

AI READY DATA 2026

현장에서 본 AI Ready 데이터의 조건

AI를 동료로 만들기 위해, 우리가 먼저 알아야 하는 것

구름 · (주)빅밸류 공동대표 · 2026. 9. 3 · 포스코타워 역삼 이벤트홀

PART 1

우리는 왜 AI를 믿지 못하는가

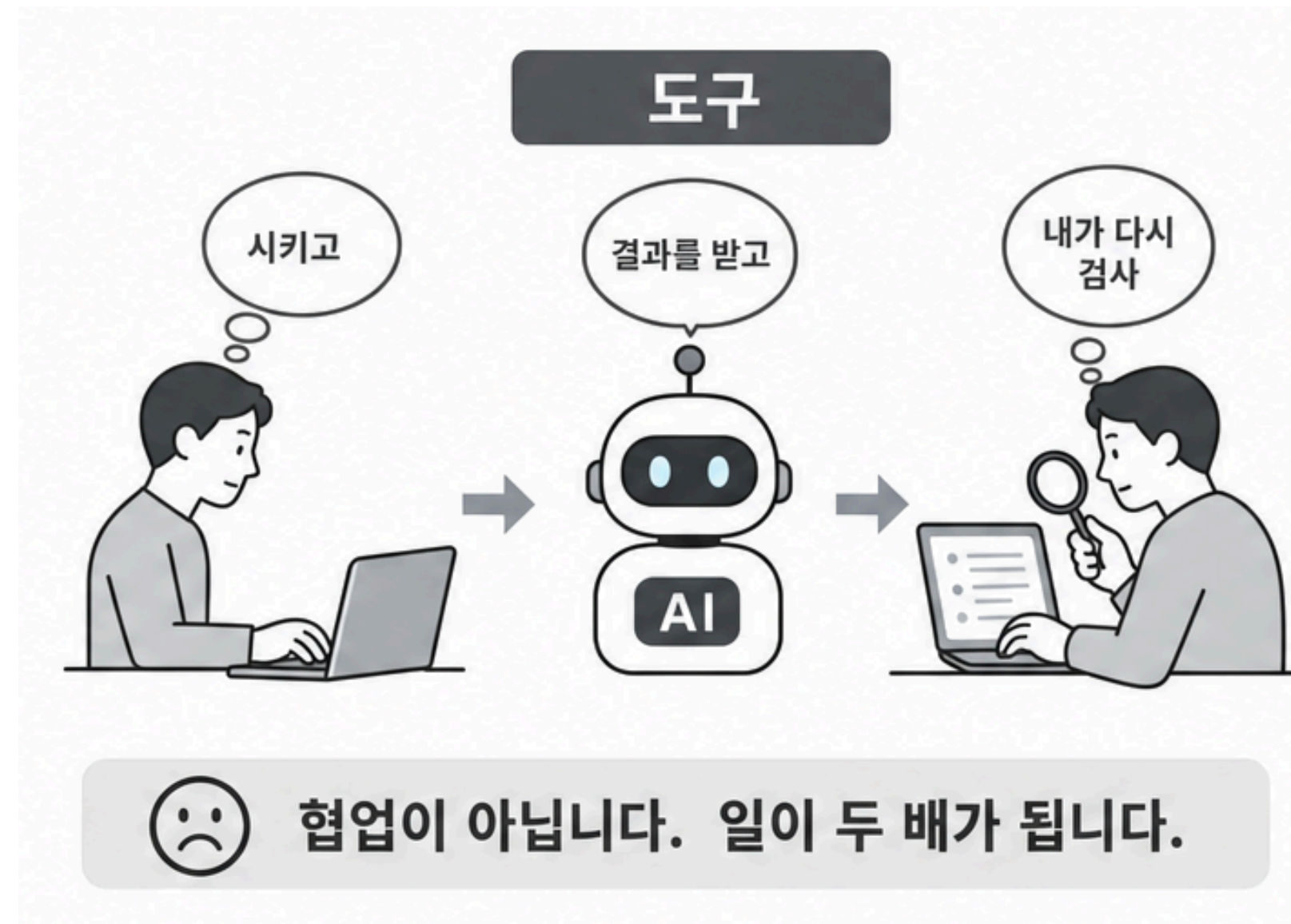
AI 기술은 이렇게 빠르게 좋아지는데,

**우리는 왜 아직도
데이터 이야기를 하고 있을까요?**

AI와 함께 일한다는 것

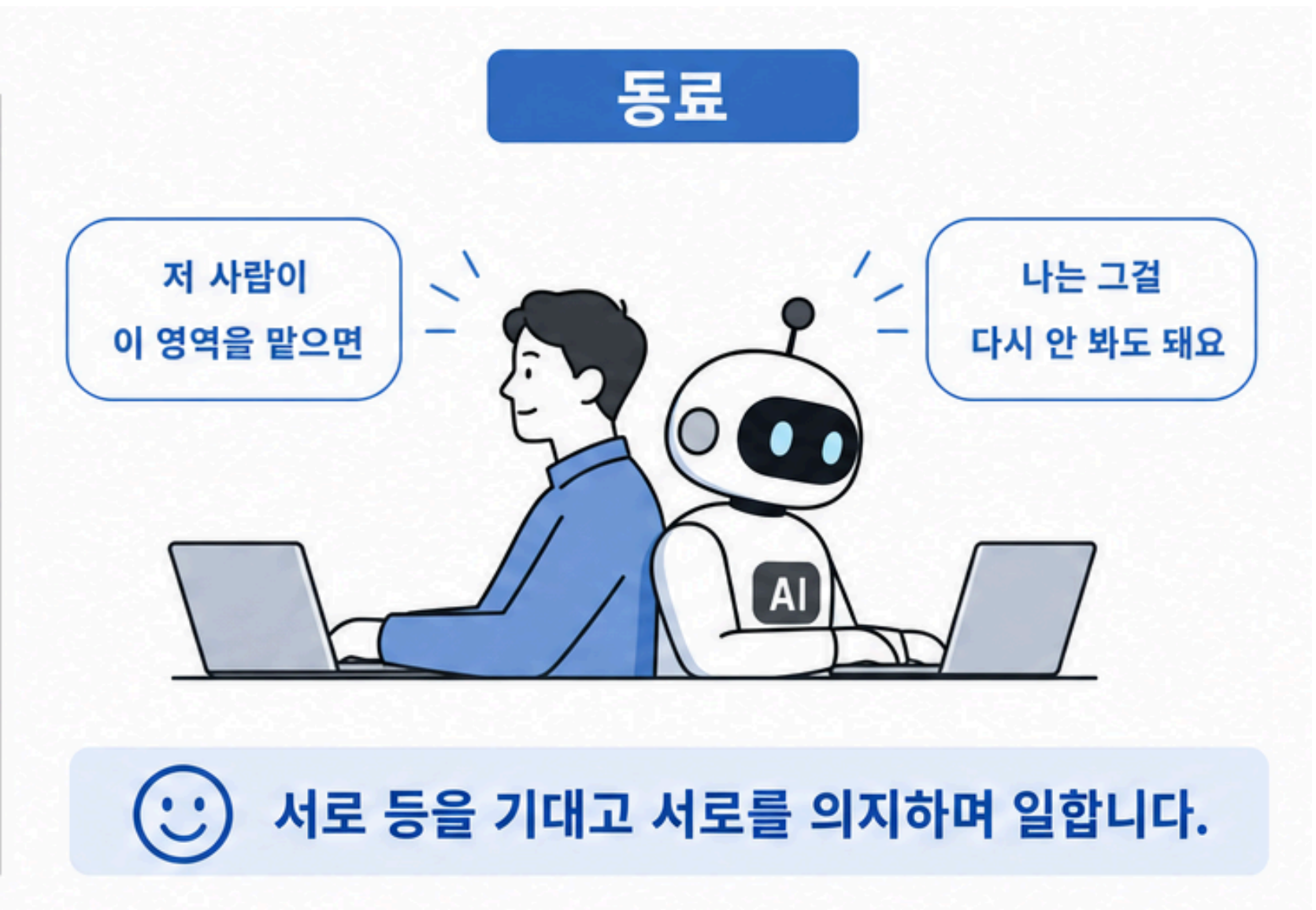
도구

시키고, 받고, 내가 다시 검사한다 — 일이 두 배가 된다



동료

서로 등을 기대고, 서로를 의지한다 — 다시 안 봐도 된다



**동료가 되려면,
믿어야 합니다.**

~~할루시네이션 때문일까~~

모델은 성실했습니다.

데이터에 있는 대로 정확히 답했습니다.

