

Coordinating supply and demand on an on-demand service platform **with impatient customers**

Bai, J., So, K. C., Tang, C. S., Chen, X., & Wang, H. (2019)
<https://doi.org/10.1287/msom.2018.0707>

A paper presentation by Adam Spannbauer

“on-demand service platform with impatient customers”

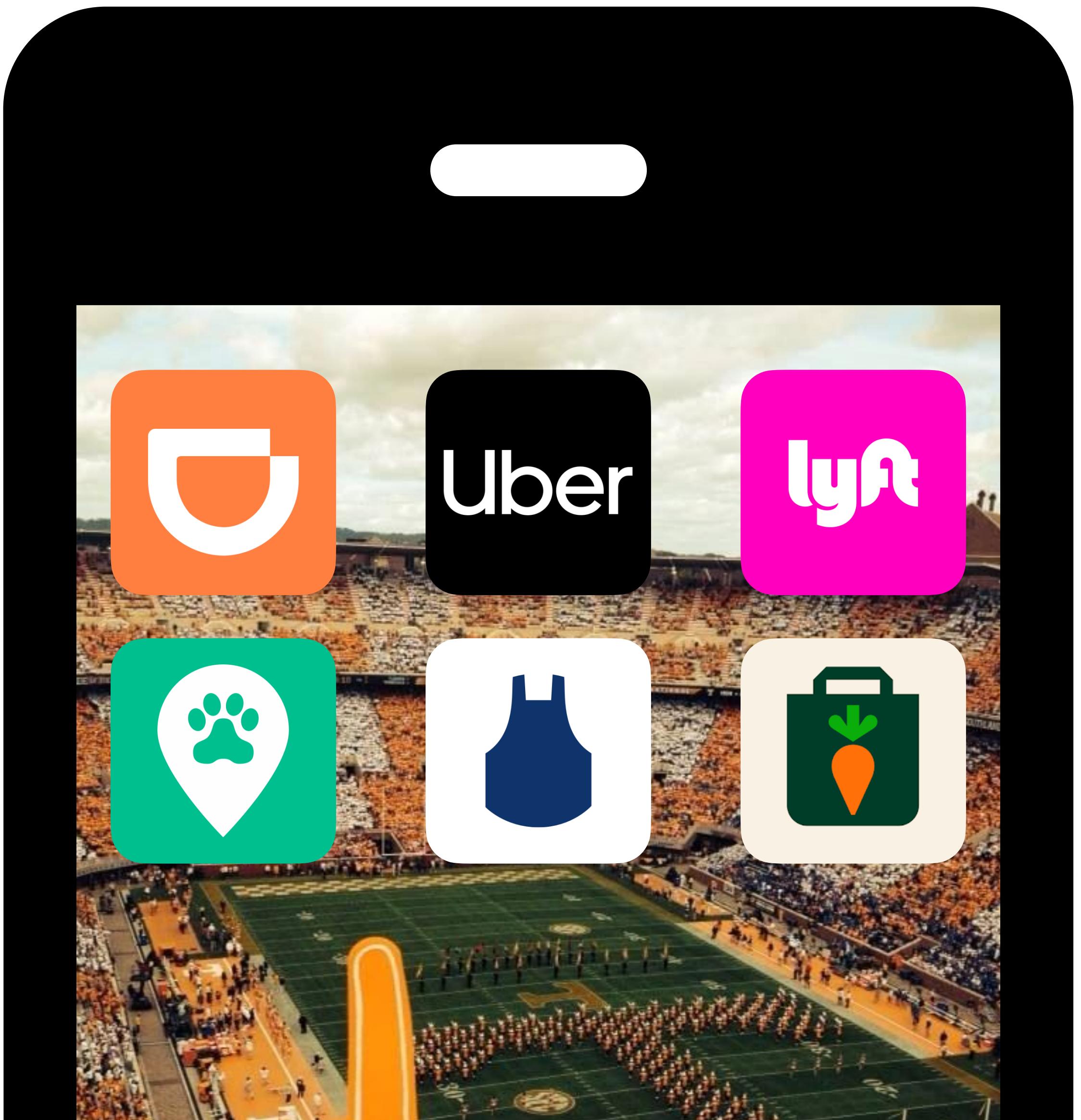
Example companies with this business model



“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

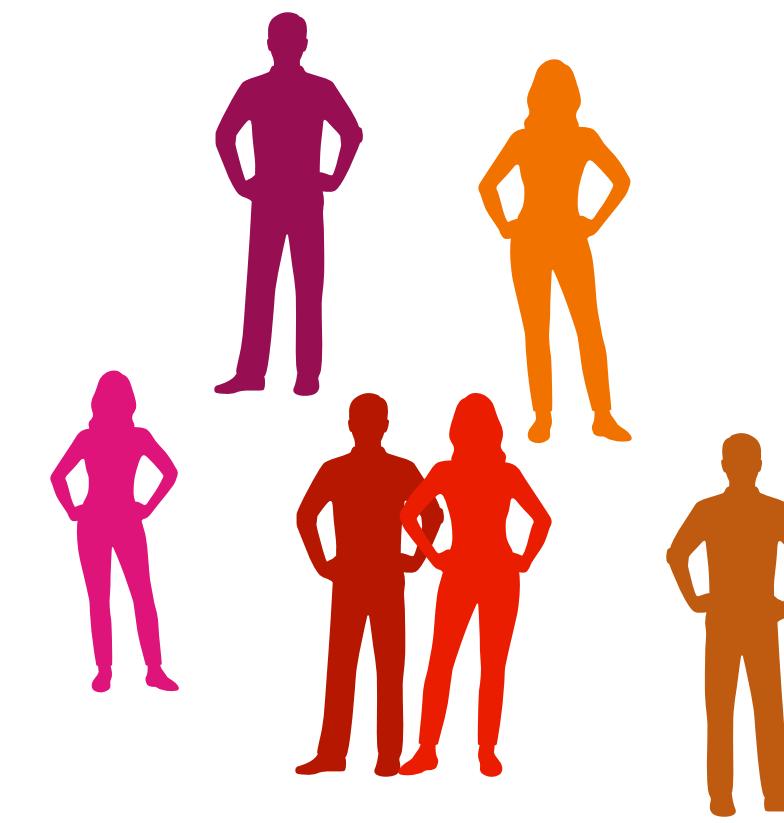
Coordinating supply and demand for on-demand services

The supply (independent providers)

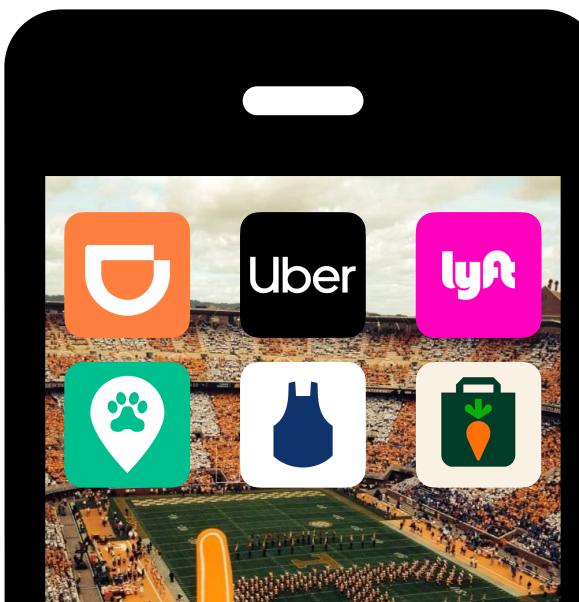


Goal: Maximize earnings

The demand (impatient customers)



Goal: Minimize cost and wait time



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

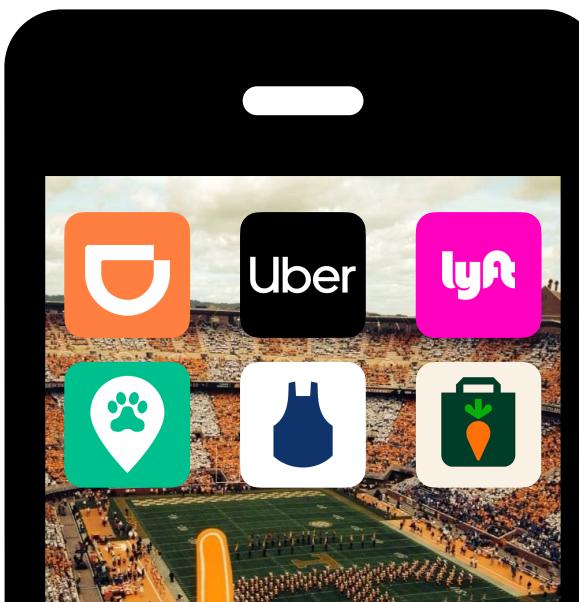
Depends on:

- wage rate ←

Decided by the firm

Depends on:

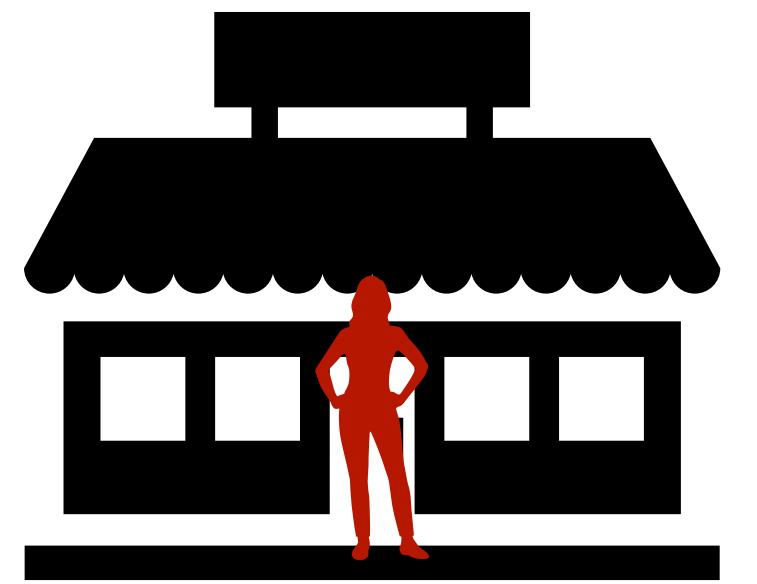
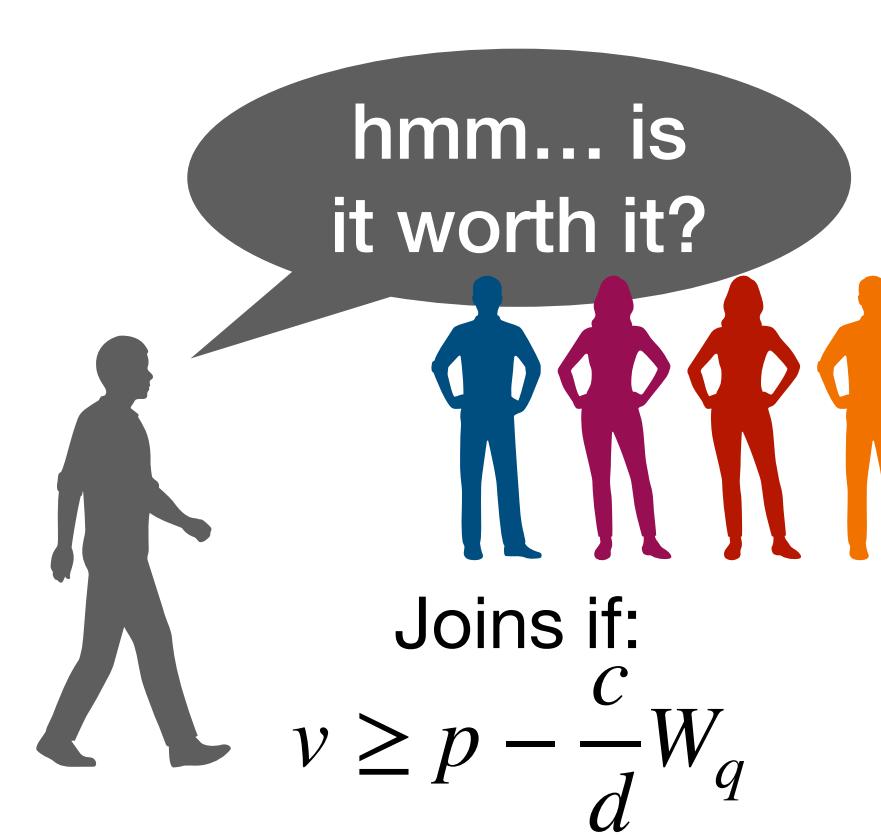
- price



