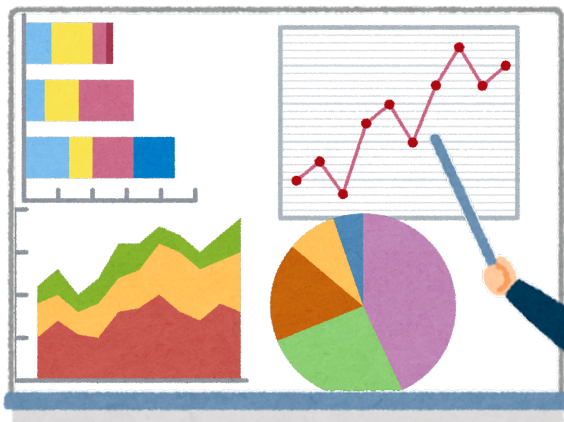




CITY
UNIV.

データサイエンス・リテラシー

データの可視化



イラスト：©いらすとや

Table of Contents

データの可視化

- データの可視化とは
- 可視化の事例
- 可視化が必要な場面と方法

代表的な可視化手法

- | | |
|---------|---------|
| ●棒グラフ | ●円グラフ |
| ●ヒストグラム | ●箱髭グラフ |
| ●折れ線グラフ | ●ネットワーク |
| ●散布図 | |

演習:

- データの整理・可視化演習



Table of Contents

データの可視化

- データの可視化とは
- 可視化の事例
- 可視化が必要な場面と方法

代表的な可視化手法

- 棒グラフ
- ヒストグラム
- 折れ線グラフ
- 散布図
- 円グラフ
- 箱髭グラフ
- ネットワーク

演習:

- データの整理・可視化演習

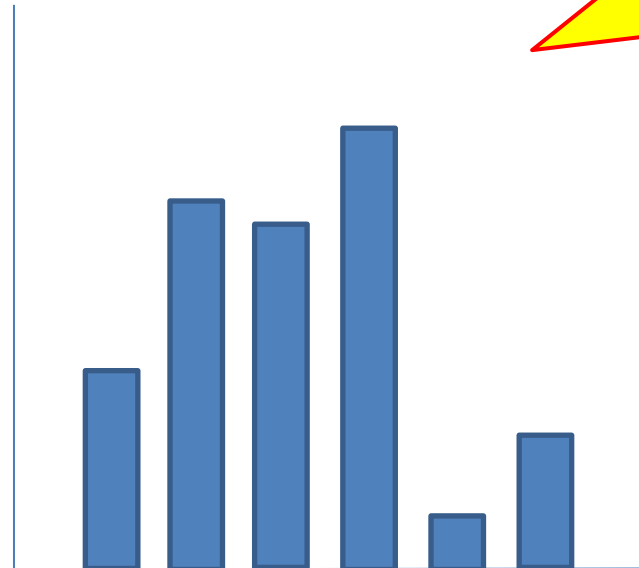


データの可視化とは

データ

コンピュータでは
扱いやすい

可視化



人間が
見やすい図に



気づき
知識

イラスト：©いらすとや

データの可視化とは

例えば、「東京の新型コロナウイルス感染症対策サイト」

<https://stopcovid19.metro.tokyo.lg.jp/>

都内の最新感染動向

新型コロナウイルス感染症が心配なときに

新型コロナウイルス感染症で入院・宿泊療養される方へ

新型コロナウイルス感染症の患者発生状況に関するよくあるご質問

ご家庭でのマスク等の捨て方

お子様をお持ちの皆様へ

妊産婦の皆様へ

都民の皆様へ

企業の皆様・はたらく皆様へ

東京都新型コロナウイルス感染症対策本部報

感染状況・医療提供体制（サマリ） 2021年6月20日時点

感染状況 新規陽性者 376 人 / 検査数 6,400.3 件（2021年6月18日 参考値（3日間移動平均））、うち65歳以上の高齢者数 25 人、死亡者数 1 人、都外からの持込検体による陽性数 2

医療提供体制 入院数 1,270 人（確保病床数 5,594 床）、うち重症者数 45 人（うち重症病床数 373 床）

感染状況・医療提供体制の分析

新規陽性者数の増加比は上昇しており、今後100%を超えることが強く懸念される。加の抑制、感染防止対策を徹底し、感染の再拡大を防止しなければならない。再び増加に転じれば、医療提供体制の逼迫を招く。

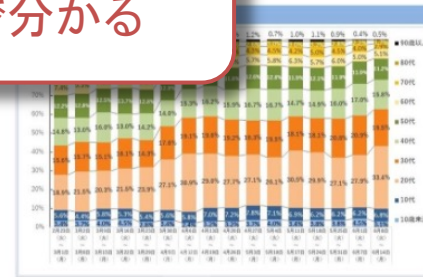
感染状況

感染の再拡大の危険性が高いと思われる

医療提供体制

通常の医療が大きく制限されていると思われる

感染状況や医療提供体制の状況がひと目で分かる



＞ 専門家による分析・総括コメントの詳細はこちら

都内の滞在人口の増減状況

＞ 詳細はこちら

池袋

[2021年6月19日時点]
2020年1月比 ↓46 %

自分や家族の症状に不安や心配があれば
まずは電話相談をどうぞ

相談の手順を見る

データの可視化とは

例えば、「東京の新型コロナウイルス感染症対策サイト」

<https://stopcovid19.metro.tokyo.lg.jp/>

検査陽性者の状況

陽性者数 (累計)	170,945 人
入院	1,360 人
軽症・中等症	1,317 人
重症	43 人
宿泊療養	921 人
自宅療養	822 人
入院・療養等調整中	675 人
死亡	2,216 人
退院等 (療養期間経過を含む)	164,951 人

死亡日別による死亡者数の推移はこちら

- (注)
- チャーター機帰国者、クルーズ船乗客等は含まれていない
 - 「重症」は、人工呼吸器管理 (ECMOを含む) が必要な患者数を計上。重症基準の考え方はこちら
 - 退院者数の把握には一定の期間を要しており、確認次第数値を更新している

オープンデータを入手
2021年6月24日 18:00 更新

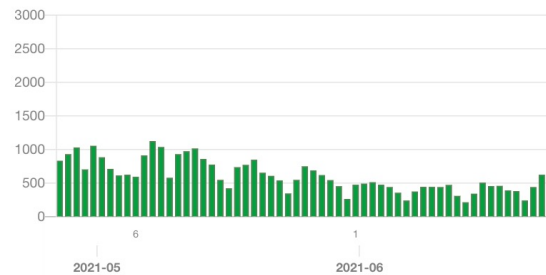
報告日別による陽性者数の推移

570 人

2021年6月24日 日別値 (前日比: -49 人)

発症日別による陽性者数の推移はこちら

日別 累計



65歳以上の新規陽性者数の推移及び届出保健所別の内訳

- (注)
- 保健所から発生届が提出された日を基準とする
 - 医療機関等が行った検査も含む
 - チャーター機帰国者、クルーズ船乗客等は含まれていない

> テ

統計ダッシュボード

- 直感的にわかりやすく情報を表示
- ニーズや目的に応じて必要な情報を追加したり、深掘り可能

モニタリング項目

感染状況

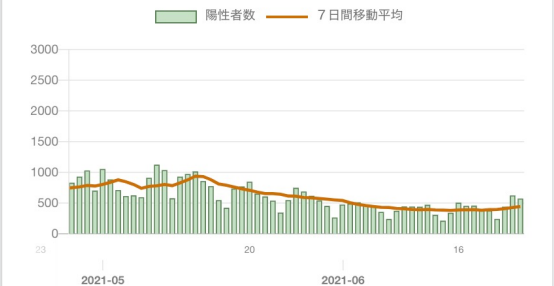
(1)新規陽性者数	439.6 人
潜在・市中感染	
(2)#7119 (東京消防庁救急相談センター) における発熱等相談件数	60.3 件
(3)新規陽性者における接触歴等不明者	
人数	272.7 人
増加比	110.8 %

医療提供体制

検査体制	
(4)PCR・抗原検査	
陽性率	4.6 %
検査人数	6,796.3 人
受入体制	
(5)救急医療の東京ルール適用件数	39.0 件
(6)入院患者数	1,360 人

モニタリング項目(1) 新規陽性者数

439.6 人

2021年6月24日 の数値 (7日間移動平均)
(前日比: +16.9 人)

- (注)
- 保健所から発生届が提出された日を基準とする
 - 医療機関等が行った検査も含む
 - チャーター機帰国者、クルーズ船乗客等は含まれていない
 - 集団感染発生や曜日による数値のばらつきにより、日々の結果が変動するため、こうしたばらつきを平準化し全体の傾向を見る趣旨から、過去7日間の移動平均値を折れ線グラフで示す (例えば、2020年5月7日の移動平均値は、2020年5月1日から5月7日までの実績値を平均したものの)

> テーブルを表示

収集された生データは表形式（行列）

No	全国地方公共団体コード	都道府県名	市区町村名	公表_年月日	曜日	発症_年月日	患者_居住地	患者_年代	患者_性別	患者_属性	患者_状態	患者_症状	患者_渡航歴の有無フラグ	備考	退院済フラグ
1	130001	東京都		2020/1/24	金		湖北省武漢市	40代	男性						1
2	130001	東京都		2020/1/25	土		湖北省武漢市	30代	女性						1
3	130001	東京都		2020/1/30	木		湖南省長沙市	30代	女性						1
4	130001	東京都		2020/2/13	木		都内	70代	男性						
5	130001	東京都		2020/2/14	金		都内	50代	女性						1
6	130001	東京都		2020/2/14	金		都内	70代	男性						1
7	130001	東京都		2020/2/15	土		都内	80代	男性						1
8	130001	東京都		2020/2/15	土		都内	50代	女性						1
9	130001	東京都		2020/2/15	土		都内	50代	男性						1
10	130001	東京都		2020/2/15	土		都内	70代	男性						1
11	130001	東京都		2020/2/15	土		都内	70代	男性						1
12	130001	東京都		2020/2/15	土		都内	40代	男性						1
13	130001	東京都		2020/2/15	土		都内	60代	女性						1
14	130001	東京都		2020/2/15	土		都内	40代	男性						1
15	130001	東京都		2020/2/16	日		都内	60代	男性						
16	130001	東京都		2020/2/16	日		都内	30代	男性						1
17	130001	東京都		2020/2/16	日		都内	60代	男性						1
18	130001	東京都		2020/2/16	日		都外	60代	男性						1
19	130001	東京都		2020/2/16	日		都内	30代	男性						1
20	130001	東京都		2020/2/18	火		都内	80代	男性						1
21	130001	東京都		2020/2/18	火		都内	20代	男性						1
22	130001	東京都		2020/2/18	火		都内	50代	男性						
23	130001	東京都		2020/2/19	水		都内	70代	男性						1
24	130001	東京都													1
25	130001	東京都													1

この表を見て，様々な傾向を見つけたり
解析する方針を決めるのは難しい！

生データ → 情報

No	全国地方公共団体コード	都道府県名	市区町村名	公表_年月日	曜日	発症_年月日	患者_居住地	患者_年代	患者_性別	患者_属性	患者_状態	患者_症状	患者_渡航歴の有無フラグ	備考	退院済フラグ
1	130001	東京都		2020/1/24	金		湖北省武漢市	40代	男性						1
2	130001	東京都		2020/1/25	土		湖北省武漢市	30代	女性						1
3	130001	東京都		2020/1/30	木		湖南省長沙市	30代	女性						
4	130001	東京都		2020/2/13	木		都内	70代	男性						
5	130001	東京都		2020/2/14	金		都内	50代	女性						
6	130001	東京都		2020/2/14	金		都内	70代	男性						
7	130001	東京都		2020/2/15	土		都内	80代	男性						
8	130001	東京都		2020/2/15	土		都内	50代	女性						
9	130001	東京都		2020/2/15	土		都内	50代	男性						
10	130001	東京都		2020/2/15	土		都内	70代	男性						1

