제 2회 KAIST-POSTECH-UNIST

데이터 사이언스 경진대회 설명회

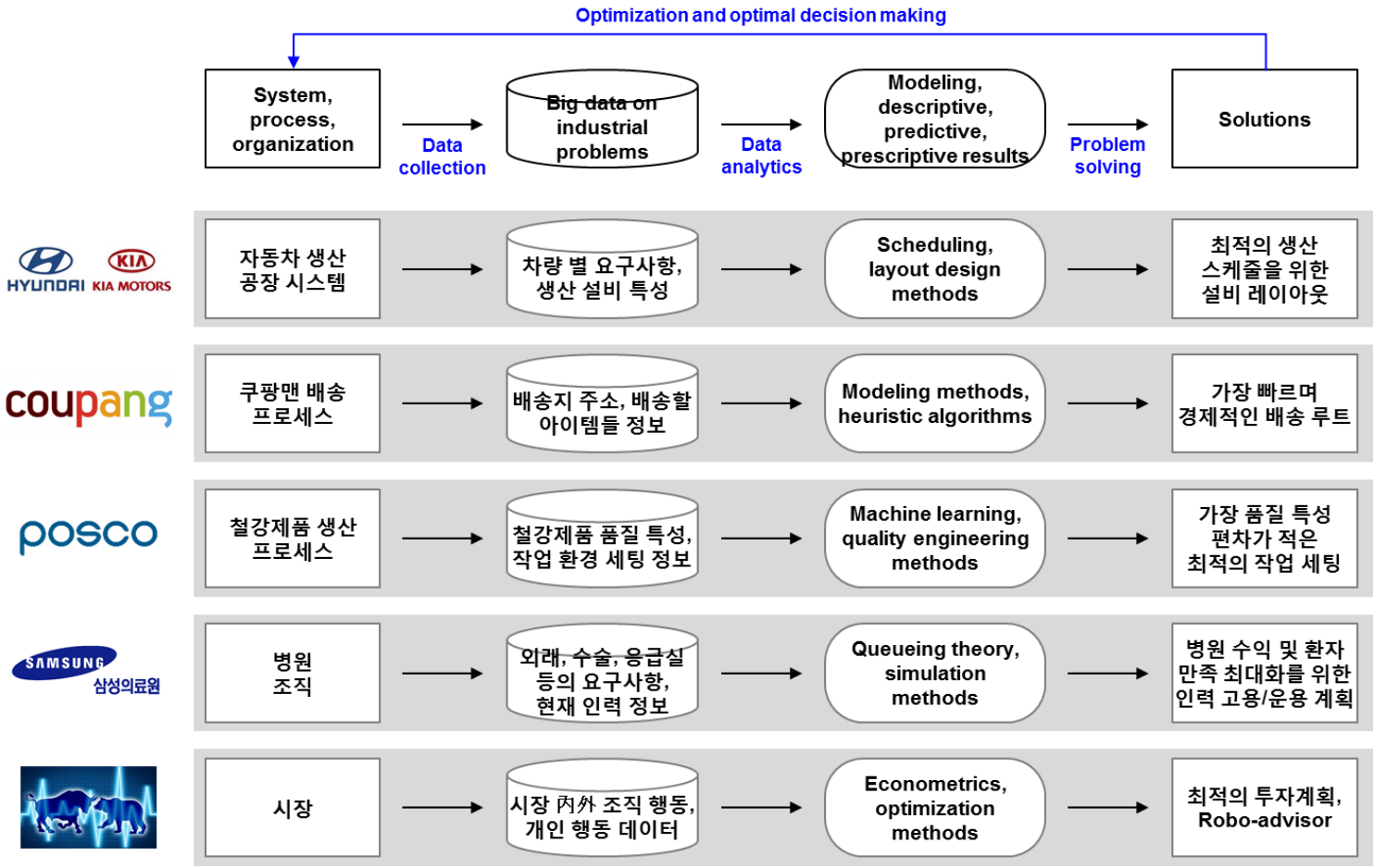
공동주최 KAIST POSTECH UNIST
주관 ISysE KAIST GSDS
후원 하나은행 Lunit MakinaRocks Fast campus ECMiner

