Optimal Multiserver Scheduling

with Unknown Job Sizes in Heavy Traffic

Ziv Scully
Isaac Grosof
Mor Harchol-Balter

Carnegie Mellon University

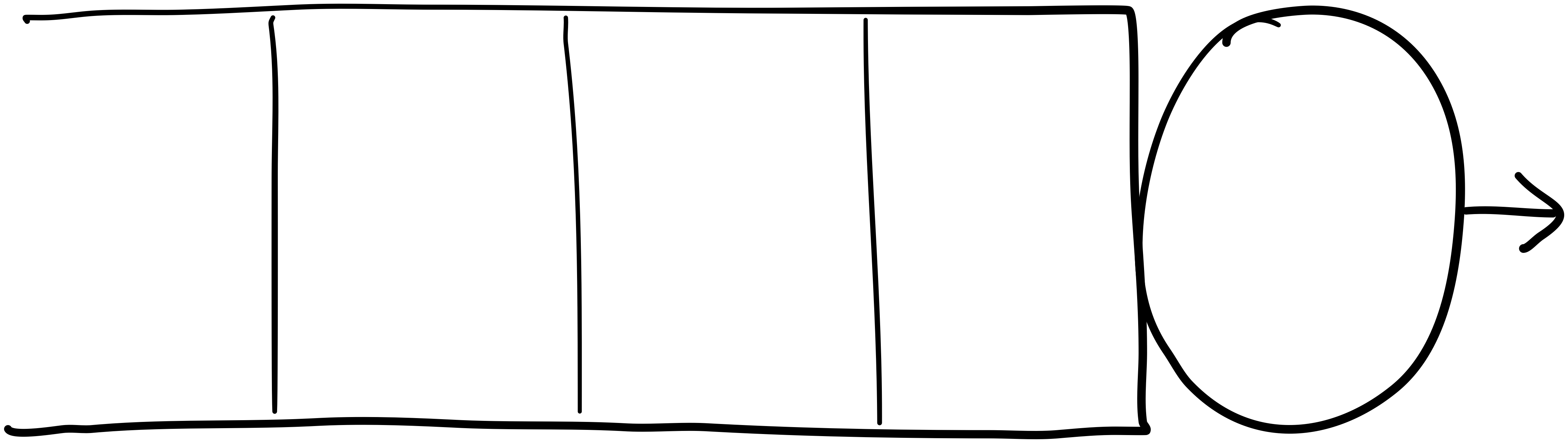
Optimal Multiserver Scheduling

with Unknown Job Sizes in Heavy Traffic

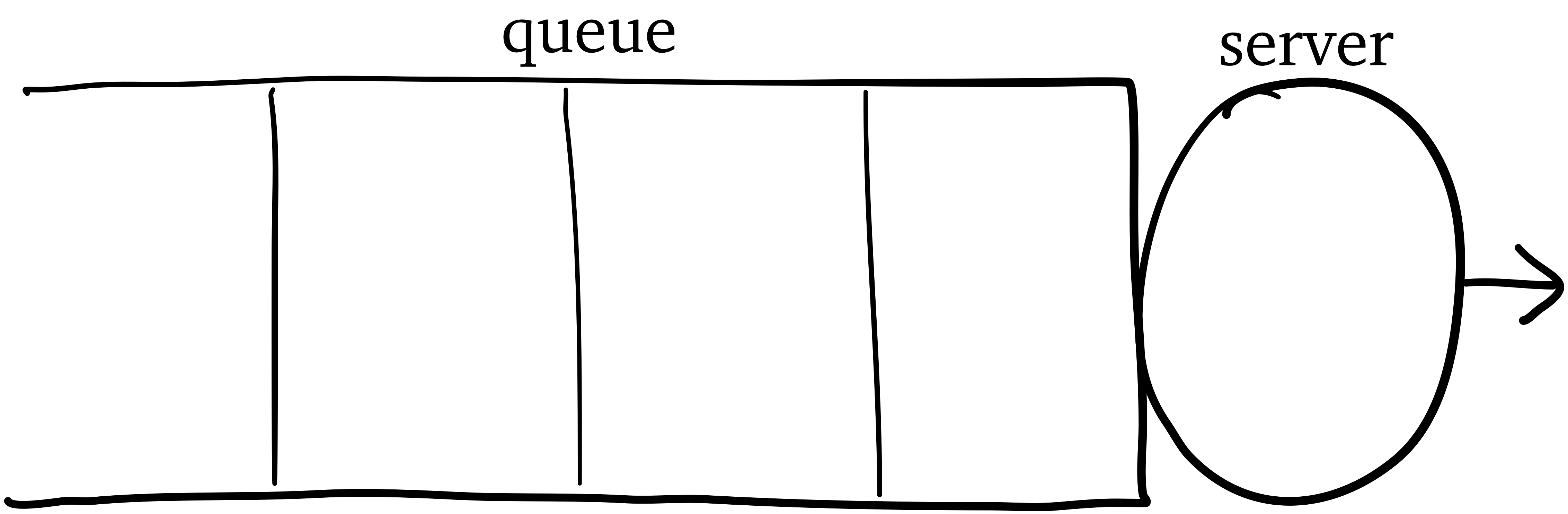
Ziv Scully
Isaac Grosof
Mor Harchol-Balter

Carnegie Mellon University

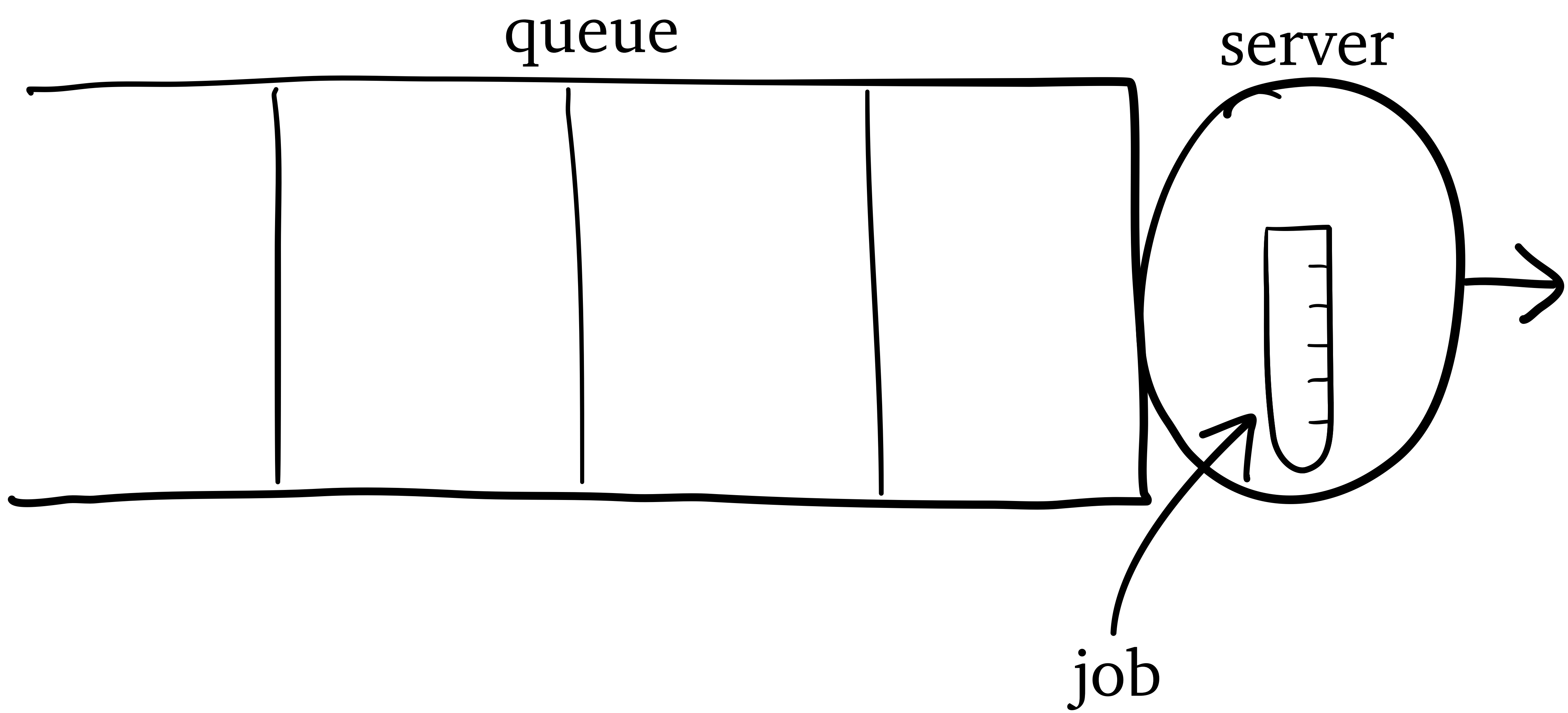
M/G/1 Queue



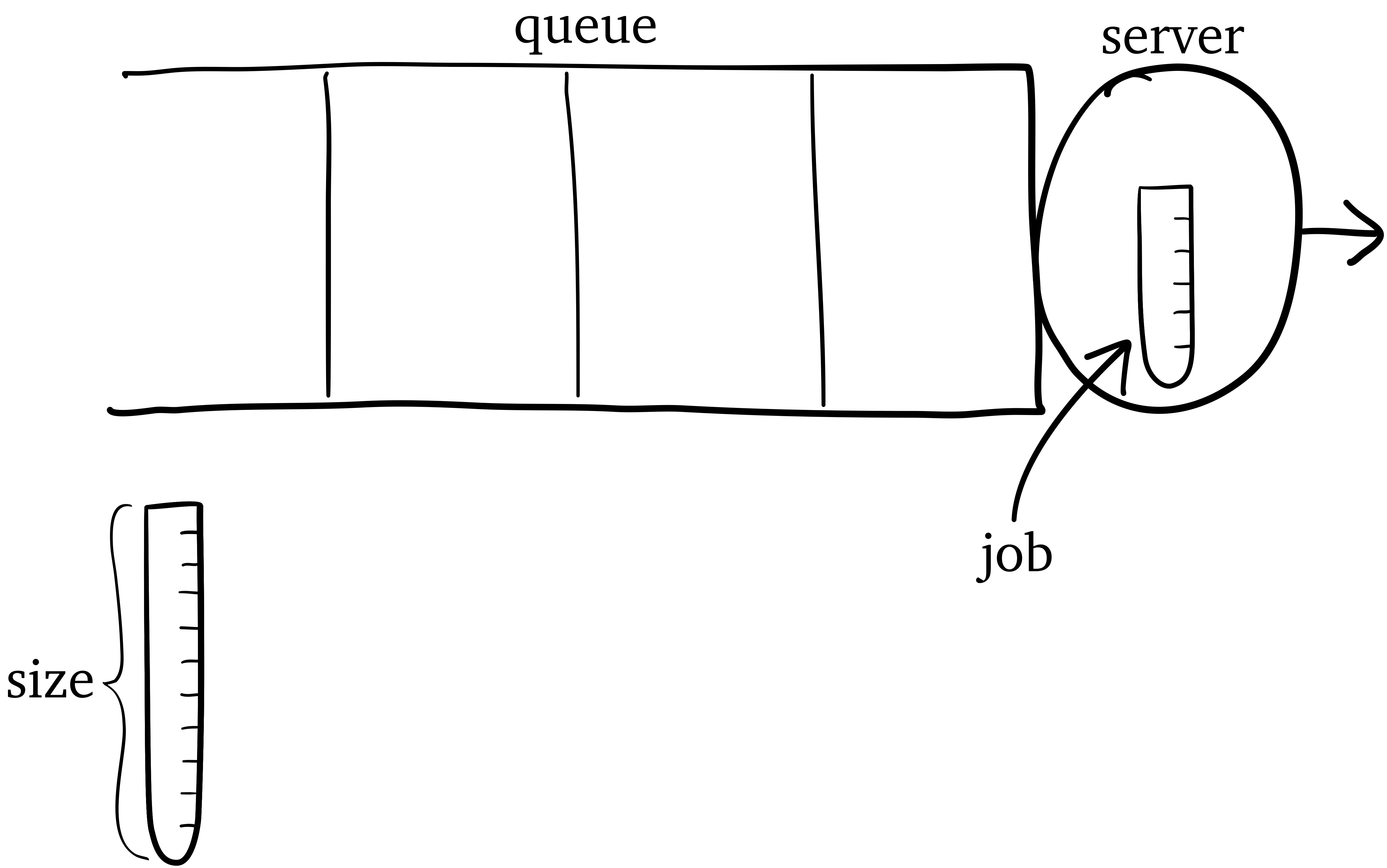
M/G/1 Queue



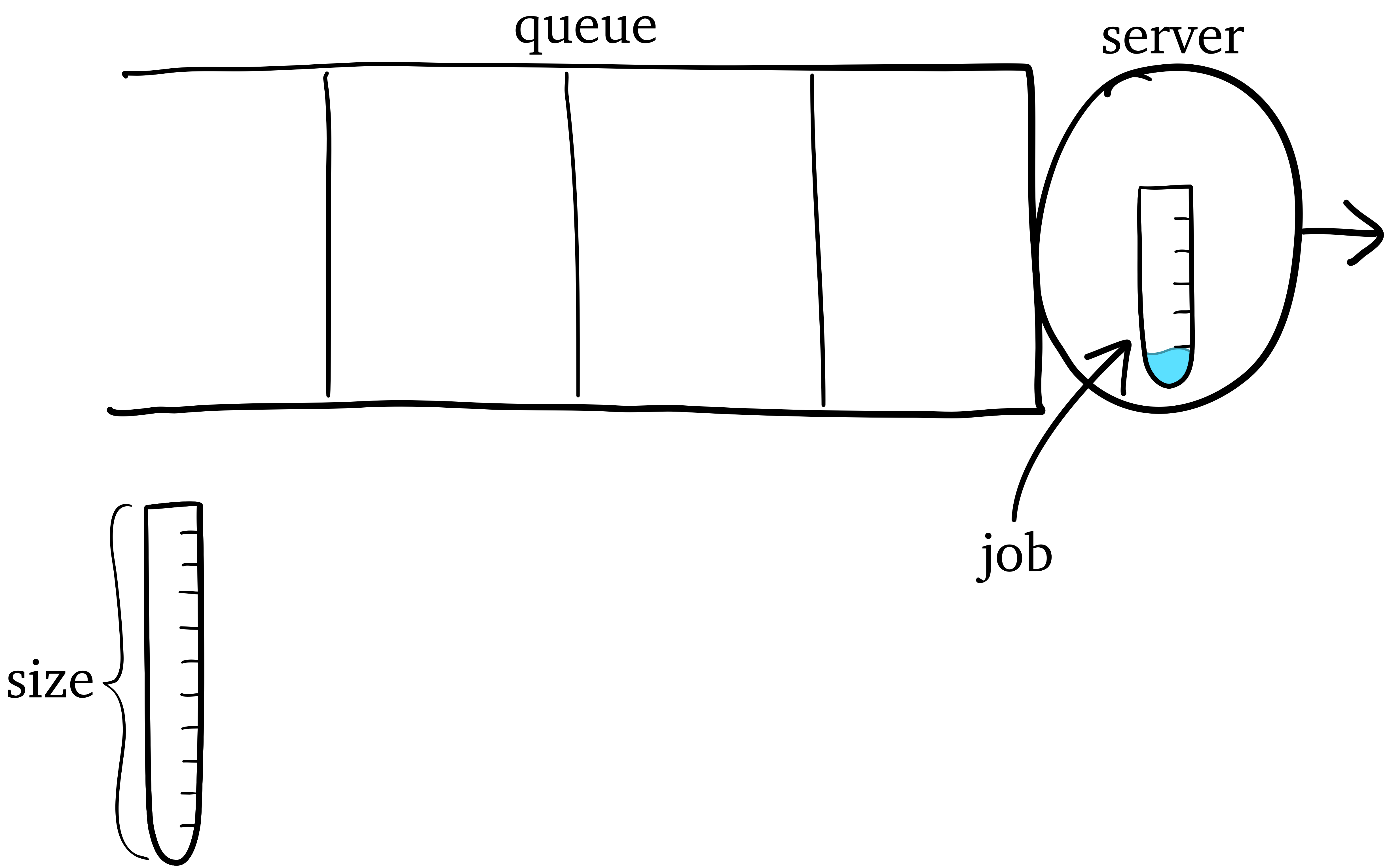
M/G/1 Queue



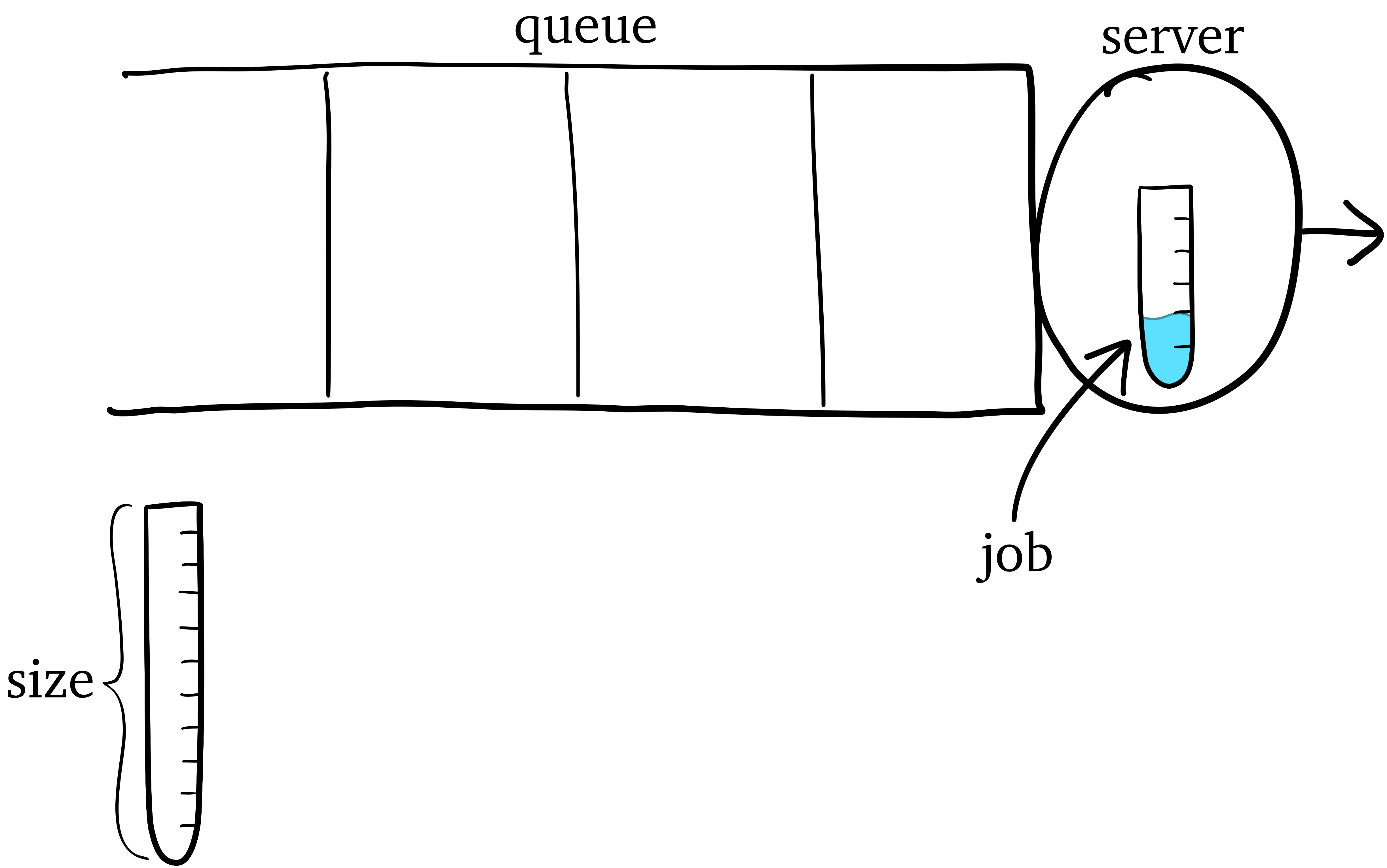
M/G/1 Queue



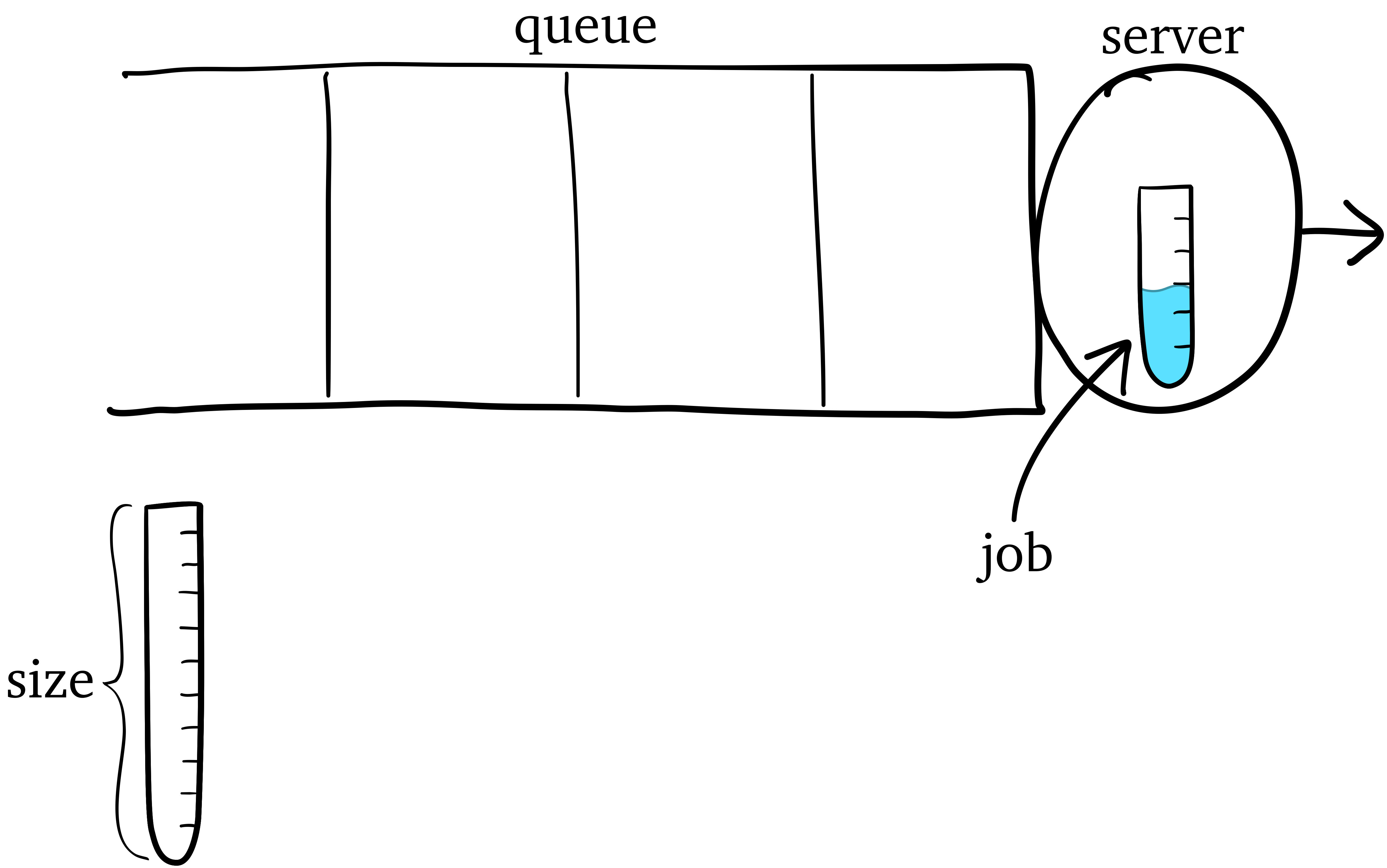
M/G/1 Queue



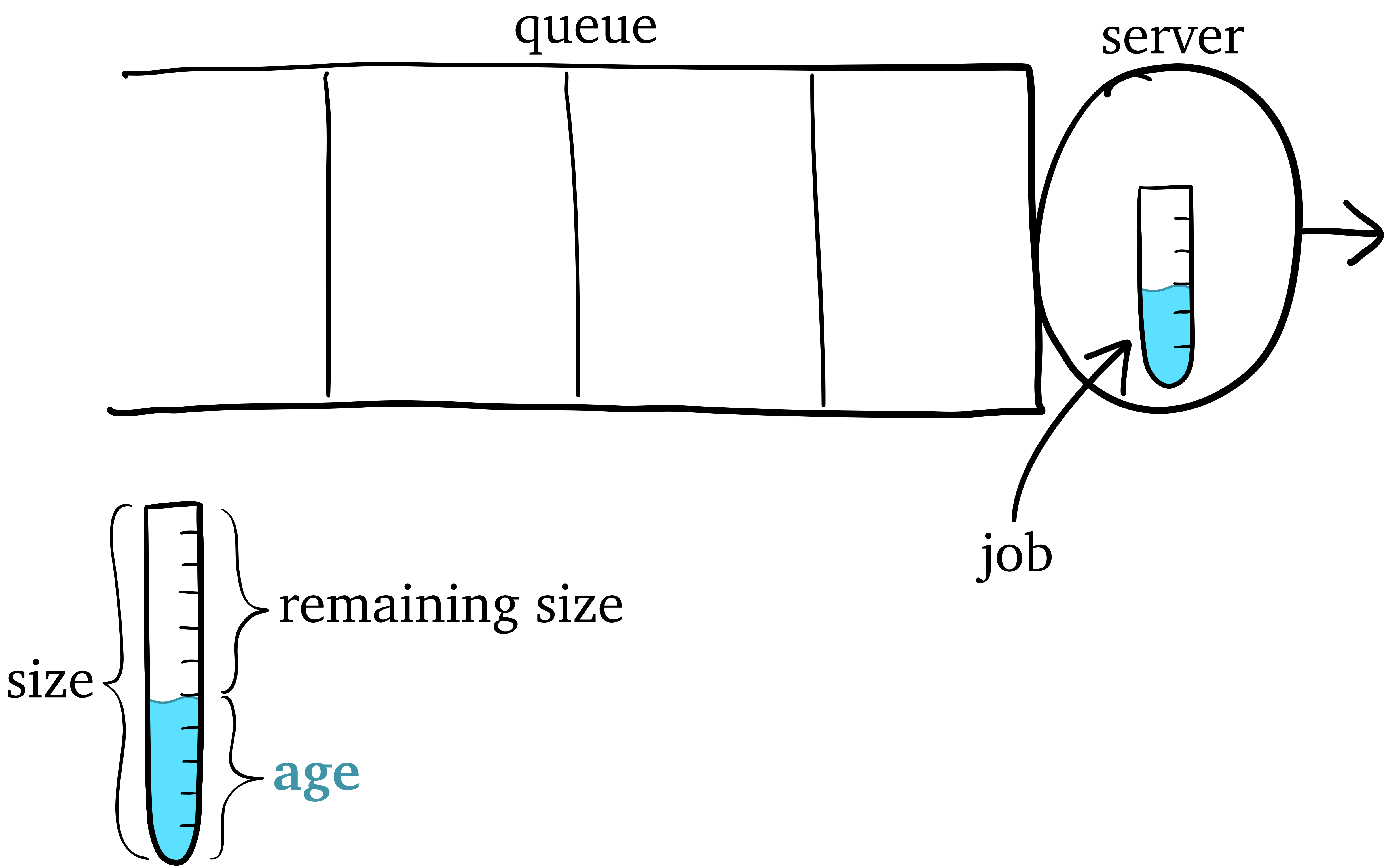
M/G/1 Queue



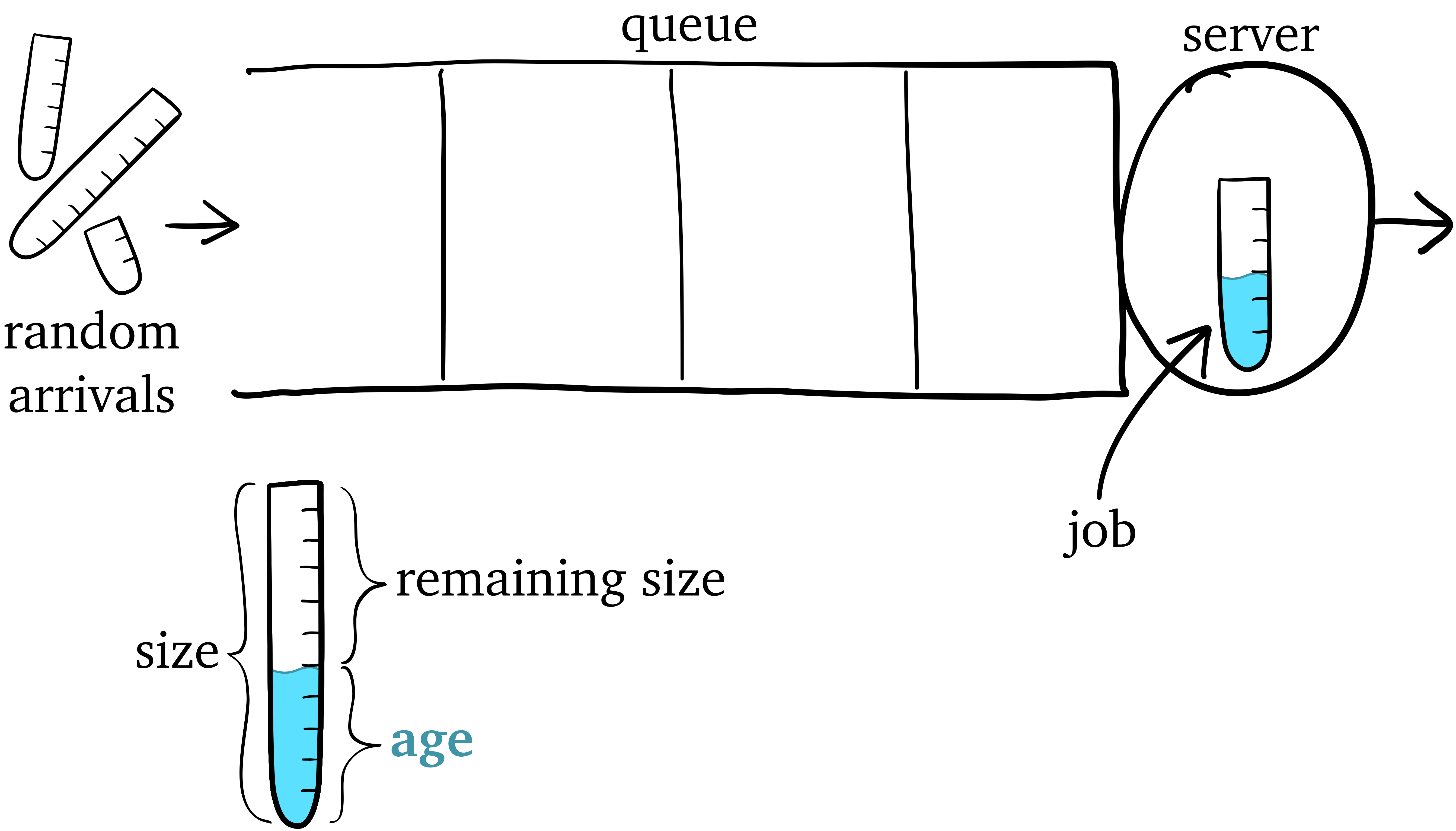
M/G/1 Queue



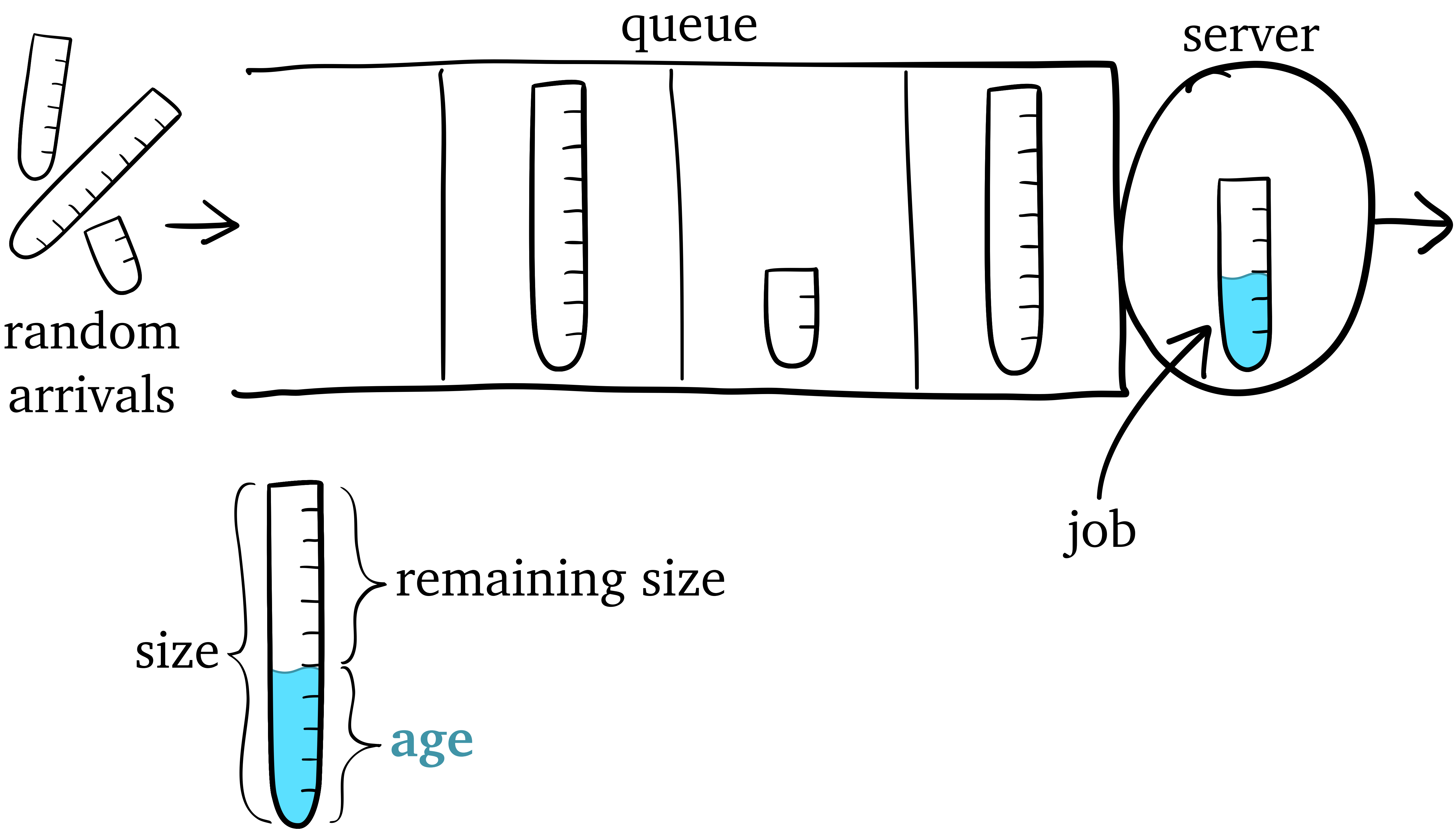
M/G/1 Queue



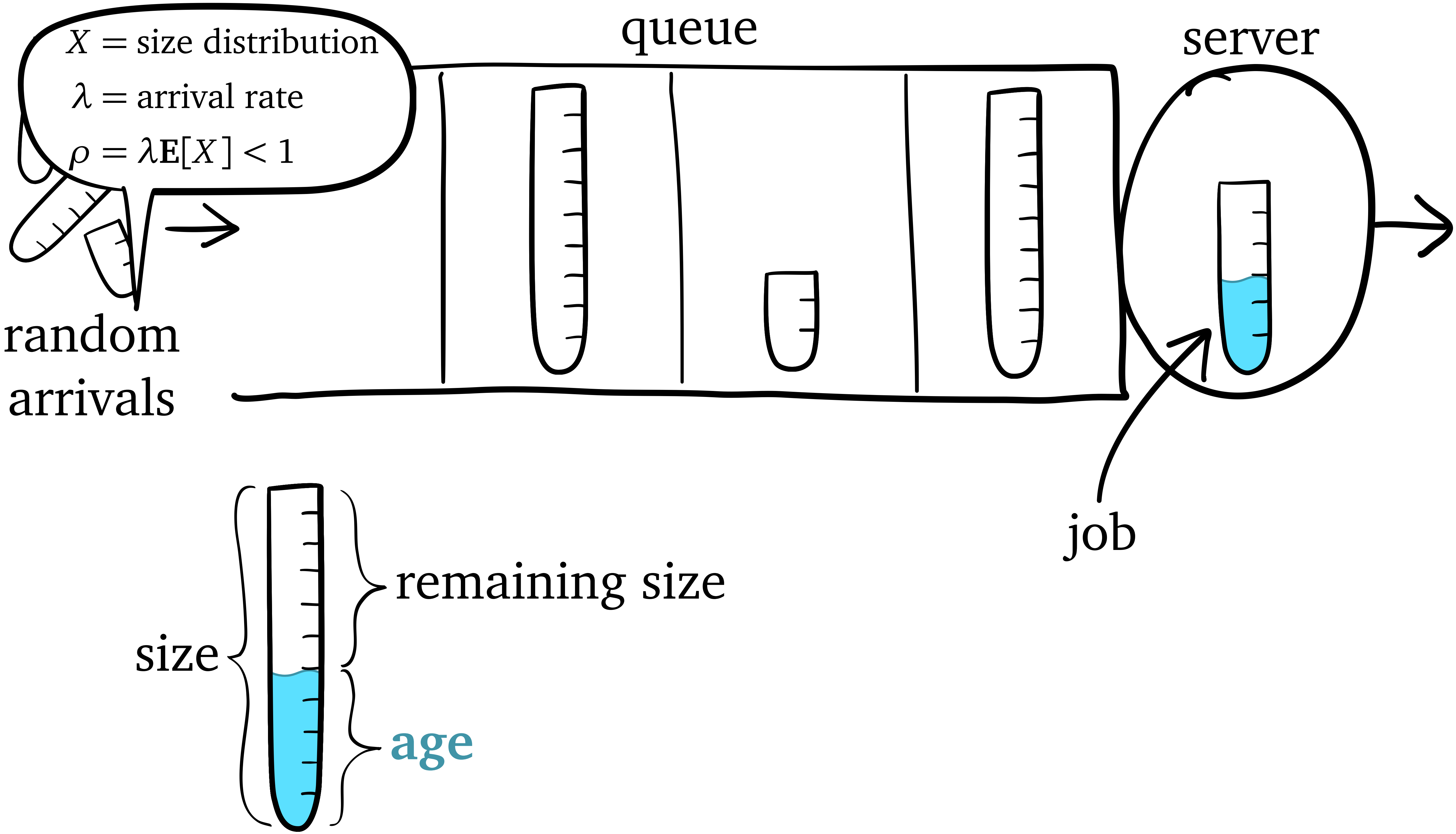
M/G/1 Queue



M/G/1 Queue

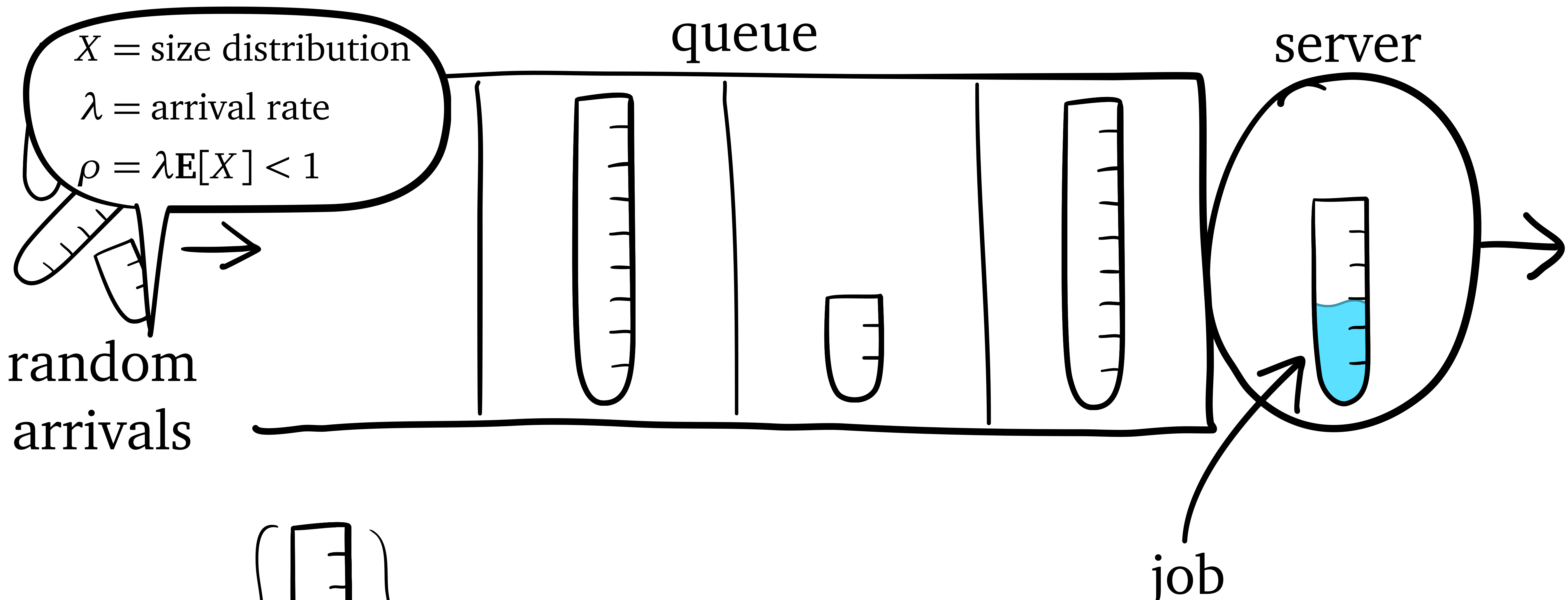


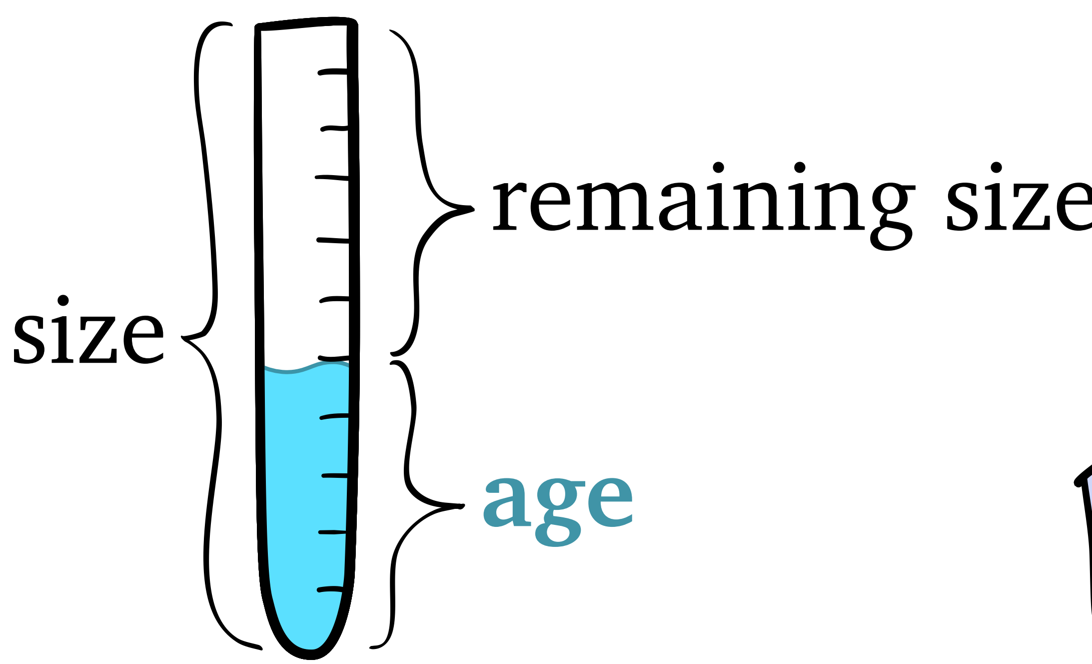
M/G/1 Queue

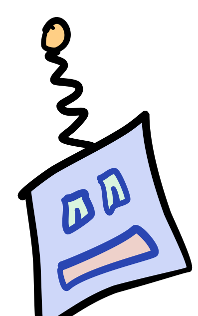


X = size distribution
λ = arrival rate
$\rho = \lambda \mathbf{E}[X] < 1$
random arrivals
queue
server
job
size
remaining size
age

M/G/1 Queue

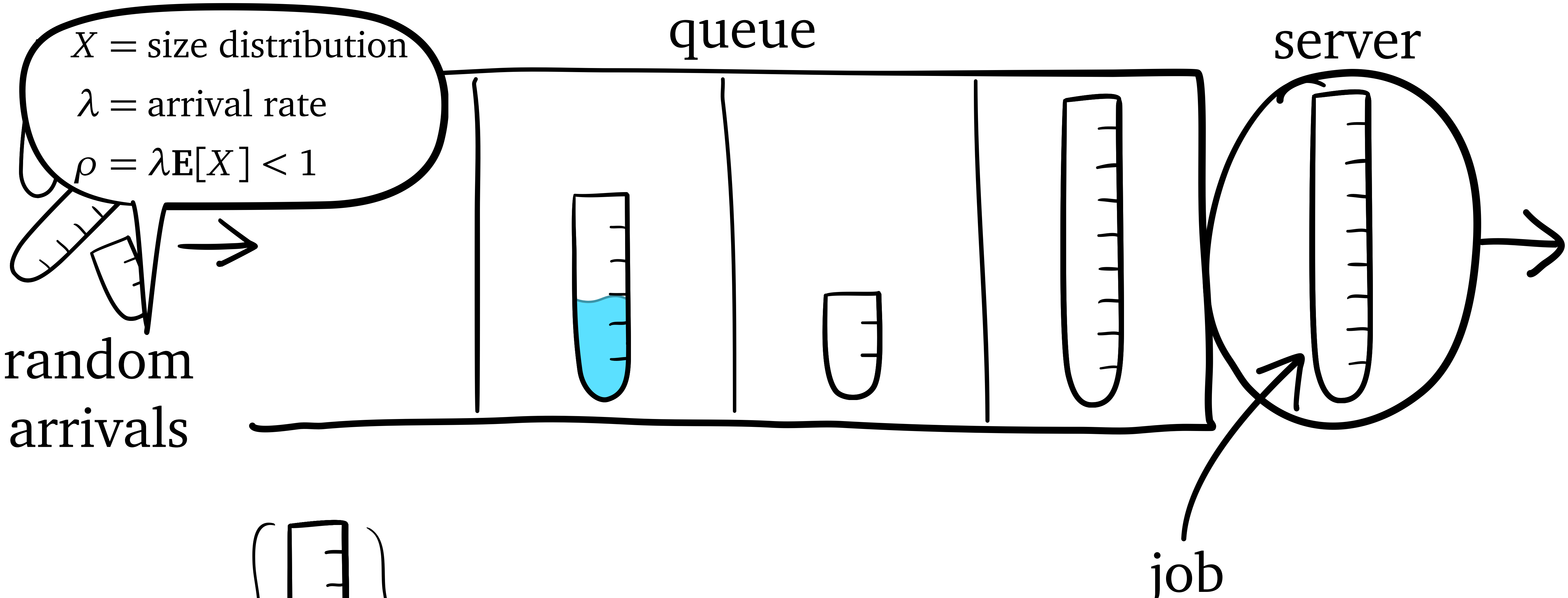


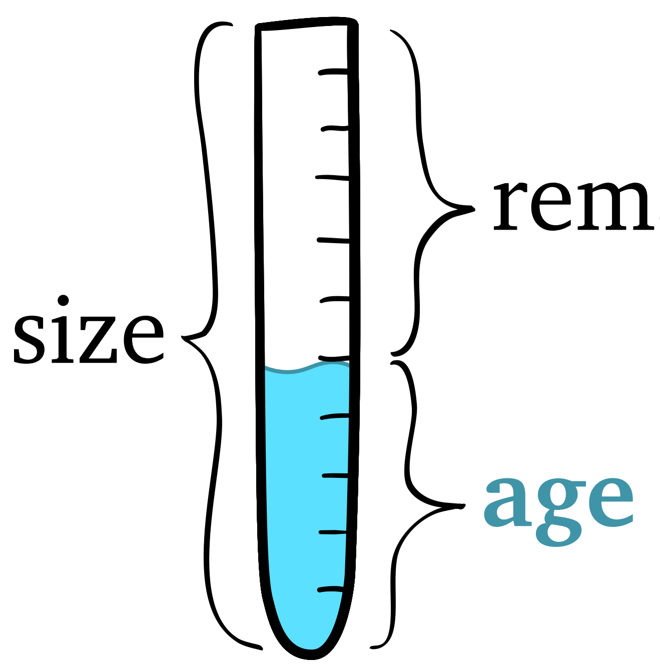


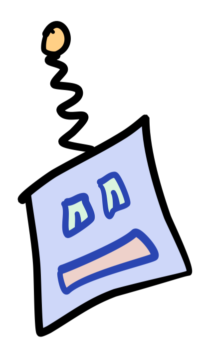


Scheduling policy:
picks which job to serve

M/G/1 Queue

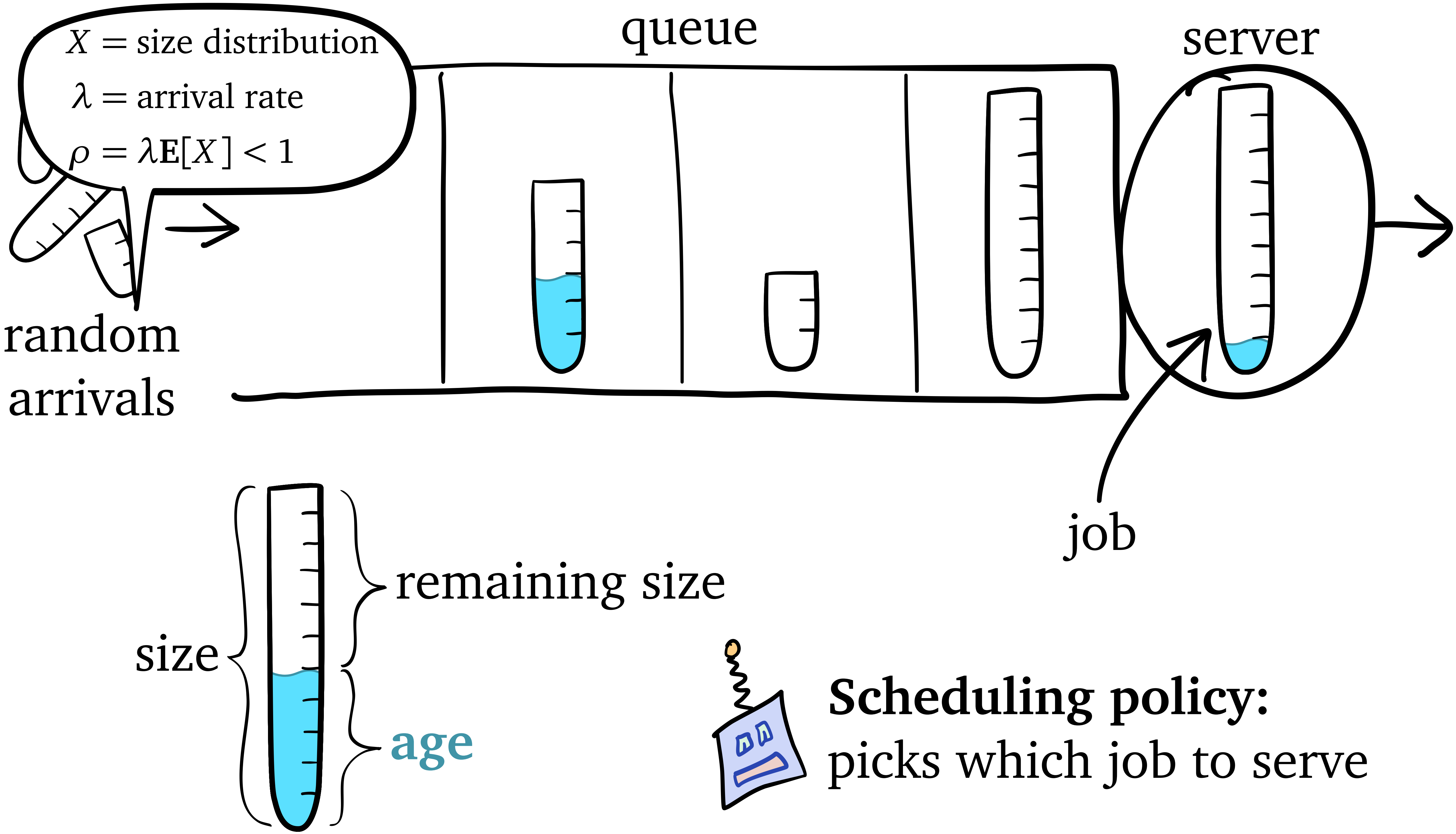






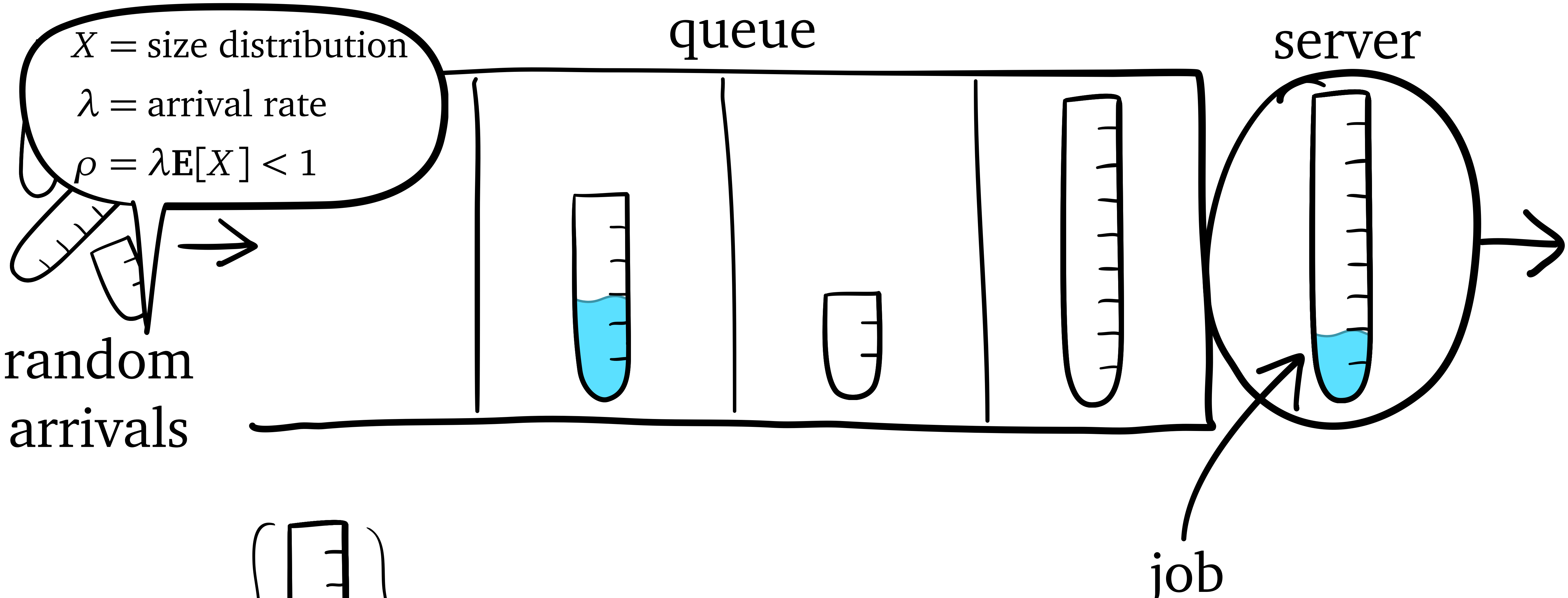
Scheduling policy:
picks which job to serve

