실험 노출 증가시키기:
속도, 품질 및 위험의 트레이드오프

(Ramping Experiment Exposure: Trading Off Speed, Quality, and Risk)

가짜연구소 인과추론팀
온라인 통제 실험 연구자로 거듭나기

2024-06-11

이재호

What Is Ramping?

Causal-Lab

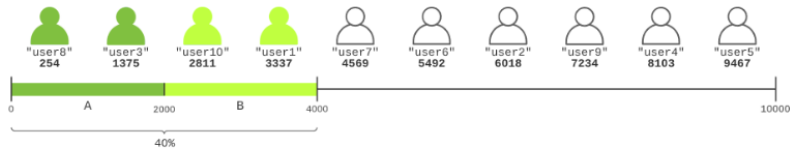
Ramp Process

- 통제된 노출 (aka. controlled exposure)
- 새로운 기능 출시(변수)와 관련해서 실험에서 새로운 기능들에 대해 트래픽을 점차 증가시키는 과정 (불확실한 위험을 통제)
- 실험 규모가 확장됨에 따라 제품 안정성을 저하시킬 수 있기 때문에 원칙이 필요하며 크게 3가지 사항을 고려할 수 있음.
  - 속도 (Speed, 시간과 자원의 낭비 vs 사용자에게 악영향 및 최적 의사결정의 어려움)
  - 품질 (Quality)
  - 위험 (Risk)
- 이 챕터는 전반적으로 4장에서 언급하는 실험플랫폼이 전제된 챕터(democratize experimentation with a fully self-served platform)
  - 실험 수행자들을 위해 Ramp Process에 대한 원칙이 필요
  - (이상적으로는) 대규모 실험에도 도구를 통해 프로세스 자동화 및 원칙 적용이 가능

Ramp Up

- 소수의 사용자에게만 Feature 노출 → 지표 및 시스템 확장성 안정적 → 실험에서 목표한 수준까지 노출 대상 확대

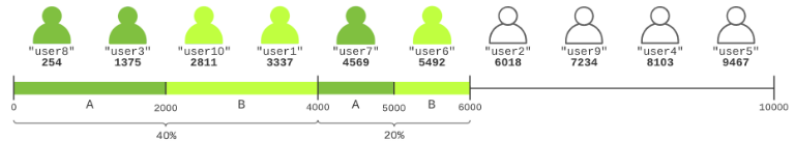
테스트 그룹 분배



트래픽 할당 40% / 그룹분배 A(50) : B(50)

40%를 할당한 A/B테스트가 A, B 두 개의 그룹을 가지고 있고 그룹 분배 비율이 50:50인 경우는 할당된 4000개의 슬롯을 그룹 분배 비율에 맞게 분배합니다.

0~1999 슬롯에 맵핑된 사용자는 A 그룹으로 할당되고, 2000~3999 슬롯에 맵핑된 사용자는 B 그룹으로 할당됩니다.



트래픽 할당 60% / 그룹분배 A(50) : B(50)

트래픽을 60%로 증가시키면 추가로 20%에 해당하는 슬롯을 더 할당해야 합니다. 남은 슬롯 2000개(20%)를 그룹 분배 비율에 맞게 분배해서 할당합니다.

핵클(Hackle) - 테스트 그룹 분배 원리

What Is Ramping?

Causal-Lab

온라인 종합 대조 실험을 수행하는 이유

- Treatment가100% 출시 되었을 때 (New Feature Launched) impact와 Return-On-Investment (ROI) 측정
- 사용자 및 비즈니스에 대한 피해와 비용을 최소화하여 위험을 줄이기 위해. (즉, 부정적인 영향이 있을 때 위험을 줄이기 위해.)
- 사용자의 반응을 (이상적으로는) 세그먼트별로 파악하여 잠재적인 버그를 파악하고 향후 계획에 반영 (5장)

    - 특정 기능의 성능이 얼마나 중요한지?
    - 성능 향상이 사용자의 참여와 전환에 얼마나 영향을 주는지?
    - 성능 개선으로 인한 장기적 영향이 있는지?

