Chapter 10: Product Sense

왜 프로덕트 센스인가?

데이터 사이언티스트는 PM, 비즈니스 이해관계자와 **교차 기능적으로**(cross-functionally) 협업함

프로덕트 로드맵 수립과 비즈니스 문제의 근본 원인 파악에 기여해야 함

프로덕트 인터뷰 — 4가지 핵심 유형

유형	핵심 질문
메트릭 정의	신규 제품/기능의 성공을 어떤 지표로 측정할 것인가?
메트릭 변화 진단	지표가 올랐거나 내렸다 — 근본 원인은?
제품 기능 브레인스토밍	새 제품을 출시해야 하는가? 어떤 기능을 추가할 것인가?
A/B 테스트 설계	새 기능의 효과를 어떻게 실험으로 검증할 것인가?

프로덕트 인터뷰 — 6가지 행동 원칙

1. **명확화 질문**: 사용자 흐름, 최종 사용자, 이해관계자, 비즈니스 목표를 먼저 파악
2. **문제 범위 설정**: 큰 문제를 면접 시간 내 해결 가능한 크기로 축소
3. **소리 내어 생각하기**: 사고 과정 자체가 평가 대상 — 머릿속에 담아두지 말 것
4. **대화형 진행**: 면접관에게 "이 방향이 맞는지" 중간중간 확인
5. **목표를 잊지 말 것**: 기술적 디테일에 빠져 회사의 미션/비전을 놓치면 안 됨
6. **외부 경험 적절히 활용**: 과거 도메인 경험을 유연하게 언급하되, 오만하지 않게

> 핵심 마인드셋: **"이미 입사한 직원처럼"** — 면접이 아니라 동료와의 업무 미팅이라고 상상하면 모든 팁이 자연스러워짐

프로덕트 센스 개발 — 일상 속 습관

제품을 만났을 때 물어볼 5가지

- 이 제품은 **누구를 위해** 만들어졌는가?
- 해결하려는 **핵심 문제**는 무엇인가?
- 사용자에게 주는 **궁극적 혜택**은?
- 디자인과 카피가 **목적을 얼마나 잘 전달**하는가?
- 회사의 **미션/비전과 어떻게 연결**되는가?

Snapchat vs Instagram 사례

- Snapchat: 카메라 바로 실행 → 즉시 표현, 메시지 자동 삭제 → 부담 없는 공유
- Instagram: 피드 중심 → 소비 유도, 영구 게시물 → 완벽함에 대한 압박
- **같은 "사진 앱"이지만 핵심 가치가 완전히 다름**

비즈니스 센스 & 리뷰 보정

매일의 습관 — 비즈니스를 만났을 때

- **비즈니스 모델**: 어떻게 수익을 내는가? 어떤 레버를 당기면 수익이 늘어나는가?
- **핵심 지표**: 이 회사 DS라면 어떤 KPI를 추적할 것인가?
- **생태계**: 경쟁사는? 파트너는? 업계 내 위치는?

리뷰로 직관 보정

- **긍정 리뷰** → 내가 놓친 가치, 사용자가 인지하는 혜택 발견
- **부정 리뷰** → 미충족 니즈, 브레인스토밍 아이디어 원천
- 소스: Reddit, App Store, G2 Crowd, Amazon
- 돈의 흐름을 이해해야 **고레버리지 프로젝트** 선택 가능

도메인 경험 해킹 — 면접 전 준비법

준비 항목	방법
사업 모델 이해	"company name business model" 검색, 공개 실적 보고서
핵심 메트릭 파악	동종 상장사 earnings report에서 추적 KPI 확인
기술 블로그 읽기	엔지니어링 블로그 → 내부 시스템과 비즈니스 로직 파악
직접 제품 사용	사용자뿐 아니라 판매자/드라이버 등 **모든 사이드의 UI** 탐색

> Uber Eats 면접이라면: 주문자 앱 + Uber Driver 앱 + 레스토랑 가입 페이지까지 확인
> 면접 끝 질문에서도 빛남: "블로그에서 배달 시간 알고리즘을 봤는데..." → **상위 1% 후보자**

