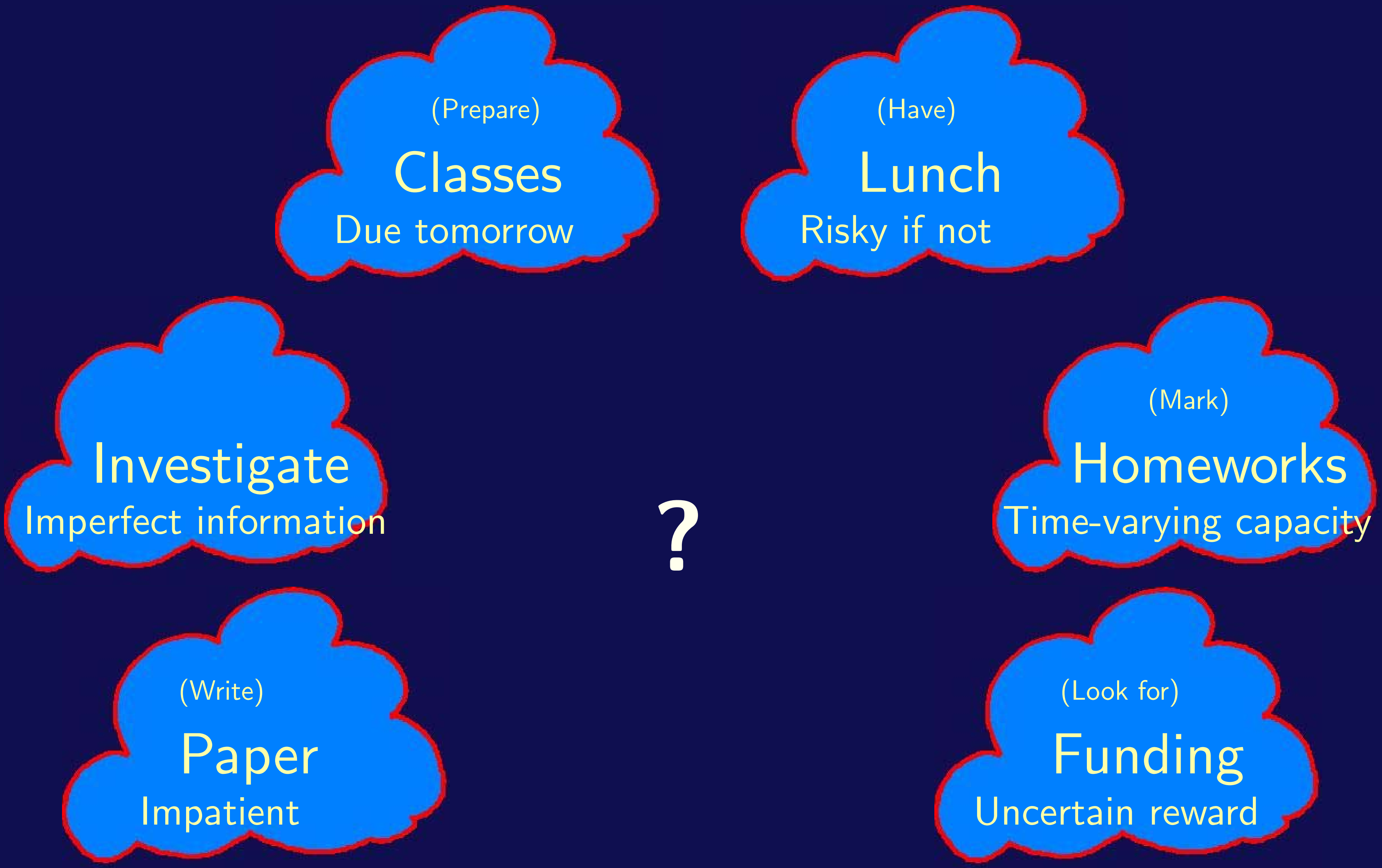
The Finite-Horizon Two-Armed Bandit Problem with Binary Responses

