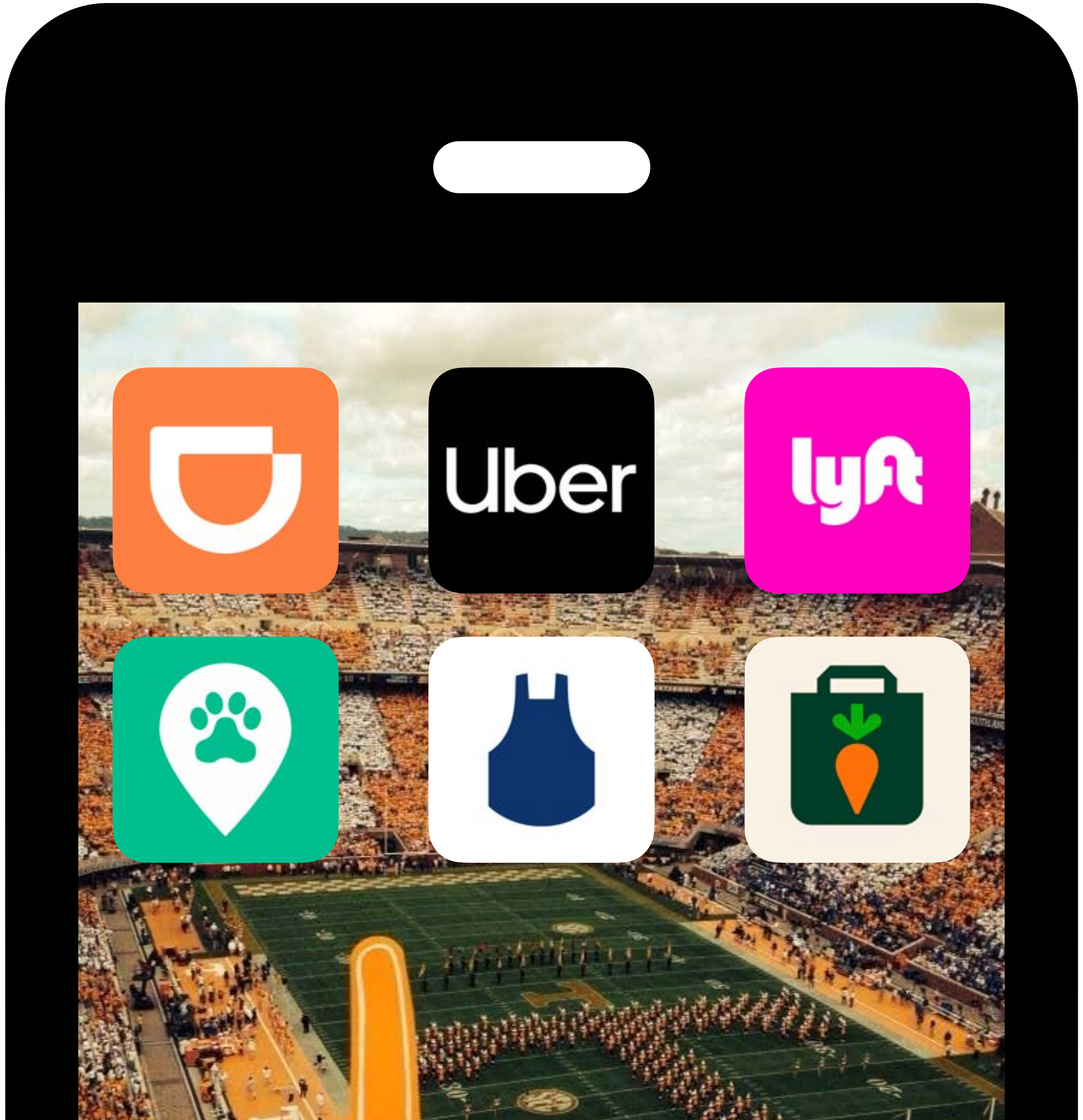
“on-demand service platform with impatient customers”

Driven by mobile apps

- **Customers** use an app to queue and eventually purchase a service for a **price**

- Service **providers** use an app interact with their customers and eventually receive a **wage**

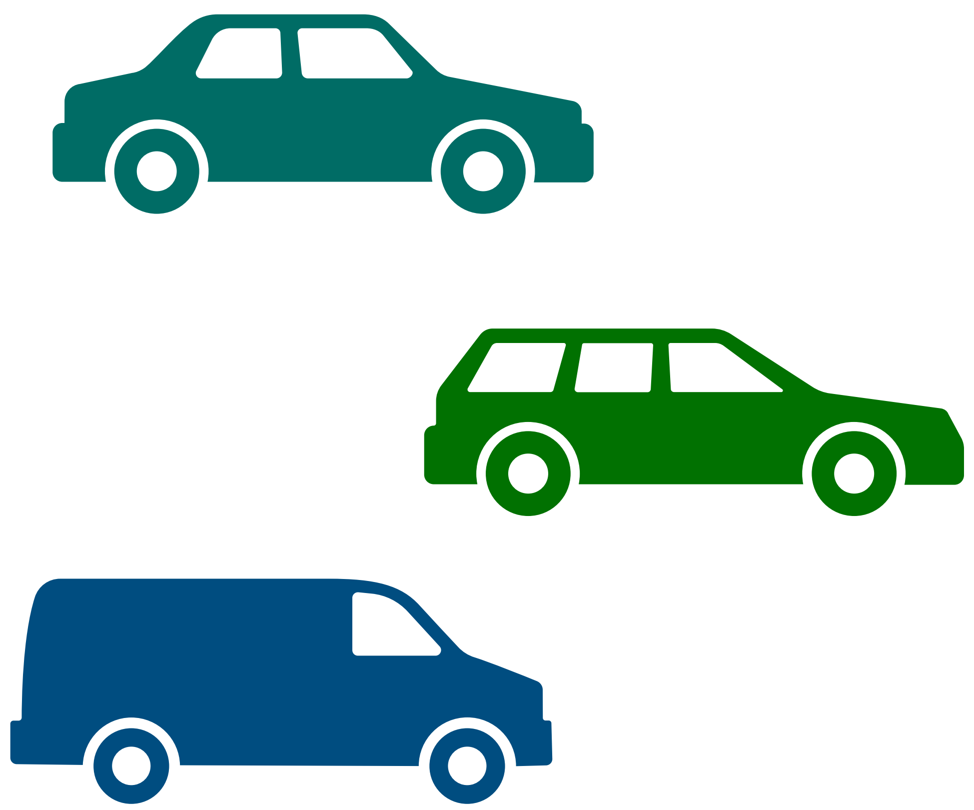
- Firms want to maximize profit, how to set **price** and **wage**?



Overview of the system being modeled

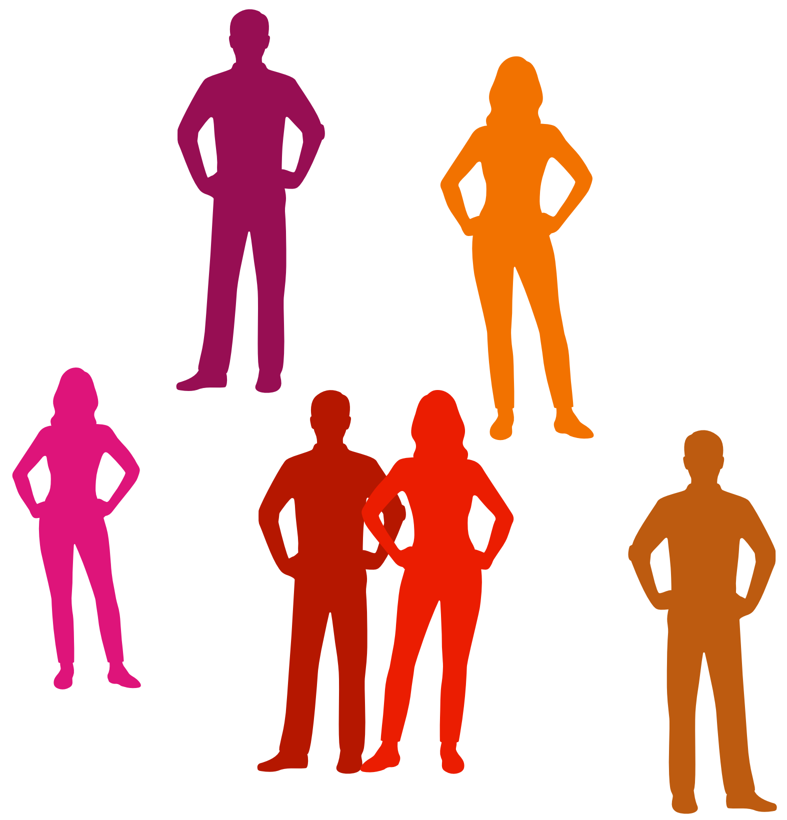
Coordinating supply and demand for on-demand services

The supply (independent providers)



Goal: Maximize earnings

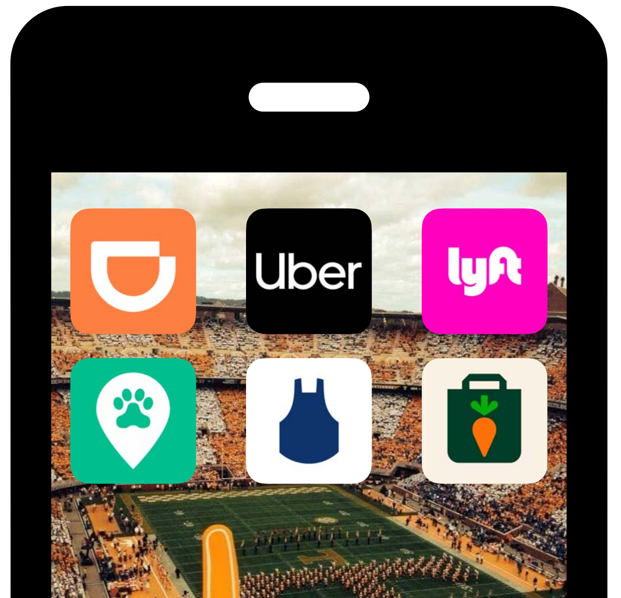
The demand (impatient customers)



Goal: Minimize cost and wait time

Depends on:
- wage rate ← **Decided by the firm** →

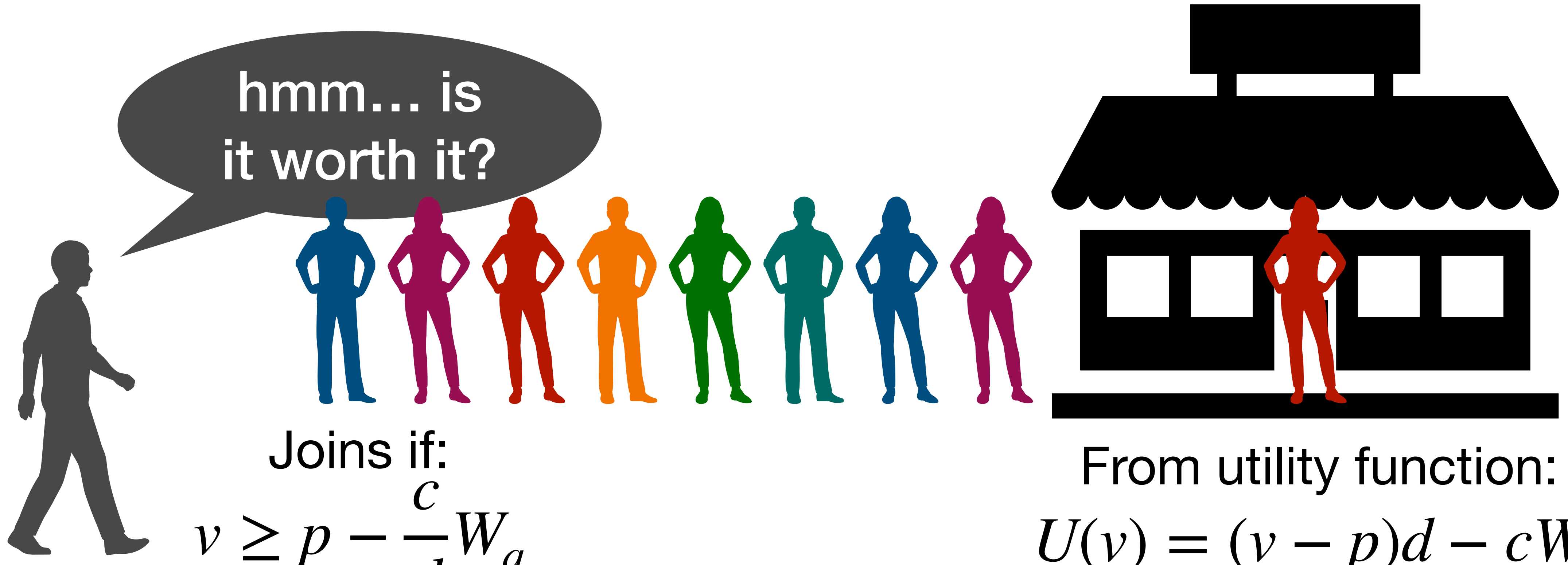
Depends on:
- price



Customers only act when utility > 0

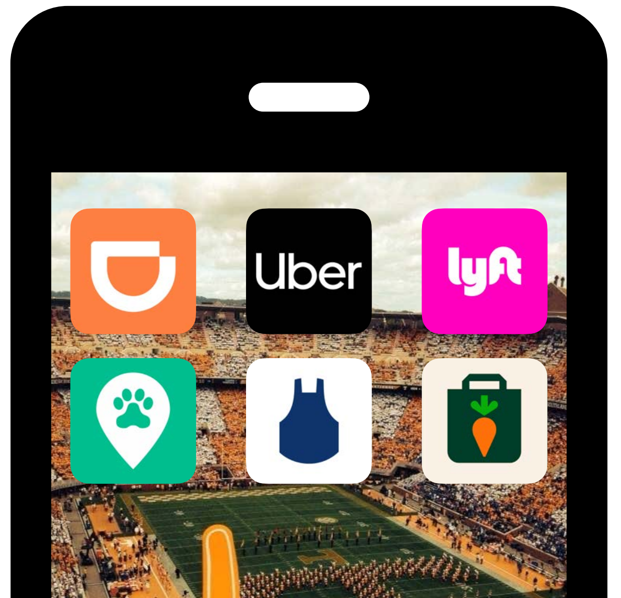
Classic Naor queueing ideas applied to $M/M/k$ queues

