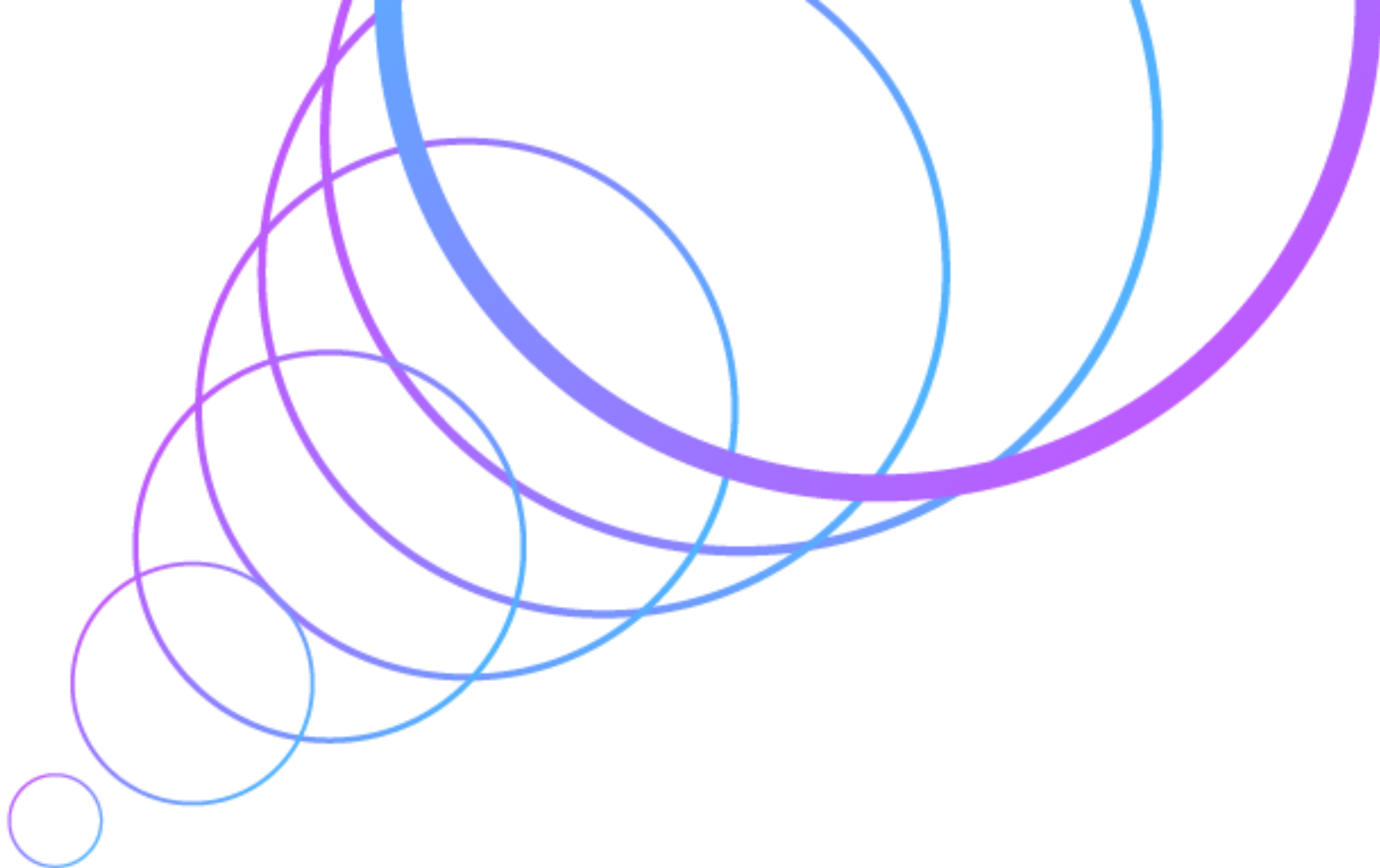




No measure, No SRE

장애율 0%대 달성 후 새로운 SRE KPI 개발기



박정률 NAVER
이권재 NAVER
이미현 NAVER

CONTENTS

1. 네이버 검색과 SRE
2. SRE KPI
3. 새로운 KPI 개발 과정
4. 결론

1.네이버 검색과 SRE

1.1 네이버 검색 SRE

SRE

- Site Reliability Engineering

1.1 네이버 검색 SRE

SRE

- Site Reliability Engineering
- 서비스를 **안정적**이고, **효율적**으로 운영할 수 있는 엔지니어링 방식

1.1 네이버 검색 SRE

SRE

- Site Reliability Engineering
- 서비스를 안정적이고, 효율적으로 운영할 수 있는 엔지니어링 방식
- Search Reliability Engineering

1.2 왜 SRE를 도입해야 하는가

네이버 검색은 365일/24시간 무중단 제공되어야 한다

- 온 국민이 사용하는 공공재와 같아서 절대 서비스가 중단되면 안된다

1.2 왜 SRE를 도입해야 하는가

네이버 검색은 365일/24시간 무중단 제공되어야 한다

- 온 국민이 사용하는 공공재와 같아서 절대 서비스가 중단되면 안된다

서비스 규모가 커지는 만큼 엔지니어는 늘어나지 않는다

- 서버는 무한히 늘어나도, 사람은 무한히 늘어날 수 없다

1.2 왜 SRE를 도입해야 하는가

네이버 검색은 365일/24시간 무중단 제공되어야 한다

- 온 국민이 사용하는 공공재와 같아서 절대 서비스가 중단되면 안된다

서비스 규모가 커지는 만큼 엔지니어는 늘어나지 않는다

- 서버는 무한히 늘어나도, 사람은 무한히 늘어날 수 없다

지진과 같은 재난 상황에서는 기존의 대응 방안이 동작하지 않는다

- 기존의 DevOps 방법론으로는 **규모도, 재난도** 감당할 수 없다

1.3 걸어온 길 돌아보기

DEVIEW 2018

- Search Reliability Engineering (부제: 지진에도 흔들리지 않는 네이버 검색시스템)
- SRE를 도입하게 된 계기와 맨땅에 헤딩하면서 겪었던 경험 소개

1.3 걸어온 길 돌아보기

DEVIEW 2018

- Search Reliability Engineering (부제: 지진에도 흔들리지 않는 네이버 검색시스템)
- SRE를 도입하게 된 계기와 맨땅에 헤딩하면서 겪었던 경험 소개

DEVIEW 2019

- Fail Fast, Learn Faster SRE (실패에서 배워나가는 SRE)
- SRE가 정착되는 과정에서 겪었던 여러가지 시행착오와 고도화를 위한 방법론들을 통해 장애 지속 시간, 사후 처리 시간을 어떻게 단축시켰는지 소개

1.3 걸어온 길 돌아보기

DEVIEW 2018

- Search Reliability Engineering (부제: 지진에도 흔들리지 않는 네이버 검색시스템)
- SRE를 도입하게 된 계기와 맨땅에 헤딩하면서 겪었던 경험 소개

DEVIEW 2019

- Fail Fast, Learn Faster SRE (실패에서 배워나가는 SRE)
- SRE가 정착되는 과정에서 겪었던 여러가지 시행착오와 고도화를 위한 방법론들을 통해 장애 지속 시간, 사후 처리 시간을 어떻게 단축시켰는지 소개

DEVIEW 2020

- **A hands-on SRE (손에 잡히는 SRE)**
- 시스템을 운영해오면서 어떻게 SRE가 검색 서비스 운영에 녹아들 수 있도록 했는지 사례 공유

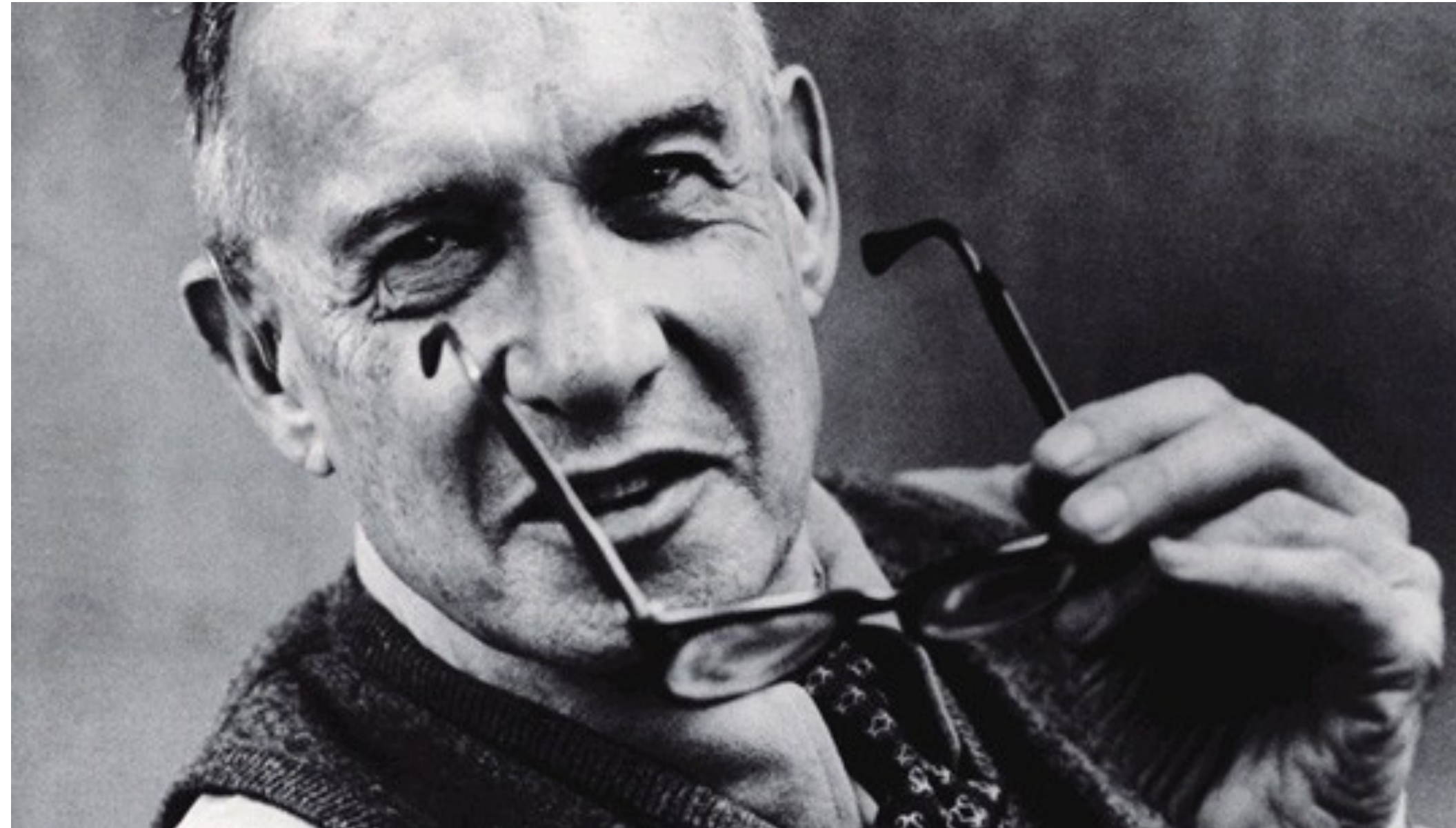
1.3 걸어온 길 돌아보기

DEVIEW 2021

- No Measure, No SRE (장애율 0%대 달성 후 새로운 SRE KPI 개발기)
- 장애율 0%대 달성 후에도 SRE를 개선하기 위해 개발한 지표 공유

2.SRE KPI

2.1 KPI의 필요성



You can't **manage**
what you don't **measure**.

Peter F. Drucker

2.1 KPI의 필요성

측정하지 않으면 개선할 수 없다

- 검색 시스템 신뢰성 개선을 위해 여러 지표를 측정한 것처럼,
- 우리의 SRE를 개선하려면 이를 측정할 수 있어야 한다

2.2 기존 KPI 소개 - 장애율

전사 변경/장애 관리 시스템

변경

장애

2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

- 간단하면서도 효과적인 지표(오컴의 면도날)

2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

- 간단하면서도 효과적인 지표(오컴의 면도날)
- 아무것도 하지 않으면 아무 일도 발생하지 않는다

2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

- 간단하면서도 효과적인 지표(오컴의 면도날)
- 아무것도 하지 않으면 아무 일도 발생하지 않는다

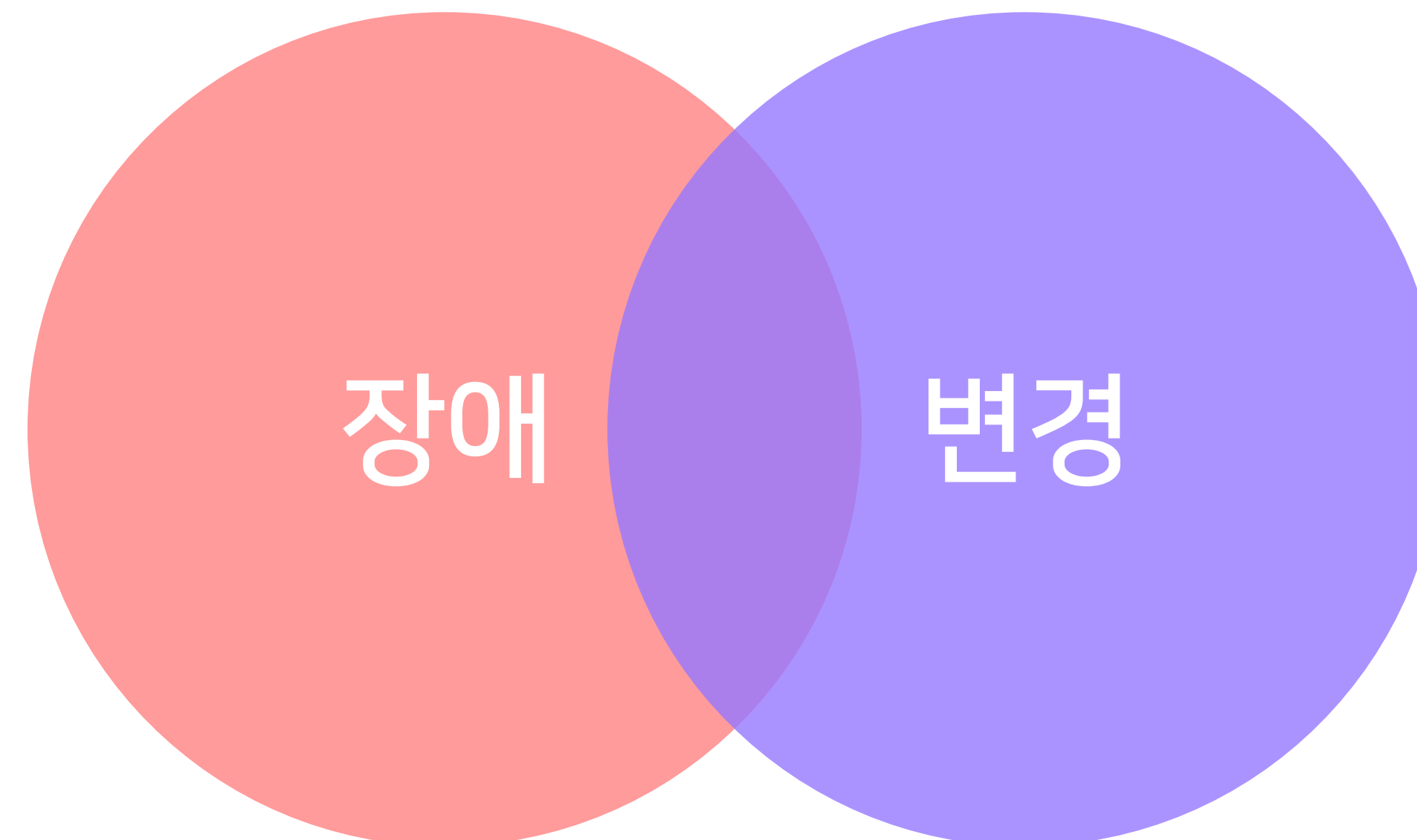


장애

2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

- 간단하면서도 효과적인 지표(오컴의 면도날)
- 아무것도 하지 않으면 아무 일도 발생하지 않는다
- 변경 건수를 통해 Feature Freshness를 함께 알 수 있음



2.2 기존 KPI 소개 - 장애율

Feature Freshness

2.2 기존 KPI 소개 - 장애율

Feature Freshness

- SREcon19 Americas: Extending the Error Budget Model to Security and Feature Freshness
- 단순히 장애가 적게 난다고 좋은 것은 아니다
- 적절한 밸런스가 중요

2.2 기존 KPI 소개 - 장애율

Feature Freshness

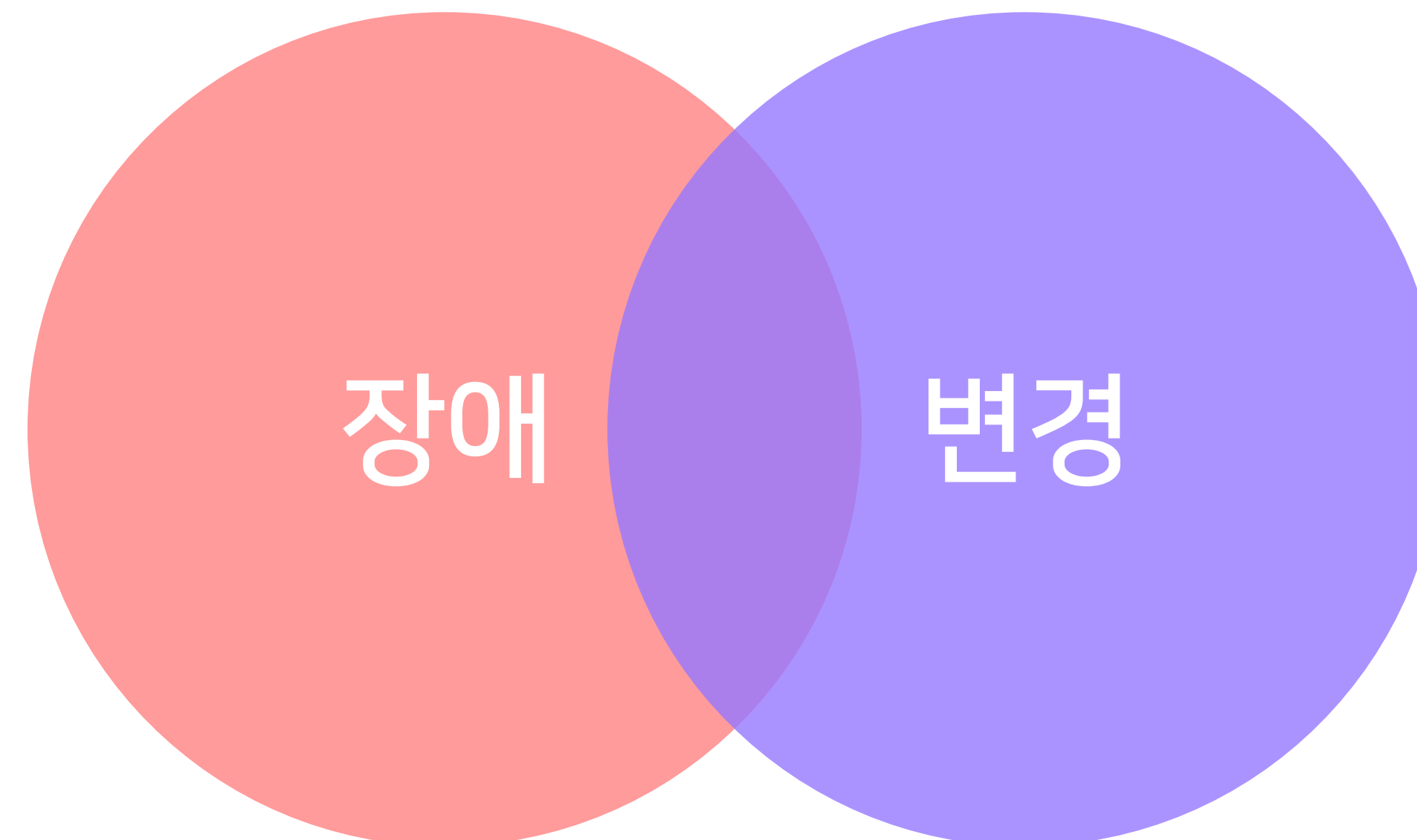
- SREcon19 Americas: Extending the Error Budget Model to Security and Feature Freshness
- 단순히 장애가 적게 난다고 좋은 것은 아니다
- 적절한 **밸런스**가 중요
- 잘 굴러가는 구식 자동차와, 첨단 기술이 들어간 신형 자동차를 비교



