

aa-13. ビッグデータと AI

金子邦彦



「人工知能」第13回の内容



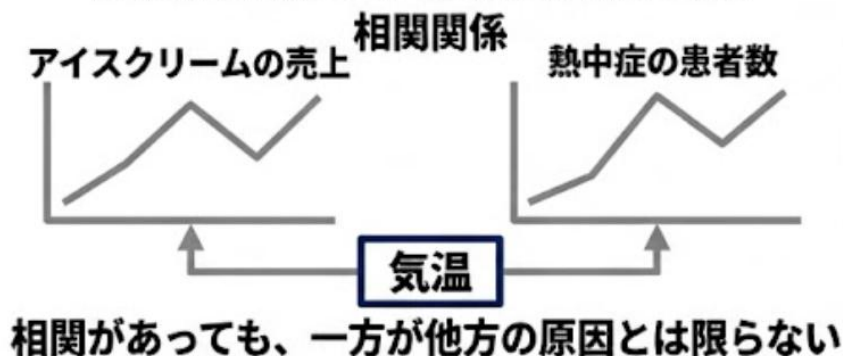
ビッグデータの3つの特徴

量（データの大きさ）

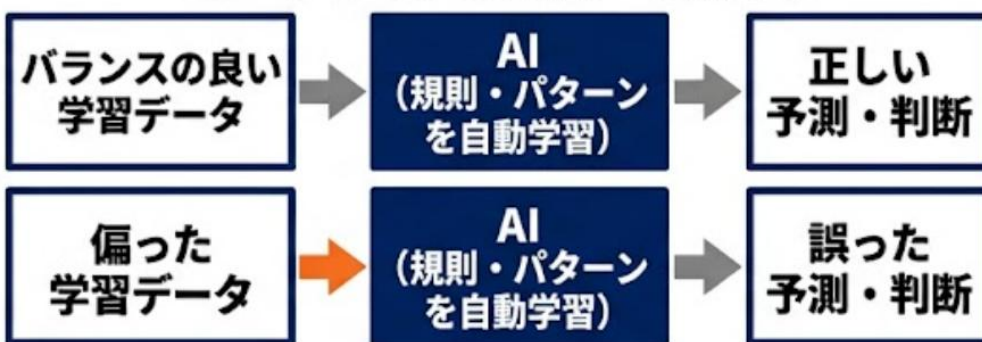
速度（データが生まれる速さ）

多様性（データの種類の豊かさ）

相関関係と因果関係は別



AIの学習と偏りの影響

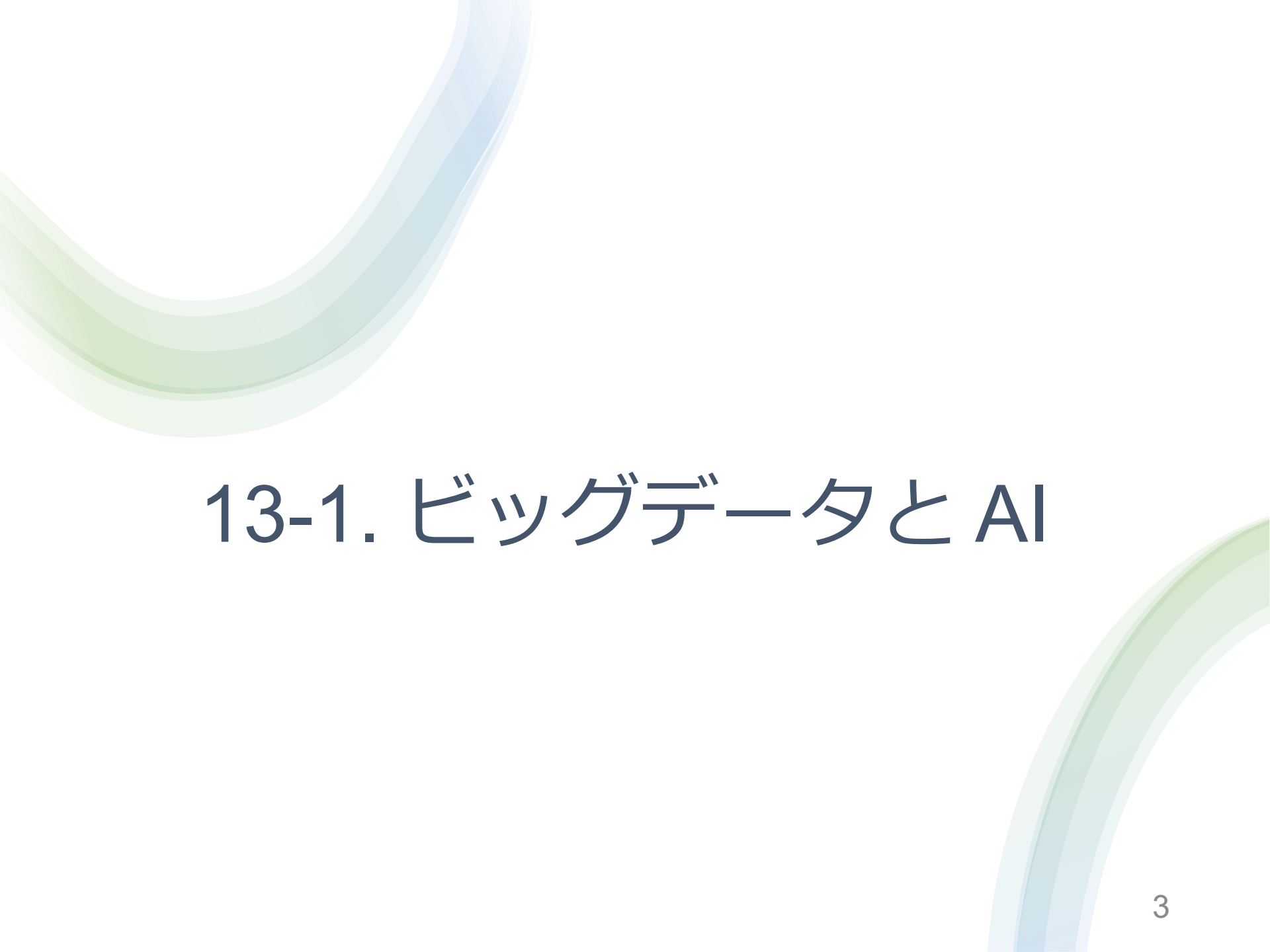


データの質を高める3つの工夫

表記の統一
(例: "東京"と"Tokyo"のような
表記のばらつきをなくす)

欠損値の除外
(空白のデータを取り除く)

外れ値の検出 (IQR法):
四分位範囲から大きく外れた
値を見つけて取り除く



13-1. ビッグデータと AI



ビッグデータとは、巨大かつ複雑なデータ群のこと

3つの 特徴

量

データの総量が大きい

速度

データが生成・記録・処理
される速さが速い

多様性

データの形式や種類が多様
(数値・テキスト・画像・音声など)

