

번역논문

사회과학에서의 데이터 혁명에는 질적 연구가 필요하다^{* **}

저자: 니콜리차 그리고로폴루^{***} · 마리오 L. 스몰^{****}

번역자: 이지원^{*****}

인간 행동에 관한 연구에서 대용량 데이터가 갈수록 더 빈번하게 사용되고 있지만 오늘날 연구자들은 양질의 사회과학 지식을 생산하는 데 있어 대용량 데이터가 가지는 여러 한계를 인지하고 있다. 질적 연구는 이러한 문제 중 일부를 예방할 수 있다. 질적 연구는 우리가 수집한 데이터의 질을 더 잘 이해하고, 더 나은 연구 설계와 분석 결정을 도출하며, 분석 결과를 더 잘 해석하게끔 도와줄 수 있다.

번역자 주제어: 빅데이터, 질적 연구, 전산사회과학, 사회과학 방법론

《조사연구》는 조사연구 분야에 도움이 되는 서적이거나 해외 논문을 소개하기 위해 편집규정 3조 6항을 신설하여 서평과 번역논문을 싣는다. 익명의 동료심사를 통해 게재를 결정하는 독창적 연구, 즉 연구논문/연구단편/조사동향 등과는 달리 편집위원회에서 소개가 필요하다고 판단할 경우 서평이나 번역을 의뢰 혹은 자체 심사를 진행한다.

* 원문 서지정보: Grigoropoulou, Nikolitsa and Mario L. Small. 2022. "The Data Revolution in Social Science Needs Qualitative Research." *Nature Human Behaviour* 6(7): 904-6.

** 저자들은 본 논문의 출판을 지원한 브레멘 대학 우수학자 지원(the University of Bremen Excellence Chairs) 프로그램에 감사드린다. 또한 이 연구에 이해관계 충돌의 여지가 존재하지 않음을 밝혀 둔다.

*** 독일 브레멘 대학 SOCIUM 불평등 및 사회정책 연구센터 박사후연구원.

**** 미국 컬럼비아 대학 사회학과 교수 (mario.small@columbia.edu), 교신저자.

***** 미국 뉴욕 주립 올버니대학 사회학과 박사과정 수료 (jlee28@albany.edu).

I. 서론

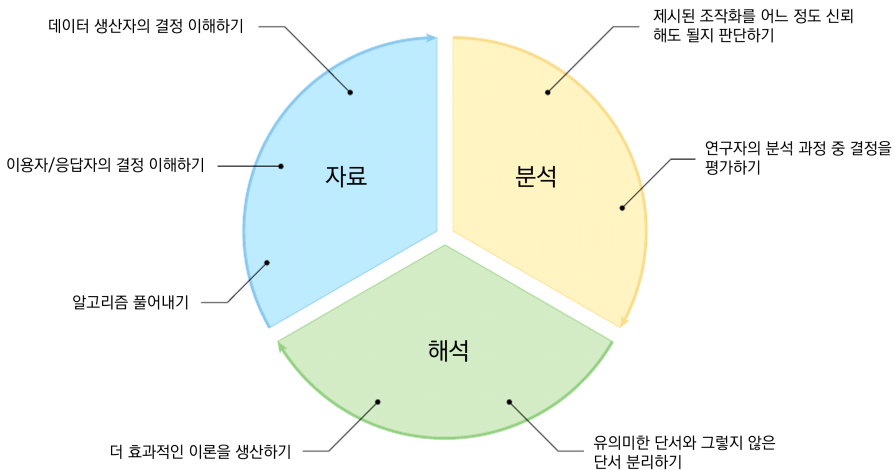
인간 행동에 관한 정보를 대용량으로 담은 데이터¹⁾의 확산은 사회과학에서 데이터 혁명의 시작을 가능하게 해주었다. 이러한 변화는 연결망의 진화, 정치적 양극화, 인종 간 불평등 같은 주제를, 과거에는 시도하기가 어려웠거나 불가능했던 방식으로 연구하도록 연구자를 자극하였다(Small et al. 2021). 그러나 오늘날 연구자들은 기업이나 기관에서 생산한 대용량 데이터셋가, 의도하지는 않았더라도, 사회현실을 잘못 재현하고 질이 낮은 과학 지식을 생산하게 만드는 특성을 가질 가능성을 우려하고 있다(boyd & Crawford 2012; Lazer et al. 2014). 우리는 이러한 문제가 표준적인 과학에 필수적이지는 않다고 여겨져 온 방법 즉, 심층 인터뷰나 현장관찰 같은 질적 연구방법을 통해 예방될 수 있다고 생각한다. 질적 방법의 사용은 연구의 배경과 맥락을 보여주기 위해서뿐만이 아니라 수집한 데이터를 더 잘 이해하고, 더 나은 연구설계와 분석 결정을 도출하며, 분석 결과를 더 잘 해석하기 위해 꼭 필요하다. 우리는 질적 연구가 ‘빅데이터’ 사회과학에 필수불가결한 이유가 최소한 일곱 가지 존재한다고 본다(<그림 1> 참조).

