「　MI²社会実装に向けた
スクール機能の試行を振り返って　」

Materials research by Information Integration

Mi2i 副PL　真鍋　明

MaDIS
NIMS MATERIALS DATA and INTEGRATED SYSTEM

背景　ML,DS,MI　その萌芽は20年以上前から

1940　1960　1980　2000　2005　2010　2015　2020

◆1943 Neural Network

◆1959 **Machine Learning** /Arthur Samuel

◆1979 Neocognitron
/Kunihiko Fukushima

◆1962 Data Analysis
/John Tukey

◆1997 **Data Science**
/C.F.Jeff Wu

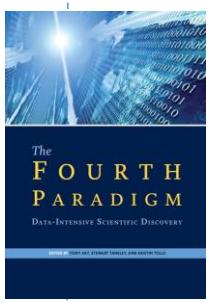
◆1998 Chemoinformatics

◆1999 **Materials Informatics**
/John R. Rodgers

◆2012 Google's Cat

◆2016 AlphaGo

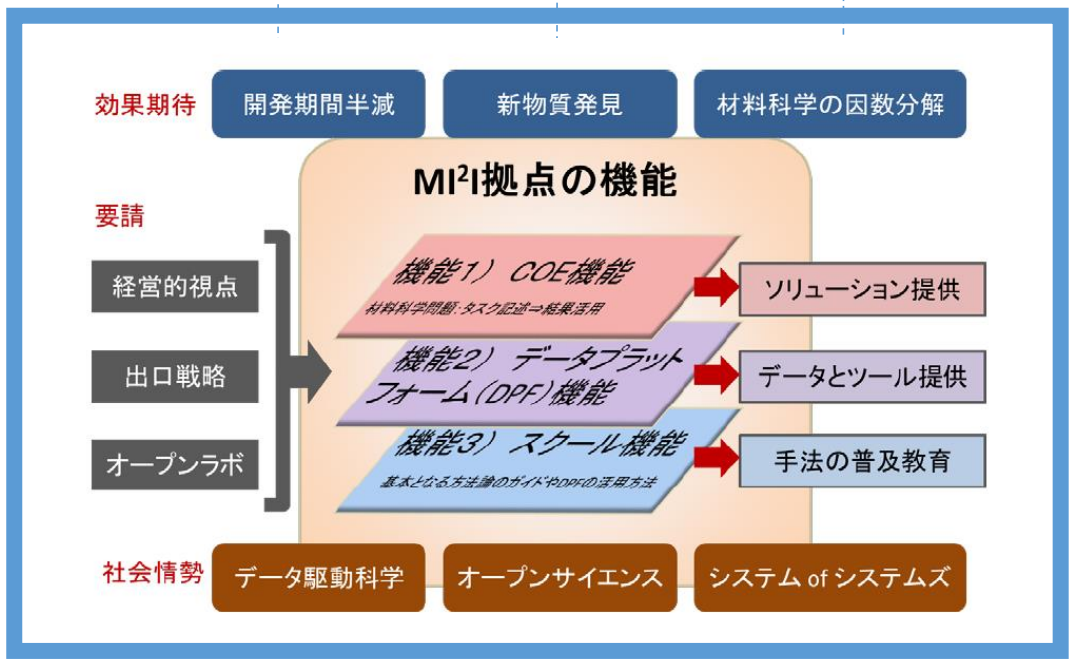
◆2007 **Data Intensive Science**
/Jim Grey



◆2019
Sycamore

MGI

MI2I



背景　実質的な進展はここ10年

巨大データとIT

1940　1960　1980　2000　2005　2010　2015　2020

◆1943 Neural Network

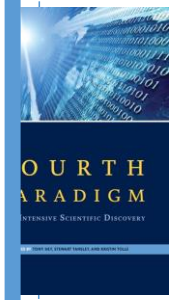
◆1959 **Machine Learning** /Arthur Samuel