Contents

산업공학 : 산업 시스템의 설계와 운영

Domain knowledge → 연역적 추론 기반 모델 → 최적 의사결정

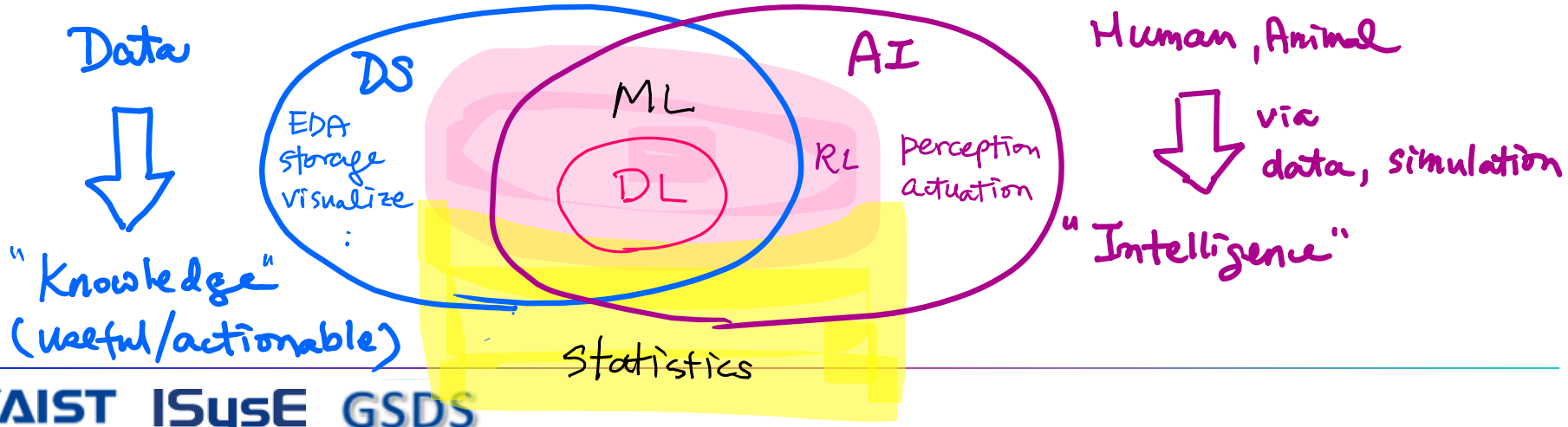
Big data → 귀납적 추론 기반 모델 → 최적 의사결정



Data Science?

- **Data science** is an interdisciplinary field (* mainly from Wikipedia with some modifications)
  - that uses scientific methods, processes, algorithms, and systems
    - math, statistics, programming, analytics, AI, ML, …
  - to extract knowledge and insights from noisy, structured and unstructured data
    - inductive knowledge (vs deductive knowledge)
  - to apply knowledge from data across a broad range of application domains

- **Data science is related to**
  - **data mining**: old name for DS, narrower than DS
  - **statistics**: key foundation of DS
  - **machine learning**: core technology for DS
  - **AI**: overlaps with DS, but a little different objective



데이터사이언스 경진대회

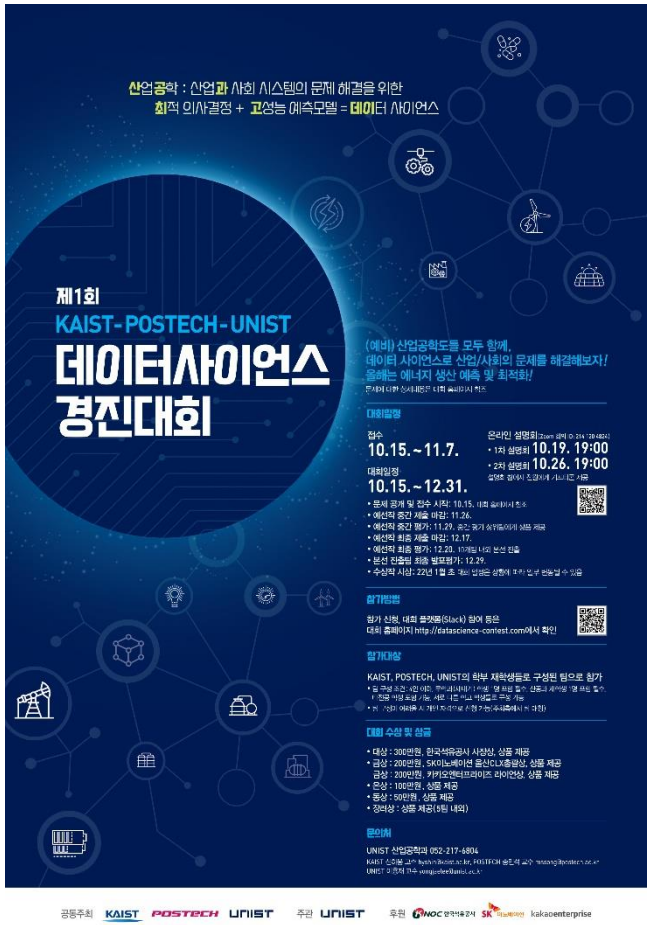
- 본 경진대회의 목적

  – 기존에 많은 경진대회 : ‘예측’에 중점

  – 산업공학 관점에서 본 데이터사이언스 = **예측** + **의사결정**
    - '**예측**' 결과를 이용한 ‘**의사결정**’ → ‘**문제 해결**’
    - ‘산업/사회 시스템'의 설계와 운영 문제해결 → 산업공학

