1-5.　データの可視化

東京大学 数理・情報教育研究センター

2021年4月12日

概要

- 本節では，データを可視化する（グラフにして観察する）方法を学びます.
- まず，棒グラフ，折れ線グラフ，散布図など数値データを可視化する方法を学びます.
- さらに，人やモノの間の関係性（つながり）や地図上のデータなどビッグデータを可視化する方法を学びます.

本教材の目次

1．データの可視化の目的

データを持っているだけでは意味がありません.
データが持つ情報を読み取るためにグラフを作成します.

グラフ作成の目的は以下の通りです.

A. 比較

・大きいか小さいかを比較したい

B. 変化

・増えているか、減っているかを知りたい

C. 構成比

・全体の中での割合 (構成比) を知りたい

D. 分布

・データの散らばりの度合いを知りたい

E. 相関

・データ間に関係がありそうか知りたい

2．グラフによる可視化：A．比較（1）

データを表やグラフにまとめると大小関係がわかります．
棒グラフ（右）を作成すると比較しやすくなります．

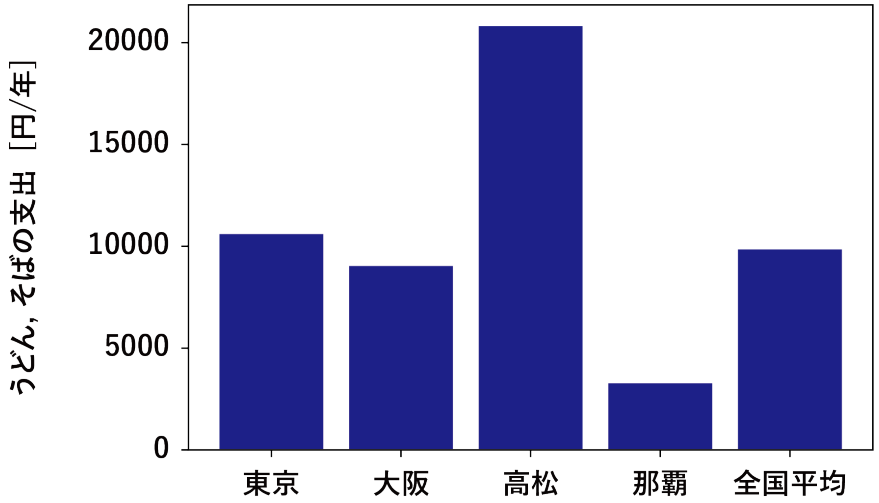
表：1世帯のうどん，そばの平均支出金額
（購入+外食）

都市名	年間支出金額
東京都（区部）	10,594
大阪市	9,026
高松市	20,807
那覇市	3,271
全国平均	9,838

総務省統計局「2019(令和元)年家計調査」
（調査年月：令和元年）

可視化

棒グラフ：1世帯のうどん，そばの平均支出金額
（購入+外食）



東京都、大阪市は全国平均に近く，
高松市は全国平均の2倍以上うどん，そばを
消費していることがわかります．

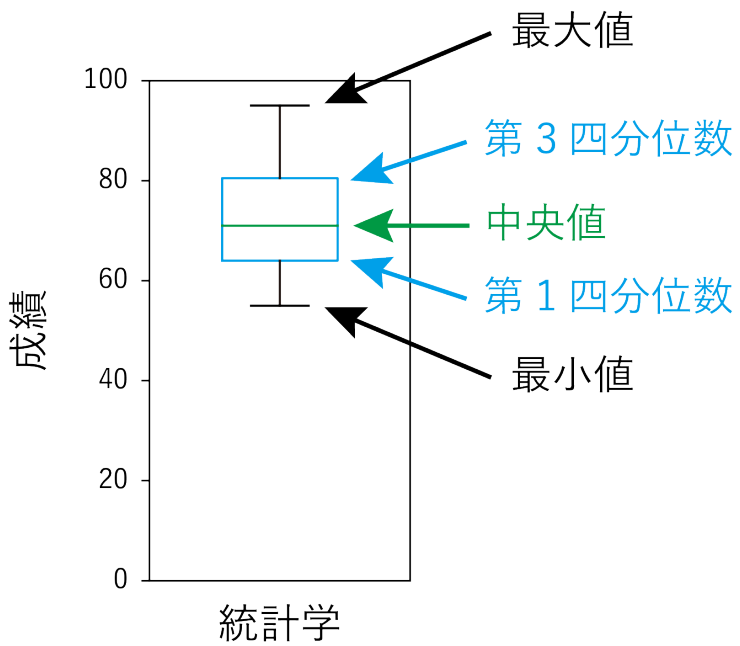
参考情報：うどんに関する統計情報

2．グラフによる可視化：A．比較（2）

箱ひげ図を作成すると，集団間（たとえば，大学生 vs 高校生，日本 vs アメリカなど）の大小関係を比較できます．

箱ひげ図の例

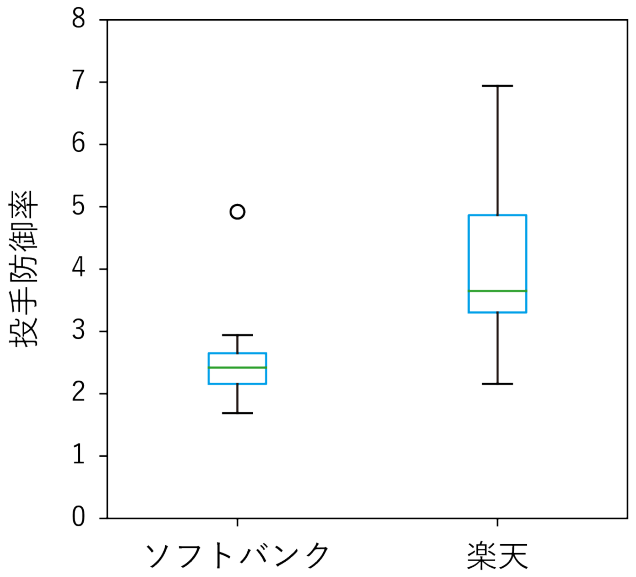
例1：研究室の学生の成績（仮想データ）



第1（3）四分位数から大きく離れたデータ点は外れ値 ○ と示されます．
（例2：ソフトバンク参照）

例2：ソフトバンクと楽天の投手陣の成績

出典：プロ野球データフリーク（ 2020年） リンク



ソフトバンク投手陣は成績のバラつきが少なく，半数以上の投手が良い防御率（2.5以下）です．一方，楽天投手陣は成績のバラつきが大きく，半数以上の投手が悪い防御率（3.5以上）です．
ただし，防御率は投手が1試合で取られる平均得点です．

