

# 빅데이터 분석 플랫폼 설계 & 구축

1주차 (4/16)

# 강사 소개 | 김영욱



### 경력

- LG Electronics
- NHN (현 NAVER)
- Accenture
- 서울대학교 화학공정 신기술연구소

### 좋아하는 것

- 책, 음악, 드라마, 영화, 소셜미디어
- 걸으며 음악 듣기
- 경치 좋은 곳에서 공상하기
- 믿음 주고 사랑 받는 동료들
- 설명하는 것 (설명병)

### 싫어하는 것

- 대한민국 정치
- 대한민국 환경 (특히 미세먼지, 황사)
- 앞뒤 다른 사람
- 교수님이라고 부르는 것 (님으로만..)

궁금한 건 편하게 물어보세요. ^^

## 강의를 시작하기 전에...

---

### **Learning by doing**

학습의 가장 중요한 것은 lecture가 아닙니다. 각자 경험을 하고, 질문을 하고, 논의하면서 개개인이 발전하는 것입니다.

### **학습 목표 & Post-mortem**

주 별 학습 목표를 정의하고 각자 목표가 얼마나 달성되었는지 post-mortem을 통해 확인합니다. 부족한 부분은 차주 수업에 반영되도록 노력할 것입니다.

### **강사의 한계를 인정해 주세요.**

강사는 모든 것을 아는/알아야 하는 사람이 아닙니다. 특히나 특정 도메인 지식은 많이 부족합니다. 현장의 문제를 풀면서 느끼고 배운 것을 benchmarking해 간다고 생각하면서 8주의 시간을 보내야 합니다.

## Goal

실제적 비즈니스 문제 해결을 위한 데이터 분석 역량  
데이터 분석 플랫폼 Architecture 설계 역량

## Activities

문제정의 → 데이터 수집 / Prototyping → 핵심 모듈 정의 및 Architecture 설계  
→ 데이터 분석 플랫폼 구축 → 데이터 수집/저장/처리 → Dashboard 구현 →  
Report 작성

## Products

Architecture (optional)  
데이터 플랫폼 (mandatory)  
Dashboard (mandatory)  
분석리포트 (optional)  
예측모델 (optional)

## Assessments

산출물 기반의 교수 평가 50% - 별도의 시험은 없음  
산출물을 만드는 과정에서의 참여도와 역량에 대한 Peer Review 50%

## 강의 구성 및 진행 방식

- 조별 프로젝트 형태로 진행함.
- 풀고자 하는 문제를 정의하고 문제를 풀기 위해 필요한 데이터를 찾고, 이 데이터를 처리, 저장, 분석할 수 있는 데이터 분석 플랫폼 Architecture를 정의한 후에 이를 구현함.
- 조별로 [문제정의 / 데이터 / Architecture / 분석플랫폼 / Dashboard / 분석 리포트]를 만들어야 함
- 강의는 기본적으로 전체를 대상으로 하는 필수 강의와 소수의 수강생들을 대상으로 하는 선택 강의로 나누어 짐. 선택 강의는 니즈가 있는 수강생만을 선별적으로 모아서 진행함. 그 외에 프로젝트 단위로 필요한 강의는 상황에 맞추어서 진행함.
- 프로젝트 과정에서 발생하는 산출물은 모두 Github에서 관리됨. 소스 코드는 Git에서 나머지 모든 문서들은 Wiki를 통해 관리함. (지식 축적과 공유 차원에서 중요)