프로덕트 메트릭의 기초

AARRR 퍼널, 좋은 메트릭, North Star

AARRR — 해적 메트릭 프레임워크

단계	핵심 질문	대표 지표 (Pinterest 사례)
Acquisition	사용자가 어떻게 찾아오는가?	신규 다운로드 수, 가입 전환율, CAC
Activation	첫 "아하 모먼트"를 경험했는가?	프로필 작성 완료율, 첫 핀(pin) 저장
Engagement	얼마나 자주, 깊게 사용하는가?	DAU/WAU/MAU, 세션당 핀 저장 수
Retention	계속 돌아오는가?	28일 리텐션, 월간 이탈률(churn)
Referral	다른 사람에게 알려주는가?	K-factor (추천 수 × 전환율), K>1 = 지수 성장
Revenue	어떻게 돈을 버는가?	LTV, LTV/CAC 비율

AARRR 퍼널 — 단계별 핵심 개념

Acquisition	Activation	Retention	Revenue
사용자가 제품을 "발견"하는 단계. 신규 가입, 전환율, CAC 추적. "Top of funnel"이라 불림.	사용자가 제품의 **핵심 가치를 처음 경험**하는 순간. DoorDash: 첫 주문, Instagram: 첫 게시물 또는 피드 10개 조회.	사용자가 **장기간** 계속 돌아오는지 추적. 신규 고객 획득 > 기존 유지 → 비용 더 큼. 월간 리텐션, 월간 churn으로 측정.	LTV(고객 생애 가치) = churn 전까지 가져오는 총 수익. **LTV/CAC 비율이 높아야** 지속 가능한 비즈니스.

Engagement & 면접 전 메트릭 숙제

User Engagement

- AARRR에 명시적으로 없지만 **Activation~Retention 사이**에 위치
- 일정 기간 내 핵심 행동을 수행한 **고유 사용자 수**로 측정
- Facebook: DAU/WAU/MAU, Uber: 라이드 수, YouTube: 시청 시간
- **빈도(frequency) 반영**이 단순 사용 여부보다 유의미함

비즈니스 모델별 핵심 용어

- **SaaS**: ACV, MRR, Churn Rate
- **마켓플레이스**: GMV, Take Rate, Liquidity
- **광고 기반**: ARPU, CPM, CTR, DAU/MAU 비율
- 면접 전 해당 업종 메트릭 용어를 반드시 학습

나쁜 메트릭의 5가지 유형 — Facebook Dating 사례

유형	문제점	예시
Vanity	보기엔 좋지만 의미 없음	프로필 조회 수 — 앱 작동 여부를 판단할 수 없음
Irrelevant	비즈니스 목표와 무관	앱 내 체류 시간 — 데이팅 앱의 핵심 가치와 거리가 멂
Impractical	측정 자체가 불가능	3번째 데이트 횟수 — 추적 방법이 없음
Complicated	변화 시 원인 파악 어려움	복합 지수 — 하위 구성요소를 분해해야 해석 가능
Delayed	수집까지 시간이 너무 오래 걸림	결혼 건수 — 매칭 후 수년 소요

> 핵심: **선행 지표(leading indicator)**를 추적해야 빠르 의사결정이 가능

좋은 메트릭의 4가지 조건

조건	설명
Meaningful	비즈니스 목표와 직결, 쉽게 조작 불가, 의사결정을 이끌어냄
Measurable	일관되고 신뢰성 있게 추적 가능
Understandable	이해관계자가 이름만 보고 무엇을 측정하는지 직관적으로 파악
Timely	합리적인 시간 내에 수집 가능

> 나쁜 메트릭의 결함을 **뒤집으면** 좋은 메트릭의 조건이 됨

North Star 메트릭과 Guardrail(카운터) 메트릭

North Star 메트릭

- 제품/팀의 **가장 중요한 단일 지표**
- 비즈니스 목표와 직접 연결
- 면접에서 제일 먼저 제안해야 할 지표
- 예: Facebook Groups → "유해 콘텐츠 노출률 감소"

Guardrail(카운터) 메트릭

- North Star를 최적화할 때 **훼손되면 안 되는 지표**
- 99% 게시물 삭제 → 유해 콘텐츠 0% 달성! 하지만 피드가 비어버림
- 예: 게시물 수, 조회 수, 좋아요/댓글 수
- **면접에서 자발적으로 언급하면 높은 점수** — 제품을 총체적으로 사고하는 증거

메트릭 정의 — 3단계 프레임워크

Step 1: 명확화

제품과 목적을 정의.
사용자 흐름, 타겟 유저, 비즈니스 목적 파악.
"왜 이 제품을 만들었는가?"

