공개SW(Open Source SW)를 중심으로 하는

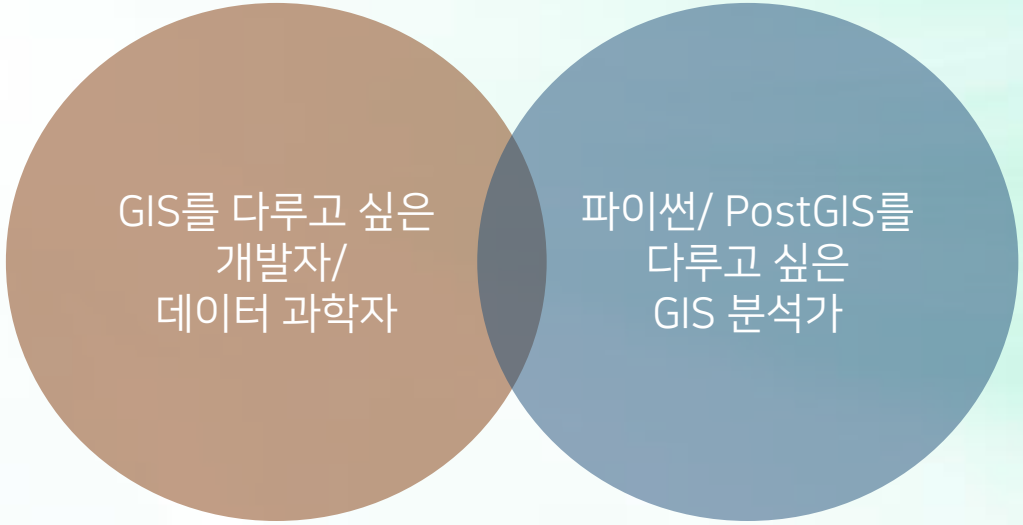
공간정보 빅데이터 분석 및 실습

01. 공간정보

GU GIS UNITED 이동훈 (thlee33@gmail.com)

개요

- **대상**
  - 공간 빅데이터를 다루고 싶은 개발자
  - 파이썬, PostGIS 등 다양한 공개 SW를 다루고 싶은 GIS 분석가
- **주요 범위 및 내용**
  - 공간빅데이터이지만 Hadoop 또는 NoSQL을 직접 다루지는 않고
    공공데이터포털과 같이 주소/좌표가 포함된 CSV 형태로 가공된 데이터에서 시작
  - 공간정보(주소/좌표)가 포함된 정보의 전처리, 가공 및 분석 과정을
    오픈소스 툴(Python, QGIS, PostGIS) 위주로 진행

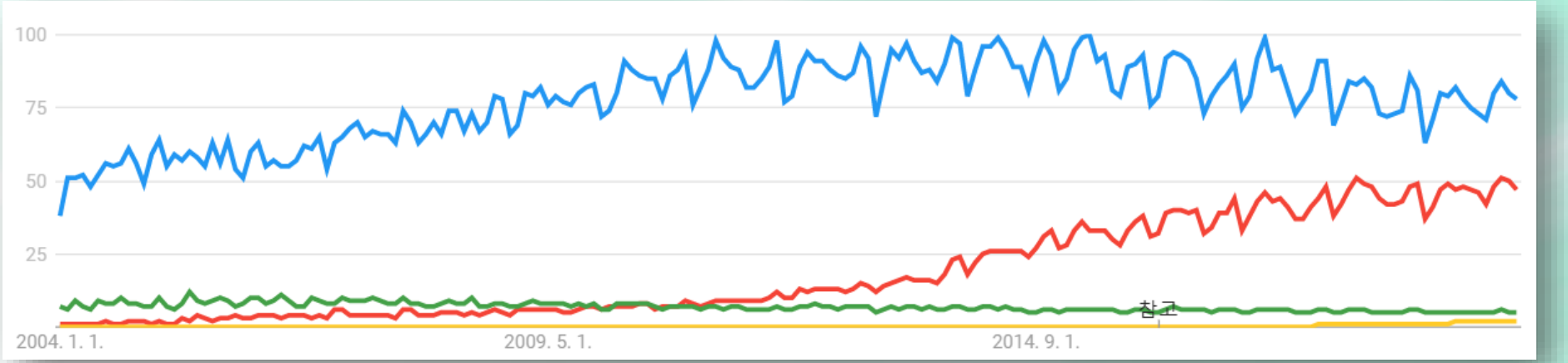


GIS 부문 트렌드 분석

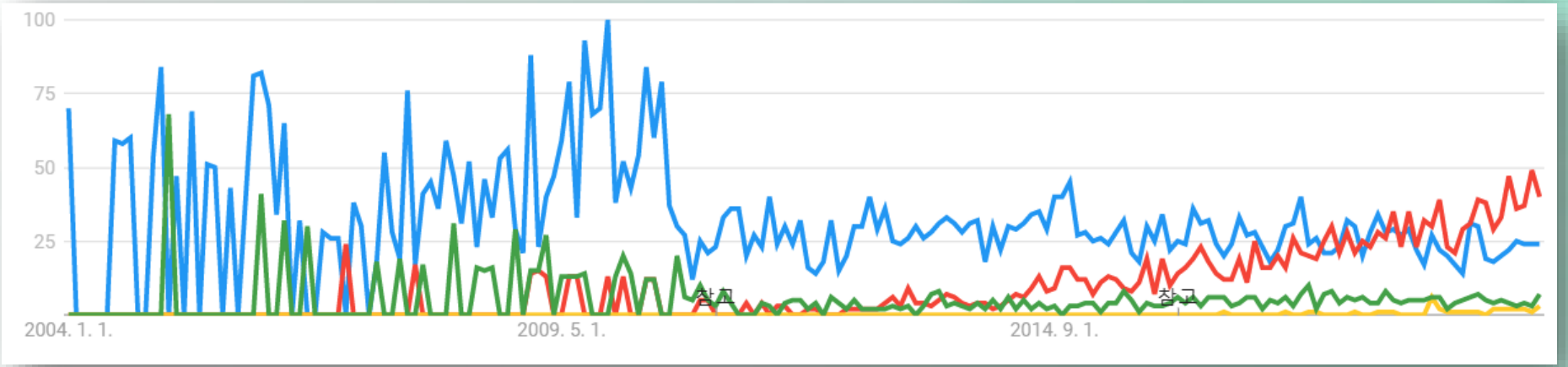
- 구글 트렌드(trends.google.co.kr). 2004년 이후 ~