산업공학 :산업**과** 사회 시스템의 문제해결을 위한
최적 의사결정 + **고**성능 예측모델 = **데이**터 사이언스

데이터사이언스 경진대회



- 산업공학 데이터사이언스 경진대회
  - KAIST, POSTECH, UNIST 과학기술특성화대학 산업공학과에서 공동 주최
- 제1회 대회 : UNIST 산업공학과 주관
  - 3개 대학에서 약 38개 팀 (150+명) 참여
  - 대회 문제 : Shale gas 유정의 생산량 예측 및 유정 구입 결정
  - 총 상금 약 1000만원
    - 후원사:한국석유공사, SK이노베이션, 카카오엔터프라이즈
- 향후 3개 대학이 돌아가면서 주관
  - 제2회 대회 : KAIST에서 주관
    - 산업및시스템공학과
    - 데이터사이언스 대학원 (GSDS)

제2회 경진대회

KAIST ISysE KAIST GSDS

제2회 대회 일정

대회 일정은 상황에 따라 일부 변동 될 수 있습니다.
대회 홈페이지 및 슬랙 통해 지속적인 확인 부탁드립니다.

- 1차 온라인 설명회: 10/26, 19:00
- 2차 온라인 설명회: 11/2, 19:00
- 접수: 10/24 ~ 11/9
  - 대회 홈페이지(datascience-contest.com)에서 신청
  - 참가신청 완료된 팀은 Slack에 초대 후 데이터 제공 예정
- 예선 : 정량 평가
  - 예선작 중간 제출: 마감 11/26, 평가발표 12/1
  - 예선작 최종 제출: 마감 12/17, 예선 결과 발표 12/22
- 본선 : 발표 평가 (10개팀 내외)
  - 예전 정량 평가를 통해 10개 팀 내외 본선 진출팀 결정
  - 본선 진출 팀 최종 발표평가: 12/29(목)
- 수상작 시상: 2023년 1월 초

참가자격 및 팀구성

- KAIST, POSTECH, UNIST의 학사과정 학생으로 구성된 팀
  - 4인 이하
  - 최소 1인 이상의 새내기(무학과) 학생 포함
  - 최소 1인 이상의 산공과 학부생 포함
    - 복수 전공, 휴학생 포함 가능
  - 타전공 학생 포함 가능
  - 서로 다른 학교 학생들로 구성 가능
- 팀 구성이 어려울 시 개인 자격 (또는 부분 팀)으로 신청 가능
  - 신청 후 주최측에서 팀 매칭
- 그외 팀 구성에 있어서 특이 상황은 주최측에 문의 요망

대회 시상 및 상금

후원

하나은행 Lunit MakinaRocks Fast campus ECMiner

- **대상** (1 팀): 300만원
  - 하나은행 상
- **금상** (3 팀): 각 200만원
  - 루닛 상
  - 마키나락스 상
  - 패스트캠퍼스 상
- **은상** (1 팀): 100만원
  - ECMiner 상
- **동상** (2 팀): 각 50만원
  - ECMiner 상
- **장려상** (5팀 이내)
  - 상품

*** 상금 이외의 특전**

- (희망자에 대하여) 후원사 Tool 사용 교육
- 마키나락스 특전
  - Link™ 사용 본선진출 팀에 대해 Team building 지원 (20만원/팀)
  - 수상팀 인턴지원시 서류 평가 면제
  - 대회 참가자를 위한 Link™ 사용법 설명 Session
    * 11/4(금) 11am에 아래 URL로 접속
    * https://meet.google.com/hpj-fxsy-ywv
- 그 외에도 특전이 있을 경우 및 특전의 상세한 내용은 홈페이지를 통해 공지 예정

문의처

- 홈페이지
  - https://datascience-contest.com
  - 대회 자료
  - 참가신청
- 슬랙채널
  - 참가신청자에게 공개

- 담당 교수
  - KAIST 신하용 교수 (hyshin@kaist.ac.kr),
    박찬영 교수 (cy.park@kaist.ac.kr), 민승기 교수 (skmin@kaist.ac.kr)
  - POSTECH 고영명 교수 (youngko@postech.ac.kr)
  - UNIST 이용재 교수 (yongjaelee@units.ac.kr)

