재현데이터의 활용 및 한계
정책적 시사점

임종호

연세대 통계데이터사이언스학과

한국금융학회

2022. 06. 10

발표 개요

Part I　Why 재현데이터

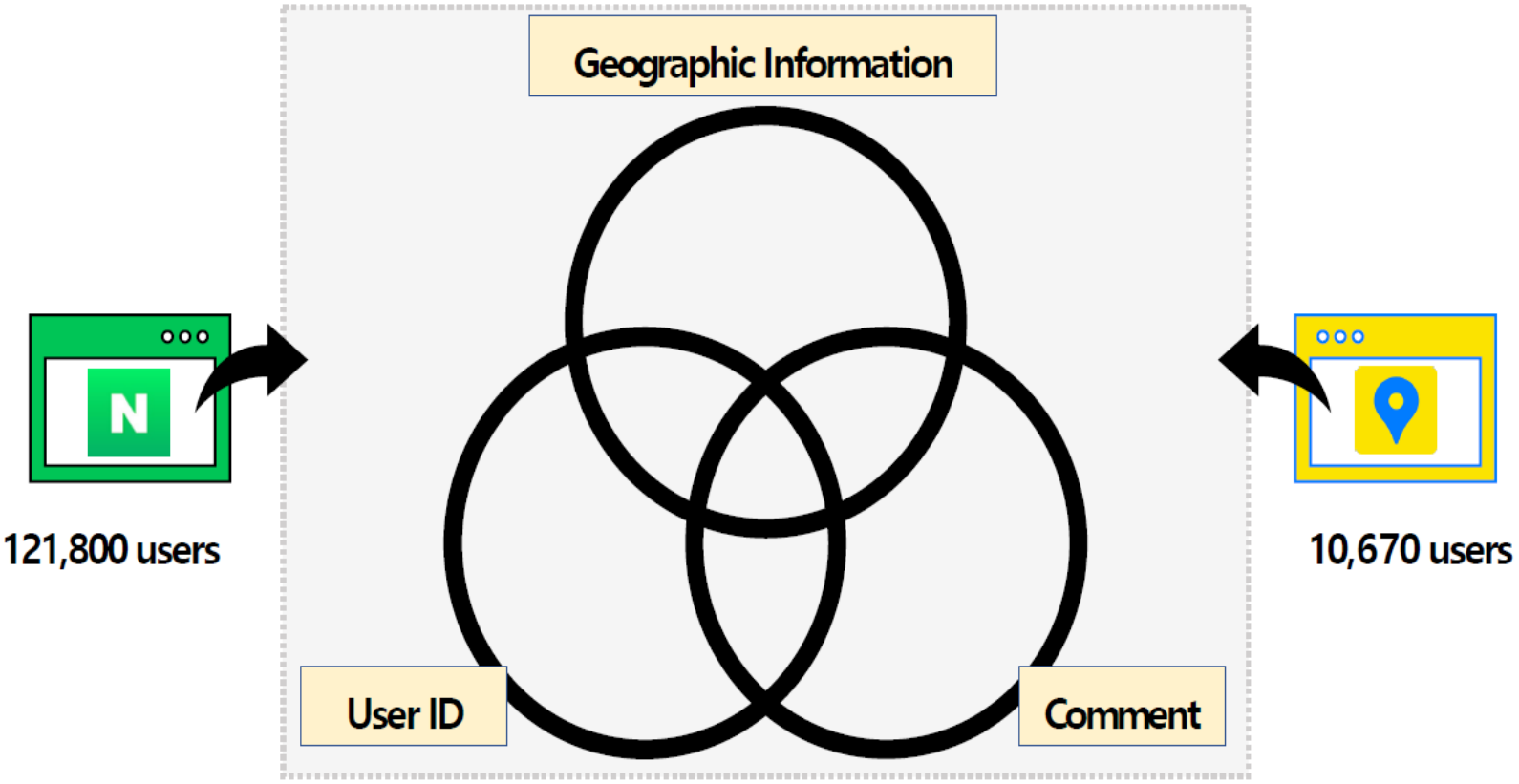
Part II　재현데이터의 한계 및 문제점

Part III　정책적 시사점

Part I Why 재현데이터?

