データサイエンス特別講義

Machine Learning in Cosmology and Fundamental Physics

第5回 ハイパフォーマンスコンピューティング

世界スパコンランキング　最新版(2025.6)

1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581	
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607	
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698	
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	793.40	930.00	13,088	
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84		
6	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461	

More sources for one energy

Energy diversification and technological neutrality guide our strategic choices on our path to decarbonization.

Eni for 2024. Check out our interactive feature
More sources for one energy
Born to be unique
HVO biofuels, the Tour d'Europe makes a stop in Rome

Press releases

See all >

04 JUNE 2025 - 12:32 PM CEST
Eni: report on the purchase of

28 MAY 2025 - 4:35 PM CEST
Eni and the Ministry of

28 MAY 2025 - 3:00 PM CEST
Eni: report on the purchase of

28 MAY 2025 - 10:00 AM CEST
Plenitude and Marelli sign

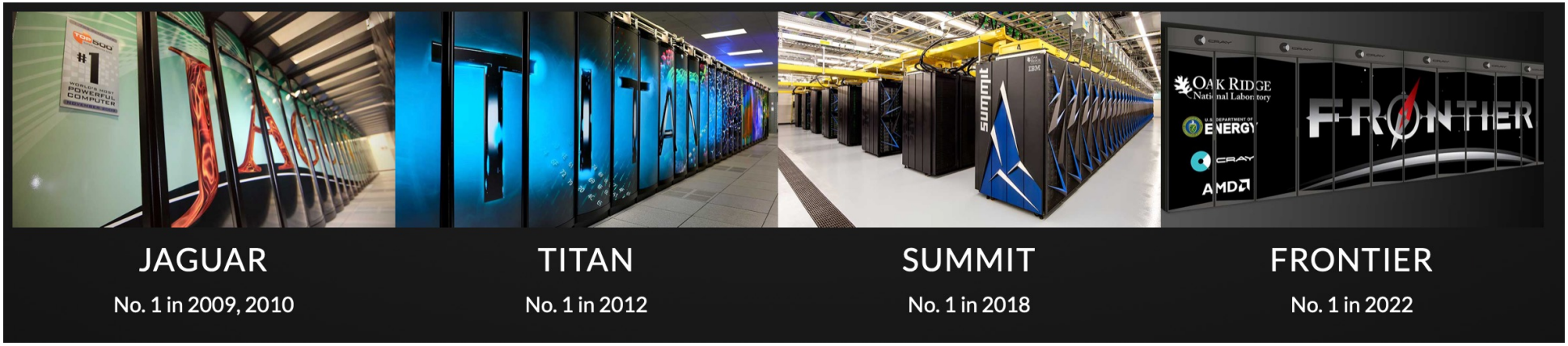
AIはエネルギー問題の救世主か，あるいは問題を悪化させるのか

これ以上優れたAIは必要なのか. Why ?

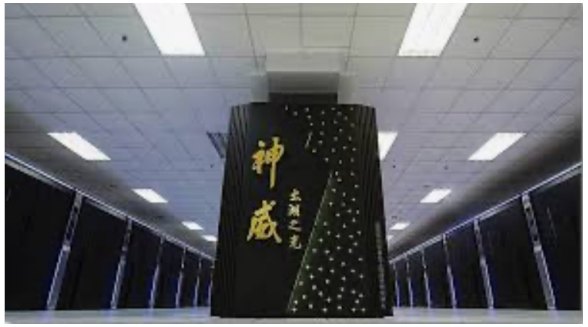
これ以上 _CO2を出してまで_ 何を知りたいのか

後ほどの議論

歴代ナンバー１



この他の年は日本の「京」や「富岳」
中国の「天河」「神威太湖之光」



何をする速さがすごいのか

LINPACK: Linear equation package [線形方程式ソルバー]

$$x + y = 2$$
$$2x + 3y = 5$$

また出た!
つるかめ算を思い出そう

行列の形で書くと

$$\begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

$$\boldsymbol{A\,x = b}$$

答えは

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\boldsymbol{x = A^{-1} b}$$

逆行列

LU分解という
手法で逆行列を
構成し，解を
求める．その速さ

スーパーコンピュータ 富岳

158,976 ノード (52 CPU/ノード)
= 763万個のCPU
160 ペタバイトディスク + クラウド
442 ペタフロップス
2020年-2021年世界最速

地球上の78億人全員が電卓をもち、24時間365日休みなく
計算し続け、それでも1000年かかる計算を10分でできる,
らしい (よく分からん! がとにかく速い)

