

# 애자일 A/B 테스트 소개

A/B 테스트 개관

A/B 테스트 프로세스 소개

# Contents

1. 데이터 팀의 미션과 발전 단계
2. A/B Test란 무엇인가?
3. 왜 A/B Test를 하고 왜 애자일 A/B Test가 필요한가?
4. 전체적인 A/B Test 프로세스
5. A/B Test 분석을 위해 필요한 데이터
6. A/B Test 관련 문제들

# 데이터 팀의 미션과 발전 단계

데이터 팀의 일반적인 미션과 이상적인 발전 단계를 통해  
데이터 팀이 어떻게 회사의 발전/성장에 도움이 되는지  
살펴본다

# 데이터 조직의 미션은?

- 신뢰할 수 있는 데이터를 바탕으로 부가 가치 생성
  - Data is the new oil
  - But data can be a liability
    - 데이터의 잘못된 노출과 사용으로 인한 위험을 줄여야 함

# 데이터 조직이 하는 일 (1) - Decision Science

- 고품질 데이터 기반으로 의사 결정권자에게 입력 제공
  - 데이터를 고려한 결정(data informed decisions)을 가능하게 해줌
    - vs. 데이터 기반 결정(data driven decisions)
  - 예를 들면 데이터 기반 지표 정의, 대시보드와 리포트 생성 등을 수행



Data Literacy (데이터 문해력)

