

인과추론팀: 커리어 성장곡선을 바꾼 처치

발표자 : LG CNS 김성수
2024.05.25

발표자 소개

안녕하세요,
이번 가짜연구소
인과추론과 실무의
빌더 김성수 입니다.



Data Scientist
Enterprise Data 분석팀 소속
가짜연구소 3기 & Causal inference Lab 4-8기

From Lab To Life
배운 것을 어떻게 실무에 적용할지 고민하고 있음

관심 분야
MLOps & Machine Learning System &
Experiment Platform & Causal inference

가짜연구소 인과추론 팀의 성장

한국어 자료가 많지 않은 인과추론을 대중에게 쉽게 접할 수 있도록 기여, 성장



Chanran Kim 2023. 02. 11.

인과추론 계의 연예인! 잘 부탁드립니다 😊



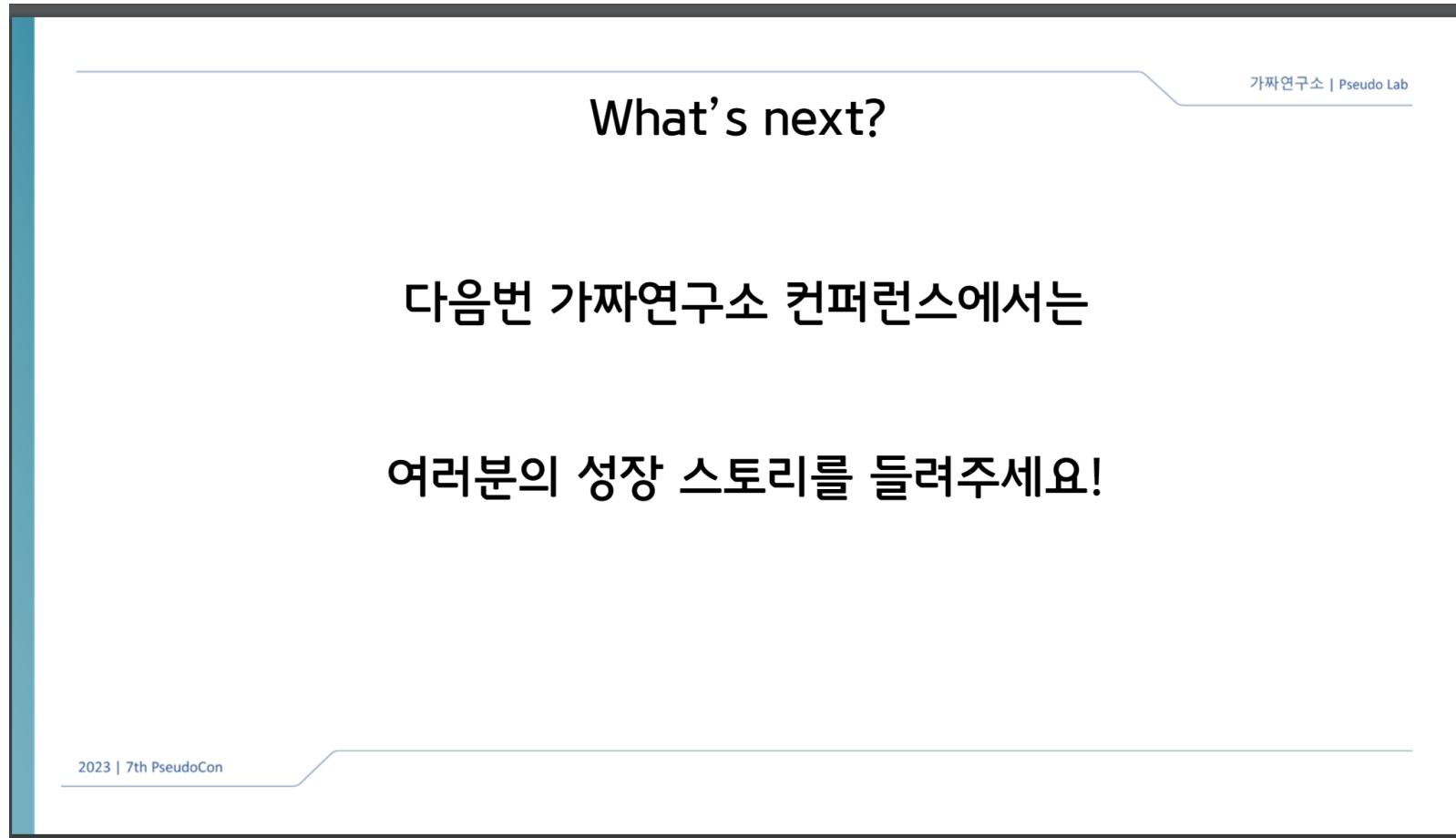
1



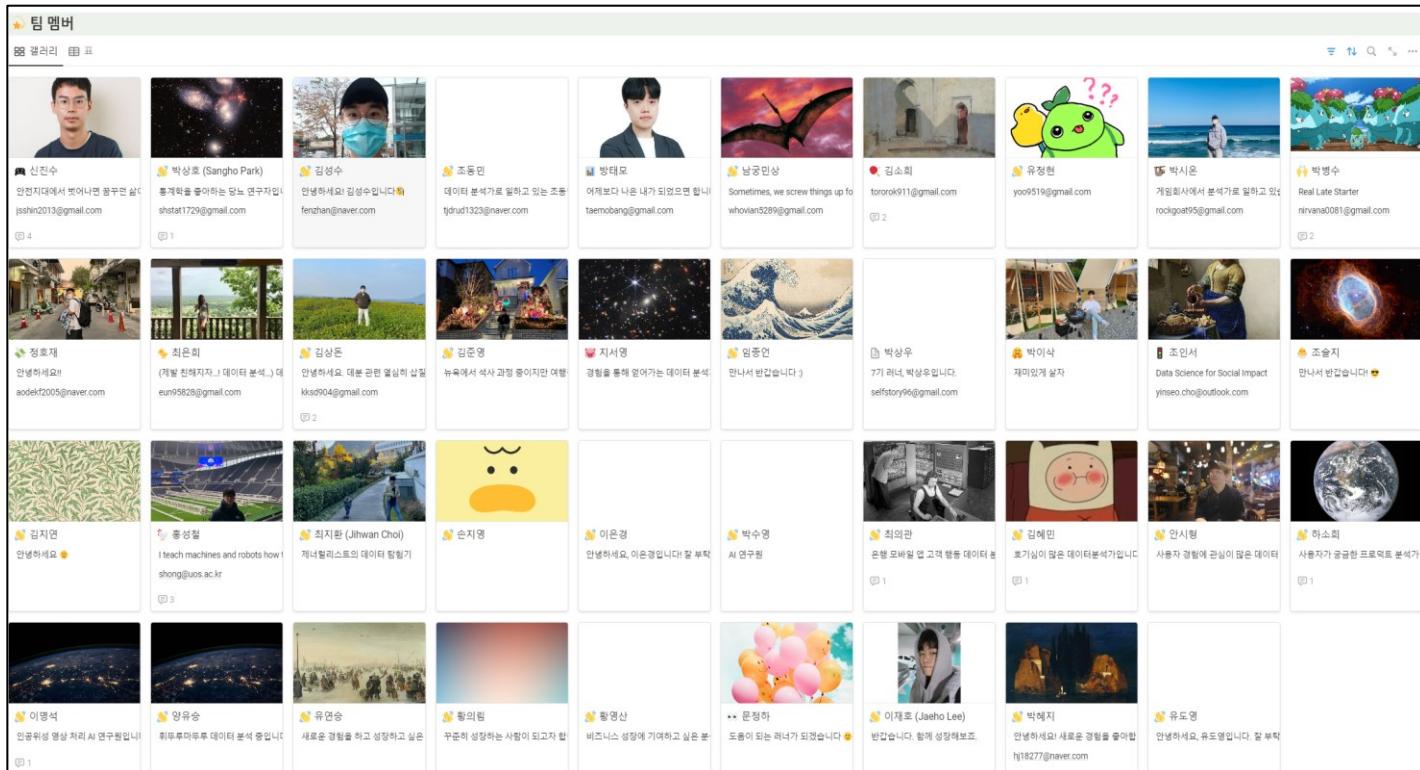
Jinsoo Shin 2023. 02. 14.