様々な角度で情報を抽出、可視化(表・グラフなど)することでデータの直感的な理解を助ける

検査陽性者の状況

(注) チャーター機帰国者、クルーズ船乗客等は含まれていない
(注) 「重症」は、集中治療室 (ICU) 等での管理又は人工呼吸器管理が必要な患者数を計上
(注) 退院者数の把握には一定の期間を要しており、確認次第数値を更新している

陽性者数 (累計)	11,214 人
入院	1,165 人
軽症・中等症	1,147 人
重症	18 人
宿泊療養	179 人
自宅療養	452 人
入院・療養等調整中	966 人
死亡	328 人
退院等 (療養期間経過を含む)	8,124 人

新規患者に関する報告件数の推移

239 人

7/26 日別値 (前日比: -56 人)

(注) 保健所から発生届が提出された日を基準とする
(注) 医療機関等が行った検査も含む
(注) チャーター機帰国者、クルーズ船乗客等は含まれていない

日別 累計


[テーブルを表示](#)
[オープンデータを手取り](#)

モニタリング項目

感染状況

項目	数値
(1)新規陽性者数	257.6 人
潜在・市中感染	
(2)#7119 (東京消防庁救急相談センター)における発熱等相談件数	67.0 件
(3)新規陽性者における接触歴等不明者	人数 148.9 人 増加比 135.3 %

医療提供体制

項目	数値
検査体制	
(4)PCR・抗原検査	陽性率 6.3 % 検査人数 2,206.7 人
受入体制	
(5)救急医療の東京ルール適用件数	31.0 件
(6)入院患者数	1,165 人
	確保病床 1000床 → 2800床
(7)重症患者数	18 人
	確保病床 100床

[最新のモニタリング項目の分析・総括コメントについて](#)
知恵
Wisdom知識
Knowledge情報
Informationデータ
Data

DIKWピラミッド

<https://stopcovid19.metro.tokyo.lg.jp/>

DIKWピラミッド

データ・情報・知識・知恵の関係を分かり易く体系化した図

□ Data (データ)

- それ自体では意味を持たない数字、記号など
 - 生データ、ローデータ

データから情報を得ることが
データサイエンスのはじめの目的

□ Information (情報)

- データを何らかの基準で整理して意味づけしたもの

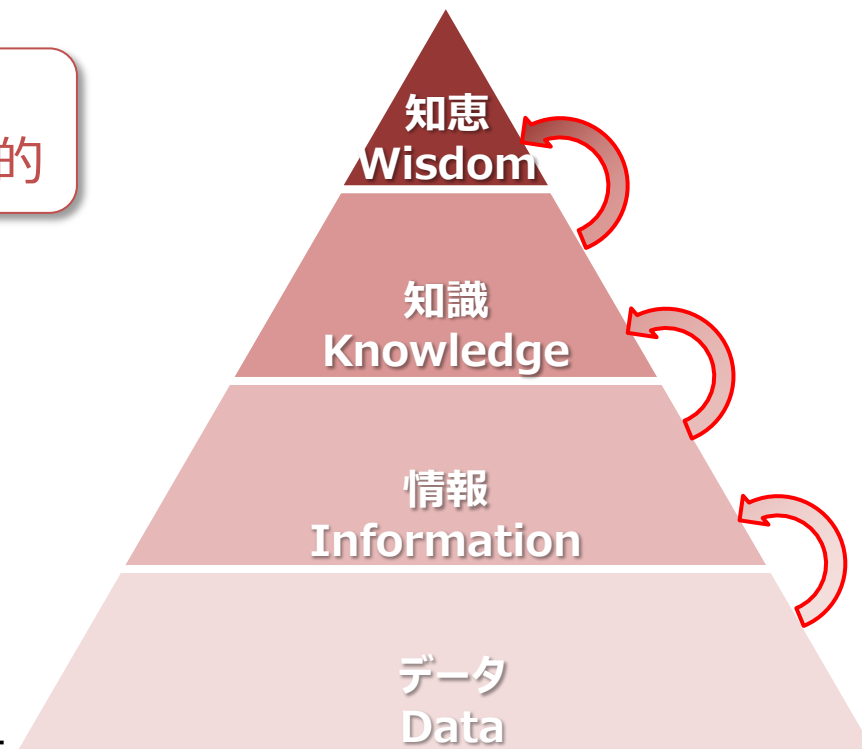
□ Knowledge (知識)

- 情報をまとめて体系化・構造化することで、規則性、傾向、知見を導き出したもの。

□ Wisdom (知恵)

- 知識を正しく認識して判断し、価値観やモラルに昇華させたもの。

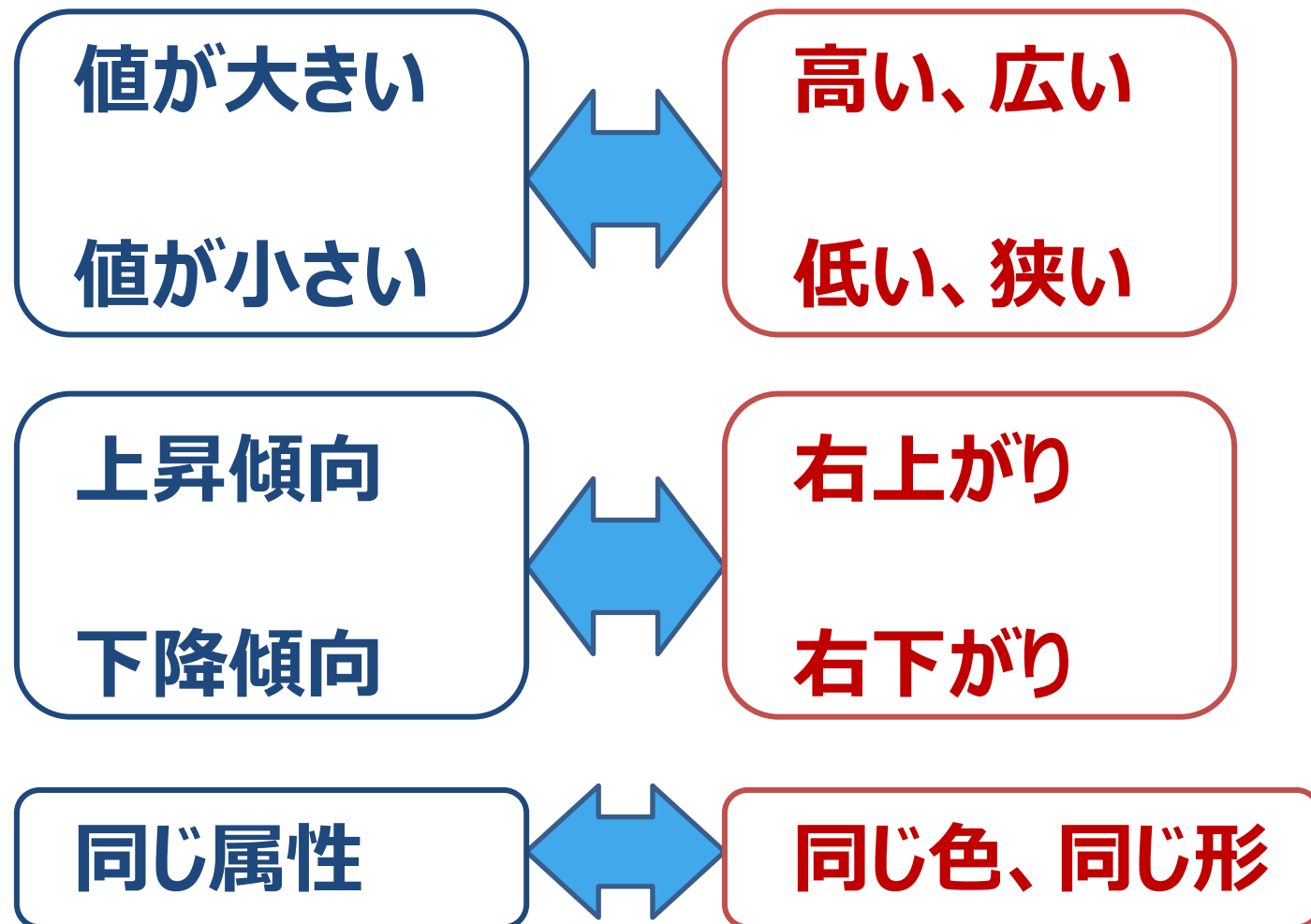
課題の解決につながるアイデア・指針



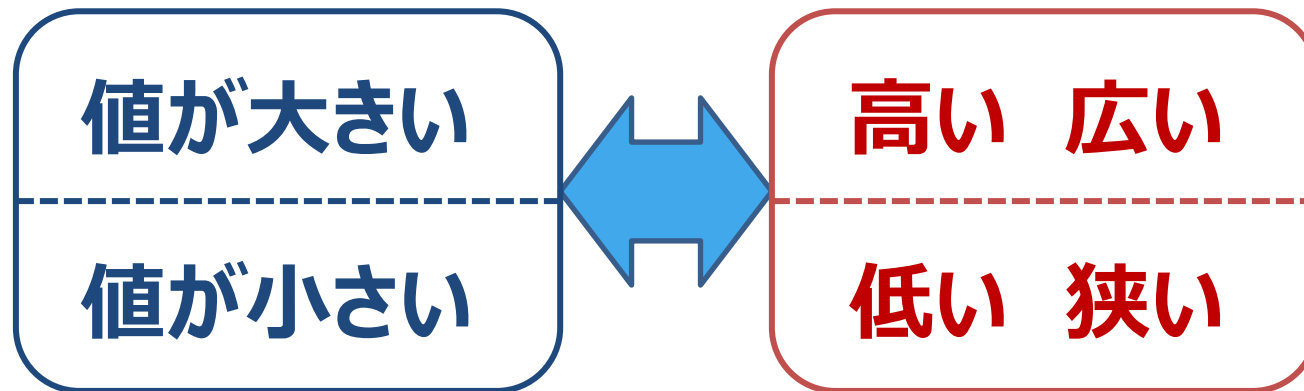
DIKWピラミッド

なぜ可視化すると理解しやすいか

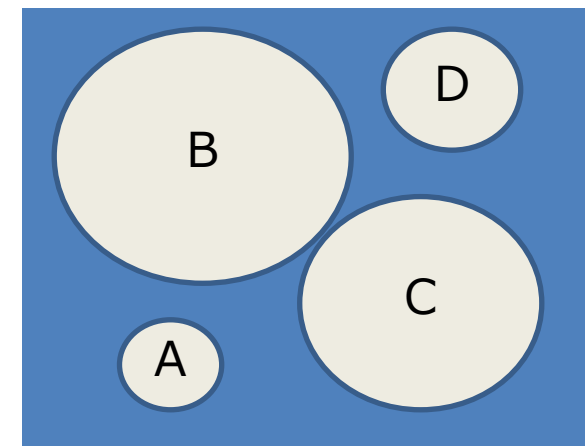
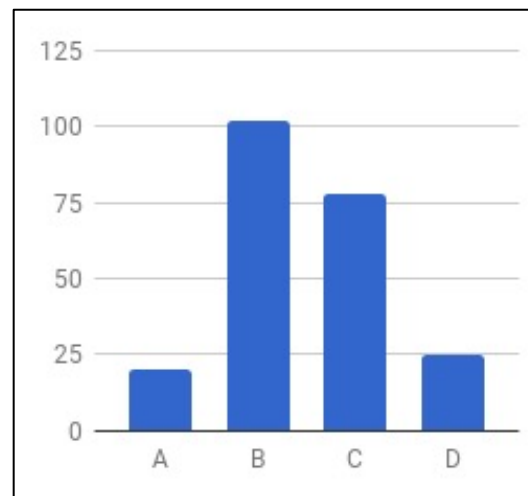
経験・認知（直感）による理解を最大限に利用することができる