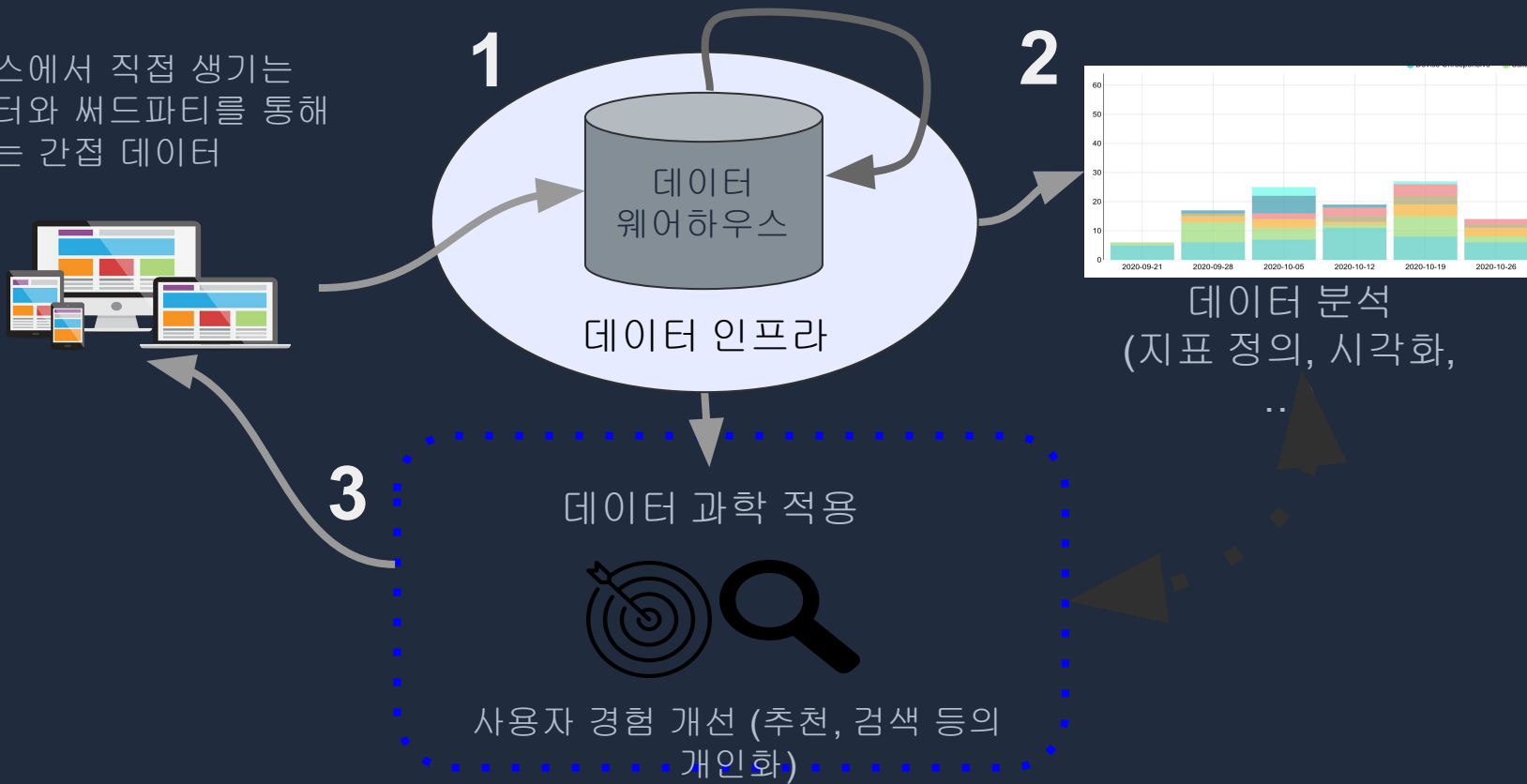
## 데이터 조직이 하는 일 (2) - Product Science

- 고품질 데이터를 기반으로 사용자 서비스 경험 개선 혹은 프로세스 최적화
  - 머신 러닝과 같은 알고리즘을 통해 사용자의 서비스 경험을 개선
    - 예) 개인화를 바탕으로 한 추천과 검색 기능 제공
    - 공장이라면 공정 과정에서 오류를 최소화 혹은 기기 고장 예측등을 수행



# 데이터의 흐름과 데이터 팀의 발전 단계

서비스에서 직접 생기는  
데이터와 써드파티를 통해  
생기는 간접 데이터



# A/B Test란 무엇인가?

A/B 테스트를 사용해야하는 경우와 사용하면 안되는 경우를  
구분하자



## A/B Test란 무엇인가?



[Image source: <https://www.invespcro.com/blog/what-is-ab-testing-split-testing/>]

- A/B 테스트 = 실험 (Split Test or Bucket Test)
  - Randomized Controlled Trial의 온라인 버전
- 다수의 Variant로 구성됨
  - 하나의 컨트롤 (기존 버전)과 하나 혹은 그 이상의 테스트

## A/B Test란 무엇인가? (1)

- 객관적으로 새로운 기능이나 변경을 측정/비교하는 방식
- 큰 위험없이 새로운 기능을 테스트하고 빠르게 배우는 방법

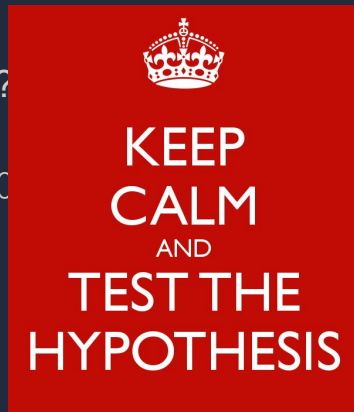


If You Can't Measure,  
You Can't Improve it!

