

第6回情報科学系セミナー

2019年12月4日

自然科学研究の道具としての機械学習

瀧川 一学 (たきがわ・いちがく)

ichigaku.takigawa@riken.jp

- 理化学研究所 革新知能統合研究センター (AIP)
iPS細胞連携医学的リスク回避チーム@京都
- 北海道大学 化学反応創成研究拠点 (WPI-IReDD)



革新知能統合研究センター
Center for Advanced Intelligence Project



自己紹介：瀧川 一学(たきがわ・いちがく)

専門：機械学習・データマイニングとその科学での利活用

「データからの学習」をどう問題解決に活用できるのか？



10年 北大
(1995～2004)

統計的信号処理とパターン認識 (工学研究科)

"劣決定信号源分離のL1ノルム最小解の理論分析"



7年 京大
(2005～2011)

バイオインフォマティクス (化学研究所)

ケモインフォマティクス (薬学研究科)



7年 北大
(2012～2018)

データ駆動科学・離散構造を伴う機械学習
(情報科学研究科)

+ JSTさきがけ: 材料インフォマティクス



?年 理研(京都)
(2019～)

AIPセンター iPS細胞連携医学的リスク回避チーム
(北大 化学反応創成研究拠点とクロアポ)

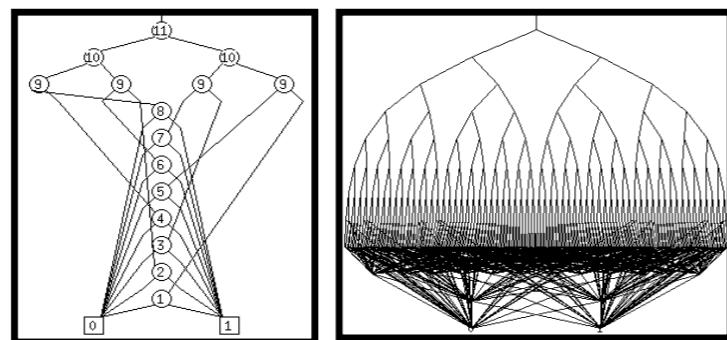


最近の研究対象

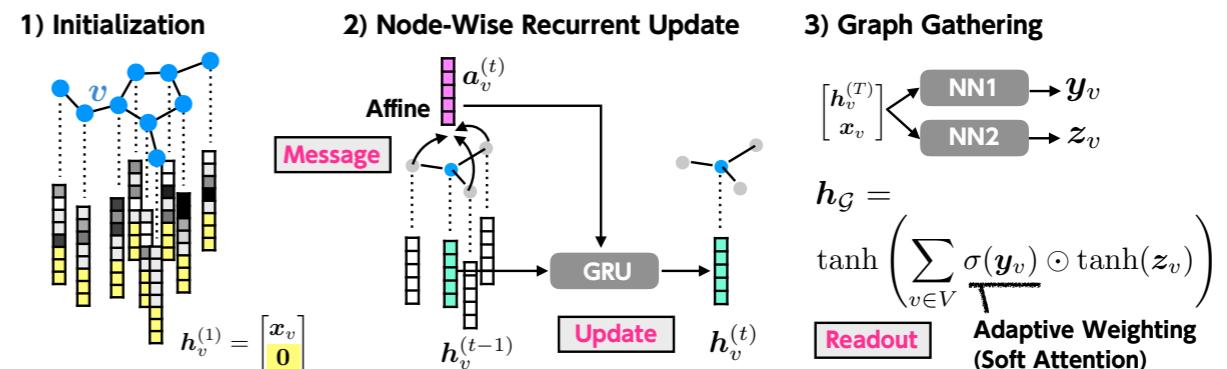
離散構造

—構造と知識—

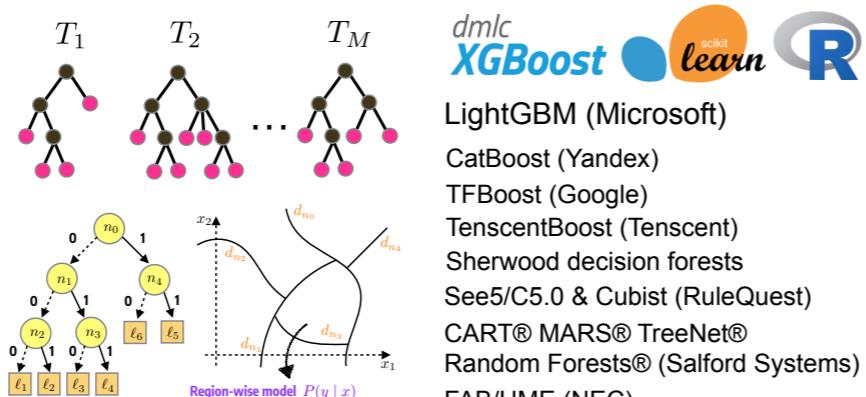
離散構造の表現と構成法



離散構造を入力・制約とする機械学習



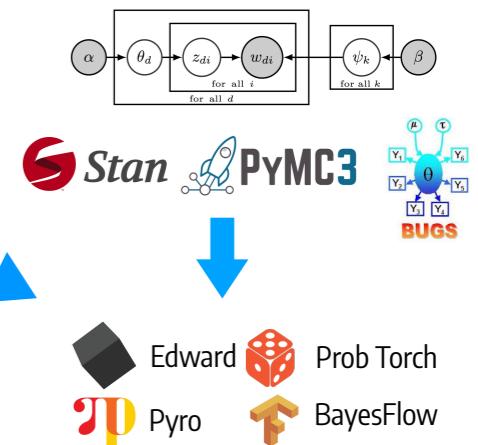
木構造アンサンブル



深層学習/計算グラフ



確率的プログラミング



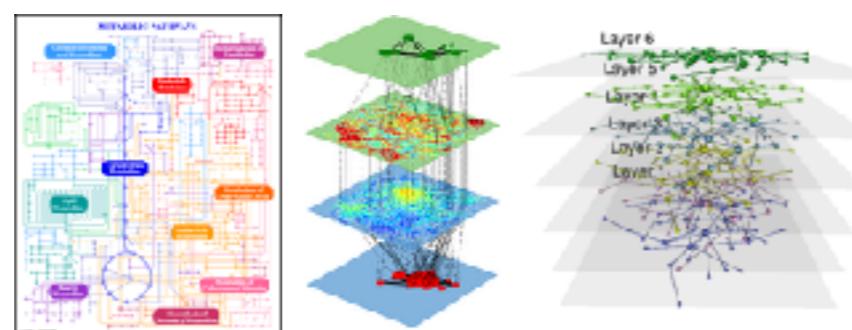
機械学習

—学習と知能—

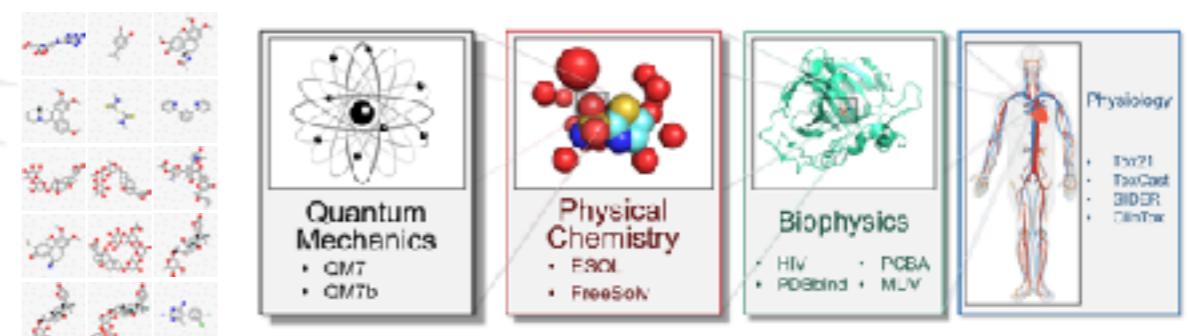
データ
駆動科学

—情報と科学—

生命科学/医・薬・生物



化学(量子化学・触媒化学・生化学・有機化学)

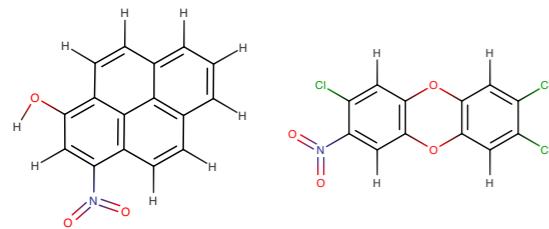


物質・材料科学

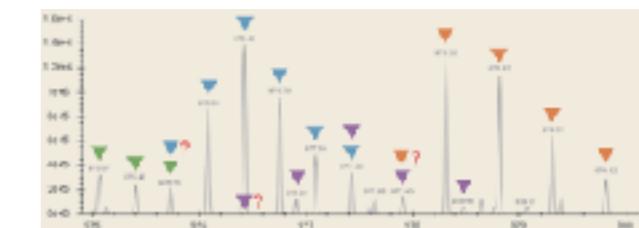
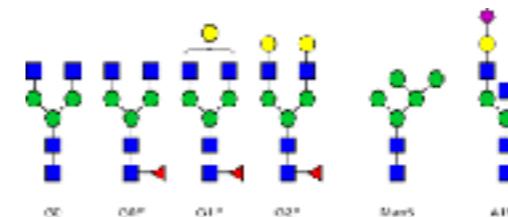
「離散構造」を伴う機械学習

- 対象が「離散構造」を持つ

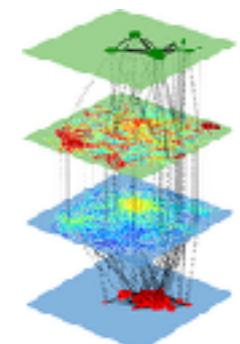
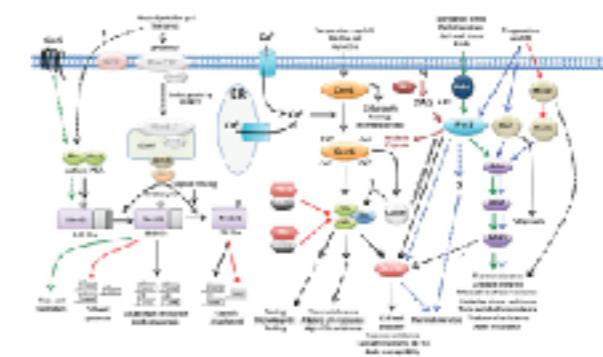
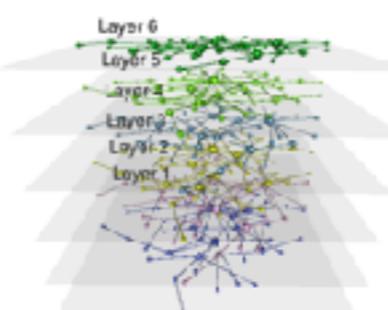
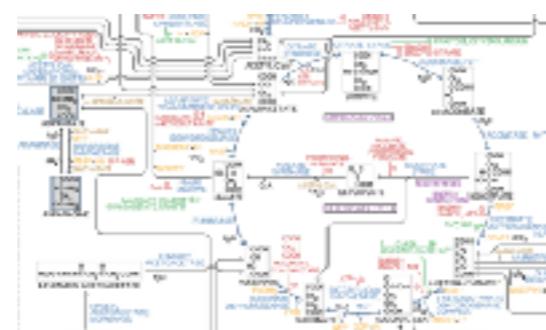
集合、系列、組合せ、置換、分岐(木)、ネットワーク(グラフ)、…



GTATT(146)-TGGATGAAAGATATT-(591)-CTCAC
CTTGCTTCACTGCA (6)-GCGTTCTGGCTTCAAG (2)-AT
TCCGTGACCTCAAG-(15)-ICCCAAAGTCCTGGGATTA
ATGCCAGSACTTTGGGA (16)-GATCACGAGGTCAAGG
AGT(17)-GGCTGAGGCAGGGADAT (17)-GGTGTGAACCGG
AATAAT-(18)-ACTTCCATCTAAAC (137)-TGCTGCTG
CTTGCGCTGCTG-(19)-CTGAAATCCACGACT (15)-
AACGCCAAACT (5)-TTAGCCACGGCTGGTC (16)
CTCCAGCTGGG-(1)-ALAGAGTGAGAUCCCC (52)-
TAACAACATTACAT-(37)-AGCAATTATTTTAAA-
G (2)-TGTAGTCCTGGCTACT (15)-GGAGGATOGCTI

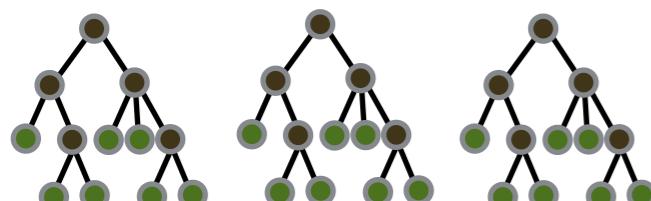


- 対象の関係が「離散構造」を持つ

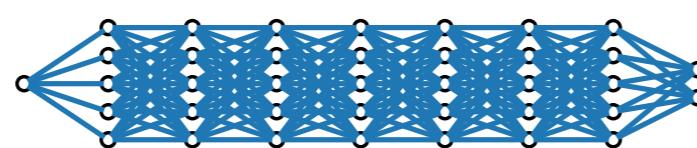


- モデルが「離散構造」を持つ

決定木・決定DAG



ニューラルネット



確率的プログラミング



統計数理研究所 DAT講座 L-B2

【リーディングDAT講座】L-B2 機械学習とデータサイエンスの現代的手法

▶ 申込みに関するQ&A

内 容	<p>● 内容</p> <p>「イントロダクション」伊庭幸人（統計数理研究所） 「カーネル法：基礎から最近の発展まで」福水健次（統計数理研究所） 「行列データ・テンソルデータの機械学習」今泉允聰（統計数理研究所） 「ガウス過程の基礎と応用」持橋大地（統計数理研究所） 「ニューラルネットワーク入門」庄野 逸（電気通信大学） 「決定木に基づくアンサンブル学習」瀧川一学（理化学研究所、北海道大学）</p> <p>※講師プロフィールは こちら</p> <p>● 受講者に期待する予備知識やレベル 簡単な微積分・行列計算・確率の計算の知識は前提とします。</p> <p>● 参考書 福水健次「カーネル法入門—正定値カーネルによるデータ解析—」（朝倉書店） 持橋大地、大羽成征「ガウス過程と機械学習（機械学習プロフェッショナルシリーズ）」（講談社） Carl Edward Rasmussen、Christopher K. I. Williams「Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)」（The MIT Press）</p>
	<p>※やむを得ない事情により、断りなく講師が変更となる可能性があります。</p>
	<p>日 時 2019年12月19日(木)～12月20日(金) 10時～17時（開場9時30分）</p>
	<p>申込受付期間 2019年10月21日(月)10時～11月5日(火)10時 申込みの受付は終了しました。</p>
	<p>定 員 50名（抽選）</p>

統計数理研究所 DAT講座 L-S

【リーディングDAT講座】L-S 決定木とアンサンブル学習の基礎と実践

② 申込みに関するQ&A

④ 内容

現在のデータ社会では多種多様なデータの利活用が求められています。決定木アンサンブルは高速、高精度、非線形で柔軟な機械学習法の一つとしてデータサイエンティストの道具箱に定着してきました。母集団の分布や生成過程を仮定しその未知母数を標本から推定するデータモデリング型の統計的推定と異なり、決定木アンサンブルは極めてアルゴリズム的な手法です。本講義では、決定木学習の多様な背景・歴史、その仕組みと利点・欠点の整理から始めて、決定木を基底とするアンサンブル学習の基礎と実践について、関連手法のライブラリの紹介を含め勘所を1日で講義します。

※2019年度L-B2講座内の「決定木に基づくアンサンブル学習(講師：瀧川一学)」と30%程度重複します。

内 容

④ キーワード

分類木・回帰木・モデル木、プロダクションルールと論理推論、CART、C4.5とM5、バギングとブースティング、ランダム部分空間法、ランダムフォレスト法、ランダム木とExtraTrees、勾配ブースティング法、確率的勾配降下、XGBoostとLightGBM

④ 受講者に期待する予備知識やレベル

簡単な微積分・行列計算・確率の計算の知識は前提とします。

④ 参考書

適宜講義内で紹介します。

講 師

瀧川一学（理化学研究所、北海道大学）

④ 講師プロフィール

離散構造を伴う機械学習・データマイニング、生命科学・量子化学・材料科学でのデータ駆動型研究

※やむを得ない事情により、断りなく講師が変更となる可能性があります。

日 時

2020年3月3日(火)10時～17時（開場9時30分）

転職：2019年4月1日～

北海道大学情報科学研究科の研究室をcloseし下記2組織の
「クロスアポイントメント」へ

- **理化学研究所**

革新知能統合研究センター (AIP)

70%

iPS細胞連携医学的リスク回避チーム 研究員

- **北海道大学**

化学反応創成研究拠点 (WPI-ICReDD) 特任准教授

30%



北海道大学 化学反応創成研究拠点 (WPI-ICReDD)



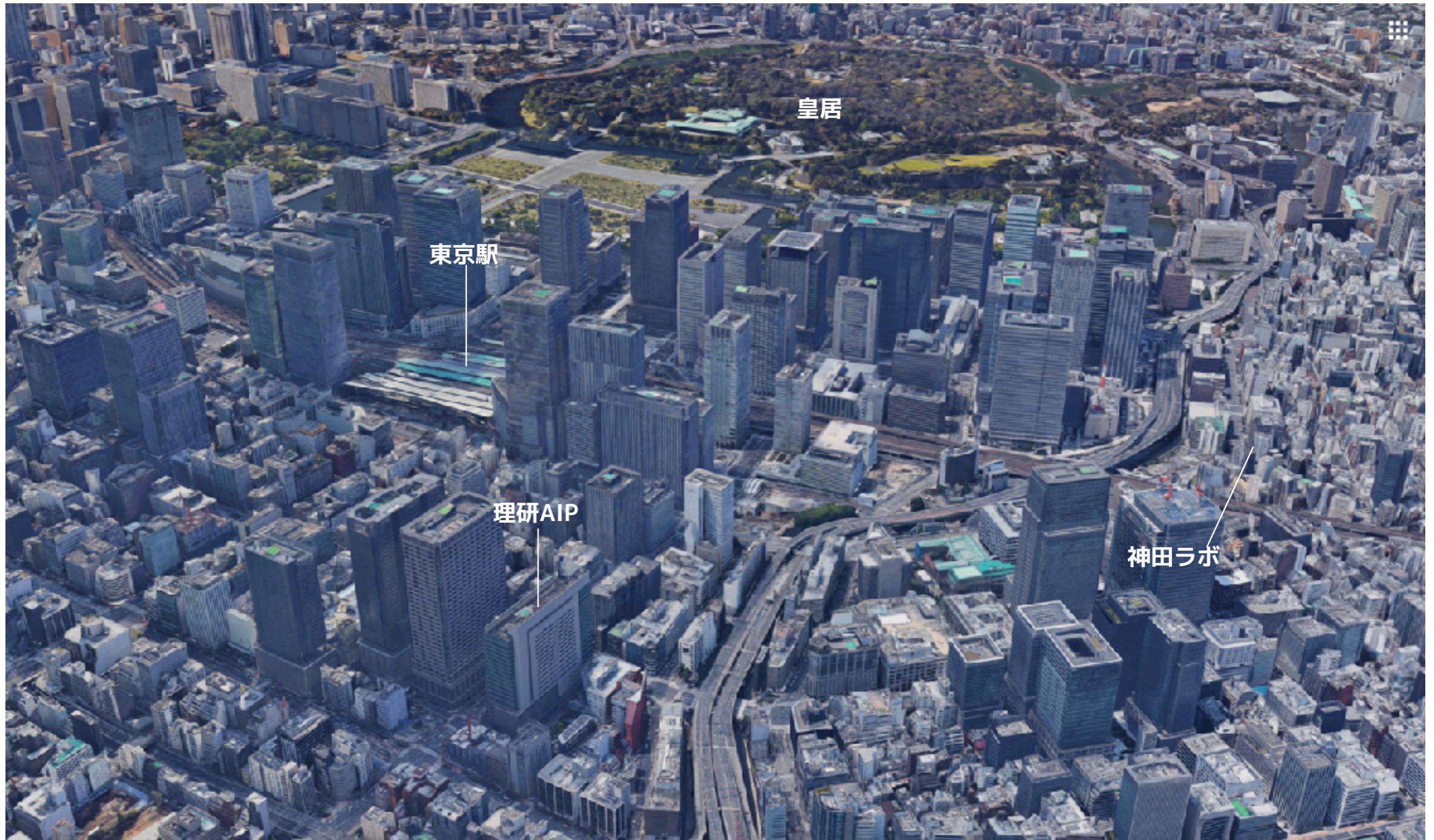
創成研究機構棟 05-116



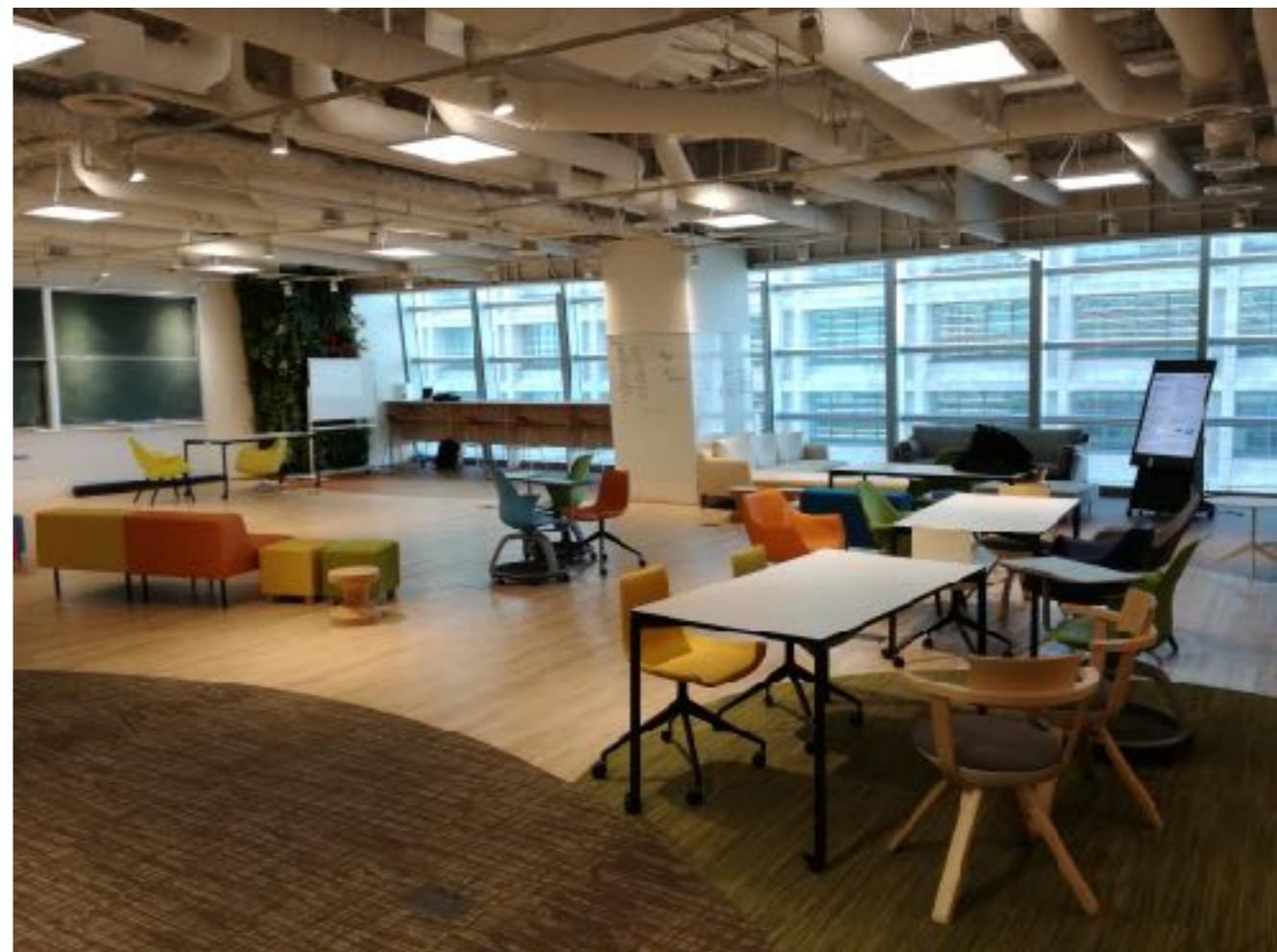
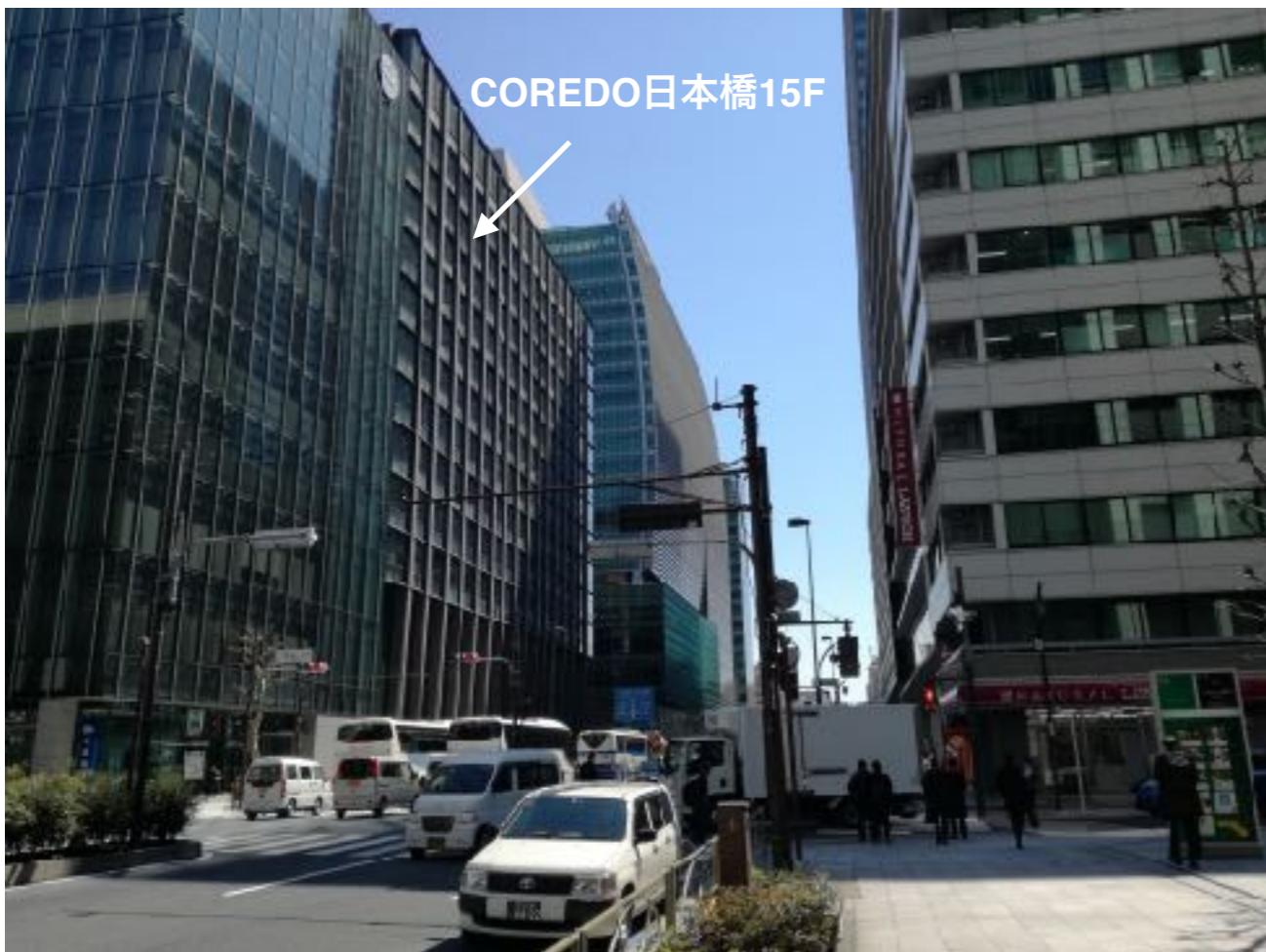
創成研究機構棟 02-106



理化学研究所 革新知能統合研究センター



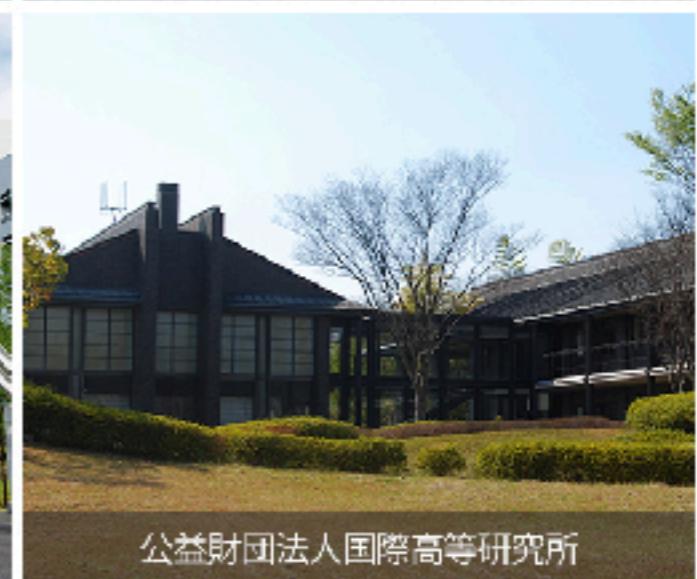
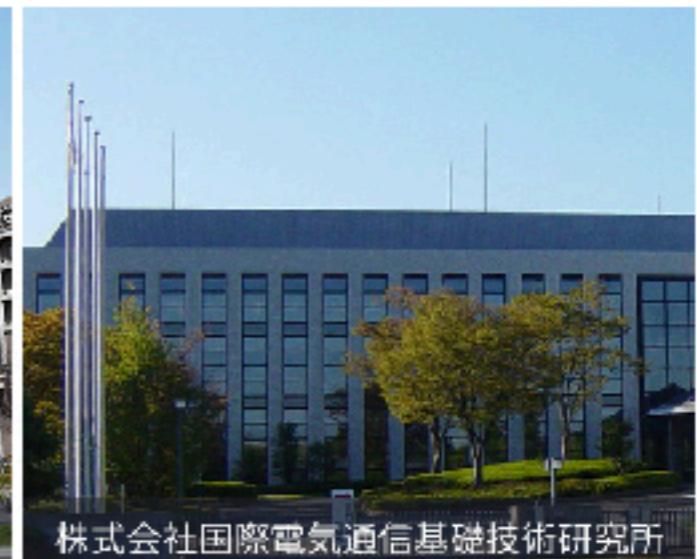
理化学研究所 革新知能統合研究中心