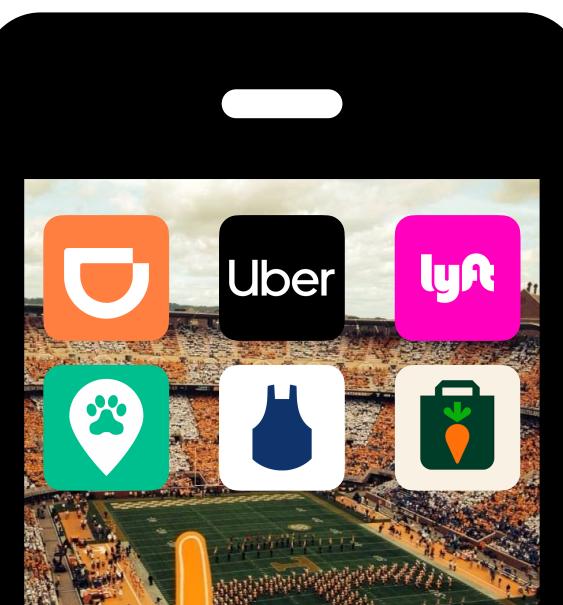
Customers only act when utility > 0

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)



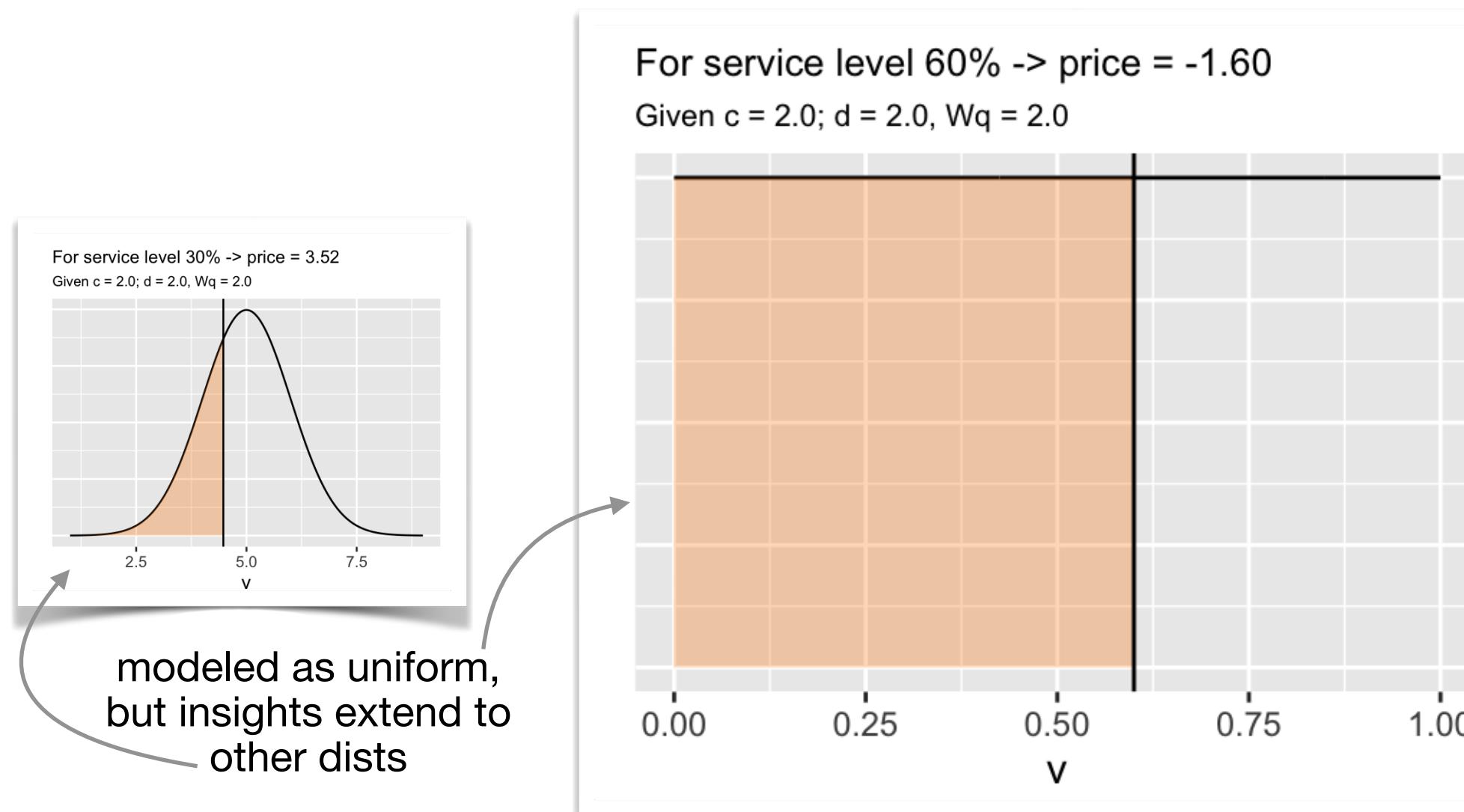
From utility function:
 $U(v) = (v - p)d - cW_q$



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)



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

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

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



The effects of inputs on price

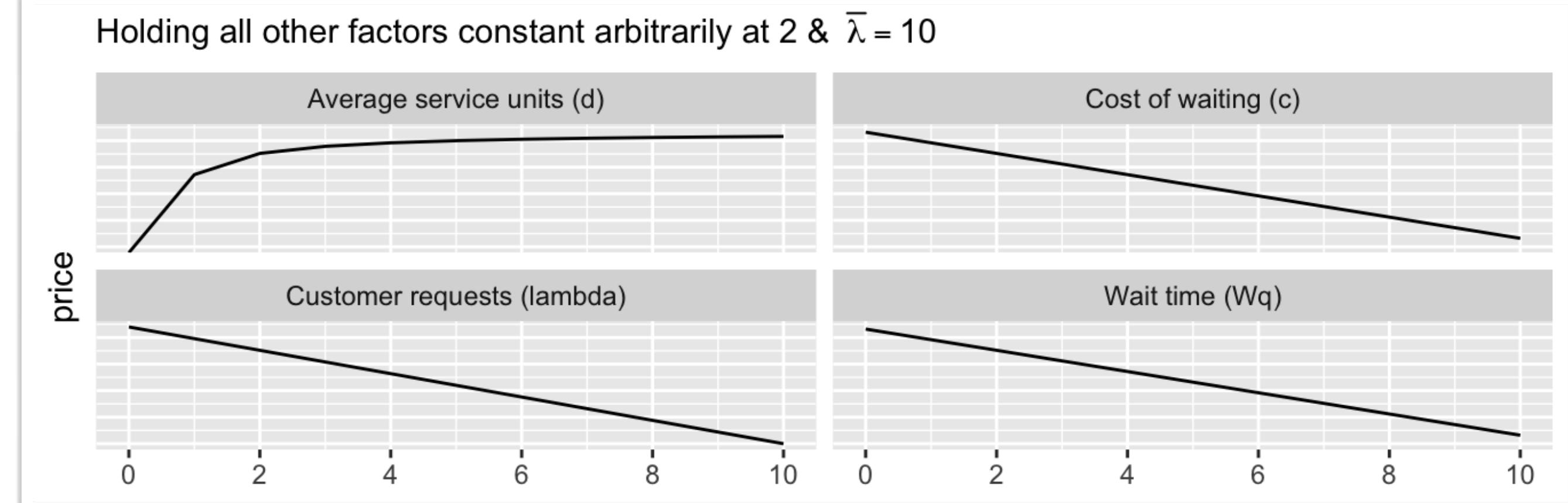
Classic Naor queueing ideas applied to $M/M/k$ queues

- 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



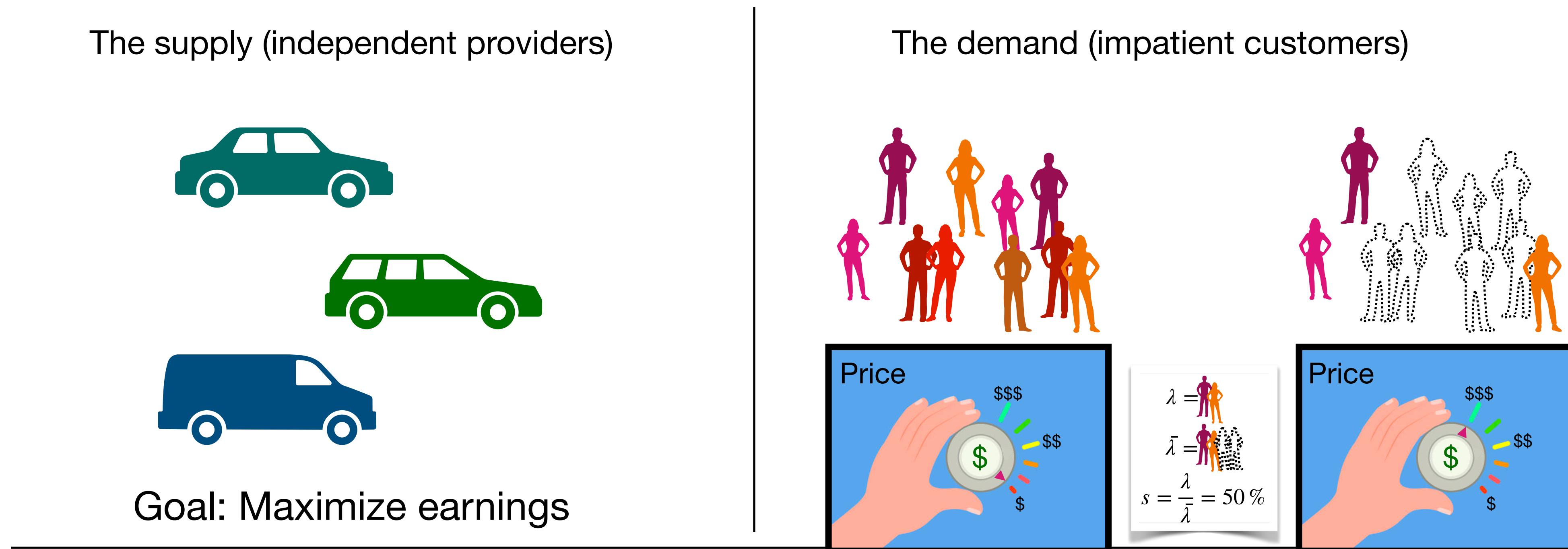
$$p = F^{-1} \left(1 - \frac{\lambda}{\bar{\lambda}} \right) - \frac{c}{d} W_q$$

