『한국사회학』 제49집 제2호(2015년), pp. 161~192

연구논문

빅데이터와 사회과학하기:
자료기반의 변화와 분석전략의 재구상*

한 신 갑**

빅데이터와 사회과학하기의 관계에 논의의 초점을 맞추는 이 글에서는 이론–방법–자료를 꼭짓점으로 하는 삼각형의 구도를 사회과학하기의 기본 틀로 잡고 논의를 시작하여, 빅데이터가 가져온 자료기반의 변화와 그 특성을 살펴보고, 그에 따른 분석전략을 가늠해보고자 한다. 사회과학에서 자료가 차지하는 위치와 그동안의 변화 추세, 그리고 빅데이터의 출현을 가져온 배경과 그 영향에 대해 논의하면서 자료의 양과 질에 있어서의 변화가 오랜 기간 계속되어온 것이라는 점을 짚고, 그 연장선상에 빅데이터의 출현도 자리매김한다. 빅데이터 시대가 가져온 새로운 사회과학의 시제품 역할을 하면서 그 가능성과 한계를 극명하게 보여준 구글 독감추세예측 서비스를 취사선택의 필요성과 근거, 방향을 경험적으로 보여주는 유용한 실제 사례로 검토한다. 이 논의들을 원론적 사회과학하기의 틀 안에서 다시 정리하고 이 비판적 검토에 근거하여 앞으로의 전망을 제시하는 것으로 글을 맺는다. 글머리에 제시한 삼각구도의 요소들 각각의 내용과 상대적 비중, 그들 간 연결의 구조와 방향 등은 재조정되어야 하겠지만, 크게는 이 기본 틀의 개조와 확장을 통해 빅데이터의 잠재력을 수용할 수 있을 것이라 조심스러운 진단을 내린다.

주제어: 빅데이터, 사회과학 방법론, 자료기반/자료환경, 분석전략, 인과관계/상관관계

* 이 문제에 대해 생각할 기회를 마련해주고, 토론을 통해 내용을 구체화시켜준 연구팀의 동료 교수들(김청택, 이봉주, 이재현, 강정원)에게 감사드린다. 원고작성 과정에서는 이상직과 김영진이 도움을 주었다. 심사과정에서 빈 곳을 지적해주시고 제언해주신 논평자들께도 감사드린다. 이 논문은 2013년 서울대학교로부터 <기초학문 사회과학분야 인재육성 기반연구> 지원을 받아 수행된 연구이다.

** 서울대학교 사회학과 교수(shinkaphan@snu.ac.kr)

Ⅰ. 들어가기

> 볼만하다는 마술 기술들은 모두 세 단계로 구성되어 있다. “플렛지(Pledge)”라고 부르는 내어놓기가 그 첫 번째 단계로 카드 뭉치나 비둘기 또는 사람 같은 일상적인 것을 내어놓는다. 그것들을 보여주면서 진짜인지, 손 댄 곳은 없는지, 정상적인 것인지 잘 살펴보라고 한다. 물론 대개 그것들은 보기와는 다르다. 두 번째 단계는 “턴(Turn)”이라고 부르는 바꾸어 놓기다. 마술사가 그 일상적인 것들을 무언가 예상치 못했던 것으로 바꾸어 놓는다. 이 단계에서 보고 있던 사람들은 마술의 비밀을 알아내고 싶어 하지만 답을 찾지는 못한다. 제대로 보고 있지 않기 때문이다. 구태여 답을 꼭 찾으려는 것도 아니고, 마술에 속아 넘어가도 좋다고 생각하고 있기도 하다. 하지만 아직 박수를 치기에는 이르다. 왜냐하면 뭔가를 사라지게 하는 것만으로는 충분치 않기 때문이다. 그것을 되돌려 놓기까지 해야 한다. 이것이 바로 모든 마술에서 가장 어려운 “프레스티지(Prestige)”라는 세 번째 단계다.
>
> —크리스토퍼 프리스트(Christopher Priest),『프레스티지(The Prestige)』(1995)

매년 ‘사회구조연구’라는 과목을 강의한다.[1] 대학원생들을 상대로 사회과학은 이렇게 하는 것이라고 가르치는 강의이다. 그 강의에서 여러 번 강조하는 메시지 중의 하나가 좋은 연구를 위해서는 이론, 방법, 자료가 모두 필요하고, 이 세 요소가 서로 잘 맞물리고 어울려 선순환하는 구조를 이룰 때 가장 힘 있는 사회과학 연구의 짜임새가 만들어 진다는 것이다.[2] 그림이 조금씩 다를 수는 있지만, 이런 삼각형(三角形) 정립(鼎立)모형의 기본적인 틀은 기존의 여러 방법론 교과서에서 쉽게 찾아볼 수 있다. 내 강의에서처럼 방법과 자료의 현실적 비중이나 능동적 역할을 훨씬 더 강조하는 경우도 있기는 하지만(Duncan, 1984), 이 세 요소 중에 가장 크고 무거운 자리가 주어지는 것은 (명시적으로든, 암시적으로든) 대부분 이론이다(Babbie, 2012). 세 요소 간의 관계에 주목해 그 중 어느 하나가 바뀌면 다른 요소들도 따라 변화하면서 새로운 균형을 만들어 낸다는 동태적인 시각을 취할 때에도, 이 순환고리의 시작과

1) 강의의 영어명은 “Making an Argument for Social Structure”이다.
2) 사회학 분야로 한정해 본다면, 이런 선순환의 성공적 사례로는 지위획득모형(Status Attainment Model)의 위스콘신학파, 조직생태모형(Organizational Ecology Model)의 스탠포드학파를 들 수 있다.

끝이 이론이라는, 또는 이론이어야 한다는 것이 전통적인 시각이다.[3]

이런 입장을 축약적으로 잘 담아내고 있는 것이 "돌을 쌓아 집을 만드는 것처럼 과학은 사실(자료)로 만들어 진다. 그렇지만 돌무더기가 집이 아닌 것처럼 사실(자료)의 더미가 과학은 아니다"라는 19세기의 과학자 앙리 프앙카레(Henri Poincaré)의 말이다(2011[1901]).[4] 물론 돌이 많으면 그만큼 큰 집을 지을 수 있는 여지가 생기지만 그 자체가 좋은 집을 짓는 충분조건은 아니다. 오히려 돌이 많아질수록 그것들을 어떻게 쌓느냐가 더 중요해진다. 즉 자료를 연결하는 틀과 조직하는 시각이 — 다시 말하면, 넓은 의미의 이론과 방법론이 — 여전히 핵심이라는 말이다(Abend, 2008; Hunt, 1985).

비록 교과서적 규범에 그치는 것이었을지라도 이런 식으로 사회과학하기의 원칙을 배우고 가르치기 시작했던 세대들에게 '빅데이터(Big Data)'를 둘러싼 최근의 논의는 그저 낯선 정도를 넘어 혼돈스럽기까지 하다. 한편에는 그동안의 많은 실리콘 밸리발(發) "혁명적 신기술들"이 그랬던 것처럼 "과장광고 싸이클"을 타고 반짝하다가 곧 철 지난 유행이 되어 버릴 수도 있다는 회의론이 있다.[5] 새로운 것에 대한 막연한 호기심과 기대, 몇몇 성공사례에 근거한 확대추론이 뒤섞여 있다고 보는 것이다. 하지만 빅데이터가 커다란 변화의 가능성을 제시한다는 것도 맞는 것 같다. 망원경이 천체의 세계를, 현미경이 세균의 세계를 들여다 볼 수 있게 해 준 것처럼, 빅데이터가 불과 얼마 전까지 상상도 할 수 없었던 규모의 자료를 수집하고 분석할 수 있게 함으로써 그동안 모르고 있던 방식으로 인간의 세계를 들여다 볼 수 있게 해 줄 것이라는 기대는 널리 퍼져 있다(Lazer, Pentland, Adamic, Aral, Barabási, Brewer, Christakis, Contractor, Fowler, Gutmann, Jebara, King, Macy, Roy, and Van Alstyne, 2009). 그러나 그 새로운 시대가 "자료가 스스로 말하고, 따라서 이론

3) 즉 이론이 "primus inter pares (first among equals)"의 위치를 차지한다. 이런 입장을 가장 전형적으로 표현한 것이 배비(Earl Babbie)의 대표저서인 *The Practice of Social Research*이다. 1975년 처음 출간된 이후, 학부 방법론 과목에 교과서로 널리 쓰이면서 베스트셀러가 되었고, 그 13판이 2012년 출간되었다. 이 책은 국내에서도 2007년과 2013년에 『사회조사방법론』이란 제목으로 번역되어 널리 사용되고 있다.

4) Science is built with facts as a house is with stones — but a collection of facts is no more a science than a heap of stones is a house.

5) 가트너(Gartner)사에서 제시한 "과장광고 싸이클"의 다섯 단계라는 틀로 보면(www.gartner.com/technology/research/methodologies/hype-cycle.jsp) 빅데이터는 지난 몇 년간 두 번째 단계인 과잉기대의 정점(Peak of Inflated Expectation)에 있었다고 볼 수 있고, 2014년에 들어서는 그 다음 단계인 내리막 길(Trough of Disillusionment)로 접어들기 시작했다는 느낌도 든다(Fenn and Raskino, 2008).

은 불필요해지는", 과학하기의 틀이 지금까지와는 근본적으로 달라지는 그런 모양의 것일지는 분명치 않다.

이렇게 서로 엇갈리는 극단적인 주장들이 만들어 내는 현재의 "소란"(이재현, 2013) 속에서 잃지 말아야 할 것은 빅데이터라는 현상, 그리고 그 출현의 배경에 있는 과학기술 환경의 변화를 비롯한 여러 가지 상황이 제기하는 문제들을 '사회과학을 어떻게 할 것인가?'라는 기본적인 문제의 시각에서 따져보는 차분한 자세이다. 빅데이터의 가능성은 열린 마음으로 수용하되 거품은 빼자는 말이고, 교과서를 고쳐 써야 할지 아니면 새로 써야 할지를 판단할 기회로 삼자는 말이다.

빅데이터와 사회과학하기의 관계에 논의의 초점을 맞추는 이 글에서는 빅데이터와 관련된 다른 측면들, 즉 기술공학적 측면이나 산업 · 경영적 측면, 그리고 사회문화적 측면들에 대한 내용은 깊이 다루지 않는다. 크게 보면 직간접적으로 서로 관련되어 있어서 여러 가지로 영향을 미치기는 하지만, 그 측면들에 대한 논의는 현재 시중에 나와 있는 많은 소개문헌에 미루고자 한다(매일경제 기획팀 · 서울대 빅데이터 센터, 2014; 시로카 마코토, 2013; 정우진, 2013; 함유근 · 채승병, 2012; Smolan and Erwitt, 2012; LaValle, Lesser, Shockley, Hopkins, and Kruschwitz, 2011).