2.2 기존 KPI 소개 - 장애율

장애율 = 장애 건수 / 변경 건수

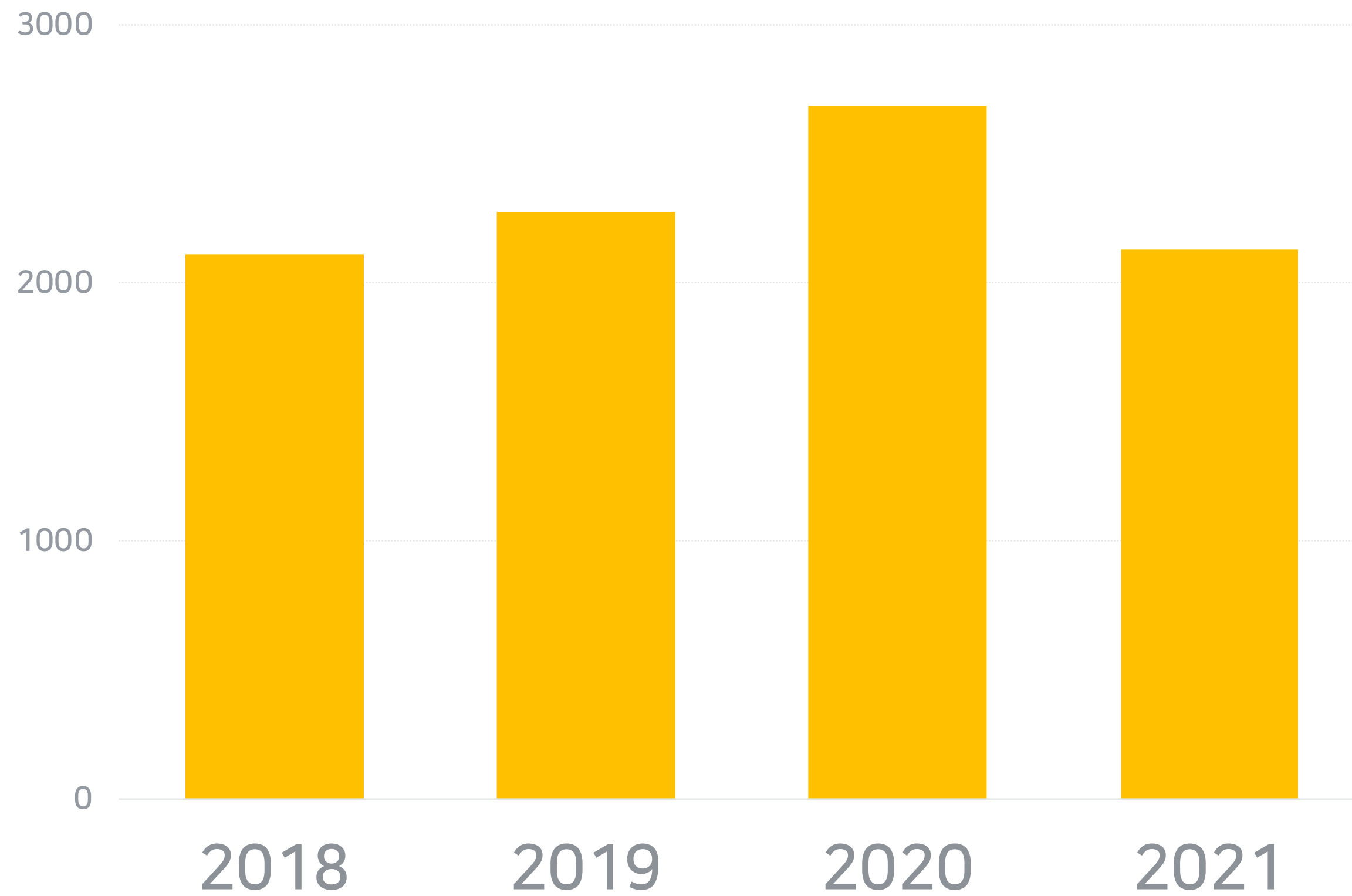
- 간단하면서도 효과적인 지표(오컴의 면도날)
- 아무것도 하지 않으면 아무 일도 발생하지 않는다
- 변경 건수를 통해 Feature Freshness를 함께 볼 수 있음



2.2 기존 KPI 소개 - 장애율

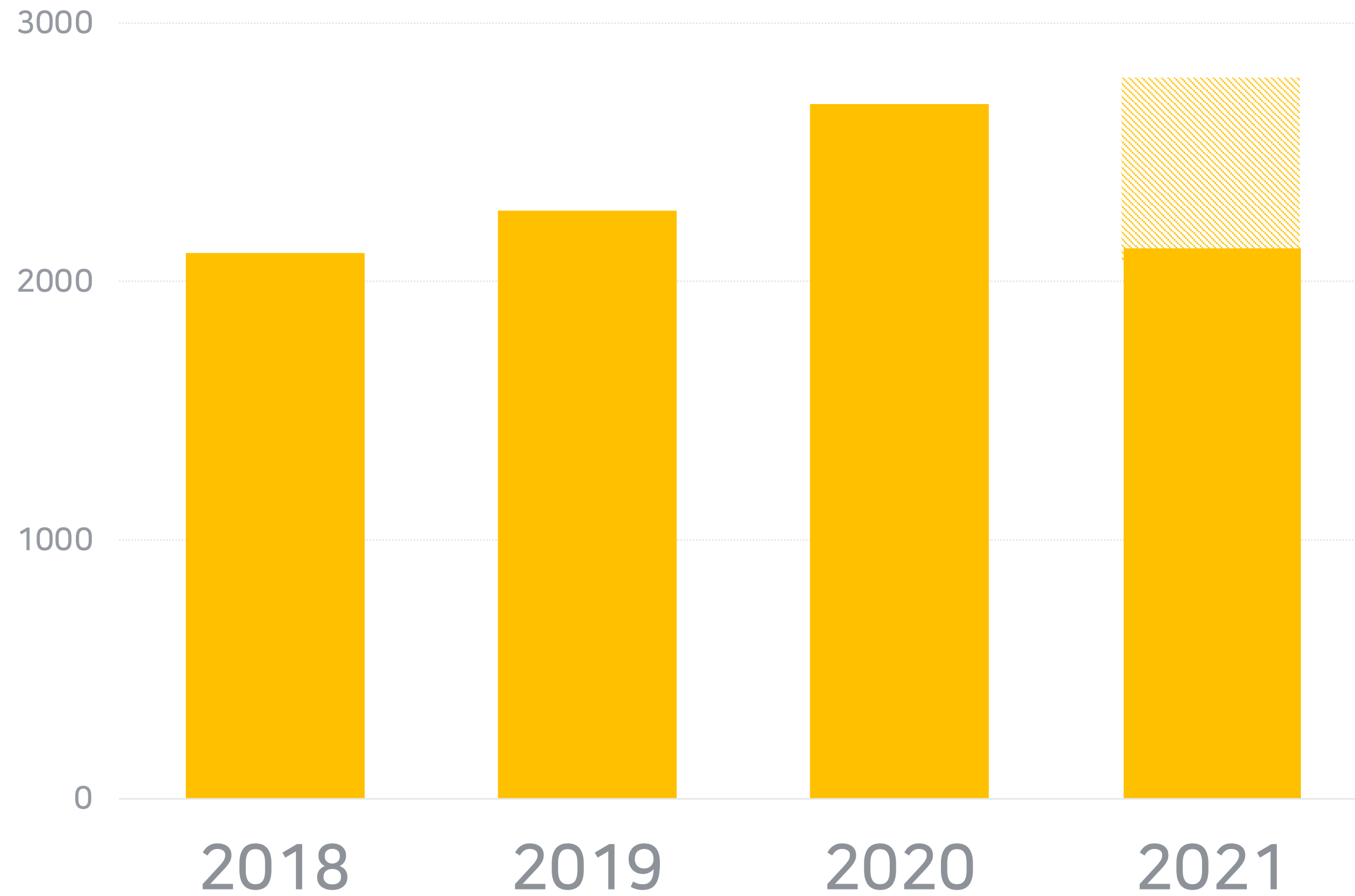
2.2 기존 KPI 소개 - 장애율

변경 건수



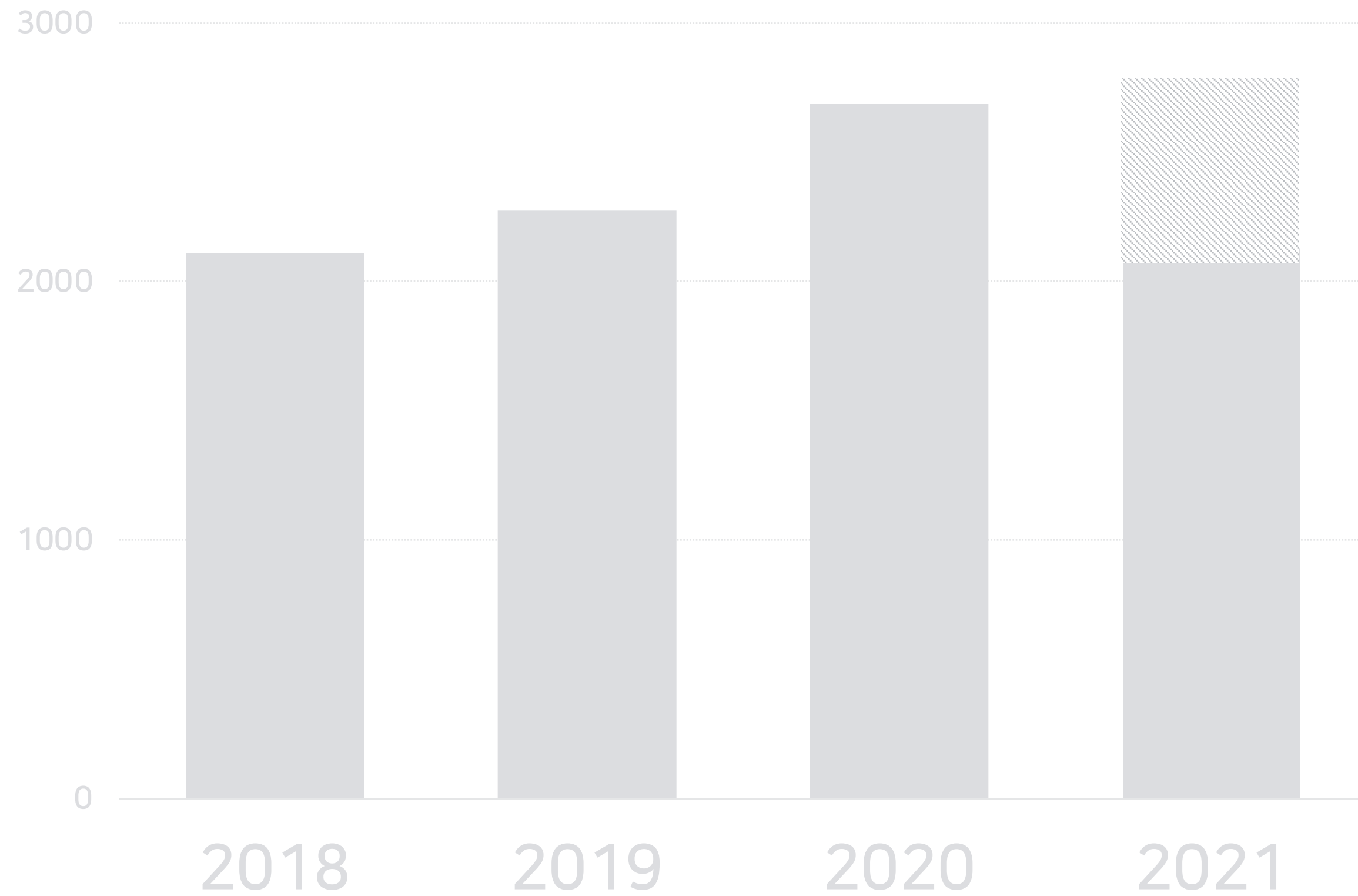
2.2 기존 KPI 소개 - 장애율

변경 건수

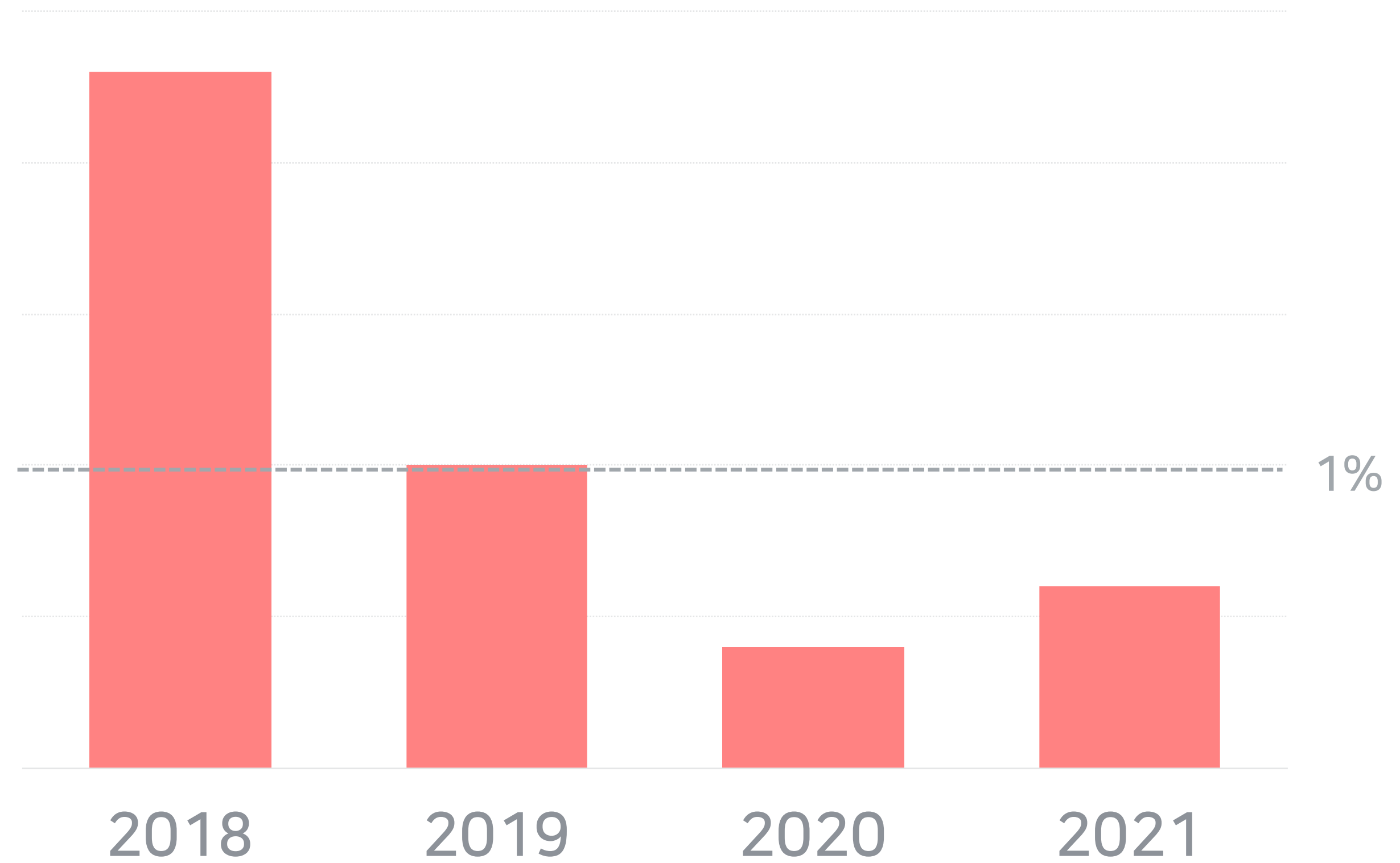


2.2 기존 KPI 소개 - 장애율

변경 건수



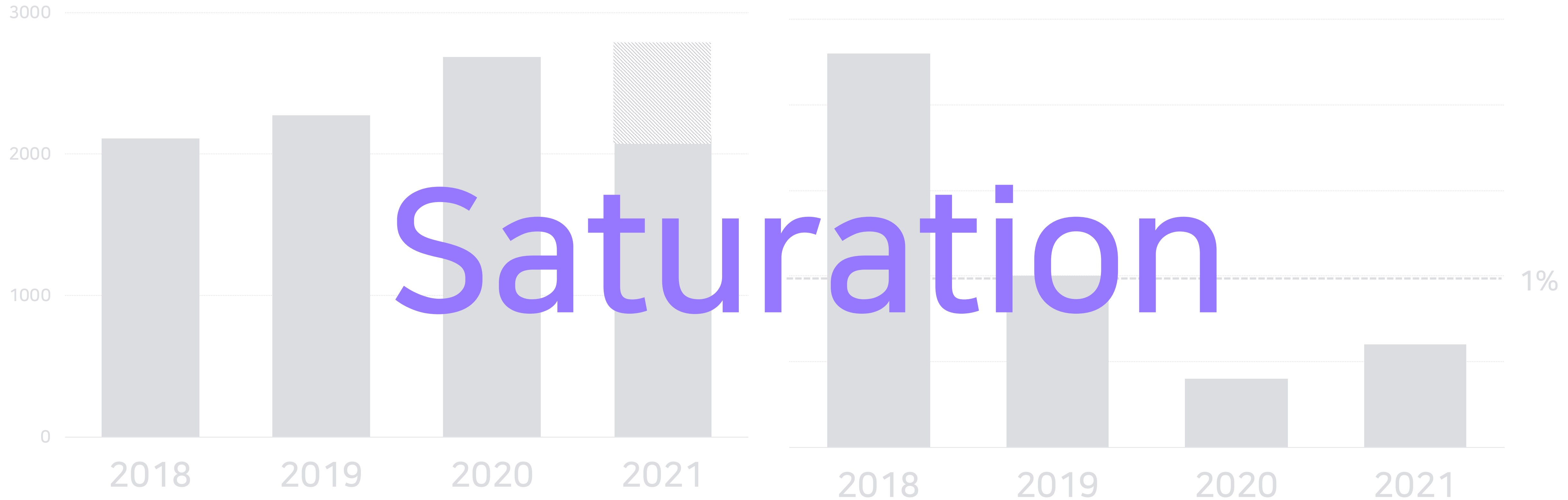
장애율



2.2 기존 KPI 소개 - 장애율

변경 건수

장애율



2.2 기존 KPI 소개 - 장애율

그럼 우리 SRE는 완벽한가…?

- No! 해상도를 높여 더 자세하게 보면 개선의 여지가 있다 → 새 KPI의 필요성

2.2 기존 KPI 소개 - 장애율

그럼 우리 SRE는 완벽한가...?

- No! 해상도를 높여 더 자세하게 보면 개선의 여지가 있다 → 새 KPI의 필요성
- Observability (관측 가능성)

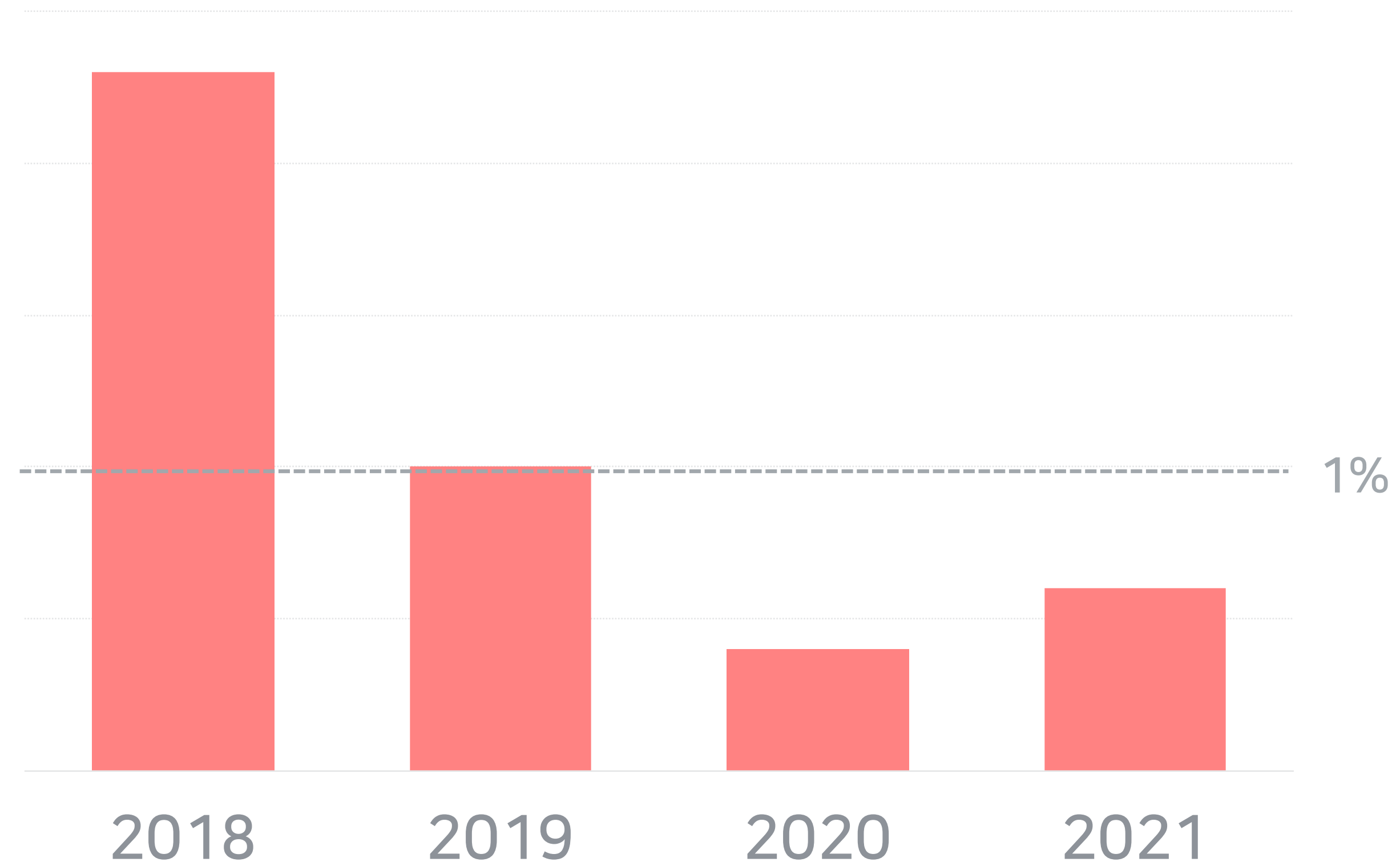
3. 새로운 KPI 개발 과정

3.1 새로운 KPI 사용 필요성 대두

장애 발생 (건) / 변경 배포 (건)

- 0.45% 달성 (네이버 검색 시스템, 2020 기준)
- 200건 당 약 1건의 장애 발생을 의미
- 현실적으로 달성하기 어려운 수치를 달성
- 우리 노력도 있었지만 운도 따라주었음
- 그러나 앞으로 KPI로 사용할 수 있을까?