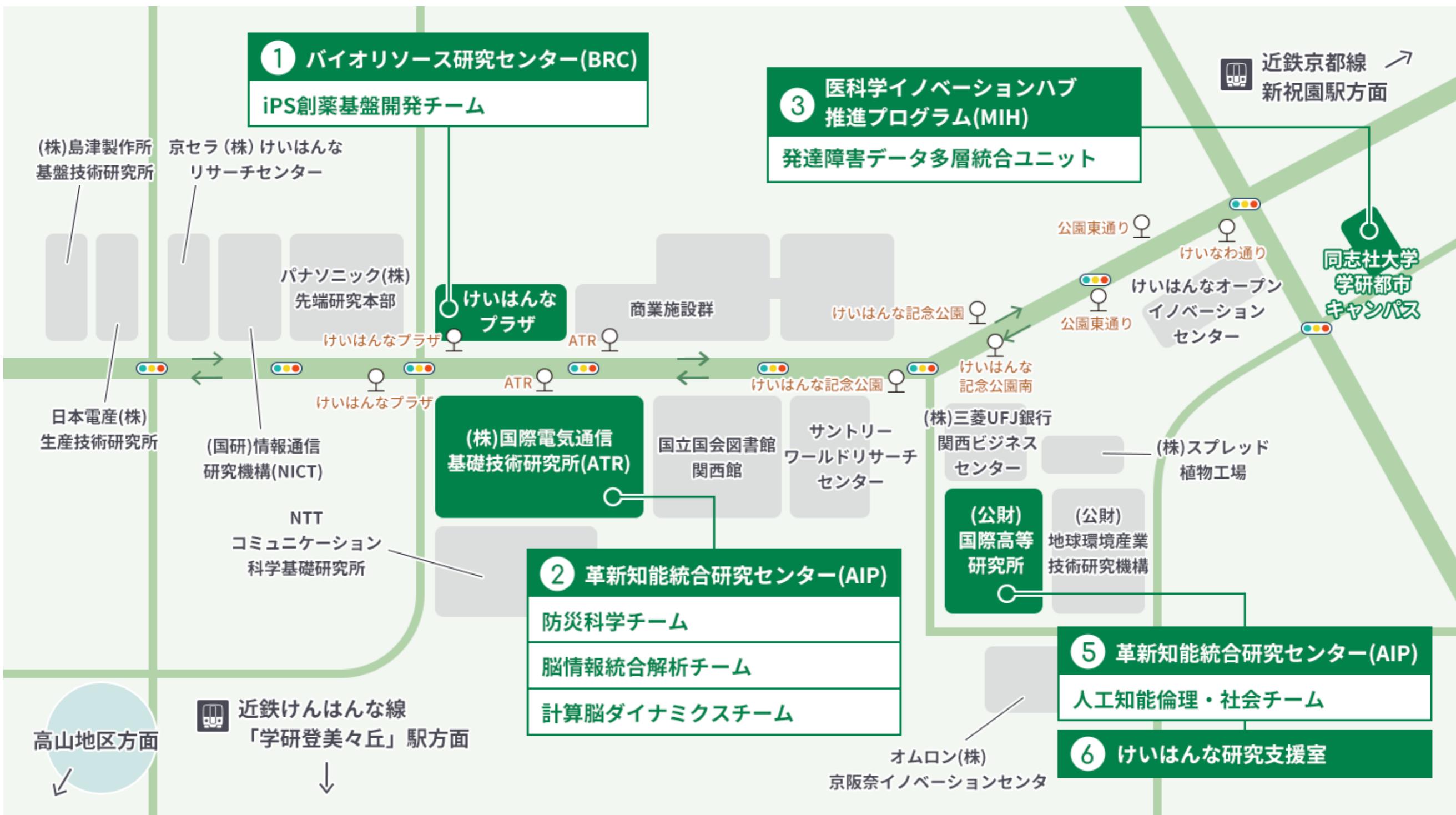
勤務地：京阪奈地区(京都府相楽郡精華町)

<https://www.kobe.riken.jp/about/map/keihanna/>

けいはんな地区



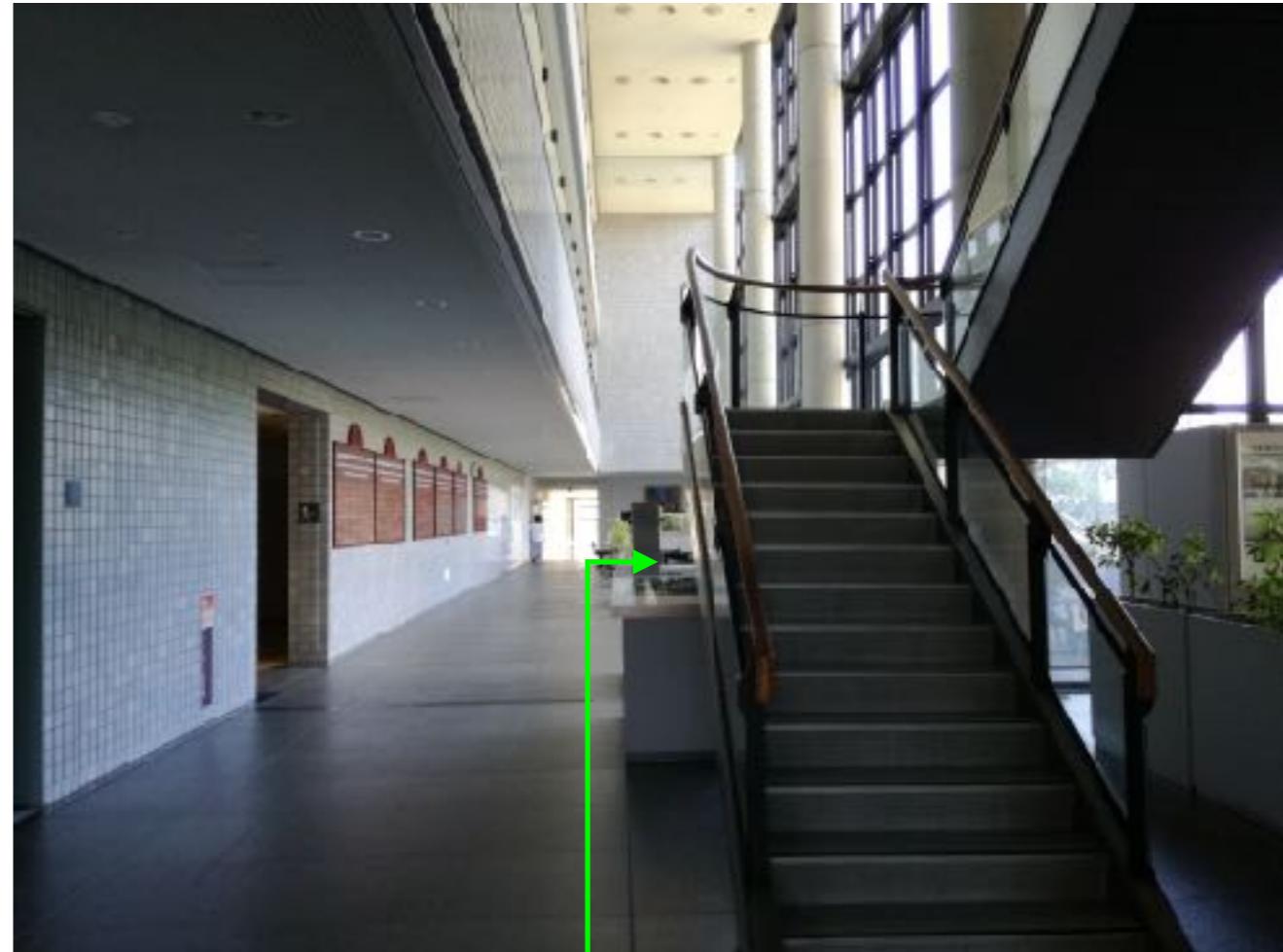
勤務地：京阪奈地区(京都府相楽郡精華町)



勤務地：京阪奈地区(京都府相楽郡精華町)



国際電気通信基礎技術研究所 (ATR)

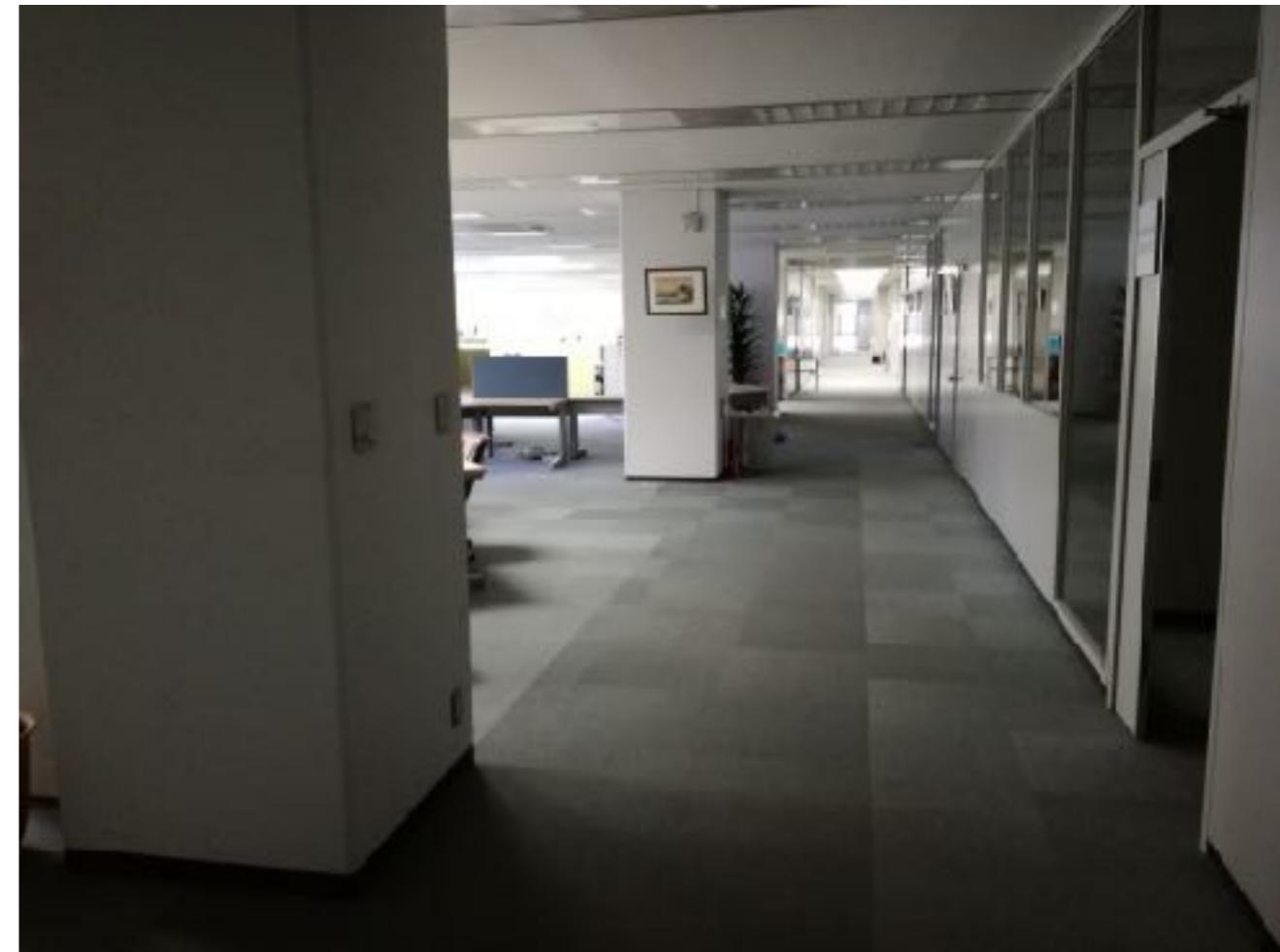


- 脳情報通信総合研究所
- 知能ロボティクス研究所
- 適応コミュニケーション研究所
- 波動工学研究所
- 石黒浩特別研究所 (石黒ERATO)
- 佐藤匠徳特別研究所 (佐藤ERATO)



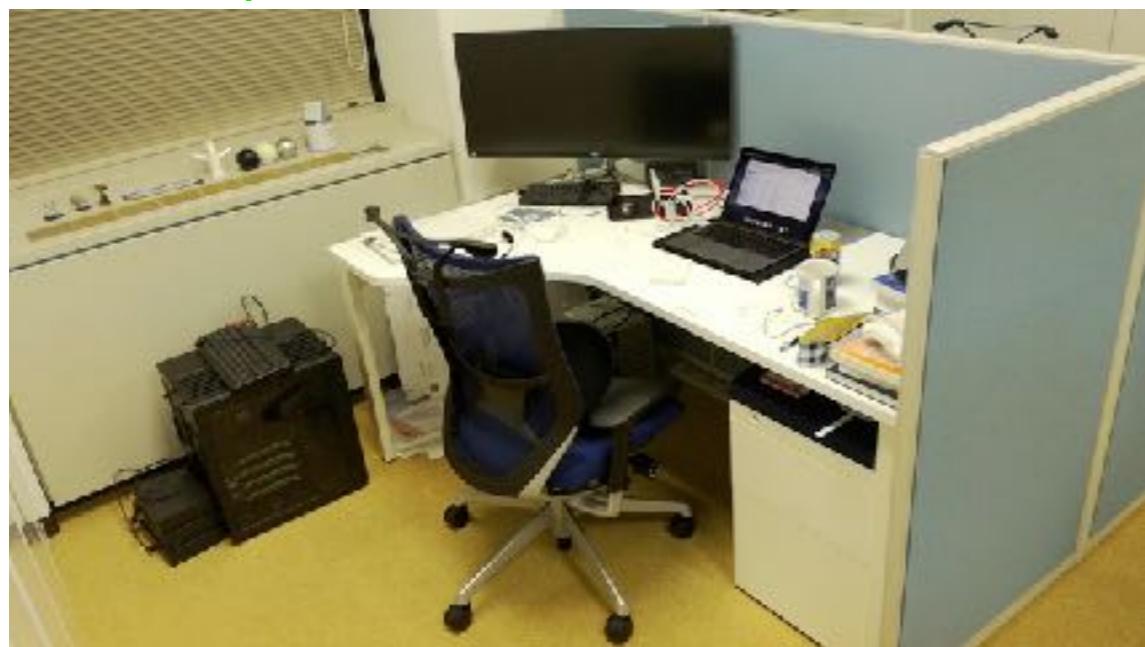
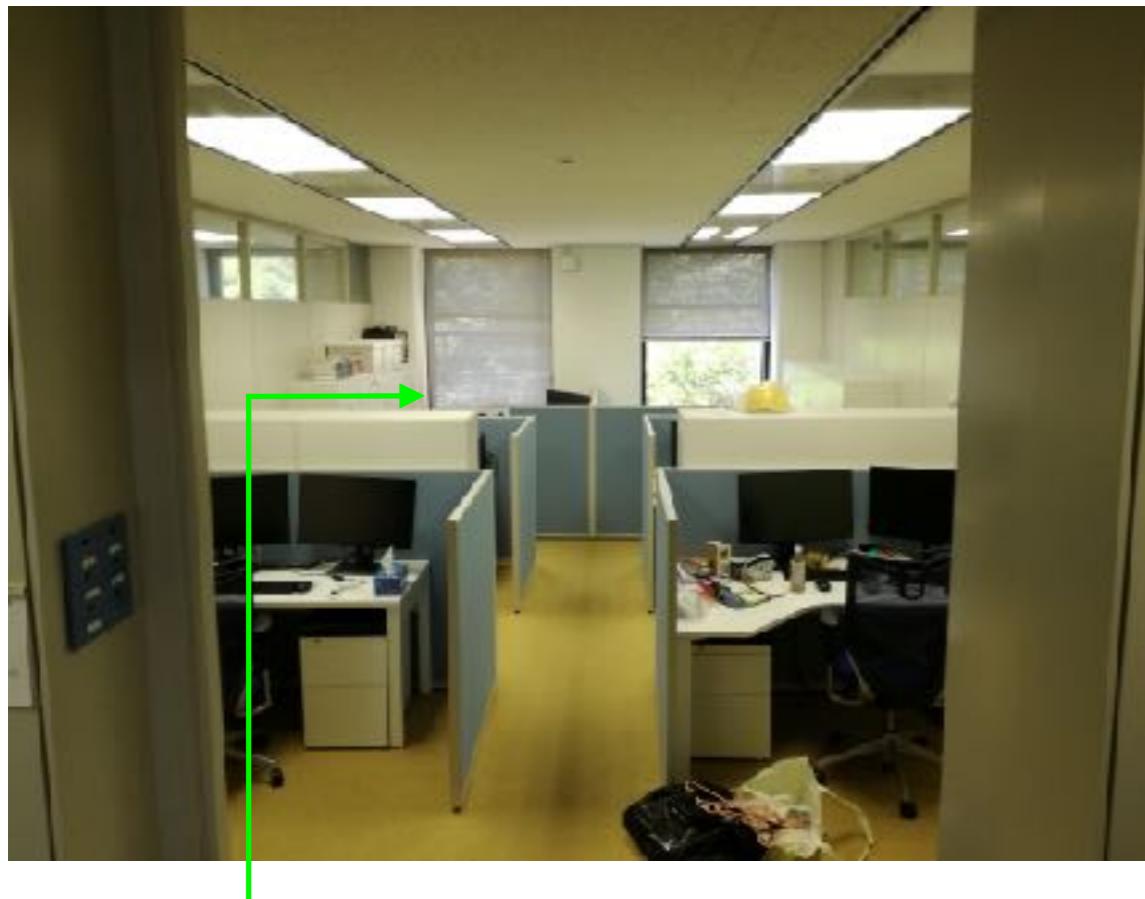
石黒特研の
アンドロイド
「エリカ」様が鎮座
(実物は撮影禁止)

國際電氣通信基礎技術研究所 (ATR)



- 脳情報通信総合研究所
 - 脳情報研究所 (CNS)
 - 認知機構研究所 (CMC)
 - 脳情報解析研究所 (NIA) →
 - 動的脳イメージング研究室 (DBI)
 ≡ 理研AIP 脳情報統合解析チーム (川鍋T)
 - 計算脳イメージング研究室 (CBI)
 ≡ 理研AIP 計算脳ダイナミクスチーム (山下T)

ATR内 理研AIP



- 防災科学チーム (上田 修功)
- 脳情報統合解析チーム (川鍋 一晃)
- 計算脳ダイナミクスチーム (山下 宙人)
- iPS細胞連携医学的リスク回避チーム (上田 修功)



私と上田TL以外は京大CiRAの方々...

自己紹介：瀧川 一学(たきがわ・いちがく)

専門：機械学習・データマイニングとその科学での利活用

「データからの学習」をどう問題解決に活用できるのか？



10年 北大
(1995～2004)

統計的信号処理とパターン認識 (工学研究科)

"劣決定信号源分離のL1ノルム最小解の理論分析"



7年 京大
(2005～2011)

バイオインフォマティクス (化学研究所)

ケモインフォマティクス (薬学研究科)



7年 北大
(2012～2018)

データ駆動科学・離散構造を伴う機械学習
(情報科学研究科)

+ JSTさきがけ: 材料インフォマティクス



?年 理研(京都)
(2019～)

AIPセンター iPS細胞連携医学的リスク回避チーム
(北大 化学反応創成研究拠点とクロアポ)



「データ利活用技術」は科学研究の道具の一つに

Science is changing, the tools of science are changing. And that requires different approaches. —— Erich Bloch, 1925-2016

Nature, 559
pp. 547–555 (2018)

REVIEW

<https://doi.org/10.1038/s41586-018-0337-2>

Machine learning for molecular and materials science

Keith T. Butler¹, Daniel W. Davies², Hugh Cartwright³, Olexandr Isayev^{4*} & Aron Walsh^{5,6*}

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

Science, 361
pp. 360-365 (2018)

SPECIAL SECTION FRONTIERS IN COMPUTATION

REVIEW

Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling¹ and Alán Aspuru-Guzik^{2,3,4*}

The discovery of new materials can bring enormous societal and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for achieving inverse design, which aims to discover tailored materials from the starting point of a particular desired functionality. Recent advances from the rapidly growing field of artificial intelligence, mostly from the subfield of machine learning, have resulted in a fertile exchange of ideas, where approaches to inverse molecular design are being proposed and employed at a rapid pace. Among these, deep generative models have been applied to numerous classes of materials: rational design of prospective drugs, synthetic routes to organic compounds, and optimization of photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

act properties. In practice, approximations are used to lower computational time at the cost of accuracy.

Although theory enjoys enormous progress, now routinely modeling molecules, clusters, and perfect as well as defect-laden periodic solids, the size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience (1). In the context of materials, ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice. Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploiting patterns in massive datasets, these models can distill average and salient features that characterize molecules (2,3).

Inverse design is a component of a more complex materials discovery process. The time

Science, 293
pp. 2051-2055 (2001)

VIEWPOINT Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

Machine learning (ML) (1) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the framework

creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern recognition. A recent example is event analysis

教訓 "low input, high throughput, no output science." (Sydney Brenner)

→ 雜な設定・系で網羅的なハイスループット実験をいくらしても何も得られない

本日の話の前置き

今日は、その後の話を反映した再々編版です。

MI²I・JAIST合同シンポジウム

(情報統合型物質・材料開発イニシアティブ・北陸先端科学技術大学院大学) データ科学における予測と理解の両立を目指して 一分かるとは何か? —

開催日 2018年5月21日（月）13:00～17:40 【終了】

会場 JST東京本部別館1階ホール（東京都千代田区五番町K's五番町）

趣旨

アルファ碁が世界最強棋士に勝利したニュースは大きな話題となった。その根幹である深層学習手法は、非常に高い予測能を有する一方、その選択に関し理解することは不能とされる。古来、科学における最終ゴールは説明・理解できることであった。今、新規物質探索にも最先端の深層学習等のデータ科学手法が利用されつつある。理解は一旦諦め、高い予測能による新たな科学技術の可能性を探るのか、その選択が現実問題となりつつある。最近の研究成果を紹介しつつ、この問題への理解を深めたい。

主催

国立研究開発法人物質・材料研究機構 (NIMS)

Deepな情報系の方には

人工知能学会誌

Vol.34 No.5 (2019/9)

特集：「研究会紹介」

人工知能基本問題研究会 (FPAI)

<https://sig-fpai.org/>

オープンアクセス



ところで、今月の人工知能学会誌の瀧川さんのSIG-FPAI記事、
ただの研究会の紹介記事のハズなのに、えらい見識になっていて、
他の研究会の記事との落差に困惑しました。
みなさまも是非。

kashi_pong 教授 (京都大学)

人工知能学会 人工知能基本問題研究会(SIG-FPAI)

SIG-FPAI

人工知能学会 人工知能基本問題研究会(旧：人工知能基礎論研究会)

Special Interest Group on Fundamental Problems in Artificial Intelligence

運営メンバー

主査	瀧川一学	理研/北海道大学
幹事	井智弘	九州工業大学
	大久保好章。	北海道大学
	杉山磨人	国立情報学研究所
	戸田貴久	電気通信大学
	西野正彬	NTTコミュニケーション科学基礎研究所

○印：主幹事

人工知能学会誌 Vol.34 No.5 (2019/9)

特集：「研究会紹介」 人工知能基本問題研究会 (FPAI)
<https://sig-fpai.org/>

オープンアクセス Permalink : <http://id.nii.ac.jp/1004/00010296/>

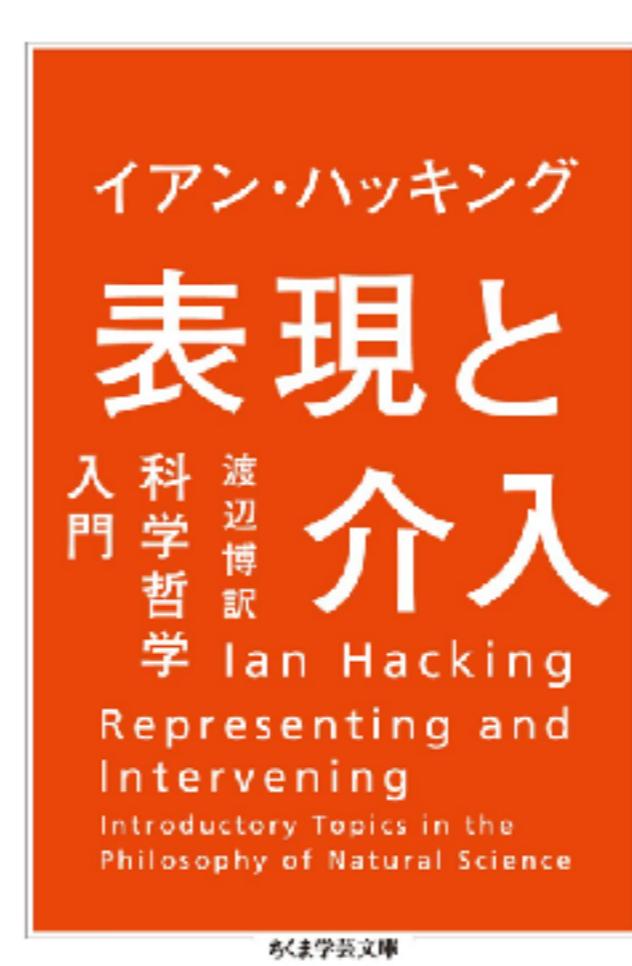
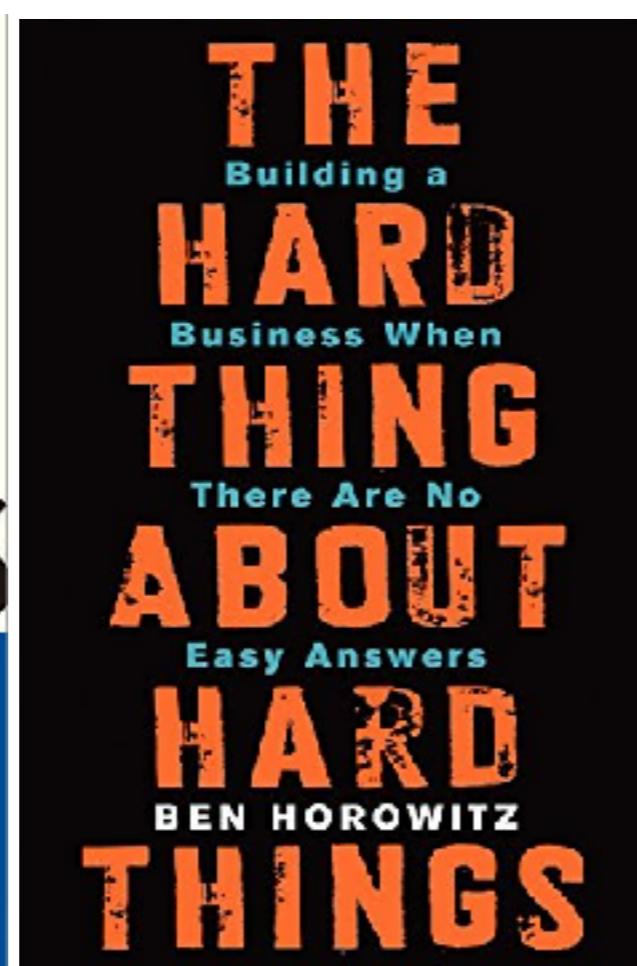
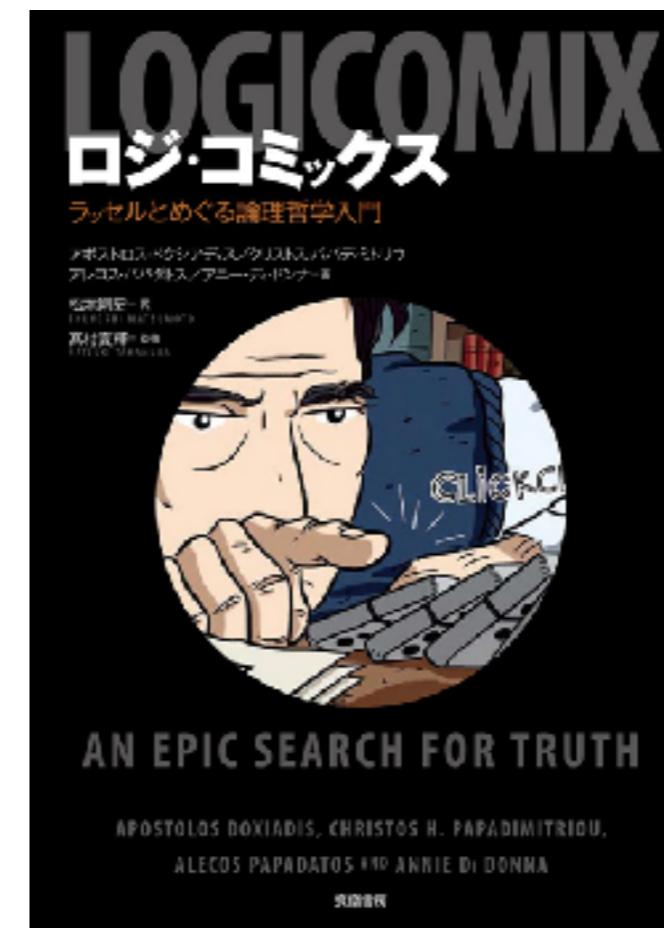
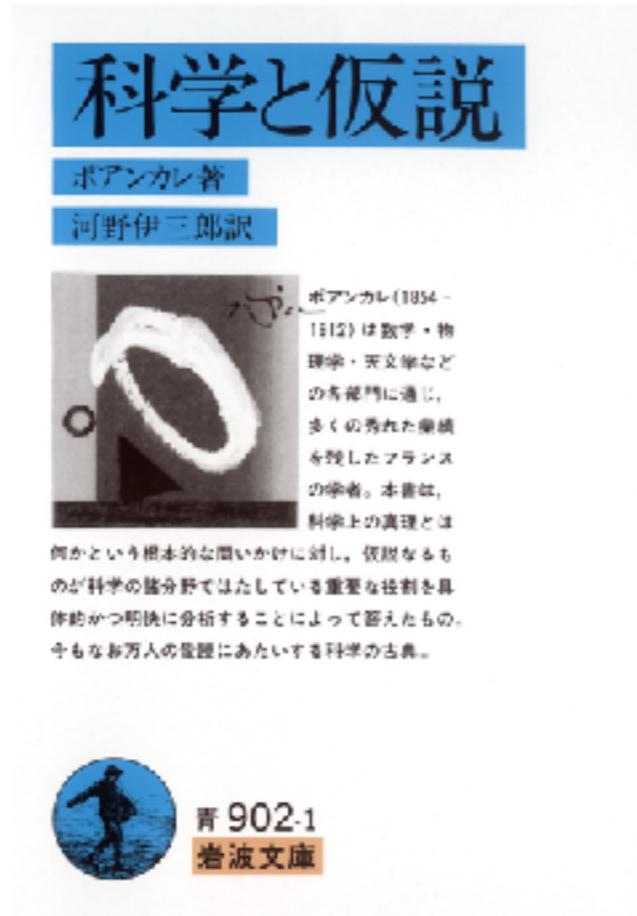
章立て

1. はじめに
2. その上に物の建たないものは基礎とはいわない
3. 変わるもの, 変わらないもの
4. **The Hard Thing about Hard Things**
5. 機械学習と自動プログラミング: 選択と学習の間
6. 組合せの汎化: 離散と連続の間
7. 機械発見と自動化の夢: 学習と発見の間
8. 表現と介入: 経験論と合理論の間
9. 過程と実在: 有限と無限の間
(参考文献100件)

古い文献サーベイ
によるFAI/FPAIの歴史

当時の話題と現代の
話題の私なりのリンク

趣味的雑感と展望



Take Home Message

科学が求めること: 分からないことが分かる(科学的発見)

理解

原因と結果(因果関係)を見出す

$x \rightarrow y$ の過程を理解し(人間が)発見する

発見

今まで見出されていない良い対象を見出す

$x \rightarrow y$ を利用して良い y を持つ x を発見する

今日伝えたいたった3つのこと

1. 単純に機械学習を使うだけでは**いざれも解けない**
2. **機械学習以外のもの**(介入やドメイン知識)が原理上必須
3. 最近**まさに研究が進行中**の未解決領域だが研究は色々ある

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

Empirical optimization or "Edisonian empiricism"



問題：時間とコストは有限！！
理論的に可能なあらゆる候補を
この方式で検証することは不可能



次の実験計画へfeedback

既知の知見・
観測(データ)