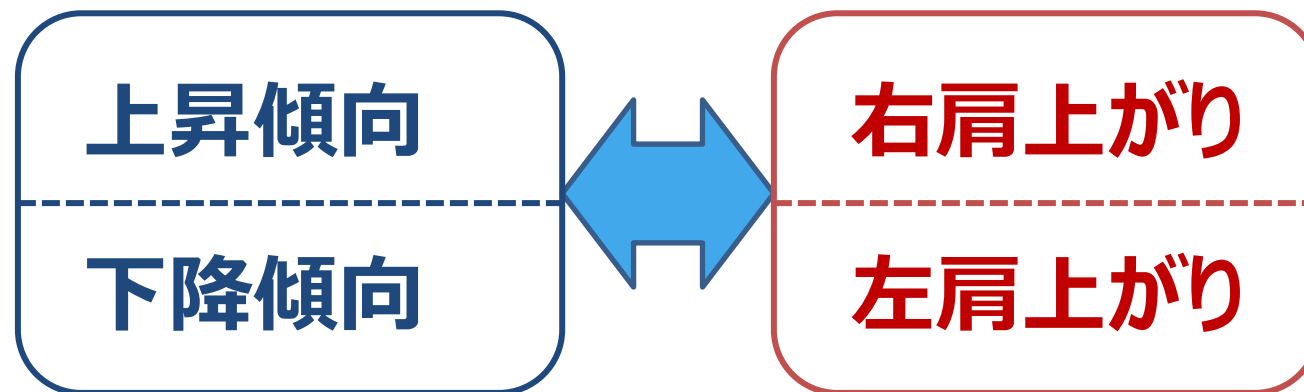
可視化と認知



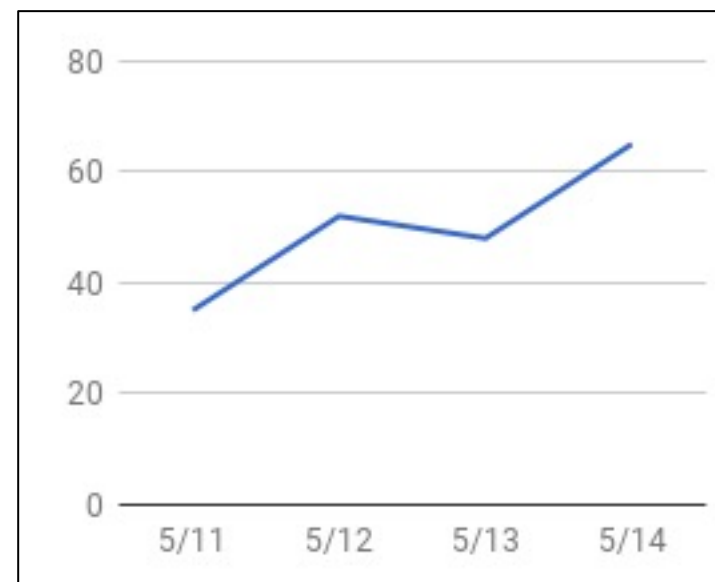
物	数値
A	20
B	102
C	78
D	25



可視化と認知

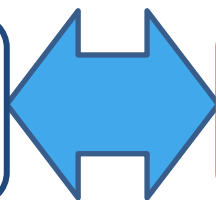


時間	数値
5/11	35
5/12	52
5/13	48
5/14	65



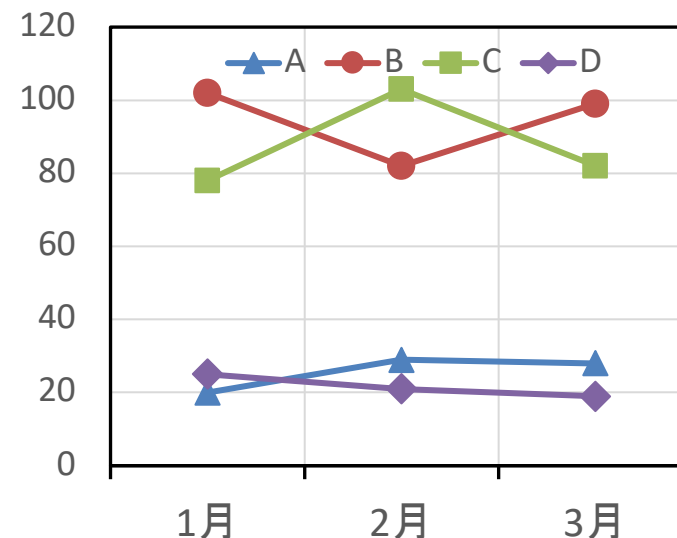
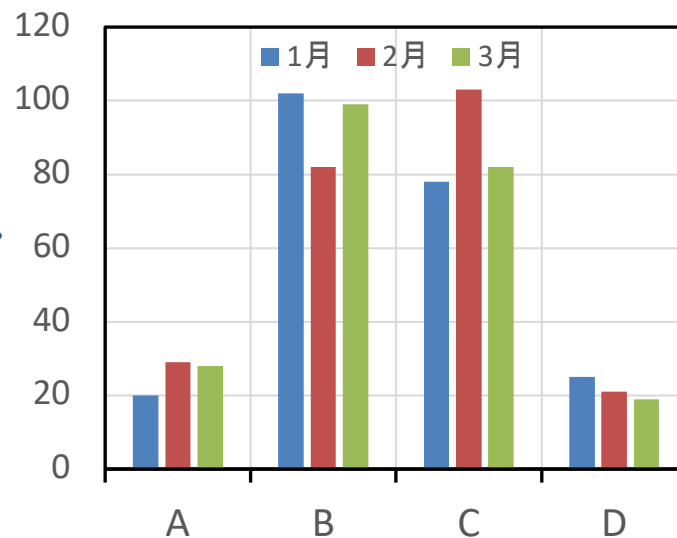
可視化と認知

同じ属性



同じ色、同じ形

	1月	2月	3月
A	20	29	28
B	102	82	99
C	78	103	82
D	25	21	19



データ可視化が役に立った事例（１）

□1854年ロンドンのソーホー地区、ブロード通りでコレラが大流行
□当時、「瘴気」と呼ばれる空気中の粒子が原因とされていたが、
コレラが**なぜおこるのか、よく分かっていなかった。**

※**1884年**にコッホによりコレラ菌発見
ちなみにコレラは菌により汚染された
水や食物により経口感染する。



データ可視化が役に立った事例（１）

CITY
UNIV.

ジョン・スノー（1813-1858）のコレラ地図

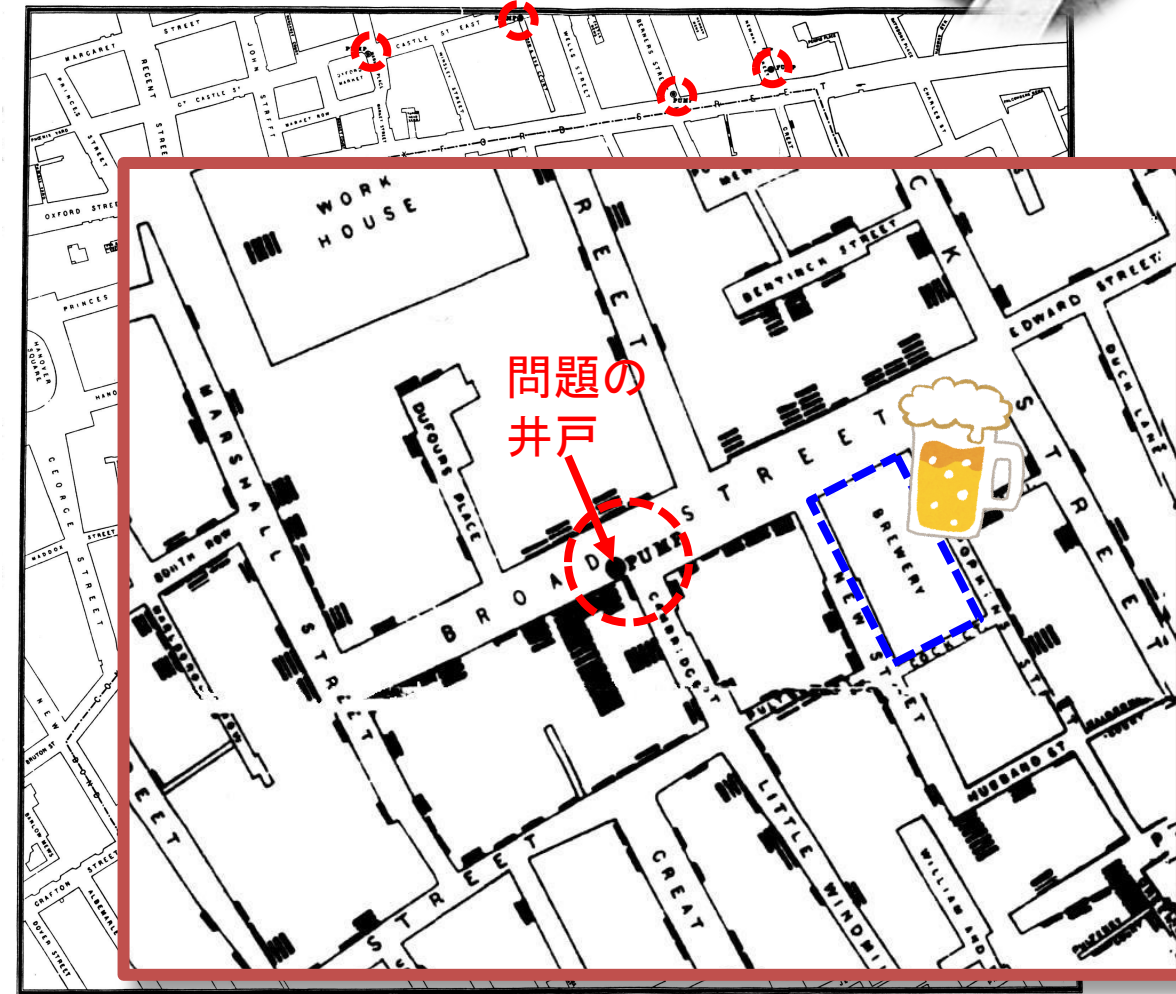


ロジョン・スノーは、コレラ患者の発生数を地図で可視化した

ロひとつの井戸の周囲に患者が集中し、その井戸を利用する家に患者が発生することが判明



井戸の利用を中止させたところ
流行はおさまった！



データ可視化が役に立った事例（2）

CITY
UNIV.

