

データサイエンス教育・研究の効率をX倍にするにはどうすればよいか

第2回 千葉大学データサイエンスコア オープンカフェ

大田達郎

千葉大学 国際高等研究基幹 / 大学院医学研究院

千葉大学データサイエンスコア 副ダイレクター



CHIBA
UNIVERSITY



データサイエンスコア
DATA SCIENCE CORE

自己紹介

- 1986 福岡県福岡市生まれ
- 2009 神戸大学農学部卒
- 2011 東京大学大学院農学生命科学研究科修了
- 2011-2023 ライフサイエンス統合データベースセンター
 - 2019 総研大にて論文博士取得
- 2023- 千葉大学 国際高等研究基幹 / 医学研究院 人工知能医学 所属
 - 2024/7- 千葉大学データサイエンスコア 副ディレクター
- 専門: バイオインフォマティクス
 - バイオデータベース構築, ゲノムデータ解析システム開発



千葉大学大学院医学研究院
人工知能(AI)医学

データベース統合

トーゴの日シンポジウム

2024

AI+ロボティクス + データベース が 変える 生命科学

10.5 土 品川ザ・グランドホール
一部ライブ配信

#togo2024

パネルディスカッション「生命科学の未来を予想するーデータベースはもう要らなくなる...ってコト?!」

AIがあればデータベースはもういない？

い
ら
な
い
と
は

い
ら
な
い

い
ら
な
い

心配事の多い時代

- AIによって研究者はお払い箱になるのか？
- 自我を持つAIの出現、シンギュラリティの可能性は？

What if

つよつよAI「私が研究しますからあなたは研究費の申請書を書き続けてください」

現実

- まだそうはなってない
- 早く無駄で退屈な仕事を奪ってほしい



AIに奪われるのを待ってられない、奪ってもらえないならこちらから仕事を手放してやる

「データサイエンス教育・研究の効率をX倍にするにはどうすればよいか」

第2回オープンカフェ 開催のお知らせ



DATA
SCIENCE
CORE

2024年7月発足したDSC（データサイエンスコア）の活動内容・今後の事業展開について、学内外からスピーカーをお招きし、カジュアルなミニセミナーをシリーズ実施します。第2回の講師は大田副ダイレクター、データサイエンスを駆使して教育・研究を革新する戦略を熱く語って頂き、ご参加の皆様とご一緒に考える機会としたいと思います。奮ってご参加下さい！

○日 時： 2024年 10月 9日(水) 17:30～19:00
[17:00過ぎ開場、17:30頃プレゼン開始予定]

○会 場： 千葉大学西千葉キャンパス・BiZCAFE
(IMO棟Ⅱ・1F)

○登壇者： 大田 達郎 DSC副ダイレクター
(国際高等研究基幹・医学研究院准教授)



○テーマ： データサイエンス教育・研究の効率をX倍にするにはどうすればよいか

○対象者： 本学学生、教職員（事前申込み制；先着順）

○備 考：

- ・ 開会前にBiZCAFEの簡単な利用案内等を行います。
- ・ 講師プレゼン等は、当日オンライン配信計画中の他、後日DSCウェブサイト（www.dsc.chiba-u.jp）で動画配信の予定です。
- ・ オープンカフェは、当面1～2ヶ月に1回程度開催予定。

