

医科学とビッグデータ **Medical Science and Big Data**

高田則雄/Takata, Norio

慶應義塾大学医学部・先端医科学研究所・脳科学研究部門（リサーチパーク2F-2N3）

Division of Brain Sciences, Institute for Advanced Medical Research,
Keio University School of Medicine

In this class, you would ...

- learn how to utilize Big Data in Medical Science
- get familiar with innovations achieved with Big Data
- be able to describe differences between Big Data- and traditional-approaches in Medical Science

Contents

- **Big Data in General**
 - Major players
 - Definition of Big Data
 - Revolutions enabled with Big Data
- **Big Data in Medical Science**
 - Differences between Big Data- and classical-analysis
 - Ongoing efforts
 - Future directions

Leaders | Regulating the internet giants

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

The data economy demands a new approach to antitrust rules



Major players in Big Data



A survey by IBM around 2017:
2.5 EB (exabyte) data is produced
everyday

$$\begin{aligned}2.5 \text{ EB/day} &= 2.5 \times 10^{18} \text{ byte/day} \\&\sim 1 \times 10^{21} \text{ byte/year}\end{aligned}$$

The data size is approaching the maximum prefix, Y (yotta, **10^{24}**)...

The world generates so much data that new unit measurements were created to keep up

November 19, 2022 · 3:02 PM ET

SI接頭語



接頭語	記号	10^n	十進数表記	漢数字表記	short scale	制定年
クエタ (quetta)	Q	10^{30}	1 000 000 000 000 000 000 000 000 000 000 000	百穰	nonillion	2022年
ロナ (ronna)	R	10^{27}	1 000 000 000 000 000 000 000 000 000 000 000	千杼	octillion	2022年
ヨタ (yotta)	Y	10^{24}	1 000 000 000 000 000 000 000 000 000 000 000	一杼	septillion	1991年
ゼタ (zetta)	Z	10^{21}	1 000 000 000 000 000 000 000 000 000 000 000	十垓	sextrillion	1991年
エクサ (exa)	E	10^{18}	1 000 000 000 000 000 000 000 000 000 000 000	百京	quintillion	1975年
ペタ (peta)	P	10^{15}	1 000 000 000 000 000 000 000 000 000 000 000	千兆	quadrillion	1975年
テラ (tera)	T	10^{12}	1 000 000 000 000 000 000 000 000 000 000 000	一兆	trillion	1960年
ギガ (giga)	G	10^9	1 000 000 000 000 000 000 000 000 000 000 000	十億	billion	1960年
メガ (mega)	M	10^6	1 000 000 000 000 000 000 000 000 000 000 000	百万	million	1960年
キロ (kilo)	k	10^3	1 000 000 000 000 000 000 000 000 000 000 000	千	thousand	1960年
ヘクト (hecto)	h	10^2	100 000 000 000 000 000 000 000 000 000 000	百	hundred	1960年
デカ (deca)	da	10^1	10 000 000 000 000 000 000 000 000 000 000	十	ten	1960年
		10^0	1 000 000 000 000 000 000 000 000 000 000 000	一	one	
デシ (deci)	d	10^{-1}	0.1 000 000 000 000 000 000 000 000 000 000 000	一分	tenth	1960年
センチ (centi)	c	10^{-2}	0.01 000 000 000 000 000 000 000 000 000 000 000	一厘	hundredth	1960年
ミリ (milli)	m	10^{-3}	0.001 000 000 000 000 000 000 000 000 000 000 000	一毛	thousandth	1960年
マイクロ (micro)	μ	10^{-6}	0.000 001 000 000 000 000 000 000 000 000 000 000	一微	millionth	1960年
ナノ (nano)	n	10^{-9}	0.000 000 001 000 000 000 000 000 000 000 000 000	一塵	billionth	1960年
ピコ (pico)	p	10^{-12}	0.000 000 000 001 000 000 000 000 000 000 000 000	一漠	trillionth	1960年
フェムト (femto)	f	10^{-15}	0.000 000 000 000 001 000 000 000 000 000 000 000	一須臾	quadrillionth	1964年
アト (atto)	a	10^{-18}	0.000 000 000 000 000 001 000 000 000 000 000 000	一刹那	quintillionth	1964年
ゼプト (zepto)	z	10^{-21}	0.000 000 000 000 000 000 001 000 000 000 000 000	一清净	sextillionth	1991年
ヨクト (yocto)	y	10^{-24}	0.000 000 000 000 000 000 000 001 000 000 000 000		septillionth	1991年
ロント (ronto)	r	10^{-27}	0.000 000 000 000 000 000 000 000 001 000 000 000		octillionth	2022年
クエクト (quecto)	q	10^{-30}	0.000 000 000 000 000 000 000 000 000 000 000 001 000		nonillionth	2022年

New metric prefixes approved in Nov. 2022

- **Q** (quetta) 10^{30}
- **R** (ronna) 10^{27}

"The change was largely driven by the growing requirements of **data science and digital storage**, which is already using prefixes at the top of the existing range"

[National Physics Lab. in the U.K.](#)

* SI: International System of Units

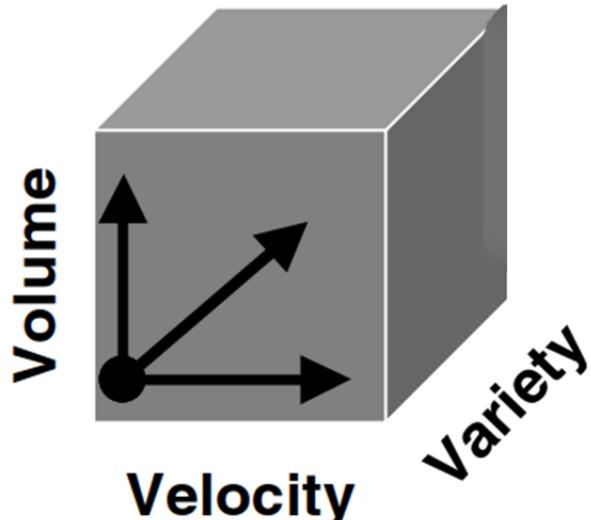
Brown (2019) Measurement

[SI接頭語 - Wikipedia](#)

What's special with “Big Data”?

- Big Data originally referred to just large amount of digital data after the invention of World Wide Web (1989).
- Meaning of “Big Data” has changed.

D. Lancey (2001) is often cited as the first to define “Big Data” with 3 V’s



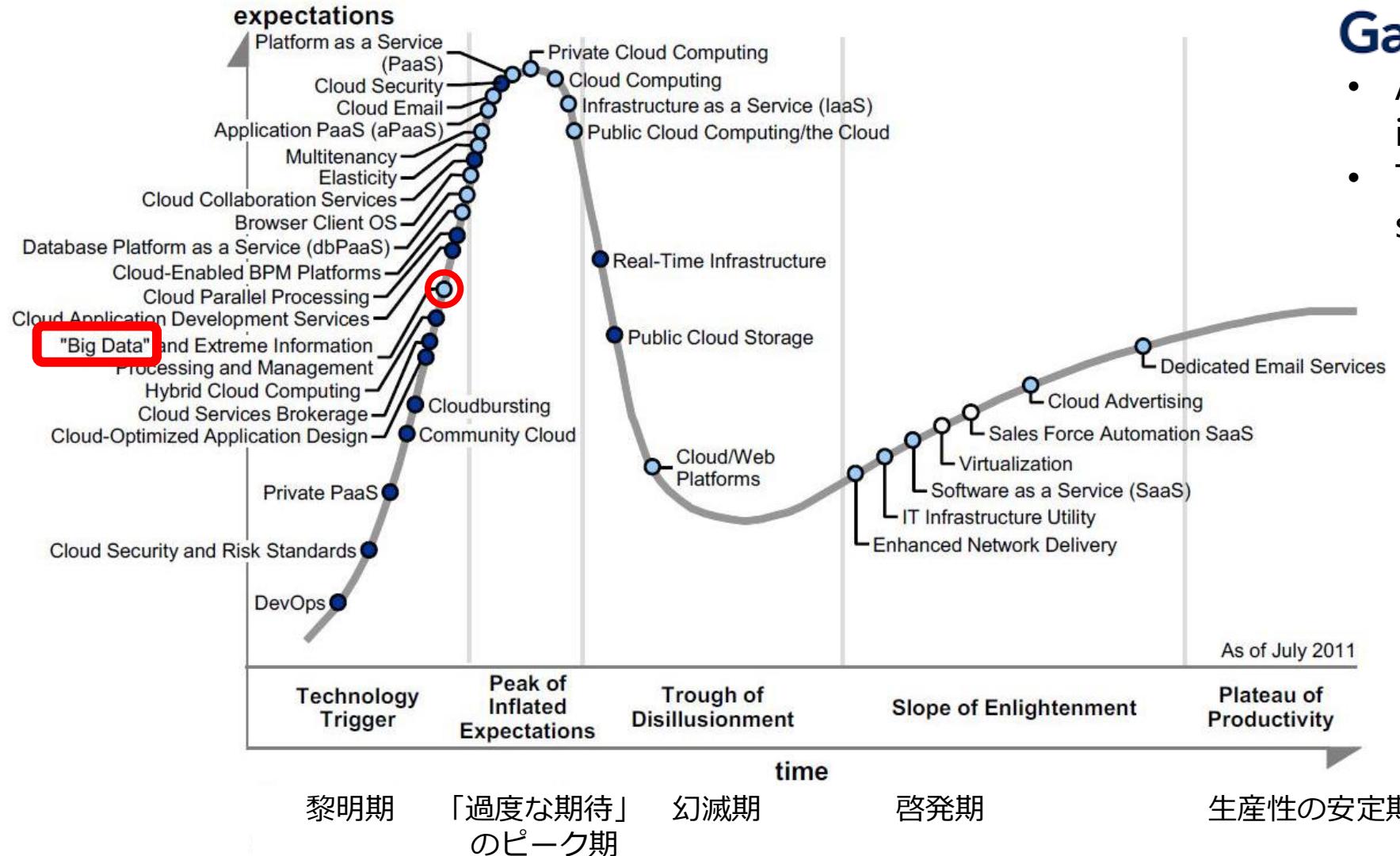
- **Volume**
 - Standards changes over time. Google searches 3.5 G/day in 2012 would satisfy the criteria: amount of data difficult to handle with traditional data management principles.
 - **Velocity**
 - e.g. Continuous data generation by sensors in automated vehicles
 - **Variety**
 - e.g. Structured data: spread sheets, ...
 - Unstructured data: photos, videos, ...
 - Semi-structured data: email, tweets, ...
- (More V's proposed now: veracity (真実性), value, ...)

“traditional data management principles” were at their limit

Actually, the term “Big Data” did not appear in the report.

When did the term “Big Data” first come into use?

Gartner's Hype Cycle Chart put “Big Data” in **2011**

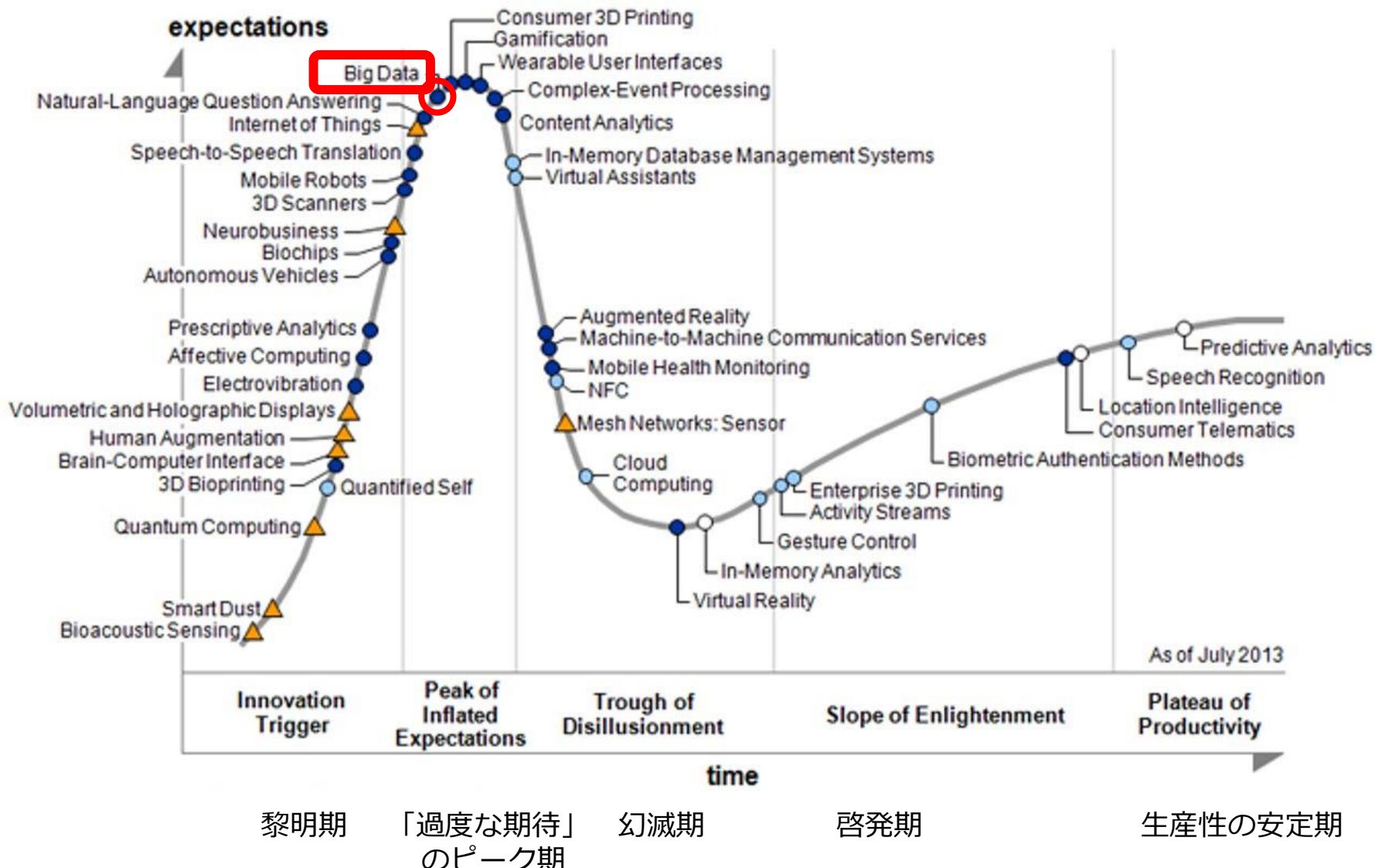


Gartner

- A technological consulting firm in the U.S.
- The chart intends to show the social stages of a new technology.

