ディープラーニングとオープンサイエンス
研究の爆速化が引き起こす摩擦なき情報流通へのシフト

北本 朝展（KITAMOTO Asanobu）
国立情報学研究所
情報・システム研究機構
人文学オープンデータ共同利用センター（CODH）
http://researchmap.jp/kitamoto/
@KitamotoAsanobu

自己紹介

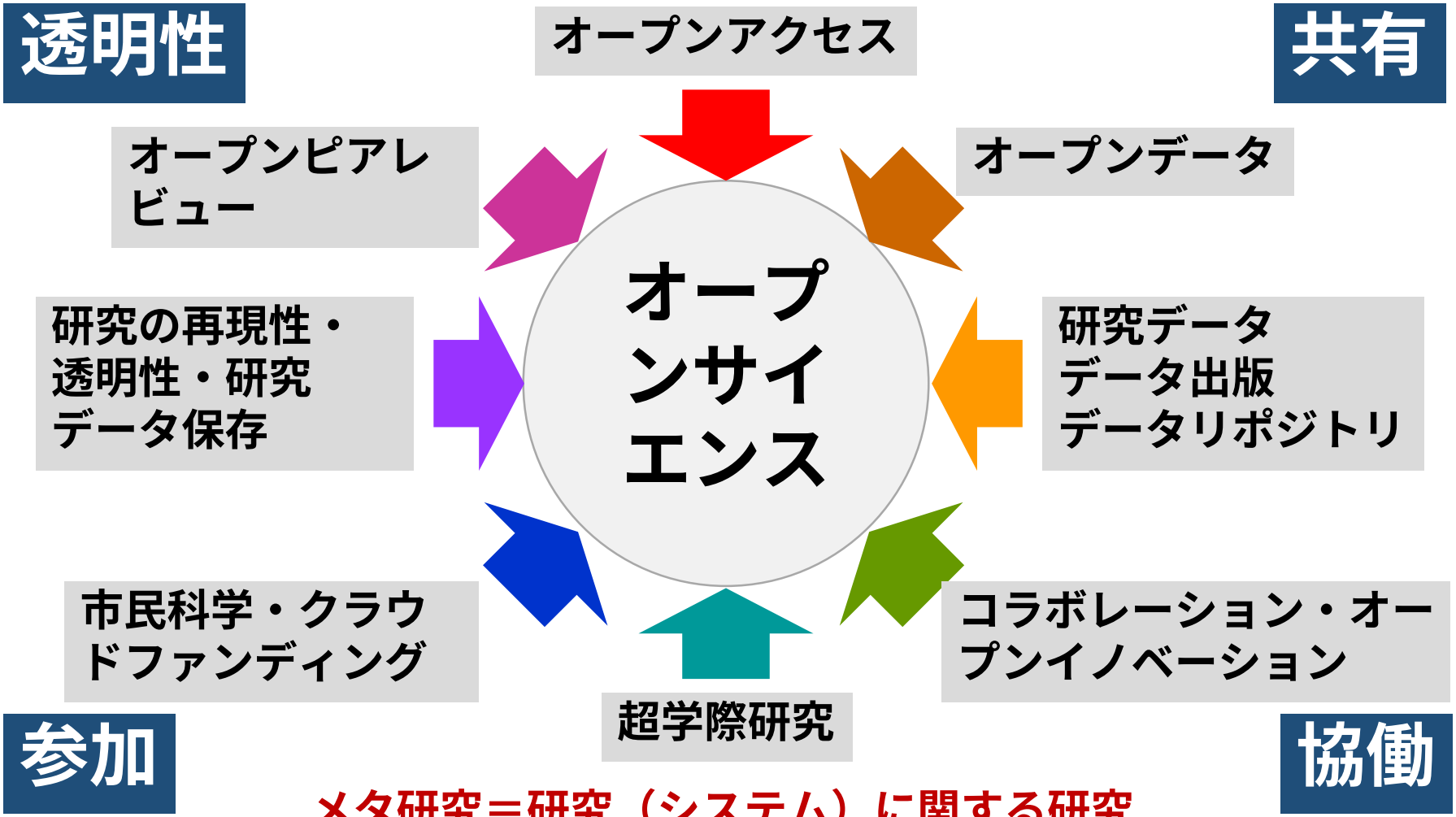
- 研究分野は**情報学**。画像処理や画像データベースの研究が出発点。
- データ駆動型サイエンスに発展し、**気象情報**、**地球環境情報**、**人文科学情報**などの分野で研究。
- 最近はオープンサイエンスにも関わる。

オープンサイエンスの背景

オープンサイエンスとは？

- 「オープン」という言葉を梃子にして、サイエンス（研究）の方向を変える。
- 「よりオープンに」という方向性を共有する活動を、一語で束ねると見える世界。
- 個々の活動ごとに「オープンサイエンス」の意味は異なり、単一の定義は困難。
- 大同団結？同床異夢？個々の活動を超える新しい目標を示せるかが問われる。

