データサイエンス入門

小森　理
成蹊大学 理工学部 データ数理専攻
対象校：下妻第一高等学校
日時：2023 年 6 月 2 日 (金)

本日の内容

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

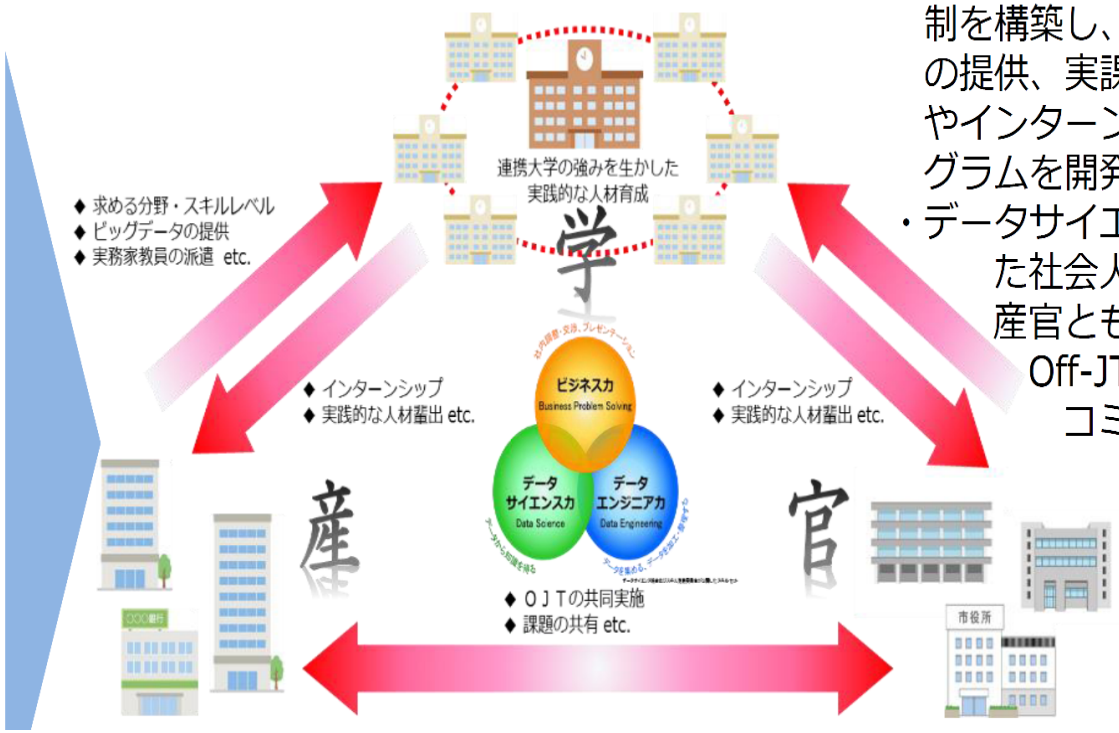
石油からデータの時代へ

- 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



・産業界や地方公共団体と強力な連携体制を構築し、必要となるビッグデータの提供、実課題によるPBL（共同研究）やインターンシップ等からなる教育プログラムを開発・実践
・データサイエンスを学ぶ必要に駆られた社会人の学び直しの場を提供し、産官ともに人材不足の中で、Off-JTの産官共同実施の機会やコミュニティ形成を醸成

※Off-JT：
Off-the-JobTraining
（職場外でのセミナーや講義による研修）

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

Society5.0

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？

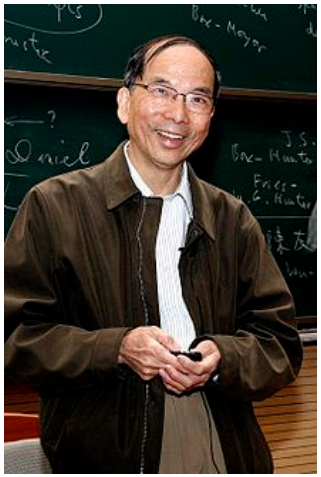
- データを科学 (サイエンス) する，または調理する学問.
- 見えなかった (気づかなかった) ものを見えるようにする学問．新たな価値の創生.
- ビッグデータ時代に急速に必要性が認識されてきた．膨大なデータの中から価値のある情報を取り出すことが主な目的.
- データマイニング (data mining) もその 1 つ.
- データの不確実性 (ランダム性) を扱う統計学と確率論と密接な関係がある.

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

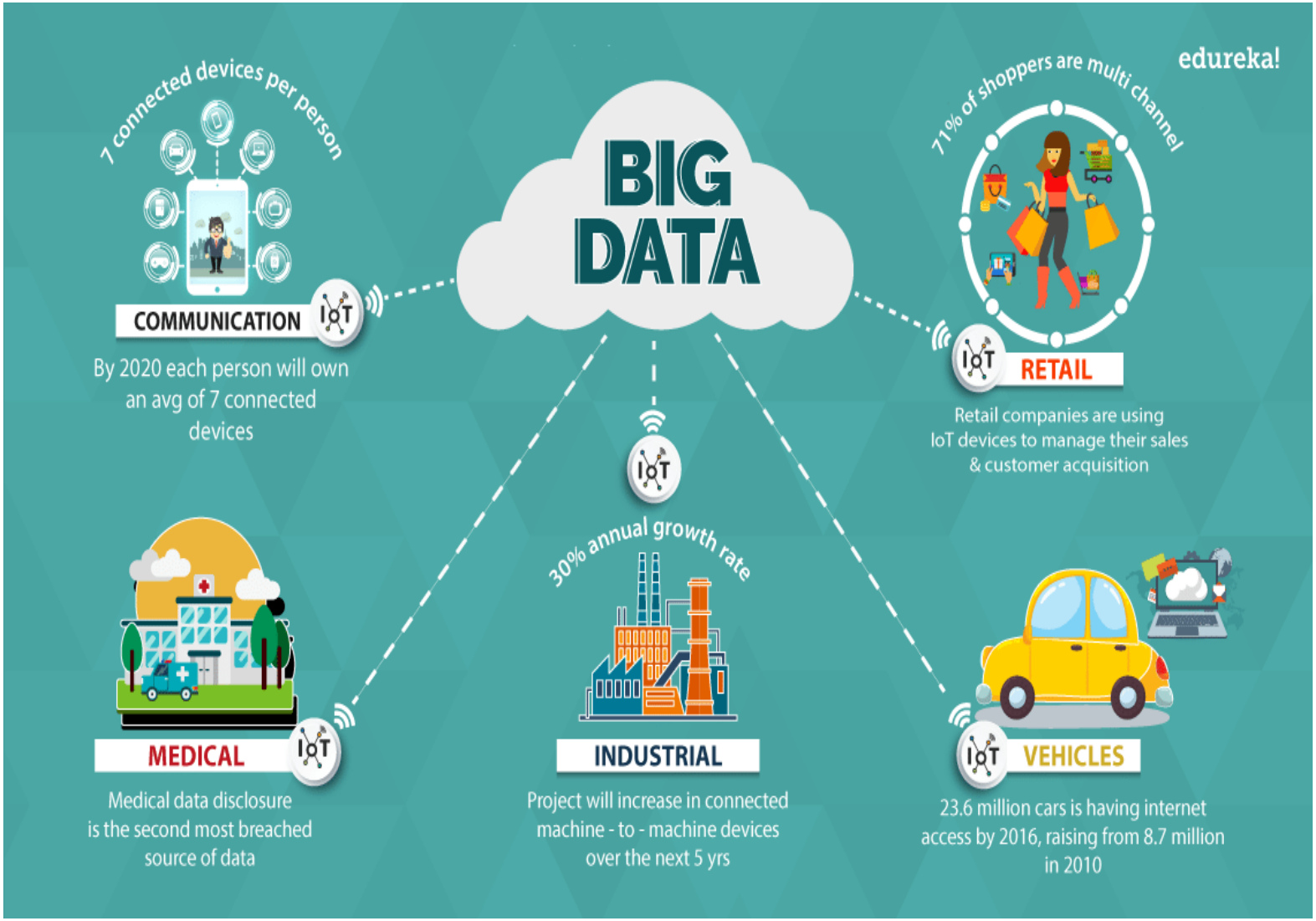


ジョージア工科大学教授．専門は統計科学 (データサイエンス)，統計数理研究所の Akaike Memorial Lecture Award の最初の受賞者．
そもそも統計 (statistics) はラテン語の status に由来．つまり国の状態を調べ，国民の幸福を増大させるためのもの (主に数を数えて表にまとめる作業)．しかし昔の統計とは違い，現代の統計の仕事は次の 3 つからなる．

- 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ク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**

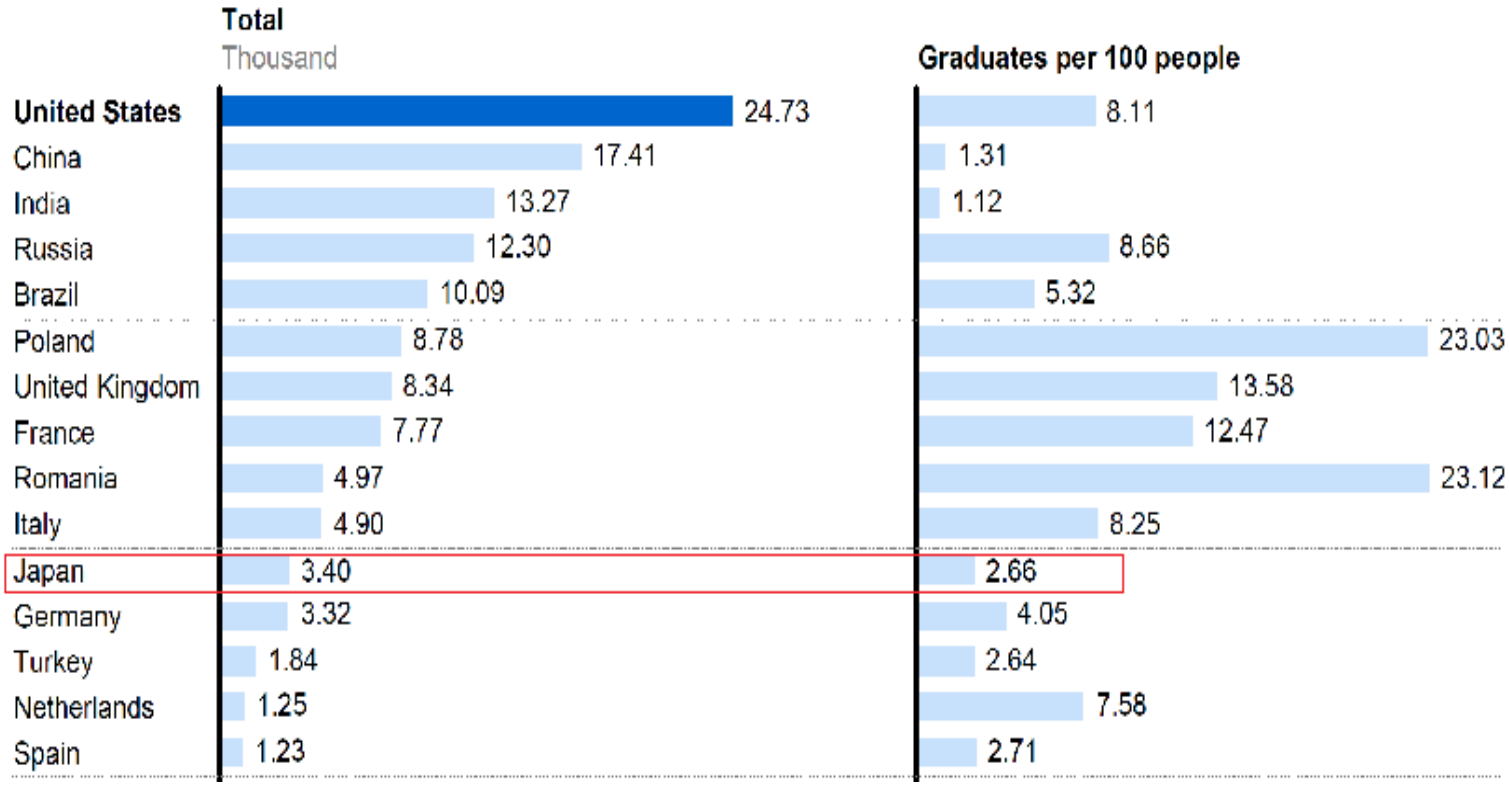
北川源四郎（6 大学コンソーシアムを中心とした数理・データサイエンス教育強化の取り組み）．データソース
http://datascience.community/colleges