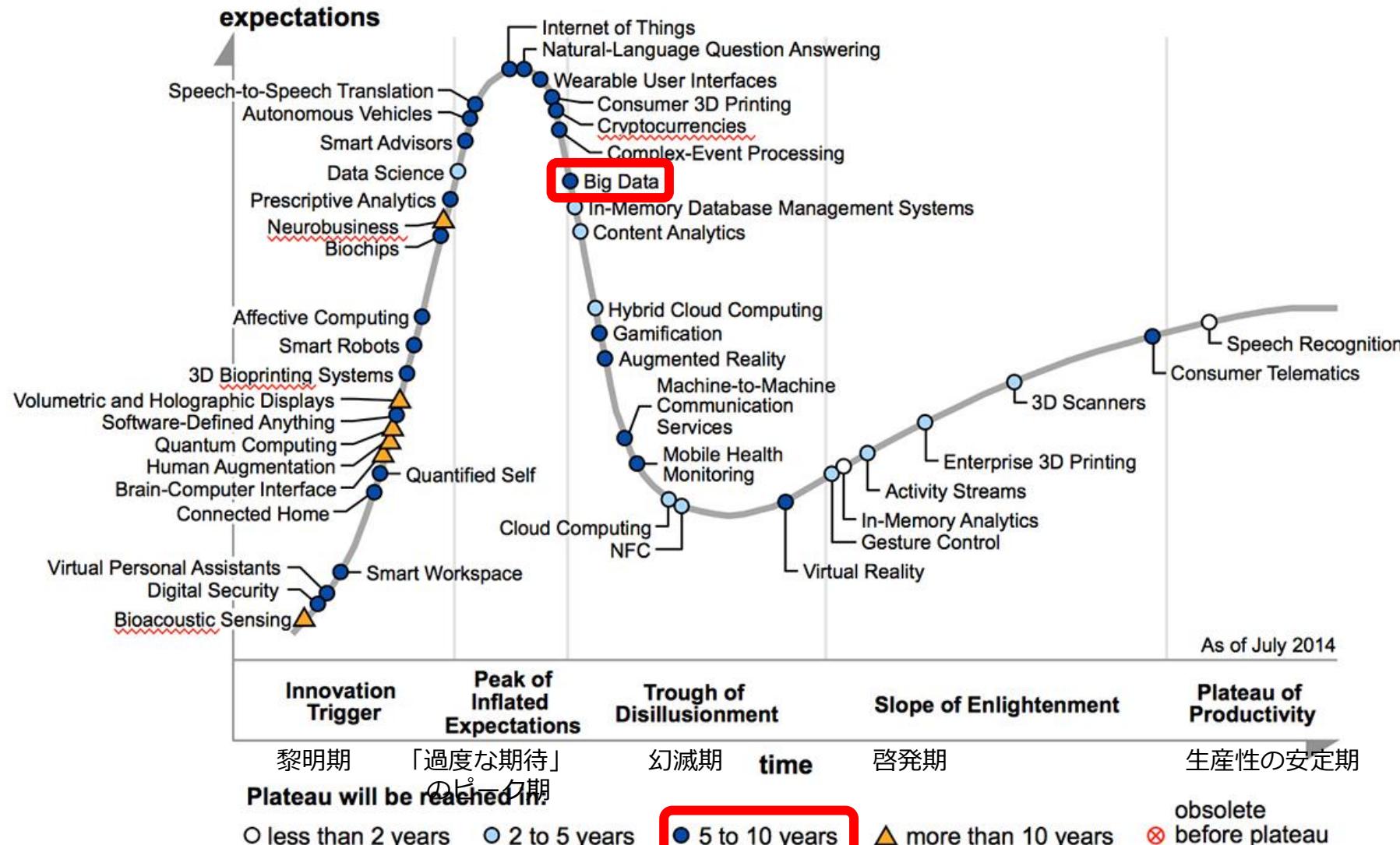
<https://www.gartner.com>

“Big Data” reached its peak in 2013



<https://www.gartner.com>

"Big Data" started to lose its momentum in 2014



7 of 11 TED talks on “Big Data” were presented during 2013-2014

Tim Smith: Big Data



2013/6

There is a mind-boggling amount of data floating around our society. Physi
their ever more massive data for decades- stimulating globalization of the i
Tim Smith plots CERN's involvement with big data from fifty years ago to to

https://www.ted.com/talks/tim_smith_big_data

Tricia Wang: The human insights missing from big data



2016/9

Why do so many companies make bad decisions, even with access to unpr
Netflix to the oracles of ancient Greece, Tricia Wang demystifies big data ai
"thick data" -- precious, unquantifiable insights from actual people -- to ma

https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_

Kenneth Cukier: Big data is better data



2014/6

Self-driving cars were just the start. What's the future of big data-driven te
Cukier looks at what's next for machine learning -- and human knowledge.

https://www.ted.com/talks/kenneth_cukier_big_data_is_better_data

Joel Selanikio: The big-data revolution in health care



2013/2

Erick Brethenoux: The big data behind complex human emotions



2014/10

Intriguing fact: adults who share a recovery room with children recover faster than those who don't. But why? Erick Br
deconstruct complex human emotions by applying big data analytics, so that positive learning like this can be applied
he also sees the potential for abuse, and explains why we need to tread care...

https://www.ted.com/talks/erick_brethenoux_the_big_data_behind_complex_human_emotions

Susan Etlinger: What do we do with all this big data?



2014/9

Does a set of data make you feel more comfortable? More successful? Then your interpretation of it is likely wrong. In
moving talk, Susan Etlinger explains why, as we receive more and more data, we need to deepen our critical thinking :
hard to move beyond counting things to really understanding them.

https://www.ted.com/talks/susan_etlinger_what_do_we_do_with_all_this_big_data

Erin Baumgartner: Big data, small farms and a tale of two tomatoes



2019/1

The path to better food is paved with data, says entrepreneur Erin Baumgartner. Drawing from her experience running
business, she outlines her plan to help create a healthier, zero-waste food system that values the quality and taste of
harvests over factory-farmed produce.

https://www.ted.com/talks/erin_baumgartner_big_data_small_farms_and_a_tale_of_two_tomatoes

Monika Blaumueller: How we could use big data to forecast the next global outbreak



2014/10

Gunjan Bhardwaj: How blockchain and AI can help us decipher medicine's big data



2018/12

When diagnosed with a disease, it's often overwhelming to sort through mountains of medical data to figure out what th
available, pinpoint where they're offered and identify the best experts to help. Complexity specialist Gunjan Bhardwaj re
mining this information may best be done using a system of artificial intelligence...

https://www.ted.com/talks/gunjan_bhardwaj_how_blockchain_and_ai_can_help_us_decipher_medicine_s_big_data

Cathy O'Neil: The era of blind faith in big data must end



2017/4

Algorithms decide who gets a loan, who gets a job interview, who gets insurance and much more -- but they don't autor
things fair. Mathematician and data scientist Cathy O'Neil coined a term for algorithms that are secret, important and han
of math destruction." Learn more about the hidden agendas behind the formulas.

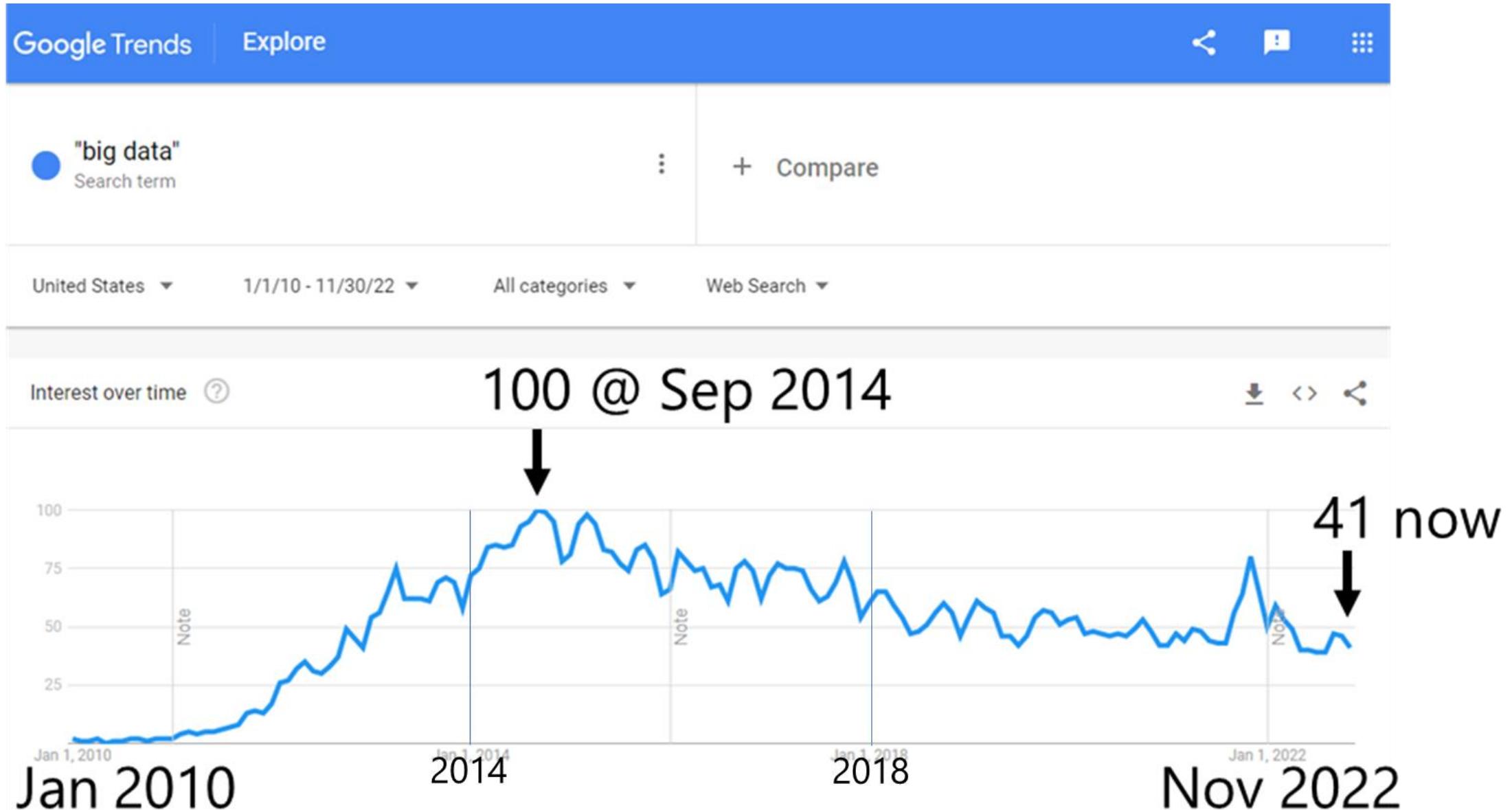
https://www.ted.com/talks/cathy_o_neil_the-era_of_blind_faith_in_big_data_must_end

Ben Wellington: How we found the worst place to park in New York City -- using big data



2014/11

Search volume for “Big Data” peaked in 2014.
Now it’s less than a half



Ministry of Education in Japan Decided a New Curriculum on “AI and Big Data” for Medical School

2022/11/4 読売新聞

医学部の新カリキュラム「AI・ビッグデータ」活用に重点…感染症教育も大幅に充実

文部科学省などは、大学医学部の新たなカリキュラムに人工知能（AI）やビッグデータなど情報科学技術の活用を柱として盛り込む方針を固めた。7日の医療関係者らの会議で決定する。医療現場では、医療機器や画像データの解析にAIが利用されており、最新技術を適切に扱える医療人材の育成が急務となっている。

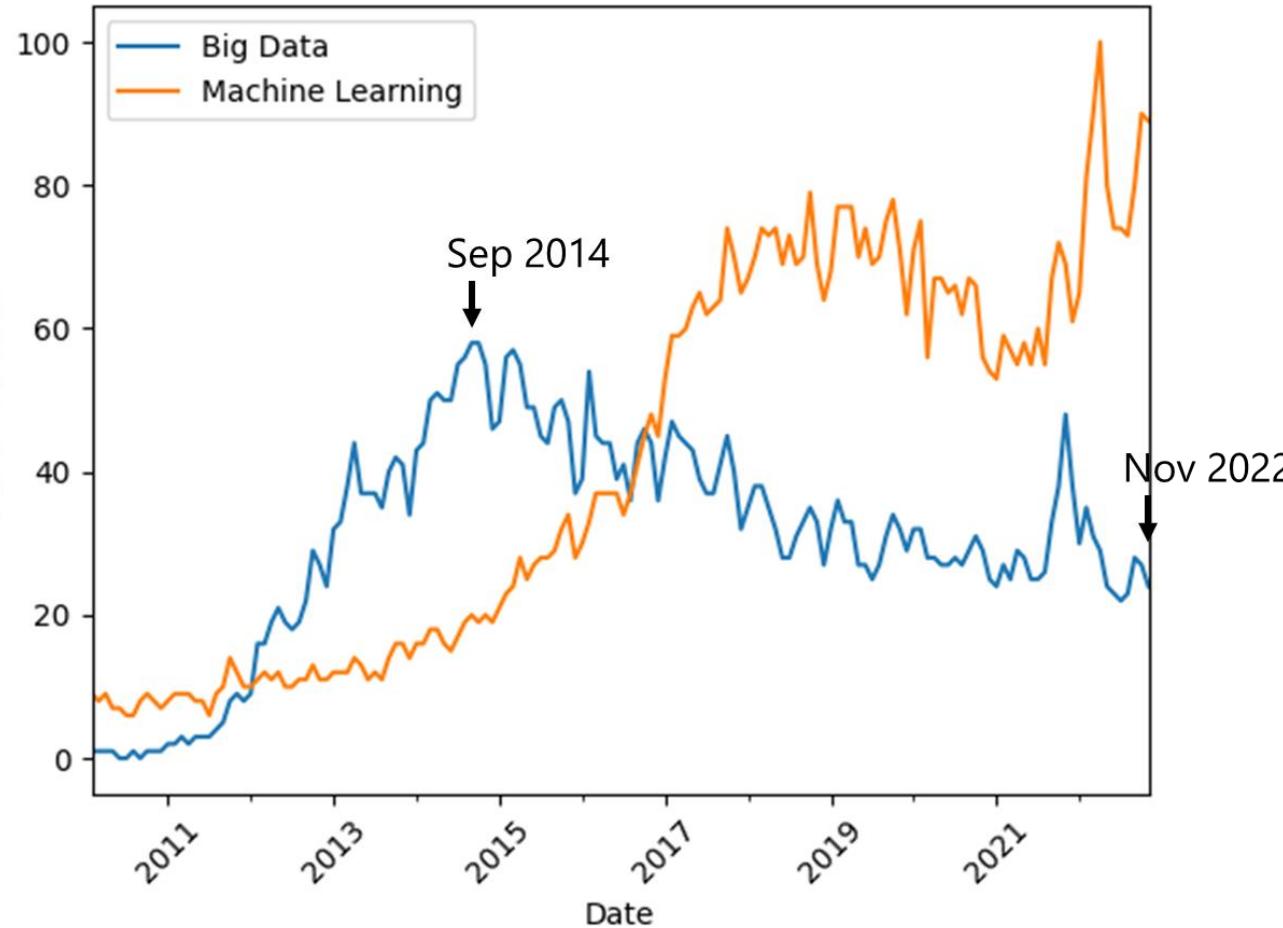
国内82の大学医学部で共通して学ぶカリキュラムは6年に1度改定され、医師に必要な資質や能力、目標などを示す。新しい内容は2024年度以降の入学者が対象。このカリキュラムを踏まえ、各大学は6年間の教育内容を個別に定める。

- **The new curriculum starts from 2024**
- Enable students to use AI, Big Data, and other information technologies to solve medical problems.

