

仮想人体生成モデルと ライフケアの民主化

丸山宏

Twitter; @maruyama

自己紹介:

- 1983 - 2009 日本IBM東京基礎研究所
 - 1996-1997 米国インターネット事業部
 - 1997-2000 東京工業大学客員助教授
 - 2004-2005 IBMビジネス・コンサルティング・サービス



- 2009 - 2010 キヤノン株式会社 デジタルプラットフォーム開発本部

- 2011 - 2016 統計数理研究所



- 2016 - Preferred Networks
 - 2020 - 東京大学 人工物工学研究センター
 - 2020 - 花王株式会社 エグゼクティブフェロー



今ココ

黒字: 研究開発
青字: アカデミア
緑字: ビジネス



花王のライフケア・テクノロジー例： 皮脂RNAモニタリング



<https://beautytech.jp/n/n6aa590a0767b>

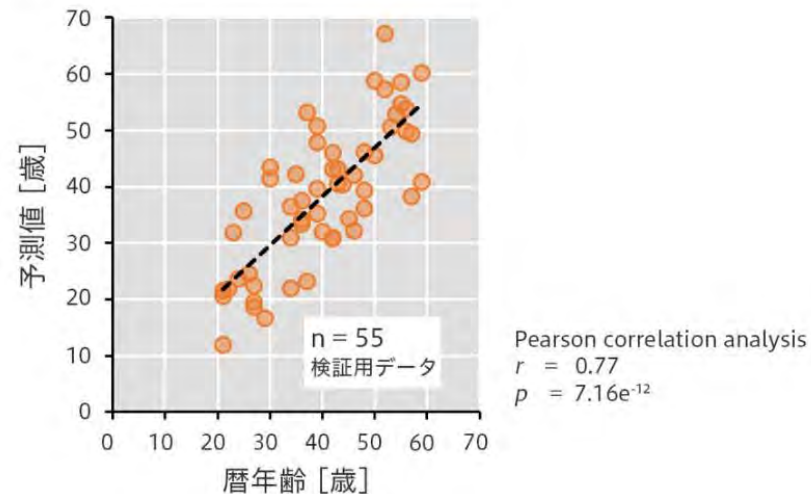


図1 暦年齢と機械学習モデルから算出された予測値の相関

<https://www.kao.com/jp/corporate/news/rd/2021/20210816-001/>

蚊の忌避剤



*1 Ikura, H., Takizawa, H., Ozawa, S., Nakagawa, T., Matsui, Y., Nambu, H. Mosquito repellence induced by tarsal contact with hydrophobic liquids. *Sci. Rep.* 10, 14480 (2020).

<https://doi.org/10.1038/s41598-020-71406-y>

花王の仮想人体生成モデル

イノベーションのレシピ

花王の新たな“道しるべ”となるか、PFNと仮想人体生成モデルを共同開発

(1/2 ページ)

2022年03月01日 07時30分 公開

【林尚洙, MONOIS



印刷する



クリップする



通知する



10



Share



0

花王とPreferred Networks（以下、PFN）は2022年2月28日、オンラインで会見を開き、健康診断などで得られる身体データから、食事、運動、睡眠などのライフスタイル、性格傾向、嗜好性、ストレス状態など1600以上の項目を網羅し、ある項目のデータを入力すると別の項目の推定データを出力できる「仮想人体生成モデル」を開発したと発表した。花王は、PFNの協力を得て、仮想人体生成モデルをAPI経由で提供する新規デジタルプラットフォーム事業の開始に向けた準備を進める方針で、最初の協業パートナーとしてNTTドコモとヘルスケア領域で連携していくことを明らかにした。

花王 代表取締役 社長執行役員の長谷部佳宏氏は「花王が手掛ける生活用品は、コンパクトな消費でありながらQOL（Quality of Life）を高められることが重要になっている。これを実現していく上で役立つ仮想人体生成モデルは、多種多様な企業との連携が可能であり、当社が構想するデジタル・ライフ・プラットフォームにおいて重要な役割を果たすものだ」と語る。



会見の登壇者。左から、花王の長谷部佳宏氏、PFNの岡野原太輔氏、丸山宏氏（クリックで拡大） 出所：花王、PFN

2022年2月28日、花王とPFNは「仮想人体生成モデル」を発表



モノ売りでないビジネスモデルとは？

例えば、こんなライフケアアプリを作りたい「明るい引退生活」

ペルソナ：

- 60歳台男性、数年後に引退を迎える
- 今は健康、生活習慣病に若干の不安あり（血圧、糖尿など）
- 引退後に、少なくとも10年程度は健康な生活を送りたい
- でも、シリアスな努力は嫌だ

定例血液検査

健康食品サブスク



孫からの声援



運動記録

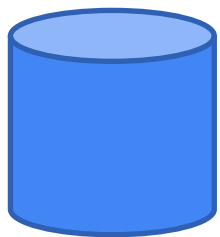


RNA検査



イノベータを阻む2つの壁

ライフケアデータ



ソリューションアイデアを持つイノベータ



最新の計測・介入技術



データの壁

- データ設計
- データ収集コスト
- 個人情報保護法, ...

技術の壁

- 多様な最新技術
- 個別の交渉, 知財, ライセンス料, ...