オープンサイエンスへの収束

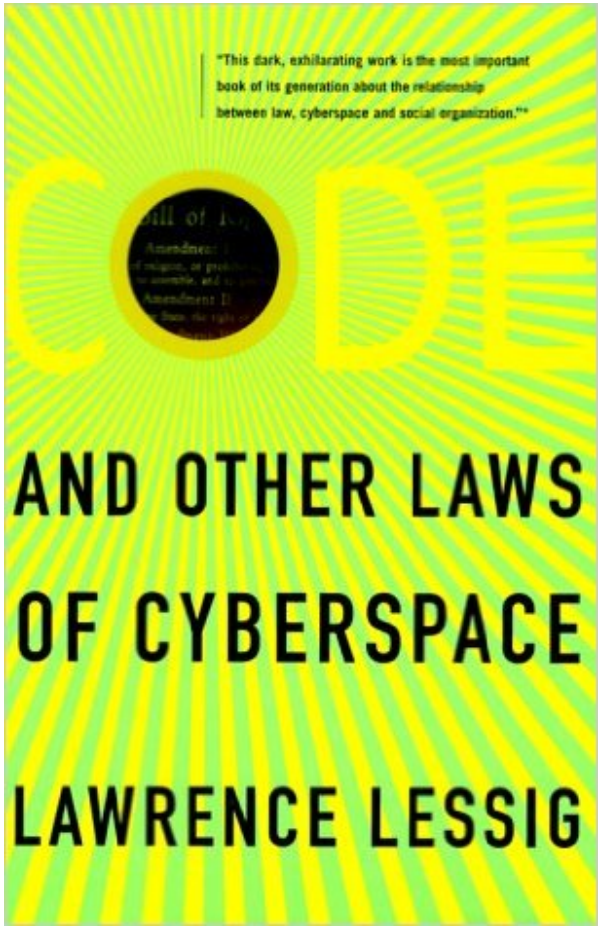


メタ研究＝研究（システム）に関する研究

「オープン」の3つの側面

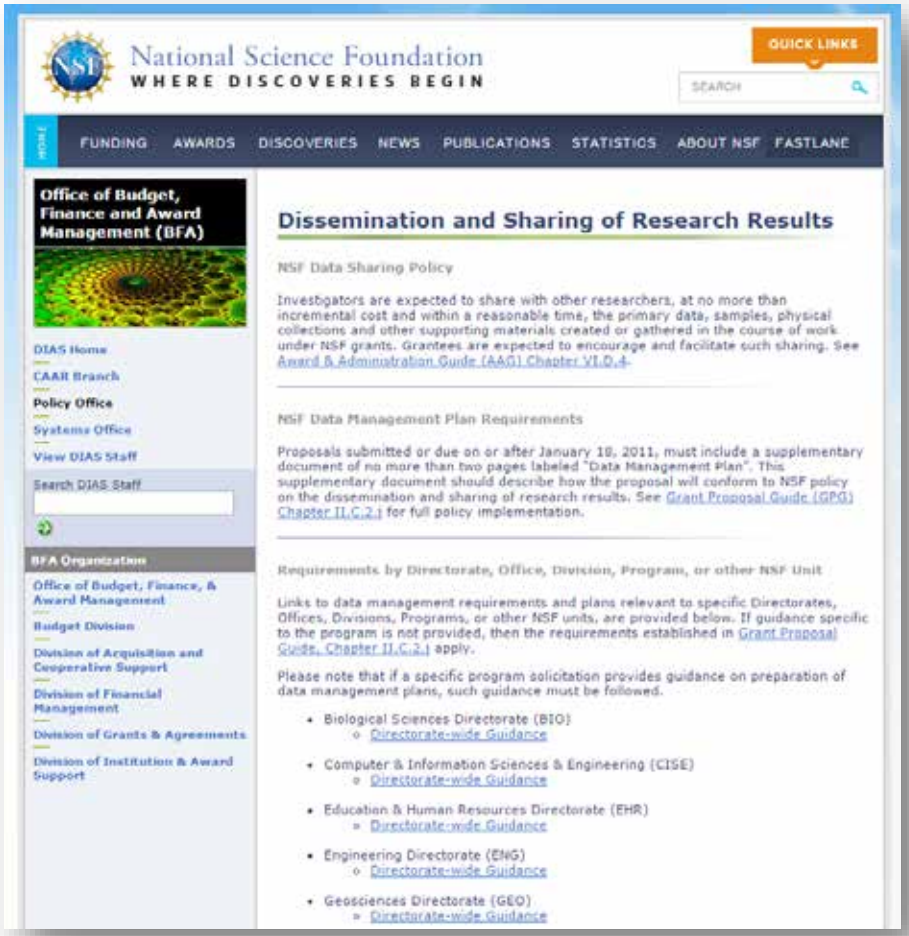
1. **他者が使える（再利用）**
   - オープンデータやオープンアクセスなど。外部の人が研究結果を自分の目的に再利用できる。
2. **他者が検証できる（透明性）**
   - オープンガバメントや研究再現性など。外部の人がエビデンスを検証し、正当性を判断できる。
3. **他者を受け入れる（参加）**
   - オープンイノベーションや市民科学など。外部の人を招きいれ、共に価値を生み出す。

制度を分析する4つの視点



- Lawrence Lessig (Founder of Creative Commons), *Code: And Other Laws of Cyber Space* (first edition 1999)
- **法＝しなければならない**
- **規範＝すべきである**
- **市場＝した方が利益がある**
- **アーキテクチャ＝せざるを得ない**

「法」によるオープン化



NSF Data Management Plan

- **資金提供機関**が「データ管理計画」などを義務化。
- **研究評価**でも、オープン化の進展を指標に含める。
- 上からの押し付けには**副作用**も大きく、見極めが必要。

「規範」によるオープン化

https://www.icsu-wds.org/

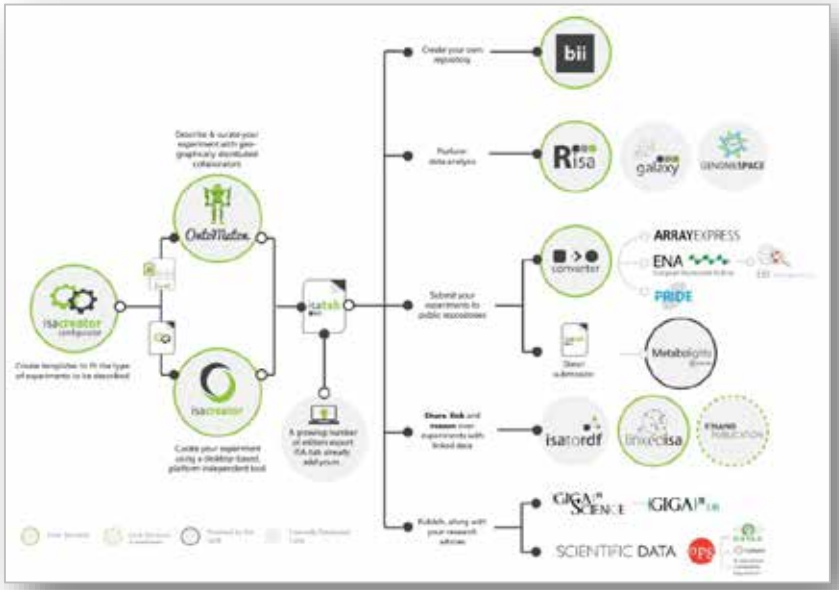
- **オープンな文化**：データ共有が不可欠な分野もある。
- **世代の差**：若い世代では共有文化の経験がより強い。
- **文化の差**：異なる文化圏に対する説得力が弱い。