Part 1

Data Description

Dataset: Overview

- **Provided by Hana Bank**
  - Usage record of Hana 1Q App (mobile banking app)
  - for 8 months (2022.01.01 ~ 2022.08.26, 238 days)

하나원큐

- **Data description**
  - Input features (raw) : 3 + 238*3 = 717 columns per user

feature name	meaning	remark
gender	Male(0) or female(1)	binary
age_code	age band (1:<20, 2~13:20~79, 14:>=80)	
region_code	region code (1~18)	categorical
c20220101~ c20220826	Number of logins for each day	time series (cardinal)
t20220101~ t20220826	Number of logins with money transfer	time series (cardinal)
s20220101~ s20220826	Duration of staying with the app	time series (cardinal)

  - Label = {1: small business owner, 0: general (non-business)}
    - The proportion of the small business owner (label=1) is approximately 6.55%

Dataset: Example

- **Example**
  - Input features

Gender	age_code	region_code	c20220101	…	s20220101	…	t20220101	…
2	4	4	4		121		2	
1	8	1	1		15		-	

- YYYYMMDD : 20220101 ~ 20220826
- cYYYYMMDD : # logins on YYYYMMDD
- sYYYYMMDD : Duration of stay on YYYYMMDD
- tYYYYMMDD : # logins with money transfer on YYYYMMDD

  - Label

Small business owner
1
0

- 1: A small business owner
- 0: Not a small business owner

Dataset: Files

- Dataset for modeling (D_{model}): input features + label
  - 800,000 user records
  - can be used for developing your model

    • X_model.csv: Input features
    • Y_model: Label

- Dataset for competition (D_{exam}) : input features only
  - 200,000 user records

    • X_exam.csv: Input features

- File to submit : prediction and decision results for D_{exam}

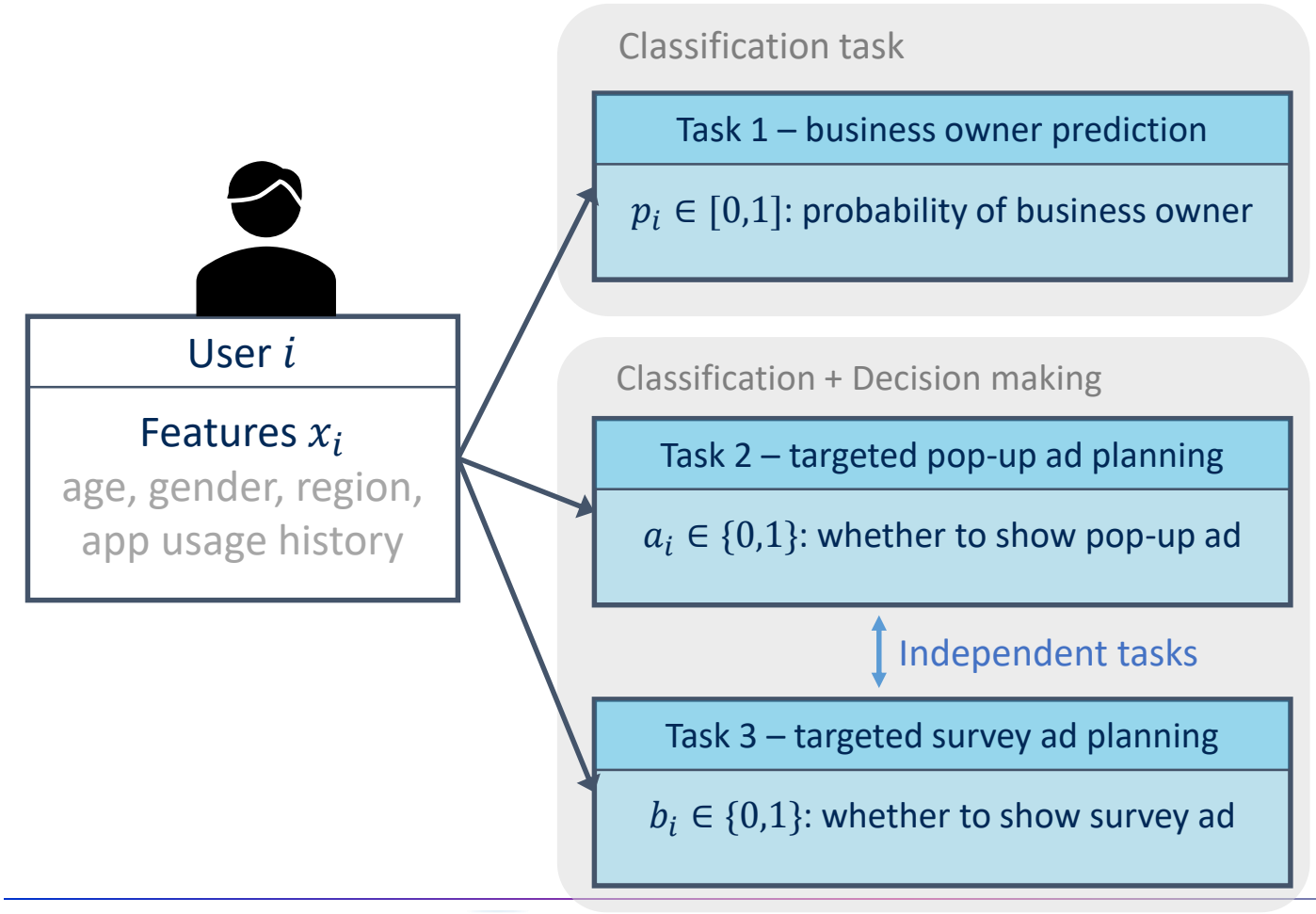
  • submission.csv: prediction and decision results
    • a sample file will be given
    • $\{p_i, a_i, b_i\}$ in the same order as X_exam.csv

