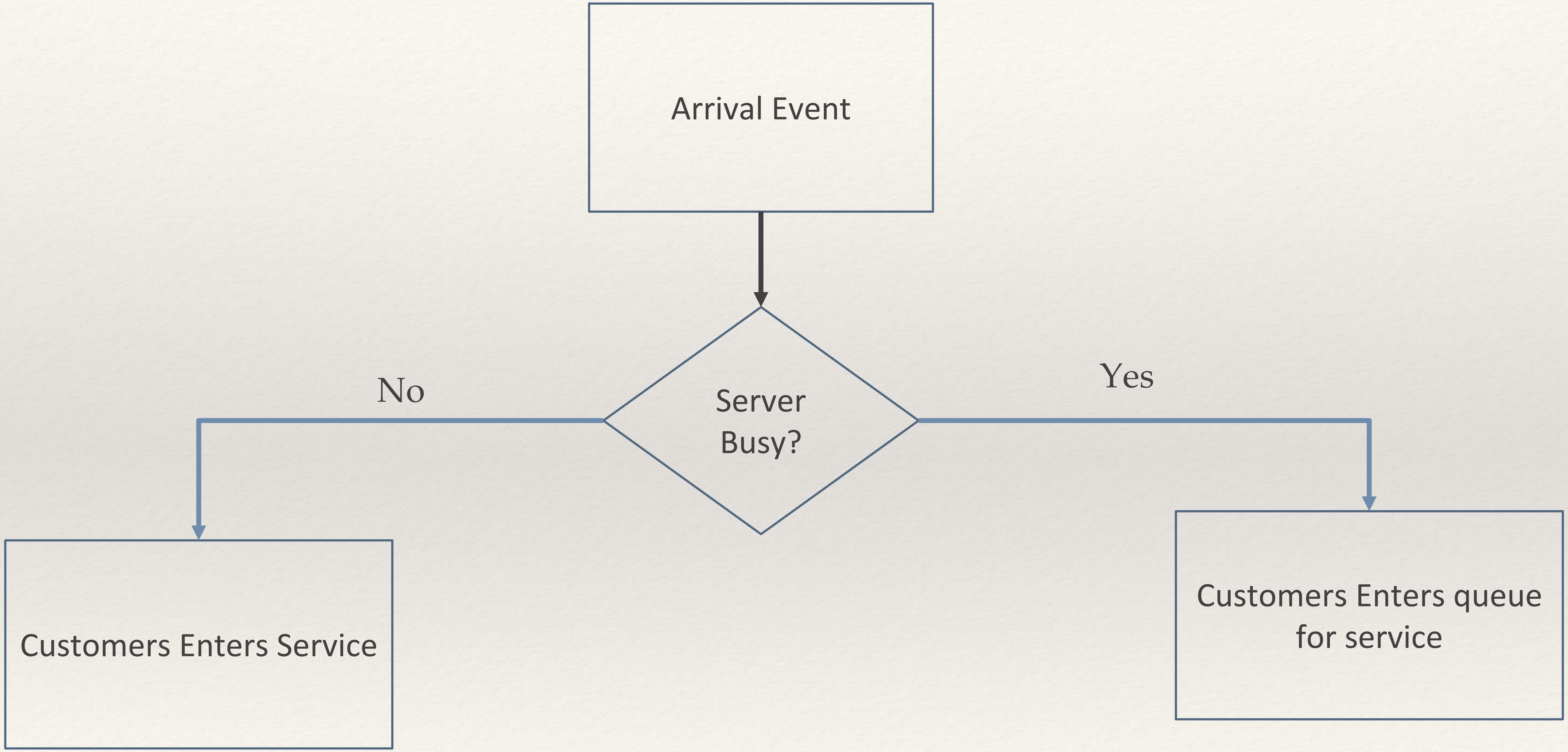


Simulating the clock

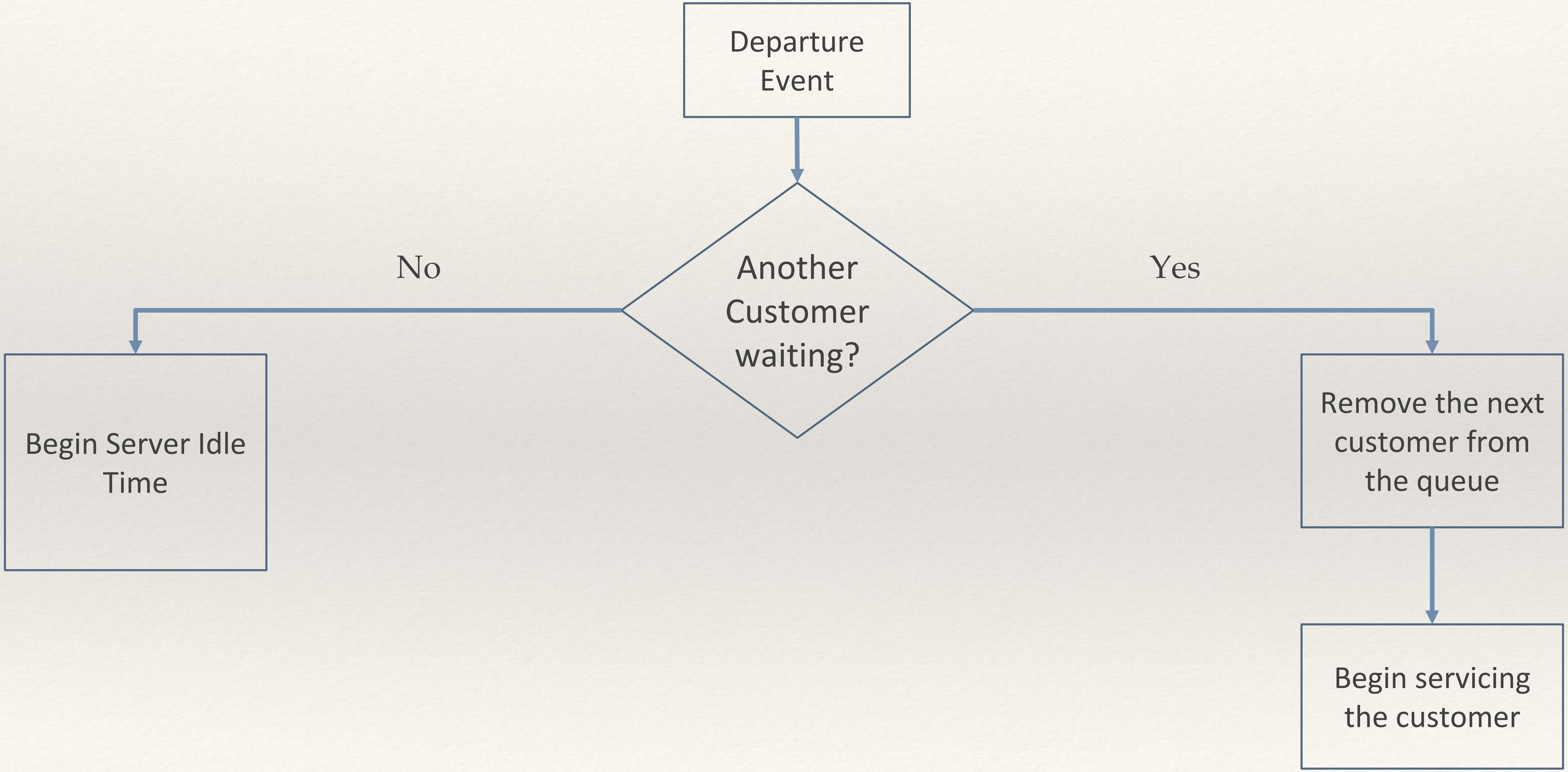
- Obvious choice: Start counting 1,2,3…
  - But it is inefficient in most cases

- Solution:Discrete Event Simulation
  - Get to the earliest event and update state and simulation clock

Simulation of a Queueing System



Simulation of a Queueing System



Simulation of a Queueing System

- Events usually occur at random times
- The randomness in simulation imitates uncertainty in real life. It is not known with certainty
  - When the next customer will arrive to a grocery store
  - How long for a bank teller will take to complete a transaction
- A statistical model developed from historical data

Simulation of a Queueing System

- For simplicity assume that interarrival times (time between arrivals) are the result of rolling a die.
  - 1 to 6 equally likely
- Assume that service times are generated from a distribution with four values; 1, 2, 3, 4 minutes. They are equally likely
- Assume that there is a customer using the ATM machine (server) at time 0.
- State: Number of customers in the system
- Let's simulate this simple system

