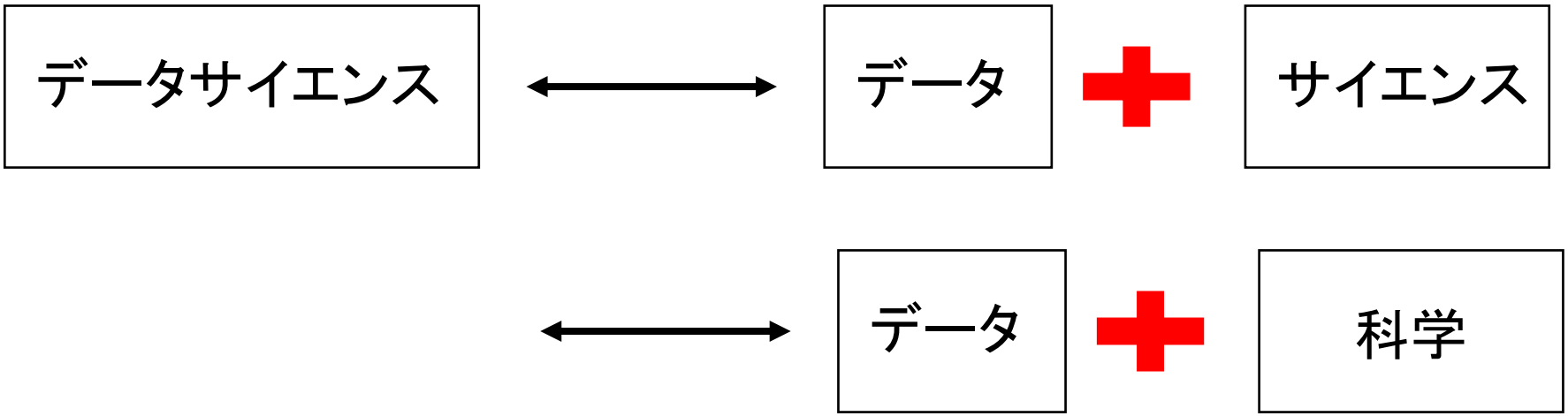
Excelで学ぶデータ分析

第１回

「データの要約と可視化」

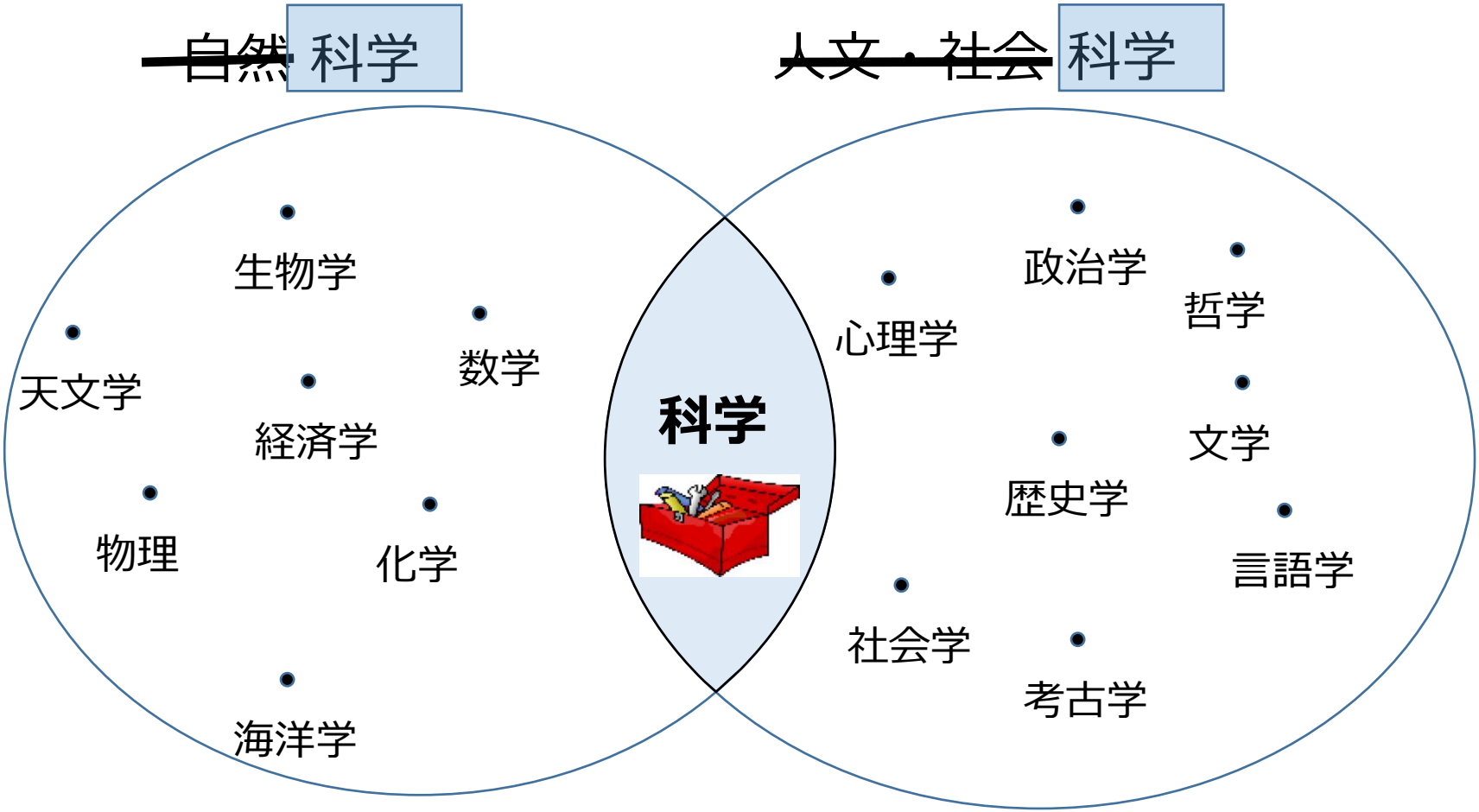
和から株式会社
WAKARA

データサイエンス

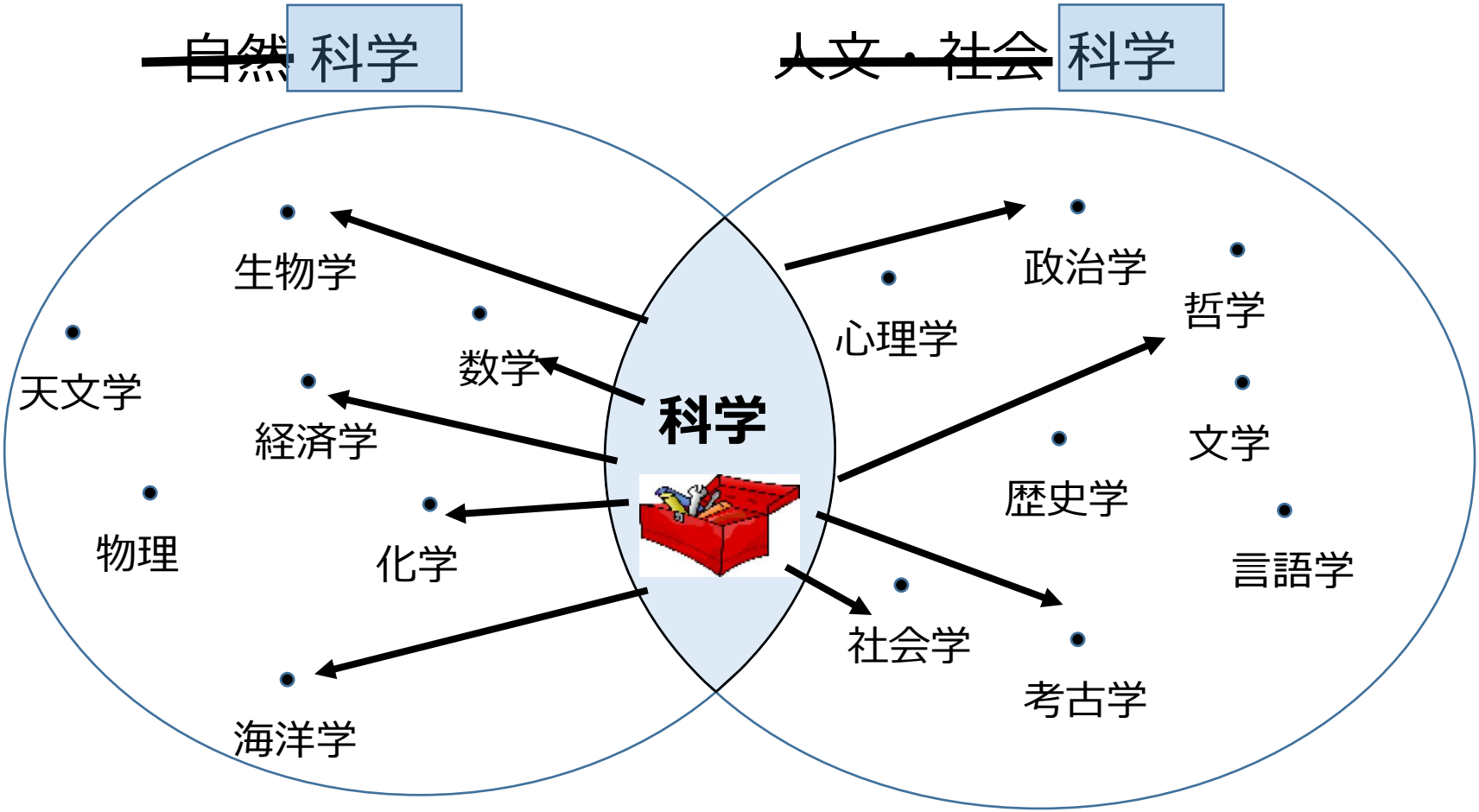


「科学する」とは？

科学するとは？



科学するとは？



科学理論の構造

定義/公理/仮説

F=ma

論理

直接法
背理法
帰納法
演繹法　etc

定理
エネルギー保存則

定理
力積保存則

定理

理論

統計検定の論理構造

仮説

背理法

結論

科学による問題解決方法とは？

「分解と統合」
の哲学

化学

医療

数学

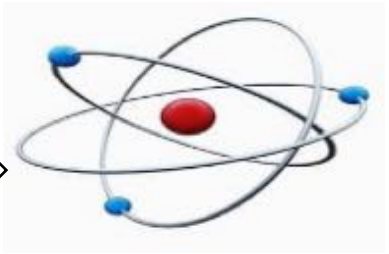
ビジネス

問題解決の為の
共通アプローチ？

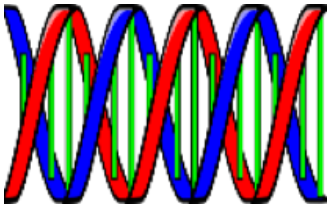
ルネ・デカルト
（1596-1650）

「分解と統合」の哲学

分子



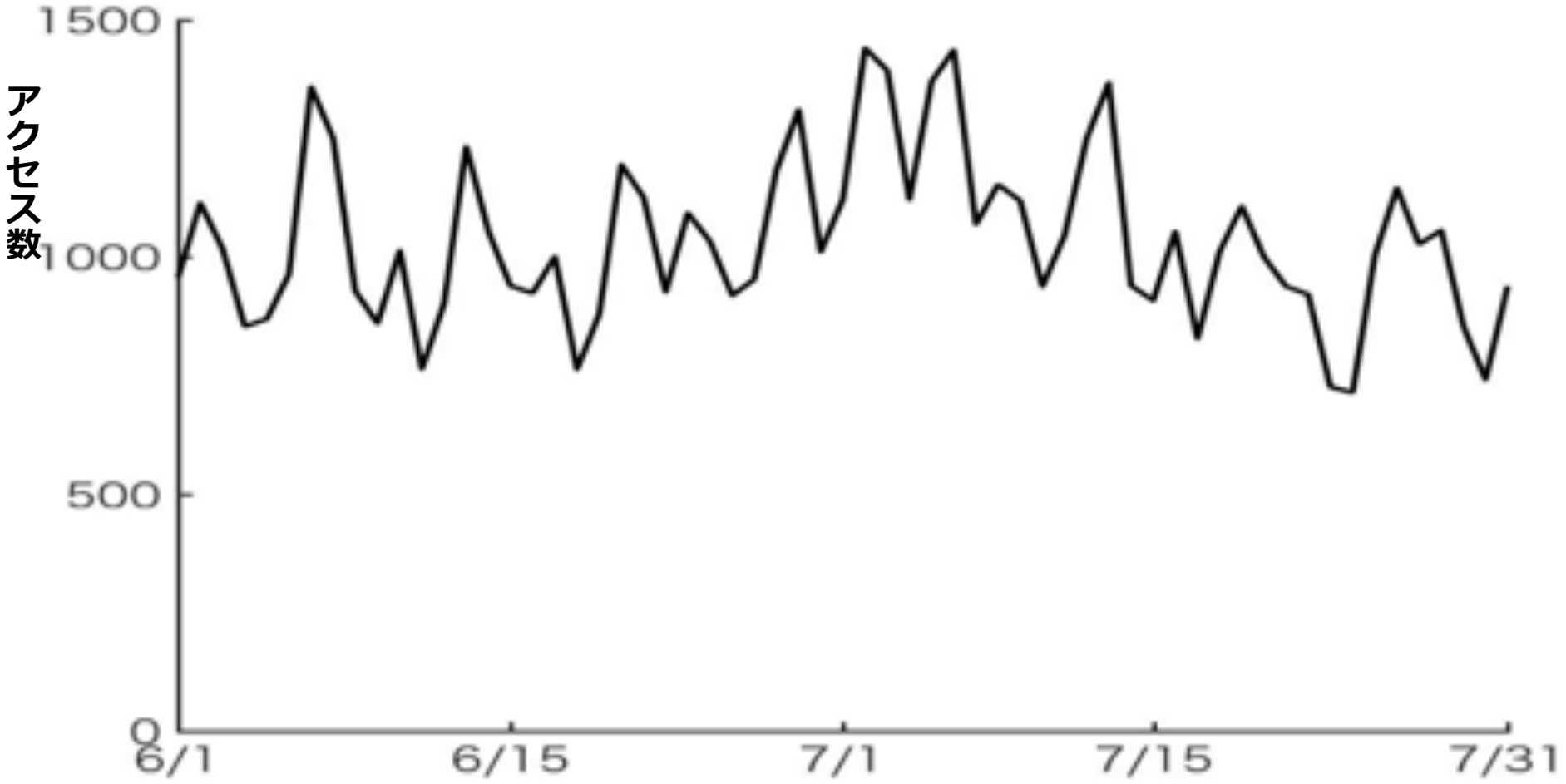
DNA



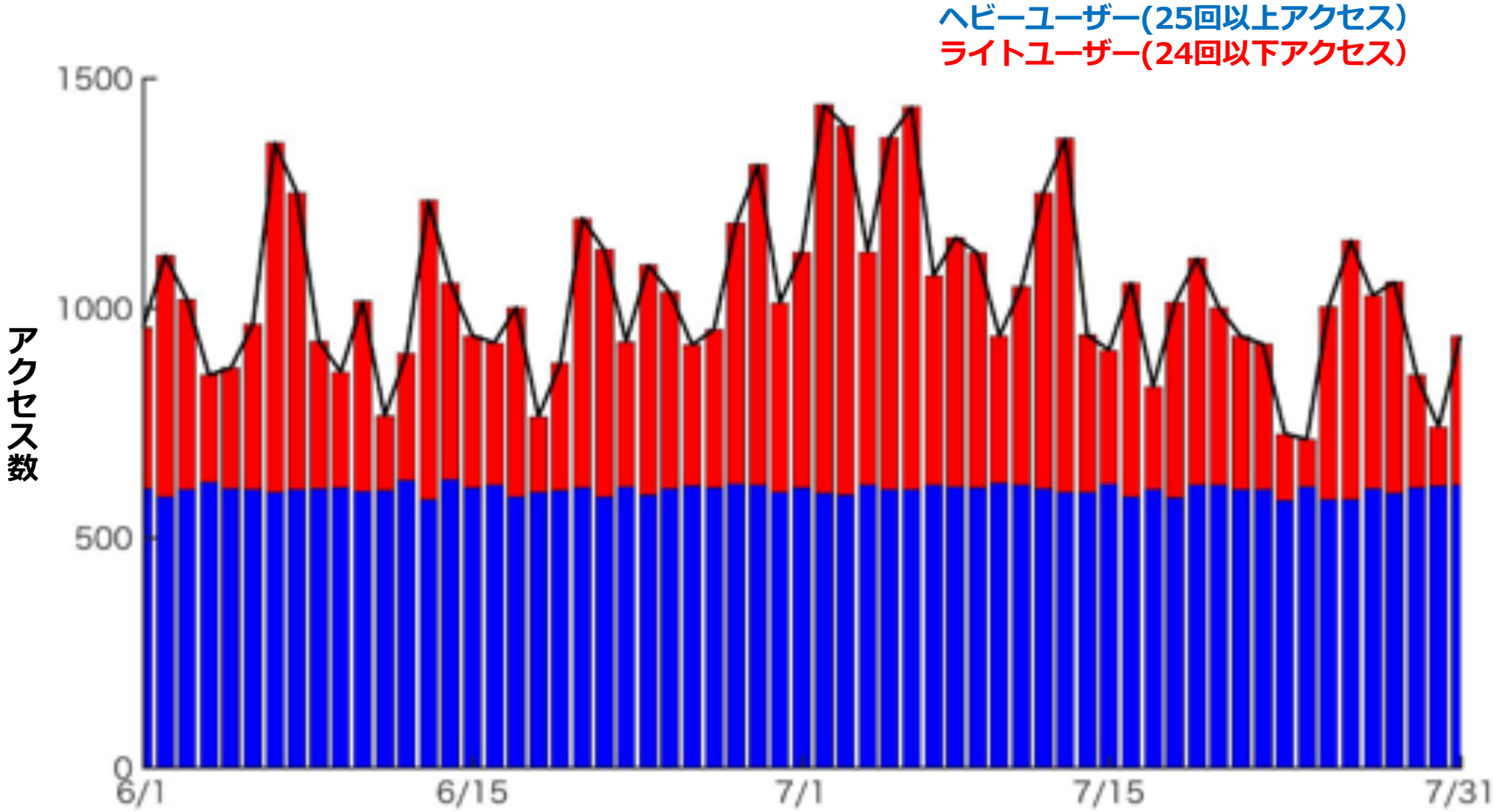
素因数分解　42　2, 3, 7　42=2・3・7

科学的視点による分析

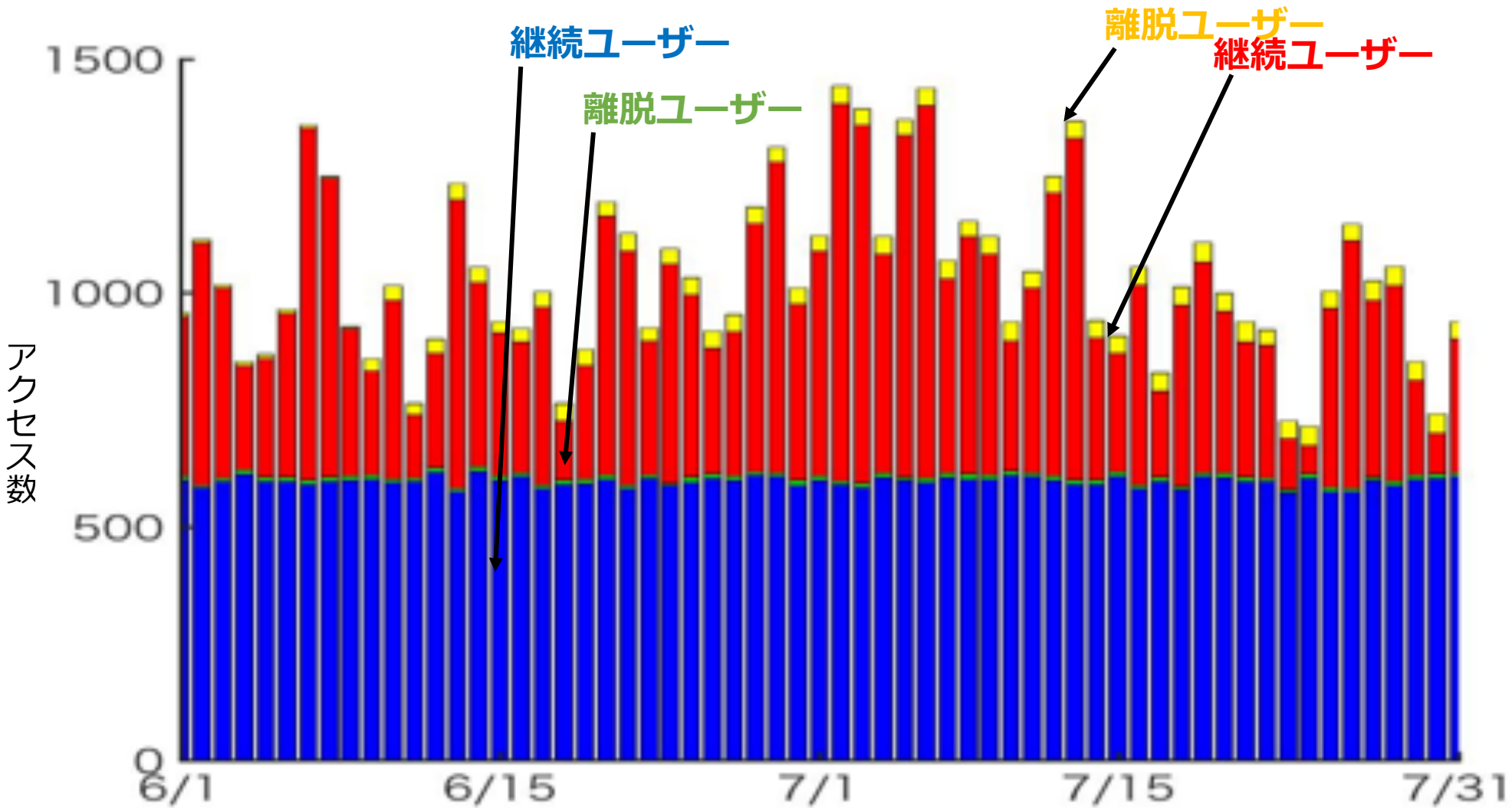
課題：「サイトへの登録者数が減少しているようだが、
アクセス数からその原因を調査できないか？」