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

$$185(2014/4) \Rightarrow 279(2015/7) \Rightarrow 505(2016/2) \Rightarrow 603(2019/5)$$

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

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

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

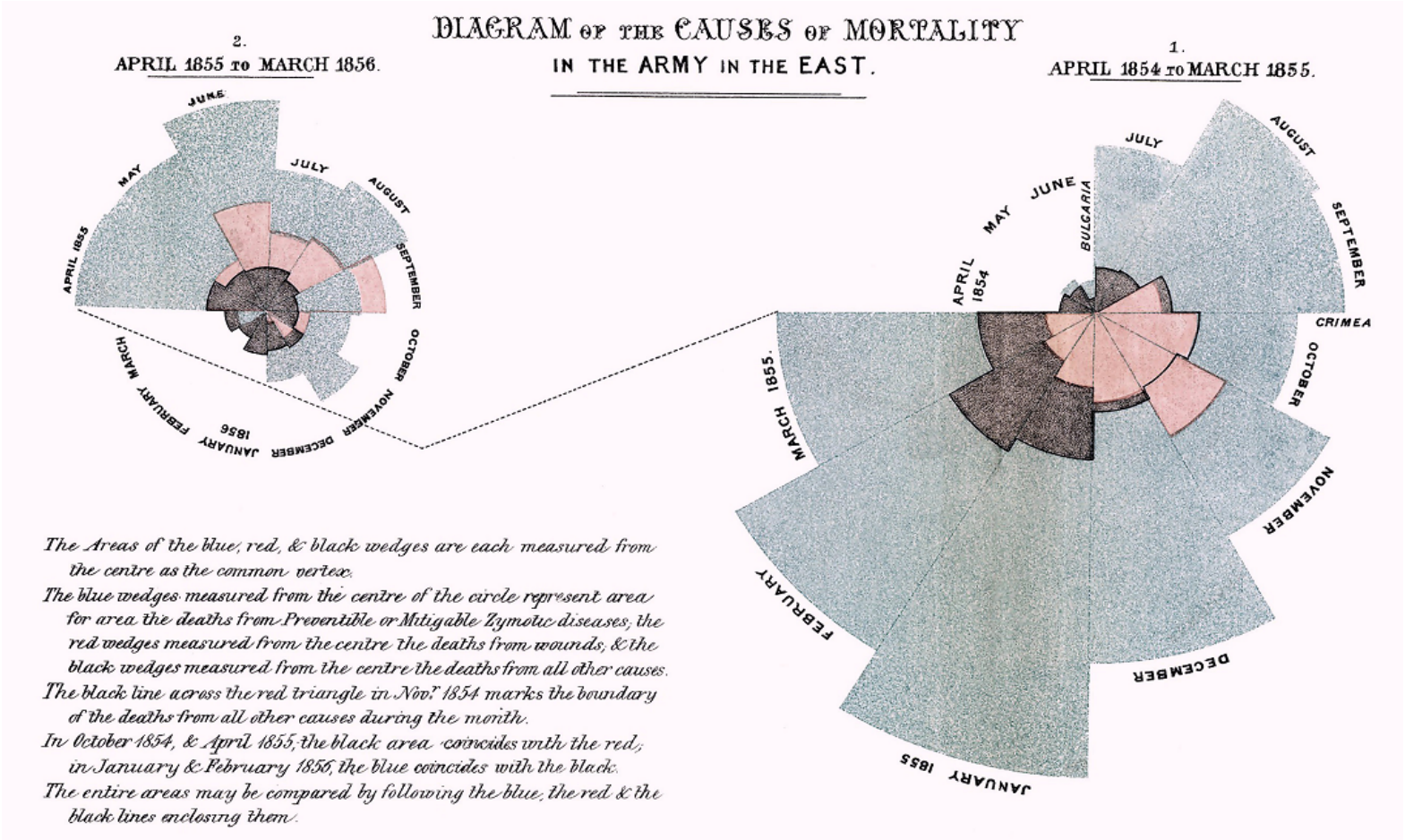
フローレンス・ナイチンゲールの言葉

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)

鶏頭図



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 grammer 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．

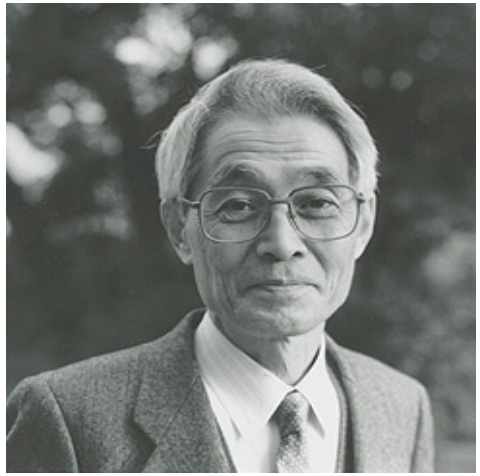
R. A. Fisher(1890-1962)

- 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.

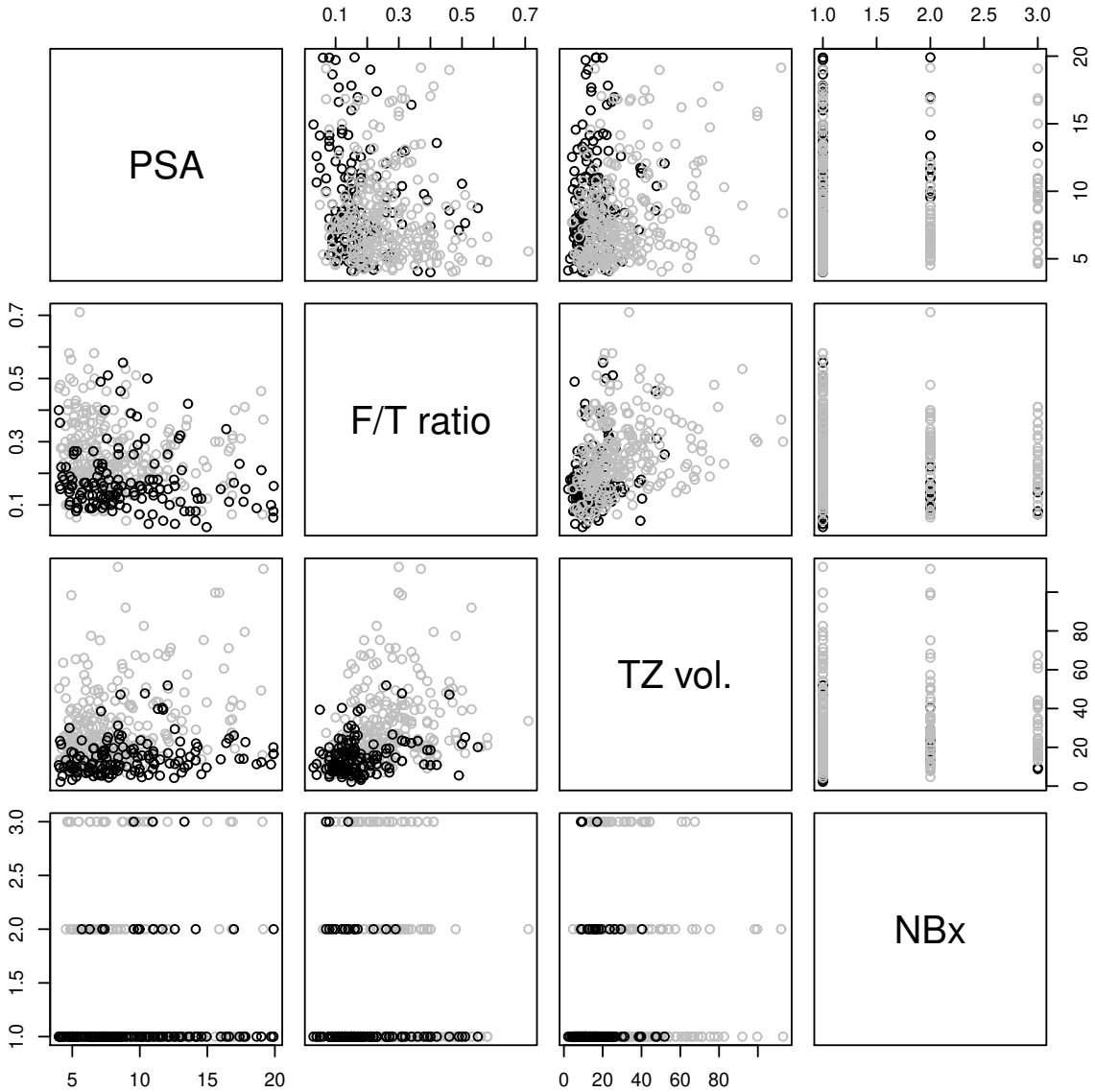
病院の腫瘍のデータ（実例）

- 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」

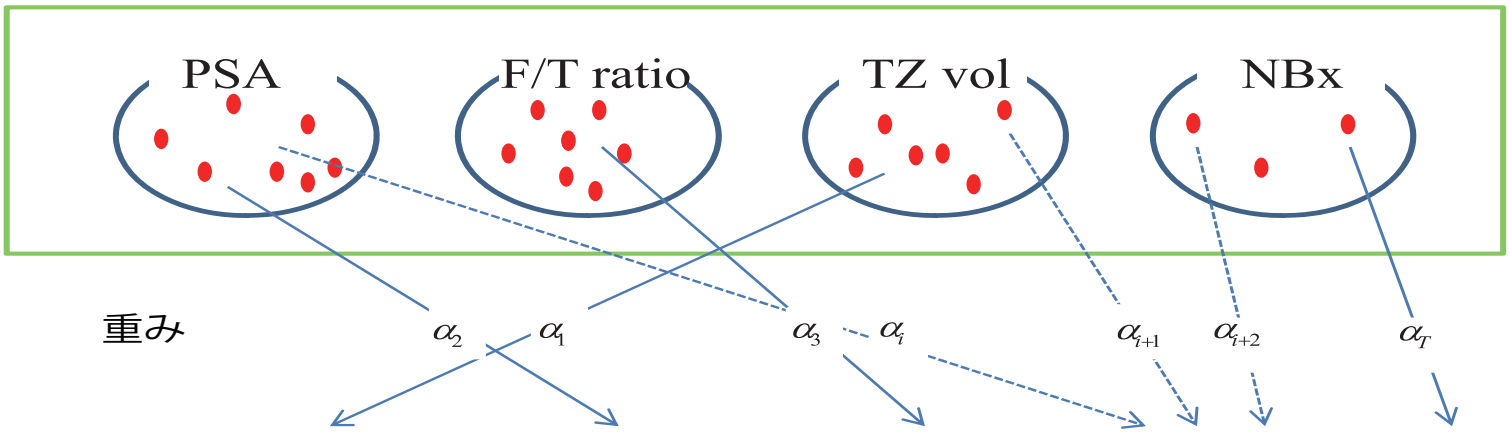
⇓

「ROC の下側面積 (AUC) 注目し，それを最大化するブースティングの手法を提案 (Komori, 2011)」



AUCBoost

４つの弱判別機の集合

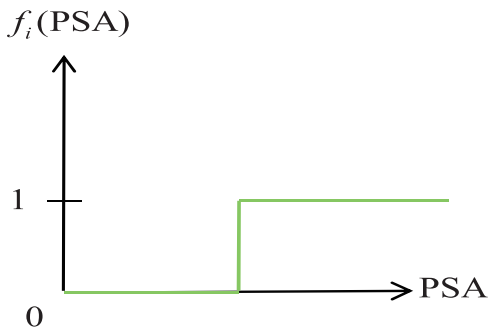


$$F(\boldsymbol{x}) = \alpha_1 f_1(\text{TZ vol}) + \alpha_2 f_2(\text{PSA}) + \alpha_3 f_3(\text{F/T ratio}) + \cdots\cdots + \alpha_T f_T(\text{NBx})$$
$$= F(\text{PSA}) + F(\text{F/T ratio}) + F(\text{TZ vol}) + F(\text{NBx})$$