問合せ先：DSCダイレクターオフィス・マネージャー satonao1@chiba-u.jp



申込みフォームはこちら

<https://forms.office.com/r/U2fNkBLPW6>

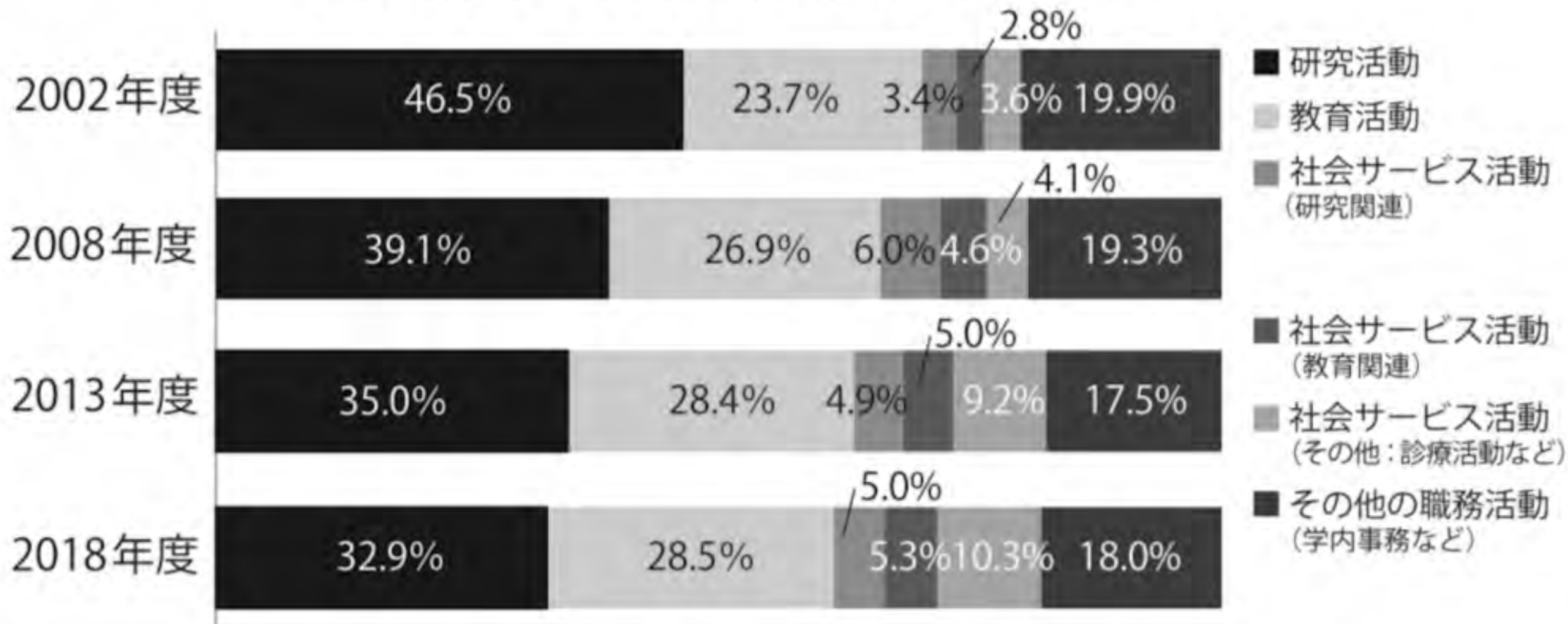
データサイエンス教育・研究の効率をX倍にするにはどう
すればいいかみんなで考えよう会議

なんでこんな話をするのか

- DSCができたおかげで私の（教育|研究）の（速度|質）が急上昇！となるのが理想。主役は学生・教員・事務職員のみなさん。DSCはそれを全力でサポート。
- 自分の目の前の困りごとには気付きやすいが、本質的に解決すべき問題は人と話し合えないとわからない
- 「何を解決すべきか」を見誤らないための議論をしましょう

研究・教育の効率とはなにか、それは測定できるのか

(図) 大学などの教員の職務活動時間割合の推移



目的関数の候補

- 授業の準備時間
- 学会発表のための資料作成時間
- 実験に着手してからデータが出るまでの時間
- データを手渡されてから解析が完了するまでの時間
- 期間あたりの論文数
- 事務的な手続きに必要な書類を作成する時間
- 特定ドメインの問題解決能力を会得するために必要な授業時間数
- 人類が新たな知識を得るまでにかかった金銭的成本
- 人類誕生から新技術の発明まで何年かかったか

これらを X 倍にできるか？ X のオーダーをどれくらい変化させられるか？

変数

- 1つの教育プログラム・研究プロジェクトに関わる人間の数 (労働集約・人的並列化)
- 決断・アイディアに至るまでの思考の速さ・深さ
- 試行錯誤のイテレーションを回すスピードと回数
- ミスのない作業 or 手戻りによるロス
- DX・自動化
- 外注 (時間・知識・技術を金で買う)