目指す姿：ライフケアの民主化



仮想人体生成モデル
様々な部品をデータで結びつける「糊」としての存在



API基盤上で提供される汎用の部品たち（T&Cが公開されている）

計測技術



データ連携



介入サービス



様々な既存ヘルスケアサービスへのコネクタ



医療サービスへの接続



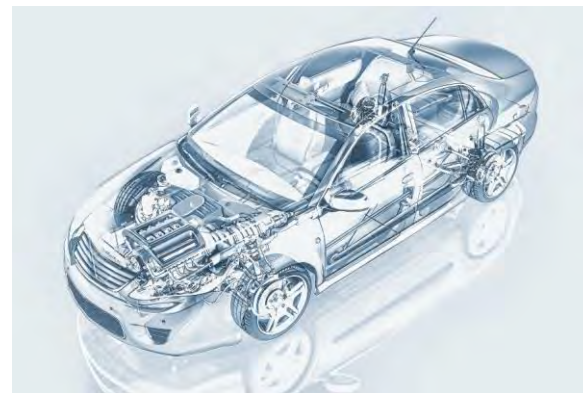
人のカラダの不思議 – 世界は分けてもわからない*

人体



様々な細胞が集まって、総体として機能を発現する

工学的システム



それぞれのパーツが与えられた機能を果たす



* 福岡伸一著「世界は分けてもわからない」
ISBN : 4062880008

総体としての発現をどのように表現するか ⇒ 仮想人体生成モデル

体格・体組成

ストレス／疲労傾向

血液検査値

年齢、性別

性格特性

精神的健康状態／抑うつ傾向

運動機能

血管機能

睡眠状況

生産性

認知機能

皮膚状態

幸福感

食事／栄養素

免疫指標

毛髪状態

東洋医学の体質分類

手指バリア機能

糖代謝機能

薄毛／脱毛状態

体臭

頭皮菌叢

腸内菌叢

皮脂RNA

月経関連

更年期障害

尿失禁/過活動膀胱傾向



生成モデルとは「現実世界を写し取った統計的パターン」

画像の場合

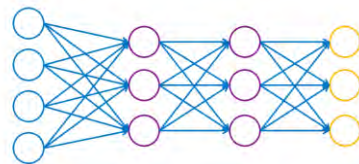
実在の人物（訓練データ）



訓練

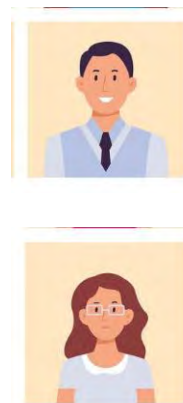


生成モデル



実在の人物のパターンで訓練
ただし、個别人物の情報は含
まない

生成された仮想的な人物



人体における生成モデル

画像の場合の生成モデル

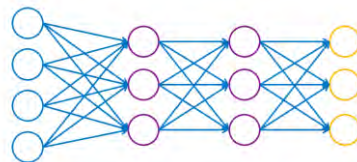
実在の人物画像（訓練データ）



訓練



生成モデル



実在の人物のパターンで訓練
ただし、個別人物の情報は含まない

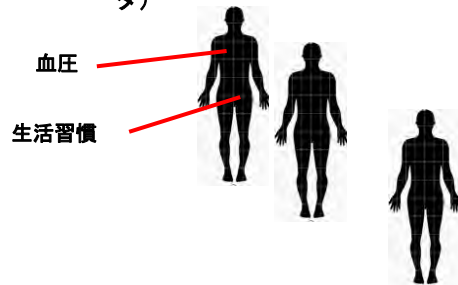


生成された仮想的な人物画像



仮想人体生成モデル

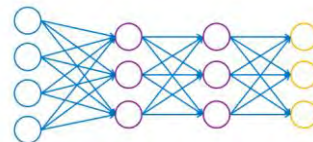
実在の人物（訓練データ）



訓練



仮想人体生成モデル



個別の人物の情報は含まない



生成された仮想的な人体



同時確率分布としての仮想人体生成モデル



$P(X)$: 属性 X を持つ人の同時確率分布



周辺分布を取れば、年齢、身体情報、生活習慣などの分布が得られる

観測された属性 $x_1, x_2, \dots, x_n$: x_i は性別・年齢・生活習慣など入力値

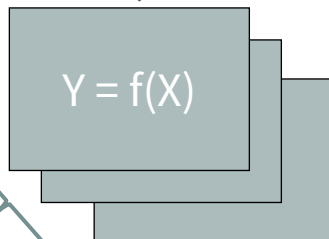
未知の属性 y_j の条件付き分布 $P(y_j | x_1, x_2, \dots, x_n)$



仮想人体生成モデルの汎用性 ある属性値を、他の属性値から推定するには？

通常の機械学習モデルの場合
(10属性の入力から出力を推定する)

$x_1, x_2, \dots, x_{10}$



y

${}_{1600}C_{10} \simeq 3 \times 10^{25}$ 通りの
モデルが必要

仮想人体生成モデル (同時確率分布)
の場合

$x_1, x_2, \dots, x_n$



$y_1, y_2, \dots, y_m$

どんな入出力の組み
合わせに対してもモ
デルは1つ

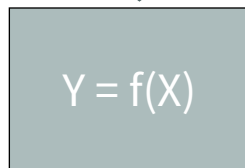
仮想人体生成モデルは汎用の機械学習モデルとみなせる！

事後分布としての出力

出力の不確実性をどのように解釈するか？

通常の機械学習モデルの場合

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{10}$



y

出力は1つの値

当たり!
ハズレ!

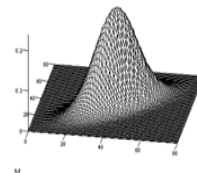


仮想人体生成モデル（同時確率分布）
の場合

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$



出力は確率分布



$P(y_1, y_2, \dots, y_m)$

リスク
情報量



仮想人体生成モデルの訓練データ設計

必要なサンプル数と項目数を、複数の試験／データでカバーする階層的なデザイン



個人情報 は 訓練に必要か：複数のデータソースの結合

	個人ID	収縮期血圧	血糖値	...
	12345	140mmHg	120mg/dl	...

	個人ID	運動習慣	飲酒習慣	...
	12345	なし	1日3合以内	...

個人IDをキーにして結合



	個人ID	収縮期血圧	血糖値	運動習慣	飲酒習慣	...
	12345	140mmHg	120mg/dl	なし	1日3合以内	...

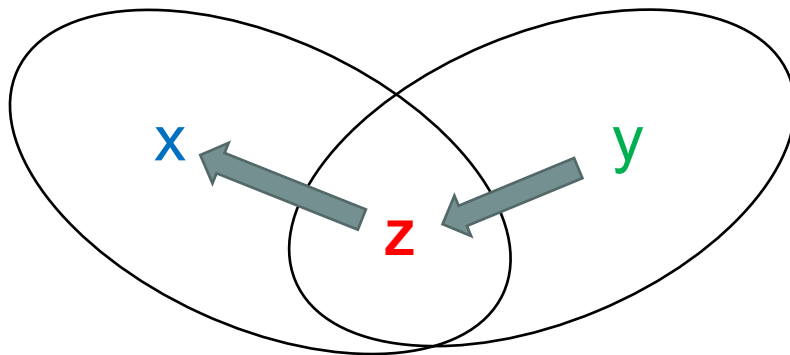
個人情報がないと結合できない！

仮想人体生成モデルの考え方：共通属性を使った結合

	$z_1 z_2 \dots z_n$	$x_1 x_2 \dots x_m$
	$p(Z, X)$	

	$z_1 z_2 \dots z_n$	$y_1 y_2 \dots y_k$
	$p(Z, Y)$	

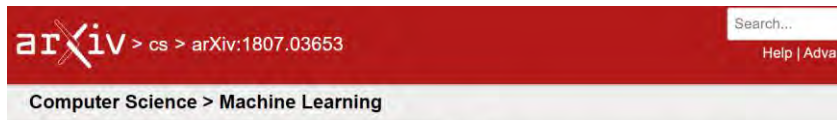
個人を識別できる情報がない



あるYの下でのXの分布は？
YからZの分布を求め、その
条件の下でのXの分布を求める

訓練アルゴリズム

欠損値のある、連続・離散変数混合の生成モデリング

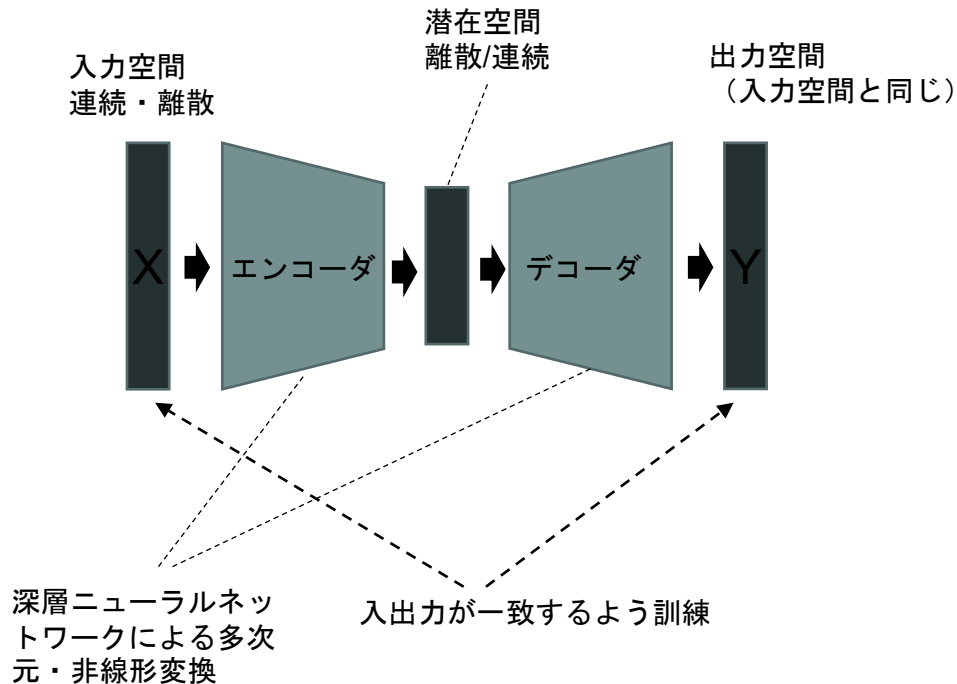


[Submitted on 10 Jul 2018 (v1), last revised 22 May 2020 (this version, v4)]