아래에서는 이론-방법-자료를 꼭짓점으로 하는 삼각형의 구도를 사회과학하기의 기본 틀로 잡고 논의를 시작하여, 빅데이터가 가져온 자료기반의 변화와 그 특성을 살펴보고, 그에 따른 분석전략을 가늠해보고자 한다. 다음 절에서는 사회과학에서 자료가 차지하는 위치와 그동안의 변화 추세, 그리고 빅데이터의 출현을 가져온 배경과 그 영향에 대해 논의한다. 자료의 양과 질에 있어서의 변화가 오랜 기간 계속되어온 것이라는 점을 짚고, 그 연장선상에 빅데이터의 출현도 자리매김한다. 빅데이터 시대가 가져온 새로운 사회과학의 시제품 역할을 하면서 그 가능성과 한계를 극명하게 보여준 구글 독감추세예측 서비스에 대한 검토가 제3절을 구성한다. 취사선택의 필요성과 근거, 방향을 경험적으로 보여주는 유용한 실제 사례이다. 앞의 두 절에서 논의하고 검토한 내용을 원론적 사회과학하기의 틀 안에서 다시 정리하는 것이 제4절이다. 마지막으로 이런 비판적 검토에 근거하여 앞으로의 전망을 제시하는 것으로 글을 맺는다. 물론 세 요소 각각의 내용과 상대적 비중, 요소 간 연결의 구조와 방향 등은 재조정되어야 한다. 그러나 크게는 이 기본 틀의 개조와 확장을 통해 빅데이터의 잠재력을 수용할 수 있을 것이라 조심스러운 진단을 내린다.

Ⅱ. 내어놓기(Pledge)

1. 사회과학의 전통적 자료기반

사회구성원들이 무엇을 생각하고 어떻게 행동하는지를 관찰하고 기록할 수 있게 하는 도구로서의 사회조사는 현대 사회과학의 근간을 이루는 핵심 요소 중 하나다. 사회조사를 통해 생산된 자료는 앞에서 언급한 사회과학하기의 삼각형 구도에서 일차적 원료로서 물적 기반의 역할을 하고(한신갑, 2012), 계량적 분석방법과 짝을 이루면서 전후(戰後) 실증적, 경험적 사회과학의 급속한 성장과 발전을 이루어 냈다(Hunt, 1985). 이 틀이 현재까지 널리 쓰이고 있는 사회과학하기의 기본적인 틀이다.

제2차 세계대전이 끝난 후 발간된『미군 병사들: 병영생활에의 적응 (제2차 세계대전 중의 사회심리연구, 제1권)』(Stouffer and Suchman, 1949)은 전후 미국 사회과학의 전범(典範)으로 꼽힌다(Ryan, 2013). 현대적 자료와 방법의 조합이라는 기본 틀을 성공적으로 적용한 초기의 사례로 그 뒤를 따른 많은 연구들에게 새로운 가능성의 지평을 열어 주었기 때문이다. 설문자료의 기록, 정리와 분석에 당시로서는 최첨단의 전산기술을 사용했다는 점이 그렇고, 전시(戰時)라는 상황에서만 가능했던 것일 수는 있지만, 대규모의 자원동원과 조사대상 집단의 접근가능성을 활용해 오십만 명이 넘는 병사들을 조사했다는 점도 그렇다. 책임연구자였던 스토퍼도 이 연구가 가지는 함의에 대해 충분히 인식하고 있었고, 책의 서문에 "사회심리학이나 사회학 분야에서 단일 연구로는 지금까지 그 유례를 찾을 수 없을 만큼 크고 풍부한 규모의 자료("a mine of data")를 앞에 두고 있다"라고 쓰고 있다(Stouffer and Suchman, 1949: 29-30).

이 연구에서 사용되었던 기술들은 그 후로도 지속적으로 확산되어 간다. 그러나 자료의 규모라는 측면에서는 곧 현실적 제약, 특히 비용의 문제에 부딪히게 된다(Dillman, Smyth, and Christian, 2009). 당시 빠른 속도로 발전하고 있던 표집을 통한 조사방법이 이에 대한 해법으로 제시되고, '모집단 전체를 대상으로 한 무작위 확률표본 추출'이라는 통계학적 기본 틀을 연결고리로 하여 자료의 계량화와 자료수집 방법의 체계화를 이루어 간다. 그에 따라 최근까지도 자료의 크기 자체는 상대적으로 덜 강조되어 왔다. 20세기 후반에 걸쳐 집약적으로 진행된 이 발전과정에서 자료수집 방법의 체계화를 통해 자료의 대표성을 효율적으로 확보할 수 있다는 입장이

규범으로 받아들여졌기 때문이다.

현재 대부분의 자료는 이런 표집의 틀을 통해 만들어진다. 사회과학 연구자들에게 가장 널리 알려진 예는 미국과학재단(NSF)의 지원을 받아 미국여론조사연구소(NORC)가 1972년부터 정기적으로 수행해오고 있는 <일반사회조사>(General Social Survey)의 틀이다. 이 조사는 앞에서 언급한 이론적 기반 위에서 1,500명 정도의 표본으로 미국사회 전체를 관찰하고, 또 그 변화를 오랜 기간 성공적으로 추적해왔다(Smith, Marsden, Hout, and Kim, 2013).[6] 국내에서는 성균관대학교 서베이리서치센터가 한국연구재단의 지원으로 2003년부터 매년 실시하고 있는 전국표본조사인 <한국일반사회조사>(KGSS)를 통해 상응하는 역할을 수행하고 있고[7], 이 틀을 공유하는 여러 나라들이 포함된 <국제사회조사프로그램>(ISSP)을 통해 국제비교도 가능하다.[8] 물론 표집을 통한 자료수집의 전형으로 일반에 가장 널리 알려진 것은 각종 여론조사이다. 일억 오천만 명이 넘는 미국 유권자들을 대상으로 천 명 안팎의 표본을 가지고 오차범위 5% 내에서 결과를 예측해내는 선거여론조사의 예에서 보듯 이 방법의 효율성은 놀라운 것이다.

다른 한편으로 이렇게 자료의 규모가 작아짐으로써 생기는 제약에 대한 지적과 그 해결책의 하나로 대규모 자료('large-scale data')의 필요성에 대한 논의는 계속 있어 왔다. 실제로 1960년대를 넘어서면서 사회과학에서 사용되는 자료의 규모는 지속적으로 늘어났고, 그 추세는 지금도 지속되고 있다. 이런 큰 규모의 자료는 사회과학이 발전하면서 새로운 영역에서 새로운 형태로 더 상세하고 복잡한 내용을 담아낼 수 있는 자료를 요구하게 되면서 생겨나는 것이다.

1990년 미국 학술원에서 내놓은『사회과학과 행동과학의 최첨단』이란 책에 실린 한 보고서는 이 변화의 이유로 네 가지를 들고 있다(Miller, Davis, Clubb, Russett, David, and Morgan, 1990: 588- 589). 첫째, 통시적 변화에 대한 관심의 증가로 패널조사나 연속횡단조사와 같이 누적적으로 자료를 수집하는 연구들이 늘어났다는 것이다. 반복을 통한 시간적 확장이다. 둘째는 인간행동의 복잡성/복합성(complexity)이다. 사회과학자들이 다루는 대부분의 현상들이 여러 요소들이 복합적으로 관련되는 것들이고, 그 요소들 각각도 다수의 변수로 측정될 수밖에 없기 때문

6) www3.norc.org/GSS+Website
7) www.kossda.or.kr
8) www.issp.org/index.php

에 점차 변수의 수가 늘어나면서 자료의 크기도 커졌다는 것이다. 한편으로는, 앞의 두 번째 이유와 관련하여, 이렇게 변수의 수가 늘어나면서 생겨나는 방법론적 문제를 해결하기 위해서, 또 한편으로는 작은 규모의 표집으로는 포착되지 않는 소수(또는 하위)범주집단이나 희귀특정사건 등을 연구하기 위해서 자료의 규모를 확대할 수밖에 없다는 것이 그 세 번째 이유이다. 네 번째 이유는 앞의 것들과는 좀 다른 차원의 것인데, '관성적 첨가'라고 부를 수 있다. 즉, 현대 사회과학 연구방법의 귀납적 구조가 현존하는 변수를 현상의 설명에 더 이상 필요치 않은 것으로 밀어내기보다는 새로운 변수를 계속 추가하는 쪽으로 작동한다는 것이다. 여기에 연구 프로그램 간의, 특히 시계열적 비교를 가능하게하기 위한 측정연속성의 요구가 더해지면서 이런 관성적 첨가의 추세는 지속되고, 그 결과 자료의 규모는 계속 불어나게 된다. 이런 다양한 기제들이 때로는 따로, 때로는 함께 작동하면서 사회과학 분야뿐만 아니라 정부를 비롯한 공공부문과 기업 등 민간부문에서 계속 자료의 규모를 불려온 것으로 보인다. 하지만 이런 대규모 자료의 대표적 예로 제시된 <일반사회조사>, <전국선거연구(National Election Studies)>, 그리고 <소득변화패널연구(Panel Study of Income Dynamics)>에서 보는 것처럼 이 규모의 확대는 그 전부터 있어온 추세의 연장선상에서 사회과학 방법론의 근본적인, 전통적인 틀은 그대로 유지한 채 점진적으로 이루어졌다.

2. 빅데이터의 등장과 자료기반의 변화

'빅데이터'라는 용어가 처음 등장한 1997년을 빅데이터의 시발점으로 본다면 그 무렵 이후의 특기할 만한 사항들은 뒤에서 제시할 <그림 1>의 연대표로 정리할 수 있다. 하지만 그 전사(前史)에 해당하는 부분에서도 오늘의 현실에서 맞닥뜨리게 되는 기본적인 이슈들은 선명하게 나타난다. 비록 빅데이터라는 용어 자체는 아직 생겨나기 전이지만 정보량의 급속한 증가와 그것을 어떻게 관리하고 사용할 것인가에 대한 문제의식은 "정보폭발(information explosion)"이란 용어가 쓰이기 시작한 1940년대 초반까지 거슬러 올라갈 수 있다. 그 당시에도 이미 여러 영역에서 이 현상이 포착되고 있었던 것이다(Gleick, 2011). 특히 컴퓨터와 인터넷의 등장 이후 정보통신기술의 발달이 모든 영역으로 확산되고 그를 통해 부수적으로 파생되고 축적되는 정보의 양이 기하급수적으로 증가하면서 이 문제의 심각성은 더욱 더 커져 갔고, 문제

를 해결해야만 한다는 현실적 필요성은 더욱 절실해진 것이다. 1961년 프라이스(Price)가 보여준 과학분야 학술지와 논문의 수에 있어서의 증가추세와 1997년 레스크(Lesk)가 보여준 전자매체에 담긴 정보량의 증가추세가 기본적으로 같은 형태의 것이라는 점은 그런 면에서 시사적이다(Lyman and Varian, 2000; Pool, Inose, Takasaki, and Hurwitz, 1984).

하지만 최근의 정보량 증가추세는 단순히 증가라고 부르기에는 그 규모나 속도에 있어서 기존의 그것과 너무나 달라 자료의 "쓰나미"라고 부를 만큼 기존의 도구로는 감당하기 힘든 정도의 것이 되었다(Mayer-Schönberger and Cukier, 2013; Berry, 2011).[9] 이처럼 빅데이터라는 별도의 새로운 용어가 필요해질 정도로 자료기반의 물적 토대가 바뀌는 데 원동력을 제공한 것은 디지털 혁명이다. 저장매체의 고용량화, 저비용화가 이루어지고, 스마트 기기를 포함한 자료수집 기기가 소형화, 저렴화, 보편화되고, 네트워크의 보급확산과 고속화로 정보의 이동과 수집이 활성화되었으며, 연산능력이 향상되고, 인공지능, 기계학습 등 자료처리 기술이 발달하는 등 빅데이터의 등장 배경에는 하드웨어와 소프트웨어 부문 모두에서 일어난 기술환경의 진화가 자리 잡고 있다. 다분히 기술결정론적으로 들리지만 이런 환경의 변화가 자료의 역할과 비중을 바꿔 놓게 된 것이다.