Ramping이 필요한 이유

- 변경에 의한 영향을 통제하면서 잠재적인 위험을 완화하기 위해서는 적은 노출로 시작 해야함. (pre-MPR)
- 선택적으로 노출의 확대 과정에서 50%(MPR) ~ 100% 사이의 중간 램핑 단계가 필요할 수 있음. (post-MPR)

    - 운영상 이유로 Traffic을 감당할 수 없을 때
    - 학습이 필요한 경우, 실험을 장기적으로 측정 해야 하는 경우
        - 특정 기간 동안 새로운 Feature(Treatment)에 노출되지 않게 함 (Hold Out Ramp)

SQR Ramping Framework

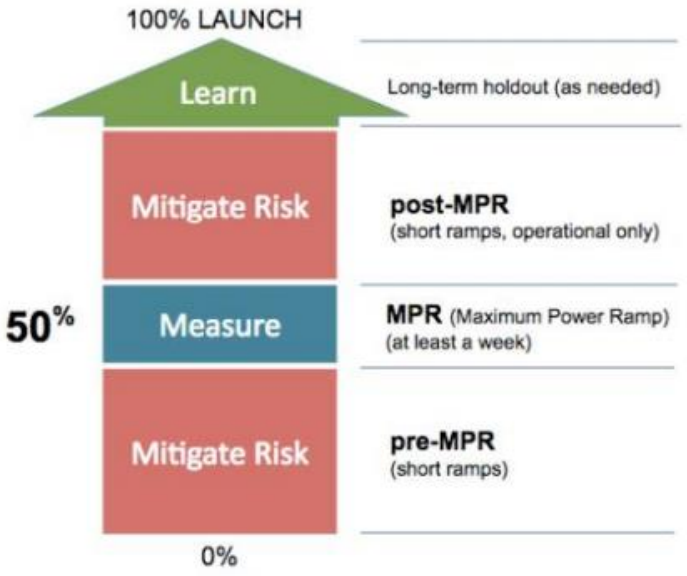
Causal-Lab

SQR Ramping Framework :

SQR: Balancing Speed, Quality and Risk in Online Experiments: https://arxiv.org/pdf/1801.08532

Background

- 2015년 LinkedIn에서 실험을 위해 300개 이상의 독특한 램프 시퀀스가 존재했으며, 실험당 평균 4개의 램프가 있었음.
- 그림 1은 각 실험들이 각 Ramp마다 소요된 일자들의 분포를 의미함.
- 램프마다 비슷한 시간이 걸리는 경향이 있으며 (평균 6일), 비효율을 해결할 원칙이 필요하다고 판단
- 추가적으로 램프 프로세스에 자동 알고리즘을 추가하여 실험기간을 절반으로 줄일 수 있다면, 위험을 통제하고, 의사결정 품질을 올리면서, 새로운 시도를 두배 더 할 수 있는 것으로 확인.



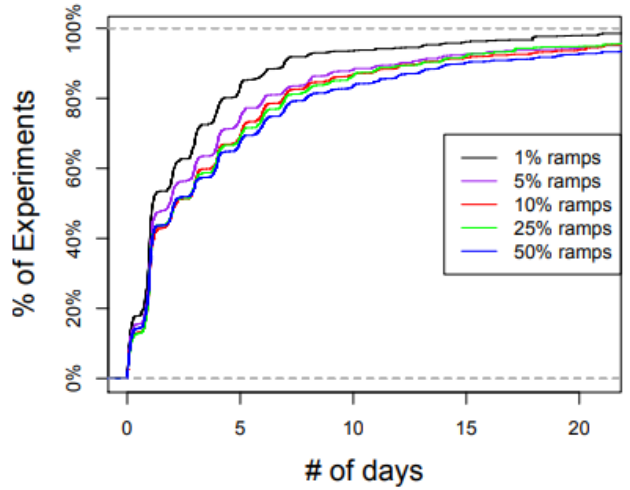


Figure 1: Cumulative distribution of ramp duration, by ramp%.