- **Customers** use an app to queue and eventually purchase a service for a **price**
  - Customers have different valuations ($v \sim F(\cdot)$) of the service
  - Customers queue if their valuation (v) of the service is greater than the price (p) + cost of waiting ($\frac{c}{d}W_q$)
    - W_q is average wait time
    - c is cost of waiting per unit time
    - d is average number of service units (e.g. average km per ride)



Joins if:
$v \geq p - \frac{c}{d}W_q$

From utility function:
$U(v) = (v - p)d - cW_q$

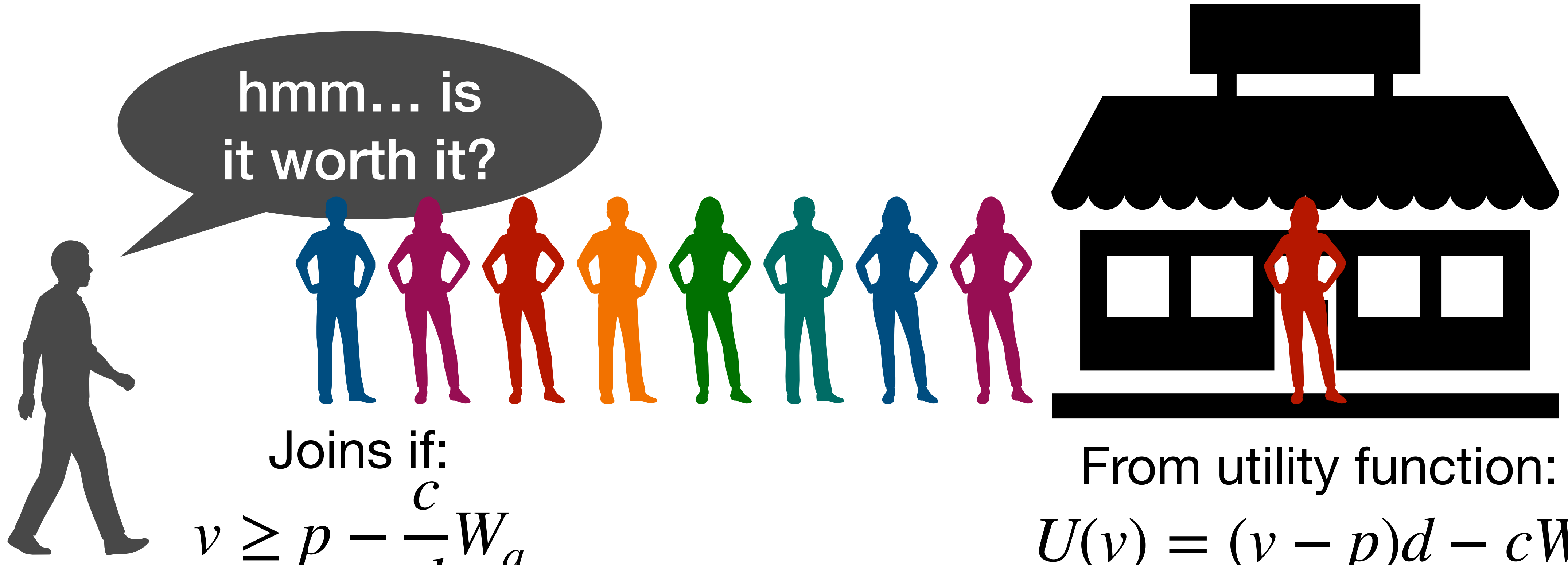


We can adjust price to realize a service level (s)

Classic Naor queueing ideas applied to $M/M/k$ queues

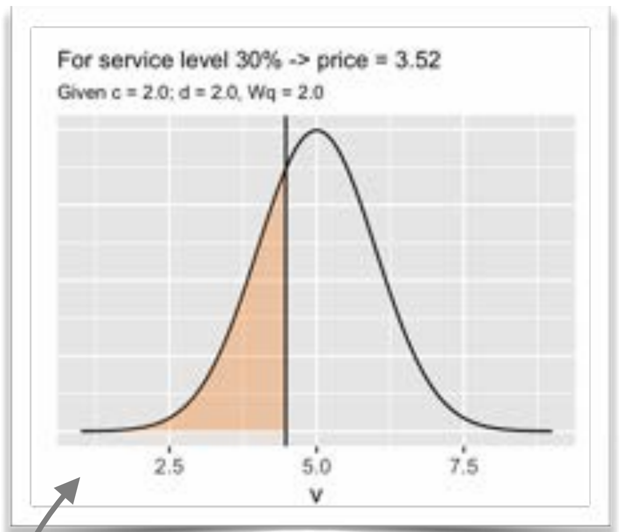
- Depending on the price (p), a different number (λ) of the possible $(\bar{\lambda})$ customers will request ($\frac{\lambda}{\bar{\lambda}}$ is the request rate)

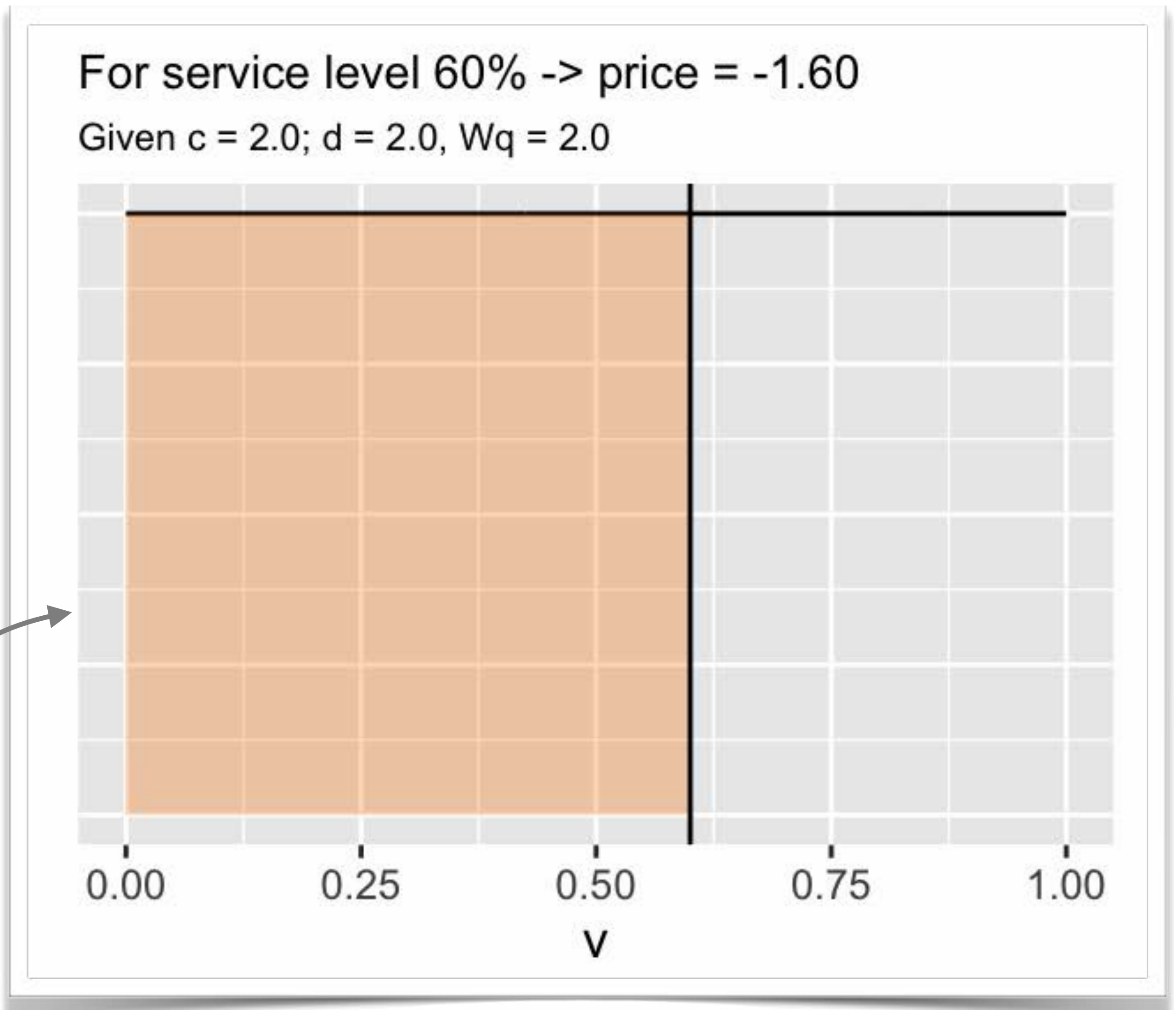
- Customers have different valuations of the service (v; the minimum rate at which they'll participate)



Joins if:
$v \geq p - \frac{c}{d} W_q$

From utility function:
$U(v) = (v - p)d - cW_q$





modeled as uniform, but insights extend to other dists

$$p = F^{-1}\left(1 - \frac{\lambda}{\bar{\lambda}}\right) - \frac{c}{d} W_q$$

How to set price (p) based on target service level ($s = \frac{\lambda}{\bar{\lambda}}$)

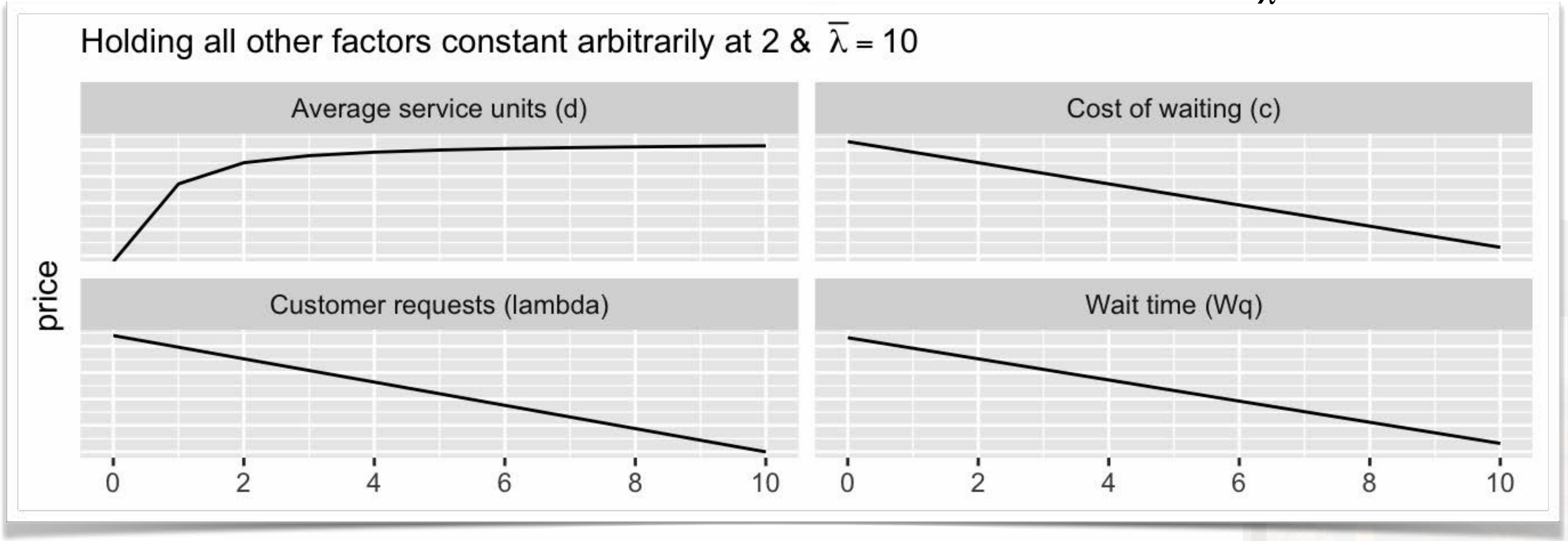
The effects of inputs on price

Classic Naor queueing ideas applied to $M/M/k$ queues

- c (cost of waiting) increases, we charge less
  - The cost of waiting per unit of service $(c/d)W_q$ increases with c, if we kept same price we would be over priced and we'd be below target service.
- d (avg service units) increases, we charge more
  - The cost of waiting per unit of service $(c/d)W_q$ decreases with d, if we kept same price we would be under priced and we'd be above target service rate.
- W_q (avg wait time) increases, we charge less
  - The cost of waiting per unit of service $(c/d)W_q$ increases with c, if we kept same price we would be over priced and we'd be below target service.
- λ increases (aka service rate increases), we charge less
  - An increase in λ leads to an increased wait time W_q