活用 のフロー

大量かつ
多様な
データ

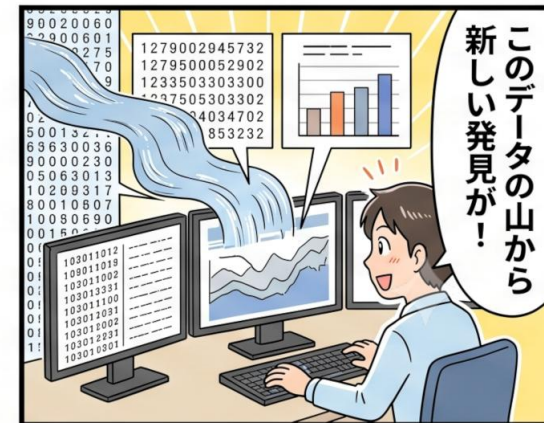
分析

見えなかった
傾向・相関を
発見

精度の高い
予測・
意思決定

注意点

- ① 個人情報を含むことが多い
→ プライバシー保護と情報セキュリティの確保が不可欠
- ② データの量が多いだけでは価値は保証されない
→ データの質 (正確性・信頼性) が低ければ、
分析結果も誤ったものになる



毎日使っているビッグデータ



動画配信のおすすめ = 性質の違うデータを組み合わせて次に見る作品を予測



多様性 (Variety) = 形式の違うデータが混ざっていることが特徴

たった1人の遺伝子でも、巨大データ



たった1人分の遺伝情報でも約200ギガバイト = 量の桁違いさ



ふつうの書類や
スマホ写真1枚



ヒトゲノム1人分
=約200GB (米国
米国NHGRI公表)

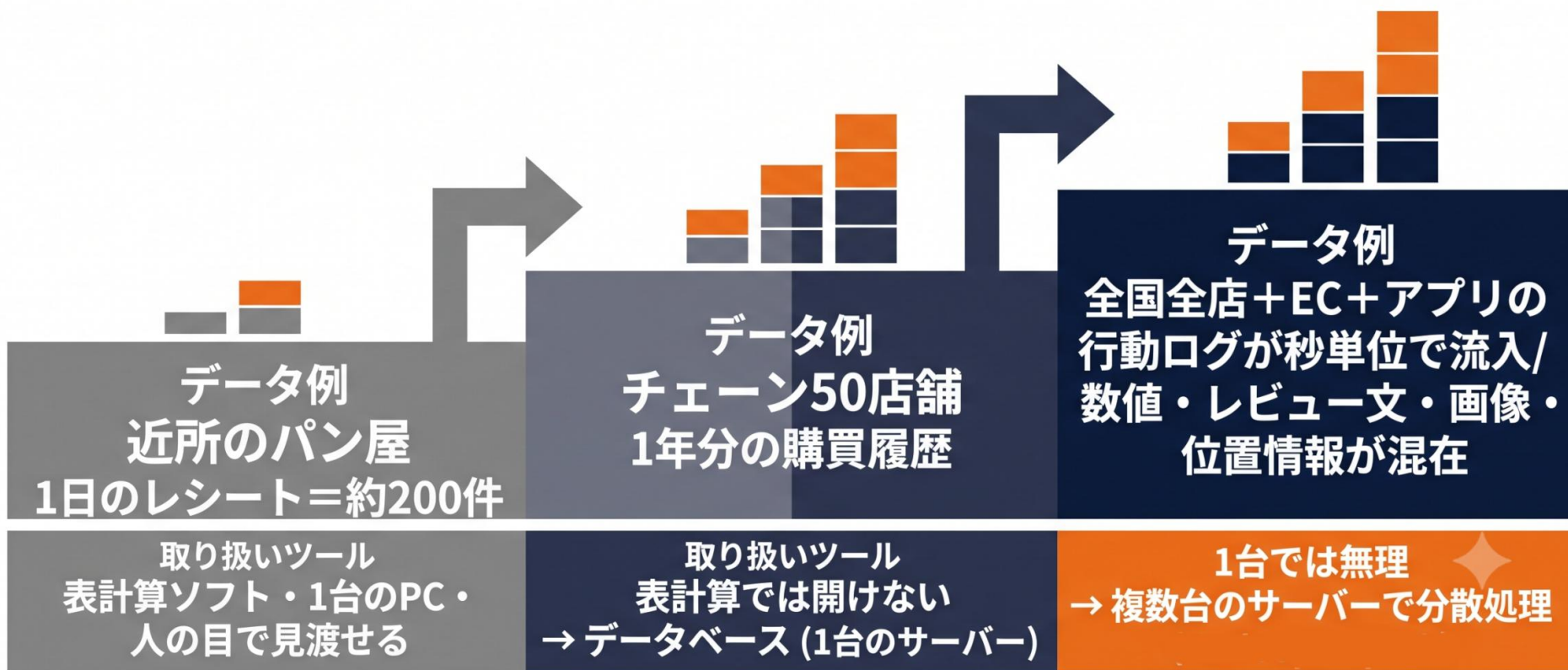


多人数分を集めて分析

→ 精密医療 (一人ひとりに合った治療) へ

量 (Volume) = 1件でもう巨大/
それが数百万～十数億人分に積み上がる

ビッグデータとは"件数"で決まるのではない = 今までの道具では扱いきれなくなったデータ



1台では、もう無理



1台のPCでは扱えなくなった = ではどうやって処理するのか？

全国全店+EC
+アプリの
ログが
秒単位で流入

全国全店+EC
+アプリのログが
秒単位で流入

大きすぎる仕事を、1台で
やろうとするから無理が出る

答え =
分けて
大勢で
同時に



データを複数台に分けて、結果を集める



分散処理 = 分けて・大勢で同時に処理し・最後に集める

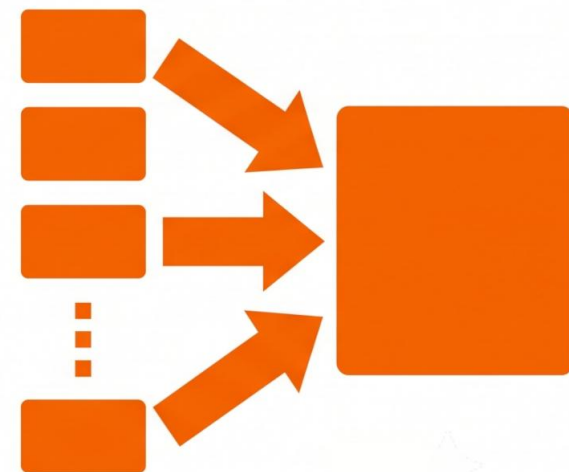
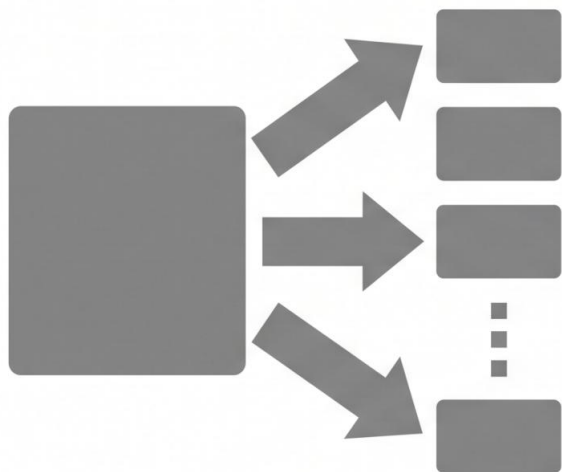
配る (Map)



