LG SDC 2022
CONNECT GROW OUTSTAND

Track #2. SW전문가 Session

데이터 기반 강화학습을 통한 세탁기 탈수 성능 개선

강찬석 선임
LG전자 ICT센터 인공지능연구소 제어지능TP

In >seed {2018}...

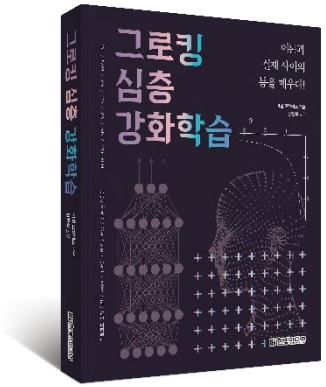

{2018} LG 소프트웨어 개발자의 날
시 : Gridworld
pplying Q-Learning
Episode
>seed {2018}
LG
CLOUD HOSTING

개인 약력

- **LG**전자 **ICT**센터 인공지능연구소 제어지능**TP**

- **LG**전자 **AI Specialist**

- 이력
  - **2018**년 차세대표준(연) **5G Base Platform** 개발 및 검증
  - **2020**년 인공지능(연) 냉장고 부하대응 정온 정밀제어 **PoC** 개발 및 검증
  - **2022**년 인공지능(연) 세탁기 강화학습 알고리즘 자동화 개발 및 검증



- 감수/번역
  - "텐서플로를 활용한 머신러닝" (한빛미디어, **2018**년)
  - "그로킹 심층 강화학습" (한빛미디어, **2021**년)

조직 소개

- **LG**전자에도 인공지능을 범용적으로 연구하는 조직이 있다**??**

- **ICT**센터 인공지능**(**연**)** 제어지능**TP**
  - **Mission**

제어(control) + 지능(Intelligence)

  - **Rule & Responsibility**

    제어를 필요로 하는 다양한 **Domain**에 인공지능을 접목시키는 업무 수행

    (세탁기**,** 냉장고**,** 에어컨**, BEMS** 등**….)**

세탁기

- 자사 대표 생활가전 중 하나

- 다양한 기술 **Domain**의 집합체
  - 세탁물의 관성 및 원심력
  - 내부 유체(물)의 움직임
  - 물리적 모터 제어
  - 전체 동작을 제어할 **SW**
  - **…**

- **실제 세탁기** 기반으로 **수많은 연구원**들의 **반복된 실험**을 통한 성능 개선

⇨

- **인공지능** 기반으로 **소수의 인원**으로 **빠르게** 성능 개선**?**

세탁기

- 탈수

Source : FYPhysics

목표

- 실환경 데이터 기반 모델 성능 향상 기술 개발

  - 데이터 기반 학습 알고리즘 개발
  - 모델 성능 개선 기간 단축

  } 데이터 기반 모델 성능 개선을 위한 강화학습 원천기술 확보

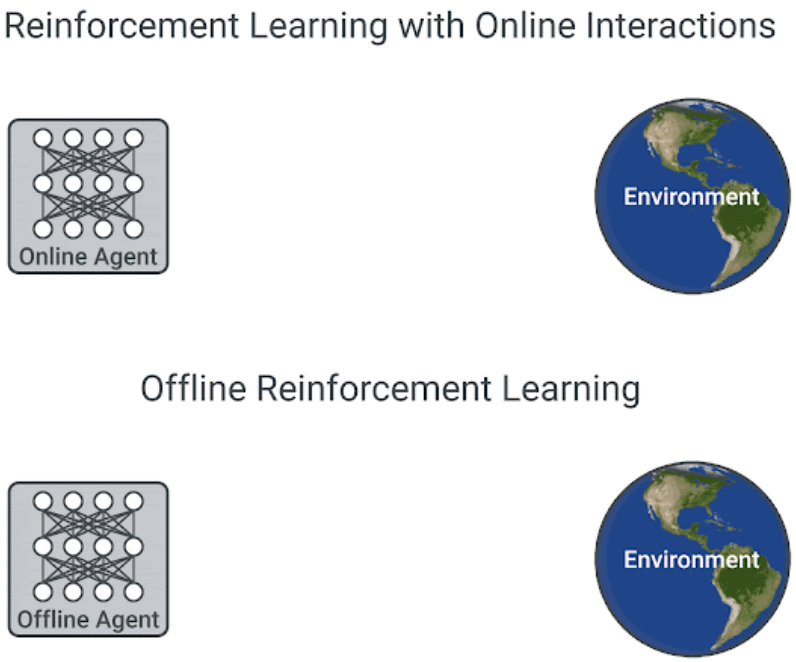
  - 데이터 기반 실환경 적응기술 개발 → 고객 환경 맞춤형 모델 업그레이드를 위한 선행 기술 탐색

⇩

Offline RL!

Offline RL

- **Data-Driven RL method**



- **Online RL**
  - 환경과 **interact**할 **policy**(π_β)와 학습할 **policy** (π_θ) 일치 여부에 따라서 **on-policy/off-policy**로 나눠짐
  - **Exploration, Data-efficiency, …**
- **Offline RL**
  - 환경과는 **interaction**없이 특정 **policy** (π_β)가 쌓은 **dataset**내에서 **policy**(π_θ)를 학습시키고 이를 환경에 **deploy**함
  - **Off-policy Evaluation (& Hyperparameter tuning), Over-Conservatism, …**