$$p = F^{-1}\left(1 - \frac{\lambda}{\bar{\lambda}}\right) - \frac{c}{d}W_q$$

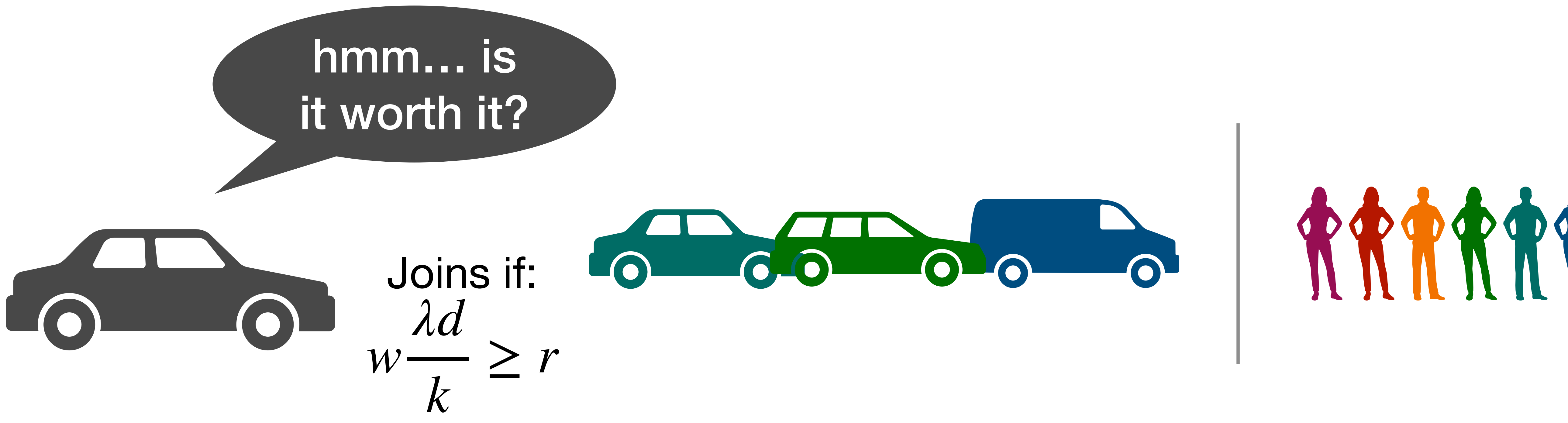
How to set price (p) based on target service level ($s = \frac{\lambda}{\bar{\lambda}}$)

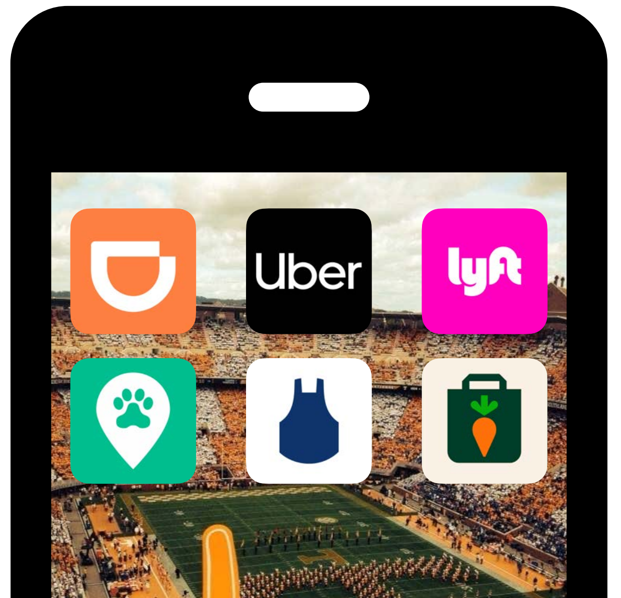


Providers only enter when it's worth their time

Higher wages (w) and utilization (ρ) leads to more providers

- Service **providers** use an app interact with their customers and eventually receive a **wage**
  - Providers have different reservation rates for which they'll participate ($r \sim G(\cdot)$) of the service
  - Providers join if their reservation rate (r) is less than the wage rate (w) $\times$ expected units of service utilized
    - λ is number of requesting customers
    - k is number of participating providers
    - d is average number of service units (e.g. average km per ride)

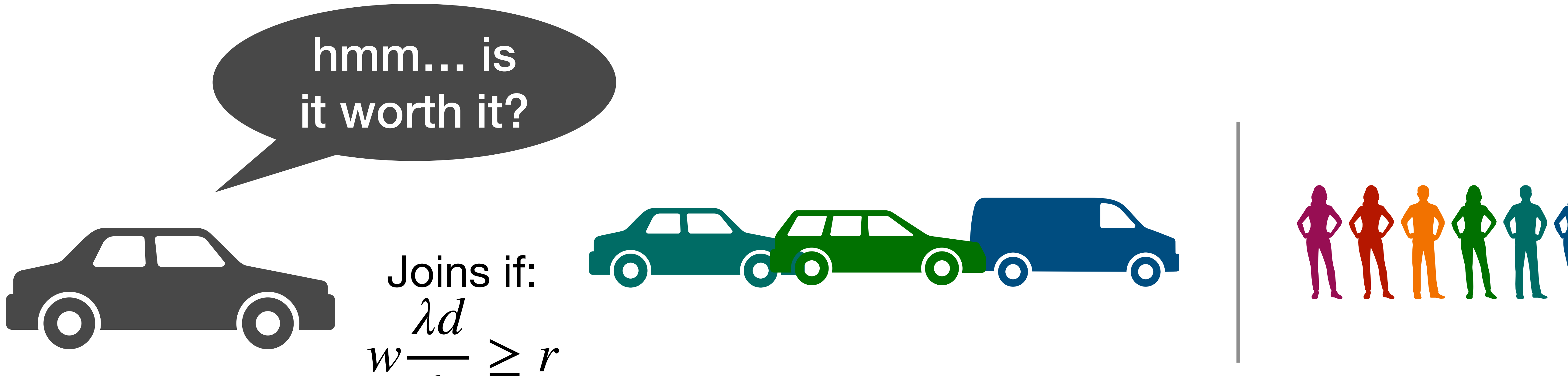




Providers only enter when it’s worth their time

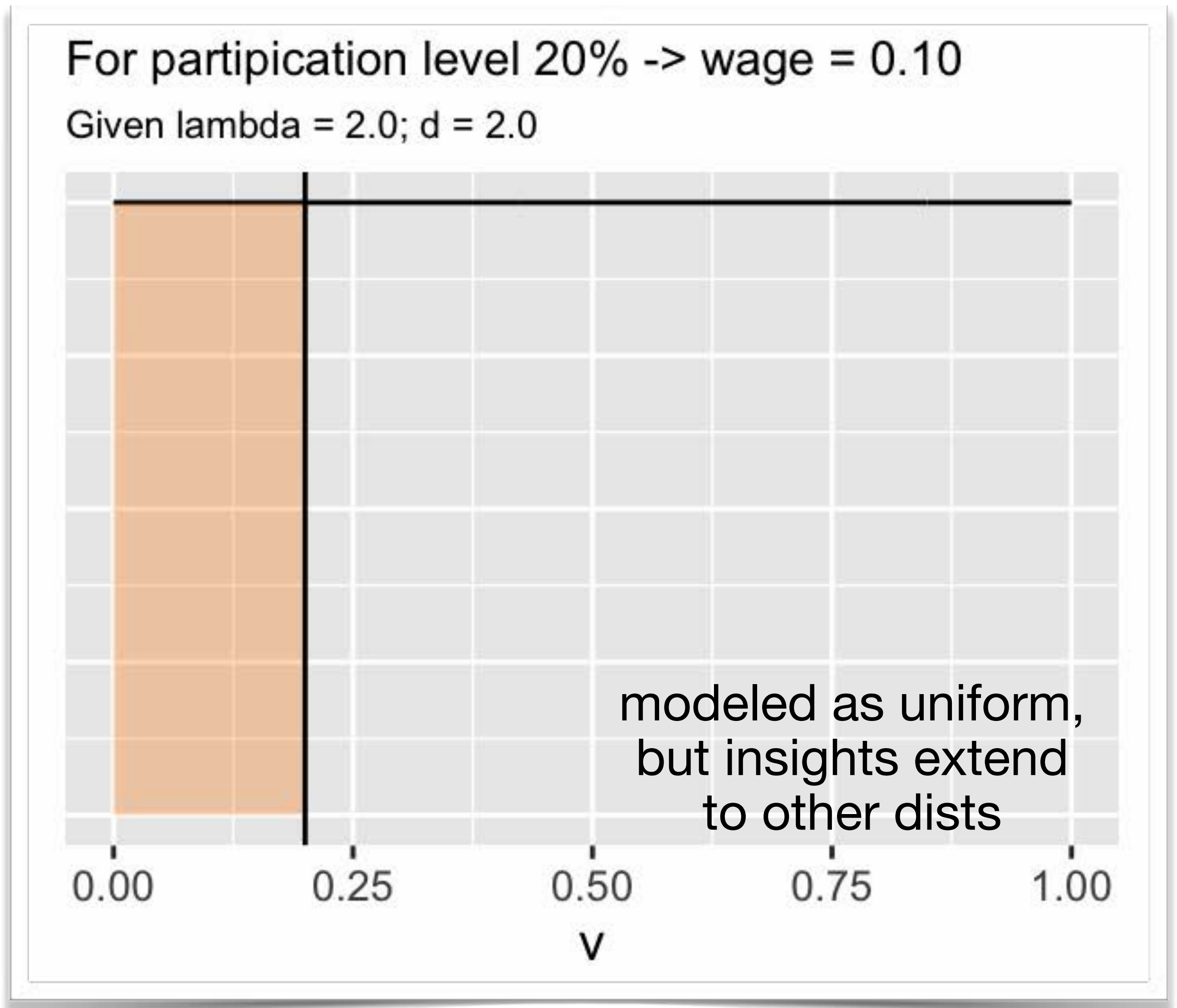
Higher wages (w) and utilization (ρ) leads to more providers

- Depending on the wage (w), a different number (k) of the possible (K) providers will participate ($\frac{k}{K}$ is the partication rate)
- Providers have different reservation rates (r; the minimum rate at which they’ll participate)



Joins if:

$$w\frac{\lambda d}{k} \geq r$$



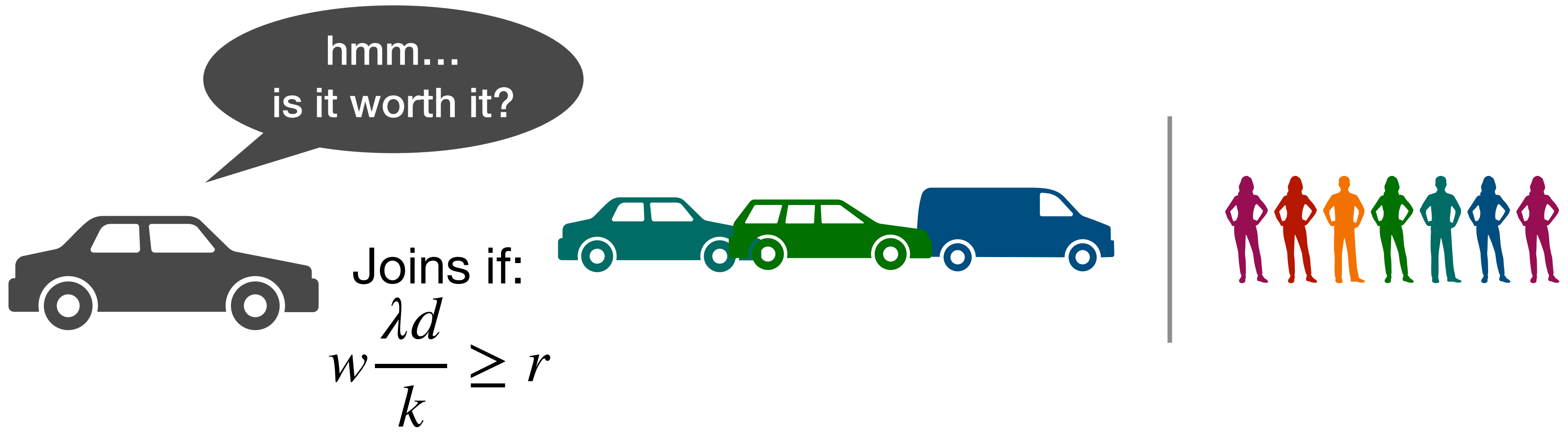
$$w = G^{-1}(\frac{k}{K})\frac{k}{\lambda d}$$