# 강의 계획

주차	날짜	강의	목표
1주차	4/16(토)	문제 및 데이터 정의	<ul style="list-style-type: none"> <li>한 학기 동안 풀 문제를 정의한다.</li> <li>문제를 풀기 위한 데이터 정의 및 검증 한다.</li> </ul>
2주차	4/22(금), 4/23(토)	데이터 분석 플랫폼 Prototyping	<ul style="list-style-type: none"> <li>데이터 분석 플랫폼의 기본적인 Component들을 활용한다. (AWS, Splunk, R 등)</li> <li>데이터 초기 탐색을 통해 비즈니스 문제를 풀 수 있는지 검증한다.</li> </ul>
3주차	4/29(금)	데이터 분석 플랫폼 Architecture 설계	<ul style="list-style-type: none"> <li>데이터 수집/처리/저장/활용에서 필요한 프로세스 &amp; 기능을 정의한다.</li> <li>핵심 컴포넌트를 정의하고 데이터 분석 플랫폼 Architecture를 설계한다.</li> </ul>
4주차	5/13(금), 5/14(토)	데이터 분석 플랫폼 구축 1차	<ul style="list-style-type: none"> <li>Public Cloud(AWS)에 데이터 분석 플랫폼을 위한 기본 컴포넌트를 구성한다.</li> <li>데이터 처리/분석/모델링에 필요한 라이브러리나 모듈을 구현한다.</li> <li>데이터를 올리고, 핵심 컴포넌트들이 정상 동작하는지 검증한다.</li> </ul>
5주차	5/20(금)	중간 발표회	<ul style="list-style-type: none"> <li>구현한 산출물을 전체 공유한다.산출물의 개선 방향과 앞으로의 계획을 수립한다.</li> </ul>
6주차	5/21(토), 5/27(금)	데이터 분석 플랫폼 구축 2차	<ul style="list-style-type: none"> <li>데이터 분석 플랫폼의 결과 적합성을 높이고, 데이터 프로세싱을 고도화 하기 위한 모듈을 구현한다.</li> </ul>
7주차	5/28(토)	Dashboard 구현 및 결과 발표회를 위한 산출물 정리	<ul style="list-style-type: none"> <li>비즈니스 목적에 맞는 Dashboard를 구현한다.</li> <li>Dashboard를 전체 공유하고 조별 Critic을 진행한다.결과 발표회를 위한 산출물 정의 및 정리 한다.</li> </ul>
8주차	6/3(금)	분석 리포트	<ul style="list-style-type: none"> <li>Dashboard를 기반으로 의사결정 하는 분석 리포트를 작성한다.</li> </ul>
9주차	6/4(토)	결과 발표회	<ul style="list-style-type: none"> <li>조별 프로젝트 결과를 발표한다.</li> <li>부족한 부분과 개선할 부분을 이해한다.</li> </ul>

## 1주차 목표

---

- 한 학기 동안 풀기 위한 문제를 정의한다.
- 함께 문제를 풀어갈 팀을 구성한다.
- 문제를 푸는데 필요한 데이터를 확인한다.
- 개발 환경과 활용할 툴을 이해한다.



Break Time

## 문제 정의

---

- 비즈니스 목적과 이를 해결하기 위한 문제를 정의하는 것이 중요함.  
관심 있는 분야의 문제를 정의하는 목적은 내적 동기를 끌어내기 위함임
- 문제 정의하는 과정은 Biz Function 단위로 진행  
(마케팅, 품질, 서비스, 제품개발, 상기 등)

why biz function 단위로?

데이터 분석을 통해 실질적인 문제를 푸는 단위가 Biz function임.

꼭 Biz Func. 단위로 해야만 하나요? NO!

- **개인별로 관심 있는 biz function과 그 안에서 풀고자 하는 문제를 정의합니다.** (마케팅, 품질, 서비스, 제품개발, 상기, 생산 등)
- **세부 목표에 대한 카테고리를 정의합니다.**  
세부목표 예시: 마케팅 효과성 검증 / 고객 Pain Point 이해 / 사용패턴 분석 / 신규 서비스 개발 / 업무 효율 개선 / 기타
- **문제는 가능한 명확하게 정의하는 것이 좋습니다.**  
'크롤링 데이터로 고객을 이해하고 싶습니다'라는 식은 좋지 않습니다.

- 한 명씩 정의한 문제에 대해서 이야기 합니다.
- 이 때 비슷한 분야를 이야기 한 것 같다고 생각되는 분들이 있으면 손을 들어 자신이 정의한 문제를 이야기 하고 옆 자리로 옮겨 옵니다.
- 그냥 저 사람이 좋다고 생각하시는 분이 오셔도 좋습니다.  
(조가 만들어 질 때는 여러 동기가 있을 수 있으니까요..)

Break Time



Lunch Time

- **개인의 문제를 공유하고, 함께 풀 만한 문제를 정의합니다.**
- **데이터 분석이 Biz Purpose 비즈니스 목적을 달성하는데 도움이 되는지를 확인해야 합니다.** 데이터 분석을 통해 해결할 수 없는 문제를 가지고 온다면 다음으로 이어지는 활동이 아무 의미가 없습니다.
- **문제를 정의할 때 해결 방안이 어느 정도 feasible한지도 생각을 하면 좋습니다.**
- **풀려는 문제가 어느 정도 일치하고, 사람들도 좋고, 기술 세트도 어느 정도 갖추어진 4~5명의 조를 만들어 주십시오.**



- 조별로 풀려는 문제를 1차로 확정해 주십시오.
- 조별로 만들고자 하는 산출물을 정의해 주십시오.  
Architecture (optional) / 데이터 플랫폼 (mandatory)  
Dashboard (mandatory) / 분석리포트 (optional) / 예측모델 (optional)
- 정의한 문제를 끝까지 풀어야 하는 것은 아닙니다. 언제든지 목표는 바꿀 수 있습니다.