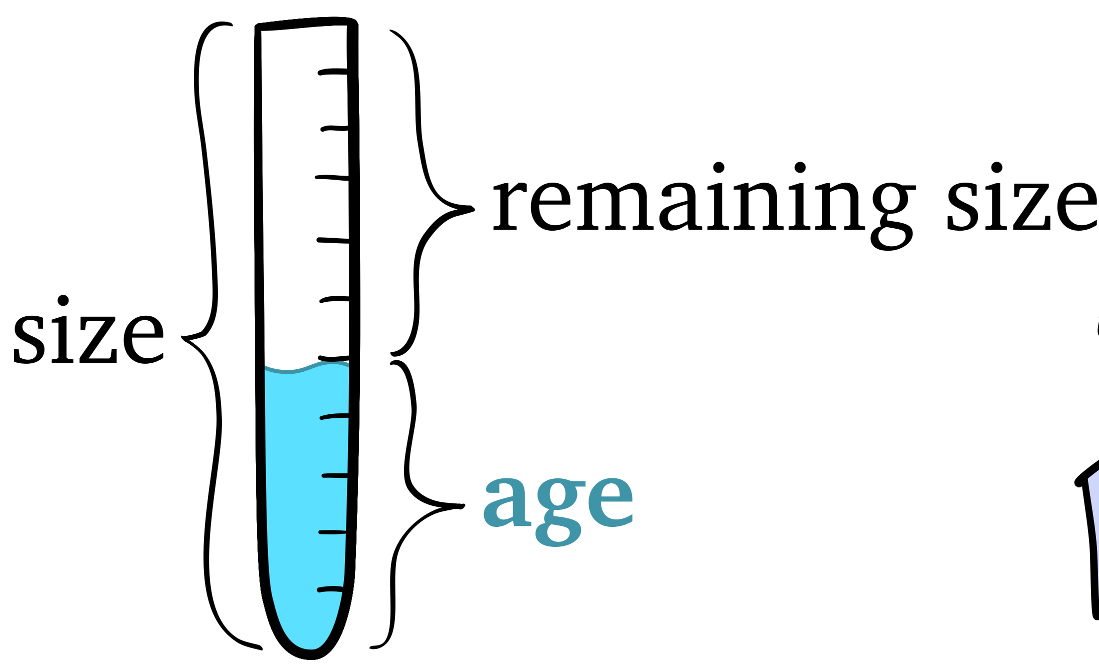
M/G/1 Queue

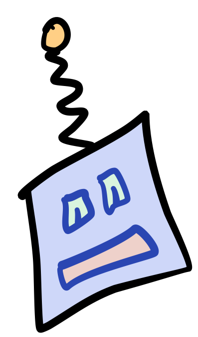


Scheduling policy:
picks which job to serve

M/G/1 Queue

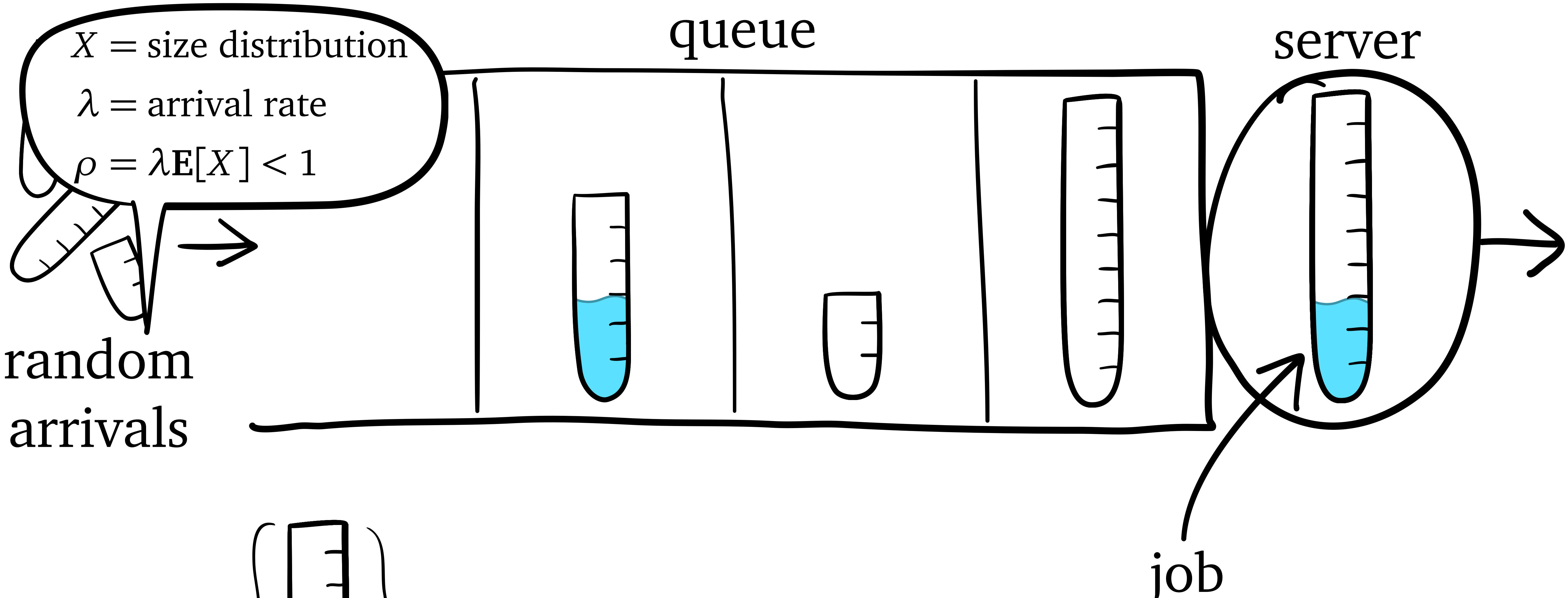


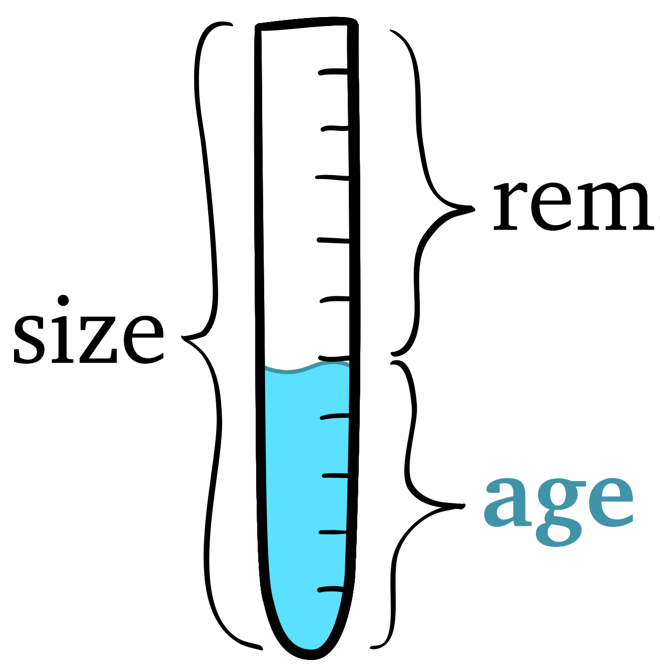




Scheduling policy:
picks which job to serve

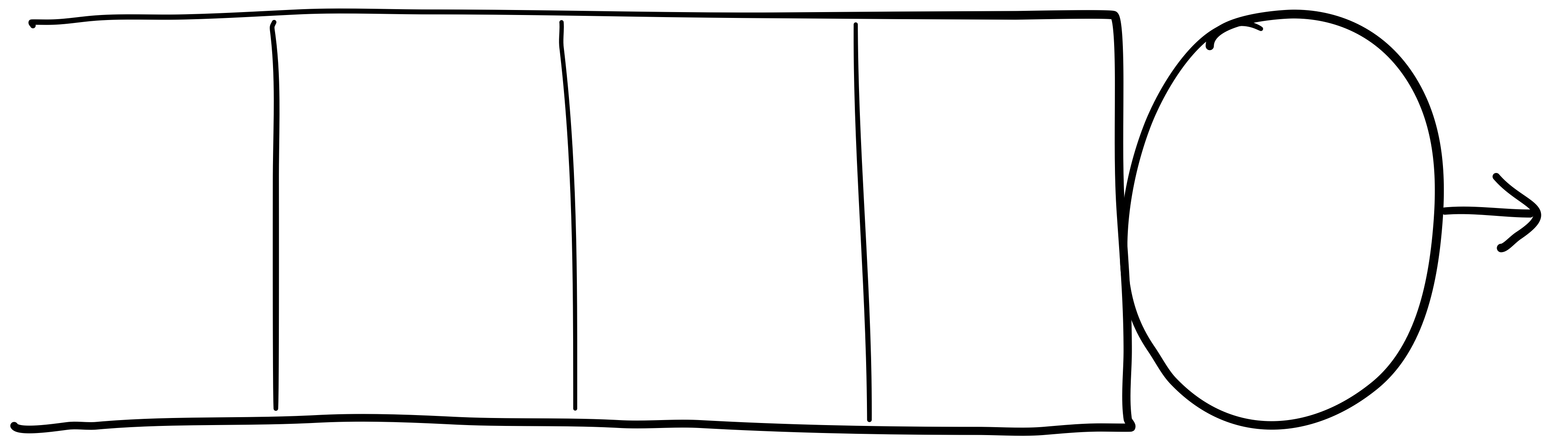
M/G/1 Queue



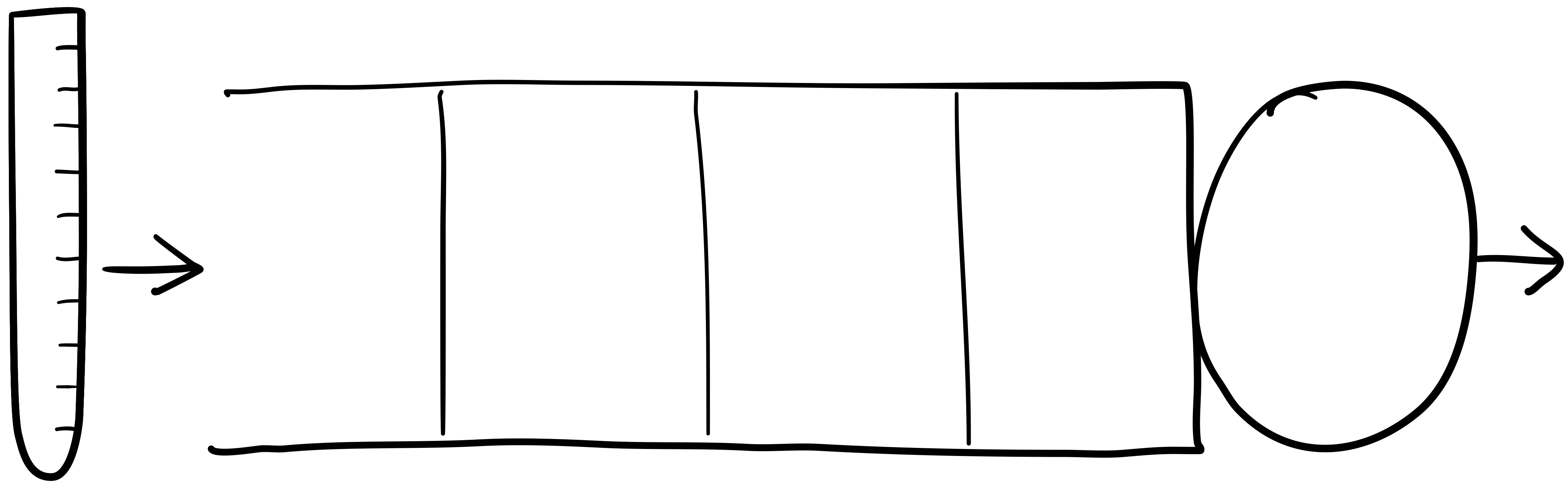


Scheduling policy:
picks which job to serve

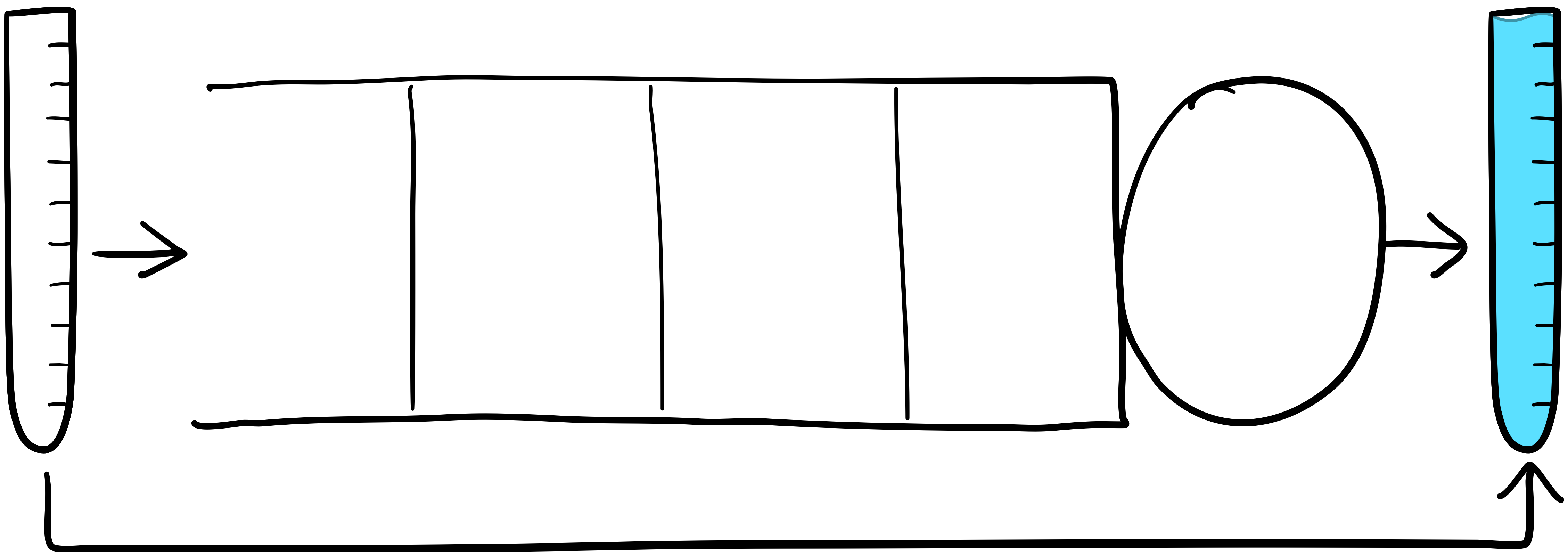
Response Time

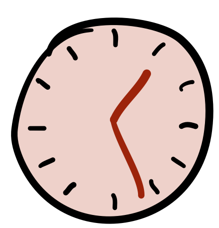


Response Time

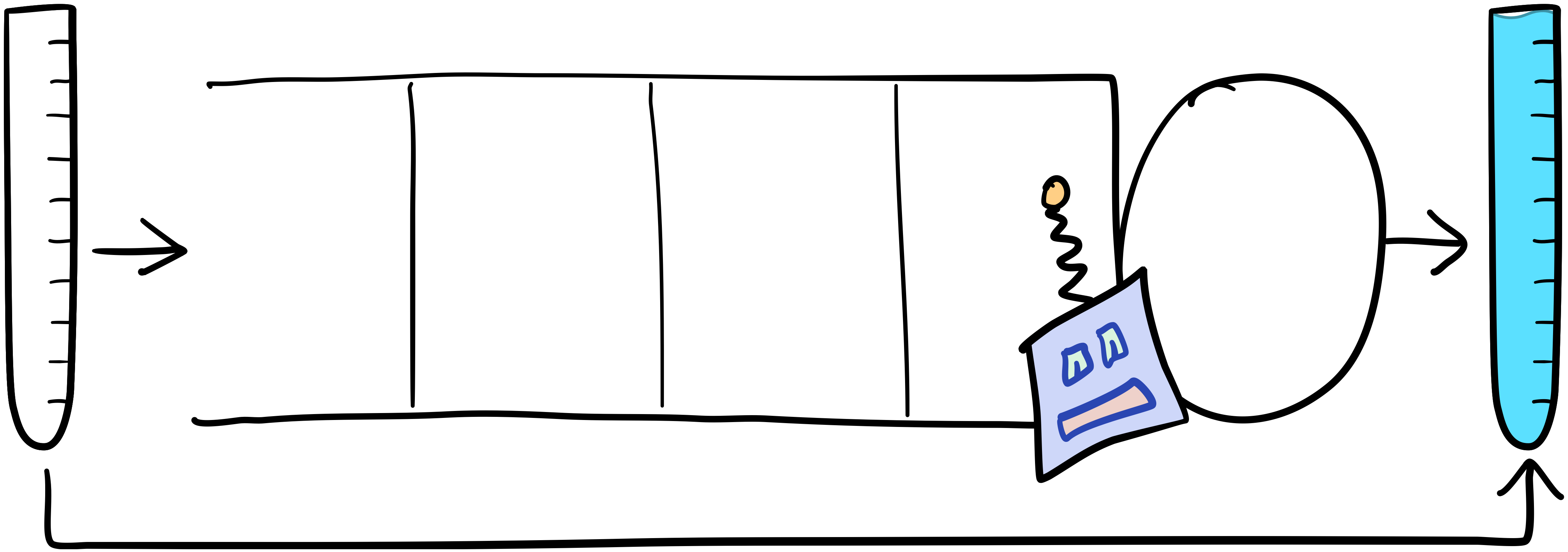


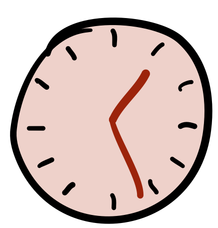
Response Time



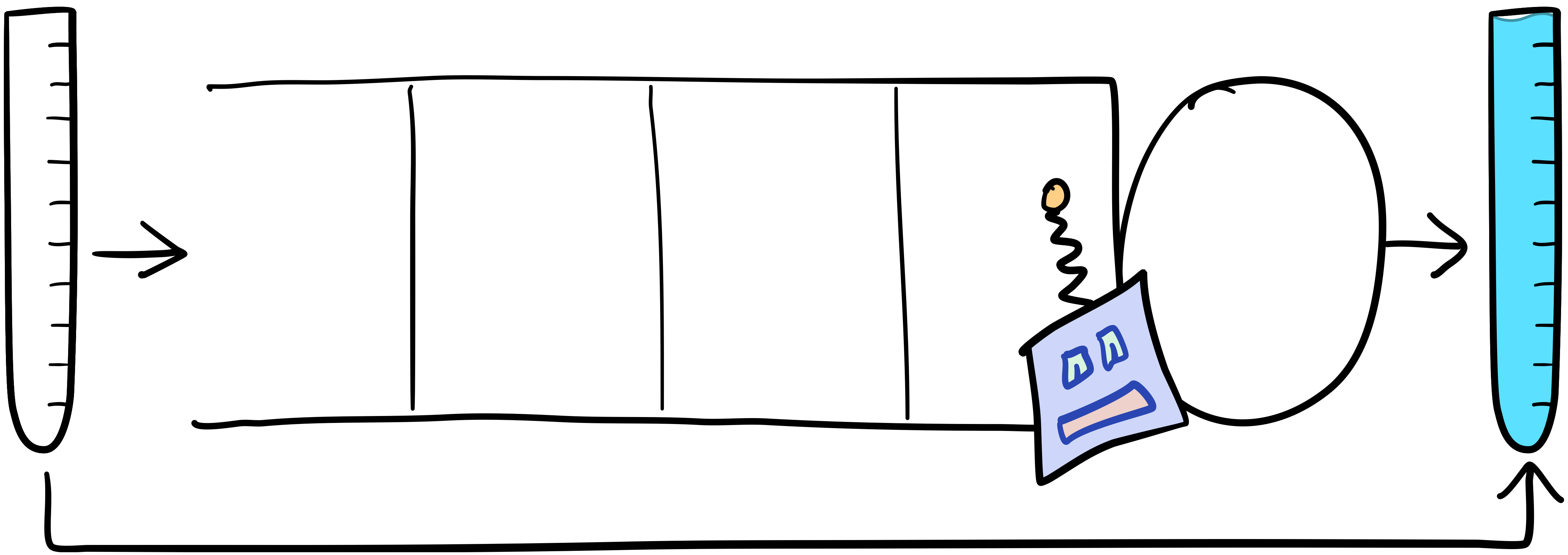
 $= T =$ *response time*

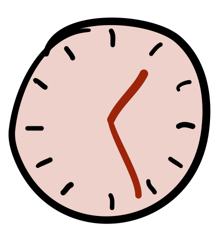
Response Time



 $= T = \textit{response time}$

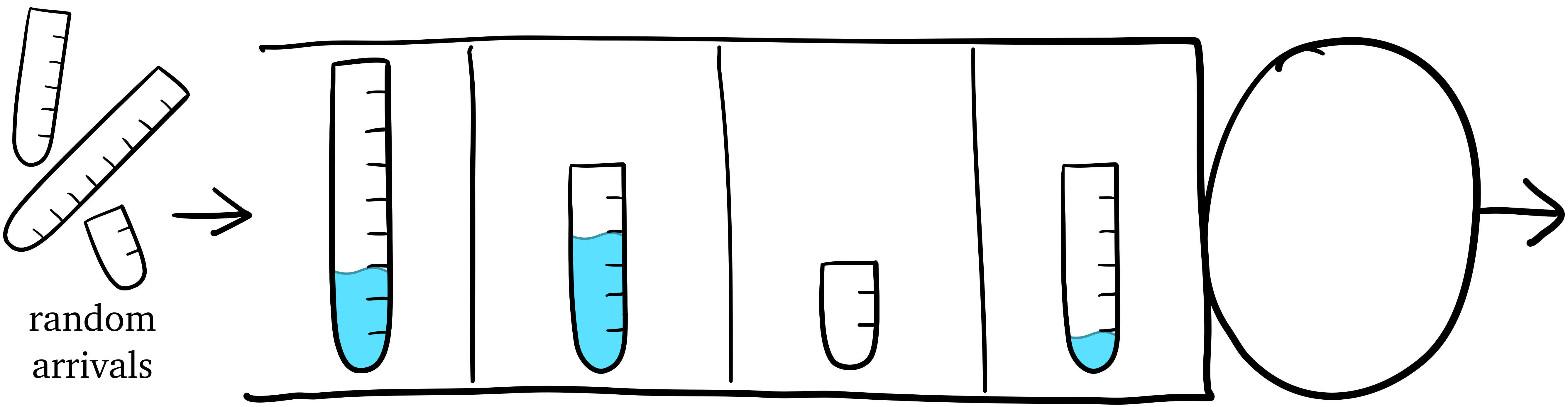
Response Time



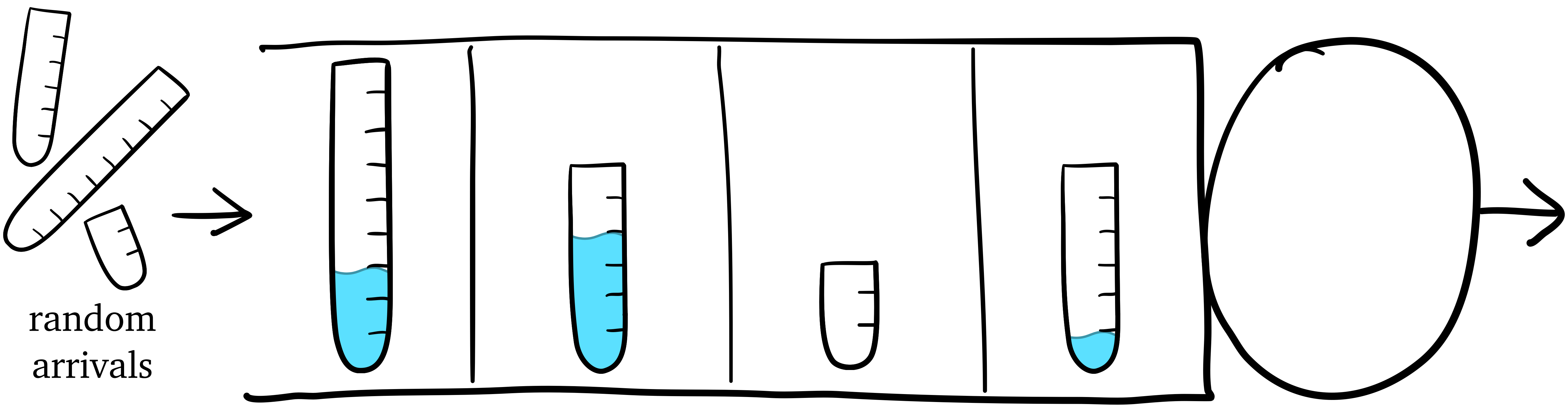
 = T = *response time*

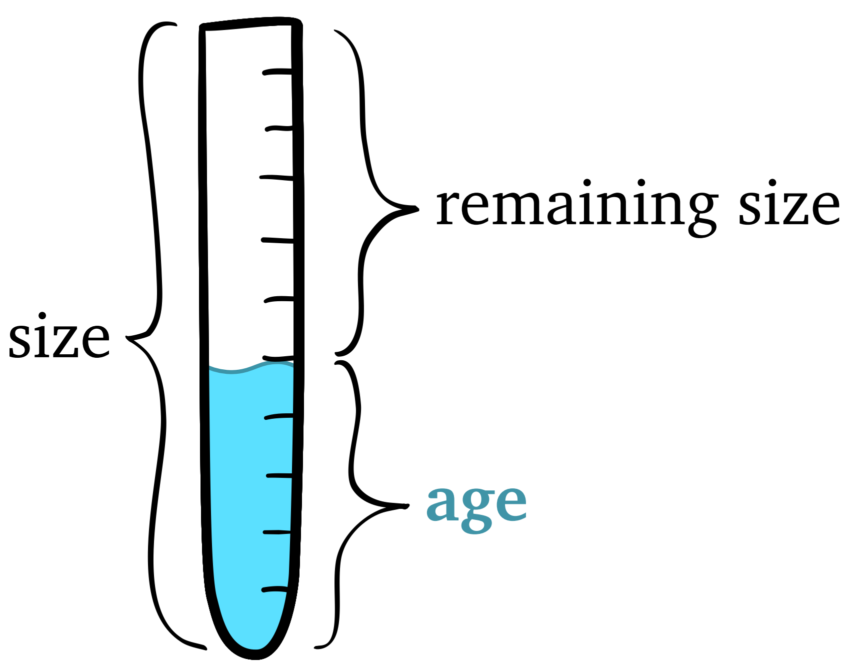
Goal: schedule to minimize
mean response time $\mathbf{E}[T]$

How to Schedule?

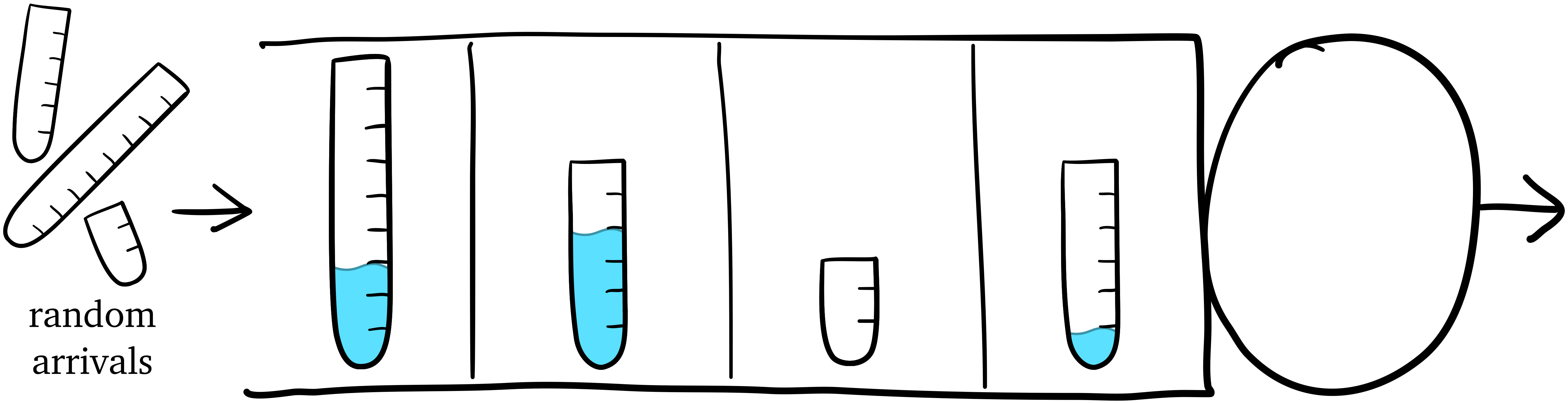


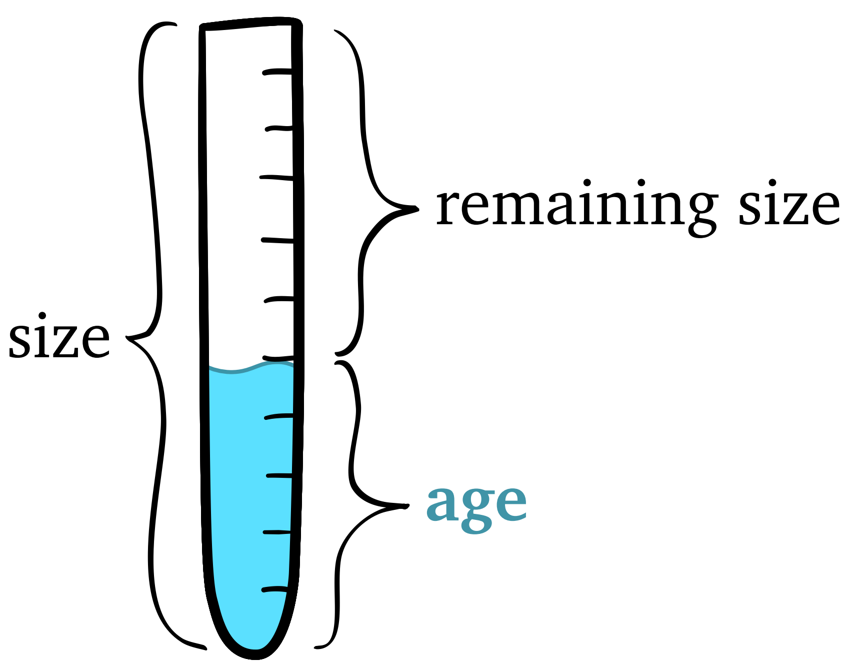
How to Schedule?

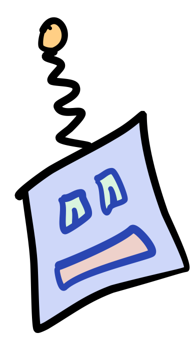




How to Schedule?

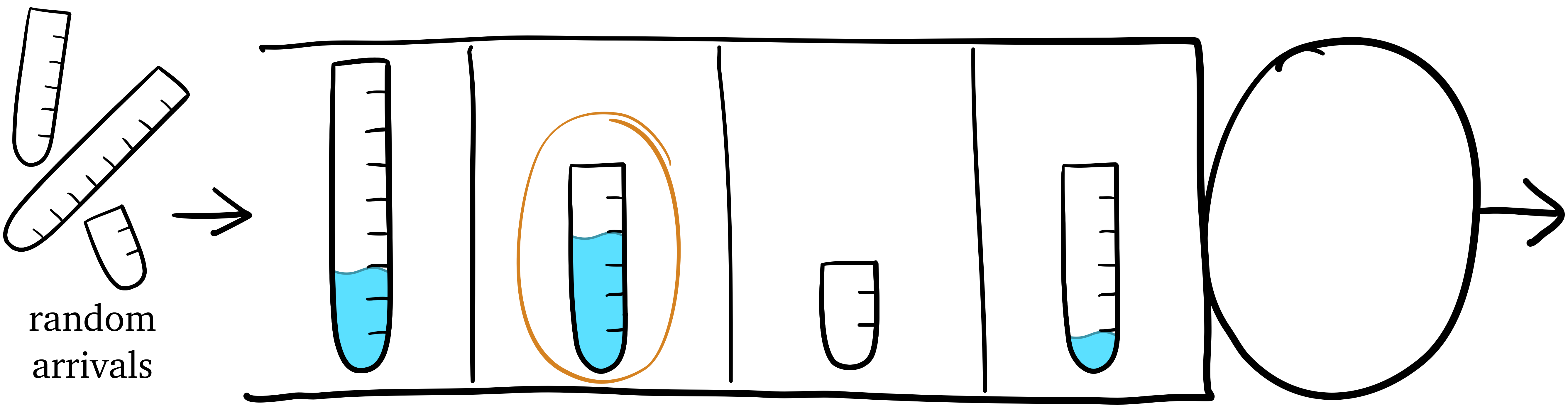


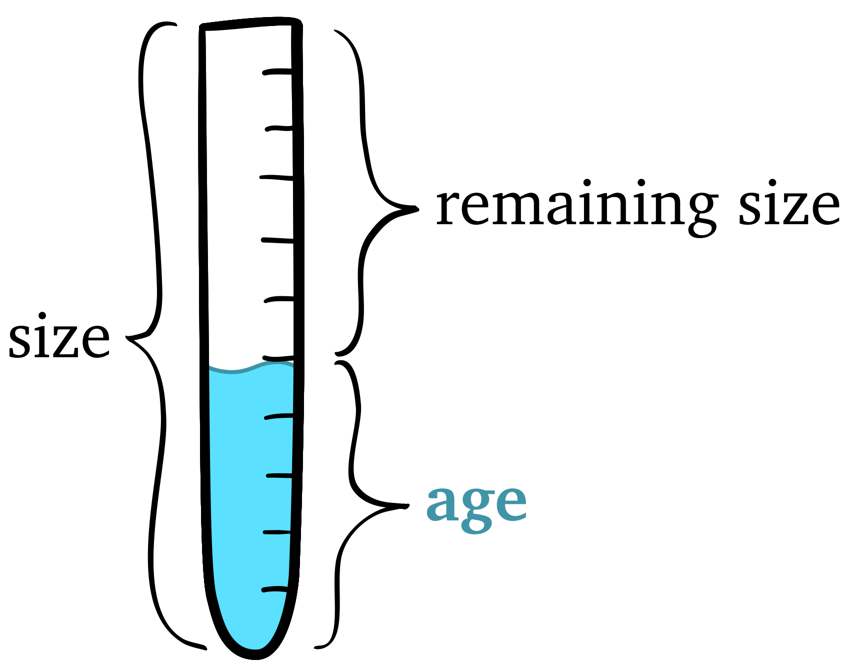


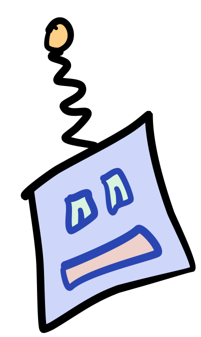


SRPT: always serve job of *least remaining size*

How to Schedule?

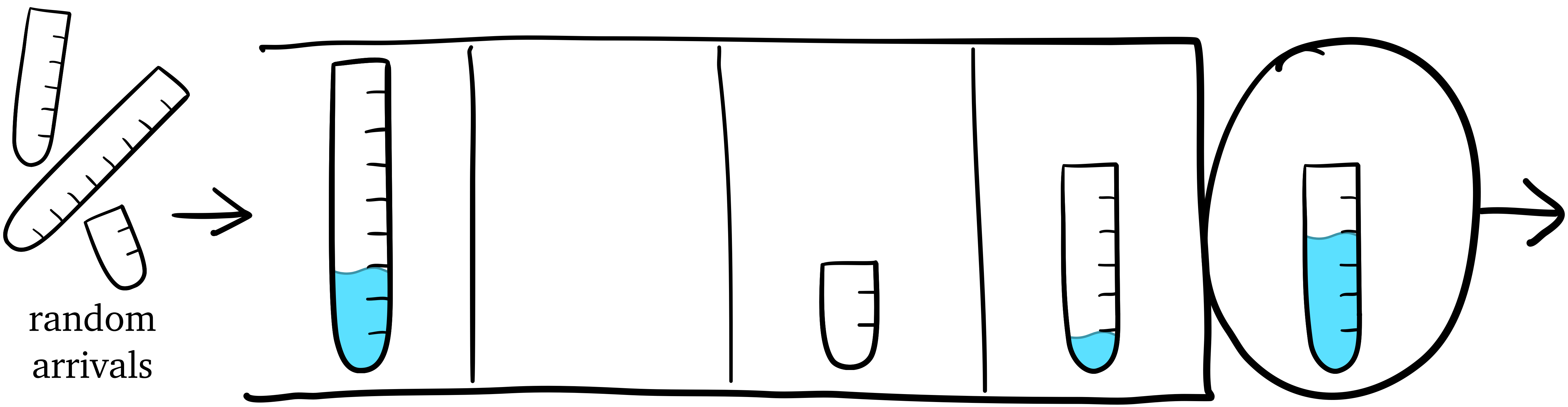


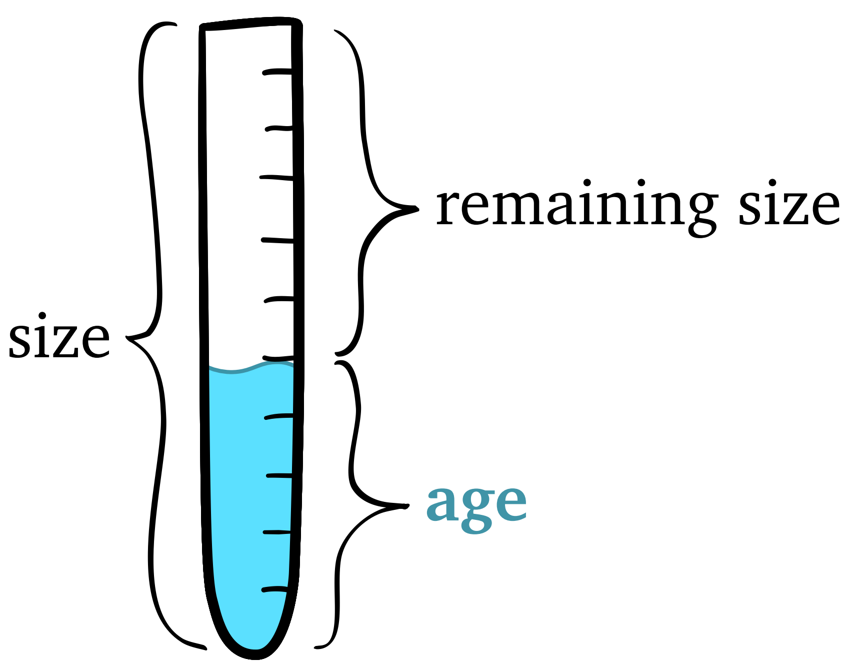


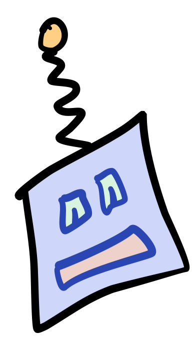


SRPT: always serve job of *least remaining size*

How to Schedule?

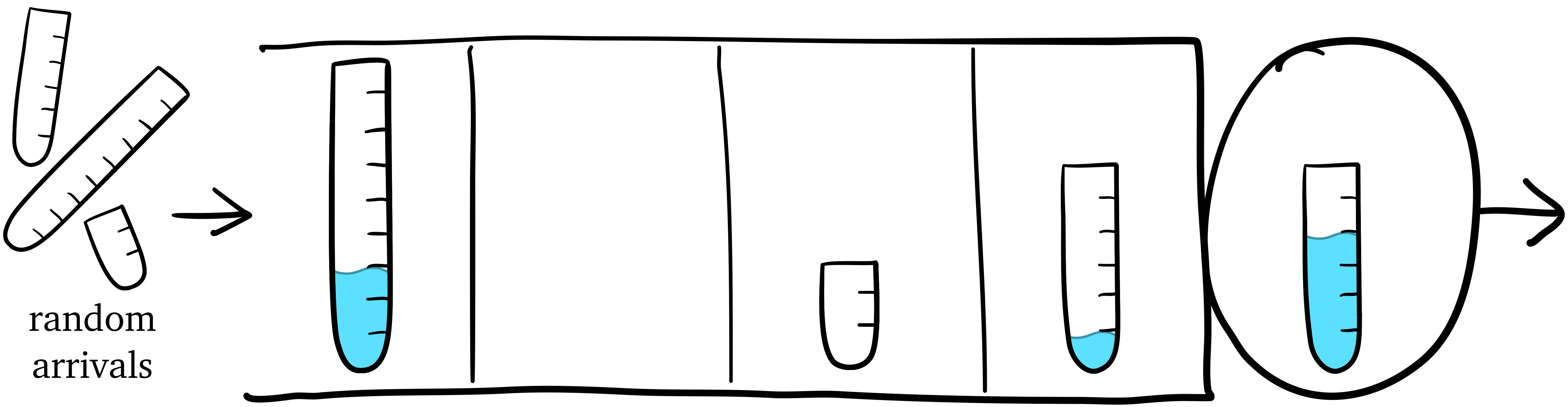


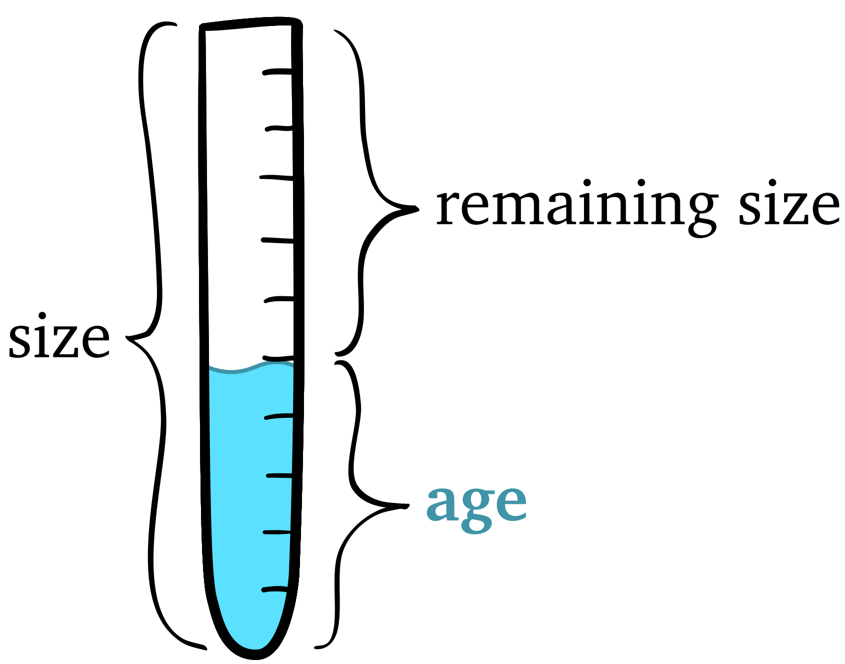


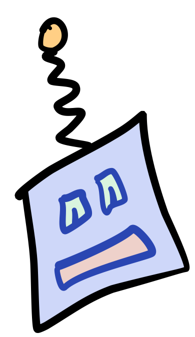


SRPT: always serve job of *least remaining size*

How to Schedule?