何が一番手っ取り早い？

「人間を研究のプロセスから排除する」

研究フローからの人間排除の事例

1. ラボラトリー・オートメーション
2. データ解析ワークフローシステム
3. 大規模言語モデルによるテキストアノテーション

1. ラボラトリー・オートメーション



ご利益

- 人間には難しい細かすぎるパラメータチューニングが可能
- 労働基準法が適用されない



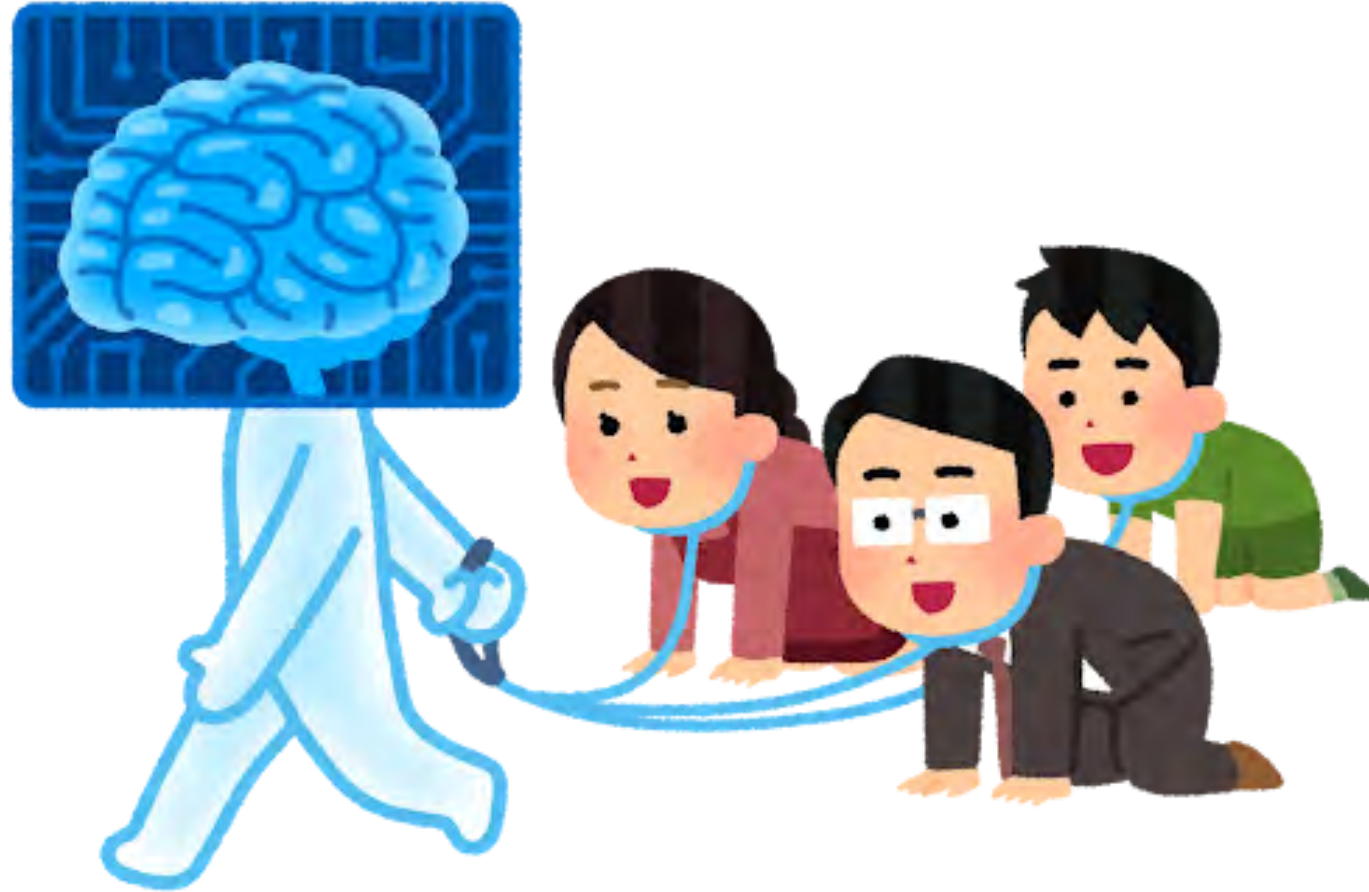
図6 近未来の研究室像

さまざまなロボットやAIが実験の遂行を担当し、研究者は知的創造に専念する。

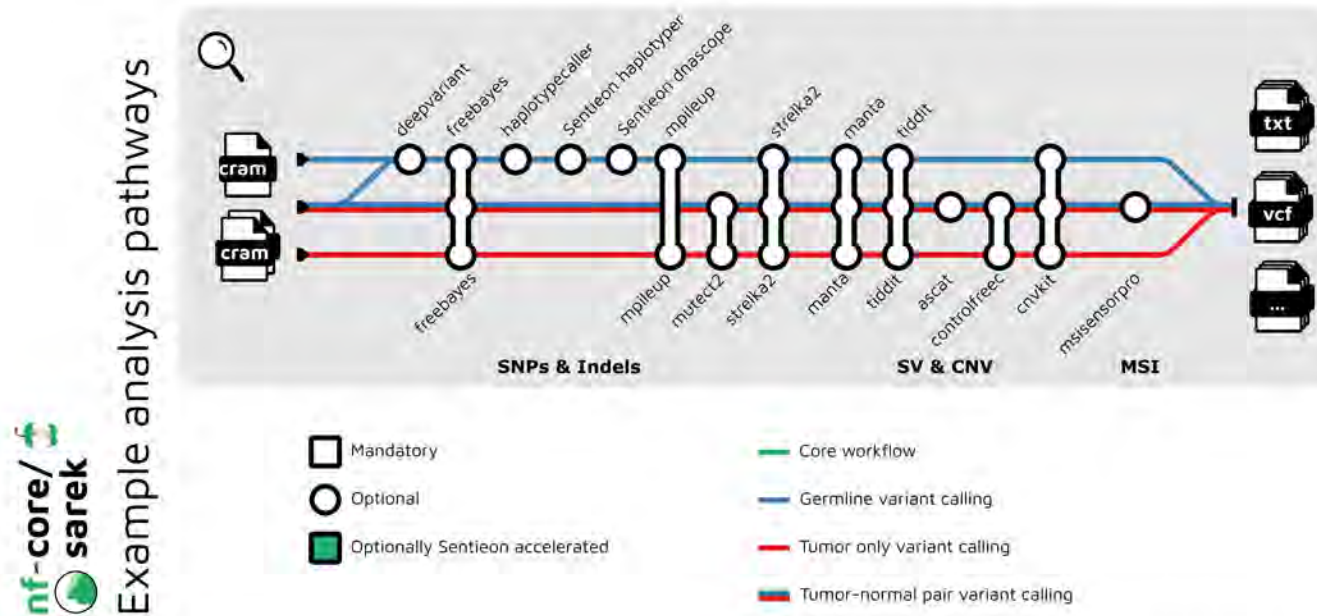
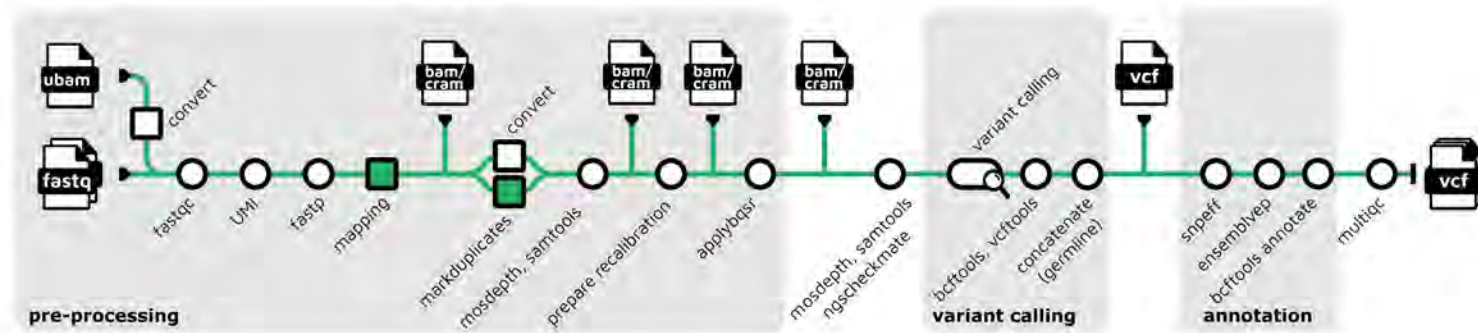
課題

- ロボットが高価
- 人間は安い

実際はこう



2. データ解析ワークフローシステム



