DEVIEW 2021

네이버서치ABT: 신뢰할 수 있는 A/B 테스트 플랫폼 개발 및 정착기

김진영

DATA & ANALYTICS @ NAVER SEARCH

발표자 (김진영) & 개발팀 소개

- 네이버 서치의 Data & Analytics팀(*) 리더 (AB Test / Causal Inference / Crowdsourcing)
- 네이버 서치 미국 오피스 1호 (Co-Director)
- 과거 Microsoft와 Snap에서 검색 및 추천 시스템의 평가 업무
- SIGIR / WWW / WSDM등에 관련 논문 & 튜토리얼 진행

*** 한국/미국에서 Data Scientist & Engineer 채용중입니다! (jin.y.kim@navercorp.com)**

네이버 서치 ABT 개발팀

Leadership
김재헌 / 김진영

Platform Dev
전용우 / 서동유

Search Server
하한누리 / 이원진

Project Management
조한나 / 최수연

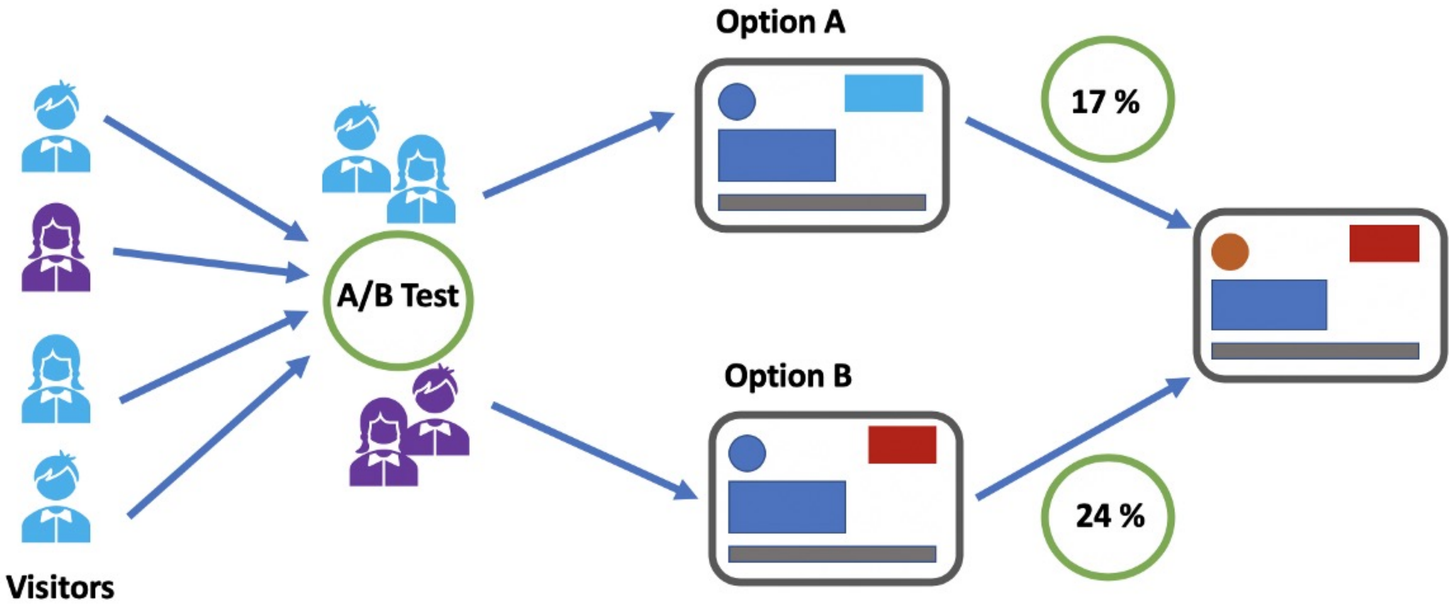
Data Pipeline & Analysis
박형애 / 박동현

CONTENTS

A/B 테스트:
보기보다 어려운 이유

A/B 테스트의 기본 개념은 단순함

1. 사용자를 무작위로 나누어 실험군(A/B)을 설정한다
2. 실험군 별로 다른 서비스를 제공하고 반응을 기록한다
3. 사용자의 반응을 서로 비교하여 결론을 도출한다



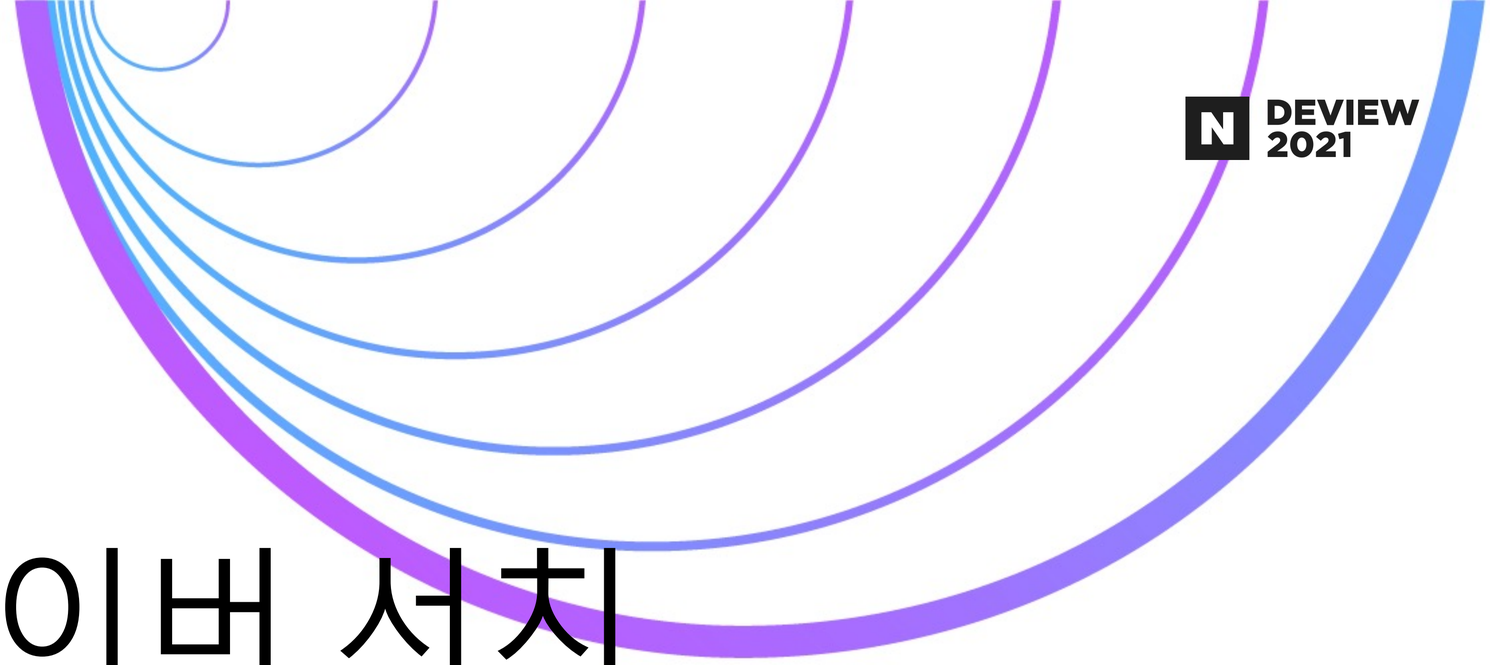
이미지 출처: Medium

A/B 테스트 과정에 존재하는 다양한 함정

1. 사용자군 설정이 완전 무작위로 이루어지지 않음
2. 실험군에 따른 분기가 제대로 이루어지지 않음
3. 사용자 로깅이 일부 누락되거나 오류가 발생
4. 의사결정 기준 지표가 사전에 충분히 정의되지 않음
5. 분석 결과에 대한 유의성 검증이 제대로 되지 않음
6. 실험을 너무 일찍 중단하고 의사결정을 내림

현대적인 A/B 테스트 플랫폼의 조건?