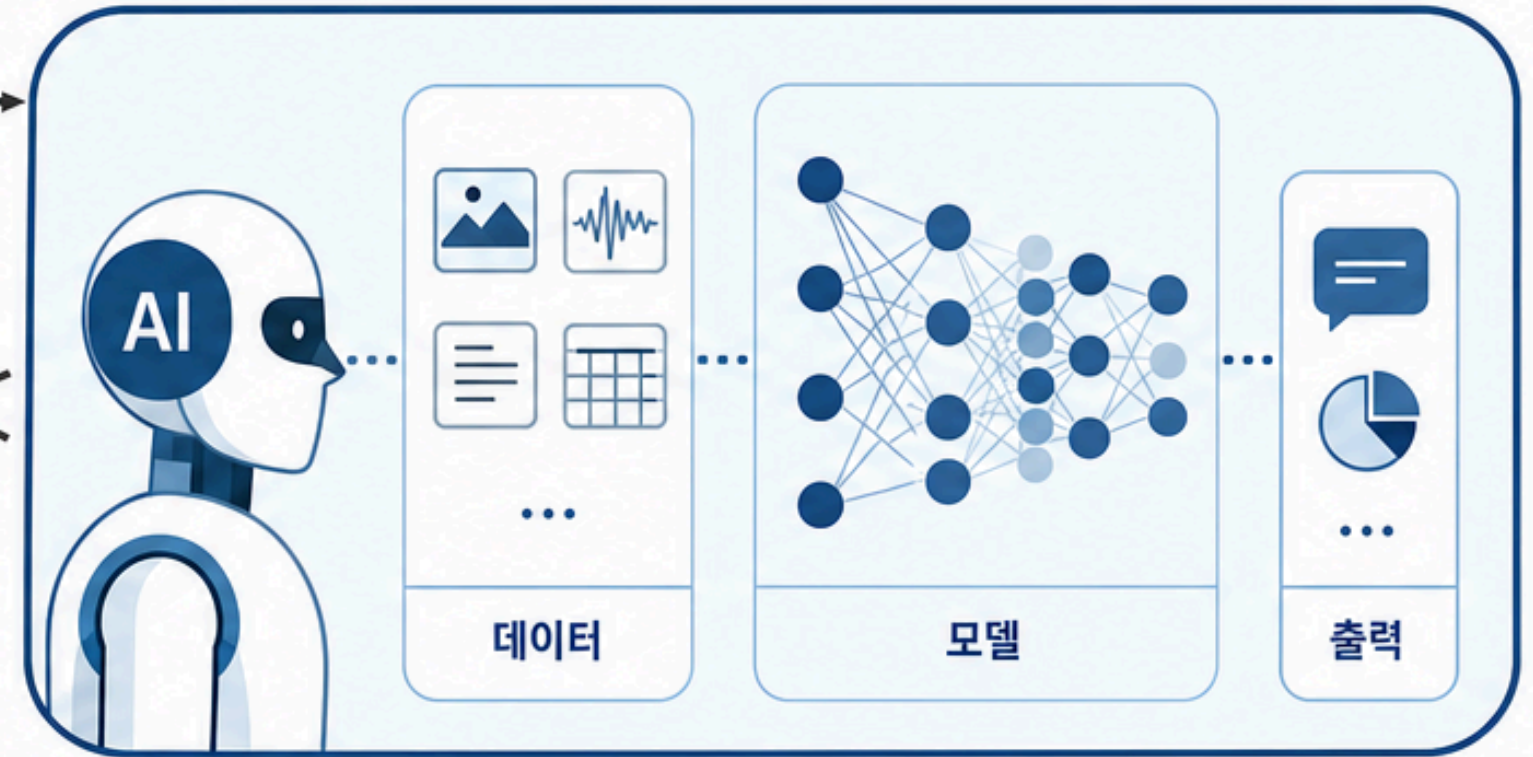
AI는 현실이 아니라 데이터 세계에 삽니다

현실 세계 (사람)



직접 보고, 듣고, 겪는 현실에 삽니다

데이터 세계 (AI)



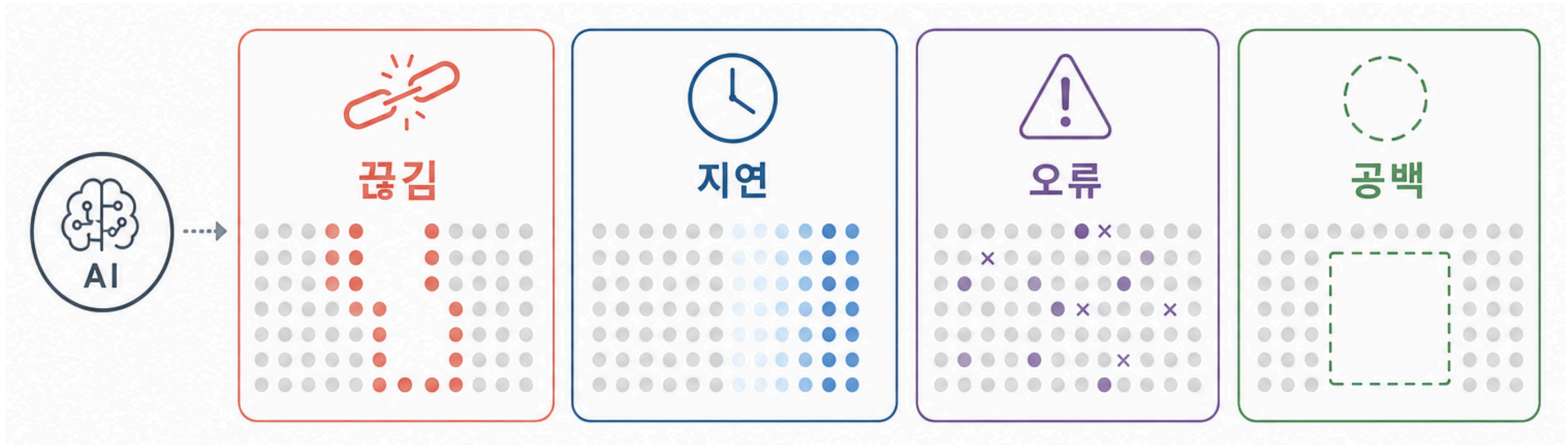
데이터로 만들어진 세계에 삽니다



같다고 전제하고 협업을 시작해서
자꾸 어긋납니다

그 데이터 공간은 온전하지 않습니다

끊긴 곳 · 시간이 늦은 곳 · 오류로 오염된 곳 · 아예 비어 있는 곳



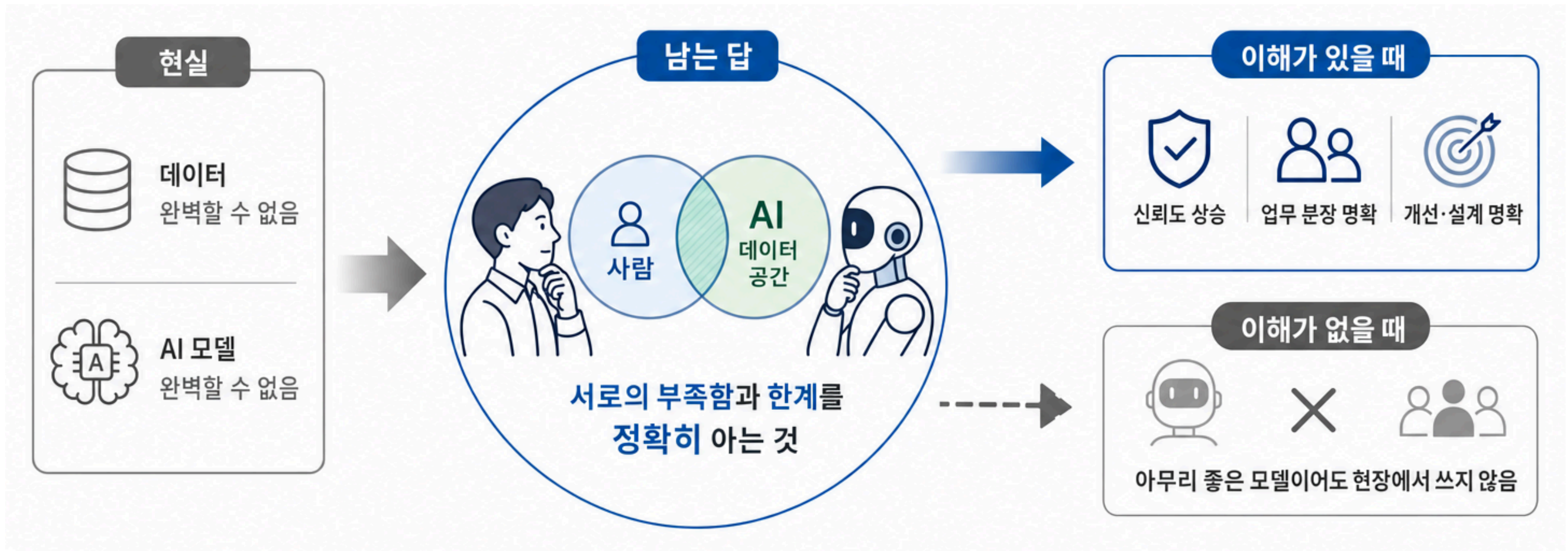
AI는 그걸 확인할 렌즈가 없습니다

사람은 나가서 보고, 현장에 가고, 담당자에게 물어봅니다.
이게 AI와 사람의 가장 큰 차이입니다.

~~데이터를 현실과 완벽히 일치시키면 되지 않나요?~~

**필요한 건 현실의 복제본이 아니라,
이 판단을 내리기에 충분한 데이터입니다**

남는 답은 하나입니다. 서로의 한계를 정확히 이해하는 것

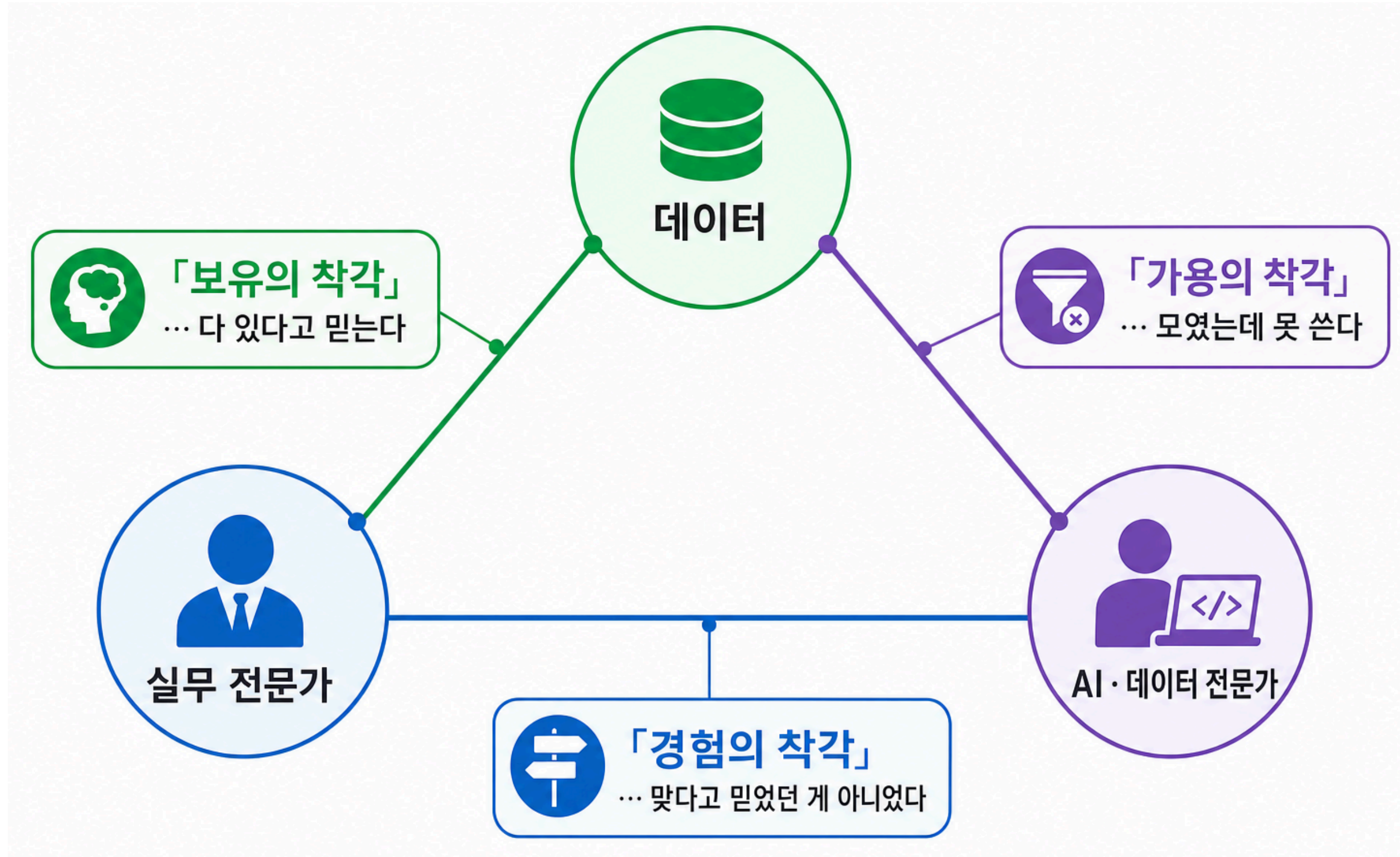


DEFINITION

AI 도입이란?