연예인이라뇨ㅋㅋㅋㅋㅋ 올해도 잘 부탁드려요 (_-)

Call for presentation by 진수



좋은 조직 != 개인 성장의 보장, 인과추론팀에 속한 개개인의 성장에 초점



Topic

Question 1. 데이터 커리어와 인과추론

Question 2. 인과추론 학습 과정

Question 3. 인과추론 팀과 앞으로 나의 비전

들어가기 전에,
여러분은 왜 인과추론에 관심을 가지게 되었나요?

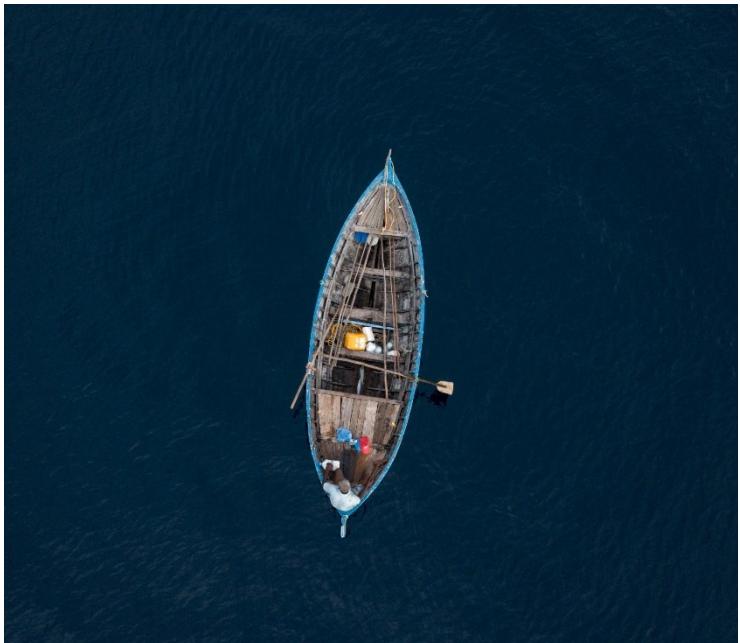
인과추론이 어떤 것인지
궁금해서

교수님 때문에
그냥 따라서

그렇게 어렵다는데
얼마나 대단 한지 보려고

조금 더 쉬운 질문.
어떤 데이터 커리어의 배를 타고 싶으신가요?

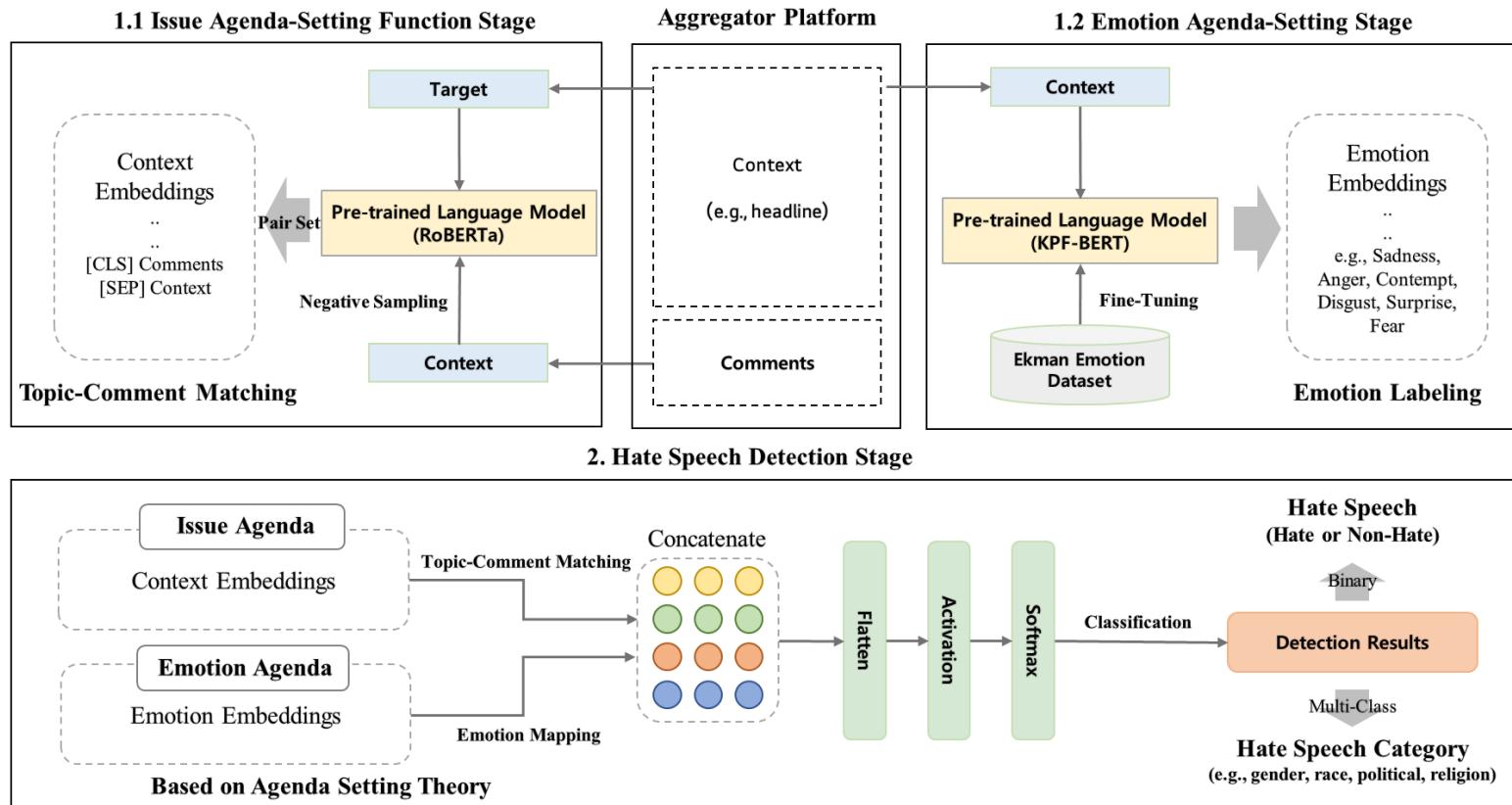
#막막 #방향성 없음 '표류'



#명확 #뚜렷한 목표 '항해'



저는 대학원 시절에, 분류 성능을 높이기 위해 Design Choice 연구를 했고,



저는 대학원 시절에, 의사결정을 위한 분석 프레임워크와 시스템 구축을 했습니다.

