データサイエンス特別講義

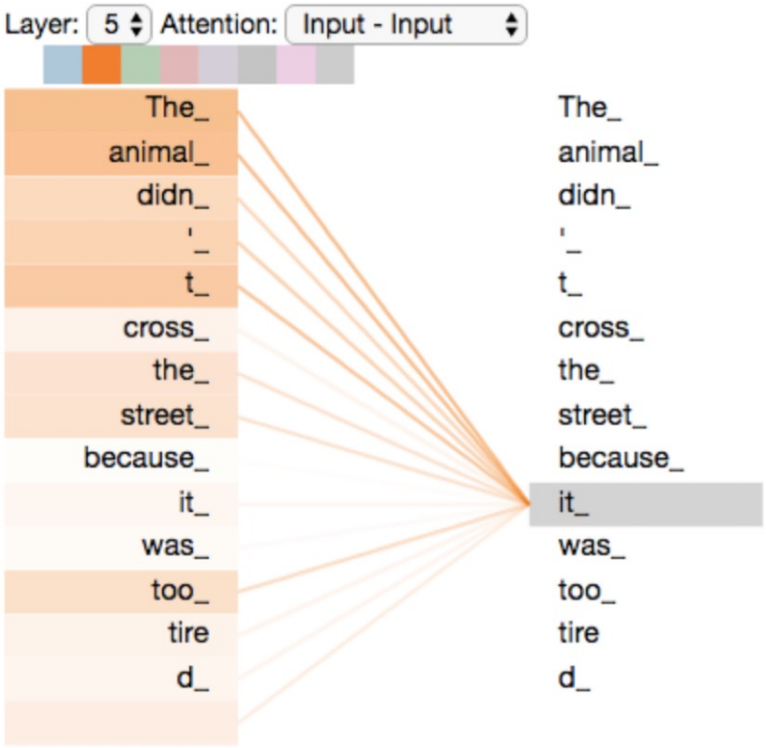
Machine Learning in Cosmology and Fundamental Physics

吉田直紀
東京大学理学系研究科
理化学研究所 革新知能統合研究センター

前回の復習

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.

大規模言語モデルは膨大な文献やwebテキストデータを学習し，
attention 機構を使うことで翻訳や回答の生成を行うことができる．

言語モデルは実は言語の違い（英語，日本語）は気にしていない．
I am a student と単語に分かれていなくとも文字や言葉を埋め込み表現（数値）にして取り扱っているだけ．元気ですか，what’s up？と書かれていても，何の違和感もなく解釈する（他の文字との相関を使う）．

上記の「文献」や「テキスト」を日常会話に置き換えると，言語モデルが音声モデルとなることも可能．実際に研究や実商品が（部分的）に進んでいる．Meta の seamless や google のAudioLM

ということは将来的には…

google の earth species project

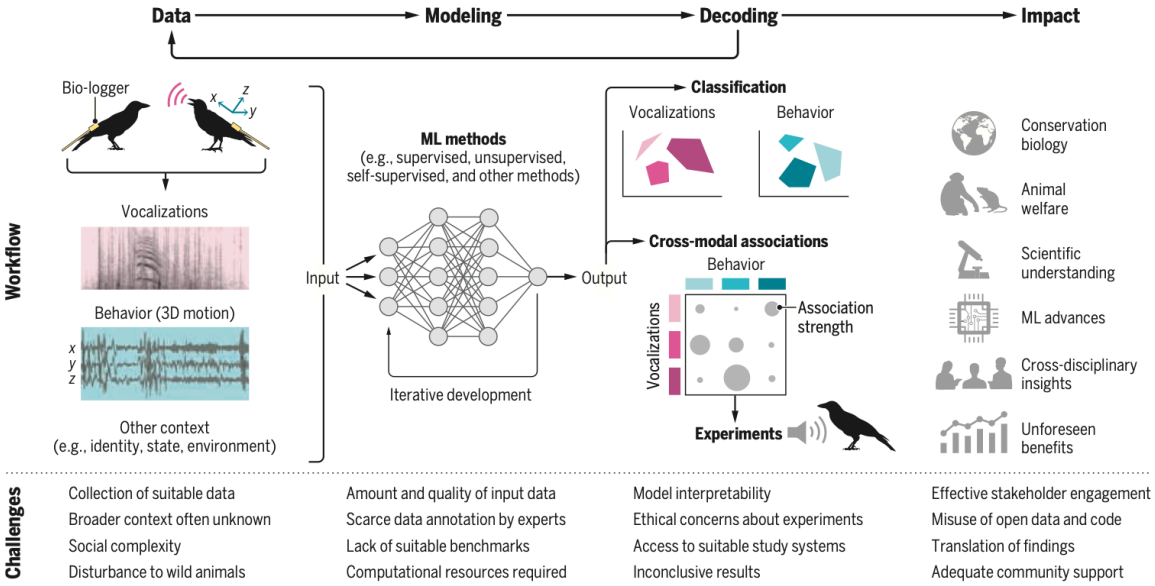
2023年04月26日 06時00分　　サイエンス

Googleが協力する「AIを使った動物とのコミュニケーション」を実現させる試みが進行中

近年はAIを利用した画像生成や高性能なチャットが注目を集めていますが、AIは芸術だけでなくさまざまな科学的研究にとっても役に立つ可能性を秘めています。新たに、GoogleのクラウドサービスであるGoogle Cloudが、機械学習を用いて動物とのコミュニケーションを実現しようとする非営利団体・Earth Species Projectの取り組みについてブログで紹介しました。

「本当に人間以外の動物も言語的なコミュニケーションを取っているのか？」という疑問もあるかもしれませんが、鳥の歌は世代を超えて伝承されており新たに生まれた歌が3000kmもの距離を伝わって流行することなども知られているほか、シジュウカラは単語や文法を持つ言語を操るという研究結果も報告されています。また、チンパンジーの膨大な鳴き声を解析した研究では、12種類の異なる鳴き声を組み合わせて390通りの構文を作っていることも示されました。

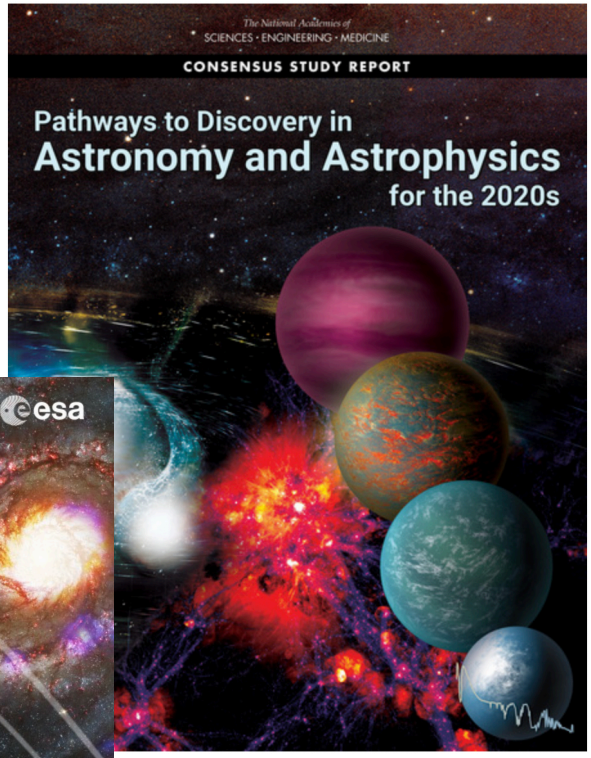
チンパンジーが390もの構文を使って会話をしていることが鳴き声5000回の録音から示唆される -

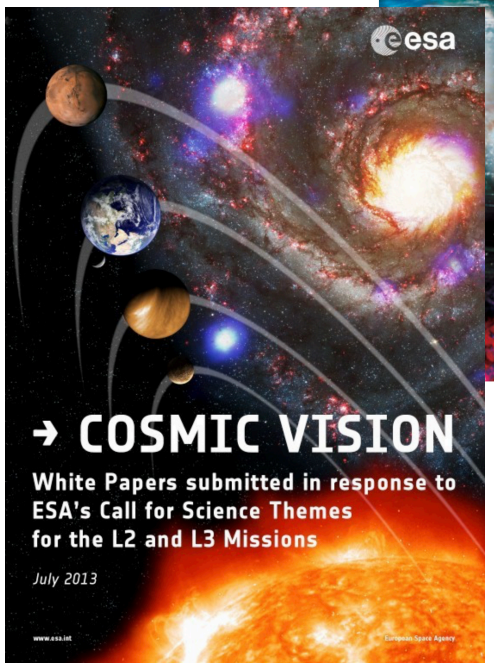


Using machine learning to decode animal communication

C. Rutz et al., Science, 381（2023）152

今後の宇宙研究と「発見」





全天級の宇宙論サーベイ観測

可視光、赤外線、電波

- ダークマター、ダークエネルギー
- 原始重力波の痕跡

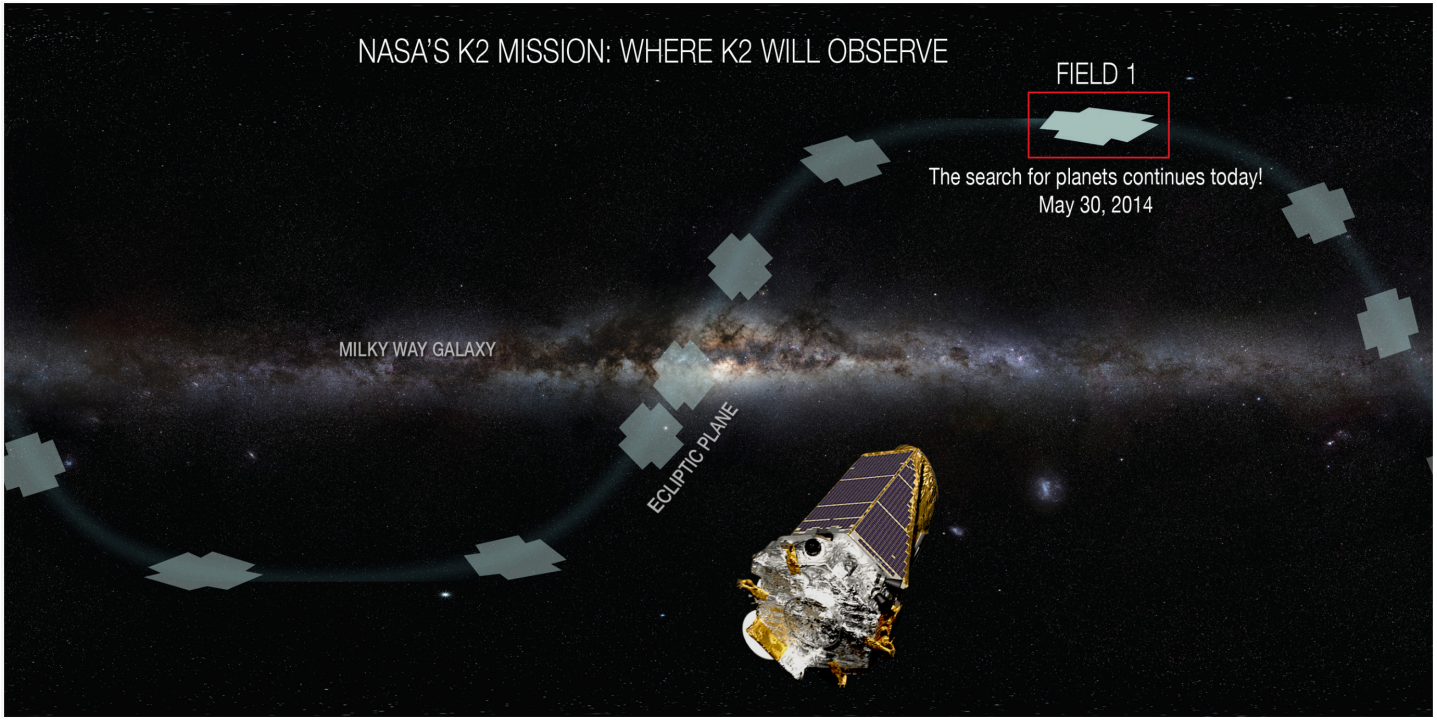
マルチメッセンジャー宇宙物理学

- 重力波、ニュートリノ、高エネルギー粒子
- 時間領域天文学
- ブラックホールの観測

系外惑星

- 第二の地球、太陽系は特殊か普遍か
- 宇宙生物学

NASAのケプラー衛星



初のハビタブルゾーン惑星
Kepler-22bの発見

小型惑星が一般的であることを証明

初の地球サイズ惑星の発見

連星系の惑星Kepler-16bの発見
映画「スター・ウォーズ」の惑星タトゥイーンのような二重の夕日が見える世界

2009年に打ち上げられ，2018年まで観測を続けた.
はくちょう座とこと座の領域にある15万個以上の恒星を継続的にモニタリング
2,662個の系外惑星を発見!