◆1979 Neocognitron

◆2012 Google's Cat

◆2016 AlphaGo

Data Intensive Science
/Jim Grey



◆2019
Sycamore

MGI

MI2I

Buzzword

効果期待　開発期間半減　新物質発見　材料科学の因数分解

MI²I拠点の機能

要請
- 経営的視点
- 出口戦略
- オープンラボ

機能1）COE機能
材料科学問題：タスク記述⇒結果活用
→ ソリューション提供

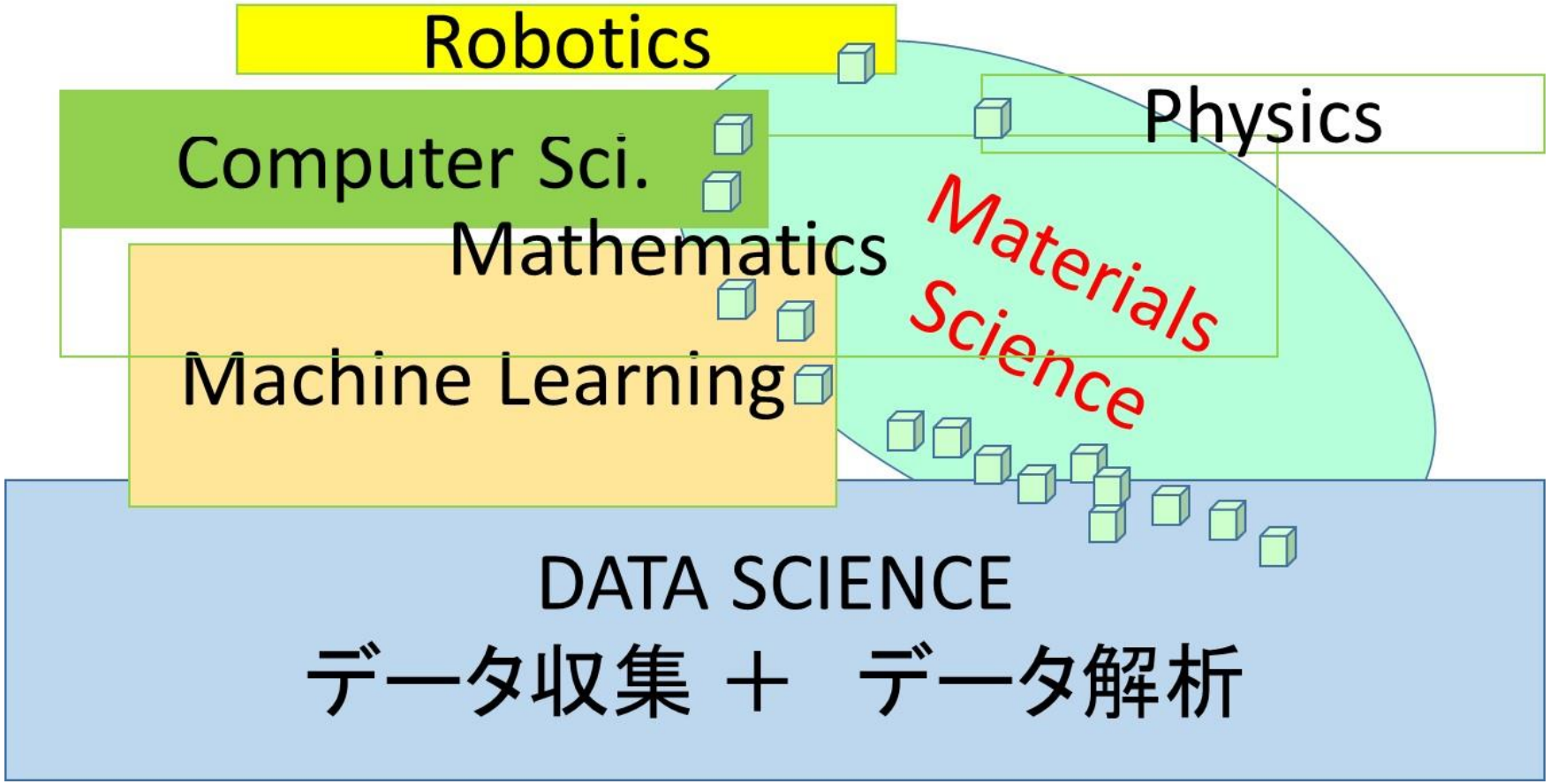
機能2）データプラットフォーム（DPF）機能
→ データとツール提供

機能3）スクール機能
基本となる方法論のガイドやDPFの活用方法
→ 手法の普及教育

社会情勢　データ駆動科学　オープンサイエンス　システム of システムズ

Materials Informatics （Buzzword状態だった）

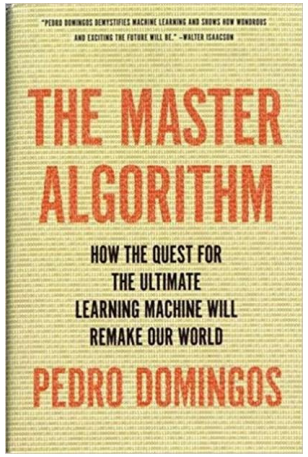
＝　過剰期待･断片情報の集合体



材料研究者にとって「データ収集＋データ解析」は普通にやってきたこと。

いままでと何が違うのか？　＝＞　機械学習ツールの敷居が著しく低くなった。

数限りないアルゴリズム　究極はいつ？



The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World
Pedro Domingos
Basic Books(2015/9/22)