Simulation of a Queueing System

- Simulation Clock t=0, Number of customers in the system, St=1
- Assign service time to the first customer.
  - 2 minutes (t=2)
- Assign interarrival time for next customer (time until next arrival)
  - 2 minutes (t=2)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0		

Simulation of a Queueing System

- Simulation Clock t=0, Number of customers in the system, St=1
- Assign service time to the first customer.
  - 2 minutes (t=2)
- Assign interarrival time for next customer (time until next arrival)
  - 2 minutes (t=2)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2

Simulation of a Queueing System

- Simulation Clock t=0, Number of customers in the system, St=1
- Assign service time to the first customer.
  - 2 minutes (t=2)
- Assign interarrival time for next customer (time until next arrival)
  - 2 minutes (t=2)
- Advance clock to the closest event time, update state

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2			

Simulation of a Queueing System

- Simulation Clock t=2
- Customer 1 departs the server, Customer 2 arrives (interarrival time was 2) and starts using server since the first customer left, St=1
- Assign service time for Customer 2: 1 minute (t=3)
- Assign interarrival time for Customer 3: 4 minutes (t=6)
- Move clock to the nearest event (departure)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2			

Simulation of a Queueing System

- Simulation Clock t=3
- Customer 2 departs, no waiting patients in the queue, server is idle, St=0
- Advance the clock to the next closest event (arrival)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6			

Simulation of a Queueing System

- Simulation Clock t=6
- Customer 3 arrives, starts using server, St=1
- Assign service time for Customer 3: **3 minutes (t=9)**
- Assign interarrival time for Customer 4: **1 minute (t=7)**
- Move the clock to nearest event (arrival)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7			

Simulation of a Queueing System

- Simulation Clock t=7
- Customer 4 arrives, Customer 3 in server, Customer 4 joins the queue, St=2
- Assign interarrival time for Customer 5**: 2 minutes (t=9)**
- Advance clock to the nearest event (arrival and departure)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7			
5	9			

Simulation of a Queueing System

- Simulation Clock t=9
- Customer 5 arrives, Customer 3 departs, Customer 4 starts service, customer 5 joins the queue, St=2
- Assign a service time for Customer 4: **2 minutes (t=11)**
- Assign an interarrival time for Customer 6: **6 minutes (t=15)**

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7	9	2	11
5	9			
6	15			

Simulation of a Queueing System

- Simulation Clock t=11
- Customer 4 departs, Customer 5 in server starts service, St=1
- Assign a service time for Customer 5: **1 minute (t=12)**
- Arrival for Customer 6 is at t =15
- Advance to the next event (departure)

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7	9	2	11
5	9	11	1	12
6	15			

Simulation of a Queueing System

- Simulation Clock t=12
- Customer 5 departs, no customer in the queue, St=0
- Advance to arrival of customer 6