장애율

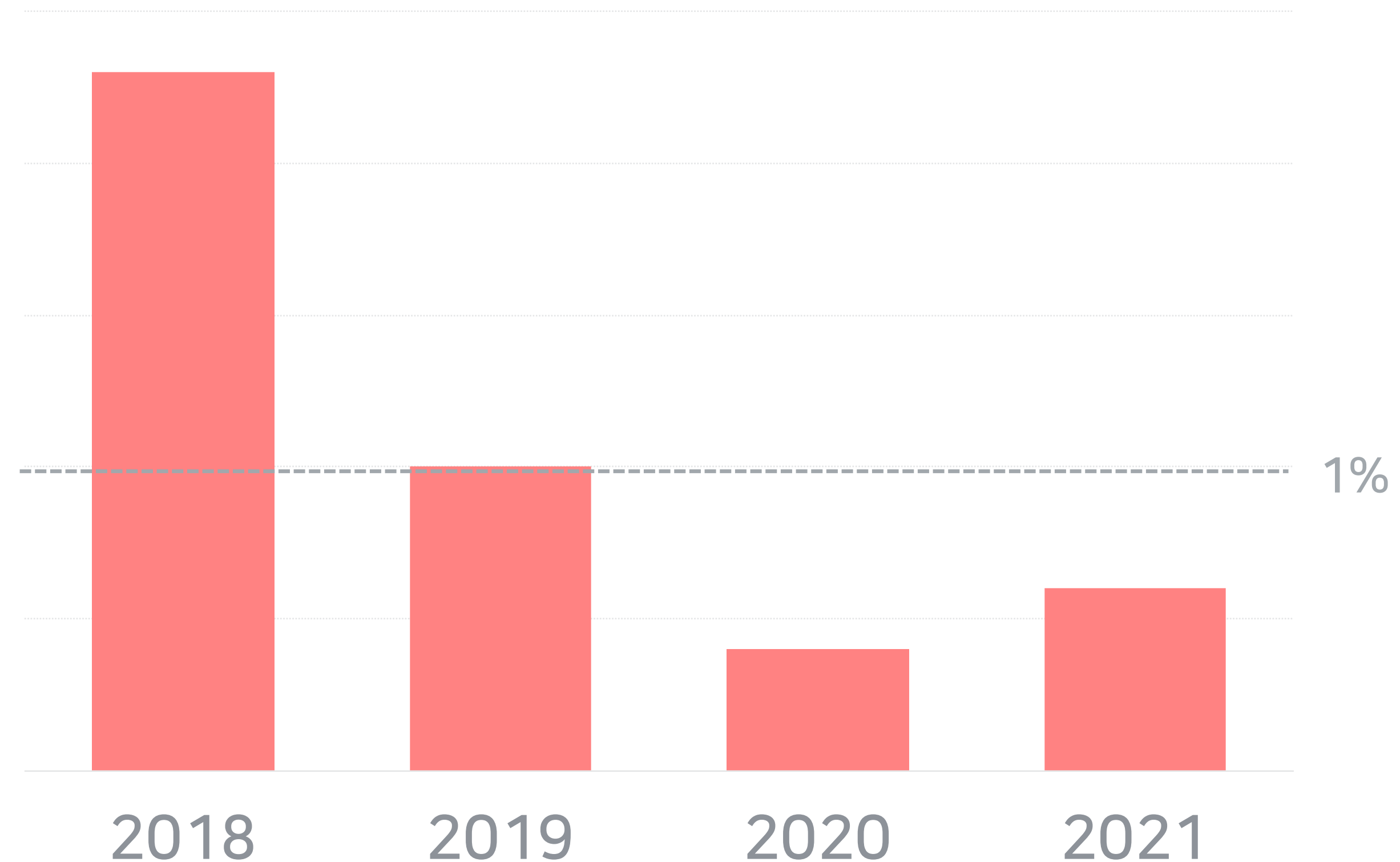


3.1 새로운 KPI 사용 필요성 대두

장애 발생 (건) / 변경 배포 (건)

- 0.45% 달성 (네이버 검색 시스템, 2020 기준)
- 200건 당 약 1건의 장애 발생을 의미
- 현실적으로 달성하기 어려운 수치를 달성
- 우리 노력도 있었지만 운도 따라주었음
- 그러나 앞으로 KPI로 사용할 수 있을까?

장애율



3.1 새로운 KPI 사용 필요성 대두

장애율을 줄이는 것은 중요하지만
장애율을 0%로 줄이는 것은 불가능

결국 장애는 발생할 수 밖에 없는 것

3.1 새로운 KPI 사용 필요성 대두

장애율은 더이상 개선하기 어려운 지점에 도달

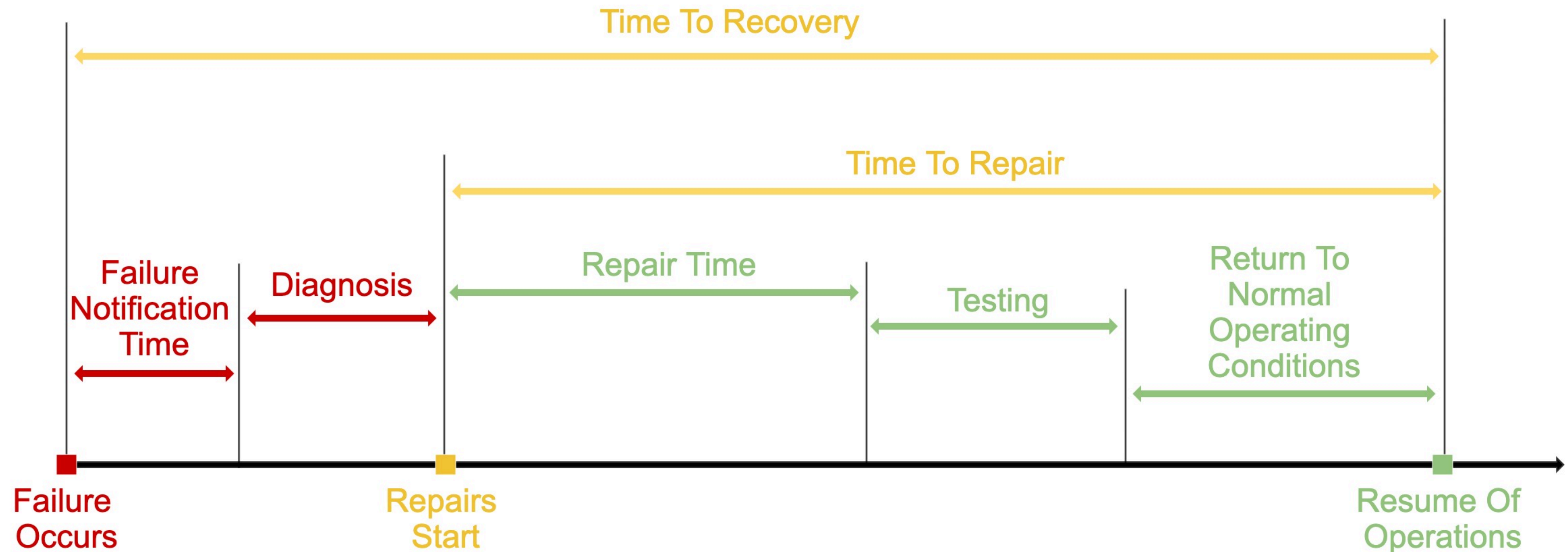
장애가 발생하였을 때 신속하게 복구하는 것도 중요

평균 장애 복구 시간 (MTTR)을 KPI로 사용하면 되지 않을까?

3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

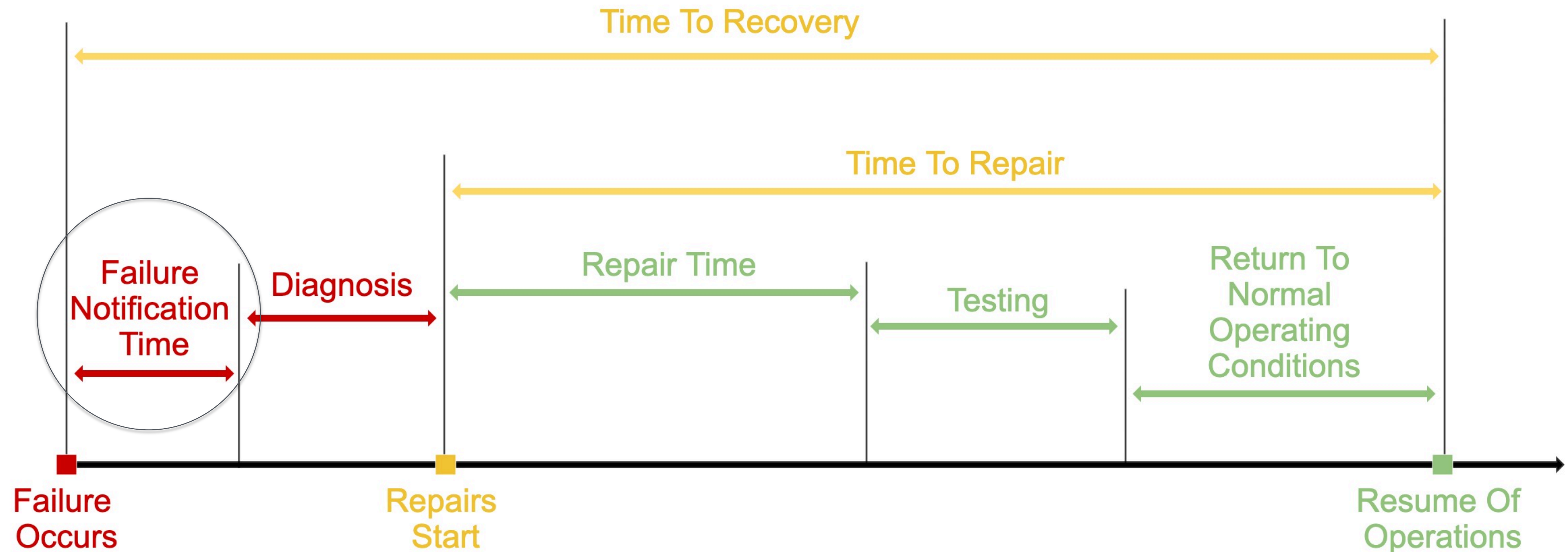
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

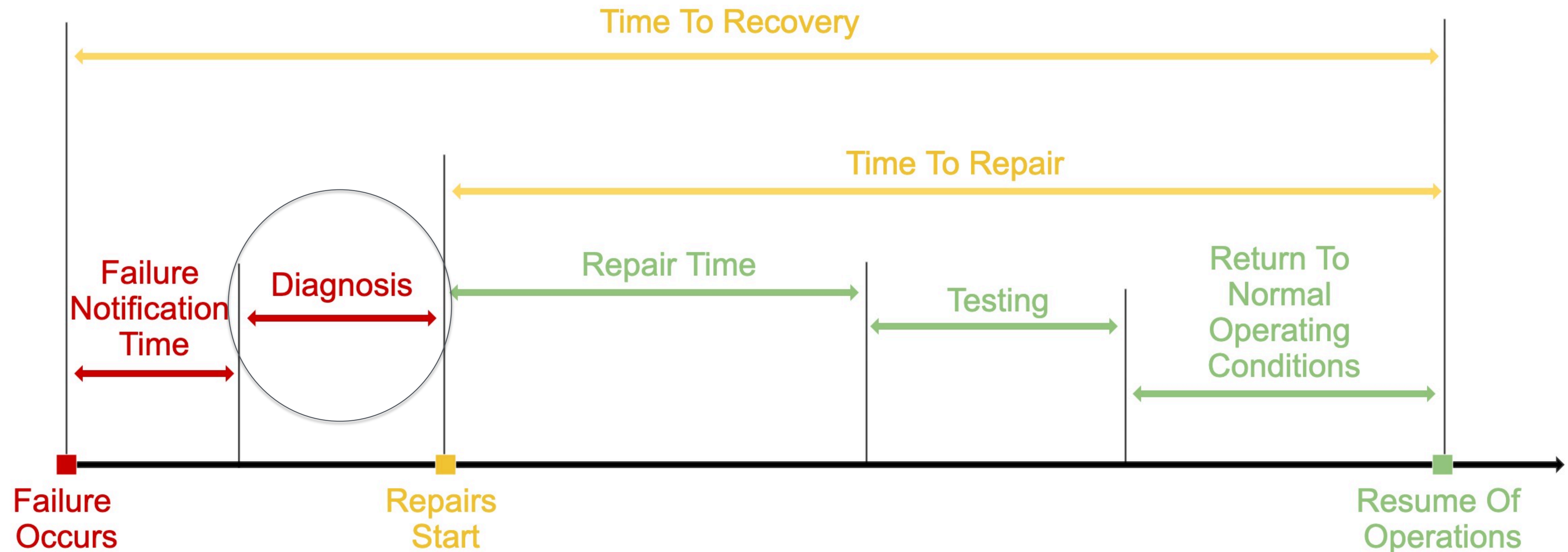
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

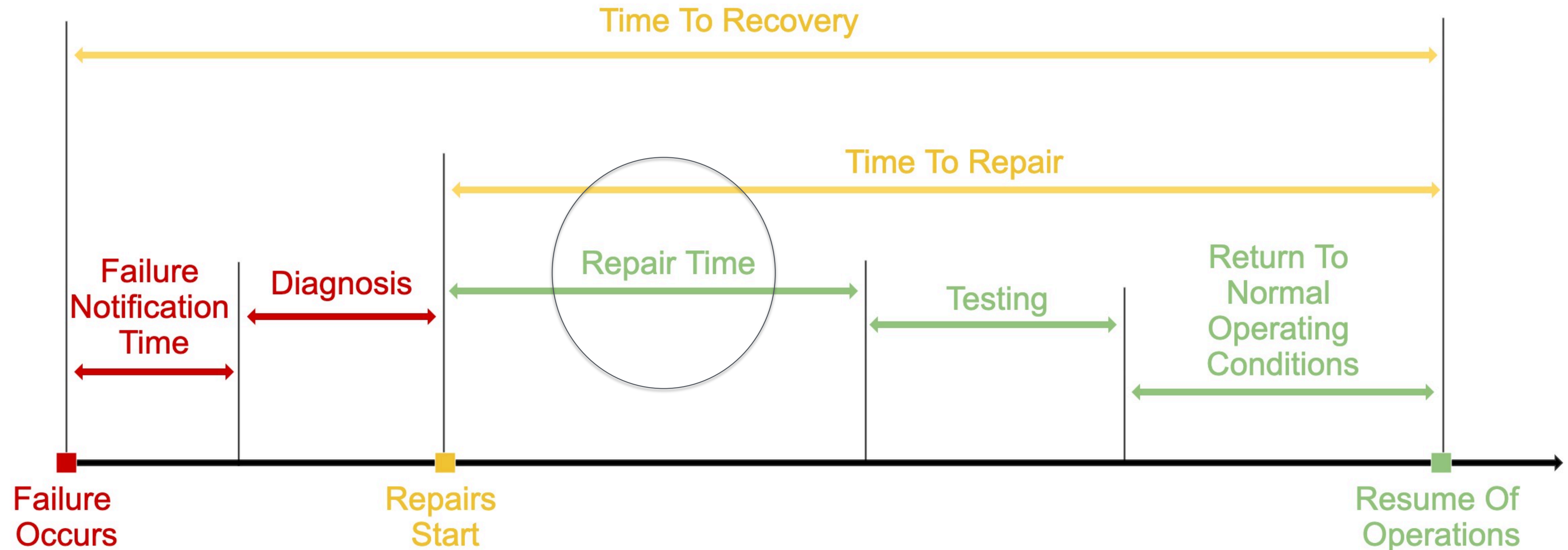
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

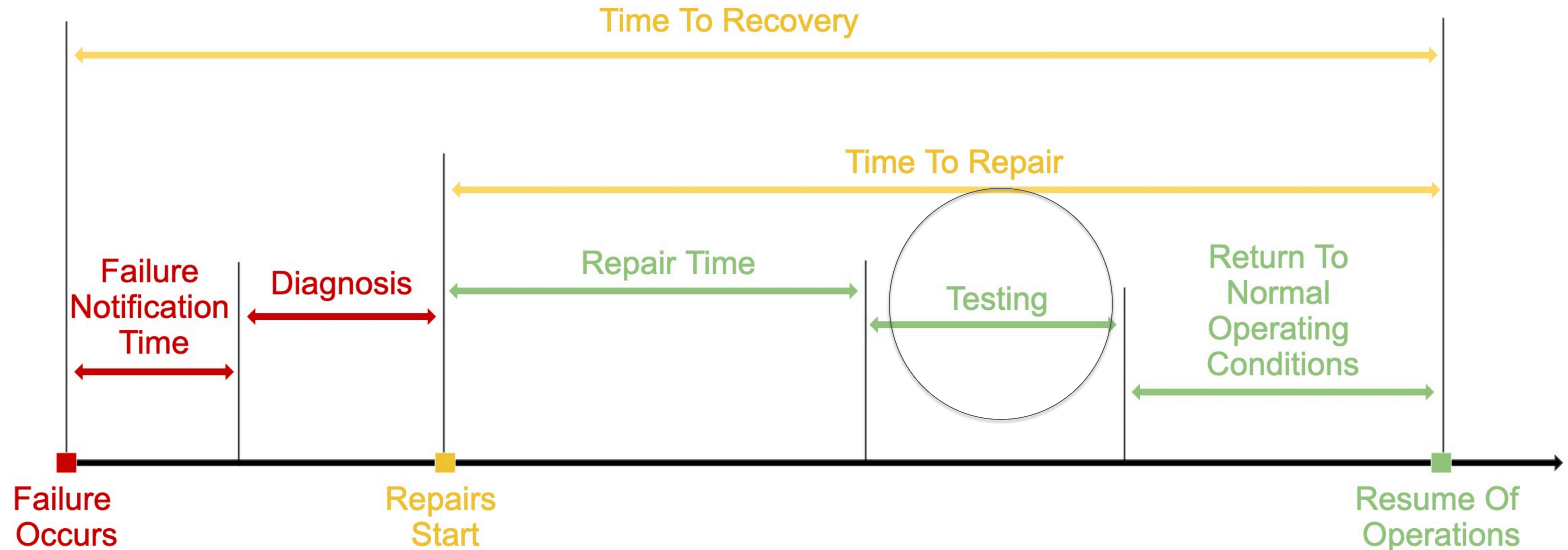
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

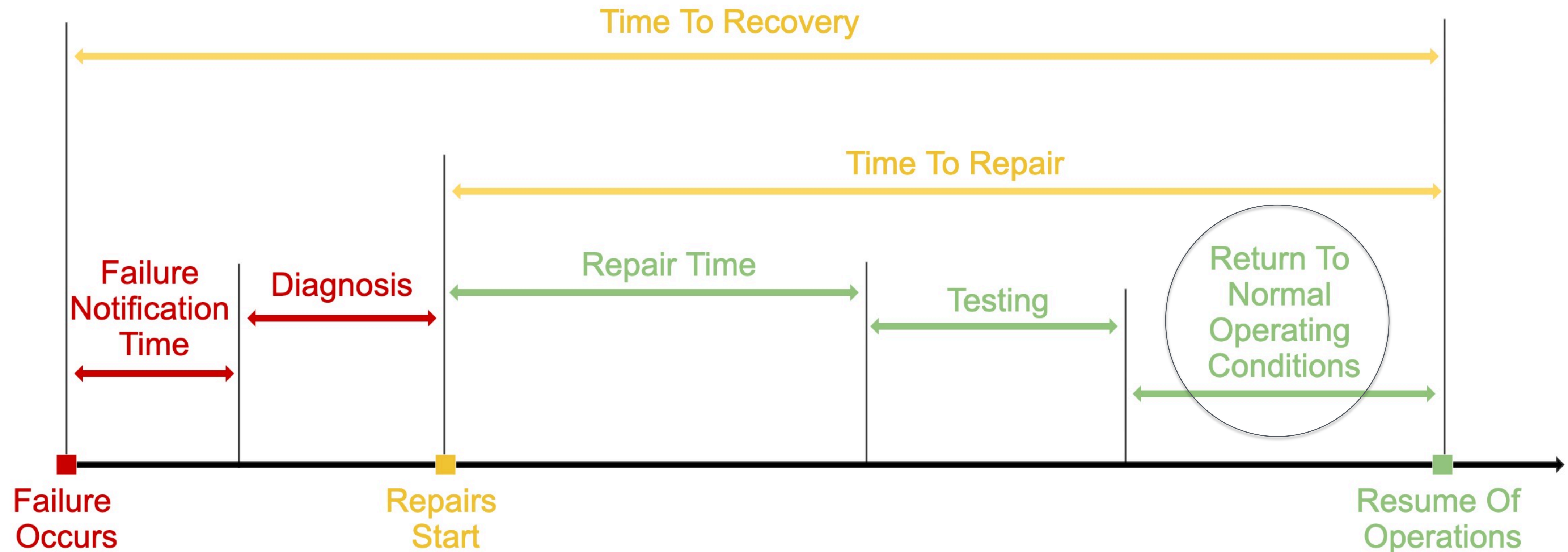
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

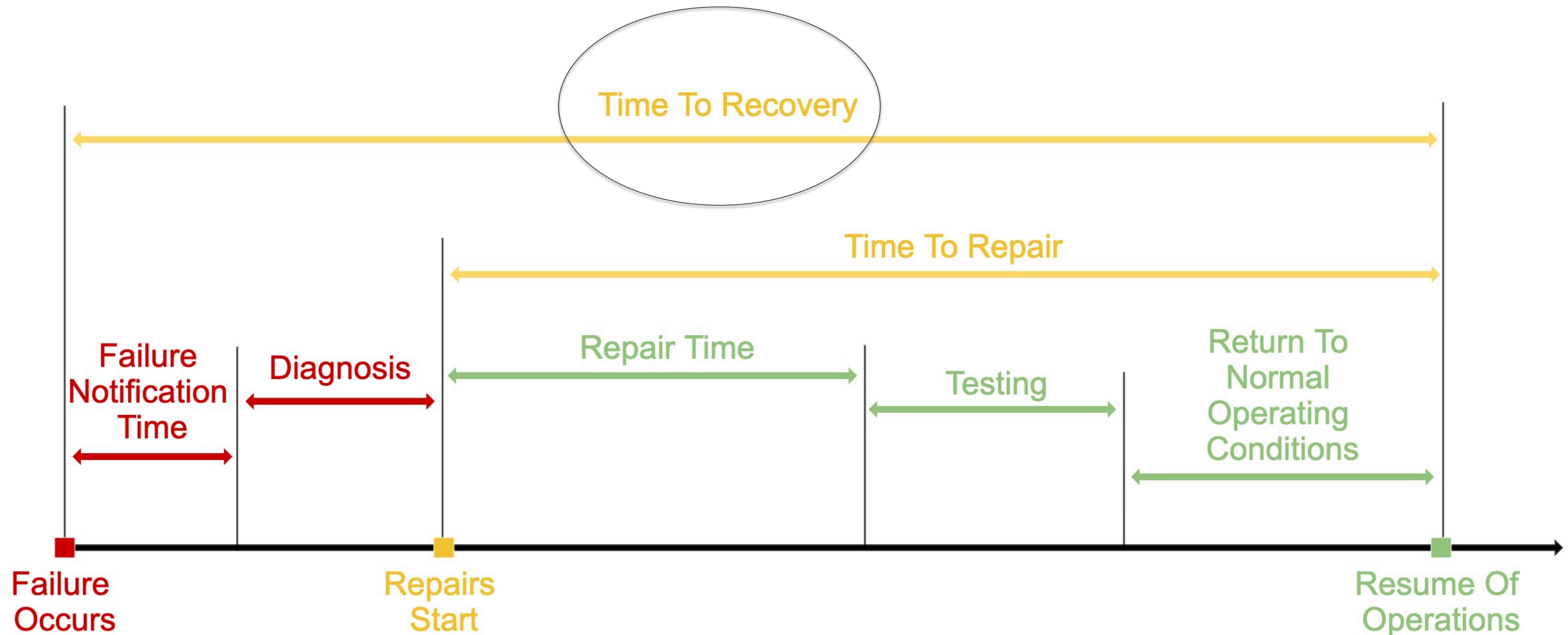
장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

MTTR - Mean Time To Recovery

장애의 발생 시점부터 복구까지 걸리는 시간의 평균



3.2 새로운 KPI 후보 및 선정

네이버 검색 시스템의 연도별 MTTR을 측정해보았다.