Step 2: 목표 설정

제품/비즈니스 목표를 명시.
회사 미션과의 연결고리 제시.
"이 제품이 비즈니스에 어떻게 기여하는가?"

Step 3: 메트릭 정의

AARRR 퍼널을 따라 각 단계별 지표 제안.
목표와의 정렬을 재확인.
카운터 메트릭도 함께 언급.

핵심 원칙

AARRR 퍼널이 분석의 가이드 프레임워크 역할.
목표 ↔️ 메트릭 연결을 **매번 명시적으로 재확인**.

Step 1: 제품과 목적 명확화 — Facebook Dating 사례

- **왜?** Facebook이 데이팅 제품에 관심을 갖는 이유는? 앱 참여 증대? 광고 수익?
- **목적은?** 캐주얼 데이팅인가, 장기 관계 지향인가?
- **누구를 위한 것인가?** 전 연령? 특정 연령/성향?
- **어떻게 작동하는가?** 독립 앱인가 Facebook 내장인가? Tinder 스타일 스와이프?

> 가정 예시: Tinder와 유사하지만 더 **상세한 프로필**로 의도적 매칭 유도.
> 뉴질랜드에서 먼저 출시 — 영어권 시장 테스트.

Step 2: 제품 & 비즈니스 목표 설정

- 면접관이 목표를 알려주길 기대하지 말 것 — **직접 합성(synthesize)**해야 함
- Facebook Dating의 목표:
  - **앱 참여 증대**: 매칭 확인 → 알림 클릭 → 그룹/생일 알림도 함께 확인
  - **에코시스템 강화**: Facebook → Instagram → Messenger → WhatsApp 순환
  - **미션 연결**: "의미 있는 관계 형성" — 가족, 커뮤니티, 친구 → 로맨틱 관계도 자연스러운 확장
- **회사 미션과의 연결을 명시하면 보너스 점수**

Step 3: 성공 메트릭 정의 — AARRR 퍼널 적용

Acquisition

가입자 수, 프로필 작성 완료율.
Facebook의 프라이버시 이슈가 데이팅 신뢰에 영향 → 모니터링 필수.

Activation

온보딩 완료(프로필+사진 업로드), 프로필 10개 조회, 첫 매칭.
사용자 기쁨(delight)의 순간에 도달했는지 확인.

Engagement & Retention

매칭 수/유저, 총 매칭 수, 스와이프 수, 실제 만남(전화번호 교환).
28일 후 스와이프 비율, 월간 매칭 유지율.

Revenue

유료 멤버십, 업그레이드 구매, 광고 임프레션 수.
다만 Facebook은 **PMF(제품-시장 적합) 우선** → 초기엔 수익 지표보다 Engagement/Retention.

메트릭 변화 진단 — 4단계 프레임워크

Step 1: 범위 파악

메트릭 정의 확인 (비율이면 분자/분모 분리).
중요도, 시간 프레임, 변화 크기 파악.

Step 2: 가설 수립

가능한 원인을 4가지 버킷으로 분류.
제품 퍼널을 거슬러 올라가며 브레인스토밍.

Step 3: 검증

데이터 슬라이싱 — 인구통계, 디바이스, 기간.
상위 메트릭(upstream) 확인으로 범위 축소.

Step 4: 분류

각 요인을 근본 원인 / 기여 요인 / 상관 결과 / 무관으로 분류.

Step 1: 범위 파악 — 무엇을 명확히 해야 하는가

- **메트릭 정의**: 비율 지표라면 분자 변화인지 분모 변화인지 분리
- **중요도**: 팀에 실제로 중요한 메트릭인지? 수천 개 중 하나일 수 있음
- **시간 프레임**: 갑작스러운 변화 vs 장기 추세? 일/주/월 단위?
- **크기**: 상대적·절대적 변화량은? 전주 대비? 전년 동기 대비?

> Instagram **사례**: 게시물당 평균 댓글 수 감소
> → 게시물은 증가, 댓글도 증가, 하지만 **댓글 증가 속도 < 게시물 증가 속도**
> → 6개월간 전년 대비 10% 하락

Step 2: 가설 수립 — 4가지 원인 버킷

Accidental (우발적)

데이터 수집 버그, 로깅 오류, 계측(instrumentation) 문제.
지표 하락이 **실제 사용자 행동 변화가 아닐 수 있음**.