How to set price (w) based on target participation rate ($\beta = \frac{k}{K}$)

Providers only enter when it’s worth their time

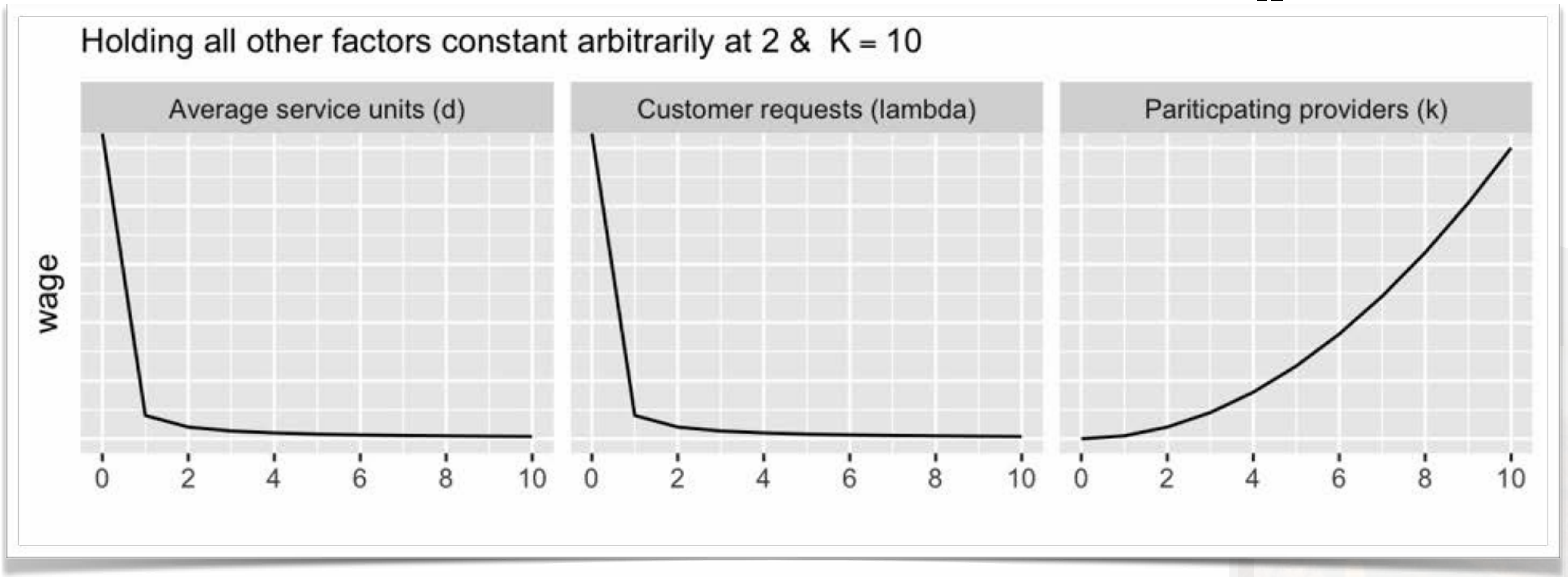
Higher wages (w) and utilization (ρ) leads to more providers

- d (avg service units) increases, we pay less
  - Average utilization & earnings increase with d, if we kept same wage we would be over paying and we'd be above target participation.
- λ increases (aka requesting customers), we pay less
  - Average utilization & earnings increase with λ, if we kept same wage we would be over paying and we'd be above target participation.
- k increases (aka participation rate increases), we pay more
  - Average utilization & earnings decrease as more providers enter, if we want more providers we need to pay more.



$$w = G^{-1}(\frac{k}{K})\frac{k}{\lambda d}$$

How to set price (w) based on target participation rate ($\beta = \frac{k}{\bar{K}}$)

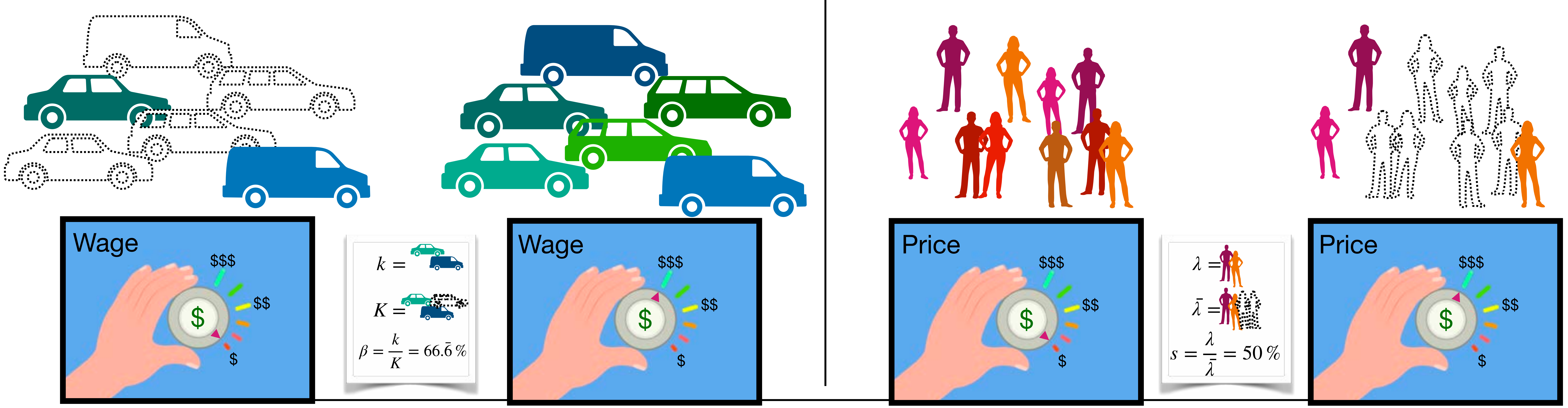


Overview of the system being modeled

Coordinating supply and demand for on-demand services

The supply (independent providers)

The demand (impatient customers)

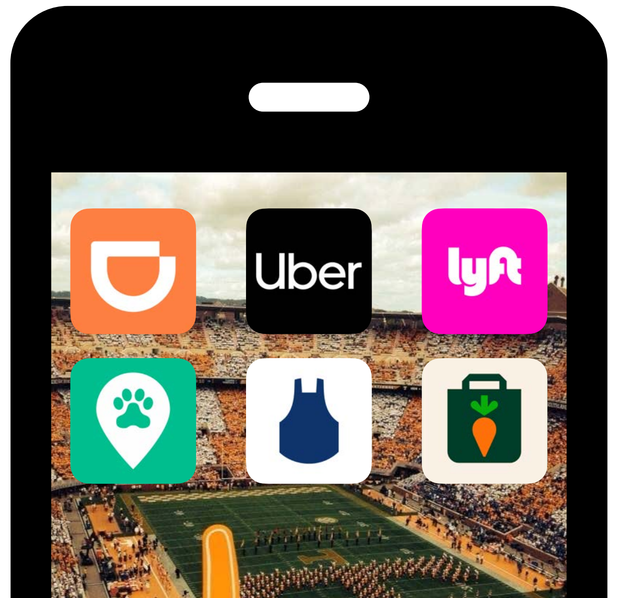


Depends on:

- wage rate ← **Decided by the firm** → price
- utilization (demand) ← **Heavily intertwined** → supply

Depends on:

- price
- supply

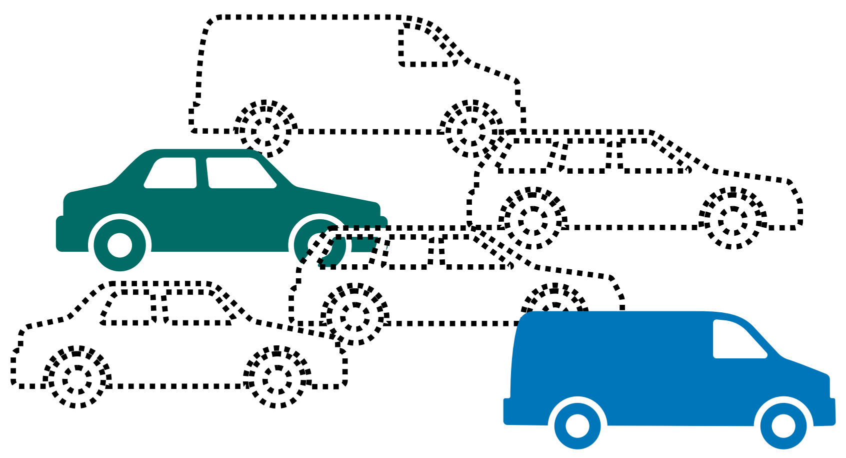


Overview of the system being modeled

Coordinating supply and demand for on-demand services

The supply (independent providers)

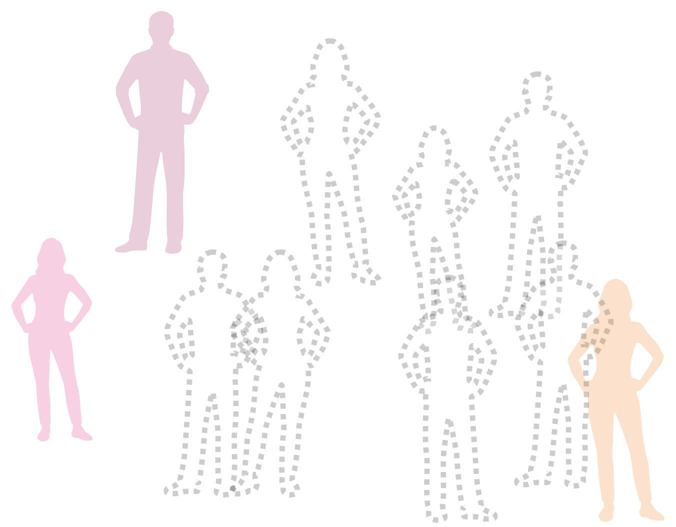
demand >> supply



The demand (impatient customers)

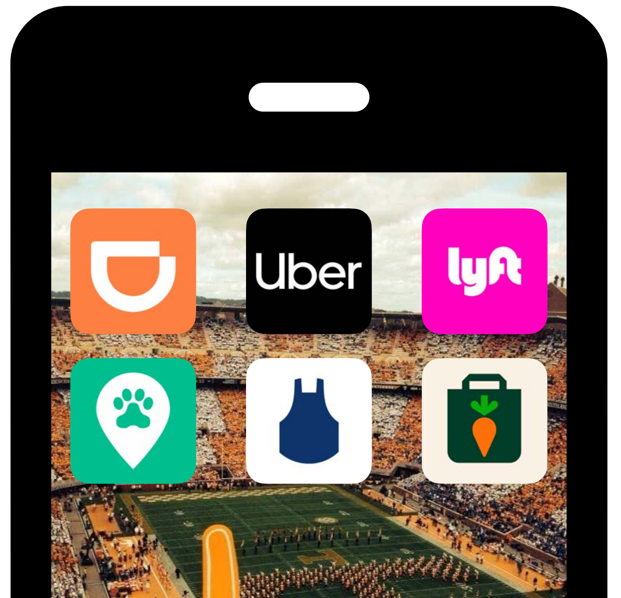
This is too long to wait.. I'll use a different service

I don't make money sitting around idle...



supply >> demand

Depends on:		Depends on:
• wage rate	← **Decided by the firm** →	• price
• utilization (demand)	← **Heavily intertwined** →	• supply



