

これからの社会での データサイエンスの役割

小森 理

成蹊大学 理工学部

場所：大妻高等学校

日時：2026年6月11日(木)

本日の内容

① データサイエンス

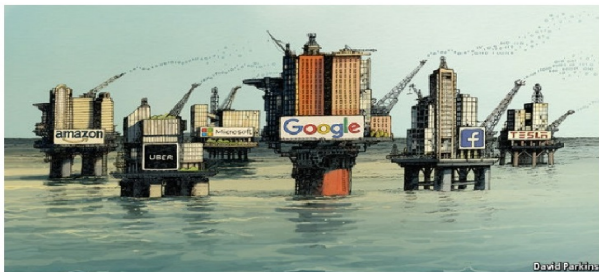
② 統計科学の歴史と実例

③ データサイエンスの最前線（人工知能，AI）

データ社会

The world's most valuable resource is no longer oil,
but data

The data economy demands a new approach to antitrust rules



引用：The Economist May 6th 2017 page 7

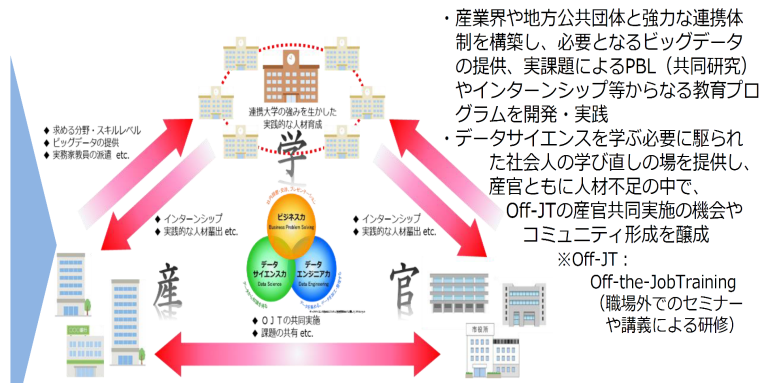
石油からデータの時代へ

- Alphabet (Google), Amazon, Apple, Facebook, Microsoft だけで 250 億ドルの純利益
- 20 世紀初頭は石油の時代 (Standard Oil, 1911 年 34 の会社に). 21 世紀はデータの時代
- 石油もデータもそれ自体だと価値がない。精製し加工して初めて本来の価値が出てくる。

超スマート社会の実現(文部科学省, 平成30年度)

背景

膨大なデータが溢れる次代において、**数理的思考やデータ解析・活用能力を持つ人**が戦略的にデータを扱うことによる経営等の影響はきわめて大きい。我が国が国際競争力を強化し、世界に先駆けて **Society5.0** を実現していくためには、データから新しい価値の創造を見出せる人材 (**データサイエンティスト**) の育成が急務



http://www.mext.go.jp/a_menu/koutou/kaikaku/miraikachisouzou/1403520.htm



内閣府 https://www8.cao.go.jp/cstp/society5_0/index.html

社会の変遷

- Society1.0: 狩猟採集社会。家族単位で狩猟をしながら移動生活を行っていた。 **飢えとの戦い**。
- Society2.0: 農耕社会。磨製石器が使われ、鋤や鍬などの農具を使い村を形成。新石器時代とも称され、紀元前 10000～7000 年ごろから始まった。 **食料の安定供給**
- Society3.0: 工業社会。蒸気機関の発明 (1769 年, ワット)。産業革命が始まる。農民が減少し、 **商工業従事者が急増**。とくに石炭の需要が増えたため鉱工業が盛んになった。資本家と労働者の対立が顕著に。
- Society4.0: 情報社会。インターネット, 携帯電話, スマートフォンの普及。 **情報が価値を持つようになる**。個人情報保護法 (2003 年)。
- Society5.0: **超スマート社会**。AI が頭脳, データが食料, センサーが口, IoT(Internet of Things) が神経ネットワーク, ロボットが筋肉

持続可能な開発目標 (Sustainable Development Goals)



国連 https://www.unic.or.jp/activities/economic_social_development/sustainable_development/2030agenda/

持続可能な開発目標

貧困や飢餓をなくし、質の高い教育を目指し、地球を保護し、平等、平和、豊かさを目指す普遍的な行動。国連開発計画の2030年に向けた行動指針。

「Society 5.0」を実現するとともに、これによりSDGsの達成に寄与する(未来投資戦略2018, 安倍内閣による成長戦略)

→ 但し実際はグローバルな競争が激しくなり、貧富の格差が広がっているのが現状

データサイエンス

データサイエンスとは？

データサイエンス

データサイエンスとは？

- データを科学(サイエンス)する, または調理する学問.
- 見えなかった(気づかなかった)ものを見えるようにする学問. 新たな価値の創生.
- ビッグデータ時代に急速に必要性が認識されてきた. 膨大なデータの中から価値のある情報を取り出すことが主な目的.
- データマイニング (data mining) もその1つ.
- データの不確実性(ランダム性)を扱う統計学と確率論と密接な関係がある.

データサイエンス

データサイエンスとは？

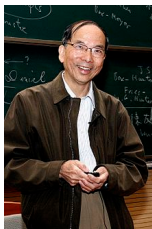
- データを科学 (サイエンス) する, または調理する学問.
- 見えなかった (気づかなかった) 物を見えるようにする学問. 新たな価値の創生.
- ビッグデータ時代に急速に必要性が認識されてきた. 膨大なデータの中から価値のある情報を取り出すことが主な目的.
- データマイニング (data mining) もその1つ.
- データの不確実性 (ランダム性) を扱う統計学と確率論と密接な関係がある.

データ分析とデータサイエンス [柴田里程, 2015]

データは現象の放つ光である

この光を適切に捉え, 背後にあるメカニズムを明らかにする. そのためには**数学**, **統計科学**, **機械学習**, **コンピュータ科学**を駆使する必要がある. もともとは1997年のC. F. Jeff Wu教授がミシガン大学での就任演説 (Statistics=Data Science?) が始まりだとされている.

C. F. Jeff Wu 教授



ジョージア工科大学教授. 専門は統計科学 (データサイエンス), 統計数理研究所の Akaike Memorial Lecture Award の最初の受賞者.

そもそも統計 (statistics) はラテン語の **status** に由来. つまり国の状態を調べ, 国民の幸福を増大させるためのもの (主に数を数えて表にまとめる作業). しかし昔の統計とは違い, 現代の統計の仕事は次の 3 つからなる.

- データ収集
- データモデリング・データ解析
- 問題解決, 問題の理解, 意思決定

Statistics(統計) ⇒ Statistical Science (統計科学) ⇒ **Data Science(データサイエンス)**

Big Data



<https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/>
そもそも従来の統計学はビッグデータを扱っていなかった。新しい解析手法 (AI 技術, 機械学習, 深層学習), 解析環境 (スパコン, 大型計算機) が必須になった。

日本でのデータサイエンス I

データサイエンスはこれからますます重要になってくる。データサイエンスとは？

滋賀大学作成ビデオ：<https://www.youtube.com/watch?v=J60nT185sio>

キーワード

- ビッグデータ
- 分析, モデリング, 予測
- 料理 (素材, いろんなレシピ, 一工夫)
- データから価値を創り出す
- データを学習する.

さまざまな分野に活用されている。ビジネス, マーケティング, 天文学, 医療, 生物学, ロボット, 自動車産業, ...

データサイエンスで必要なこと

- 理論的な思考力 ⇒ 数学
- データを扱える能力, 情報処理 ⇒ コンピュータサイエンス, 計算機科学
- 探究心, データと対話する力 ⇒ 真摯な研究姿勢
- データに内在する不確実性を捕らえる ⇒ 統計学, 確率論
- いろいろな解析手法 ⇒ 統計学

日本でのデータサイエンス II

どこで学べるの？

- 統計数理研究所：統計数理の日本の拠点（立川市），大学院教育もしている。
<http://www.ism.ac.jp/>
- 滋賀大学：日本で最初のデータサイエンス学部 (2017)
<https://www.ds.shiga-u.ac.jp/>
- 横浜市立大学：岩崎学先生 (以前成蹊大学に所属), データサイエンス学部 (2018)
<https://www.yokohama-cu.ac.jp/academics/ds/index.html>
- 武蔵野大学：データサイエンス学部 (2019)
https://www.musashino-u.ac.jp/academics/faculty/data_science/
- 成蹊大学：2020年にデータ数理コース新設。2022年にはデータサイエンス副専攻を新設予定。文系理系だれでも受講可能。
<https://tinyurl.com/ycvtsg52>

その他にも山形大学，武蔵大学，広島大学，京都産業大学などにもデータサイエンスコースがある。

海外のデータサイエンス教育プログラム数

教育プログラム数（国別・学位別）2019年5月

	US	GB	IE	ES	NL	FR	—	Total
Bachelors	46	5	1	1	0	1	4	59
Masters	301	40	8	8	7	10	46	420
Doctorate	19	1	1	0	0	0	2	23
Certificate	100	0	1	0	0	0	1	102
Total	466	46	11	9	7	11	53	603

