MGT 422 Operations Engine

Lecture 6: City Hospital ED Discussion & More on Variability!

2026 // Spring 1 // Core

Yale SCHOOL OF MANAGEMENT

Announcements

- No pre-slides today
- Littlefield simulation game:
  - Simulation System Registration is Open!
    - Instructions have been posted on Canvas
    - **Please register an account and your team ASAP (by next Wednesday Feb 11)**
    - **More instructions on this in the next couple of slides**
  - Action Plan Writeup due at 5pm EST on **Sunday, Feb 15**
    - One page
    - Submit via Canvas, one submission for each team
  - Littlefield Simulation starts at 5pm EST on **Sunday, Feb 15**
  - Simulation will run for 5 days, ending at 5pm EST on Friday, Feb 20

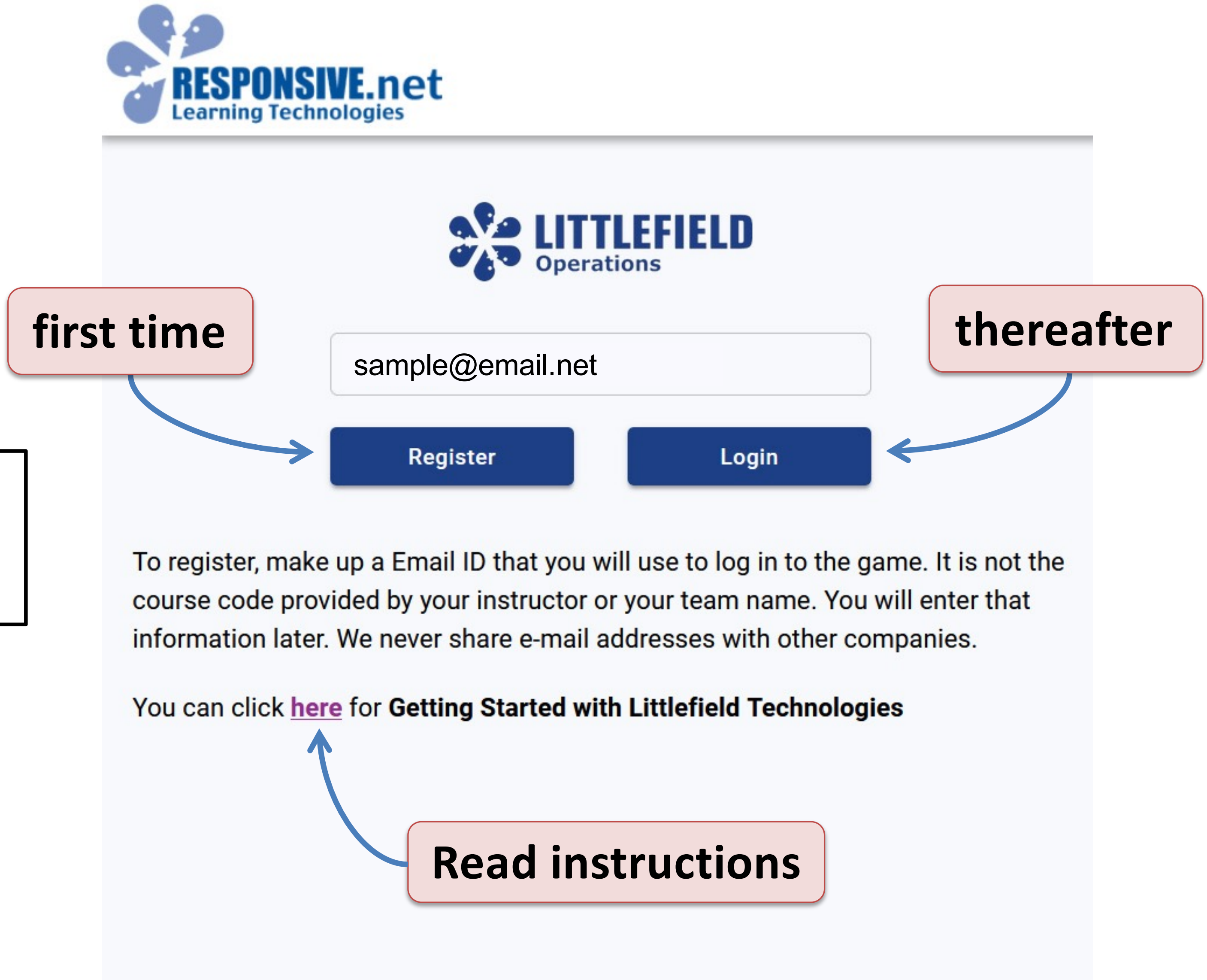
- Assignment #3 has been posted: Individual, due next Wed/Thurs before class

Littlefield Registration

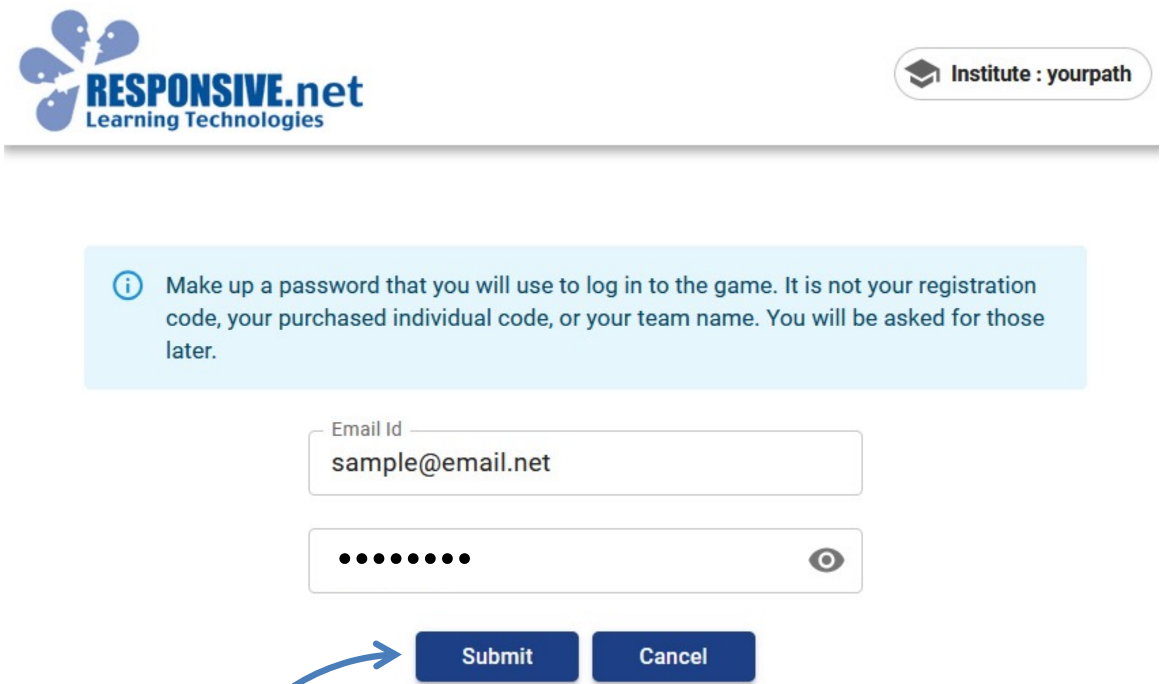
Website for Game + Registration: op3.responsive.net/lf/yale/
Access Code for Operations Engine: “bulldog”

(all of this information is also on the Littlefield Instructions Guide on Canvas!)

Account registration (individual)



Account registration (individual)



submit

RESPONSIVE.net
Learning Technologies

Institute : yourpath

Please enter the course code provided by your instructor.