nf-core/
sarek

Example analysis pathways

- ☐ Mandatory
- ☐ Optional
- ☒ Optionally Sentieon accelerated

- Core workflow
- Germline variant calling
- Tumor only variant calling
- Tumor-normal pair variant calling

Adapted from: Fellows Yates, James A., et al. PeerJ 9 (2021).

ご利益

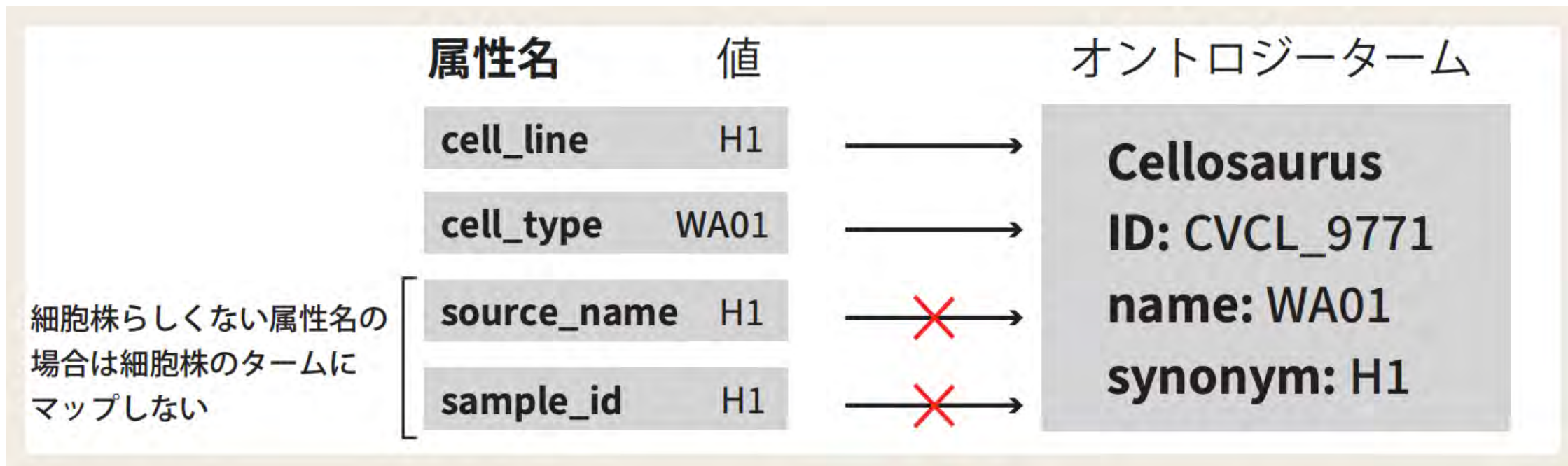
- レビューワーに「このパラメータも試してみたら」と言われても動じない強い心が得られる
- 先輩からもらうコードが「読めないし動かない」から「読めないがちゃんと動く」になる

