Multi-armed Bandits

James Grant (he/him)

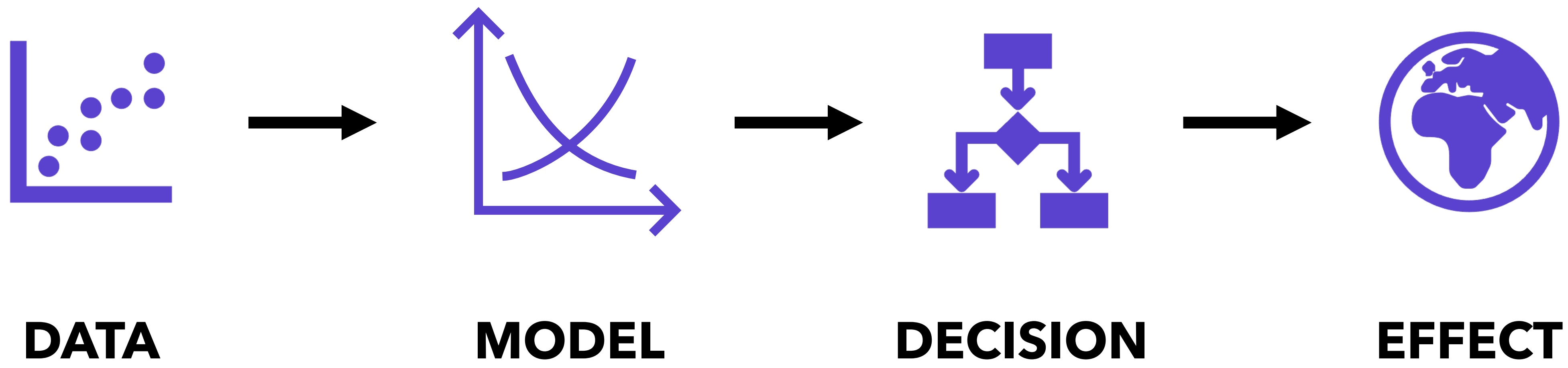
Lancaster University

j.grant@lancaster.ac.uk

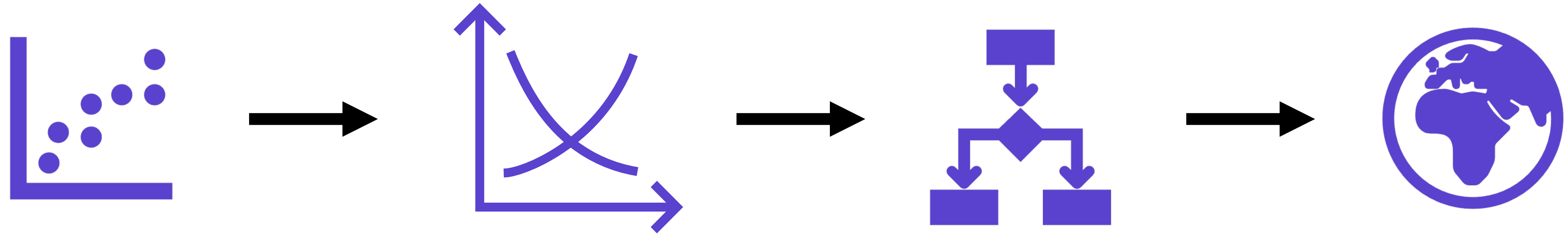
@james_a_grant

Tommy Flowers Network – Thursday 14th Oct

Using data to make decisions



Example: what to send customers?



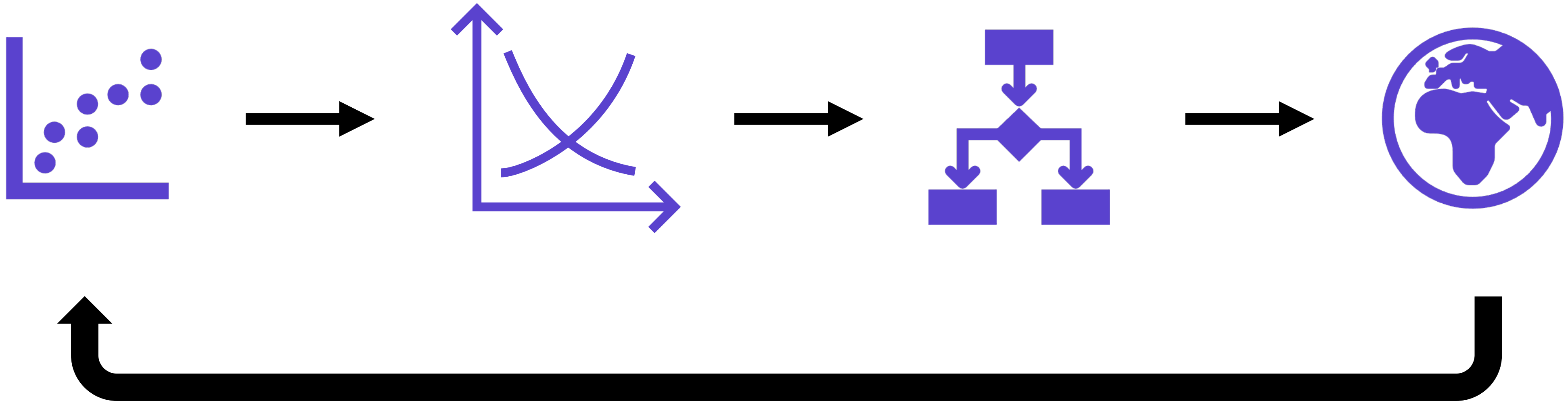
Test some different messaging strategies

Model customer response to strategies

Optimise for return, satisfaction etc.

