환경 빅데이터 분석 및 서비스 개발

중간자문회의(2018.6.27)

한국 환경정책·평가연구원

강 성 원

1. 연구 일반

2. 연구 진행 상황

3. 향후 계획

1. 연구 일반

개관

구분	내용	
연구성격	일반사업(연구형), 계속사업	
연구기간	2018.1 ~2018.12	
연구진	강성원 선임연구위원(책임) 진대용 부연구위원(부책임) 명수정 연구위원 홍한움 부연구위원	한국진 선임전문원 김진형 연구원 김도연 위촉연구원 강선아 위촉연구원 이동현 한국산업기술대 교수(위탁)
자문위원	내부	공성용 선임연구위원 김호정 연구위원 하종식 연구위원 신동원 부연구위원
	외부	김종률 정책관 (환경부 대기환경정책관) 강희찬 교수 (인천대학교 경제학과) 이성호 박사 (한국개발연구원) 오세영 박사 (한국행정연구원)
자문일정	착수자문회의: 2018년 3월 중간자문회의: 2018년 7월 최종자문회의: 2018년 10월	

기간, 인력, 예산

- 기간: 2018년 1월 – 2017년 12월
- 인력: 박사급 연구원 4명(1명 원외), 선임전문원 1명, 연구원 1명, 위촉연구원 2명 투입
- 예산: 2억 7천6백만 원 책정
  - 위탁연구비 4천 만원 책정: '컨벌루션 신경망(CNN)을 통한 미세먼지 예측'
    - 위탁과제 책임자: 한국 산업기술대학교 이동현 교수

목적: 빅데이터 연구방법론 환경연구 적용 가능성 모색

- **세부목적 1**: 환경 빅데이터 연구 수행
  - 주제선정 → 데이터 수집 및 가공 → 데이터 분석 → 결과 전달 전 과정 빅데이터 분석 기법 도입
- **세부목적 2**: 환경 빅데이터 연구 인프라 구축
  - 환경연구에 특화된 빅데이터 연구 플랫폼 구축
    - 데이터 수집 및 가공 → 데이터 분석 을 일괄 처리할 수 있는 연구환경 제공
  - 환경 빅데이터 연구 결과 축적된 자료 및 알고리듬 공유
  - 원내외 환경자료 수집·추출 사례 축적 및 공개
- **세부목적 3**: 원내외 빅데이터 서비스 개발
  - 환경 빅데이터 연구성과를 활용하여 연구정보 서비스 및 공공 서비스 개발

환경 빅데이터 연구 목적

◆ **빅데이터 연구 단계**

주제선정 → 자료수집 → 자료분석 → 결과전달

1. 환경빅데이터 연구
   - 주제선정 ~ 자료 분석 단계를 적용한 환경연구 수행
2. 환경 빅데이터 인프라
   - 자료수집~자료분석 단계를 수행할 수 있는 작업 공간 구축
3. 환경 빅데이터 서비스
   - '결과 전달' 단계를 확장하여 수요자 중심 서비스 개발

연속사업: 3년 단위 연구단계 설정

- 1단계(2017-19): 환경 빅데이터 연구 시작/ 환경 빅데이터 분석 플랫폼 설계
- 2단계(2020-22): 환경 빅데이터 분석 플랫폼 구축/빅데이터 활용 공공 서비스 설계
- 3단계(2023-25): 환경 빅데이터 분석 플랫폼 자동화 시도/공공환경 서비스 시범 사업

환경 빅데이터 분석 및 서비스 개발 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1기 (2017-19)	• **환경 빅데이터 연구 시행**	• **환경 빅데이터 분석 플랫폼 설계**	• 원내 연구정보 서비스
2기 (2020-22)	• 발신주기 단축	• 환경 빅데이터 분석 플랫폼 구축	• 연구기획 평가 및 준비 서비스 • 공공 서비스 설계
3기 (2023-25)	• 시의성 중심 발신체계 개편	• 환경 빅데이터 분석 플랫폼 지능화 시도	• 공공 서비스 시범 사업

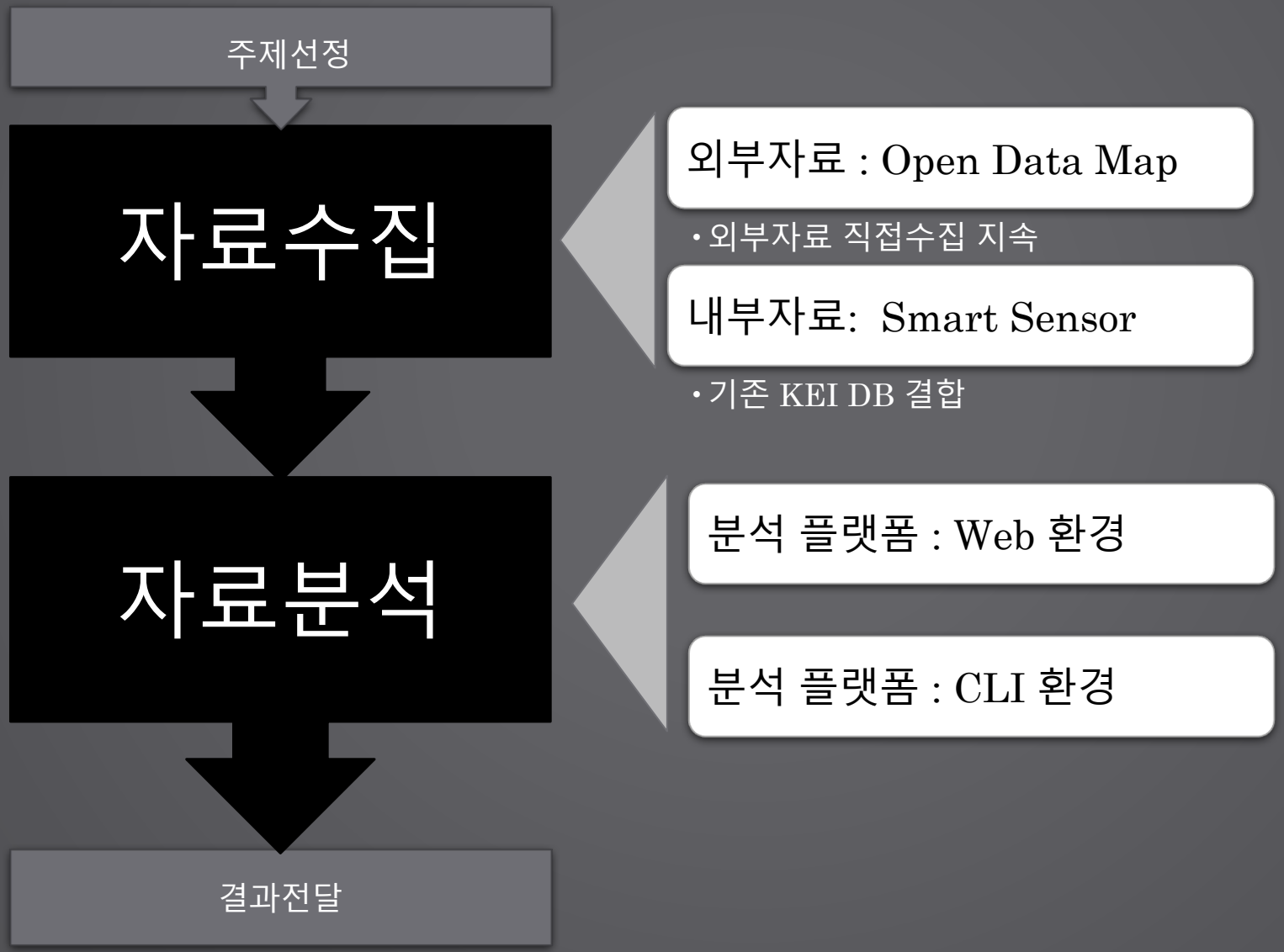
2017-19년 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1단계	**환경 빅데이터 연구 시행**	**환경 빅데이터 플랫폼 설계**	**원내 연구정보 서비스**
2017	• 환경연구 알고리듬 개발 - 전산화된 자료 + Deep Learning	• 환경분야 기초데이터 수집방법 • 자료 및 알고리듬 축적/공개	• 연구동향 파악 서비스
2018	• 환경연구 알고리듬 개발: - 비정형자료 + Deep Learning	• 환경 빅데이터 플랫폼 설계 - 대용량 자료 저장-분석 기능 구비 - 연구결과 자료 및 알고리듬 공유 - 환경 기초데이터 수집 결과 축적	• 연구동향 파악 서비스 원내 • 환경 데이터 포털(Open Data Map) 구축
2019	• 환경연구 알고리듬 개발 지속 • 딥러닝 기반 연구수요 분석 상시화	• 환경 빅데이터 플랫폼 설계 완료 - 연구결과 자료 및 알고리듬 공유 지속 - 환경분야 기초데이터 수집 1단계 완료	• 연구동향 파악 서비스 원외공개 • 환경 데이터 포털(Open Data Map) 원내 공개
2단계	**발신주기 단축**	**연구 과정 자동화/플랫폼 구축**	**연구기획 서비스/공공 서비스 설계**
3단계	**시의성 중심 발신체계**	**분석 플랫폼 지능화 시도**	**공공 서비스 시범 사업**

2018년 연구목표 1: 대용량 자료 활용 연구 플랫폼 설계

- ‘수요자 맞춤 지원행정’ 인프라 역할을 수행할 수 있는 환경 빅데이터 플랫폼 설계
  - 환경 빅데이터 플랫폼: 환경 데이터 활용 연구 및 환경 빅데이터 분석기법 개발 연구를 연구자가 수행할 수 있는 연구 환경
- 자료 수집, 축적: 환경 데이터 안내지도(Open Data Map)를 구축하고 기관 자체 자료를 결합하여 환경 데이터 안내지도를 보완
  - 환경 데이터 안내지도(Open Data Map) 구축 : 데이터의 목록과 Link를 제공
  - 기관 자체 자료를 사용하여 환경 데이터 안내지도를 보완: 자체수집, 기존 DB
- 자료분석: 대용량 자료 분석 및 빅데이터 분석 알고리듬 개발 환경
  - 기존 알고리듬 사용자: 사용자 편이성이 높은 Web기반 환경 제공
  - 알고리듬 개발자: 개발자의 자유도가 높은 CLI 기반 환경 제공
- 시험운영: 과제 참여자들이 설계된 플랫폼을 구현한 서버를 1년간 시험운영하여 플랫폼의 실용성을 점검

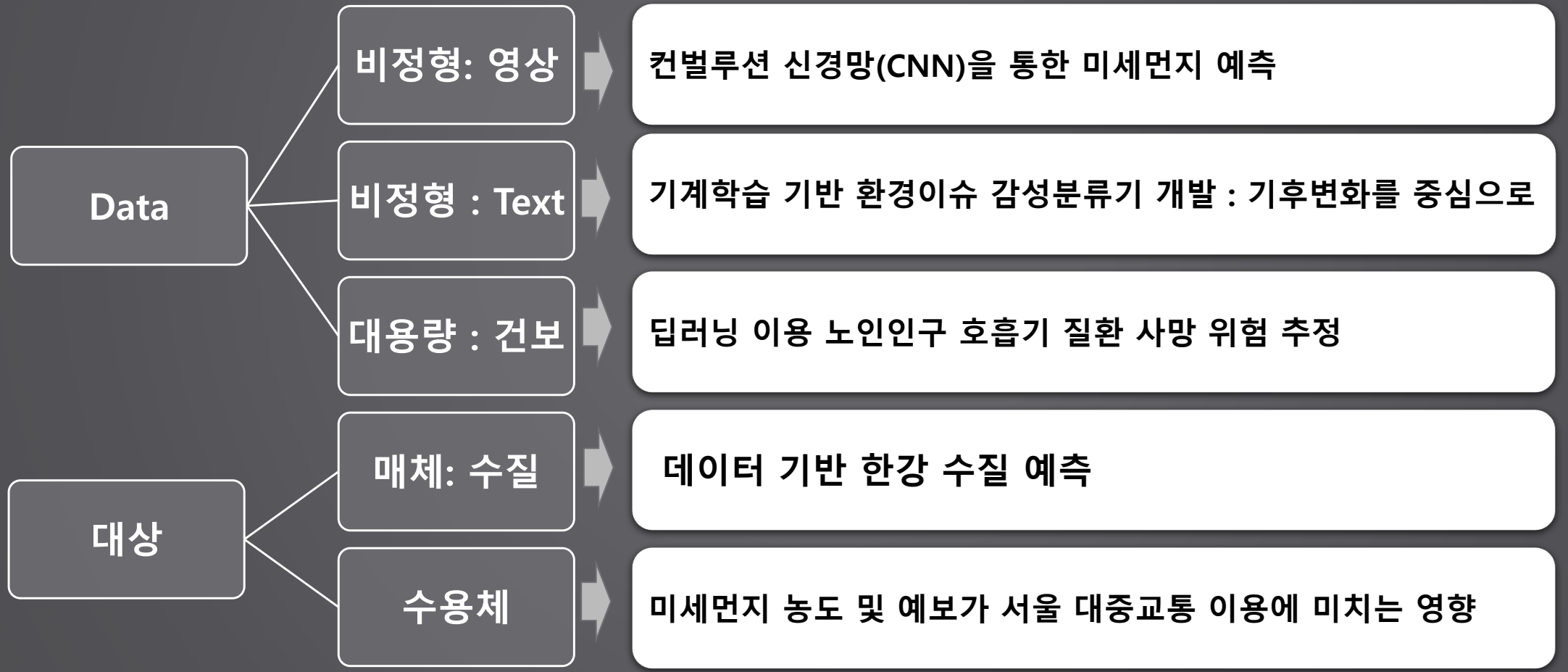
2018년 환경 빅데이터 연구 인프라 관련 연구 목표



2018년 연구목표 2: 환경 빅데이터 연구영역 확대

- 빅데이터 연구기법이 비교우위를 나타내는 비정형 대용량 자료 분석을 추진하고 연구영역을 확대
  - 2017년 연구: 전산화된 수치-텍스트 데이터/기후-대기 영역에 집중
- 비정형 대용량 자료 분석 : 화상(Image) 분석, 건강보험 자료, SNS 자료 분석
  - 화상 분석: 미세먼지 오염도를 이미지로 전환하여 컨벌루션 신경망 모형(CNN)을 적용
  - 건강보험 자료 분석: 건강보험 코호르트 자료를 이용하여 개인별 환경성질환 분석
  - SNS 자료 분석 : 환경이슈와 관련된 SNS 자료를 이용하여 환경이슈에 대한 감성을 분석
- 연구영역 확대: 수질오염 예측 및 환경위험에 대한 수용체 반응 분석
  - 수질오염 : 한강수계 측정소별 주간 수질오염 오염도(부영양화 정도) 예측 알고리듬 개발
  - 수용체 반응 : 미세먼지 오염도가 서울시 유동인구에 미친 영향 분석 알고리듬 개발

환경 빅데이터 연구영역 확대



2018년 연구목표 3: 연구동향 서비스 원내 공개 추진

주제선정 | 자료수집 | 자료분석 | **결과전달**

- 2017년 연구성과 '텍스트마이닝을 이용한 KEI 연구동향 분석'에서 개발한 연구동향 분석 알고리듬을 이용하여 연구동향 서비스를 개발하고 원내 공개
- LDA 기반 토픽 클러스터링 : 연구보고서를 유관성이 높은 토픽으로 분류하고 연간 토픽 구성을 파악하여 개괄적 연구 동향을 파악
  - KEI 보고서 및 NAVER News 제목
- 네트워크 분석: KEI 보고서의 연관어를 파악하여 연관 빈도가 높은 단어들의 네트워크를 도출
  - 네트워크 구성의 시간 별 추이를 파악하여 연구 동향을 파악
  - KEI 보고서 및 NAVER News 제목

2. 연구 진행 상황 (1)

환경 빅데이터 연구 인프라 구축(진대용)

착수 자문회의 자문의견 반영

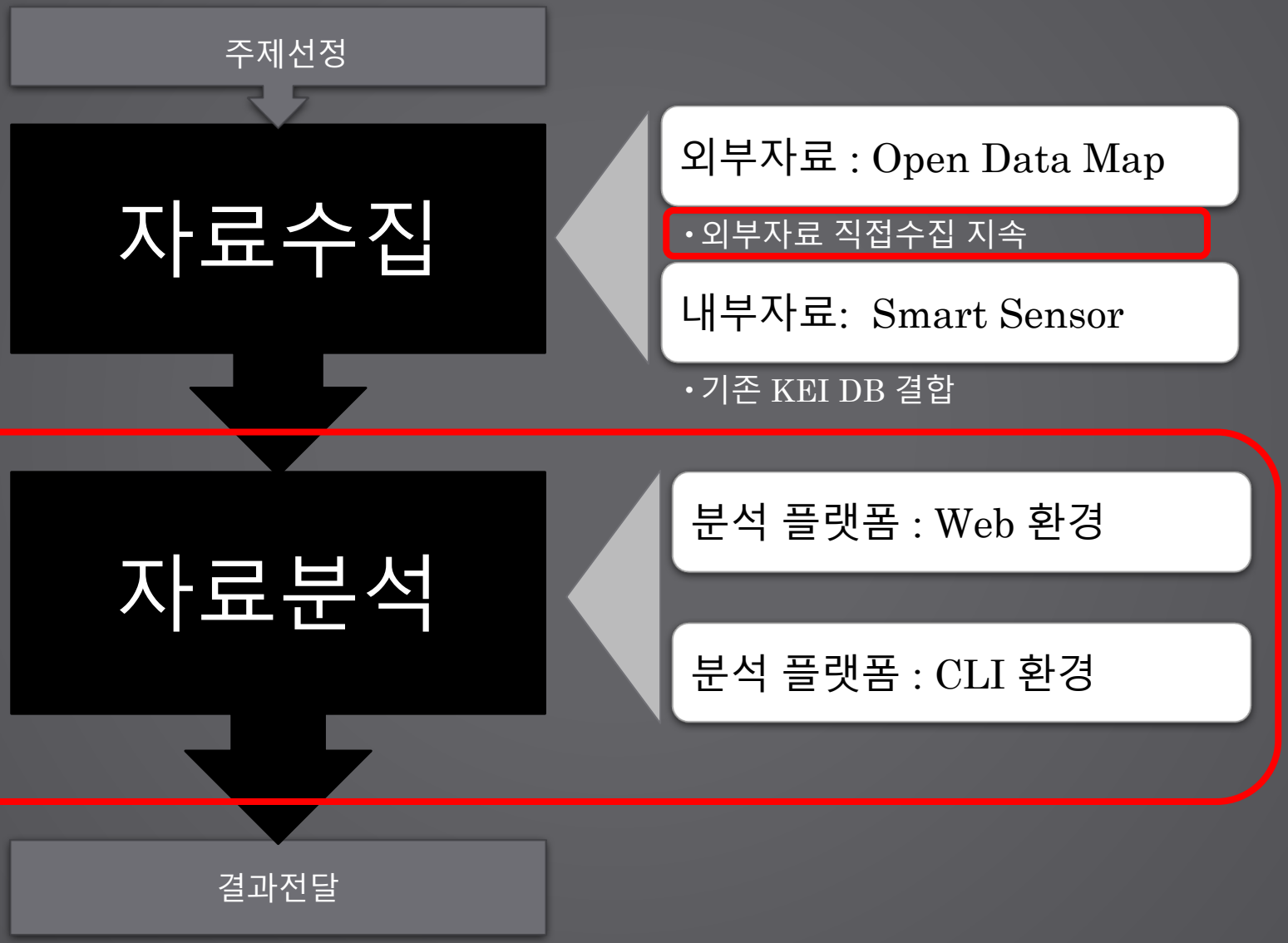
◆ 세부과제 연구 목적을 명확히 설정하고 부문 전문가 의견 청취 : 대기, 의료분야 전문가 접촉 중