仮説形成

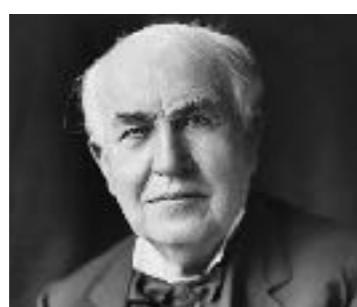
- 実験
- シミュレーション

仮説検証

結果の確認と
検証

"観察と帰納 (empirical/inductive)"

"論理と演繹 (rational/deductive)"



Thomas Edison先生

- Genius is 1% inspiration and 99% perspiration.
- There is no substitute for hard work.
- I have not failed. I've just found 10,000 ways that won't work.
- :

よく考えるとブラックなことしか言ってない！

科学的発見とセレンディピティ

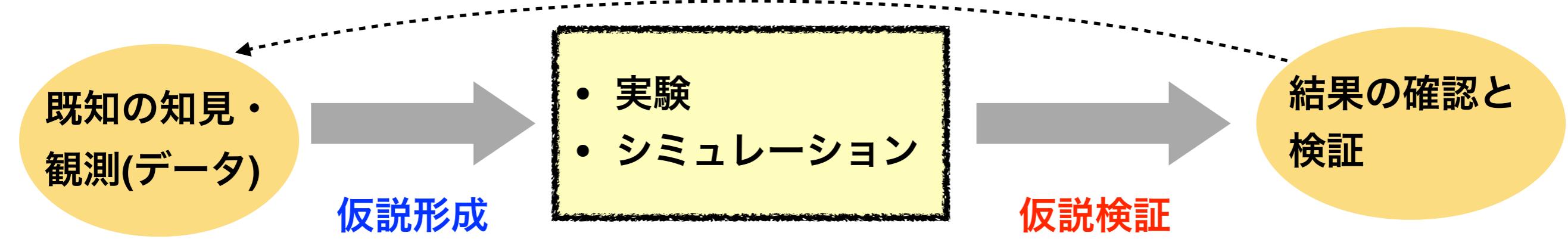


問題：時間とコストは有限！！

理論的に可能なあらゆる候補を
この方式で検証することは不可能



次の実験計画へfeedback



- それゆえ「研究者のセンス・腕の見せ所」 + 「幸運(セレンディピティ)」に依存する筋の良さそうな候補を選ぶ、今まで試されてない全く新しいやり方を思いつく、etc
- 候補が**あまりに膨大(実質ほぼ無限)**なので(数多く試すのは有利だとは言え...)必ずしも「力技とお金と人海戦術で数多く試した者が勝つ」とは限らない

機械学習はシミュレーション・実験と相補的



次の実験計画へfeedback

既知の知見・
観測(データ)

高速・高精度な
Data-Driven予測

結果の確認と
検証

仮説形成

(機械学習・データマイニング)

- どういう実験・シミュレーションを次に行うかの計画立案
- 時間のかかる計算の高精度高速近似
- 曖昧な因子や実験条件の最適化
- Multilevelの情報統合

仮説検証

(シミュレーション+実験)

- 再現性を担保する高精度・高速実験系
- 仮想化検証が可能な因子のシミュレーション(計算科学)による探索
→ 望ましい対象のさらなる絞り込み

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

「機械学習」的シチュエーション

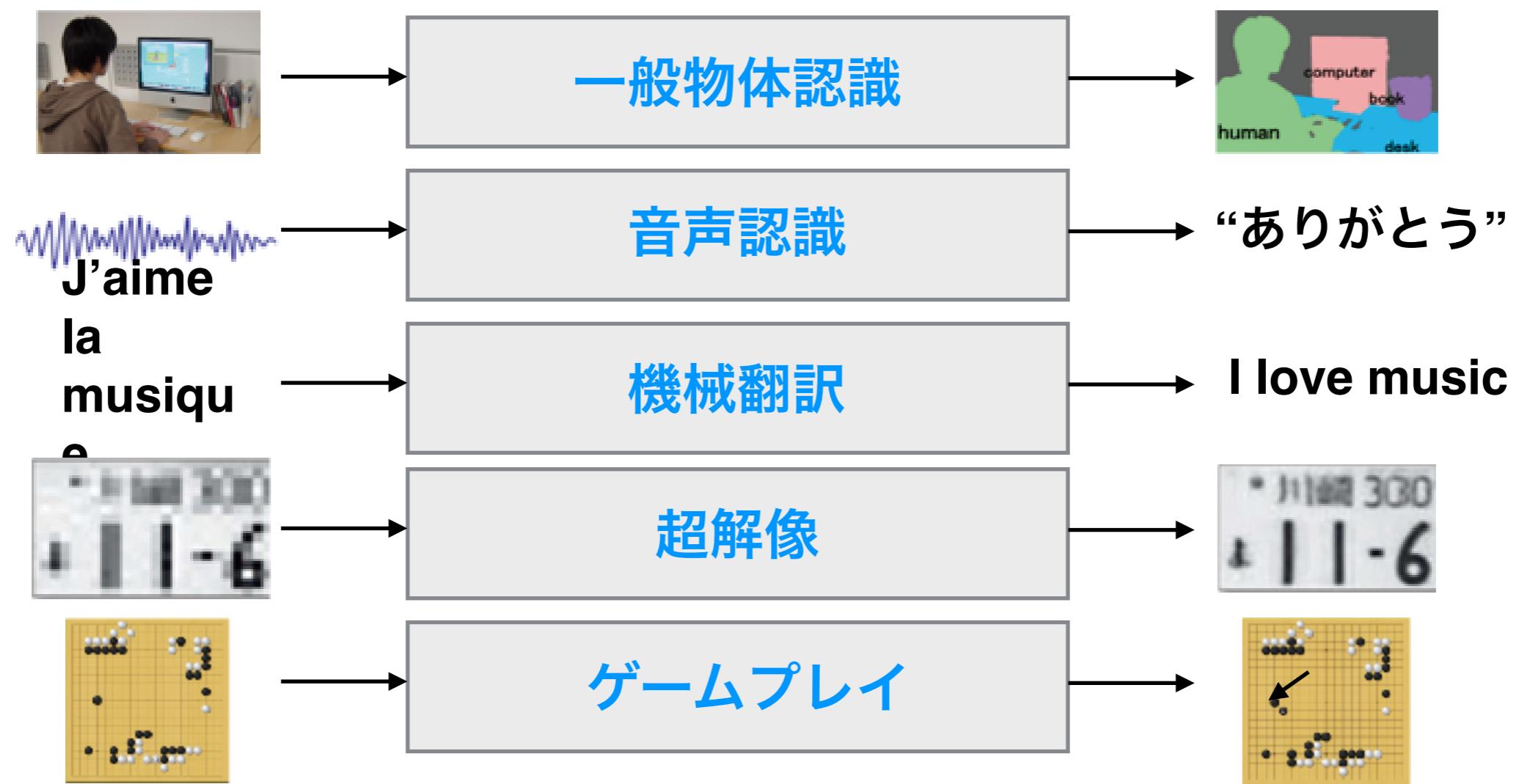
事例：写真をAさんかBさんかに分類するコンピュータプログラムを作りたい。(大量の写真を人手分類するの嫌)



- 人間は簡単にできる
- が、どうやってやっているのか原理は不明確
- 髮型、角度、照明、背景、表情、化粧、年齢、などを考えると明示的なプログラミングはとても難しい

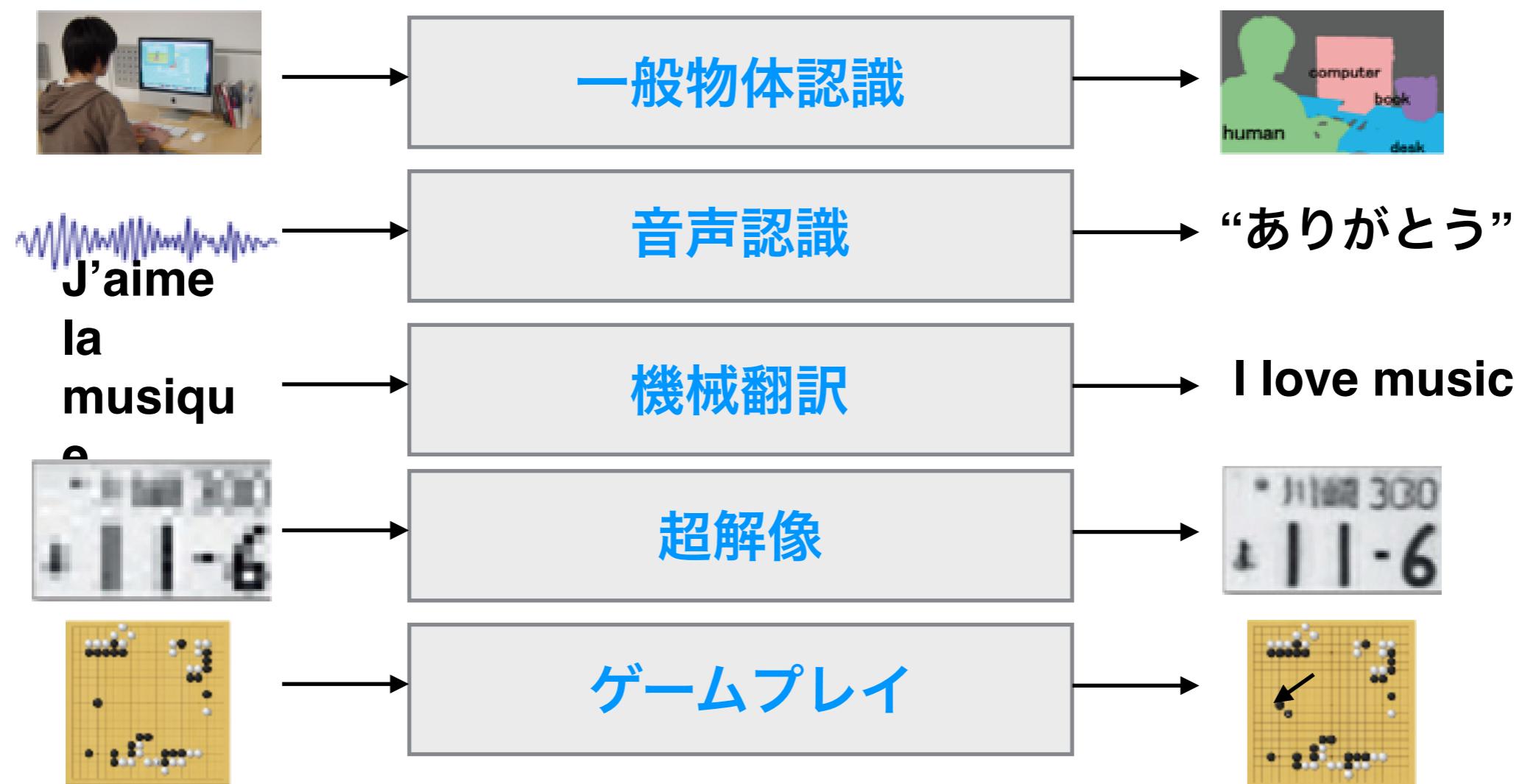
機械学習: 新しいプログラミングのかたち

入出力の関係がよく分からない変換過程(関数)を大量の入出力の見本例から明示的にプログラミングすることなく構成する技法



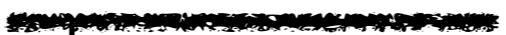
機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



「理解」編のポイント！

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る

1はOKだが
2はとても微妙！


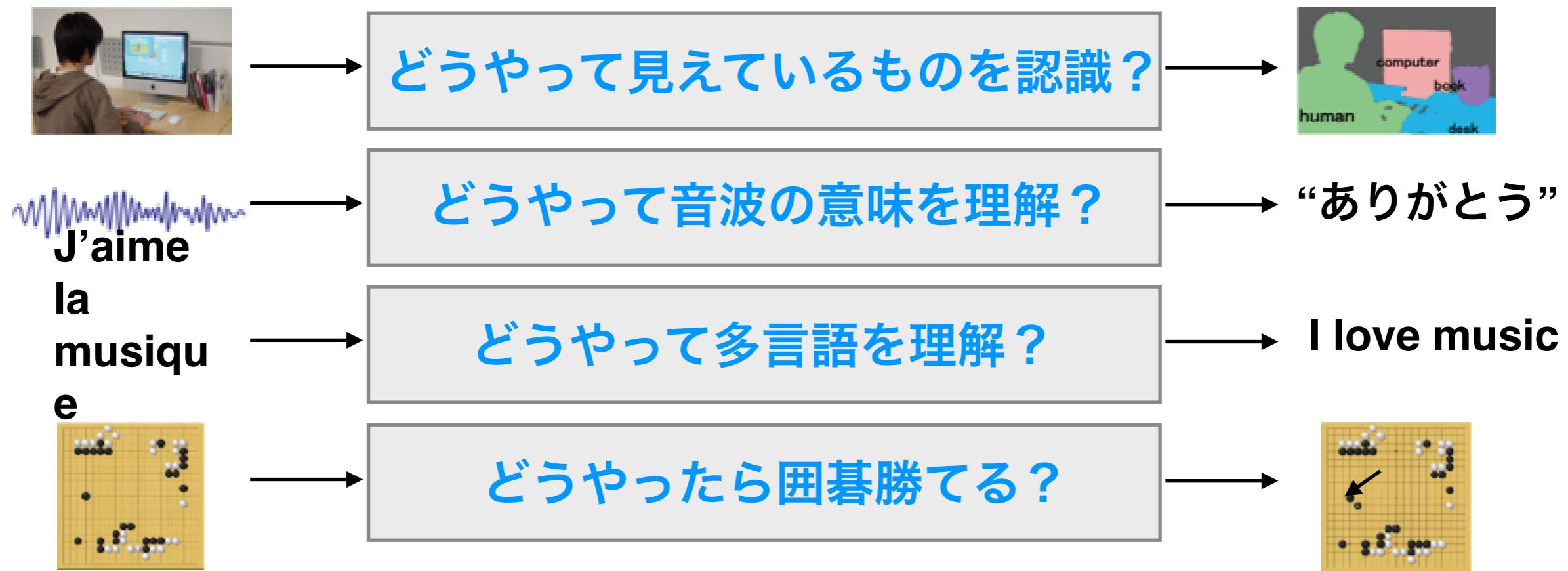


そしてその理由もまた微妙なのですが
さしあたって...

機械学習予測とその理解の「深い溝」

下記はどれも機械学習でかなり高精度な予測ができますが、
果たしてその仕組みの理解が得られたのでしょうか...?

実情：予測は当たるんだけど理由はよくわからない！



予測があれば理由はわからなくともOK？

どう考えてもOKなわけないやろ... 😭 と思うかもしれません
が

- それは用途による (予測が高い精度で当たるという前提で)

現在の商業的成功を牽引する多くの用途では要らない場合も。
検索、広告、推薦、センサー/IoT、画像・音声認識、芸術など

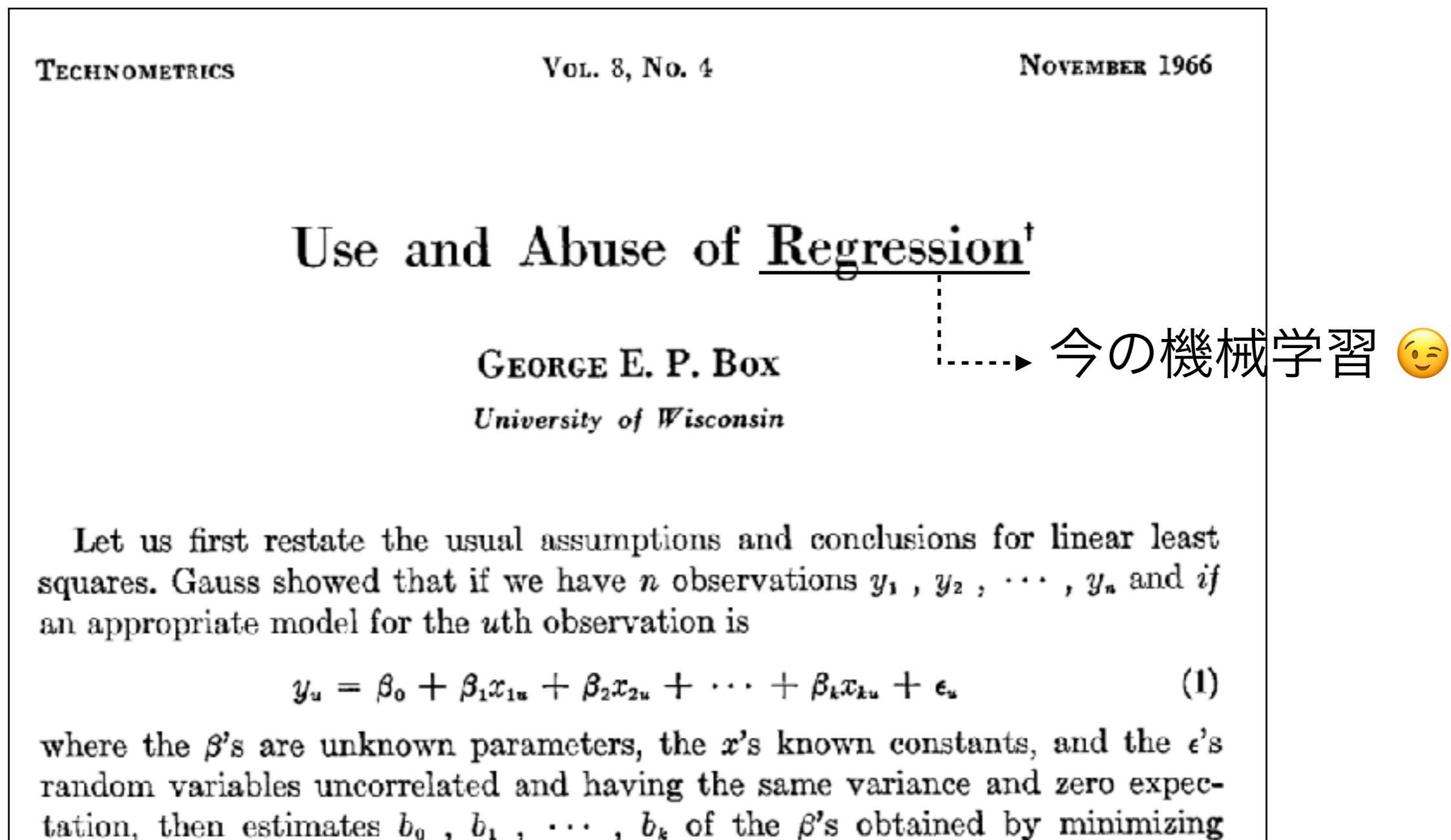
- 最近、適用先が広がり 「理由」 を求められるように

社会的にクリティカルな目的に投入するなら 「理由」 必要。
医療、自動制御、インフラ制御、雇用、政策決定、融資など

└→ 説明責任・透明性・公平性・安全性・倫理を担保可能？

論文：Use and Abuse of Regression (1966)

コンピュータが初めて卑近な道具になり、人々はあらゆる目的に
Data-driven(回帰分析)を多用するようになってしまった... 😅



論文：Use and Abuse of Regression (1966)



大統計学者 George E. P. Box (1919-2013)

"one of the great statistical minds of the 20th century"

**"Essentially, all models are wrong,
but some are useful"**

https://en.wikipedia.org/wiki/All_models_are_wrong

回帰分析の目的: (教師付き学習一般に当てはまる) ...苦言? 😅

1. 説明変数を観測したときの目的変数の予測 → **Use** 😊

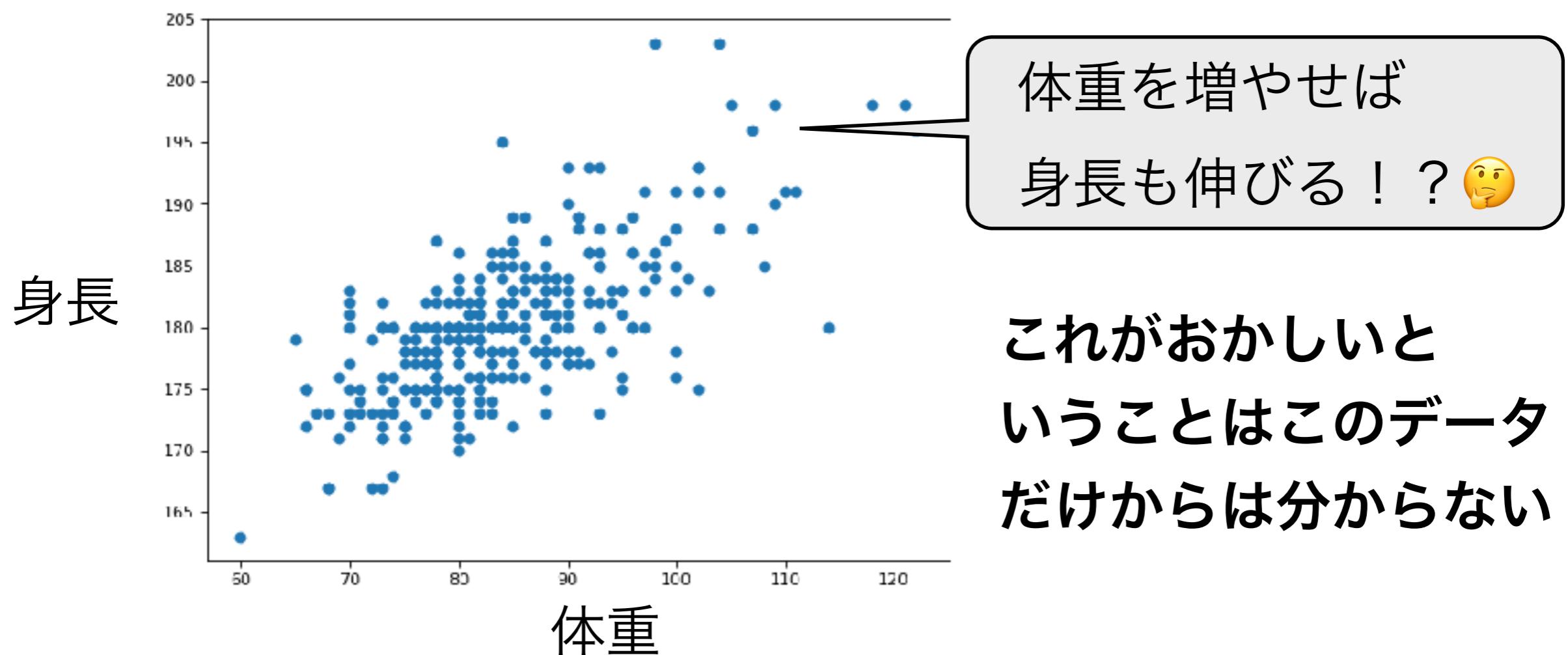
2. 説明変数に外的操縦を加えたときの
目的変数への因果的効果の発見 → **Abuse** 😣

相関関係は必ずしも因果関係を意味しない

応用統計学のイロハ : Correlation does not imply causation

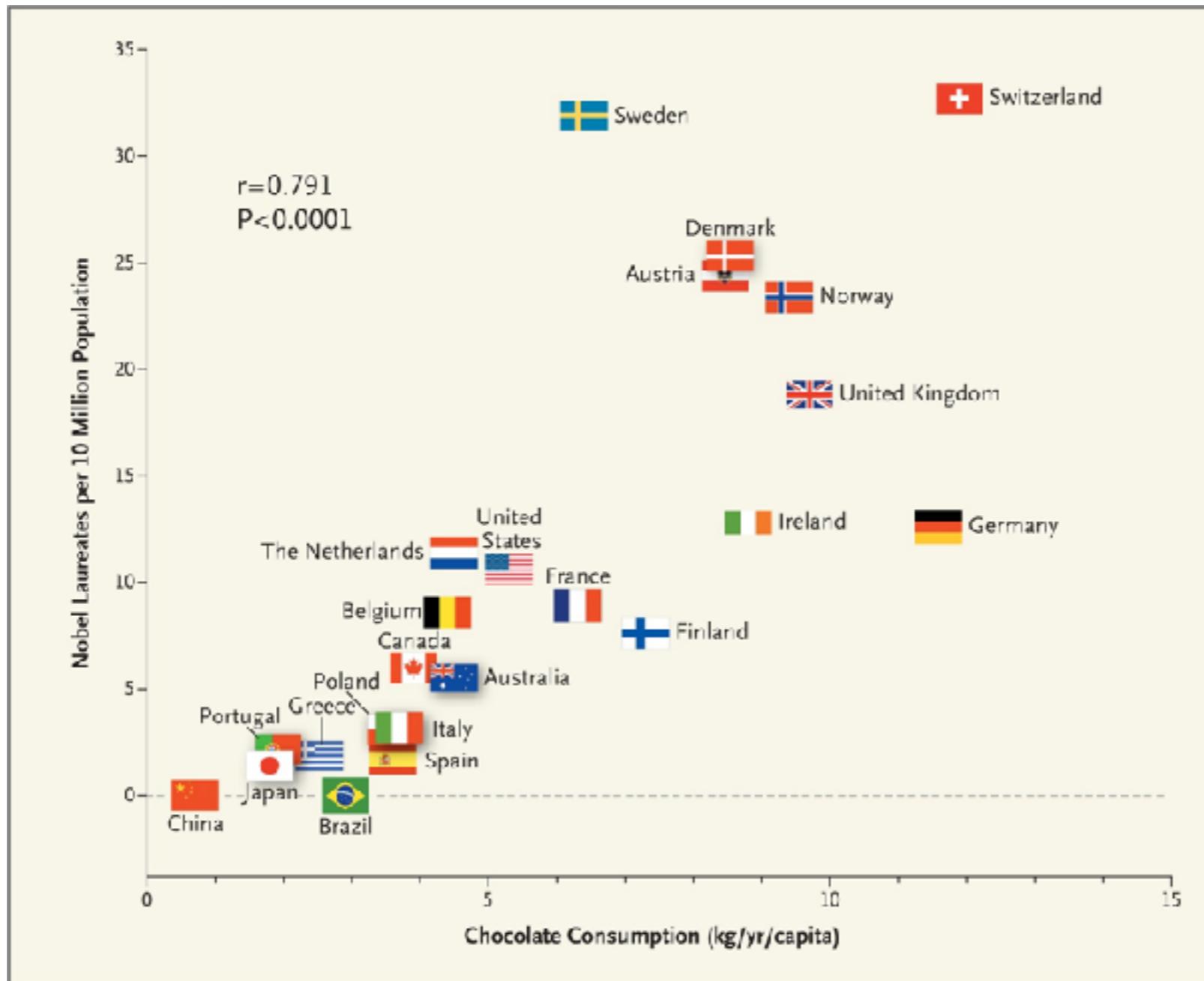
→ そして機械学習はデータに内在する相関関係の利活用技術

日本プロ野球開幕一軍選手の身長・体重データ
(2016年球団公式サイト選手データより自作)



人口1千万人あたりの
ノーベル賞受賞者数

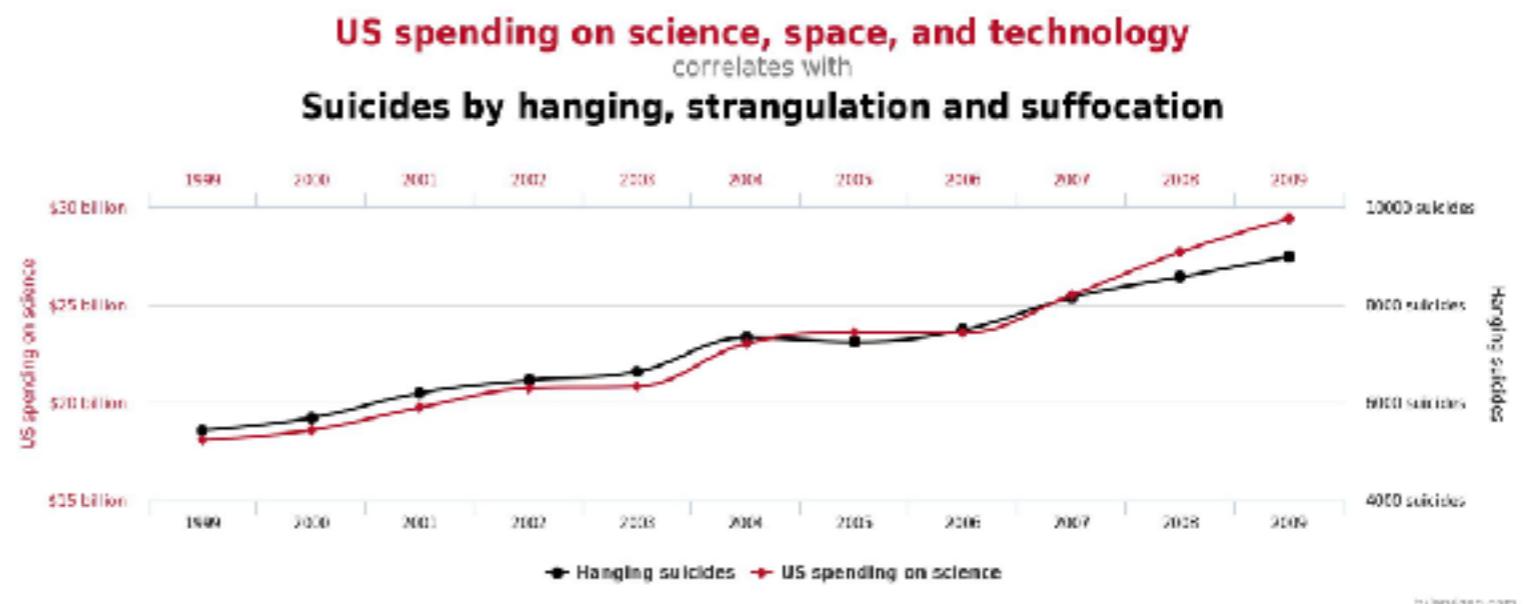
→ IF(2018) 70.670 😳 の最も歴史と権威のある医学誌



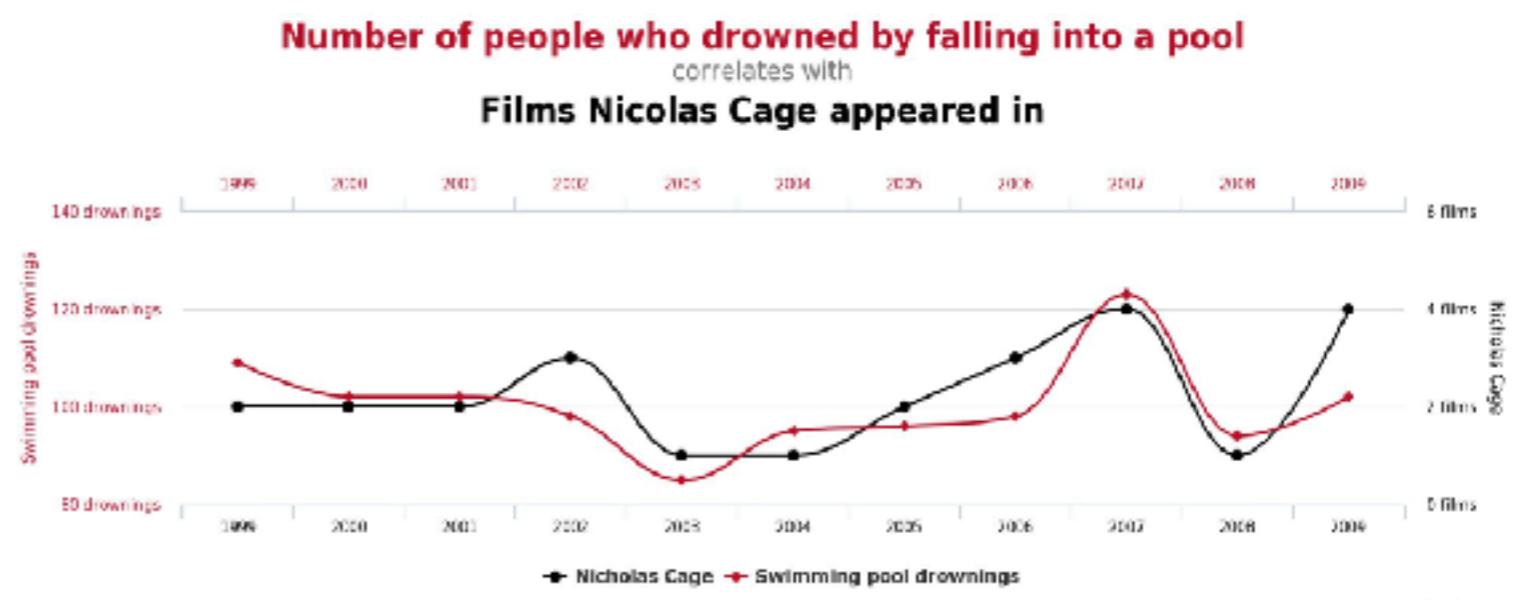
チョコレート消費量

観察データだけからは因果は分からぬ

アメリカの科学予算 vs
首吊りによる自殺者数



プールでの溺死者数 vs
ニコラスケイジの映画出演数



<http://phenomena.nationalgeographic.com/2015/09/11/nick-cage-movies-vs-drownings-and-more-strange-but-spurious-correlations/>

いつ相関と因果は乖離しうるのか？

因果関係判定のHillのガイドライン (Hill, 1965)

相関に
ついて

それ以外
(Check難)

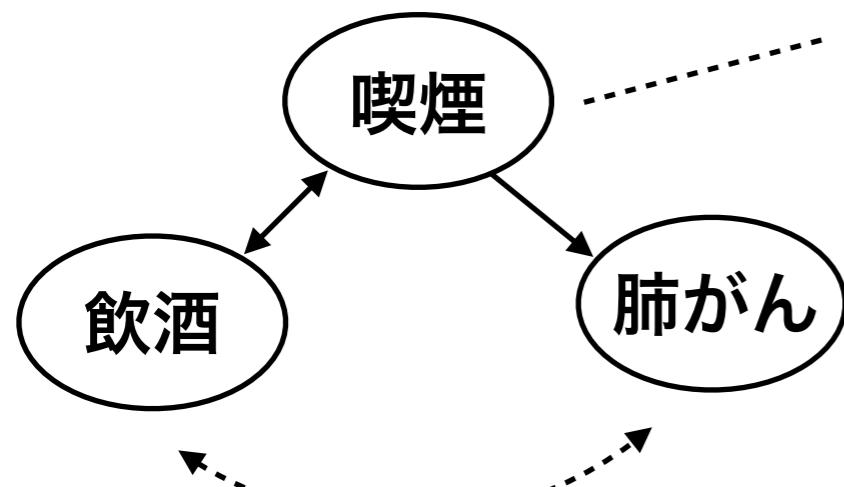
事象 A が事象 B の原因であると結論づけるためには以下の基準に適合しているかを確認することを推奨する。

- | | |
|--------------|---|
| 1. 相関関係の強さ | A の生起と B の生起の間に強い相関関係がある。 |
| 2. 相関関係の一致性 | 相関関係の大きさは様々な状況で、対象や実証に利用する手法が違っても一致している。 |
| 3. 相関関係の特異性 | B と「A 以外に原因として想定される変数」の相関は高くない。また A と「B 以外の結果変数」の相関も高くない。 |
| 4. 時間的な先行性 | A は B に時間的に先行する。 |
| 5. 量・反応関係の成立 | 原因となる変数 A の値が大きくなると、単調に結果となる変数 B の値も大きくなる。 |
| 6. 妥当性 | A が B の原因となっているという因果関係が生物学的に(または各分野の知見にもとづいて)もっともらしい。 |
| 7. 先行知見との整合性 | これまでの先行研究や知見と首尾一貫している。 |
| 8. 実験による知見 | 動物実験などの実験研究による証拠がある。 |
| 9. 他の知見との類似性 | すでに確立している別の因果関係と類似した関係・構造を有している。 |

Hill, A. B., The Environment and Disease: Association or Causation?, Proc. R. Soc. Med., 58, 295-300, 1965. および星野崇宏『調査観察データの統計科学』岩波書店, 2009, p. 140 より作成。

交絡因子と見せかけの相関

例：飲酒は肺がんのリスク要因である(?)