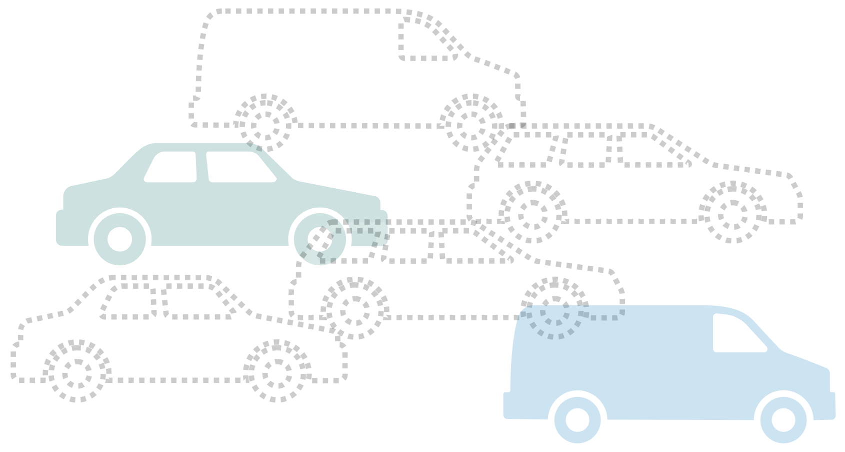
Overview of the system being modeled

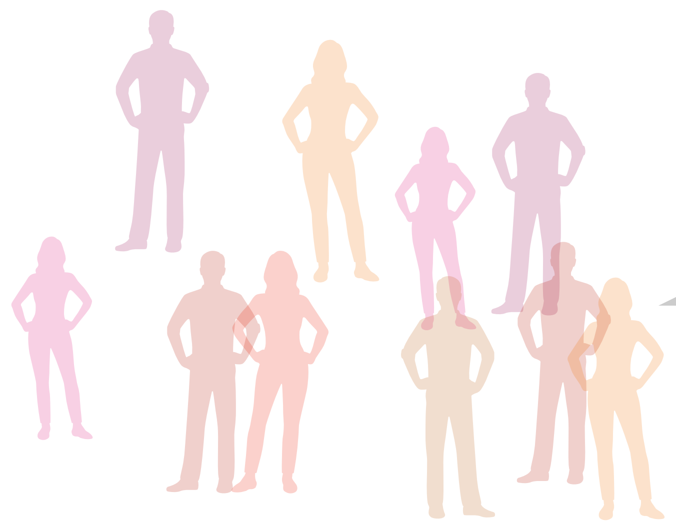
Coordinating supply and demand for on-demand services

The supply (independent providers)

The demand (impatient customers)

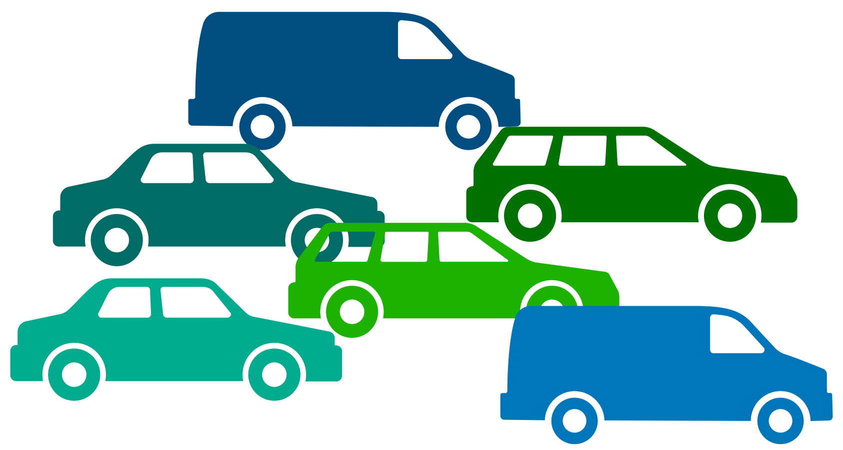
demand >> supply

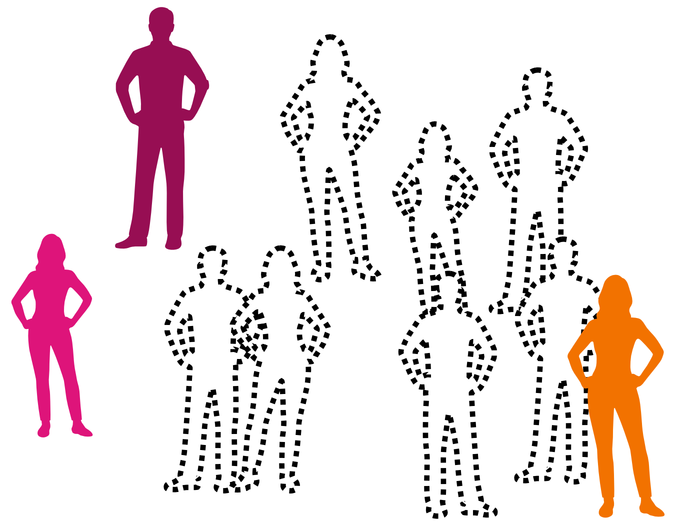




This is too long to wait.. I'll use a different service

I don't make money sitting around idle…





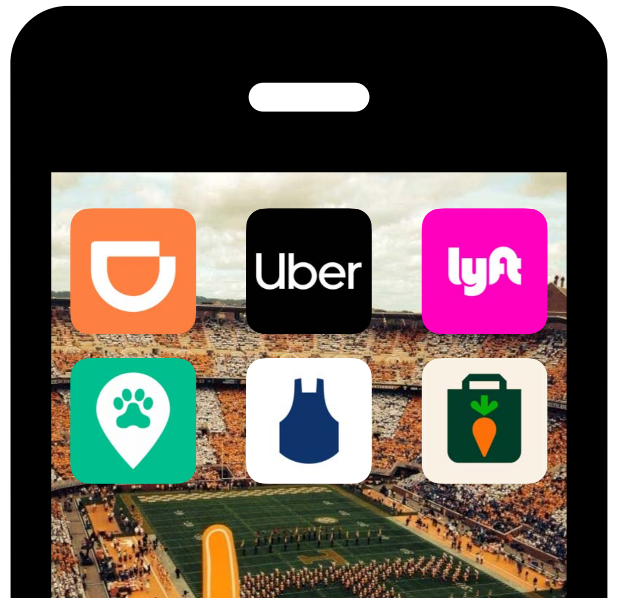
supply >> demand

Depends on:
- wage rate
- utilization (demand)

Decided by the firm

Heavily intertwined

Depends on:
- price
- supply



The firms want to maximize profit

Find optimal requests (λ) and optimal providers (k) to max profit (π)

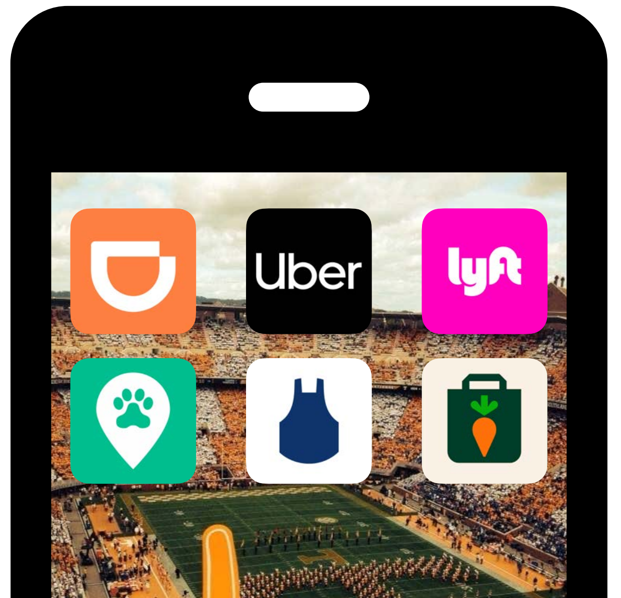
Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

$$\max \pi = \lambda(p - w)d$$

$$\max \pi = \lambda(\frac{w}{\alpha} - w)d$$

From now on, wage (w) will be a proportion (α) of price (p)

$$w = \alpha p$$



The firms want to maximize profit

Find optimal requests (λ) and optimal providers (k) to max profit (π)

Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

$$\max \pi = \lambda(p - w)d$$

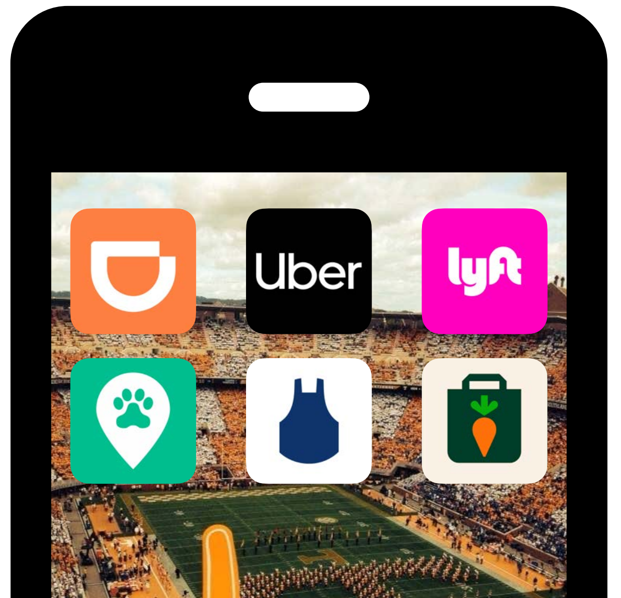
$$\max \pi = \lambda\left(\frac{w}{\alpha} - w\right)d$$

After manipulation we end up with this objective

$$\max_{k,\lambda} \pi(k, \lambda) \equiv \frac{k^2(1-\alpha)}{K\alpha}$$

Clear to see we want to maximize participation rate (k)

From this we can find optimal k^* & λ^*, and then use price p^* & wage w^*



The firms want to maximize profit

Find optimal requests (λ) and optimal providers (k) to max profit (π)

Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

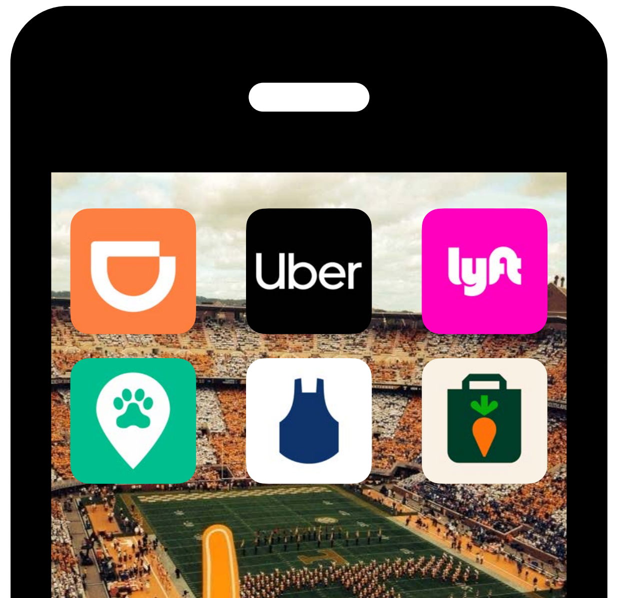
$$\max_{k,\lambda} \pi(k, \lambda) \equiv \frac{k^2(1-\alpha)}{K\alpha}$$

From this we can find optimal k^* & λ^*, and then use price p^* & wage w^*

This table shows results for all parameters fixed except the ones shown

Not clearly mentioned in paper, these are not unique solutions; they appear to prefer solutions with higher p^* and shorter W_q

	W_q is given by exact formula (9)			
$\bar{\lambda}$	k^*	λ^*	p^*	π^*
10	7	2.71	0.72	0.98
20	10	5.79	0.69	2.00
30	11	6.20	0.78	2.42
40	12	7.14	0.81	2.88
50	13	8.32	0.81	3.38
60	14	9.80	0.80	3.92
70	14	9.29	0.84	3.92
80	15	11.16	0.81	4.50
90	15	10.62	0.85	4.50
100	15	10.36	0.87	4.50



The firms want to maximize profit

Find optimal requests (λ) and optimal providers (k) to max profit (π)

Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

$$\max_{k,\lambda} \pi(k,\lambda) \equiv \frac{k^2(1-\alpha)}{K\alpha}$$

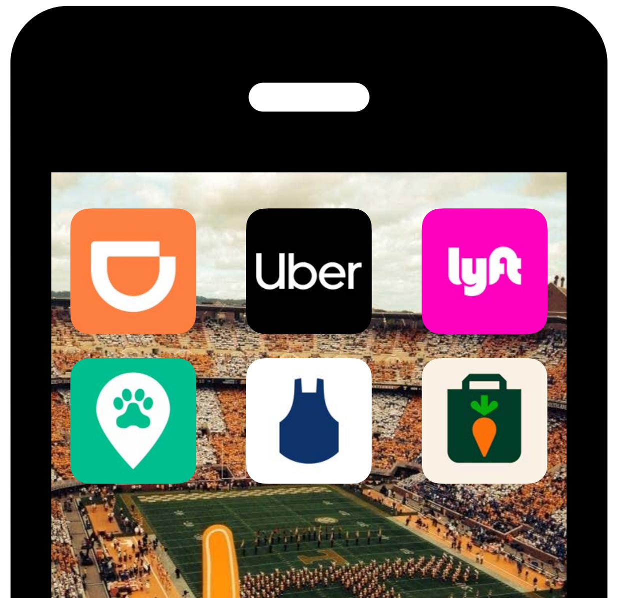
From this we can find optimal k^* & λ^*, and then use price p^* & wage w^*

This table shows results for all parameters fixed except the ones shown

Not clearly mentioned in paper, these are not unique solutions; they appear to prefer solutions with higher p^* and shorter W_q

