PhD Conclusion Talk

Adaptive Recommendations with Bandit Feedback

Mengyan Zhang

Computational Media Lab, Australian National University
Machine Learning Research Group, Data61,CSIRO

Supervisors: Cheng Soon Ong, Lexing Xie, Eduardo Eyras

Multi-armed Bandits:
Sequential decision-making

In each round $t \epsilon \{1, .., N\}$,

1. an agent selects an arm $A_t = i \epsilon \mathcal{K}$ according policy π

2. then receive a reward $X_{i,T_i(t)}$ sampled from unknown distribution F_i

3. update estimations over distribution F_i based on historical observations

Best choice with
the current information

Other possibilities
have not been tried or
with high uncertainty

Exploitation vs. Exploration ?

How to allocate samples adaptively ?

Why Bandits?

- Challenge: **sequential decision-making with uncertainty**
- Provide model for **E & E dilemma**
- **Applications**:
  - Adaptive experimental design: clinical, drug, food
  - Configure web interfaces: item recommendations, dynamic pricing, ad placement
  - Plays a role in algorithms like Monte Carlo Tree Search
- Rich structure connecting to other branches of **math**: concentration analysis, information theory, etc.

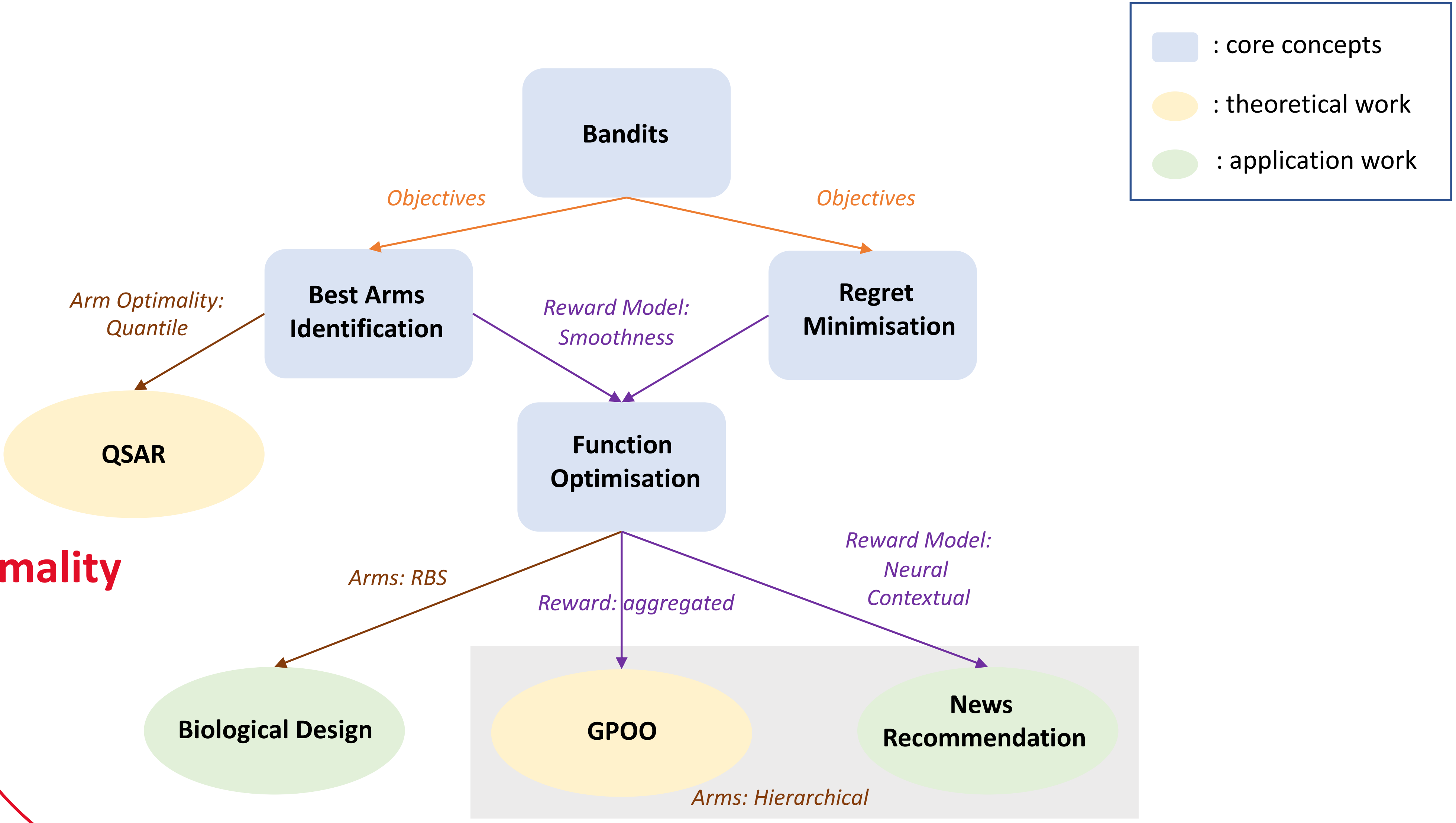
Tor Lattimore and Csaba Szepesvari, 2020. Bandit Algorithms. Cambridge University Press.

Three Design Choices of bandit tasks

- Objectives:
  - What's the goal of designing a policy?
  - How to evaluate the performance?
- Rewards
  - How to model the rewards? – Smoothness, context
- Arms
  - How to define an arm?
  - Can we form new arms based on single arms?
  - How to define the optimality of arms?

Outline

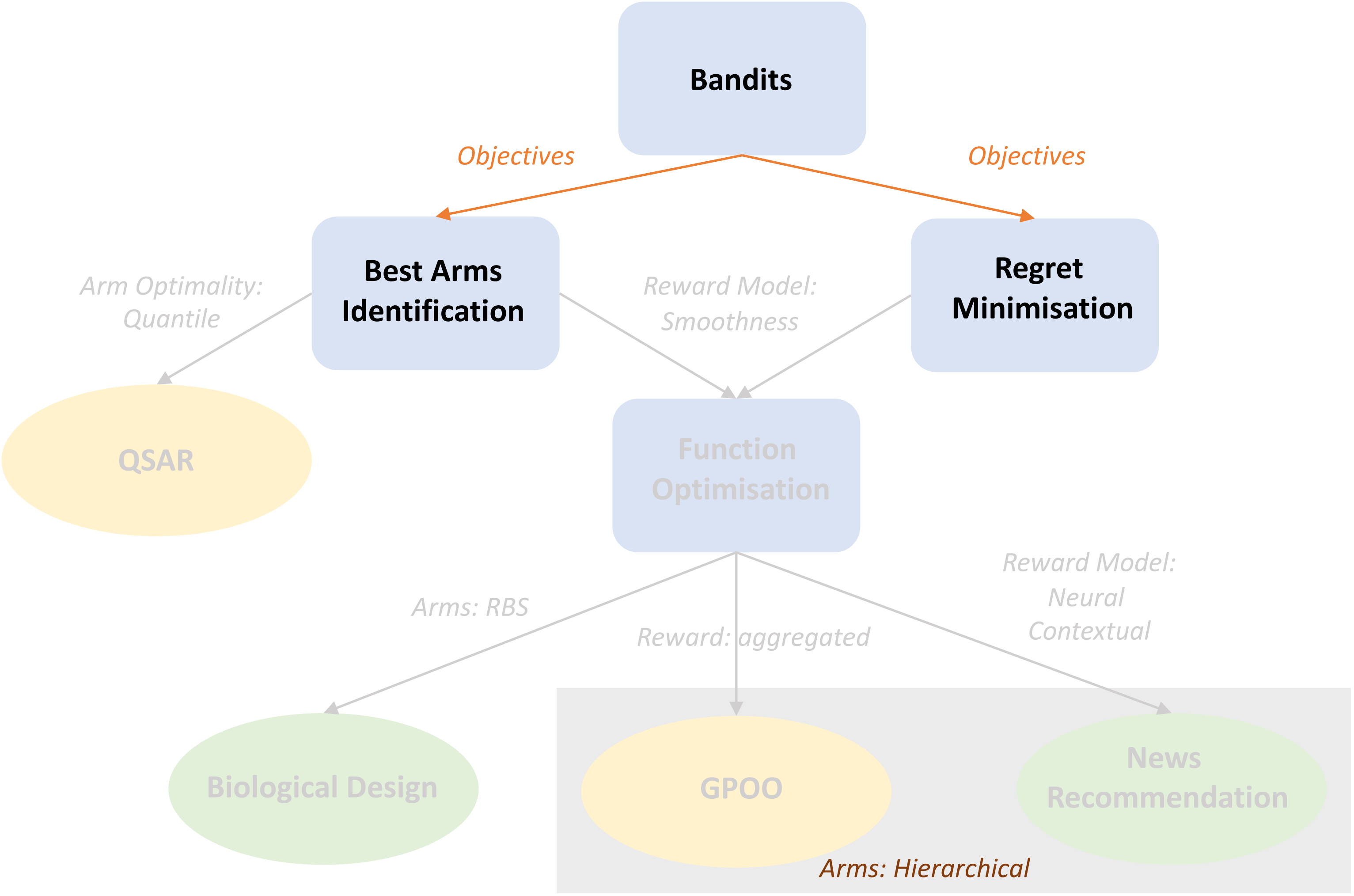
This Talk:



Arm: Optimality

Arm: Bio → Arm: Hierarchical

Outline


Bandits
Objectives
Objectives
Best Arms Identification
Regret Minimisation
Arm Optimality: Quantile
Reward Model: Smoothness
QSAR
Function Optimisation
Arms: RBS
Reward: aggregated
Reward Model: Neural Contextual
Biological Design
GPOO
News Recommendation
Arms: Hierarchical

Objectives

- **Best arms identification (BAI):** have separate exploration stage

  identify the best m items when the exploration stage ends, e.g. with fixed budget N

Simple regret $r_N = \sum_{i=1}^{m}(\mu_{o_i} - \mathbb{E}[\mu_{A_N^i}])$ where $\mu_{o_1} \geq \cdots \geq \mu_{o_K}$

Probability of error $e_N := \mathbb{P}\left(\mathcal{S}_m^N \neq \mathcal{S}_m^*\right)$

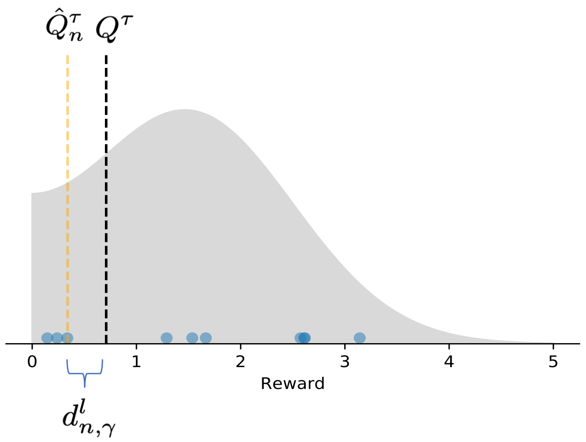
- **Regret minimization:** have no separate exploration stage

  recommend items sequentially to users with the goal of minimising cumulative regret

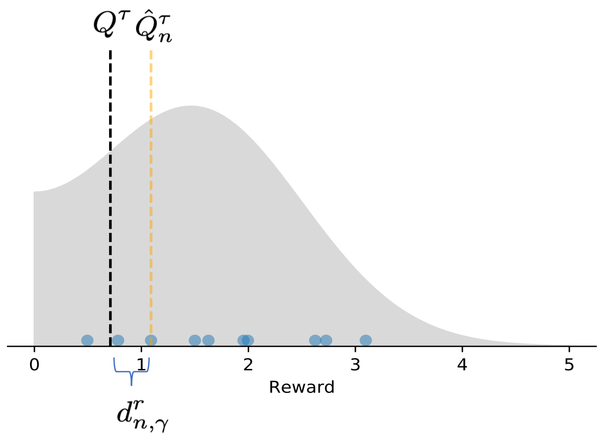
Cumulative regret $R_N = N\mu_* - \mathbb{E}\left[\sum_{t=1}^{N} X_{A_t}\right]$

What matters and how to achieve? – In Theory

- Regret bounds (in expectation, or in high probability)
  - Sublinear regret e.g. $\lim_{N\to\infty} \frac{R_N}{N} = 0$
  - Probability or error e.g. decrease exponentially wrt budget, $O(\exp(-N))$
- How: utilise concentration inequalities:



$$\mathbb{P}\left(Q^\tau - \hat{Q}_n^\tau \geq d_{n,\gamma}^l\right) \leq \exp(-\gamma)$$

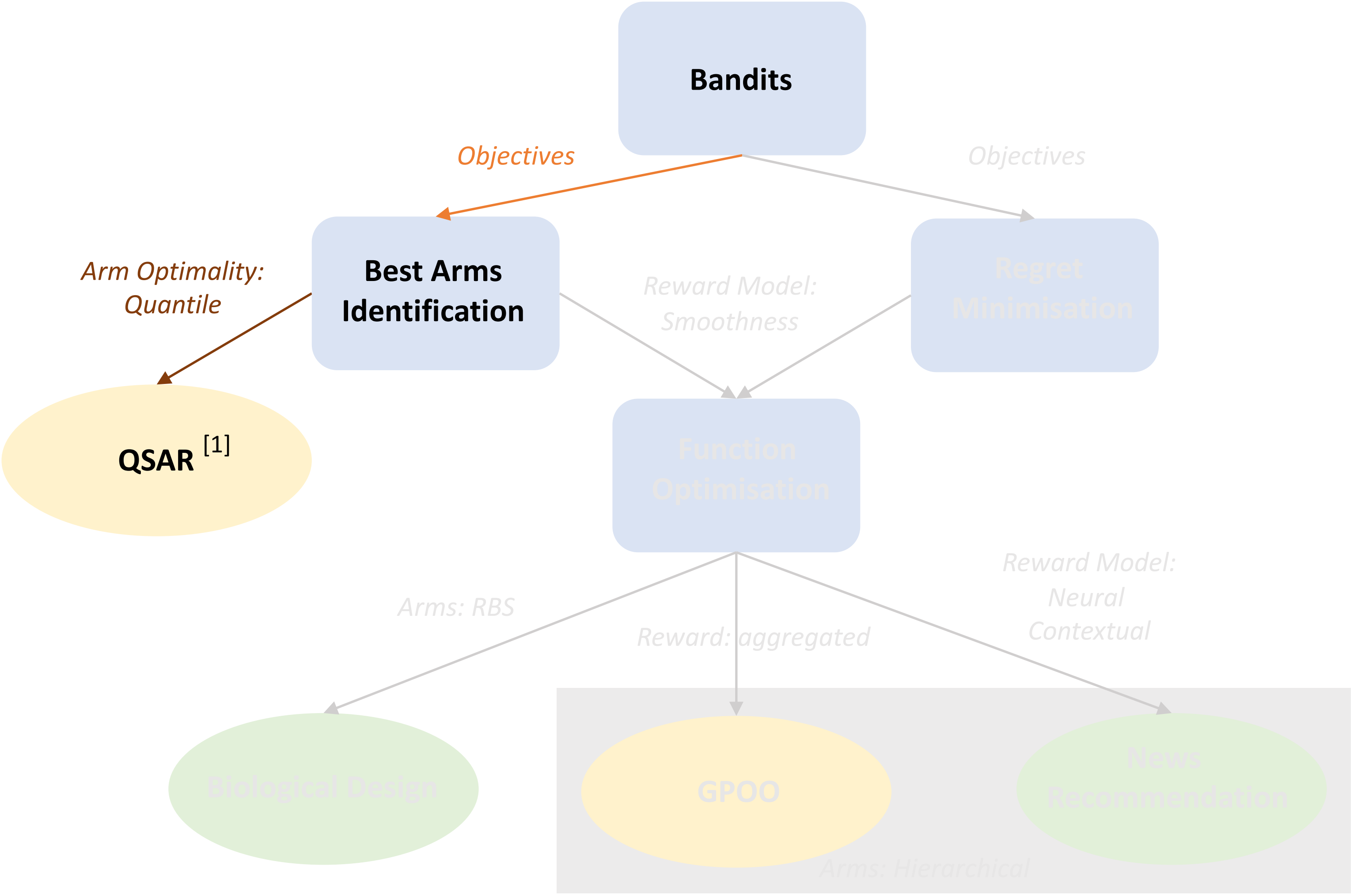


$$\mathbb{P}\left(\hat{Q}_n^\tau - Q^\tau \geq d_{n,\gamma}^r\right) \leq \exp(-\gamma)$$