	자문의견	반영내역
이성호	- 기존 방법론의 성과가 더 좋을 수도 있음 - IoT 센서를 활용하여 자체 수집 DB로 data map을 보완할 계획인데, 공식 데이터와 센서 데이터 간의 경합을 통해 데이터 신뢰성을 구축하면 어떤가?	- 기존 방법론과 딥러닝 방법론 성과 비교 - 센서 데이터와 측정소 데이터 비교 예정
강희찬	- 작년 연구에 대한 피드백을 이번 연구에 반영해야 함 - KEI 원내 설문조 결과로 얻어진 시사점 혹은 needs를 올해 연구에 반영해야 함 - 플랫폼 서비스 공개 범위를 정의해야 함 - DB 용량 확보 방안에 대한 계획 수립이 필요 - 감성분석 알고리즘 결과를 무엇에, 어떻게 활용할 것인지 사전에 규정 - 수용체 대응 연구들에서 선택한 topic이 "필요부급"한 것인지에 대한 고민 필요	- 2018년 과제 평가 반영 대규모 부정형 Data 분석 - 설문결과 반영 2018년 데이터 수집 범위 설정 - 정회원/준회원/객원 형태로 접근성 차등 공개 - (미반영) 2019년 예산에 반영 - 감성분석 알고리즘 개발 세부과제 분석과정에서 반영 - 연구 목적을 명확히 하여 보고서에 반영하겠음
오세영	- 성격이 다른 두 데이터 간의 결합이 가능한지 검토해야 함 - SNS 분석의 경우, 스토리 구성에 대한 논의가 추가적으로 필요함 - Open Data Map 개념 정의를 명확하게 해야 함 - 대부분의 연구가 단기 예측에 초점을 맞추고 있는데, 정책 대안 고민 필요	- 설문조사 데이터와 SNS 데이터는 상호 보완적 - (미반영) 2019년도에 연구를 확장하는 과정에서 반영하겠음 - 연구자가 데이터를 찾을 때 참고할 수 있는 지도 역할 - 민감도 분석 등을 추가하여 정책 시사점 도출
김종률	- 미세먼지 농도 예보에 따른 대중교통 이용 변화 연구에 있어서, 분석시점의 미세먼지 오염도, SNS상 국민 관심도 또는 감정변화, 이에 따른 대중교통 이용 변화 등에 대해 3각 분석을 할 경우 보다 심층적, 다각적 분석으로 정책적 활용도를 높일 것임	- -(미반영) 2018년에 대중교통 이용 변화 연구를 시작하는 단계이므로 복합 정보 분석을 수행할 여력이 부족. 2019년 연구에 반영하도록 하겠음.
공성용	- 세종시에 IoT 센서를 구축하여 데이터를 수집하고 있는데 데이터 품질 검증 필요 - 연구진행과정에서 알고리즘 개발 또는 분석 시 각 분야의 전문가 의견을 들어야	- 우선 데이터를 수집하여 평가하고 부족한 부분을 보완 - 자문회의, 인터뷰를 통해 전문가의 의견 반영하겠음
하종식	- 환경 빅데이터 분석 파트에서 관련 전문가들의 추가 자문이 필요 - 사망 위험 추정의 경우에 변수, 공간, 시간, 인구 등 각 요소들을 구체화해야 함 - 환경성 질환 예측 개인 단위 분석에서 대기오염 및 배출량 등 노출시간 구체화 필요	- 자문회의, 인터뷰를 통해 전문가 의견 반영하겠음 - 변수선정 부분에 반영 - 변수 구축 과정에서 반영
김호정	- 수질 예측에서 댐 방류량 등 수질에 영향을 미치는 수문 데이터를 고려해야 함 - 녹조, 바이러스 및 박테리아 등 기존 방법론으로 예측이 어려운 분야를 예측	- 수위 및 유량 자료 확보 중 - **용존 산소량에서 부영양화 정도로 분석 대상을 전환**
신동원	- 기존 빅데이터 한계가 무엇이며, 어떻게 극복하였는지에 초점을 맞춰서 분석 - 세부 연구에 대해 명확한 연구목적 설정이 필요	- 기존 빅데이터 연구의 일회성 극복 - 세부 연구에 대한 명확한 연구목적을 보고서에 반영

1. 환경 빅데이터 분석 플랫폼

2. Open Data Map

3. 스마트 센서 활용 Data 수집


주제선정
자료수집
외부자료 : Open Data Map
• 외부자료 직접수집 지속
내부자료: Smart Sensor
• 기존 KEI DB 결합
자료분석
분석 플랫폼 : Web 환경
분석 플랫폼 : CLI 환경
결과전달

(1) 환경 빅데이터 분석플랫폼

◆ 연구 내용: 연구자가 환경 데이터 수집-저장-분석을 수행할 수 있는 분석플랫폼 구축
- 2017년 서버 1기 도입(56-core, 192GB, 28TB) :오픈소스 기반의 분석플랫폼 구축

◆ 진행 방식: 연구진 공개 및 활용을 통한 요구사항 반영 및 유지관리 등 운영 노하우 축적
- 실제 연구활용 후 오류, 개선사항, 요구사항 수렴 및 반영 : 깃허브(github)를 통해 사용 가이드 배포
- 데이터 수요조사(실태조사)를 통한 데이터 서비스 개선
  - 2018년 목표 70건 중 현재 24건 수행

◆ 진행 상황: 연구자별 차별화된 연구환경 구축 중
- 정회원(Core) : 개발 – 터미널(Command Line Interface) 접근, 개인환경 구성
  - FTP(Filezilla), 프로그램 언어(Python, R, Tensorflow), 개발환경(Vim, Emacs) 구현
- 준회원(Member) : 분석 – 이미 구축된 웹 개발환경에서 알고리듬 사용 분석
  - Jupyter notebook(python), Rstudio(R) 개발환경 구현
- 객원(Guest) : 서비스 활용 – 로그인 없이 데이터 및 원내 빅데이터 서비스 활용
  - 데이터 : 수집된 공공 데이터 (AirKorea, 기상청, 네이버 환경뉴스)를 파일 형태로 등재, Smart Sensor Data 등재 예정
  - 원내 빅데이터 서비스(추가 예정) : 오픈 데이터 맵, 연구동향 서비스

CORE

MEMBER

GUEST

환경 빅데이터 분석플랫폼 시범운영 서비스입니다.

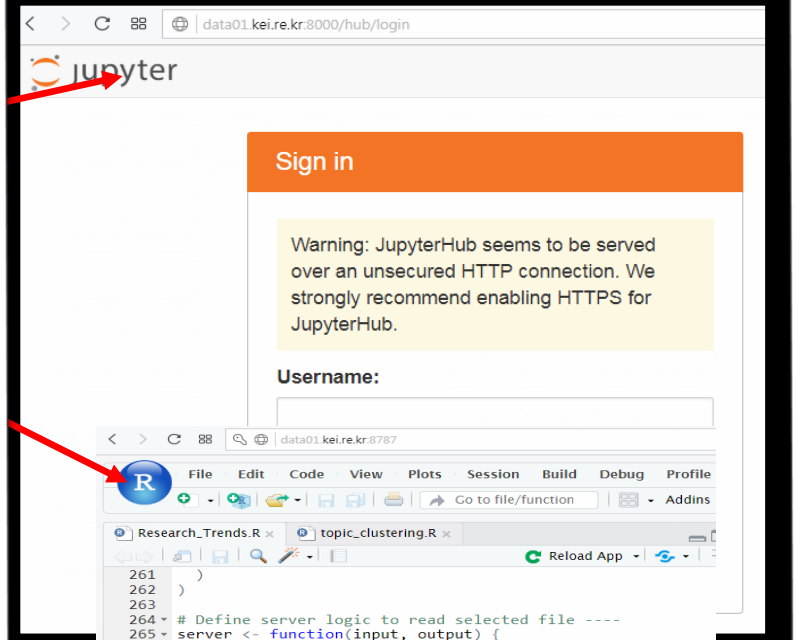
주피터 노트북 - http://data01.kei.re.kr:8000/
파이썬은 터미널(SSH) 사용을 권장드리며
virtualenv로 개인설정하셔서 주피터 노트북 구성하시기 바랍니다.

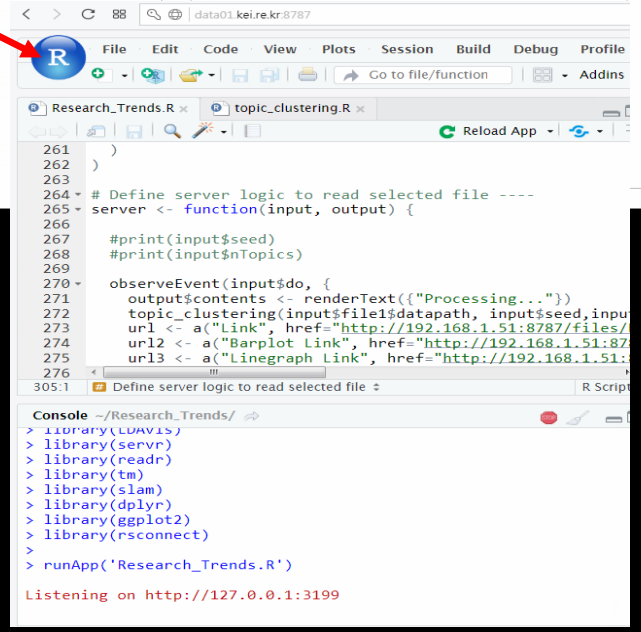
RStudio Server - http://data01.kei.re.kr:8787/

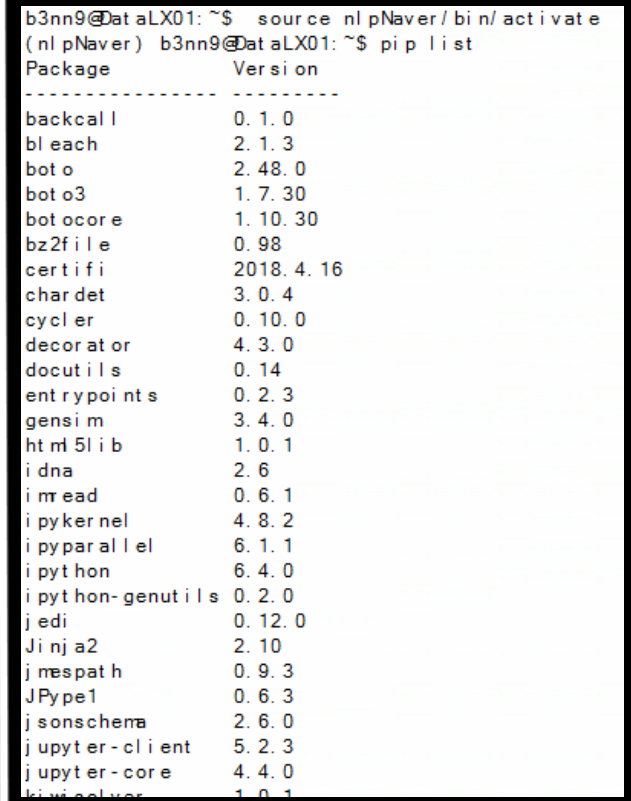
환경관련 데이터 서비스 - http://data01.kei.re.kr:8080/

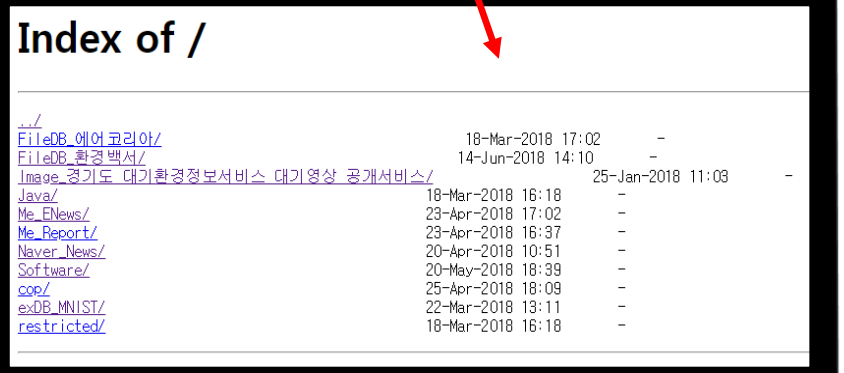
한글 파일명 사용은 지양해 주시고
필요시 FTP 소프트웨어를 사용하시기 바랍니다. - Filezilla

문의는 dataq@kei.re.kr로 주시기 바랍니다.





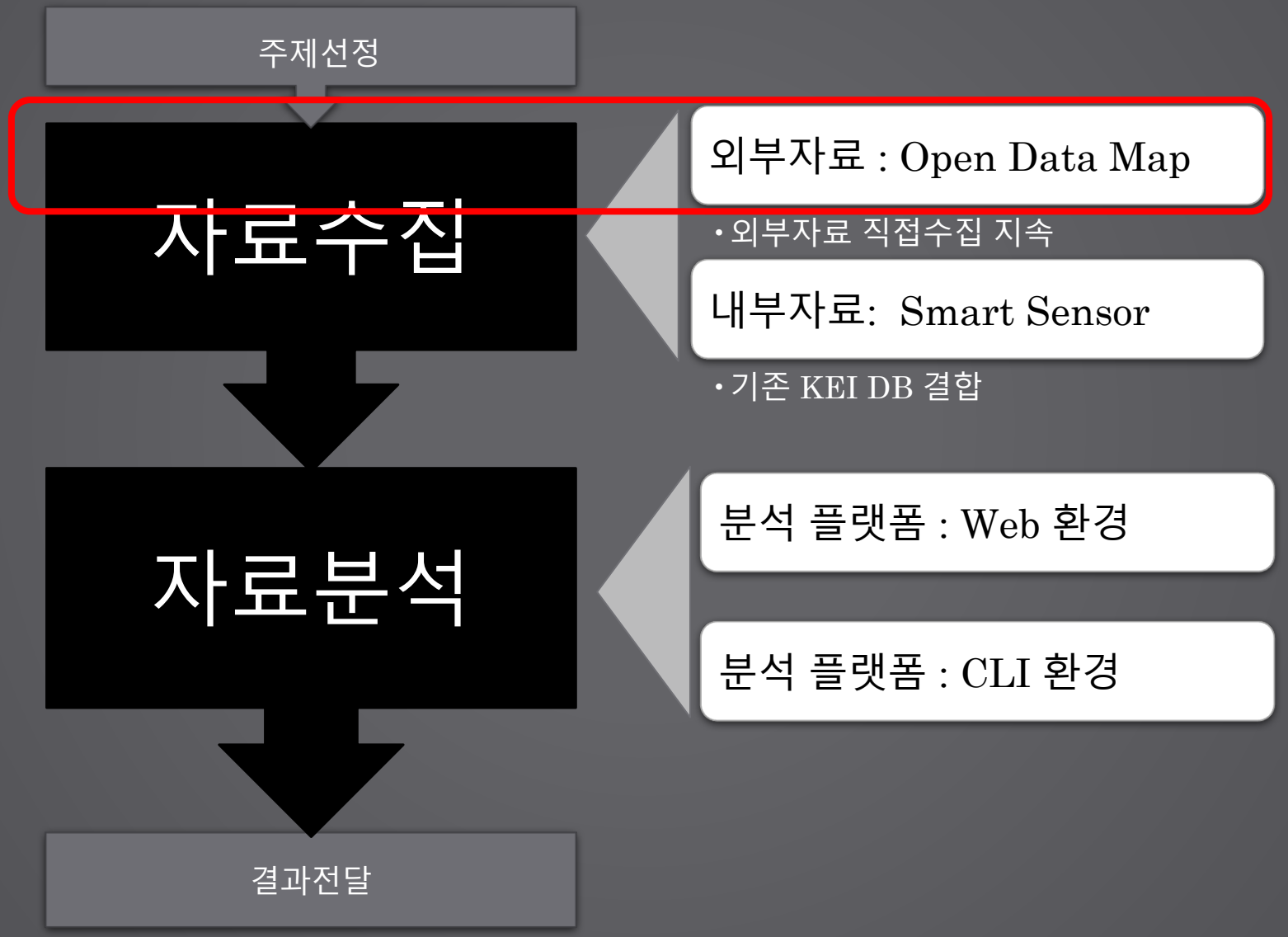




2018. 6. 25.