하지만 규모는 빅데이터를 특징짓는 요소 중 하나일 뿐이다. 물론 상대적인 것이기는 하지만 앞에서 언급했던 것처럼 크기 자체로만 본다면 빅데이터 시대 이전에도 큰 규모의 자료는 존재했다. 문제는 규모의 변화와 더불어 달라진 자료의 성격이다. 즉, 예전과 같은 형태와 내용의 자료가 단순히 양적으로만 늘어난 것이 아니라, 그렇게 늘어나는 자료들이 그 원천과 생산방식, 구성과 사용방식에 있어서 지금까지와는 근본적으로 다른 새로운 형태와 내용의 것이라는 질적 변화의 측면에도 주목해야 한다.

빅데이터 시대에는 생활의 모든 측면에서 정보가 생산되고, 측정되고, 기록되고, 저장된다. 바로 이 점이 자료의 질적 측면에서의 변화이다. 이런 자료의 대부분은 일상생활에서의 부산물로 나온다. 별다른 생각없이 이메일과 메시지를 주고받고, 상품을 주문하고, 파일을 공유하는 과정에서 남기게 되는 전자흔적들('digital traces')이

9) 빅데이터라고 부를 수 있으려면 규모가 얼마나 커야 하는지에 대해 딱히 정해진 답이 있는 것은 아니다. 2012년의 논의들을 보면, 단일 자료의 규모가 수십 테라바이트(terabyte [TB] = 10^{12} bytes = 1000 gigabytes)를 넘어서서 수 페타바이트(petabyte [PB] = 10^{15} bytes = 1000 terabytes)에 이르는 경우들이었다. 참고로 1993년에 인터넷을 통해 움직인 정보의 연간총량은 100TB 정도였던데 반해 2008년 추정된 규모는 초당 160TB였다.

자동적으로, 실시간으로 자료로 구성된다. 그래서 이를 "데이터 배기가스"(data exhaust)라고 부르기도 한다. 자동차가 움직일 때 배출되는 배기가스처럼 인간의 활동이 남기는 흔적이라는 뜻이다. 인위적 개입에 의해 별도로 만들어지는 것이 아니라, '자연스럽게' 만들어진, 있는 그대로를 반영하는 자료라는 점에서 기존의 사회과학 자료와 차별화된다. 그 결과 자료의 형태도 지금까지처럼 정형화된 것뿐만 아니라 이런 과정에서 생산되는 자연언어 텍스트, 사진이나 음악, 동영상, 위치 정보 등 다양한 형태의 비정형적인 자료까지 포함하게 된다. 또 스마트폰을 통해 생산되는 다양한 자료들의 경우처럼 이 자료들을 서로 연결시키는 것이 가능해진다. 지금까지는 수집할 수 없었거나, 수집 대상으로 삼지 않았거나, 수집은 하더라도 분석할 수단이 없어 버려지던 자료들이 모두 포함된다는 점에서 성격이 다른 자료를 만들어 낸다.[10] <그림 1>에 표시한 사건들은 이런 점에서 특기할 만한 것들이다.

이처럼 정보의 양과 질이 달라지자 이를 어떻게 다루어야 할 것인가 하는 문제에 대한 시각도 달라지기 시작했다. 이 변화를 어쩔 수 없이 떠안아야 하는 부담으로만 보는 것이 아니라 오히려 그것을 이용할 수 있고, 또 이용해야 하는 자원으로 보는 적극적 시각이 나타나고 기술적인 해결책을 찾기 시작한 것이다. 이렇게 빅데이터가 가지는 잠재력과 가능성에 주목하는 사람들은 자료의 성격 자체가, 또는 자료의 정의 자체가 변했다고 본다. 새로운 복합체로서의 빅데이터에서 그저 양적인 변화뿐만 아니라, 그와 더불어, 또는 그로 인한, 질적 변화까지 생긴다는 것이다. 그리고 이렇게 만들어진 자료는 미시 수준에서나 거시 수준에서나 종래에는 가능하지 않았던, 즉 전통적인 사회과학의 틀에서는 할 수 없었던 새로운 형태와 내용의 분석을 가능하게 할 것이라는 점을, 또 기술환경의 변화추세가 앞으로도 더 큰 규모로, 더 빠른 속도로 지속될 것이라는 점을 감안할 때 이렇게 생산되는 자료의 비중은 계속 커질 것이라는 점을 강조한다(Lazer et al., 2009). 그리고 이렇게 자료가 양적으로, 질적으로 변화하면, 그것에 맞는 새로운 분석 방법이 나와야 한다고 주장한다.

3. 빅데이터 시대의 분석전략

이런 자료가 만들어지고 있다는 것 자체가 새롭고, 중요한 현상인 것은 분명하고,

10) 빅데이터를 정의하면서 흔히 사용되는 3Vs(Volume, Variety, Velocity)는 이런 양적, 질적 변화의 내용을 축약하고 있다(함유근·채승병, 2012; Laney, 2001).

그에 따라 '이렇게 기존의 자료와는 양적, 질적으로 달라진 자료를 어떻게 포착하고, 분석하고, 해석할 수 있을까?'하는 질문이 생겨나는 것도 자연스러운 일이다. 즉 빅데이터가 기존의 자료에 기반한 관찰대상의 구성, 연구문제의 설정, 분석도구의 사용의 틀에서 크게 벗어나는 만큼 그것들도 새롭게 만들어져야 한다는 요구가 생기게 되고, 이는 다시 어떻게 사회과학을 할 것인가 하는 근본적인 문제로 이어진다. 빅데이터란 개념을 보다 넓게 이해하고 사용하자는 입장은 바로 이런 맥락에 근거한다. 즉, 빅데이터는 단순히 새로운 자료만을 의미하기 보다는 그것의 출현이 야기하는 '무엇을(what), 어떻게(how), 왜(why)'라는 복합적인 질문들에 대한 답을 제시하는 넓은 의미에서의 체계적 방법론까지를 포함하는 새로운 틀, 새로운 패러다임이라는 것이다(이상구, 2014; 이재현, 2013; Macy and Golder, 2014; Kitcin, 2014; Lazer et al., 2009).

이런 시각에서 볼 때 빅데이터의 시대로 들어서면서 무엇이 근본적으로 바뀌는가는 크게 다음의 세 가지 측면으로 정리된다: 자료규모("More"), 자료구성("Messy"), 분석준거("Good Enough")(Mayer-Schönberger and Cukier, 2013). 이 세 가지는 서로를 지원하고 증폭시키는 고리로 연결되어 있기도 하다. 첫 번째는 자료의 양적인 측면이다. 지금까지 상상하지도 못했던 규모와 정확도로 수십억 인구의 일상생활을 관찰할 수 있게 되고, 특히 이를 통해 기존의 표집에서는 접근할 수 없었던 소수하위집단들에 대해서도 세분화된 관찰이 가능해진다는 점이 강조된다.

두 번째는 자료의 구조와 구성에 대한 태도의 변화이다. 빅데이터는 예전의 자료처럼 일관되게, 체계적으로 축약되고 정리된 자료가 아니다. 이는 빅데이터의 엄청난 규모 자체가, 또 자료의 생성과정이 가져오는 불가피한 측면이기도 하고, 이 자료가 실제 세상의 복잡다기함을 직접적으로, 즉각적으로 반영하기 때문에 생기는 측면이기도 하다. 따라서 예전의 작은 규모에서만 가능했던 엄밀한, 통제된 정확성을 더는 요구할 수도 충족시킬 수도 없게 된다. 하지만 기준을 바꾸어보면, 즉 이런 단점에 집착하지 않고 빅데이터의 장점에 주목한다면, 잃는 것보다 얻는 것이 더 많다는 것이다.

이 두 가지 변화와 연결되어 일어나는 세 번째 변화가 그동안 사회과학 연구의 틀을 정립하는 데 있어서 이상적 준거가 되었던 인과관계로부터 상관관계로의 방향전환이다. 한편으로는 더 커지고, 더 다양해지고, 더 직접적이 된 자료와 그를 다룰 수 있는 향상된 연산능력에 힘입어 상관관계 자체가 더 강력해졌다는 것이고, 또 한편

으로는 이렇게 많고, 복잡하고, 빠르게 생산축적되는 자료를 효율적으로 분석하는 현실적인 대안으로서의 상관관계가 그 입지를 더 강화했다는 것이다.

즉 빅데이터라는 새로운 자료의 출현으로 인한 자료기반의 변화가 사회과학하기의 틀을 근본적으로, 총체적으로 바꾸어 놓게 되고, 따라서 분석전략도 새로운 방식으로 재구성되어야 한다는 체제변혁적인 주장이다. 이런 입장의 선봉에 선 것이 앤더슨(Anderson, 2008)이다. 앤더슨은 이제 이론의 시대는 끝났다고 주장한다. 지난 시대를 조건지웠던 물질적, 기술적 제약들이 빅데이터, 슈퍼컴퓨터, 기계학습과 인공지능에 의해 사라진 "페타바이트의 시대(Petabyte Age)"에는 이론에 근거한 가설과 모형을 중심으로 짜여졌던 전통적 과학하기의 틀(Box, 1987)이 불필요해진다는 것이다. 이제 자료 자체가 과학하기의 중심에 서게 되고 다른 요인들은 그것에 종속되는 구도로 바뀌었다는 것이다.

이런 방법론적 대전환에 대한 논의에서 핵심적인 축은 인과관계에서 상관관계로의 (역)전환이다. 지난 수 세기에 걸쳐 지속되어온 '과학적 방법'의 작동구조는 관찰된 자료로부터 두 변수 사이의 관계가 원인과 결과의 관계라는 결론을 엄밀한 사고를 통해 도출해내는 과정으로서의 인과관계 추론(causal inference)을 중심에 두고 있다(Hume, 1978). 1850년대 런던의 콜레라 창궐이 오염된 우물을 식수원으로 사용했던 데에서 비롯되었다는 것을 밝힌 스노우(Snow, 1855)의 연구가 이런 방법의 가장 고전적인 전형으로 제시된다. 이 과정에서 상관관계는 인과관계를 규정하는 데 있어서의 세 가지 기본 필요조건 중 하나일 뿐이다. 가장 중요한 것은 어떻게 원인이 결과를 만들어내는지를 설명하고, 그 기제를 제시하는 이론적 논의가 있어야 한다는 것이다(Blossfeld, Golsch, and Rohwer, 2007; Marini and Singer, 1988). 따라서 이 기존의 틀에서는 인과관계가 자료에서 경험적으로 찾는 상관관계보다 한 단계 위에 놓인다.

이런 사고의 틀에서 벗어날 수 있게 되었다는, 따라서 벗어나야 한다는 급진적 빅데이터 주창자들의 입장을 가장 단적으로 표현한 것이 아래에 인용된 앤더슨(Anderson: 2008)의 주장이다.