ArcGIS QGIS Geopandas PostGIS

전 세계

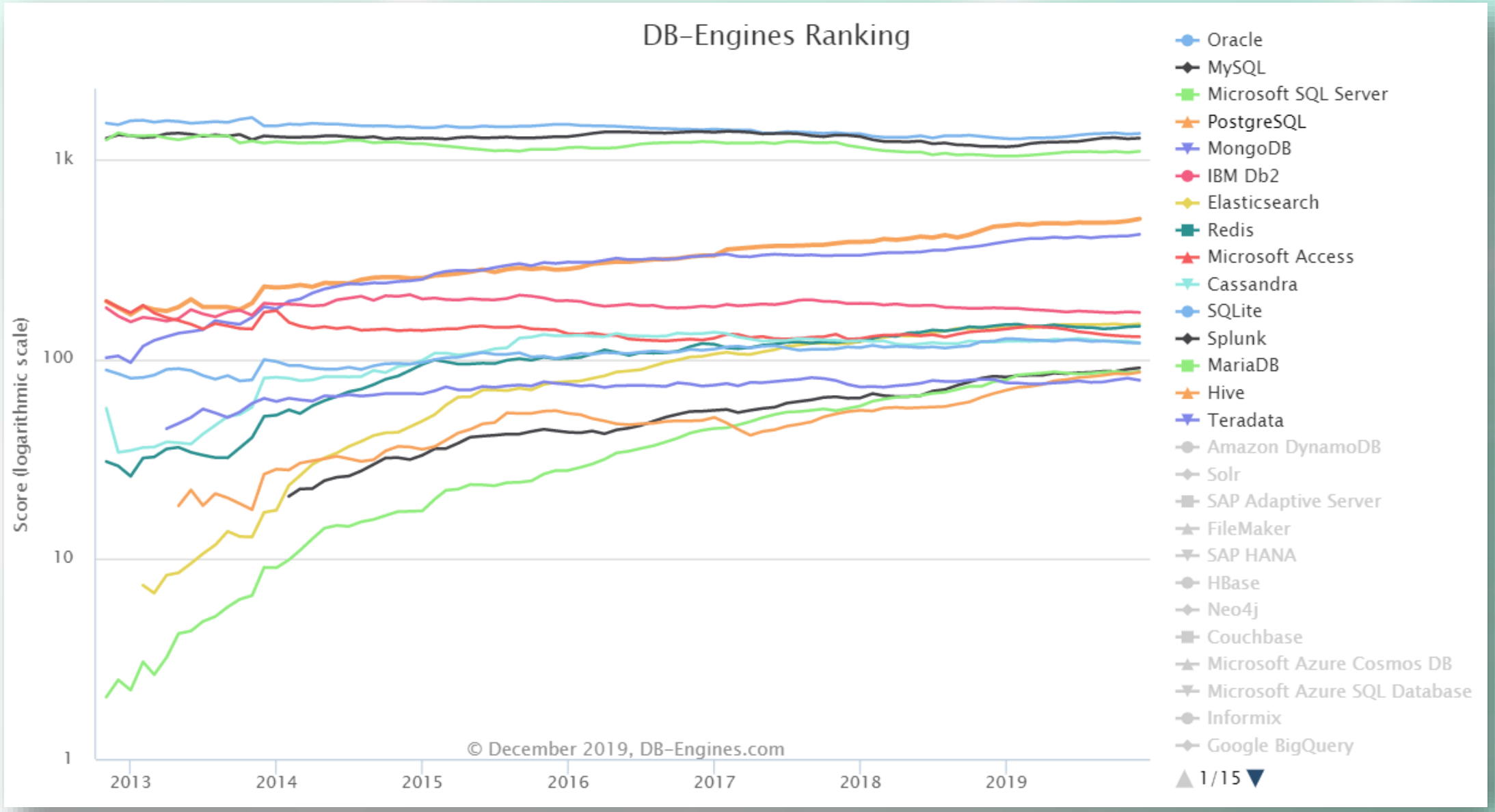


대한민국



DBMS Ranking Trends

https://db-engines.com/en/ranking_trend

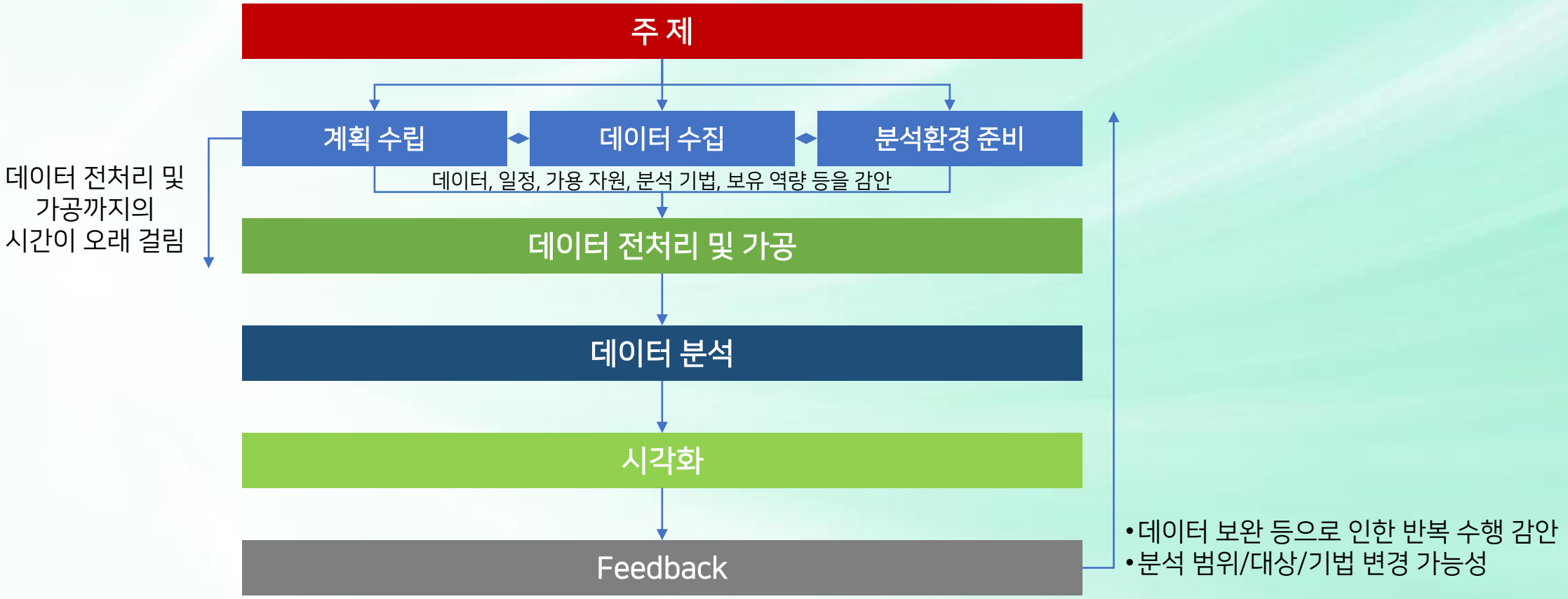


공간빅데이터 분석 관련 분야

- 공간 빅데이터 분석을 위해서는 다음과 같은 다양한 분야를 다루게 됨
- 본 자료에서는 제약상 스토리에 따른 핵심 기법을 짚어보는 방식으로 진행
- 새로운 Insight의 밑받침이 되는 아이디어와 관련 업무에 대한 이해도도 매우 중요

주요 분석 과정

- 전반적인 분석 과정은 다음과 같음
  - 주제가 주어진 분석과 데이터가 주어지고 데이터의 분석을 통해 새로운 인사이트를 도출하는 방식으로 나눠볼 수 있음
  - 방향 설정부터 데이터의 수집, 전처리 및 기초 가공 등 본격적인 분석 전의 과정이 가장 시간이 오래 걸릴 수 있음
  - 1차 분석 결과를 검토하여 추가 데이터 수집 및 분석 기법을 보완하여 다시 분석을 진행할 수 있음



차시별 주요 내용

1차시	2차시	3차시	4차시	5차시	6차시
공간정보 개요	분석 준비	QGIS 기반 분석	파이썬 기반 분석	PostGIS 기반 분석	시각화 기타
• 주요 공간데이터 • 공간데이터 제공 사이트 • 좌표계 • 공간 분석 • 공간 시각화	• 분석환경 준비 • 주요 분석 과정 • 데이터 수집	• 데이터 전처리 • 공간분석 • 시각화	• 데이터 전처리 • 공간분석 • 시각화	• 데이터 전처리 • 공간분석 • 시각화	• 시계열 • 네트워크 • 3D

주요 공간 데이터 및 공간정보 포맷

구분		특성	비고
CSV	주소	• CSV 항목에 도로명주소 또는 지번주소(행정동 주소 포함)가 포함된 경우 • 상세주소(건물번호/ 지번)까지 포함되어 있는지 확인 필요 • 포털사이트의 API는 대부분 도로명/지번주소가 섞여 있어도 처리되는데, 서비스에 따라 구분이 필요한 경우도 있음 • **정제된 주소이더라도 () 내 또는 부가 상세주소에 구분자(,)가 포함된 경우로 인한 문제가 가장 빈번하므로 상황에 따라 부가적인 상세주소는 지우고 이용하는 게 더 나을 수도 있음** • 또한 어떤 **지오코딩** 서비스를 이용하느냐에 따라 좌표 반환이 실패하거나 유사 주소 좌표로 제공되는 경우가 있음 따라서, 지오코딩 신뢰도를 높이려면 2가지 이상의 지오코딩 서비스를 이용할 필요도 있음	↓ 지오코딩