[1] https://ai.googleblog.com/2020/04/an-optimistic-perspective-on-offline.html

Offline RL

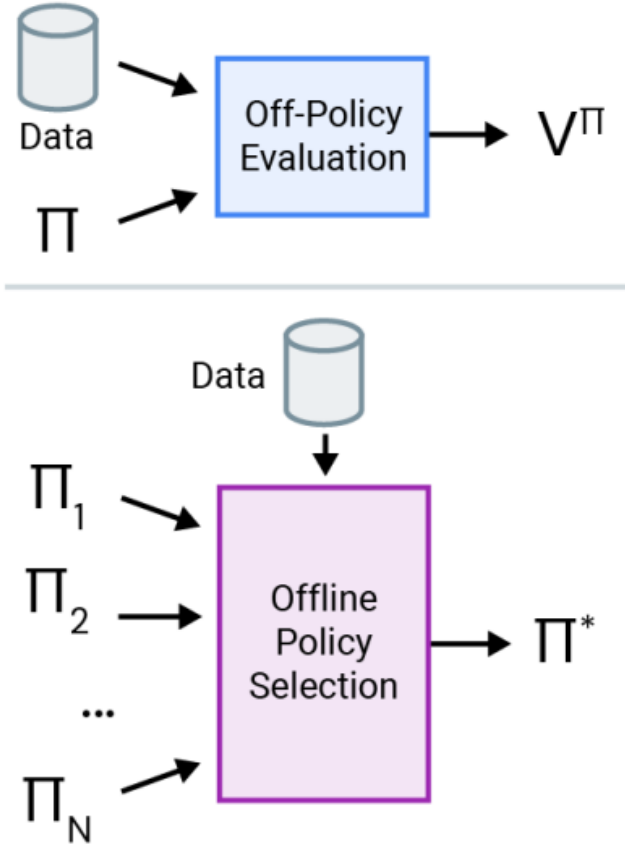
- 궁극적인 목표

$$arg\max_{\pi} \sum_{t=0}^{T} E_{S_t \sim d^{\pi}(s), a_t \sim \pi(a|s)}[\gamma^t r(s_t, a_t)]$$

[2] https://ai.googleblog.com/2020/04/an-optimistic-perspective-on-offline.html

Offline RL

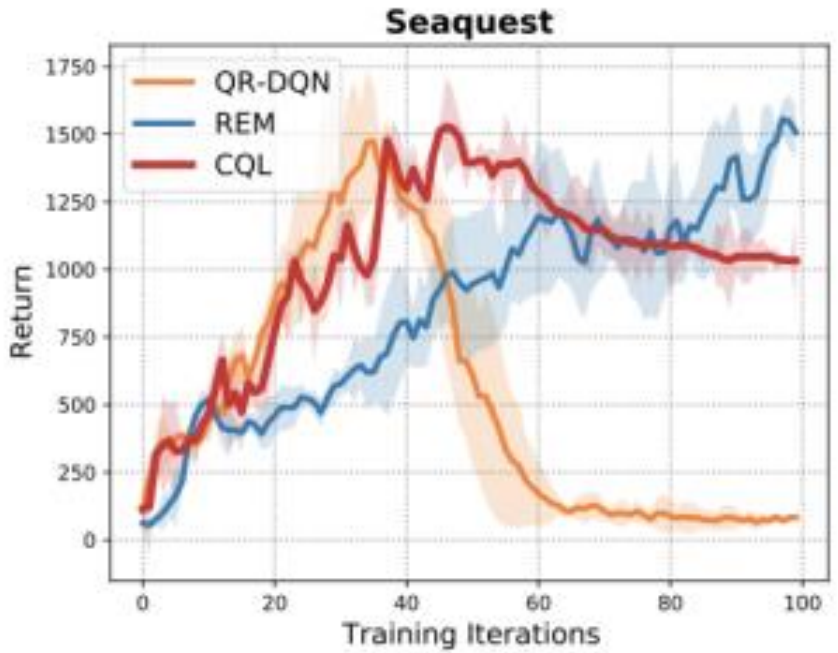
- **Off-Policy Evaluation (OPE)**
  - **Behavior policy**와 **Target policy**간의 불일치한 상황에서 해당 **policy**의 좋고 나쁜 정도를 판단해야 함
    → V^{π}
  - 후보 **policy** 집단 중에서 최적의 성능을 보이는 **policy**를 찾는 **policy Selection** 과정이 필요
    → π^{*}



[3] BENCHMARKS FOR DEEP OFF-POLICY EVALUATION

Offline RL

- **Limitation**
  - 최적의 **policy**를 찾는 과정에서 다양한 **hyperparameter combination**들이 나옴 **(Base algorithm, Training iteration, Reward function, …)**
  - **Offline RL**에서의 성능 개선은 위해서는 반복적인 실험은 위한 **Simulator**나 **OPE**용 **Dataset**의 존재가 필요



그런데 세탁기는 **Simulator**가 없네.....?

[4] Offline Reinforcement Learning: From Algorithms to Practical Challenges

문제 정의

- **Offline RL in** 세탁기 탈수

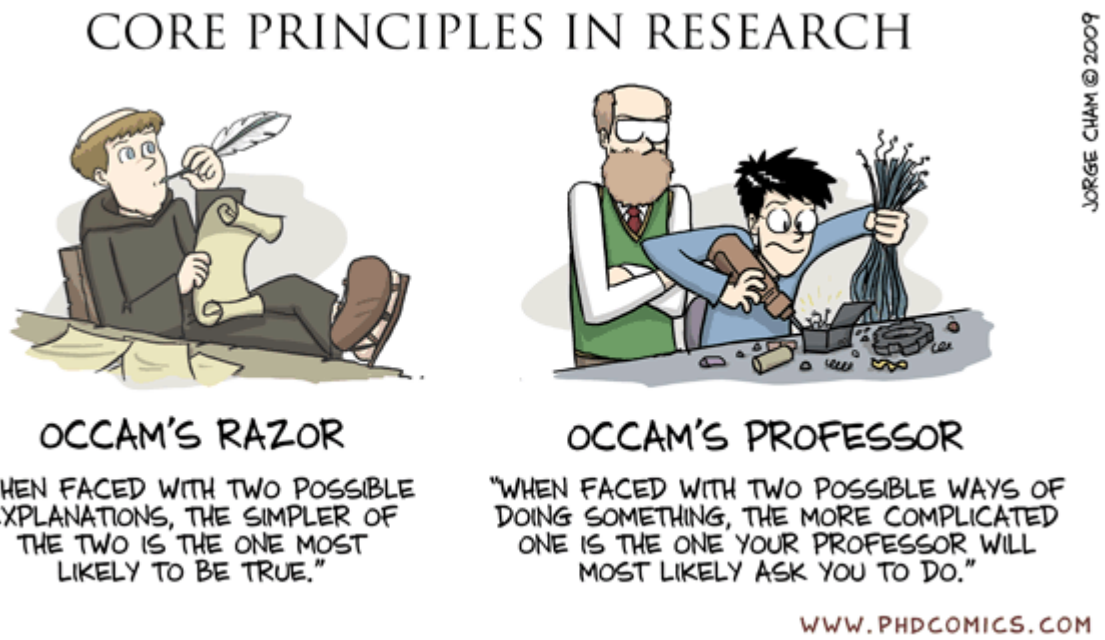
1. **[Data Collection] Online RL**은 동해 데이터를 처음부터 수집
2. **[Model Training]** 데이터도부터 샘플딩한 **Batch**도 **Offline RL** 수행
3. **[~~Off-policy~~ Online Evaluation]** 여러 후보군 모델을 **test case**에 넣어서 성능 추춛
4. **[Hyperparameter Selection]** 성능이 잗 나오는 **combination set** 탐색
5. 과정 반복