1. (앞서 언급한) 실험의 오류가능성을 사전에 최소화
2. 여러 팀/프로젝트에서 동시에 간섭없이 실험을 진행
3. 개별 실험의 버전업을 통해 여러 차례 반복 개선을 지원
4. 사용자 지표에 악영향을 주는 실험을 자동 경고 & 셧다운
5. 실험 결과에 대한 대시보드와 Custom Analysis 템플릿 지원



DEVIEW 2021

네이버 서치
ABT 프로젝트 소개

Motivation: 네이버 검색을 위한 현대적인 ABT

AS-IS: 개별 팀에서 ABT를 독자적으로 진행

전체적인 사용자 영향의 확인이 어려움

실험 결과의 검증 및 고급 분석의 어려움

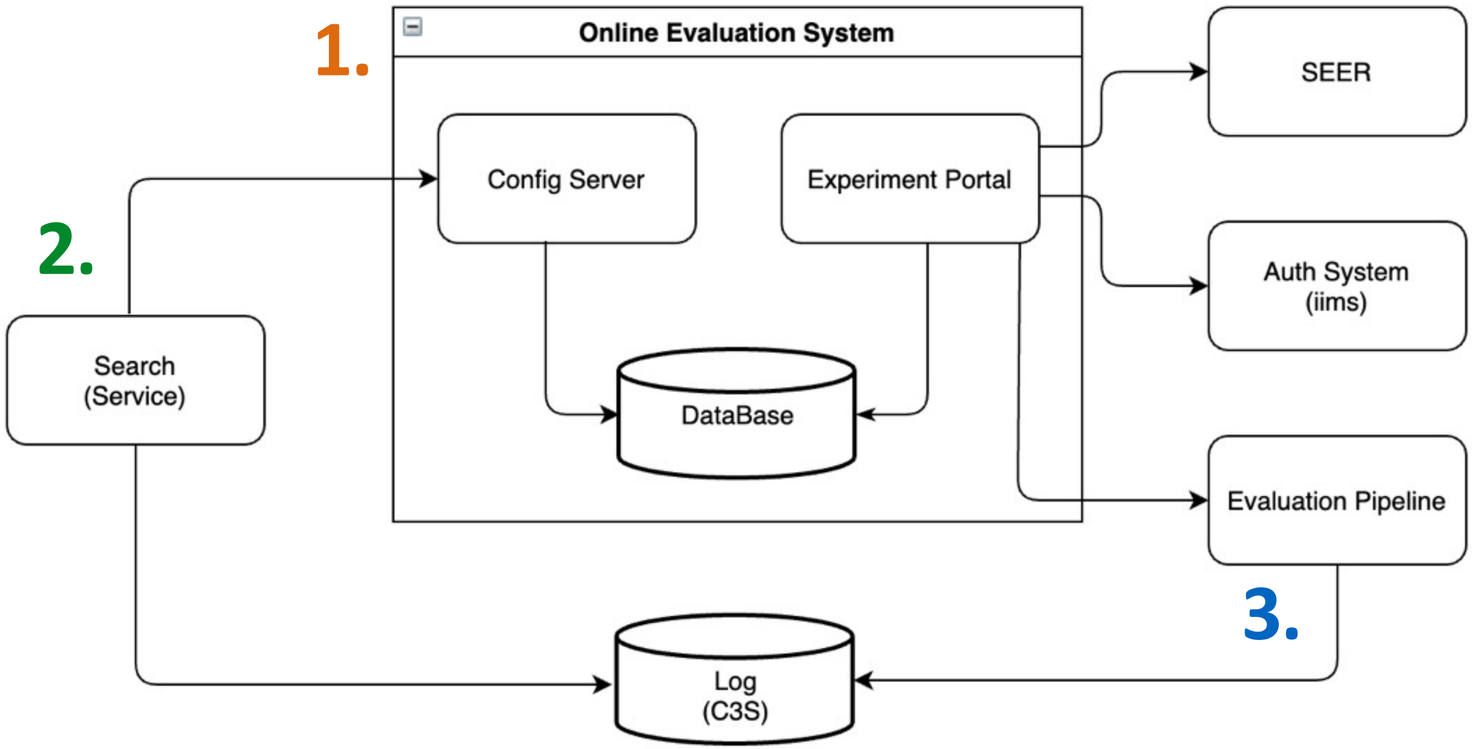
TO-BE: 네이버 검색 전체를 위한 현대적인 ABT 개발

실험 설계/수행/분석의 일원화 (Data & Analytics팀)

모든 팀이 같은 지표/프로세스를 통해 의사결정

네이버 검색 ABT 아키텍쳐

1. 실험 관리자가 등록/변경한 실험 설정 정보를 제공하는 OES 서버
2. 검색 서버에서 OES의 버킷 정보를 받아 서비스 분기 및 로깅
3. 검색 로그를 정제하여 실험 결과를 생성하는 파이프라인



네이버 검색 ABT: 실험 생성 UI

DEVIEW 2021

Create Experiment

layer*

선택

state

enable

name*

owner

start date*

mm/dd/yyyy, --:-- --

end date*

mm/dd/yyyy, --:-- --

traffic assign(%)*

salt key

실험에 속한 사용자 안에서 variant에 속한 사용자를 선정할 때 key로 사용됩니다. 자동 생성됩니다.

description

네이버 검색 ABT: 실험 설정 UI

DEVIEW 2021

List

enable

Id		Name	Modify
1	통합 검색 UX		Modify

Id	Version	Traffic	Name	Modify	Version Up	State
8	v1	12%		Modify	Version up	enable

Id		Name	Modify
2	쇼핑 검색		Modify

Id	Version	Traffic	Name	Modify	Version Up	State
2	v1	9%		Modify	Version up	enable
7	v1	9%		Modify	Version up	enable