Do not enter the individual code you purchased. You will be asked for that later.

bulldog

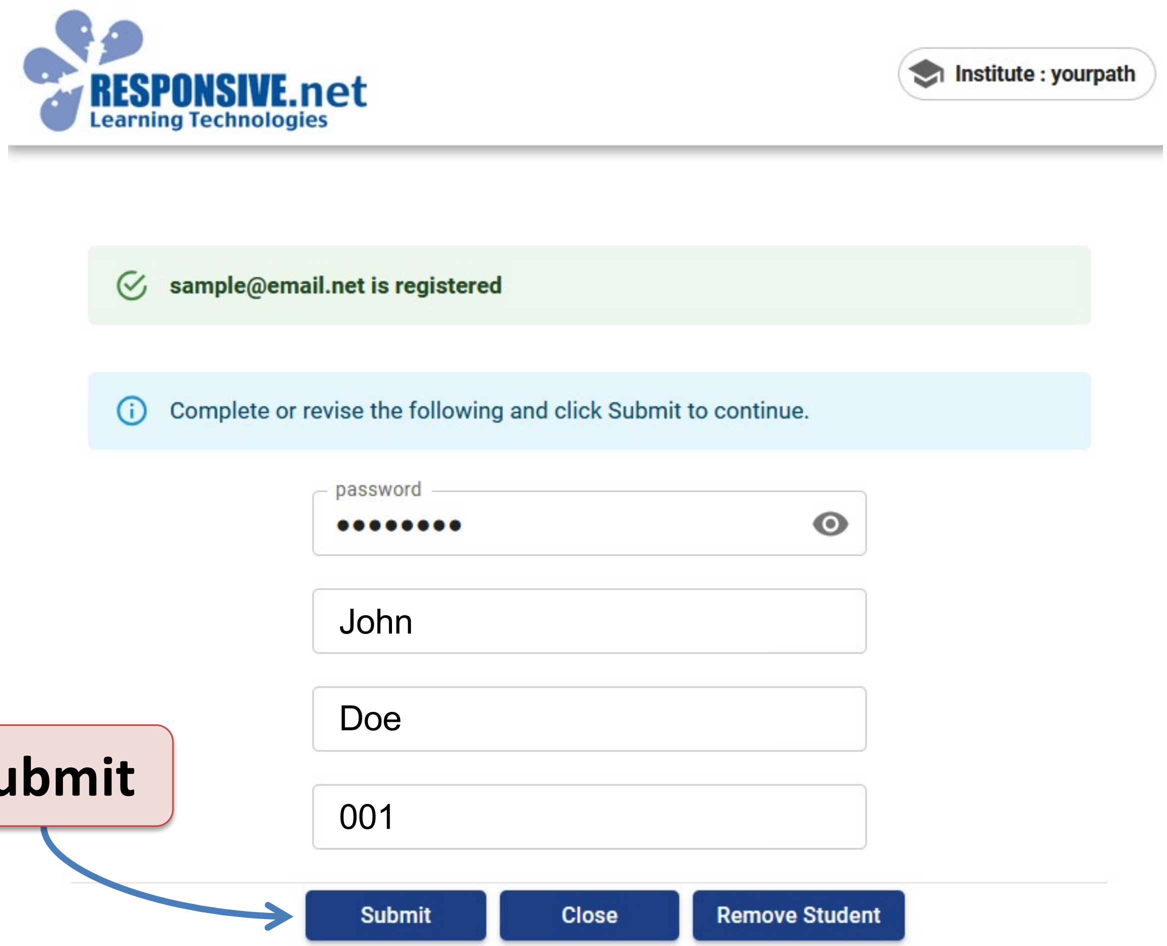
Submit Cancel

Account registration (individual)

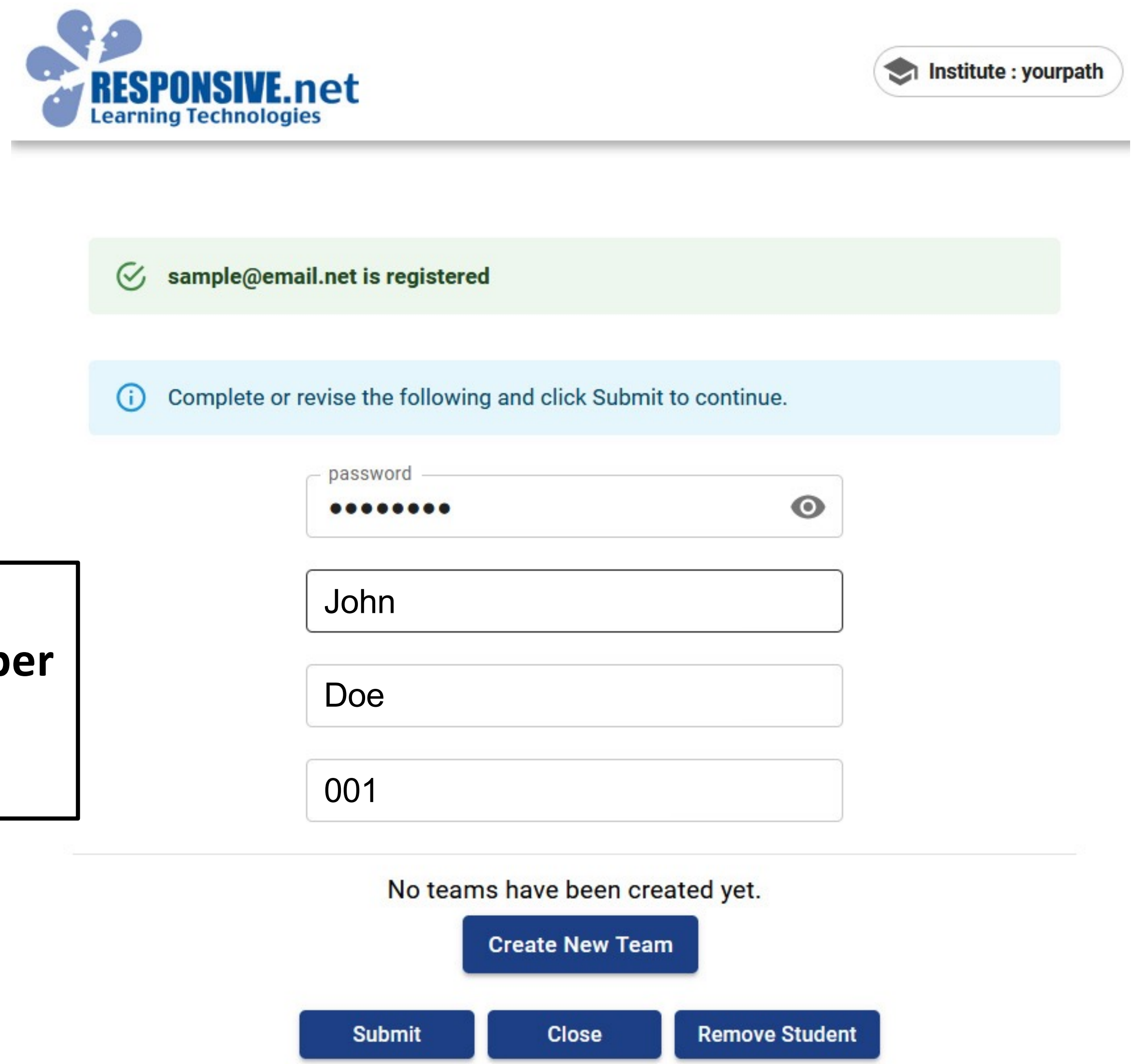
submit

Account registration (individual)