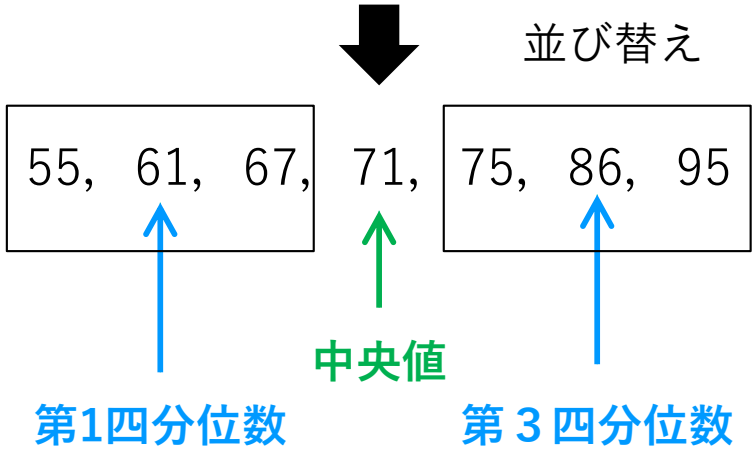
2．グラフによる可視化：A．比較（2）

箱ひげ図の作り方

1. データを小さい順に並び替えます．
2. 最大値，最小値，中央値，第1，3四分位数を計算します．

例：研究室の学生（7名）の統計学の成績（仮想データ）

71点，95点，75点，55点，67点，61点，86点



中央値：データを小さい順に並び替えた時の中央にある値．ただし，
データが偶数個の場合、中央順位の2つのデータの平均をとります．

第1（3）四分位数：中央値より小さい（大きい）データ群の中央にある値．

2．グラフによる可視化：A．比較（3）

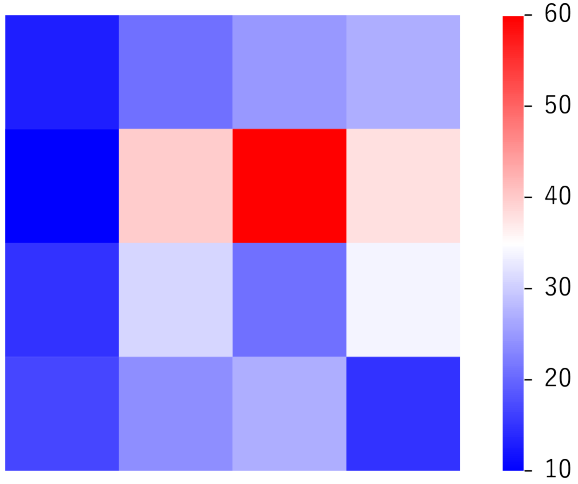
ヒートマップを作成すると，地図上のデータや行列など，
2次元に並んだデータの大小関係を比較できます．

人口密度（4×4区画）のデータ
（仮想データ）

13	21	25	27
10	40	60	38
15	31	21	34
17	24	27	15

可視化

ヒートマップ



右図（ヒートマップ）は，左のデータ（行列）を，大きい数ほど赤に近く，小さい数ほど青に近くするというルールで可視化したものです．

ヒートマップ を見ると，データ中の数値が大きい部分がわかりやすくなります．

2．グラフによる可視化：B．変化（1）

折れ線グラフ（右図）を作成すると，対象となる量の変化（増えている，減っている）を観察できます．

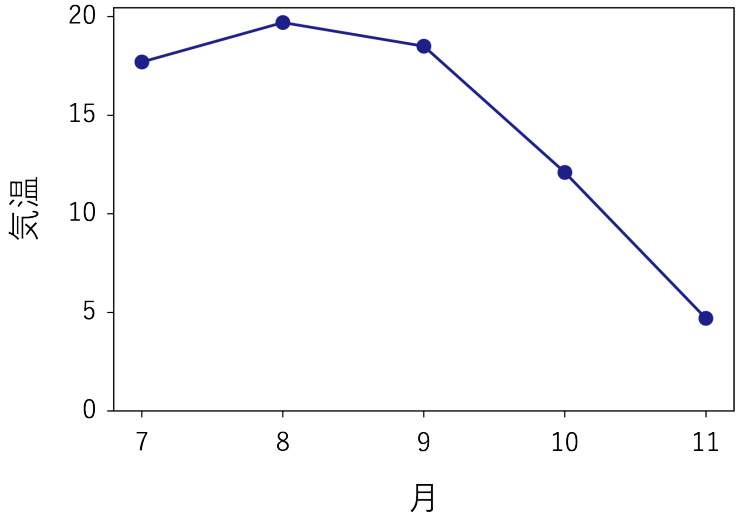
表：稚内市（北海道）の月間平均気温（2020年）

月	平均気温 (°C)
7月	17.7
8月	19.7
9月	18.5
10月	12.1
11月	4.7

出典：気象庁「過去の気象データ・ダウンロード」　リンク

可視化

折れ線グラフ



稚内市では，8月にそれほど気温が上がらず、9月以降、気温が急に下がることがわかります．

2．グラフによる可視化：B．変化（2）

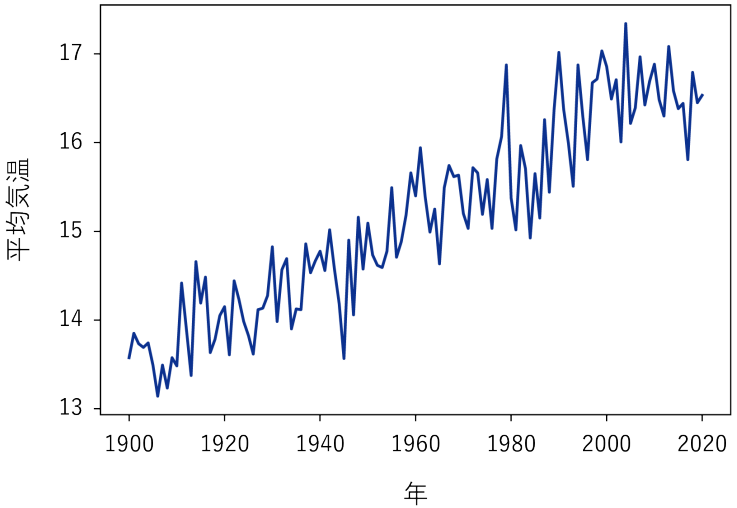
折れ線グラフの例 ①

年間平均気温（東京）

年	平均気温
1900	13.6
1901	13.9
…	…
2020	16.5

出典：気象庁「過去の気象データ・ダウンロード」　リンク

可視化



年によってばらつきがあるものの,
東京の年間平均気温は100年で3 (℃) 程度,
増加していることがわかります.