가명정보 – 식별 가능성

- 가명정보는 추가정보와의 결합에 따른 식별가능성을 염두해두지 않기 때문에 정보보호수준 측면에서는 매우 취약함



① User ID (닉네임, ****처리)

② 리뷰 작성 스타일

③ 지리정보

①, ②, ③ 정보를 엮으면 특정인을 식별할 수 있음 (네이버 사용자 기준 약 5% 내외)

같은 방식으로 인스타그램 등 SNS에 올라오는 정보를 붙일 수 있음

가명정보 vs 익명정보

가명 정보

- 개인을 식별할 수 없도록 자료의 일부/전체를 삭제/대체하여 얻어진 정보

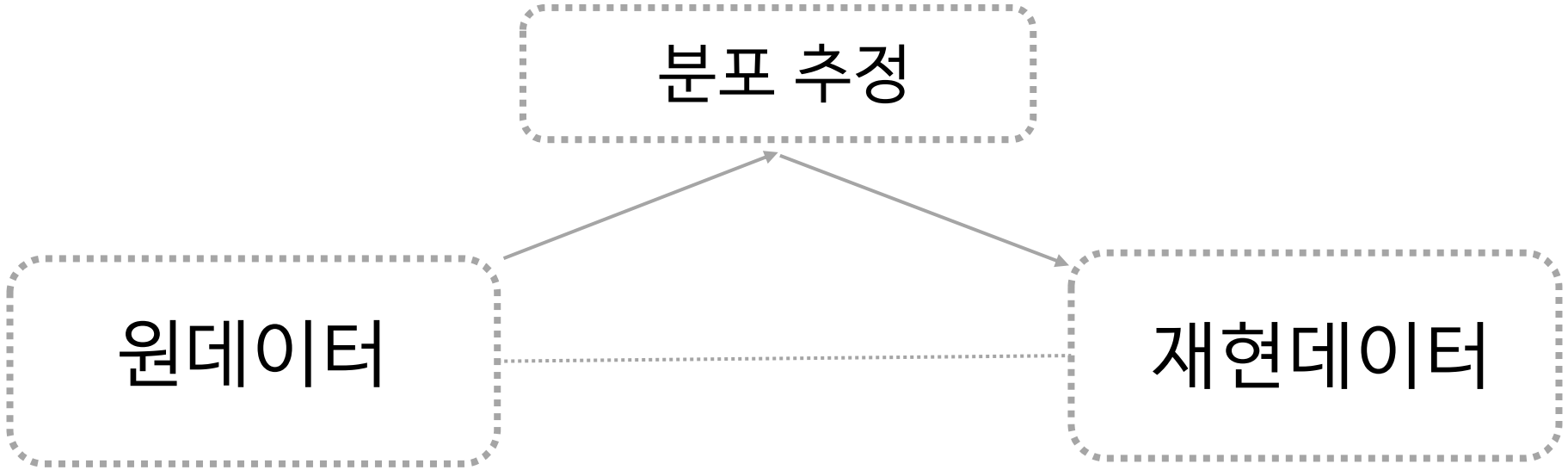
  보통 인구사회학적 정보에 마스킹을 적용하여 생성
- **추가정보**를 사용하면 식별가능
- 데이터의 **노출위험 수준**을 제어하는 것에 초점이 맞춰져 있음
- 통계작성, 과학적 연구, 공익적 기록목적 등과 관련하여 처리