- 고려 사항
  - 실환경 실험에 최적화된 **Evaluation** 방법이 있을까**?**
  - **Evaluation** 시간을 단축핟 수 있을까**?**
  - 실제 실험하지 않은 케이스에 대한 안정성 보장**?**

접근 방법

- Occam's Razor

“Simpler solutions are more likely to be correct than complex ones.”

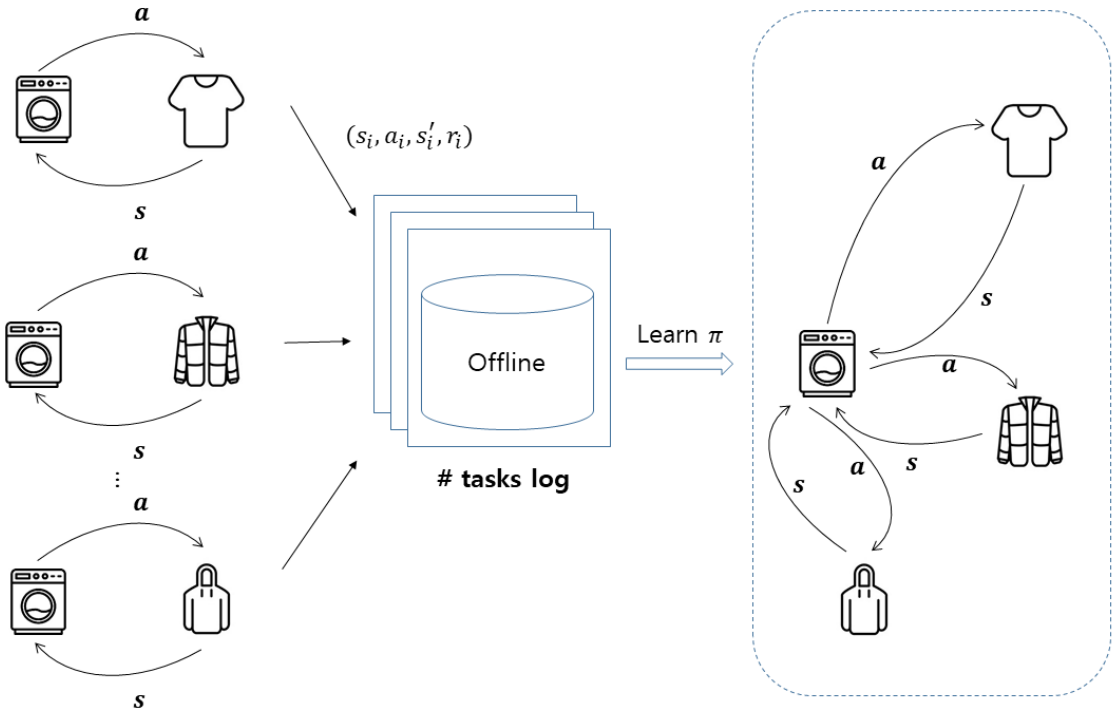


접근 방법

- **Multi-task Fusion**

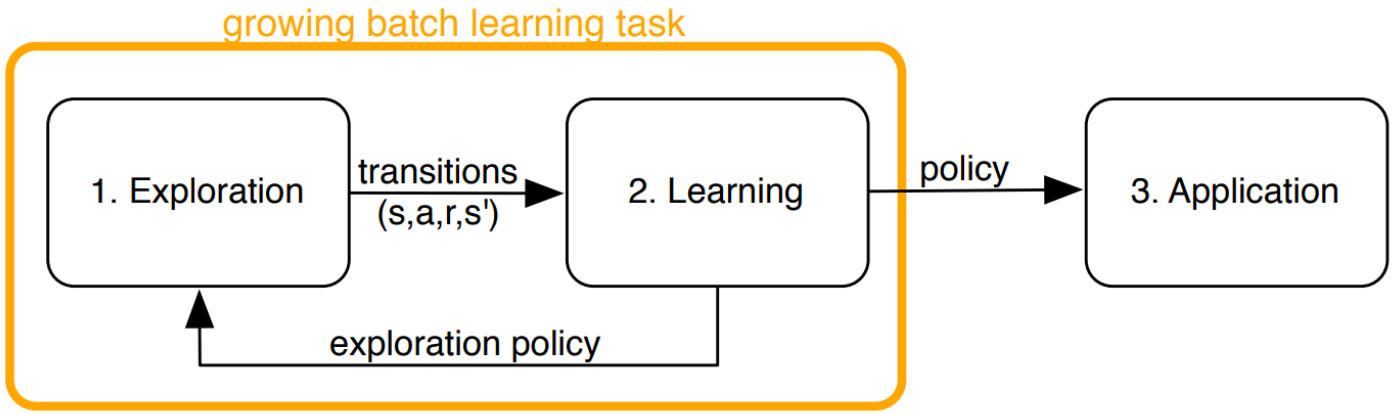
(a) T-Shirt (b) Jean (c) JumperA Towels (d) JumperB Towels (e) Towel (f) Jeans Towel

- 사업부에서 **QA**시험시 활용하는 세탁포 중 대표성을 띄는 **6**개 포 대상
- **Online RL**도 학습시 생성되는 **log**도 **Offline RL** 수행
- 포별 구분 없이, 하나의 동합 모델을 만드는 것!