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



Overview of the system being modeled

Coordinating supply and demand for on-demand services



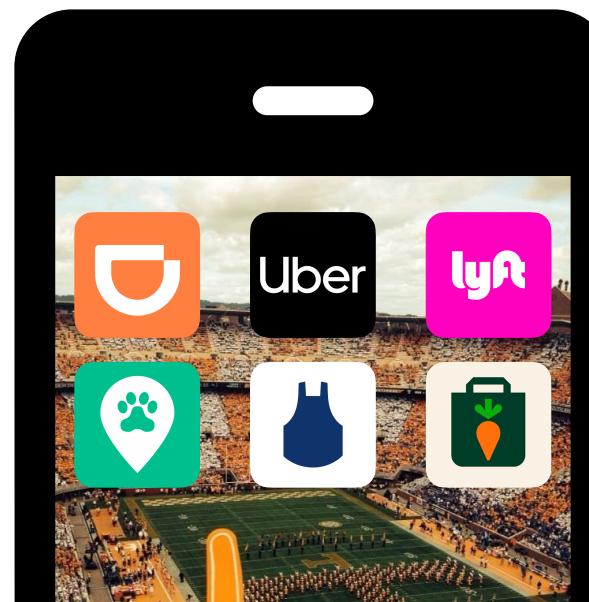
Depends on:

- wage rate ←

Decided by the firm →

Depends on:

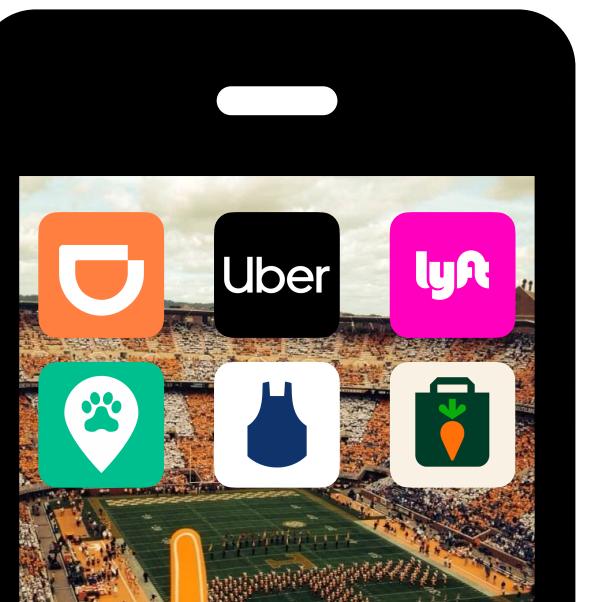
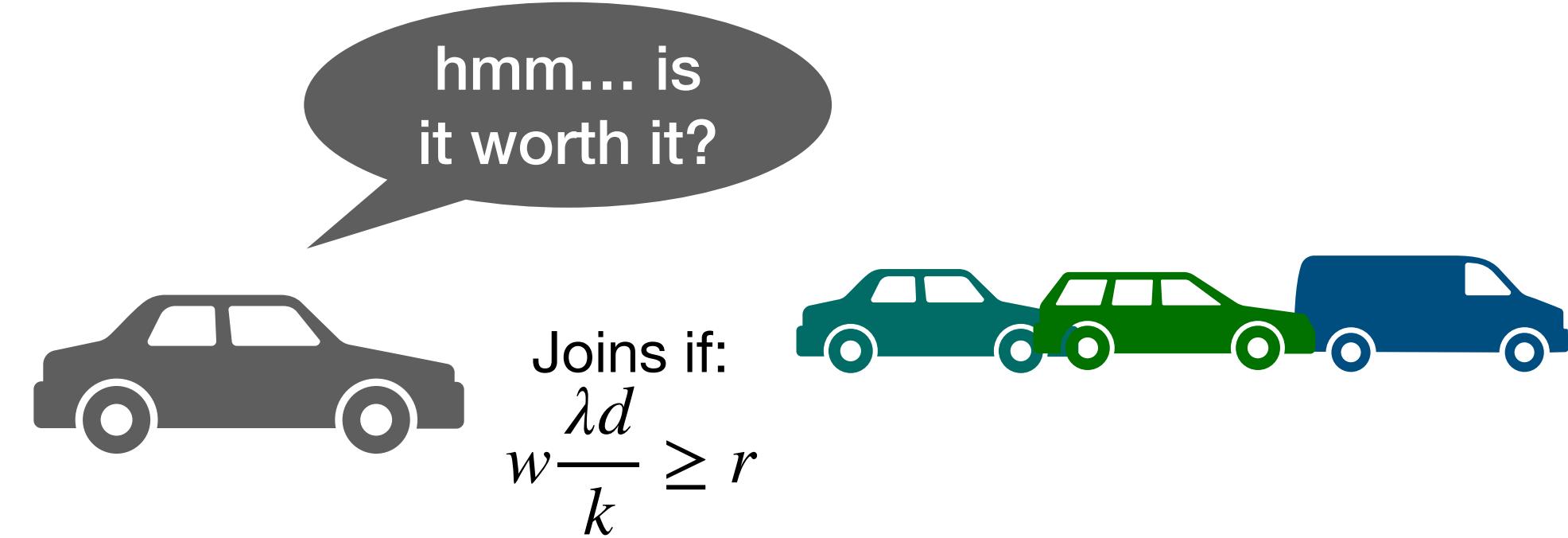
- price



Providers only enter when it's worth their time

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

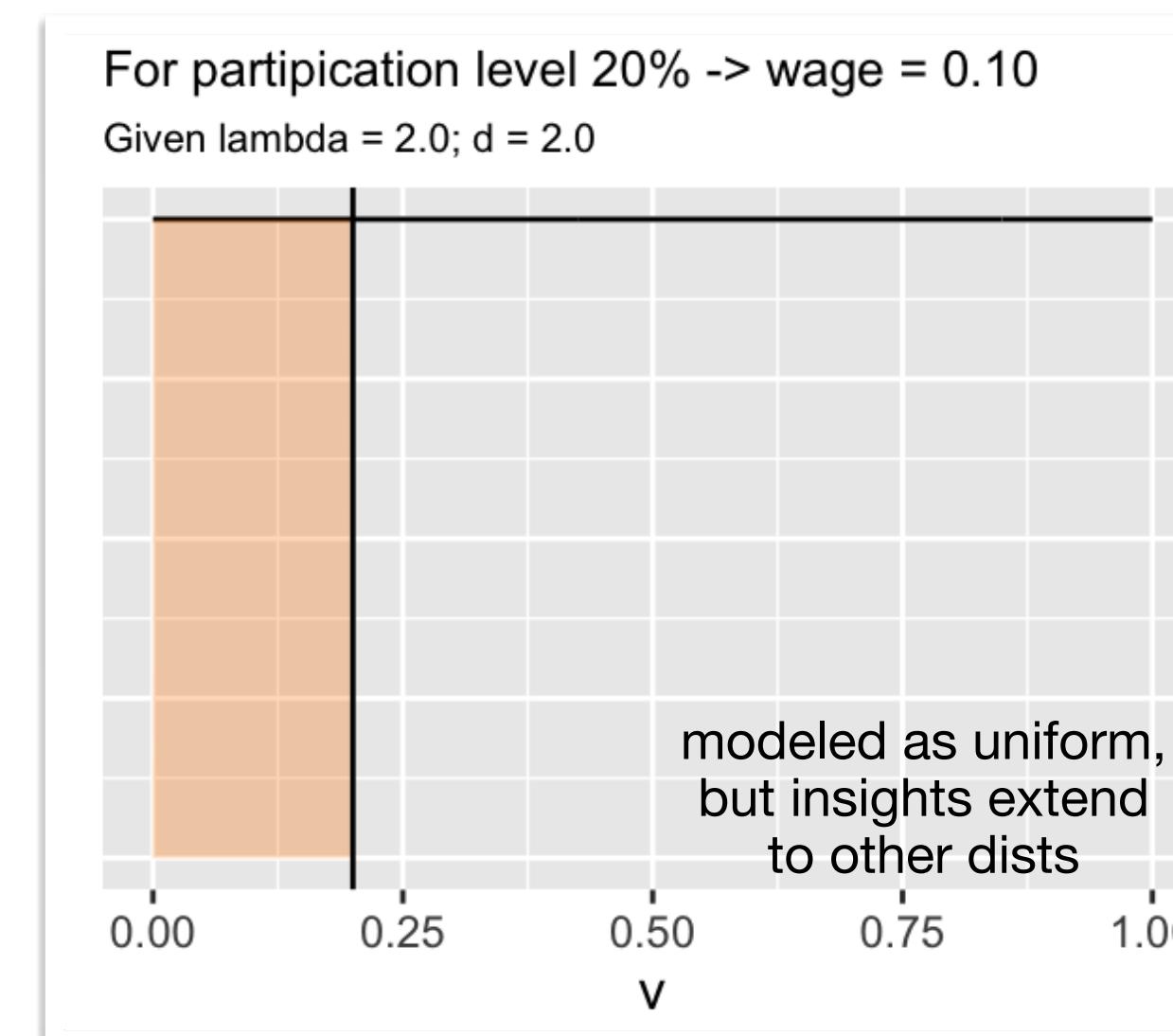
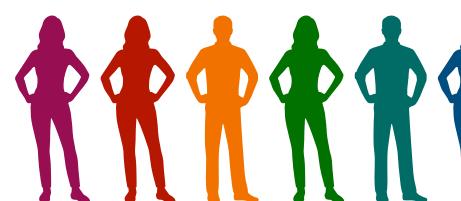
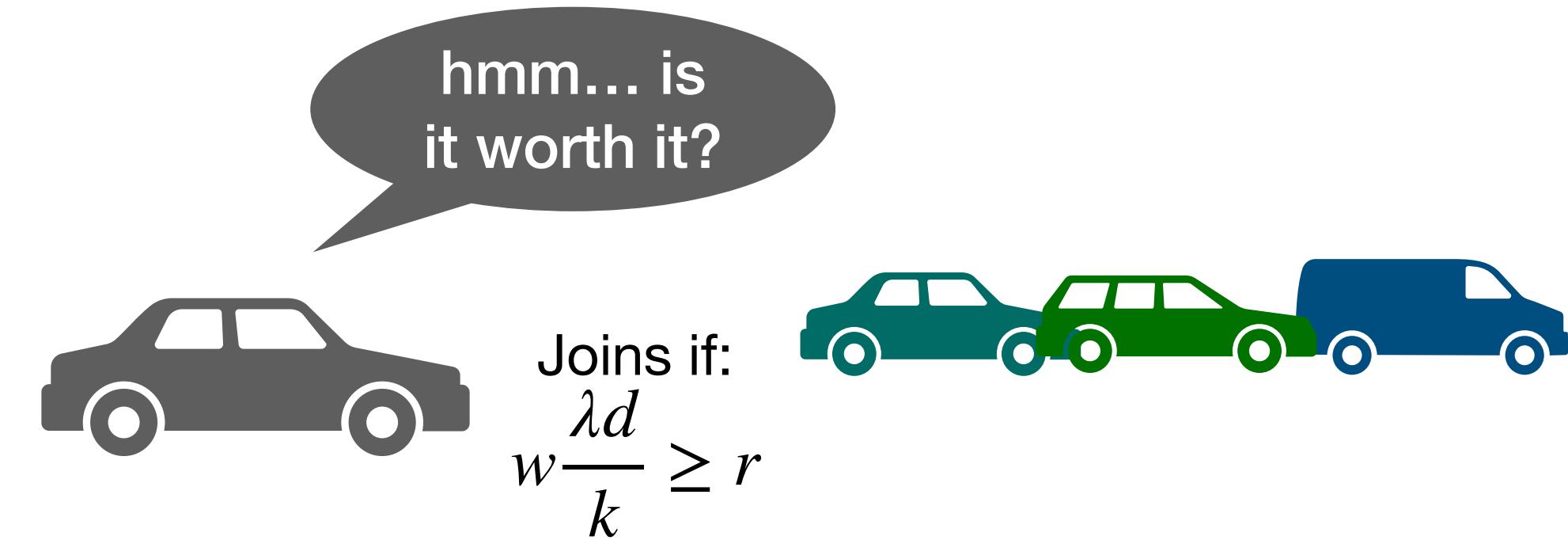
- Service **providers** use an app interact with their customers and eventually receive a **wage**
 - Providers have different reservation rates for which they'll participate ($r \sim G(\cdot)$) of the service
 - Providers join if their reservation rate (r) is less than the wage rate (w) \times expected 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 participation rate)
- Providers have different reservation rates (r ; the minimum rate at which they'll participate)