향후 계획: 성능개선, 활용 편이성 촉진, 점검 및 보수

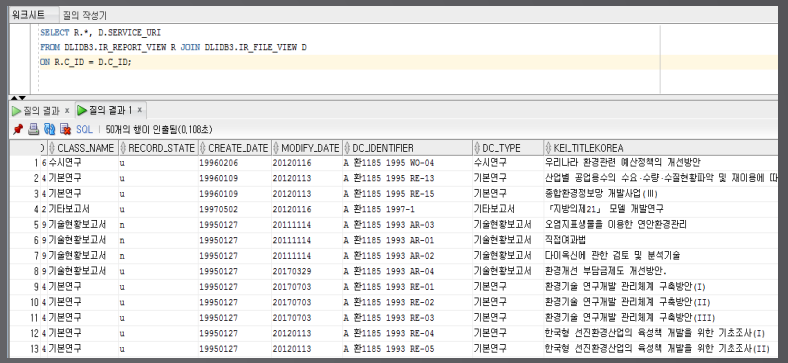
- 성능개선 : 서버 메모리 추가 증설
  - 텍스트 마이닝 분석 시 메모리 부족 오류 개선
- 활용 편이성 증진: 연구진 활용 feedback 반영 및 온라인 설명서 작성(Github)
- 주 1회 점검 및 보수


주제선정
자료수집
외부자료 : Open Data Map
• 외부자료 직접수집 지속
내부자료: Smart Sensor
• 기존 KEI DB 결합
자료분석
분석 플랫폼 : Web 환경
분석 플랫폼 : CLI 환경
결과전달

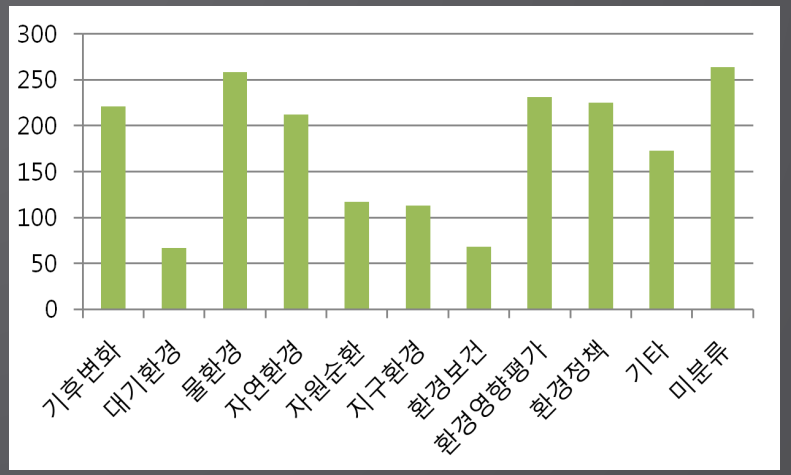
(2) Open Data Map

- Open Data Map : 환경연구 활용 빈도가 높은 온라인 데이터 hub
  - 카테고리 별 목록 /링크/ 내용소개를 제공
  - KEI 보고서에사 자주 인용된 데이터 source 우선 파악
- KEI 도서관 DB로 부터 KEI에서 발간된 문서정보 추출
  - 문서 경로, 문서 제목, 키워드, 카테고리 등
- 문서 수집 및 변환
  - PDF → TXT
  - 총 1979개 문서 중 1949개 문서 활용
    - 다운로드가 안되거나 txt 문서로 전환이 안되는 30개 문서 제외
- 문서 카테고리 재분류
  - 총 22가지 카테고리 → 11개 카테고리

KEI 도서관 DB



문서 카테고리



환경연구 활용 빈도가 높은 온라인 데이터 Source 파악

◆ KEI 발간된 문서 DB의 문서들로부터 데이터 리스트 추출

– 전처리

→ 정규표현식 활용 데이터 리스트 추출

→ 추가규칙 적용

→ 문서 카테고리별 카운트

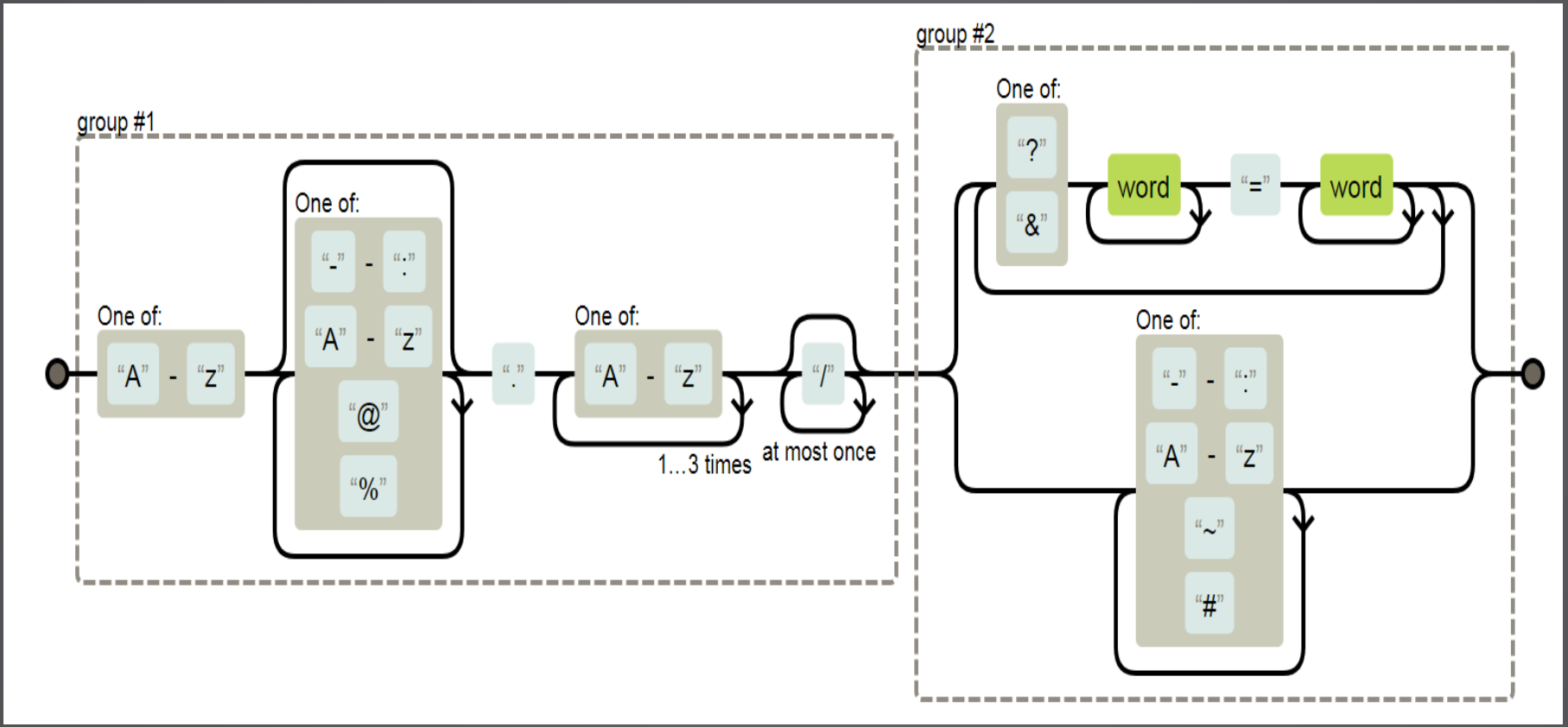
1단계: 정규표현식

- 정규표현식 핵심규칙
  - 영어로 시작
  - 콤마(.) 이후 1글자 – 3글자의 영어로 구성된 부분
    예) .kr .com 등
  - 그 이후에 내용이 있을경우 반드시 '/'

2단계: 추가 규칙

- 추가 규칙
  1. 프로토콜이 없는 경우 -> http:// 추가
  2. https://, http:// -> https://로 통일
  3. .kr, .com, .gov 등 국가기관 코드 없는 경우 제외

STEP 1 : 정규표현식


group #1
One of:
"A" - "z"
One of:
"-" - "."
"A" - "z"
"@"
"%"
"."
One of:
"A" - "z"
1...3 times
"/"
at most once
group #2
One of:
"?"
"&"
word
"="
word
One of:
"-" - ":"
"A" - "z"
"~"
"#"

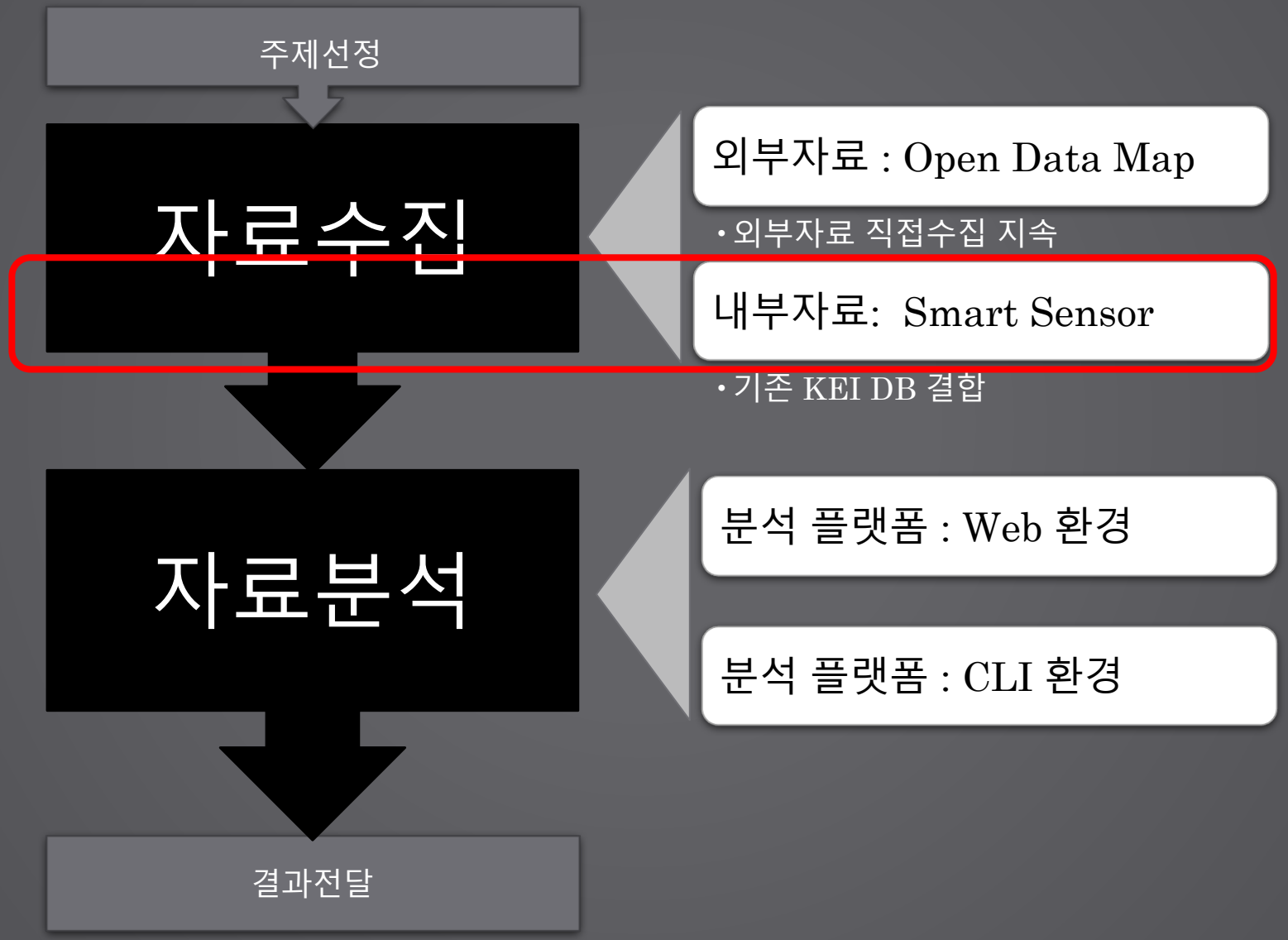
파악된 온라인 데이터 Source 범주 부여 : 인용된 문헌의 범주

각 카테고리 별 데이터가 사용된 문서의 갯수

순위	데이터	환경정책	환경보건	물환경	대기환경	자원순환	환경영향평가	기후변화	자연환경	지구환경	미분류	기타	합	분류	연구자 검증
1	https://www.me.go.kr	17	3	20	8	11	15	19	23	5	10	18	149	종합	종합
2	http://www.law.go.kr	1	3	8	0	5	10	11	17	4	8	14	81	종합	종합
3	https://www.epa.gov	5	4	15	4	8	4	4	11	0	6	9	70	종합	종합
4	https://kosis.kr	0	5	10	1	5	1	18	2	1	2	20	65	종합	종합
5	http://water.nier.go.kr	1	0	25	0	2	8	6	4	1	4	4	55	물환경	물환경
6	http://www.kma.go.kr	2	5	9	2	2	2	21	4	4	1	1	53	기후변화	종합
7	https://www.wamis.go.kr	4	1	20	0	0	2	11	11	1	1	2	53	물환경	물환경
8	https://www.eiass.go.kr	0	1	1	0	3	20	6	6	2	10	2	51	환경영향평가	환경영향평가
9	http://www.kosis.kr	1	4	4	0	1	3	13	3	1	1	11	42	기후변화, 기타	종합
10	https://www.index.go.kr	0	0	5	0	3	2	14	7	2	0	7	40	기후변화	종합
11	**https://www.airkorea.or.kr**	**2**	**8**	**2**	**1**	**2**	**1**	**9**	**3**	**3**	**3**	**4**	**38**	**기후변화, 환경보건**	**기후변화, 대기환경**
12	https://egis.me.go.kr	0	0	2	0	1	6	8	13	2	1	4	37	자연환경, 기후변화	종합
13	http://www.env.go.jp	3	2	1	4	3	4	2	2	4	6	4	35	종합	종합
14	http://www.moleg.go.kr	3	0	4	1	3	6	3	8	0	3	3	34	종합	종합
15	http://kosis.nso.go.kr	10	0	2	3	1	2	1	6	2	5	0	32	환경정책	종합
16	https://ecos.bok.or.kr	3	1	1	0	4	0	7	0	1	1	14	32	기타	기타
17	http://www.nier.go.kr	2	0	3	1	1	2	9	5	0	1	7	31	종합	종합
18	http://www.nso.go.kr	17	0	0	0	2	0	2	2	2	2	0	27	환경정책	종합
19	https://www.gims.go.kr	1	1	10	0	0	3	4	7	1	0	0	27	물환경, 자연환경	물환경
20	**http://airemiss.nier.go.kr**	**2**	**3**	**0**	**0**	**0**	**1**	**8**	**0**	**0**	**2**	**8**	**24**	**기후변화, 기타**	**기후변화, 대기환경**

향후 계획

- 오픈 데이터맵 데이터 리스트 보완 : 카테고리 랭킹 알고리듬
  - 카테고리 미분류 문서 → 카테고리별 분류
  - 카테고리 별 단순 빈도수 정렬 → 전체 빈도수 및 비율 등을 고려한 랭킹 알고리즘 적용
  - 원내 설문조사를 활용한 데이터 리스트 보완
- 오픈 데이터맵 원내 서비스화
  - 오픈 데이터맵 웹 서비스화
  - 오픈 데이터맵 서비스 원내 테스트 및 공개


주제선정
자료수집
외부자료 : Open Data Map
• 외부자료 직접수집 지속
내부자료: Smart Sensor
• 기존 KEI DB 결합
자료분석
분석 플랫폼 : Web 환경
분석 플랫폼 : CLI 환경
결과전달

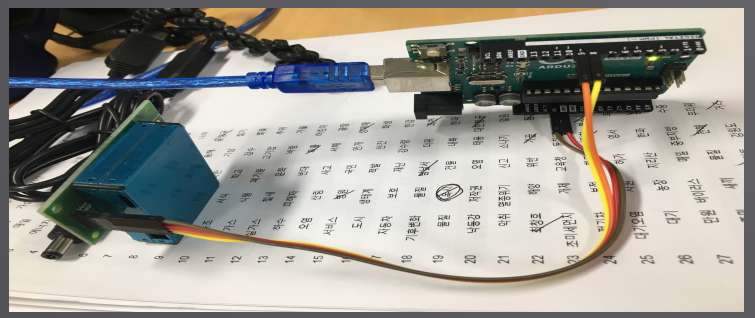
(3) 스마트센서 활용 Data 수집

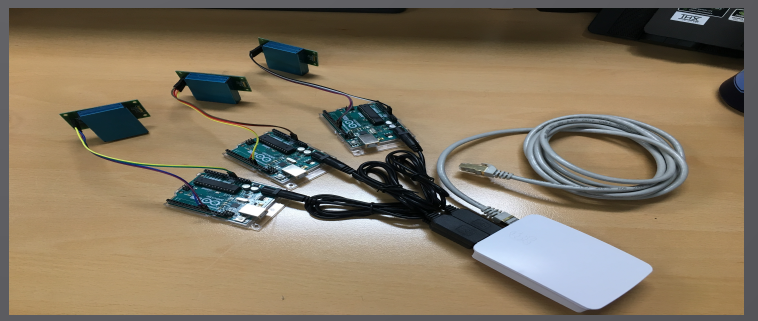
- 의의: 데이터 음영지역 및 소규모 연구대상 지역 데이터 수집 수단으로서의 가능성 모색
  - 미세먼지 측정기 미설치 지역 데이터 수집 : 데이터의 질을 평가하고 측정소 데이터와 비교
  - 저비용으로 자체 데이터 수집 및 수집-저장체계 마련 : 개별 연구/시민 참여형 연구 활용 가능성 점검

- 아두이노, 라스베리파이와 미세먼지 센서(PMS7003)를 활용한 스마트 센서를 구축하고 데이터를 수집 중
  - 스마트 센서 1기 : 센서 3기, 데이터 컨트롤러 3기(아두이노), 데이터 측정소 1기(라즈베리파이)
    - 미세먼지 센서[측정] - 아두이노 [데이터 전달] – 라즈베리파이 [임시저장]
  - 설치장소 : KEI(10층), 세종 새롬동(28층)
  - 수집된 Data 를 실시간으로 서버에 저장 : 10건 수집 Data 중 최대, 최소값 제외하고 평균값을 기록

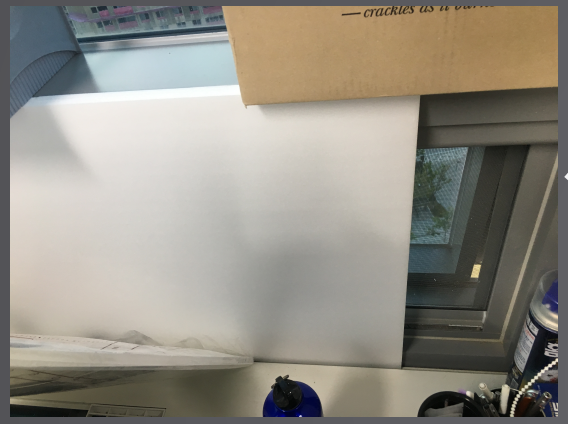
- 향후 계획 : 실시간으로 수집 데이터 파일을 생성하여 플랫폼에 수록 및 수집자료의 신뢰성 점검
  - 신뢰성 점검 : 3개 센서의 측정치 상호 비교 등 관측 데이터의 정보를 활용하고 전문가 상담을 진행

스마트 센서 시스템 Architecture

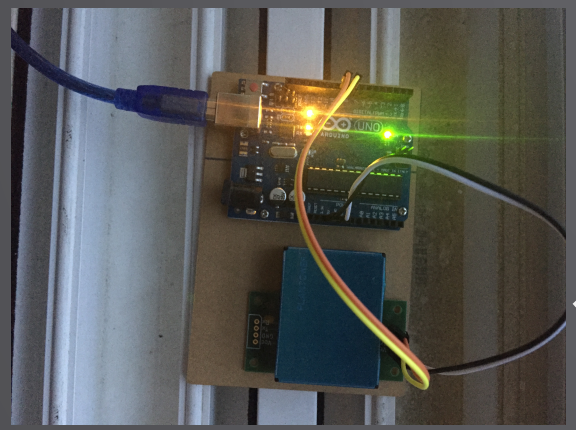




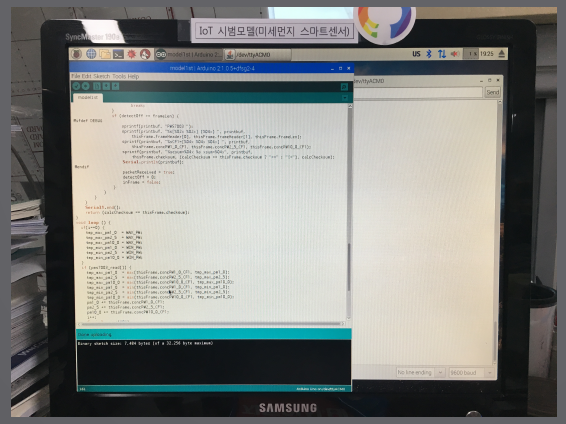
스마트센서 활용 데이터 수집 현황



설치(KEI)



