CC BY

データサイエンス概論I＆II

データとデータ分析

九州大学 数理・データサイエンス教育研究センター

データとは何か？

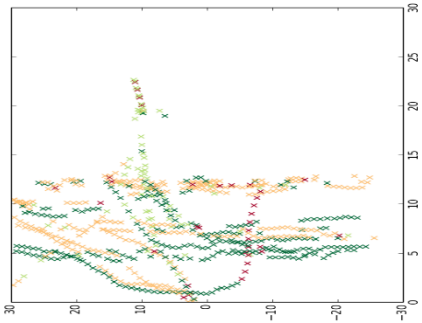
「データ」とは？ (デジタル大辞泉より)

1. 物事の推論の基礎となる事実。また、参考となる資料・情報。「―を集める」「確実な―」

2. コンピューターで、プログラムを使った処理の対象となる記号化・数字化された資料。

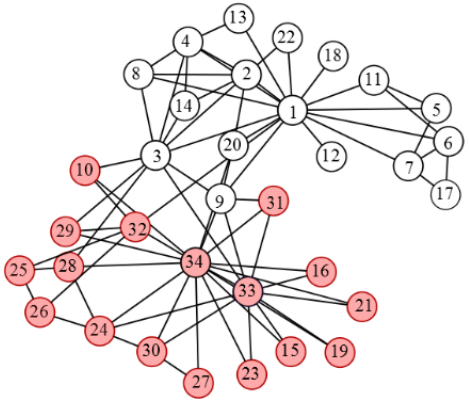
データとは？

- 測定値
  - 体温，体重，消費カロリー，人流
- メディアデータ
  - 画像(次スライド)，動画像（ビデオ），音声
- ラベルデータ
  - 患者の病名，地点名・駅名，生物種
- ネットワーク(関係データ)
  - 空手クラブメンバーの仲良し関係