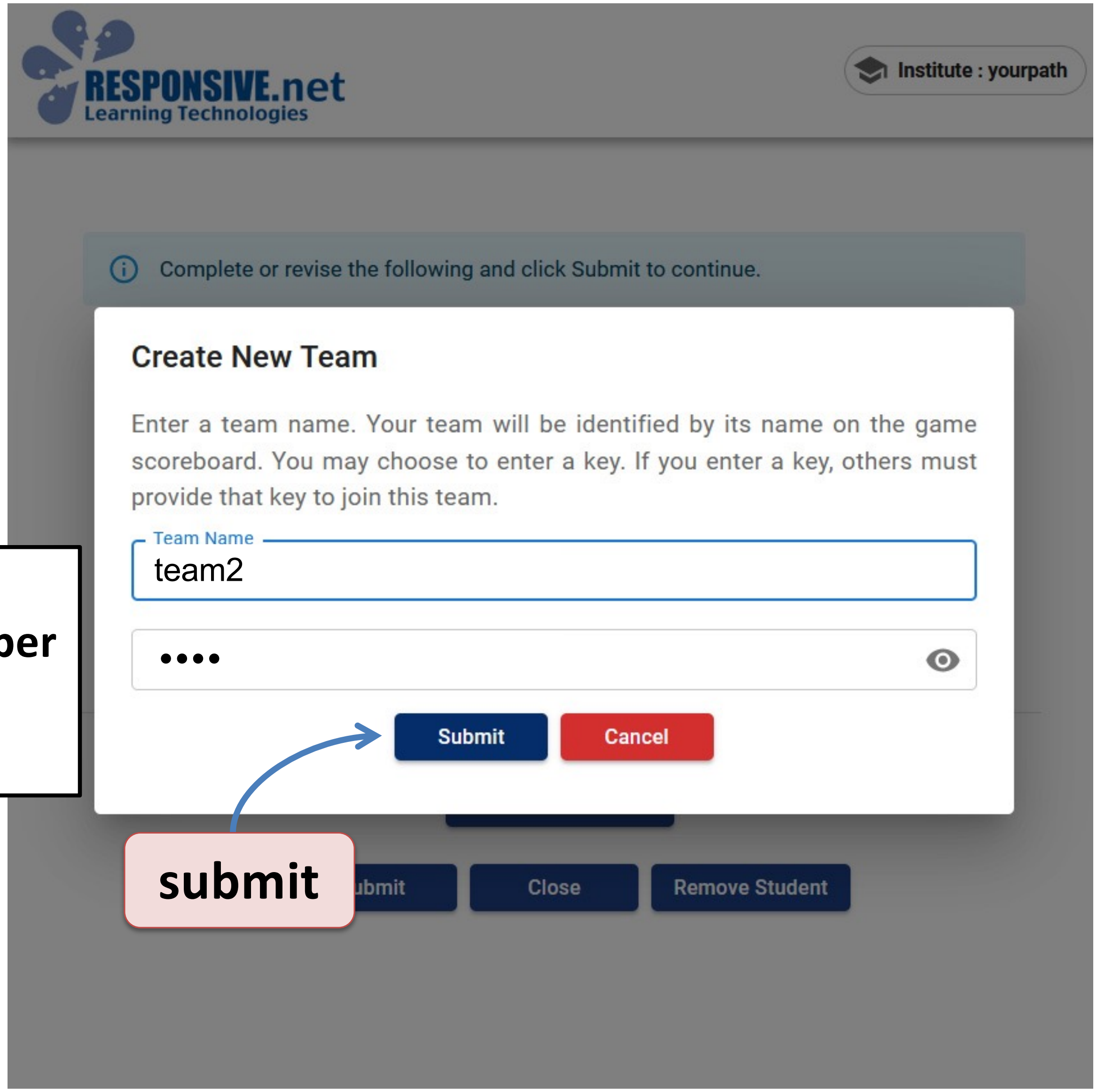
submit

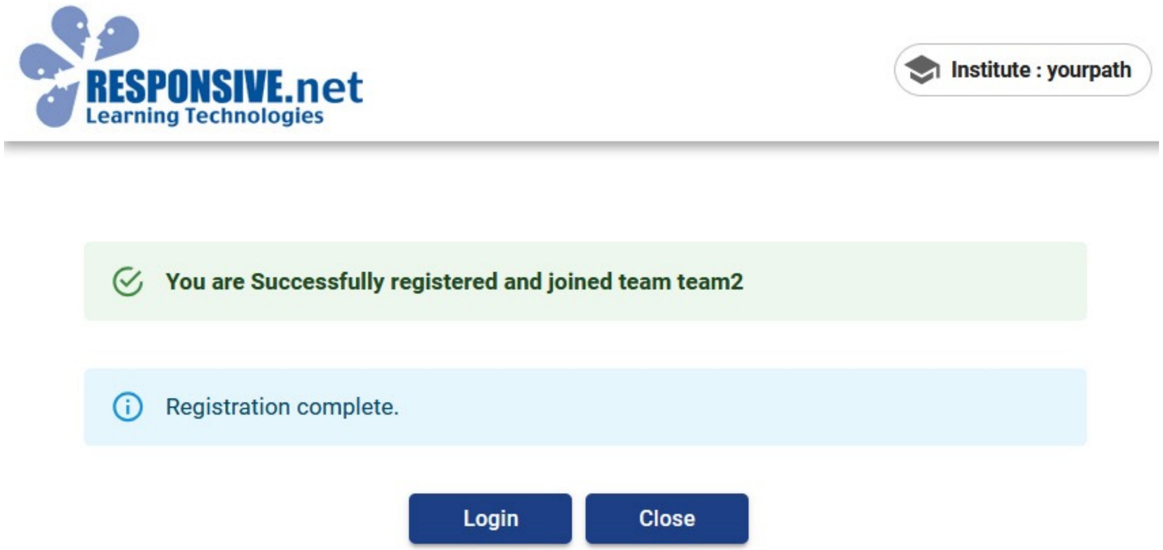


Creating a team (Only one person per group needs to do this)



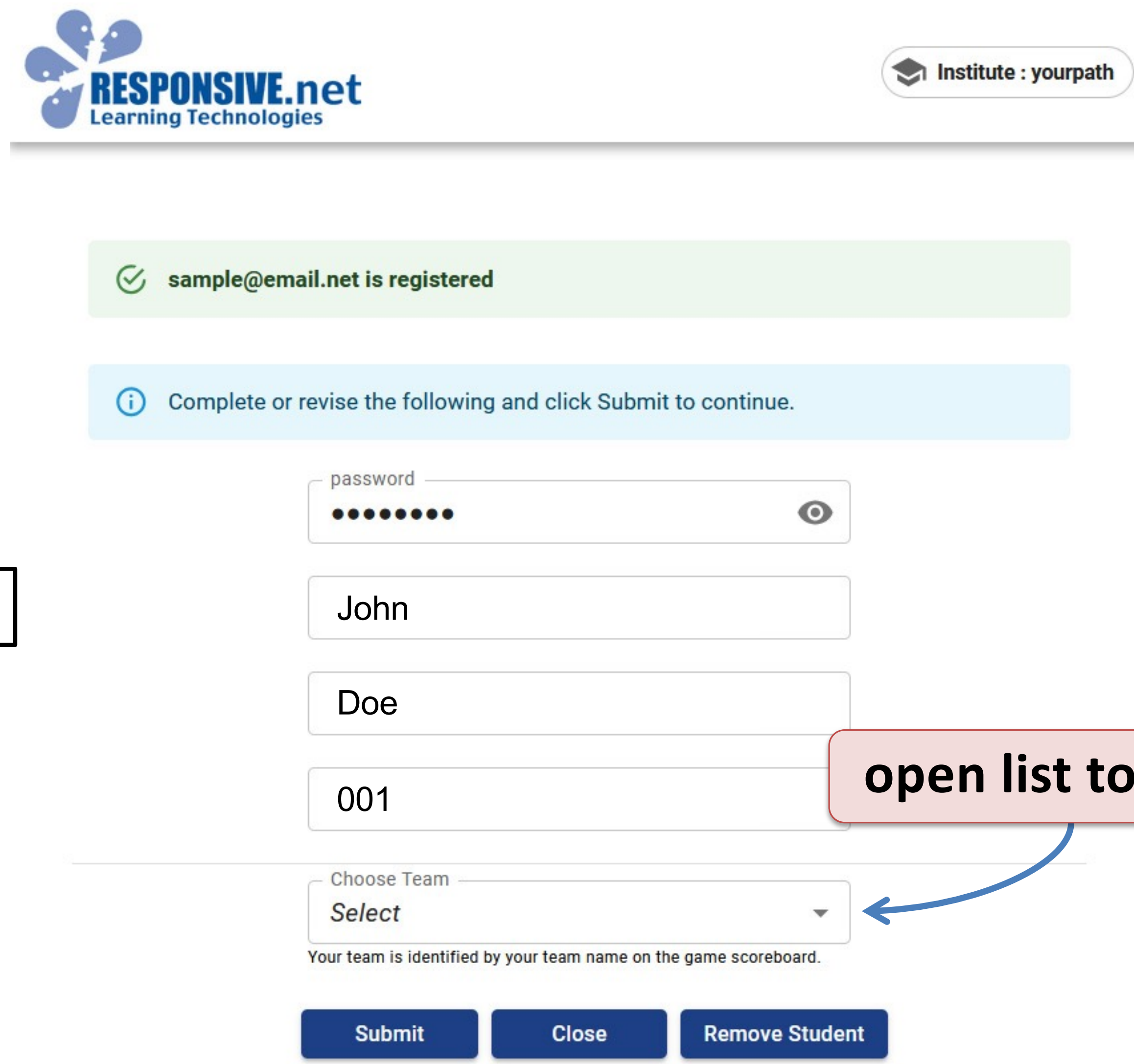
Creating a team (Only one person per group needs to do this)





