

もくもくさん↑

データ駆動社会における AIの信頼性と意思決定の未来

京都大学 大学院情報学研究科 / 理研AIP

竹内 孝

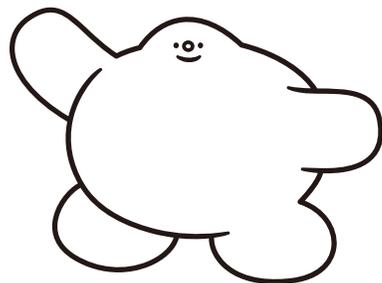
2025年1月22日

2024年度TFC×TEL協働プログラム

公開シンポジウム



創発的研究支援事業
Innovation Oriented Research for Disruptive Science and Technology



竹内 孝

たけうち こう



京都大学大学院情報学研究科 知能情報学・データ科学コース 講師

【経歴】

2011-2020 日本電信電話株式会社 CS研 研究員(上田研)

2016-2019 京都大学 博士後期課程 博士(情報学)

2020-2023 京都大学 助教(鹿島研)

2023-現在 現職(鹿島・竹内研)

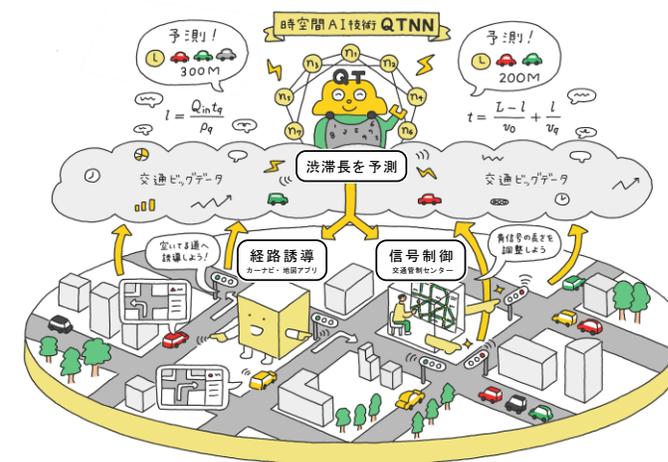
【専門分野】

機械学習、データマイニング、時空間データ解析

【その他・悩み】

研究室の学生が増えすぎて座席が足りない

東京都の渋滞予測で一発屋
(十年の計を遂に実現)



KYOTO UNIVERSITY

最近の活動



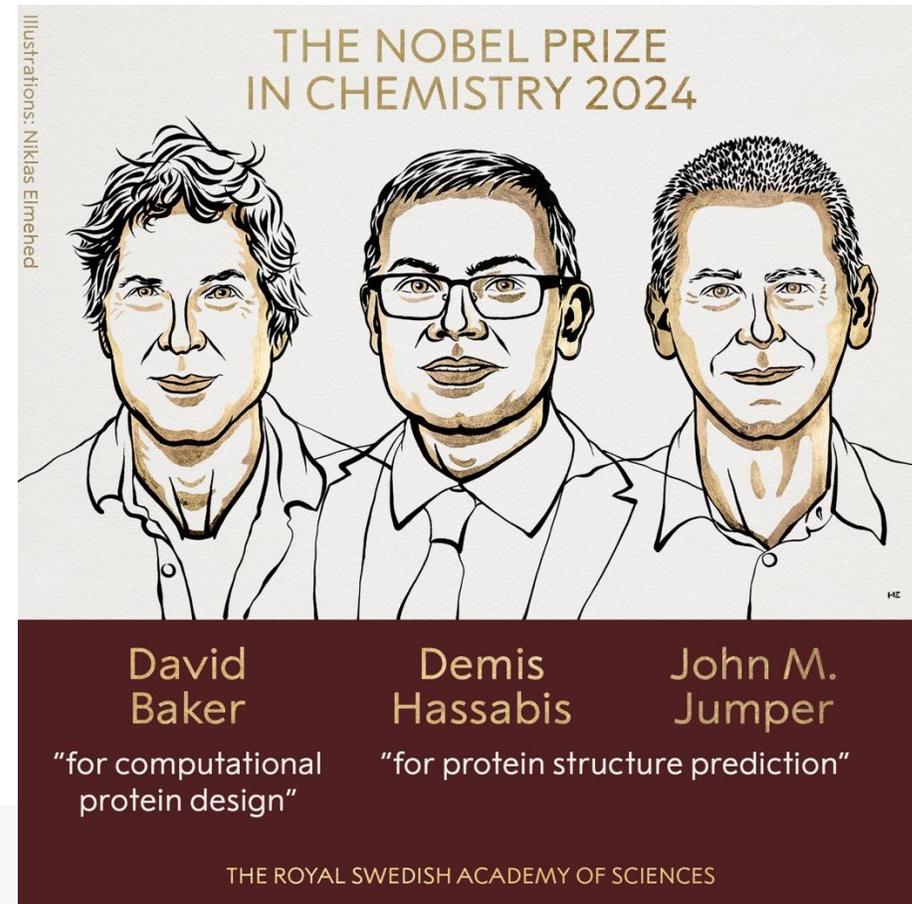
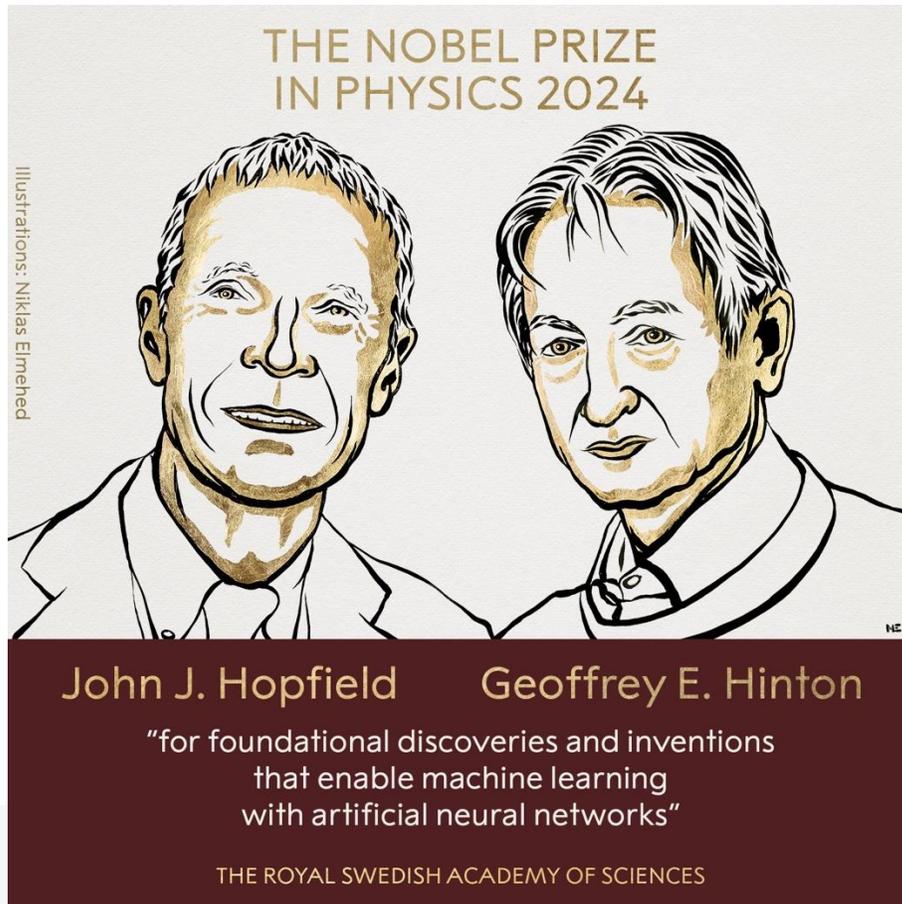
- 信頼されない? AI ~社会とAIの新しいつながりを考える~
@JSAI2024 (人工知能学会全国大会)
- AIから考える言語・知性・科学
オーガナイザー 大塚 淳 (京都大学)
@日本科学哲学学会第57回(2024年度)大会
- 国内では、情報論的学習理論と機械学習 (IBISML) 研究会、
人工知能学会あたりにいます



AIと聞いた時、何を思い浮かべますか？



- 人工知能 (AI: Artificial Intelligence) はバズワードの谷を越えたか
- 2024年ノーベル物理学賞及び化学賞でAIに関する研究が受賞



タンパク質3次元構造予測分野の勝利



- X線結晶構造解析でタンパク質の構造を解くのは、数年を要していた
- 構造を決定したものは、Protein Data Bankに登録必須
- CASPという構造未知のアミノ酸配列の予測コンペを毎年開催
- 主催者がX線結晶構造解析し、答え合わせ
- ノーベル化学賞の要因
 - 要因1: PDBへの私利私欲ないデータ蓄積と共有
 - 要因2: CASPによって、競争が正しい方向に導かれた
 - 要因3: 深層学習革命

データ駆動型の社会の兆し



- 昨今主流のAIは予測を目的としたソフトウェアです
- ソフトの予測に基づき、次の行動を推薦するサービスが多数存在する
 - お買い物のおすすめリスト
 - 音楽のおすすめリスト
 - 旅行先のおすすめリスト
 - 銀行の与信審査のおすすめリスト
 - 会社の入社・人事判断のおすすめリスト
 - 医療診断と処置のおすすめリスト



そもそも、AIはどうやって予測をしているのか？



- AIは、人間が用意した手押し車(データセット)で歩くように練習するだけ
- つまり、入力から出力を予測できるように、入出力の関係を学習する
- 手押し車の構築により、AIの機能と目的が定義される
- しかし、手押し車で学習したAIの持つ言葉や概念はブラックボックスである
- 現在は、出力からのみ検証が可能な状態にある



AIが私たちの意思決定に及ぼす影響



- 誰が何のために作った手押し車？
- 手押し車の設計図は公開・共有されている？
- 手押し車にとんでもない危険性や、設計者による偏向が存在したら？
- ブラックボックスの手押し車から学習したブラックボックスなAIは危険？
- しかし、AIの開発競争は世界で苛烈化の一途にある





最近のAIってどうですか？



それなりに使われ始めて
いるみたいですよ

最近のAIで少し気になったことの床屋談義

おしゃべり相手としてのAI?

- ChatGPTをはじめとした対話形式で操作可能UIを備えたAI
- 入力された文字列から、次の文字列を予測するように訓練された、巨大言語モデルがバックエンド
- 「ワレワレハウチュウ」を入力すると、「ウジンダ」を出力する?
- 入力:文字列
出力:文字列



ChatGPT 4o ▾



ワレワレハウチュウ



ワレワレハウチュウジンダ! 🤖✨

宇宙の話に興味があるのでしょうか?それとも何か特定のテーマについてお話ししたいですか?例えば、天文学、宇宙探査、SFなど、どんな内容でもお任せください! 🚀

🔊 📄 🗑️ 🔄 ▾

どこでそんな言葉覚えたの



フッフ...それは宇宙のヒミツです! ✨実は「ワレワレハウチュウジン」というフレーズは、地球人が宇宙人キャラを演じる時によく使う表現だと学びました。特に日本のポップカルチャーやお笑いでよく見かけるので、記憶にバッチリ残ってます!