Handling Incomplete Heterogeneous Data using VAEs

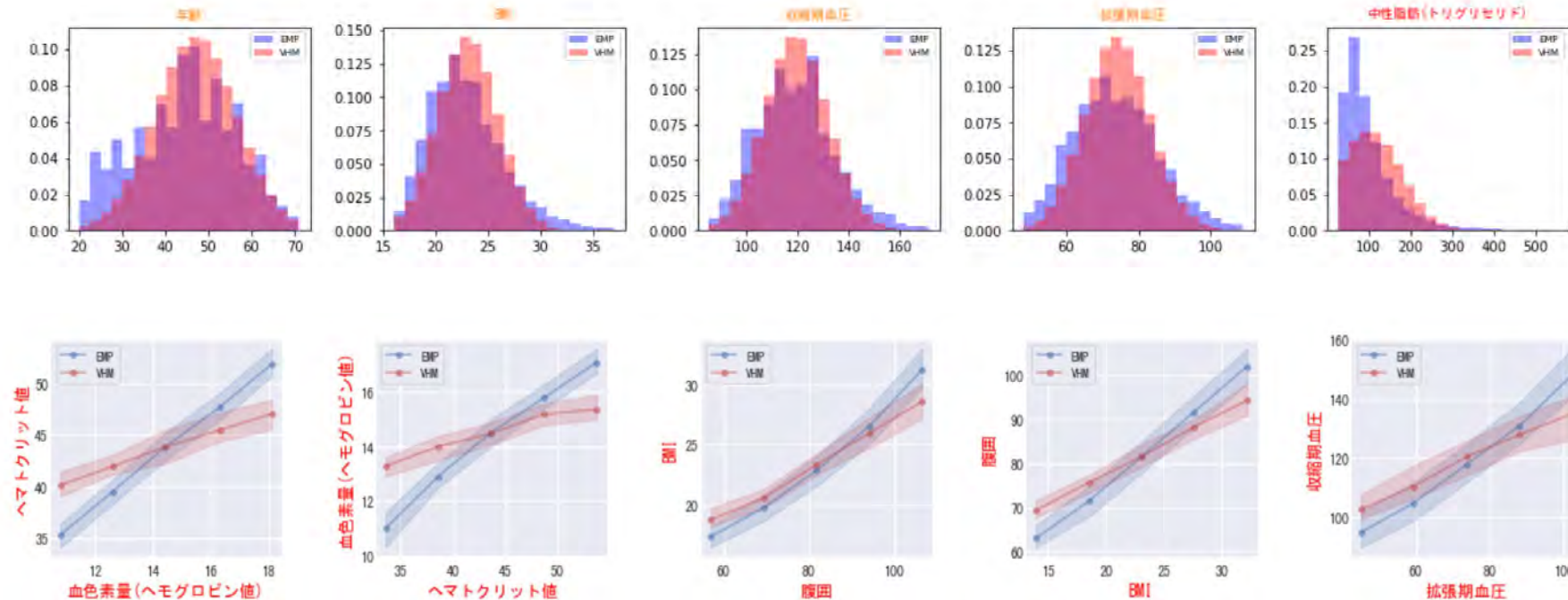
Alfredo Nazabal, Pablo M. Olmos, Zoubin Ghahramani, Isabel Valera

Variational autoencoders (VAEs), as well as other generative models, have been shown to be efficient and accurate for capturing the latent structure of vast amounts of complex high-dimensional data. However, existing VAEs can still not directly handle data that are heterogeneous (mixed continuous and discrete) or incomplete (with missing data at random), which is indeed common in real-world applications. In this paper, we propose a general framework to design VAEs suitable for fitting incomplete heterogeneous data. The proposed HI-VAE includes likelihood models for real-valued, positive real valued, interval, categorical, ordinal and count data, and allows accurate estimation (and potentially imputation) of missing data. Furthermore, HI-VAE presents competitive predictive performance in supervised tasks, outperforming supervised models when trained on incomplete data.



評価：経験分布とモデル分布の乖離

AS OF MAY 2022*



*モデルは毎月改良されていきます

すこやかペルソナ 標準版

モデル選択202110

探索条件選択Select...

全項目花王食生活設問抜粋

項目	入力値
性別	
年齢	
身長	
体重	
BMI	
内臓脂肪面積 (男:0-100, 女:0-100)	
収縮期血圧 (座位による測定) (男:90-139, 女:90-139)	
拡張期血圧 (座位による測定) (男:0-90, 女:0-90)	
中性脂肪(トリグリセリド) (-) (男:35-149, 女:35-149)	
hdlコレステロール (男:40-86, 女:40-96)	
ldlコレステロール (男:70-119, 女:70-119)	
got(ast) (GOT (グルタミン酸オキサロ酢酸トランスアミナーゼ)) (男:0-40, 女:0-40)	
gpt(alt) (GPT (グルタミン酸ピルビン酸トランスアミナーゼ)) (男:0-45, 女:0-45)	
γ-gt(γ-gtp) (γ-グルタミルトランスペプチターゼ) (男:0-75, 女:0-45)	
空腹時血糖 (男:70-99, 女:70-99)	
随時血糖	
hba1c (男:4.6-5.5, 女:4.6-5.5)	
ヘマトクリット値 (男:39.8-51.8, 女:33.4-44.9)	
赤血球数(単位:10 ¹² /L)	