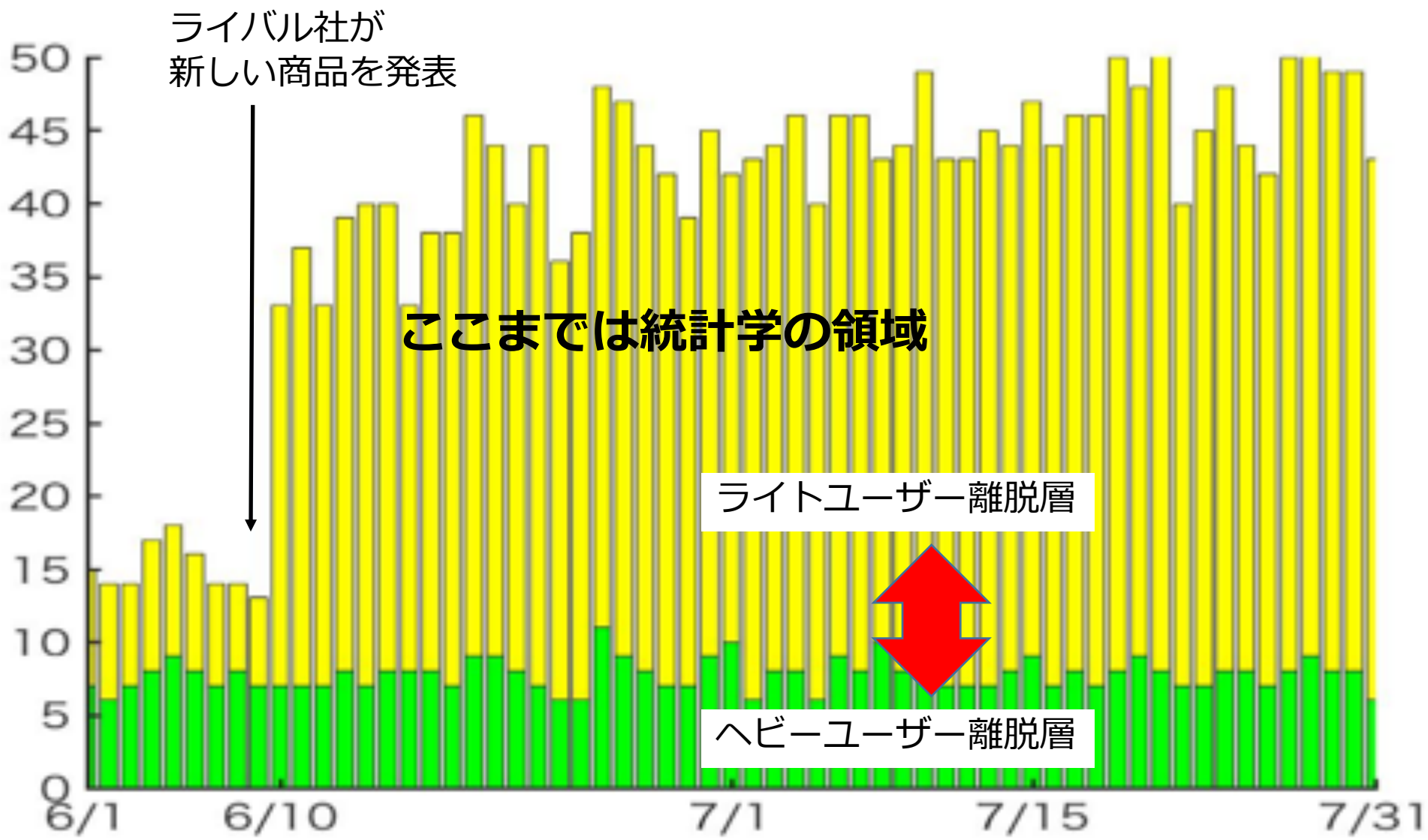
分解と統合



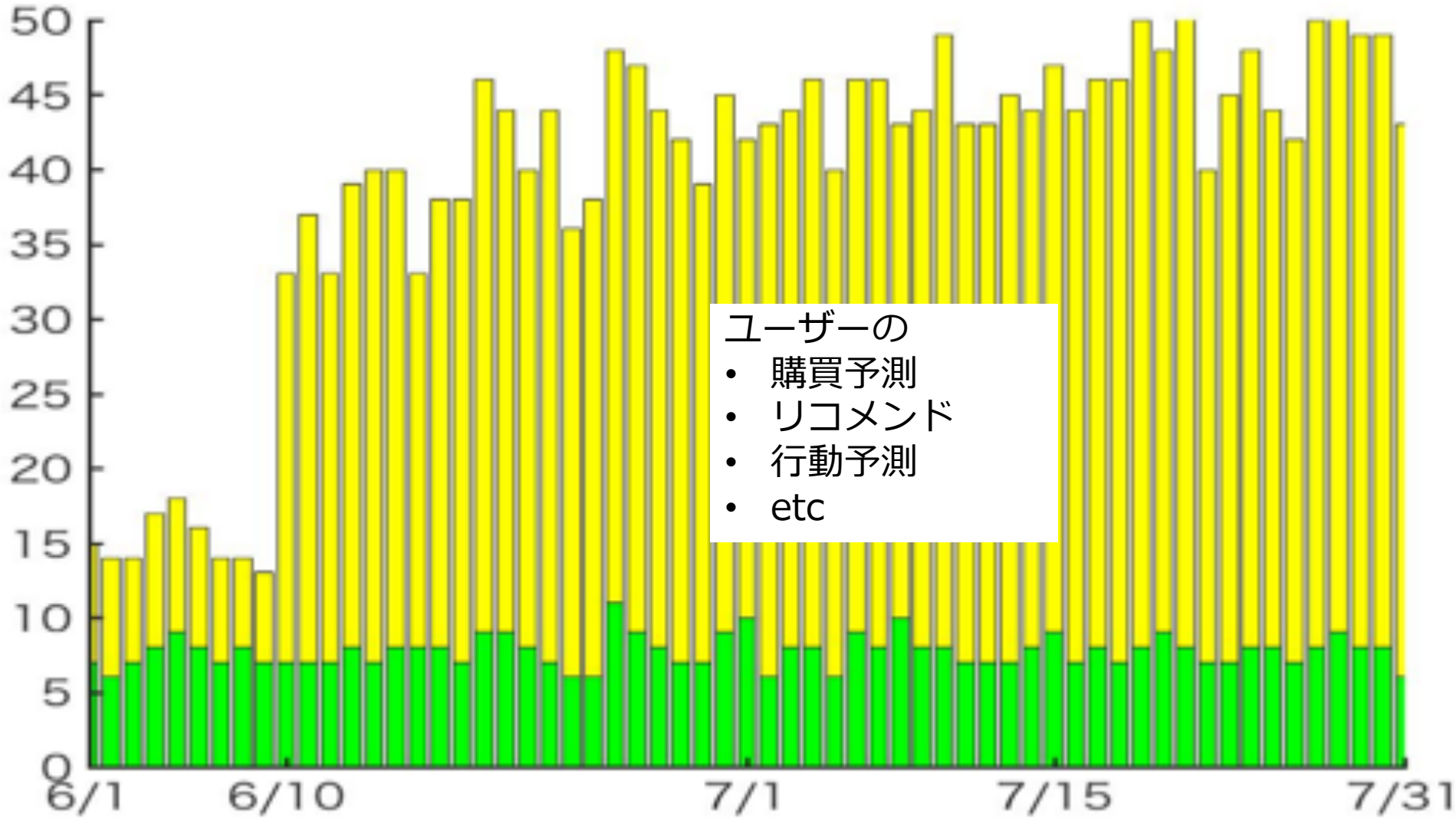
分解と統合



分解と統合

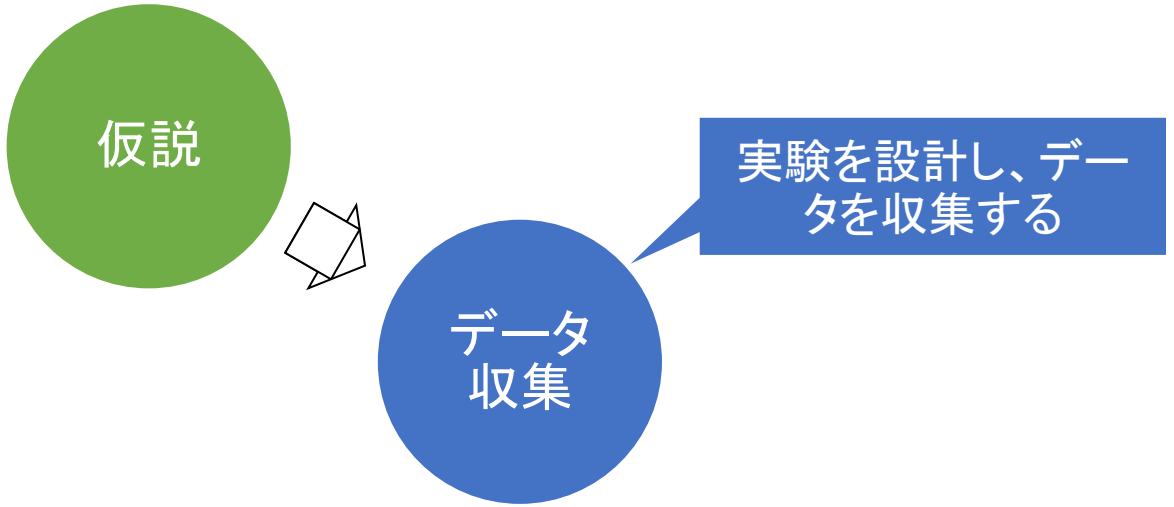


ここから先が機械学習

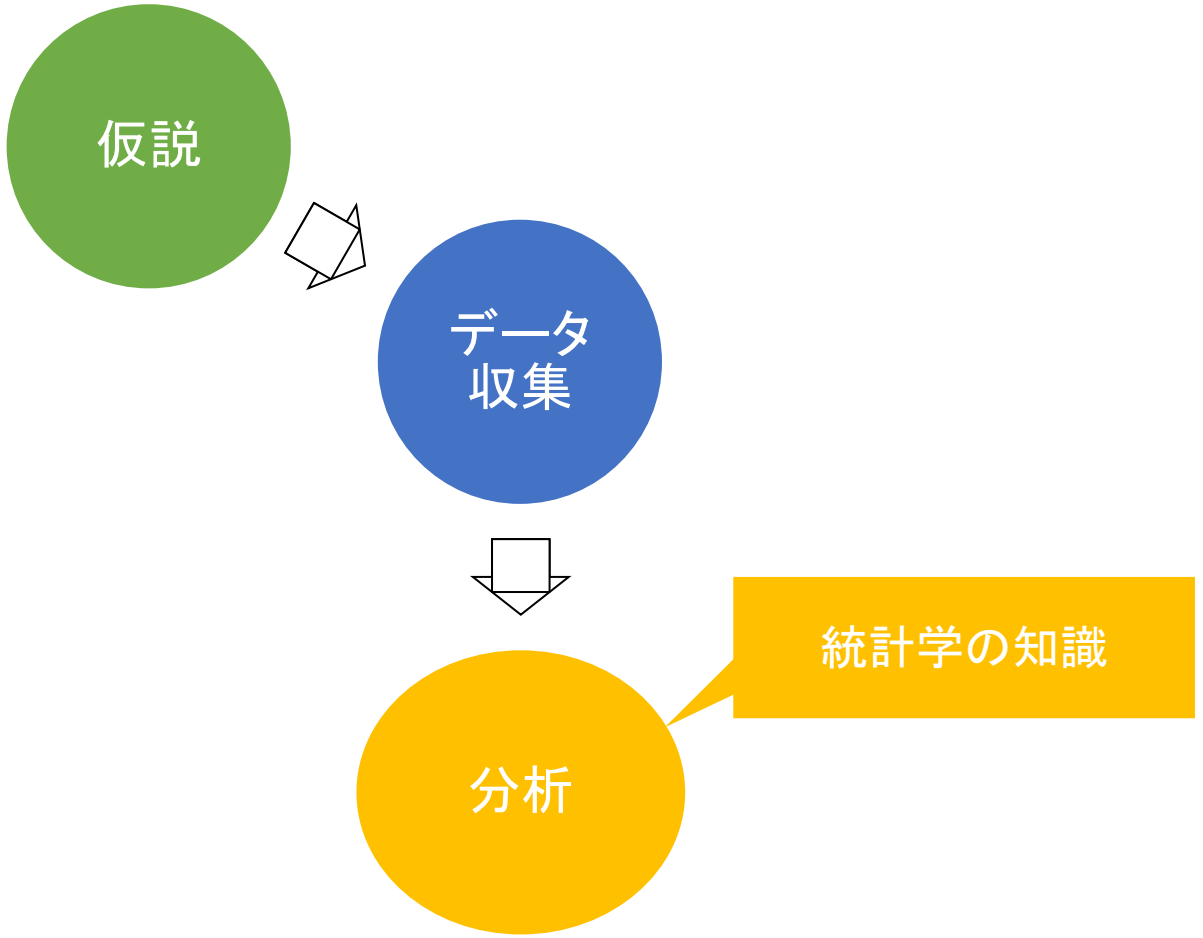


問題解決に必須なスキル

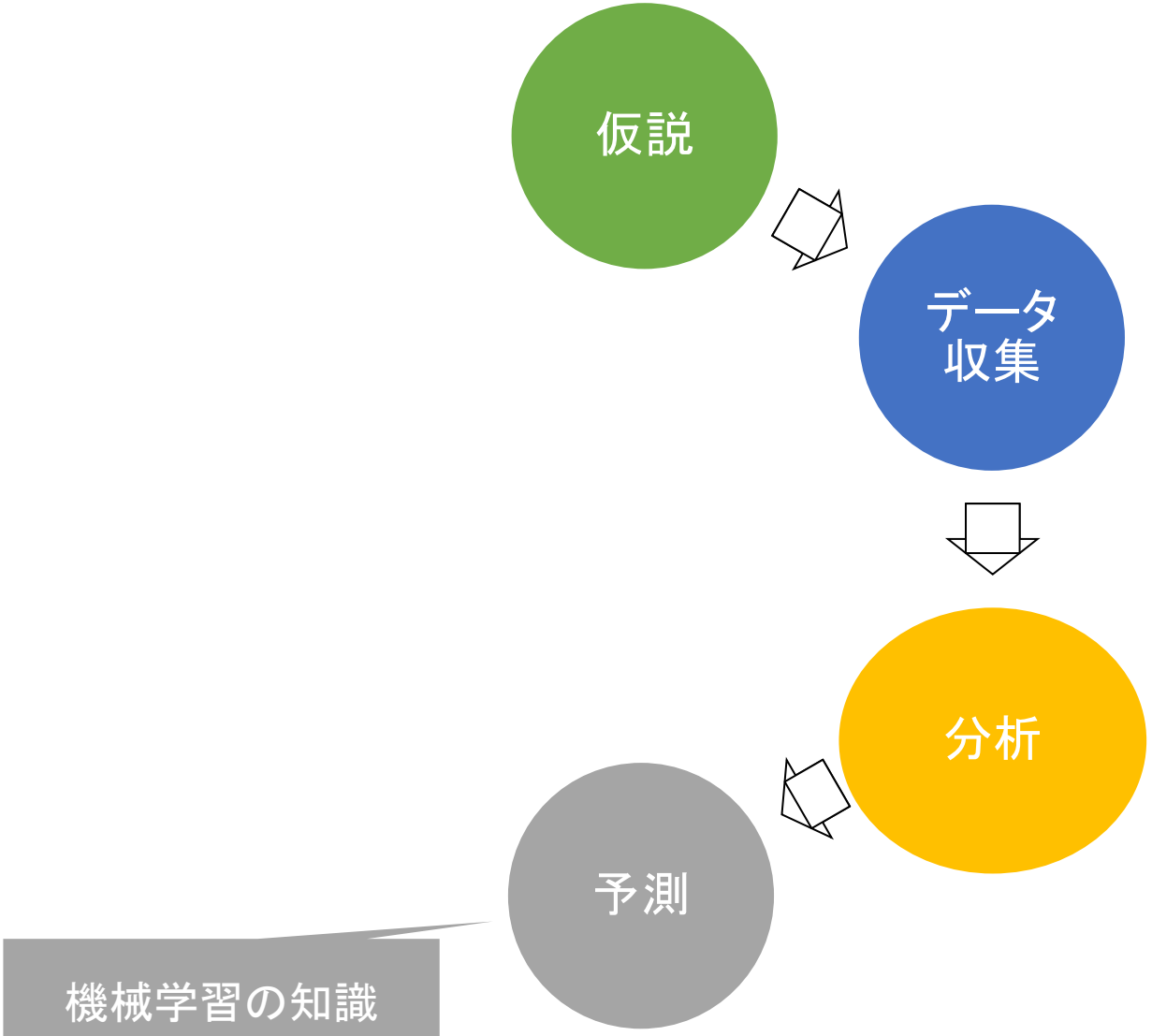
問題解決に必須なスキル



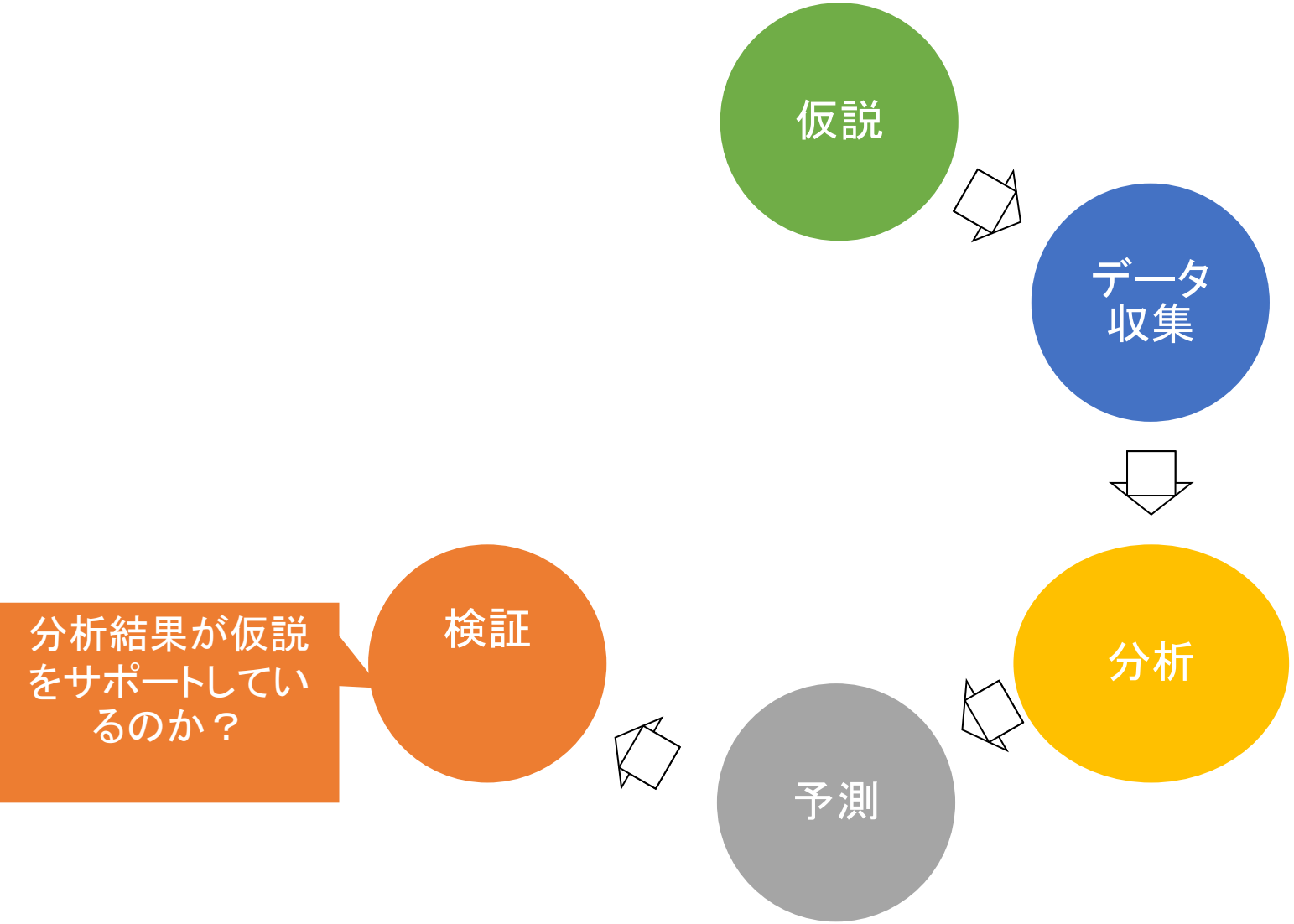
問題解決に必須なスキル



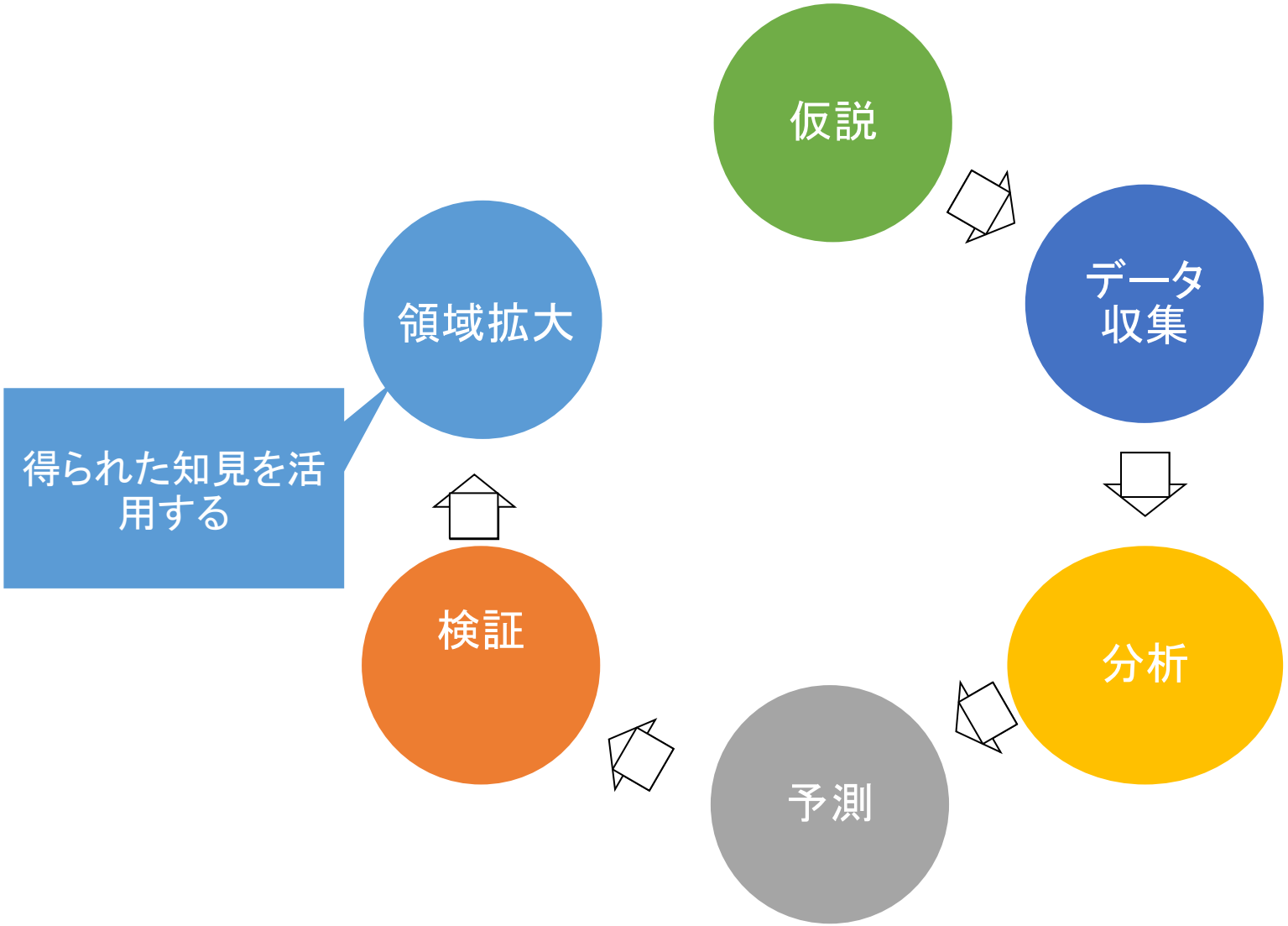
問題解決に必須なスキル



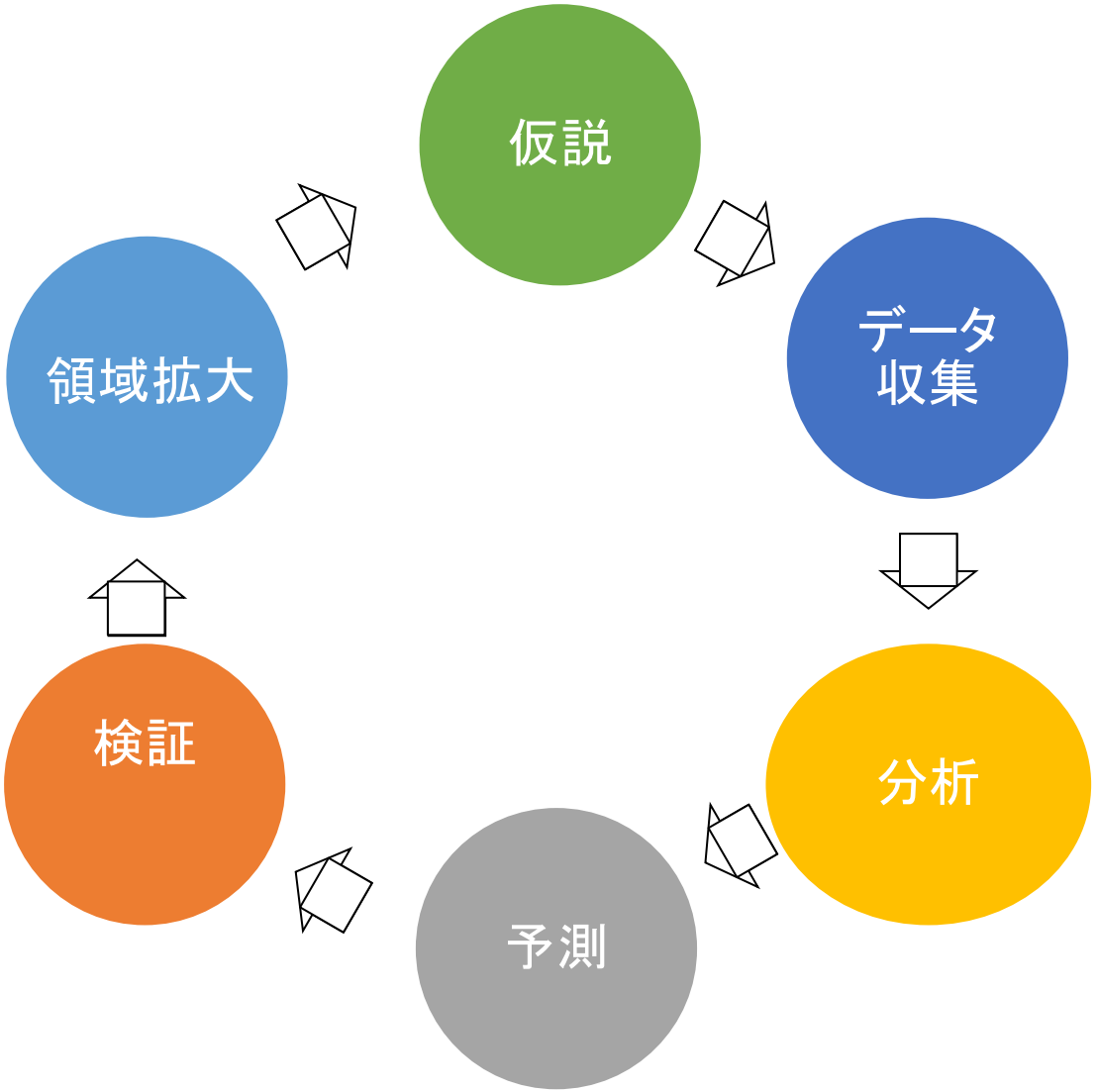
問題解決に必須なスキル



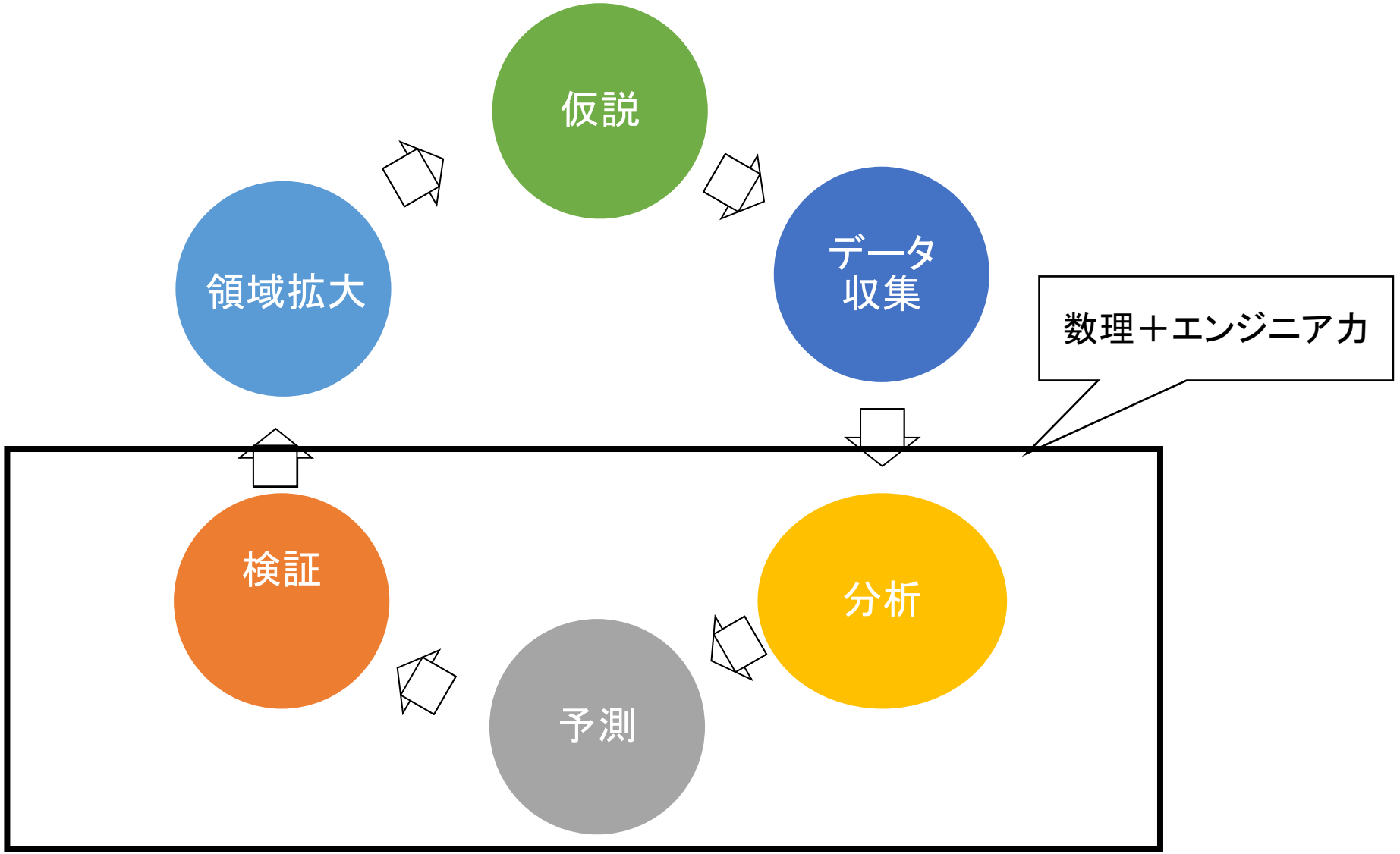
問題解決に必須なスキル



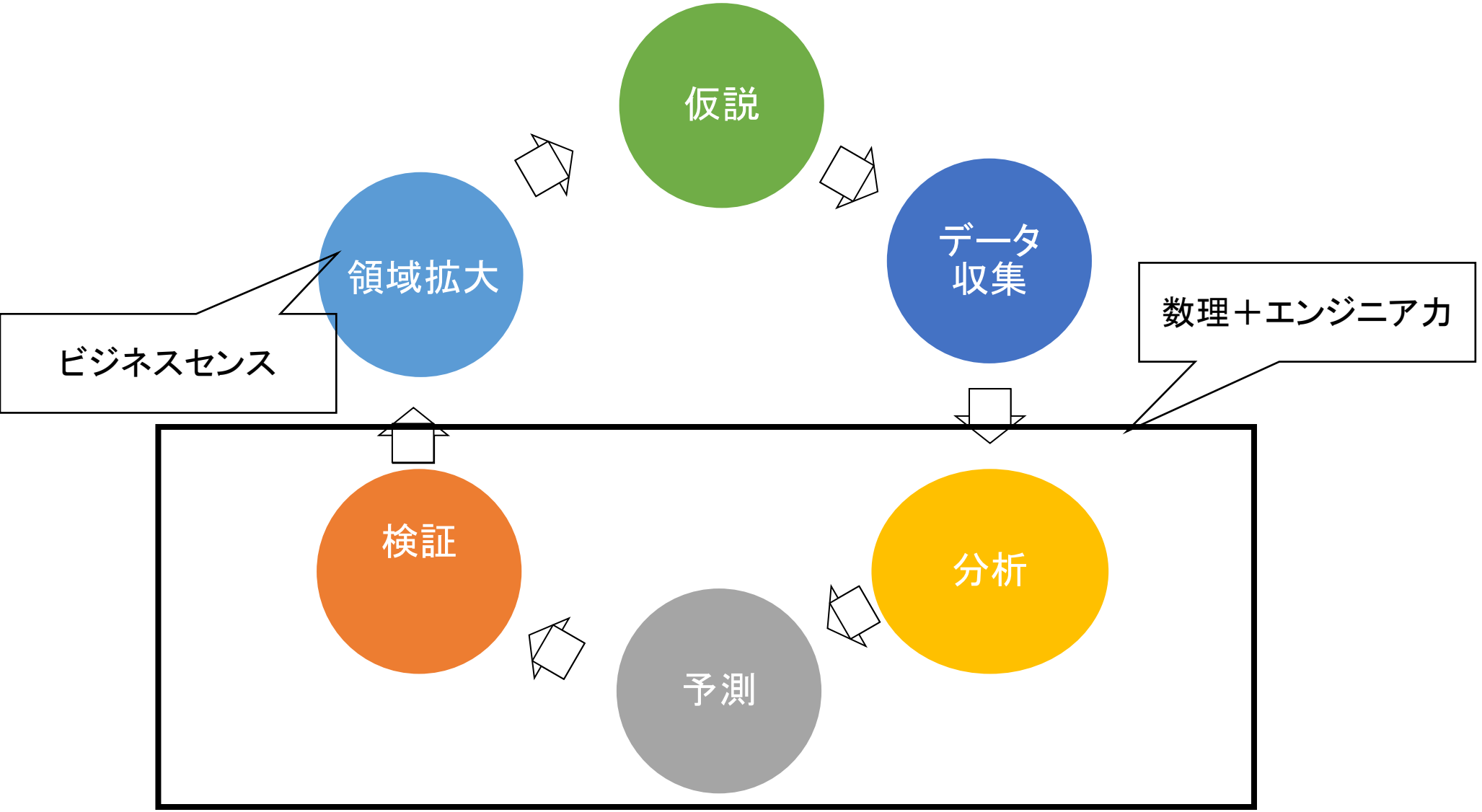
問題解決に必須なスキル



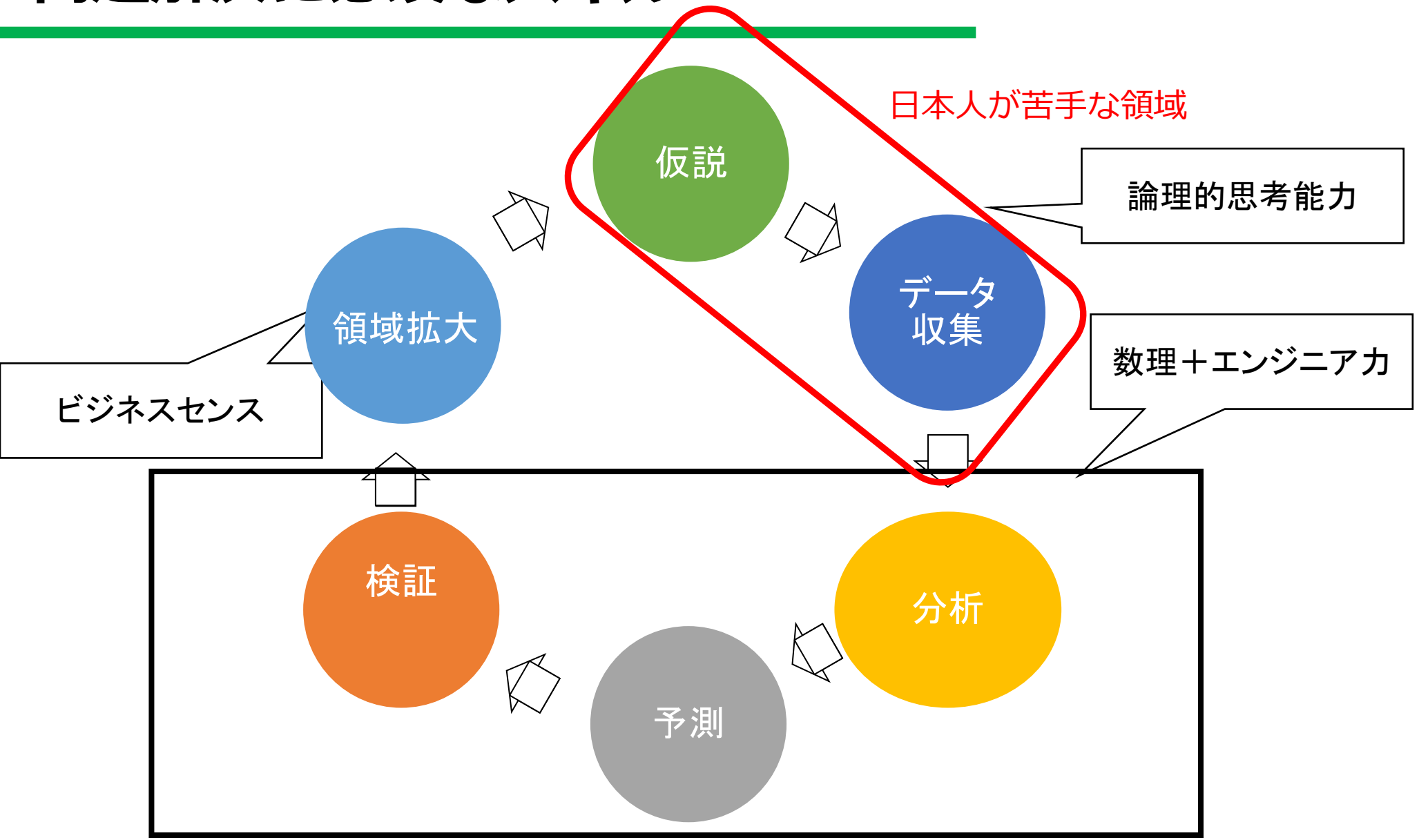
問題解決に必須なスキル



問題解決に必須なスキル



問題解決に必須なスキル



統計の３つの源流

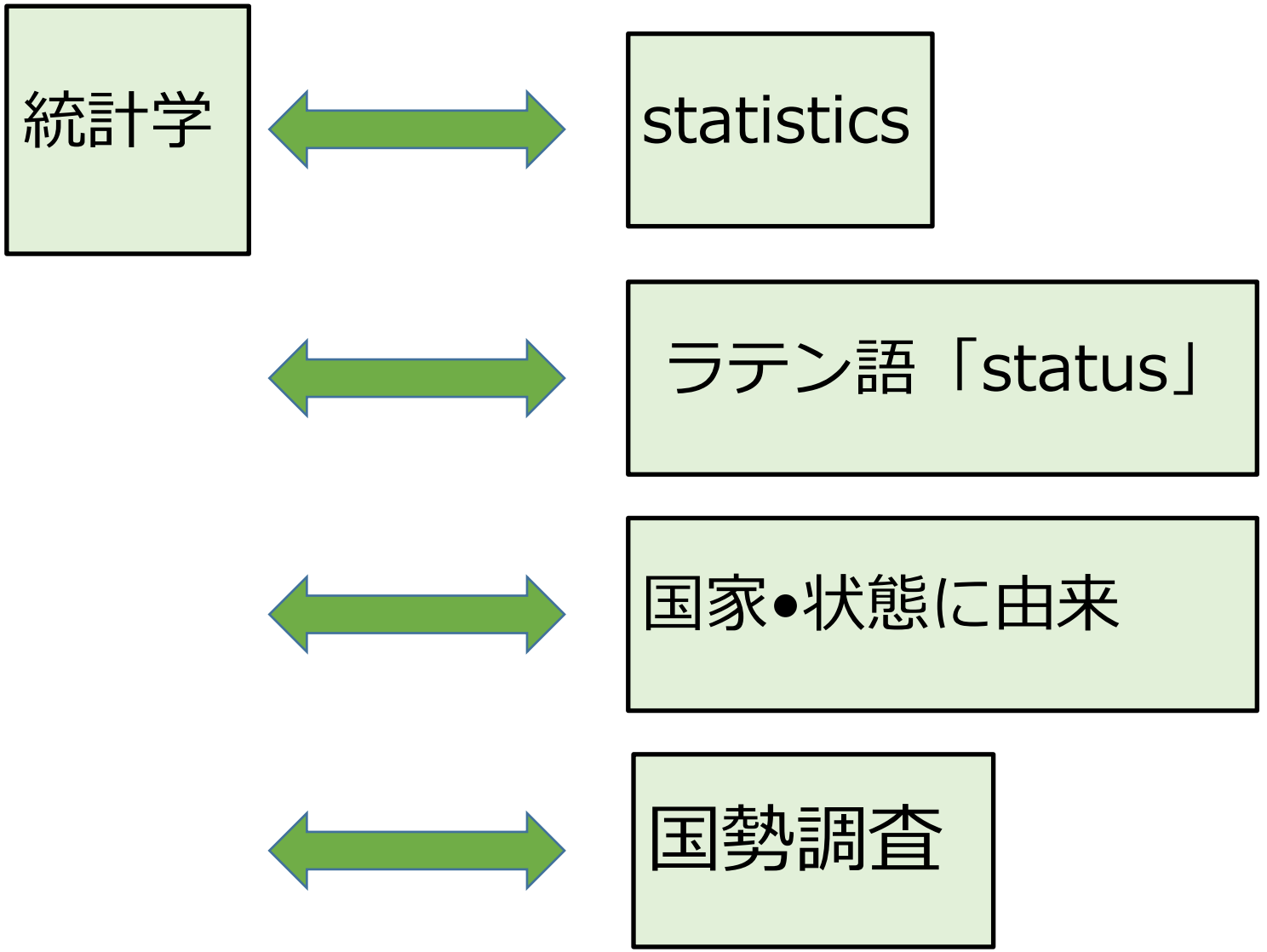
1.　国の実態をとらえるための「統計」

2.大量の事象をとらえるための「統計」

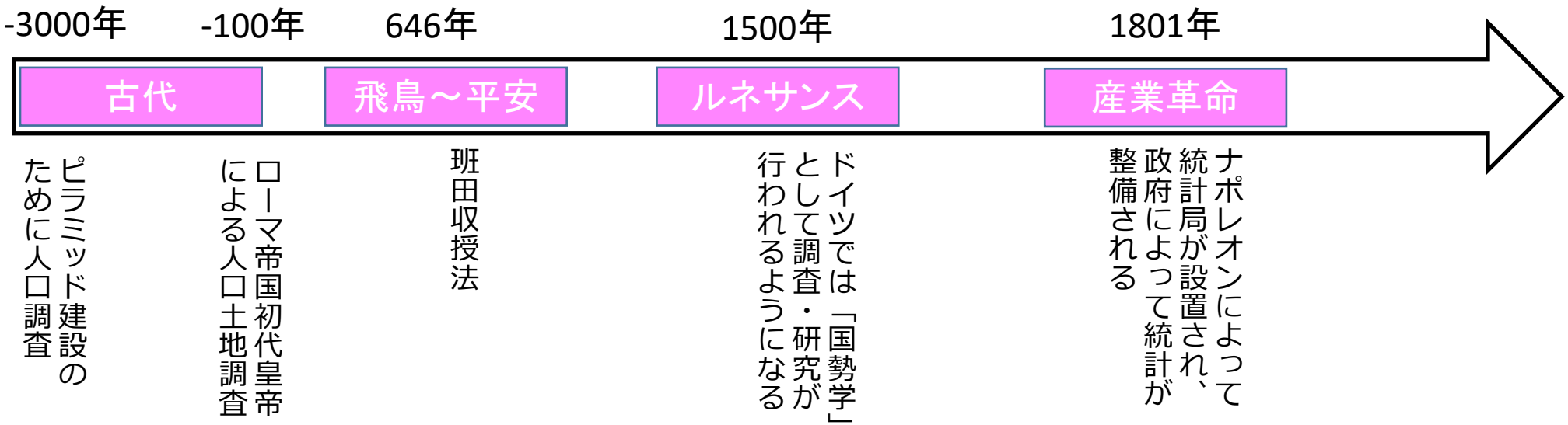
3. 確率的事象をとらえるための「統計」

現代統計学

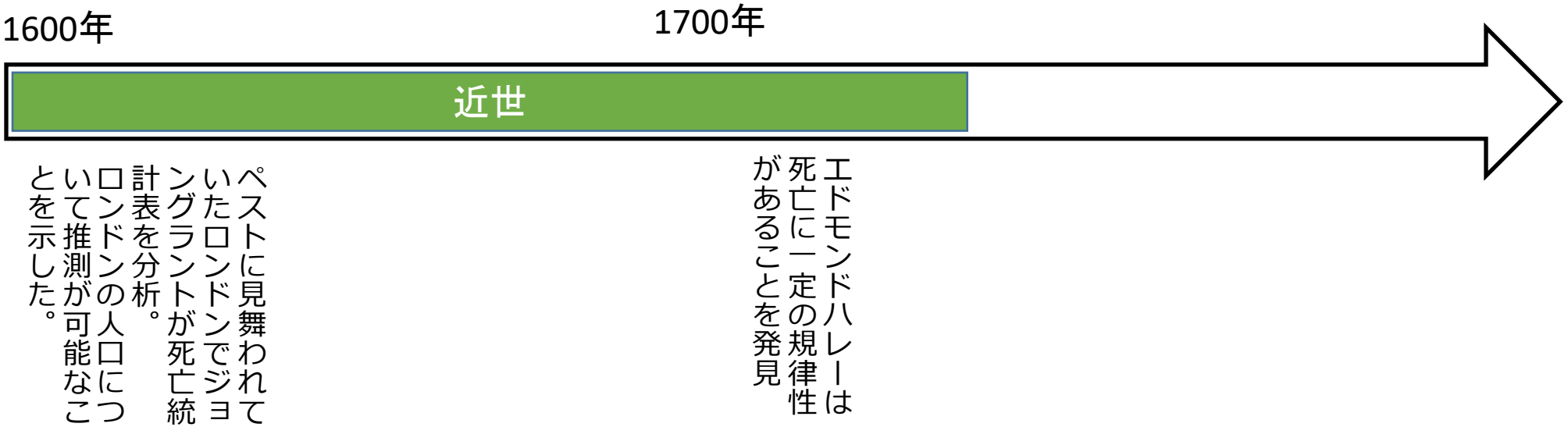
①国の実態をとらえるための「統計」



①国の実態をとらえるための「統計」



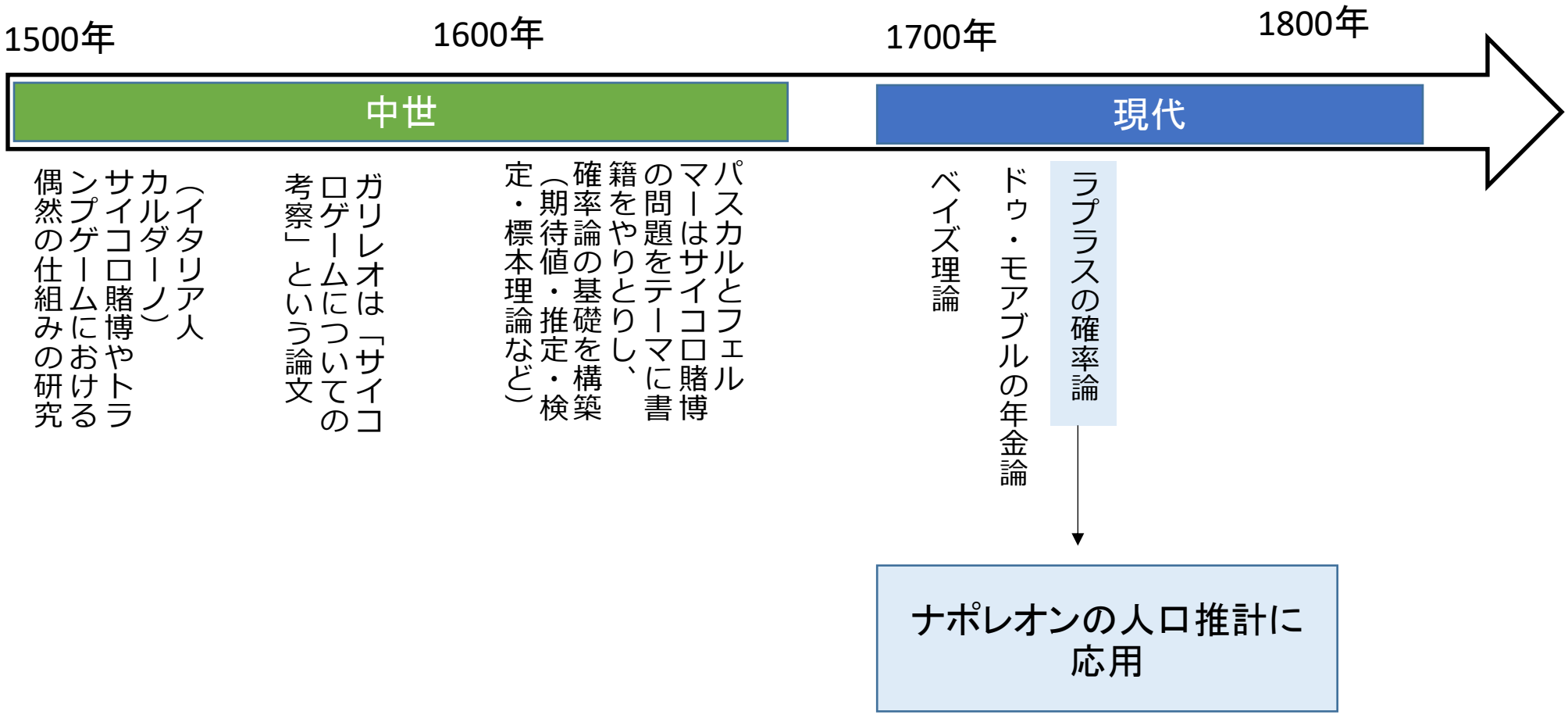
②大量の事象を捉えるための「統計」



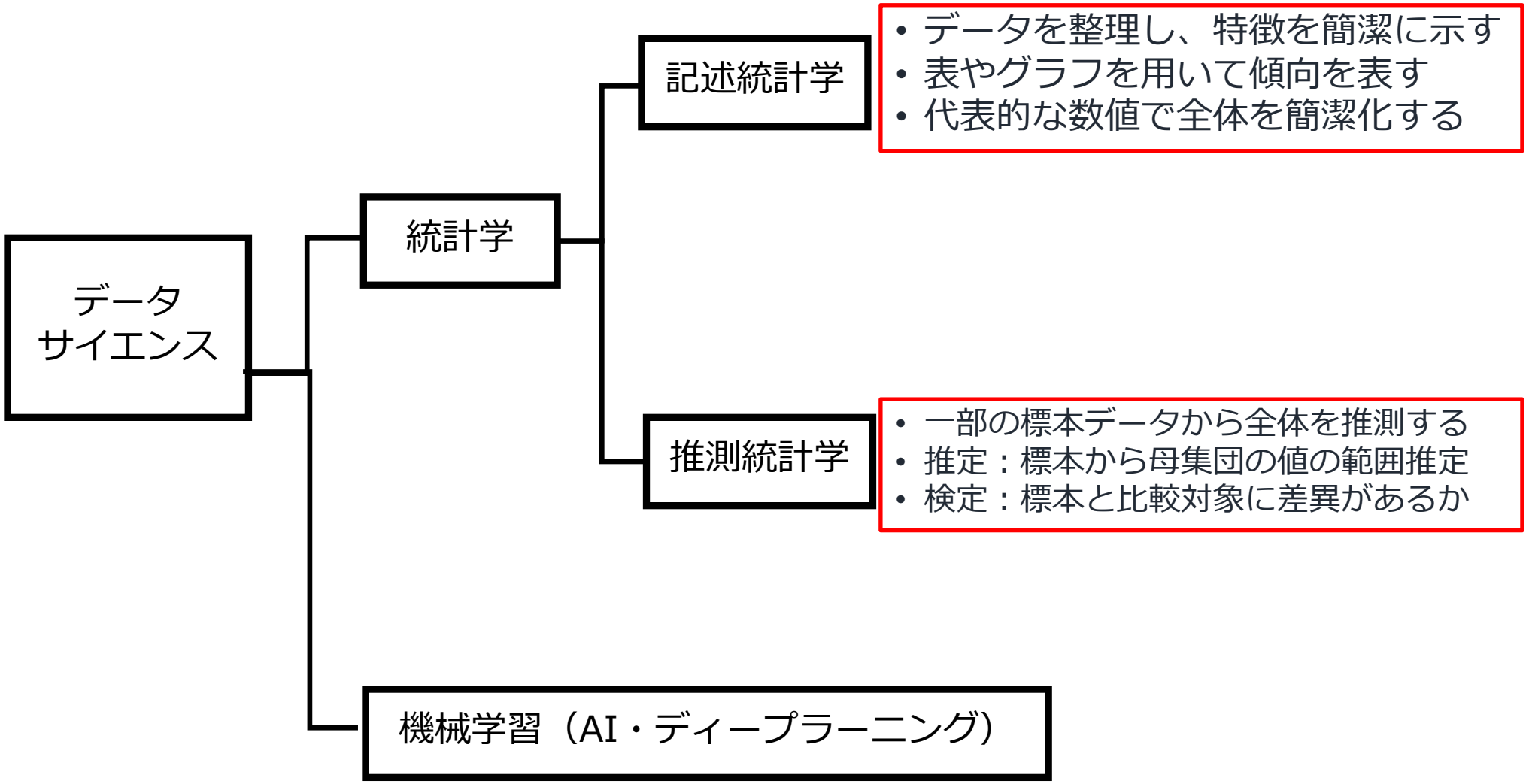
偶然と見られる現象に規律を探求する手法としての統計

「母集団」「標本」の概念

③確率的事象を捉えるための「統計」



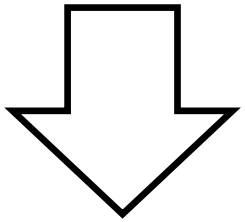
データ分析マップ



分析するとは？

データを分析する

① データの要約
② データ間の関係性
③ 予測する
④ 結果の検証



①～④のスキルを体得する

分析するとは？

Aさんの分析を依頼された

何を目的として分析するのか？

「結婚相手としてのふさわしいのか？」

「採用すべき人材なのか？」

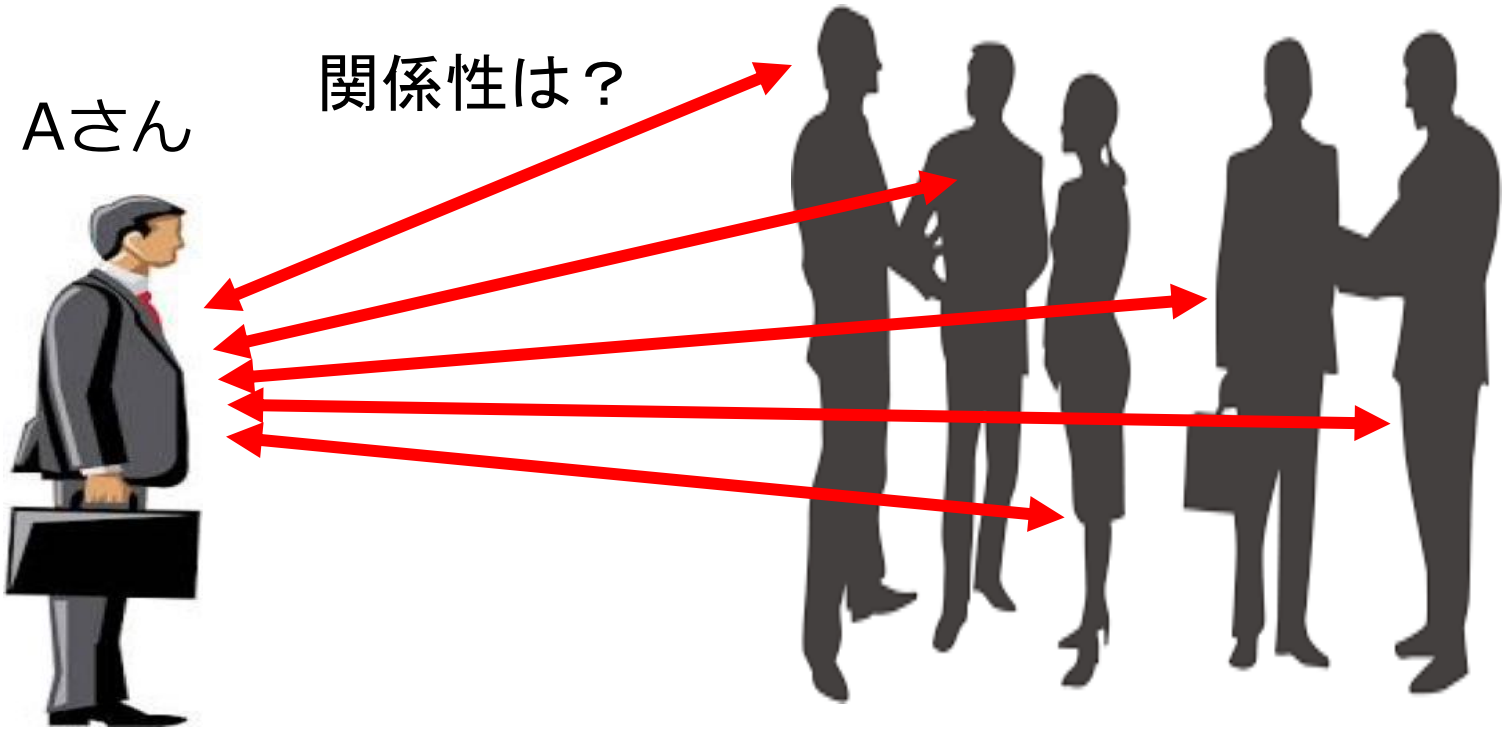
「どの部署に配属すべき人材なのか？」

「定着するのか？」

① データの要約

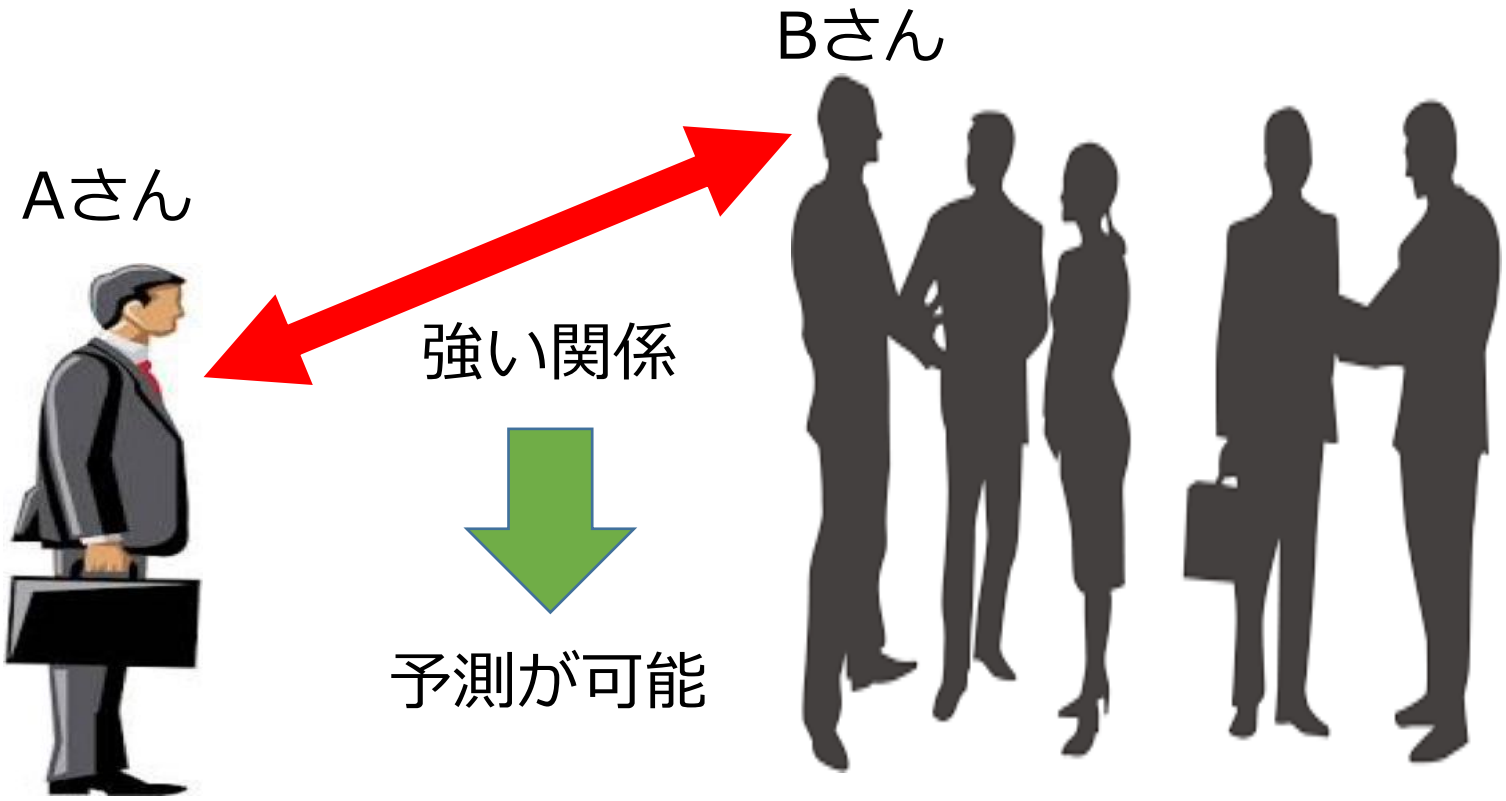
「Aさんを一言で言うとどんな人なのか？」

分析するとは？



②データ間の関係性
　Aさんとその友人・家族間などの関係を調べる

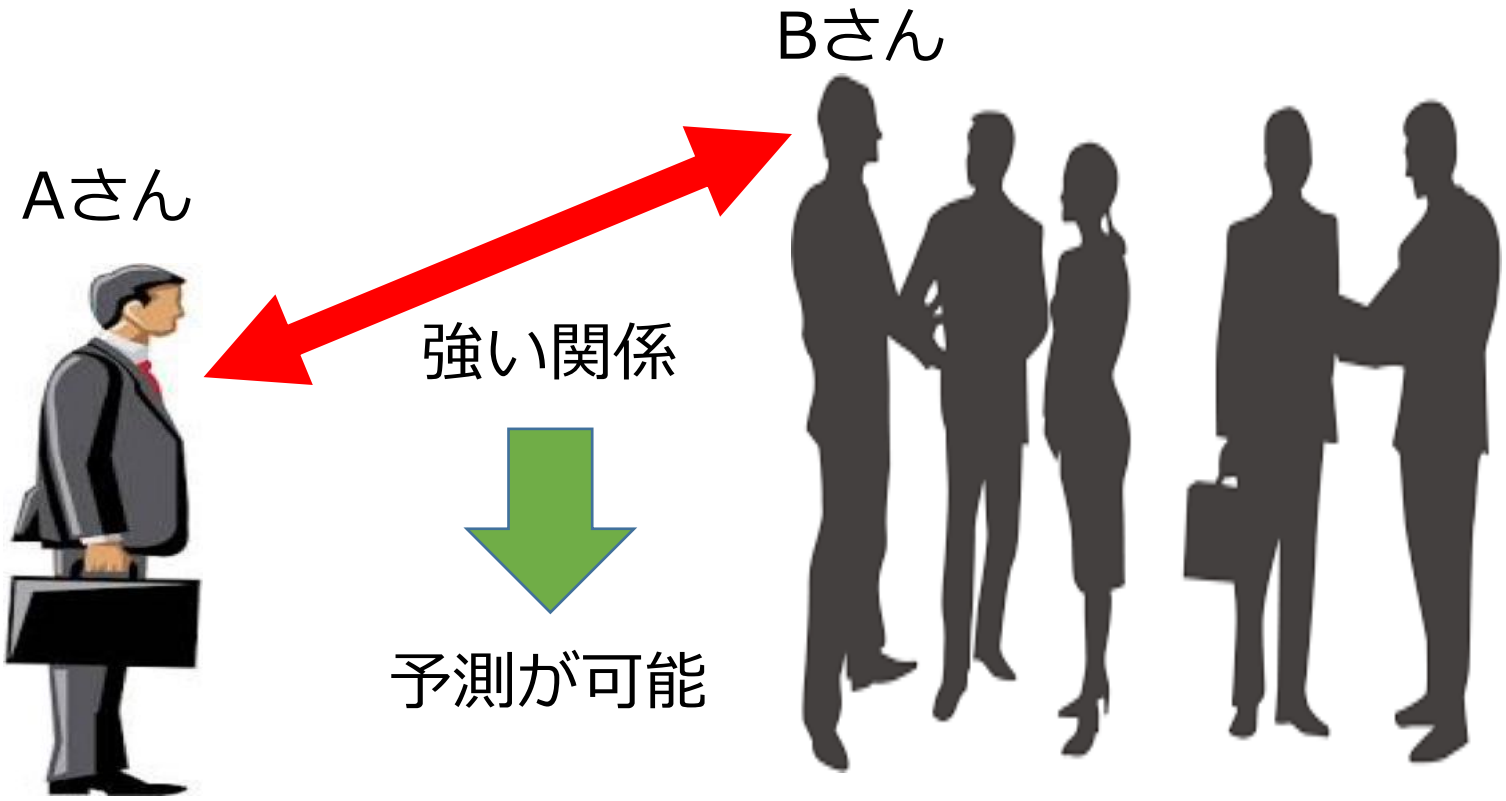
分析するとは？



③ 予測する

人間関係を把握することで、Aさんの行動を予測することが可能になる場合がある

分析するとは？



④信頼性の検証

自分の出した結論がどれくらい信頼できるものなのかを検証する必要がある。

データの分類

データ

質的データ

✓数値でないデータ（文字型）

量的データ

✓数値データ（整数型・実数型）

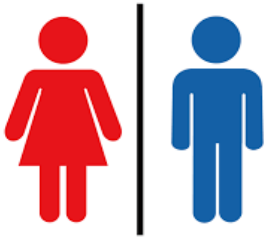
メール送信数

（１５通）

資格の有無

(有・無)

性別のデータ

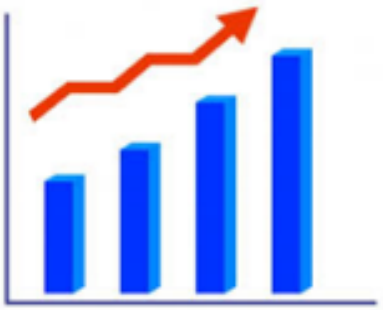


（男性・女性）

適性検査の結果

（適正・不適正）

売上



(56.7億円)

量的データと質的データの要約

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職か在職	過去５年昇進（有無）	所属部署	給料