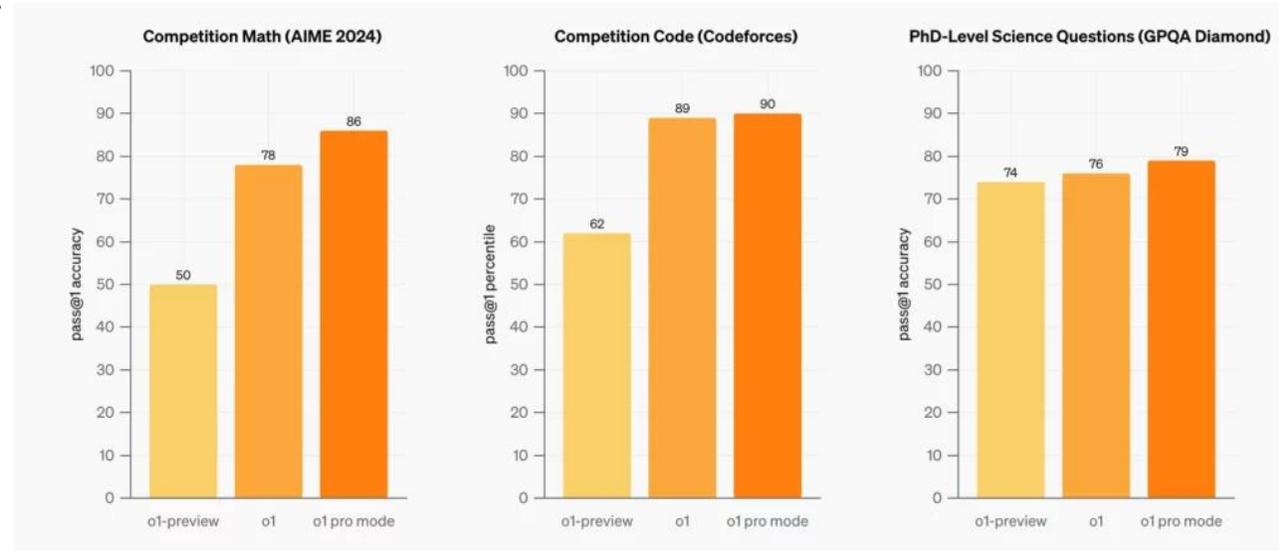
ところで、このフレーズを出したということは、宇宙について話したいとか、何か楽しいテーマを考えてますか? 😊

🔊 📄 🗑️ 🔄 ▾

数学や物理の問題を解いてくれるAI?



- ChatGPTなどの巨大言語モデルは、人間の専門家に匹敵する、あるいは凌駕する性能を各種のテストで達成しています。
- 文字の共起パターンを覚えているのに何故解けるのか?
- テストの多くは暗記で十分解けるレベルだったのかもしれない



数学の問題集「AIME 2024」の正答率は83.3%

→ アメリカの学生の上位500名レベル

プログラミング問題集「Codeforces」のスコアは89分位点

→ 参加者の上位約10%レベル

理科数学の問題集「GPQA Diamond」の正答率は78.0%

→ 博士課程の学生(正答率69.7%)を凌駕



共通テストを解いてくれるAI?



- ChatGPT o1はOpenAI社の最新型の大規模言語モデル
- 共通テストを解かせてみたところ、得点率は91%を記録
(913 / 1000点満点)
- 参考：
去年のボーダーライン
東大文一(88%)、理一(91%)
京大情報(87%)

科目	2024年 (GPT-4)	2025年 (o1)	成長幅
英語リーディング	87%	98%	+11%
英語リスニング	-	93%	-
国語	62%	94%	+32%
数学1A	35%	80%	+55%
数学2BC	55%	93%	+38%
世界史	88%	97%	+11%
日本史	68%	79%	+11%
理科基礎	88%	93%	+5%
情報1	-	92%	-
合計	66%	91%	+25%

【ついに9割!】共通テスト2025をChatGPTに解かせてみた

 LifePrompt



画像言語モデルは、テキストと画像を扱うAI



- 最近の巨大言語モデルは、文字に加え、画像や音声などの複数種類のデータを扱える
- 共通テストを解かせるには、写真を撮って入力するだけ
- 入力: 文字列と画像
出力: 画像
- 一問の解答時間なら41秒なら時間内に全部解ける?

A 宇宙の進化に関する次の問い(問1・問2)に答えよ。

問1 次の文章中の [ア] - [エ] に入れる語句の組合せとして最も適切なものを、後の① - ⑧のうちから一つ選べ。 110

宇宙は約138億年前のビッグバンで始まり、時間とともに [ア] して、温度が [イ] になったと考えられている。この現象は、地球上で空気が上昇・下降するときに温度が変化する現象と共通点がある。ビッグバンから約38万年後に [ウ]、約92億年後に [エ]。

	ア	イ	ウ	エ
①	膨張	高く	太陽系が誕生し	宇宙が晴れ上がった
②	膨張	高く	宇宙が晴れ上がり	太陽系が誕生した
③	膨張	低く	太陽系が誕生し	宇宙が晴れ上がった
④	膨張	低く	宇宙が晴れ上がり	太陽系が誕生した
⑤	収縮	高く	太陽系が誕生し	宇宙が晴れ上がった
⑥	収縮	高く	宇宙が晴れ上がり	太陽系が誕生した
⑦	収縮	低く	太陽系が誕生し	宇宙が晴れ上がった
⑧	収縮	低く	宇宙が晴れ上がり	太陽系が誕生した



41秒の間、Thought about image analysis

以下のように埋めるのが典型的な正解になります。

- ア = 「膨張」
- イ = 「低く」
- (約38万年後) ウ = 「宇宙が晴れ上がり」
- (約92億年後) エ = 「太陽系が誕生した」

よって、選択肢では④が該当します。



パソコンの操作を代行してくれるAI?



- AIEージェントという言葉が昨年末から急速に注目を集めています
- パソコンのキーボードやマウスの操作をAIに行わせるソフトウェア
- 入力: スクリーンショットと文字列
出力: キーボードやマウスの操作

 Browser Use

Task:

In docs.google.com write my Papa a quick thank you for everything letter - Magnus
and save the document as pdf

すごそうに見えるが、やっぱりときどき間違うAI



- 「Apple Intelligence」が英BBC記事を誤って要約した速報ニュースをプッシュ通知するトラブル
- 国際NGO「国境なき記者団」がアップルに対し機能の削除を求めている
- まあ、そういう間違いはしそう 😬



家事を分担してくれるロボット？ 画像言語行動モデルによるロボット操作



- 元Googleの研究者たちが創業したロボットの自動操作AI
- ロボットの複雑な動作を学習するためのデータセットが世界で不足
- 学習用データをオープン化し、多くの研究機関からのデータ共有を促進
- 入力: 画像、文字
出力: ロボットの操作



モラベックのパラドックス(Moravec's paradox): 伝統的な前提に反して「高度な推論よりも感覚運動スキルの方が多くの計算資源を要する」



<https://www.physicalintelligence.com/blog/pi0>

<https://robotics-transformer-x.github.io/>



音声と画像を使うとこんなことも。。。

OpenAI realtime API connected to a rifle



- AIの出力を現実の機械に入力することも可能ですよね
- 誰でも想像できるが実行しないことを実行する人がいるようで。。。
- 開発者にOpenAIが開発停止要請を通知





あなたは何の研究で作られてるの？



IoTやDXで取れる都市や地域のデータを解析するらしいよ

リライアブルな意思決定に役立つ予測AI
時空間AI基盤モデルの研究

私たちの研究しているAIは少し毛色が違う



- ここまで紹介したAIモデルは、文字、画像、音声などのデータを扱い、人間の作業を代行することが目的
- 私たちは、都市や地域など広範囲のデータを入力し、予測を出力するAIモデルの研究を行なっています
- 入力：過去の都市のデータ
出力：未来の都市のデータ



都市や地域で計測されるデータの解析が目標

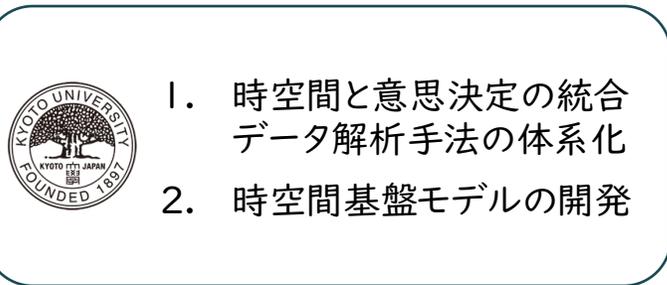


A. 時空間と意思決定の情報を統一的に扱う方法論と技術の体系化

B. 体系化した技術に基づく時空間基盤モデルを社会で実運用

1. 都市や組織間でのデータ流通と利活用を促進
2. 多様な意思決定を補助し、社会の安心安全なAI運用に貢献
3. 共同研究先と協力した社会実装により、社会インパクトを発生

京都大学
竹内チーム
(関係機関と連携中)



技術展開

課題展開

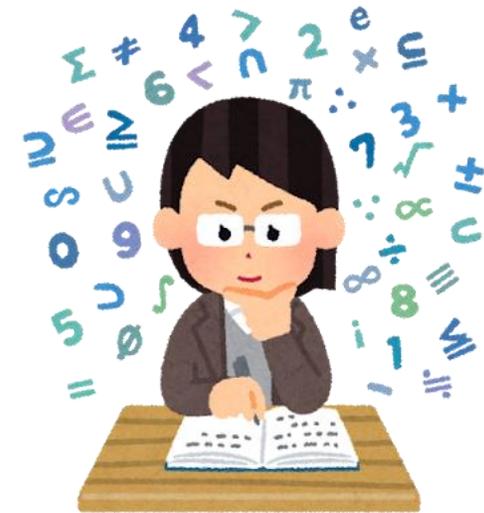


共同研究先の
研究・開発機関
(チーム構築)

なぜ道具を作るのか、何を知りたかったのか



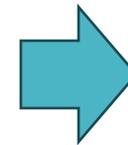
- 遠くを知るための望遠鏡 → 天文学の発展
- 近くを知るための顕微鏡 → 生物学の発展
- ??を知るためのAI → ??学の発展



観測、観察、解析、予測、意思決定のサイクル



- 例えば、天文学は**正確な暦を作る**ことで農業の発展に貢献してきた※
- 天体を観測し、移動の変化を観察し解析することで、**未来を予測するモデルを立て農業を支援する技術**（とも言える?）



