데이터에서 가치를 만드는 방법 그 어려움에 대하여

Coupang
이주형

발표자 소개

- *유학생*
- *Quant*
- *Data scientist*

들어가기에 앞서...

- *본 발표는 쿠팡과 관련이 없음을 미리 밝힙니다.*
  - *쿠팡에 입사한지 이제 2주일...*
  - *사실 아직 제가 무슨일 하는지도 몰라요...*
- *제가 아는 것들이 일반적이지 않을 수 있음을 주의하셔야 합니다.*
  - *특히 시스템이나 DB등에 대한 경험이나 지식이 약해요...*
- *발표가 매우 지루하오니 미리 양해 부탁드립니다.*
  - *기술에 대해서 얘기하면서 재미있는 예제들을 보였어야 하는데...*
  - *주제를 잘못 잡았어요...*
- *그리고 오늘은 제 생일입니다...*

데이터 활용의 시대

- 데이터는 21세기 원유, 분석기술은 엔진 (Gartner 2011, 박근혜 2015)
- 데이터 과학자, 21세기 가장 섹시한 직업 (*Harvard business review 2012)*
- 딥러닝이 세상 모든 것을 바꿀 것 *(Forbes 2016)*

발표의 목표 – 데이터에서 가치를 만드는 법

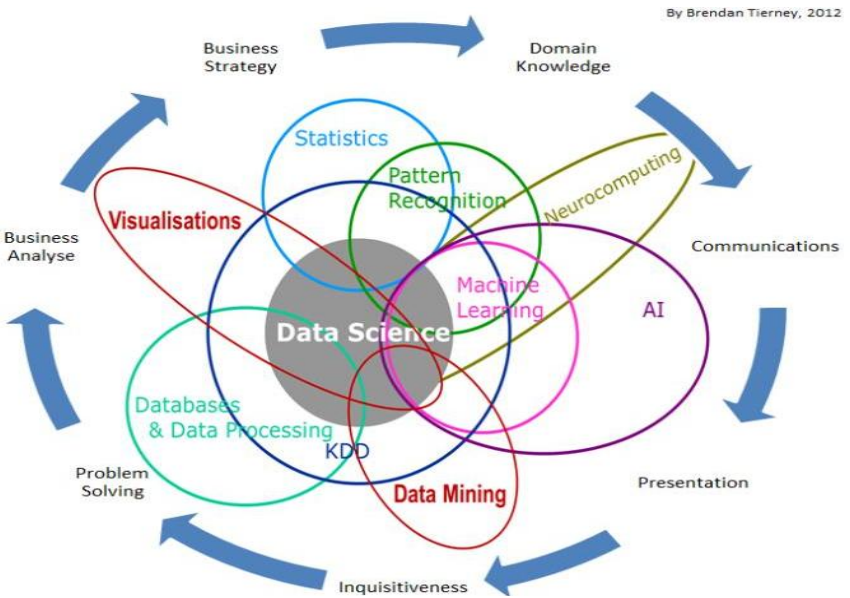
- *일반적인 발표*
  - *매우 좋은(좋아보이는) 실제 사례를 가지고 성공스토리 공유*
    - *데이터 기반 의사 결정이 이렇게 좋아요!*
  - *으리뻔쩍한AI 알고리즘에 대한 이론적인 설명과 적용사례*
    - *머신러닝 써서 데이터만 넣고 컴퓨터가 알아서 다 해요!*

- *이 발표의 목표*
  - *데이터에서 가치를 만드는 실제 프로세스에 대한 설명*
  - *각 프로세스가 어떻게 어렵고, 실제로 어떤 문제가 생기곤 하는지.*

- *기대되는 결과*
  - *데이터를 다루면서 나타날 수 있는 문제들을 미리 알고 ~~대처한다.~~ 버텨낸다.*
  - *데이터 팀 안에서 역할배분이나 일의 진행을 더 효율적으로 할 수 있다.*

데이터 과학

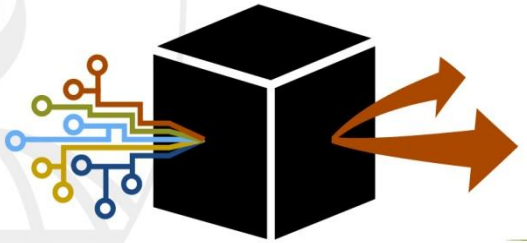
- 데이터에서 가치를 만드는 작업
  - *데이터를 다루고 무엇인가 쓸만한 것을 찾아내었다면 모두 데이터 과학*
  - *매우 많은 분야에 걸쳐 있으며, 그 경계 또한 모호*
  - *꿀 빨수도 있었지만… 분야가 너무 유명해졌어요…*