MTTR > 1hour

3.2 새로운 KPI 후보 및 선정

네이버 검색 시스템의 연도별 MTTR을 측정해보았다.

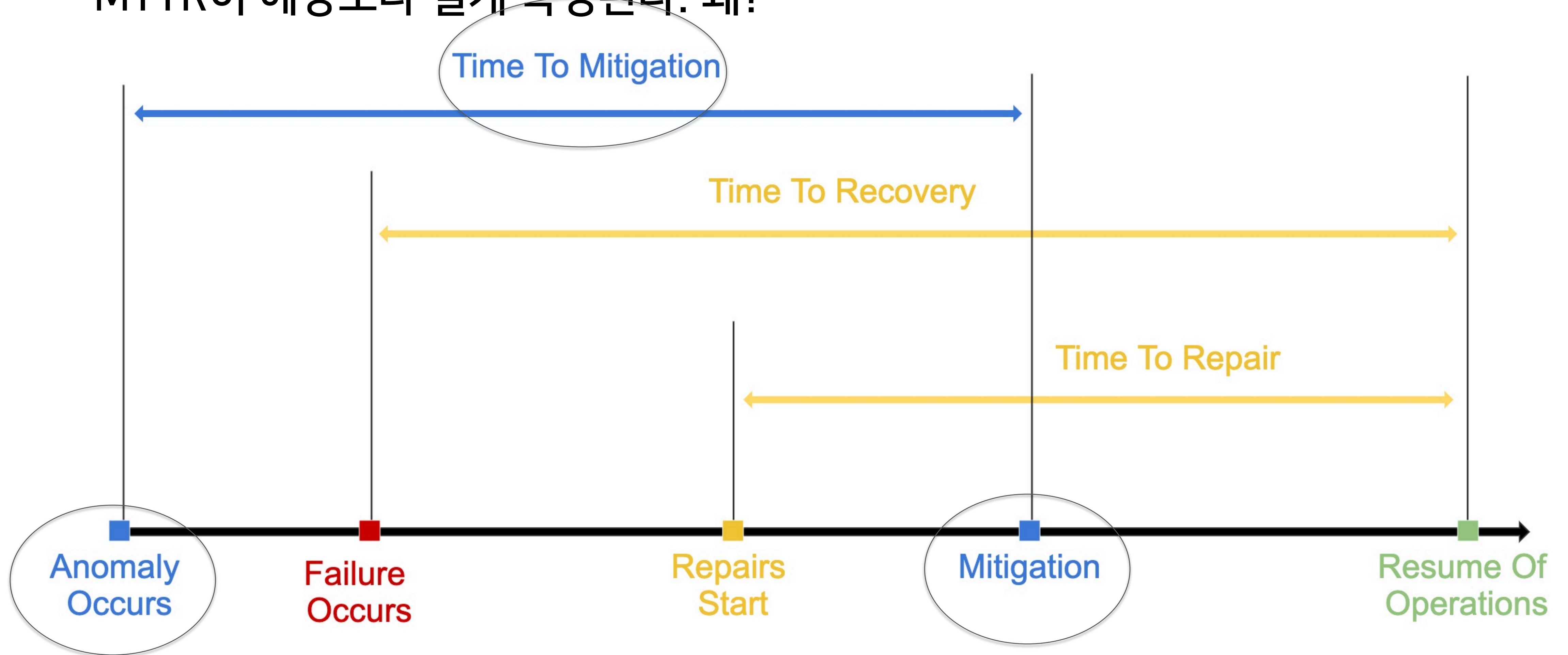
MTTR > 1hour

‘검색 서비스가 실제로 1시간 이상동안 장애가 지속되었던적이 많이 있었던가..?’

예상했던 값보다 너무나 긴 수치들이 나옴

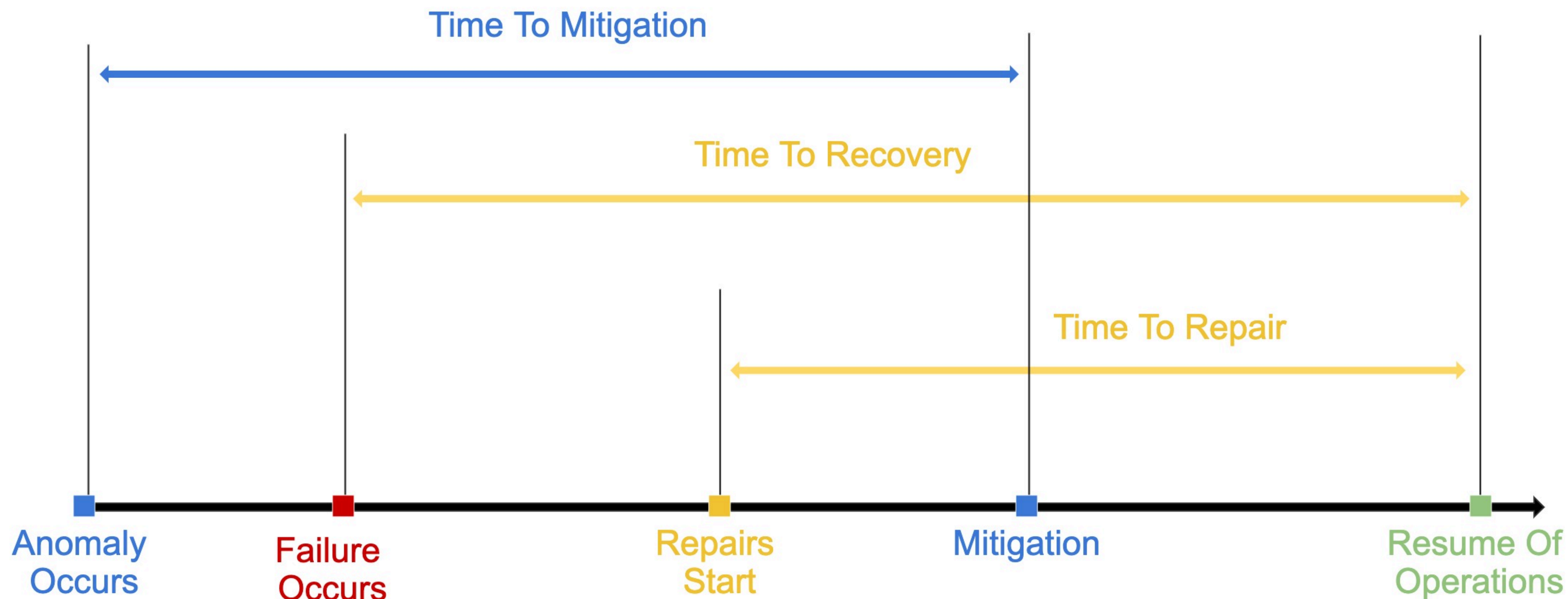
3.3 MTTM 소개

MTTRO이 예상보다 길게 측정된다. 왜?

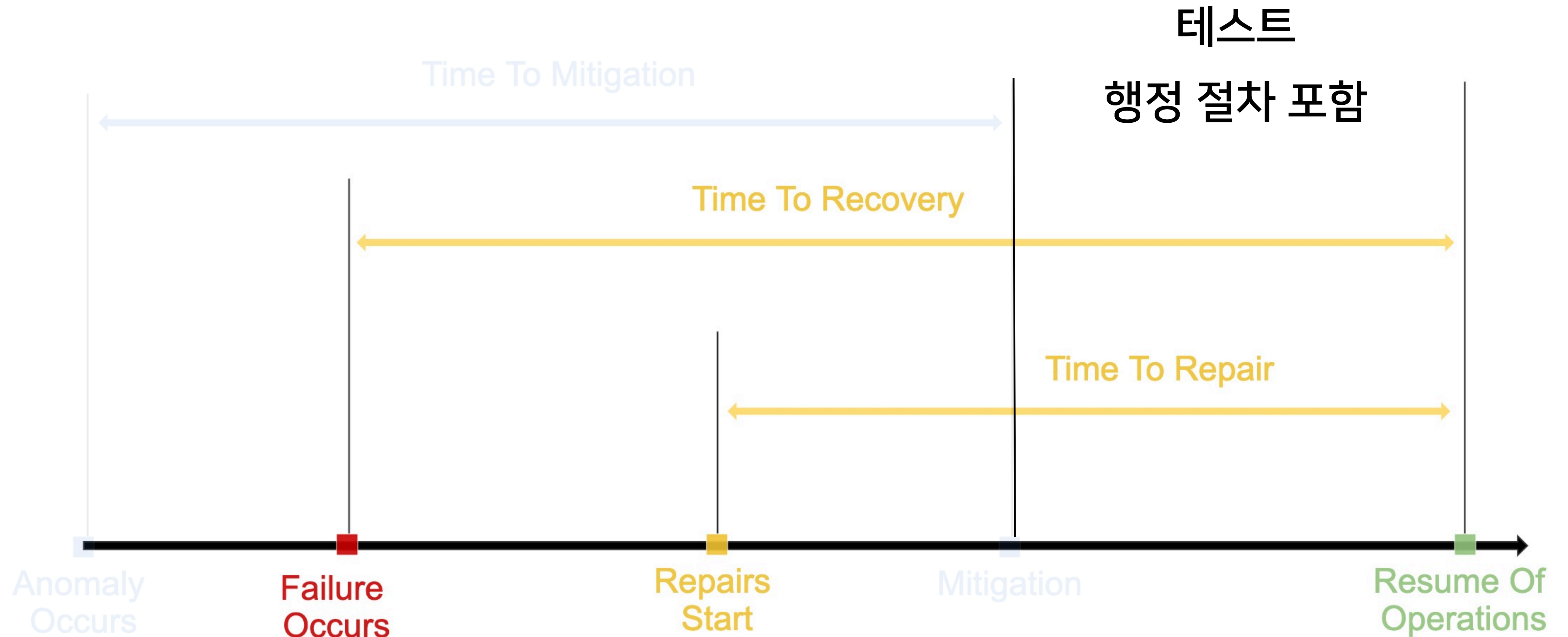


3.3 MTTM 소개

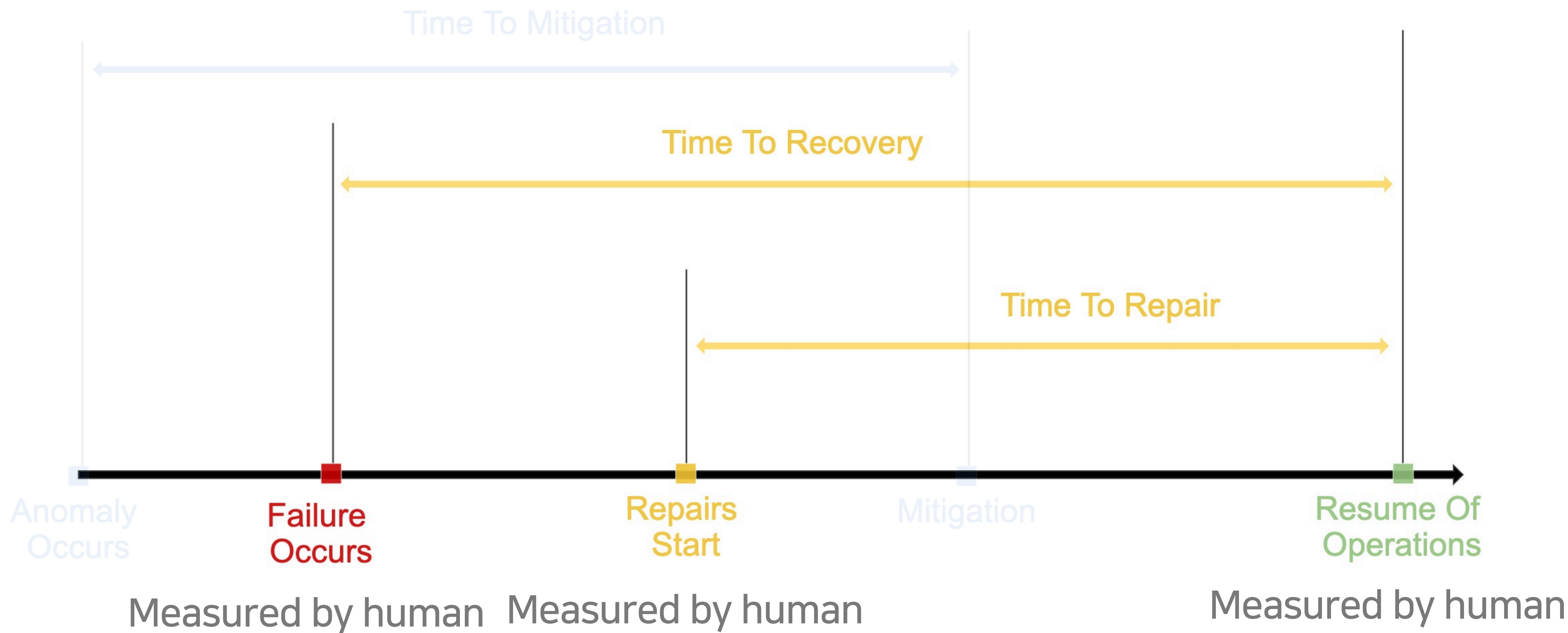
MTTRO이 예상보다 길게 측정된다. 왜?