「市場」によるオープン化

Scientific Data (Nature publishing group)

- **報酬への期待**：研究成果をオープン化すると、引用も増加する [要出典]。
- **損失への不安**：他者に成果を横取りされるんじゃないの？報酬は労力に見合うの？

「アーキテクチャ」による オープン化



http://www.isa-tools.org/software-suite/

- **選択と誘導**：プラットフォームを選ぶと、可視／不可視なルールによって誘導される。
- **苦痛の軽減**：オープン化は大変だから、有償サービスにお任せ？
- **ベンダーロックイン**：良くも悪くも企業のビジネスチャンス。

研究データのオープン化

研究資源データ	研究の入力となるデータ。観測データや評価用データセットなど。再利用のためのオープン化が求められる。
論文付属データ	研究の出力となるデータ。論文のエビデンスとなるデータなど。透明性のためのオープン化が求められる。
研究過程データ	研究の入力と出力の間で生み出されるデータ。日々の研究活動のエビデンスとなるデータなど。研究不正防止（透明性）のためのクローズドな長期保存が求められる。

データ駆動型研究からの要請

PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

https://www.whitehouse.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence

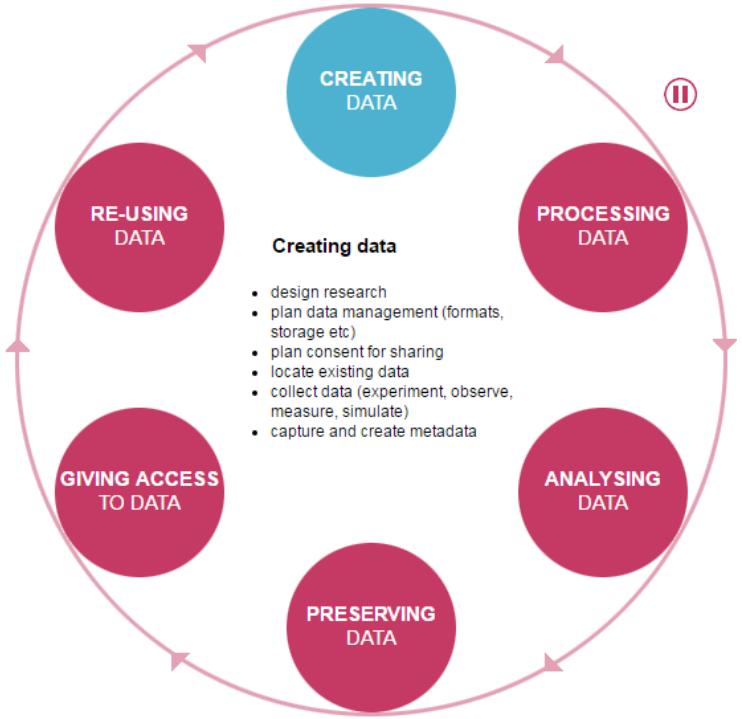
- Recommendation 1: **Private and public institutions are encouraged to examine whether and how they can responsibly leverage AI and machine learning in ways that will benefit society**.
- Recommendation 2: **Federal agencies should prioritize open training data and open data standards in AI**.

データライフサイクル

- **データが誕生してからたどるプロセス**を、段階の遷移として表すモデル。
- ライブラリアン／キュレーターは、**サイクルの一部の段階**に関わる。例えばデータ管理計画（DMP）の作成支援は、計画段階に関わることを意味する。
- **唯一の正解はなく、プロジェクトに適したライフサイクル**を考える必要がある。

データライフサイクルの例

https://www.dataone.org/data-life-cycle



http://www.data-archive.ac.uk/create-manage/life-cycle

ライブラリアンの役割変化

従来のライブラリアン

「本」という**最終生成物**を対象とし、利用者（読者）のために整理していればよかった。

データライブラリアン

データ作者へと出向き、**作品**の意図を汲み取りメタデータを付与するなど、**編集者や出版者**に近い役割も担う。

データキュレーター

編集スキルを活用して、データに**付加価値**を付けて利用者に売り込む。

整理スキルを活用して、データに**標準価値**を与え、利用者に提供する。

求められるスキル@2017年

1. **データライフサイクル**に関する知識とDOIなどによるデータ管理の基礎。
2. **メタデータ**規格や語彙に関する国際標準との相互運用性を踏まえた知識。
3. **データ共有と利活用**を促進する方法に関する基本的な動向の把握。
4. 【上級】研究コミュニティとの議論やエスノグラフィ的観察に基づき、**データ管理システム**を設計し構築する能力。

求められるスキル@2027年？

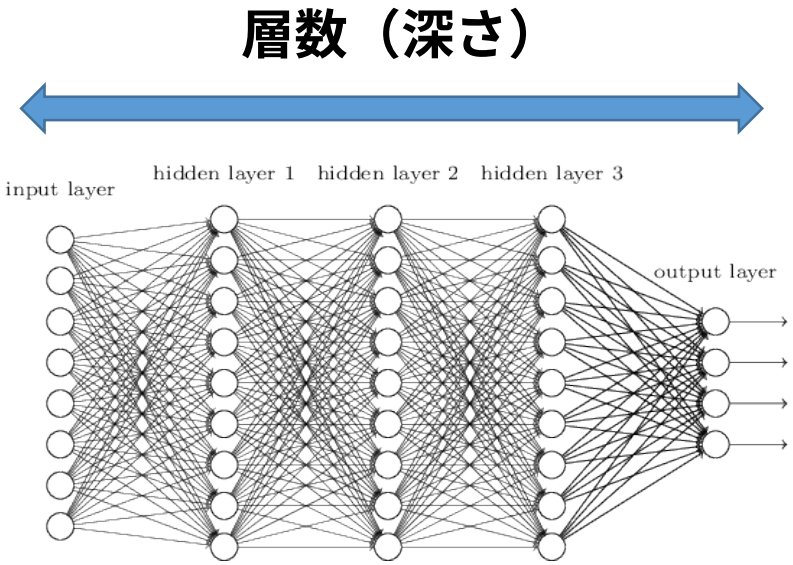
1. AIによる人間の代替はどこまで進む？代替可能なスキルを学ぶのは、人生戦略上の損失。
2. メタデータは人間が付与するか？アクセスが目的になると、項目や品質の考え方も変わる。
3. AIを使いこなす：データライブラリアン兼データサイエンティストというスーパーマン。
4. AIの下で働く：現実世界のダーティデータを、AIが使いやすいクリーンデータに整理する。
5. 問い：10年後も維持できる価値は何か？