北川源四郎（6 大学コンソーシアムを中心とした数理・データサイエンス教育強化の取り組み）. データソース

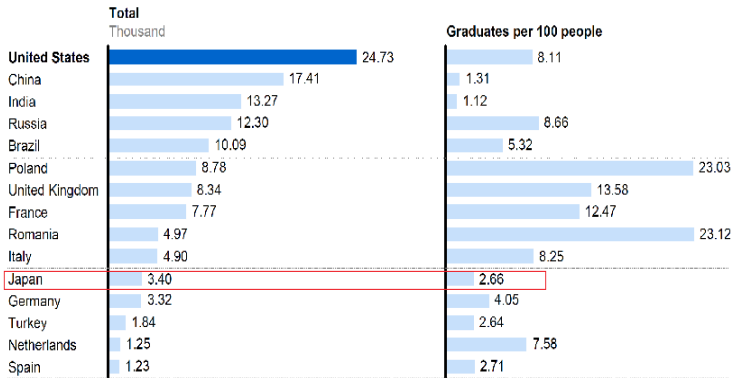
<http://datascience.community/colleges>

海外のデータサイエンスプログラムの数は年々増加傾向

185(2014/4) ⇒ 279(2015/7) ⇒ 505(2016/2) ⇒ 603(2019/5)

US: United States, GB: Great Britain, IE: Ireland, ES: Spain, NL: Netherlands, FR: France

データサイエンティストの人材不足



Bit data: The next frontier for innovation, competition, and productivity (McKinsey&Company)

These data count new graduates, i.e., a flow of deep analytical talent, which we define as people with advanced training in **statistics and/or machine learning** and who conduct data analysis.

2008 年時点で日本は $3.40 \times 1000 = 3400$ 名程度の卒業生しか輩出していない！

数理・データサイエンス教育強化拠点コンソーシアム

数理・データサイエンス教育強化拠点コンソーシアム

平成28年に文部科学省より数理及びデータサイエンスに係る教育強化の拠点校として、北海道大学、東京大学、滋賀大学、京都大学、大阪大学、九州大学が選定される。拠点校6校は、数理・データサイエンスを中心とした全学的・組織的な教育を行うセンターを整備して、各大学内での数理・データサイエンス教育の充実に努めるだけでなく、全国の大学に取組成果の波及を図るため、地域や分野における拠点として他大学の数理・データサイエンス教育の強化に貢献することが期待されている

<http://www.mi.u-tokyo.ac.jp/consortium/>

- 全国的なモデルとなる標準カリキュラム・教材の作成
- その標準カリキュラム・教材の他大学への普及方策（例えば全国的なシンポジウムの開催等）の検討及び実施
- センターの情報交換等を行うための対話の場の設定（各大学のセンターにおける教育内容・教育方法の好事例を共有し、より取組を発展させるための議論など）
- センターの取組の成果指標や評価方法の検討

協力校 20 校

北見工業大学、東北大学、山形大学、筑波大学、宇都宮大学、群馬大学、千葉大学、お茶の水女子大学、新潟大学、長岡技術科学大学、静岡大学、名古屋大学、豊橋技術科学大学、神戸大学、島根大学、岡山大学、広島大学、愛媛大学、宮崎大学、琉球大学

統計検定

合格者の声

1級合格
この検定が、ビジネス界の一つの指針となればと思っています！

2級合格
データを扱うセンスを磨くことができたのが私にとって一番の収穫でした。

3級合格
今までの業務知識を再確認するいい機会になったと思います。

4級合格
これからはもっと統計の知識を深め世の中の情報を正しく判断していきたいです。

受験のきっかけや勉強方法、合格時の感想などを掲載しています。

<http://www.toukei-kentei.jp/>

統計検定

統計に関する知識と活用力を評価する全国統一試験。 2015年から実施。英国の王立統計学会 (The Royal Statistical Society, RSS) の統計検定を模範としている。1級から4級まであり、中学生、高校生、大学生、社会人に必要なスキルを評価。日本統計学会が運営母体。

全国の会場で実施。紙媒体を使った試験方式と CBT(Computer Based Testing) を実施 (ただし CBT は2級まで)。日本統計学会公式認定テキストや、過去問も販売しています。

<http://www.toukei-kentei.jp/info/books/>

データサイエンスは魅力的な仕事？



2012年に Harvard Business Review にデータサイエンティストに関する記事が出版された。その中で、Google のチーフエコノミストの Hal Varian さんのコメントでは

The sexy job in the next 10 years will be **statisticians**. People think I'm joking, but who would've guessed that **computer engineers** would've been the sexy job of the 1990s?

Google や Facebook などでは年に数千万円を稼ぐ人も、様々なところでデータサイエンスに関する講習会が開かれている。データサイエンス・ブートキャンプ

<https://exploratory.io/training-jp>

3日間の内容で受講料 19 万 8 千円ほど

しかし膨大なデータを扱うには地道な作業も多いことは世間ではあまり認知されていない。

発表内容

1 データサイエンス

2 統計科学の歴史と実例

3 データサイエンスの最前線（人工知能，AI）

近世実証科学

「自然という書物は我々の前に開かれている。それは我々のアルファベットとは違った文字で書かれている。その文字は三角形や角や円や球である。」 (Galileo Galilei, 1564-1642)

- Galileo Galilei: 質量によって落下速度は変わらないことを実証「落下の法則」。緻密な観測によりアリストテレスの自然哲学体系を覆す。
- Johannes Kepler: Tycho Brahe が残した膨大で精密な天体観測データをもとに、惑星の運動は楕円運動であることを発見「Kepler の惑星運動法則」。地動説を確立。天動説を覆す。
- Isaac Newton: Kepler の惑星運動法則を力学的に解明「万有引力の法則」。Newton 力学、微積分法を創始。

「現代の自然科学は自然（データ）を緻密に観察し、規則性を発見することから始まった」

データリテラシー

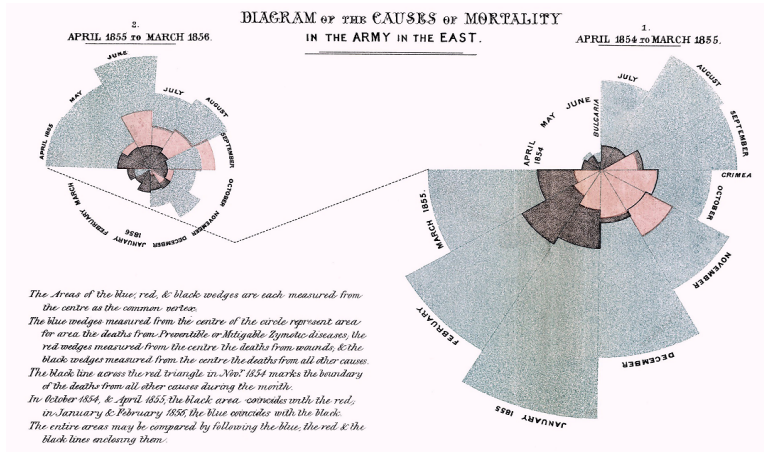
フローレンス・ナイチンゲールの言葉

神の御心を知るには統計学を学ばなくてはならない。

イギリスの看護師、近代看護教育の生みの親。クリミア戦争(1853年10月1856年3月フランス、オスマン帝国(トルコ)、イギリス、サルデーニャ(イタリア) v.s. ロシア)に看護師として従軍し、兵士の死亡データを集計・分析した。その結果をもとにそれまで劣悪だった戦地の病院の環境を改善したところ、負傷した兵士の死亡率が大きく下がった。つまり病院内での死亡は負傷が主な原因ではなく病院内の不衛生(感染症)が原因だった。もともとは裕福な家庭で育ち、さまざまな学問を修得している教養人。ギリシア哲学にも深い造詣がある。1859年にイギリス王立統計学会の初の女性メンバーに選ばれた。「犠牲なき献身こそ真の奉仕」を基本的な考えとしていた(マザー・テレサと同じ)。37歳のときに心臓発作で倒れ、その後50年近くはベッドの上で過ごした。イギリスの英雄の一人。旧10ポンド札の裏側はフローレンス・ナイチンゲール(新10ポンド札の裏側はチャールズ・ダーウィン)。10ポンド札の表はエリザベス女王。結局クリミア戦争はロシアが敗北。産業革命(1760年ごろ~1830年ごろ)によってイギリス、フランスの方が国力が上回っていた。



フローレンス・ナイチンゲール
(1820年~1910年)



Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army

青の面積が眠り病 (Zymotic disease) による死者数。赤が戦争の負傷による死者数。黒がその他の死者数。眠り病 (感染症) が死因の主な原因。また負傷兵が夜間に多く死亡していることに気づき、夜間の巡回介護を始めて行った！

