データサイエンスについて

森本　孝之（関西学院大学　理工学部　数理科学科）

2021 年度研究室紹介特設ページ用資料

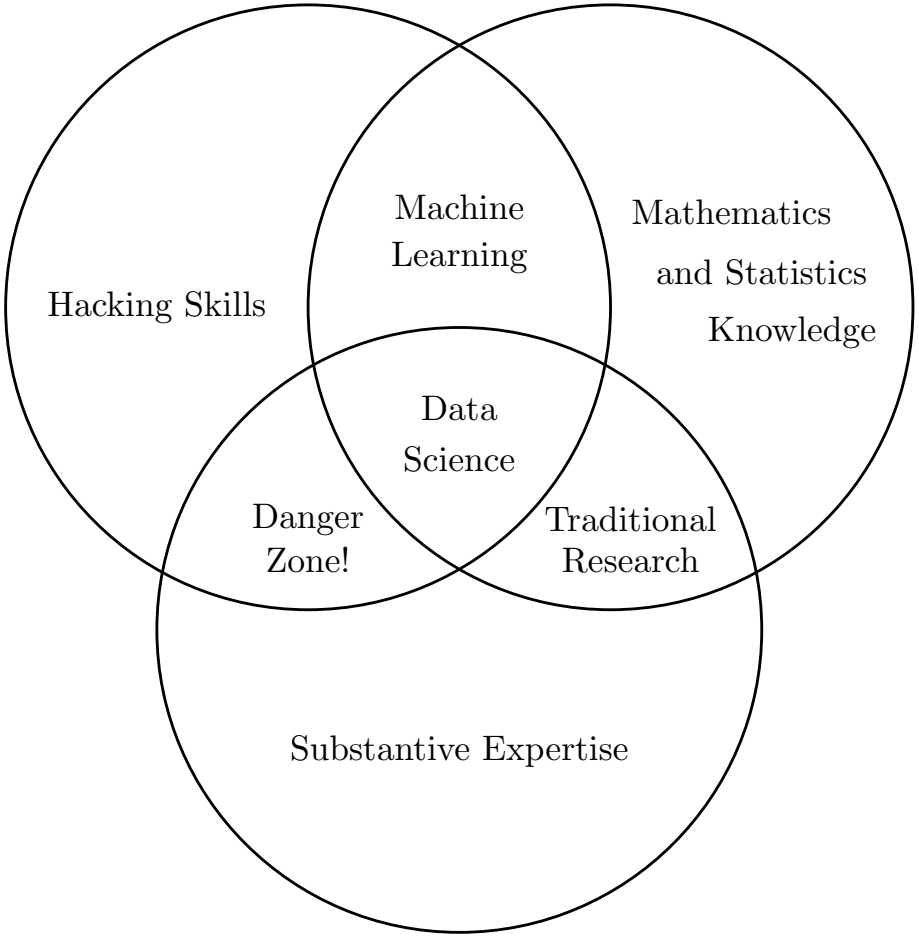
2020年11月29日

目次

目次	1
データサイエンスのベン（Venn）図（1）	2
データサイエンスのベン（Venn）図（2）	3
修正版データサイエンスのベン（Venn）図	4
データサイエンスの定義（1）	5
データサイエンスの定義（2）	6
統計学の定義，統計学の特性，統計学を学ぶ本質的な意義	7
PPDAC サイクル	8
フィッシャーによる三原則	9
統計実験のための三原則	10
統計，データサイエンスの歴史的推移（1）	11
統計，データサイエンスの歴史的推移（2）	12
統計における歴史的な画期的貢献に関する表	13
ランダムネス（randomness）の定義	14
ランダムネスの懐柔と活用	15
無作為標本，無作為割付およびランダム回答法	16
データサイエンスにおけるデータの種類	17
量的データの分類と統計量	18
データの質，データの量およびデータの種類	19
3 つの V	20
5 つの V	21
参考文献	22

