統計的因果探索:
領域知識とデータから因果仮説を探索する

清水昌平

理研AIP 因果推論チーム

統計的因果探索とは

・ データを用いて因果グラフを推測するための方法論

仮定

- 関数形
- 分布
- 未観測共通原因の有無
- 非巡回 or 巡回 など

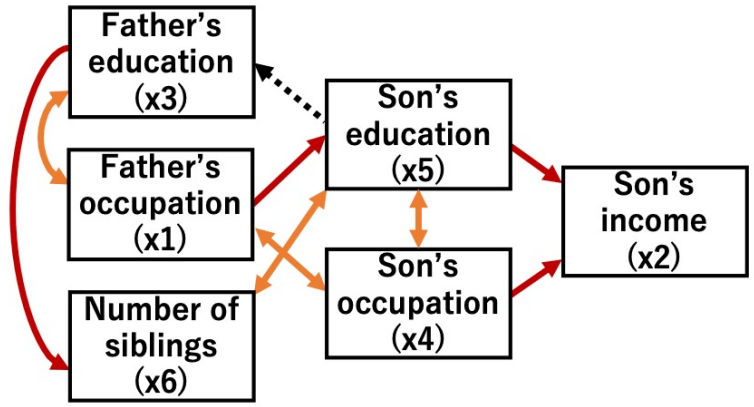
＋

データ

x1	x2	x3	x4	x5	x6
87.9	45433	17	76.3	17	1
87.9	55071	16	86	18	2
62.1	113159	16	87.9	16	0
78.5	30289	16	30.1	14	4
32.3	113159	20	63.5	20	7
60.6	55071	17	83.7	17	1
76.4	55071	16	78	14	2
63.5	37173	12	63.2	16	3
63.2	113159	14	86.5	17	1
36.5	37173	12	83.7	12	4

推測

因果グラフ

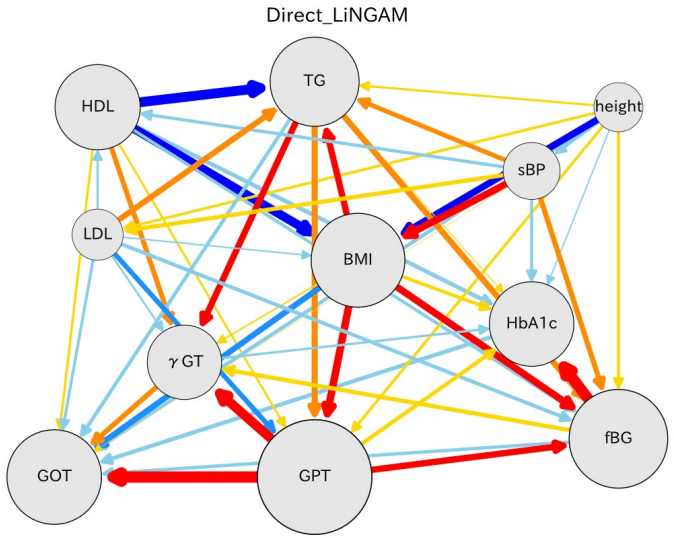


Maeda and Shimizu (2020)

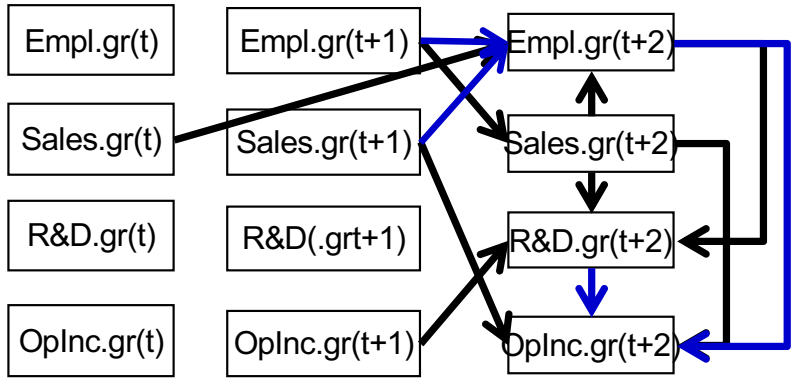
因果探索の適用例: ターゲットの原因候補の探索

https://www.shimizulab.org/lingam/lingampapers/applications-and-tailor-made-methods