SQR Ramping Framework

Causal-Lab

Three Mistakes Most People Make

- 1. Let's keep it running to get statistical significance.
  - LinkedIn에서는 회사 전체가 통계적 중요성"에 관심을 갖도록 하는 데 매우 성공적 (유의미한 결과에 의미부여)
  - 실험을 충분히 오래 실행하면 통계적으로 유의미한 결과를 얻을 것이라고 생각함.
    - 장기간에 걸친 여러 테스트로 인해 false positive rate이 높기 때문에 사실일 수 있음
    - 실험을 오래 진행할 수록 샘플크기가 커지며, 효과크기가 동일하다면 유의미한 결과를 얻을 가능성이 더 커짐
  - 하지만, 실험기간과, 노출 사이의 균형은 고려되지 않았음
- 2. The cost is lower if I keep the experiment at a smaller ramp (even for a longer time).
  - 통계적 검정력은 거의 얻지 못하지만 오랫동안 실험을 유지하는 데는 비용이 많이 들게 됨
    - 기회 비용
    - 플랫폼 비용
    - 사업 비용
- 3. We have enough users at the 10% ramp. Let's ramp to 100%
  - 실제 실험 중 60%는 10% 미만의 활성 사용자를 대상으로 실행되었으며, 실제 표본 크기는 더 작게 됨.
  - 많은 지표, 특히 수익관련 지표의 변동성이 매우 크며, 정규성 가정이 타당해지려면, 높은 볼륨이 필요함.
  - 낮은 램프에서 결론을 도출하기에 분산이 너무 높기에 더 나은 해결 방법이 필요
    (구글, 마이크로소프트에서도 트래픽 규모에 대해 유사한 논의를 하였음.)

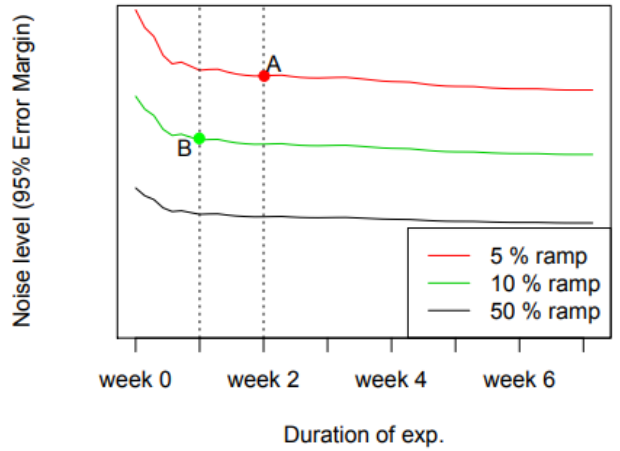


Figure 2: Tradeoff between ramping up (B) vs. running longer (A).

SQR Ramping Framework

Causal-Lab

SQR Ramping Framework

- Ramp Process에서 SQR(Speed, Quality, Risk) 사이에서 효과적으로 균형을 맞출 수 있는 램핑 프레임워크를 구축.
- 실험자들이 저지르는 가장 흔한 실수를 식별하는 것부터 시작한 다음, 실험의 4가지 램프 단계에 해당하는 4가지 SQR 원리를 소개.
- SQR을 모든 실험으로 확장하기 위해 우리는 모든 실험을 실행하는 프로세스에 내장되어 자동으로 램프 결정을 권장하는 통계 알고리즘을 개발

MPR (Maximum Power Ramp, 최대 검정력 램프)

