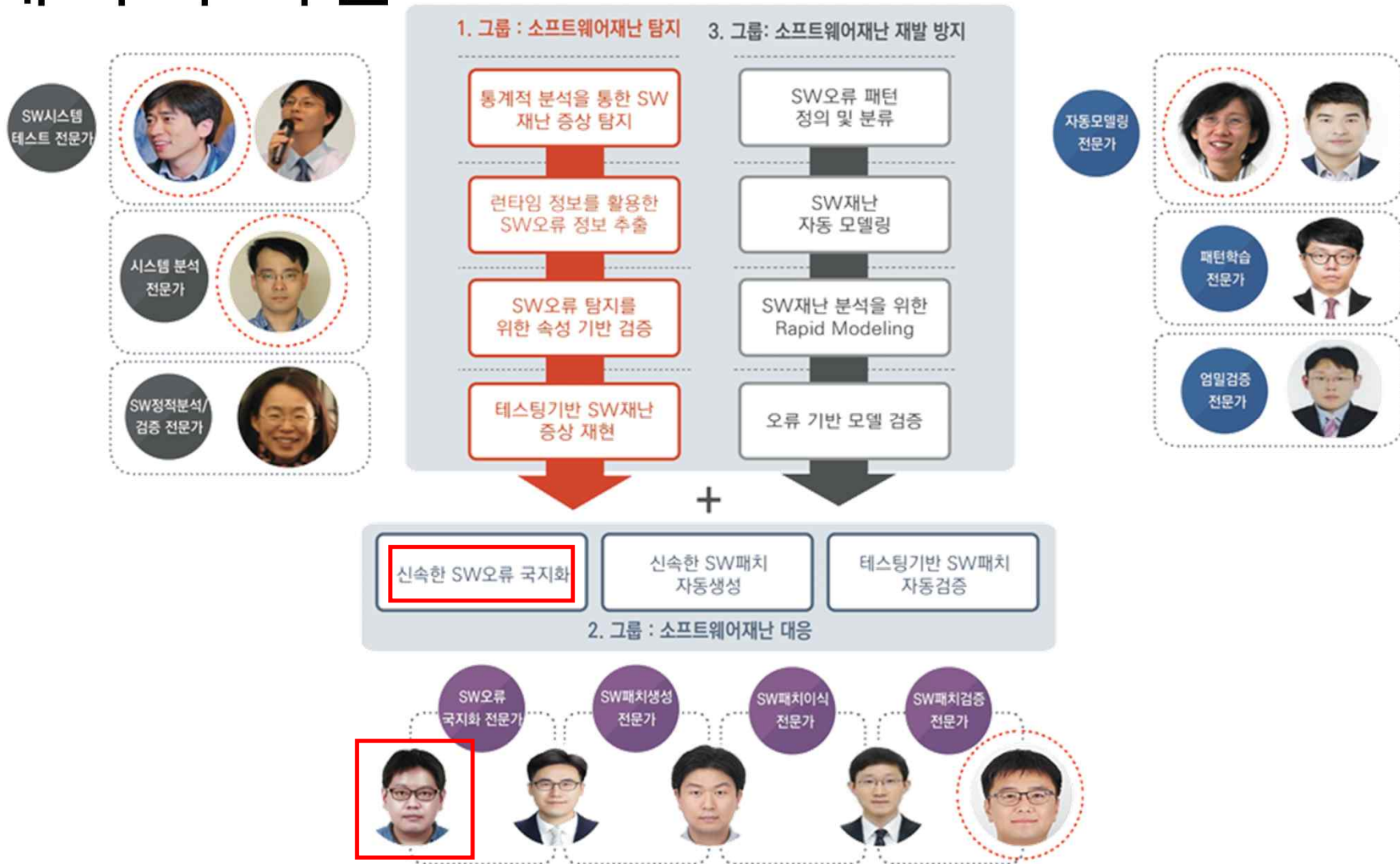


효율적이고 효과적인 오류 위치 추정 기법을 찾아서

김윤희

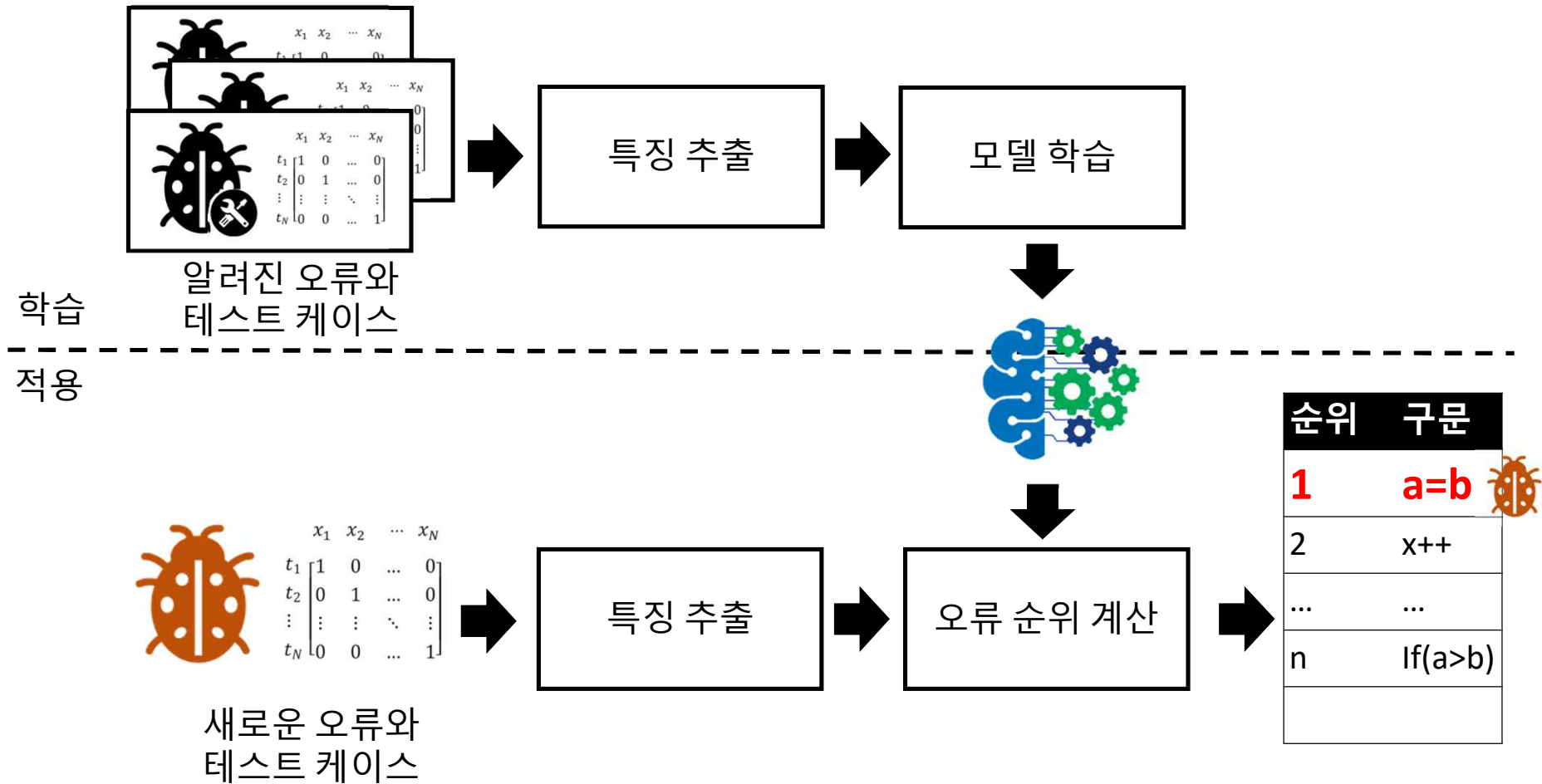
한양대학교 컴퓨터소프트웨어학부

센터에서의 역할



머신 러닝 기반 오류 위치 추정 기법

- 오류 구문의 다양한 특징을 학습하여 오류 위치 추정



충분한 학습 데이터를 사용하고 있는가?

- 더 많은, 더 좋은 데이터가 더 좋은 모델을 만든다
- 머신 러닝 기반 오류 위치 추정 기법은 충분한 데이터를 사용하고 있는가?

연구	오류 데이터셋 크기	사용 벤치마크
PRINCE[Kim et al, TOSEM19]	154	SIR, CoREBench
DeepFL[Li et al., ISSTA19]	395	Defects4J 1.2
DeepRL4FL[Li et al., ICSE 21]	395	Defects4J 1.2

- 다른 분야의 데이터셋은 어떤가?

