테스트 도구

🥑 데이터 드리븐이란

데이터 중심의 데이터 기반의 의사결정 행위를 의미해요.

즉 사람의 직관이 아닌 객관적인 지표 설계와 테스트로 이를 증명하는 방법이고

결과적으로 리스크를 줄이는 효율적인 근거있는 일처리가 될겁니다.

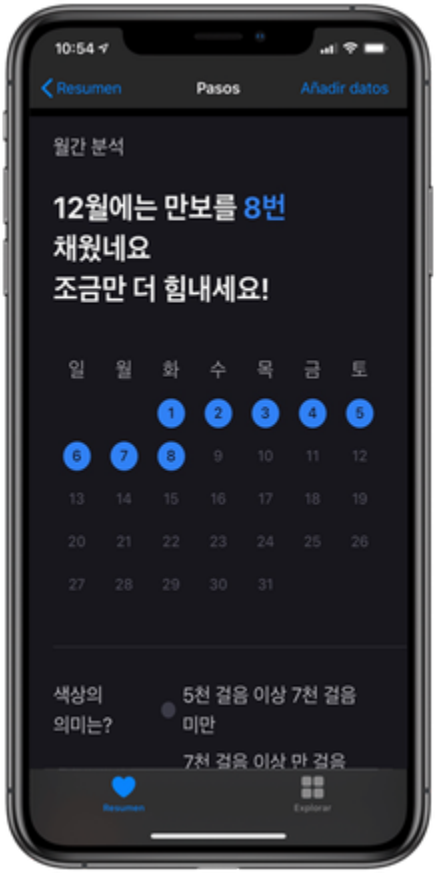
그럼 데이터 드리븐 ( ) 에는 어떤 것들이 있을까요?

- 피트니스 앱 리포트에서
  - 일간 세부 리포트 주간 집계 리포트 월간 리포트 순서 정하기
- 좋아할 만한 콘텐츠 추천
  - 앱 사용자 사용량을 늘려 광고 매출 증대 시키기

> *아래 두가지 리포트 중에서 어떤 그래프가 의미 있을까요?*

두 그래프 모두 일단위 걸음수를 표시한 내용이에요.

다만 정보의 함축 정도에 따라 두가지로 표현할 수 있을 것 같아요.



데이터 드리븐의 경우 지표를 설계하고 유저 테스트로 이를 측정해요.

즉 페이지에 얼마나 자주 방문했는지?

또는 활용률 수치를 바탕으로 의사결정을 할 수 있을거에요.

> *어떤 것들인지 이해가 가시나요?*

그런데 곰곰히 생각해보면

이렇게 객관적인 방법을 왜 우리는 사용하지 않을까요?

사실 데이터 드리븐은 개발 방법론과 관계가 있어요.

즉 서비스 모델과 관련이 있다는거죠.

> *갑자기요?*

그럼 간단히 소프트웨어 생명주기를 알아보고

데이터 드리븐에 대해 자세히 알아볼게요.

🏈 소프트웨어 생명주기란

시스템 구축 운영을 체계적으로 나타낸 절차라고해요.

- 시스템 요구분석 → 설계 → 구현 → 테스트 → 유지보수

이러한 생명주기를 어떤식으로 개발할지

고민한 것이 개발 방법론이구요.

개발론은 소프트웨어를 어떻게 만들지에 대한 관심으로써

단계별 산출물뿐만 아니라

산출물은 누가 어떤 순서로 어떻게 만들어야 하는지 구체적으로 정의해요.

보다 깊은 내용을 원한다면 국제 표준을 확인해 볼 수 있어요.

하기 문서는 번역이 되어 있답니다.

- 시스템 그리고 소프트웨어 엔지니어링 소프트웨어 생명 주기

그럼 잘 알려진 방법론에는 어떤 것들이 있을까요?

- 폭포수 모델
- 애자일 스크럼 칸반
- 린스타트업

우리는 개발할 때 상황에 맞는 방법론 선택이 필요할거에요.

만약 대규모 인원이 개발을 진행한다면

애자일이 아닌 폭포수 모델이 맞을 거에요.

기획부터 개발까지 하나의 프로세스로 반복하기엔 무리가 있다는거죠.

> *애자일 많이 들어보셨죠?*

애자일은 기민한 · 날렵함을 뜻하고 낭비없게 만드는 방법을 말해요.

> *서비스 설계에서 완벽한 설계를 해보신적 있어요?*

애자일은 확실한 것들에 대해서는 계획을 세우고요.