상호 이해를 설계하고 문화로 만드는 과정

세 주체 사이의 세 개의 괴리

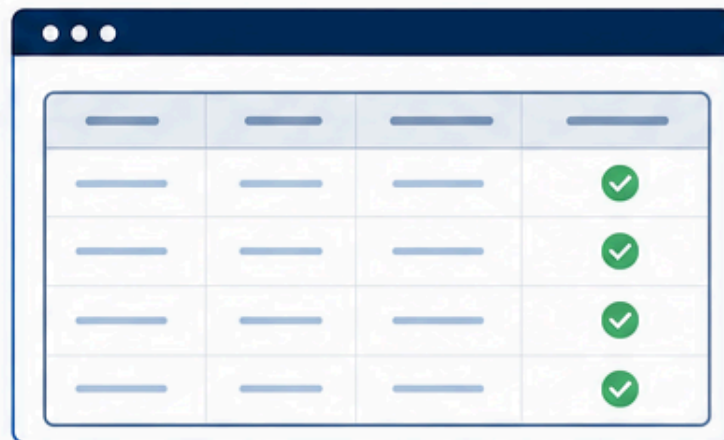


관측 창 의 문제

「우리 그 데이터 다 있어.
이렇게 쓰면 돼.」

— 실무 전문가는
자기 데이터를 잘 안다고 생각합니다.

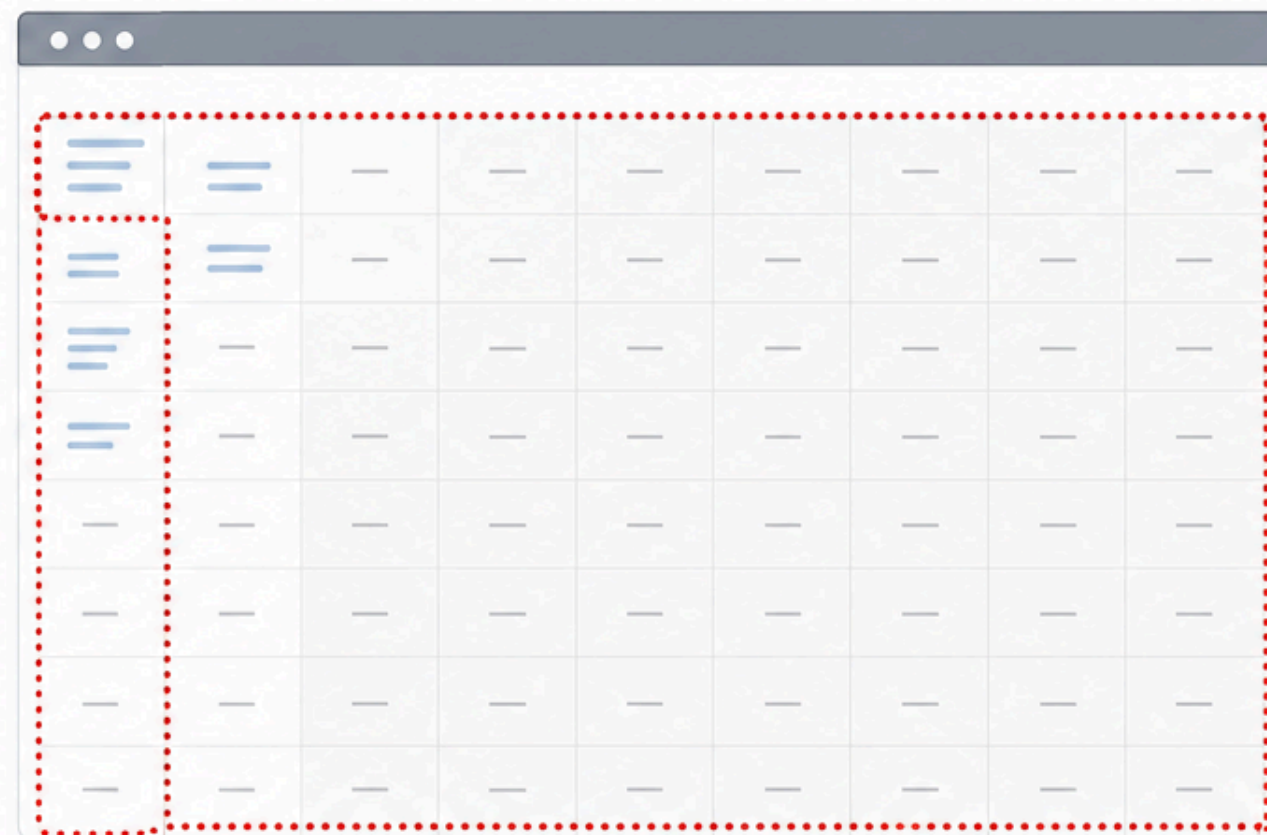
그런데 대부분 자기 업무 화면으로만
봐 왔습니다. 전체 모양을 한 번도 본 적이 없습니다.



			✓
			✓
			✓
			✓

좁은 업무 화면으로 보는 일부

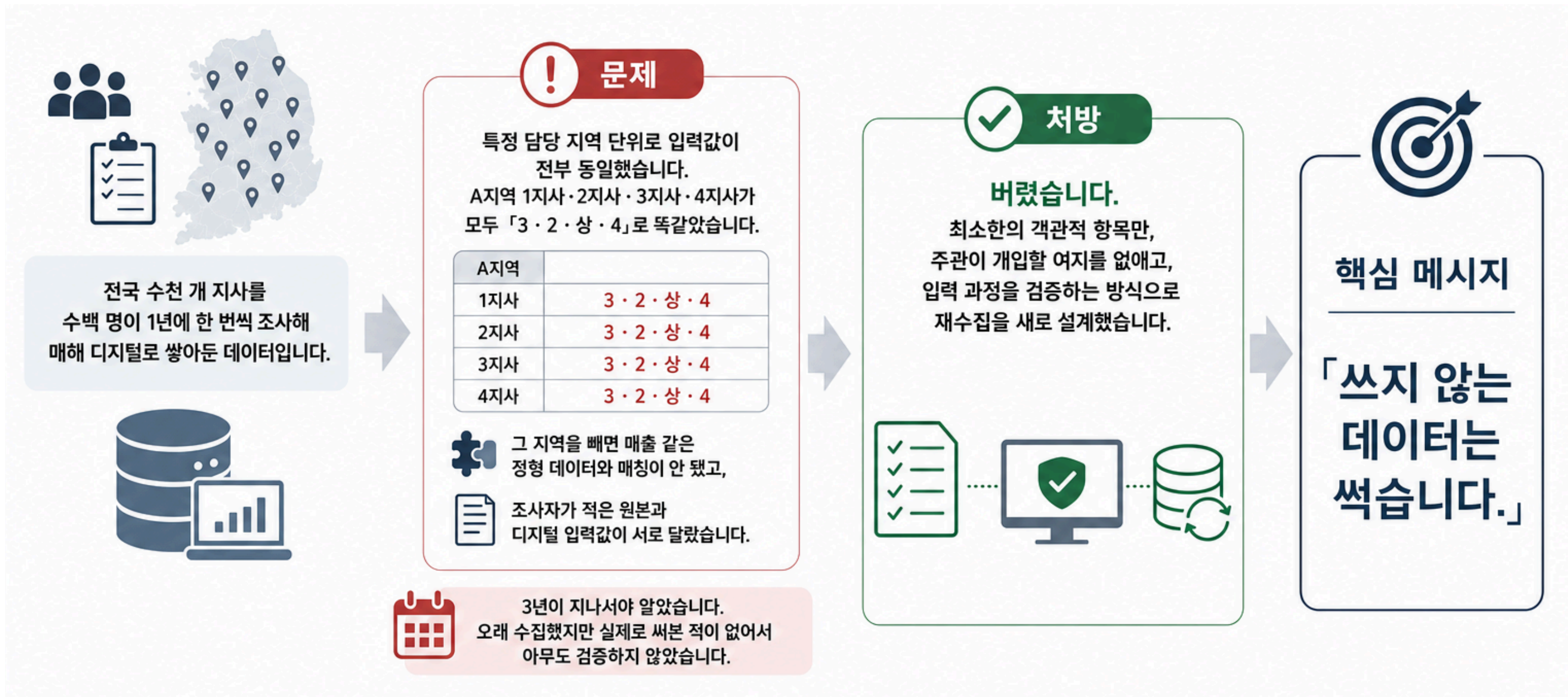
본사에서 늘 완벽하다던 대리점 데이터를 열어보니,
5% 이내를 제외하고 **전부 공란**이었습니다.



실제 데이터 전체의 모양

실무자가 거짓말한 게 아닙니다.
나머지는 열어볼 이유가 없었을 뿐입니다.