Part 2

Task Description

Overview

- For each user i in the exam dataset (200,000 users), you have to submit three numbers p_i, a_i, b_i



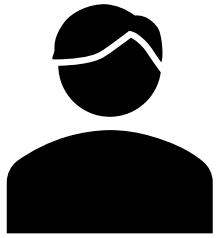
submission.csv

Task 1 (p_i) Task 2 (a_i) Task 3 (b_i)

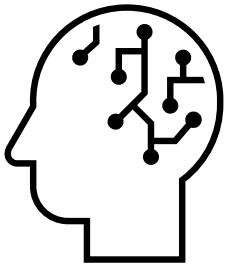
business_prob	popup	survey
0.61845	1	0
0.97296	1	1
0.78073	1	0
0.02992	0	0
0.25522	0	0
0.92836	1	1
0.89172	1	0
0.63428	1	0
0.78102	1	0
0.93564	1	1
0.2145	0	0
0.97235	1	1
0.64393	1	0
0.36104	0	0
0.51847	1	0
0.53315	1	0
0.43127	0	0
0.28834	0	0

Task 1. Small Business Owner Prediction

- **Context:** the company wants to predict which users are small business owners based on user profiles & app usage history



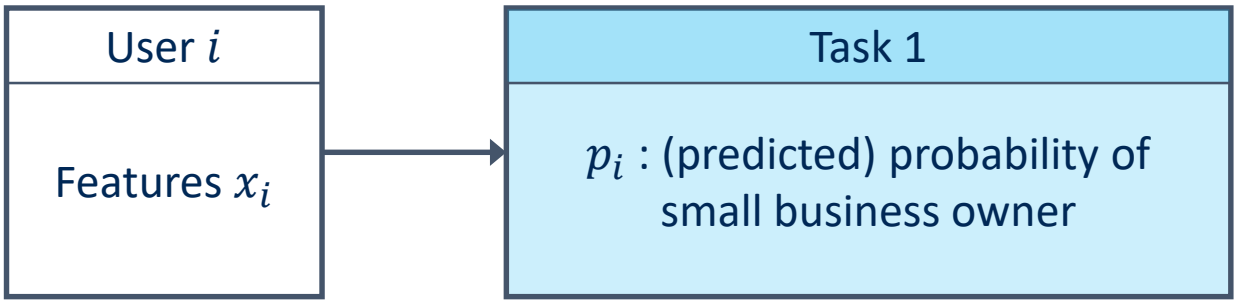
- Gender: male
- Age: in 20s
- On Jan 1st,
  - logged in 2 times
  - transferred money once
  - used app for 10 mins
- On Jan 2nd , no activity
- …



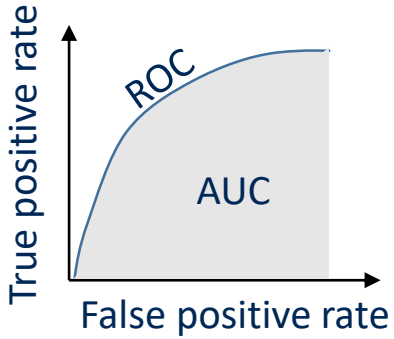
This user will be
a small business owner
with 60% probability