フローレンス・ナイチンゲール (1820-1910)の「鶏のトサカ」ダイアグラム

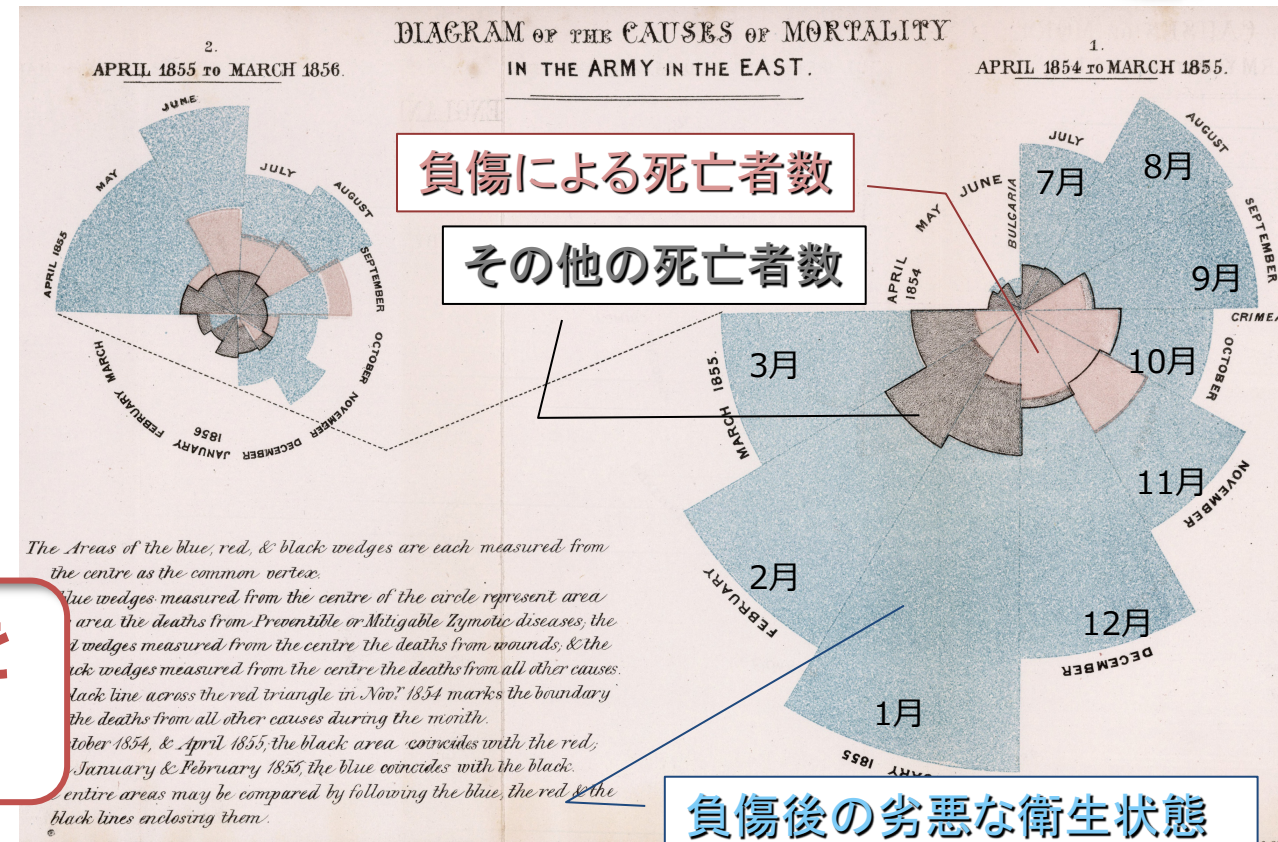


「病人を救うのは、宗教者の愛よりも衛生環境である」

- クリミア戦争中では、兵士たちが劣悪な衛生状態と栄養失調で死亡
- 病院での死亡者数を綿密に記録



データから分かりやすいダイアグラムを作成し、衛生状態の改善などを要求



<http://chaton2.blogspot.com/2011/10/blog-post.html>

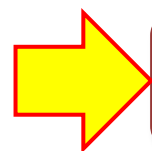
データ可視化が必要なとき

1. 探索的データ解析＜自分で作って自分で見る＞

- 仮説生成やモデル選択の前段階での「データそのものを理解する」場面で必要となる.
- 視覚的にデータのいろいろな面を表示することで, 「データに語らせる」

2. プレゼンテーション＜他の人に見せて納得させる＞

- データ解析の結果を用いて, なんらかの主張を第三者に納得させるための可視化
 - Ex. 一般市民@広告, 上司@会社など



データ可視化により, 直観的かつ効率的に主張を伝える

データ可視化の手順

1. 伝えたい主張に使うデータ解析結果を選び抜く
2. 解析結果の各要素に最も合う視覚表現を選ぶ
3. 解析結果に最も合う可視化手法を選ぶ



不適切な可視化手法を選ぶと誤解を招くので注意

- 巷にあふれているグラフなどでは誤解を招く表現や間違い多数
- データに嘘をつかせようとしている可視化を見抜く力も大切

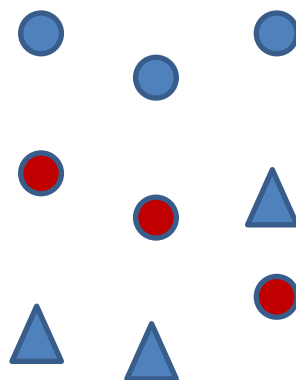
視覚変数

CITY
UNIV.

対象や特性を区別して表すための要素

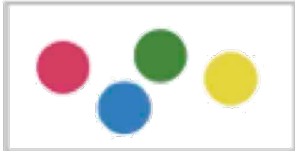
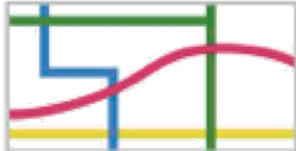
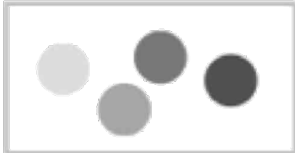
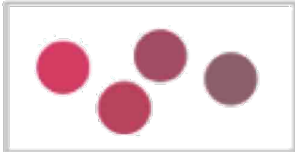
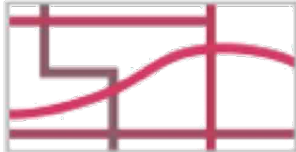
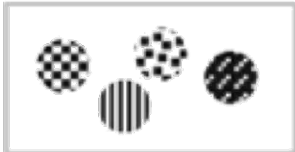
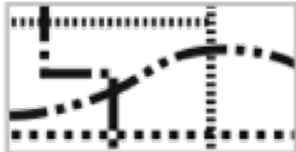
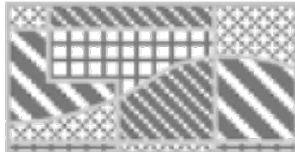
□網膜の変数

- 形
- 大きさ
- 色相
- 明度
- 彩度
- テクスチャー



青丸と赤丸は色で区別されるが、
丸という共通要素を持ち、
青丸と青三角は形で区別されるが
青という共通要素を持つ

点・線・エリアで様々な性質を表現 —視覚変数—

	<i>Points</i>	<i>Lines</i>	<i>Areas</i>	<i>Best to show</i>	
<i>Shape</i>		<i>possible, but too weird to show</i>	<i>cartogram</i>	<i>qualitative differences</i>	形：質の違い
<i>Size</i>			<i>cartogram</i>	<i>quantitative differences</i>	大きさ：量の違い
<i>Color Hue</i>				<i>qualitative differences</i>	色調：質の違い
<i>Color Value</i>				<i>quantitative differences</i>	明度：量の違い
<i>Color Intensity</i>				<i>qualitative differences</i>	色の強さ：質の違い
<i>Texture</i>				<i>qualitative & quantitative differences</i>	テクスチャ：質と量の違い

<http://understandinggraphics.com/visualizations/information-display-tips/>

データの可視化

データの可視化

- データの可視化とは
- 可視化の事例
- 可視化が必要な場面と方法

代表的な可視化手法

- 棒グラフ
- ヒストグラム
- 折れ線グラフ
- 散布図
- 円グラフ
- 箱髭グラフ
- ネットワーク

演習:

- データの整理・可視化演習



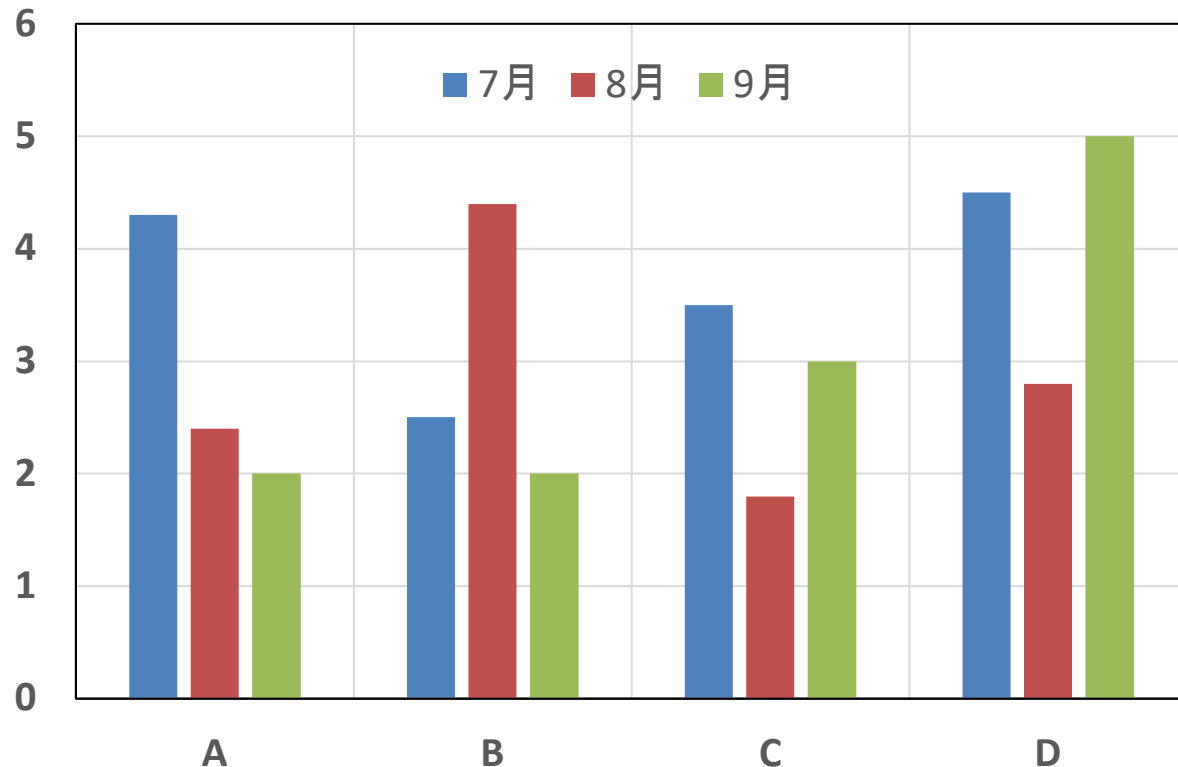
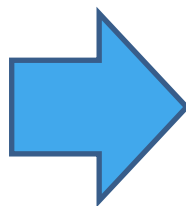
棒グラフ