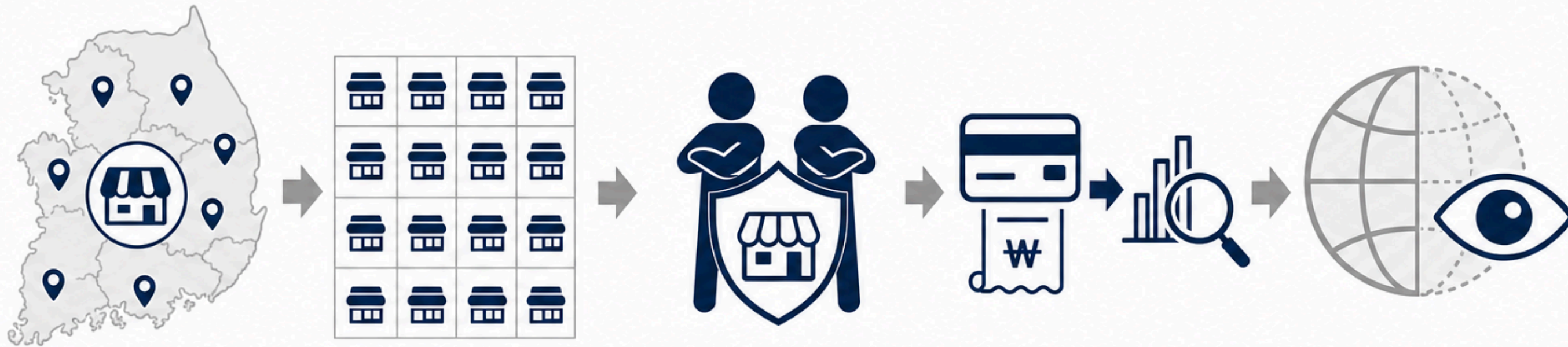
전국 조사 — 결국 버렸습니다



논밭 갱신 지연 — 위성으로 만들었습니다



자사 매출만 — 사 왔습니다



전국 점유율 30%인
프랜차이즈가
자체 매출만으로
진출지를 골랐더니
가는 족족 실패했습니다.

결측률은 **0%**였습니다.
칸은 다 차 있었습니다.
그런데 세상의 **70%**를
못 봤습니다.

원인:
경쟁 매장이 강력한
충성도를 확보한
지역을 구분하지
못한 것이었습니다.

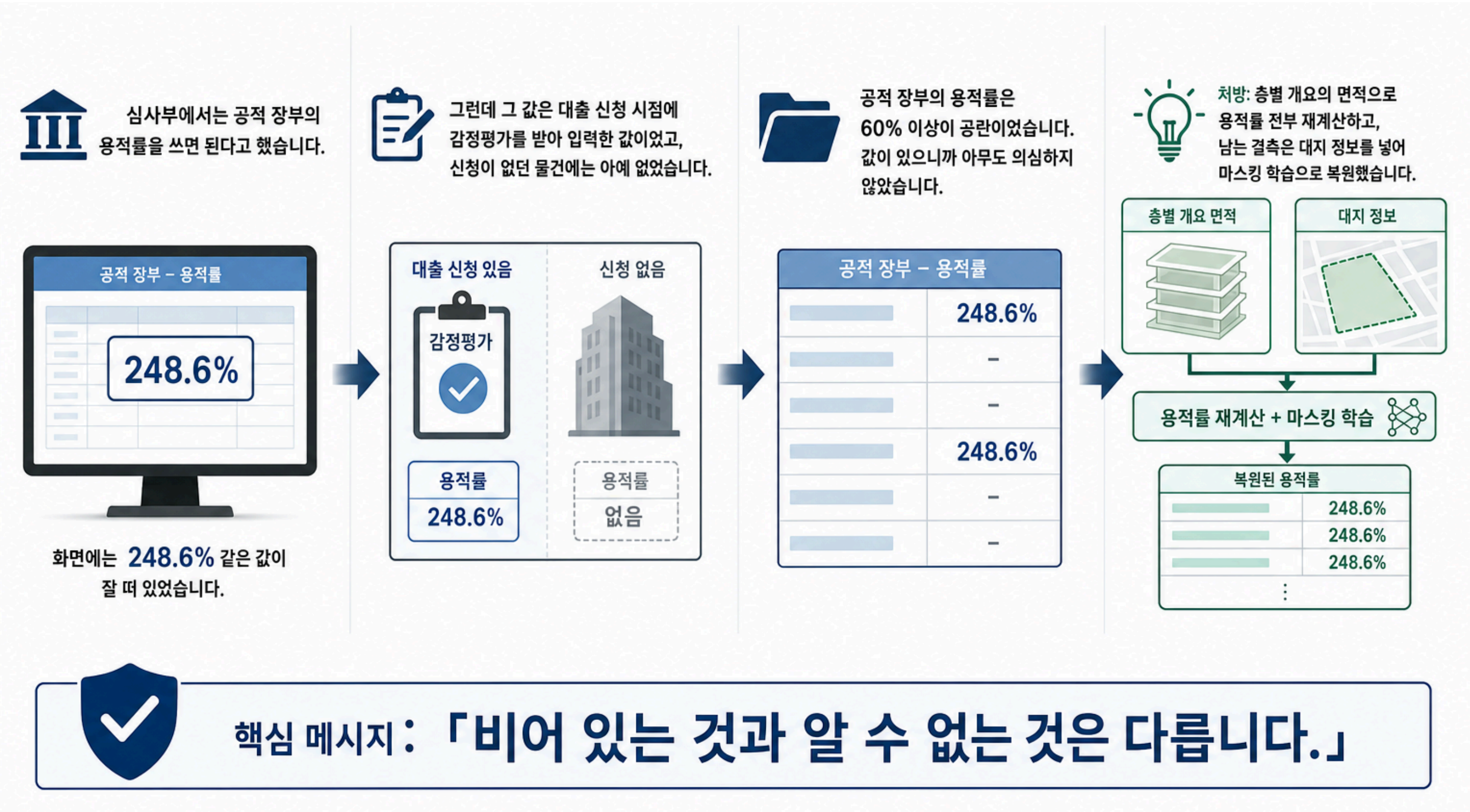
처방:
카드사 매출액을 구매해
경쟁 지점의 매출을
추정하고 변수로
반영했습니다.

핵심 메시지:
**「나를 선택하지 않은 사람이
왜 선택하지 않았는지는
데이터에 없습니다.」**



「나를 선택하지 않은 사람이 왜 선택하지 않았는지는 데이터에 없습니다.」

용적률 — 계산하고 복원했습니다





핵심 메시지: 「비어 있는 것과 알 수 없는 것은 다릅니다.」

데이터가 없을 때 할 수 있는 네 가지



전국 조사



버린다



논밭 갱신



만든다



자사 매출



사 온다



용적률



계산한다



매일 보고서로 확인해
데이터가 있다고 믿었는데,
남은 건 그날그날 정리된
보고서뿐이고 원본이
사라진 경우입니다.



그 기간은 포기하고
지금부터 원본을
남기기로 했습니다.



「데이터 품질이
나쁩니다」는
방어를 부릅니다.



「저희도 이걸
놀랐습니다」로 가야
같은 편이 됩니다.

경험이 이론이 될 때



두 번째 괴리는 사람 사이에 있습니다.
실무 전문가가 확신하는 인과가,
분석해보면 아닌 경우입니다.

원인은 둘 중 하나입니다.



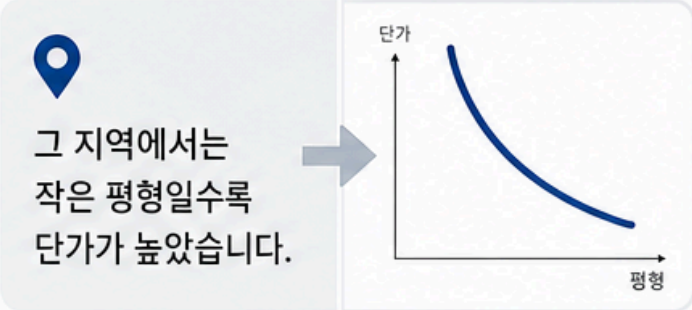
몇 번의 강력한
경험이 강력한
이론으로 굳었거나



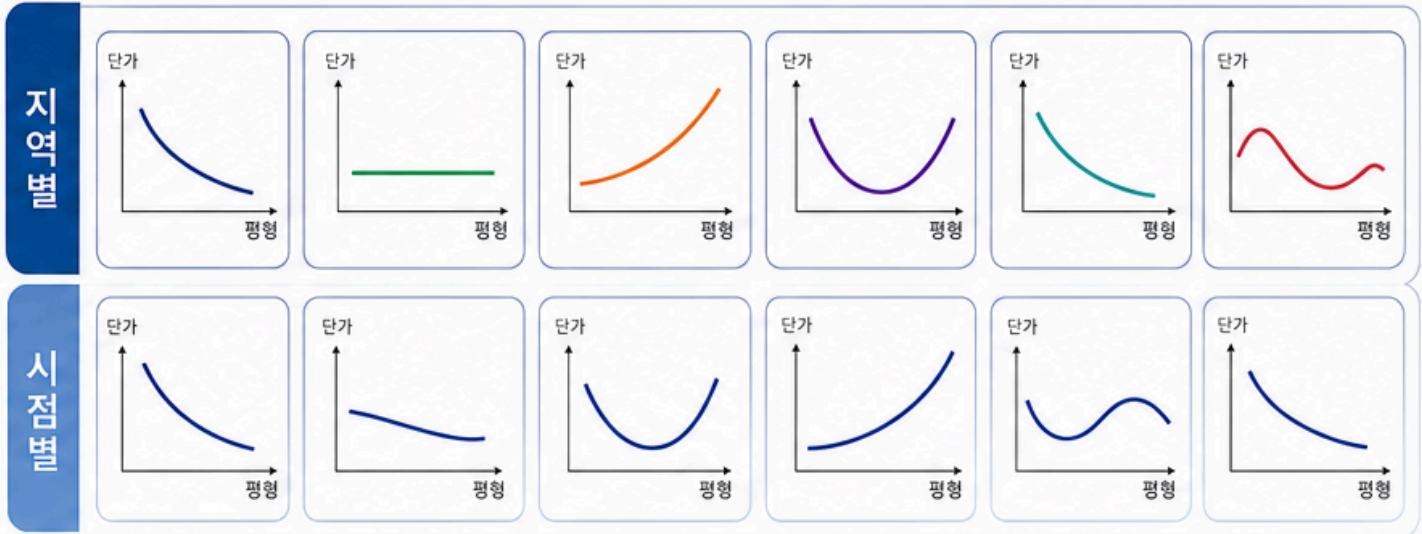
표본으로 본 것이
전수에는 맞지
않았거나

감정평가 — 모델로 넘겼습니다

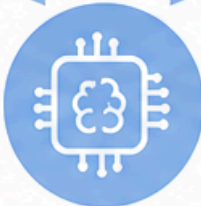
 감정평가사께서
「작은 평형일수록 단가가 높다」고 하셨습니다.
그 지역에서는 정확히 맞았습니다.