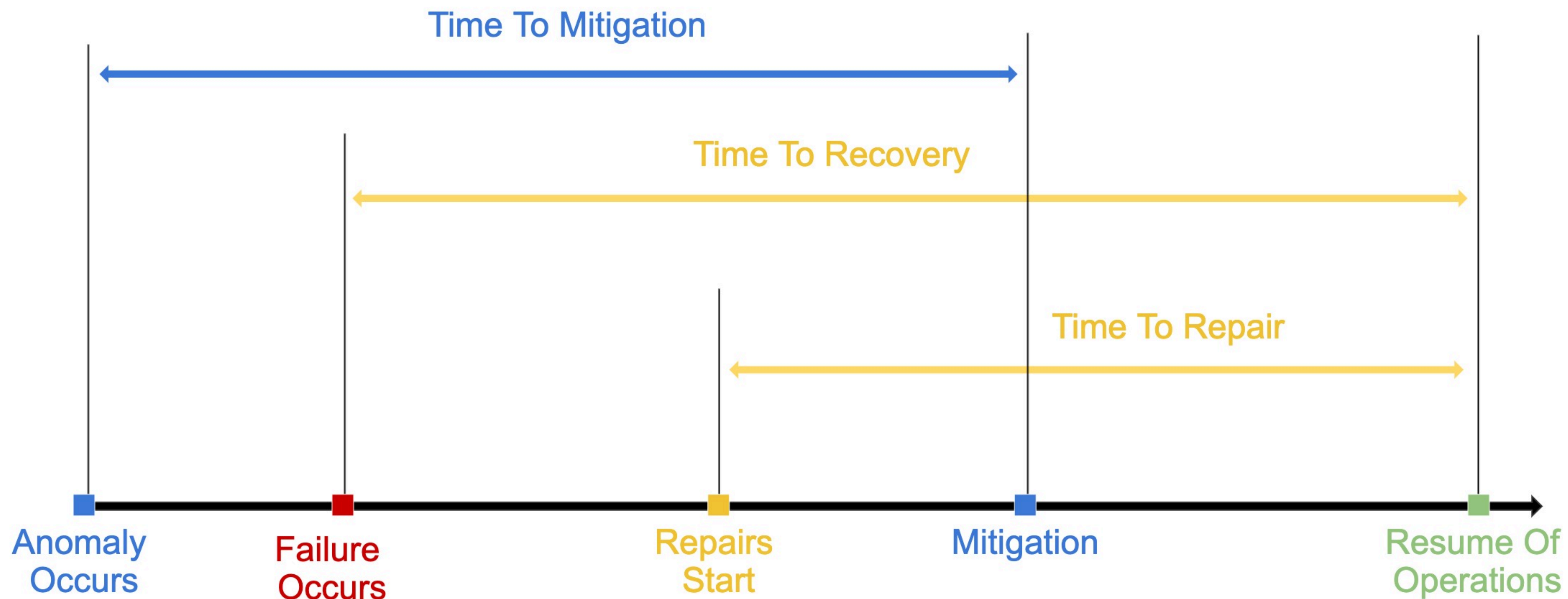
3.3 MTTM 소개



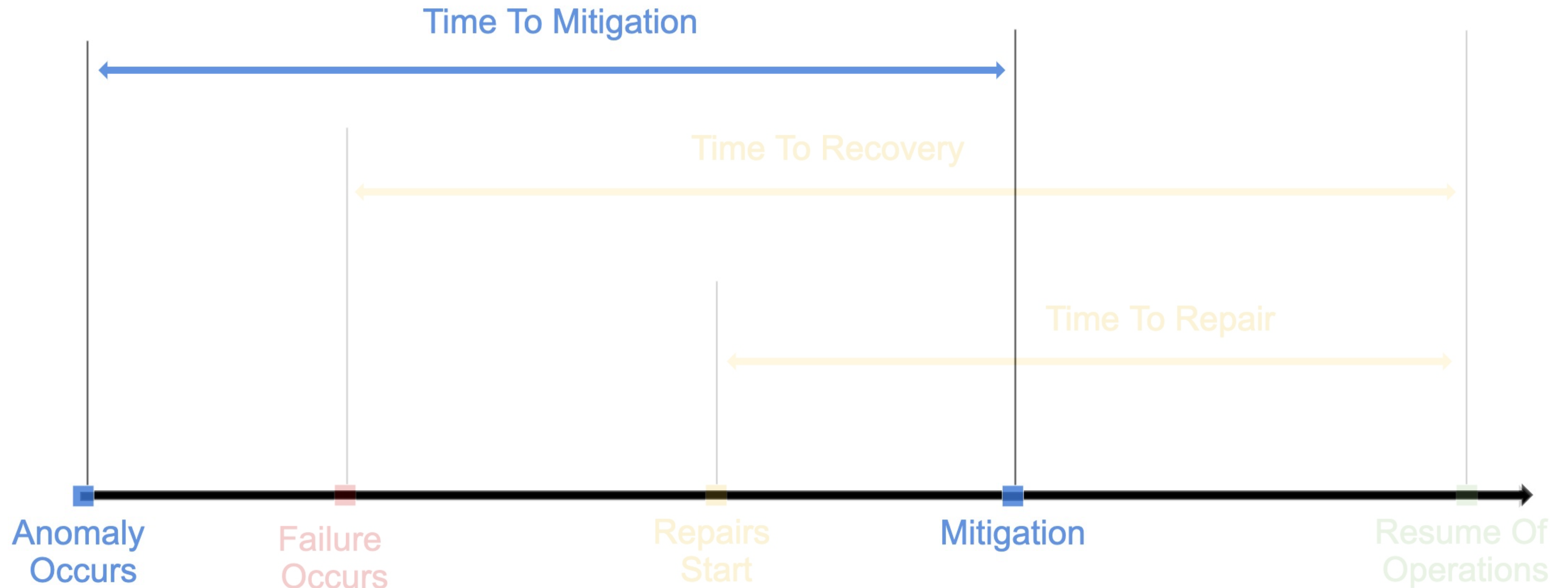
3.3 MTTM 소개



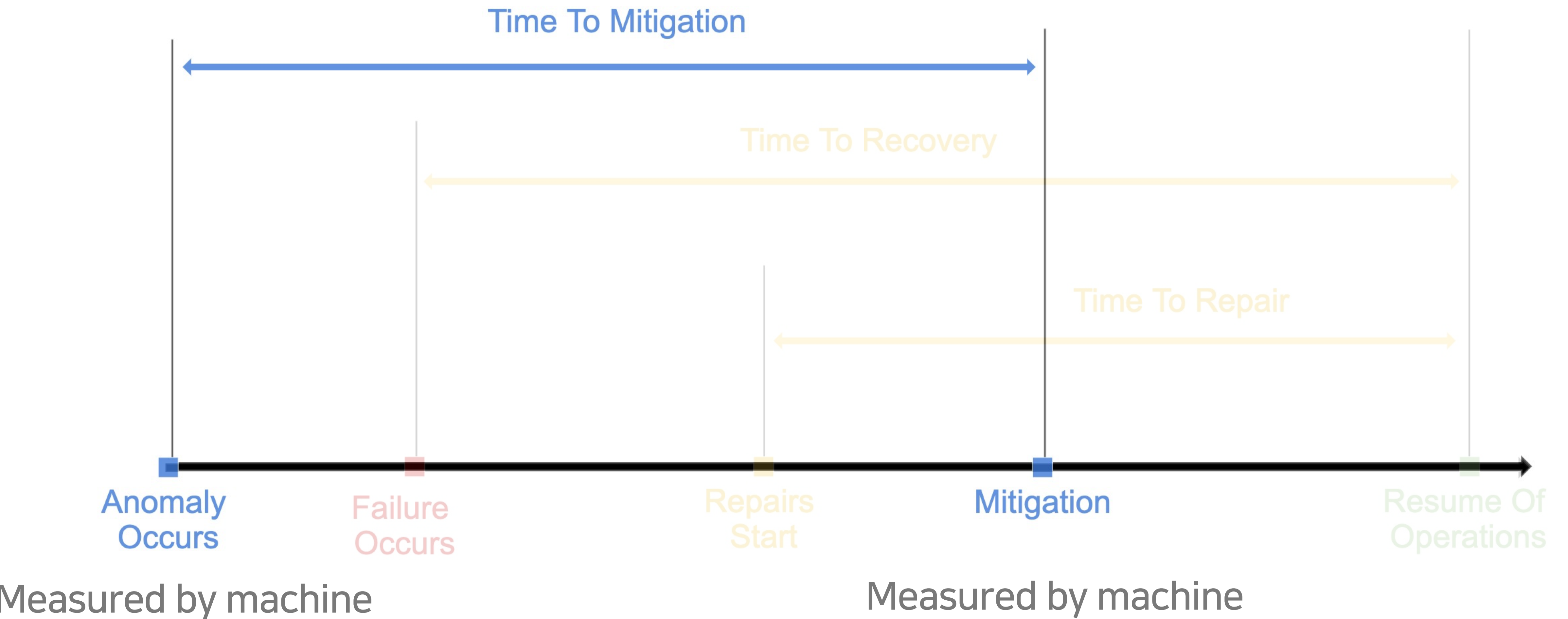
3.3 MTTM 소개



3.3 MTTM 소개



3.3 MTTM 소개

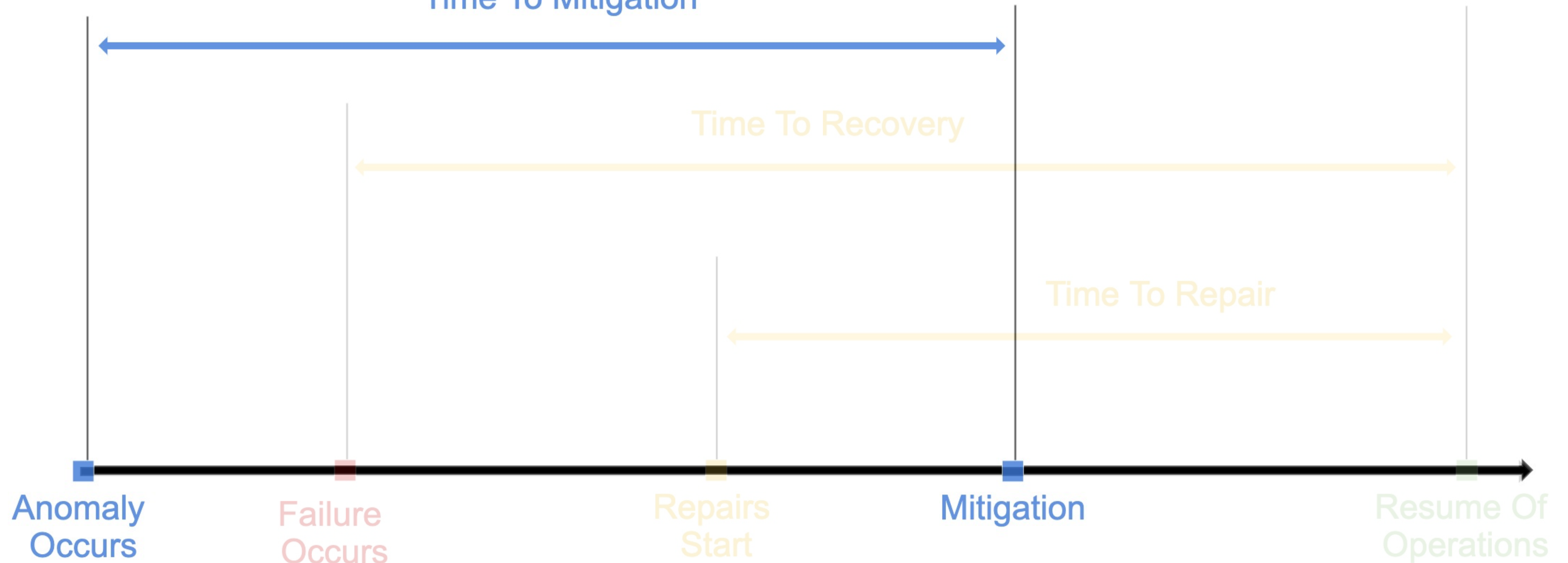


3.3 MTTM 소개

Impact On Customer

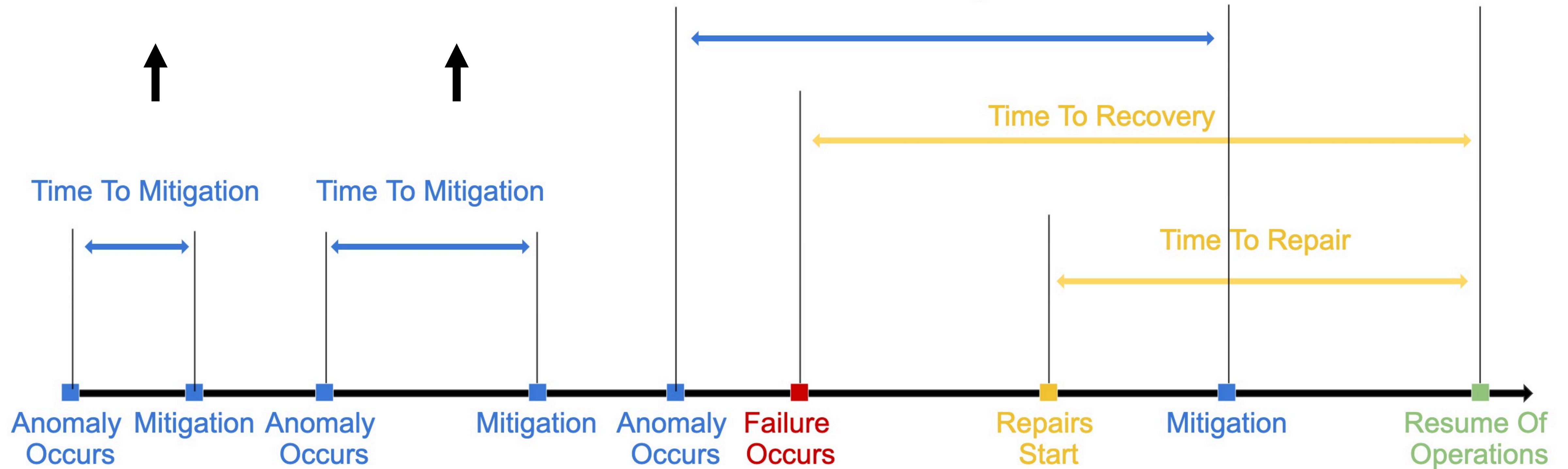


Time To Mitigation



3.3 MTTM 소개

장애(Failure)로 이어지지 않은
이상증상(Anomaly) 또한 지표로 포함



3.3 MTTM 소개

1. 감지 대상을 **장애(Failure)**로 보는 것이 아닌,

이상 증상 (Anomaly)

예) 응답시간 지연
타임아웃 발생
가용량 부족

장애 (Failure)
예) 검색중단

3.3 MTTM 소개

1. 감지 대상을 **장애(Failure)**로 보는 것이 아닌, 서비스의 **이상 증상(Anomaly)**을 측정 데이터로 사용

이상 증상 (Anomaly)

예) 응답시간 지연
타임아웃 발생
가용량 부족

장애 (Failure)
예) 검색중단

3.3 MTTM 소개

1. 감지 대상을 **장애(Failure)**로 보는 것이 아닌, 서비스의 **이상 증상(Anomaly)**을 측정 데이터로 사용
2. **이상 증상(Anomaly)**의 감지(Detection)는 네이버 검색 모니터링 시스템을 통해서

이상 증상 (Anomaly)

예) 응답시간 지연
타임아웃 발생
가용량 부족

장애 (Failure)
예) 검색중단

3.3 MTTM 소개

MTTM을 통한 개선점 (as-is MTTR)

사용자
관점

측정
자동화

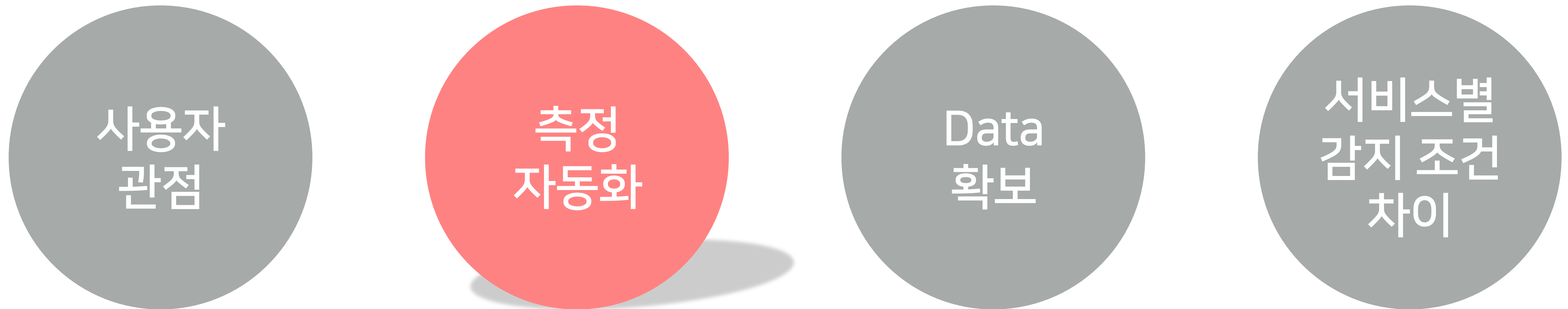
Data
확보

서비스별
감지 조건
차이

- MTTR은 운영자 관점에서의 영향도
- MTTM은 실제 고객(사용자)에 대한 영향도

3.3 MTTM 소개

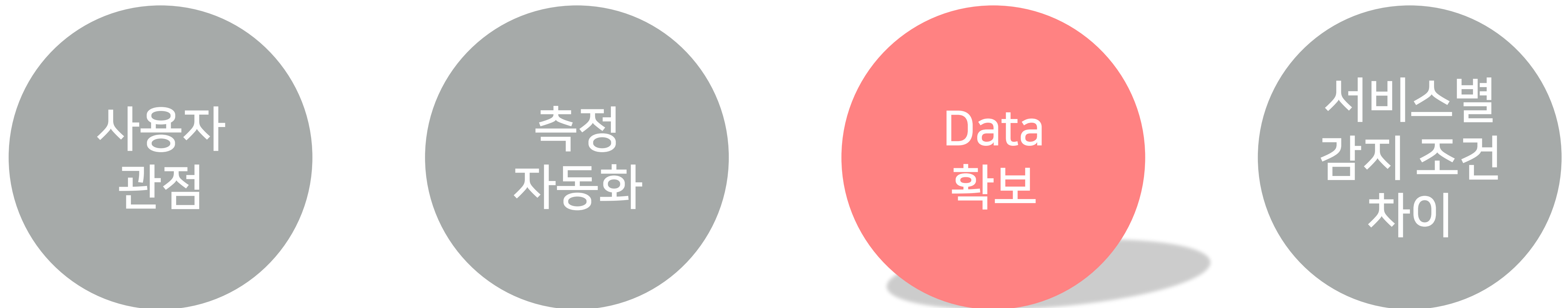
MTTM을 통한 개선점 (as-is MTTR)



- MTTR은 사람의 개입이 많은 방식 => 반영 지연 및 오류 발생 가능
- MTTM은 자동적이고 기계적인 방식

3.3 MTTM 소개

MTTM을 통한 개선점 (as-is MTTR)

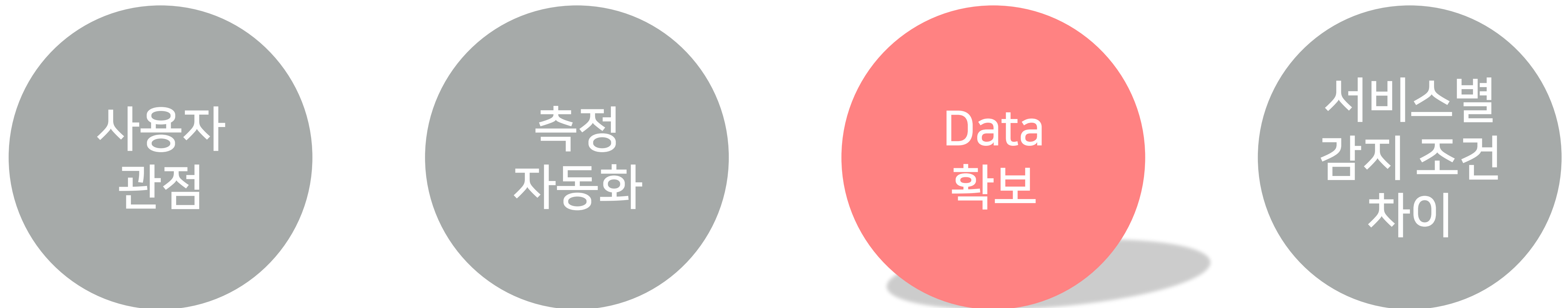


- MTTR - 연간 10~ 20 건 장애(Failure) 발생
- MTTM - 연간 수천 건 이상 증상(Anomaly) 발생

(Search Platform 조직, 2020 기준)

3.3 MTTM 소개

MTTM을 통한 개선점 (as-is MTTR)



- MTTR - 연간 10~ 20 건 **장애(Failure)** 발생
- MTTM - 연간 수천 건 **이상 증상(Anomaly)** 발생

(Search Platform 조직, 2020 기준)

3.3 MTTM 소개

MTTM을 통한 개선점 (as-is MTTR)



- MTTR - 서비스별 **장애(Failure)** 감지 조건의 차이 발생
- MTTM - 검색팀의 자체적인 기준으로 **이상증상(Anomaly)** 감지