데이터 과학 업무 종류

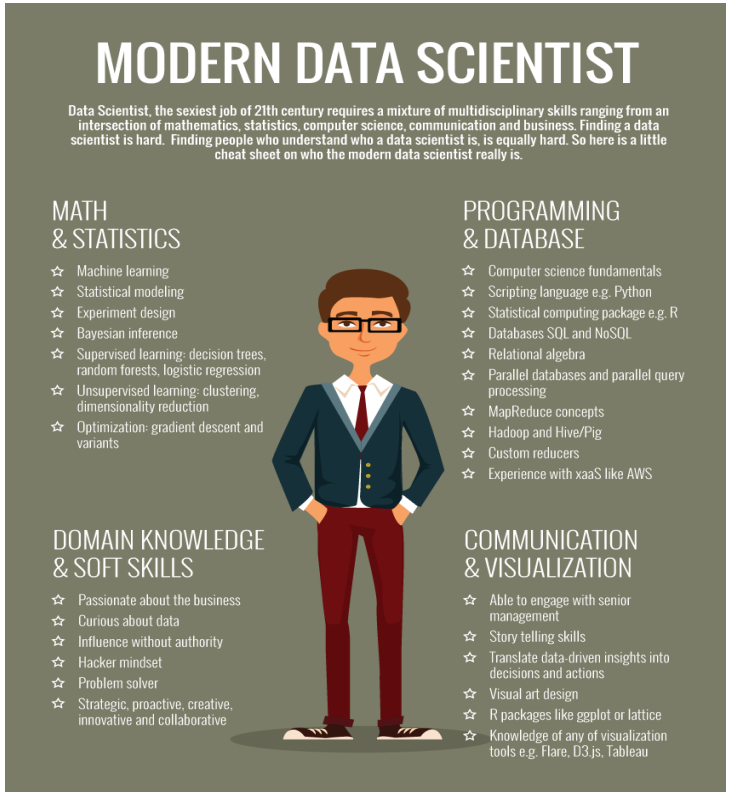
- 데이터 분석
  - 데이터에서 의미를 추출하는 과정

- 데이터 예측
  - 데이터에서 구조화된 지식을 바탕으로 새로운 데이터에 대한 추론을 하는 과정



- 실제 업계에서는…
  - 일반적으로 분석과 예측을 함께 하게 되는 경우가 많음.

데이터 과학자



Data scientist

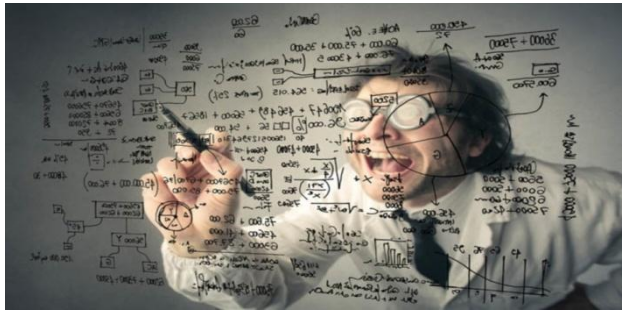
- ***Data engineer***
- *Data Processer*
- *Data miner*
- *Data conditioner*
- *Data Visualizer*
- ***Algorithm developer***
- *Data analyst*
- *Statistician*
- *Machine learning engineer*
- *Machine learning scientist*
- ***Data business person***
- *Data product manager*
- *Business intelligence analyst*
- ***Researcher***
- *Data Innovator*
- *Data creator*
- *….*

참고 글 : Data scientist 는 누구인가?

https://brunch.co.kr/@data/4

데이터 과학의 난관

- *현실은 시궁창*



배워야 할 것은 너무 많고요,

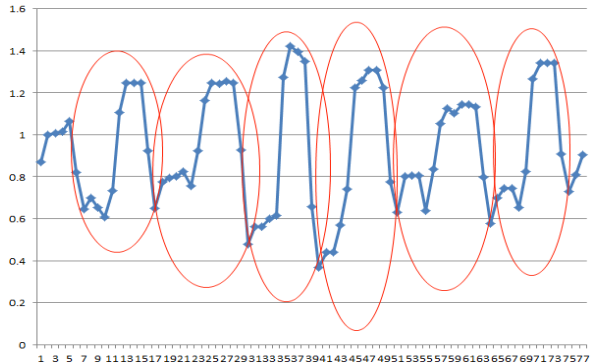
사람들은 말을 안들어먹고요,

나도 내가 뭘 하는지 모르겠어요...

