Data Linkage & Combination between Survey Data and Rich (Big) Data for Powerful Data

2024

김 선 웅

본 슬라이드에서는

첫째, Callegaro의 미국 네브라스카 경영대학에서의 세미나 내용을 간단히 요약한다.

둘째, 빅데이터 품질 평가 및 개선 방법론을 제시한다.

셋째, Survey 데이터와 Rich 데이터 연계 및 결합 분석 방법론을 제시한다.

빅데이터 시대 ‘서베이’의 역할(Callegaro, 2017)

Importance of Surveys in the Era of Big Data

Mario Callegaro

Senior Survey Research Scientist
Ads and Commerce User eXperience team
Google London

UNL Business seminar
November 20, 2017

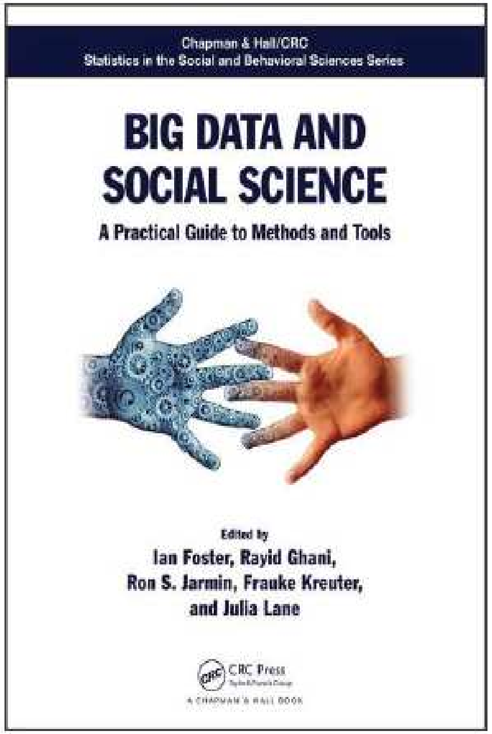
Relationship between surveys and big data

Rich data vs. BIG data

'Big Data Total Error (BDTE)'

The TE framework identifies all major sources of error contributing to data and or estimator inaccuracy (Biemer, 2016)

- **Generation**
  - Missing data
  - Self selection
  - Coverage
  - Low signal to noise ratio
  - Non representative
- **Extract**
  - Specification
  - Matching
  - Coding
  - Editing
- **Analyze**
  - Sampling
  - Selectivity
  - Weighting
  - Modeling
  - Computation



빅데이터 품질 평가 및 개선 방법론

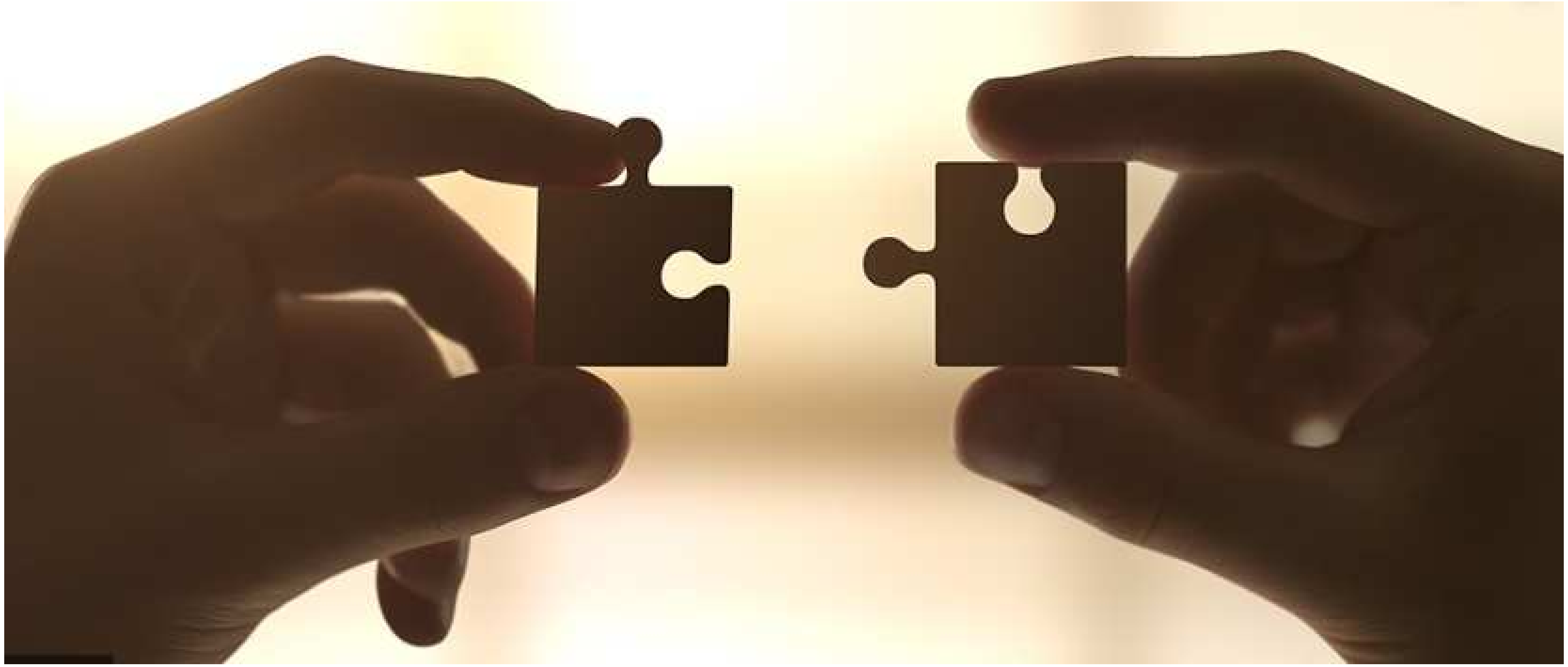
- 빅데이터의 품질(예: 대표성, 편향, 정확성)을 평가하기 위해 서베이를 직접 사용
  - 빅데이터 자체에 대한 주기적, 반복적 서베이 실시
  - 동일한 모집단을 대상으로 한 서베이 데이터 분석 결과와 빅데이터 분석 결과 비교