オープンサイエンス再考

オープン化の第四の軸

- これまでオープンサイエンスの意義を**利便性、透明性、参加**という3軸から考えてきた。ここに「**スピード**」という新たな軸を導入する。
- **研究のスピードが極限まで高速化**すると、情報流通もそれに追従して高速化せねばならないため、**情報流通の妨げとなる摩擦**を取り除く方向に進化し、結果的に科学研究がオープン化する。
- この仮説を、**ディープラーニング（深層学習）**の現状から検証する。

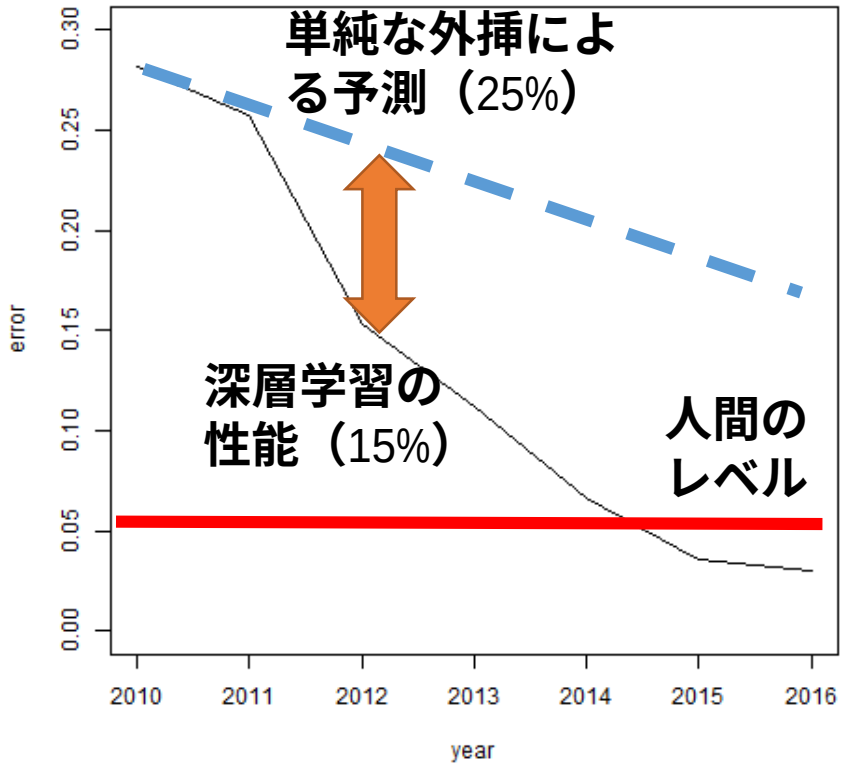
ディープラーニング登場



Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015, CC BY-NC

- **ニューラルネットワーク**の中で特に層が多いもの（深層）。
- **原理は1980年代**から知られている。
- ビッグデータとアルゴリズム改良で**画期的な性能向上**を達成。
- **第3次人工知能ブーム**の中心的存在。

物体認識の画期的性能向上



物体認識タスクの誤認識率の低下。
ImageNet, https://arxiv.org/abs/1409.0575

ディープラーニングが**圧倒的な性能**でコンテストに勝利。ここから快進撃が始まった。

AlphaGoの衝撃

https://deepmind.com/research/alphago/

アルファ碁観戦ツイート
https://togetter.com/li/983741

- ディープラーニングは、**人間とは異なる戦略を用いて**、人間のチャンピオンに勝利した。
- 過去データを学ぶだけでなく、**自己対戦で戦略を深化**させた。
- 開発：DeepMind社（Googleが買収）

オープンソース競争

ディープラーニングの最先端ライブラリを、各社が競ってオープンソース化。

- **知的財産のオープン化**：知的財産のオープン化が、協力者を「おびき寄せる」一つの戦略になった。
- **コミュニティの形成**：協力者が増えれば、創出される価値も増える。
- **競争領域と協調領域**：差別化できる部分は守りつつ、外部の力を使えるところは使う。