Natural (자연적)

계절성, 요일 효과, 공휴일, 날씨 변화.
"매년 이 시기에 동일한 패턴이 있는가?"

Internal (내부적)

새 기능 출시, 버그 수정, 의도적 제품 변경, 마케팅 캠페인.
가장 흔한 원인 — 제품 퍼널을 거슬러 올라가며 탐색.

External (외부적)

경쟁사의 신제품 출시, 팬데믹, 경기 침체 등 거시적 사건.
사용자 행동 전반에 영향을 줄 수 있음.

Step 2 심화: 퍼널 역행(Walk-Up) 브레인스토밍

Instagram 댓글 감소 사례를 **로컬 → 글로벌** 순으로 추적:

1. **댓글 UI 변경?** 댓글 작성기가 덜 눈에 띄게 변경, 또는 댓글 자동 삭제 강화
2. **피드백 UI 변경?** 좋아요/공유 강조 → 댓글 cannibalization
3. **피드 콘텐츠 변경?** 광고(댓글 비활성화), 릴스(소비 중심) 비율 증가, 랭킹 모델 변경
4. **피드 자체의 변화?** 스토리, 릴스, 탐색 탭이 피드의 참여를 잠식

> 20개 원인을 나열하는 것보다 **검증 방법을 설명하는 것이 핵심**

Step 3: 가설 검증 — 데이터 슬라이싱

- **인구통계 세분화**: 연령, 성별, 위치, 언어, 디바이스별 교차 분석
- **상위 메트릭(upstream) 확인**: 퍼널을 점점 넓혀가며 이상 지점 탐색

Instagram 댓글 사례에서 확인할 지표:

- 댓글 작성기 / 댓글 섹션의 **뷰 수** → 댓글 노출이 줄었는지
- 좋아요 대비 댓글 비율, 공유 대비 댓글 비율 → 댓글만의 문제인지 전체 참여 문제인지
- 피드 vs 스토리 vs 릴스의 **참여량 추이** → 피드 자체가 밀려나고 있는지

> 프로덕트 센스와 사전 리서치로 **가장 가능성 높은 가설부터 검증**

Step 4: 요인 분류 — 4가지 카테고리

분류	의미	Instagram 사례
Root Cause	메트릭 변화의 직접적 원인	Trust & Safety 팀이 추가한 "댓글 끄기" 안내 인터스티셜
Contributing Factor	근본 원인은 아니나 기여	릴스 비중 증가로 댓글 가능 게시물 비율 감소
Correlated Result	근본 원인의 증상일 뿐	신규 계정의 댓글 수 감소(결과이지 원인 아님)
Unrelated Factor	메트릭 변화와 무관	같은 시기의 스토리 기능 업데이트

Instagram 사례 — 근본 원인 발견 과정

1. 전체 피드 활동(좋아요, 공유) 정상 → **댓글에 국한된 문제**
2. 계정 연령별 슬라이싱 → **신규 계정의 댓글 감소가 더 심함**
3. 코호트 분석 → **6개월 전부터** 신규 유저 댓글이 기존 유저와 괴리
4. PM 확인 → Trust & Safety 팀이 **"댓글 비활성화 가능" 안내**를 온보딩에 추가
5. 댓글 꺼진 게시물 제외하고 재분석 → **댓글 수 변화 없음!**

> 구조적 접근(퍼널 역행 + 세분화 + 코호트)이 **직감보다 신뢰할 수 있는 진단**을 만듦

메트릭 트레이드오프 — 접근법

- 경제학자 Thomas Sowell: **"해결책은 없다, 트레이드오프만 있을 뿐"**
- 면접 예시: "LinkedIn에 광고를 더 보여주면 **매출 +1%, 체류시간 -3%** — 출시할 것인가?"