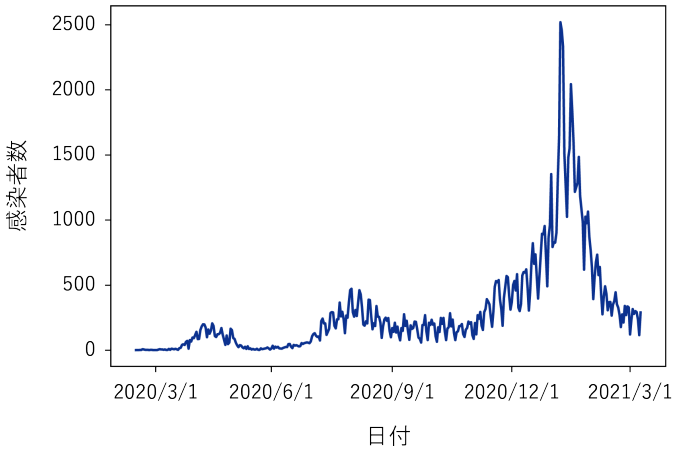
2．グラフによる可視化：B．変化（3）

折れ線グラフの例 ②

COVID-19感染者数（東京都）

日付	感染者数
2020/1/24	1
2020/1/25	1
…	…
2021/3/9	290

可視化



出典：東京都 新型コロナウイルス対策サイト　リンク

東京都の感染者数には3つの大きなピーク
2020年4月 (第1波)，　2020年7月 (第2波)，
2021年1月 (第3波)
があることがわかります．

詳しくデータを分析する場合は，回帰分析や時系列分析 (1.4節) を使います．

2．グラフによる可視化：C．構成比（1）

円グラフ（上）や帯グラフ（下）を作成すると，
全体に占める割合（構成比）を比較しやすくなります．

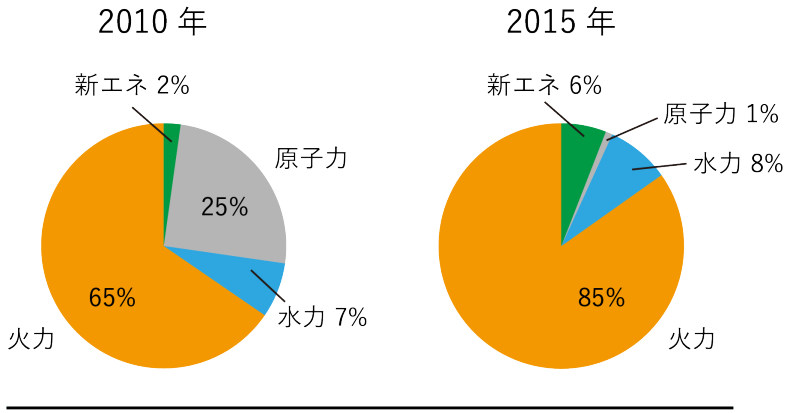
日本の発電量

	2010年	2015年
火力	65.4%	84.8%
水力	7.3%	8.4%
原子力	25.1%	0.9%
新エネなど	2.2%	5.9%

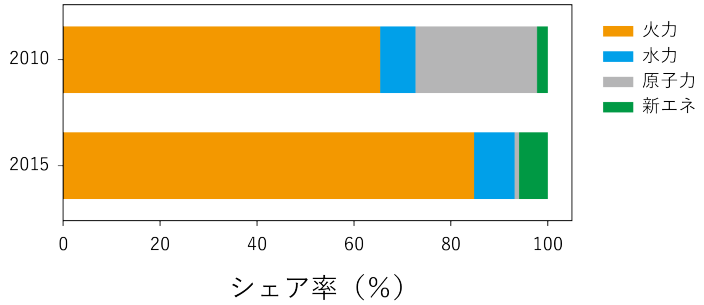
出典：資源エネルギー庁/エネルギー白書2019 リンク
ただし，石炭，石油，LNGを合わせたものを火力とした．

可視化

円グラフ



帯グラフ



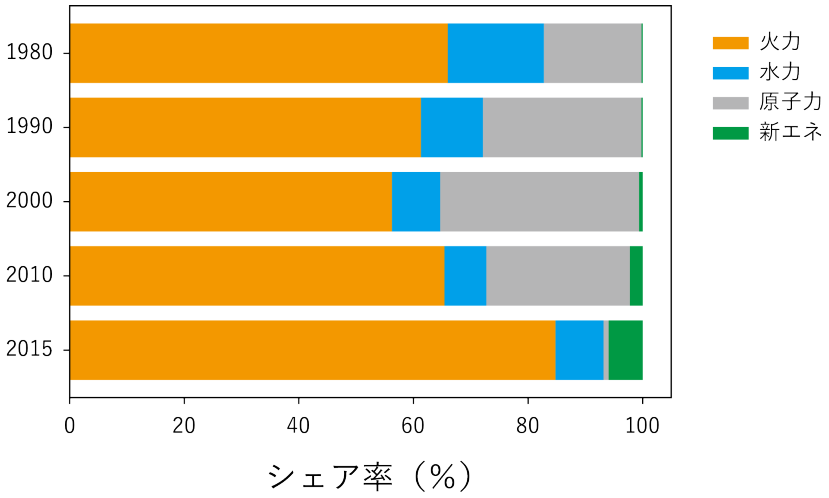
2015年は2010年に比べて，原子力発電の割合が大幅に減り，
火力発電の割合が大きく増加していることがわかります．

2．グラフによる可視化：C．構成比（2）

比較対象が多い場合には，
帯グラフ（左）や積み上げ縦棒グラフ（右）が便利です．

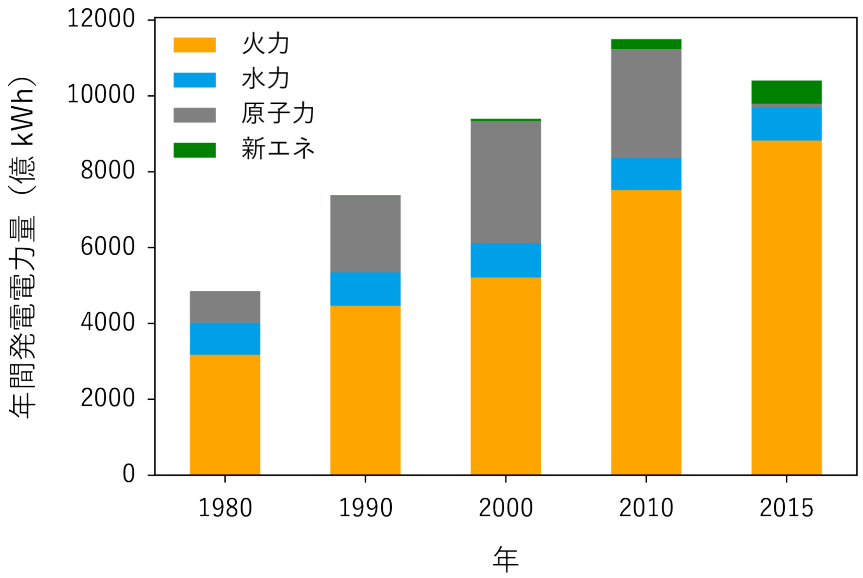
例：1980, 1990, 2000, 2010, 2015 年の日本の発電量の比較

帯グラフ



出典：資源エネルギー庁/エネルギー白書2019 リンク

積み上げ縦棒グラフ



どちらのグラフからも，火力発電の割合は2000年まで減少，それから増加していること，
新エネの割合は2010年以降急増していることがわかります．
積み上げ縦棒グラフ（右）から，日本の発電量は2010年がピークであったことがわかります．