4

利用例1: 疾病リスクの網羅的推定

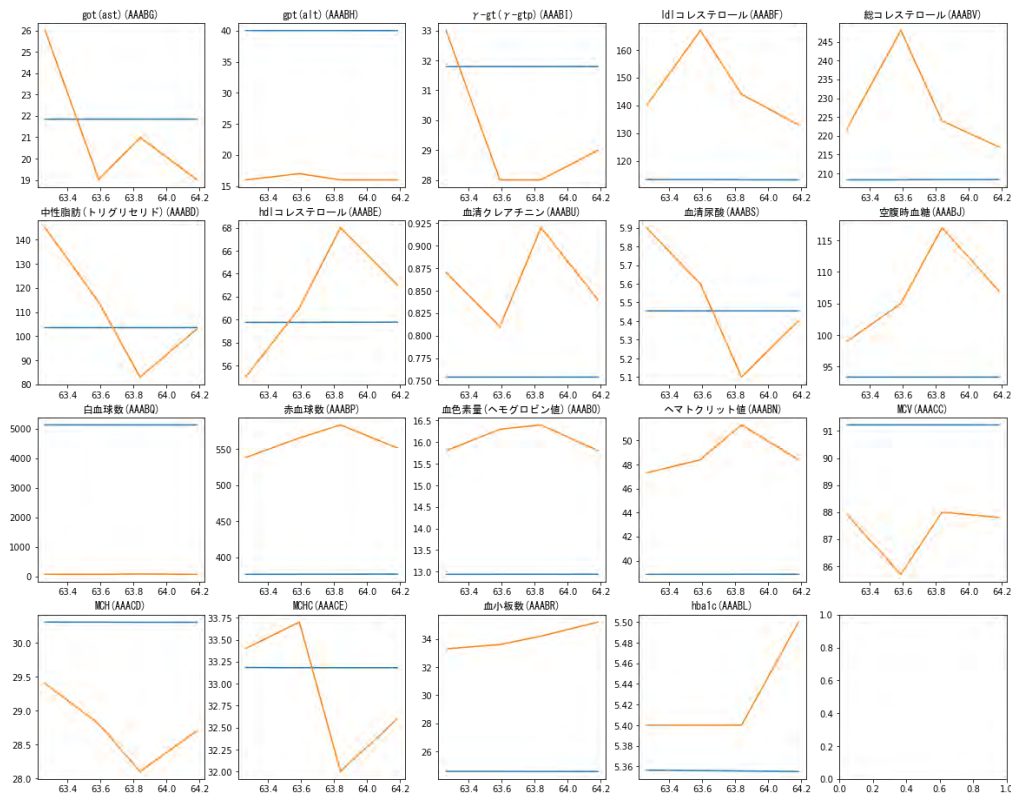
丸山の血液検査データ

- 64歳・男性
- 高血圧で、2005年から降圧剤を定期的に服用
- 3ヶ月に1回、クリニックで血液検査
- 4回分の血液検査の結果を、スキャン・アップロード

ユーザーID	12637			
お名前	丸山宏			
ダウンロード日	2022/7/27			
日付	2021/7/10	2021/11/6	2022/2/5	2022/6/11
検査名	検査報告書	検査報告書	検査報告書	検査報告書
病院名	依田医院（大田区）	依田医院（大田区）	依田医院	依田医院(大田区)
メモ				
総ビリルビン(T-BIL)	1.1	1.1	1.1	1.2
AST(GOT)	26	19	21	19
ALT(GPT)	16	17	16	16
乳酸デヒドロゲナーゼ(LD・LDH IFCC)	325	205	228	236
γ-GT・γ-GTP (GGT・G-GT)	33	28	28	29
コリンエステラーゼ(ChE)	365	368	382	348
LDL-コレステロール(LDL-C)	140	167	144	133
総コレステロール(T-CHO・TC)	222	248	224	217
中性脂肪(TG・トリグリセリド・トリグリセライド)	145	114	83	103
HDL-コレステロール(HDL-C)	55	61	68	63
クレアチニン(Cre)	0.87	0.81	0.92	0.84
尿酸(UA)	5.9	5.6	5.1	5.4
グルコース（血糖・空腹時血糖）(BS・GLU)	99	105	117	107