	좌표	• CSV 항목에 경도와 위도 항목이 포함된 경우 • 경도는 대한민국에서 127 등 3자리로 표현되는 동서축 좌표, 위도는 36 등 2자리로 표현되는 남북축 좌표 • **경위도 좌표계의 경우 소수자리가 최소 6자리 이상 되어야 정확한 위치에 매핑됨** • 경위도 좌표가 도/분/초(DMS)로 표현되는 경우가 간혹 있음. 분/60, 초/3600하여 합산하면 됨 • 간혹 TM (6-6자리 숫자 또는 7-7자리) 좌표로 된 경우가 있음. **메타데이터 확인 또는 좌표계 추정** 필요 • 공간적 범위(대상 행정구역/ 한반도)를 벗어나는 **공간적 이상치** 확인 필요	↓ 집계 (속성) ↓ 매핑 (원데이터) 공간조인 (행정구역)
	(행정)구역	• 주소/좌표가 아닌 행정구역 또는 통계적 구역(국가기초구역, 격자 등)에 집계화된 상태 • 국가통계포털 등을 찾아보면 시군구/ 읍면동 단위로 집계화된 정보들이 있으므로 먼저 확인 필요 • 반면 시군구 이상 단위로 집계화된 정보를 가지고는 더 세부적인 정보 파악은 어려움	↓ 속성조인
SHP		• GIS 분야의 산업 표준(de facto) 포맷 • 최소 3개의 동일 명칭 파일(shp/ shx/ dbf)이 한 경로(폴더) 내에 있어야 함 • dbf 기반이기 때문에 1백만 건 이상의 데이터를 저장할 수 없음. 8자리 이상의 컬럼명 등 제약사항이 있음 • 좌표계를 정의하면 .prj 파일이 생겨서 다음에 GIS 툴에서 로딩시 좌표계를 인식함 • .cpg 파일을 통해 캐릭터셋 인코딩을 인식함	GeoPackage (OGC 표준) / PostGIS (Postgres)
GeoJson		• Json 포맷을 기반으로 점/선/면 Geometry(도형 좌표정보)를 저장하는 포맷 • Tool/ 개발언어간 호환시 유용하나 상대적으로 용량이 커서 대용량 데이터 표출(Display)시 빠르지 않음	Topojson
TMS (배경지도)		• Google 및 국내 포털 등에서 지도서비스를 제공할 때 배경지도를 축척별 타일 이미지로 제공하는 지도서비스 방식 • QGIS/ Python-Folium 등에서 배경지도를 표현할 때 이용	OSM, 브이월드, 포털지도