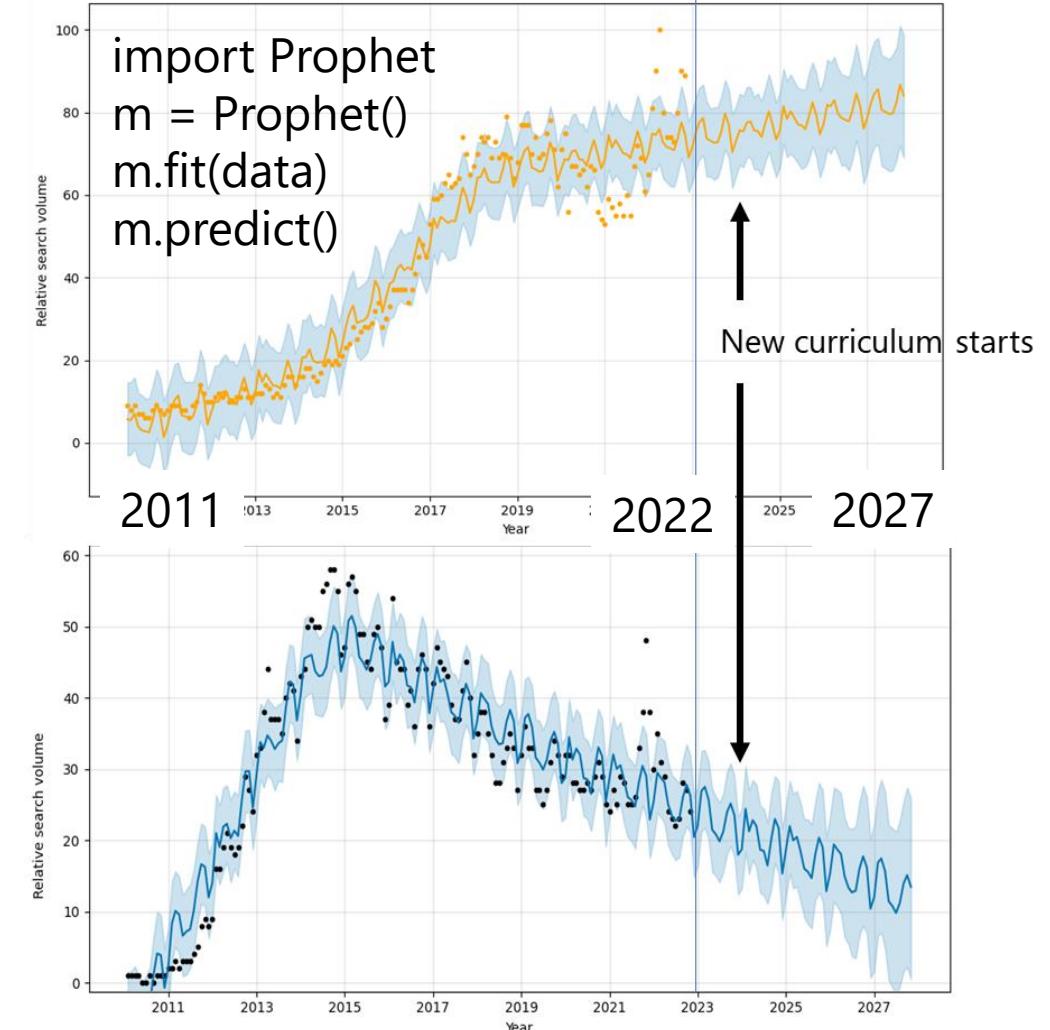
<https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20221104-OYT1T50117/>

Correlation suggests utilization of “**big data**” with “**machine learning**” is important

Google Trends: Big Data vs. Machine Learning



It has become very easy to try prediction.
Just 4-line python codes using **Prophet** by Facebook.



Big Steps achieved with Big Data

1. Text-to-Image
2. Text-to-Speech
3. Text-to-Text (ChatBot)
4. Programming code suggestion and generation
5. Journal article search

1. Text-to-Image generation by Stable Diffusion

arXiv > cs > arXiv:2112.10752

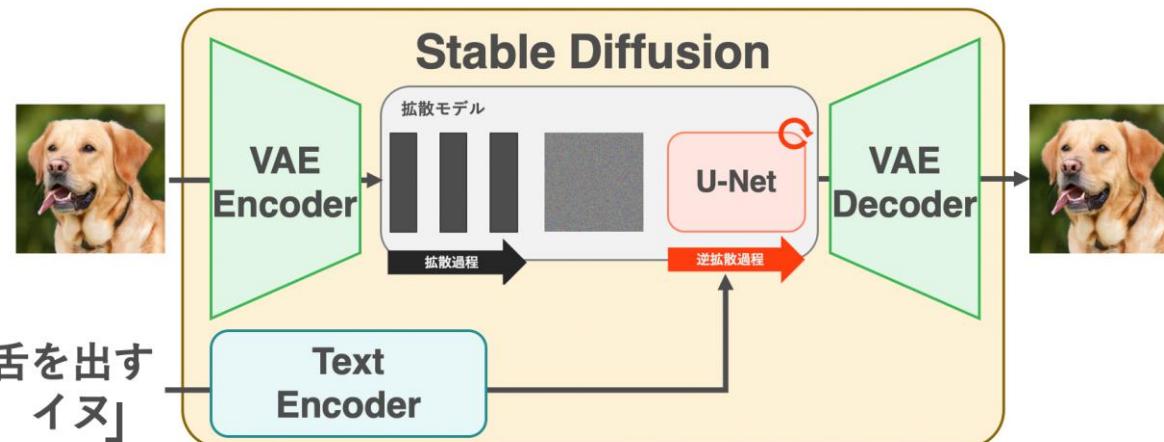
Computer Science > Computer Vision and Pattern Recognition

[Submitted on 20 Dec 2021 (v1), last revised 13 Apr 2022 (this version, v2)]

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer

- A deep learning, text-to-image neural network model by the CompVis Group at LMU Munich.

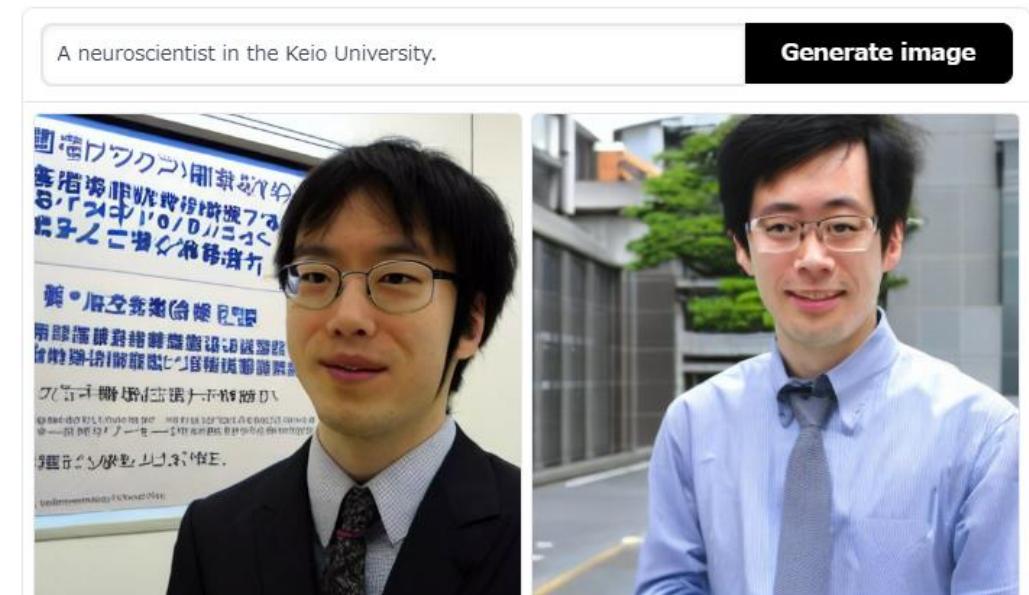


[世界に衝撃を与えた画像生成AI「Stable Diffusion」を徹底解説！ - Qiita](#)

- Free codes & pre-trained weight available at GitHub
- <https://stablediffusionweb.com/>

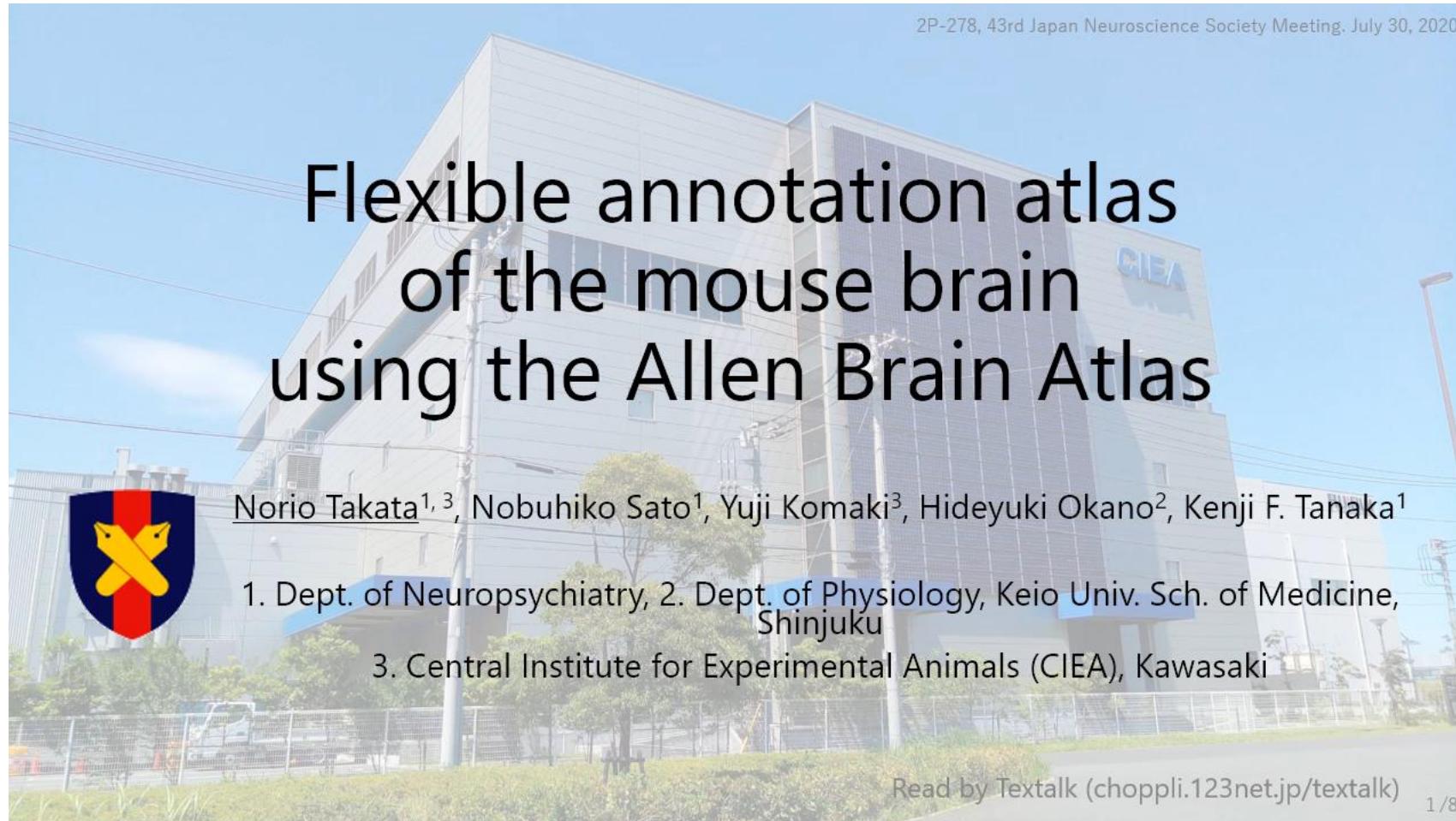
Stable Diffusion Demo

A neuroscientist in the Keio University.



2. Text-to-Speech in 2020 vs. 2022

Poster presentation has been changed to
an online 3-min PowerPoint presentation in Japan Neuroscience Meeting **2020**.