불확실성에 대해서는 변화에 잘 대응하도록 개발하자는 방법론이에요.

> *그럼 설계없이 프로젝트를 시작하는 것도 애자일일까요?*

無 계획 無 설계가 애자일이라고 잘못 생각하는 경우도 많아요.

애자일은 크고 간헐적인 변화를

보다 작고 빈번한 변화로 전환시키는데 목적이 있음을 잊어서는 안되요.

애자일은 일정한 주기로 요구사항을 추가하거나 수정해나아가요.

여기서 일정한 주기는 스프린트이고

주기를 반복하는게 큰 사이클은 스크럼이라 불러요.

스크럼은 시간을 통제함으로써 개발 효율을 높이자는 거에요.

> *우리는 아침마다 데일리 미팅을 해요. 이것을 왜하는 걸까요?*

저는 한참 동안 이 사실을 몰랐어요.

반면 칸반은 작업을 통제해서 효율을 높여요.

한번에 한가지만 작업만을 허용 합니다. 즉 선택과 집중을 하겠다는 거에요.

제가 연구실에서 종종 하던 말이 있어요.

> *하나라도 잘하자.*

잘 생각해보면 두가지 개발론의 목적은 같아요.

개발 효율을 높이자는 것이고

이는 프로젝트에서 무엇을 통제하느냐에 따라 나뉘게 되는 거죠.

근데 비슷한게 개념으로 린스타트업이 있습니다.

여기부터 좀 혼란스러울 수도 있어요.

> *스크럼과 린스타트업은 어떠한 차이가 있을까요?*

애자일은 비즈니스 전문가를 개발 프로세스에 참여시켜요.

즉 고객이 본인이 어떤 문제를 풀고 싶은지는 알고는 있지만

방법은 모르는 경우가 이에 해당해요.

그렇다면 불확실성이 너무 크거나

고객의 이해도가 낮은 경우에는 어떻게 해야할까요?

이때 린스타트업을 사용하할 수 있어요.

기획해서 만들고 평가받고 이를 반복하면서

고객은 정말 원하는 것이 무엇인지 찾을 수 있을거에요.

자 이제 데이터 드리븐과 연계해서 살펴볼게요.

> *고객이 정말 원하는 것은 어떤 것일까요?*

우리는 그것을 찾기 위해 우리는 데이터 드리븐을 합니다.

이에 대한 행위는 테스트라 불리고요.

> *와 두가지 그래프 중에서 사용자는 어떤 것을 원할까요?*

사용자에게 피드백 받는 방법은 다양한 편이에요.

두 그래프를 동시에 보여주며 설문을 할 수도 있고

이번주에는 를 노출시키고 다음주에는 를 보여주며

좋아요 수를 확인할 수도 있을거에요.

다만 모든 사용자들이 피드백을 주진 않을거에요.

이러한 명확한 피드백을 형식지라 부르고

이는 극히 드물게 반응하곤해요.

그래서 우리는 간접적으로 이를 확인할 수밖에 없고

이런 수치를 암묵지라 불러요.

> *영상 추천이 얼마나 잘 이루어지고 있을까요?*

영상에 머물러 있는 시간을 체크해요.

즉 우리의 서비스를 개선하고자 한다면 테스트 시스템과 지표가 필요할거에요.

하기는 테스트 단계별 프로세스인데요.

- 목표를 구체화한다.
- 지표를 선정한다.
- 가설을 세운다.
- 실험 설계와 실행을 한다.
- 결과를 분석한다.

> *마치 냉장고에 코끼리 넣는 방법 같아보여요.*

먼저 기획진 분석진은 목표와 가설을 세워요.

- 현재 운영 화면을 신규 기획 화면으로 변경한다면

- 현재 대비 이용률이 퍼센트 증가할 것으로 예상

샘플 개수와 테스트 기간 산정

기획자 개발진 분석가는 지표를 선정합니다.

- 이용률에 대한 정의 : 머물러 있는 시간 페이지 방문 횟수

수집 가능 여부에 따른 지표 수립

그리고 개발진은 실험 설계와 실행을 수행하구요.

프론트엔드 : 지표 수집이 가능하도록 화면 구현

백엔드 : 기간 동안 테스트 그룹 분할과 유지 기능 구현

모니터링 : 수집된 통계 데이터의 시각화 구현

마지막으로 분석가는 결과를 도출하는데요.