지오코딩 방법

- GEEPS 주소 → 좌표변환서비스

  http://geeps.krihs.re.kr/geocoding/service_page

  - 포털 등의 지오코딩 API들을 활용하여 변환
  - 속도가 느린 편임

- 비즈GIS 지오코딩 툴(Geocoder-Xr)

  http://geeps.krihs.re.kr/geocoding/service_page

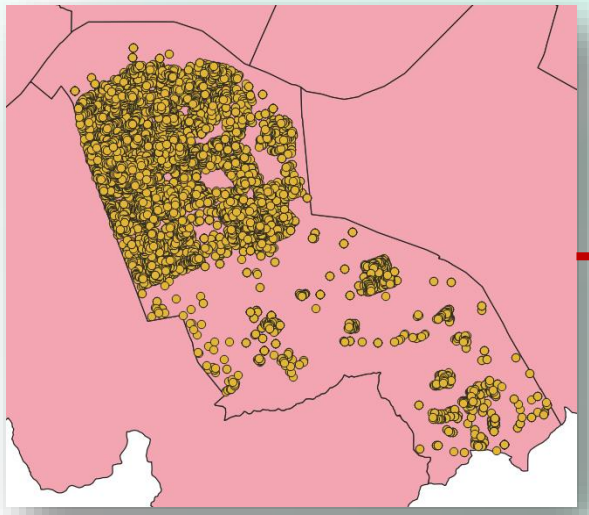
  - 아파트 동별 주소까지 지원하나 라이선스가 없으면 일 1만 건까지만 변환 가능

- 포털 및 브이월드 API
  - 네이버, 카카오, 브이월드에서 제공하는 지도 API의 지오코딩 함수를 이용
  - 웹프로그래밍 또는 파이썬 등을 기반으로 코드 구현이 필요

포인트/행정구역 속성 데이터와 공간영역 데이터간의 Join

- 포인트 데이터는 공간조인을 통해 행정구역 또는 영역(그리드 등) 데이터와의 공간조인을 통해 추출/속성 조인
- 주소/관리 행정기관 등의 정보가 담긴 속성데이터는 행정구역과 속성조인을 통해 연계/집계 가능
  (행정/법정구역 및 관리 코드 주의)

포인트 또는 행정구역 속성/주소가 포함된 데이터

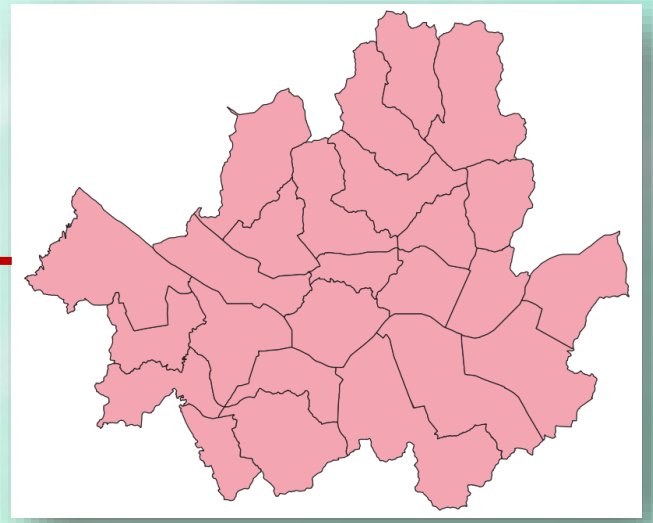


	id	addr	lat	lon
4	개포1동 660-13	서울특별시 강남구 개포1동 660-13	37.47798548...	127.05832355...
5	개포1동 660-5	서울특별시 강남구 개포1동 660-5	37.48208362...	127.05760766...
6	개포1동 660-5	서울특별시 강남구 개포1동 660-5	37.48208362...	127.05760766...

공간조인

속성조인

행정구역/ 영역 데이터



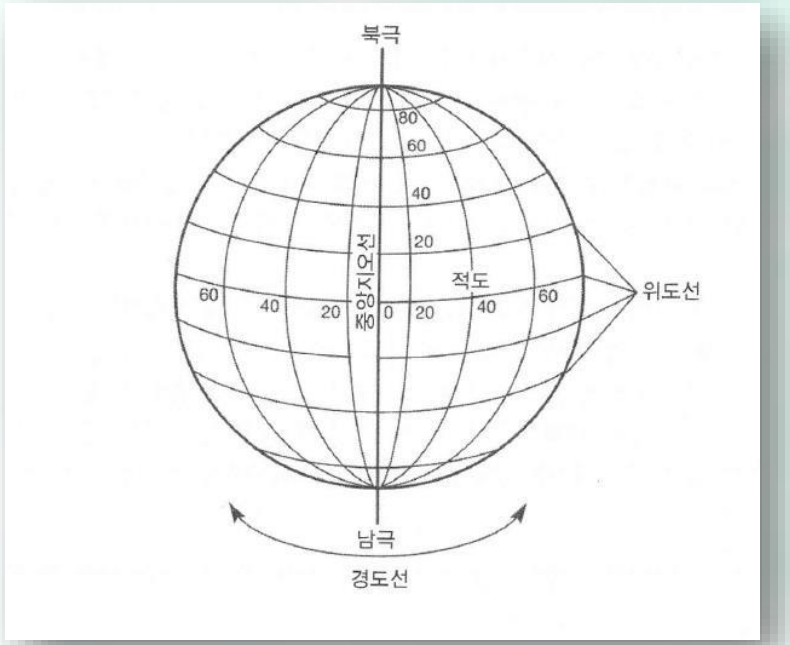
	SIG_CD	SIG_ENG_NM	SIG_KOR_NM
1	11545	Geumcheon-gu	금천구
2	11560	Yeongdeungpo...	영등포구