1	0.38	0.53	2	157	3	0	退職	無	sales	low
2	0.8	0.86	5	262	6	0	退職	無	sales	medium
3	0.11	0.88	7	272	4	0	退職	無	sales	medium
4	0.72	0.87	5	223	5	0	退職	無	sales	low
5	0.37	0.52	2	159	3	0	退職	無	sales	low
6	0.41	0.5	2	153	3	0	退職	無	sales	low

数量データ

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差
- ２５%、７５点
- ヒストグラム

質的データ

- 円グラフ・ヒストグラムetc
- クロス集計

量的データと質的データの要約

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職か在職	過去5年昇進（有無）	所属部署	給料
1	0.38	0.53	2	157	3	0	退職	無	sales	low
2	0.8	0.86	5	262	6	0	退職	無	sales	medium
3	0.11	0.88	7	272	4	0	退職	無	sales	medium
4	0.72	0.87	5	223	5	0	退職	無	sales	low
5	0.37	0.52	2	159	3	0	退職	無	sales	low
6	0.41	0.5	2	153	3	0	退職	無	sales	low

質的なのか量的なのか？

労災（有） 1
労災（無） 0

質的変数として扱う必要がある

量的データの要約

A社とB社の１週間における売上のデータです。どちらかの店に融資するとしたら、どういう理由でA社もしくはB社を選びますか？

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

データを要約するために使う統計量

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	1400	50	20	120	240	100	-79
B社	87	970	120	104	112	112	117

どの統計量から計算する？

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差

最大値・最小値から何が分かる？

1. データに異常値がないのか把握するために使う。
2. データの上限・下限を把握することで、データに関する理解を深めるために使う。

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

どの統計値から計算する？

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差

最大値・最小値から何が分かる？

1. データに異常値がないのか把握するために使う。
2. データの上限・下限を把握することで、データに関する理解を深めるために使う。

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

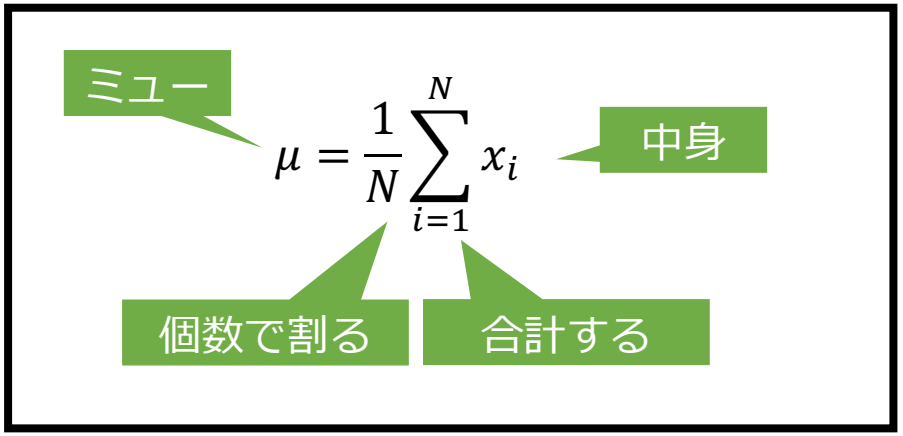
どの統計値から計算する？

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差

1. データの中心を把握するために使う
＝データを代表する値は？

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117



A社の平均値を求める

$$\frac{140 + 50 + 20 + 120 + 240 + 100 + 79}{7} = 107$$

平均値に騙されるな！

平均値に騙されるな！

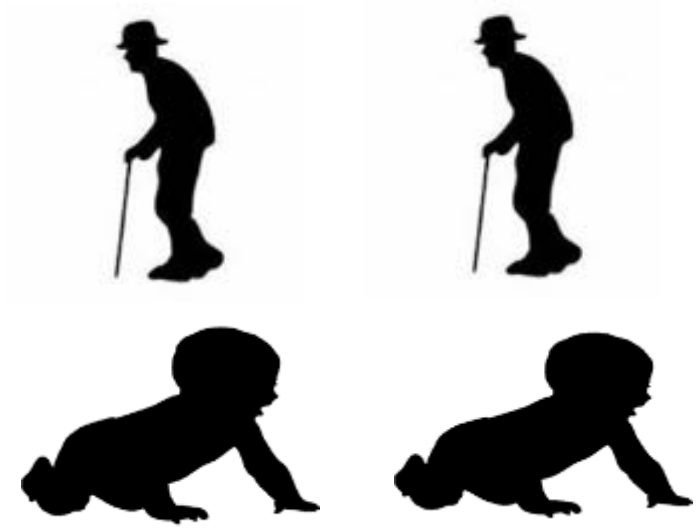
２０兆５９００億円の資産を保有
（ The Richest.com ）

平均資産１億円はベントンビル市住民の資産を代表している値ではない

平均値に騙されるな！

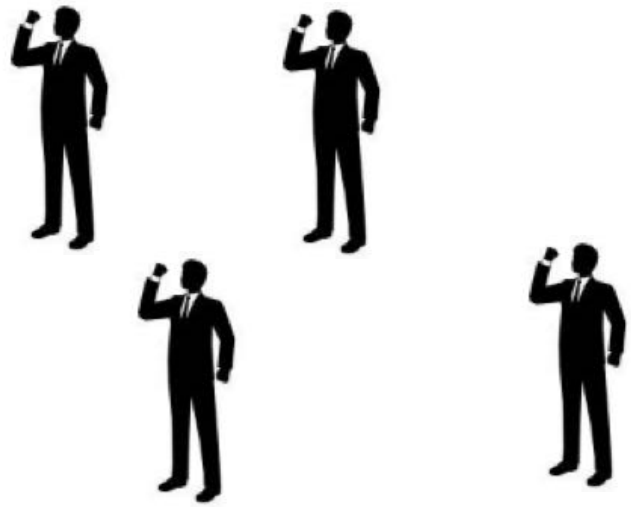
顧客Aグループと顧客Bグループの平均年齢は４０歳だった。
この２つの顧客グループは類似してるだろうか？