・アルゴリズムは5つの派閥に分類できる。

・この5つを統合した究極のアルゴリズムの可能性に言及

https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

“SCHOOL” 実践スキルの獲得を目指して

～そのきっかけづくり （しかできません）

・チュートリアルセミナー（座学）

＜様々な材料研究でのMI活用シーンを紹介＞

シリーズで全体像

講演映像はDVD教材化

・ハンズオンセミナー（講義＋実習）

＜とにかくMLツールを触ってみる＞

基本コース：Python系のML基本ツール

実践コース：材料専用MLツール@各拠点

多くの方に参加・活用いただきありがとうございました。

詳細はハブ構築事業のノウハウレポートをご覧ください　https://www.jst.go.jp/ihub/seika.html

	概要	実施回数	実施規模	備考
Mi2I チュートリアルセミナー	MI2に関する座学	10回	延べ 1042名	教材DVD配布 DVD：1629枚
Mi2Iハンズオンセミナー 基本コース	Pythonコード習得 Orange Scikit-learn	15回	延べ 597名	ライブ配信 教材CD配布
Mi2Iハンズオンセミナー 実践コース	プロジェクトで開発したコード習得 Nap, XenonPy, CrySPY, COMBO, HomCloud	11回	延べ 137名	コードは GitHubで公開

これまでは、プロジェクトからの情報発信（一方向）

今後は各企業での実践ステージへ（個別解法の探求）

◆ ML活用場面は多種多様

1）既存材料を凌駕する画期的な新材料発見

2）材料諸現象の理解深化

3）開発効率化・自動化に向けた代理モデル活用

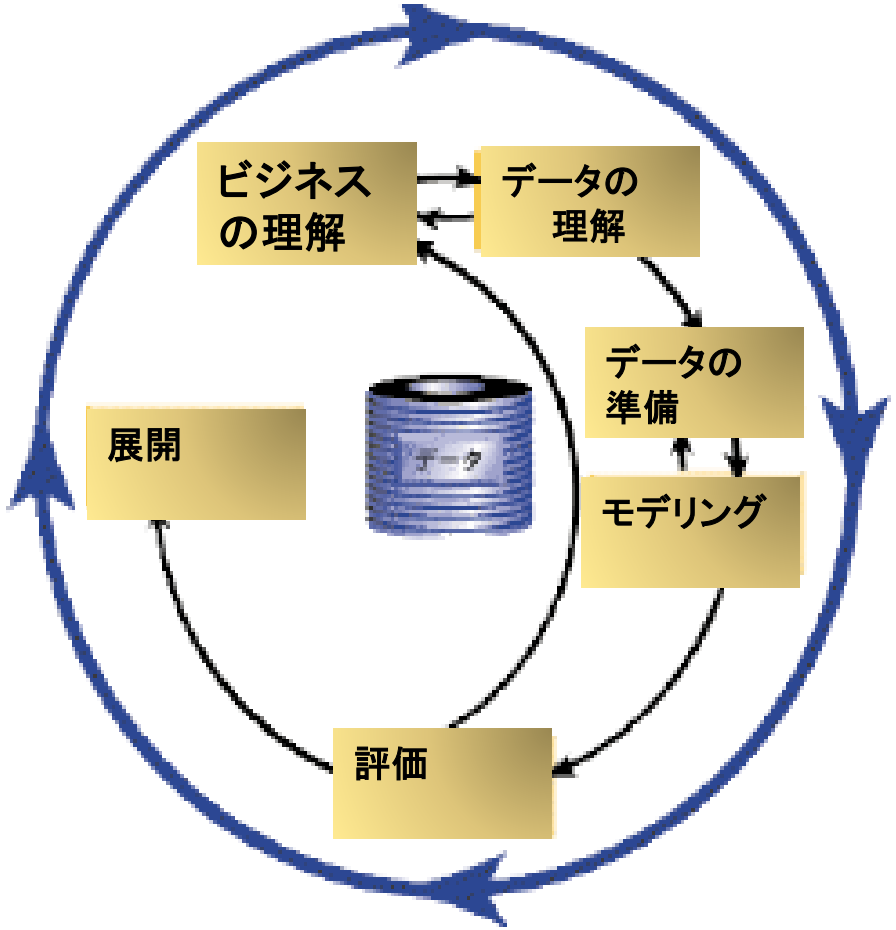
◆ MLタスクは研究業務フローの一部

「ビジネスの理解」が絶対条件

CRISP-DM (Cross-Industry Standard Process for Data Mining)