What matters and how to achieve? – In Practice

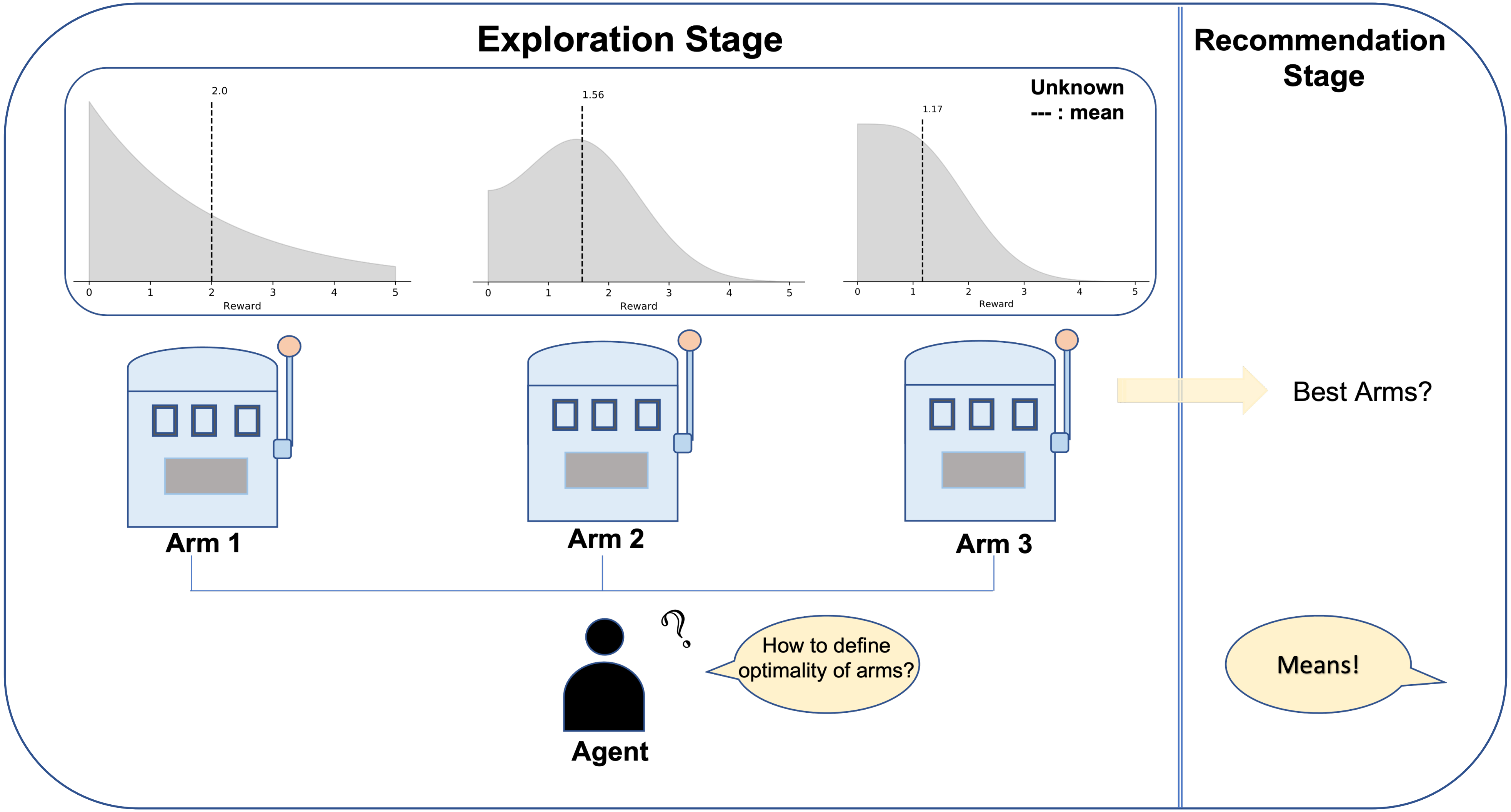
- Performance
  - regret, probability of error
  - The improvement over random/baselines
- How:
  - Model assumption – fits the real applications
  - Quality of predictions of labels and uncertainty – representations, neural, Bayesian methods
  - Large design space – hierarchical design

Outline

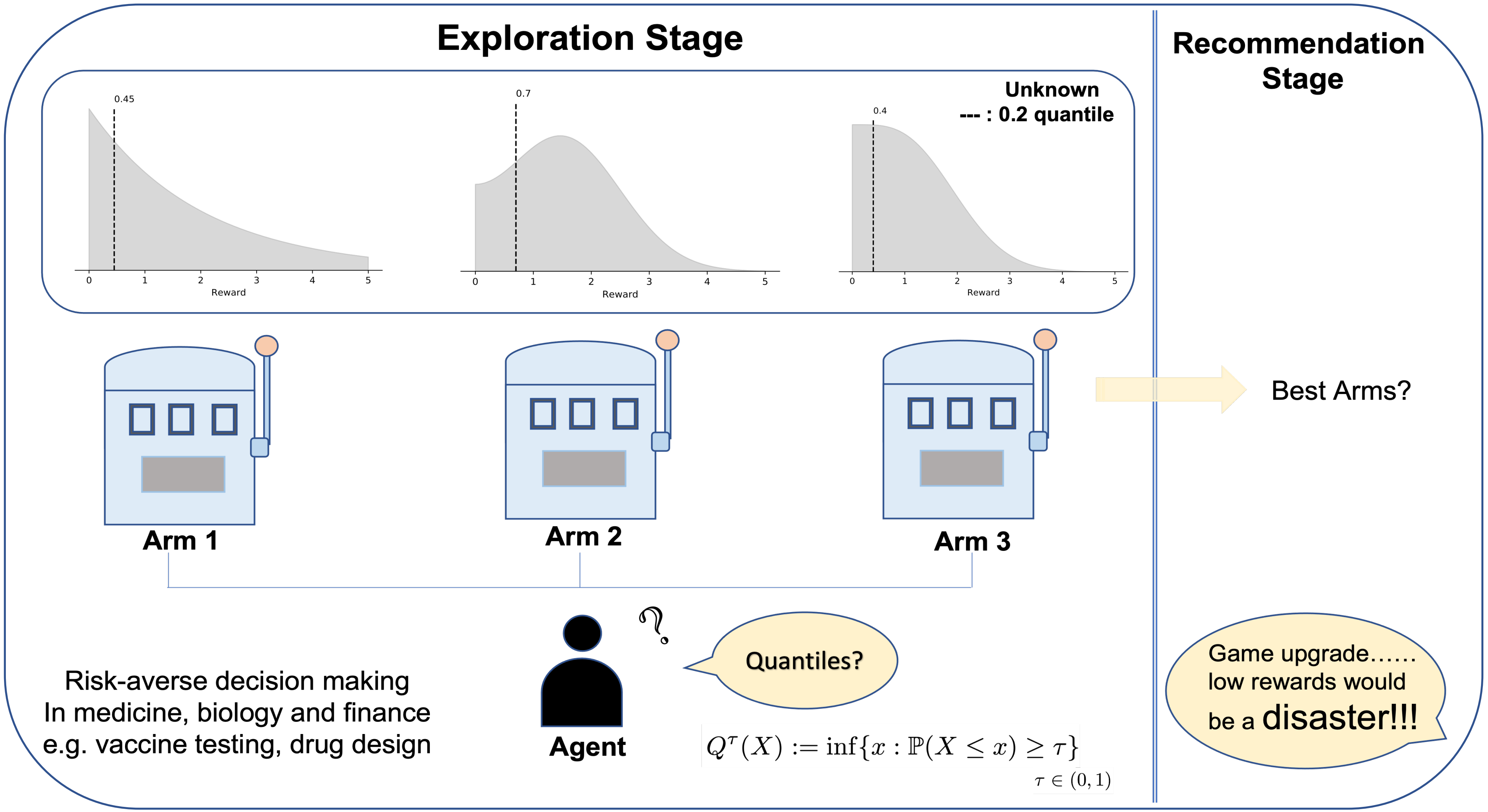


[1] Quantile Bandits for Best Arms Identification. **Mengyan Zhang**, Cheng Soon Ong. International Conference on Machine Learning 2021.

Best Arms Identification with Fixed Budget

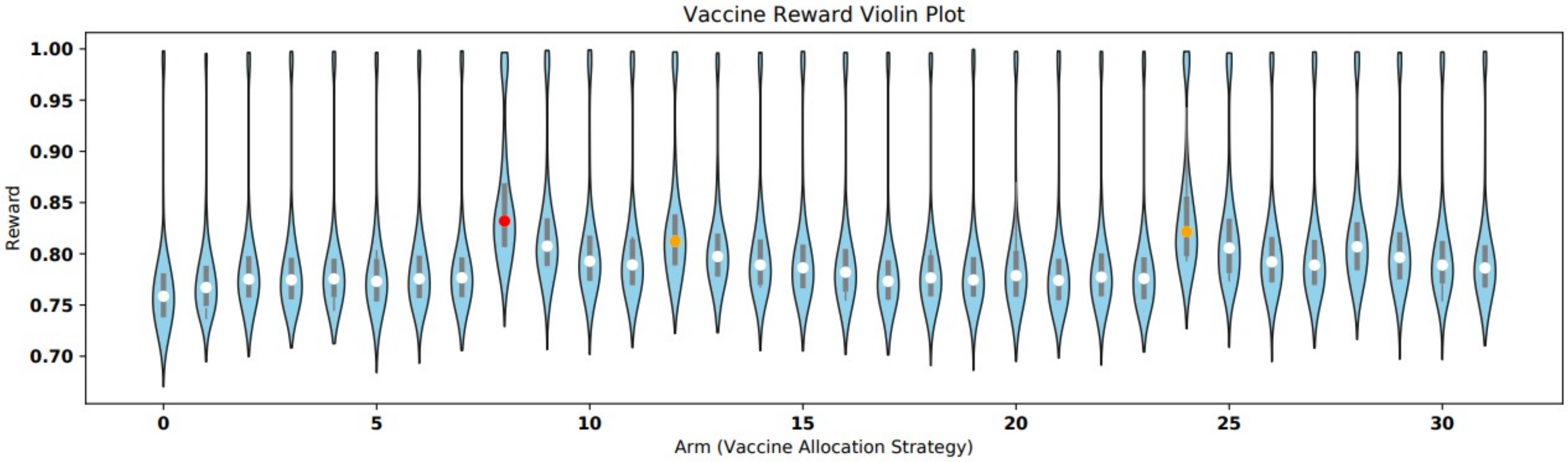


BAI with Quantiles

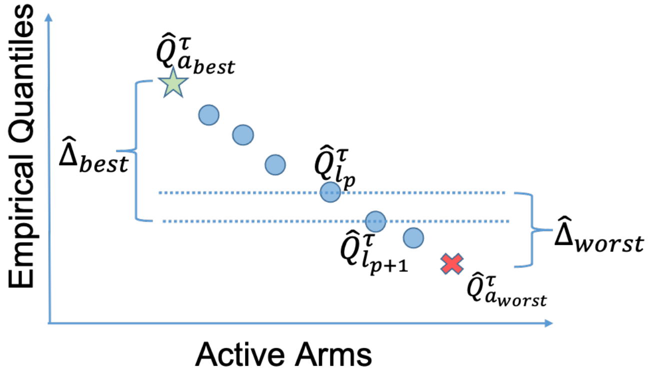


Applications: vaccine allocation

- Identify optimal strategies (highest **median** reward) for vaccine allocation
- **Arm**: vaccine allocation strategy (Allocate 100 vaccine doses to 5 age groups -- all combinations as arms)
- **Reward**: proportion of individuals that did not experience symptomatic infection



Contributions on Quantile BAI



$\hat{\Delta}_{best} > \hat{\Delta}_{worst}$: Accept

$\hat{\Delta}_{best} \leq \hat{\Delta}_{worst}$: Reject

New Algorithm: Quantile-based Successive Accepts and Rejects (Q-SAR)

Extends Bubuck et al. 2013: Mean-based SAR

New Concentration inequalities

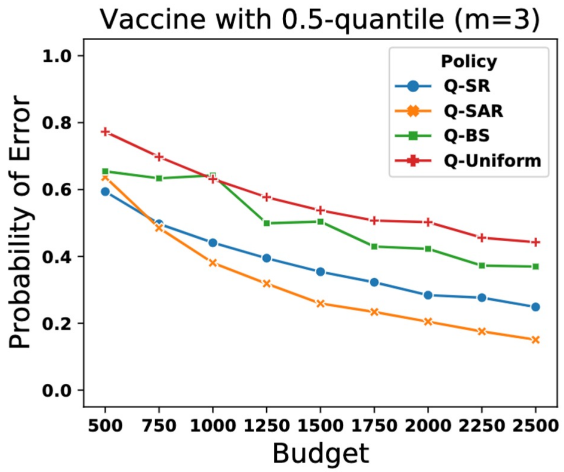
$$\mathbb{P}\left(Q^{\tau} - \hat{Q}^{\tau}_{n} \geq d^{l}_{n,\gamma}\right) \leq \exp(-\gamma)$$

$$\mathbb{P}\left(\hat{Q}^{\tau}_{n} - Q^{\tau} \geq d^{r}_{n,\gamma}\right) \leq \exp(-\gamma)$$