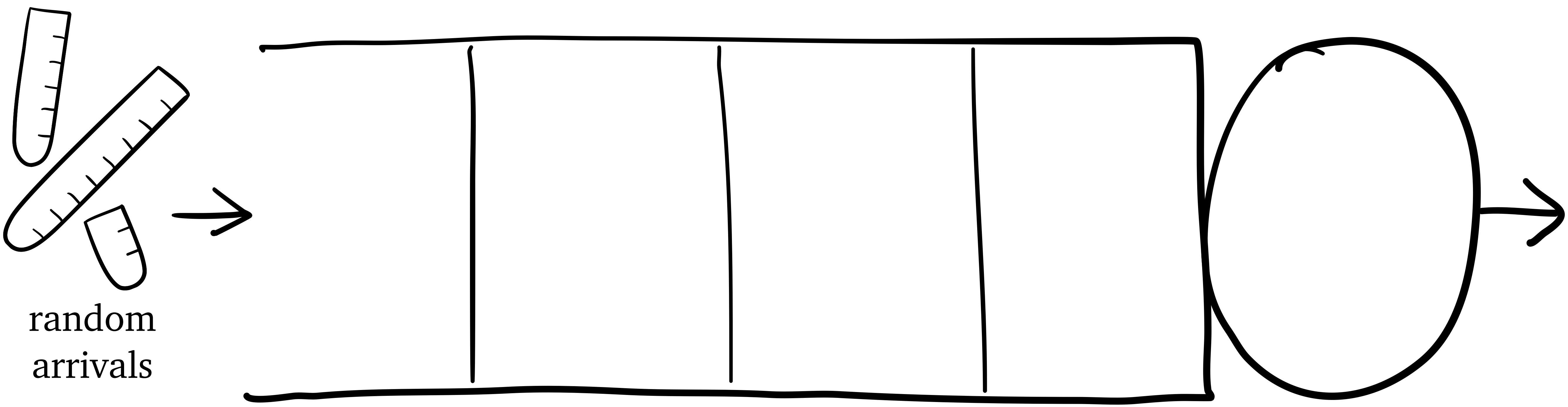




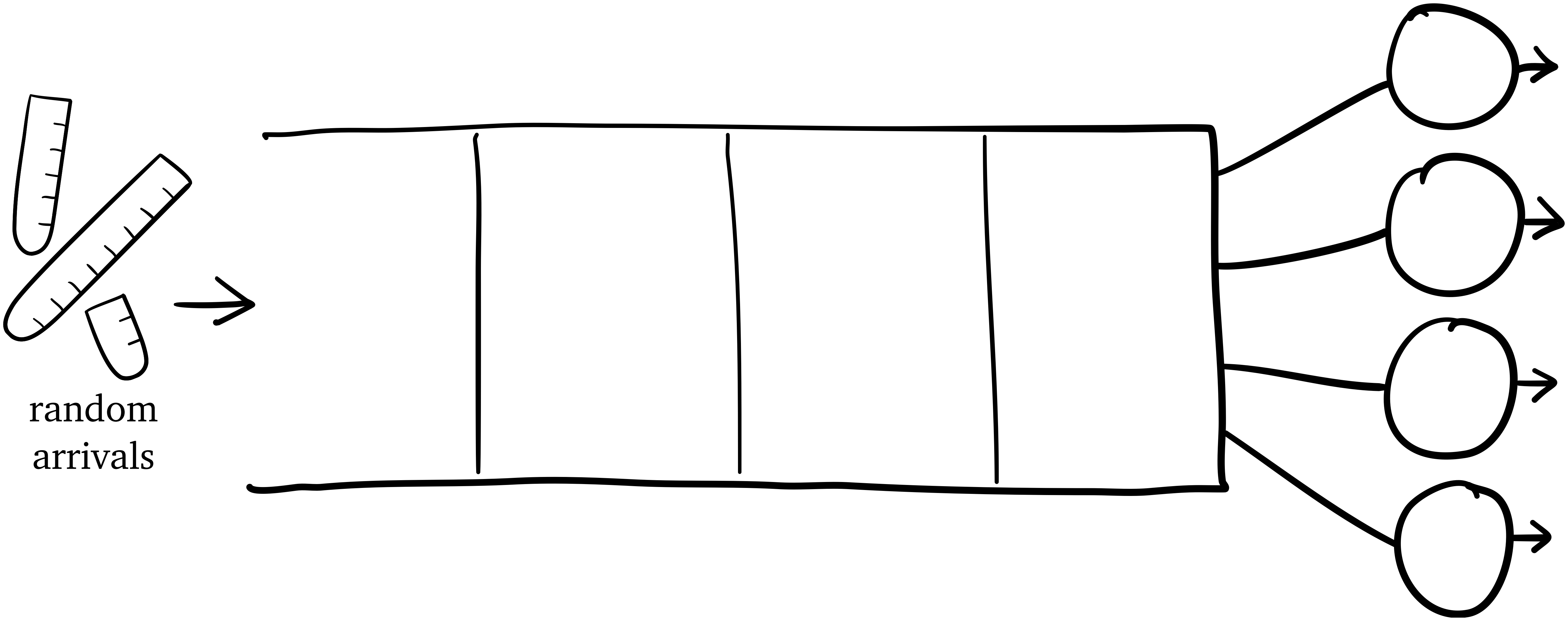
SRPT: always serve job of *least remaining size*

SRPT minimizes $\mathbf{E}[T]$

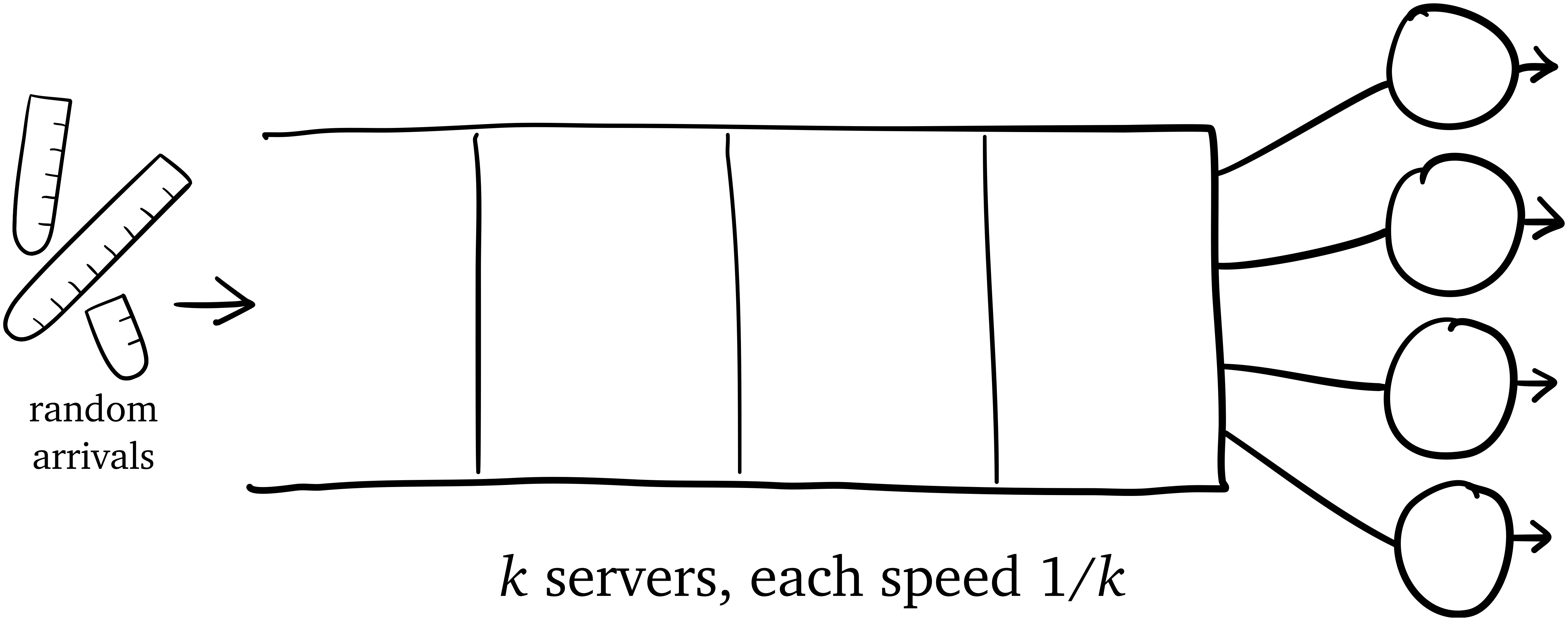
M/G/k Queue



M/G/k Queue

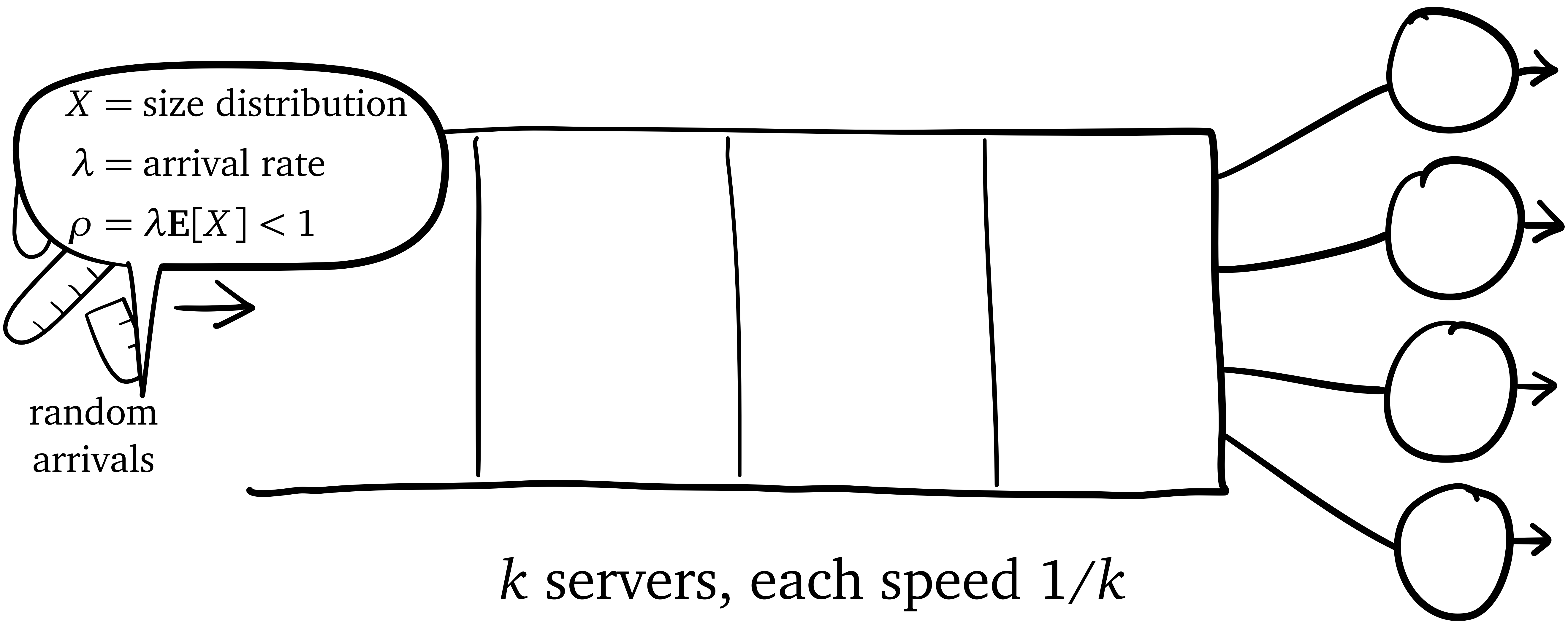


M/G/k Queue



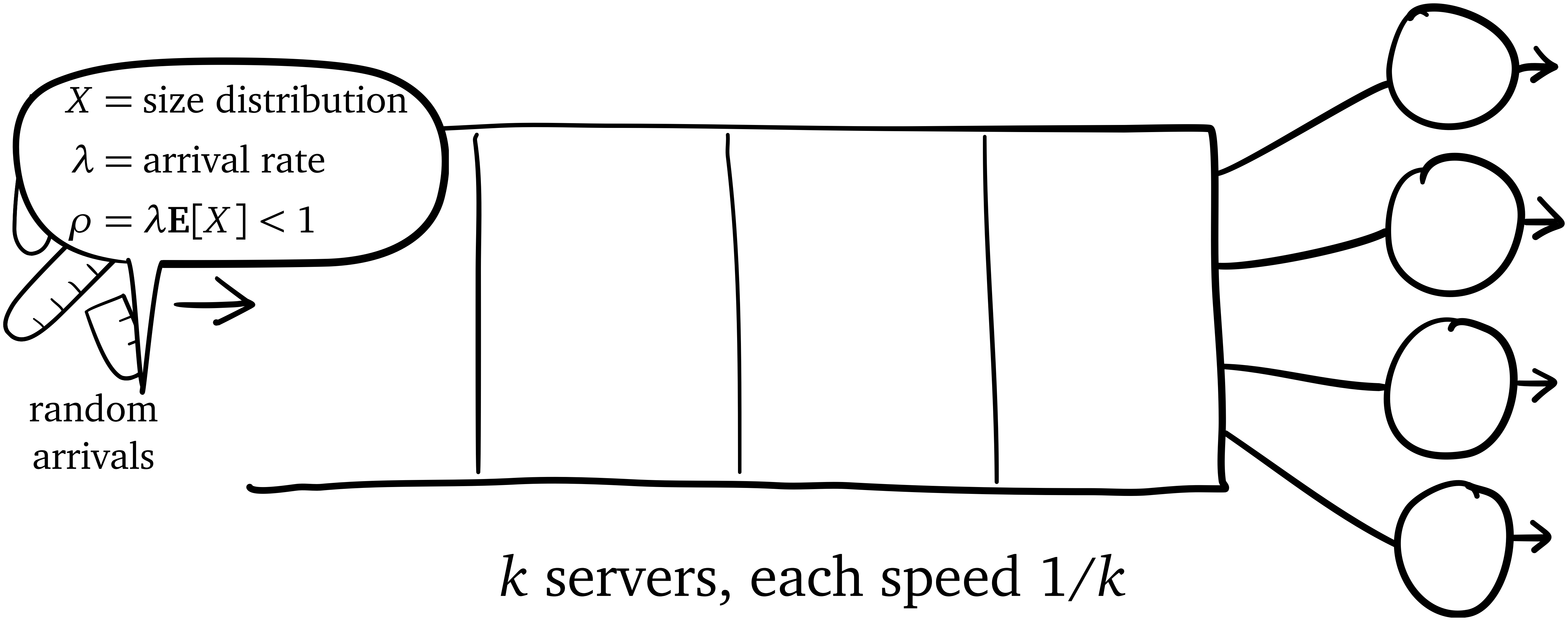
k servers, each speed $1/k$

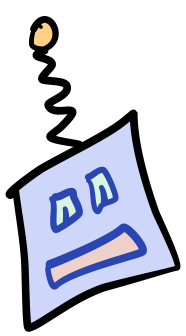
M/G/k Queue



X = size distribution
λ = arrival rate
$\rho = \lambda \mathbf{E}[X] < 1$
random arrivals
k servers, each speed $1/k$

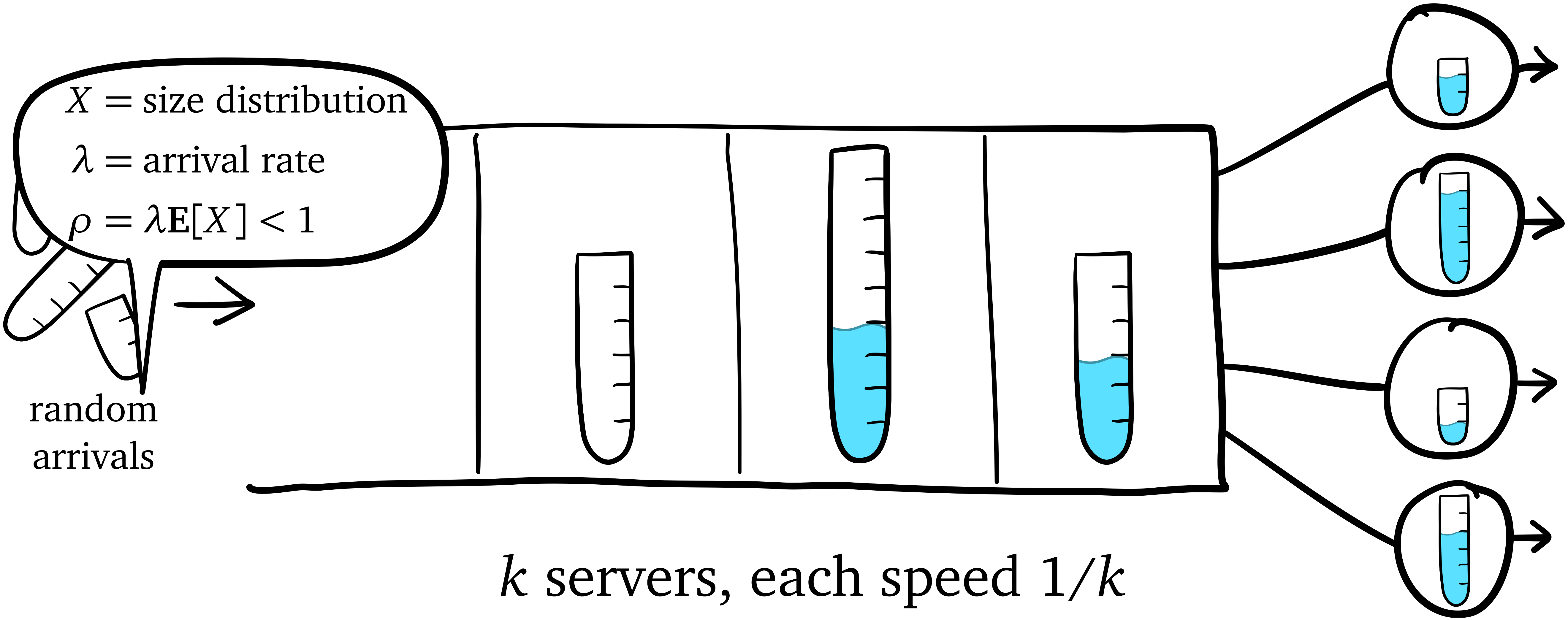
M/G/k Queue

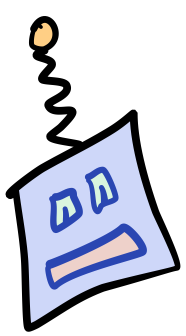




SRPT-*k*: always serve k jobs of *least remaining size*

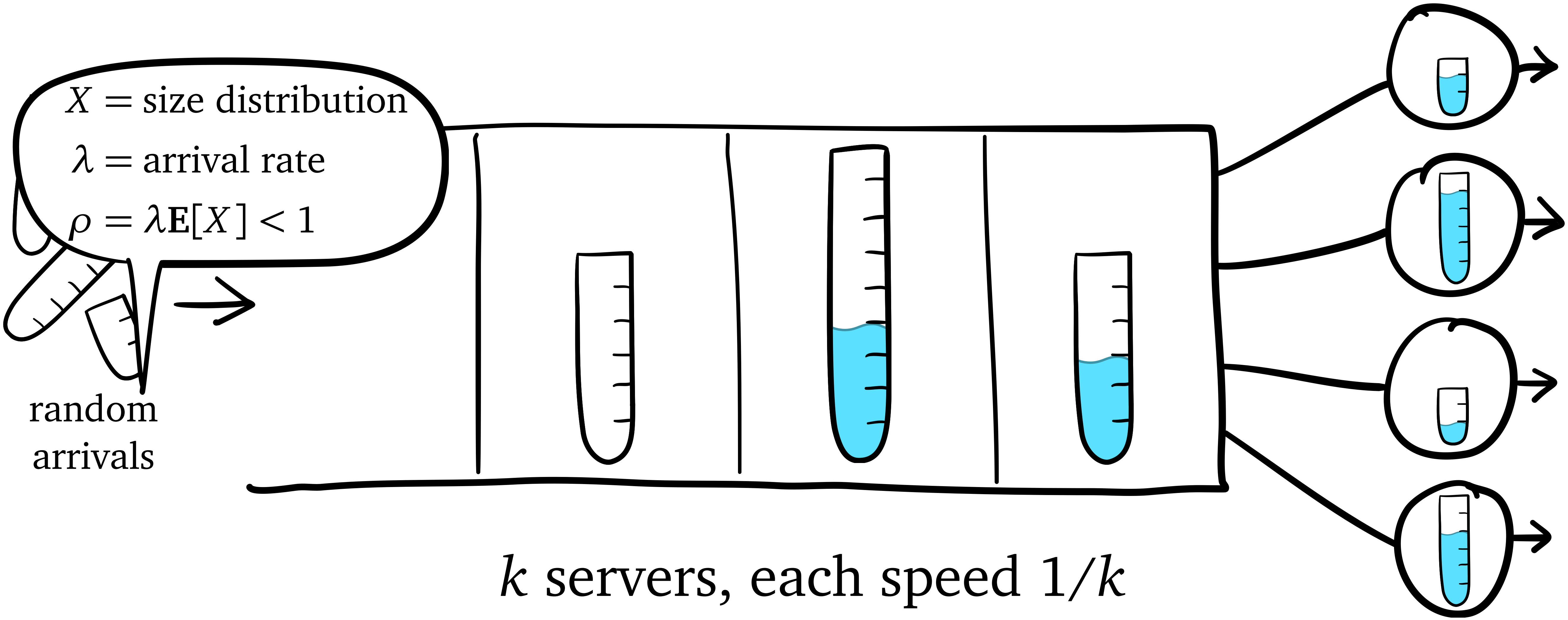
M/G/k Queue

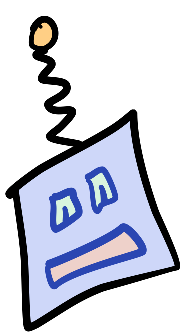




SRPT-k: always serve k jobs of *least remaining size*

M/G/k Queue

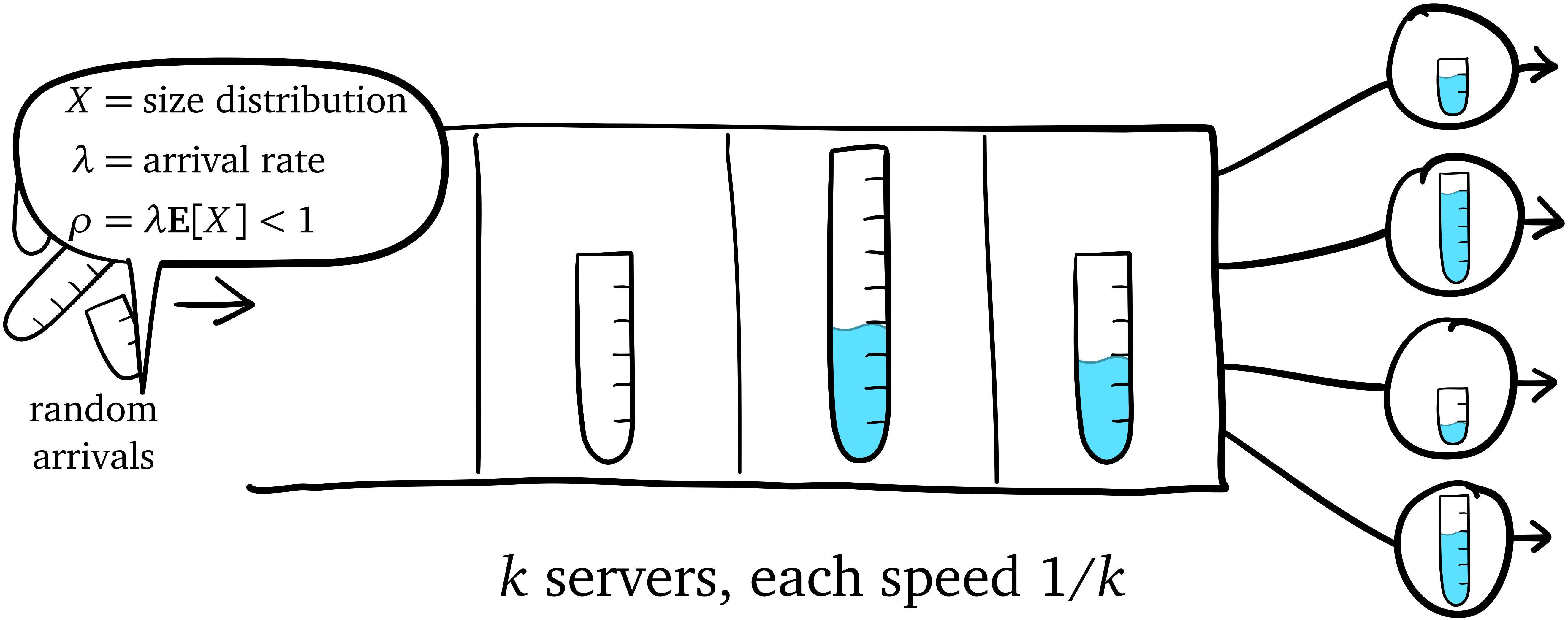


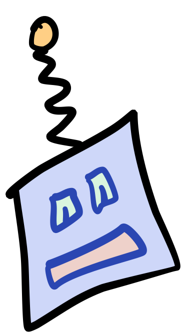


SRPT-k: always serve k jobs of *least remaining size*

SRPT-k minimizes $\mathbf{E}[T]$ in heavy traffic

M/G/k Queue



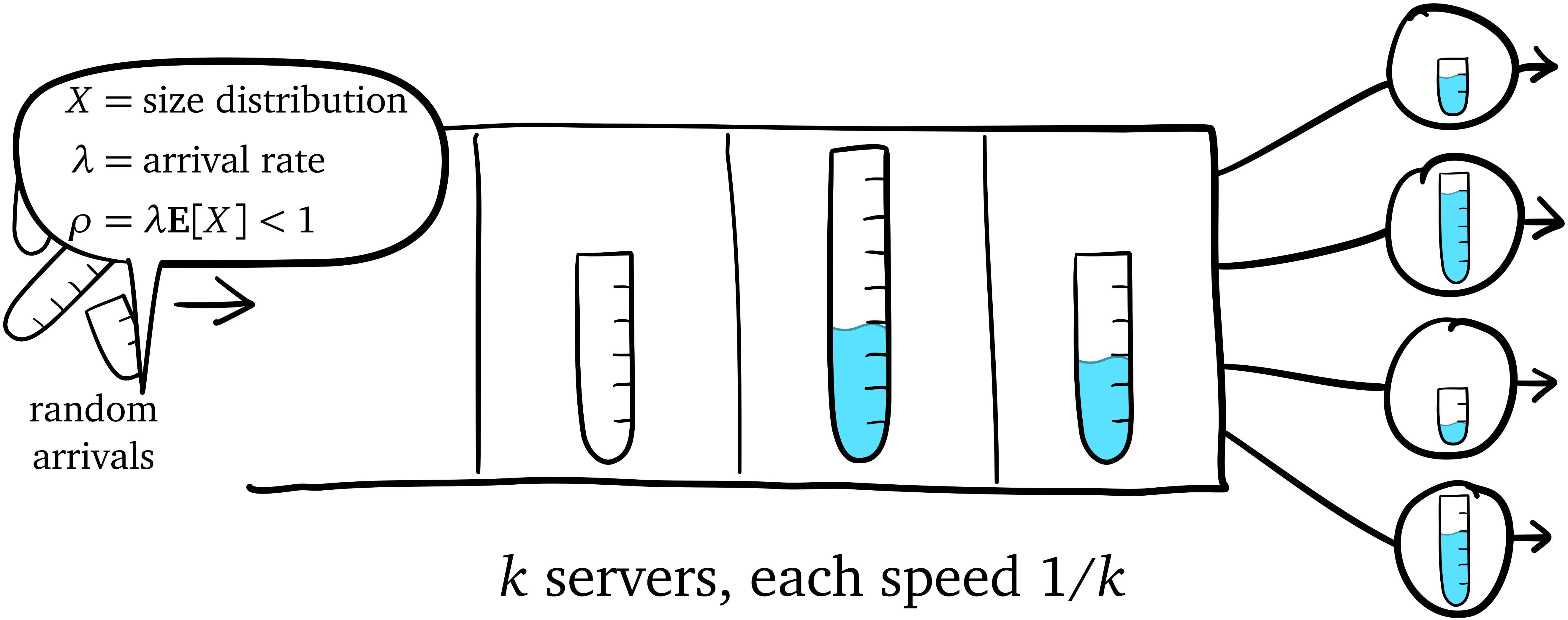


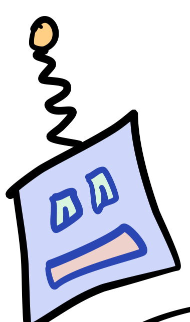
SRPT-k: always serve k jobs of *least remaining size*

SRPT-k minimizes $\mathbf{E}[T]$ in heavy traffic

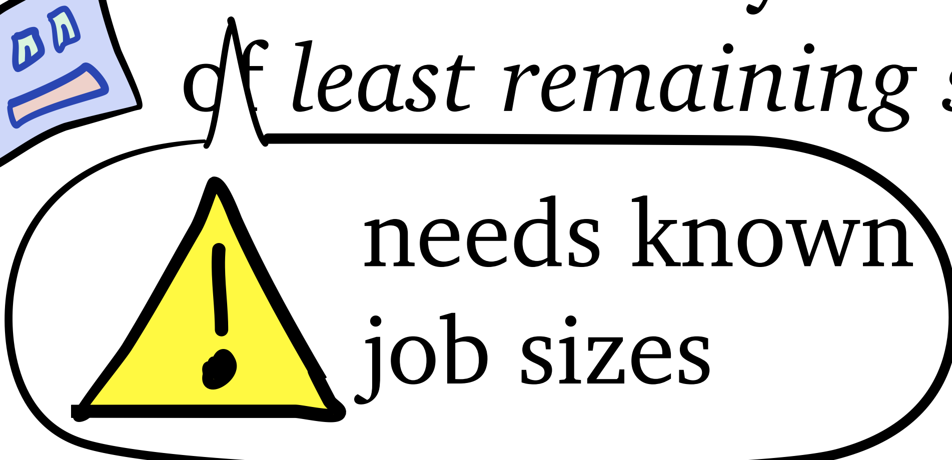
$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{SRPT-}k}]}{\mathbf{E}[T_{\text{SRPT-1}}]} = 1$$

M/G/k Queue





SRPT-k: always serve k jobs of *least remaining size*



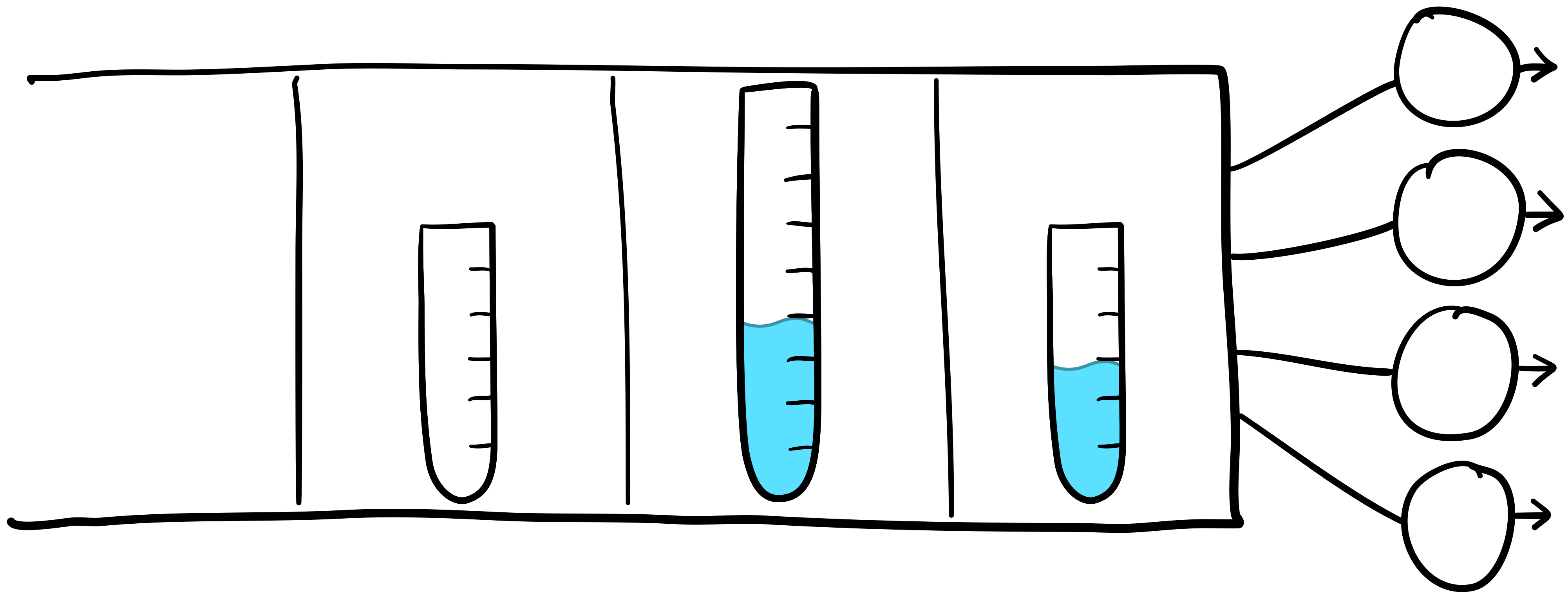
SRPT-k minimizes $\mathbf{E}[T]$ in heavy traffic

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{SRPT-}k}]}{\mathbf{E}[T_{\text{SRPT-1}}]} = 1$$

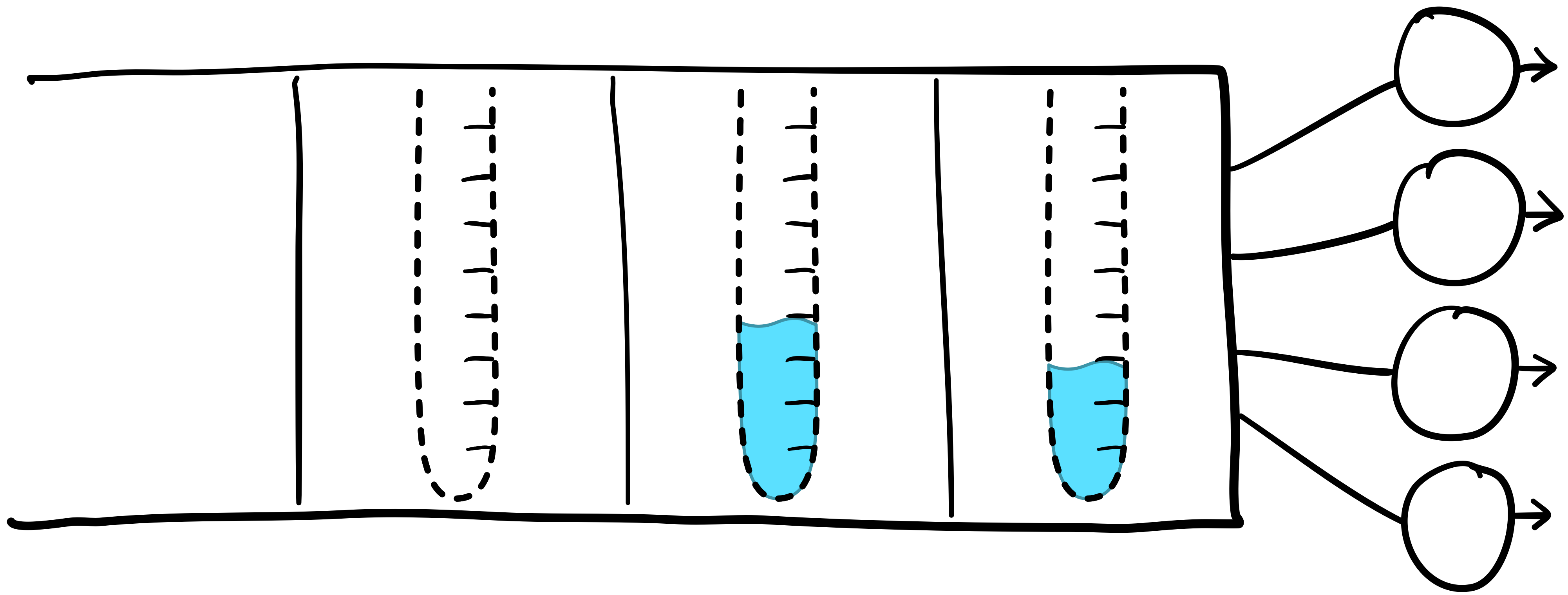
This talk:

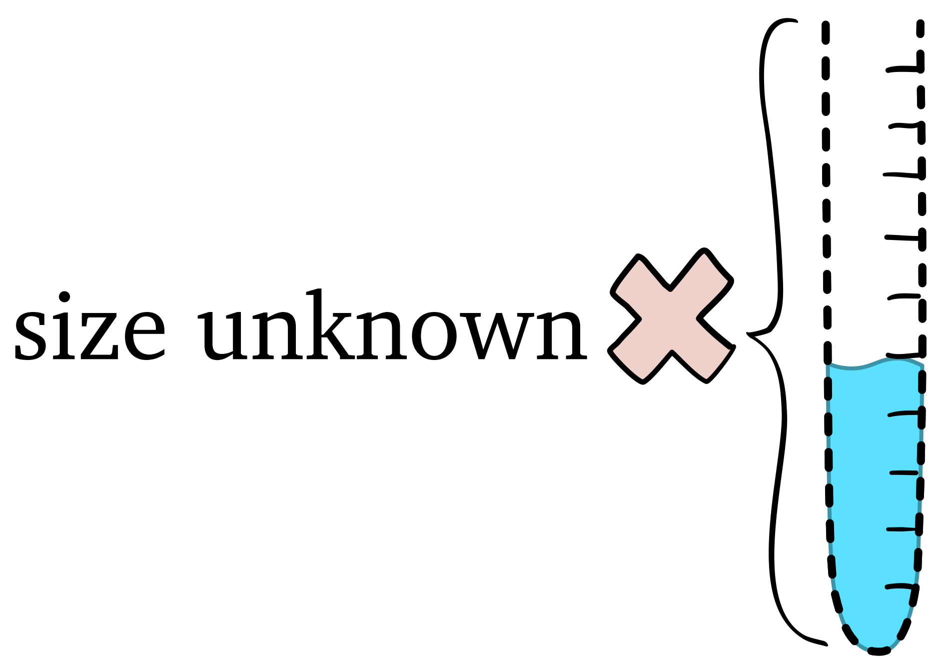
unknown job sizes

Unknown Job Sizes

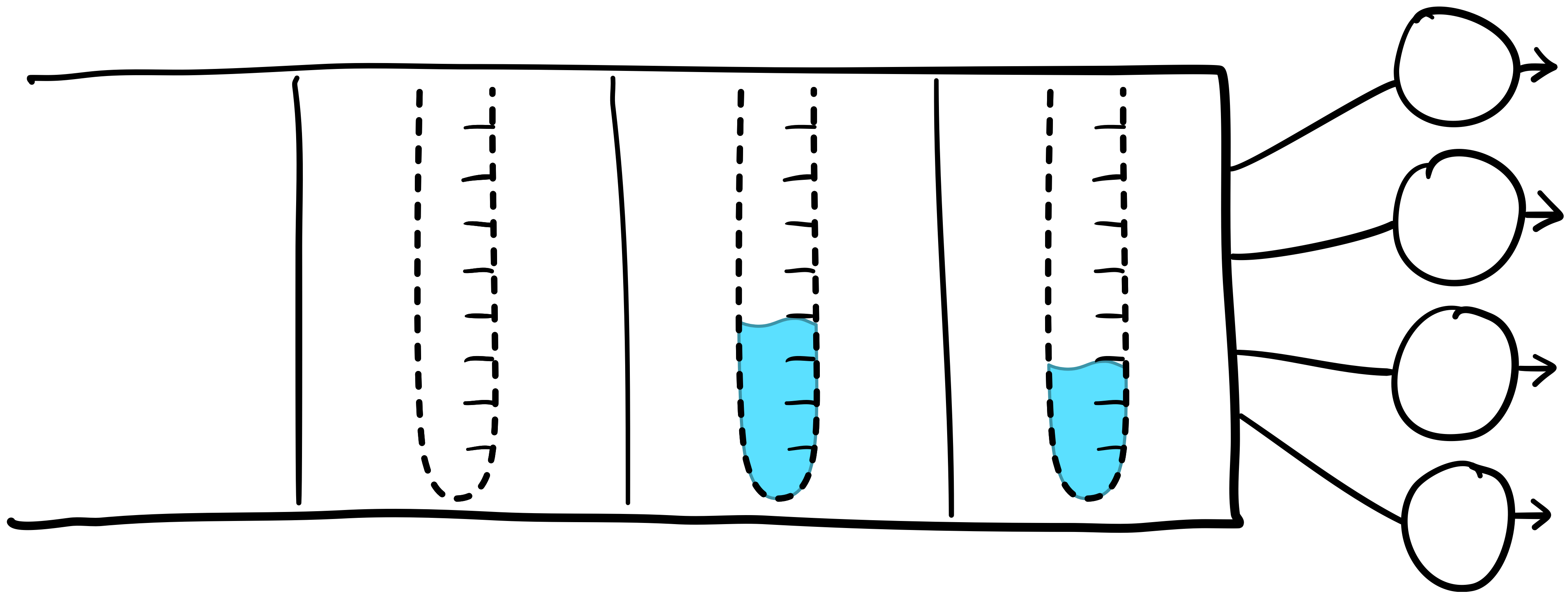


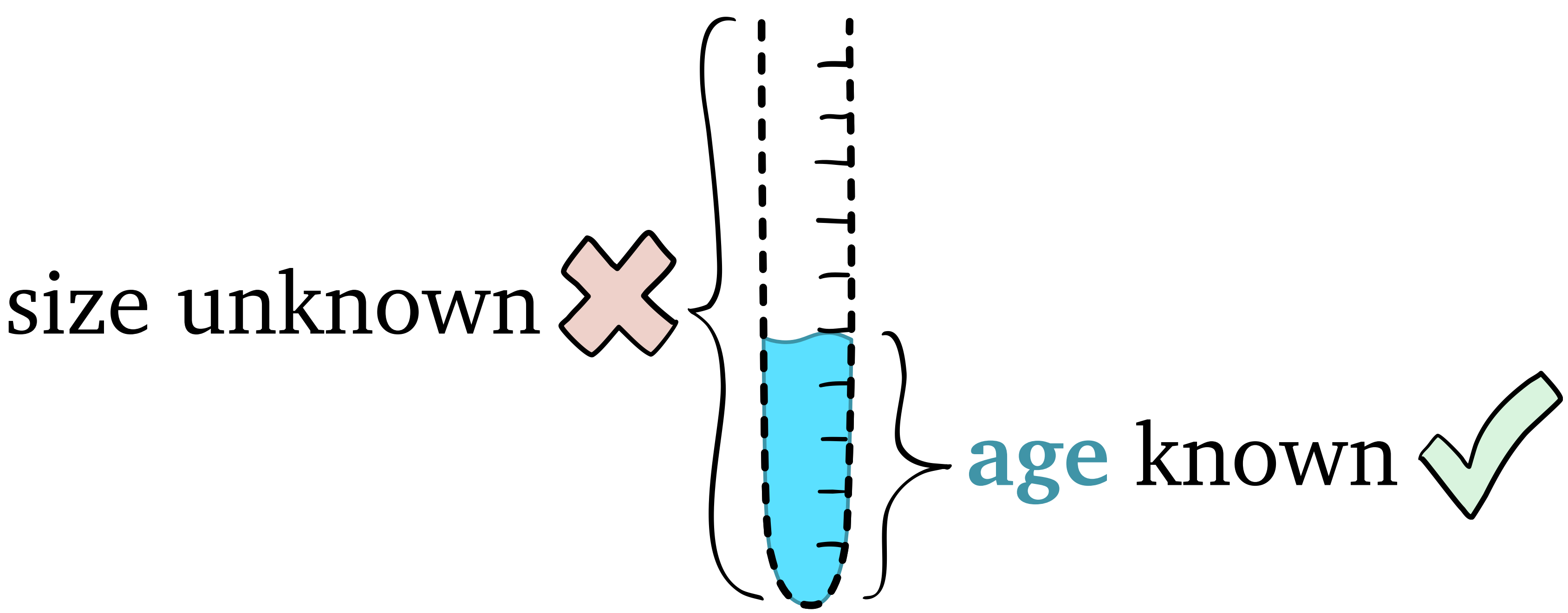
Unknown Job Sizes



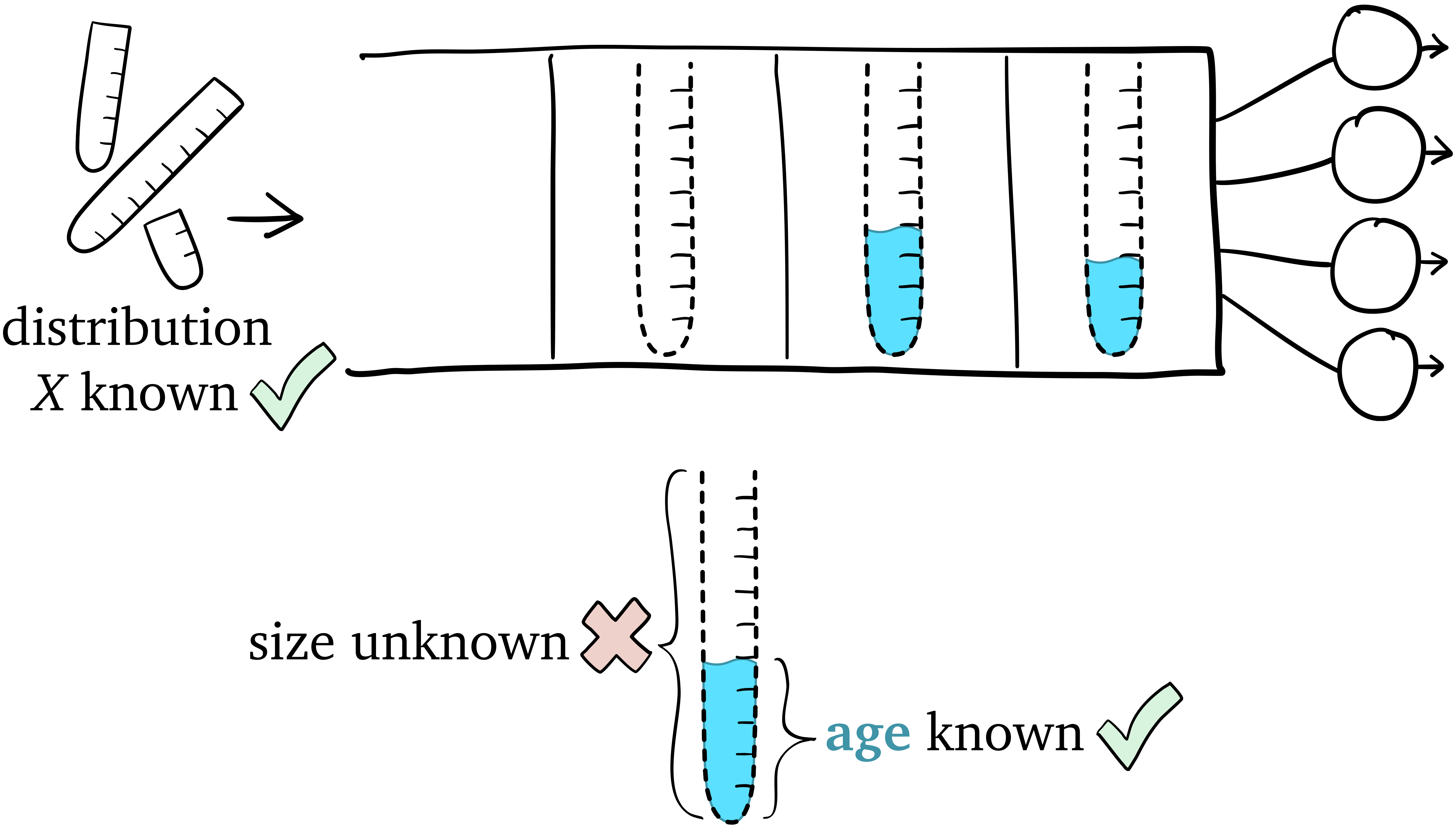


Unknown Job Sizes





Unknown Job Sizes



Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1		
M/G/k		

Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	
M/G/k	**SRPT-k**	

Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	
M/G/k	**SRPT-*k***	

in heavy traffic,
X "finite variance"

Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	**Gittins**
M/G/k	**SRPT-k**	

in heavy traffic,
X "finite variance"

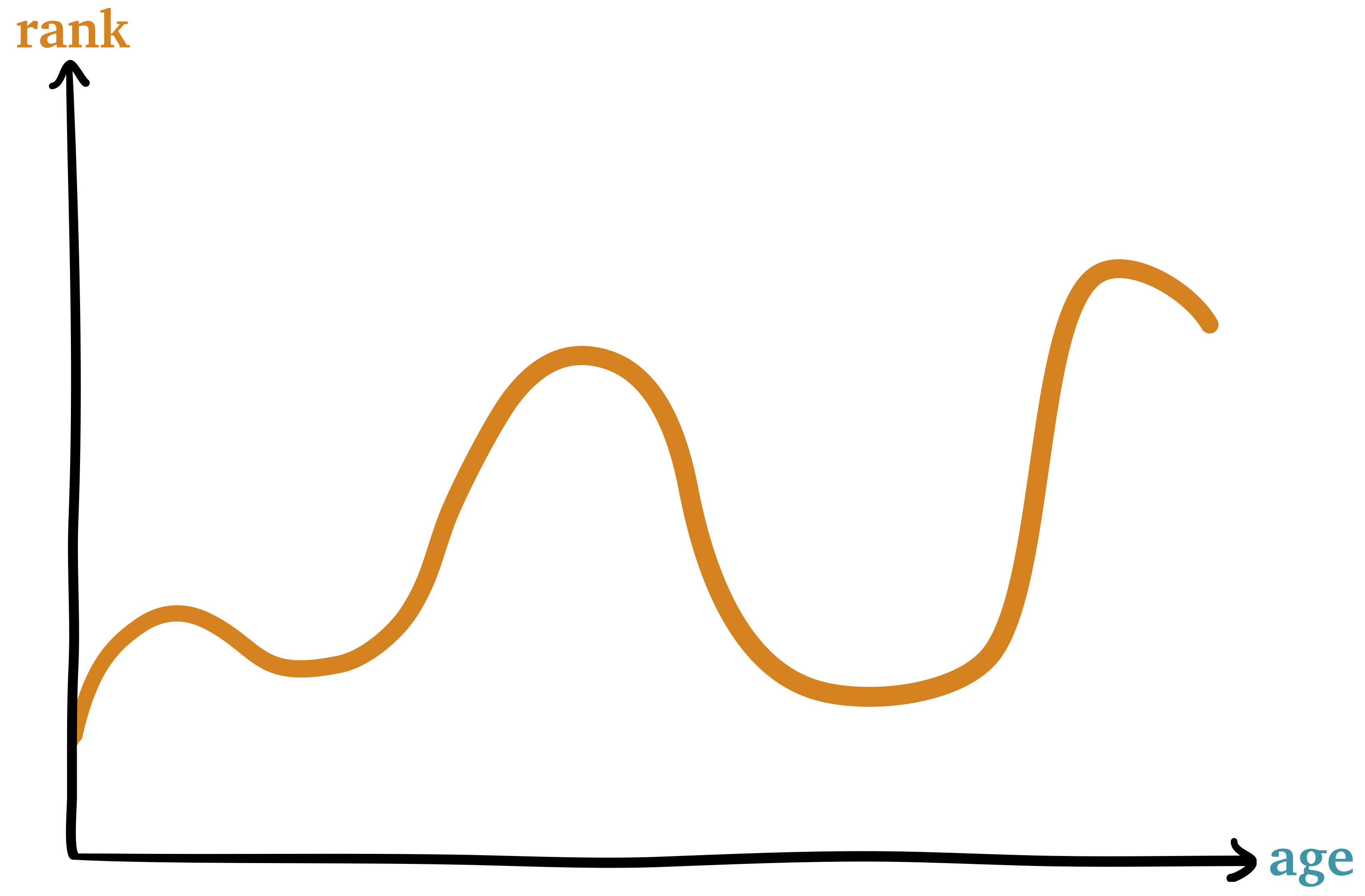
Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	**Gittins**
M/G/k	**SRPT-k**	**???**

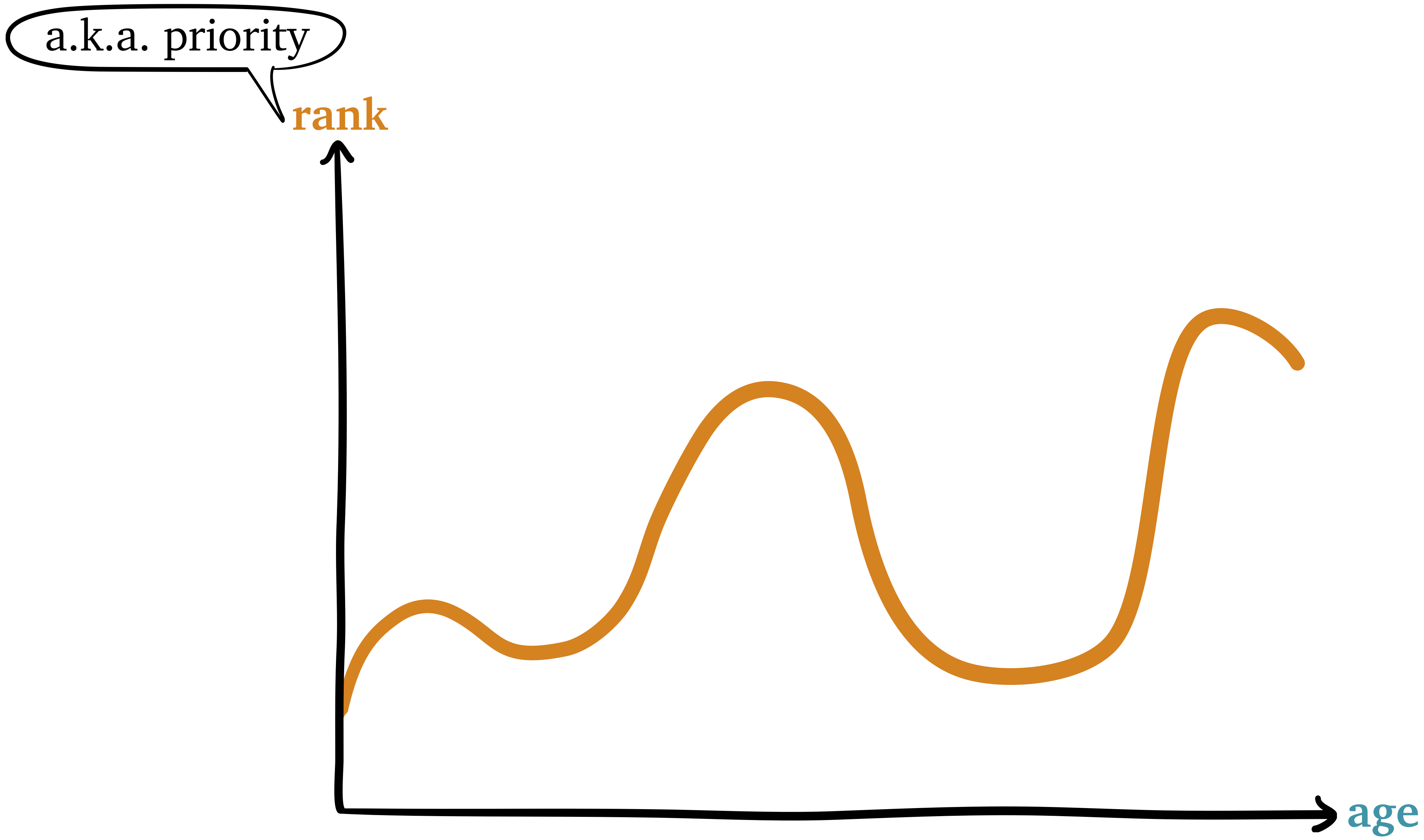
in heavy traffic,
X "finite variance"

What is **Gittins**?

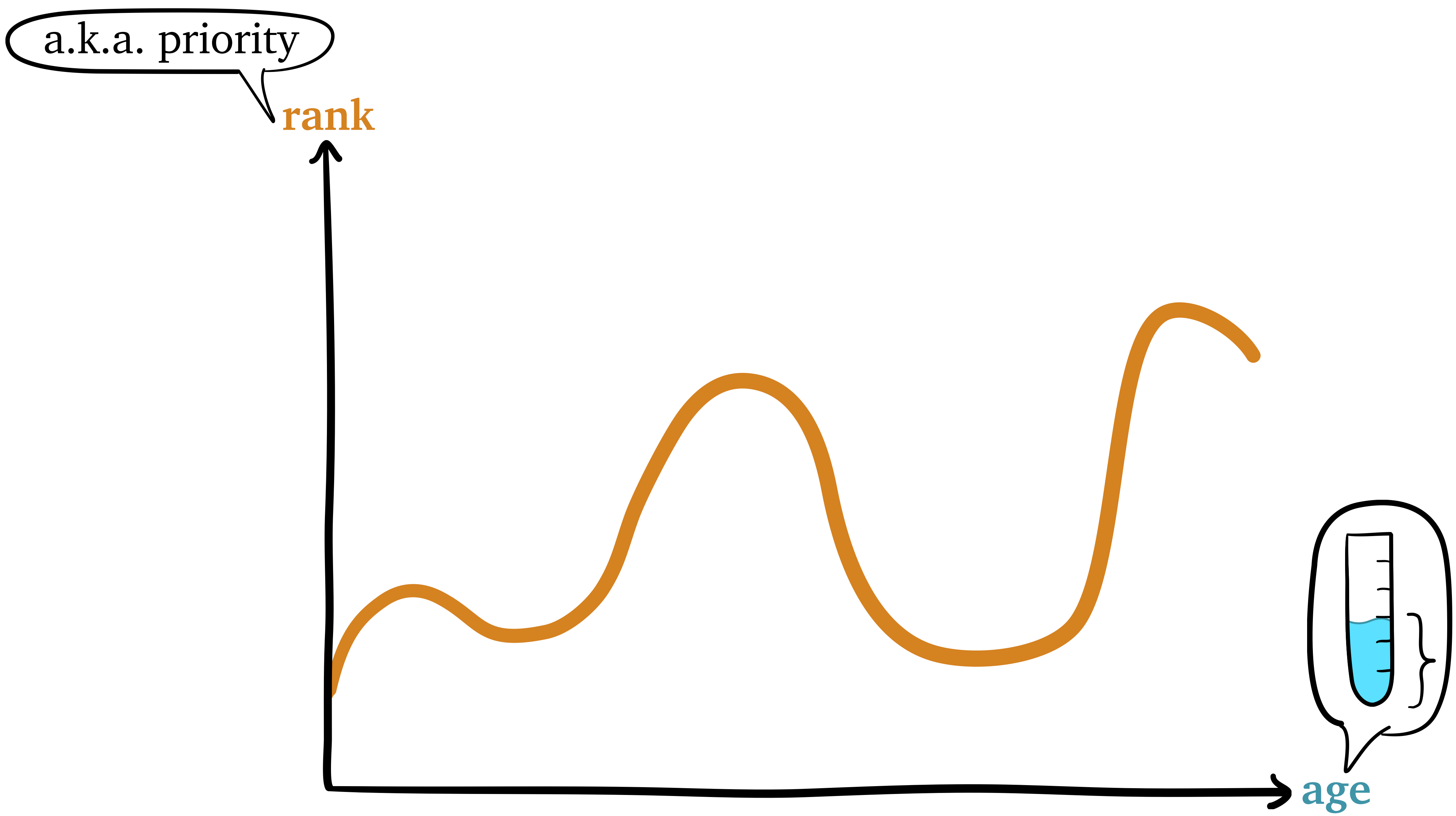
What is **Gittins**?