Peter Jacko*

Loughborough University
4 March 2020

*Lancaster University Management School, UK

Academic Task Management



(Prepare)
Classes
Due tomorrow

(Have)
Lunch
Risky if not

Investigate
Imperfect information

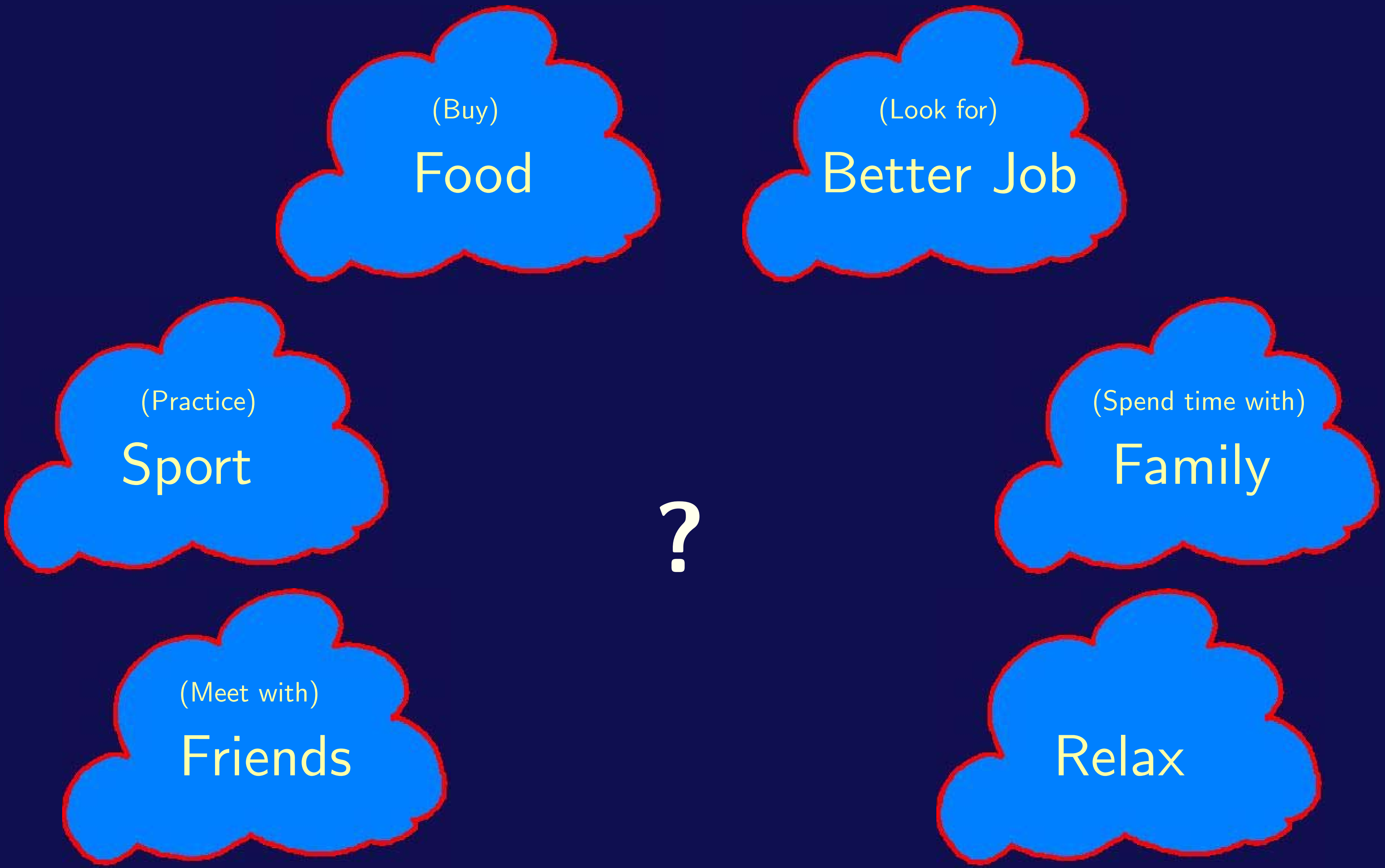
(Mark)
Homeworks
Time-varying capacity

?

(Write)
Paper
Impatient

(Look for)
Funding
Uncertain reward

Everyday Decision-Making



(Buy)

Food

(Look for)

Better Job

(Practice)

Sport

(Spend time with)

Family

?