課題

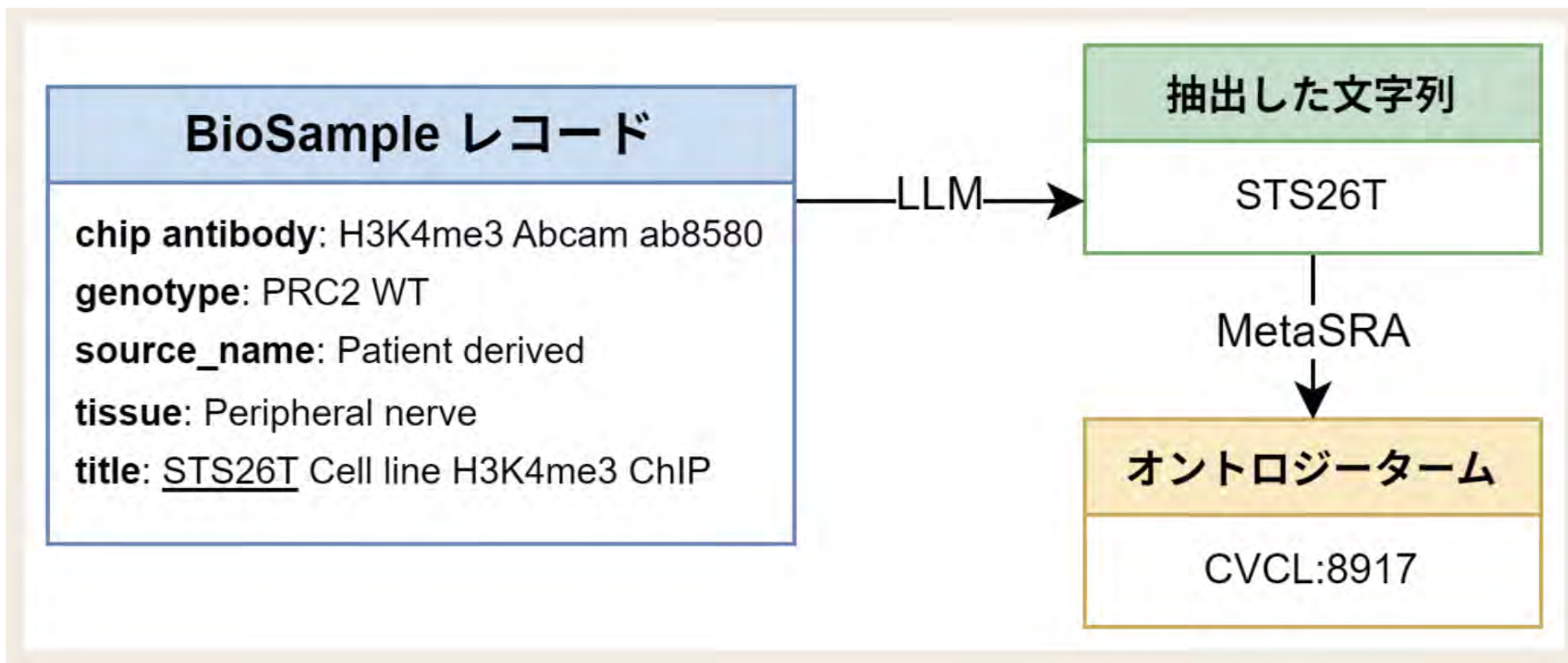
- 組み上げた自動化パイプラインは色々な理由で壊れる
- メンテナンスやアップデートで結局コストゼロにはならない

3. 大規模言語モデルによるテキストアノテーション

文献や自然言語テキストからの情報抽出



一度LLMに投げて抽出させるステップを挟むことで改善



複数のプロンプトで精度を比較

① 基本

細胞株の説明

A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.

入力の説明と
タスクの指示

I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.