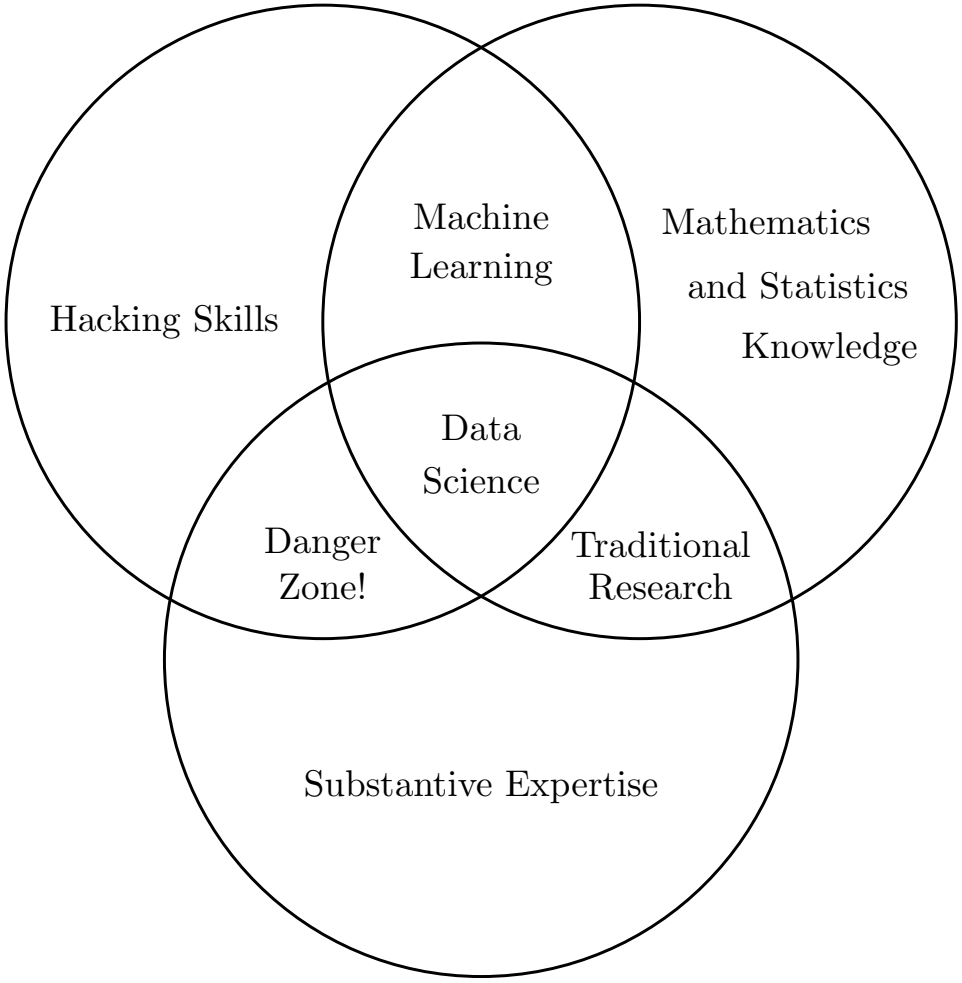
データサイエンスのベン（Venn）図（1）

- データサイエンスの定義を考えるとき，「データ」とか何か，「サイエンス」とは何か，そして「データサイエンス」とは何か，というように定義にもいろいろなものが提唱されているが，有名なものにドリュー・コンウェイ（Drew Conway, 2013）によるデータサイエンスのベン（Venn）図がある．