Task 1. Small Business Owner Prediction

- **Task:** for each user in the exam dataset, predict the probability that the user is a small business owner
  - **Prediction**: $p_i \in [0,1]$

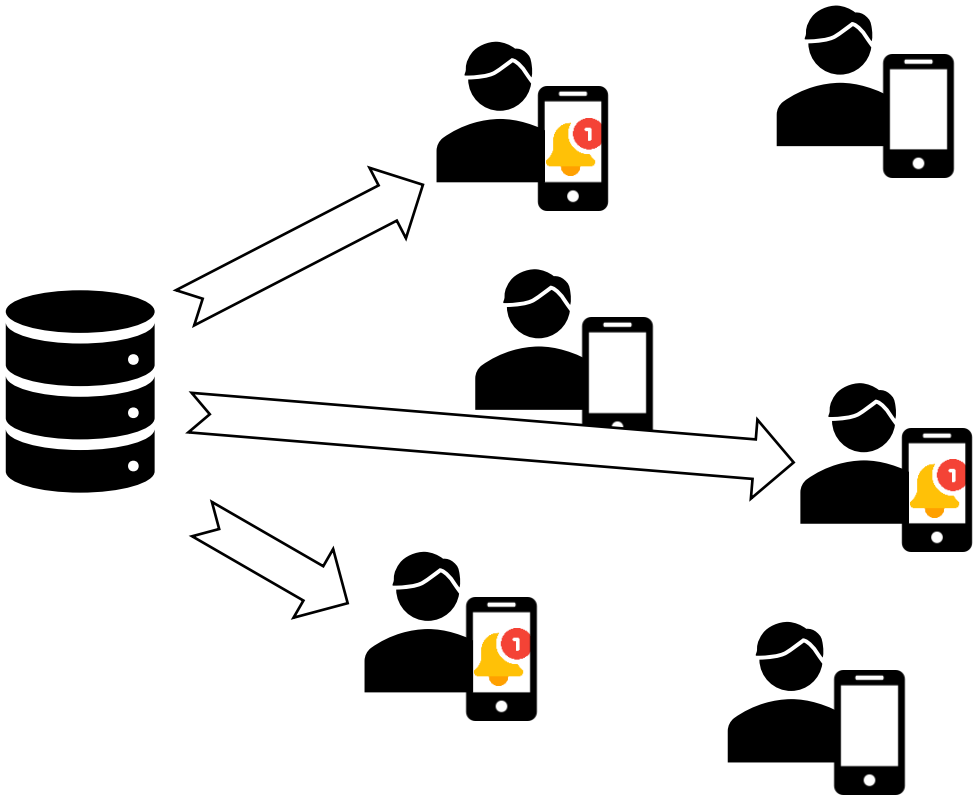


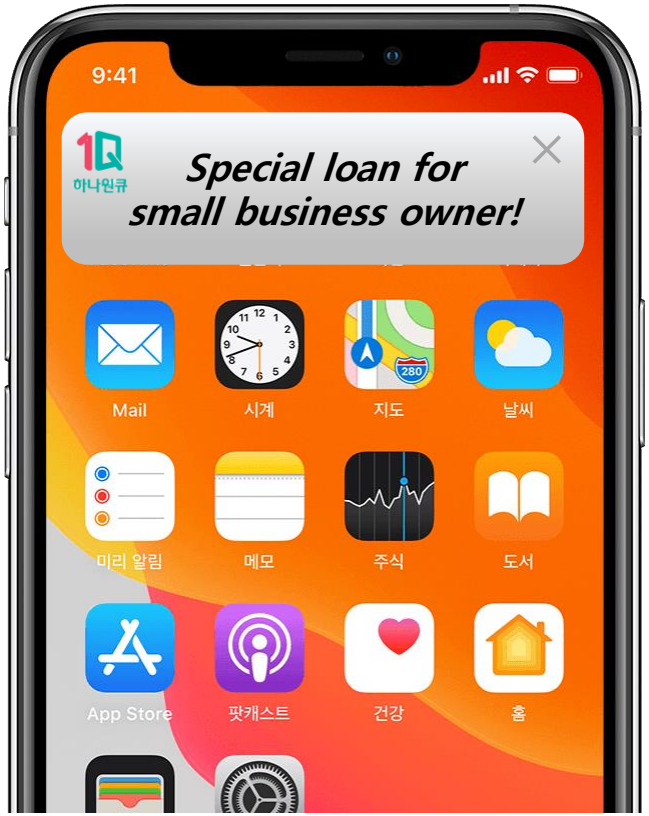
- **Evaluation**: AUC value of ROC curve