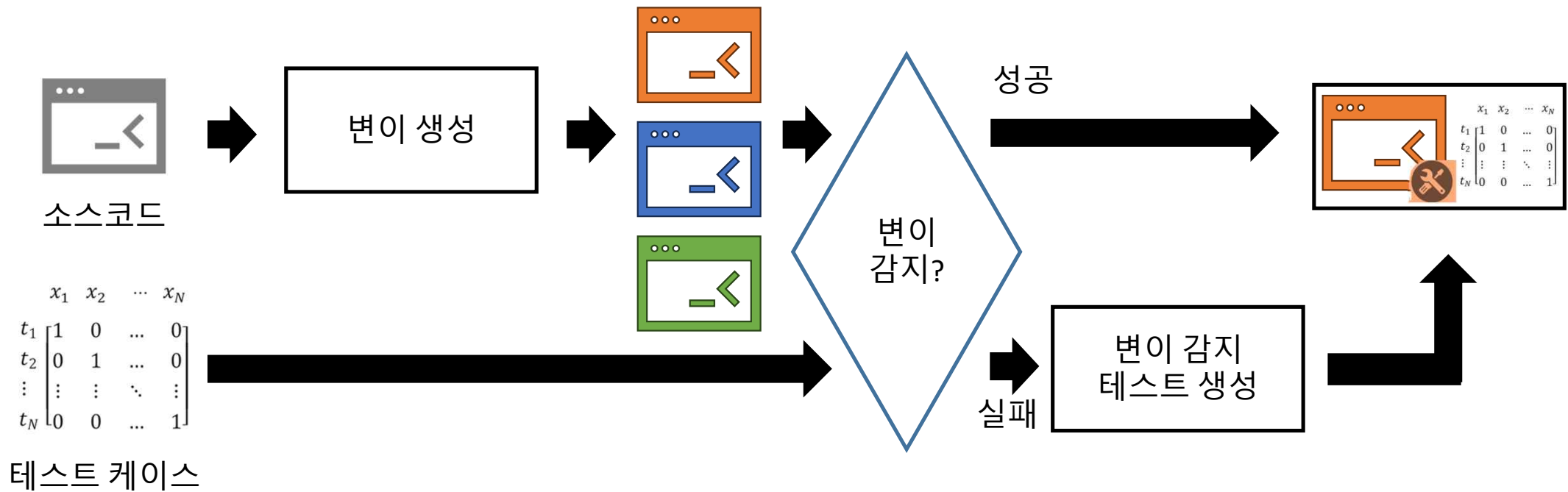
데이터 셋	데이터셋 크기
CIFAR	학습: 50,000, 테스트: 10,000
ImageNet	학습: 1,200,000, 테스트: 50,000
MNIST	학습: 60,000, 테스트: 10,000

- 더 많은 학습 데이터를 사용해서 정확도를 높여보자!

더 많은 오류 데이터를 어떻게 확보할 것인가?

- 방법 1: 더 많은 실제 오류 데이터 수집
 - 장점: 양질의 오류 데이터 확보 가능
 - 단점: 버그 수집, 테스트 확보, 버그 재현성 확인 등 막대한 노동력 필요
- 방법 2: 변이 생성으로 인위적으로 오류 데이터 생성
 - 장점: 다수의 오류 데이터를 손쉽게 확보 가능
 - 단점: 인위적으로 생성한 오류 데이터의 품질이 실제 오류 데이터보다 높은지 의문

인위적인 오류 데이터 생성 과정



연구 질문

- 더 많은 오류 데이터가 얼마나 모델을 향상시킬것인가?
 - RQ1: 오류 데이터를 추가로 사용하여 얼마나 정확하게 오류 위치를 찾을 수 있는가?
 - RQ2: 오류 데이터의 갯수에 따른 정확도가 얼마나 차이가 나는가?

실험 설정

- Defects4J v1.2의 395개 오류 버전을 대상으로 실험

프로그램	버그 수
Chart	26
Closure	133
Lang	65
Math	106
Mockito	38
Time	27

- 사용 모델: DeepFL 에서 제안한 오류 위치 추정 머신 러닝 모델 사용
 - 4 타입의 특징 중 SBFL, MBFL 피쳐만 사용하고 코드 복잡도, 유사도 특징은 제외
- 오류 데이터 생성 구현
 - 변이 도구: Major 활용
 - 테스트 생성 도구: Evosuite 활용

실험 결과 – RQ1:정확도 비교

- DeepFL 보다 TOP1이 크게 향상되고 DeepRL4FL과 유사한 Top1, Top5 정확도 달성
 - Top1, Top5: 오류 구문이 1등, 5등 이내로 탐지된 오류 버전의 수
 - MFR: 오류 구문의 평균 탐지 순위
- DeepFL, DeepRL4FL이 어려워했던 오류를 상당히 잘 찾음
 - MFR의 향상도가 높음

	Top1		Top5		MFR
DeepFL	39	9.9%	129	32.7%	24.1
DeepRL4FL	71	18.0%	142	35.9%	20.3
39,500개 오류 데이터 사용 모델	68	17.2%	138	34.9%	14.7

실험 결과 – RQ2: 데이터 크기 효과

- 오류 데이터 크기가 커질수록 모델 정확도 향상
 - 다만, 한계효용도 있음: 395,000개의 오류 데이터가 39,500개 오류 데이터보다 더 좋은 모델을 학습하는 것은 아님

모델 학습/추론에 사용한 오류 갯수	Top1		Top5		MFR
395	29	7.3%	102	25.8%	28.8
3950	47	11.9%	119	30.1%	20.3
39500	68	17.2%	138	34.9%	14.7
395000	66	16.7%	139	35.2%	14.3

결론: 더 많은 학습 데이터가 필요하다

- 더 많은 오류 데이터 -> 더 정확한 모델 학습
- 단, 한계 효용은 있다
 - x100은 x10 보다 좋지만 x1000은 x100보다 꼭 좋진 않다

향후 계획

- 하지 못한 이야기: 더 **좋은** 데이터를 어떻게 확보할 것인가?
 - 무엇이 **좋은** 데이터인가?
 - **좋은** 데이터를 어떻게 확보할 것인가?

[illegible]

- 퍼저 선택 문제를 Multi-Armed Bandit 추천 문제로 풀어보자