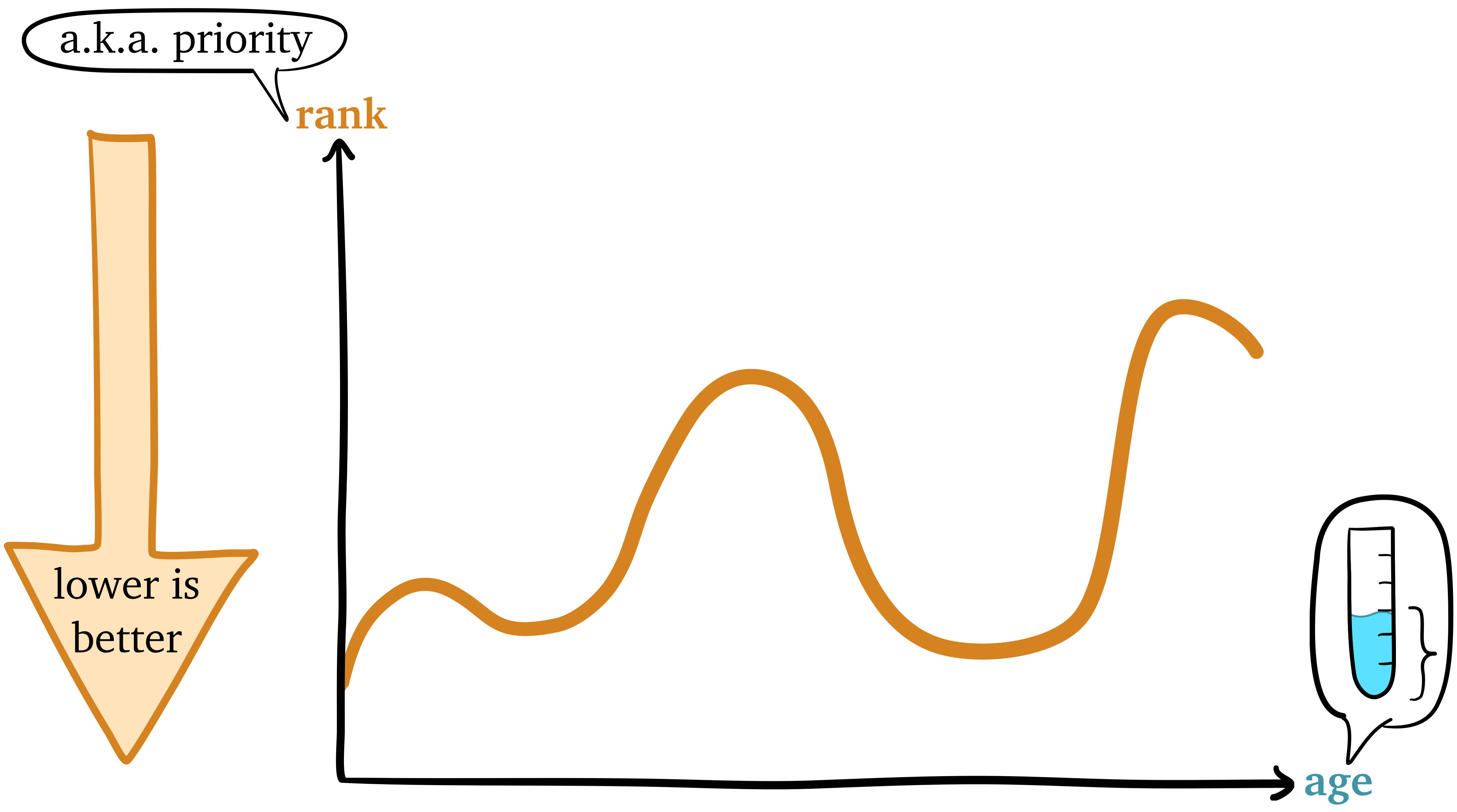
What is Gittins?



What is Gittins?

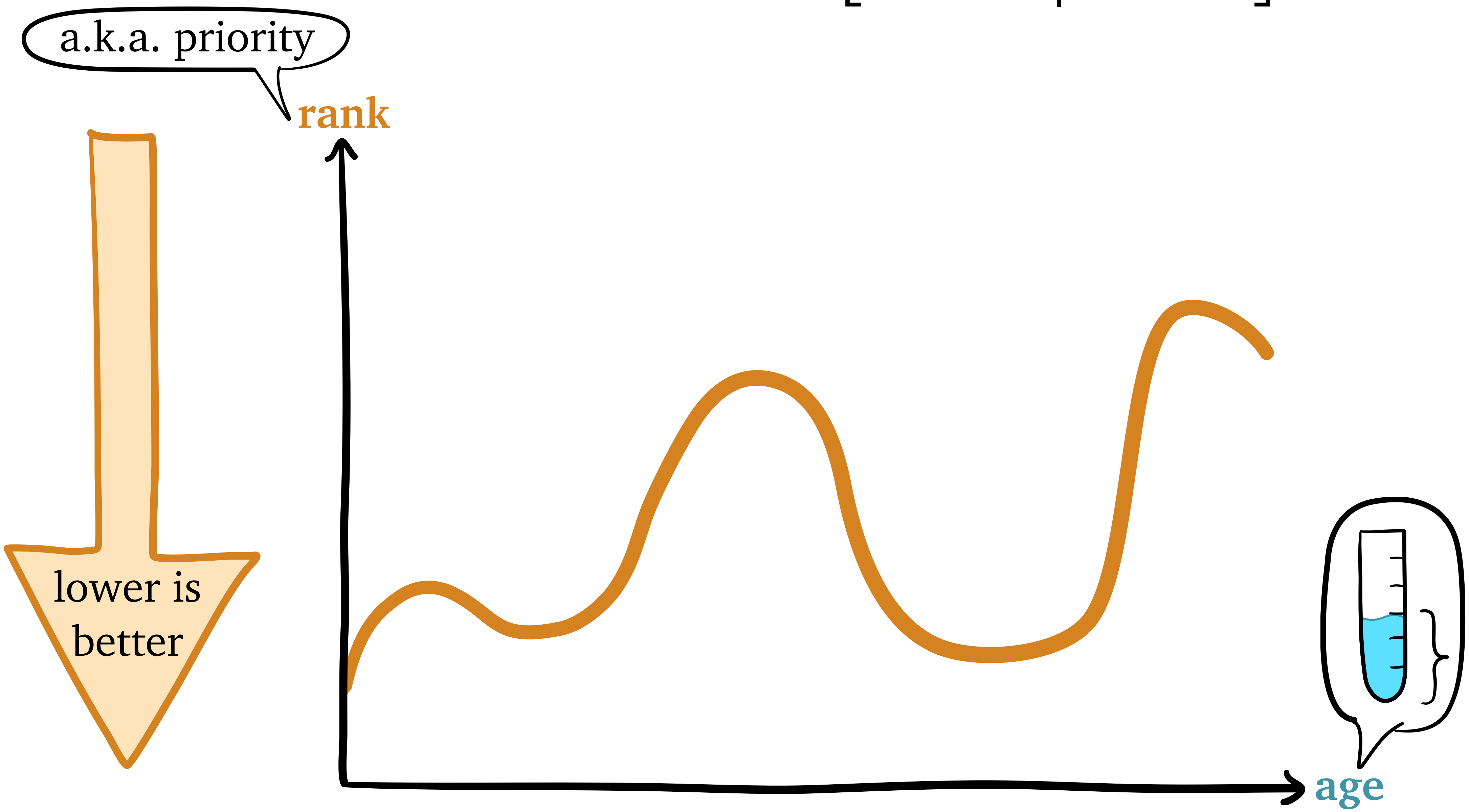


What is **Gittins**?



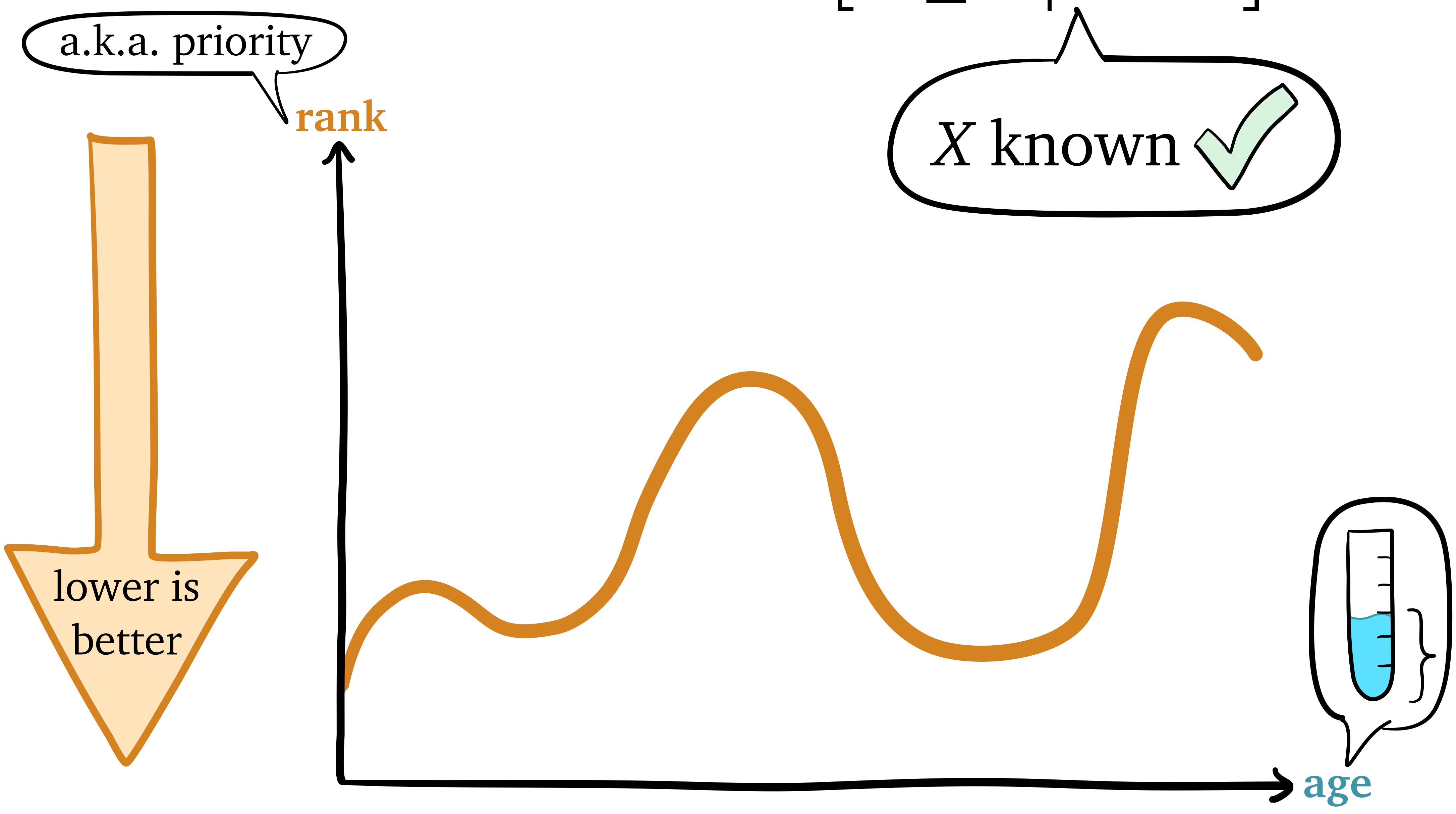
What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbf{E}[\min\{X, b\} \mid X > a]}{\mathbf{P}[X \le b \mid X > a]}$$



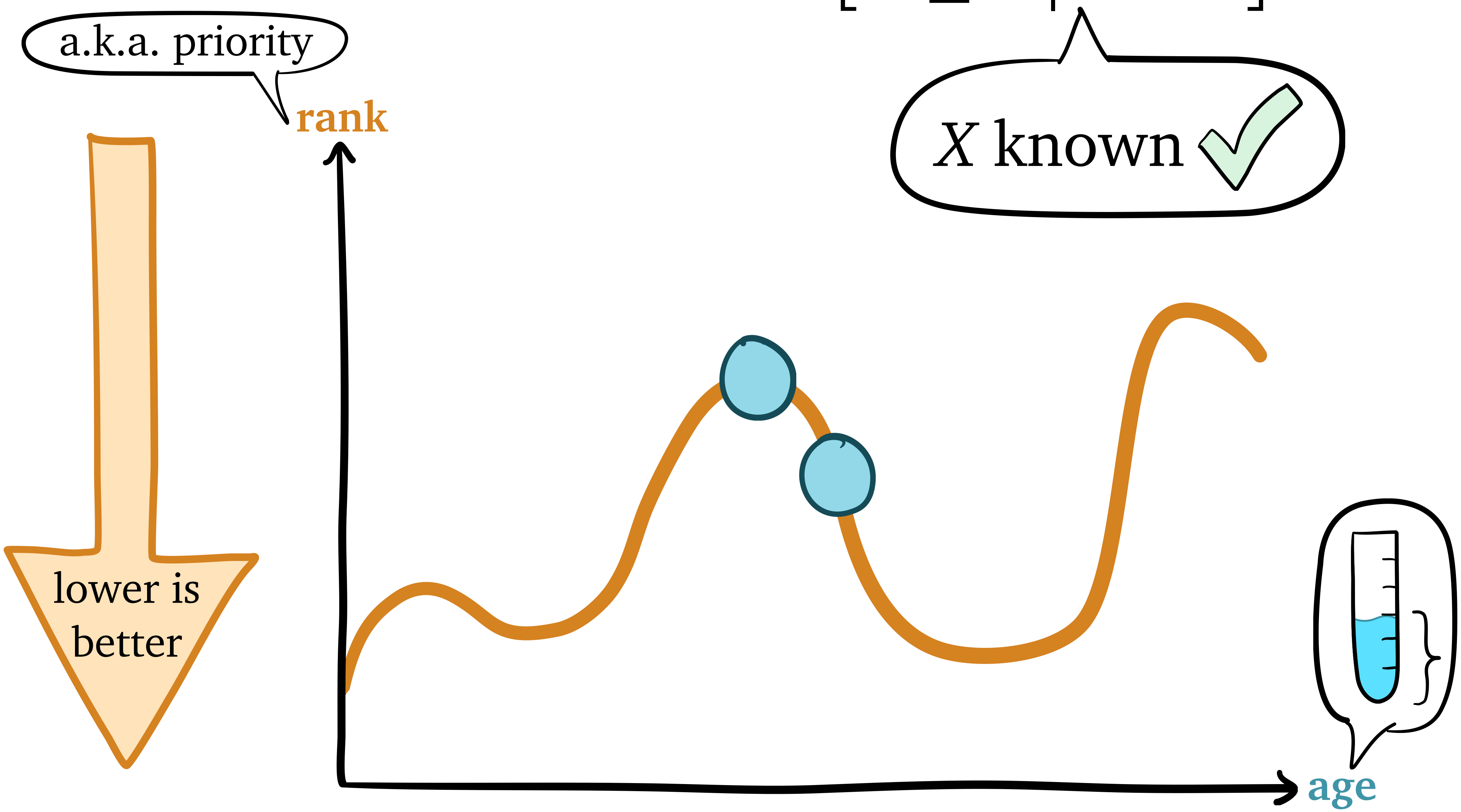
What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbf{E}[\min\{X, b\} \mid X > a]}{\mathbf{P}[X \le b \mid X > a]}$$



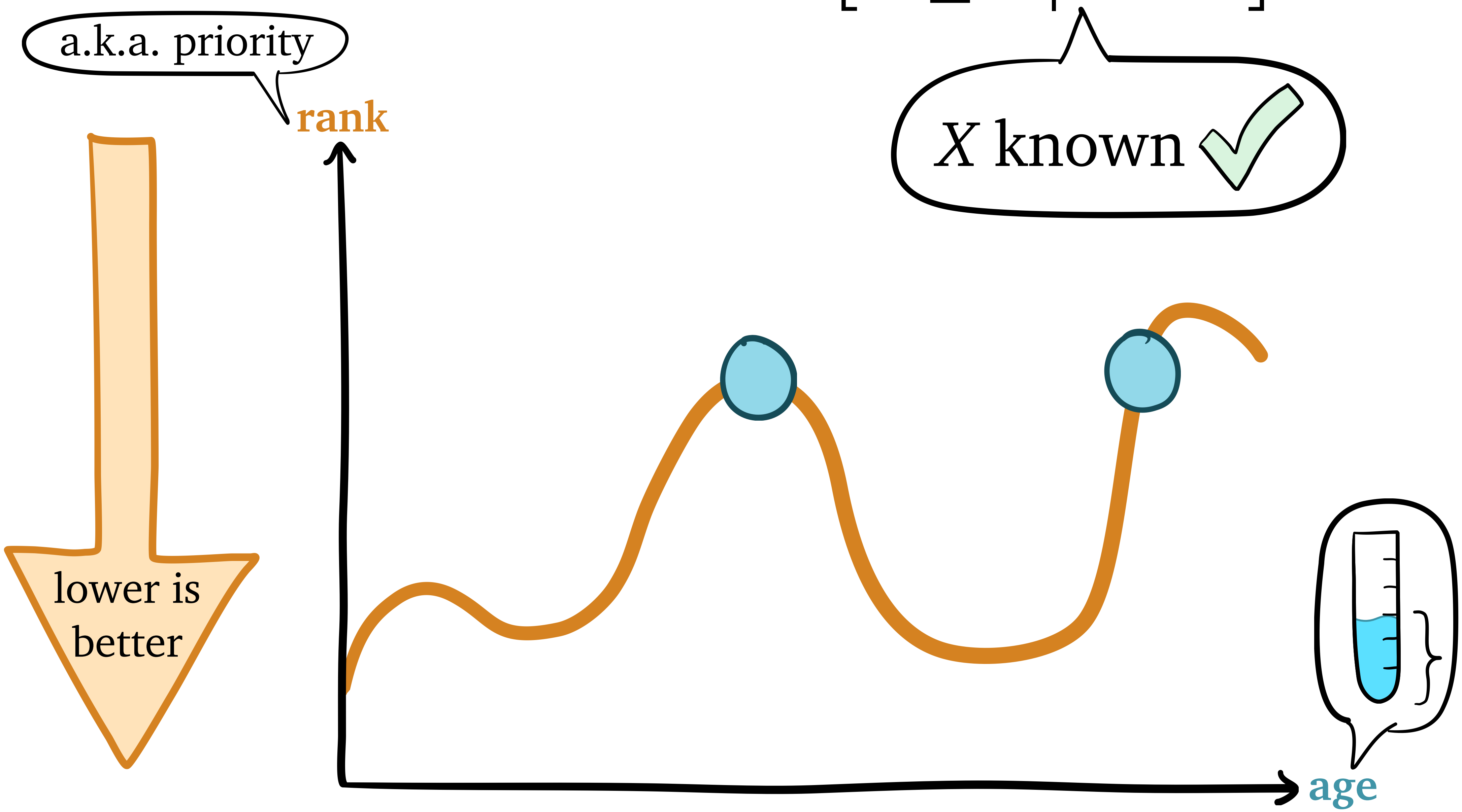
What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbf{E}[\min\{X, b\} \mid X > a]}{\mathbf{P}[X \leq b \mid X > a]}$$



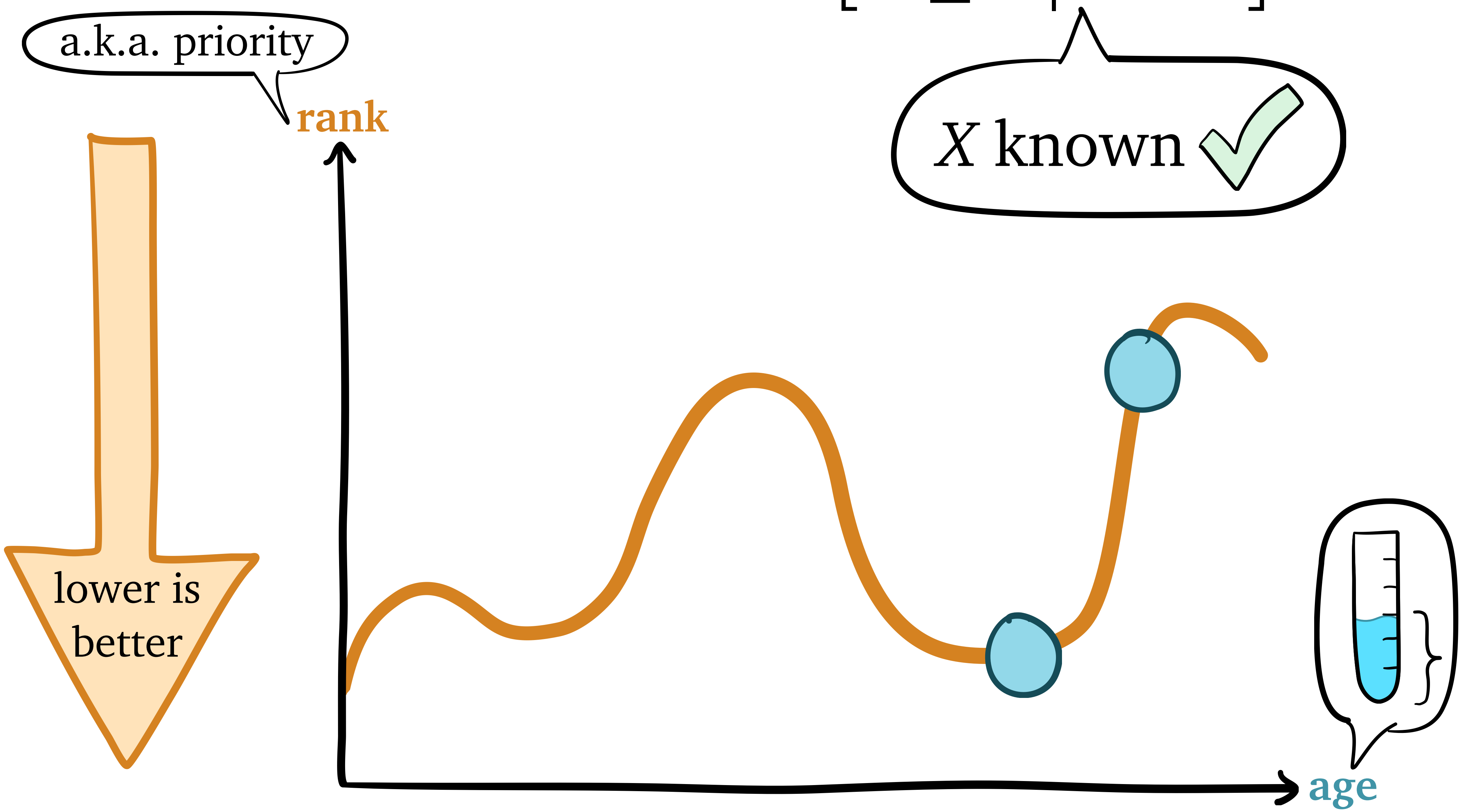
What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbf{E}[\min\{X, b\} \mid X > a]}{\mathbf{P}[X \le b \mid X > a]}$$



What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbf{E}[\min\{X, b\} \mid X > a]}{\mathbf{P}[X \leq b \mid X > a]}$$



Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	**Gittins**
M/G/k	**SRPT-*k***	**???**

in heavy traffic,
X “finite variance”

Minimizing $\mathbf{E}[T]$

	Known job sizes	Unknown job sizes
M/G/1	**SRPT**	**Gittins**
M/G/k	**SRPT-*k***	**???**

in heavy traffic,
X "finite variance"

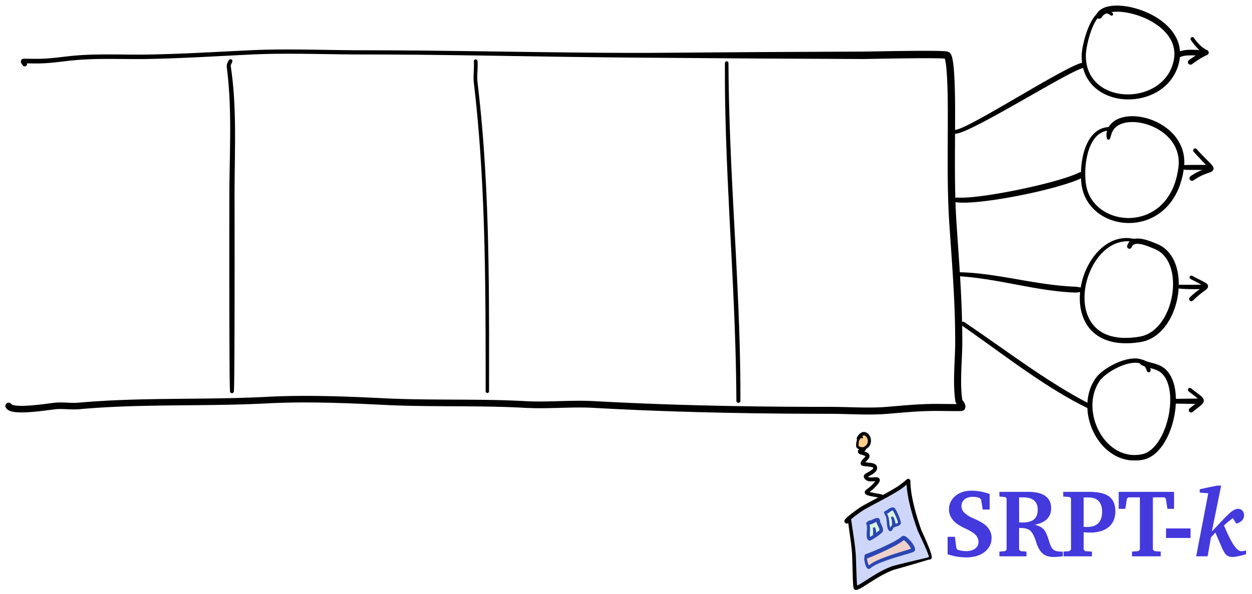
Does **Gittins-*k*** work?

Background:
SRPT-*k* optimality

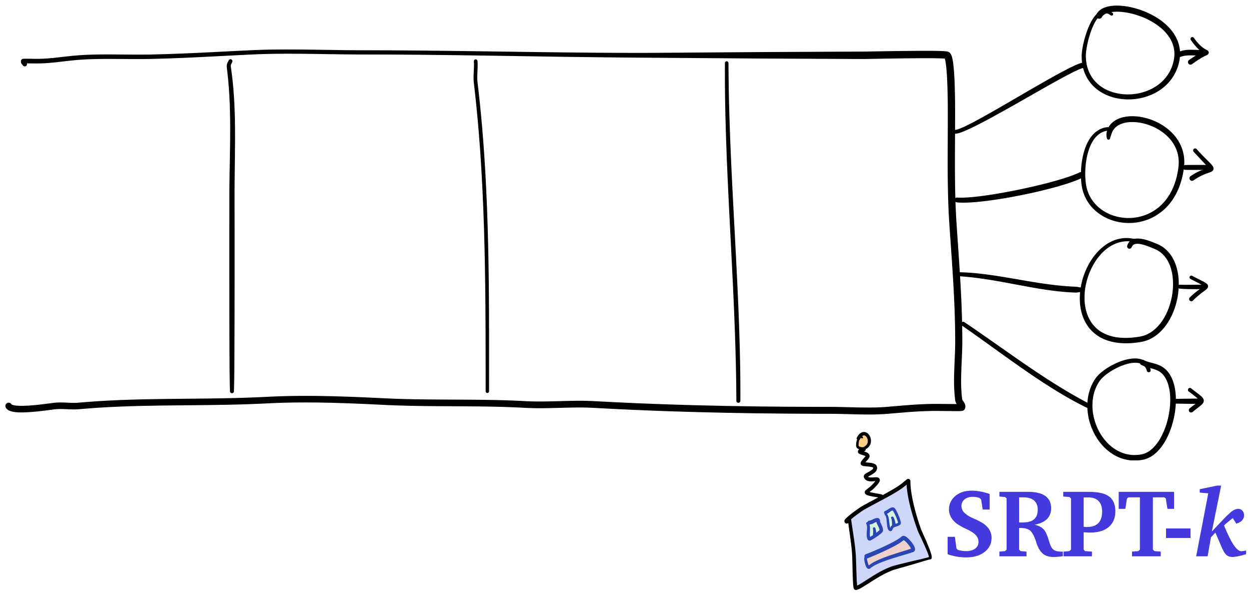
Background:
SRPT-*k* optimality

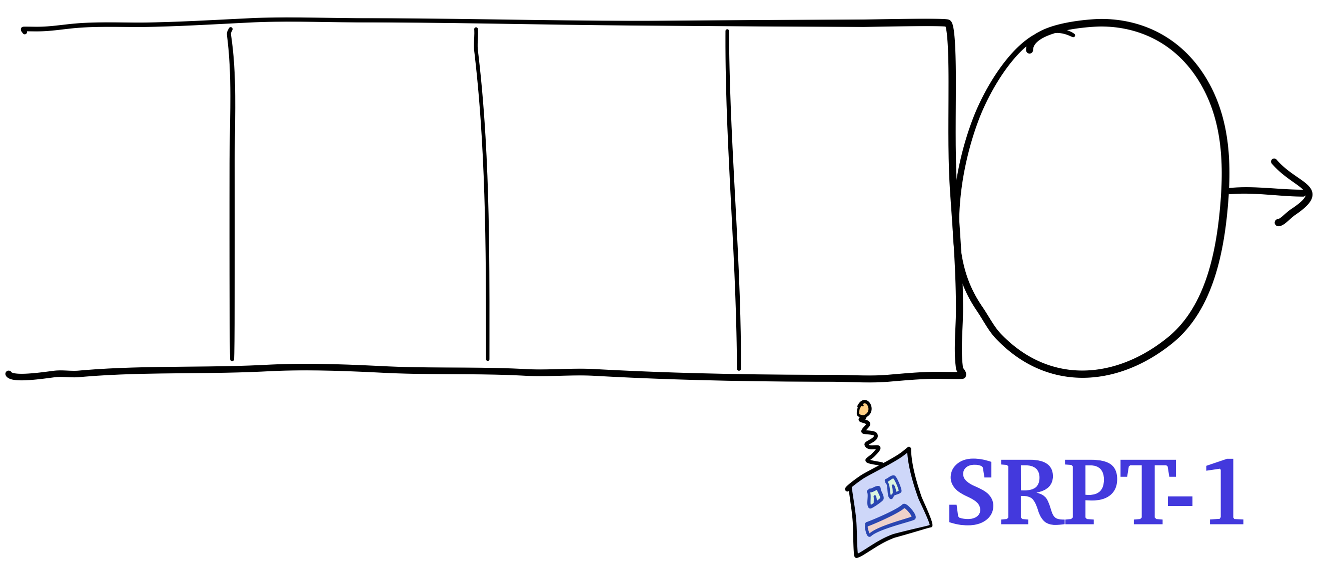
(Grosof, Scully, & Harchol-Balter, 2018)

SRPT-*k* Proof Sketch

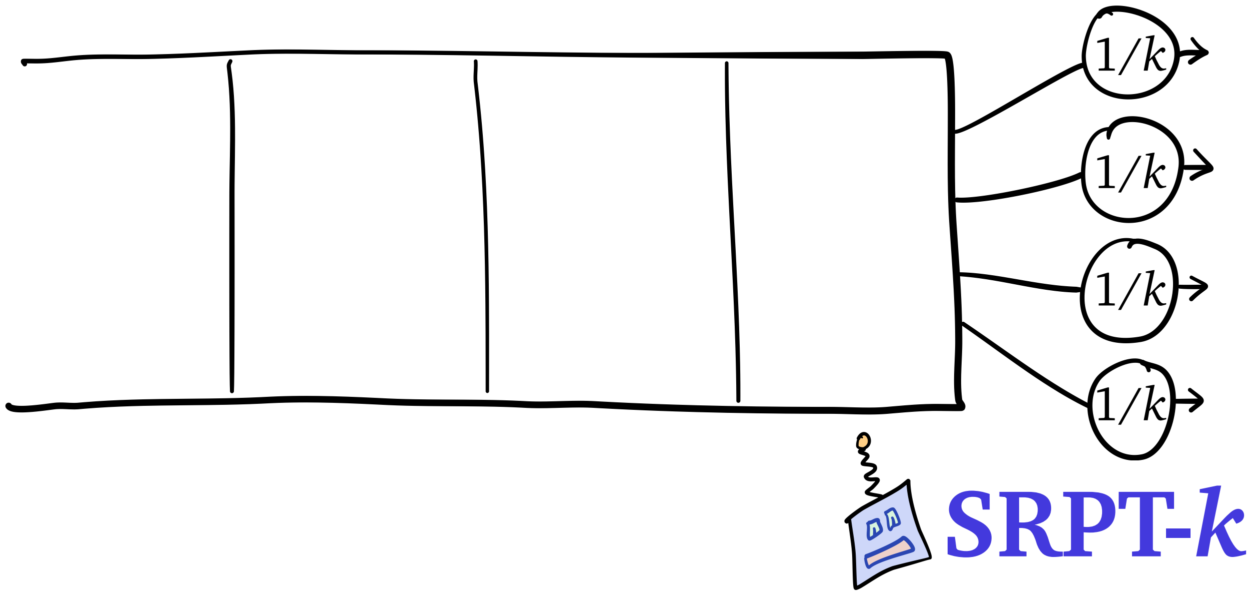


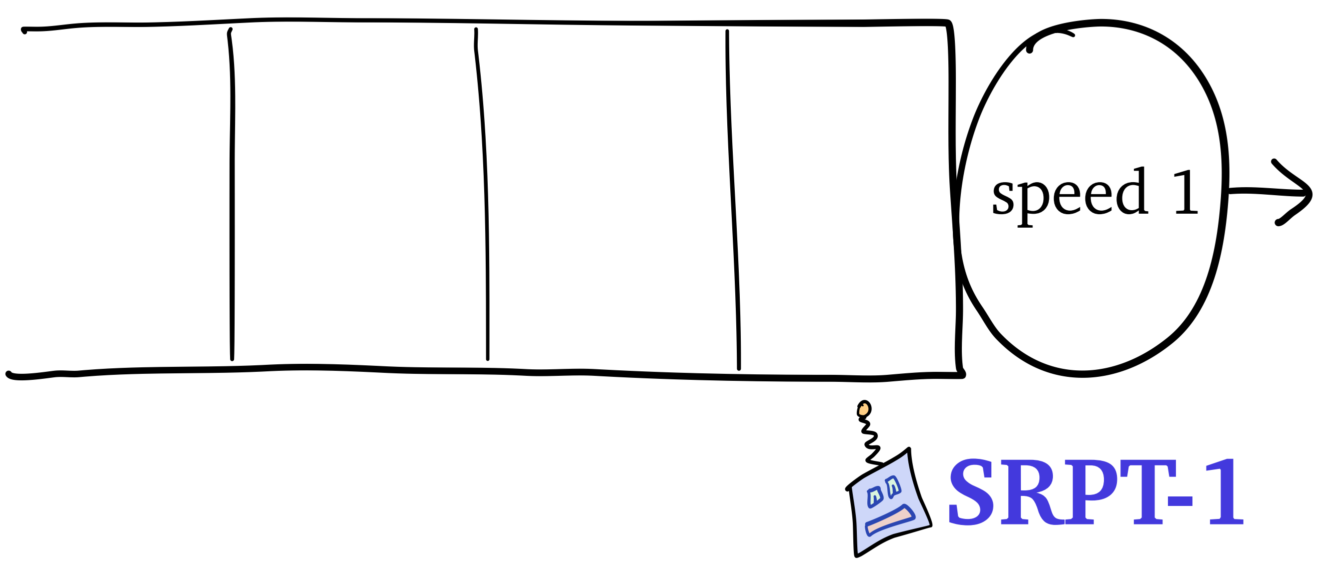
SRPT-*k* Proof Sketch



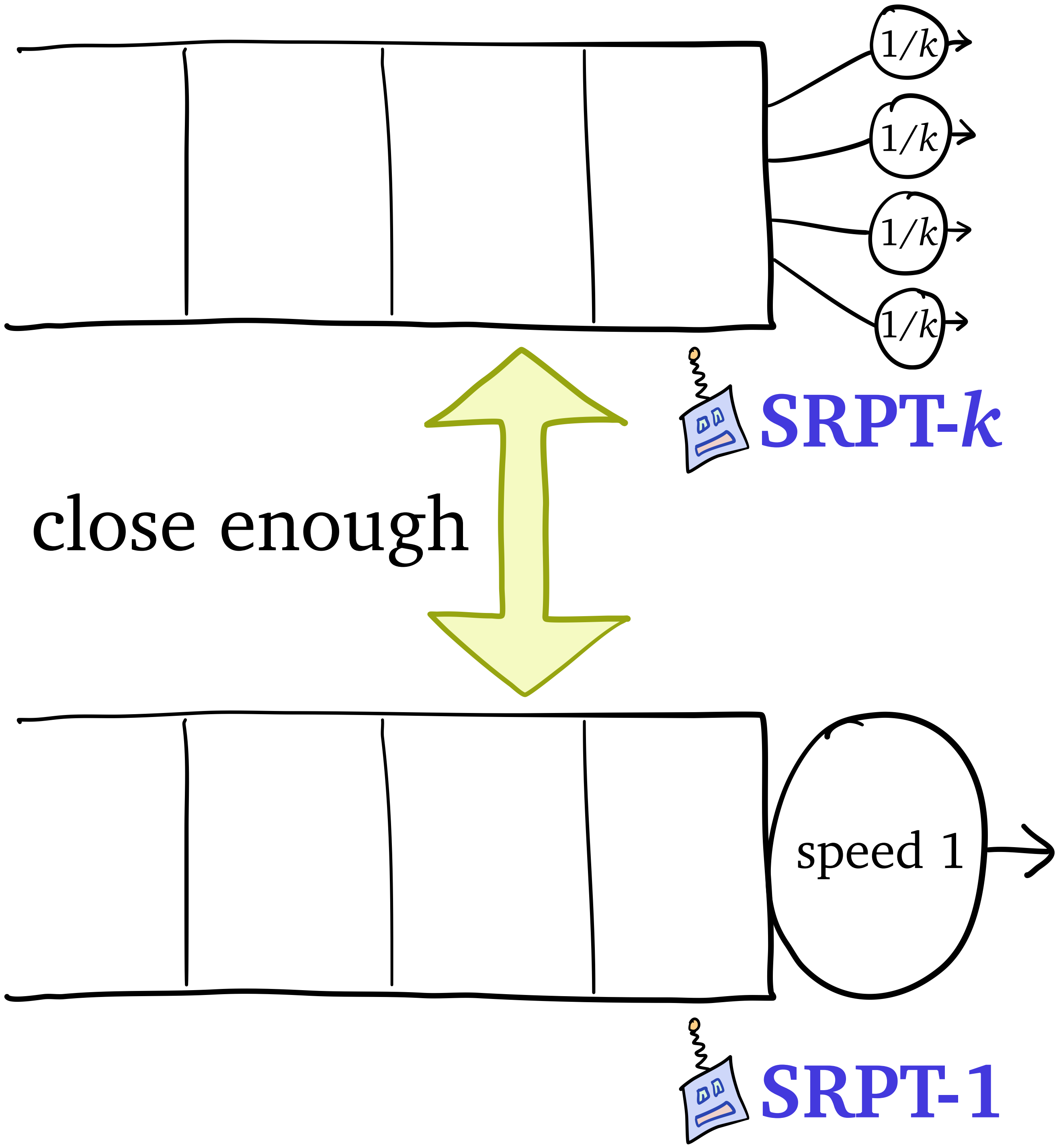


SRPT-k Proof Sketch

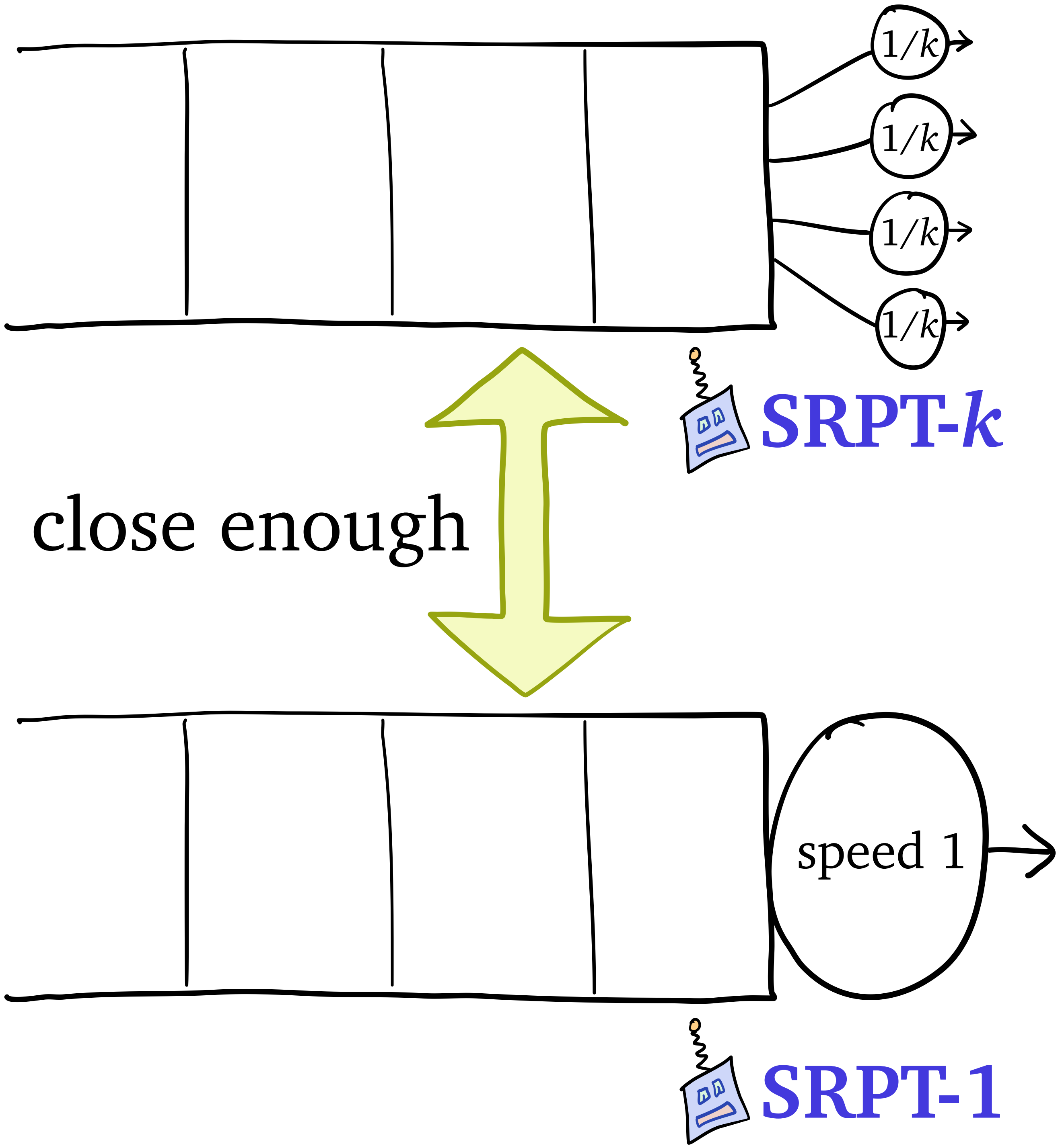




SRPT-*k* Proof Sketch



SRPT-*k* Proof Sketch

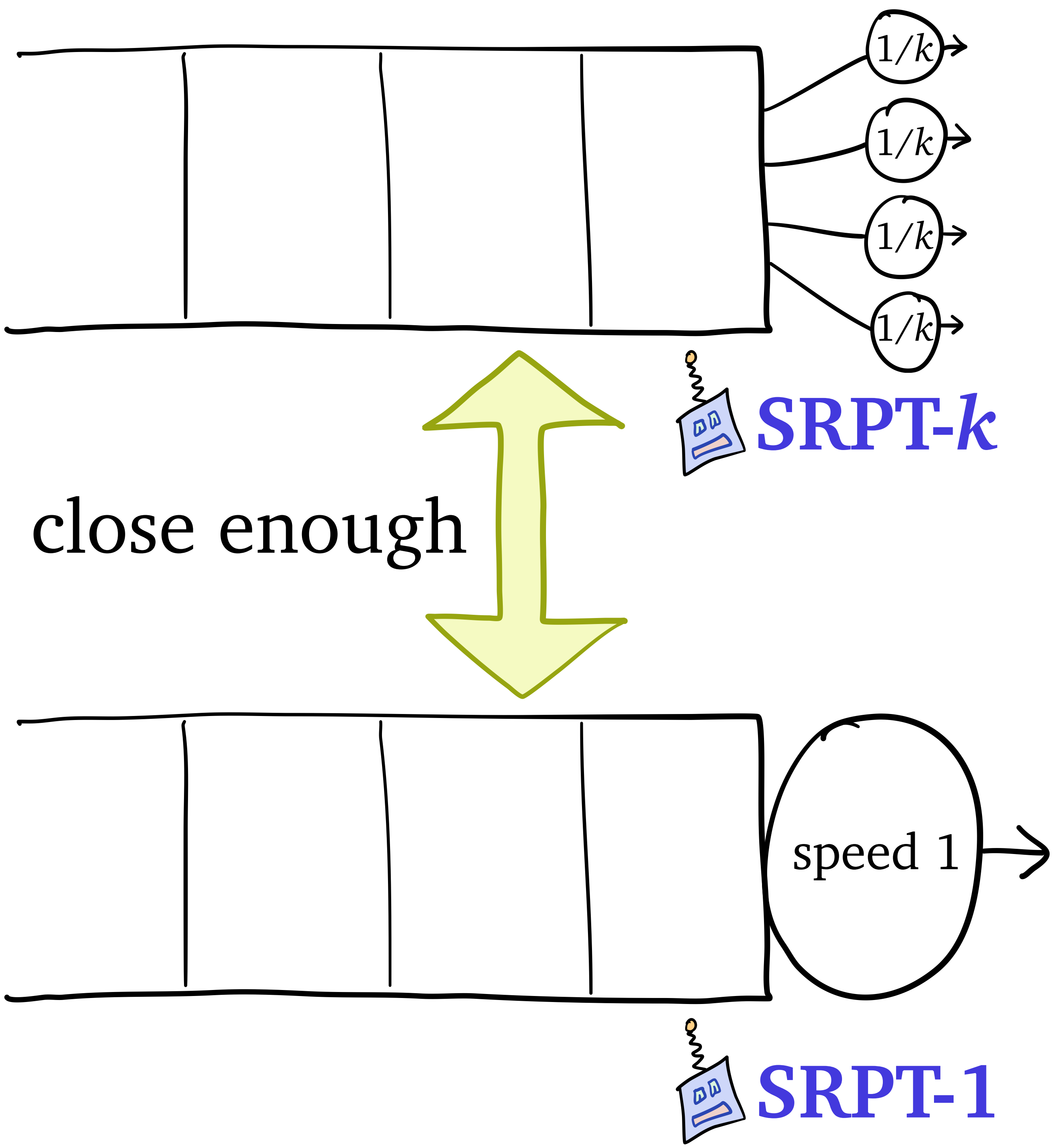


Step 1:

link **SRPT-*k*** to **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-}k}] \le \mathbf{E}[T_{\text{SRPT-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

SRPT-k Proof Sketch



Step 1:

link **SRPT-k** to **SRPT-1**

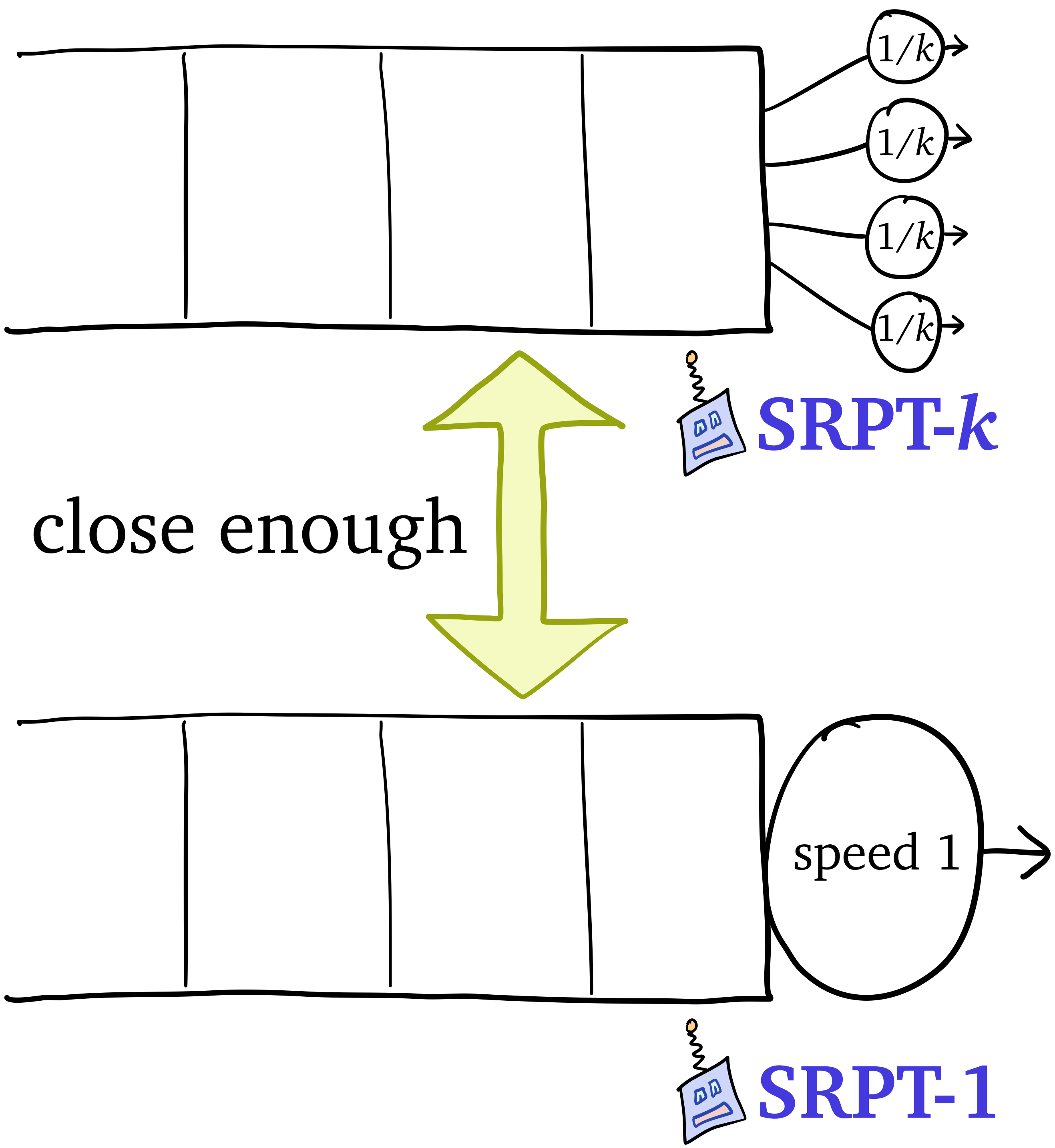
$$\mathbf{E}[T_{\text{SRPT-}k}] \leq \mathbf{E}[T_{\text{SRPT-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

Step 2:

analyze heavy-traffic **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

SRPT-k Proof Sketch



Step 1:

link **SRPT-k** to **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-}k}] \leq \mathbf{E}[T_{\text{SRPT-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