(Meet with)

Friends

Relax

Questions to Answer

- [Economic] For a given joint goal, is it possible to define sound dynamic quantities for each task that can be interpreted as priorities? And if yes,
- [Algorithmic] How to calculate such priorities quickly?
- [Mathematical] Under what conditions is there a priority rule that achieves optimal resource capacity allocation?
- [Experimental] If priority rules are not optimal, how close to optimality do they come? And how do they compare to alternative policies?

The Scheduling Problem

Need for Generalisations

- Contact centres
  - customers are impatient
- Wireless networks
  - customers move and signal time-varies due to fading
- Retail industry
  - products are perishable and not homogeneous
- Healthcare
  - unknown novel treatments appear over time

Approach

- Problems intractable for finding an optimal solution

- Use of (dynamic) priorities in decision making
  - ▷ easy to interpret
  - ▷ easy to implement
  - ▷ often well-performing

- A divide and conquer solution approach

- Model: constrained stochastic dynamic programming
  - ▷ optimizing under the discounted or average criterion
  - ▷ subject to a sample path resource capacity constraint
  - ▷ e.g.: multi-armed restless bandit problem

Multi-Armed Restless Bandit Problem

JACKPOT PAYS
777 $50
JACKPOT PAYS
AAA $10
27 WAYS TO WIN Scattered PAYS 20 SEVENS
27 WAYS TO WIN Scattered PAYS 10 ACES
MONEYTREE Scatters SEVENS AND ACES
ARISTOCRAT REGAL SUPER '88'
PRESS TO RESERVE
175
BR 1476
REJECT COINS
MONEYTREE

Index Rules

- Priorities defined by (dynamic) index values

- Index rule: assign the resource to the customer with highest actual index value

- Proposed in increasingly more general settings by
  - ▷ Smith (1956): job scheduling (optimal)
  - ▷ Gittins (1970's): classic bandits (optimal)
  - ▷ Whittle (1988): restless bandits (asympt. optimal)
  - ▷ Glazebrook et al., Jacko et al. (2005–): dynamic resource allocation (asympt. optimal)
    – index-knapsack heuristic

- Index rules are often tractable solutions

Multi-disciplinary Bandits

- Different terminology across disciplines

Anecdotic	strategy	choice	pull	arms
OR	policy	allocation	resource	projects
CS/ML	algorithm	decision	time step	actions
Biometrics	design	randomisation	patient	treatments
Telecom	scheduler	allocation	server	jobs
Universal (?)	design	allocation	subject	interventions

Clinical Trials

- The gold standard design: randomised controlled trial
  - ▷ 50% vs 50% fixed equal randomisation
  - ▷ avoids all types of biases
  - ▷ in use since 1948 (advocated since Hill 1937)

- Its main goal is to learn about intervention effectiveness with a view to prioritise future outside subjects
  - ▷ maximises power of an intervention effect difference
  - ▷ if approved, future subjects are, say, 95% confident that the novel intervention is better than the control

- A half of trial subjects gets the inferior intervention

Randomised Controlled Trial

- Statistical testing based on randomised equal allocation is a widespread state-of-the-art approach in the design of experiments, under different names:
  - ▷ randomised controlled trial in biostatistics
  - ▷ between-group design in social sciences
  - ▷ A/B testing in Internet marketing

Bayesian Decision-Theoretic Trial

"...there can be no objection to the use of data, however meagre, as a guide to action required before more can be collected ... Indeed, the fact that such objection can never be eliminated entirely—no matter how great the number of observations—suggested the possible value of seeking other modes of operation than that of taking a large number of observations before analysis or any attempt to direct our course... This would be important in cases where either the rate of accumulation of data is slow or the individuals treated are valuable, or both."