Probability of error

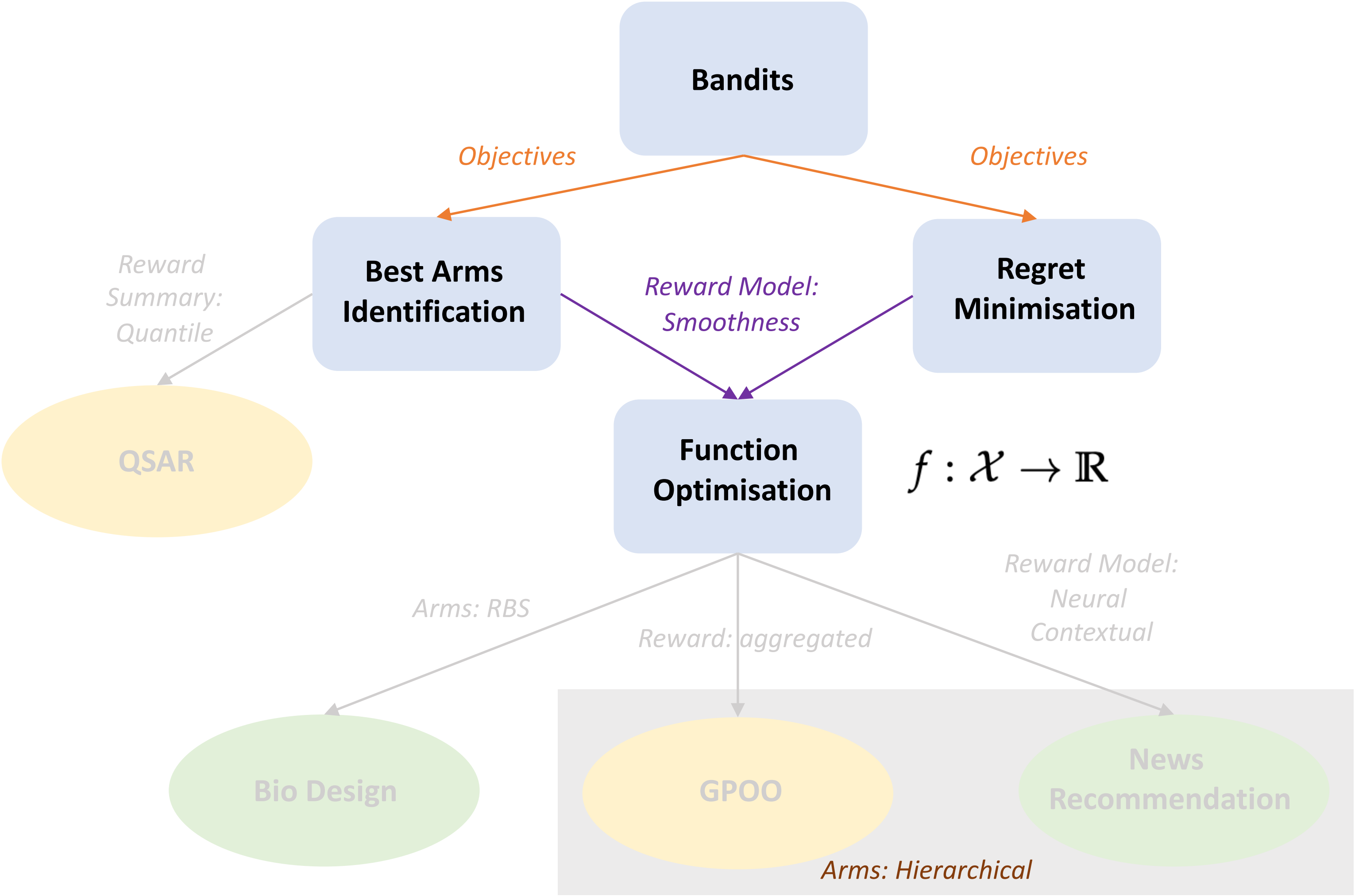
$$e_N := \mathbb{P}\left(\mathcal{S}^{N}_{m} \neq \mathcal{S}^{*}_{m}\right) \leq 2K^2 \exp\left(-\frac{N-K}{\overline{\log}(K) H^{\tau}}\right)$$

Experiments on vaccine allocation



Bubeck, S.; Wang, T.; and Viswanathan, N. Multiple identifications in multi-armed bandits. ICML 2013.

Outline

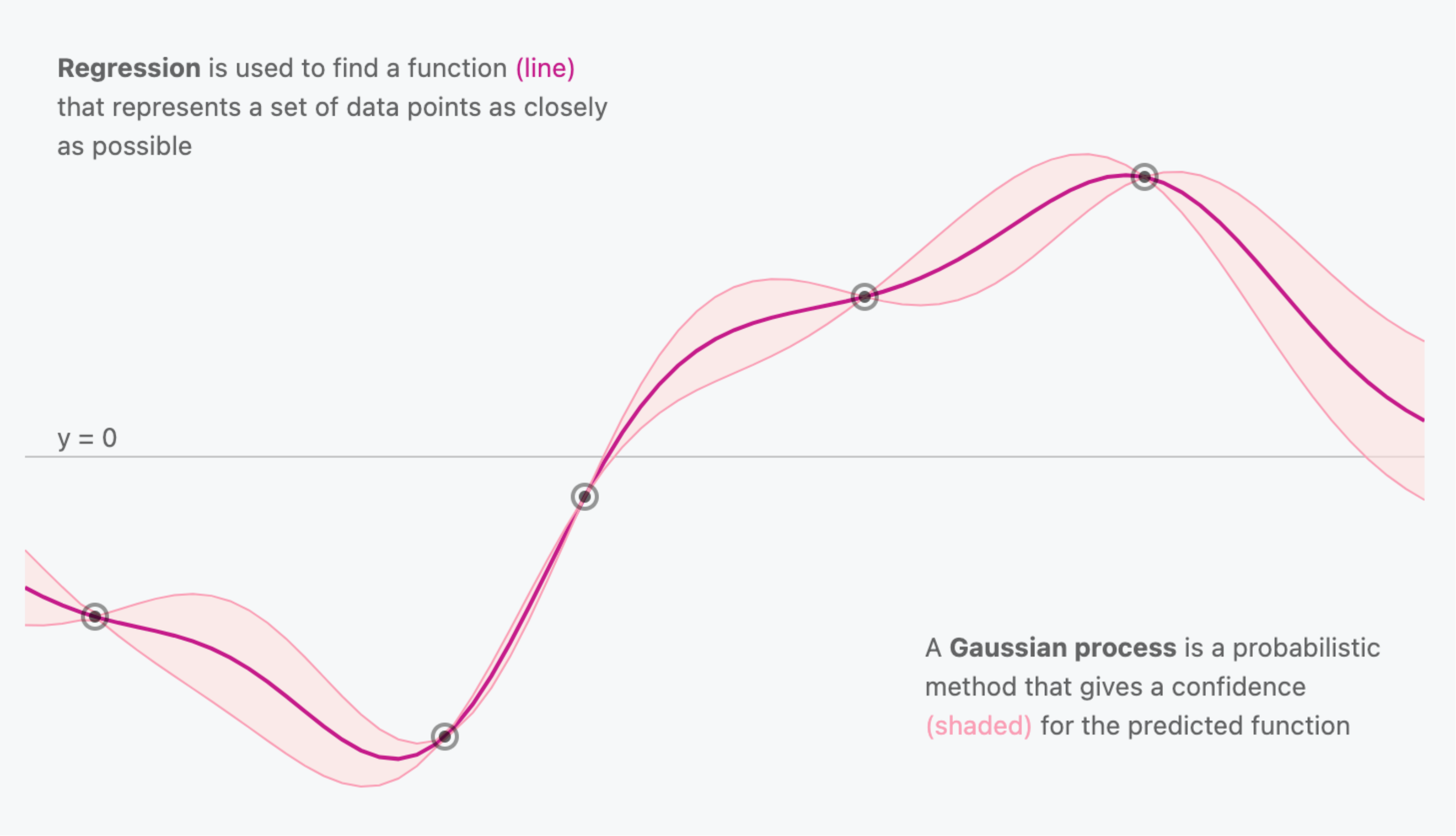


Bandits
Objectives
Objectives
Best Arms Identification
Regret Minimisation
Reward Summary: Quantile
Reward Model: Smoothness
QSAR
Function Optimisation
$f : \mathcal{X} \to \mathbb{R}$
Arms: RBS
Reward: aggregated
Reward Model: Neural Contextual
Bio Design
GPOO
News Recommendation
Arms: Hierarchical

Reward Smoothness – e.g. Gaussian Process

$$f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right)$$

$$\mu(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})] \quad \text{and} \quad k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - \mu(\boldsymbol{x}))(f(\boldsymbol{x}') - \mu(\boldsymbol{x}'))].$$



https://distill.pub/2019/visual-exploration-gaussian-processes/

Acquisition Function: e.g. GP-UCB

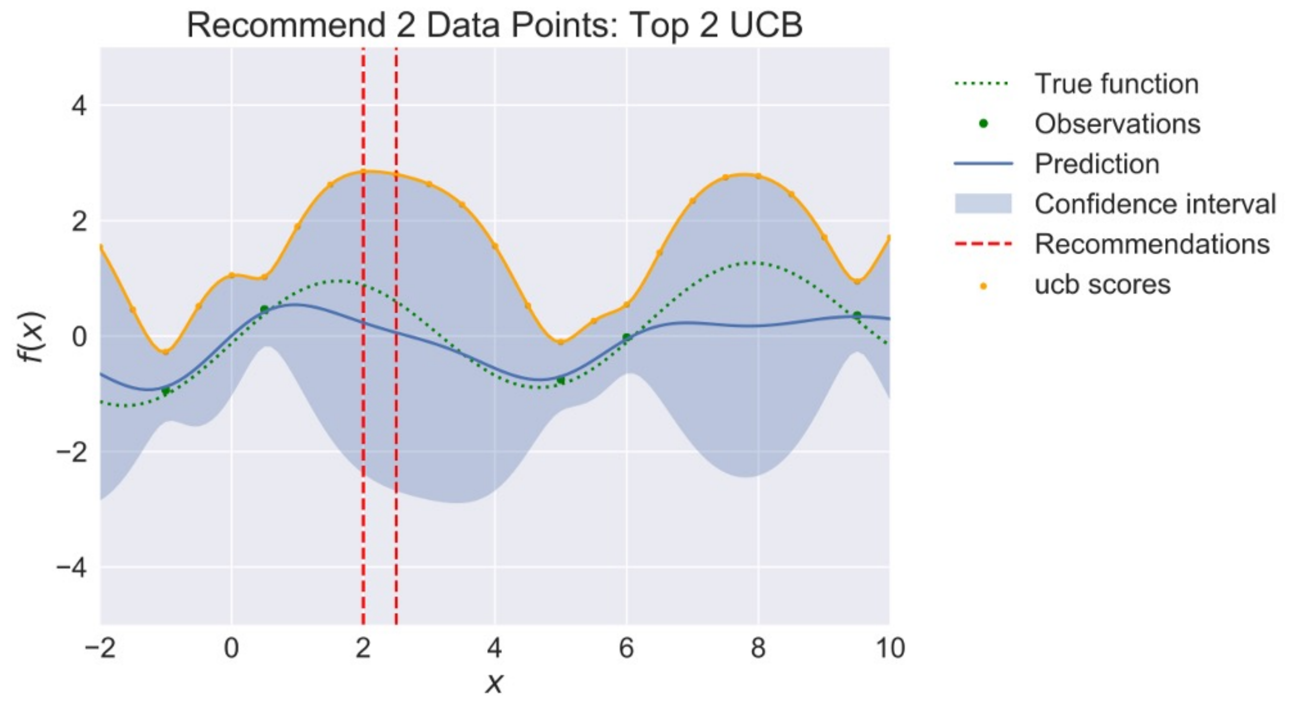
Gaussian Process Upper Confidence Bound (GP-UCB) [1]