II. 자료의 생산에 영향을 주는 결정

질적 연구는 데이터셋 생산 과정 중 사람이 내린 결정에 대한 이해를 위해 필요하다. 전통적인 형태의 사회과학 자료는 과학적 지식 생산을 목표로 하는 연구자에 의해 만들어졌다. 반면 오늘날 대용량 행정자료는 그렇지 않은 경우가 많다. 이

1) (역주) ‘Data’는 맥락에 따라 ‘데이터’와 ‘자료’로 번갈아 가며 번역하였다. 저자들이 이 단어를 설문조사, 인터뷰, 참여관찰, 행정기관 데이터셋, 온라인 게시물 등 넓은 의미에서의 ‘자료’를 지칭하는 데 쓰고 있지만 최근 대량의 행정정보나 온라인 게시물 등 특정 자료를 “데이터”라고 칭하는 경향이 국내와 국외 모두에 있음을 함께 고려하였다.

러한 자료는 기업이나 정부 기관에 의해 주로 생산되며 이런 조직은 자기 자신의 이해관계를 염두에 두고 어떤 정보를 모을지, 그리고 연구자에게 얼마를 받고 공개할지를 결정한다. 사회과학자는 조직 내부에서 내려지는 이러한 결정에 대해 거의 잘 모를 것이고 이런 정보 부족은 연구에 심각한 영향을 줄 수 있다.



<그림 1> 질적 연구와 빅데이터: 대용량 데이터를 활용한
사회과학에서 질적 연구의 일곱 가지 역할

예컨대 페이스북은 자사가 보유한 자료에 연구자가 접근할 수 있게끔 소셜 사이언스 원(Social Science One)이라는 플랫폼을 만들었다. 그러나 2021년 9월 언론은 페이스북이 페이스북 활동으로 정치적 성향을 분명하게 식별할 수 없는 사람들에게 대한 정보, 미국 이용자의 약 절반에 해당하는 양의 정보를 의도 여부는 확실치 않지만 공개하지 않았다고 지적했다(Alba 2020). 페이스북 제공 자료로 허위 정치정보(political misinformation) 같은 주제를 연구했던 학자는 자신도 모르게 편향된 표본을 분석하고 연구 결과에서 잘못된 결론을 도출하고 있었던 것이었다.

대용량 데이터가 어떤 과정을 통해 생산되었는지 이해하는 절차는 데이터 분석을 위한 기초 절차가 되어야만 한다. 다시 말해, 해당 자료를 생산하기 위해 관리자,

프로그래머, 그 조직의 다른 직군 사람들이 무엇을 했고 왜 그렇게 했는지를 이해해야만 한다. 자료 생산 과정에서 결정 권한을 가진 인물과의 질적 인터뷰는 자료를 이해하는 데 꼭 필요할 것이다. 보건 분야에서 자료의 질을 평가해온 연구자들은 이런 종류의 인터뷰 방법을 통해 보건 자료 생산자의 개인적 판단, 조사 기준 차이 등이 최종 자료의 질에 미치는 영향을 밝혀낼 수 있다는 점을 보여준 바 있다 (Strong, Lee, & Wang 1997).