Problem 1. 기본 능력

Q. Algorithm developer가 되려면 Job description 에 나열되어 있는 시스템과 알고리즘들을 다 알아야 하나요?

A. 아니요, 데이터 문제 풀이 능력(어떻게든)이 있으면 됩니다.



Problem 2. 프로세스 이해도 부족 & 팀워크 실패

- 우선 이해해야 하는 3개의 원칙이 있습니다.

- 제 1 원칙 : 모든일은 같이 해야 한다.
  - 일이 나눠져 있는것 같죠? 사실 다 같이 봐야 하는 일이에요.
- 제 2 원칙 : 과정의 반복을 당연시 해야 한다.
  - 순차적으로 완료되는 프로세스가 아니라 내부에서 무수히 많은 반복이 있을 수 밖에 없어요.
- 제 3 원칙 : 자기 일을 계속 설명해야 한다.
  - 제발 좀 설명과 홍보를 귀찮아 하지 마세요.
  - 결정권자와의 소통 뿐 아니라 서로 다른 직군의 프로젝트 팀원들 사이에서도 소통이 필수적이에요.

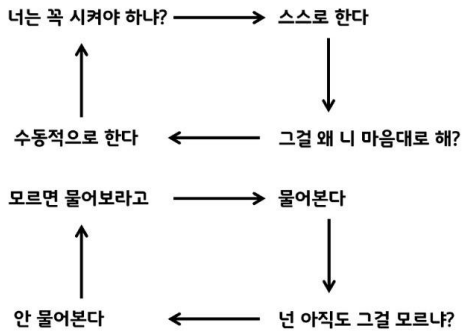
Problem 2. 프로세스 이해도 부족 & 팀워크 실패

- 합쳐놓고 보면....