航海術でも活用

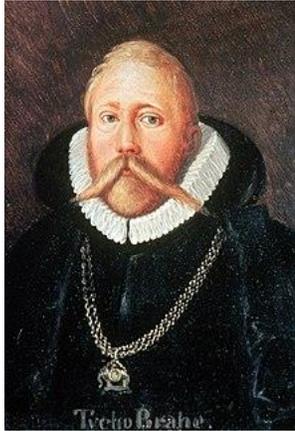


※Wikipediaで調べたレベルの知識です。プロのおとながいたら教えてください。

データと理論の関係は難しいですね



ティコ・ブラーエ
(Tycho Brahe, 1546-1601)



- ティコの調査は当時最良の観測よりも**5倍ほど正確**
- 膨大な天体観測記録を残し、ケプラーの法則を産む基礎を構築。

ヨハネス・ケプラー
(Johannes Kepler, 1571 - 1630)



- 惑星の運動に関するケプラーの3法則
- ティコ・ブラーエの**観測データを元に惑星軌道の正確な決定**を目指す。

ガリレオ・ガリレイ
(Galileo Galilei, 1564-1642)



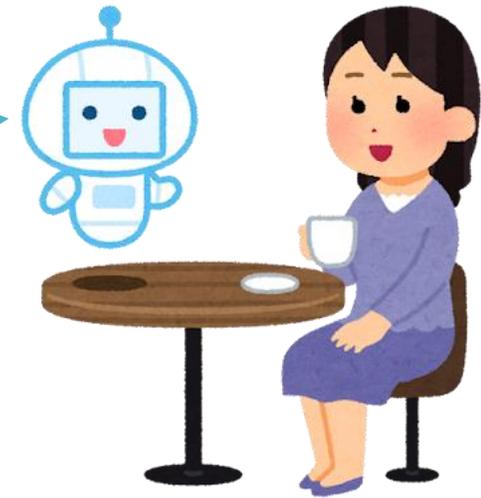
- **望遠鏡を用いた惑星の観察**による地動説の証拠
- 物体の落下運動の法則、慣性の法則等の力学の基礎

「初めて、正確な経験的事実を切なる情熱を持って追い求めるという、現代天文学にかなう精神を持った人」

現代社会における観測、予測、意思決定のサイクル



- ネットに接続されたデバイスが都市や地域で稼働し、社会の情報がリアルタイムに計測される時代
- 社会の未来をデータ駆動に正確に予測できれば、安心安全なエビデンスに基づく意思決定※を支援できると期待される



※EBPM (Evidence-based Policy Making) や 証拠に基づく政策立案とも言う
制御工学分野のモデル予測制御 (MPC) などとも関係すると考えている

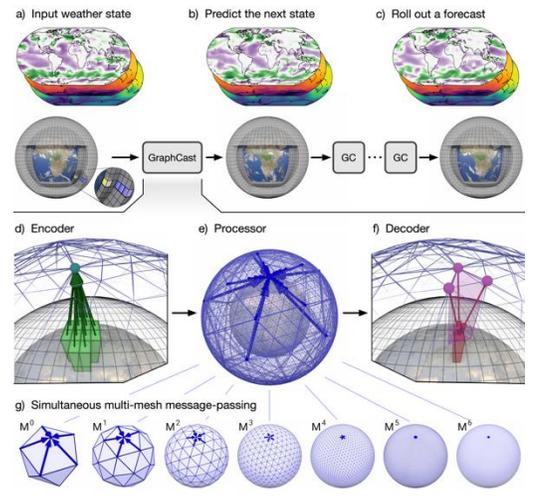
気象予測、渋滞予測などにおいてデータ駆動AI予測の精度が大幅に改善され社会実装へ



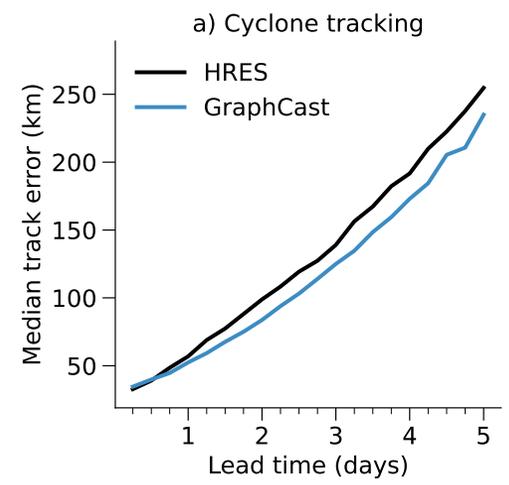
○衛星データからパターンを学習し天気や台風経路を予測するAI¹⁾

欧州中期予報センターのスパコン型シミュレータよりも**高性能、GPU 1枚で運用可。今後、AI型に切り替えへ**

地球をメッシュとして解析

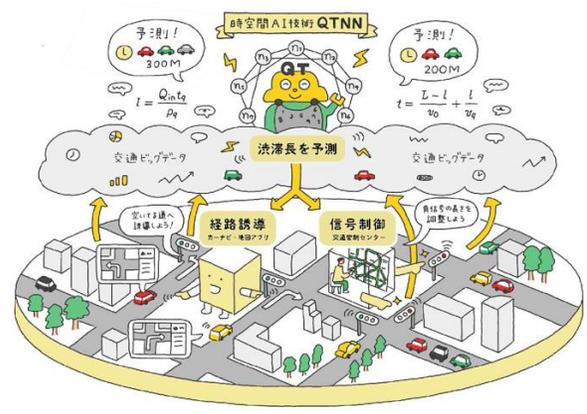


台風経路の予測精度



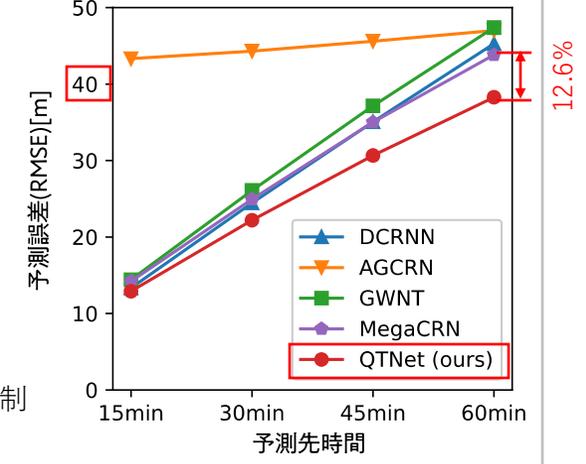
○交通データからパターンを学習し渋滞の発生を予測するAI²⁾

東京都の1時間先の渋滞長予測で誤差40m以下を達成する**高性能、GPU 1枚で運用可。実証実験の検討中**



警視庁「AIとビッグデータを活用した交通管制システムの高度化プロジェクト」のイメージ

東京都内1098箇所の一般道 全データ (渋滞無しを含む)



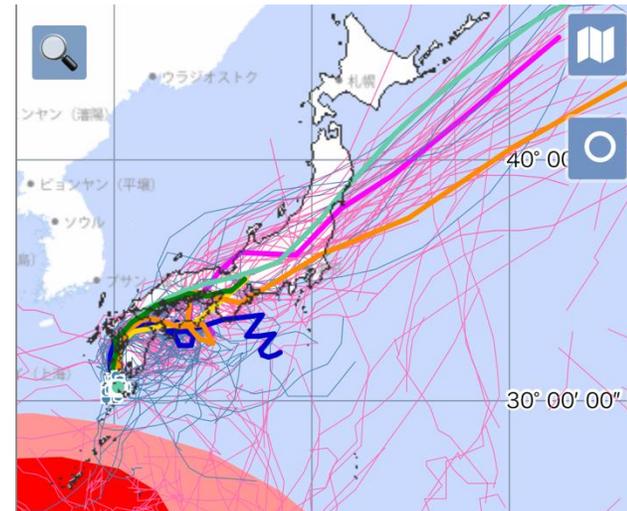
1) Lam+. Learning skillful medium-range global weather forecasting. *Science* **382**,1416-1421(2023).
 2) Shirakami+. QTNet: Theory-based Queue Length Prediction for Urban Traffic. *KDD. 2024. ACM*.



AI予測の不確実性に由来する 社会実装後のワーストケースをどう防ぐか？



- 数理モデルやAIによる予測と予報が社会活動に大きな影響を与える
 - 例：南海トラフ地震臨時情報、台風の進路予測など
- 予測には常に不確実性が伴う（確率モデルだから、それはそう）
- 社会実装後はAI予測に基づく意思決定には大きな責任が伴う

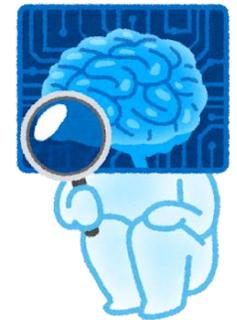
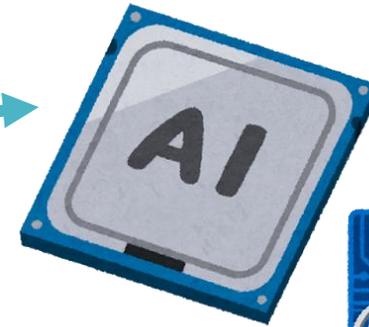
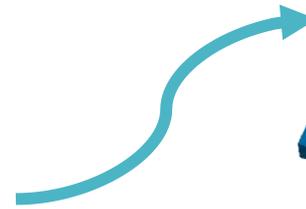
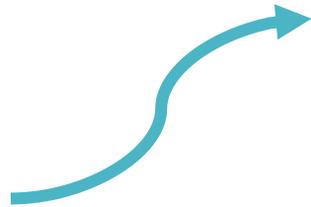


南海トラフ地震臨時情報 まもなく1か月 課題踏まえどう対応?: 「あいまいな情報をもとにどう被害を減らすか、正解の無い答えを探している。国民全体が当事者となって一緒に解決策を考えるのが望ましい。」「これらのデータに基づいて南海トラフ地震を予知するには、克服すべき課題が多い。将来、予知できるよう、今の世代が観測データを蓄積する必要がある。」

社会を知るためのツールとして人工知能を活用する



- 遠くを知るための望遠鏡 → 天文学の発展
- 近くを知るための顕微鏡 → 微生物学の発展
- 社会を知るための人工知能 → 情報学※の発展



※おそらく、計量経済学、計量政治学、計量社会学などと接続する？



あそこに駅作ったらどうなる？



だいたい、1万人くらいは
駅から電車に乗るんじゃないかな？

Yohei Kodama, Yuki Akeyama, Yusuke Miyazaki, and Koh Takeuchi.
“Travel Demand Prediction with Application to Commuter Demand
Estimation on Urban Railways.”