診断

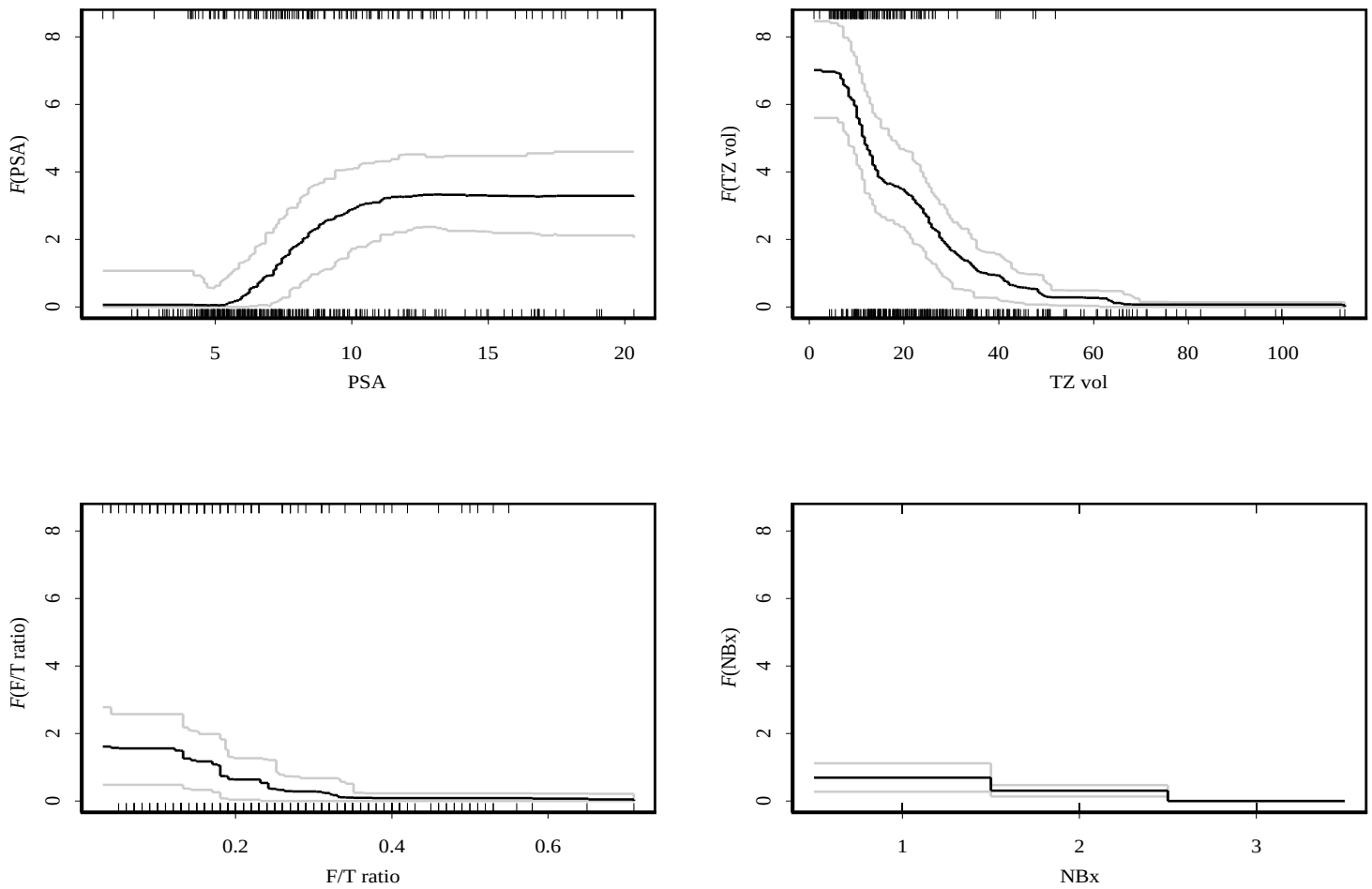
$F(\boldsymbol{x}) \geq c \Rightarrow$ 悪性腫瘍

$F(\boldsymbol{x}) < c \Rightarrow$ 良性腫瘍



- 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)

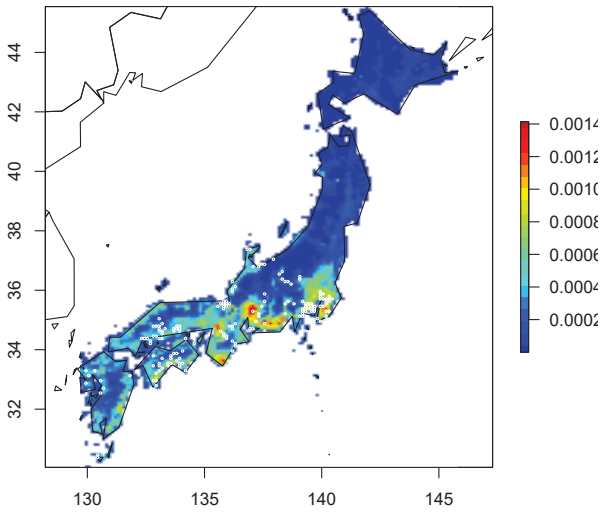
Table 2. Individualized PSA cut-off values for men aged ≥70 years old with normal DRE (≤5% risk of missing intermediate/high-risk cancer)

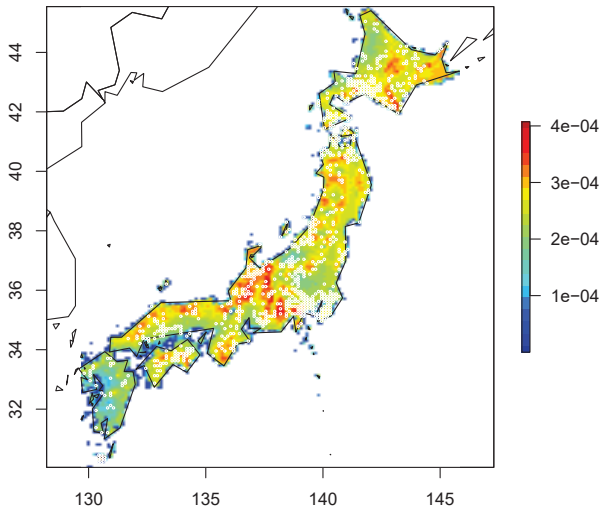
Initial biopsy

%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

Maxent による生物多様性パターン予測





Maxent のモデル式

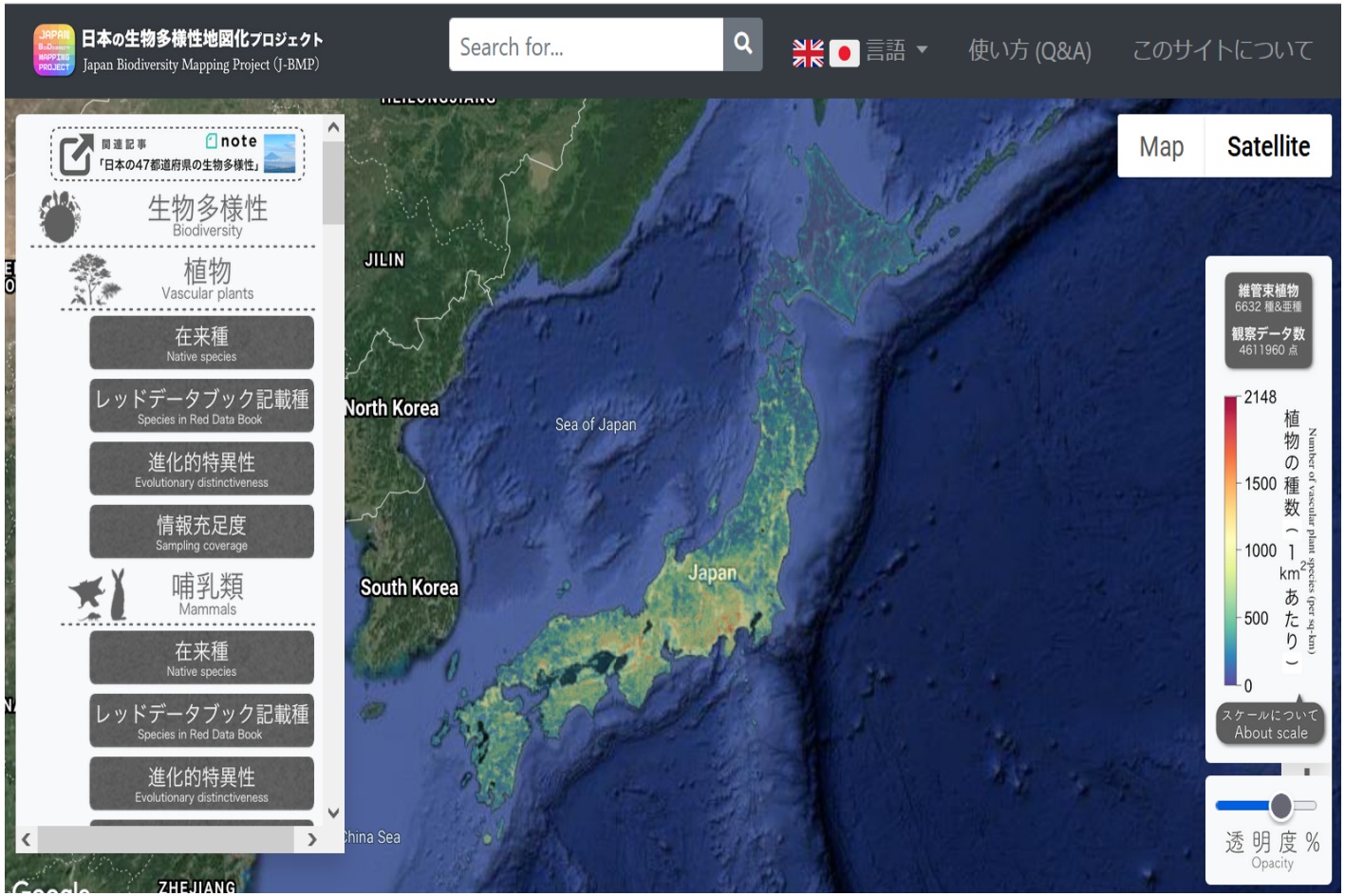
$$p(x) = \frac{\exp(\lambda^\top f(x))}{Z(\lambda)},$$

但し $Z(\lambda)$ は規格化項.

地点 x の気温，湿度，降水量などの環境変数 $f(x)$ と生息確率 $p(x)$ を結びつける統計モデル．パラメータ λ は最尤法で推定する．

日本の生物多様性地図化プロジェクト

琉球大学の久保田先生が主導するプロジェクト．生物ビックデータと Maxent を用いた生物多様性の可視化を実現．種数を正確に推定することで自然保護区の効果的な制定にも利用できる．



日本の生物多様性地図化プロジェクト：https://biodiversity-map.thinknature-japan.com/

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).
- サンプリ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，いつもど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．

モデル I

モデルとは現象の本質的な部分を抽出した骨格または模型である．骨格は単純なほどよい．その際に有用となる考えがケチの原理である．

ケチの原理

ケチの原理とはオッカムの剃刀 (Ockham's razor) とも呼ばれる．実際の現象をなるべく単純なモデルで説明しようとする考え方．

- １４世紀の哲学者オッカムが提唱した考え．
- 情報量に見合った複雑さのモデルを考えることが大切 (データが少ないときは単純なモデルを優先すべき)．
- モデルを通して現象を理解する．つまりモデルはデータを理解するうえでのレンズの役割を果たす．
- 但しあくまでモデルは近似でしなかないことに注意．

モデル II

モデリングには４つのステップがある．

- 現象のどの側面をモデリングするかを決める（目的を絞る）
- データにその目的を達成するのに十分な情報が入っているかを見極める
- モデルの近似精度を吟味する (実用性があるのか？)．
- モデルの実用性を見極める．

逆に何も有益な情報が入っていないデータを，しっかりと有益な情報なしと判断できることも重要．データの蓄積収集方法の改善につながる．

データエンジニアリング I

データサイエンスと対比する言葉にデータエンジニアリングがある.

データエンジニアリングとデータサイエンス

- データエンジニアリング：データ工学とも呼ばれる．明確な目的があり（売り上げ増など），それを達成すること目指す.
- データサイエンス：データ科学とも呼ばれる．目的（売り上げ増など）を達成するだけでなく，何故，どのようにその目的が達成されたかの仕組みを解明することを目指す.

機械学習 (再掲)

データ工学の 1 つ．確率論を基礎とした統計科学の枠組みを拡張した分野で，データに内在する規則性をさまざまなアルゴリズムを駆使して学習する．人間の五感を模倣し膨大なデータを機械的に処理することが出来る．ブースティング，SVM, Lasso, 遺伝的アルゴリズムや深層学習などが代表例.

最近では物理学の基本方程式と実際のデータを融合させ，より精度の高い予測を地中規模の現象

(気象現象，宇宙現象) の予測に適用することも試みられている (データ同化).