スピードへの強烈な圧力

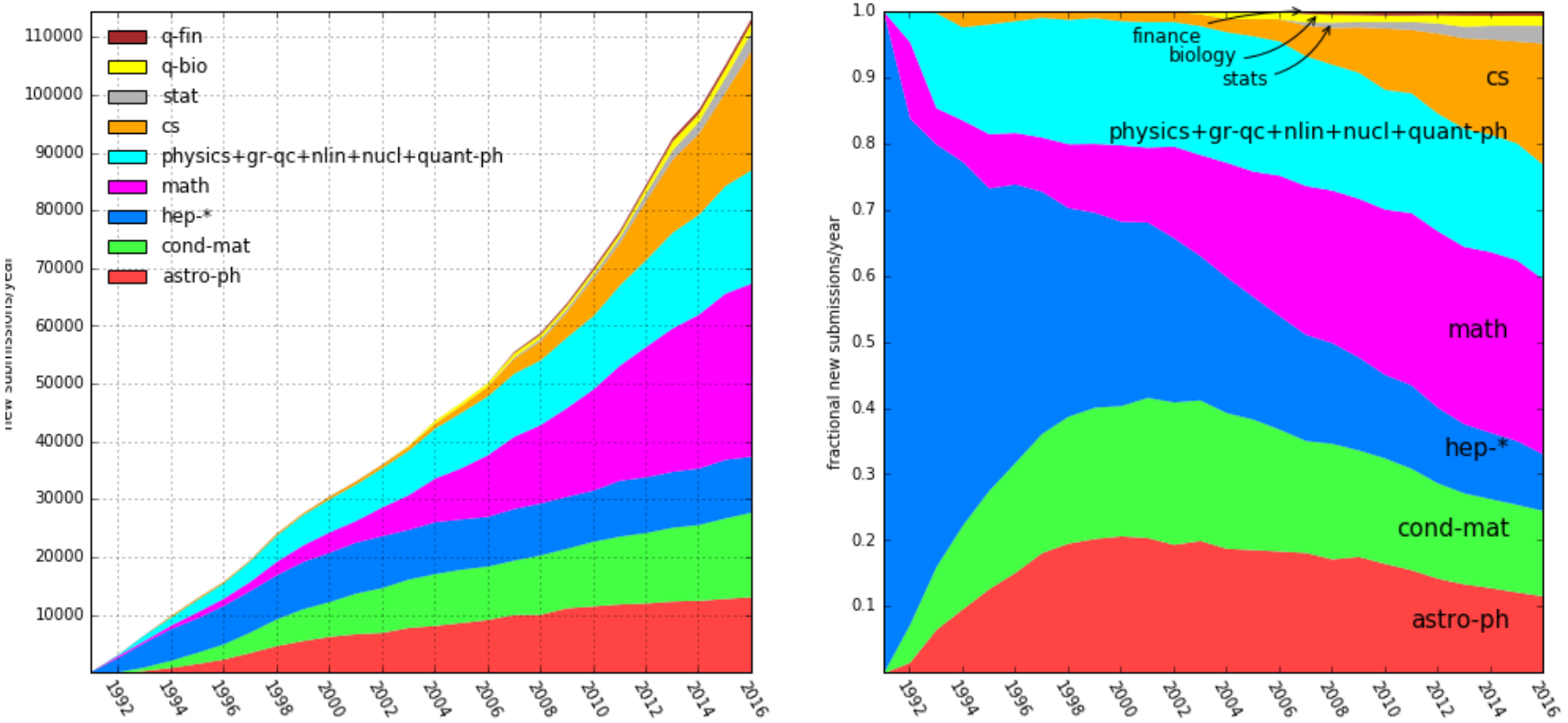
- 各種ライブラリが**オープンソース**化され、各種実験コードも再利用や再現性の観点からオープンソース化。誰でも試せる。
- **共通基盤データ**（例ImageNet）がオープン化。誰でも性能が比較できる。
- **応用分野（囲碁その他）が急速に広がり**、他分野研究者や一般の人々も大挙参入。
- **一刻も早く成果を世界に公表せねば！！**

arXivの利用

https://arxiv.org/

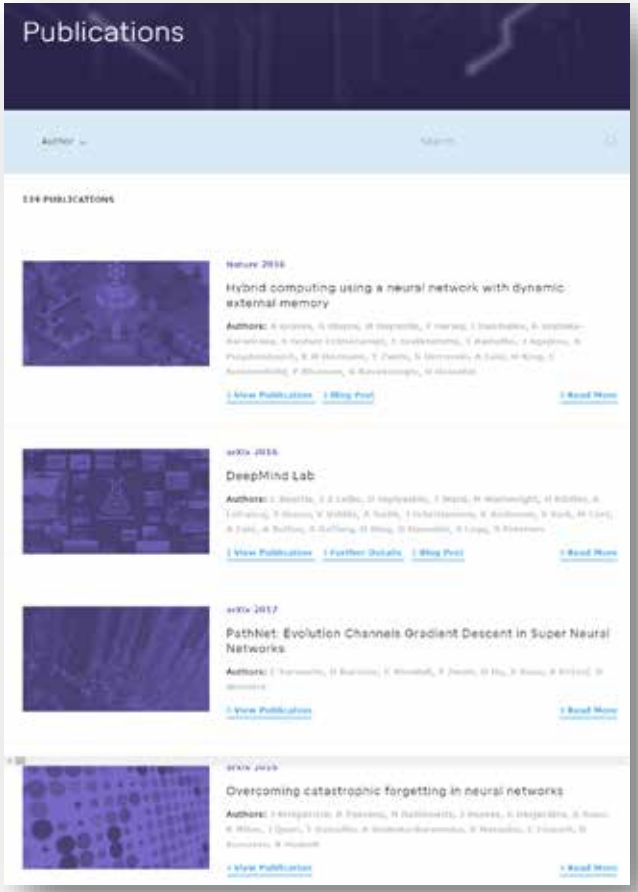
- 1991年登場の元祖プレプリントサーバ。現在はコーネル大学運営。
- 元々は物理学論文対象、後に他分野に拡大。
- 査読前論文をオープンアクセス化。よほど不適格な論文以外は掲載。

arXivへの投稿推移



Left: number of new submissions/year as a function of calendar. Right: submission rates divided by the total for each year, giving the fractional submission rates for each of the domains. https://arxiv.org/help/stats/2016_by_area/index

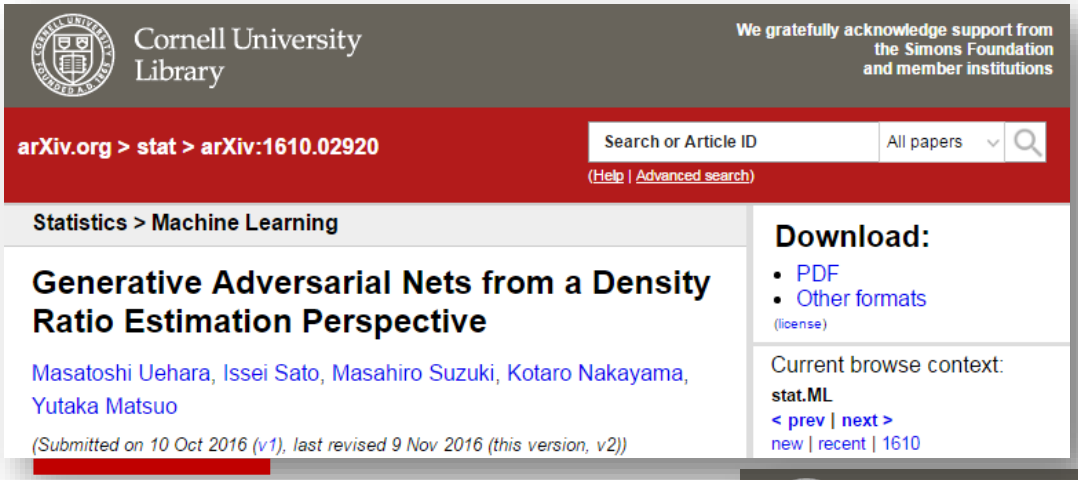
arXivが主戦場



- DeepMindの出版134件中、雑誌10件、arXiv系34件、著名国際会議90件（2017-02-11現在）。
- Natureや国際会議とarXivが同格で並ぶ。
- arXivにまず最新の成果を流す文化を反映？