Bayesian Decision-Theoretic Trial

- Proposed in Thompson 1933 (pre-dates Hill 1937)

- The goal is to provide higher benefit to both in-trial subjects and after-trial subjects

  ▷ as opposed to the RCT's learning goal of reliable intervention effect estimation

- It is done by deciding the allocation, i.e., the randomisation probabilities for every subject (or for a group of subjects)

  ▷ response-adaptive: decisions based on the responses accumulated so far, i.e. Bayesian

Bayesian Decision-Theoretic Trial

- In theory, can be solved to optimality by decision theory
- In practice, optimal decisions are computed numerically
  - ▷ it is often believed to be tractable only for small trials
- Milestones IMHO (w.r.t. clinical trials)
  - ▷ Thompson (Biometrika 1933)
  - ▷ Glazebrook (Biometrika 1978)
  - ▷ Gittins & Jones (Biometrika 1979)
  - ▷ Armitage (ISR 1985)
  - ▷ Cheng, Su & Berry (Biometrika 2003)
  - ▷ Berry (Nature 2006), Cheng & Berry (Biometrika 2007)
  - ▷ Villar, Bowden & Wason (Statistical Science 2015)

Health Benefit Approach

- Important because healing patients is the ultimate goal of new treatment development
- Bayesian decision-theoretic model
  - ▷ optimally solving learning/healing trade-off
  - ▷ both learning and healing takes place during the trial
- This kind of general problem became known as the multi-armed bandit problem

Bayesian Bernoulli Bandit Model

- Finite horizon: n sequentially arriving subjects
- Two-armed: intervention A or B for each subject
- Binary endpoints: success (1) or failure (0)
- Let X_i and Y_i denote subject i's response from intervention A and B respectively (for $i = 1, ..., n$). Then,

$$X_i \sim \text{Bernoulli}(1,\ \theta_A) \ \text{ and } \ Y_i \sim \text{Bernoulli}(1,\ \theta_B),$$

where θ_A and θ_B are the unknown success probabilities of interventions A and B respectively

Bayesian Approach

- Beliefs $\widehat{\theta}_A$ and $\widehat{\theta}_B$ to be updated over the trial

- Prior Distribution: $\widehat{\theta}_A \sim \text{Beta}(a,\ b)$, $\widehat{\theta}_B \sim \text{Beta}(c,\ d)$ where we take $a = b = c = d = 1$ (uninformative)

- Posterior Distribution: After observing i (j) successes (failures) on intervention A, and k (l) successes (failures) on intervention B, the posterior distribution is represented by another Beta distribution (by conjugacy)

$$\widehat{\theta}_A \sim \text{Beta}(a+i, b+j),\ \widehat{\theta}_B \sim \text{Beta}(c+k, d+l)$$

DP Design

- We use dynamic programming (DP) to obtain an optimal adaptive intervention allocation sequence

- Optimal in the sense of maximising the expected total number of successes in the trial

- Specifically, we use backward induction algorithm

- Let $\mathcal{F}_m(i, j, k, l)$ be the expected total number of successes under an optimal policy after m subjects

- If $m = n$, there is nothing to do: $\mathcal{F}_n(i, j, k, l) = 0$ $\forall i, j, k, l$

Backward Recursion

- If $m = n - 1$ (one subject left):

1. If intervention A, we compute the expectation

$$\mathcal{F}^{A}_{n-1}(i,j,k,l) = \frac{i}{i+j} \cdot 1 + \frac{j}{i+j} \cdot 0$$

2. If intervention B, we compute the expectation

$$\mathcal{F}^{B}_{n-1}(i,j,k,l) = \frac{k}{k+l} \cdot 1 + \frac{l}{k+l} \cdot 0$$

- We wish to choose the optimal allocation such that

$$\mathcal{F}_{n-1}(i,j,k,l) = \max\{\mathcal{F}^{A}_{n-1}(i,j,k,l),\ \mathcal{F}^{B}_{n-1}(i,j,k,l)\}$$