III. 응답자가 내리는 결정

질적 연구는 자료의 부분 부분을 구성하는 사람이 내린 결정을 이해하기 위해 필요하다. 자료를 입력하는 시스템을 만드는 것은 조직이다. 그렇지만 SNS에 무엇을 게시할지, 병원에서 제공되는 서류에 무슨 정보를 입력할지, 세무원에게 무엇을 보고할지 등 어떤 정보가 입력될지는 결국 자료의 부분 부분을 직접 채워나가는 고객, 손님, 이용자에게 달려 있다. 이들의 결정에 영향을 주는 심리학적 그리고 사회적으로 요인-터놓고 말하자면, 편견, 공포, 프라이버시 민감도, 동료로부터의 압력 등-은 자료의 성격에 영향을 준다. 이 영향은 해당 자료를 연구에 활용할 때 고려될 필요가 있다.

예를 들어보자. SNS 이용자들은 각자 다른 강도로 자신이 속한 사회적 연결망 내에서 자신에 대한 타인의 평판을 걱정한다(Ruths & Pfeffer 2014). 따라서 이용자들은 서로 비슷한 생각을 하고 있더라도 무엇을 게시할지, 어디에 게시할지 각자 다른 결정을 내릴 것이며, 그로 인해 연구자가 관찰하는 게시물은 이용자의 생각을 진정으로 반영하고 있지 못할 것이다. SNS 자료를 활용한 허위 정치정보 연구는 이 문제로 인하여 어려움에 직면할 수 있다. 이 문제를 해결하기 위해 연구자가 이용자의 실제 생각을 알아보는 설문조사를 성공적으로 수행하고자 한다면 우선 개방형 인터뷰를 포괄적으로 시행해야 한다. 연구자와 인터뷰 참여자 사이에 천천히 쌓아 올린 신뢰에 기반한 인터뷰는 연구자가 인터뷰 전에는 물어볼 생각조차 못 했을 질문에 대한 답변까지 찾아낼 수 있을 것이다.

IV. 알고리즘의 특성

질적 연구는 알고리즘 뒤에 숨겨진 요소를 파악하기 위해 필요하다. 알고리즘이란 전산시스템이 특정한 기능을 수행하거나 문제를 해결하게 만들기 위해 입력해야 하는 지시사항을 순서대로 묶어놓은 것이다. 기업은 이용자에게 노출될 화면을 구성하고, 의심스러운 이메일을 걸러내며, 신용점수를 결정하는 등의 일을 하기 위해 알고리즘을 이용한다. 그리하여 알고리즘은 이용자의 행동과 추후 연구자가 활용하게 될 데이터의 특성에 큰 영향을 준다.

많은 경우 알고리즘은 전매특허품이기 때문에 알고리즘이 데이터에 어떤 영향을 주는지에 대해 연구자로서는 거의 알아내기 어려울 것이다. 그렇기에 자료에서 관찰되는 패턴이 근본적으로는 알고리즘 자체에 의해 조형된 것인지 확인하기도 어려울 것이다(Lazer et al. 2014; Ruths & Pfeffer 2014). 예컨대 트위터의 알고리즘은 흑인의 얼굴 이미지를 처리하는 방식 때문에 ‘인종주의적’이라는 지적을 자주 받아왔다. 관련된 여러 문제 중 하나는 트위터 이용자가 백인과 흑인 얼굴이 모두 포함되어 있고 용량이 너무 큰 사진을 올리면 알고리즘이 이용자의 타임라인에 나오는 사진 미리보기에서 흑인은 잘라내고 백인의 얼굴만 등장시킨다는 것이다. 알고리즘의 이 결정이 연구에 어떤 영향을 줄지 고려해 보자. 만약 연구자가 트위터 데이터를 프로그래밍으로 일괄적으로 긁어모았다면 업로드된 사진도 수집했을 것이고, 그 사진은 이용자가 타임라인에서 보는 미리보기 버전이 아니라 업로드된 원본일 것이다. 이 경우, 미리보기 자동 편집 과정에서 잘라내기 알고리즘의 영향은 데이터세트에 반영되지 않는다. 이 데이터세트를 활용하는 연구자는 이용자가 단체 사진에서 흑인의 얼굴을 무시한다고 잘못 추론할 위험에 처하게 된다. 이용자는 그저 알고리즘의 결정 때문에 흑인의 얼굴을 덜 보게 되었을 뿐임에도 말이다.