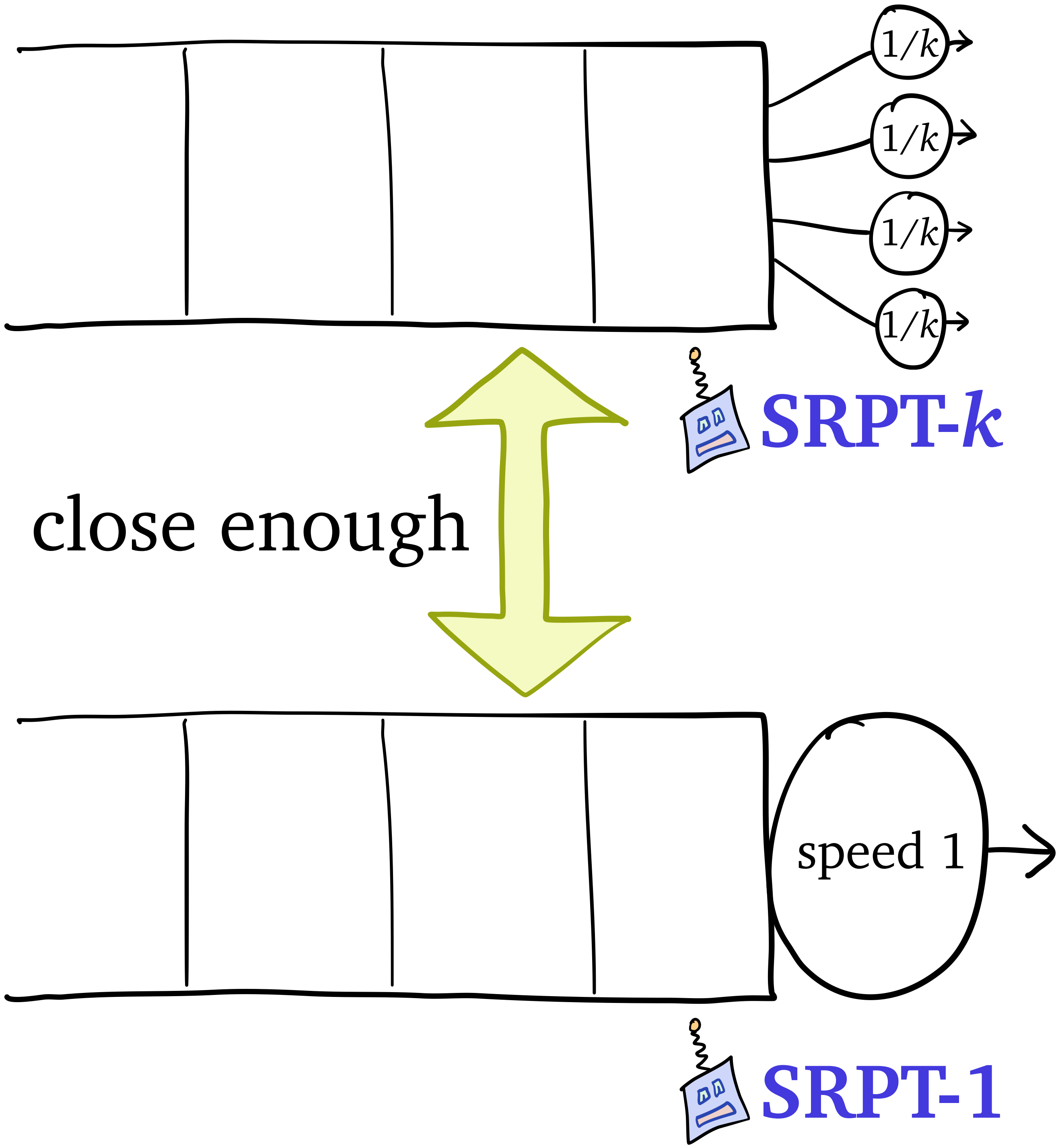
Step 2:

analyze heavy-traffic **SRPT-1**

(Lin, Wierman, & Zwart, 2011)

$$\mathbf{E}[T_{\text{SRPT-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

SRPT-*k* Proof Sketch



Step 1:

link **SRPT-*k*** to **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-}k}] \le \mathbf{E}[T_{\text{SRPT-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

(Lin, Wierman, & Zwart, 2011)

Step 2:

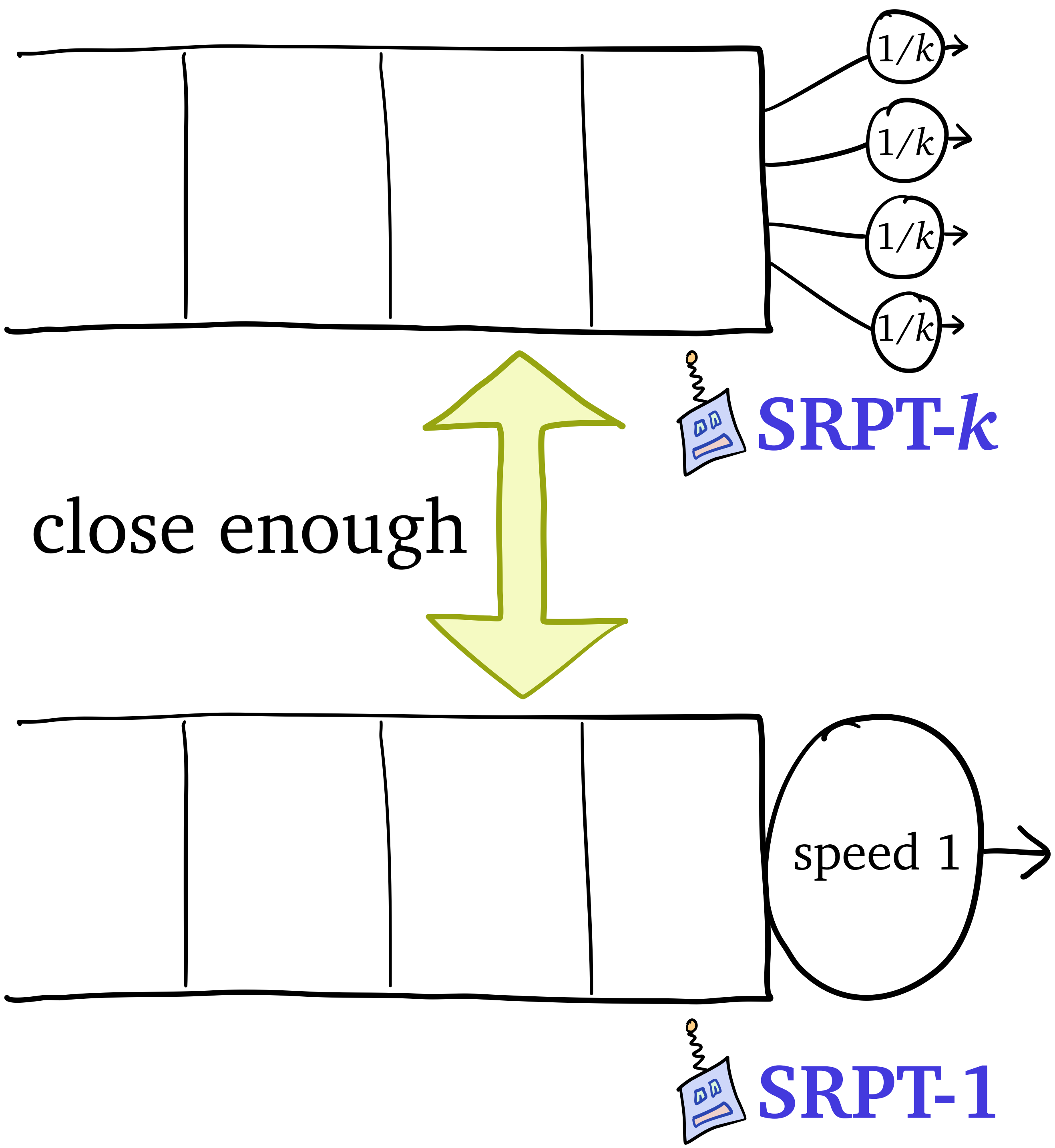
analyze heavy-traffic **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

Theorem: for any fixed k,

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{SRPT-}k}]}{\mathbf{E}[T_{\text{SRPT-1}}]} = 1$$

SRPT-k Proof Sketch



Step 1:

link **SRPT-k** to **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-}k}] \leq \mathbf{E}[T_{\text{SRPT-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

(Lin, Wierman, & Zwart, 2011)

Step 2:

analyze heavy-traffic **SRPT-1**

$$\mathbf{E}[T_{\text{SRPT-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

needs X to be "finite variance"

Theorem: for any fixed k,

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{SRPT-}k}]}{\mathbf{E}[T_{\text{SRPT-1}}]} = 1$$

Step 1: link **Gittins-*k*** to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

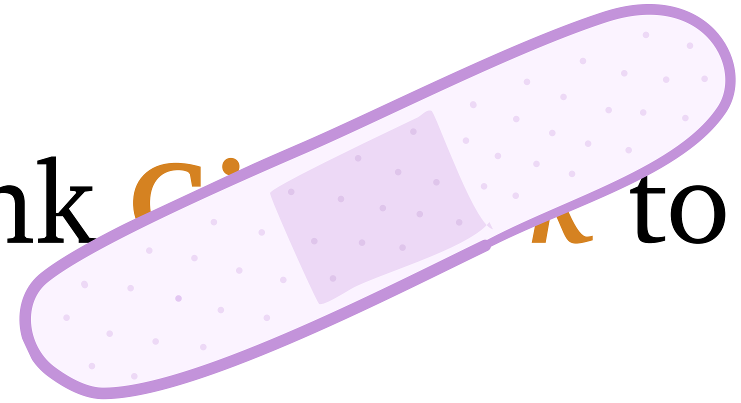
Step 1: link **Gittins-*k*** to **Gittins-1**

✓ **Step 2:** analyze heavy-traffic **Gittins-1**

Step 1: link **Gittins-*k*** to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

Step 1: link **Gi[illegible]k** to **Gittins-1**



Step 2: analyze heavy-traffic **Gittins-1**

Step 1: link M-Gittins-k to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

new policy!

M-Gittins-k

Step 1: link G…k to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

new policy!

Step 1: link **M-Gittins-k** to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

new policy!

Step 1: link M-Gittins-k to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

Theorem: for any fixed k,

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1$$

new policy!

Step 1: link **M-Gittins-*k*** to **Gittins-1**

- **Gittins-*k*** vs. **Gittins-1**
- What **M-Gittins** is
- **M-Gittins-*k*** vs. **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

Theorem: for any fixed k,

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1$$

new policy!

Step 1: link M-Gittins-k to **Gittins-1**

- **Gittins-k** vs. **Gittins-1**
- What **M-Gittins** is
- **M-Gittins-k** vs. **Gittins-1**

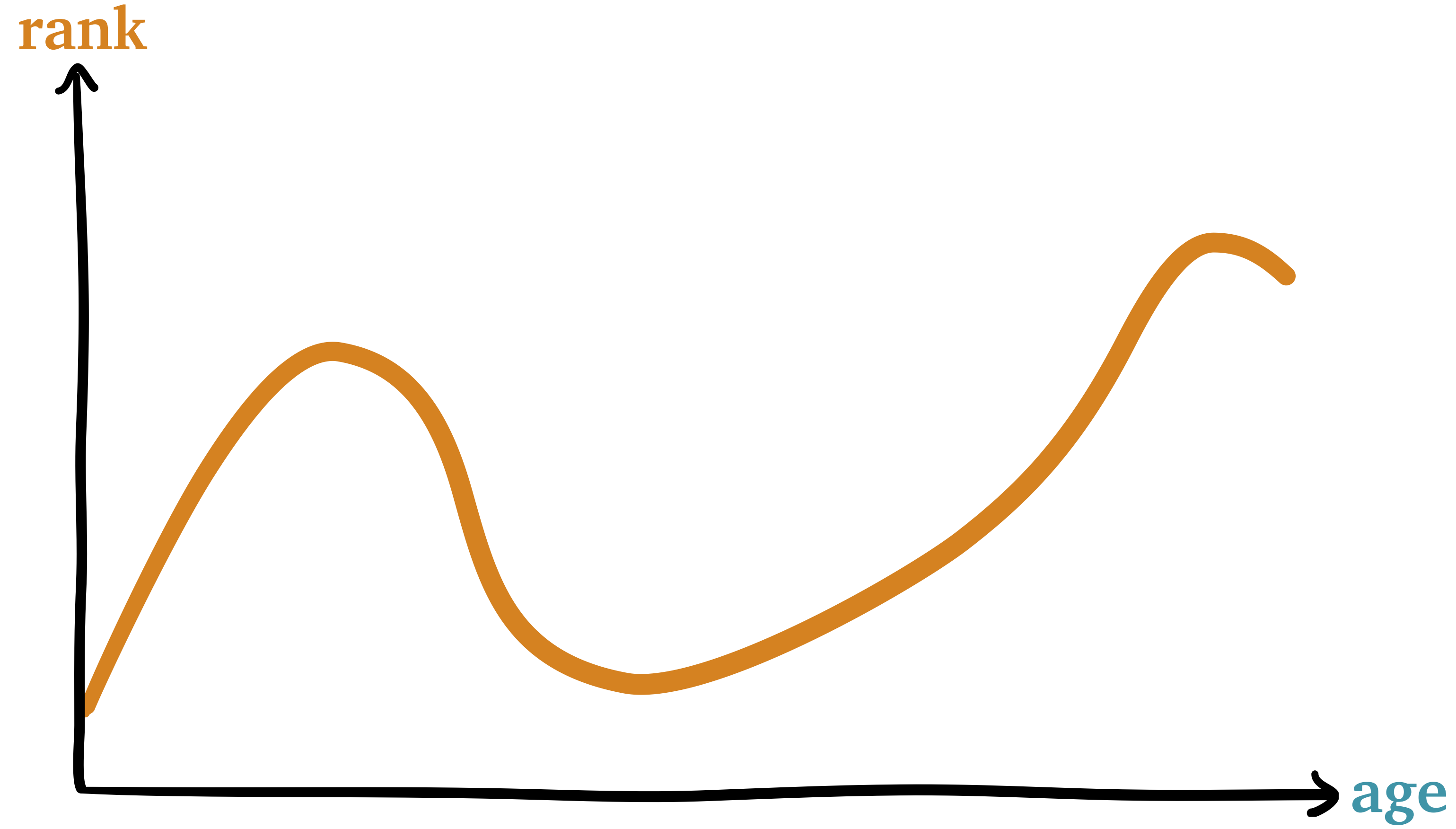
Step 2: analyze heavy-traffic **Gittins-1**