접근 프로세스:

1. 두 메트릭의 **정의를 정확히 파악** (메트릭 진단 Step 1과 동일)
2. 제품/비즈니스 목표 기준으로 **어느 쪽이 더 중요한지** 판단
3. 결론 도출 — 3가지 선택지 중 하나

트레이드오프 — 3가지 결론 선택지

선택지	설명
되돌리기(Revert)	트레이드오프가 수용 불가 → 기능 변경 철회
영향 최소화	새로운 제품 개입을 브레인스토밍해서 부정적 영향 축소
수용	트레이드오프가 정당화됨 → 그대로 출시

- 면접에서 "정답"은 없음 — **어떤 시나리오에서 어떤 선택을 할지 설명**하는 것이 핵심
- LinkedIn 사례: 매출 +1%가 다른 방법으로 쉬운가? 체류시간 -3% 회복은 어려운가?
- 추가 아이디어: 저품질 광고가 원인이면 → "이 광고 숨기기" 기능으로 **둘 다 개선 가능**

A/B 테스트 설계 — 4단계 프로세스

Step 1: 메트릭 선택

핵심 메트릭 + Guardrail 메트릭 결정.
사후에 유의미한 것을 골라내는 **체리피킹 금지**.

Step 2: 임계값 설정

유의수준 α = 0.05, 검정력 1 - β = 0.8.
MDE(최소 감지 효과)를 이해관계자와 협의.

Step 3: 표본 크기 & 기간

MDE + 검정력 + 분산으로 필요 샘플 산출.
최소 **2주** 이상 → 요일 효과 고려.

Step 4: 그룹 배정

무작위 배정(randomization)으로 교란변수 제거.
대형 회사는 A/B 테스트 인프라가 자동 처리.

A/B 테스트를 하지 말아야 할 때

인프라 부족 (Lack of Infrastructure)

소규모 회사에서 신뢰할 수 있는 실험 인프라가 없을 때.
잘못된 인프라 위의 실험은 잘못된 결론을 낳음.

트래픽 부족 (Lack of Traffic)

사용자 수가 적어 합리적 기간 내 통계적 결론 불가.
작은 MDE를 감지하려면 막대한 표본 필요.

영향 부족 (Lack of Impact)

테스트 성공해도 비즈니스 영향이 미미한 경우.
엔지니어링 자원 대비 기대 효과를 **사전에 추정**.

격리 불가 (Lack of Isolation)

로고 변경 같은 전사적 변경은 대조군 설정 불가.
대안: 포커스 그룹, 설문, 정성적 리서치.

A/B 테스트 불가 시 대안

방법	설명
UX 리서치	포커스 그룹, 설문조사로 정성적 판단
사용자 행동 로그 분석	기존 로그에서 패턴 추출
Retrospective Analysis	변경 후 과거 데이터와 비교하여 효과 추정

> "확신 부족(Lack of Conviction)"도 주의 — 원숭이가 다트 던지듯 무작위 실험은 비효율. Amazon "Buy Now" 버튼 수준의 고트래픽이 아니면 **가설 기반 실험**이 핵심.

비정규성 처리 — A/B 테스트의 통계적 함정

- A/B 테스트 = 본질적으로 Z/t-test → **정규 분포 가정**
- 대형 테크 기업은 대수의 법칙(CLT)으로 보통 충족
- 충족되지 않을 때의 대안:

방법	원리
Bootstrapping	각 그룹에서 리샘플링 후 평균 → CLT 적용
비모수 검정	Wilcoxon rank-sum test — 정규성 가정 불필요
데이터 추가 수집	예산/시간이 허용하면 표본 확대로 대표성 확보

다중 검정 문제 (Multiple Testing Problem)

- 100개 A/B 테스트를 동시 실행 → **순수 확률로 일부가 유의미하게 나옴**
- α = 0.05일 때 100개 중 약 5개는 false positive

해결 방법:

방법	원리
Bonferroni Correction	α를 테스트 수로 나눔 (보수적)
FDR 제어	False Discovery Rate = FP/(FP+TP) 관리
FWER 제어	Familywise Error Rate — 1개 이상 Type I 오류 확률 관리

> 통계적 보정만으로 한계 → **활성 실험 간 상호작용**을 직접 조사하는 것이 현실적

네트워크 효과 (Network Effects)

- **가정 위반**: A/B 테스트는 각 유저가 독립이라 가정 — 소셜 네트워크에서는 성립하지 않음
- **오염(Spillover)**: Treatment 그룹의 행동이 Control 그룹에 영향
  - 예: Facebook Live 신기능 → 실험군 참여 증가 → 대조군 친구와 더 많은 상호작용 → 대조군 참여도 상승
  - 결과: **실험 효과가 과소평가**됨 (대조군이 오염되어 차이가 줄어듦)

해결: 연결된 사용자를 **클러스터로 묶어** 클러스터 단위로 실험/대조 배정

- 그래프 파티셔닝(normalized cuts 등)으로 소셜 서브네트워크 분리

신규성 효과 & 변화 저항 (Novelty vs Primacy)