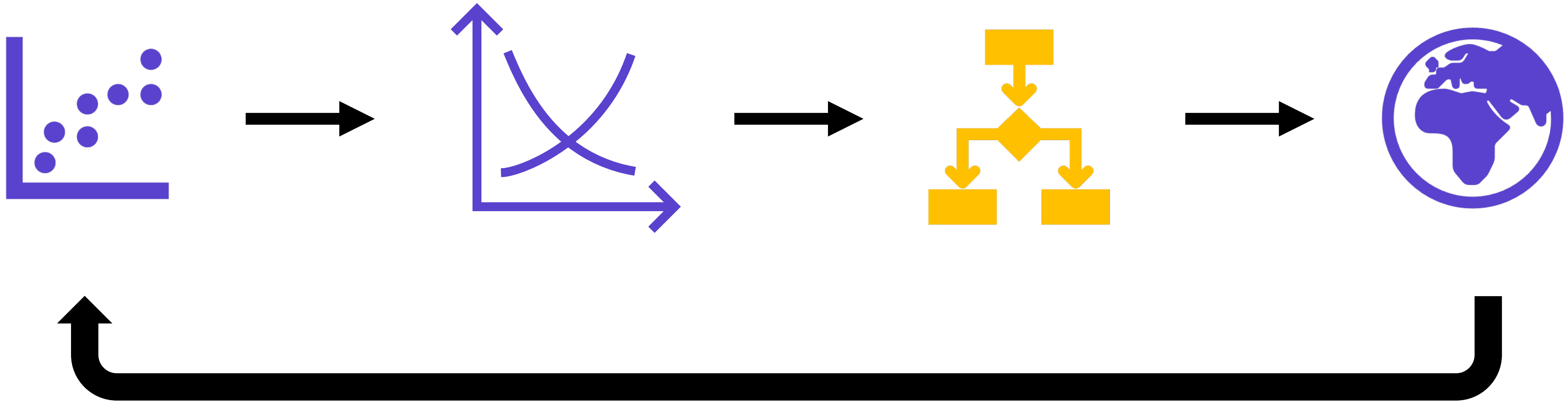
Hope that's a good choice?

We can observe effects, then iterate,



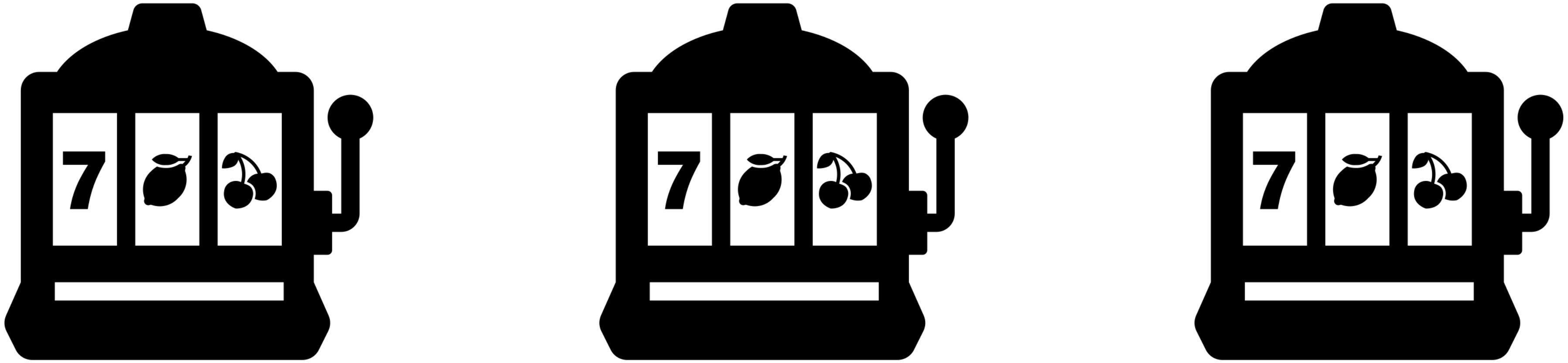
and converge to an optimal decision.

We can observe effects, then iterate,



and converge to an optimal decision, if we design our intermediate decisions wisely.