설치(새롬동)



화면연결

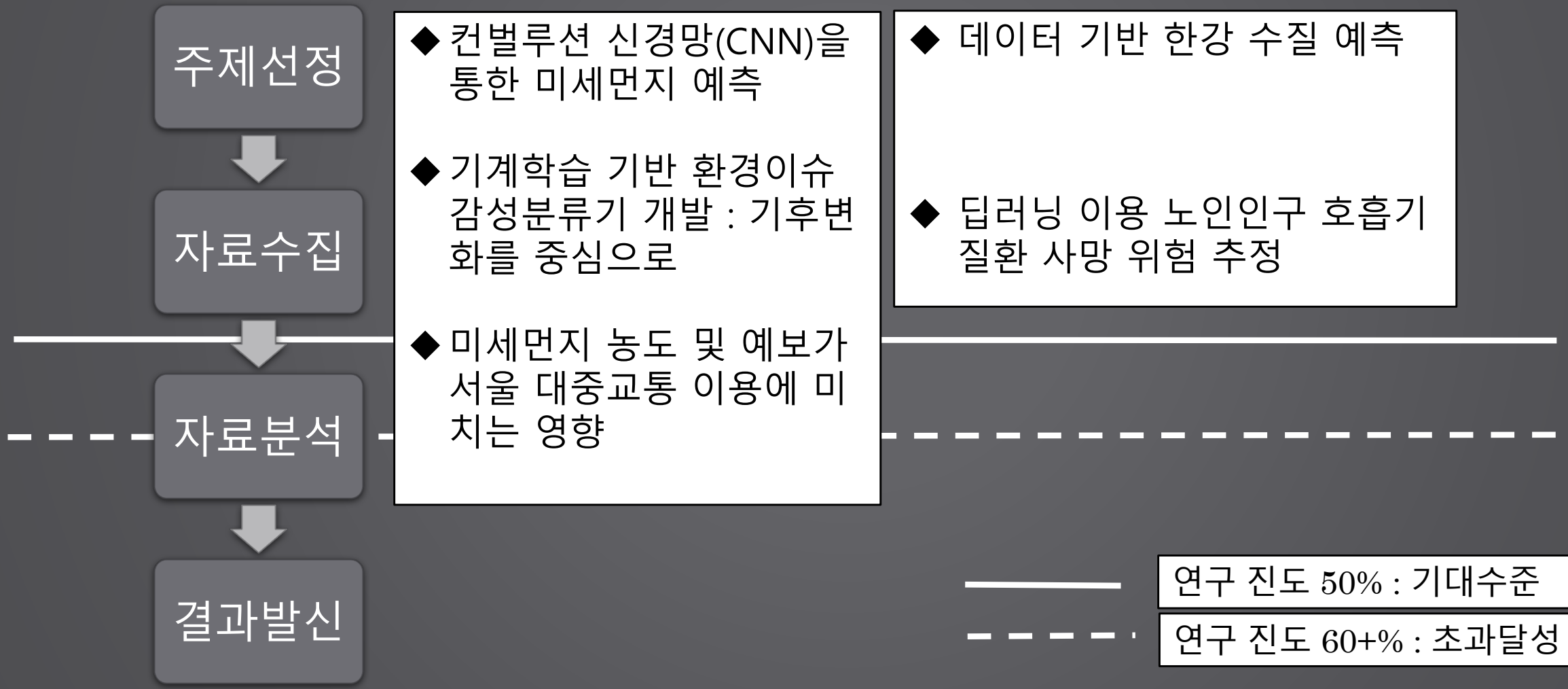
```
[Checksum OK] PM1.0 = 1b, PM2.5 = 27, PM10 = 28 [max =1c,28,2a, min = 1b,26,26]
0000000000
[Checksum OK] PM1.0 = 1b, PM2.5 = 26, PM10 = 28 [max =1c,27,2b, min = 1b,26,28]
0000000000
[Checksum OK] PM1.0 = 1b, PM2.5 = 26, PM10 = 2a [max =1c,27,2b, min = 1b,26,28]
0000000000
[Checksum OK] PM1.0 = 19, PM2.5 = 24, PM10 = 26 [max =1b,26,28, min = 19,24,26]
0000000000
[Checksum OK] PM1.0 = 17, PM2.5 = 20, PM10 = 22 [max =18,23,25, min = 16,20,21]
0000000000
[Checksum OK] PM1.0 = 17, PM2.5 = 21, PM10 = 21 [max =18,22,23, min = 17,21,21]
0000000000
[Checksum OK] PM1.0 = 17, PM2.5 = 22, PM10 = 24 [max =18,23,25, min = 17,21,23]
0000000000
[Checksum OK] PM1.0 = 17, PM2.5 = 23, PM10 = 25 [max =19,24,26, min = 17,23,25]
0000000000
[Checksum OK] PM1.0 = 19, PM2.5 = 24, PM10 = 27 [max =1b,27,2a, min = 17,23,25]
0000000000
[Checksum OK] PM1.0 = 1b, PM2.5 = 28, PM10 = 29 [max =1c,28,2a, min = 1b,26,29]
0000000000
[Checksum OK] PM1.0 = 1b, PM2.5 = 29, PM10 = 2b [max =1c,2a,2c, min = 1b,29,2a]
0000000000
[Checksum OK] PM1.0 = 1a, PM2.5 = 26, PM10 = 28 [max =1c,29,2b, min = 1a,25,26]
0000000000
[Checksum OK] PM1.0 = 18, PM2.5 = 23, PM10 = 27 [max =1a,26,28, min = 18,23,27]
0000000000
[Checksum OK] PM1.0 = 19, PM2.5 = 23, PM10 = 27 [max =19,24,28, min = 19,23,26]
0000000000
[Checksum OK] PM1.0 = 18, PM2.5 = 24, PM10 = 27 [max =19,25,28, min = 18,23,25]
0000000000
[Checksum OK] PM1.0 = 19, PM2.5 = 23, PM10 = 24 [max =1a,24,25, min = 18,23,24]
0000000000
[Checksum OK] PM1.0 = 19, PM2.5 = 23, PM10 = 23 [max =1a,24,26, min = 19,23,23]
```

2. 연구 진행 상황 (2)

환경 빅데이터 분석

- 컨벌루션 신경망(CNN)을 통한 미세먼지 예측
- 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화를 중심으로
- 데이터 기반 한강 수질 예측
- 딥러닝 이용 노인인구 호흡기 질환 사망 위험 추정
- 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향

연구 진행 상황 : 3건 시범 분석 시도, 2건 자료 수집 중



(1) 컨벌루션 신경망(CNN)을 통한 미세먼지 예측 [이동현]

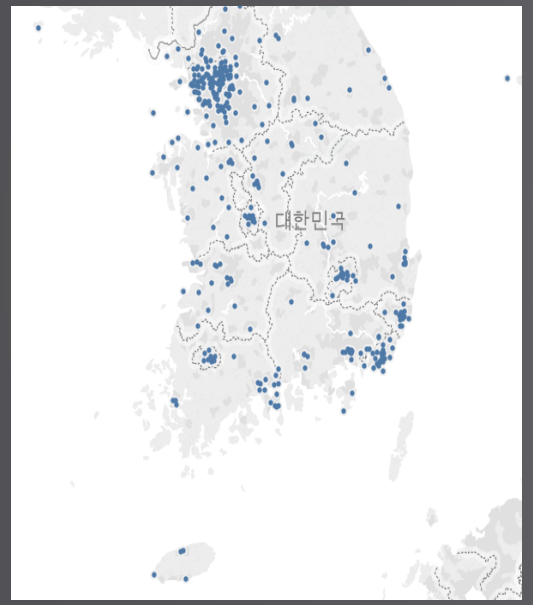
- 주변 지역 정보를 반영하여 대기오염을 예측하는 컨벌루션 신경망 모형
  - 전국 측정소 별 미세먼지를 거리를 반영하여 10 x 10 격자로 보간하고 컨벌루션 신경망(CNN: Convolution Neural Network)을 적용
  - 위도, 경도, 대기 및 기상자료, 미세먼지 오염도 → 미래 미세먼지 오염도 추정

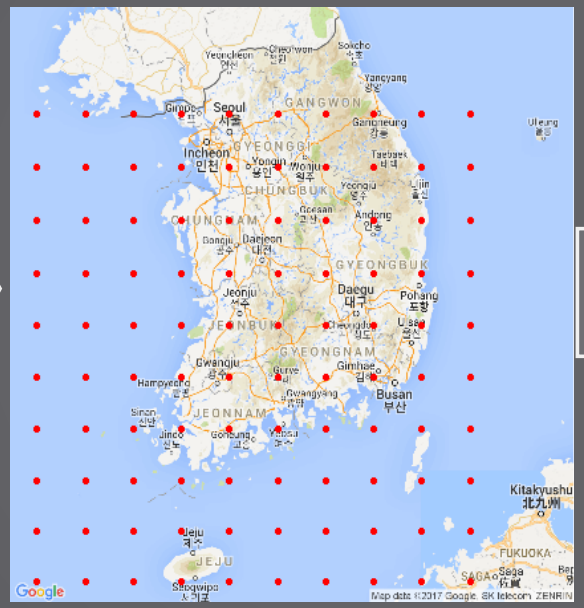
- 연구 내용 : 자체 개발 CNN 에 4가지 Architecture를 구축하여 미세먼지 오염도 예측
  - CNN Algorithm : 동일시점 과거정보 집적 + 주변공간 정보 반영 + 여타 변수 정보 반영
    - 첫 번째 아키텍처 (Arc1) : 이전 7시간의 PM10으로 다음시간의 PM10을 예측
    - 두 번째 아키텍처 (Arc2) : 이전 1시간의 요인들로 다음 1시간의 PM10을 예측
    - **세 번째 아키텍처 (Arc3) : 이전 7시간의 요인들로 다음 1시간의 PM10을 예측**
    - 네 번째 아키텍처 (Arc4) : 이전 7시간의 요인들로 다음 1시간의 지역별 PM10을 예측
  - Stochastic Gradient Descent (SGD) + Adaptive Moment Estimation (ADAM) 최적화 기법 이용

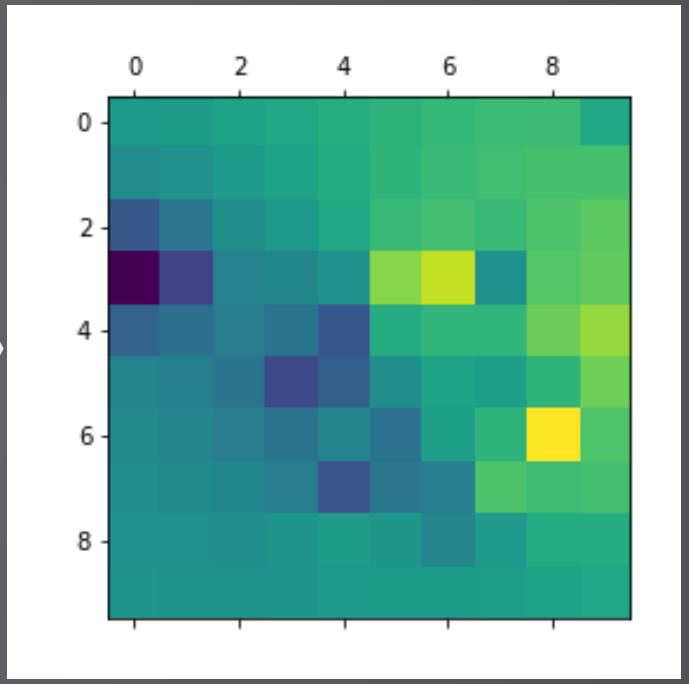
- 연구 성과 : 미세먼지 7시간 예측치의 평균제곱근오차(RMSE)를 2.21 $\mu g/m^3$ 까지 축소

- 향후 계획 : Hyper Parameter 조정을 통한 예측 오차 개선 가능성 점검 및 강건성(Robustness) 검증

데이터 변환 : 측정소 자료를 격자형 자료로 보간



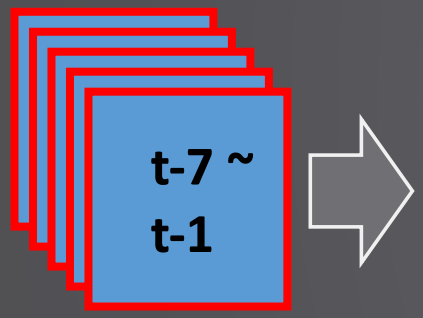




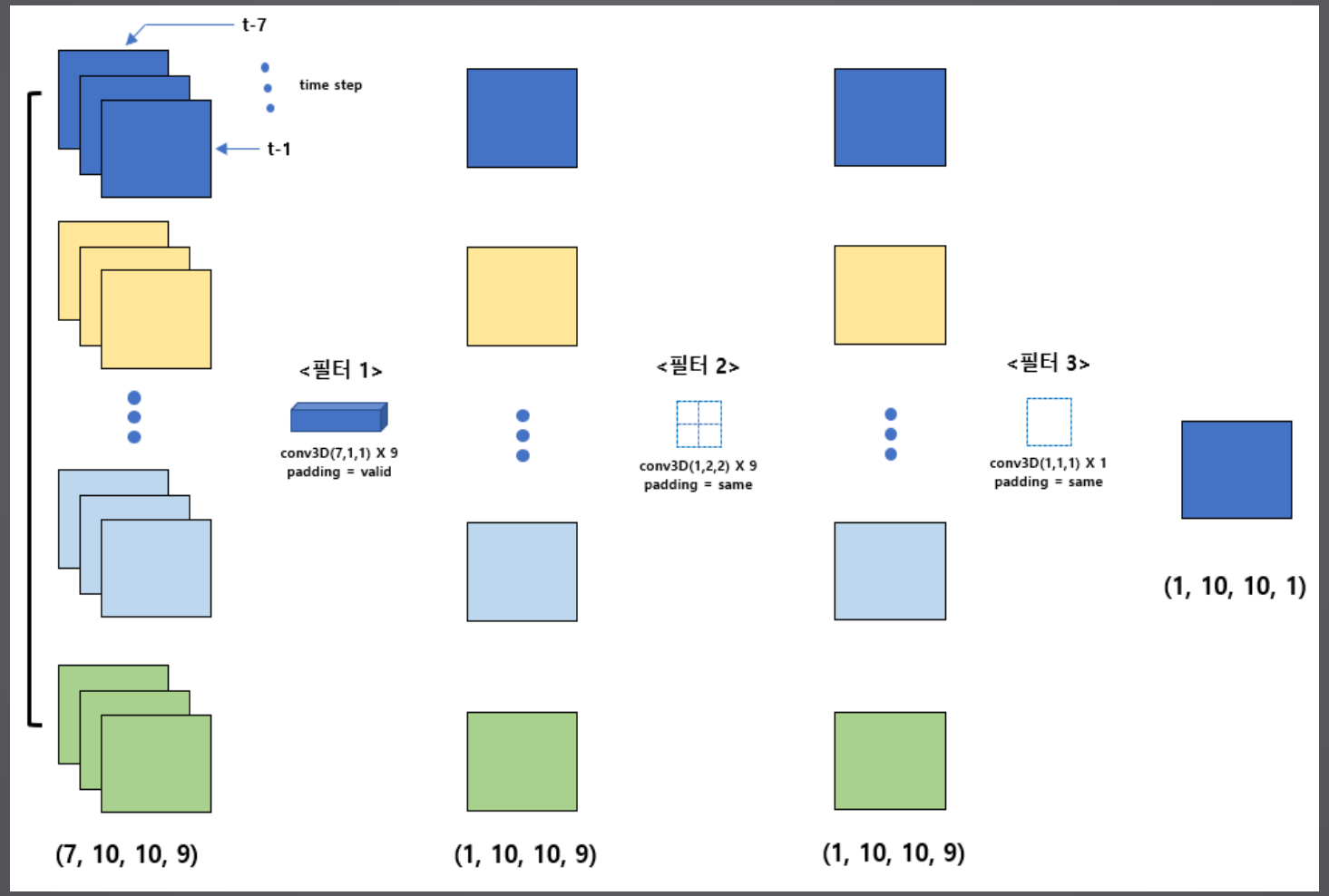
데이터 격자 보간
(IDW: Inverse Distance Weighted)

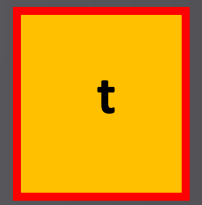
CNN (Convolution Neural Network) 을 활용한 예측

CNN 알고리듬



$x_1 \sim x_9$:
SO2, CO, O3, NO2,
Temp,
Precipitation,
Wind_Speed,
Wind_Direction,
PM10

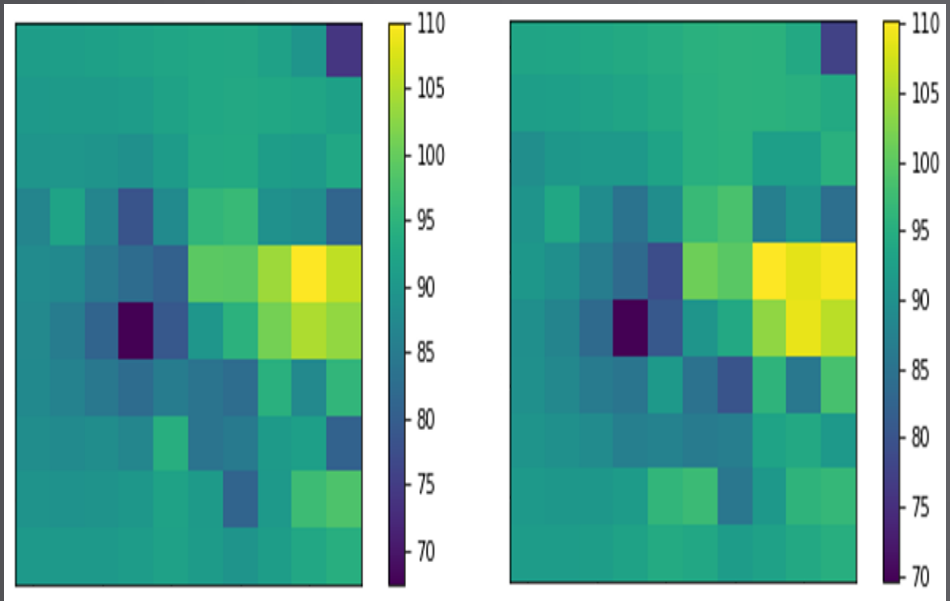




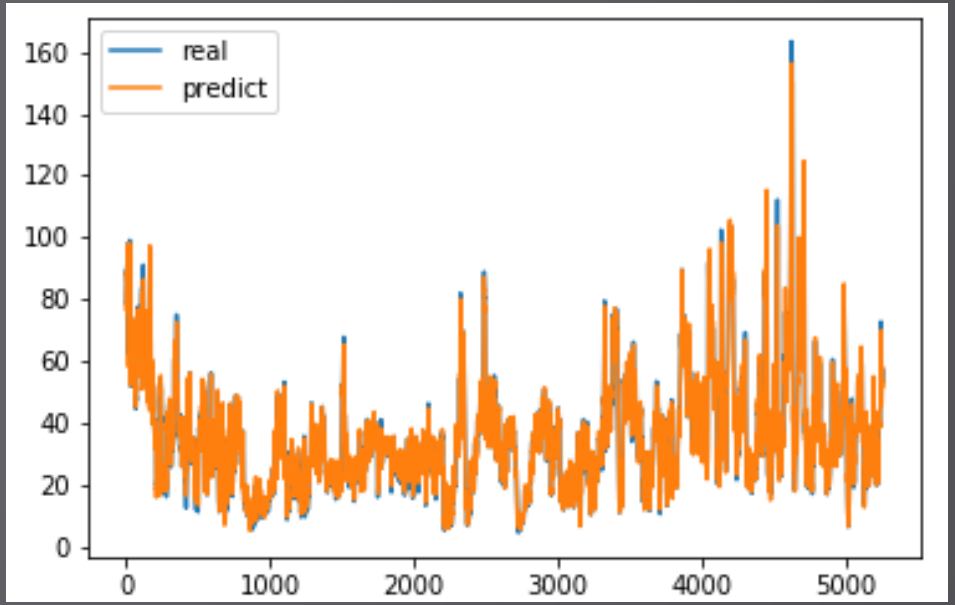
Y:
PM10

미세먼지오염도 예측 결과 (Arc3: 7시간 전 모든 변수)