- 生命科学 (Maathuis et al., 2010)
- 医学 (Kotoku et al., 2020)
- 化学 (Campomanes et al., 2014)
- 材料 (Nelson et al., 2021)
- 気候学 (Liu et al., 2020)
- 経済学 (Moneta et al., 2013)
- 心理学 (von Eye et al., 2012)
- 政策 (高山ら, 2021)
- ネットワークデータ (Jarry et al., 2021)



Kotoku et al. (2020)

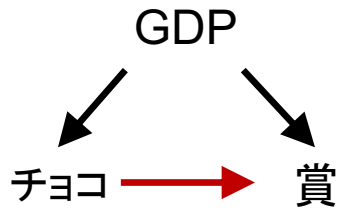


Moneta et al. (2013)

統計的因果推論では因果グラフが要(かなめ)

- データから介入効果を推定
  - チョコ消費量を変えると
    ノーベル賞受賞者の数はどのくらい増えるのか(減るのか)
  - 機械学習のする予測
    - チョコ消費がこのくらいならノーベル賞数このくらい？
    - ノーベル賞数がこのくらいならチョコ消費このくらい？

- 介入効果を「正しく」推定するには
  因果グラフが必要 (e.g., バックドア基準)



E(賞 | do(チョコ))=E_{GDP} [E(賞 | チョコ, GDP)]