2P-278, 43rd Japan Neuroscience Society Meeting, July 30, 2020

Flexible annotation atlas of the mouse brain using the Allen Brain Atlas

Norio Takata^{1, 3}, Nobuhiko Sato¹, Yuji Komaki³, Hideyuki Okano², Kenji F. Tanaka¹

1. Dept. of Neuropsychiatry, 2. Dept. of Physiology, Keio Univ. Sch. of Medicine,
Shinjuku

3. Central Institute for Experimental Animals (CIEA), Kawasaki

Read by Textalk (choppli.123net.jp/textalk) 1/8

<https://gui.jp.net/textalk/>

<https://murf.ai/>

3. The chat bot, ChatGPT

Dec. 5, 2022. Tweet from a CEO of **OpenAI**, founded in 2015 by Elon Musk and others



Sam Altman ✅

@sama

ChatGPT launched on wednesday. today it crossed 1 million users!

ツイートを翻訳

午後4:35 · 2022年12月5日

<https://chat.openai.com/chat>

- Generative Pretrained Transformer (**GPT**) is a deep-learning based language model for producing human-like text ([proposed by Google in 2017](#))
- ChatGPT uses GPT-3
 - Trained using **45 TB texts on the web**.
 - $1.75 * 10^{11}$ parameters [[10¹⁵ synapses in the human brain](#)]

N

日本語の質問にも答えられますか？



はい、日本語の質問にも答えることができます。私は言語モデルであり、多くの言語に対応しています。私はどのような質問をしても、できる限りお答えするよう心がけます。どうぞ、お気軽に問い合わせください。

N

If you were teaching a class to graduate students in Medical school about Big Data in Medical Sciences, what would you talk about?



If I were teaching a class to graduate students in medical school about big data in medical sciences, I would focus on the following topics:

1. The basics of big data and its potential applications in the medical field, including how large data sets can be used to improve medical research and patient care.
2. The challenges and limitations of working with big data in the medical field, including issues of data quality, privacy, and security.
3. The role of machine learning and AI in analyzing and interpreting large data sets in medical research and patient care, including the potential benefits and drawbacks of using these technologies.
4. The ethical and legal considerations of using big data in medical research and patient care, including issues of consent, confidentiality, and bias.
5. Examples of successful applications of big data in medical research and patient care, including case studies of how big data has been used to improve medical outcomes and advance medical knowledge.

Overall, my aim would be to provide a comprehensive overview of the field of big data in medical sciences, and to help students understand the potential benefits and challenges of using big data in the medical field. I would also encourage students to think critically about the ethical and legal implications of using big data in medical research and patient care, and to consider the potential risks and benefits of these technologies.

4-1 Code suggestion by Copilot

The image shows two parts of the GitHub Copilot interface. The top part is a landing page with a dark background. It features the GitHub logo, a search bar, and navigation links for Pull requests, Issues, Codespaces, Marketplace, and Explore. Below these are tabs for Features, Actions, Packages, Security, Codespaces, Copilot (which is underlined), Code review, Search, Issues, and Discussions. A large white title 'Your AI pair programmer' is centered. Below it is a subtitle: 'GitHub Copilot uses the OpenAI Codex to suggest code and entire functions in real-time, right from your editor.' At the bottom are two buttons: 'Start my free trial >' and 'Explore docs'. The bottom part shows a code editor window with several tabs at the top: 'sentiments.ts', 'write_sql.go', 'parse_expenses.py', and 'addresses.rb'. The 'parse_expenses.py' tab is active. The code editor displays a Python script for parsing expense strings. The script includes imports for `datetime` and `re`, defines a function `parse_expenses` that processes a string of expenses, and initializes an `expenses` list with some sample data.