> 사람들이 왜 어떤 행동을 하는지를 누가 다 알 수 있단 말인가? 중요한 점은 사람들이 그런 행동을 한다는 것이며, 이제 그 행동들을 지금까지 상상도 하지 못했던 정도로 정교하게 추적하고 측정할 수 있게 되었다는 것이다. 데이터가 충

분히 주어진다면, 숫자가 스스로 자기 얘기를 할 수 있게 된다.
상관관계가 인과관계를 대체하게 되고, 일관성을 갖춘 모형이나 통합된 이론 등의 체계적 설명이 없이도 과학은 발전해나갈 수 있다.

양적으로, 또 질적으로 근본적으로 달라진 자료가 이런 극단적 형태의 귀납적 경험주의를 가능하게 하며 이제 상관관계만으로도 충분한 시대가 왔다는 것이다(Halevy, Norvig and Pereira, 2009).

4. 빅데이터의 현재와 미래: 중간점검

이제 자료기반의 양적, 질적 변화 자체는 누구도 부인할 수 없는 현실로 우리 앞에 서 있다. 그렇지만 상관관계로의 전환이라는 주제를 둘러싸고는, 특히 새로운 패러다임 안에서 이론이 차지하는 위치와 관련하여 현재 사회과학 내에서는 여러 가지 질문이 나오고 있다(Taylor, 2013): 빅데이터 패러다임에서 이론은 어디에, 얼마나, 어떻게 필요한 것인가? 만약 필요하다면, 새로운 이론이 필요한가, 아니면 기존의 이론으로 충분한가? 상관관계를 찾는 기술적(記述的) 작업이 결국에는 나름의 이론을 만들어 낼 것인가? 더 나아가서는, 빅데이터 패러다임의 연구라는 것이 사회과학 연구의 과학화 과정에 있어서 자연스러운 다음 단계인가, 아니면 지금까지와는 근본적으로 다른, 하지만 불가피한, 새로운 것인가? 이런 질문들에 대한 답을 찾는 데 있어서 빅데이터를 그저 양적, 질적으로 다른, 새로운 자료라는 좁은 의미에서의 기술적 현상으로만 보지 않고, 자료와 그 자료를 둘러싼 생산과 소비의 양식까지를 포함하는 복합적 구성체로 보면서 이 현상을 이론-방법-자료의 상호작용이라는 사회과학하기의 틀 안에서 보아야 한다는 데는 이론(異論)의 여지가 없다.

유전자 정보를 다루는 생물학, 천체자료를 모으는 천문학, 자연언어 사용을 연구하는 언어학 등의 분야에서 부분적으로 적용사례를 찾을 수는 있지만, 아직 사회과학의 틀을 흔들어 놓을 만큼 밀접하게 연결된 성공사례를 찾기는 쉽지 않다. 그런 면에서 현재의 논의에서 위의 복합체의 중요한 일부로 녹아들어 있는 것 중 하나는 빅데이터에 대한 신화(神話)다(boyd and Crawford, 2012). 빅데이터가 단지 빅데이터이기 때문에 그 자체로서 진실성, 객관성, 정확성을 가질 것이라거나, 그 규모만큼 거기서 얻을 수 있는 아이디어도 클 것이라거나, 자료의 새로움이 지금까지 볼 수

없었던 보다 강력하고, 보다 차원 높은 새로운 지식으로 이어질 것이라는, 아직 충분히 확인되지 않은 기대가 깔려 있다.

다음 절에서는, 한 편으로는 이런 기대를 만들어내는 데 큰 역할을 하였지만, 다른 한편으로는 현재와 미래 사이의 녹녹치 않은 거리를 보여주기도 한 구체적 사례 하나를 통해 지금까지 이루어 놓은 것과 앞으로 이루어야 할 것들을 검토해 보고자 한다.

Ⅲ. 바꾸어 놓기(Turn): 구글 독감추세예측(GFT)

1. 성공신화 만들기

<그림 1>의 길지 않은 연대표가 보여주는 것이 바로 이 과정에서 중요한 의미를 갖는 사건들이다. 빅데이터의 기본적 속성이 일반적으로 정리되고(Laney, 2001), 하둡(Hadoop)을 포함한 소프트웨어의 개발과 보급을 통해 기술적 해법이 본격적으로 제시되면서 시각의 변환은 급격하게 이루어 졌고, 특히 비즈니스 분야에서는 전폭적으로 빅데이터의 개념과 기술을 수용하게 된다. 2008년에 들어서서는 네이쳐(*Nature*)지가 빅데이터 특집호를 냈으며, 와이어드(*Wired*)지에 앤더슨(Anderson)의 도발적인 논설이 실린다. 2010년에는 이코노미스트(*The Economist*)지가 특집호를 내면서 빅데이터에 대한 시각은 모양을 갖춰 간다.[11]

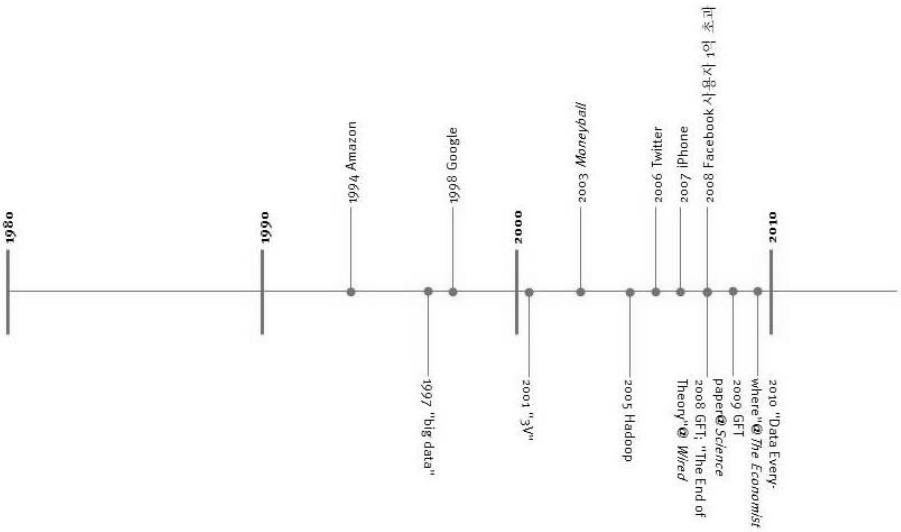
그런 면에서 2008년은 분기점을 이룬다. 그 해 "구글 독감추세예측(Google Flu Trends [GFT])"이라는 서비스가 공개되었다.[12] 하루 5억 건이 넘는 구글 검색기록이라는 엄청난 규모의 자료에서 사용자들이 40여개의 독감의 징후들과 관련된 검색어를 얼마나 사용하는가를 살펴봄으로써 독감의 추세를 예측해보고자 하는 도구였다. GFT는 검색엔진을 통해 자동적으로 집계되는 자료의 위력을 질병 확산의 추세라는 구체적, 실질적 주제를 살피는 데 사용한 눈에 띄는 프로젝트였다. GFT가 예측한 값은 당시까지의 자료와 95 퍼센트가 넘는 일치도를 보였고, 특히 공식자료보다 일주일에서 열흘 정도나 앞서 독감 유행상황을 예측하는 것이 가능하다고 알려졌으

11) Special Report: Managing Information (February 25, 2010). http://www.economist.com/printedition/specialreports?page=1&year[value][year]=2010&category=76984.

12) 현재 웹사이트(http://www.google.org/flutrends/intl/en_us/)에서 볼 수 있는 것은 그 후 수차례에 걸쳐 개정된 것이다.

며, 개발과정에 참여했던 미국 질병예방통제센터(CDC)도 그 효용을 인정하고 만족해했었다. 뉴욕타임즈와 CNN 등 여론에 영향력이 큰 대중매체들도 이를 중요한 뉴스로 다뤘다.

〈그림 1〉 빅데이터에 관련된 중요 사건 연대표



GFT가 구글 사용자들이 만들어낸 자료에 새로운 시각으로 접근해 지금까지는 볼 수 없었던 새로운 것을 만든다는 빅데이터 활용의 이상적 전형을 보여준 것은 분명하다. 그리고 당시가 '빅데이터'란 말이 지금처럼 널리 유행하기 전이었다는 것을 고려하면 이것이 GFT 자체의 성패를 떠나 빅데이터가 가질 수 있는 가능성에 대해 시사하는 바는 컸다. 간명하게 정의하기 어려운 빅데이터라는 말에 대표적인 성공사례로서 얼굴을 붙여 준 셈이다.[13] GFT의 방법론에 대한 논문이 그 다음 해에 네이쳐(*Nature*)지에 발표되면서(Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant, 2009), 이 전환은 마무리되는 것으로 보였고, 그 전환의 영향권은 과학 전반을 포함하게 된다.

표면에 드러나지는 않지만 실제 분석의 구조와 과정에서 보이는 두 가지 특징에

13) 최근에 나온 빅데이터에 대한 대표적인 대중용 소개서인『빅데이터: 우리가 어떻게 살고, 일하고, 생각하는지를 근본적으로 바꾸어 놓을 혁명』(Mayer-Schönberger and Cukier, 2013)은 바로 이 GFT를 가장 성공적인 사례로 들면서 책을 시작한다.

주목할 필요가 있다. 이 둘은 GFT 뿐만 아니라 빅데이터의 성공적인 적용사례로 꼽히는 넷플릭스(Netflix)(Auletta, 2014; Madrigal, 2014)나 『머니볼』(Lewis, 2004) 등의 다른 사례들에서도 찾아볼 수 있는 것들로 자료보다는 분석의 측면에서 빅데이터를 고려할 때 더 중요해 지는 부분이다.

그 하나는 앞에서 다룬 인과관계와 상관관계 사이의 긴장과 연관되는 부분이기도 하다. 과거를 설명하는 것과 미래를 예측하는 것은 서로 대칭적인 것이 아니라 근본적으로 서로 다르다는 것이 사회과학자들에게 일반적으로 받아들여지는 명제이다. 그러나 빅데이터의 틀에서는 이 문제를 다른 시각에서 접근한다. 빅데이터의 틀에서 분석의 주된 목적은 설명이 아니라 예측인 경우가 많다. 그리고 그 예측을 위해 과거에 대한 설명력을 높이는 데 집중한다. 과거의 설명과 미래의 예측을 논리적으로 동일시하기 때문이다. 이 과정에서 과거는 예측을 위한 자료로서의, 준거로서의 역할이 주어지고, 이 기준에 가장 가까운 모형을 찾는다. 이렇게 얻어진 모형은 과적합화(overfit)될 위험이 높고, 특히 사회현상처럼 자기 교정과 확대재생산의 경향이 큰 경우 그로 인한 실패의 가능성도 더욱 커진다.