접근 방법

- **Growing Batch Learning Approach**



  - **Offline RL process** 중 **policy selection**시 생성된 **log**를 다시 **offline RL** 학습에 사용함

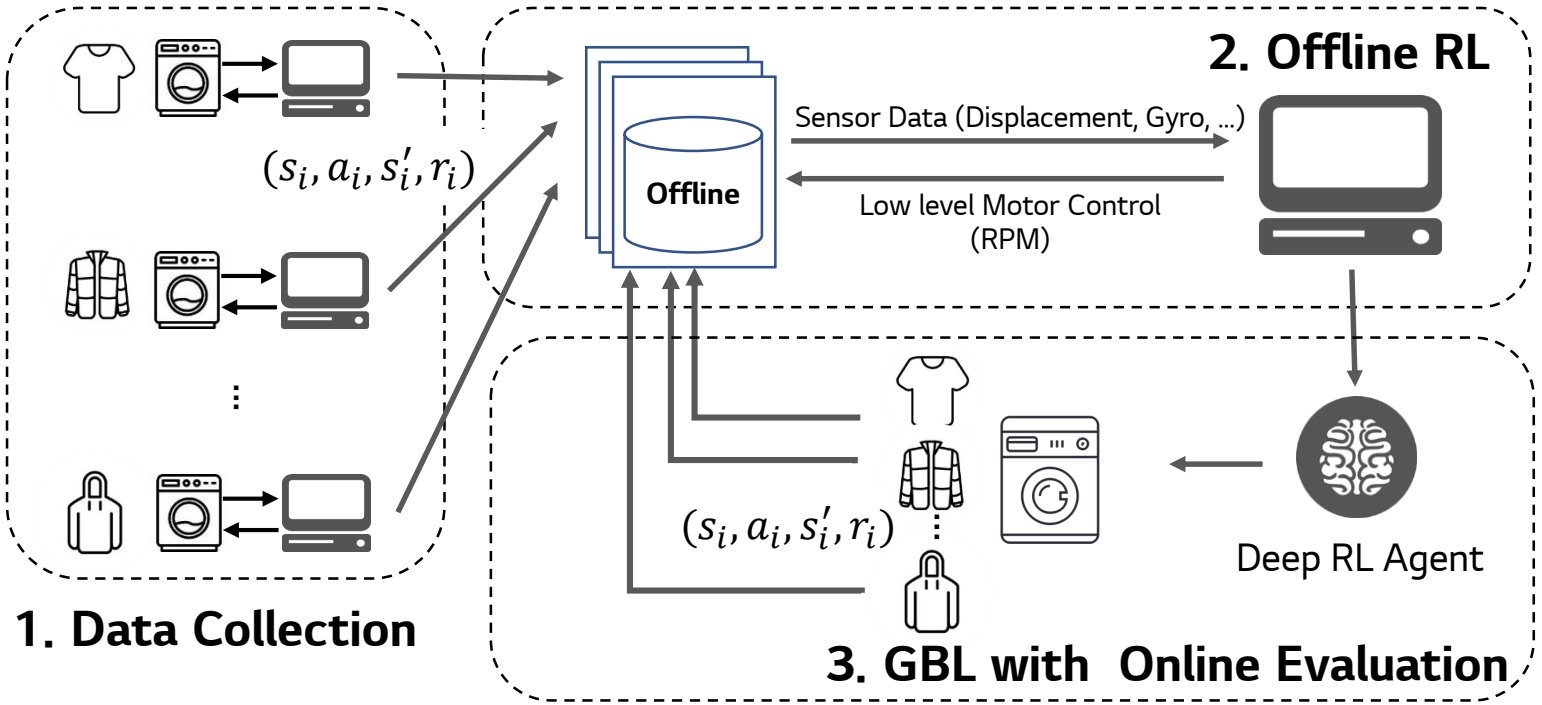
    → 새도운 **Data**를 동한 **exploration coverage** 확보