Multi-armed Bandit



Multi-armed Bandit

0.1

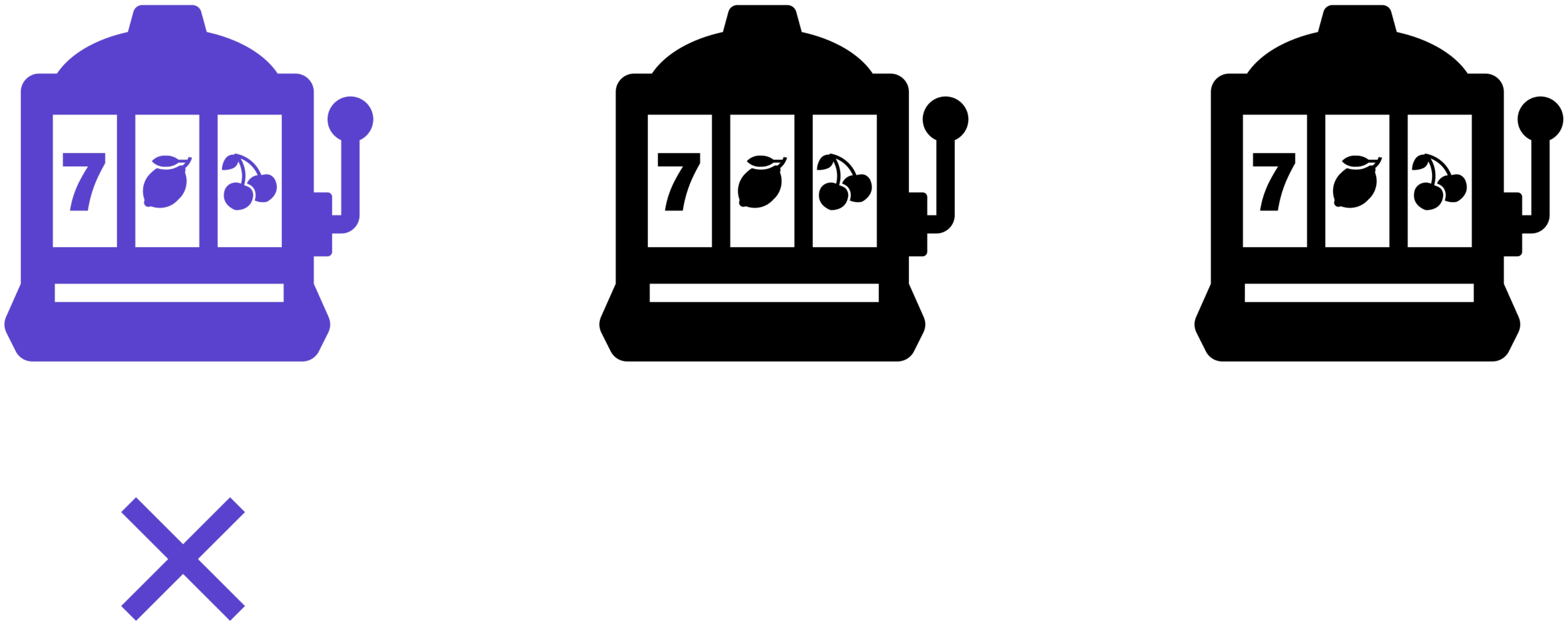
0.5

0.2

Multi-armed Bandit

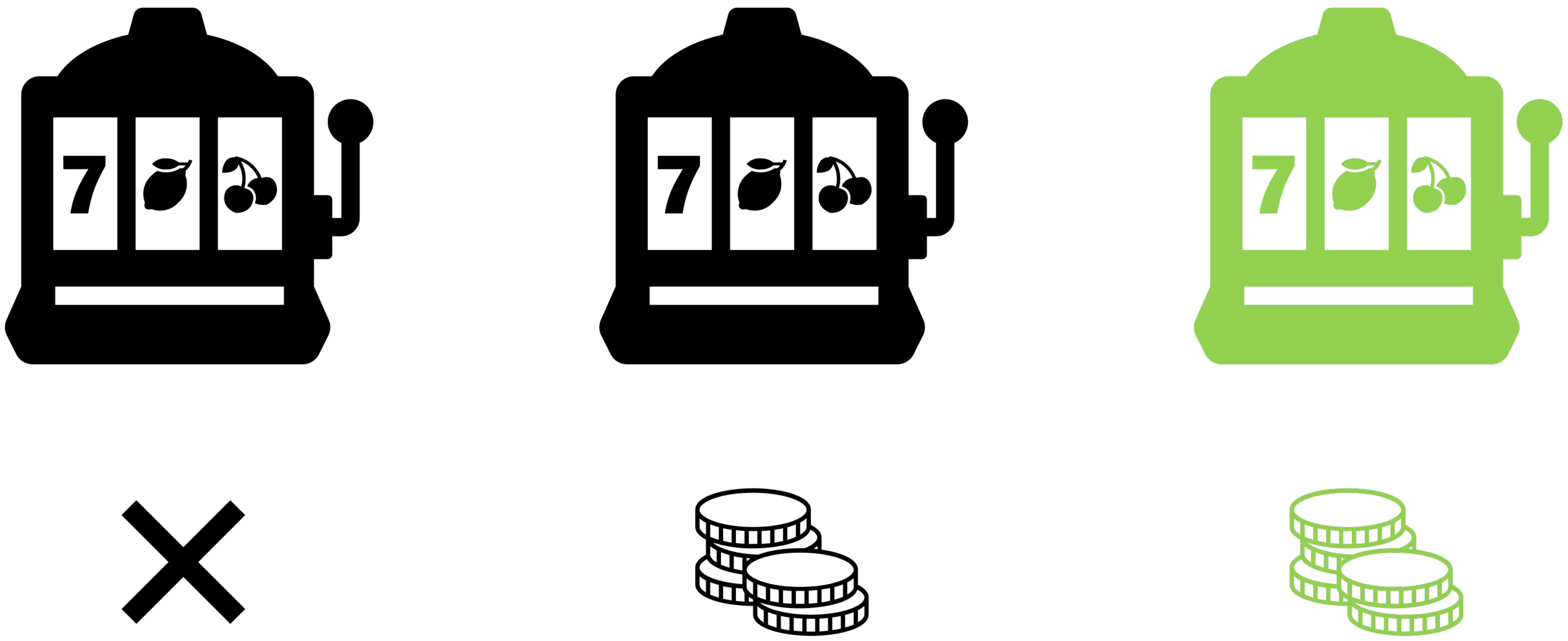
? ? ?

