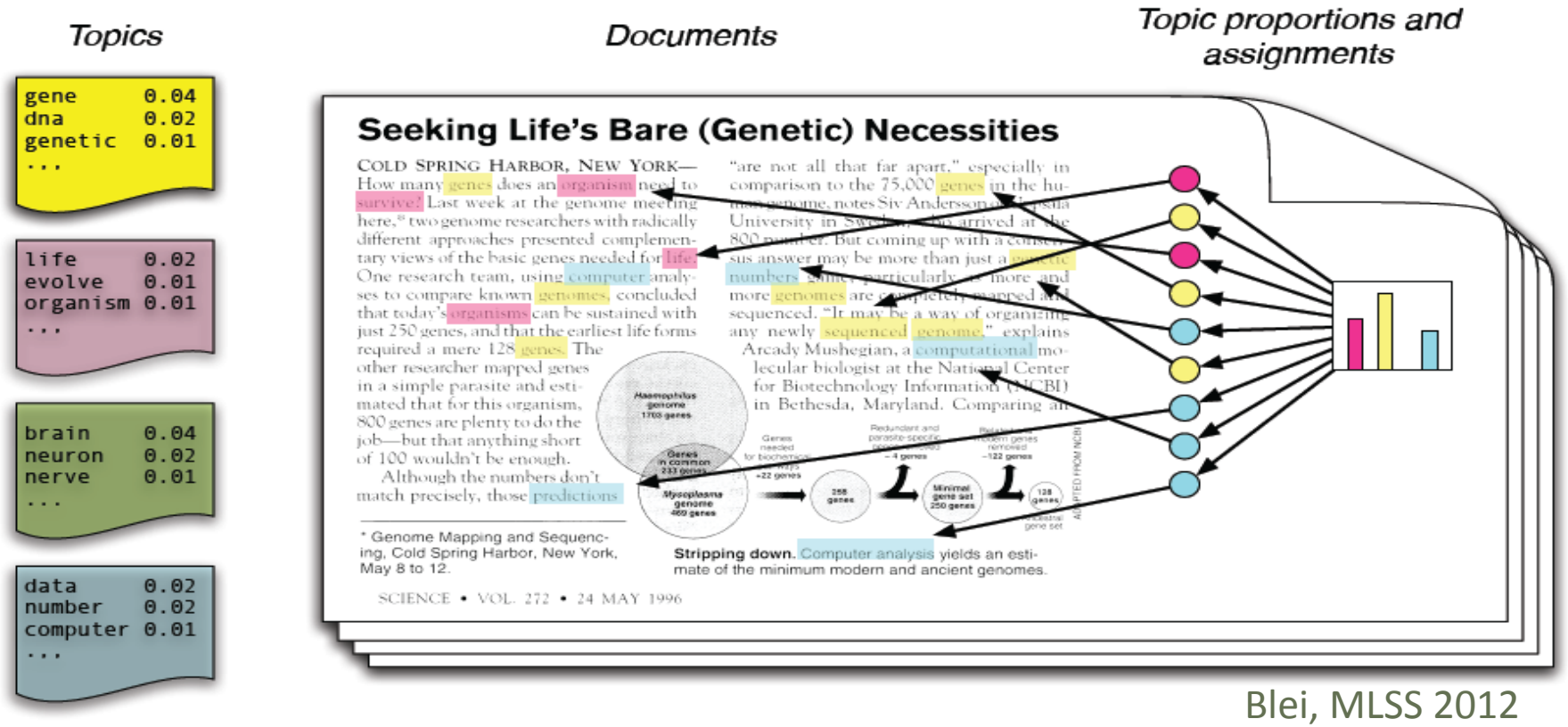
토픽모형을 이용한 마이크로 세그멘테이션

서울대학교 통계학과 박사과정
정구환

토픽모형

- 출현하는 단어의 빈도를 기반으로 문서를 분류하기 위하여 개발된 분석 방법



Blei, MLSS 2012

토픽모형

- 쇼핑자료(구매이력자료)에 적용 가능

문서자료	**쇼핑자료**
n개의 문서	n명의 고객
j번째 문서는 n_j개의 단어 포함	j번째 고객은 n_j개의 상품을 구매
단어의 종류는 W개	상품의 종류는 W개
x_{ji} : j번째 문서의 i번째 단어	x_{ji} : j번째 고객이 구매한 i번째 상품

토픽모형

- 토픽모형에서의 '토픽'은 군집분석에서의 '세그먼트'와 유사함
- 군집분석은 하나의 객체를 하나의 세그먼트에만 할당
- 토픽모형에서는 하나의 객체를 여러 개의 토픽에 할당 (토픽의 조합들이 세그먼트를 형성)

마이크로 세그멘테이션

- 한 고객이 여러 개의 토픽에 할당될 수 있음 (다중 멤버십)