또 하나는 독감유행추세를 예측하기 위해 동원된 분석력이 단지 슈퍼컴퓨터만이 아니었다는 점이다. 상관관계만으로 충분하다는, 자료가 스스로 말한다는 주장을 그대로 받아들였을 때 가지게 되는 분석과정의 그림은 기계가 자료를 생산하고, 기계가 그것을 분석해 결론을 내는 기계중심적 투입-산출의 틀이다. 하지만 이는 분석과정의 실재와는 많이 다른, 지나치게 단순화된 것임을 자신들의 방법론을 소개한 2009년의 논문에서 쉽게 볼 수 있다. GFT의 예측은 4억 5천만 개에 달하는 반복된 분석과 정교한 모형을 통해 만들어진 것이다. 그리고 이 과정에서 쓰이는 통계학적 도구들은 다양한 수준에서 이론에 연결되어 있다(Duncan, 1984). 마치 유전자에 관한 이론이 있었기에 유전자배열분석이 가능했던 것처럼, 아무 이론적 틀 없이 기계의, 기계에 의한, 기계를 위한 분석은 이루어질 수 없다. GFT의 성공도 그에 의한 것이었다.

2. 성공신화 깨기

하지만 이 결과, 특히 예측의 정확성에 대한 의문들이 2009년의 자료를 검토한 쿡 등의 연구로부터 제기되기 시작했고(Cook, Conrad, Fowlkes, and Mohebbi, 2011),

2013년 네이쳐(*Nature*)지에 실린 기사에서도 이 문제가 다루어진다(Butler, 2013). 이런 문제점들을 보다 포괄적, 체계적으로 정리해 지적하고, GFT에 대한, 그리고 넓게는 빅데이터 패러다임의 연구들에 대한 재성찰을 요구한 것이 최근에 싸이언스(*Science*)지에 발표된 레이저 등(Lazer, Kennedy, King, and Alessandro, 2014)의 논문이다. "GFT의 우화(寓話): 빅데이터 분석의 함정"이라는 제목을 단 이 논문에서 제시한 분석결과는 GFT의 등장 때만큼이나 큰 충격을 가져 온다. 그들이 검토한 2011년에서 2013년 사이의 108주 동안 100주에 걸쳐 GFT의 예측치가 실제값을 넘어 섰고, 심할 때는 두 배나 되는 경우도 있었다는 것이다.[14)]

물론 대중매체들은 바로 등을 돌렸다. 학계에서도 "실리콘밸리식의 아마츄어 사회과학이 만들어 낸 예상했던 실패"라는 터프티(Tufte)의 입장처럼 비판적인 견해를 찾아볼 수 있다(2014). 어떻게 보면 새로운 것의 출현과 그것에 반짝 집중하는 여론의 관심, 그러나 얼마 지나지 않아 노출되는 문제점과 그에 따른 급속한 몰락이라는 전형적인 줄거리를 따른 것 같아 보인다. 하지만 이렇게 표면상의 굴곡만을 따라가다 보면 놓치는 점들이 있고, 바로 그것들이 빅데이터의 현재와 미래를 넓고, 깊게 이해하는 데 더 중요한 의미를 갖는다.

우선 이 논문에 깔려 있는 기본적인 태도는 그들이 2009년 논문에서 보여준 것과 같다. 즉, 레이저 등의 2014년 논문이 GFT를 비판하고 있는 것은 맞지만, 그 비판의 초점은 빅데이터 분석이라는 GFT의 근본적인 틀에 대한 것은 아니다(Lazer et al., 2014). 기본적으로 저자들은 빅데이터의 과학적 잠재력과 가능성을 부정하지 않는다(Lazer et al., 2009). 하지만 데이터의 양이 많다고 해서 측정과 개념의 타당성과 일관성, 자료 간의 비독립성 등의 기초적인 문제를 간과할 수는 없다는 것이다. 양이 질을 보완할 수는 있지만, 보장하거나 대체할 수는 없기 때문이다. 특히 자료의 타당성 문제는 모든 자료 분석에 공통되는 가장 근본적인 것이고, 자료의 규모가 커진다고 해서 해소되는 것이 아니다. 게다가 빅데이터는 이런 기초적인 문제들에 대한 엄밀한 검토가 이루어진 도구와 절차에 의해 생산된 것이 아니라는 점도 다시 환기해야 한다. 빅데이터 분석을 보통 '짚더미 속에 떨어진 바늘 찾기'에 비유하곤 하는데, 그만큼 더 어렵고, 그만큼 실수의 가능성이 더 높다는 경고의 뜻으로 받아들일 필요가 있다.

---

14) 이런 차이는 긴스버그 등(Ginsberg et al., 2009)의 논문에 실린 두 번째, 세 번째 그림과 버틀러(Butler, 2013)의 기사에 실린 첫 번째 그림을 대조해보면 더할 수 없이 확연해진다.

대중매체에서 본 표면상의 단순한 줄거리와 저자들이 갖고 있는 이런 기본적인 시각의 기저(基底) 사이에는 반전(反轉)이 있다. 그 반전은 GFT의 결과가 실제의 값과 어긋난다는 것을 보여주는 데 그치지 않고 어떤 대안이 있는지도 검토하면서 찾게 된다. 그것은 GFT의 빅데이터와 CDC에서 전통적인 방식으로 수집하는 자료를 합쳐서 쓰면 각각의 결과보다 더 나은 결과를 얻을 수 있다는, 즉 기존의 자료를 대체하기보다는 상호보완하는 것일 때 빅데이터의 효용이 나타난다는 결론이다. GFT의 초기 개발팀의 일원이었던 모헤비(Mohebbi)의 증언에 의하면 당시 자신들의 목적도 바로 그것이었다고 한다(Madrigal, 2014). 단지 CDC의 자료와는 다른 자료로 그리고 다른 방법으로 지표를 만들어 CDC의 지표계산 방식에는 포함되지 않는 다양한 다른 현상들에 무게를 둠으로써 (독립변수 간의 다중공선성을 피하는 것과 같은 논리로) 최대한 상호보완성을 높이려 했다는 것이다. 새로운 자료가 가지는 보완효과를 통해 전통적 방법으로는 다룰 수 없었던 영역으로, 또 전통적 방법으로는 이룰 수 없었던 수준으로 분석을 확장하고 제고할 수 있다는 것이 레이저 등의 결론적인 제안이다.

Ⅳ. 되돌려 놓기(Prestige)

해수의 수온이 바뀌면 어종(魚種)이 바뀌고, 그에 따라 어구(漁具)도 바뀌어야 한다. 낚시로 잡는 고기, 그물로 잡는 고기, 작살로 잡는 고기가 따로 있기 때문이다. II와 III에서 살펴본 것처럼 빅데이터의 출현으로 인한 자료기반의 변화가 사회과학 생태계에 큰 변화를 가져왔다는 것은 돌이킬 수 없는 사실이고, 그렇다면 그에 맞춰 분석전략도 바뀌어야만 할 것이다. 그런 의미에서 지금 사회과학은 전환기를 맞고 있다. 이 전환의 시기는 기존의 틀이 흔들리고 있다는 점에서 위기이기도 하고, 새로운 가능성을 찾아볼 수 있다는 점에서 기회이기도 하다(Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, and Byers, 2011; Bollier, 2010). 이 위기와 기회의 문제를 고민하고 있는 여러 사람들의 공통된 결론은 어떻게 사회과학을 할 것인가 하는 넓은 의미의 방법론적 틀을 생각하면서 새 길을 찾아야 한다는 것이다. 이 틀의 한 축은 사회과학하기가 사회현상에 대해 질문을 던지고 그에 대한 답을 찾는 연속적 과정이고, 그 과정에서 끊임없이 이론과 자료, 자료와 이론 사이를 오가는 것이라고

보는 것이다. 자료를 독립적인 별개의 것으로 보기보다는, 사회과학하기라는 구조의 일부로, 또 그 안에서 이론의 상대역을 하는 짝으로 보는 것이다(Gitelman, 2013; Merton, 1968). 아래에서는 이런 틀에서 빅데이터에 접근할 때 고려해야 할 몇 가지 문제들을 살펴보고자 한다.

1. 성공사례

빅데이터를 반기는 이유 중 하나는 많은 자료를 손쉽게 구할 수 있게 되면 더 나은 사회과학이 나올 수 있을 것이라는 막연한 기대이다. 수십억 인구의 일상생활을 시시각각으로 관찰하고 기록할 수 있게 되면서 지금까지 상상도 하지 못했던 형태와 내용의 자료가 주어지는 엄청난 기회라는 것이다. 하지만 이 자료를 사회과학에서 어떻게 사용할 수 있을지는 아직 탐색 중인 것으로 보인다. 사회과학이 관심을 두는 연구문제들과 이 새로운 자료를 효과적으로 결합시키기 위해서는 그 문제들을 새로운 방식으로 다시 써야하기 때문이다.

자연과학 분야에서 빅데이터 활용의 대표적인 성공사례로 꼽히는 것들은 천문학에서의 천체측량이나 생물학의 유전자배열처럼 연구대상 자체가 큰 규모의 자료를 요구하는 경우들이다. 다른 분야에서 빅데이터라는 새로운 자료기반에 근거해 분석전략의 핵심이 바뀐 경우로는 언어학의 자연언어처리 분야를 들 수 있다. 특히 문법구조에 기반했던 종래의 방식에서 언어사용 자료에 중심을 두는 방식으로 번역 문제를 다루는 새 틀은 자료의 규모와 연산능력의 가속화를 효과적으로 사용한 경우이다.

하지만 위의 경우들처럼 빅데이터의 1차적인 속성만을 토대로 새로운 분석의 틀을 짠 경우를 사회과학 분야에서는 아직 보기 어렵다. 자료의 규모는 커졌지만 분석의 틀 자체는 바뀌지 않은 경우가 대부분이다. 기존의 연구 틀에 빅데이터를 수용해 기존의 연구를 확장·심화시키는 것이 최근 보이기 시작하는 추세인데, 특히 이런 추세는 그동안 자료의 집적과 전산처리가 상대적으로 어려웠던 텍스트를 대상으로 한 연구들에서 두드러진다(예, Craig, 2013; Michel, Shen, Aiden, Veres, Gray, The Google Books Team, Pickett, Hoiberg, Clancy, Norvig, Orwant, Pinker, Nowak, and Aiden, 2011). 또 하나 흥미로운 예는 전통적인 조사자료의 문제점을 빅데이터를 이용해 진단한 경우이다(Ansolabehere and Hersh, 2012).

사회과학 분야에서 새로운 연구문제를 찾는다는 시각을 통해 보았을 때 빅데이터

의 속성 중 특히 주목할 만하고 효과가 클 것으로 보이는 것으로 두 가지를 꼽겠다. 하나는 자료가 커지면서 생겨나는 '깊이'를 활용하는 것이다. 그동안 표집을 통해 얻은 자료의 한계로 지적되어 온 소수집단과 희귀사건의 연구에 특히 활용의 가능성이 높아 보이는데, 산업과 경영 부문에서는 이미 상당한 진전을 보이고 있는 부문이다 (Auletta, 2014; Madrigal, 2014). 또 하나는 달라진 자료의 생성과정이 자료간의 연계를 현실적으로 가능하게 한다는 점이다. 그동안 논의만 되던 다양한 측면들 간의 관련성과 복합성을 실질적으로 포착해낼 수 있는 물적 토대가 마련되었다는 것이다 (boyd and Crawford, 2012). 빅데이터로 인해 가능하게 된 이 두 속성을 활용하려면 모집단의 대푯값을 중심으로 하고 자료의 복합적 다차원성을 축약해내도록 만들어진 현재의 분석틀은 다시 구상해야 할 필요가 생긴다.