同時にやる



集める (Reduce)



巨大な1台ではなく、普通のコンピュータを多数並べて協力させる

データから、AIがパターン・ルールを発見

大量かつ
多様な
データ



AI（機械学習）＝
データからパターン・
ルールを自動で発見

規則が見えない

データが少ない



データが多い

データが多いほど学べるパターンが豊かになる

"分けて同時に"は、データ処理だけでなくAIの学習にも使う

AIを学習させるとき

分散学習（複数のプロセッサで同時に学習）



データを分けて複数台でAI学習を行い結果を集める

大量に処理できる ≠ 正しく処理できる



便利になるほど弱点も拡大：『大量に処理できる』と『正しく処理する』は別

課題1：質



課題2：プライバシー・セキュリティ





13-2. データサイエンス のための AI



データの基礎知識があるほど、AIとよりの的確に対話できる

データの基礎知識

平均 / 中央値 / 外れ値 / 欠損値
相関 / 因果 / 標本

AIへの的確な指示ができる

例1：外れ値を除いた場合の傾向も教えて

例2：相関だけでなく因果関係の可能性も検討して

AIの結果を鵜呑みにせず 見極める力になる

例1 相関関係を因果関係と混同する

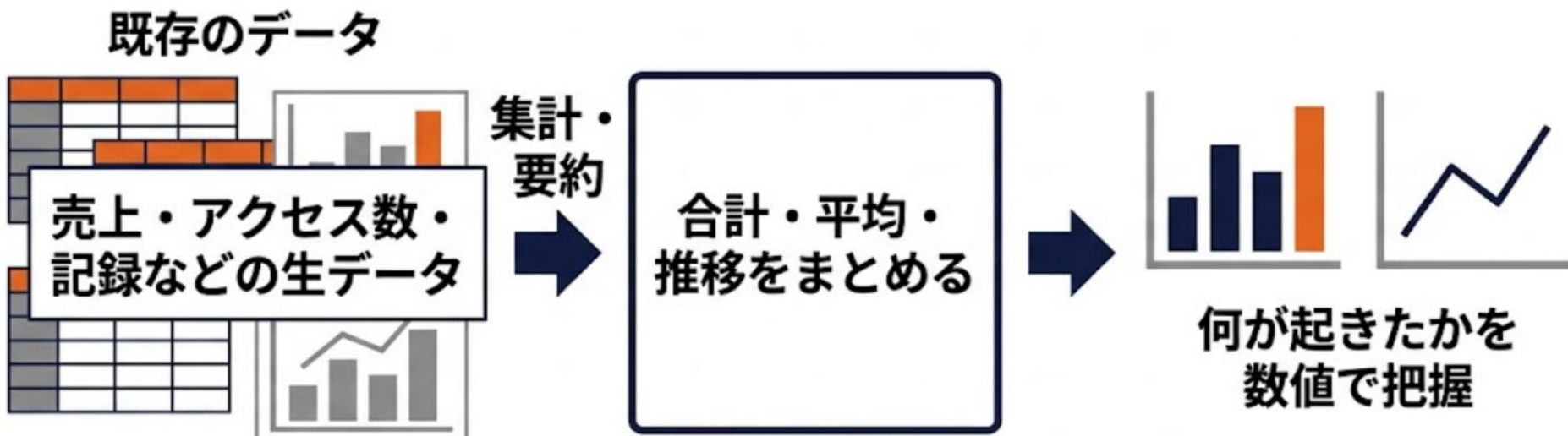
例2 少数の外れ値に平均値が引きずられる

AIは自信を持って誤った結論を提示することがある

記述的分析



記述的分析：過去・現在のデータを集計・要約し『何が起きたか』を示す分析



できること（有用性）

現状を数値で正確に把握できる

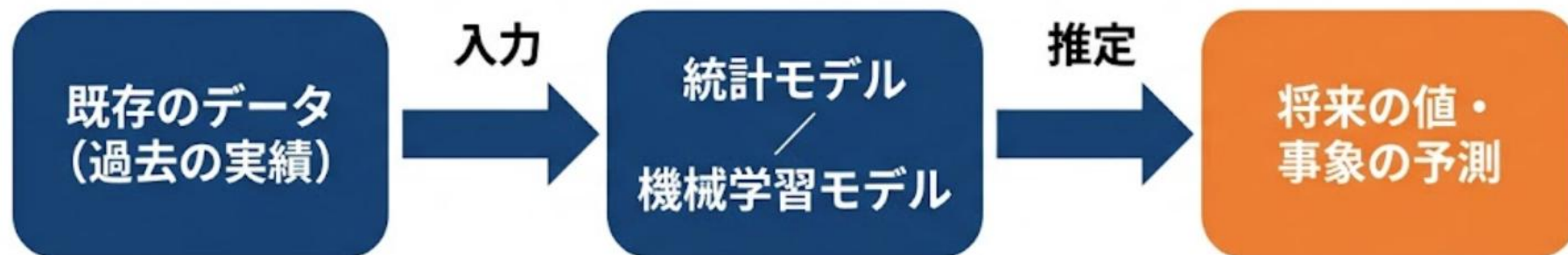
できないこと（限界）

なぜ起きたか（因果）や
今後どうなるかは示さない
相関があっても、そのまま因果ではない

データを用いた予測



データを用いた予測：過去のデータからモデルが将来の値や事象を推定する分析



有用性

過去のデータを基に、将来の見込みを数値で示せる

限界

精度は入力データの量と質に依存する
過去と異なる状況では予測は保証されない
相関は示せても因果関係は示せない

65 70 72 75 80

平均値：すべての値を足して
個数で割った値

$$(65+70+72+75+80)\div 5=72.4$$

データ全体の合計をならした値。すべての数字が計算に使われる

中央値：小さい順に並べたとき、ちょうど真ん中に来る値

65 70 **【72】** 75 80

大きさの順番だけに注目した値。真ん中の位置にある数字がそのまま答えになる

平均値と中央値は、必ずしも同じ値になるわけではない

平均値と中央値：外れ値による影響の違い



外れ値が1つ混ざるだけで、平均値は大きく引きずられるが、中央値はほとんど変わらない

外れ値：他の値から極端に離れた値
入力ミスや特殊事情で生じる (75の入力を750と誤入力)