データの上流から下流まで I

データサイエンスの実践のプロセスは川の流れに例えられる．データが生まれるところを源流とし，新たな価値の発見創造が川の下流となる．上流から下流までを俯瞰する大域的な視点が必要．

解析，モデル構築

データ解析をする際，解析者の知識（会得している理論）と経験によってその解析方針が大きく変わる．また理論には 2 種類あり，データが取得された分野特有の理論 (医療データなら医学に関する知見，金融データなら経済金融の知見など) と，データ解析をするうえでどの分野でも必要になる共通の理論である．前者はその都度勉強する他ないが，後者は数理統計，データサイエンスを勉強することで習得が可能．

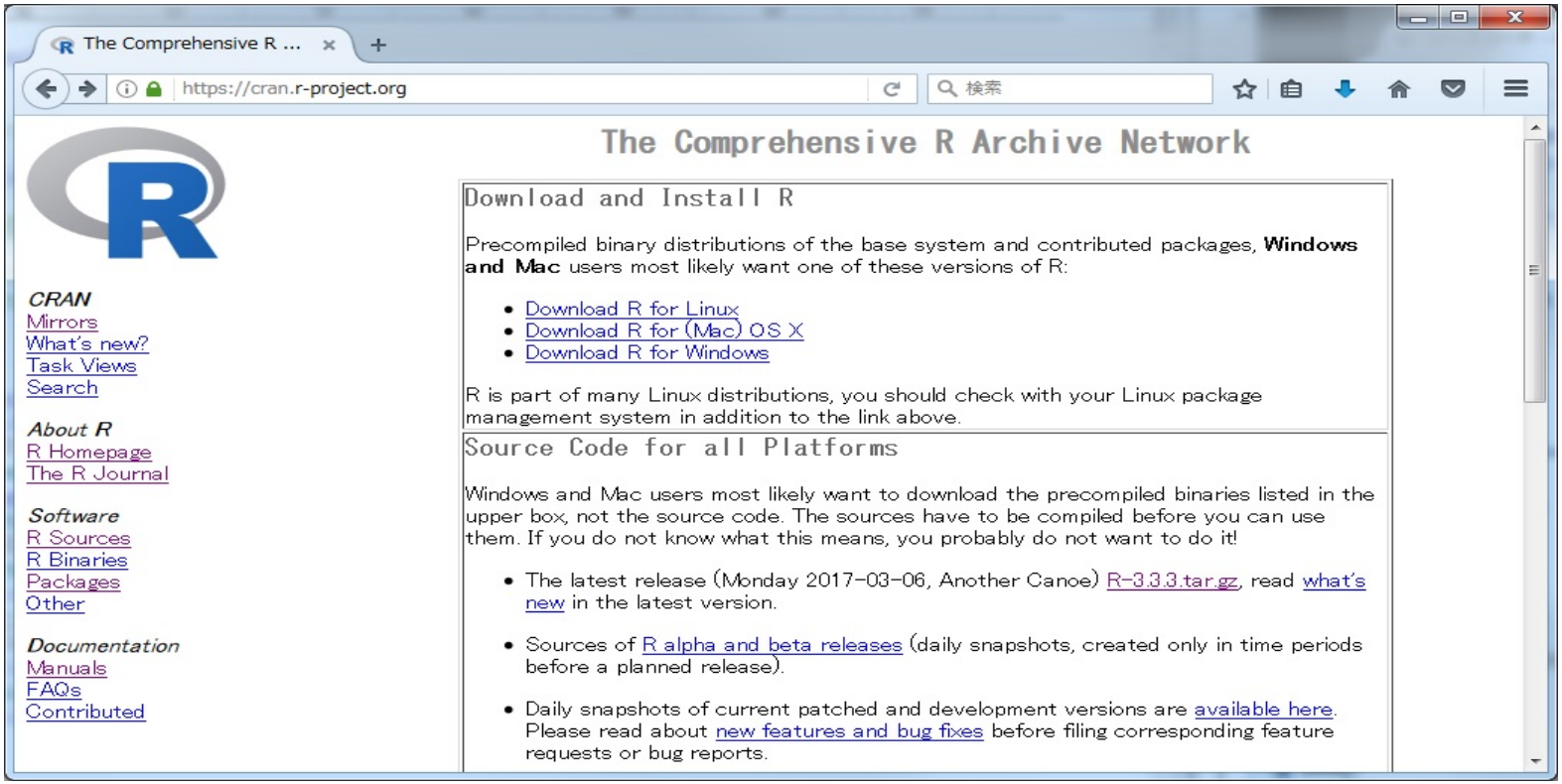
プレゼンテーション

データサイエンスの実践の最後は，データサイエンスで得られた新しい知見を第三者に発表し伝えることである．相手の心理を的確に読み取り，興味を持ち続けてもらうことが重要．１つのストーリを作るようにプレゼンテーションも工夫する必要がある．必ず伝えるべき項目を 1 つか 2 つ選び，それを中心にまとめるのがよい．その際の原則は「作成する図は簡潔に，その説明は丁寧に」が基本．

散布図，垂線プロット，棒グラフ，ヒストグラム，箱ひげ図．円グラフなど．

発表内容

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．

ベル研究所

- 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)

PRINCETON UNIVERSITY

- 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 (⇔ 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．

R のインストール I

最新の R をインストールする.

- Windows 版　https://qiita.com/FukuharaYohei/items/8e0ddd0af11132031355
- Mac 版　https://www.typea.info/blog/index.php/2020/12/07/x_install_r/

注：Mac 版は「XQuartz」も一緒にインストールしておくこと（描画に必要）
https://www.typea.info/blog/index.php/2020/12/07/x_install_r/

RData の保存方法

```
>x=1
>ls()  #作成したオブジェクトを表示
[1] "x"
> getwd()  #カーレントディレクトの表示
[1] "C:/Users/owner/Desktop"
>save.image(file="a.RData") #カーレントディレクトリに a.RData ができる.
```

- a.RData のファイル名 a は任意. 自分の好きな名前で保存する.
- ファイルはこまめに保存することを勧める.
- a.RData を立ち上げ，ls() を実行すると作成したオブジェクトが保存されていることを確認できる.

便利なショートカット I

- Ctrl+C: コピー
- Ctrl+V: 貼り付け
- Ctrl+S: 保存
- Ctrl+X: 切り取り
- Ctrl+Z: やり直し (間違えたときに元の状態に戻す)
- Ctrl+A: 全て選択

茨城県のオープンデータ I

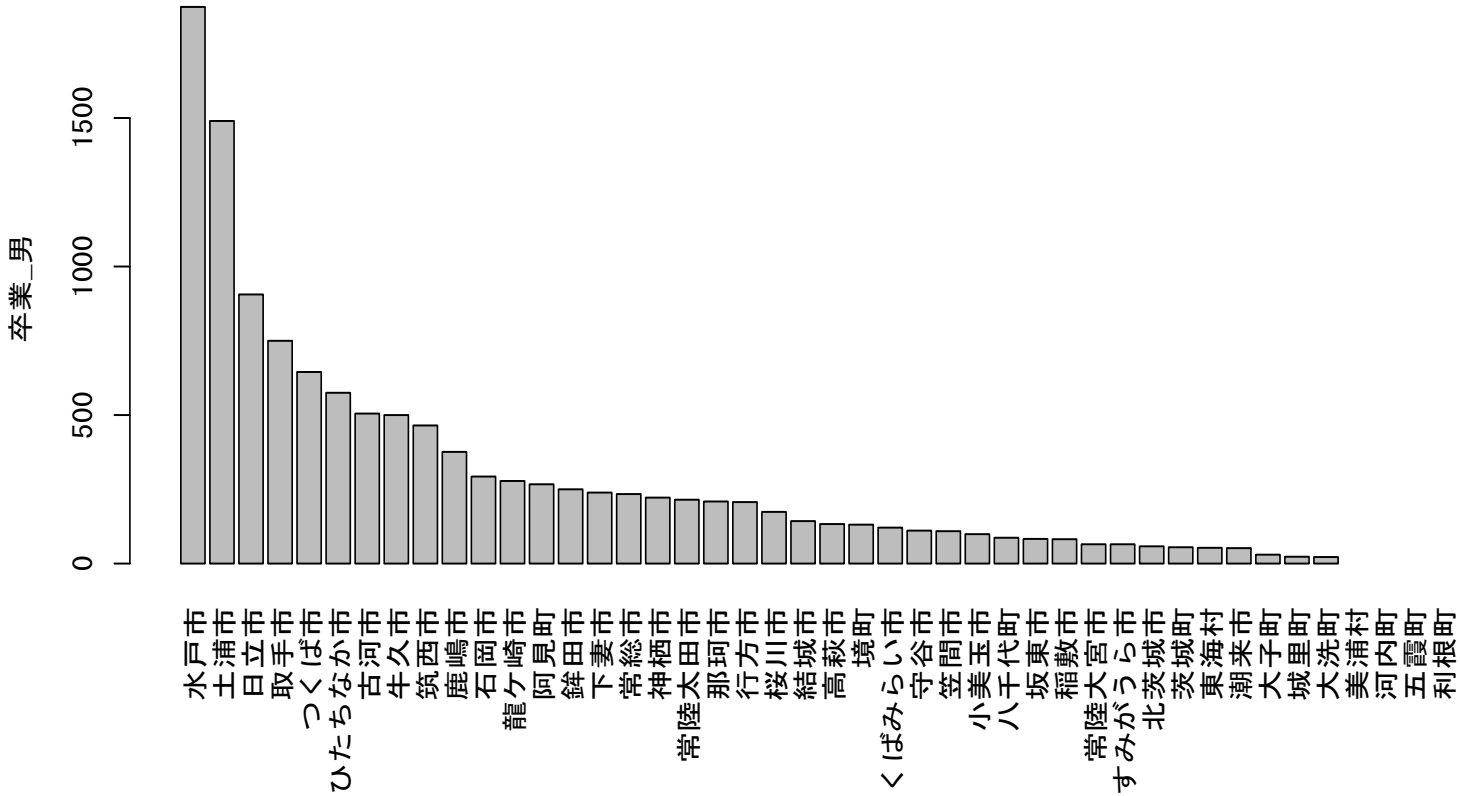
茨城県では様々なデータが公開されている.

https://www.pref.ibaraki.jp/kikaku/tokei/fukyu/tokei/betsu/kyoiku/gakucho2022k/hyou.html#hyou1-01

今回は「高等学校〔全日制・定時制〕エクセル：85 キロバイト)」のデータを使う.

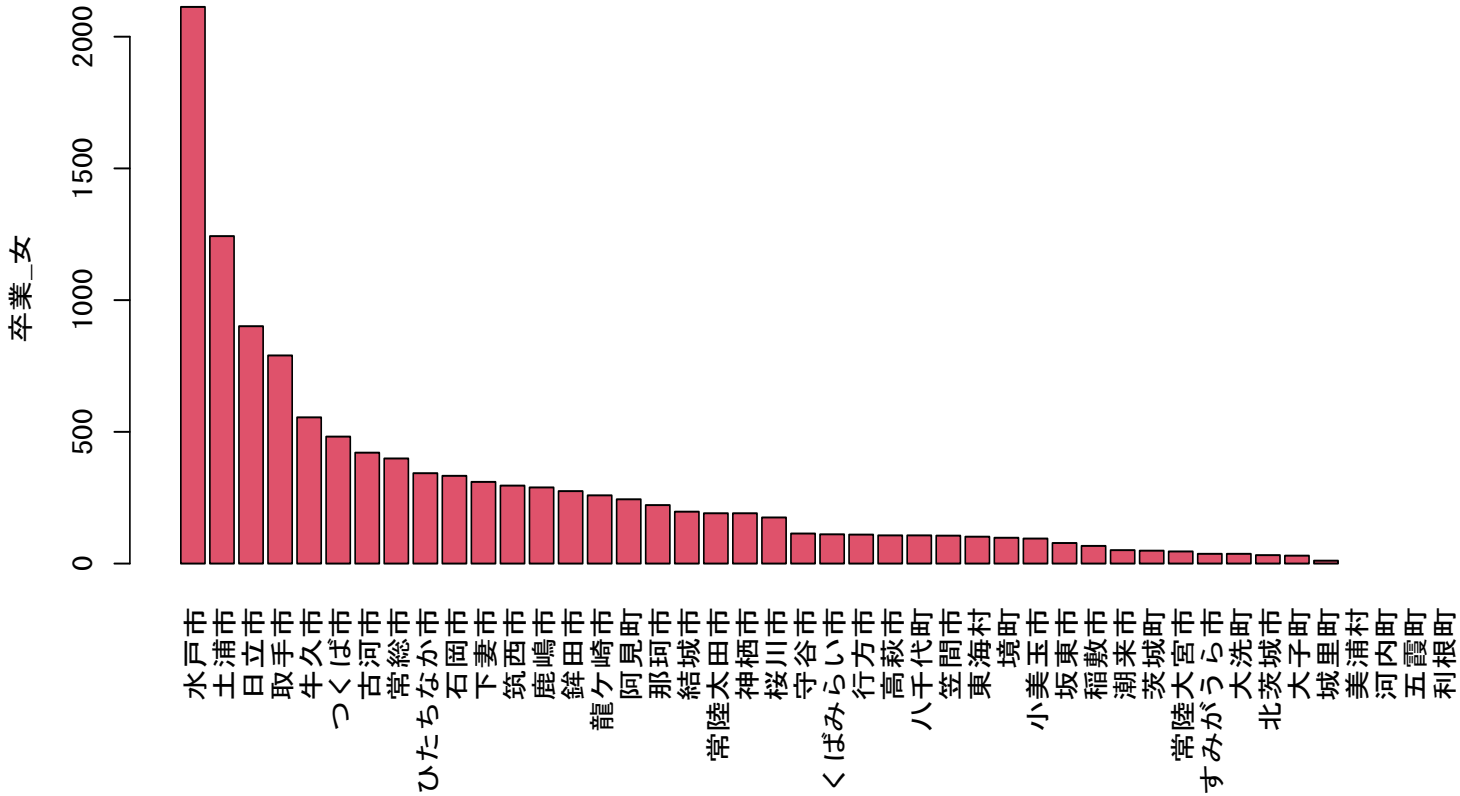
- 各市町村ごとの卒業者数の分布は？
- 進学率が上位の市町村はどこ？
- 男子と女子で進学率に違いがあるのか？