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

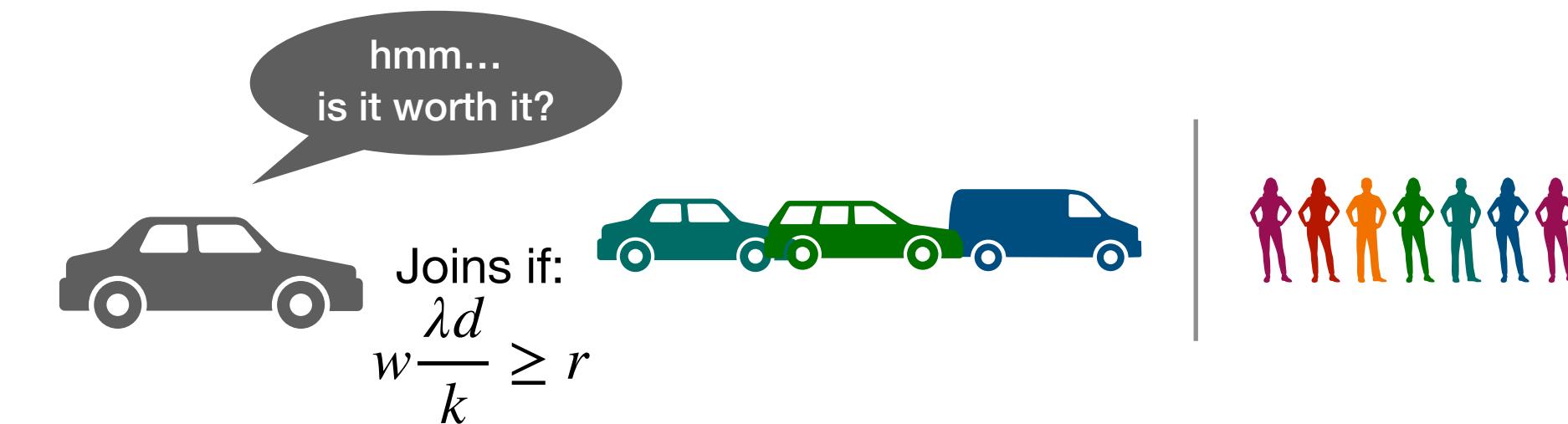
How to set price (w) based on target participation rate ($\beta = \frac{k}{K}$)



Providers only enter when it's worth their time

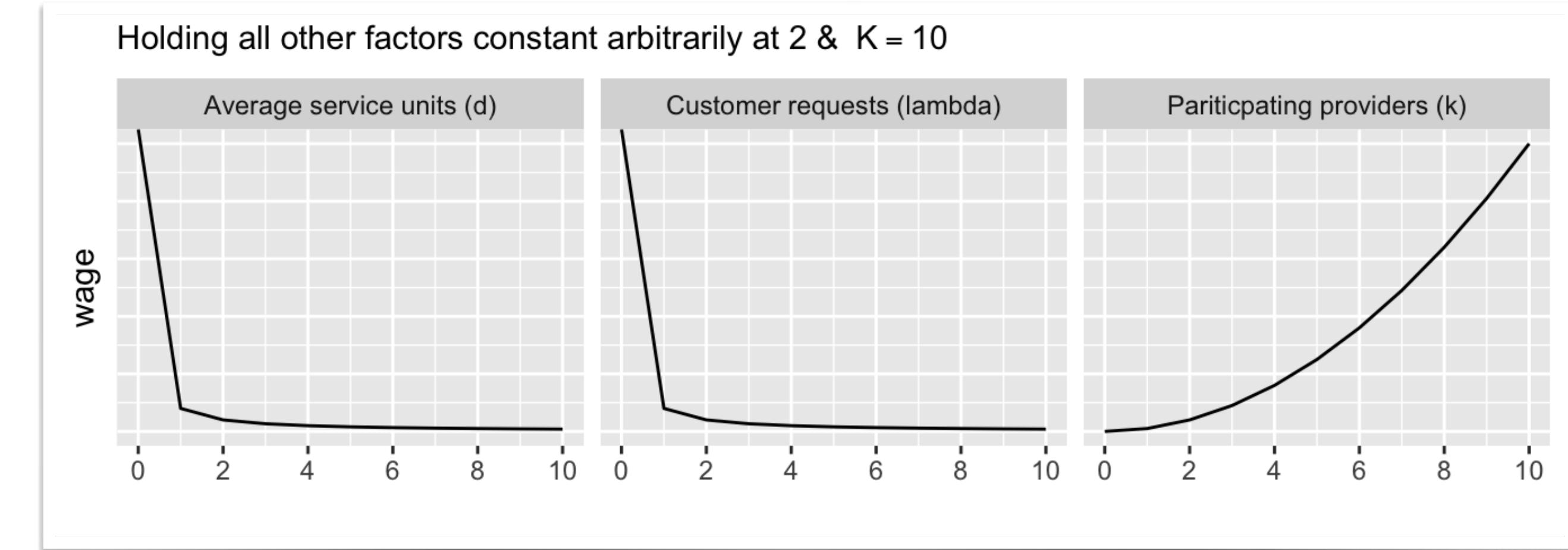
Higher wages (w) and utilization (ρ) leads to more providers

- 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.



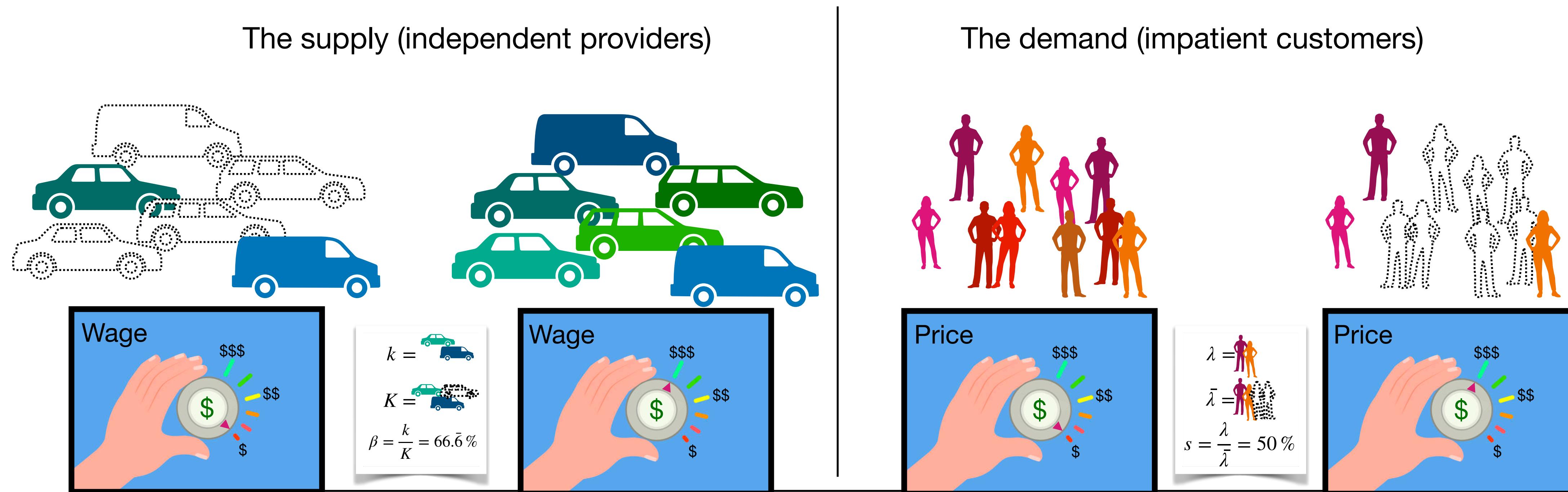
$$w = G^{-1}\left(\frac{k}{K}\right) \frac{k}{\lambda d}$$

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



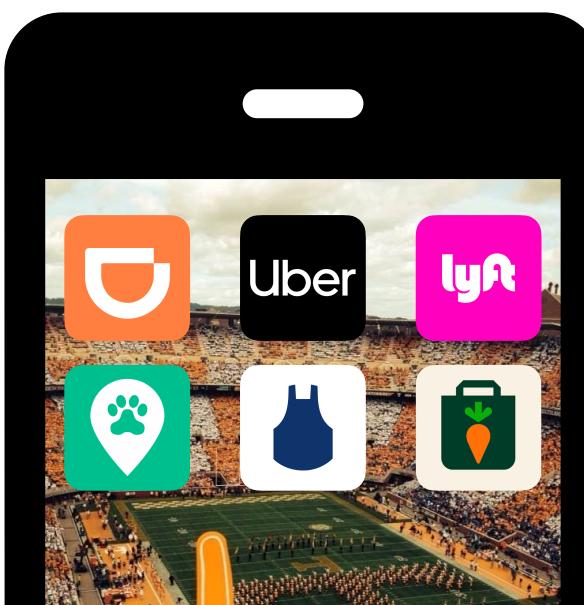
Depends on:

- wage rate ←

Decided by the firm

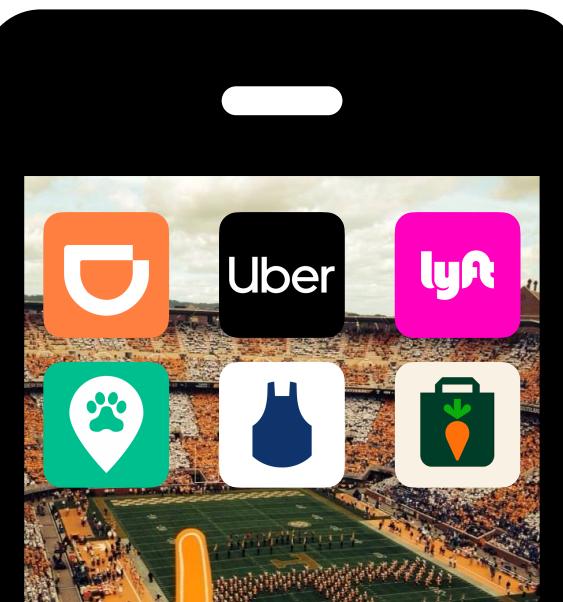
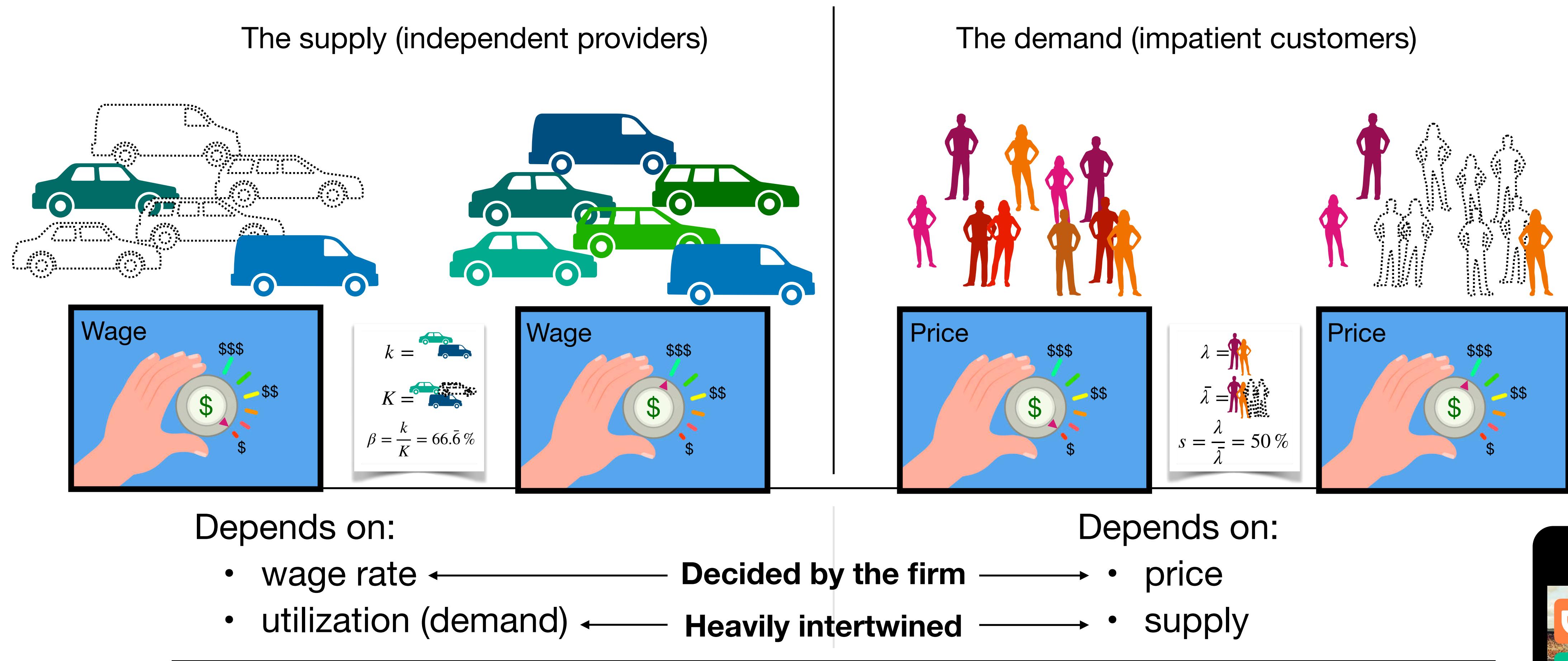
Depends on:

- price



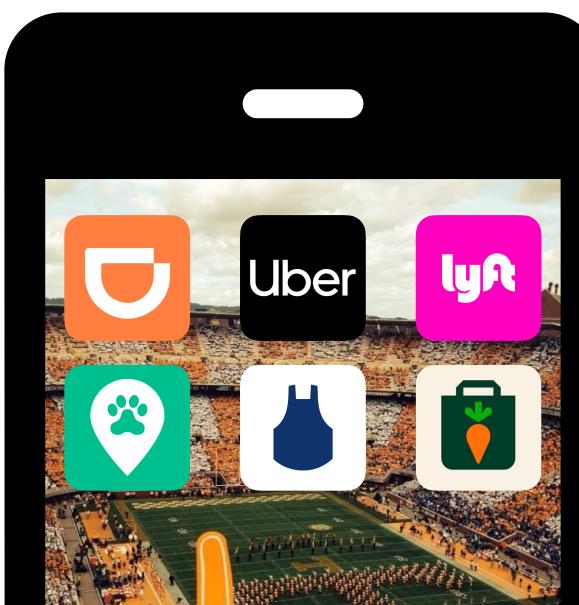
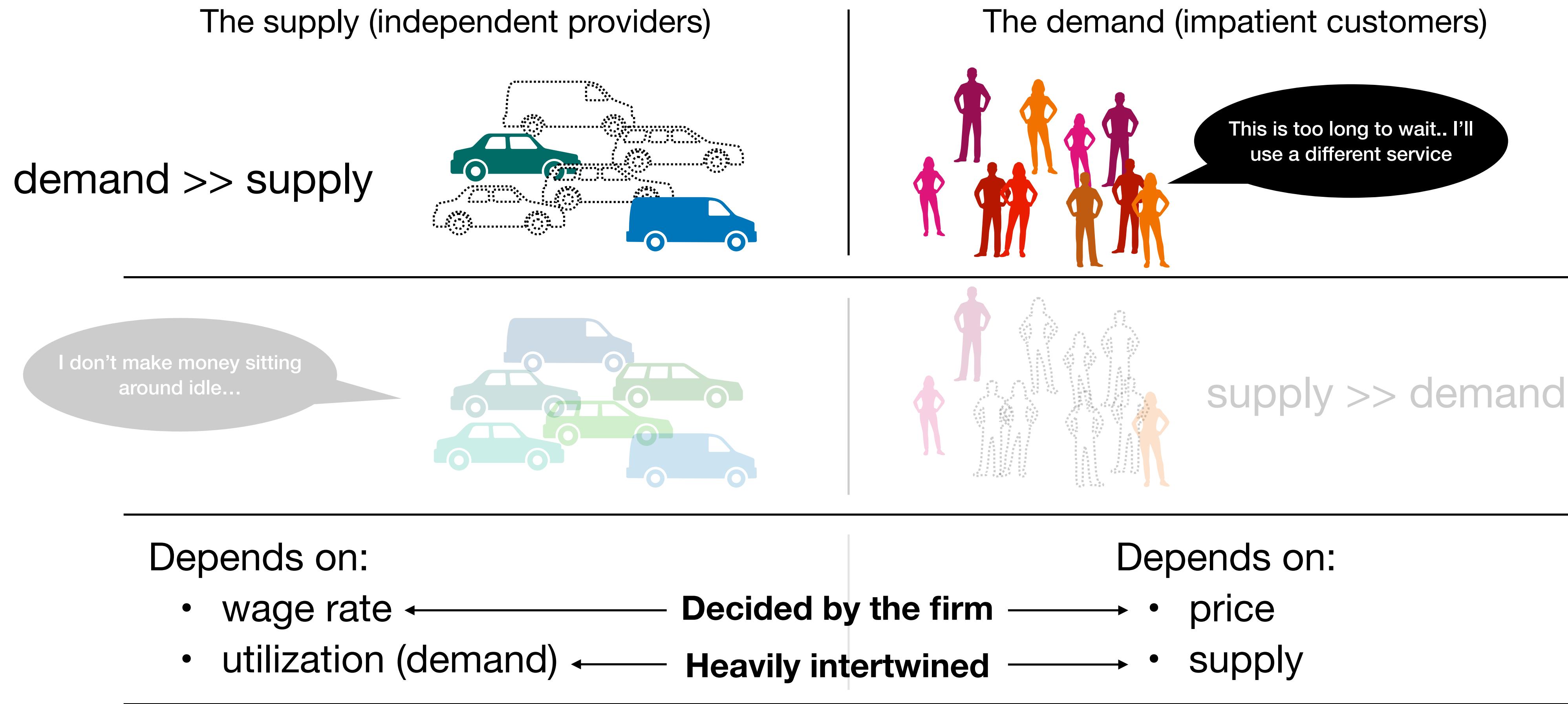
Overview of the system being modeled

Coordinating supply and demand for on-demand services



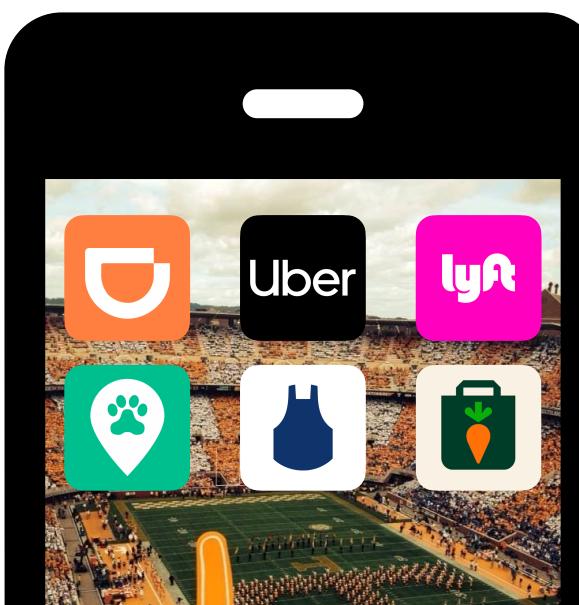
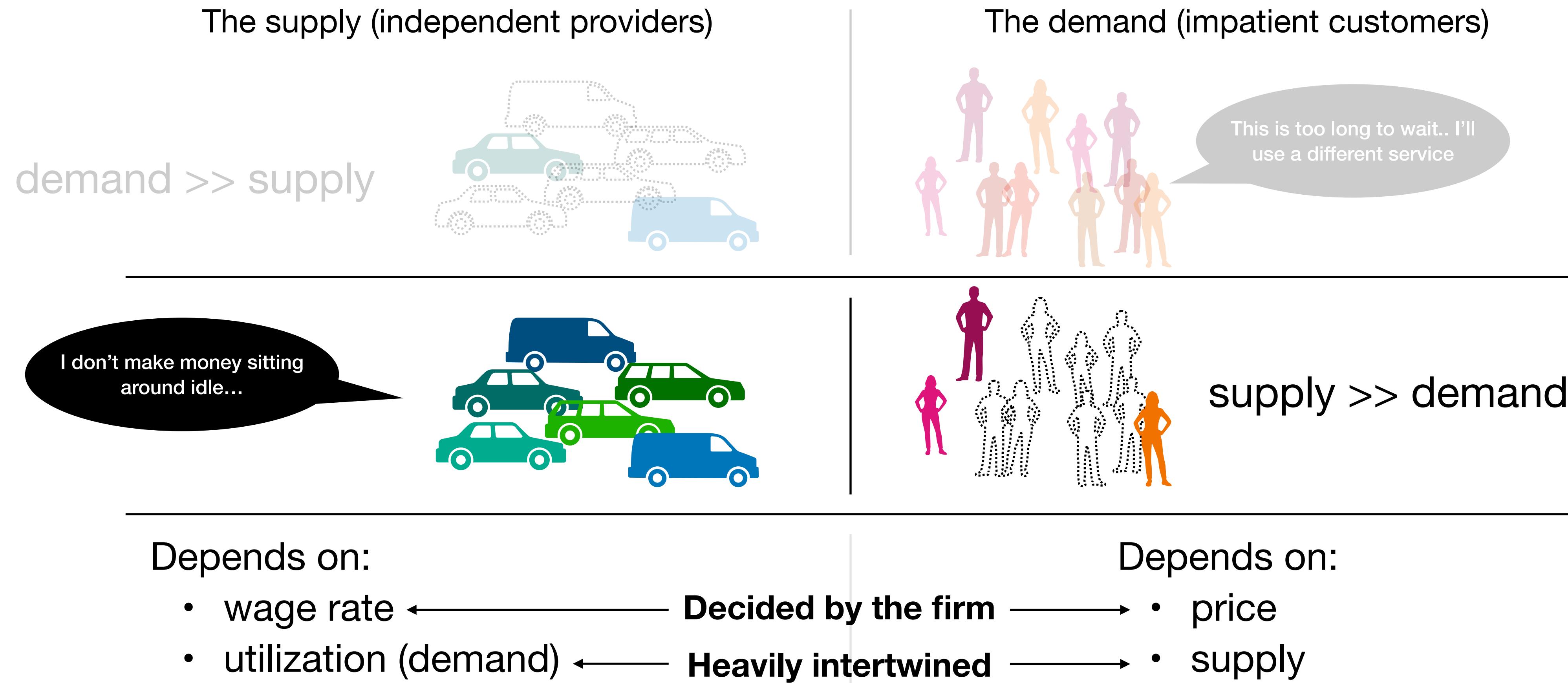
Overview of the system being modeled