가설 검정에 따른 리포트

여기서 결과가 좋다면 반영할 것이고 미미하다면

다시 기획 테스트 분석 과정을 반복합니다.

> *이러한 프로세스가 뭐라고 했죠?*

🎨 테스트란

두가지 방법을 비교해서 어느 방법이 좋은지 결과를 도출하는 과정이에요.

예를 들어 스포츠 뉴스 기사가 있습니다.

이때 어떤 상황에서 더 많이 클릭하는지 테스트하고자 해요.

- 가설 : 이미지 썸네일 링크는 단순 텍스트보다 상대적으로 클릭이 많다.
- 실험 설계와 실행 : 는 단순 텍스트 링크이고 는 이미지 썸네일 링크 구현

실제로 이미지 썸네일이 더 효율적인 결과를 얻을 수 있지만

해당 결과만으론 무조건 좋다고 단정 지을 순 없어요.

왜냐하면 화면에 보여줄 수 있는 컨텐츠 개수는 줄어들었기 때문이에요.

즉 테스트 결과를 쉽게 단정 지어서는 안될거에요.

다만 오늘은 테스트에 대해 쉽고 얕게 알아볼 것이기 때문에

테스트 주의사항과 단점은 나중에 이야기하도록 하죠.

📌 개발진 역할로

위와 같은 상황에서 이러한 고민을 할 수 있어요.

> *어떻게 하면 효율적으로 와 군을 나눌 수 있을까?*

현재 우리 서비스는 화면을 사용중이라면

기획진는 썸네일 이미지가 클릭율이 더 높은 것이란 가정으로 테스트를 진행해요.

운영중인 서비스이기 때문에 너무 갑장스런 변화는

사용자에 혼란을 줄 수 있으므로

전체 사용자 중 5 % 비율만 썸네일 화면을 제공해 주기로 설계해요.

- 그룹 : 그룹 = 95 % : 5 %

> *여기서 의문이 생기죠?*

절반으로 나눠야 정확한 측정이 가능할 텐데요.

위 예시는 일반적인 방법으로 특별한 이유가 있어보입니다.

> *생각을 해보자.*

이렇게 나뉘어진 그룹에서 반드시 지켜야될 부분이 있습니다.

 그룹 유저의 경우 테스트 기간 동안에는 화면 만이 노출되어야 해요.

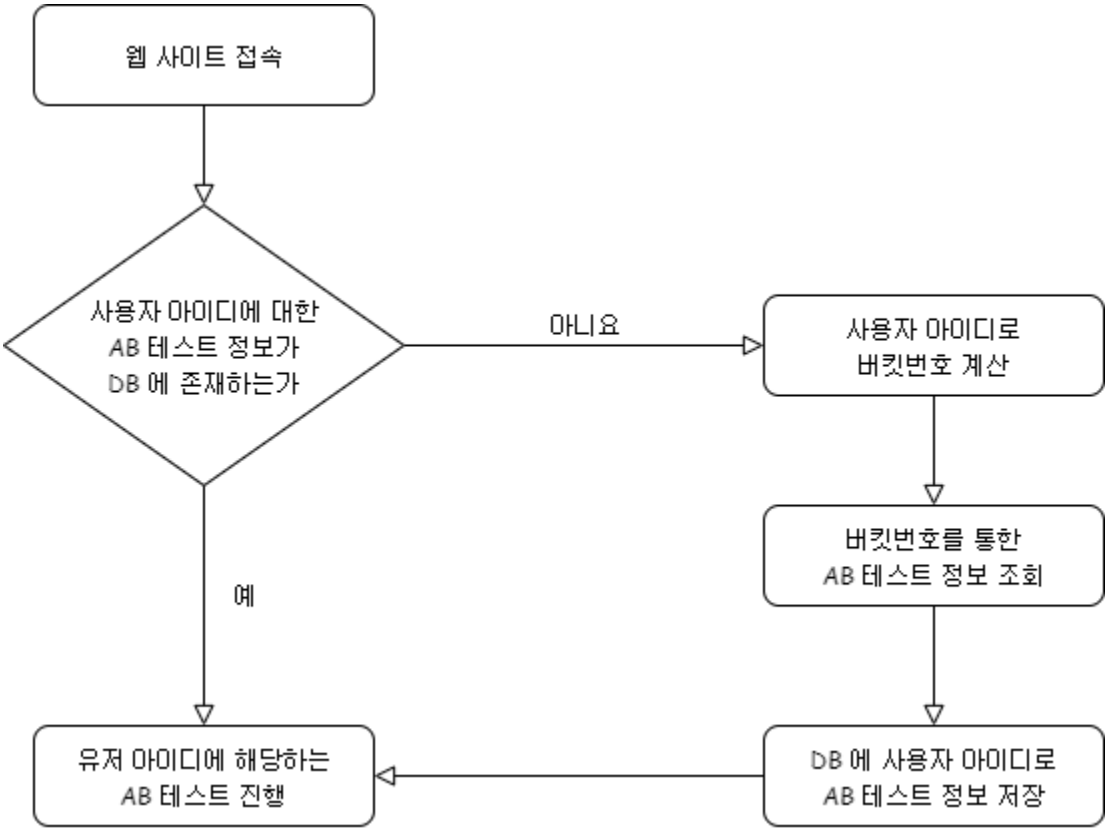
재접속 했을 경우에서 마찬가지구요.

