토론: 허 명 회 (고려대학교 교수, 통계학)

빅데이터에 대하여 저는, 또한 통계학계는 발표자 및 전산학계 만큼 열정적이진 않습니다. 빅데이터는 최근에 처음 나타난 것이 아니고 당대의 기술상황에서 도전적인 데이터 관련 문제를 당대의 빅데이터로 봅니다. 현재의 빅데이터 관련 기술적 문제들이 최근 극복되어 가고 있으므로 조만간 "빅데이터"는 없어지고 "데이터"가 남을 것으로 예상합니다.

최근의 빅데이터 관련 통계적 이슈는 다음 두 가지로 줄일 수 있습니다.

첫째, 변수(feature) 수 가 관측 수 n보다 큰 데이터의 모형화 문제입니다. 대다수의 이전 통계적 분석 방법에서는 $p < n$임을 가정하였습니다. 최근 통계학계에서 lasso 등 여러 성과를 냈습니다.

둘째, 빅데이터는 대부분 이차적 자료이므로 대단히 조심하지 않으면 분석결과가 심각한 잘못일 수 있습니다. 분석에서 고려되지 않은 제3의 변수가 교락효과(confounding effect)를 미칠 수 있기 때문입니다. 사회에서는 빅데이터가 신뢰롭다고 생각하는 편인데, 일반 상식과는 다르게 실제로는 특정 연구목적에 맞춰 수집된 스몰데이터가 신뢰로운 편입니다.

셋째, 빅데이터 및 기계학습, 인공지공이 예측(prediction)에 머물러서는 기술일 뿐 과학일 수 없습니다. 설명(explanation)이 되어야 합니다. 이런 방향으로 관심과 노력이 필요합니다.

정리하자면, 빅데이터는 현재 다소 과잉된 흥분상태입니다. 차분하게 접근할 필요가 있습니다. 빅데이터는 전산학과 통계학 등 도구 개발자들의 전문영역이라기보다는 실제 사회 및 과학기술 연구의 방법론 추세입니다. 각 분야 영역으로의 확산은 글로벌 트렌드입니다.