Task 2. Pop-up Ad Planning

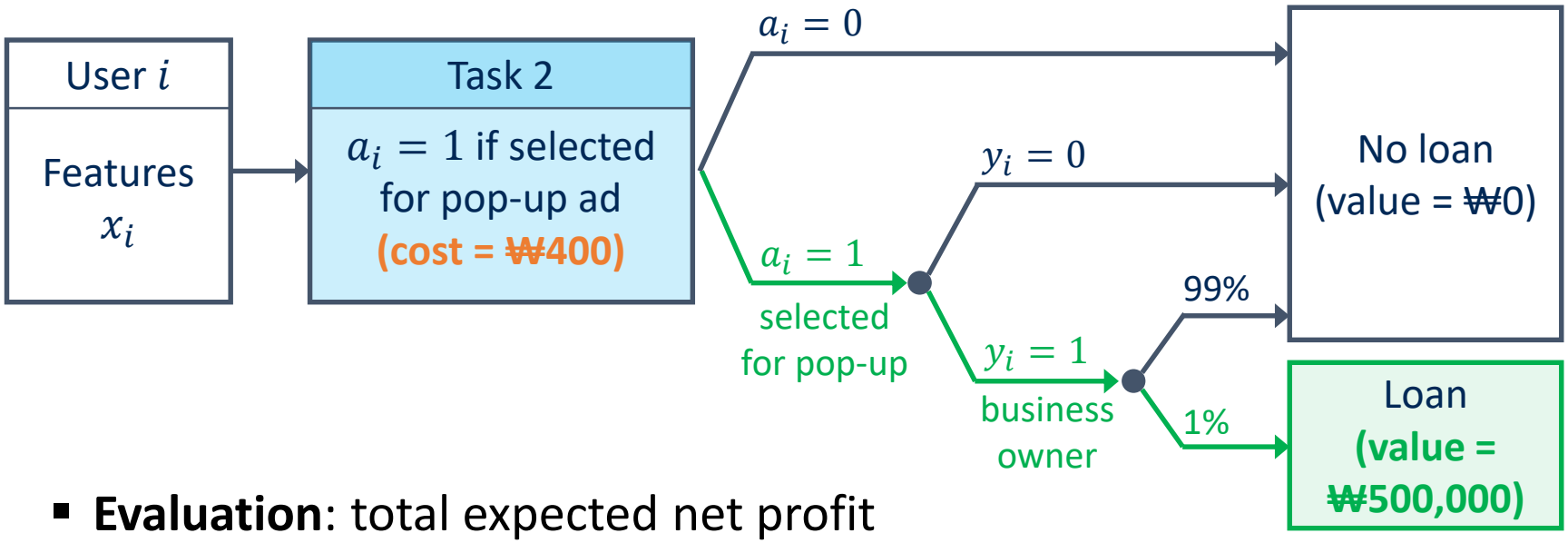
- **Context**: the company is planning to advertise the loan service by sending ***pop-up notifications*** to selected users





Task 2. Pop-up Ad Planning

- **Task**: for each user in the exam dataset, decide whether to send a pop-up notification
  - **Decision**: $a_i = 0$ or 1
  - **Cost**: ₩400 per user (pop-up add setup cost)
  - **Value**: 1% of small business owners who viewed pop-up ad will use the loan service, yielding value of ₩500,000