Messerli, (2012), New England Journal of Medicine

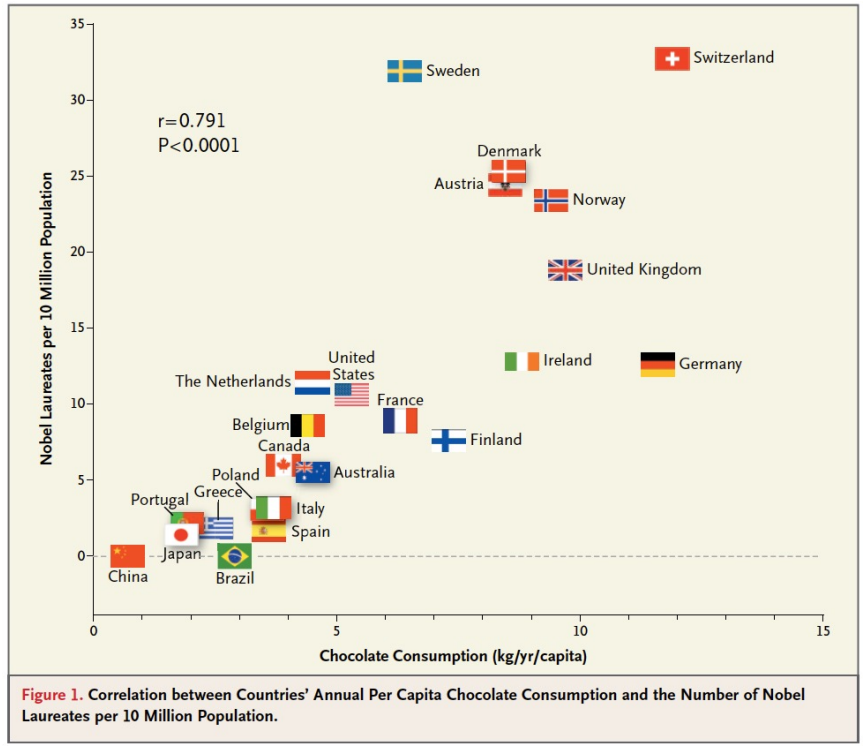


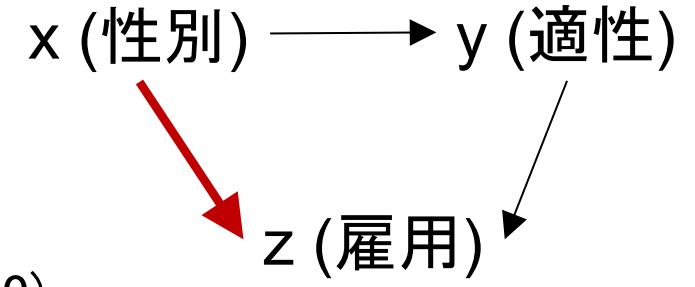
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

ノーベル賞受賞者の数

チョコレート消費量

AIのための因果推論でも要（かなめ）

- 公平性 (Kusner et al., 2017)
- 説明性: 原因の確率 (Galhotra et al., 2021)
- 予測メカニズム解析 (Blobaum et al., 2017; Sani et al., 2020)
- 個体レベルの最適介入 (Kiritoshi et al., 2021)
- 転移学習 (Zhang et al., 2013; Zhang et al., 2020; Bareinboim et al., 2016)
- 科学的知識の取り込み (Teshima et al., 2021)



(Pearl, 2001)

- さまざまな因果に関するクエリー(介入効果等) に答えられるか を判定するために因果グラフが必要

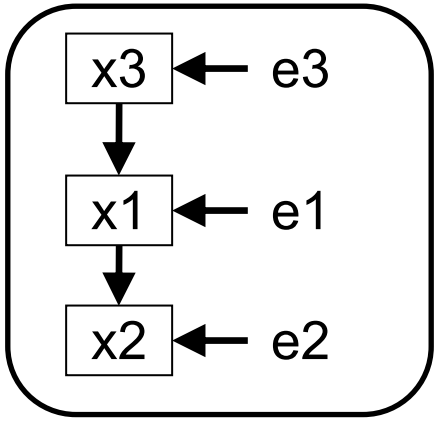
統計的因果探索の方法

フレームワーク

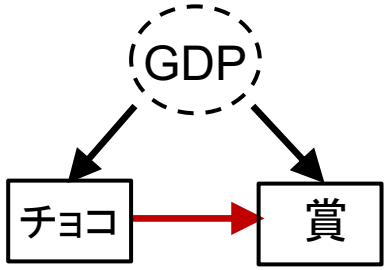
- 構造的因果モデル (Pearl, 2001)

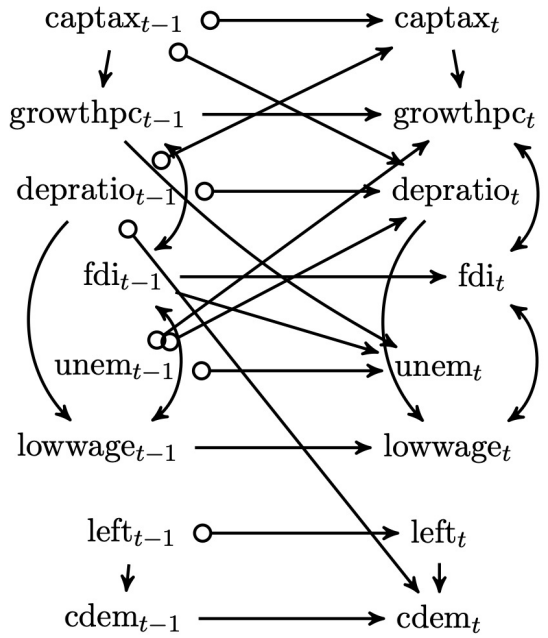
$$x_i = f_i(x_i\text{の親}, e_i)$$

誤差変数



- 因果モデルに仮定をおき、
  その中でデータとつじつまの合うモデルを探す
  - 典型例:
    - 非巡回有向グラフ
    - 潜在共通原因なし(すべて観測されている)
      or 潜在共通原因あり