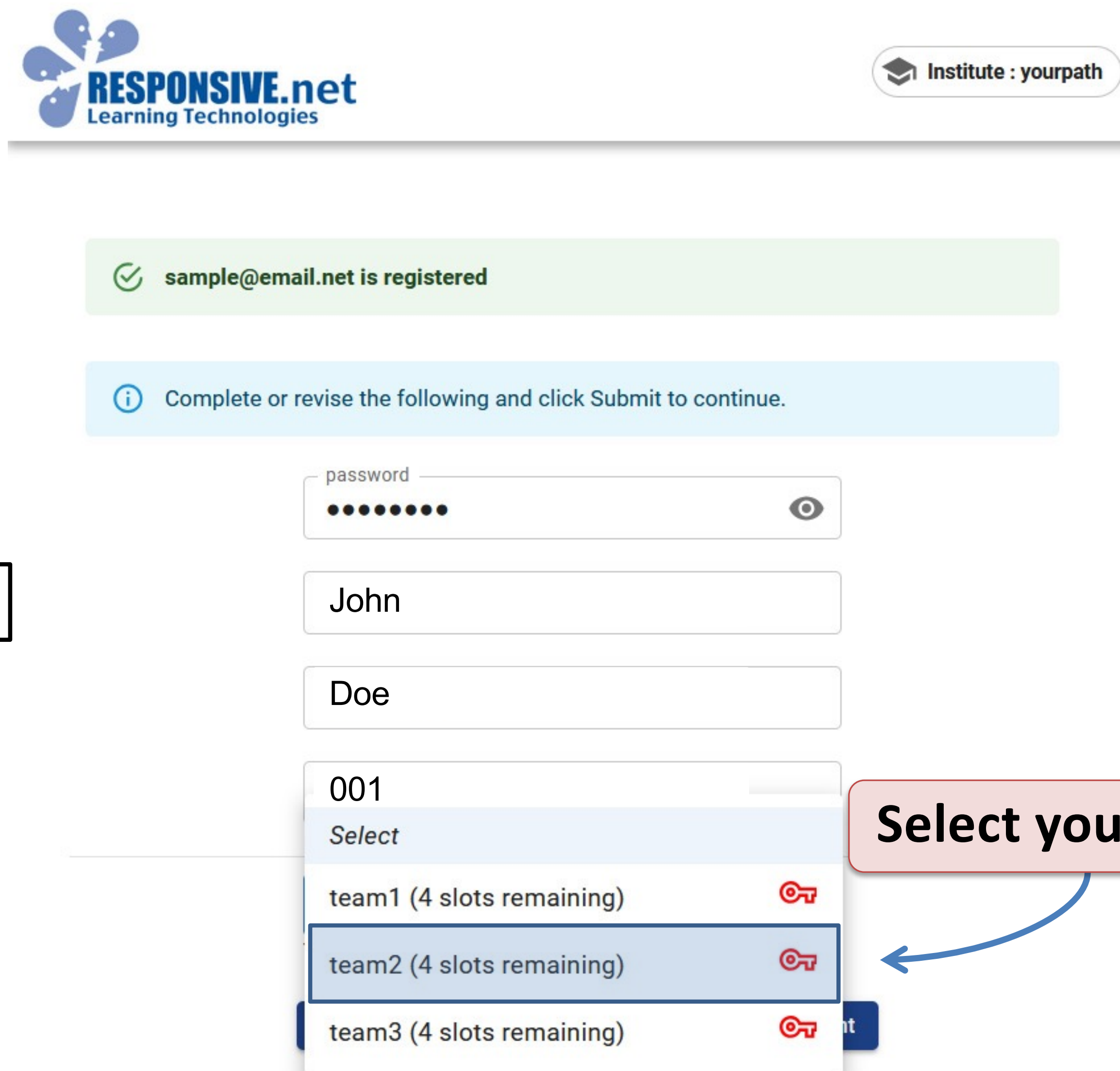
Creating a team (Only one person per group needs to do this)