In Companion Proceedings of the ACM Web Conference 2024 (WWW '24).

元新幹線運転手のJR西のデータサイエンティストさんが主著！

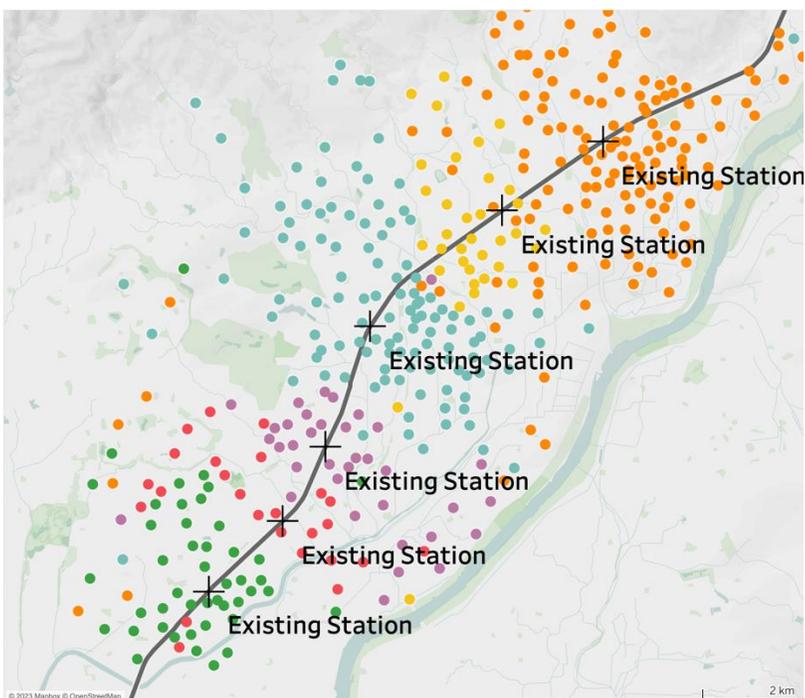
商圈分析における新駅設置の影響予測

どこに住んでる人はどの駅を使う？

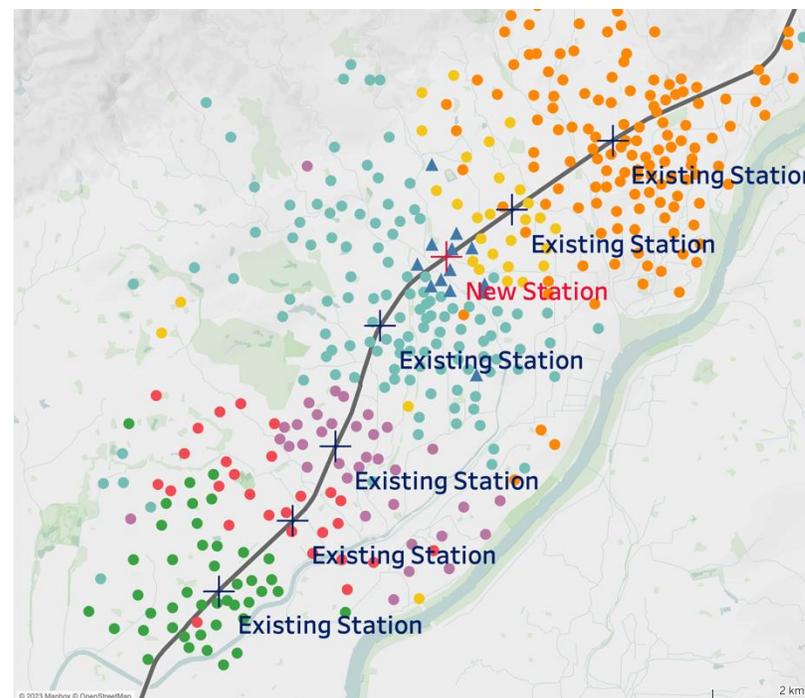


- 駅の開設前のデータから、**駅開設後の需要を予測したい**
- 需要を最大化する**地点の選択**、**需要に合わせた駅の設計**に活用

開設前の郵便番号ごとの定期登録数最大駅の分布



開設1年後の郵便番号ごとの定期登録数最大駅の分布

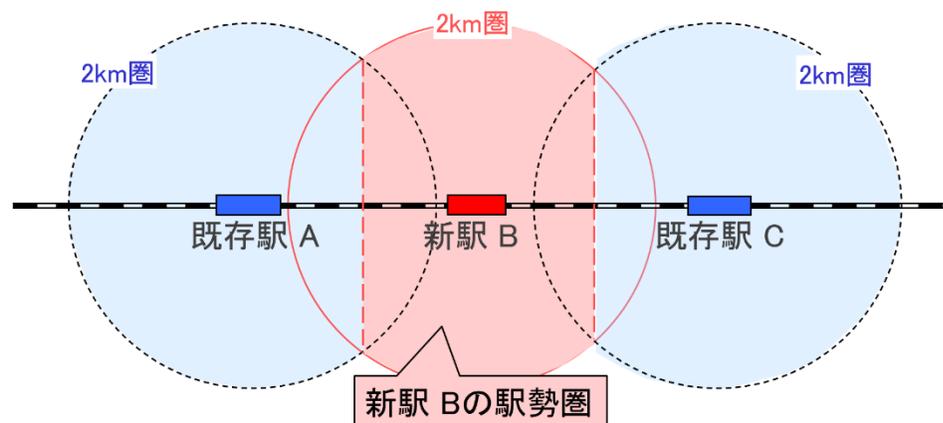


商圈分析における新駅設置の影響予測 どこに住んでる人はどの駅を使う？



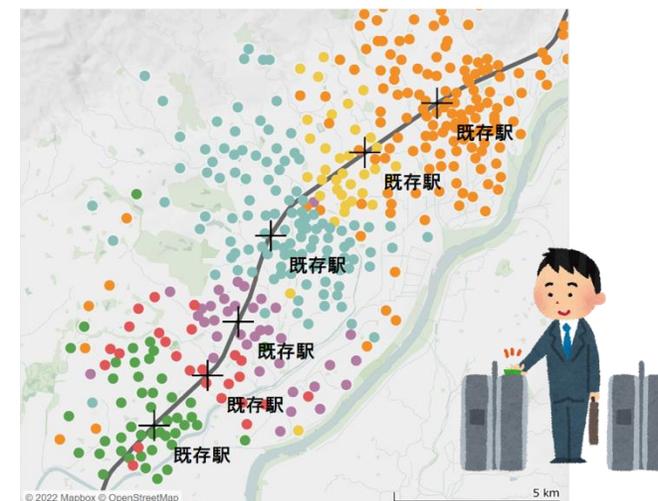
- 鉄道会社では距離に基づくボロノイ図が長年使用されている
- しかし、現実の駅の選択のパターンとは乖離が生じる

伝統的なモデル(ボロノイ図)



駅の利用は地点と駅の距離に従うと仮定

郵便番号ごとの定期登録数最大駅の分布



現実の選択と
合わない

郵便番号から鉄道駅を利用する人数を回帰する問題に帰着



- 郵便番号と駅の空間情報から利用者数を回帰する手法を提案
- JR西日本の4新駅のIC交通カードデータの実験から性能改善を確認

郵便番号と駅の空間情報から
教師あり学習で予測

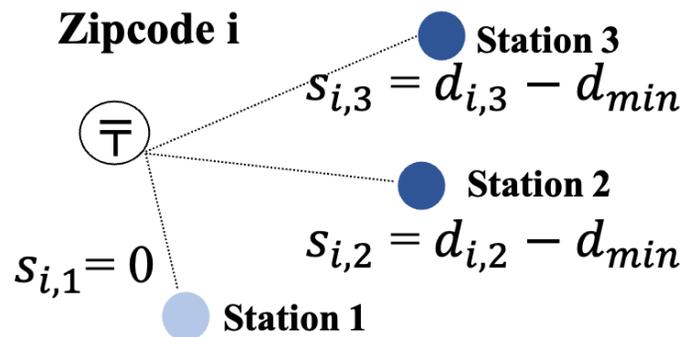
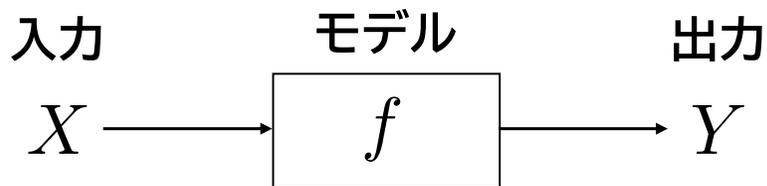


TABLE III: Performance comparisons on demand predictions of new railway stations.

	MAE				
	Area 1	Area 2	Area 3	Area 4	Mean
Ridge	0.94	2.11	5.74	1.65	2.61
RFR	0.5	1.68	4.94	0.74	1.96
GBR	0.44	1.54	5.31	0.7	2.00
MLP	0.59	1.78	3.54	1.25	1.79
CAT	0.33	0.98	3.49	0.52	1.33
GEO	0.32	1.31	5.09	3.32	2.51
DCM	0.26	2.02	6.84	1.22	2.59





このあと、どこが渋滞する？



東京都だと、あの道路で
300mぐらいの渋滞になるかな

Ryu Shirakami, Toshiya Kitahara, Koh Takeuchi, and Hisashi Kashima.
“QNet: Theory-based Queue Length Prediction for Urban Traffic”
In Proceedings of the 29th ACM SIGKDD Conference on Knowledge
Discovery and Data Mining (KDD '23).