2．グラフによる可視化：D．分布（1）

ヒストグラムを作成すると，データの散らばり具合（分布）を観察できます．

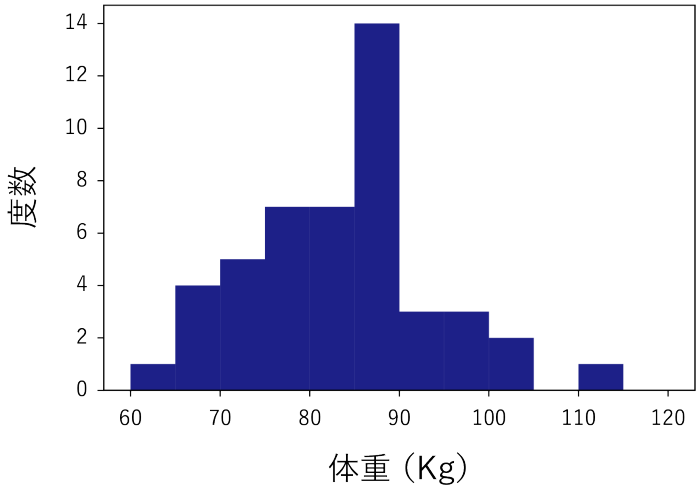
プロ野球選手の体重
（ソフトバンク，2020年，n=47）

選手	体重
バレンティン	100 Kg
柳田悠岐	96 Kg
…	…

出典：プロ野球データフリーク　リンク

可視化

ヒストグラム



上のグラフから，

① 体重 85～90 Kg の選手が多いこと，
② 90 Kg 以上の選手は少ないこと

がわかります．

2．グラフによる可視化：D．分布（2）

ヒートマップ を作成すると、2種類からなるデータの
散らばり具合（分布）を観察できます．

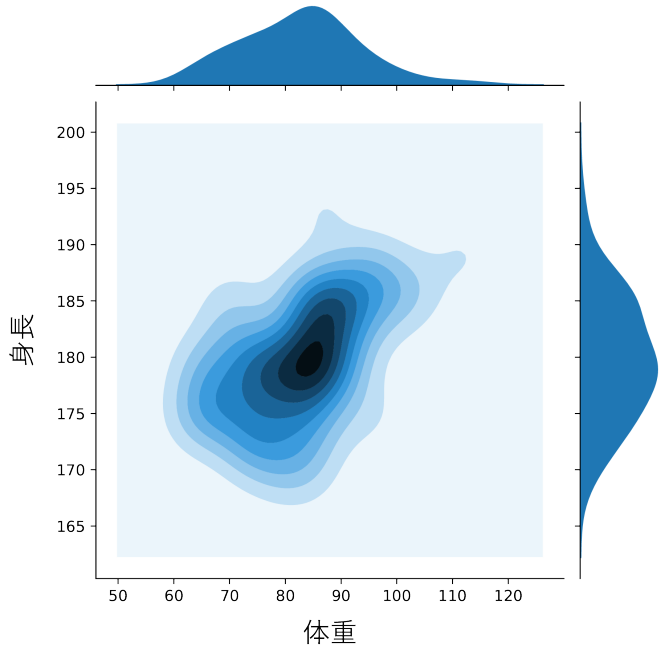
プロ野球選手の体重
（ソフトバンク，2020年）

選手	身長	体重
バレンティン	185 cm	100 Kg
柳田悠岐	188 cm	96 Kg
…	…	…

出典：プロ野球データフリーク　リンク

可視化

ヒートマップ



ヒートマップ から，身長 180 cm，体重 85 Kg 周辺
の選手が多いことがわかります．

2．グラフによる可視化：E．相関（1）

散布図を作成すると，2つの変数間（データ間）の関係を観察できます．

例１：プロ野球のチーム成績
（パリーグ，2015～19年，n= 30）

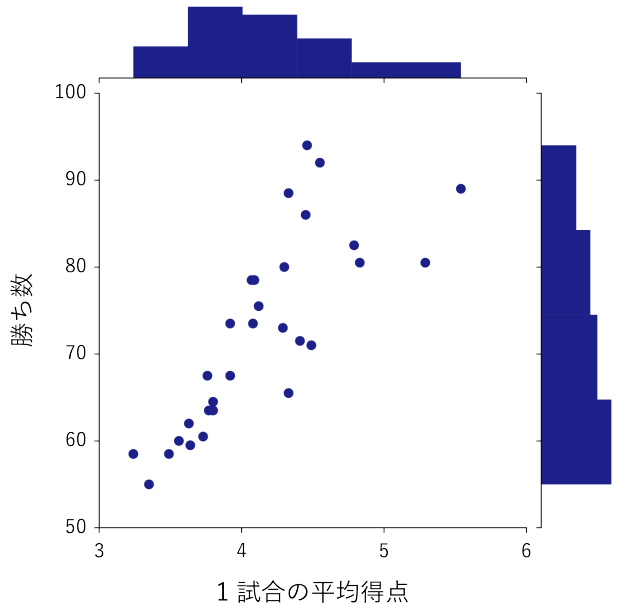
チーム名（年）	勝ち数	平均得点
西武（2019）	80.5	5.29
ソフトバンク（2019）	78.5	4.07
…	…	…

出典：プロ野球データフリーク　リンク

簡単のため、勝ち数＝ 勝利数 + 0.5 引き分け数とした．

可視化

散布図



各点（青丸）は1つのデータに対応します．平均的に多くの得点を取るチームほどシーズンの勝ち数が多くなっていることがわかります．ピアソン相関係数 0.79 はこの関係が強いことを示しています．

相関が高いからと言って因果関係があるわけではありません．つまり，平均得点と勝ち数の相関が高いからといって，得点を多くとるようにすればチームが強くなるとは限りません．逆に，因果関係が強ければ相関は強くなるはずなので，有力な因果関係候補を探すために散布図を使うことはできます．

2．グラフによる可視化：E．相関（2）

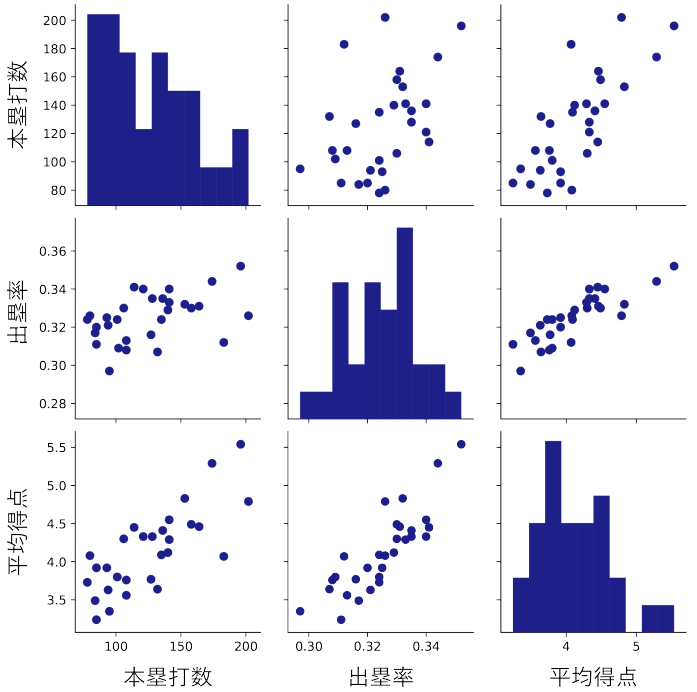
3つ以上の変数間（データ間）の関係を観察したい場合は，
散布図行列を使うと便利です．

例２：プロ野球のチーム成績
（パリーグ，2015～19年）

チーム名	本塁打数	出塁率	平均得点
西武	174	0.344	5.29
ソフトバンク	183	0.312	4.07
…	…	…	…

可視化

散布図行列



散布図行列（右）から，本塁打（ホームラン）数や出塁率が高いチームほど多くの得点を取っていることがわかります．

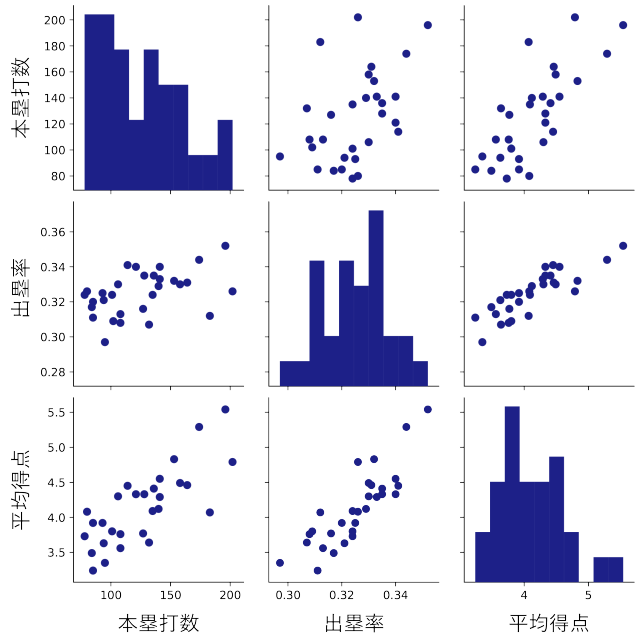
平均得点に関係が強いのは，本塁打数と出塁率のどちらでしょうか？ 相関行列を計算すると調べることができます（次ページ）．