出力形式の指定

Your output must be JSON format, like {"cell_line": "NAME"}. "NAME" is just a place holder. Replace this with a string you extract.

When input sample data is not of a cell line, you are not supposed to extract any text from input. If you can not find a cell line name in input, your output is like {"cell_line": "None"}.

② 分化した幹細胞株に注意

A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.

I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.

Your output must be JSON format, like {"cell_line": "NAME"}. "NAME" is just a place holder. Replace this with a string you extract.

When input sample data is not of a cell line, you are not supposed to extract any text from input. If you can not find a cell line name in input, your output is like {"cell_line": "None"}.

Note that some samples are cells differentiated from a stem cell line. In this case, the stem cell line name is mentioned in the metadata, but the sample is not the cell line itself. Therefore, the output must be {"cell_line": "None"}.

③ 幹細胞株そのものなら 抽出することを強調

A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.

I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.

Your output must be in JSON format, like {"cell_line": "NAME"}. "NAME" is just a placeholder. Replace it with a string you have extracted.

If the input sample data is not of a cell line, you should not extract any text from the input and your output must be like {"cell_line": "None"}.

Note that some samples are cells differentiated from a stem cell line. In this case, the stem cell line name is mentioned in the metadata, but the sample is not the cell line itself. Therefore, the output must be {"cell_line": "None"}. Of course, if the sample is considered to be the stem cell line itself, extract the stem cell line name and include it in your output.

ご利益

- 人間の業の深さに魂を削られることがなくなる

Active ammonia oxidizers in an acid soil are phylogenetically closely related to neutrophilic *Nitrososphaera viennensis*

Identifiers	BioSample: SAMD00009749; SRA: DRS012638	
Organism	soil metagenome unclassified entries; unclassified sequences; metagenomes; ecological metagenomes	
Attributes	sample name	DRS012638
	ACAGACAGCGT	CO2-day-56-F9-replicate2
	ACAGACTATAT	CO2-day-56-F10-replicate2
	ACAGAGACTCT	CO2-day-56-F14-replicate2
	ACAGCTCGTGT	CO2-day-56-F11-replicate2
	ACAGTGTCGAT	CO2-day-56-F12-replicate2
	ACGACACGTAT	13CO2-day-56-F7-replicate2
	ACGAGCGCGCT	CO2-day-56-F13-replicate2
	ACGAGTAGACT	13CO2-day-56-F7-replicate1
	ACGATGAGTGT	CO2-day-56-F3-replicate3
	ACGCGAGAGAT	CO2-day-56-F4-replicate3
	ACGCGTCTAGT	13CO2-day-56-F8-replicate1, 13CO2-day-56-F8-replicate2
	ACGCTCTCTCT	CO2-day-56-F5-replicate3

A. adenivorans genomic DNA

Identifiers BioSample: SAMEA2024666; SRA: ERS169829

Organism [Blastobotrys adenivorans](#)
cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Saccharomycotina; Dipodascomycetes; Dipodascales; Trichomonascaceae; Blastobotrys

Attributes	isolation source	wood hydrolysate, Sibiria
	sample name	Arxula adenivorans LS3
	ENA-FIRST-PUBLIC	2014-05-15T17:02:12Z
	ENA-LAST-UPDATE	2018-03-09T10:01:39Z
	External Id	SAMEA2024666
	INSDC center name	IPK-Gatersleben
	INSDC first public	2014-05-15T17:02:12Z
	INSDC last update	2018-03-09T10:01:39Z
	INSDC status	public
	Submitter Id	Arxula adenivorans LS3
	TODO: TAG NAME	TODO: TAG VALUE
	scientific_name	Blastobotrys adenivorans

課題

- 「抽出させたものが正しいか」は判断できるが「正しく抽出できたか」の評価には人間が読まなくてはならない = カバレッジ100%達成の判断ができない
- 商用LLMは強力だが、大量のデータを処理させるにはAPI利用の値段が高すぎる
- オープンなLLMは小型のものとレスポンスが速いが非力、大型のものは計算機パワーが必要な上レスポンスが遅くなるので大量データの処理が難しい
- 今月の電気代がやばい

確かにまだ課題はあるが、便利な時代はもうとっくに来ている

機械が全部勝手にやってくれる！



🤔 全部自動化されると我々はハッピーなのか？

あなたの研究はどこから？

- キュリオシティ・ドリブン
 - 探索の過程がなくなるのは味気ない
- ミッション・ドリブン
 - 省力化できるに越したことはない

大学教育の意義

- 大学とは最先端の学問が実施され、それを学ぶことができる場所
- インターネットでアクセスできるAIが最先端になったとき、大学教育は必要なのか？

みんなで考えよう会議

今の我々に足りない「X倍効率化」を実現するもの

Google Maps がなかったときにどうやって目的地まで行っていたか、覚えていますか？
Google Maps がなかったときに、Google Maps が想像できましたか？