Multi-armed Bandit

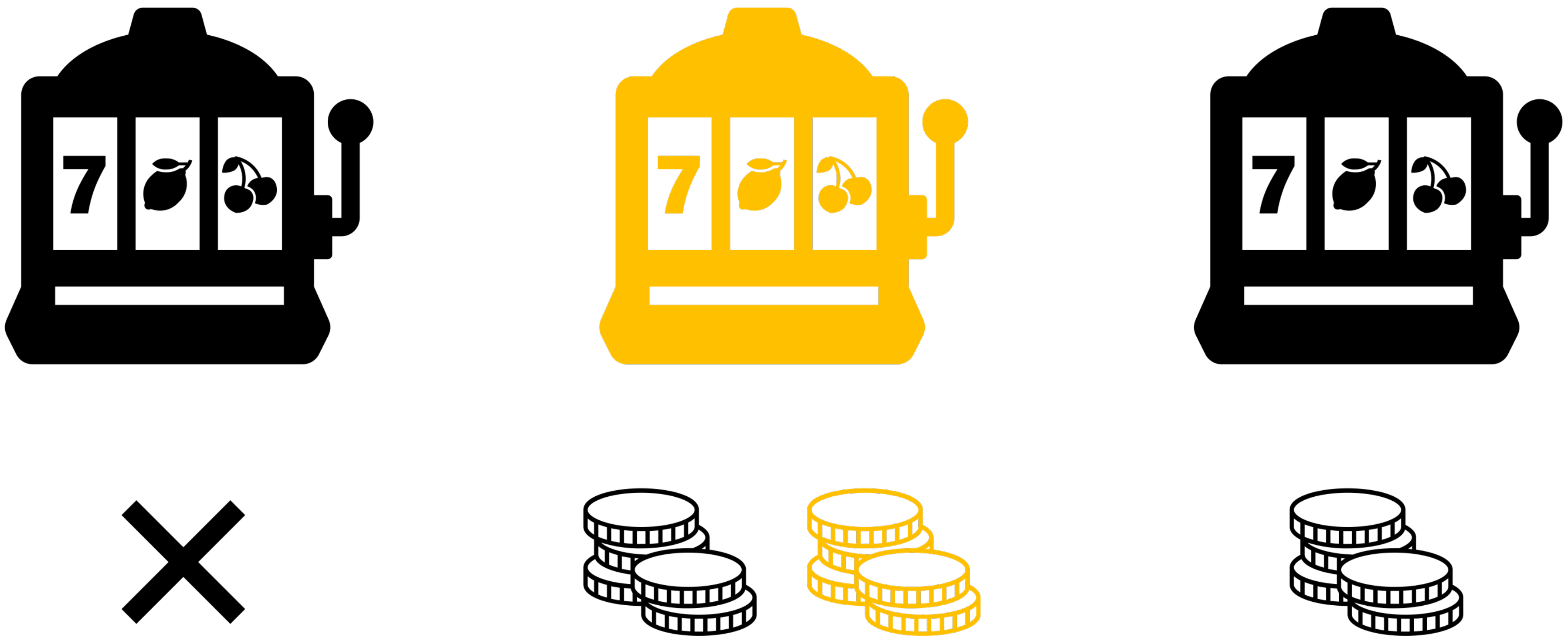


Multi-armed Bandit

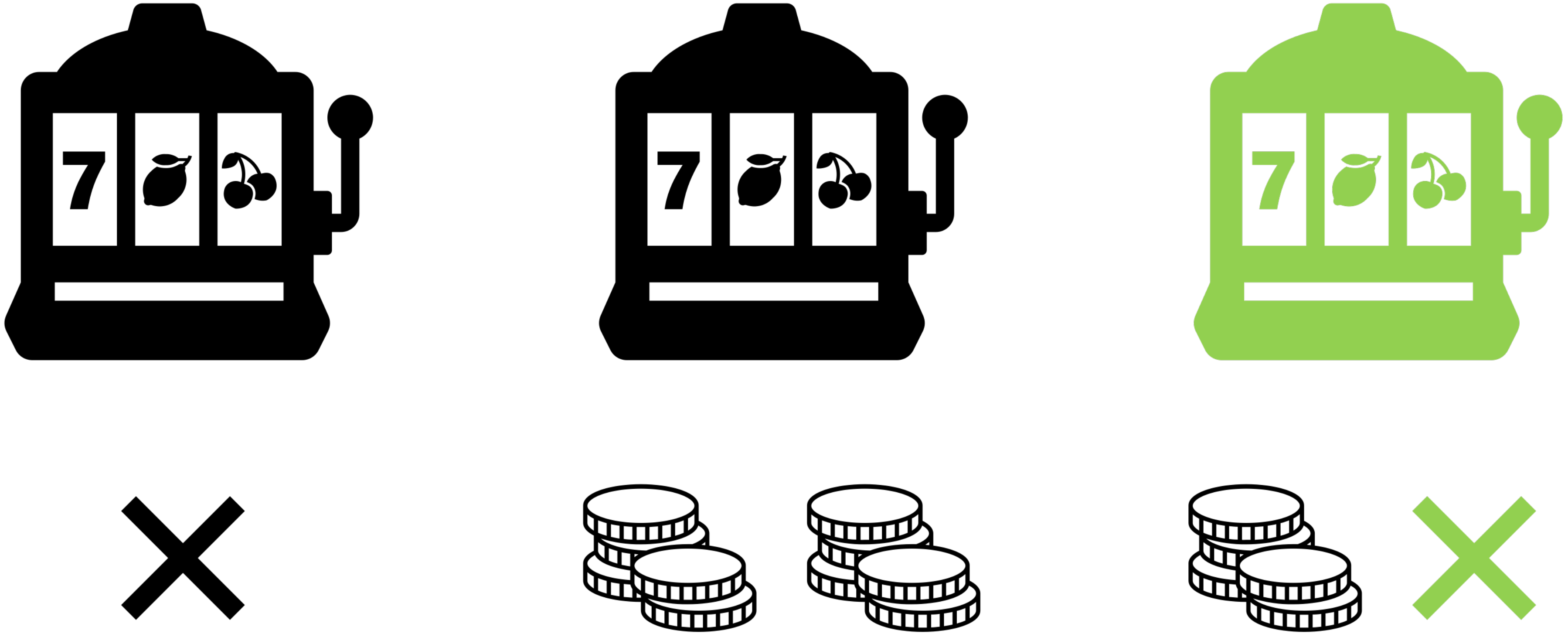
Multi-armed Bandit



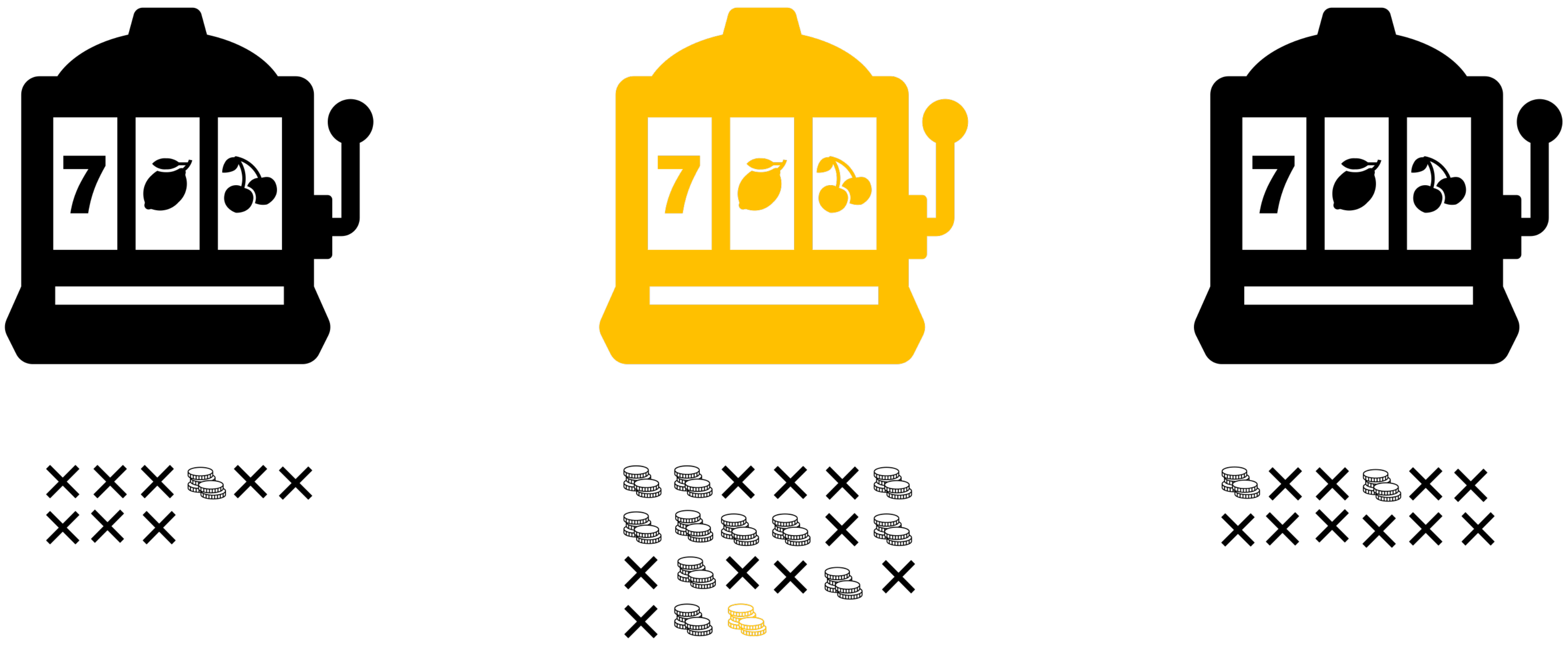
Multi-armed Bandit