研究成果の詳細　AIと機械学習の活用

トランジット法による惑星検出

- CNNがBLS法より高精度で検出
- ノイズ入りデータから地球型惑星を検出
- Keplerデータで96%の真惑星を識別

GoogleとNASAの発見

- 2017年にKepler-90i惑星を発見
- 太陽系と並ぶ8惑星系を確認
- 15,000シグナルで訓練し96%精度達成

TESS衛星データの解析

- Transformerモデルを活用
- 214個の新惑星系候補を発見
- 周期性を仮定せず検出可能

大気組成の特性評価

- ランダムフォレストで大気分析
- 従来法より大幅に時間短縮
- 分子存在量と雲の不透明度を推定

MNRAS **000**, 1–19 (0000) Preprint 10 March 2025 Compiled using MNRAS LATEX style file v3.0

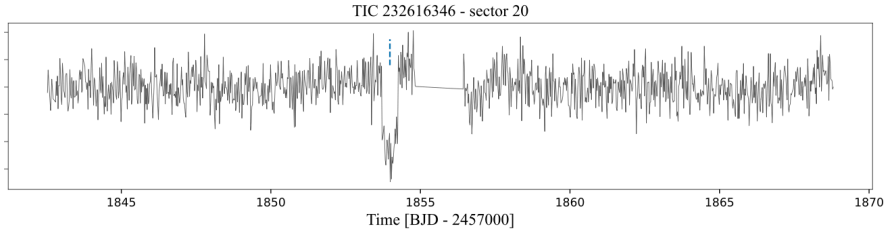
Exoplanet Transit Candidate Identification in TESS Full-Frame Images via a Transformer-Based Algorithm

Helem Salinas[1,2]★, Rafael Brahm[4,5,6], Greg Olmschenk[3,7]†, Richard K. Barry[3]‡, Karim Pichara[1,5], Stela Ishitani Silva[3,8]§, Vladimir Araujo[9]

ABSTRACT

The Transiting Exoplanet Survey Satellite (TESS) is surveying a large fraction of the sky, generating a vast database of photometric time series data that requires thorough analysis to identify exoplanetary transit signals. Automated learning approaches have been successfully applied to identify transit signals. However, most existing methods focus on the classification and validation of candidates, while few efforts have explored new techniques for the search of candidates. To search for new exoplanet transit candidates, we propose an approach to identify exoplanet transit signals without the need for phase folding or assuming periodicity in the transit signals, such as those observed in multi-transit light curves. To achieve this, we implement a new neural network inspired by Transformers to directly process Full Frame Image (FFI) light curves to detect exoplanet transits. Transformers, originally developed for natural language processing, have recently demonstrated significant success in capturing long-range dependencies compared to previous approaches focused on sequential data. This ability allows us to employ multi-head self-attention to identify exoplanet transit signals directly from the complete light curves, combined with background and centroid time series, without requiring prior transit parameters. The network is trained to learn characteristics of the transit signal, like the dip shape, which helps distinguish planetary transits from other variability sources. Our model successfully identified 214 new planetary system candidates, including 122 multi-transit light curves, 88 single-transit and 4 multi-planet systems from TESS sectors 1-26 with a radius > 0.27 $R_{\rm Jupiter}$, demonstrating its ability to detect transits regardless of their periodicity.

恒星TIC232..の明るさ変動



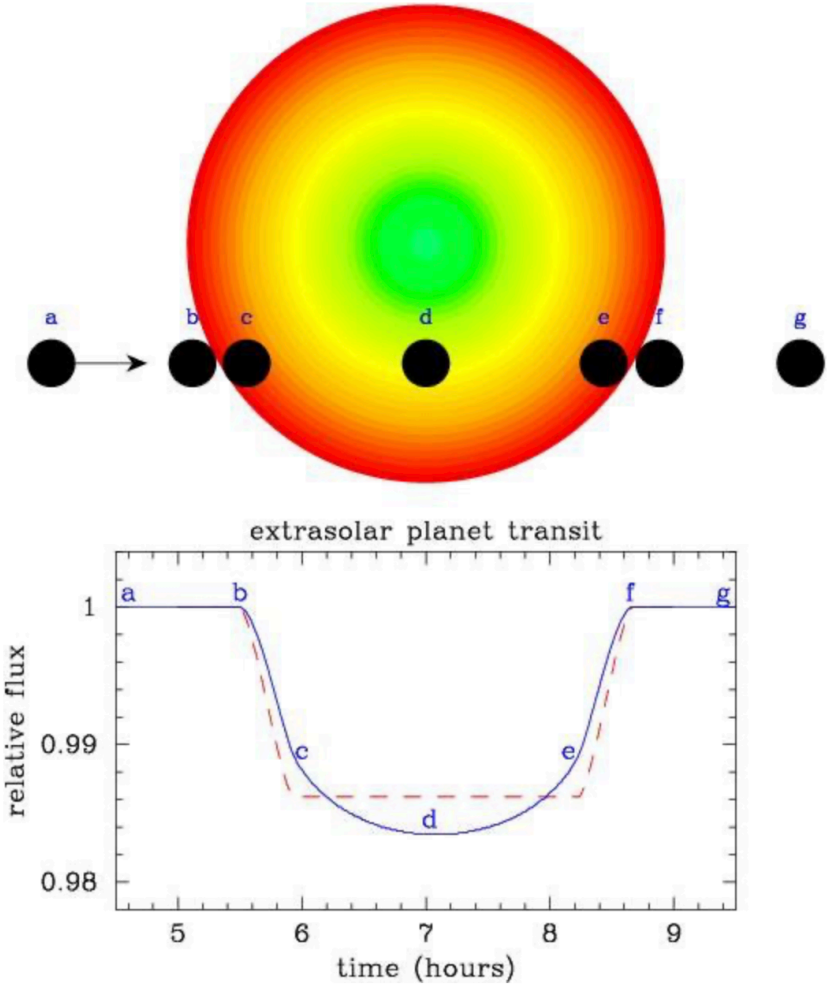
各光度曲線は1000データポイントの統一された長さに処理され、値は正規化されて-1から1の範囲に線形変換

↓

Transformer をこれを「文章」として扱い，self-attention により重要部分を検出

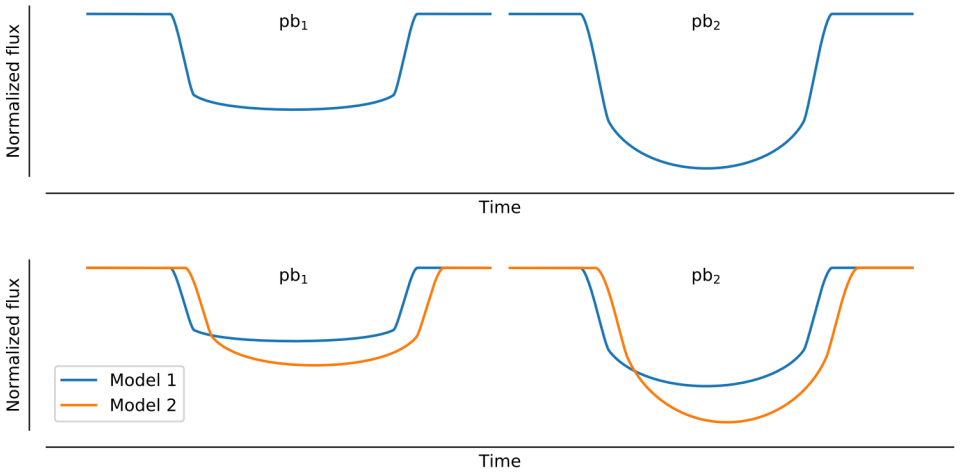
TESS衛星データから224個の新しい惑星系候補を発見！

データ分析のための理論モデル



観測された明るさの変動（光度曲線）から惑星やその軌道の特徴を特定するには，大量の理論モデル，理論テンプレートが必要

　稀に，複数の惑星が食を起こすこともある．



光度曲線は”比較的”簡単な物理モデル(左）で再現できるが，観測によっては大規模な非線形シミュレーションが必要なものもある

宇宙論・大規模構造形成のエミュレーション

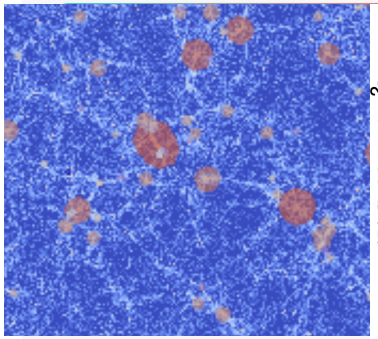
▶ 課題と解決策

従来の問題：
N体シミュレーションは計算コストが非常に高く、多数のパラメータ探索が困難

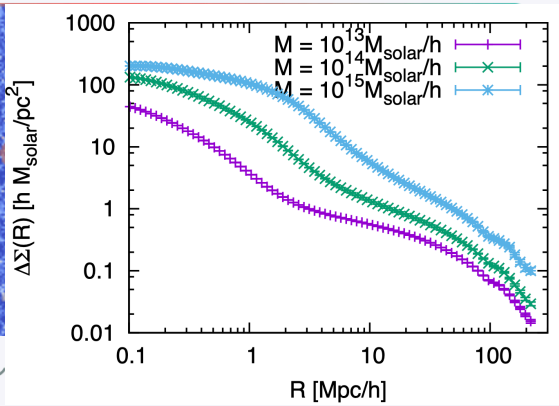
エミュレーション手法：
- 畳み込みニューラルネットワーク (CNN)
- 機械学習による密度場進化の予測
- ガウス過程回帰との組み合わせ

主な応用：
- 宇宙論パラメータの推定
- 観測データとの比較研究
- Dark Quest プロジェクト

密度場の進化シミュレーション



N体シミュレ



3桁の高速化達成
フルN体シミュレーション → 約1000倍高速

🌍 気候モデルの統計的ダウンスケーリング

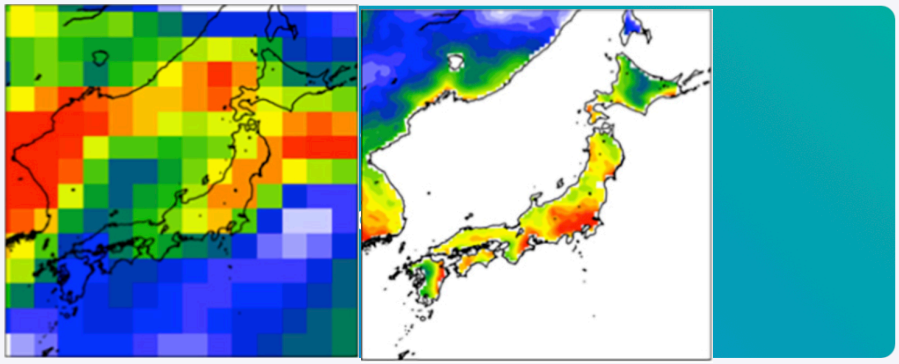
▶ 技術概要

ハイブリッドダウンスケーリング：
- 力学的手法 + 統計的手法
- 少数の高解像度シミュレーション
- 統計モデルによる関係性学習

RCMエミュレーター：
- ニューラルネットワークによる関数学習
- 大規模予測子と局所変数の関係
- 12km解像度での日単位予測

応用分野：
- 地域気候変動予測
- 影響評価研究
- 政策分析支援

地域気候ダウンスケーリング



GCM → 統計的ダウンスケーリング → 高解像度予測

瞬時の気候予測が可能
統計モデル学習後は即座に結果出力

GCM → NN → RCM

🏗️ 地震動予測のエミュレーション

▶ 従来手法の限界

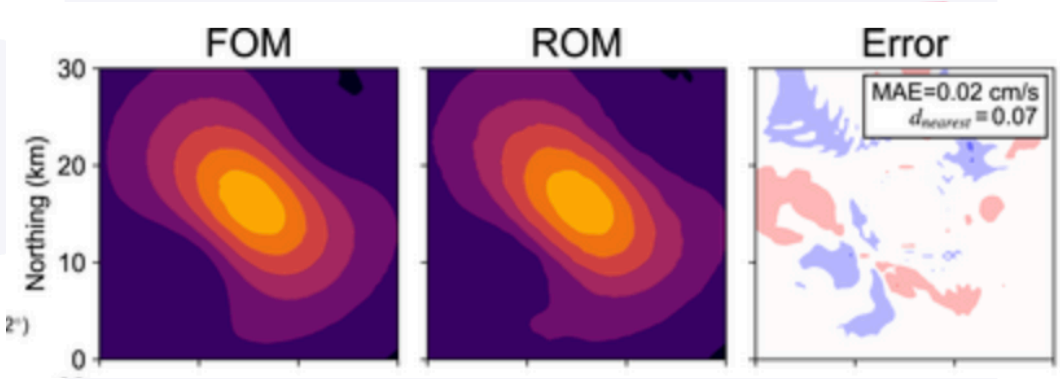
物理ベースシミュレーション：
- 3次元波動伝播計算
- 高い計算コストと時間
- リアルタイム適用困難