交絡因子(cofounders)

因果の上流に共通因子が存在

見せかけの相関(spurious correlation)

因子「喫煙」が交絡している

交絡にどう対処するか？

理想「実験する(介入する)」：

介入するか否かを無作為に割り付けるランダム化比較試験(RCT)



観察研究ではできない：喫煙するかどうかを割り付けできない

観察研究による因果推論の基本

必要な前提：興味の対象の関係因子と交絡因子がすべて測定されている（さらに因子の間の因果構造も分かっている）

→ よく分からない対象では現実的に満たされづらい…

① 層別(Stratification)

喫煙=有の群と、喫煙=無の群に分け、各々解析したあと統合

② 回帰分析の利用

「喫煙」を説明変数に含めて回帰分析で有意性検定

→ 交絡しそうな因子は全て説明変数に入れておけば良いが
サンプル数によっては回帰分析が破綻してしまう

- ・「傾向スコア」によって多数の共変量を1次元に変換する
- ・「バックドア基準」によって取り入れるべき説明変数を選ぶ

統計学と機械学習の「溝」

注意：機械学習屋は因果をあまり気にしない

- 相関関係の利活用で"予測がめっちゃ当たるんならええやん..."

しばしば「ええわけないやろ」という軋轢を生んできた...

e.g. 言語学の巨人 Chomsky vs Google研究部門長 Norvig

<http://norvig.com/chomsky.html>

- 統計学と機械学習の違い: データや変数に対する仮定が違う

統計学: 制御された実験計画 (臨床試験, 社会調査, 農業試験,...)

機械学習: 制御されないデータ (画像, 音声, テキスト, 信号, ...)

特徴量 vs 説明変数: 因子的意味はない場合も(画像のピクセル)

 Statistical Modeling: The Two Cultures (Breiman, Statist. Sci. 16(3), 199-231, 2001)

データ駆動科学: 科学も因果(理由)が主たる関心

科学の関心は「仕組みや原理がよく分からぬ現象」



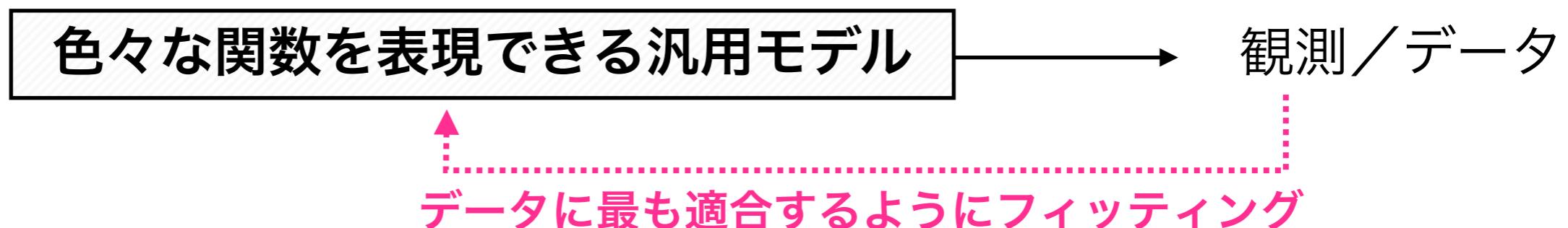
演繹

- **Theory-driven / Hypothesis-driven (自然科学)**



帰納

- **Data-driven (機械学習、人工知能、統計学など)**



Data-driven vs Theory-driven

Data-drivenはTheory-drivenと考え方・役割が異なるので注意

All models are wrong, but some are useful
(George Box)

Theory-driven models can be wrong
But data-driven models **cannot be wrong**



David Hand

Data-driven vs Theory-driven

Data-drivenはTheory-drivenと考え方・役割が異なるので注意

All models are wrong, but some are useful
(George Box)

Theory-driven models can be wrong
But data-driven models **cannot be wrong**

or right



David Hand

Data-driven vs Theory-driven

Data-drivenはTheory-drivenと考え方・役割が異なるので注意

All models are wrong, but some are useful

(George Box)

Theory-driven models can be wrong

But data-driven models **cannot be wrong**



or right

David Hand

Data-driven are **not trying to describe an underlying reality.**

But are merely intended to be **useful**
so they could be **poor or useless, but not wrong**

*If data can speak for themselves,
they can also lie for themselves*

David Hand

So it's critically important to

- exercise caution
- do not claim too much
- understand the data
- and its quality

cf.

With enough data, the
numbers speak for
themselves.

Chris Anderson (2008)

WIRED

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE
DATA DELUGE MAKES THE
SCIENTIFIC METHOD OBSOLETE

KDD2018



「理解」編: まとめ

データの相関関係の利活用技術である機械学習だけでは
対象現象の背後にある仕組みを理解するのは原理上困難

- 相関関係は必ずしも因果関係を意味しない
- 因果の検証には観察研究ではなく介入研究が必要
- 医療や脳科学など倫理的に介入研究が難しい場合も多く
因果推論の理論・手法は長らく研究されてきている
- 因果推論では関連因子や因果構造がすべて分かっている
などの現実的には難しい前提が満たされる必要がある
- 相関関係は因果の示唆ではあるので注意深く考えよう！

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

「発見」は学習の発展系？

発見 = 今までにないもの・ことを見つける

- 今までにない画期的な新薬
- 今までのデータのどれよりも長持ちする電池材料
- 今まで誰も試さなかった画期的な会社経営戦略
- 今まで未発見だった画期的な科学法則や科学理論
- 今まで対戦した誰よりも強いボードゲーム勝利戦略

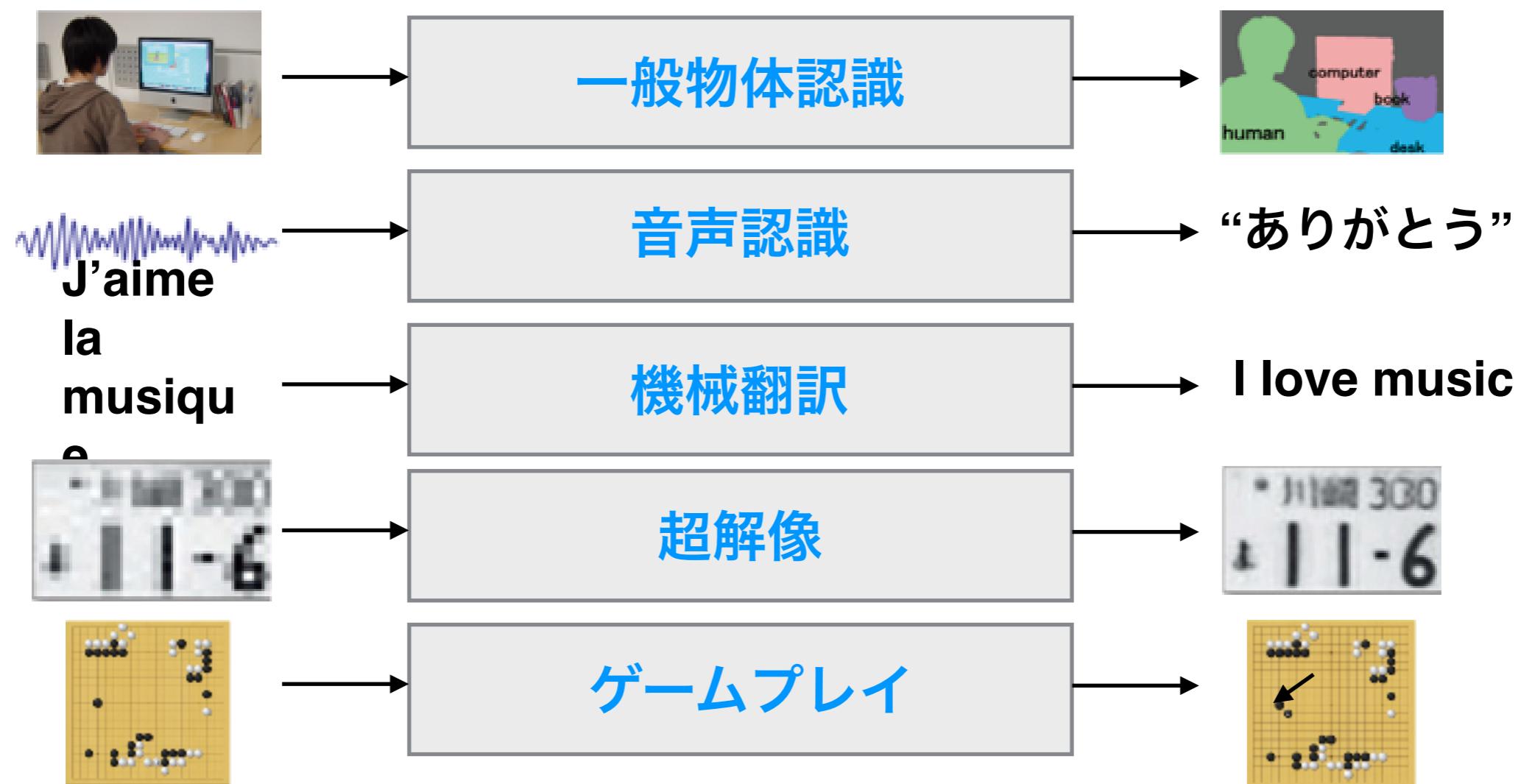
発見はふつう完全に行き当たりばったりではない。

「勘と経験」が非常に大切 「幸運は準備された者に降りる」

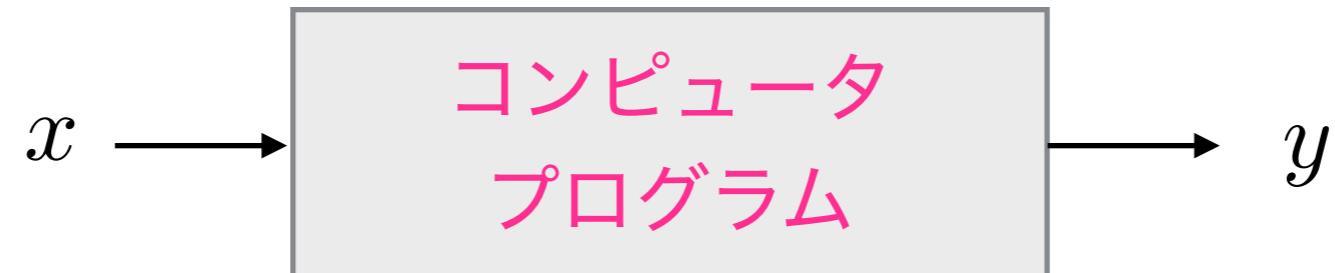
経験(過去のデータ)から学習した勘(法則性) = 機械学習
と考えるとなんだかイケそうな気がする～？

一方、機械学習とは何だったか

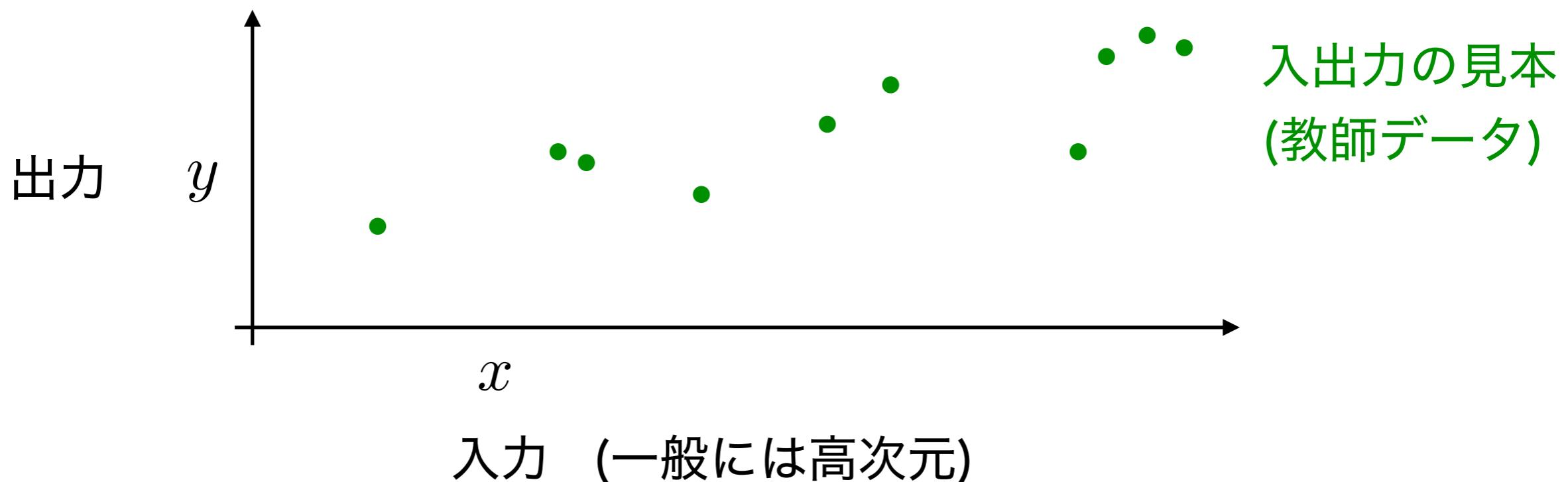
入出力の関係がよく分からない変換過程(関数)を大量の入出力の見本例から明示的にプログラミングすることなく構成する技法



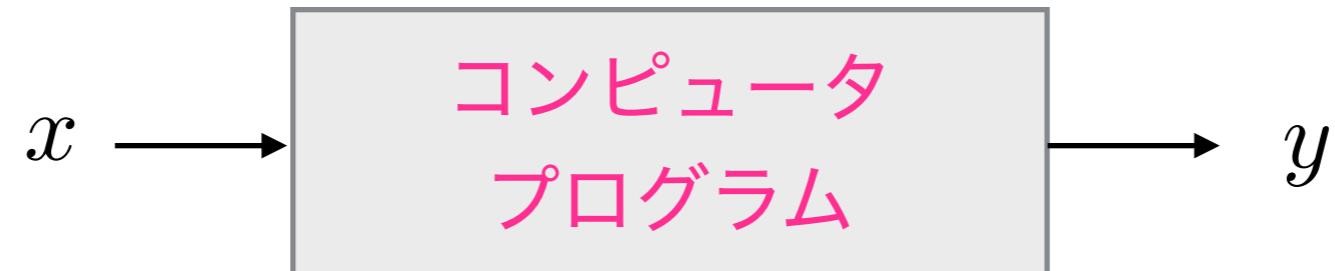
機械学習の仕組み = 高次元での曲面あてはめ



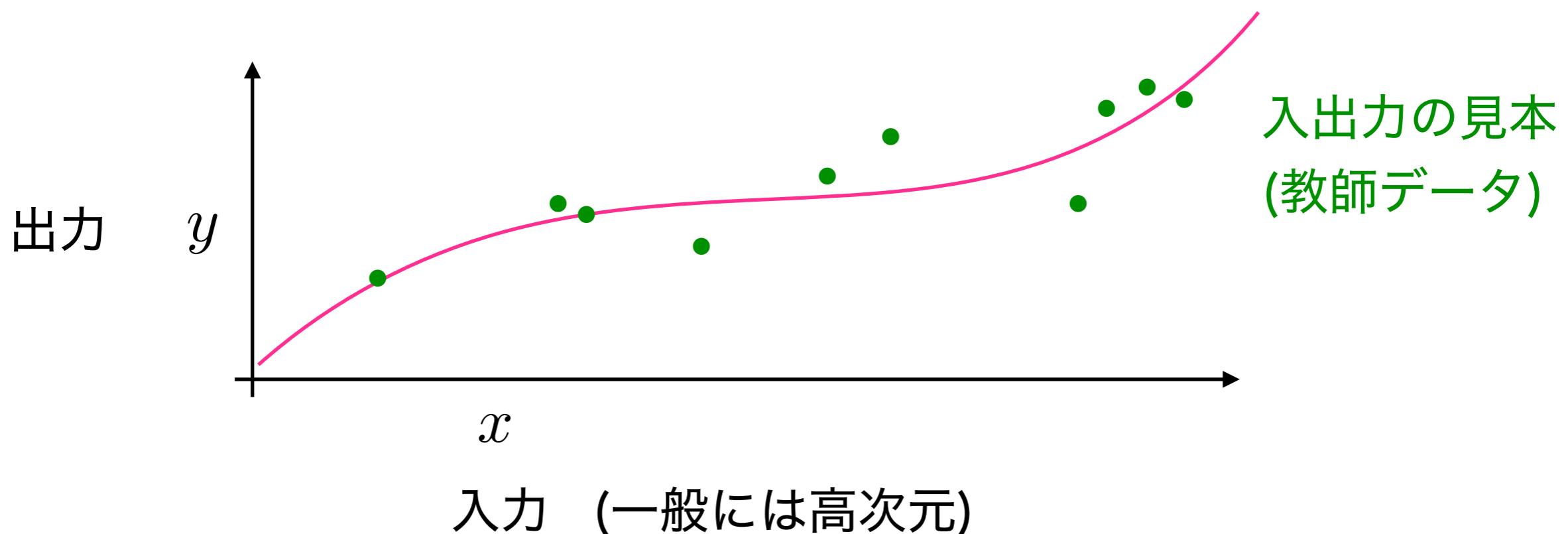
入出力の見本: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$



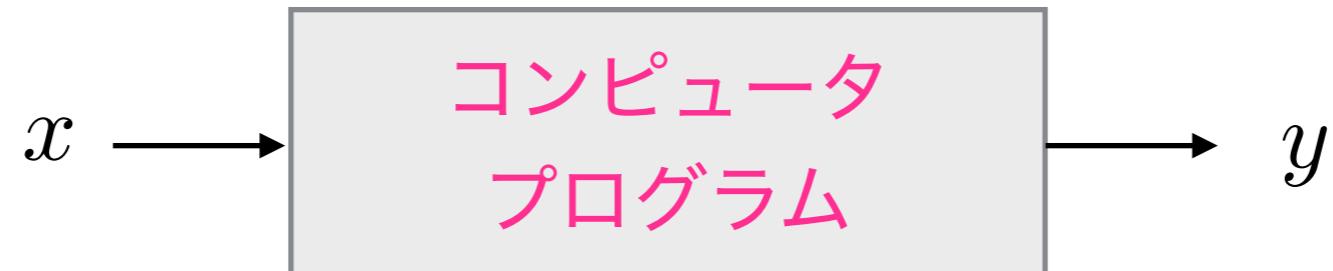
機械学習の仕組み = 高次元での曲面あてはめ



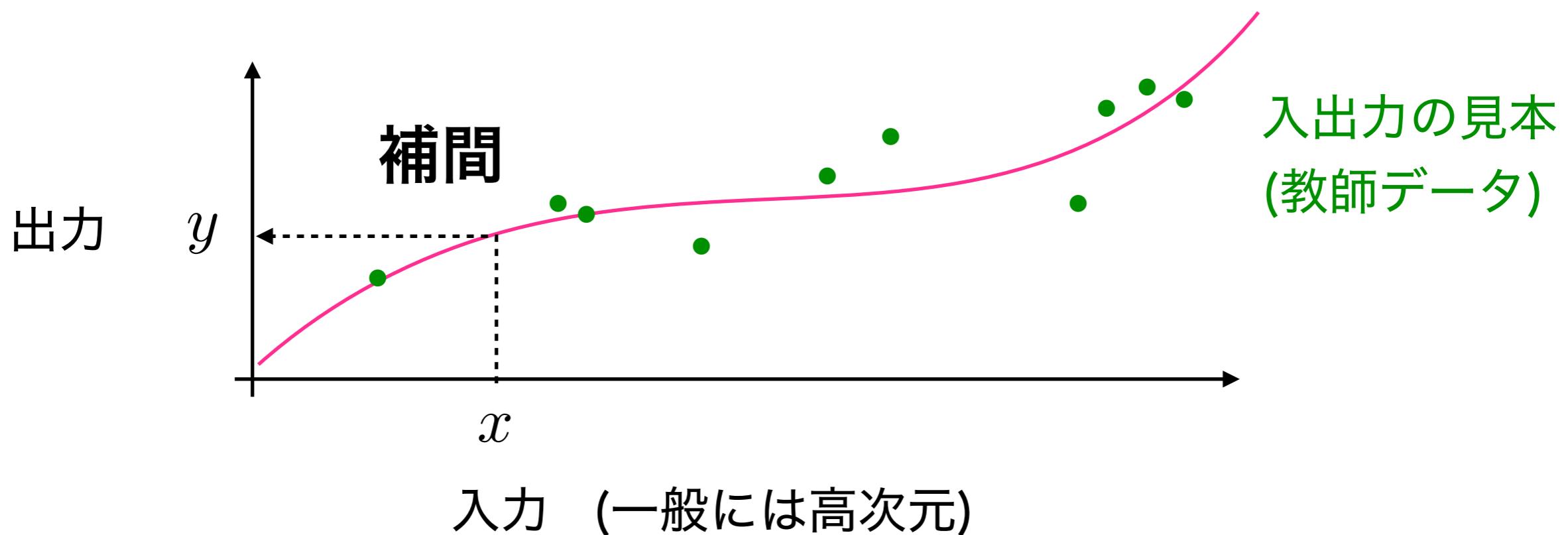
入出力の見本: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$



機械学習の仕組み = 高次元での曲面あてはめ



入出力の見本: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

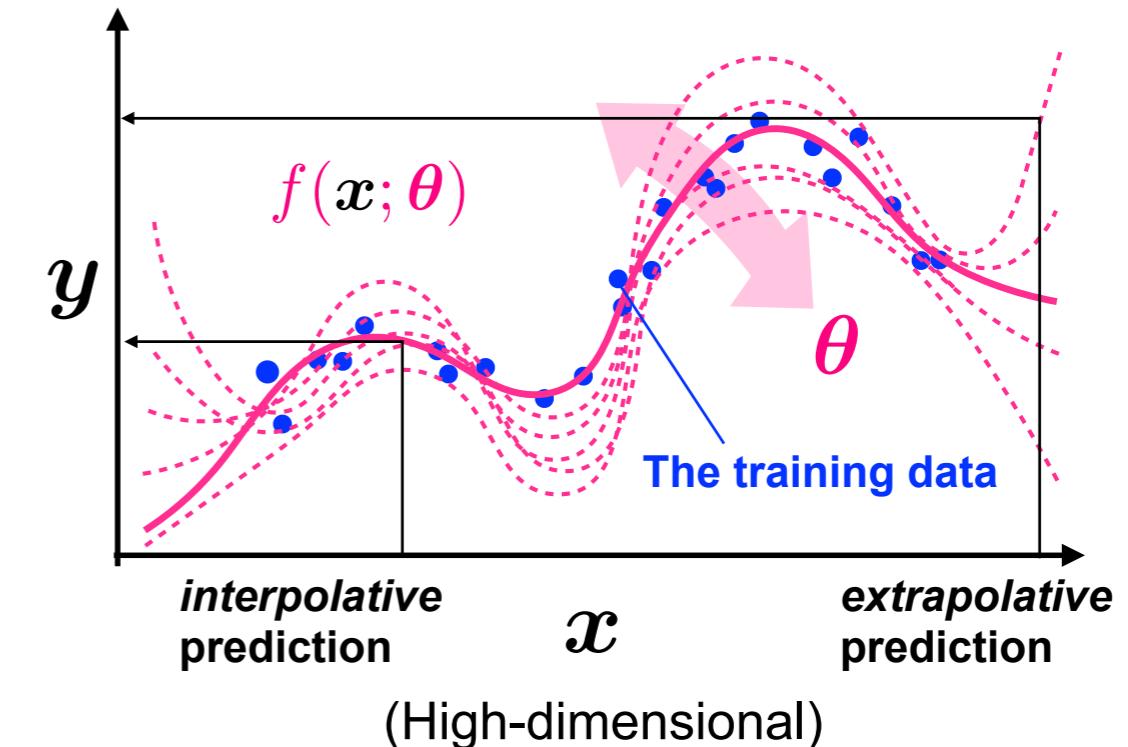


機械学習の仕組み = 高次元での曲面あてはめ

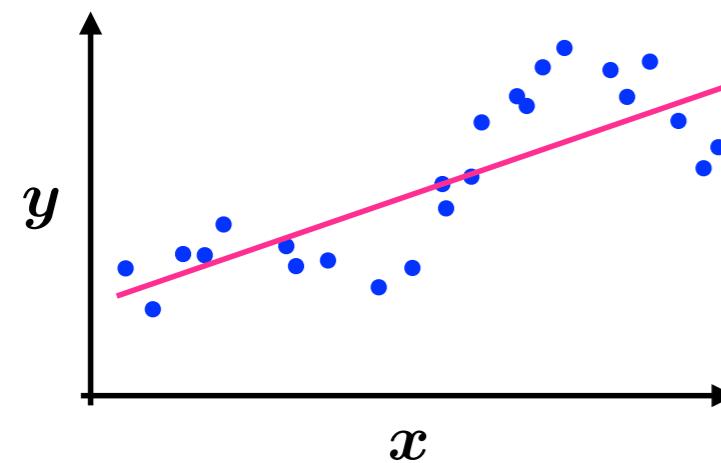


A function $f(x; \theta)$ best fitted to a given set of example input-output pairs (the training data).