알고리즘을 이해하는 일은 중요하다. 그러나 알고리즘이 어떻게 기능하고 데이터 세트에 어떤 편향을 삽입하는지 평가하기 위해 꼭 기업이 특허권을 가진 알고리즘 코드를 알아내야만 하는 것은 아니다. 알고리즘은 “기업 소속 프로그래머” 그리고

그 모델을 훈련하고 정교화시킨 사람들이 “내린 수천 번의 결정”(Lazer et al. 2014)을 통해 만들어진 결과물이다. 이렇게 사람들이 내린 결정을 이해하는 데 인터뷰와 현장 관찰은 우리가 사용할 수 있는 가장 강력한 도구에 속한다. 인터뷰는 사람들이 어떻게 결정을 내리는지를 밝혀내기 위해 고안된 방법이고, 수개월 동안 관련 직장의 구성원으로 생활하며 알고리즘의 생산을 내부자의 관점에서 이해하는 에스노그래피가 수행되기도 했기 때문이다(Christin 2018).

V. 변수를 개념에 연결하기

질적 연구는 특정 개념의 조작화 방식을 어느 정도 신뢰할 수 있을지 이해하는 데 필요하다. 대용량 데이터를 사용할 때 연구자는 과학적 연구를 목적으로 생성되는 많은 변수를 과학적 목적을 위해 재해석하곤 한다. 그러나 이렇게 변수에 새롭게 부여된 의미는 부정확하거나 부적합할 수 있다. 일례로 비공식적 멘토링 관계가 과학자의 커리어에 주는 영향에 관한 최근의 연구 하나를 검토해 보자(AlShebli, Makovi, & Rahwan 2020). 연구진은 주로 10개의 STEM(Science, Technology, Engineering, and Mathematics) 분야에서 출판된 논문의 인용 자료를 획득했고, 그중에서 신진 학자와 중견 학자의 공동 저술을 비공식적 멘토링 관계의 지표로 해석했다. 논문의 발견 중 하나는 여성 신진 학자와 여성 중견 학자 사이의 공저-연구진의 조작화에 따르면 여성 간 비공식 멘토링 관계-가 멘토에게나 멘티에게나 본인의 연구가 더 낮은 빈도로 인용되는 현상과 상관관계를 보인다는 것이었다. 이 결과는 연구진에게 여성 학자 사이의 멘토링을 권장하는 관련 정책의 유효성을 의문시하게끔 만들었다.

이 논문은 비공식 멘토링의 범주를 공동 저술로 축소하고 다른 여러 차원은 무시했다는 비판을 받았으며 지금은 논문 게재가 철회되었다. 해당 연구진의 입장을 나름 고려해 보자면 그들은 공동 저술이 좋은 지표인지 판단하기 위하여 설문조사 자료를 사용했었다. 그러나 이 타당성 확인 설문조사는 질문이 포괄하는 범위(scope)가 너무 좁고 사전테스트를 시행하지 않았기에 성공하지 못했다. 더 효과적이었을

접근법은 잠재적 응답자가 비공식 멘토링을 어떻게 이해하는지, 무엇이 적절한 지표가 될지, 설문조사 질문이 어떻게 해석되는지에 관해 인터뷰를 한 번 가져보는 것이다. 익히 알려져 있듯이 이러한 ‘인지 인터뷰(cognitive interviewing)’는 더 효과적인 설문조사 질문지를 만들고 분석 결과를 더 섬세하게 해석할 수 있게 도와주었을 것이다. 개념의 조작화는 과학의 핵심을 구성한다. 그리고 효과적인 인터뷰는 많은 경우-새로운 목적을 위해 재해석되는 대용량 데이터의 변수와 관련하여-해석의 신뢰도를 증진하는 데 도움이 될 것이다.