伊都キャンパス内センサで獲得した人流データ

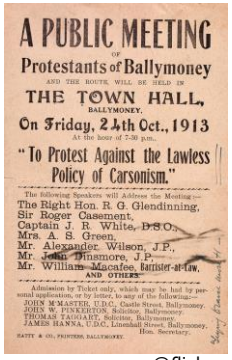
シロイヌナズナ by Alberto Salguero@Wikipedia



Zachary's Karate Club by Cuneytgurcan @Wikipedia

メディアデータの代表例：画像

- カメラ画像



commons@flickr

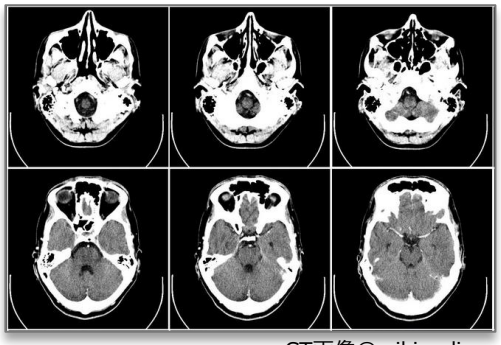
- 文字，文書，記号，標識，ナンバープレート

- 顔，指紋，虹彩，耳，唇，掌の静脈

IAM face dataset

@wikipedia

- CT・MRI・X線などの医用画像



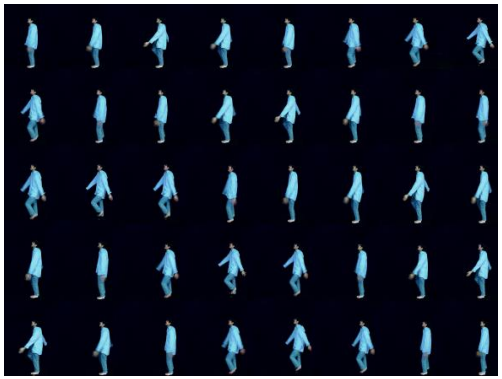
CT画像@wikipedia

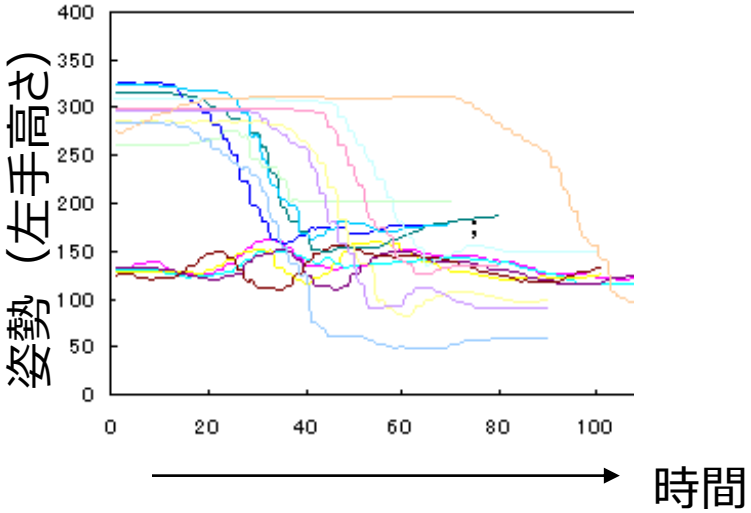
MRI画像@wikipedia

X線画像@wikipedia

データの種類～別の角度から：前後関係のあるデータ＝「系列データ」

- 時々刻々と得られる系列データ（時系列データ）
  - 動画像
  - 行動，ジェスチャ，歩行，ゲーム操作
  - 音声信号．対話系列
  - 心拍数変化，呼気量変化
  - 環境中のNOx濃度変化，気温変化
  - 10年ごとに測定した世界人口





- 時間とは関係のない系列データ
  - 文字列（文章）
  - DNA系列

cgcacagtgg atcctaggcg ttactaggtc
ttcaattctt gaactaattg ttttcgggtt …

別の角度からの分類：構造化データと非構造化データ

- 構造化データ
  - 簡単に言えば，表形式のデータ
  - 例えば，「横に月・縦に都道府県」を並べた表を作り，それを「ある月のある県での平均降水量」で埋めたとすれば，それは構造化データ

1-8 降水量（平年値）（昭和56年～平成22年）

観測地点	降水量（mm）												
	年計	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
札幌	1,107	114	94	78	57	53	47	81	124	135	109	104	112
青森	1,300	145	111	70	63	81	76	117	123	123	104	138	151
高知	2,548	59	106	190	244	292	346	328	283	350	166	125	58
福岡	1,612	68	72	113	117	143	255	278	172	178	74	85	60
佐賀	1,870	57	78	129	156	198	339	339	197	180	76	76	48
長崎	1,858	64	86	132	151	179	315	314	195	189	86	86	61
熊本	1,986	60	83	138	146	196	405	401	174	170	79	81	54
大分	1,645	45	65	112	129	150	274	253	172	220	121	69	34
宮崎	2,509	64	91	182	213	239	429	309	290	355	182	95	60
鹿児島	2,266	78	112	180	205	221	452	319	223	211	102	92	71
那覇	2,041	107	120	161	166	232	247	141	241	261	153	110	103

資料　気象庁「2010年平年値」

- 非構造化データ
  - 文章，画像，音がその代表例
  - 表形式にはならないので「非構造化データ」と呼ばれる
  - スマートフォンやパソコンで日々読んだり見たり聞いたりしているが，これらもデータ

データの一般的な4分類 (1/3)

- 量的データ
  - 比例データ（比率データとも）
    - 積や除算ができる．和や差もできる．Ex．体重．年収．長さ
  - 間隔データ
    - 積や除算に意味がない．ただし和や差はできる．Ex．(華氏・摂氏で測る)気温，西暦年
- 質的データ
  - 順位データ
    - 四則演算（加減乗除）すべて意味がない．ただし並べることはできる．
    - Ex．アンケート結果（５:非常によい，4:よい，3:ふつう，2:わるい，1:非常に悪い)．成績順
  - カテゴリデータ
    - 形式的に数字になっているだけ．
    - Ex.「1:女性, 2:男性」，電話番号，背番号，バスの系統番号