  - **Online Evaluation**에서만 성능 측정이 가능한 **Real-world problem**에서의 현실적인 방법

[4] Batch Reinforcement Learning

접근 방법

- **Multi-task Growing Batch Learning (MTGBL)**



접근 방법

1. **[Data collection]**
   6개의 세탁포를 각각 넣은 세탁기로 **Online RL** 수행

2. **[Offline RL]**
   학습시 생성된 **log**로 후보 **model**들을 생성

3. **[Online Evaluation]**
   후보 **model**을 실제 세탁기에 넣어 성능 측정

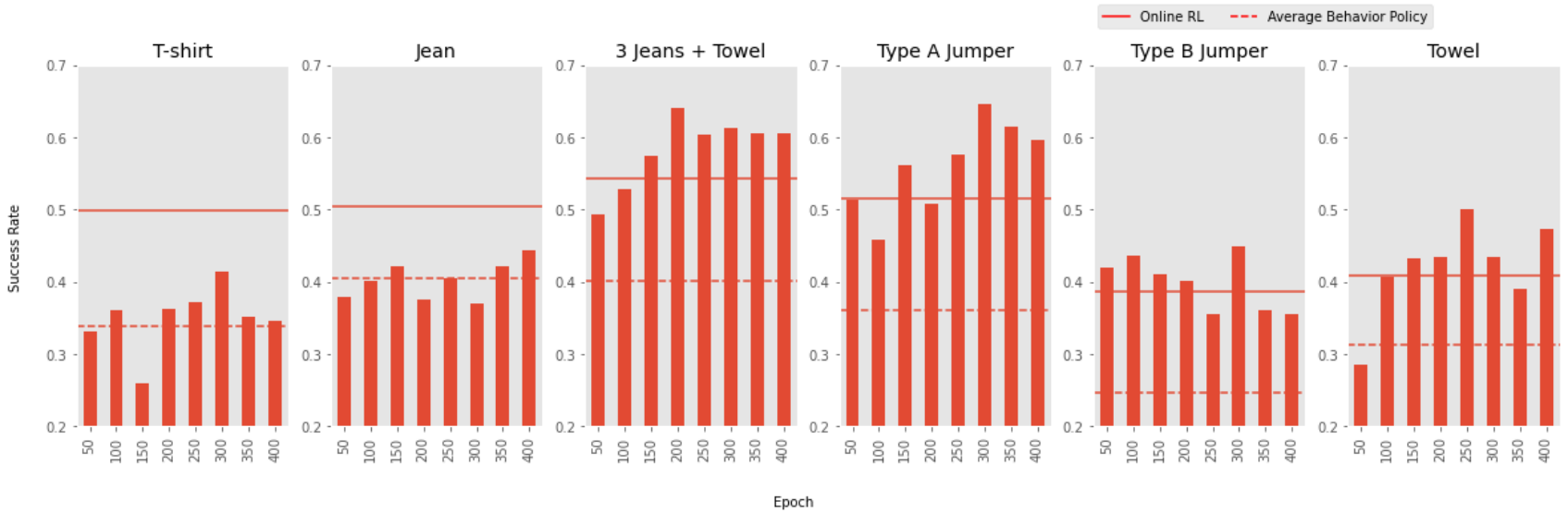
4. **[Growing Batch Learning]**
   Online Evaluation시 좋은 **model**의 **log**를 재학습

5. **3**번 과정 반복

실험 결과

- **Policy Selection (1)**

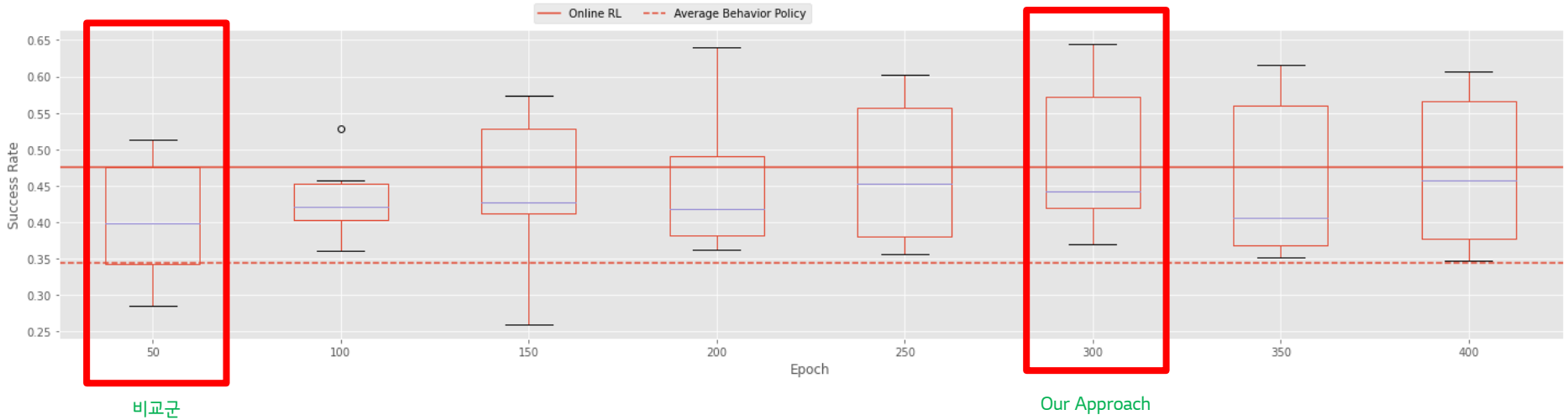
Average Behavior Policy : Dataset 내의 policy (Behavior policy) 성능



- 세탁포마다 최적의 **epoch**이 다 다른 것이 관찰됨
- 일부 세탁포의 경우 **Online RL (전문가) model**보다 성능이 개선됨을 확인함

실험 결과

- **Policy Selection (2)**



  - **300 epoch** 모델 **(49.4%)**이 **online RL** 모델**(47.5%)**보다 좋은 성능을 보여줌
    → 이 때의 **trajectory log**를 재학습 데이터로 추가함

  - **Data**의 **Quality** 측면에 대한 실험을 위해서 **50 epoch** 모델**(41.6%)**의 **log**를 비교군으로 삼음
    → **Growing Batch Learning**의 **Robustness** 평가용

데모 영상


Motor Control in Washing Machine Example - Towel (Heuristic Based)