形状：棒の長さにより大小を表す二次元グラフ。

棒の向きは縦と横がある。

目的：大小の比較

	7月	8月	9月
A	4.3	2.4	2
B	2.5	4.4	2
C	3.5	1.8	3
D	4.5	2.8	5

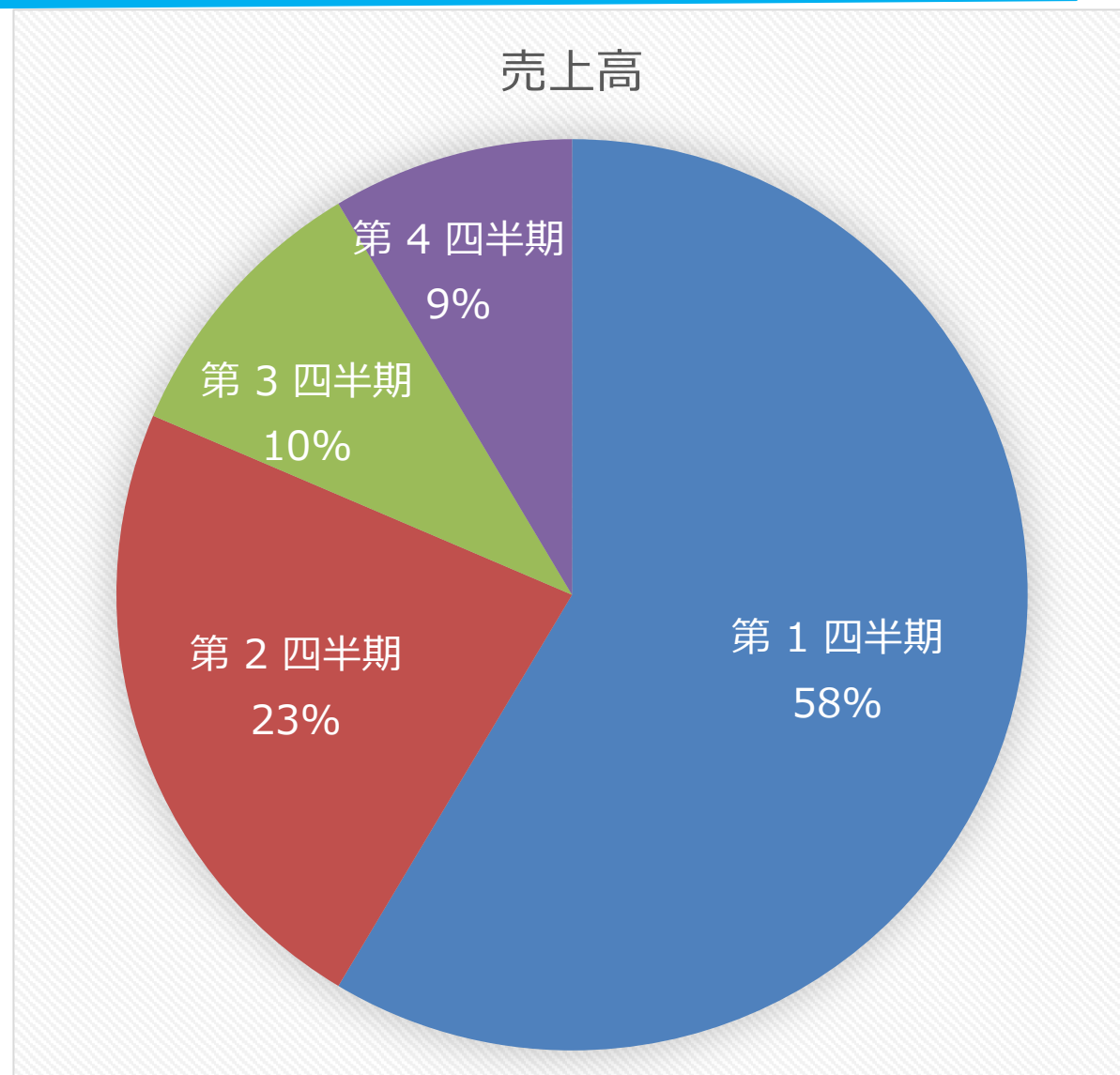
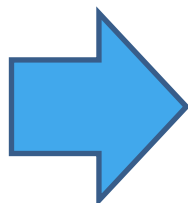


円グラフ

形状：丸い図形を扇形に分割したグラフ。

目的：全体に対する割合の視覚化

	売上高
第 1 四半期	8.2
第 2 四半期	3.2
第 3 四半期	1.4
第 4 四半期	1.2

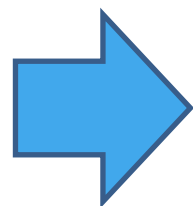


ヒストグラム

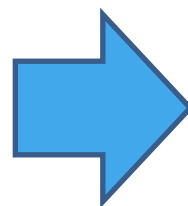
形状：縦軸に度数，横軸に階級をとる棒グラフ

目的：分布の視覚化

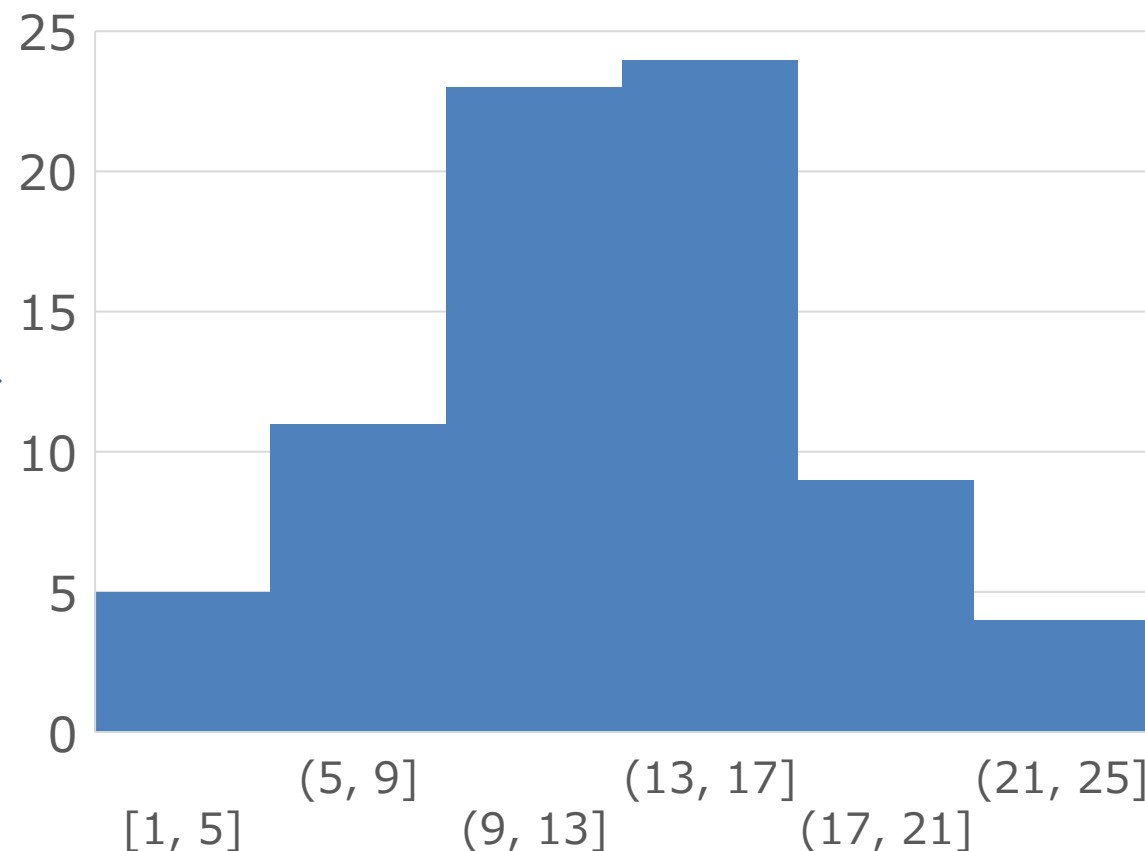
系列1	
	1
	3
	3
	3
	5
	6
	6
	6
	7
...	
	22
	24
	24



度数分布表



ヒストグラムの例



箱ひげ図（の説明のための分位数の説明）

分位数（パーセンタイル：percentile）

□データを小さい順に並べたとき、初めから数えて全体の $(100\alpha)\%$ 番目に位置する値を α パーセンタイルと言う。

- 例：25人の試験結果のデータについて
- 50パーセンタイルは最低点から13番目の人の点数（中央値）
- 0パーセンタイルは最低点から1番目の人の点数（最低点）
- 100パーセンタイルは最低点から25番目の人の点数（最高点）