먼저 단순한 아이디어로 그룹을 나눠볼까 합니다.

특정 사용자가 테스트 기간중에 웹사이트를 처음 방문했어요.

그리고 그룹으로 설정했다면

다음 접속에도 화면이 노출되도록 사용자 아이디로 구현을 할 수 있어요.



해당 방법은 매우 심플하지만

데이터베이스에 사용자 테스트 정보를 모두 저장해야 한다는 단점이 있죠.

만약 사용자 아이디에 대한 테스트 정보가 항상 동일하다면

굳이 데이터베이스에 저장할 필요는 없어보여요.

예를 들어 사용자 아이디의 인덱스가 홀수라면 그룹으로

짝수라면 그룹으로 구분할 수 있을거에요.

> *좀 더 나이스한 방법은 뭐가 있을까요?*

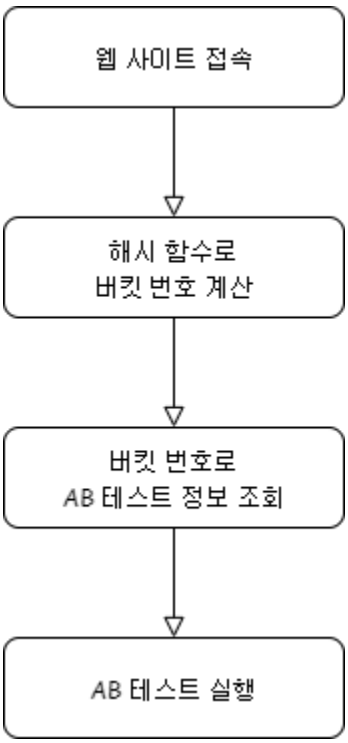
바로 해시 함수를 사용하여 샘플링하는 방법이 있다고 해요.

해시 함수는 입력 데이터를 고정 길이값으로 생성하는

일방향 알고리즘인데요.

해시 함수를 사용해서 일관된 버킷 번호를 구하고

이에 해당하는 테스트 그룹을 선택하는 방법이에요.



만약 실험마다 테스트 집단이 동일하지 않다는 조건이라면

버킷 번호를 계산하는 해시 함수에 변화가 필요합니다.