▶ エミュレーション技術

縮約次数モデリング (ROM)：
- 適切直交分解 (POD)
- 補間に基づく高速予測
- 震源パラメータ依存性の学習

実用化例：
- SCEC Broadband Platform
- 地震早期警報システム
- 確率論的地震ハザード評価

地震動強度分布予測



物理ベース波動伝播 → 縮約次数モデル → 瞬時予測

ミリ秒単位での
物理ベース地震動マップ生成

🌊 海洋循環モデルのエミュレーション

▶ 対象システム

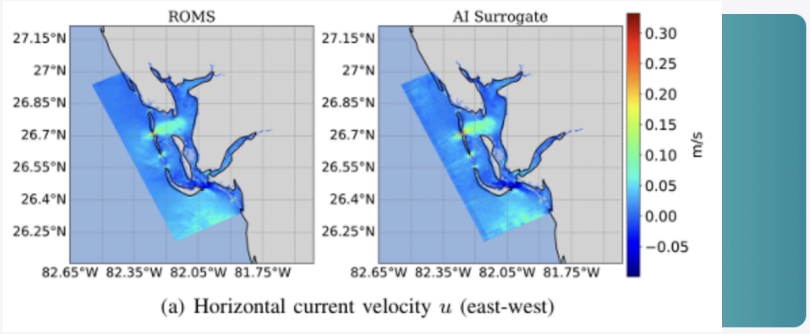
ROMS (Regional Ocean Modeling System)：
- 沿岸海洋循環モデル
- 非一様格子による高精度計算
- 早期警報システムでの活用

▶ AI技術の適用

4D Swin Transformer：
- 潮汐波伝播シミュレーション
- 時空間データの効率的処理
- 物理制約による結果検証

社会的意義：
- 9億人の沿岸住民保護
- ハリケーン・高潮災害対策
- リアルタイム災害対応

沿岸海洋循環予測



ROMS → AI Surrogate → 高速循環予測

450倍以上の高速化
512 CPUコア: 9,908秒
↓
1 A100 GPU: 22秒

💨 流体力学・CFDのサロゲートモデリング

▶ 対象と課題

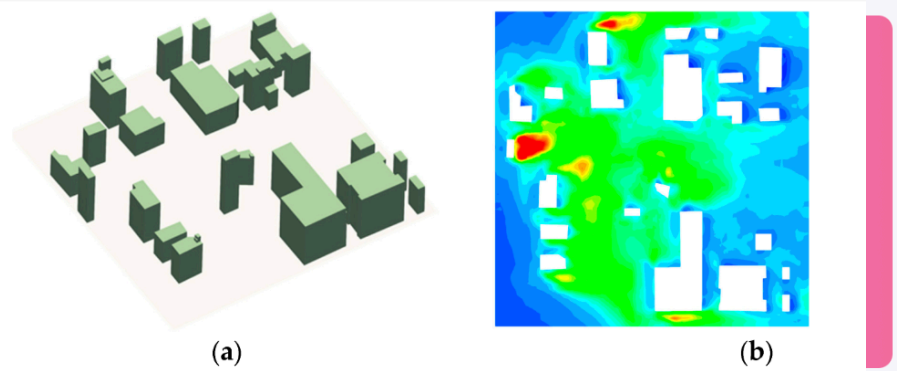
計算流体力学 (CFD)：
- 複雑な偏微分方程式系
- 大規模3次元格子計算
- 長時間計算の制約

▶ 機械学習アプローチ

CNN ベースモデル：
- 畳み込みニューラルネットワーク
- マルチグリッド構造の活用
- U-Net-MG アーキテクチャ

建築環境への応用：
- 室内気流分布予測
- 換気システム設計
- リアルタイム環境制御

室内気流分布予測

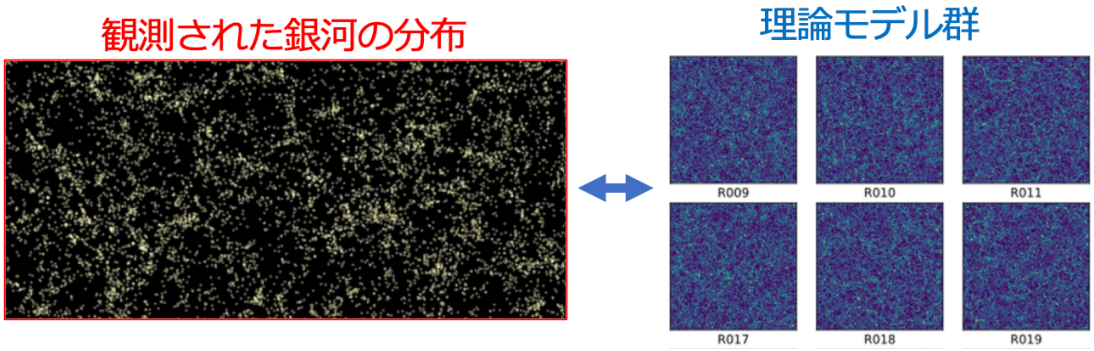


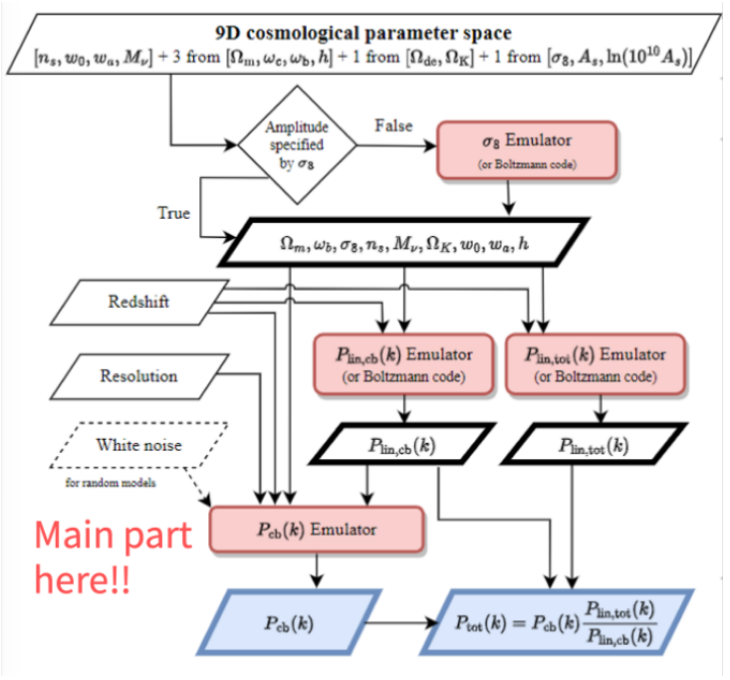
CFD → CNN学習 → 高速流体予測

計算時間を0.05%に削減
従来CFD比で大幅な高速化達成

宇宙の構造のエミュレーション

1つの理論モデルの非線形シミュレーションに富岳で3日かかる。
サマリ統計量を計算するだけなら我々の「エミュレータ」で0.1秒で出力可能.
6-10次元のパラメータ空間のどの点でも精度保証





古くはクリギングや諸分野のガウス過程など.
最近の計算能力とアルゴリズムの向上により
かなり複雑なことができ、統計解析の強力なツール
2019年に発表した ”Dark Emulator” は物質分布に特化した
ものだが25,000ダウンロード達成. 統計データ解析のプレイグラウンド

エミュレータによるベイズ推定の流れ

従来の解析パイプライン

観測量

測定

要約統計量

順モデル:
各点でスパコン計算

MCMCサンプラー

多次元パラメタ空間から約
100万パラメタセットを探査

新たな解析パイプライン

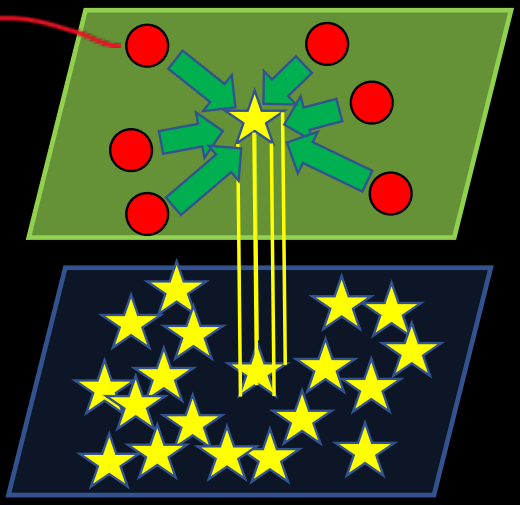
観測量

測定

要約統計量

高精度シミュレーション
(100パラメータセットを
データベース化)

エミュレータ
（ベイズ推定）



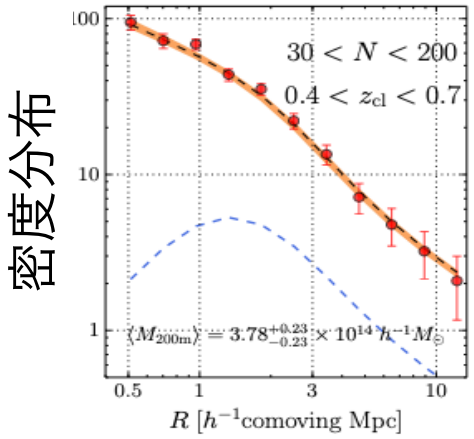
計算コストの高い直接数値シミュレーションの結果を
近似やフィット無しにベイズ統計解析に利用可能

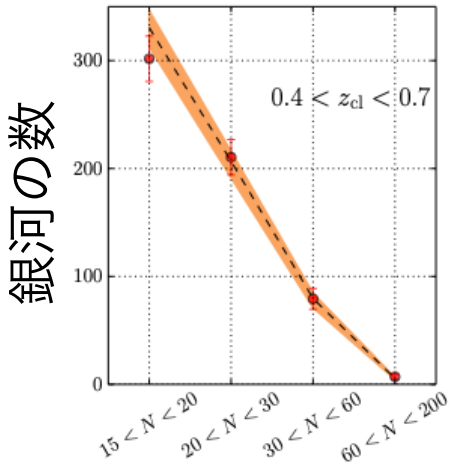
シミュレーションデータベースからエミュレータ構築

ハイブリッド統計モデリング

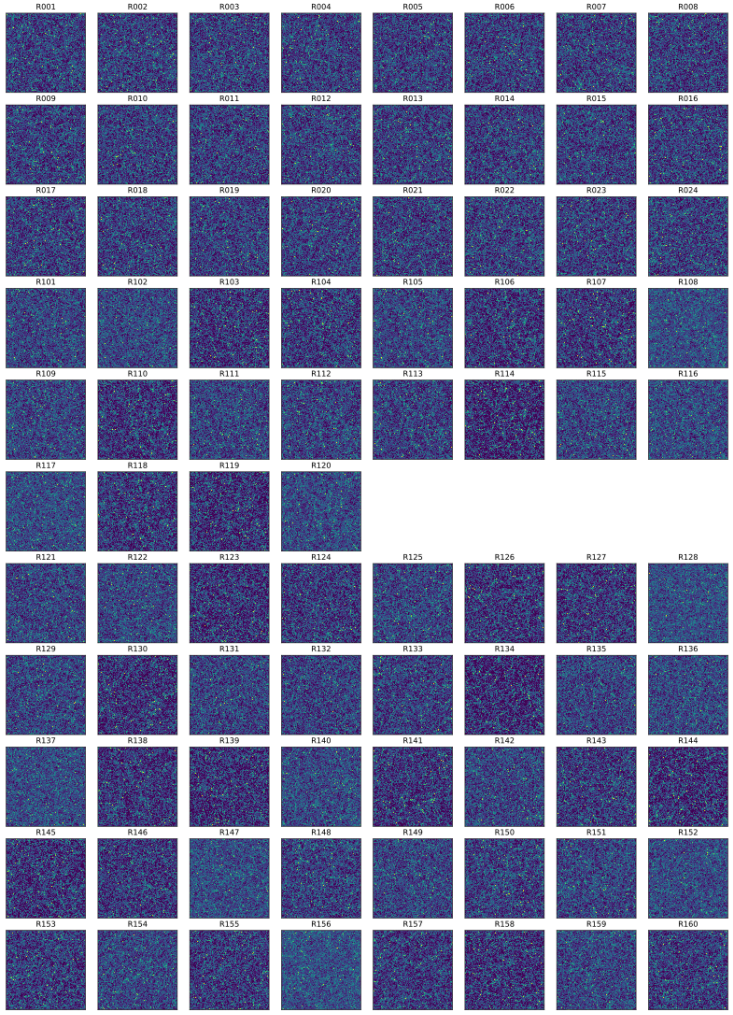
- ガウス過程回帰 (nonparametric Bayes)
- 重みつき主成分分析
- パラメトリック銀河・ハローモデル