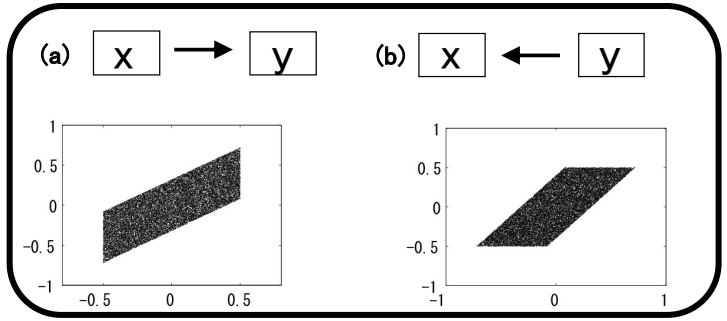
時系列モデル: Malinsky and Spirtes (2018)

2つのアプローチを使い分け

- 関数形や分布に仮定を“おかない”アプローチ (Spirtes et al., 1993)
  - 条件付き独立性
  - 同値類

これ以上は区別できない

- 関数形や分布に何らかの仮定をおくアプローチ (Shimizu et al., 2006)
  - 例えば、線形性+非ガウス連続分布: *LiNGAM*
    一意に識別可能 (or より小さい同値類)

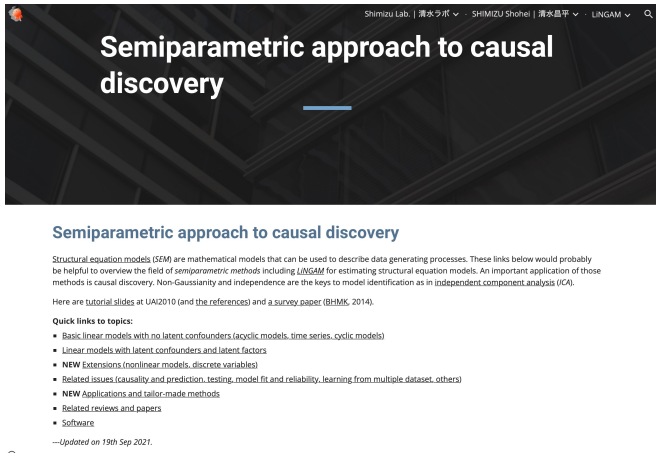


区別できる

DirectLiNGAMアルゴリズム (Shimizu et al., 2011)

- 潜在共通原因なし（すべて観測されている）
- 回帰分析と独立性の評価を繰り返す
- Guaranteed to converge in finite steps（変数の数）

- p>nの場合への拡張 (Wang & Drton, 2020)
- 並列化+GPUで高速化 (Shahbazinia et al., 2021)
- 数百から数千変数くらい



関連論文: https://www.shimizulab.org/lingam

他の識別可能なモデル

- 非線形 +“加法”誤差 (Hoyer et al., 2008; Zhang et al., 2009; Peters et al., 2014)
  - $x_i = f_i(\mathrm{par}(x_i)) + e_i$
  - $x_i = g_i^{-1}(f_i(\mathrm{par}(x_i)) + e_i)$

- 離散: ポワソンDAGモデルと拡張 (Park+18JMLR)

- 離散と連続の混在: LiNGAM + ロジスティック型モデル
  (Wei et al. 2018; Li & Shimizu, 2018)

$$x_i = \begin{cases} 1, & e_i + c_i + \sum_{j \in \mathrm{pa}(i)} b_{ij} x_j > 0 \\ 0, & \text{otherwise} \end{cases}, \quad e_i \sim Logistic(0,1)$$