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

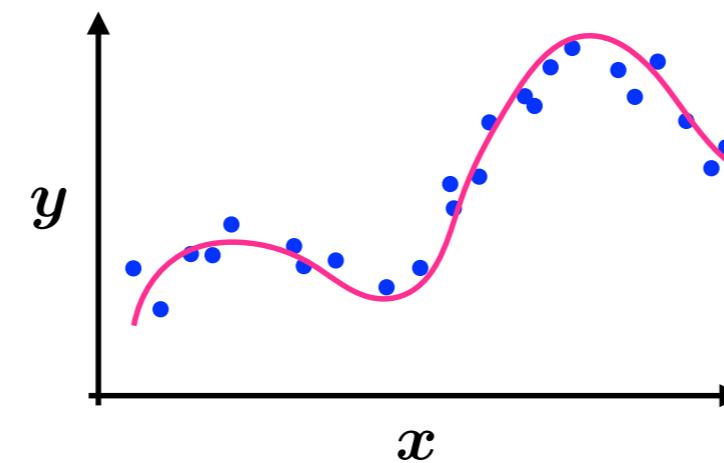


Underfitting
(High bias, Low variance)



"The bias-variance tradeoff"

Overfitting
(Low bias, High variance)



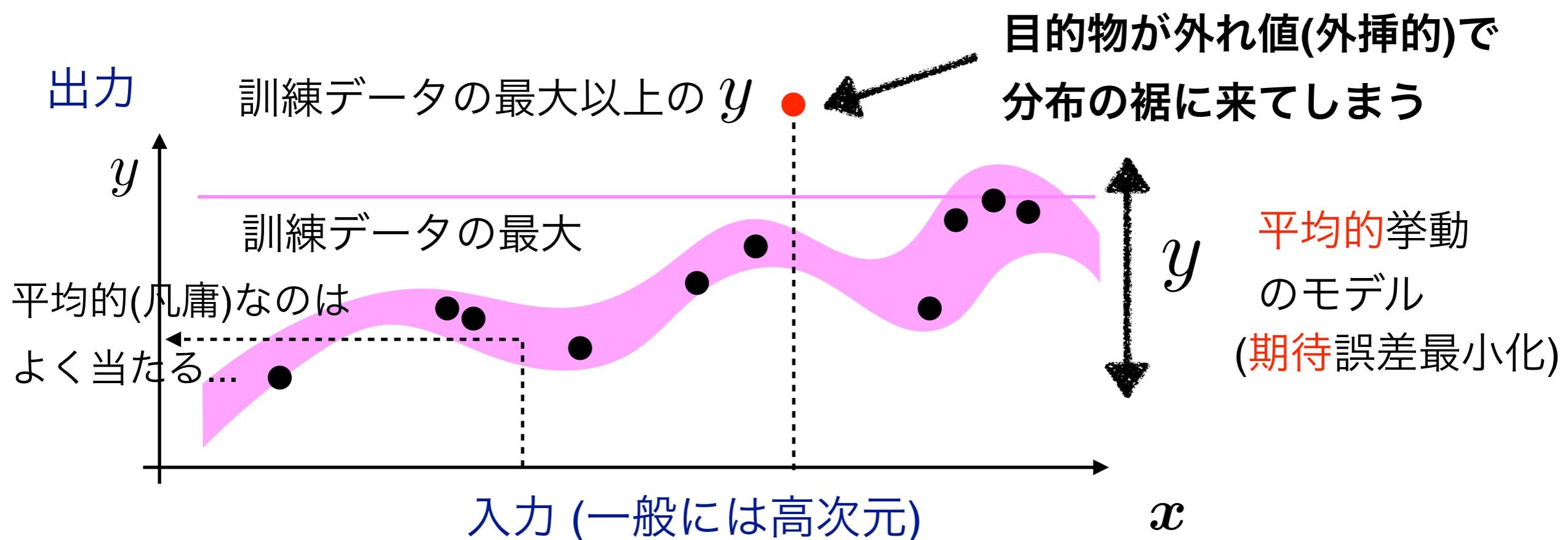
Low

Model Complexity

High

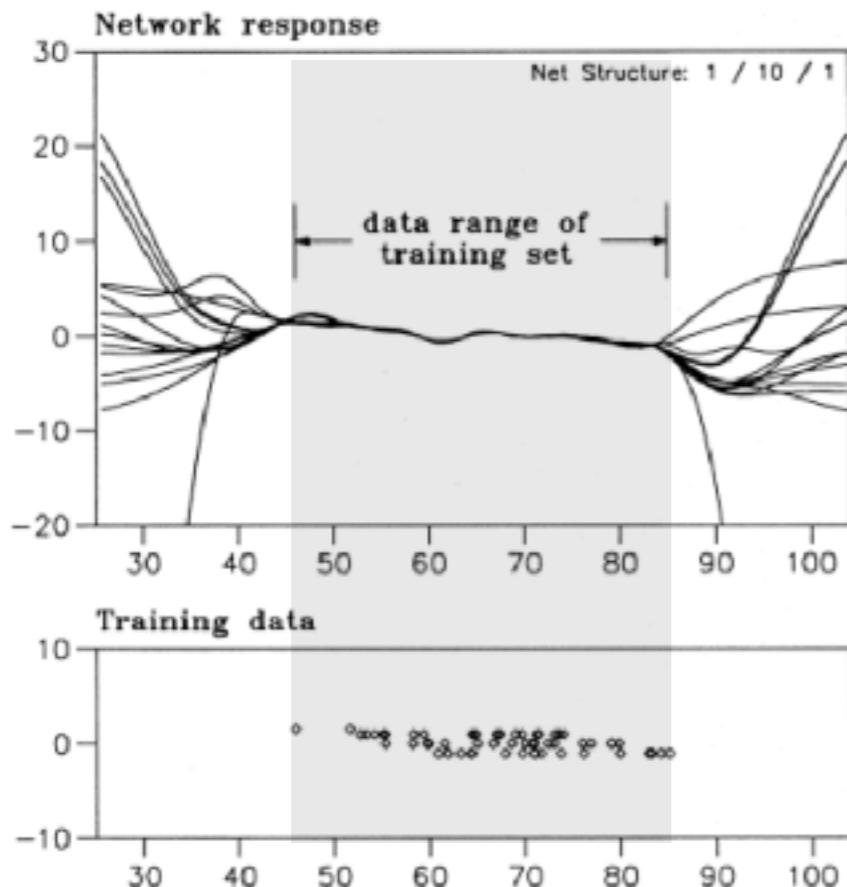
注意：機械学習は「発見」に向いていない

機械学習 = 訓練データの平均的法則性をとらえる
目的が不整合 → 予測モデルとの誤差の「期待値」を最小化 = 汎化
**発見 = 見本データの中にはないものを見つけたい
「外れ値」**



機械学習は与えた訓練データを代表するだけ

Highly Inaccurate Model Predictions
from Extrapolation (Lohninger 1999)



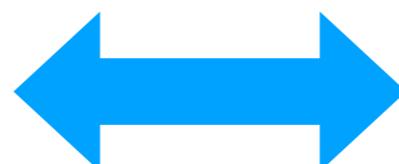
与えたデータの傾向を(曲線あてはめで)
表すだけでデータがない外挿領域では
無根拠な予測を返す



探索 "exploration"

新しい知識/データを獲得

トレードオフ



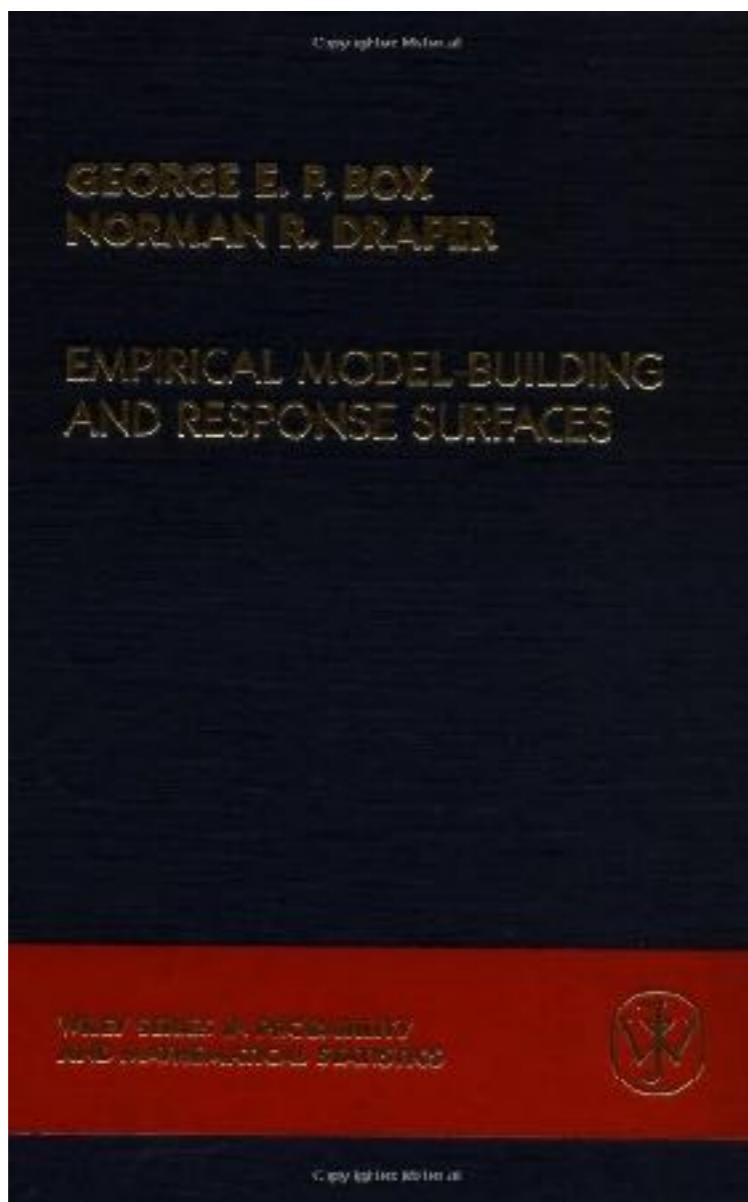
利用 "exploitation"

獲得した知識/データを利用

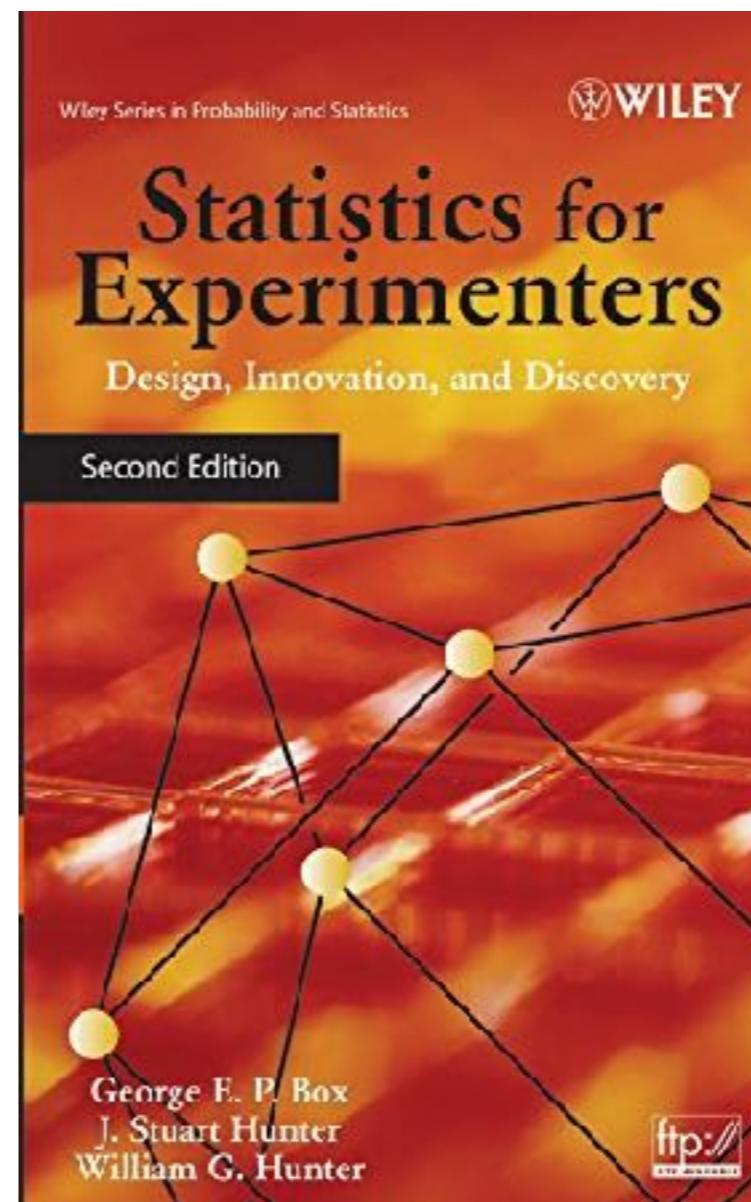
介入実験の計画：実験計画法と応答曲面法

当然Boxはどのように回帰分析を使えば良いか探求済み！ 😊

Empirical Model-Building and Response Surfaces (1987)



Statistics for Experimenters: Design, Innovation, and Discovery (2005)



Box-Wilsonの応答曲面法

応答曲面法 (Box & Wilson, 1951)

1. 応答曲面(Response Surface)をモデル化
(e.g. 二次多項式回帰)
2. 上記モデルを当てはめるための実験計画(e.g. 中心複合計画)で検査点を得る
3. 応答曲面を検査点に当てはめそれが最大になる点を求める

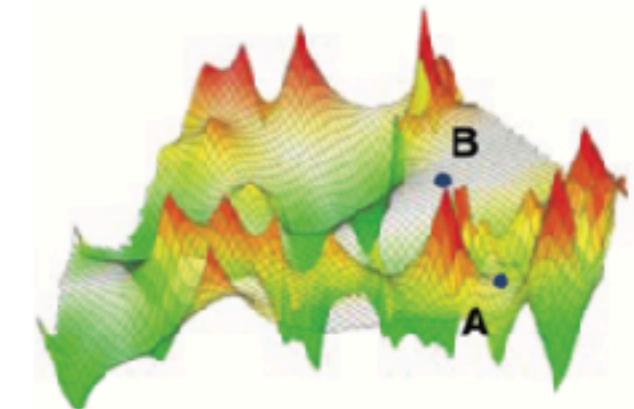
探索空間(関心領域)の内挿になるよう実験計画で事例点を得る



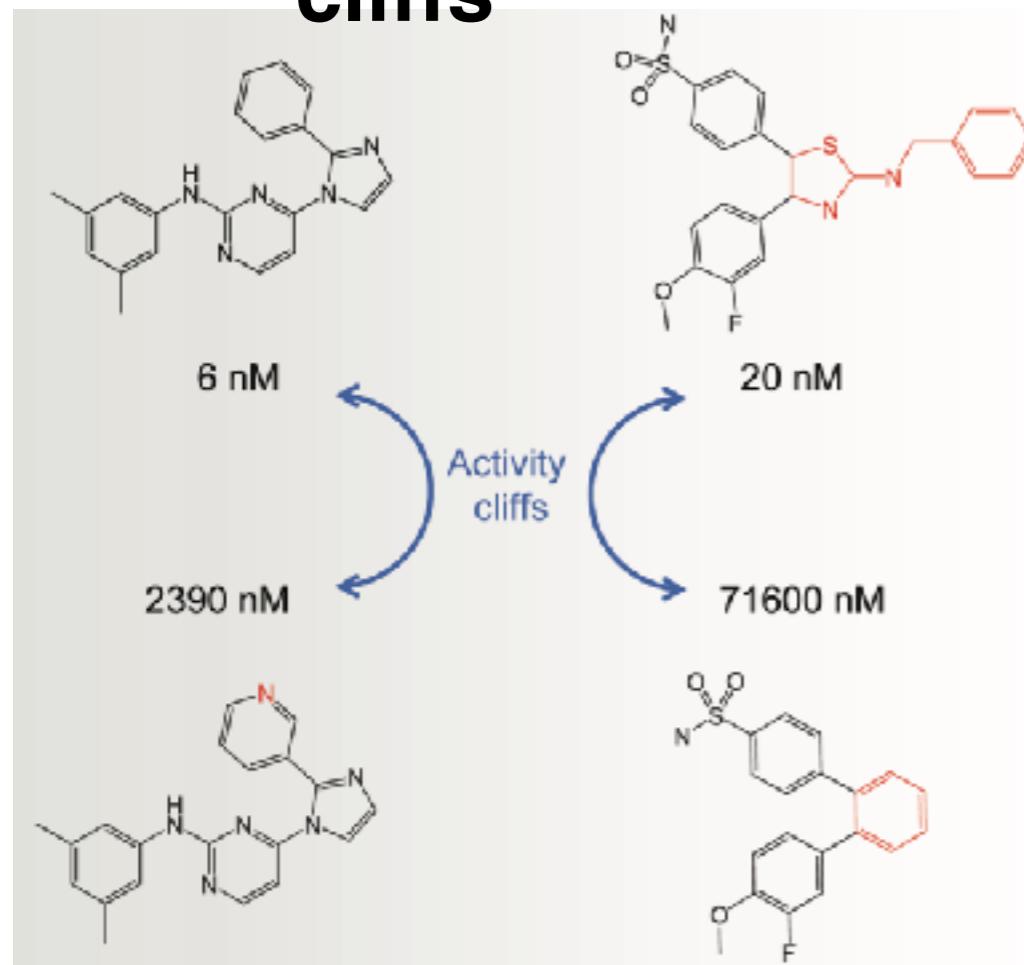
変数の数が少なく統計学的な仮定がある程度有効ならこれでOK。
しかし、現代の実問題のデータは...

入力が多様 + 少しの変化で出力が変わり得る

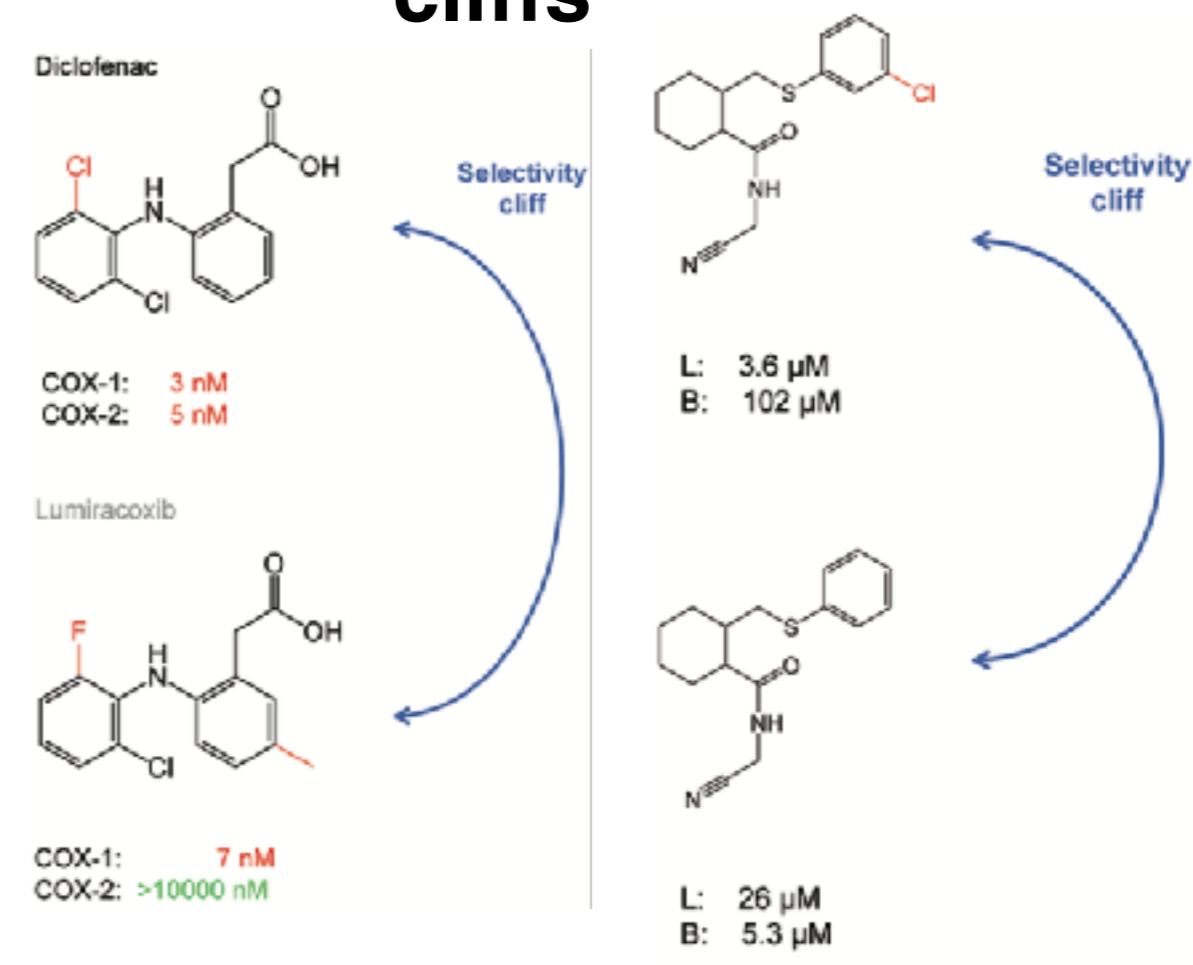
物性や活性のランドスケープは非平滑的
(少しの構造変化が急峻な影響をもたらす)



Activity cliffs



Selectivity cliffs



さらに高次元では外挿か内挿かの判定すら難しい...



与えられた訓練データ

内挿 or 外挿？

高次元空間は非直感的な性質を持つ

- 偽相関：変数の数があまりに多いと訓練データすべてをとおる曲面が自由に作れてしまい偽相関が生じやすくなる
- 測度の集中現象：見本点の間の距離が全てほぼ同じになる

成功例も内挿的だと直感するのは非常に困難

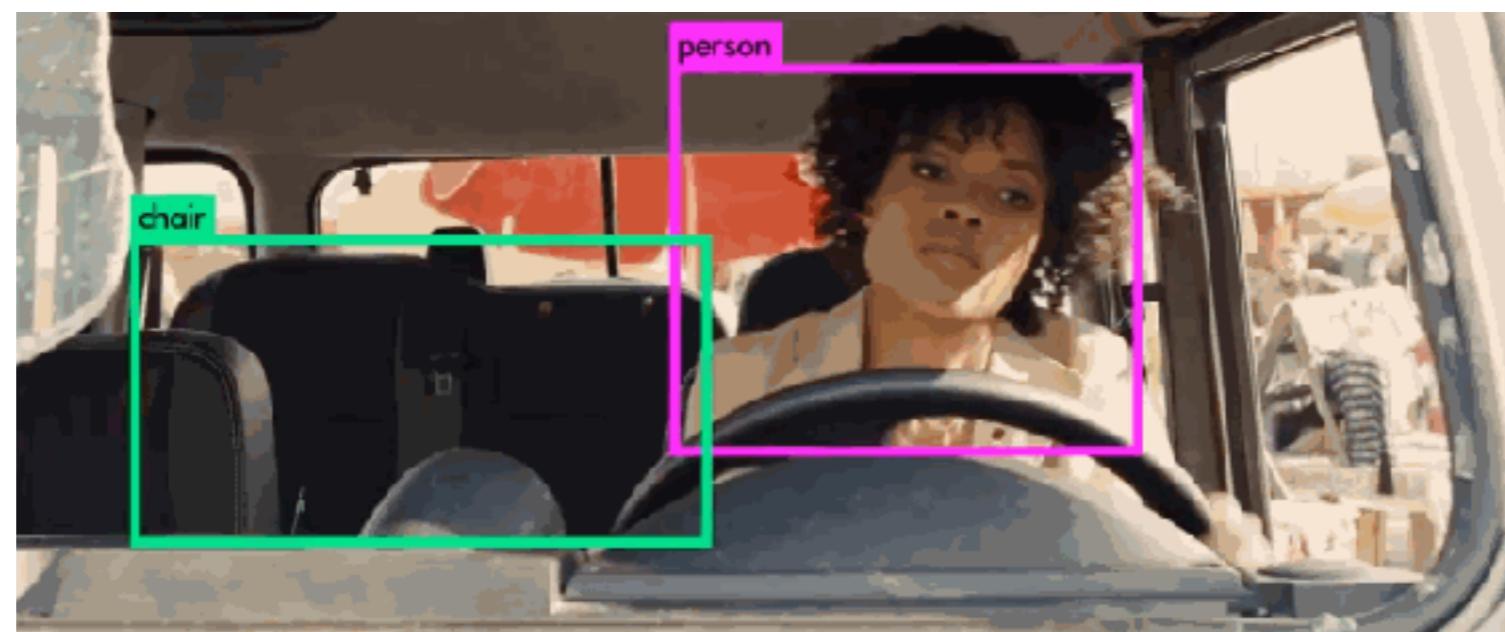
pix2pix



**CycleGAN
(e.g. DeepFake)**



YOLO



成功例も内挿的だと直感するのは非常に困難

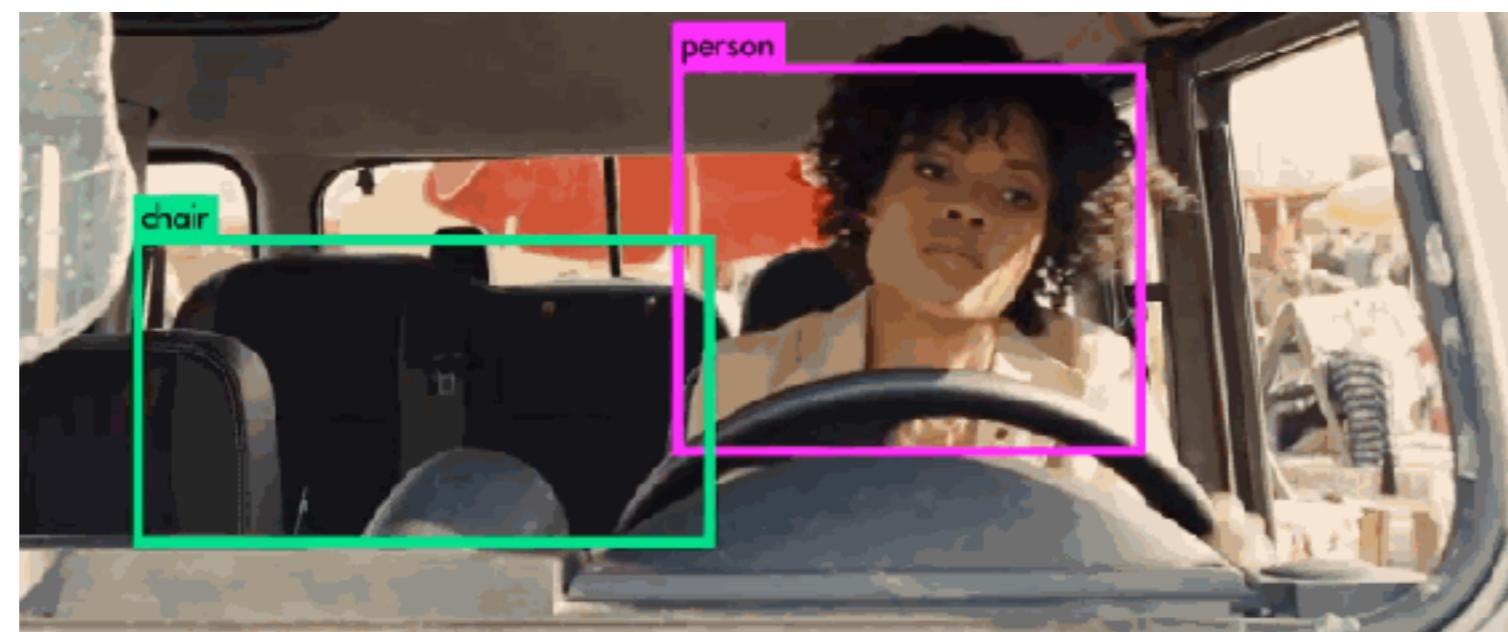
pix2pix



**CycleGAN
(e.g. DeepFake)**

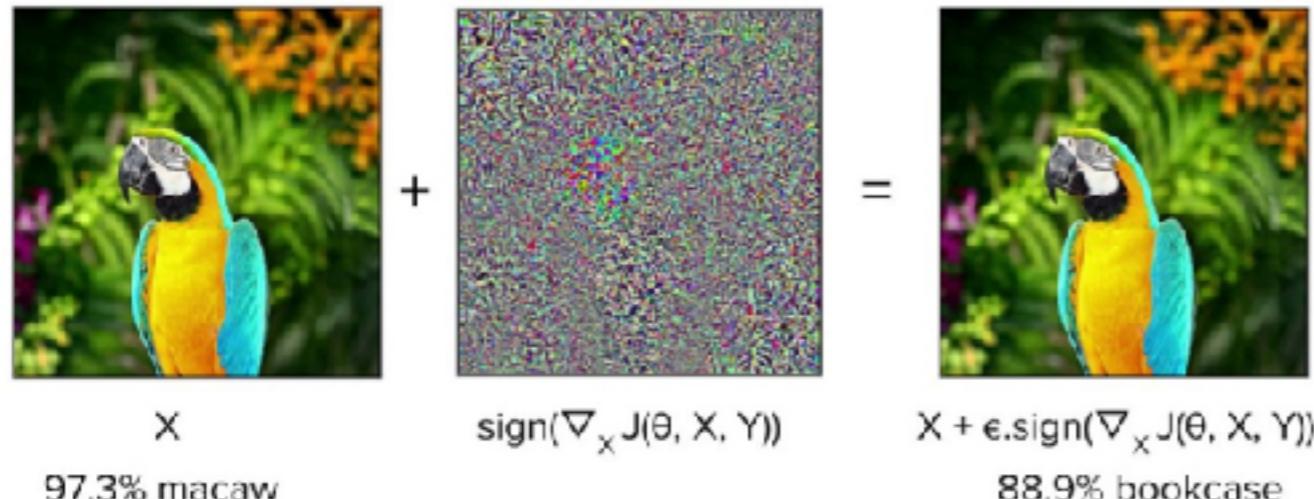


YOLO



あてはめる曲面のほうも高次元では非直感的

e.g. Adversarial examples (GANの発想)、内挿 or 外挿の判定の難しさ



"stop"
to "30m speed limit"

"80m speed limit"
to "30m speed limit"



"go right"
to "go straight"

Direct Encoding					Indirect Encoding				
brambling	redshank	robin	cheetah	king penguin	starfish	baseball	electric guitar		
armadillo	lesser panda	centipede	jackfruit	freight car	remote control			peacock	African grey

「発見」編: まとめ

データの曲面あてはめによる内挿である機械学習だけでは
訓練データに全くない新規発見をするのは原理上困難

- 機械学習は曲面あてはめで訓練データを代表するだけ
- 曲面は見本データにあうようフィットされるので
外挿的なトレンドは予測根拠がきわめて薄くなる
- 発見(知識獲得)には知識の利用と探索のトレードオフの
考慮が必須
- 最近のデータは複雑で多様で制御されていないので古典
的な実験計画や曲面応答方ではなかなか十分ではない

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

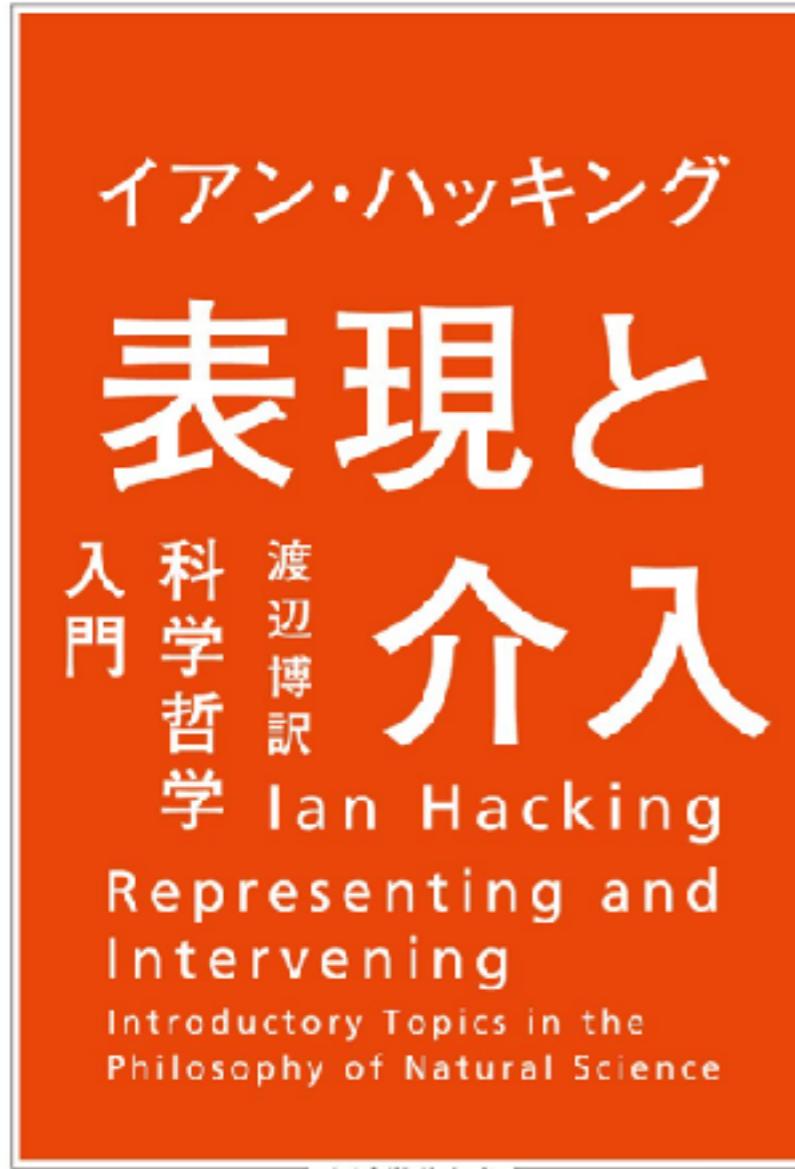
Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