Joining your team



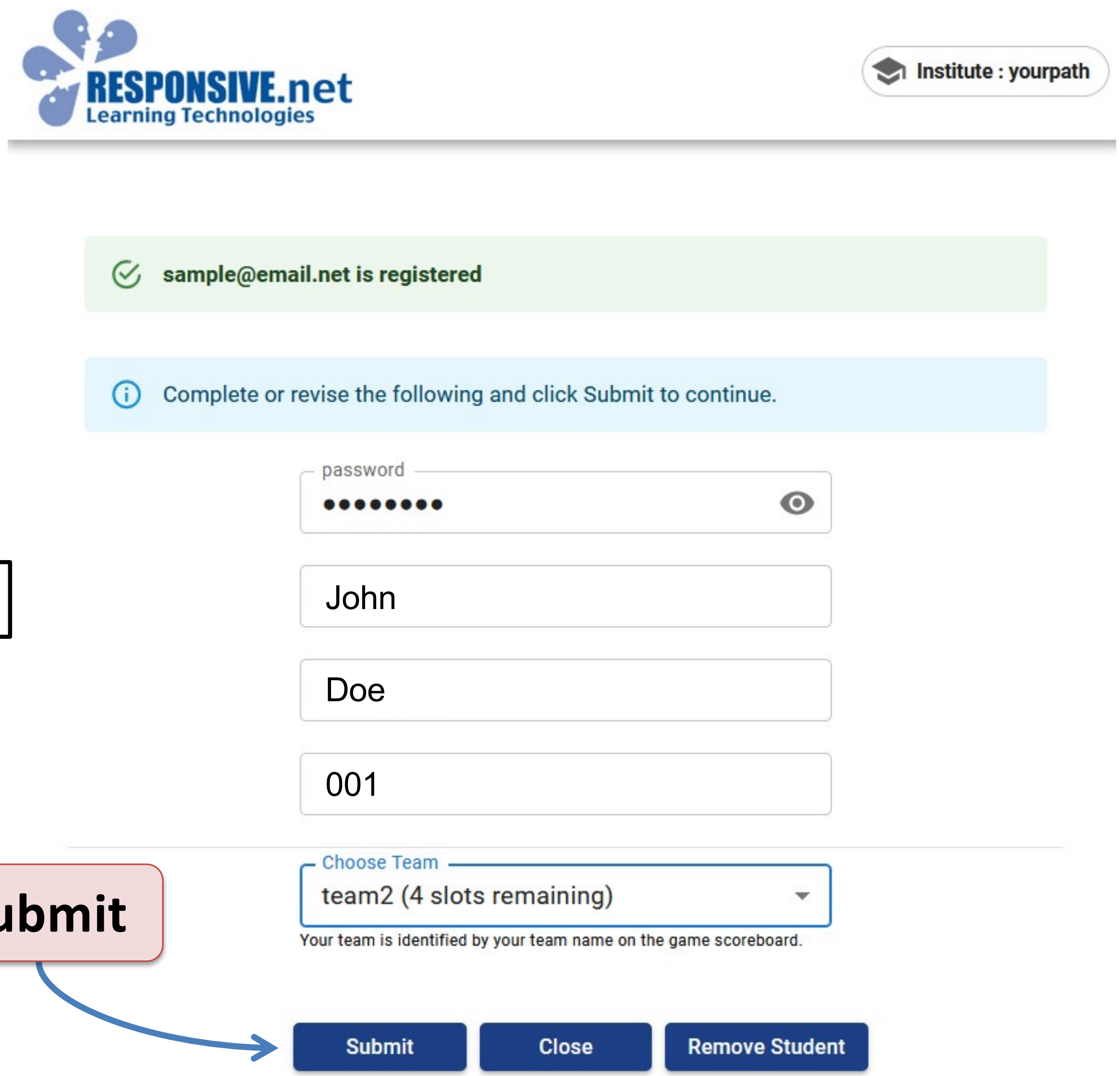
open list to select

Joining your team



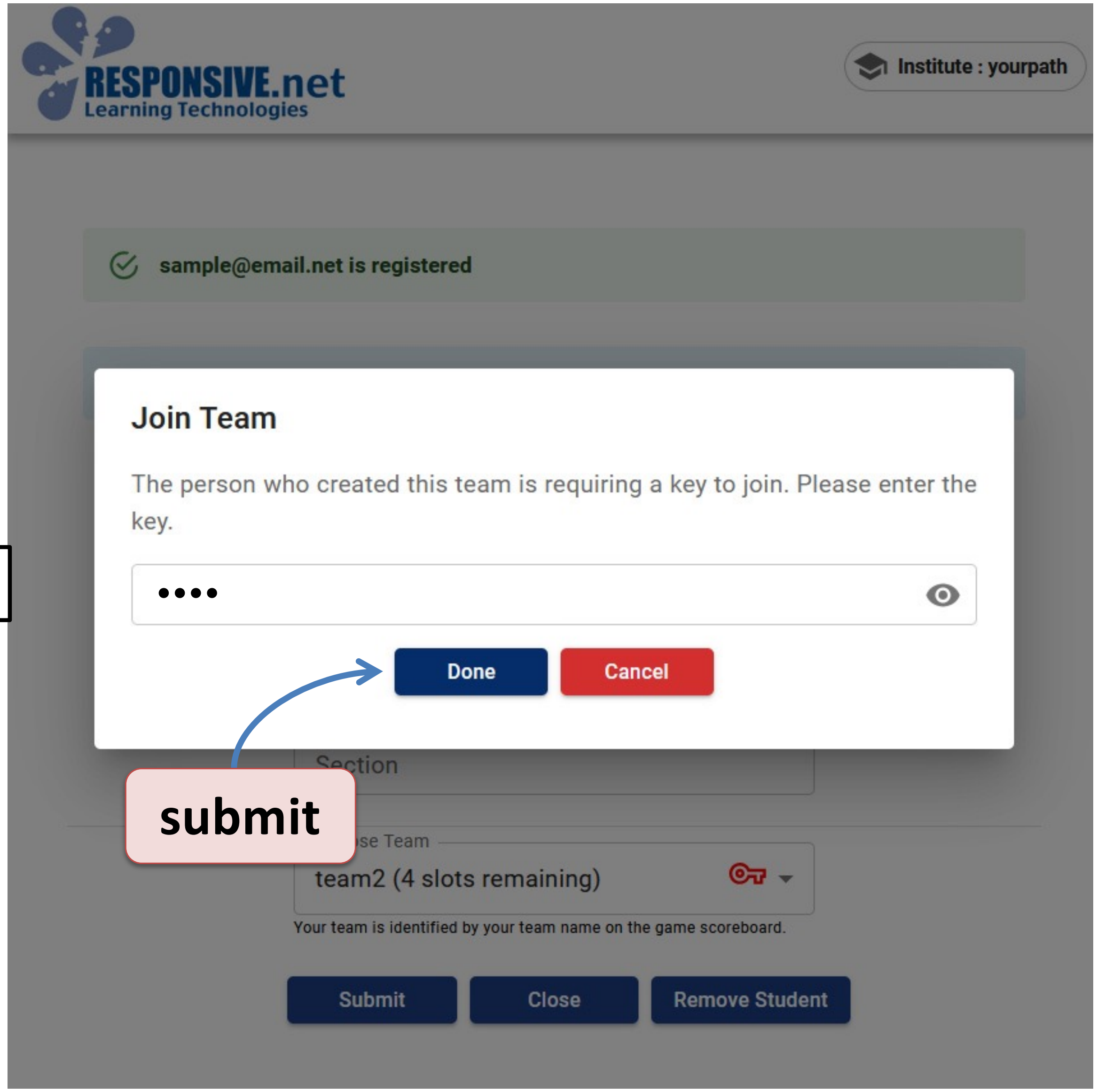
Select your team

Joining your team



submit

Joining your team



Littlefield Expert TAs

- If you have any questions about registration or Littlefield during the duration of the simulation, please feel free to reach out to our Littlefield Expert TAs:

Gai Sawant (gai.sawant@yale.edu)

Alex Hunt (alex.hunt@yale.edu)

Martin Cheong (martin.cheong@yale.edu)

Some True/False Review Questions

- The bottleneck capacity of a process is always the flow rate of that process. True or False?
- If an arrival process is a Poisson process, we need to know the mean and the standard deviation to calculate the coefficient of variation. True or False?
- If the interarrival time of a process is exponentially distributed, we know that the arrival rate is normally distributed. True or False?
- If the service time of a step in the process is exponentially distributed, the coefficient of variation of arrivals (CV_a) is then equal to 1. True or False?
- When modeling random interarrival times, you should always select the exponential distribution. True or False?

(answers on next slide so you have the opportunity to try and solve these yourself first!)

Some True/False Review Questions + Answers

- The bottleneck capacity of a process is always the flow rate of that process. True or False?
  - **False**. The flow rate is the minimum of the arrival rate and the process (bottleneck) capacity. If you are in a demand-constrained system, the arrival rate (or demand) is the flow rate.
- If an arrival process is a Poisson process, we need to know the mean and the standard deviation to calculate the coefficient of variation. True or False?
  - **False**. If an arrival or service process is a poisson process, the interarrival times or process times would be exponentially distributed, which means the mean would equal the standard deviation and the relevant CV would be equal to 1.
- If the interarrival time of a process is exponentially distributed, we know that the arrival rate is normally distributed. True or False?
  - **False**. If the interarrival time of a process of exponentially distributed, we would know that the arrival rate is a poisson process.
- If the service time of a step in the process is exponentially distributed, the coefficient of variation of arrivals (CV_a) is then equal to 1. True or False?
  - **False**. If the service time of a step in the process is exponentially distributed, the coefficient of variation for the PROCESS (CV_p) would be equal to 1. This says nothing about the CV_a (you would need additional information on the arrival process).
- When modeling random interarrival times, you should always select the exponential distribution. True or False?
  - **False**. It depends on the historical data and what you know about the process and variability!

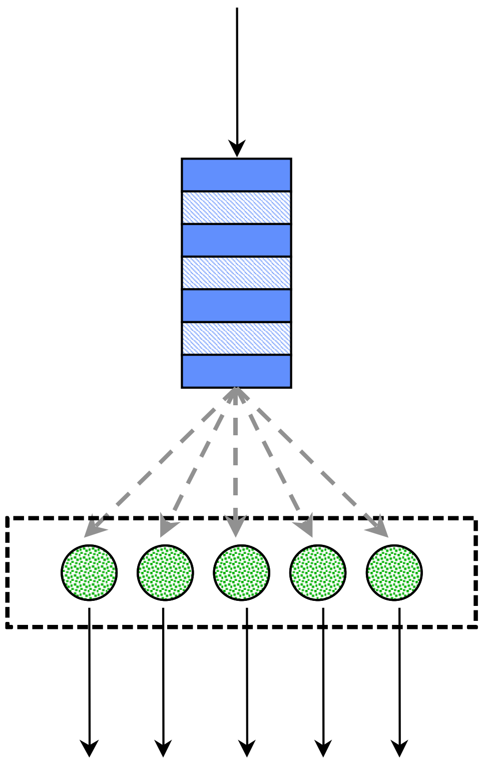
Outline for Today's Class

Waiting Time Analysis

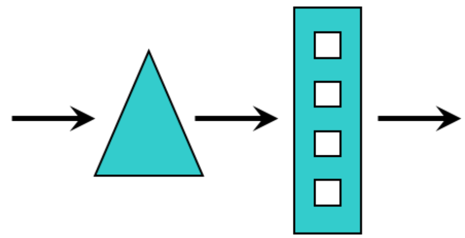
- Review of queuing concepts
- City Hospital case
- Properties of waiting times
- Mitigating waiting times

Quick Review of Queuing Model Basics

- Variability in interarrival times
  - ✓ a = average interarrival time ➔ $1/a$ = arrival rate
  - ✓ CV_a = coefficient of variation of interarrival time
  - ✓ Poisson process = exponentially distributed interarrival time ($CV_a = 1$)
- Variability in service times
  - ✓ p = average service time ➔ $1/p$ = capacity of each server
  - ✓ CV_p = coefficient of variation of service time
  - ✓ Exponentially distributed service times ($CV_p = 1$)
- Utilization
  - ✓ m = number of servers
  - ✓ Utilization: $u = p / (a \times m)$
- Average waiting time in queue: T_q



Average Waiting Time in Queue T_q



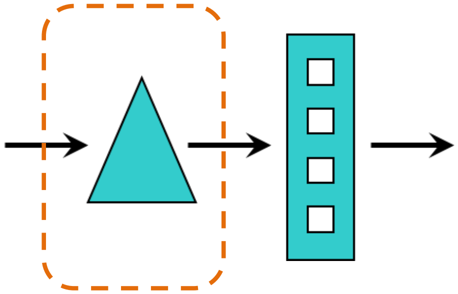
T_q = Average waiting time in queue

p = Average service time of each server

Flow Time = $T = T_q + p$

$$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\,2} + CV_p^{\,2}}{2}\right) \qquad \text{where} \qquad u = \text{Utilization} = \frac{p}{a \times m}$$

From Average Waiting Time to *Queue Length*

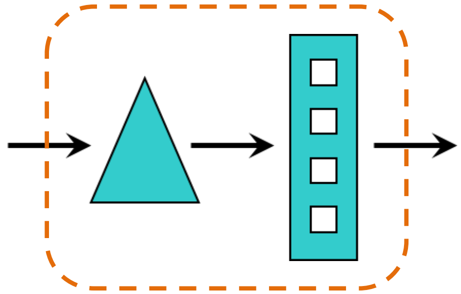


$I_q = R \times T_q$ **(Little's Law)**

- Recall the three performance measures of a process:
  - T = Flow Time
  - R = Flow Rate
  - I = Inventory

- In the queue part of the stable queuing system:
  - T_q = Average waiting time in the queue
  - R = Arrival rate (= $1/a$)
  - I_q = Average number of customers waiting in queue ("queue length")

Little's Law for the *Entire* Queuing System



$I = R \times T$ **(Little's Law)**

- Recall the three performance measures of a process:
  - T = Flow Time
  - R = Flow Rate
  - I = Inventory

- In the entire queuing system:
  - T = Average time for a customer to go through the system (flow time)
  - R = Arrival rate (= $1/a$)
  - I = Average number of customers in the system (those waiting in the queue + those being served)