2. 자료의 속성과 분석의 방법

앞에서도 언급한 것처럼 자료의 양이 지나치게 빨리 늘어나면서 자료의 홍수, 자료의 폭주 상태를 만들어내고 있다는 입장이 있다. 최근에 자주 쓰이는 "정보비만(infobesity)"이나 "정보과부하(information overload)"등의 용어가 이 입장을 대변하는데, 이들의 논의도 정보와 자료의 양 자체나 그 증가에 대한 것이라기보다는, 그로 인해 생겨나는 문제점들에 대한 지적이다. 이런 문제들의 대부분은 빅데이터 논의에 때로는 명시적으로, 때로는 암시적으로 깔려있는 '자료의 양이 많아지면 그에 따라서 질도 좋아진다'는 양에서 질로의 전이(轉移)가정에서 비롯된다. 객관성이나 정확성 등 사회과학 자료로서 갖춰야 할 속성들이 다다익선(多多益善)의 논리에 의해 확보될 것이라는 가정이다. 이 가정이 간과하는 것은 양과 질 사이의 충돌지점들이다 (Deming, 1960). 이런 맥락에서 아직은 잠재적 가능성 수준에 머무르고 있는 빅데이터를 효과적으로 수용하기 위해 비판적으로 검토해야 할 사항들 몇 가지를 간략하게 짚어보려고 한다.

우선, 모든 자료는, 크기와 상관없이, 신호(signal)와 잡음(noise)의 두 부분으로 구성되어있다. 즉 자료에서 실제로 얻을 수 있는 유용한 정보의 양은 자료의 총량만으로 결정되는 것이 아니라 신호 대 잡음의 비율(signal-to-noise ratio, 종종 약칭 SNR 혹은 S/N)에 의해서도 영향을 받는다. 그래서 이 둘을 분리해내고 잡음 부분을 제거하거나 처리하는 것이 과학하기의 핵심적 단계 중 하나이다. 빅데이터 시대에 들어

서도 이 단계를 생략할 수는 없다. 예전과 같이 자료수집 단계에서의 기획과 조직, 통제가 이루어지지 않은 자료이기 때문에 더욱 그렇다. 인터넷이 자동으로 만들어내는 자료들도 여러 가지 잡음, 또는 오차의 문제를 갖고 있고, 자료의 규모가 커지면 이 문제는 오히려 더욱 증폭될 수도 있기 때문이다. 그래서 역설적으로 자료가 많아질수록 쓸 만한 정보를 찾는 일이 더욱 어려워지고 있다는 말도 나온다(함유근·채승병, 2012: 7).

둘째, 클수록 좋다는 빅데이터의 논리에 내포된 또 하나의 오류는, 자료의 크기가 커지면 그만큼 포괄적이 되고, 따라서 자료의 대표성도 확보된다는 가정이다. 자료의 근거가 수백이나 수천에 그쳤던 지금까지의 제한된 사회과학 표집자료와는 달리 빅데이터의 근거는 수백만에 이른다는 점이 자주 강조되고 있고, "*N* = All"이란 표현도 자주 쓰인다. 표집의 한계, 특히 표집오차로부터 자유로워진다는 것이다. 하지만 빅데이터와 전수조사는 개념적으로 별개의 것이고, 연구대상의 완전한, 또는 체계적인 대표성이 보장되지는 않는다(한신갑, 2012). 가령, 트위터(Twitter) 자료를 사용한 연구에서 볼 수 있는 것은, 그 숫자가 얼마나 큰가에 상관없이, '트위터 사용자'라는 특정한 범주에 관한 것일 뿐, 전체 인구를 대상으로 한 것이 아니고, 따라서 그 자료가 누구를 대표하는지의 문제에 주의해야만 한다.

셋째, 이와 관련해 주의해야 할 또 하나의 문제는 관찰 또는 분석의 단위(unit of observation, unit of analysis)가 무엇인가 하는 것이다. 다시 트위터의 예를 들면, 실제 트위터 사용자와 트위터 계정은 반드시 일인일계정으로 짝 지어지지는 않는다. 자료의 구조가 관찰하고자 하는 실제 사회와는 다른 방식으로 짜여 있기 때문이다. 이 상황에서 어느 쪽을 기준으로 *N*을 집계할지는 쉽게 결정할 수 없다. 스마트폰에서 생성되는 자료나 구글 독감추세예측에 사용된 검색의 경우도 마찬가지이다.

넷째, 빅데이터 시대의 분석방법으로 '재발견'된 상관관계가 가질 수 있는 문제점들도 고려해야 한다. 자료의 규모가 커질수록 통계적으로 유의한 상관관계를 찾을 가능성은 기하급수적으로 커진다. 하지만 이렇게 찾아지는 상관관계들의 대부분은 실제 상황을 이해하는 데 도움이 되지 않는 의사(擬似) 상관관계들이다. 짚더미가 커지면, 바늘처럼 보이는 지푸라기들도 많아지고, 결과적으로 진짜 바늘 찾기는 그만큼 더 어려워진다. 즉 자료의 양 증가가 오히려 분석을 저해할 수 있다는 것이다. 이 점을 보여주기 위해 자주 드는 예 중 하나는 빅데이터 분석을 통해 주가지수(S&P 500)와 방글라데시의 버터 생산량 사이에서 높은 상관관계를 찾은 경우이다

(Leinweber, 2007). 빅데이터 분석이 오히려 더 쉽게 아포페니아(apophenia, 실제로는 아무 관련이 없는 것들 현상들 사이에 연관성을 부여하는 일)에 빠질 수 있다는 점을 경고하는 예이다(boyd and Crawford, 2012).[15] 이중 특히 문제가 되는 것이 없는 것을 있다고 하는, 즉 <표 1>의 제4사분면에 해당하는 경우이다.[16] 이렇게 늘어나는 통계적으로는 유의미한 상관관계 중 어느 것이 실질적으로 의미있는 것인지를 자료 자체만을 기준으로 가려낼 수는 없다.[17] 즉 빅데이터에서는 상관관계를 너무나 쉽게, 너무나 많이 찾을 수 있기 때문에 오히려 분석과 해석은 더 어려워지는 것이고, 따라서 자료의 속성, 방법의 선택, 분석틀의 기본 가정 등에 대해 더 많은 질문을 할 필요가 생기는 것이다.

〈표 1〉 상관관계의 적부(適否)

		y (반응)	
		없음 (-)	있음 (+)
x (자극)	있음 (+)	(2) False Negative	(1) Correct Positive
	없음 (-)	(3) Correct Negative	(4) False Positive

자료의 내용과 관련하여, 다섯째로, 온라인 환경에서 만들어지는 디지털 기록들이 무엇을 보여주는 것인지, 만약 그것이 사람들의 의견, 태도, 행위 등을 보여주는 것이라면 그것을 얼마나 어떻게 보여주는지 등의 문제도 다시 생각해봐야 한다. 사회조사 자료를 다루는 연구자들에게 이미 잘 알려져 있는 것 중 하나가 사람들이 뭘 하는지, 뭘 생각하는지, 어떻게 얘기하는지가 각각 서로 다르다는 것이다. 여기에 무엇을 어떻게 찾는지(search), 무엇을 얼마나 보는지(view), 어디에서 어디로 움직여 가는지(click) 하는 온라인 환경에서 생산되는 행위기록까지 더해지면 자료는 그만큼

15) 나보코프(Nabokov)의 단편 '상징과 기호(Symbols and Signs)'에 나오는 주인공은 나뭇가지, 구름의 움직임, 포개 놓은 유리잔들 등 그가 주위에서 보는 모든 것에 숨은 의미가 있다고 생각하고 그것을 인지하고, 해석해 내고자 하는 편집 증상("referential mania")에 시달리며 미쳐간다.

16) "False hit" 또는 "false alarm"이라고도 부른다.

17) 이 문제에 대한 기술적 해법(Maximal Information Criteria)을 개발하려는 연구들이 통계학 분야에서 진행되고 있기는 하다.

더 복잡해진다. 서치, 뷰, 클릭과 같은 행위들이 무엇을, 얼마나 어떻게 보여주는지에 대한 기초적인 연구가 더 있어야 할 것이다.

여섯째, 디지털혁명을 통해 그동안 볼 수 없었던 새로운 영역들을 포함하는 다양한 생활의 영역들이 자료화되고 있기는 하지만 빅데이터가 보여주는 내용 또는 대상 영역은 사실상 매우 제한되고 편향되어 있다. 티라노사우루스(T-Rex)라고 하는 공룡이 현재 많이 알려진 이유가 그 공룡이 중요하거나 특이해서가 아니라, 단지 그 공룡의 화석이 많이 남아있어서 쉽게 발견되고 채집되기 때문이라는 점을 환기할 필요가 있다. 마찬가지로 빅데이터로 자료가 만들어지는 영역과 그렇지 않은 영역을 구분하는 기준은 기술적 가능성, 상업적 필요성 등 자의적인 것들이다. 사회과학에서 다루는 영역 중 일부만이 전자에 속해 있고, 나머지 영역 중 어떤 것들은 앞으로도 디지털화되지 않을 것이다. 따라서 선택된 영역에서 자세한 정보들을 항시적으로 생산하기는 하지만 그런 영역들은 생활의 제한된 일부만을 보여주는 "깊기는 하지만 넓지는 않은" 역설적 상황이 생겨나는 것이다(Arbesman, 2012).

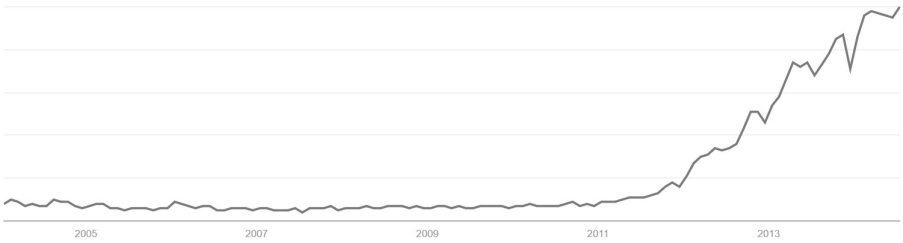
마지막으로, 그냥 있는 그대로의 것이라는 뜻에서의 "원(原)자료(raw data)"는 논리적으로도, 실제적으로도 있을 수 없다(Gitelman, 2013; Bowker, 2005). 자료는 항상 만들어진다. 그리고 그것이 만들어지는 역사적, 현실적 맥락과 과학기술적, 사회문화적 환경이 자료의 성격을 규정한다. 따라서 어떤 자료도 그런 의미에서의 특수성, 의존성을 띠지 않을 수 없다. "자동적으로" 생산된다는 빅데이터의 경우에도 그 생산과정을 "자동"이게 하는 알고리즘이 있게 마련이고, 그것을 통해 온라인상에 남겨지는 전자적 흔적들과 그것을 필요하게 하고, 가능하게 하고, 해석하게 하는 사회문화적 논리가 연결된다(Korb, 2004). 그리고 이런 배경이 위에서 논의한 어떤 영역이 분석의 대상으로 선택되는지의 기준이 되고, 제약으로 작용한다. 아마존(Amazon)이나 넷플릭스(Netflix) 등의 회사가 만들어내는 자료, 트위터나 페이스북(Facebook) 등의 소셜 미디어를 통해서 만들어지는 자료 등의 빅데이터도 그것이 생산된 맥락과의 연결 속에서, 환경과의 관계 속에서 분석하고 이해하여야 하는 이유이다.