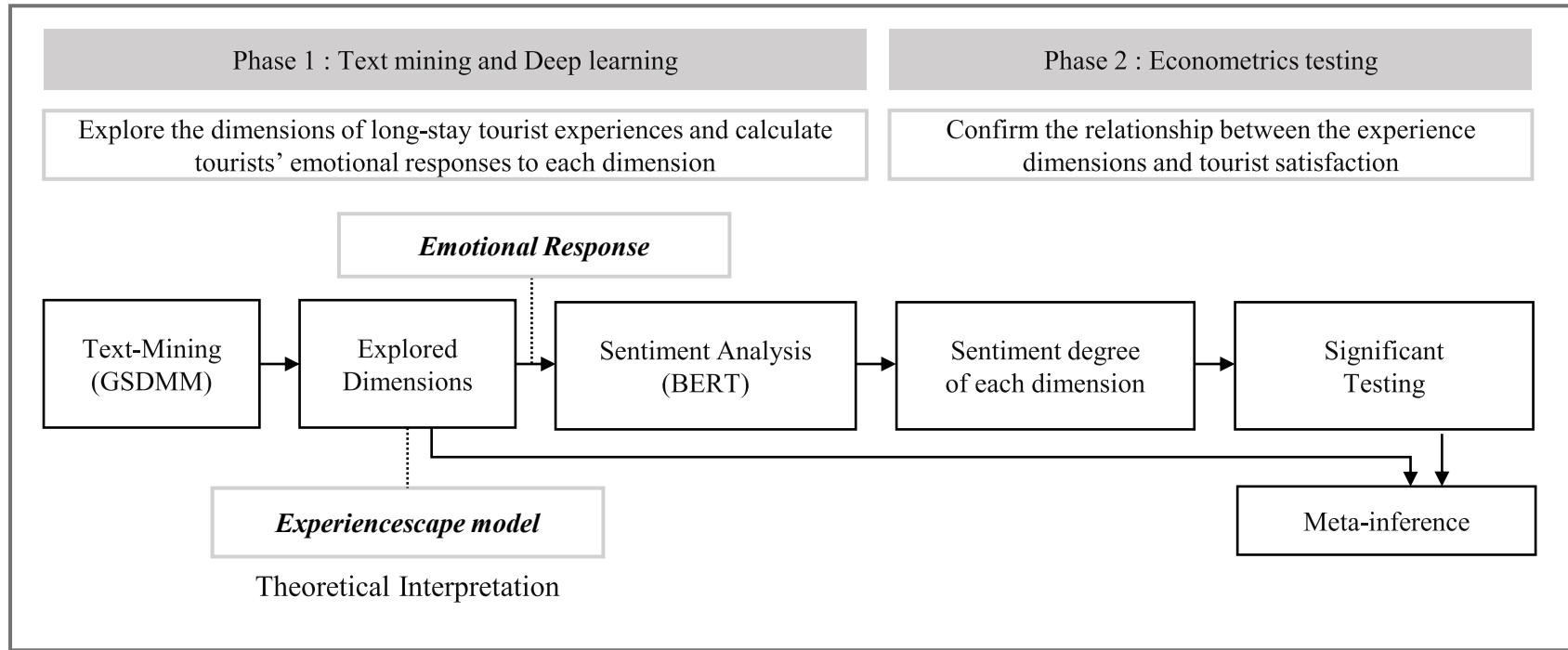


Figure. Research Procedure

2022.02.22,

가짜연구소 4기 인과추론 팀 나의 신청서 : 막연한 관심

지원 동기를 간단히 작성해주세요! *

데이터 분석가에게 인과 관계는 데이터 분석의 핵심이자, 비즈니스에서 끊임없이 고민해야 하는 영역이라고 생각하였습니다.

이렇게 업무를 하며 인과추론의 중요성을 깨닫고 스스로 공부를 하기 시작한 지 벌써 3개월이 지났지만, 세미나에 참여하여 학습하기엔 알았고, 동료 없이 스스로 유튜브를 통해 Quasi-Experiments 등 인과관계 연구방법론을 공부하기에는 내용이 광범위하여 갈피를 잡지 못하였습니다.

이렇게 3개월 동안 많은 시행착오를 겪었고, 스스로 정말로 제대로 공부하고 있는지 의문이 들기 시작했습니다.

그리고 이번에 열린 "Casual하게 Causality 이해하기" 스터디는 이런 저에게 인과관계의 기초 지식을 채우기 위한 새로운 도전이 되리라 생각하였기에 이번 스터디에 지원하게 되었습니다.

좋은 사람들, 좋은 빌더와 함께 서로 부족한 점을 채우고 함께 성장해나가는 것을 목표로 삼고 끝까지 해내도록 하겠습니다.

좋은 교육의 기회 제공해주셔서 감사합니다.



내용 갈피 X : 머라카는지 도통 모르겠다
인과추론 너무 광범위..

#막막 #방향성 없음 '표류'

저 머신러닝만 해본 사람인데요.. 인과추론 들어는 봤어요.

인과추론을 바라보는 관점

머신러닝은 비즈니스 목표 달성의 수단이자 하나의 도구 동의 하시나요?

> 기업의 머신러닝 도입의 목표는 비즈니스 성과 창출입니다.

인과추론도 비즈니스 목표 달성의 수단이자 하나의 도구

> 기업의 인과추론 활용의 목표도 현상을 이해하고, 올바른 의사결정을 통한
비즈니스 성과 창출입니다.

인과추론은 예측 모델링과 목표가 다른 데이터 과학의 수단이자 하나의 도구

데이터 과학 질문 예시

(비지도 학습)

60-80세 여성 중 뇌졸중 과거력을 가진 여성을 그들의 특성에 따라 분류할 수 있는가?

(지도 학습)

내년에 뇌졸중을 겪을 확률은 어떻게 되나요? 여성의 Feature에 따른 확률은?

(인과 추론)

스타틴(Statin)을 줄이면 평균적으로 여성의 뇌졸중 위험을 줄일 수 있나요?

분석 방법 예시

클러스터 분석, ...

회귀분석, 의사결정나무, 랜덤 포레스트,
서포트 벡터 머신, 신경망, ...

매칭, 역확률 가중치, Difference in
Difference, RDD, 도구변수, ...

참고: Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

어떤것이 다른가요?

(1) 추구하는 목표가 다르고 (2) 접근 프로세스가 다릅니다.

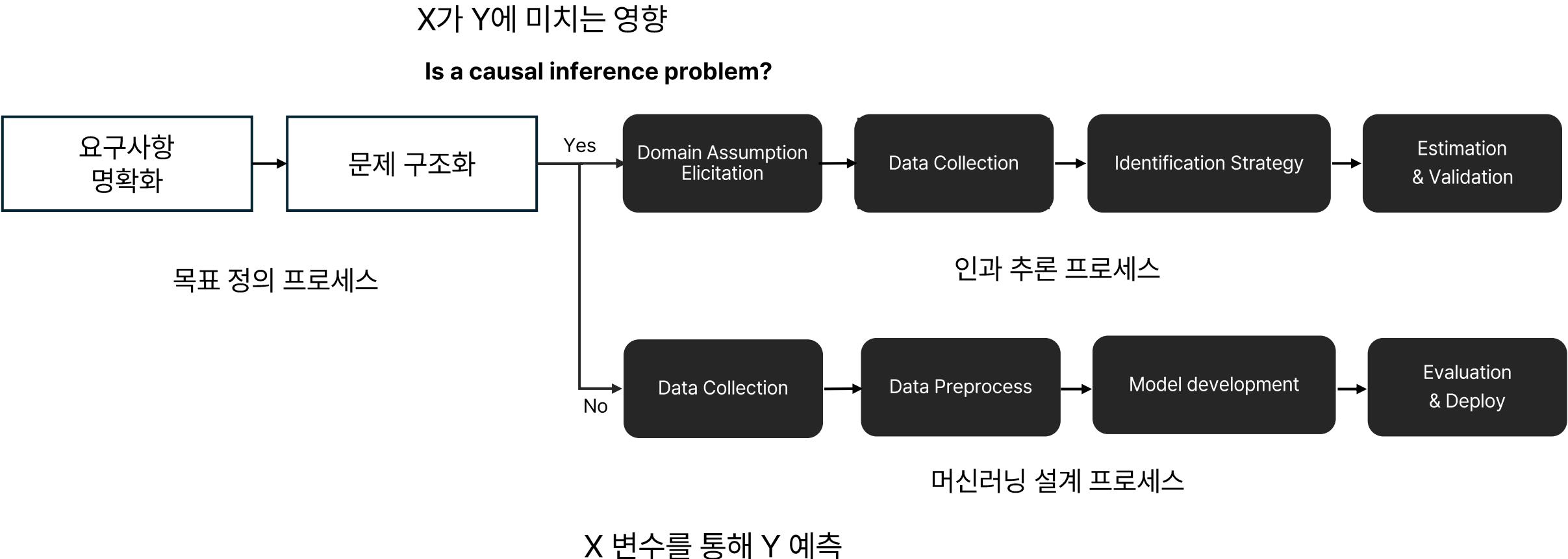
제약회사 입장에서 현업의 비즈니스 문제

Question 1 : 스타틴(Statin)을 줄이면 평균적으로 여성의 뇌졸중 위험을 줄일 수 있나요?

Question 2 : 이 환자가 내년에 뇌졸중을 겪을 확률은 정확히 어떻게 되나요?

1. (다른 변수는 고정하고) 특정 X 변수 변화에 따른 Y의 변화를 확인하고 싶어요. -> X가 Y에 미치는 영향
2. X 변수를 통해 Y 예측을 잘 했으면 좋겠어요 -> 정확한 Y 예측값

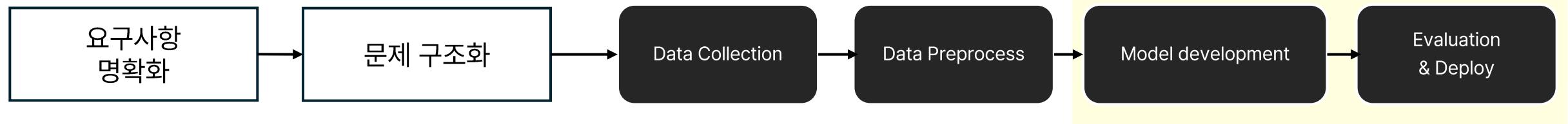
목표 정의: 데이터를 통해서 어떠한 목표를 달성하고자 하는가?