- 고객의 니즈는 다양하기 때문에 여러 개의 토픽에 할당되는 것이 자연스러움

  ex) 고객 A는 스포츠 용품과 가구를 주로 구매함 ➡ 고객 A $\in$ $Topic_{스포츠}$, $Topic_{가구}$

- 10개의 토픽을 사용하면 1024 ($= 2^{10}$)개의 세그먼트를 만들 수 있음

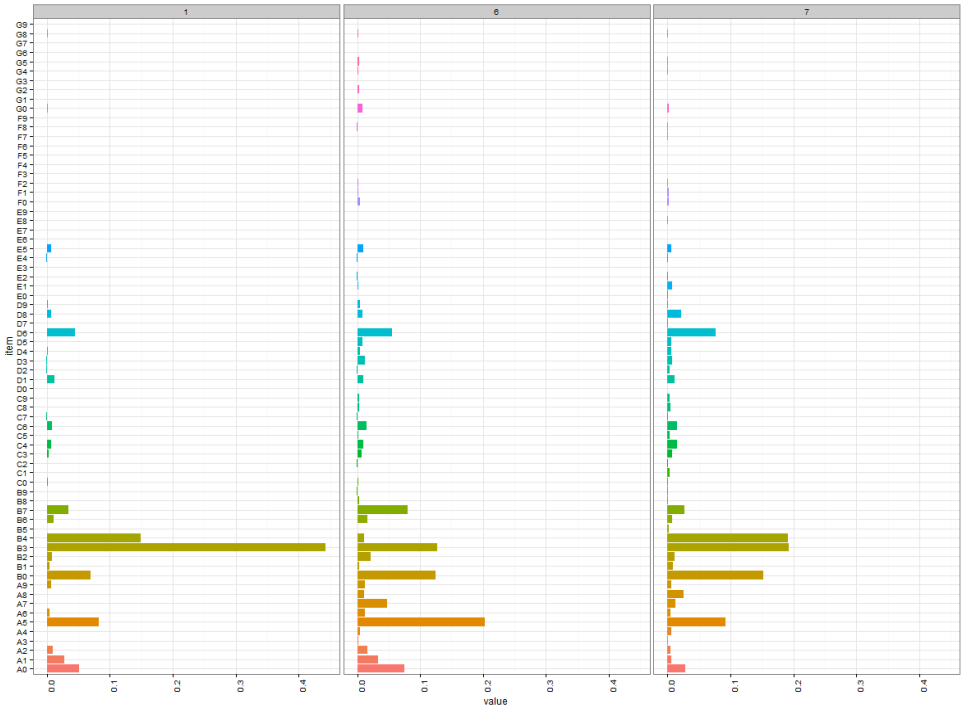
- 수천 개 혹은 수만 개의 고객 세그멘테이션을 위해서는 수십 개의 토픽이면 충분함

LDA 모형(Blei, 2003)

- 가장 널리 알려진 토픽모형인 Latent Dirichlet Allocation (LDA) 모형으로 쇼핑자료를 분석
- 쇼핑자료는 고객 10,000명의 70가지 상품에 대한 구매이력으로 구성되어 있음
- 토픽의 개수를 11개로 하여 LDA 분석 시행

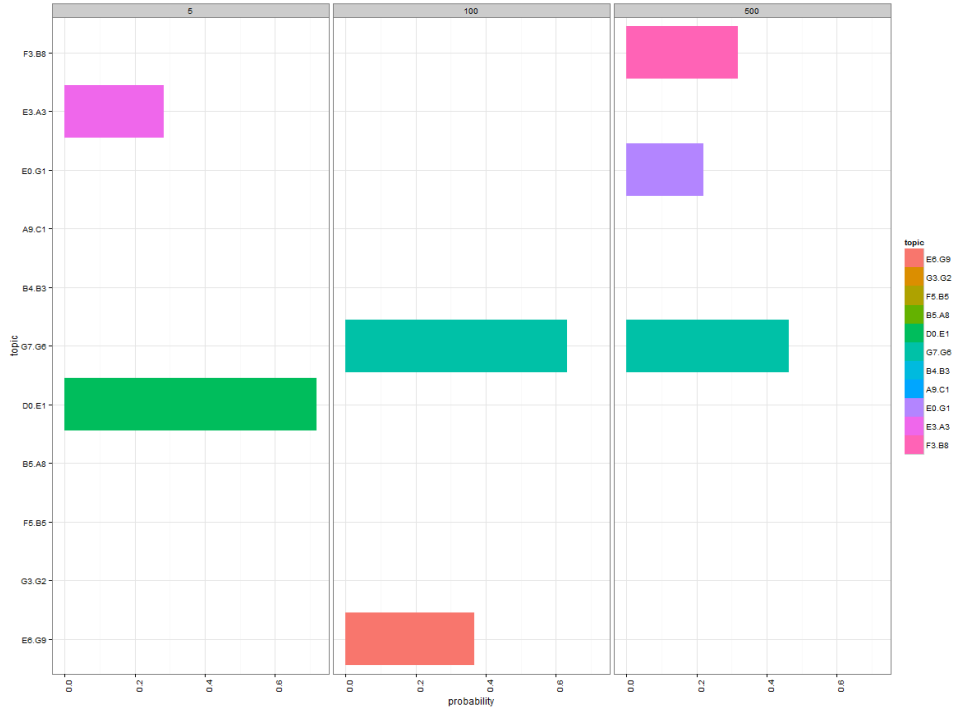
LDA 모형(Blei, 2003)

- 1, 6, 7번째 토픽의 확률분포와 리프트



LDA 모형(Blei, 2003)

- 5, 100, 500번째 고객의 토픽



토픽의 개수 선택

- LDA 모형에서는 토픽의 개수를 미리 정해주어야 함
- 쇼핑자료에서는 여러 가지 토픽의 개수에 대해 분석을 해 본 후, 가장 타당해 보이는 토픽의 개수를 선택함
- 여러 가지 토픽의 개수에 대해 분석하는 것은 시간 낭비가 심하고, 토픽의 개수를 선택하기 위한 기준이 불명확함

HDP 토픽 모형 (Hierarchical Dirichlet Processes, 2006)

- 비모수 베이지안 방법인 Hierarchical Dirichlet Processes (HDP) 토픽모형은 토픽의 개수가 무한히 많다고 가정

- 자료를 설명하는데 필요한 만큼의 토픽 개수를 스스로 찾음

연구 내용

- ✓ HDP 토픽 모형의 병렬 알고리즘
- ✓ HDP 토픽 모형의 온라인 학습 알고리즘
- ✓ HDP 토픽 모형의 다이나믹 알고리즘