Coordinating supply and demand for on-demand services



Overview of the system being modeled

Coordinating supply and demand for on-demand services

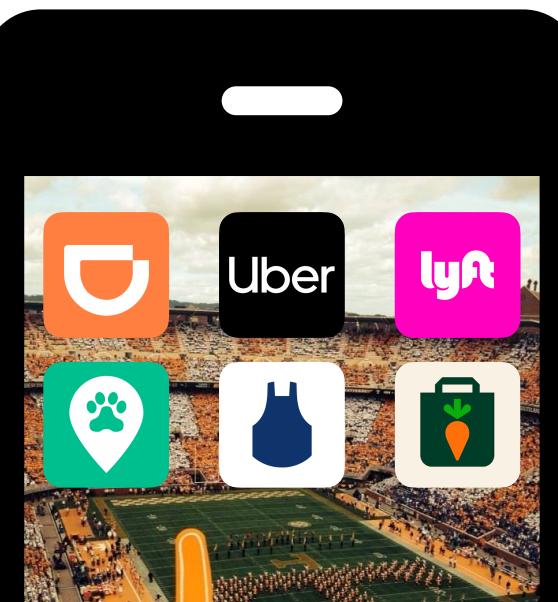


The firms want to maximize profit

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)

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



The firms want to maximize profit

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

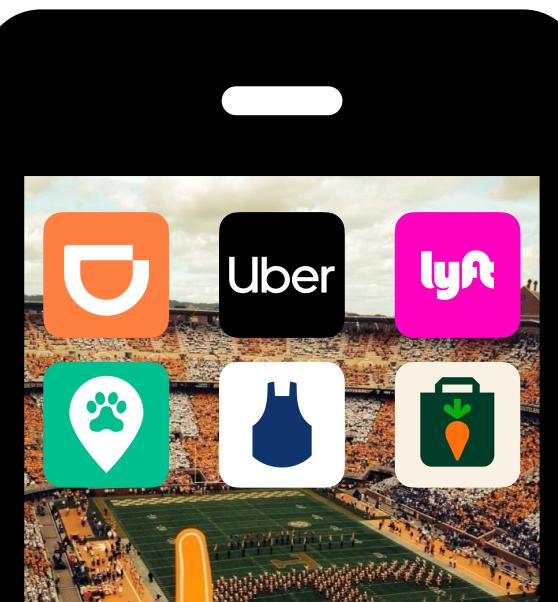
Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

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

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

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)

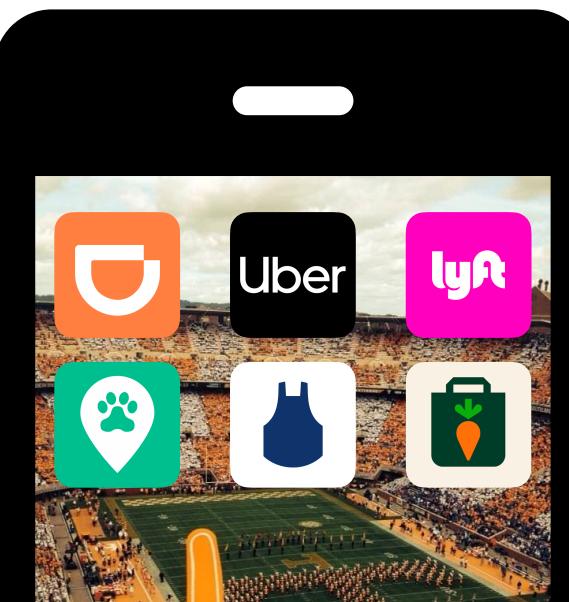
$$\max \pi = \lambda(p - w)d$$

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

After manipulation
we end up with this
objective

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

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



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)

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

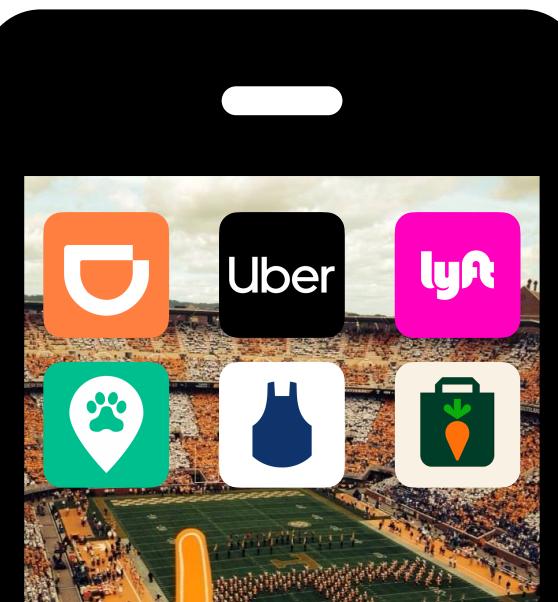
$$\max \pi = \lambda\left(\frac{w}{\alpha} - w\right)d$$

After manipulation
we end up with this
objective

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

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

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)

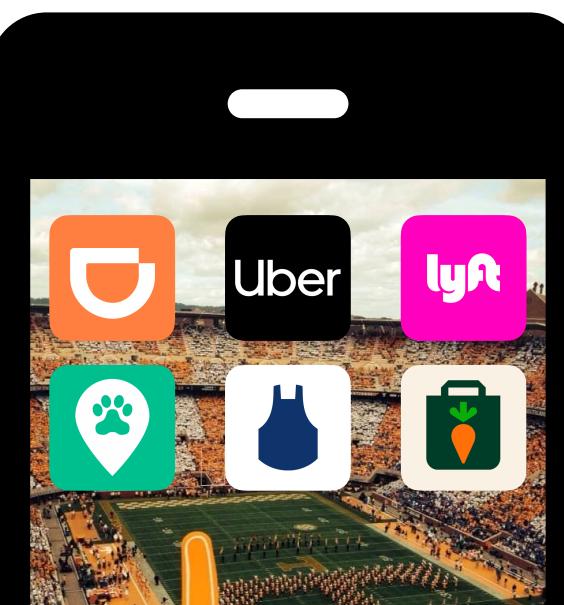
Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

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

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

$\bar{\lambda}$	W_q is given by exact formula (9)			
	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)

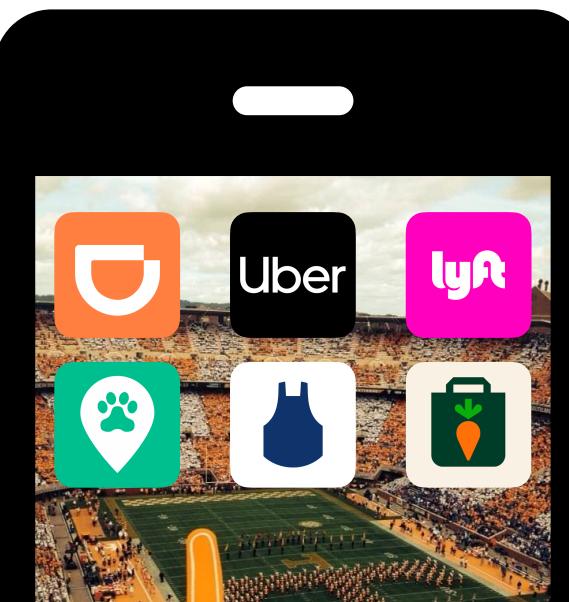
Use price (p) and wage (w) to drive requests (λ) and participating providers (k)

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

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

$\bar{\lambda}$	W_q is given by exact formula (9)			
	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

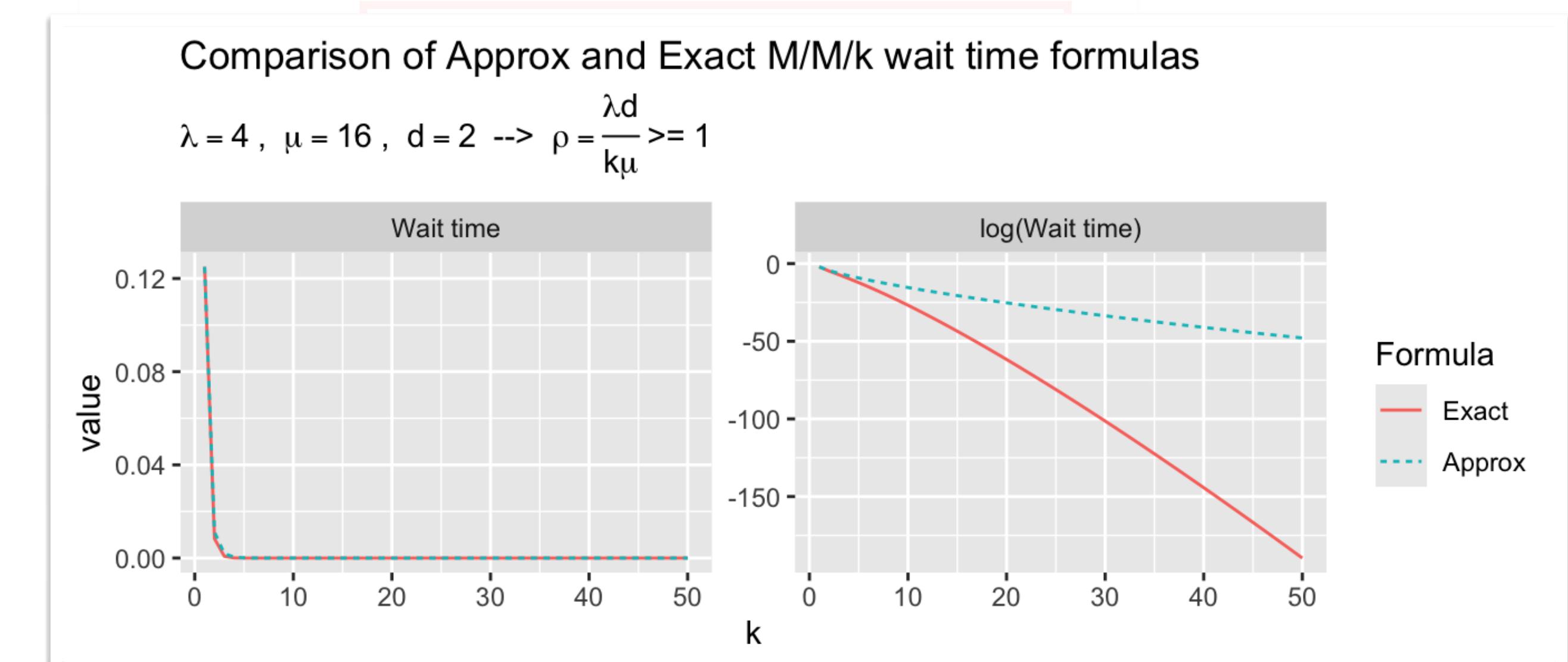
To calculate p^* we need W_q , which is defined as:

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

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

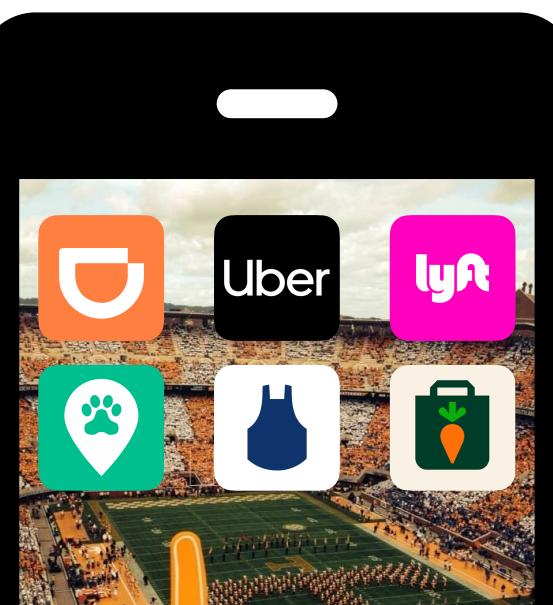
How inputs affect the optimal solution parameters

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).