対角成分（本塁打数, 本塁打数成分など）には散布図ではなく，ヒストグラムが示してあるので注意しましょう．
これは作図ツール seabornの仕様によるものです．

2．グラフによる可視化：E．相関（2）

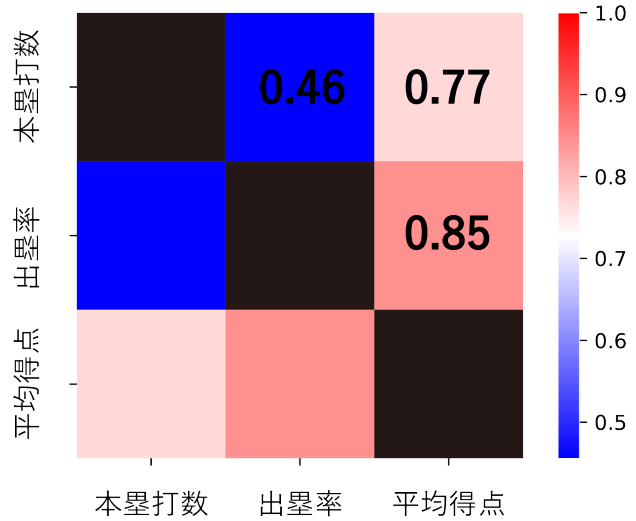
3つ以上の変数間（データ間）の関係性の強さを比較したい場合は，相関行列を計算しましょう．