머신러닝: 일반화 성능 평가 용이

- 내년에 뇌졸중을 겪을 확률은 어떻게 되나요? 모델의 정확도는? Deploy시에 성능의 Drift가 나는가?
- 머신러닝은 일반화 성능의 검증이 간단하기 때문에 큰 성공을 거둠 (aka. Train-Valid-Test 매커니즘)

머신러닝의 흐름



문제 정의 프로세스

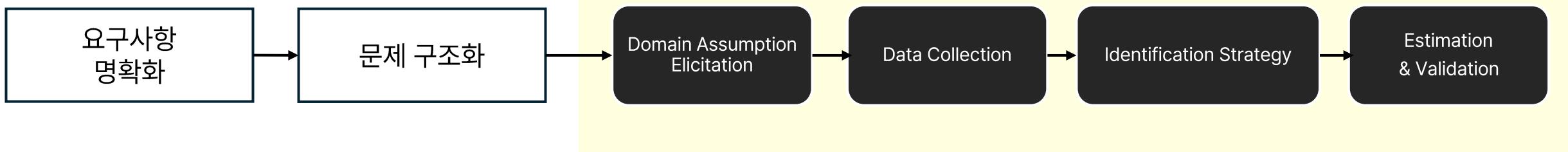
머신러닝 시스템 설계 프로세스
Development -> Evaluation

인과추론: 평가와 일반화가 어려움

- 스타틴을 줄이면 평균적으로 여성의 뇌졸중 위험을 줄일 수 있나요?
> 네. 다른 변수를 고정했을 때, 스타틴을 1단위 올리면 뇌졸중 위험에 0.25만큼 영향을 미칩니다

(그런데.. 최대한 시스템에 영향을 미치는 Endogeneity를 제어하기는 했는데.. 혹시 모르는 불확실성이 존재하고요.
또 그게 다른 도메인과 상황에서는 똑같이 적용 안될 수도 있습니다.)

인과추론 흐름



관찰할 수 없는 숫자를 추정하는데 관심이 있기 때문에 평가가 어려움
Identification -> Estimation

인과추론: 평가와 일반화가 어려움

the validity of causal inferences depends on structural knowledge, which is usually incomplete, to supplement the information in the data.

As a consequence, no algorithm can quantify the accuracy of causal inferences from observational data

인과추론의 타당성은 데이터의 정보를 보완하기 위해 일반적으로 불완전한 구조적 지식에 의존합니다.
따라서 어떤 알고리즘도 관찰 데이터로부터 인과추론의 정확도를 정량화할 수 없습니다.

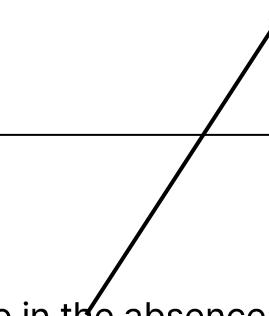
인과추론은 관찰할 수 없는 양을 추정하는데 관심

- 추론한 값을 인과로 해석할 수 있는지에 대한 설득 (빌드업이 중요)
(1) 처치 배정 매커니즘이 랜덤? (2) 처치 이전의 두 그룹이 비교 가능한가?
-> 인과적 해석을 부여하기 위한 정당성 제공
- 성능으로 증거를 제시하는 머신러닝에 익숙한 우리에게 적응이 안되는 과정
- 결과의 Robustness Check

RCT: A/A Test(Randomization Check), SRM
Quasi (DID): Parallel trends

인과적 해석을 부여하기 위한 설득

- What influences the assignment to treatment?
(Is it **as good as random**? Is it independent to outcomes?)
- How similar is the **counterfactual** to the treatment group in the absence of treatment?
→ Prove parallel trends (Key Assumption of DID)



예측 방법론과 인과추론은 다른 목적을 바탕으로 데이터를 분석하는 관점을 제공

데이터 과학: 인과추론 방법론 + 예측 방법론



다시 돌아와서, 인과추론은 예측 모델링과 추구하는 목표와 프로세스가 다른 수단이자 도구

머신러닝 시스템 개발
(지도& 비지도 학습)



다시 돌아와서, 인과추론은 예측 모델링과 추구하는 목표와 프로세스가 다른 수단이자 도구

머신러닝 시스템 개발
(지도& 비지도 학습)

인과추론



Question 1. 데이터 커리어와 인과추론?

현업의 요구사항 : 효과 추정 & Why

2년 전 나



Q. 그래서 이번에 진행한 이벤트가 효과가 있었어요?
이벤트를 Binary Feature로 넣고 Feature Importance
한 번 돌려볼게용

관점과 도구 장착:인과추론 창 +



Q. 그래서 이번에 진행한 이벤트가 효과가 있었어요?
- 인과 추론해보겠습니다 (자신감)

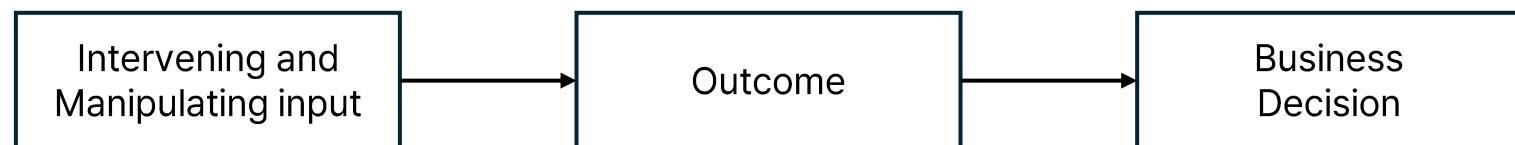
Question 1. 데이터 커리어와 인과추론?

데이터 커리어의 방향성: 현업 난제 해결과 의사 결정을 도와줄 수 있는 데이터 과학자

Define what kind of data scientist you want to be and what your focus is

인과추론은 증거를 바탕으로 설득하는 과정
비즈니스 의사결정의 과정과 유사

LAB TO LIFE :
Business Decision에 영향을 주는 데이터 실무자



Question 1. 데이터 커리어와 인과추론

Question 2. 인과추론 학습 과정

Question 3. 인과추론 팀과 앞으로의 비전

데이터 직무 여러분의 후기

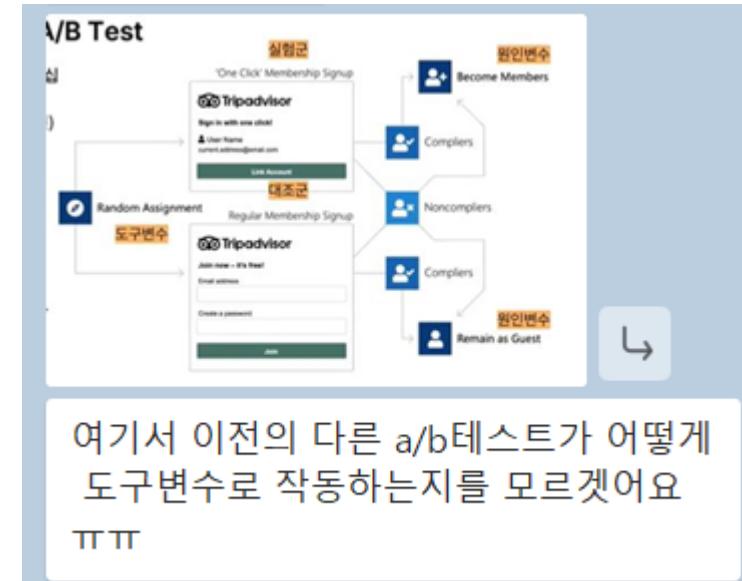
어떻게 적용하는지 또는 시작해야 하는지 모르겠다

음 저희가 스터디하고 잇긴한데

아직 기본적인 내용도 소화를 못하고 잇
긴해요ㅠ

인과추론에서

잠재적결과와 처치가 독립이라는 말이
잘 이해 안가요ㅠ



여기서 이전의 다른 a/b테스트가 어떻게
도구변수로 작동하는지를 모르겟어요
ㅠㅠ

First Step, 도구가 무엇으로 구성되었는지 구경

2022년 3월 - 8월 가짜연구소 4기

인과추론 백그라운드가 없는 이들에게 인과추론은 가까이서 보면 지옥

인과추론팀 정호X 曰: 못하겠어요. 무슨 말인지 모르겠어요.. 어떻게 적용하죠? 하..