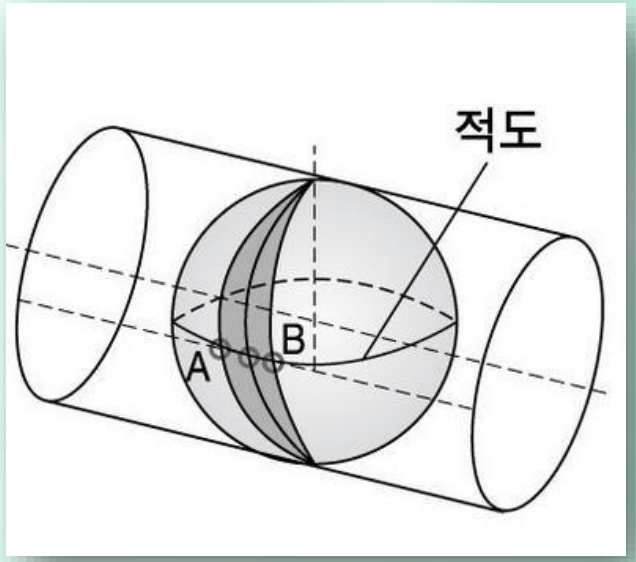
주요 공간데이터 제공 사이트/기관

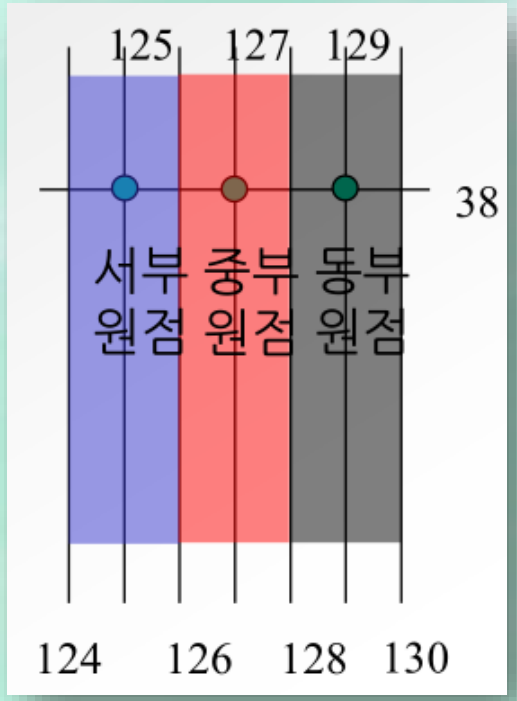
사이트	주요 자료	URL
공공데이터포털	표준데이터, 중점데이터 등 공간/비공간 공공데이터 포털 사이트	www.data.go.kr
국가공간정보포털	**공간정보 위주의 포털 사이트**	**nsdi.go.kr**
도로명주소	도로명주소를 관리하기 위한 건물 및 도로, 국가기초구역 공간데이터	www.juso.go.kr
통계청 통계지리정보	**행정구역, 집계구(통계구역)**	**sgis.kostat.go.kr**
기상자료개방포털	기상 관측 데이터	data.kma.go.kr
데이터스토어	가공 처리된 데이터 포함 (유료 포함)	www.datastore.or.kr
지방행정인허가 데이터개방	지자체에서 인허가 하는 업종별 업소 정보	localdata.kr
서울열린데이터광장	서울시 관할 각종 데이터	data.seoul.go.kr
서울시 빅데이터캠퍼스 *	서울시열린데이터 등을 분석할 수 있는 분석환경 제공	bigdata.seoul.go.kr
K-ICT 빅데이터센터 *	공공데이터 온라인 및 오프라인 분석환경 제공	kbig.kr
산림공간정보서비스	**임상도, 산사태위험지도 등 산림관련 주제도 제공**	**www.forest.go.kr**
문화재공간정보서비스	**문화재 정보 제공**	**gis-heritage.go.kr**
환경공간정보서비스	**각종 환경 관련 주제도 제공**	**egis.me.go.kr**
국토정보플랫폼	**수치지형도, 항공사진, DEM 등 제공**	**map.ngii.go.kr**
브이월드 *	**주요 공간데이터의 목록 및 제공기관을 조회할 수 있음**	**www.vworld.kr/data/v4dc_svcdata_s001.do**

지리좌표계와 평면직각좌표계(TM)

- 좌표계 정의가 안되어 있으면 지도 상의 제 위치에 표시되지 않을 수 있음
- 다수 레이어의 도시(Map Display)는 물론, 특히 공간연산 시에는 연산하려는 공간데이터들을 동일한 좌표계로 통일해야 빠른 성능 및 오류를 방지할 수 있음
- TM은 지리(경위도) 좌표를 원통에 투영하여 평면직각좌표계로 나타낸 것으로, m 단위로 되어 있어 단위 계산 및 공간 연산에 유리







출처: 공간정보체계의 이해와 활용, 국토정보공사 공간정보아카데미

주요 좌표계 목록

범위	타원체/적용시기	좌표계/원점	EPSG code	적용 지도서비스/ 비고	원점 좌표	가상이동원점 좌표
한국	Bessel (~90년대)	TM 서부	5173			
		TM 중부	5174	연속지적도/ 건물통합정보	경도 127.0028902777778, 위도 38	200000, 500000
		TM 동부	5176	(동해, 제주 원점은 편의상 목록에서 제외)		
세계	GRS80 (2000년대)	TM 서부	5180			
		TM 중부	5181		경도 127, 위도 38	200000, 500000
		TM 동부	5183	(동해, 제주 원점은 편의상 목록에서 제외)		
	GRS80 **(2010년 이후 현재 표준)**	TM 서부	5185			
		TM 중부	5186		경도 127, 위도 38	200000, 600000
		TM 동부	5187	(동해 원점은 편의상 목록에서 제외)	경도 129, 위도 38	200000, 600000
		UTM-K	5179/ 102080	**도로명지도/ KTDB 교통주제도/ 정밀도로지도**	경도 127.5, 위도 38	1000000, 2000000
		KATECH	-	KTDB 교통주제도	경도 128, 위도 38	400000, 600000
	WGS84	경위도	4326	**GPS**		
		UTM 52N	32652	한반도 지역 51N~52N/ 정밀도로지도		
		Google	3857	**OpenStreetMap, 구글맵, 브이월드맵**		