データからのパターン抽出のスタートは **「ビジネスの理解」**

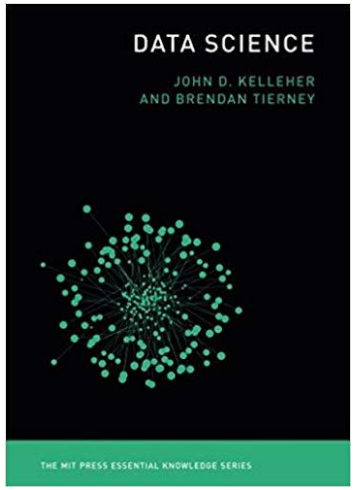
各企業はそれぞれのビジネス（研究業務）を完全に把握している。
具体的展開法は、それぞれが完成させることとなります。



IBM SPSS Modeler CRISP-DM ガイドより引用

データ科学を活用して意味あるタスクを

“DATA SCIENCE” “データサイエンス”日本語版もあり
/ MIT PRESS essential knowledge series



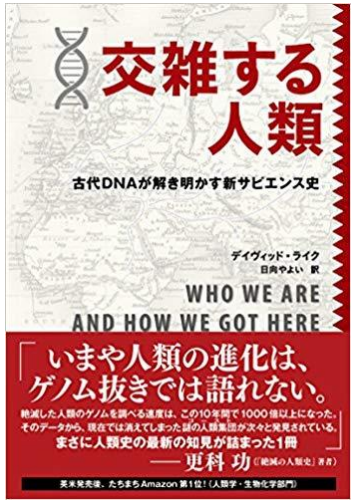
DATA SCIENCE
John D. Kelleher, Brendan Tierney
MIT PRESS essential knowledge series(2018/4/6)

「専門家が容易にパターンを考え出せるなら、それを見出すためにデータサイエンスで時間と手間をかけるのは無駄なこと。」

要領よくデータサイエンスの基本がまとめられた本。

述べられていることは厳しい。
「本格的に取り組まないと失敗する」ということのようだ。

データ科学を活用して意味あるタスクを



“交雑する人類” / 日本語版 NHK出版

WHO WE ARE AND HOW WE GOT HERE

交雑する人類　ディヴィッド・ライク
NHK出版(2018/7/25)

古代DNAデータで人類進化の歴史を語る
主成分分析とデンドログラム（樹形図）

→ 良質なデータがあれば従来手法でも「雄弁」

参考）
今年の日本国際賞受賞のペーボ先生のDNA解析手法を高精度、ハイスループット化し数千体の古代DNAデータベースにしたのが著者のライク先生ら

JAPAN PRIZE

スバンテ・ペーボ 博士
（スウェーデン）

授賞対象分野「生命科学」分野

授賞業績

古代人ゲノム解読による古人類学への先駆的貢献

スバンテ・ペーボ博士　1955年4月20日生まれ(64歳)
マックス・プランク進化人類学研究所 教授

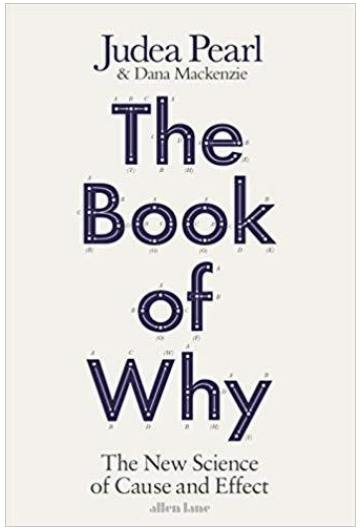
現生人類の祖先とネアンデルタール人は交雑していた

ネアンデルタール人は、かつて存在していた人類です。今から50万年ほど前にアフリカを出てヨーロッパから中近東にかけて広く住んでいましたが、4万年ほど前に絶滅しました。そのため現生人類とは無関係と考えられてきたネアンデルタール人ですが、博士が、発掘されたネアンデルタール人の骨のDNAを解析したところ、そのDNAが現生人類に受け継がれていることがわかりました。古代人DNAの扱いには細心の注意が必要なため、博士はDNA抽出法の確立やクリーンルームの設置など新しい研究法を工夫しました。