公用 자료 DB ...				
날짜	Aa 이름	작성자	기수	작성 팀/프로젝트
	Chapter 13. Transfer learning and Transportability	홍 성철 Junyoung Kim	4기	Causal하게 Causality 이해하기
	Chapter 14. Counterfactual Mediation	Jinsoo Shin	4기	Causal하게 Causality 이해하기
	Chapter 12. Causal Discovery from Interventions	Sohee Kim 최은희	4기	Causal하게 Causality 이해하기
	Chapter 11. Causal Discovery from Observational Data	Don Don Junyoung Kim	4기	Causal하게 Causality 이해하기
	Chapter 10. Difference-in-Differences	BradyKim Minsang Namgoong	4기	Causal하게 Causality 이해하기
	Chapter 9. Instrumental Variables	Minsang Namgoong 정호재	4기	Causal하게 Causality 이해하기
	Chapter 8. Unobserved Confounding Analysis	Don Don 최은희	4기	Causal하게 Causality 이해하기
	Chapter 7. Estimation	Minsang Namgoong hyoseok BANG	4기	Causal하게 Causality 이해하기
	Chapter 5 & 6. Randomised Experiments & Nonparametric Identification	Sehoon Kim Minsang Namgoong	4기	Causal하게 Causality 이해하기
	Chapter 4. Causal Models	BradyKim 정호재	4기	Causal하게 Causality 이해하기
	Chapter 3. Graphical Models	홍 성철	4기	Causal하게 Causality 이해하기
	Chapter 2. Potential Outcomes	Jinsoo Shin	4기	Causal하게 Causality 이해하기
	Chapter 1. Motivation	Jinsoo Shin	4기	Causal하게 Causality 이해하기

First Step, 도구가 무엇으로 구성되었는지 구경

중꺾마 마인드셋:

비즈니스 의사결정에 도움을 주는 인과추론이라는 도구 구경왔습니다.



Second Step, 도구를 배웠으니, 무라도 썰어보자

2022년 8월 – 12월 가짜연구소 5기

Causal inference brave and true 번역 – Github star 377개...!

한국어 번역 일정

Causal Inference for The Brave and True에 대한 한국어 번역은 아래와 같은 일정에 따라 진행되었습니다. Part2의 경우, 저자의 부탁에 따라 추후 진행될 예정입니다.

순서	완료여부	Chapter	완료일	작성자
1	✓	1. Introduction To Causality	2022-08-20	신진수
2	✓	2. Randomised Experiments	2022-12-04	최은희
3	✓	3. Stats Review: The Most Dangerous Equation	2022-11-10	김준영
4	✓	4. Graphical Causal Models	2022-11-23	김소희
5	✓	5. The Unreasonable Effectiveness of Linear Regression	2022-11-23	남궁민상
6	✓	6. Grouped and Dummy Regression	2022-11-05	정호재
7	✓	7. Beyond Confounders	2022-11-26	김상돈
8	✓	8. Instrumental Variables	2022-12-04	최은희
9	✓	9. Non Compliance and LATE	2022-11-19	김성수
10	✓	10. Matching	2022-11-04	김상돈
11	✓	11. Propensity Score	2022-11-11	김성수
12	✓	12. Doubly Robust Estimation	2022-10-28	홍성철
13	✓	13. Difference-in-Differences	2022-08-27	신진수
14	✓	14. Panel Data and Fixed Effects	2022-11-26	신진수
15	✓	15. Synthetic Control	2022-09-03	정호재
16	✓	16. Regression Discontinuity Design	2022-11-22	남궁민상
21	✓	21. Meta Learners	2023-12-31	박병수



가짜연구소 Causal Inference Team
안녕하세요. 가짜연구소 Causal Inference 팀입니다!

84 followers Korea, South https://causalinferencelab.github.io

Pinned

 [Causal-Inference-with-Python](#) Public

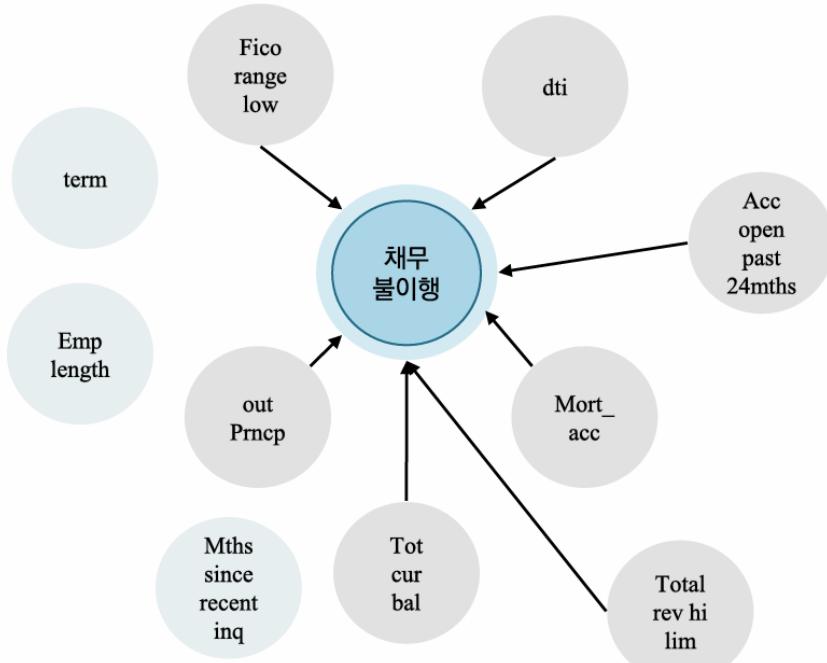
Causal Inference for The Brave and True 책의 한국어 번역 자료입니다.