Galton(1822-1911)による相関のデータ

上腕の長さ y(インチ)	身長 x(インチ)																
	59-60	60-61	61-62	62-63	63-64	64-65	65-66	66-67	67-68	68-69	69-70	70-71	71-72	72-73	73-74	74-75	
21.0-21.5																1	1
20.5-21.0													1	1			2
20.0-20.5														1			1
19.5-20.0									2				1		2		5
19.0-19.5									2	4	6	11	8	4	2	1	38
18.5-19.0					1		2	6	8	7	15	13	2	1			55
18.0-18.5						3	8	15	28	14	25	5	2	2			102
17.5-18.0				2	1	2	12	18	15	7	2	1	1				61
17.0-17.5			1	3	6	11	10	7	7	3	1						49
16.5-17.0			1	5	6	5	4	1	1	1	1						25
16.0-16.5	1	1	1	3	2												8
15.5-16.0		1															1
	1	2	3	13	16	21	36	47	61	38	50	30	15	9	4	2	348

Regression towards mediocrity in hereditary stature. Galton 1886

相関係数]

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

はじめて「相関」という考え方が示された。

Karl Pearson(1851-1936)



- Francis Galton の弟子. イギリスの数理統計学者
- 統計学は科学の文法「The grammar of science」. アインシュタイン, 夏目漱石, 寺田寅彦に影響を与える.
- 相関係数, カイ二乗適合度検定, ヒストグラム, 標準偏差
- 母集団という概念の提示
- その後 R. A. Fisher により母集団に対する推定の理論と仮説検定の基礎が構築される.

R. A. Fisher(1890-1962)



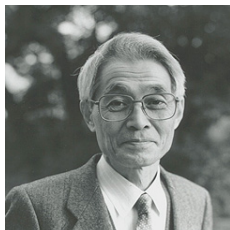
- 近代統計学（統計的推測）の創設者，頻度論者，ベイズ流の恣意性（パラメータに対して確率構造を考える）を強く批判.
- 「仮説的無限母集団」を想定し，母集団と標本の違いを明確化.
- 推定方式の良さの基準として一致性，有効性，十分性の概念を考案し，それらをすべて満たす**最尤法を確立した**（ただし不偏性は保証されない）.
- 良い標本を得るためにはどうするか？⇒実験計画法構築される.

田口玄一 (1924-2012)



- 品質管理工学の創始者。設計段階でバラツキが小さくなることを重視。そのあと目標値に近づけることを行う（2段階設計法）。制御因子（材質，加工方法，長さ，角度，重さ，厚さなど）と誤差因子（設計の段階でうまく制御できないもの。ノイズ。温度，湿度，振動など）。
- バラツキの指標として SN 比 (signal to noise ratio) を用いる。実験の組み合わせを減らすため直交表を用いる。EDA(Exploratory Data Analysis) の先駆け。統計学は実学。はじめアメリカの産業界で業績が評価。

赤池弘次 (1927-2009)



- 知識・経験・常識（直感）を客観的に評価できる尺度を提案した (Akaike Information Criterion). 2006年京都賞受賞 (基礎科学部門)
- 現実世界 (自然界, 産業界, 人間社会) と数理の世界 (モデル) の結びつきの重要性を説かれた. 統計数理研究所の基本理念.
- 常に現場の方たちと関わり, 実際の問題の解決に尽力された (船舶の制御, セメント工場の制御など). <https://www.ism.ac.jp/akaikememorial/>

「現場の人達が直面している問題, 望んでいることを的確に把握し, 実際に役に立つ仕事をしなさい. やっぱり統計に携わる者は実際の問題に触れなきゃいかん。」 (赤池弘次, 統計科学を語る)

数理統計学

「確率モデルを想定することによりデータに潜在する構造を探索」

- 母集団の概念

- ▶ 数字のカラクリ・データの真実 (NHK) 「心筋梗塞で死亡した人の“95%”が、この食べ物を摂取していた」。この食べ物は禁止すべきか？

- 確率変数と確率分布

- ▶ 誤差を記述する分布，稀少な事象を記述する分布。

- 検定

- ▶ プロ野球のホームラン数の増加は偶然か否か (NHK)？

- 回帰分析

- ▶ 身長から体重を予測するには？売上に影響しているのは値段，品揃え，立地条件？

機械学習（パターン認識）

パターン認識とは「計算機アルゴリズムを使い、膨大なデータに潜在する**規則性**（モデル）をうまく**学習**すること」。機械学習の枠組みに含まれ、統計科学の新しい一分野として位置づけられている。

- 人間が五感（視覚，聴覚，触覚，嗅覚，味覚）を使い行っている複雑な情報処理をコンピュータを使い工学的に実現することを目指す。文字認識、音声認識、顔認識など。
- 現在では医療，心理，自然言語処理など幅広い応用範囲を持つ。
- 判別に重要な情報抽出（特徴抽出）と判別の2段階からなる学習方法。

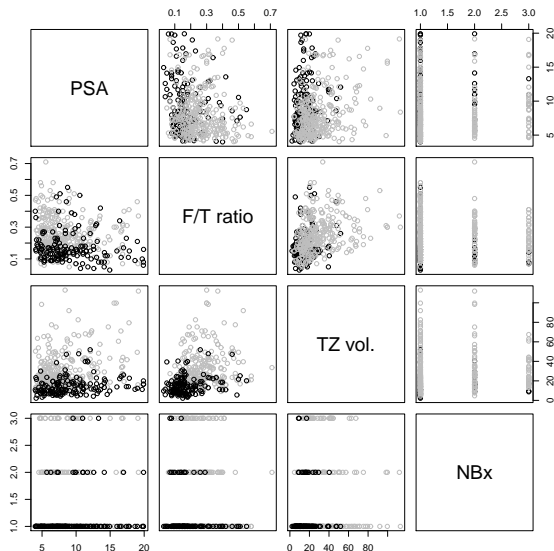
病院の腫瘍のデータ (実例)

- 灰色: 良性の腺腫瘍
- 黒色: 悪性の腺腫瘍

「機械学習の手法であるブー
スティングを用い、**良性と
悪性の腫瘍を特徴づける規
則性を抽出したい**」

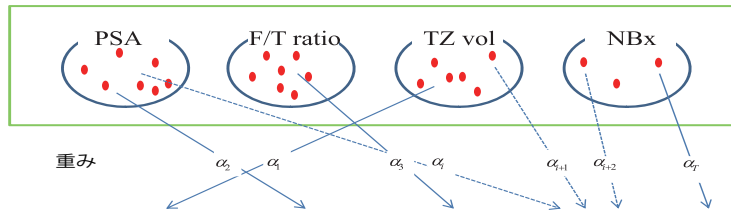


「ROC の下側面積 (AUC) 注
目し、それを最大化する
ブーシングの手法を提
案 [Komori, 2011]」



AUCBoost

4つの弱判別機の集合

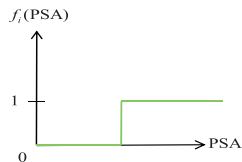


$$F(\mathbf{x}) = \alpha_1 f_1(\text{TZ vol}) + \alpha_2 f_2(\text{PSA}) + \alpha_3 f_3(\text{F/T ratio}) + \dots + \alpha_T f_T(\text{NBx})$$
$$= F(\text{PSA}) + F(\text{F/T ratio}) + F(\text{TZ vol}) + F(\text{NBx})$$



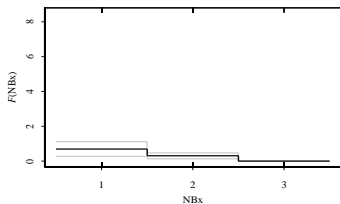
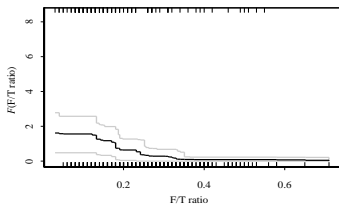
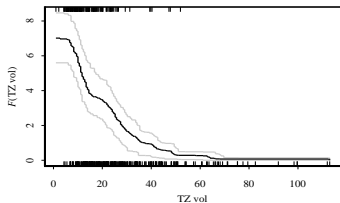
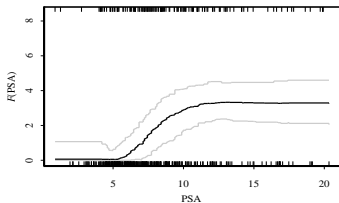
診断

$$\left\{ \begin{array}{l} F(\mathbf{x}) \geq c \Rightarrow \text{悪性腫瘍} \\ F(\mathbf{x}) < c \Rightarrow \text{良性腫瘍} \end{array} \right.$$



- PSA, F/T ratio, TZ vol と NBx を使った AUCBoost.