VI. 분석 과정 중 결정

질적 연구는 연구자가 분석 과정 중 내리는 결정을 이해하는 데 필요하다. 많은 대용량 데이터는 거의 무한히 많은 수의 각기 다른 방식으로 연구될 수 있다. 그렇기에 연구자는 반드시 주관적인 결정을 여러 번 내려야만 한다. 이 유연성, 최근에 ‘연구자의 자유도’(Simmons, Nelson, & Simonsohn 2011)라고 불리게 된 대용량 데이터 분석의 특징은 의심스러운 연구 결과를 만들어 낼 수도 있다. 이러한 대용량 데이터의 특성을 비판하는 사람들은 가설의 사전등록 등을 해결책으로 제시하는데 이는 타당한 지적이다. 하지만 이런 해결책은 연구자들이 애초에 왜 서로 다른 결정을 내리는지를 설명하는 데는 거의 도움이 되지 않는다. 이를 설명하기 위해서는 연구자의 분석 결정에 관한 연구가 필요하다.

최근 한 연구는 같은 가설을 같은 자료로 검증한 73개 팀의 분석 코드를 수집하여 분석 결정이 결론에 어떻게 영향을 주는지 이해하고자 하였다(Breznau et al. 2021). 이 팀들은 서로 다른 분석 경로를 택했고 이들의 핵심 결론은 우려스럽게도 상이하며 심지어 서로 충돌하기까지 했다. 여기서 주목할 점은 이 연구가 통계적 분석으로는 이러한 결론의 차이 중 일부, 연구팀의 전문성, 본래 가지고 있던 생각과 기대에 기초한 일부만을 설명할 수 있었다는 것이다. 연구진은 분석된 연구팀들의 결정 뒤에 있는 특이성을 더 잘 이해할 필요가 있었으며, 심층 인터뷰와 초점집단 인터뷰 연구 방법은 이런 주제를 연구하는 데 최적화된 방법이다.

VII. 자료에서 이론으로, 그리고 다시 자료로

질적 연구는 효과적인 이론의 생성을 위해 필요하다. 전산사회과학(computational social science)의 성장은 대용량 데이터의 가치를 증진했으며, 이러한 자료는 인간 행동에 대한 우리의 예측 역량을 향상시켜 주었다. 그러나 좋은 이론이 없다면 양질의 자료를 최첨단의 방법으로 분석해도 예측은 실패할 것이다. 최근 치러진 한 데이터 분석 경연의 예를 살펴보자(Salganik et al. 2020).

이 경연은 160개 연구팀에게 4,000가구 이상의 미국 가족을 대상으로 약 13,000개에 가까운 변수를 조사한 종단형 자료를 이용해 이들의 생애 결과를 예측하는 모델을 제출하도록 했다. 연구팀은 설문조사 자료 중 1회차에서 5회차까지의 가족 배경 자료와 예측해야 하는 생애 결과가 들어 있는 6회차 설문조사 자료의 약 절반을 전달받았고 모델 생성에 있어서는 폭넓은 재량권을 보장받았다. 결과는 실망스러웠다. 가장 우수한 모델들도 가장 잘 예측된 결과에서의 변이를 20% 이상 설명하지 못했다(Salganik et al. 2020). 이 경연을 조직한 양적 연구자들은 예측에 더 효과적인 이론을 생산하기 위해서는 심층 인터뷰, 특히 독특한 사례에 대한 인터뷰가 필요하다는 결론을 내렸다(Salganik, Maffeo, & Rudin 2020). 여러 종류의 사회적 결과를 성공적으로 예측하는 데 효과적인 이론은 꼭 필요할 것이다. 그리고 대용량 데이터세트만으로는 그런 이론을 만들 수 없다. 질적 연구는 대용량 데이터의 어떤 부분에 주의를 기울여야 할지 그리고 관찰된 요소를 어떻게 이론화할지에 관한 통찰을 제공해 줄 수 있다.