Jupyter Notebook ★ 377 ⚡ 68

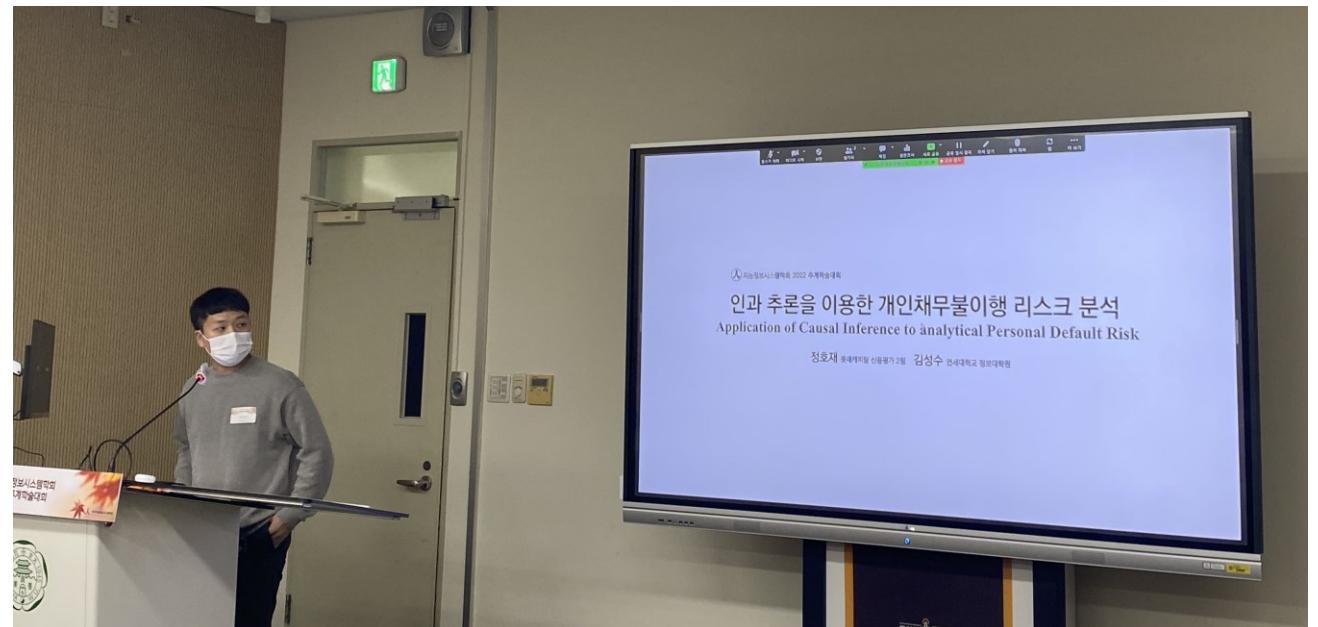
Second Step, 도구를 배웠으니, 무라도 썰어보자

2022년 8월 – 12월 가짜연구소 5기

인과추론 금융 도메인에 적용해보기 – 개인채무불이행 리스크 분석



채무불이행에 영향을 미치는 인과 방향 by Notears



2022 지능정보시스템 추계학술대회

Third Step, 고명을 얹기

2023년 3월 - 8월 가짜연구소 6기

인과추론으로 Top conference 페이퍼 제출과 Reject

인과추론은 설득과 반박의 과정 – Endogeneity

Overall Rating 2, 3, 4 -> Reject

Both reviewers have described **issues concerning endogeneity** in your analysis. To summarize, it is very important to rule out **potential endogeneity** because properties which receive the label are surely very different from properties that do not receive the label. I encourage you to explore other approaches, **beyond PSM, to address endogeneity concerns**. In addition, **the current sample size** is probably....

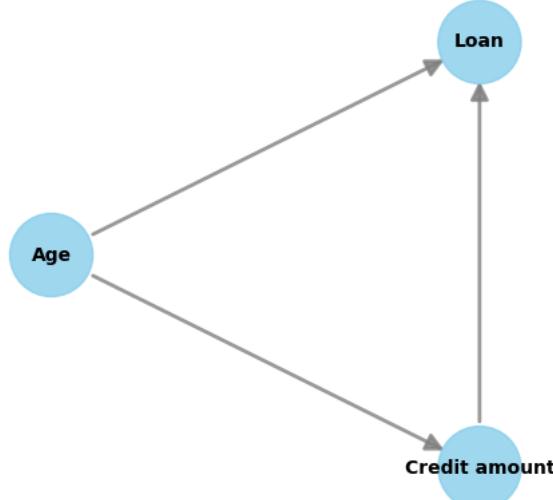
One major concern I noticed in the current study is the issue of endogeneity. Considering the presence of **selection biases**, the paper adopts a matching approach, which, however, **may not be sufficient to address the endogenous concerns....**



Third Step, 고명을 얹기

2023년 8월 - 12월 가짜연구소 7기

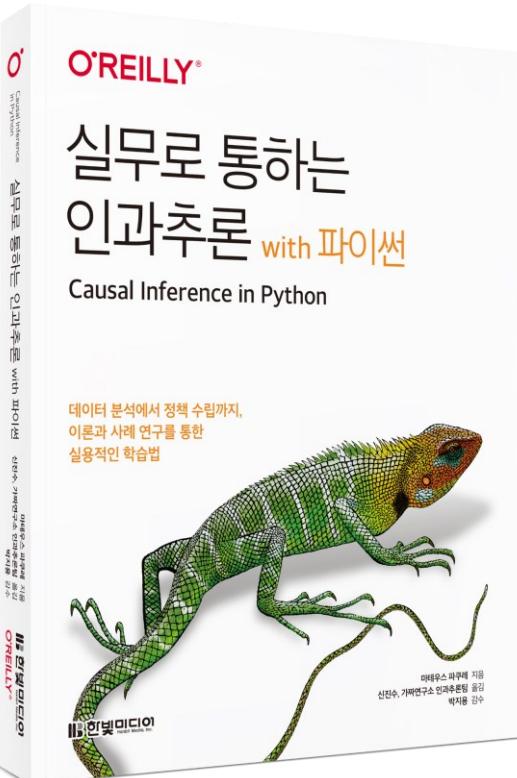
인과추론으로 머신러닝 알고리즘 Fairness 연구



Third Step, 고명을 얹기

2023년 8월 - 12월 가짜연구소 7기

진수님의 번역서 작업에 약간의 기여와 최종 TF (책 정말 좋습니다 – 홍보 맞음)



**9장

344p(위에서 3번째 줄): 예를 들어, 온라인 마케팅은 고객의 행동을 더 잘 추적할 수 있게 해주며, 더 나은 어트리뷰션(기여도)attribution이 가능합니다.

-> 예를 들어, 온라인 마케팅은 고객의 행동을 더 잘 추적할 수 있게 해주며, 더 나은 어트리뷰션 측정이 가능합니다.

346p(밑에서 8번째 줄): 이제 사용할 패널데이터 분석에 사용할 몇가지 표기법을 복습해보겠습니다.

"사용할 반복" -> 이제 패널데이터 분석에 사용할

347p(9.2 행렬 표현 바로 윗 줄) 보았듯이, 가상의 통제집단이 필요한 이유가 바로 여기에 있습니다. 이는 과거 결과를 사용하여 (조건부가 아닌) $E[Y_0 | D = 1, Post = 1]$ 을 추정하는 아주 현명한 방법이죠.

-> 1. "보았듯이 삭제"

347p(9.2 행렬 표현) 이 행렬은 가상의 통제집단에 직접적으로 적용되므로 다시 살펴보겠습니다. 이 행렬에서 행은 기간이고 열이 도시(실험 대상)라고 가정해보겠습니다.

"이 반복" 이러한 행렬 표현은 가상의 통제집단에 직접적으로 적용되므로 다시 살펴보겠습니다.

349p (위에서 첫번째 줄): 작업 중 무엇을 다루고 있는지 잊었다면 이 함수로 돌아오세요.

-> 삭제

350p (위에서 2번째 줄) 그리고 $E(Y_0 | D=1, Post=1)$ 을 추정하고

"윗 단락과 그리고 반복" -> 이때, $E(Y_0 | D=1, Post=1)$ 을 추정하고

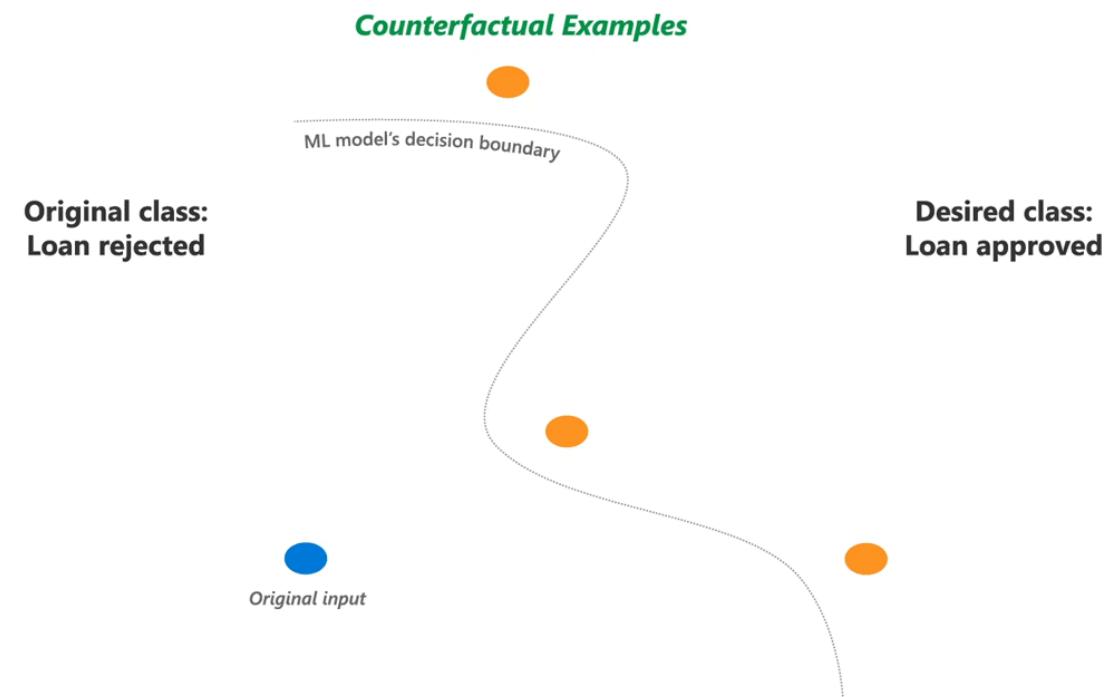
Third Step, 고명을 얹기

2023년 8월 - 12월 가짜연구소 7기

Counterfactual Explanation 시스템 개발

- 머신러닝 모델의 결정을 변경하는데 필요한 최소한의 데이터의 변화를 설명

Example: “설비 온도가 10도 더 높았다면 이상 알람이 울리지 않았을 거에요” – What if?