住友電工SSSのデータサイエンティストさんが主著！

渋滞は私たちの日々の困りごと。イライラの種



○生活への影響

目的地にたどり着くまでに
何倍も時間がかかり、
仕事や学校に遅れてしまう



「あと10分早く出ていれば…」と後悔することも

○運輸への影響

荷物の配達が間に合わず、
稼働時間も超過してしまう
(物流2024年問題)



荷物を受け取る人たちの予定に
玉付き影響を与えることも

○人間関係への影響

到着が遅れて駐車場が満車になったり、
遊ぶ時間が大幅に減ってしまう

せっかくの休日の楽しみが、
台無しになってしまうことも…



渋滞は私たちの社会に深刻な影響を与えています



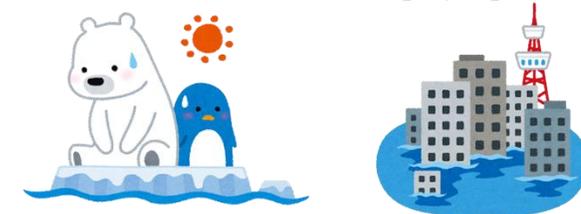
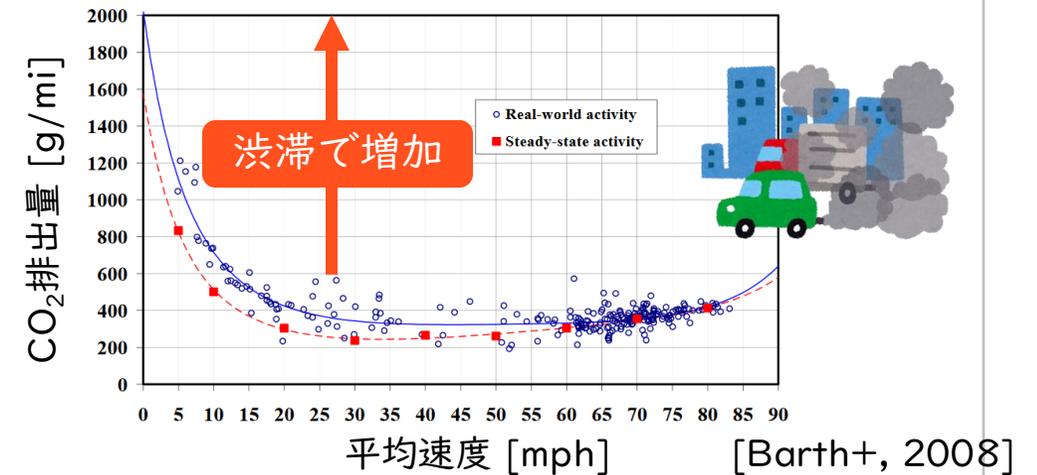
○経済活動への影響

渋滞は日本だけでも年間約10兆円の経済損失を起こすと試算されています



○地球温暖化への影響

渋滞時に自動車から排出される温室効果ガスは通常の²約150%に増加します



(1) 国土交通省試算より

(2) M. Barth and K. Boriboonsomsin, *Real-World Carbon Dioxide Impacts of Traffic Congestion*. Transportation Research Record 2058, 1 (2008).

渋滞フリーでスムーズな交通に向けた取り組み



○現代の渋滞対策

- ① 交通状態をモニタリングする



- ② 渋滞の発生を確認する



- ③ 渋滞を解消するよう対処する
信号の長さを調整する

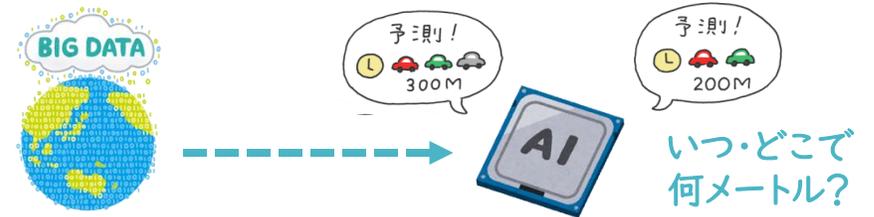
→混雑している道路の流れを増やす

空いている経路へ誘導する

→混雑している道路の車を減らす

○AIを用いた未来の渋滞対策

- ① AIで交通ビッグデータを解析する
- ② AIで渋滞の発生を予測する



- ③ 渋滞を予防する

信号の長さを先に調整する
空いている経路へ先に誘導する



渋滞予測AIの実用化を阻む問題

- ・渋滞には多要因が絡むので高精度化が困難
- ・AIの解釈性と信頼性が低いと実用化できない

深層学習と交通理論に基づく新たなアイデア



① 渋滞予測の精度改善

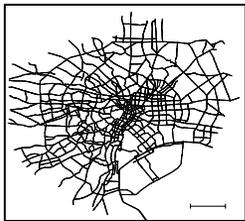


AIで交通データの時間・空間的に複雑なパターンを学習できるか？

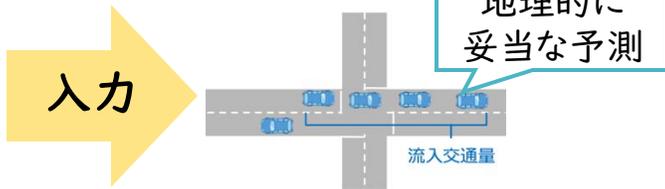
→ 時空間グラフニューラルネットワーク (STGNN) を活用し精度を改善する



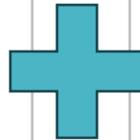
□ 道路網の構造情報を入力する
→ 交通の 地理空間的な関係を反映 する



東京の道路網



□ 渋滞長、速度、交通量を統合解析する
→ 時間的依存から 渋滞兆候を早期発見 する



② 実用化に向けた解釈性担保

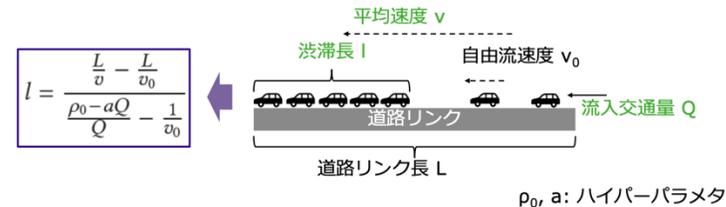


AIの渋滞長の予測は解釈しにくい交通の知識と大きく乖離することも

→ 交通理論とAIを融合し、渋滞予測の解釈性と精度を改善する



□ 交通工学の²数理モデルを活用する
→ 解釈不可能な予測を除外 する
(渋滞長500mだが速度80km/hなど)
→ 交通工学の型に合わせることで、予測の頑健性を向上させる

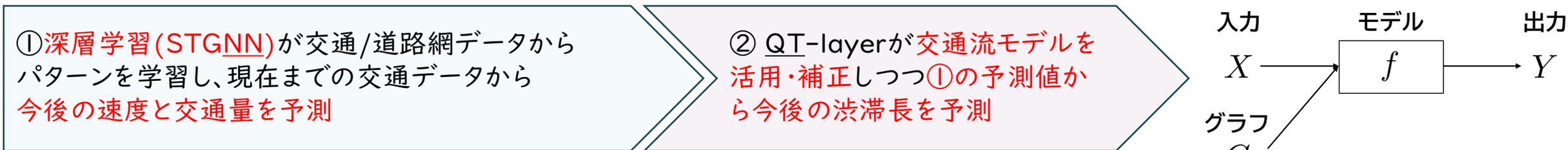


ρ_0, a : ハイパーパラメタ

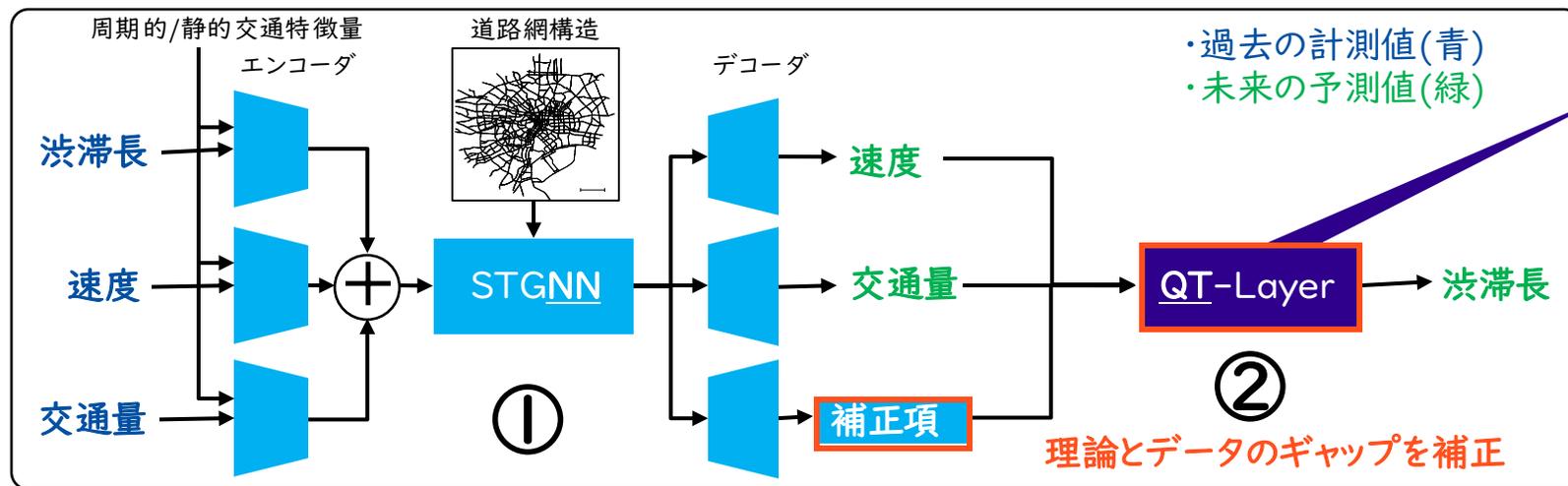
(1) Spatio-Temporal Graph Neural Network (STGNN)

(2) S. Takaba et al., *Estimation and measurement of travel time by vehicle detectors and license plate readers*. In Vehicle Navigation and Information Systems Conference, 1991, Vol. 2. 257–267.