Posterior mean: exploitation

Posterior standard deviation: exploration

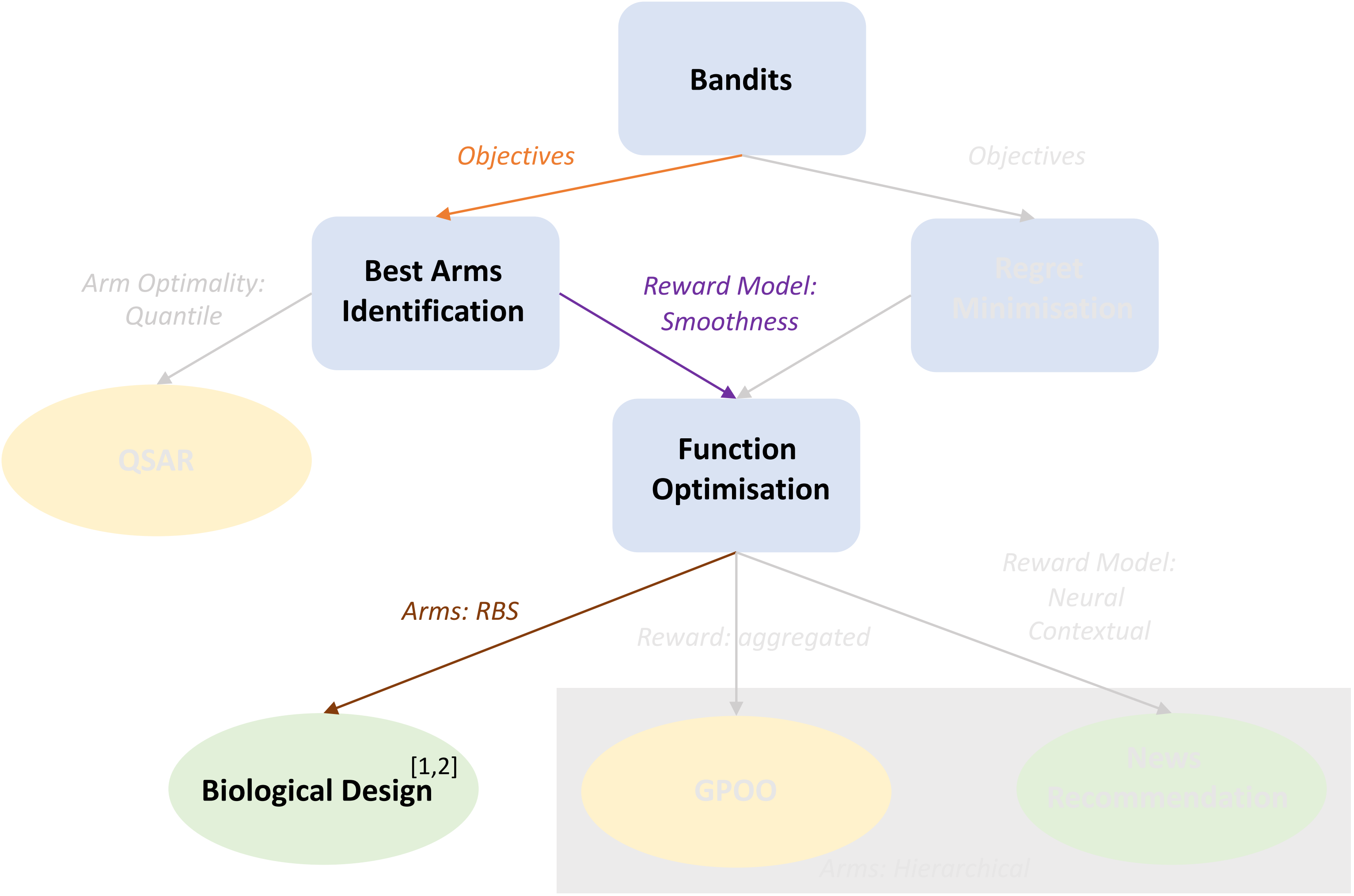
$$\operatorname{argmax}_{\mathbf{x}_i \in \mathcal{K}} \left(\mu_t(\mathbf{x}_i) + \beta_t \sigma_{t-1}(\mathbf{x}_i)\right)$$

Balancing term



[1] Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M., Gaussian process optimization in the bandit setting: No regret and experimental design. ICML 2009.

Outline



[1] Machine learning guided batched design of a bacterial Ribosome Binding Site.

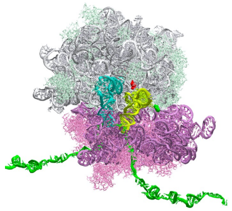
Mengyan Zhang, Maciej Bartosz Holowko, Huw Hayman Zumpe, Cheng Soon Ong. ACS Synthetic Biology Journal 2022.

[2] Opportunities and Challenges in Designing Genomic Sequences. **Mengyan Zhang**, Cheng Soon Ong. ICML Workshop on Computational Biology 2021.

Bandits for Synthetic Biology

With fixed budget (450), design **Ribosome Binding Site (RBS)** sequences in batches (4)
(300 for bandit groups)

Optimize the protein expression level (translation initiation rate)
Identify the RBS sequences with highest possible protein expression level



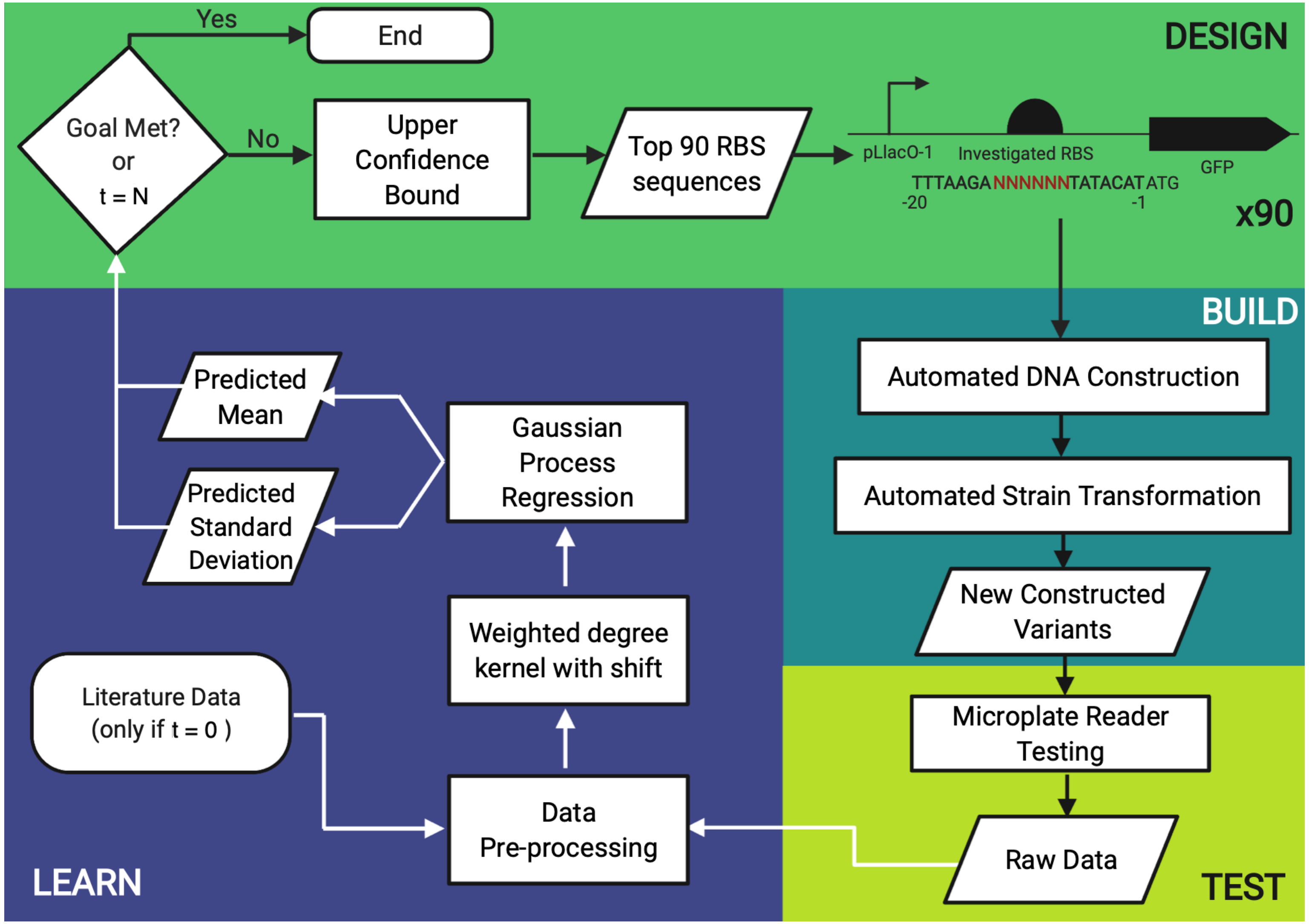
Arm: RBS sequence	Reward: Normalized* Translation Initiation Rate
TTTAAGAGTTATATATACAT	1.58
TTTAAGAATATGCTATACAT	1.42
TTTAAGACTCGGATATACAT	0.14
TTTAAGAGTTTTTTATACAT	2.88

- Design space: 4096 sequences

Green Fluorescent Protein (GFP)

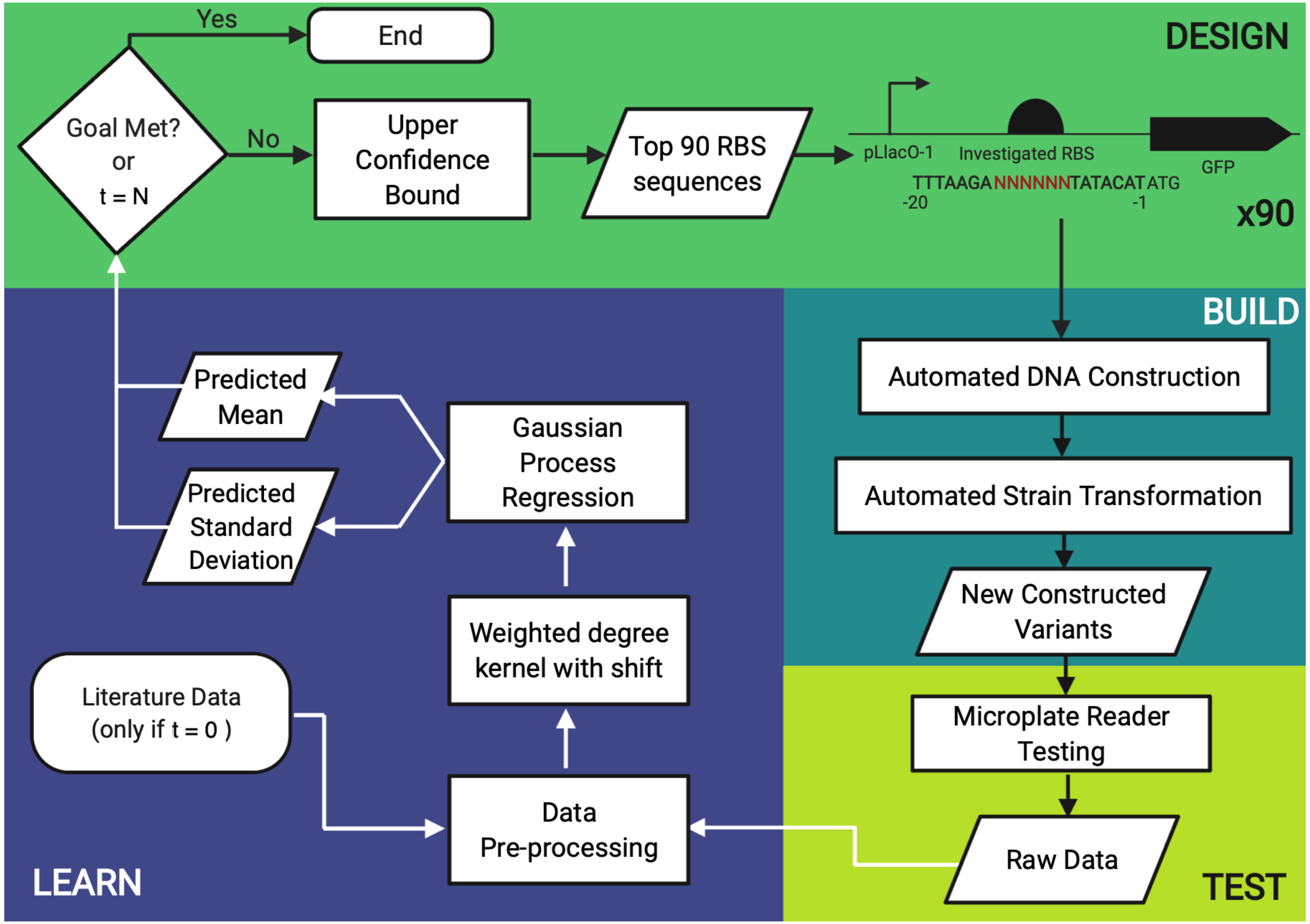
* zero mean and unit variance normalization $z = \frac{x-\mu}{\sigma}$

