データサイエンス
なにそれおいしいの？

KOF2021

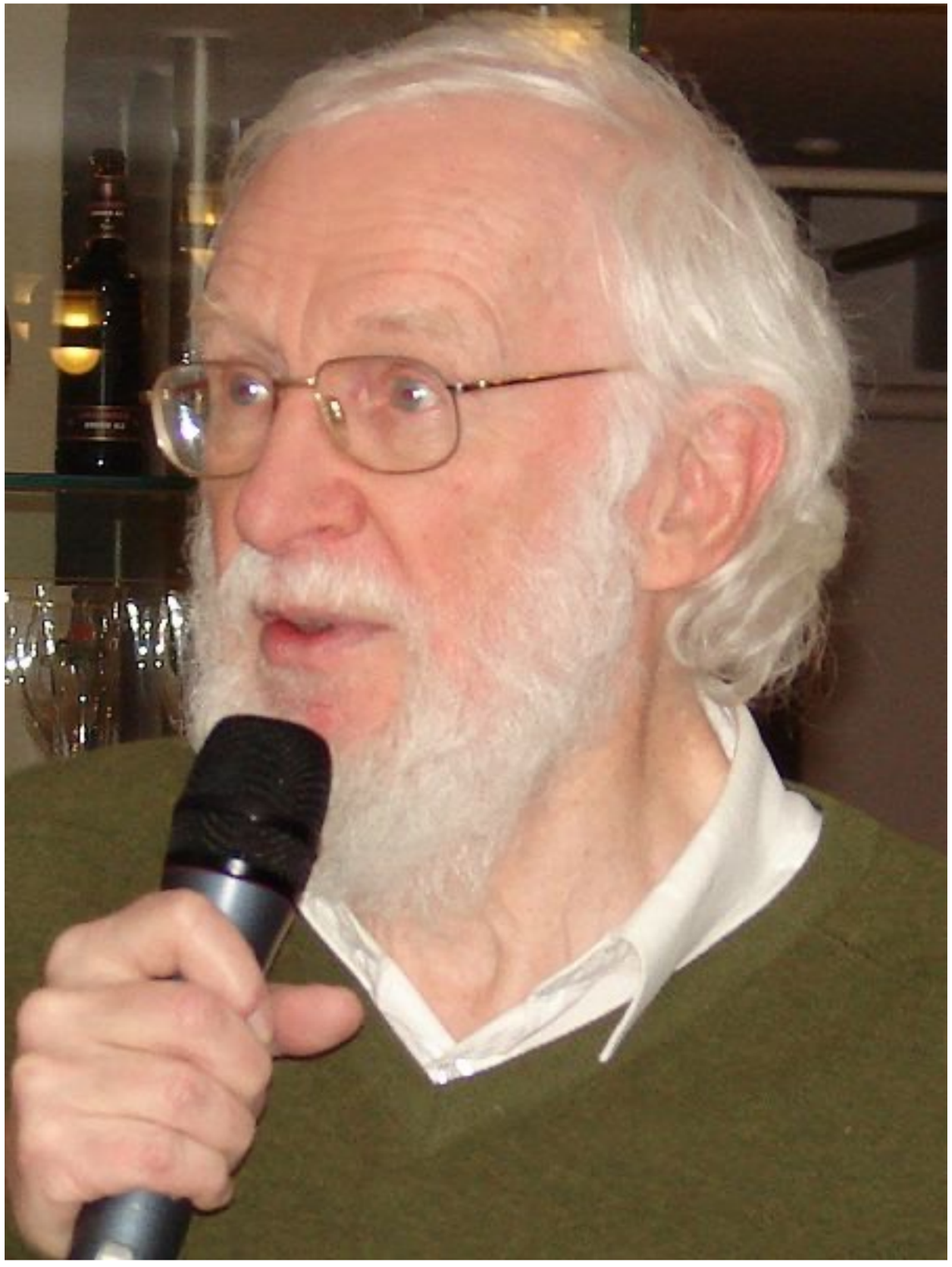
2021-11-13 14:00-14:50

奥村晴彦@三重大学教育学部特任教授　Twitter: h_okumura

To understand God's thoughts we must study statistics, for these are the measure of his purpose.

Florence Nightingale
1820–1910

Hal Varian (2008):
the sexy job in the next ten years will be statisticians



Peter Naur (1928–2016)

The Science of Datalogy

EDITOR:

This is to advocate that the following new words, denoting various aspects of our subject, be considered for general adoption (the stress is shown by an accent):

datálogy, the science of the nature and use of data,

datum&tics, that part of datalogy which deals with the processing of data by automatic means,

datámaton, an automatic device for processing data.

In this terminology much of what is now referred to as "data processing" would be datamatics. In many cases this will be a gain in clarity because the new word includes the important aspect of data representations, while the old one does not. Datalogy might be a suitable replacement for "computer science."

The objection that possibly one of these words has already been used as a proper name of some activity may be answered partly by saying that of course the subject of datamatics is written with a lower case d, partly by remembering that the word "electronics" is used doubly in this way without inconvenience.

What also speaks for these words is that they will transfer gracefully into many other languages. We have been using them extensively in my local environment for the last few months and have found them a great help.

Finally I wish to mention that datamatics and datamaton (Danish: datamatik and datamat) are due to Paul Lindgreen and Per Brinch Hansen, while datalogy (Danish: datalogi) is my own invention.

PETER NAUR
A/S Regnecentralen
Falkoner Alle 1
Copenhagen F, Denmark

Peter Naur
"The Science of Datalogy"
Letters to the Editor
Communications of the ACM
Vol. 9, No. 7, July 1966

large industrial machinery, where data processing is used extensively to prepare

Peter Naur, "Data and their applications" (1974), in *Computing, A Human Activity* (1992)

seems too artificial for the taste of the reader, he may of course add the control of machinery to the list of sensible uses of the data produced in data processing.

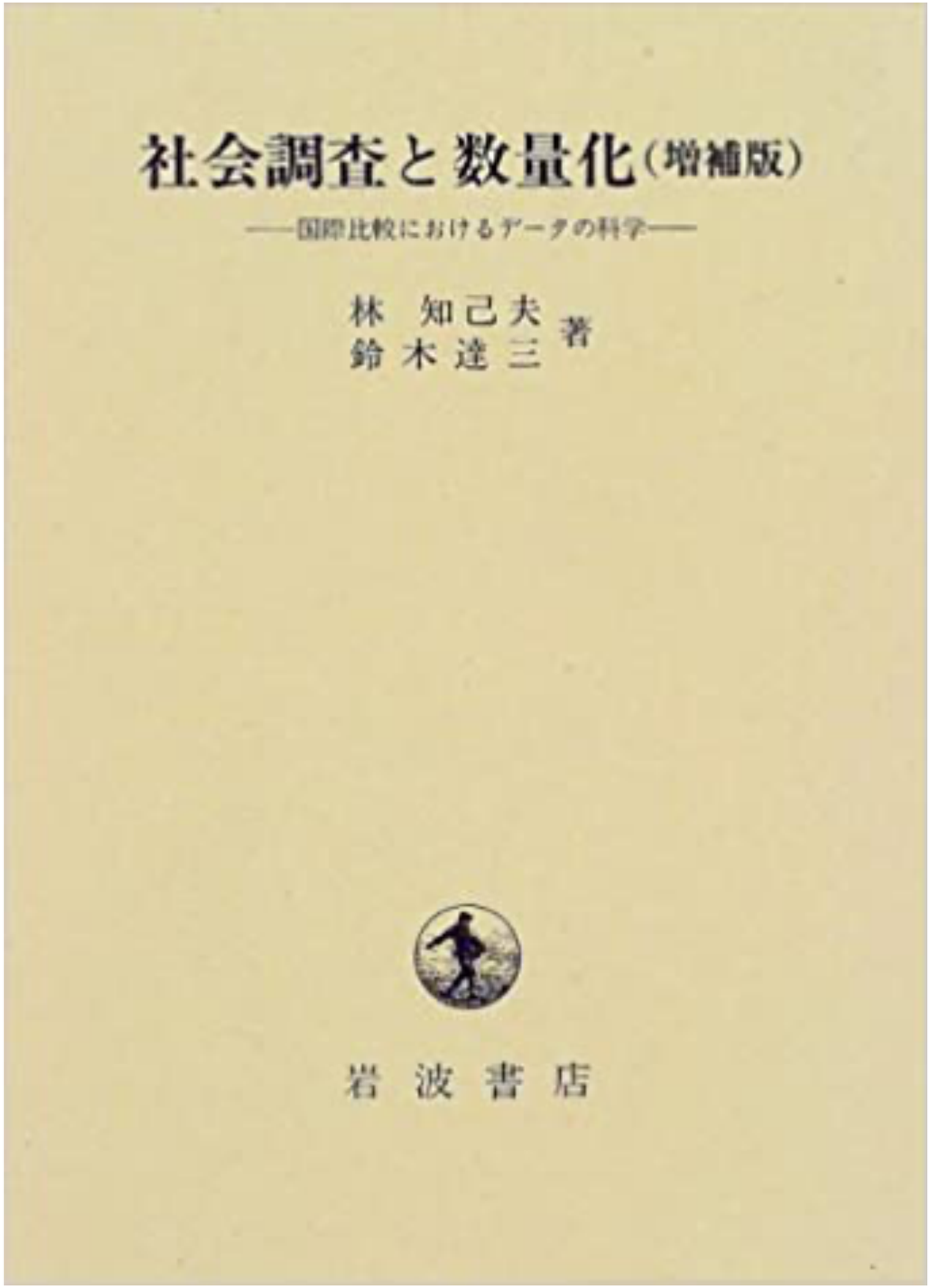
1.2.8 A Basic Principle of Data Science

A basic principle of data science, perhaps the most fundamental that may be formulated, can now be stated:

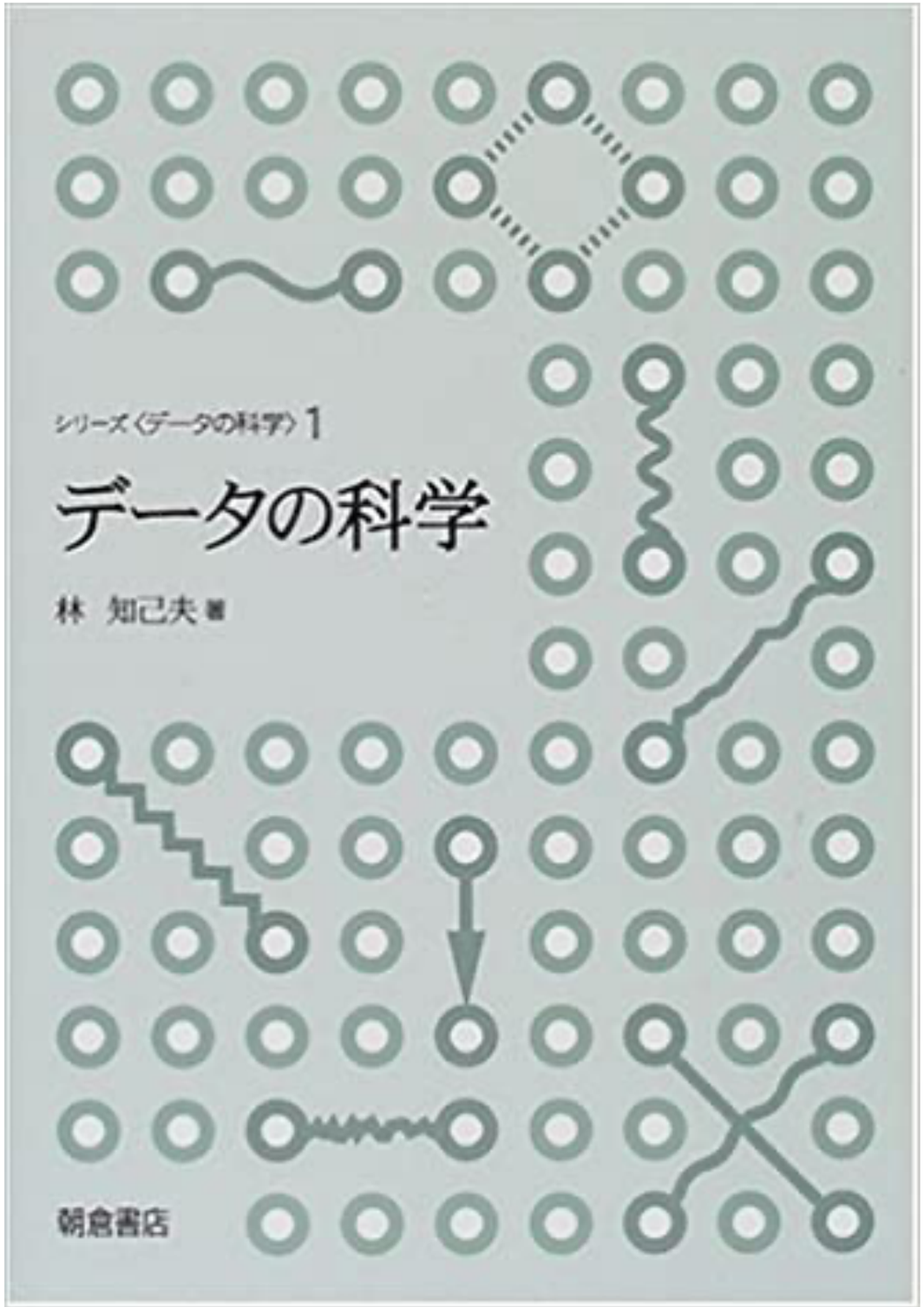
The data representation must be chosen with due regard to the transformation to be achieved and the data processing tools available.

Three remarks are relevant: (1) Since data science is concerned with methods of construction of data processes, it is consistent that its basic principles come as design guidance. (2) The principle is consistent with the idea of the freedom to choose the data representation. (3) The regard to the data processing tools is consistent with the notion that basically data are things to be processed.

Examples of applications of this principle will appear again and again in the following chapters of this text.



林知己夫・鈴木達三『社会調査と数量化（増補版）——国際比較におけるデータの科学——』(1997)



林知己夫

『データの科学』

（2001年）

William S. Cleveland (1943–)

International Statistical Review (2001), **69**, 1, 21–26, Printed in Mexico

Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Laboratories, 600 Mountain Avenue, Murray Hill NJ07974, USA
E-mail: wsc@research.bell-labs.com

Summary

An action plan to enlarge the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department, and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations.

Key words: Future; Applications; Computing; Methods; Models; Theory.

1 Summary of the Plan

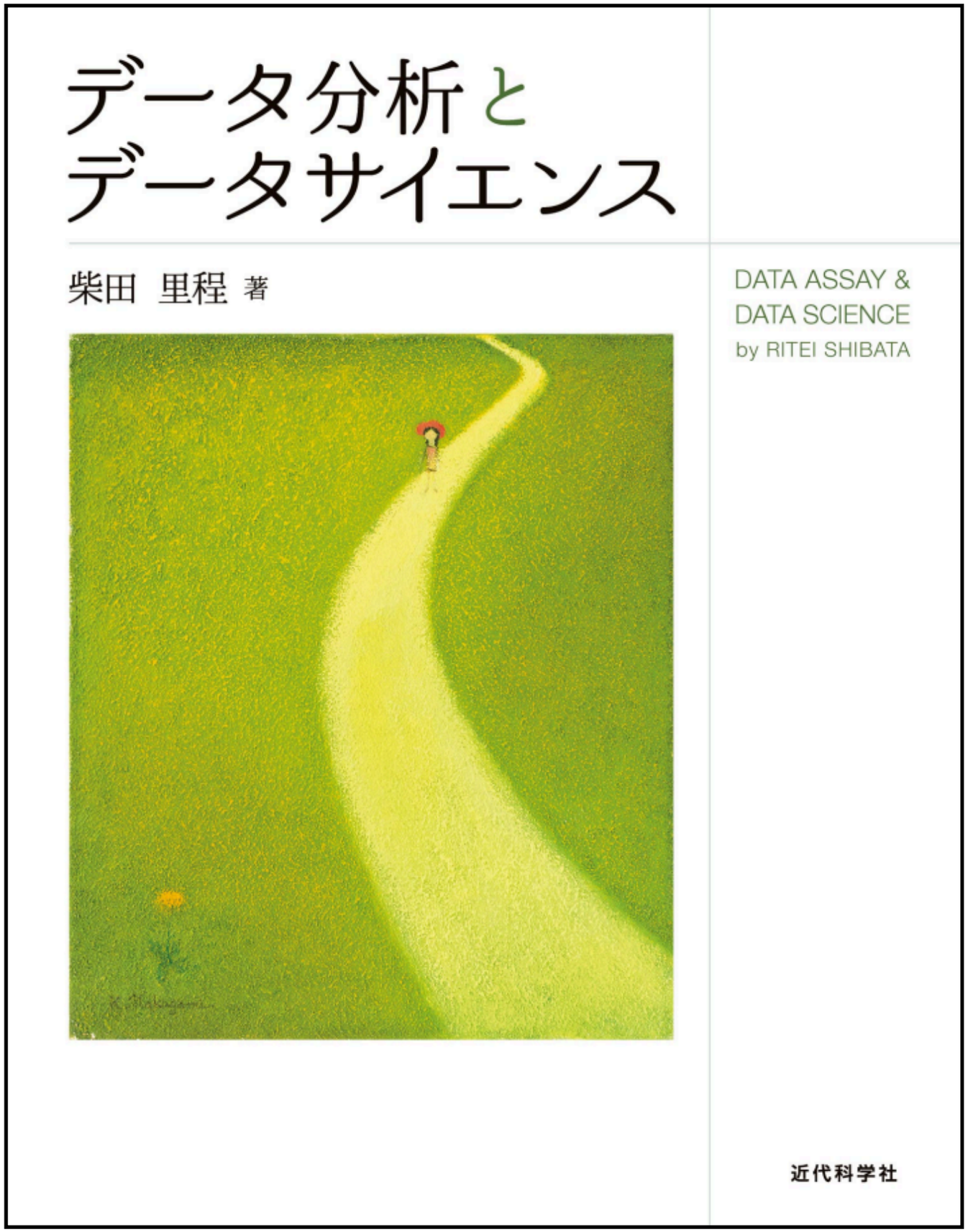
This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called "data science".

The focus of the plan is the practicing data analyst. A basic premise is that technical areas of data science should be judged by the extent to which they enable the analyst to learn from data. The benefit of an area can be direct or indirect. Tools that are used by the data analyst are of direct benefit. Theories that serve as a basis for developing tools are of indirect benefit. A broad successful theory can have a wide-ranging benefit, affecting data analysis in a fundamental way. For example, the Bayesian theory of statistics affects all methods of estimation and distribution.

The plan sets out six technical areas for a university department, and advocates a specific allocation of resources to research and development in each area as a percent of the total resources that are available beyond those needed to teach the courses in the department's curriculum. Furthermore, the allocation applies to the make-up of the curriculum; that is, the allocations are a guideline for the percentage of courses in each of the technical areas. The six areas and their percentages are the following:

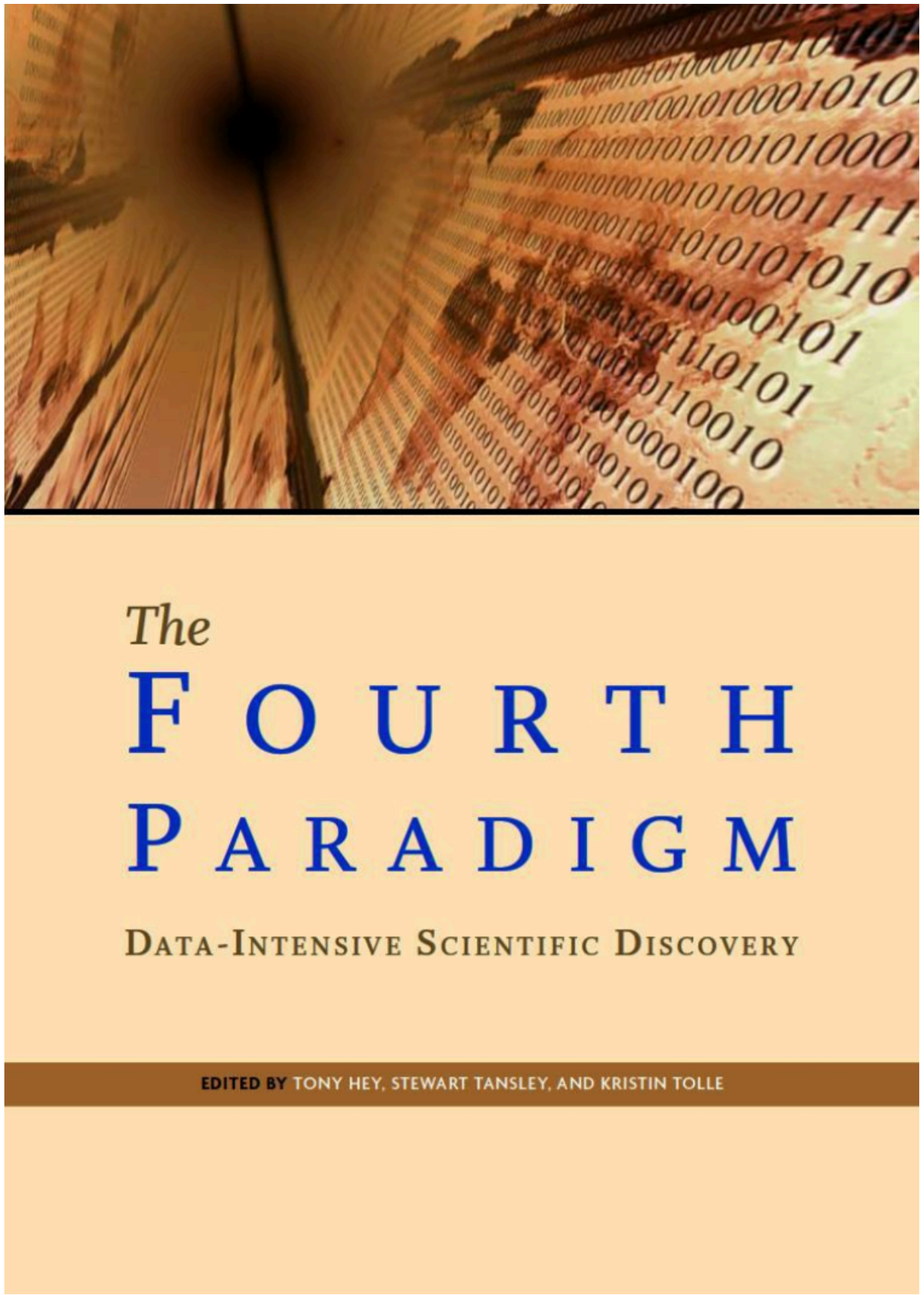
- **(25%) Multidisciplinary Investigations**: data analysis collaborations in a collection of subject matter areas.
- **(20%) Models and Methods for Data**: statistical models; methods of model building; methods of estimation and distribution based on probabilistic inference.
- **(15%) Computing with Data**: hardware systems; software systems; computational algorithms.
- **(15%) Pedagogy**: curriculum planning and approaches to teaching for elementary school, secondary school, college, graduate school, continuing education, and corporate training.
- **(5%) Tool Evaluation**: surveys of tools in use in practice, surveys of perceived needs for new