Id		Name	Modify
3	이미지 검색		Modify

Id	Version	Traffic	Name	Modify	Version Up	State

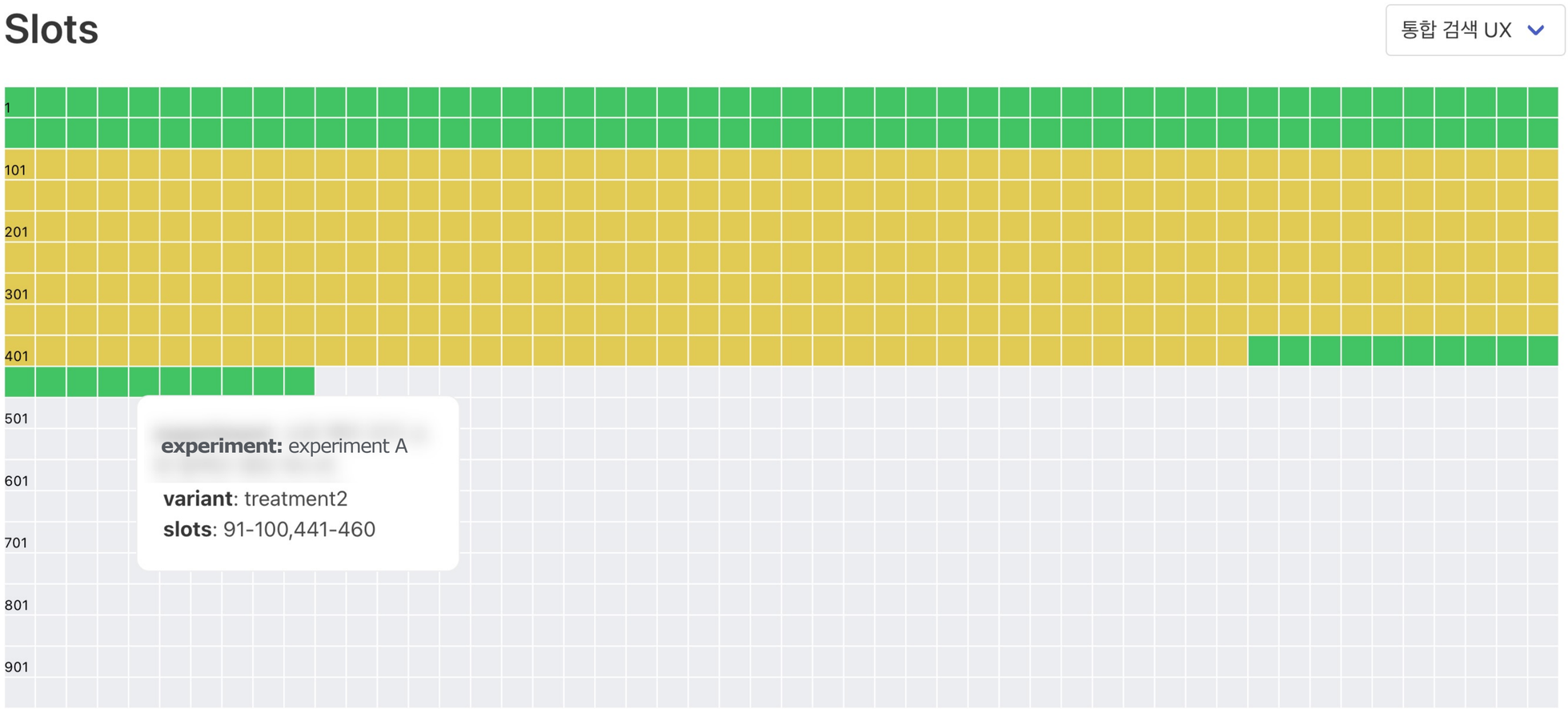
Id		Name	Modify
4	평가전용		Modify

Id	Version	Traffic	Name	Modify	Version Up	State

네이버 검색 ABT: 실험 설정 UI

DEVIEW 2021

레이어 별로 개별 실험에 할당된 사용자 트래픽을 시각화

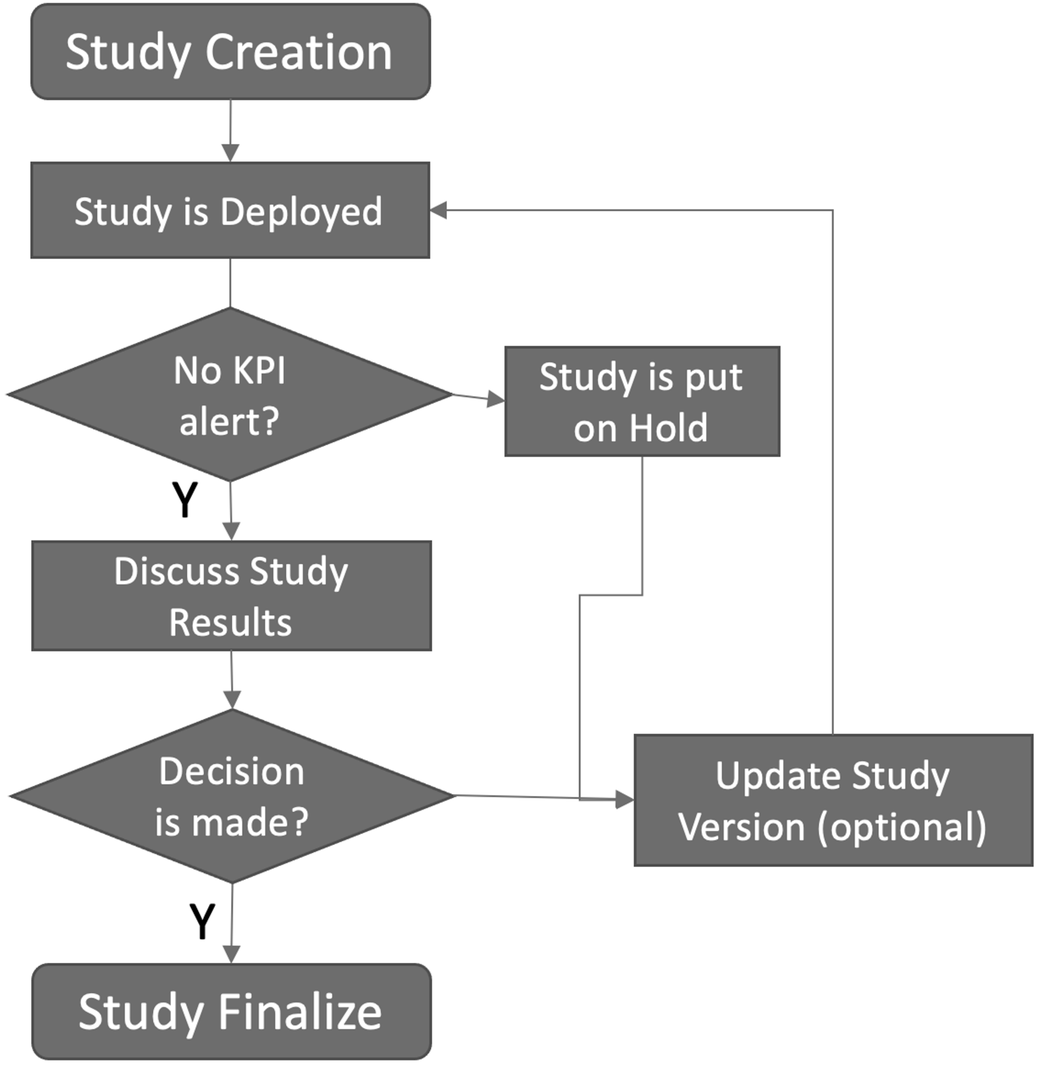


네이버 검색 ABT: 실험의 Lifecycle

실험 버전업의 필요성
대부분의 실험은 여러 iteration을 거침

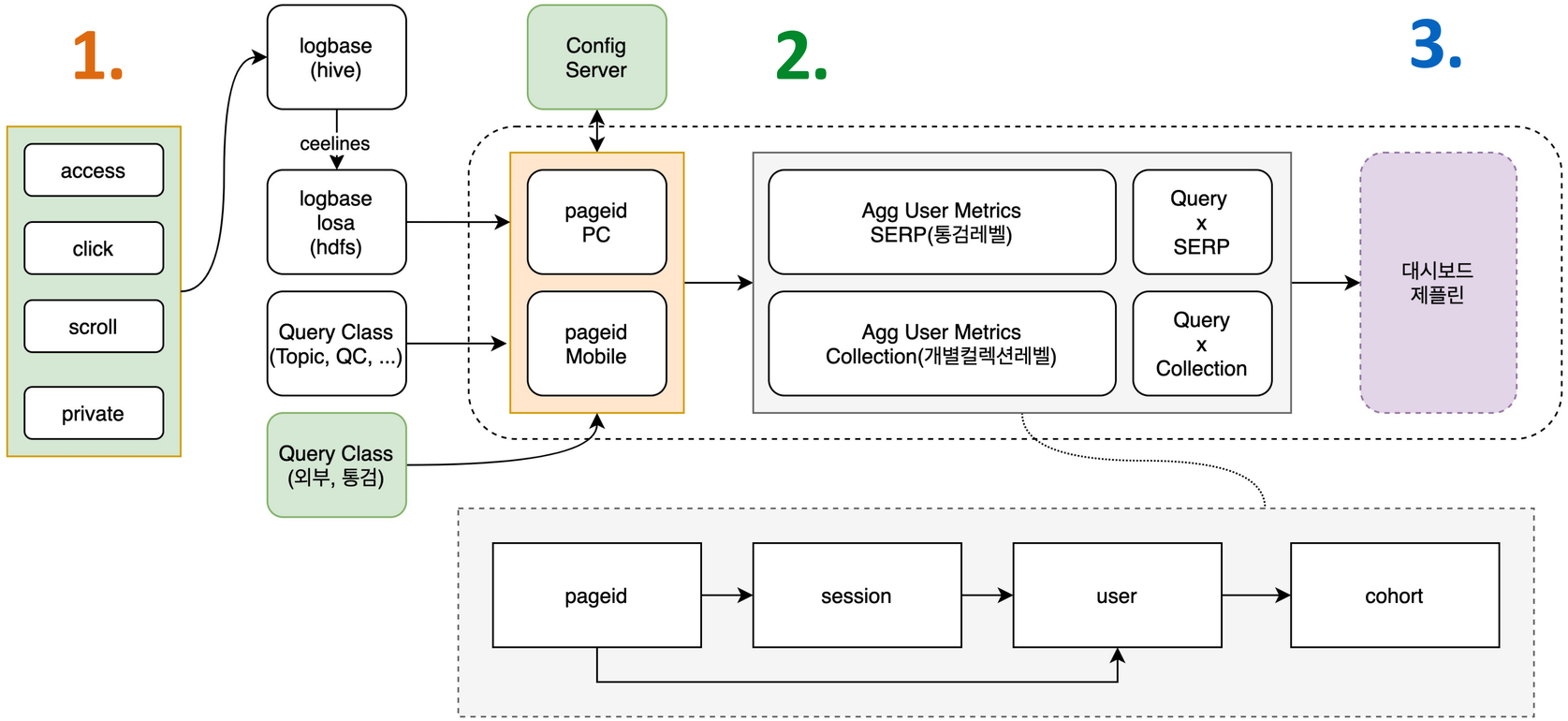
실험 버전업의 조건
실험군의 설정 혹은 구현이 바뀌는 경우
트래픽 재분배(re-shuffle)가 필요한 경우

실험 버전업의 프로세스
보통 담당자들이 회의를 거쳐 의사결정



네이버 검색 ABT: 실험 분석 파이프라인

1. 사용자 로그 및 메타데이터가 수집되어 저장
2. 대용량 데이터 파이프라인에서 기본적인 정제 작업
3. 실험군별 지표가 계산되어 대시보드 및 노트북으로 결과 확인



네이버 검색 ABT: 실험 진행 프로세스

실험을 위해 서비스 담당자 및 실험 담당자가 협업

실험 요청 / 구현 / 테스트 / 배포 / 리포팅의 순서

주간 회의를 통해 계획/진행중인 실험에 대한 논의

런치 2달만에 여러 파트너 팀과 두자리수 실험 진행

실험 준비에서 완료까지 보통 3-4주 가량 소요

네이버 서치 ABT:
플랫폼 신뢰도 검증하기

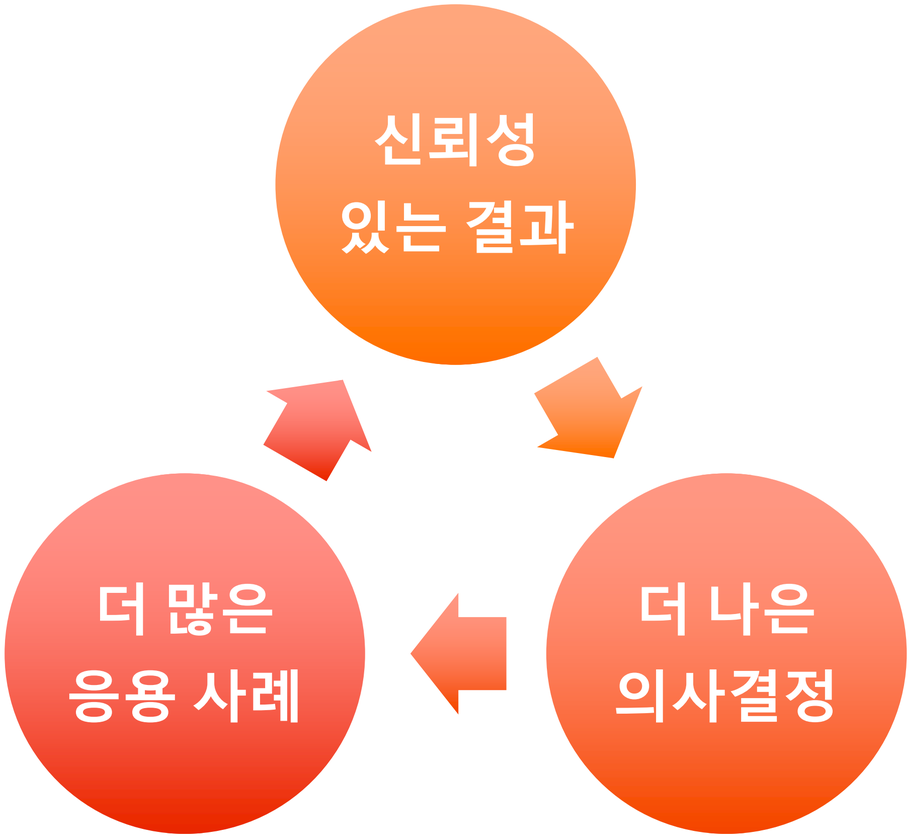
네이버 검색 ABT: 플랫폼 신뢰도 검증하기

실험 레이어간 결과의 간섭은 없을까?
- 실험 레이어 간 트래픽 배분 비율 검증

개별 실험의 트래픽은 제대로 분배되었을까?
- 실험 내부의 트래픽 배분 비율 검증

실험 트래픽 배분은 무작위(random)로 되었을까?
- Large-scale A/A 테스트 수행 결과

같은 실험은 반복하면 같은 결과가 나올까?
- 재현(replication) 실험을 통한 검증



신뢰성있는 플랫폼이 만드는 선순환

실험 레이어간 트래픽 교차 분배 검증

목표: 서로 다른 레이어의 실험은 트래픽 교차 배분을 통해 독립성 보장