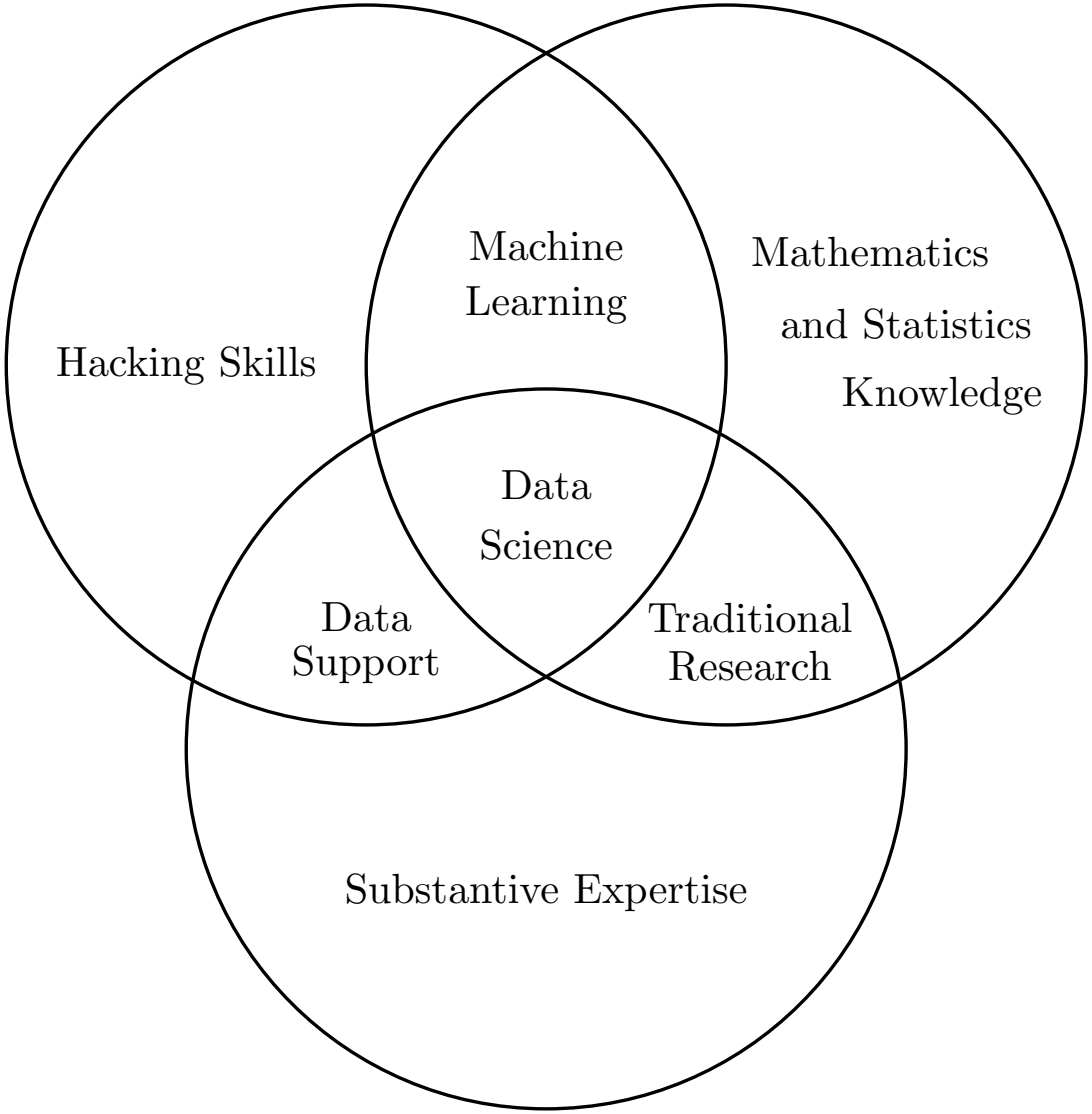
データサイエンスのベン（Venn）図（2）

- Substantive Expertise は実質的な専門知識，Traditional Research は従来の研究，Hacking Skills はプログラミング技術の意味である．
- Danger Zone!（危険地帯）はコンウェイの洒落でもあろう．



修正版データサイエンスのベン（Venn）図

- ベイリー（Bailey, 2017）は Danger Zone!（危険地帯）ではなくて Data Support（データ補助）であると修正提案をしている.