	Shown in Paper	Alternative solution
k	7.000	7.000
K	50.000	50.000
λ	2.710	4.730
$\bar{\lambda}$	10.000	10.000
ρ	0.387	0.676
W_q	0.005	0.116
p^*	0.724	0.411
w^*	0.362	0.207
α	0.500	0.500
profit	0.980	0.980

It might make sense to implement both strategies for vertically differentiated services



The firms want to maximize profit

Find optimal requests (λ) and optimal providers (k) to max profit (π)

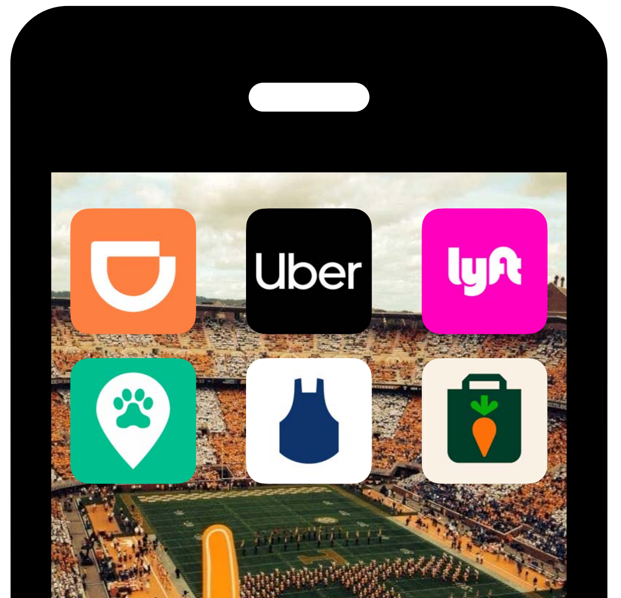
Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

$$\max_{k,\lambda} \pi(k, \lambda) \equiv \frac{k^2(1-\alpha)}{K\alpha}$$

From this we can find optimal $k*$ & $\lambda*$, and then use price $p*$ & wage $w*$

This table shows results for all parameters fixed except the ones shown

	W_q is given by exact formula (9)			
$\bar{\lambda}$	k^*	λ^*	p^*	π^*
10	7	2.71	0.72	0.98
20	10	5.79	0.69	2.00
30	11	6.20	0.78	2.42
40	12	7.14	0.81	2.88
50	13	8.32	0.81	3.38
60	14	9.80	0.80	3.92
70	14	9.29	0.84	3.92
80	15	11.16	0.81	4.50
90	15	10.62	0.85	4.50
100	15	10.36	0.87	4.50



The firms want to maximize profit

It's untractable to find analytical results with exact W_q

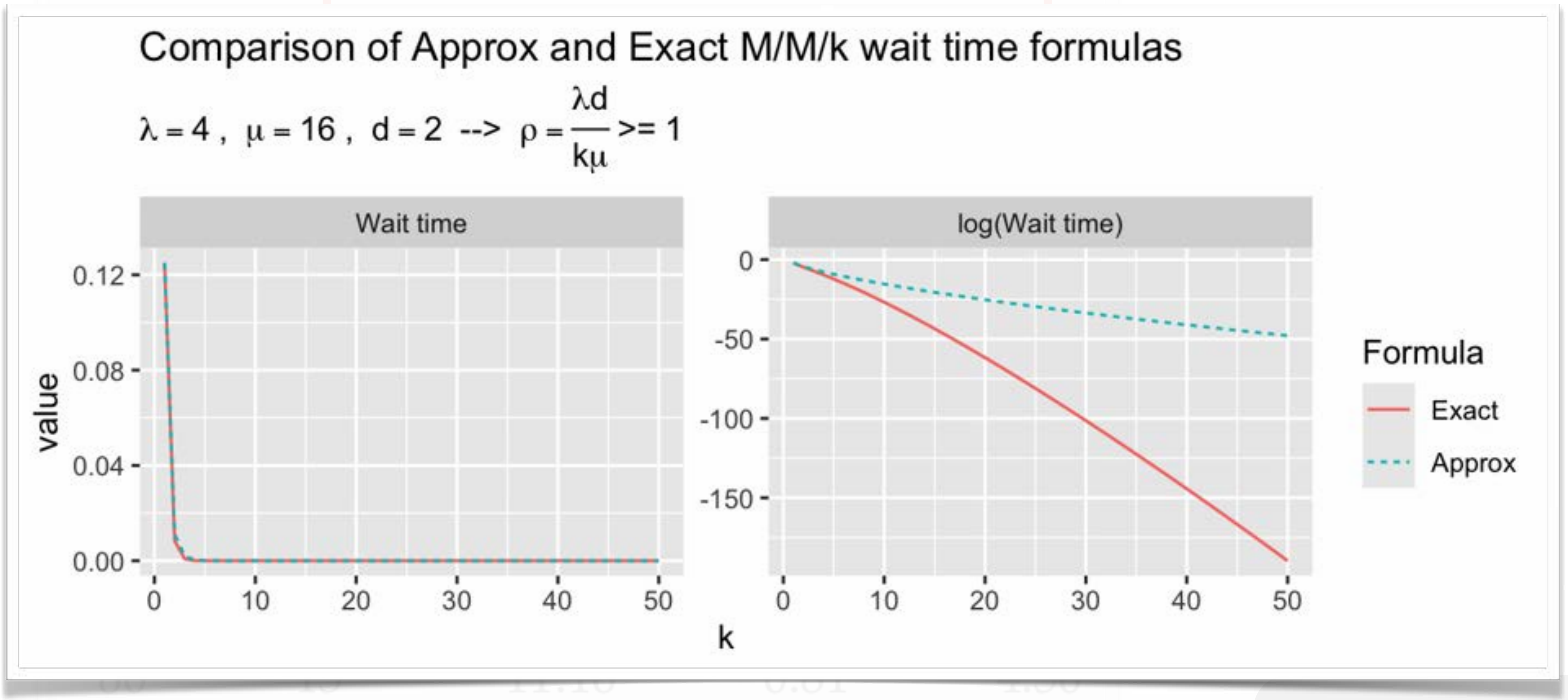
To calculate p^* we need W_q, which is defined as:

$$W_q = \frac{1}{1 + \left(\frac{k!(1-\rho)}{k^k\rho^k}\right)\sum_{i=0}^{k-1}\frac{k^i\rho^i}{i!}}\left[\frac{\rho}{\lambda(1-\rho)}\right]$$

This is obviously not easy to work with for analytically optimal results.

An approximation is given by:

$$W_q = \frac{\rho^{\sqrt{2(k+1)}}}{\lambda(1-\rho)}$$



They are identical for $k = 1$, and then diverge with approximation overestimating W_q

(still a good, commonly used approximation)

The firms want to maximize profit

How inputs affect the optimal solution parameters

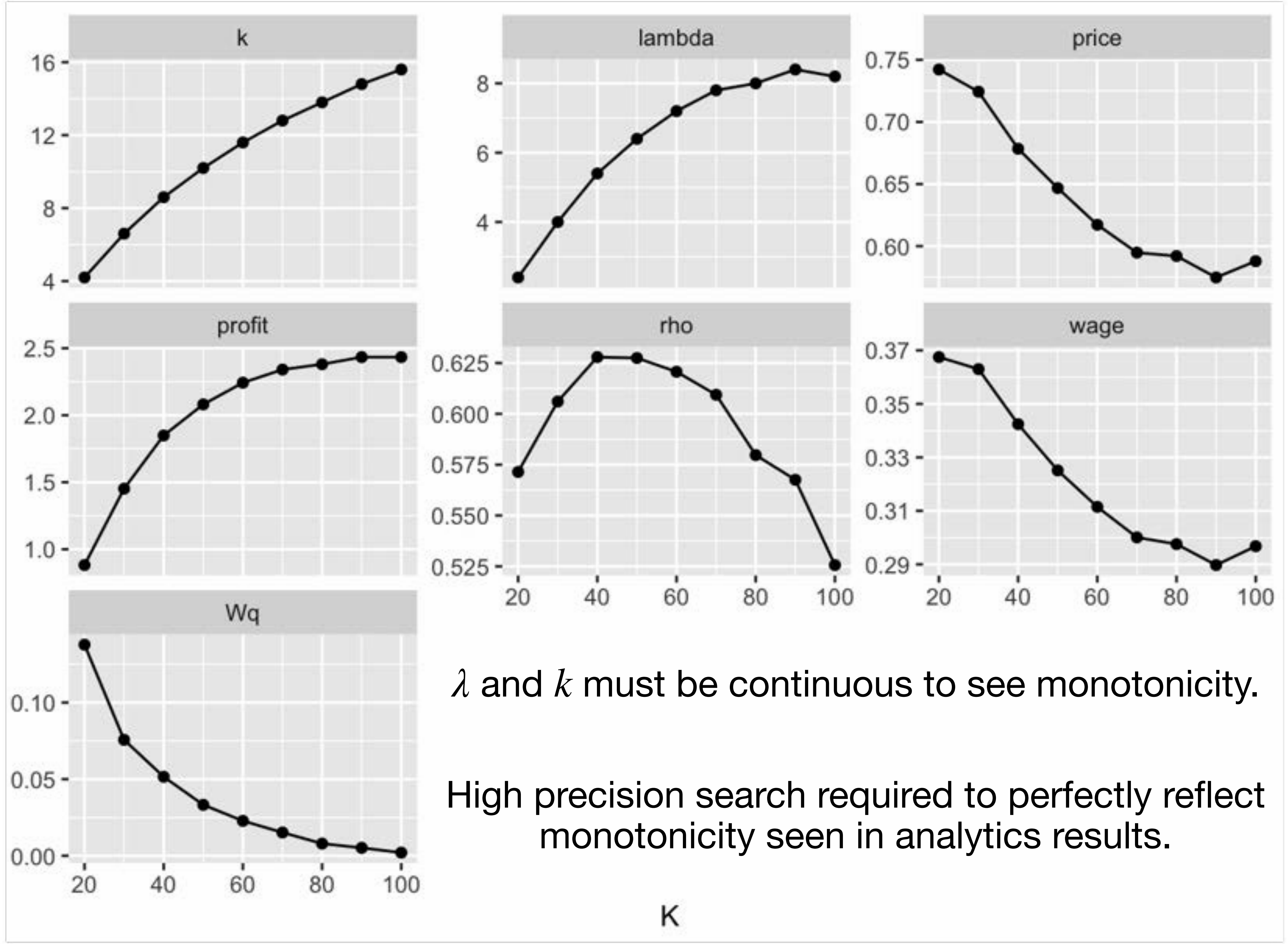
Paper's analytical results using approximate W_q

Table 5. Impact of Model Parameters on s^*, k^*, W_q^*, λ^*, and ρ^*

Variable	s^*	k^*	W_q^*	λ^*	ρ^*
K	↑	↑	↓	↑	×
μ	↑	×	↓	↑	×
c	↓	×	↓	↓	↓
$\bar{\lambda}$	↓	↑	↑	↑	↑
d	↓	↑	↑	↓	↑

↑ (increasing); ↓ (decreasing); × (nonmonotonic).

My results optimizing with exact W_q



λ and k must be continuous to see monotonicity.

High precision search required to perfectly reflect monotonicity seen in analytics results.