スパコンで2日を
ノートPCで1秒に！





シミュレーションデータベース

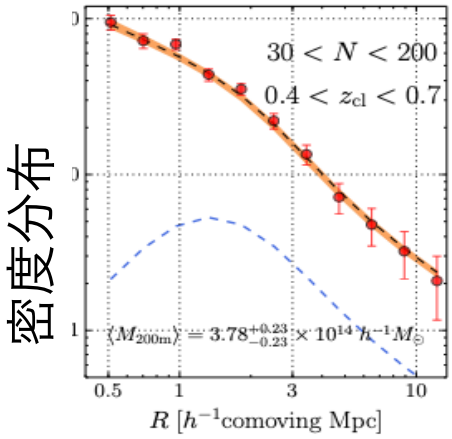


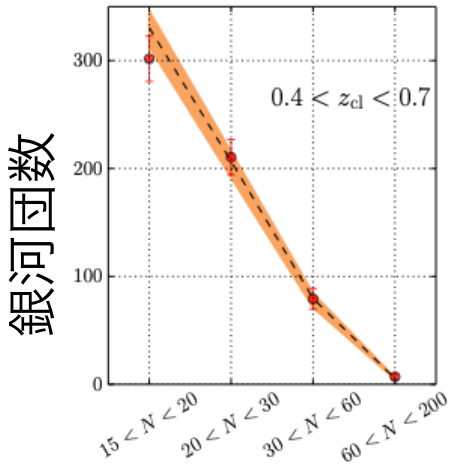
シミュレーションデータベースからエミュレータ構築

ハイブリッド統計モデリング

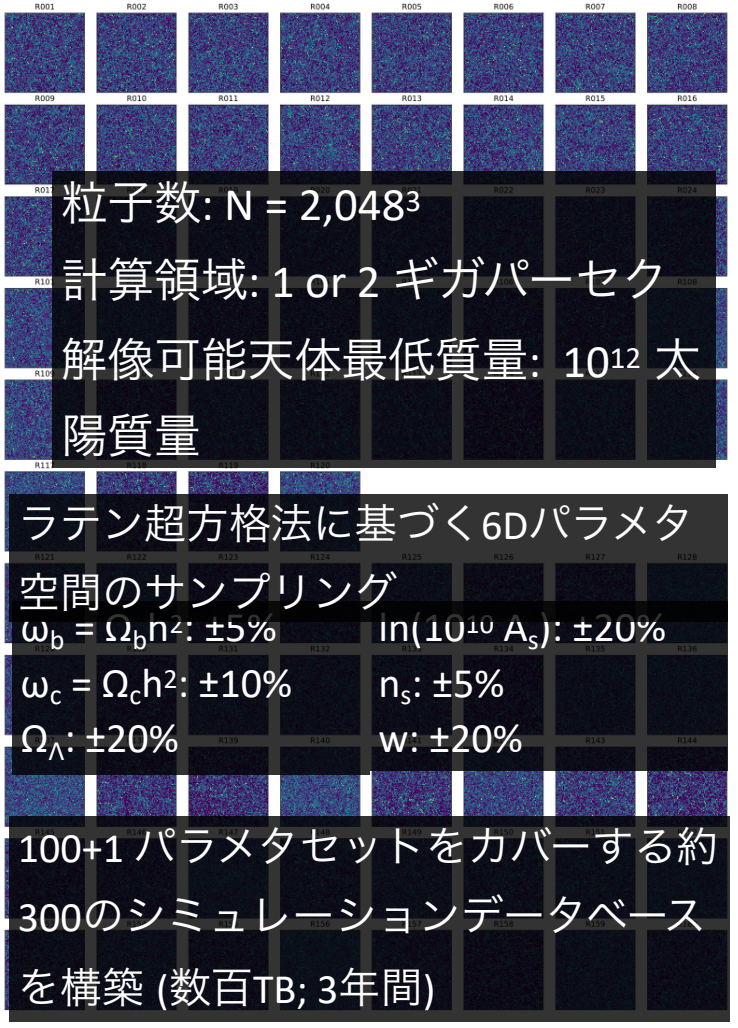
- ガウス過程回帰 (nonparametric Bayes)
- 重みつき主成分分析
- パラメトリック銀河・ハロー結合模型

スパコンで2日をラップトップで1秒に短縮！





シミュレーションデータベース



粒子数: $N = 2{,}048^3$

計算領域: 1 or 2 ギガパーセク

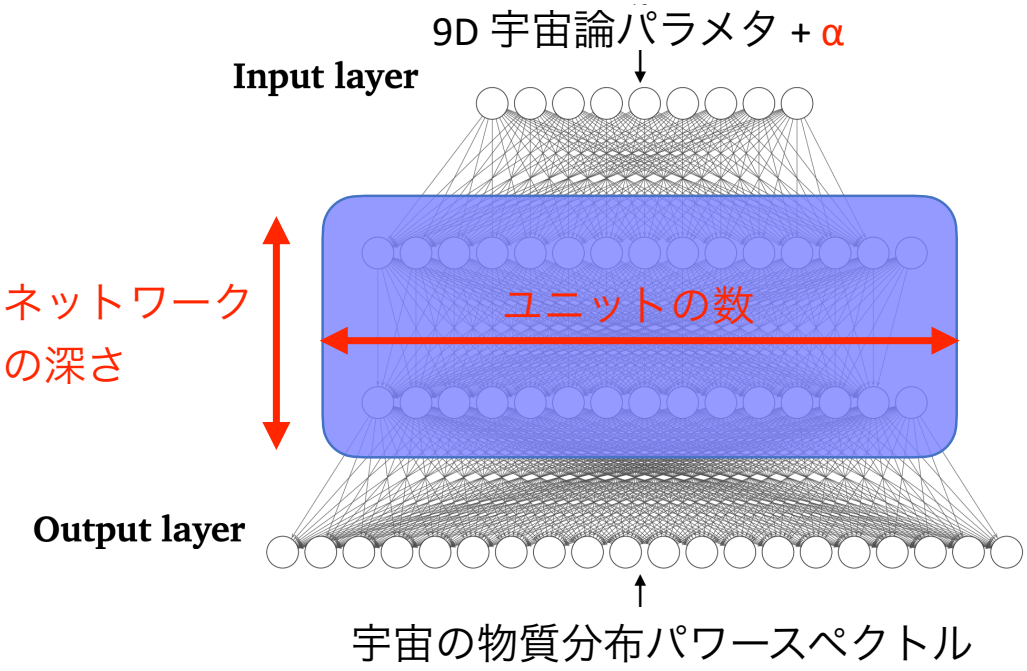
解像可能天体最低質量: 10^{12} 太陽質量

ラテン超方格法に基づく6Dパラメタ空間のサンプリング

$\omega_b = \Omega_b h^2$: ±5%	$\ln(10^{10} A_s)$: ±20%
$\omega_c = \Omega_c h^2$: ±10%	n_s: ±5%
Ω_Λ: ±20%	w: ±20%

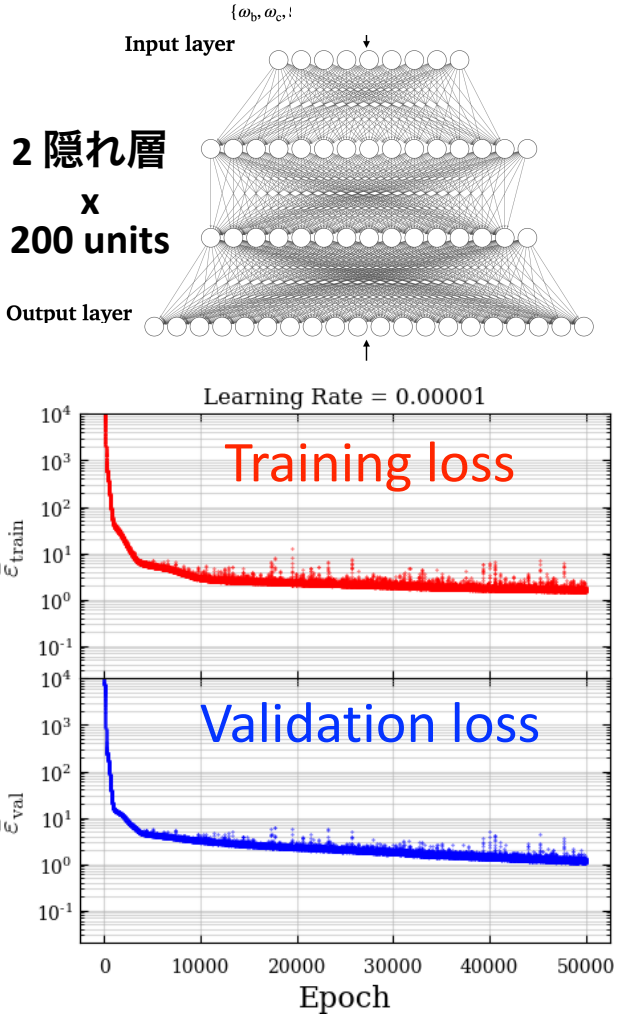
100+1 パラメタセットをカバーする約300のシミュレーションデータベースを構築 (数百TB; 3年間)