1997年、まずネアンデルタール人のミトコンドリアDNAの配列の一部を決定、その後、全配列を決めました。ミトコンドリアは、細胞内小器官の一つで、核とは別のDNAをもっています。ミトコンドリアDNAは16000塩基対と短い上に、1個の細胞に数千個あって量を確保しやすかったため、当時のPCR法とDNA解析技術で配列を決めることができました。

日本国際賞HPより　https://www.japanprize.jp/prize_past_2020_prize02.html

データ科学を活用して意味あるタスクを



“The Book of Why” 因果推論の新科学

The Book of Why The New Science of Cause and Effect
Judea Pearl and Dana Mackenzie
Allen Lane (2018/5/5)

フィッシャーの実験計画法 ～1920
ランダム化比較試験やラテン方格の話題

→ 温故知新 （古典的）実験計画法は因果が得意

ACS NANO
ACS Nano 2018, 12, 7434−7444
Cite This: ACS Nano 2018, 12, 7434−7444
www.acsnano.org

How To Optimize Materials and Devices *via* Design of Experiments and Machine Learning: Demonstration Using Organic Photovoltaics

Bing Cao,[†,‡] Lawrence A. Adutwum,[*,†,§,‡] Anton O. Oliynyk,[†,‡] Erik J. Luber,[†,‡] Brian C. Olsen,[*,†,‡] Arthur Mar,[*,†,‡] and Jillian M. Buriak[*,†,‡]

PERSPECTIVE

- プロセスパラメータを含むので 新たな実験でデータを準備する
- 実験計画法（直交表）で効率よく
- 応答曲面で最適解探索 （ここで最新のMLを使う）

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS
2019, VOL. 20, NO. 1, 1010–1021
https://doi.org/10.1080/14686996.2019.1673670