Ⅴ. 마무리 짓기

<그림 2>는 이 글과 관련하여 두 가지를 보여준다. 그 하나는 현재 빅데이터를 사

용하는 분석의 예를 보여준다는 것이다. 인터넷에 올라와 있는 자료들을 정리해 기술(記述)통계를 내는 이런 종류의 분석의 가장 전형적인 예이다. 그보다 중요한 것은 이 그림이 보여주는 내용이다. 즉 빅데이터에 대한 현재의 높은 관심은 2011년, 2012년 무렵에서 지금까지의 불과 2-3년이라는 매우 짧은 기간 동안 빠르게 달아오른 결과라는 것이다. 몇 년 사이의 이런 급속한 증가가 변화의 토대에 대한 차분한 검토와 함께 이루어지지 않았고, 그 부분을 채워보겠다는 것이 이 글의 목적이었다.

〈그림 2〉 빅데이터("big data")에 대한 관심 추세*



주: Google Trend에서 2014년 7월 30일에 잡힌 자료로 가장 관심이 높은 점을 기준(100)으로 한 상대적 추세임.

관련된 분야에 따라, 또는 다루는 문제에 따라 달라지기는 하겠지만 경험적 연구의 지평을 열어주는 기반으로서 이런 자료가 만들어진다는 것은 자료의 질을 높여주는 것뿐만 아니라 새로운 방식의 연구를 시도해 볼 수 있는 기회를 제공하는 것이기도 하다. 자료수집방법과 측정기술의 개선은 과학을 하는 데 있어서 필수적인 부분이라는 점에서 이 측면의 의의는 크다. <표 2>에서 보듯 이런 변화는 아직은 산업이나 경영, 컴퓨터 공학 등 몇몇 분야에 한정된 것으로 보인다. 그런 면에서 사회과학 분야에서도 이런 변화에 맞춰 여러 가지 논의가 활발해 지고 있는 것은 반길만한 일이다. 하지만 <표 3>에 정리한 사회학 분야에서의 반응처럼 아직은 문제제기 수준에 머물거나, 몇몇 실험적 사례의 소개, 낙관적인 예측에 대부분의 논의가 제한되어 있는 것으로 보인다. 그래서 이 글에서는 빅데이터 시대의 사회과학 패러다임에 초점을 맞추고 자료의 속성과 분석의 방법과 관련하여 짚어봐야 할 여러 문제점들을 검토해보았다.

〈표 2〉 "빅데이터"로 검색되는 국내출간도서 (출간년도 × 분야)

출간년도	분야					합계
	경제/경영	컴퓨터/IT + 과학/공학	사회 + 인문	잡지	기타※	
2011	3	1	0	4	0	8
2012	32	22	8	11	3	76
2013	59	60	19	22	13	173
§2014	(76)	(72)	(36)	(42)	(30)	(256)
합계	170	155	63	79	46	513

주: 네이버의 '책'메뉴에서 2014년 7월 28일자로 검색된 총 445건 중 중복되는 아이템을 제외한 숫자임. '기타' 부분은 다음 분류항목들을 포함함: 자기계발, 학습/참고서, 취업/수험서, 해외도서, 청소년, 종교, 어린이, 국어/외국어. ( )안은 검색일자를 감안하여 추정한 2014년 말의 예측값(=7월까지의 출간도서 수 × 2).

〈표 3〉 한국사회학회에서의 빅데이터 관련세션

2014년 6월 전기사회학대회 기획 및 조직: 이원재 (KAIST)	
"자연과학이 열어가는 빅데이터전환"	
정하웅 KAIST 물리학과	Anatomy of Science and Technology Using Google
문수복 KAIST 전산학과	전산학도가 갖는 인문사회학적 질문들
송길영 다음소프트	일상, 그 변화에 대한 관찰
2014년 6월 전기사회학대회 기획 및 조직: 김선업 (고려대), 신종화 (서울과학종합대학원)	
"데이터 네트워크 시대의 새로운 조사연구의 플랫폼 모색: 건강, 소비자, 여성, 선거 빅데이터 사례 활용"	
김선업 고려대학교 한국사회연구소	빅데이터 시대의 경험 사회학의 도전과 과제
신종화 서울과학종합대학원 빅데이터 연구센터	건강 빅데이터 분석사례: <지역사회건강조사>를 활용한 경력단절 여성의 일과 삶 연구
정희태 TSIS 전략서비스본부	기업 빅데이터의 사회학적 연구 활용방안
이택면 한국여성정책연구원 여성일자리·인재센터	공공 'Big-Enough Data' 분석 사례 및 과제

임상렬	리서치플러스	선거 빅데이터 사례
		2013년 12월 후기사회학대회 기획 및 조직: 장덕진 (서울대), 이원재 (KAIST)
“빅데이터와 한국 사회, 1” (서울대학교 사회발전연구소 SNCC-KAIST 문화기술대학원 SCL 공동세션)		
한규섭	서울대학교 언론정보학과	Party-based Selective Following on Twitter over Time: A Test of the Depolarization Hypothesis
차미영	KAIST 문화기술대학원	온라인 루머의 전파
이준환	서울대학교 언론정보학과	소셜미디어에서 사람들은 어떻게 행동하는가?
박주용	KAIST 문화기술대학원	A Neurotic’s World: Where the Big Fish Eat Small Fish
		2013년 12월 후기사회학대회 기획 및 조직: 장덕진 (서울대), 이원재 (KAIST)
“빅데이터와 한국 사회, 2” (서울대학교 사회발전연구소 SNCC-KAIST 문화기술대학원 SCL 공동세션)		
신종화	서울과학종합대학원 빅데이터 연구센터	한국인의 생활시간과 빅데이터 연구: 통계청 ‘생활시간조사’ 다년 분석을 중심으로
임성우	서울시 정보기획단	공공으로부터의 데이터 확장: 서울시 사례를 근간으로
김도훈	TREUM	빅데이터, 사회과학의 새로운 패러다임인가?
송길영	다음소프트	Mining Minds

이런 기술적 측면에 매달리다 보면 놓치기 쉬운, 여기서 꼭 물어야 할 질문은 이런 식의 연구가 왜 중요한지, 이런 식의 연구가 이전과 다른 점은 과연 무엇인지 하는 것이다. “1월의 국민여동생은 김연아였고, 2월에는 아이유다”라는 식의 ‘분석’에 15억이 넘는 규모의 자료를 썼다고 해서 그 결과가 더 중요해지지도 않고, 자료의 규모를 제외하면 지금까지의 시장분석과 딱히 다를 것도 없다(송길영, 2012). 지금까지는 상상하지 못했던, 지금까지는 불가능했던 그런 질문, 발견, 아이디어를 보여주지 못한다면 앞에서 논의한 문제점들을 고려할 때 구태여 새로 분석의 틀을 짤 필요성이 있는지를 질문하게 된다.

이처럼 “클수록 좋다”는 식의 빅데이터에 대한 무조건적 낙관론에 물음표를 던질 수는 있지만, 빅데이터가 가져온 자료기반의 변화는 이전과는 근본적으로 달라진 면

이 있다는 점에서, 그 영향이 넓고 크다는 점에서, 그리고 다시 돌이킬 수 없는 것이라는 점에서 '획기적(劃期的)'이라고 볼 수밖에 없다. 새로운 것에 대한 논의가 항상 그렇듯, 빅데이터 시대의 사회과학하기에 대한 논의도 한편으로는 위기에 대한 것이지만, 또 한편으로는 기회에 대한 것인 이유가 바로 여기에 있다고 본다. 중요한 것은 지금까지의 과학하기에서 얻은 분석의 기본을 지키면서, 지금까지의 틀을 확장하고 재구성하려는 시도이다. 지금이 바로 그런 시도의 기회이고, 그런 노력이 없으면 위기가 될 것이다(Kitcin, 2014; Macy and Golder, 2014; Lazer et al., 2009; Savage and Burrows, 2007).

앞에서 다룬 망원경이나 현미경의 예에서처럼 과학하기의 과정에서 도구가 차지하는 비중과 역할은 크다. 고전적 방법론에서 사회과학하기의 이론-방법-자료의 삼각형 정립(鼎立)구조는 이론에서 가설로, 그리고 조작화와 측정으로 이어져 검증에 이르는 연결고리로 구체화된다. 이 과정의 모든 단계에 도구는 맞물려 있다. 이런 의미에서 도구가 우리가 연구하는 현실을 구성하고, 따라서 도구가 바뀌면 이 틀 전체가 바뀐다는 입장까지도 생각해볼 수 있다(Latour, 2009: 9; du Gay and Pryke, 2002: 12-13). 이렇게 시각을 넓혀 잡으면 보통 자료라고 묶어 부르는 사회현상 관찰의 기록이 이 과정의 다양한 단계에서 어떤 역할을 하는지에 초점을 맞출 수 있게 된다. 새로운 자료환경에서의 사회과학하기에 대해 전망해 보는 것은 이런 의미에서 근본적인 과제이다. 그런 검토의 결과 필요하다면, 즉 그 변화가 가져올 새로운 틀이 더 나은 사회과학을 할 수 있게 한다면, 기존의 관행이나 인식의 틀을 깨는 것이 마땅하다. 하지만 이 결정이 기사의 헤드라인이나 잡지의 표지에 의한 것이어서는 안 된다. 물어야 할 것은 묻고, 따져야 할 것은 따져야 한다. 새 술은 새 부대에 담아야 한다면, 이 술이 과연 새 술인지를 확인해야 하고, 그렇다면 새 부대는 어떻게 만들어야 하는지에 대해서도 고민해야 한다.

현재의 패러다임 전환에 대한 논의에 더해져야 할 부분이 바로 이 '어떻게 사회과학을 할 것인가?'하는 근본적인 질문이다. 빅데이터의 이론적, 방법론적 기반에 대한 논의가 다분히 기술결정론적인 방향으로 흐르면서 깊이 있는 논의가 이루어지지 않고 있다. 자료의 양과 상관관계를 중심에 두는 이 '새' 틀이 그동안 다듬어온 사회과학하기의 입장에서는 오히려 후퇴한 것으로 보이는 이유가 그것인 것 같다. 그런 면에서 자료의 비중과 역할이 늘어난다고 했을 때, 어느 정도, 어떤 식으로 늘어나는지, 그렇게 늘어난 자료의 비중과 역할이 과학하기의 다른 요인들에는 어떤 영향을

주는지, 그리고 그렇게 새로 짜여진 틀의 구조와 기제는 어떤 것인지 등이 이런 패러다임 전환 논의의 핵심이 되어야 할 것이다.[18]

빅데이터라는 새로운 자료가 어떤 식으로 얼마나 기존의 과학하기의 틀을 바꾸어 낼 수 있는지를 살펴보는 것이 이 글에서 시도한 것이다. 특히 이론에서 자료로 분석틀의 중심을 옮기고, '이론의 종언'을 선언하는 새로운 틀로서의 빅데이터 경험주의에 대한 비판적 검토를 했다. 그들의 주장과 몇몇 사례에서 전통적인 연구의 틀과는 차별적인 모습을 볼 수는 있지만 그것이 대안적인 패러다임을 제시한다고 보기에는 이른 것 같다. 그런 대안이 가능하다면, 그것은 이런 "비판적 논의와 실제적 활용 사이의 대화"를 통해 만들어져야 한다고 본다(이재현, 2013).