William S. Cleveland (2001) Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics



柴田里程『データ分析とデータサイエンス』(2015) p.77 傍注

17) W.S. Cleveland はデータサイエンスのアイディアを，1996 年に国際分類学会出席のため来日したとき，著者から得たようである．このとき著者は同時開催の日本統計学会年会で共通テーマとして「データサイエンス I, II, III」を企画していた．



The Fourth Paradigm:
Data-Intensive Scientific Discovery
(2009)

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

学習指導要領

中学校（2008年告示，2012年度～全面実施）
「資料の活用」返り咲く（資料＝データ）

高校（2009年告示，2012年度～学年進行）
必履修の「数学I」に「データの分析」が入る

2016年の入試センター試験

数学I（全問必答）

13ページ中7ページが統計分野

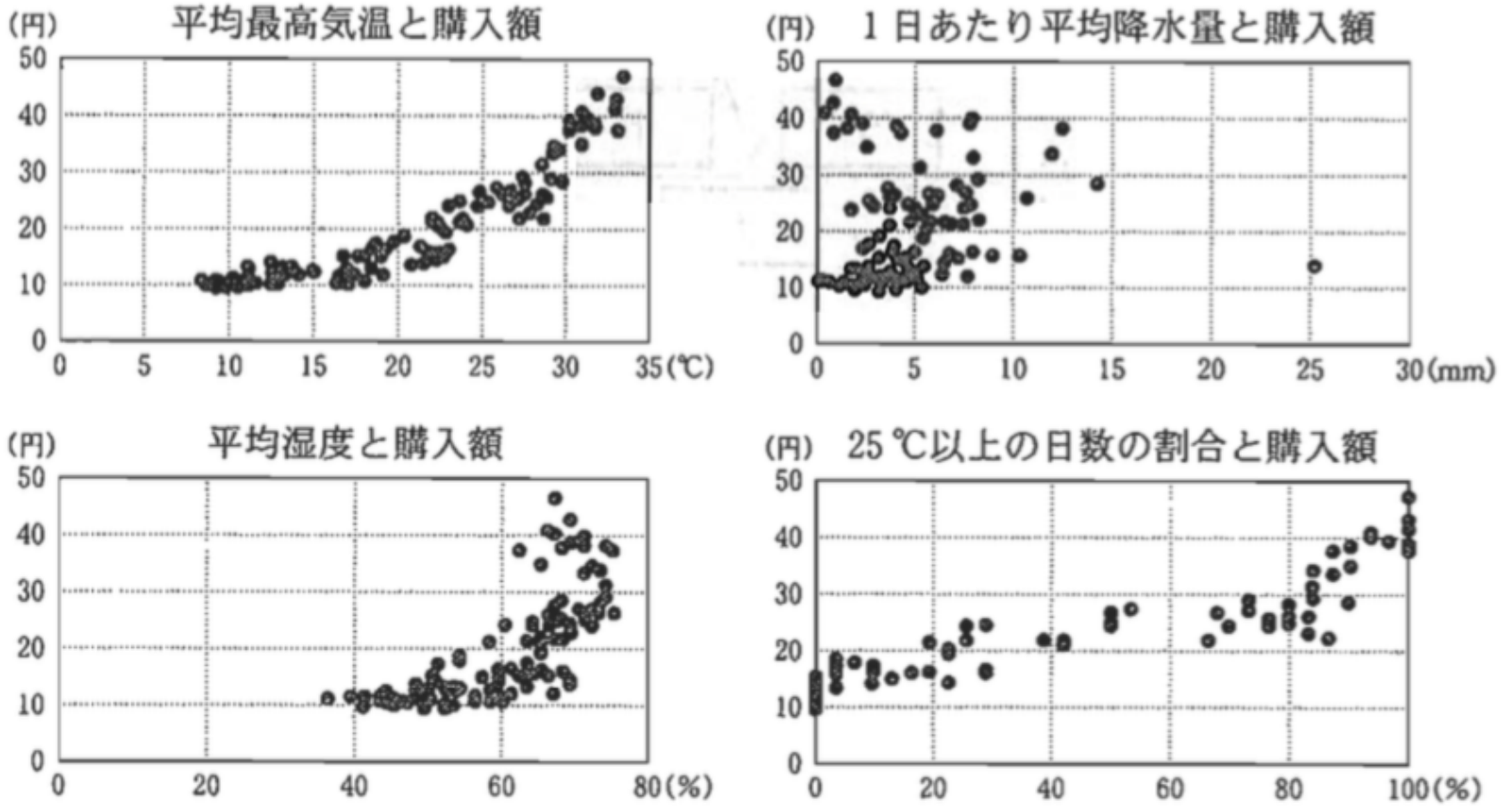
数学I・A

必答問題8ページ中4ページが統計分野

選択問題（3問中2問）の一つは確率

第4問 (配点 20)

〔1〕 次の4つの散布図は，2003年から2012年までの120か月の東京の月別データをまとめたものである。それぞれ，1日の最高気温の月平均(以下，平均最高気温)，1日あたり平均降水量，平均湿度，最高気温25℃以上の日数の割合を横軸にとり，各世帯の1日あたりアイスクリーム平均購入額(以下，購入額)を縦軸としてある。

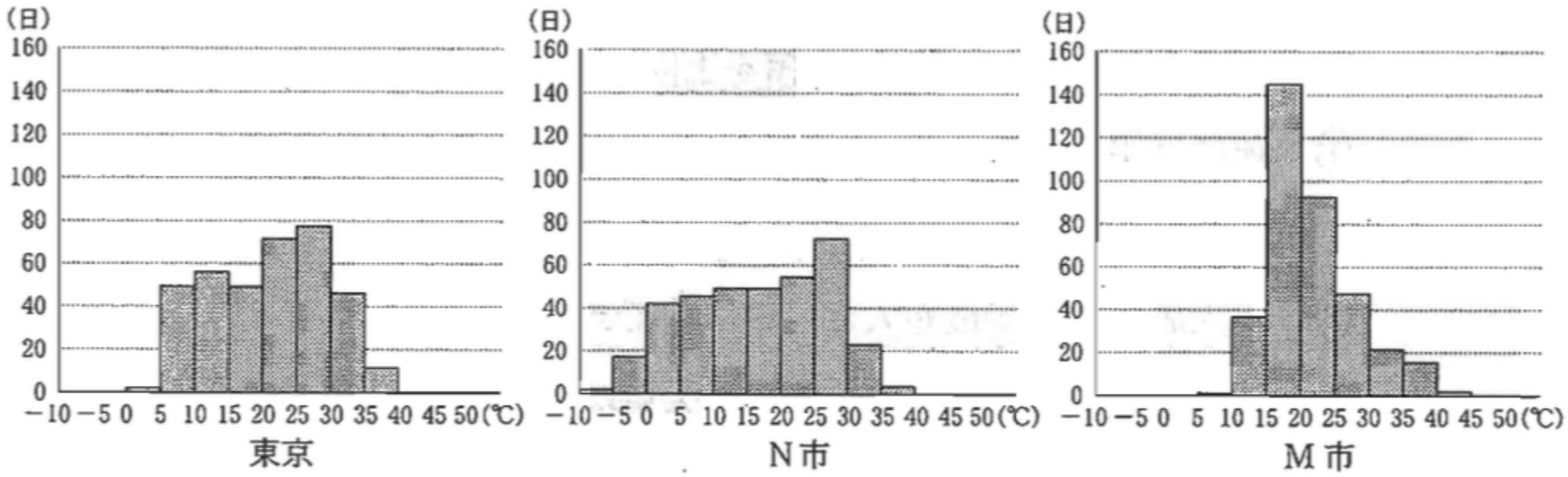


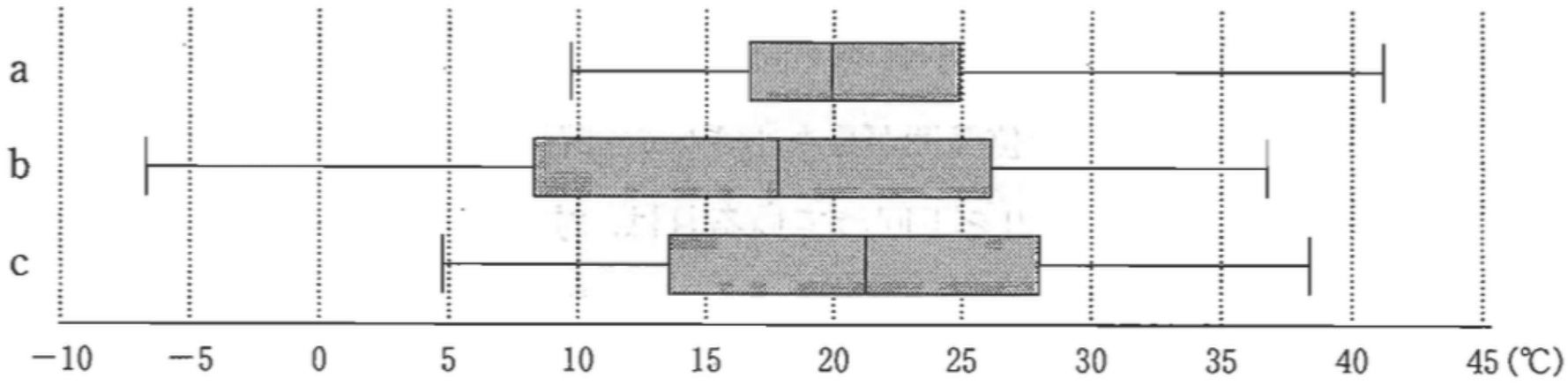
出典：総務省統計局(2013)『家計調査年報』，『過去の気象データ』(気象庁Webページ)などにより作成