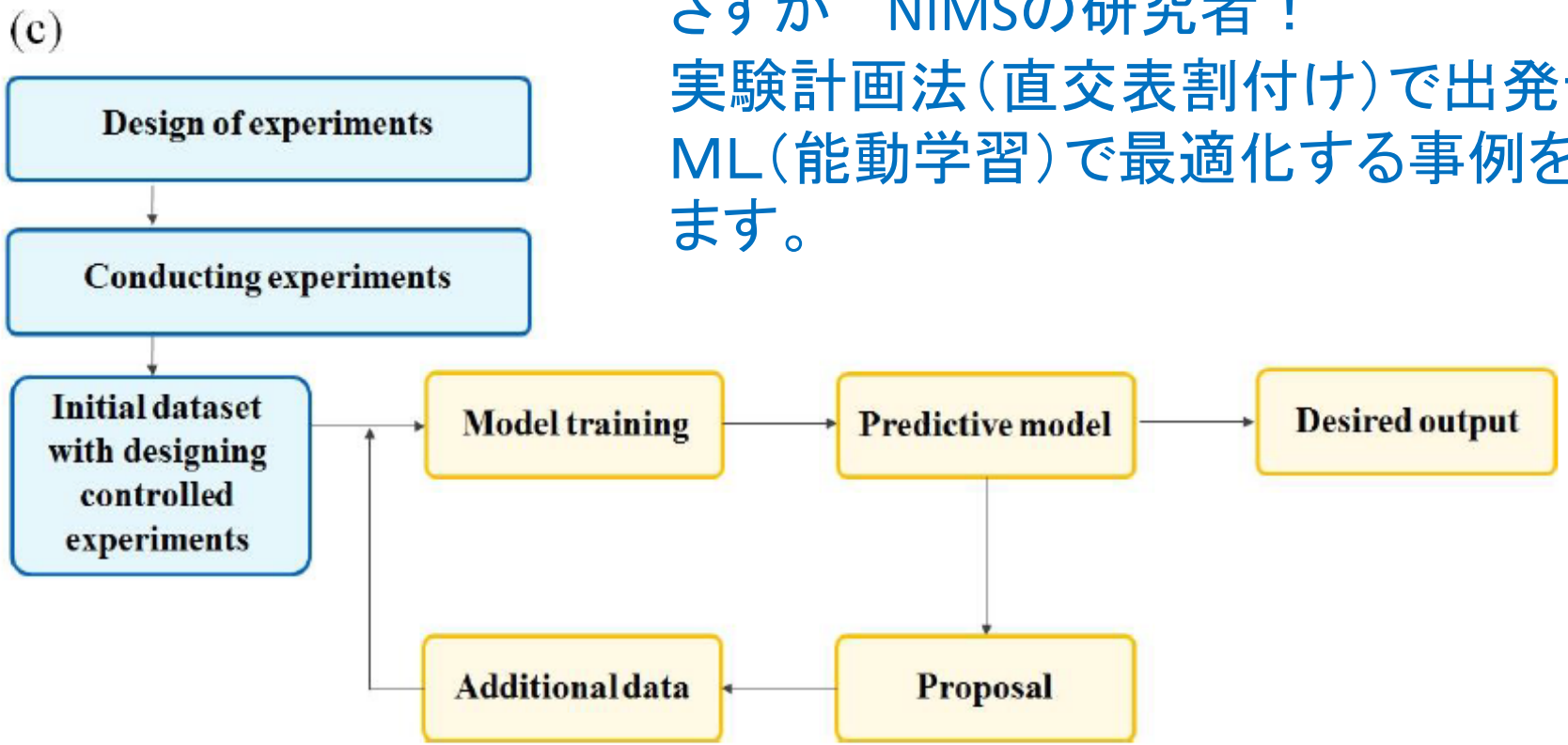
Taylor & Francis
Taylor & Francis Group

Mi²i

Prediction and optimization of epoxy adhesive strength from a small dataset through active learning

Sirawit Pruksawan[a,b], Guillaume Lambard[c], Sadaki Samitsu[a], Keitaro Sodeyama[c] and Masanobu Naito[a,b,d]

さすが　NIMSの研究者！
実験計画法（直交表割付け）で出発データを生成し、ML（能動学習）で最適化する事例を論文にしております。



研究コミュニティは必要か？

ML誤用・乱用

1 AUG 2019 | VOL 572 | NATURE | 27

Three pitfalls to avoid in machine learning

As scientists from myriad fields rush to perform algorithmic analyses, Google's **Patrick Riley** calls for clear standards in research and reporting.

Google's Patrick Riley

ML誤用・乱用への警告

1. データ分割法不適切
2. 潜伏要因への配慮不足
3. 目的不明瞭
（MLとしては正しいが目的は？）

提言として
機械学習の実行・まとめ方に関し、それぞれの専門分野で明確な基準をつくる必要がある。

1. ML基準・規範づくり

・論文～学会・協会

・商取引～各業界・工業会

2. 上記に関する教育・普及活動

研究コミュニティは必要か？　超えるべき共通課題

Chem. Mater. 2019, 31, 9579-95811
Citrine Informatics　Bryce Meredig

1．**材料特有のML手法開発**
small, noisy, multiscale, heterogeneous
multidimensional dataset
very large, high dimensional design spaces.

2．**外挿予測モデルの実験による検証**
3．**ハイスループットデータ取得　計算、自動ラボ、論文**
4．**物理モデルとMLの統合**
5．**科学者支援**（可視化や計測インフォマ等）

左記課題の実施主体は？

アカデミア主導か？
産業主導か？

情報共有のための
コミュニティの形成

課題はSTAGEで異なる

NIMSフォーラム　2015.10.9

	STAGE I 新物質創成 例）C_{60}	STAGE II 物性極値化 Materials Genome	STAGE III 材料最適化 Integrated Computational Materials Engineering	STAGE IV 適用研究開発
内容	従来の特性限界超 物質探索	結晶構造あり 元素置換　ドープ 極値を探す	材料化 プロセス・組織構造 の最適化	システム設計 試作実証 信頼性確保
ポイント	コンセプトひらめき 実験発見 Abduction	傾向予測と実験 Deduction Induction	実験検証 特性ﾄﾚｰﾄﾞｵﾌ克服 Induction主体	Virtual　Prototype シミュレーション Deduction
データ共有	Unknown OPEN　知識	OPEN　一部CLOSE 物質データベース	特性CLOSE/OPEN プロセスCLOSE	固有材料CLOSE 一般材料OPEN
MIの期待	逆問題 特性→構造予測	結晶構造・特性相関 QSPR	特性・組織相関 プロセス・組織相関	短期間化 （時間、費用）
課題	**方法論研究**	**事例研究 手法選択**	**組織構造データ化 データ形式統一** （メタデータ）	**各種シミュレーション**

出口

多くの課題がまだ残っております。

レシピ＋支配方程式

SCHOOL (一方向) から

COMMUNITY
（双方向の情報発信）のステージへ