データサイエンスの定義（1)

- 統計学者のドノホー（Donoho, 2017）はデータサイエンスを以下の６項目の結合として定義した:

  (1) Data gathering, preparation, and exploration
  (データの収集，前処理，そして調査や吟味)

  (2) Data representation and transformation
  (データの表現や変換)

  (3) Computing with data（データを用いた計算)

  (4) Data modeling（データに対するモデリング)

  (5) Data visualization and presentation
  (データの視覚化とその説明)

  (6) Science about data science（データサイエンスに関連する科学)

データサイエンスの定義（2）

- 統計学者の柴田里程（2018）は

> 『辞書によれば，データとは「数値，記号で表した推論の根拠となるもの」である． この定義のポイントは「推論」という目的が含まれている点で，これが無ければ単なる「情報」でしかない． その上で，データサイエンスは「データに関するなぜを追求するサイエンス」が定義となる.』

という定義を提唱し，また

> 『データサイエンス実践にあたっては，まず「データの総体的な理解」を基本に据えることが欠かせない． 特段の推論という目的無しに単にサマリーを作ったり，逆に特定の目的にむかってデータをつまみ食いするだけなら，なにもデータサイエンスは必要ない． 「データを的確にとらえ理解する」ことの助けとなるのがデータサイエンスであり，「どのようにしたらデータから新たな価値を発見できるか」その指針を与えるのがデータサイエンスであるからである.』

とデータサイエンスの実践を定義している．

統計学の定義，統計学の特性，統計学を学ぶ本質的な意義

- 以上のことからも，データサイエンスがいかに総合的な取り組みであるかがわかり，データの本質の理解およびデータからの新たな価値の発見という目的から鑑みてもデータサイエンスの必要性は明らかであり，今後ますます重要性は増していくであろう．

- ここで，日本学術会議における数理科学委員会での「統計学分野の参照規準検討分科会」によって作成された平成 27 (2015) 年 12 月 17 日の報告書からの抜粋を示しておく．

  統計学の定義: 統計学は，データを元に現象を記述し，現象のモデルを構築し知識を獲得するための方法論である．

  統計学の特性: 帰納的推論の中に演繹的論理の過程を導入することにより科学的な結論を導くことにある．

  統計学を学ぶ本質的な意義: 自然や人間社会における不確実性の理解とそれへの対処法の習得，課題解決型思考力の獲得等である．

PPDAC サイクル

- データサイエンスにおけるプロシージャとして PPDAC サイクルも一般的に提唱されている.

(1) P: Problem (問題設定)
(2) P: Plan (計画設定)
(3) D: Data (データ収集)
(4) A: Analysis (データ分析)
(5) C: Conclusion (結論導出)

- PPDAC サイクルを繰り返すことで, 先のデータサイエンスにおける定義で『データに関するなぜを追及するサイエンス』(柴田, 2018) を実践できる.

- PPDAC サイクルは, カール・ポパー (Popper) の提唱する科学的研究方法との対応が可能である.

(1) 仮説理論
(2) 演繹的推論 (Problem)
(3) 理論の結果
(4) 実験計画 (Plan)
(5) 実証データ (Data)
(6) 帰納的推論 (Analysis)
(7) 理論の実証
(8) 啓発的推論 (Conclusion)

フィッシャーによる三原則

- 20 世紀の統計学者で影響力の大きかったのがイギリス人のフィッシャー (R. A. Fisher) 卿である.
- そのフィッシャーが統計学とは何かというのを示したのが以下のものである.

フィッシャーによる三原則

(1) 母集団 (population) に関する研究: Statistics

(2) 変動や多様性に関する研究: Variation

(3) データの縮約方法に関する研究: Reduction of Data

- 母集団に関する研究というのは，ある国全体の人口，生産量などの，ある想定している集団全体の状況を把握するということで，日本でいえば法律で定められている国勢調査などによって母集団を把握し，最近の状況がどうなっているかを調べることである.