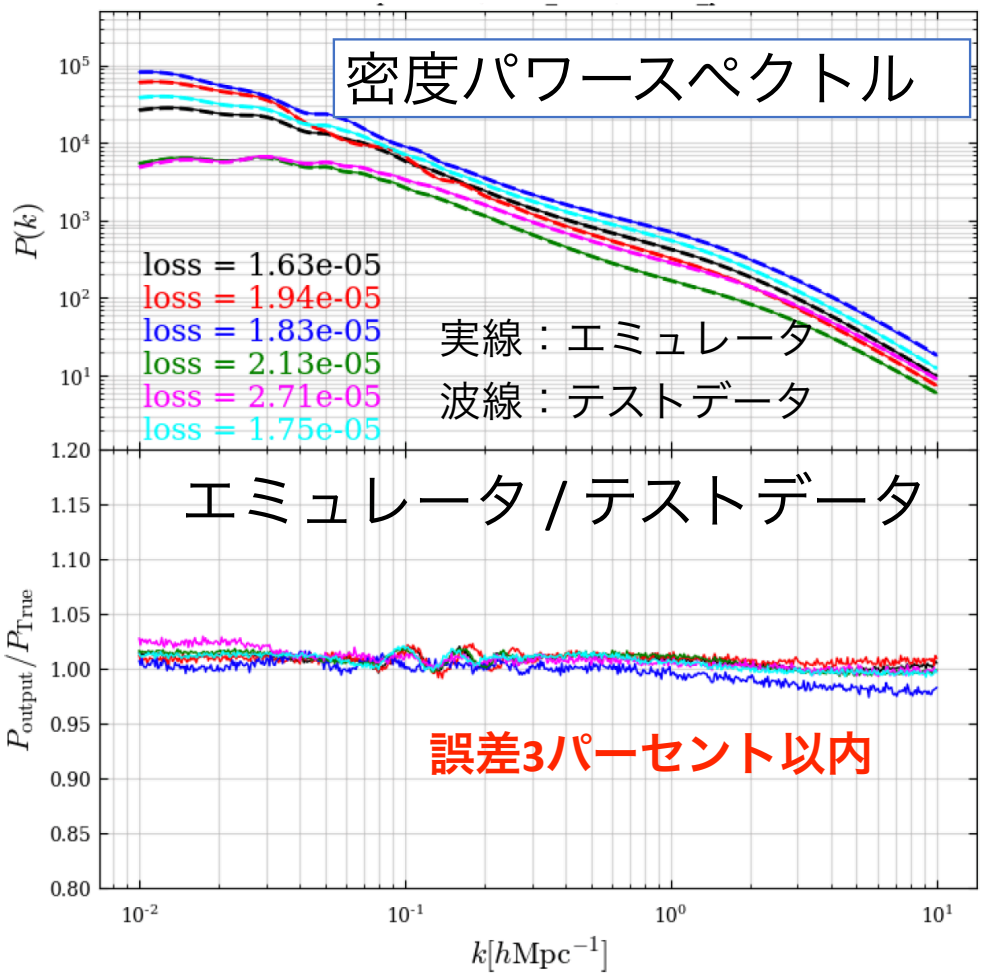
エミュレータのデザイン



- 活性化関数
- オプティマイザー
- ミニバッチサイズ
- エポック数
- …

学習の結果と統計量の精度





ガウス過程

関数の回帰、推定

ベイズ統計に基づき、関数形の仮定なし

非線形回帰

$$f(x) \sim P[\mu(x), k(x, x')]$$

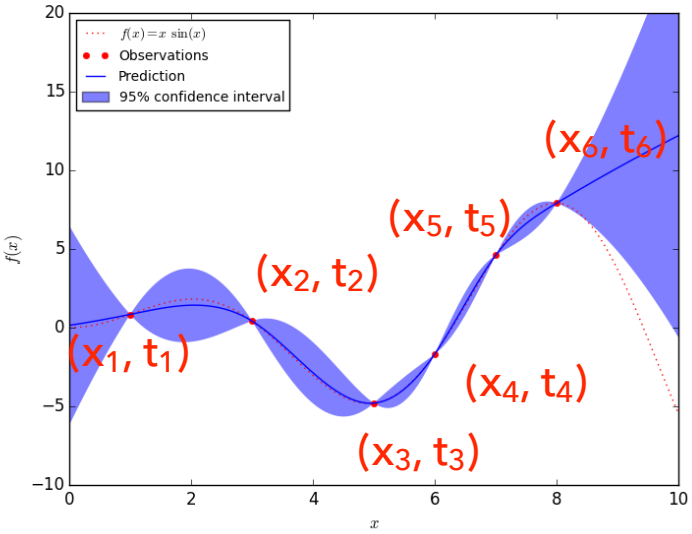
平均 μ

共分散関数 k (カーネル関数)

共分散関数として、パラメータθを含む形を仮定

$$C(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \theta_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{I}\frac{(x_i - x_i')^2}{r_i^2}\right] + \theta_2.$$

"length scale" r_i for each x_i



✓トレーニングデータ $\{x_i\}$ から

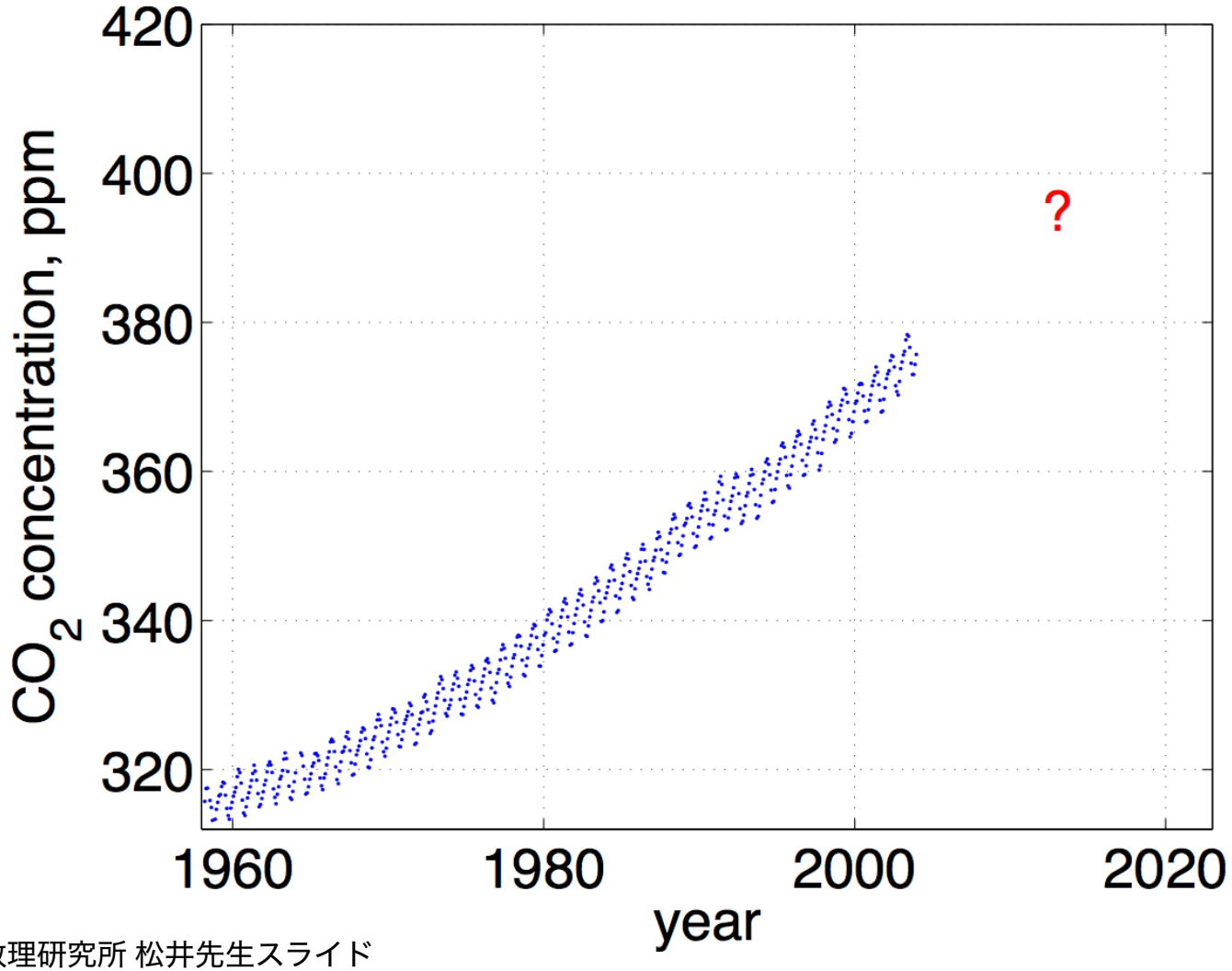
ハイパーパラメータ θを決定

✓データ $\{x_i\}$と上記のθから x_{n+1}

$$P(t_{N+1} \,|\, \mathbf{t}_N) \propto \exp\left[-\frac{1}{2}\left[\mathbf{t}_N\ t_{N+1}\right]\mathbf{C}_{N+1}^{-1}\begin{bmatrix}\mathbf{t}_N \\ t_{N+1}\end{bmatrix}\right].$$

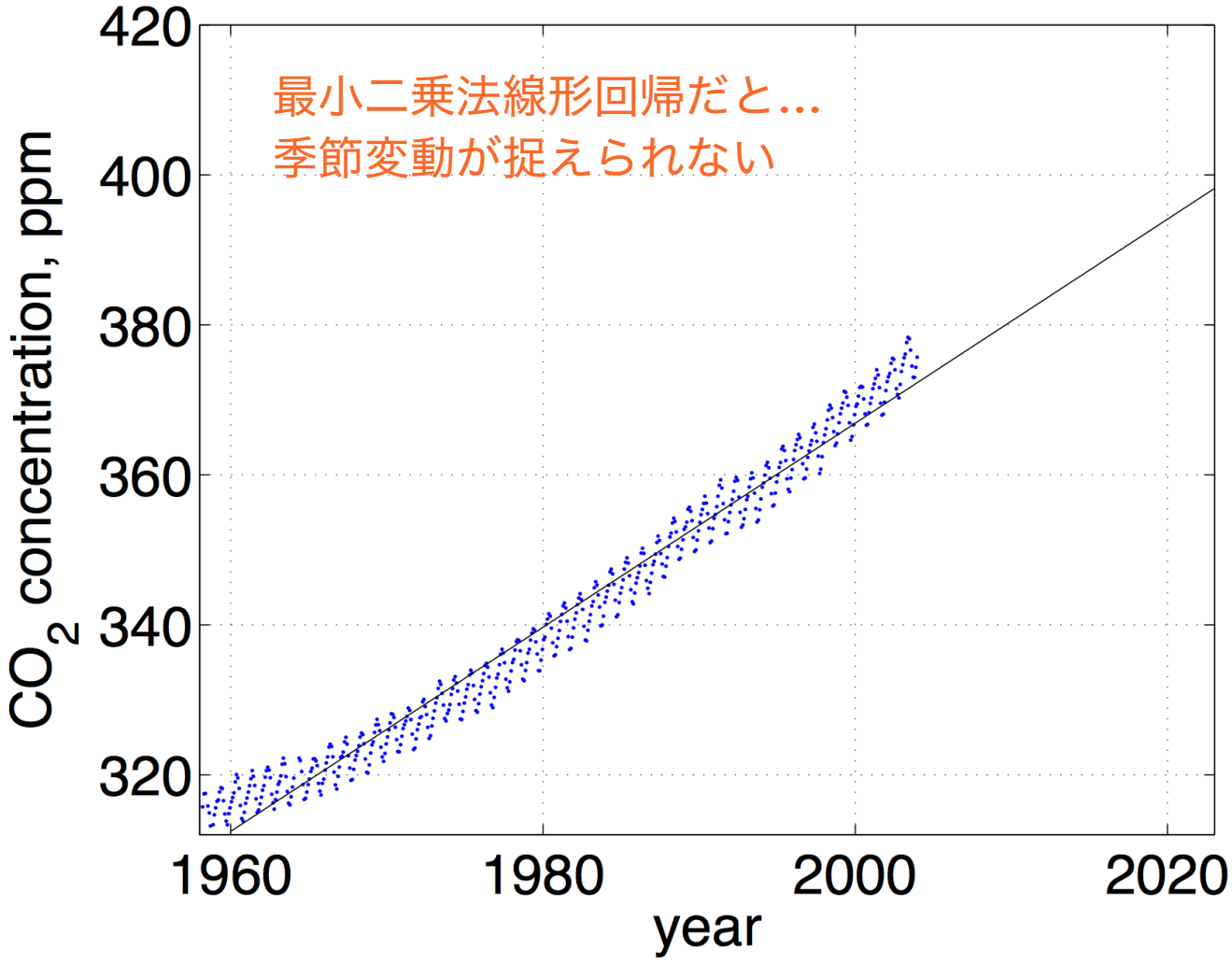
$$\begin{aligned}\hat{t}_{N+1} &= \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{t}_N \\ \sigma_{\hat{t}_{N+1}}^2 &= \kappa - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}.\end{aligned}$$

なんでガウス過程?