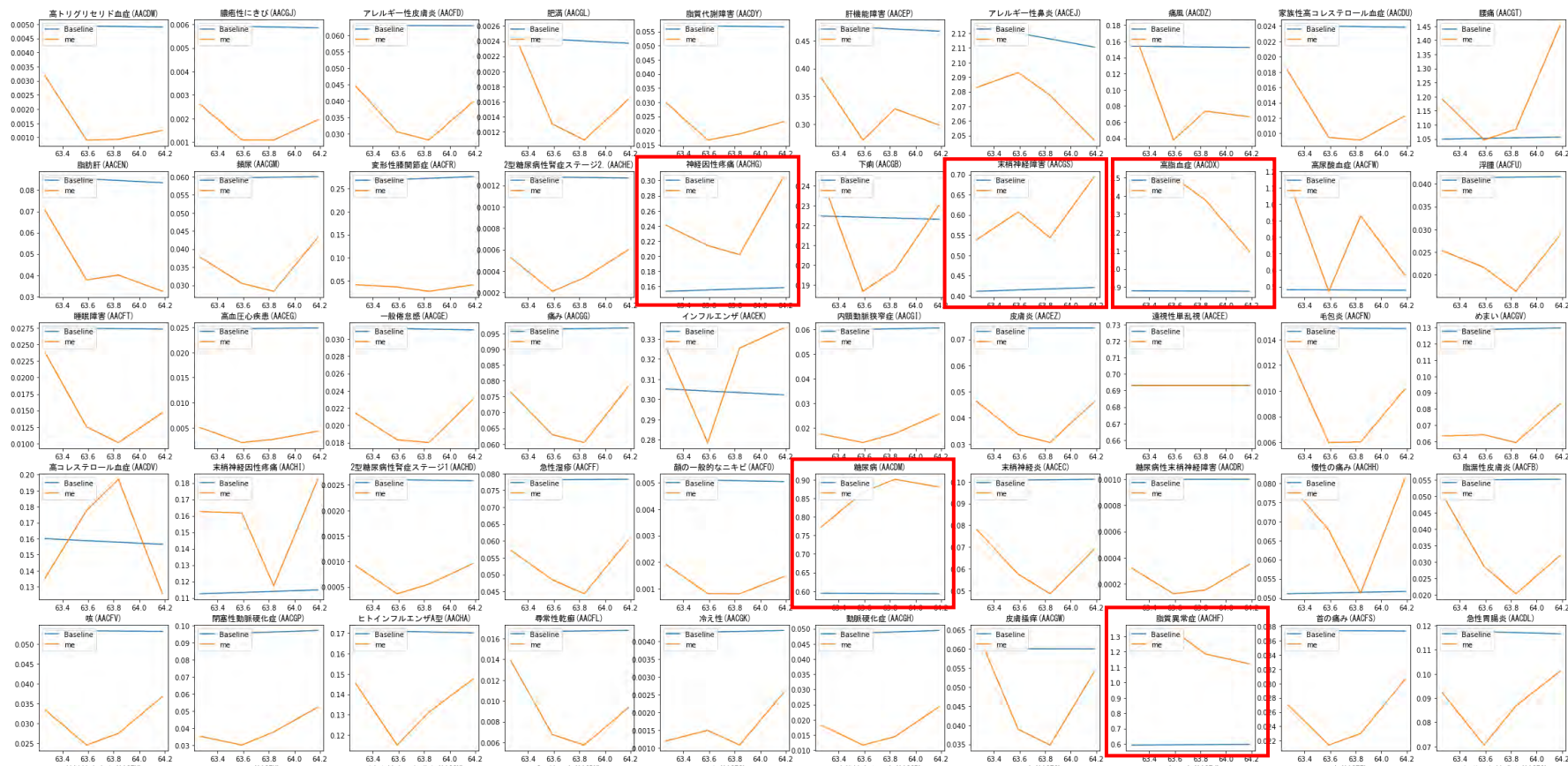
利用例1: 疾病リスクの網羅的推定

丸山の血液検査データの推移（過去1年、4回分）

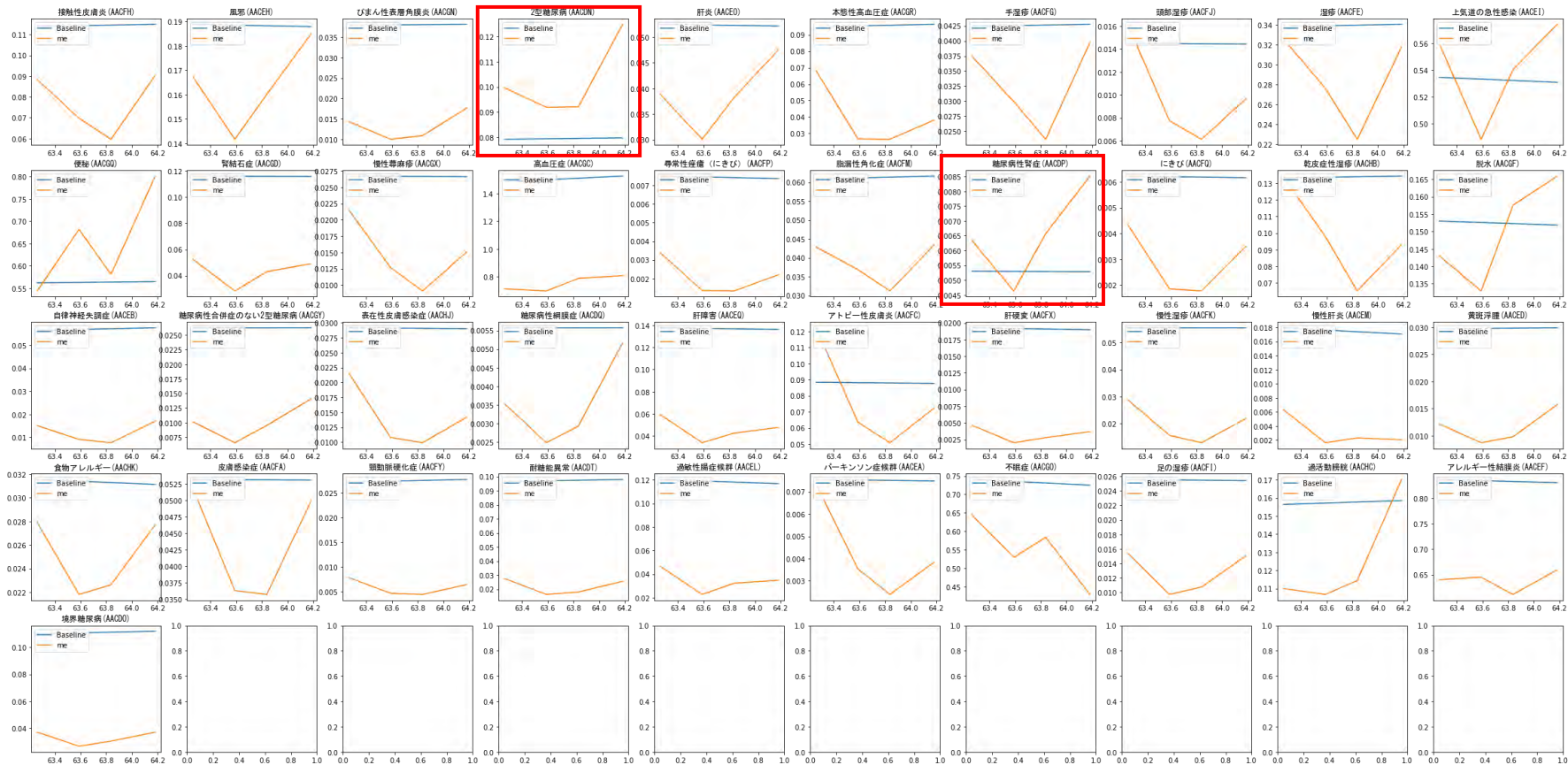


利用例1: 疾病リスクの網羅的推定 丸山の疾病リスク一覧 (1/2)

• Model: chalawan-a



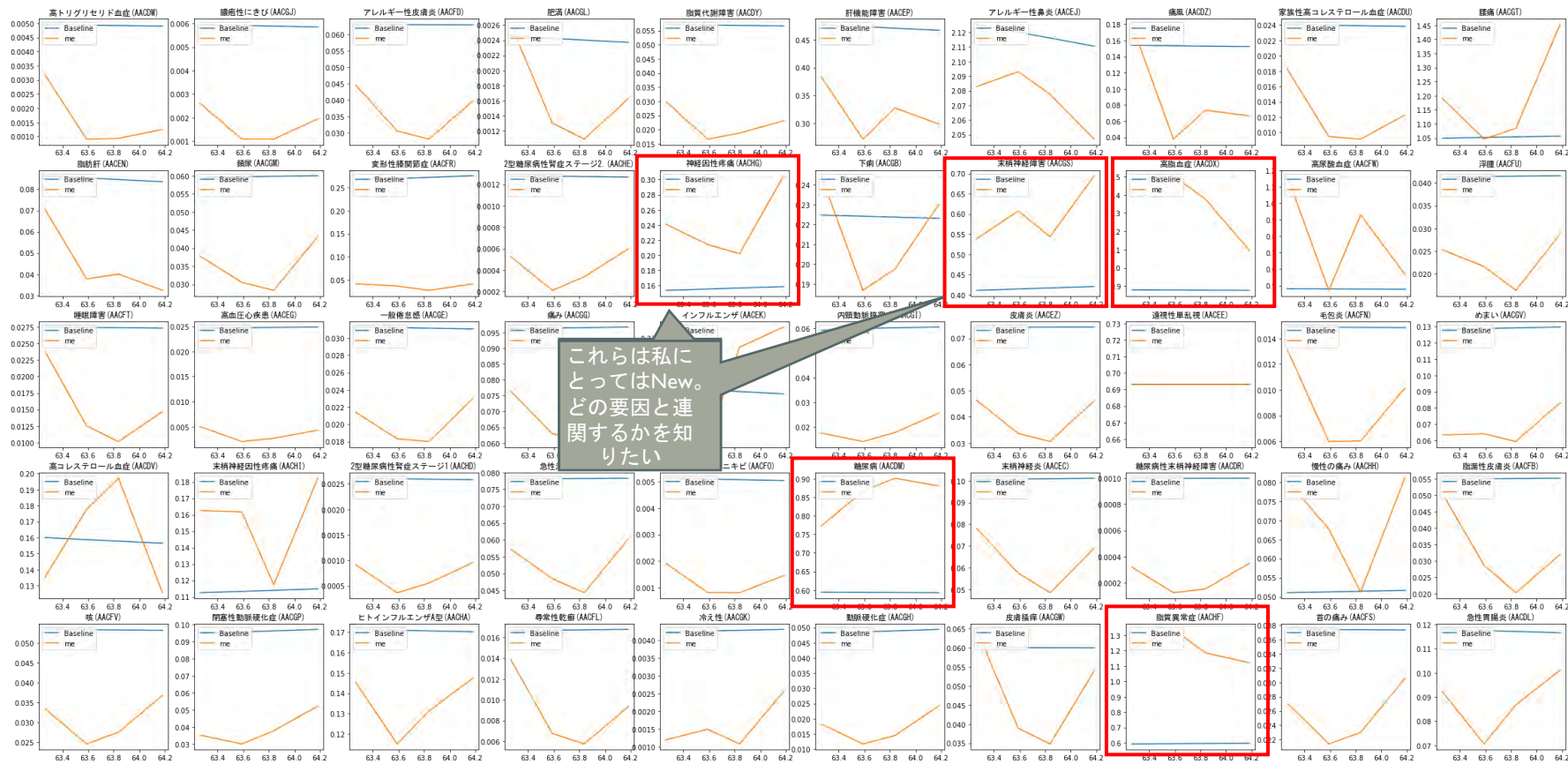
丸山の疾病リスク一覧 (2/2)