 그런데 전국으로 펼쳐보니
시군구마다 평형-단가 곡선의 모양이 달랐고,
어떤 지역은 아예 반대였습니다.
게다가 시점에 따라서도 변했습니다.



 처방:
곡선을 실제로 그려 보여드리고,
이 부분은 전문가 직관이 아니라
학습 모델로 넘겼습니다.



학습 모델로 넘겼습니다.

피자 — 브랜드별로 분해했습니다

? 신규 피자 브랜드가 대형 프랜차이즈 가까이 가면 매출이 줄어든다고 보셨습니다. 데이터로 확인해보니 브랜드마다 반대였습니다.

A • A 브랜드 인접:
가까울수록 매출이 떨어짐 → 경쟁자

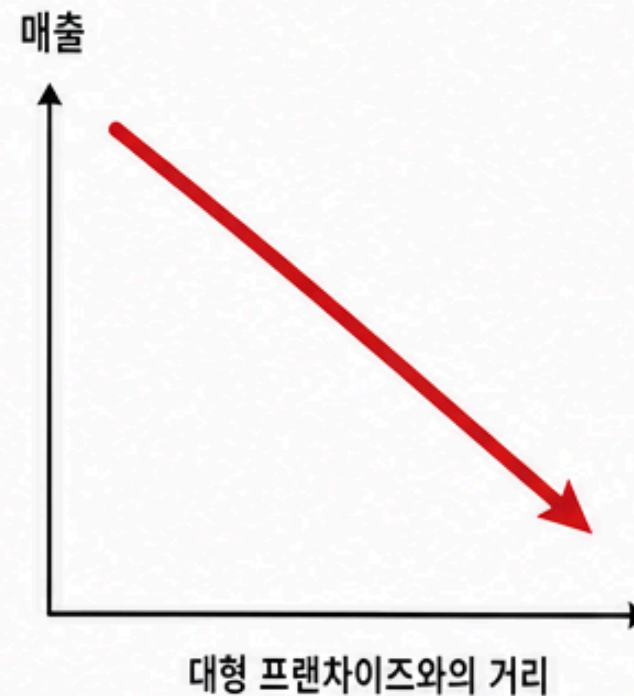
B • B 브랜드 인접:
가까울수록 매출이 오름 → 상권 조성자

💡 오래 자리 잡은 지역은 피자 상권 자체가 만들어져 있어서 소비자가 새로운 맛에 더 빠르게 반응했습니다. 경쟁자와 상권 조성자를 구분하지 못한 것입니다.

📋 처방: 경쟁사를 하나로 묶지 않고 브랜드별 인접성을 각각 변수로 넣었습니다.

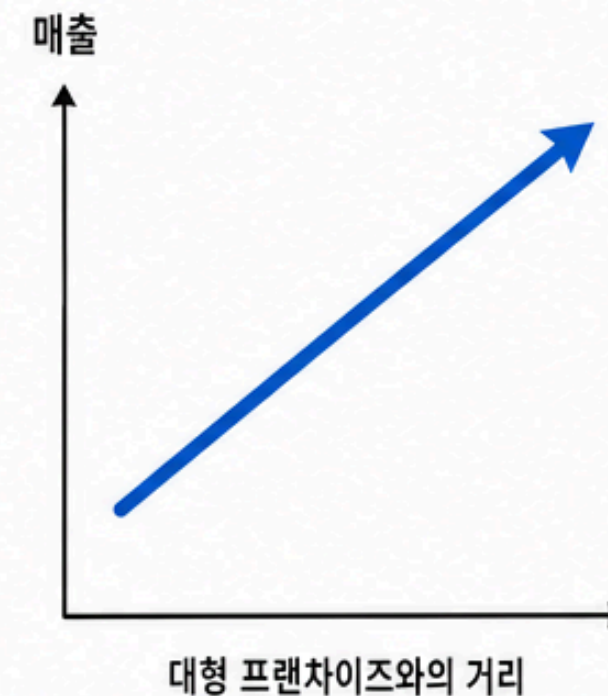
A 브랜드 인접

가까울수록 매출이 떨어짐
→ 경쟁자



B 브랜드 인접

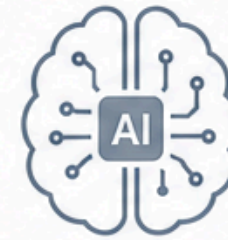
가까울수록 매출이 오름
→ 상권 조성자



오래 자리 잡은 지역은 피자 상권 자체가 만들어져 있어서 소비자가 새로운 맛에 더 빠르게 반응했습니다. 경쟁자와 상권 조성자를 구분하지 못한 것입니다.

📋 처방: 경쟁사를 하나로 묶지 않고 브랜드별 인접성을 각각 변수로 넣었습니다.

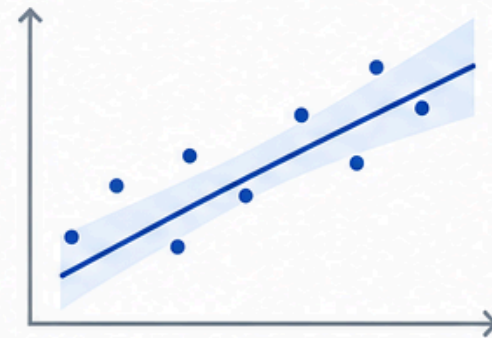
데이터의 양이 모델을 결정합니다



실제로는 가진 데이터의 양과 성격이
쓸 수 있는 모델을 이미 정해버립니다.

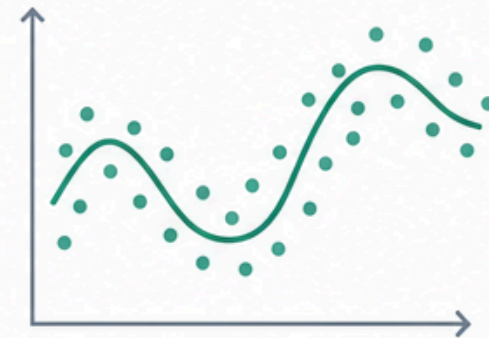
데이터가 적을 때 → 선형 모델

모집단을 가정해야 하고,
그 가정이 곧 모델의 한계가 됩니다.



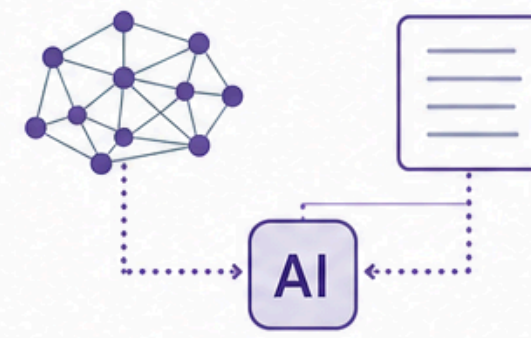
충분할 때 → 비선형 모델

가정 없이 데이터가
관계를 말하게 할 수 있습니다.



적지만 패턴이 있을 때 → 파운데이션 파인튜닝

사전학습된 표현을
빌려 씁니다.



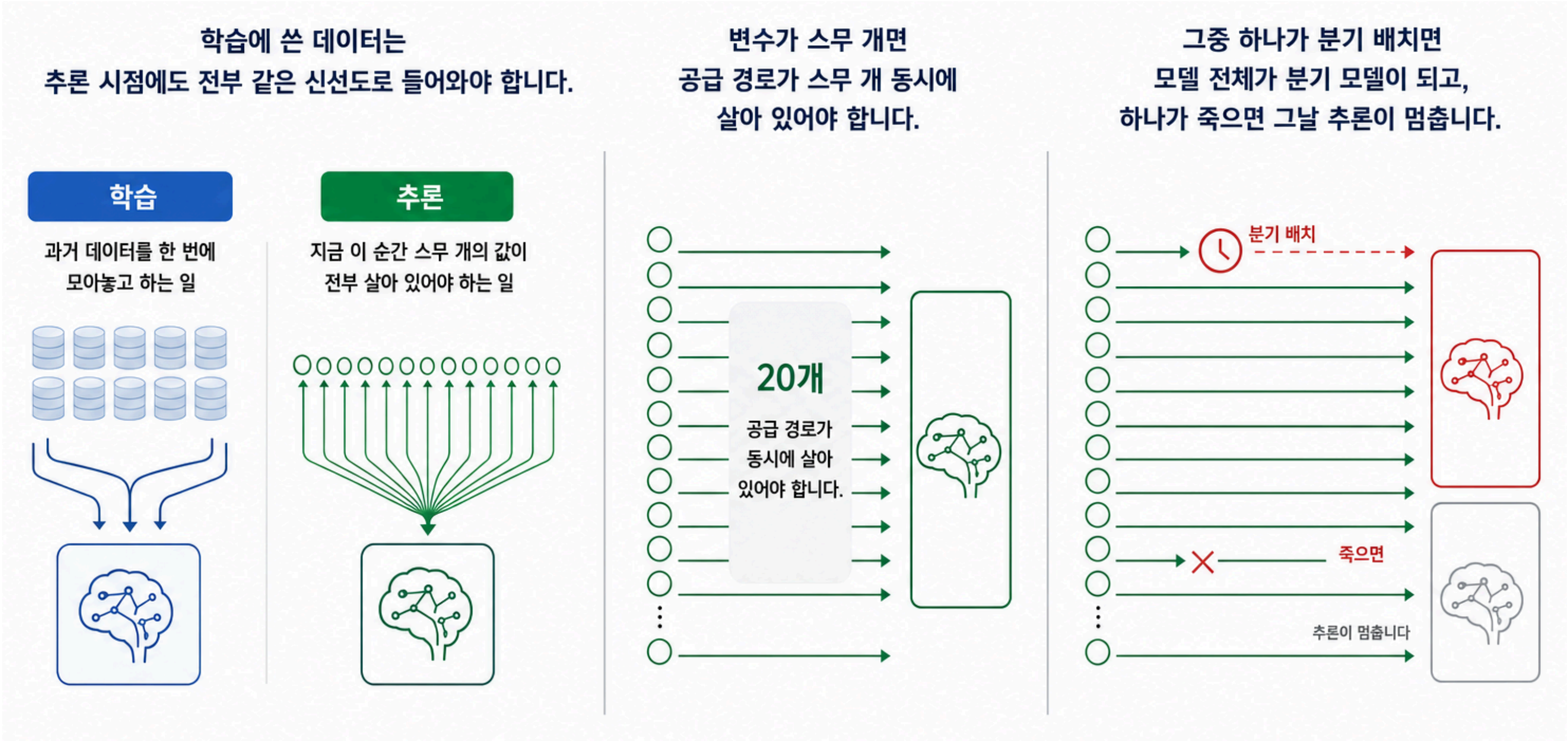
덧붙임:

몇 번의 경험으로 만든
이론은, 표본이 적을 때
사람 머릿속에서
일어나는 **과적합**입니다.