Optimal Designs

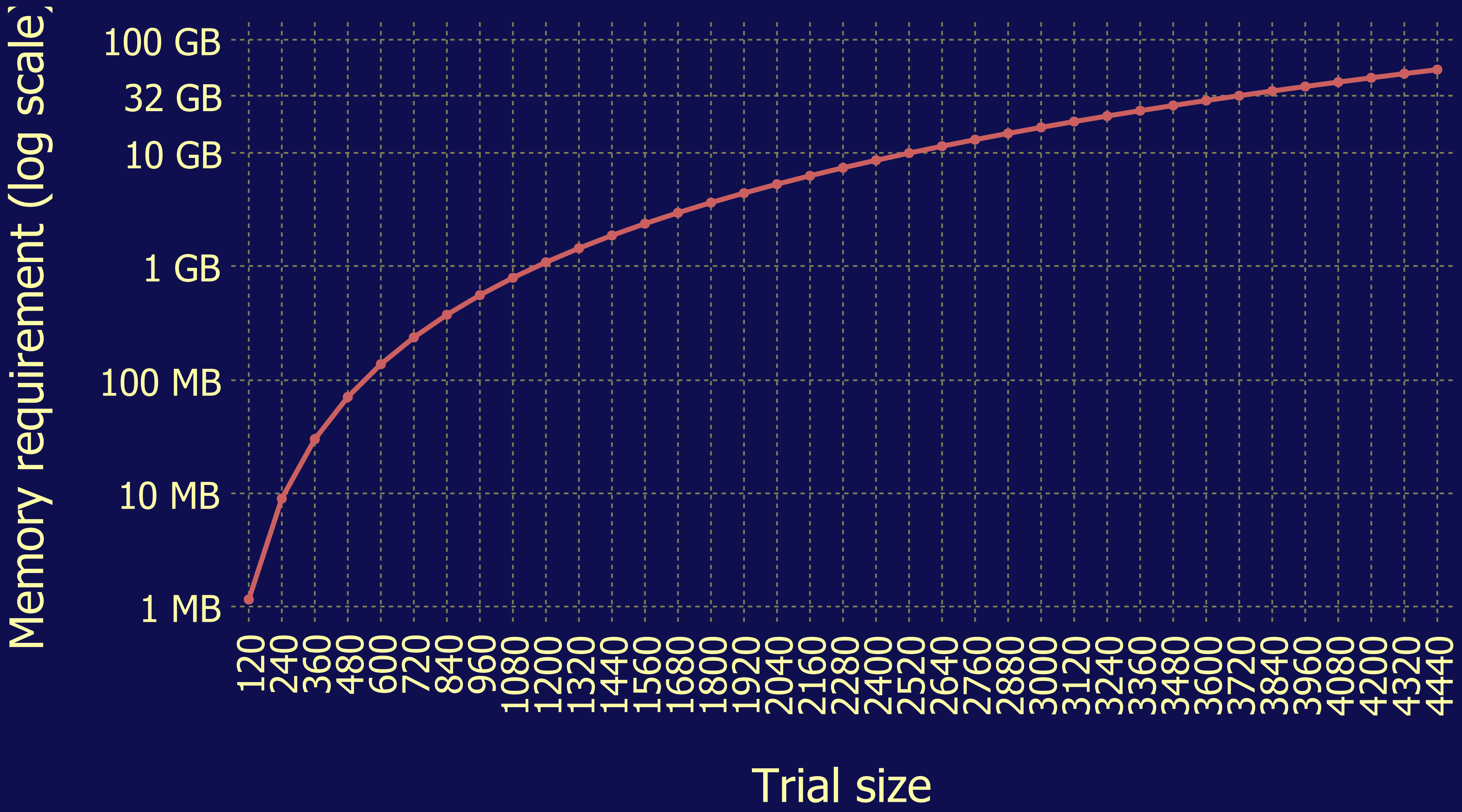
- The designs computed using DP are optimal, i.e. provide the maximum benefit given their respective restrictions

- For the two-armed case (on a standard laptop):
  - ▷ a basic R code can design trials of size up to 200
  - ▷ an efficient Julia code up to $1,500$
  - ▷ a during-the-trial computation allows even larger trials