Multi-armed Bandit



Multi-armed Bandit



Multi-armed Bandit

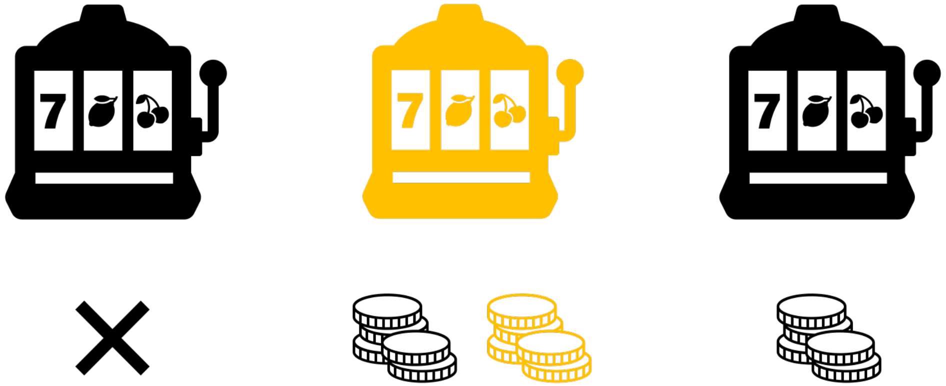
A balance between **exploration** and **exploitation** is required to maximise rewards.