茨城県のオープンデータ II



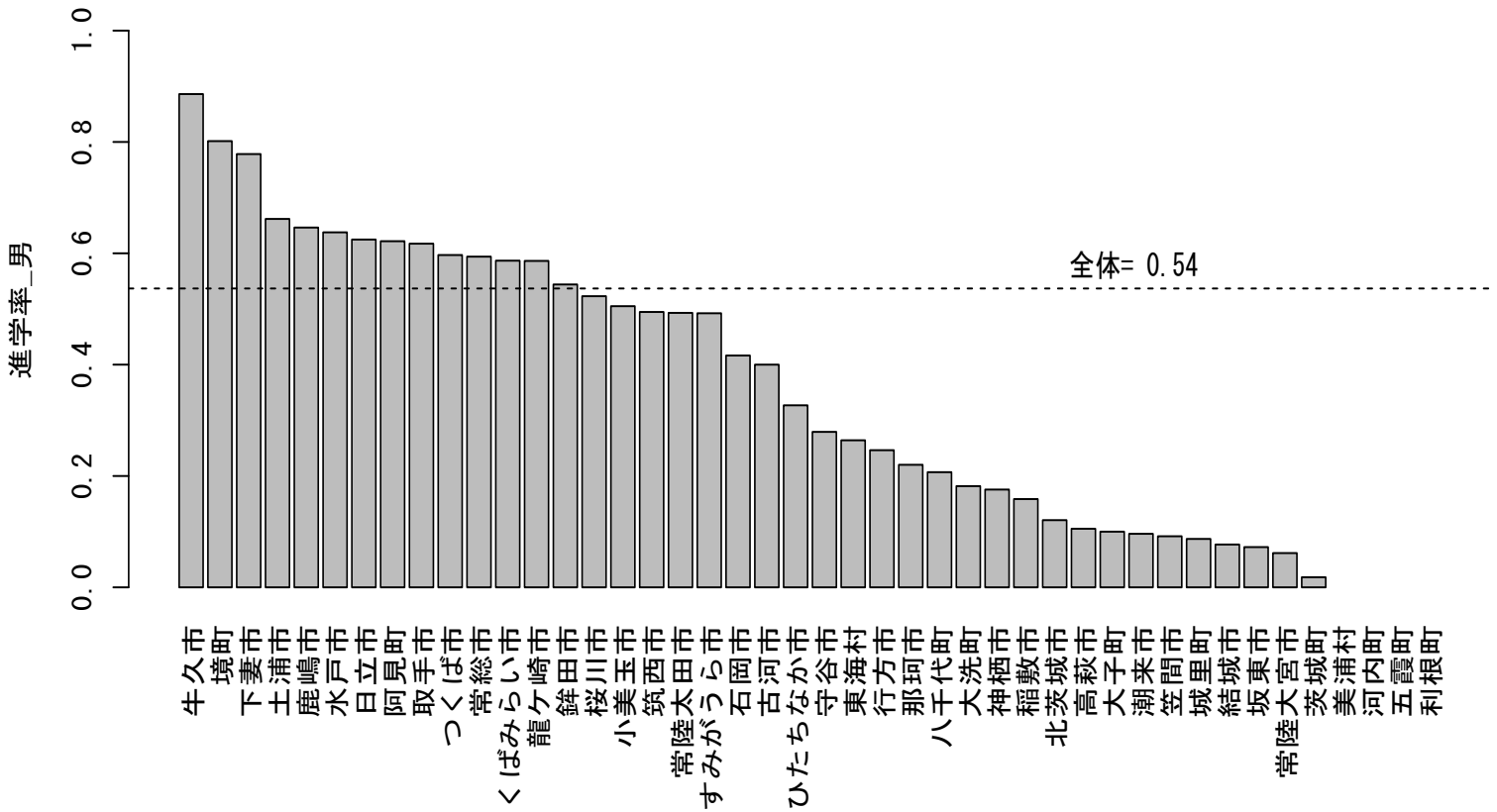
- 男子の卒業者数水戸市が多い.

茨城県のオープンデータ III



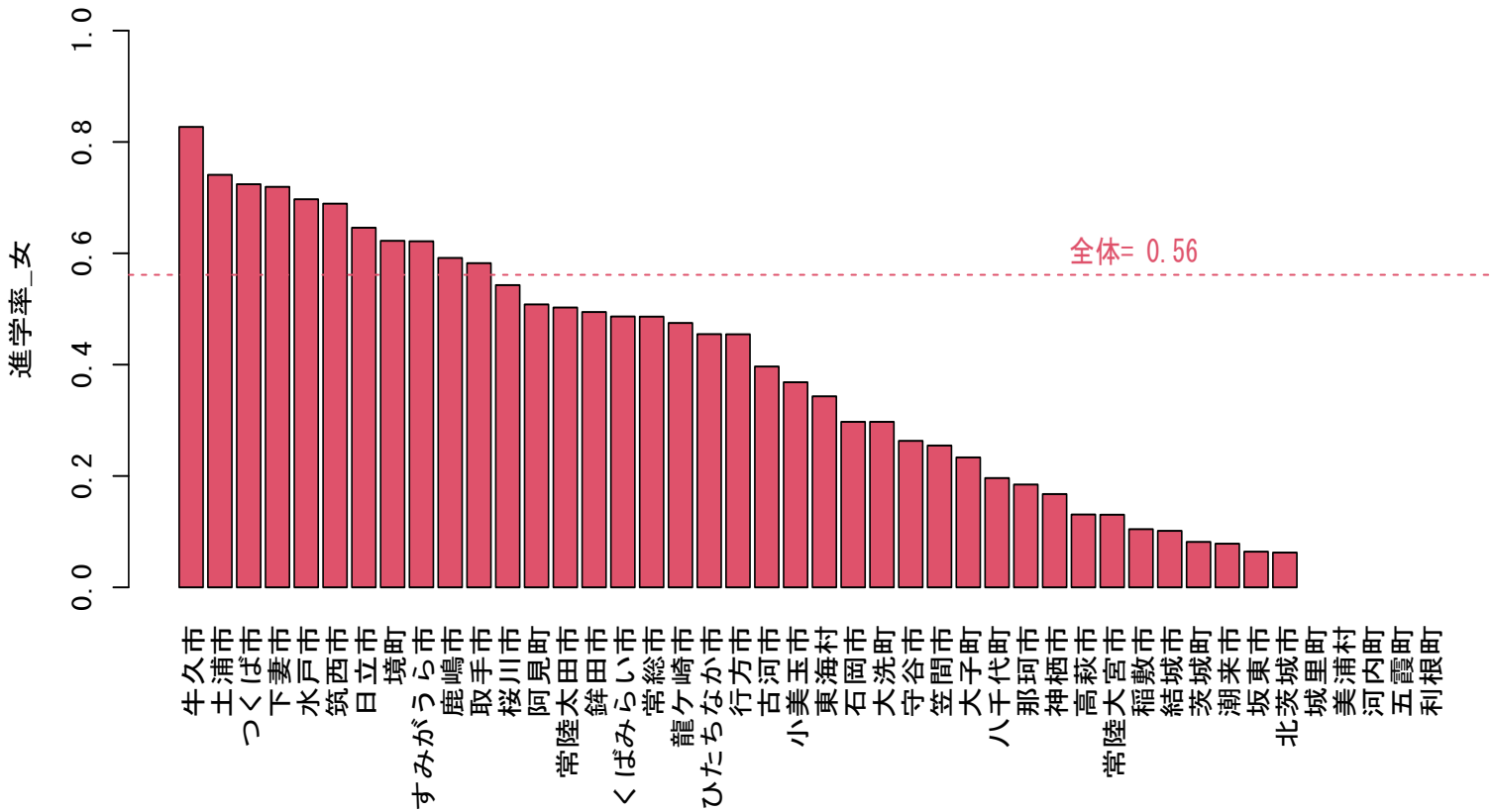
- 女子も同じ

茨城県のオープンデータ IV



- 進学者数は牛久市が多い.

茨城県のオープンデータ V



- 女子も同じ

哺乳類の体と脳の重さのデータ I

Asian elephant

Rhesus monkey(赤毛猿)