실험A (Layer1)와 실험B에 (Layer2) 할당된 사용자의 교차 배분 비율을 비교

Chi-Square Test의 p-value를 사용해 이상 여부를 검증

실험B의 Variants

실험A의 Variants	1 (p-value: 0.09)		2 (p-value: 0.69)		3 (p-value: 0.05)	
	실험B의 유저수	실험A의 유저수	실험B의 유저수	실험A의 유저수	실험B의 유저수	실험A의 유저수
4	853558	34144	853577	34267	851643	34483
5	853558	25257	853577	25535	851643	25670
6	853558	25631	853577	25568	851643	25638

실험군별 트래픽 분배 정합성 검증

DEVIEW 2021

목표: 실험군별 트래픽이 실험 설정에 맞게 분배되었는지 확인

날짜별/전체 기간 동안 실험군별 사용자 수의 비율을 비교

Chi-Square Test의 p-value를 사용해 이상 여부를 검증 [관련논문]

settings

Date	A	B	A'	Total	A ratio(%)	B ratio(%)	A' ratio(%)	P-Value
20210909	682188	682487	681654	22740781	3.0	3.001	2.997	0.749
20210910	864872	864470	863220	28789043	3.004	3.003	2.998	0.267
20210911	853878	854669	852258	28444572	3.002	3.005	2.996	0.151
20210912	853819	853632	852220	28421095	3.004	3.004	2.999	0.221
20210913	868453	868183	867608	28912576	3.004	3.003	3.001	0.342
20210914	861072	861959	859960	28672418	3.003	3.006	2.999	0.095
20210915	856126	856282	855474	28505358	3.003	3.004	3.001	0.263
20210916	608044	608763	606333	20236127	3.005	3.008	2.996	0.029

settings

Date	A	B	A'	Total	A ratio(%)	B ratio(%)	A' ratio(%)	P-Value
20210909 ~ 20210916	1604915	1605627	1604269	53462206	3.002	3.003	3.001	0.257