顧客Aグループ



≠

顧客Bグループ



平均に騙されるな！

問題

A社とB社の給料に関するデータです。このデータに基づいて、２つの会社の特徴について判断し、どちらに就職したいか理由を述べてください。

A社	年齢	年収
1	22	255
2	23	250
3	24	255
4	25	283
5	60	3000

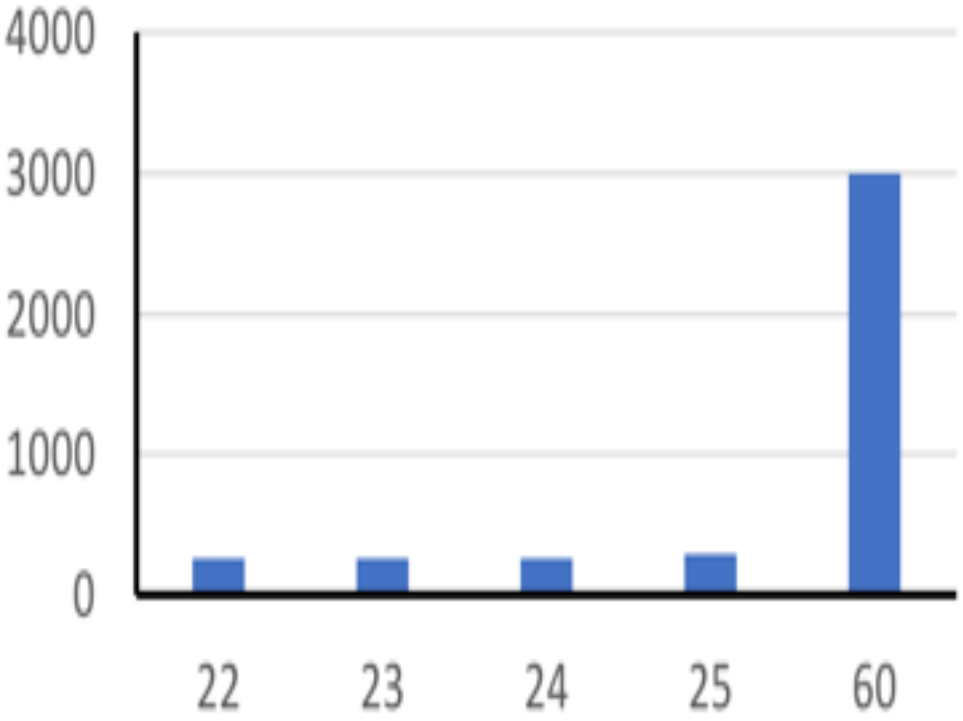
平均	30.8	800

B社	年齢	年収
1	20	300
2	23	400
3	26	600
4	40	800
5	50	1000

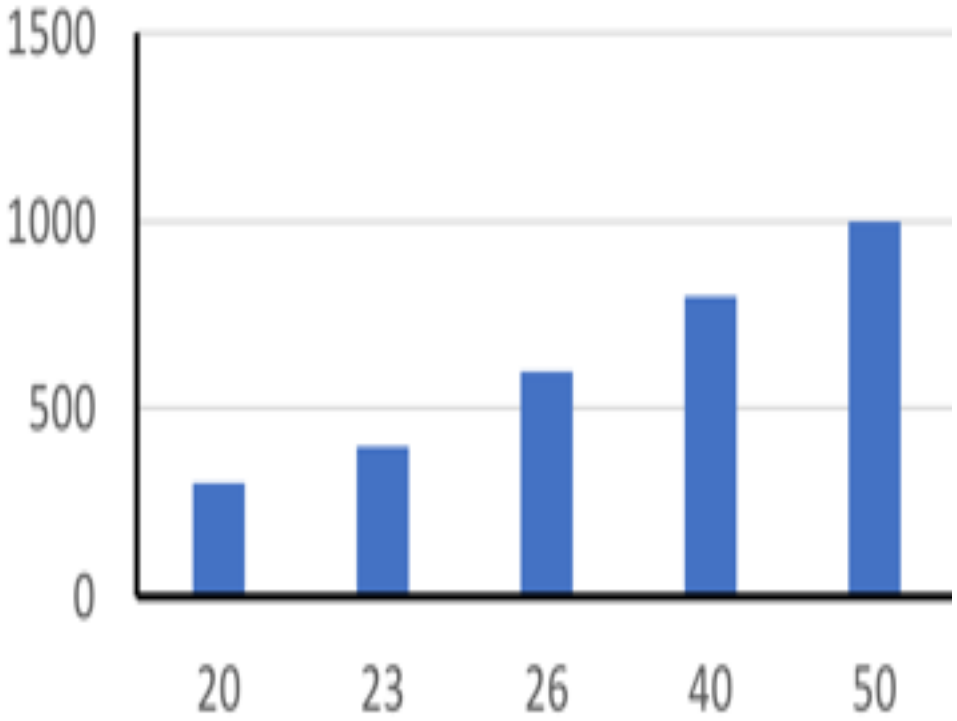
平均	31.8	620

可視化して比較

A社



B社



平均値に騙されるな！

サンプル数が少ない時、データの中に外れ値（異常値）が存在すると平均値は代表値としての役割をなさないことがある

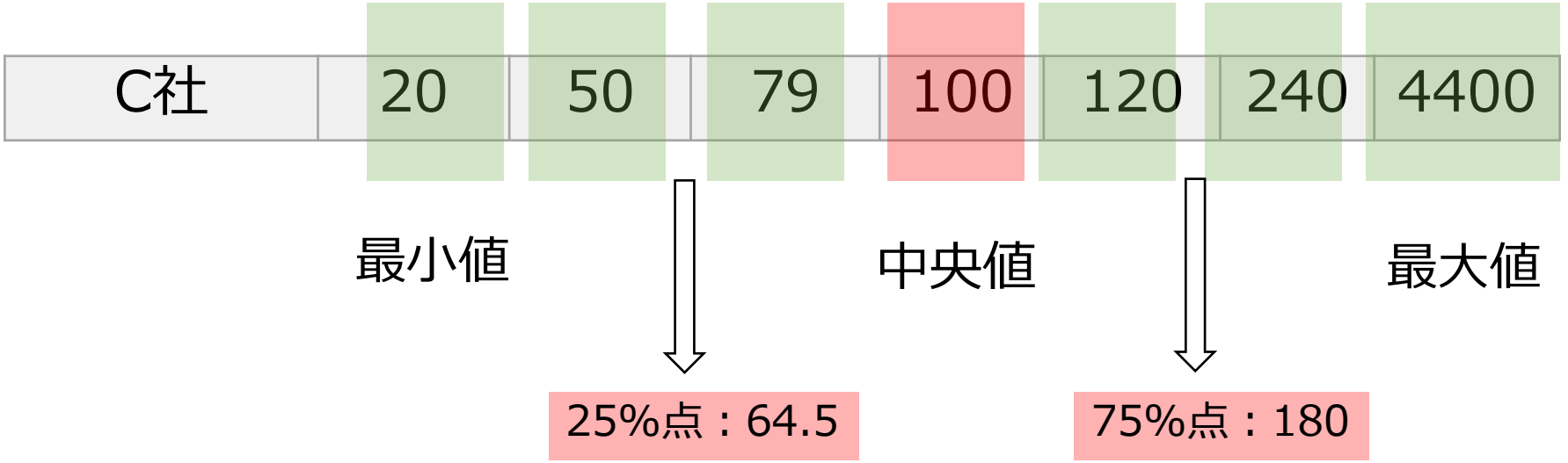
	月曜	火曜	水曜	木曜	金曜	土曜	日曜
C社	4400	50	20	80	90	50	79

平均値=681.3　　上のデータを代表している値とは言い難い

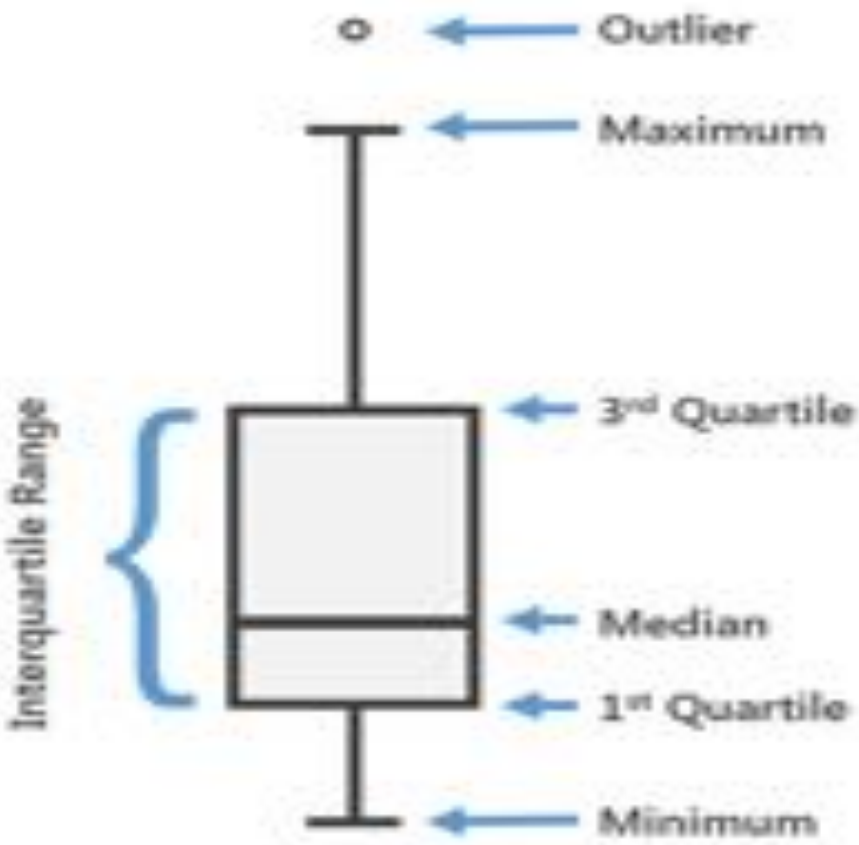
平均値が要約値として使えない場合[中央値(median)]

中央値：データを小さい順に並べた時、中央に位置する値

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
C社	4400	50	20	120	240	100	79



４分位



（最大値）

（７５%）

（中央値）

（２５%）

（最小値）

代表値　[最頻値（Mode)]

最頻値：データの中で最も頻繁に出現する値

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A店	140	50	20	120	240	100	79
B店	87	97	120	104	112	112	117

以下のデータの最頻値を求めよ

45	99	22	60	45	70	33

45	99	22	22	45	70	22

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

次にどの統計値を計算する？

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差

A社の平均値＝１０７

B社の平均値＝１０７

データの散らばり具合を調べる

平均値が同じ場合、データ間の違いをどう表現するか？

-3	-2	-1	1	2	3

平均＝ 0

-300	-200	-100	100	200	300

平均＝ 0

データが平均値を中心にどれくらいバラ付いているかを測る

標準偏差　！

平均値が同じ場合、データの違いをどう表現するか？

-3	-2	-1	1	2	3

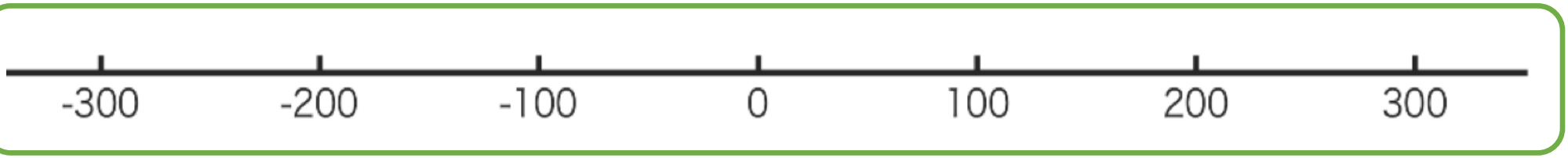
平均周りにデータがかたまっている

標準偏差が小さい

-300	-200	-100	100	200	300

平均の周りにデータは散らばっている

標準偏差が大きい



標準偏差

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

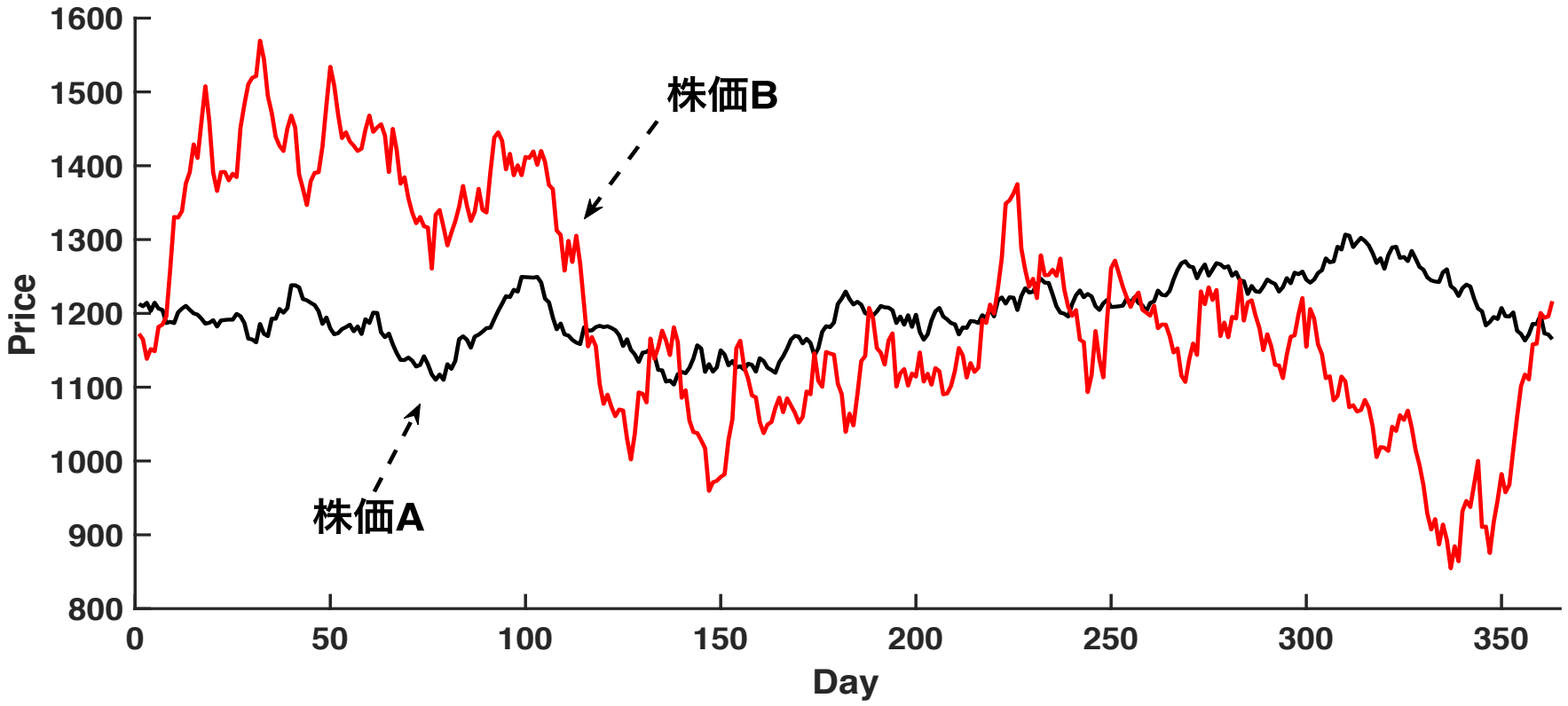
視覚的なイメージを持つことが重要

標準偏差をイメージする

株価A　2017年度の平均株価　1200円/株
株価B　2017年度の平均株価　1200円/株

標準偏差をイメージする

株価A　2017年度の平均株価　1200円/株
株価B　2017年度の平均株価　1200円/株

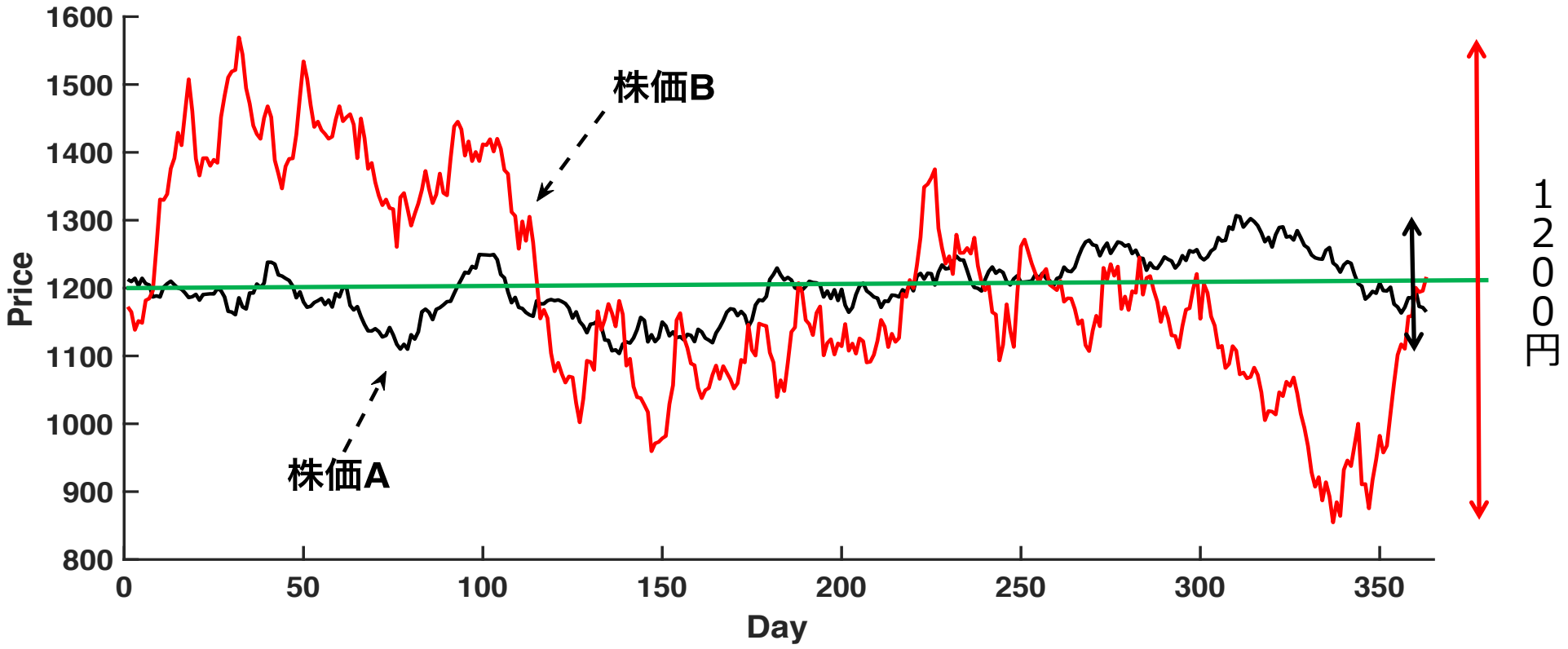


標準偏差をイメージする

株価Aのばらつき　＜　株価Bのばらつき

標準偏差が小さい
リスクが小さい

標準偏差が大きい
リスクが大きい



量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

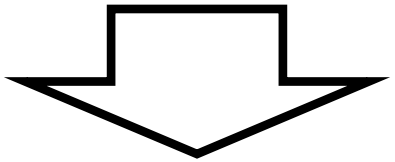
標準偏差が大きいのはA社、それともB社？

「A社」の標準偏差 ＝ 71.4

「B社」の標準偏差 ＝ 11.7

量的データの要約

	月曜	火曜	水曜	木曜	金曜	土曜	日曜
A社	140	50	20	120	240	100	79
B社	87	97	120	104	112	112	117

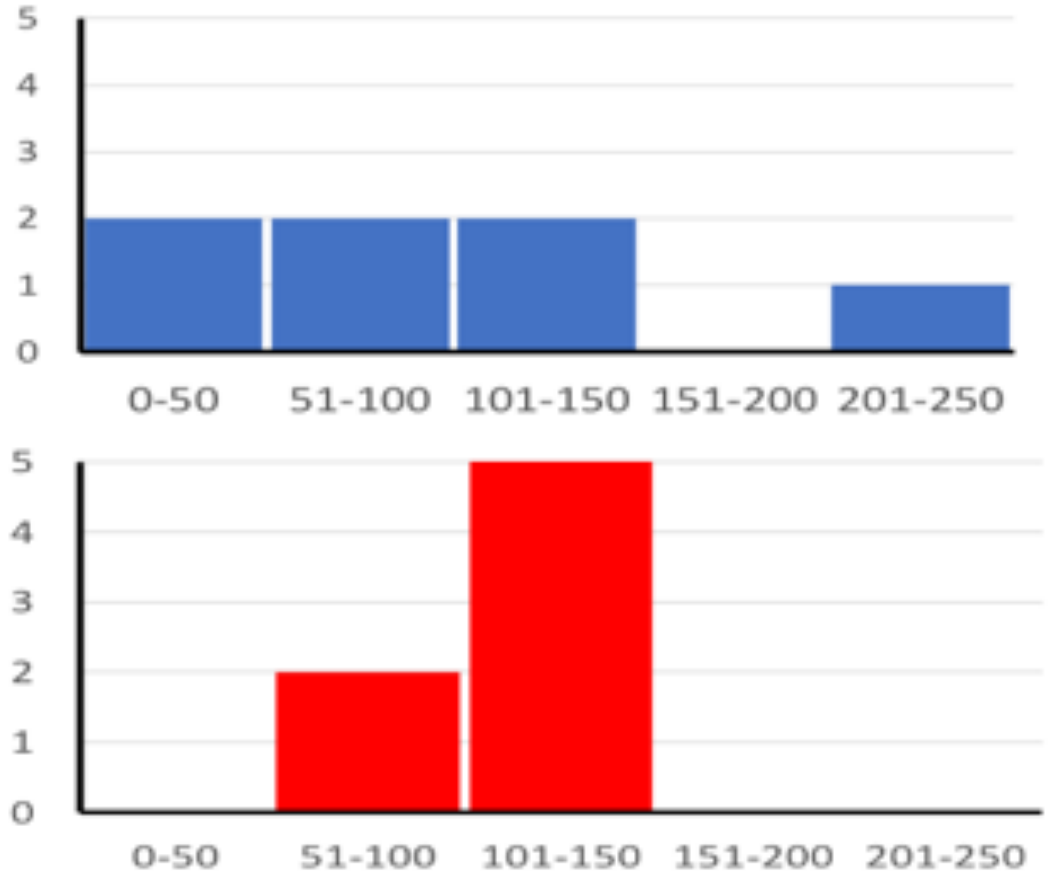


	A社	B社
平均値	107	107
中央値	100	112
最大値	240	120
最小値	20	87
標準偏差	71.4	11.7

データを要約する

A社とB社の１週間における売上データです。どちらかの会社に融資するべきですか？その理由は？

	A社	B社
平均値	107	107
中央値	100	112
最大値	240	120
最小値	20	87
標準偏差	71.4	11.7



データ分析の実例

問題：このデータから何がわかるのか？

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.66	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low

データの分類

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

数量データ

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差
- ２５%、７５点
- ヒストグラム

質的データ

- 円グラフ
- クロス集計

虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

データを分析するとは

データの要約

データ間の関係性

予測する

結果の検証

問題解決のための哲学

分解と統合

虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

社員は会社に満足しているのだろうか？

（レベル１　集計）

虎の巻（データ分析）

量的データの集計

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

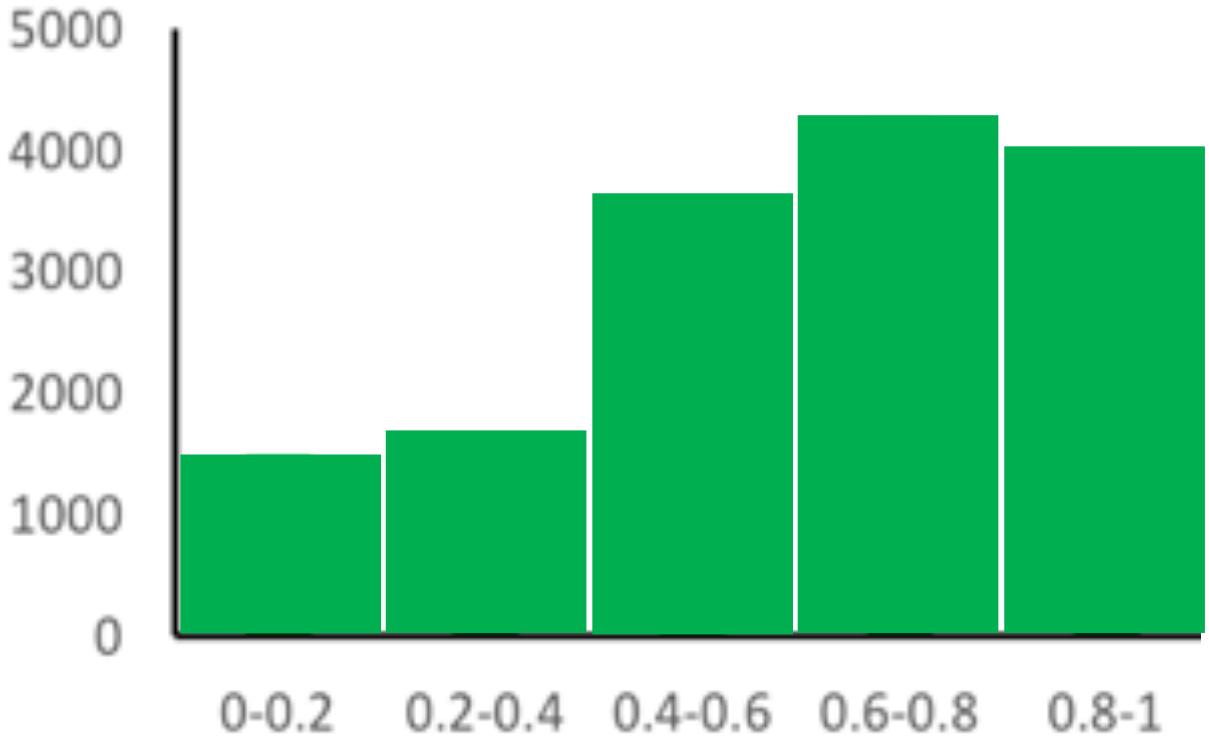
虎の巻（データ分析）

データを分析するとは

データの要約

満足度

データ区間	頻度
0~0.2	1478
0.2~0.4	1646
0.4~0.6	3605
0.6~0.8	4268
0.8~1.0	4002



虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

このデータからどの社員が退職するか予測することは可能なのか？

（レベル２　検定）
（レベル３　予測モデルの設計）

虎の巻（データ分析）

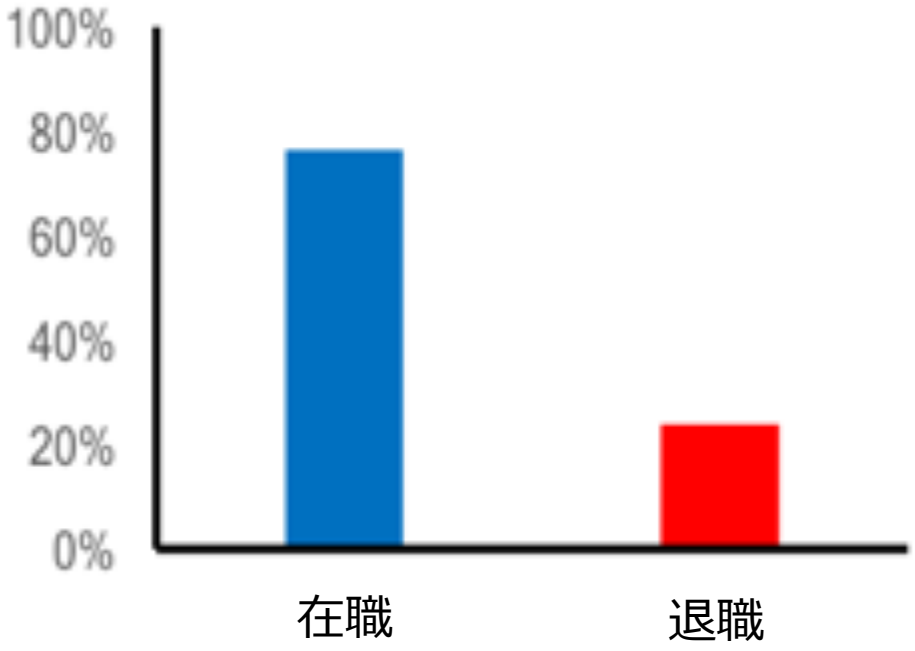
質的データの集計

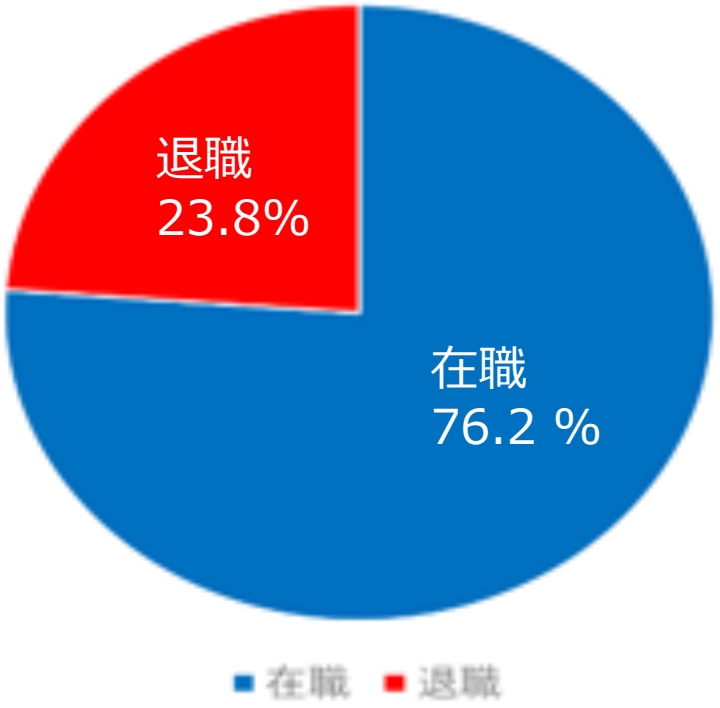
ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

データの可視化

退職	在職
3571	11428

23.8%	76.2%





虎の巻（データ分析）

データを分析するとは

データ間の関係性

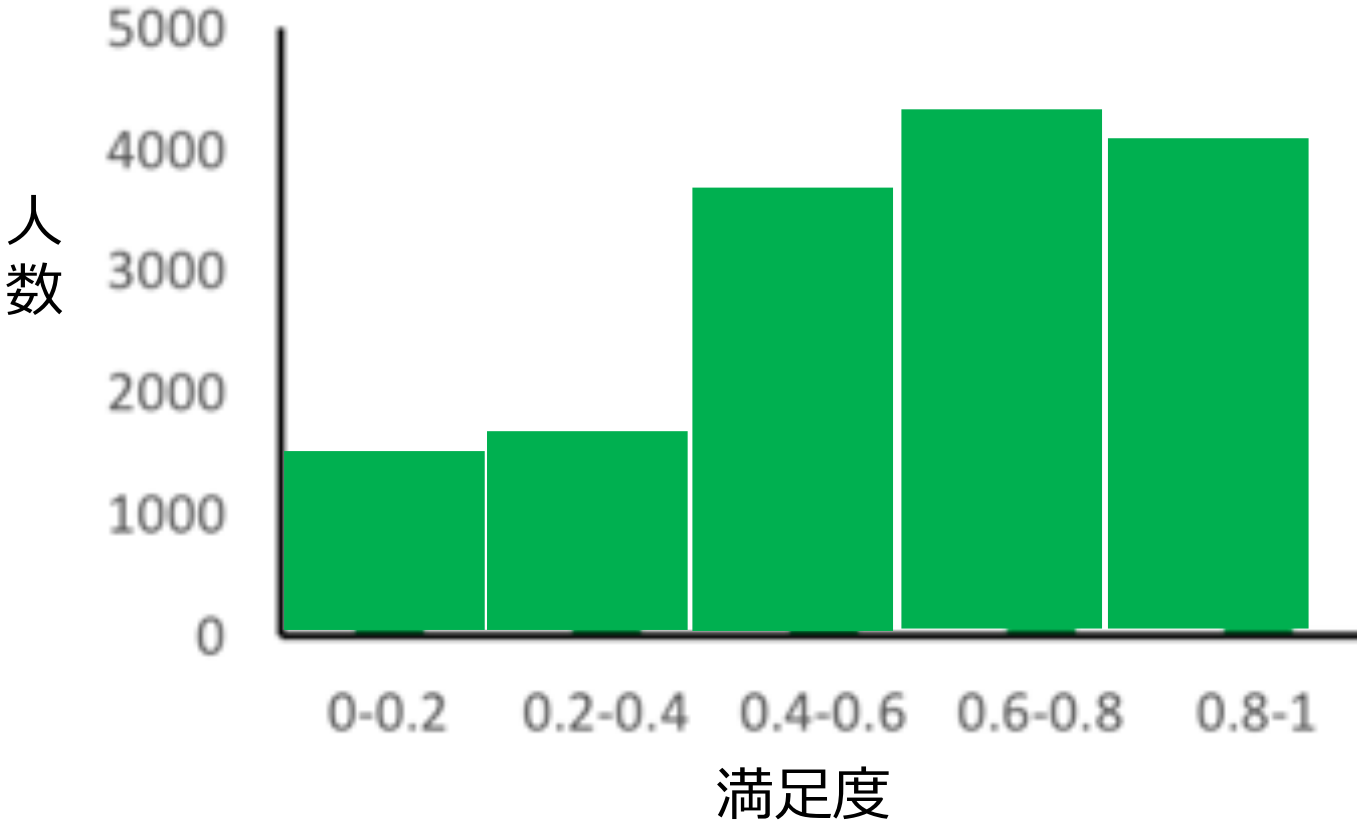
ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

虎の巻（データ分析）

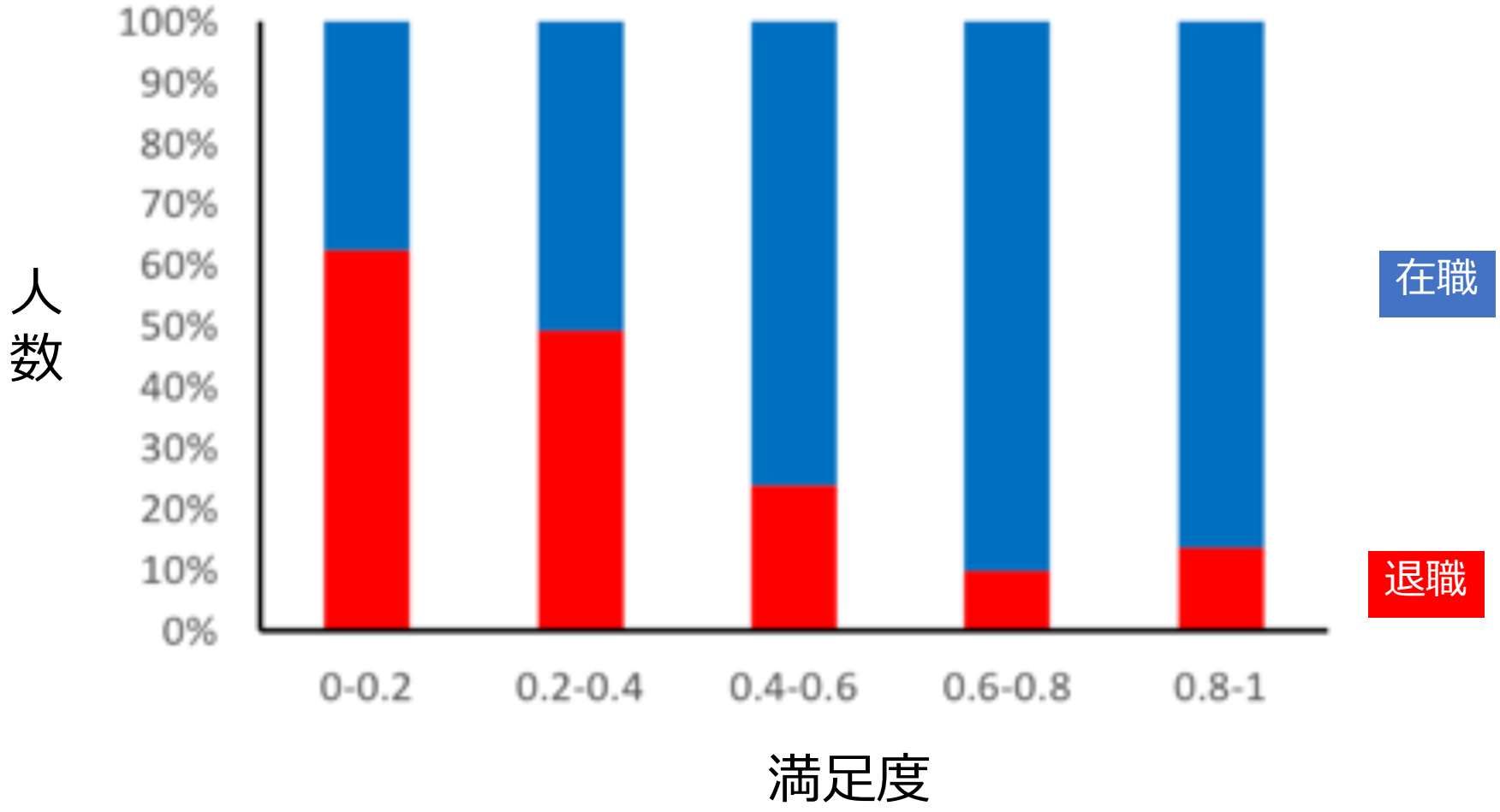
問題解決のための哲学

分解と統合

具体的にどうする？



虎の巻（データ分析）



虎の巻（データ分析）

データを分析するとは　検定

統計学的に退職者と在職者で満足度に差があるのかを検証する

t-検定: 分散が等しくないと仮定した２標本による検定

	在職者	退職者
平均	0.664468641	0.443011364
分散	0.045235724	0.07403649
観測数	1148	352
仮説平均との差異	0	
自由度	490	
t	14.01362063	
P(T<=t) 片側	4.61969E-38	
t 境界値 片側	1.647969283	
P(T<=t) 両側	9.23938E-38	
t 境界値 両側	1.964817132	

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?