(数学Ⅰ第4問は次ページに続く。)

〔2〕 世界4都市(東京，O市，N市，M市)の2013年の365日の各日の最高気温のデータについて考える。

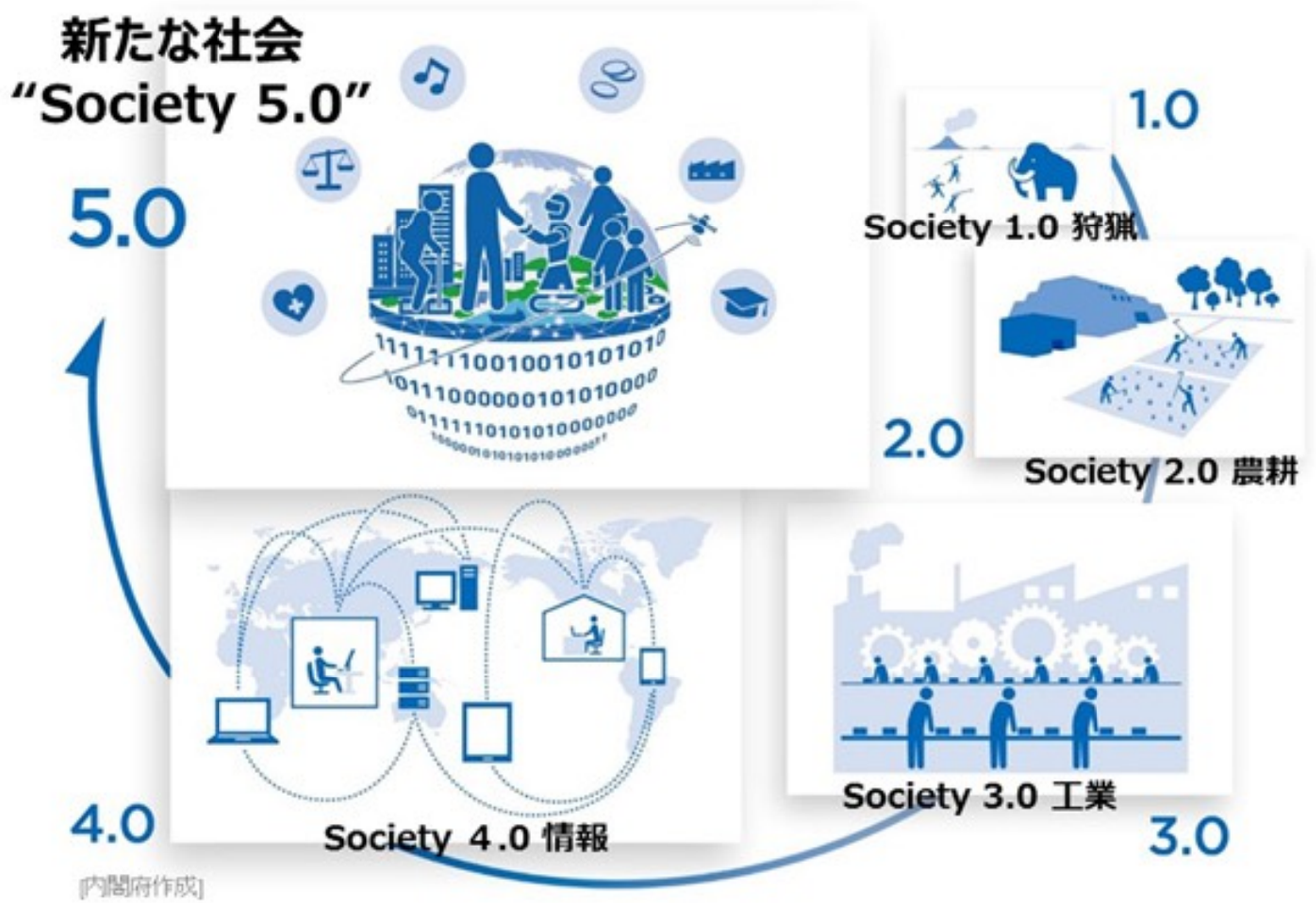
(1) 次のヒストグラムは，東京，N市，M市のデータをまとめたもので，この3都市の箱ひげ図は下のa，b，cのいずれかである。





出典：『過去の気象データ』(気象庁Webページ)などにより作成

第5期科学技術基本計画（2016年1月22日閣議決定）



Society 1.0 狩猟社会
Society 2.0 農耕社会
Society 3.0 工業社会
Society 4.0 情報社会
Society 5.0 サイバー空間（仮想空間）とフィジカル空間（現実空間）を高度に融合させたシステムにより、経済発展と社会的課題の解決を両立する、人間中心の社会

スマート社会計画「ソサエティー5.0」、肝はデータ

新井紀子 日経新聞2016-04-12

https://www.nikkei.com/article/DGXKZO99346090W6A400C1X12000/

……

私が参加しているシステム基盤技術検討会では、その実現のための基盤となる仕組みを検討している。例えば、3次元地図情報。全地球測位システム（GPS）の盲点となっている地下街の地図も含めて、機械可読（機械が意味を理解できるよう）な地図情報を整備することにより、震災やテロの中でも、スマートフォンや電動車椅子を通じて人々を安全に誘導することが期待されている。

この話を動かすための肝は何か。つい技術革新に目が向かいがちだが、それよりも大切なことがある。それはデータである。出発点になるデータが集まらなければこの話は絵に描いた餅になる。

集まってきたデータは、人ではなく機械に処理させるので、機械可読な形式に整っていなければ意味がない。つまり、データの量と質を確保できるか否かに、この計画の成否がかかっているのである。

2016年8月：文部科学省に「数理及びデータサイエンス教育の強化に関する懇談会」が設置される

2016年12月：「大学の数理・データサイエンス教育強化方策について」が公表される

2016年12月：「数理及びデータサイエンスに係る教育強化」の拠点校（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

うちが騒ぎ出したのは2018年10月・・・

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が未来社会を拓く

数理・データサイエンス教育強化を目的として国立大学に設置されたセンターが結集して，各大学内での数理・データサイエンス教育の充実のための取組成果を全国への波及させるための活動を推進し，数理・統計・情報を基盤として未来世界を開拓できる人材の育成を目指します.

「数理・データサイエンスと大学」インタビュー

統計的問題解決力をどう育むか　カギを握る多様な「経験価値」

第14回
立正大学データサイエンス学部教授
渡辺 美智子氏

小学校　プログラミング

小・中学校　1人1台端末

小学校学習指導要領（平成 29 年告示）解説

算数編

平成 29年 7月

(3)「Dデータの活用」の領域で育成を目指す資質・能力

①目的に応じてデータを収集，分類整理し，結果を適切に表現すること

統計的な問題解決活動

目的に応じてデータを収集，分類整理し，結果を適切に表現するとは，統計的な問題解決活動を指しているが，統計的な問題解決活動においては，「問題－計画－データ－分析－結論」というような段階からなる統計的探究プロセスと呼ばれるものがある。

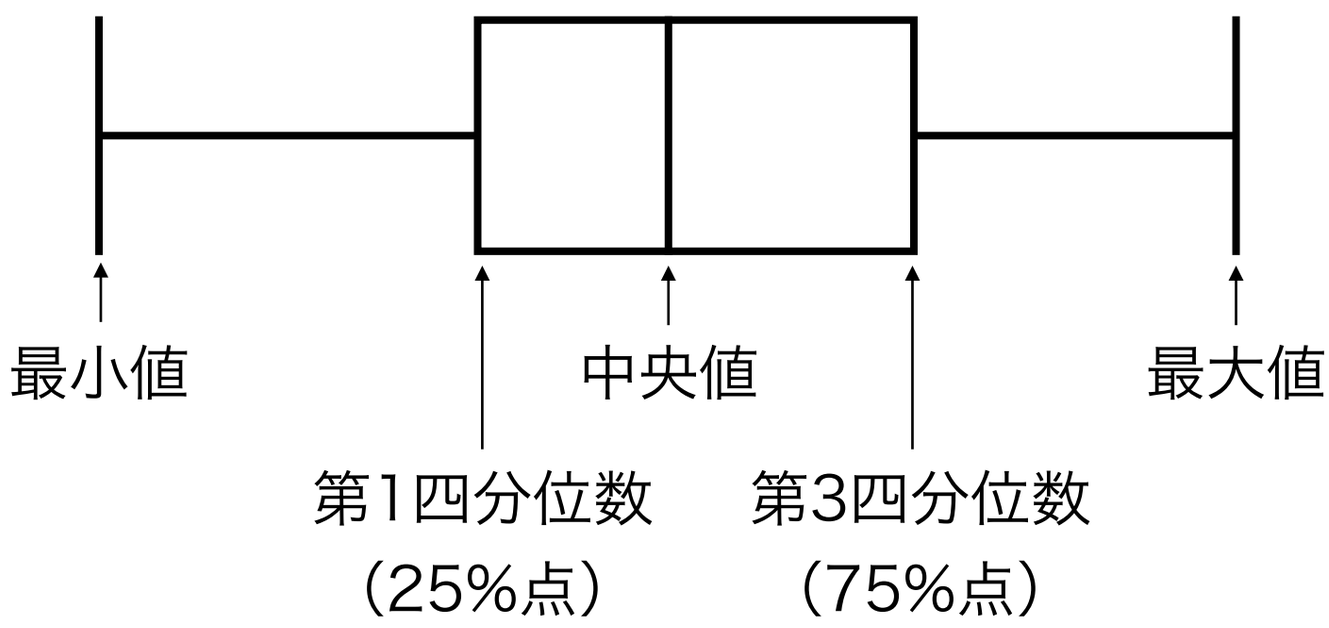
PPDACサイクル！

問題	・問題の把握	・問題設定
計画	・データの想定	・収集計画
データ	・データ収集	・表への整理
分析	・グラフの作成	・特徴や傾向の把握
結論	・結論付け	・振り返り