1. 같이 구르고 보니 진흙탕

+

2. 답이없는 무한반복

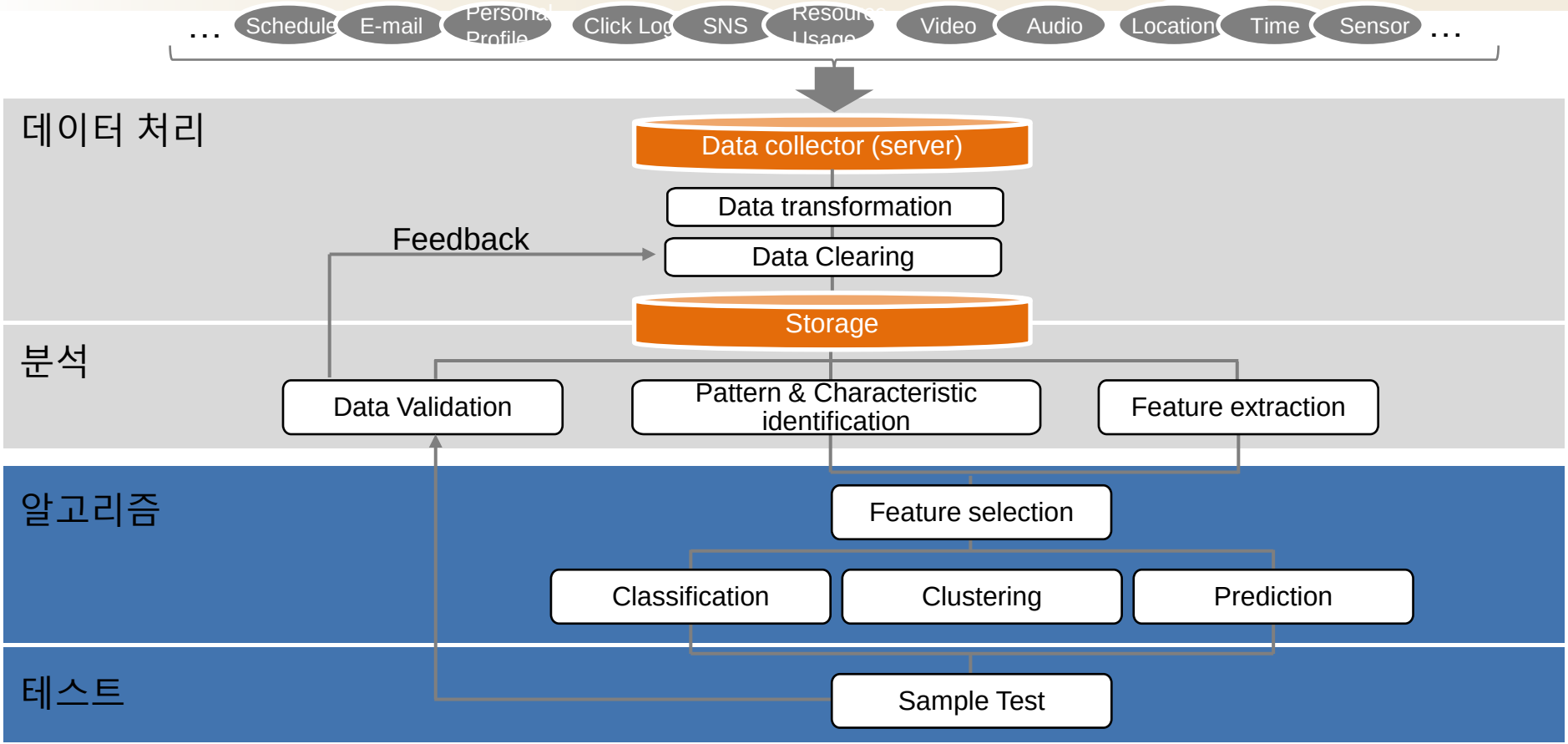


+

3. 너무 솔직한 소통

- 그럼 지금부터 데이터를 활용하는 프로세스를 설명하겠습니다.

데이터 활용 시스템


...
Schedule
E-mail
Click Log
SNS
Video
Audio
Location
Time
Sensor
...
데이터 처리
Data collector (server)
Data transformation
Feedback
Data Clearing
Storage
분석
Data Validation
Pattern & Characteristic identification
Feature extraction
알고리즘
Feature selection
Classification
Clustering
Prediction
테스트
Sample Test

1. Objective and data setup

- 프로젝트의 목적은 항상 애매한 언어적 표현으로 시작됩니다.
  - 등장인물 : 결정권자 , Data business person
- 1) 목표 설정
  - 수식으로 정의되는 Metric 또는 Objective function 을 설정.
  - 최대한 구체적인 레벨로 결정권자와의 합의가 중요.
- 2) 데이터 결정
  - 데이터를 구체화 해 나가며 예상되는 문제와 비용에 대한 해결방안 준비.
  - 여러 제한 조건들을 단계적으로 정해 데이터의 범위를 제한.

- 중요능력: Feasibility 확인
  - 엔지니어와 알고리즘 제작자는 빠르게 제작 가능 여부를 판단해 알려주어야 합니다.
  - 중요한 요소 : 엔지니어의 느낌, 제작자의 느낌 …

    (Resource, Context, Condition, False tolerance, User intervention …)

2. Data gathering

- 어떤 데이터를 어떻게 모아 어떻게 저장할 것인가요?
  - 등장인물: Data engineer, Algorithm developer, Data business person
- 1) 데이터의 양과 종류의 결정.
  - 프로젝트를 진행할 때 충분한 정확도를 보장할 수 있는 양의 데이터 확보.
  - 조금이라도 관계가 있을 것 같은 모든 Data의 확보.
- 2) 데이터 수집 방법 결정
  - System logging 및 online 수집.
  - 실험군의 사람들로 User test 진행
  - Web crawling 등의 외부 소스 활용.
- 3) 데이터 저장 방법 결정
  - HW/SW 시스템의 결정, DB의 결정
  - 저장하는 데이터를 계속적으로 추가할 수 있는 시스템 필요.
  - 데이터 항목들에 대한 엔지니어와 알고리즘 제작자의 긴밀한 협업 필요.