Exploration/Exploitation Balance

Naïve strategies can fail to converge

Follow the Leader ('Greedy')

- Try each machine once to initialise
- Then proceed playing on the machine with the largest average pay-out so far.



Exploration/Exploitation Balance

Naïve strategies can fail to converge

Follow the Leader ('Greedy')

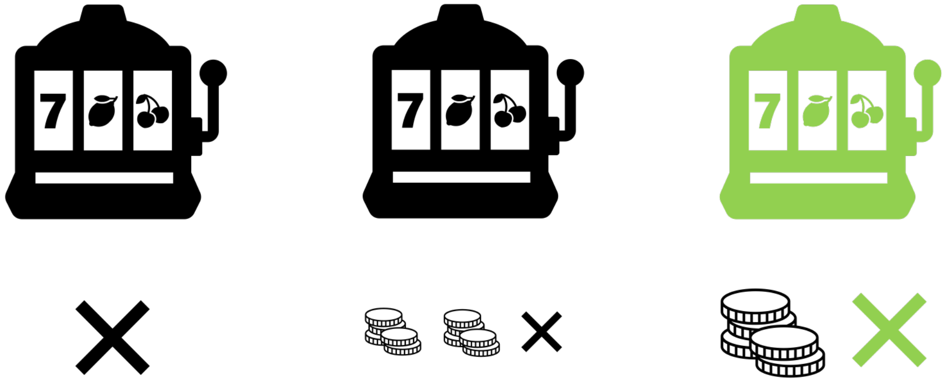
- Try each machine once to initialise
- Then proceed playing on the machine with the largest average pay-out so far.

Exploration/Exploitation Balance

Naïve strategies can fail to converge

Follow the Leader ('Greedy')

- Try each machine once to initialise
- Then proceed playing on the machine with the largest average pay-out so far.

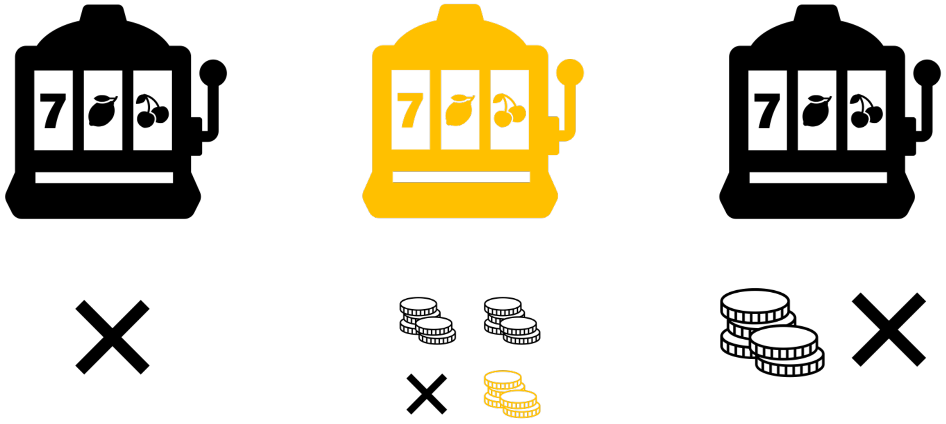


Exploration/Exploitation Balance

Naïve strategies can fail to converge

Follow the Leader ('Greedy')

- Try each machine once to initialise
- Then proceed playing on the machine with the largest average pay-out so far.

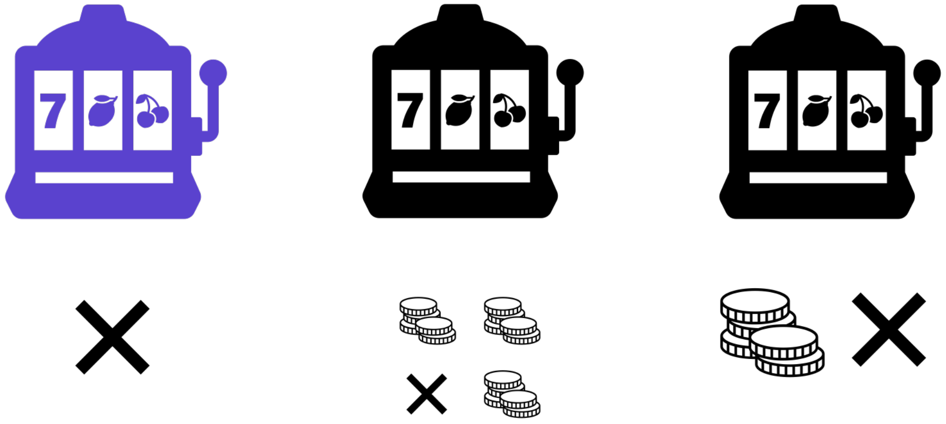


Exploration/Exploitation Balance

Naïve strategies can fail to converge

Follow the Leader ('Greedy')

- Try each machine once to initialise
- Then proceed playing on the machine with the largest average pay-out so far.