4つのマーカーのスコアプロット



- 4つのマーカーがどのように判別に効いているかを読み取ることができる。

PSA のノモグラム [Kanao et al., 2014]

Table 2. Individualized PSA cut-off values for men aged ≥ 70 years old with normal DRE ($\leq 5\%$ risk of missing intermediate/high-risk cancer)

%fPSA	TZ volume (cc)										
	<12.0	12.0–15.9	16.0–19.9	20.0–23.9	24.0–27.9	28.0–31.9	32.0–35.9	36.0–39.9	40.0–43.9	44.0–47.9	≥ 48.0
<8	4 (4-4)	4 (4-4)	4 (4-4)	4 (4-4)	4.2 (4-6.8)	4.6 (4-7.6)	4.9 (4-8.1)	5.2 (4-8.4)	5.4 (4-9.4)	5.9 (4-9.5)	6.1 (4-9.5)
8–11	4 (4-4)	4 (4-4)	4 (4-4)	4 (4-4)	4.3 (4-6.8)	4.8 (4-7.6)	5.2 (4-8.4)	5.7 (4-8.8)	6 (4-9.4)	6.5 (4-9.5)	6.6 (4-9.8)
12–15	4 (4-4)	4 (4-4)	4 (4-4)	4 (4-4)	4.3 (4-7)	4.9 (4-7.6)	5.3 (4-8.4)	5.9 (4-8.8)	6.1 (4-9.4)	6.6 (4-9.6)	6.7 (4-10)
16–19	4 (4-4)	4 (4-4)	4 (4-4)	4.1 (4-5.6)	4.7 (4-7.4)	5.7 (4-8.2)	6.4 (4-8.8)	6.9 (4-9.5)	7.2 (4-9.6)	7.7 (4-10.2)	7.8 (4-11)
20–23	4 (4-4)	4 (4-4)	4.1 (4-5.9)	4.5 (4-6.8)	6.1 (4-7.9)	7.5 (4-9.5)	8 (4-10)	8.6 (6.4-11.5)	9 (6.4-20)	10 (7.3-20)	10.8 (7.3-20)
24–27	4 (4-4)	4 (4-4)	4.3 (4-6.7)	5.2 (4-7.4)	7 (4-8.8)	8.4 (4-10.8)	9.2 (7.3-20)	10.4 (7.4-20)	11 (7.5-20)	12.2 (7.7-20)	12.9 (8-20)
28–31	4 (4-4)	4.1 (4-4)	4.4 (4-6.8)	5.3 (4-7.4)	7.2 (4-8.8)	8.7 (7.1-12.4)	9.7 (7.4-20)	11 (7.6-20)	11.6 (7.7-20)	12.9 (8.1-20)	13.6 (8.4-20)
32–35	4 (4-4)	4.1 (4-5.9)	4.5 (4-7.1)	5.6 (4-7.8)	7.4 (4-9.5)	8.9 (7.1-16.9)	10.1 (7.4-20)	11.3 (7.7-20)	12 (7.9-20)	13.5 (8.4-20)	14.1 (8.4-20)
36–39	4 (4-4)	4.1 (4-6.3)	4.7 (4-7.3)	5.8 (4-8.1)	7.8 (4-10)	9.4 (7.1-20)	10.6 (7.4-20)	12.1 (7.9-20)	12.8 (8-20)	14.6 (8.4-20)	15 (8.4-20)
≥ 40	4 (4-4)	4.1 (4-6.3)	4.7 (4-7.3)	5.8 (4-8.3)	7.8 (4-10)	9.5 (7.1-20)	10.7 (7.4-20)	12.2 (7.9-20)	13 (8-20)	14.6 (8.4-20)	15.1 (8.4-20)

● 患者ごとに最適な PSA の閾値を導出. 括弧内は 95% 信頼区間

データ解析の戦略 I

情報とは (塩野七生, ローマ亡き後の地中海世界)

情報とは、量が多ければそれをもとにして下す判断もより正確度が増すとはまったくの誤解である。情報はたとえ与えられる量が少なくともその意味を素早く正確に読み取る能力を持った人の手に渡ったときに初めて活きる。

- つまりまずはデータをよく見ることが大事 (ブラウジング, browsing).
- サンプルング手法が使えると、ビッグデータの取り扱いも楽になる。
- 不要なデータを除き、解析しやすいように整える (クリーニング, cleaning またはクレンジング, cleansing).
- データの整備が非常に大変。整備をしながらデータの特徴を捉えることができると、その後の解析方針が決めやすい。

データ解析の戦略

データサイエンスの実践においては、いつものような切り口で解析をするのか、どこまで掘り下げるのかといった戦略 (ストラテジー) をよく練り直しながら進める必要がある。具体的な目標を定め、**さまざまな角度からデータを無心に眺め廻すことが欠かせない。**

哺乳類の体と脳の重さのデータ I



Asian elephant



Rhesus monkey(赤毛猿)



Kangaroo



lesser short tail shrew



Water opossum



Musk shrew



Owl monkey



Golden hamster



Ground squirrel



Pig

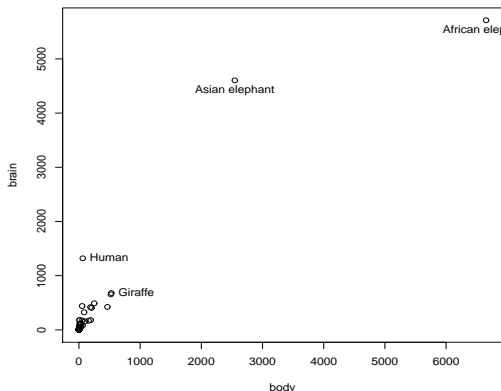


Giraffe



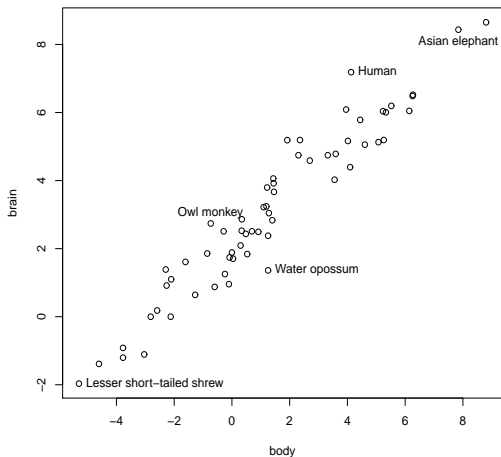
Donkey

哺乳類の体と脳の重さのデータ II



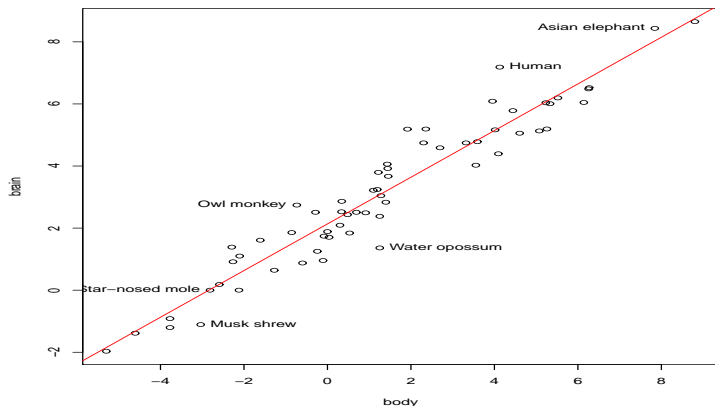
- 25 個体の陸上動物のデータ (cow, wolf, goat, pig, horse, monkey,...). 統計ソフトウェア R の MASS package から引用
- 横軸：体重 (kg), 縦軸：脳の重さ (g)
- 一見体重と脳の重さには関係性は見られない。

哺乳類の体と脳の重さのデータ III



脳と体の重さの散布図 (対数変換後)

哺乳類の体と脳の重さのデータ IV



Human, Owl(フクロウ) は上に, Musk shrew(トガリネズミ), Water opossum(ネズミの一種) は下に予測がずれている。Star-nosed mole(鼻が星の形をしたモグラ) はほぼ予測が的中

推定されたモデル I

結局推定されたモデルは

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= 2.13 + 0.75x_i\end{aligned}$$

ただし $\hat{y}_i = \log(\text{脳の重さ}_i)$, $x_i = \log(\text{体の重さ}_i)$ であった。よって元のスケールに戻すと

$$\begin{aligned}\text{脳の重さ}_i &= \exp(2.13) \text{体の重さ}_i^{0.75} \\ &= 8.45 \times \text{体の重さ}_i^{0.75}\end{aligned}$$

の関係式が哺乳類の動物に関し成り立つことが分かった。一般にこのような関係式 ($y = ax^b$) は、生物の2つの部分 (x と y) の間でよく成立する関係式であり、**アロメトリー (allometry) の関係式**と呼ぶ。陸上哺乳類において、心臓重量と体重の関係では $b = 1.0$, 骨重量と体重の関係では $b = 1.09$, 心拍数と体重の関係では $b = -0.25$ などが知られている [八木光晴 and 及川信, 2008].