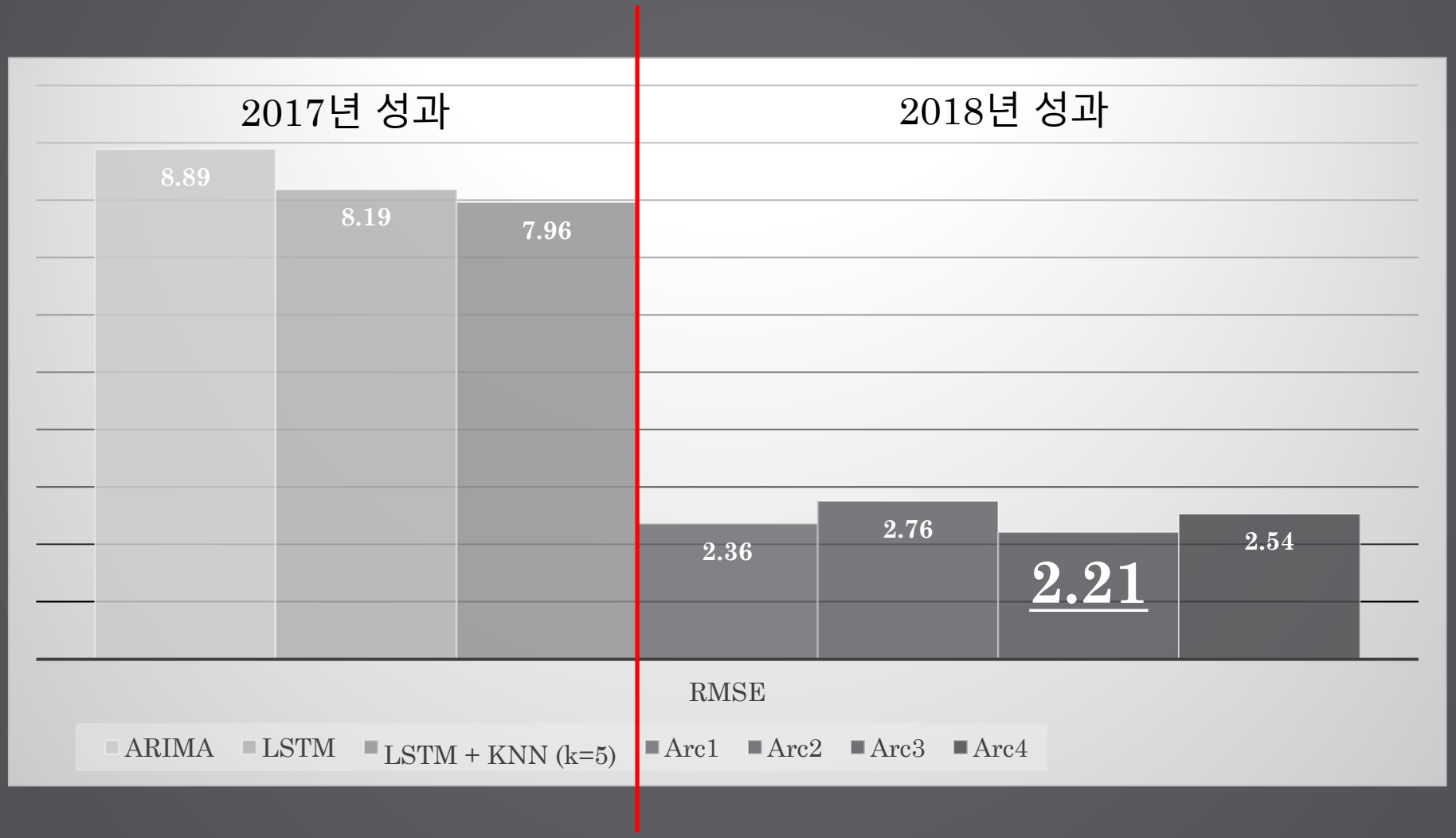
예측 이미지(좌측) vs. 실측 이미지(우측)



미세먼지 농도 예측-실측치 (4,4 지역)



미세먼지 예측 평균제곱근오차(RMSE)



(2) 데이터 기반 한강 수질 예측 [홍한움]

- 인공지능 및 공간통계모형을 이용한 데이터 기반 수질 예측 알고리듬 개발
- 연구내용: 측정소 및 인근지역 수질, 수위, 기상자료에 RNN 적용 부영양화 정도 예측
  - 분석 대상: 2013-17 수도권 수질 측정 지역 중 주 단위 측정 자료 380건 이상 9개 측정소 선택
    - 수질 일반측정망 자료(물환경정보시스템), 기상자료(기상자료개방포털) 수집
    - 수위 자료(한강홍수통제소) 확보 진행 중 : 기관 간 협의 필요
  - 인근지역 정보를 시차를 두고 반영하여 예측 정확도 제고
    - 과거 정보 반영: RNN(Recurrent Neural Network)
    - Long Memory 특성 반영: GRU(Gated Recurrent Unit)
  - 전처리 : 결측치 보간 및 수질 측정소-기후측정소간 거리 차이 반영
- 진행상황: 데이터 수집 및 전처리 작업 진행
  - 수질 및 기상자료 전처리 완료 상태

분석 대상 관측소

◆ 대상 관측소

– 9개 한강 수질측정 주요지점

– 관측 위치별 자료 수

관측 위치	자료 갯수
노량진	403
섬강4-1	403
경안천5	400
팔당댐	398
강천	390
이포	390
가양	389
경안천5A	384
강상	381
삼봉리	381

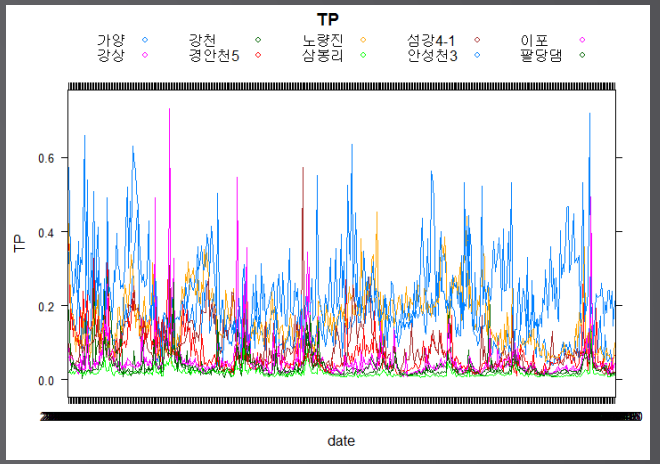
주요 수질 관측지점: 왼쪽부터 순서대로 가양, 노량진, 팔당댐, 경안천5, 삼봉리, 강상, 이포, 강천, 섬강4-1 (출처: 물환경정보시스템 메인페이지)
http://water.nier.go.kr/main/mainContent.do

◆ 시간

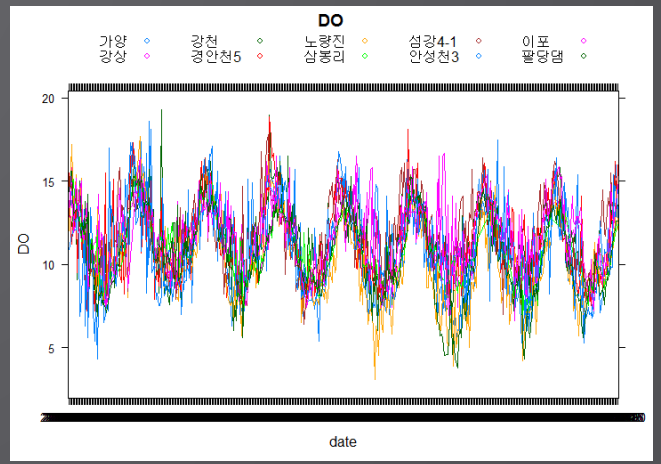
– 2010/02/01 – 2017/12/31 주별 자료

추정 대상 변수 및 독립변수

- 추정 대상 변수: 부영양화지수 기준 구분 부영양화 정도
  - 부영양화 지수: TSI_{KO} = COD, 클로로필-a, 총인의 가중 평균
  - 부영양화 정도
    - 빈영양 ($TSI_{KO} \leq 30$)
    - 중영양 ($30 < TSI_{KO} \leq 50$)
    - 부영양 ($50 < TSI_{KO} \leq 70$)
    - 과영양 ($TSI_{KO} \geq 70$)



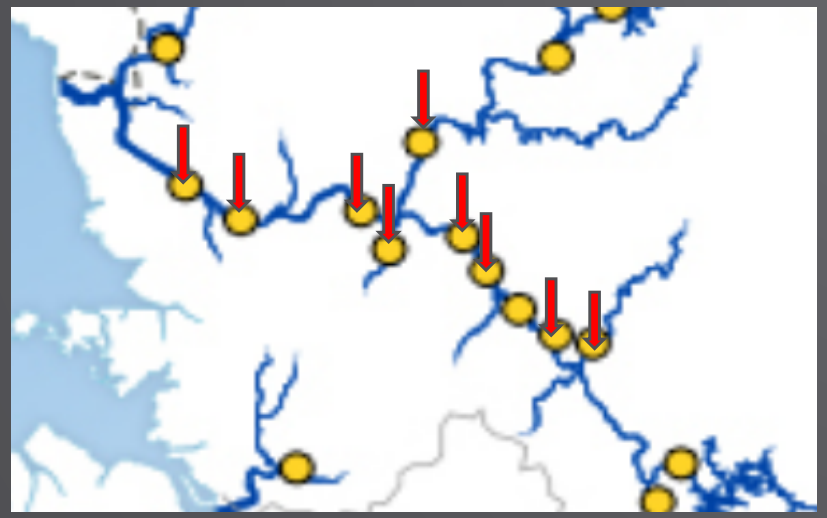
- 독립변수 : 수질, 기상, 유량 자료
  - 수질 : COD, 클로로필-a, 총인 제외 수질자료 및 상류지점 부영양화지수
  - 기상 : 강우량, 습도, 해면기압
  - 유량 : 수위, 유량
  - 각 출력변수, 독립변수의 lag 변수



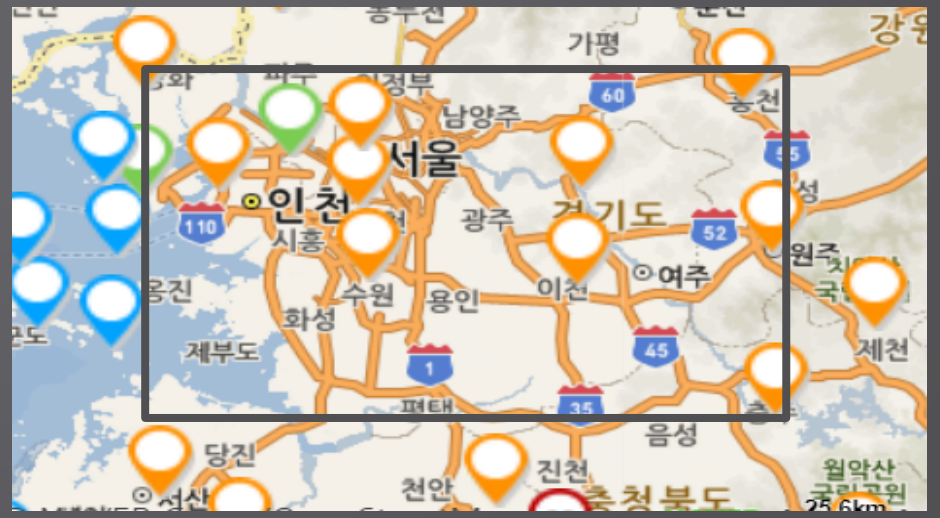
자료 전처리 : 수질 관측지점 – 기상 관측지점 거리차 반영

- 수질 관측 지점 주변 기상 관측지점 자료를 거리 역순 가중 평균
  - 관악, 서울, 인천, 수원, 양평, 이천, 원주, 충주, 제천
  - 수위: 수질 관측지점 주변 수위를 사용하거나 기상정보와 같은 방식으로 처리

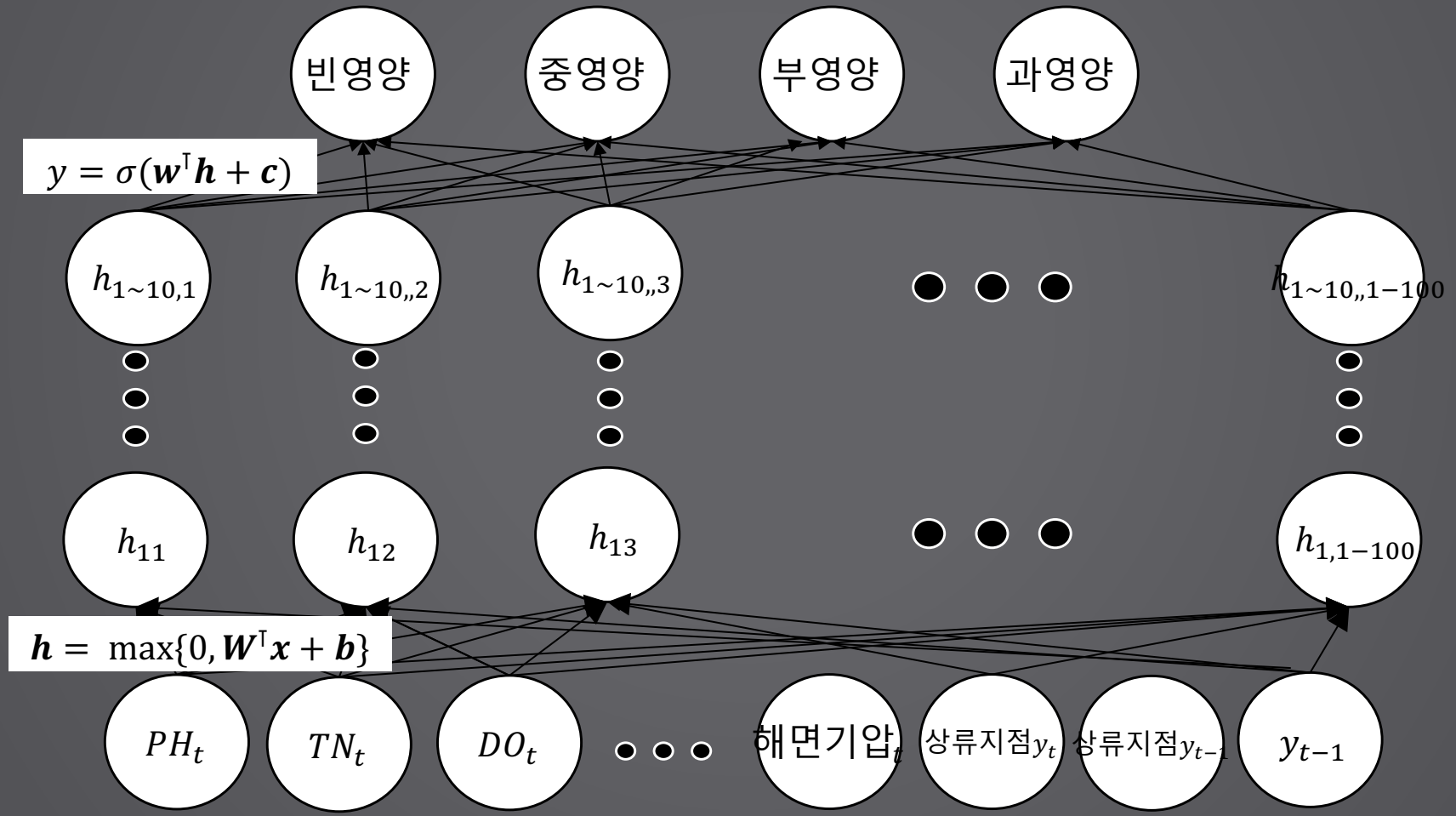
수질 관측지점



기상 관측지점



분석 알고리듬 : RNN and GRU



데이터 기반 한강 수질 예측 연구 향후 계획

◆수위 데이터 확보, 이후 추정 작업 수행

- 최적화 알고리듬으로 Adagrad, RMSProp, ADAM 등의 적응적 학습률 기법 사용
- 하이퍼 매개변수 최적화
- 오류율 비교를 통한 최적 예측모형 선정