利用例2: 特定疾患に対する生活習慣最適化

丸山の疾病リスク一覧 (1/2)

• Model: chalawan-a



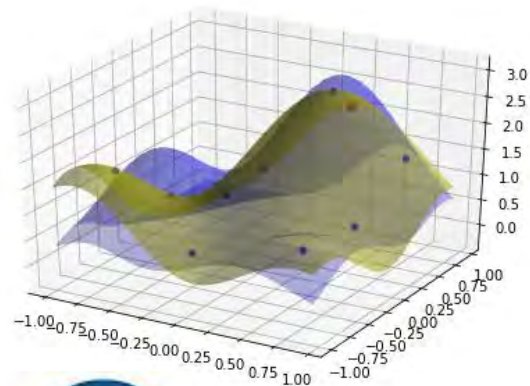
利用例2: 特定疾患に対する生活習慣最適化

問： 11種の生活習慣のどの組み合わせだったら、神経因性疼痛のリスクが最小だっただろうか？

神経因性疼痛(AACHG)のリスクを最小化する生活習慣, 最小化されたリスク=0.04642033949494362

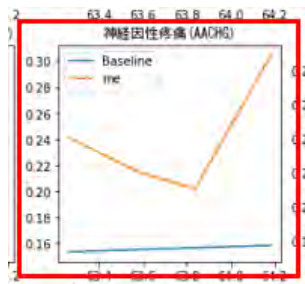
1. 喫煙(AAADZ), 1=はい
2. 飲酒(AAAEB), 2=時々
3. 飲酒量(AAAED), 3=2~3 合未満
4. 歩行又は身体活動(AAAFC), 2=いいえ
5. 歩行速度(AAAFD), 2=いいえ
6. 30分以上の運動習慣(AAAFB), 2=いいえ
7. 食べ方1(早食い等)(AAAEN), 3=遅い
8. 食べ方2(就寝前)(AAAEQ), 2=いいえ
9. 食べ方3(間食)(AAAEQ), 2=時々
10. 食習慣(AAAER), 1=はい
11. 睡眠(AAAES), 2=いいえ

こんなアドバイスは、医者からは絶対に出て来ないだろうが...



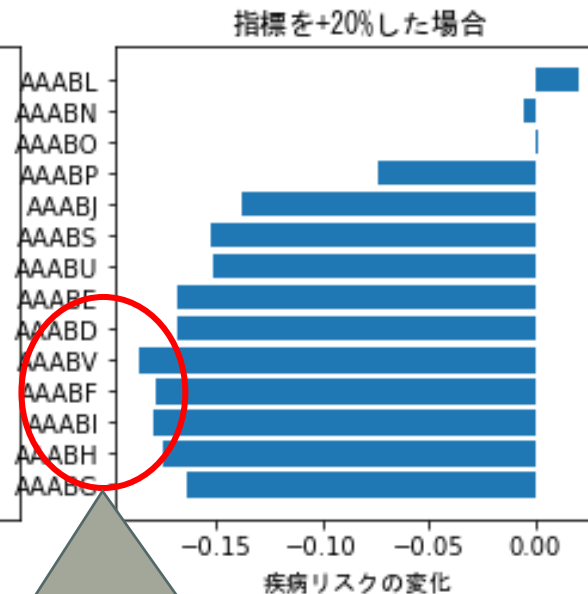
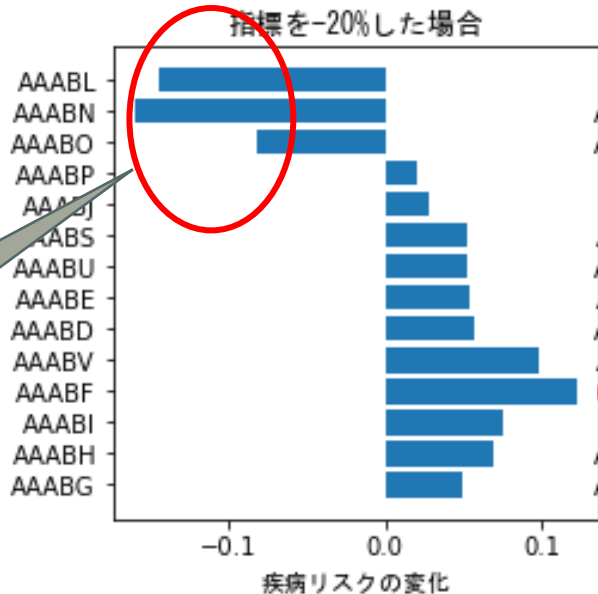
ベイズ最適化ツール
Optuna

血液検査と神経性疼痛（AACHG）のリスク



HbA1c(AAABL),
ヘマトクリット値(AAABN),
ヘモグロビン(AAABO)
が小さいと、リスクも小さい

神経因性疼痛に影響を与える因子



総コレステロール(AAABV),
ldlコレステロール(AAABF),
γ-gtp(AAABI)が大きいほうがリスクが小さい

利用例3: 行動変容につなげる、シナリオ可視化

仮想人体ベクトル可視化ツール



現状の仮想人体ベクトル

M	64	170	62.3	120	70	1.1	26	...
---	----	-----	------	-----	----	-----	----	-----

シナリオ1:

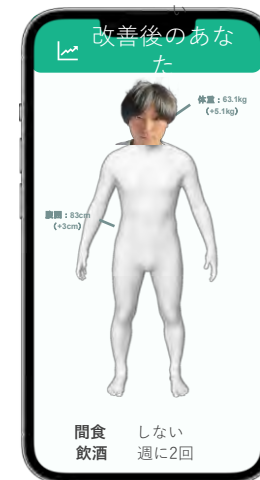
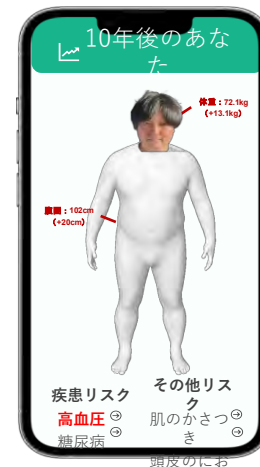
- 年齢 +10
- 運動習慣 : なし
- 飲酒 : 毎日
- 食事習慣 : ...

M	64	170	62.3	120	70	1.1	26	...
---	----	-----	------	-----	----	-----	----	-----

シナリオ2

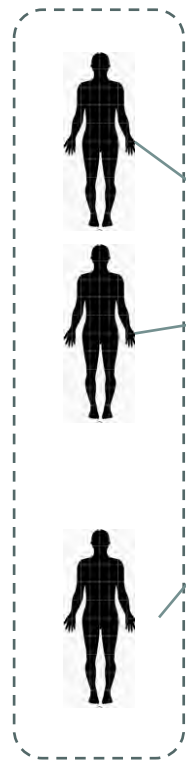
- 年齢 +10
- 運動習慣あり
- 飲酒 : 時々
- 食事習慣 ...

M	64	170	62.3	120	70	1.1	26	...
---	----	-----	------	-----	----	-----	----	-----



利用例4: チーム役割の最適化

チームメンバー



駅伝メンバー割当

第1区	M	64	170	62.3	120	70	1.1	26	...
第2区	M	64	170	62.3	120	70	1.1	26	...
:									
第10区	M	64	170	62.3	120	70	1.1	26	...