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12
13    expenses = []
14    for line in expenses_string.splitlines():
15        if line.startswith("#"):
16            continue
17        else:
```

- Copilot (2021-)
 - developed by GitHub and OpenAI
 - based on the Generative Pre-trained Transformer 3 (GPT-3)
 - Trained on **159 GB Python codes** (text files) on GitHub
- I felt like my mind was being read by Copilot during my coding with it.

<https://github.com/features/copilot>

4-2 Competition-level Code Generation by AlphaCode



One of problems solved by AlphaCode

D. Backspace

time limit per test: 2 seconds
memory limit per test: 256 megabytes
input: standard input
output: standard output

You are given two strings s and t , both consisting of lowercase English letters. You are going to type the string s character by character, from the first character to the last one.

When typing a character, instead of pressing the button corresponding to it, you can press the "Backspace" button. It deletes the last character you have typed among those that aren't deleted yet (or does nothing if there are no characters in the current string). For example, if s is "abcbd" and you press Backspace instead of typing the first and the fourth characters, you will get the string "bd" (the first press of Backspace deletes no character, and the second press deletes the character "c"). Another example, if s is "abcaa" and you press Backspace instead of the last two letters, then the resulting text is "a".

Your task is to determine whether you can obtain the string t , if you type the string s and press "Backspace" instead of typing several (maybe zero) characters of s .

Input

The first line contains a single integer q ($1 \leq q \leq 10^5$) — the number of test cases.

The first line of each test case contains the string s ($1 \leq |s| \leq 10^5$). Each character of s is a lowercase English letter.

The second line of each test case contains the string t ($1 \leq |t| \leq 10^5$). Each character of t is a lowercase English letter.

It is guaranteed that the total number of characters in the strings over all test cases does not exceed $2 \cdot 10^5$.

Output

For each test case, print "YES" if you can obtain the string t by typing the string s and replacing some characters with presses of "Backspace" button, or "NO" if you cannot.

You may print each letter in any case (YES, yes, Yes will all be recognized as positive answer, NO, no and nO will all be recognized as negative answer).

Example

input	Copy
4 ababa ba ababa bb aaa aaaa aababa ababa	
output	Copy
YES NO NO YES	

Note

Consider the example test from the statement.

In order to obtain "ba" from "ababa", you may press Backspace instead of typing the first and the fourth characters.

There's no way to obtain "bb" while typing "ababa".

There's no way to obtain "aaaa" while typing "aaa".

In order to obtain "aababa" while typing "aababa", you have to press Backspace instead of typing the first character, then type all the remaining characters.

Given strings S and T,
answer Yes or No
if S can be converted to T
by "backspace operation(s)" on S.
($1 \leq |s| \leq 10^5$, $1 \leq |t| \leq 10^5$)

Examples:

S: abc₁defg

T: afg

Yes

S: elfbediglbdseatiablqgaib

T: bigdata

?

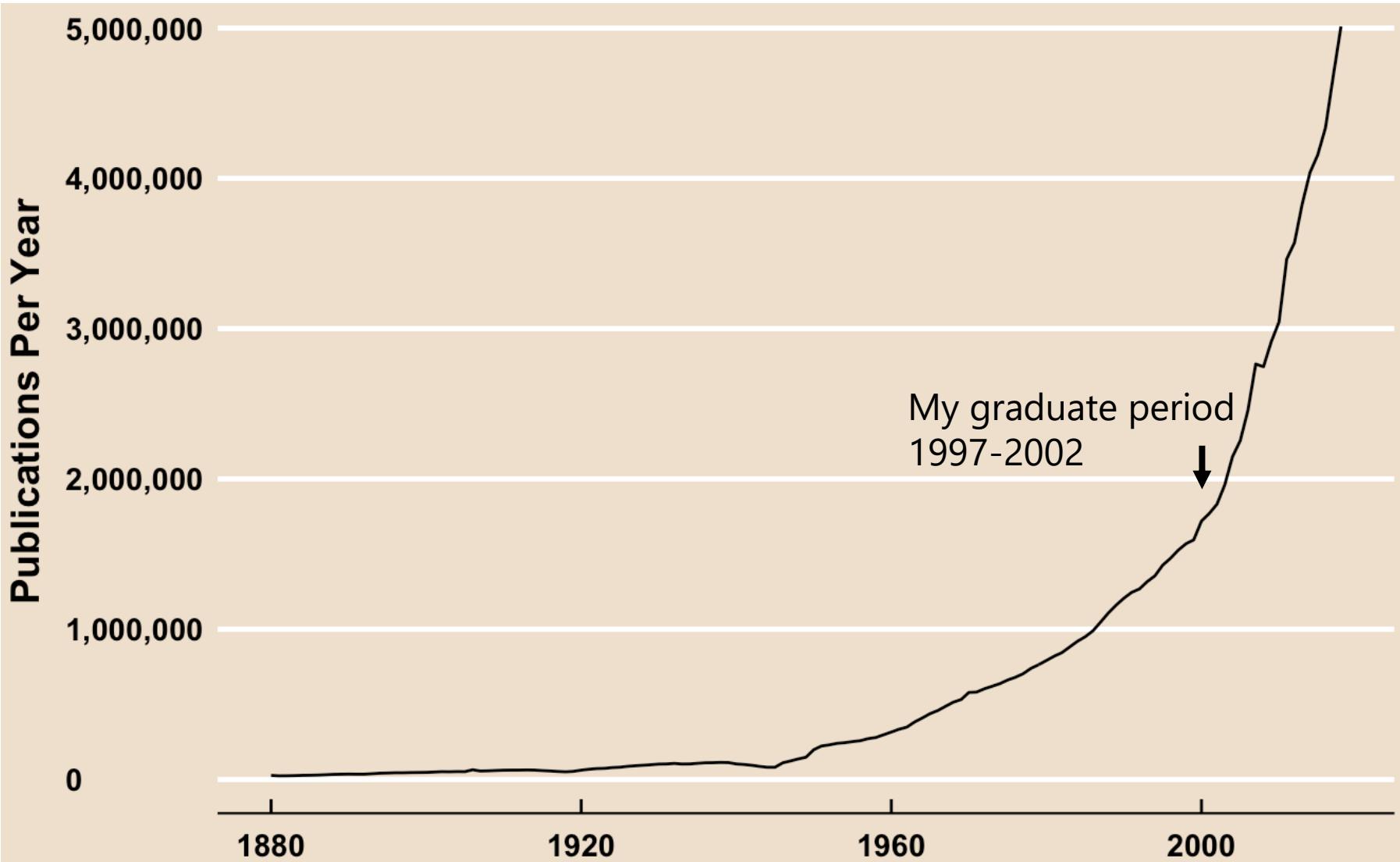
715-GB codes by human on GitHub for pretraining

2.6-GB code by human on **CodeContests** for fine-tuning

4.14×10^{10} parameters

(blog post on Feb. 2, 2022)

5 Scientific Publications Per Year



PubMed (1996-) lists papers based on keywords.

PubMed.gov

big data

Advanced Create alert Create RSS

Save Email Send to

Sorted by: Best match Display options

1 of 3,463

MY NCBI FILTERS: 1 Biological Data

RESULTS BY YEAR: 1939 2023

1 TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. Mol Plant. 2020 Aug 3;13(8):1194-1202. doi: 10.1016/j.molp.2020.06.009. Epub 2020 Jun 23. PMID: 32585190 Free article.

Share

2 Big data analytics in medical engineering and healthcare: methods, advances and challenges. Wang L, Alexander CA. J Med Eng Technol. 2020 Aug;44(6):267-283. doi: 10.1080/03091902.2020.1769758. Epub 2020 Jun 5. PMID: 32498594 Review.

ARTICLE ATTRIBUTE: Associated data

Emergence of AI-assisted paper search

Google Scholar (2004-) ranks papers according to, mainly, citation counts.

Google Scholar search results for "big data". The search bar shows "big data". The results page displays approximately 17,800 items. The first result is a PDF titled "Big data: the management revolution" by McAfee et al., published in 2012. Other results include surveys on big data from Chen et al. (2014) and Gandomi & Haider (2015), and a paper on big data concepts by Chen et al. (2012).

Semantic Scholar (2015-) by Allen Institutes lists papers and provides a summary.