**平均値：すべての値を足して
個数で割った値**
 $(65+70+72+75+750) \div 5 = 206.4$

外れ値に大きく引きずられて、
実際の点数からかけ離れた値になる

**中央値：小さい順に並べたとき、
ちょうど真ん中に来る値**

65 70 **72** 75 750

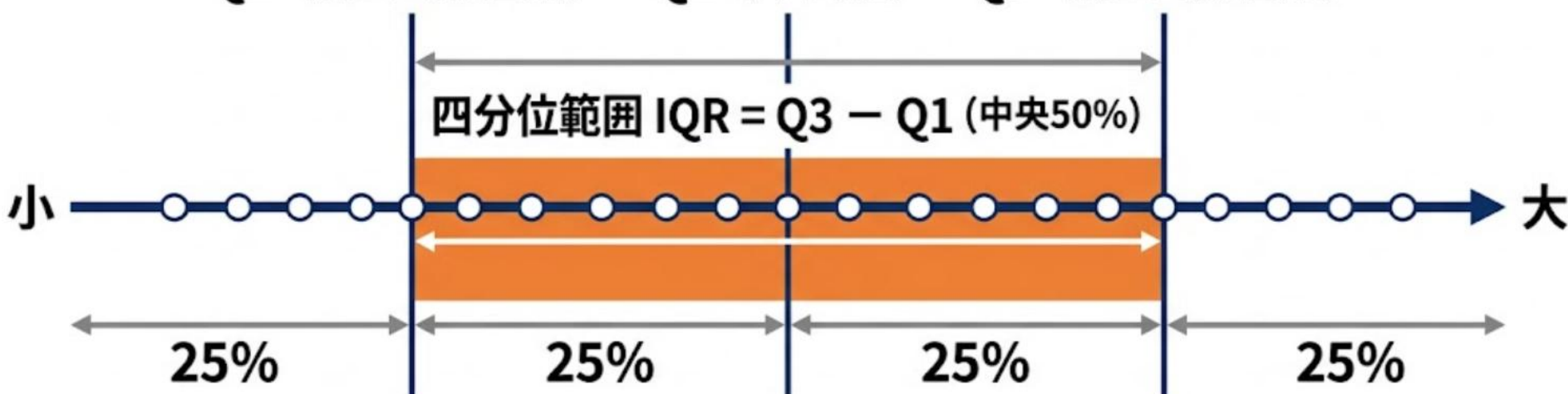
外れ値 (750) があっても、真ん中の位置
は変わらないためほとんど影響を受けない

四分位数



データを昇順に並べて4等分し、中央50%の範囲をつかむのが四分位数

Q1 (第1四分位数) Q2 (中央値) Q3 (第3四分位数)



有用性

外れ値の影響を受けにくい／
データ中央50%の範囲が分かる

限界

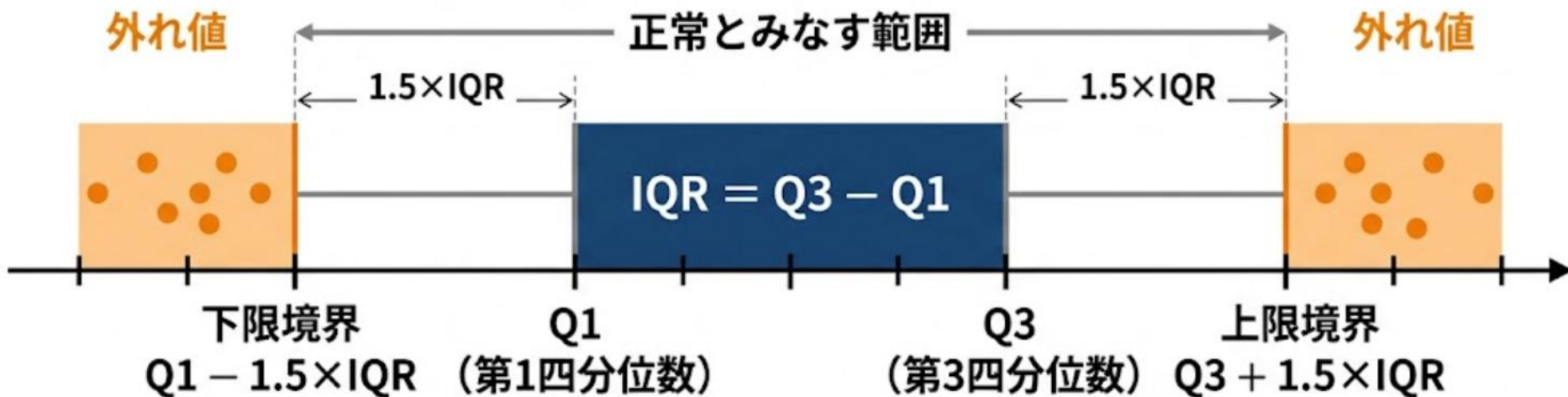
中央50%の範囲を示すだけ／
個々の値の大小は分からない

Q3 - Q1 (IQR) は次のIQR法の計算に用いる

四分位範囲 IQR を用いた外れ値の推定



四分位範囲IQRを使い、外れ値を主観でなくデータから機械的に判定する手法



定義

$IQR = Q3 - Q1$ 。境界の外側を外れ値と判定

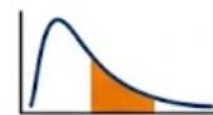
有用性

主観に頼らず、データから機械的に基準を算出できる

限界



対称な分布：
うまく機能



偏った分布：
正常値を誤判定しやすい

係数1.5は慣習値。偏った分布では正常値を外れ値と誤判定することがある

表の中の空欄（欠損値）は、対処が必要

空値（欠損値）とは？

氏名	身長	睡眠時間	好きな食べ物	学食の満足度
リク	175	(空欄)	ハンバーガー	3
ナナ	152	6.5	(空欄)	4
モモ	(空欄)	(空欄)	パスタ	2
トモヤ	176	7.0	(空欄)	4

空値（欠損値）＝回答がなく、データが記録されていない箇所

対処方法を結果が変わる
睡眠時間の平均
を計算する

方法A：
空欄の行を
無視する
(除外する)