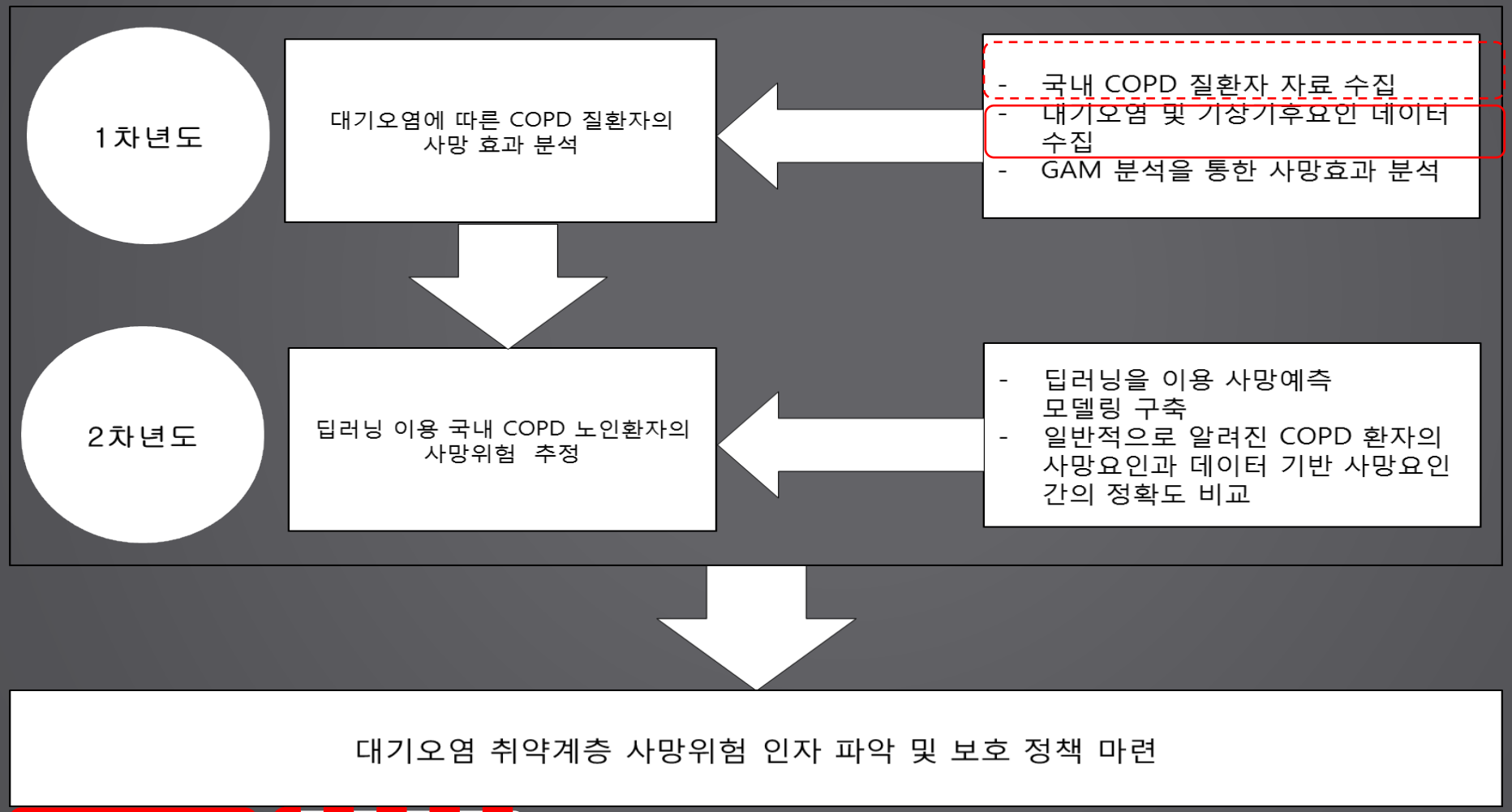
◆ 기존 시공간 예측 모형과 예측 오차 비교

- 시공간 예측 모형에서는 부영양화 지수 TSI_{KO} 를 직접 추정하여 부영양화 정도 예측
- 오류율 비교를 통해 딥러닝 예측 모형과 예측 오차 비교

(3) 딥러닝 이용 노인인구 호흡기 질환 사망 위험 추정 [강선아]

- 만성폐쇄성 폐질환 사망 위험을 딥러닝을 이용하여 추정
  - 연구 대상 :65세 이상 만성폐쇄성폐질환(COPD) 환자
- 연구내용 : 1단계 사망요인 파악 연구 , 2단계 사망확률 추정 연구 2년차로 확대
  - 자료 : 건강보험 맞춤형연구 DB , 2006-2015년 건강보험 코호트 DB version 2.0, 인구, 기후, 대기오염도 및 대기오염물질 배출량 자료를 연계
  - 1차년도 (2018) : 대기오염에 따른 COPD 질환자 사망효과 분석 (맞춤형 연구 DB)
    - GAM 분석을 통해 사망에 영향을 미치는 주요 변인을 발견하고 사망 효과를 분석
  - 2차년도 (2019) : COPD 노인 질환자 사망위험 추정(코호트 DB 2.0)
    - 딥러닝 이용 사망예측 모델 구축
    - 일반적으로 알려진 COPD 환자의 사망요인 .vs 데이터 기반 파악 사망요인 : 사망위험 추정치 정확도 비교
- 연구 진행 상황 : 자료 구축 작업 및 의료 전문가 인터뷰 진행 중
  - 환경 및 기후자료 구축 수행, 건강보험 맞춤형 DB 구축 중
  - 사망요인 선정 과정에서 의료 전문가 인터뷰 진행 (서울삼성병원 호흡기내과 박혜윤박사)

연구 프레임워크


1차년도
대기오염에 따른 COPD 질환자의 사망 효과 분석
- 국내 COPD 질환자 자료 수집
- 대기오염 및 기상기후요인 데이터 수집
- GAM 분석을 통한 사망효과 분석
2차년도
딥러닝 이용 국내 COPD 노인환자의 사망위험 추정
- 딥러닝을 이용 사망예측 모델링 구축
- 일반적으로 알려진 COPD 환자의 사망요인과 데이터 기반 사망요인 간의 정확도 비교
대기오염 취약계층 사망위험 인자 파악 및 보호 정책 마련
완료
진행 중

분석 자료 및 전처리 과정

- 설명변수: 국민건강보험공단 맞춤형 DB 진료 기록, 기상기후 데이터, 대기오염 데이터
- 대기 및 기상 자료: 건강보험 DB 와 연계 목적 시공간 해상도 조정
  - GIS 기법 중 Kriging을 이용하여 점(point)데이터인 측정소 자료를 면 데이터인 시군구자료로 변형
  - 시간단위 대기오염 데이터를 일간 평균을 취하여 일단위 건강보험 DB와 연계
- 건강보험 맞춤형 연구DB : 추출대상을 지정하여 추출 작업 수행 중

설명변수 데이터

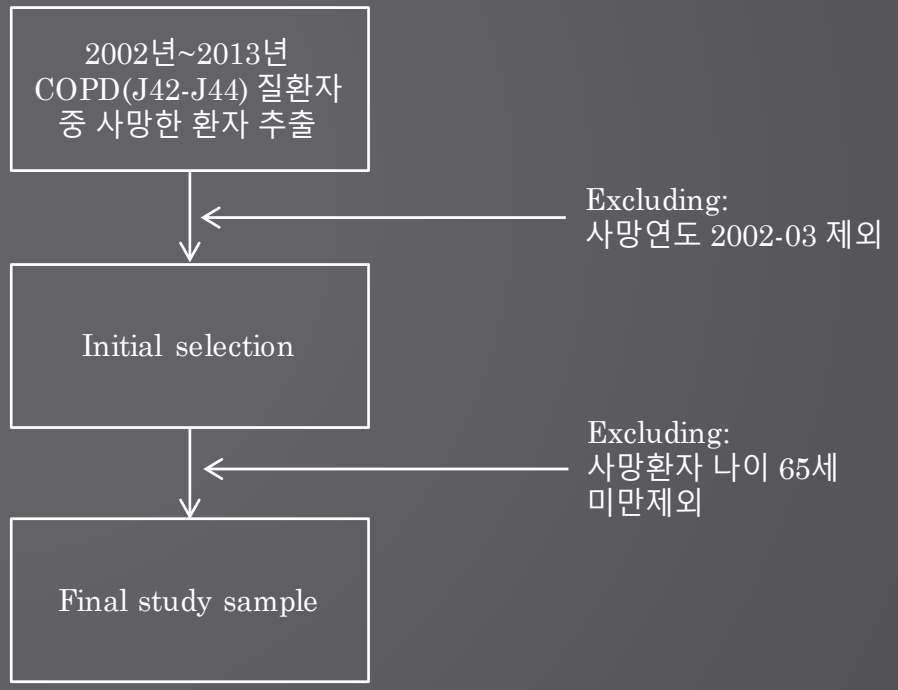
데이터	출처	기간	항목	시간 단위	공간 단위
국민건강보험공단 맞춤형 DB 진료기록	국민건강보험공단	2002-2013	개인특성 데이터/진료기록		시군구
기상기후 데이터	기상개방포털	2002-2013	기온, 습도	일단위	측정소
대기오염 데이터	한국환경공단	2002-2013	PM10, O_3	1시간	측정소

건강보험 맞춤형 DB 신청항목 및 전처리 내용

DB 신청내용

조건	내용
연도	2002년~2013년
상병코드	J42, J43, J44
주상병/부상병	주상병 및 모든 부상병
산정특례 특정기호구분	없음
의과/한방/치과/약국	의과
입원/외래	입원, 외래
행위수가코드	전체자료
약제주성분코드	전체자료
기타	없음

전처리 내용

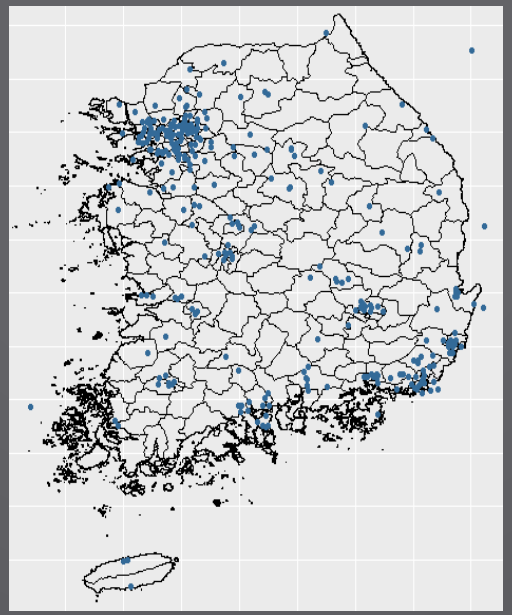


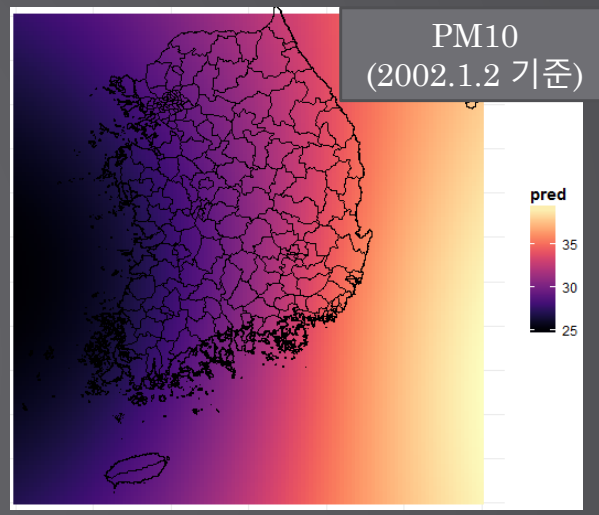
대기 및 기후자료 시공간 해상도 전처리 작업 : 대기오염 데이터

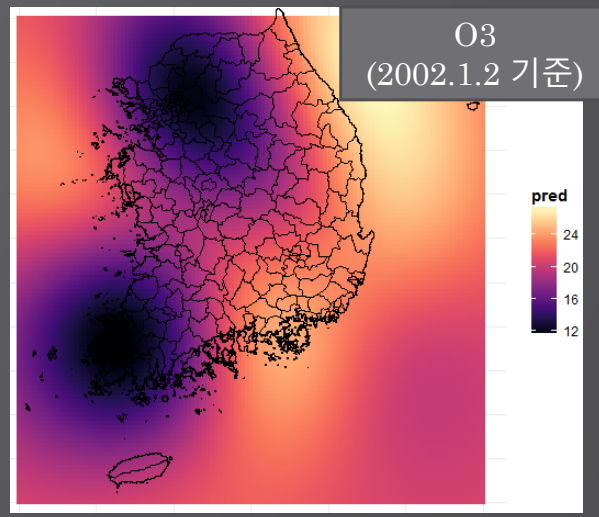
대기오염 데이터

대기오염물질 측정소
CODE
TYPE
* NAME
LOCATION
ADDRESS
LAT
LNG
ALT

대기오염물질 측정 데이터
* DATE_TIME
* NAME
YEAR
MONTH
DAY
HOUR
O3
PM10





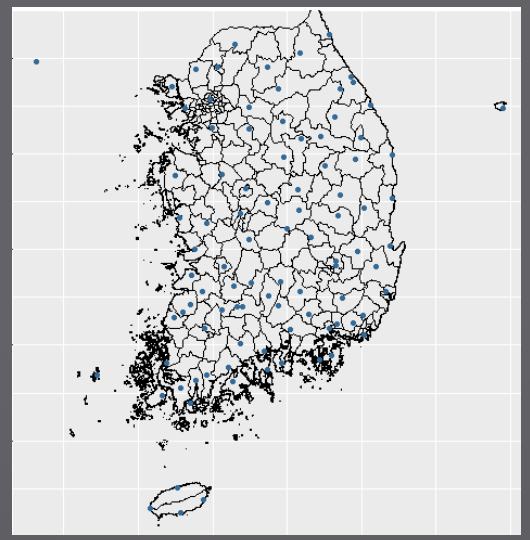


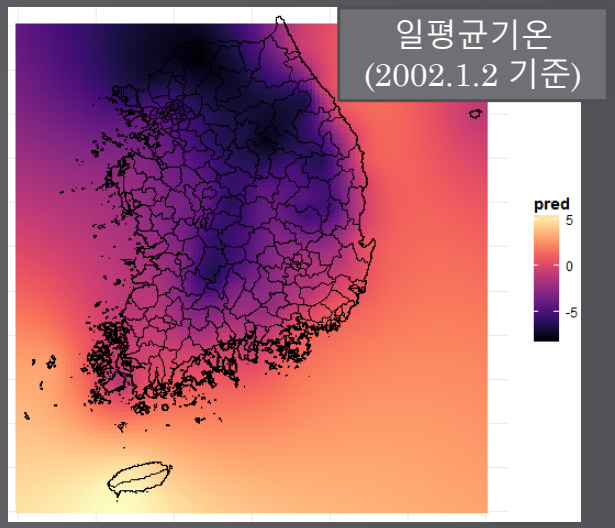
대기 및 기후자료 시공간 해상도 전처리 작업 : 기상기후 데이터

기상기후 데이터

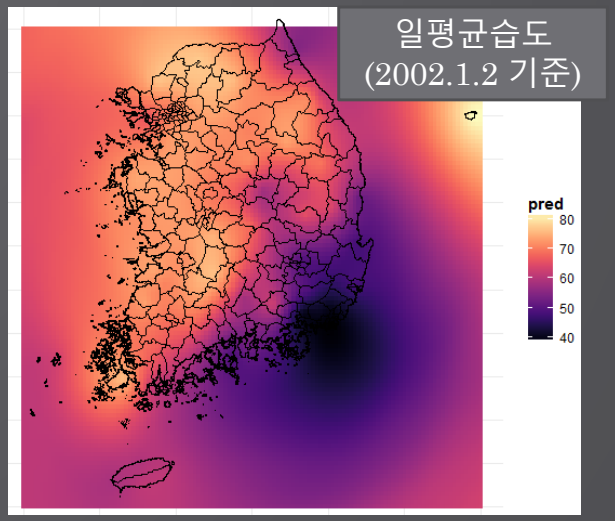
기상기후 측정소
* CODE
TYPE
NAME
LOCATION
ADDRESS
LAT
LNG
ALT

기상기후 데이터
* DATE_TIME
* CODE
YEAR
MONTH
HOUR
TEMP
Humidity





일평균기온
(2002.1.2 기준)



일평균습도
(2002.1.2 기준)

COPD 사망영향 분석 향후 계획 : 변수 추출, 사망효과 분석, 결과 해석

- 설명변수 조작(대기오염 데이터)-코딩화
  - 노출 기간(단기: leg0, leg1, leg2, average/ 장기: 1년, 5년 평균(사망일 기준))
  - 노출 정도: the daily maximum concentration of air pollution
  - 중증도 : 약재 주성분 Code의 복용 기간을 이용해서 중증도 변수 추출 예정
- 데이터 기반 노인 COPD 질환자의 사망효과 분석 수행(GAM 이용)
  - HEAT package 이용(임연희, 2015)
- 결과 해석
  - 전문가 집단 자문

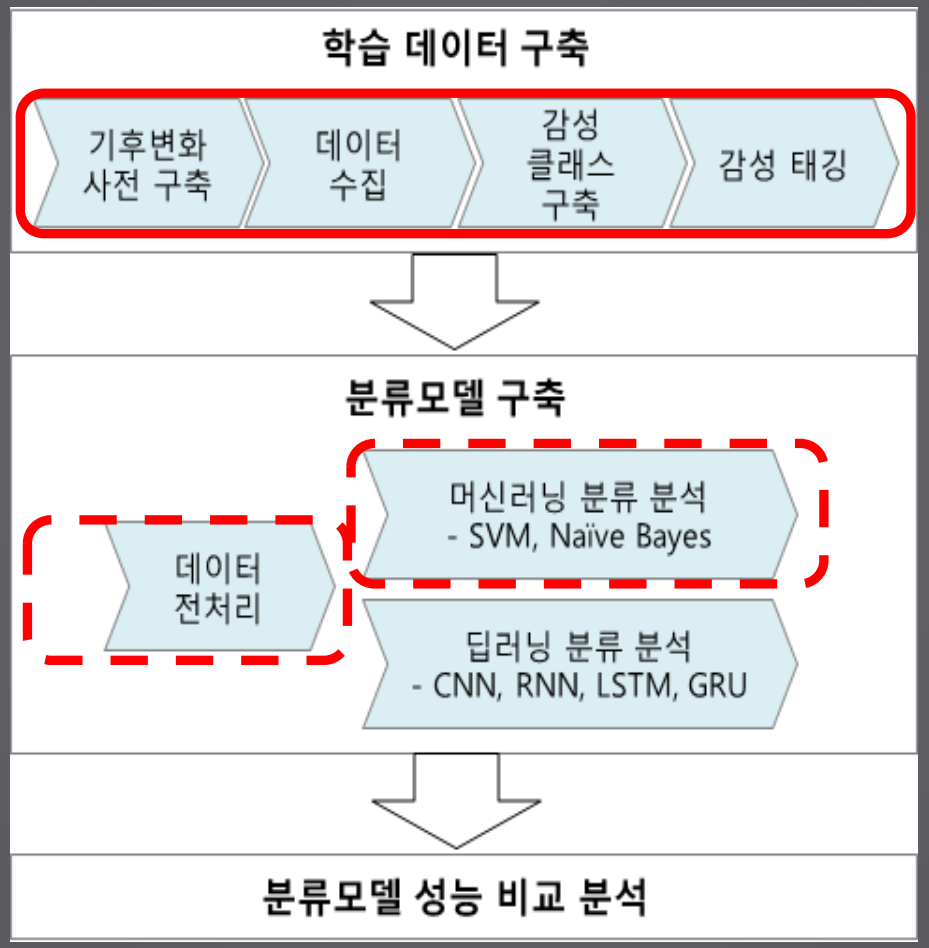
(4) 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화 중심으로 [김도연]