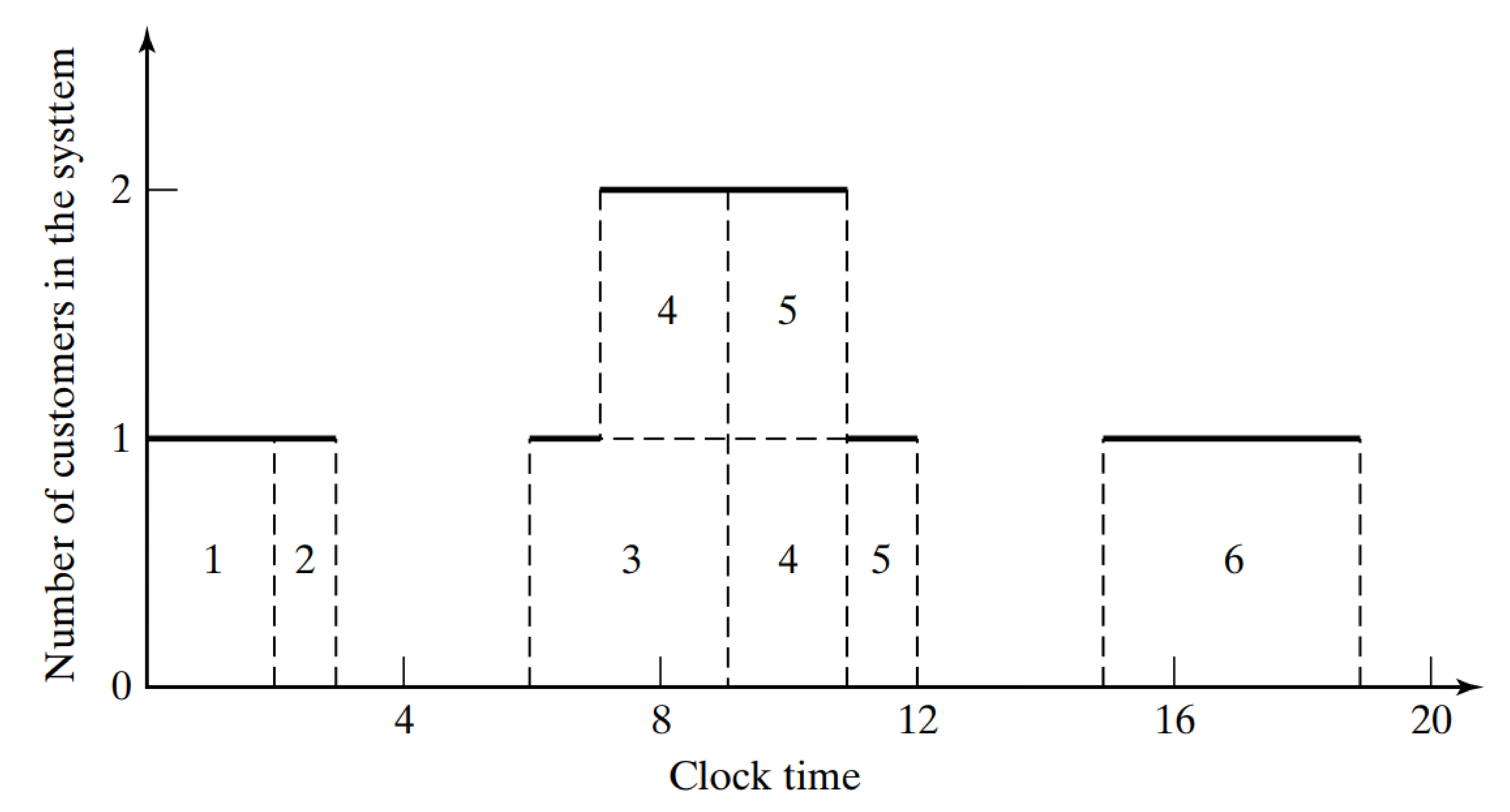
Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7	9	2	11
5	9	11	1	12
6	15			

Simulation of a Queueing System

- Simulation Clock t=15
- Customer 6 arrives, starts getting service, St=1
- Assign a service time for Customer 6: **4 minutes (t=19)**
- **END OF SIMULATION**

Customer	Arrival Time	Time Service Begins	Service Time	Time Service Ends
1	0	0	2	2
2	2	2	1	3
3	6	6	3	9
4	7	9	2	11
5	9	11	1	12
6	15	15	4	19

Simulation of a Queueing System



Simulation of a Grocery Checkout

- A small grocery store has only one checkout counter. Customers arrive at this checkout at random from 1 to 8 minutes apart. Each possible value of interarrival time has the same probability of occurrence. The service time vary from 1 to 6 minutes with the probabilities shown below. Analyze the system by simulating the arrival of 100 customers.

Service Time	Probability	Cumulative Probability
1	0.10	0.10
2	0.20	0.30
3	0.30	0.60
4	0.25	0.85
5	0.10	0.95
6	0.05	1.00

Simulation of a Grocery Checkout

- How can we generate random interarrival times and service times in accordance with the given distributions?
- Generate a uniform random number R, between 0 and 1, then return service time x, if it falls in the interval (F(x-1),F(x)]

Service Time	Probability	Cumulative Probability	Random Number Interval
1	0.10	0.10	(0-0.10]
2	0.20	0.30	(0.10-0.3]
3	0.30	0.60	(0.30-0.60]
4	0.25	0.85	(0.60-0.85]
5	0.10	0.95	(0.85-0.95]
6	0.05	1.00	(0.95-1.00]