統計実験のための三原則

- 変動や多様性に関する研究というのは，1 つには時間に関係する変動で昨年と今年の経済活動の変動を把握することであり，同一の場所でいえばある農地における作物の収穫量の多様性の原因を探ることである．
- データの縮約方法の研究というのは，覚えておいて損はないキーワードで，例えば 100 人の身長データがあった場合，それらの数値をただ単に眺めてもそこから情報らしいことは得られにくいが，標本平均と標本分散を計算し求めるだけでもデータの状況を把握しやすくなる．
- フィッシャーは統計実験のための三原則も以下のように提唱している．

統計実験のための三原則

(1) 局所管理 (local control): 系統的な誤差をできるだけ排除

(2) ランダム化 (randomization): 残った系統誤差を偶然誤差に転化

(3) 反復 (replication): その誤差を評価

統計，データサイエンスの歴史的推移（1）

- コックス（D. R. Cox）卿による 20 世紀の統計学の流派を分類したものが以下のものである.

頻度論的確率に基づく

- フィッシャー学派の帰納的推論の理論
- ネイマン・ピアソン・ワルド学派の帰納的行動の理論

ベイズ（Bayes）統計理論に基づく

- 論理的確率のジェフリーズ学派
- 主観的確率のサベージ学派

統計，データサイエンスの歴史的推移（2）

- 頻度論的確率の枠組みにおいて，帰納的推論というのは

> 得られたデータからデータの発生する母集団の状況を帰納的に推定する理論のこと

- 帰納的行動というのは

> 得られたデータからデータの発生する母集団の状況を統計的検定という手段によって採択するか棄却するか，という行動に出ること

- ベイズ統計の枠組みにおいて，論理的確率は

> 無情報事前分布やジェフリーズ分布といった客観性に則った客観的事前分布のこと

- 主観的確率は

> ある意味恣意的な主観に則った主観的事前分布のこと

統計における歴史的な画期的貢献に関する表

年代	概念	貢献者，貢献分野	関連語
19世紀	Average Man	ケトレー	ガウス・ラプラス
1900	χ^2-test	ピアソン	Biometrika
1908	t 統計量	スチューデント（ゴセット）	小標本
1925	フィッシャー情報量	フィッシャー	最尤推定量
1933	統計的検定	ネイマン，ピアソン	フィッシャーの反対
1937	信頼区間	ネイマン	統計的推測
1950	統計的決定理論	ワルド，ベイズ推定	
1962	データ解析	モステラー，テューキー	探索的データ解析
1972	比例ハザード	コックス	生存解析
1979	ブートストラップ	エフロン	MCMC
1995	False-discovery Rates	ベンジャミニ，ホッホバーグ	LASSO
2000	Large-scale 推測	Microarray technology	微生物学的データ
2001	Random forests	機械学習	Boosting
2016a	データサイエンス	Data Science Association	ビッグデータ
2016b	Personalized medicine	Genome-wide association studies	ビッグデータ

参考文献: Efron and Hastie (2016)

- 19 世紀から 1950 年代までは，統計関係の関心は応用的関心から数学的関心へと変遷した結果として統計的決定理論を確立し，そこから，計算機の発達とともに，応用的な関心と計算機的な関心を併せ持つような領域へと変遷していったことがわかる．

ランダムネス（randomness）の定義

- 統計実験の三原則の 1 つにランダム化というものがあった.
- 系統誤差を偶然誤差に転化するためにランダム化を行うときの偶然誤差に関して，誤差項をランダム項という.
- ランダムネス（randomness）の定義は，Oxford Advanced Learner's dictionary 9th edition (2015) によると

  “the fact of being done, chosen, etc. without somebody deciding in advance what is going to happen, or without any regular pattern (誰かが前もって何が起こるかを決めることなく，若しくはどんな規則的なパターンもなく，なされたり選ばれたりなどすること)”

★ このようにどんな規則的なパターンもないようなランダムネスに関して，統計学やデータサイエンスにおいては，確率的な挙動という枠組みによって次のような 2 通りの対応を行う.