- 기후변화 주제의 SNS 및 주요 포털 댓글 데이터 기반 감성분류 알고리듬 개발
  - 분석 대상 : Facebook, Instagram, Twitter, Naver 뉴스 댓글 4개 SNS 텍스트 자료
  - 텍스트 데이터의 감성을 8개 감성으로 분류
    - 8개 감성 Category : Robert Plutchik의 Wheel of Emotions 활용

- 연구 내용: 기후변화 사전 구축, 감성 분류 학습 데이터 구축, 감성분류 알고리듬 개발
  - 기후변화 사전 : 기후변화에 따른 현상을 4개의 범주(온도, 강수, 토지, 해양) 분류 후 구축
    - 환경관련 문서에 워드 임베딩 방법(LDA, Word2Vec) 적용 후보군 추출
    - 전문가(최희선,명수정) 및 SNS 이용자 의견 반영
  - 감성분류 학습데이터 : 약 5만 건 단문 데이터에 감성을 수작업으로 파악 (4인 1개월 full time 시간 투입)
    - 기후변화 사전 기준 5만건을 수집하여 1만 건의 감성 클래스 파악
  - 감성분류 알고리즘 : 다양한 기계학습 기반 분류 알고리듬 성능 평가를 통해 선정 예정
    - SVM, CNN, RNN, LSTM, GRU, Ensemble

- 연구 성과 : 기후변화 사전 및 감성분류 학습데이터 구축을 완료하여 시험적 분석 실시
  - 감성분류를 긍정, 부정, 중립으로 간소화하고 SVM, Naive Bayes를 적용하여 46.1%~51.2% 정확도 도출

연구 흐름도 및 진행 상황



완료

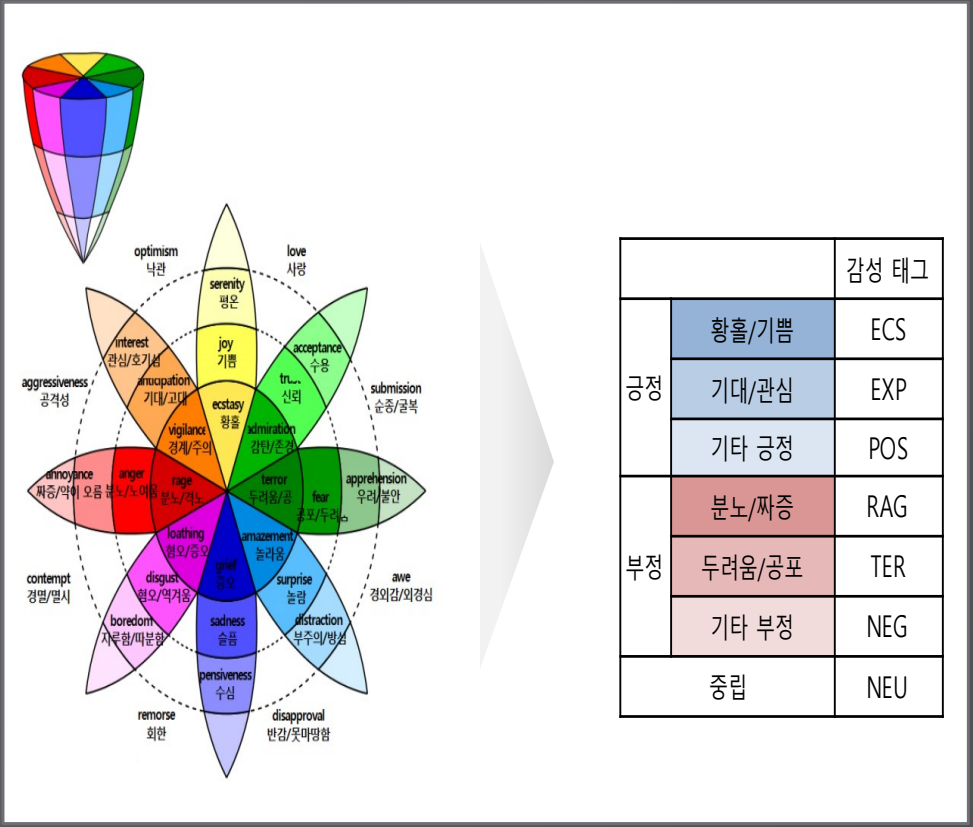
진행 중

기후변화 사전 및 감성클래스

기후변화 사전

기후변화에 따른 '자연재해'					
구분	no.	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우		풍수해
	10	이상기온	홍수		해랑
	11	이상저온	황사비		해수면
	12	폭염			해일
	13	한파			
	14	혹서			
	15	혹한			
비전문가	16	짱춥	눈난리	갈라진땅	괴물파도
	17	짱덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	
	19	졸덥	물난리	찢어진땅	
	20	넙덥	비폭탄	흔들리는땅	
	21	넘춥	흙비		
	22	너무춥	흙탕물비		
	23	너무덥			
	24	개춥			
	25	개덥			

감성 클래스



		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	기타 긍정	POS
부정	분노/짜증	RAG
	두려움/공포	TER
	기타 부정	NEG
중립		NEU

SNS 학습 데이터 구축

1. 기후변화에 따른 현상 사전 구축

기후변화에 따른 '현상' 키워드					
구분	no.	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우		풍수해
	10	이상기온	홍수		해랑
	11	이상저온	황사비		해수면
	12	폭염			해일
	13	한파			
	14	혹서			
	15	혹한			
비전문가	16	짱춥	눈난리	갈라진땅	괴물파도
	17	짱덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	
	19	졸덥	물난리	찢어진땅	
	20	넙덥	비폭탄	흔들리는땅	
	21	넘춥	홈비		
	22	너무춥	흙탕물비		
	23	너무덥			
	24	개춥			
	25	개덥			

2. 감성분류 기준 구축

		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	기타 긍정	POS
부정	분노/짜증	RAG
	두려움/공포	TER
	기타 부정	NEG
중립		NEU

3. 학습데이터 구축 / 4. 크로스체크: 감정 태그 확정

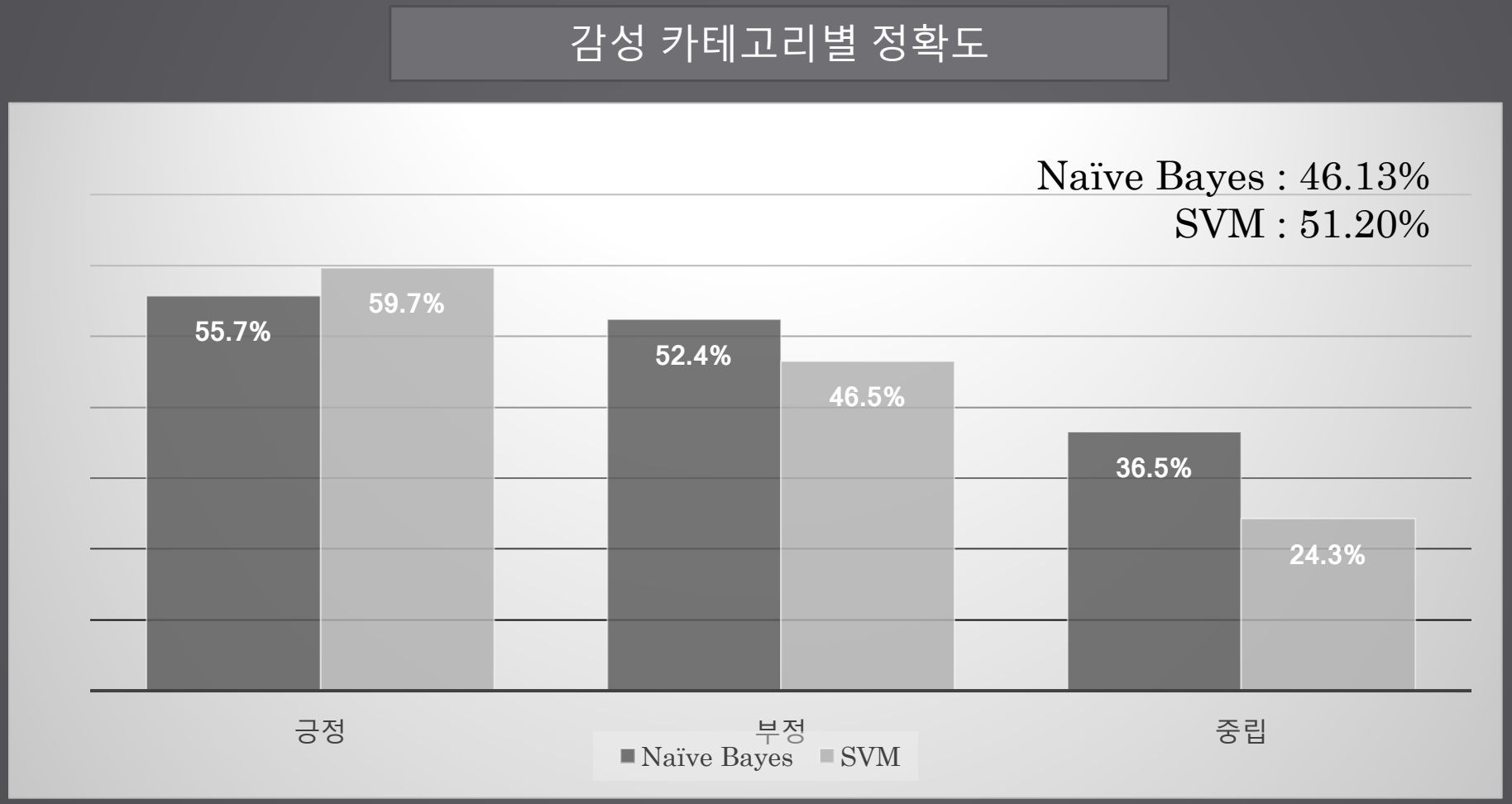
					cross-check					
No	Chanel	Keyword	Keyword1	Content	Tag	tag1	tag2	tag3	tag4	final Tag
1	뉴스 댓글	온도	강추위	난 추울때 겨울 냄새가 넘 좋아	ECS	ECS	POS	ECS	ECS	ECS
2	뉴스 댓글	온도	무더위	밖에서 일하시는 이 세상에 아버지 어머니들 힘내세요!!!	POS	POS	POS	ECS	ECS	POS
3	뉴스 댓글	온도	무더위	더 뜨거워지면 좋겠다	EXP	EXP	EXP	EXP	EXP	EXP
4	뉴스 댓글	온도	강추위	동요에 찬바람 불어도 괜찮아요 가사있죠? 이 날씨에 괜찮은 사람 없을듯	RAG	RAG	RAG	TER	TER	RAG
5	뉴스 댓글	온도	무더위	하루하루가 더위때문에 넘 힘드네요..내일은 얼마나 또더울까..이런생각에..	TER	TER	TER	TER	TER	TER
6	뉴스 댓글	온도	너무춥	의정부 -13 너무 춥다..ㅠㅠ	NEG	NEG	NEG	NEG	NEG	NEG
7	뉴스 댓글	온도	강추위	백수라 춥든말든.....	NEU	NEU	NEG	NEG	NEG	NEG
8	뉴스 댓글	온도	강추위	오늘 모스크바 -11 대관령 -21.7	NEU	NEU	NEU	NEU	NEU	NEU
9	Facebook	온도	열대야	샤워하고먹는아이스아메리카노가 진리임 열대야극복완료 단순해ㅋㅋㅋ	ECS	ECS	NEU	ECS	POS	ECS
10	Facebook	온도	무더위	밖에 내다 놓은 화초들이 무더위를 견디다니..허허..튼튼한것들	POS	POS	POS	NEU	NEU	POS
11	Facebook	온도	열대야	열대야가 이번주가 끝이래요!!!! 밖으로 놀러가즈아~~~~	EXP	EXP	EXP	POS	EXP	EXP
12	Facebook	온도	결빙	이젠 진심 눈이 싫어 지네요	RAG	RAG	TER	TER	RAG	RAG
13	Facebook	온도	결빙	역대급추위.....ㄷㄷㄷ올 첫 영산강 결빙현상 발생 11	TER	TER	TER	TER	TER	TER
14	Facebook	온도	무더위	응? 무더위 본격적 시작이라는데 ,, 그리고 10월까지 덥다는데 실화냐 ㅠㅠ	NEG	NEG	TER	TER	TER	TER
15	Facebook	온도	강추위	이거봐 매일매일이 강추위임	NEU	NEU	NEU	NEG	NEU	NEU

SNS Data 전처리

전처리 단계	전처리 내용
1) 이모지 한글로 변환 2) 이모티콘(특수문자) 전처리 3) ID 삭제 4) 정규화 : 함축어, 신조어, 은어 등	- SNS 특성을 반영한 전처리 단계 - 이모지 전처리: 약 1,200개 이모지 한글로 변환 예) 😍 😱
5) Document Term Matrix(DTM) 생성 6) 말뭉치(Corpus) 생성 : 단어길이 최소 2글자 이상 7) Sparse Terms 삭제 : 출현빈도가 매우 낮은 단어 삭제	- DTM:
8) 데이터 프레임(Data frame) 형태로 변환	- 기계학습 분석에 적합한 형태로 변환

	Term1	Term2	…	TermM
Doc1	2	1	…	0
Doc2	0	4	…	2
…				
DocN	3	1	…	1

3개 감정 class 분류 성과 : 정확도 제고 작업 필요



감성분류기 연구 향후 계획

◆ SNS 데이터 전처리
- 추가적인 정규화 전처리 : 신조어, 함축어, 은어, 두음 문자 , 띄어쓰기 및 맞춤법 오류 등

◆ 채널과 현상을 고려한 분류기 구축
- 4개 채널과 4개 현상 데이터의 성격이 다르므로 각 부분집합 내에서 감성분석을 시도

◆ 딥러닝 기반 분류기 구축
- CNN, RNN, LSTM, GRU 등의 기계학습 기반 분류 알고리듬 구축
- 분류 정확도 성능 평가를 통해 최종 분류모델 선정

채널 및 현상 구분

채널 / 현상	온도	강수	토지	해양
인스타그램				
페이스북				
트위터				
뉴스댓글				

(5) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향 [김진형]

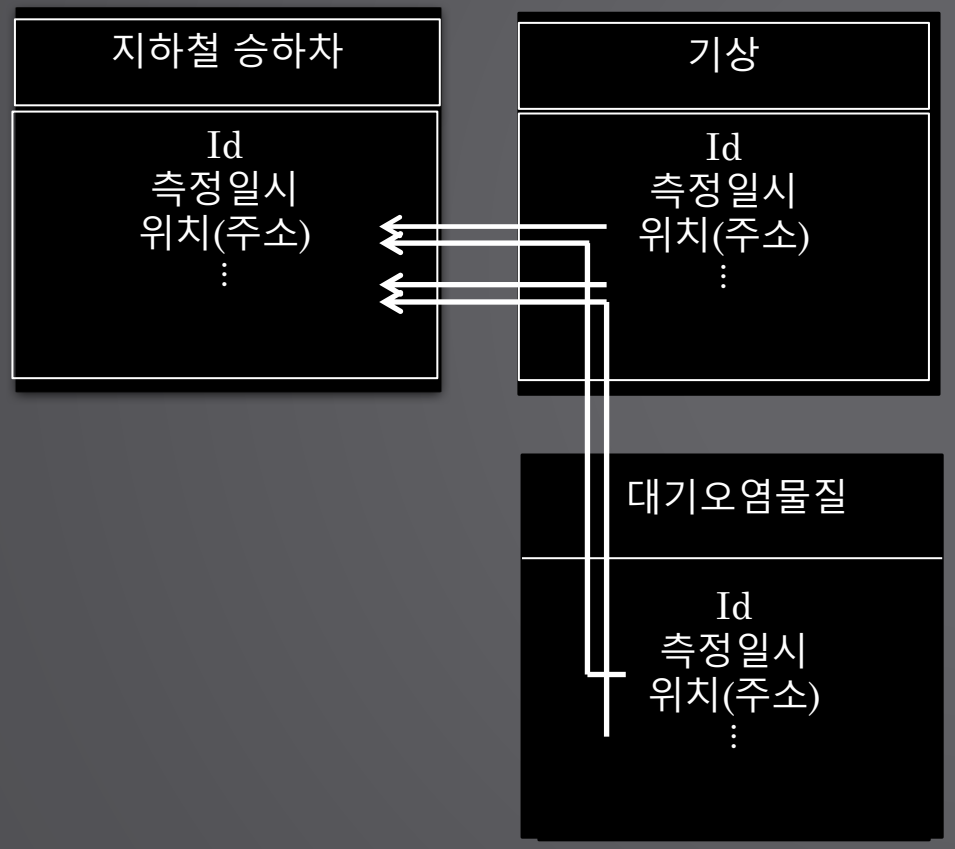
- 미세먼지 농도 및 예보가 사회적 행위(대중교통 이용)에 미치는 영향 파악
  - 미세먼지로 인한 외부활동의 감소 및 대중교통 수요 증가 현상 등을 정량적으로 파악
- 연구 내용: 미세먼지 농도 및 예보가 지하철 이용에 미치는 영향 분석 및 지하철 이용
  - 자료: 서울시 지하철 승하차 정보(서울 열린데이터 광장, 공공데이터 포털) , 기상기후 데이터(기상자료개방포털), 미세먼지 데이터(에어코리아)
  - 미세먼지 농도 변화에 따른 지하철 이용의 변화를 의사결정 나무 방법론을 적용하여 분석
    - 방법론 후보군: 회귀분석, SVM(Supporting Vector Mechanism), Boosted Tree
  - 실시간으로 변화하는 자료의 특성을 반영하여 추정 결과를 상시적으로 갱신하는 발신 방식을 고민
- 진행 상황: 대기, 기상, 교통량 자료 전처리가 완료 및 의사결정 나무 방법론 적용 시범 분석
  - Boosted Decision Tree algorithm: 서울역 승하차 인원 Peak 값을 실측치와 근사하게 추정