see paper

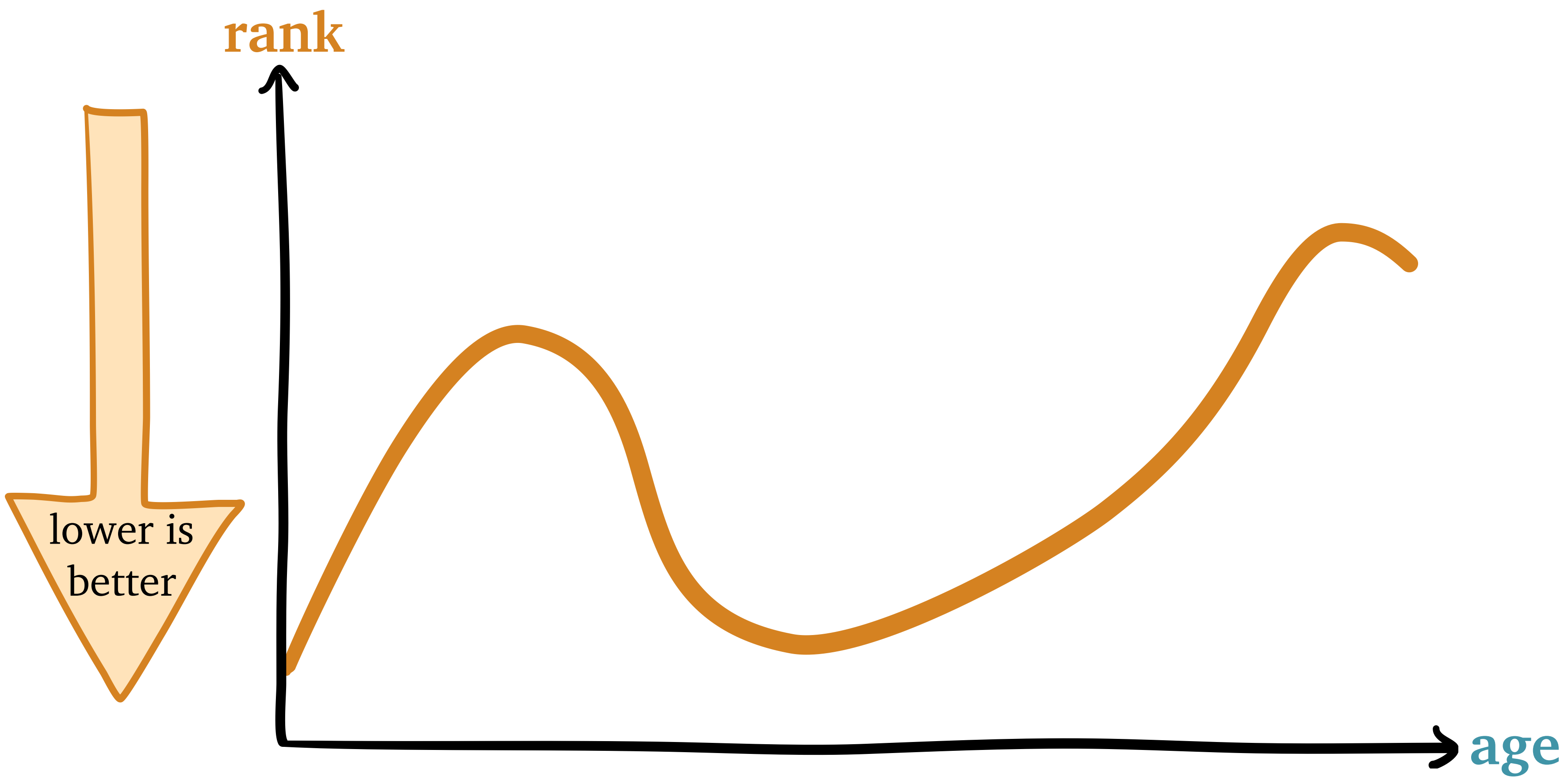
Theorem: for any fixed k,

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1$$

Analyzing **Gittins-1**

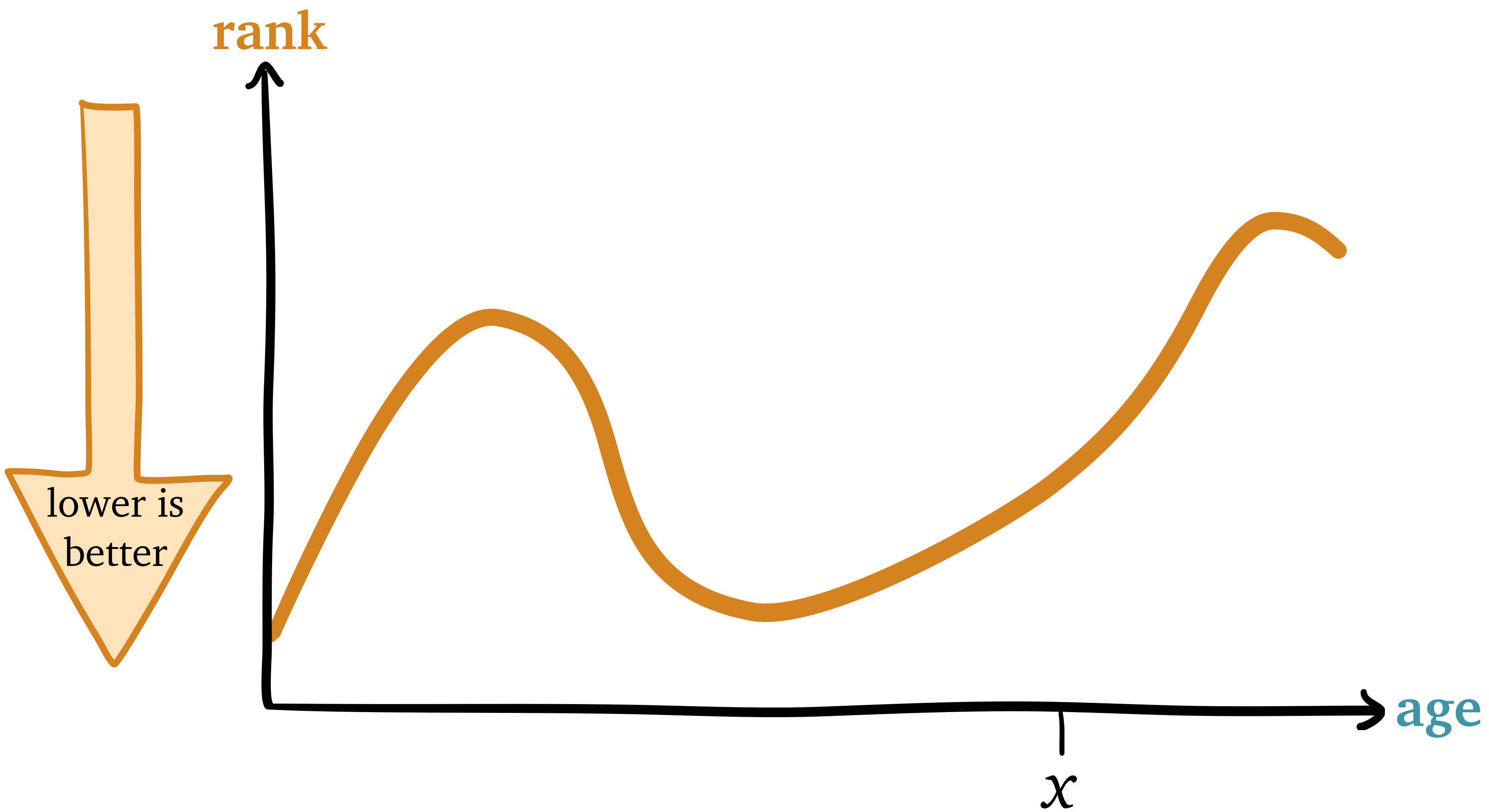


Analyzing **Gittins-1**



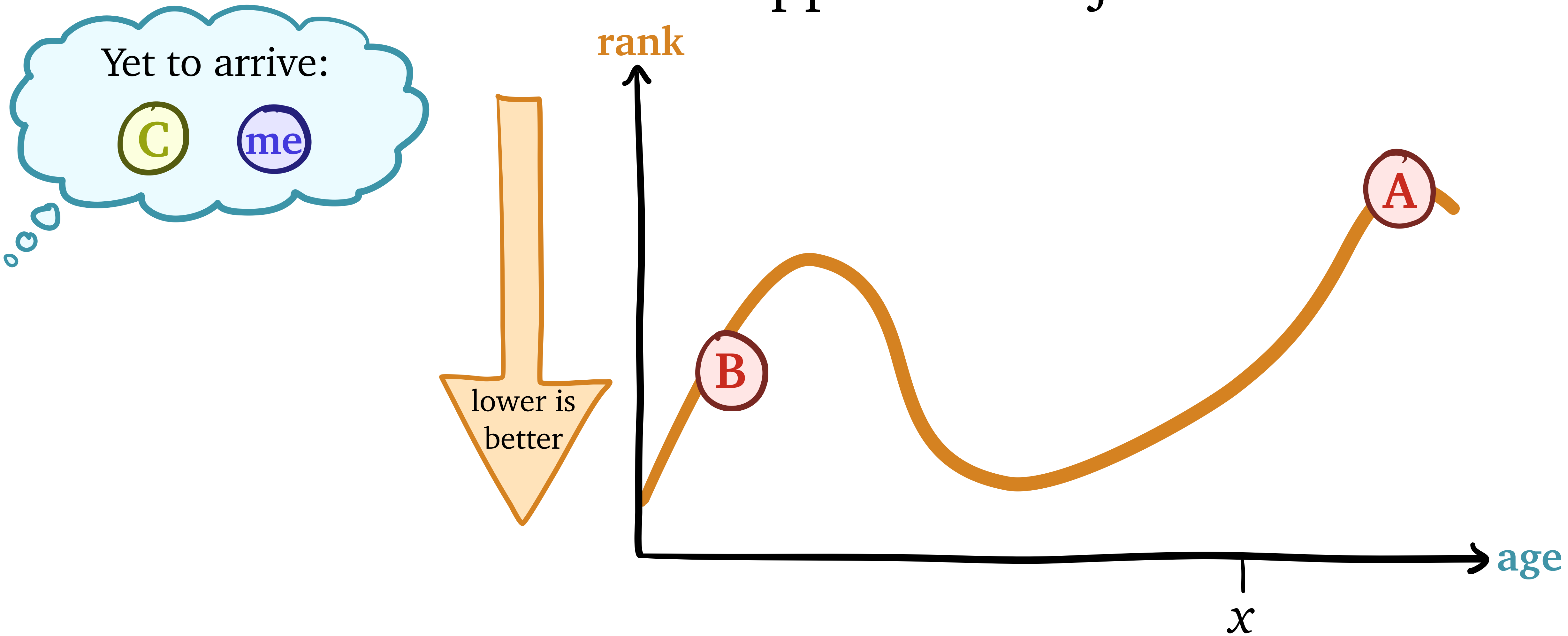
Analyzing **Gittins-1**

Suppose I'm a job of size x



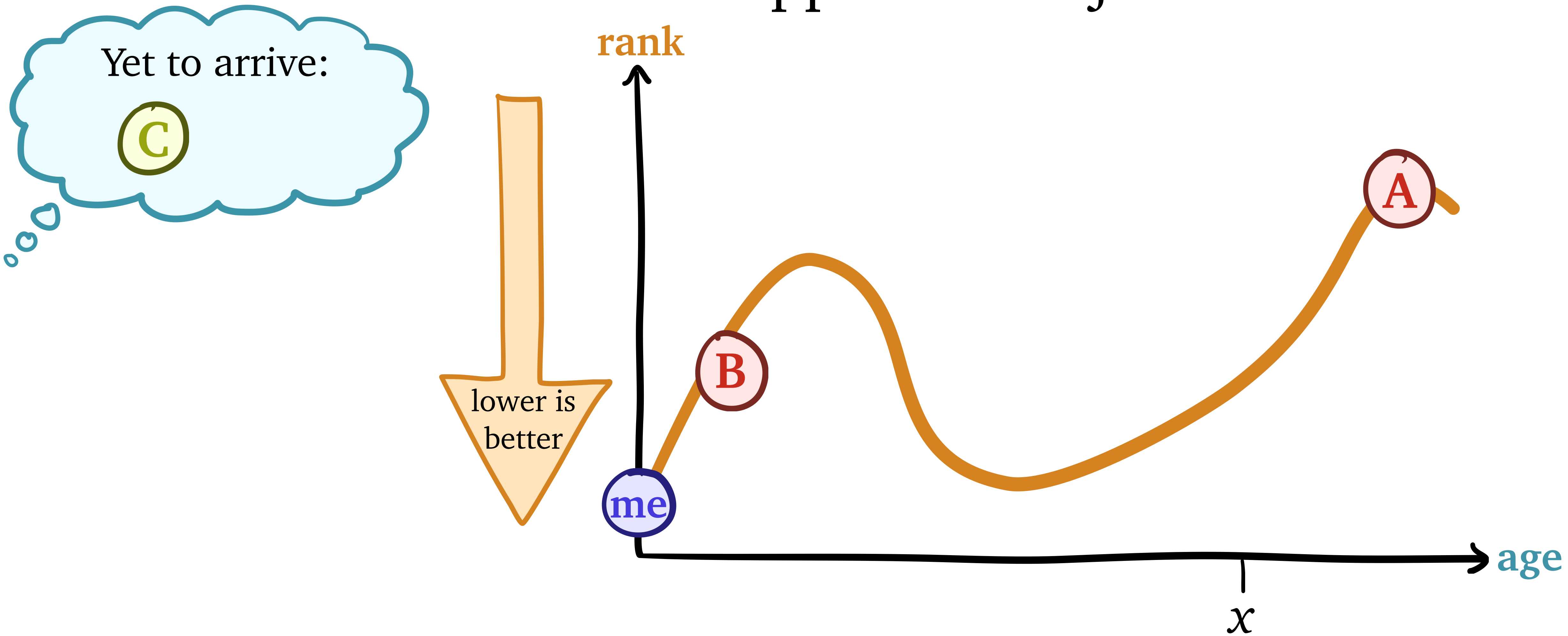
Analyzing Gittins-1

Suppose I'm a job of size x

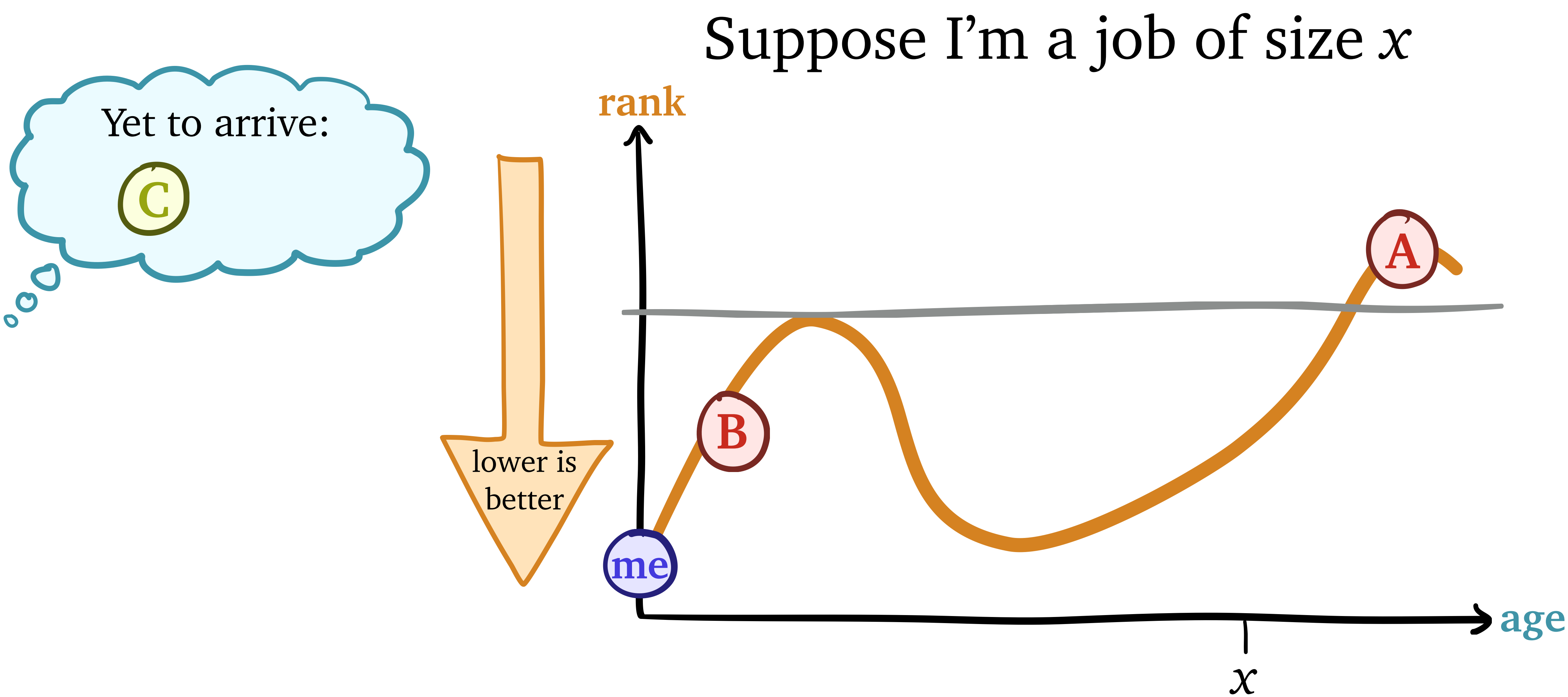


Analyzing **Gittins-1**

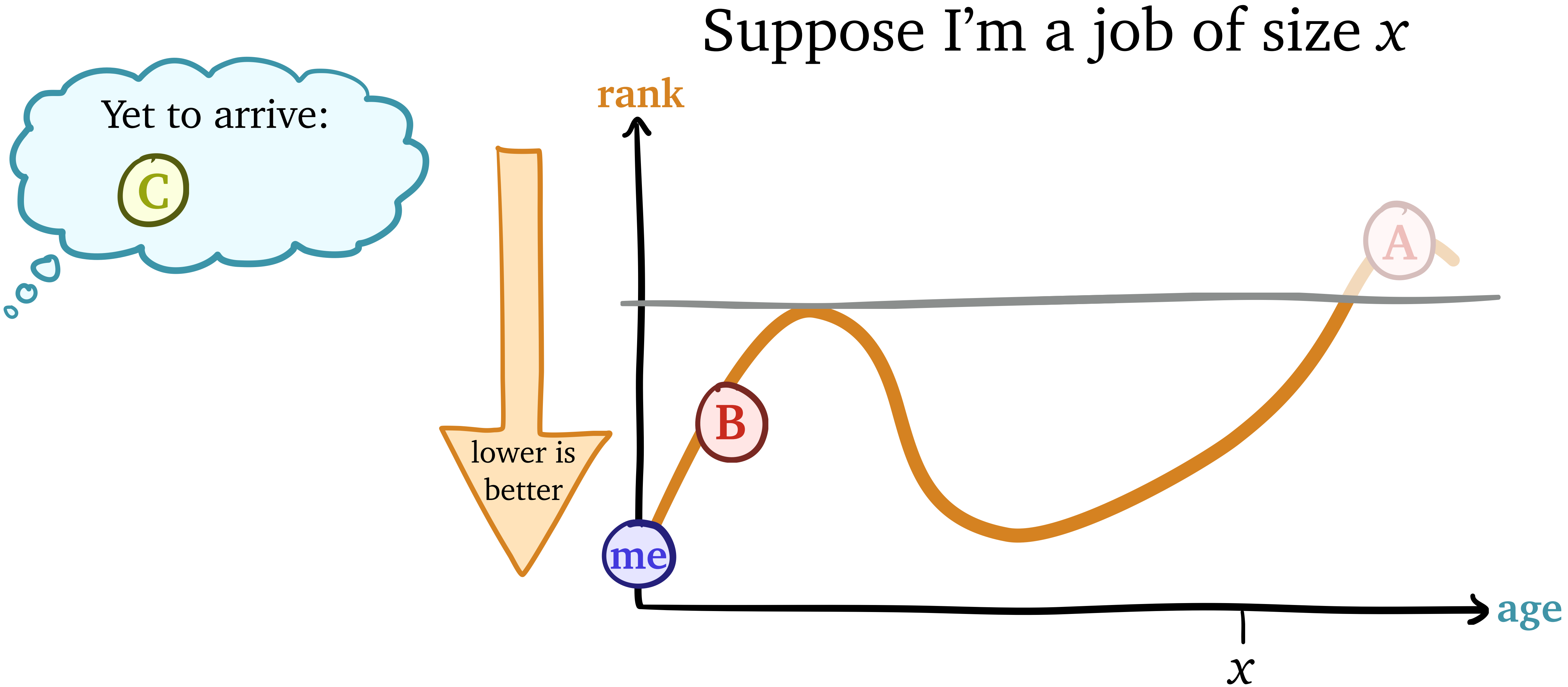
Suppose I'm a job of size x



Analyzing **Gittins-1**

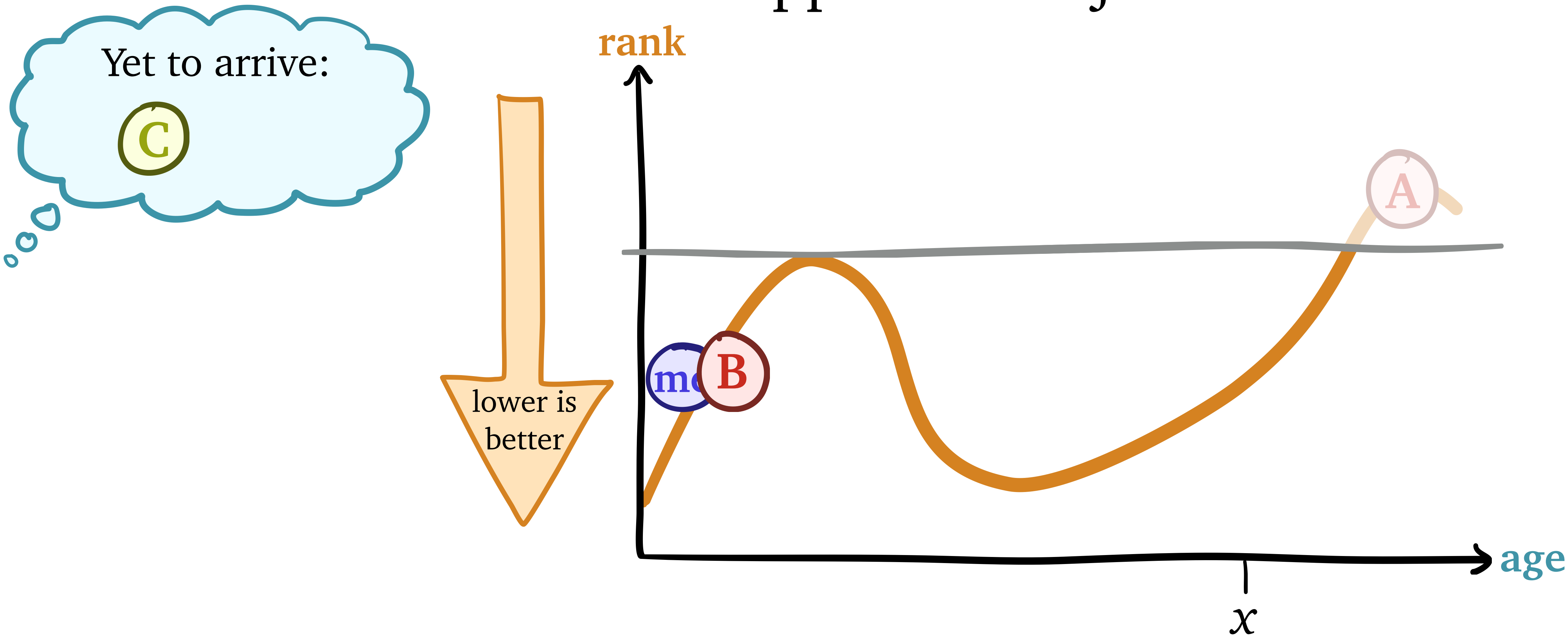


Analyzing **Gittins-1**



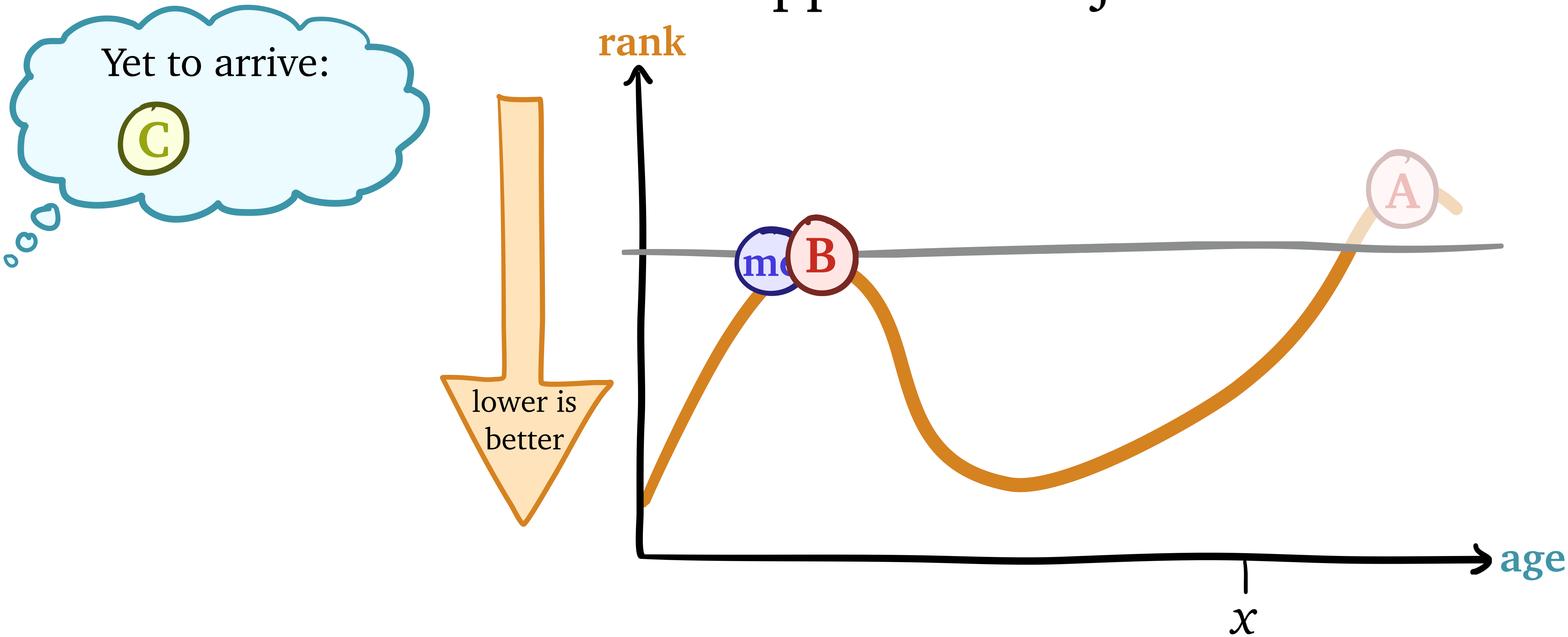
Analyzing **Gittins-1**

Suppose I'm a job of size x



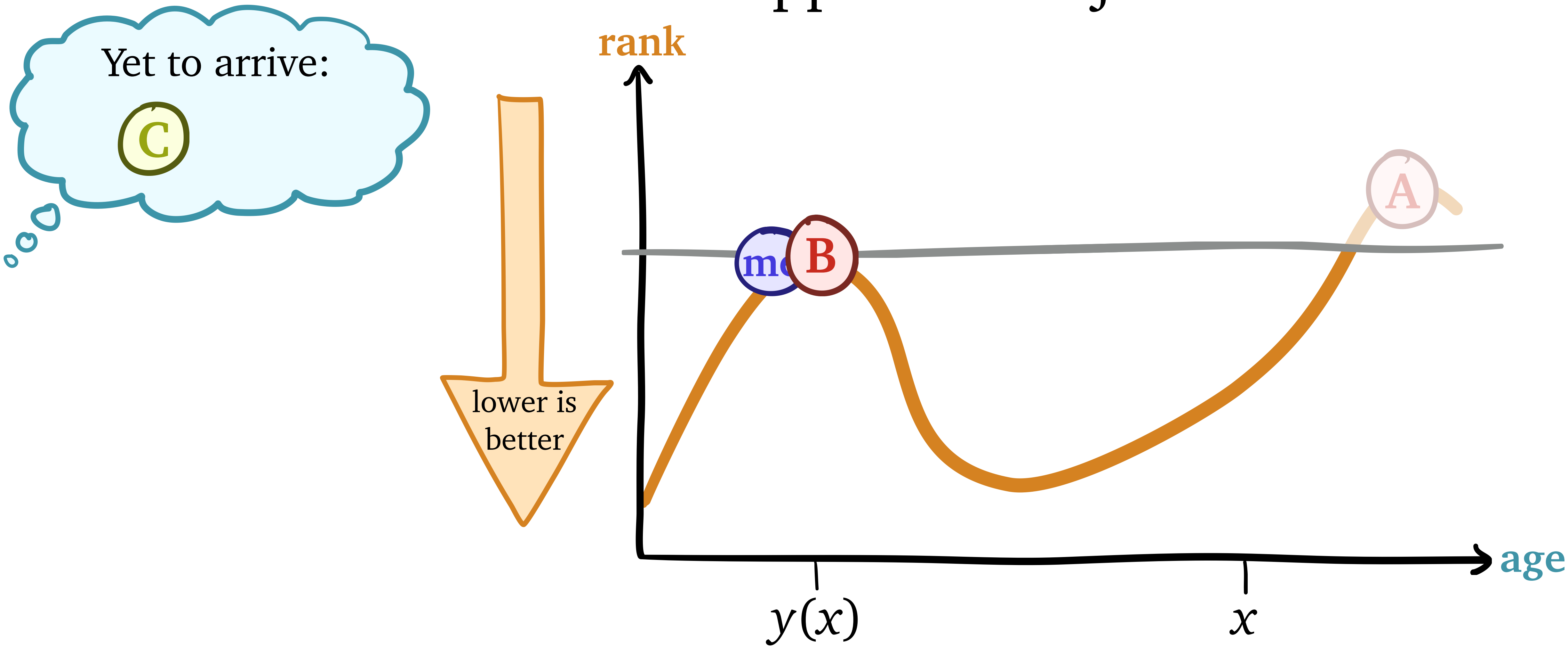
Analyzing **Gittins-1**

Suppose I'm a job of size x



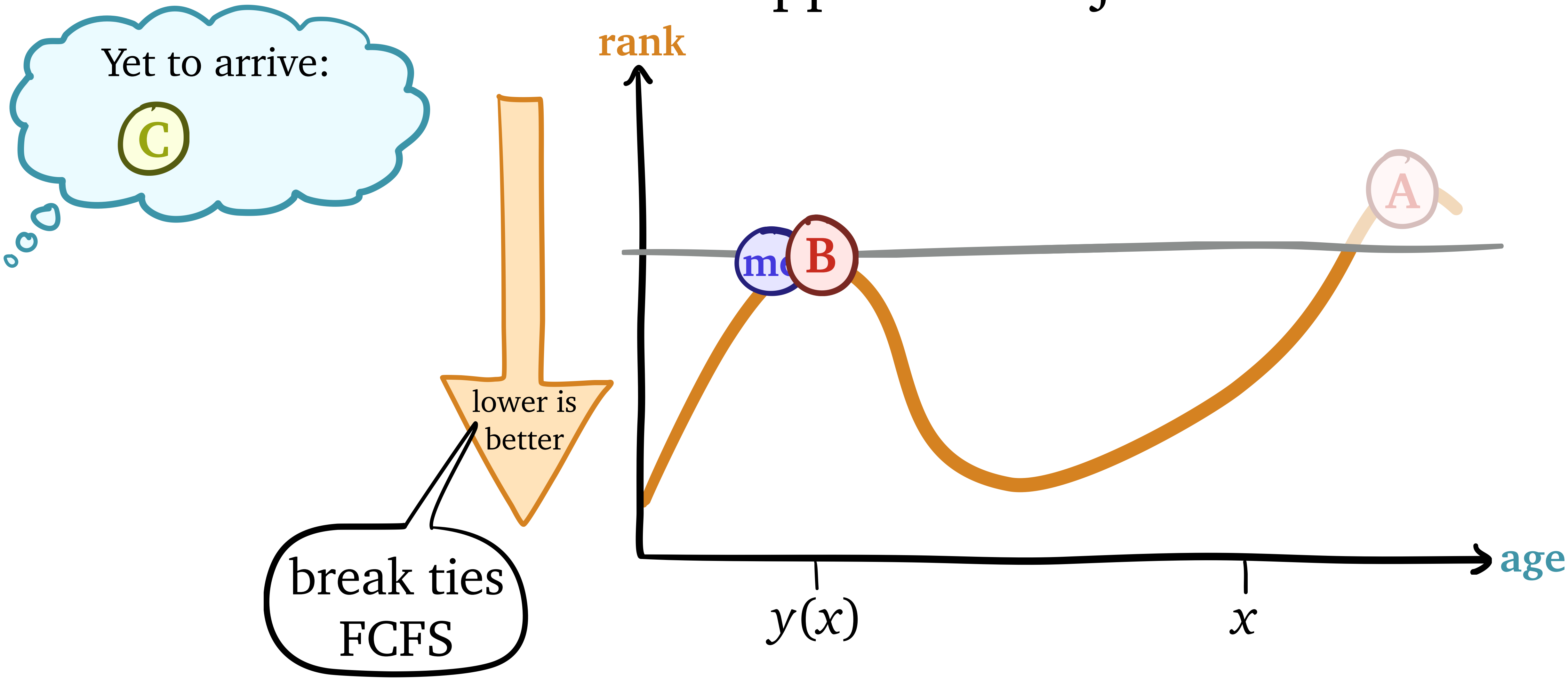
Analyzing **Gittins-1**

Suppose I'm a job of size x



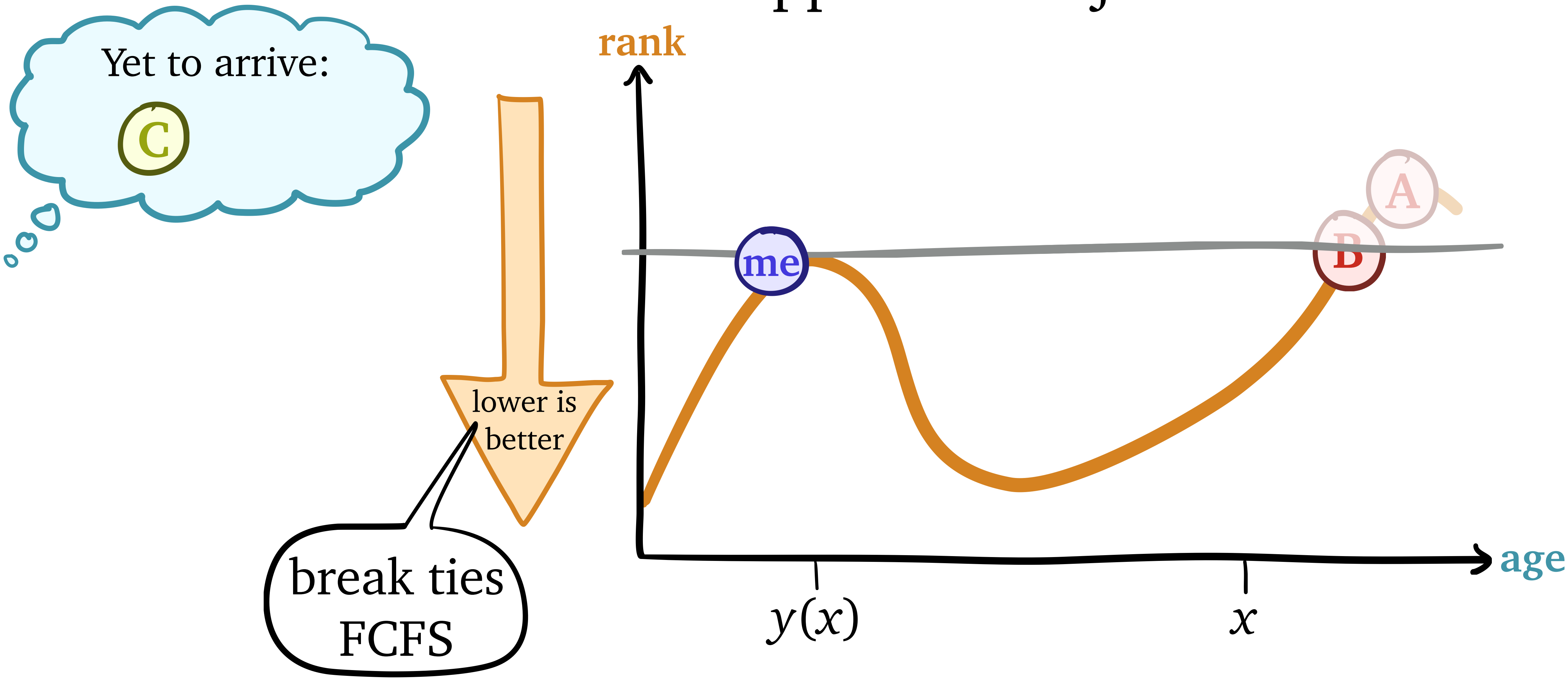
Analyzing **Gittins-1**

Suppose I'm a job of size x



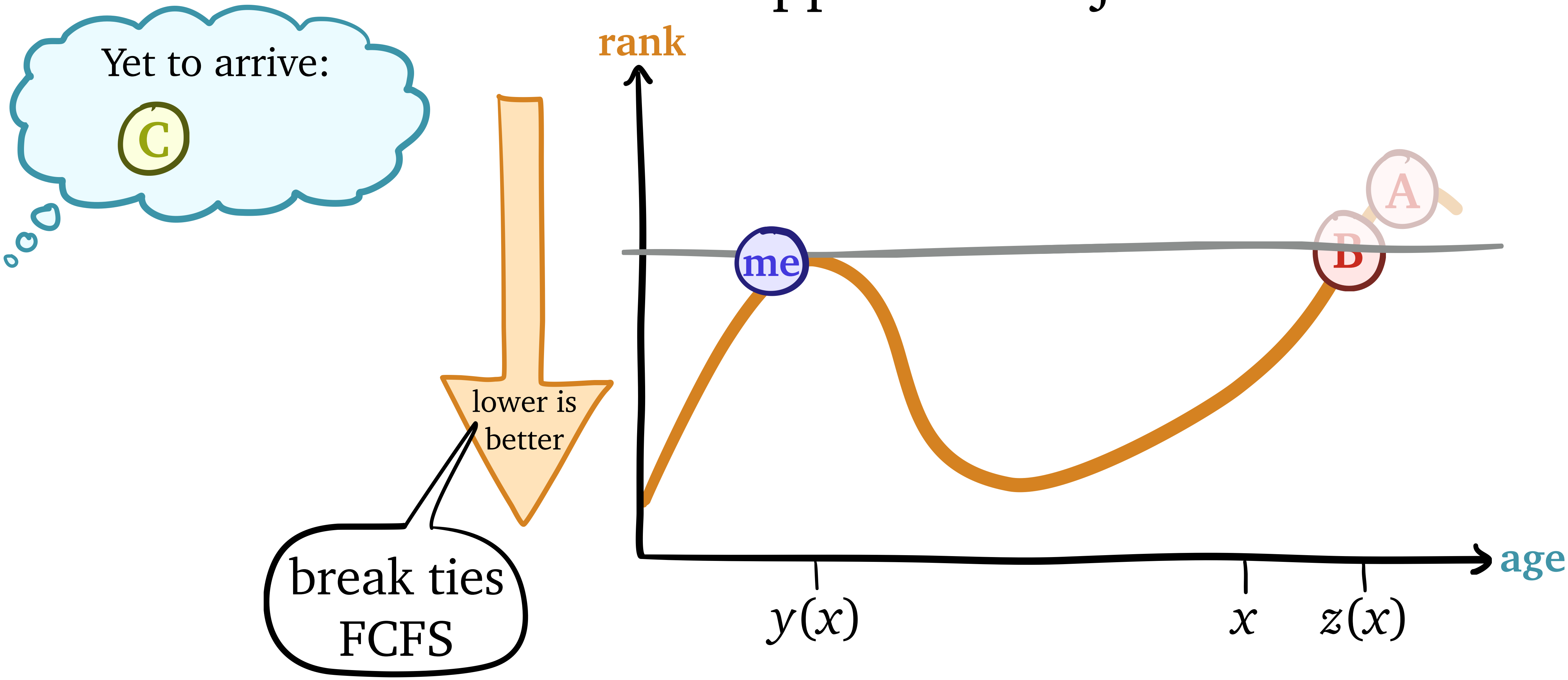
Analyzing **Gittins-1**

Suppose I'm a job of size x



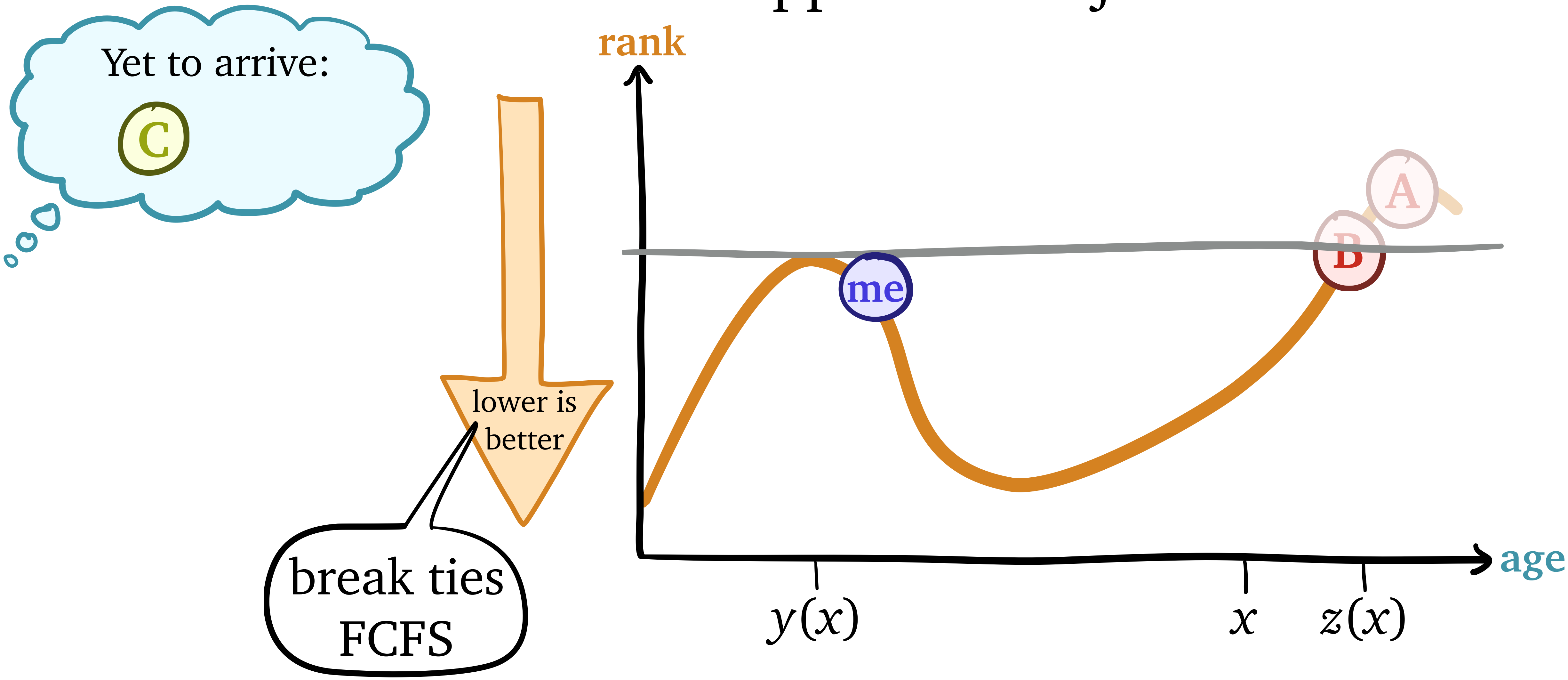
Analyzing **Gittins-1**

Suppose I'm a job of size x



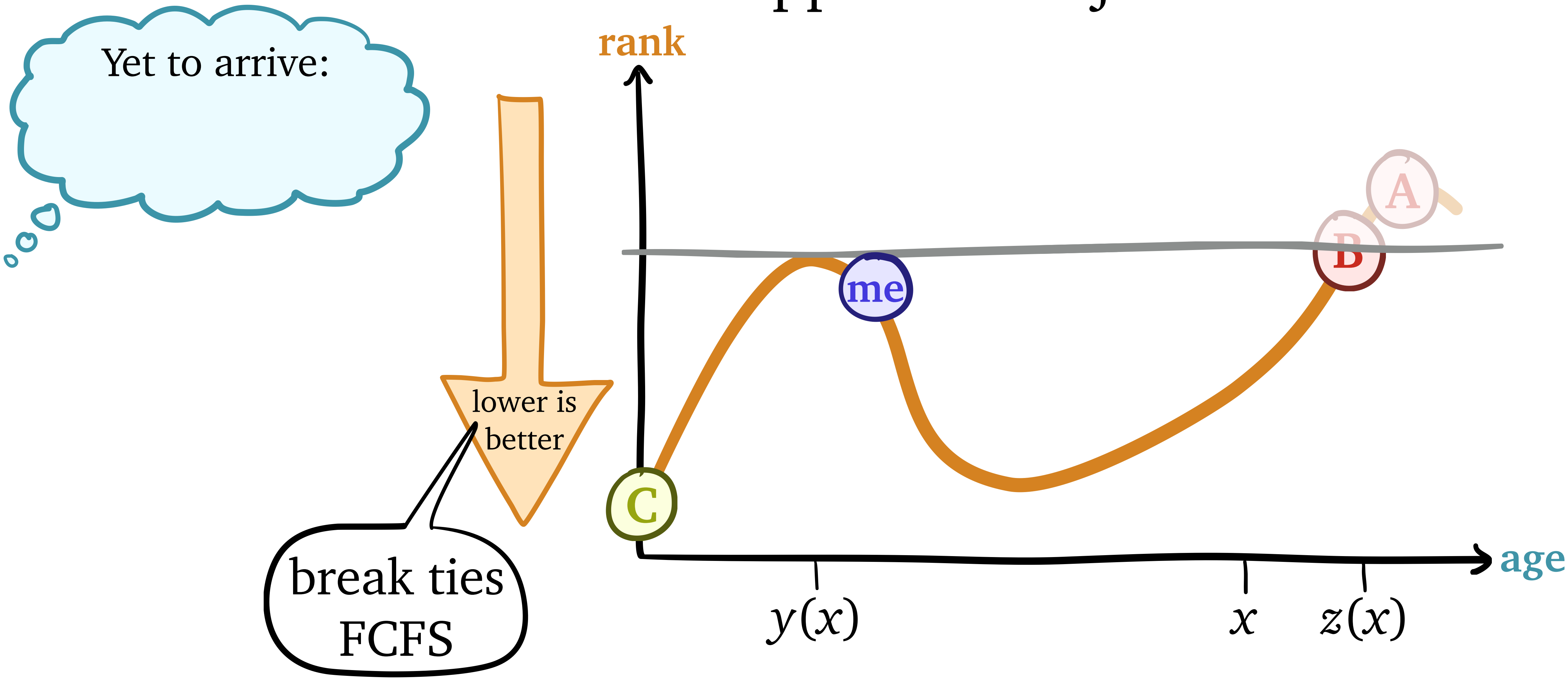
Analyzing **Gittins-1**

Suppose I'm a job of size x



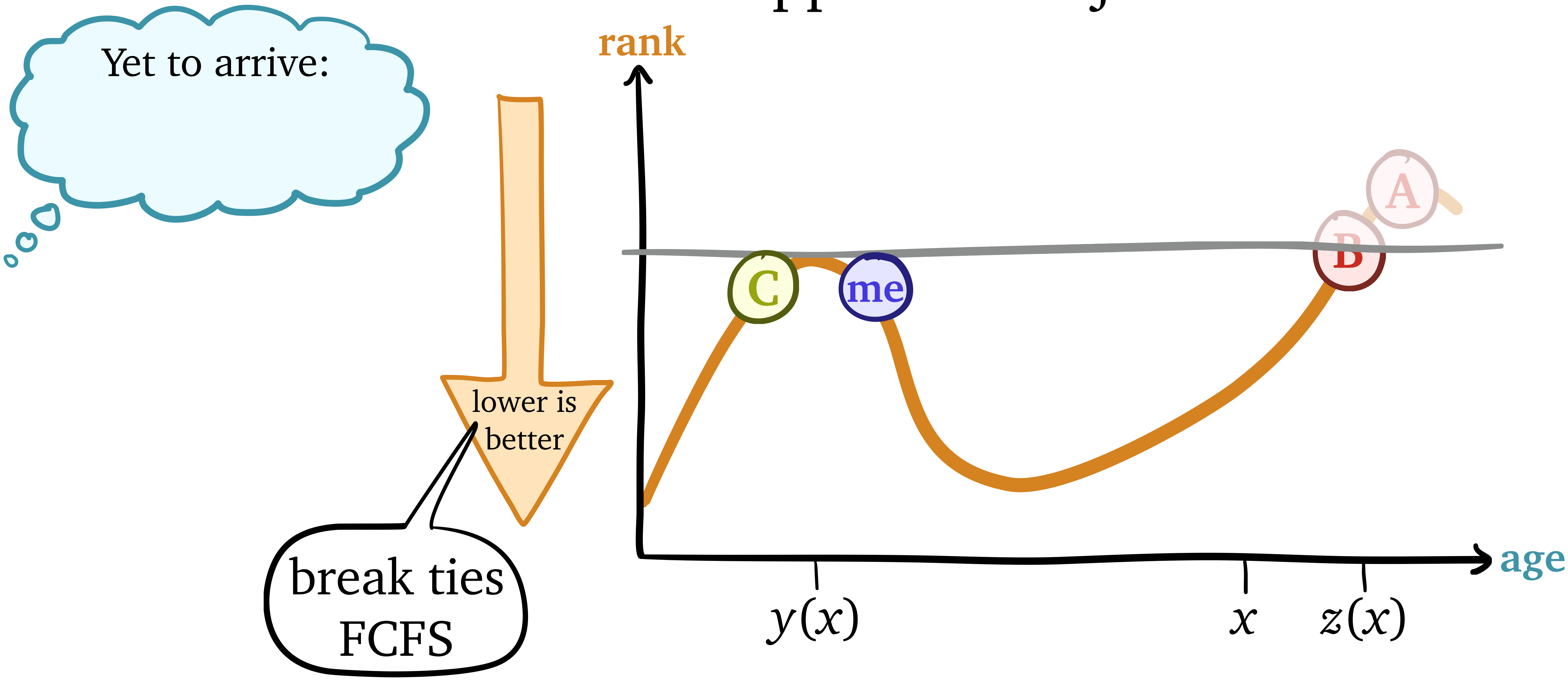
Analyzing **Gittins-1**

Suppose I'm a job of size x



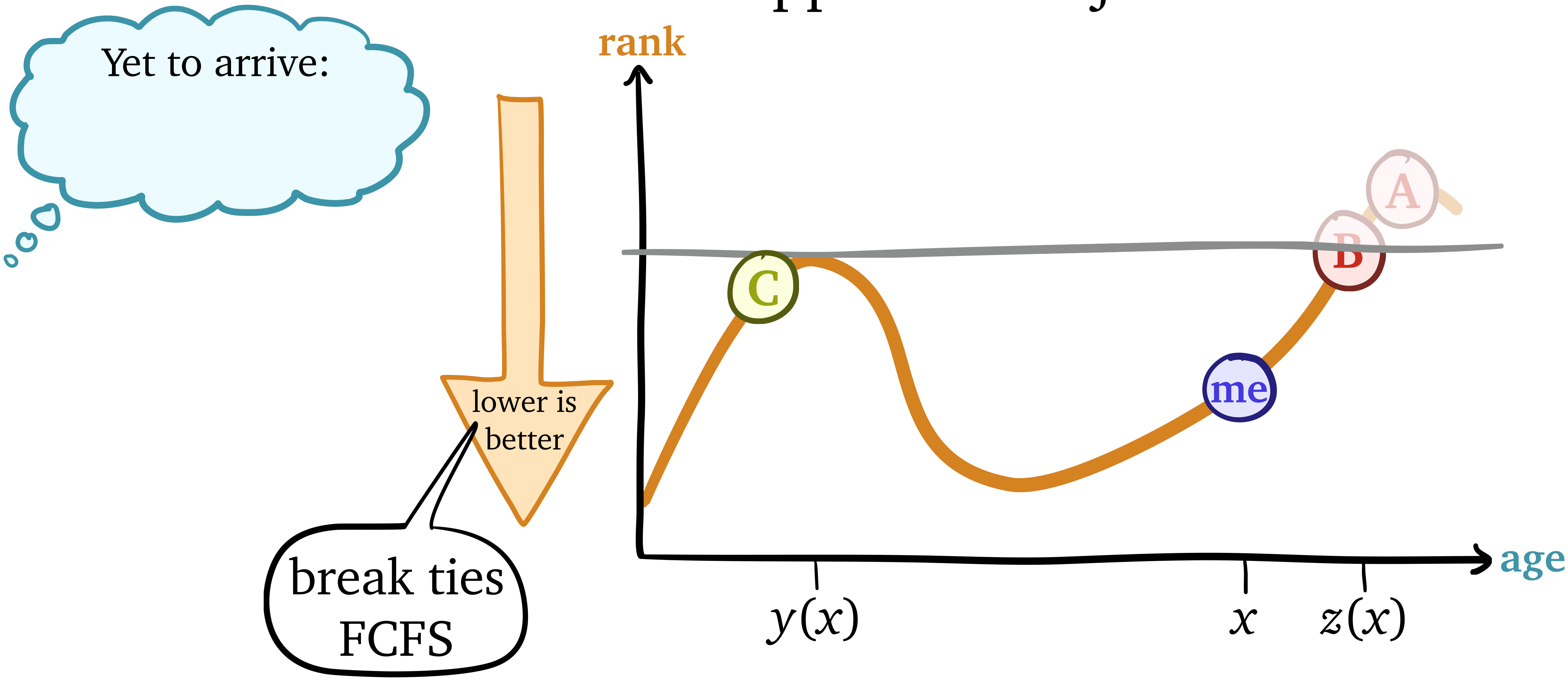
Analyzing **Gittins-1**

Suppose I'm a job of size x



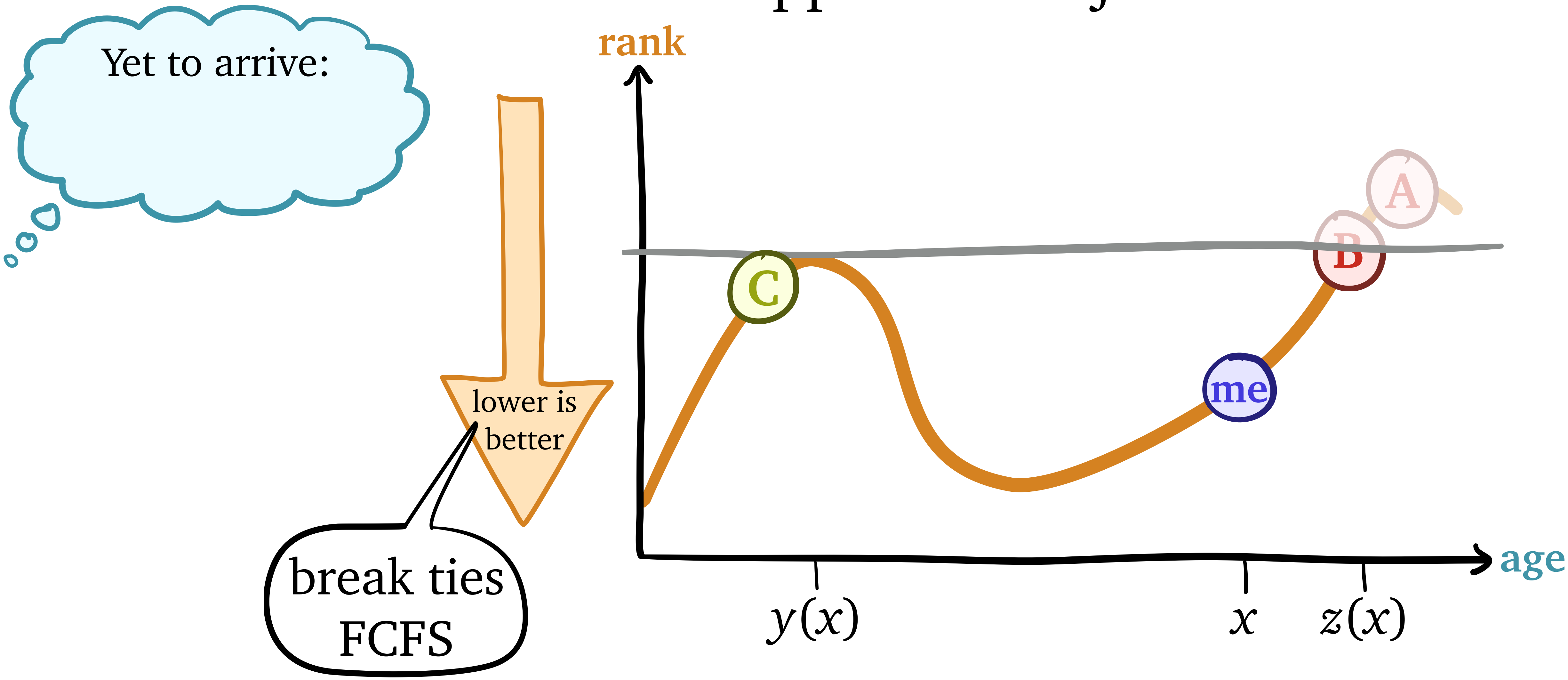
Analyzing **Gittins-1**

Suppose I'm a job of size x



Analyzing **Gittins-1**

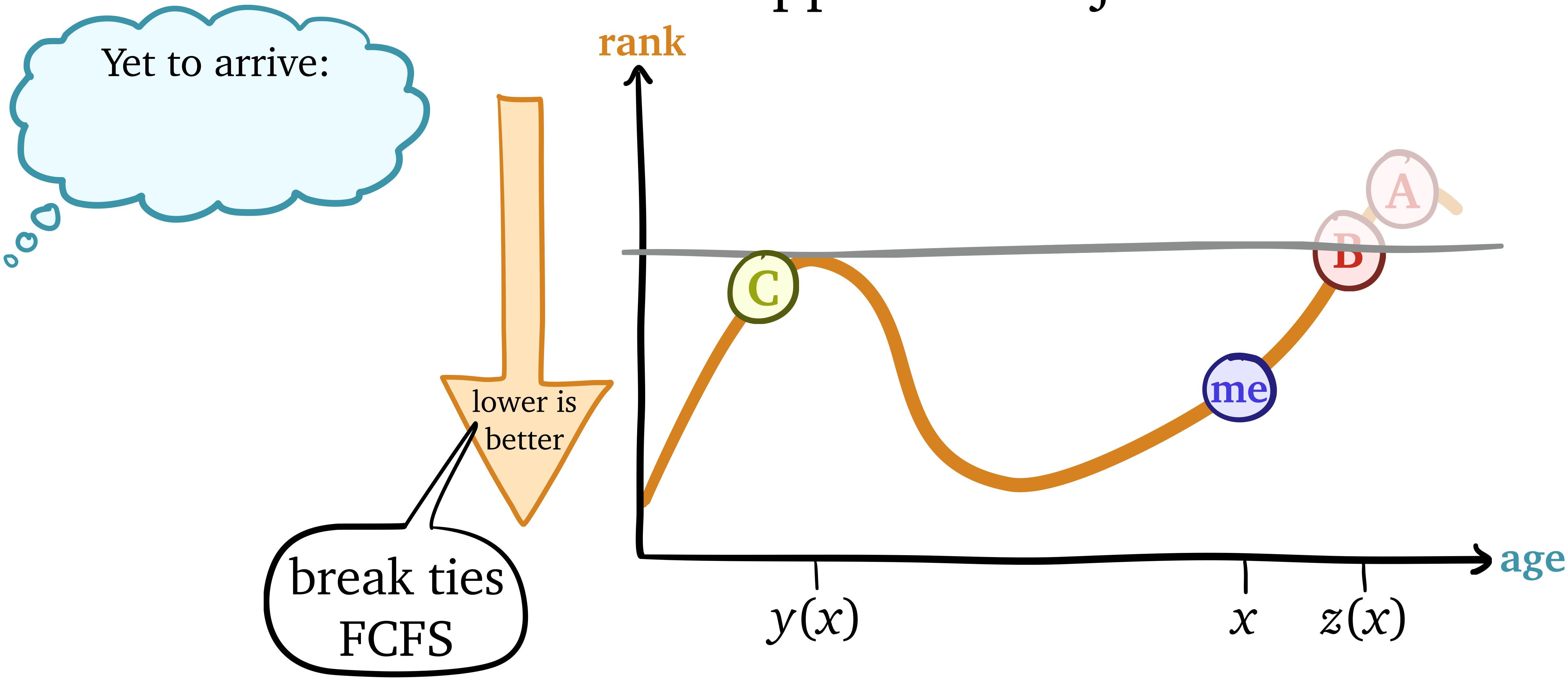
Suppose I'm a job of size x



I can ignore { *old* jobs (**A** & **B**) / *new* jobs (**C**) } after age { ??? / ??? }

Analyzing **Gittins-1**

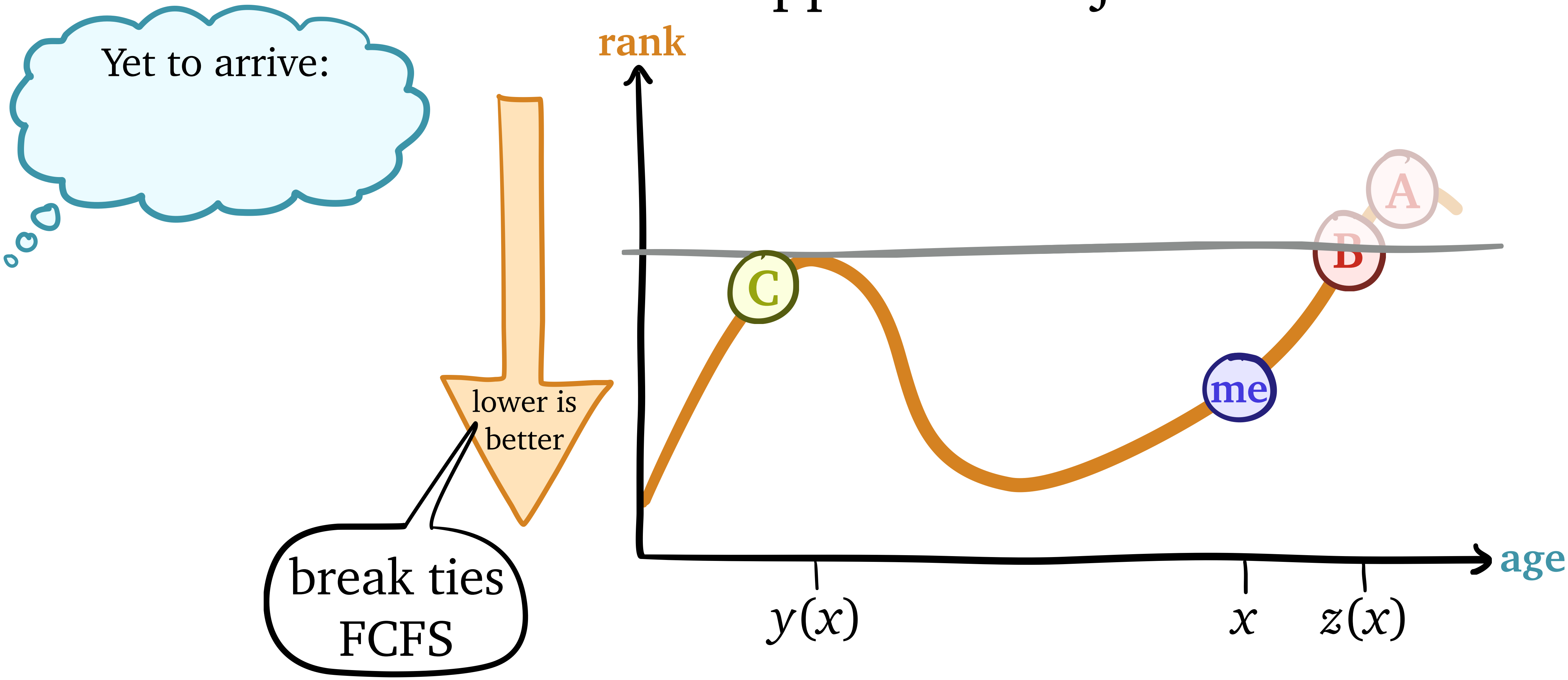
Suppose I'm a job of size x



I can ignore $\left\{\begin{array}{l} \textit{old} \text{ jobs (A \& B)} \\ \textit{new} \text{ jobs (C)} \end{array}\right\}$ after age $\left\{\begin{array}{l} z(x) \\ ??? \end{array}\right\}$

Analyzing **Gittins-1**

Suppose I'm a job of size x

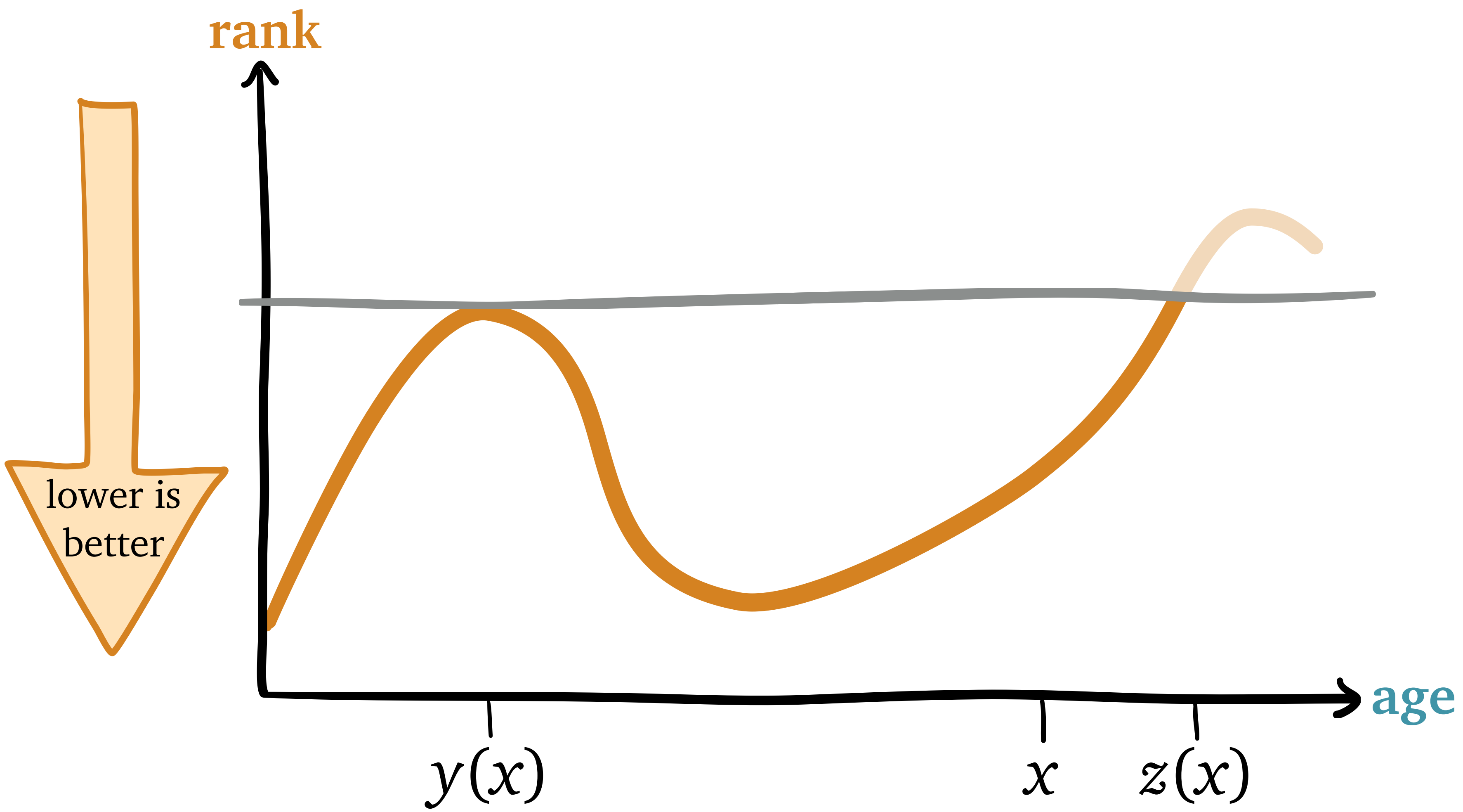


I can ignore { *old* jobs (**A** & **B**) / *new* jobs (**C**) } after age { $z(x)$ / $y(x)$ }

Question:
What goes wrong for **Gittins-*k***?

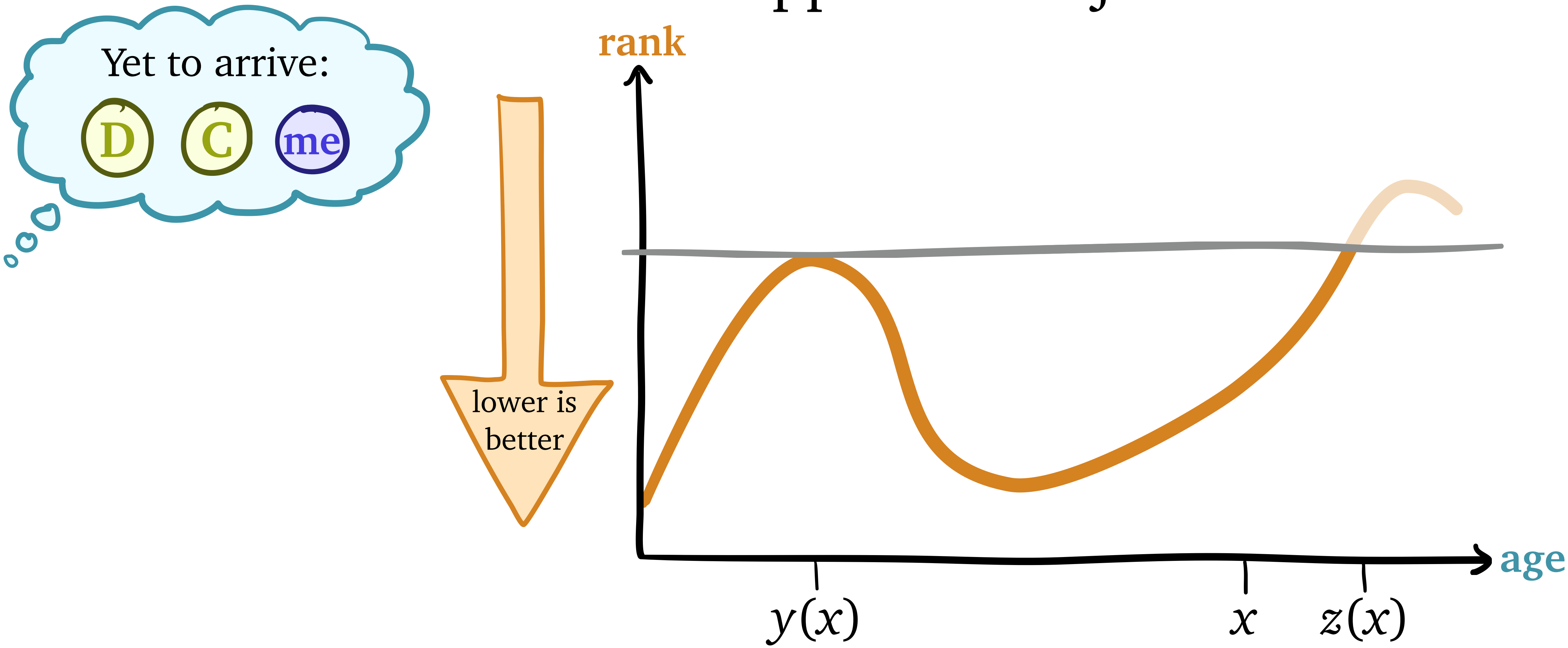
Analyzing **Gittins-*k***

Suppose I'm a job of size x



Analyzing **Gittins-*k***

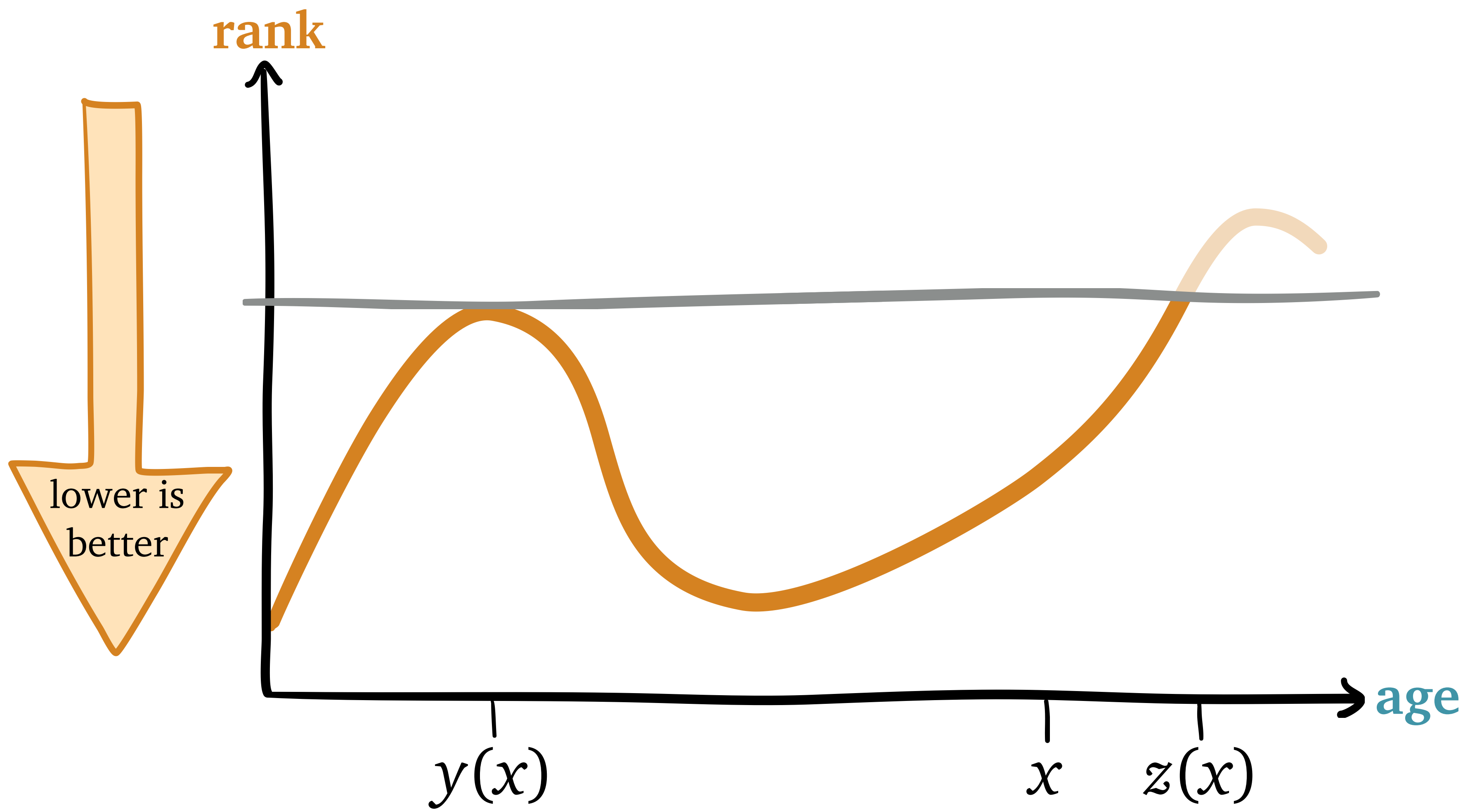
Suppose I'm a job of size x



I can ignore $\left\{ \begin{array}{c} \textit{old} \text{ jobs} \\ \textit{new} \text{ jobs (C \& D)} \end{array} \right\}$ after age $\left\{ \begin{array}{c} ??? \\ ??? \end{array} \right\}$

Analyzing **Gittins-k**

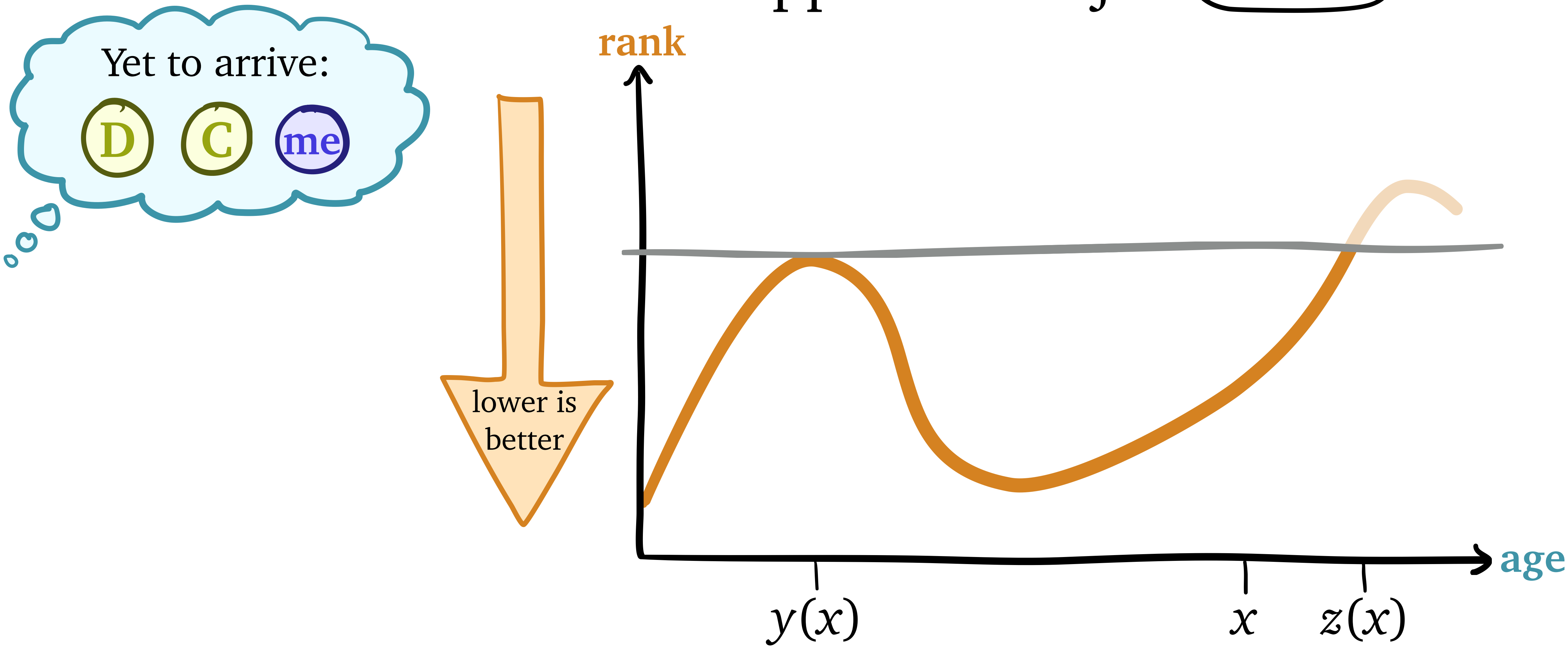
Suppose I'm a job x ($k = 2$)



I can ignore { *old* jobs / *new* jobs (C & D) } after age { ??? / ??? }

Analyzing **Gittins-k**

Suppose I'm a job $k = 2$ x

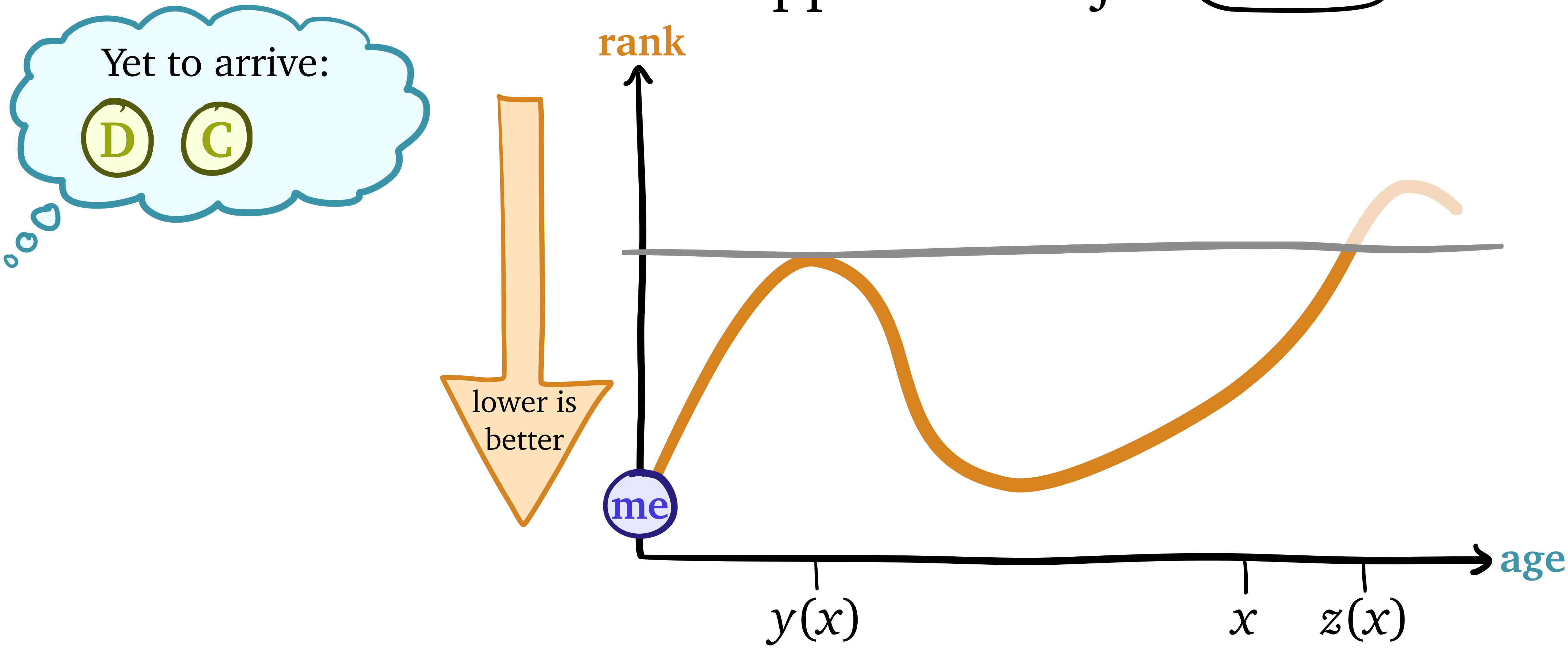


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

Analyzing **Gittins-k**

Suppose I'm a job x

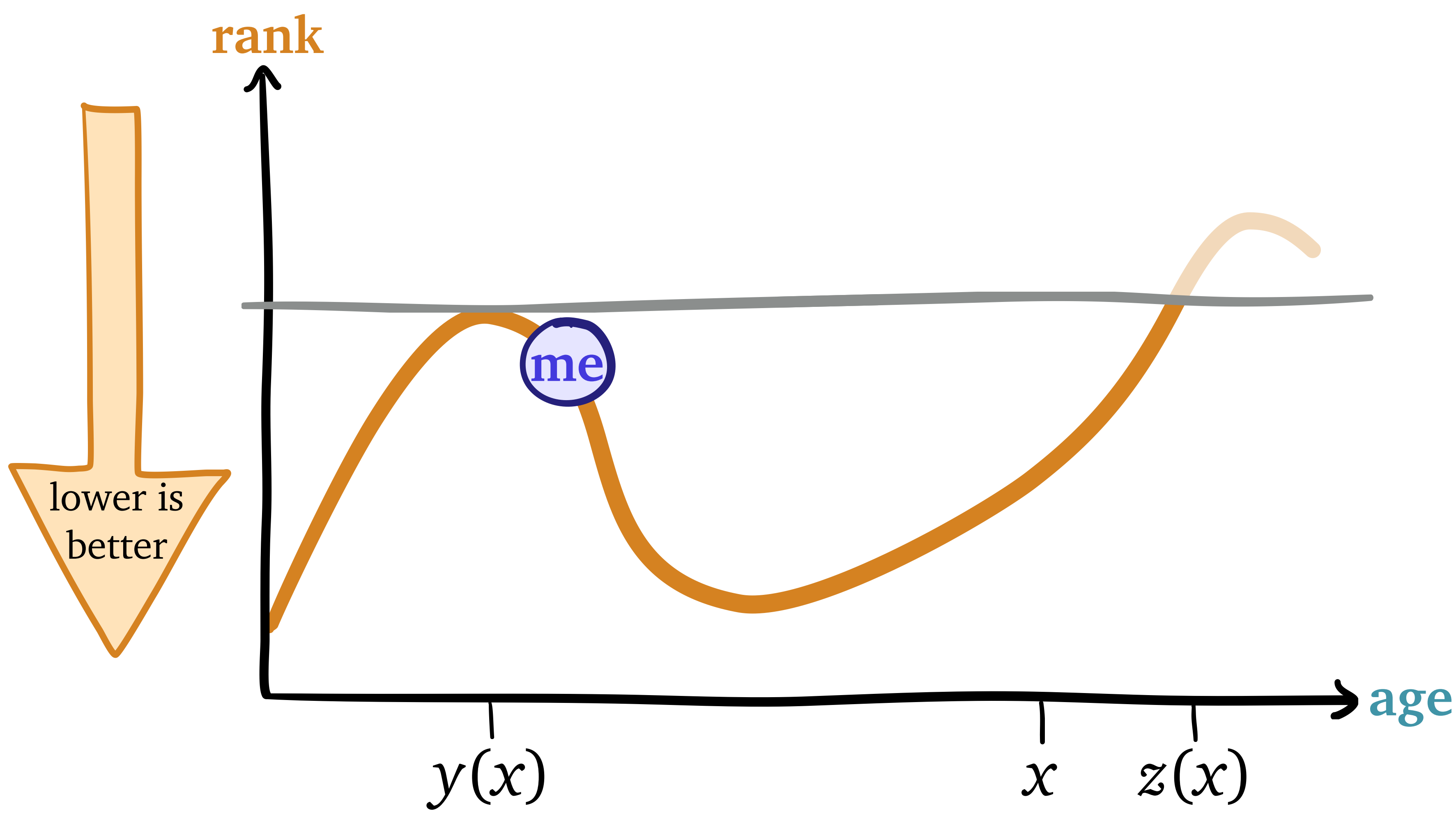
$k = 2$



I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

Analyzing **Gittins-*k***

Suppose I'm a job ($k = 2$) x

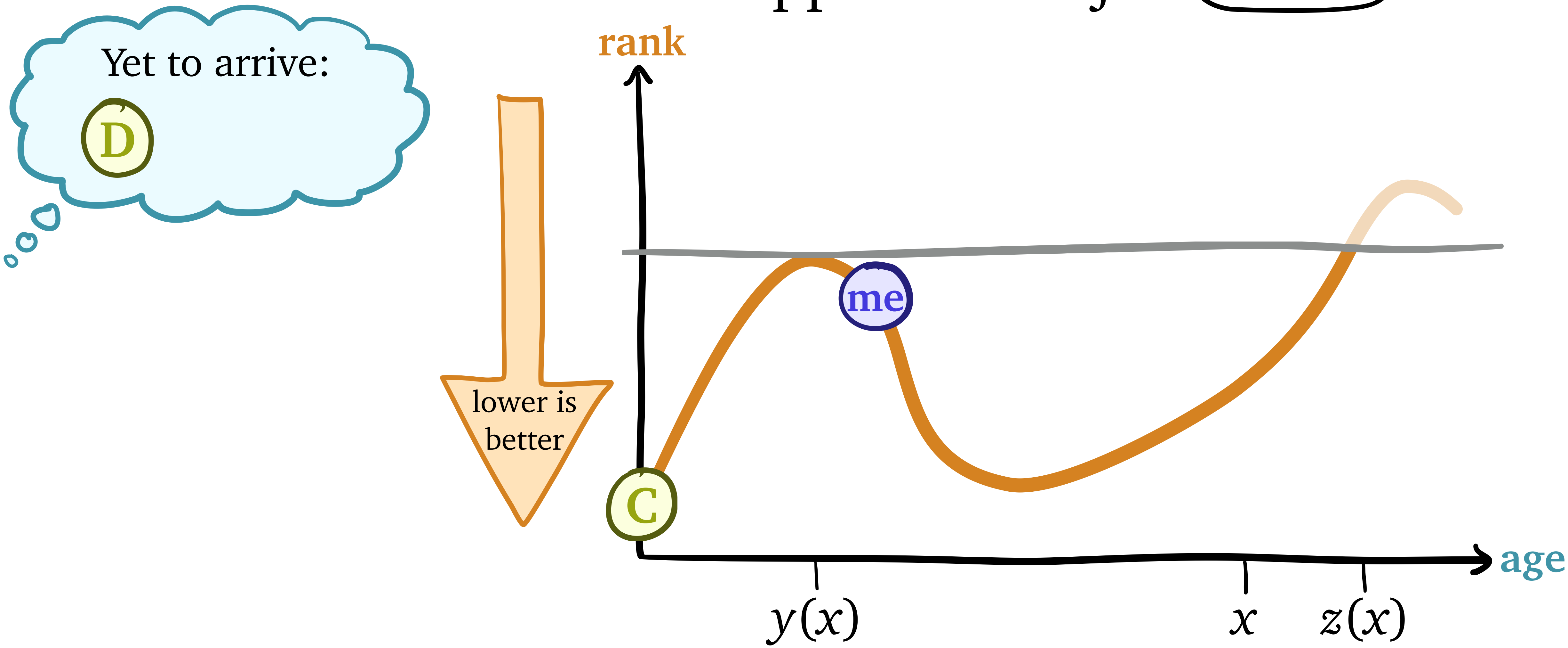


I can ignore { *old* jobs / *new* jobs (**C** & **D**) } after age { $z(x)$ / **???** }