Paper's analytical results using approximate W_q



The firms want to maximize profit

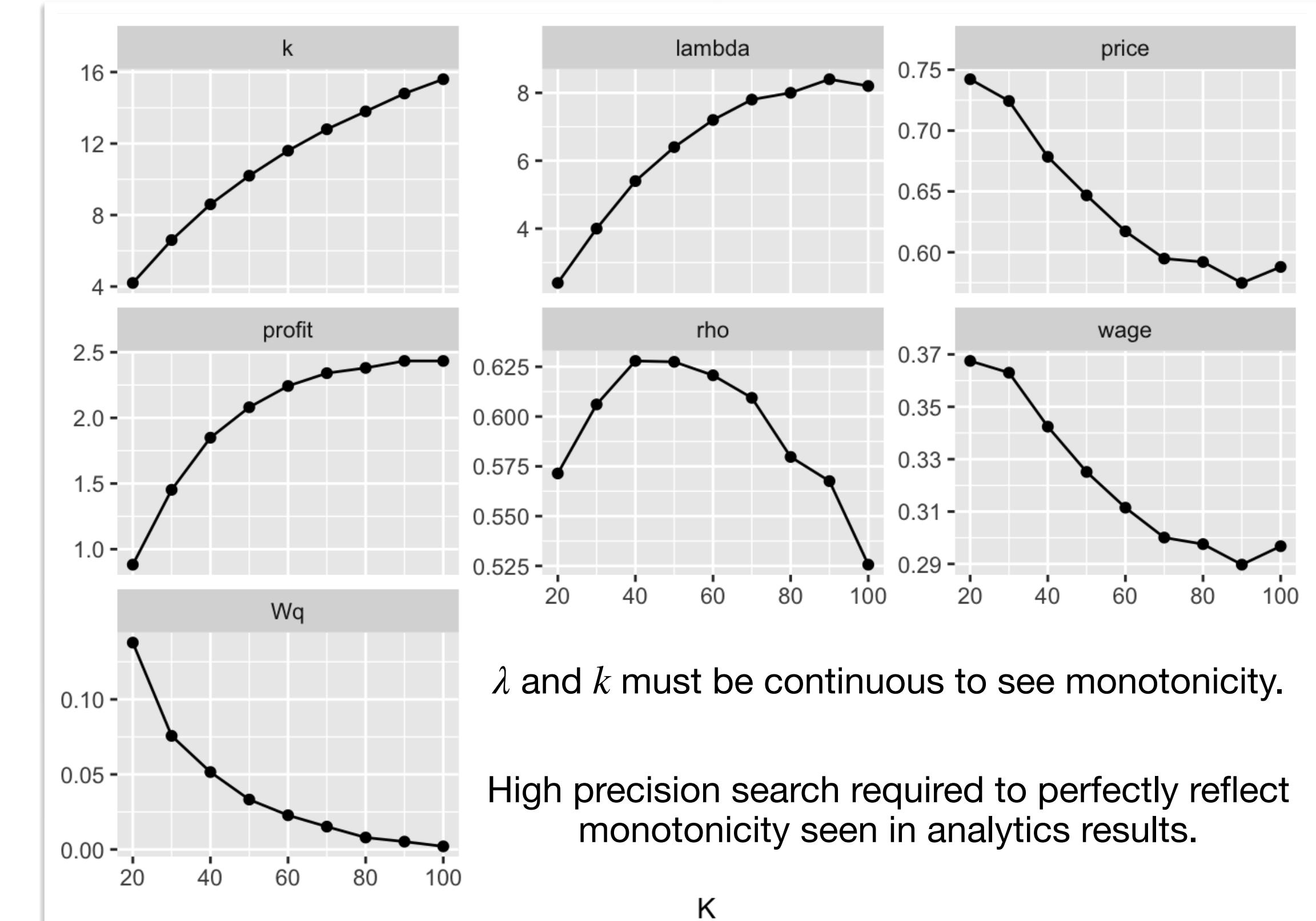
How inputs affect the optimal solution parameters

Paper's analytical results using approximate W_q^*

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

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

My results optimizing with exact W_q



The firms want to maximize profit

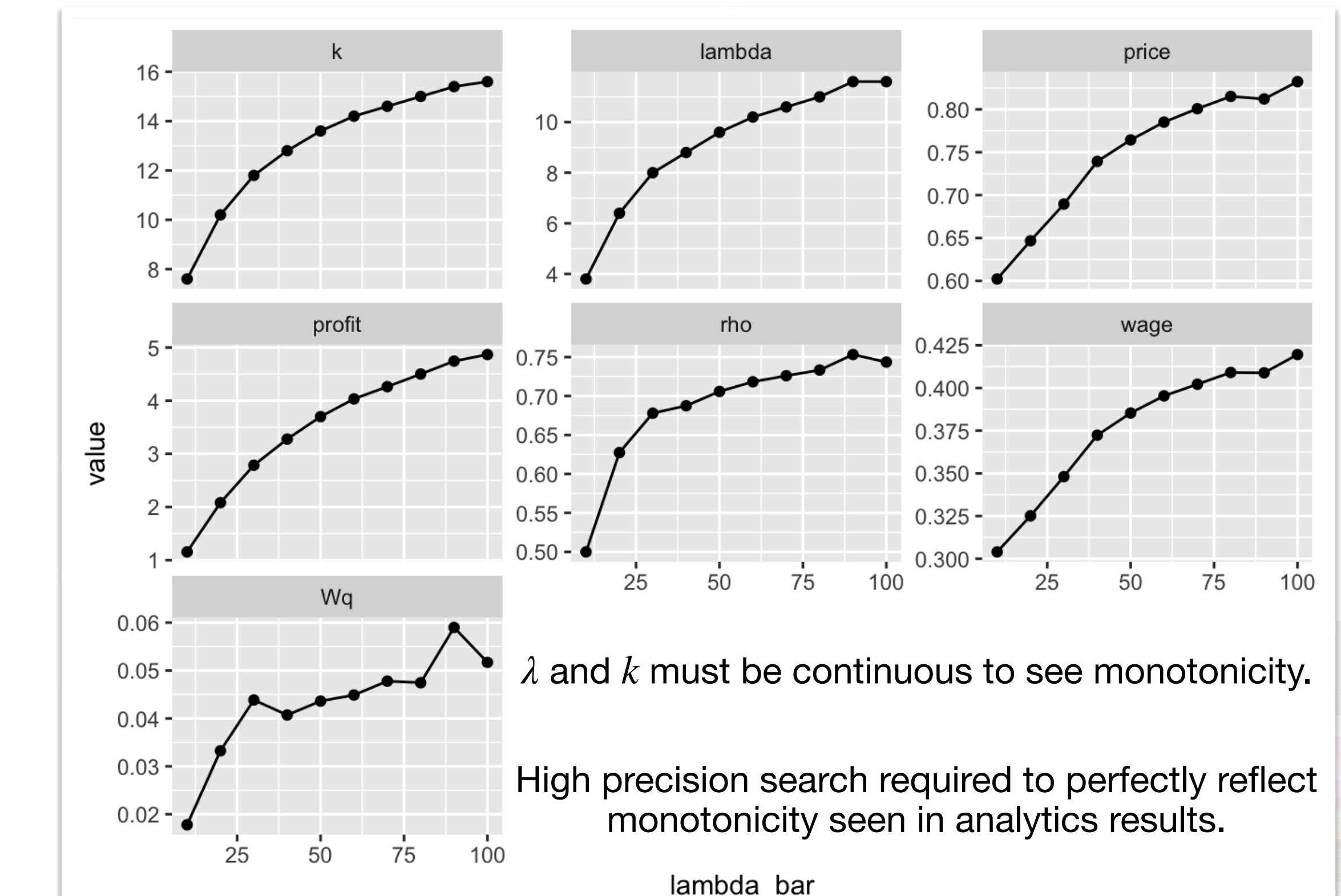
How inputs affect the optimal solution parameters

Paper's analytical results using approximate W_q^*

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

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

My results optimizing with exact W_q



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	↓	↑	↑	↓	↑

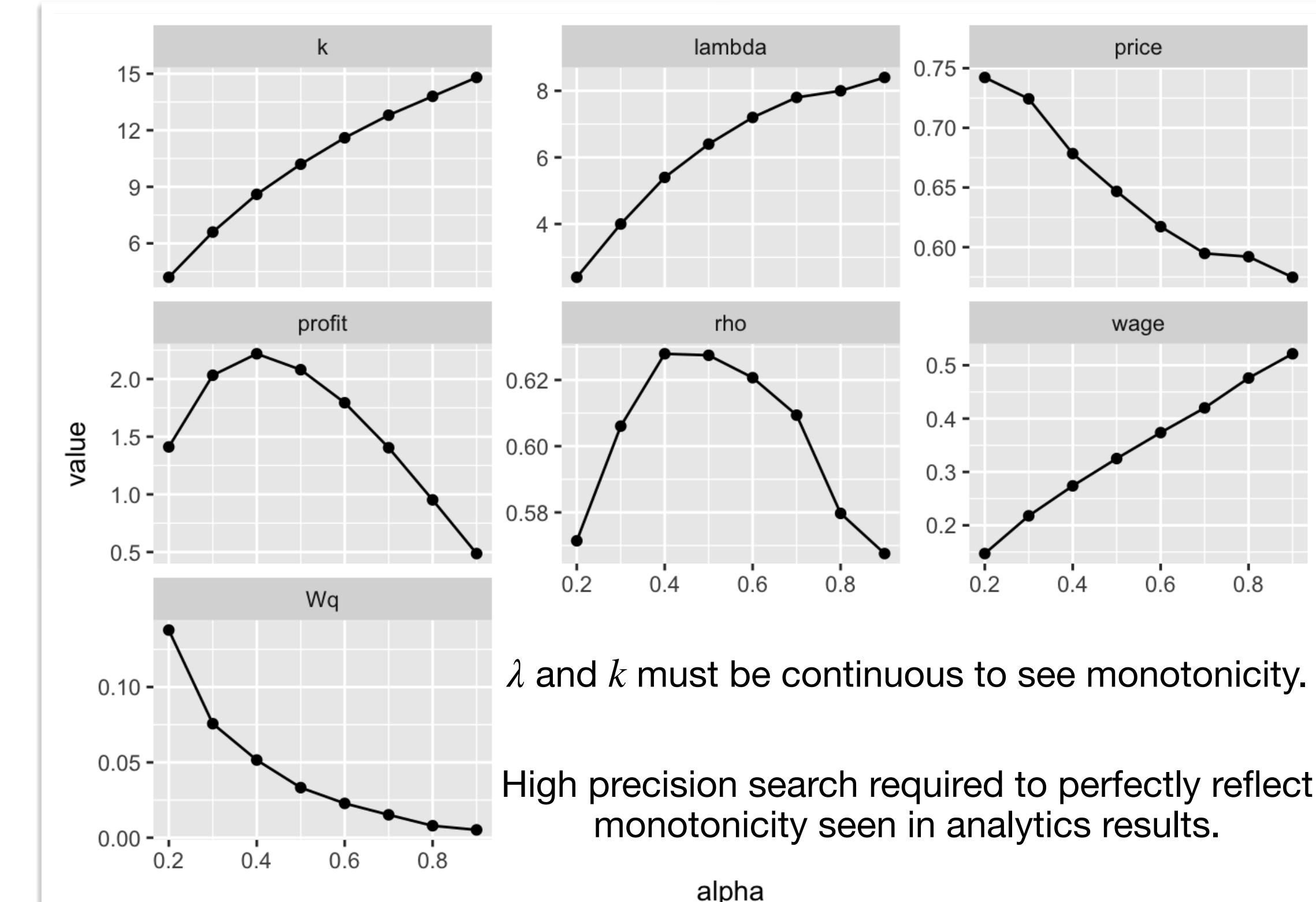
↑ (increasing); ↓ (decreasing); × (nonmonotonic).