懐かしの行政事業レビュー 2009

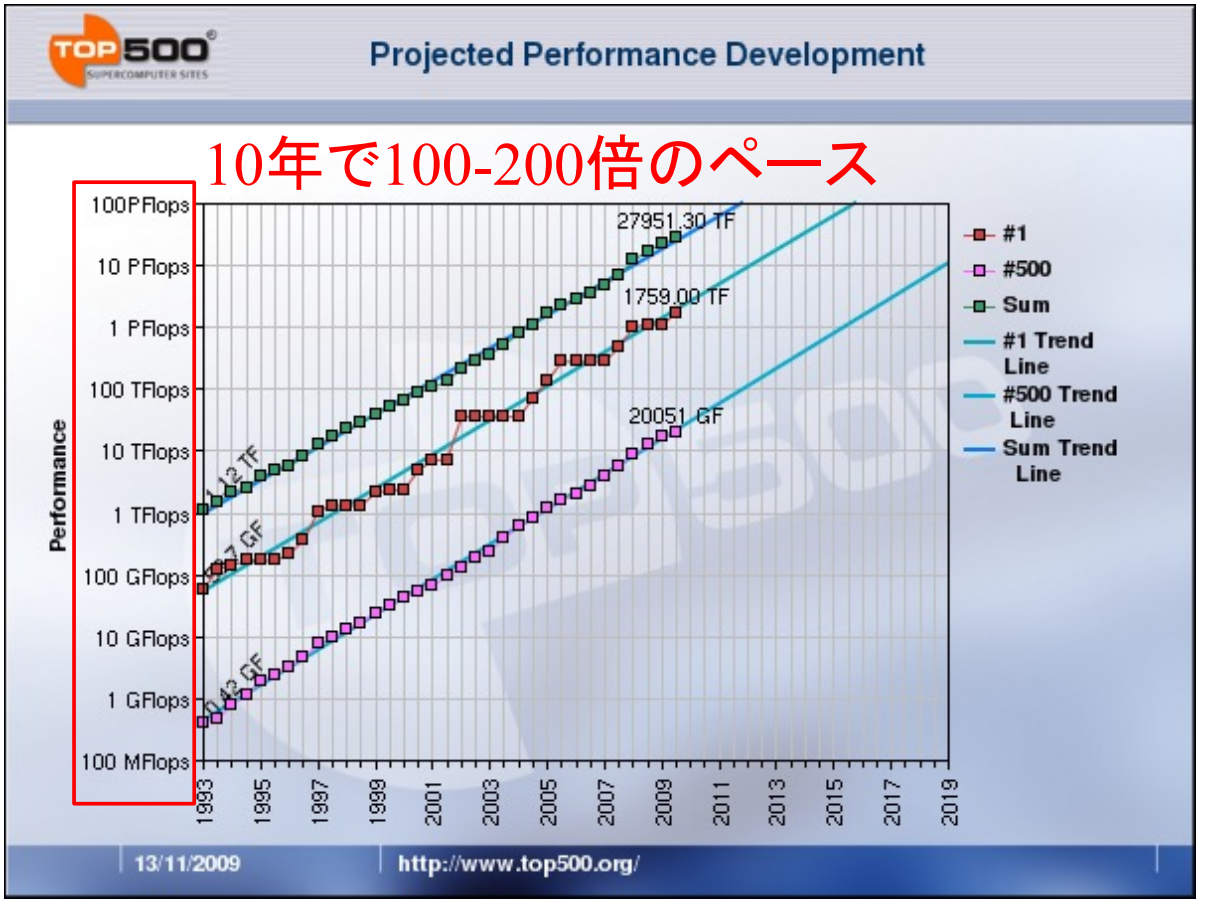
「2位じゃだめなんですか？」

数ある事業の中で最も説明がなされてこなかった

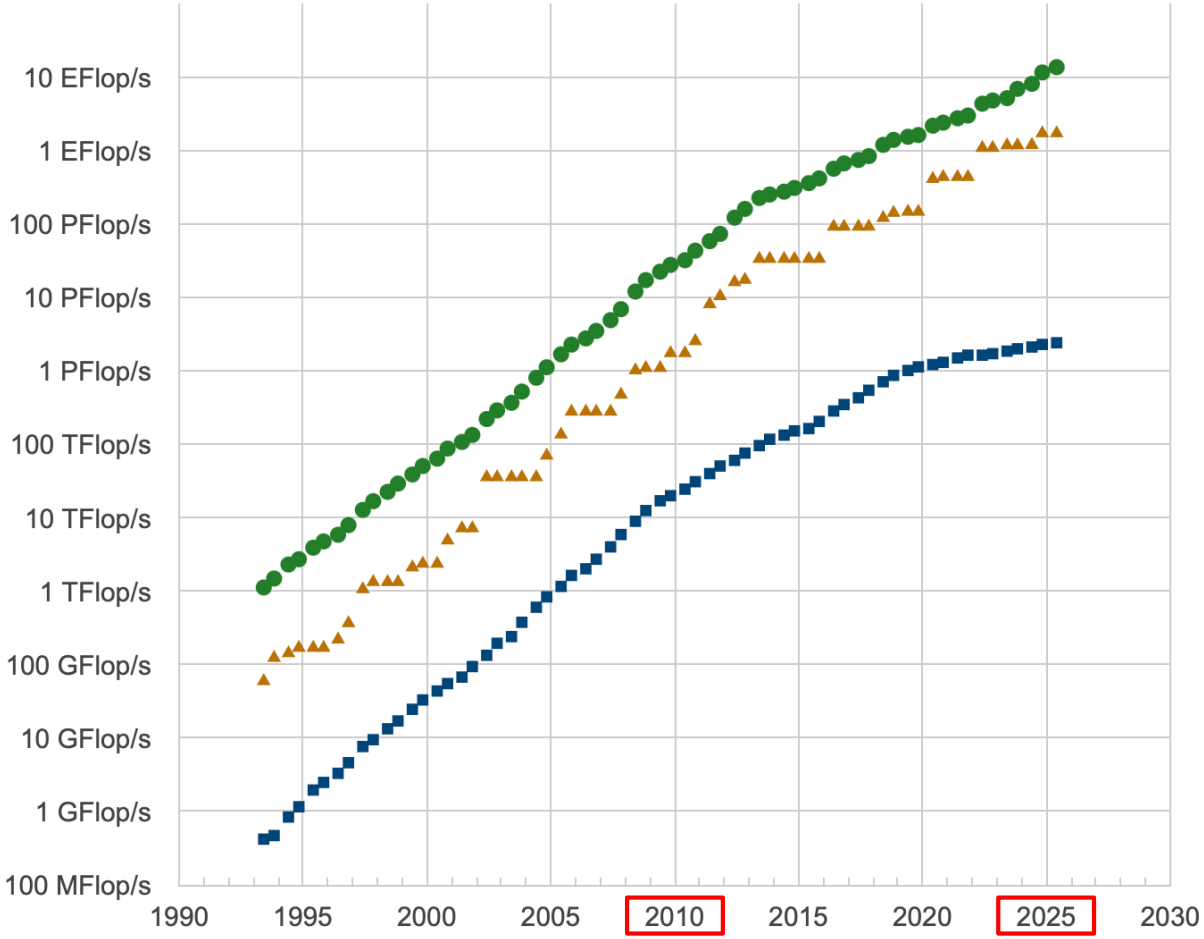
社会にどんなインパクトが
あったのか数字で示すことが
大事、として年間110億円に上る
維持費に見合う成果を要求