Kangaroo

lesser shourt tail shrew

Water opossum

Musk shrew

Owl monkey

Golden hamster

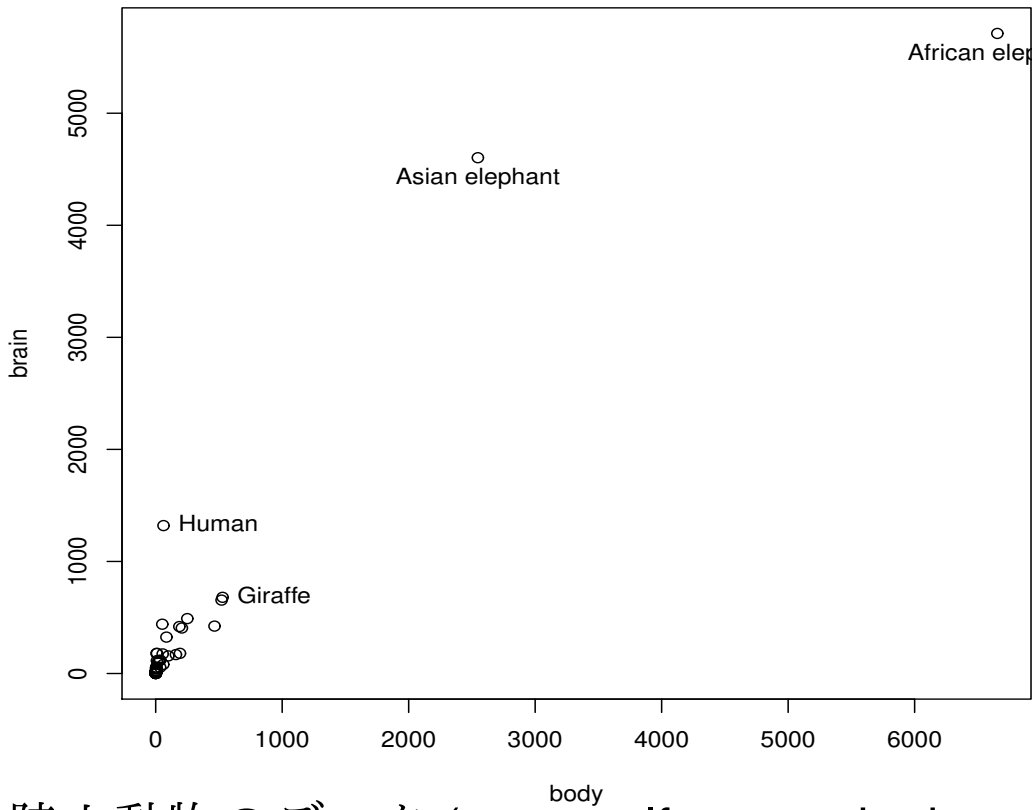
Ground squirrel

Pig

Giraffe

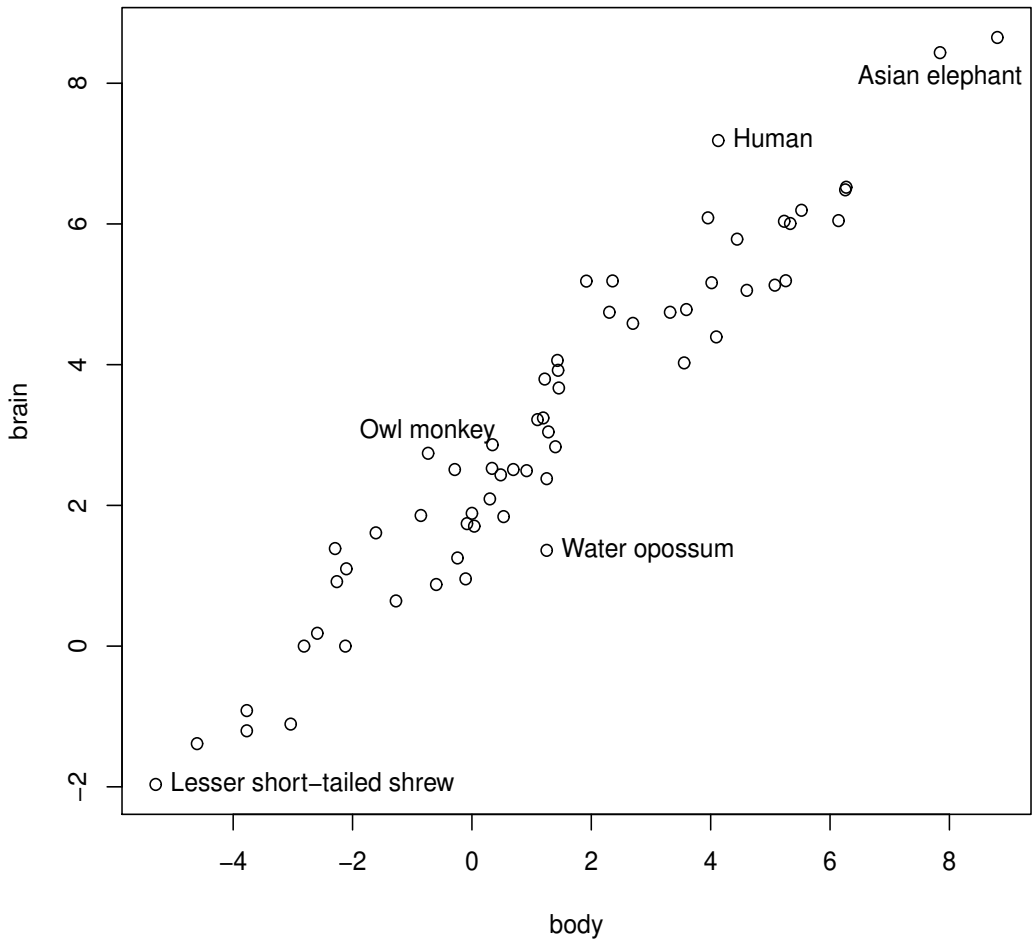
Donkey

哺乳類の体と脳の重さのデータ II



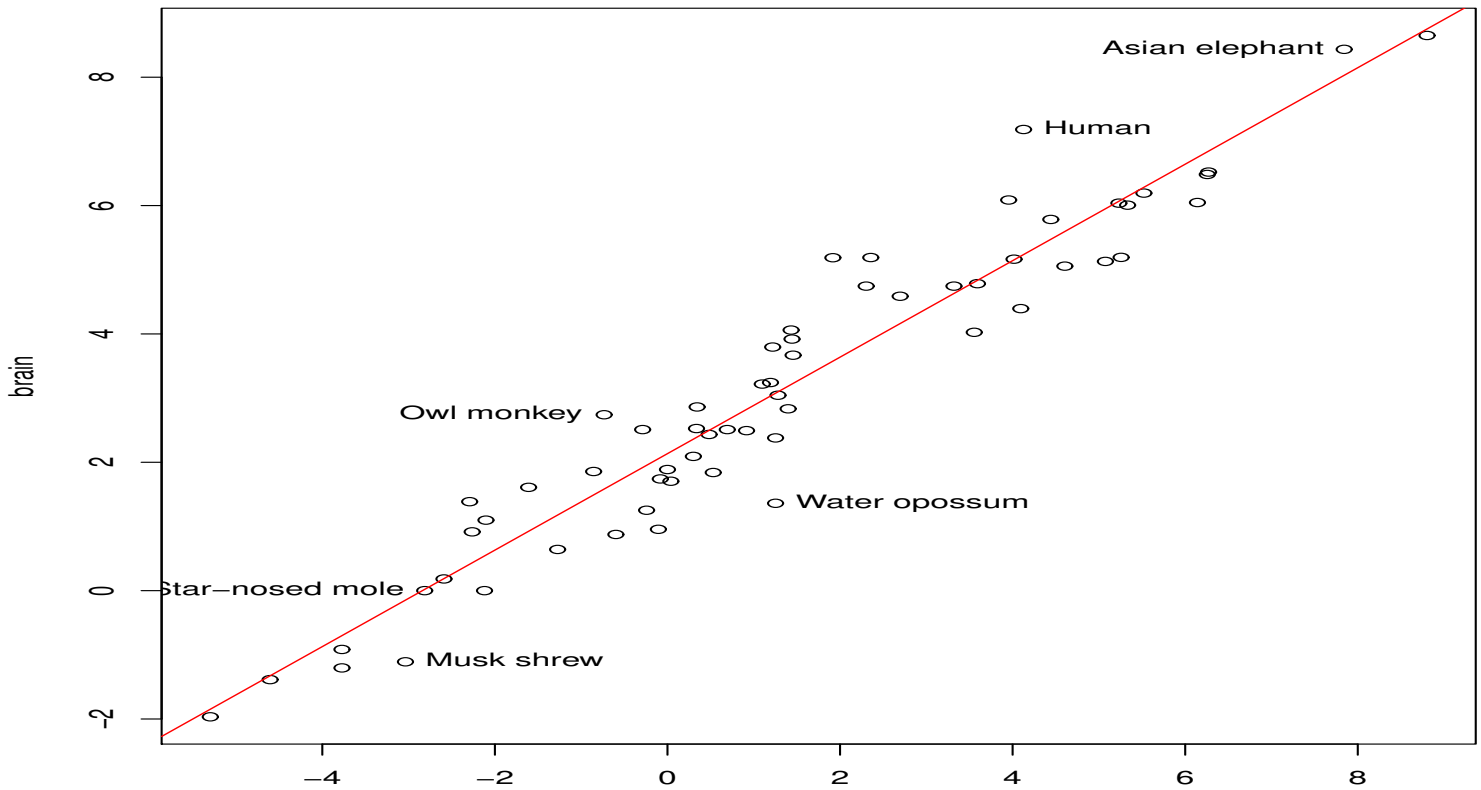
- 25 個体の陸上動物のデータ (cow, wolf, goat, pig, horse, monkey,...). 統計ソフトウェア R の MASS package から引用
- 横軸：体重 (kg), 縦軸：脳の重さ (g)
- 一見体重と脳の重さには関係性は見られない.

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-nozed 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

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

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).

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

```
>ls()
[1] "fix"    "fun1"   "fun2"   "fun3"   "fun4"   "jpn"
[7] "Mammal" "Shigen"
```

プログラムを少し修正すると，他の県の地図も描ける．

パッケージのインストールとプログラムの修正

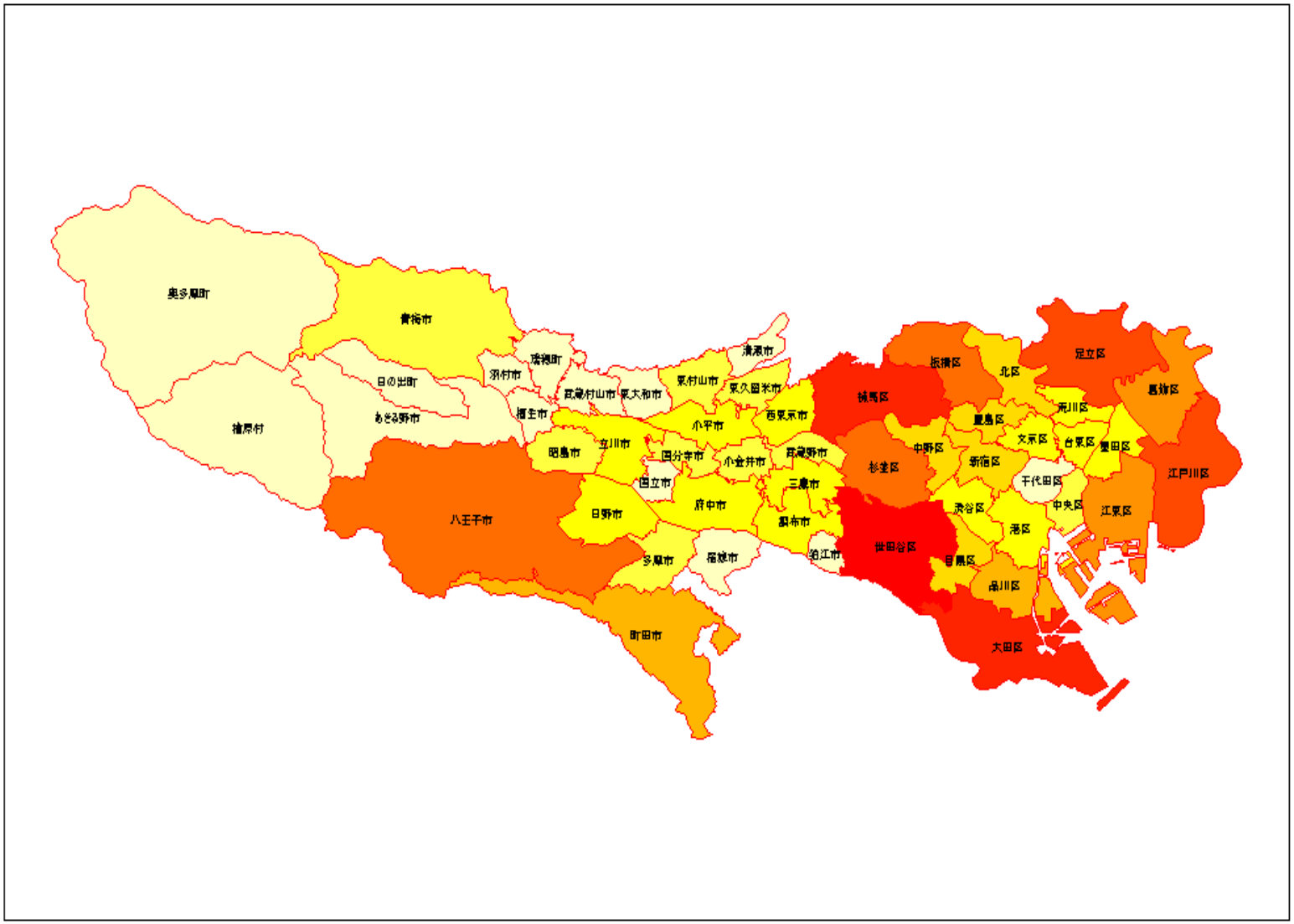
```
>install.packages("maptools")
>fix(fun4)
```