方法B：
空欄を
平均値で
埋める

欠損値をどう扱うかで分析結果は変わる。無視するか、値で埋めるかを目的に応じて正しく選ぶ必要がある。

相関・因果…
似てるようで
全部違うんだな。
ちゃんと見分けよう



相関の定義と種類（正・負・なし）

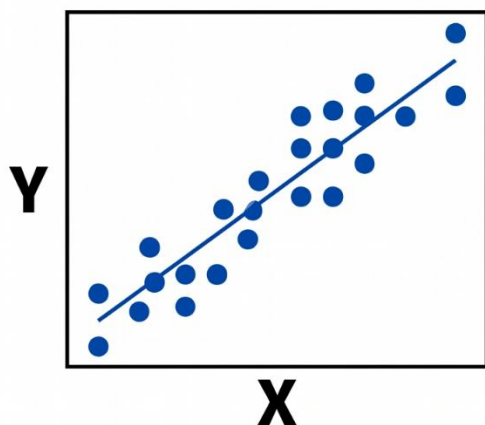


2つの変数の関係进行分析（数値データ用）

相関がある場合，一方が変化するともう一方も変化する傾向にある

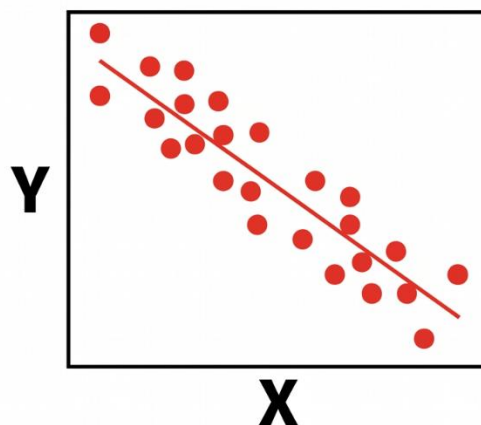
種類

正の相関



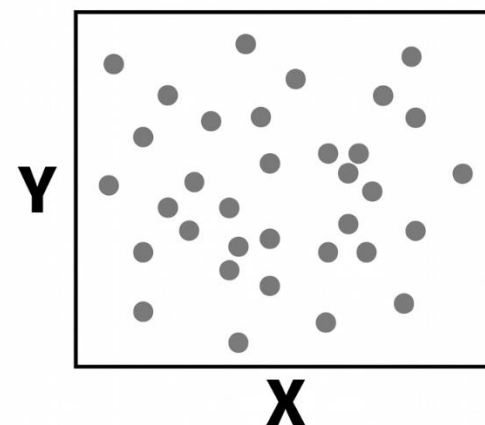
Xが増えるとYが増える傾向
例：勉強時間が増えると
得点上がる

負の相関



Xが増えるとYが減る傾向
例：ガソリン代が上がると
車の利用が減る

相関なし

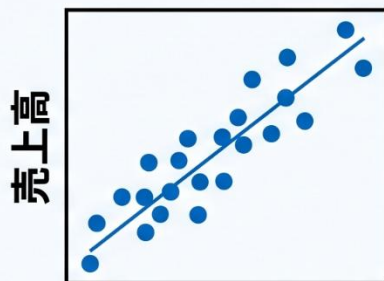


XとYに関係がない
例：足のサイズと勉強時
間に関係がない



2つの量の間の関係性の分析に活用される

① マーケティング



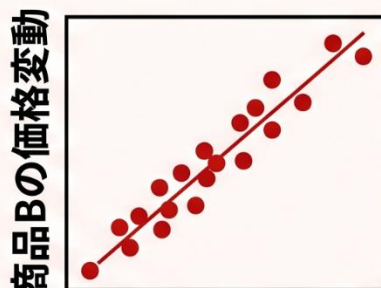
売上高

広告費

相関係数 $\doteq 0.8$

広告を増やすと売上高
が増えそうか

② 金融



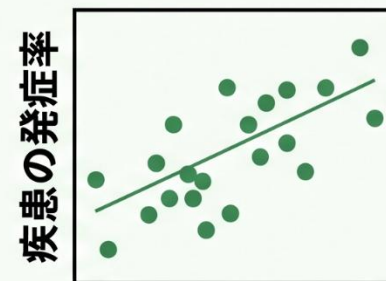
商品Bの価格変動

商品Aの価格変動

相関係数 $\doteq 0.9$

相関が高い複数の金融商
品を扱うとリスクが高いか

③ 医療・生命科学



疾患の発症率

遺伝子の発現量

相関係数 $\doteq 0.6$

遺伝子と疾患に
関係がありそうか

いずれも2つの量の間の関係性の分析に活用される

相関（2つのデータの関連）と因果（原因と結果の関係）は違う



一緒に動いて見えるだけの関係と、原因と結果でつながる関係は違う

相関



因果

一方が原因で、もう一方が結果になる関係



気温

アイスクリームの売上

一緒に増える（相関）

水の事故の件数

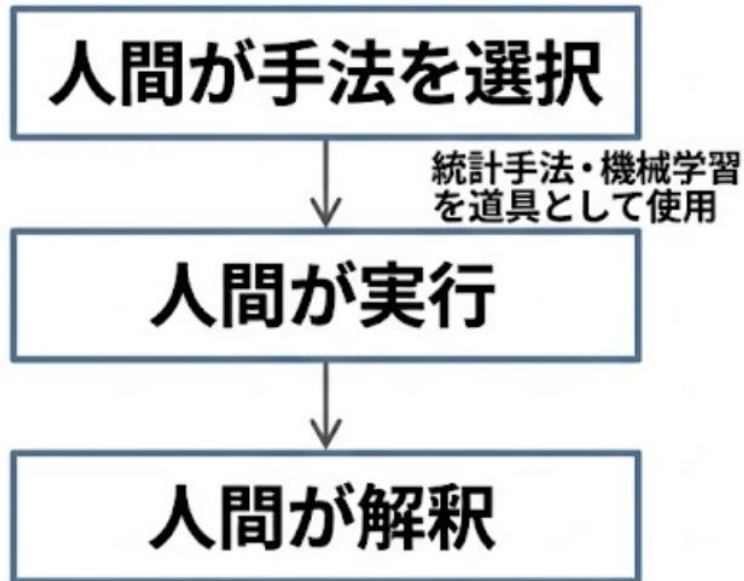
相関があるからといって、因果があるとは限らない

対話型AIの登場により、手法選択から解釈までの一連の作業を、人間とAIが対話しながら進められるようになった



データからのパターン・規則の発見

従来型



AIとの対話



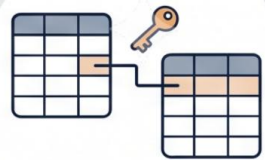
発見のプロセス自体を
人間とAIが協働で進める

手段：手作業・Excel・プログラミング → AIがプログラミングを支援

13-3. データエンジニア リングの実践

表記がバラバラだと
同じものが別々に数えられちゃう…
まずは表記をそろえて、キーでつなげよう

え？



表記のゆれを一つに統一し、正しく比較・集計するための処理

定義：

表記を一つに統一する

アップル (全角)

アップル (半角カナ)

あっぷる (ひらがな)

アップル

同じ対象を指す複数の表記
→ 1つに統一

有用性：

誤った集計を防ぐ

アップル：3件

アップル：2件



× 本来同じものが
別々に数えられる

アップル：5件



正しく集計できる

限界：

同一の基準は人が決める

どこまでを『同じ』と
みなすか＝人が定義する

アップル = アップル

りんご (果物) ≠ りんご (人名)