Large-scale A/A 실험 분석을 통한 검증

목표: 서로 동일한 실험군에 대해서 실제로 지표 차이가 없는 것을 증명하자

- 서로 동일한 실험군 100개를 (A01~A100) 갖는 실험을 생성
- 주요 지표에 대해 실험군간에 유의미한 차이가 있는지를 검증
- 결과: P-value가 기준치 이하로 나타난 실험군은 5% 미만

Average of sess	Column Labels									
Row Labels	A11	A12	A13	A14	A15	A16	A17	A18	A19	Grand Total
A12	0.073									0.073
A13	0.943	0.084								0.5135
A14	0.303	0.449	0.336							0.362666667
A15	0.31	0.431	0.344	0.981						0.5165
A16	0.228	0.55	0.256	0.869	0.849					0.5504
A17	0.466	0.283	0.51	0.757	0.773	0.633				0.570333333
A18	0.322	0.417	0.357	0.961	0.98	0.83	0.793			0.665714286
A19	0.311	0.433	0.345	0.982	0.999	0.851	0.773	0.979		0.709125
A20	0.515	0.251	0.561	0.701	0.716	0.58	0.939	0.735	0.716	0.634888889
Grand Total	**0.385666667**	**0.36225**	**0.387**	**0.8752**	**0.8634**	**0.7235**	**0.835**	**0.857**	**0.716**	**0.588333333**

Average of qc	Column Labels									
Row Labels	A11	A12	A13	A14	A15	A16	A17	A18	A19	Grand Total
A12	0.044									0.044
A13	0.782	0.079								0.4305
A14	0.114	0.653	0.188							0.318333333
A15	0.239	0.393	0.364	0.683						0.41975
A16	0.315	0.313	0.461	0.571	0.871					0.5062
A17	0.232	0.397	0.355	0.691	0.99	0.861				0.587666667
A18	0.216	0.43	0.332	0.731	0.949	0.822	0.959			0.634142857
A19	0.137	0.597	0.222	0.933	0.748	0.632	0.756	0.797		0.60275
A20	0.063	0.863	0.11	0.78	0.491	0.398	0.497	0.533	0.718	0.494777778
Grand Total	**0.238**	**0.46563**	**0.290285714**	**0.7315**	**0.8098**	**0.67825**	**0.737333333**	**0.665**	**0.718**	**0.518**

Average of cc	Column Labels									
Row Labels	A11	A12	A13	A14	A15	A16	A17	A18	A19	Grand Total
A12	0.02									0.02
A13	0.182	0.314								0.248
A14	0.3	0.197	0.771							0.422666667
A15	0.091	0.549	0.698	0.503						0.46025
A16	0.108	0.474	0.775	0.567	0.916					0.568
A17	0.101	0.518	0.734	0.534	0.963	0.954				0.634
A18	0.094	0.513	0.725	0.523	0.966	0.949	0.996			0.680857143
A19	0.188	0.316	0.996	0.777	0.697	0.773	0.732	0.724		0.650375
A20	0.009	0.794	0.201	0.119	0.389	0.326	0.364	0.357	0.204	0.307
Grand Total	**0.121444444**	**0.45938**	**0.7**	**0.5038**	**0.7862**	**0.7505**	**0.697333333**	**0.541**	**0.204**	**0.511133333**

재현(replication) 실험을 통한 검증

목표: 결과의 재현성을 확인하여 ABT 플랫폼 및 실험의 신뢰성을 검증

- 특정 실험에 대해 동일한 조건에서 서로 다른 시기에 2회 반복 실시
- 1회와 2회 결과를 비교하여 큰 경향성이 거의 유사함을 확인
- P-value 가 유의성 임계치 (0.05) 근처에 있는 지표의 경우 결과가 재현되지 않음

	V2 Results		V1 Results	
Metric	**%Delta A/B**	**P-Value A/B**	**%Delta A/B**	**P-Value A/B**
metric 1	0.106	0.287	-0.154	0.112
metric 2	-0.059	0.654	-0.267	0.038
…	0.047	0.754	-0.166	0.242
	0.079	0.529	-0.216	0.077
	-0.033	0.809	-0.288	0.025
	0.02	0.693	0.002	0.971
	0.157	0.189	0.135	0.216
	-0.486	0.057	-0.186	0.441
	-4.586	0	-3.713	0
	3.729	0	3.04	0
	2.917	0	2.178	0
	-4.076	0	-3.296	0
metric n	-0.358	0	-0.509	0

네이버 서치 ABT: 실험 분석 방법론

네이버 서치 ABT: 실험 분석 방법론 개요

DEVIEW 2021

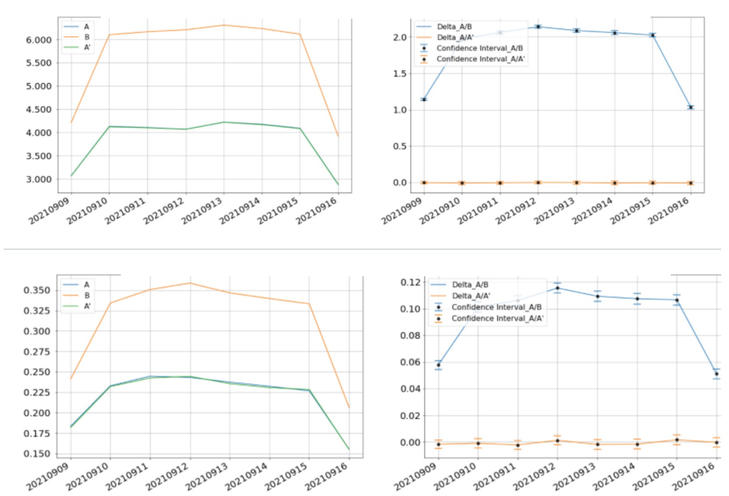
실험이 전체적으로
어떤 결과를 보이는지?

이 결과가 시간이
지나도 바뀌지 않을지?

이 결과의 원인을 어떻게
설명할 수 있는지?

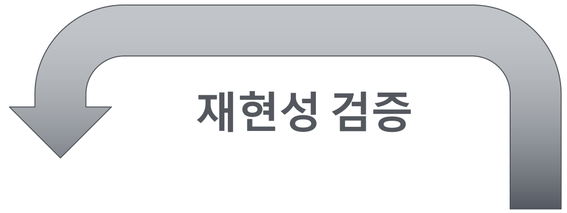
실험 전체 지표 → **일별 지표 트랜드** → **부문별 세부 지표**

	V2 Results		V1 Results	
Metric	%Delta A/B	P-Value A/B	%Delta A/B	P-Value A/B
metric 1	0.106	0.287	-0.154	0.112
metric 2	-0.059	0.654	-0.267	0.038
...	0.047	0.754	-0.166	0.242
	0.079	0.529	-0.216	0.077
	-0.033	0.809	-0.288	0.025
	0.02	0.693	0.002	0.971
	0.157	0.189	0.135	0.216
	-0.486	0.057	-0.186	0.441
	-4.586	0	-3.713	0
	3.729	0	3.04	0
	2.917	0	2.178	0
	-4.076	0	-3.296	0
metric n	-0.358	0	-0.509	0



- 사용자군별 지표
- 질의군별 지표
- 결과 유형별 지표

실험 전체 지표 해석하기



- 서로 다른 단위를 가지는 지표간의 비교를 위해 절대적인 차이를 상대적인 차이(%Delta)로 환산
- 이런 차이의 통계적인 유의성을 확인하기 위해 p-value를 계산
- 모든 실험에 대조군 2개(A와 A')를 설정하여 A/A'테스트를 수행