Data driven decision vs. Data informed decision

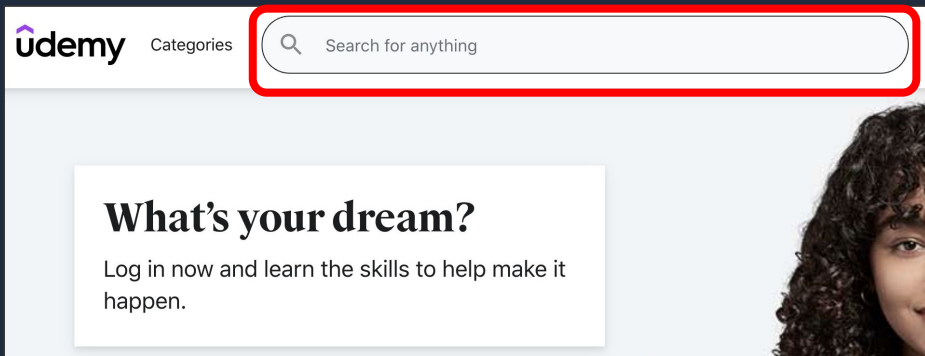
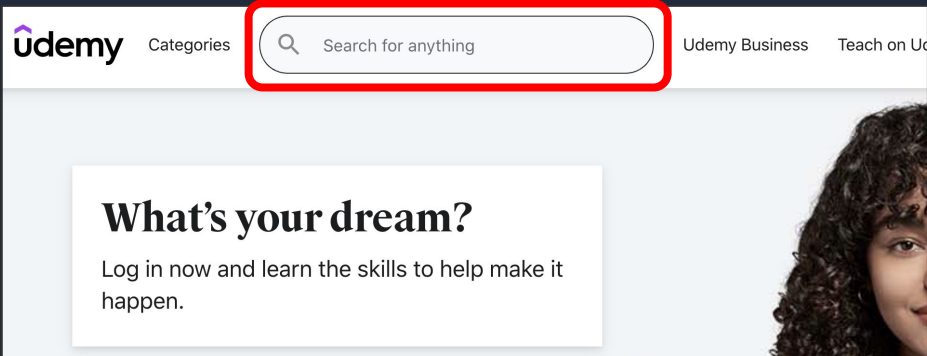
## A/B Test란 무엇인가? (2)

- 가설없는 A/B Test는 불가!
  - A/B test란 기본적으로 가설을 실험하고 검증하는 것
  - 예1. 새로운 추천방식이 기존의 추천방식보다 매출을 증대시키는가?
    - 어떤 지표에서 어느 정도의 임팩트가 예상되는가?
    - 가설을 나중에 결과에 비교하면서 생각지 못했던 다양한 배움
  - 예2. 상품 체크아웃 페이지의 스텝을 줄이면 결제가 더 올라가는가?
    - 스텝을 줄이면 정말 매출이 올라갈까?
    - 사용자 관점과 개발자 관점은 굉장히 다를 수 있음



## A/B Test란 무엇인가? (3)

- 보통 프로덕션 환경에서 2개 혹은 그 이상의 버전(Variants)을 비교
  - 베이스라인 버전 (“control”) vs. 하나 혹은 그 이상의 테스트 버전 (“test”)
    - “control”: 현재 버전
    - “test”: 새 버전
  - 보통 서비스내의 다른 영역을 테스트하는 A/B Test들은 독립적이라 생각하고 다수의 AB Test를 동시 실행하는 것이 일반적
    - 하지만 상호작용(Interactions)이 있을 수 있음



## A/B Test를 사용하면 안되는 경우는?

- No Data No A/B Test
- 버그 수정을 임팩트를 측정하는 경우
  - 빨리 고치는 것이 좋음
- 아직 구체적이지 않은 아이디어 테스트
  - A/B test isn't cheap and affects real traffic
  - You will have to do offline testing (user survey for example)
  - Fake door testing
- 가설없이 굉장히 랜덤한 아이디어 테스트
- 비교대상없이 굉장히 새로운 기능 테스트



# 왜 A/B Test를 하고 왜 애자일 A/B Test가 필요한가?

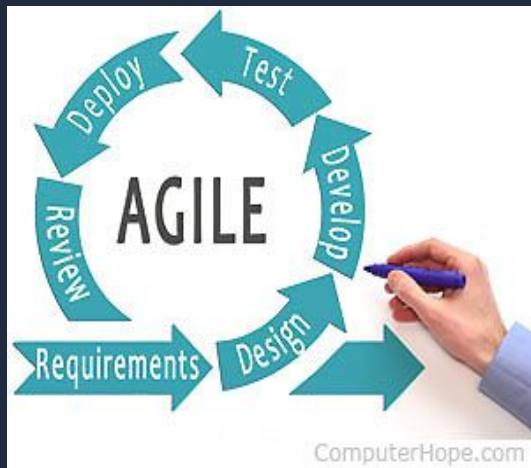
A/B 테스트가 왜 필요한지 그리고 A/B 테스트가 왜  
애자일해야하는지 이야기해보자