何かその、動く心臓を見せるのが悪いとは言えませんけれども、
それがいったいどういう科学的成果なのか全くわからないですし、
何かその、ろう人形館に行って、本物のアートでもなく、何か凄い技術をもって何か
ものまねをしたものを見せられているようなところがある

スーパーコンピュータの性能: 2010年頃の講義スライド



その後の進歩



top500.org
HPより

スパコンで何を計算をしているのか

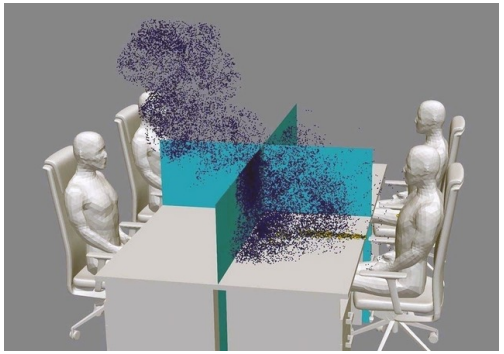
富岳 成果創出加速プログラム

領域１　人類の普遍的課題への挑戦と未来開拓
　　脳, がん, プラズマ, 宇宙

領域２　国民の生命・財産を守る取組の強化
　　創薬, 地震, 気象

領域３　産業競争力の強化
　　エネルギー, 電池, 機械

領域４　研究基盤
　　医療支援技術



10年前くらいの重点研究課題

創薬

パンデミック

ゲノム

ゲリラ豪雨

地震・津波災害

大規模文献

農業フィールド
センシング

宇宙（基礎物理）

JST CREST ビッグデータ 2014-2020　JST AIP加速 2020-2023

使用時間はどう決めてるの?

日本では基本的には公募制で、課金は多くはない

たとえば富岳コンピューターの場合、先述の重点課題枠、一般公募枠、若手育成枠、
商用貸し出し枠などいくつかおおまかな割り振りがあり、その中で課題科学目標や必要資源
を記した申請書をだしてもらい、審査(peer review).
審査にとおると、一定のCPU時間があたえられ、お金は(さほど)払わなくてよい。
欧州はいろいろ。一部のスパコンでは審査+課金(アメリカ式). 中国はよくわからない.

ちなみに、最近のスパコンでは1日を丸ごと借り切って計算するということはほとんどない。
だいたいは全体の10%ほどを何日か使う、くらいが大口ユーザー。
逆に、そういう100%使用の(かつ意味のある)需要があるなら、おそらくそれ専用の
スパコン1台が会社や組織できちんと調達されることに。金融とか軍事とかAIとか。

基礎科学分野では、使用を競合するのは宇宙、素粒子、プラズマ、気象など。
これらでCPU時間を分け合う(というか取り合う).

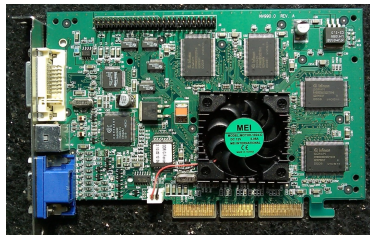
世界スパコンランキング　最新版(2025.6)

1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581	
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607	
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698	
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	793.40	930.00	13,088	
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84		
6	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461	