VIII. 노이즈의 존재를 인지하기

질적 연구는 모호한 패턴에서 의미를 잘못 추론하는 위험을 회피하는 데 필요하

다. 학자들은 대용량 데이터가 모호한 패턴에 마주했을 때 얼핏 보기에 의미 있어 보이는 연결을 찾아내려는 인간의 인지적 경향을 자극할 수 있다고 지적해 왔다 (boyd & Crawford 2012). 질적 연구는 모호한 패턴을 정확하게 이해하는 데 도움을 줄 수 있다. 예를 들어, 파티 참가자의 스마트폰 위치추적 자료를 활용한 최근의 한 연구는 핸드폰 사이의 물리적 근접도를 기준으로 사회적 접촉을 연구했다(Blok et al. 2017). 그러나 연구진은 사람들이 때로는 서로 밀착해 서 있어도 대화하지는 않는 등 높은 물리적 근접도가 항상 유의미한 사회적 접촉을 의미하지는 않다는 것을 깨닫게 되었다. 분석가가 모든 접촉에 손쉽게 의미를 가져다 붙이면 우연한 만남도 패턴 일부를 구성한다고 잘못 해석하게 된다는 점을 알게 된 것이다. 이후 연구진은 유관성 있는 정보를 유관성 없는 정보로부터 구분하고, 사건이 발생한 현장의 분위기를 재구성하고, 관찰된 패턴에 대해 신뢰할 만한 해석을 제시하는 데 도움을 얻기 위해 직접 현장에서 에스노그래피 자료를 수집했다.

IX. 결론

우리는 대용량 데이터가 신뢰성 높은 사회과학 지식을 생산하는 데 쓰일 수 있도록 질적 연구를 활용할 수 있는 방법의 목록을 제시했다. 이 목록이 가능한 모든 질적 연구 활용법을 포괄하고 있지는 않다. 대용량 데이터를 다루는 모든 연구가 여기 제시된 모든 방법을 써야만 하는 것도 아니다. 하지만 이 목록은 가용할 수 있는 대용량 데이터가 증가하고 있는 현재 상황이 질적 연구를 덜 중요하게 만들지 않았으며 오히려 더 중요하게 만들었다는 점을 분명하게 보여준다. 사회과학 지식의 생산을 위해서는 데이터에서 패턴을 읽어낼 수 있는 능력 이상이 필요하다. 데이터가 적합한지 그리고 잘 이해되었는지, 발견된 패턴이 유의미한지 또 적절하게 해석되었는지, 과학자 스스로가 자신이 잠재적으로 가지고 있는 편향과 한계에 대해서 의식하고 있는지에 대한 지식도 필요하다. 이를 고려할 때, 인터뷰 연구자와 에스노그래피의 연구 활용은 사회과학에서의 데이터 혁명 성공을 위한 핵심 조건이다.

참고문헌

- Alba, Davey. 2020. "Tracking Viral Misinformation." *The New York Times*, <https://go.nature.com/3Kszjyw> (검색일자: 2021년 9월 15일).
- AlShebli, Bedoor, Kinga Makovi, and Talal Rahwan. 2020. "RETRACTED ARTICLE: The Association between Early Career Informal Mentorship in Academic Collaborations and Junior Author Performance." *Nature Communications* 11(1): 5855.
- Blok, Anders, Hjalmar B. Carlsen, Tobias B. Jørgensen, Mette M. Madsen, Snorre Ralund, and Morten A. Pedersen. 2017. "Stitching together the Heterogeneous Party: A Complementary Social Data Science Experiment." *Big Data & Society* 4(2): 2053951717736337.
- boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15(5): 662-79.
- Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Hrik Kenneth Andersen, Daniel Auer, Flavio Azevedo, Oke Bahnsen, Dave Balzer, Gerrit Bauer, Paul Bauer, Markus Baumann, Sharon Baute, Verena Benoit, Julian Bernauer, Carl Berning, Anna Berthold, Felix S. Bethke, Thomas Biegert, Katharina Blinzler, Johannes Blumenberg, Licia Bobzien, Andrea Bohman, Thijs Bol, Amie Bostic, Zuzanna Brzozowska, Katharina Burgdorf, Kaspar Burger, Kathrin Busch, Juan Carlos Castillo, Nathan Chan, Pablo Christmann, Roxanne Connelly, Christian S. Czymara, Elena Damian, Alejandro Ecker, Achim Edelmann, Maureen A. Eger, Simon Ellerbrock, Anna Forke, Andrea Gabriele Forster, Chris Gaasendam, Konstantin Gavras, Vernon Gayle, Theresa Gessler, Timo Gnambs, Amélie Godefroidt, Max Grömping, Martin Groß, Stefan Gruber, Tobias Gummer, Andreas Hadjar, Jan Paul Heisig, Sebastian Hellmeier, Stefanie Heyne, Magdalena Hirsch, Mikael Hjerm, Oshrat Hochman, Andreas Hövermann, Sophia Hunger, Christian Hunkler, Nora Huth-Stöckle, Zsafia Ignacz, Laura Jacobs, Jannes Jacobsen, Bastian Jaeger, Sebastian Jungkunz, Nils Jungmann, Mathias Kauff, Manuel Kleinert, Julia Klinger, Jan-Philipp Kolb, Marta Kolczyńska, John Seungmin Kuk, Katharina Kunißen, Dafina