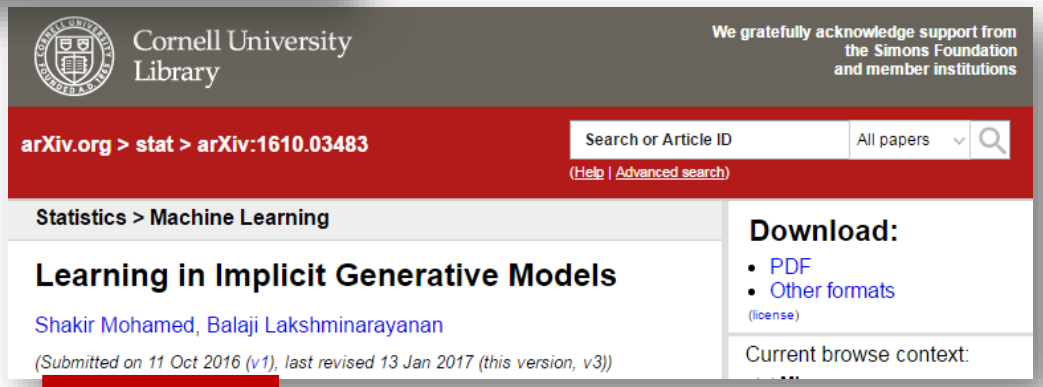
https://deepmind.com/research/publications/

引用の爆速化



https://arxiv.org/abs/1610.02920

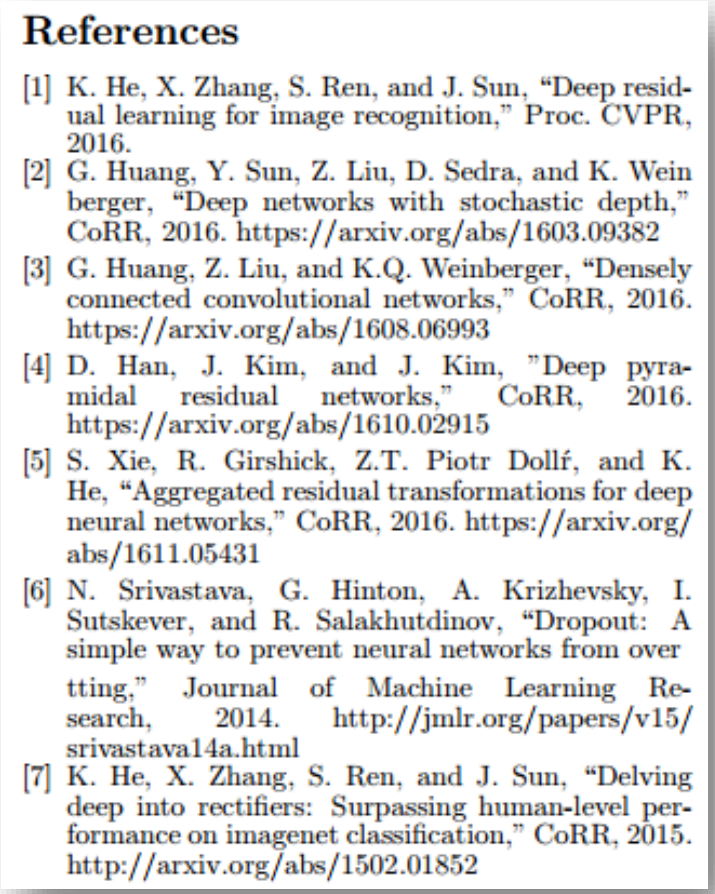
https://arxiv.org/abs/1610.03483



M. Uehara, I. Sato, M. Suzuki, K. Nakayama, and Y. Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

2016年10月10日投稿の論文（上）が、翌10月11日投稿の論文（右）に引用されている！

最新論文の引用への偏り



References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. CVPR, 2016.

[2] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," CoRR, 2016. https://arxiv.org/abs/1603.09382

[3] G. Huang, Z. Liu, and K.Q. Weinberger, "Densely connected convolutional networks," CoRR, 2016. https://arxiv.org/abs/1608.06993

[4] D. Han, J. Kim, and J. Kim, "Deep pyramidal residual networks," CoRR, 2016. https://arxiv.org/abs/1610.02915

[5] S. Xie, R. Girshick, Z.T. Piotr Dollí, and K. He, "Aggregated residual transformations for deep neural networks," CoRR, 2016. https://arxiv.org/abs/1611.05431

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from over tting," Journal of Machine Learning Research, 2014. http://jmlr.org/papers/v15/srivastava14a.html

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," CoRR, 2015. http://arxiv.org/abs/1502.01852

https://arxiv.org/abs/1612.01230

- 技術が急速に進展しているため、**引用は最近数年の論文**が多くなる。
- 必然的にarXivなどのリポジトリの引用が多くなり、**過去論文データベースの重要性が低下**する。

二重投稿とarXiv

[NIPS] Prior submissions on arXiv.org are permitted.

https://nips.cc/Conferences/2016/CallForPapers

[ICML] Submission is permitted for papers that are available as a technical report (or similar, e.g., in arXiv).