新学習指導要領　中学校

・資料の活用→データの活用
・四分位範囲・箱ひげ図が高校「数学I」から中学2年に降りてきた

箱ひげ図



2022年度からの学習指導要領　高校数学

現行	
数学III	**（数学活用）**
平面上の曲線と複素数平面	数学と人間の活動
極限	社会生活における数理的な考察
微分法	
積分法	
数学II	**数学B**
いろいろな式	確率分布と統計的な推測
図形と方程式	数列
指数関数・対数関数	ベクトル
三角関数	
微分・積分の考え	
数学I（必履修）	**数学A**
数と式	場合の数と確率
図形と計量	整数の性質
二次関数	図形の性質
データの分析	

改訂案	
数学III	**数学C**
極限	ベクトル
微分法	平面上の曲線と複素数平面
積分法	数学的な表現の工夫
数学II	**数学B**
いろいろな式	数列
図形と方程式	統計的な推測
指数関数・対数関数	数学と社会生活
三角関数	
微分・積分の考え	
数学I（必履修）	**数学A**
数と式	図形の性質
図形と計量	場合の数と確率
二次関数	数学と人間の活動
データの分析	

2022年度からの学習指導要領　高校情報

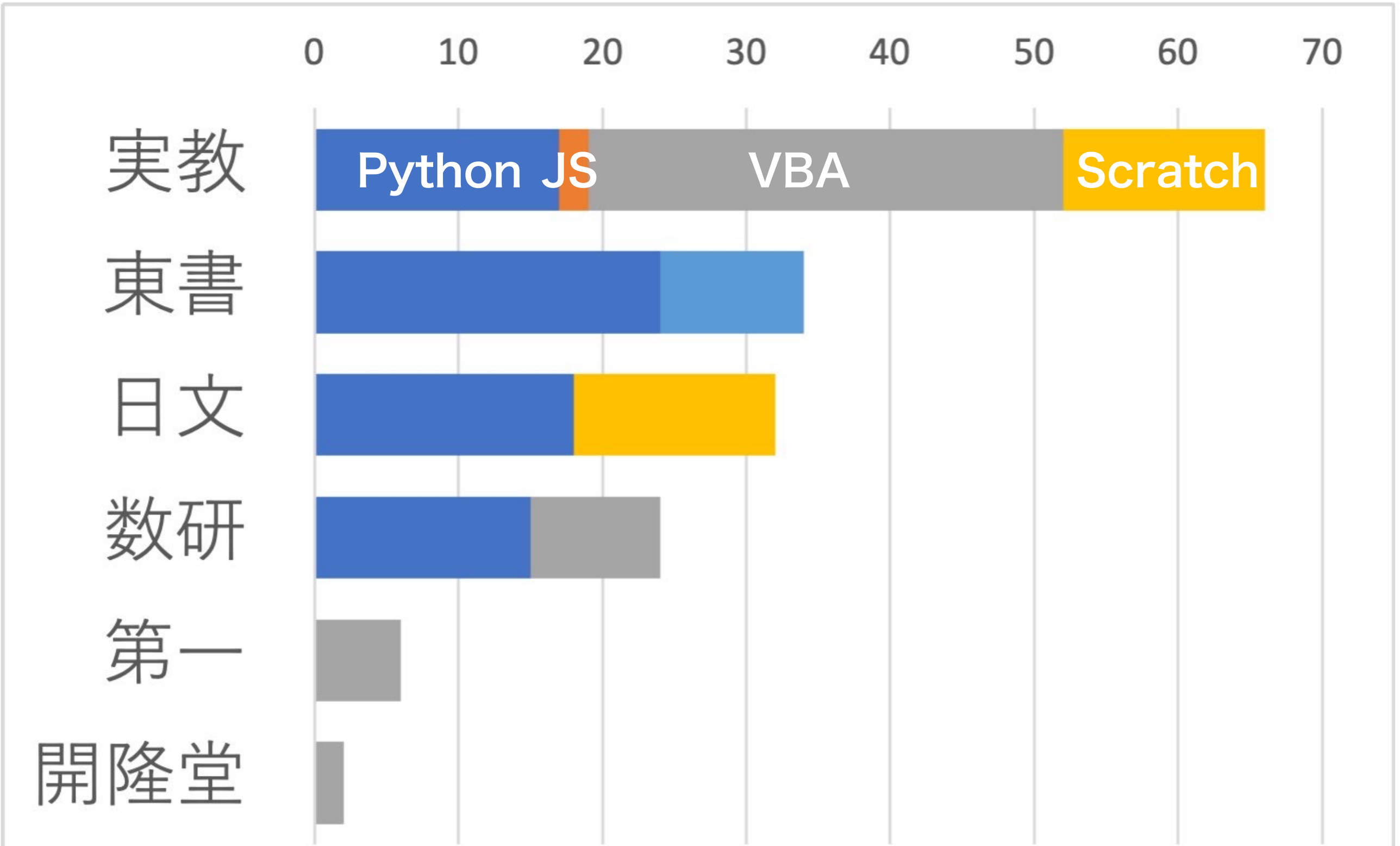
情報I（必履修）

(1) 情報社会の問題解決
(2) コミュニケーションと情報デザイン
(3) コンピュータとプログラミング
(4) 情報通信ネットワークと**データの活用**

情報II（選択）

(1) 情報社会の進展と情報技術
(2) コミュニケーションとコンテンツ
(3) 情報と**データサイエンス**
(4) 情報システムとプログラミング

東京都立高校情報I教科書別学校数



データサイエンス・カリキュラム標準（専門教育レベル）の公開

データサイエンス・カリキュラム標準（専門教育レベル）の公開

2021年4月15日

報道関係者各位
プレスリリース

データサイエンス・カリキュラム標準（専門教育レベル）の公開

一般社団法人 情報処理学会

一般社団法人情報処理学会（会長：江村 克己）は，データサイエンス分野における大学レベルの専門教育を対象としたカリキュラム標準の策定を進めています．このたび，同カリキュラム標準がまとまりましたので，公開いたします．

政府が推進しているAI戦略2019や欧州EDISON Data Science Framework等にも見られるように，データサイエンス教育やデータサイエンティストの育成は社会的にも重要性が高いことが認識されています．情報処理学会のデータサイエンス・カリキュラム標準（専門教育レベル）は，関連する様々な取り組みを参照して策定されており，以下に挙げる様々な特徴を持っています．

・ ACM Data Scienceカリキュラムおよび欧州EDISON Data Science Frameworkの参照を通じて，国際的通用性を確保する．

2019年4月18日　第43回総合科学技術・イノベーション会議

https://www8.cao.go.jp/cstp/siryo/haihui043/haihu-043.html

資料1

AI 戦略（人材育成関連）

平成３１年４月18日

内閣府特命担当大臣（科学技術政策）平井卓也

AI 時代に求められる人材育成に関する主な取り組み

デジタル社会の「**読み・書き・そろばん**」である「**数理・データサイエンス・AI**」の基礎などの必要な力を**全ての国民**が育み、あらゆる分野で人材が活躍

主な取組

エキスパート

先鋭的な人材を発掘・伸ばす環境整備

若手の自由な研究と海外挑戦の機会を拡充

実課題をAIで発見・解決する学習中心の**課題解決型AI人材育成**

応用基礎

AI応用力の習得

AI×専門分野のダブルメジャーの促進

AIで地域課題等の解決ができる人材育成（産学連携）

認定制度・資格の活用

大学等の**優れた教育プログラムを政府が認定**する制度構築

国家試験（ITパスポート）の見直し、高校等での活用促進

リテラシー

学習内容の強化

大学の標準カリキュラムの開発と展開（**MOOC※活用**等）

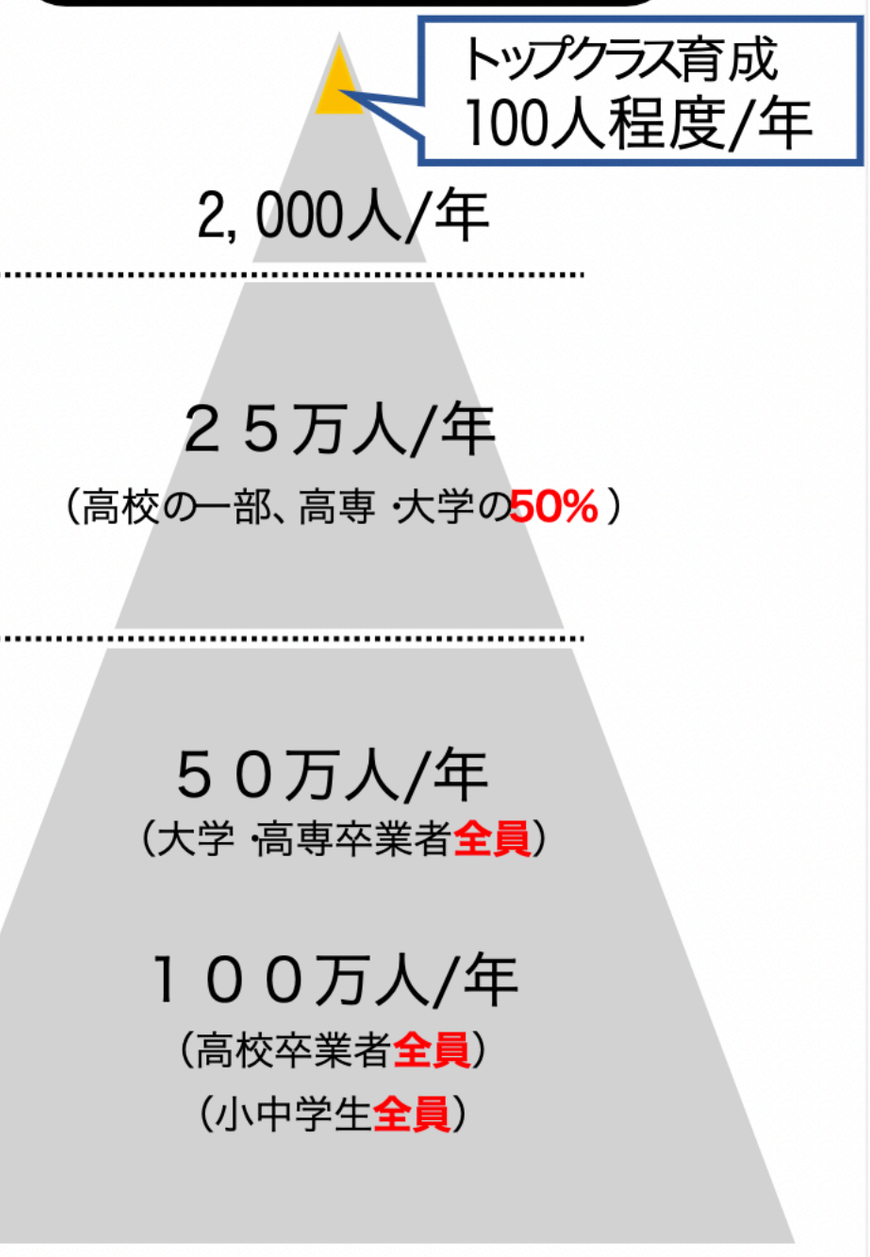
高校における**AI**の基礎となる**実習授業**の充実

小中高校における教育環境の整備

多様な**ICT人材の登用**（高校は1校に1人以上、小中校は4校に1人以上）

生徒一人一人が端末を持つICT環境整備

育成目標【2025年】



※Massive Open Online Course：大規模公開オンライン講座

リテラシー教育

デジタル社会の「**読み・書き・そろばん**」である**数理・データサイエンス・AI**」の定着に向けて、**小学生から社会人まで**各段階において長期的に取り組む