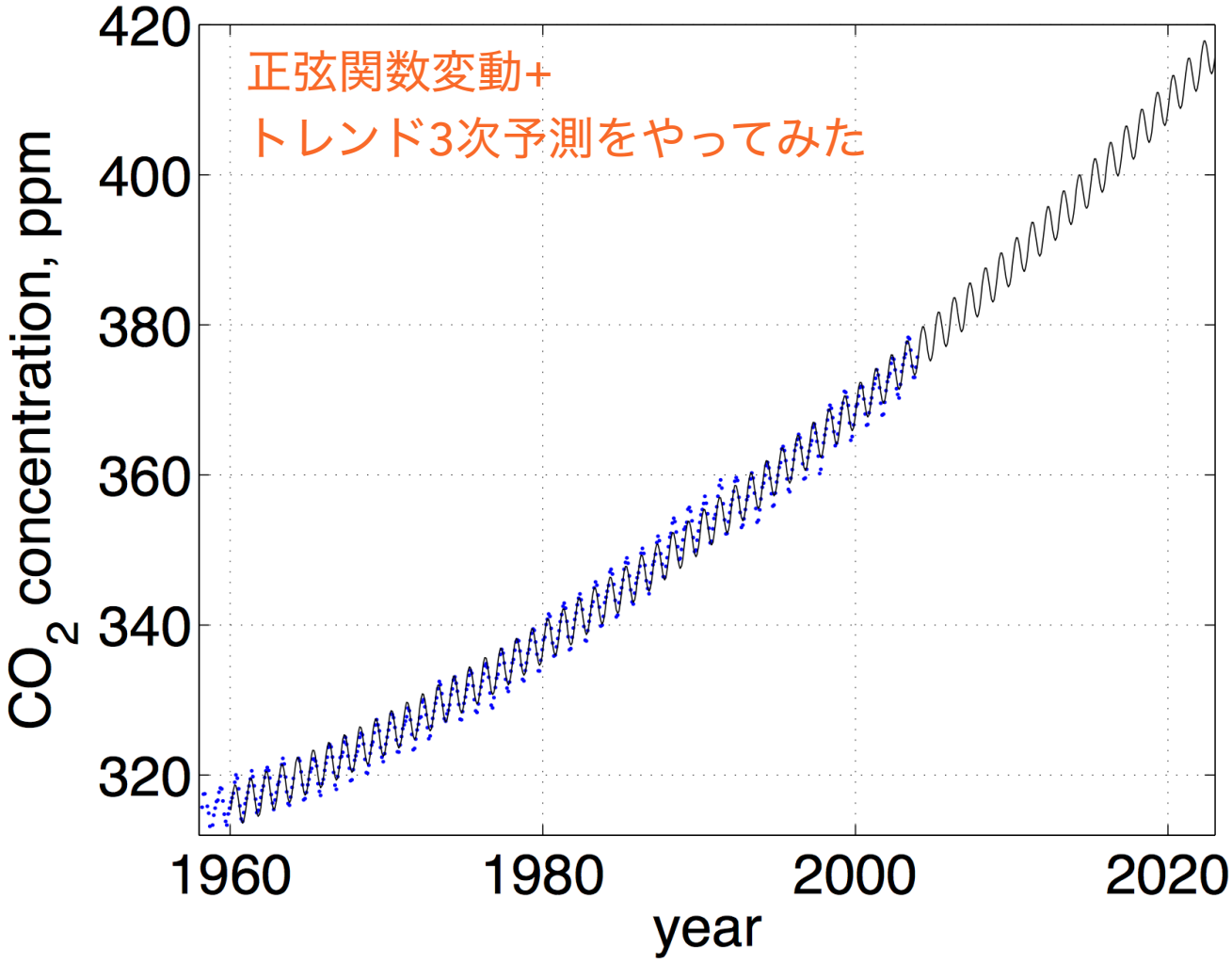


統計数理研究所 松井先生スライド

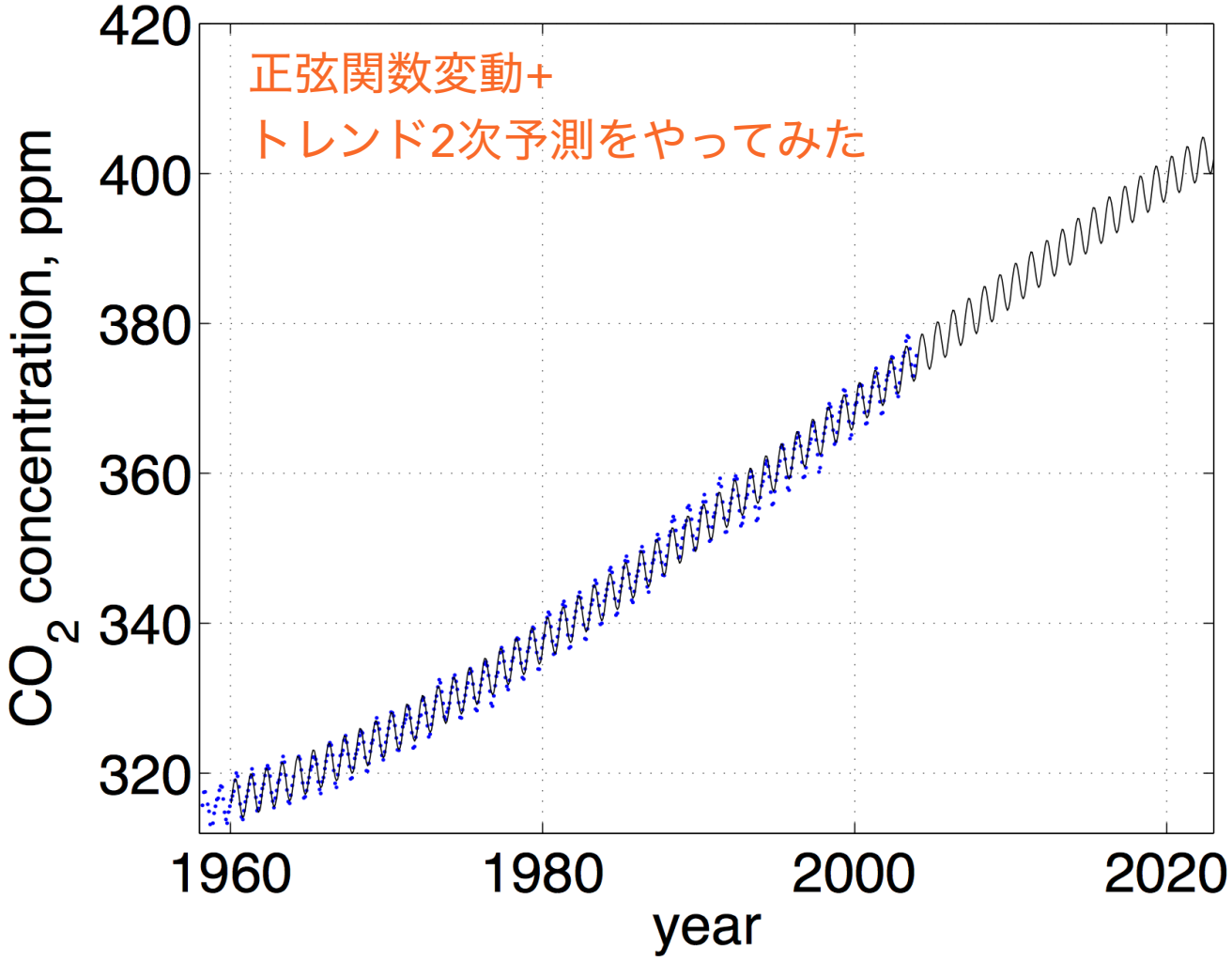
複雑な変動を予測したい



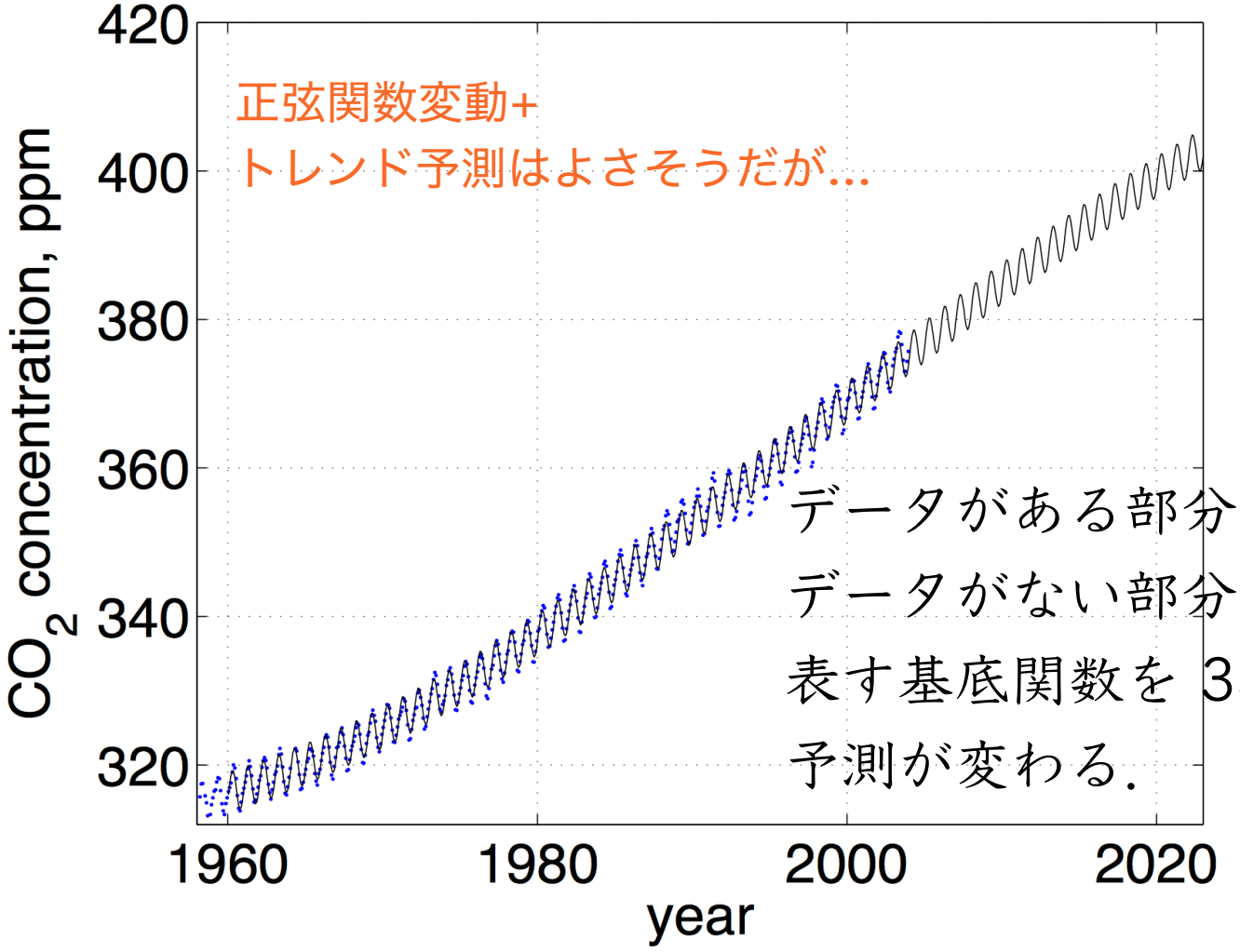
複雑な変動を予測したい



複雑な変動を予測したい



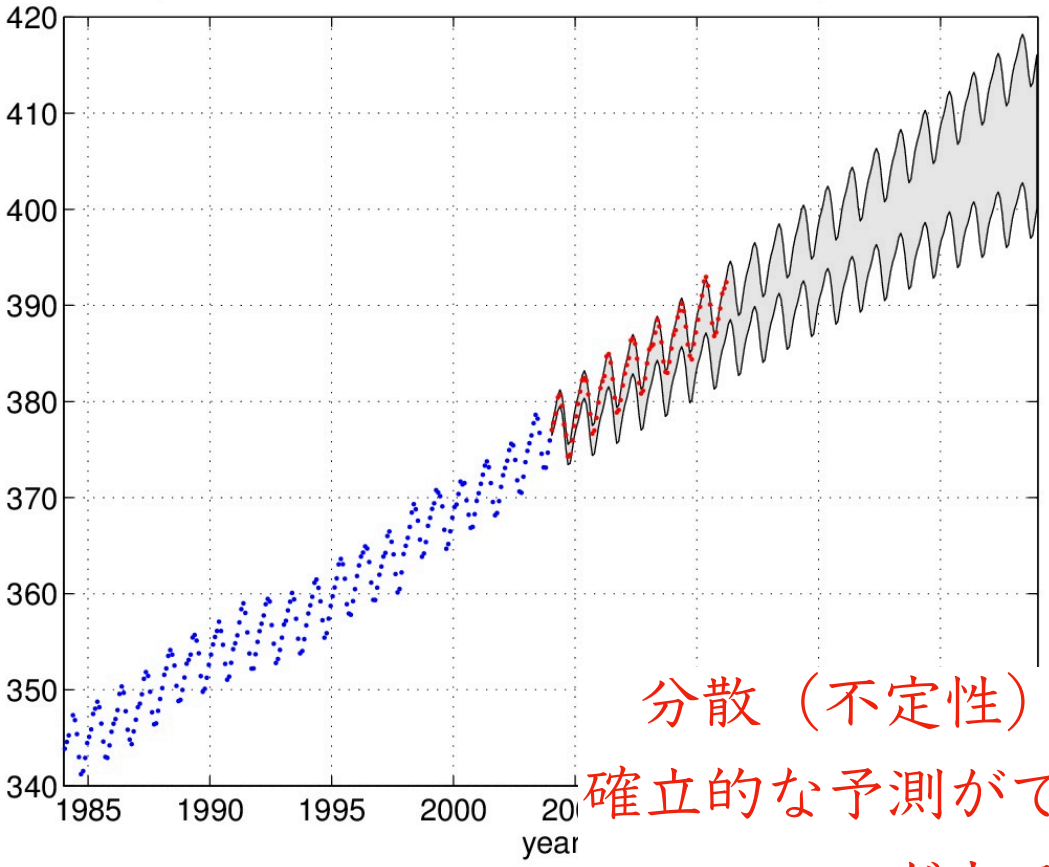
なにがいけないの?



データがある部分についてよく適合している。
データがない部分については、トレンド成分を
表す基底関数を 3次から2次多項式に変更すると
予測が変わる.