데모 영상

실험 결과

- **Growing Batch Learning**의 효과

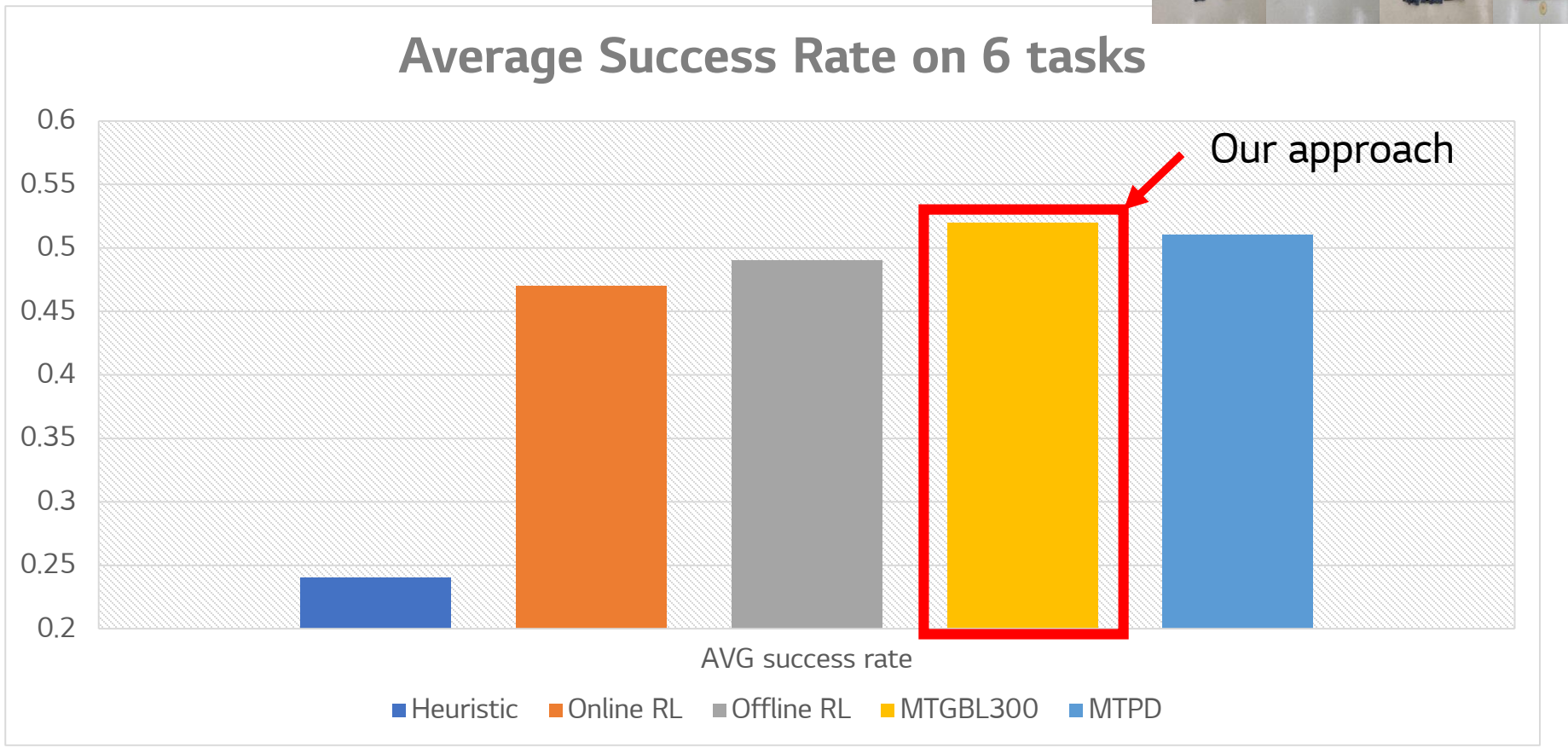
bad: 50 epoch 학습모델
good: 300 epoch 학습모델

- 좋은 모델의 **log**로 학습한 모델은 단순히 **Offline RL**을 했던 것에 비해 전반적으로 성능이 개선되었음
  (최대 성공율: 150 epoch - 52.4%, online RL 대비 5%↑, offline RL 대비 3%↑)
  |
- 안 좋은 모델의 **log**로 학습한 모델도 거의 비등하게 성능이 개선됨을 확인할 수 있었음
  (최대 성공율: 100 epoch - 51.5%, online RL 대비 4%↑, offline RL 대비 2%↑)

굳이 **policy selection**을 통해서 좋은 모델을 찾지 않아도,
Offline RL에서 **data**를 추가하는 효과만으로도 성능 개선이 가능하다!