	V2 Results			V1 Results		
Metric	%Delta A/B	P-Value A/B	P-Value A/A'	%Delta A/B	P-Value A/B	P-Value A/A'
metric 1	0.106	0.287	0.867	-0.154	0.112	0.057
metric 2	-0.059	0.654	0.902	-0.267	0.038	0.185
...	0.047	0.754	0.749	-0.166	0.242	0.359
	0.079	0.529	0.986	-0.216	0.077	0.183
	-0.033	0.809	0.871	-0.288	0.025	0.193
	0.02	0.693	0.83	0.002	0.971	0.333
	0.157	0.189	0.895	0.135	0.216	0.953
	-0.486	0.057	0.249	-0.186	0.441	0.138
	-4.586	0	0.125	-3.713	0	0.643
	3.729	0	0.013	3.04	0	0.879
	2.917	0	0.473	2.178	0	0.23
	-4.076	0	0.105	-3.296	0	0.5
metric n	-0.358	0	0.039	-0.509	0	0.215

A/A 검증

[DeepDive] False Discovery Rate 보정

- 일반적인 AB테스트 환경에서는 수백/수천개의 지표에 대해서 동시에 통계적 유의성을 테스트
- 이는 개별 결과에 대해 우연히 유의성을 발견하게 될 확률 (False Positive Rate)을 높임
- 네이버서치ABT에서는 이런 잘못된 발견의 비율 (False Discovery Rate)을 통제하기 위해 p-value를 오른쪽과 같이 보정 (Work in Progress)

$$False\ Discovery\ Rate\ (FDR)$$

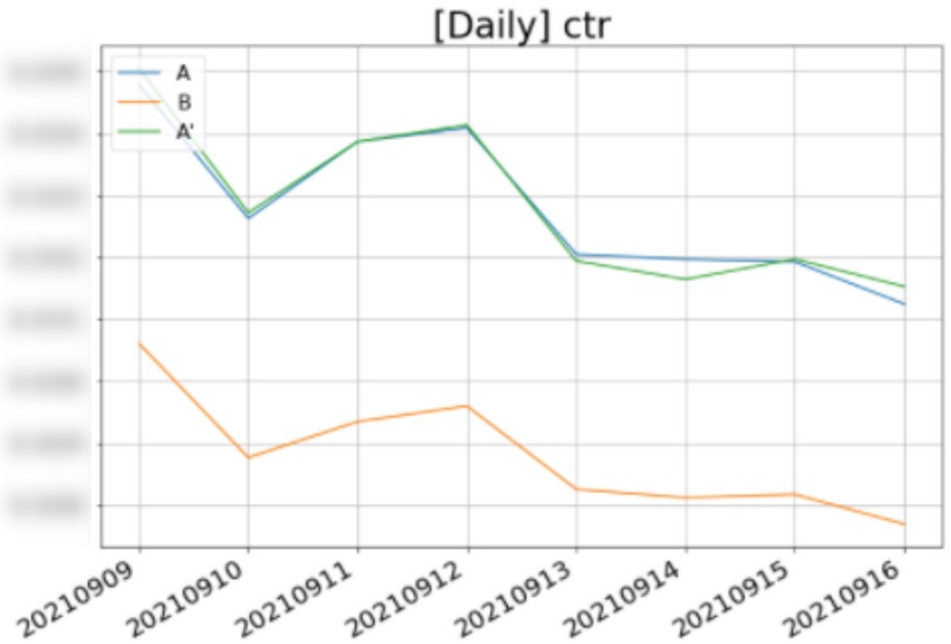
$$= E\left(\frac{\#FalsePositive}{\#FalsePositive + \#TruePositive}\right)$$

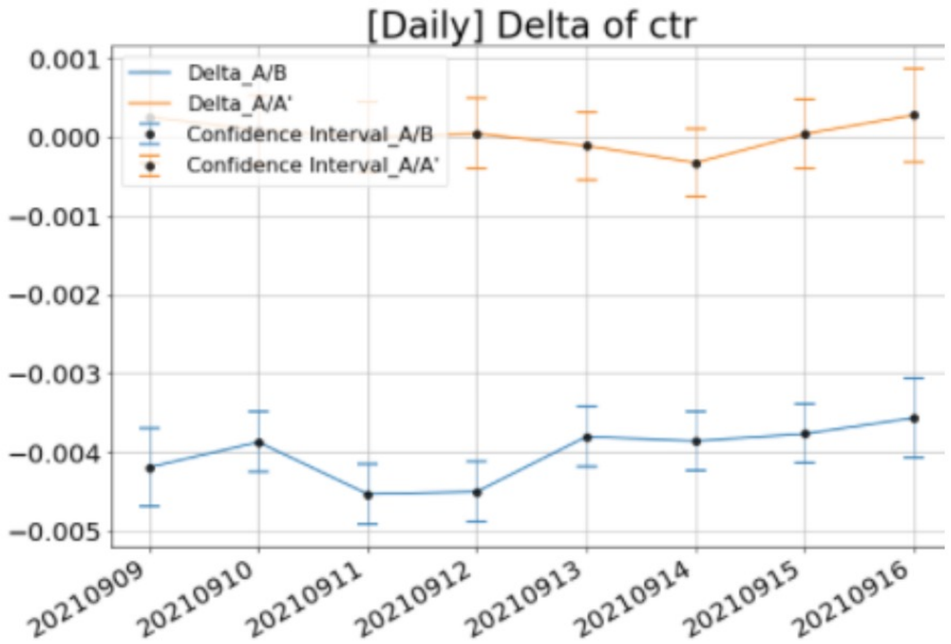
Original p-value (largest first)	Adjusted p-value
$p_{(n)}$	$p_{(n)}$
$p_{(n-1)}$	$\min(p_{(n)}, p_{(n-1)} * N/(N-1))$
$p_{(n-2)}$	$\min(p_{(n-1)}, p_{(n-2)} * (N-1)/(N-2))$
…	…

Computing FDR-adjusted p-value

일별 지표 트렌드 해석하기

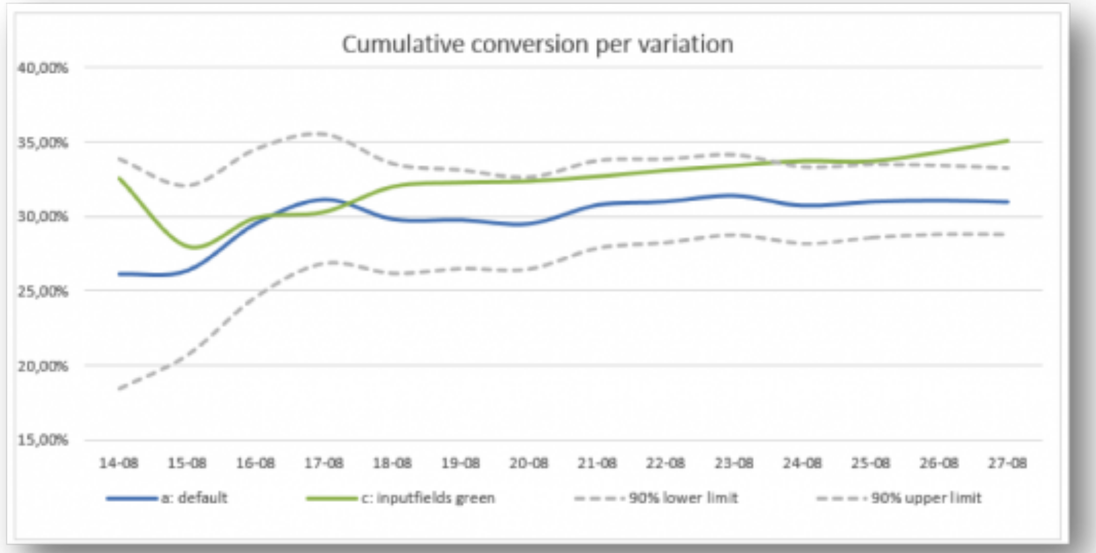
- 통합검색 내 특정 영역의 일별 CTR 트렌드 (좌: CTR절대값 / 우: CTR차이)
- 일별 트렌드에서도 A/A'의 차이가 거의 없음
- A/B군의 차이 역시 비교적 안정적으로 보임