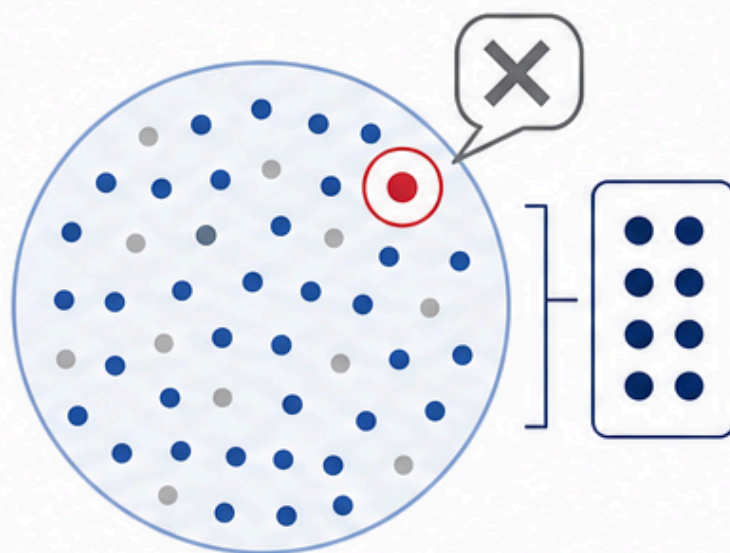
핵심 메시지: 「어떤 AI를 쓸 것인가는 선택이 아니라 **데이터가 정해주는 결과**입니다.」

적시성은 테크닉이 아니라 인프라입니다



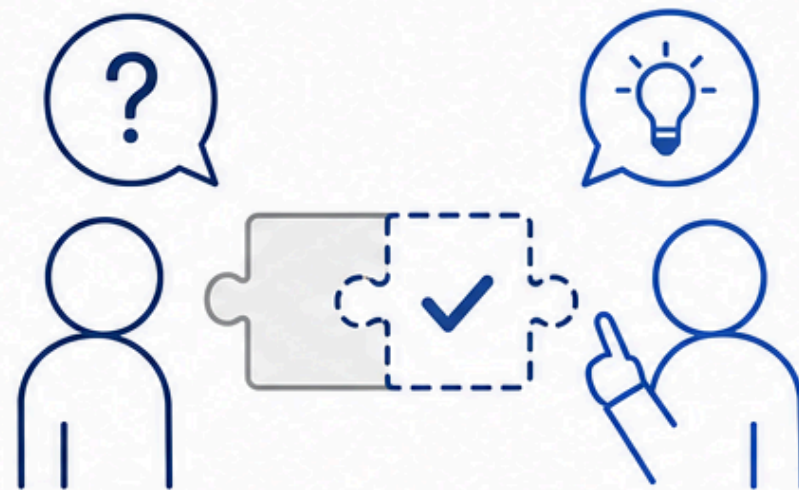
모델은 가장 느린 데이터의 속도로 움직입니다.

사람과 이야기할 때의 세 가지 원칙



1 반례 하나씩은 금지

매번 「예외」로 처리되고
방어만 단단해집니다.
전수 분포를 한 번에 보여드립니다.



2 폐기가 아니라 조건 붙이기

대부분 틀린 게 아니라
조건이 빠져 있었습니다.
적용 범위를 찾아드리면
방어자가 아니라 공동 발견자가 됩니다.



3 인과 검증을 기록으로

도메인 지식이
데이터로 증명됐는지 남겨두면
다음 사람이 같은 논쟁을
반복하지 않습니다.



이 셋 다 안 통하는 경우가 있습니다.

그때는 인정할 때까지 데이터로 계속 입증하는 수밖에 없습니다.

모였는데 못 쓴다

“잘 모여 있는데, 학습에 못 쓴다.”



데이터도 있고



전문가도 왔는데



왜 안 되지



세 번째 괴리는 밖에서 안 보입니다.
실무자 눈에는
「데이터도 있고 전문가도 왔는데 왜 안 되지」
로만 보입니다.

다루는 순서가 있고, 그 순서대로 난이도가 올라갑니다:

연결 → 정확성 → 완결성 → 일관성



연결



정확성

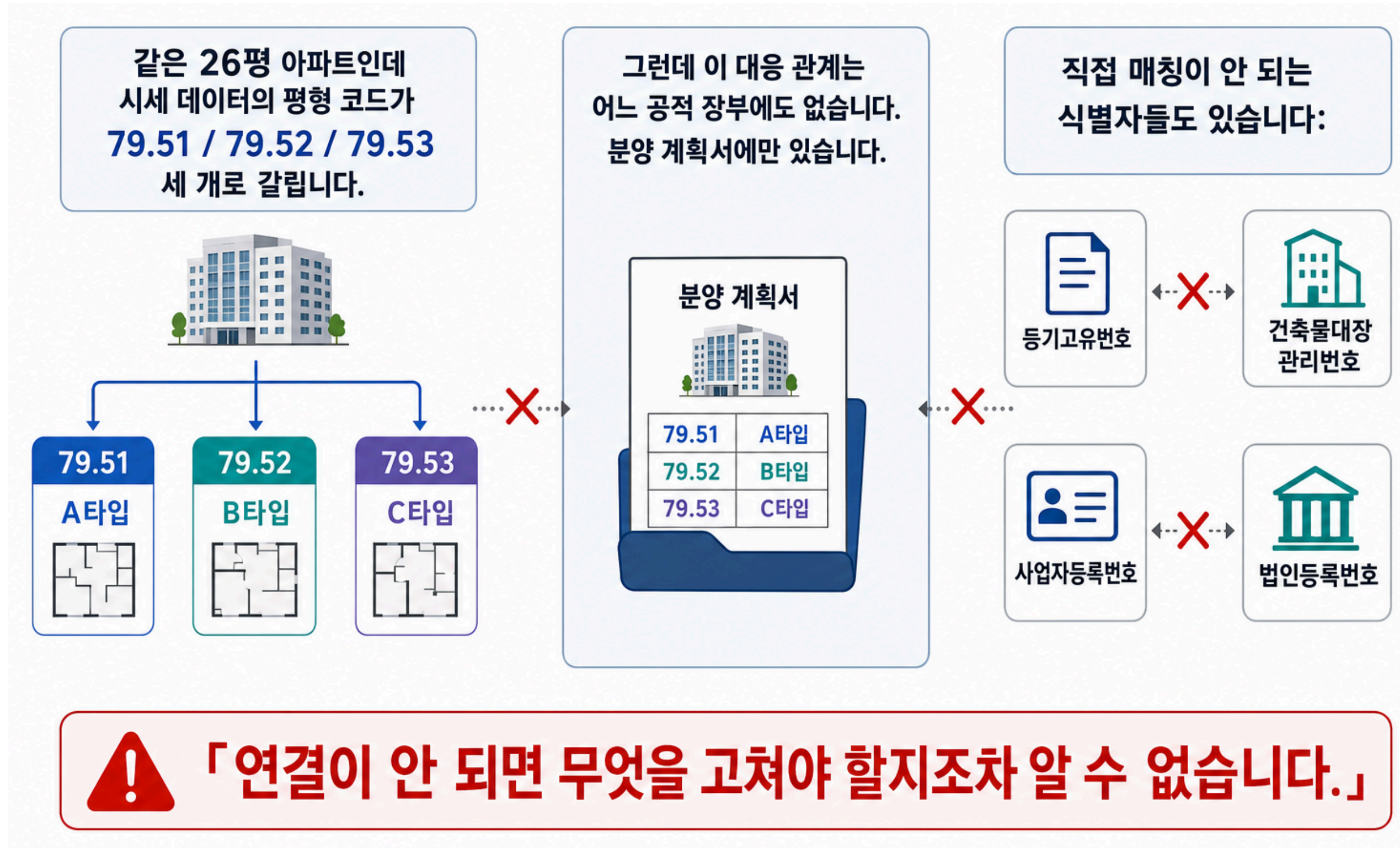


완결성



일관성

불지 않는 키들



같은 필지, 네 개의 대지면적



넷 중 하나라도 어긋나는 경우가
30%가 넘습니다.

그런데 따로 보면 전부 정상입니다.
형식도 맞고, 범위도 맞고,
결측도 없습니다.
어느 하나만 놓고는
틀렸다는 걸 알 방법이 없습니다.