Summary

**목표: 인과추론 그 자체?
NO : 문제 해결과 의사 결정을 위한 적용**

1. 표준적이고 잘 정립된 방법론을 활용
2. 필요한 부분을 습득하고 적용

Summary

인과추론 항해기

인과추론 도구 체험
:인과추론 개념 학습

인과추론 도구 도메인 적용
:Causal Discovery +
Brave and True 번역

인과추론으로 도전
TOP Conference + 번역서
Causal Fairness

Industry 커리어 시작
:Counterfactual
Explanation

2022.02

인과추론 팀 4기 합류

2022.08

인과추론 팀 5기

2023.02

인과추론 팀 6기

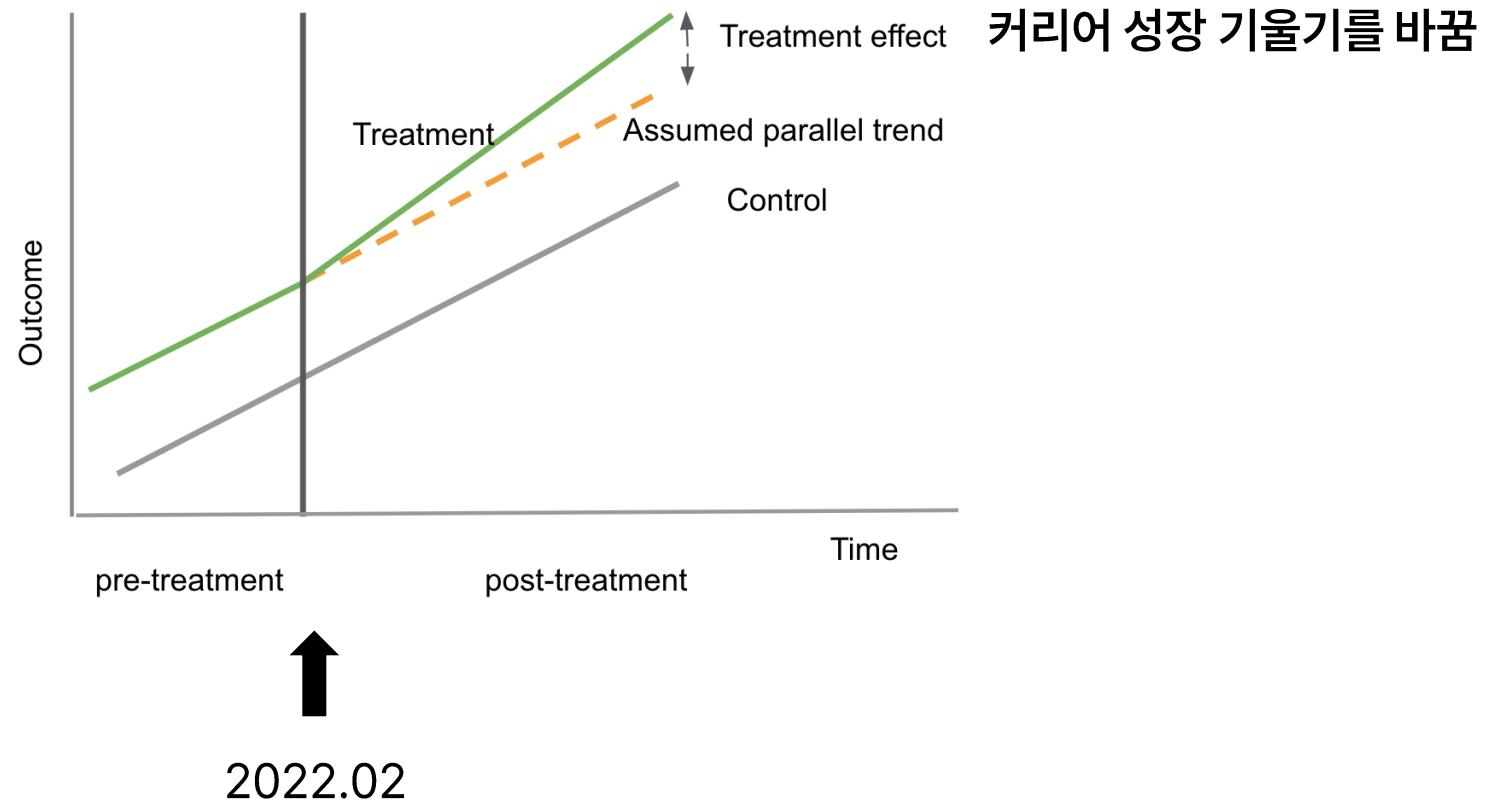
2023.08

인과추론 팀 7기



Summary

Treatment: 인과추론팀 합류



Question 1. 데이터 커리어와 인과추론

Question 2. 인과추론 학습 과정

Question 3. 인과추론 팀과 앞으로의 비전

2024년 4월 – 8월 가짜연구소 8기

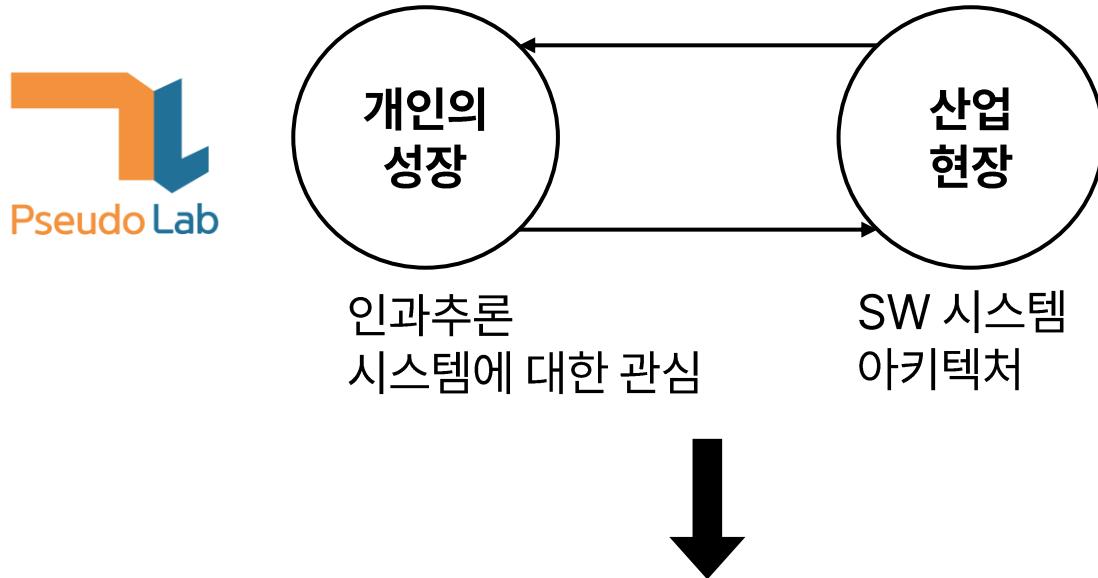
1. 가짜연구소 인과추론팀 빌더

문제 상황에 따른 인과추론 활용 사례 템플릿화 (일부)