× 別のものを同じ扱いに
してしまう危険

自動では判定されない。基準を
誤ると別物を同一にしてしまう

表のキー列とデータ結合



キー列は複数の表を対応づける基準となる列

表A

日付	気温
2026/10/10	20
2026/01/19	6
2026/01/20	8

共通の列（キー列）を基準に、
別々の表の情報を1つに結びつける



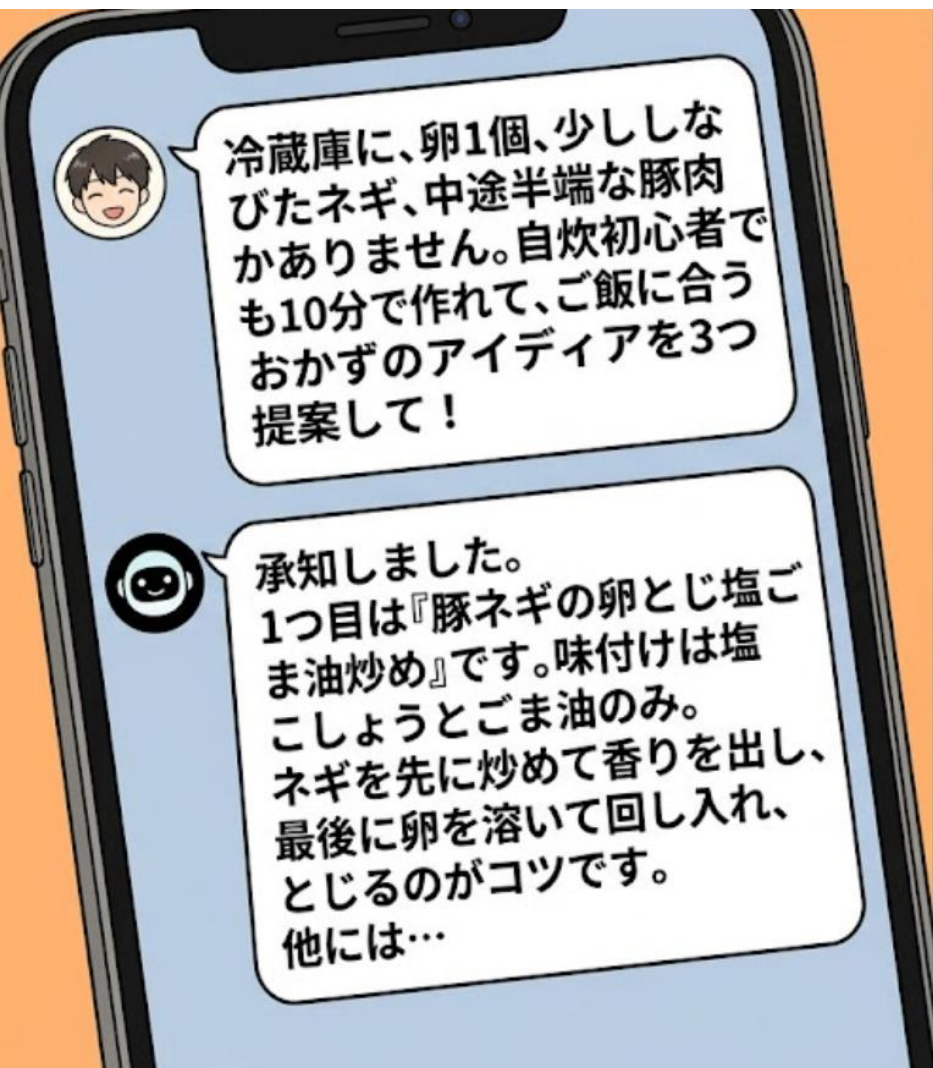
表B

日付	売上
2026/10/10	2,000
2026/01/19	5,000
2026/01/20	1,000

結合後の表（日付・気温・売上）

日付	気温	売上
2026/10/10	20	2,000
2026/01/19	6	5,000
2026/01/20	8	1,000

演習



対話型 AI の特質



AIの回答は、次の3点で注意が必要

指示（プロンプト）の出し方で答えは変わり、時には自信满满だが事実や実際の内容と異なる回答が起こることもある。

プロンプトの工夫で 結果が変わる

『AIについて教えて』
(ざっくりとした指示)

曖昧で漠然とした回答

『生成AIのしくみを
3つのポイントで説明して』
(具体的な指示)

整理された明確な回答

指示の出し方で
答えの質が大きく変わる

AIは自信满满に間違えること もある (ハルシネーション)

富士山の高さは？

5000メートルです

実際の正しい高さ：3776メートル

もっともらしく自信满满に答える
が、事実とは異なることがある
＝ハルシネーション

実際の内容とズレた提案を してしまうこともある

『Python入門の授業
レポートを考えて』

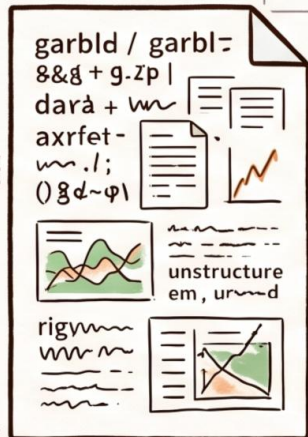
『try は信頼性向上に役立つ』
など、授業でまだ習っていない
内容を含めてしまう

実際の授業で習った内容について、
自分で考えて作成することで
実力が伸びるもの

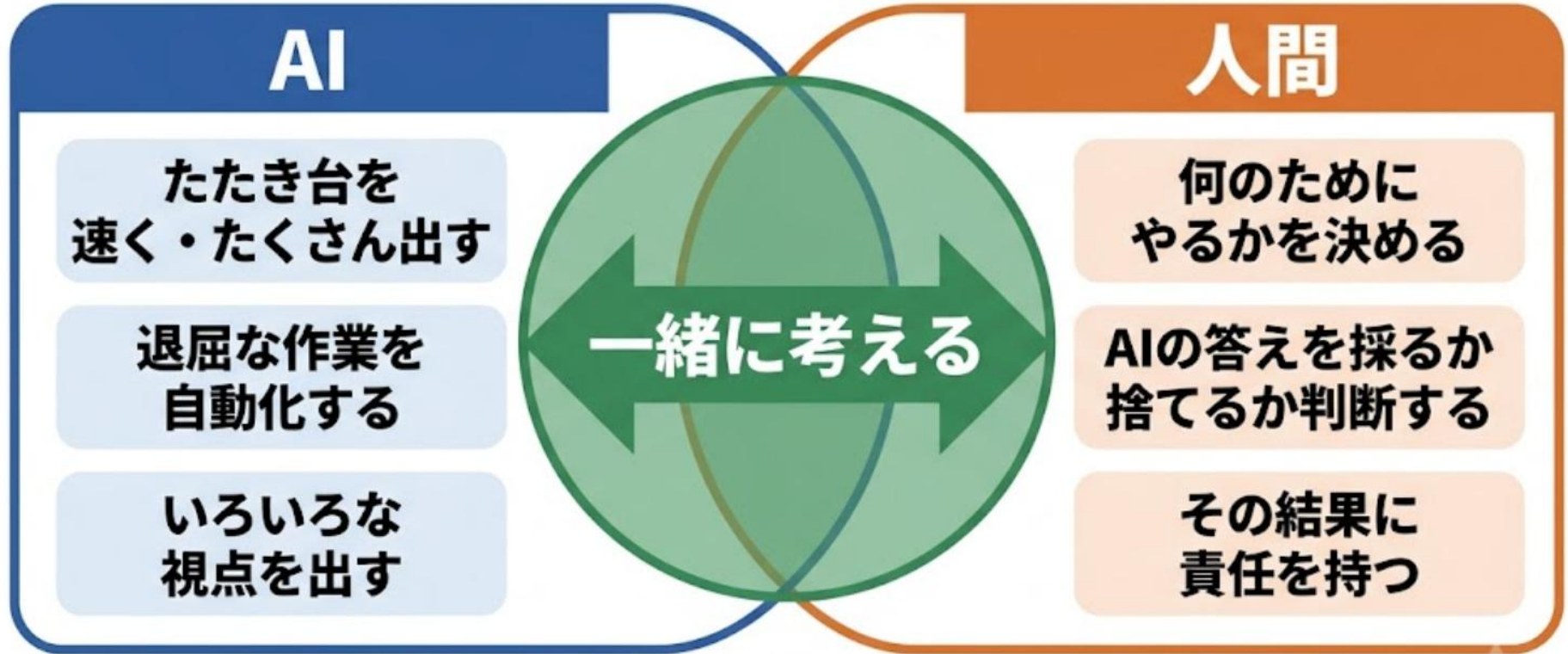
AIはあくまで相談相手であり、
正解をそのまま出す存在ではない。
内容は自分で確認・調整する