The firms want to maximize profit

How inputs affect the optimal solution parameters

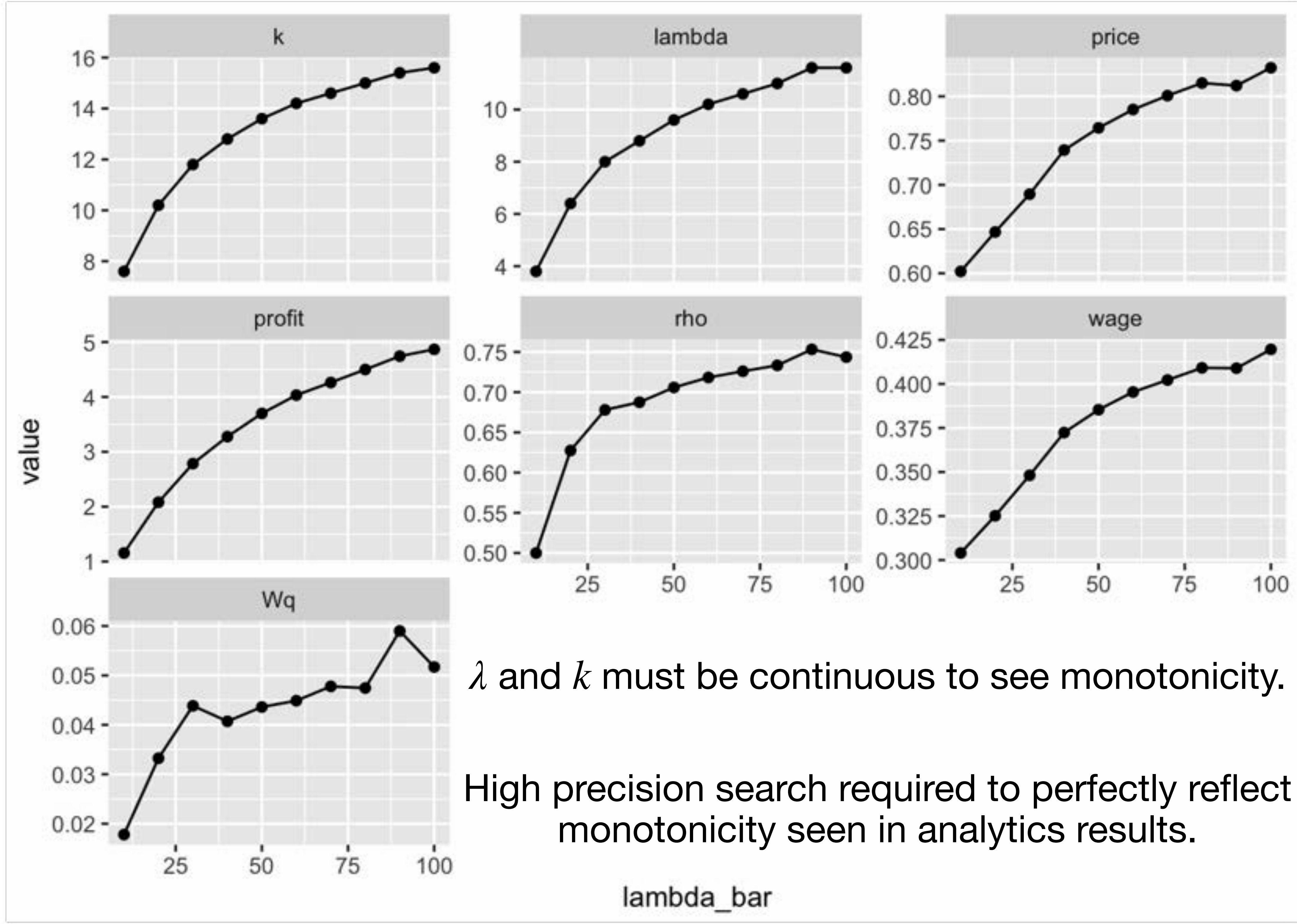
Paper's analytical results using approximate W_q

Table 5. Impact of Model Parameters on s^*, k^*, W_q^*, λ^*, and ρ^*

Variable	s^*	k^*	W_q^*	λ^*	ρ^*
K	↑	↑	↓	↑	×
μ	↑	×	↓	↑	×
c	↓	×	↓	↓	↓
$\bar{\lambda}$	↓	↑	↑	↑	↑
d	↓	↑	↑	↓	↑

↑ (increasing); ↓ (decreasing); × (nonmonotonic).

My results optimizing with exact W_q



λ and k must be continuous to see monotonicity.

High precision search required to perfectly reflect monotonicity seen in analytics results.

The firms want to maximize profit

How inputs affect the optimal solution parameters

Paper's analytical results using approximate W_q

Table 5. Impact of Model Parameters on s^*, k^*, W_q^*, λ^*, and ρ^*

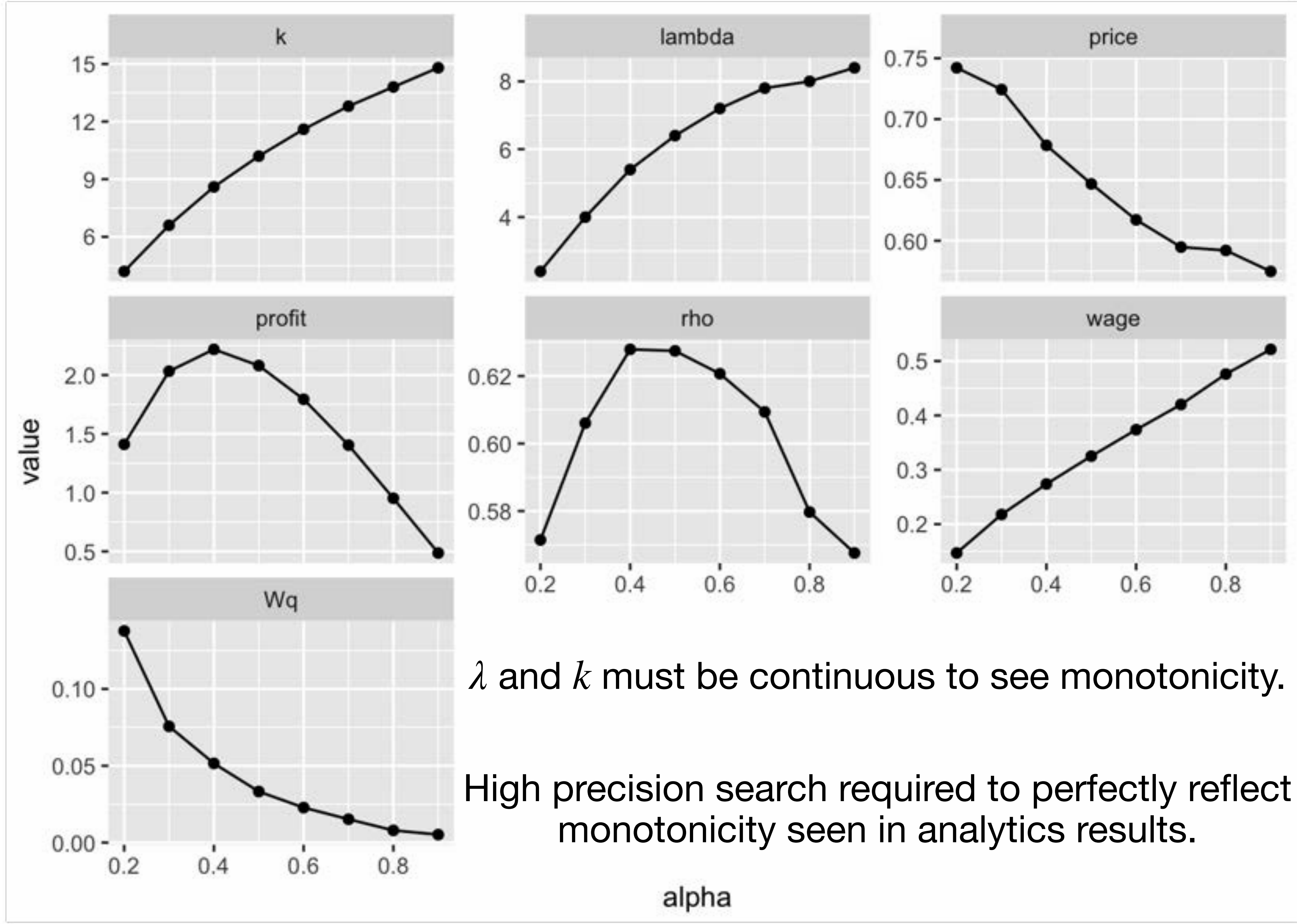
Variable	s^*	k^*	W_q^*	λ^*	ρ^*
K	↑	↑	↓	↑	×
μ	↑	×	↓	↑	×
c	↓	×	↓	↓	↓
$\bar{\lambda}$	↓	↑	↑	↑	↑
d	↓	↑	↑	↓	↑

↑ (increasing); ↓ (decreasing); × (nonmonotonic).

Not shown in table is wage's proportion of price (α).

Recall: $w = \alpha p$

My results optimizing with exact W_q



λ and k must be continuous to see monotonicity.

High precision search required to perfectly reflect monotonicity seen in analytics results.

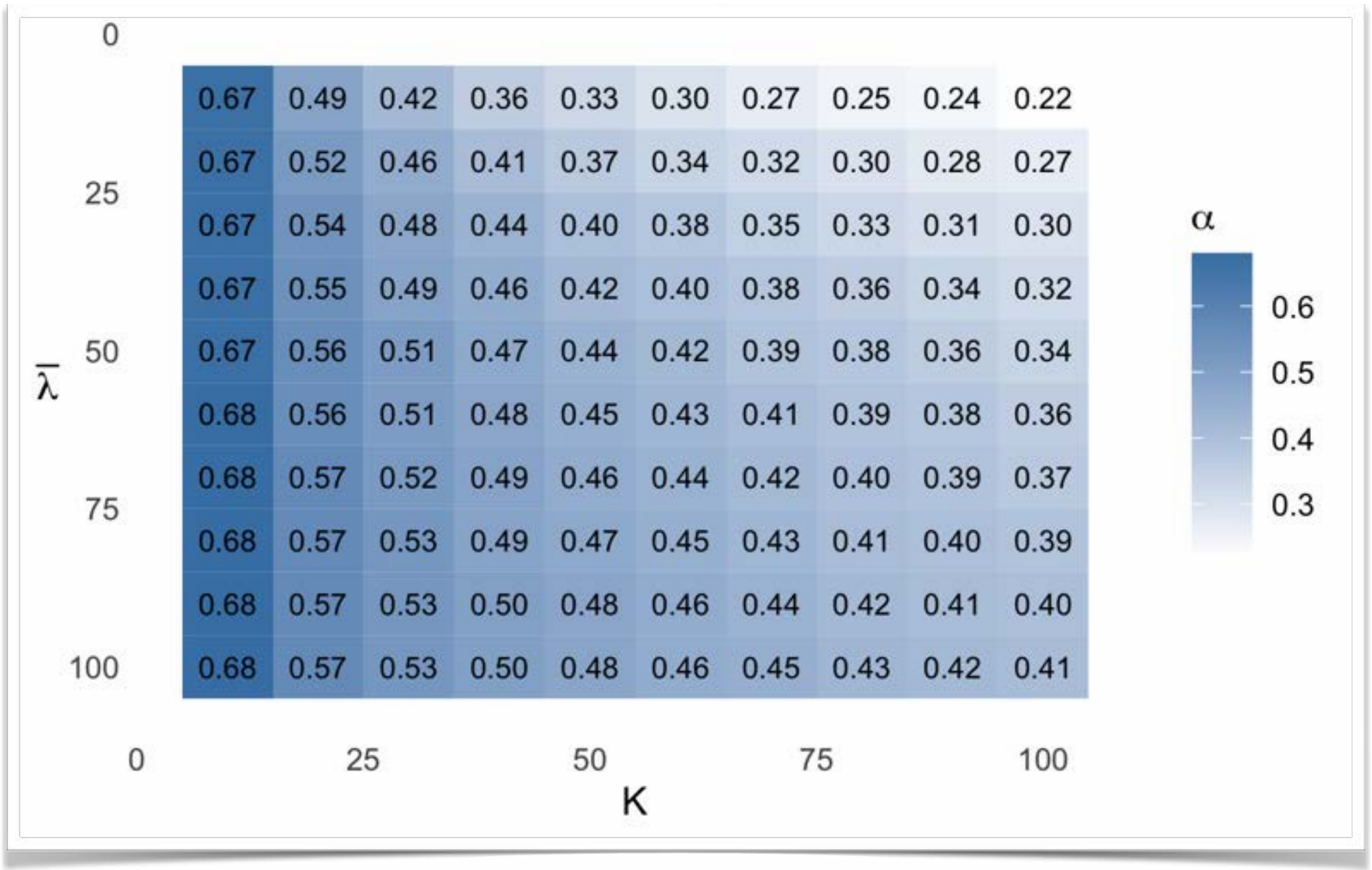
The firms want to maximize profit

How to set payout percentage (α)

re-colored

Paper's analytical results using approximate W_q to find optimal α for given:

- K (count of potential providers)
- $\bar{\lambda}$ (count of potential customers)