- 균등한 할당을 의미하며 균등한 할당이 되었을 때, 가장 작은 MDE를 감지할 수 있음. (50:50)
- (논문) 분산이 가장 작으므로 측정된 델타의 정밀도가 가장 좋음.
  실험에 하나의 처리만 사용하여 전체 100% 트래픽이 있는 경우 2-표본 t-검정의 분산은 $1/q(1-q)$에 비례함. (여기서 q는 처리 트래픽 백분율)
- Example) q=0.5 라면 분산은 1/0.5(1−0.5)=4, q=0.2 라면 분산은 1/0.2(1−0.2)=25 이므로 50%일 때 분산이 더 작음.
- 독립적인 두 집단의 평균 차이에 대한 분산 식

$$\mathrm{Var}(\bar{X}-\bar{Y}) = \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y} = \frac{\sigma^2}{nq} + \frac{\sigma^2}{n(1-q)} = \frac{\sigma^2}{n}\left(\frac{1}{q} + \frac{1}{1-q}\right) = \frac{\sigma^2}{n}\frac{1}{q(1-q)}$$

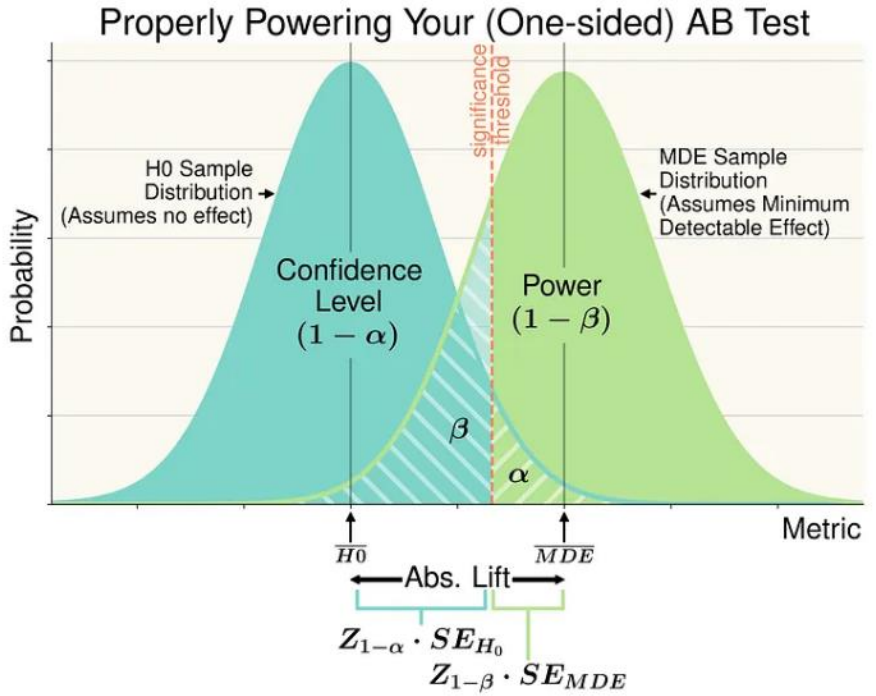
- AB Test에서 실험군과 대조군이 독립적이라는 가정과 이점
  - 무작위 할당 (Random Assignment): 각 그룹의 표본은 모집단의 특성을 대표할 확률이 높아지며 독립성을 보장함
  - 그룹간 차이가 실제 처리 효과임을 보기위한 가정이며, 독립성을 가정하면 교란변수의 영향을 최소화 할 수 있음

SQR Ramping Framework

Causal–Lab

MDE(Minimal Detectable Effect):

- A/B 테스트에서 실험이 감지할 수 있는 최소한의 효과 크기를 의미.
- MDE는 실험 설계 시 중요한 요소로, 특정 통계적 검정력과 유의 수준 하에서 검출할 수 있는 효과의 크기 = 두 분포가 구별될 수 있는 최소한의 차이



효과 크기 (Effect Size, Δ) =μ1−μ2 = 두집단의 평균차이

표본 크기 (Sample Size, n), 표본의 분산 (Variance, $\sigma 2$)

독립인 두 집단의 표준 오차

$$\mathrm{SE} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$ 두 샘플의 크기가 같다면 우측과 같이 단순화 $$\mathrm{SE} = \sqrt{\frac{2\sigma^2}{n}} = \frac{\sigma\sqrt{2}}{\sqrt{n}}$$

$$t = \frac{\Delta}{\mathrm{SE}} = \frac{\Delta}{\sigma\sqrt{2/n}} = \frac{\Delta\sqrt{n}}{\sigma\sqrt{2}} \qquad \mathrm{MDE} = (Z_{\alpha/2} + Z_{\beta}) \cdot \mathrm{SE}$$