比例と間隔の違い
(わかりにくい？)は
次スライド！

「よい-ふつう=わるい-非常に悪い」
とはならない

データの一般的な4分類 (2/3)
比例データと間隔データの違いをもう少し

- 見分け方① ゼロが絶対的か相対的か
  - 気温は間隔データ．摂氏0度は人間が適当に決めたもの(＝相対的)なので
  - 標高も間隔データ．現在の標高0mは人間が決めたもの

  - 重さや長さは比例データ．ゼロは本当にゼロ（何もない状態）なので
  - 絶対温度は比例データ．絶対0度は全ての運動がゼロになるので

- 見分け方② 比に意味があるか？
  - 西暦は間隔データ．　西暦1500年は，西暦1000年の1.5倍ではない
  - 気温は間隔データ．気温30度は，気温20度の1.5倍ではない

  - 重さは比例データ．30グラムは20グラムの1.5倍
  - 収入は比例データ．給料40万は20万の2倍

データの一般的な4分類(3/3)
表としてまとめると…

	名称	可能な演算	主な代表値	主な事例
量的データ	比例データ	＋－×÷	各種平均	質量，長さ，年齢，時間，金額
	間隔データ	＋－	算術平均	温度（摂氏），知能指数
質的データ	順位データ	＞＝	中央値，最頻値	満足度，選好度，硬度
	カテゴリデータ	度数カウント	最頻値	電話番号，性別，血液型

Slide by H. Hayashi@KU

データの種類によって使える手法が
大きく異なってくる

あらゆる分野で
データ分析の必要性が
高まっている

データの分析
＝データから意味のある情報を引き出す

- 直感的には
  - データ＝コーヒー豆
  - 分析結果＝(おいしい)コーヒー

焙煎→粉砕→湯による成分抽出

- 適切な分析方法を用いなければ，意味のある情報は抽出できない
  - コーヒー豆を炒めて食べても，おいしくない

なぜ「今」データ分析なのか？

- 学術的・社会的要請
  - 客観性・再現性のある(＝だれがやっても同じになる)根拠が必要
  - さらにその根拠を数値として表現する必要
- データからの要請
  - データが大規模・複雑・多様化→手計算では無理
  - 分析が待たれるオープンデータの蓄積

- データ分析技術の進展
  - 計算機リソースの大規模化
  - 数値分析法，機械学習(特に深層学習）の進歩
  - オープンソース化，無料ライブラリ，技術解説サイト

皆さんもそのうちデータ分析に関わる可能性大：九大で実施されているデータ解析に基づく研究の例

- 新規強誘電体材料のインシリコ・オンデマンド探索
- 計算・観察・機械学習による電池開発の高速化支援
- 階層ノンパラメトリックベイズトピックモデルの開発
- プラズマプロセス分野における small data を基盤とした機械学習モデルの構築
- 植物の比較トランスクリプトームによる地球環境変化へのフェノロジー応答機構の解明
- Machine Learningによる電子教材使用時の学習活動パターンと教材内容理解度の関連性に関する研究
- 工学部におけるデータサイエンティスト養成のため教材開発
- 日本列島地殻内部の時空間モニタリングと人工知能を用いた危険予測
- 磁性柔軟材料による生体模倣運動の最適設計
- 夏季建設工事現場における脱水・熱中症災害防止のための労働従事者の生体情報の機械学習によるクラス分類
- 植物３次元形態データセットの作成と全体構造を記述する特徴量の開発
- 摂動した多体力学系のダイナミックモード分解による解析
- 訓点資料本文データベース作成のためのシステム構築
- 地域社会における相互文化理解と多文化理解教育に関する包括的研究
- 感覚間における時間情報統合の心理物理学的検討
- 深層学習を用いた被害写真に基づく震災マンションの被災度判定・復旧費用概算システムの開発
- 付加詞条件の普遍性と統語構造―日・中・英語データに基づく実証的研究
- 都市工学と経済学の融合：持続可能な発展へ向けた将来設計