武蔵野市人口データ I

武蔵野市の人口データを使う。

<https://tinyurl.com/ycwqr4mw>

武蔵野市の 51 の地域の世帯数と、男女それぞれの年齢ごとの人口のデータである (平成 29 年 10 月 1 日現在).

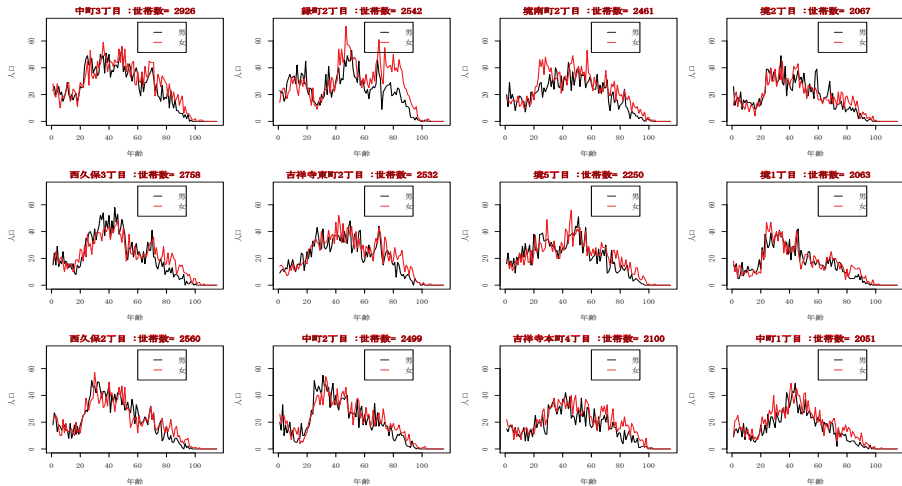
- データにはどのような特徴があるか？
- 世帯数が一番多い地域は？
- 特徴的な地域はどこ？

統計解析ソフトが有用. データサイエンティストの大工道具. シェフの調理器具.

統計ソフト R

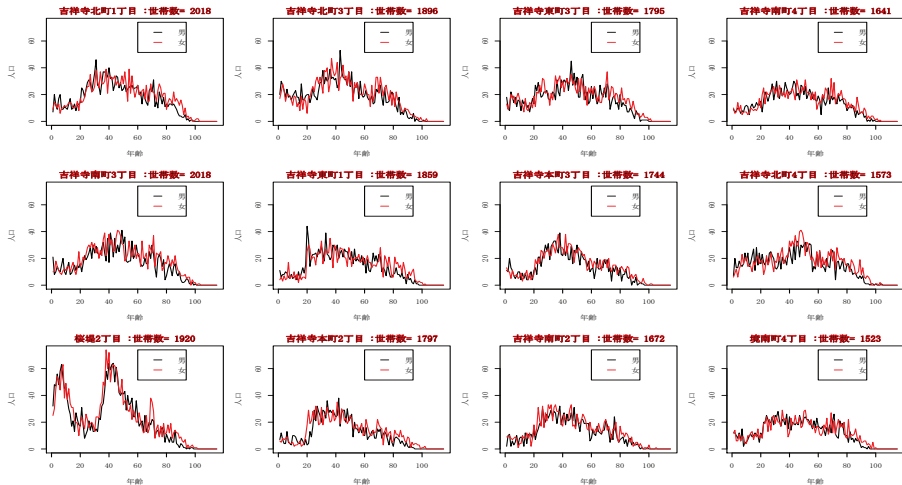
```
>source("jinkou_plot.R")
>ls()
[1] "jinkou_plot"
>jinkou_plot()
```

武蔵野市人口データ II



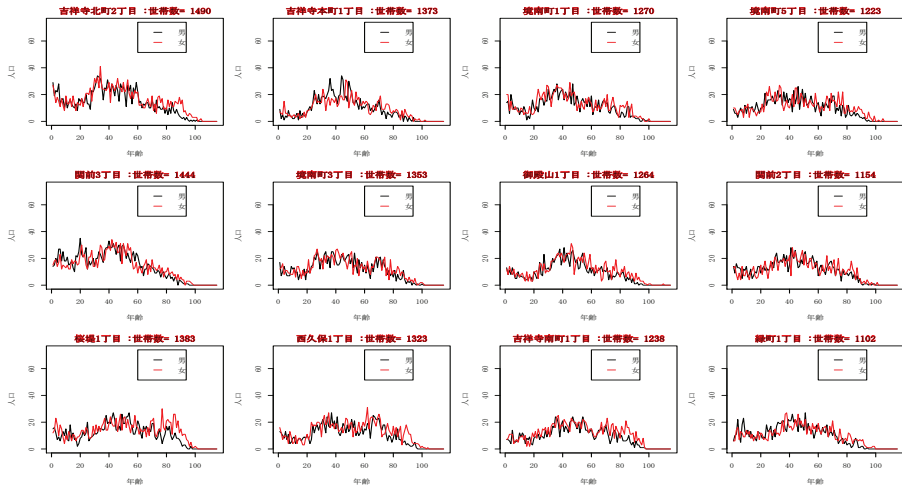
年齢と人口のグラフ

武蔵野市人口データ III



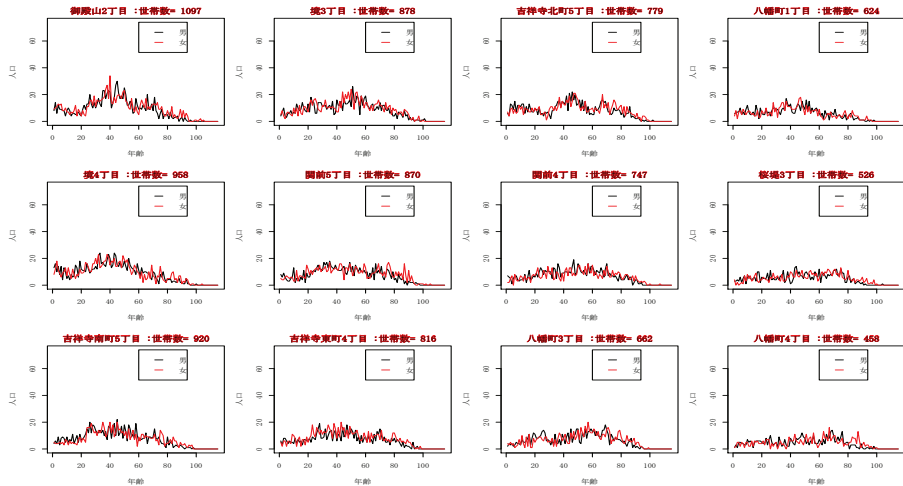
年齢と人口のグラフ

武蔵野市人口データ IV



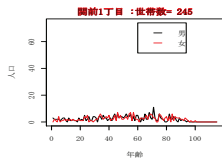
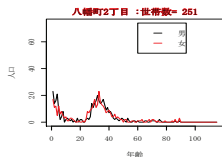
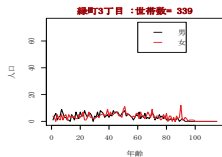
年齢と人口のグラフ

武蔵野市人口データ V



年齢と人口のグラフ

武蔵野市人口データ VI



年齢と人口のグラフ

東京都の地図の描画 I

日本の地図情報を使ってコロプレス図 (階級区分図) を描いてみる。コロプレス図とは人口などの統計データに合わせて地域を塗り分けた地図のこと

統計ソフト R

```
>source("choropleth.R")
>ls()
[1] "choropleth"
>choropleth()
```

プログラムを少し修正すると、他の県の地図も描ける。

プログラムの修正

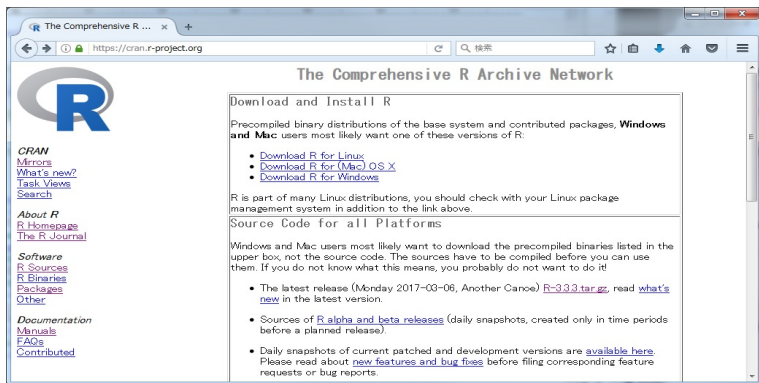
```
>fix(choropleth)
```

東京都の地図の描画 II

choropleth 関数

```
function(){
library(maptools)
  jpn=readShapePoly("japan_ver81.shp")
  region=jpn[662:714,] #東京都の島は除く
  #region=jpn[jpn$KEN=="福井県",]
  plot(region,border="red")
  jinko=region$P_NUM# 市または町ごとの東京の人口データ
  class.num=10# 色分けの数
  classes=cut(jinko,seq(min(jinko),max(jinko),length=class.num+1),
  include.lowest=T)# 地域を人口の数で 10 に分ける
  cols=rev(heat.colors(class.num))# 白から赤で色分け
  plot(region,border="red",col=cols[classes])
  box()# 枠を付ける
  loc.x=NULL
  n=53
  #n=sum(jpn$KEN=="福井県")
  for(i in 1:n)loc.x=c(loc.x,region[7]@polygons[[i]]@labpt[1])
  loc.y=NULL
  for(i in 1:n)loc.y=c(loc.y,region[7]@polygons[[i]]@labpt[2])
  text(loc.x,loc.y,paste(region[[6]]),cex=0.5)
}
```