GPUの発展史：誕生から汎用計算, AI革命へ

1970年代～1990年代：グラフィックス専用時代

画面描画専用チップとして誕生。**3Dゲーム**とともに進化。



2000年代前半：プログラム可能性の獲得

研究者が**並列処理能力**に着目し、グラフィックス以外の用途を模索。

2007年：CUDA登場 - 汎用計算革命

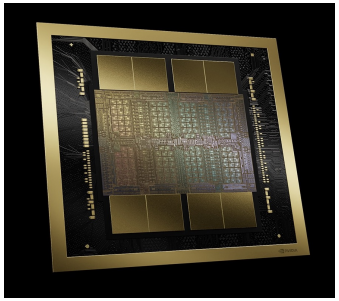
NVIDIAがCUDAを発表。**直接科学計算が可能**に。

2007年～2012年：科学計算の主流へ

スパコンへ搭載開始。気象・創薬など**科学技術計算**で活用。

2012年：深層学習革命の幕開け

AlexNetがImageNet競技会で圧勝。**GPUによる学習の優位性**を実証。



2020年～現在：生成AI時代の主役

ChatGPT等の**大規模言語モデル**の学習に不可欠な存在に。

第1問:「富岳」はどのぐらい広い部屋に置かれているでしょう？

A 学校の教室2つ分ぐらい
B サッカーコート半分ぐらい
C 野球場3つ分ぐらい

第２問:「富岳」が全力で計算しているとき、1日にどのぐらい電気を使う？

A　4人家族の家の10年分
B　4人家族の家の100年分
C　4人家族の家の1000年分

スパコン，AI, … エネルギー問題の解決? 悪化？

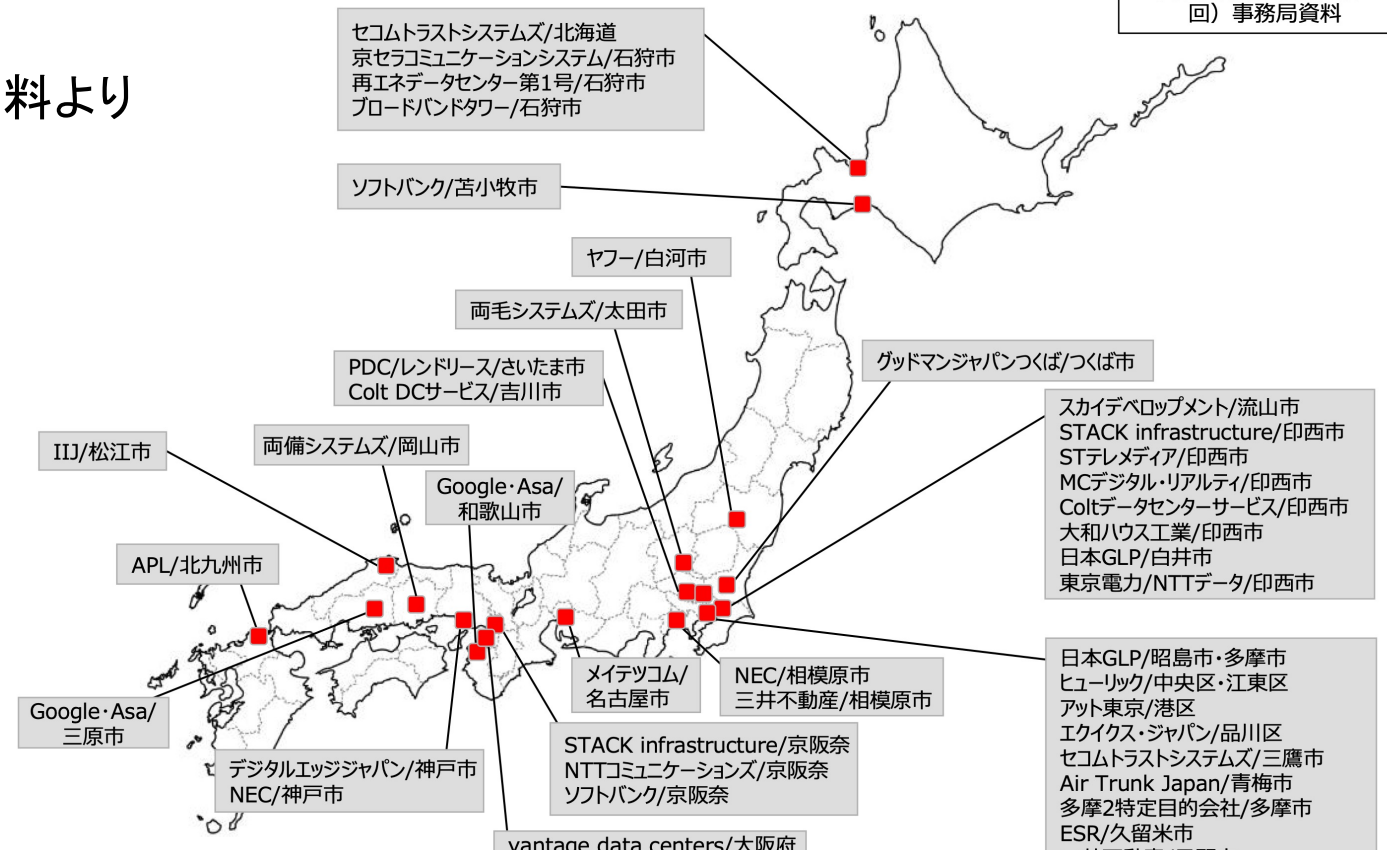
近年相次いで建設されている「ハイパースケールデータセンター」は数十MW（メガワット）を超え、なかには100MWを超える規模のデータセンターも計画。一般家庭の消費電力に換算すると:
家庭の消費電力は電力会社と交わしている「契約アンペア数」が指標。単身世帯だと30A、少し大きめの家だと50A程度の契約になります。一般家庭の電圧は100Vなので、3,000VA～5,000VA、つまり3kW～5kWになります。
ハイパースケールデータセンターの50MWは、一般家庭の契約容量に換算すると1万～1万6千軒分。