仮定する関数に依存する.

ガウス過程を使ってみた

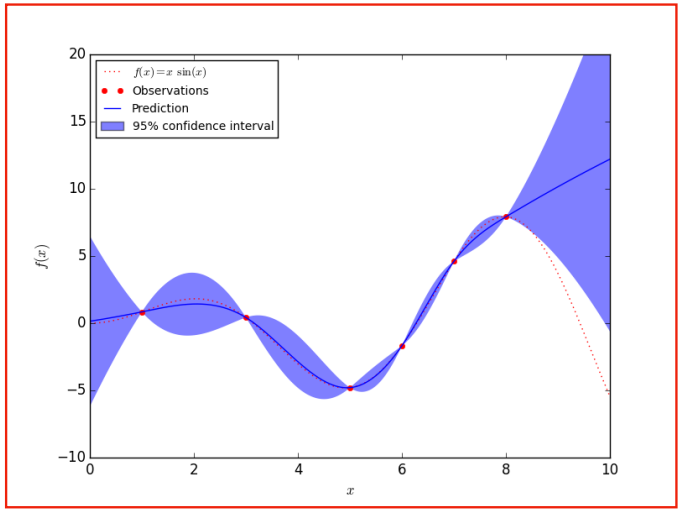


ノンパラメトリック
（関数形を仮定しない）

n番目の入力 x_n と y_n

1- n 番目までのデータを学習して予測

分散（不定性）もあわせて
確立的な予測ができることが
ガウス過程の強み



データの相関，確率的と聞くと…　最近はこんなのもある

Transformer Gaussian Process

観測データ {data $(x_1,y_1),\ldots,(x_n,y_n)$}

↓

Self-Attention Encoder（データ点間の関係を学習）

↓

Cross-Attention Decoder（予測点との関係）

↓

予測分布 $p(y^* \mid x^*, \text{data})$

Transformer Neural Processes:
Uncertainty-Aware Meta Learning Via Sequence Modeling

Tung Nguyen [1] Aditya Grover [1]

Abstract

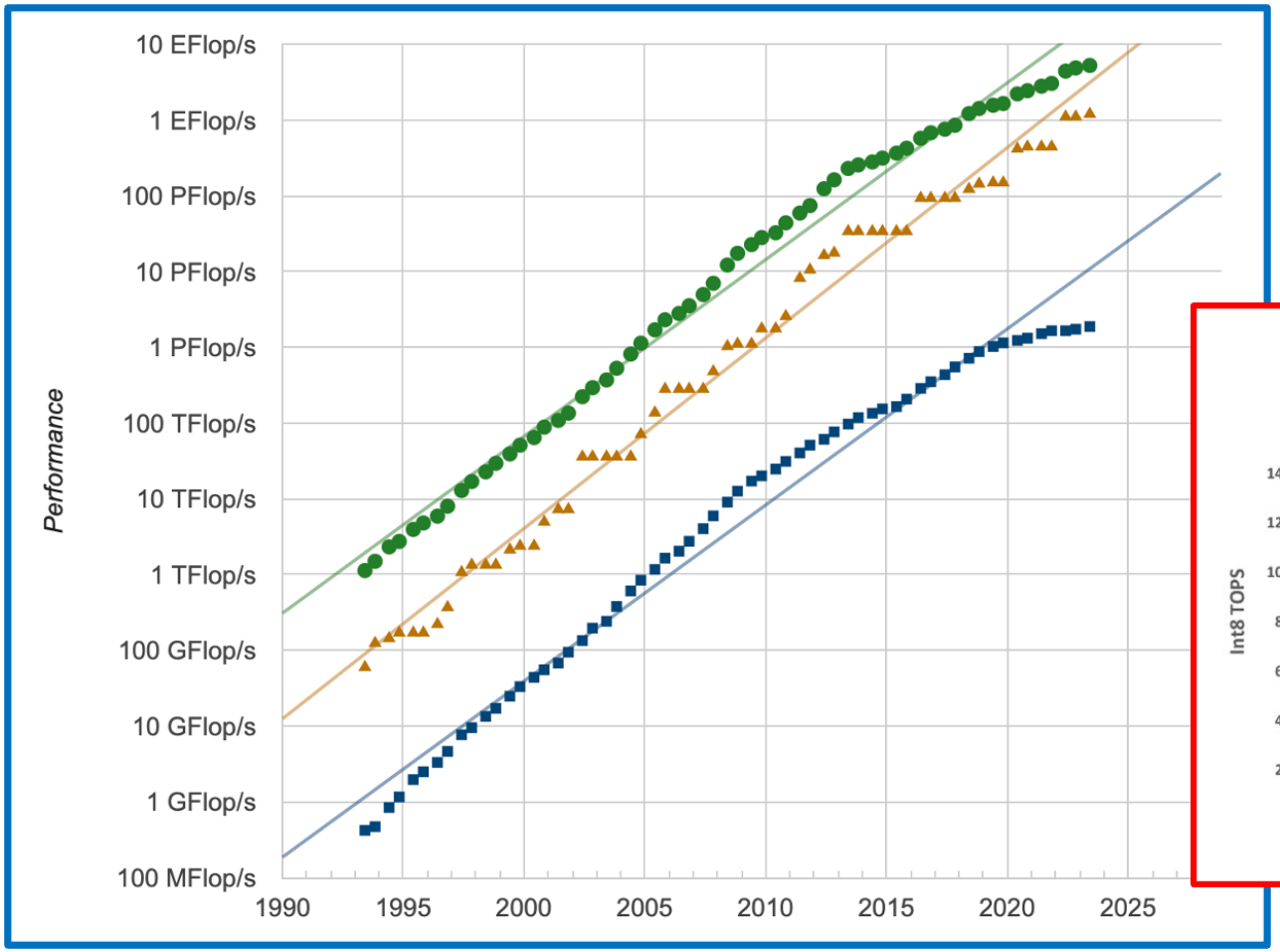
Neural Processes (NPs) are a popular class of approaches for meta-learning. Similar to Gaussian Processes (GPs), NPs define distributions over functions and can estimate uncertainty in their

Shahriari et al., 2015; Hakhamaneshi et al., 2021) and multi-armed bandits (Cesa-Bianchi & Lugosi, 2006; Riquelme et al., 2018), where the quantified uncertainty can guide data acquisition. We call this paradigm uncertainty-aware meta learning, which is the main focus of our paper.

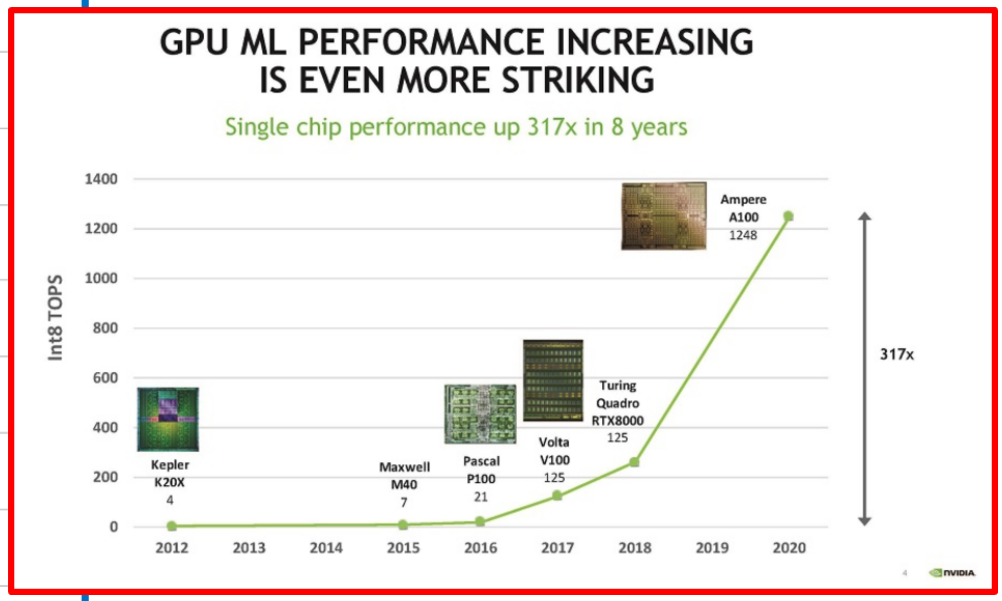
テキスト

前回の復習

スパコンの進化 1990-2025



GPUとAI

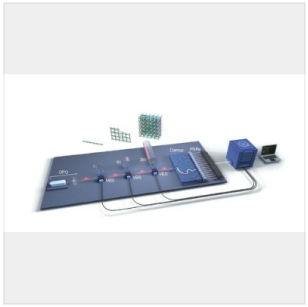


量子コンピュータ

「量子超越性」を持つ光量子コンピュータ、AWSで利用可能に　スパコン富岳で9000年かかる計算を36マイクロ秒で

6/3(金) 16:50 配信 67

ITmedia NEWS



Natureに掲載されたBorealisの構成図

カナダの量子ベンチャーXanadu（ザナドゥ）は6月1日（現地時間）、特定のタスクで世界最高性能のスーパーコンピュータの計算速度を上回るとする光量子コンピュータ「Borealis」をAmazon Web Services（AWS）上で提供すると発表した。

【画像】AWSが掲載したお知らせ

XanaduはBorealisを使って、量子コンピュータの計算能力が従来のスーパーコンピュータを上回ることを示す「量子超越性」を持つことを実証。「初めての完全にプログラマブルな光量子コンピュータであり、量子超越性を持つマシンがクラウドで一般に公開されたのも初めてだ」と同社は説明している。この成果は、英科学雑誌「Nature」に6月1日付で掲載された。

Borealisは、ユーザーが指定したプログラムに従い、3次元的に絡み合った216個のスクイズド状態（量子ゆらぎを抑えた状態）の光量子ビットを合成し、計算を行う。スーパーコンピュータ「富岳」が9000年かけて行う計算をBorealisでは、36μ秒（1μ秒＝100万分の1秒）で計算できるという。

量子コンピュータの発展と将来展望

2020年以降の主要マイルストーン

🚀 2019年：量子超越の実証

Google が53量子ビットのマシンでスーパーコンピュータが1万年かかる計算を200秒で完了

📈 2020-2024年：急速な技術進展

- **IBM**：2023年に1,121量子ビット「Condor」プロセッサ発表
- **量子ビット数の指数的増加**：2016年の5量子ビットから2024年には1,000量子ビット超へ
- **日本**：2023年3月、理化学研究所が国産初の量子コンピュータを公開

164億ドル

2027年予想開発投資額

76億ドル

2027年予想市場規模

現在の技術状況と主要課題

技術の現在地

量子コンピュータは現在**「真空管レベル」**の発展段階にあり、実用化に向けて多くの技術的ブレークスルーが必要です。

🔧 主要な技術課題

- **量子エラー訂正**：信頼できる量子ビット1つに1,000個の物理量子ビットが必要
- **ノイズとの戦い**：極低温環境（-250度以下）での安定動作が必要
- **スケーラビリティ**：実用レベルには数百万〜数千万量子ビットが必要
- **人材不足**：量子プログラミングができる専門人材の深刻な不足

開発方式の多様化

超伝導方式
IBM、Google採用
高速動作が特徴

イオントラップ方式
高精度動作
長期安定性に優位

2030-2050年の将来展望

実用化ロードマップ

2025-2030	2030-2035	2035-2050
限定的実用化	量子優位性確立	社会実装拡大
NISQ段階での特殊用途活用	実用的性能を持つ量子コンピュータ	産業インフラの一部として定着

2035年までの主要目標（専門家予測）

- **ハードウェア**：数百〜数千量子ビットの安定動作
- **アプリケーション**：材料開発分野での実用的成果
- **エコシステム**：量子プログラミング人材の本格育成