Semantic Scholar search results for "big data". The search bar shows "big data". The results page displays approximately 18,500,000 results. A summary box for "I. Data" shows 2 publications, 2 citations, and materials science. The top result is a summary titled "Big data: The next frontier for innovation, competition, and productivity" by Manjika et al. (2011). Other results include a survey on business intelligence and analytics by Chen et al. (2012) and a survey on data collection for machine learning by Roh et al. (2018).

Elicit (2020-) is a research assistant using a language model like GPT-3. You ask a question; Elicit provides a relevant paper-list and a summary even if they don't match keywords.

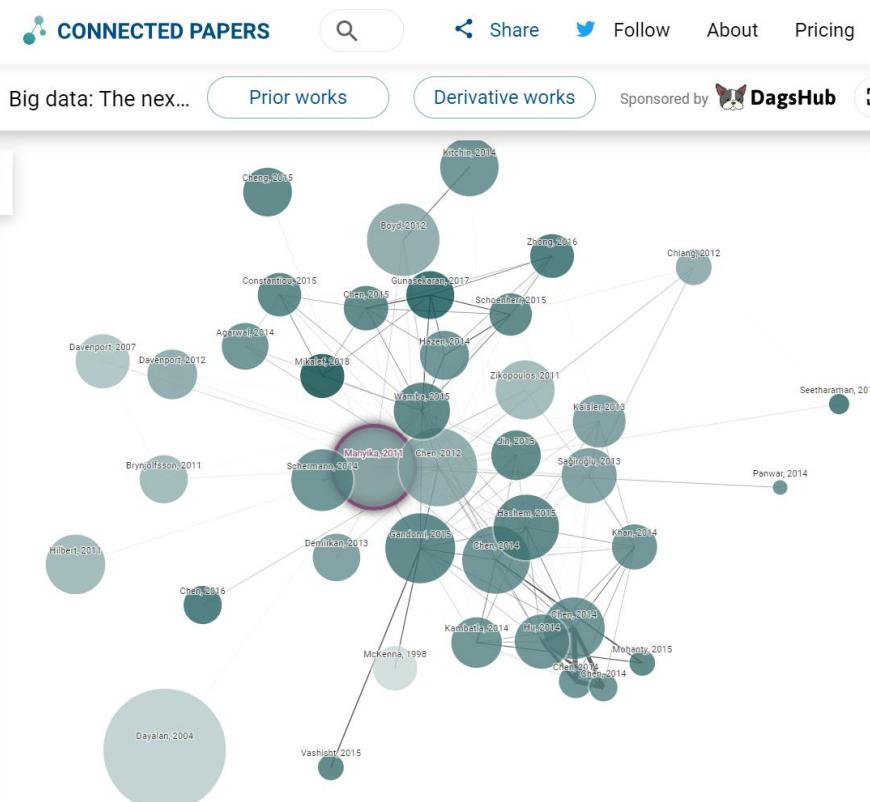
Elicit search results for "What is the next question in big data?". The search bar shows the query. The results page displays a list of top 4 papers. The top result is a summary titled "Big data: The next frontier for innovation, competition, and productivity" by Manjika et al. (2011). Other results include a paper on big data challenges by Cannataci et al. (2020), a paper on big data for statistics by Ieva et al. (2015), and a paper on legal challenges of big data by Cannataci et al. (2016). A sidebar notes that leaders in every sector will have to grapple with the implications of big data.

The increasing spread of automated technologies involves all sectors and different areas of the world with different cultures and legal systems.

Today, an enormous amount of data is being generated in all walks of life by all kinds of devices and systems.

Visualization and suggestions of papers

Connected papers (2020-) is a visual tool for papers according to their similarity (not citation)



Email alerts by
Google Scholar Recommendation
Modified in 2021-

Recommended articles

Google Scholar Alerts <scholaralerts-noreply@google.com>

Neuroscience-Bulletin <neurosci.bull@ion.ac.cn>
To 自分 ▾

This paper might be of interest to you:

[Neurosci Bull](#), 2022 May;38(5):474-488. doi: 10.1007/s12264-021-00782-w. Epub 2021 Oct 26.

[Entrainment of Astrocytic and Neuronal Ca²⁺ Population Dynamics During Information Processing of](#)

Zhu Lin ¹, Feng You ¹, Ting Li ¹, Yijia Feng ¹, Xinyue Zhao ¹, Jingjing Yang ¹, Zhimo Yao ¹, Ying Gao ², Jiang-Fan Chen ³

Abstract: Astrocytes are increasingly recognized to play an active role in learning and memory, but whether neural inputs due to the difficulties in directly monitoring astrocytic Ca²⁺ dynamics in animals performing tasks. Here, using fiber photometry (centered at the turning point of the T-maze) and modified by the reward delivery during the encoding and retrieval phases. Ca²⁺ dynamics with a 3-s phase difference. Furthermore, there was a robust synchronization of astrocytic Ca²⁺ at the population level may contribute to the modulation of information processing.

Keyword: Astrocyte; Ca²⁺ dynamics; Fiber photometry; Hippocampus; Neuron; Working memory.

Cite: Lin Z, You F, Li T, Feng Y, Zhao X, Yang J, Yao Z, Gao Y, Chen JF. Entrainment of Astrocytic and Neuronal Ca²⁺ Population Dynamics. *Neurosci Bull*. 2022 May;38(5):474-488. doi: 10.1007/s12264-021-00782-w. Epub 2021 Oct 26. PMID: 34699030; PMCID: PMC9106780.

Big Data in Medical Science

Characteristics of Medical Big Data

- **Comprehensive** molecular information

- Genome/Omics-based Medicine

- Volume

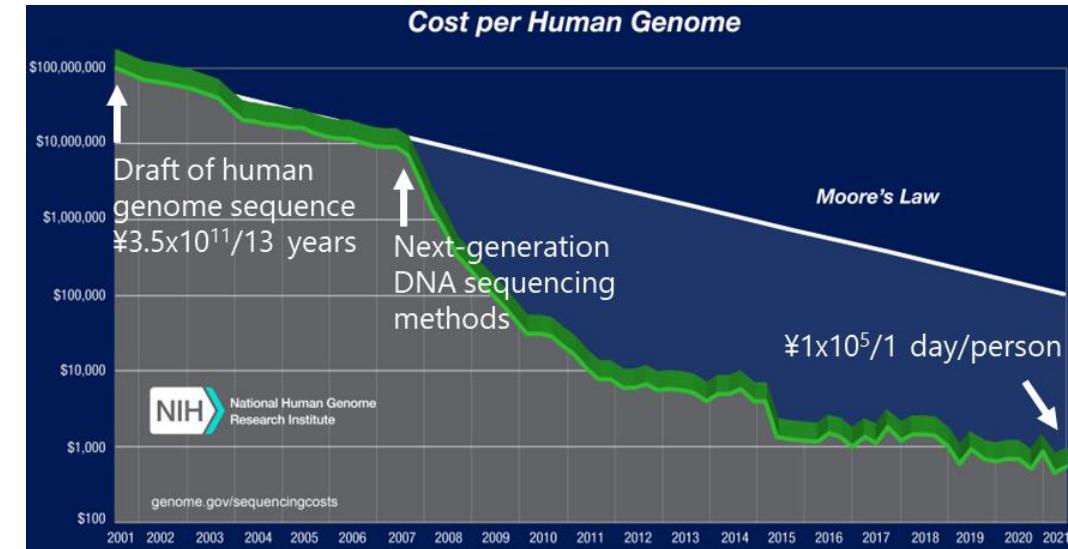
- DNA sequencing data by Sequence Read Archive (SRA)
> 100 Terabases in 2011.

- Velocity

- Data size of European Nucleotide Archive doubles/year

- Variety

- Omics data can be linked to various data of patients.



- **Unobstructed** monitoring

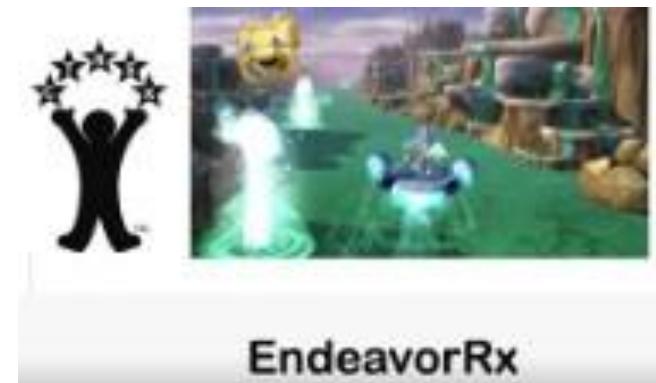
- Mobile health (mHealth) data

- Continuous time-series data (> days~years) obtained with mobile phones and monitoring devices of patients.

- Digital therapeutics (DTx):

- EndeavorRx, the first FDA approved video game prescription for treating children with ADHD.

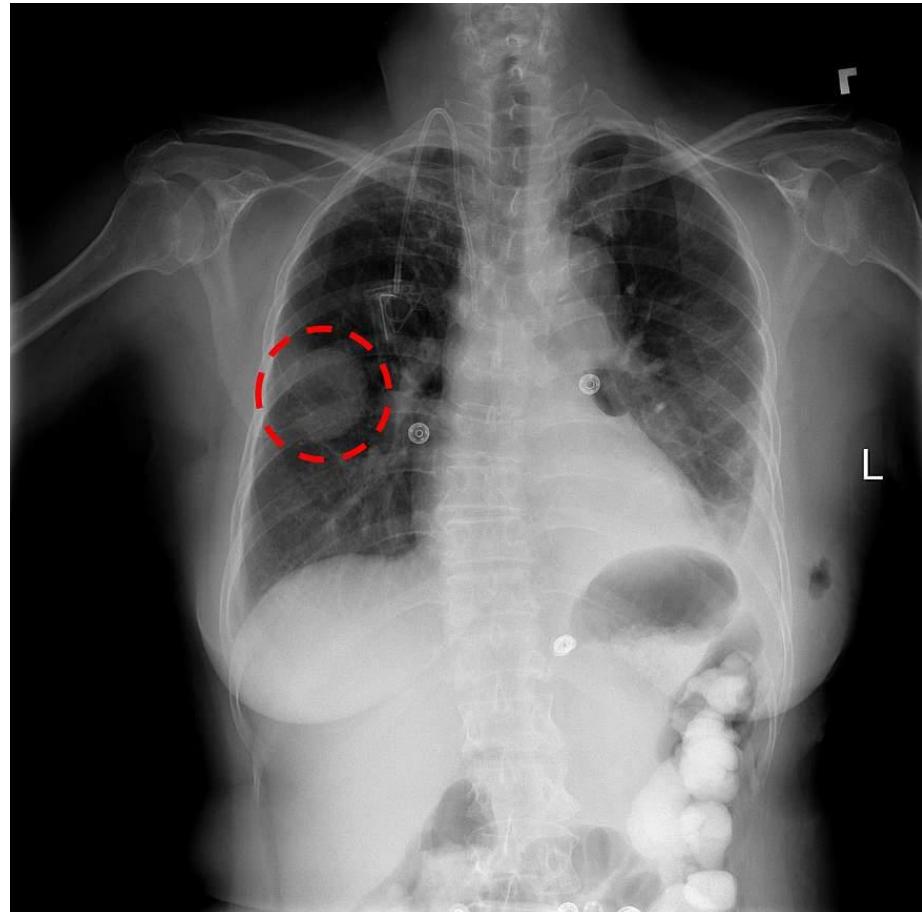
- Japanese version by Shionogi is in phase 3 trial (August 2022).



Comparison of Big Data- and classical-analysis

	Medical Big Data analysis	Classical statistical analysis
Application	Hypothesis-generating	Hypothesis-testing
Domain knowledge	Important for interpretation of the results	Important both for formulation of a hypothesis and interpretation of the results
Sources of data	Any kind of sources; frequently multiple sources	Carefully specified collection of data; usually single source
Coverage of data	Substantial fraction of entire population	Small data samples from a specific population with assumptions of their distribution
Nature and structure of data	Unstructured and structured; Wide format data 	Mainly structured; Long format data
Data quality	Rarely clean as in real world data	Quality controlled as in random controlled trial (RCT)
Assumption of the model	Frequently absent	Based on various underlying probability distribution function