Analyzing **Gittins-*k***

$k = 2$

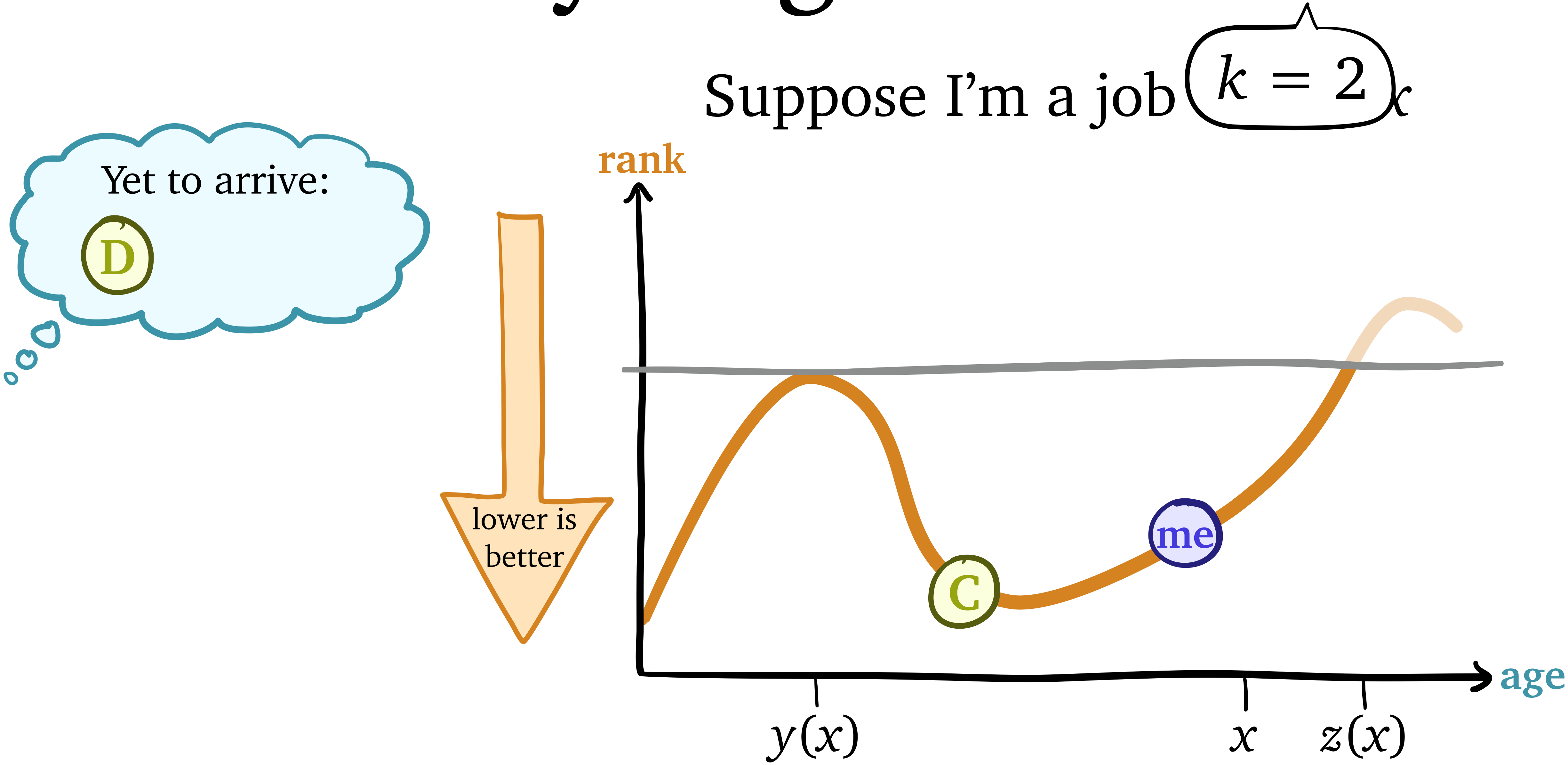
Suppose I’m a job x



I can ignore $\left\{ \begin{array}{c} \textit{old} \text{ jobs} \\ \textit{new} \text{ jobs (C \& D)} \end{array} \right\}$ after age $\left\{ \begin{array}{c} z(x) \\ ??? \end{array} \right\}$

Analyzing **Gittins-*k***

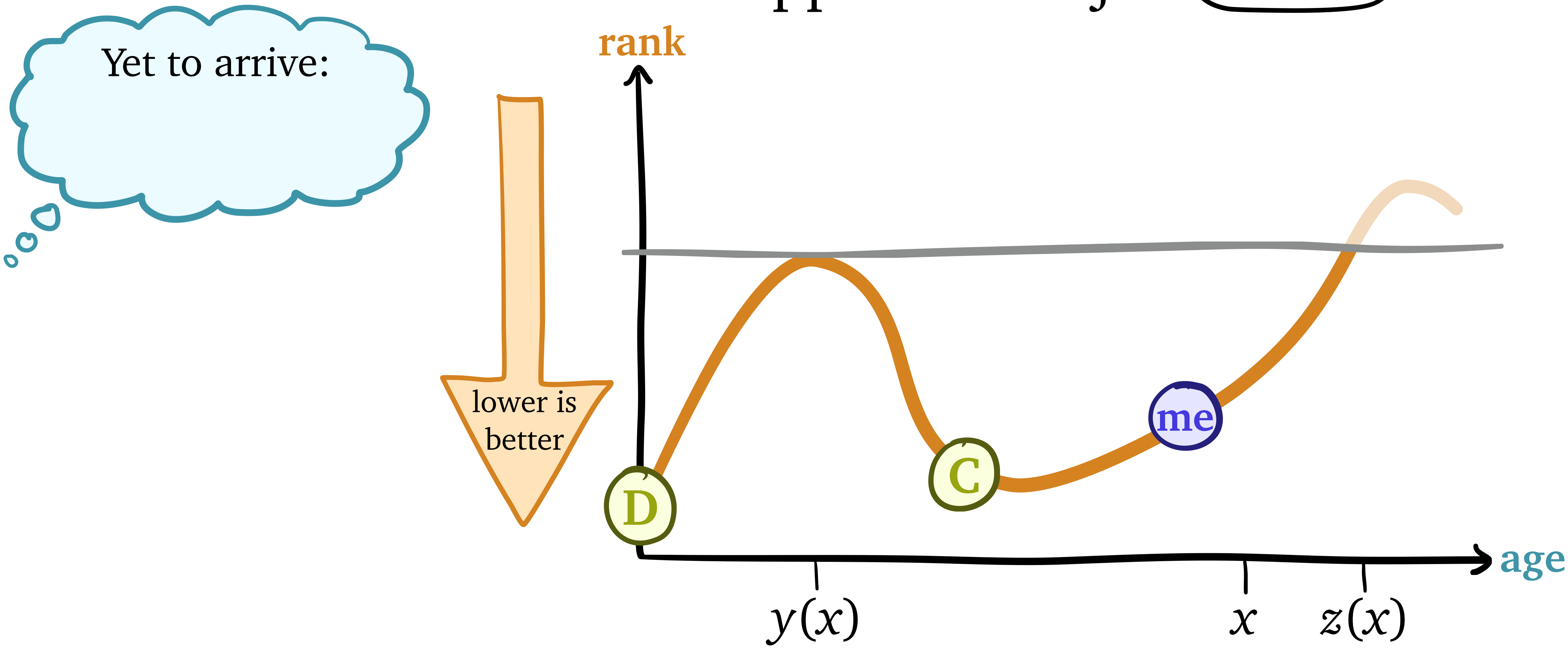
Suppose I'm a job x



I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

Analyzing **Gittins-k**

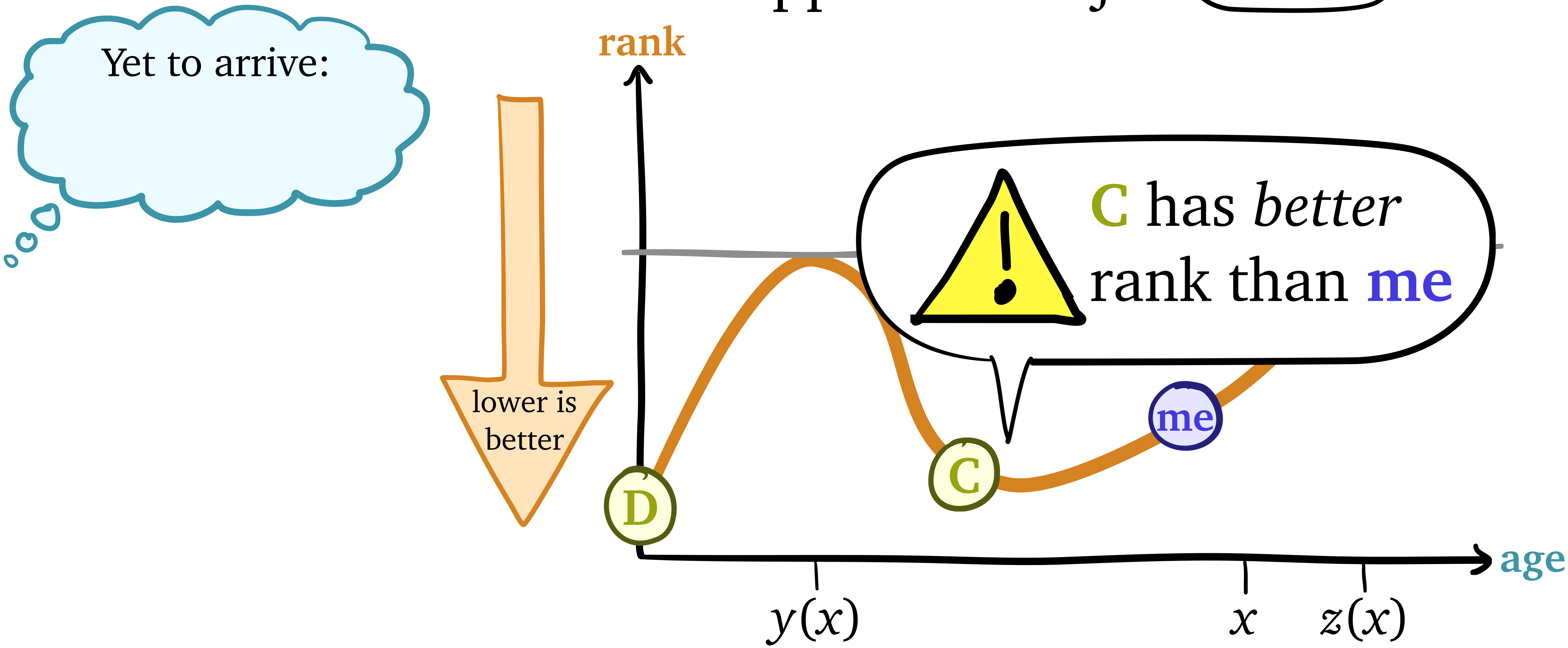
Suppose I'm a job $k = 2$ x



I can ignore { *old* jobs / *new* jobs (**C** & **D**) } after age { $z(x)$ / **???** }

Analyzing **Gittins-*k***

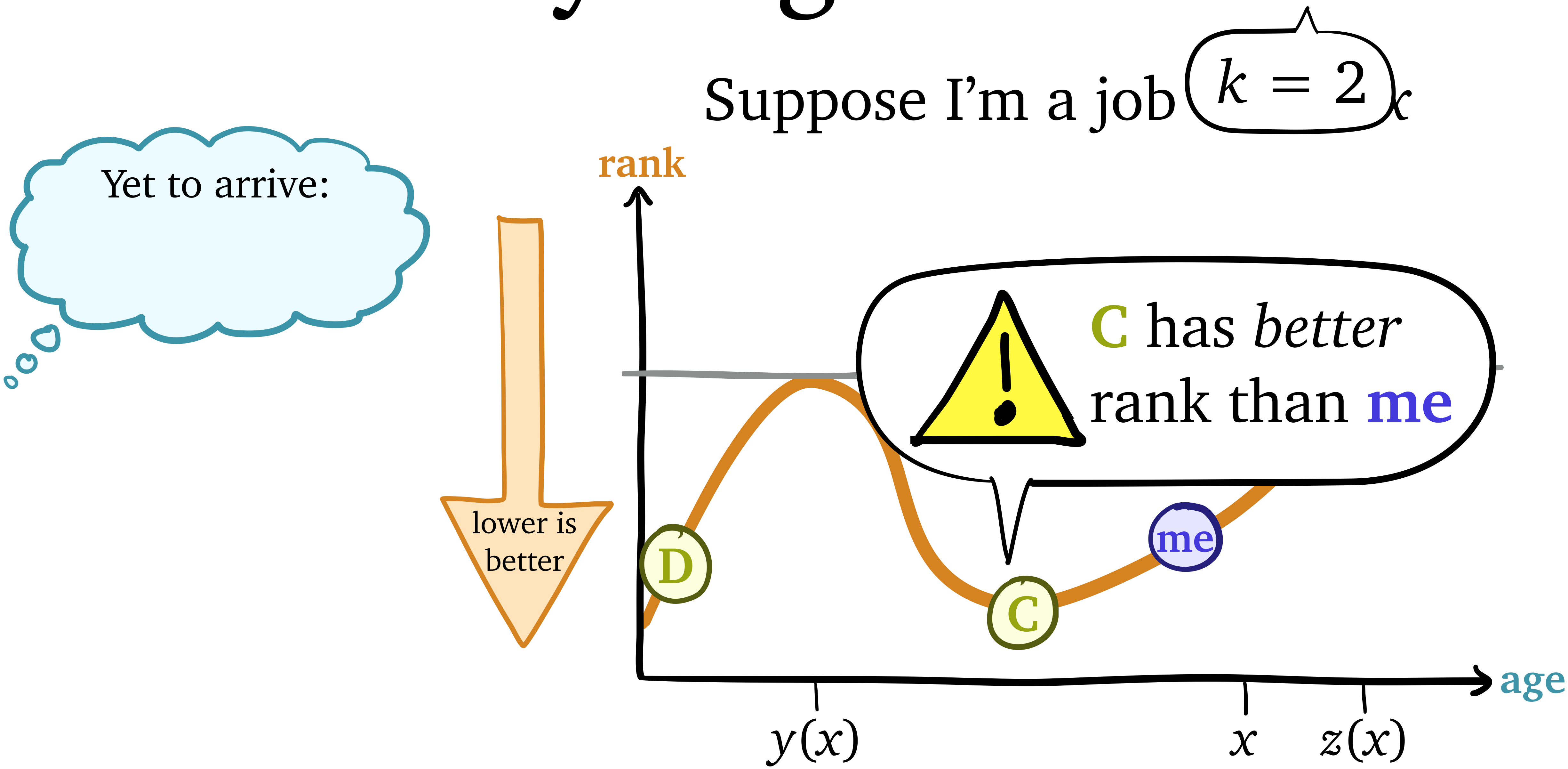
Suppose I'm a job x ($k = 2$)



I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

Analyzing Gittins-k

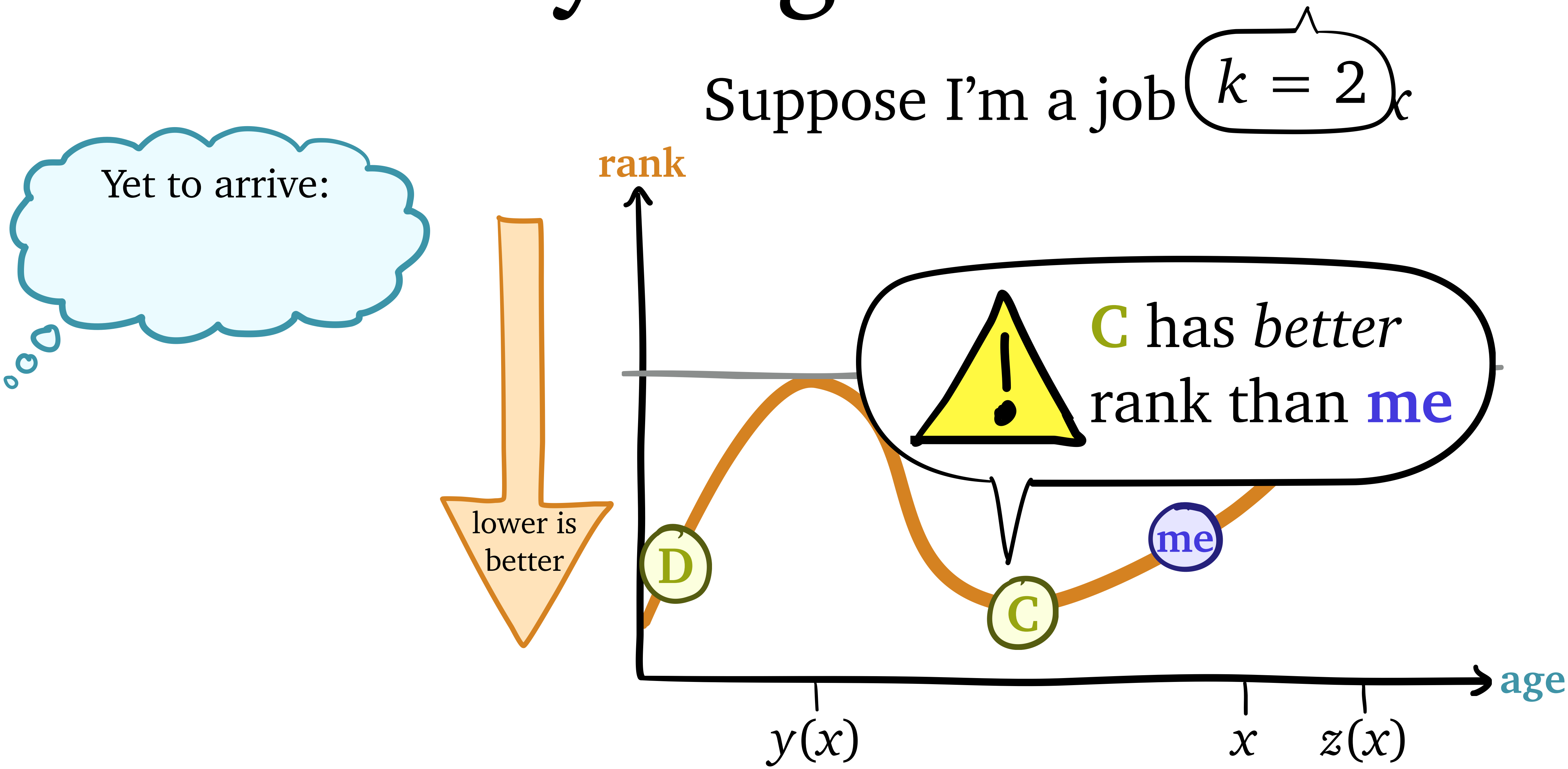
Suppose I'm a job x



I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

Analyzing Gittins-k

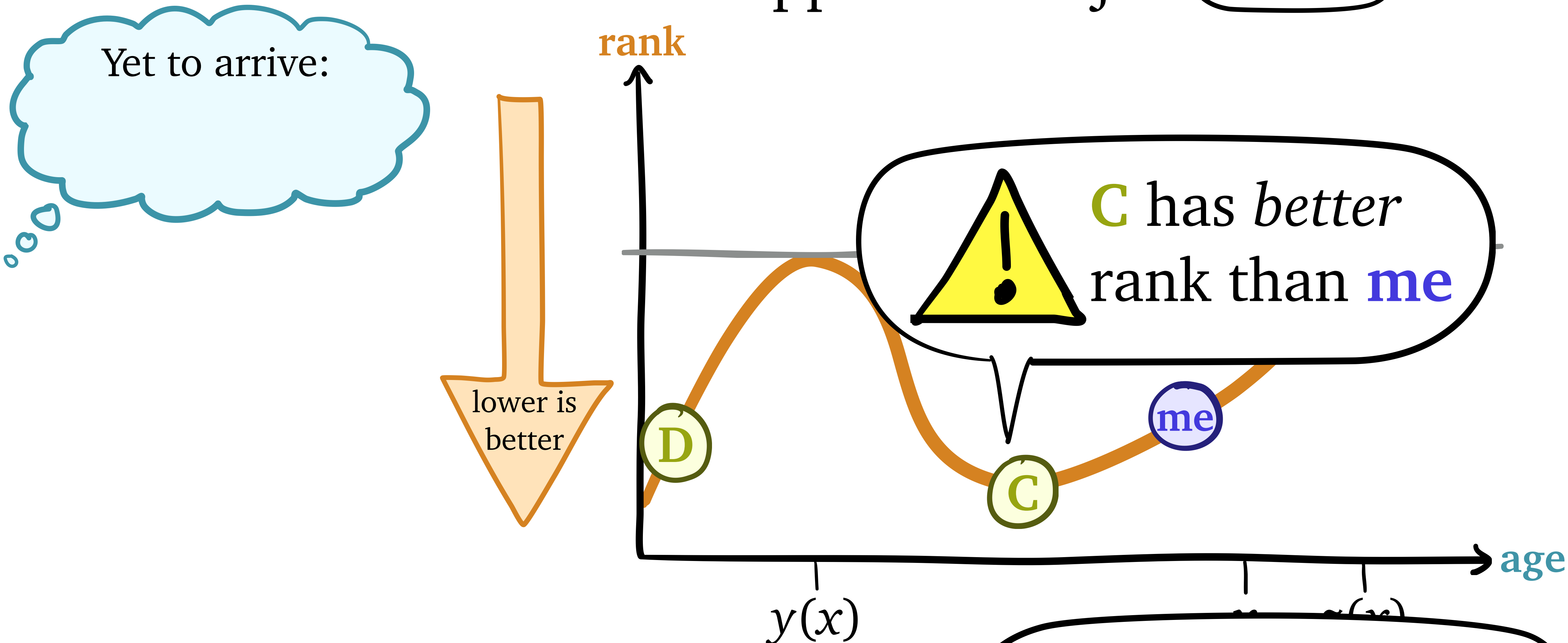
Suppose I'm a job x



I can ignore $\left\{\begin{array}{c}\textit{old}\text{ jobs} \\ \textit{new}\text{ jobs (C \& D)}\end{array}\right\}$ after age $\left\{\begin{array}{c} z(x) \\ x \end{array}\right\}$