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

Recall:

$$w = \alpha p$$

My results optimizing with exact W_q



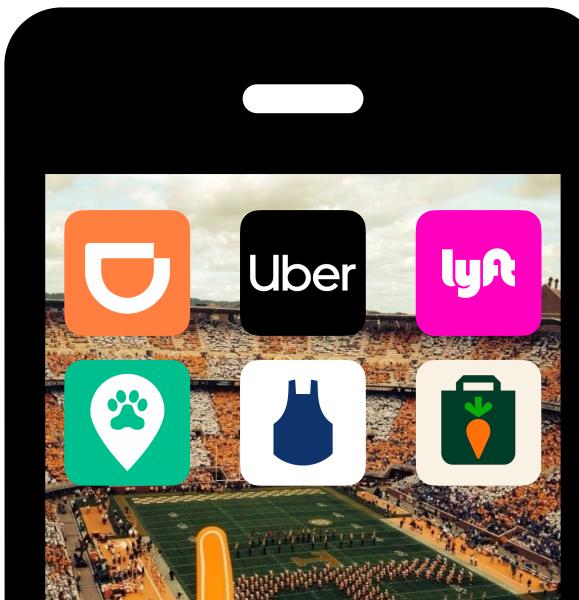
The firms want to maximize profit

How to set payout percentage (α)

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)

$\bar{\lambda}$	K									
	10	20	30	40	50	60	70	80	90	100
10	0.67	0.49	0.42	0.36	0.33	0.30	0.27	0.25	0.24	0.22
20	0.67	0.52	0.46	0.41	0.37	0.34	0.32	0.30	0.28	0.27
30	0.67	0.54	0.48	0.44	0.40	0.38	0.35	0.33	0.31	0.30
40	0.67	0.55	0.49	0.46	0.42	0.40	0.38	0.36	0.34	0.32
50	0.67	0.56	0.51	0.47	0.44	0.42	0.39	0.38	0.36	0.34
60	0.68	0.56	0.51	0.48	0.45	0.43	0.41	0.39	0.38	0.36
70	0.68	0.57	0.52	0.49	0.46	0.44	0.42	0.40	0.39	0.37
80	0.68	0.57	0.53	0.49	0.47	0.45	0.43	0.41	0.40	0.39
90	0.68	0.57	0.53	0.50	0.48	0.46	0.44	0.42	0.41	0.40
100	0.68	0.57	0.53	0.50	0.48	0.46	0.45	0.43	0.42	0.41

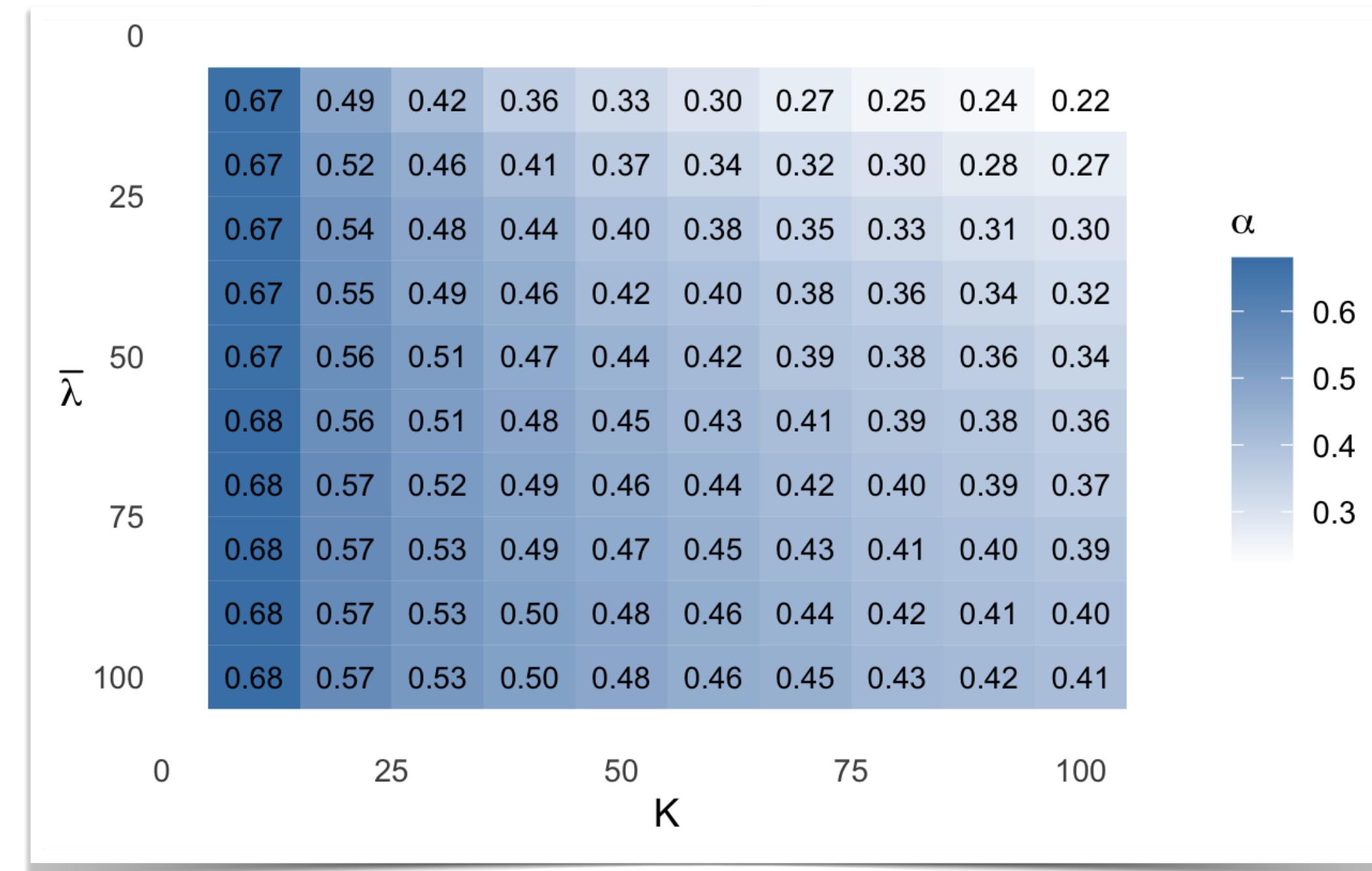


The firms want to maximize profit

How to set payout percentage (α)

Paper's analytical results^{re-colored} using approximate \hat{W}_q to find optimal α for given:

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

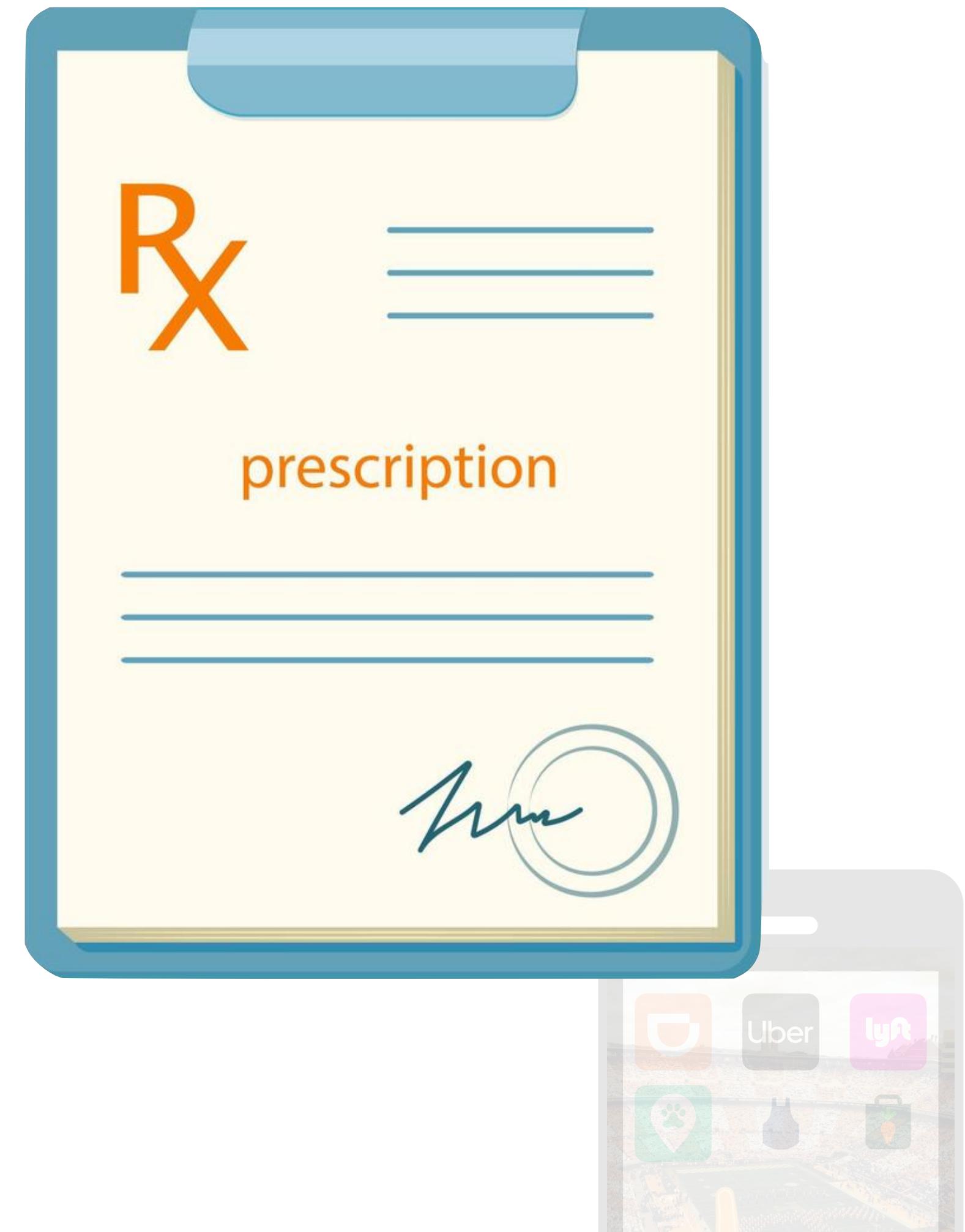


- When there are few potential providers (K), entice them with higher percentage of earnings (α)
- When there are many potential customers ($\bar{\lambda}$), use a higher percentage (α) to encourage more providers to participate and ensure wait times (W_q) stay down

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