科学的「理解・発見」に必要な二大要素？



「表現」

- 対象をどう表現するか？何を測るか？
- 問題を内挿的にする表現の学習
(いまのところ設計に要ドメイン知識)
- 背景過程について分かっている
ことの反映や活用 (帰納バイアス)

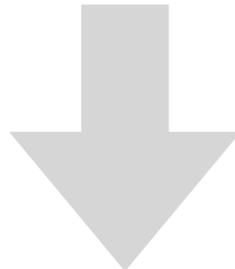
「介入」

- 機械学習に「実際に追加データを取りに行く」仕組みを融合
- 次に何を実験するかの最適計画

→ 科学的「理解」や「発見」とは何か(何であるべきか)は
科学哲学の問題

機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



1.について

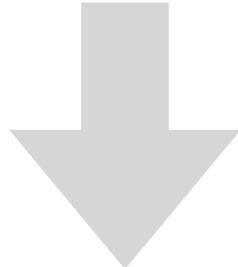
- **問題が内挿的になるよう工夫** (表現学習・隠れ構造同定)
- **内挿・外挿判定** (予測の信頼度計算)
- **モデルベース最適化と探索** (最適実験計画)

2.について

- **ポストホック解析と解釈性モデル** (学習済みモデル分析)

機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



1.について

- **問題が内挿的になるよう工夫 (表現学習・隠れ構造同定)**
- **内挿・外挿判定 (予測の信頼度計算)**
- **モデルベース最適化と探索 (最適実験計画)**

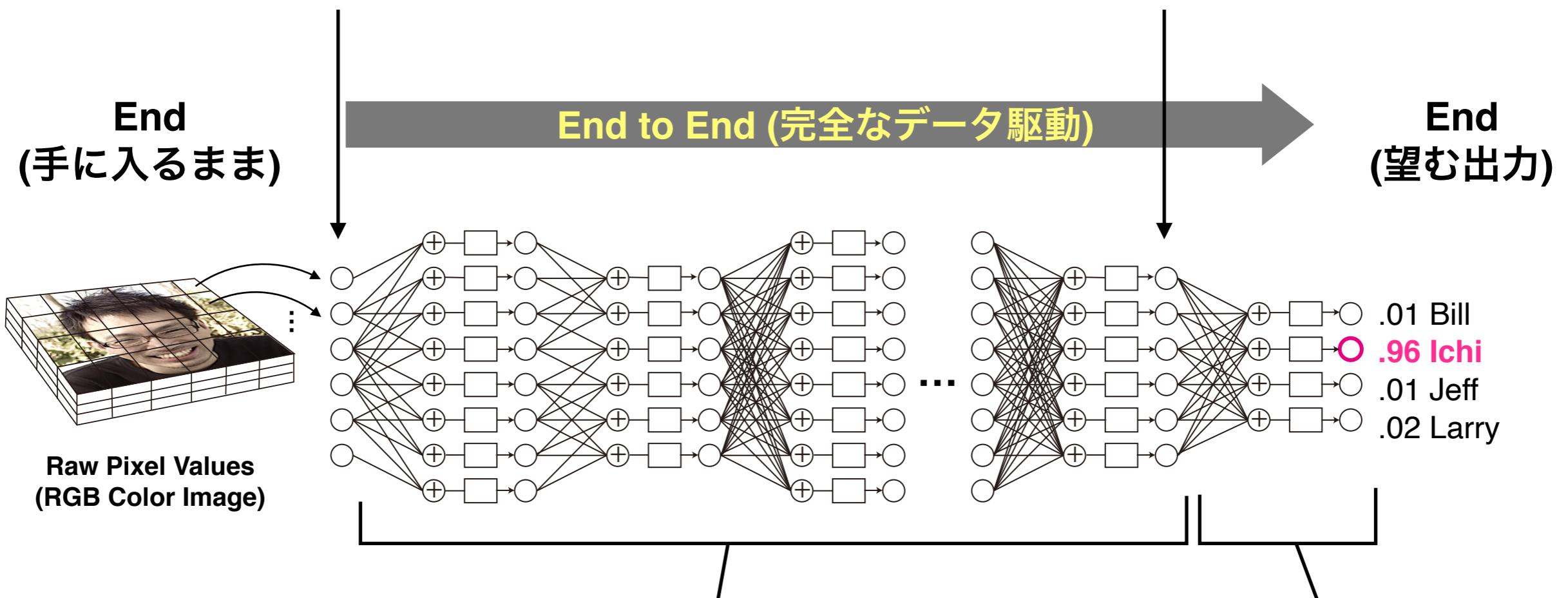
2.について

- **ポストホック解析と解釈性モデル (学習済みモデル分析)**

深層学習による表現学習

入力変数の段階では
内挿的じゃなくても...

予測に使う中間表現(隠れ
構造)で内挿的であればOK!

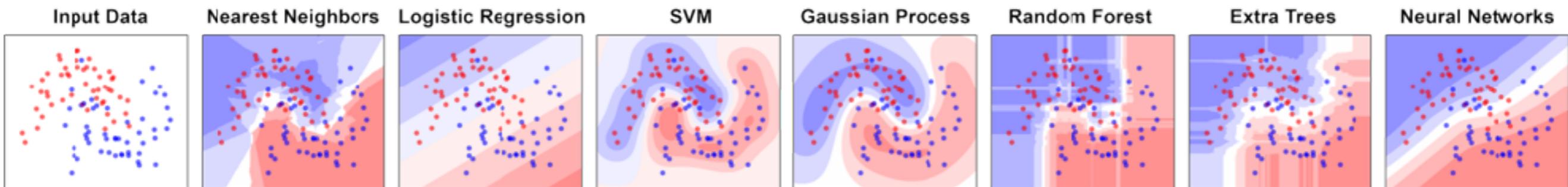


入力の「多次元の数値組(ベクトル)」を少しづつ
別の「多次元の数値組(ベクトル)」へ変換するプロセス

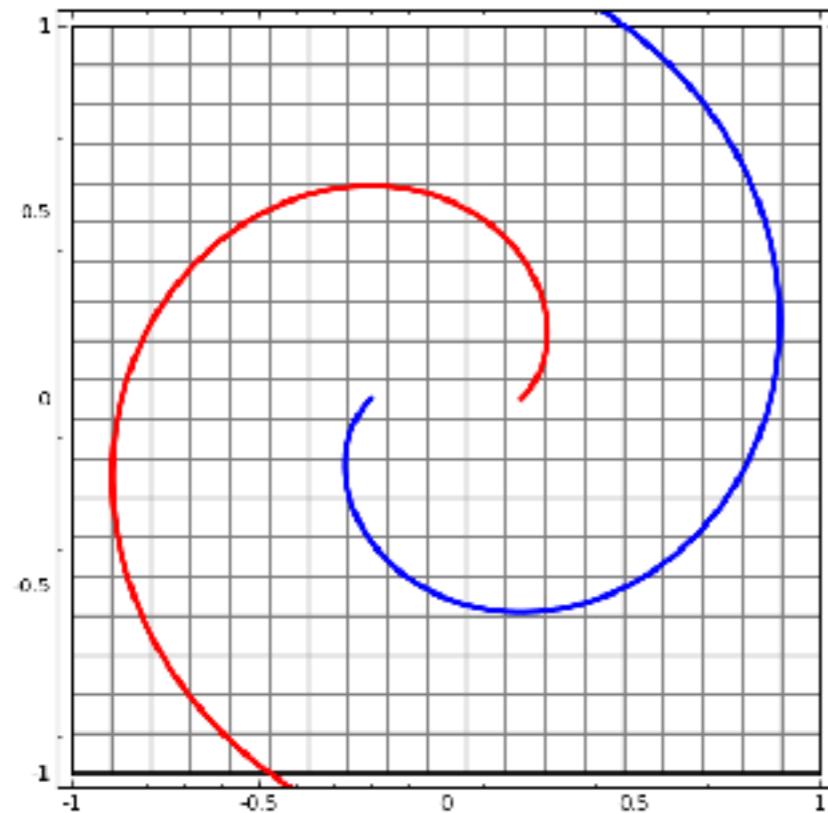
変換された最終量
について予測

例) 深層学習は入力の多段の変換プロセスを学習

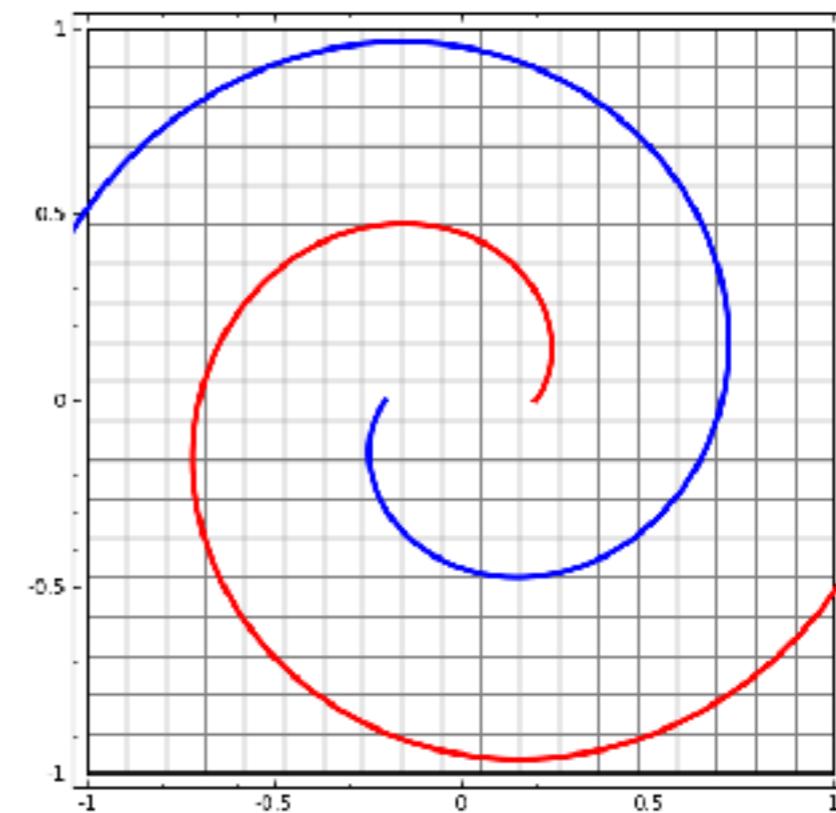
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>



線形分離できるような変換を学習

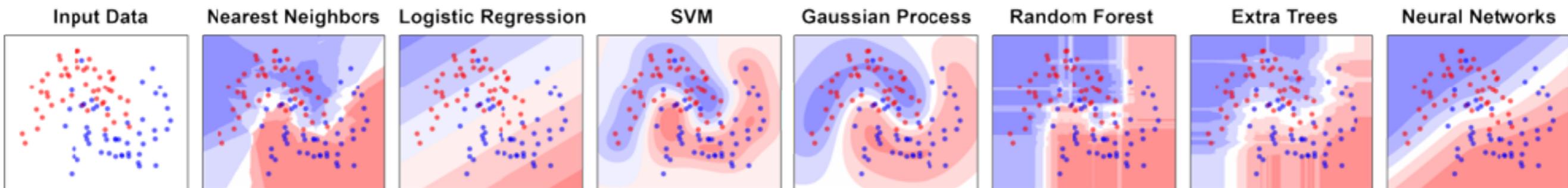


失敗例(常に成功するとは限らない)

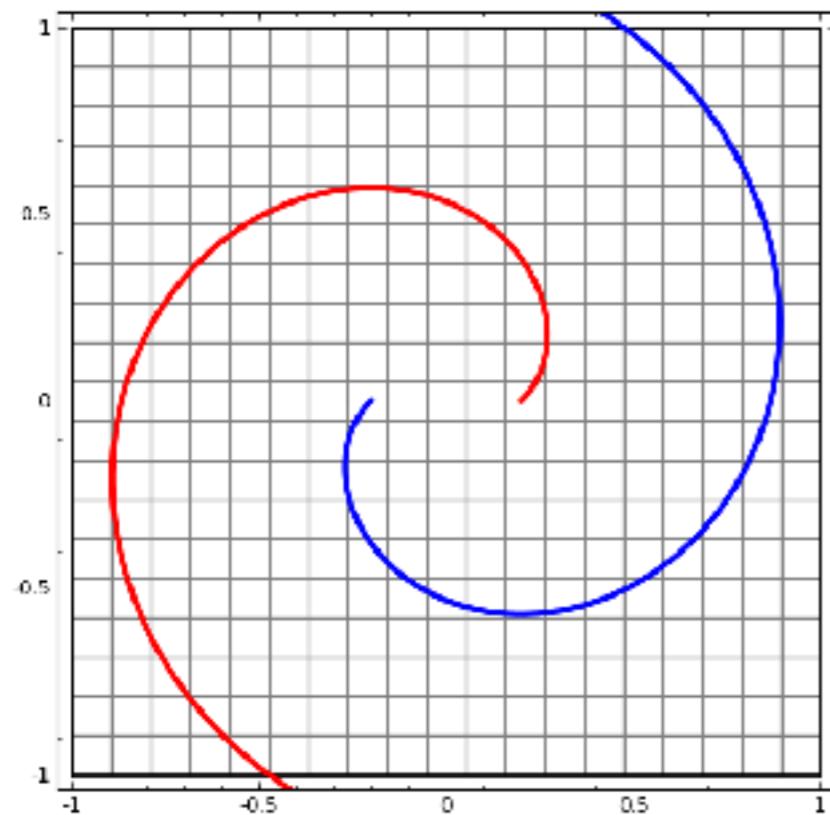


例) 深層学習は入力の多段の変換プロセスを学習

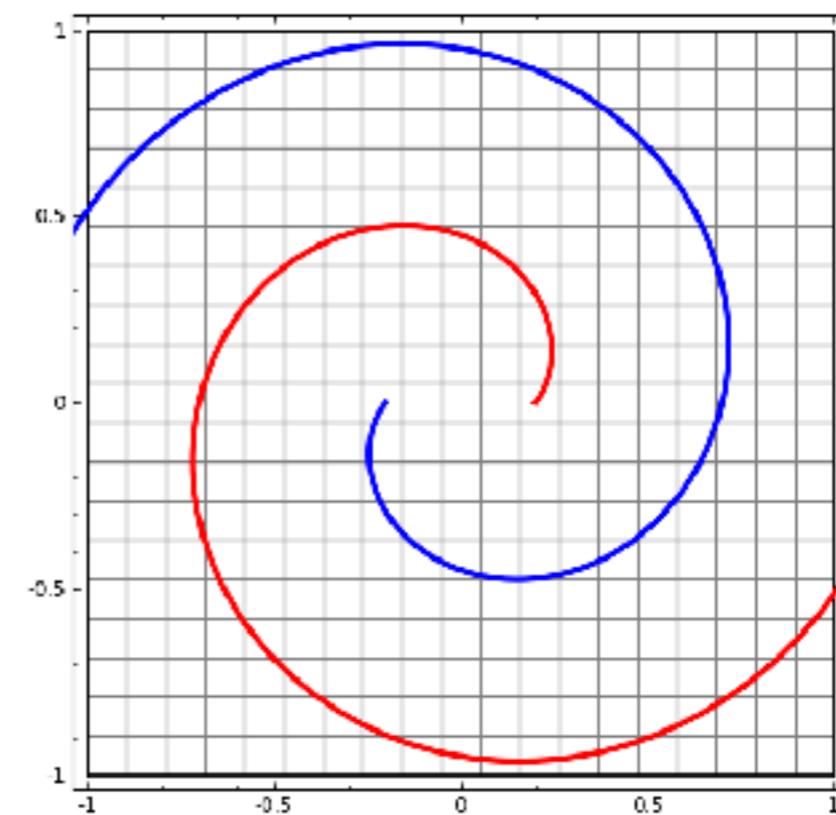
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>



線形分離できるような変換を学習

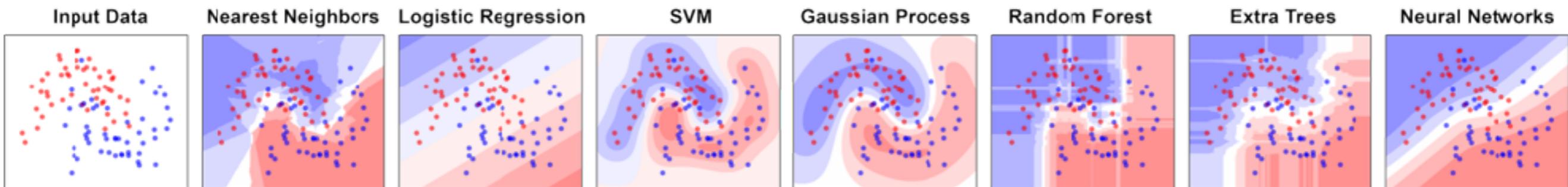


失敗例(常に成功するとは限らない)

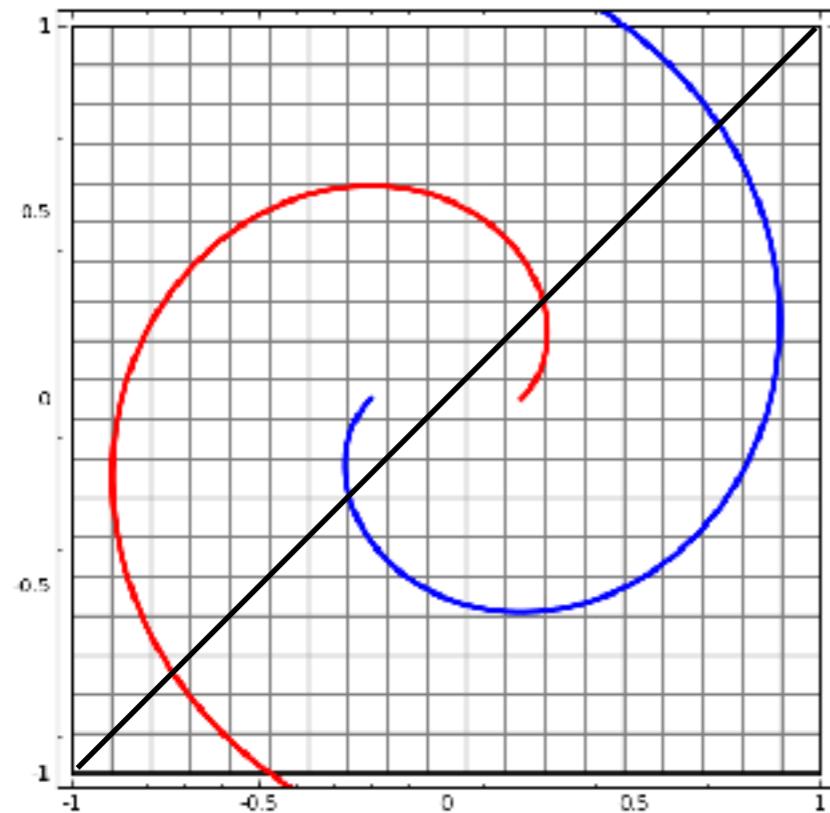


例) 深層学習は入力の多段の変換プロセスを学習

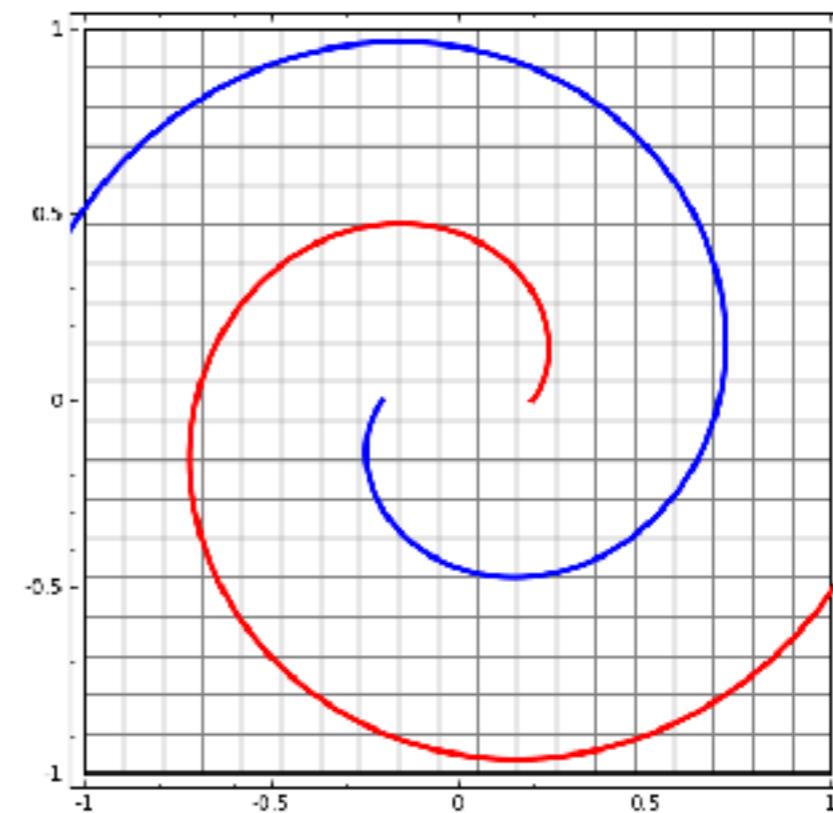
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>



線形分離できるような変換を学習

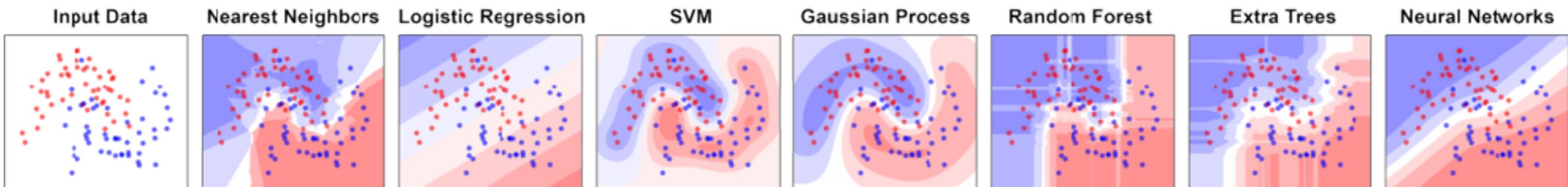


失敗例(常に成功するとは限らない)

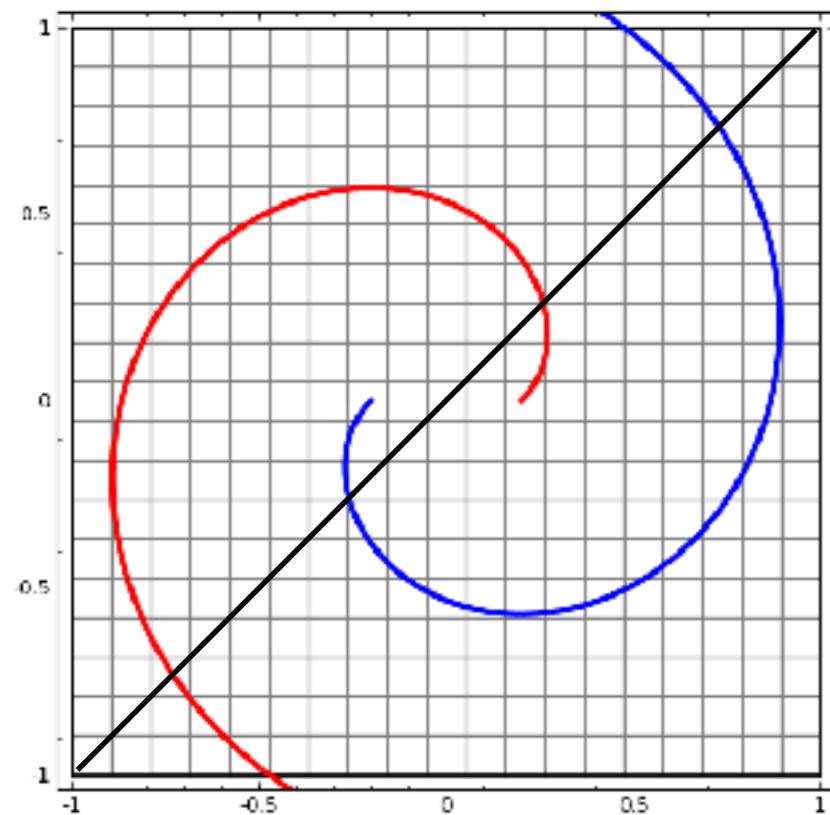


例) 深層学習は入力の多段の変換プロセスを学習

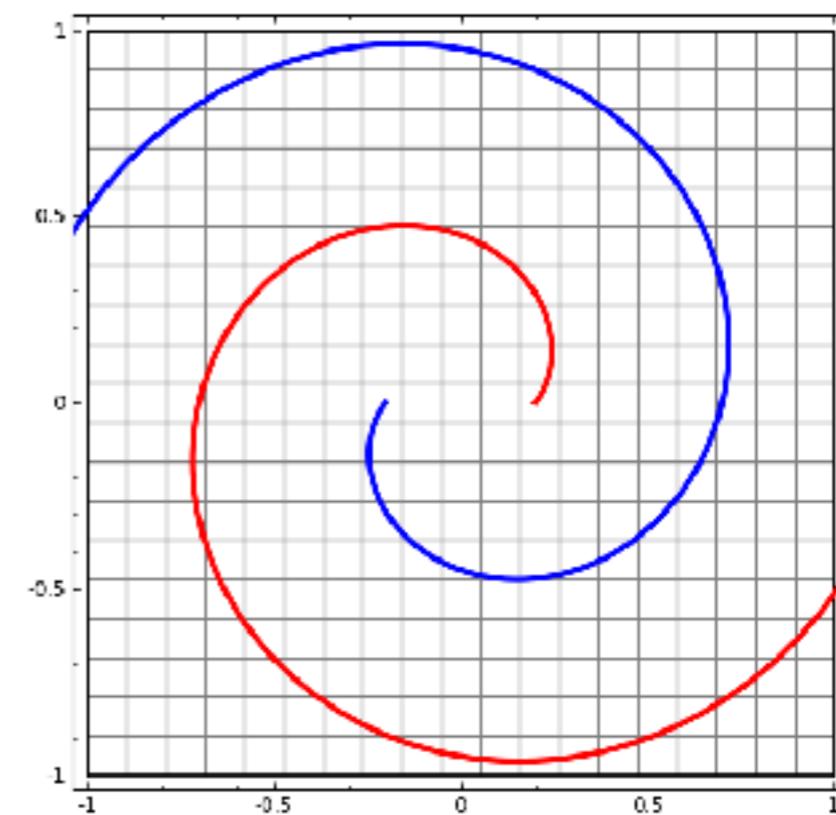
<https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/>



線形分離できるような変換を学習



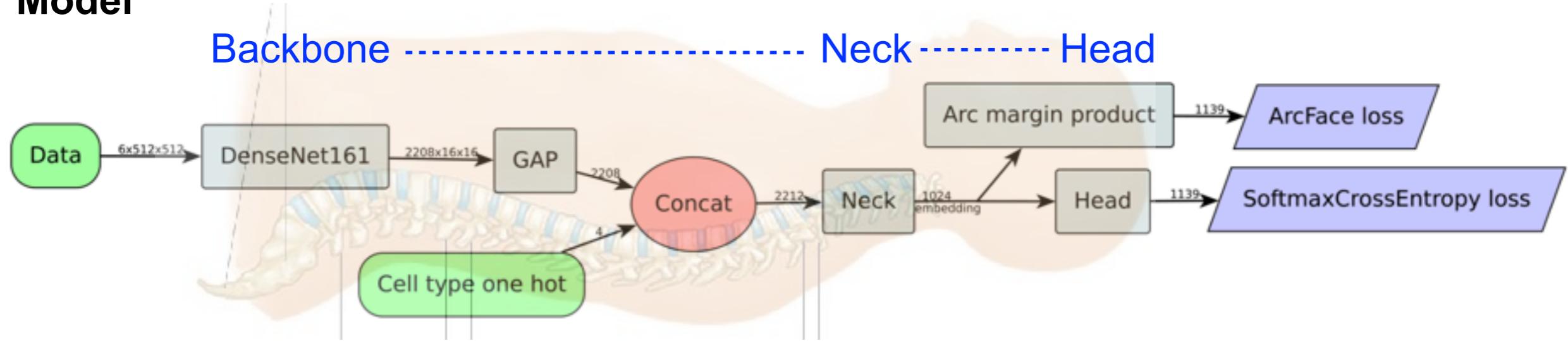
失敗例(常に成功するとは限らない)



モデル構造へのドメイン知識のエンコード

とにかく微分可能な演算の(深い)合成関数にさえなっていれば
モデルパラメタは自動微分(aka backprop)で学習できるので
背景過程を理解しそれにそって柔軟にモデル構造を設計

Model



- Backbone is pre-trained on ImageNet and first convolution is replaced with 6 input channel convolution
- Neck: BN + FC + ReLU + BN + FC + BN
- Head: FC

事前学習からの転移も有効

<https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110543#latest-637352>

事前学習が効かないとされた言語タスクでも...

かつて言語タスクではRNN(LSTM/GRU)→CNNの流れだったが...

Attention Is All You Need

arxiv.org ▾

by A Vaswani - 2017 - Cited by 3294 - Related articles

Jun 12, 2017 - Attention Is All You Need. ... The best performance was achieved by stacking multiple layers of the encoder and decoder through an attention mechanism. We

Googleの華麗な論文を契機にRNNやCNNよりTransformerが支配的に!?