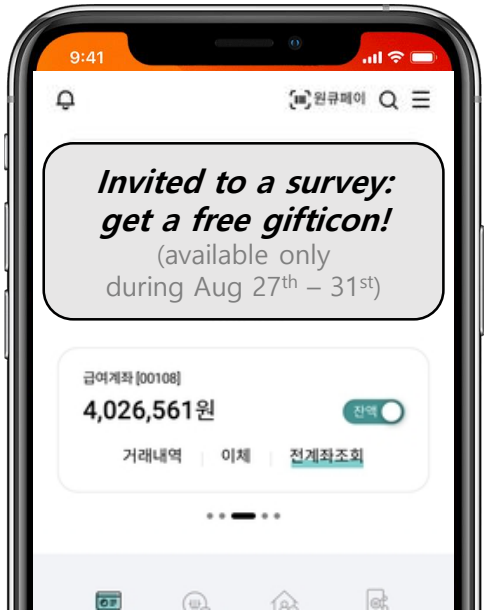


- **Evaluation**: total expected net profit

  = total expected value – total cost

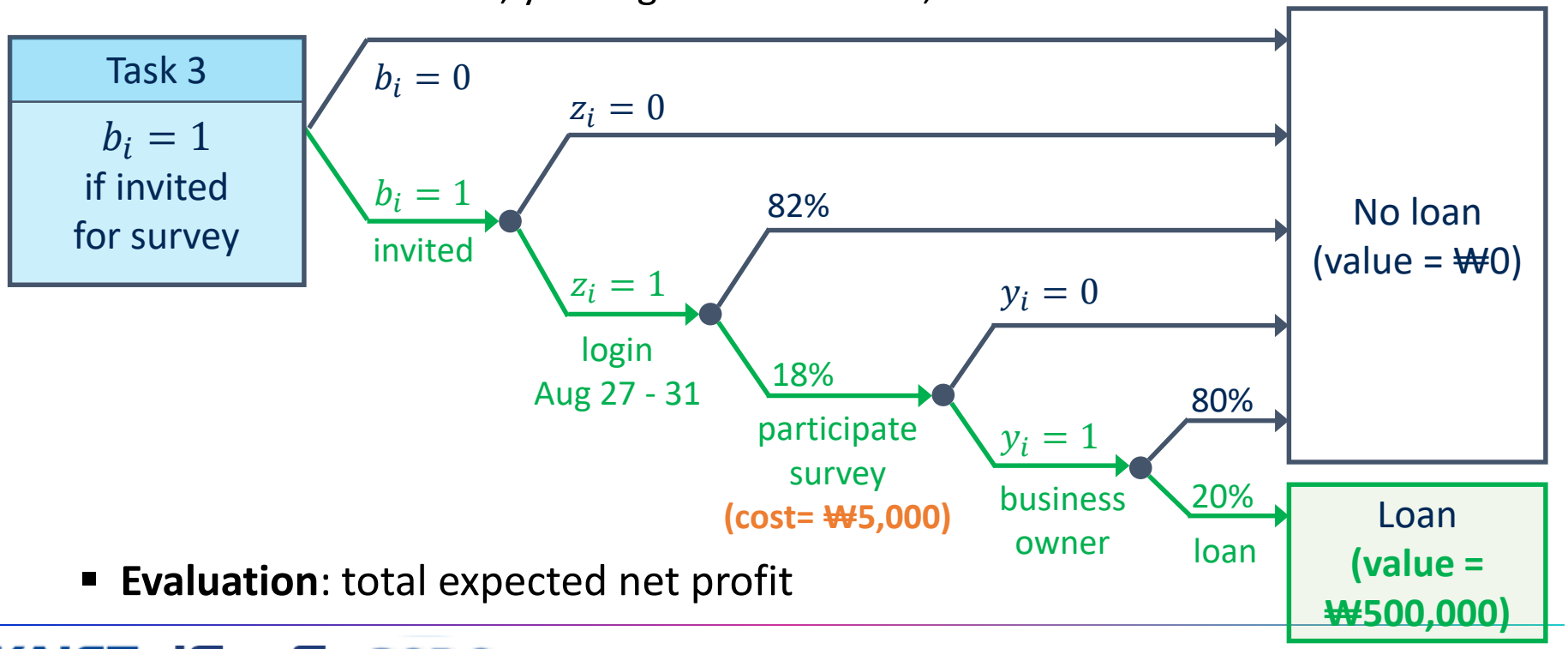
Task 3. Survey Ad Planning (indep. of Task 2)

- **Context:** the company is planning to advertise the loan service through ***in-app survey-based promotion***
  - over the next 5 days (August 27th – 31st)
  - only for selected users (up to 50,000 users)
  - gifticon for survey participants



Task 3. Survey Ad Planning (indep. of Task 2)

- **Task:** for each user in the exam dataset, decide whether to invite him/her to the survey promotion
  - **Decision**: $b_i = 0$ or 1 (up to 50,000 users: i.e. $\sum_i \ b_i \leq 50000$)
  - **Cost**: upon login, an invited user will participate the survey with 18% chance, incurring ₩5,000 cost (gifticon)
  - **Value**: upon participation, a small business owner will use the loan service with 20% chance, yielding value of ₩500,000



- **Evaluation**: total expected net profit

Conclusion

- 대회 진행 관련 세부 사항은 아래 참조
  - 대회 홈페이지 (https://datascience-contest.com)
    - 참가신청 페이지 (대회 홈페이지 통해 접속 가능) https://docs.google.com/forms/d/e/1FAIpQLSdwOdkCwmWmsoTQ5Q0x7r_XVobrRS3j4kSGWVE8v8g4SlraFA/viewform
  - 오픈카톡방 (대회 참여 전, https://open.kakao.com/o/gaALZVGe)
  - 슬랙 (대회 참여 후)
- 질문사항은 오픈카톡방과 슬랙 통해서 편하게 물어봐주세요

Thank you for your attention!