Analytic tools	Frequently automated with data mining algorithm including machine learning	Manually by expert with classical statistics
Main outputs of analysis	Prediction, models, patterns identified	Statistical score contrasted against random chance
Objectives	Personalized medicine: Proactive (予知的) and life-course-oriented	Population medicine: Reactive (対応的) and occasional (機会主義的)

Medical Big Data repositories are expanding



- **Publicly available Chest x-ray datasets** by National Institutes of Health (NIH)
 - 2017 Sept.
 - 46 GB
 - 100k images (1024 x 1024 PNG) from > 30k patients
- [Wang et al. \(2017\) IEEE Conf. on CVPR](#)
- <https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community>

Google BigQuery: 46 Medical Datasets



American Community Survey
(ACS)

United States Census Bureau

Detailed US demographic data at various geographic resolutions

broadstreet

Area Deprivation Index (ADI)

BroadStreet

ADI: An index of socioeconomic status for communities



Births Data Summary

Centers for Disease Control

Nativity Data from CDC Births



Broad References

Broad Institute of MIT and Harvard

Human genomics reference files used for sequencing analytics



COVID-19 Public Forecasts
BigQuery Public Datasets Program
COVID-19 forecasts for the next 28 days in the US and Japan

Google Health

COVID-19 Search Trends

symptoms dataset

BigQuery Public Datasets Program

Trends in symptom search volumes due to COVID-19

Google Health

COVID-19 Vaccination Access

BigQuery Public Datasets Program

Metrics quantifying travel times to COVID-19 vaccination sites.

Google Health

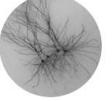
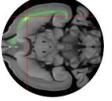
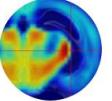
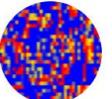
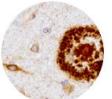
COVID-19 Vaccination Search Insights

BigQuery Public Datasets Program

Trends in Google searches for COVID-19 vaccination information

Allen Institute for Brain Science/Cell Science/Immunology

Allen Brain Atlases and Data

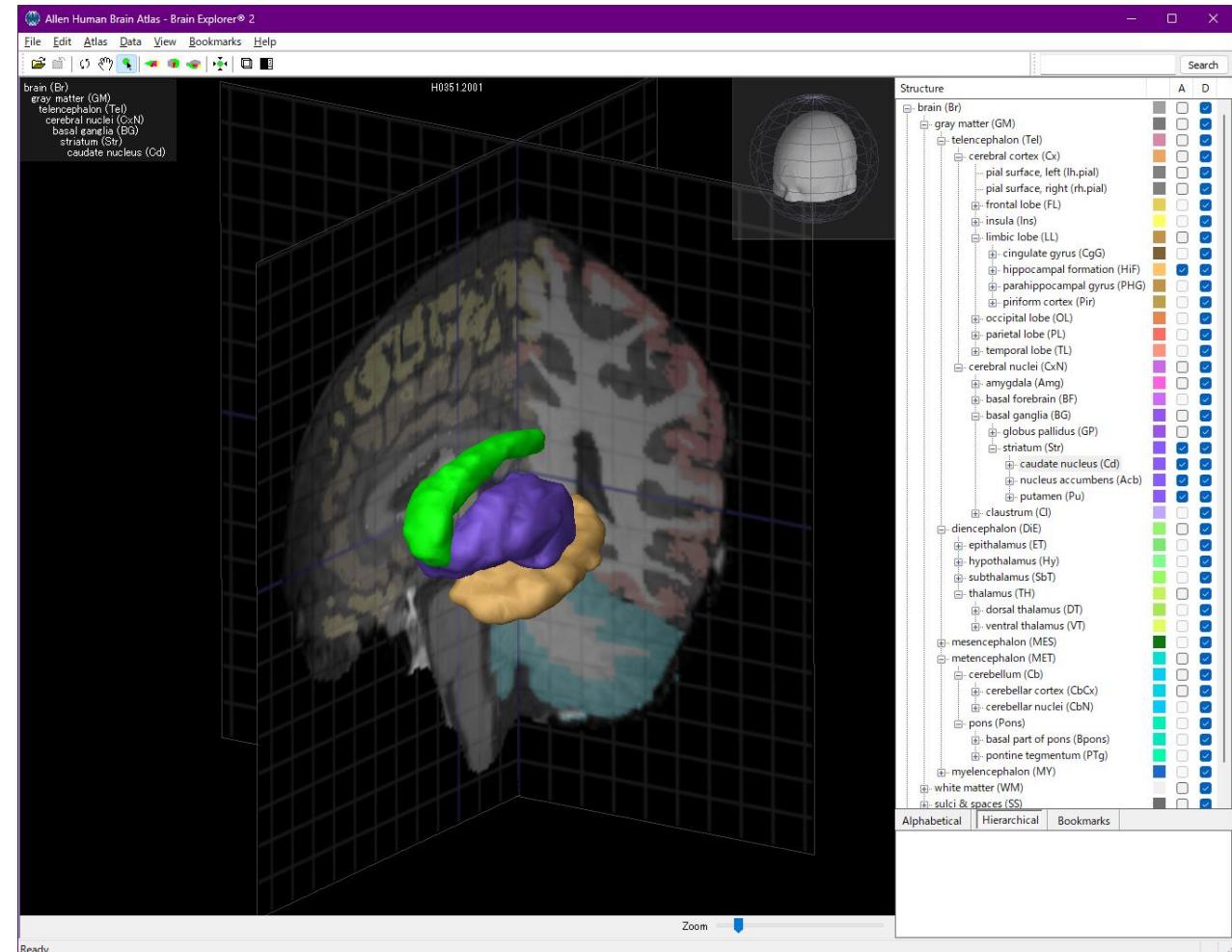
 CELL TYPES DATABASE A database of biological features derived from single cells, from both human and mouse. View Data →	 BRAIN OBSERVATORY A new approach to open data, featuring a survey of <i>in vivo</i> recordings from the mouse visual cortex. View Data →	 MOUSE BRAIN CONNECTIVITY ATLAS A brain-wide map of neural projections, including cell class-specific data. View Atlas →	 REFERENCE ATLASES High resolution anatomical reference atlases and histology for mouse and human. View Atlases →
 MOUSE BRAIN ATLAS A unique multimodal atlas of the adult mouse brain, featuring anatomic and genomic data. View Atlas →	 DEVELOPING MOUSE BRAIN ATLAS A detailed atlas of gene expression across 7 stages of development. View Atlas →	 MOUSE SPINAL CORD ATLAS A detailed atlas of gene expression across the adult and juvenile mouse spinal cord. View Atlas →	 ADULT AND DEVELOPING NHP ATLAS The NIH Blueprint Non-Human Primate Atlas characterizes the developing rhesus macaque brain. View Atlas →
 HUMAN BRAIN ATLAS A unique multimodal atlas of the adult human brain, featuring anatomic and genomic data. View Atlas →	 DEVELOPING HUMAN BRAIN The BrainSpan project is a detailed atlas of gene expression across human development. View Data →	 AGING, DEMENTIA AND TBI A dataset for exploring the neuropathology and genomic features of disease and aging. View Data →	 IVY GLIOBLASTOMA ATLAS PROJECT IvyGAP is a dataset for exploring the anatomic and genomic basis of glioblastoma. View Atlas →



SEATTLE ALZHEIMER'S DISEASE BRAIN CELL ATLAS

A multifaceted resource for exploring the cellular and molecular basis of Alzheimer's Disease.

[View Data →](#)



[Brain Explorer :: Allen Brain Atlas: Mouse Brain \(brain-map.org\)](#)

[Brain Map - brain-map.org](#)

Flexible atlas of mouse brain utilizing Allen Database

ntakata / flexible-annotation-atlas Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master · 1 branch · 0 tags Go to file Add file · < Code

ntakata Update README 511bad · on Mar 18, 2021 · 17 commits

FAAs Modified README.md for FAAoocd 2 years ago

Supplementary_Tables added supplementary tables and yaml for Mac 2 years ago

data initial commit 3 years ago

fiber_data/From_AI_To_CP initial commit 3 years ago

gene_data/74881161 initial commit 3 years ago

notebooks initial commit 3 years ago

yaml added supplementary tables and yaml for Mac 2 years ago

.gitignore added FAAoocd 2 years ago

Divide_nodes.ipynb modified separator 3 years ago

Divide_nodes.py added scripts 2 years ago

LICENSE.md initial commit 3 years ago

Prepare_AObaseAVbase.ipynb fix for nbparameterise timeout 3 years ago

Prepare_AObaseAVbase.py added scripts 2 years ago

README.md Update README 2 years ago

Flexible annotation atlas (FAA) of the mouse brain