0.553

Generated Service time 3 minutes

Simulation of a Grocery Checkout

- Same idea for interarrival times

IA Time	Probability	Cumulative Probability	Random number interval
1	0.125	0.125	(0.000-0.125]
2	0.125	0.250	(0.125-0.250]
3	0.125	0.375	(0.250-0.375]
4	0.125	0.500	(0.375-0.500]
5	0.125	0.625	(0.500-0.625]
6	0.125	0.750	(0.625-0.750]
7	0.125	0.875	(0.750-0.875]
8	0.125	1.000	(0.875-1.000]

Generated Inter-arrival Times

Customer	Random Digits	Time between Arrivals (Minutes)	Customer	Random Digits	Time between Arrivals (Minutes)
1	—	—	11	413	4
2	064	1	12	462	4
3	112	1	13	843	7
4	678	6	14	738	6
5	289	3	15	359	3
6	871	7	16	888	8
7	583	5	17	902	8
8	139	2	18	212	2
9	423	4	⋮	⋮	⋮
10	039	1	100	538	5

Generated Service Times

Customer	Random Digits	Service Time (Minutes)	Customer	Random Digits	Service Time (Minutes)
1	84	4	11	94	5
2	18	2	12	32	3
3	87	5	13	79	4
4	81	4	14	92	5
5	06	1	15	46	3
6	91	5	16	21	2
7	79	4	17	73	4
8	09	1	18	55	3
9	64	4	⋮	⋮	⋮
10	38	3	100	26	2

Exercise: Complete the Table

- Use the generated numbers given in the previous slides to run the simulation until the arrival of the 5th customer
- Use the table template below:

Cust.	Arrival Time	Time Service Begins	Service Time	Waiting Time in Queue	Time Service Ends	Time in system	Server idle time
1	0	0					0

Final Simulated Table

Customer	Interarrival Time (Minutes)	Clock Arrival Time	Service Time (Minutes)	Clock Time Service Begins	Waiting Time in Queue (Minutes)	Clock Time Service Ends	Time Customer Spends in System (Minutes)	Idle Time of Server (Minutes)
1		0	4	0	0	4	4	
2	1	1	2	4	3	6	5	0
3	1	2	5	6	4	11	9	0
4	6	8	4	11	3	15	7	0
5	3	11	1	15	4	16	5	0
6	7	18	5	18	0	23	5	2
7	5	23	4	23	0	27	4	0
8	2	25	1	27	2	28	3	0
9	4	29	4	29	0	33	4	1
10	1	30	3	33	3	36	6	0
11	4	34	5	36	2	41	7	0
12	4	38	3	41	3	44	6	0
13	7	45	4	45	0	49	4	1
14	6	51	5	51	0	56	5	2
15	3	54	3	56	2	59	5	0
16	8	62	2	62	0	64	2	3
17	8	70	4	70	0	74	4	6
18	2	72	3	74	2	77	5	0
19	7	79	1	79	0	80	1	2
20	4	83	2	83	0	85	2	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	5	415	2	416	1	418	3	0
Total	415		317		174		491	101

Analysis Using the Simulation Results

- The average waiting time for a customer is

$$\frac{\text{Total time customers waited in queue}}{\text{Number of customers}} = \frac{174}{100} = 1.74 \text{ mins}$$

- The average service time is (E[Service Time] =?)

$$\frac{\text{Total service time}}{\text{Number of customers}} = \frac{317}{100} = 3.17 \text{ mins}$$

- The average time spent in system

$$\frac{\text{Total time spent in system}}{\text{Number of customers}} = \frac{491}{100} = 4.91 \text{ mins } (1.74+3.17)$$

- The proportion of times that server is idle

$$\frac{\text{Total time server idle}}{\text{Total Time simulation Run}} = \frac{101}{418} = 0.24$$

- The probability that a customer has to wait in the queue

$$\frac{\text{Number of customers waited in queue}(>0 \text{ mins})}{\text{Total number of customers}} = \frac{46}{100} = 0.46$$

Note: 46 can be found by counting the number of customers who waited in the queue more than 0 minutes. This cannot be directly observed from the table presented in the slides, since the table is not complete.

Excel Time

- Now Let’s see how we can simulate this grocery store example in Excel…