3.3 MTTM 소개

Mean Time To Recovery

장애(Failure)가 발생한 시점부터
복구(Recovery) 완료 시점까지 걸린 시간의 평균

3.3 MTTM 소개

Mean Time To Mitigation

서비스의 이상 증상(Anomaly)이 감지(Detection)되어
경보가 발생한 시점부터 증상의 완화(Mitigation)까지 걸린 시간의 평균

3.4 MTTM 개발 과정



3.4 MTTM 개발 과정

2020년도 1년치 데이터 검수

3.4 MTTM 개발 과정

발열

수면 장애

알 수 없는 통증

잦은 피로

건강에 이상이 있는지
확인할 수 있는 증상



3.4 MTTM 개발 과정

트래픽 이상 경고

가용량 임계 경고

Timeout 경고

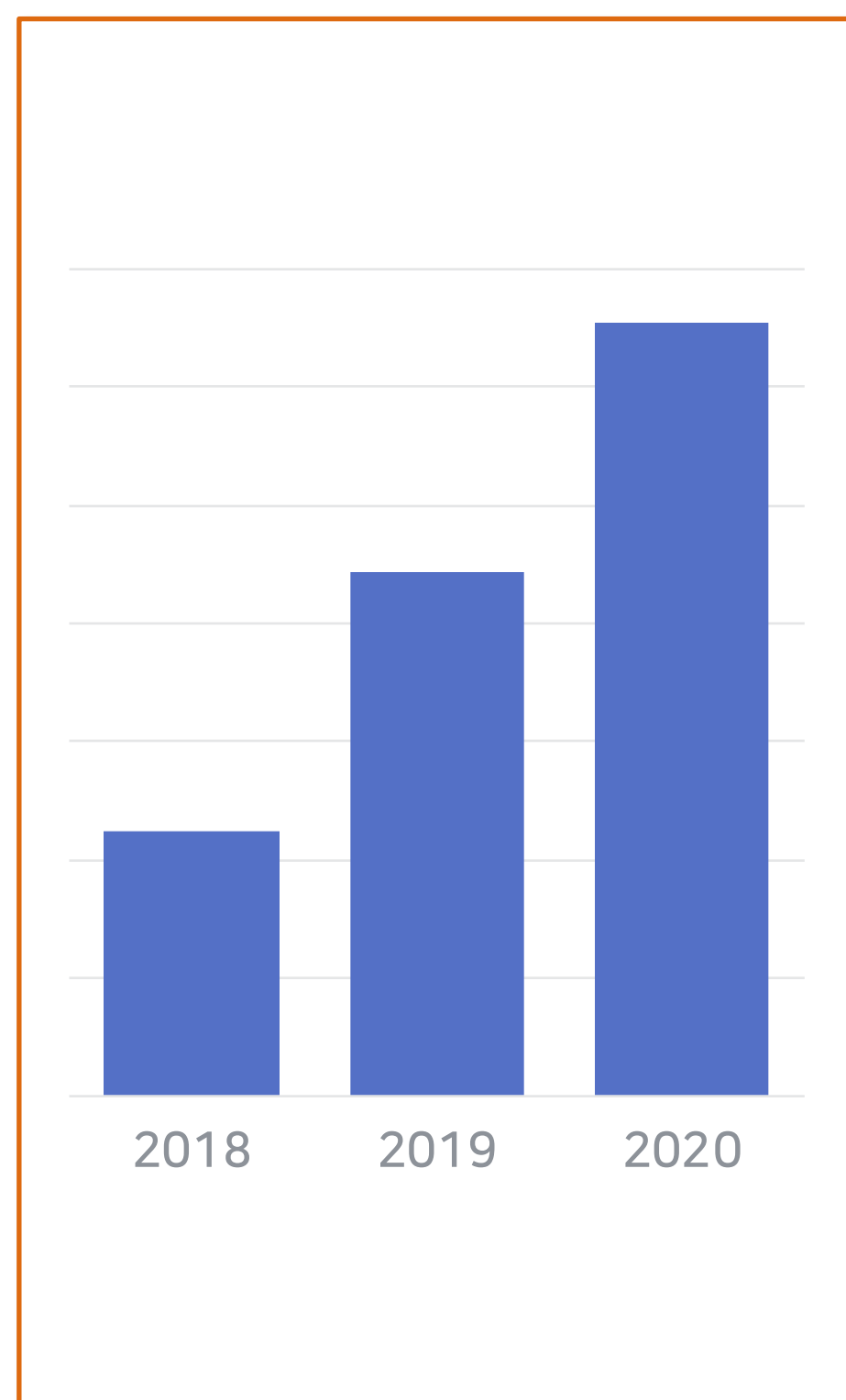
평균응답시간 경고

장애로 번질 수도 있는
이상증상 경고

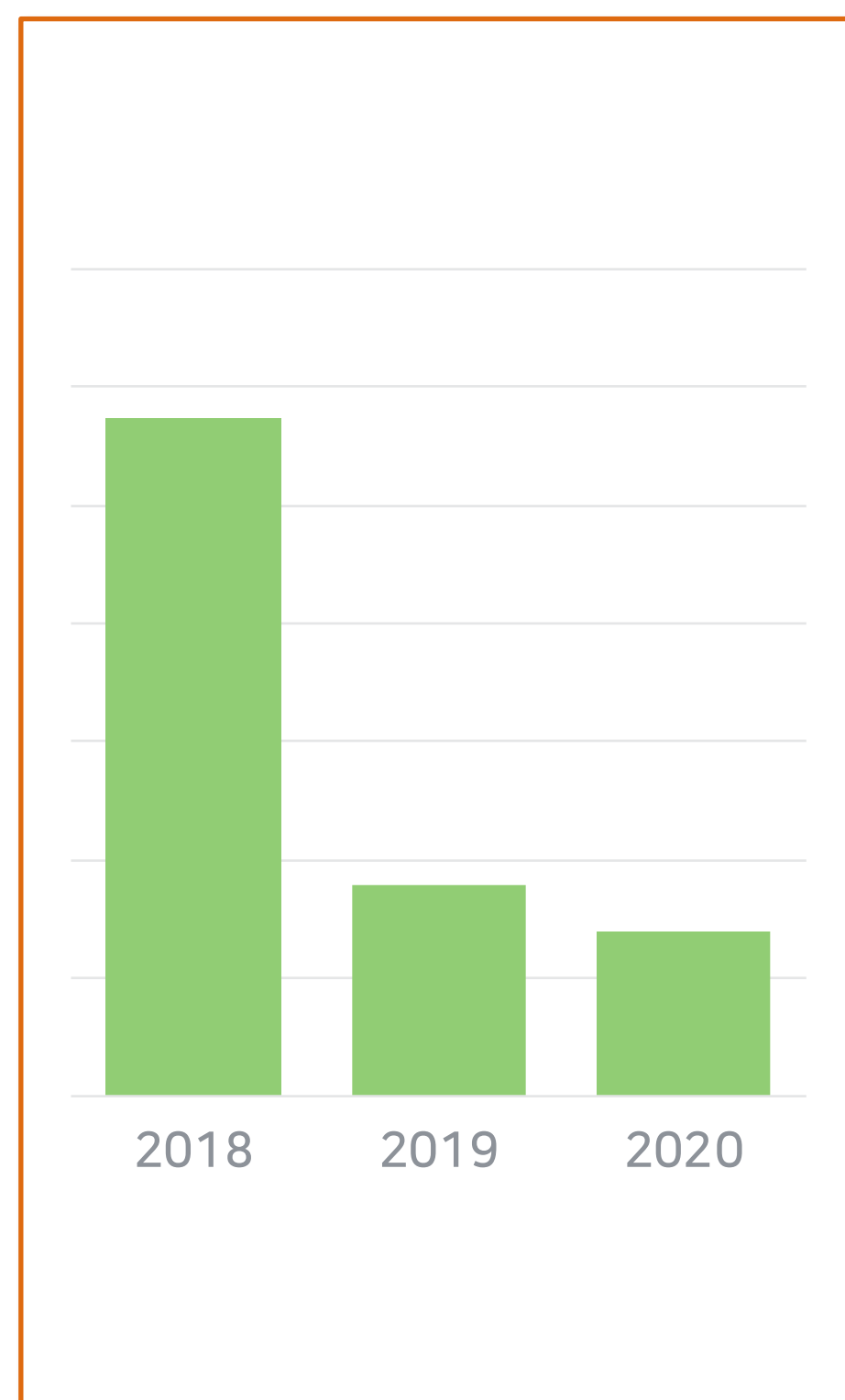


3.4 MTTM 개발 과정

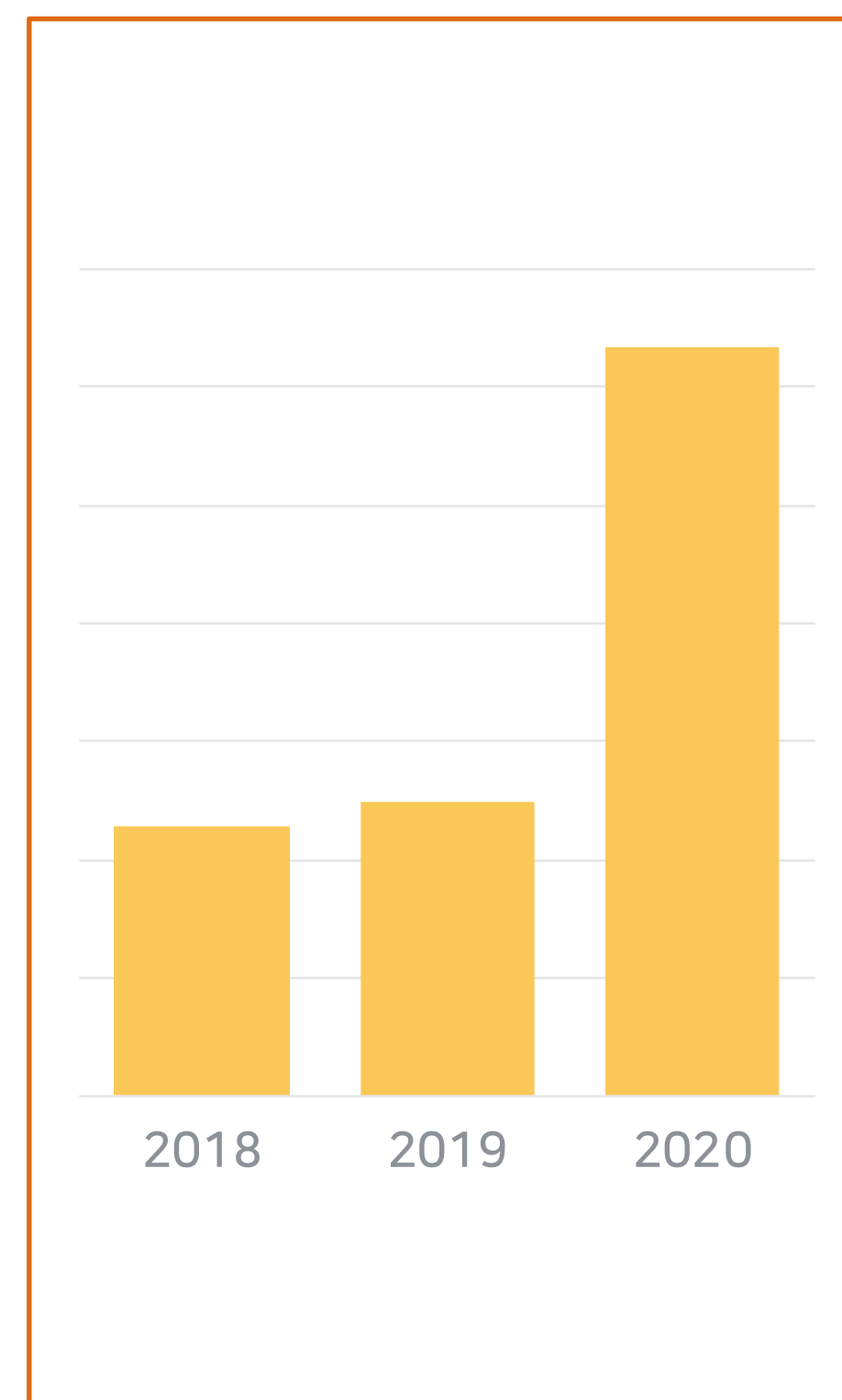
트래픽 이상 경보



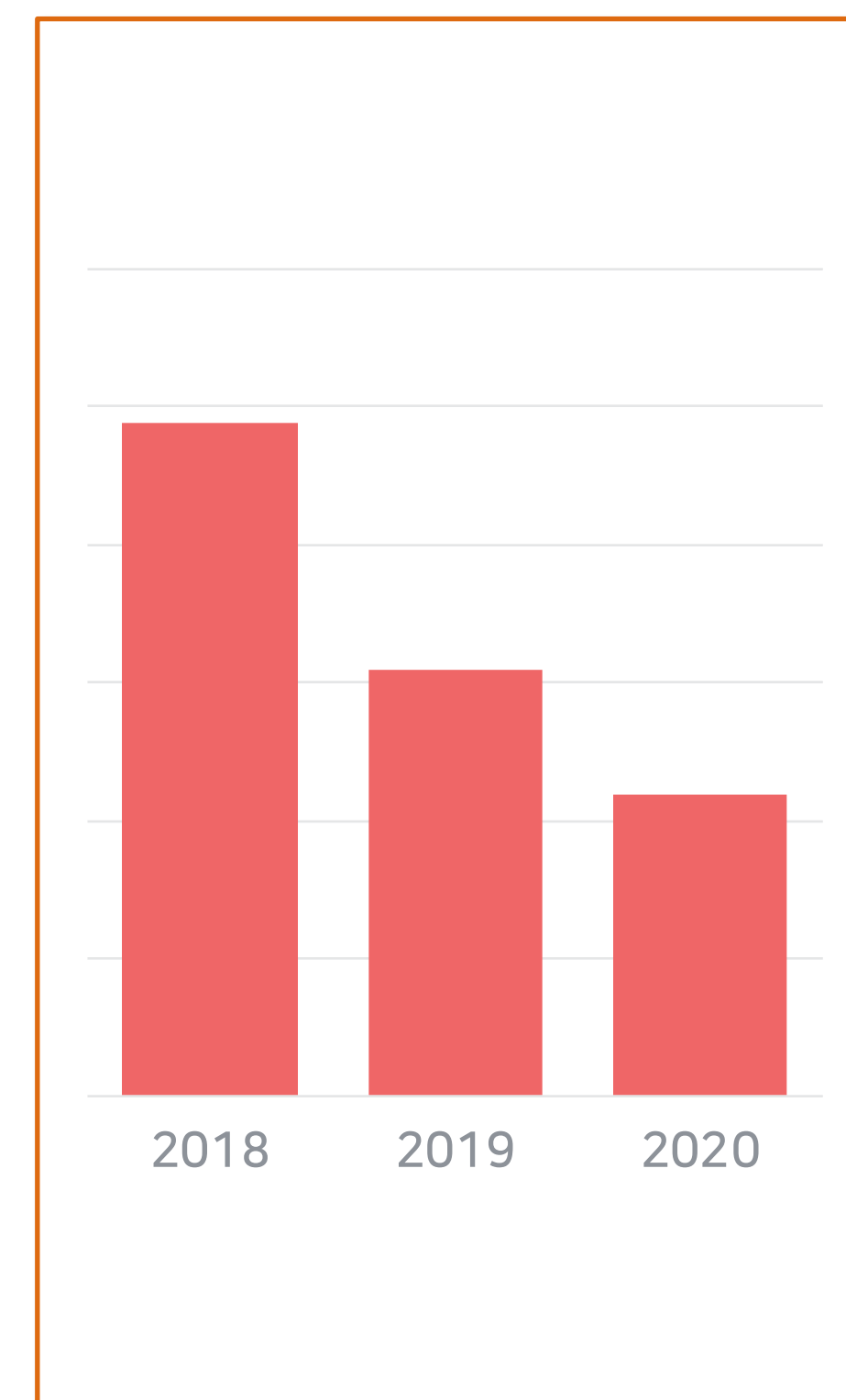
가용량 임계 경보