Example

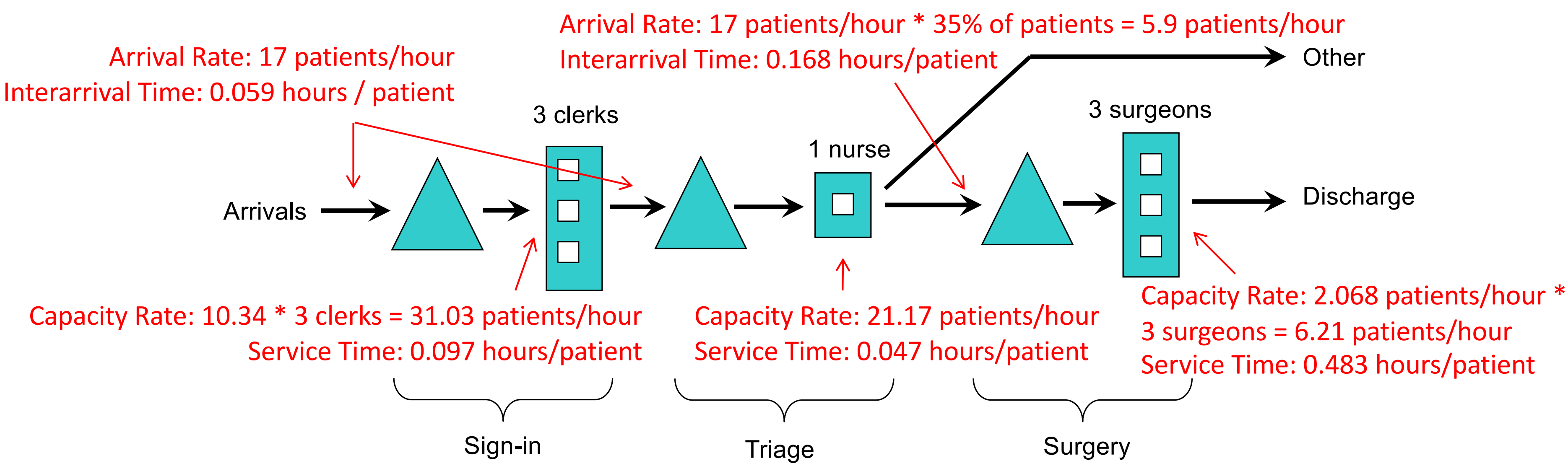
Suppose that customers arrive at a service facility as a Poisson process with rate equal to 2 customers/min. There are 6 service representatives, each taking exactly 2.5 minutes to serve a customer. On average, how many customers are waiting in queue?

- $1/a$ = 2 customers/min ➔ a = 0.5 min
- Poisson arrivals ➔ $CV_a = 1$
- 6 servers ➔ $m = 6$
- Constant service time of 2.5 min ➔ p = 2.5 min, $CV_p = 0$

- Utilization: $u = p / (a \times m) = 2.5 / (0.5 \times 6) = 2.5/3 = 0.833$
- Average waiting time in queue: $$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\,2} + CV_p^{\,2}}{2}\right) = 0.758 \text{ min}$$
- Average number of customers waiting in queue: $I_q = R \times T_q = (1/a) \times T_q$ = (2 customers/min) × (0.758 min) = 1.52 customers

City Hospital Emergency Room Case

City Hospital Emergency Room (Simplified)



- Assumptions:
  - ✓ Patients arrive at ER based on a Poisson process
  - ✓ All service times are exponentially distributed
  - ✓ Ample room to wait in each stage

- What are some limitations of this representation?

What Happens in the Surgery Waiting Area with Two Surgeons?

Table 1: Arrivals to the Surgery Area

July	8-10	10-12	12-2	2-4	4-6	6-8	8-10	10-12	Average
1	15	11	9	3	3	10	11	11	9.1
2	7	5	6	8	6	9	10	9	7.5
3	8	10	8	4	11	9	14	10	9.3
4	11	12	13	9	14	15	13		12.0
5	23	11	9	15	18	13	12	11	14.0
6	6	13	14	11	15	15	17	17	13.5
7	10	8	15	11	8	11	8	6	9.6
8	10	14	10	13	14	12	13	11	12.1
9	5	12	4	8	12	8	12	6	8.4
10	12	14	8	7	7	13	11	7	9.9
11	11	11	12	11	14	17	10	12	12.5
12	5	18	8	14	12	10	13	5	10.6
13	13	5	5	8	5	13	14	4	8.4
14	5	8	6	11	5	5	15	5	7.5
15	13	11	12	13	12	9	10	5	10.6
16	9	10	9	9	12	7	6	3	8.1
17	8	13	9	8	10	9	11	6	9.2
18					Missing Data				
19	9	11	9	14	14	8	7	5	9.6
20	5	4	9	13	14	6	10	14	9.6
21	5	9	8	11	11	10	9	9	9.0
22	8	10	11	7	9	9	9	8	8.9
23	1	5	11	7	2	10	9	9	6.8
24	7	10	10	9	8	9	7	11	8.9
25	7	10	14	10	7	13	7	5	9.1
26	2	6	17	8	9	8	13	14	8.4
27	5	7	8	7	9	10	11	7	8.0
28	13	6	15	11	17	12	12	8	11.8
Average*	8.3	9.4	9.6	9.3	9.9	10.0	10.5	8.1	

* This is the average number of arrivals in a two-hour interval. The average number of arrivals per hour is one-half of this number.

These values represent the average arrivals of patients in a two-hour window. We divide by 2 to get the hourly arrival average.

Hourly arrivals range from **4.05 patients/hour and 5.25 patients/hour**.

With two surgeons, the capacity at the surgery stage is 2.068 patients/hour * 2 = **4.138 patients/hour**.

Therefore, except for during the last two-hour arrival window, all other arrival rates exceed the surgery area capacity with 2 surgeons, so the queue size would grow over time. (supply constrained system!)

With Three Surgeons

	Sign-in	Triage	Surgery
Utilization (u)	0.55	0.80	0.95
Avg. Waiting Time (T_q)	0.02 hour (1.2 mins)	0.19 hour (11.4 mins)	2.96 hour (177.6mins)
Avg. Waiting Time + Service Time ($T_q + p$)	0.12 hour (7.2 mins)	0.24 hour (14.4 mins)	3.44 hour (206.4 mins)

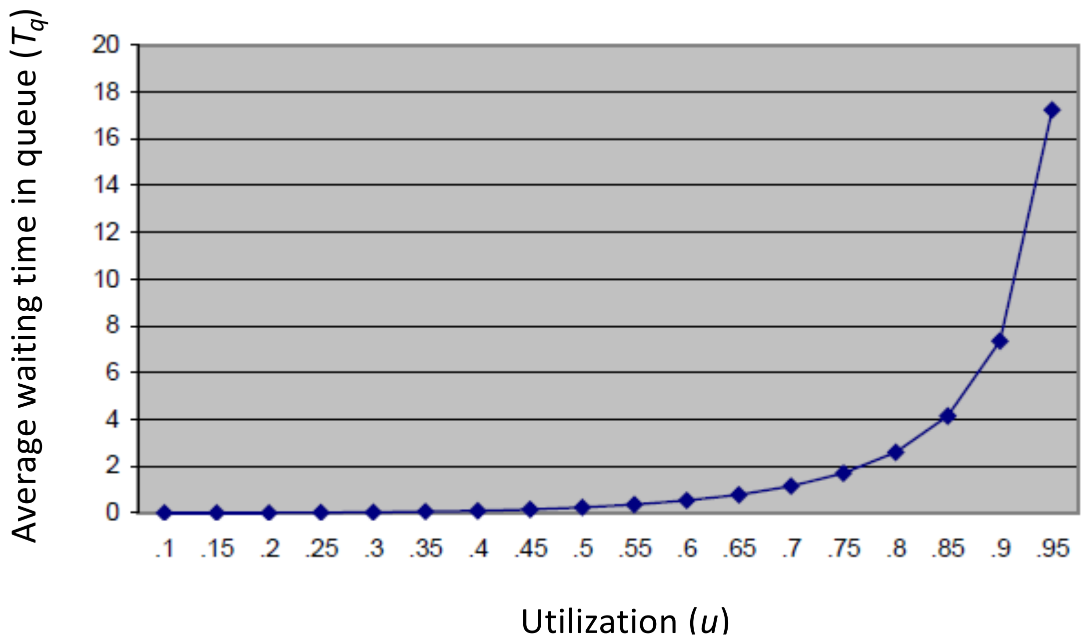
Flow Time with 3 surgeons = 0.12 + 0.24 + 3.44 = **3.80 hours** (228 mins)

With Four Surgeons

	Surgery with 3 surgeons	Surgery with 4 surgeons
Utilization (u)	0.95	0.71
Avg. Waiting Time (T_q)	2.96 hour (177.6 mins)	0.20 hour (12 mins)
Avg. Waiting Time + Service Time ($T_q + p$)	3.44 hour (206.4 mins)	0.69 hour (41.4 mins)