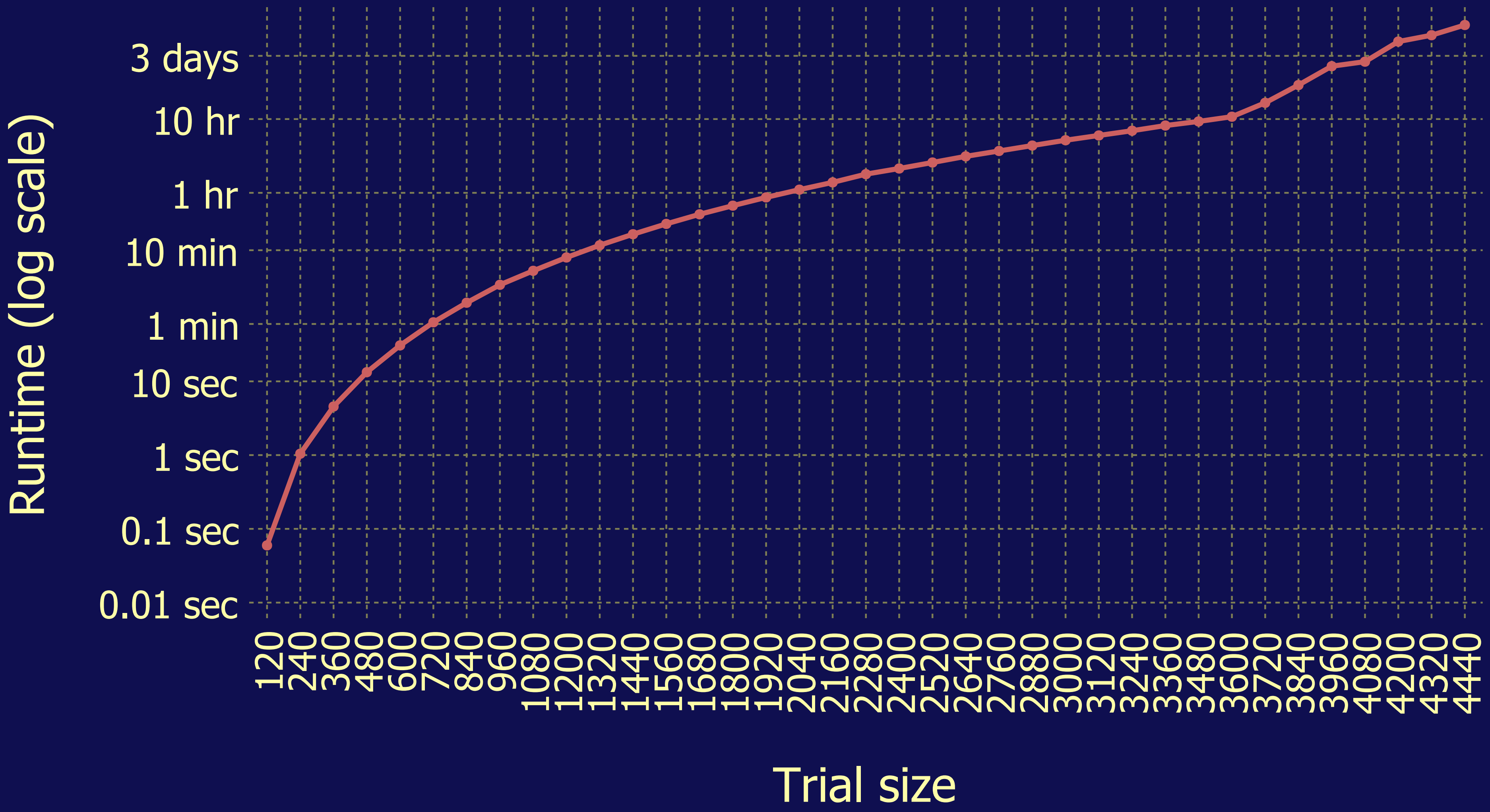
- Longer trials can be designed on a workstation/cloud

- More complex trials with much smaller sizes

Optimal Designs: Memory



Optimal Designs: Runtime



Conclusion

- Powerful approach to omnipresent intractable problems
  - ▷ elegant, easy to interpret/implement
  - ▷ index rules optimal for relaxations
  - ▷ suggests structure of (asymptotically) optimal policies
  - ▷ valuable for both researchers and practitioners

Conclusion

- Powerful approach to omnipresent intractable problems
  - elegant, easy to interpret/implement
  - index rules optimal for relaxations
  - suggests structure of (asymptotically) optimal policies
  - valuable for both researchers and practitioners
- Stochastic literacy
  - priorities are better to use than stereotypes
  - intuition what is probability
  - very poor, including among mathematicians (!)
  - more important than exact mathematical literacy (?)

Thank you for your attention

Ďakujem za Vašu pozornosť