좌표계 확인 방법

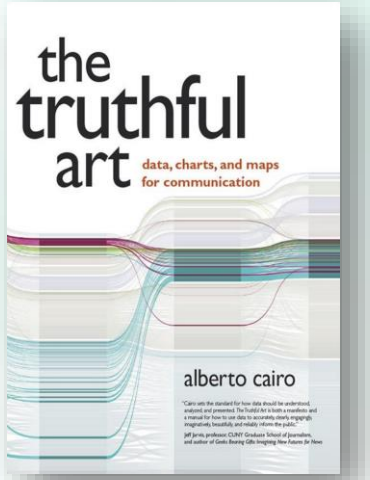
- **좌표계 정보 파일(.prj)이 있는 경우**
  - QGIS 등에 로딩하여 확인. 정상적으로 정의된 좌표계 정보 파일인 경우 QGIS에서 바로 로딩되고 좌표계 정보를 확인할 수 있음

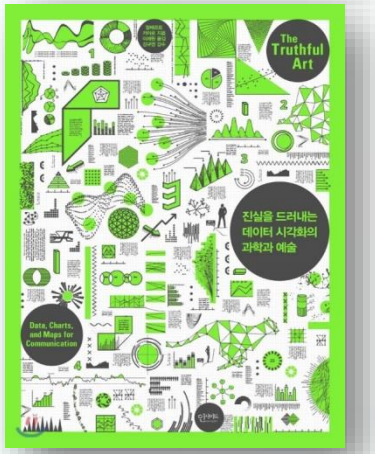
- **좌표계 정보 파일이 없는 경우**
  - 먼저 데이터의 출처/관리기관을 확인 : 도로명주소 주제도는 EPSG:5179, GPS 데이터는 EPSG:4326, 지적도는 EPSG:5174일 가능성이 높음
  - QGIS에 좌표계 정의하지 않고 로딩하여 도형 자체의 좌표를 보고 좌표계를 추정
  - 정수 자리가 127, 38 등으로 표시되는 것은 EPSG:4326 (경위도 좌표계)
  - 남북(Y)축 좌표 정수 자리 수가 7자리면 UTM-K (EPSG:5179)
  - 남북(Y)축 좌표 정수 자리 수가 6자리면 TM 중 하나로 임의의 TM 좌표계를 지정해보면서 추정
  - 위의 과정에서 중부원점으로 정의시, 모든 데이터가 배경지도와 동서로 일정 거리가 떨어진 경우 동부원점으로 바꿔서 정의
  - 속성 정보의 행정구역/주소 등 확인하여 추정(부산 등 동쪽이면 동부원점 가능성)

좌표계 확인 방법

- 일부 공간데이터가 아직 한국측지계인 경우가 있는데 이는 www.osgeo.kr/17에 정의된 좌표계 정의를 QGIS/Geopandas/PostGIS 상에서 커스텀 좌표계로 정의해줘야 함
  - 배경지도와 300여 미터 정도 비스듬하게 떨어져서 보이는 경우는 한국측지계 (Bessel 타원체)인 EPSG 5174/5176으로 변환계수를 포함한 좌표계 정의를 QGIS 또는 Geopandas에서 정의해줘야 함
  - 정의 방법은 2차시에서 설명

공간정보(점/선/면) 시각화





이산/ 범주형 데이터는 심볼모양으로

위계별 행정경계/ 도로는 선의 패턴으로 구분

범주화 영역은 각기 다른 색상 계열을 적용 (Categorized)

다양한 범주값을 각기 다른 심볼과 색상으로 인지하기 쉽도록

동일 항목의 정량적 차이를 심볼의 크기와 색상의 농도 차이로 인지할 수 있도록

정성

정량

Point	Line	Area	Volume
Qualitative	Qualitative	Qualitative	Qualitative
Quantitative	Quantitative	Quantitative	Quantitative

Figure 10.9 Symbols to encode data on maps.

연속형 데이터는 심볼 크기/ 도형표현도

연속형 데이터는 숫자/선의 두께로 표현

연속형 영역은 동일 색상의 농도차로 표현 (Choropleth/ Graduated)