Design-Build-Test-Learn (DBTL) Cycle

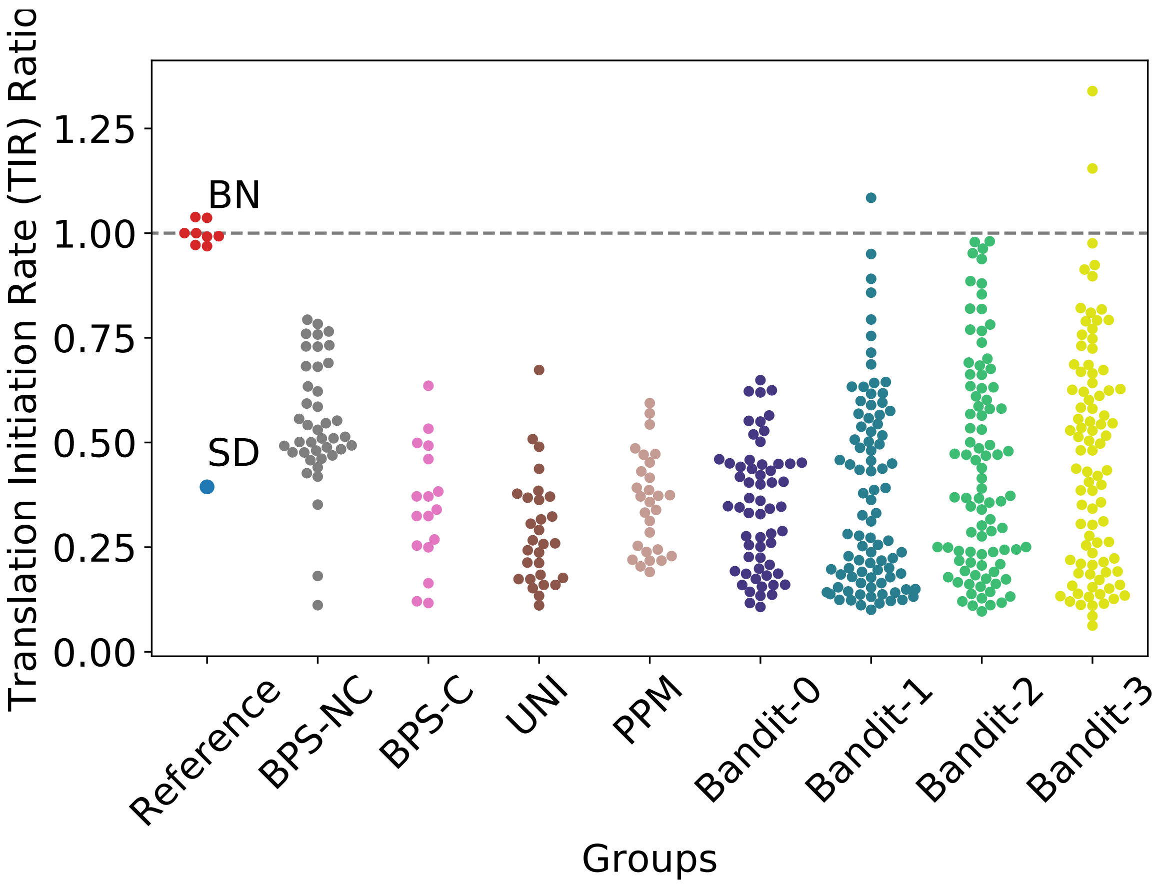


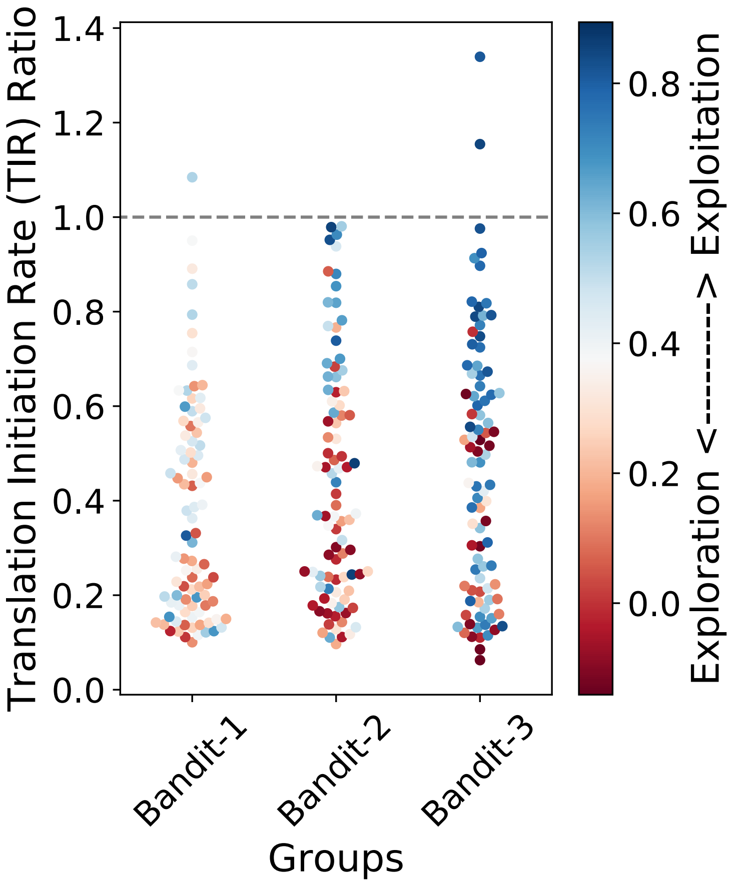
CSIRO BioFoundry Lab

Design-Build-Test-Learn (DBTL) Cycle



Results: swarmplot



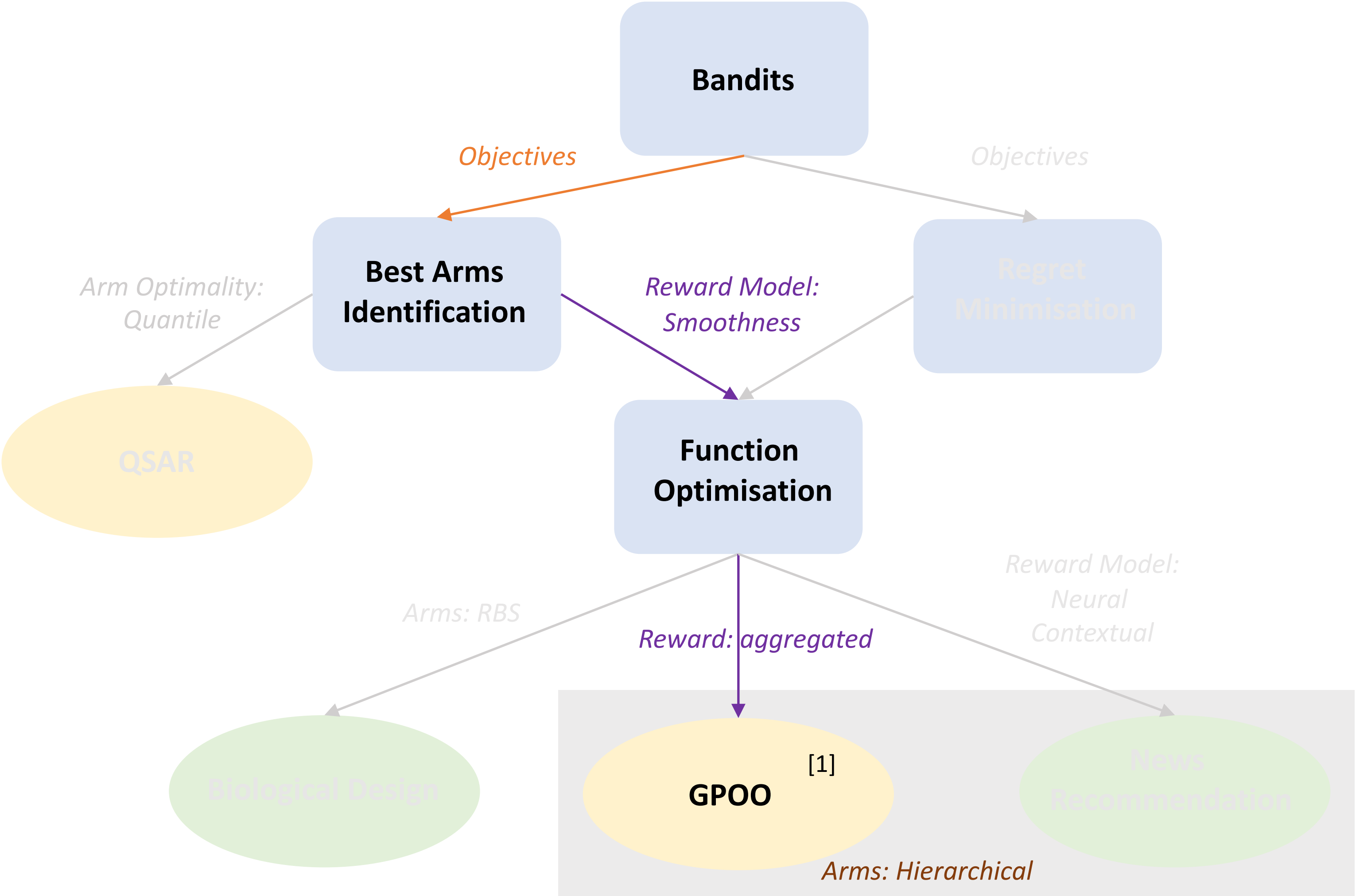


see more our ICML CompBio Workshop discussion.

Lessons learned and future opportunities

- ML (Bandits) guided DBTL cycle -- increase expression of our target protein by up to 35%, compared to a strong benchmark RBS
- Generalisation of our workflow: target on larger design space, more complicated genetic elements, e.g. promoters

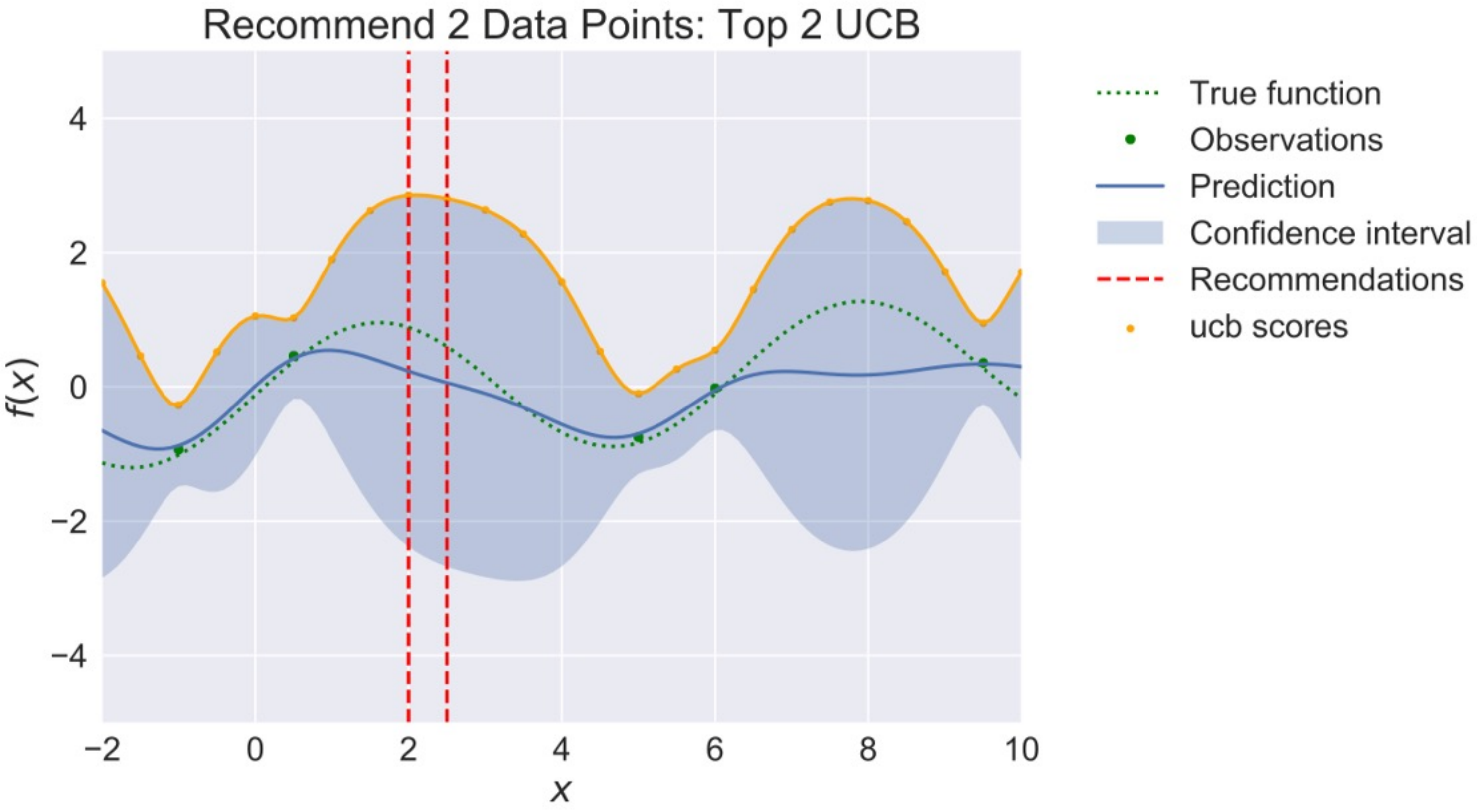
Outline



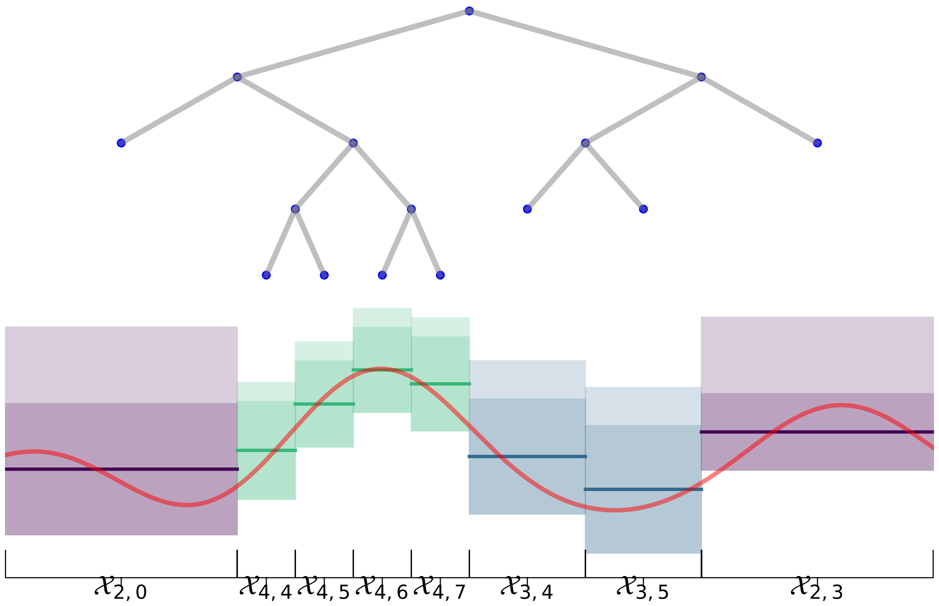
[1] Gaussian Process Bandits with Aggregated Feedback. **Mengyan Zhang**, Russell Tsuchida, Cheng Soon Ong. AAAI 2022.

Computational cost for continuous space

$$\text{argmax}_{\mathbf{x}_i \in \mathcal{K}} \left(\mu_t(\mathbf{x}_i) + \beta_t \sigma_{t-1}(\mathbf{x}_i)\right)$$



Problem Setting



- Arms: a leaf node, corresponding to a subset of continuous $[0,1]^d$
- Rewards: sampled from GP, only average reward for a node

$$r_t = \bar{F}(X_{h_t,i_t}) + \epsilon_t, \quad \bar{F}(X_{h_t,i_t}) := \frac{\sum_{\boldsymbol{x} \in \mathcal{C}_{h_t,i_t}} f(\boldsymbol{x})}{|\mathcal{C}_{h_t,i_t}|}$$

with **Representative points** $\mathcal{C}_{h,i} = \{\boldsymbol{x}_{h,i^s}\}_{1 \le s \le S}$, where $\boldsymbol{x}_{h,i^s} \in \mathcal{X}_{h,i}$.

Why Aggregated Feedback?