ランダムネスの懐柔と活用

(1) ランダムネスの懐柔

- ランダム誤差をうまく飼い慣らすことによって不可知であることを評価することができるので，各種統計モデルへの適用がなされている．

- 例えば，$\mathbf{x}$ という説明変数に対してある関数 f によって被説明変数である $\boldsymbol{Y}$ が得られるとしたときの統計モデルは

$$\boldsymbol{Y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

のように表現することができ，ここでの $\boldsymbol{\varepsilon}$ がランダム誤差である．

(2) ランダムネスの活用

- ランダムネスの懐柔は，どちらかといえばランダムネスのネガティブな取り扱いということができるが，その反対にランダムネスの活用というのは，どちらかといえばランダムネスのポジティブな取り扱いということができる．

無作為標本，無作為割付およびランダム回答法

無作為標本（random sample）

ある母集団から無作為に（ランダムに）サンプルが抽出された無作為標本によって，帰納的にその母集団の姿を推測することができる．

無作為割付（random assignment）

ある医師の診察を受けるインフルエンザの患者に対して，ランダムにタミフルを投与するか投与しないか，という無作為割付を行うことで，タミフルの投薬効果を得ることができる．

ランダム回答法

あるアンケート調査で調査員に回答結果を知られることなく回答することができることでプライバシーに関する調査を行うことが可能となる．

データサイエンスにおけるデータの種類

- データ (data), 標本 (sample), 観測値 (observation):

  実験や調査, 観察, 観測等の結果得られた数値や属性

  - 質的データ (qualitative data) [属性]
    - 名義 (nominal) 尺度データ (カテゴリカルデータ)
      男女, 国籍, 職種など
    - 順序 (ordinal) 尺度データ (反応カテゴリーなど)
      嗜好度合, 5 段階アンケート調査, 新生児指数など
  - 量的データ (quantitative data) [変量]
    - 間隔 (interval) 尺度データ (観測値間の和や差に意味がある)
      気温, テストの点数, 方角など
    - 比率 (ratio) 尺度データ (2 つの観測値の比が意味を持つ)
      距離, 重量, 価格など

量的データの分類と統計量

- 量的データの分類
  - 離散型データ (discrete data)，計数値データ (count data)
    家族の人数，月間事故件数など
  - 連続型データ (continuous data)，計量値データ (metrical data)
    身長，体重，摂取カロリーなど
- 統計量 (statistic)
  - 母集団から得られた標本データの意味するところをうまく表現する量，またはデータの関数
- こうした分類を踏まえて，データサイエンスの関わるデータとしては，最近ではビッグデータ (big data) などの巨大なデータ群が存在する.
- その際，データにおいて注意すべき点などを列挙しておく.

データの質，データの量およびデータの種類

データの質

無作為標本されたようなきれいなデータばかりではなく，データに欠損値が多いとか，データ収集において偏り（バイアス）が掛かっているとか，とにかくデータの質に関しては非常に注意する必要がある．