Timeout 경보

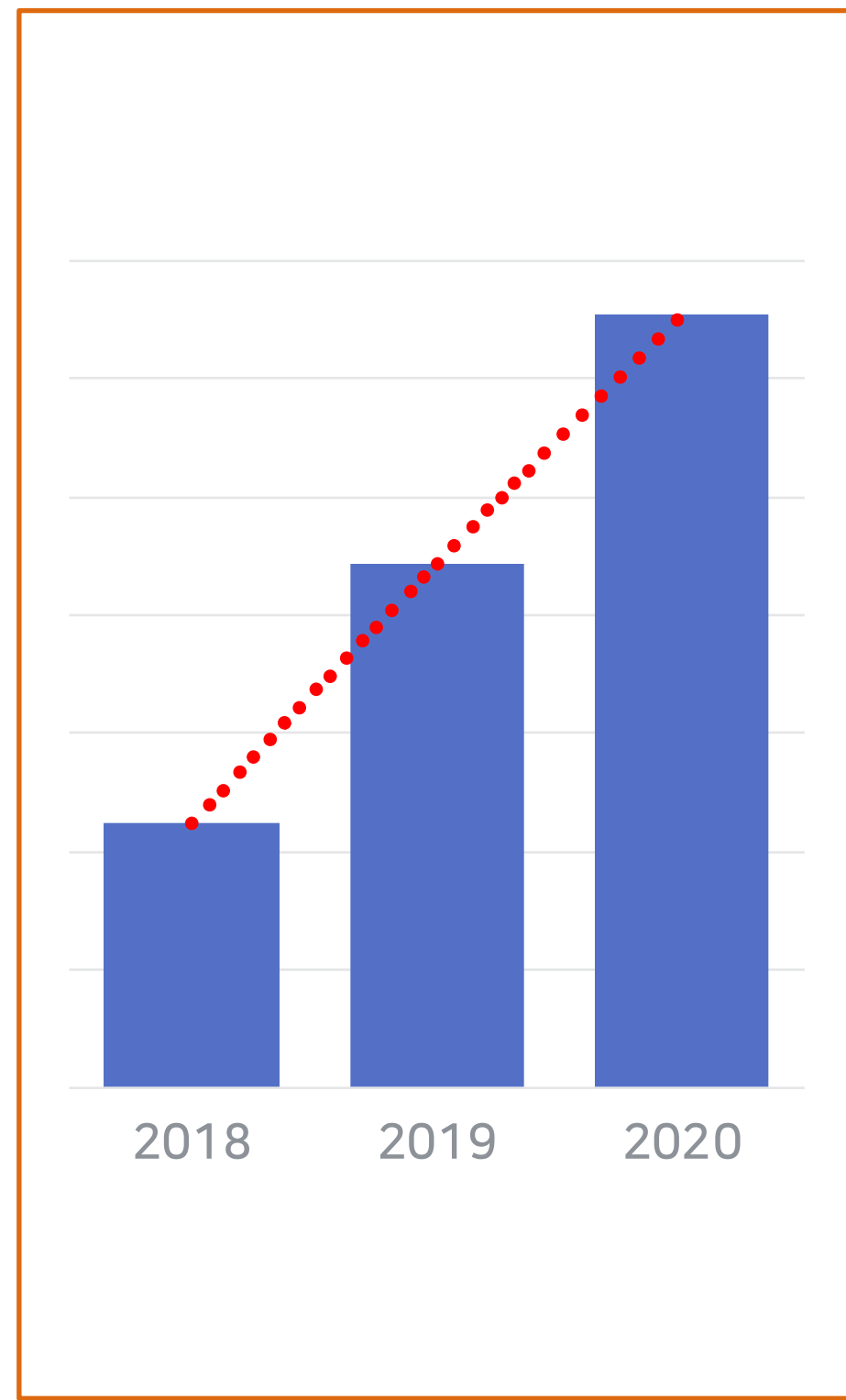


평균응답시간 경보

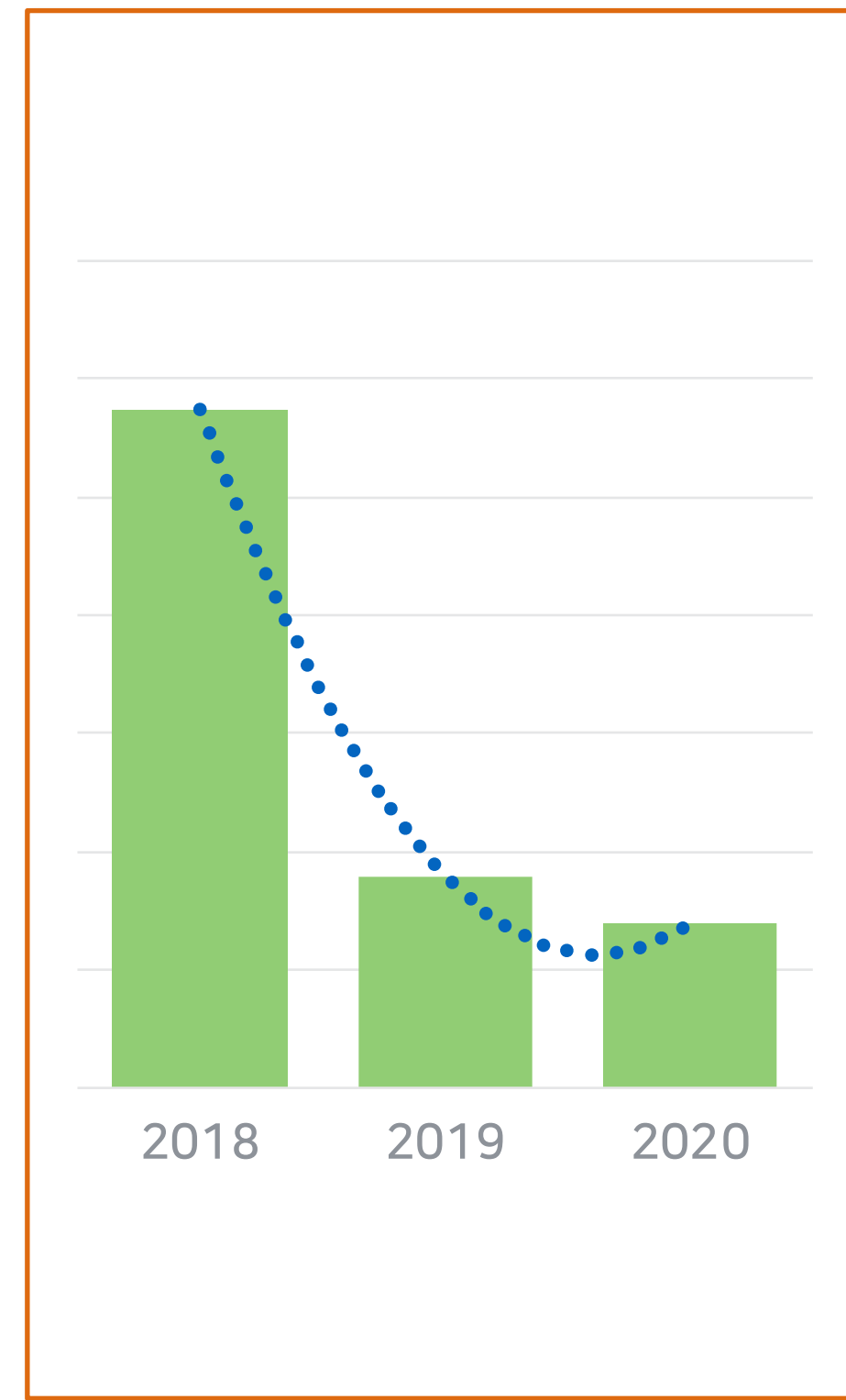


3.4 MTTM 개발 과정

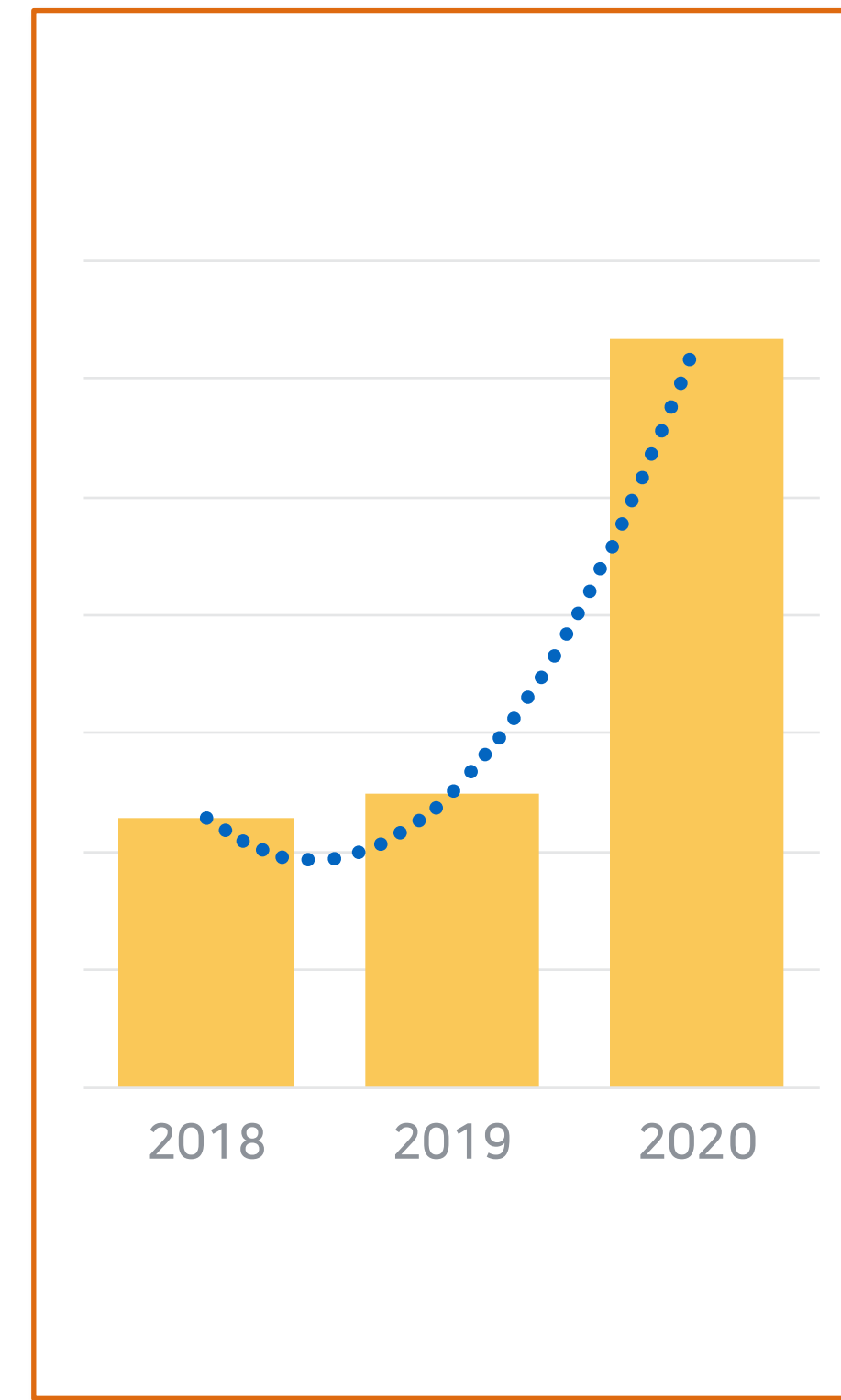
트래픽 이상 경보



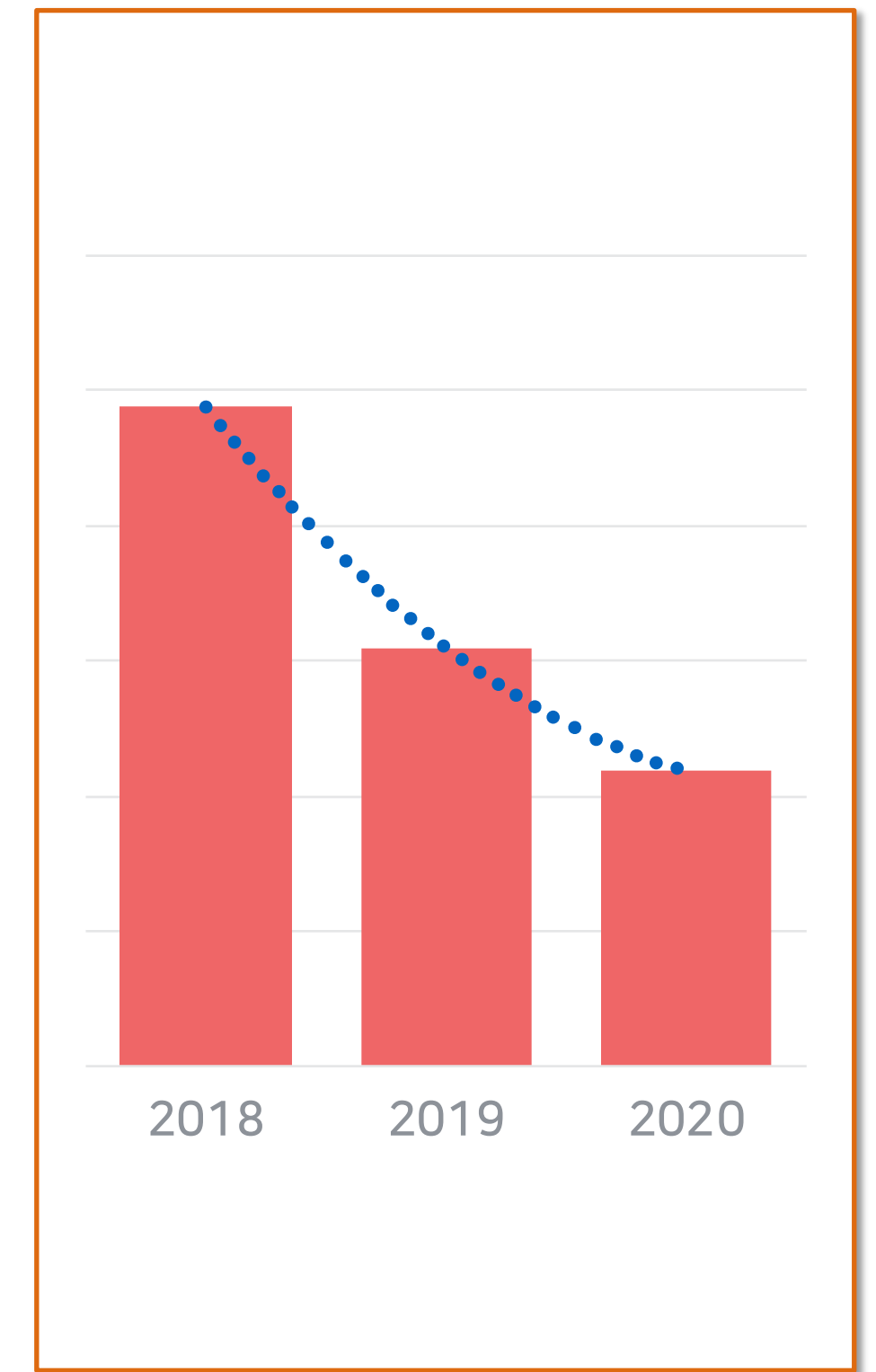
가용량 임계 경보



Timeout 경보

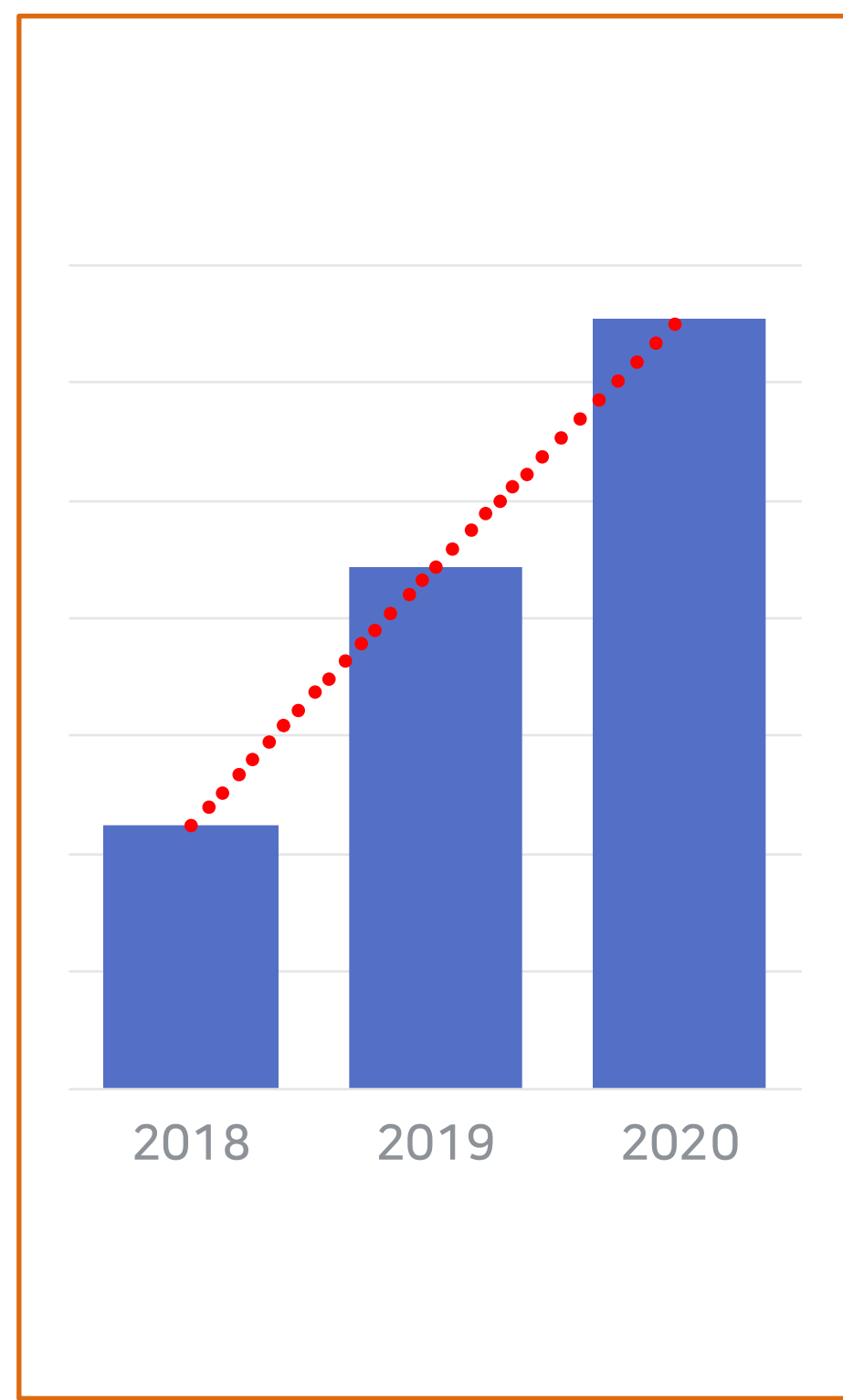


평균응답시간 경보

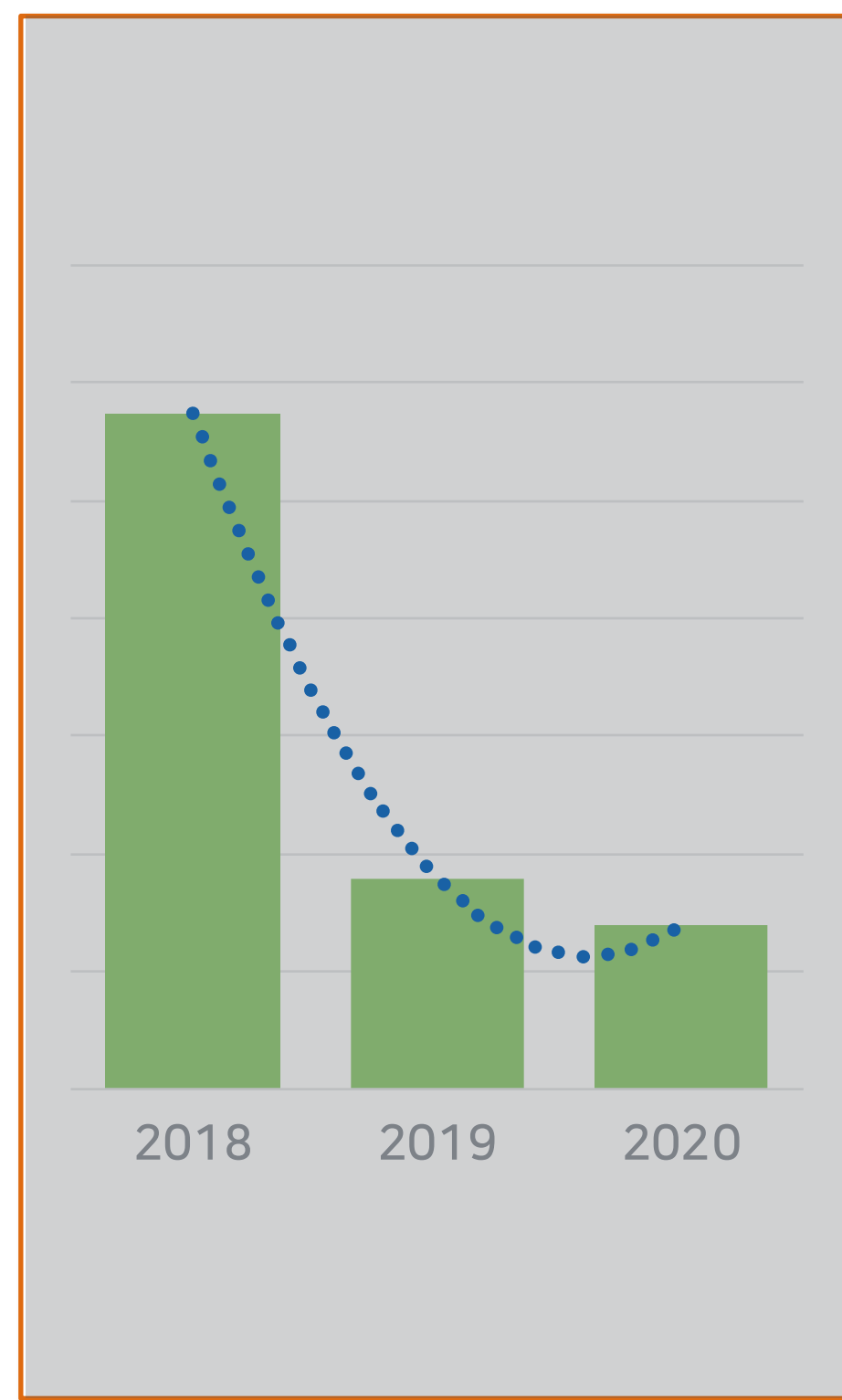


3.4 MTTM 개발 과정

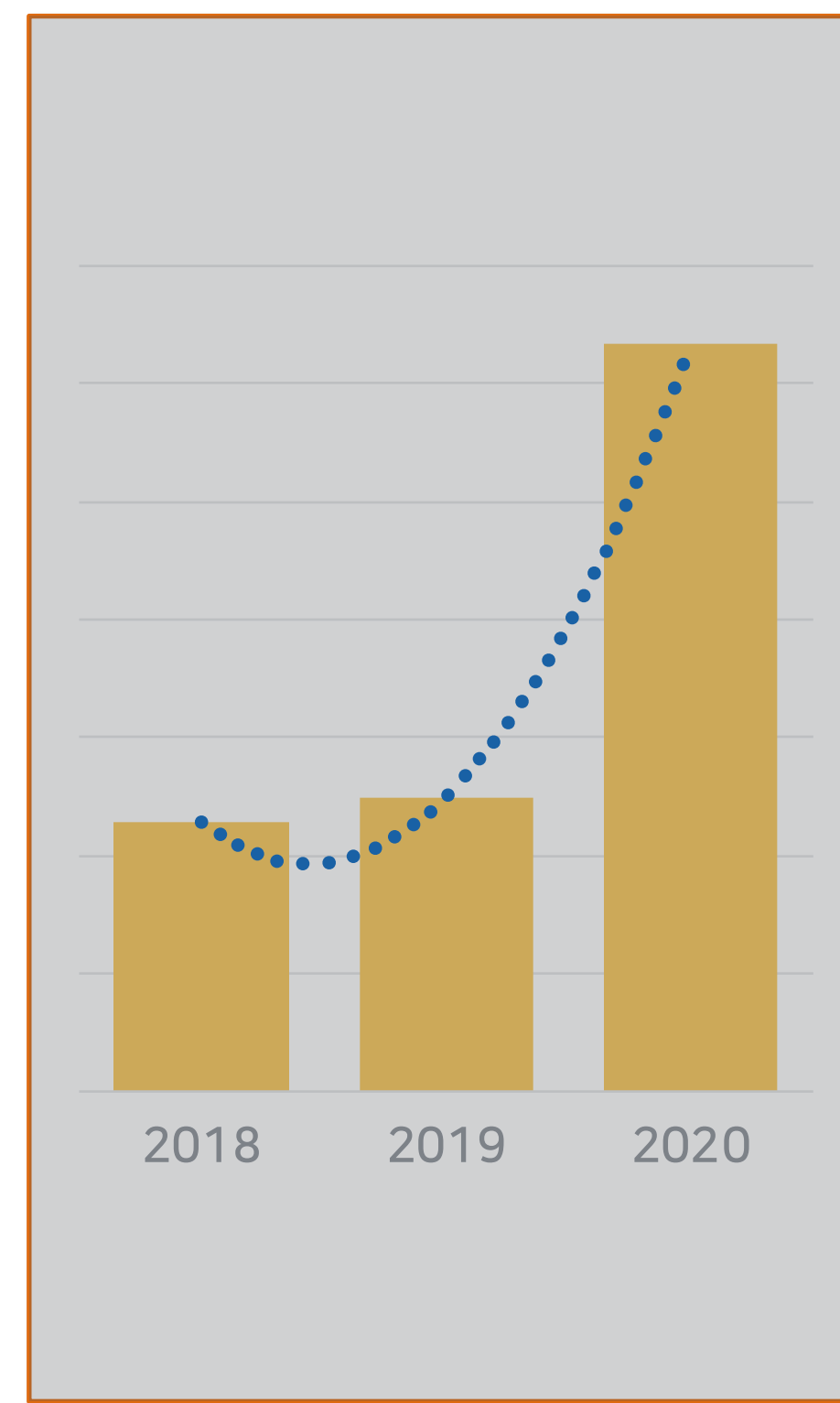
트래픽 이상 경보



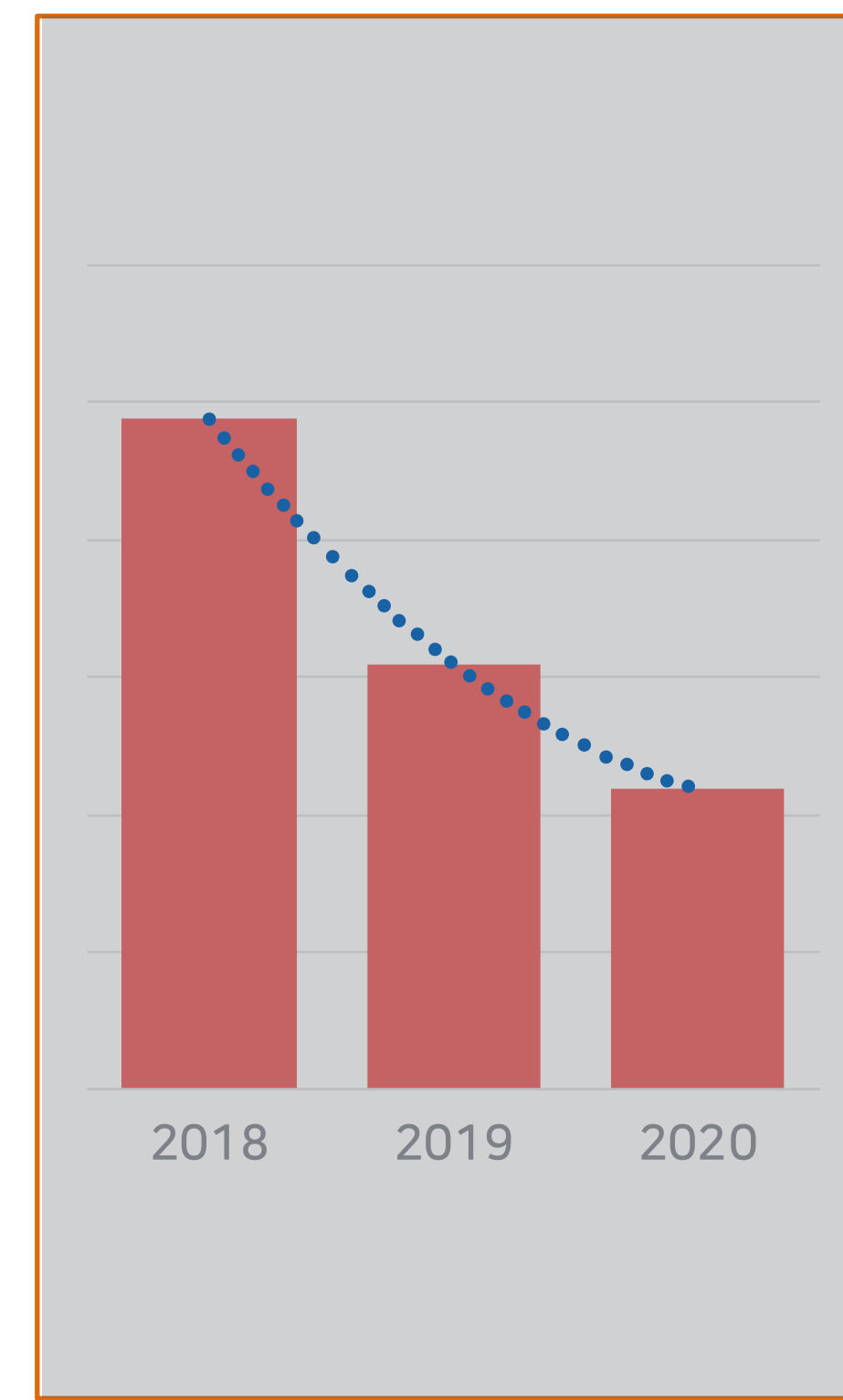
가용량 임계 경보



Timeout 경보



평균응답시간 경보



3.4 MTTM 개발 과정

트래픽 이상 경보

외부요인

가용량 임계 경보

Timeout 경보

평균응답시간 경보

내부요인

3.4 MTTM 개발 과정

트래픽 이상 경고

외부요인

- 올림픽
- 지진상황
- ...