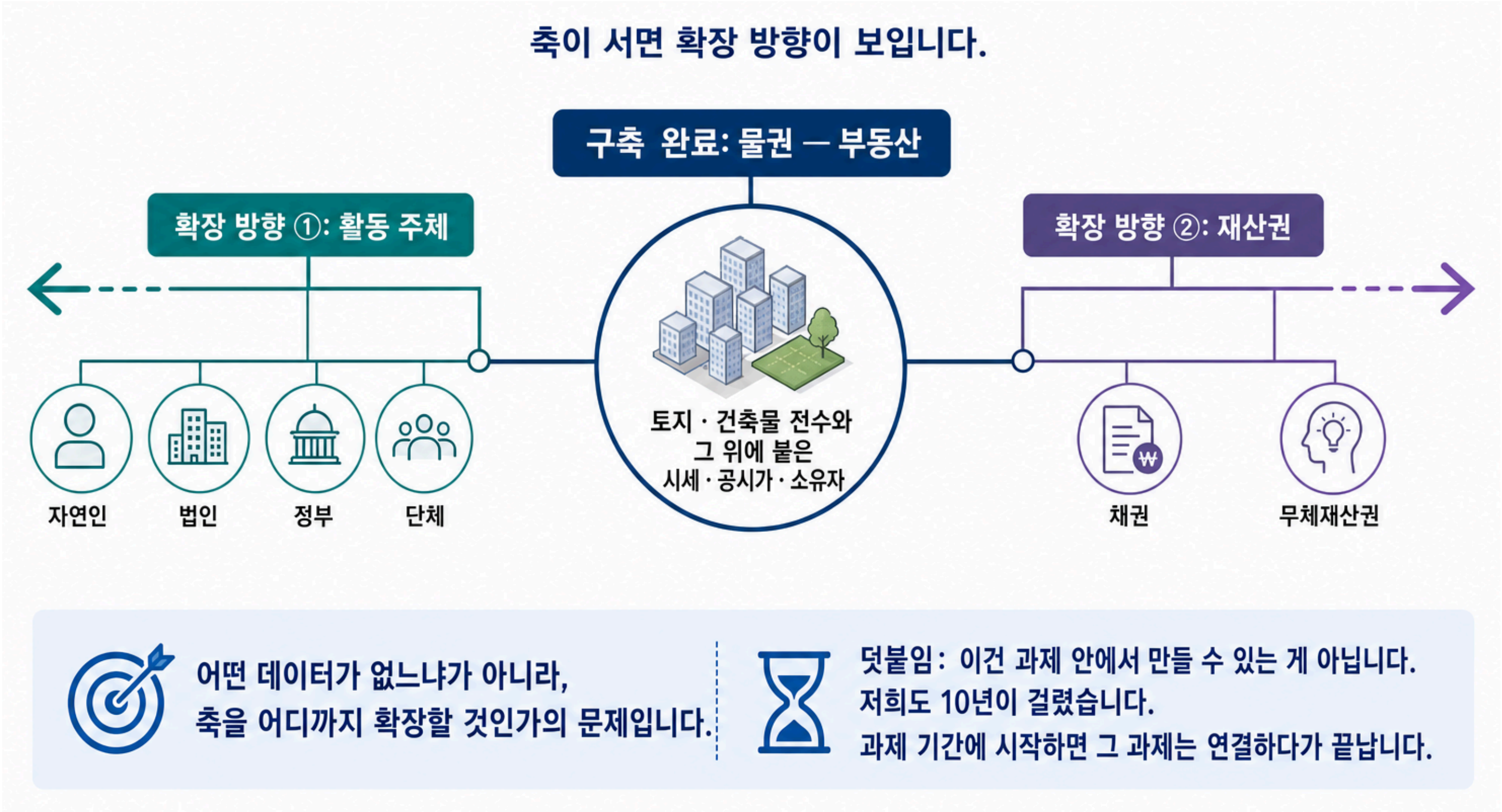


처방: 불일치를 계속 확인해 관리하고,
자동 정제와 정기 정제
프로세스를 고도화하는
문제가 됩니다.

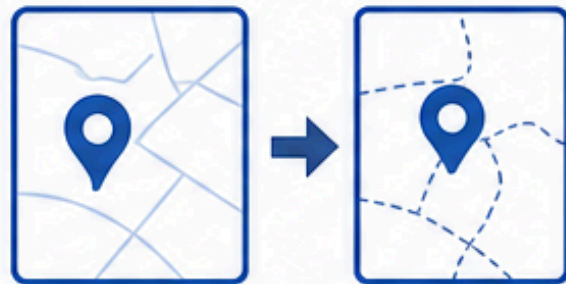
축이 하나 있으면



축을 어디까지 확장할 것인가

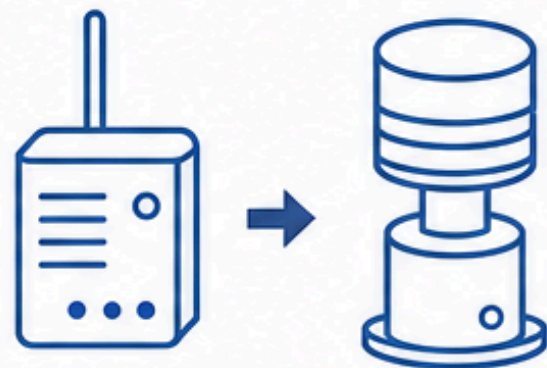


셋은 잡히고, 하나는 안 잡힙니다



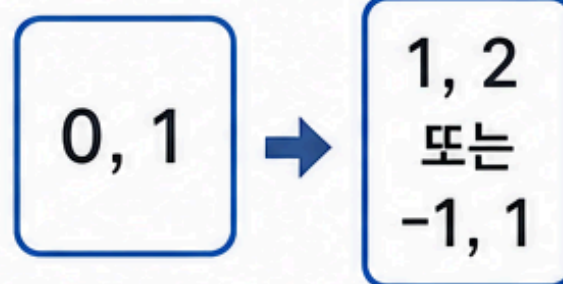
✓ 행정구역 개편

주소·좌표의
시계열 연속성이
끊깁니다.
자체 로직과 코드
체계로 시계열을
유지합니다.



✓ 센서 장비 교체

다른 기계로 바꿔
측정하면 값의
기준이 달라집니다.



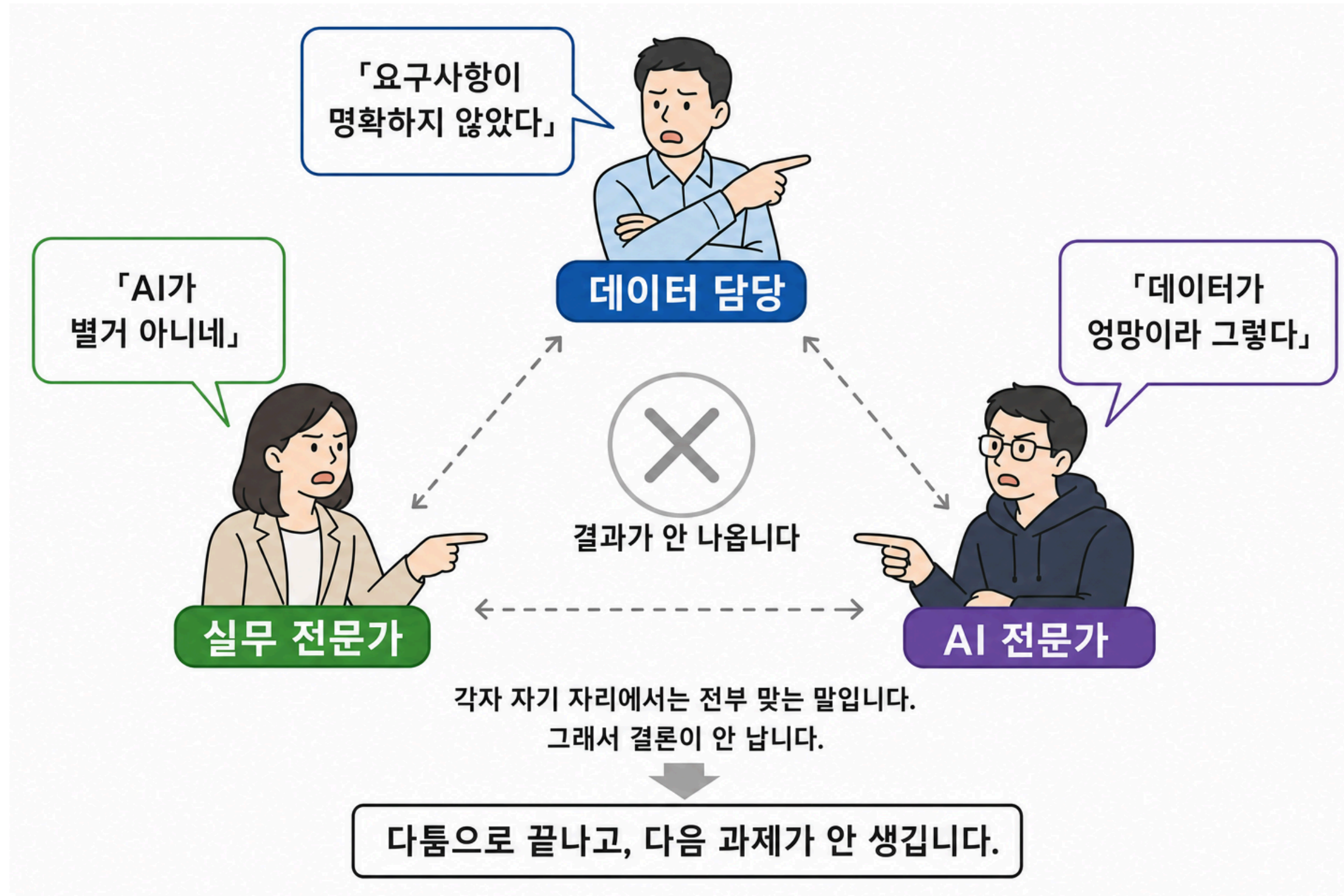
✓ 코드 정의 변경

0과 1로 구분하던 걸
1과 2로, 또는
-1과 1로
바꿔놓습니다.



? 현실 행동의 변화
스키마도 코드도 그대로,
결측도 없고 값도
정상입니다.
바뀐 건 데이터가 아니라
현실에서의 행동입니다.