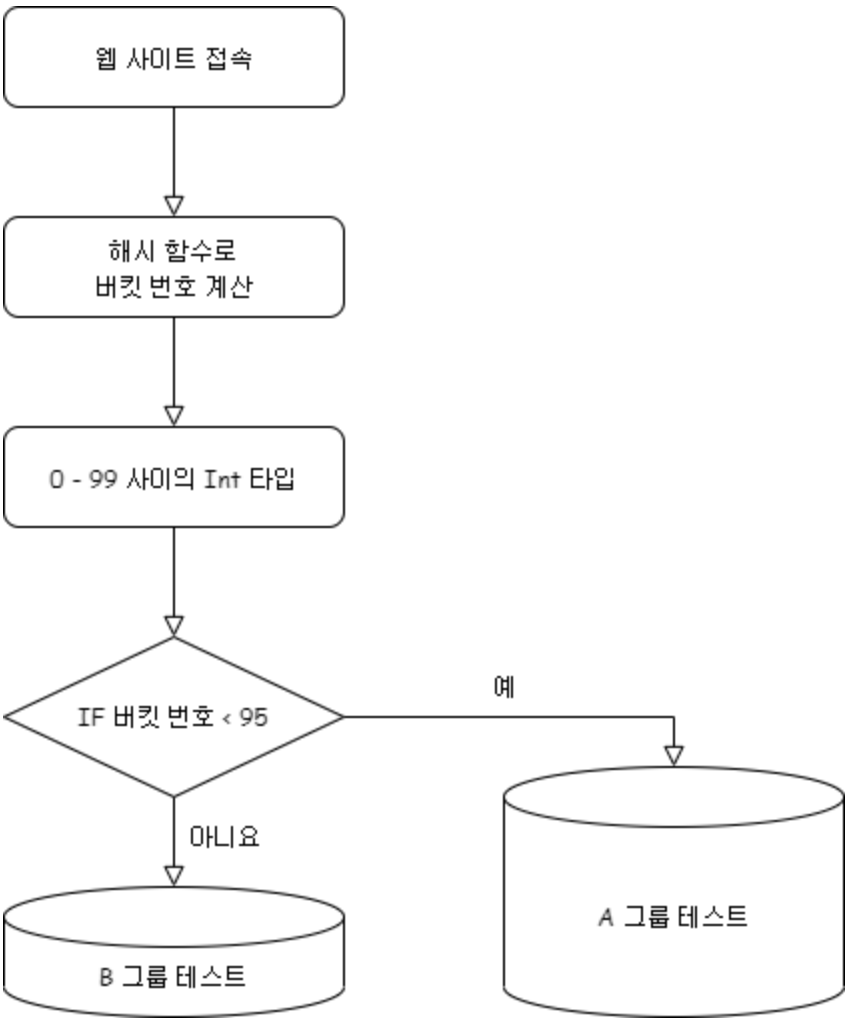
왜냐하면 사용자 아이디를 기준으로

해시 함수를 계속 수행하면

항상 동일한 버킷 번호가 조회되기 때문이에요.

테스트를 95 대 5 비율로 진행 한다고 가정하면

가 속하는 범위는 0 부터 94 까지이고 는 95 에서 99 가 될 것이다.



이 경우 해시 함수에 입력되는 데이터를

사용자 아이디와 함께 여러 조건을 조합해서 문자열을 만들 수 있어요.

```
class HashFunction:
    ...

bucket_number = HashFuntion(user_id) = 30

test_number = str(1)
bucket_number = HashFuntion(user_id + test_number) = 5

test_number = str(2)
bucket_number = HashFuntion(user_id + test_number) = 95
```

위와 같이 실험 번호를 붙인다면

해시 함수에 입력되는 문자열이 계속 변경이 되고

계산된 버킷 번호 역시 변경될거에요.

🥝 분석진 역할은

테스트를 위한 목표와 가설을 세우고 결과를 분석을 진행합니다.

> *사실 테스트 설계는 그리 간단하지 않다.*

표면상으로 방안과 방안 중에서 어느 것이 효과적인지 찾는 것이지만

표면상 질문 뒤에 숨은 진짜 질문은 이것보다 더욱 복잡하기 때문이죠.

> *몇가지 예시를 살펴보면 바로 느낌이 올거다.*

첫째로 우리는 암묵적으로 테스트 결과가 상당히 크길 바라고 있어요.

다음은 학습 리포트 사용률에 대한 테스트 결과인데요.

- 는 그룹 보다 1 퍼센트로 승리

> *1 퍼센트는 무언가 찜찜하다.*

다시 말해 테스트 결과 차이가 상당히 크거나 확실하길 바란다는거죠.

둘째로 우리는 실험 진행이 공정하길 바란다는 거에요.

> *공정함의 실체가 무엇인지 아무도 모를거다.*

만약 최대한 트래픽을 균등하게 나누려 했지만

어떤 이유로 그룹의 사용자 비율에 차이가 생겼다고 가정하자.

그룹은 리포트 사용률이 55 퍼센트이지만 300 명의 사용자이고

반면 그룹은 54 퍼센트이며 1,300 명의 사용자가 있다.

> *비슷하다면 답변이 많은 그룹이 낫지 않을까?*

> *애시당초 비교 대상이 맞는건가?*

이처럼 우리는 테스트를 진행할 때

공정함의 실체가 무엇인지 모르면서 진행이 공정하길 바란다는 거죠.

그래서 자신도 모르게 실험 배경이 5 대 5 로 동일했다고 생각을 하는 경우가 많아요.

셋째로 실험 결과가 이례적이거나 우연이 아니길 바란다는 거에요.

몇일간 모아서 분석한 결과가

앞으로도 확실히 효과가 있었으면 하는거에요.