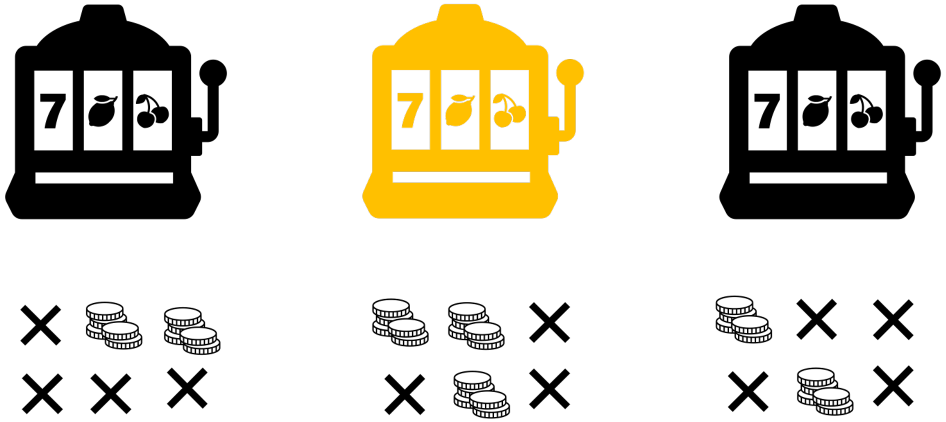
What if this machine had the largest pay-out probability?

Exploration/Exploitation Balance

Naïve strategies can fail to converge

Explore then Greedy

- Try each machine **many times** to initialise
- Then proceed playing on the machine with the largest average pay-out so far.

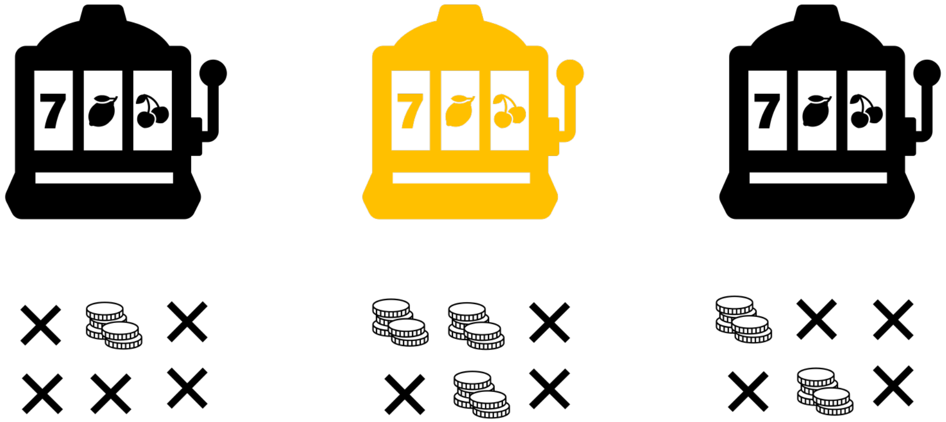


Exploration/Exploitation Balance

Naïve strategies can fail to converge

Explore then Greedy

- Try each machine **many times** to initialise
- Then proceed playing on the machine with the largest average pay-out so far.



Better, but need to specify a non-adaptive 'many'..

Getting the Balance Correct

One successful strategy is **optimism**.

Upper Confidence Bound Algorithm

- Make decisions based on an optimistic estimate of the mean payoff (**upper confidence bound**)

$\hat{p} = 0.2$

plays = 5

$CI: (0, 0.7)$

$\hat{p} = 0.5$

plays = 150

$CI: (0.45, 0.55)$

$\hat{p} = 0.3$

plays = 50

$CI: (0.2, 0.4)$

Getting the Balance Correct

One successful strategy is **optimism**.

Upper Confidence Bound Algorithm

- Make decisions based on an optimistic estimate of the mean payoff (**upper confidence bound**)

$\hat{p} = 0.2$

plays = 5

$CI: (0, 0.7)$

$\hat{p} = 0.5$

plays = 150

$CI: (0.45, 0.55)$

$\hat{p} = 0.3$

plays = 50

$CI: (0.2, 0.4)$

Getting the Balance Correct

One successful strategy is **optimism**.

Upper Confidence Bound Algorithm

- Make decisions based on an optimistic estimate of the mean payoff (**upper confidence bound**)