[DeepDive] 일별 지표 트렌드의 원인

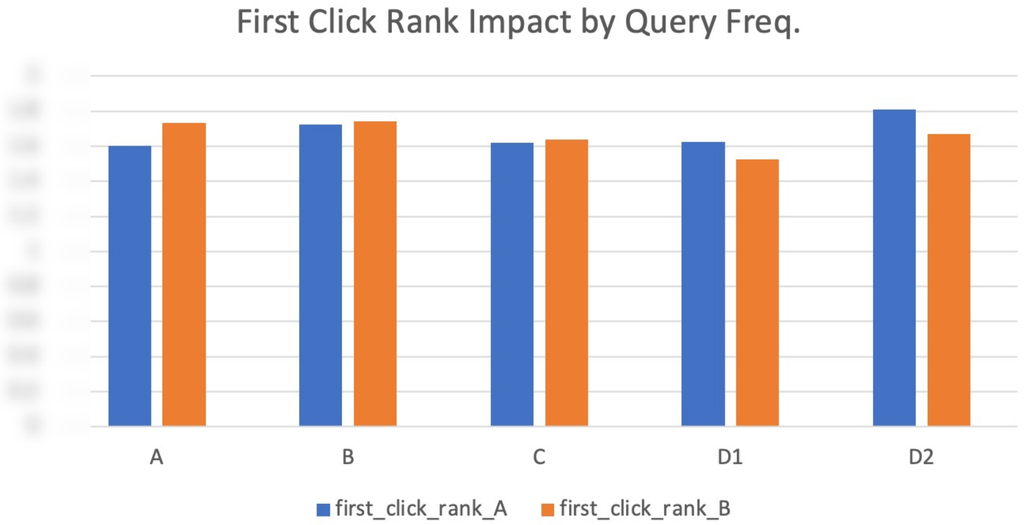
- Novelty / Learning Effect:
  새로운 UX 피처에 대한 새로움/학습 효과
- User Mix Shift:
  날짜별로 서비스 방문 사용자 군의 차이
- Seasonal / DoW Effect:
  계절/요일별 사용패턴 차이
- Carryover Effect:
  예전 실험의 효과가 신규 실험에 전이

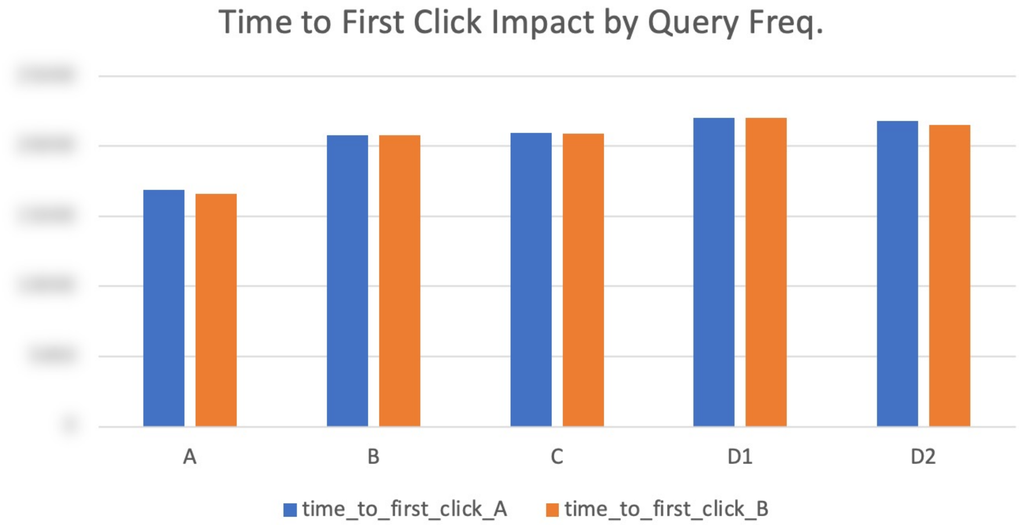


출처: https://cxl.com/blog/visualize-ab-test-results/

부문별 세부 지표 해석하기

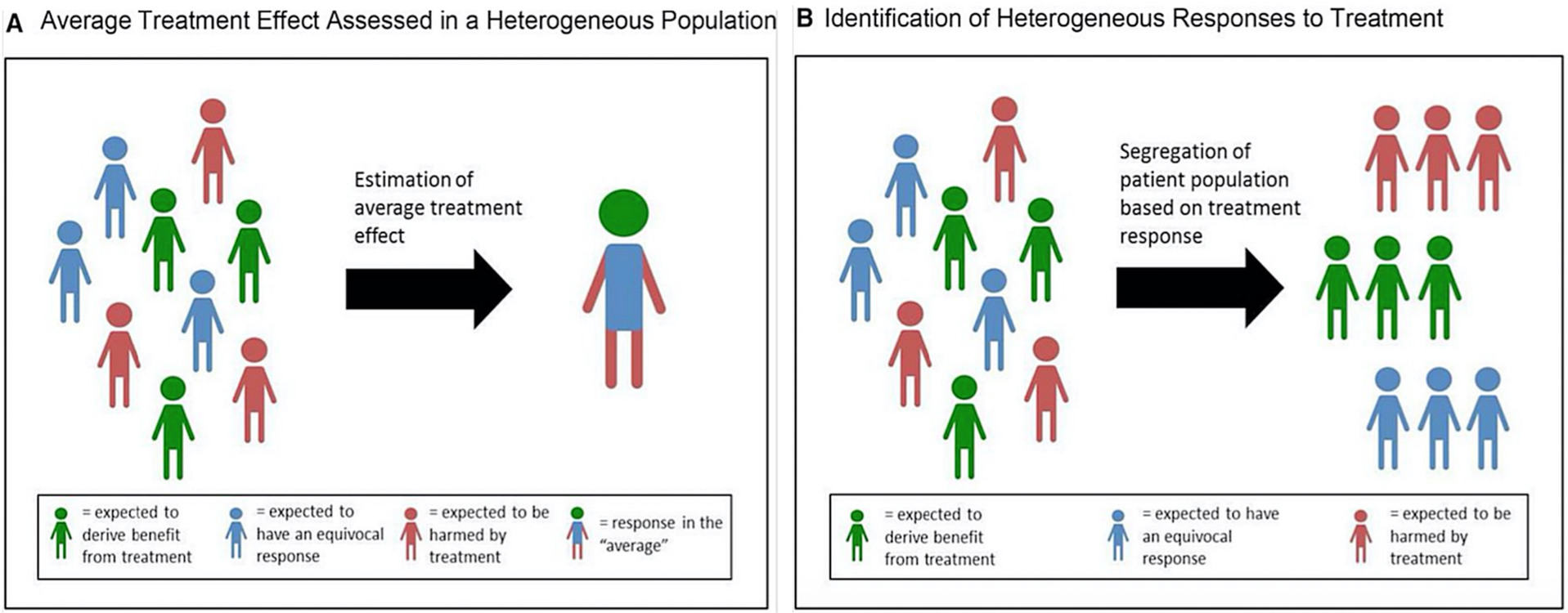
- 개별 실험의 결과를 이해하기 위해 다양한 기준으로 결과를 나눠볼 수 있어야
- 오른쪽 차트의 결과는 위 지표에 대해서 (First Click Rank) 질의군별로 다른 영향
- 아래 지표에 대해서는 (Time to First Click) 질의군에 관계 없이 거의 같은 영향
- 비슷한 Breakdown을 사용자군 / 검색 랭킹 및 검색 결과 유형별로도 제공