ハルシネーションとは、AIが誤った内容を、自信を持って正しいかのように述べてしまう現象を指す

同じデータでも、指示の出し方で
結果はこんなに変わるのか。
演習1・2・3、じっくり比べてみよう！

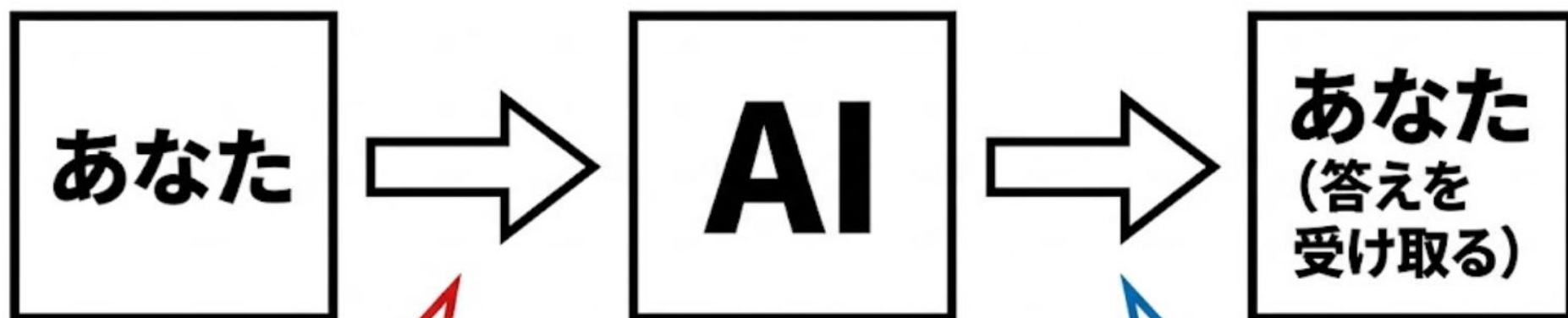


AIの得意と人間の役割を分けて、一緒に考える



AIに考えさせるのではなく、AIと一緒に考える

AI を使うときは、情報の『入れ方』と『答えの使い方』に注意が必要です



AI に入れてはいけない情報

- ・自分や他人の個人情報
- ・他人の秘密
- ・公開されていない機密情報

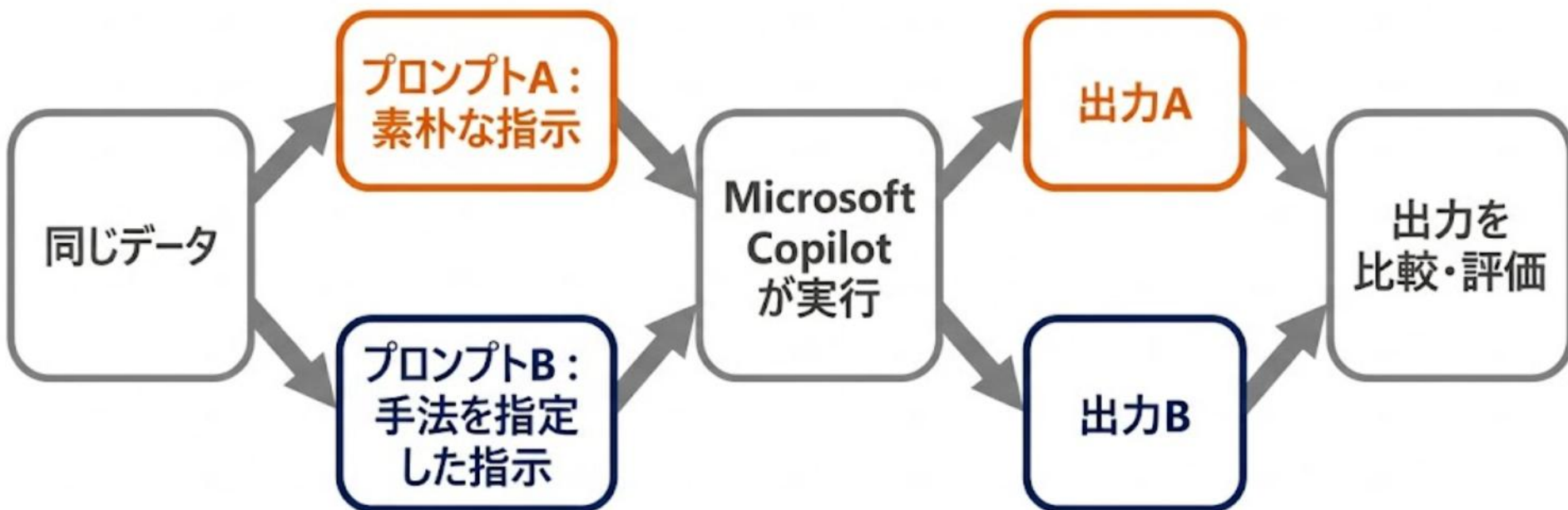
AI の答えを使うときに守ること

- ・AI の回答をそのまま信じない
- ・最終的判断は人間の役割

演習 1 , 2 , 3 では Microsoft Copilot を使用



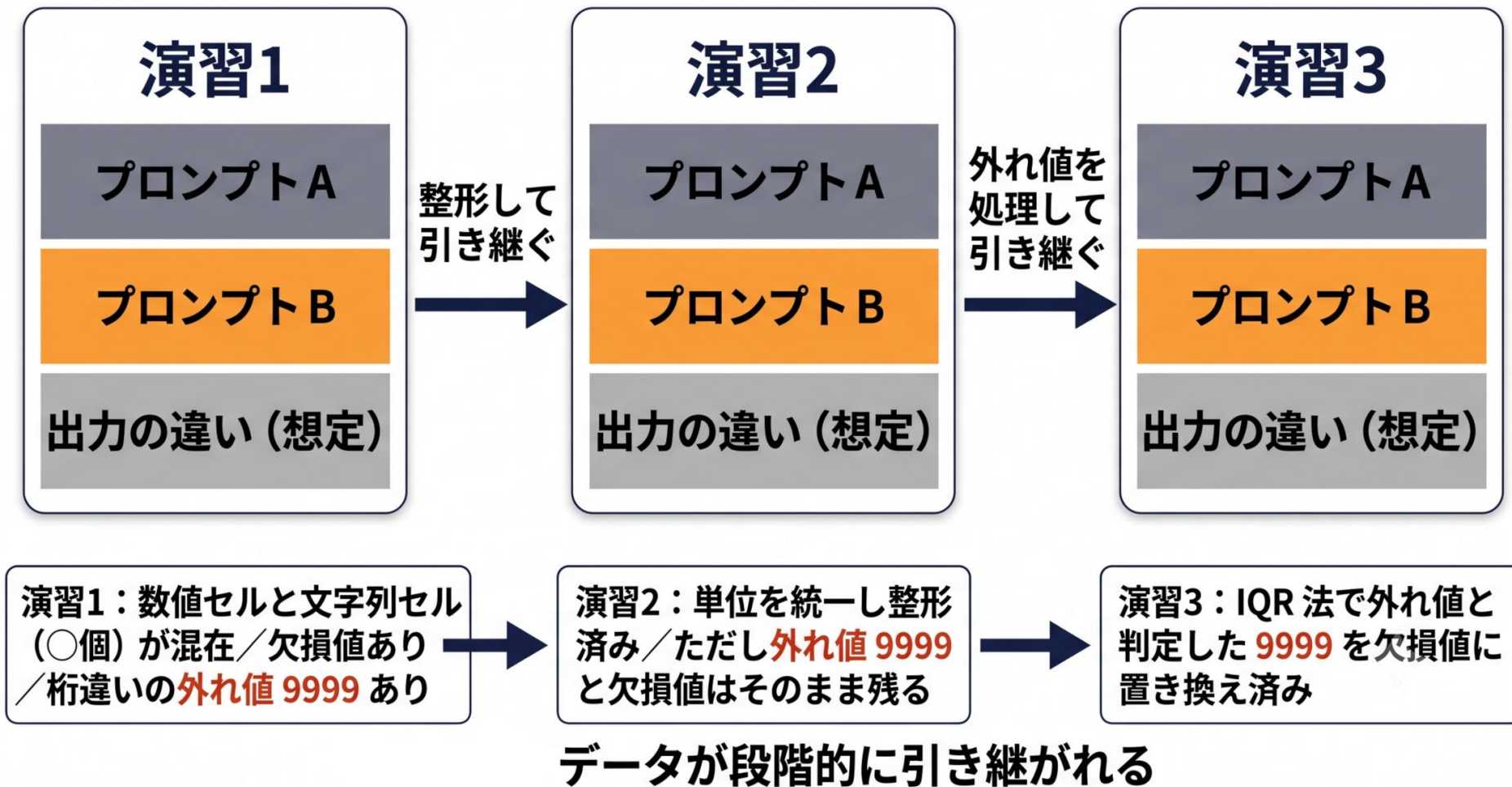
計算を実行するのはCopilot。あなたの役割は、指示による出力の違いを観察し、評価すること



Copilotの役割：計算・処理を実行する

あなたの役割：出力を観察し、比較・評価する

演習 1 , 2 , 3 の流れ



演習 1 . 表記ゆれを意識した指示



同じデータでも、AIへの指示の出し方で結果は変わる

生データ（たこ焼き店の売上記録）

日付	商品	個数	単価
2026/10/12	たこ焼き	15	300
10月12日	たこやき	10個	300
2026-10-13	たこ焼き	(空欄)	300
2026/10/13	タコ焼き	9999	300

…全30件

このデータに
潜む5つの問題

日付表記が
3系統で混在

商品名の
表記ゆれ

個数の型が混在

欠損値

桁違いの外れ値

プロンプトA：素朴な指示
このデータを分析して



問題を見落としやすく、処理基準
が指示者ごとに異なりうる

プロンプトB：手法を指定した指示

- 各列のデータ型を判定
- 表記ゆれを検出し統一
- 日付を統一形式YYYY-MM-DDに変換

問題を的確に検出して整形できる

指示が具体的なほど、AIの出力は正確になる

演習2. 外れ値、欠損値を意識した指示



同じデータでも、指示（プロンプト）の書き方で分析結果は変わる

元データ：たこ焼き店の売上記録

日付	個数
2022/7/20	66
2022/7/21	9999
2022/7/22	55
2022/7/23	
2022/7/24	19
...	...

外れ値