□25%ごとに分けて「四分位数」で表すことがある。

- 第一四分位数：25パーセンタイル
- 第二四分位数：50パーセンタイル
- 第三四分位数：75パーセンタイル

箱ひげ図

□ 箱の部分

- 第一四分位数から第三四分位数を表し、第二四分位数（中央値）のところに線を引く

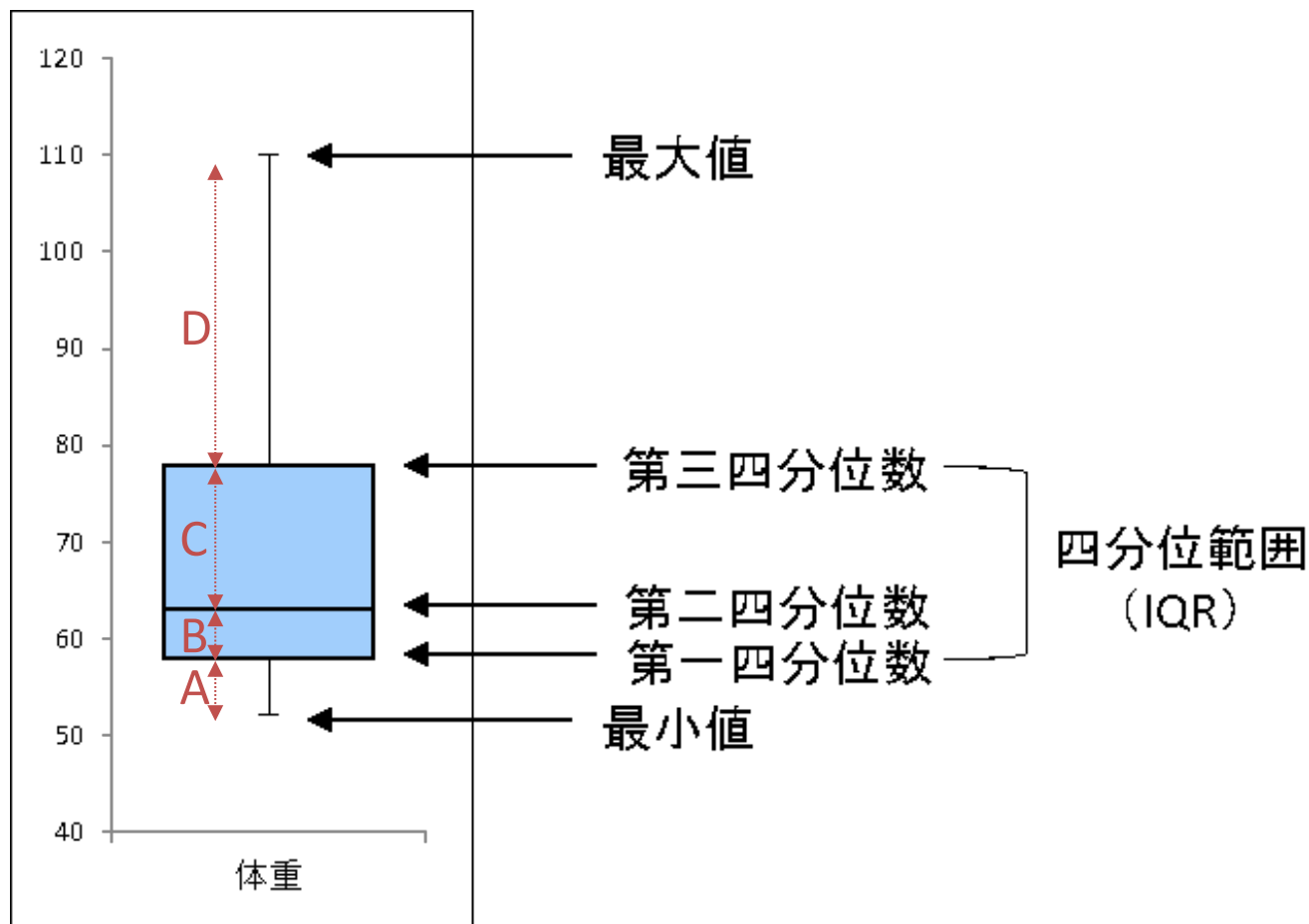
□ ひげの部分

- 最大値、最小値の値まで箱から線を引く

- A,B,C,Dの各区間には同数のデータ

→ データのばらつき具合が分かる

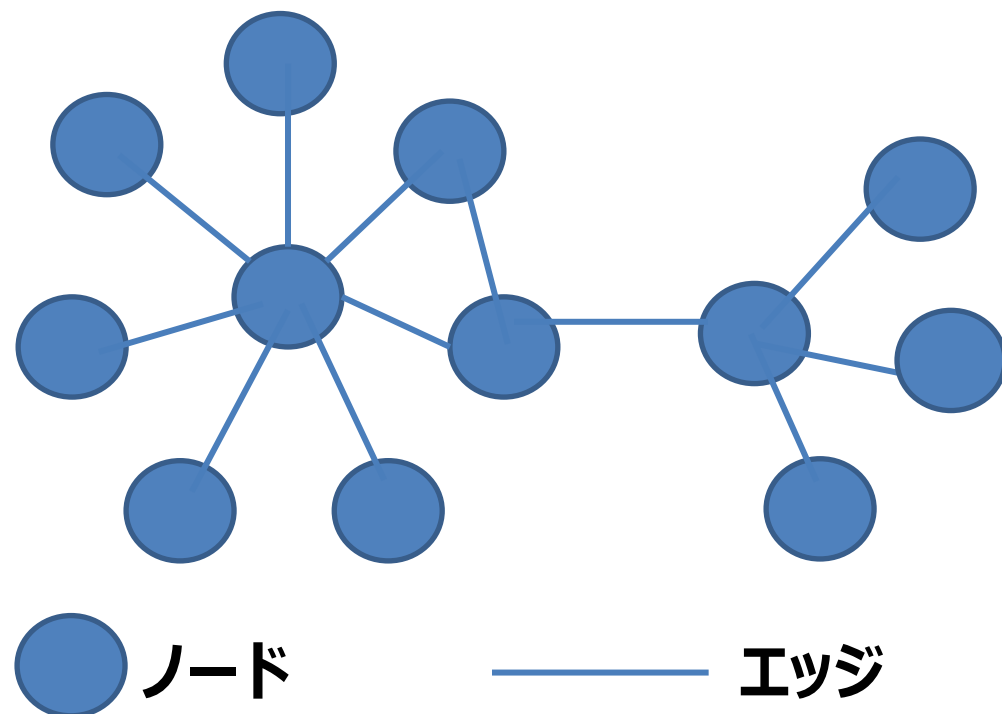
- 区間Bにデータが偏っている
- 区間BとCの箱の領域にデータの半分が存在



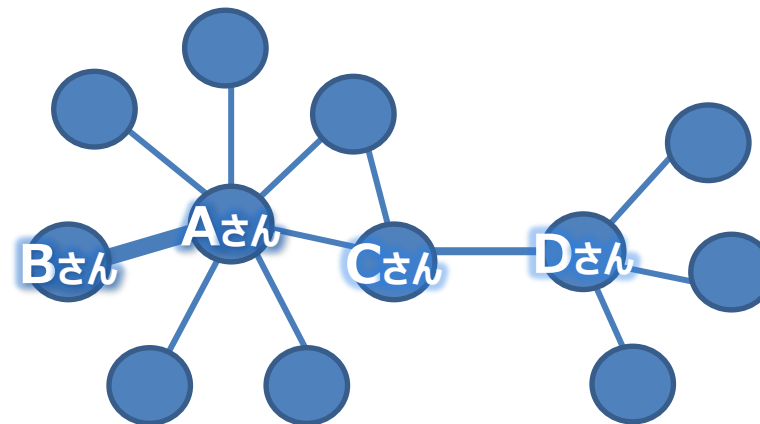
グラフ・ネットワーク

形状：頂点と辺からなるネットワーク図。

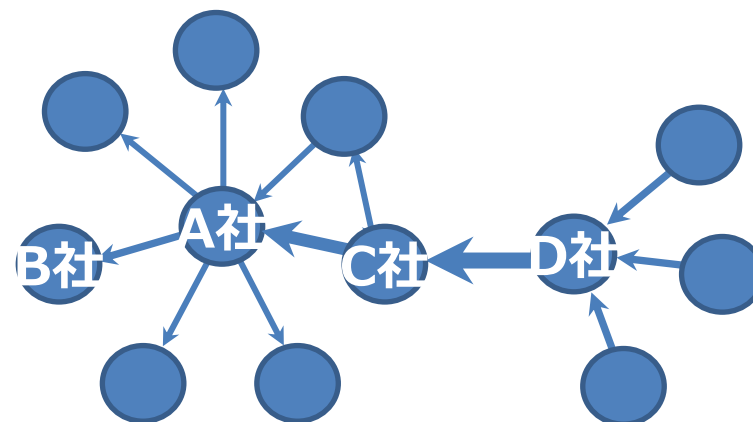
目的：ネットワーク構造の視覚化



SNSでのユーザーの関係



会社間の取引



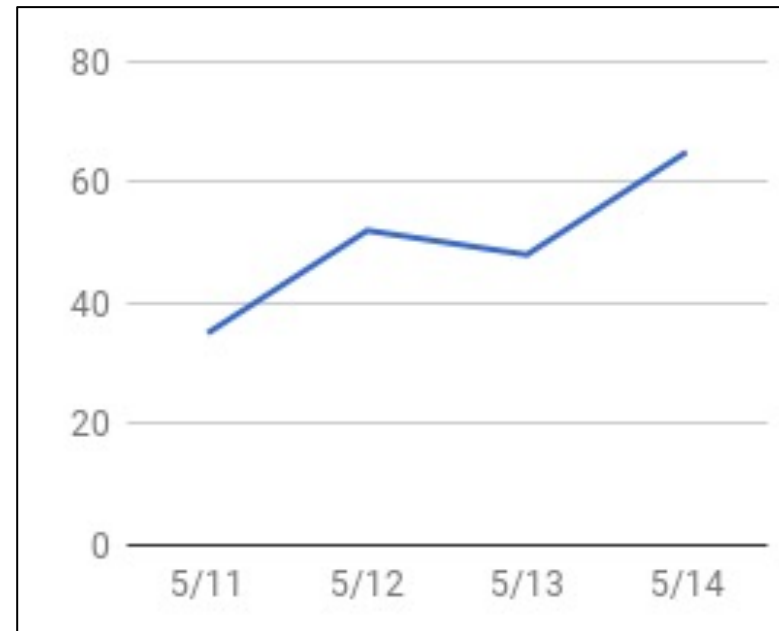
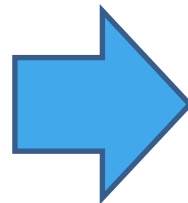
折れ線グラフ

形状：二つの項目の値を順に線でつないだ二次元グラフ。

横軸に時系列が来ることが多い

目的：二つの項目の関係，特に時系列の変化の視覚化

時間	数値
5/11	35
5/12	52
5/13	48
5/14	65

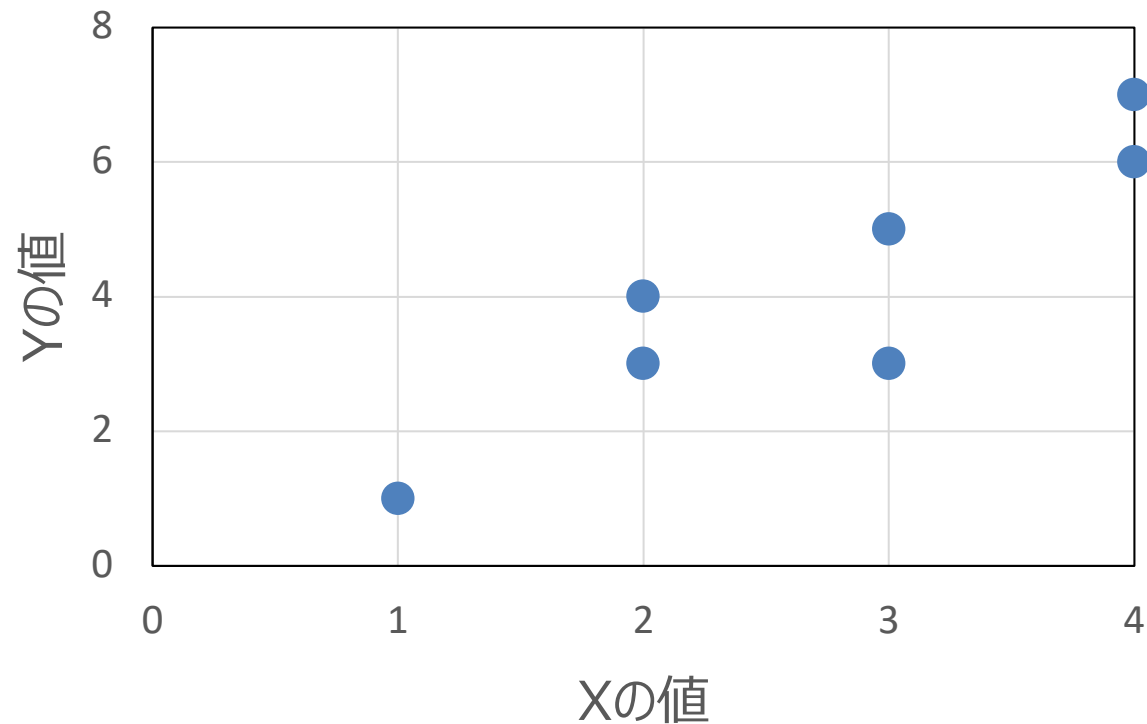
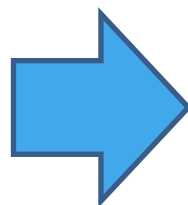