This repository contains Python codes for our papers ([Sci Rep](#) and [bioRxiv](#)) "Flexible annotation atlas of the mouse brain: combining and dividing brain structures of the Allen Brain Atlas while maintaining anatomical hierarchy".

Several FAAs are available in [an FAAs folder](#) including FAAoocd that was used in [NeuroImage](#) for cortico-striatal-thalamo (CST) circuitry analysis.

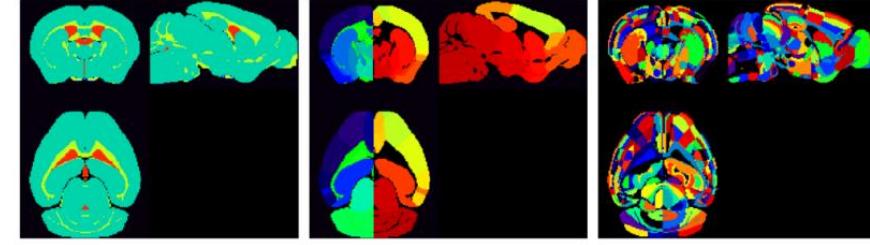
Examples of FAA

FAA consists of a JSON-formatted text file (anatomical ontology, AO) and a three-dimensional volume file of the mouse brain (annotation volume, AV). Examples of FAA are shown below: FAAsegment (total node count: 4), FAAcoarse (101), and FAAdetailed (1381). In FAAdetailed, nodes for the hippocampus and striatum were divided to two (dorsal and ventral regions) based on gene expression and fiber projection, respectively.

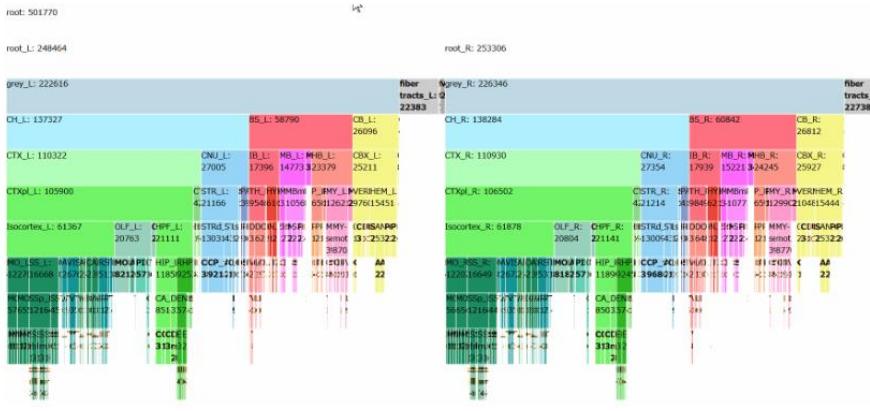
README.md

CH_R: 138284 stands for a brain structure "Cerebrum" in the right hemisphere "R" with a volume size of 138,284 nl. In an FAAs folder, some FAAs and an [HTML](#) file for zoomable visualization of anatomical hierarchy are available.

FAAsegment **FAAcoarse** **FAAdetailed**



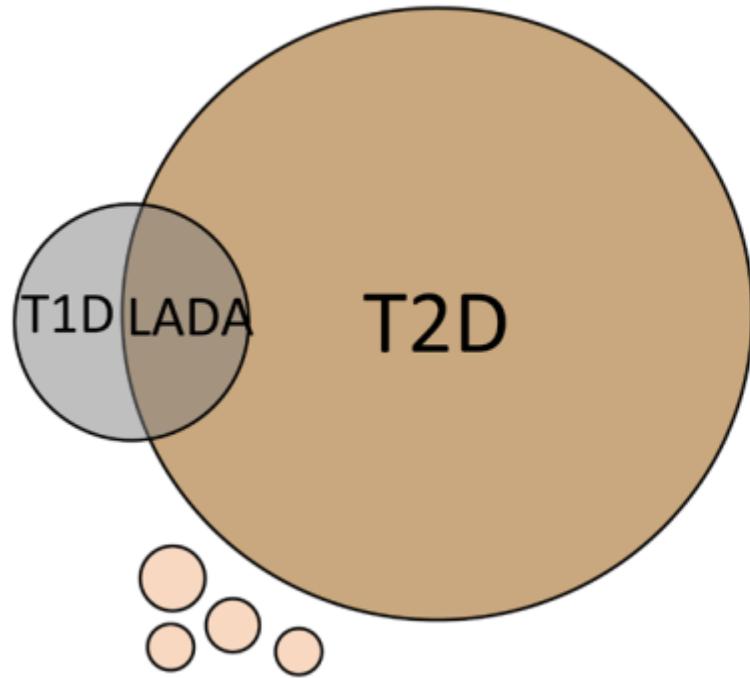
Anatomical hierarchy of FAAdetailed



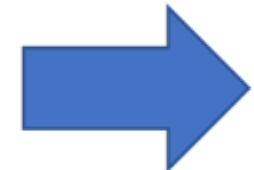
[ntakata/flexible-annotation-atlas: a pipeline for constructing a mouse brain atlas](#)

Achievements with Medical Big Data

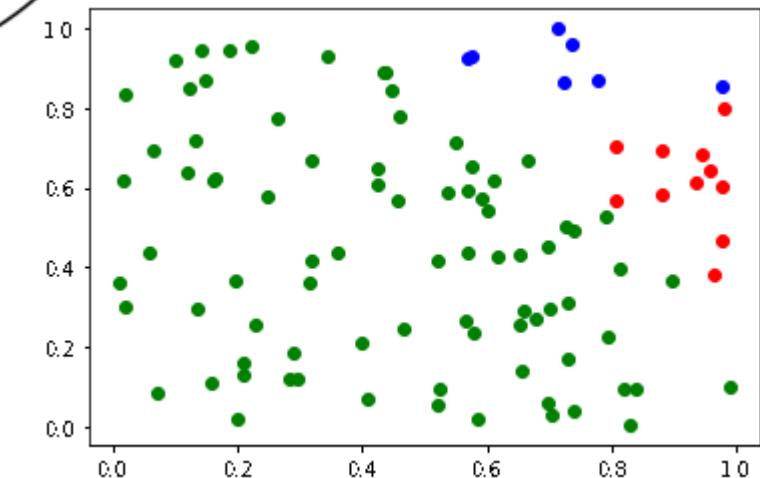
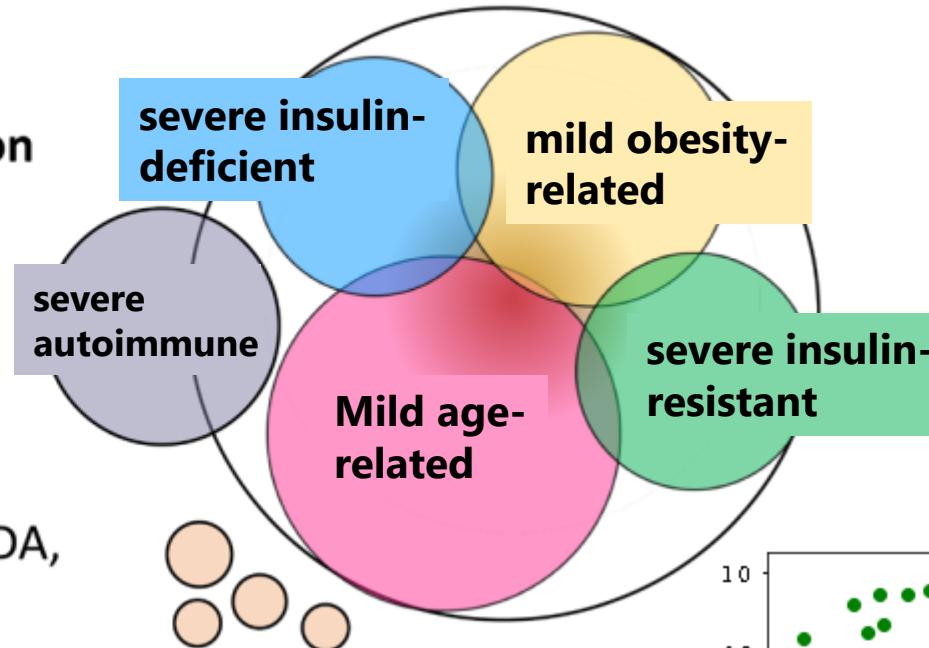
Discovery of novel subgroups of diabetes by a data-driven cluster analysis



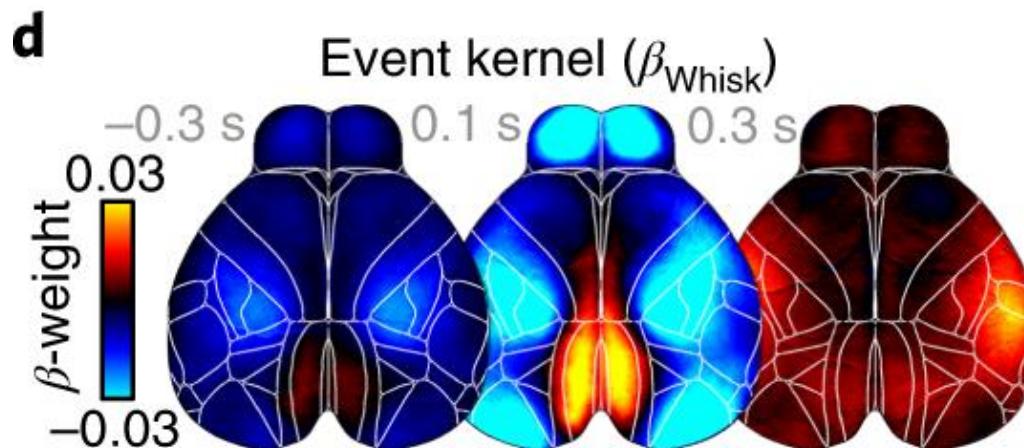
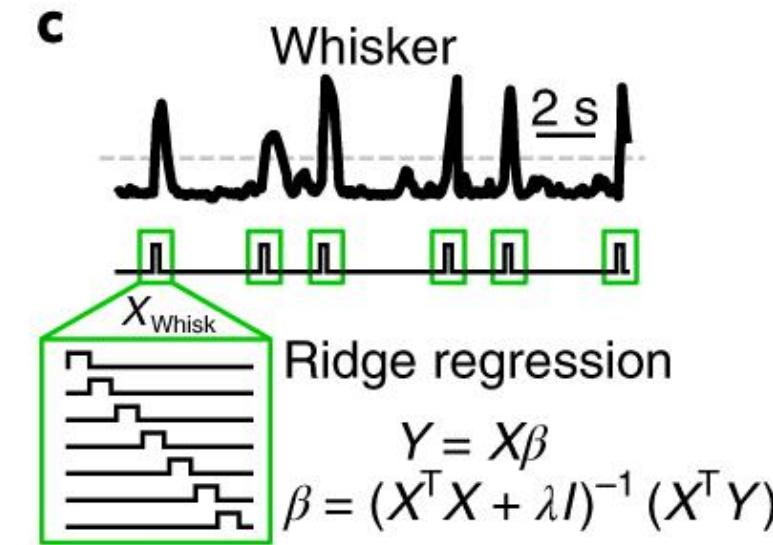
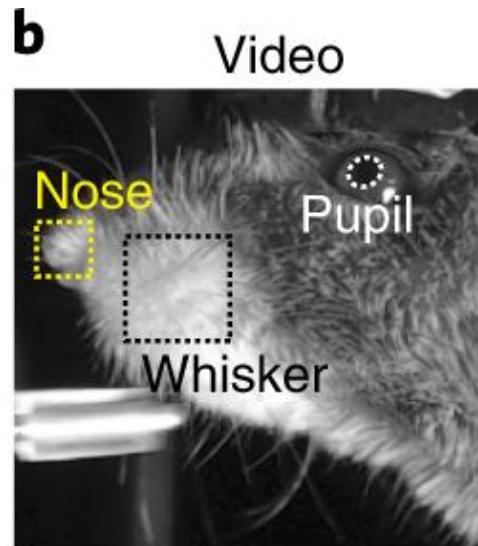
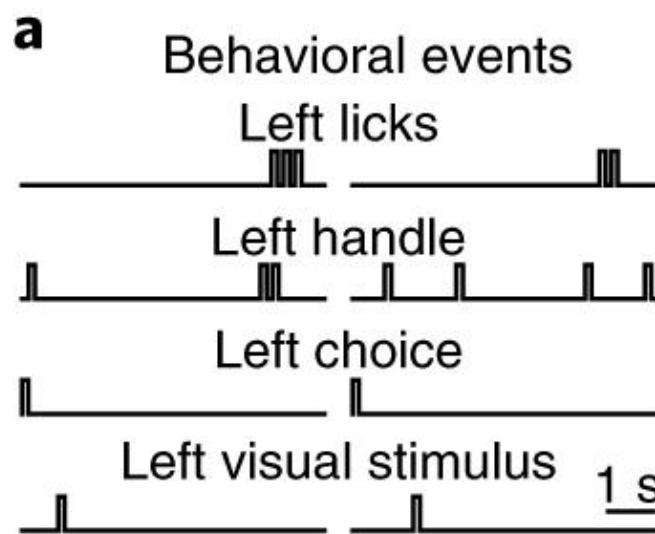
Clustering/
classification
(K-means)



Age, BMI,
HbA1c, GADA,
HOMA2-B,
HOMA2-IR

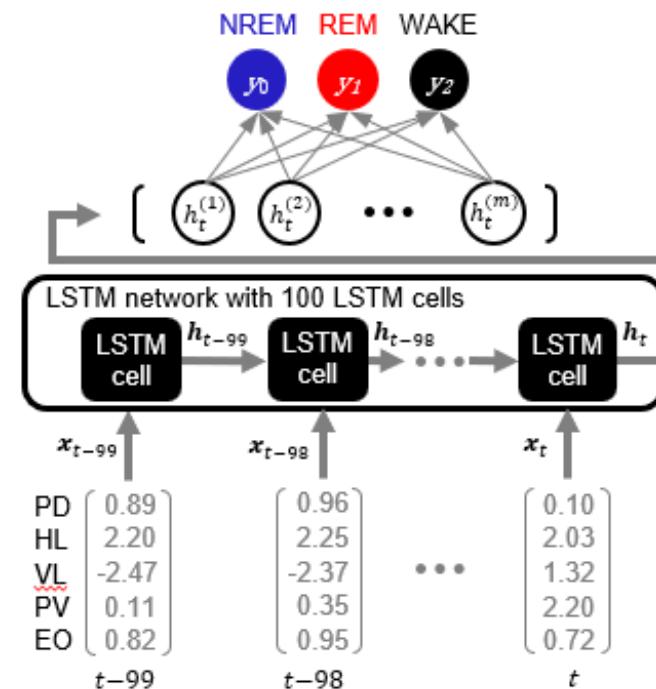
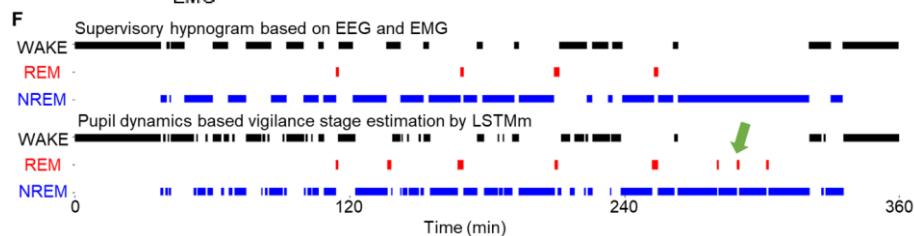
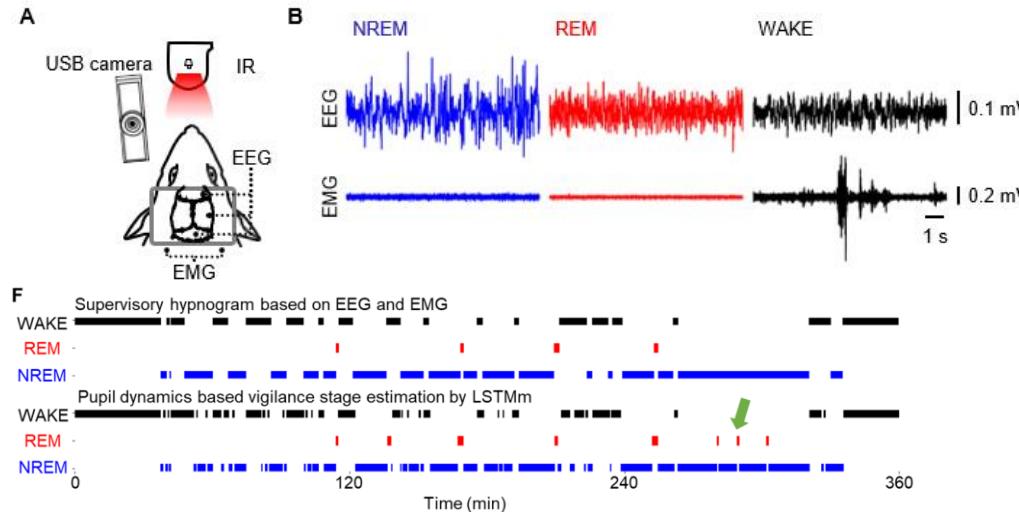


Realtime movies of mouse face and body revealed that most brain activity during a task reflected irrelevant movements



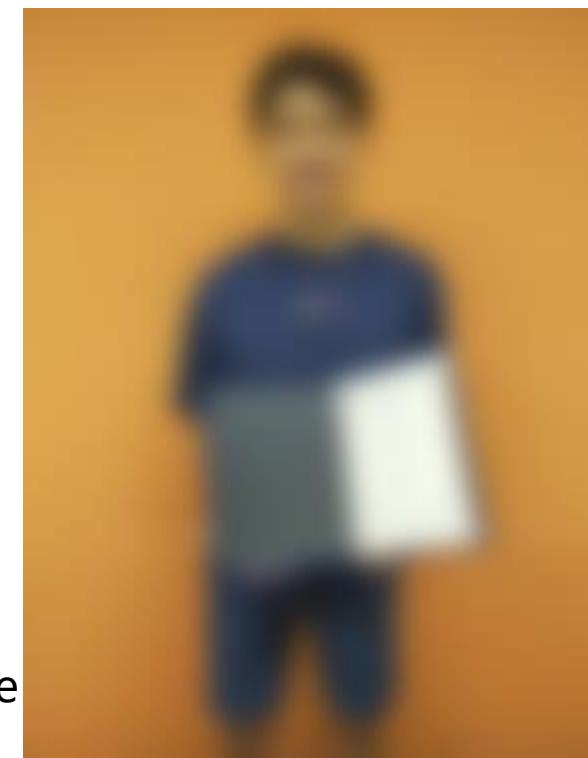
Musall et al (2019) Nature Neurosci "Single-trial neural dynamics are dominated by richly varied movements"

Sleep stage estimation based on pupil dynamics using a recurrent neural network, LSTM



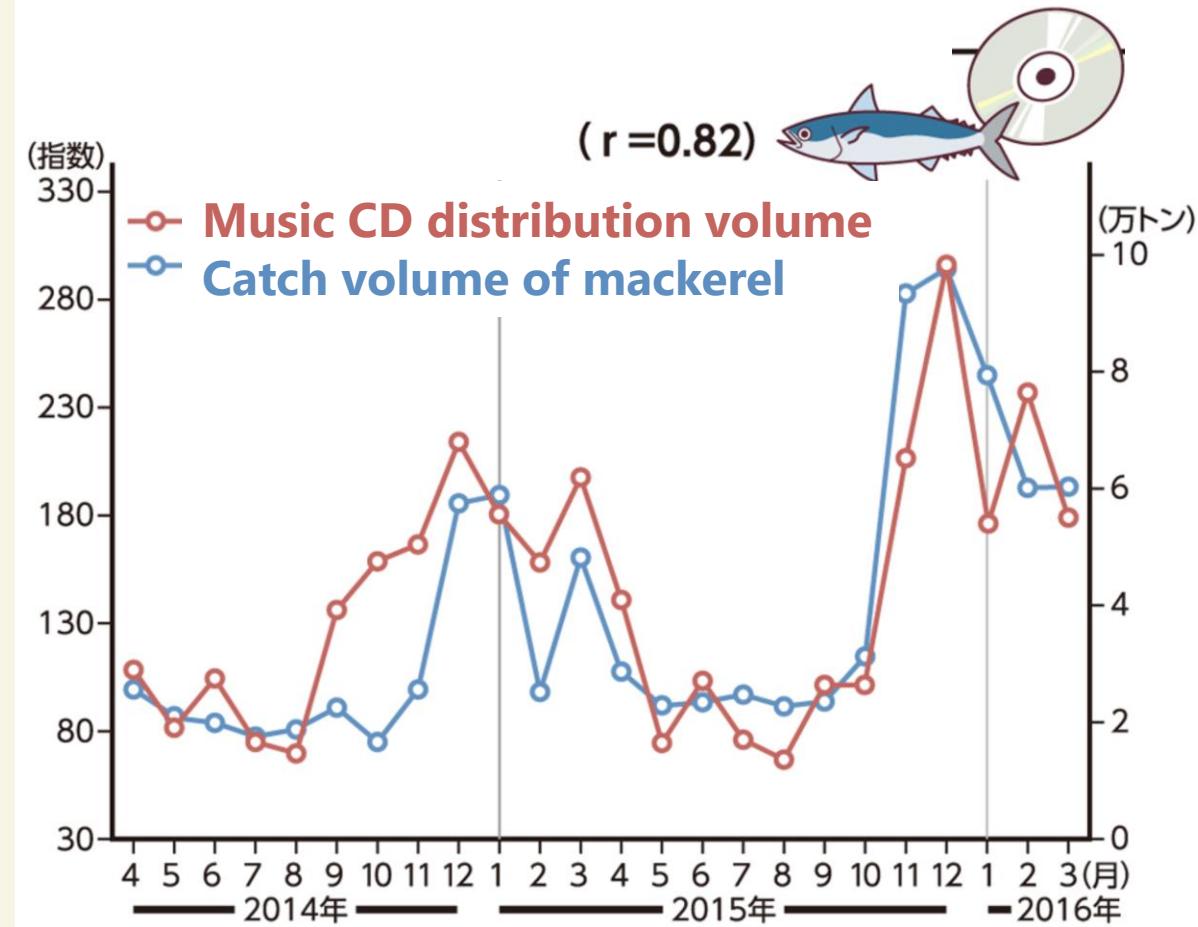
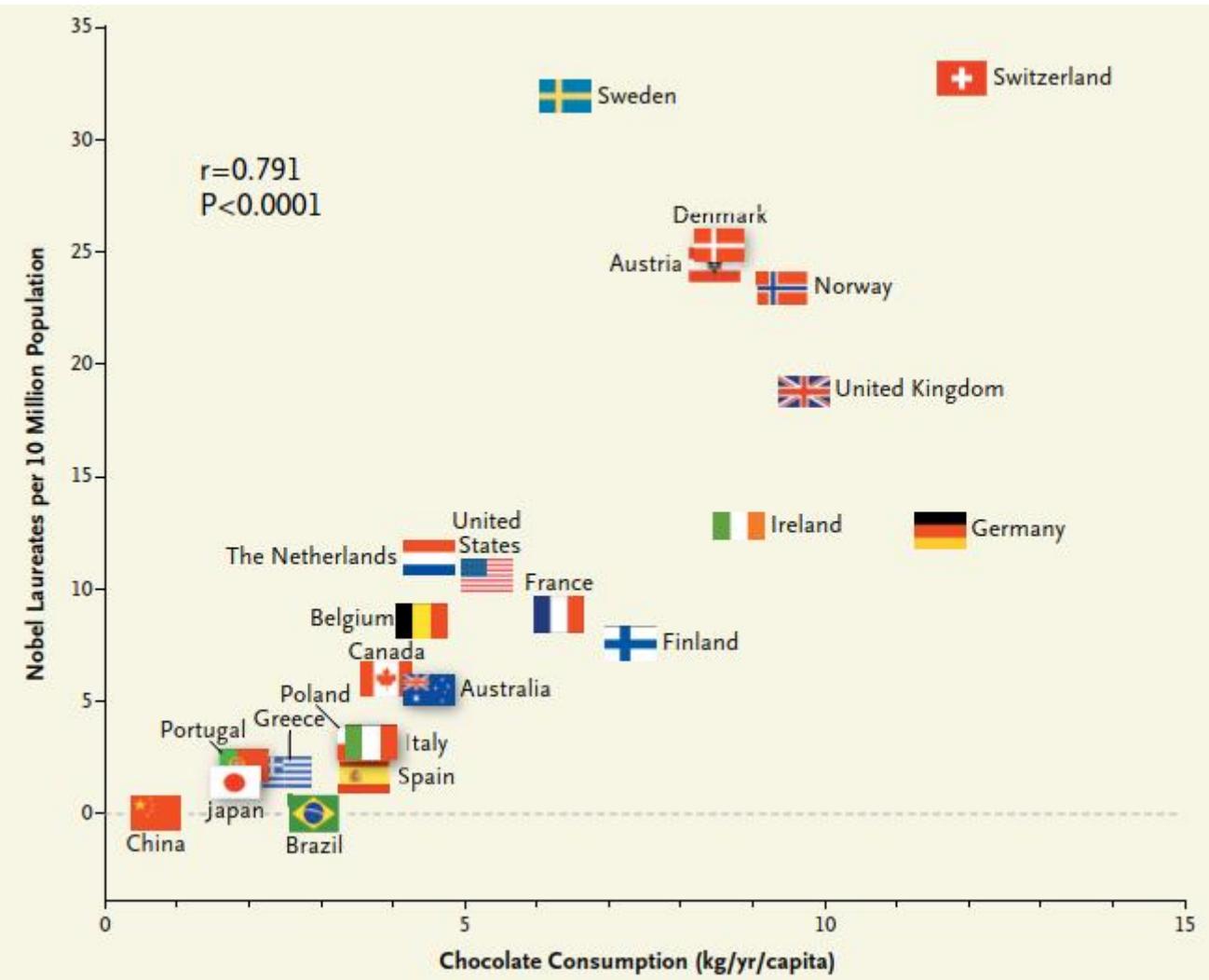
- 30 Hz movies for ~6 hr
 1.08×10^5 pupil images/rec. of mouse

5th grade med. student
Presentation award at
Neurosci conf. 2022



Future direction

Correlation and causation

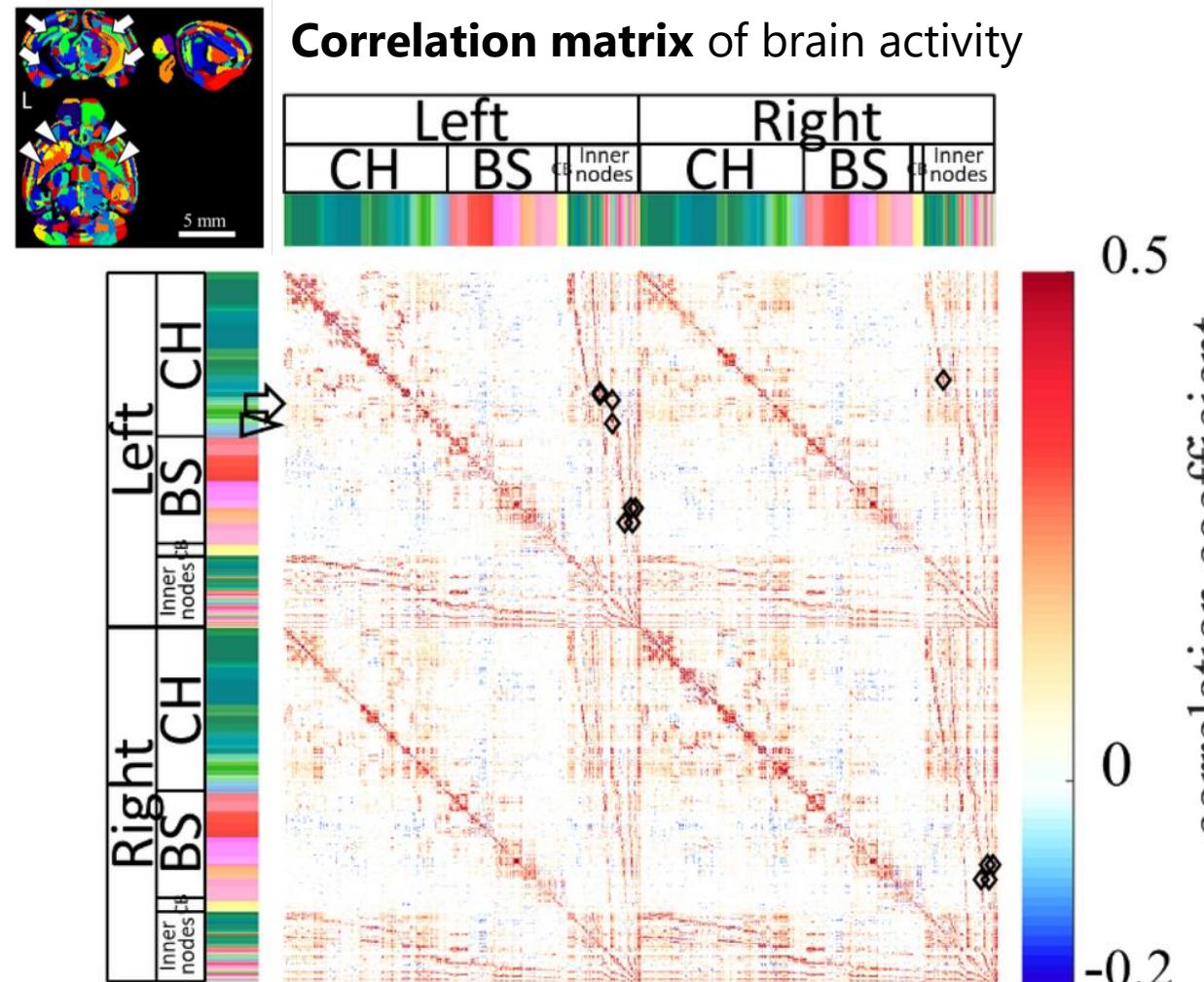
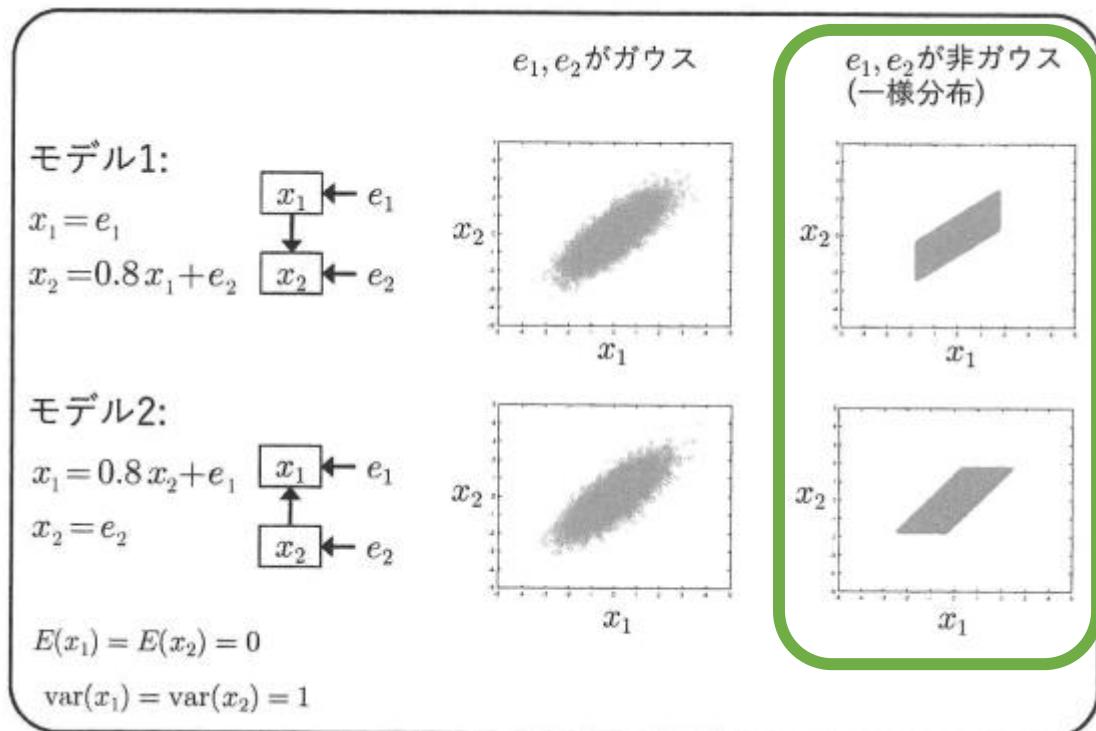


Four types of correlation without causation

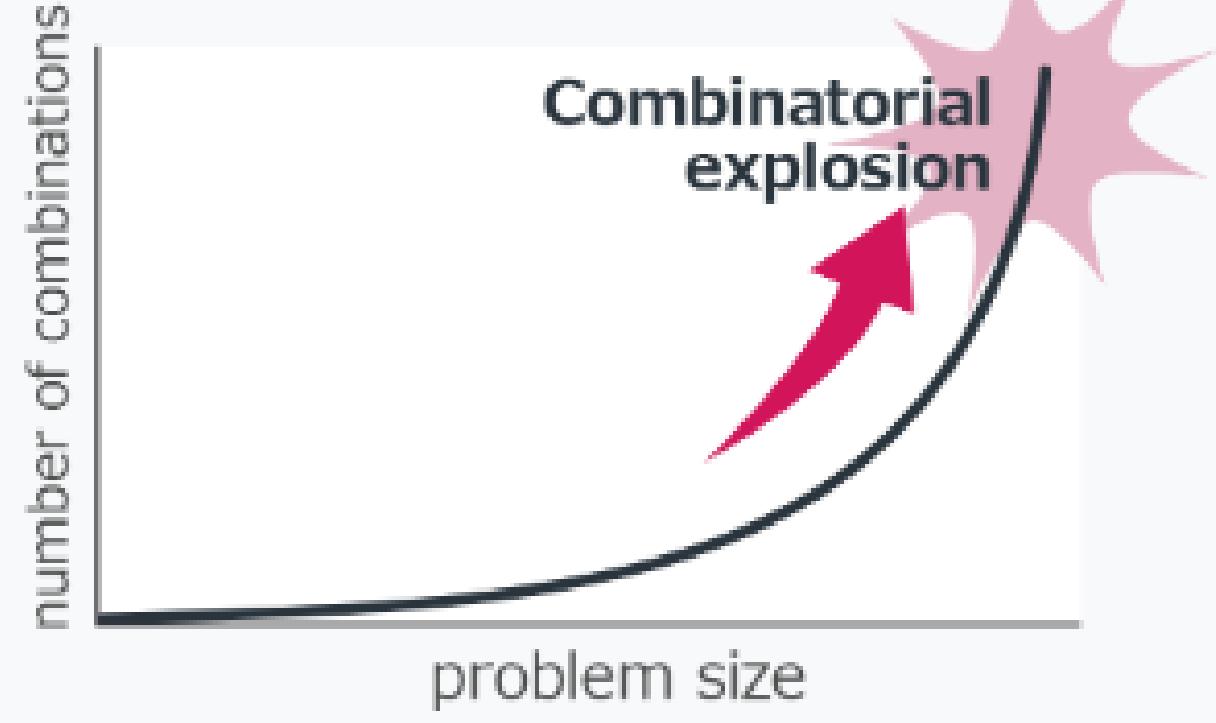
Common causes		Higher height, higher academic score	In elementary school, height and score generally improves with grades
Reversed causation		More warning signs, more accidents	Signs are usually put at places with many accidents
Coincidence		More fish catches, more jacket sales	Just happened to be correlated
Selection bias		More rains, more train use (in Tokyo)	Biased samples caused less common correlation

Causal discovery

Assuming non-gaussian residual (noise) e_1 & e_2
and linearity between variables x_1 & x_2 ,
causal order become identifiable.



Combinatorial explosion in the use of Big Data

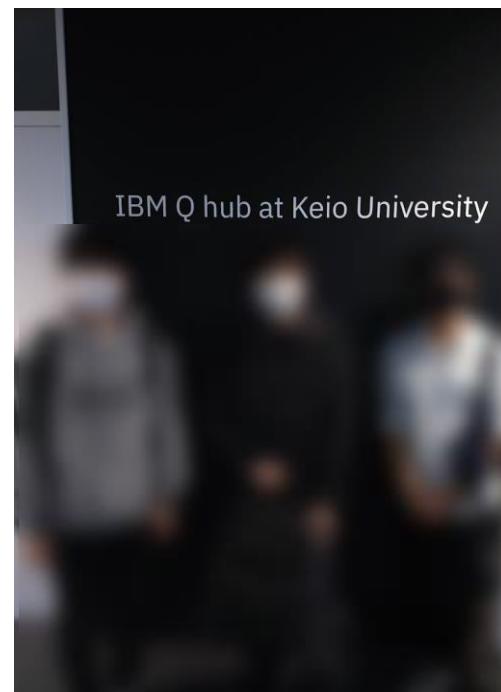


[Large-scale problems by Toshiba](#)

Keio Becomes First Private University to Be Selected under World Premier International Research Center Initiative

October 20, 2022

Keio University's Center for "Human Biology-Microbiome¹-Quantum Research Center (Bio2Q)" has been chosen to be a part of the 2022 World Premier International Research Center Initiative (WPI). It is the first such center among private universities.



@ Keio Univ.
Quantum Computing Center
2020 Oct. with the 1st grade students

Advice from one of the developers of the supercomputer Fugaku



医学の世界もAI使って、画期的なことが起こるのは間違いないので
院生もコンピュータに興味持ってね、っというか
大発見のチャンスが目の前にゴロゴロ転がってますよー。

There is no doubt that groundbreaking things will happen in the world of medicine using AI, so graduate students should also be interested in computers.
The opportunities for great discoveries are lying right in front of you!



#1 in Top500 Super Com. list during 2020-2021. Now #2.

References

Big Data

- Dawn E. Holmes (2017) "Big Data: A Very Short Introduction"
岩崎学・訳 (2020) 「ビッグデータ超入門」東京化学同人)
- [Google Cloud's topic on "Big Data"](#)

Medical Big Data

- 田中博 (2020) [「医療ビッグデータ」](#)羊土社
- 田中博 (2017) [「ビックデータとAIによる創薬」](#)Webスライド

Analysis of Big Data

- Lee and Yoon (2017) Kidney Res Clin Pract "Medical big data: promise and challenges"
- Fan et al. (2014) NSR "Challenges of Big Data analysis"

Other topics

- Combinatorial explosion [日本科学未来館 「フカシギの考え方」](#) YouTube
- AlphaCode [テレ東BIZ・橋本幸治の理系通信 \(2022/2/22\)](#) YouTube