超巨大な事前学習モデルの
(Attentiveな)「転移学習」へ

2018/10/18

→ Googleの言語モデルBERT

GLUEベンチマークの全言語理解タスクでぶっちぎりのSOTA！

質疑応答タスクのSQuADでもSOTA！

2019/01/31

→ Microsoftの言語モデルMT-DNN

2019/02/14

→ OpenAIの言語モデルGPT-2

作文性能が高すぎてオープンソースとして公開してしまうとフェイクニュースが作り放題になってしまふ懸念から研究者向けに規模縮小版のみを公開

2019/06/19

→ CMUのXLNet

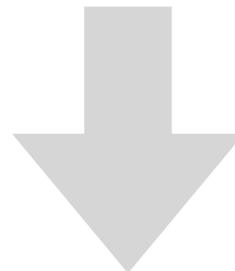
外挿的領域での機械学習の活用

見本点がない外挿的領域で機械学習(曲面あてはめ)を活用する。

- 類似した問題・データでの関係性を横断的に活用して類推
(転移学習、半教師つき学習、マルチタスク学習、注視的学习)
- 近さの測り方を適切に学習して問題を内挿的に (計量学習)
- 背景過程の第一原理モデル(シミュレーション)を不確実な因子を含めて立てその最適な推定値を機械学習する (データ同化)
- シミュレーションデータや経験者の教示を用いてデータを増やしできるだけ問題を内挿的に (敵対的生成、模倣学習)
- シミュレーション結果から実現象のギャップを機械学習する
(転移学習、メタ学習)

機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



1.について

- 問題が内挿的になるよう工夫 (表現学習・隠れ構造同定)
- 内挿・外挿判定 (予測の信頼度計算)
- モデルベース最適化と探索 (最適実験計画)

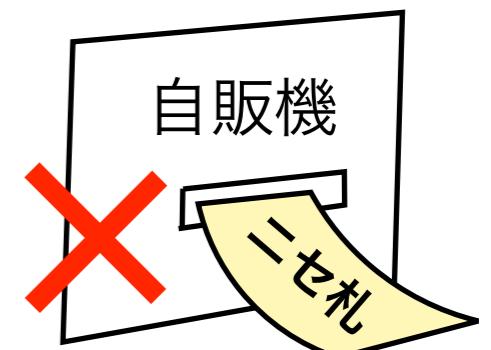
2.について

- ポストホック解析と解釈性モデル (学習済みモデル分析)

外挿判定もしくは信頼領域推定

予測したい検査入力点の近くに見本点が全然ない(or 非常に少ない)場合は基本的に外挿的

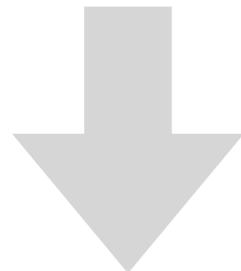
- パターン認識における棄却オプション
- 信頼区間(Confidence Interval)
- 信頼領域(Trust Region)
- ベイズにおける予測分布
- CheminfoにおけるApplicability Domain(AD)



😊科学と機械学習のあいだ：変量の設計・変換・選択・交互作用・線形性
<https://www.slideshare.net/itakigawa/ss-69269618>

機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



1.について

- 問題が内挿的になるよう工夫 (表現学習・隠れ構造同定)
- 内挿・外挿判定 (予測の信頼度計算)
- モデルベース最適化と探索 (最適実験計画)

2.について

- ポストホック解析と解釈性モデル (学習済みモデル分析)

機械学習の利活用による最適実験計画？



次の実験計画へfeedback

既知の知見・
観測(データ)

高速・高精度な
Data-Driven予測

結果の確認と
検証

仮説形成

(機械学習・データマイニング)

- どういう実験・シミュレーションを次に行うかの計画立案
- 時間のかかる計算の高精度高速近似
- 曖昧な因子や実験条件の最適化
- Multilevelの情報統合

仮説検証

(シミュレーション+実験)

- 再現性を担保する高精度・高速実験系
- 仮想化検証が可能な因子のシミュレーション(計算科学)による探索
→ 望ましい対象のさらなる絞り込み

2000-2010年頃から創薬/生命科学で先行

NATURE REVIEWS | DRUG DISCOVERY
VOLUME 17 | FEBRUARY 2018 | 97

PERSPECTIVES

INNOVATION

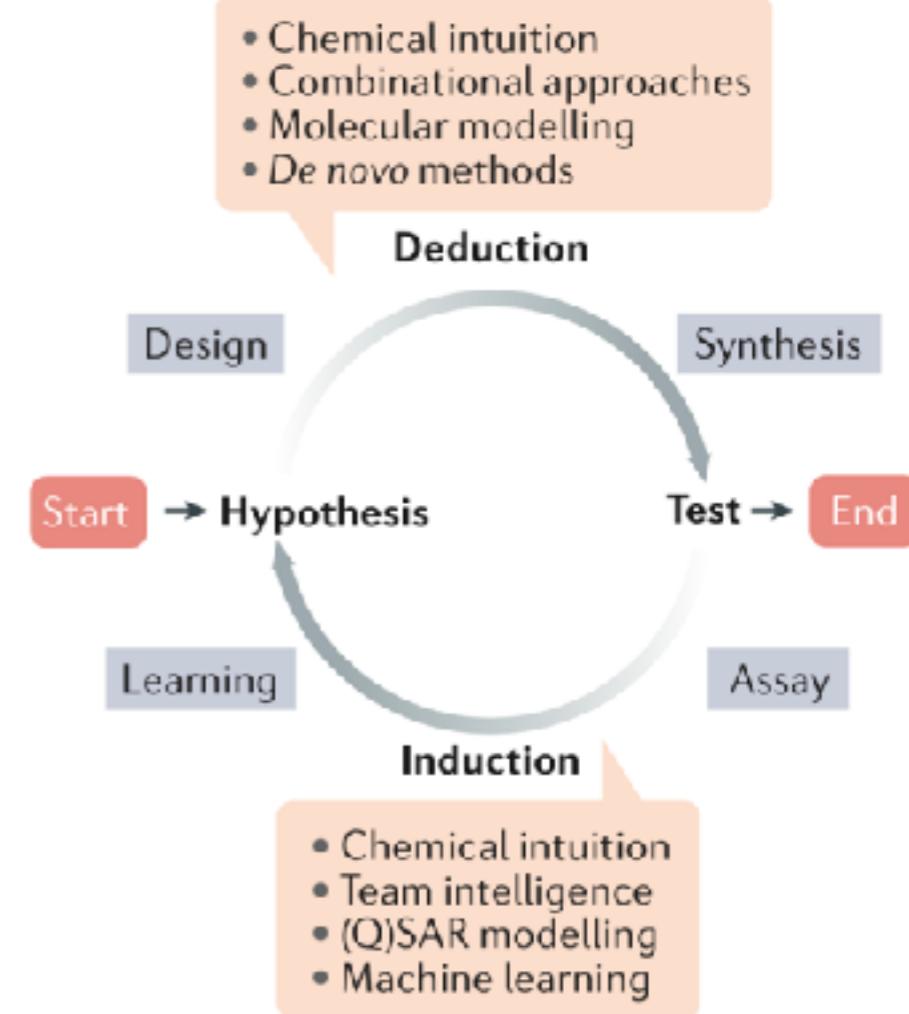
Automating drug discovery

Gisbert Schneider



Figure 2 | Automated drug discovery facilities. a | Millions of compound samples are stored in compact high-capacity facilities and handled by robots. b | Robot systems perform both high-throughput and medium-throughput screening of up to ten thousand samples per day to determine the activity against the biological target of interest. Multiple arms and flexible workstations enable fully automated liquid dispensing, compound

preparation and testing. These storage and screening systems have become cornerstones of contemporary drug discovery. c | A prototype of a novel miniaturized design-synthesize-test-analyse facility for rapid automated drug discovery at AstraZeneca is shown. Images a and b courtesy of Jan Kriegel, Boehringer-Ingelheim Pharma; Image c courtesy of Michael Kossenjans, AstraZeneca.



材料開発へも展開

Toyota teams with China's CATL and BYD to power electric ambitions

Automaker diversifies battery source and moves up electrification goal by 5 years

YUKIHIRO OMOTO, Nikkei staff writer

JUNE 07, 2019 02:00 JST • UPDATED ON JUNE 07, 2019 14:39 JST



↓品質を担保する大規模化の技術的必須要素：
製造ラインに人が(ほとんど)いない



CATL

研究開発体制

材料開発、セルデザインにおいて
最先端シミュレーション技術を活用

生産体制

完全自動化によるフレキシブルな生産
IoTやビッグデータを活用した生産管理

材料開発へも展開

Toyota teams with China's CATL and BYD to power electric ambitions

Automaker diversifies battery source and moves up electrification goal by 5 years

YUKIHIRO OMOTO, Nikkei staff writer

JUNE 07, 2019 02:00 JST • UPDATED ON JUNE 07, 2019 14:39 JST



↓品質を担保する大規模化の技術的必須要素：
製造ラインに人が(ほとんど)いない



CATL

研究開発体制

材料開発、セルデザインにおいて
最先端シミュレーション技術を活用

生産体制

完全自動化によるフレキシブルな生産
IoTやビッグデータを活用した生産管理

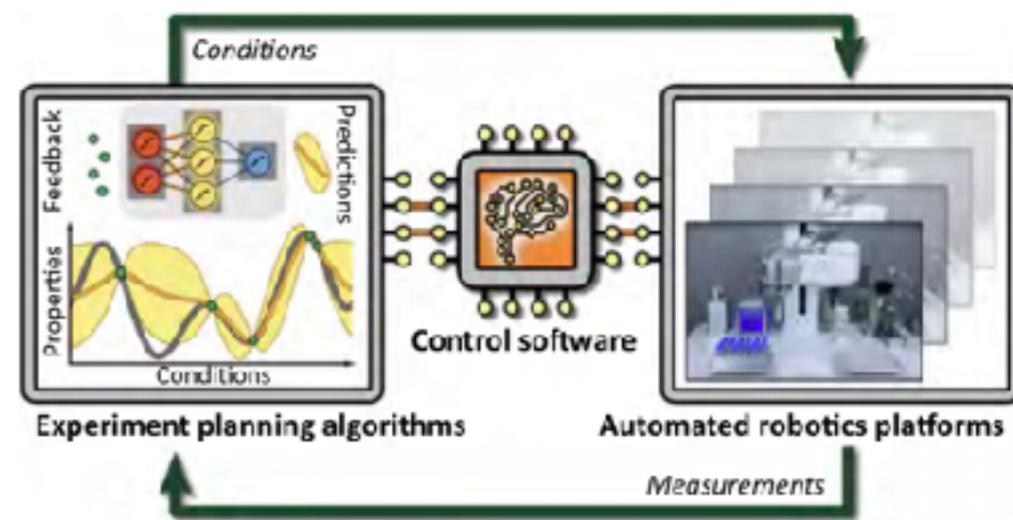
化学にも波及

Trends in Chemistry, June 2019, Vol. 1, No. 3 [10.1016/j.trechm.2019.02.007](https://doi.org/10.1016/j.trechm.2019.02.007)

Opinion

Next-Generation Experimentation with Self-Driving Laboratories

Florian Häse,^{1,2,3,4} Loïc M. Roch,^{1,2,3,4} and Alán Aspuru-Guzik^{1,2,3,4,5,*}



Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101
German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski*

Angew. Chem. Int. Ed. 2016, 55, 5904–5937

Angewandte
Chemie
International Edition

Machine-Assisted Chemistry Special Issue 150 Years of BASF

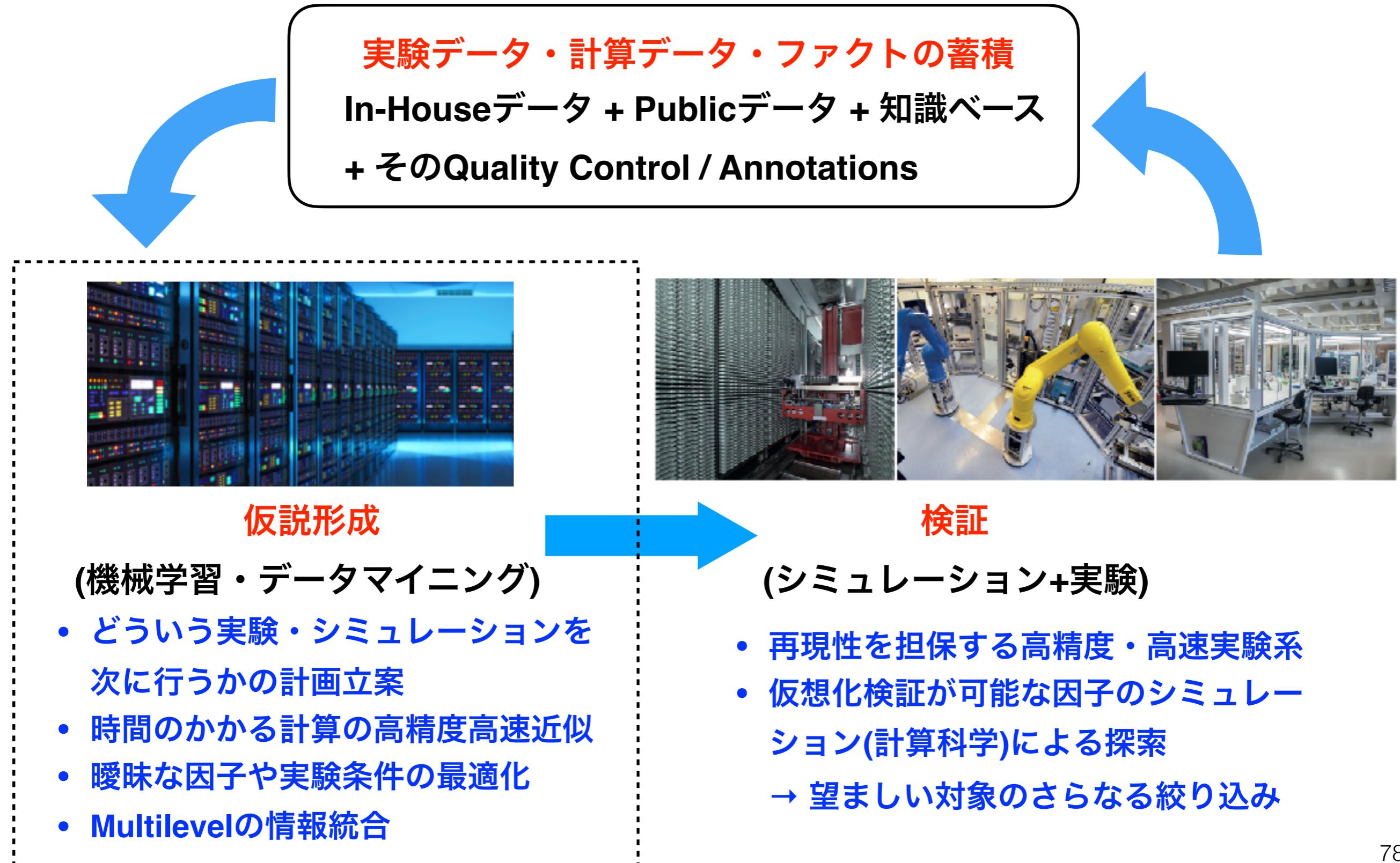
DOI: 10.1002/anie.201410744

Organic Synthesis: March of the Machines

Steven V. Ley,* Daniel E. Fitzpatrick, Richard J. Ingham, and Rebecca M. Myers

Angew. Chem. Int. Ed. 2015, 54, 3449–3464

Key：蓄積された「計算・実験データ」の利活用



「データ利活用技術」は科学研究の道具の一つに

Science is changing, the tools of science are changing. And that requires different approaches. Eric Bloch, 1925-2016

Nature, 559
pp. 547–555 (2018)

REVIEW

<https://doi.org/10.1038/s41586-018-0337-2>

Machine learning for molecular and materials science

Keith T. Butler¹, Daniel W. Davies², Hugh Cartwright³, Olexandr Isayev^{4*} & Aron Walsh^{5,6*}

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

Science, 361
pp. 360-365 (2018)

SPECIAL SECTION FRONTIERS IN COMPUTATION

REVIEW

Inverse molecular design using machine learning: Generative models for matter engineering

Benjamin Sanchez-Lengeling¹ and Alán Aspuru-Guzik^{2,3,4*}

The discovery of new materials can bring enormous societal and technological progress. In this context, exploring completely the large space of potential materials is computationally intractable. Here, we review methods for achieving inverse design, which aims to discover tailored materials from the starting point of a particular desired functionality. Recent advances from the rapidly growing field of artificial intelligence, mostly from the subfield of machine learning, have resulted in a fertile exchange of ideas, where approaches to inverse molecular design are being proposed and employed at a rapid pace. Among these, deep generative models have been applied to numerous classes of materials: rational design of prospective drugs, synthetic routes to organic compounds, and optimization of photovoltaics and redox flow batteries, as well as a variety of other solid-state materials.

act properties. In practice, approximations are used to lower computational time at the cost of accuracy.

Although theory enjoys enormous progress, now routinely modeling molecules, clusters, and perfect as well as defect-laden periodic solids, the size of chemical space is still overwhelming, and smart navigation is required. For this purpose, machine learning (ML), deep learning (DL), and artificial intelligence (AI) have a potential role to play because their computational strategies automatically improve through experience (1).

In the context of materials, ML techniques are often used for property prediction, seeking to learn a function that maps a molecular material to the property of choice. Deep generative models are a special class of DL methods that seek to model the underlying probability distribution of both structure and property and relate them in a nonlinear way. By exploiting patterns in massive datasets, these models can distill average and salient features that characterize molecules (2, 3).

Inverse design is a component of a more complex materials discovery process. The time

Science, 293
pp. 2051-2055 (2001)

VIEWPOINT

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

Machine learning (ML) (1) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business, and science. A generalized version of the stan-

correlate surprisingly well with subsequent gene expression analysis (3). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (4), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to ameliorate every aspect of a working scientist's

creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semi-automatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern recognition. A recent example is event analysis

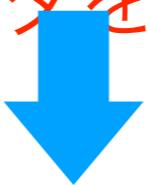
一方で、生命科学でも得られた教訓として、（お金もかかるので、）
実効性をともなう方式の確立にはまだ要素技術の改良と「良い」データの蓄積が必要

"low input, high throughput, no output science." (Sydney Brenner)

→ 雜な設定・系で網羅的なハイスループット実験をいくらしても何も得られない

「(筋の良い)仮説形成」と機械学習・データマイニング

- 自動化のもう一つの恩恵は、「再現性」「結果の質」の担保
属人性が残っているとデータの質にも(予測にも)ばらつきが生じる
- 近年の自動化で「高速にできるだけたくさん試す」のは
探索を効率化する王道だが、考えられる候補がほぼ無限にありえる
ので、「何を試すか」の選択の問題は残る(全部は試せない...)
- 自動化をするかしないかによらず、実験でも計算でも、筋の良い
ターゲット、実験条件、パラメタを決めるステップはボトルネック

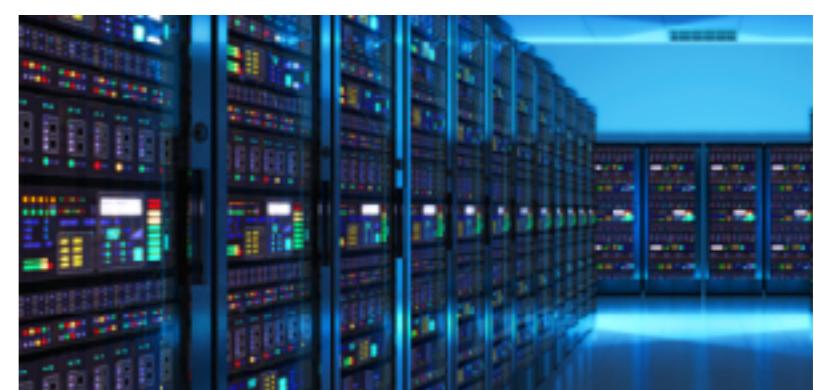


実験データ・計算データ・ファクトの蓄積

In-Houseデータ + Publicデータ + 知識ベース

+ そのQuality Control / Annotations)

+



機械学習・データマイニング

モデルベース最適化(代理モデル最適化)

航空宇宙機のような流体機械設計など、計算時間がかかるシミュレーションを用いた設計最適化技術として発展

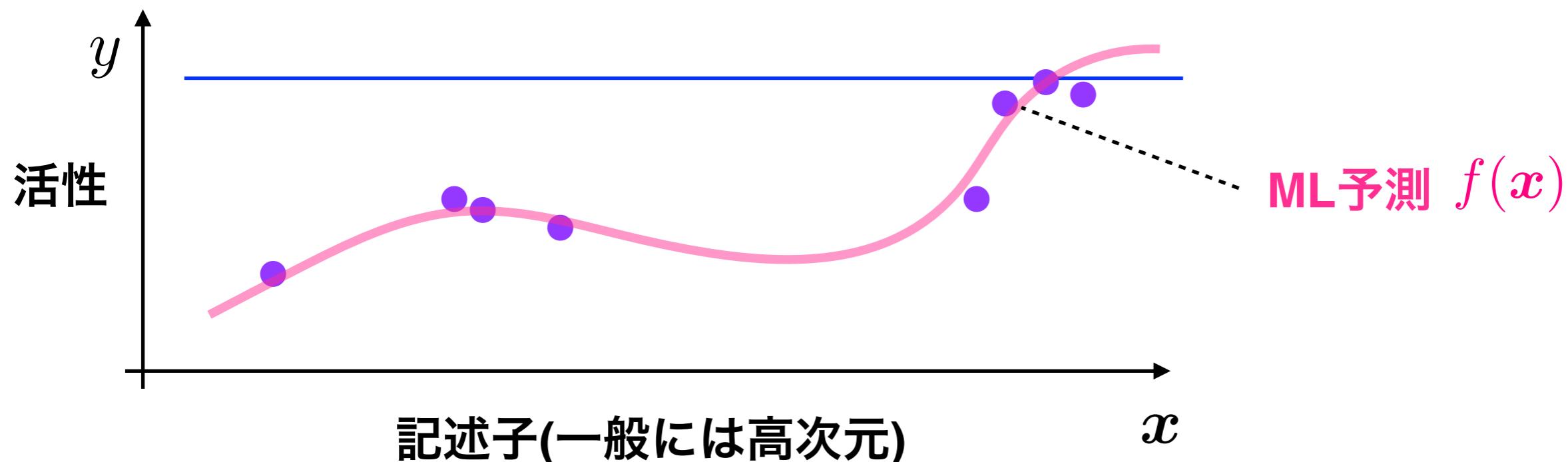
計算時間がかかるシミュレーションの代理(surrogate)
として、機械学習 $x \rightarrow y$ を活用する

- | | |
|---------------------------------|---|
| 1. Initial Sampling | 実験計画
e.g.
Latin hypercube sampling (LHS) |
| 2. Loop: | 機械学習
適応サンプリング
e.g.
Expected improvement (EI) |
| 1. Construct a Surrogate Model. | |
| 2. Search Infill Criterion. | |
| 3. Add new samples. | |

Recent advances in surrogate-based optimization (Forrester & Keane, 2009)
<https://doi.org/10.1016/j.paerosci.2008.11.001>

Infill基準と最適実験計画

"exploitation" と "exploration" のバランス制御にもMLを用いる

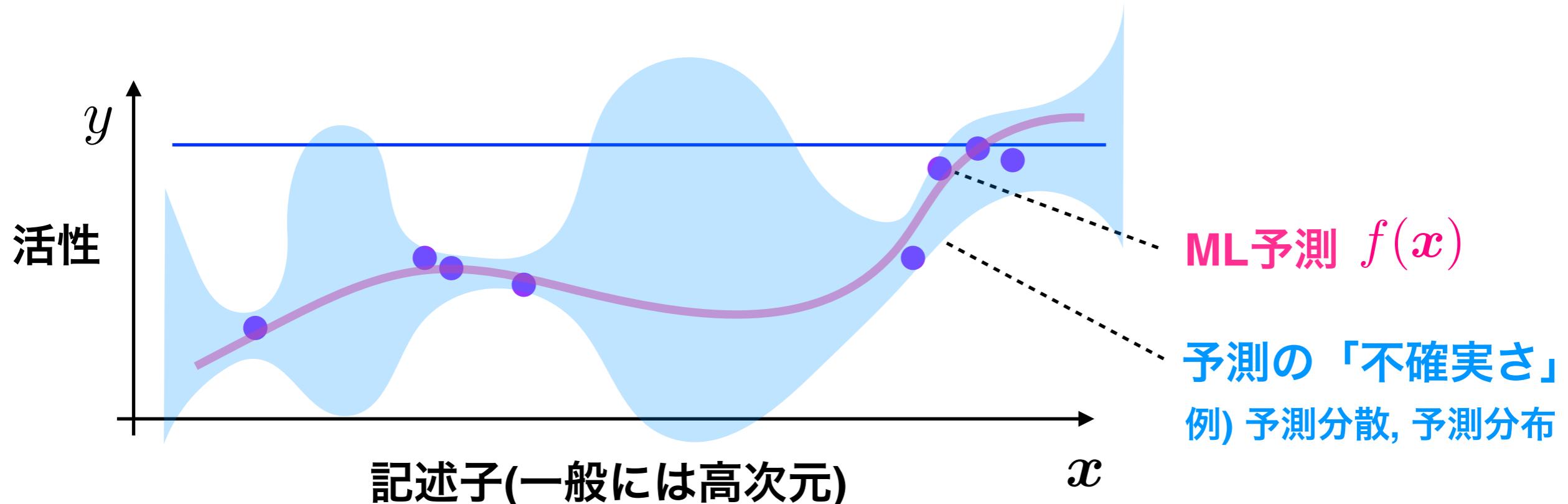


どうやってバランス制御するのかには色々なやりかたがあり Hot Topics

- 強化学習 + 探索
- ブラックボックス最適化
- ベイズ最適化
- 逐次実験計画
- アクティブラーニング
- 多腕バンディット
- 進化計算
- ゲーム理論 (CFRなど)

Infill基準と最適実験計画

"exploitation" と "exploration" のバランス制御にもMLを用いる

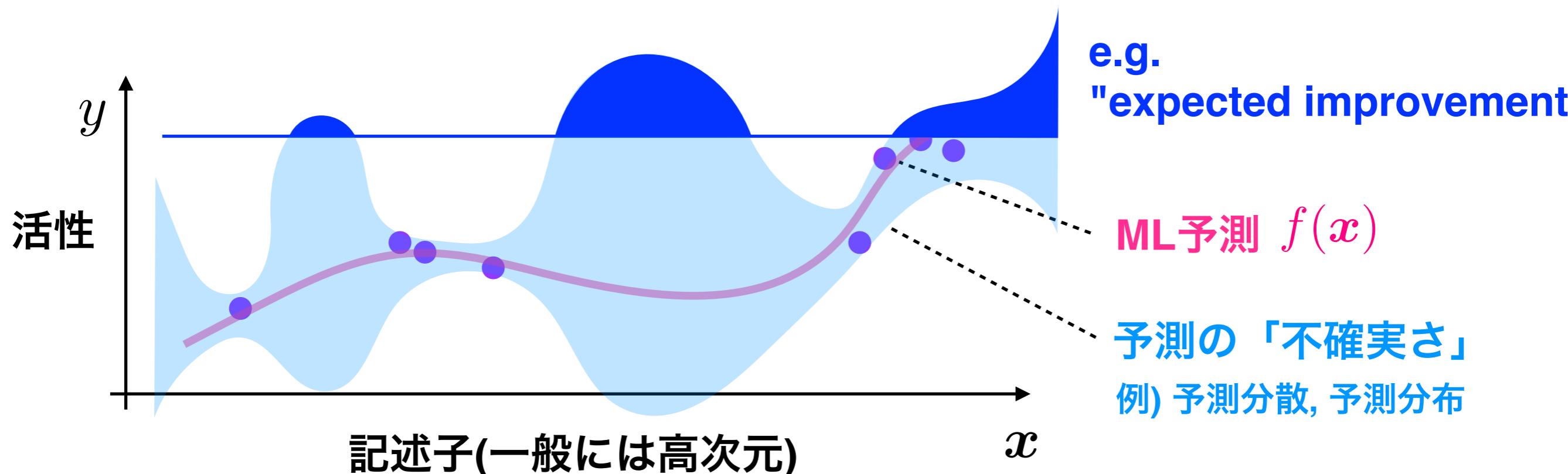


どうやってバランス制御するのかには色々なやりかたがあり Hot Topics

- 強化学習 + 探索
- ブラックボックス最適化
- ベイズ最適化
- 逐次実験計画
- アクティブラーニング
- 多腕バンディット
- 進化計算
- ゲーム理論 (CFRなど)

Infill基準と最適実験計画

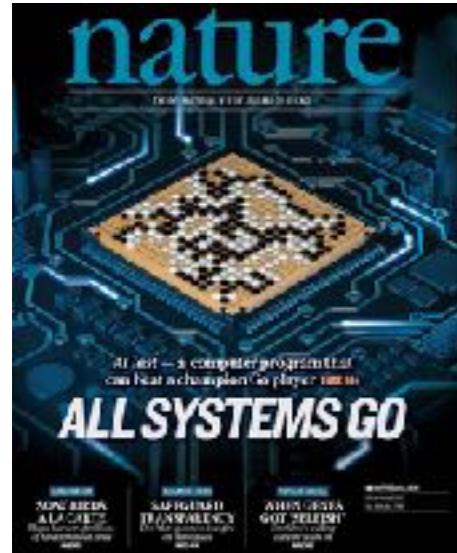
"exploitation" と "exploration" のバランス制御にもMLを用いる



どうやってバランス制御するのかには色々なやりかたがあり Hot Topics

- 強化学習 + 探索
- ブラックボックス最適化
- ベイズ最適化
- 逐次実験計画
- アクティブラーニング
- 多腕バンディット
- 進化計算
- ゲーム理論 (CFRなど)

機械學習分野自体でもHot Research Topic



AlphaGo
(Nature, Jan 2016)

ARTICLE

Mastering the game of Go without human knowledge

David Silver^{1,2*}, Julian Schrittwieser^{3,4}, Karen Simonyan¹, Thomas Hubert¹, Lucas Baker¹, Matthew Lai¹, Daan Hendrikse¹, Valerii Denysenko¹, Timothy Lillicrap¹, David Hardcastle¹, Tomáš Pávlik¹, Oriol Vinyals¹, Daan Wierwille¹, Dharshan Kumara¹, Thore Graepel¹, Timothy Lillicrap¹, Karen Simonyan¹, Demis Hassabis¹

A long-standing goal of artificial intelligence is an algorithm that learns to play more superhuman Go without any explicit knowledge of the game. In this article, we present AlphaZero, a self-play reinforcement learning algorithm that masters Go without any knowledge of the game. The algorithm uses a general reinforcement learning framework that combines two main ideas. First, it performs self-play, where the algorithm learns to play Go against itself, without any knowledge of the game rules. Second, it uses a learned model to predict the outcome of Go games, which allows the algorithm to evaluate moves and select the best ones. The algorithm is trained to play Go without any knowledge of the game rules, and it achieves superhuman performance in less than 40 hours of self-play. The algorithm also achieves superhuman performance in chess and shogi, and it is the first AI program to learn all three games from scratch.

AlphaZero is the first AI program to learn Go from scratch, without any knowledge of the game rules. The algorithm uses a general reinforcement learning framework that combines two main ideas. First, it performs self-play, where the algorithm learns to play Go against itself, without any knowledge of the game rules. Second, it uses a learned model to predict the outcome of Go games, which allows the algorithm to evaluate moves and select the best ones. The algorithm is trained to play Go without any knowledge of the game rules, and it achieves superhuman performance in less than 40 hours of self-play. The algorithm also achieves superhuman performance in chess and shogi, and it is the first AI program to learn all three games from scratch.

AlphaGo Zero
(Nature, Oct 2017)

Silver et al., *Science* **362**, 1140–1144 (2018)

7 December 2018

Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model

Julian Schrittwieser,^{1*} Ioannis Antonoglou,^{1,2*} Thomas Hubert,¹ Karen Simonyan,¹ Laurent Sifre,¹ Simon Schmitt,¹ Arthur Guez,¹ Edward Lockhart,¹ Demis Hassabis,¹ Thore Graepel,^{1,2} Timothy Lillicrap,¹ David Silver^{1,2}

¹DeepMind, 6 Pancras Square, London WC1E 4AG,
²University College London, Gower Street, London WC1E 6BT.
*These authors contributed equally to this work.

Abstract

Controlling agents with planning capabilities has long been one of the main challenges in the field of artificial intelligence. Tree-based planning methods have enjoyed large success in challenging domains, such as chess and Go, where a perfect simulator is available. However, in real-world problems the dynamics governing the environment are often complex and unknown. In this work, we present the AlphaZero algorithm, which, by combining a tree-based search with a learned model, achieves superhuman performance in a range of challenging and visually complex domains, without any knowledge of their underlying dynamics. AlphaZero is a model that, when applied iteratively, provides the quantities most directly relevant to planning: the reward, the action-selection policy, and the value function. When evaluated on 27 different Atari games – the evaluation space provides a convenient testbed for testing AI techniques, in which model-based planning approaches have historically struggled – our new algorithm achieves new state-of-the-art. When extended to Go, chess and shogi, without any knowledge of the game rules, AlphaZero matches the superhuman performance of the AlphaGoZero algorithm that was supplied with the game rules.