統計解析ソフト R とは？



- 「探索的データ解析」(Exploratory Data Analysis: EDA) を実践するために作られた対話型計算機環境。グラフィックスの機能も充実している。EDA は 1970 年代に John Wilder Tukey 博士によって初めて提唱された。 <https://cran.r-project.org/>
- 前身は AT&T(American Telephone & Telegraph Company) のベル研究所で開発された S である。



- 1925年設立. 電気通信基礎技術に関する研究が盛ん. 1940年代まではニューヨーク市内に本拠地があったが, 今現在ニューヨーク郊外のニュージャージー州に移転. さまざまな革新的技術を開発してきた (電波望遠鏡, トランジスタ, 情報理論, C, C++言語). ノーベル物理学賞受賞者を数多く輩出.

探索的データ解析 (EDA)



(Tukey 博士)



- 探索的データ解析は1970年代に John Wilder Tukey 博士によって初めて提唱された。データと対話（知的やり取り）をしながらデータの背後にあるメカニズム（本質）を探り出すための能動的解析方法（ \Leftrightarrow 記述統計）。Bell Labs と Princeton University に勤務。データサイエンスの提唱者 W.S. Cleveland も Tukey の弟子の一人。

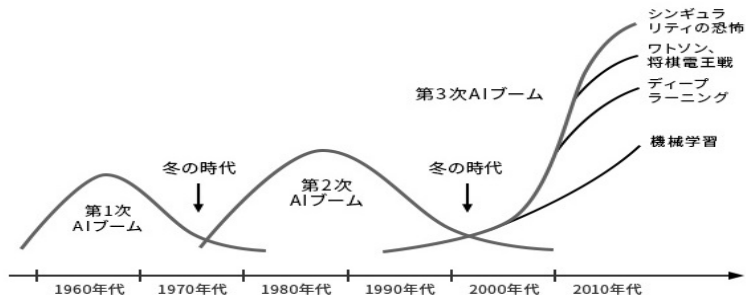
発表内容

1 データサイエンス

2 統計科学の歴史と実例

3 データサイエンスの最前線（人工知能，AI）

人工知能とは？

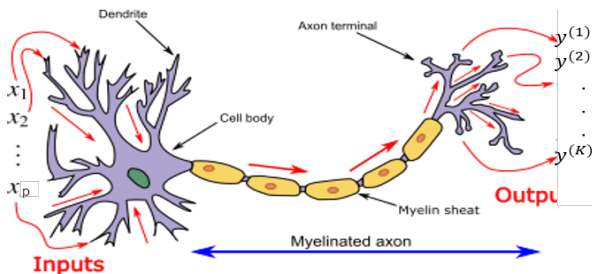


引用：「人工知能は人間を超えるか ディープラーニングの先にあるもの」松尾豊 著

人工知能とは

- 1956年のダートマス会議で初めて提唱。
- 第二次ブームでは膨大な専門知識を利用したエキスパートシステムが注目される。但し膨大な知識を収集し、定式化するのが大変だった (冬の時代)
- **特徴量を学習できる深層学習法 (機械学習)** が提案され、再度注目される。
- 人によって定義がまちまち。「人工的につくった知能をもつ実体」, 「人間と区別がつかない人工的な知能」, 「知能, 心をもつ機械」, 「人間の脳をシミュレートするシステム」

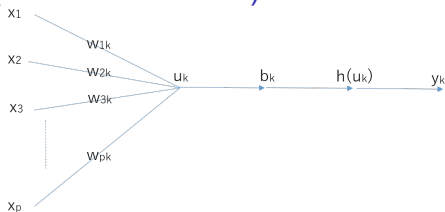
神経細胞 (neuron) I



<https://ja.wikipedia.org/wiki/神経細胞>

- Dendrite(樹状突起)：複数の信号 (input, x_1, \dots, x_p) を受け取る場所. 信号の重み付けもここで行われる.
 - Cell body(細胞本体)：Soma とも呼ばれる. 樹状突起で受け取った複数の信号を1つにまとめる.
 - Axon(軸索)：軸索の末端でシナプスが形成され, 神経細胞同士の信号のやり取りが行われる. つまり活性化された信号のみ $\text{output}(y^{(1)}, \dots, y^{(K)})$ として次の神経細胞に伝達される.
- シナプス：神経細胞間の信号のやり取りをする部位またはその構造.

人工神経細胞 (artificial neuron) I



基本構造

$$y_k = h(u_k), \quad u_k = \sum_{j=1}^p w_{jk} x_j + b_k, \quad k = 1, \dots, K$$

$h(u_k)$: 活性化関数, u_k : neuron, b_k : バイアス, K : ユニット数 (neuron の数), p : 入力数

- 階段関数 $h(u_k) = H(u_k)$: パーセプトロン ($k = 1, 2$)
- 恒等関数 $h(u_k) = u_k$: 線形モデル, 主成分分析
- softmax 関数 $h(u_k) = \exp(u_k) / \sum_{k=1}^K \exp(u_k)$: ロジスティック回帰モデル
- ReLU(Rectified Linear Unit) 関数 $h(u_k) = \max(0, u_k)$: 深層学習

2024年ノーベル物理学賞 I



<https://ja.wikipedia.org/wiki/ジョン・ホップフィールド>

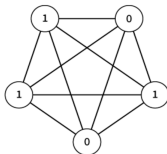


<https://ja.wikipedia.org/wiki/ジェフリー・ヒントン>

ジョン・ホップフィールド氏（アメリカ・プリンストン大学名誉教授）と、ジェフリー・ヒントン氏（カナダ・トロント大学名誉教授）がノーベル物理学賞を受賞した。ホップフィールドは神経細胞同士の結合強度を適切に設定することで、脳の「連想記憶」の仕組みを数理的に実現。ヒントンは入力パターン（画像）を学習し、類似したパターンを「生成」するボルツマンマシンを提案。どちらも人工知能の研究分野で重要な貢献をした。

ホップフィールド・ネットワーク I

ホップフィールド・ネットワーク



https://www.aist.go.jp/aist_j/magazine/20241204.html

- N 個のニューロンに n 個のパターン (状態) を記憶させる $V_i^s \in \{0, 1\}$, ($i = 1, \dots, N, s = 1, \dots, n$).
- ニューロン同士の結合強度 (重み) を以下の式で計算する

$$T_{ij} = \sum_{s=1}^n (2V_i^s - 1)(2V_j^s - 1), \quad i, j = 1, \dots, N$$

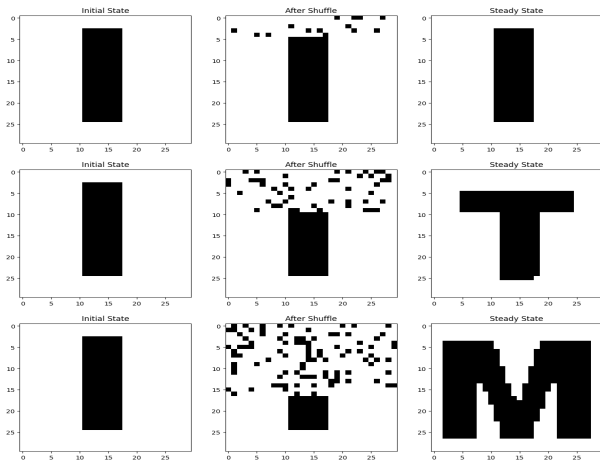
- 新しい (ノイズの入った) パターン V_i , ($i = 1, \dots, N$) に対し

$$V_i \leftarrow \begin{cases} 1 & \sum_{j \neq i} T_{ij} V_j > 0 \\ 0 & \sum_{j \neq i} T_{ij} V_j < 0 \end{cases}$$