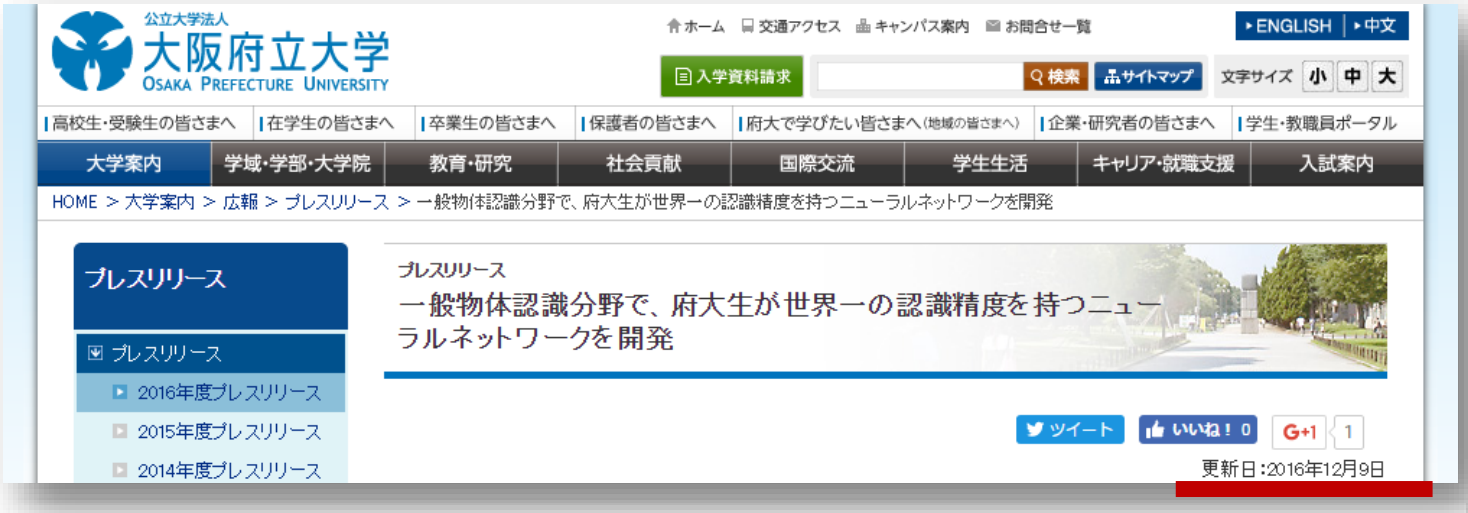
https://2017.icml.cc/Conferences/2017/CallForPapers

[CVPR] Note that such a definition does not consider an arXiv.org paper as a publication because it cannot be rejected. It also excludes university technical reports which are typically not peer reviewed.

http://cvpr2017.thecvf.com/submission/main_conference/author_guidelines

NIPSは**遅くとも2010年**に、CVPRは2012年からarXivに言及しており、2012年前後が普及への**臨界点**？

プレスリリースとarXiv



なお、本研究成果は2016年12月5日に「Computing Research Repository」に公開されました。

論文タイトル：Deep Pyramidal Residual Networks with Separated Stochastic Depth

掲載論文（Cornell University Library「Computing Research Repository」）

https://www.osakafu-u.ac.jp/news/publicity-release/pr20161209/

成果の迅速な公開 ⇔ Ingelfinger Ruleとの兼ね合い

オープンピアレビュー

著者は論文をarXivに投稿しリンクを連絡する。

https://sites.google.com/site/representationlearning2013/program-details/publication-model

2016年まではarXiv投稿形式。

https://openreview.net/group?id=ICLR.cc/2017/conference

The friction in our publication system is slowing the progress of our field.

SNSやまとめサイトでの共有

http://www.itmedia.co.jp/news/articles/1701/30/news065.html

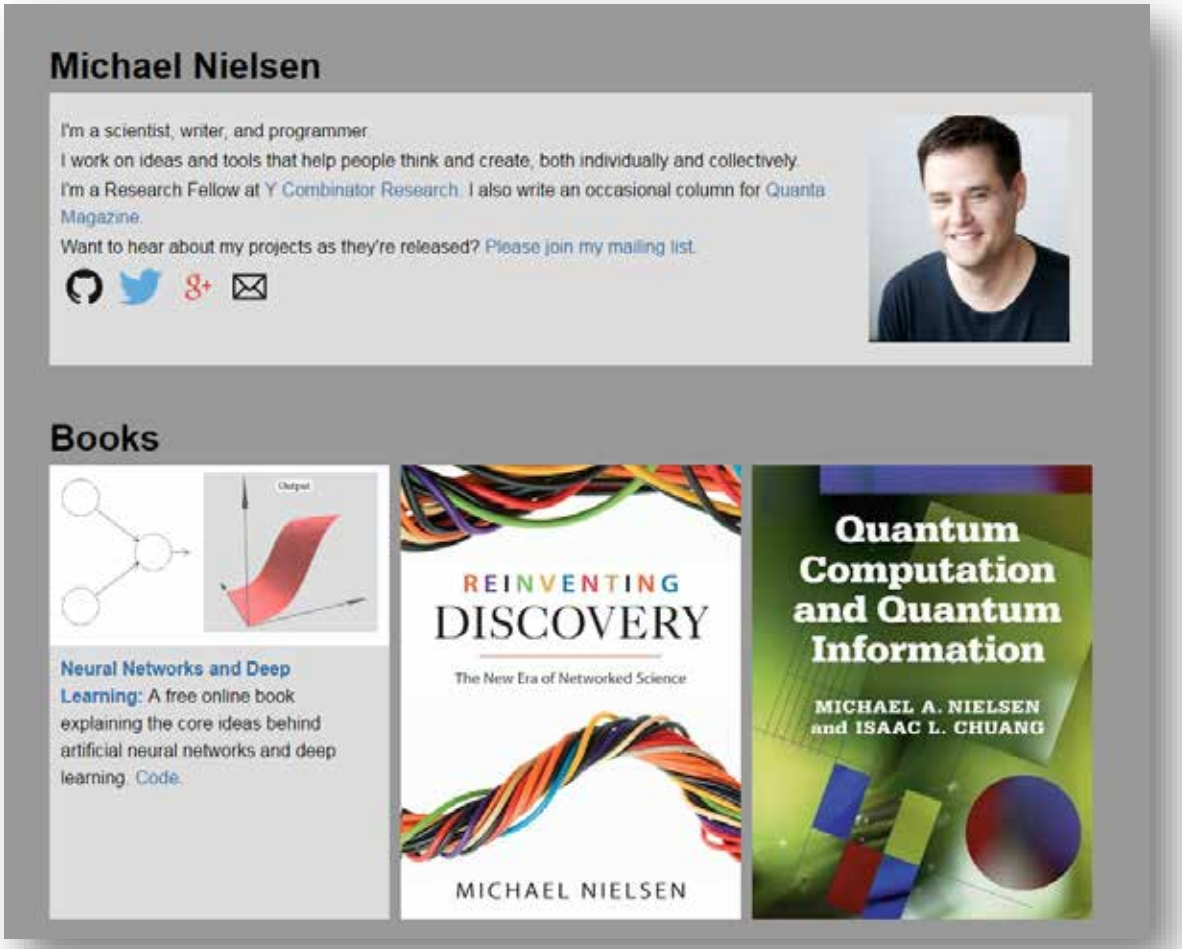
http://qiita.com/shinya7y/items/8911856125a3109378d6