認定制度・資格の活用

必要な素養・スキル（出口）に応じた**人材の質を担保**する仕組みを構築

小中学校

理数の興味関心向上

高校

全ての高校生が基礎的リテラシーを習得

大学入試

情報Ⅰ」の採用の拡大

大学・高専

初級レベルの数理・データサイエンス・AIを習得

社会人

AI活用スキル習得
就職・待遇への活用

初等中等教育、高等教育、リカレント教育の長期的取組

大学におけるリテラシー・応用基礎教育

2025年までに**全学生（50万人/年規模）**が**「数理・データサイエンス・AI」**の基礎を習得可能となる大学教育へ改革

AI ×専門分野のダブルメジャーの促進等による**AI を応用する基礎力の習得**

入試

- 大学入学共通テストでの**「情報Ⅰ」出題**
- 取組状況に応じた**インセンティブ**（助成金等）

学習

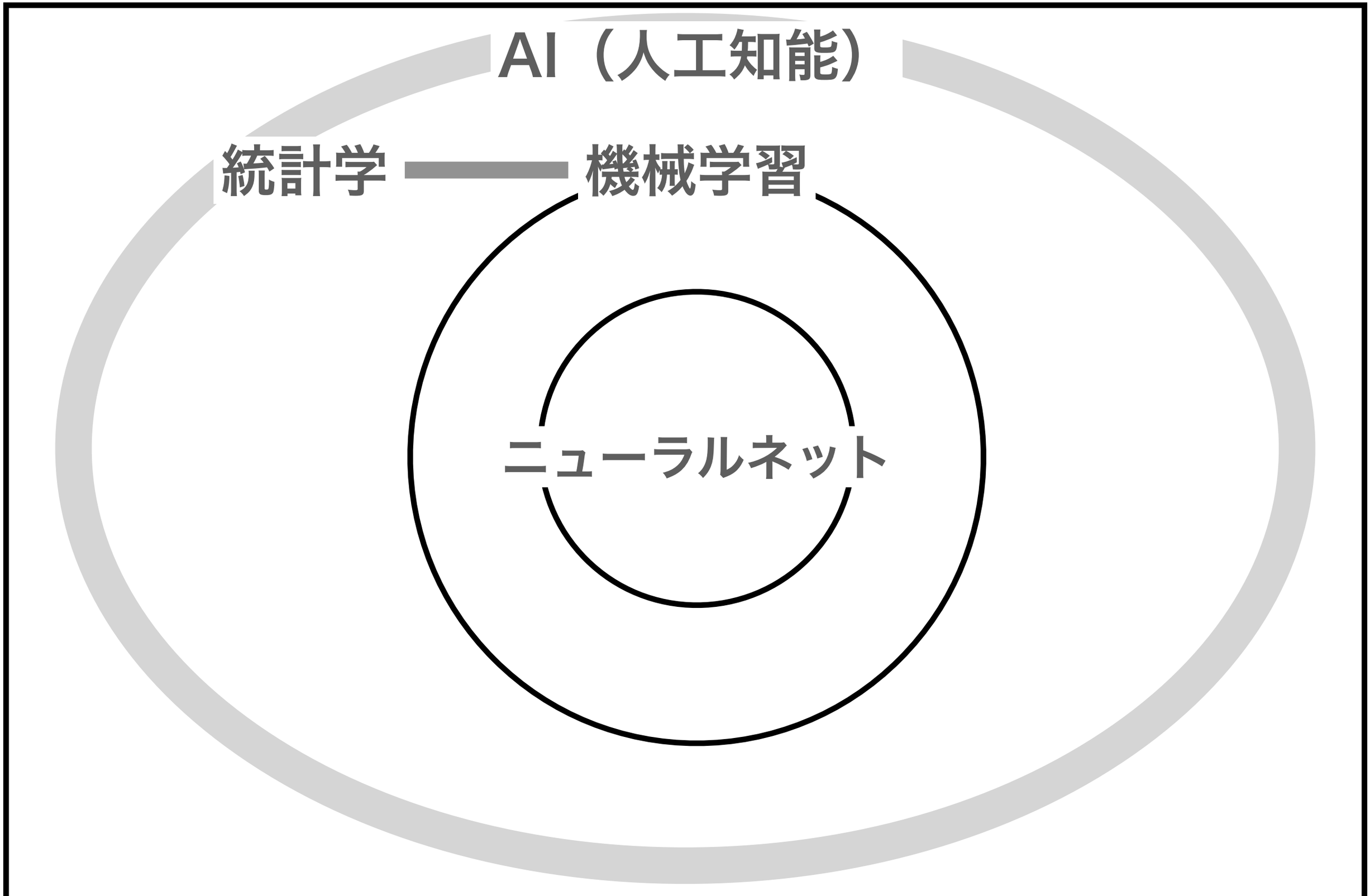
- 標準カリキュラム・教材・教育プログラムの**開発と全国展開**
- 優れたコースの**政府認定**
- 取組状況に応じた**インセンティブ**（運交金、助成金等）

教員

- **拠点大学**での**研修**
- **産業界**、研究機関等との**連携**（講師派遣、共同研究等）

環境

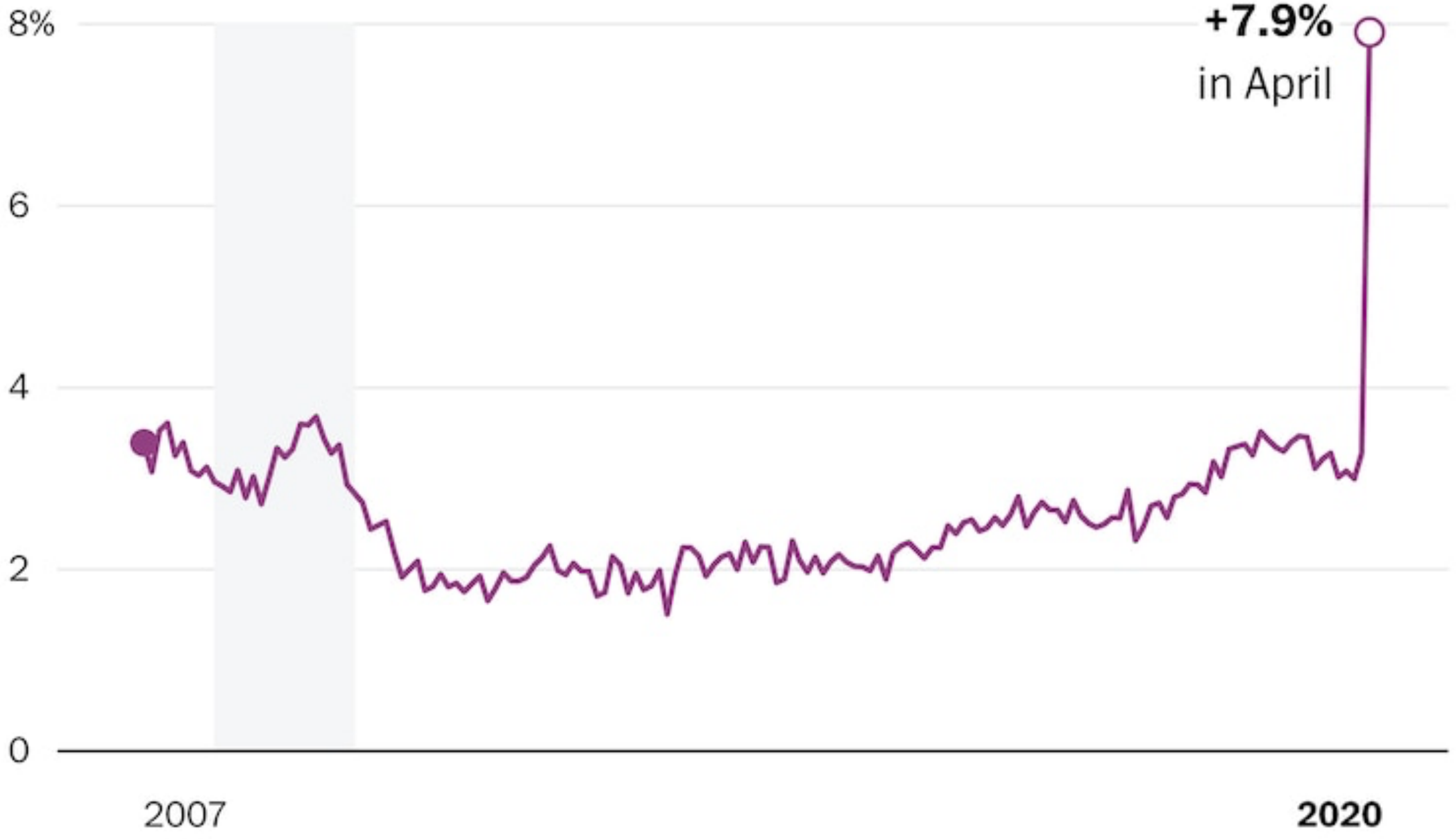
- **AI ×専門分野のダブルメジャー**を可能とする**制度改正**
- **MOOC**、放送大学等の**活用支援**
- 留学生の**受け入れ促進**