をエネルギー関数 $-1/2 \sum_{i \neq j} T_{ij} V_i V_j$ が最小化 (定常化) するまで繰り返す。

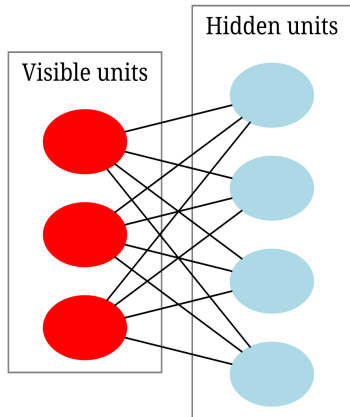
ホップフィールド・ネットワーク II

MITAKA の最初の 3 文字「M, I, T」をニューロンに記憶させると。



制約ボルツマンマシン I

ボルツマンマシンのニューロン同士の結びつきを簡素化したもの。可視層のニューロン v_i と隠れ層のニューロン h_j のみが結合している。



https://en.wikipedia.org/wiki/Restricted_Boltzmann_machine

制約ボルツマンマシン II

ニューロン同士の結合強度 (重みを) w_{ij} とすると, エネルギー関数は

$$E(\mathbf{v}, \mathbf{h}) = - \sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j$$

となり, 可視層のニューロンに対する確率分布に以下のボルツマン分布 (ギブス分布) を仮定する.

$$P(\mathbf{v}) = \frac{\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))}{Z}$$

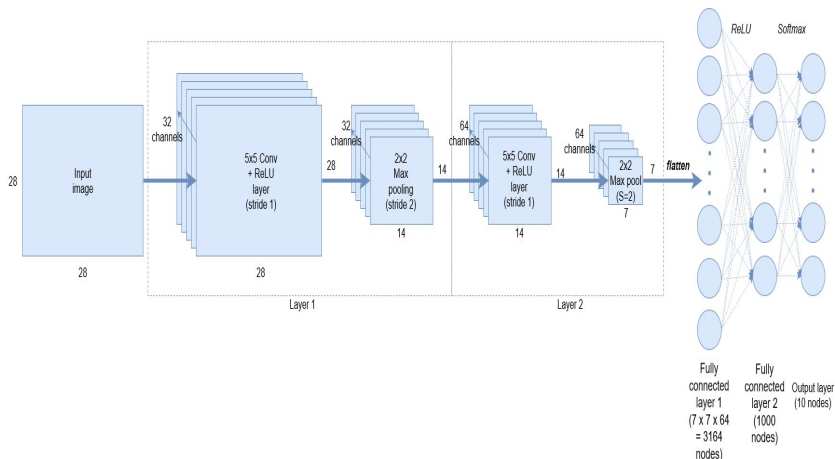
但し $Z = \sum_{\mathbf{v}, \mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$. パラメータ (a_i, b_j, w_{ij}) の推定は尤度

$$\prod_{\mathbf{v}} P(\mathbf{v})$$

を最大化することによって求める (Fisher の最尤法!).

畳み込みニューラルネットワーク (CNN) I

通常のニューラルネットワーク (NN) を拡張したものに、畳み込みニューラルネットワークがある (Convolutional Neural Network).

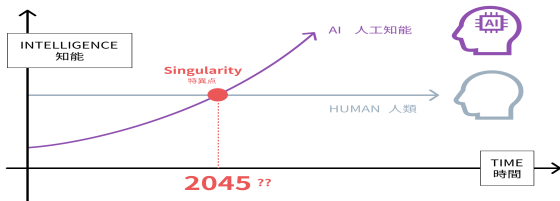


<https://medium.com/@bian0628/image-classification-cifar-10-dc1c23db46d5>

畳み込みニューラルネットワーク (CNN) II

- 畳み込み層では、**カーネル (kernel)** を使い画像の局所的なパターンを学習する。
- このカーネルが推定するパラメータとなる。
- Keras では `kernel_size` で高さと幅を指定する。
- カーネルによって作成される 2 次元テンソルを**特徴量 (フィルタ)** と呼ぶ。
- 特徴量を集めたものを**特徴量マップ** と呼ぶ (3 次元テンソル)。
- **特徴量 (フィルタ)** の数をフィルタ数と呼び、Keras では `filters` で任意の値を指定できる。
- 特徴マップのチャンネル数はフィルタ数と呼ぶ。
- Mnist の入力画像データはフィルタ数 1 (白黒画像) の特徴量マップである。
- カラー画像の場合は、画像データはフィルタ数 3 (RGB) の特徴量マップとなる。
- 1 枚の入力画像からたくさんの特徴量が生成されるのが、畳み込みニューラルネットワークの特徴。

技術的特異点



引用：「社会の変化に合わせてニーズが多様化する現代において、今、大人がまなび続ける理由」イベントレポート後編

技術的特異点

- 人工知能 (AI) が人類に代わって文明の進歩の主役となる地点。
- レイ・カーツワイルが提唱。「1000ドルのコンピュータで全ての人間の英知を凌ぐ」。2045年ごろとされる。
- 自ら考え、自ら行動する AI が誕生。自己改善サイクルの発生。AI と人類の融合 (人類のサイボーグ化)
- どのような事態が起こるのか、だれも予測できない。

Which Mindset Are You Using?

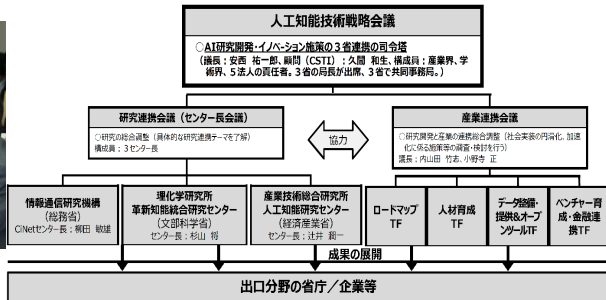
Incremental Mindset – 10%	Exponential Mindset – 10x
Set realistic goals	Set ambitious goals
Follow the plan	Follow the vision
Minimize risk	Maximize learning
Standardize	Personalize
Centralize decision-making	Empower decision-making
Expand authority	Expand influence
Make your numbers	Grow your network

Source: Mark Bonchek, Shift Thinking

壮大なビジョンを持ち、従来の常識や発想に捕らわれない革新的で野心的な思考のこと。それを推し進める教育機関「[シンギュラリティ大学](#)」をAIの権威[レイ・カーツワイル](#)氏が創設。世界の最重要課題（国連の持続可能な開発目標）に取り込んでいる。

- 普通の目標：「二酸化炭素5年後までに10%削減するには？」
- 野心的な目標：「温暖化問題を根本的に解決するには？」

AI技術戦略会議 I



人工知能技術戦略会議の取組状況 (資料13-1)

平成28年(2016年)4月12日に開催された第5回「未来投資に向けた官民対話」における安倍総理の指示を受け、産学官の叢智を集め、縦割りを排した「人工知能技術戦略会議」が創設。「生産性」「健康・医療・介護」「空間の移動」が主な重点分野。また横断的な重点分野として「情報セキュリティ」も加えている。

AI 技術戦略会議 II

通信ソサイエティマガジン (No.45 夏号 2018)

● 生産性

- ▶ ものづくり・流通・サービスの融合が進み、エネルギー・食料なども含めた社会全体としての生産性を高めた究極のエコシステムの構築。人とロボットの協調生産工場、匠の技を再現するロボット、AI × 農作業ロボット、家電の AI 活用など。

● 健康、医療・介護

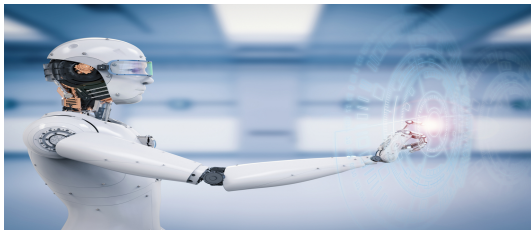
- ▶ 医療・介護の膨大な情報をビックデータ化し、AI を使って世界一の医療技術先進国・介護技術先進国を構築。また 80 歳でも就業を希望する高齢者が元気に働いている社会を実現。遠隔診療・在宅診療、歩行支援、見守り対話型ロボット、AI 補助による各種疾病早期発見、生体内医療ナノロボットなど。

● 空間の移動

- ▶ 人の移動時間・移動空間を「移動」そのものではなく、「作業」、「生活」、「娯楽」を行う時間・空間にする。移動のエコ社会を実現。移動のエンターテイメント産業、自動運転、自律型輸送・配送サービス、シェアリングエコノミーなど

近年の AI は、機械学習、特に深層学習（ディープラーニング）に基づくものが中心であるが、AI 関連の技術は急速に進展しており、AI に利用される技術に限定して AI の定義とすることはしない。

強い AI, 弱い AI



<https://iedge.tech/article/8835/>

人工知能の分類

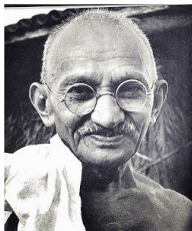
- 弱い AI: 特定のことしか出来ない AI. 特化型人工知能. 与えられた課題に対して優れた性能を発揮する. AlphaGo, IBM Watson, 自動運転技術
- 強い AI: あたかも人間のような自意識を備えている AI. **汎用性と自律性がある**. 汎用人工知能. 鉄腕アトム, ドラえもん

データサイエンスの目指すところ・役割



<https://careerhack.en-japan.com/report/detail/559>

- 膨大なビックデータから有用な情報 (鉱山の金) を抽出し、**普遍的な法則を発見**すること (Kepler の惑星運動法則).
- お金儲けではなく (ビジネス目的ではなく)、**科学 (サイエンス) の進展**に貢献すること.
- 全人類の幸福を実現すること. **国連の持続可能な開発目標の達成.**



7つの社会的罪 (マハトマ・ガンジー (1869-1948))

- 労働なき富：不当な手段で手に入れた富. 詐欺, 脱税, ...
- 献身なき信仰：信仰には行動も必要. 上辺だけの信仰はない.
- 良心なき快楽：他人を不幸にして手にする快楽. 浪費, いじめ, ...
- 理念なき政治：政治はビジョンが大事
- 人格なき学識：学識がある人はそれを社会のために生かすべき. 銜学者,..
- 道徳なき商業：利益だけを追求する商業活動. ブラック企業, ..
- **人間性なき科学**：科学は諸刃の剣.