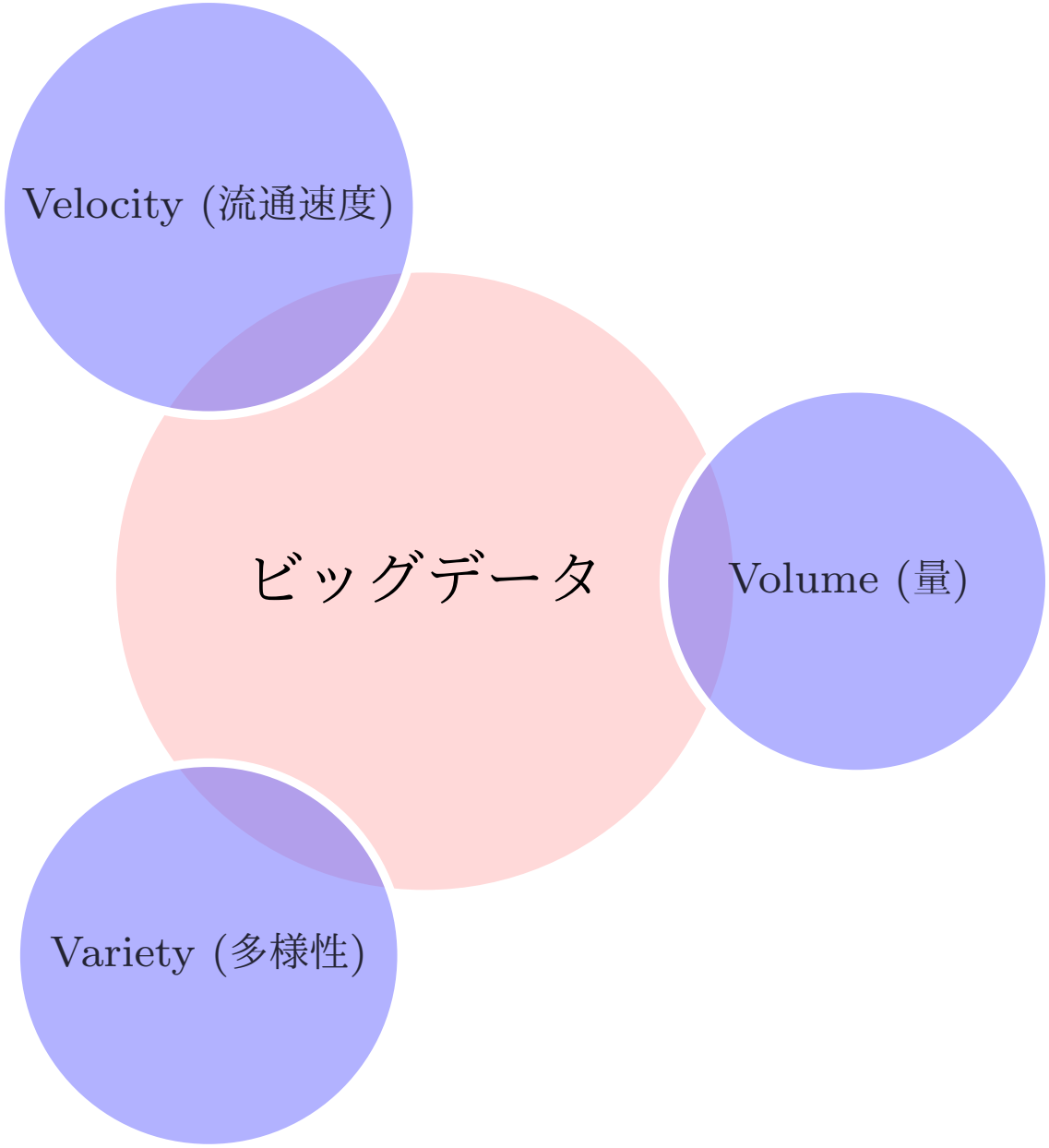
データの量

従来の小標本と大標本という括りに加えて，ビッグデータと呼ばれる巨大データ量がある．

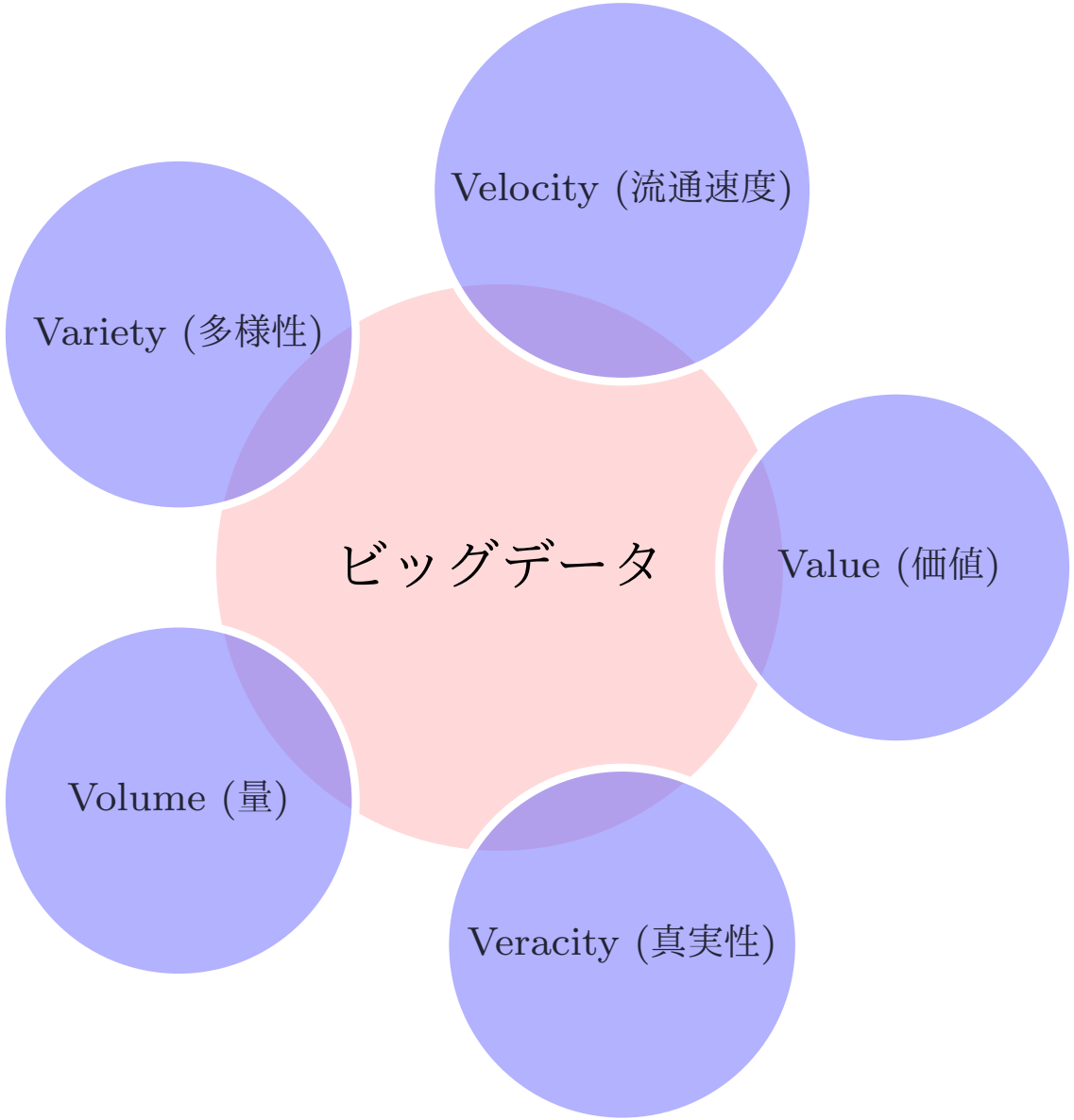
データの種類

従来の質的データと量的データに加えて，画像データのような，それらの混合データなどもあり，さまざまな種類に対応する必要がある．

ビッグデータ（3 つの V）



ビッグデータ（5 つの V）


Velocity (流通速度)
Variety (多様性)
ビッグデータ
Value (価値)
Volume (量)
Veracity (真実性)

参考文献

- 濱田 悦生，狩野裕（編集）(2019) データサイエンスの基礎，講談社
- 柴田 里程 (2018) データサイエンス普及の隘路，2018 年度統計関連学会連合大会予稿集
- D.Donoho (2017) 50 years of data science, *Journal of Computational and Graphical Statistics*, **26**, 745-766.
- B.Efron and T.Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press.
- A.S.Hornby (2015) *Oxford Advanced Learner's Dictionary of Current English, 9th edition*, Oxford University Press.

(順不同)