즉, 앞서 확인한 트래픽 비율이 50%에 가까워 질 수록 두분산은 작아지게 되며

각 집단이 가질 수 있는 최대 트래픽을 가지게 됨, 분산이 작아지게 되므로, 검정 통계량은 커지게 되어 통계적 검정력이 높아지며

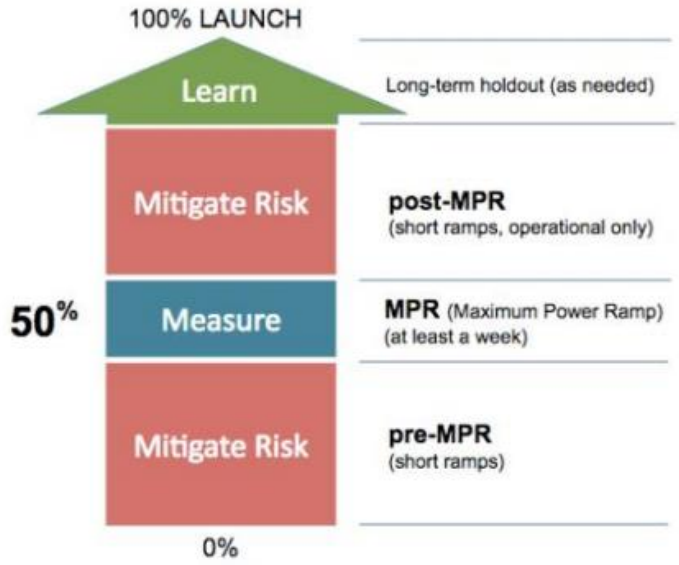
실험이 감지할 수 있는 최소한의 효과크기는 작아져서 작은 효과라도 검출할 수 있는 확률이 높아짐.

Calculating Sample Sizes for A/B Tests - Timothy Chan

SQR Ramping Framework

Causal–Lab

Phase1 (pre-MPR)



- 위험에 대한 탐색 및 완화 (속도를 올리고 위험을 상쇄)
- 실험(Test) 모집단의 링(Ring)을 만들고 Treatment를 서서히 적용해 나감
  - 실험 모집단 (Rings of testing population)
    - 해당 기능 개발팀 (화이트리스트)
    - 서비스 충성도가 높은 베타 테스터, 내부자
    - 특정 단일 데이터 센터

- 원칙1. 위험이 작다고 판단되는 즉시 MRP까지 빠르게 램프 한다.

- 모든 실험이 모든 사용자에게 영향을 미치는 것이 아님 (앞선 배경에서 3번째 문제)

- 가장 작은 Ring에서는 트래픽이 충분하지 않기 때문에 Statistical Power가 낮고 정성적인(qualitative) 피드백만 확인 가능
  - 조직원(화이트리스트)로부터 직접 피드백

- 확대하여 트래픽이 조금 더 커진다면, 정량적(quantitative)인 측정도 가능하겠지만, 여전히 Power는 낮은 상태 (통계적 유의성을 기다릴 필요는 없음)

- 주요 가드레일 지표에 대한 실시간에 가까운 지표를 생성 해야함. (빨리 파악할 수로 다음 단계로 넘어갈 수 있음)