あなたが「これは自分の教育・研究スタイルを変えた(これがなかったら今の自分の研究はない)！」と思った知識や技術、ツールはなんですか？

それは研究や教育におけるどんなボトルネックを解決したと思いますか？

今抱えているボトルネックはなんですか？

それを解決するものはどんな姿をしていますか、誰が作りますか、いつごろできますか？

生命科学のボトルネックとその解消に必要なもの

- 賢く誠実で献身的な科学者 (教育 🎓)
- 新たな現象の発見と観察・仮説構築 (資金 💰 時間 ⌚ 知識 🧠)
- 新たな計測技術の開発 (資金 💰 時間 ⌚)
- 生体試料のサンプリング (資金 💰 社会的合意 🗣️)
- 再現可能な実験とその結果 (資金 💰 時間 ⌚ 技術 🔬)
- 論文の出版・およびサイエンスというゲームのルール (社会的合意 🗣️)
- 人類社会における需要 (予測・制御が難しい 🤔)

まとめ

- DSCでは継続的にこういった機会を設けますのでぜひご発言ください
- DSCをうまく使ってください

前提: 生成AIは凄いが、解決すべき問題もある

SaaSでの提供

- 強力なプロプライエタリ・モデルへの依存 => 透明性・セキュリティ・データ保護

計算機

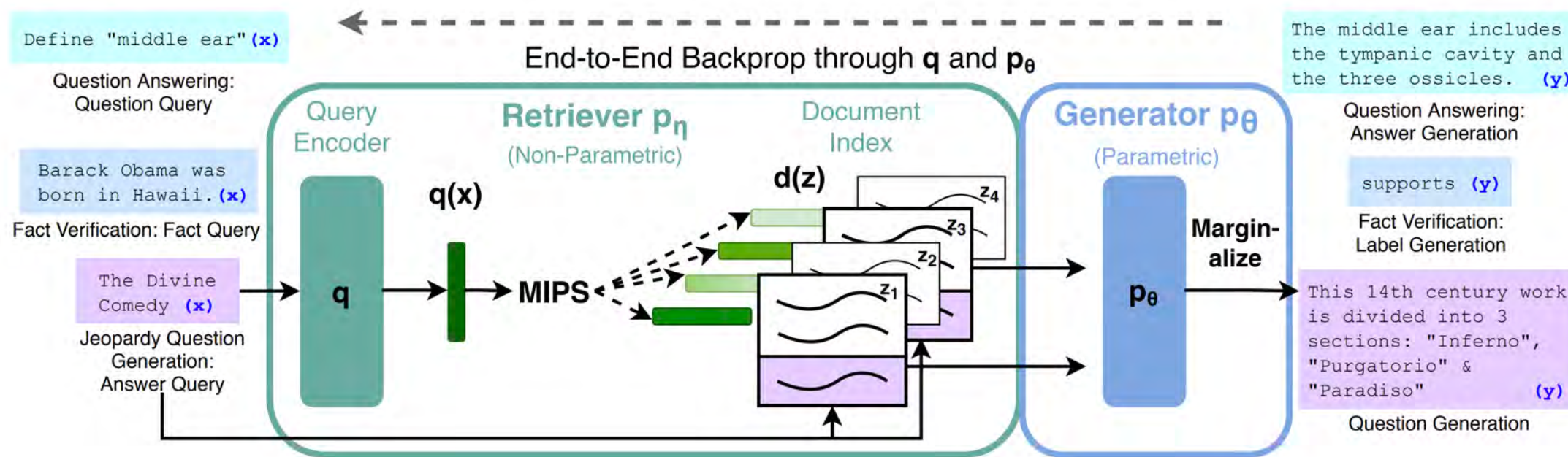
- 高性能GPU計算機への依存 => 特定企業による供給の寡占・膨大な電力消費

応用可能性

- 医療応用における事実誤認など正確性に関するリスク
- 人間とAIの価値観の相違について（倫理）

実応用への障壁: 出力の正確さの担保

Retrieval Augmented Generation (RAG): データベースを基盤にLLMを「強化」



一般にはドキュメントを蓄積したベクトルDBが用いられるが 知識グラフでもRAGは可能

LLMは人における"直感"？

- Daniel Kahneman の著書に曰く、人の思考は「直感的・感情的な速い思考」システム1と「時間のかかる論理的な思考」システム2との組み合わせによって成立している
- 膨大な学習データからパターンを見出すLLMは「AIにおけるシステム1」と見なせる
- 論理的な推論を行うシステム2にあたるのは、本来データベースが担う機能ではないか