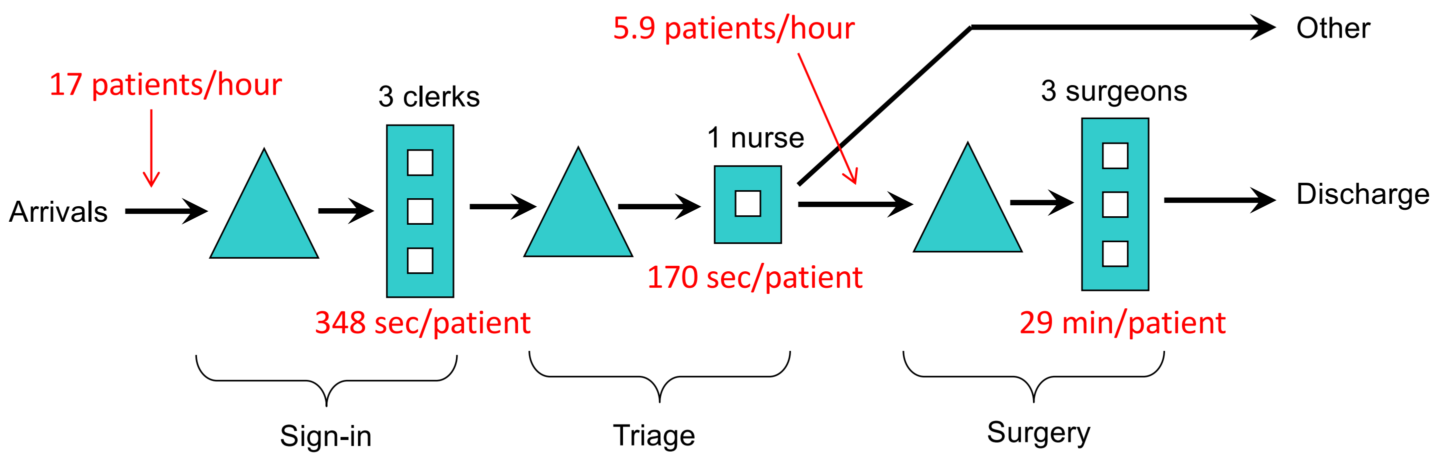
Flow Time with 4 surgeons = 0.12 + 0.24 + 0.69 = **1.05 hours** (63 mins)

Average Waiting Time Is NOT Linear in Utilization!



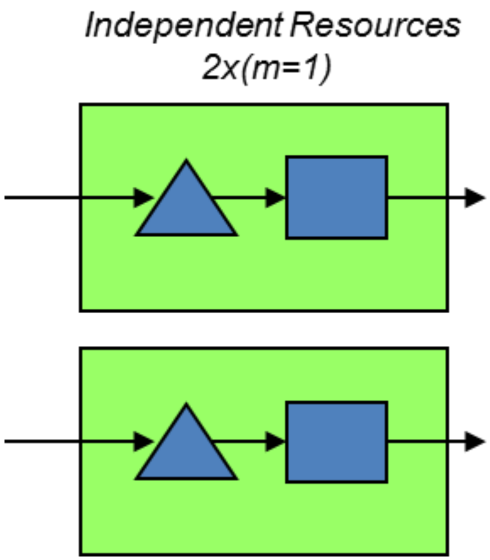
$$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\ 2} + CV_p^{\ 2}}{2}\right)$$

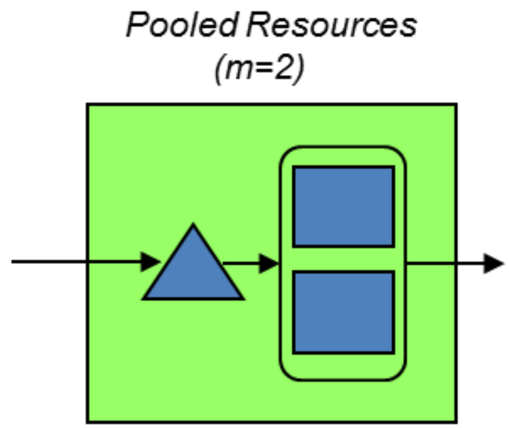
Improvement Options



- Besides adding a doctor, what other suggestions do you have for reducing waiting time in the surgery area?

Capacity Pooling





Separate queue system

Pooled queue system

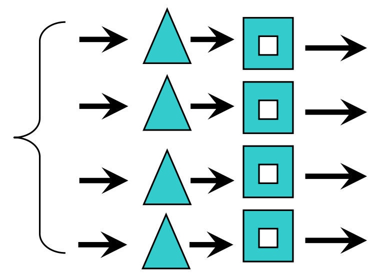
Separate System vs. Pooled System

For these two types of systems, assume:

- ✓ Same variability: $CV_a = 1$, $CV_p = 1$
- ✓ Same rate of customer arrivals to the system: $1/a = 1/35$ customers per second
- ✓ Same average service time of each server: $p = 120$
- ✓ **Utilization is the same**: $p \,/\, (a \times m) = 85.7\%$

Separate queue system:

Four servers, each having a separate queue



For each of the 4 identical queues:

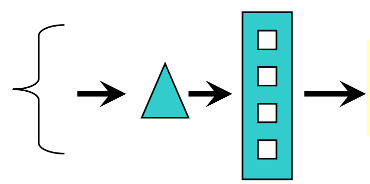
$a = 35 \times 4 = 140$

$m = 1$

Holding utilization constant!

Pooled system:

Four servers, sharing a common queue



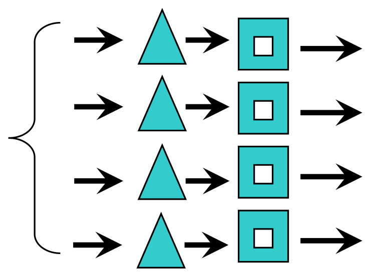
$a = 35$

$m = 4$

We also assume that customers cannot switch queues.

Effect of Pooling

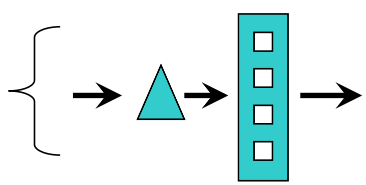
Separate queue system



For <u>each</u> of four queues in the system:

- ✓ $a = 140$, $p = 120$, $m = 1$, $CV_a = 1$, $CV_p = 1$
- ✓ Queue waiting time: $\boldsymbol{T_q = 720}$
- ✓ Queue length: $\boldsymbol{I_q =} R \times T_q = (1/140) \times 720 \boldsymbol{= 5.14}$

Pooled system

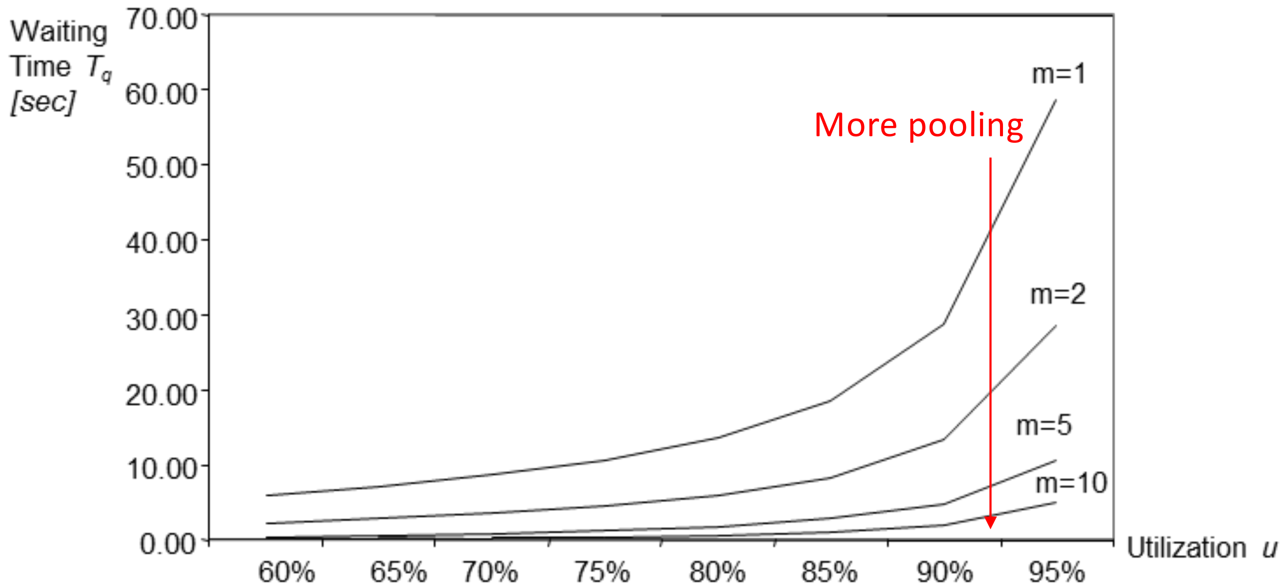


For the single queue in the system:

- ✓ $a = 35$, $p = 120$, $m = 4$, $CV_a = 1$, $CV_p = 1$
- ✓ Queue waiting time: $\boldsymbol{T_q = 150}$
- ✓ Queue length: $\boldsymbol{I_q =} R \times T_q = (1/35) \times 150 \boldsymbol{= 4.29}$