맞추지 않으면 다툼으로 끝납니다



가이드라인과 FAIR은 어디까지 해결하는가

행정안전부 「AI 친화적 데이터 관리 가이드라인」

6개 지표



완전성



일관성



정확성



적시성



유효성



유일성



진단 기준



진단 방법

FAIR 4원칙



Findable



Accessible

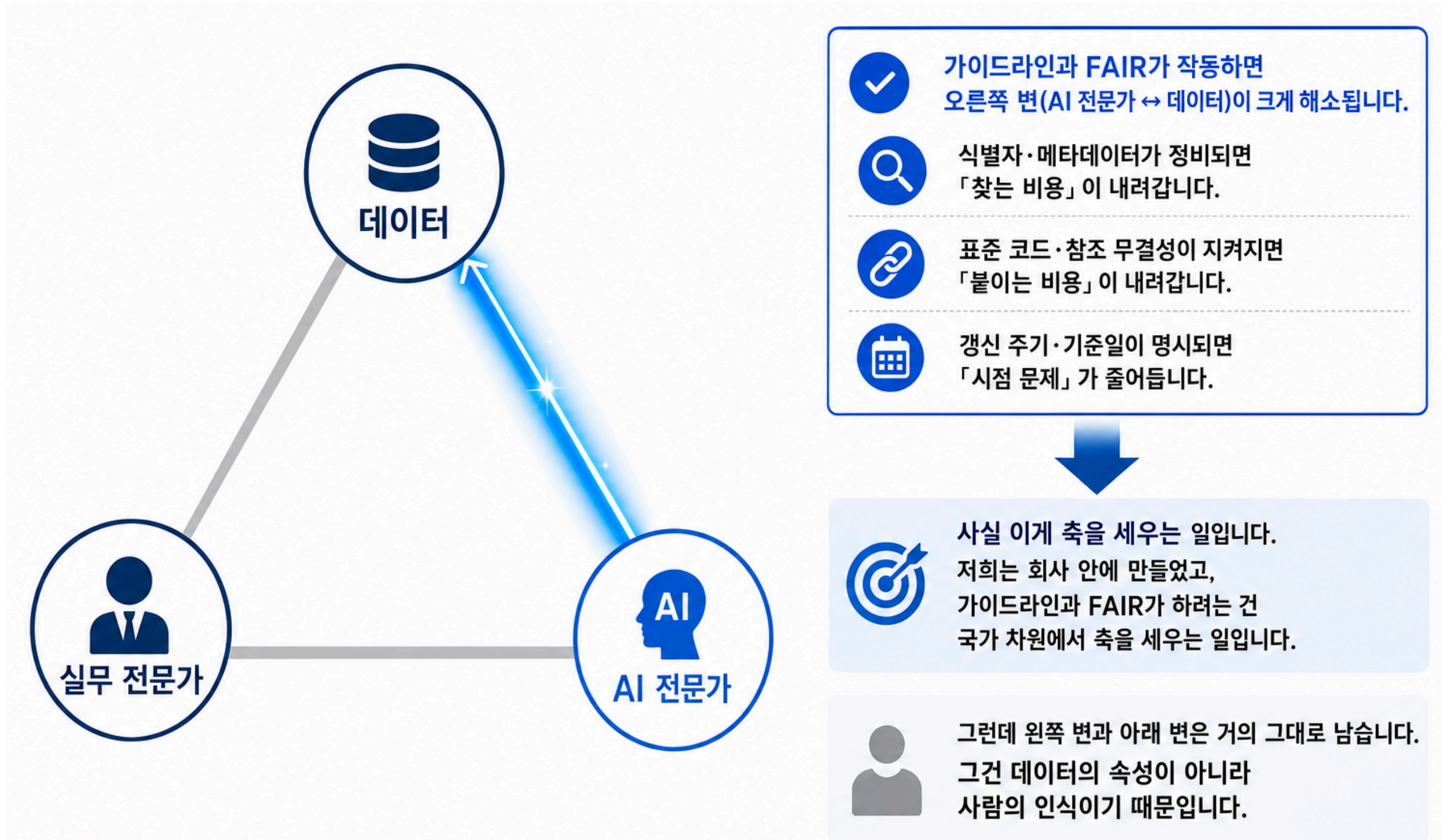


Interoperable

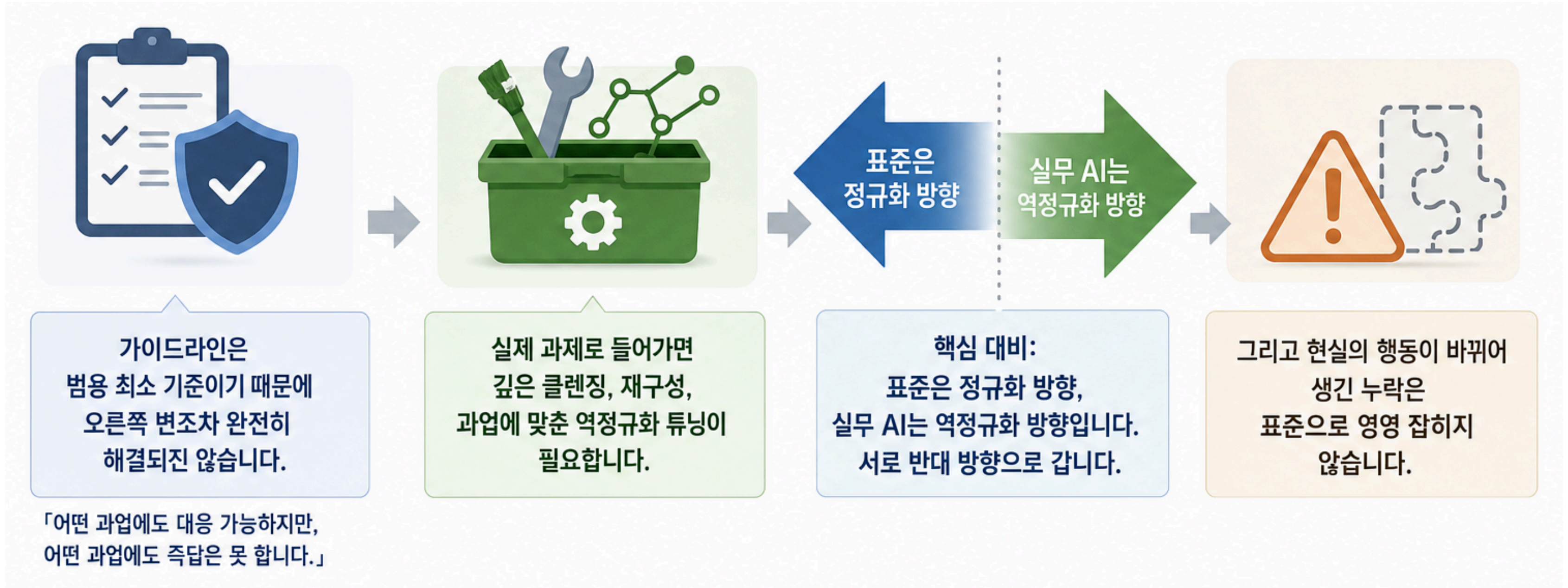


Reusable

해소되는 곳은 오른쪽 변입니다



오른쪽 변도 완전히는 아닙니다



표준은 목적지가 아니라 출발선입니다

표준을 안 지킨 데이터

이 과업이 되는지 안 되는지
판정에만 **몇 주**.
비용이 아까워 시도 자체가 사라집니다.



VS

표준을 지킨 데이터

된다·안 된다가
며칠에 나옵니다.
그래서 시도 횟수가 늘어납니다.



AI에서 가치가 나오는 건 정확도가 아니라 **시도 횟수**입니다.

표준의 가치는 완성품을 주는 게 아니라,
더 많이 도전할 수 있게 **문턱을 낮추는 것**입니다.

AI는 사람을 대체하지 않았습니다

이 과정을 거친 과제들의 결과입니다.



은행 감정평가

2~3시간 → 10분.

무거운 일은 AI가,
판단은 사람이.



점포 개발

**신규 매장 매출
예측 정확도 2배.**

자사 데이터의 한계를 넘어서.



HPAI 방역

매년 +10%.

예측 정확도 개선.

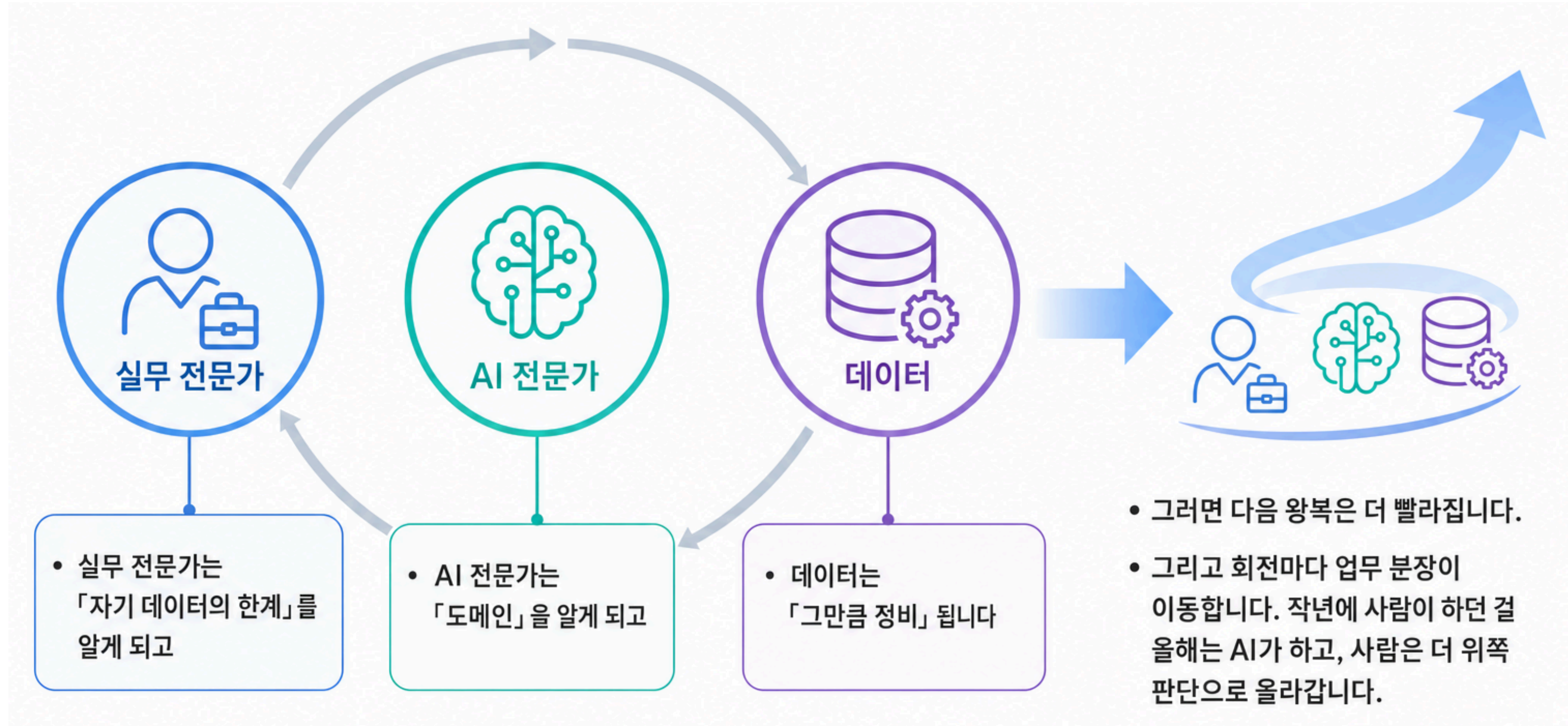


금융 · 유통 · 방역. 공통점이 하나도 없는 세 도메인을
같은 기반에서 시작할 수 있게 해준 것이 그 연결(기준축)입니다.

GPS 기반 이동 경로 데이터 수집 변화



왕복이 한 번 돌 때마다 셋이 바뀝니다



그래서 이걸 한 번 납품하고 끝나는 일이 될 수가 없습니다.

AI는 데이터 세계에 삽니다.

실무자와 AI 전문가와 데이터,
셋이 서로의 한계를 정확히 알 때 비로소 등을 기대할 수 있습니다.

그 선순환을 만드는 것이 AX 전환의 핵심 프로세스입니다.

데이터테크 기업, 빅밸류

 BigValue