Application	Arm	Reward	Goal: to design a policy such that…	Why Aggregated?
DNA Design	DNA sequences	average protein expression level in a mixed culture	identify DNA sequences with the highest protein expression level with a given budget	expensive; search space is large
Census Querying	Respondent	average age of respondents inside queried area	identify the region with the highest average age with a fixed amount of querying	privacy concerns
Radio Telescope	spatial-frequency coordinates of objects in the sky	average radio wave energy from the queried area	identify the region with the highest average radio energy with a fixed amount of querying	hardware constraint

Gaussian Process Optimistic Optimisation (GPOO)

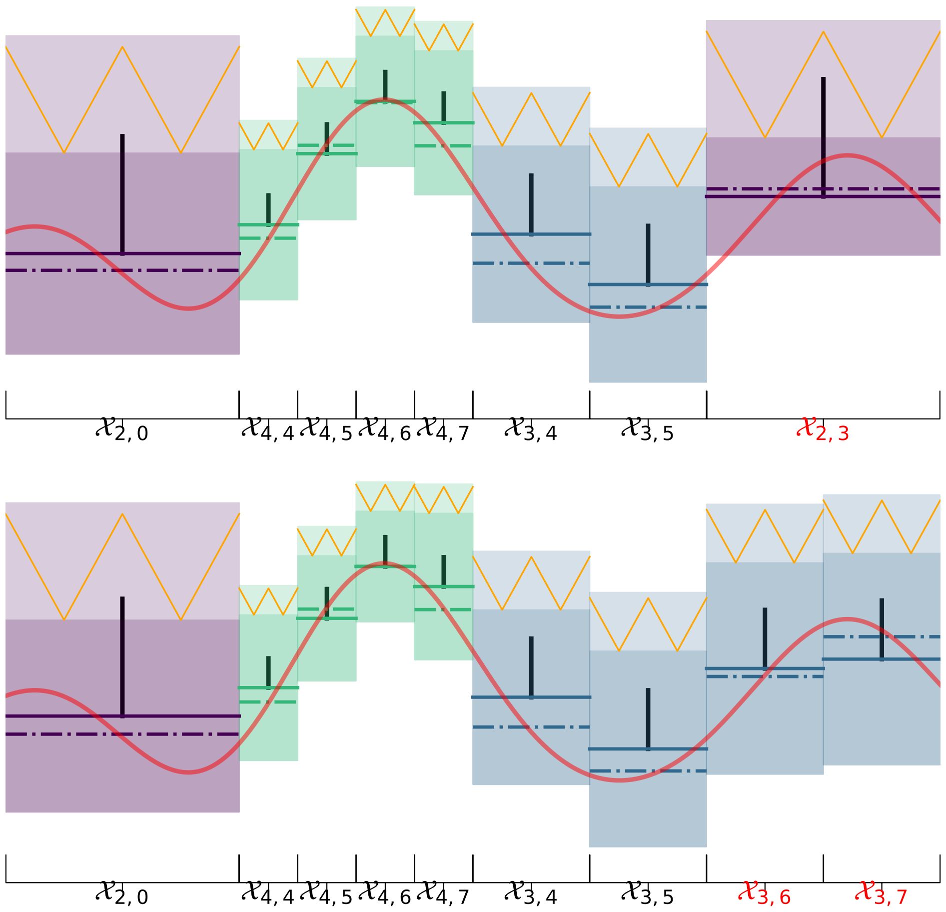
How to choose node and when to split?

Assumption: **Decreasing Diameters:** $\sup_{\boldsymbol{x} \in \mathcal{X}_{h,i}} L\ell\left(\boldsymbol{x}_{h,i}, \boldsymbol{x}\right) \leq \delta(h)$ some decreasing sequence $\delta(h) > 0$.

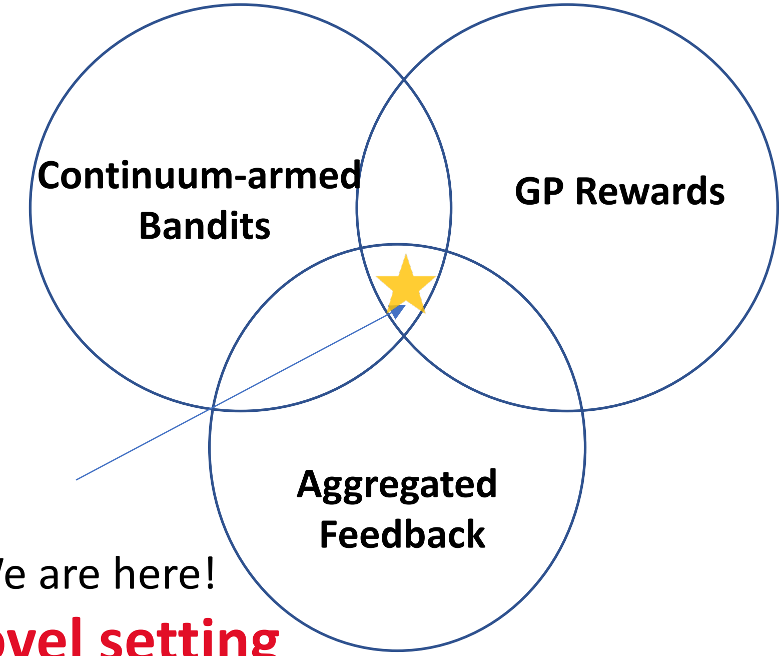
- **Select leaf node with largest b-value:**

$b_{h,i}(t)$ = posterior mean + confidence interval + diameter $\delta(h_t)$

- **Expand:** if $\delta(h_t)$ > confidence interval



Contributions of GPOO

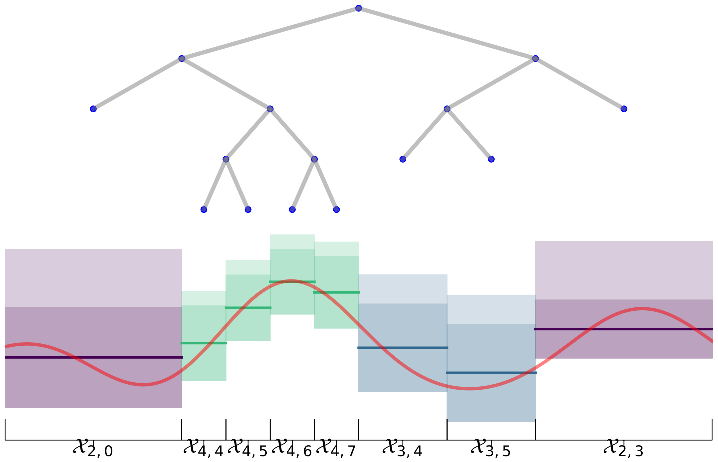


We are here!

A novel setting

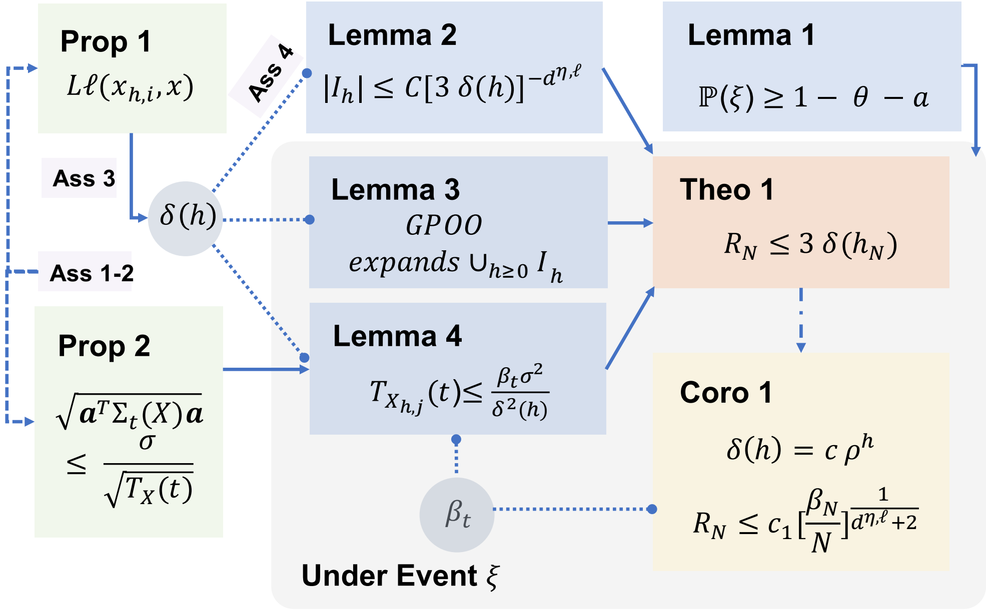
New Algorithm:

GPOO



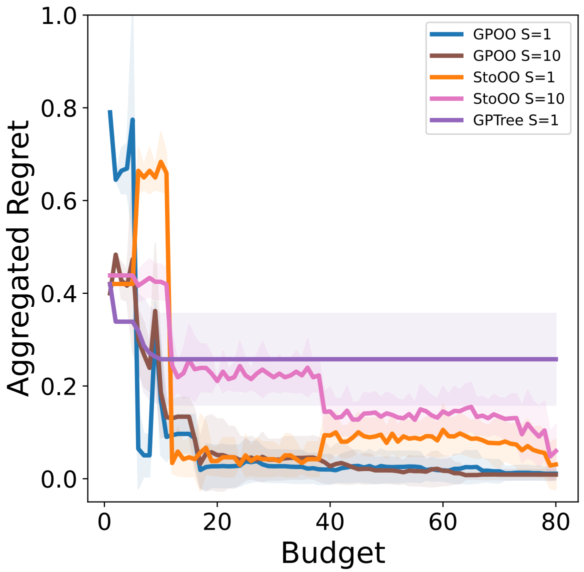
Theoretical results:

Upper bound on (aggregated) regret



Simulation results:

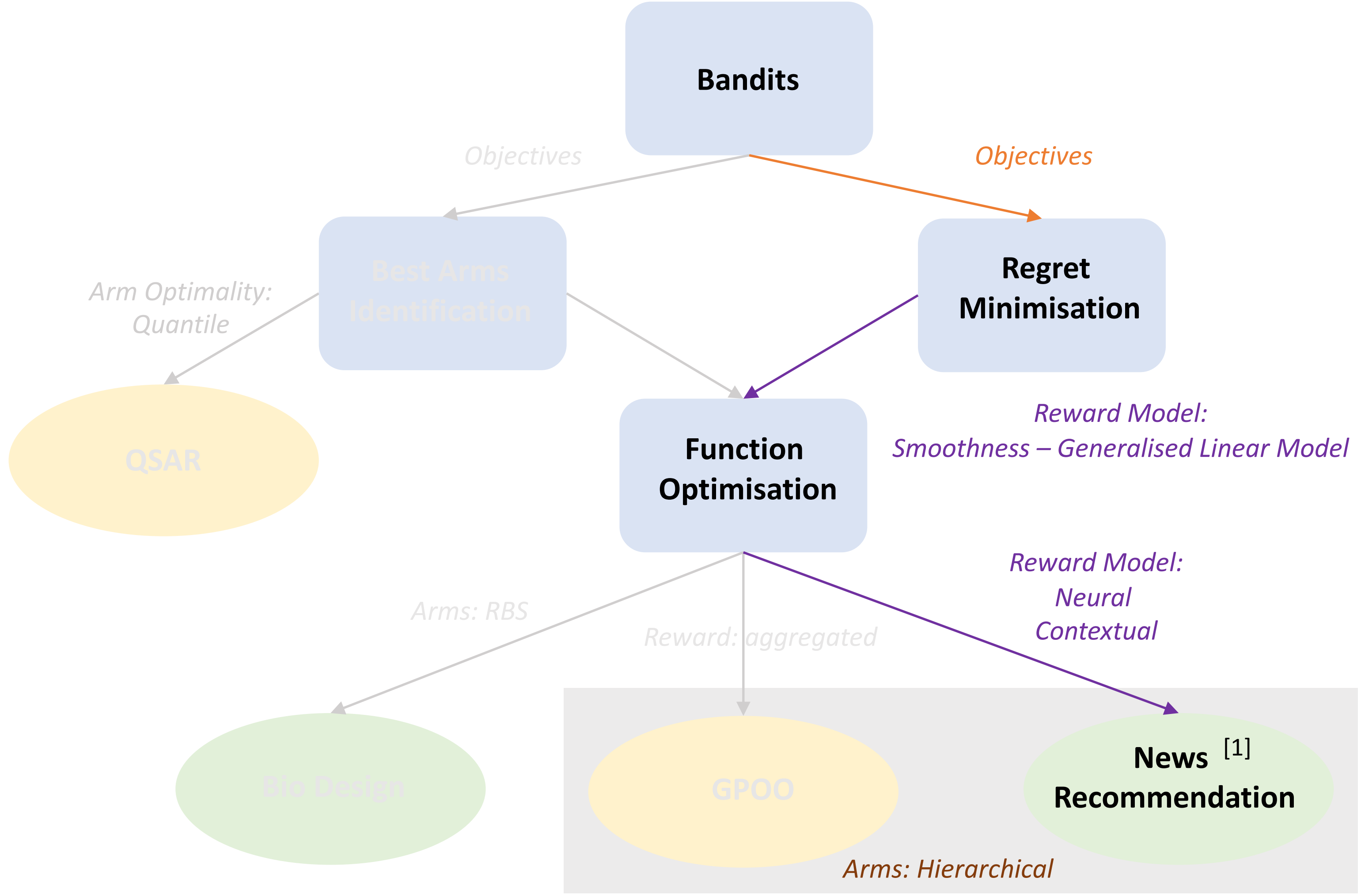
Outperform baselines



Lessons learned

- Hierarchical design of arms -> computational efficient for large/continuous design space
- Average reward feedback -> same regret upper bound rate as single arm feedback (in our setting)

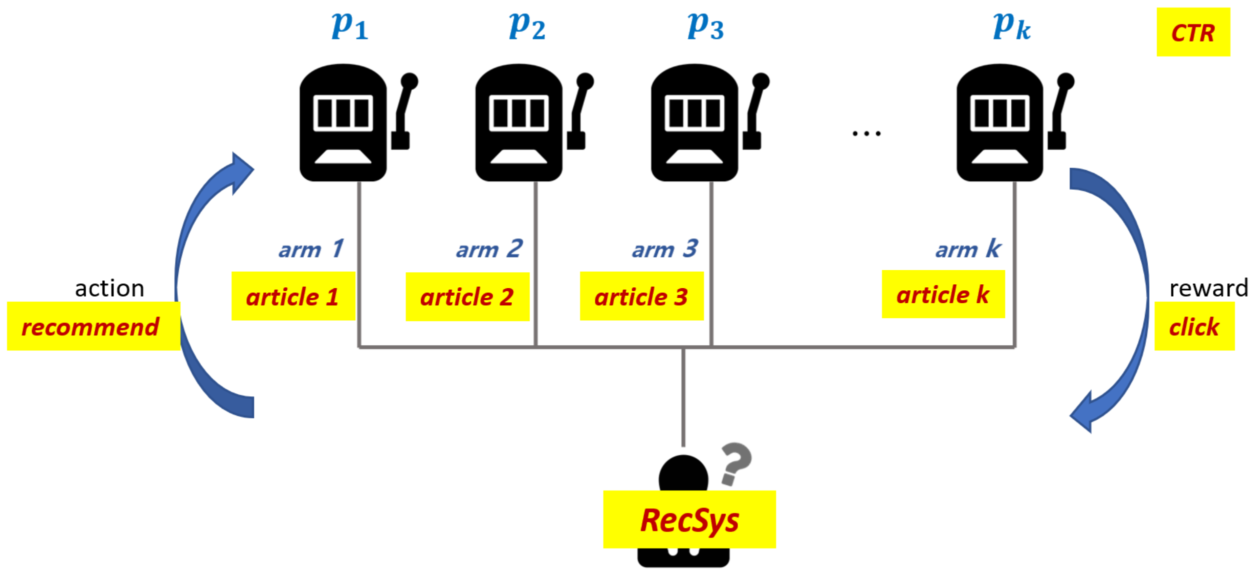
Outline



[1] Two-Stage Neural Contextual Bandits for Personalised News Recommendation.
Mengyan Zhang, Thanh Nguyen-Tang, Fangzhao Wu, Zhenyu He, Xing Xie, Cheng Soon Ong. Under Review 2022
(work conducted during the internship in Microsoft Research Asia)