가용량 임계 경고

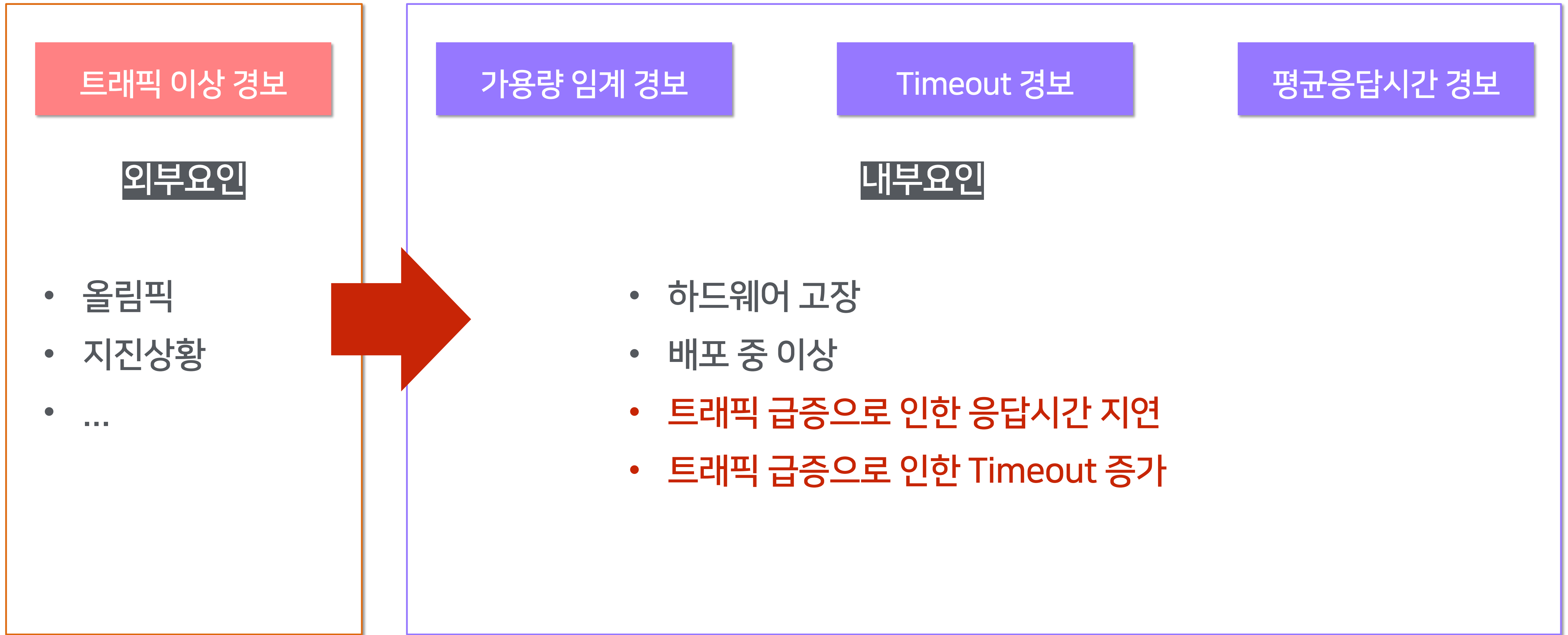
Timeout 경고

평균응답시간 경고

내부요인

- 하드웨어 고장
- 배포 중 이상

3.4 MTTM 개발 과정



3.4 MTTM 개발 과정

트래픽 이상 경고

외부요인

불가항력

장애 위험

가용량 임계 경고

Timeout 경고

평균응답시간 경고

내부요인

Action 가능

개선 가능

3.4 MTTM 개발 과정

트래픽 이상 경고

외부요인

트래픽 이상 비율

가용량 임계 경고

Timeout 경고

평균응답시간 경고

내부요인

MTTM

3.4 MTTM 개발 과정

연간 트래픽 이상 비율

1년 중 트래픽 이상이 발생한 총 시간의 합

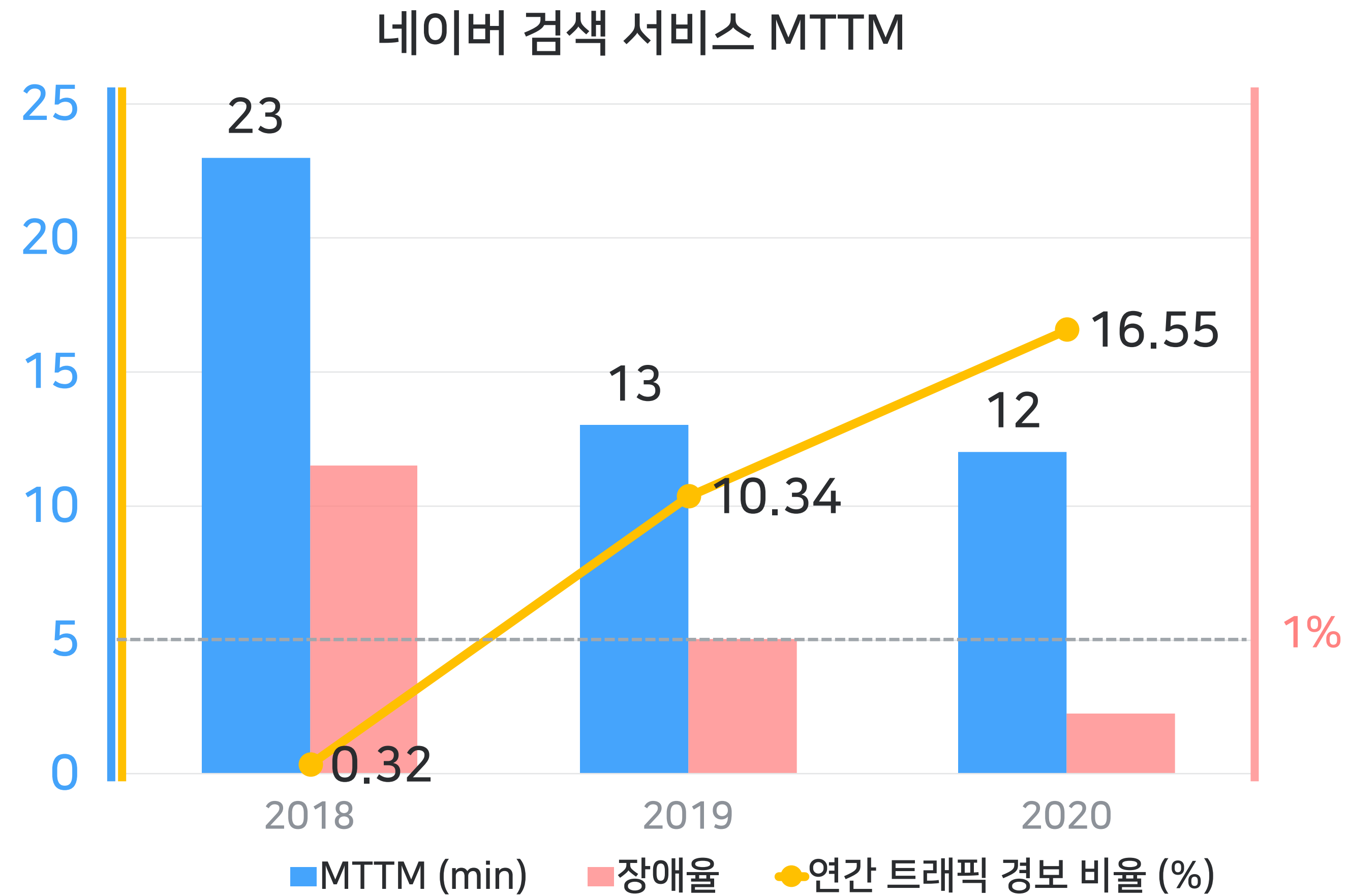
1년

3.4 MTTM 개발 과정

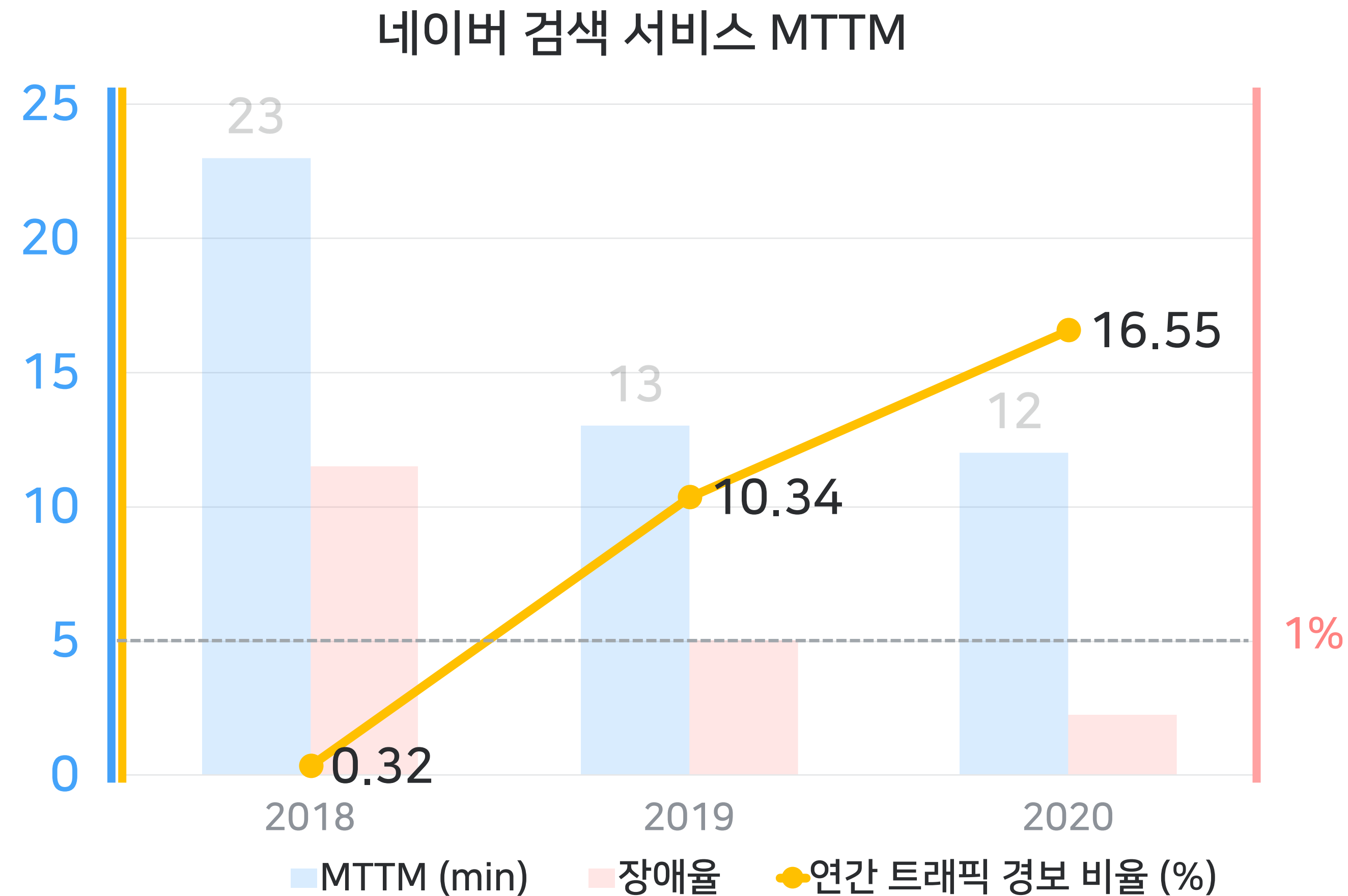
- TTM (Time To Mitigation): 이상 증상이 발생해서 해결되기까지의 시간

- $$\text{MTTM (Mean Time To Mitigation)} = \frac{\text{TTM 시간 총 합}}{\text{TTM 개수}}$$

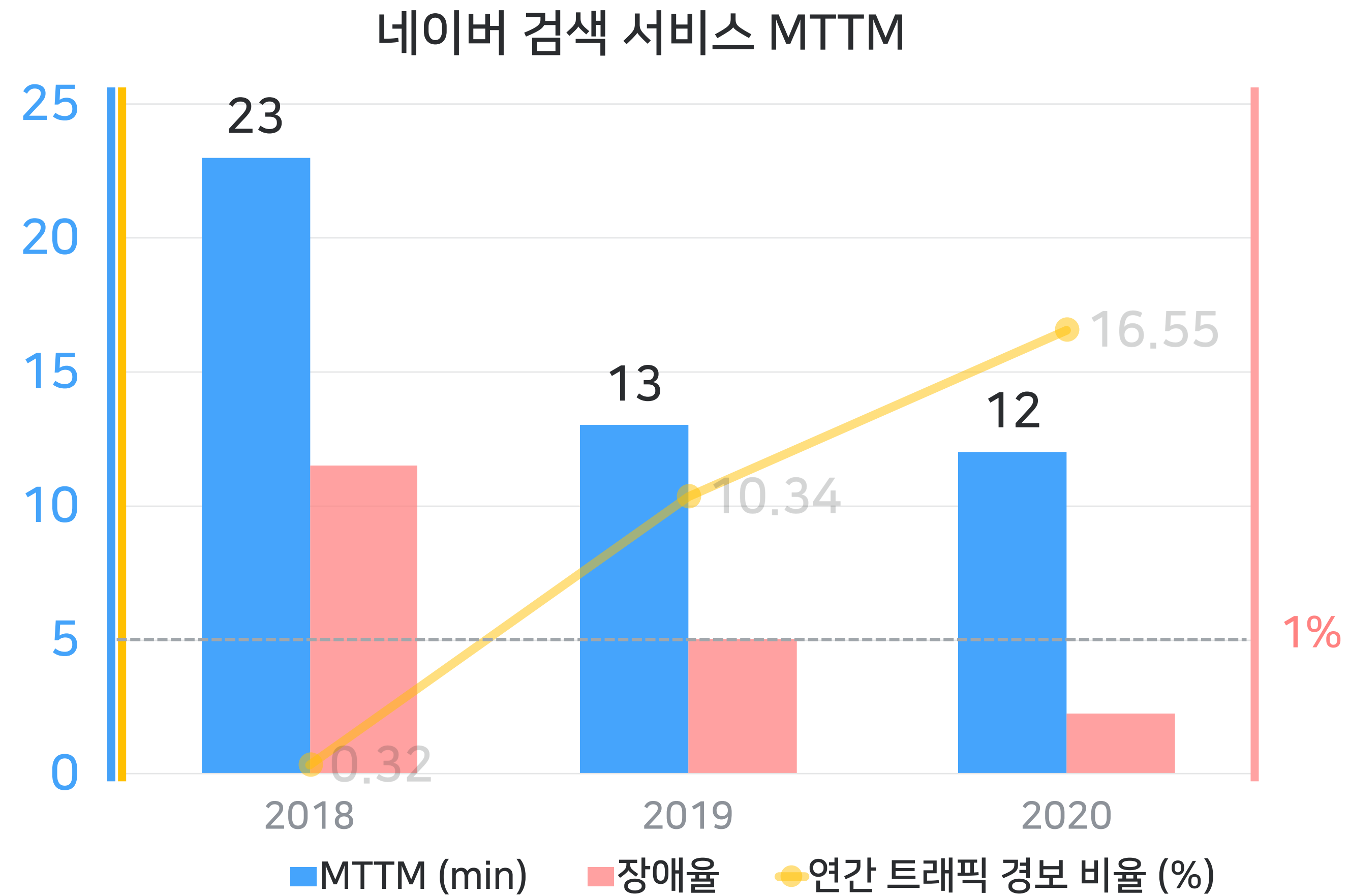
3.4 MTTM 개발 과정



3.4 MTTM 개발 과정



3.4 MTTM 개발 과정



결론

4. 결론

MTTM (Mean Time To Mitigation)

- 장애가 없어도 적용 가능한 KPI
- 시스템 건강도 측정 가능
- 장애 관제 개선도 파악에 도움
- 장애 감지 해상도를 높임

4. 결론

MTTM의 감소

: 우리 서비스의 이상증상이 빨리 완화된다

- 우리가 가용량 관리를 적절하게 잘 하고 있음
- 우리 시스템의 건강도가 높게 유지되고 있음

4. 결론

사용자
관점

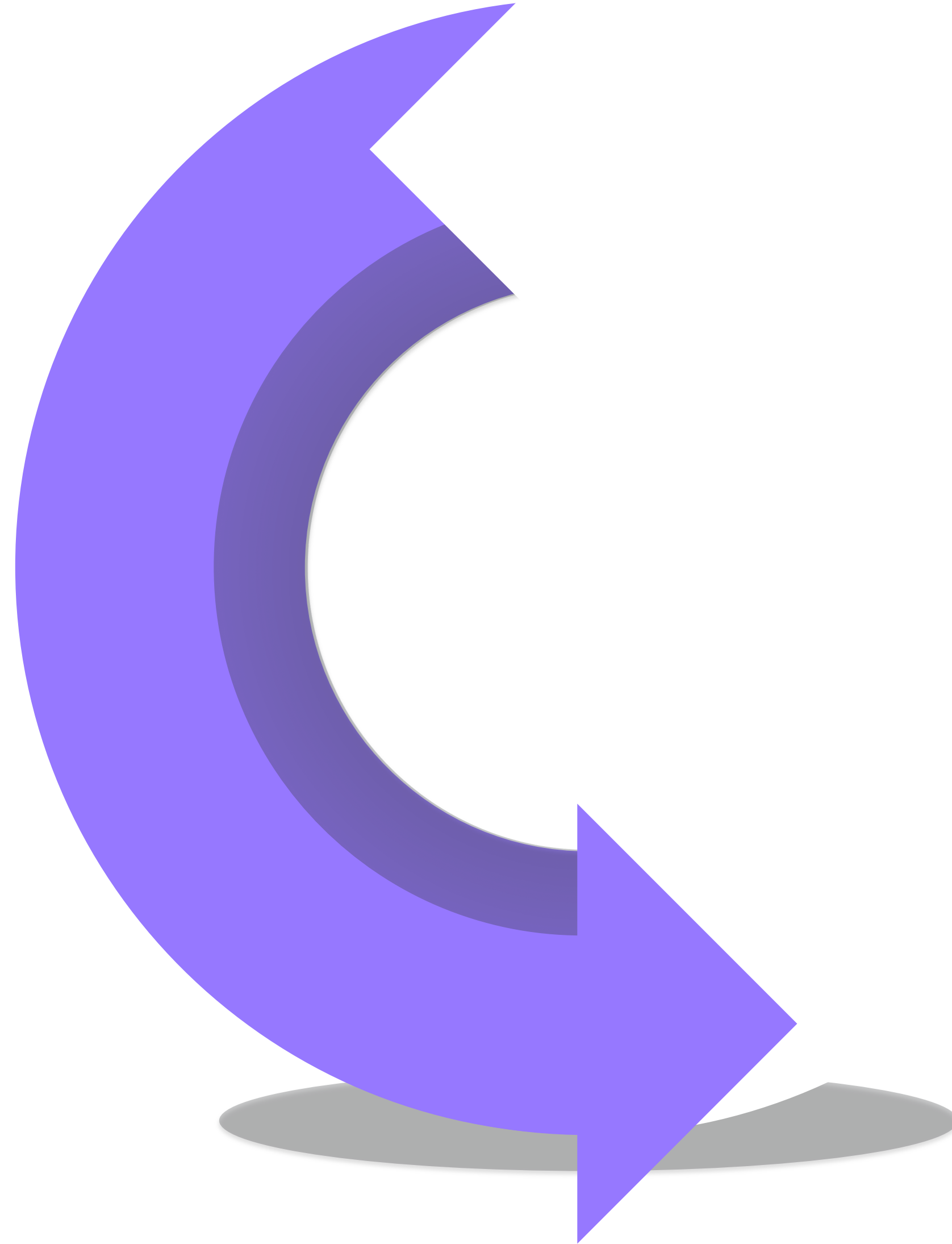
측정
자동화

Data
확보

서비스별
감지 조건
차이

4. 결론

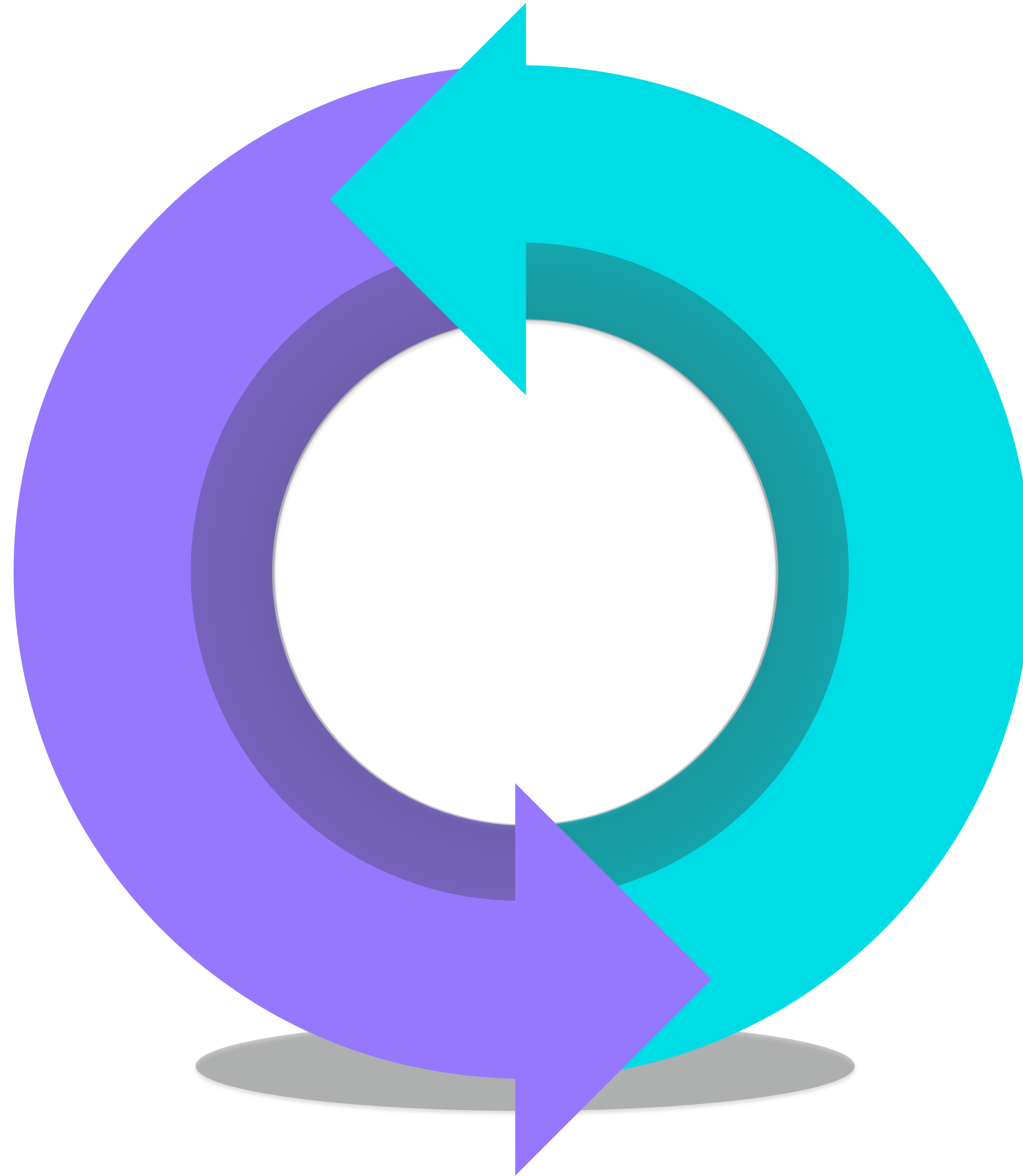
KPI



4. 결론

KPI

측정/개선



4. 결론

KPI

KPI saturation을 넘어서서
안정성과 신뢰성 향상
완전 자동화 목표

측정/개선



플러스!

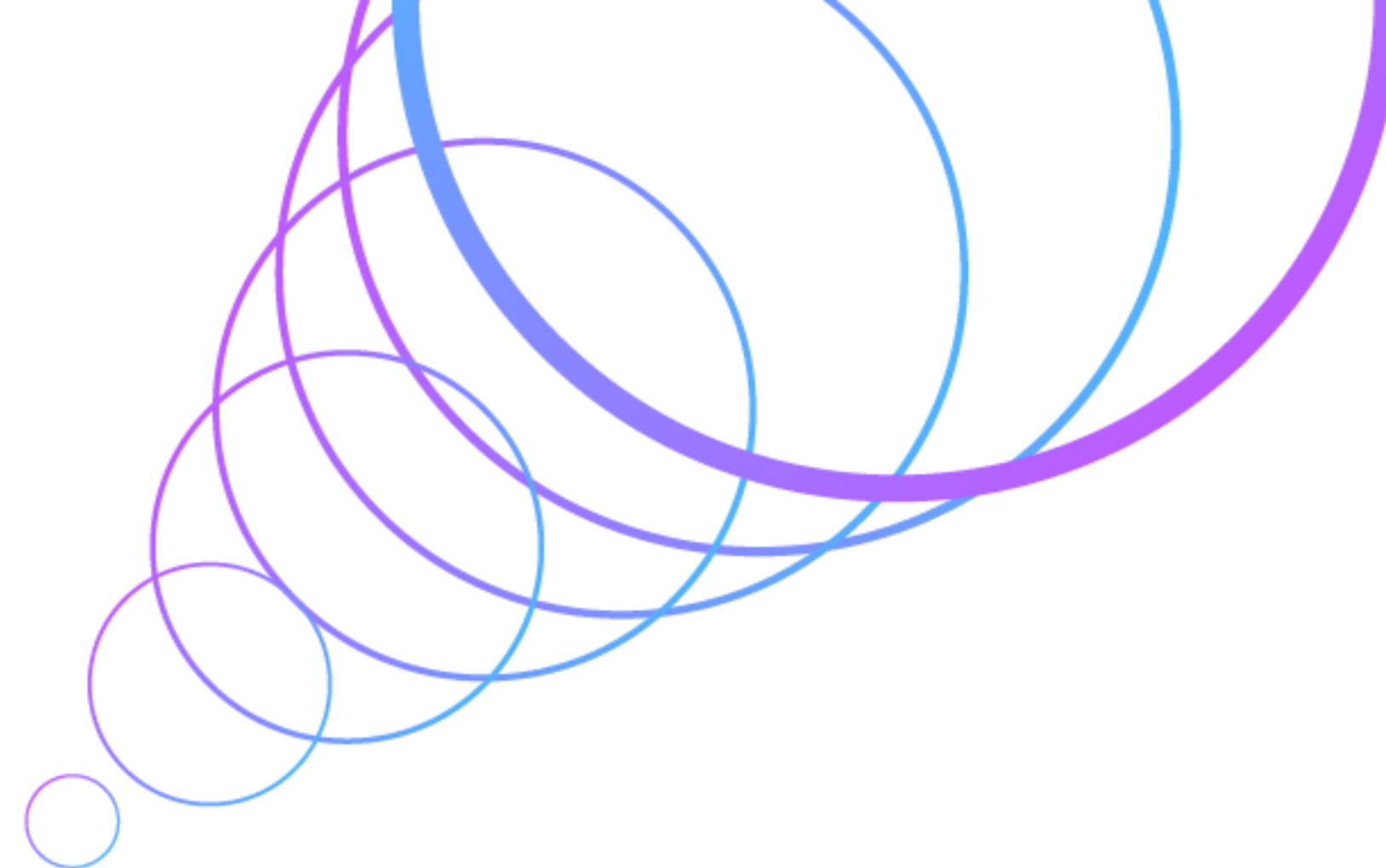
채용합니다. 😊

QR코드 찍어주세요

<https://naver-career.gitbook.io/kr/service/search/reliability-platform/seer>



Q & A





Thank You