- 문제를 푸는데 필요한 데이터를 어떻게 수집할 것인지 정의합니다.
- 데이터가 없으면 할 수 있는 것도 없습니다. 조원 중에서 내부 데이터를 가지고 올 수 있던가 아니면 외부에 있는 데이터를 수집해 올 수 있는 방법을 구상해야 합니다.
- 이 시간에 모든 것을 확인할 수는 없습니다. 일차적으로 확인해 보고 어떻게 구체적으로 확인해 볼 것인지를 논의하고, 각자 추가적인 시간을 갖고 데이터를 가지고 올 수 있는 방법을 찾아야 합니다.

Break Time




Break Time

조별로 아래 사항을 정의하고, 공유합니다.

- 조명
- 정의한 문제
- 산출물
- 데이터 수집 및 활용 방안 (벤치마킹 하셔야 겠죠..)
- 조별 운영 철학 (가능한 재미있게..)

# Lecture | 공공 데이터 소개

- Public Data 페이지 : <https://github.com/assistbig/bigdata/wiki/Public-Data>

 assistbig / bigdata

Unwatch 3 Star 0 Fork 0

[Code](#) [Issues 0](#) [Pull requests 0](#) **Wiki** [Pulse](#) [Graphs](#) [Settings](#)

## Public Data

assistbig edited this page 41 minutes ago · 11 revisions

[Edit](#) [New Page](#)

### 데이터 포털

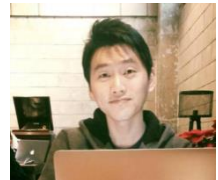
1. 공공데이터 포털
  - 한국정보화진흥원에서 운영하는 공공데이터 통합제공 시스템
  - 교육, 문화관광, 공공행정, 재정금융등 정부가 보유한 다양한 공공데이터를 개방
  - 파일 다운로드 / Open API 방식으로 운영
2. 보건 의료 빅데이터 개방 시스템
  - 건강보험심사평가원에서 보유한 보건의료관련 공공데이터
  - 데이터 소개 Page
3. 서울 열린데이터 광장
  - 서울시가 보유한 공공데이터를 개방한 공유 플랫폼
  - 파일 다운로드 / Open API 방식으로 운영
4. K-ICT 빅데이터 센터
  - 빅데이터 관련 창업 및 중소기업을 지원하는 K-ICT 센터(한국정보화진흥원 주관)에서 관리하는 데이터 시스템
5. Quandl - 금융 데이터
  - 전 세계 금융 관련 데이터를 가지고 올 수 있는 데이터 플랫폼
  - R, Python 등의 프로그램 언어 SDK로 데이터 제공
  - Wiki-Quandl
6. 국가지표체계
  - 국가 발전 상황의 장기적인 추세를 전망하기 위해 연 단위의 데이터 공개
  - 구가 주요 정책 수립에 긴요하게 활용될 수 있는 성과 중심 지표로 주요지표 139개 선정 및 공개

**PAGES** 17

**MENU**

- GAPA
- 강의계획서
- 1주 (4/16) - 문제 및 데이터 정의
- 2주 (4/22, 4/23) - 데이터 플랫폼 Prototyping
- 3주 (4/29) - Architecture 설계
- 4주 (5/13-14) - 데이터 분석 플랫폼 구축 1차
- 5주 (5/20) - 중간 발표회
- 6주 (5/21, 5/27) - 데이터 분석 플랫폼 구축 2차
- 7주 (5/28) - Dashboard 구현 및 산출물 정리
- 8주 (6/3) - 분석 리포트
- 9주 (6/4) - 결과 발표회
- Public Data
- 자료실
- 활용 툴
- 기타

Clone this wiki locally



# Homework

---

- 조별로 풀고자 하는 문제를 detail하고 solid하게 정의합니다.
- 문제를 실제 데이터로 풀 수 있는지를 재 검증합니다.
- 문제를 푸는데 필요한 데이터를 수집할 수 있는 방안을 마련하고 샘플 데이터를 수집해 옵니다.
- 조에서 정의한 내용과 유사한 문제를 푼 사례를 조별로 1가지씩만 조사해서 다음 주에 공유합니다. 유사한 문제가 아니더라도 인상적인 사례를 찾아서 공유해도 좋습니다.
- Github에서 human으로 인정 받아 오시길..



## 1주차 목표 Post-mortem

---

- Google Survey: <http://goo.gl/forms/On4Vy1DuMC>
- 조를 입력합니다.
- 주차 목표에 대한 목표 달성 수준을 평가합니다. (1~7점)
  - 1) 한 학기 동안 풀기 위한 문제를 정의한다.
  - 2) 함께 문제를 풀어갈 팀을 구성한다.
  - 3) 문제를 푸는데 필요한 데이터를 확인한다.
  - 4) 개발 환경과 활용할 툴을 이해한다.
- 주차 목표 별 Comment 사항을 작성합니다.
- 전체 Comment 사항을 작성합니다.