이렇게 숨은 질문은 생각하면

트레픽을 절반으로 나누거나 지표를 하나만 사용하는 식으로

단순히 정의할 수 없다는거에요.

그래서 테스트 고민을 조금 정리했는데요.

- 샘플 사이즈는 얼만큼 필요한가?

- [ ] 결과 기대값 설정하기
- [ ] 신뢰구간 설정하기 : 우연이 아닐 확률

- 샘플 수집 기간은 얼마로 할까?

- [ ] 비즈니스 주기 또는 통계치 바탕 : 장기간 진행하면 릭 발생

- 어떤 분포를 가정하고 가설 검정을 할 것인가?

- [ ] 가정 분포 설정하기 : 정규 분포 파레토 분포 등에 따라 결과가 상이

> *우선 테스트 예시를 먼저 살펴볼까요?*

우리는 서로 다른 이미지인 안과 안을 화면에 노출시켜 구매 여부를 체크했어요.

와 그룹 사용자는 8 만명으로 동일해요.

그리고 그룹은 1,600 의 구매 그룹은 1,690 의 구매가 이루어졌다네요.

- 테스트 결과 분석 도구
  - 데이터 샘플수 구매 여부 합계
  - 가설 : 단측 검정 기각한다면 ?? 안이 안보다 구매율이 높다.
  - 신뢰 수준 : 95 퍼센트 우연일 확률은 5 퍼센트를 의미

이제 분석 도구를 사용할 준비가 되었어요.

다만 여전히 샘플 사이즈는 얼마로 해야할지 감이오질 않는데요.

정해진 최소 샘플 사이즈가 있을까요?

> *뭐 기본값 같은건 없나.*

결론부터 말하자면 샘플 사이즈는 매번 다를거에요.

안과 안 사이의 차이가 어느 정도 되기를 기대하는지에 따라

동일한 신뢰 수준이라도 가설 검정 결과가 달라질 가능성이 있어요.

가령 안과 의 차이가 작다면 샘플 사이즈는 많이 필요하고

반대로 안과 의 차이가 크다면 샘플은 작아도 된다는 거에요.

- 테스트 샘플 분석 도구
  - 안의 기본 구매율이 20 퍼센트일 때
  - 안과 1 퍼센트 차이가 의미있으려면 : 그룹당 샘플이 최소 25,555 개가 필요
  - 안이 5 퍼센트 차이가 난다면 : 그룹당 샘플이 최소 1,030 개 필요

결국 샘플 사이즈는 얼마나 필요한가에 대한 질문은

테스트에서 두 방안이 몇 퍼센트가 차이날 것을 기대하는가로

바꾸어 말할 수 있는거죠.

이는 경험을 바탕으로 할 수 밖에 없어요.

다만 샘플 사이즈가 어느정도 되어야 충분한 차이인가라는 질문으로 바꾼다면

아래와 같은 상상을 해볼 수 있겠죠.

고객이 100,000 명이라면 구매율 차이가 1 퍼센트만 되어도

그 차이는 1,000 명이 될 것이고

객 단가가 100 만원이라면 1 퍼센트라는 작은 차이에도

매출액 차이는 1 억원이 될거에요.

반대로 1,000 명이라면 5 퍼센트가 되더라도

50 명 밖에 안되고

객 단가가 1 만원이라면 50 만원 뿐이란 거죠.

> *아마 인건비가 더 들겁니다.*

이외도 여러가지 고려 사항들이 있을겁니다.

- 테스트는 같은 날짜에 시작하기
- 안이 이기고 있더라도 조기 종료하지 않기 의도적 오류

다음으로 분포에 대해 이야기하고 싶지만

내용이 너무 길어지는 것 같아서 여기서 끝내도록 할게요.

> *가장 중요한 부분은 이야기 안하네.*

이상 테스트 도구에 대한 설명이었구요.

아마 클라이언트 서비스를 시작했거나 계획하고 있다면

언젠가는 만나게 될 녀석인 것 같아요.

레퍼런스는 하기에 나열되 있는데요.

좋은 자료이니 읽어보고 찾아보고 한발 자국 나아가시길 바랄게요.

- AB 테스트 스프링부트 구현 바로가기
- AB 테스트를 보완하는 알고리즘 바로가기
- AB 테스트 자동 분석툴 개발 바로가기
- 개발자를 위한 AB 테스트 해시 샘플링 바로가기
- AB 테스트 설계시 우리의 진짜 질문 바로가기
- AB 테스트 계산기의 세팅과 해석 바로가기