Analyzing Gittins-k

Suppose I'm a job ($k = 2$) x



Was $y(x)$ for **Gittins-1**

I can ignore { *old* jobs / *new* jobs (C & D) } after age { x }

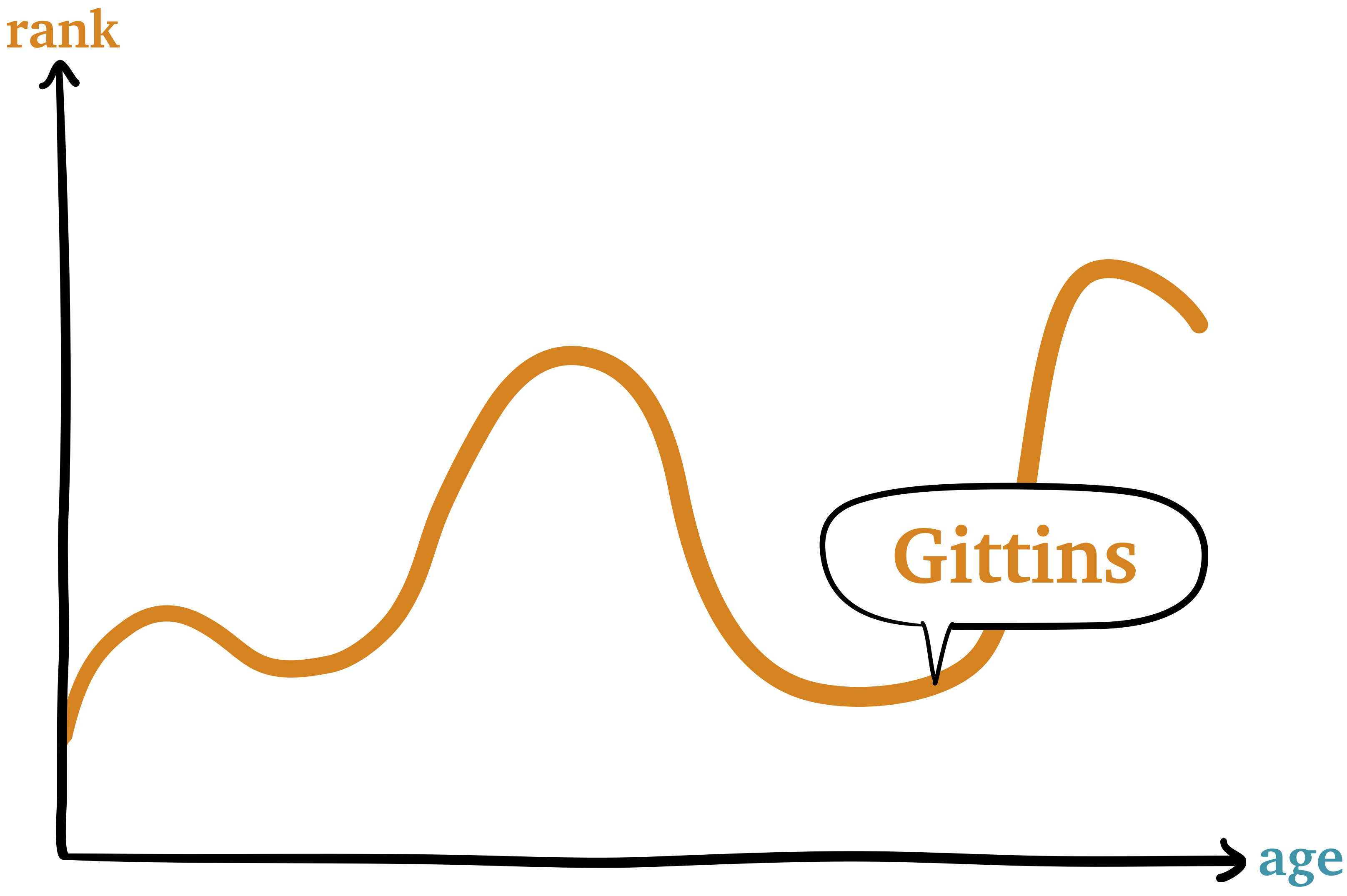
New Policy: **M-Gittins**

New Policy: **M-Gittins**

monotonic

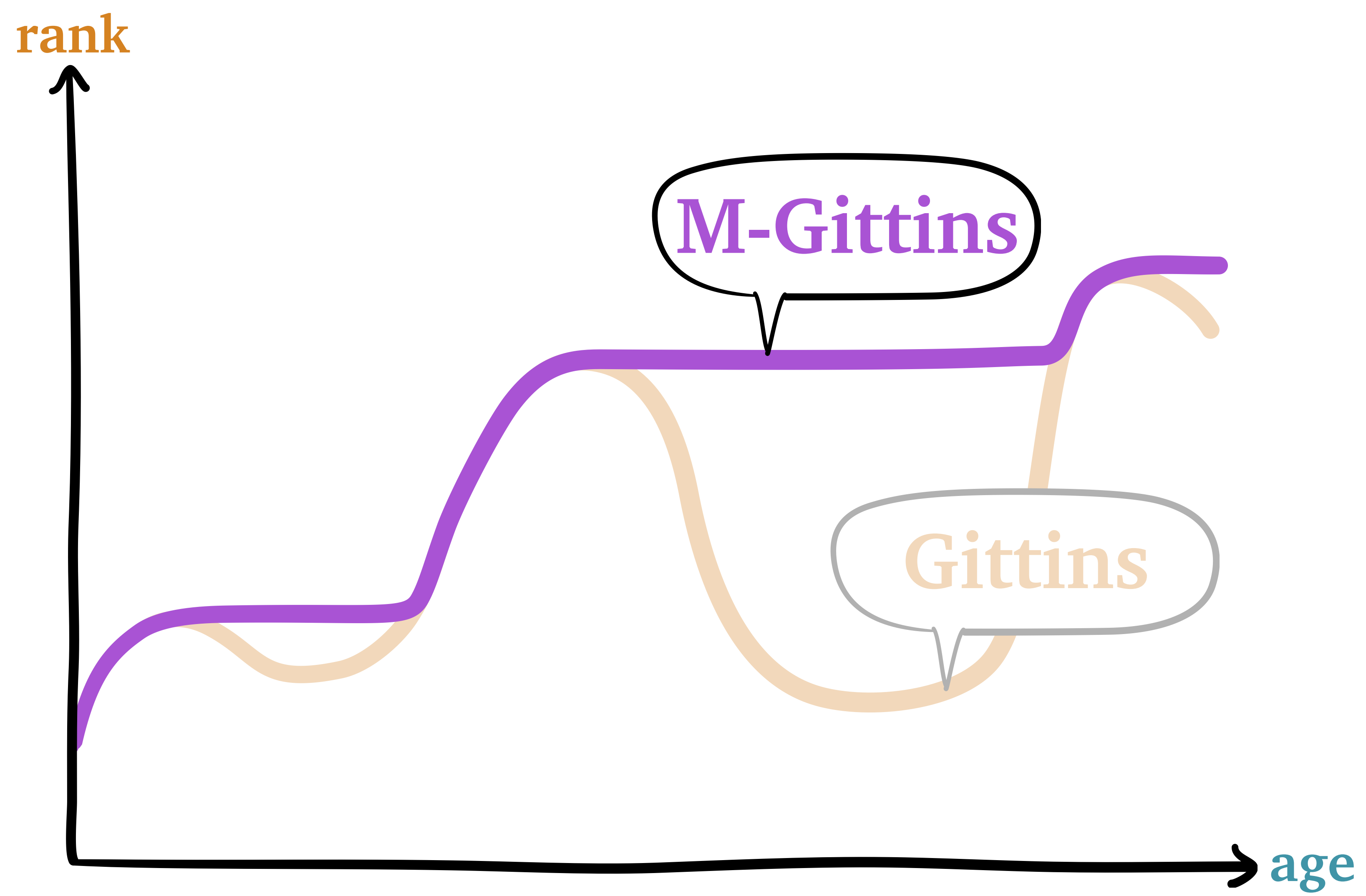
New Policy: M-Gittins

monotonic



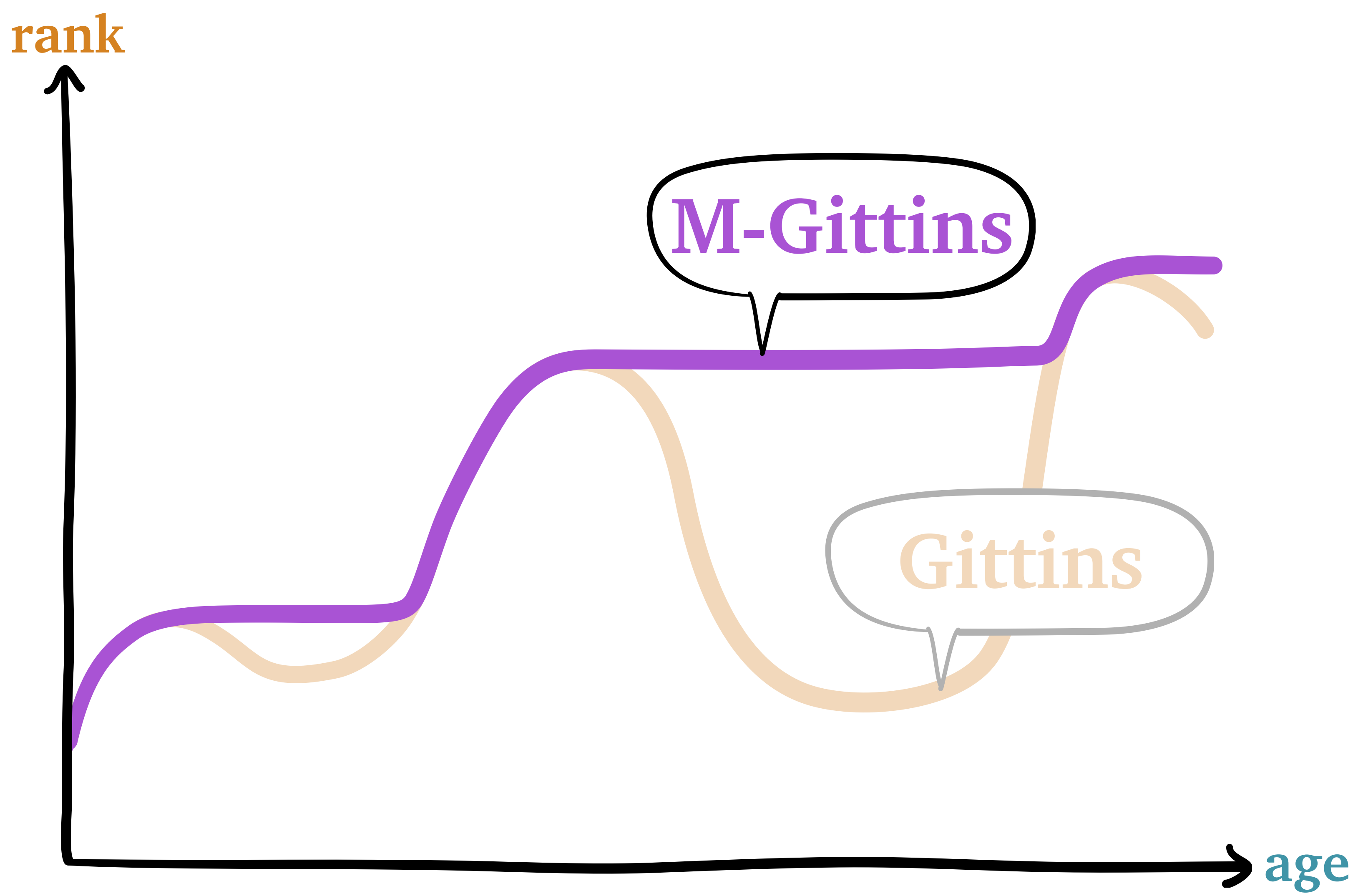
New Policy: M-Gittins

monotonic



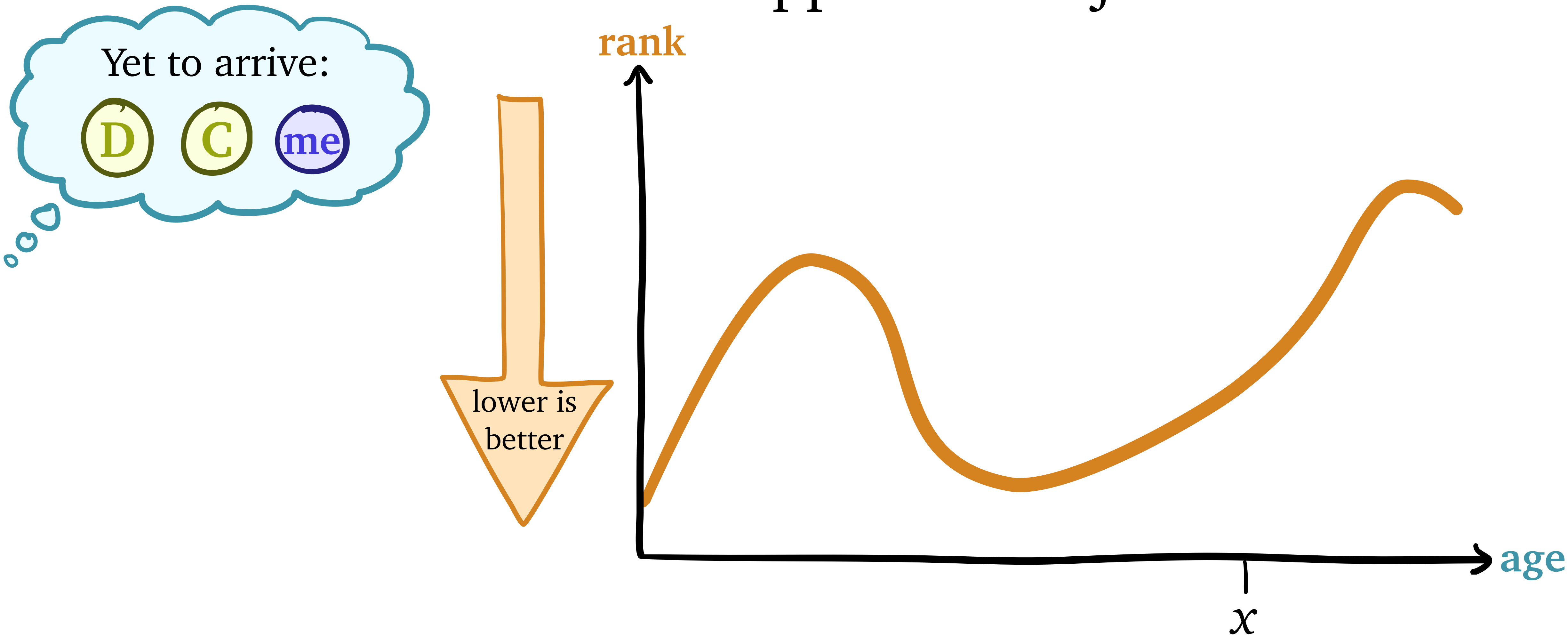
New Policy: M-Gittins

$$r_{\text{M-Gittins}}(a) = \max_{0 \le b \le a} r_{\text{Gittins}}(b)$$



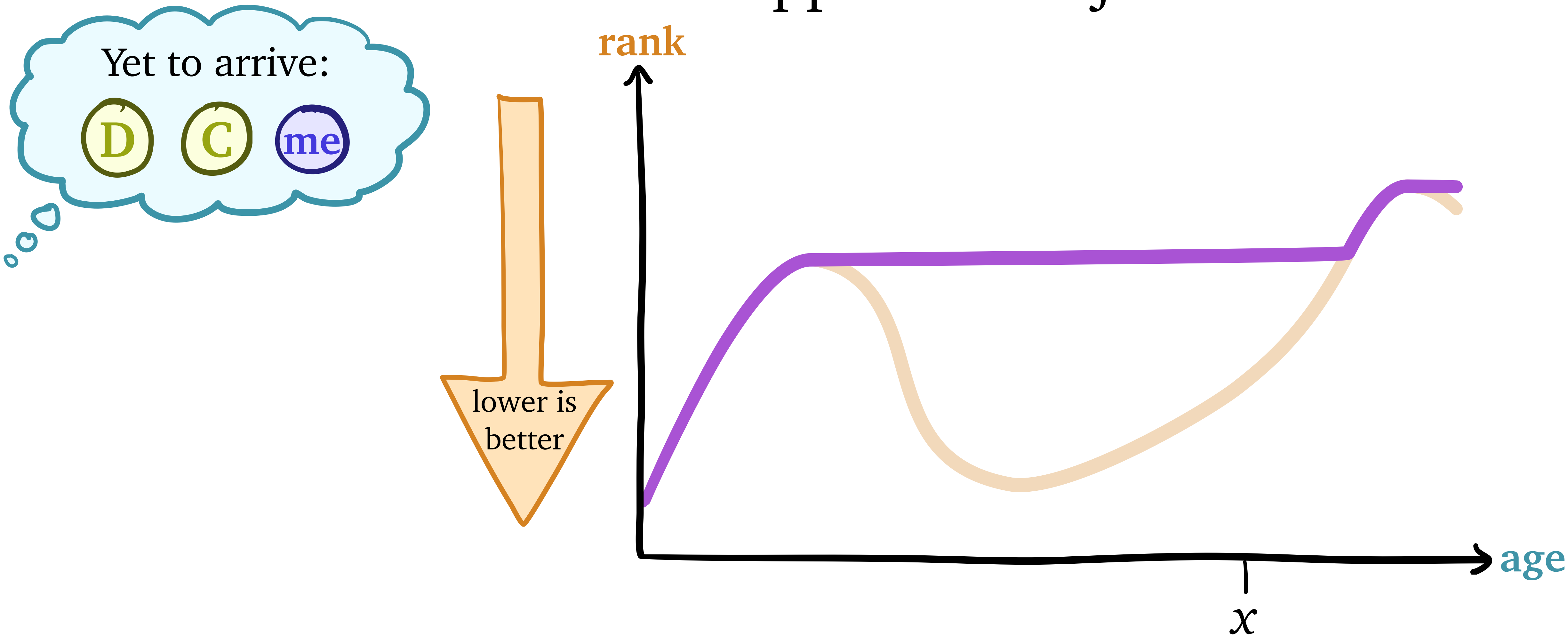
M-Gittins-*k* Saves the Day

Suppose I'm a job of size x



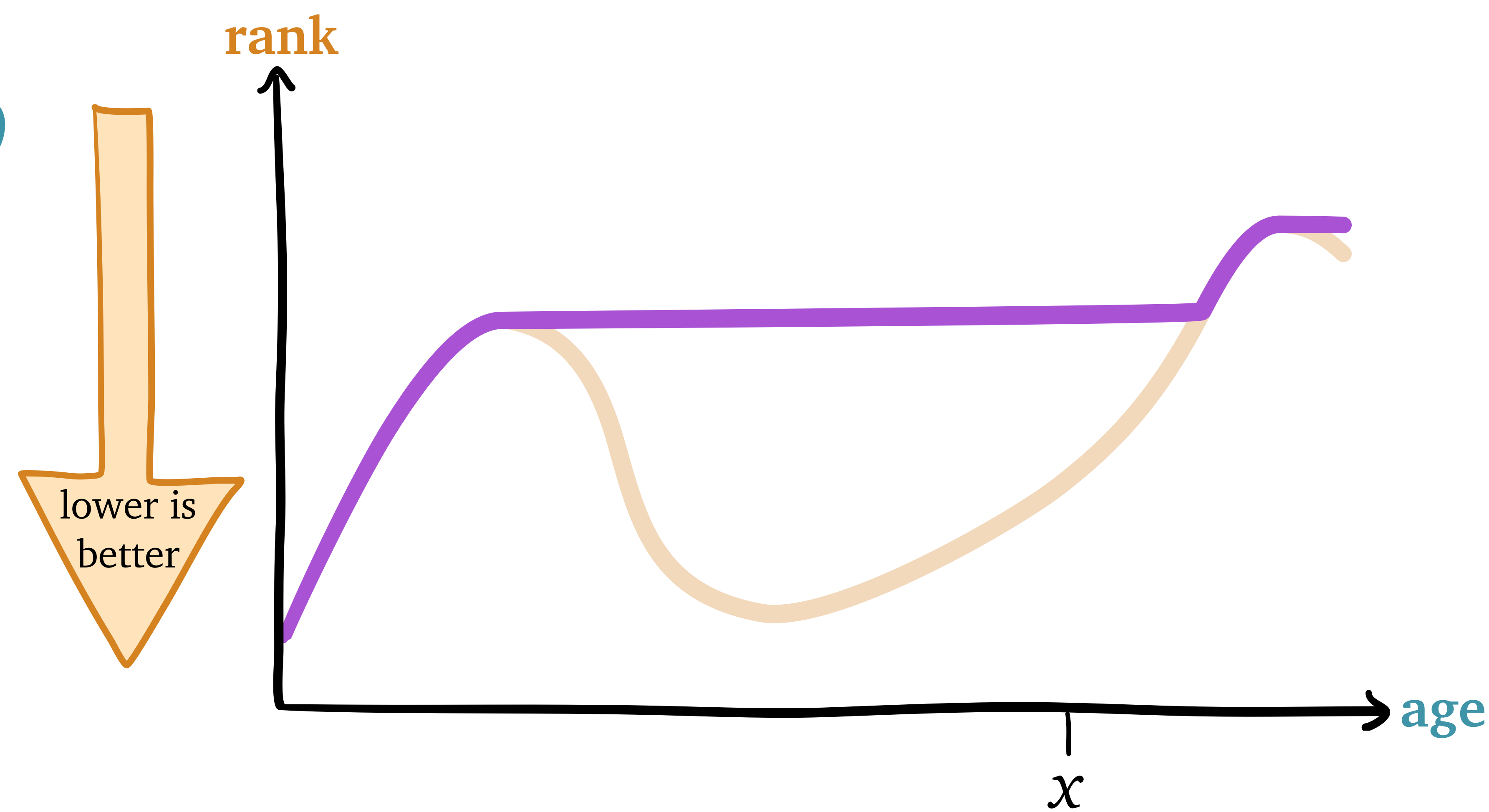
M-Gittins-*k* Saves the Day

Suppose I'm a job of size x



M-Gittins-*k* Saves the Day

Suppose I'm a job of size x

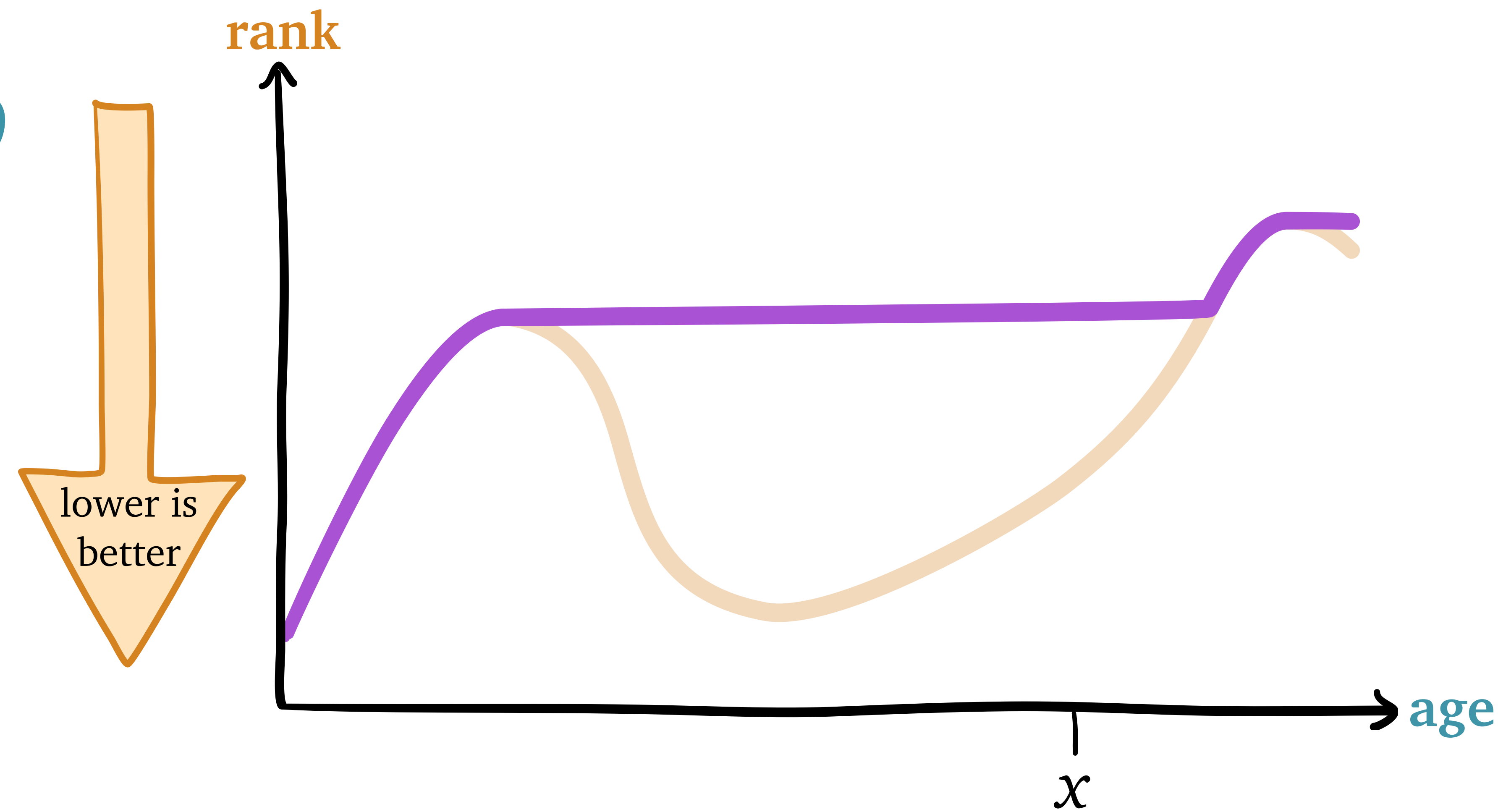


I can ignore { *old* jobs / *new* jobs (C & D) } after age { ??? / ??? }

M-Gittins-k Saves the Day

$k = 2$

ppose I’m a job of size x

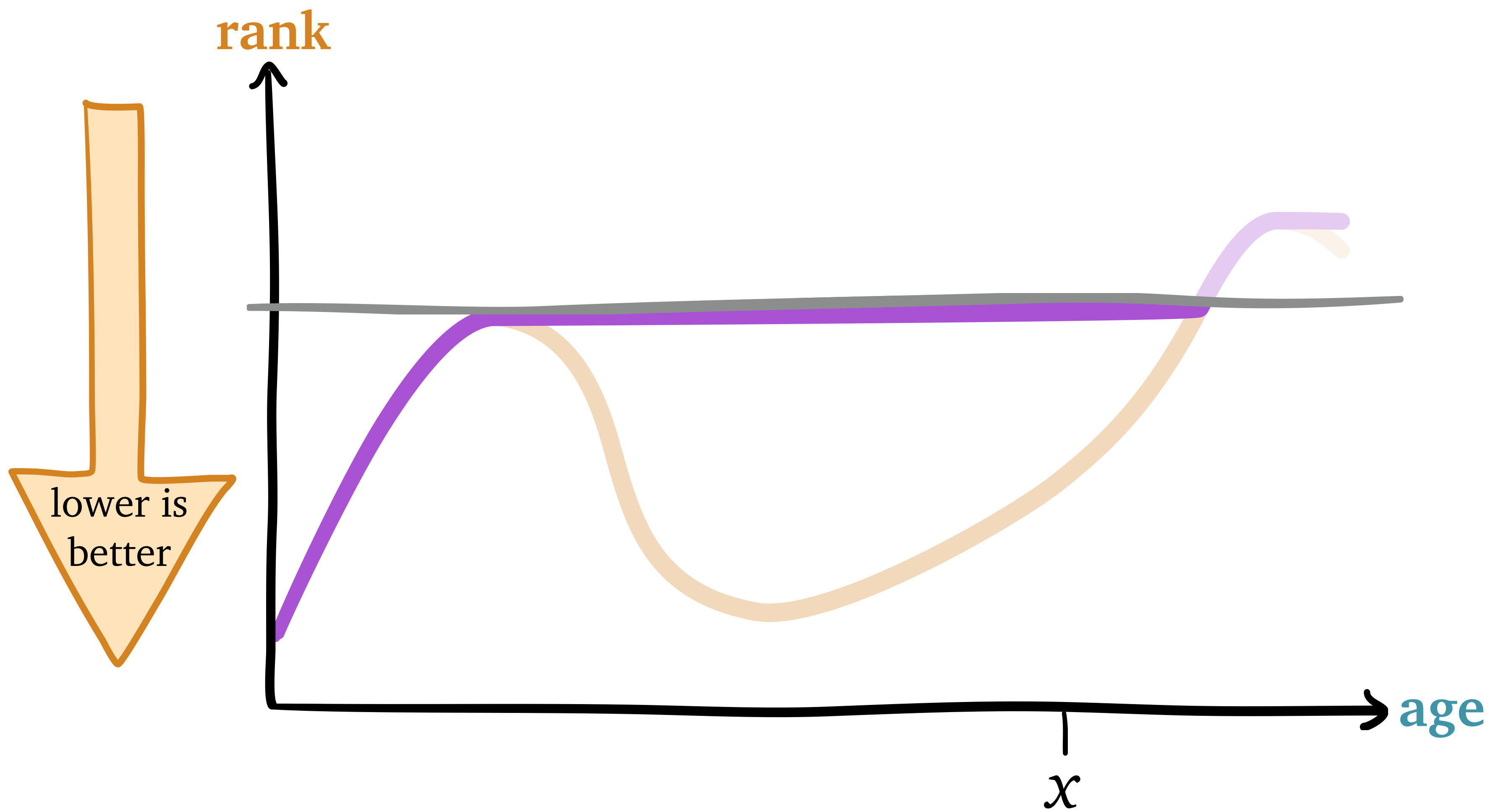


I can ignore { *old* jobs / *new* jobs (C & D) } after age { ??? / ??? }

M-Gittins-k Saves the Day

$k = 2$

ppose I'm a job of size x

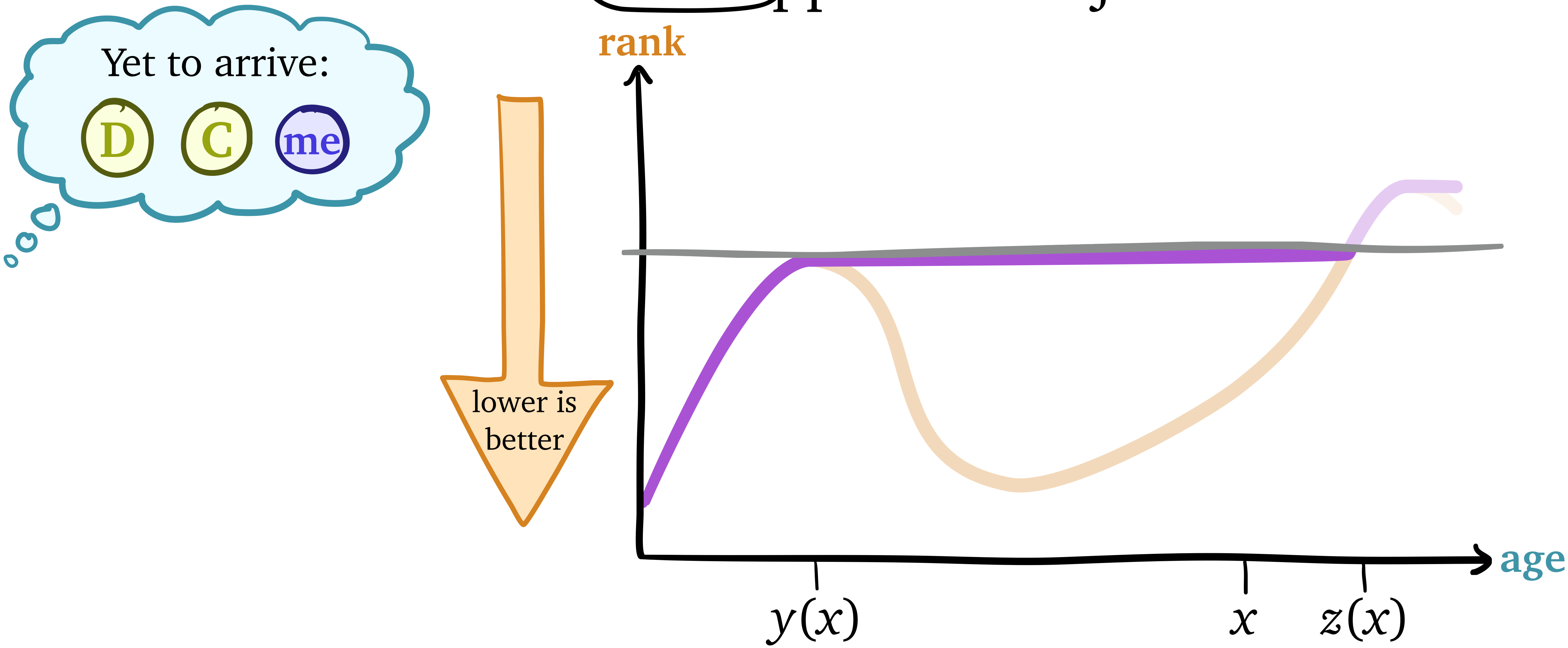


I can ignore { *old* jobs / *new* jobs (C & D) } after age { ??? / ??? }

M-Gittins-*k* Saves the Day

$k = 2$

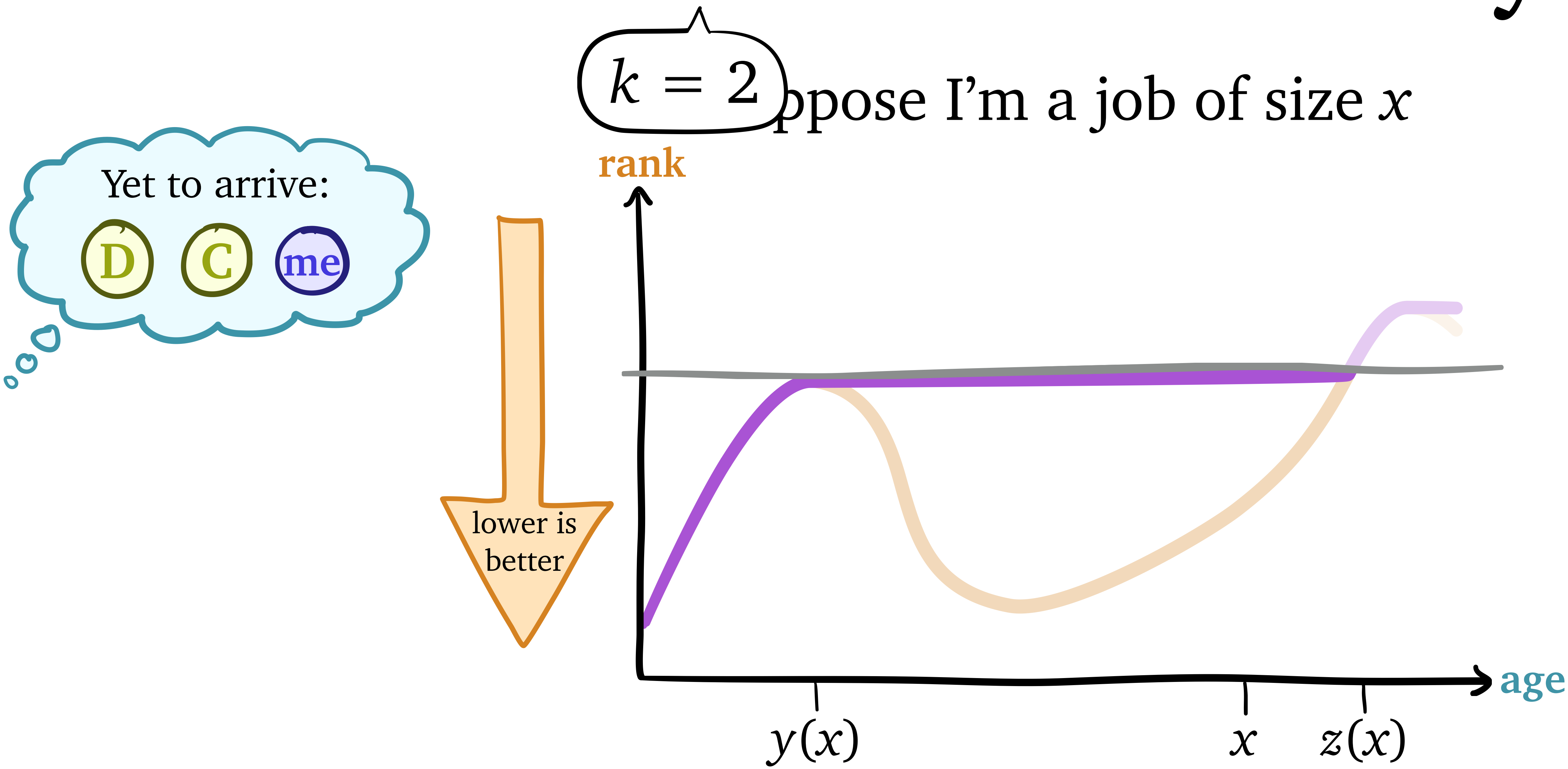
ppose I'm a job of size x



I can ignore { *old* jobs / *new* jobs (C & D) } after age { ??? / ??? }

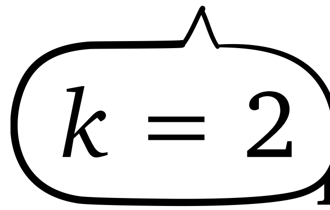
M-Gittins-*k* Saves the Day

ppose I'm a job of size x

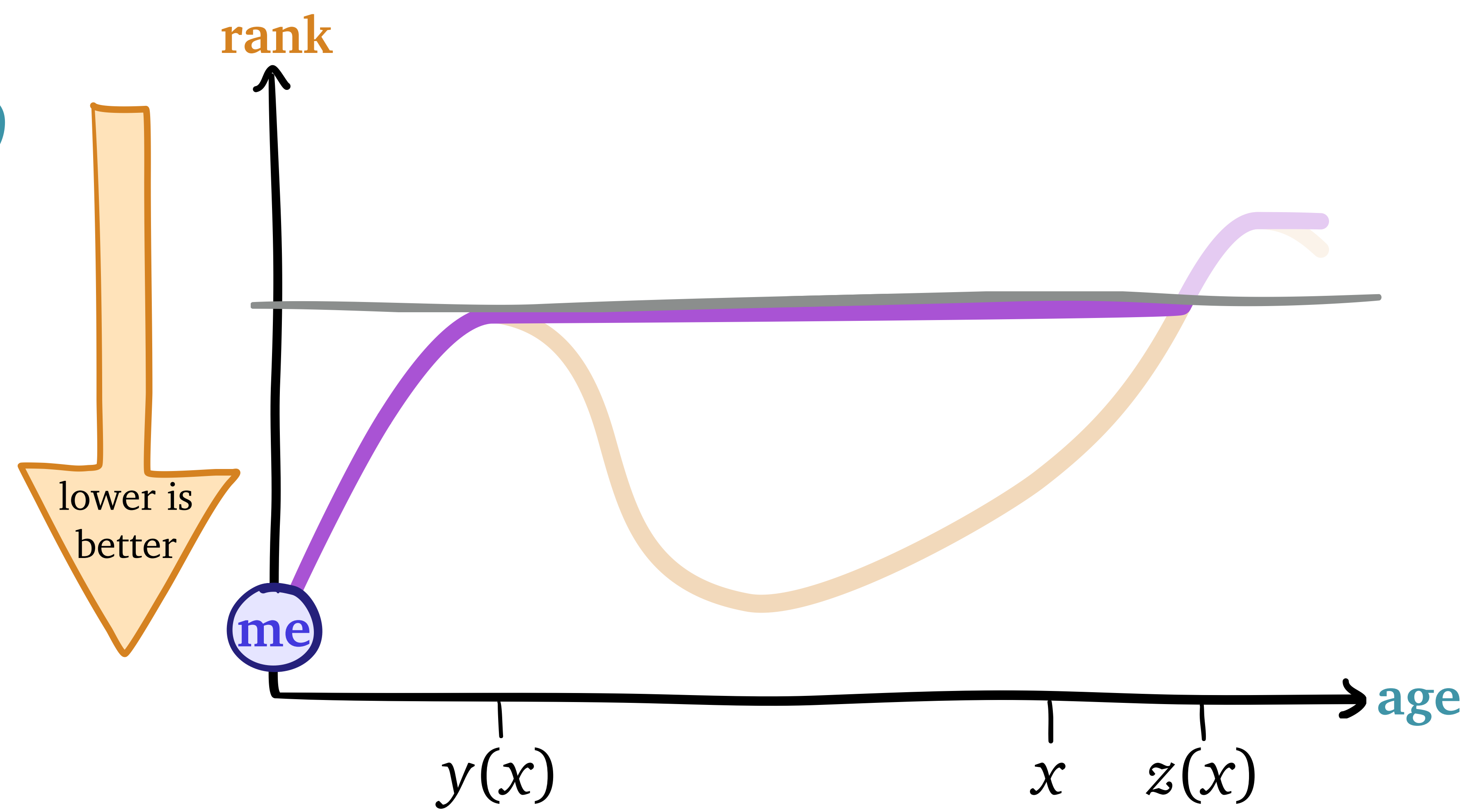


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-k Saves the Day



ppose I'm a job of size x

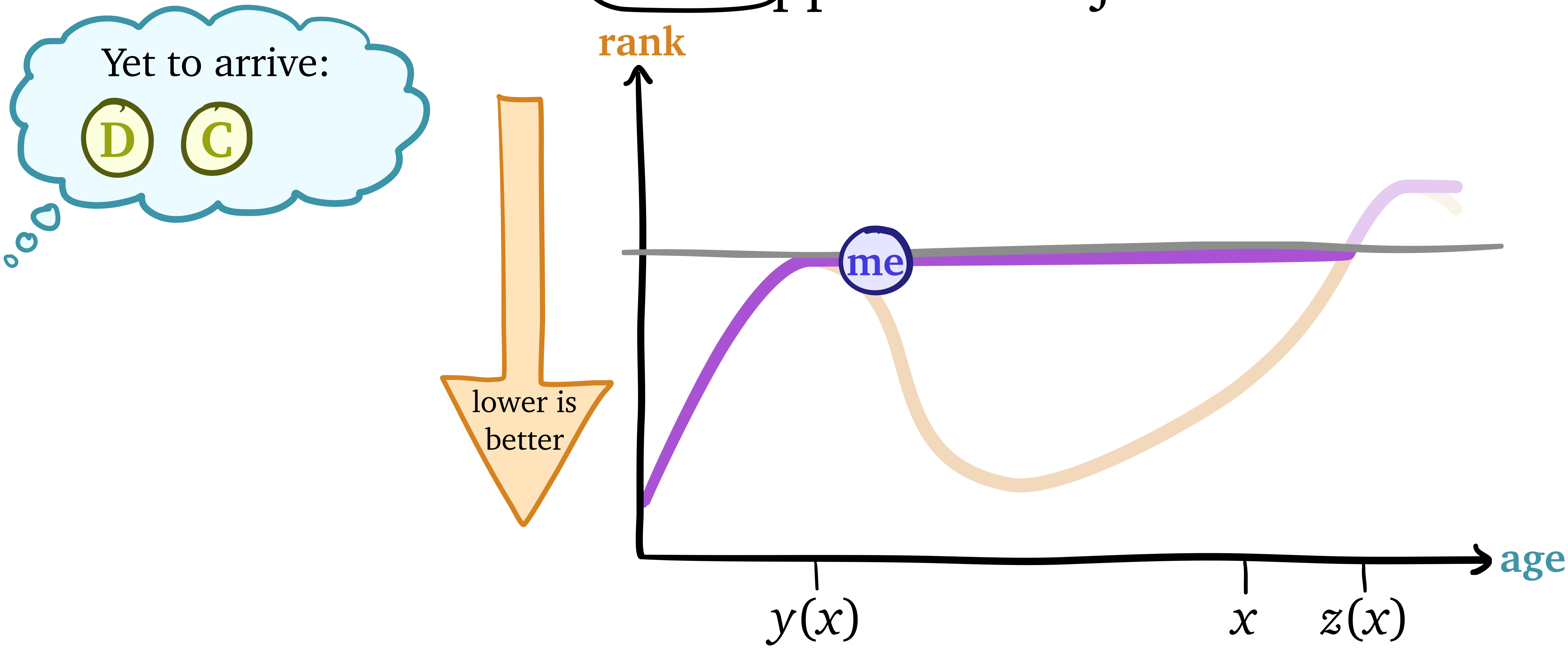


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-k Saves the Day

$k = 2$

ppose I'm a job of size x

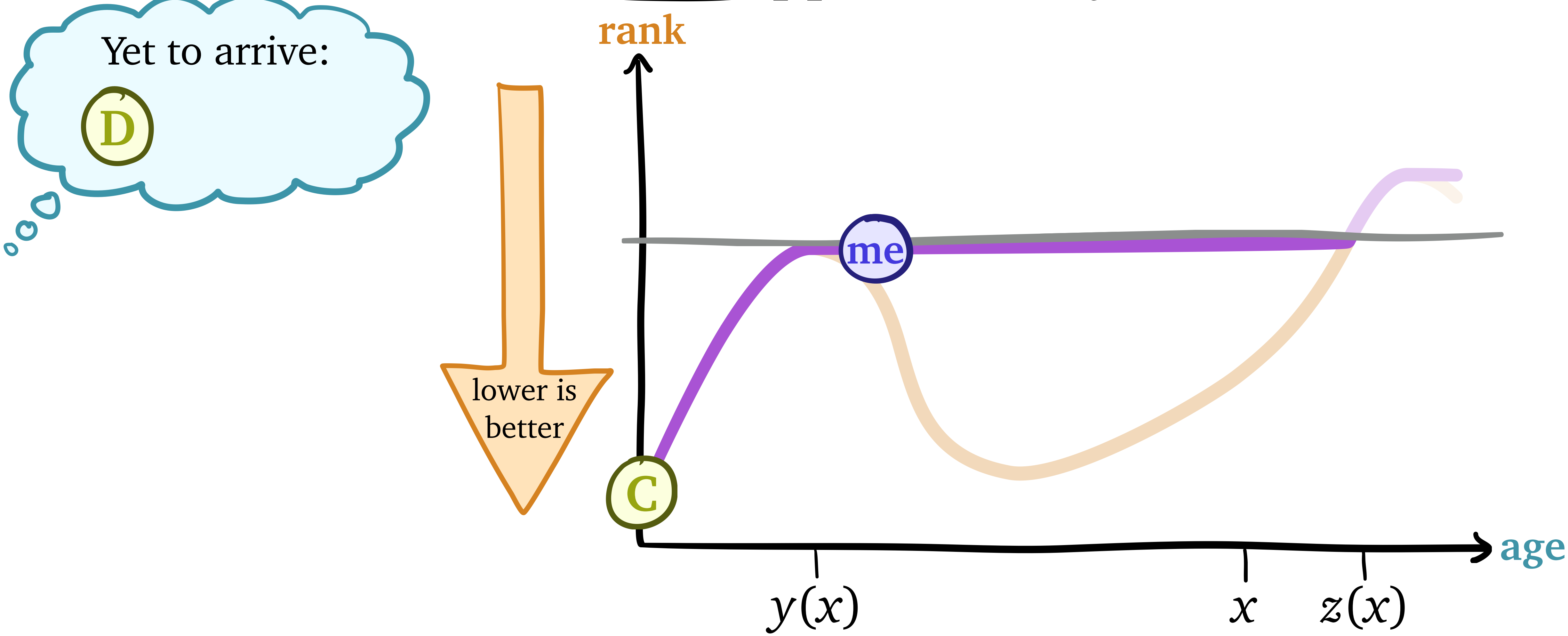


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-*k* Saves the Day

$k = 2$

ppose I’m a job of size x

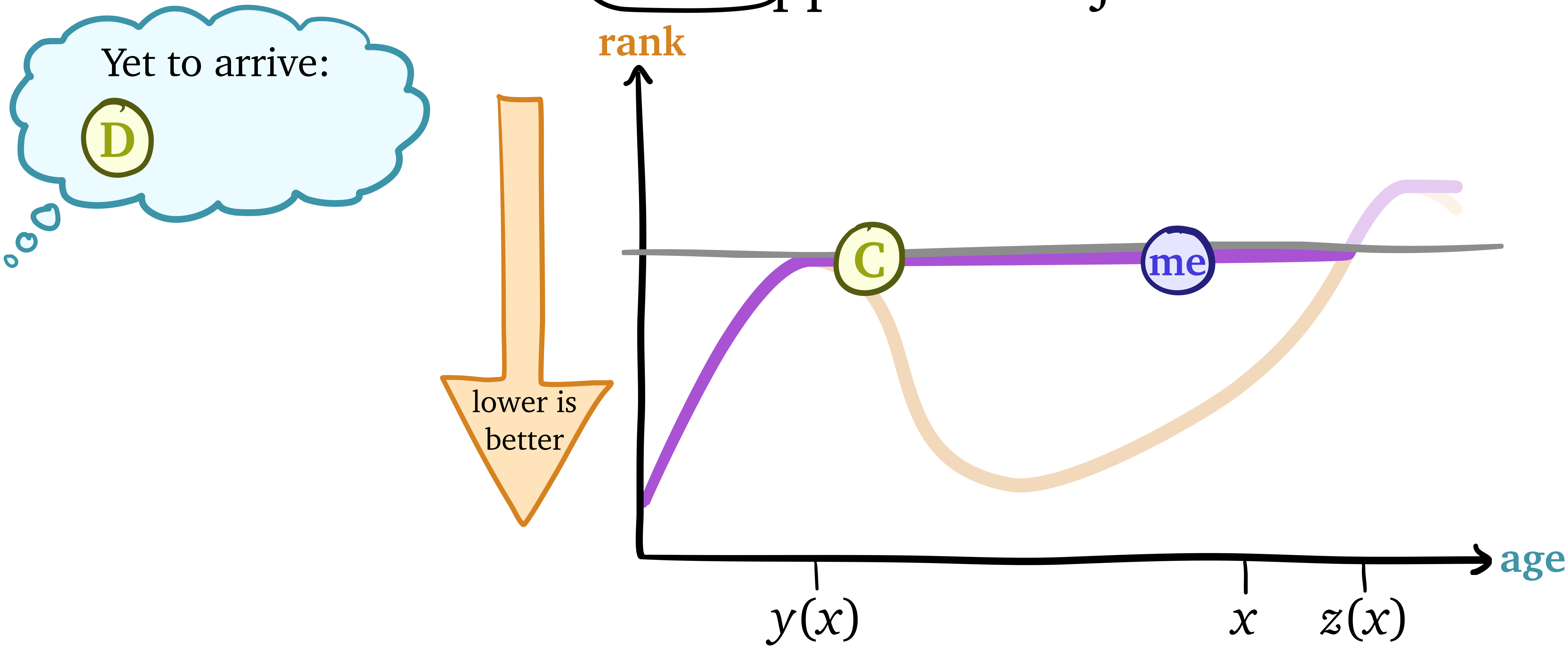


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-k Saves the Day

$k = 2$

…ppose I'm a job of size x

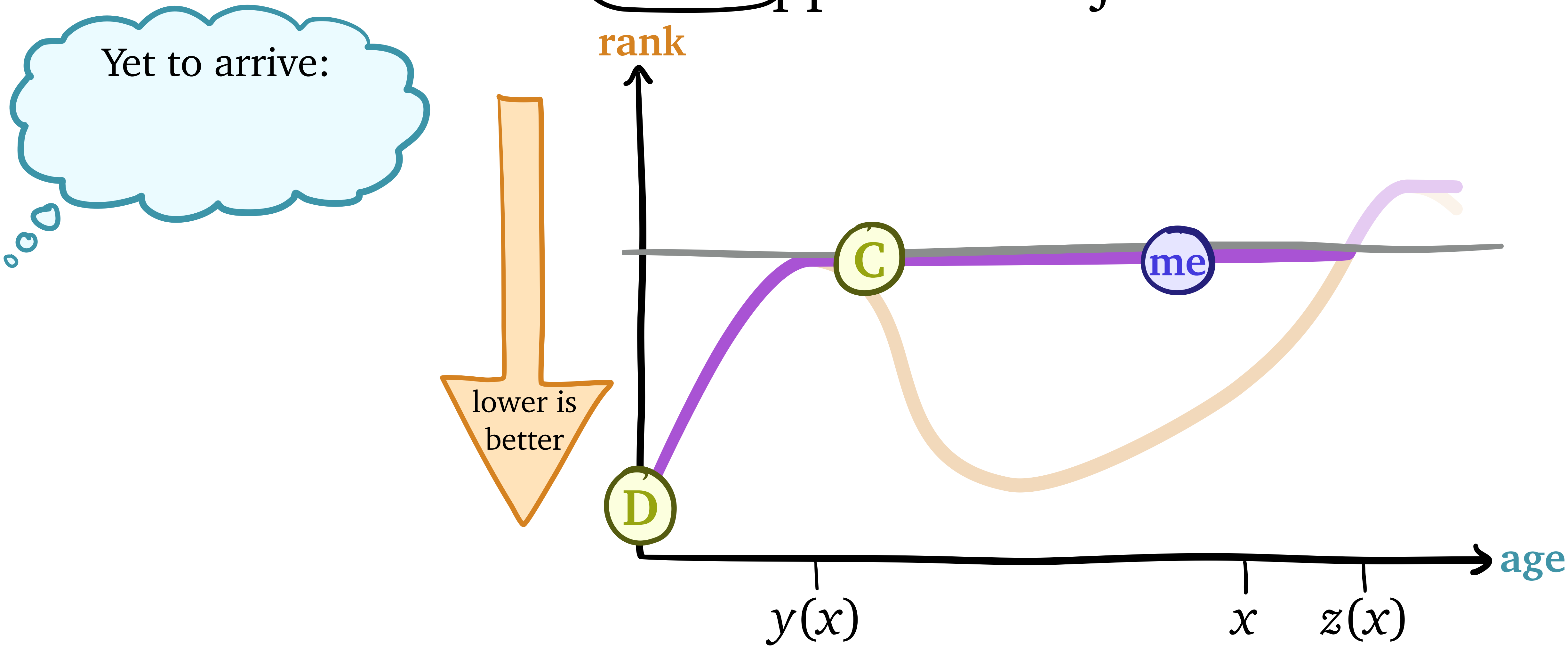


I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-*k* Saves the Day

$k = 2$

ppose I'm a job of size x

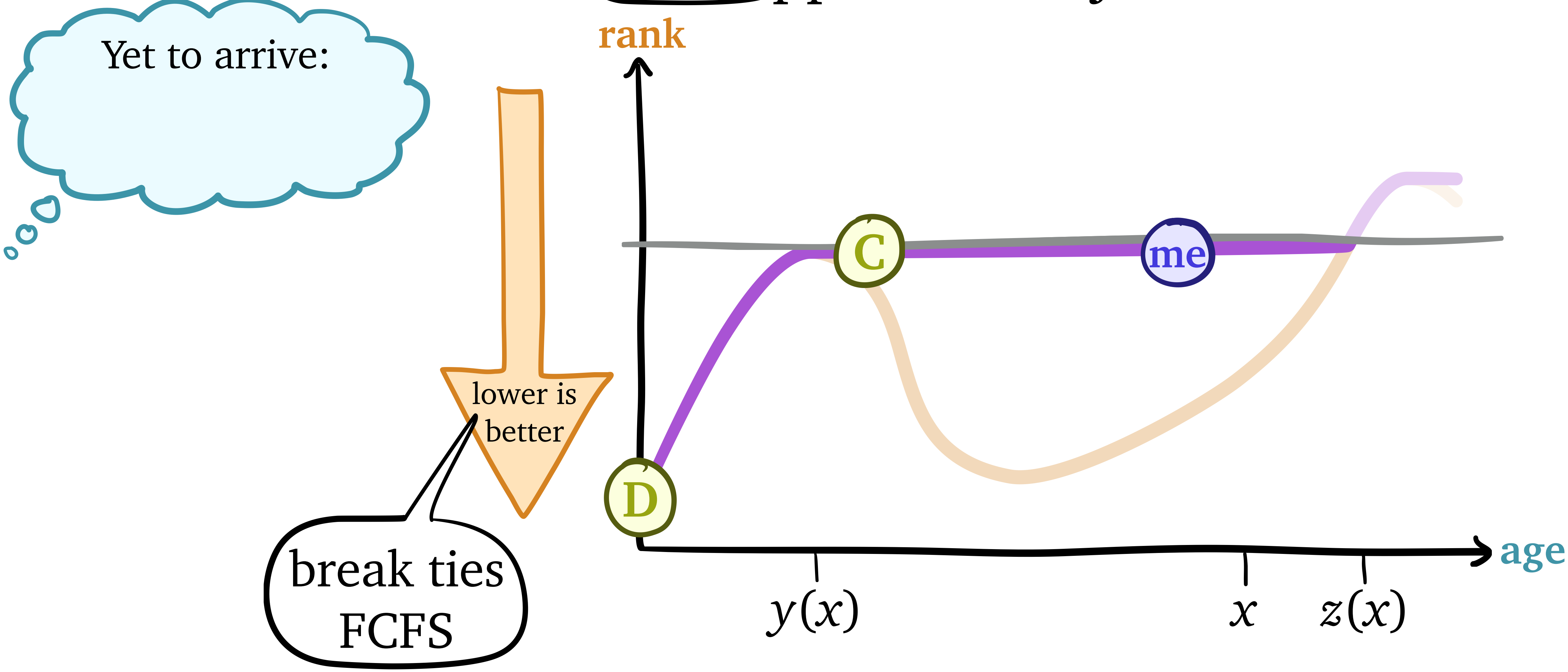


I can ignore { *old* jobs after age $z(x)$; *new* jobs (C & D) after age ??? }

M-Gittins-*k* Saves the Day

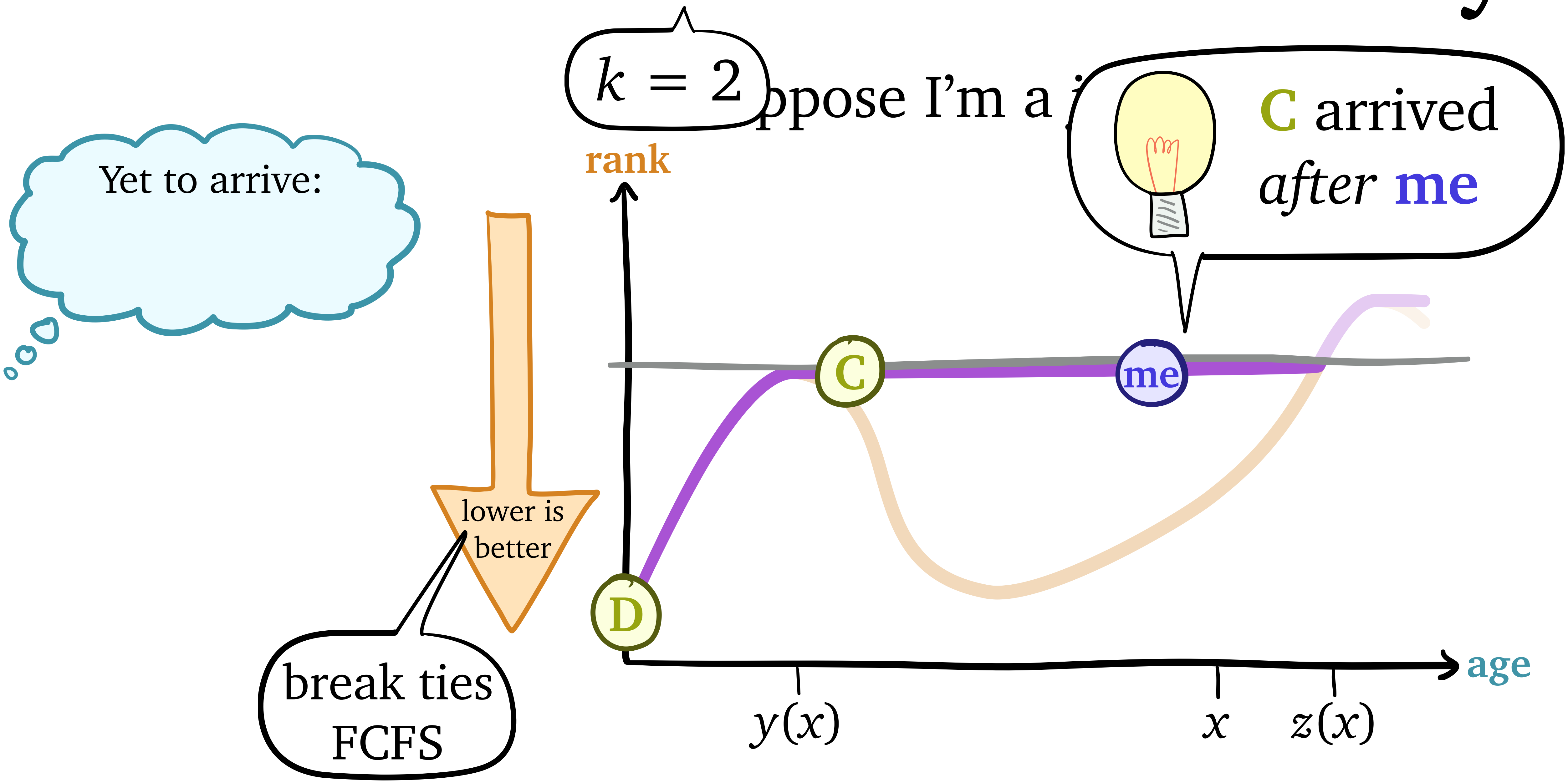
$k = 2$

ppose I'm a job of size x



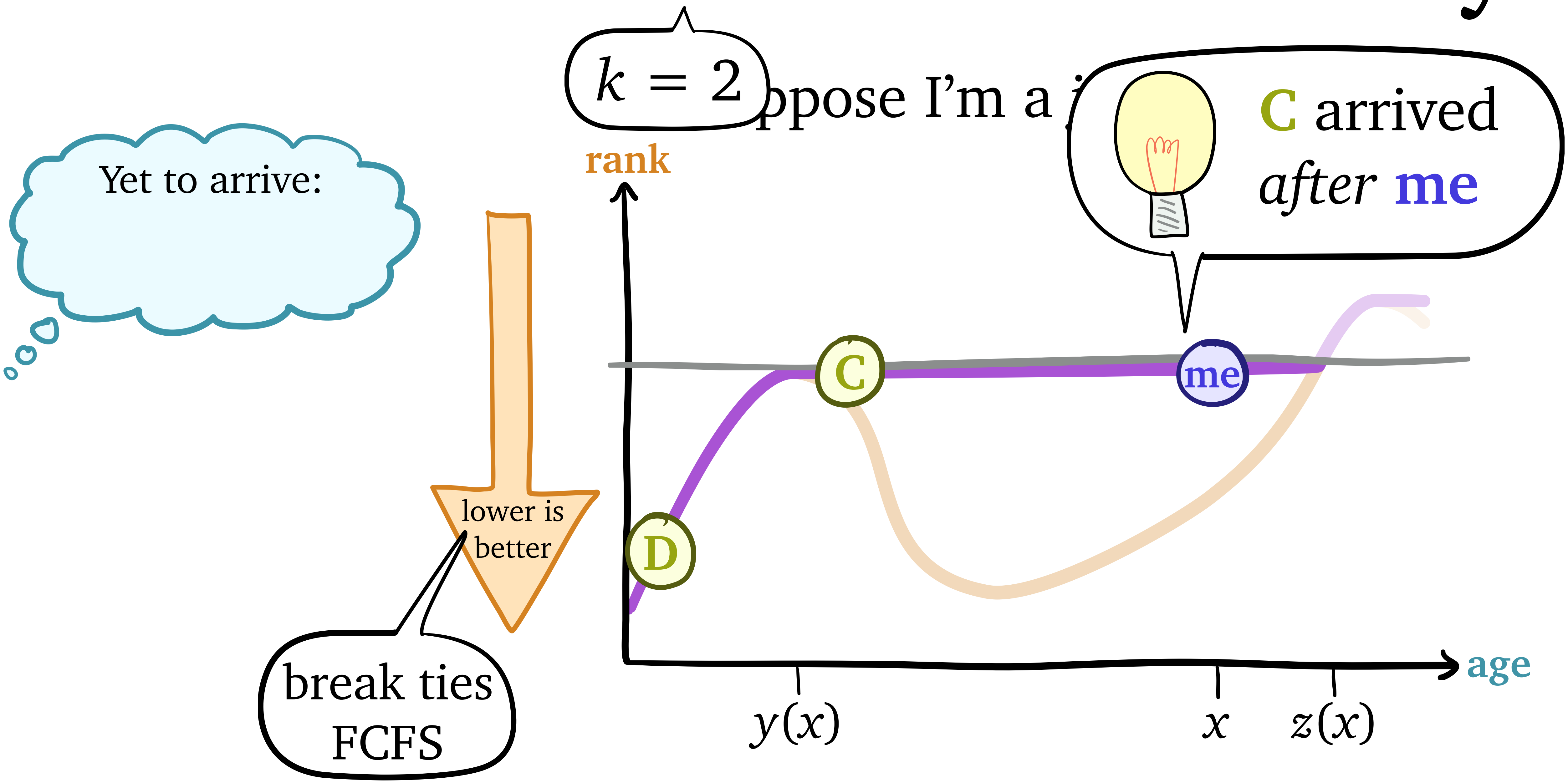
I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-k Saves the Day



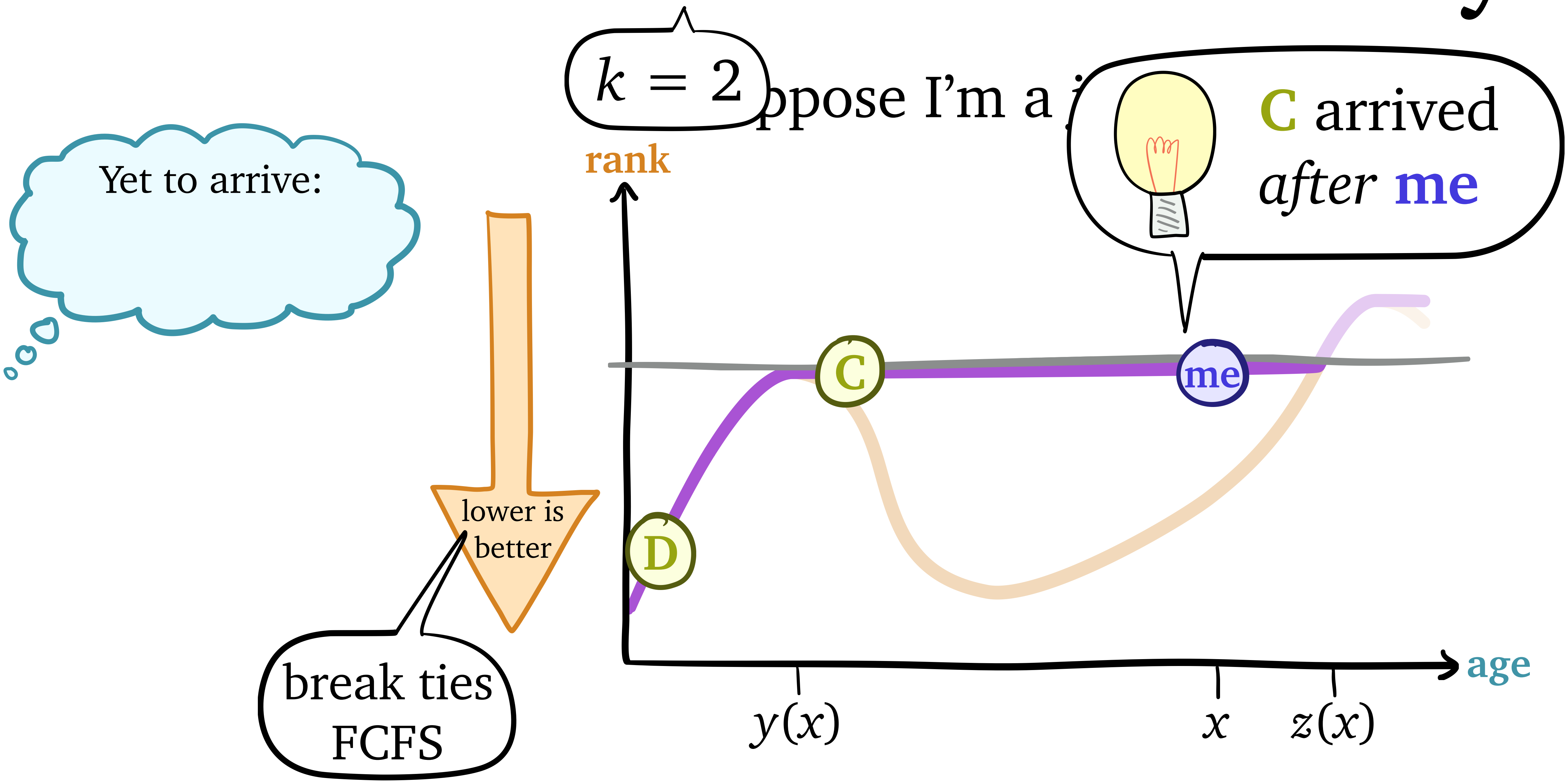
I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-*k* Saves the Day



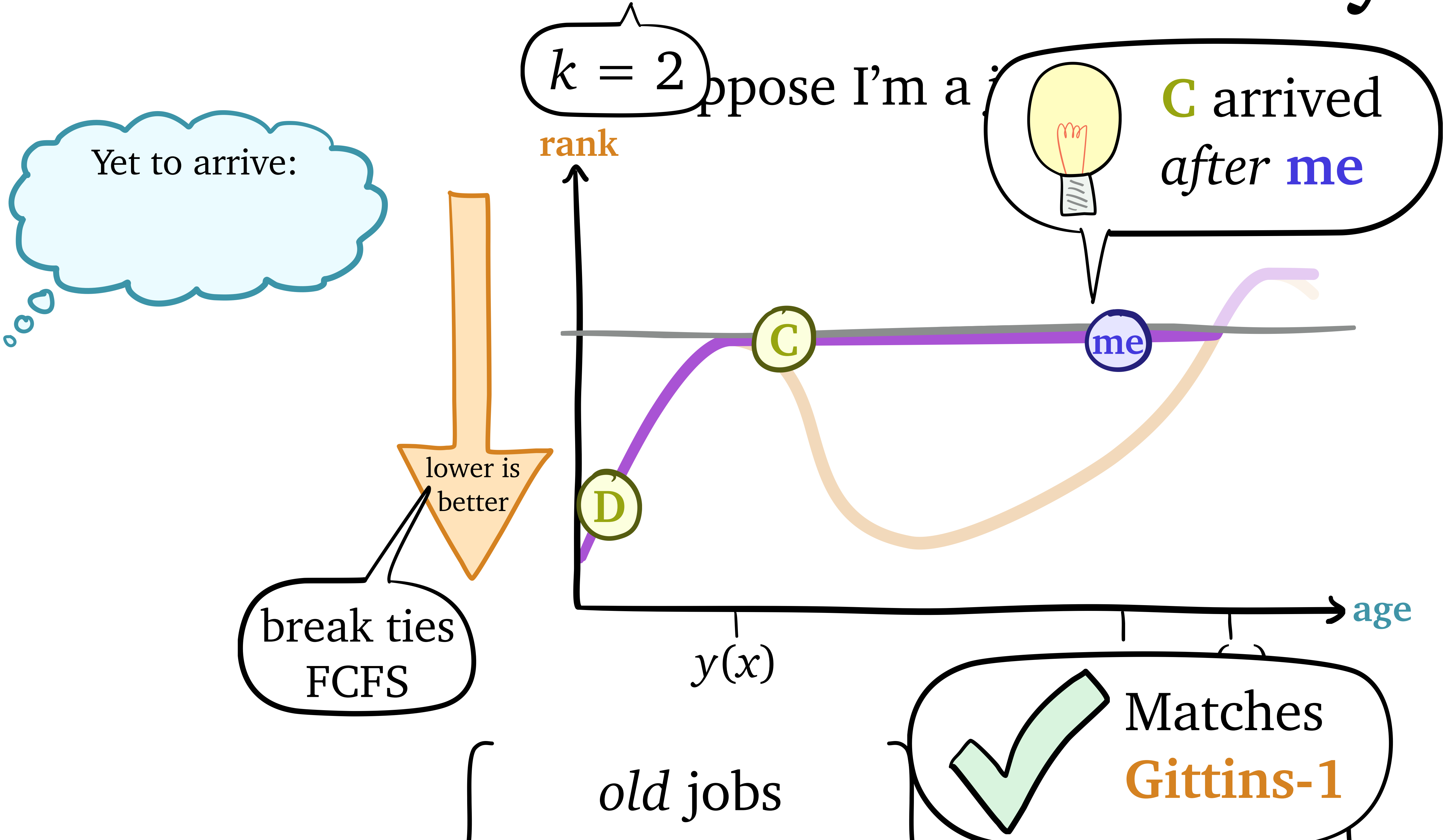
I can ignore { *old* jobs / *new* jobs (C & D) } after age { $z(x)$ / ??? }

M-Gittins-k Saves the Day



I can ignore { *old* jobs after age $z(x)$; *new* jobs (C & D) after age $y(x)$ }

M-Gittins-*k* Saves the Day



I can ignore { *old* jobs / *new* jobs (C & D) } after age { $y(x)$ }

Main Results

Suppose X is heavy-tailed with finite variance

Main Results

Suppose X is heavy-tailed with finite variance

similar results for some light-tailed X (see paper)

Main Results

Suppose X is heavy-tailed with finite variance

similar results for some light-tailed X (see paper)

Step 1: link **M-Gittins-*k*** to **Gittins-1**

Step 2: analyze heavy-traffic **Gittins-1**

Main Results

Suppose X is heavy-tailed with finite variance

similar results for some light-tailed X (see paper)

Step 1: link **M-Gittins-k** to **Gittins-1**

$$\mathbf{E}[T_{\text{M-Gittins-}k}] \le \mathbf{E}[T_{\text{Gittins-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

Step 2: analyze heavy-traffic **Gittins-1**

Main Results

Suppose X is heavy-tailed with finite variance

similar results for some light-tailed X (see paper)

Step 1: link **M-Gittins-k** to **Gittins-1**

$$\mathbf{E}[T_{\text{M-Gittins-}k}] \le \mathbf{E}[T_{\text{Gittins-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

Step 2: analyze heavy-traffic **Gittins-1**

$$\mathbf{E}[T_{\text{Gittins-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

Main Results

Suppose X is heavy-tailed with finite variance

similar results for some light-tailed X (see paper)

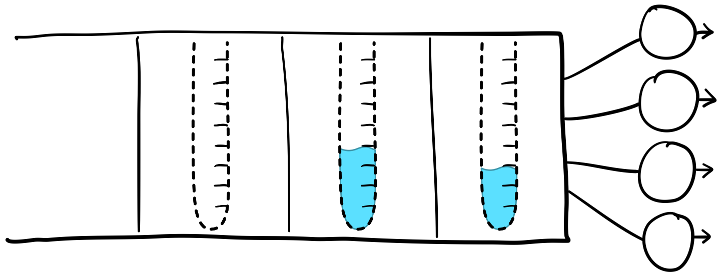
Step 1: link **M-Gittins-k** to **Gittins-1**

$$\mathbf{E}[T_{\text{M-Gittins-}k}] \le \mathbf{E}[T_{\text{Gittins-1}}] + k \cdot O\left(\log \frac{1}{1-\rho}\right)$$

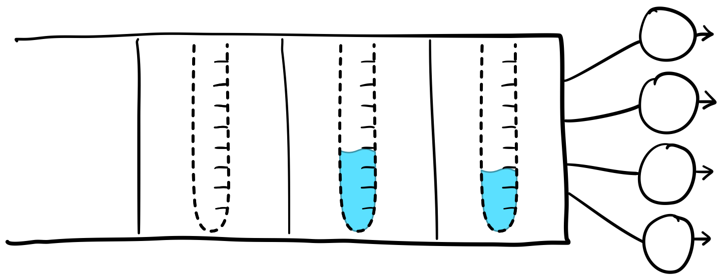
Step 2: analyze heavy-traffic **Gittins-1**

$$\mathbf{E}[T_{\text{Gittins-1}}] = \omega\left(\log \frac{1}{1-\rho}\right)$$

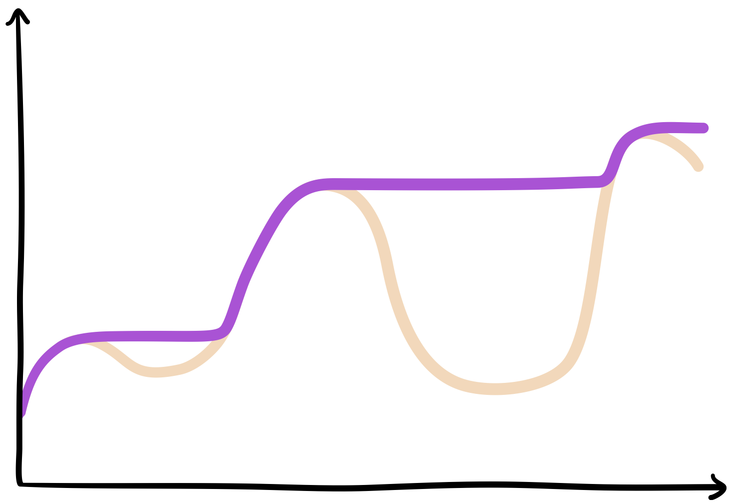
$$\Theta\left(\frac{1}{1-\rho} \middle/ \max_{0 \le b \le \bar{F}_e^{-1}(1-\rho)} \mathbf{E}[X - b \mid X > b]\right)$$



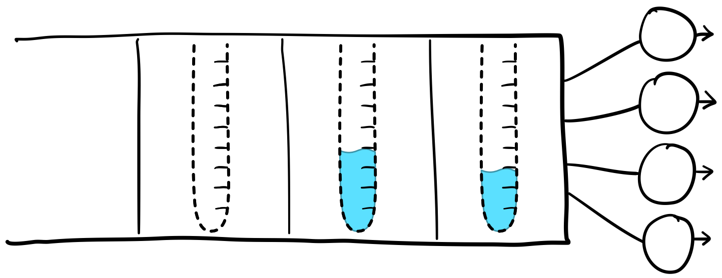
Goal: minimize $\mathbf{E}[T]$ in M/G/k with unknown job sizes



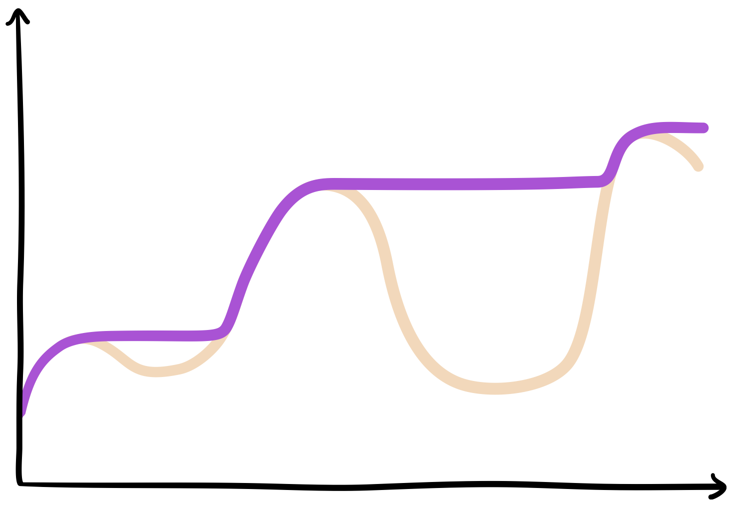
Goal: minimize $\mathbf{E}[T]$ in M/G/k with unknown job sizes



Key idea: new *monotonic* variant of **Gittins**, namely **M-Gittins**

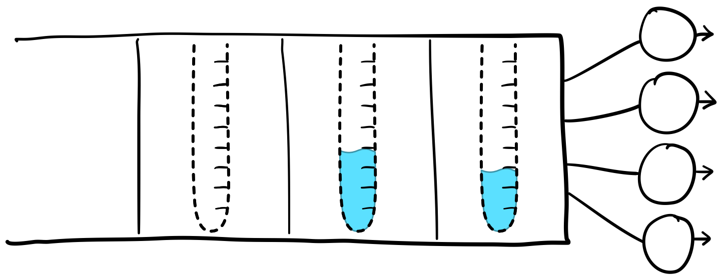


Goal: minimize $\mathbf{E}[T]$ in M/G/k with unknown job sizes

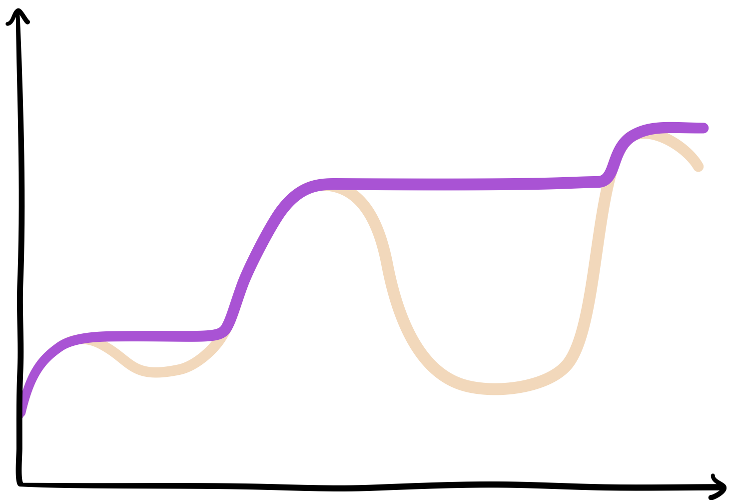


Key idea: new *monotonic* variant of **Gittins**, namely **M-Gittins**

Theorem: $\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1$



Goal: minimize $\mathbf{E}[T]$ in M/G/k with unknown job sizes



Key idea: new *monotonic* variant of **Gittins**, namely **M-Gittins**

Theorem: $\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1$

Get in touch: `zscully@cs.cmu.edu`

Bonus Slides

Heavy-Traffic Optimality

Heavy-Traffic Optimality

Theorem:
M-Gittins-*k* is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\textbf{M-Gittins-}k}]}{\mathbf{E}[T_{\textbf{Gittins-1}}]} = 1,$$

Heavy-Traffic Optimality

Theorem:
M-Gittins-*k* is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1,$$

if X is in *any* of the following classes:

Heavy-Traffic Optimality

Theorem:
M-Gittins-k is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1,$$

if X is in *any* of the following classes:

- bounded

Heavy-Traffic Optimality

Theorem:
M-Gittins-k is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1,$$

if X is in *any* of the following classes:

- bounded
- "finite-variance heavy-tailed"
  (O-regularly varying with Matuszewska indices less than –2)

Heavy-Traffic Optimality

Theorem:
M-Gittins-k is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\text{M-Gittins-}k}]}{\mathbf{E}[T_{\text{Gittins-1}}]} = 1,$$

if X is in *any* of the following classes:

- bounded
- "finite-variance heavy-tailed"
  (*O*-regularly varying with Matuszewska indices less than –2)
- MDA(Λ) with "quasi-decreasing hazard rate", e.g. $h(x) = \Theta(x^{-\gamma})$

Heavy-Traffic Optimality

Theorem:
M-Gittins-*k* is *heavy-traffic optimal* in the M/G/k, specifically

$$\lim_{\rho \to 1} \frac{\mathbf{E}[T_{\textbf{M-Gittins-}k}]}{\mathbf{E}[T_{\textbf{Gittins-1}}]} = 1,$$

if X is in *any* of the following classes:

- bounded
- "finite-variance heavy-tailed"
  (*O*-regularly varying with Matuszewska indices less than –2)
- MDA(Λ) with "quasi-decreasing hazard rate", e.g. $h(x) = \Theta(x^{-\gamma})$

exponential, log-normal, Weibull…

M/G/1 Heavy-Traffic Scaling

M/G/1 Heavy-Traffic Scaling

Theorem:

"infinite variance" {

"finite variance" {

M/G/1 Heavy-Traffic Scaling

Theorem:

"infinite variance"
If $X \in \mathrm{OR}(-2, -1)$, then

$$\mathbf{E}[T_{\text{Gittins-1}}] = \Theta\left(\log \frac{1}{1-\rho}\right),$$

"finite variance"

M/G/1 Heavy-Traffic Scaling

Theorem:

"infinite variance" {

If $X \in \mathrm{OR}(-2,-1)$, then

$$\mathbf{E}[T_{\text{Gittins-1}}] = \Theta\left(\log \frac{1}{1-\rho}\right),$$

}

"finite variance" {

and if $X \in \mathrm{OR}(-\infty,-2) \cup \mathrm{MDA}(\Lambda) \cup \mathrm{ENBUE}$, then

$$\mathbf{E}[T_{\text{Gittins-1}}] = \Theta\left(\frac{1}{1-\rho} \Bigg/ \max_{0 \le b \le \overline{F}_e^{-1}(1-\rho)} \mathbf{E}[X - b \mid X > b]\right).$$

}