パフォーマンス予測



0:57:32
1:12:14
:
0:58:40



合計

合計時間を最短にする割当を探索

議論1: データの価値とモデルの価値

- データはそのままでは価値を産まない。価値を産むのはモデル
- 品質の悪いデータから良いモデルは産まれない – “Garbage-in, garbage-out”
- 目的をよく吟味し、よく計画され、制御された状況でデータを集められる者だけが、良いモデルを作れる

Harvard
Business
Review

If Your Data Is Bad, Your Machine Learning Tools Are Useless

by Thomas H. Davenport

Managing Director

of Data Science at IBM Research



Harvard Business Review 4/2の記事

<https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>

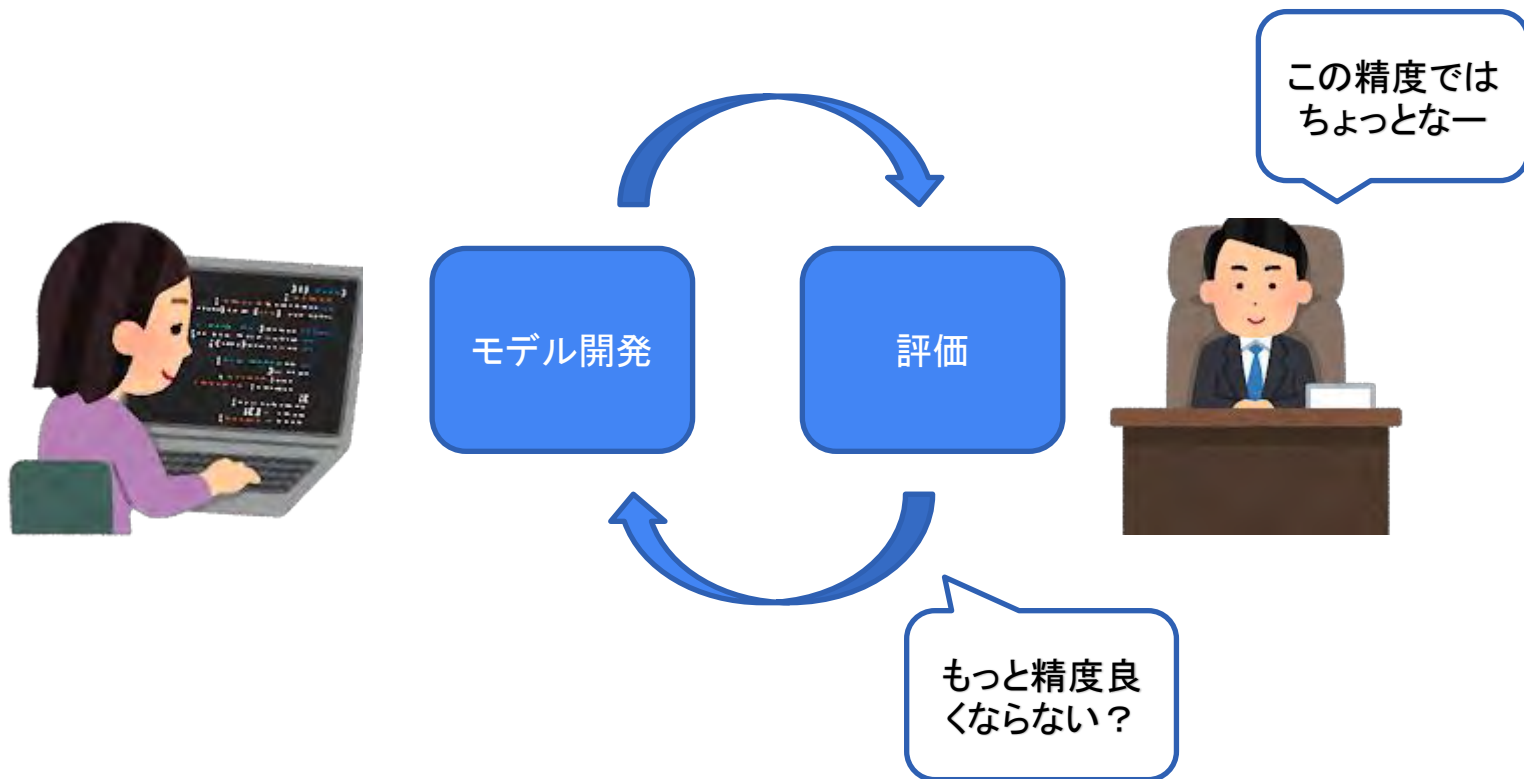
データを持っていること（ストックデータ）の価値

<

モデルを産み出す能力（+データを設計できる能力）の価値

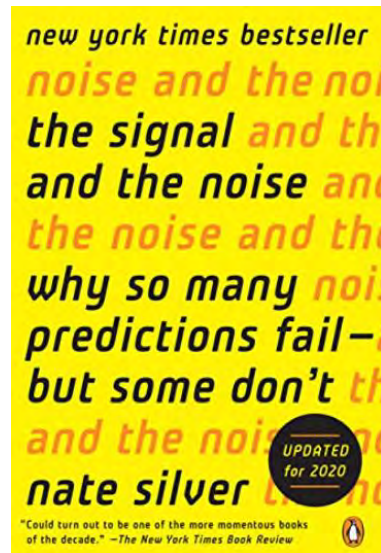
データサイエンティストの悲哀

議論2: 「予測精度」の呪縛: 「PoC貧乏」の構造



何のための「予測」か？

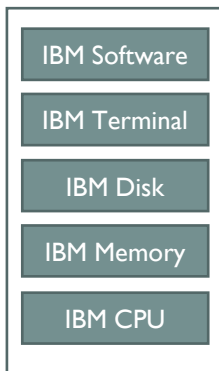
- 未来を当てたい
 - 当てられるもの（台風の予想進路、スポーツ、...）
 - 「良い予測をするためには、過去のデータの精度を語ってもあまり意味がない。繰り返し未来の予測を行い、そこから学ぶこと」
- 次の行動を決めたい ⇐ 多くの場合こちら
 - 可能な将来シナリオを描く
 - 例：西浦先生2020年4月
「感染防止しなければ最大40万人死亡」
 - シナリオ検討には、精度は必ずしも重要でない
 - モデルは意思決定・コミュニケーションの道具