満足度	他者評価	プロジェクト数	労働時間(月平均)	労働時間(会社内)	Work accident	退職か在職	過去5年昇進(有無)	所属部署	給料
0.58	0.55	4	202	3	0	在職	無	IT	medium
0.67	0.74	3	226	3	0	在職	無	product_mng	low
0.11	0.91	7	287	4	0	退職	無	sales	low
0.37	0.5	2	135	3	0	退職	無	product_mng	low
0.93	0.79	5	241	4	0	在職	無	marketing	high
0.4	0.38	3	280	2	0	在職	無	marketing	low
0.23	0.64	5	150	5	0	在職	無	hr	medium
0.83	0.98	5	189	4	1	在職	無	management	low
0.2	0.58	3	209	5	0	在職	無	hr	medium
0.95	0.7	4	257	3	1	在職	無	technical	low
0.11	0.8	6	282	4	0	退職	無	technical	medium
0.7	0.5	6	214	5	0	在職	無	support	medium
0.43	0.51	5	168	4	0	在職	無	product_mng	medium
0.46	0.75	6	276	6	0	在職	無	support	low
0.67	0.8	4	137	2	0	在職	無	support	medium
0.63	0.88	4	260	2	0	在職	無	sales	low
0.99	0.92	5	213	2	0	在職	無	hr	high
0.24	0.94	4	146	4	0	在職	無	product_mng	medium
0.55	0.82	4	134	6	0	在職	無	technical	medium

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1									
0.6-0.8									
0.4-0.6									
0.2-0.4									
0-0.2									

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1	30	28	43	14	33	38	150	91	129
0.6-0.8	12	22	26	11	20	30	128	72	62
0.4-0.6	59	80	66	25	73	56	311	154	177
0.2-0.4	44	35	50	13	36	34	186	103	126
0-0.2	59	50	88	28	41	40	239	135	203

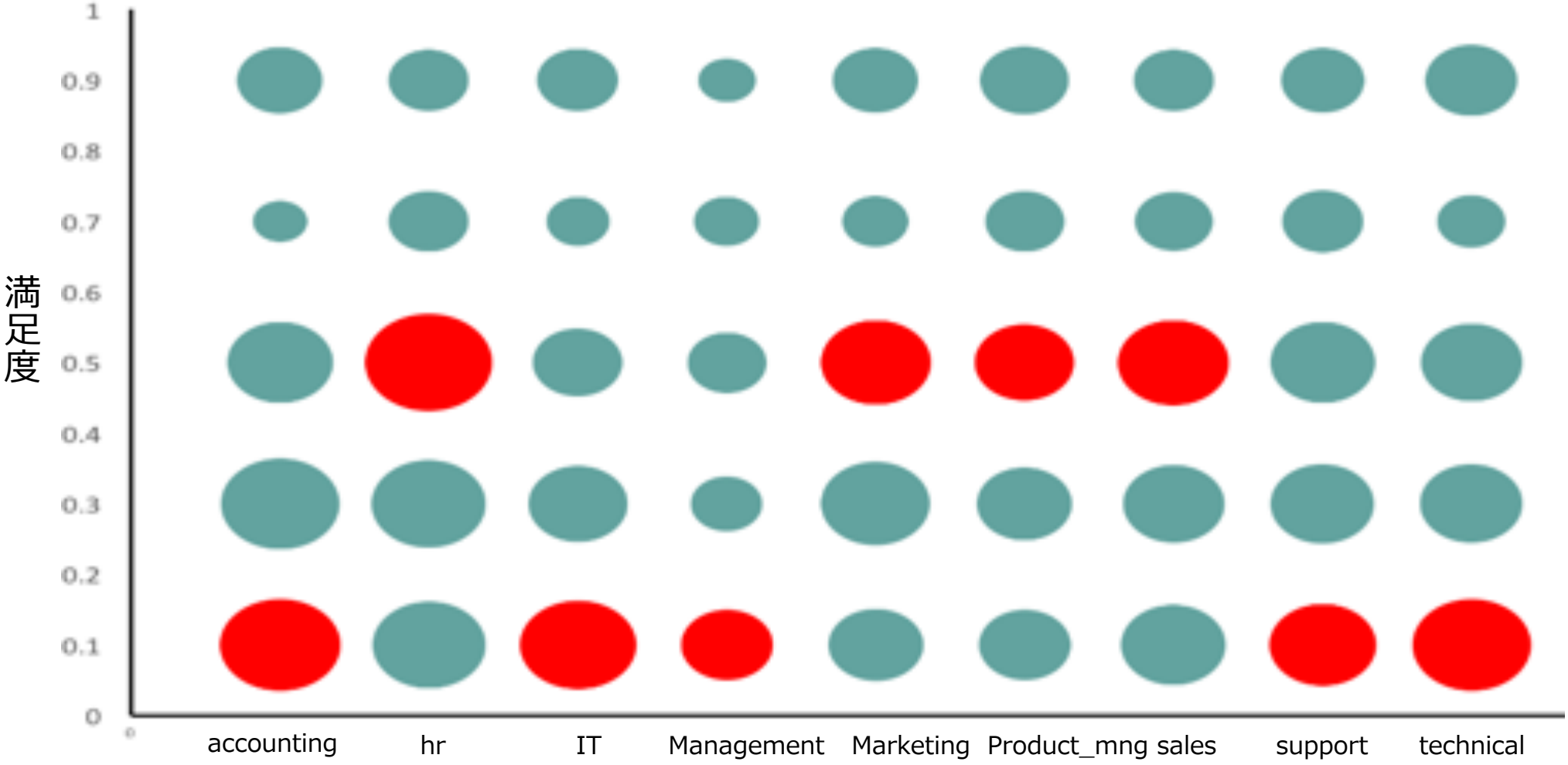
データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1	30	28	43	14	33	38	150	91	129
0.6-0.8	12	22	26	11	20	30	128	72	62
0.4-0.6	59	80	66	25	73	56	311	154	177
0.2-0.4	44	35	50	13	36	34	186	103	126
0-0.2	59	50	88	28	41	40	239	135	203

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?



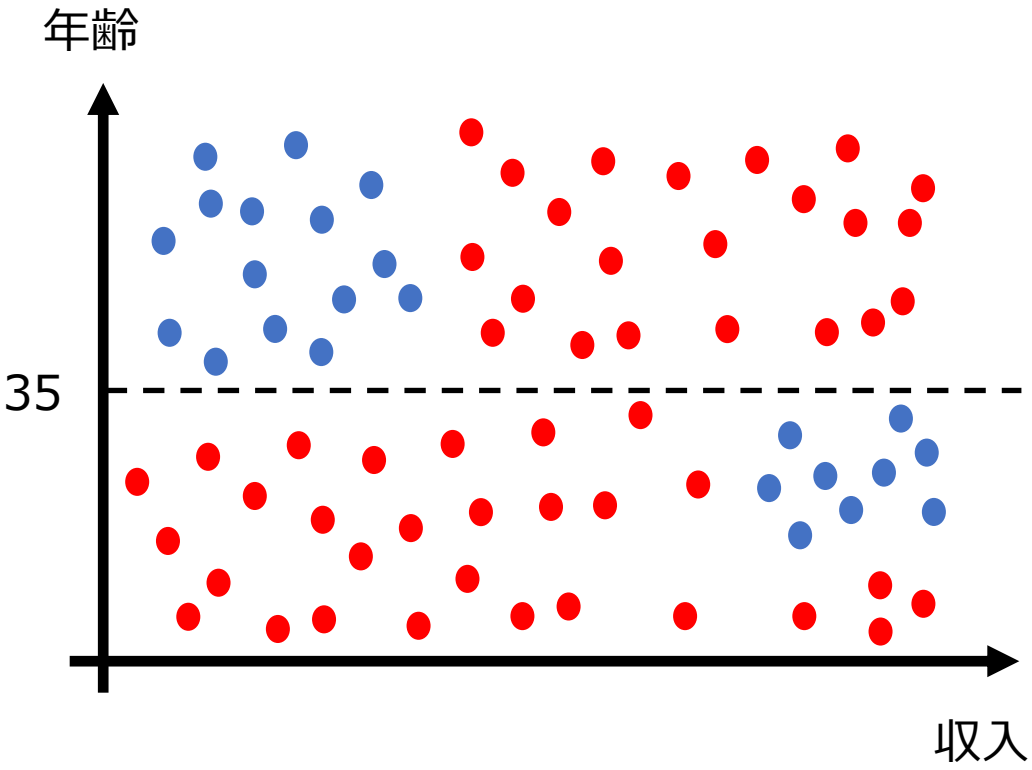
虎の巻（データ分析）

データを分析するとは　予測する

ID	満足度	他者評価	プロジェクト数	労働時間（月平均）	労働時間（会社内）	Work accident	退職・在職	過去５年の昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