익명 정보

- 개인을 식별할 수 없도록 처리된 정보를 의미함

  *보통 모의데이터 혹은 **재현데이터** (synthetic data) 형태로 생산*
- 추가정보를 사용하더라도 식별 불가능한 정보를 의미함
- 익명화 되었다는 가정하에서 **데이터의 유용성**에 초점이 맞춰져 있음
- 개인정보에 해당되지 않기 때문에 **이용범위에 제한이 없음**

재현데이터 생성



통계적 방법론 (Explicit Approach)

- 통계적 모형을 통하여 원데이터의 분포를 추정한 후에 이를 통하여 재현데이터를 생성
- 딥러닝 기반 방법론에 비하여 원자료의 구조를 더 잘 보존

딥러닝 기반 방법론 (Implicit Approach)

- 재현데이터를 우선 생성한 후에 재현데이터가 원자료와 유사해지도록 업데이트 하는 방식으로 생성
- 통계적 방법론에 비하여 대체적으로 정보보호 수준이 더 높음

재현데이터 활용사례

- 미국 인구조사국에서는 재현 데이터 기법을 활용하여 인구의 거장지역과 직장 지역 정보 등을 제공

- 미국 CMS(Center for Medicare and Medicaid Services)는 수명 주기, 인구통계, 1차 진료기록, 응급실 진료기록, 증상 기록 등을 재현데이터로 활용할 수 있도록 제공

- 한국정보화지능원-제주시 주관으로 제주시의 전입/전출 인구데이터를 KCB의 금융자료와 연계하여 재현데이터 생성

소결

- 가명정보는 식별가능성 및 데이터 왜곡 문제 등의 한계가 있어서 공익적 목적에 한정한다고 하더라도 내재된 문제가 많음
- 재현데이터는 원자료와 구조적으로 유사하도록 새로운 데이터를 생성하는 메커니즘 관점에서는 개인정보의 노출위험에서 자유로운 편
- 해외에서는 공익적 목적이 큰 경우에는 재현데이터를 적극적으로 활용하고 있음
- 통계청/통계개발원에서 재현데이터를 수년 전부터 연구를 지속적으로 하고 있으나 국가통계로 사용하는 것에 대해서는 아직까지 회의적인 시각이 많음

Part II 재현데이터의 한계 및 문제점

노출위험

- 앞선 예제처럼 **식별가능성 및 노출위험은** 데이터 생산 및 작성 시점에서 판단하기 어려운 경우가 있음
- 재현데이터는 예외가 될 수 있을까?
- 재현데이터라도 식별가능성 및 노출위험에서 자유롭지 않을 수 있음.