ISBN-13 : 978-1594204111

議論3: イノベーション誘発の構造 -- PCのケース

IBM System/360 (1960年代) – 垂直統合

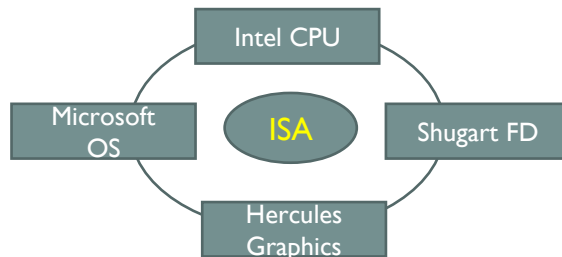


出典 : wikipedia

IBM PC (1981年) – 水平分業



出典 : wikipedia

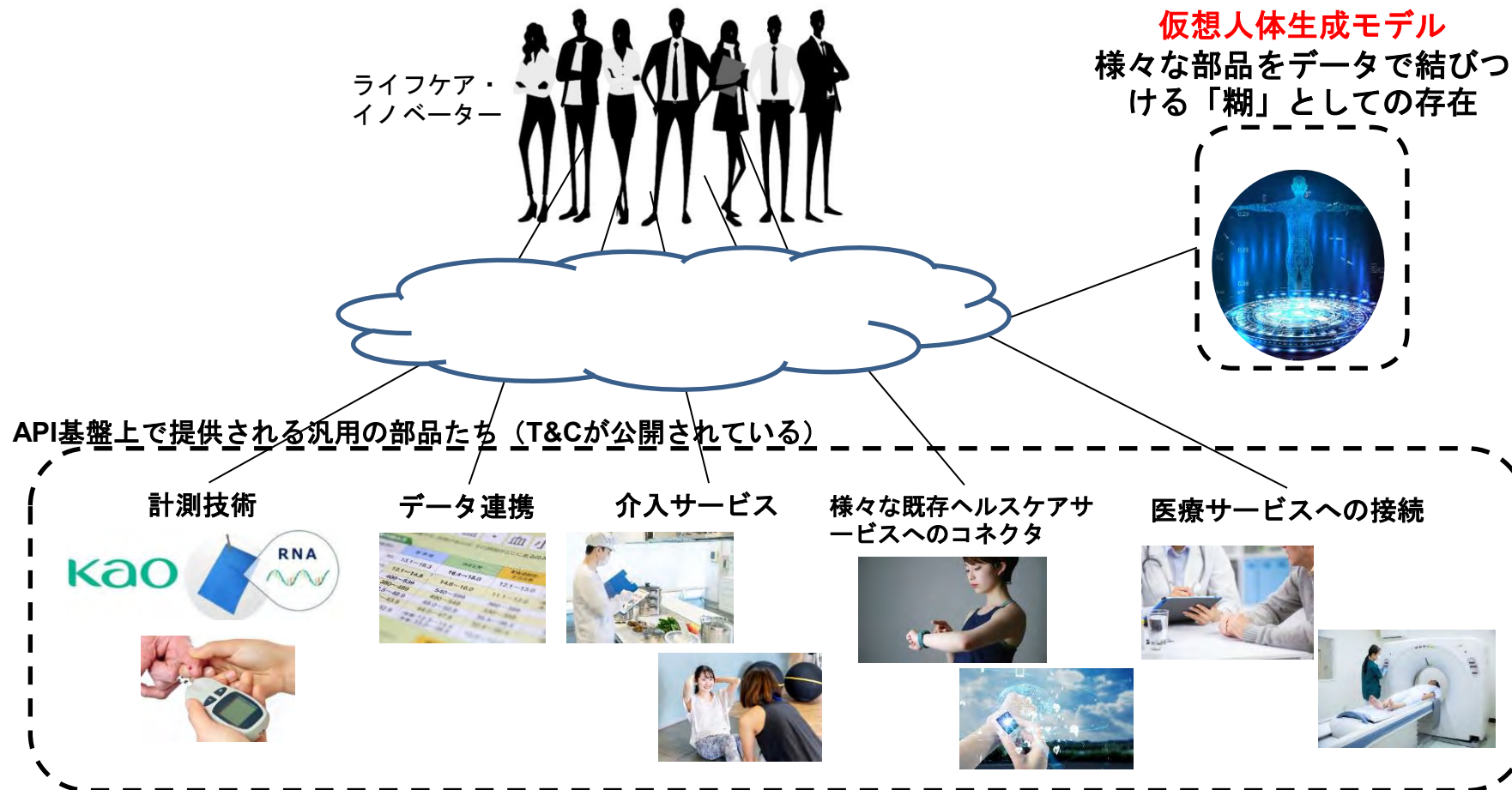


Web 2.0 (2000年-) – Net上での統合



出典 : https://gigazine.net/news/20061119_web2logo/

技術・データの民主化がライフケアイノベーションの起爆剤となる



議論4: 統計モデルは知識の一部に過ぎない

The Institute of Statistical Mathematics
大学共同利用機関法人 情報・システム研究機構

統計数理研究所

文字サイズ | 小 | 中 | 大

コラム 155

統計科学とシステム科学

宮里 義彦 (モデリング研究系)

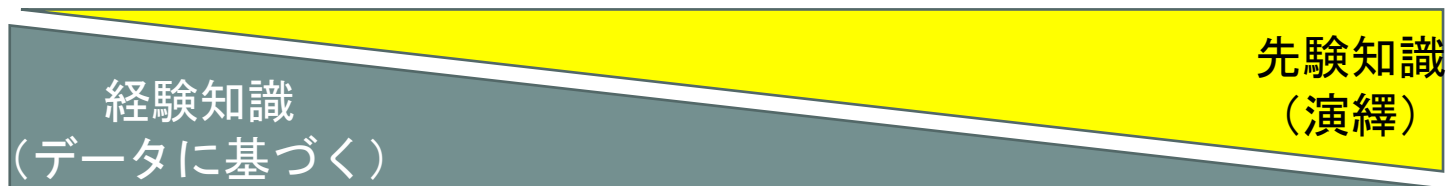
R. E. Kalmanのもとで学位を取得した日本人は数少ないが、その一人の研究室でKalmanが講演する情報を得て習志野の地に聴講に行ったことがある。統数研に入所してまだ数年の頃で、講演は経済学の時系列データから経済システムを実現する内容であったと記憶しているのだが、講演それ自体より冒頭でのKalmanの主張が強く印象に残っている。Kalmanの主張とは統計科学あるいは統計学者に対する激しい批判であった。

Kalmanの批判

“統計科学あるいは統計学者は考察の対象としているシステム本体には多くの関心を払わず、本体をモデル化したときの残差項に精緻な確率構造を当てはめて解析を行っている”

https://www.ism.ac.jp/ism_info_j/lab/column/155.html

知識には先験知識と経験知識がある



ユーザーに
関する情報

ベストプラ
クティス

工学

医学

心理学

商品に関す
る情報

社会学



これらを組み合わせることでイノベーションが起きる

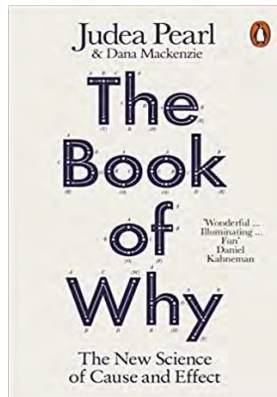
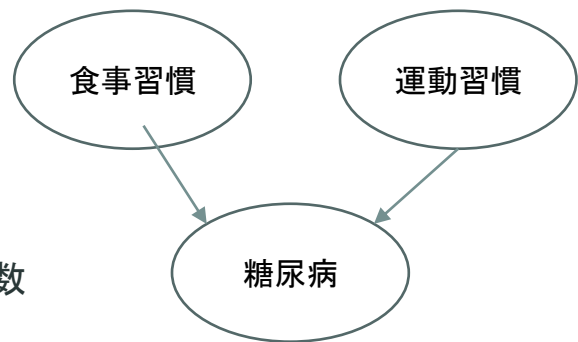
経験知識と先験知識の組み合わせが可能になりつつある

- ・ 因果推論：因果モデルとデータの組み合わせ
 - ・ e.g., “糖尿病の原因は何か”
- ・ 行動変容：行動経済学との組み合わせ
 - ・ e.g., “糖尿病リスクに関連する最も効果的（コスト最小）な変数
は何か”
- ・ :



- ・ 必要な先験知識は多くの場合状況依存
 - ・ 直接仮想人体生成モデルに組み込むことをせず、組み合わせのデ
ザインパターンを整備

でも、経験知識は仮想人体生成モデルにまかせて！



ISBN-13 : 978-0141982410



ISBN-13 : 978-4764906068

Thank You

Twitter: @maruyama