## 왜 A/B Test를 하는가?

- 비즈니스 관련 지표를 개선되는지 객관적으로 측정하기 위함
  - 가설 기반의 실제 사용자 대상 비교
- 위험을 최소화하기 위함
  - 아무리 사용자 설문등이 좋아도 실제 사용자들이 어떻게 반응할지는 알 수 없음
  - 처음에는 작은 퍼센트의 사용자들에게만 새 기능을 노출시키고 문제가 없으면 퍼센트 증가

## 왜 A/B Test는 애자일해야 하는가? (1)

- How to Measure A/B Test is important
  - but what if it takes **too long**?
- “Agile” in this case means days / weeks
  - instead of months of A/B test
  - agile A/B test configuration +  
agile A/B test analytics +  
agile ML model deployment
    - not relevant for pure front end experiments





## 왜 A/B Test는 애자일해야 하는가? (2)

- Agile A/B Test = Agile Product Improvement = Agile Company
  - If on average it takes 6 iterations to get to a launch, and each iteration takes 1 month, then it takes about a 1/2 year to launch something
- Reduction of the experiment time => faster launch time
  - For ML based A/B test, automating the whole process is critical



# 전체적인 A/B Test 프로세스

전체적인 A/B 테스트 과정을 살펴보자

# Overall A/B Test Process

- A/B Test Proposal & Approval
  - One pager with hypothesis
    - Why, Expected Impact, Confidence Level, MVP to Start, Potential Issues, Owner (Implementation, QA and Analysis)
  - Discussion with stakeholders
- Implementation and QA
- Rollout (in the next deck)
- Iterations
- Periodic Review (Weekly Experiment Review meeting)

# A/B Test Rollout Phases

## 1. Smoke test (~few days)

- 0-1% for test variant(s) to verify that nothing is broken
- validate that test variant is setup correctly

## 2. Initial ramp (~1 week)

- 5-10% for test variant(s)
- sizing depends upon any revenue (or KPI) concerns

## 3. Intermediate ramp (~few weeks)

- 25%-50%

## 4. Final ramp-up / launch

- 100% and then full launch for winning test variant

Staged Rollout Feature in  
Google Play!


# A/B Test Configuration (1)

- 코딩없이 A/B 테스트를 진행가능하게 하는 것이 목표
  - 자주 하는 A/B 테스트들은 템플릿화가 가능함
- A/B Test Configurations (뒤에서 더 설명)
  - Hashing parameter (e.g. userid, deviceid)
  - Bucket size (% of traffic)
- 보통 테스트하는 기능을 백엔드단의 **flag**로 관리하는 것이 일반적

## A/B Test Configuration (2)

- An UI to change A/B Test Parameters
  - Without code change
  - For example, bucket size change and Start/Stop of an A/B Test
  - keeping history of changes for audit purpose is important

Links to A/B Test dashboards

A/B Test Admin							
ID	Name	Page	Variants	Starts	Active	Owner	
100	Search V1	Search	200 201	11/01/2017	Yes	Dave	<u>Link</u>
101	Reco V2	Home	300 301 302	11/07/2017	No	Dan	<u>Link</u>

# A/B Test 분석을 위해 필요한 데이터

A/B 테스트 분석을 하기위해 필요한 데이터들을 알아보자

## A/B 테스트 분석을 위해 필요한 정보

### 1. 사용자별 A/B 버킷 정보

- a. 누가 A에 들어갔고 B에 들어갔는지

### 2. 사용자별 행동 정보

- a. 어떤 아이템들을 보았고
- b. 어떤 아이템들을 클릭했고
- c. 어떤 아이템들을 구매했는지
- d. ...



- 1과 2의 정보를 조인
- A와 B로 그룹핑하여 그룹간 통계 정보 계산 (매출액 등등)



# Which Users were in A (Control) and in B (Test)!

- Experiment's user bucketing info is what you want!
  - 누가 A에 들어갔고 B에 들어갔는지 알아야하며 이 정보가 로깅되어야 함
- It can be represented as tables
  - Experiment ID
  - Variant ID
  - Timestamp
  - User ID
  - ...