- LinkedIn에서는 1%, 5%, 10%, 25%, 50%를 비율로 사용한다고 함.
  - 내부적인 원칙이 필요할 것으로 보임

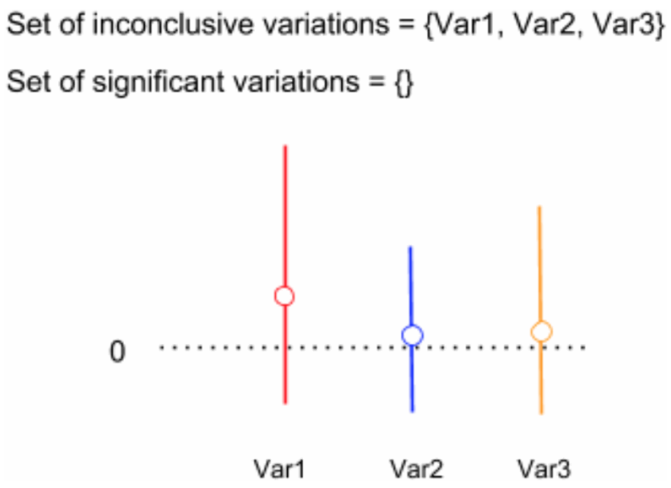
SQR Ramping Framework

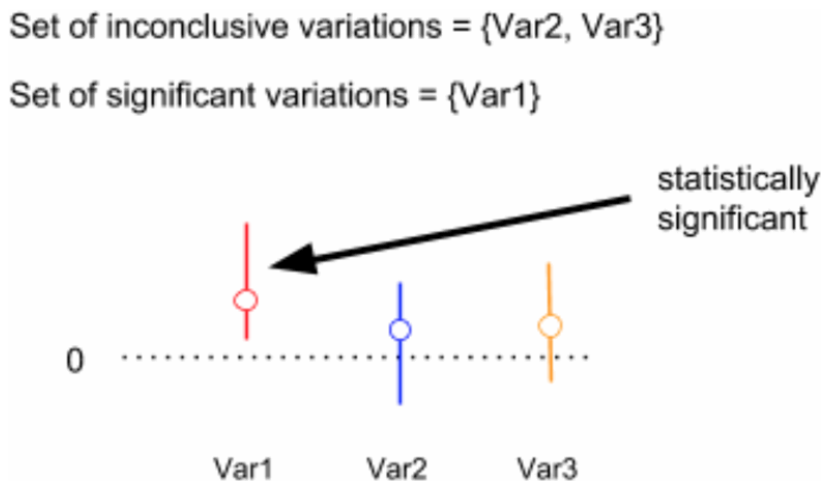
Causal–Lab

Optimizely — Controlled traffic allocation ramping (Link)

The maximum power stage looks for a winner between variations while balancing speed and precision.
After the preliminary short ramp stage, to start the maximum power stage you should:

- Reset results. This prevents Simpson's paradox and time-varying biases.
- Set the experiment into the maximum feasible traffic allocation, preferably a balanced, uniform split among the baseline and variations. See change traffic allocation for Web Experimentation and Performance Edge or updating flag rules in Feature Experimentation for more information.





SQR Ramping Framework

Causal-Lab

Phase2 - MPR

- 해당 책에서 언급한 신뢰할 수 있는 온라인 대조실험을 하기 위해 논의한 내용이 해당 과정에서 모두 적용
- 일주일 동안 유지를 권장 (신기/초두 효과가 있는 경우 더 길게)
  - 하루동안 실행되는 실험에서는 결과가 편향될 수 있음
    - 헤비유저
    - 주중/주말 방문 유저

- SQR: Balancing Speed, Quality and Risk in Online Experiments 논문에서는 MRP까지 자동으로 램핑을 하기 위한 원칙과 알고리즘에 대한 내용이 존재

- 실제로는 많은 실패와 반복이 있을 단계
  - 가드레일 지표, 실험 지표의 하락
  - 통계적으로 유의미하나 지표는 효과 없는 경우
  - 일부에만 효과가 있는 것으로 보이는 경우
    - A,B,C,D … 집단 존재
    - UI + 알고리즘을 같이 테스트하는 경우
  - 세그먼트 분석 (Break Down)
  - Feature 조정 / 수정
  - 실험 재설계 / 재할당