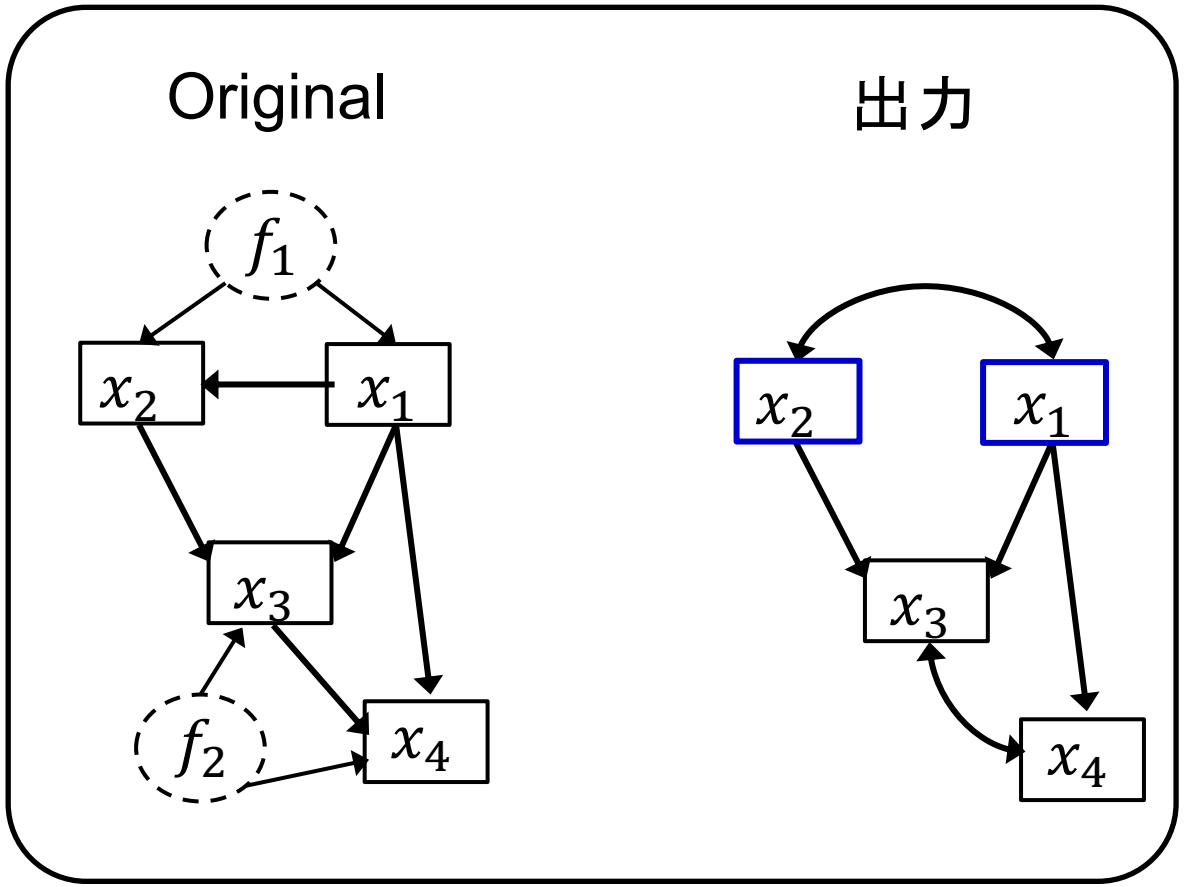
- 時系列モデル (Hyvarinen et al, 2010)