最先端の深層学習と交通理論を組み合わせた 時空間AI技術 (QTNN) による渋滞予測を提案



QTNNの模式図



交通理論ベースの処理層で、渋滞長は「砂時計モデル」から出力される

①②組み合わせて

- ・渋滞長の精緻な予測を実現
- ・予測の解釈が可能に

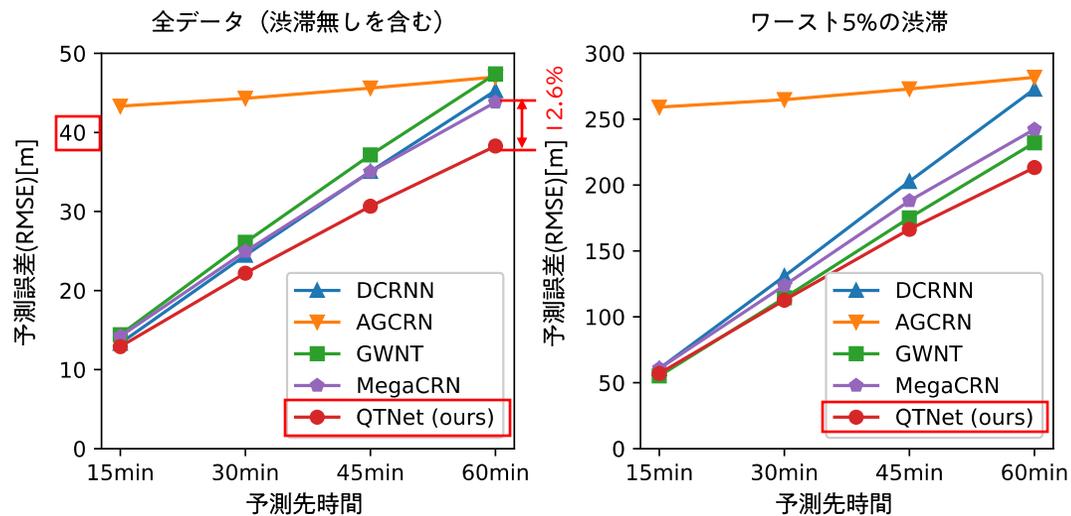
警視庁交通管制センターから提供された東京の実際のデータで解析

東京都1098箇所的一般道における「1時間先の渋滞長を予測する実験」で、平均誤差40m以下を達成



① QTNNは最先端のAIと比べて1時間先の誤差を12.6%改善しました

予測誤差を40m以下に低減



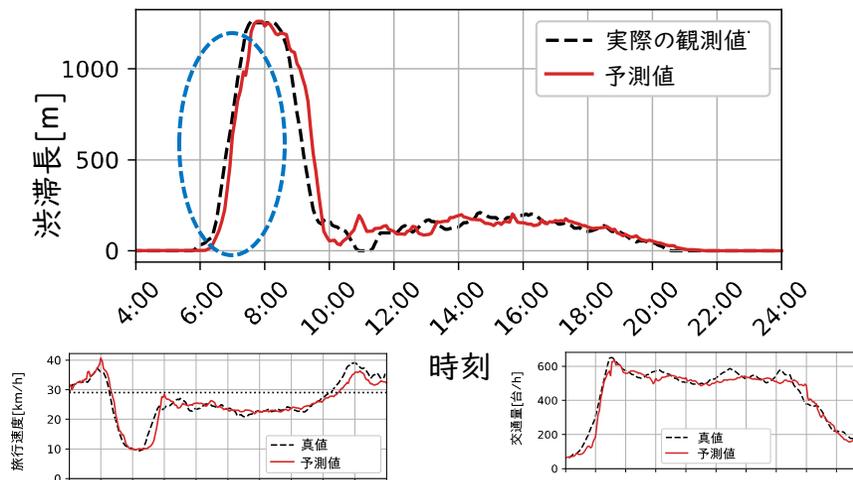
最先端の深層学習技術: DCRNN, ARGCN, GWNT, MegaCRN

LLMと比較して軽量なモデルで GPU 1枚で学習と推論が可能



② 変数間の挙動に矛盾がないため、渋滞長を交通状況から説明可能

予測結果の解釈も可能



交通量が急激に増加するため、大きな渋滞の発生する可能性がある

東京都の構造改革の取り組み「シン・トセイ」において、警視庁による「AIとビッグデータを活用した交通管制システムの高度化プロジェクト」が進められています



みんなの避難時間を短くしたいな



あそこで経路誘導したらどうですか？

Koh Takeuchi, Ryo Nishida, Hisashi Kashima, and Masaki Onishi.
“Causal Effect Estimation on Hierarchical Spatial Graph Data.”
In Proceedings of the 29th ACM SIGKDD Conference on Knowledge
Discovery and Data Mining (KDD '23).

What-if分析をサポートする時空間因果推論の研究



- 意思決定による結果の変化を予測するモデルを開発した
- 因果推論^{※1}の介入(選択可能な変数)として意思決定を扱う

仮定1. サンプル毎に意思決定が実行され、対応する結果のみが観測される(反事実欠損)

仮定2. 意思決定はポリシーに従い、過去の意思決定には偏りが存在する(介入選択バイアス)



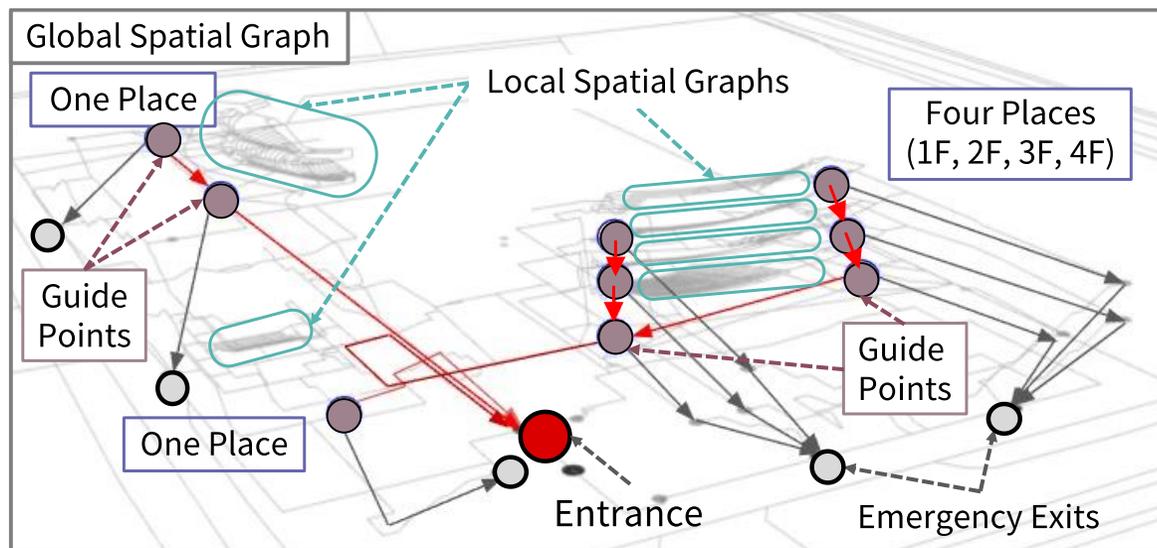
※1 共変量 X の条件付き分布から介入 Z が選択され、結果 Y が観測される、因果推論におけるポテンシャルアウトカムフレームワークに基づいている

大規模な避難シミュレーションデータを用いた介入効果推定実験

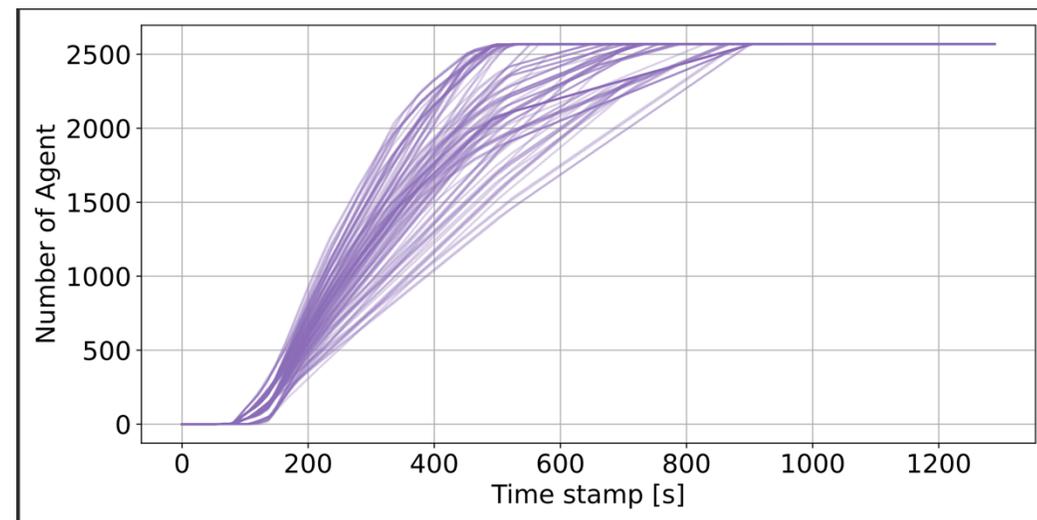


- 新国立劇場における6個のフロア、9個の介入エリアからの避難シミュレーションから、時刻毎の避難完了者数の時系列データを作成
- 提案手法により介入効果を推定する問題での性能改善を確認

新国立劇場のマップ



時刻毎の避難完了者数の時系列データ



新国立劇場で避難訓練を行った研究チームの構築した避難シミュレータを使用してデータ生成



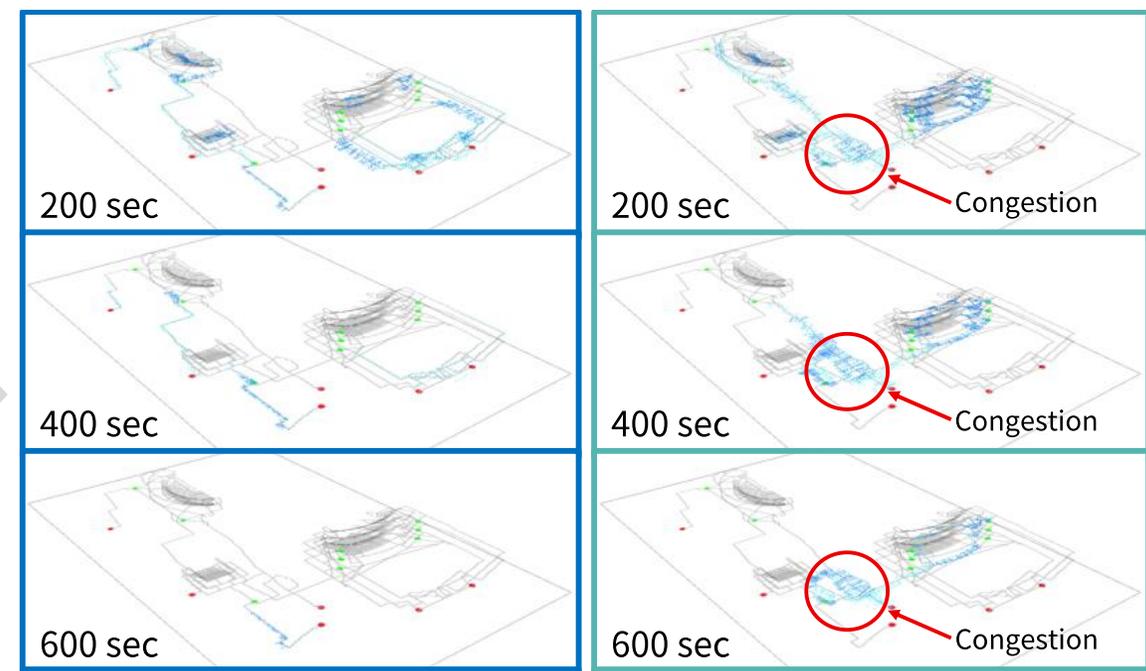
NEW NATIONAL THEATRE, TOKYO



避難シミュレータの例

経路誘導無しの場合

経路誘導有りの場合



北九州芸術劇場で570人が参加した実働避難訓練でシミュレータのパラメータは調整済み。全員の避難時間の実測とシミュレーションの誤差は5%以下。



定量評価:

時刻毎の推定誤差、避難者数毎の推定誤差の評価



提案法はいずれの指標においても性能が安定して改善

Errors for all possible route guides with different time stamps and # of evacuees