  아이러니하게 데이터 품질이 좋은 경우 식별 가능성 및 노출위험이 증가할 수 잇음

예1) 시계열 형태의 재현데이터

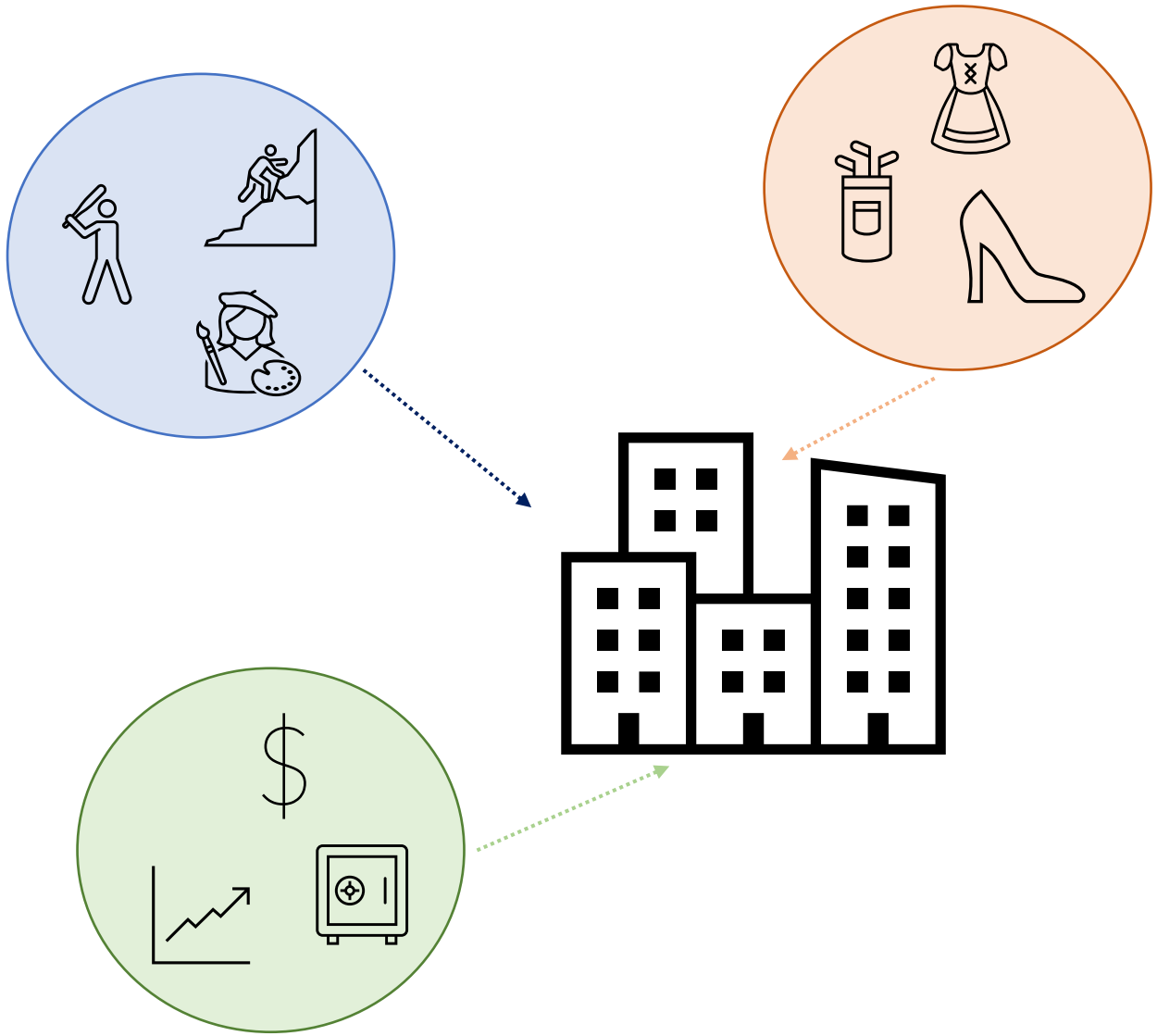
(100, 200, 300) → (95, 190, 385) 변화비 동일 → 다른 속성 정보에 대한 추론 가능

예2) 원데이터에 범주형 변수가 포함되어 있는 경우

원데이터에 없는 패턴이 재현데이터에 발생하면 데이터 품질이 떨어짐

만약에 패턴이 그대로 유지된다면 가명정보와 다를 바가 없어짐

정보독점 및 불균형



- 특정 개인을 식별하여 자료를 연계하지 않더라도 다양한 속성을 가지고 있는 데이터를 생성하는 것이 가능 (data integration, statistical matching)
- 데이터 수집 및 결합 능력에 따라서 **정보독점 및 불균형** 문제가 발생될 수 밖에 없음
- 특히 **금융/건강 관련 데이터** 중심으로 결합되어 있을 수록 **부작용**이 심각할 수 있음

예) Customer Segmentation

→ Customer Discrimination

보험 및 신용카드가입 심사

은행들의 대출심사 등

재현데이터 품질 측정 및 관리

주요 재현데이터 품질 측정 도구

ML/DL 모형 기반 평가	통계적 유사성	정보보호 수준
Accuracy AUC F1-Score	Correlation Distance Wasserstein Distance Jensen-Shannon Divergence Confidence Intervals Propensity Score Method	DCR NNDR DUPI