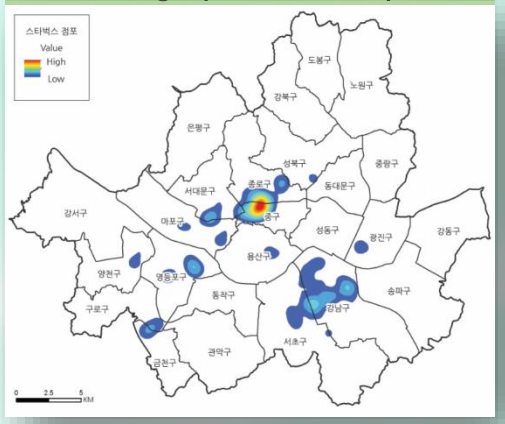
표고, 건물 높이, 양적 수치

기초 공간데이터의 목적별 가공

점묘도
공간적 분포 자체 특성 표현

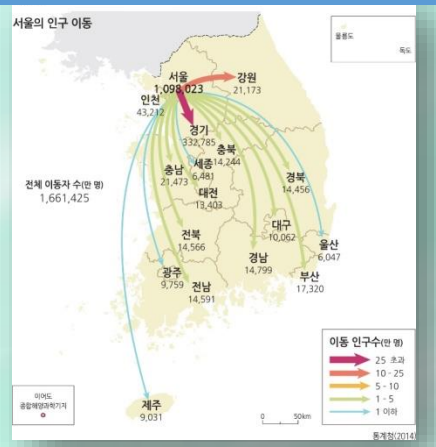
출처 : 한국경제신문 뉴스래빗

HotSpot – 공간적 밀집도
Geographic Heatmap



출처 : 지아이에스유나이티드

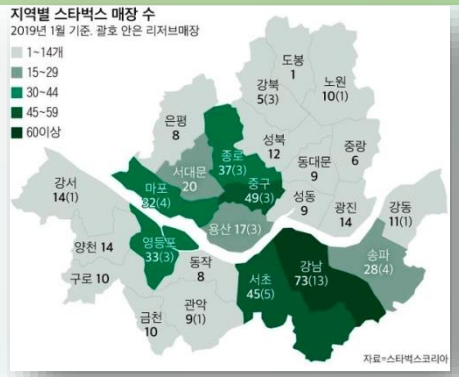
공간상 이동/흐름
Network/ Flow



출처 : 국토리지정보원

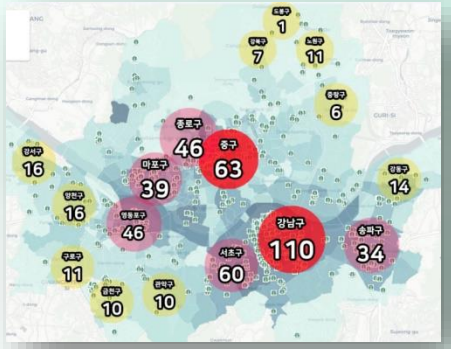
행정구역/ 격자(그리드) 단위 양적 집계

단계구분도 (Choropleth Map)
집계 결과의 순위 표현



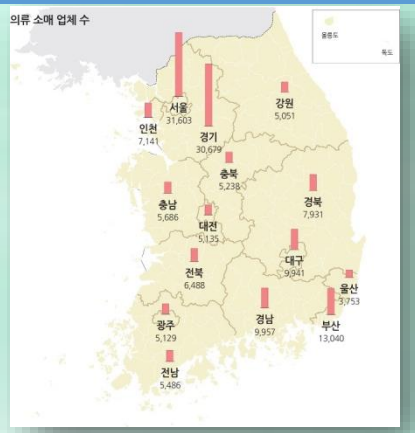
출처 : 조선비즈

Bubble Map
집계 결과의 양적 크기 표현



출처 : SPH

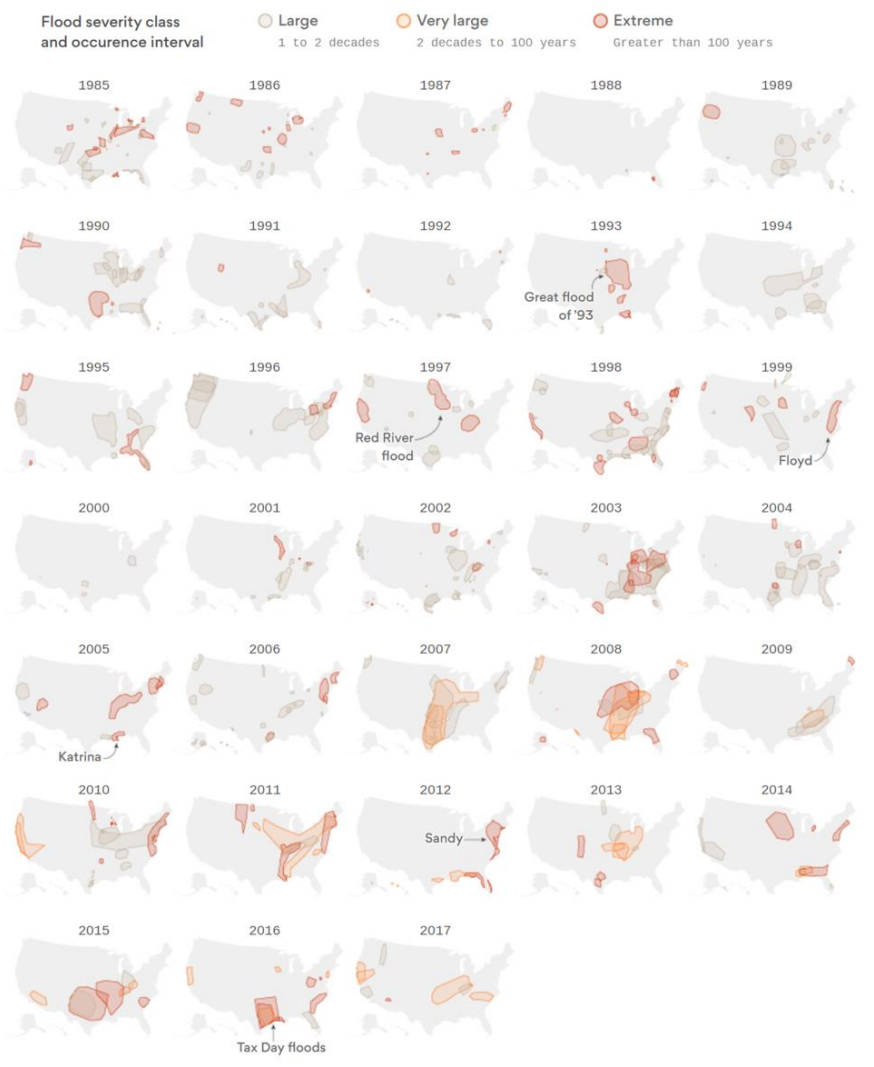
도형표현도
2개 이상의 정량 데이터를 표현 시



출처 : 국토리지정보원

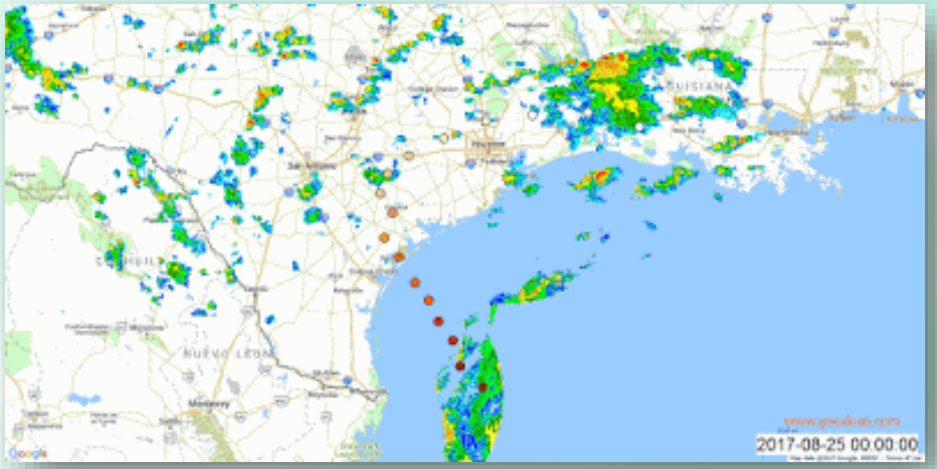
시계열 변화 공간시각화

Small Multiples

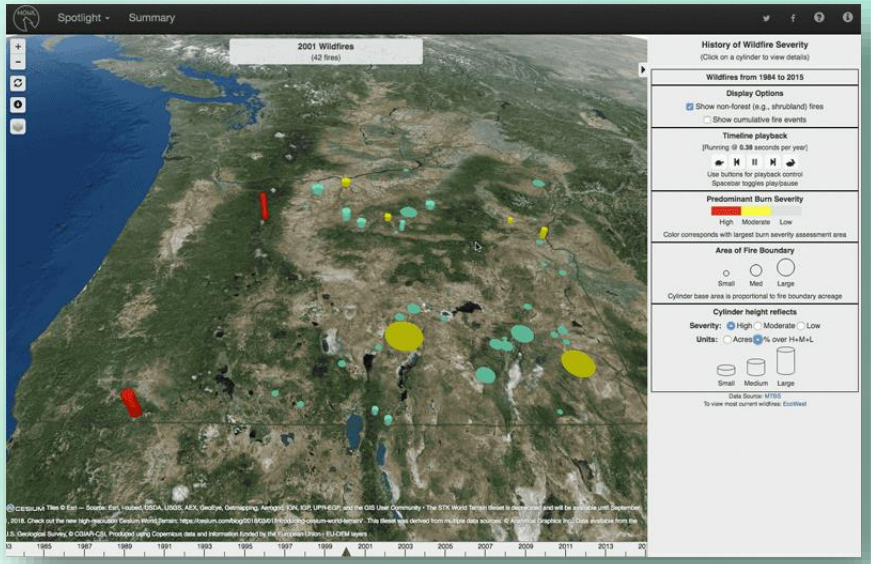


https://www.axios.com/thirty-years-of-major-flooding-in-the-us-1513305213-b0b6bb49-d101-49ef-bc71-95de30179b80.html

Time-series Animation



https://www.geodose.com/2019/11/how-to-create-animation-map.html



https://cesium.com/blog/2018/03/21/czml-time-animation/

공공데이터 주요 가공, 분석 및 시각화

일반적 인사이트 탐색 흐름

원 데이터

전처리 및 가공

지역/속성 기준 집계

상위 n% 등 특징/패턴 파악

동일 Key를 가진
데이터들을 결합 분석

회귀/ 분류/ 군집 등을 통한
특성/원인 분석

공간 데이터 분석/시각화

개략적인 데이터 용량별 분석/시각화 환경

구분	100만 건 또는 1GB 이하	GB 급	10GB 이상 (분산환경/ 클라우드)
일반	• 스프레드시트	• PANDAS • PostgreSQL	• HDFS/ SPARK • NoSQL
공간	• QGIS Desktop • (SHP)	• GeoPackage(.gpkg)* • GEOPANDAS • PostGIS	• Postgres Advanced Server
시각화	• GEOPLOT • PLOTLY • FOLIUM • kepler.gl • deck.gl		

참고문헌