- 巡回モデル(Lacerda et al., 2008)は識別可能でない場合も

潜在共通原因ありの場合

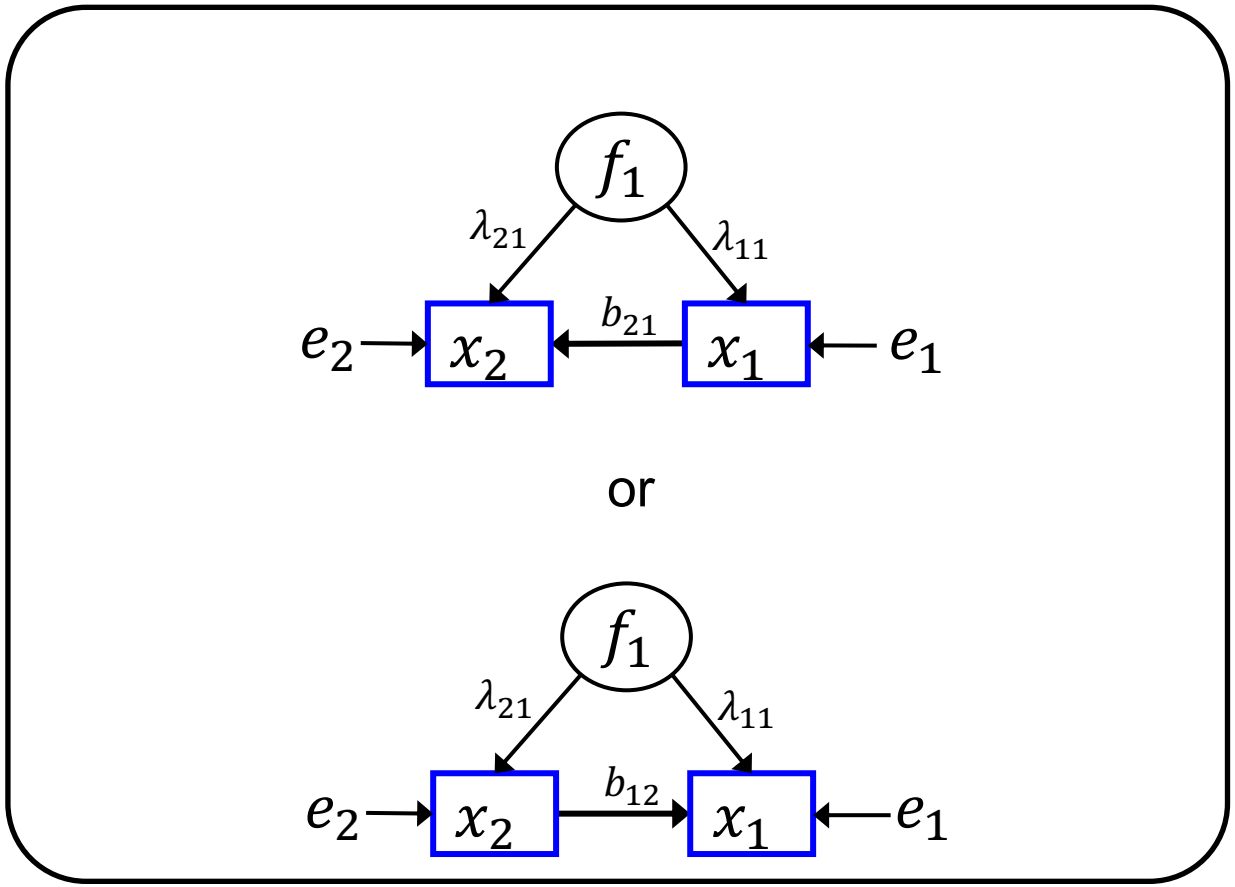
潜在共通原因のあるペアがどれか

(Maeda & Shimizu, 2020)