散布図行列（p.17と同じもの）



相関行列

ヒートマップ



赤に近いペアほど相関が高い（関係性が強い）ことを示してます．上の図から，本塁打数より出塁率の方が得点と関係が強いことがわかります．

同じもの同士の相関係数は1になるが，今回の比較対象ではないため，対角成分は黒で示しました．

3．グラフ作成のコツ

知りたいことに合わせて，グラフの作成を工夫することが重要です．ここでは，３つのグラフについて紹介します．

A. 比較：棒グラフ

D. 分布：ヒストグラム

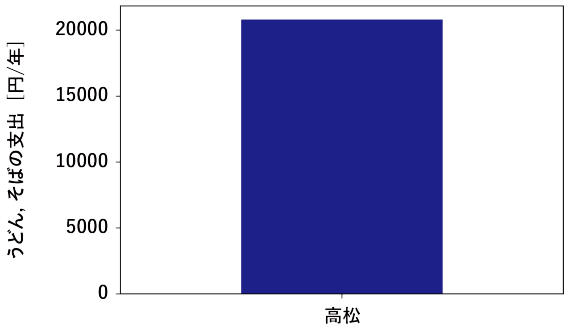
E. 相関：散布図

3．グラフ作成のコツ：棒グラフ

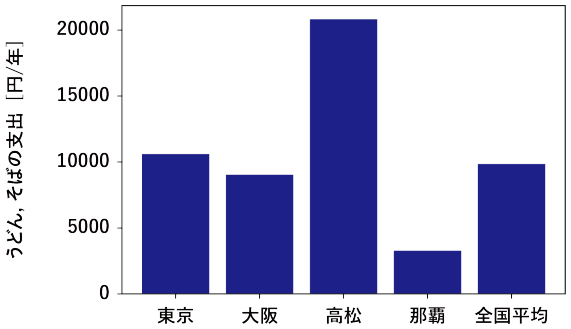
並べる棒の個数や順番に気をつけましょう．

- 棒の個数は少なすぎても（左），多すぎても（右）わかりにくくなります．
- 解釈しやすい順番（大きい順，北から南など）に棒を並べましょう．

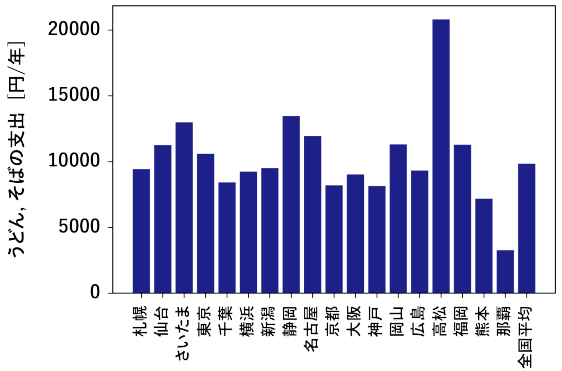
棒の数が少なすぎる



棒の数が適切



棒の数が多すぎる

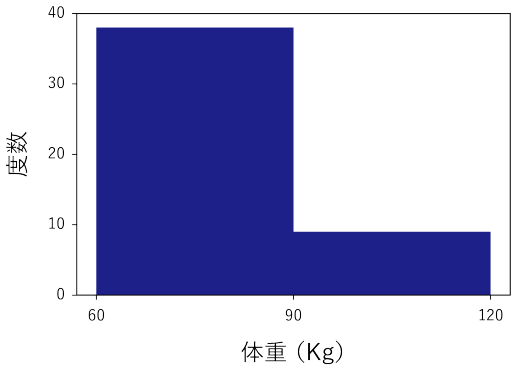


3．グラフ作成のコツ：ヒストグラム

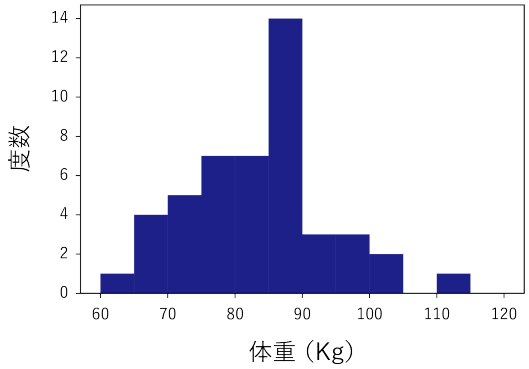
ヒストグラムを作る区間の幅（ビン幅）に気をつけましょう．

- ビン幅は太すぎても（左），細すぎても（右）
  データの散らばり具合（分布）はわかりにくくなります．

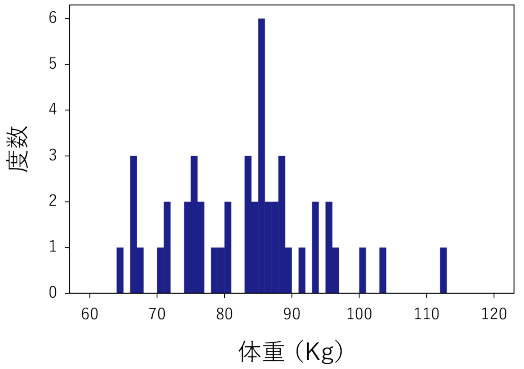
ビン幅が太すぎる



ビン幅が適切



ビン幅が細すぎる



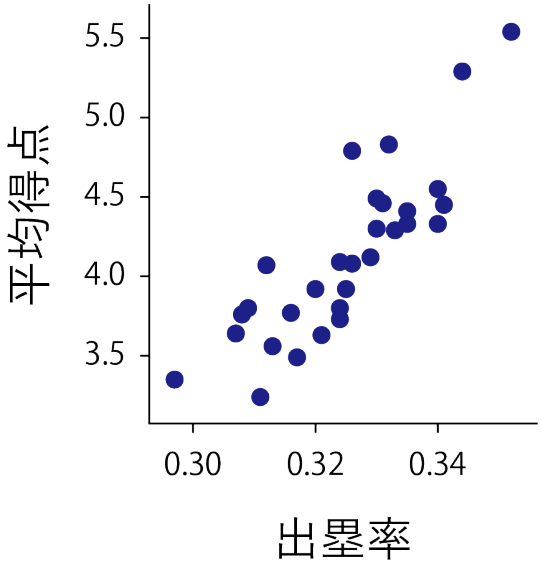
どんなデータにも最良となるビン幅はなく，いろいろなビン幅を試すことも重要です．

よく使われるビン幅の決定方法として，
スタージェスの公式，スコットの選択，島崎と篠本による方法などがあります．

参考文献：ヒストグラム (Wikipedia) リンク

3．グラフ作成のコツ：散布図

散布図：プロ野球データ（p.17）



説明したい，原因を知りたい変数（今回ではチームの平均得点）があれば，その変数を縦軸にとりしょう．

2つの変数の関係が直線的でない場合には，対数変換などを試すこともあります．

多数の変数がある場合には，適切な縦軸，横軸を選択する必要があります．

この方法として，

① 散布図行列を作成して関係性が強い変数ペアを目視で探す（p.17），
② 相関行列のヒートマップ を作成して相関係数が高いペアを探す（p.18），

より高度な方法として，

③ 予測精度やAIC （1.4節）などの指標を用いて関係性が強い変数ペアを探す，

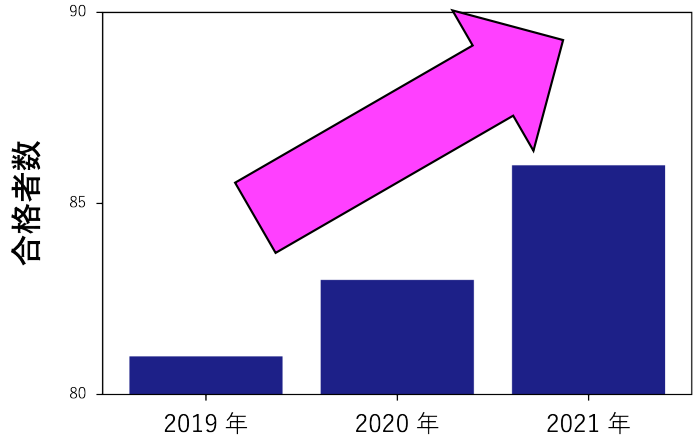
などが考えられます．

4．グラフがもたらす誤った解釈（1）

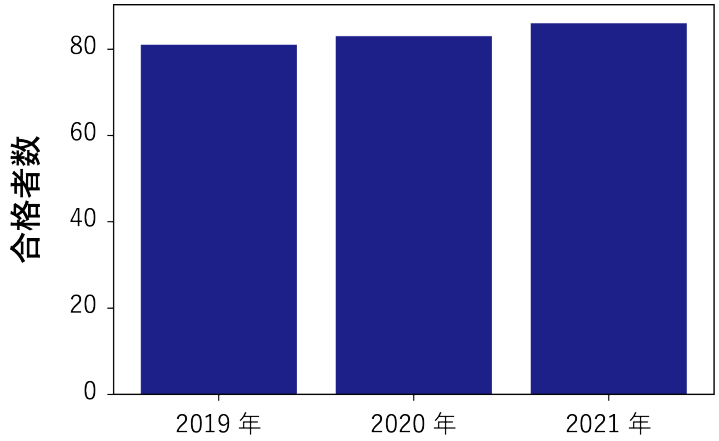
グラフはデータの誤った解釈を与えることがあります.

例1：DS予備校の● ●大学の合格者数実績（仮想データ）

DS予備校の躍進！



作り直したグラフ（0を基準）



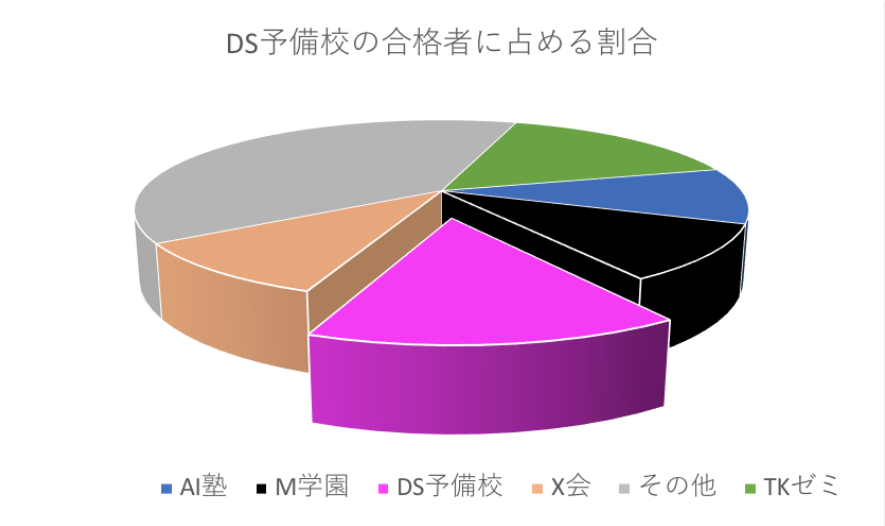
左のグラフを一見すると合格者数が年々急増しているように見えます.
グラフを作り直す（右）と合格者数はそれほど増えてない（81, 83, 86名）ことがわかります.

誤解を防ぐため，縦軸，横軸の数値やグラフの種類（対数表示等）を確認しましょう.

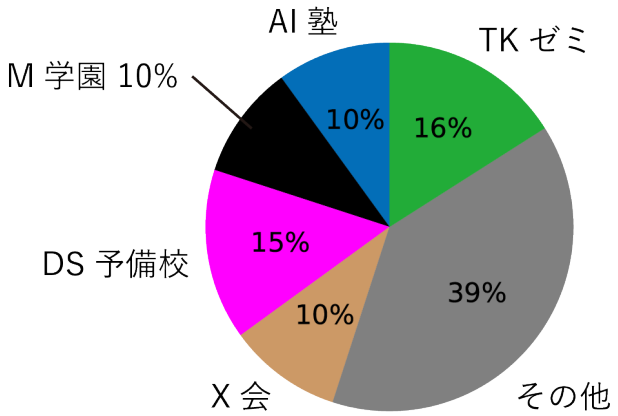
4．グラフがもたらす誤った解釈（2）

グラフはデータの誤った解釈を与えることがあります．

例2：DS予備校の● ●大学の合格者数の割合（仮想データ）



作り直したグラフ



左のグラフからは，DS予備校の割合がその他に次いで大きく見えます．
実際には，TKゼミの方がDS予備校より大きくなってます（左図）．
棒グラフや円グラフが3D になっている場合には注意しましょう．

参考文献

- 誤解を与える統計グラフ（Wikipedia）リンク
- Excel のダメなグラフでウソをつく法 リンク

5．ビッグデータの可視化（1）

ビッグデータとは，巨大で複雑なデータのことです．
ビッグデータが現れる状況として以下が考えられます．

A. **長い**：長期間あるいは高頻度のデータ

B. **種類が多い**

C. **新しいタイプ**：関係性，地図データなど

例：世界の気温

B. 種類が多い →

A. 長い ↓

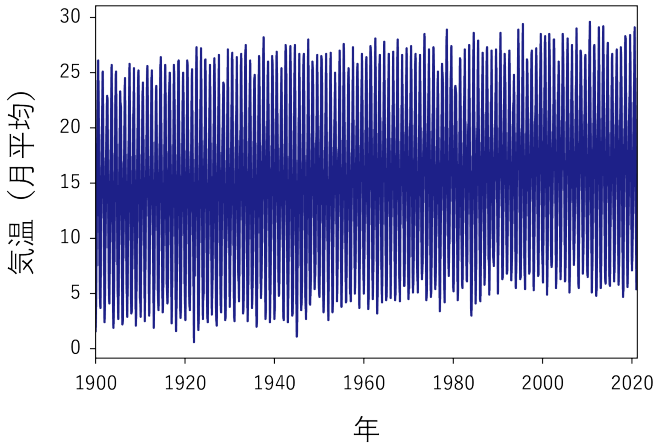
	東京	パリ	アブダビ	…
2020年12月	7.7	6.4	21.8	…
2020年11月	14.0	9.4	26.1	…
…	…	…	…	…