- 재현데이터의 품질을 측정할 수 있는 방법론은 많으나 그 방법론들이 Data Utility 혹은 Privacy Level을 제대로 측정하는지에 대한 연구 및 논의는 매우 부족함
- 이러한 환경에서 재현데이터 거래가 활성화 된다면 “악화가 양화를 구축하게 되는 현상”이 발생할 수 있음 예) 정치여론조사-통신사 제공 무선번호 데이터
- **품질이 낮은 데이터를 활용하여 만든 정책, 서비스 등의 피해는 전부 국민 혹은 소비자의 몫**

소결

- 재현데이터라고 하더라도 식별 가능성 및 노출위험에 자유롭지 않음
- 특히 재현데이터의 거래가 활성화된다면 정보 독점 및 불균형 문제가 광범위하게 발생할 수 있으며 이러한 폐해는 금융 및 건강 관련 섹터에서 가장 심각하게 나올 수 밖에 없음
- 재현데이터를 중심으로 하는 데이터 거래 및 사용을 활성화하기 위해서는 데이터 품질에 대한 측정 방법에 대한 연구 및 논의가 재현데이터 생성 방법론보다 선행되어야 함

Part III 정책적 시사점

정보독점 및 불균형 대응

- 현재 데이터 관련 정책은 “데이터 사용자”와 “데이터 수집/생산자”에 대부분 초점이 맞춰져 있으며 재현 데이터의 사용 또한 맥을 같이 함

- 데이터 수집/생산자이며 사용자인 기업의 경우 데이터 관련 규제가 약해질 수록 손쉽게 정보를 독점할 수 있으며 이로 인하여 상당한 수준의 정보 불균형 문제가 발생할 수 밖에 없으며 기업의 데이터 독점 문제와 연결됨

- 특히 데이터가 국가기관/공공기관에서 민간기관으로 한쪽으로 흐르는 경우, 공공영역과 민간영역에서의 정보 불균형 문제가 발생하기 때문에 공공-민간 영역에서의 데이터 활용에 대한 정책적 연구 및 가이드가 필요함 예) 국민건강보험공단 vs 금융회사