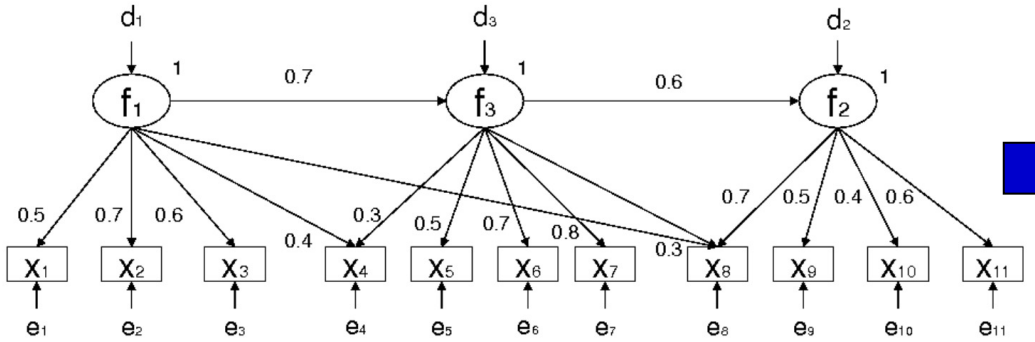
潜在共通原因のあるペアの間を推測

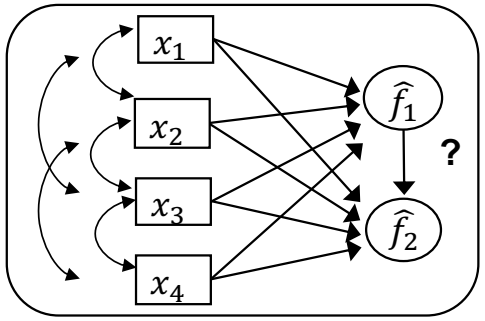
(Hoyer et al. 2008; Salehkaleybar et al., 2020)

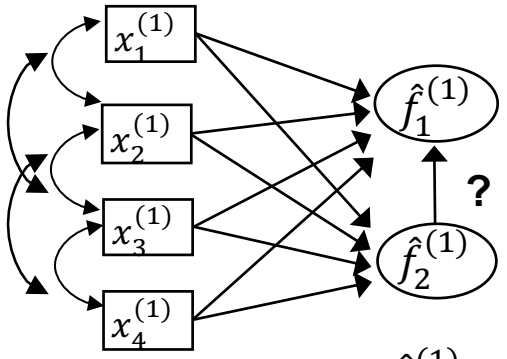


それ以外の潜在変数

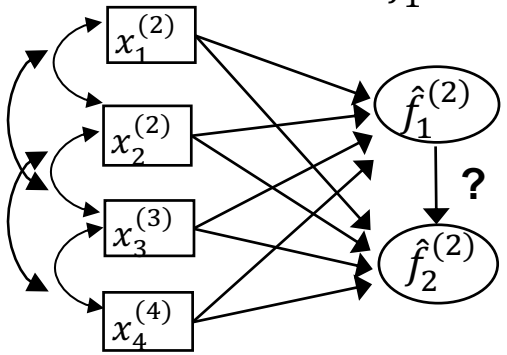
- 潜在因子 (Shimizu et al., 2007)
  - Causal representationと呼ばれることも (Adams et al., 2021)
  - 因果グラフは不変な特徴という主張 (Schölkopf et al., 2021)







$\hat{f}_1^{(1)} = \hat{f}_2^{(2)}$?



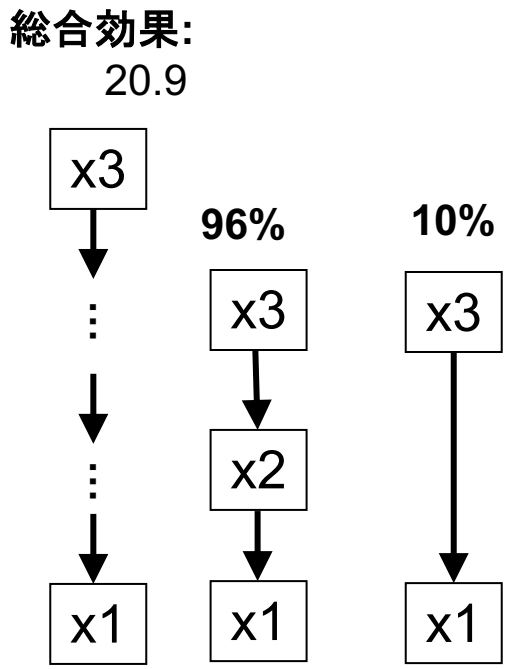
複数データセットの情報統合
複数データセットからの特徴抽出と
潜在因子の因果グラフ推測を同時に (Zeng et al., 2021)