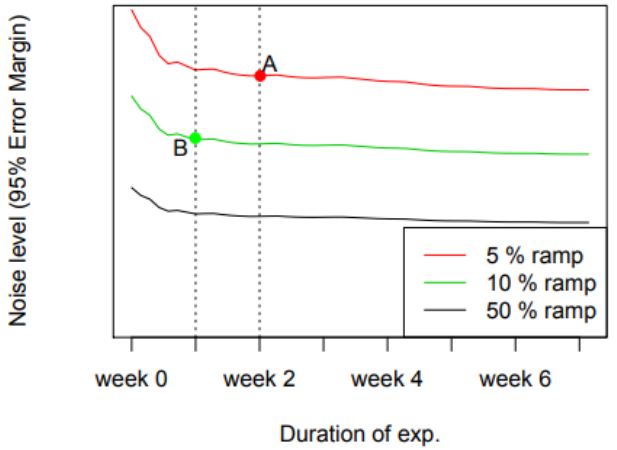
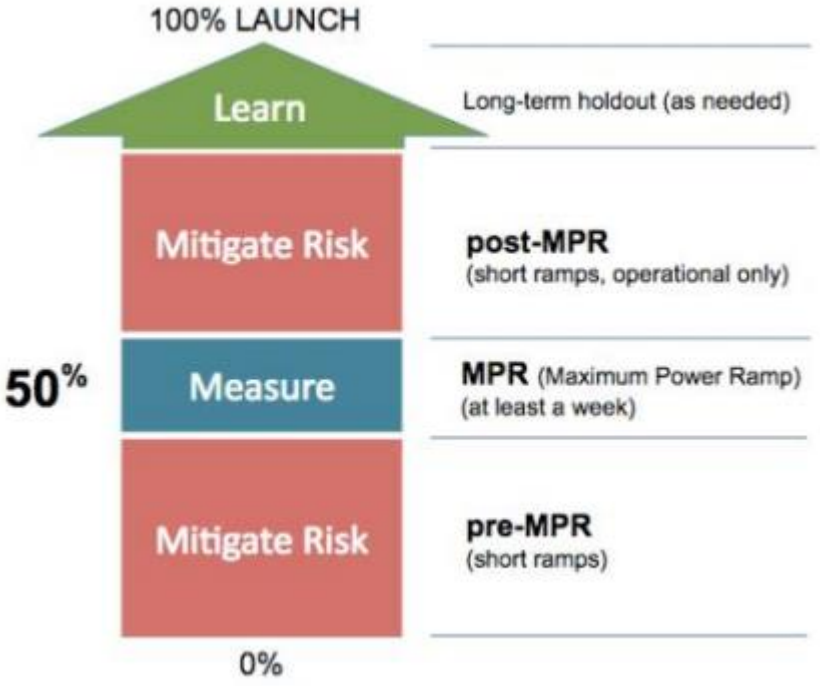


Figure 2: Tradeoff between ramping up (B) vs. running longer (A).

SQR Ramping Framework



Phase3 – post-MPR (선택적)

- LAUNCH 전 증가된 트래픽 부하에 대한 문제 같이, 운영성/확장성의 문제로 추가적인 조치가 필요한 경우, 점진적으로 램프를 증가시키며 1일 이내로 처리할 것을 권장

Phase4

- Long-Term Holdouts: 특정 사용자가 오랫동안 신규 기능에 노출되지 않게 하는 것
  - 램핑 과정의 기본 단계로 상정하지 않는 것이 좋음
  - 비용 및 고객에게 더 우수한 경험을 의도적으로 지연시키는 것 (윤리적 문제)
  - 실험군을 90%, 대조군을 10%로 놓은 채로 실험을 장기적으로 가져 감
- 그럼에도 실제로 유용할 수 있는 경우
  - 장기 실험효과가 단기와 다를 수 있는 경우
    - 신기/초두효과가 예상되는 경우
    - 주요 지표의 단기적 효과가 너무 커서 재무 예측과 같이 의도적으로 효과를 지속시켜서 측정해야 하는 경우
    - 단기적 효과가 미미하지만, 시간에 따라 효과가 나타날 것으로 예상되는 경우
    - 성공 지표자체가 장기지표인 경우
    - 분산 감소의 이점이 있는 경우
- 그 외에도 반복, 역실험을 수행할 수도 있다.

Discussion

End of Document