欠損値

外れ値と欠損値を含む

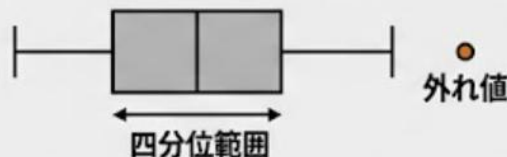
プロンプトA
素朴な指示

「売上傾向をグラフにして」
とだけ頼む

プロンプトB
手法を指定
した指示

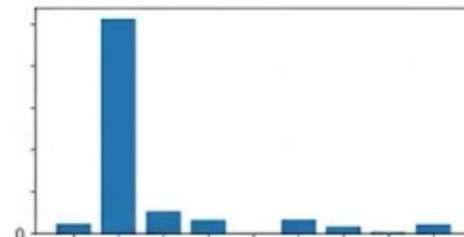
- ・ 平均・中央値・標準偏差を算出
- ・ IQR法で外れ値を検出して除外
- ・ 欠損値の件数と割合を報告し統計から除外
- ・ 日付ごとの推移を折れ線グラフで可視化

IQR法の考え方



四分位範囲の外側にある値を『外れ値』として検出する

1本だけ突出した棒グラフ



外れ値がそのまま反映され、他の日の値が相対的に小さく表示される。
欠損値の扱いは示されない

折れ線グラフ



外れ値を除いた平均が示される／
欠損値の件数・割合が明示される／時系列データが折れ線グラフとして表示される

『何を出すか』だけでなく『どう処理するか』を指示することで、再現性・妥当性のある結果が得られる

演習3. データエンジニアリングの実践



粒度（1日あたりの行数）を揃えてから結合すると正しく集計でき、揃えないまま結合すると売上が二重計上される

元データ：粒度が異なる2つの表

たこ焼き店：1日に複数行（明細）+欠損あり

日付	販売数
10-12	15
10-12	10
10-13	
10-13	

10-13は全行が欠損→合計を算出できない日

ジュース店：1日に1行

日付	販売数
10-12	30
10-13	12
10-14	40



A：そのまま日付で結合

日付	販売数	販売数
10-12	15	30
10-12	10	30
10-13		12
10-13		12

ジュース側の行が複製され、売上が過大に計上される

B：先に日次集計してから結合

集計済みの
たこ焼き店

日付	販売数
10-12	25
10-13	-

ジュース店日次表

日付	販売数
10-12	30
10-13	12
10-14	40

1-to-1

JOIN

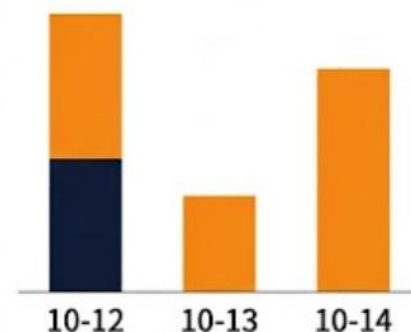
日付	たこ焼き店	ジュース店
10-12	25	30
10-13		12
10-14		40

粒度が一致し、日別の全店合計（個数×単価）を正しく算出できる

欠損は集計から除外

正しい統合テーブルから分析へ

■ たこ焼き店 ■ ジュース店



記述的分析の一例：
日別の全店売上と内訳を把握

ポイント：結合前に粒度（グレイン）を揃える。結合キーは『日付』とし、欠損は集計から除外したうえで、全欠損日は別途報告する