エネルギー資源庁資料より

（参考）2024年以降のデータセンターの新設計画

※コンテナ型データセンターなど、
小規模のものや増床案件は除く

デジタルインフラ（DC等）整備に関する有識者会合（第７回）事務局資料



出典：インプレス総合研究所　データセンター調査報告書2024
を基に総務省作成

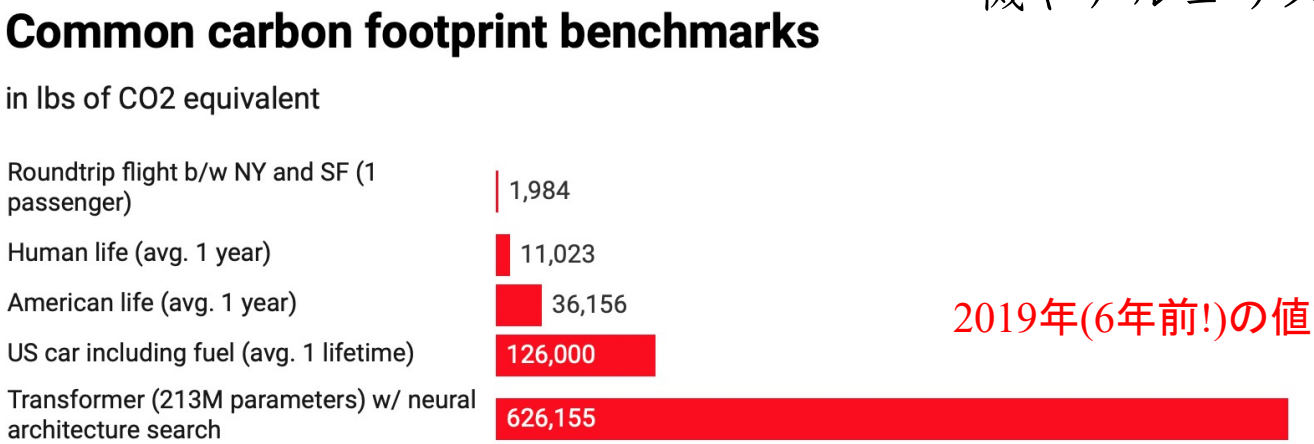
Calculating the environmental footprint of AI at Google

https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference?hl=en

Transformer

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]



"Energy and Policy Considerations for Deep Learning in NLP"
E. Strubell, A. Ganesh, A. McCallum, 2019, arxiv:1906.02243

- 論文では機械の学習にかかる時間やパラメータ依存性も報告すべき
- 研究者は皆同等の性能の計算機にアクセスできるべきである
- エネルギー消費のことも考えた計算機やアルゴリズム開発が必要

2019年(6年前!)の値であることに注意

Transformer

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
E. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell
http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

Transformerの脅威

日本経済新聞

朝刊・夕刊　ストーリー　Myニュース　日経会社

トップ　速報　マネー　経済・金融　政治　ビジネス　マーケット　テクノロジー　国際　オピニオン　スポーツ　社会・くらし　地域　文

研究者2人を解雇したGoogle AI倫理対立の深刻度

コラム　＋フォローする

2021年3月14日 2:00 [有料会員限定]

保存

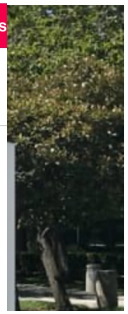
MIT Technology Review

Sign in　Subs

Topics　Magazine　Newsletters　Events

Artificial intelligence / Machine learning

We read the paper that forced Timnit Gebru out of Google. Here's what it says.



Encoding bias: Google vs Gebru

4 UNFATHOMABLE TRAINING DATA

The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics [18, 38, 42, 47, 61] resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status [11, 12, 69, 69, 132, 132, 157]. In this section, we discuss how large, uncurated, Internet-based datasets encode the dominant/hegemonic view, which further harms people at the margins, and recommend significant resource allocation towards dataset curation and documentation practices.

4.1 Size Doesn't Guarantee Diversity

The Internet is a large and diverse virtual space, and accordingly, it is easy to imagine that very large datasets, such as Common Crawl ("petabytes of data collected over 8 years of web crawling",[11] a filtered version of which is included in the GPT-3 training data) must therefore be broadly representative of the ways in which different people view the world. However, on closer examination, we find that there are several factors which narrow Internet participation, the

Encoding bias: Google vs Gebru

Gebru 論文 Section 4 のまとめ:
GPT-2 モデルのトレーニングデータ（文）はRedditからとってきた。最近(2019)の調査によると Reddit ユーザーの67 % は男性。そのうち64%は18-29歳。
- ちなみに Wikipedia を書いて投稿する人のうち女性は ~10% くらい。９割男

GPT-3モデルは Common Crawl と呼ばれる計570ギガバイト のネット上データベースで学習。その結果，「害のある」文章もたくさん生成するようになってしまっている。
GPT-2のトレーニングデータも 272,000 もの「信頼できないニュース」を含み
63,000 個もの「禁止された」ソースから拾っている。

インターネットにあふれる「豊富な」文章を使うと自然言語モデル(AI)は
“hegemonic views that are harmful to marginalized populations”
を持つようになる。

結果として、これらの「最先端の」モデルがさらに何ペタバイトもの偏向的な
文章を生成し、それが次世代の言語モデルの基礎データとなってしまう。
- encoding bias