東京都の地図の描画 II

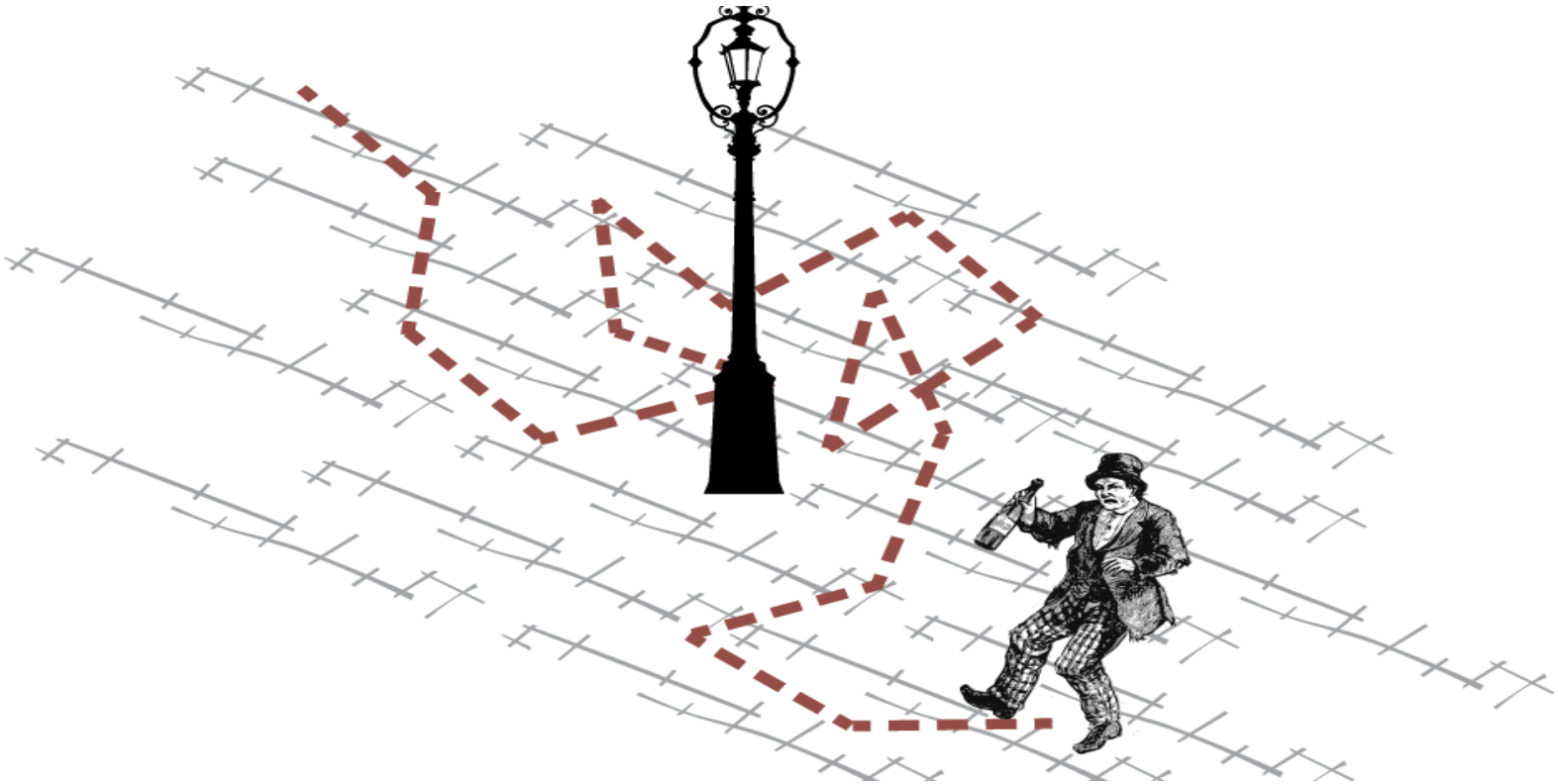
fun4 関数

```
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)
}
```

東京都の地図の描画 III



2次元のランダムウォーク (酔歩) I



2次元のランダムウォーク (酔歩) II

ランダムウォーク S_n

$Z_1, Z_2, \ldots, Z_n$ をそれぞれ独立で以下のようなカテゴリカル分布に従う 2 次元の確率変数とする

$$P(Z_i = (1,0)) = P(Z_i = (-1,0)) = P(Z_i = (0,1)) = P(Z_i = (0,-1)) = \frac{1}{4}$$

$$\Leftrightarrow \quad P(\text{右に進む}) = P(\text{左に進む}) = P(\text{上に進む}) = P(\text{下に進む}) = \frac{1}{4}$$

このとき

$$S_n = Z_1 + Z_2 + \cdots + Z_n$$

を 2 次元の対称なランダムウォークという．但し $S_0 = (0,0)$(原点) とする．

統計ソフト R

```
>fix(fun5) #プログラムを編集する
>fun5() #プログラムを実行する
```

Rでの2次元ランダムウォーク (酔歩) の実装 I

fun5

```
function(n=200000){
  set.seed(1); x=0; y=0; col=1 #x=0; y=0 で初期値を原点に設定
  plot(1,1,,xlim=c(-400,100),ylim=c(-400,300),type="n",xlab="x",ylab="y")
  for(i in 1:n){
    ran=which(rmultinom(1,1,prob=rep(0.25,4))==1) #カテゴリカル分布の乱数
    if(ran==1){ #右に進む. x 軸の正の方向に 1 進む
      x[i+1]=「?」
      y[i+1]=y[i]
    }
    else if(ran==2){  #左に進む. x 軸の負の方向に 1 進む
      x[i+1]=x[i]-1
      y[i+1]=y[i]
    }
    else if(ran==3){  #上に進む. y 軸の正の方向に 1 進む
      x[i+1]=「?」
      y[i+1]=y[i]+1
    }
    else{ #下に進む. y 軸の負の方向に 1 進む
      x[i+1]=x[i]
      y[i+1]=y[i]-1
    }
    if(i%%20000==0) col=col+1 # 線の色を変える
    segments(x[i],y[i],x[i+1],y[i+1],col=col) #線分を引く
  }
}
```

- 上記の「?」を埋めてプログラムを完成させよ.

Rでの2次元ランダムウォーク (酔歩) の実装 II

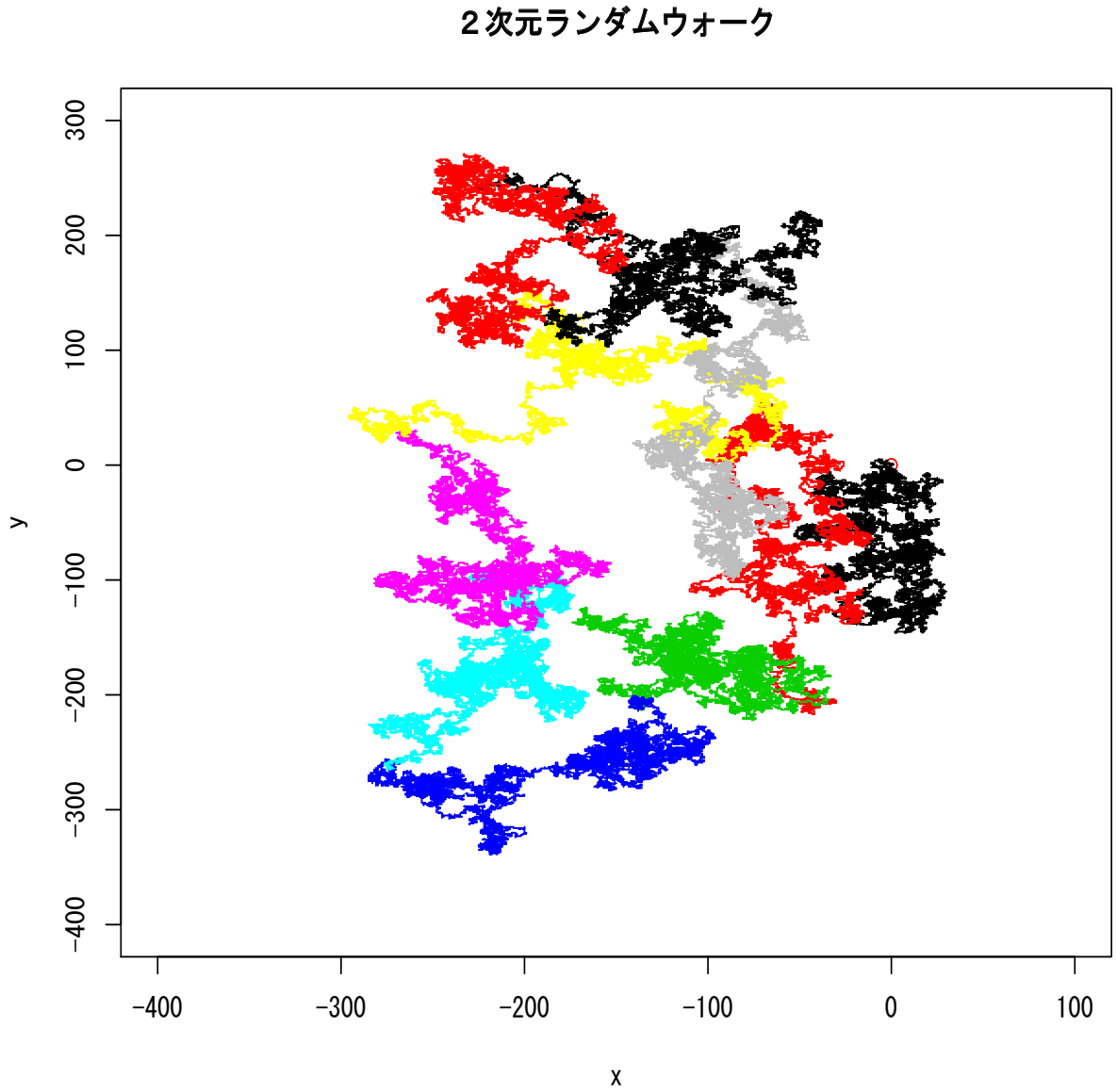


図２　２次元ランダムウォーク

コロナ陽性者予測 I

コロナ陽性者予測のモデルとして土谷モデルがある.

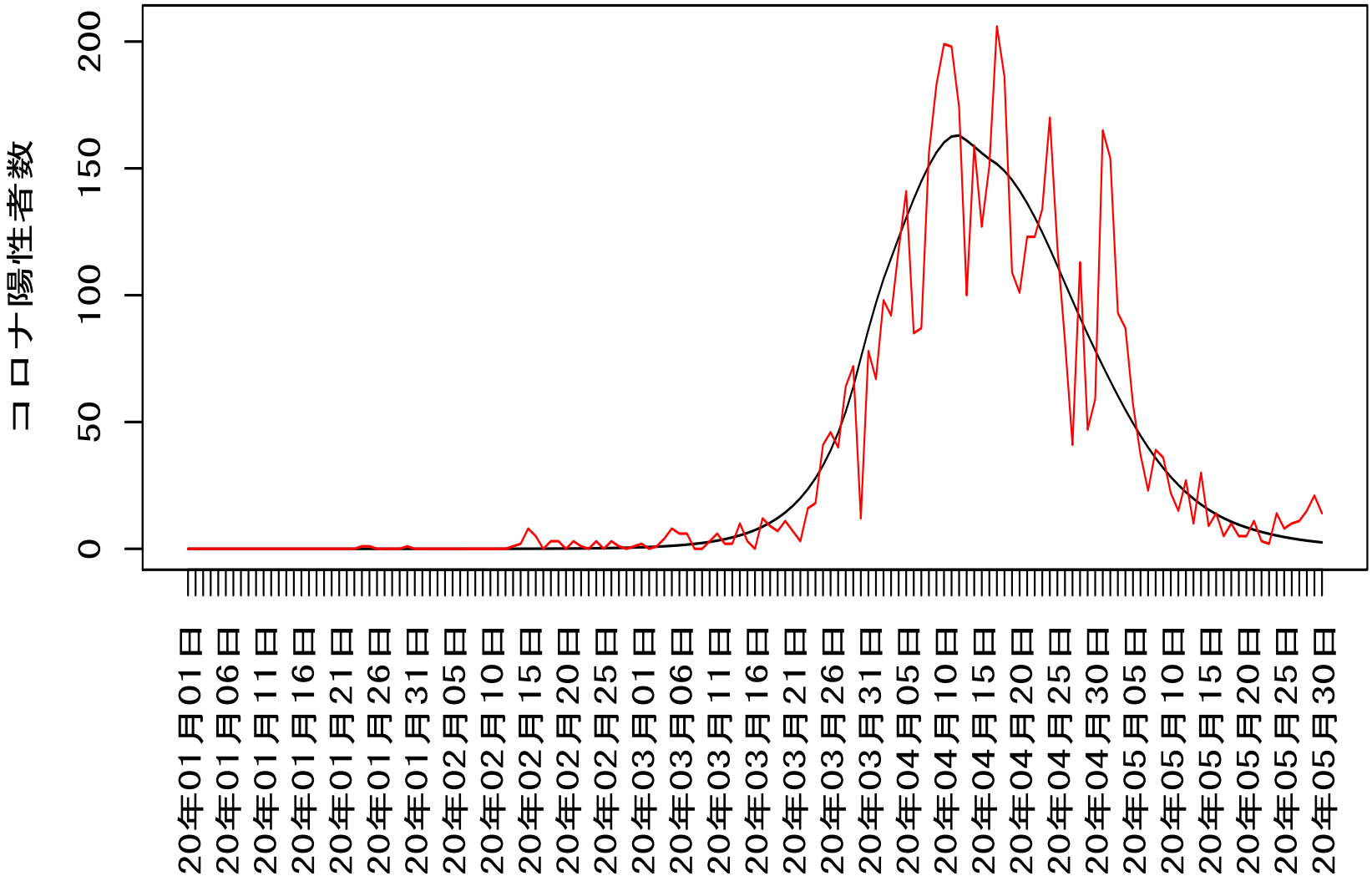
https://www.grips.ac.jp/jp/news/20200604-6443/

$$
\begin{aligned}
S(t+1) &= S(t) - \beta(t)I(t)S(t) \\
I(t+1) &= I(t) + \beta(t)I(t)S(t) - \beta(t-D)I(t-D)S(t-D) \\
R(t+1) &= R(t) + \beta(t-D)I(t-D)S(t-D)
\end{aligned}
$$