빅데이터 시대에 사회과학이 성공적으로 적응하는 데 관건이 되는 것은 이렇게 변화된 자료 환경 속에서 무엇을 어떻게 수확하고 소화해 내는가 하는 것이다(Boyd and Crawford, 2012). 현재 빅데이터의 위상은 힘은 좋지만 길들여지지 않은 야생마에 비유할 수 있을 것 같다. 그 길들이기는 '과학하기'라는 큰 틀이 기본이 되어야 한다. 그리고 여기서의 과학은 앙리 프앙카레가 "돌을 쌓아 집을 만드는 것처럼 과학은 사실(자료)로 만들어 진다. 그렇지만 돌무더기가 집이 아닌 것처럼 사실(자료)의 더미가 과학은 아니다"라고 말했을 때의 그 과학일 것 같다.

참고문헌

매일경제 기획팀 · 서울대 빅데이터 센터. 2014.『빅데이터 세상: 당신의 숨겨진 욕망까지 읽어드립니다』. 매경출판.

송길영. 2012.『여기에 당신의 욕망이 보인다: 빅 데이터에서 찾아낸 70억 욕망의 지도』. 쌤앤파커스.

시로카 마코토. 2013. 김성재 옮김.『빅데이터의 충격: 거대한 데이터의 파도가 사업 전략을 바꾼다!』. 한빛미디어.

이상구. 2014. "Big Data: Trends and Opportunities," <빅데이터와 사회과학> 초청발표

18) 이와 관련하여 다음 용어 중 어느 것이 가장 빅데이터 시대의 연구경향을 잘 표현하는지에 대해서 아직 합의된 바는 없는 것 같다: data-driven, data-first, data-intensive, data-centric.

문 (2014.2.20). 서울대학교 사회과학대학.
이재현. 2013.「빅데이터와 사회과학: 인식론적, 방법론적 문제들」.『커뮤니케이션 이론』 9(3): 129-166.
정우진. 2013.『빅데이터를 말하다』. 클라우드북스.
한신갑. 2012.「혼합식 조사와 웹패널의 (옅은) 빛과 (짙은) 그늘」.『조사연구』 13(3): 1-31.
함유근 · 채승병. 2012.『빅데이터, 경영을 바꾸다』. 삼성경제연구소.

Abend, Gabriel. 2008. “The Meaning of ‘Theory’.” *Sociological Theory* 26(2): 173-199.
Anderson, Chris. 2008. “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.” *Wired* 16.07. (23 June 2008).
Ansolabehere, Stephen and Eitan Hersh. 2012. “Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate.” *Political Analysis* 20(4): 437-459.
Arbesman, Samuel. 2012. “Big Data: Mind the Gaps.” *Boston Globe* (30 Sep 2012)
Auletta, Ken. 2014. “[Annals of Communications] Outside the Box: Netflix and the Future of Television.” *The New Yorker* (3 Feb 2014).
Babbie, Earl R. 2012. The Practice of Social Research. Cengage Learning.(고상호 · 김광기 · 김상욱 옮김. 2013.『사회조사방법론』).
Berry, D. 2011. “The computational turn: thinking about the digital humanities.” *Culture Machine* 12. http://www.culturemachine.net/index.php/cm/ article/view/ 440/470 (11 July 2011).
Blossfeld, Hans-Peter, Katrin Golsch, and Gotz Rohwer. 2007. *Event History Analysis With Stata*. Psychology Press.
Bollier, D. 2010. “The promise and peril of big data” (11 July 2011). http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.
Bowker, Geoffrey C. 2005. *Memory Practices in the Sciences*. The MIT Press.
Box, George. 1987. *Empirical Model-Building and Response Surfaces*. New York: Wiley.
Boyd, Danah, and Kate Crawford. 2012. “Critical Questions for Big Data.” *Information, Communication & Society* 15(5): 662-679.
Butler, Declan. 2013. “When Google got flu wrong.” *Nature* 494: 155-156.
Cook, Samantha, Corrie Conrad, Ashley L. Fowlkes, Matthew H. Mohebbi. 2011. “Assessing Google Flu Trends Performance in the United States during the

2009 Influenza Virus A (H1N1) Pandemic." *PLoS ONE* 6(8): e23610.

Craig, David J. 2013. "The Ghost Files." *Columbia Magazine* (Winter 2013-14): 16-23.

Deming, W. E. 1960. *Statistical Design in Business Research*. New York: Wiley.

Dillman, Don A., Jolene D. Smyth y Leah Melani Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3rd Edition). New Jersey: John Wiley & Sons, Inc.

Du Gay, P. and Michal Pryke (eds.). 2002. *Cultural Economy: Cultural Analysis and Commercial Life*. London: Sage.

Duncan, Otis Dudley. 1984. *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage Foundation.

Fenn, Jackie and Mark Raskino. 2008. *Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time*. Harvard Business Review Press.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant. 2009. "Detecting influenza epidemics using search engine query data." *Nature* 457: 1012-1014 (19 Feb 2009).

Gitelman, Lisa (ed.). 2013. *"Raw Data" is an Oxymoron*. The MIT Press.

Gleick, James. 2011. *The Information: A History, A Theory, A Flood*. New York: Pantheon Books.

Halevy, Alon, Peter Norvig and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24(2):8-12.

Hume, David. 1978. *A Treatise of Human Nature* (2nd edition). Oxford University Press.

Hunt, Morton. 1985. *Profiles of Social Research: The Scientific Study of Human Interactions*. NY: Russell Sage Foundation.

Kitcin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1(1):1-12.

Korb, Kevin B. 2004. "Introduction: Machine Learning as Philosophy of Science." *Minds and Machines* 14:433-440.

Laney, Douglas. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety." *Application Delivery Strategies* 949 (6 Feb 2001).

Latour, B. 2009. "Tarde's idea of quantification." pp. 145-162 in *The Social after Gabriel Tarde: Debates and Assessments*, edited by M. Candea. London: Routledge. http://www.bruno-latour.fr/articles/article/116-TARDE-CANDEA.pdf.

LaValle, Steve, Eric Lesser, Rebecca Shockley, Michael S. Hopkins and Nina Kruschwitz. 2011. "Big Data, Analytics and the Path from Insights to Value." *MIT Sloan Management Review* 52(2): 21-32.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323(5915): 721-723.

Lazer, David, Ryan Kennedy, Gary King, and Vespignani Alessandro. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176): 1203–1205.

Leinweber, D. 2007. "Stupid data miner tricks: overfitting the S&P 500." *The Journal of Investing* 16(1): 15–22.

Lesk, M. 1997. "How much information is there in the world?" http://www.lesk.com/mlesk/ksg97/ksg.html.

Lewis, Michael. 2014. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company.

Lyman, Peter and Hal R. Varian. 2000. "How Much Information," http://www.sims.berkeley.edu/how-much-info on (8/5/2014).

Macy, Michael W., and Scott A. Golder. 2014. "Digital Footprints: Opportunities and Challenges for Social Research." *Annual Review of Sociology* 40: 129-152.

Madrigal, Alexis C. 2014. "How Netflix Reverse Engineered Hollywood." *The Atlantic* (2 Jan 2014).

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute.

Marini, Margaret M. and Burton Singer. 1988. "Causality in the Social Sciences." *Sociological Methodology* 18:347-409.

Mayer-Schönberger, Viktor and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt.

Merton, Robert K. 1968. *Social Theory and Social Structure*. Free Press.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331(6014): 176-182.

Miller, Warren E., James A. Davis, Jerome Clubb, Bruce Russett, Martin David, and James N. Morgan. 1990. "Large-Scale Data Needs." pp. 587-610 in *Leading Edges in Social and Behavioral Science*, edited by R. Duncan Luce, Neil J.

Smelser, Dean R. Gerstein. NY: Russell Sage Foundation.

Poincaré, Henri. 2011(1901). *Science and Hypothesis*. Dover Publications.

Pool, Ithiel. S., Hiroshi Inose, Nozumo Takasaki, and Roger Hurwitz. 1984. *Communications flows: A census in the United States and Japan*. Amsterdam: Elsevier North Holland.

Price, Derek J. De Solla. 1961. *Science Since Babylon*. Yale University Press.

Priest, Christopher. 1995. *The Prestige*. London: Touchstone.

Ryan, Joseph W. 2013. *Samuel Stouffer and the GI Survey: Sociologists and Soldiers during the Second World War*. Univ Tennessee Press.

Savage, Mike, and Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41(5): 885-899.

Smith, Tom W., Peter V. Marsden, Michael Hout, and Jibum Kim. 2013. *General Social Survey Cumulative Codebook, 1972-2012* [MRDF]. Chicago: National Opinion Research Center.

Smolan, Rick and Jennifer Erwitt. 2012. *The Human Face of Big Data*. Against All Odds Productions.

Snow, John. 1855. *On the Mode of Communication of Cholera*. London: John Churchill.

Stouffer, Samuel A. and Edward A. Suchman (eds.). 1949. *The American Soldier: Adjustment During Army Life (Studies in Social Psychology in World War II, Vol. I)*. Princeton University Press

Taylor, Linnet. 2013. "Big Data: Rewards and Risks for the Social Sciences" (8 April 2013)http://linnettaylor.wordpress.com/2013/04/08/big-data-rewards-and-risks-for-the-social-sciences/

Tufte, Edward. 2014. "The Thinking Eye." University of Illinois. (10 April 2014)

한신갑(서울대학교 사회학과 교수)의 연구 분야는 사회조직(사회연결망/네트워크분석, 조직과 제도, 경제사회학, 사회계층의 구조와 과정), 경력과 생애사, 문화소비/소비문화, 사회사/역사사회학, 방법론(텍스트, 시간/공간)이다. 최근의 저서 및 논문으로 『막힌 길 돌아서 가기: 남북관계의 네트워크 분석』(2013), 「나이테 다시 그리기: 사회적 역할 구분에 따른 연령구분의 재설정」(2014, 김이선 공저), 「혼합식 조사와 웹패널의 (옅은) 빛과 (짙은) 그늘」(2012) 등이 있다.

[2015. 3. 17. 접수; 2015. 4. 20. 수정; 2015. 4. 27. 게재확정]

Doing Social Sciences in the Age of Big Data: Rethinking Analytical Strategy in the Changing Data Environment

Shin-Kap Han
Seoul National University

How do we do social sciences? It is a question that needs to be asked more often. Commonly found in the various answers is a tripartite structure of theory-method-data. I ask the question yet again in the context of a recent and drastic change in the data environment: the emergence of "Big Data." Reviewing the change and its characteristics, I consider possible analytical strategy to take advantage of the new development. In both quantity and quality, the data environment has been changing constantly. Big Data, too, can be located in the continuum. To illustrate its possibilities and limitations, I describe the case of Google Flu Trends (GFT). These critical discussions suggest what to do with the Big Data—what to take and what not to and, when you do, how to. With adjustments and extensions, the theory-method-data tripartite structure still seems to be capable of providing a solid framework to integrate the potentials of Big Data.

Key words: big data, social science methodology, data environment, analytical strategy, correlation and causation