- **개인정보**의 **경제적 가치**와 **정보독점**으로 인한 **사회적 비용**에 대한 정책적 연구가 필요함

데이터 소유 및 이익 공유

- 데이터는 **소유**에 대한 법적/경제적 개념이 명확하지 않음 ☞ 정책적 연구와 국민적 합의가 필요

  사례1) A는 열이 나서 병원에서 진료를 받았음

  ☞ 병원에 진료 기록

  ☞ 국민건강보험공단에 진료 및 건강보험 급여가 기록 되고 이를 통하여 국가 정책 개발

  ☞ (가정) 재현데이터로 타 공공기관 혹은 민간회사에 자료가 전송되고 수익 발생

  사례2) B는 C사의 통신회사에 가입을 하였음

  ☞ 통신사에서 전화번호 할당 및 저장

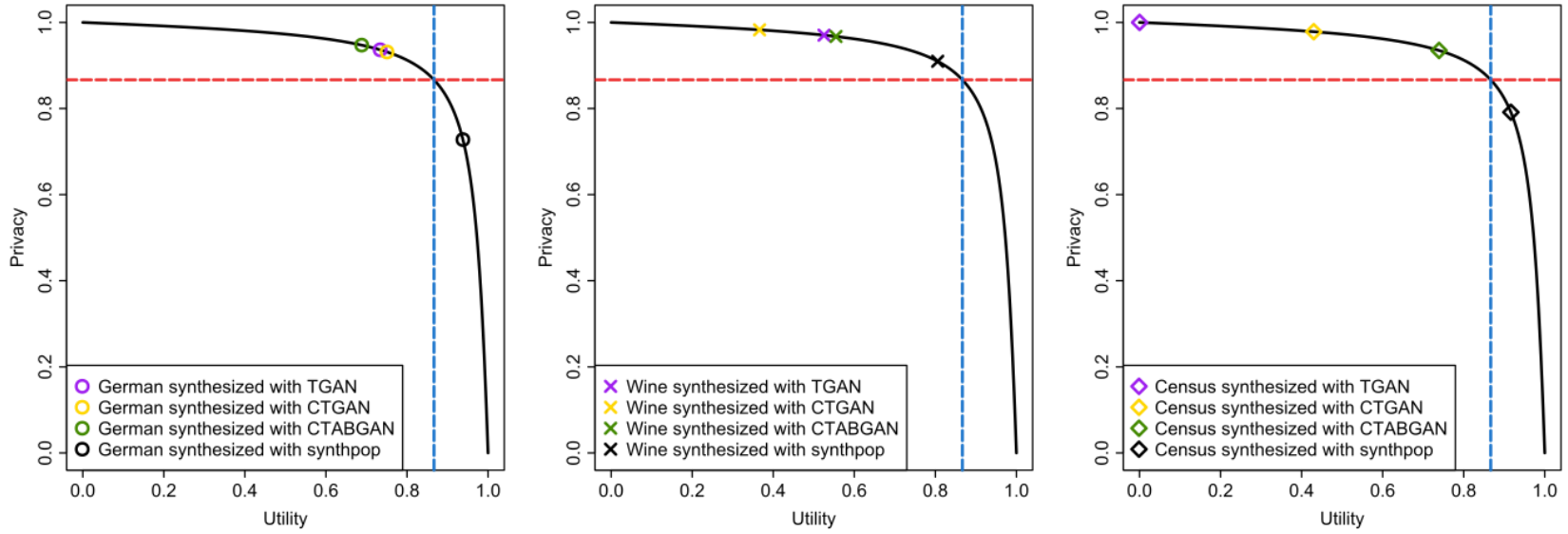
  ☞ 정치여론조사를 위하여 여론조사 회사로 전송되었으며 통신사와 여론조사회사 모두 수익 발생

- 정보제공자의 정보 사용 범위와 경제적 가치

  ☞ 재현데이터는 법적으로 개인정보가 아니기 때문에 경제적 소유권이 없는가?

  ☞ 개인정보의 경제적 가치는? 예) 마이데이터-1,000원?

비식별데이터 및 재현데이터 품질 측정



출처: Jeong, Kim, and Im (2022)

- 건전한 데이터 기반 경제를 구축하기 위해서는 **데이터 품질 측정에 대한 연구**를 독려하고 정책적으로 지향해야 함

  정보화진흥원이 최근 몇 년간 집행한 예산은 대부분 AI관련 데이터 생산 및 모형 개발에 집중되어 있었으며 데이터 관련 연구 윤리나 데이터 품질(유용성 및 정보보호 수준)에 대한 내용은 없었음

- 데이터3법이 시행된 지 2년 가까이 됐지만 여전히 K-익명성, *l*-다양성 등 단순한 측정 도구를 활용하여 정보보호 수준 및 데이터 유용성을 평가하고 있음 ☞ 2022년 IITP 재현데이터 과제

소결

- 재현데이터 사용에 대한 정보독점 및 불균형 문제는 기업의 빅데이터 규제와 관련하여 논의되어야 함
- 특히 공공영역과 민간영역에서 발생하게 될 데이터 불균형 문제에 대한 정책적 연구 및 대응 가이드 개발이 필요함
- 데이터 소유 및 권한, 경제적 가치 분배에 대한 사회적 합의 및 정책적 연구가 필요함
- 비식별 데이터 및 재현데이터에 대한 품질 측정 방법론에 대한 연구 지원 및 Unified Frame work을 제공해야 함

감사합니다 !!!

Q & A