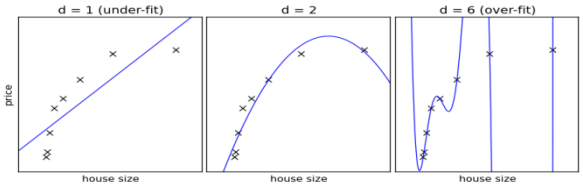
3. Data processing

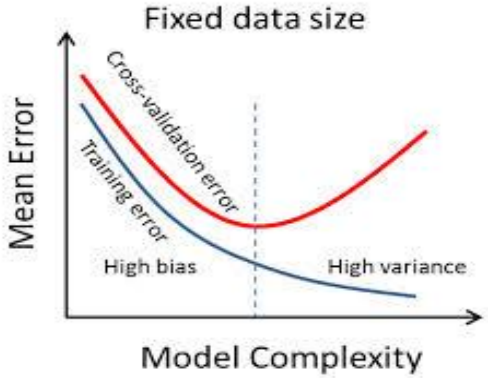
- Raw data들의 오류를 수정하고 원하는 타입으로 변경하는 것이 보통 가장 오래걸리는 과정입니다.
  - 등장인물 : Data engineer, Algorithm developer
- 1) Data cleaning
  - 모든 프로세스 중 가장 많은 시간을 소비
  - Invalid, incomplete, duplicate, noisy, inconsistent, too-many data 의 처리.
  - 한가지 방법으로만 진행하는 것이 아니라 반복 작업이 필요할 경우가 대부분.
- 2) Data wrangling
  - [Categorical, Numerical, String data] , [Event-based, State-based data]
  - [Sensor, System, User data]
  - Training & Testing data 분리

- 중요 능력: Data insight
  - 데이터를 보고 1차적인 impression을 통해 앞선 과정들을 수정할 수 있다.
  - 숫자에 대한 감

4. Algorithm development

- 어떤 모델과 어떤 요소를 가지고 효과적인 결과물을 만들건가요?
  - 등장인물: Algorithm developer
- 1) Data analytics
  - EDA, Visual analysis, Story analysis
- 2) (Predictive) Model & Feature selection
  - 관계 특성을 잘 모델링, Overfitting & Underfitting 주의.
  - Feature selection 과 feature warping 방법중 효과적인것을 사용.
- 3) Algorithm tuning
  - 앞서 정의된 Metric 또는 Objective function 을 대상으로 최적화
  - Wrap the algorithm





5. Testing and debugging

- 릴리즈 이전에 어떤 과정으로 다듬고 검증할 것인가?
  - 등장인물: Data engineer, Algorithm developer, Data business person, 결정권자
- 1) 알고리즘 성능 테스트
  - A/B testing, Quality control 부서 등에서 진행할 성능테스트의 기반 마련.
  - 간단하게 생각할 수 있는 Benchmark 알고리즘 구현하여 비교.
- 2) 시스템 테스트
  - 실제 시스템에 머지 되었을때 동작 체크, 최적화
- 3) 성능 또는 분석결과 보고
  - 이때 나온 성능과 결과들로 커뮤니케이션이 진행됨.
  - 나온 insight 들이나 visualize 된 결과들을 효과적으로 설명할 수 있어야함.

- 추가: 올바른 결과를 위해서.
  - 충분한 양의 Test 데이터 필요.
  - 비교할만한 (Benchmark) 알고리즘 필요.
  - 오류의 원인을 찾아서 다시 분석.

6. Release / Documentation

- Release 하면 끝? 사후 관리가 더 중요합니다.
  - 등장인물 : Data engineer, Algorithm developer
- Documentation 의 중요성
  - 시스템이 적용된 이후 관리 부서는 시스템을 개발한 부서와 달라질 수 있다.
  - 관리할 Engineer 들에게 알고리즘 및 분석의 세세한 이유와 관리 방법 전달이 필요함.
- 필요 설명서
  - 데이터 분석

    Analysis report, System documentation.
  - Machine learning

    Algorithm white paper 와, Code 설명서, System documentation

결론

- 1. 일단 역할에 맞는 능력은 있어야 합니다.
- 2. 모든일은 같이 해야합니다.
  - 특히 Algorithm 개발자는 Data engineer와 찰떡궁합을 이루어야 합니다.
- 3. 데이터 활용은 설득과 소통입니다.
  - Data visualization이 중요합니다.
  - Data business person 의 역할도 매우 중요합니다.
- 4. 인내심을 가지고 서로를 이해해야 합니다.
  - 프로세스는 성능이 나오거나 더 알게된 것들을 적용하기 위해 루프를 돌 수 있습니다. 모든 역할이 이를 이해해야 합니다.
- 5. 한번만 성공하면 다음일이 쉬워져 선순환고리를 만들 수 있습니다.

여담- 그럼 경영자는?

- 데이터에 기반한 의사결정 문화 도입
  - 데이터를 자산으로 여기는 자세가 필요합니다.
  - 데이터 기반 평가기준이 꼭 필요합니다.

- 다양한 배경과 능력을 가진 데이터 팀 구축
  - 한가지 역할만 있는 데이터 팀은 별로입니다.
  - 스스로 데이터 공부를 해서 Data business person 이 된다면 더 바랄게 없습니다.

- 결과의 빠르고 적극적인 활용
  - 데이터 – 분석결과 – 반영 – 데이터 – 분석결과 – 반영 의 순환고리 마련해야 합니다.

감사합니다.

혹시 Q&A?