推測された因果グラフを評価

統計的信頼性評価

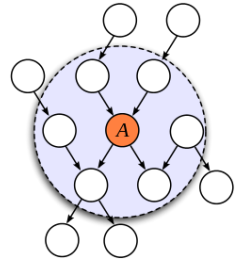
- 有向道や有向辺のブートストラップ確率
  - 例えば、閾値0.05を越えるものを解釈
  - LiNGAM Python package

	from	to	effect	probability
0	x3	x0	3.006190	1.00
1	x0	x1	3.004868	1.00
2	x2	x1	2.092102	1.00
3	x3	x1	20.931938	1.00
4	x0	x5	3.982892	1.00
5	x3	x5	12.024250	1.00
6	x2	x4	-0.887620	1.00
7	x3	x4	18.077244	1.00
8	x0	x4	7.993145	0.98
9	x3	x2	5.970163	0.96
10	x5	x1	0.011708	0.79
11	x2	x5	0.024284	0.72



モデル仮定の評価（崩れの検出）

- 誤差(残差)の独立性評価
  - 例えば、HSIC (Gretton et al., 2005)
- マルコフ境界による予測の良さで評価
  (Biza et al., 2020)
- 複数のデータセットでの結果を比較
- 領域知識による評価

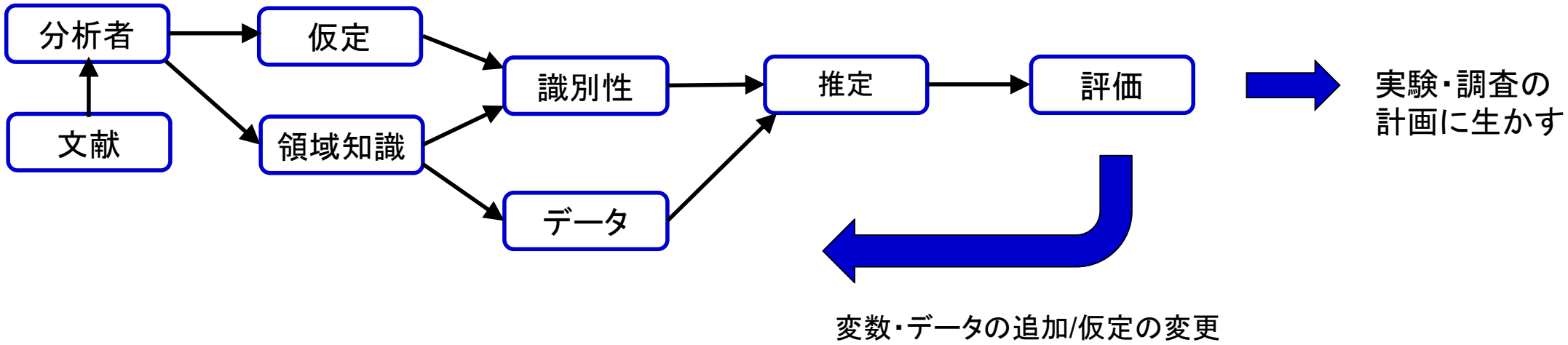


Wikipediaより

まとめ

まとめ: 因果推論「も」するAI

- 統計的因果推論: リサーチクエスチョンは予測だけではない
  - 仮定+データ+クエリー -> 回答 (ができれば)
  - フレームワークと識別性を重視. 推定の技術は機械学習と共通



- 因果探索ソフトウェア
  - GUI: TETRAD; Python: lingam, causal-learn; R: pcalg など
  - 商用: Causal analysis (NEC); Node-AI (NTTコミュニケーションズ)など