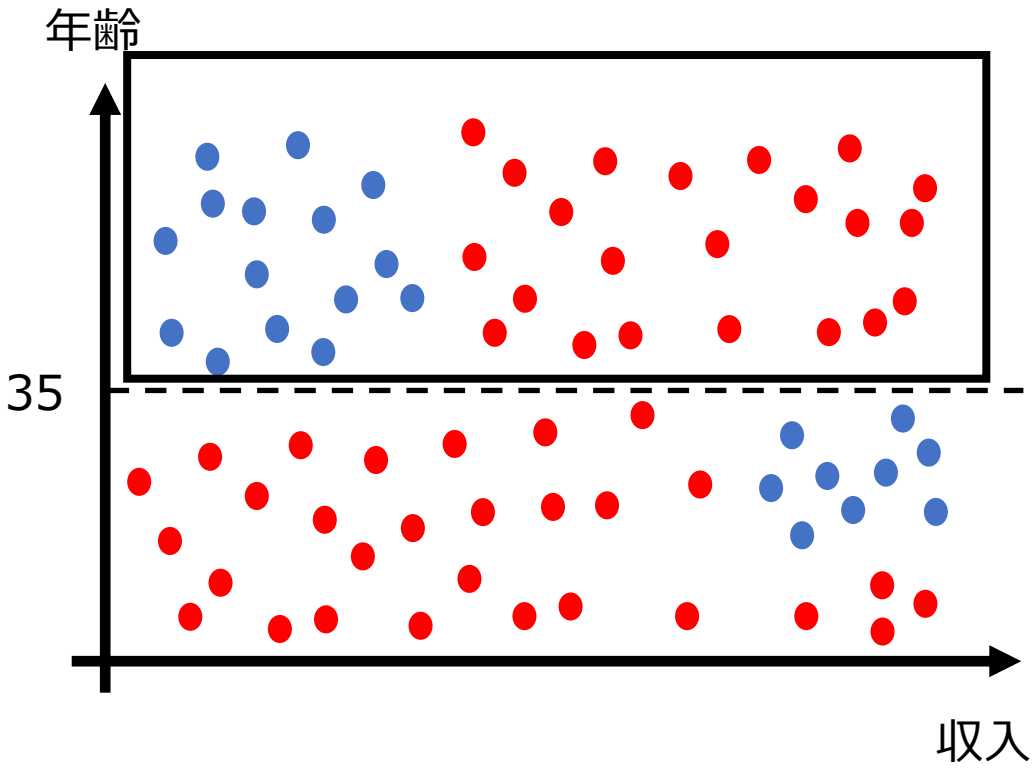
機械学習を使うと・・・

データを分類する



年齢３５以下？

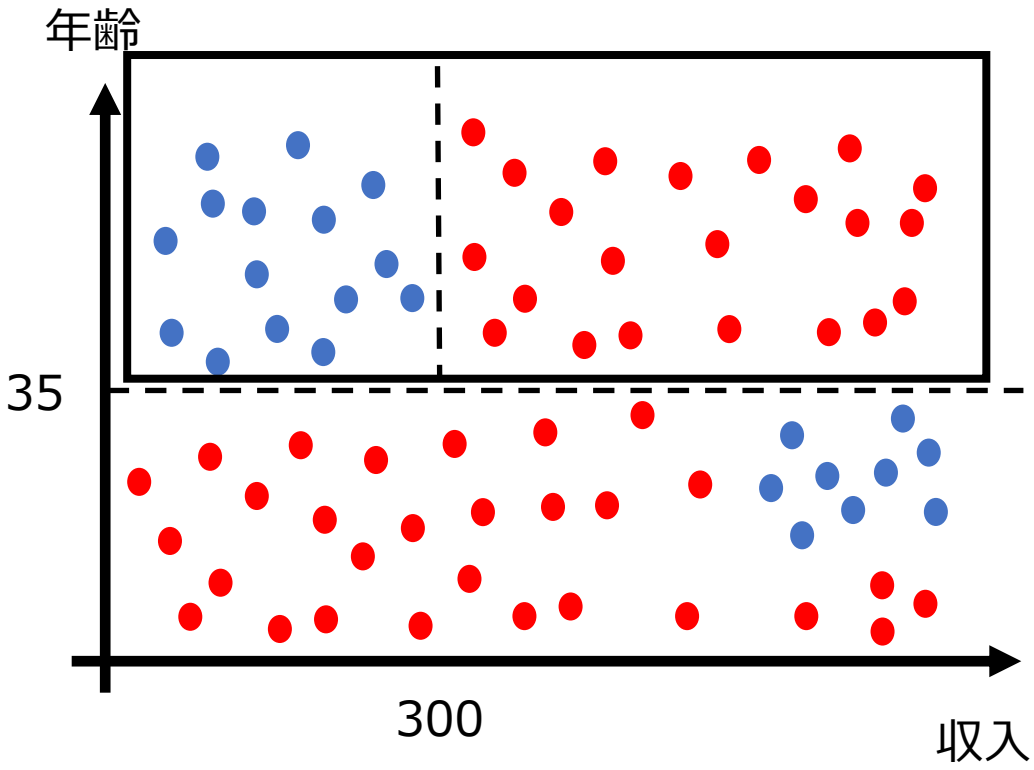
データを分類する



年齢３５以下？

No

データを分類する

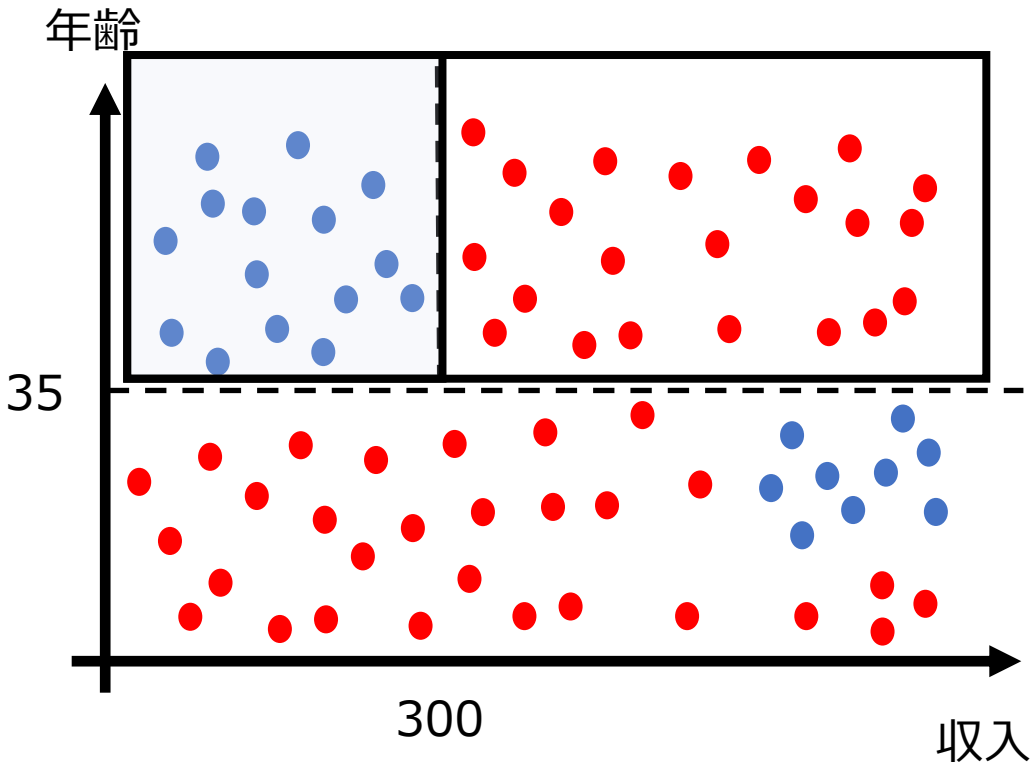


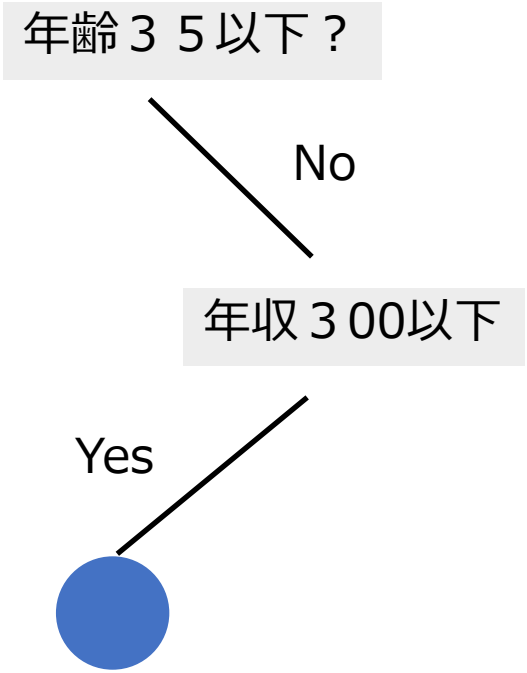
年齢３５以下？

No

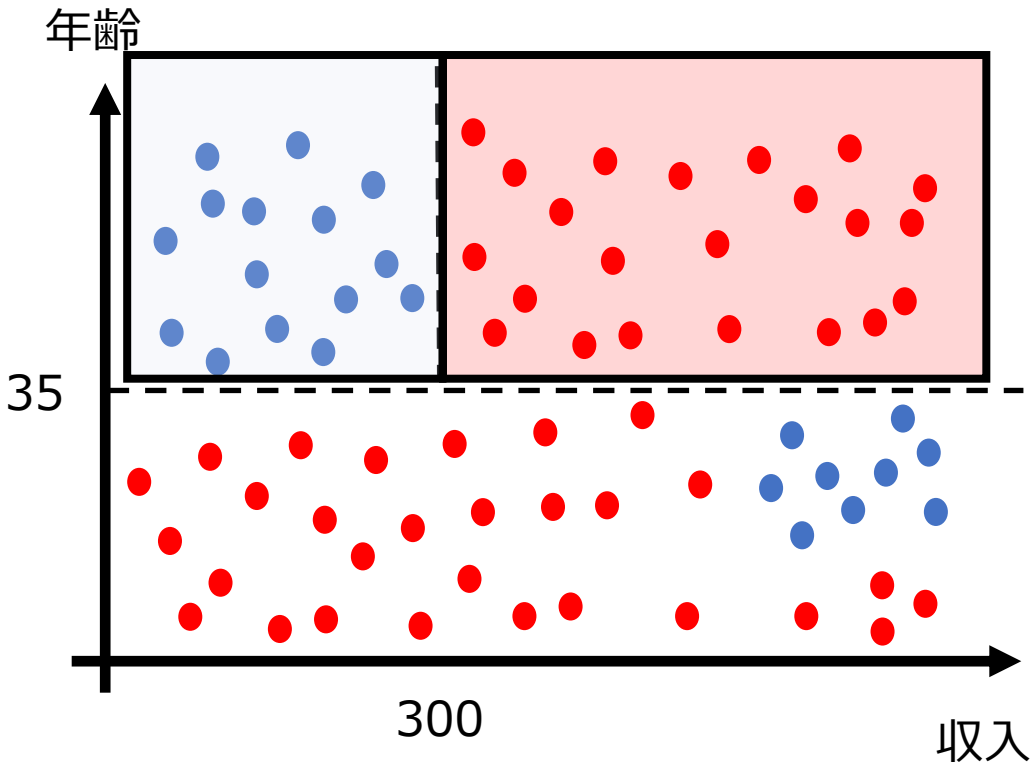
年収３００以下

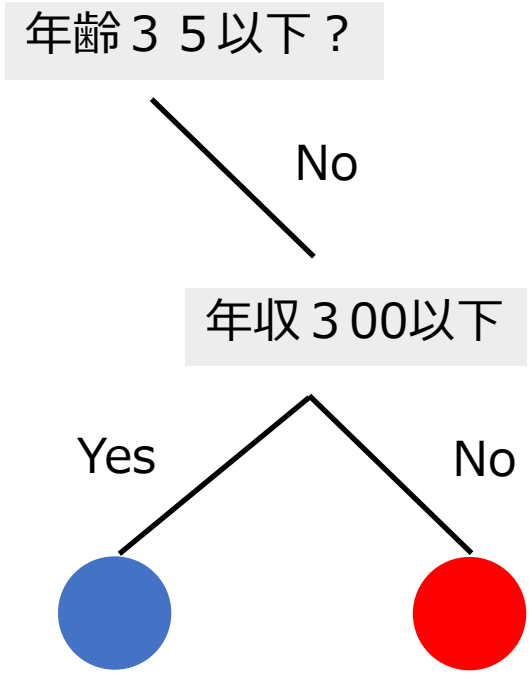
データを分類する



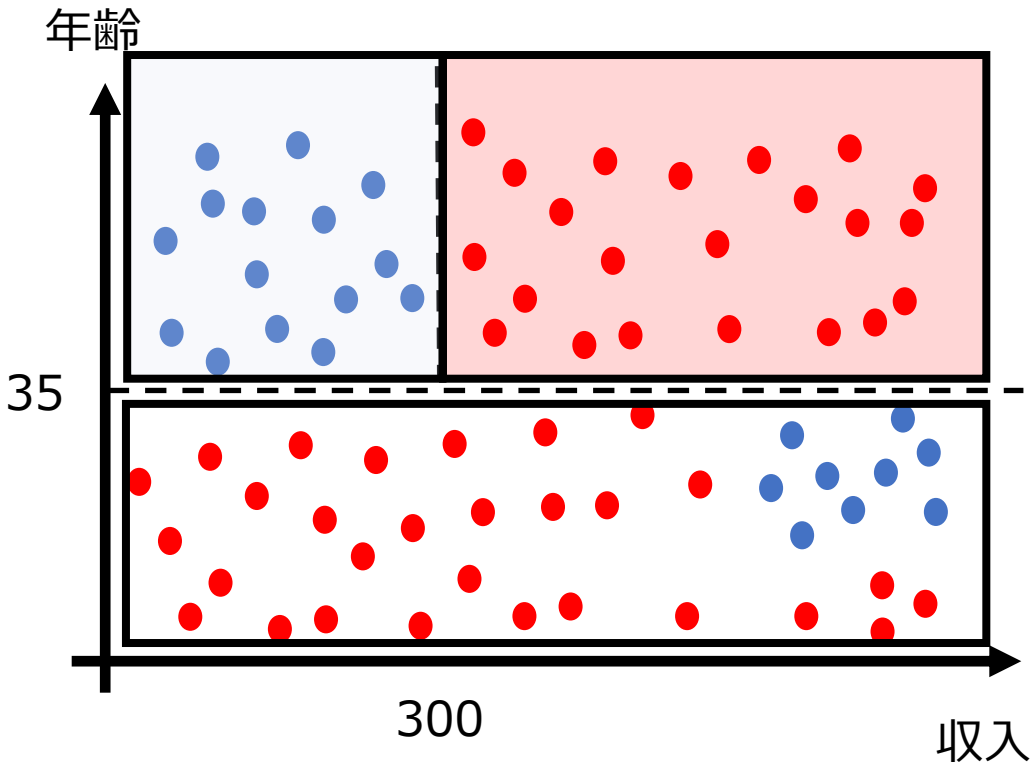


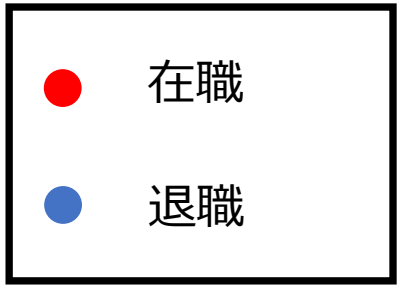
データを分類する

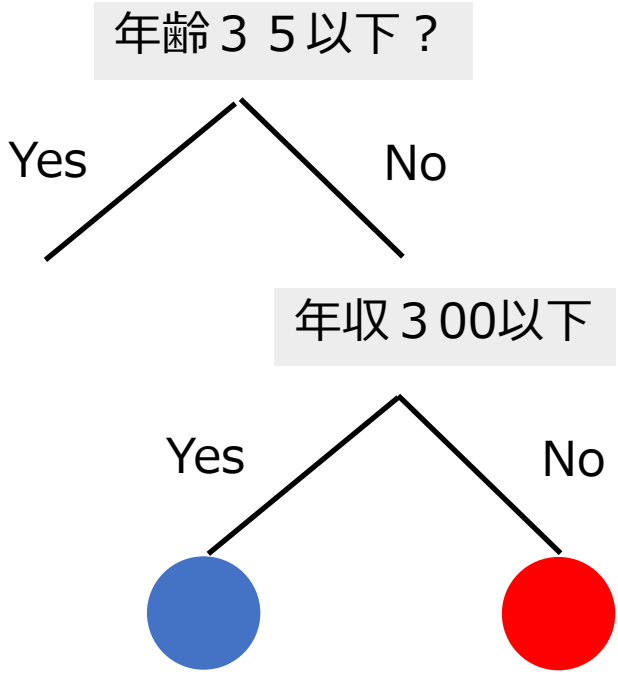




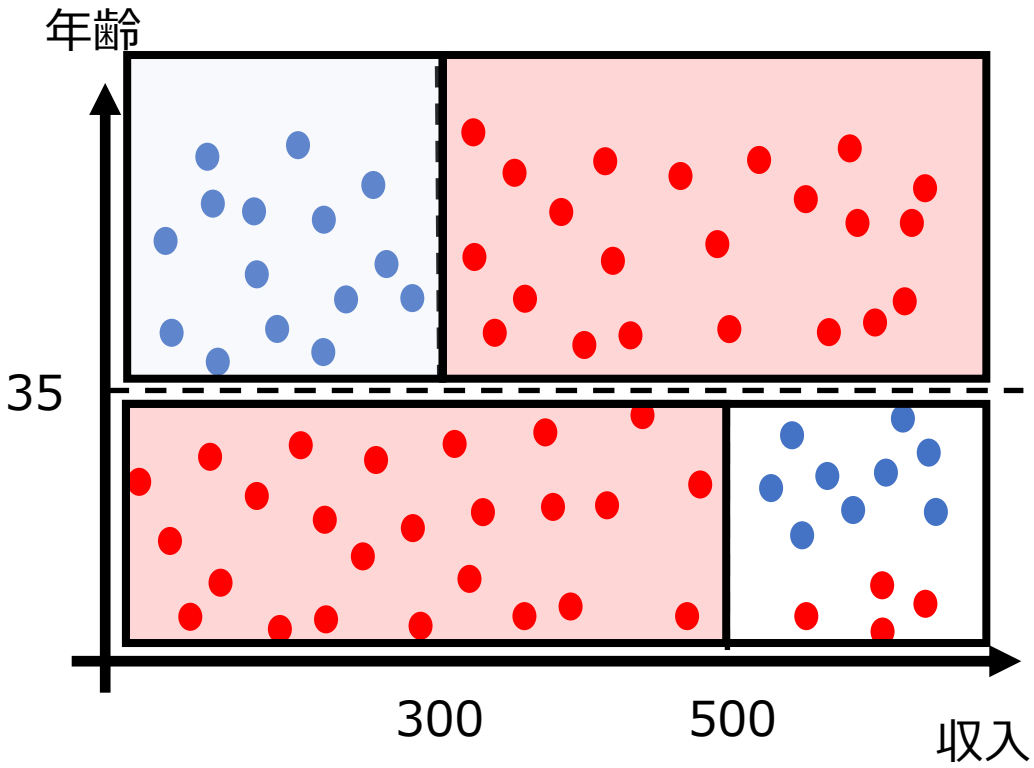
データを分類する

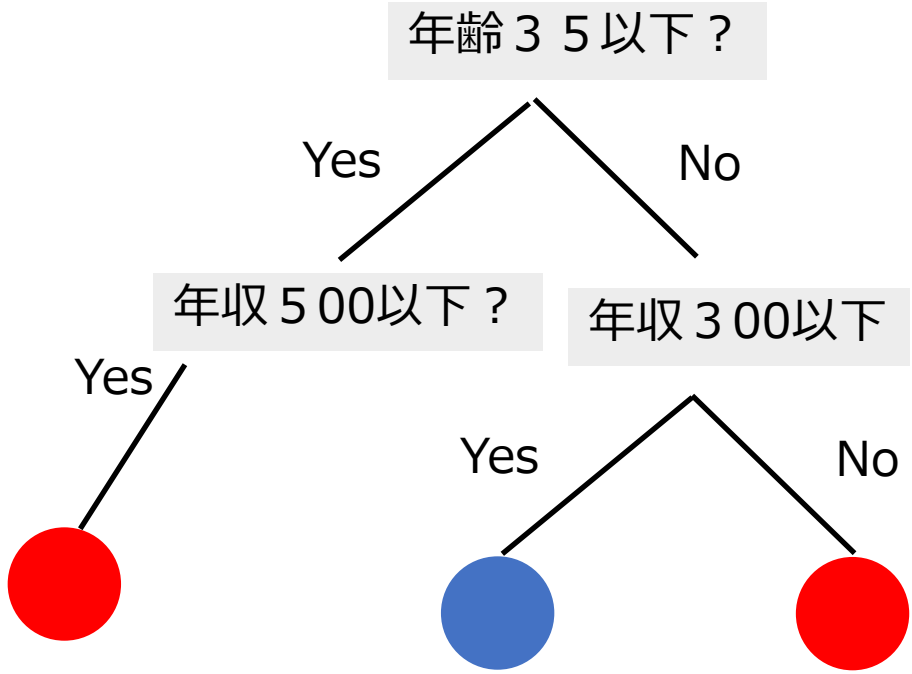




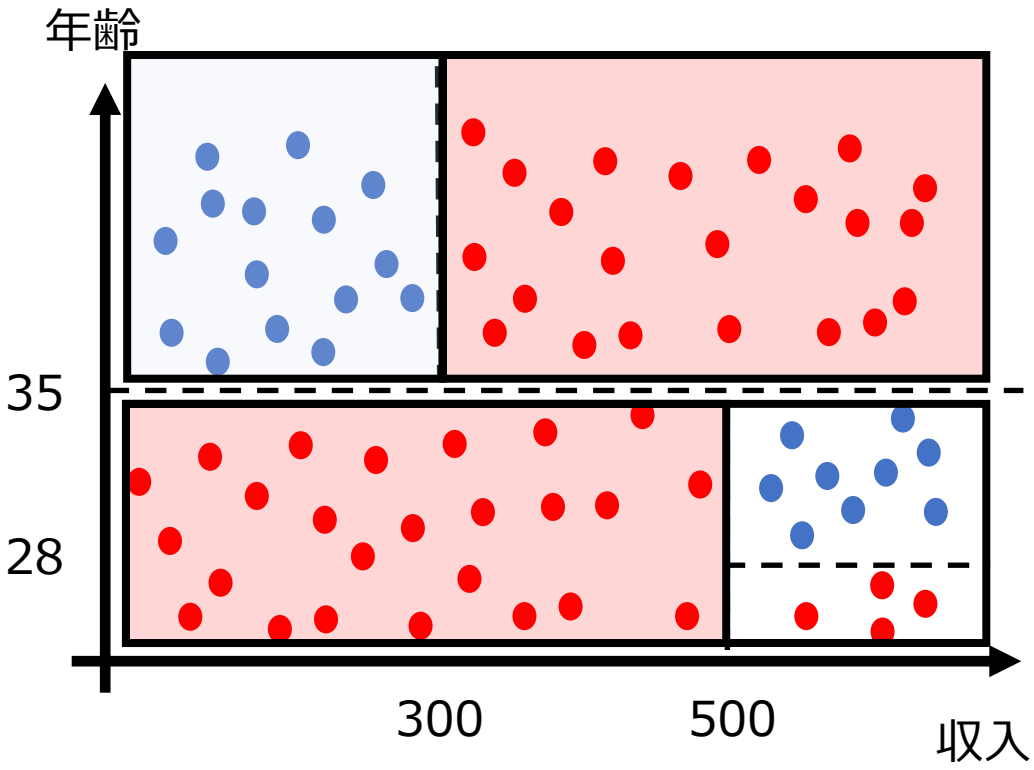


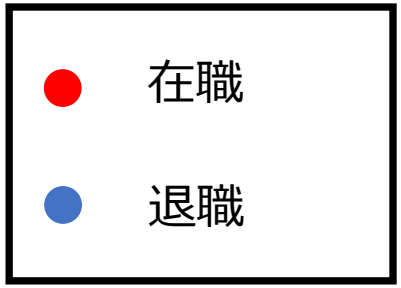
データを分類する

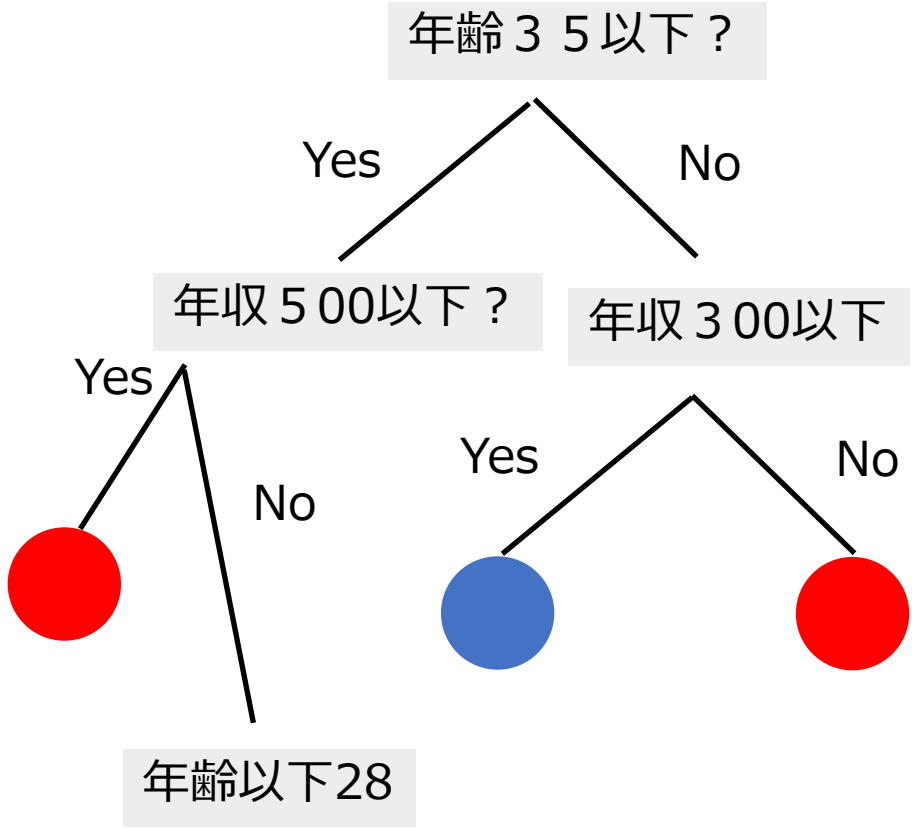




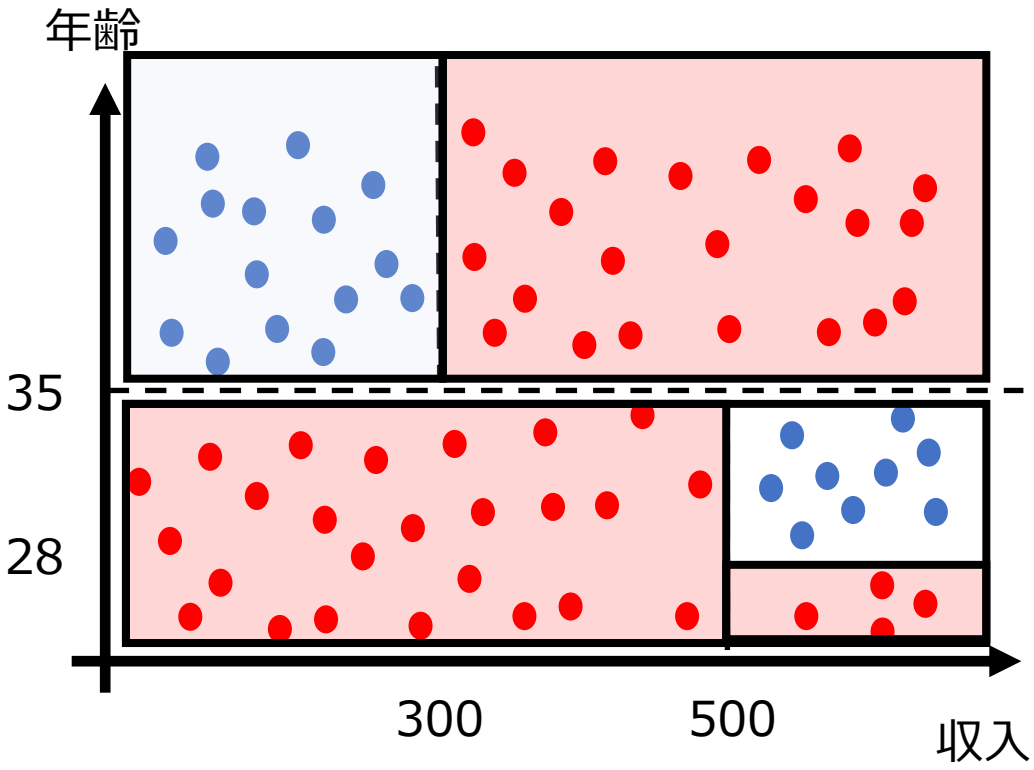
データを分類する

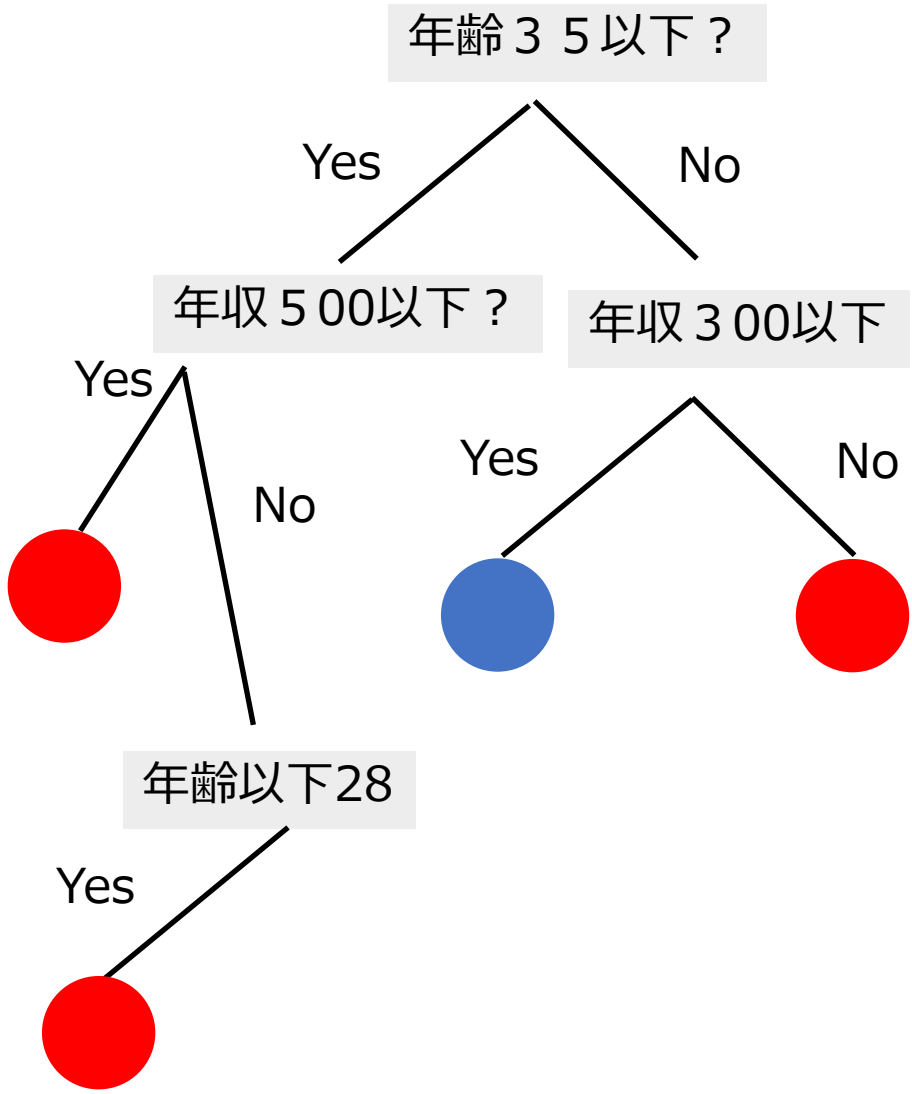




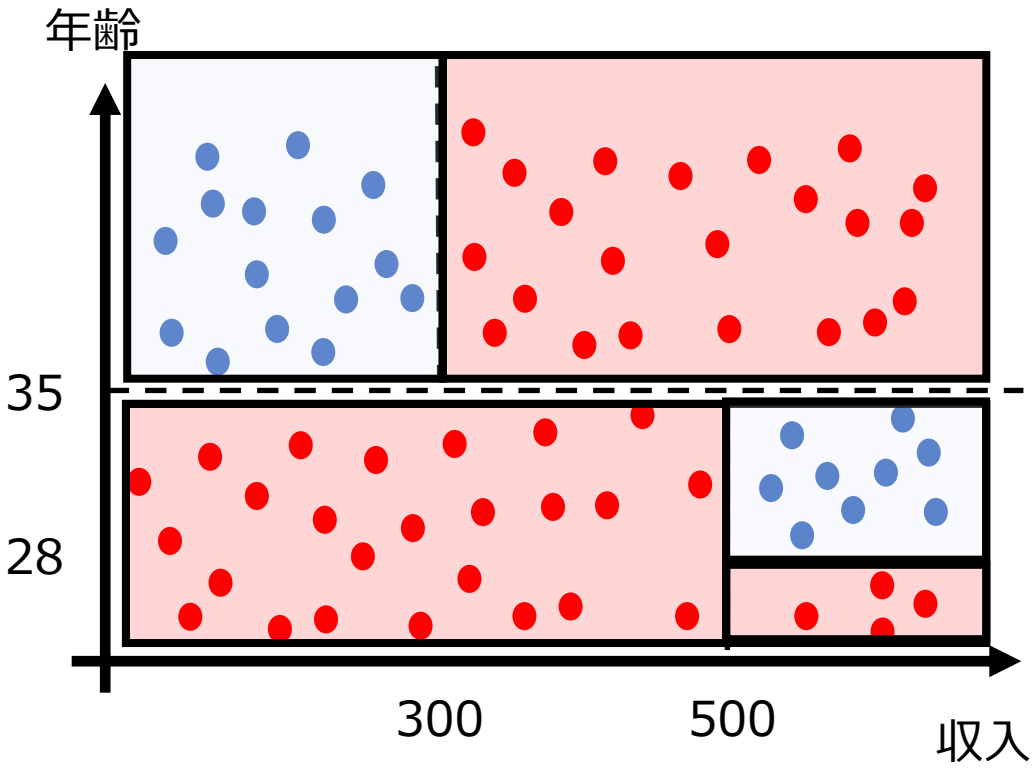


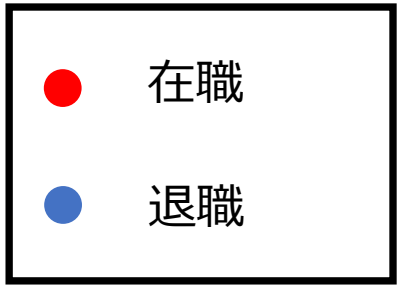
データを分類する

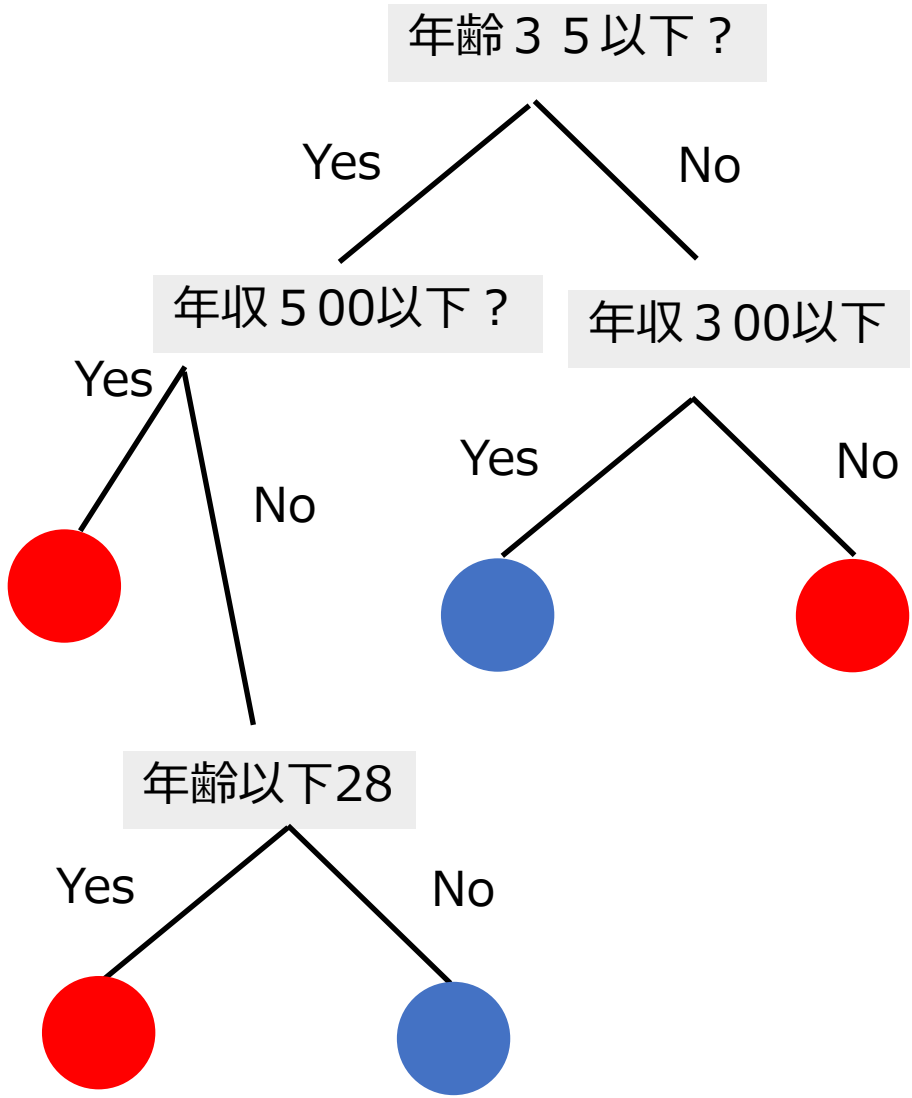




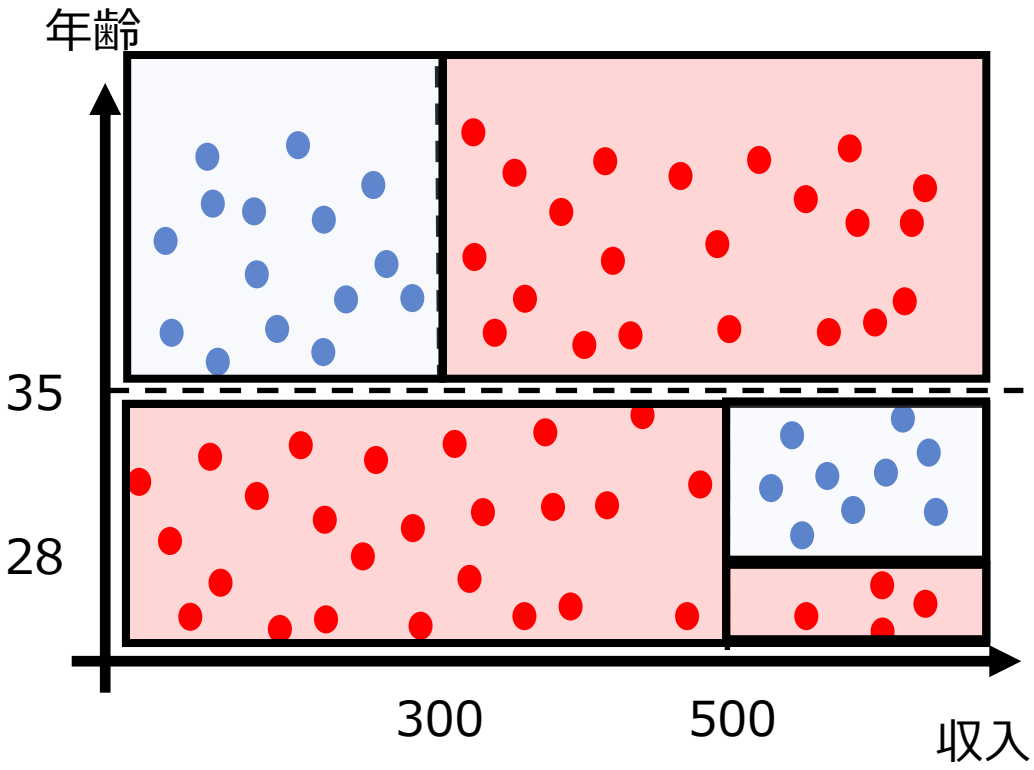
データを分類する

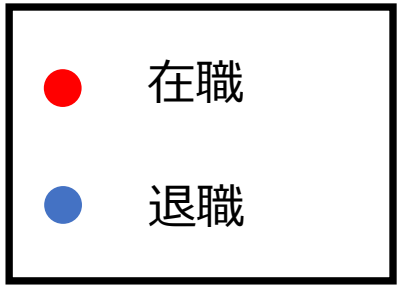




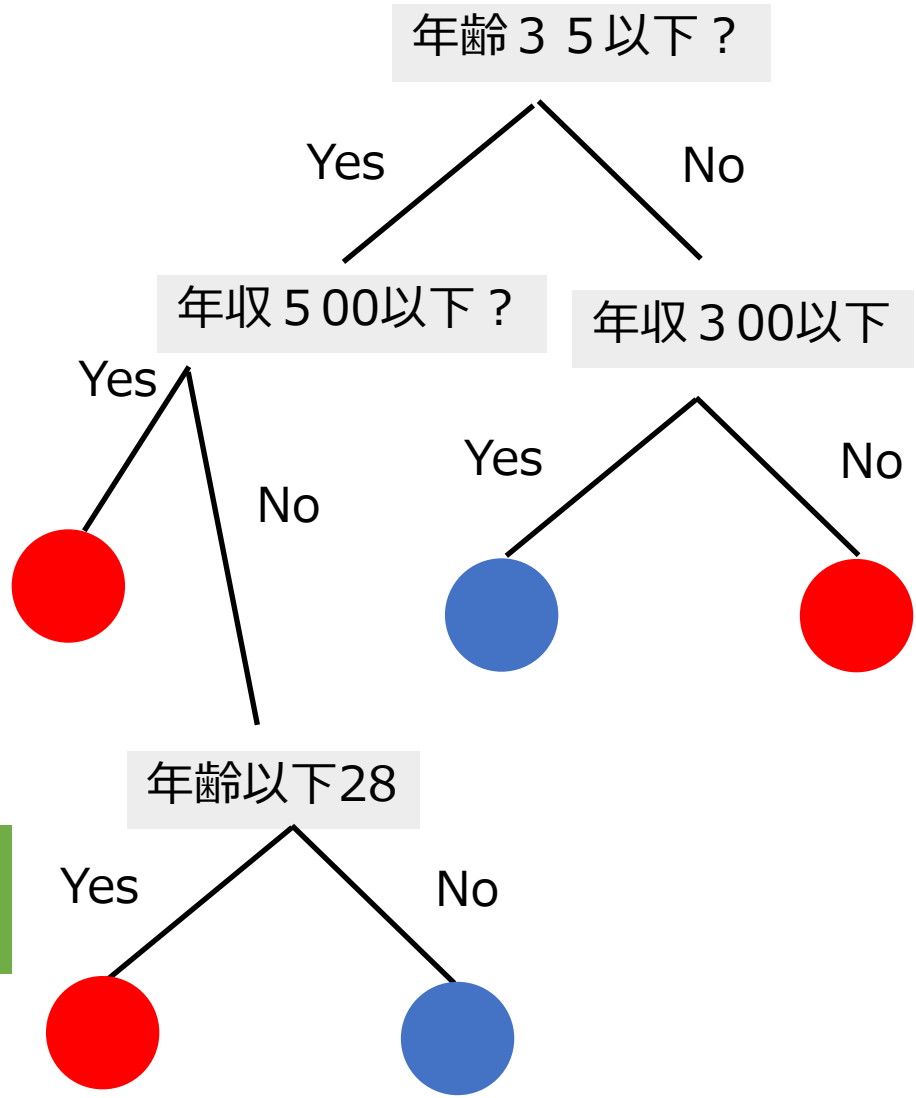


データを分類する

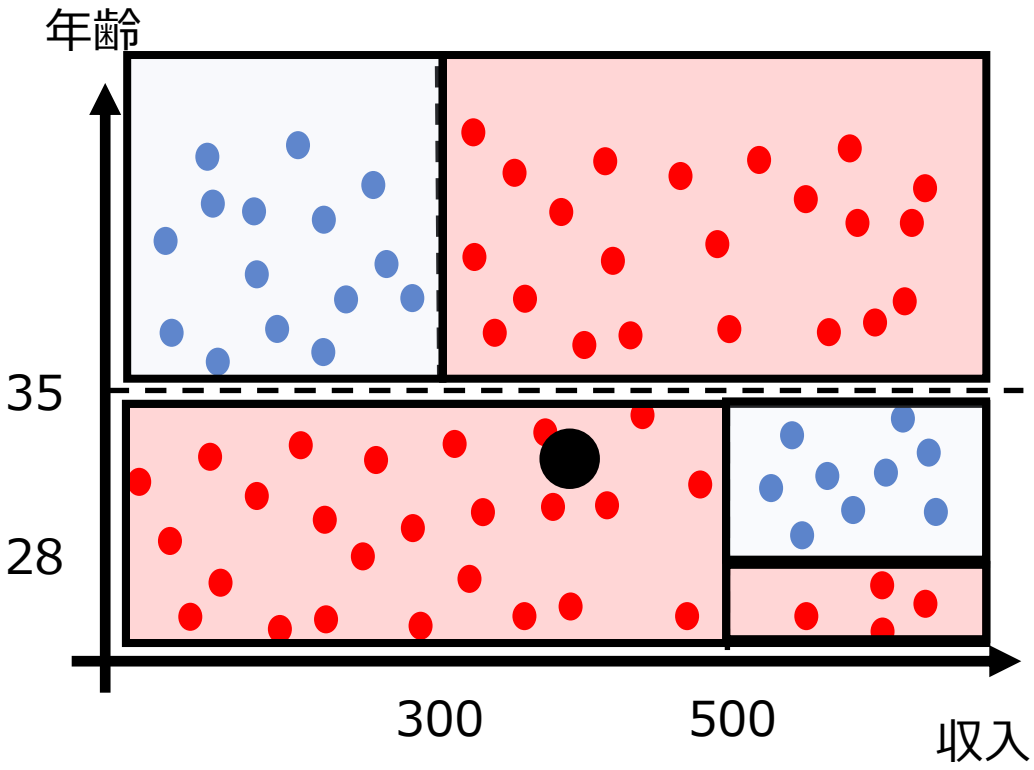


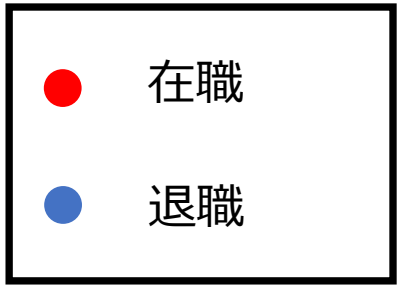


問題
年収４００万、年齢３０歳？

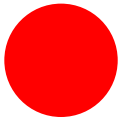


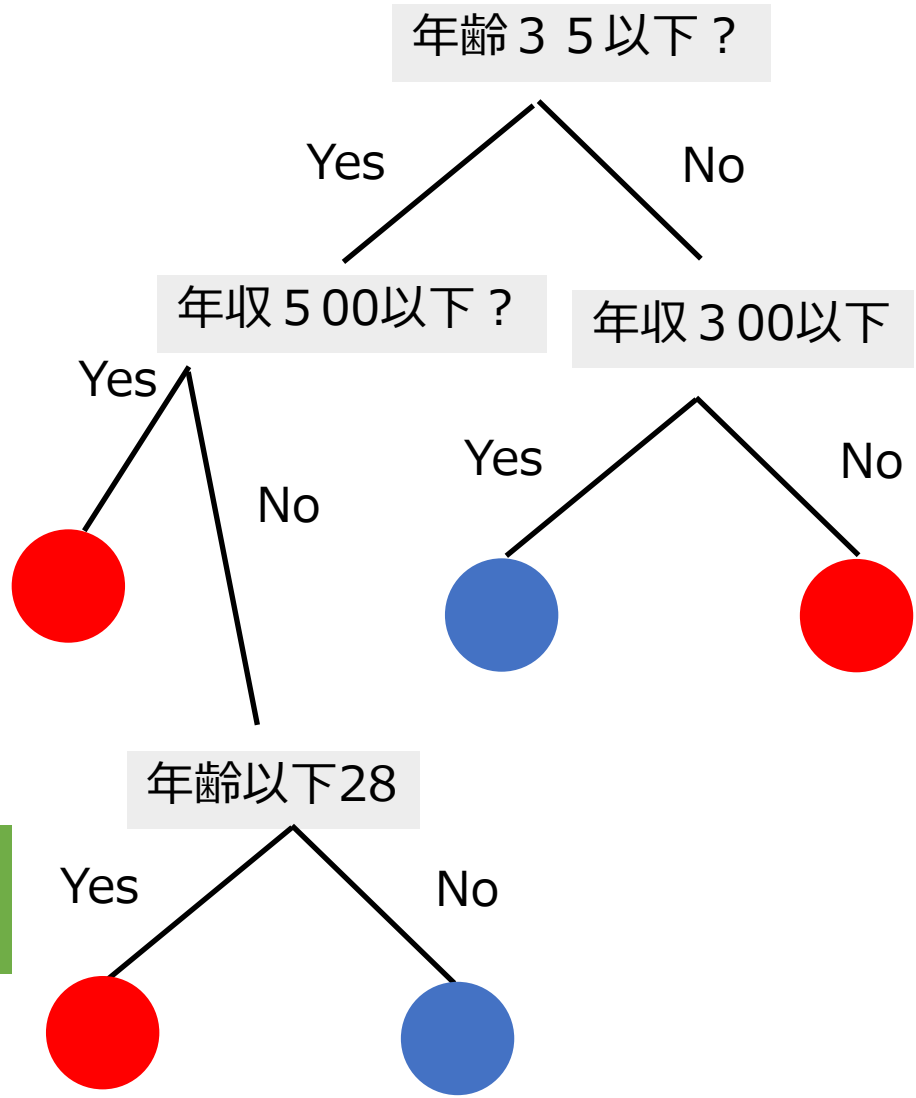
データを分類する



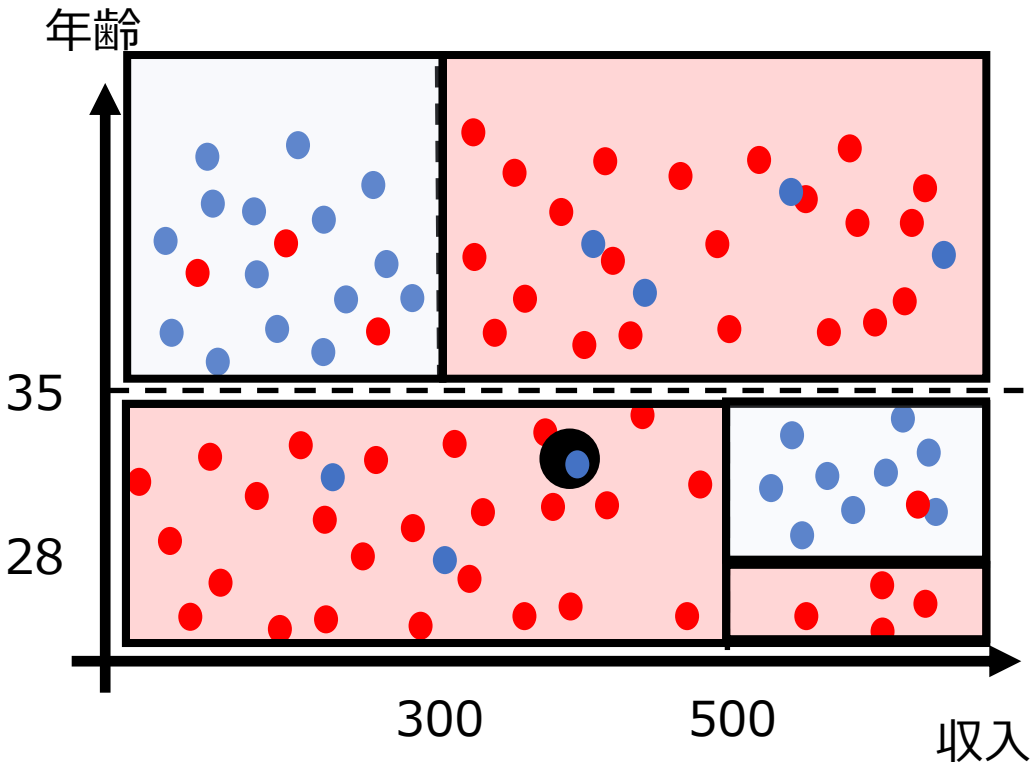


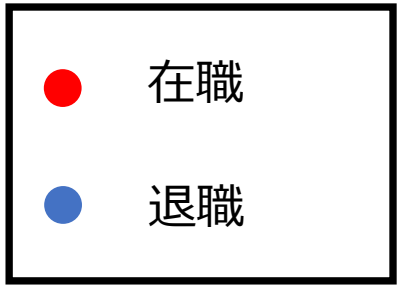
問題
年収４００万、年齢３０歳？



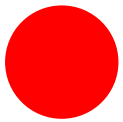


データを分類する

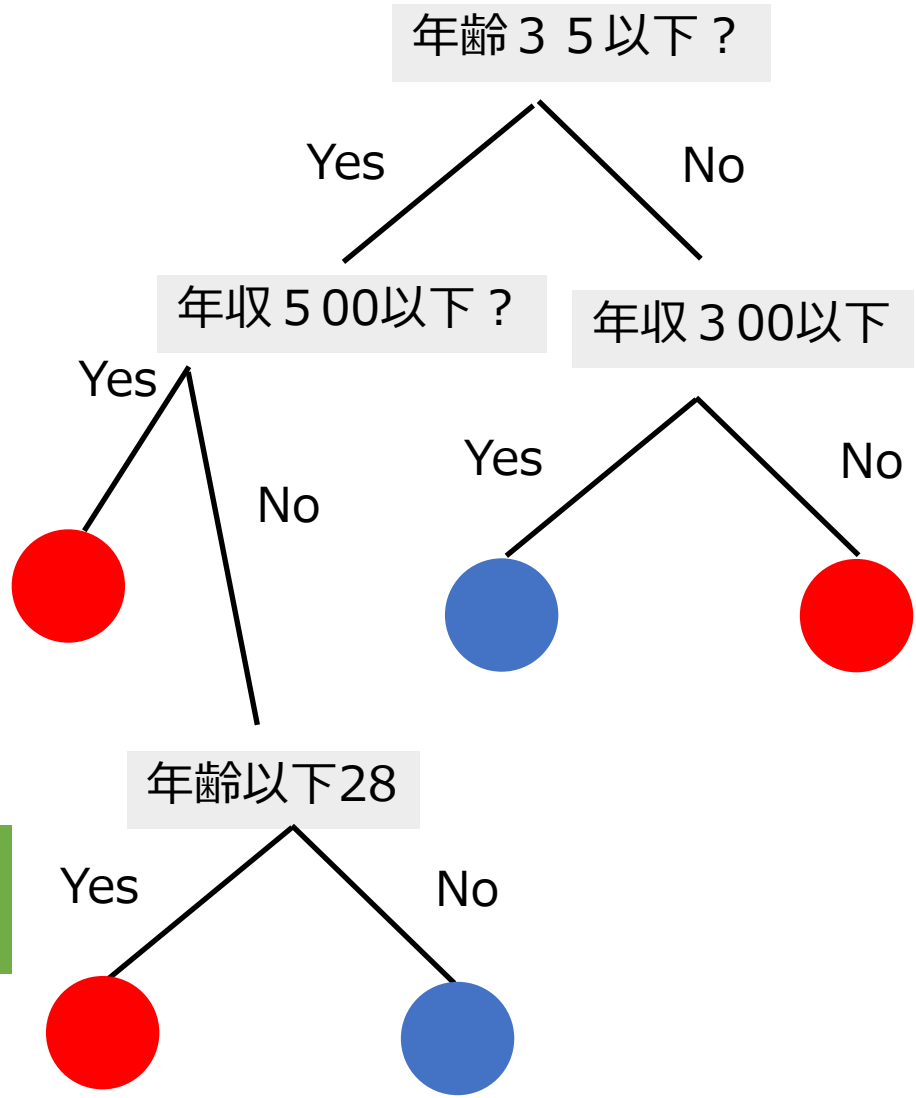




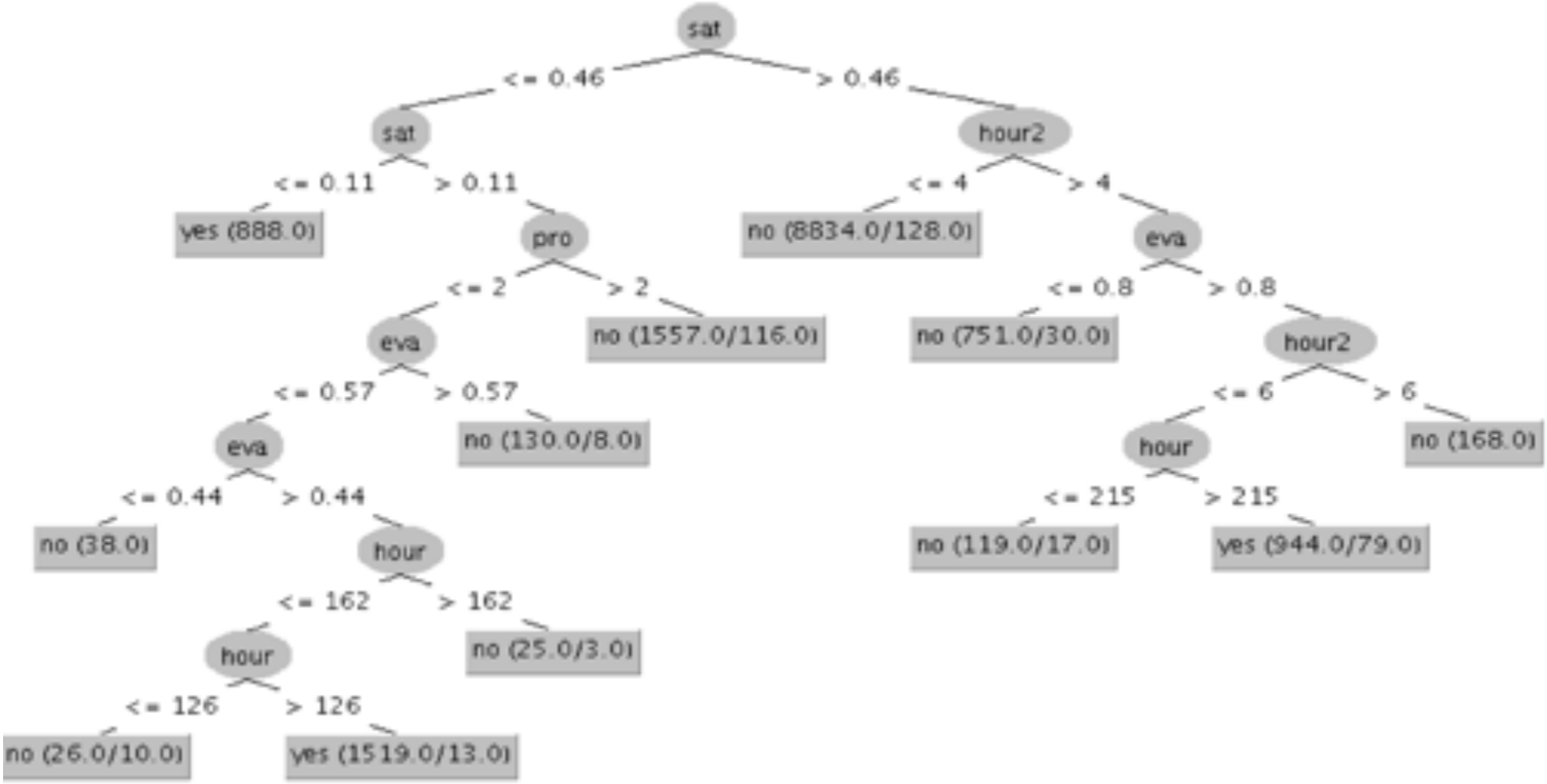
問題
年収４００万、年齢３０歳？



確率が８０％



決定木を使った分析



エクセルハンズオン

- 基本統計量の計算
- データの正規化
- クロス集計
- 移動平均法
- MVプロット

(演習1)データの要約

演習問題１

下のデータはあるサイトのアクセス数データです。
・このサイトのデータを要約せよ。

date	アクセス数
2016/9/1	3200
2016/9/2	3195
2016/9/3	3350
2016/9/4	3115
2016/9/5	3200
2016/9/6	3155
2016/9/7	3260
2016/9/8	3115
2016/9/9	3190
2016/9/10	3635
2016/9/11	3440
2016/9/12	3325
2016/9/13	3230
2016/9/14	3150
2016/9/15	3270

代表値	関数	統計量
平均値	=AVERAGE(配列)	
標準偏差	=stdev.s(配列)	
最小値	=min(配列)	
２５%値	=quartile.inc(配列,1)	
中央値	=median(配列)	
７５%値	=quartile.inc(配列,3)	
最大値	=max(配列)	

※本件は2016年データから抜粋しているため，標準偏差は=stdev.s(配列)を使用する。

(演習1)データの要約

演習問題1

下のデータはあるサイトのアクセス数データです。
・このサイトのデータを要約せよ。

date	アクセス数
2016/9/1	3200
2016/9/2	3195
2016/9/3	3350
2016/9/4	3115
2016/9/5	3200
2016/9/6	3155
2016/9/7	3260
2016/9/8	3115
2016/9/9	3190
2016/9/10	3635
2016/9/11	3440
2016/9/12	3325
2016/9/13	3230
2016/9/14	3150
2016/9/15	3270
2016/9/16	3130

代表値	関数	
平均値	=AVERAGE(配列)	=AVERAGE(C9:C130)
標準偏差	=stdev.s(配列)	
最小値	=min(配列)	
２５%値	=quartile.inc(配列,1)	
中央値	=median(配列)	
７５%値	=quartile.inc(配列,3)	
最大値	=max(配列)	

※本件は2016年データから抜粋しているため、標準偏差は=stdev.s(配列)を使用する。

Tips:
データを端まで選択するときは、
「Ctrl + Shift + ↓」

(参考)excel必要操作 | (演習1)データの要約 | (演習2)質的変数の集計 | (演習3)2変数の集計 | (演習4)人事データ可視化 | (演習5)人事データ可視化2

(演習1)データの要約

演習問題１

下のデータはあるサイトのアクセス数データです。
・このサイトのデータを要約せよ。

date	アクセス数
2016/9/1	3200
2016/9/2	3195
2016/9/3	3350
2016/9/4	3115
2016/9/5	3200
2016/9/6	3155
2016/9/7	3260
2016/9/8	3115
2016/9/9	3190
2016/9/10	3635
2016/9/11	3440
2016/9/12	3325
2016/9/13	3230
2016/9/14	3150
2016/9/15	3270
2016/9/16	3120

代表値	関数	統計量
平均値	=AVERAGE(配列)	=AVERAGE(C9:C130)
標準偏差	=stdev.s(配列)	=STDEV.S(C9:C130)
最小値	=min(配列)	=MIN(C9:C130)
２５%値	=quartile.inc(配列,1)	=QUARTILE.INC(C9:C130,1)
中央値	=median(配列)	=MEDIAN(C9:C130)
７５%値	=quartile.inc(配列,3)	=QUARTILE.INC(C9:C130,3)
最大値	=max(配列)	=MAX(C9:C130)

解答の表示

1.シートを右クリック、「再表示」を選択

2. 再表示したいシートを選択

解答の表示

3. 解答が再表示される

date	アクセス数
2016/9/1	3200
2016/9/2	3195
2016/9/3	3350
2016/9/4	3115
2016/9/5	3200
2016/9/6	3155
2016/9/7	3260
2016/9/8	3115
2016/9/9	3190
2016/9/10	3635
2016/9/11	3440
2016/9/12	3325
2016/9/13	3230
2016/9/14	3150
2016/9/15	3270
2016/9/16	3120
2016/9/17	2782
2016/9/18	2759
2016/9/19	2692
2016/9/20	2772
2016/9/21	2725

代表値	関数	統計量
平均値	=AVERAGE(配列)	=AVERAGE(C9:C130)
標準偏差	=stdev.s(配列)	=STDEV.S(C9:C130)