たとえば，
過激なAI

THE RADICALIZATION RISKS OF GPT-3 AND ADVANCED NEURAL LANGUAGE MODELS

Kris McGuffie
Center on Terrorism, Extremism, and Counterterrorism
Middlebury Institute of International Studies
kmcguffie@middlebury.edu

Alex Newhouse
Center on Terrorism, Extremism, and Counterterrorism
Middlebury Institute of International Studies
anewhouse@middlebury.edu

September 9, 2020

1 Executive Summary

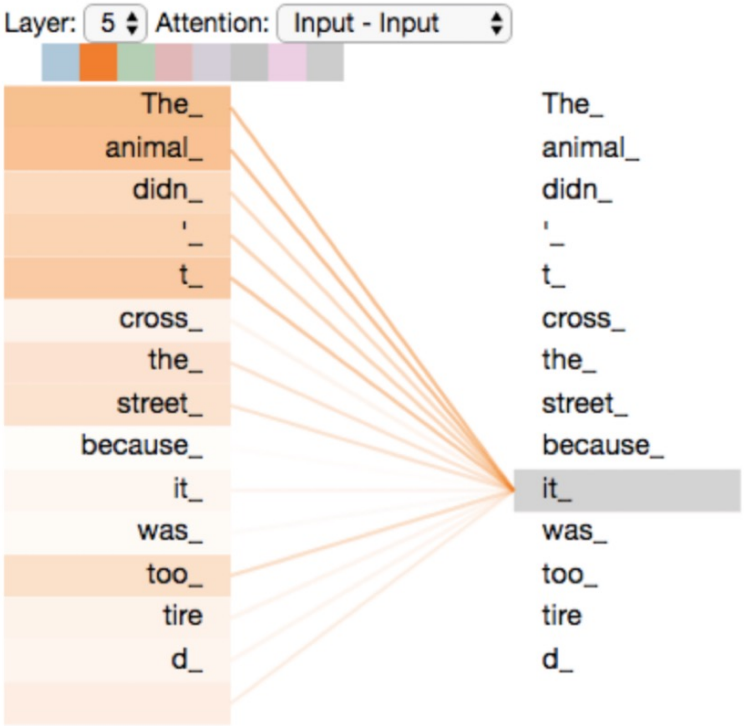
In 2020, OpenAI developed GPT-3, a neural language model that is capable of sophisticated natural language generation and completion of tasks like classification, question-answering, and summarization. While OpenAI has not open-sourced the model's code or pre-trained weights at the time of writing, it has built an API to experiment with the model's capacity. The Center on Terrorism, Extremism, and Counterterrorism (CTEC) evaluated the GPT-3 API for the risk of weaponization by extremists who may attempt to use GPT-3 or hypothetical unregulated models to amplify their ideologies and recruit to their communities. Our methods included:

1. using prompts adapted from right-wing extremist narratives and topics to evaluate ideological consistency, accuracy, and credibility, and
2. evaluating the efficacy of the model's output in contributing to online radicalization into violent extremism.

次の英文を訳してください

Transformer の根幹 - Attention 機構

“The animal didn’t cross the street because *it* was too tired.”
“The animal didn’t cross the street because *it* was too wide.”

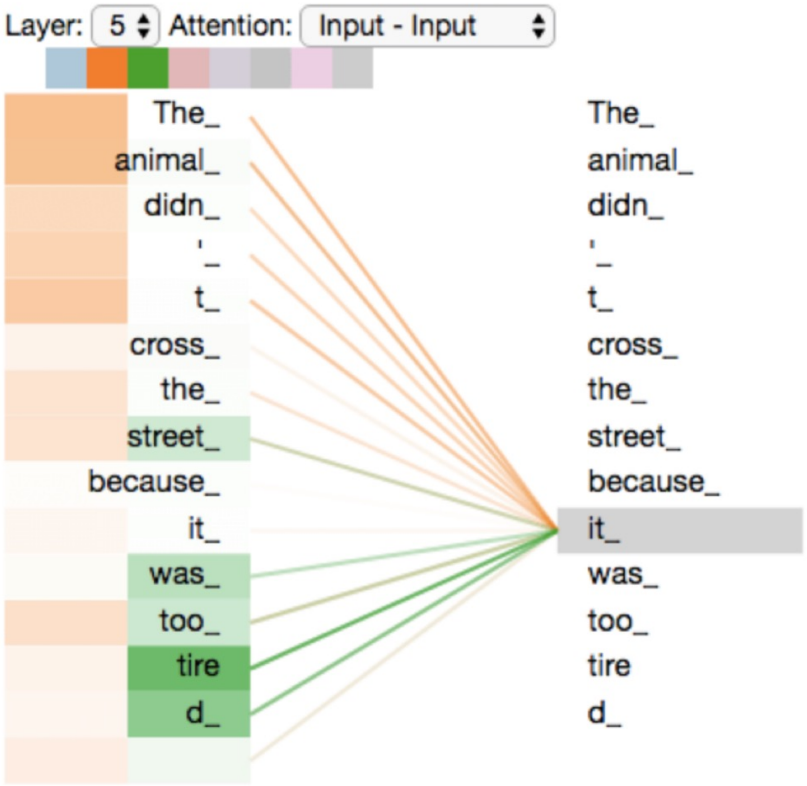


いわゆる”直訳”, word by word translation がうまくいかない例.