Contextual Bandits:

in recommender system



In each round t ϵ {1, .., N},

Given a user

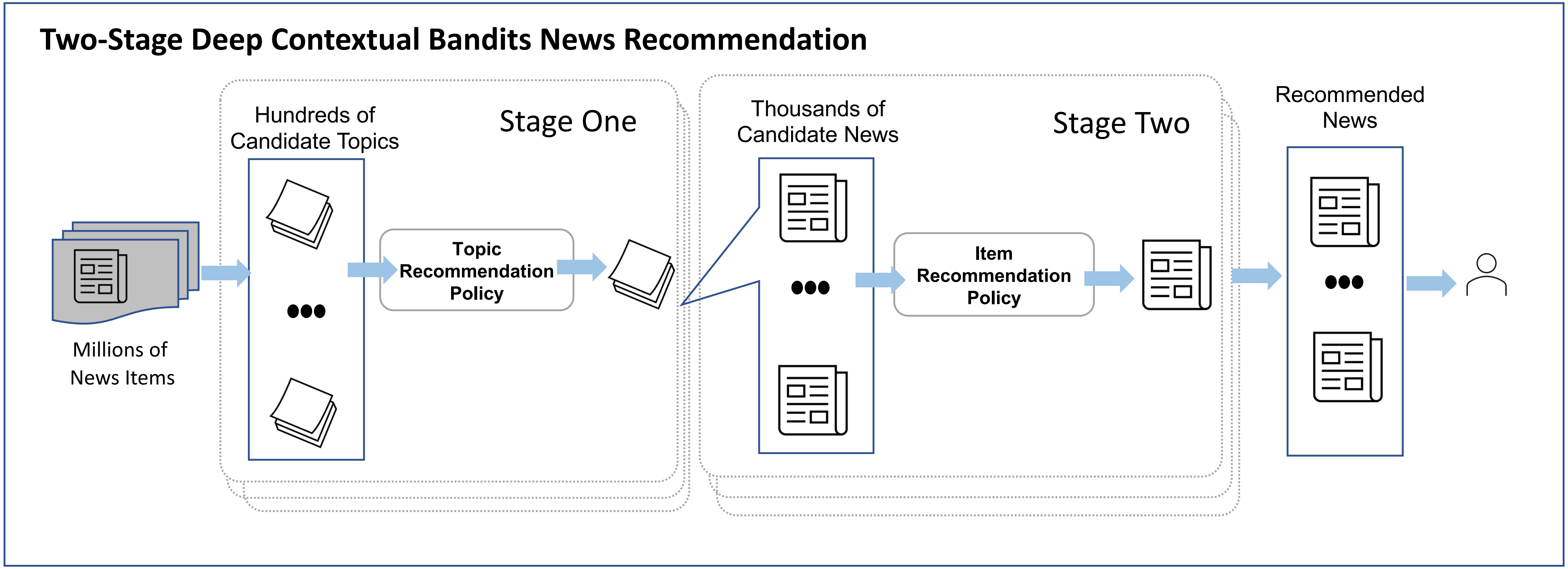
1. an agent selects an arm $A_t = i \; \epsilon \; \{1, .., K\}$ according policy π

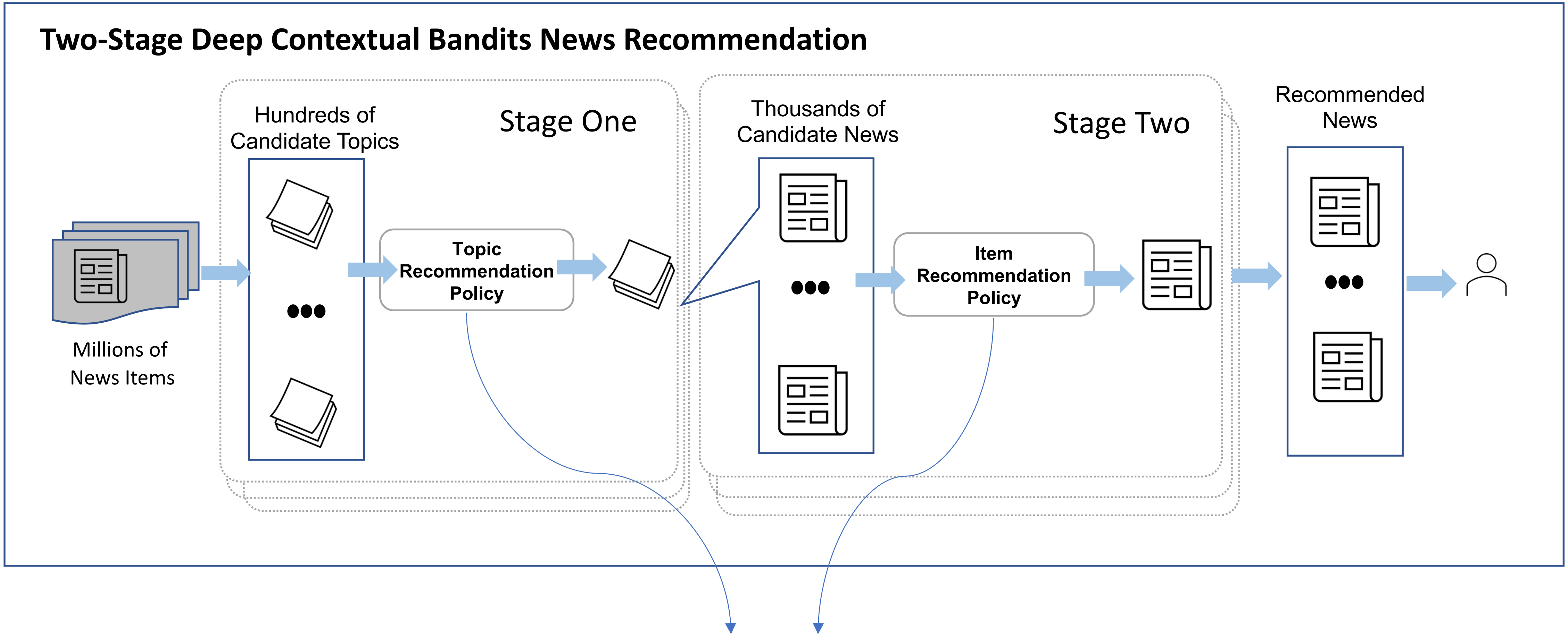
Recommender system   Item (e.g. news)   Recommendation strategy

2. then receive a reward $X_{A_t, T_{A_t}(t)}$ sampled from unknown reward distribution F_{A_t}

click; non-click

3. update estimations over reward distributions based on historical observations


Two-Stage Deep Contextual Bandits News Recommendation
Hundreds of Candidate Topics
Stage One
Topic Recommendation Policy
Thousands of Candidate News
Stage Two
Item Recommendation Policy
Recommended News
Millions of News Items



Neural Contextual Bandits Policies

Generalised Additive Linear UCB, Generalised Bilinear UCB

S-N-GALM-UCB S-N-GBLM-UCB

Large-scale experiments

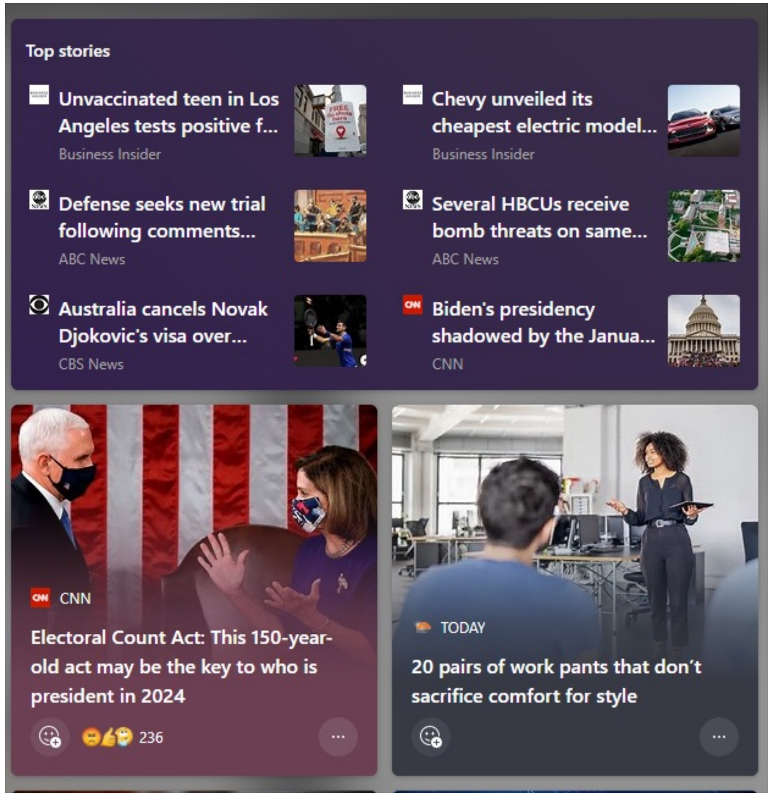
MIND: Microsoft News [1]

Users	News	Topics	Samples
1,000,000	161,013	285	24,155,470

Click-through-rate (CTR) $\mathrm{CTR}_t^\tau = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{y_t^\tau = 1\}$