経済インパクト予測

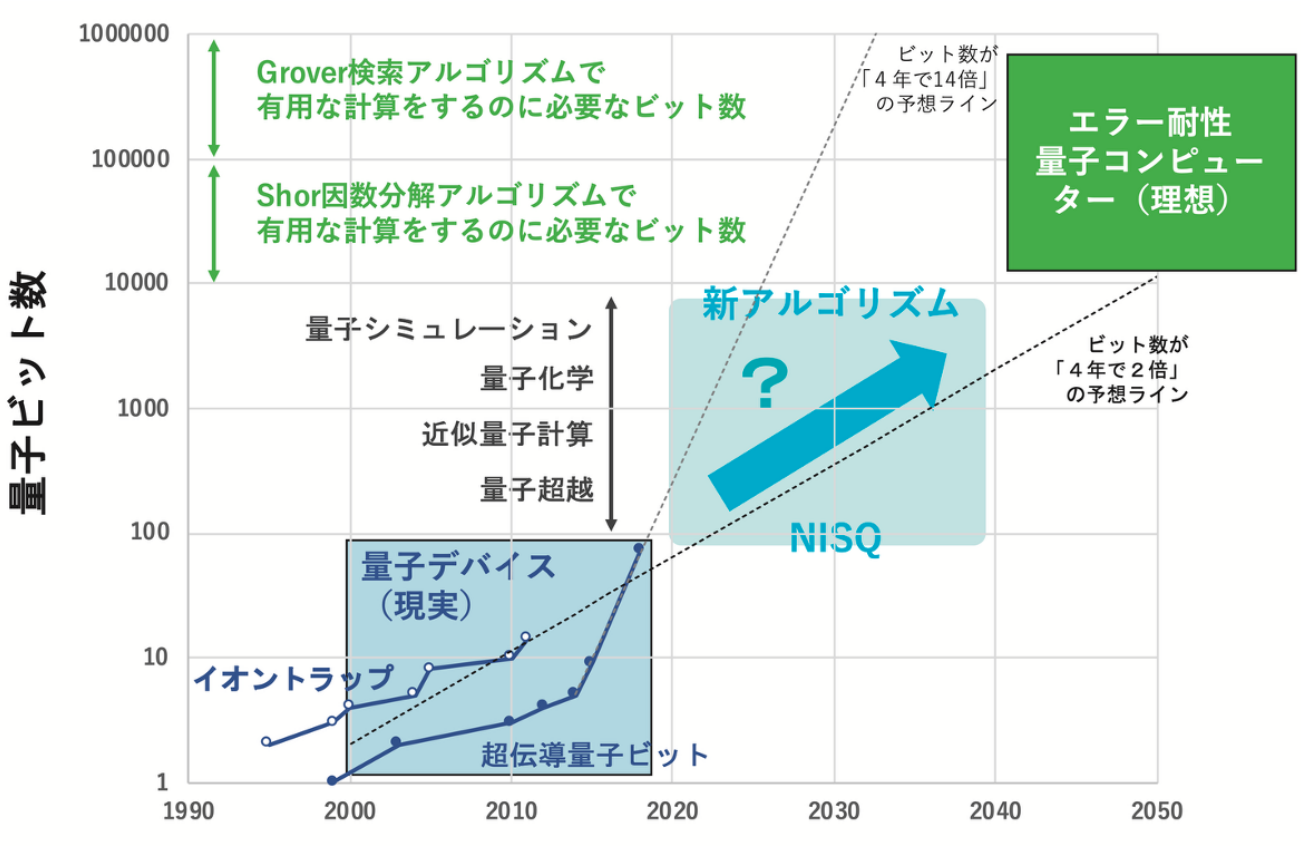
8,500億ドル

2040年までの経済価値創出
（BCG予測）

1兆円超

主要4産業での
最低経済効果

量子ビット数の推移と予測



東大、156量子ビット「IBM Heron」プロセッサ導入 スパコン「Miyabi」とも接続

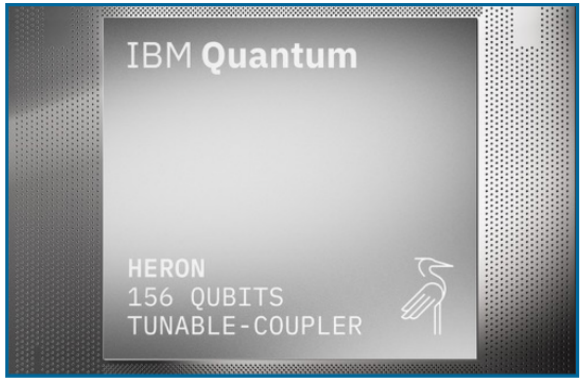
2025年05月16日 12時30分 公開 [ITmedia]

印刷 見る Share B! 2 0

PR 【驚速】取材後2週間足らずで初稿完成！ 採用サイト制作の裏側とは？

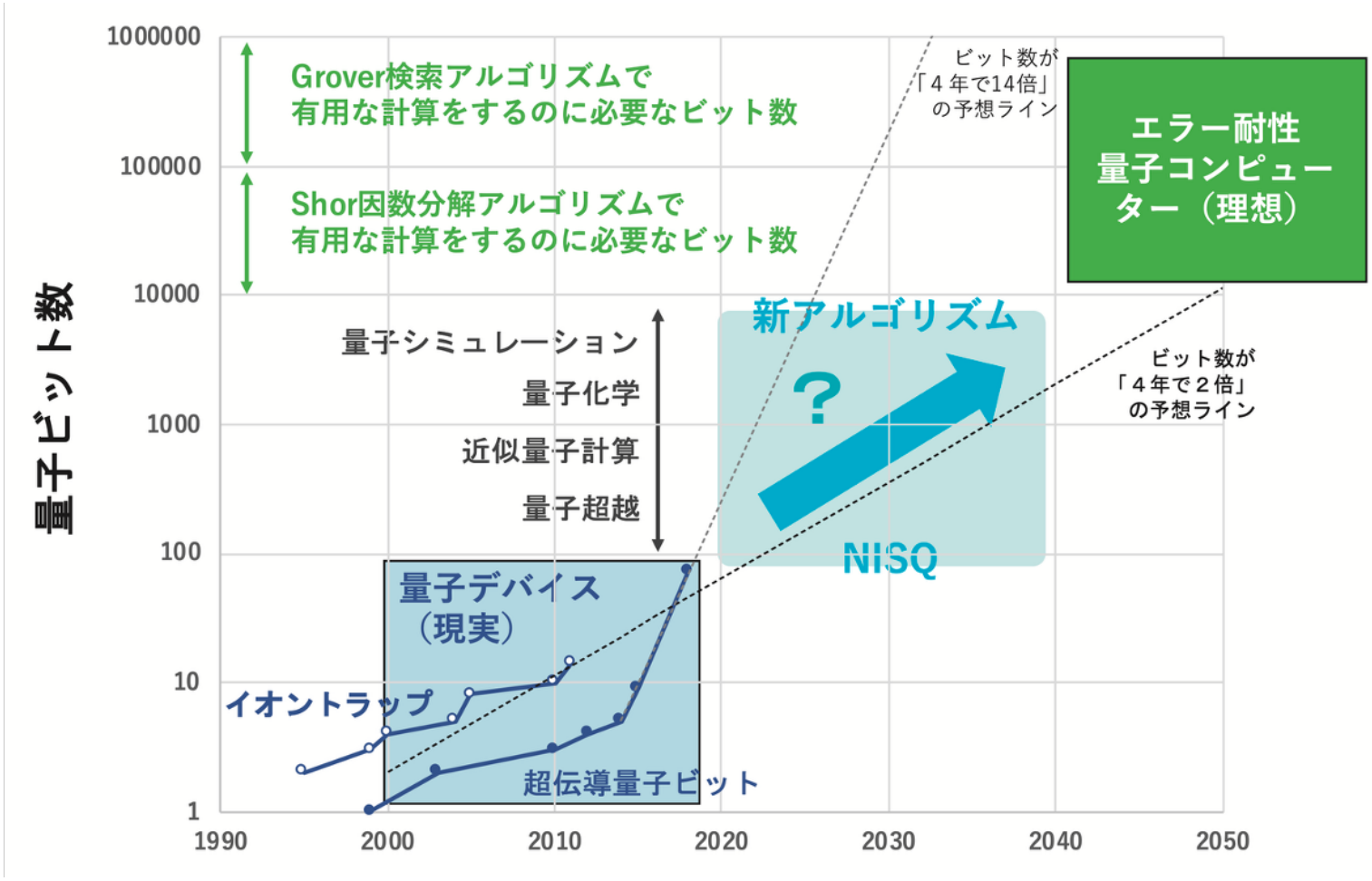
PR リスクに対処しながら生成AIを活用する「Copilotワークショップ」を開催中

東京大学とIBMは5月16日、同大に設置・運用する量子コンピュータ「IBM Quantum System One」を、最新世代の156量子ビットプロセッサ「IBM Heron」へ2025年後半にアップグレードすると発表した。2023年に同システムへ組み込まれた127量子ビット「IBM Eagle」を上回る性能を備えるという。



IBM Heronプロセッサ（出典：日本IBM）

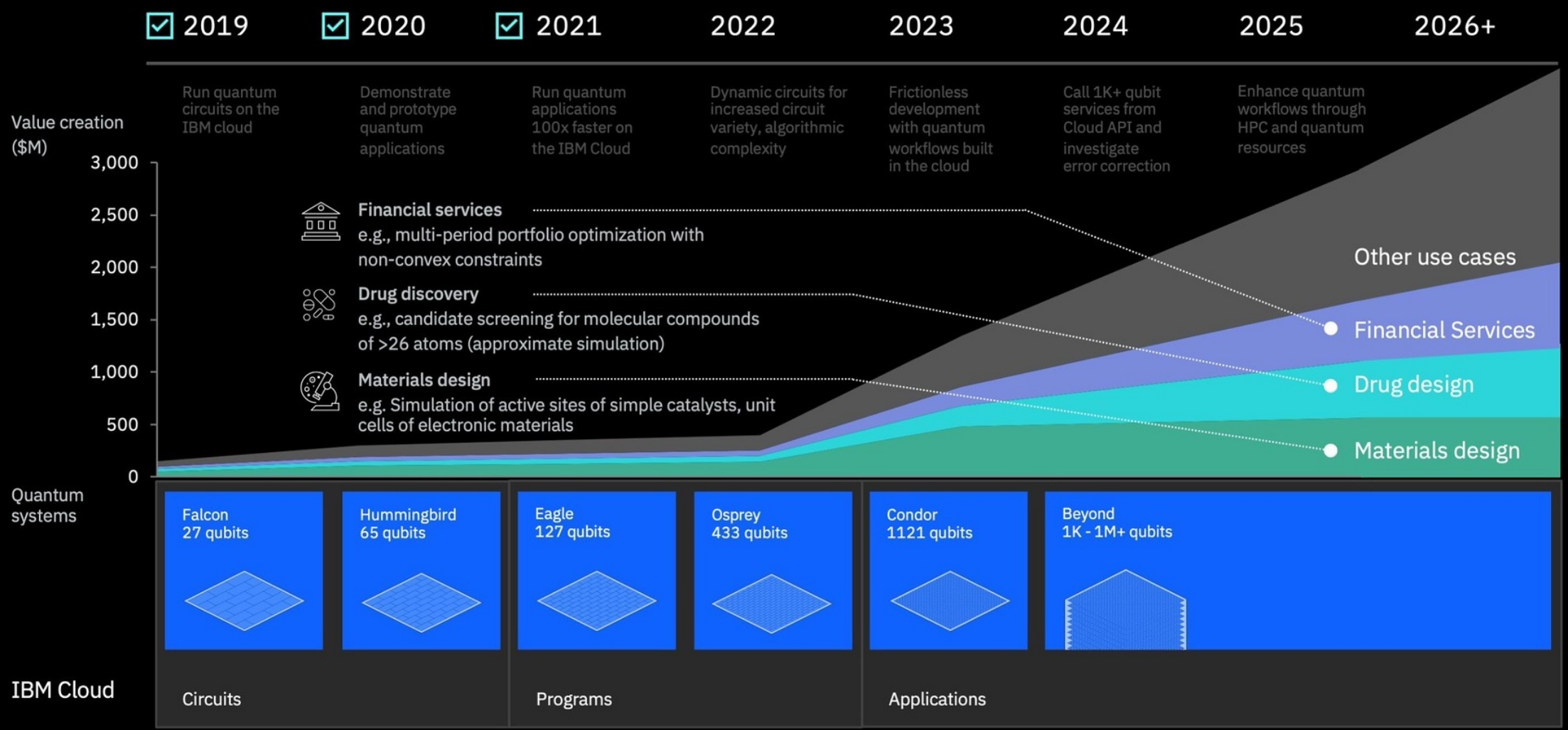
量子ビット数の推移と予測



$3B+ in near-term value creation based on IBM roadmap, with inflection points starting in 2023

IBM Quantum

BCG BOSTON CONSULTING GROUP



データサイエンス特別講義 まとめ

第一回　宇宙、地球、環境

第二回　統計計算宇宙論I　宇宙画像・動画データ分析

第三回　統計計算宇宙論II　逆問題とスパースモデリング

第四回　統計計算宇宙論III　多次元データ解析とパラメータ推定

第五回　ハイパフォーマンスコンピューティング

第六回　シミュレーションからエミュレーション，そして量子へ