$\hat{p} = 0.1$

\# plays = 10

$CI: (0, 0.52)$

$\hat{p} = 0.5$

\# plays = 150

$CI: (0.45, 0.55)$

$\hat{p} = 0.3$

\# plays = 50

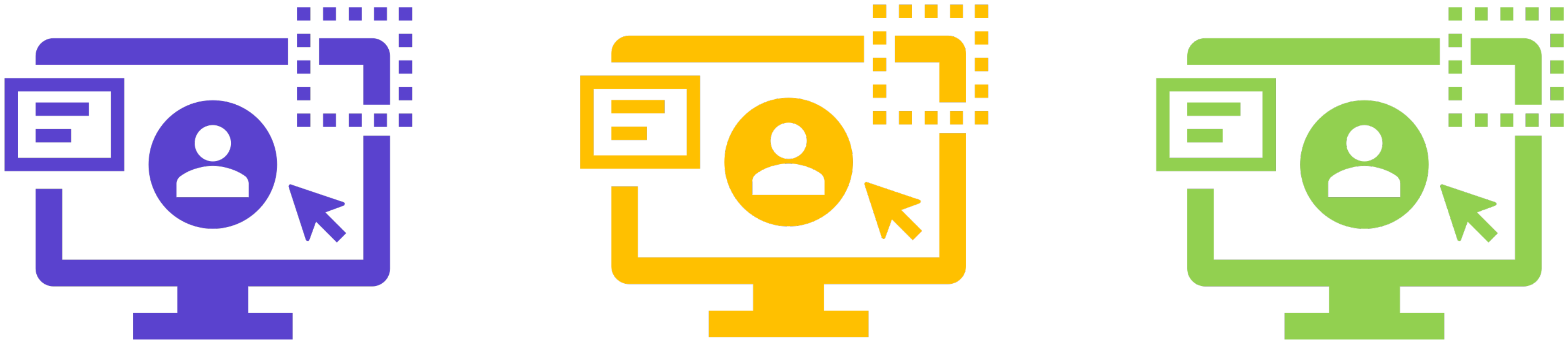
$CI: (0.2, 0.4)$

Applications

Applications

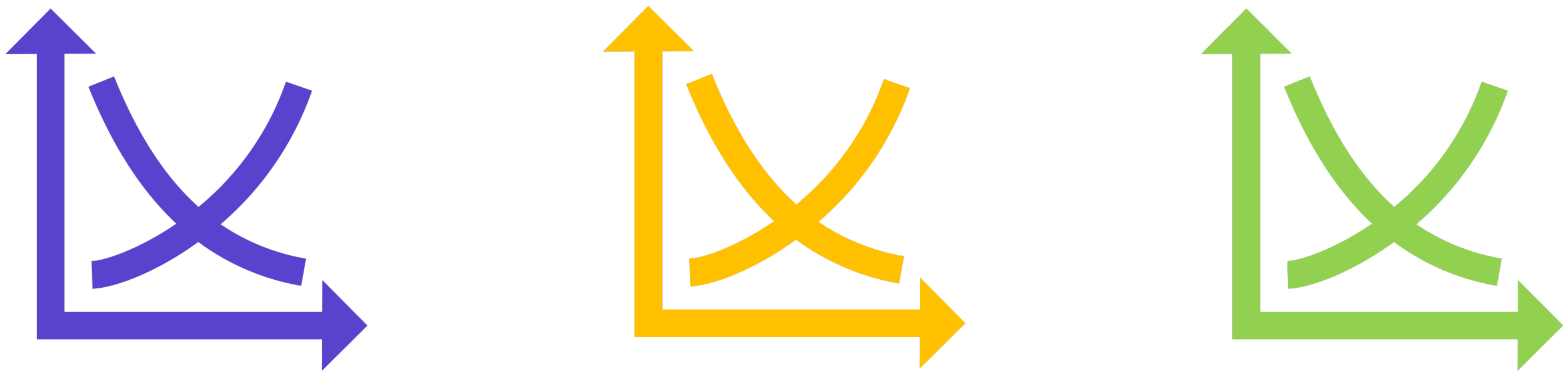
Clinical trials – particularly for rare diseases

Applications



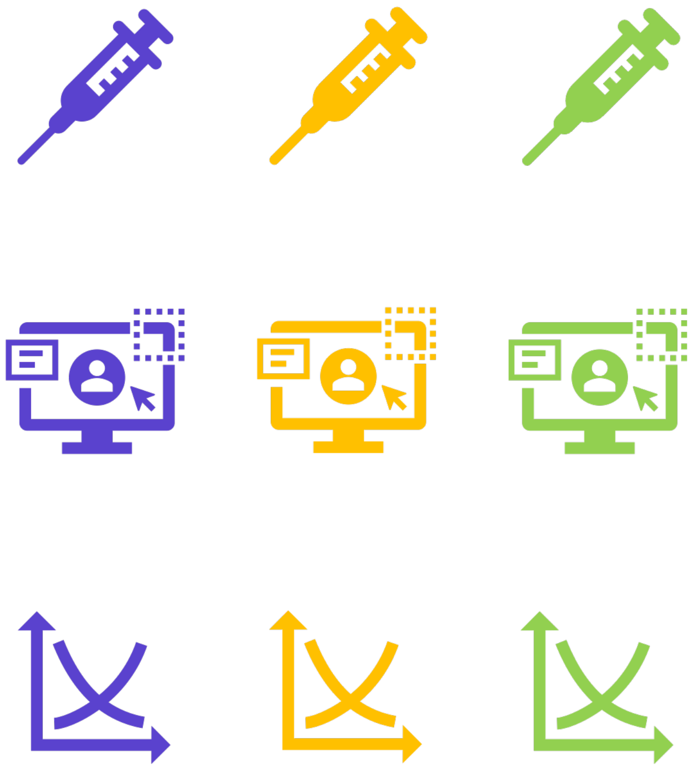
Online advertising, web design

Applications



Sequential modelling and data science

Extensions



- **Timescale restrictions**
  - Batched actions
  - Delayed feedback
- **Non-stationarity**
  - Abrupt
  - Slowly Evolving
- **Action sets**
  - Combinatorial
  - Continuous

Today's central message:

When the potential to make decisions repeatedly arises, we **can** and **ought** to do better than collecting data once, fitting a model once, and hoping for the best.

Today’s central message:

Optimistic techniques allow an appropriate, optimal balance between **exploration** (data collection) and **exploitation** (optimal decisions) to be struck.

Thank you for listening!

James Grant (he/him)
Lancaster University

j.grant@lancaster.ac.uk
@james_a_grant