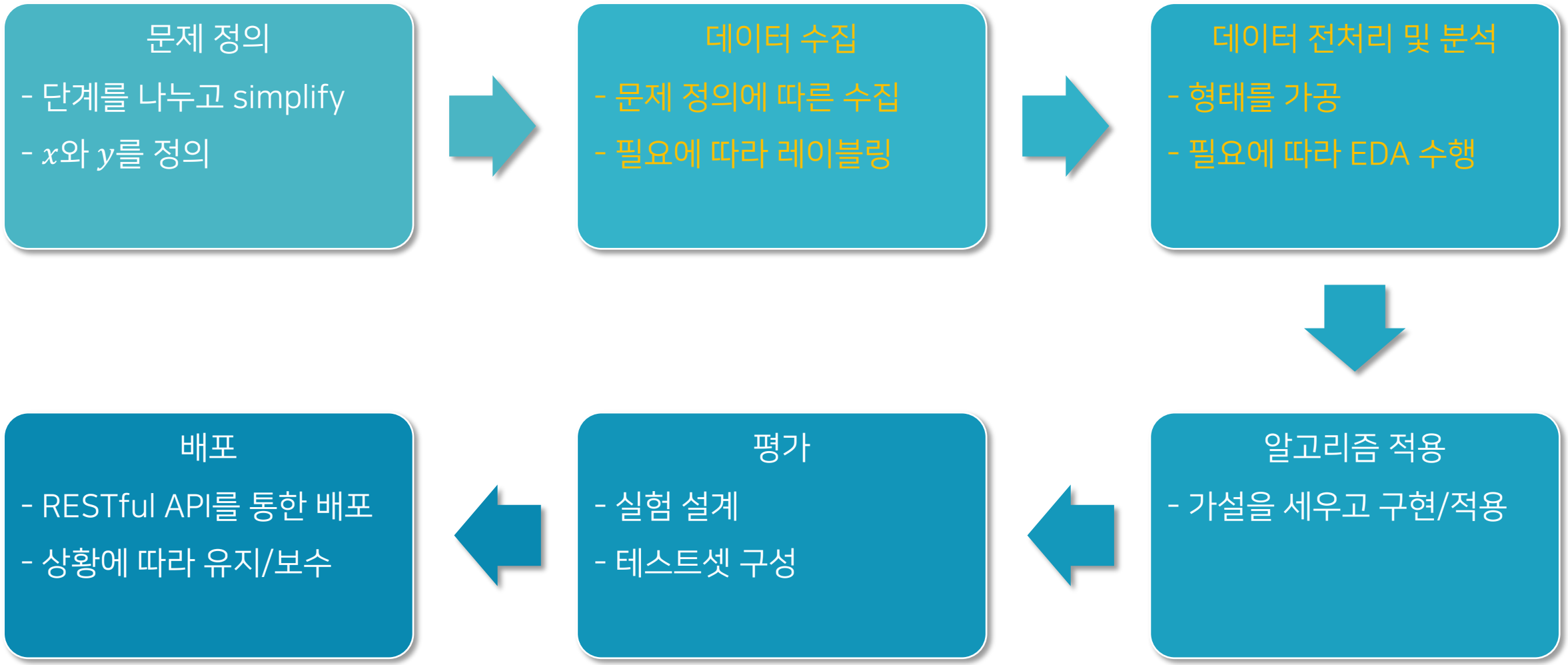
# Review: Preprocessing

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

# NLP Project Workflow



# Preprocessing Workflow

데이터(코퍼스) 수집
- 구입, 외주
- 크롤링을 통한 수집

정제
- Task에 따른 노이즈 제거
- 인코딩 변환

레이블링(Optional)
- 문장마다 또는 단어마다 labeling 수행

Tokenization
- 형태소 분석기를 활용하여 분절 수행

Subword Segmentation (Optional)
- 단어보다 더 작은 의미 단위 추가 분절 수행

Batchify
- 사전 생성 및 word2index 맵핑 수행
- 효율화를 위한 전/후처리

# Cleaning

- 기계적인 노이즈 제거
  - 전각문자 변환
  - Task에 따른 (전형적인) 노이즈 제거

- Interactive 노이즈 제거
  - 코퍼스의 특성에 따른 노이즈 제거
  - 작업자가 상황을 확인하며 작업 수행

- Therefore,
  - 전처리 과정은 Task와 언어, 도메인과 코퍼스의 특성에 따라 다르다.
  - 시간과 품질 사이의 trade-off
  - 따라서 전처리 중에서도 특히 데이터 노이즈 제거의 경우, 많은 노하우가 필요

# Tokenization

- 한국어의 경우
  1) 접사를 분리하여 희소성을 낮추고,
  2) 띄어쓰기를 통일하기 위해 tokenization을 수행

- 굉장히 많은 POS Tagger가 존재하는데,
  - 전형적인 쉬운 문장(표준 문법을 따르며, 구조가 명확한 문장)의 경우, 성능이 비슷함
  - 하지만 신조어나 고유명사를 처리하는 능력이 다름
  - 따라서, 주어진 문제에 맞는 정책을 가진 tagger를 선택하여 사용해야 함

# Subword Segmentation

- BPE 압축 알고리즘을 통해 통계적으로 더 작은 의미 단위(subword)로 분절 수행

- BPE를 통해 OoV를 없앨 수 있으며, 이는 성능상 매우 큰 이점으로 작용

- 한국어의 경우
  - 띄어쓰기가 제멋대로인 경우가 많으므로,
    normalization 없이 바로 subword segmentation을 적용하는 것은 위험
  - 따라서 형태소 분석기를 통한 tokenization을 진행한 이후,
    subword segmentaion을 적용하는 것을 권장

# Batchify