但し

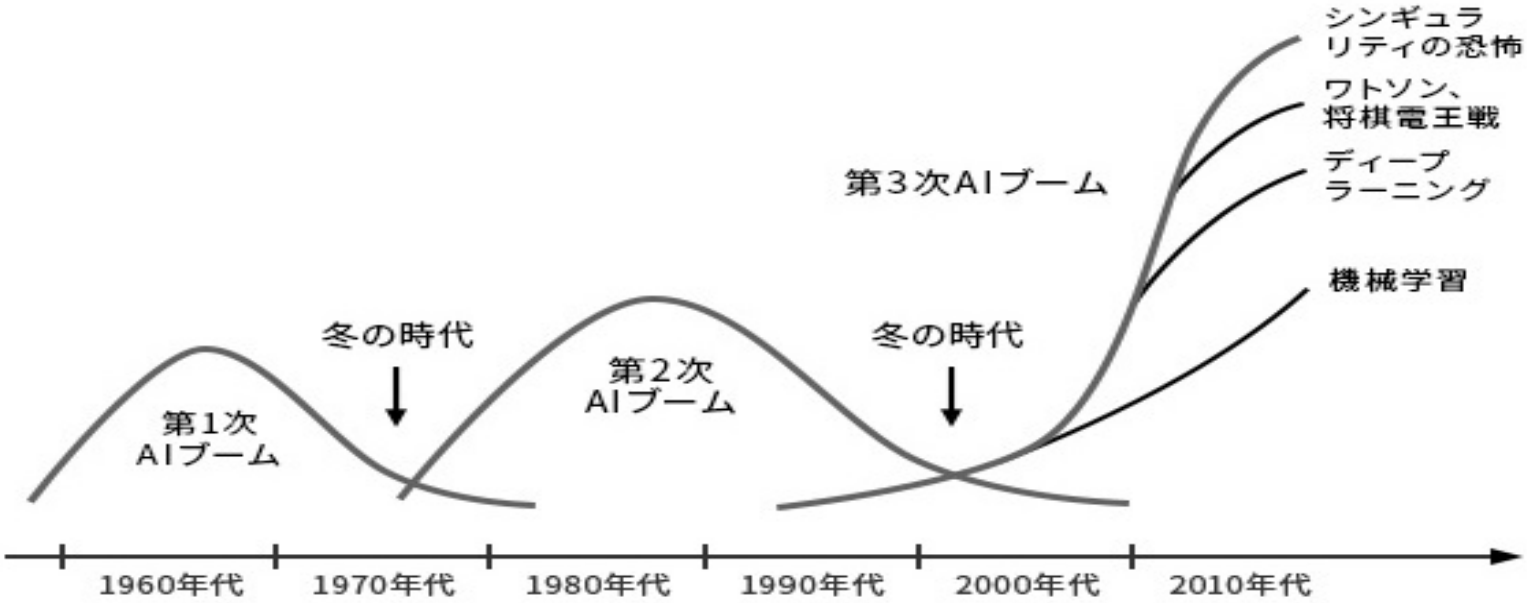
- $S(t)$: 未感染者率
- $I(t)$: 患者の比率
- $R(t)$: 治った人（+亡くなった人）の比率
- $\beta(t)$: 一日に一人の感染者が他人に感染させる人数
- D: 感染者が他人に感染させる期間 (約 15 日)
- $S(t) + I(t) + R(t) = 1$

コロナ陽性者予測 II


コロナ陽性者数
0
50
100
150
200
20年01月01日
20年01月06日
20年01月11日
20年01月16日
20年01月21日
20年01月26日
20年01月31日
20年02月05日
20年02月10日
20年02月15日
20年02月20日
20年02月25日
20年03月01日
20年03月06日
20年03月11日
20年03月16日
20年03月21日
20年03月26日
20年03月31日
20年04月05日
20年04月10日
20年04月15日
20年04月20日
20年04月25日
20年04月30日
20年05月05日
20年05月10日
20年05月15日
20年05月20日
20年05月25日
20年05月30日

発表内容

人工知能とは？

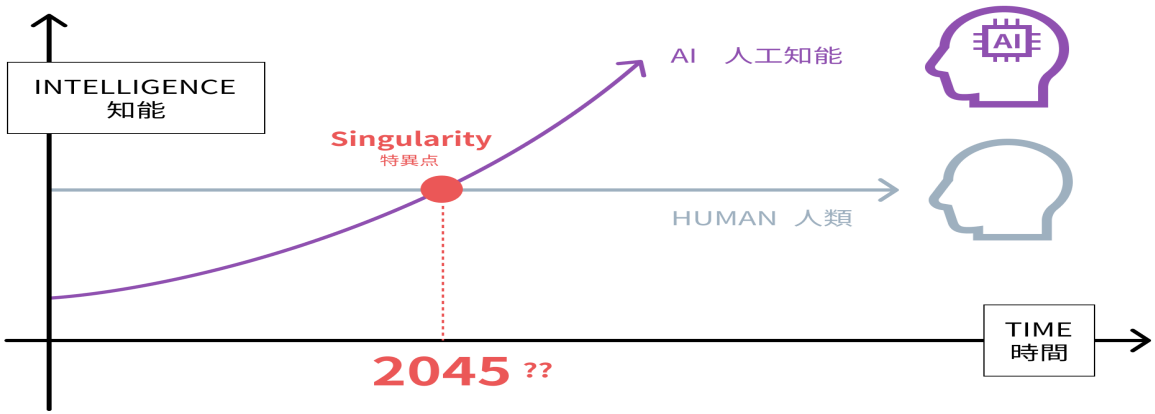


引用：「人工知能は人間を超えるか　ディープラーニングの先にあるもの」松尾豊　著

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」

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.

指数関数的思考 (Exponential mindset)

Which Mindset Are You Using?

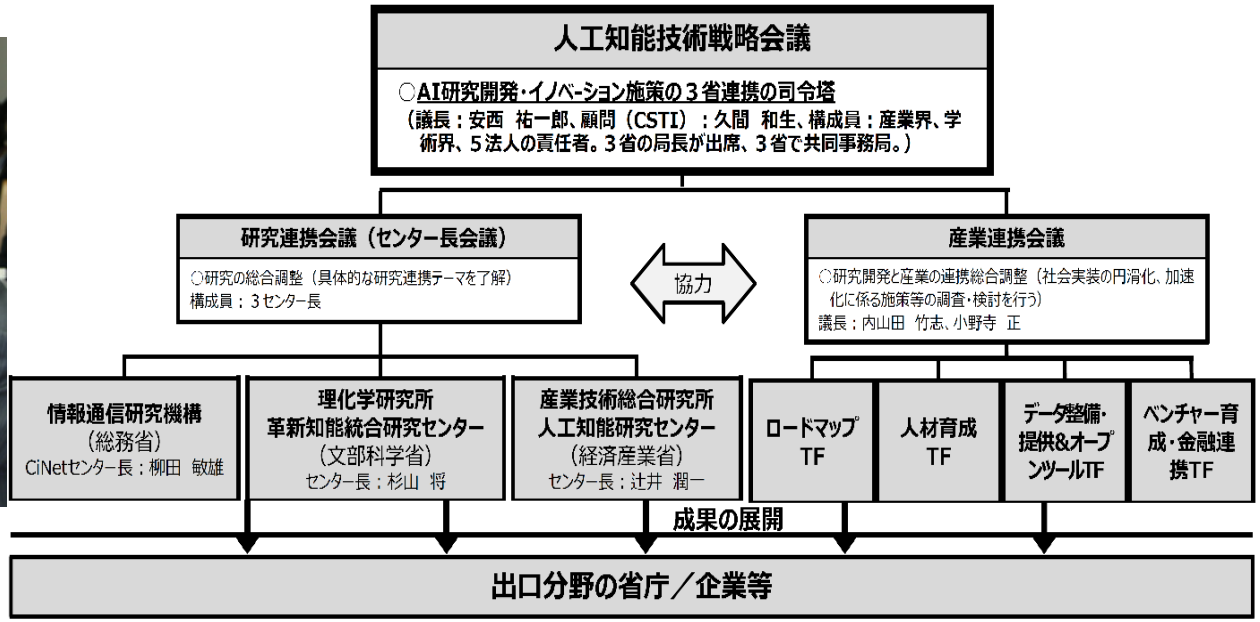
Incremental Mindset – 10%	Exponential Mindset – 10x
Set realistic goals	Set **ambitious** goals
Follow the plan	Follow the **vision**
Minimize risk	Maximize **learning**
Standardize	**Personalize**
Centralize decision-making	**Empower** decision-making
Expand authority	Expand **influence**
Make your 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、産学官の叡智を集め、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．

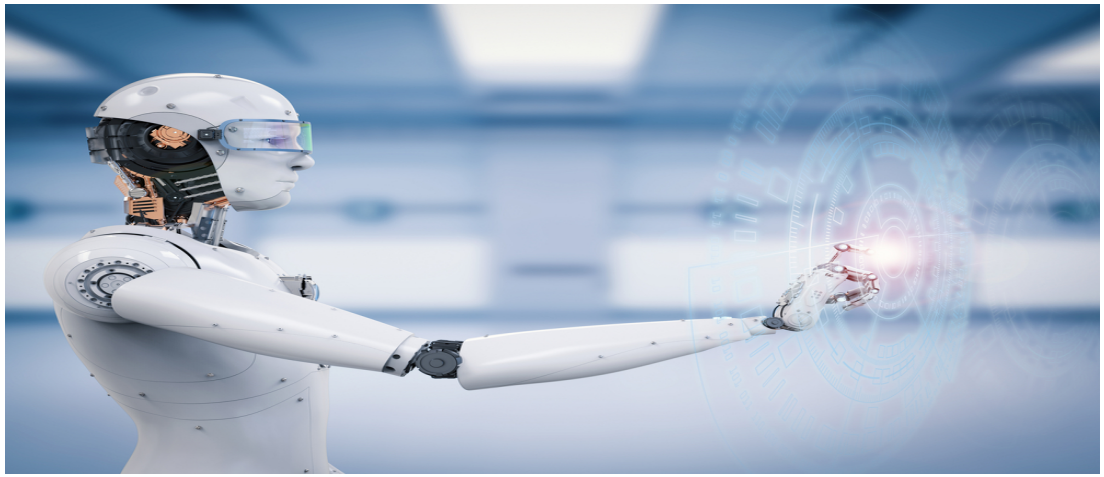
AI 技術戦略会議 II

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

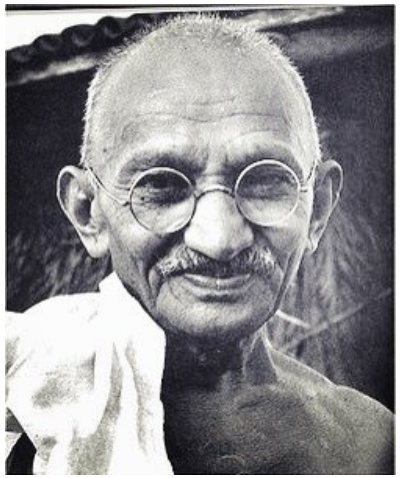
- 弱い AI: 特定のことしか出来ない AI. 特化型人工知能. 与えられた課題に対して優れた性能を発揮する．AlphaGo, IBM Watson, 自動運転技術
- 強い AI: あたかも人間のような自意識を備えている AI. 汎用性と自律性がある．汎用人工知能．鉄腕アトム，ドラえもん

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.

人間性と科学



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．衒学者,..
- 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

- Kanao, K., Komori, O., Nakashima, J., Ohigashi, T., Kikuchi, E., Miyajima, A., Nakagawa, K., Eguchi, S. and Oya, M. (2014). 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.
- Komori, O. (2011). A boosting method for maximization of the area under the ROC curve. *Annals of the Institute of Statistical Mathematics* **63**, 961–979.
- 柴田里程, . (2015). データ分析とデータサイエンス, 東京: 近代科学社.
- 渋谷政昭, and 柴田里程, . (1992). *S* によるデータ解析, 東京: 共立出版.
- 八木光晴, and 及川信, . (2008). 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.