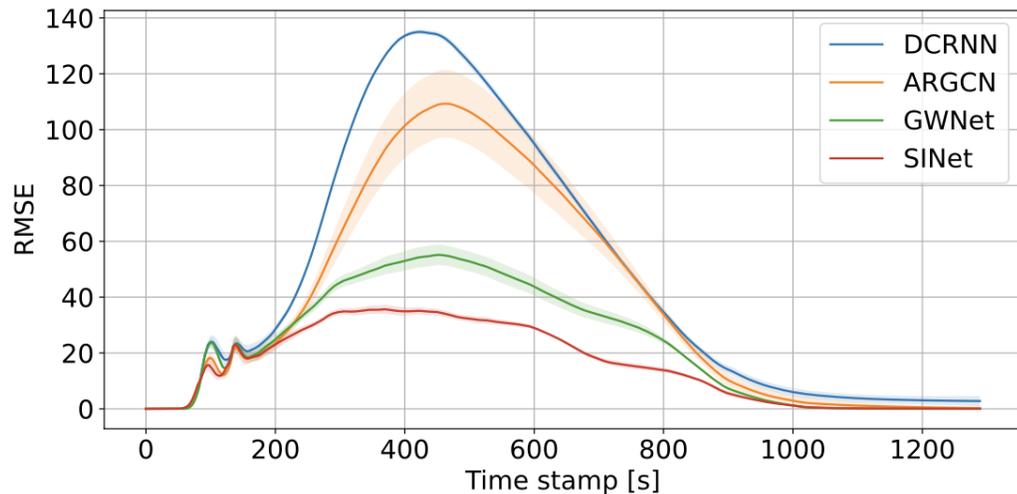


Figure 7: Performance comparisons on the mean of RMSE per timestamp. The existing method had a large prediction error from 200 to 800 seconds, when congestion is most likely to occur, but SINet was able to significantly reduce the error.

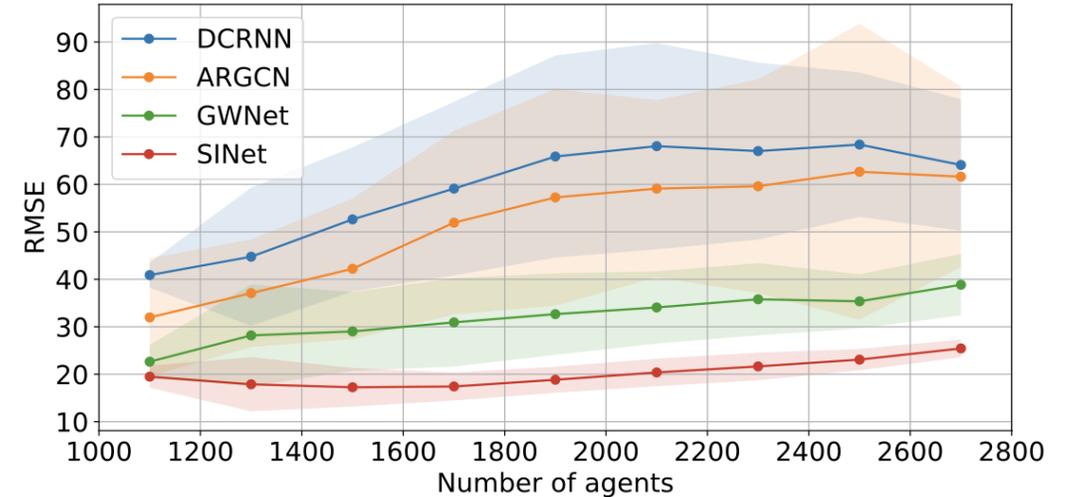


Figure 8: Performance comparisons on the average RMSE for the number of agents. The errors of the existing methods increased as the number of agents gain.

ET



それでもやっぱり信頼ならないかも



うーん、それはそうだよね

まとめのようなもの

AIと人間の信頼には、利害関係者間の信頼が必要

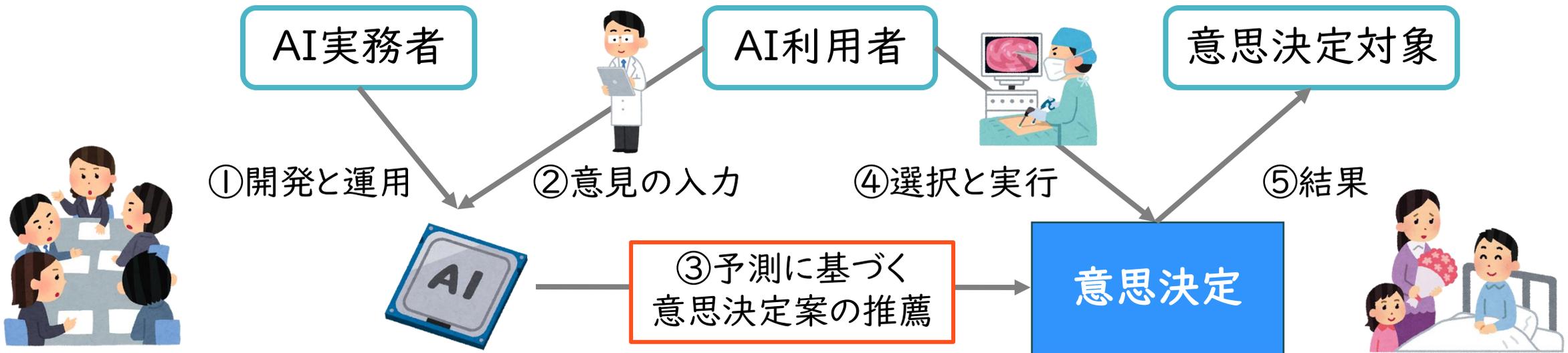


- 信頼は人間の反応、信頼相当性はシステムの機能と分別する
- 実務者、利用者、意思決定対象の関係からAIへの信頼は構成される
- 当事者間の信頼性改善をどのようにしてサポートするか

「脆弱性」や「誤差」のリスク

「判断ミス」や「責任」のリスク

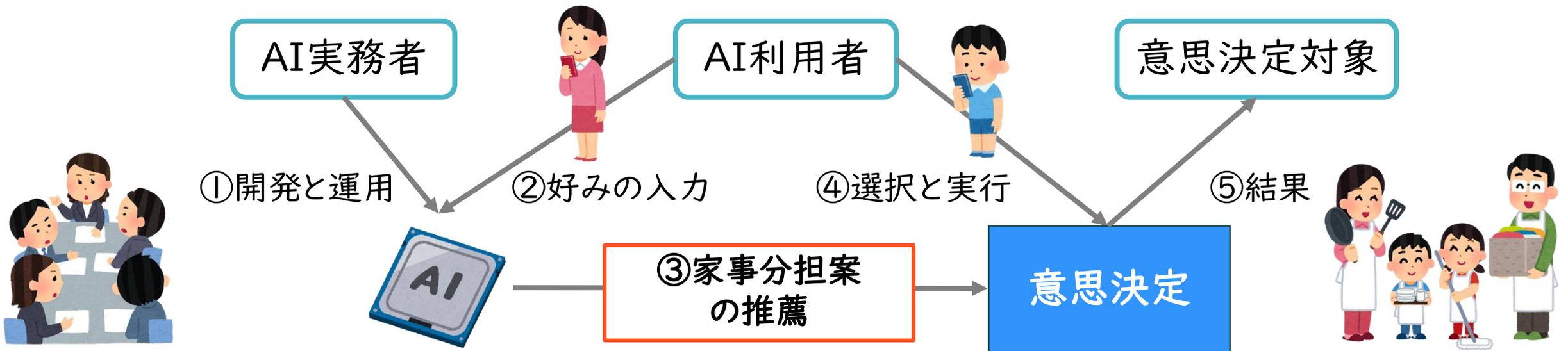
「経済的損失」や「健康被害リスク」



パートナー間の公平性を保つ家事分担の推薦



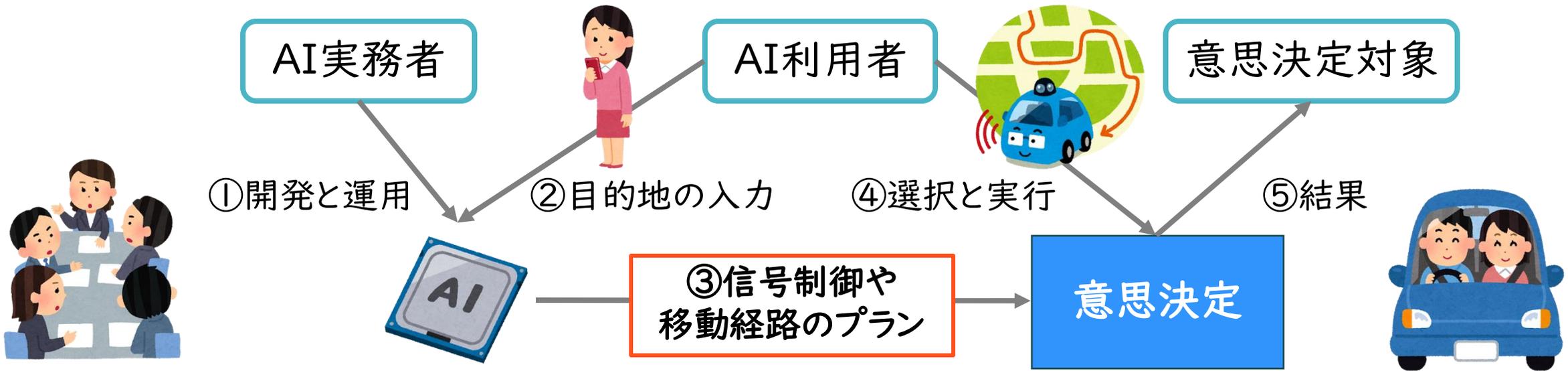
- 数学的な公平性の定義に基づき、妬みのない家事分担案を推薦するアルゴリズムが研究されている
- 利用者の感想から、「**数学的な公平性**」と「**利用者や意思決定対象の感覚**」に乖離が一部生じている可能性が示唆される？