# User Behavior Data - Funnel Data

- For **each user** (or device)

- What was seen?
- What was clicked?
- What was purchased?
- How much was it?
- Was it consumed?

Funnel Data



Impression

Click to Detail Page

Purchase

Consumption

중요지표는  
아니지만 지표  
해석이나  
디버깅에 도움이  
됩니다

# Funnel Data: Impression vs. Click

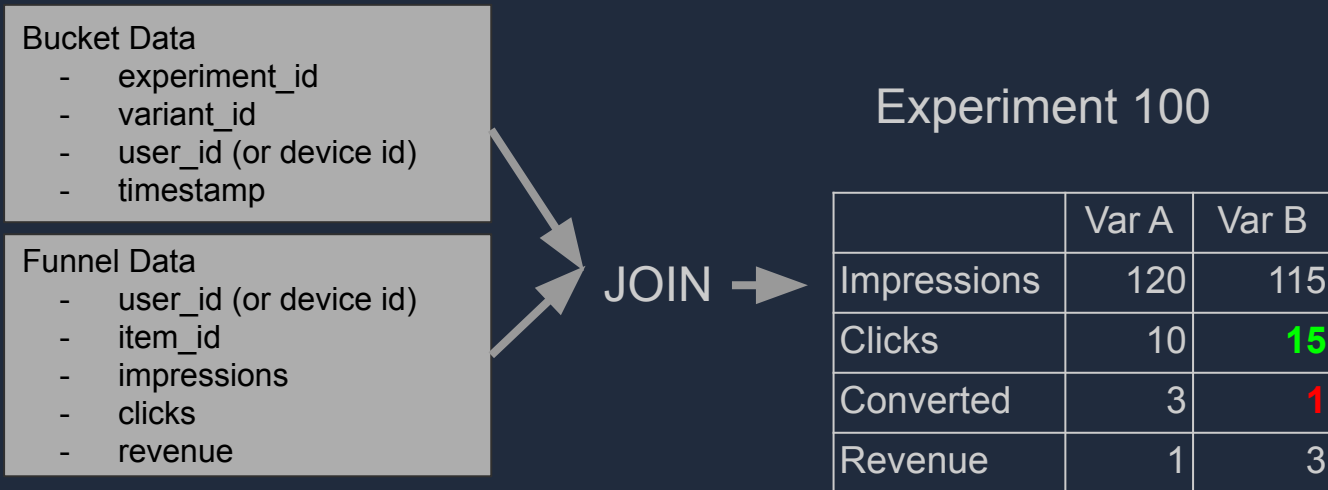
Course Title	Author	Rating (Stars)	Reviews	Original Price	Current Price
iOS 11 & Swift 4: From Beginner to Paid...	Mark Price, Developes by Mark P...	4.5	3,643	\$199.99	\$11.99
Mac Numbers: Creating and Using Spreadsheets...	Gary Rosenzweig	4.7	374	\$79.99	\$11.99
Mac Pages: Mastering Apple's Word Processing...	Gary Rosenzweig	4.3	297	\$79.99	\$11.99
How to Edit Your Videos: Official Udemmy Course	Udemmy Instructor Team	4.4	1,619		Free
Crash Course: Fundamentals Of HTML ...	Tech Lemur	4.9	261	\$19.99	\$11.99
iOS 10 & Swift 3: From Beginner to Paid...	Mark Price, Developes by Mark P...	4.0	13,277	\$199.99	\$11.99

Impressions

A user clicks the last icon

# Analysis of A/B Test Result (E-Commerce)

- Experiment Data와 Funnel Data를 조인하기 (Transform => ELT)
  - Funnel의 맨위부터 A와 B간의 숫자를 비교 가능
  - 여기에 사용자의 메타 정보를 추가하면 다양한 분석이 가능
- 보통 시니어 데이터 분석가가 이 분석을 하게 됨



# A/B Test Analysis in BI Dashboard

Age	All
Gender	All
Area	All
Context	All
Channel	All

	A (Control)	B (Test)
User Size(*)	50	51
Impressions(**)	120	115
Clicks	10	15
Converted	3	1
Revenue	1	3