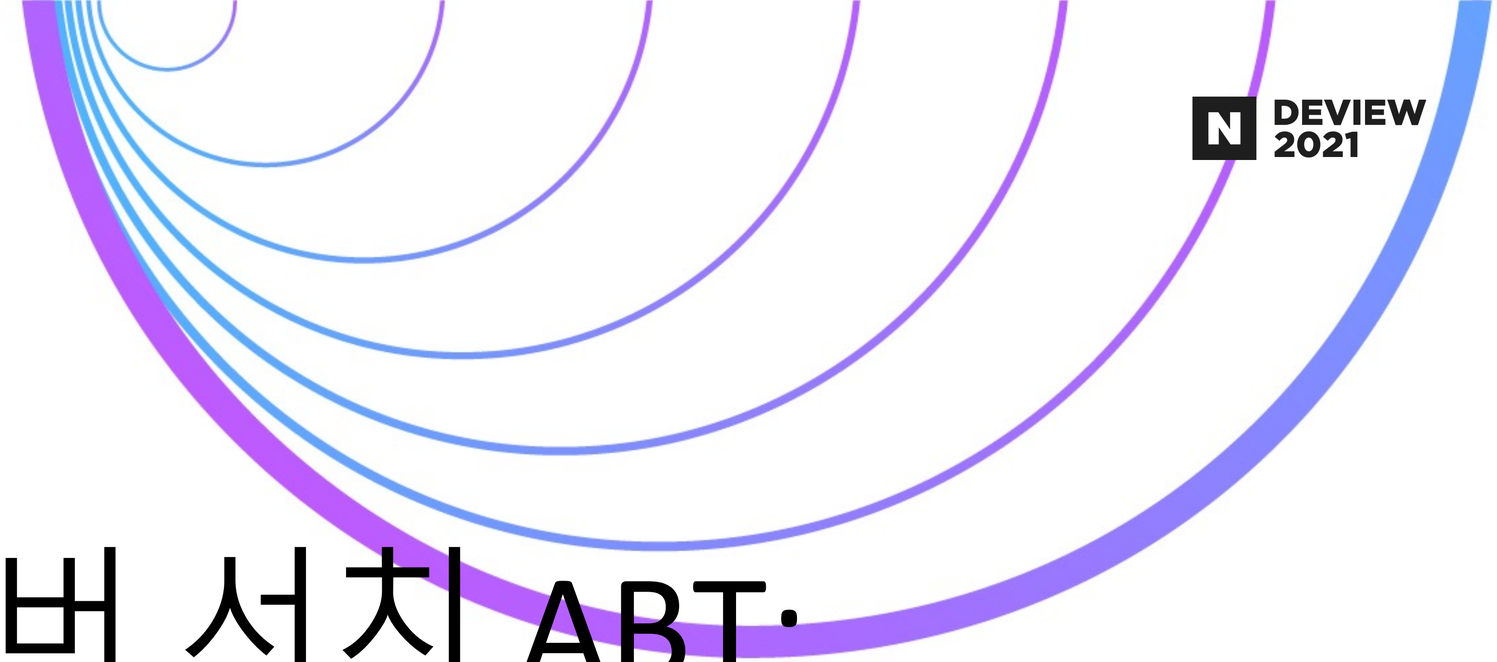


[DeepDive] 부문별 세부 지표 해석의 유의사항

실험에 따라 부문별로 다른 영향을 가져올 수 있으며, Breakdown을 통해 이를 발견할 수 있음. 단 여기서도 False Discovery 이슈를 고려해야



출처: https://arxiv.org/pdf/1808.04904.pdf



DEVIEW 2021

네이버 서치 ABT:
Summary & Next Steps

네이버 서치 ABT: Summary

네이버 검색 전체를 위한 현대적인 ABT 개발

개별 팀의 독립적인 실험 수행 및 iteration을 지원

모든 팀이 같은 지표/프로세스를 통해 의사결정

런치 두달만에 네이버 서치 대부분의 팀과 실험 수행

네이버 서치 ABT: Next Steps

고급 지표 개발 (예: 사용자의 종합적인 만족/불만족을 지표화)

고급 질의군 및 사용자군 Cohort 개발 (대용량 클러스터링 기반)

더 많은 지표에 대해 실시간 이상 탐지 및 Shutdown 기능 제공

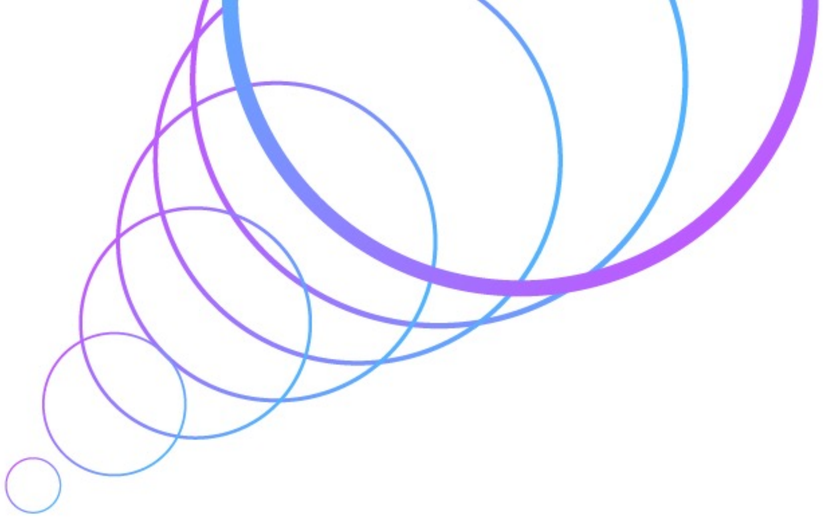
더 자동화된 실험 결과 분석 및 편리한 Custom Analysis 지원

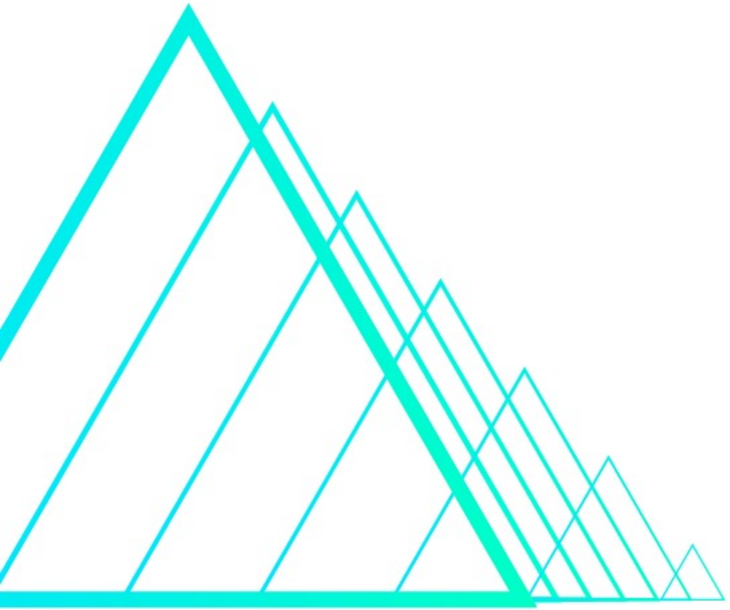
전통적인 실험이 어려운 경우에 대한 기법 연구 (Causal Inference)

평가를 넘어 온라인 환경에서의 파라메터 최적화를 지원 (Contextual Bandit)

이 여정에 함께하실 Data Scientist & Engineer 분들을 채용하고 있습니다!

(Email: jin.y.kim@navercorp.com)





Q & A

현재 ABT의 개선

고급 지표 및 Cohort 개발

모니터링/분석 자동화

고급 분석 기법 도입

Thank You