Cumulative CTR $\sum_{t=1}^{N} \mathrm{CTR}_t^\tau$

Two stage/one stage

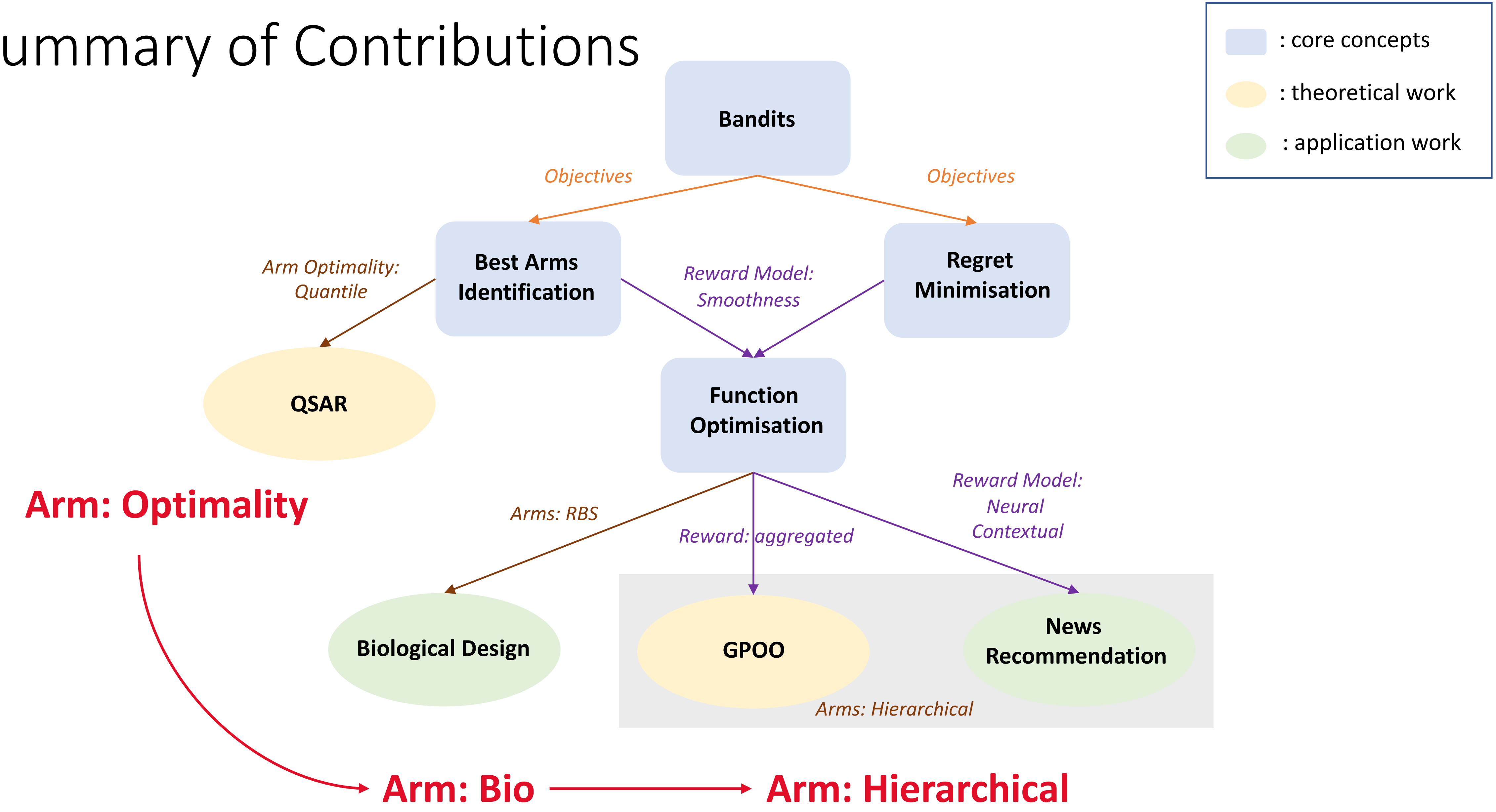


[1] https://msnews.github.io/

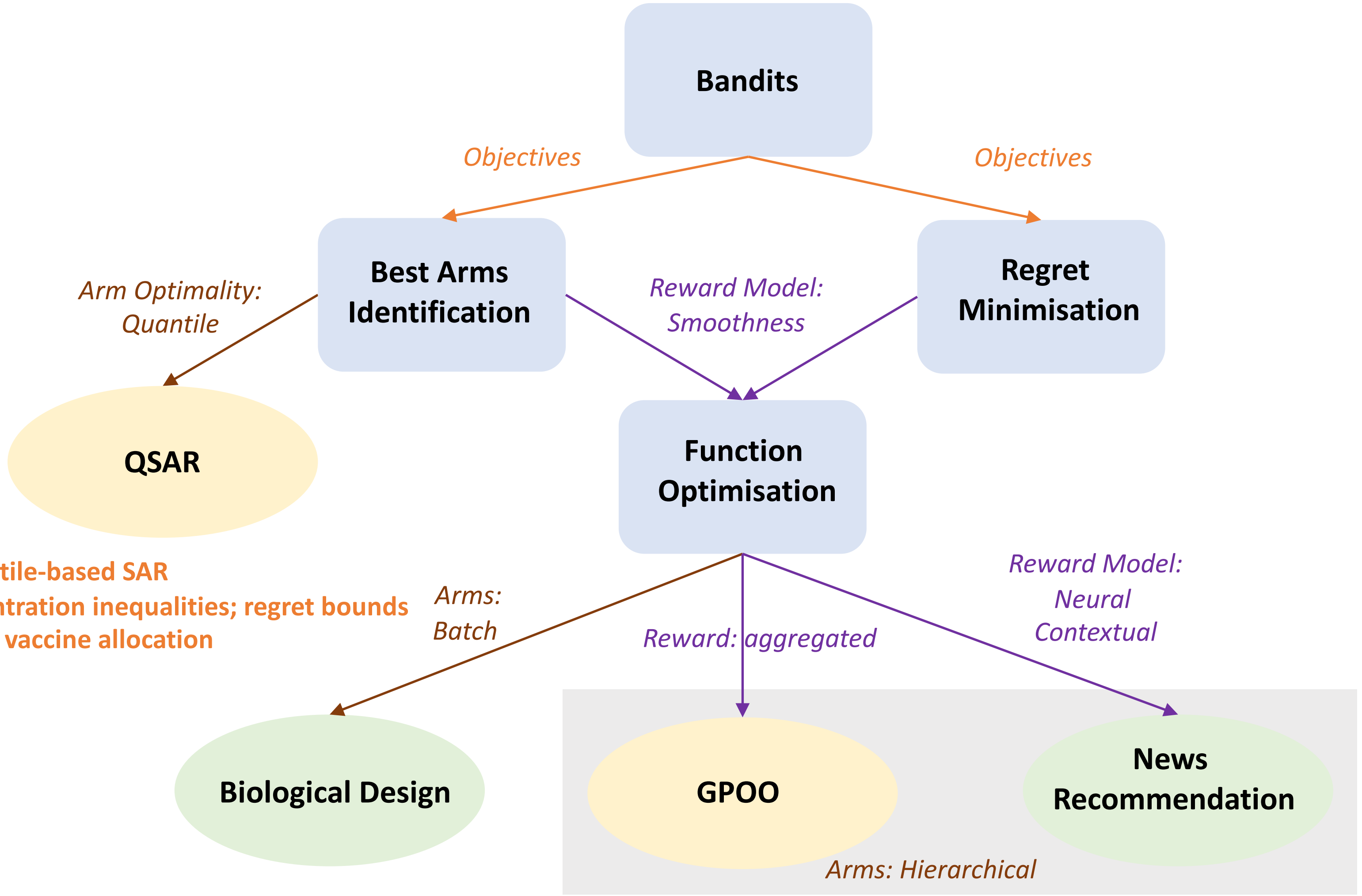
Lessons learned

- Two-stage recommendation is useful: computational efficient
- Two-tower neural representation improves the performance
- Gaps between theory and practice: neural bandits, off-policy evaluation

Summary of Contributions



Summary of Contributions



- New algorithm: quantile-based SAR
- New theories: concentration inequalities; regret bounds
- Empirical evaluation: vaccine allocation

- Real-world applications on synthetic biology
- ML-guided DBTL cycle for biological batch design
- 35% improvement on TIR

- Novel settings: aggregated feedback
- New algorithms: GPOO
- Theoretical analysis: regret upper bound
- Empirical evaluation: simulation

- Large-scale experiments on news recommendation
- Two-stage topic-item recommendation framework
- Neural Contextual Bandits policies

Publications

- Quantile Bandits for Best Arms Identification.

  Mengyan Zhang, Cheng Soon Ong. International Conference on Machine Learning 2021.

- Machine learning guided batched design of a bacterial Ribosome Binding Site.

  Mengyan Zhang, Maciej Bartosz Holowko, Huw Hayman Zumpe, Cheng Soon Ong. ACS Synthetic Biology Journal 2022.

- Opportunities and Challenges in Designing Genomic Sequences.

  Mengyan Zhang, Cheng Soon Ong. ICML Workshop on Computational Biology 2021.

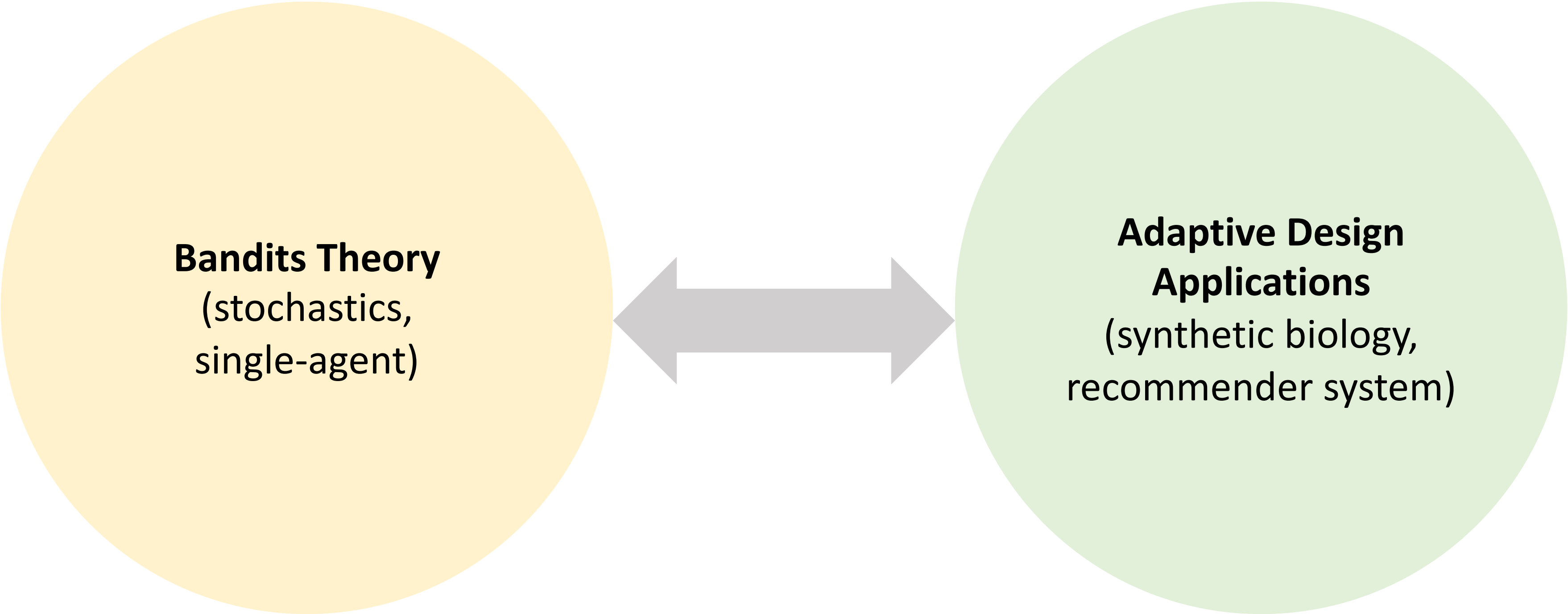
- Gaussian Process Bandits with Aggregated Feedback.

  Mengyan Zhang, Russell Tsuchida, Cheng Soon Ong. AAAI 2022.

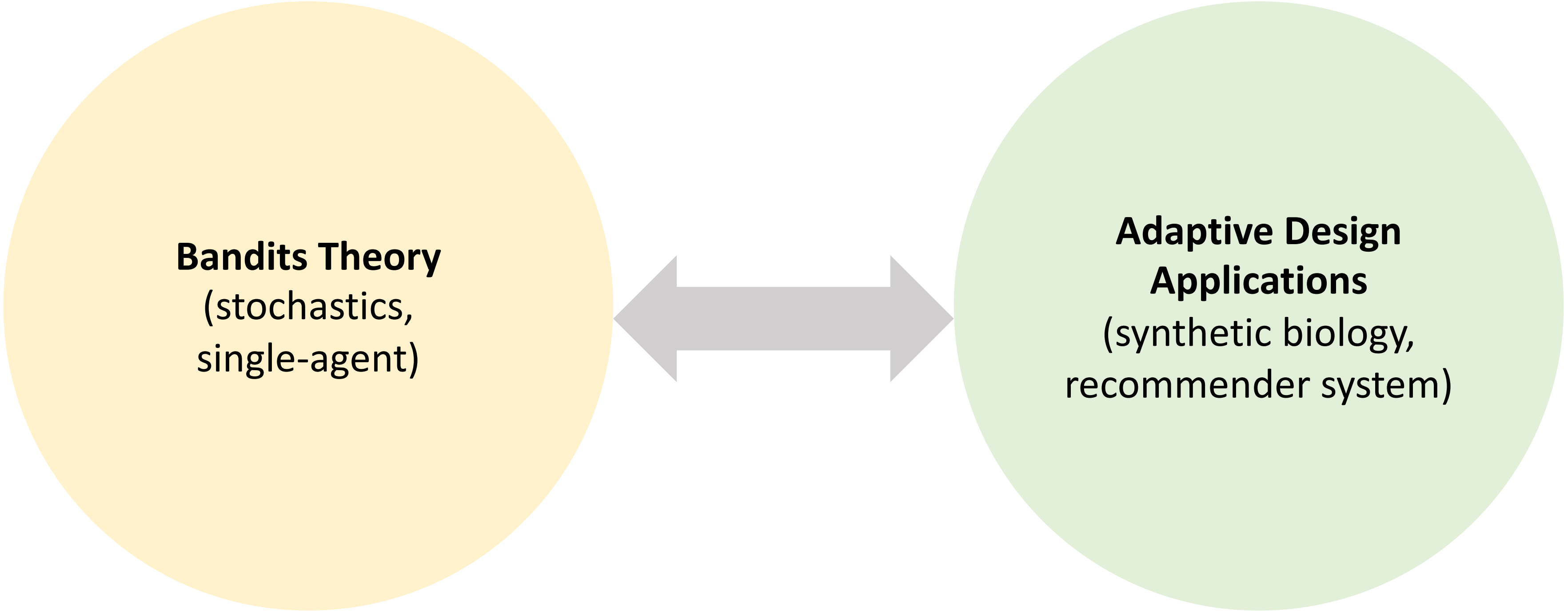
- Two-Stage Neural Contextual Bandits for Personalised News Recommendation.

  Mengyan Zhang, Thanh Nguyen-Tang, Fangzhao Wu, Zhenyu He, Xing Xie, Cheng Soon Ong. Under Review 2022.

Future Work – Gaps between Theory and Practice



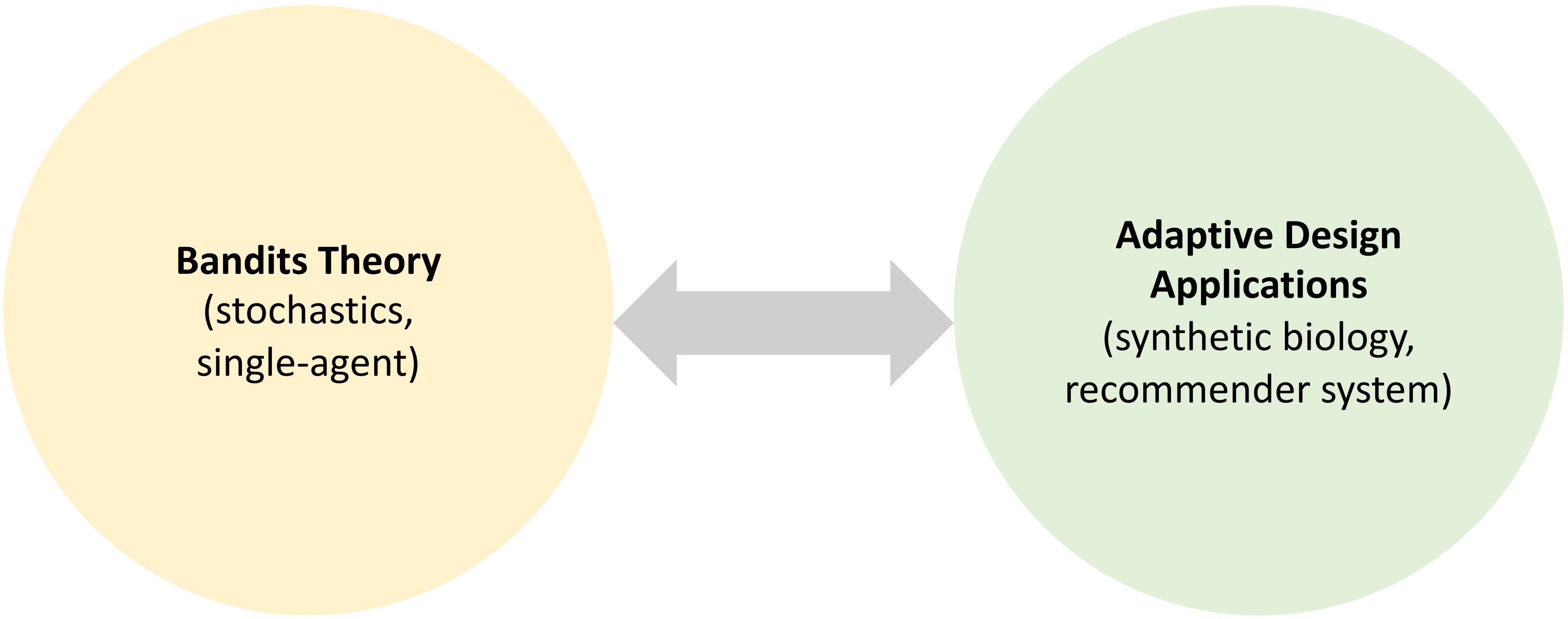
Future Work – e.g. Off Policy Evaluation



Casual Inference – unbiased simulator

Data is collected based a specific policy
How to evaluate another policy?

Future Work – Scientific Discovery



Call to action:

“scientific revolutions occur when there is cross pollination of ideas”

---- Thomas Kuhn

Acknowledgements

- **Supervisory panel**:

Cheng Soon Ong

Lexing Xie

Eduardo Eyras

- **Mentors and Collaborators**:

  Sebastien Bubeck (Microsoft Research), Xing Xie (Microsoft Research Asian), Fangzhao Wu (Microsoft Research Asian), Russell Tsuchida (CSIRO), Maciej Bartosz Holowko (CSIRO), Thanh Nguyen-Tang (Deakin University), Huw Hayman Zumpe (CSIRO), Zhenyu He (UECSTC) and many others.

- **CMLab and ML research group folks**: for interesting discussions, useful suggestions and feedback.

- **Family and friends**: for unconditional support and care.

Thanks for listening!

Mengyan Zhang
mengyan.zhang@anu.edu.au
https://mengyanz.github.io/