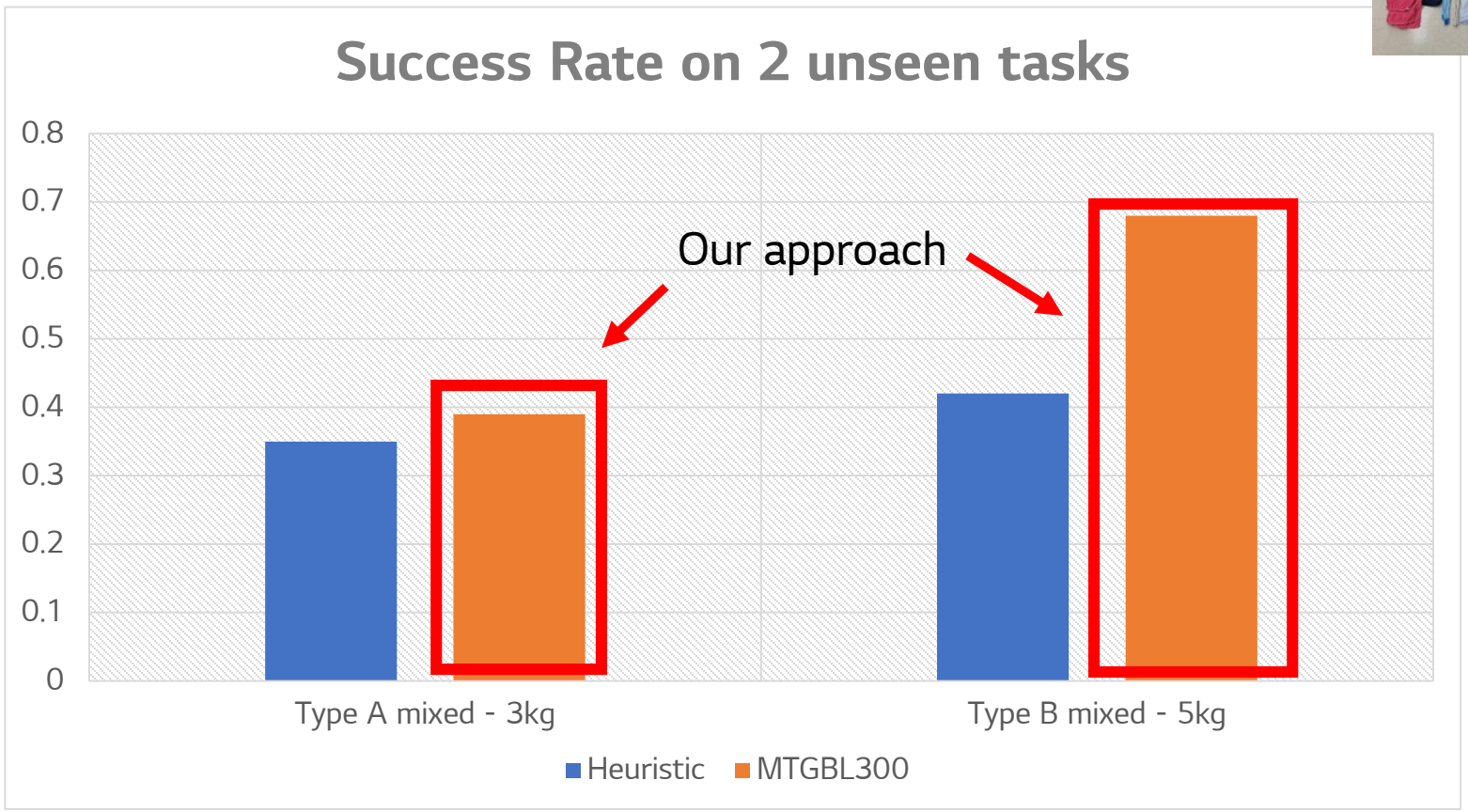
실험 결과

- 학습한 6개 세탁포에 대한 평균 성공율



실험 결과

- 학습하지 않은 2개 세탁포에 대한 평균 성공율



기대 효과 및 계획

- 다양한 세탁포를 저진동 상태로 탈수시킬 수 있는 하나의 통합 모델 생성
  (+) Policy Distillation이나 Quantization과 같은 compression 기법으로 용량 확보 극대화

- ThinQ platform을 통해 센서 데이터 수집이 가능하다면 (a.k.a superset dataset)
  실환경 데이터 추가 확보를 모델 성능 개선이 가능해짐(≈ LG UP 가전)

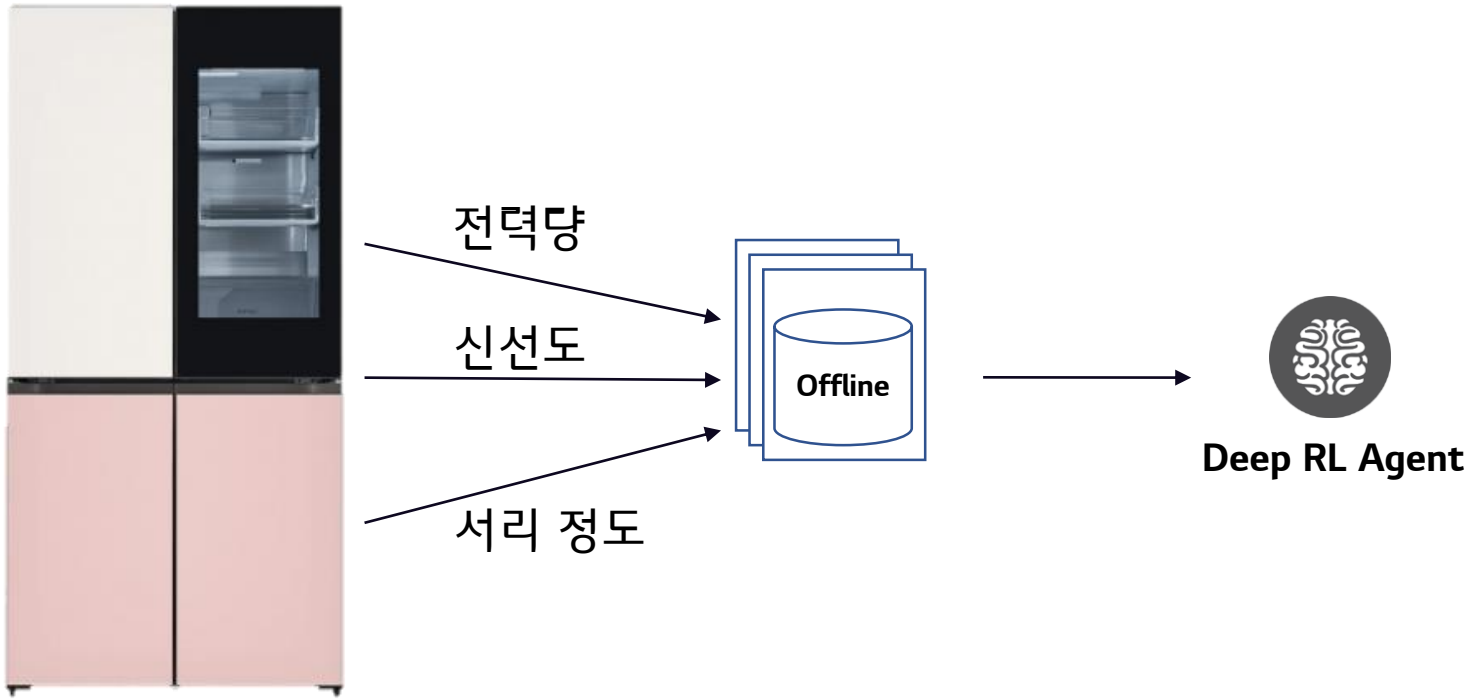
- 강화학습을 활용한 Product-level 차원에서의 본질 성능 향상 첫사례(?)