- Kurti Sinatra, Alexander Greinert, Philipp M. Lersch, Lea-Maria Löbel, Philipp Lutscher, Matthias Mader, Joan Eliel Madia, Natalia Malancu, Luis Maldonado, Helge Marahrens, Nicole Martin, Paul Martinez, Jochen Mayerl, Oscar Jose Mayorga, Patricia A. McManus, Kyle Wagner, Cecil Meeusen, Daniel Meierrieks, Jonathan Mellon, Friedolin Merhout, Samuel Merk, Daniel Meyer, Leticia Micheli, Jonathan J.B. Mijis, Cristóbal Moya, Marcel Neunhoeffer, Daniel Nüst, Olav Nygård, Fabian Ochsenfeld, Gunnar Otte, Anna Pechenkina, Christopher Prosser, Louis Raes, Kevin Ralston, Miguel Ramos, Arne Roets, Jonathan Rogers, Guido Ropers, Robin Samuel, Gergor Sand, Ariela Schachter, Merlin Schaeffer, David Schieferdecker, Elmar Schlueter, Katja Schmidt, Regine Schmidt, Alexander Schmidt-Catran, Claudia Schmiedeberg, Jürgen Schneider, Martijn Schoonvelde, Julia Schulte-Cloos, Sandy Schumann, Reinhard Schunck, Juergen Schupp, Julian Seuring, Henning Silber, Willem W.A. Slegers, Nico Sonntag, Alexander Staudt, Nadia Steiber, Nils Steiner, Sebastian Sternberg, Dieter Stiers, Dragana Stojmenovska, Nora Storz, Erich Striessnig, Anne-Kathrin Stroppe, Janna Teltemann, Andrey Tibajev, Brian B. Tung, Giacomo Vagni, Jasper Van Assche, Meta Van Der Linden, Jolanda Van Der Noll, Arno Van Hootehem, Stefan Vogtenhuber, Bogdan Voicu, Fieke Wagemans, Nadja Wehl, Hannah Werner, Brenton M. Wiernik, Fabian Winter, Christof Wolf, Yuki Yamada, Nan Zhang, Conrad Ziller, Stefan Zins, Tomasz Żółtak, and Hung Hoang Viet Nguyen. 2021. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty." preprint. MetaArXiv. doi: 10.31222/osf.io/cd5j9.
- Christin, Angèle. 2018. "Counting Clicks: Quantification and Variation in Web Journalism in the United States and France." *American Journal of Sociology* 123(5): 1382–1415.
- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176): 1203-5.
- Ruths, Derek and Jürgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346(6213): 1063-64.
- Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis,

- Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E.H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15): 8398-8403.
- Salganik, Matthew J., Lauren Maffeo, and Cynthia Rudin. 2020. "Prediction, Machine Learning, and Individual Lives: An Interview with Matthew Salganik." *Harvard Data Science Review* 2(3).
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11): 1359-66.
- Small, Mario L., Armin Akhavan, Mo Torres, and Qi Wang. 2021. "Banks, Alternative Institutions and the Spatial-Temporal Ecology of Racial Inequality in US Cities." *Nature Human Behaviour* 5(12): 1622-28.
- Strong, Diane M., Yang W. Lee, and Richard Y. Wang. 1997. "Data Quality in Context." *Communications of the ACM* 40(5): 103-10.

The Data Revolution in Social Science Needs Qualitative Research

Nikolitsa Grigoropoulou
(University of Bremen)

Mario L. Small
(Columbia University)

Translated by
Ji-won Lee
(State University of New York at Albany)

Although large-scale data are increasingly used to study human behaviour, researchers now recognize their limits for producing sound social science. Qualitative research can prevent some of these problems. Such methods can help to understand data quality, inform design and analysis decisions and guide interpretation of results.

Key words (added by the translator): big data, qualitative research, computational social science, social science methodology