最小値	=min(配列)	=MIN(C9:C130)
２５%値	=quartile.inc(配列,1)	=QUARTILE.INC(C9:C130,1)
中央値	=median(配列)	=MEDIAN(C9:C130)
７５%値	=quartile.inc(配列,3)	=QUARTILE.INC(C9:C130,3)
最大値	=max(配列)	=MAX(C9:C130)

※本件は2016年データから抽出として考えたため、標準偏差は=stdev.s(配列)を使用した。

(演習2)質点変数の集計

演習問題2

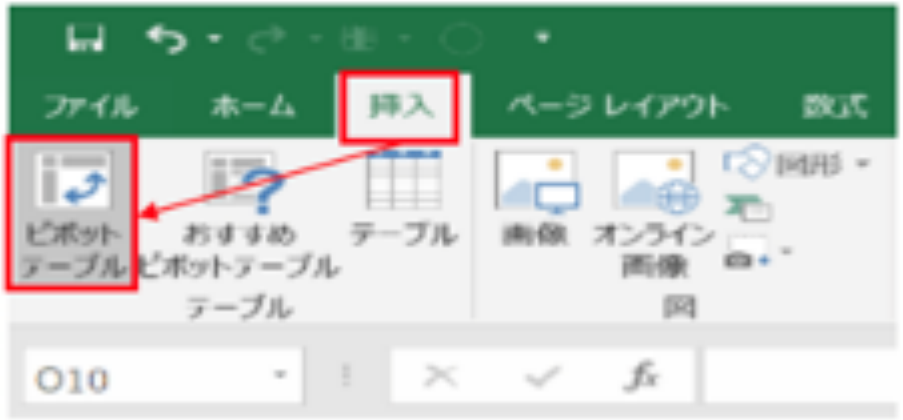
給料データを要約せよ。
（1）クロス集計（ピボットテーブル）
（2）円グラフ作成（質的データ）

ID	満足度	他者評価	プロジェクト数	月間労働時間	労働時間（会社内）	Work accident	退職・在職	過去5年昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.66	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low
6960	0.79	0.84	4	171	3	0	在職	無	sales	low
13755	0.96	0.48	4	198	7	0	在職	無	sales	medium
4754	1	0.84	3	154	3	0	在職	無	sales	medium
12906	0.97	0.9	5	262	3	0	在職	無	sales	medium
9150	0.56	0.41	6	142	3	0	在職	無	product_mng	medium
1138	0.87	0.88	5	262	6	0	退職	無	sales	low
6866	0.23	0.88	5	238	6	0	在職	無	RandD	medium
11765	0.79	0.65	3	235	10	0	在職	無	technical	low
8342	0.83	0.84	4	206	2	0	在職	無	sales	medium

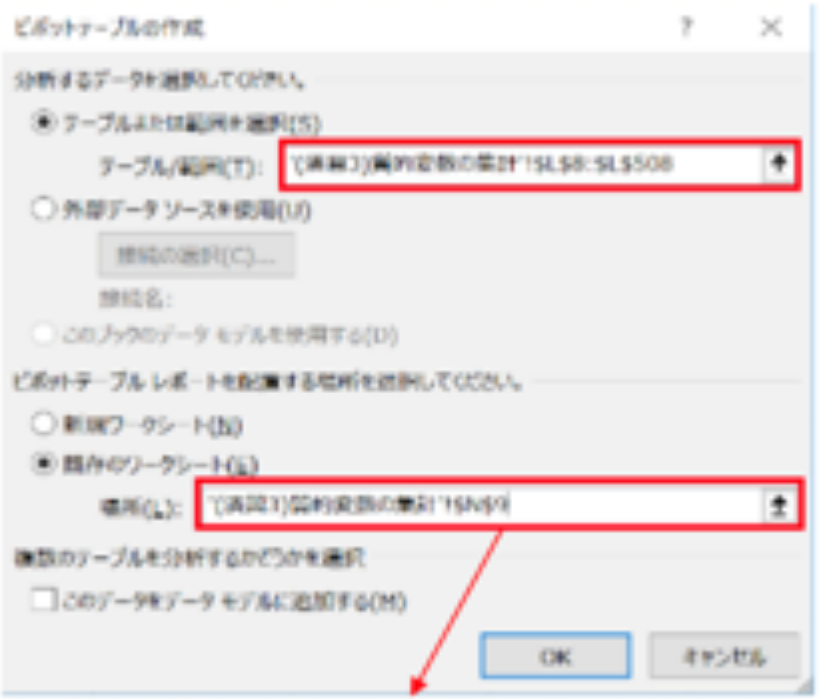
(演習2)質点変数の集計

（1）クロス集計（ピボットテーブル）

①「挿入」→「ピボットテーブル」を選択。



②ピボットテーブルでデータ範囲を選択する。



・データを抽出するセルを選択する。

（※任意の好きな場所で可）

(演習2)質点変数の集計

③ピボットテーブルに集計したい内容を選択する。

ここを選択すると集計する内容を変更できる。

※ドラッグしてそれぞれの場所に振り分ける。

④完成

行ラベル	個数 / 給料
high	41
low	247
medium	212
総計	500

(演習2)質点変数の集計

（2）円グラフ作成（質的データ）

≪ピボットグラフの場合≫

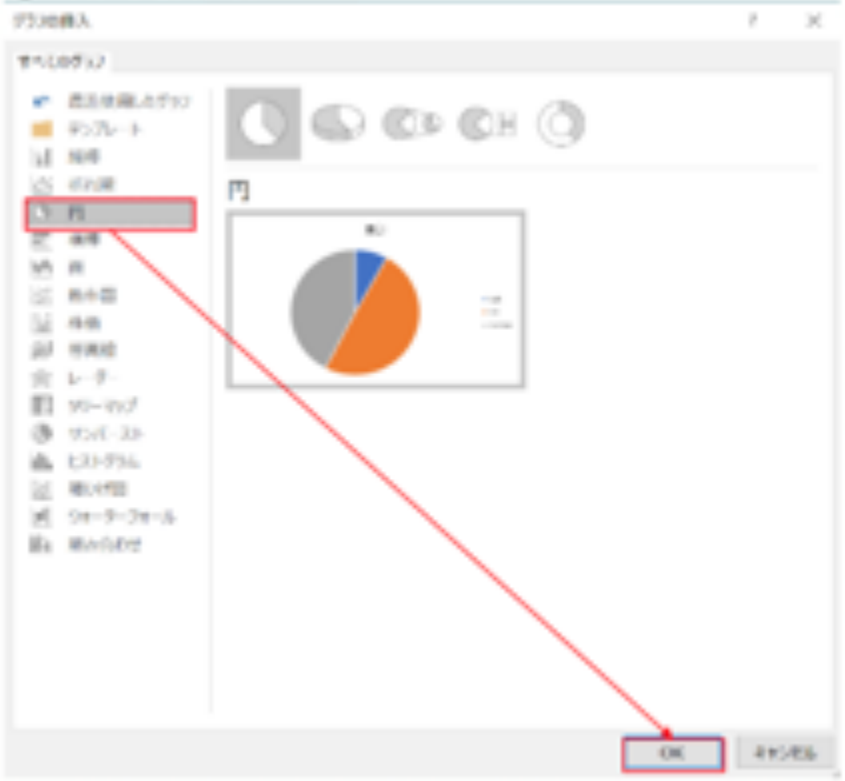
①ピボットテーブルで集計した範囲を選択する。

行ラベル	個数 / 給料
high	41
low	247
medium	212
総計	500

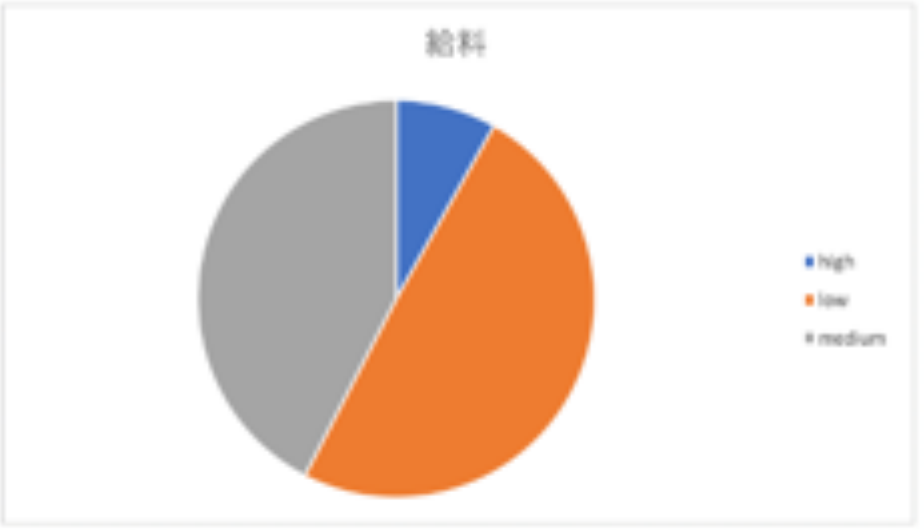
②「挿入」→「ピボットグラフ」を選択する。

(演習2)質点変数の集計

③「円」を選択する。



④完成



※必要に応じて、グラフタイトル/系列名などを修正して完成。
（グラフについての細かい作成方法は各自お調べください。）

(演習3)2変数の集計

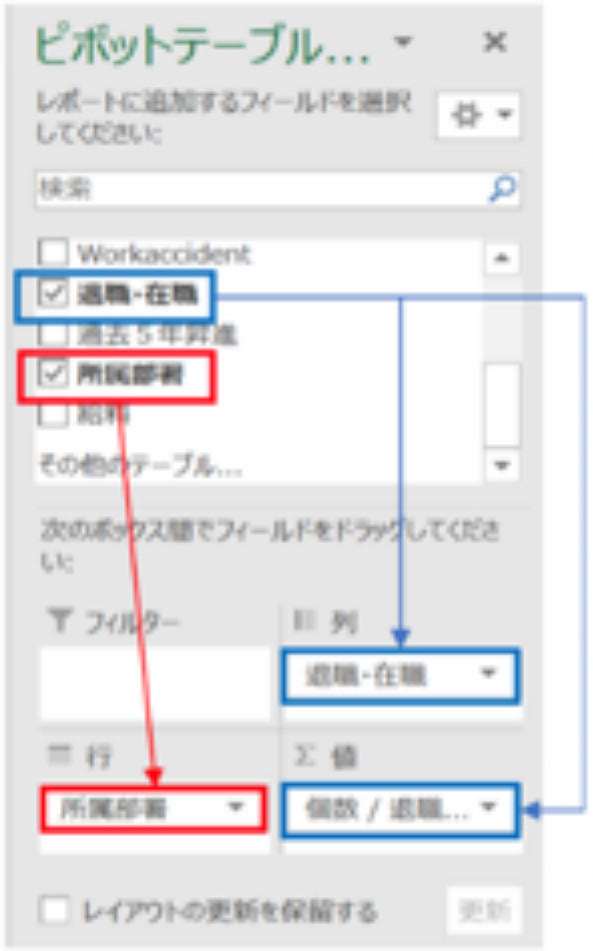
演習問題 3

部署と退職・在職データの集計を行い、グラフ化せよ。

ID	満足度	評価	プロジェクト数	月間労働時間	労働時間（会社内）	Work accident	退職・在職	過去5年昇進	所属部署	給料
1019	0.35	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.65	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low
6960	0.79	0.84	4	171	3	0	在職	無	sales	low
13755	0.96	0.48	4	198	7	0	在職	無	sales	medium
4754	1	0.84	3	154	3	0	在職	無	sales	medium
12906	0.97	0.9	5	262	3	0	在職	無	sales	medium
9150	0.56	0.41	6	142	3	0	在職	無	product_mng	medium
1138	0.87	0.88	5	262	6	0	退職	無	sales	low
6866	0.23	0.88	5	238	6	0	在職	無	RandD	medium
11765	0.79	0.65	3	235	10	0	在職	無	technical	low
8342	0.83	0.84	4	206	2	0	在職	無	sales	medium
5103	0.92	0.55	3	259	3	0	在職	無	product_mng	low
14114	0.93	0.89	3	255	7	1	在職	無	sales	medium
5075	0.81	0.75	2	100	4	0	在職	無	technical	low
11062	0.9	0.73	2	203	4	0	在職	無	support	medium

(演習3)2変数の集計

◎ピボットテーブルの選択範囲



※左記のように質的データは
列・値の2つに選択すると、
その数を集計できる。

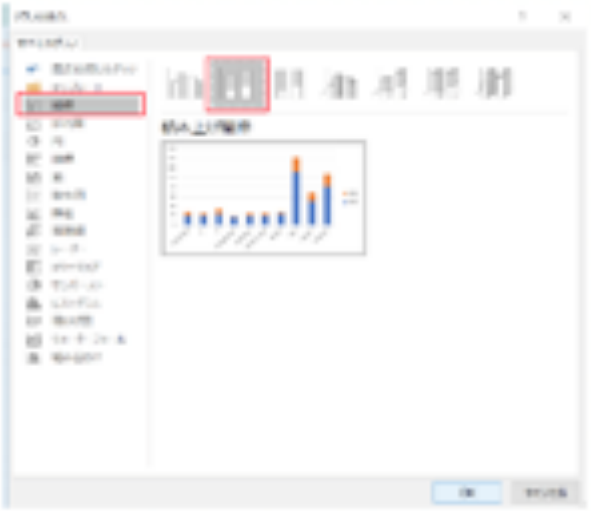
(演習3)2変数の集計

◎ピボットグラフのデータ範囲

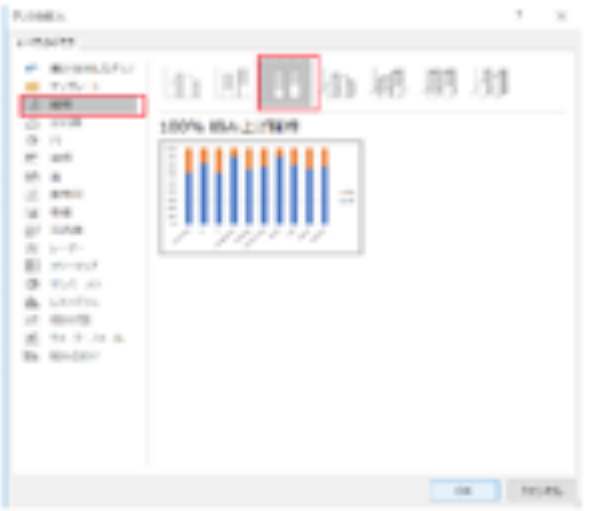
データの個数 / 退職列ラベル			
行ラベル	在職	退職	総計
accounting	19	9	28
hr	20	5	25
IT	23	11	34
management	17	2	19
marketing	18	7	25
product_mng	19	6	25
RandD			
sales			
support			
technical			
総計	382	118	500

※総計を含めてもグラフができる。

積み上げ棒グラフ



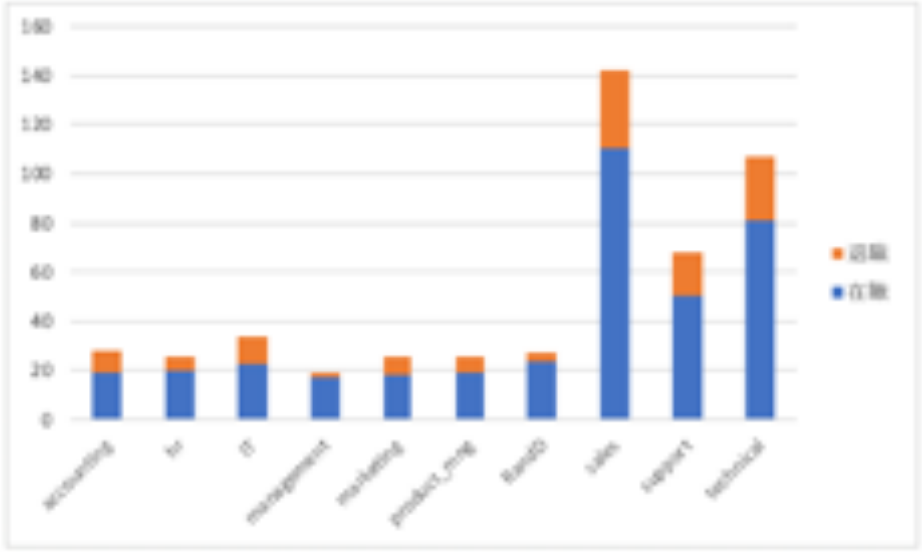
100%
積み上げ棒グラフ



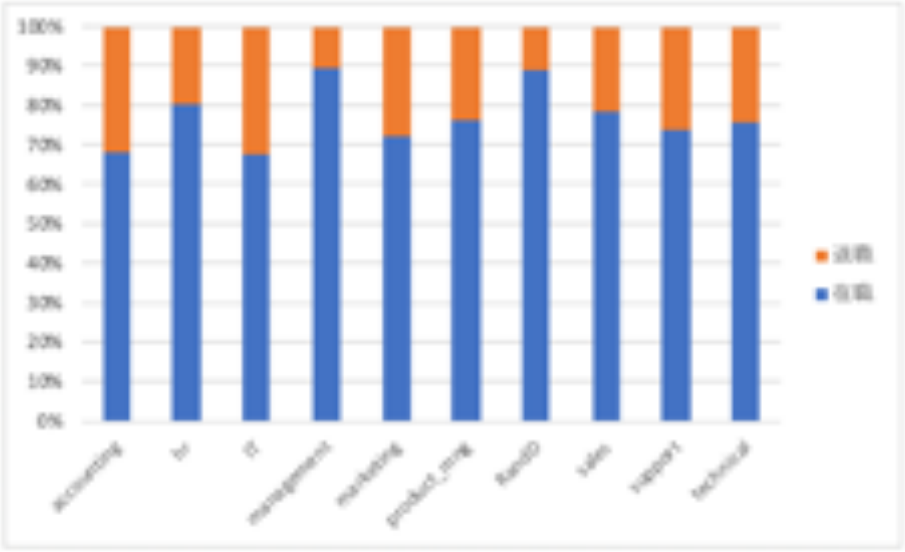
(演習3)2変数の集計

（2）グラフ作成（質的データ）

◎積み上げ縦棒グラフ



◎100%積み上げ縦棒グラフ



(演習4)人事データ可視化

演習問題4

所属部署の集計を行い、グラフ化せよ。
（1）クロス集計（ピボットテーブル）
（2）質的データ（棒グラフ・円グラフ）

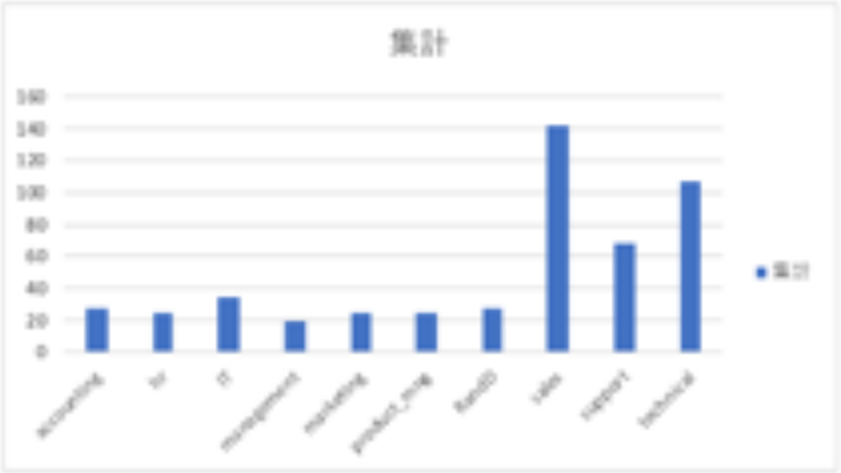
ID	満足度	他者評価	プロジェクト数	月間労働時間	労働時間（会社内）	Work accident	退職・在職	過去5年昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.66	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low
6960	0.79	0.84	4	171	3	0	在職	無	sales	low
13755	0.96	0.48	4	198	7	0	在職	無	sales	medium
4754	1	0.84	3	154	3	0	在職	無	sales	medium
12906	0.97	0.9	5	262	3	0	在職	無	sales	medium
9150	0.56	0.41	6	142	3	0	在職	無	product_mng	medium
1138	0.87	0.88	5	262	6	0	退職	無	sales	low
6866	0.23	0.88	5	238	6	0	在職	無	RandD	medium
11765	0.79	0.65	3	235	10	0	在職	無	technical	low
8342	0.83	0.84	4	206	2	0	在職	無	sales	medium

(演習4)人事データ可視化

（1）クロス集計（ピボットテーブル）

行ラベル	個数 / 所属部署
accounting	28
hr	25
IT	34
management	19
marketing	25
product_mng	25
RandD	27
sales	142
support	68
technical	107
総計	500