AlphaZero
(Science, Dec 2018)

AlphaZero
(arXiv, Nov 2019)

AutoML (全自動機械學習)

- Algorithm Configuration
- Hyperparameter Optimization (HPO)
- Neural Architecture Search (NAS)
- Meta Learning / Learning to Learn



Cloud AutoML



AutoML



AutoDL 2019



Amazon
SageMaker

例) Model-based RL (Toward sample-efficient RL)



最適制御 (Optimal Control)

対象の動的システムの挙動(物理法則など)がわかっている場合、最良行動を決定可能

何もわからない場合 (Model-free RL)

実際に環境から得られる行動・状態系列から直接的に方策や価値関数を推定する

少し当たりがつけられる(?)場合 (Model-based RL or 古典的なシステム同定の設定)

実際の行動・状態系列からまず動的システムの挙動を推定し、その推定したモデルを用いて最適行動を計画する (e.g. 将棋するとき相手の手を頭の中でシミュレートする)

Planning

例) Model-based RL or Planning with Models

Deep Planning Network (PlaNet)

Hafner+ Learning Latent Dynamics for Planning from Pixels.
arXiv:1811.04551 (Jun 2019)

MuZero

Schrittwieser,+ Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model.
arXiv:1911.08265 (Nov, 2019)

Simulated Policy Learning (SimPLe)

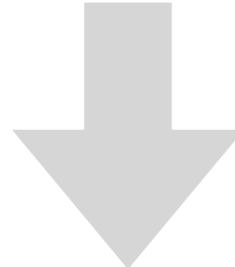
Kaiser+ Model-Based Reinforcement Learning for Atari.
arXiv:1903.00374 (Jun 2019)

Stochastic Latent Actor-Critic (SLAC)

Lee+ Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model.
arXiv:1907.00953 (Jul 2019)

機械学習できたときの2種類の期待

1. 得られた変換過程(関数)による予測を色々な目的に使う
2. 得られた変換過程(関数)を分析して背景過程の仕組みを知る



1.について

- 問題が内挿的になるよう工夫 (表現学習・隠れ構造同定)
- 内挿・外挿判定 (予測の信頼度計算)
- モデルベース最適化と探索 (最適実験計画)

2.について

- ポストホック解析と解釈性モデル (学習済みモデル分析)

(因果の理解は諦めて?)解釈/説明/仮説生成へ

Explainable AI (XAI), Interpretable ML, Causal ML

- 米DARPAのExplainable AI (XAI)プログラム
 - 機械学習業界におけるInterpretable ML
 - CausalML: 機械学習 for Causal Inference, Counterfactual Prediction, and Autonomous Action

「解釈」 vs 「理解」

人(手法)の 真実は
数だけある ひとつ?



各手法によって異なる仮説形成や示唆の提供

背後の真の法則に関する情報が得られるとは限らないので注意

私のブックマーク：機械学習における解釈性 (原 聰, 人工知能 33(3), 366-369, 2018年5月)

機械学習モデルの解釈、機械学習による解釈

- $x \rightarrow y$ の関数のPost-hoc解析
 - 回帰分析における要因分析
(回帰係数の有意性検定)
 - ベイズ予測分布
 - 変数重要度・部分従属度plot
 - グローバル/ローカル感度解析 (Sobol'法, FAST法, etc)
 - 深層学習におけるSaliencyやAttentionの利用
 - 局所説明生成: LIME (KDD16), SHAP (NIPS17)
- $x \rightarrow y$ の階層的分解による隠れ因子や階層構造の同定
 - **深層学習**
 - 計算グラフとして表現 (汎用的パラメタ推定: 逆モード自動微分)
 - **確率的プログラミング (生成的統計モデリング)**
 - 確率変数の階層的生成関係で表現 (汎用的パラメタ推定: MCMC/自動VI)

統計学手法は要前提の検証

- 説明変数の選択
- 線形の仮定
- 多重共線形性 "マルチコ"
- 残差の検討
- ⋮

今日の内容

1. イントロ

機械学習と科学(あるいは"ものづくり")

2. 機械学習で何かを「理解」できるか？

Answer: 直接的には原理上困難

3. 機械学習で何かを「発見」できるか？

Answer: 直接的には原理上困難

4. じゃあどうすんの！？何がいるの！？

Answer: 「表現」と「介入」

2と3を前提に機械学習分野のトピックを簡単に紹介

再考 統計的理解と科学の文法

科学の文法 (1892)

"Statistics is the grammar of science." (Karl Pearson)

現在の科学的疑問の多くは**100%YES/NOな答えが無い**問い！

夏目漱石より10才年上の大統計学者
↓

- ・ この薬を飲めば私の病気は治るの？
- ・ この健康食品食べていれば長生きできるの？
- ・ この化粧品つけていれば少しでも若くいられるの？
- ・ この食品たべればダイエットできるの？
- ・ 原子力は安全なの？

YES and NOの
間にきっと真実が

科学は「データの見方」と無縁ではいられない！

科学というものには、本来限界があって、広い意味での再現可能の現象を、自然界から抜き出して、それを統計学的に究明していく、そういう性質の学問なのである。「科学の方法 (中谷宇吉郎)」

Impossible to model everything...?



大統計学者 **George E. P. Box (1919-2013)**

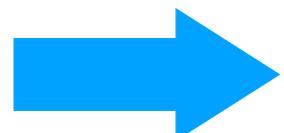
"one of the great statistical minds of the 20th century"

**"Essentially, all models are wrong,
but some are useful"**

https://en.wikipedia.org/wiki/All_models_are_wrong

モデルとは何かを捨象したものであり実世界(複雑系)とは違う。

- 真の法則が人間に理解可能なほどシンプルなモデルに要素還元できる保証はどこにもない。
- データが有限ならそれを説明できるモデルは一般に無数にある。



論述的理解/要素還元から複雑系の"統計的理解"へ?

展望：Theory-driven vs Data-drivenの解消と融合

Theory-driven 【合理論】

- 対象現象の複雑化
- シミュレーション技法も複雑化
- "経験的に決める"パラメタや初期値
- 汎関数、交換相関項の設計

(人工知能分野)

- 知識ベースと論理推論(記号AI)の限界
- 厳密推論や探索の計算爆発(NP困難性)
- 大量データの知識化の問題
- 制約プログラミングや組合せ最適化

データ同化、模倣学習、論理合成、etc

モデルベース最適化、強化学習、メタ
学習、ドメイン適応、生成モデル、etc

→ 新たな方法論へ？

Data-driven 【経験論】

- 小サンプル・低カウントの問題
- 外挿の不可能性の問題
- 帰納バイアスのモデルエンコード
- Blackbox性・解釈性の問題

(人工知能分野)

- Data-Driven手法(機械学習)と人間の
論理的思考との大きなギャップ
- Dataがない領域の探索や「ひらめき」
- モデル適用範囲と信頼性・安全性

世界トップレベル研究拠点プログラム（WPI）



World Premier International Research Center Initiative

-Science-
世界最高水準の研究

-Globalization-
国際的な研究環境の実現

4つのミッションの達成により
世界トップレベル研究拠点を構築

-Reform-
研究組織の改革

-Fusion-
融合領域の創出

WPIアカデミー拠点
【2007年度採択 5拠点】

東北大学
材料科学高等研究所 (AIMR)
小谷 元子

物質・材料研究機構
国際ナノ・キエクトニクス研究拠点 (MANA)
佐々木 高麗

京都大学
物質一細胞統合システム拠点 (iCeMS)
北川 進

大阪大学
免疫学フロンティア研究センター (IFReC)
蕨良 静明

東京大学
カブリ数物連携宇宙研究機構 (Kavli IPMU)
大栗 博司

※10年間の支援期間終了後、更に5年間の補助金支援期間延長が認められている。

最大7億円/年×10年



【2018年度採択 2拠点】

北海道大学
化学反応創成研究拠点 (ICReDD)
前田 理

京都大学
ヒト生物学高等研究拠点 (ASHBi)
斎藤 通紀

補助金支援中の拠点
【2010年度採択 1拠点】

九州大学
カーボンニュートラル・エネルギー国際研究所 (I²CNER)
Petros Sofronis

【2012年度採択 3拠点】

筑波大学
国際統合睡眠医科学研究機構 (IIIS)
柳沢 正史

東京工業大学
地球生命研究所 (ELSI)
関根 敏

名古屋大学
トランスマテイフ生命分子研究所 (TMbM)
伊丹 健一郎

【2017年度採択 2拠点】

東京大学
ニコ-ロインテリジェンス国際研究機構 (IRCN)
Takao Hensch

金沢大学
ナノ生命科学研究所 (NanoLSI)
福間 利士

北海道大学 化学反応創成研究拠点(IReDD)



拠点の機関技術：化学反応経路の自動探索

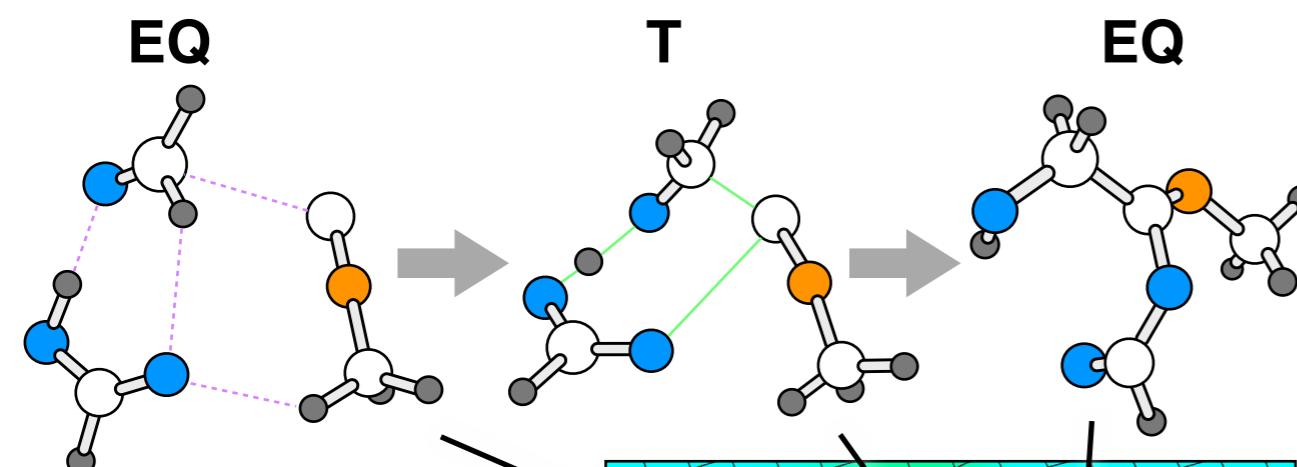
PERSPECTIVE

[View Article Online](#)
[View Journal](#) | [View Issue](#)

Cite this: *Phys. Chem. Chem. Phys.*, 2013,
15, 3683

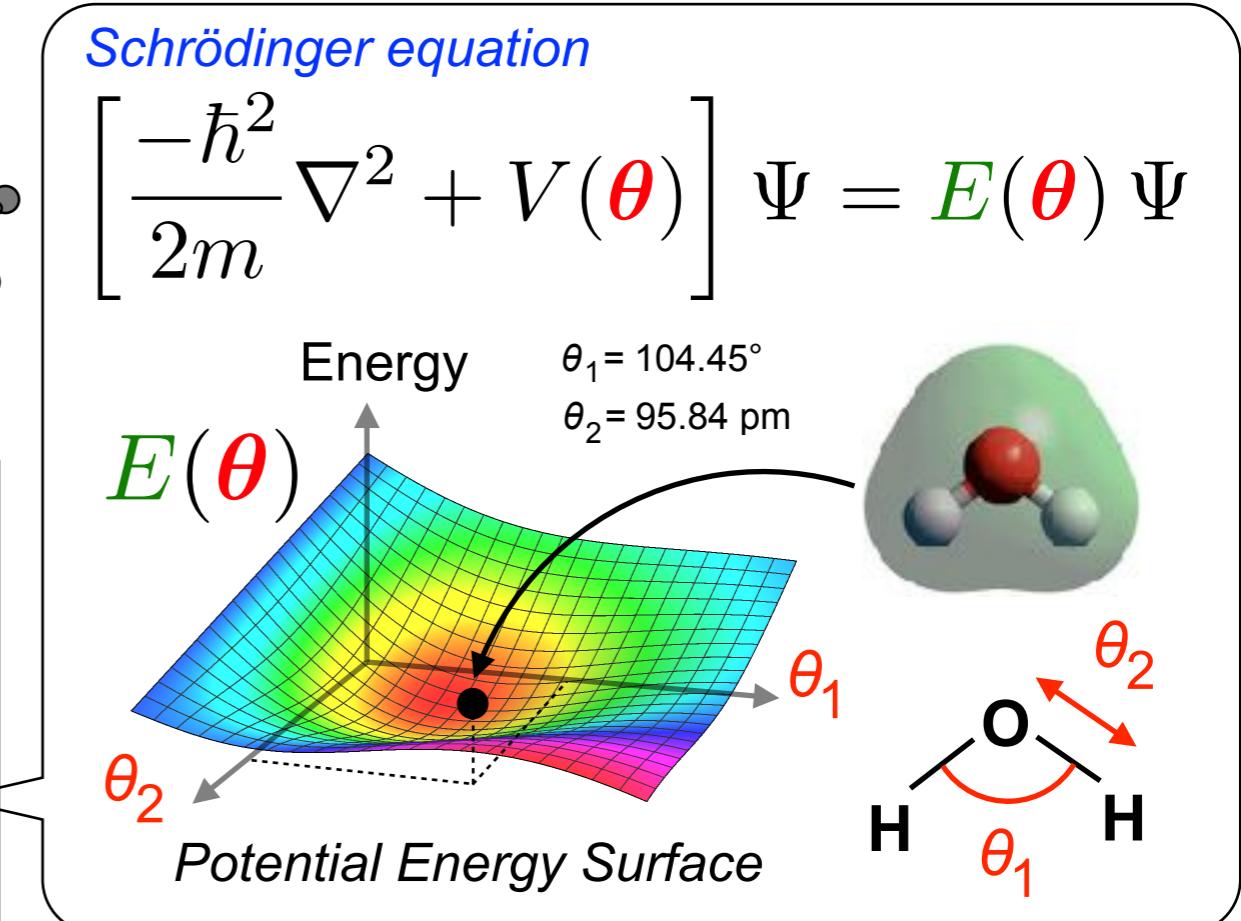
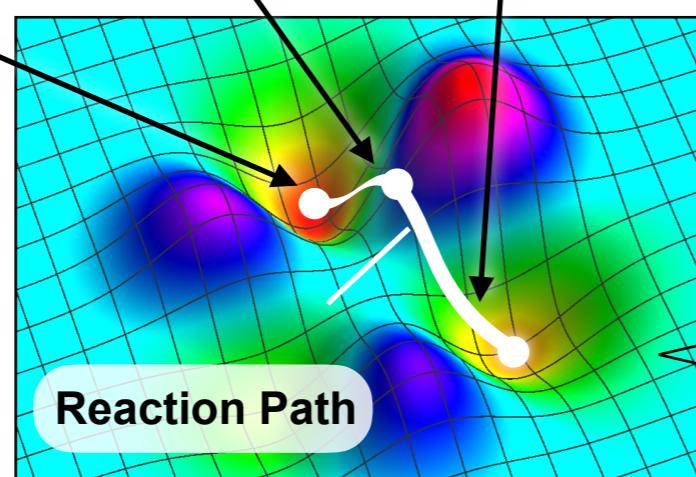
Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods

Satoshi Maeda,*^a Koichi Ohno*^b and Keiji Morokuma*^{cd}



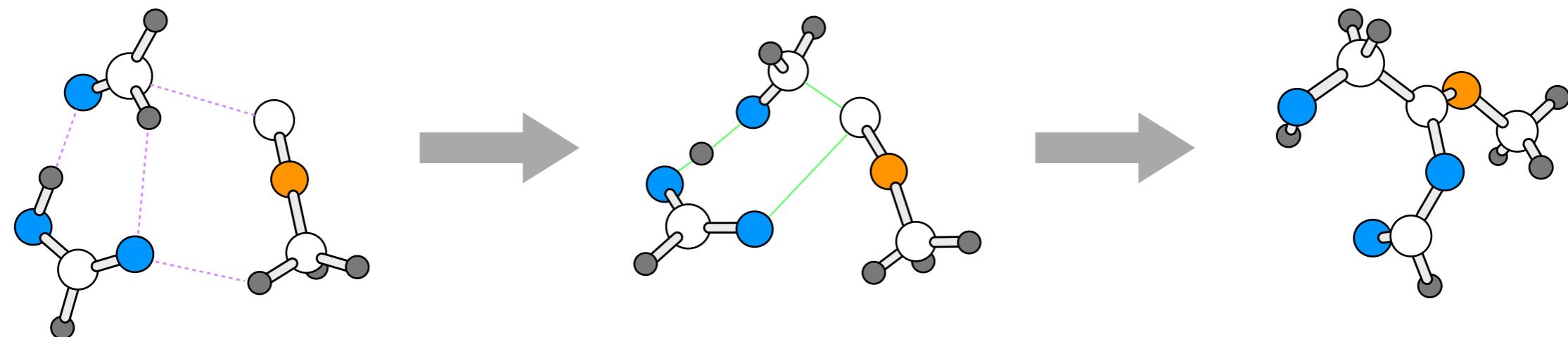
AFIR
Maeda & Morokuma, *J
Chem Phys*, 2010

ADDF
Ohno & Maeda, *Chem
Phys Lett*, 2004



化学反応の設計と探索

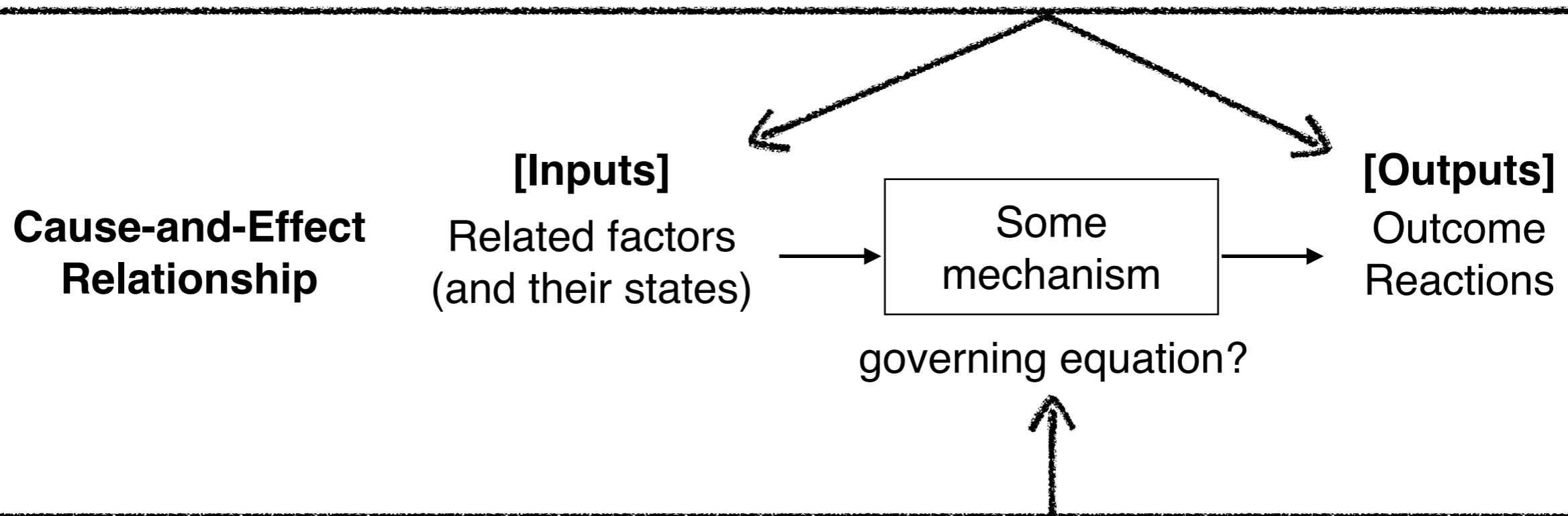
Chemical reactions = recombinations of atoms and chemical bonds
subjected to *the laws of nature*



- **Intractably large chemical space:** A intractably large number of "theoretically possible" candidates for reactions and compounds...
- **Scalability issue:** Simulating an Avogadro-constant number of atoms is utterly infeasible... (After all, we need some compromise here)
- **Complexity and uncertainty of real-world systems:**
Many uncertain factors and arbitrary parameters are involved...
- **Known and unknown imperfections of currently established theories:**
Current theoretical calculations have many exceptions and limitations...

化学反応の設計と探索

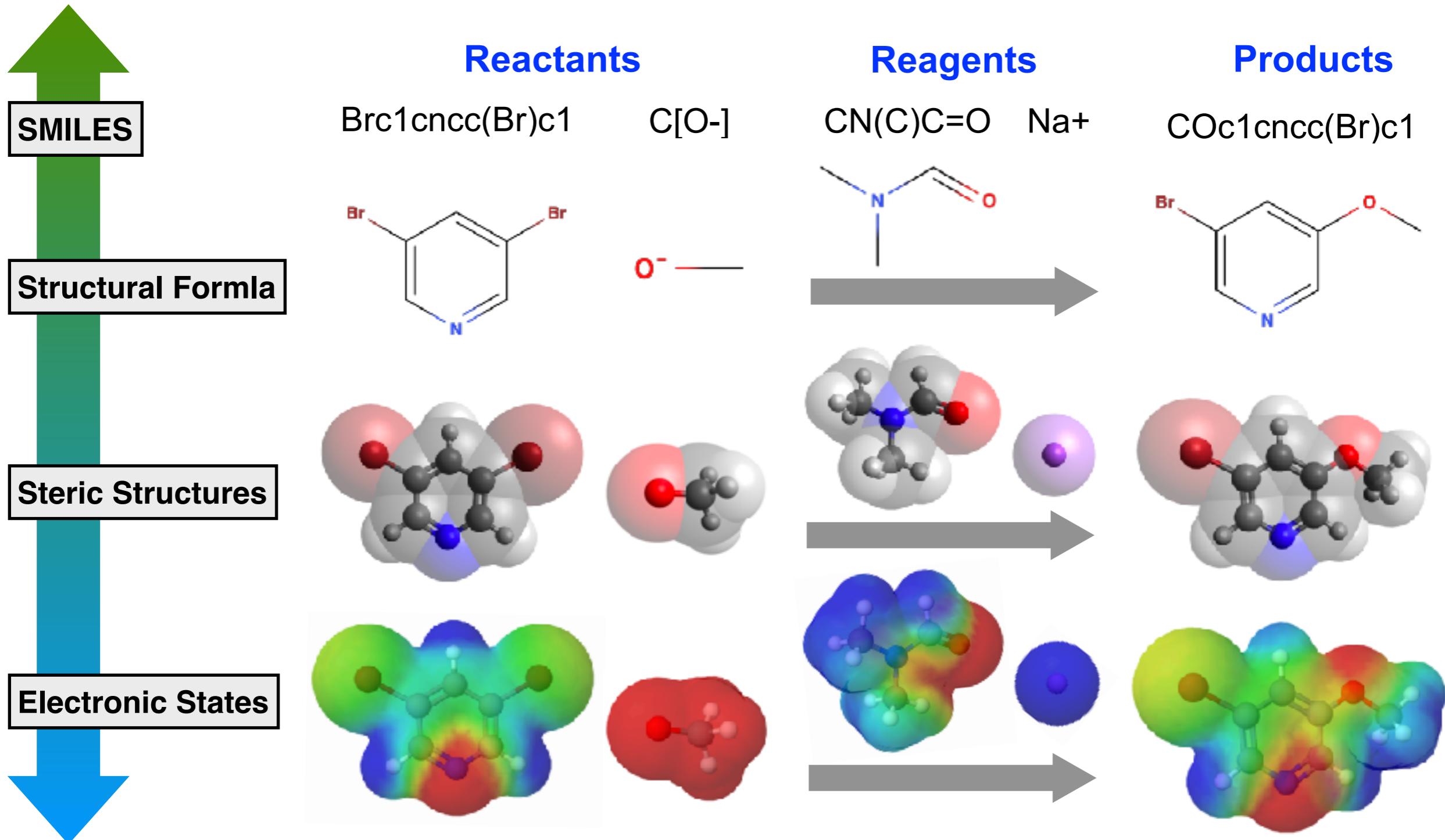
Data-driven methods try to precisely approximate its outer behavior (the input-output relationship) observable as "data".
(e.g. through *machine learning* from a large collection of data)



Theory-driven methods try to explicitly model the inner workings of a target phenomenon (e.g. through first-principles simulations)

化学反応の設計と探索

As pattern languages (e.g. known facts in textbooks/databases)



As physical entities (e.g. quantum chemical calculations)

化学反応の設計と探索

Computer-assisted synthetic planning
(path search on knowledge bases)

**or AI-Assisted Synthesis?
(with Machine Learning)**



Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101

German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald,
Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski*

Angew. Chem. Int. Ed. **2016**, *55*, 5904–5937



AI-Assisted Synthesis

Very Important Paper

International Edition: DOI: 10.1002/anie.201912083

German Edition: DOI: 10.1002/ange.201912083

Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning

Tomasz Badowski, Ewa P. Gajewska, Karol Molga, and Bartosz A. Grzybowski*

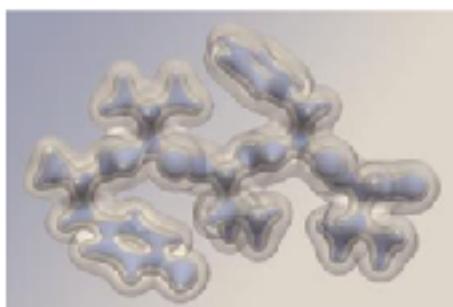
Angew. Chem. Int. Ed. **2019**, *58*, 1–7



CHEMISTRY WORLD



All machine learning articles

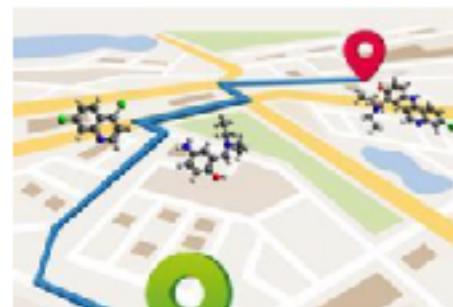


RESEARCH

Machine learning predicts electron densities with DFT accuracy

2 OCTOBER 2019

Non-covalent interactions and electron densities can be explored quickly without the need for expensive and time-consuming quantum chemical calculations



RESEARCH

Language-based software's accurate predictions translate to benefits for chemists

30 SEPTEMBER 2019

State-of-the-art design for computer language processing results in improved models for predicting chemistry

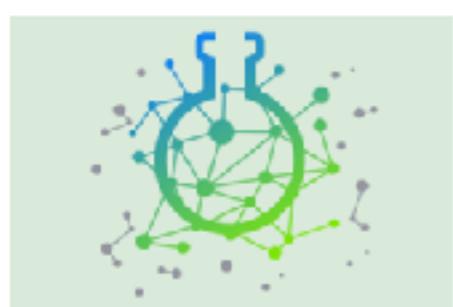


RESEARCH

Human biases cause problems for machines trying to learn chemistry

13 SEPTEMBER 2019

Including 'unpopular' reagents and reaction conditions into datasets could lead to better machine-learning models



RESEARCH

Retrosynthetic algorithm broadened to design similar, but different, molecules

26 AUGUST 2019

Chematica can now design efficient syntheses for large compound libraries

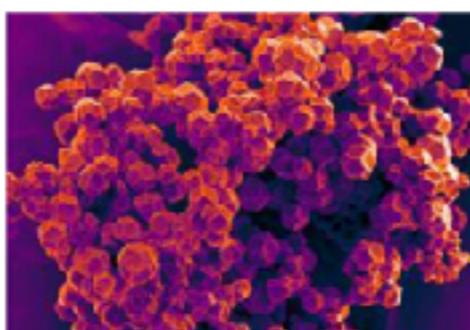


RESEARCH

Are synthetic chemists out of a job as AI meets automation?

9 AUGUST 2019

Platform can weigh up a synthetic route, plan it and then carry out it

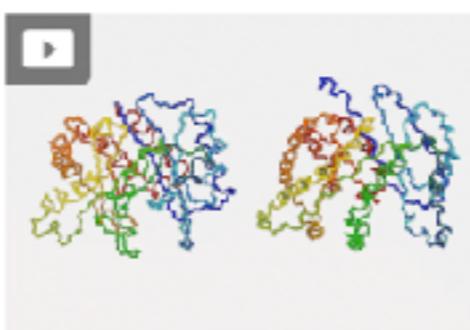


RESEARCH

Algorithm accurately predicts mechanical properties of existing and theoretical MOFs

17 MAY 2019

Machine learning could speed up the production and use of coordination polymers in industry



RESEARCH

Neural network folds proteins a million times faster than its competitors

8 MAY 2019

Machine learning algorithm that predicts protein structures in milliseconds could top next protein folding contest



RESEARCH

Dispute over reaction prediction puts machine learning's pitfalls in spotlight

18 DECEMBER 2018

Two research teams' argument over a reaction-predicting algorithm show that there is still a lot to understand when applying machine learning to chemistry

化学反応の設計と探索

ML-based chemical reaction predictions

<i>Graph NN</i>	<i>Sequence NN</i>	<i>Combined or Other</i>
WLDN Jin+ <i>NeurIPS</i> 2017	seq2seq Liu+ <i>ACS Cent Sci</i> 2017	Neural-Symbolic ML Segler+ <i>Chemistry</i> 2017
ELECTRO Bradshaw+ <i>ICLR</i> 2019	IBM RXN Schwaller+ <i>Chem Sci</i> 2018	Similarity-based Coley+ <i>ACS Cent Sci</i> 2017
GPTN Do+ <i>KDD</i> 2019	Molecular Transformer Schwaller+ <i>ACS Cent Sci</i> 2019	3N-MCTS/AlphaChem Segler+ <i>Nature</i> 2018
WLN Coley+ <i>Chem Sci</i> 2019		Molecule Chef Bradshaw+ <i>DeepGenStruct (ICLR WS)</i> 2019

ML + First-principle simulations

Fermionic Neural Network

Pfau+ Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks.
arXiv:1909.02487, Sep 2019.

Hamiltonian Graph Networks with ODE Integrators

Sanchez-Gonzalez+ Hamiltonian Graph Networks with ODE Integrators.
arXiv:1909.12790, Sep 2019.

Both from



化学反応の設計と探索



魂は細部に宿る：道具としての機械学習

God is in the details.

- Ludwig Mies van der Rohe



<http://www.900910.com/mies.php>

「機械学習」研究の意義

より理解するために技法(道具)を整備する
「良い仕事は良く手入れされた道具から」

職人魂 (技術者魂)

- ・ プロにとって使う道具は命。
- ・ 道具の特性に精通し、丁寧に扱い、手入れを怠らない。
- ・ 道具箱の中をひとめ見るだけでその職人の気質とレベルが分かる。

最後に：介入・実験も含む最適化へ

件のBoxの1966年の論文「Use and Abuse of regression」は
非常に有名なこんな一文で締めくくられる。

To find out what happens to a system when you
interfere with it **you have to interfere with it**
(not just passively observe it).

理屈から言っても機械学習屋とデータ「だけ」では何もできない
ということです。

ドメイン知識を持った専門家との協働が必須です！
どうぞよろしくお願ひします(?)

Take Home Message

科学が求めること: 分からないことが分かる(科学的発見)

理解

原因と結果(因果関係)を見出す

$x \rightarrow y$ の過程を理解し(人間が)発見する

発見

今まで見出されていない良い対象を見出す

$x \rightarrow y$ を利用して良い y を持つ x を発見する

今日伝えたいたった3つのこと

1. 単純に機械学習を使うだけでは**いざれも解けない**
2. **機械学習以外のもの**(介入やドメイン知識)が原理上必須
3. 最近**まさに研究が進行中**の未解決領域だが研究は色々ある