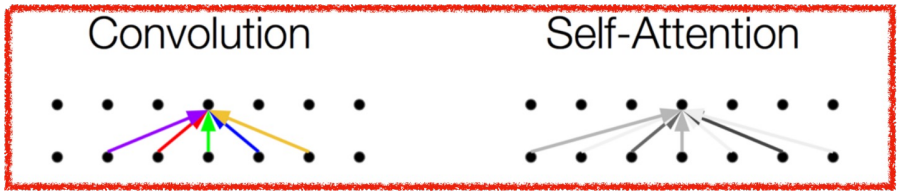
Attention 機構は他の単語に重みやスコアをつけ，たとえば “it” が何に相当するか「確率的に」推定することができる.

Attention 機構

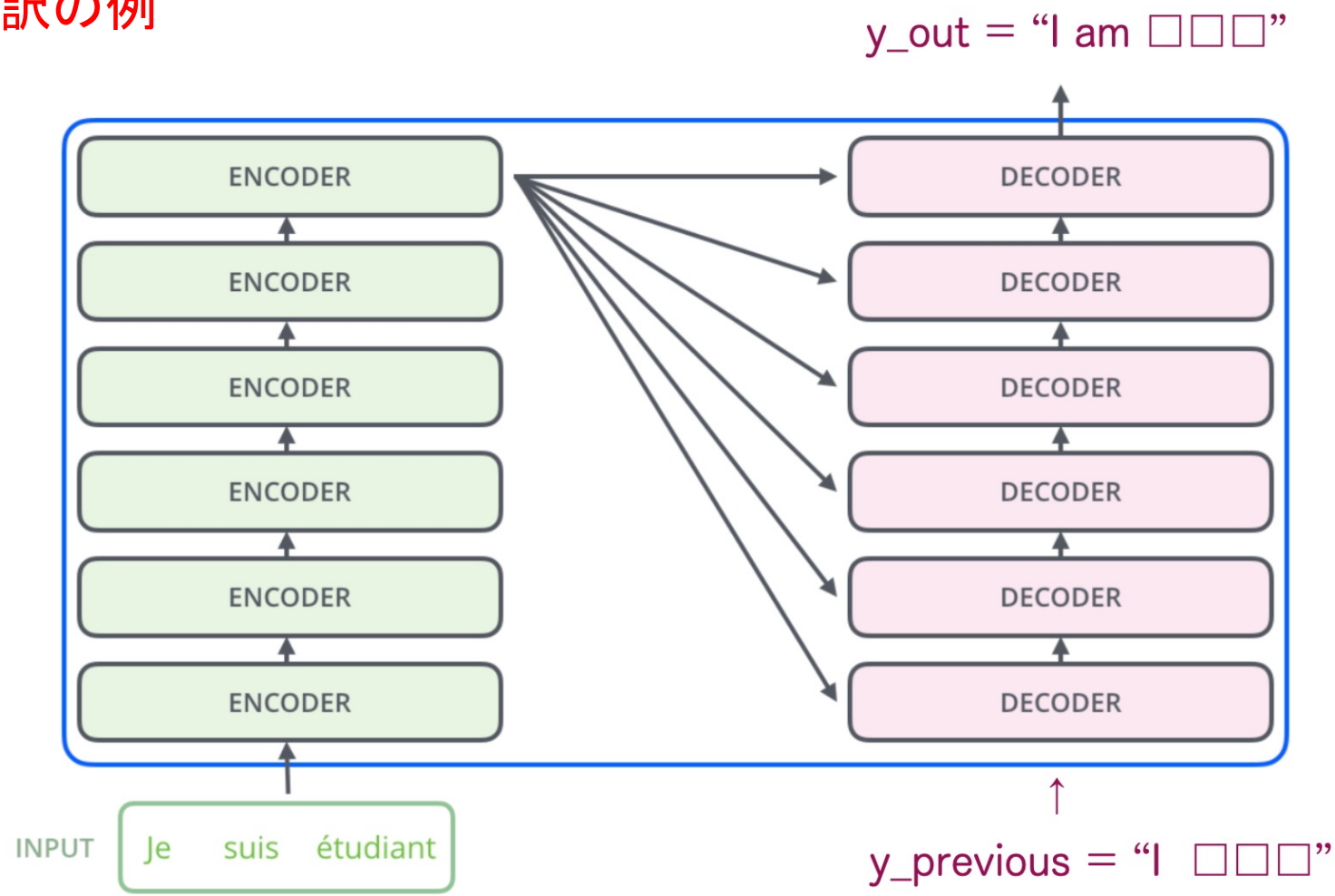
実際には多頭(多重)判断をおこなって,
システムとしてやや冗長だが汎用性と翻訳の精度を上げる.



Convolutional Neural Network
との違いはこの,
離れた箇所への“注意”能力



仏語翻訳の例


y_out = “I am □□□”
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
ENCODER
DECODER
DECODER
DECODER
DECODER
DECODER
DECODER
INPUT
Je suis étudiant
y_previous = “I □□□”

Transformerの作業

Transformerは初めに各単語の埋め込み表現（数値）を生成する.
左の図で〇印. 次に attention 機能を使って他の単語からの情報を集め,
文脈に応じた新しい表現（数値）を提案する. 左の図で●印.
これを全単語に対して行い, 次に埋めるべき単語を推定する.
たとえばフランス語だと4万-5万個の単語についてこれら基礎データを学習する.