효과	설명	사례
Novelty Effect	신기능에 호기심으로 몰려듦 → 초기 지표 과대 추정	Facebook 이모지 반응 출시 → 참여 급등 → 이후 안정화
Primacy Effect	익숙한 것에 대한 집착 → 변화에 거부 반응	2006 뉴스피드 출시 → 75만 명이 반대 그룹 가입

감지 방법:

- 신규 유저 vs 기존 유저 결과 비교 → 기존 유저만 유의미하면 novelty/primacy 가능성
- 신규 유저만 대상으로 실험 실행 (단, 신규 유저 ≠ 기존 유저일 수 있음에 주의)

A/B 테스트 결과 판단 & Holdout

p < 0.05 ≠ 자동 출시

- 대형 회사에서는 작은 차이도 유의미해질 수 있음 → **효과 크기(effect size)** 확인 필수
- Guardrail 메트릭 확인: 매출 ↑ but 리텐션 ↓ → 출시 불분명
- **인적 비용**: 배포, CS 티켓, 숨겨진 버그 → 효과가 정당화하는지

Holdout 그룹

- 출시 후에도 **소수(~수 %)의 유저**는 이전 경험 유지
- 목적: 단기 테스트로 잡지 못하는 **장기 효과(lift)** 측정
- Facebook Growth 팀: 분기별 공유 Holdout으로 팀 성과 측정
- 버그 수정/민감 콘텐츠 → Holdout 없이 전체 배포가 적절

Product Questions — 면접 질문 (1-9)

소셜 & 마켓플레이스

- 10.1 Facebook vs Twitter 소셜 그래프의 차이? 측정 메트릭은?
- 10.2 Uber 서지 프라이싱이 존재하는 이유? 효과 측정 메트릭은?
- 10.3 Airbnb에서 A/B 테스트가 어려운 요인은?
- 10.4 Google → Firefox 기본 검색엔진 계약, 2배 요구에 응해야 하는가?
- 10.5 LinkedIn 피드 참여 메트릭과 개선 아이디어는?

사용

-
-
-
-

프로덕트 질문 10-18

메트릭 진단 & 검색

- **10.10** Snapchat DAU 5% 감소(일주일 지속) — 근본 원인 탐색법은?
- **10.11** Pinterest 신규 검색 랭킹 알고리즘 — 영향 측정 메트릭은?
- **10.12** Netflix SF TV 시청 시간 감소 — 수요 문제인가 공급 문제인가?

세그먼테이션 & 가격

- **10.13** Apple 리테일 매장의 고객 세그먼테이션 → 매출 향상 방법은?
- **10.14** iOS의 70%가 Instagram 사용 vs Android 50% — 왜?
- **10.15** Capital One Quicksilver 카드의 "stickiness" 평가 방법은?
- **10.16** YouTube Premium 신규 국가 출시 — 국가별 가격 결정법은?

질문 유형 매핑 — 이 질문에 어떤 프레임워크를?

유형	해당 질문 번호
메트릭 정의	10.1, 10.2, 10.5, 10.11, 10.15, 10.18
메트릭 변화 진단	10.9, 10.10, 10.12, 10.14
제품 기능 브레인스토밍	10.4, 10.7, 10.8, 10.13, 10.17
A/B 테스트 설계	10.3, 10.6
메트릭 트레이드오프	10.8, 10.16

> 하나의 질문이 **여러 유형을 교차**할 수 있음 — 실전에서는 프레임워크를 유연하게 결합

핵심 요약 — 오늘 배운 프레임워크

메트릭 기초

AARRR 퍼널로 사용자 여정을 구조화.
좋은 메트릭: Meaningful, Measurable, Understandable, Timely.
North Star + Guardrail = 균형 잡힌 측정.

진단 & 트레이드오프

4단계 진단: Scope → Hypothesize → Validate → Classify.
4가지 원인 버킷: 우발 / 자연 / 내부 / 외부.
트레이드오프: Revert / Minimize / Accept.

A/B 테스트

4단계 설계: 메트릭 → 임계값 → 표본 → 배정.
현실 함정: 신규성, 네트워크, 다중 검정, 비정규성.
$p < 0.05 \neq$ 자동 출시 — 효과 크기와 비용을 함께 평가.