분석에 사용한 데이터

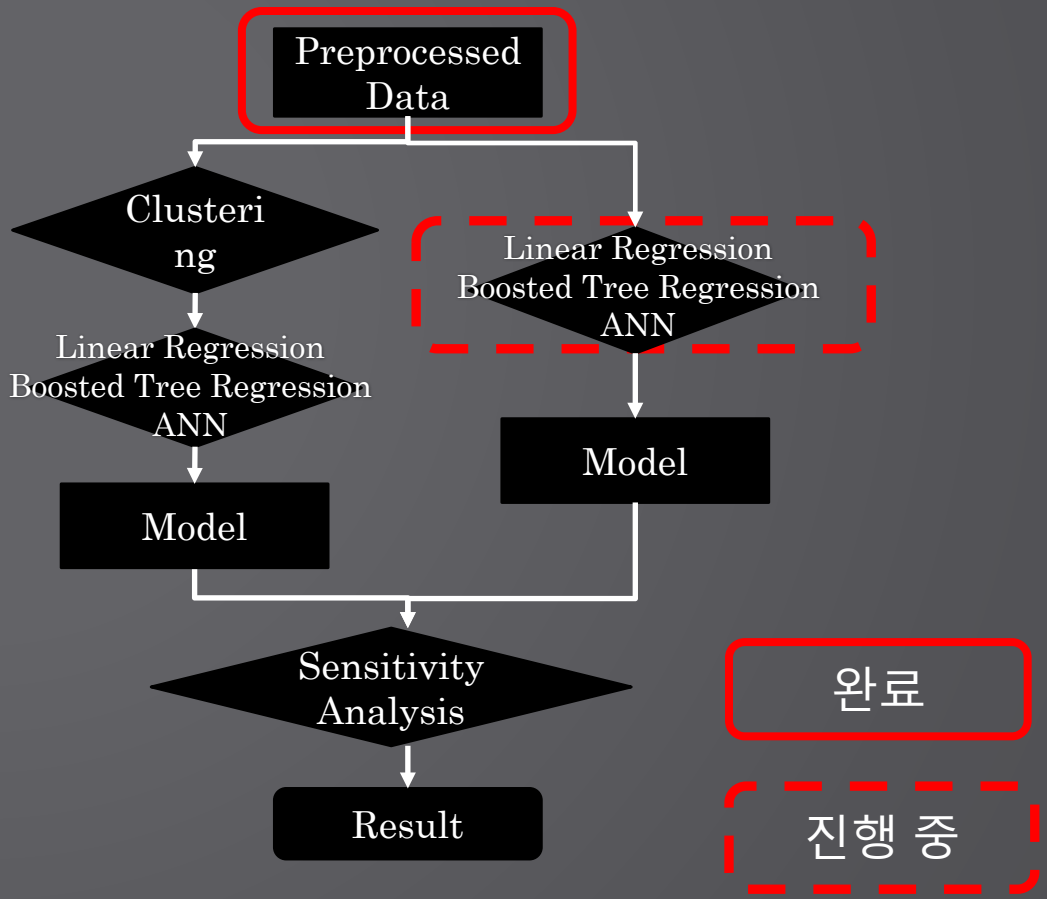
데이터명	출처	속성	비고
서울시 지하철 승하차 인구	서울교통공사 (공공데이터포털)	Id, 측정일시, 위치(주소), 승/하차, 인원	서울시 227개 역
기상 데이터	기상자료개방포털	Id, 측정일시, 위치(주소), 기온, 강수량 풍속, 습도, 적설	
대기오염물질 농도 데이터	에어코리아	Id, 측정일시, 위치(주소), PM10, SO2, CO, O3, NO2	서울시 46개 측정소
미세먼지 경보 데이터	에어코리아	일시, 지역, 주의보/경보	서울지역 대상
서울시 교통량 데이터	서울시 교통정보시스템 (TOPIS)	Id, 측정일시, 위치(주소), 유출/유입 통행량	서울시 144개 측정 지점

데이터 전처리 방식 및 분석 알고리듬

데이터 전처리



분석 알고리듬



데이터 전처리 및 시범 분석 결과

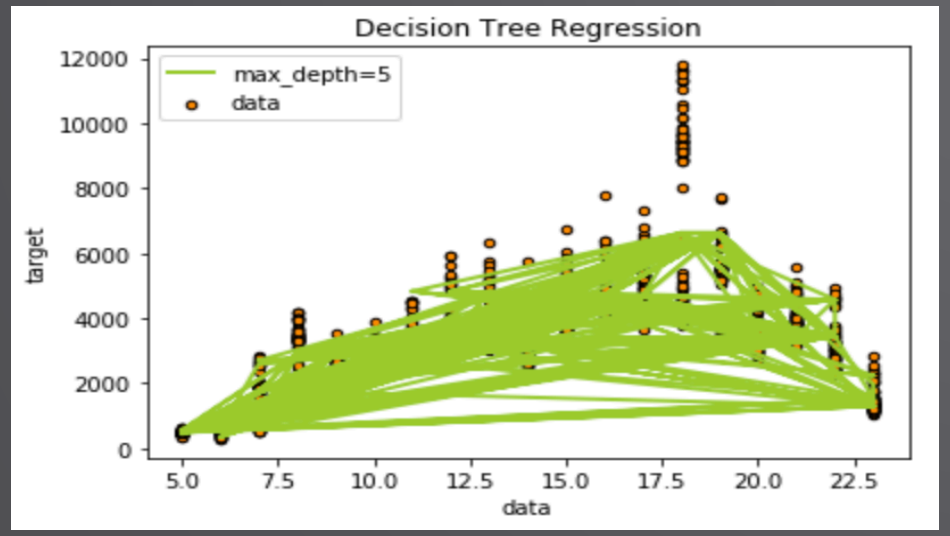
데이터 전처리 결과

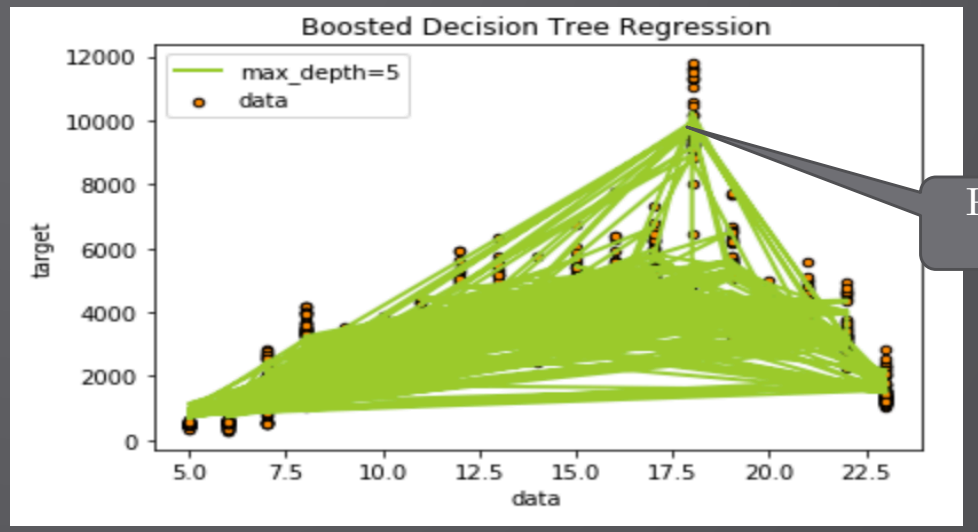
지하철 승하차 인원 | 기상 데이터 | 대기오염물질 데이터

	date	id	name	category	location	time	value	dow	지점	기온 (°C)	...	풍속 (m/s)	습도 (%)	적설 (cm)	시군구	SO2	CO	O3	NO2	PM10	PM25
0	2015-01-01	150	서울역	0	중구	5	441	Thu	108	-9.1	...	5.7	35.0	0.0	용산구	0.0045	0.25	0.0185	0.0080	111.0	19.0
1	2015-01-01	151	시청	0	중구				08	-9.1	...	5.7	35.0	0.0	중구	0.0055	0.40	0.0205	0.0075	97.5	12.0
2	2015-01-01	152	종각	0	종로구	5	898	Thu	108	-9.1	...	5.7	35.0	0.0	종로구	0.0045	0.20	0.0200	0.0055	118.5	5.0

데이터 조인

Decision Tree Regression 테스트 결과 : 서울역 승하차 인원 추정





대중교통 이용 분석 연구 향후 계획

- 예측 정확성 제고
  - 데이터 : 승하차 패턴/지역 기준으로 역 클러스터를 도출하여 동일 클러스터 내 여타 역 승하차 정보를 추정에 반영
  - 알고리듬 : 인공지능 기반 알고리듬을 적용
- 민감도 분석 : 대기오염이 지하철 승하차 인원에 미치는 영향을 정량적으로 파악
  - “ Sample 1% 변화 = 승하차 인원 예측치 … 명 (.. %) 변화 ” 파악

2. 연구 진행 상황 (3)

연구동향 파악 서비스

LDA 토픽 모델링 및 연구동향 분석 서비스

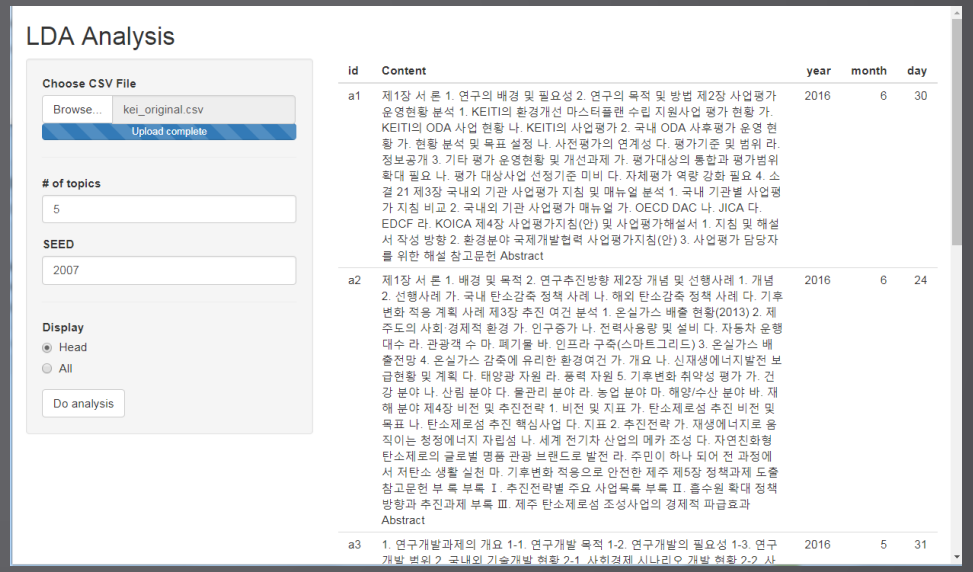
- 2017년 '텍스트마이닝을 이용한 KEI 연구동향 분석' 연구결과 재생
  - 전 연구의 소스코드와 동일 데이터를 활용하여 동일 결과를 플랫폼 위에서 출력
    - 인코딩 문제, 각종 라이브러리 세팅 등 작업 환경 차이로 인한 문제 해결
  - 기존 연구 일반화 및 서비스 구축을 위해 수정이 필요한 부분을 분석
- 동일 포맷 다른 데이터 입력 시 토픽 모델링 결과를 출력할 수 있도록 기존 소스코드 일반화
  - 형태소 분석기 추가, 파라미터 지정 부분 분리, 시각화 관련 부분 수정 등
  - 코드 간소화 : 파일과 인자를 입력 받으면 한번의 실행으로 모든 결과가 도출되게 하는 작업
- 웹서비스 구축: R shiny를 활용한 웹 기반 GUI 인터페이스 구축
  - 원내 서비스를 위한 테스트 단계
- 향후계획 : 출력 내용 조정, 원내 테스트 이후 서비스 공개

연구동향 분석 서비스 GUI

입력 파일 포맷 (csv)

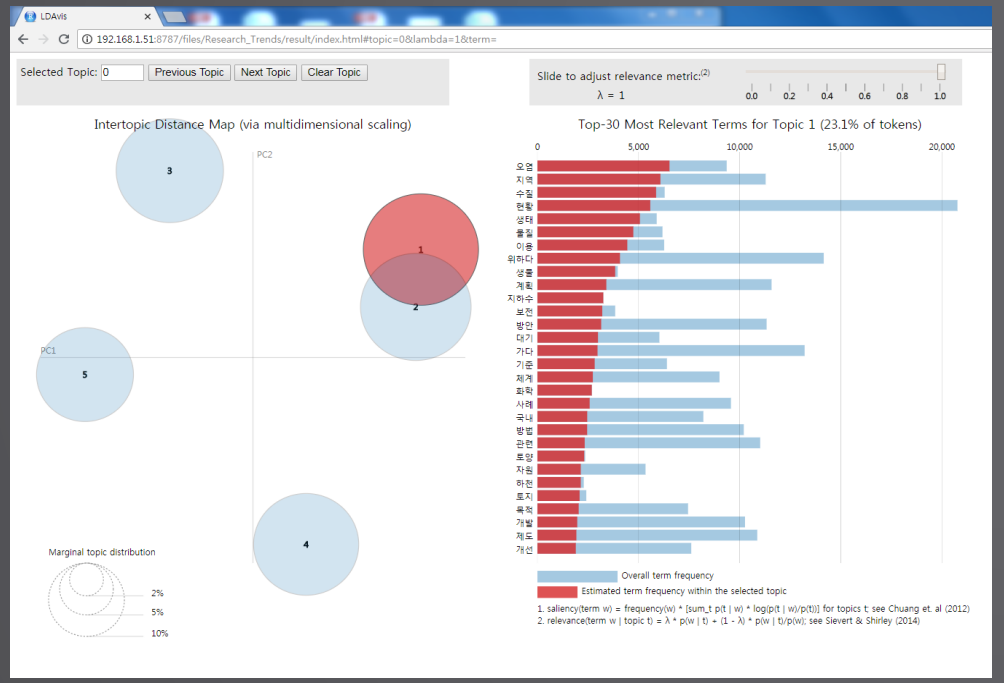
id	Content	year	month	day
a1	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 및 방법 제2장 사업평가 운	2016	6	30
a2	제1장 서 론 1. 배경 및 목적 2. 연구추진방향 제2장 개념 및 선행사례 1. 개념 2.	2016	6	24
a3	1. 연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요성 1-3. 연구개발	2016	5	31
a4	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 3. 연구 내용 및 범위 제2장	2016	4	30
a5	제1장 서 론 1. 연구의 목적 2. 연구의 배경 및 필요성 3. 연구 목표 및 범위 제2장	2016	4	30
a6	제1장 서 론 1. 연구 배경 및 목적 2. 연구 내용 및 방법 제2장 2015년 충남 서북부	2016	3	31
a7	1. 연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요성 1-3. 연구개발	2016	3	31
a8	연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요성 1-3. 연구개발 방	2016	2	28
a9	제1장 서론 1. 연구의 배경 및 필요성 2. 연구방향 3. 연구수행 방법 제2장 제네바	2015	12	31
a10	제1장 서 론 1. 연구의 배경 및 필요성 가. 사물인터넷: 혁명적 핵심기술로서 부상	2016	10	31
a11	제1장 서론 1. 연구의 필요성 및 목적 가. 연구의 배경 및 필요성 나. 연구의 목적	2016	10	31
a12	제1장 서론 1. 연구의 배경 및 목적 2. 연구의 대상 3. 연구의 구조 제2장 이론적 차	2016	10	31
a13	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 범위 제2장 국내 지역사회 환경보	2016	10	31
a14	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 제2장 파리협정의 핵심사항	2016	9	30
a15	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구의 범위 및 방법 3. 선행연구와의 차	2016	9	30
a16	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구 범위 3. 연구 내용 및 수행체계 제2	2016	9	22
a17	제1장 서 론 1. 연구의 배경 및 목적 2. 연구의 접근 및 방법 제2장 국내 "정화곤란	2016	8	31
a18	제 1 장 서 론 1. 과업 배경 및 필요성 2. 과업 내용 가. NBSAP 실행 관련 생물다양	2016	8	30
a19	지속가능발전 개념은 '87년 환경과 개발에 관한 세계위원회(WCED)가 발표한 보고	2016	7	30
a20	제1장 서 론 1. 배경 2. 필요성 및 목적 3. 연구 범위 4. 연구 내용 5. 연구 방법 및	2016	11	22
a21	제1장 서 론 1. 연구의 필요성 및 목적 2. 시스템과 네트워크 언어 제2장 미래 환경	2016	11	6

LDA 토픽 모델링 웹서비스

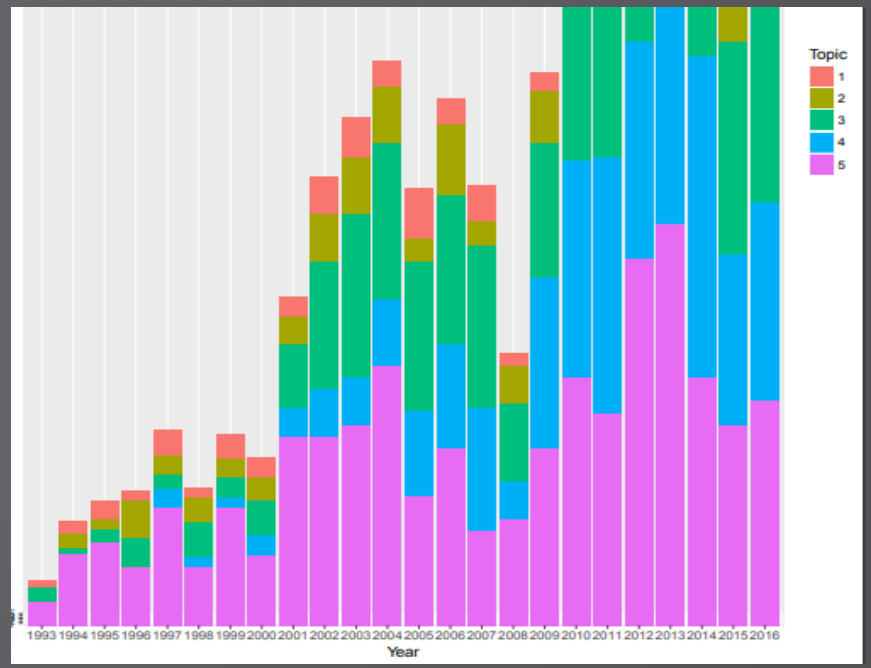


연구동향 분석 서비스 Output

LDA 토픽 모델링 결과



연구동향분석 시각화



3. 향후 계획

연구 진행 상황 요약

◆ 50% 기준으로 2개 세부과제 진도 check 필요: 데이터 수집 과정에서 예정 이상 시간 소요

– 50% 기준 : 데이터 수집 및 전처리 (연구), 분석 플랫폼 구성요소 설치(플랫폼), 서비스 투입-산출 구조(서비스)

장	절	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1. 서론	1) 필요성 및 연구 목적										
	2) 선행연구										
	3) 연구내용 및 방법론										
	4) 본문 내용										
2. 환경 빅데이터 인프라 구축	1) Open Data Map										
	2) 빅데이터 분석 플랫폼										
3. 환경 빅데이터 연구	1) 컨벌루션 신경망(CNN)을 통한 미세먼지 예측									후속	조치
	2) 데이터 기반 한강 수질 예측										
	3) 딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정										
	4) 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화를 중심으로										
	5) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향										
4. 환경 빅데이터 서비스	연구동향 파악 서비스										
5. 결론	연구결과 요약 및 시사점										

연구관리

- 격주 1회 정기 meeting : 세부과제 연구상황 공유
  - 매주 수요일 3시 : 환경 빅데이터 연구 인프라 구축
  - 매주 목요일 오전 10시: 환경 빅데이터 연구
- 월 1회 Progress Seminar 실시 : 연구진 전원 참여 및 외부 전문가 자문
- Working paper 상태의 중간 산출물을 온라인에 게시하여 피드백 기회를 확대
  - 홈페이지 (https://keibigdata.github.io/project.html),
  - GitHub (https://github.com/keibigdata/)

감사합니다