‘A lifetime’s worth of wisdom’
Steven D. Levitt, co-author of *Freakonomics*

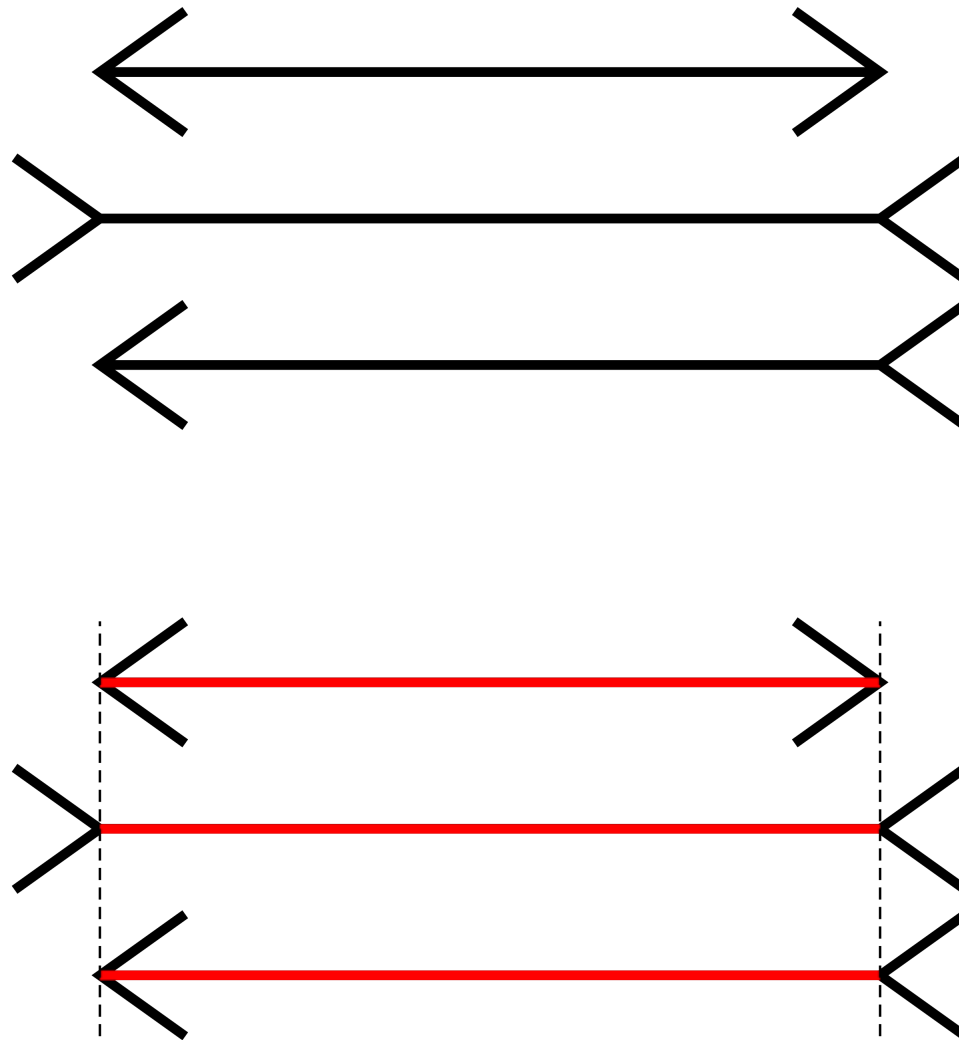
The International
Bestseller

Thinking,
Fast and Slow



Daniel Kahneman
Winner of the Nobel Prize





近い将来の生成AIは「幻覚」を見ない

完璧になったAIが解消できる「生命科学のボトルネック」は何なのか？

生命科学のボトルネックとその解消に必要なもの

- 賢く誠実で献身的な科学者 (教育 🎓)
- 新たな現象の発見と観察・仮説構築 (資金 💰 時間 ⌚ 知識 🧠)
- 新たな計測技術の開発 (資金 💰 時間 ⌚)
- 生体試料のサンプリング (資金 💰 社会的合意 🗣️)
- 再現可能な実験とその結果 (資金 💰 時間 ⌚ 技術 🔬)
- 論文の出版・およびサイエンスというゲームのルール (社会的合意 🗣️)
- 人類社会における需要 (予測・制御が難しい 🤔)

将来の生命科学を左右するもの

- 観測・予測
 - 既にあるものはAIがなんとかしてくれる。今ないもの・これから起きることは？
- インフラ・データ
 - ハードウェア・計算機・ストレージ・ネットワークの維持はますます困難に
 - 100年後、1000年後に残すべきデータは何か、どうやって残すべきか？
- プログラム・AIの自律性
 - 人間が監督すべきか、人間が監督されるべきか
 - 真に強いAI研究者が現れたとき、人間がサイエンスをやる必要があるのか？