気象庁「世界の天候データツール（ClimatView 月統計値)」 リンク

5．ビッグデータの可視化（２）

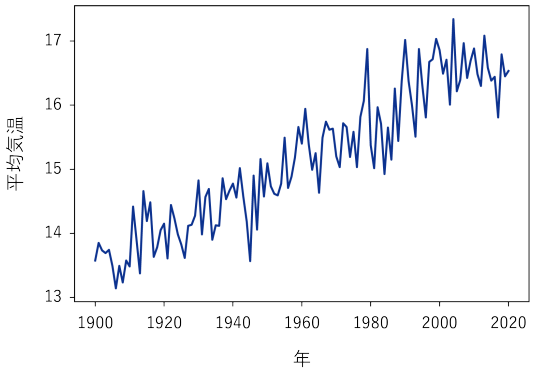
A. 長いデータは平均を計算すると可視化しやすくなります．

例：東京の気温



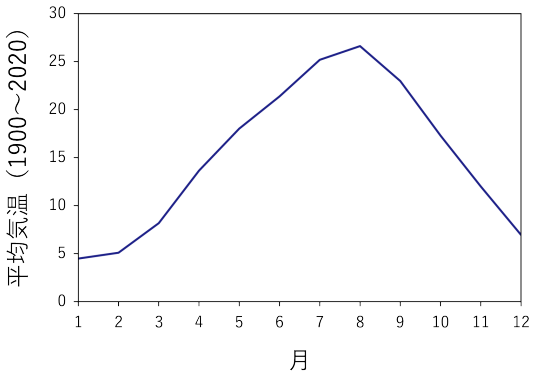
平均①

① 年ごと (1900年, 1901年…) の平均



平均②

② 月ごと (1月, 2月…) の平均



左は，東京の1900～2020年の月平均気温の折れ線グラフです．このグラフは複雑に変動しているように見えます．
右のように，年ごと（上），月ごと（下）の平均をとったグラフを作成することで，長期的な振る舞いと年単位での振る舞いを観察できます．

5．ビッグデータの可視化（３）

B. 一般に，種類が多い (4種類以上) ビッグデータの可視化は難しい問題です．この場合，次元削減技術 (主成分分析: 1.4節) を用いるとよいです．

C. 新しいタイプのビッグデータの可視化の方法

以下で3つの例を紹介します：

① グラフデータ → 関係性の可視化

② 地図データ → リアルタイム可視化

③ 映像データ → 軌跡の可視化

6．関係性の可視化（1）

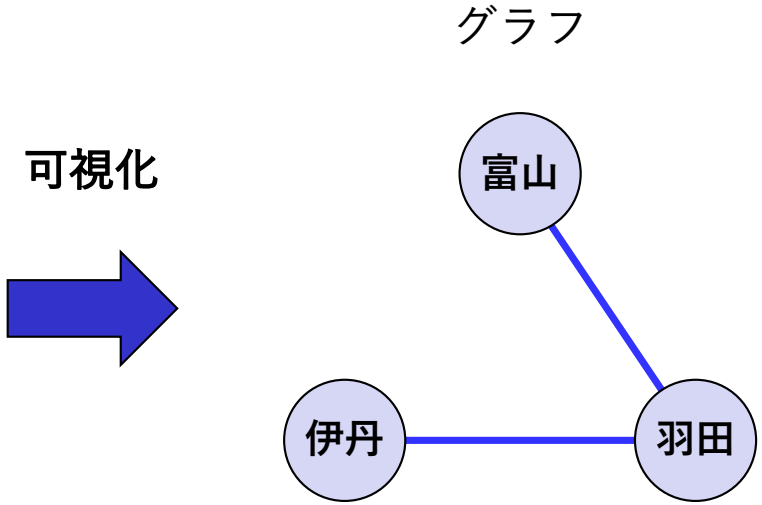
人や物の間につながりがあるかどうかを表現するため，
グラフ（ネットワーク）を用います．

例１：航空機の国内線ネットワーク

羽田（東京） ⇄ 富山：　直行便あり

羽田 ⇄ 伊丹（大阪）：　直行便あり

伊丹 ⇄ 富山：　　直行便なし



上の例では対象とする物は空港です．2つの空港を結ぶ直行便がある場合，
2つの空港はつながっているとします．

直行便についての情報は，右図のように，空港を丸，直行便を線で表現した“グラフ”で示されます．このように，丸（頂点：空港）が線（枝：つながりを示す）で結ばれた図形を“グラフ”と呼びます。

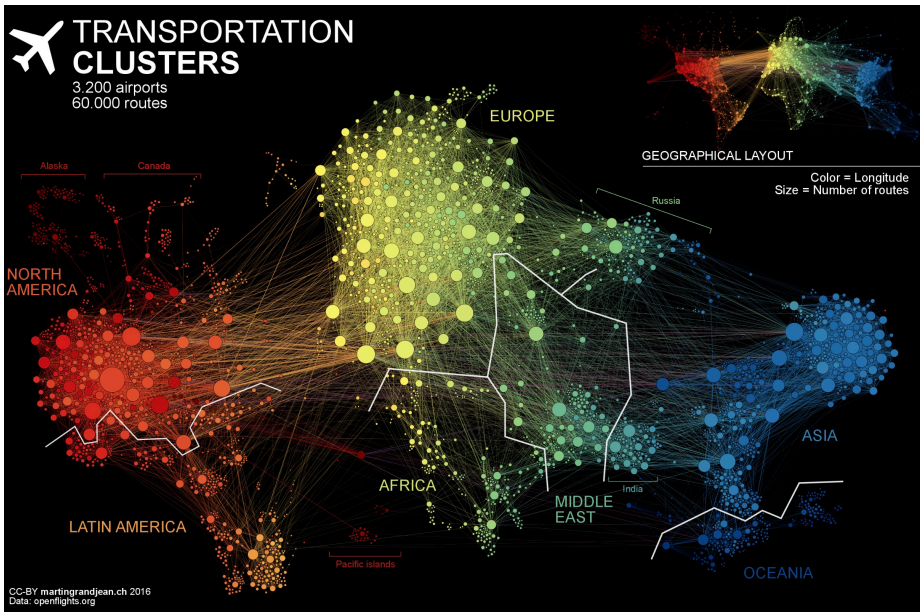
参考文献

- 複雑ネットワーク（Wikipedia）リンク

6．関係性の可視化（2）

人や物の間につながりがあるかどうかを表現するため，
グラフ（ネットワーク）を用います．

例２：航空機の国際線ネットワーク



出典：Martin Grandjean氏のホームページ　リンク

世界の3,200空港の路線図のグラフ．

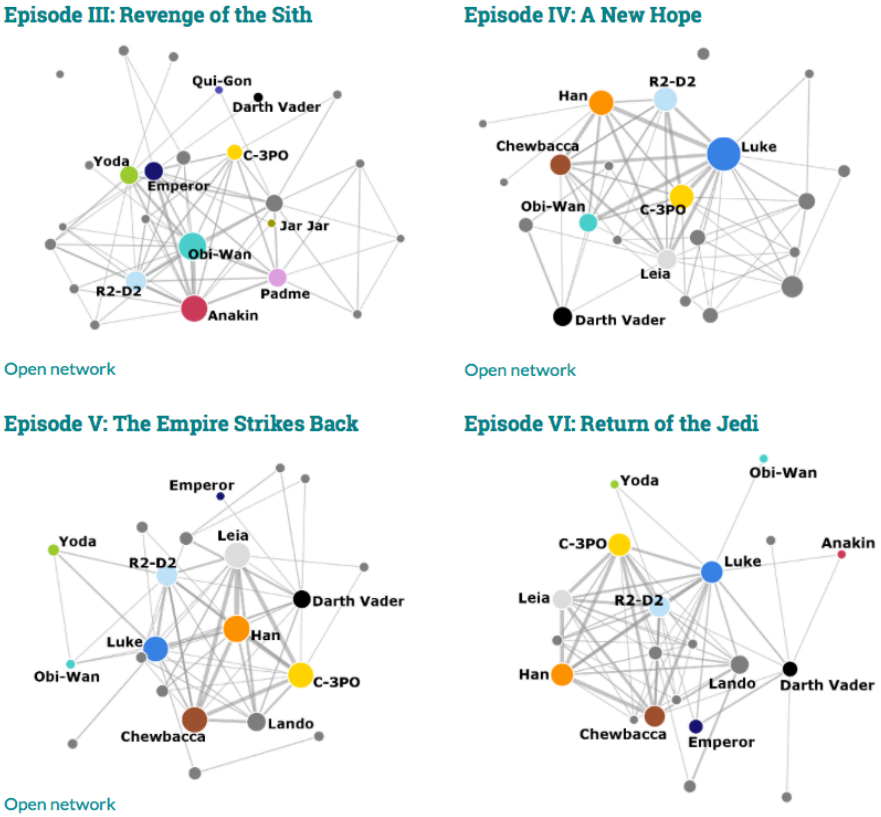
丸の大きさは空港からの路線数，丸の色は経度を表してます．路線数の多い空港はハブ空港と呼ばれます．