Multi-Task Offline Reinforcement Learning for Optimal Motor Control in Washing Machines

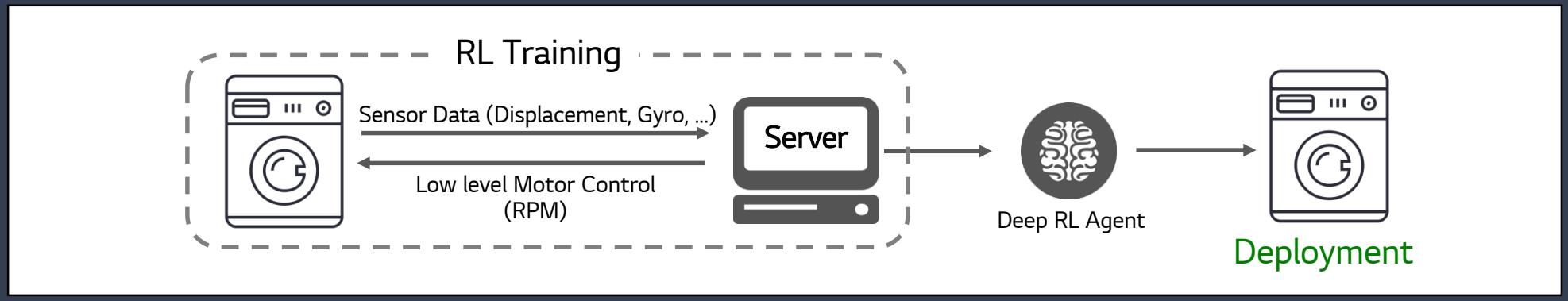
Chanseok Kang[1*], Guntae Bae[1*], Kyoungwoo Lee[1*],
Chul Lee[1], Dohyeon Son[1], Daesung Kim[1], and Jae Woong Yun[1]

응용 분야(?)

- 냉장고 부하 대응



RL in Washing Machine



- State: Displacement, Gyro, Acceleration of 3 axis on front and rear side

  Motor Information (Speed, Force)

- Action: $\begin{cases} UP \\ DOWN \end{cases}$
  - Request RPM is increased
  - Request RPM is decreased

- Reward: $\begin{cases} +1 \\ -1 \\ 0 \end{cases}$
  - Target RPM is reached
  - Terminated with overhead
  - not reached with timeout

응용 분야(?)

- 강화학습을 자사 제품군에 적용하고자 하는 노력
  - 강화학습 환경 설계 **(Markov Decision Process)**
  - 제품군의 득정 현상을 설명해줄 수 있는 **Domain Experts**
  - 다양한 시도와 이를 위한 충분한 시간 **(+ 끈기)**

회고

Q&A

LG SDC 2022
CONNECT GROW OUTSTAND

THANK YOU

Appendix

Traditional Approach

User In the Loop Process

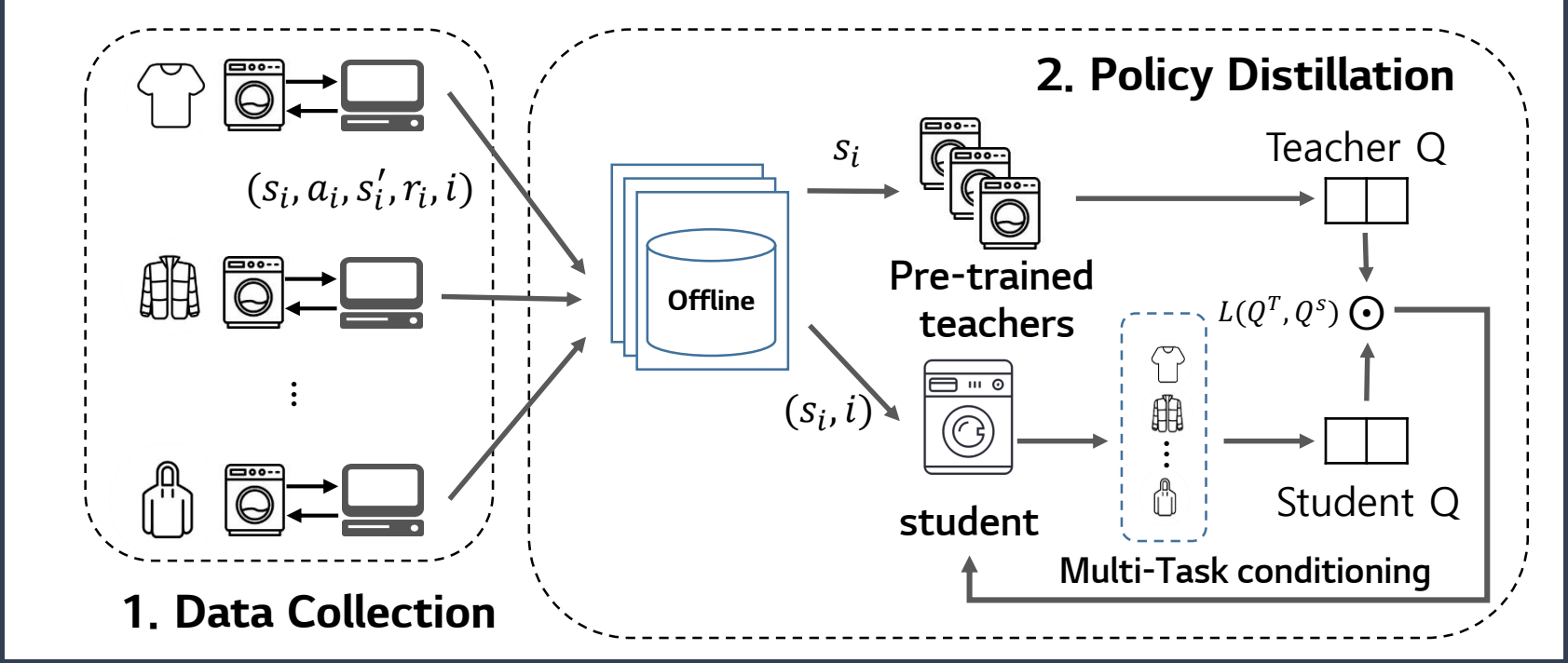
- Human Expert observed the pattern
- Modify the firmware w.r.t specific laundry

(-) Requires numerous trials and experiments

(-) Cannot handle whole laundry cases

Our Approach (MTPD)

Multi-Task Policy Distillation



1. Data Collection
2. Multi-Task Policy Distillation