- BDTE 수준 판별 및 빅데이터 품질 개선
  - Rich 데이터로 전환(변환)
  - 서베이 데이터와 연계를 위한 준비 단계

Survey 데이터와 Rich 데이터 연계 및 결합 분석 방법론

빅데이터가 서베이와 만나 결합하면?

데이터 연계(Data Linkage)의 중요성



두 데이터가 만나면?(데이터 연결, Data Linkage)

- 서베이 데이터 ➕ Rich 데이터(빅데이터) = New Data
  SD RD

- 'New Data'를 'Powerful Data (PD)' 로 만들기 위한 많은 연구가 앞으로 필요함: 새로운 개척 연구 분야

데이터 품질 예시 1

- Coverage 측면: 서베이 데이터 – Excellent
  Rich 데이터(빅데이터) – Good

- Extract 측면: Rich 데이터(빅데이터)
  – (서베이 응답자) 개인별 데이터는 사용(접근) 가능하지만,
  전체 데이터는 사용(접근) 불가능할 때

Data Linkage = PD: 예시

- PD1

What
서베이 데이터

Rich 데이터

변수 1 변수 2 변수 3 -----변수 50 + 변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
-
Case 1,000

Data Linkage = PD: 예시

- **Modified PD1**

What 서베이 데이터		What Rich 데이터
변수 1 변수 2 변수 3 -----변수 50	+	변수 2 변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
-
Case 1,000

변수2의 상호 일치 여부 판단

Data Linkage = PD: 예시

- PD2

What & Why
서베이 데이터
변수 1 변수 2 변수 3 -----변수 50

\+

What
Rich 데이터
변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
-
Case 1,000

Data Linkage = PD: 예시

- **Modified** PD2

What & Why
서베이 데이터

\+

What
Rich 데이터

변수 1 변수 2 변수 3 -----변수 50 변수3 변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
-
Case 1,000

변수3의 상호 일치 여부 판단

Data Linkage = PD: 예시

- PD3

What & Why
서베이 데이터

\+

What
Rich 데이터

변수 1 변수 2 변수 3 -----변수 50 변수2 변수 51 변수 52 -----변수6789

Case 1

Case 2 × (nonresponse, missing) O

Case 3 무응답값 확인

\-

\-

\-

Case 1,000

Data Linkage = PD: 예시

- PD3 예시

What & Why

재학생 생활실태조사 데이터 + 학교 행정자료

What

	ID	평점	전공만족도 -----	평점	변수 51	변수 52 -----
학생1	010-1534-5862					
학생2	010-5212-8431	×		3.27		
학생3	010-3867-4325					
-						
-						
-						
학생300						

학생 10,428

Note. 이 경우 서베이 데이터의 정확성을 평가하는 것도 가능함

데이터 품질 예시 2

- Coverage 측면: 서베이 데이터 – Excellent
  Rich 데이터(빅데이터) – Not Good

- Extract 측면: Rich 데이터(빅데이터) – 개인별 데이터 뿐만 아니라 전체 데이터 사용(접근) 가능할 때

Data Linkage & Combination = PD: 예시

- PD4

What
서베이 데이터
변수 1 변수 2 변수 3 -----변수 50

What
Rich 데이터
변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
Case 1,000

Case 1,283,642

Considering the under-coverage of rich data,
How to conduct a combined data analysis?
(survey data – weighted data, rich data – unweighted data)
(e.g., Using sampling techniques and post-stratification in rich data)

데이터 품질 예시 3

- Coverage 측면: 서베이 데이터 – Excellent
  Rich 데이터(빅데이터) – Good

- Extract 측면: Rich 데이터(빅데이터) – 개인별 데이터 뿐만 아니라 전체 데이터 사용(접근) 가능할 때

Data Linkage & Combination = PD: 예시

- PD5

What
서베이 데이터
변수 1 변수 2 변수 3 -----변수 50

What
Rich 데이터
변수 51 변수 52 -----변수6789

Case 1
Case 2
Case 3
-
-
-
Case 1,000

Case 23,596,218

How to conduct a combined data analysis?

(survey data – weighted data, rich data – unweighted data)

(e.g., Using sampling techniques in rich data)

질 의 응 답

감 사 합 니 다