AI, データサイエンスが進展する昨今, 科学を担う者の人間性がますます重要になってくる.

参考文献 I

- ▶ J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities.
Proceedings of the National Academy of Sciences (PNAS), 79(8): 2554–2558, 1982.
doi: 10.1073/pnas.79.8.2554.
URL <https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554>.
- ▶ University of Tokyo Institute of Mathematical Sciences.
History of ai, 2024.
URL http://www.mi.u-tokyo.ac.jp/pdf/3-1_history_of_ai.pdf.
Accessed: 2024-10-22.
- ▶ Daniel Jurafsky and James H. Martin.
Speech and Language Processing.
Draft of August 20, 2024, 2024.
- ▶ Kent Kanao, Osamu Komori, Jun Nakashima, Takashi Ohigashi, Eiji Kikuchi, Akira Miyajima, Ken Nakagawa, Shinto Eguchi, and Mototsugu Oya.
Individualized prostate-specific antigen threshold values to avoid overdiagnosis of prostate cancer and reduce unnecessary biopsy in elderly men.
Japanese Journal of Clinical Oncology, 44:852–859, 2014.

参考文献 II

- ▶ Osamu Komori.
A boosting method for maximization of the area under the ROC curve.
Annals of the Institute of Statistical Mathematics, 63:961–979, 2011.
- ▶ The Physical Society of Japan.
2024 nobel prize related information, October 2024.
URL https://www.jps.or.jp/information/2024/10/2024novelprize.php?utm_source=chatgpt.com.
Accessed: 2024-10-22.
- ▶ Japanese Neural Networks Society.
会長メッセージ「ノーベル賞に繋がる神経回路研究」, October 2024.
URL <https://tinyurl.com/4749tjvb>.
Accessed: 2024-10-22.
- ▶ The Japanese Society for Artificial Intelligence.
AI Map - Artificial Intelligence Research Map, 2024.
URL <https://www.ai-gakkai.or.jp/aimap/>.
Accessed: 2024-10-22.

参考文献 III

- ▶ 八木光晴 and 及川信.
生物の体サイズとアロメトリー：エネルギー代謝量と体サイズ.
比較生理生化学, 25:68–72, 2008.
- ▶ 柴田里程.
データ分析とデータサイエンス.
近代科学社, 東京, 2015.
- ▶ 渋谷政昭 and 柴田里程.
Sによるデータ解析.
共立出版, 東京, 1992.

基本用語

- Tokens w_i : ["Cats", "love", "fish"] のような単語. 文章 (sentence) の構成要素. $i = 1, \dots, N$.
- embedding vector (contextual vector) \mathbf{x}_i : 文字 (単語), 画像, 音声などのデータを機械学習で扱えるように数値化したもの. 文章の意味 (semantics), 構文 (syntax) の情報を保有する (埋め込んだ) 特徴量 (feature vector). 但し $\mathbf{x}_i \in \mathbb{R}^d$. chatgpt3 では $d = 12,288$.

例えば $w_1 = \text{"Cats"}$, $w_2 = \text{"love"}$, $w_3 = \text{"fish"}$ の場合は

- ▶ token embedding :
 $\mathbf{x}_1^t = [0.2, 0.4, 0.6, \dots]$
 $\mathbf{x}_2^t = [0.1, 0.3, 0.5, \dots]$
 $\mathbf{x}_3^t = [0.3, 0.5, 0.7, \dots]$
- ▶ positional embedding:
 $\mathbf{x}_1^p = [x_{1,1}^p, x_{1,2}^p, x_{1,3}^p, \dots]$
 $\mathbf{x}_2^p = [x_{2,1}^p, x_{2,2}^p, x_{2,3}^p, \dots]$
 $\mathbf{x}_3^p = [x_{3,1}^p, x_{3,2}^p, x_{3,3}^p, \dots]$ となる. 但し

$$x_{i,2j-1}^p = \sin\left(\frac{i}{10000 \frac{2(j-1)}{d}}\right), x_{i,2j}^p = \cos\left(\frac{i}{10000 \frac{2(j-1)}{d}}\right), j = 1, \dots, d/2$$

となり, 結局この二つの足し算で embedding vector を求める

$$\mathbf{x}_i = \mathbf{x}_i^t + \mathbf{x}_i^p, i = 1, \dots, N$$

Transformer 概念图

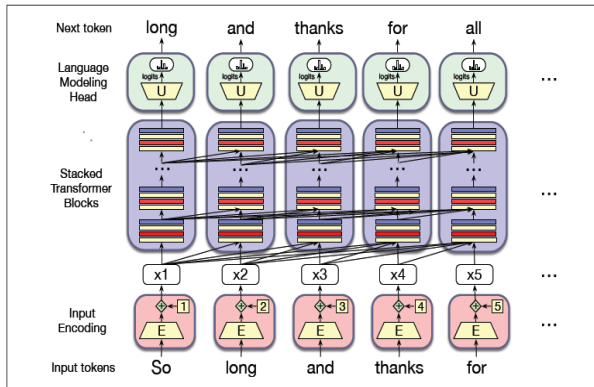


Figure 9.1 The architecture of a (left-to-right) transformer, showing how each input token get encoded, passed through a set of stacked transformer blocks, and then a language model head that predicts the next token.

引用：Speech and Language Processing. Daniel Jurafsky & James H. Martin.

Self attention layer

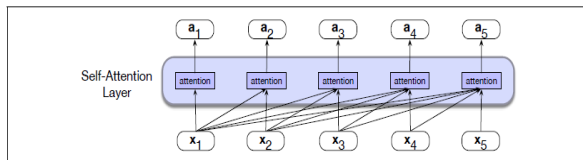


Figure 9.3 Information flow in causal self-attention. When processing each input x_i , the model attends to all the inputs up to, and including x_i .

引用：Speech and Language Processing. Daniel Jurafsky & James H. Martin.

$$q_i = x_i W_Q(\text{query}), \quad k_j = x_j W_K(\text{key}), \quad v_j = x_j W_V(\text{value})$$

$$\text{score}(x_i, x_j) = \frac{q_i \cdot k_j}{\sqrt{d}}$$

$$\alpha_{ij} = \text{softmax}(\text{score}(x_i, x_j)), \quad \forall j \leq i$$

$$a_i = \sum_{j \leq i} \alpha_{ij} v_j$$

Language modeling layer

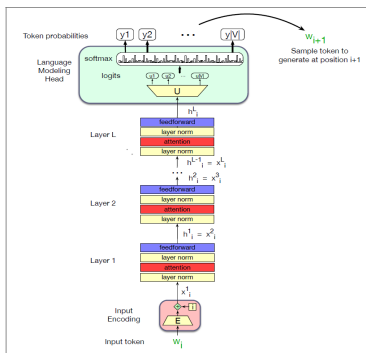


Figure 9.15 A transformer language model (decoder-only), stacking transformer blocks and mapping from an input token w_i to a predicted next token w_{i+1} .

引用：Speech and Language Processing. Daniel Jurafsky & James H. Martin.

$$\mathbf{u} = \mathbf{h}_N^L \mathbf{E}^T$$

$$\mathbf{y} = \text{softmax}(\mathbf{u})$$

one-hot vector で表現された $1 \times |V|$ の word w_i に $N \times |V|$ の embedding matrix \mathbf{E} をかけると embedding vector が生成される。よって Layer L から出力された embedding vector \mathbf{h}_N^L に \mathbf{E}^T をかける (unembedding) と、次の word w_{i+1} を予測するための $1 \times |V|$ の確率 \mathbf{y} が算出される。

Chatgpt の推定アルゴリズムの主要ステップ:

- 1 トークナイズ: テキストを token に分割
- 2 埋め込み層: token を embedding vector にマッピング。
- 3 トランスフォーマー層: self attention layer で文脈をエンコード。
- 4 言語モデル層: 次の単語を確率分布から予測。
- 5 繰り返し処理: 文章全体を生成。

応用: 機械翻訳、チャットボット、文章生成など。