渋滞の予測に基づく移動経路や信号制御の推薦



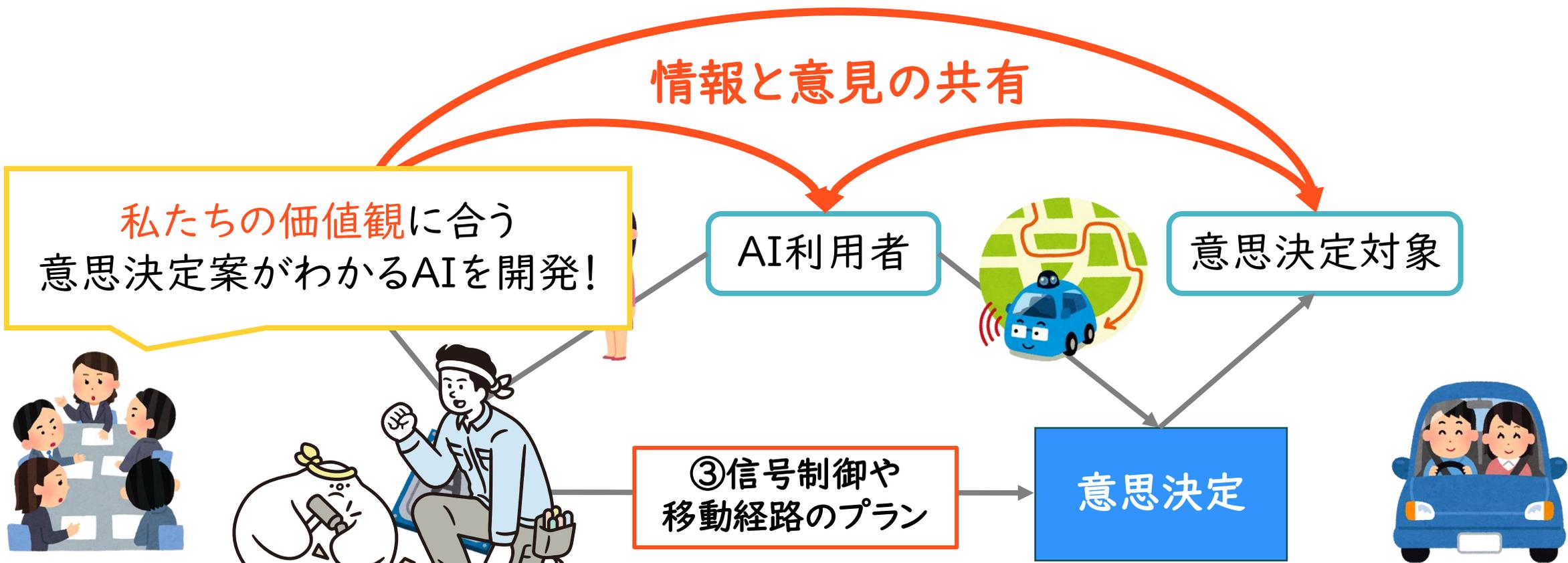
- 目的地を入力すると、今後の渋滞予測に基づき、移動経路や信号制御のプランを推薦する意思決定支援は可能？
- 渋滞を避けるつもりが、結果として逆に渋滞を発生させる場合も
- 関係者間の公平性が信頼の醸成につながると考えられる



関係者の情報共有を促進することが 人間とAIの信頼に繋がる



- 数学的な定義と人間の直感や感覚の整合性（忖度？）をとるAIが必要

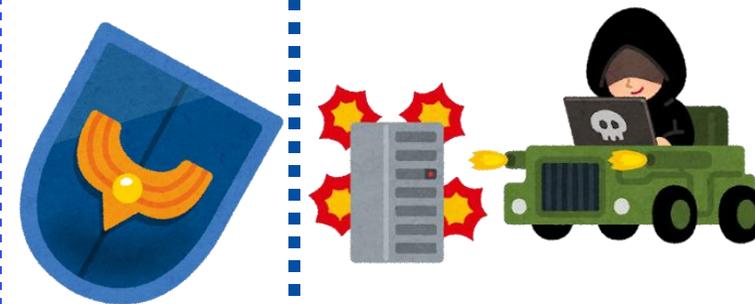
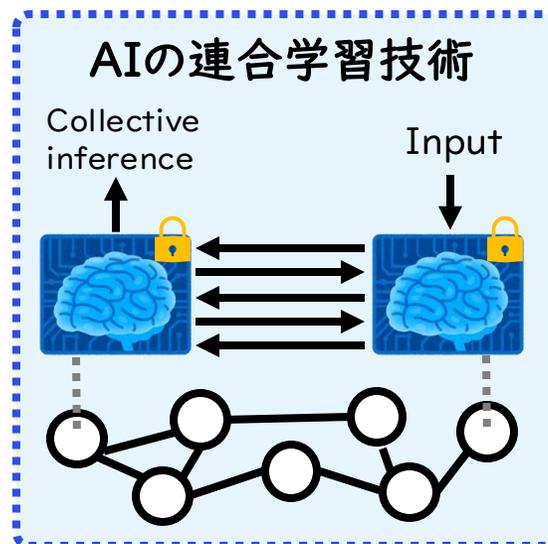
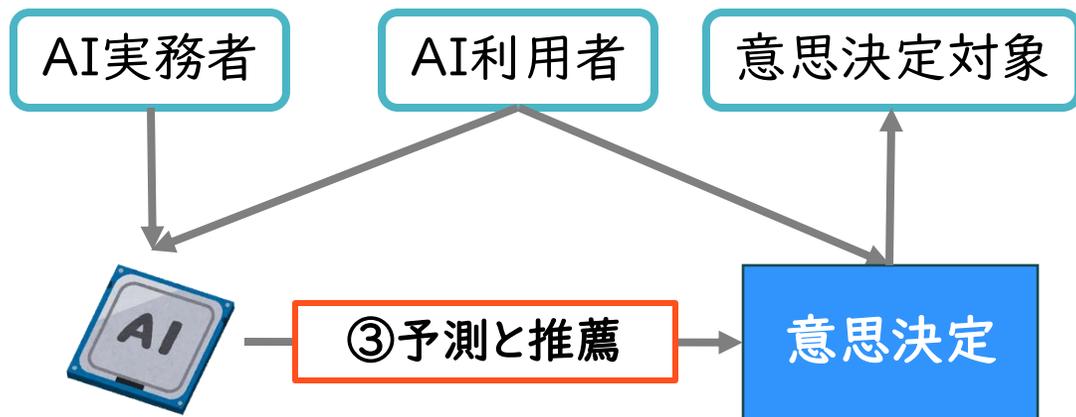


関係者の情報をセキュアに保護するAI運用が必要



- 関係者の意見(情報)は貴重でプライバシー性の高いデータ
- 学習や推論における情報のリークにより信頼性が失われる危険

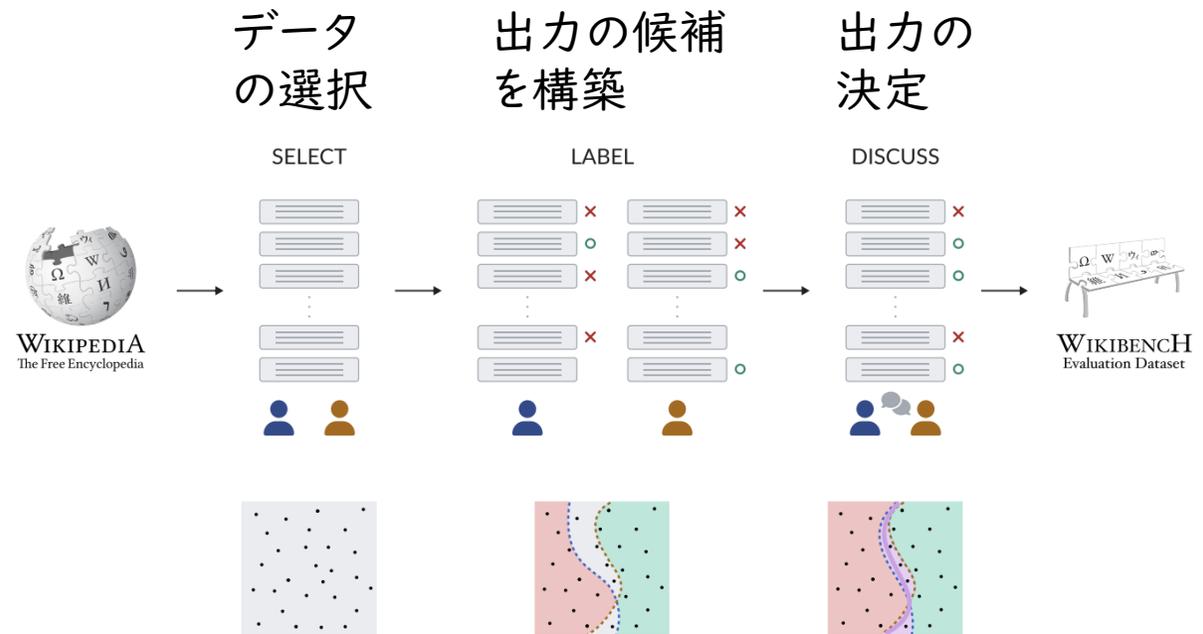
学習・推論・推薦による情報リークを防ぐネットワーク制御



手押し車の開発による価値観の補正



- コミュニティの価値観に合わせた手押し車を効率的に開発
 - 参加者がデータを選択する
 - 参加者ごとに、選択したデータの出力を設計する
 - 参加者間の議論と合意に基づき出力を集約する(手押し車の完成)
- 新たな手押し車でAIを訓練

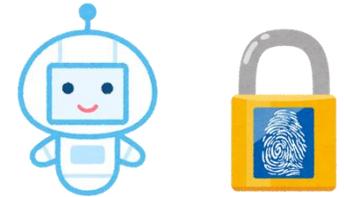


AIによる意思決定支援システムを目指すために

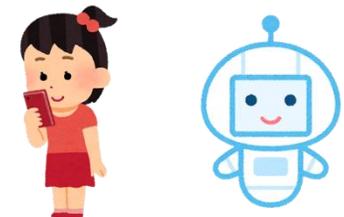


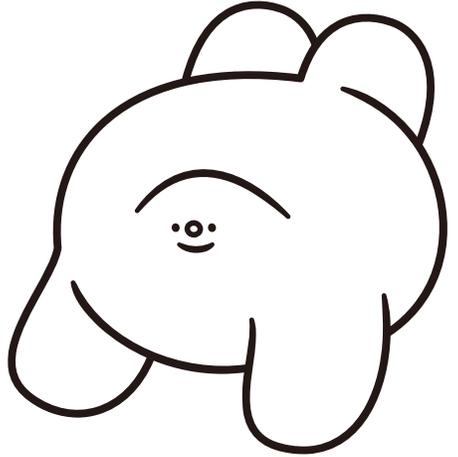
- AIの予測性能（信頼相当性）を改善するだけでは不十分な可能性
 - 巨大言語モデルのハルシネーションも問題ですが…
 - データとモデルの保護、セキュリティ、透明性、解釈可能性
- 社会はどのような意思決定の結果を目指すのか
 - 公平性、平等性、安全性、合理性、堅牢性、防護性
 - 意思決定対象のドメインによって、**リスク評価の基準は異なる**
 - 予測の意思決定の誤りの連鎖によるワーストケースの許容

AIの保護と認証？



人間社会とAIの信頼？





もくもくさん↑

データ駆動社会における AIの信頼性と意思決定の未来

京都大学 大学院情報学研究科 / 理研AIP

竹内 孝

2025年1月22日

2024年度TFC×TEL協働プログラム

公開シンポジウム