（2）質的データ（棒グラフ・円グラフ）



(演習5)人事データ可視化2

演習問題5

部署と給料データの集計を行い、グラフ化せよ。

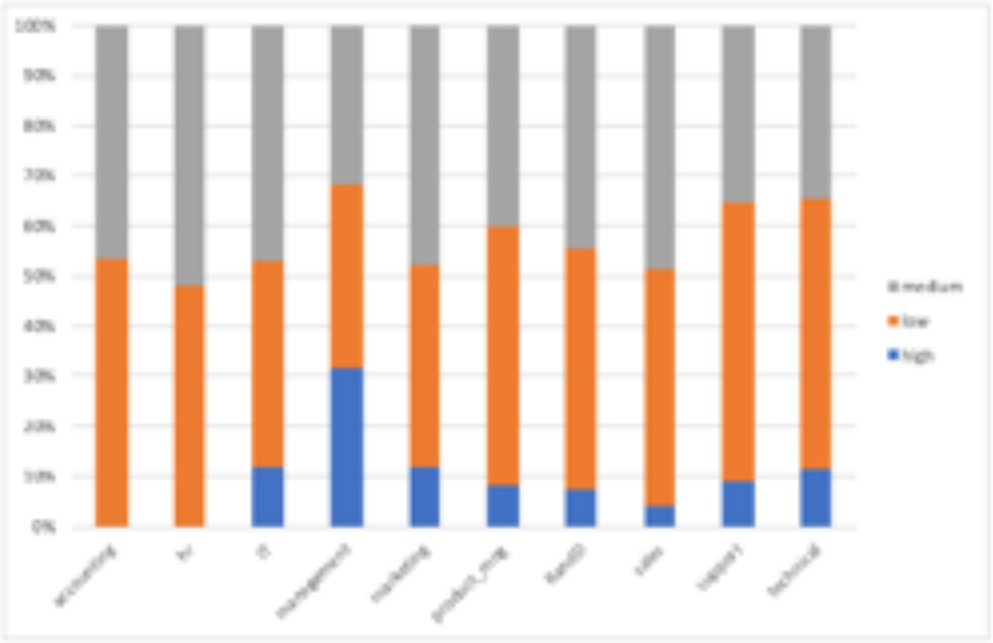
ID	満足度	社内評価	プロジェクト数	月間労働時間	労働時間（会社内）	Work accident	退職・在職	過去5年昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.66	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low
6960	0.79	0.84	4	171	3	0	在職	無	sales	low
13755	0.96	0.48	4	198	7	0	在職	無	sales	medium
4754	1	0.84	3	154	3	0	在職	無	sales	medium
12906	0.97	0.9	5	262	3	0	在職	無	sales	medium
9150	0.56	0.41	6	142	3	0	在職	無	product_mng	medium
1138	0.87	0.88	5	262	6	0	退職	無	sales	low
6866	0.23	0.88	5	238	6	0	在職	無	RandD	medium
11765	0.79	0.65	3	235	10	0	在職	無	technical	low
8342	0.83	0.84	4	206	2	0	在職	無	sales	medium

(演習5)人事データ可視化2

（1）クロス集計（ピボットテーブル）

データの個数 / 給料	列ラベル			
行ラベル	high	low	medium	総計
accounting		15	13	28
hr		12	13	25
IT	4	14	16	34
management	6	7	6	19
marketing	3	10	12	25
product_mng	2	13	10	25
RandD	2	13	12	27
sales	6	67	69	142
support	6	38	24	68
technical	12	58	37	107
総計	41	247	212	500

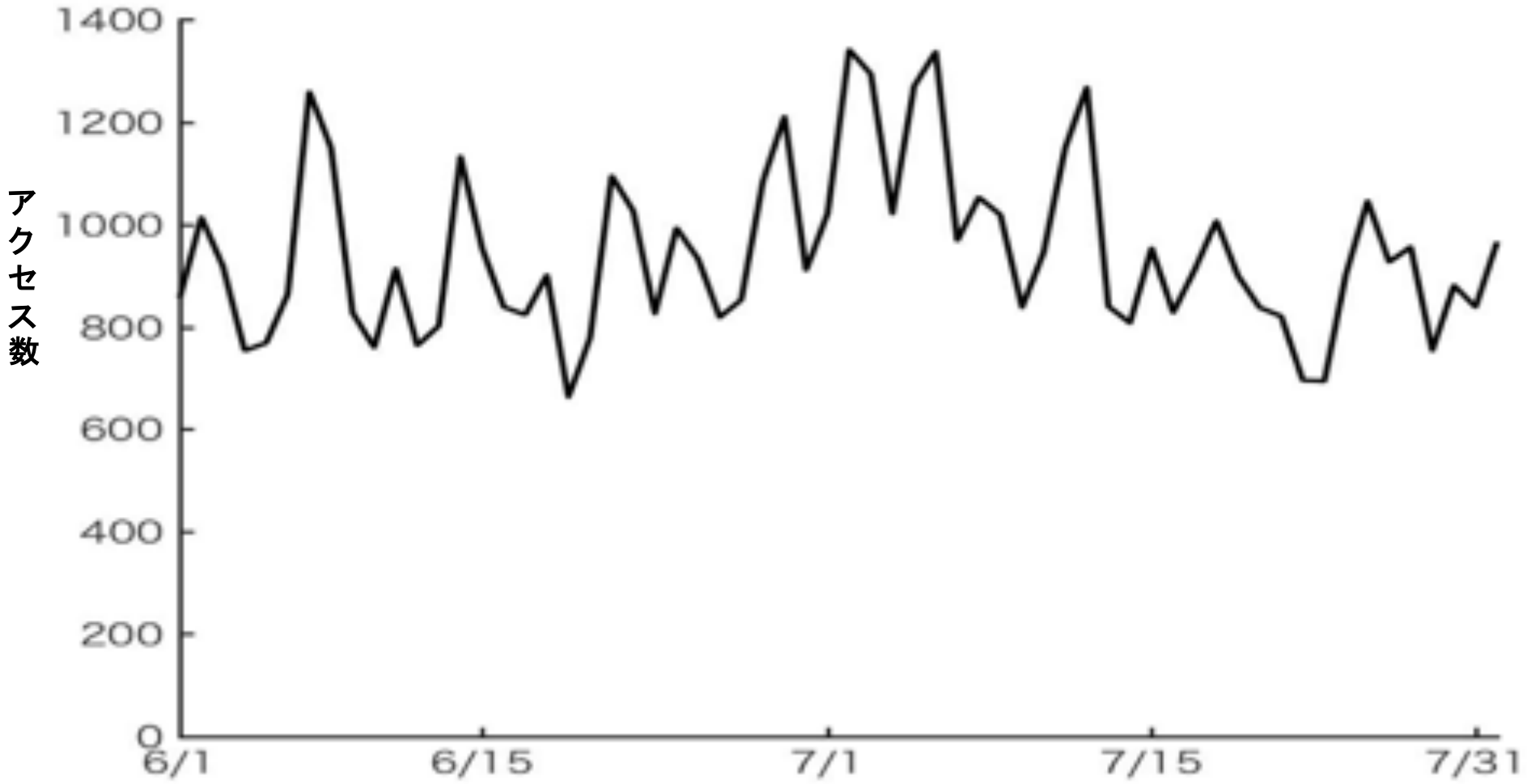
（2）質的データ（棒グラフ）



移動平均法

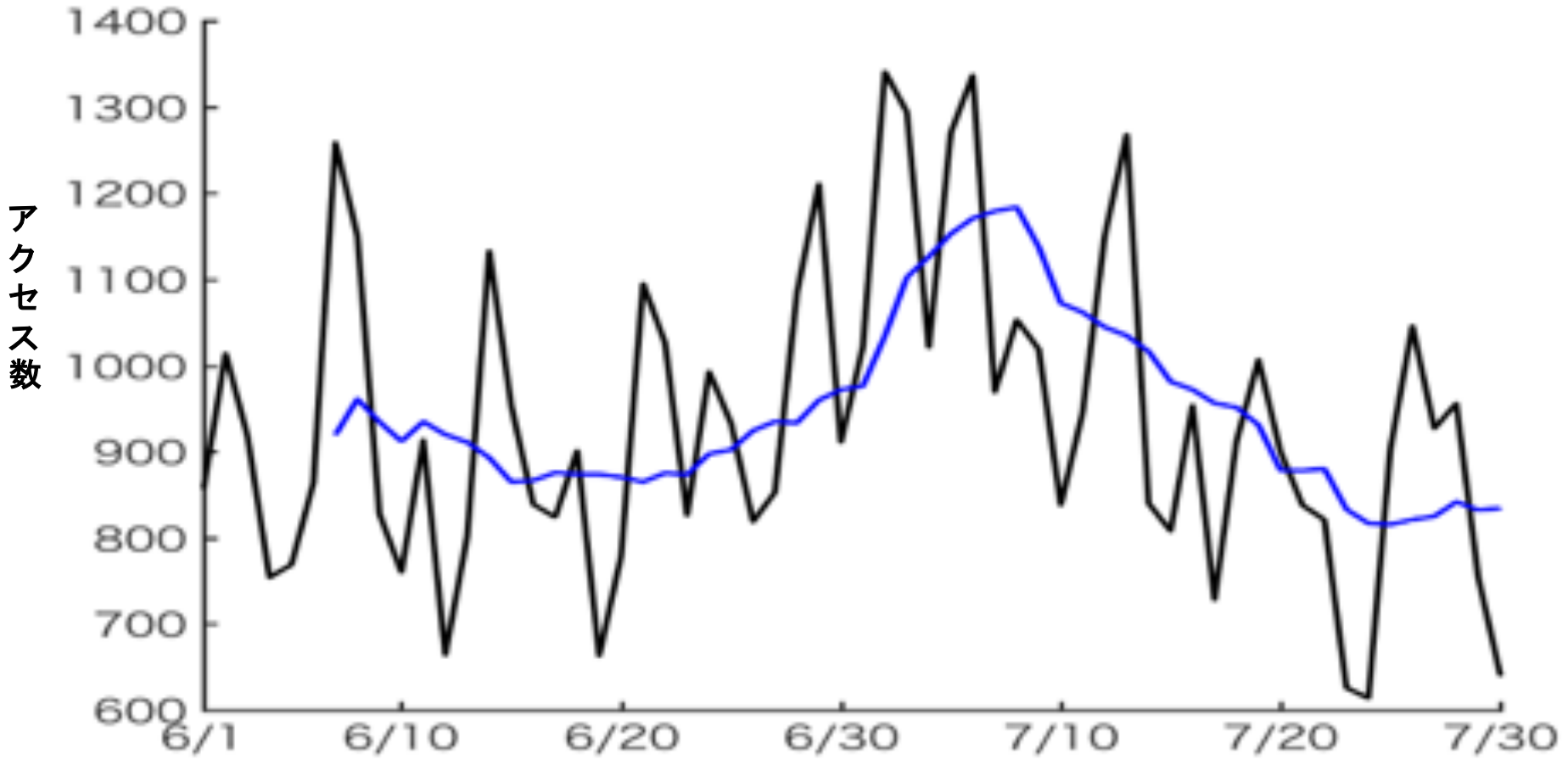
移動平均法を使ったトレンドの抽出

課題：「アクセス数のトレンドを推定せよ」

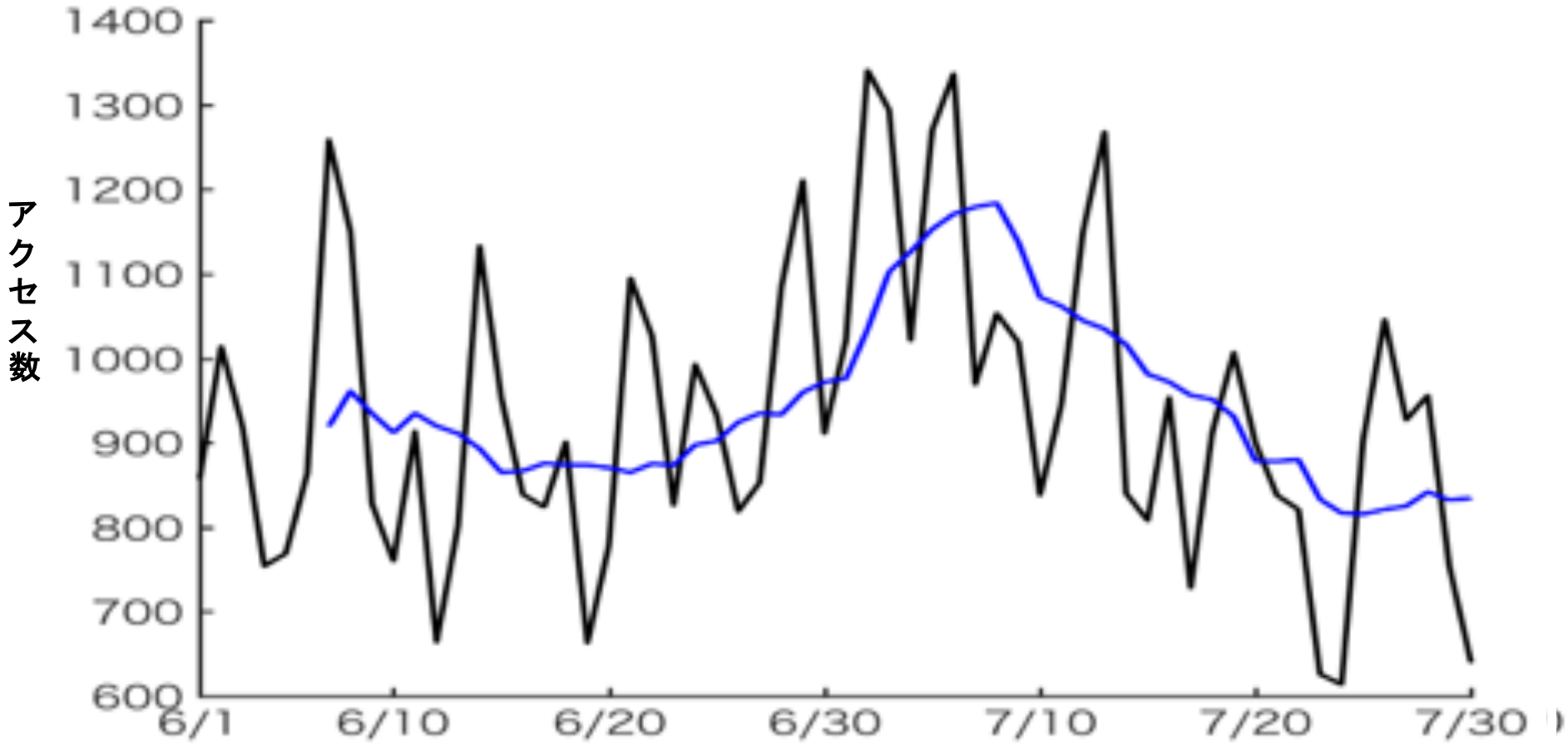


移動平均法を使ったトレンドの抽出

時系列データ　＝　トレンド＋　周期変動　＋　不規則変動



ベースラインの推定



エクセルハンズオン

- 移動平均法
- MVプロット

(演習6)移動平均法

演習問題6

下のデータはあるサイトのアクセス数データです。
・このサイトのアクセス傾向を抽出せよ。

date	アクセス数
2016/9/1	3200
2016/9/2	3195
2016/9/3	3350
2016/9/4	3115
2016/9/5	3200
2016/9/6	3155
2016/9/7	3260
2016/9/8	3115
2016/9/9	3190
2016/9/10	3635
2016/9/11	3440
2016/9/12	3325
2016/9/13	3230
2016/9/14	3150
2016/9/15	3270

(演習6)移動平均法

date	アクセス数	7日間移動平均
2016/9/1	3200	
2016/9/2	3195	
2016/9/3	3350	
2016/9/4	3115	
2016/9/5	3200	
2016/9/6	3155	
2016/9/7	3260	3210.7143
2016/9/8	3115	3198.5714
2016/9/9	3190	3197.8571
2016/9/10	3635	3238.5714
2016/9/11	3440	3285.0000
2016/9/12	3325	3302.8571
2016/9/13	3230	3313.5714
2016/9/14	3150	3297.8571
2016/9/15	3270	3320.0000
2016/9/16	3120	3310.0000
2016/9/17	2782	3188.1429
2016/9/18	2759	3090.8571
2016/9/19	2692	3000.4286
2016/9/20	2772	2935.0000

① 左記に「=AVERAGE(C9:C15)」を入力して7日間の平均を求める。
※サイトのアクセス数は1週間を1サイクルとして考え、7日間で移動平均を取った。

②上記で入力した計算式を一番下のセルまでコピーする。

≪計算式を一括でコピーする方法≫

(演習6)移動平均法

③折れ線グラフを作成する

◎グラフ作成方法

1.dateとアクセス数と7日間移動平均を選択し、「挿入」→「折れ線グラフ」を選択。

※グラフはおすすめグラフで作成しても可。

2.グラフタイトル/系列名/横軸ラベルなどを修正して完成。

（グラフについての細かい作成方法は各自お調べください。）

(演習7)二酸化炭素量

演習問題７

２０００年の１月から２０１７年５月までの大気中の二酸化炭素量を示すデータです。
（ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_mm_mlo.txt）
・このデータに移動平均法を使って、二酸化炭素の増加傾向を抽出せよ。

年	月	co2
2000	1	369.29
2000	2	369.54
2000	3	370.6
2000	4	371.82
2000	5	371.58
2000	6	371.7
2000	7	369.86
2000	8	368.13
2000	9	367
2000	10	367.03
2000	11	368.37
2000	12	369.67
2001	1	370.59
2001	2	371.51
2001	3	372.43
2001	4	373.37
2001	5	373.85
2001	6	373.21
2001	7	371.51
2001	8	369.61
2001	9	368.18

(演習7)二酸化炭素量

12ヶ月の移動平均を計算する。

年	月	co2	12ヶ月移動平均
2000	1	369.29	
2000	2	369.54	
2000	3	370.6	
2000	4	371.82	
2000	5	371.58	
2000	6	371.7	
2000	7	369.86	
2000	8	368.13	
2000	9	367	
2000	10	367.01	
2000	11	368.37	
2000	12	369.67	369.5491667
2001	1	370.59	369.6575
2001	2	371.51	369.8216667
2001	3	372.43	369.9741667
2001	4	373.37	370.1033333
2001	5	373.85	370.2925
2001	6	373.21	370.4183333
2001	7	371.91	370.5658333
2001	8	369.61	370.6791667
2001	9	368.18	370.7775

(演習8)MVプロット

演習問題８

ある会社の応募者に５つのテスト（言語、数理、論理、論文、一般）を実施しました。下のデータはそれぞれの応募者の点数データです。
１）各応募者の５つのテストの平均点と標準偏差を求め、散布図を作成せよ。
２）標準偏差と平均点の組み合わせに基づいて、どの応募者を採用するか決定せよ。

ID	言語能力	数理能力	論理力	論文	一般	標準偏差	平均点
1	87	93	61	71	50		
2	40	44	18	72	30		
3	69	74	25	71	33		
4	24	49	60	82	97		
5	42	94	69	70	54		
6	94	59	99	32	20		
7	68	90	27	28	30		
8	82	23	30	72	26		
9	89	90	68	74	69		
10	72	36	67	62	26		
11	46	71	97	78	22		
12	74	61	58	33	18		
13	31	34	25	67	46		
14	68	92	25	87	37		
15	40	65	39	78	98		
16	90	54	94	69	63		
17	74	89	56	73	87		
18	67	65	55	73	45		
19	79	43	60	51	84		

(演習8)MVプロット(1)

① 標準偏差を計算する。
② 平均点を計算する。

ID	言語能力	数理能力	論理力	論文	一般	標準偏差	平均点
1	87	93	61	71	50	=STDEV.S(C11:G11)	=AVERAGE(C11:G11)
2	40	44	18	72	30		
3	69	74	25	71	33		
4	24	49	60	82	97		
5	42	94	69	70	54		
6	94	59	99	32	20		
7	68	90	27	28	30		
8	82	23	30	72	26		
9	89	90	68	74	69		
10	72	36	67	62	26		

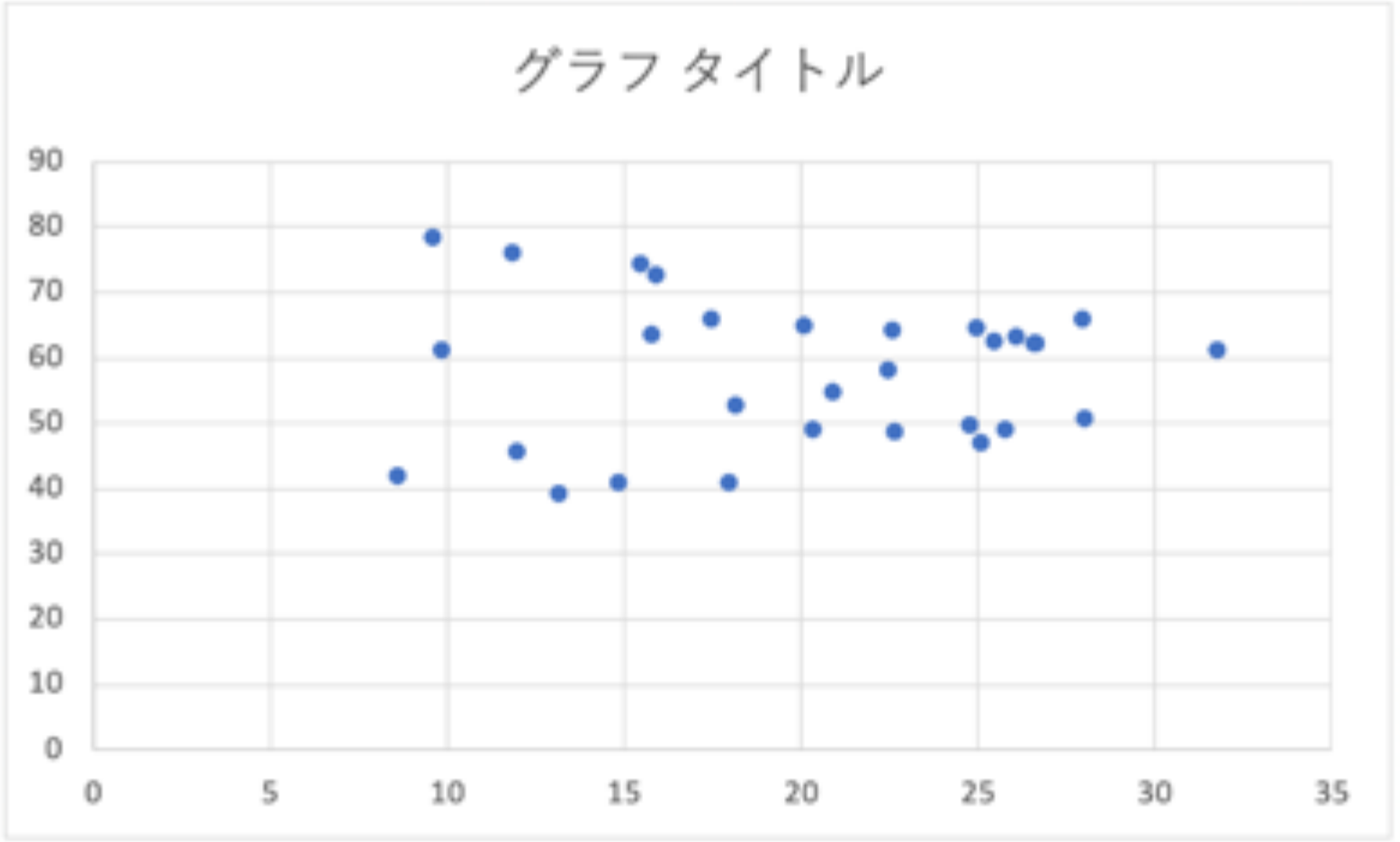
(演習8)MVプロット(1)

③ 下までコピーする

ID	言語能力	数理能力	論理力	論文	一般	標準偏差	平均点
1	87	93	61	71	50	17.8269459	72.4
2	40	44	18	72	30	20.12958022	40.8
3	69	74	25	71	33	23.42648074	54.4
4	24	49	60	82	97	28.46576892	62.4
5	42	94	69	70	54	19.54993606	65.8
6	94	59	99	32	20	35.56262083	60.8
7	68	90	27	28	30	28.84094312	48.6
8	82	23	30	72	26	28.08558349	46.6
9	89	90	68	74	69	10.74709263	78
10	72	36	67	62	26	20.34207462	52.6

(演習8)MVプロット(1)

④散布図を作成する。



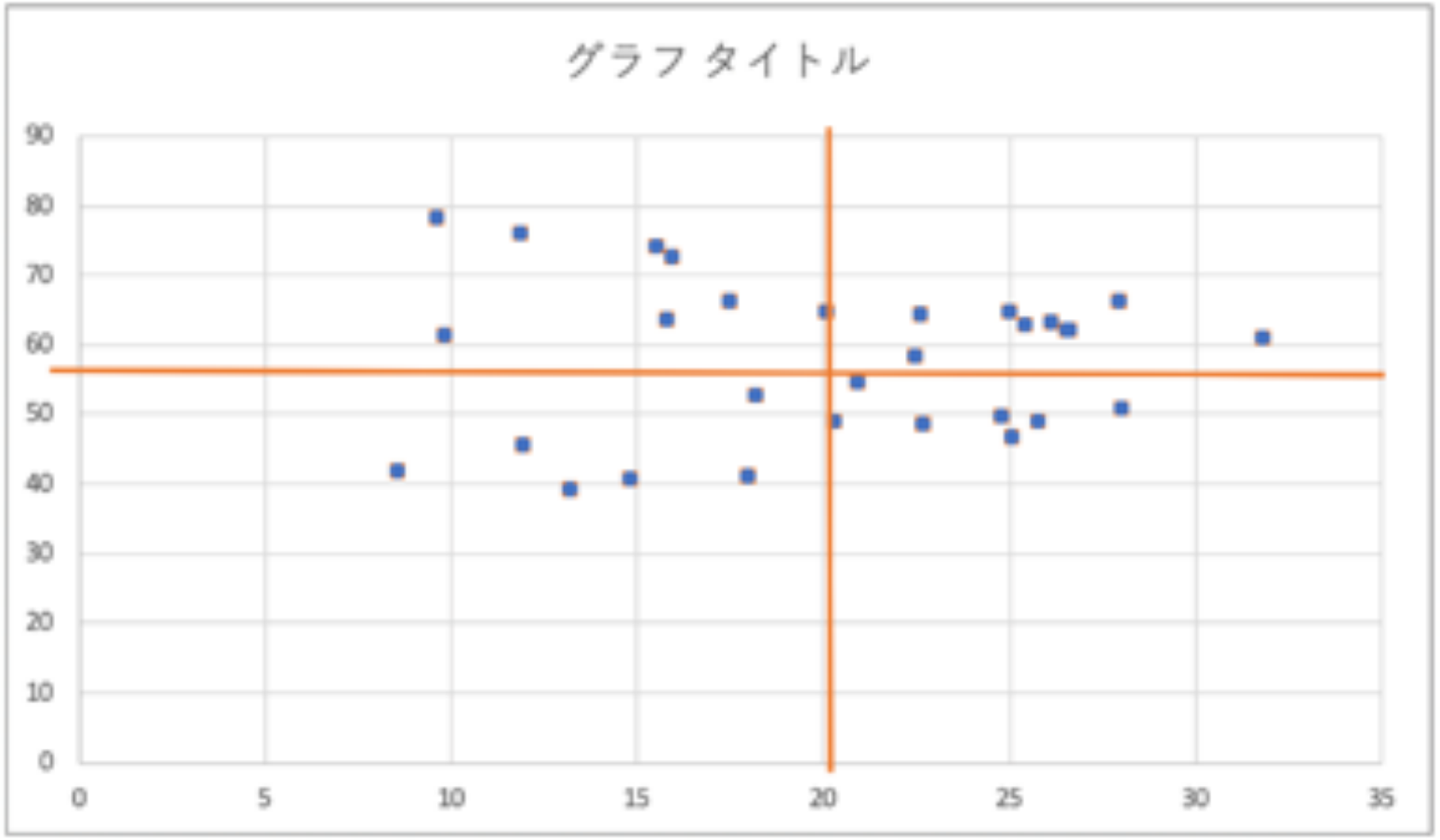
(演習8)MVプロット(2)

1. 各応募者の平均値の平均値を計算する
2. 各応募者の標準偏差の平均値を計算する

ID	言語能力	数理能力	論理力	論文	一般	標準偏差	平均点
1	87	93	61	71	50	15.94490514	72.4
2	40	44	18	72	30	18.0044439	40.8
3	69	74	25	71	33	20.95328137	54.4
4	24	49	60	82	97	25.46055773	62.4
5	42	94	69	70	54	17.4859944	65.8
6	94	59	99	32	20	31.80817505	60.8
7	68	90	27	28	30	25.79612374	48.6
8	82	23	30	72	26	25.12050955	46.6
⋮							
20	56	38	52	25	55	11.98999583	45.2
21	100	76	83	50	20	27.9885691	65.8
22	73	81	40	31	97	24.99279896	64.4
23	47	47	85	51	93	20.13554072	64.6
24	37	52	34	33	52	8.593020424	41.6
25	28	85	22	84	33	28.06136134	50.4
26	48	94	64	59	25	22.45885126	58
27	28	52	55	21	39	13.19090596	39
28	34	39	44	32	93	22.6856783	48.4
29	86	20	44	28	69	24.8	49.4
30	90	49	95	51	25	26.57818654	62
						20.11663418	57.44666667

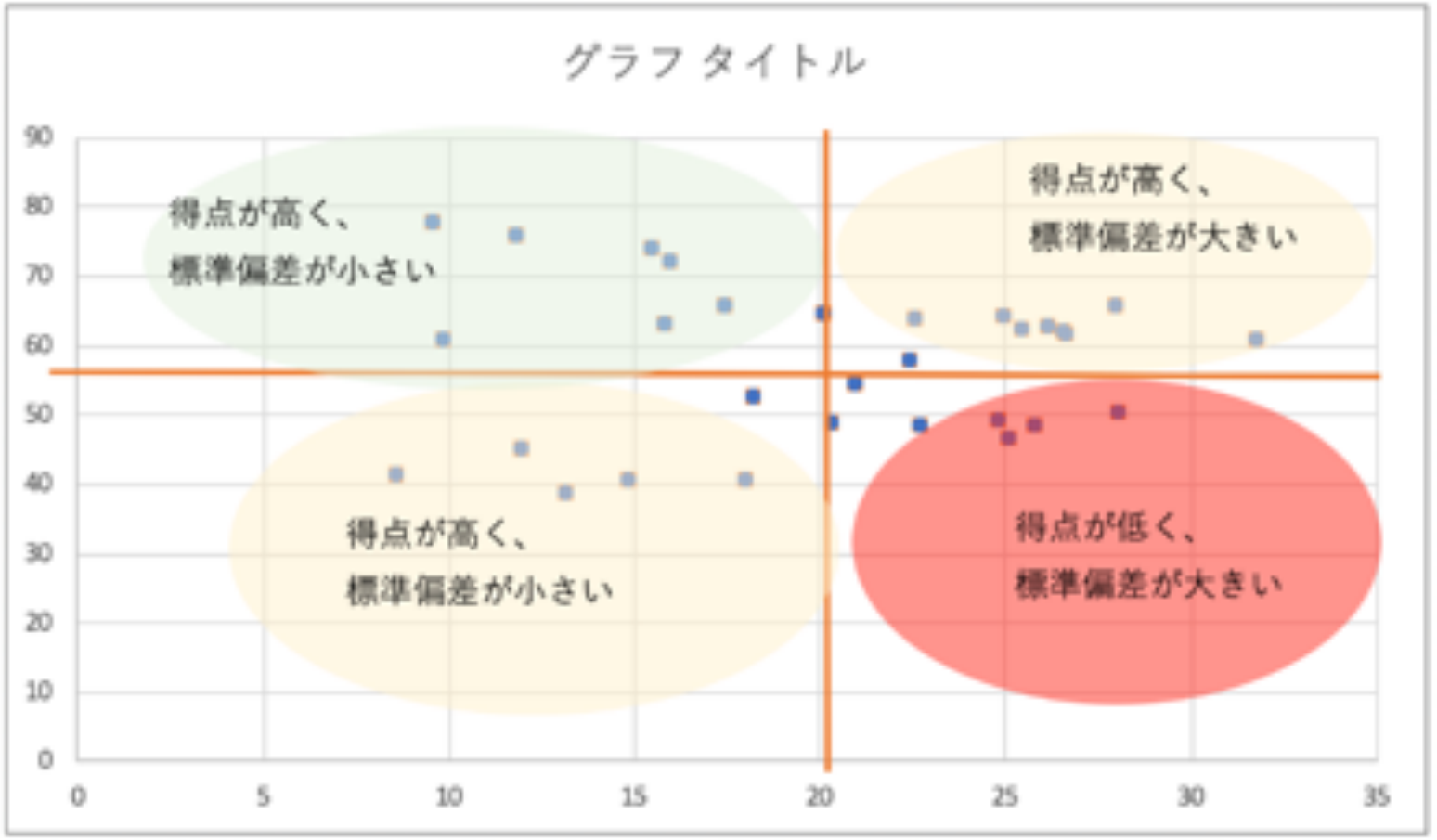
(演習8)MVプロット(2)

3. 求めた平均値を、(1)で作成した散布図に記入する。
（挿入→図形→縦棒を選択して縦棒を配置する）



(演習8)MVプロット(2)

4. データを分類する。



(演習9)タイタニック

1＝上層クラス（お金持ち）
2＝中級クラス（一般階級）
3＝下層クラス（労働階級）

チケットクラス	生存・死亡	性別	年齢	同乗している兄弟／配偶者の数	同乗している親／子供の数
3rd	survived	female	0	1	2
3rd	died	male	0	0	2
3rd	survived	male	0	0	1
2nd	survived	male	0	1	1
3rd	survived	female	0	2	1
3rd	survived	female	0	2	1
3rd	died	male	0	1	1
2nd	survived	male	0	0	2
2nd	survived	male	0	1	1
3rd	survived	male	0	0	1
1st	survived	male	0	1	2
2nd	survived	female	0	1	2
2nd	survived	male	1	2	1
2nd	survived	female	1	1	2
2nd	survived	male	1	0	2
3rd	survived	male	1	1	2
3rd	died	male	1	5	2
3rd	survived	female	1	1	1