散布図

形状：二つの項目の値を二次元平面に置いたもの。

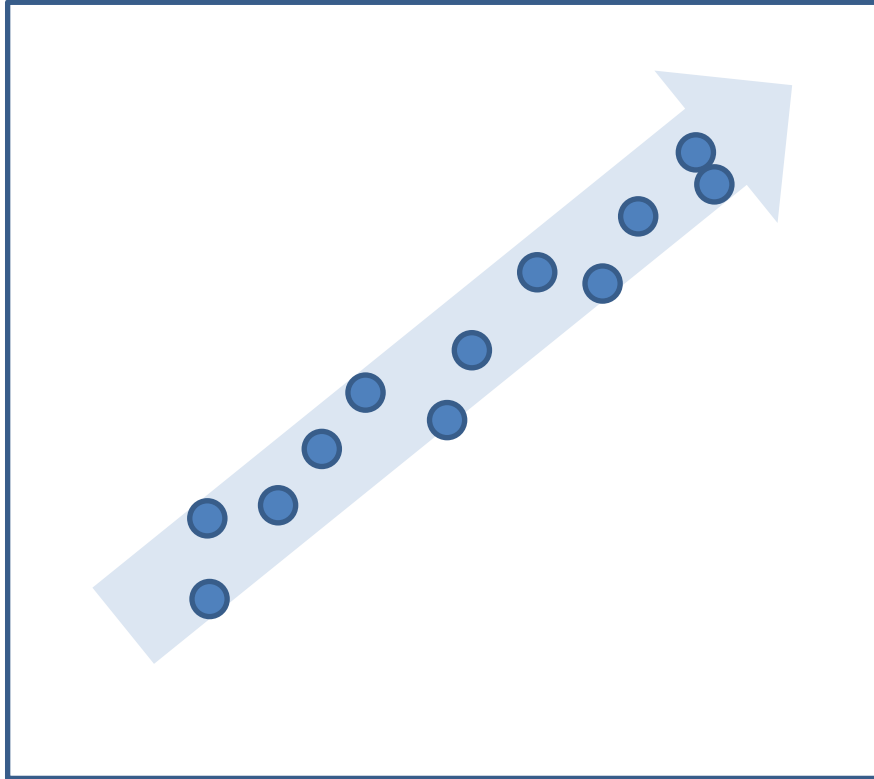
目的：二つの項目の分布関係および**相関関係**の視覚化

X の値	Y の値
1	1
2	3
2	4
3	3
3	5
4	6
4	7



相関関係：Xが増えるに従って、Yも増える（減る）！？

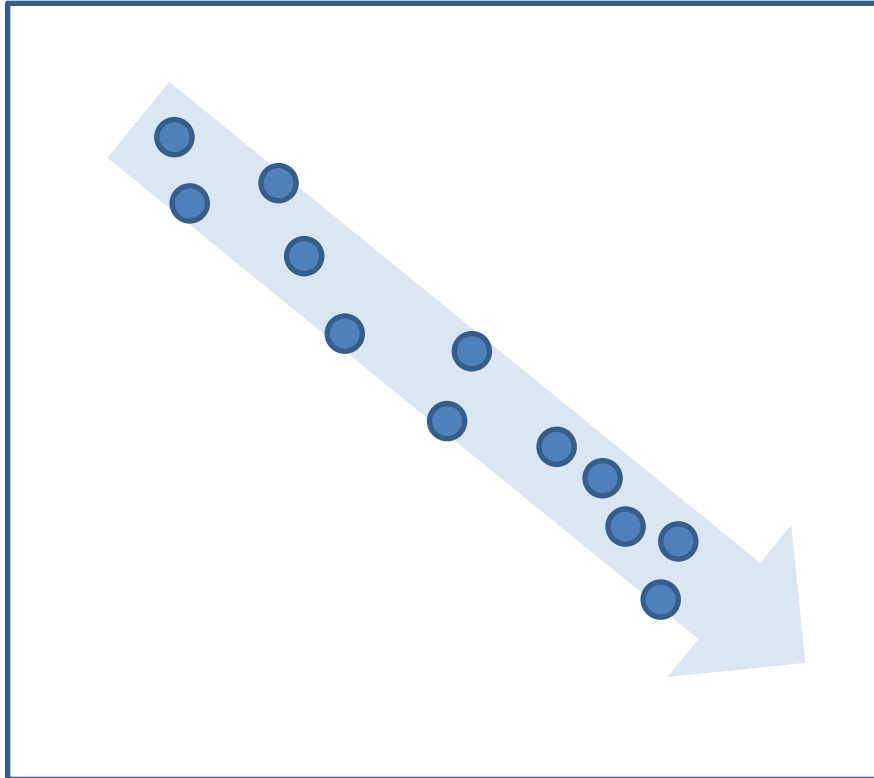
正の相関



散布図において、
左下から右上方向に
直線状にデータが並ぶとき、
「正の相関がある」
という。

直線に近いほど、
ばらつきが少なければ少ないほど、
「強い正の相関がある」という。

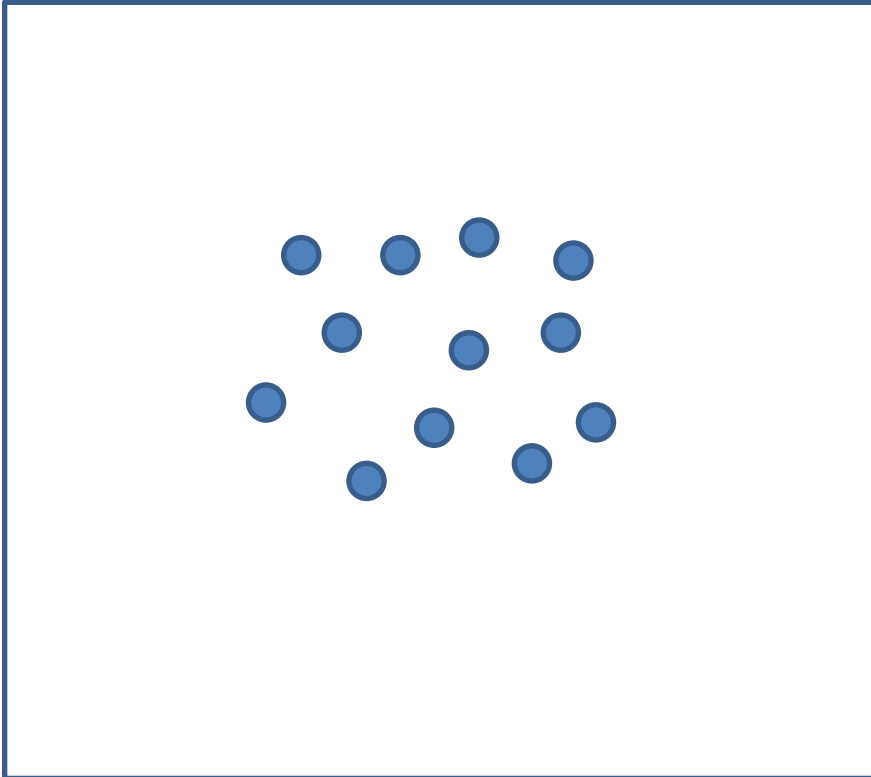
負の相関



散布図において、
左上から右下方向に
直線状にデータが並ぶとき、
「負の相関がある」
という。

直線に近いほど、
ばらつきが少なければ少ないほど、
「強い負の相関がある」という。

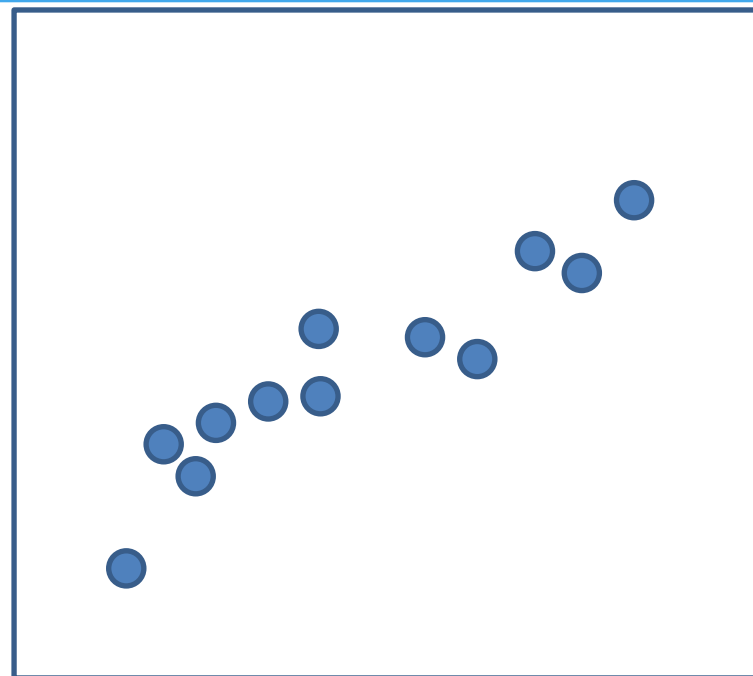
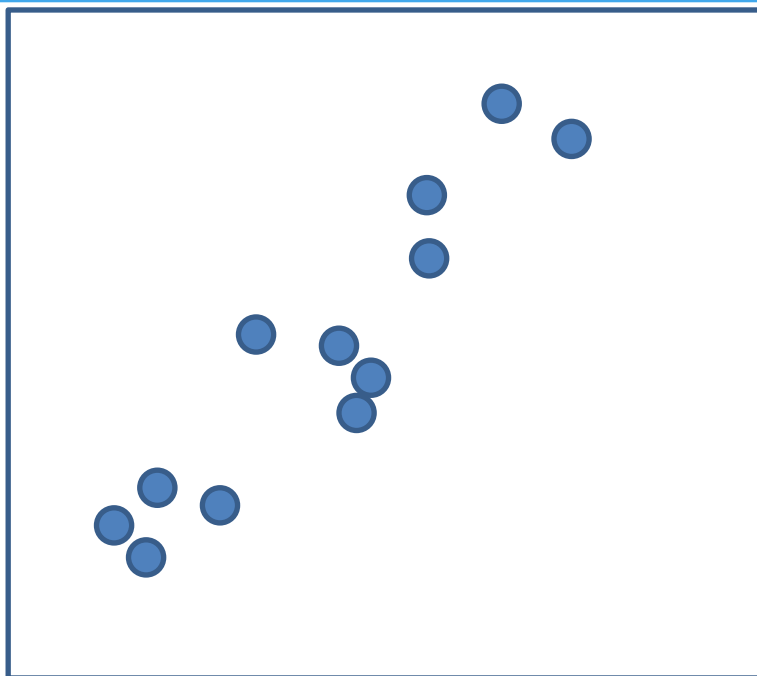
相関なし