このデータを活用すると，人や物の流れをシミュレーションできたり，効率的な輸送計画を立てたりできます．

6．関係性の可視化（3）

人や物の間につながりがあるかどうかを表現するため，
グラフ（ネットワーク）を用います．

例３：スターウォーズ（映画）の人間関係



出典： Evelina Gabašová 氏のホームページ　リンク

スターウォーズの登場人物（ロボット、宇宙人も含める）のグラフ．

映画の同じシーンに現れた登場人物たちを「つながっている」と定義して、グラフを作成した．左のグラフから、映画の重要人物を可視化できることがわかります．

上段はスターウォーズ III（左）、IV（右）、下段はスターウォーズ V（左）、VI（右）の結果を表す．グラフデータは以下のリンクから入手できる．

データへのリンク

6．関係性の可視化（4）

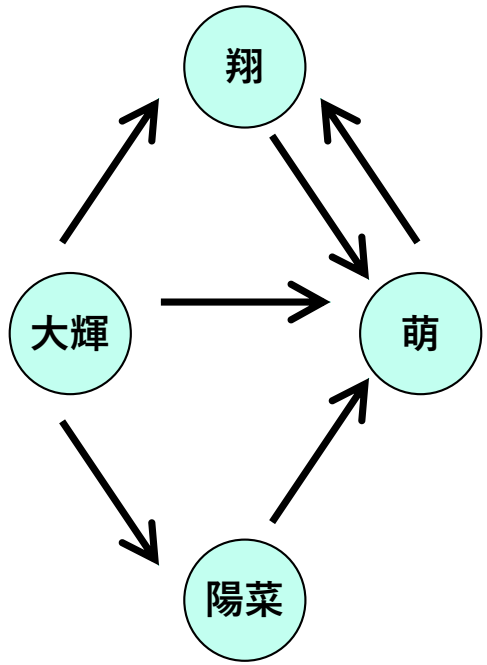
SNSでのフォロー関係，生態系での食う食われるの関係など，
つながりに向きがある場合には有向グラフになります．

例4：SNS（インスタ，Twitterなど）のフォロー関係

ユーザ	フォローしている人
大輝	翔，陽菜，萌
翔	萌
陽菜	萌
萌	翔

可視化

有向グラフ



上の例では，対象は人間であり，つながりはフォロー関係です．

このSNSデータは，右図のように，ユーザを丸，フォローを矢印で
表現したグラフになります．ツイートの拡散を考える場合には矢印
の向きが重要です．例えば，萌のツイートは全員に届きますが，
大輝のツイートは誰にも届きません．

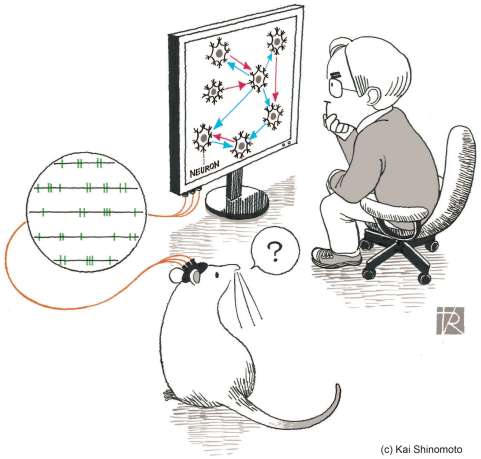
例．ツイートの拡散と政治的思想の関係を調べた研究：論文（英語）リンク、 日本語の解説　リンク

6．関係性の可視化（5）

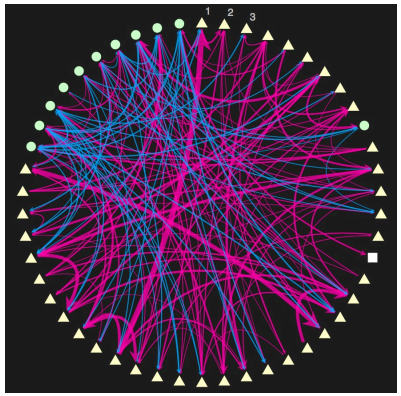
これまでは関係性についてのデータ（航空路線，人間関係など）を
持っている場合を考えました．関係性についてのデータがない場合も
別のデータを活用することで，関係性を推定，可視化できます．

① クラスタリングによる階層構造の可視化（1.4節：階層的クラスタリング参照）

② スパイクデータ（信号）から神経細胞間のつながりを推定



神経回路の可視化の例



左の図はスパイク信号（時系列）
から，50個の神経細胞間のつながり
を推定した例です．

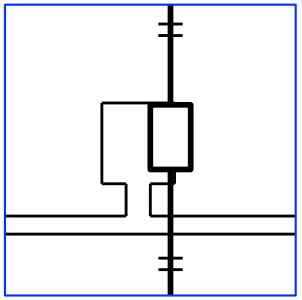
詳しくは以下を参照してください．

解説（日本語）：リンク
論文（英語）：　リンク

出典：篠本滋氏のホームページ　リンク

6．地図データの可視化（1）

地図

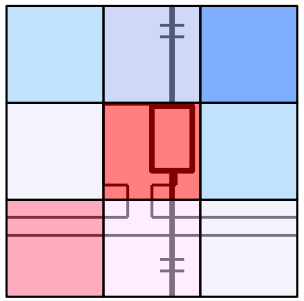


非空間データ
（例：人口密度）

4	6	2
9	18	4
14	12	8

データの統合
（地理情報システム）

地図上のデータ



地図情報システム（GIS）を使うと，
地図と非空間データを統合し，地図上のデータとして可視化や分析できるようになります．

参考文献

- GIS基礎教材：　リンク
- GISソフト（QGIS）：　リンク

6．地図データの可視化（2）

さらに，直近１時間のデータを非空間データとして用いることにより，
地図上の可視化をリアルタイムで行うこともできます.

また，1時間毎の混雑度合い（モバイル空間統計，Yahoo地図），
雨雲が動く様子（Yahoo地図），Covid-19感染者数の日毎の変化の
リアルタイム可視化を行うサービスもあります.

詳しくは以下の例を参照.

例

- モバイル空間統計（ドコモ）：リンク
- Yahoo地図（Yahoo)：リンク
- CoVid-19 感染者数（インフォマティクス）：リンク

実データ

- 関東圏の人の動きのデータ：リンク

6．軌跡の可視化

映像データをトラッキングすることにより，人やボールなどの動きの軌跡を可視化できるようになりつつあります．

ゴルフ，バレーボール，フェンシングなど様々なスポーツ競技の技術開発が進められています．

例

- ゴルフ：スィングの可視化（SPLYZA）　リンク
- バレーボール：ボールの軌跡の可視化（パナソニック）　リンク
- フェンシング：剣の軌跡の可視化（NHK）　リンク