AI（人工知能）
統計学
機械学習
ニューラルネット
データサイエンス（学際的）

なにそれおいしいの？

——コロナ下のデータサイエンス

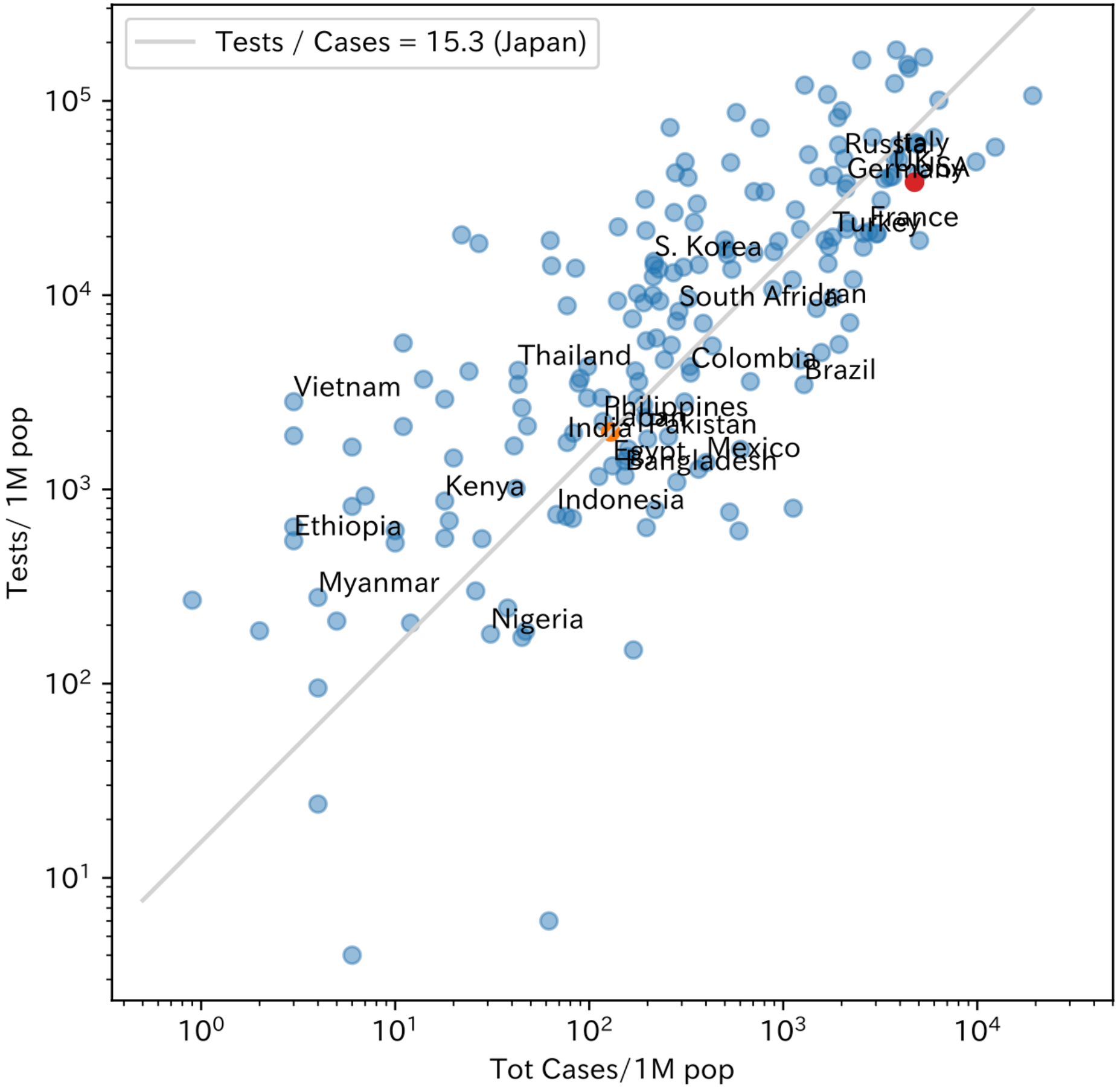
Average hourly earnings, change from a year earlier



Note: Seasonally adjusted; private sector only
Source: Labor Department
THE WASHINGTON POST

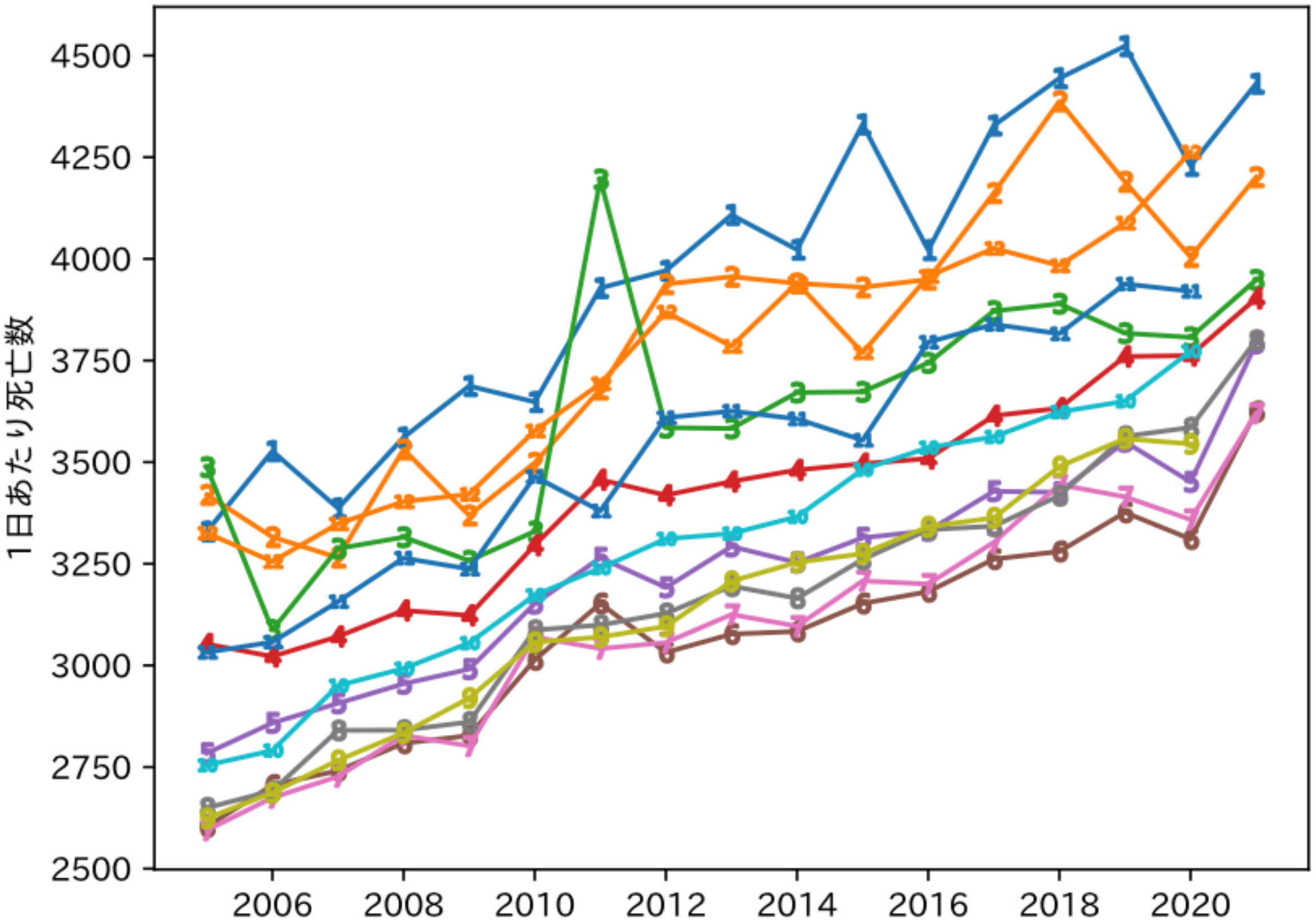
https://www.washingtonpost.com/business/2020/05/08/awful-reason-wages-appeared-soar-middle-pandemic/