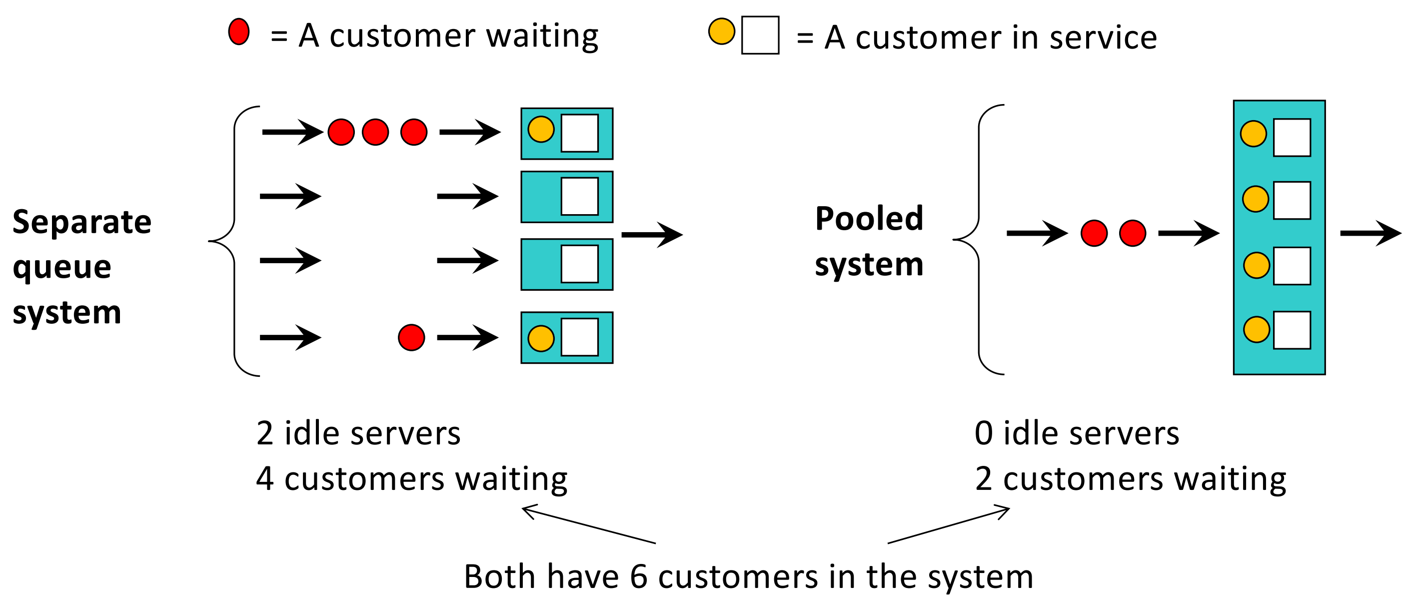
Average waiting time T_q is reduced considerably!

Impact of Pooling on Average Waiting Time



“Stochastic Economies of Scale”

How Does Pooling Work?



The separate queue system is inefficient because there can be customers waiting and some servers being idle *at the same time*

- ✧ Greater variability compared to the pooled system
- ✧ ***Pooling enables better workload balancing***

Implications of Pooling

- Benefits: Efficiency gain!
  - Reduce customer waiting time without adding resources
  - Reduce the number of resources while maintaining the waiting time
  - Combination of both

- Challenges in implementing a pooled system

But is Pooling Always Beneficial?

- It depends. New behavioral operations research suggests that sometimes, dedicated queues are quicker.

The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay

Hummy Song
Harvard University, Boston, Massachusetts 02163, hsong@hbs.edu
Anita L. Tucker
Brandeis University International Business School, Waltham, Massachusetts 02453, atucker@brandeis.edu
Karen L. Murrell
Kaiser Permanente South Sacramento Medical Center, Sacramento, California 95823, karen.l.murrell@kp.org

1. THE QUESTION

What is the **most efficient** way to handle patients waiting to see a doctor in the ER?

2. TWO OPTIONS

Queue Style 1

POOLED QUEUE

A single line in which patients wait to be seen by the next available ER physician. Most ERs use pooled queuing systems.

Queue Style 2

DEDICATED QUEUE

Every incoming patient is assigned to a specific physician's line.

3. THE FINDINGS

The team's studies showed that **dedicated queues led to faster throughput time**, for a 40-to-45-minute reduction of time in the ER.

9% decrease in waiting times

17% decrease in length of patients' stays

"This is counter to what traditional queuing theory would predict … with dedicated queues, physicians start feeling a greater sense of ownership and responsibility over those queues of patients. They're trying to actively think of ways to get people into the beds quicker, so they're doing things differently."
– Song 2017

Images Sourced from: https://magazine.wharton.upenn.edu/issues/spring-summer-2017/shorter-er-wait-times-through-smarter-queues/

But perhaps efficiency is not always the best solution…

Ryan Buell
@ryanbuell

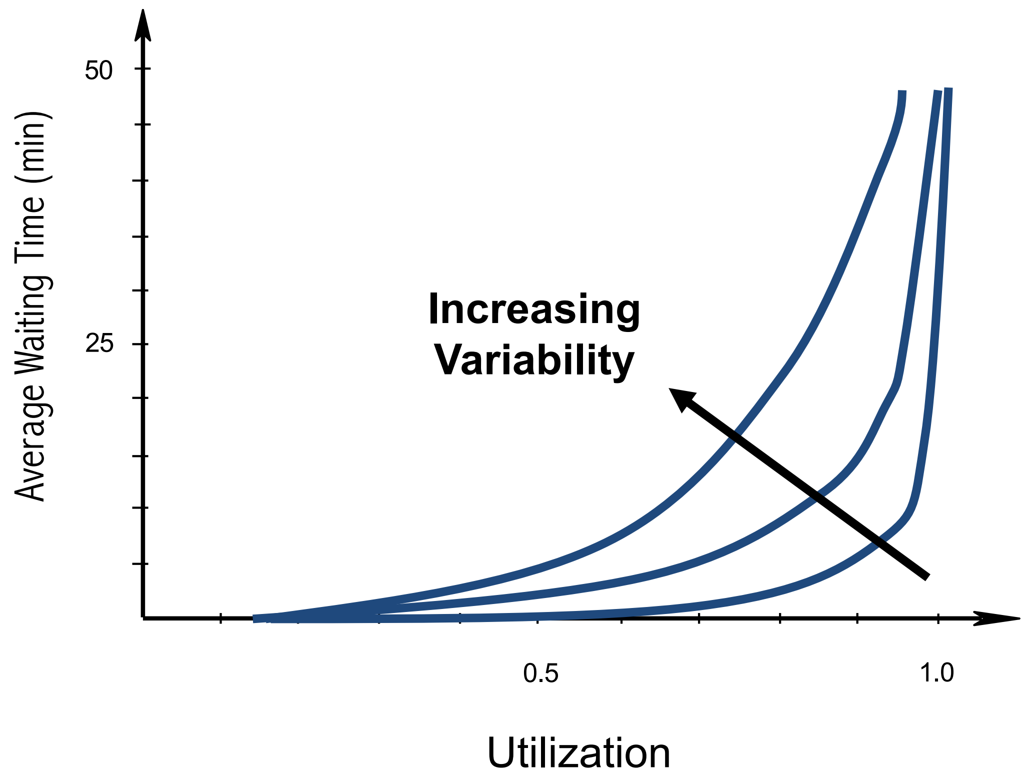
Slow checkout lanes! Customers and employees have different needs and preferences. Transparently designing solutions that better meet people where they are can set everyone up for an improved experience.

Maarten VdA @maartenvda · Jan 8

A Dutch supermarket chain introduced slow checkouts for people who enjoy chatting, helping many people, especially the elderly, deal with loneliness. The move has proven so successful that they installed the slow checkouts in 200 stores.

10:08 AM · Jan 11, 2023 · **2,533** Views

Impact of Variability



Ways to manage variability in processes

- Reduce arrival variability

- Reduce service time variability

Recap

- Variability is the norm, not the exception
- Variability leads to waiting times even though system is not capacity-constrained
- Waiting times tend to increase dramatically as the utilization of a process approaches 100%
- Benefits of pooling multiple queues
  - Reduce the waiting time with the same amount of resource
  - Use less resource to achieve the same waiting time
- Use the waiting time formula to:
  - Get a qualitative feeling of the system
  - Analyze specific recommendations/scenarios
- Managerial response to variability:
  - Understand where it comes from and eliminate what you can
  - Decide which part of the process should hold excess capacity

Next Class

- Inventory Models!
  - We'll start with the Newsvendor model on Monday/Tuesday next week
  - Please read the Ludo Case posted on Canvas

- Assignment #3 Due next Wed/Thurs
  - ***We'll cover the information you'll need to do it during next Monday/Tuesday's class***