(\*) User Size에서는 보통 둘의 크기가 통계적으로 동일하기를 바라며 그게 아니라면 테스트 설정에 무엇인가 잘못이 있음을 나타냄

(\*\*) 만일 새로운 기능이 Impressions의 수를 줄이는 영향이 있는 것이 아니라면 (\*)과 마찬가지로 통계적으로 동일해야 한다. 즉 다르다면 뭔가 실험자체에 문제가 있음을 나타낸다



# A/B Test 관련 문제들

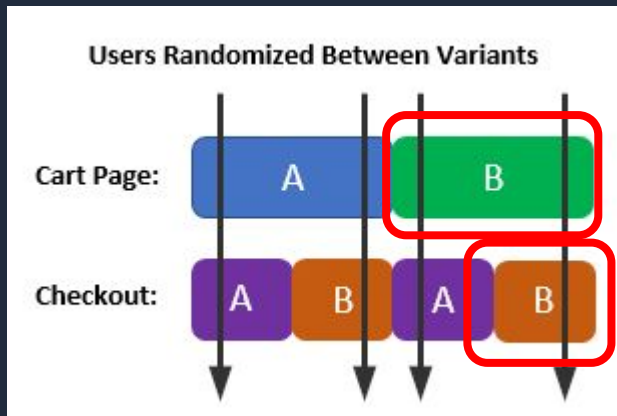
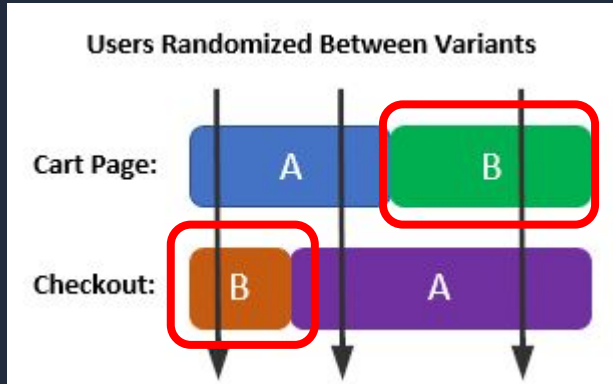
A/B 테스트에서 생길 수 있는 문제들을 살펴보자

## A/B Test시 발생하는 문제들 (1)

- 어떤 결정은 데이터로 판단할 수 없음 => Data Informed Decision
- 가설없이 혹은 대충 쓴 가설로 A/B Test를 하는 경우
  - Don't abuse A/B Test for everything
- 분석에 필요한 데이터 품질이 낮은 경우
  - Any bugs in A/B test or Funnel will lead to quality issues
  - Is sample size big enough?
  - Can bucketing be done without skewness or bias
- 결과를 선입견없이 객관적으로 분석하지 못하는 경우
  - Don't interpret to your advantage. Usually it is a good idea to wait at least a week
  - The results need to be interpreted in a **group setting** for knowledge sharing, collective learning and objective analysis

## A/B Test시 발생하는 문제들 (2)

- **Interactions (상호작용) between A/B Tests -> Multivariate Tests**
  - When you run multiple A/B tests, there can be interactions between them
    - Interactions are not easy to catch
    - The safest way is to run a test at a time but it can take too long



Source: <https://blog.analytics-toolkit.com/2017/running-multiple-concurrent-ab-tests/>



## A/B Test시 발생하는 문제들 (3)

- 데이터 인프라 비용
  - A/B Test Analysis Pipeline is very expensive usually (70% of Airbnb's data infra usage)
- 비교 대상이 하나가 아닌 경우 (Not Comparing Apple to Apple)
  - Testing more than one change at a time
    - Testing a UI change and a new recommendation at the same time
  - Control users and Test users are not the same
- 얼마나 지켜보고 결정을 내릴 것인지?
  - How long do we want to wait before determining A/B test is a success or not?

# Q & A

- 퀴즈: <https://forms.gle/73pZdL3cimBLvyDA7>