話をもとにもどして 「スパコンで実際には何を計算をしているのか」

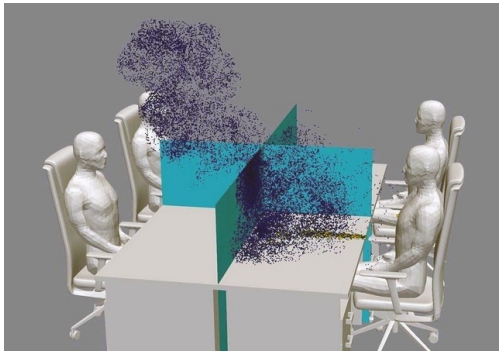
富岳 成果創出加速プログラム

領域１ 人類の普遍的課題への挑戦と未来開拓
脳, がん, プラズマ, 宇宙

領域２ 国民の生命・財産を守る取組の強化
創薬, 地震, 気象

領域３ 産業競争力の強化
エネルギー, 電池, 機械

領域４ 研究基盤
医療支援技術



コンピュータ上に宇宙を再現する　-天の川銀河の形成-

銀河形成のシミュレーション: どんな方程式を解くのか

宇宙全体の膨張 （空間の伸び）

$$H^2(z)/H_0^2 = \Omega_m(1+z)^3 + (1-\Omega_m)\, e^{3\int_0^z d\ln(1+z')[1+w(z')]}$$

密度揺らぎ(ものの集まり具合)の成長

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla_x f + \mathbf{F}\cdot\nabla_v f = 0$$

ガスの振る舞い (流体の運動方程式)

$$\frac{\partial \mathbf{u}}{\partial t} + H\mathbf{u} + \frac{1}{a}(\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{a}\nabla\phi$$

これでも
まだ全体の1部分

化学反応

$$H_2 + H^- \rightarrow H_2 + H + e,\ He + e \rightarrow He^+ + 2e$$

エネルギー方程式

$$\frac{du}{dt} = -\frac{P}{\rho}\nabla\cdot\mathbf{v} - \frac{\Lambda(u,\rho)}{\rho}$$

電磁波（光）の伝搬

$$f_{\rm rad}(\boldsymbol{n}) = \frac{X_{\rm H\,I}}{m_p c}\int d\Omega \int_{\nu_L}^{\infty} d\nu\, I_\nu \sigma_\nu \cos\theta$$

微分方程式を「解く」とは

基礎方程式を解くといっても研究分野や対象によって様々
逆行列を求める，固有値を解く，ゼロ点を求める，
解の組み合わせを求める，などなど

宇宙のシミュレーションの場合，解く，とは時間を進める事
ほとんどの場合，時間微分 ∂ を時間差分 Δt に置き換えて
時間を発展させる：$t = t + \Delta t$

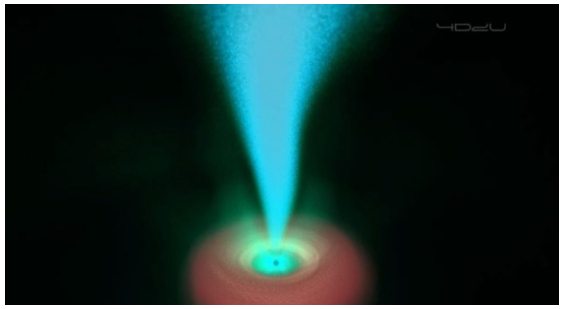
宇宙の現象には分単位で起こるものから数億年かかるもの，
キロメートル程度の現象から数億光年にわたるもの
が共存していて，これまでのようなシンプルな方法では
困難も多い

宇宙のシミュレーション：これまでにできたこと

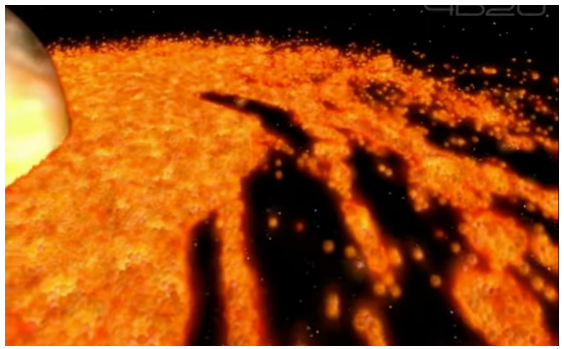
望遠鏡の画像　　シミュレーションの画像

宇宙のシミュレーション：これからしたいこと

ブラックホールの成長, ブラックホールの合体と重力波の生成　(一般相対論)
超新星爆発の爆発そのもの (ニュートリノ、乱流, 元素合成)
ブラックホールからのジェット噴出　(磁気流体力学)
惑星系形成, 月（衛星）の形成, 惑星大気の長期進化
天の川銀河の形成, その中での太陽系の形成



これから必要な事
　計算格子数　(質量粒子)　増大
　　現在は$8000^3 \sim 1$兆個くらい　→ 1京個で銀河をまるごと表現
輻射輸送
　　電磁波・ニュートリノの伝搬　7次元問題　→ 少しづつ進歩



そして、
　よく分かっていないことも多い。
　人間がよく分かっていないことをコンピュータがやって
　スッキリ分かるようになったことは…多分ない。AIならできるだろうか.