- 구글 트렌드, https://trends.google.co.kr/trends/explore?date=today%205-y&geo=KR&q=ArcGIS,%2Fm%2F0ct9z5,%2Fm%2F0ph46,%2Fm%2F02s9k80
- DBMS 랭킹, https://db-engines.com/en/ranking_trend
- 한국 주요 좌표계 EPSG코드 및 proj4 인자 정리, OSGeo한국어지부. https://www.osgeo.kr/17
- 공간정보체계의 이해와 활용, 국토정보공사 공간정보아카데미 교육
- Alberto Cairo, “The Truthful art”
- 알베르토 카이로, "진실을 드러내는 데이터 시각화의 과학과 예술", 인사이트
- 모토하시 도모미쓰, “데이터 전처리 대전”, 한빛미디어
- QGIS 설치, OSGeo한국어지부 블로그, http://blog.daum.net/geoscience/1354
- QGIS 사용자 지침서, https://docs.qgis.org/3.4/ko/docs/user_manual/
- QIGS 교육교재, https://docs.qgis.org/3.4/ko/docs/training_manual/
- QGIS Time Manager 플러그인 사용법, OSGeo한국어지부 블로그, http://m.blog.daum.net/geoscience/988
- QGIS 3D 맵뷰, OSGeo한국어지부 블로그, http://m.blog.daum.net/geoscience/1289

참고문헌

- 민형기, “파이썬으로 데이터 주무르기”, 비제이퍼블릭
- PostGIS 설치, OSGeo한국어지부 블로그, http://m.blog.daum.net/geoscience/1237?np_nil_b=2
- 김우미/유병혁, “PostGIS 시작하기”, https://www.slideshare.net/ybh0616/postgis-101511460
- PostGIS 프로젝트 운영위원회, “PostGIS 공식 가이드북”, https://www.osgeo.kr/231
- PostGIS 사용자 지침서, https://postgis.net/docs/manual-2.4/postgis-ko_KR.html
- 2016년 범죄지도, SBS 마부작침, http://mabu.newscloud.sbs.co.kr/20170308crimemap/web/index.html
- 단계구분도 분류기법, https://pro.arcgis.com/en/pro-app/help/mapping/layer-properties/data-classification-methods.htm
- 단계구분도 분류기법, http://www.qgistutorials.com/ko/docs/basic_vector_styling.html
- 단계구분도 분류기법, 지아이에스유나이티드 블로그, https://gisutd.tistory.com/7