研究者だけじゃない！

誰もが「無意識に」データを分析しながら生きている

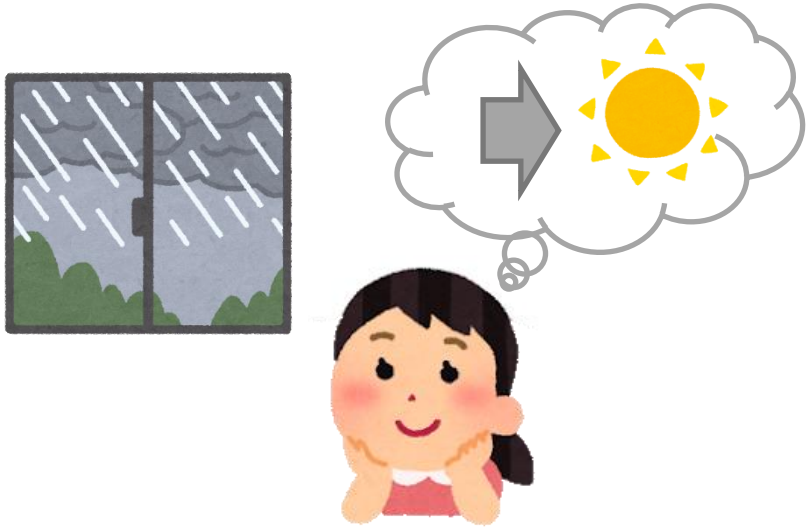
データ分析，タスクの例

- 予測
  - 未来を予測ことはできないか？
- （傾向や関連の）発見
  - これまでに気づかなかった傾向や関連などを発見できないか？
- 分類・グルーピング
  - たくさんのデータを「似たデータ」ごとにまとめられないか？

難しそう？　自分がやってるわけがない？ そんなことはない！
自らの「過去の経験」をデータとして使って，皆さん，やってますよ

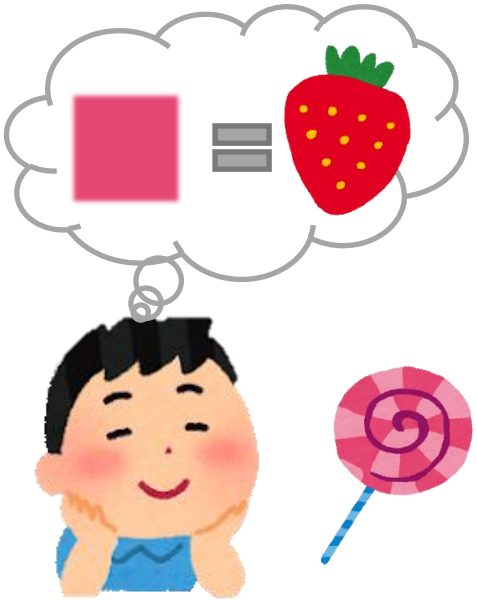
皆さんもやっているデータ分析：予測

- 「ここのところずっと雨なので，明日は晴れるだろう」
- 「あと2 時間もすれば，この宿題も終わるだろう」
- 「次はカーブを投げてくるだろう」
- 「これだけ勉強すれば，100点取れるだろう」



皆さんもやっているデータ分析：傾向や関連の発見

- 「赤いアメはイチゴ味」
- 「いい子にしていればサンタがプレゼントを持ってくる」
- 「夜更かしすると翌朝起きられない」



皆さんもやっているデータ分析：分類・グルーピング

- 「目の前の動物は犬か猫か」
- 「母親の表情がいつもとちょっと違う」
- 「私が好きなタイプの本」

- 「私とあの人は性格が似ている」
- 「ラーメンには，しょうゆ，豚骨，ミソ，塩がある」

データ分析は，数学が苦手な人どころか，算数を習ってすらいない幼児にとっても，極めて身近なもの

- 先入観は捨てよう
  - 「データ＝数字が並んだ無味乾燥なもの」→ No!
  - 「データ分析＝難しくて専門家しかできない」→ No!
  - 「自分の人生には関係ない」 → No!

- その面白さを是非理解してほしい
  - ある意味「柔らかく」「結果も色々ありうる」人間らしい話

- 自分自身が日々（無意識に）どのようなデータ分析をしながら生きているのかを考えてみると，楽しいはず！