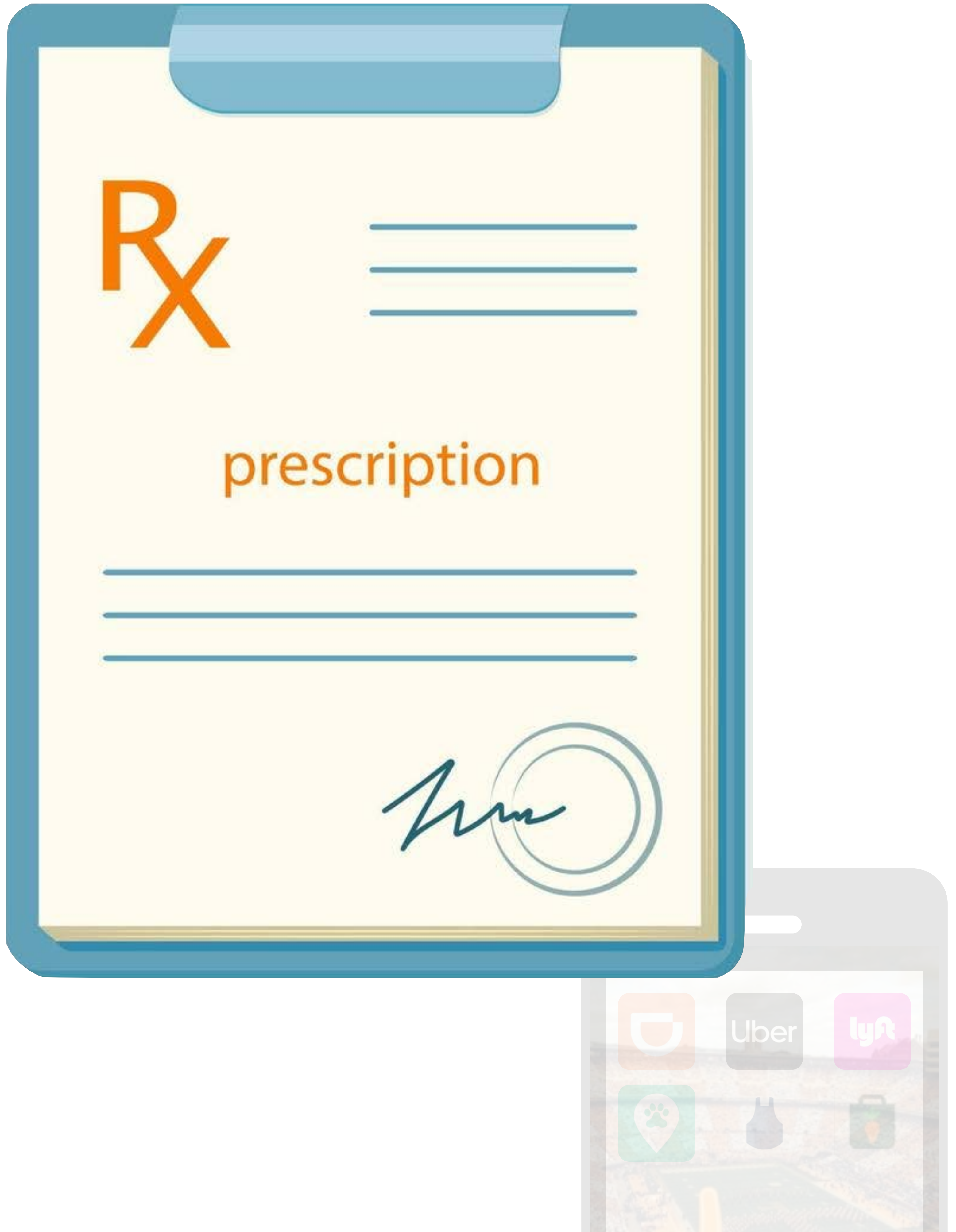


- When there are few potential providers (K), entice them with higher percentage of earnings (α)
- When there are many potential customers ($\bar{\lambda}$), use a higher percentage (α) to encourage more providers to participate and ensure wait times (W_q) stay down

Main insights

How we can apply model insights

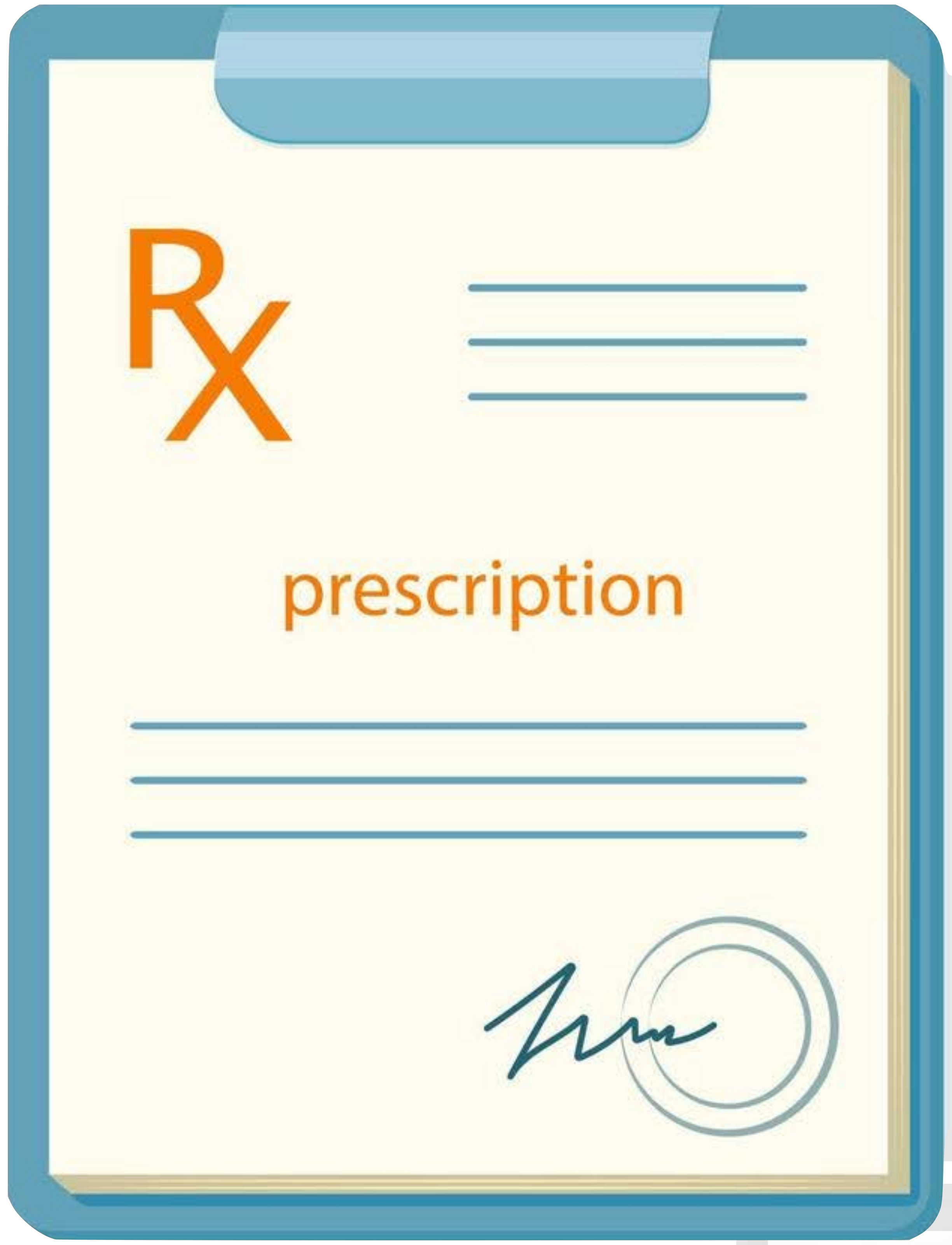
- **Platform Strategy for Providers and Service Speed**
  - As the potential number of providers (K) or service speed (μ) increases, the platform should reduce the wage rate to increase profits
  - The optimal price may increase initially with number of providers due to waiting time reductions with higher utilization but decreases later as the queueing effect diminishes at lower utilization.
- **Queueing Effect on Price with Small and Large number of K**
  - For small K, higher supply reduces waiting time significantly, leading to an increase in price.
  - For large K, waiting time reductions are marginal, so price decreases to stimulate demand.
  - This non-monotonic behavior in optimal price is attributed to nonlinear queueing effect.

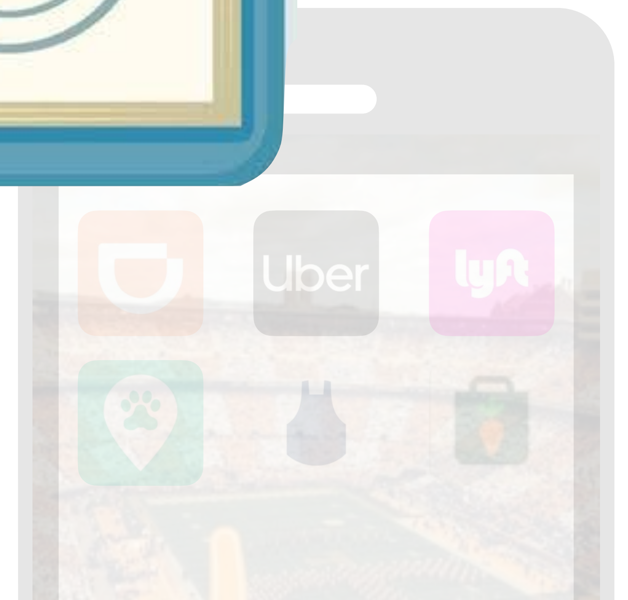


Main insights

How we can apply model insights

- **Effect of Waiting Cost (c) on Wage and Price**
  - As c increases, the platform should raise wages to attract more providers, reducing profits.
  - Price may increase initially with c due to demand reductions and waiting time improvements but decrease when c is high as marginal waiting time reductions diminish.
- **Price and Wage Adjustments for Demand and Service Units**
  - The platform should increase price and wage as customer demand rate or average service units increase, leading to higher profits.

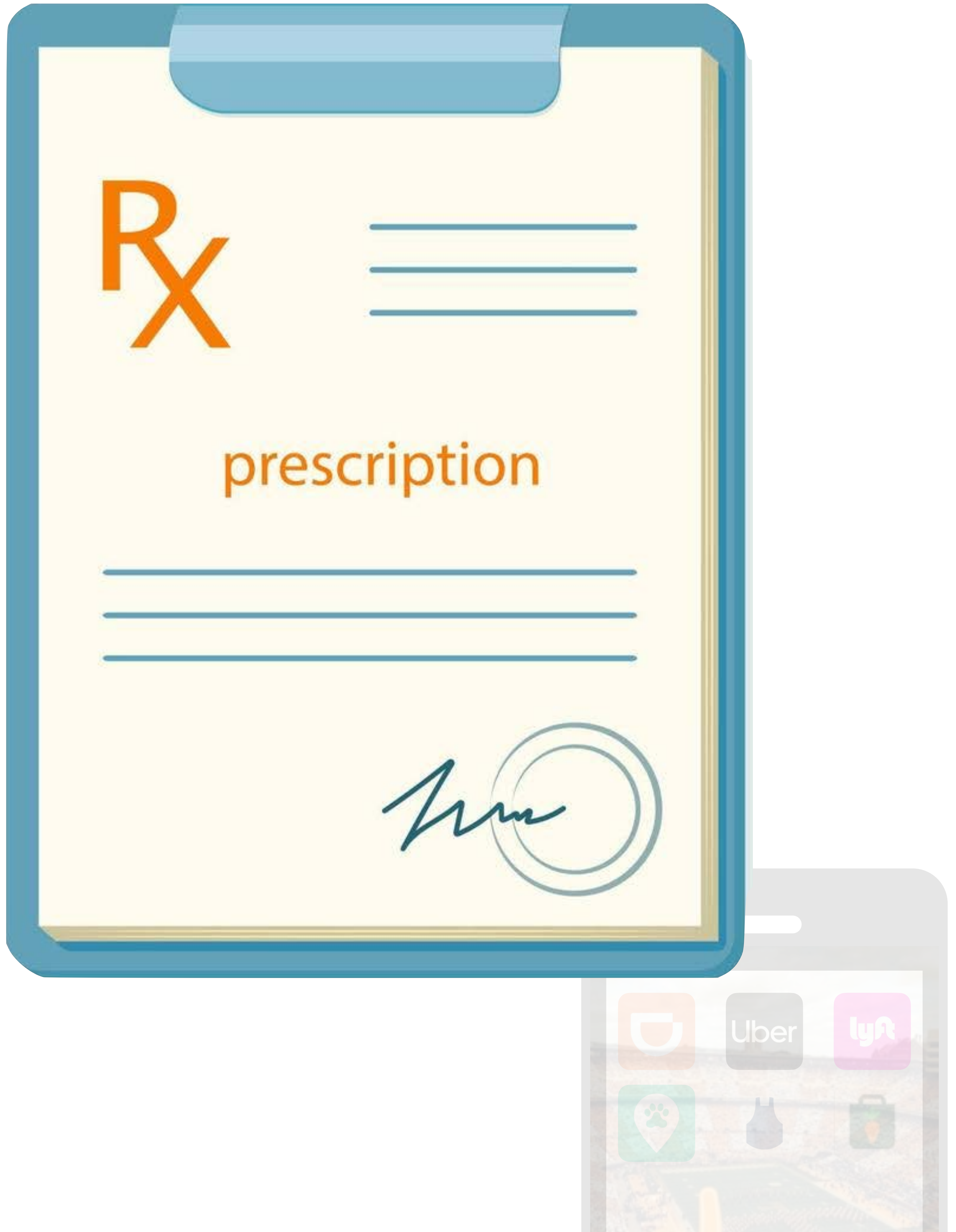




Main insights

How we can apply model insights

- **Payout Ratio Adjustments**
  - The platform should lower the payout ratio (α) as potential number of providers (K) or speed of service (μ) increases
  - The platform should increase (α) when cost of waiting (c) or potential customers ($\bar{\lambda}$) grows
    - This explains strategies like Uber's initial high payout ratios during early expansion phases, reduced later as provider and demand rates grew proportionally.



Empirical evidence

Data from Didi

- Model was compared to empirical results
- Data reflect main insights found analytically
    - Exception: The model doesn't capture competition