generated: 2020-05-20 08:22:33 JST



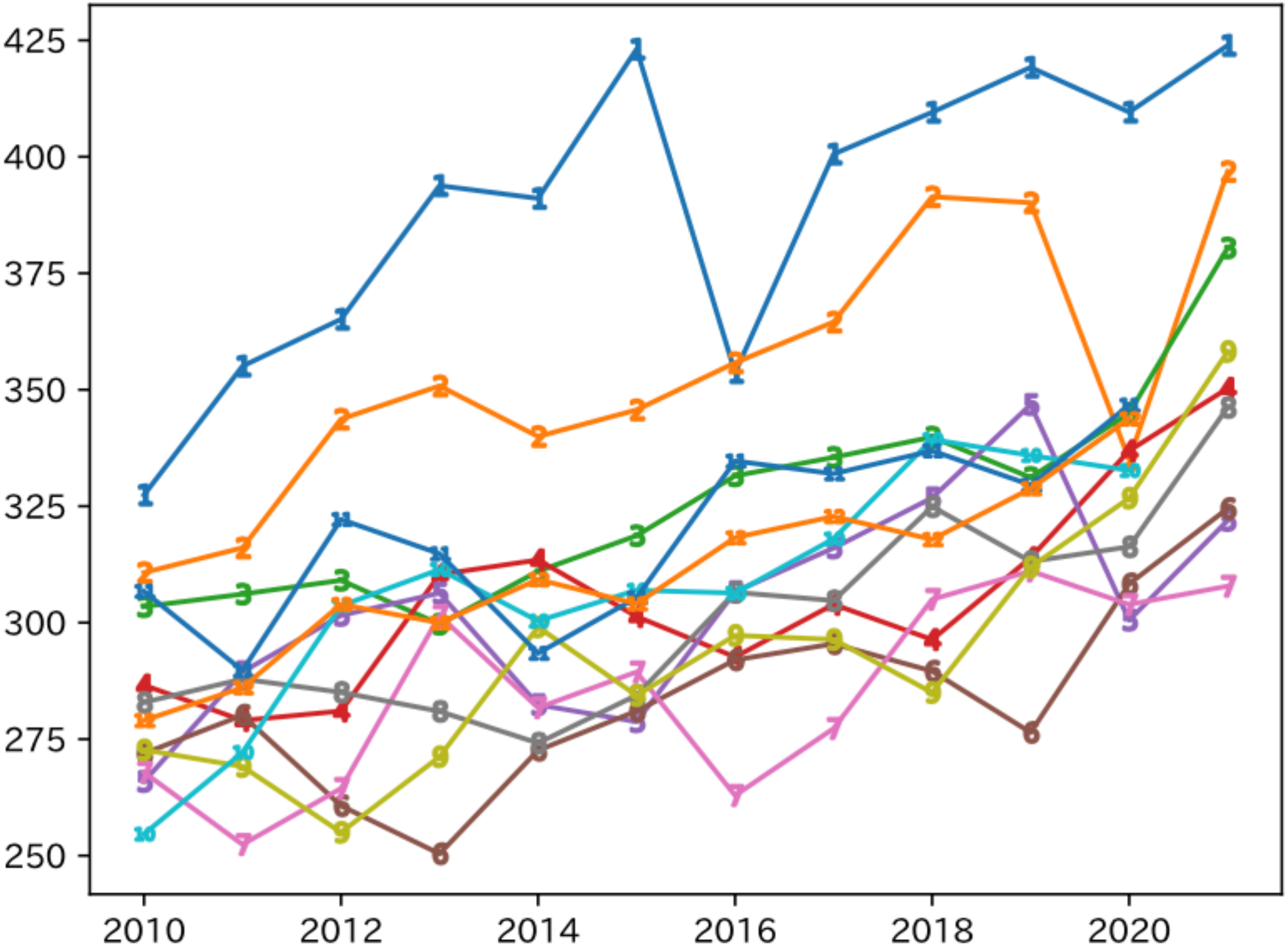
日本の超過死亡

https://oku.edu.mie-u.ac.jp/~okumura/python/japandeaths.html



東京の超過死亡

https://oku.edu.mie-u.ac.jp/~okumura/python/tokyodeaths.html



97.6%？

https://toamit.jp/products/seki_checker/

検査の感度・特異度

感度80%

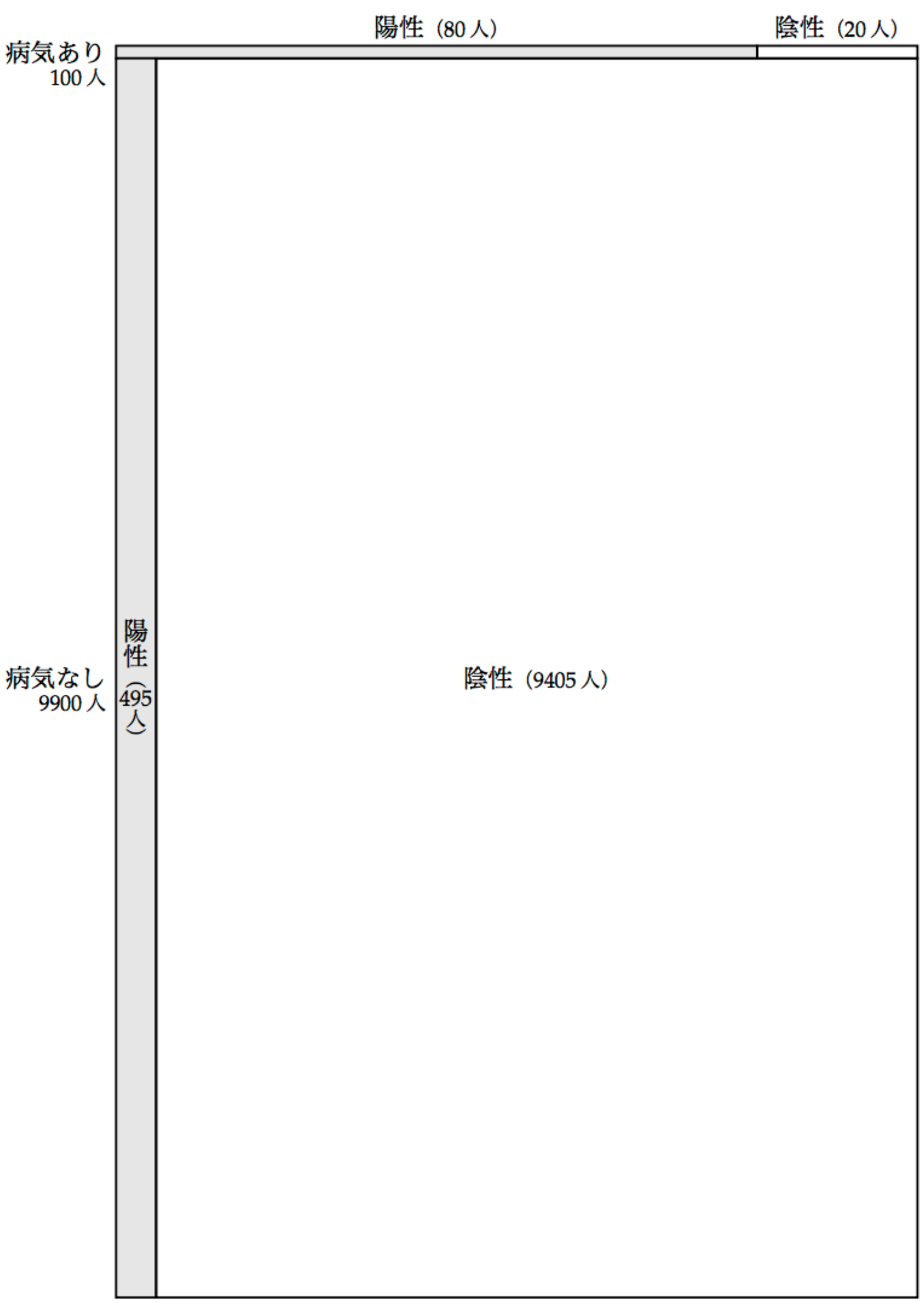
病気あり

陽性	陰性
80％	20％

特異度95%

病気なし

陽性	陰性
5％	95％

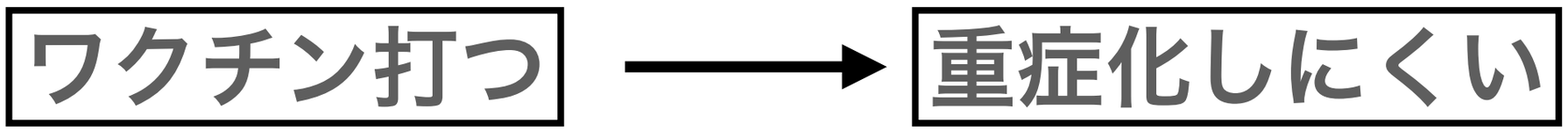

陽性（80 人）
陰性（20 人）
病気あり
100 人
陽性（495 人）
病気なし
9900 人
陰性（9405 人）

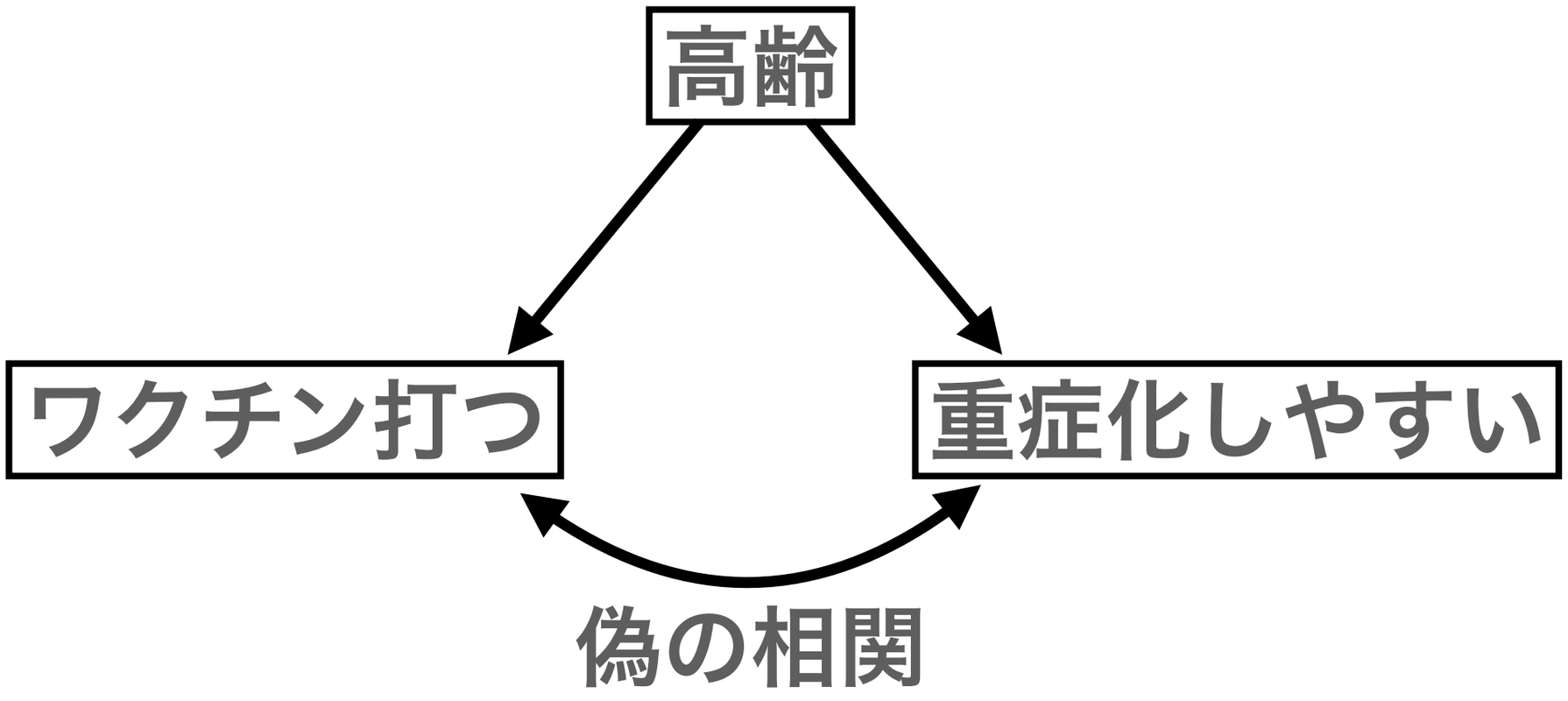
新型コロナワクチンの有効性と Simpsonのパラドックス

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

Age	Population (%)		Severe cases/100k		Severe Case Risk	Efficacy
	% Not Vax	% Fully Vax	Not Vax	Fully Vax	Ratio w/ 30-39 UnVax	vs. severe disease
12-15	62.1%	29.9%	0.30	0.00	1/20x	**100%**
16-19	21.9%	73.5%	1.60	0.00	1/4x	**100%**
20-29	20.5%	76.2%	1.50	0.00	1/4x	**100%**
30-39	16.2%	80.9%	6.20	0.20	1	**96.8%**
40-49	13.2%	84.4%	16.50	1.00	2.7x	**93.9%**
50-59	10.0%	88.0%	40.20	2.90	6.5x	**92.8%**
60-69	8.8%	89.8%	76.60	8.70	12.4x	**88.7%**
70-79	4.2%	94.6%	190.10	19.80	30.7x	**89.6%**
80-89	5.6%	92.6%	252.30	47.90	40.7x	**81.1%**
90+	6.1%	90.5%	510.9	38.60	82.4x	**92.4%**


ワクチン打つ
重症化しにくい


高齢
ワクチン打つ
重症化しやすい
偽の相関