Table. 인과추론 실무 사례 템플릿화

No	Criteria	Method	Case
1	(RCT) 그룹 간의 사전에 존재한 차이 O	A/A Testing, SRM(Sample Ratio Mismatch)	TBD
2	(RCT) 민감도를 높이기 위한 방법	CUPED / Triggering	TBD
3	(RCT) 처치 그룹 중에 Non-Complier 발생	Compiler Average Causal Effect (CACE)	Week 2
4	(RCT) 처치 효과가 하위 그룹에 따라 다르게 나타나는가?	Quantile Testing	TBD
5	(RCT) Interference가 존재하는가?	Network Clustering	TBD
6	(Quasi) 처치 전후의 시계열 데이터는 있지만 대조군은 없음	Interrupted Time-Series Analysis	TBD
7	(Quasi) 외생적 충격/사건 + 평행 추세 가정 O	Difference-in-Differences (DID)	TBD
8	(Quasi) 내생적 충격/사건 + 평행 추세 가정 O	Matching + DID	TBD
9	(Quasi) 처치에 임의의 기준점 O	Regression Discontinuity	TBD
10	가정에 맞는 도구변수가 존재 - 처치가 Continuous	Two-Stage Least Squares or Selection Bias Correction Method	TBD

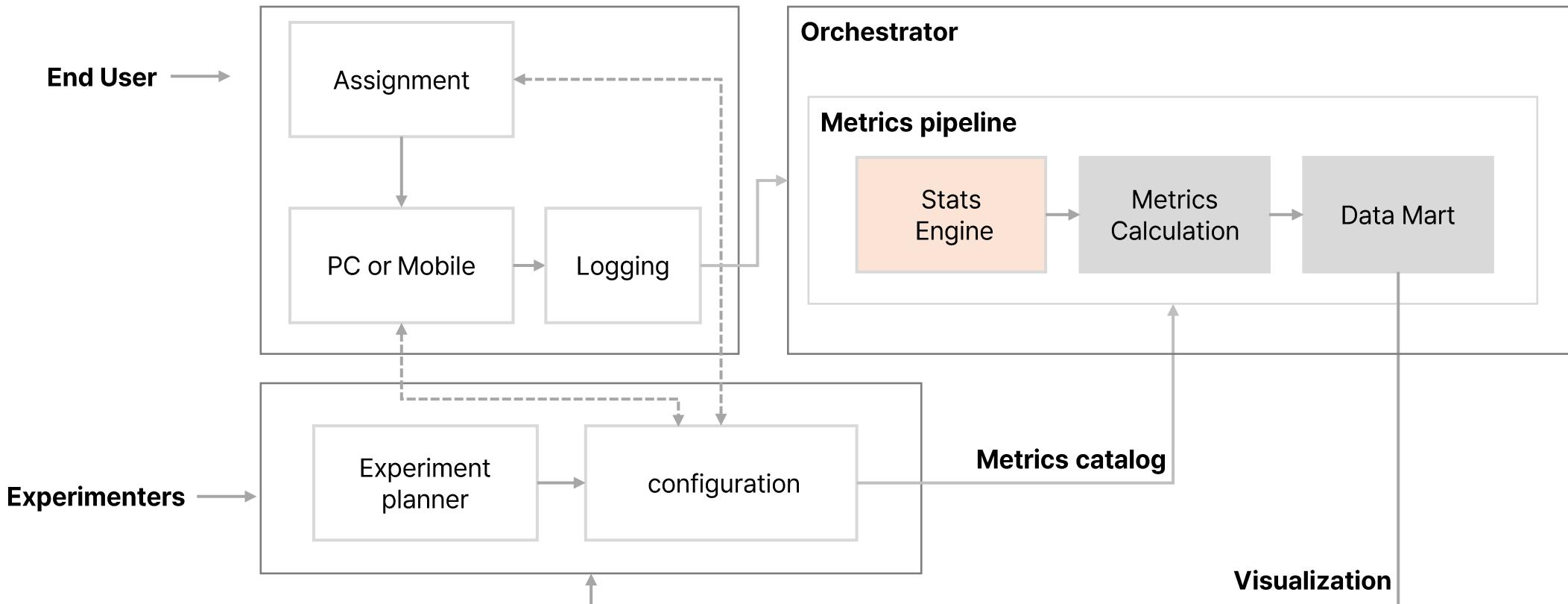
2. Toward Automated Causal inference



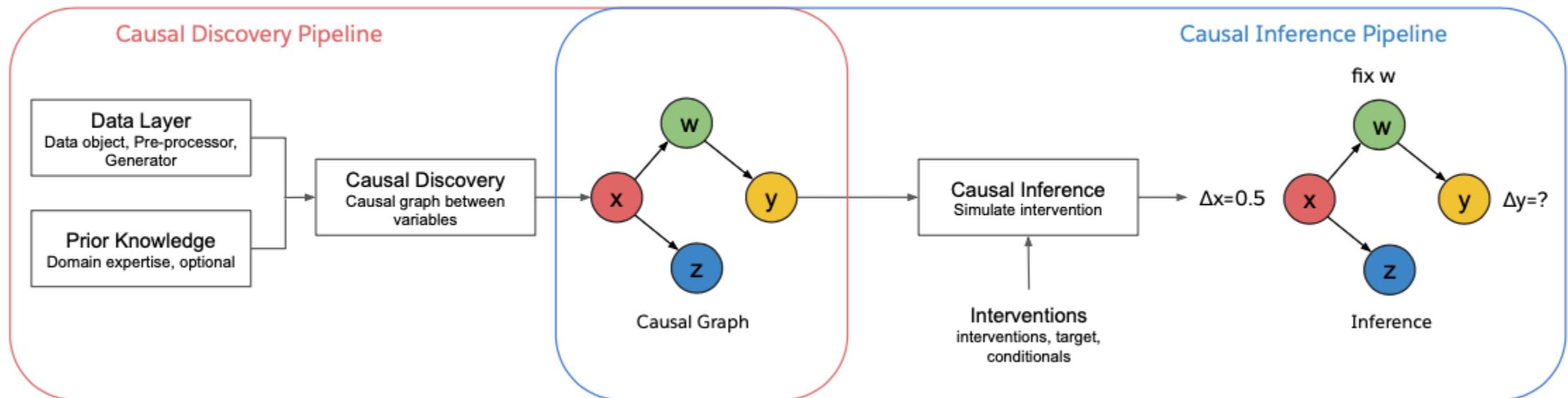
현업이나 분석가가 인과추론을 쉽게 할 수 있는 플랫폼
1. **실험 플랫폼(Experiment Platform)**
2. **CausalOps**

2. Toward Automated Causal inference 실험 플랫폼(Experiment Platform)

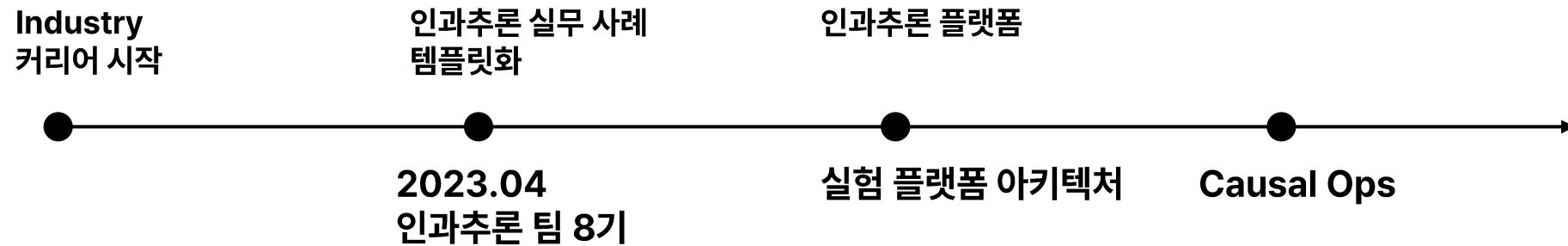
- 할당 그리고 Config와 연계된 화면과 로깅 시스템
- 실험 관리와 Config 서버 - 실험 정보와 실험 정책 저장
- 지표 계산 및 대시보드 결과 생성



2. Toward Automated Causal inference CausalOps



현업 난제 해결과 의사 결정을 도와줄 수 있는 데이터 과학자



저에게 인과추론팀은

커리어 방향성

취업

좋은 팀원

길잡이

성장곡선을 바꾼 처치

아까 그 질문 여러분은 어떤 배를 타고 싶으신가요?

#막막 #방향성 없음 '표류'



#명확 #전략 '항해'





Let's Join 가짜연구소 & 인과추론팀

감사합니다





Causal - Lab

인과추론으로 네트워킹 하기

Email: bradykim@gmail.com

Linkedin:



김성수

Data Scientist at LG CNS