散布図において、
直線状にデータがならないとき、
あるいは、
縦一直線か横一直線にならぶとき、
相関がないという。

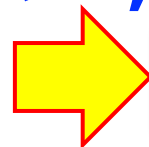
相関の比較

相関が強い／弱いと出てきたけれど、
その強さ／弱さは数値化できないの？



□どちらも正の相関がありそうだけど、どちらの正の相関が強い？

□数値化できれば比較できる。



相関係数

相関係数

2つの変数の間にある線形な関係の強弱を測る指標

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$\vec{x}'$ と $\vec{y}'$ の内積

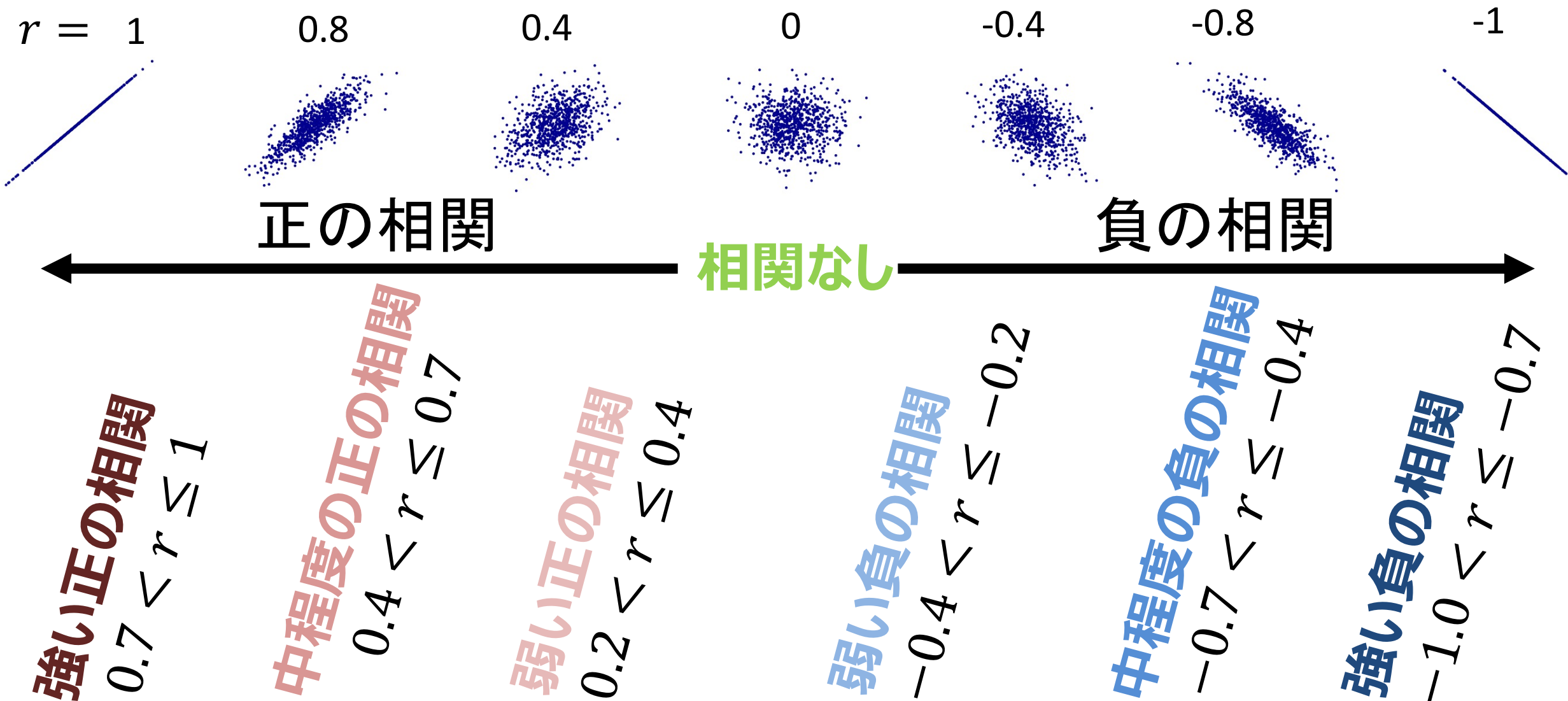
$$= \frac{\vec{x}' \cdot \vec{y}'}{|\vec{x}'| |\vec{y}'|}$$

$\vec{x}', \vec{y}'$: 平均を引いて中心を原点に揃えた n 次元のベクトル

$$= \cos \theta \quad (-1 \leq r \leq 1)$$

□ $\vec{x}', \vec{y}'$ がなす角 $\rightarrow \vec{x}', \vec{y}'$ が同じ方向を向いているかどうか

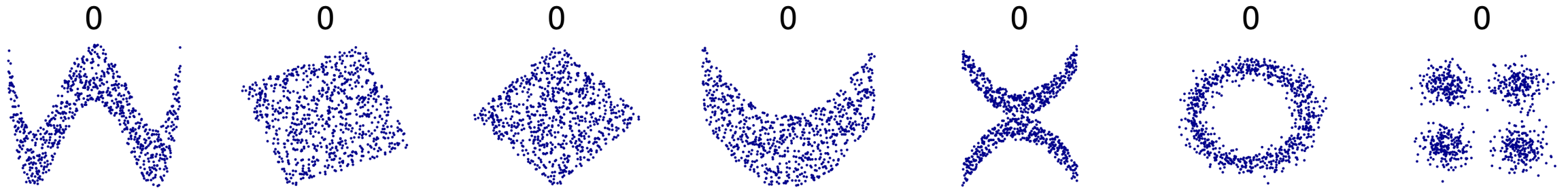
相関係数に対応する散布図と相関の強さのめやす



相関係数を使うときの注意点

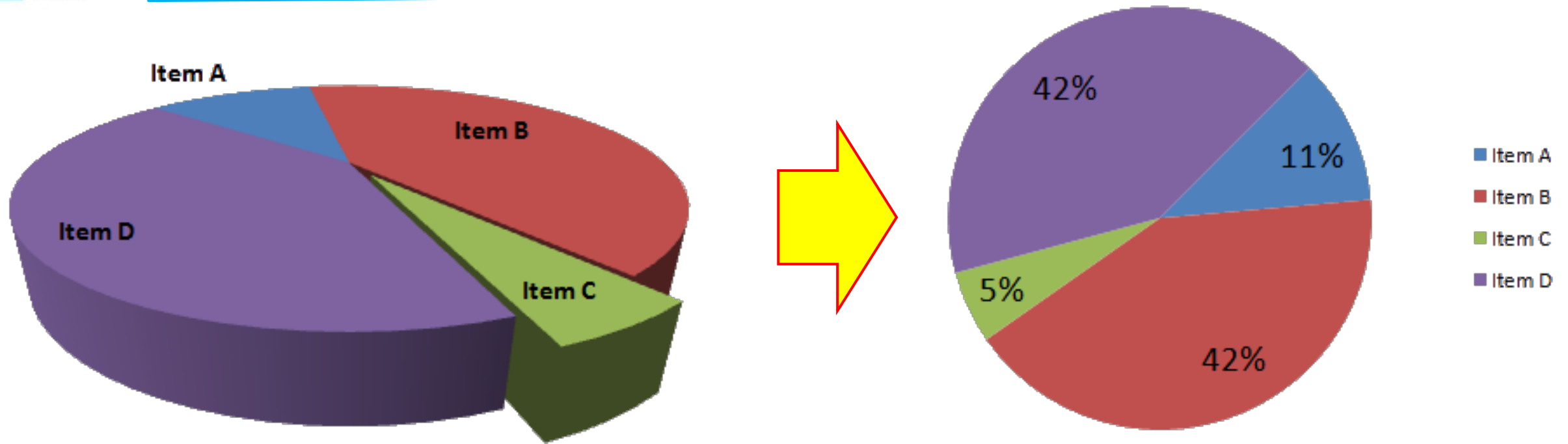
□相関係数は直線的な正の相関／負の相関の強さを調べるときには非常に有効だが、直線的でないデータ間の関係の強さを調べるためには、役に立たない。

●直線的でない関係の例



➔ 相関係数が0でも、データ間に関係がないとは限らない

誤解を招くグラフの例



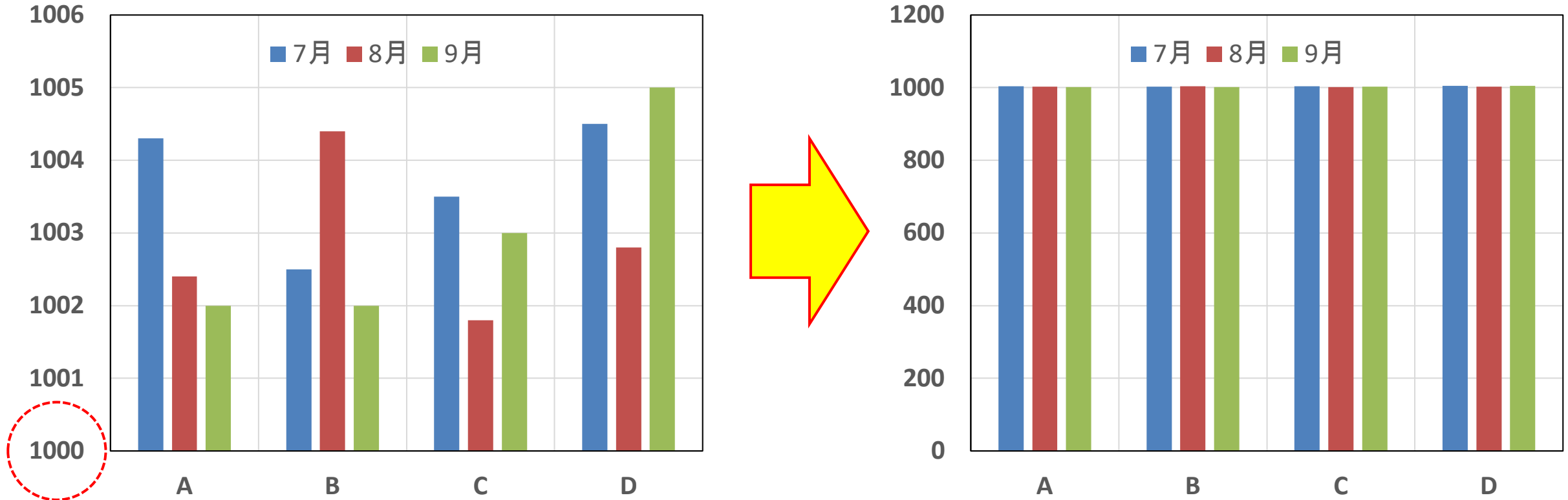
問題点

- 3次元の円グラフでは、Item C は Item A とほぼ同じ大きさに見えるが、実際には半分以下。
- Item DはItem Bよりも大きく見えるが実際には同じ。

→円グラフは面積の大小で量を比較するためのグラフなので 3 Dにはしない

<https://ja.wikipedia.org/wiki/誤解を与える統計グラフ>

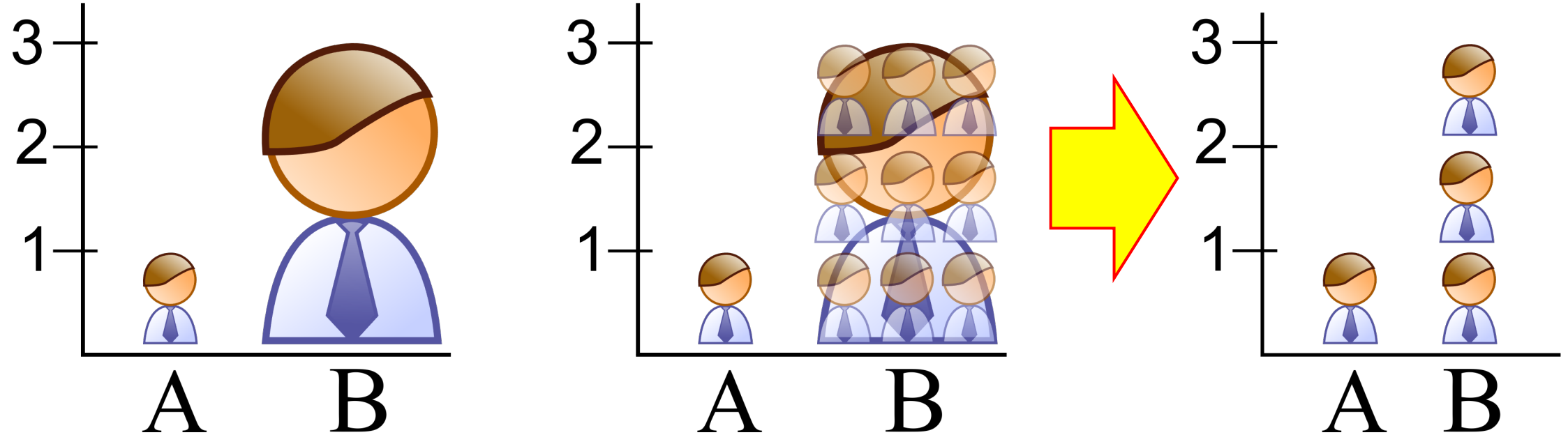
誤解を招くグラフの例



問題点

- どちらのグラフも同じデータを表している。左のグラフではデータに差があるように見られるが、縦軸のスケールを全体に広げると右側のように差はほとんど見られない。

誤解を招くグラフの例



問題点

- 左の絵文字の棒グラフでは、Bの画像はAの9倍になってしまっている。
- BがAの3倍であることを示すには絵文字のサイズは変えず、右のようなグラフに表す。

<https://ja.wikipedia.org/wiki/誤解を与える統計グラフ>