約200個の○○Netが紹介されている。
もう誰も全貌を把握できない。。

「市場」によるオープン化

- **オープンソースとオープンデータ**：研究コミュニティの基盤を共有化。
- **オープンアクセスリポジトリ**：研究成果は最初に「出版」、その後に査読へ。
- **ソーシャル化**：研究成果はSNSで迅速に共有され、公開サービスも自主的に誕生。
- **再現性（頑健性）**：公開サービスを通して、多数のユーザが迅速に評価可能。

マイケル・ニールセンさん

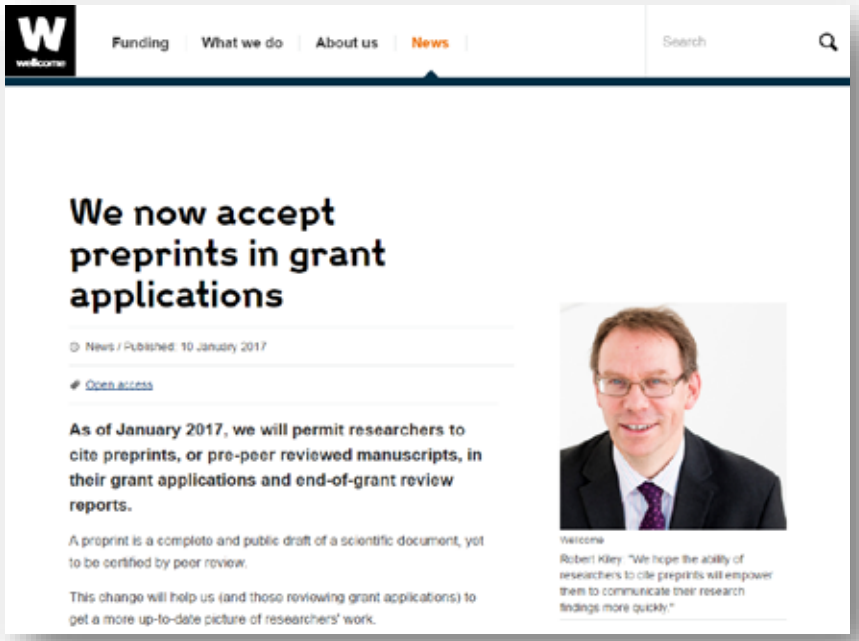


http://michaelnielsen.org/

「法」によるオープン化

http://www.nature.com/news/gates-foundation-research-can-t-be-published-in-top-journals-1.21299

資金提供機関の方針は、オープンサイエンスに大きな影響を及ぼす。



https://wellcome.ac.uk/news/we-now-accept-preprints-grant-applications

摩擦なき学術情報流通へのシフト

人々が本当に欲しいもの

People don‘t want to buy a quarter-inch drill. They want a quarter-inch hole!
（人々が欲しいのは、ドリルではなく穴である）- Theodore Levitt

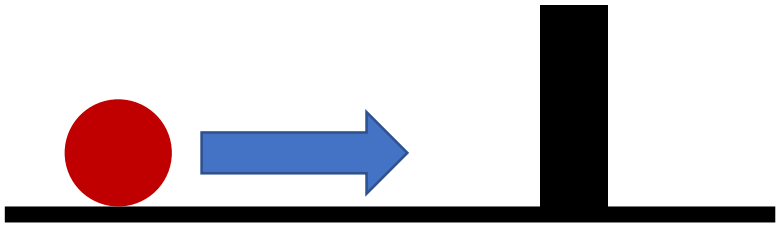
学術情報流通に関して、人々が本当に欲しいものは何だろう？現在の仕組みは、目的ではなく手段なのではないか？

学術情報流通の「摩擦」

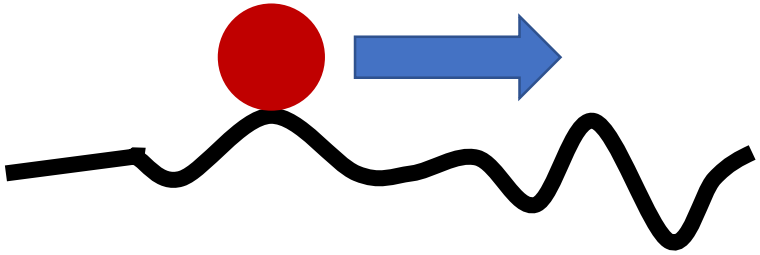
「研究のバリア」を打破する研究基盤デザインと研究データ利活用、SPARC Japan 2015、2015年10月

- 障壁だけではなく摩擦も問題である。
- インターネットは多分野で情報流通の摩擦を減らしてきた。
- 摩擦なきfrictionless情報流通は、研究の高速化が進むにつれて重要性を増す。

障壁と摩擦



- 障壁をなくす：クローズドからオープンへの大転換を求めるもの。
- 選択肢が2つしかないとの錯覚を抱くこともある。



- 摩擦を減らす：よりオープンさを増すような漸進的な変化を求めるもの。
- 律速段階を改善することで全体のフローを高速化する。

摩擦の代表例「査読」

- 学術情報流通における**摩擦**の代表例「**査読**」の問題点を解消できるか？
- **公表の遅れ**：査読に要する時間によって、研究成果の共有スピードが遅くなる。
- **不当な査読**：査読者の不適合や不正によって、研究成果が出てこなくなる。
- **システムの持続性**：査読者の負担が増加し、「**規範**」だけで維持できるか不透明。

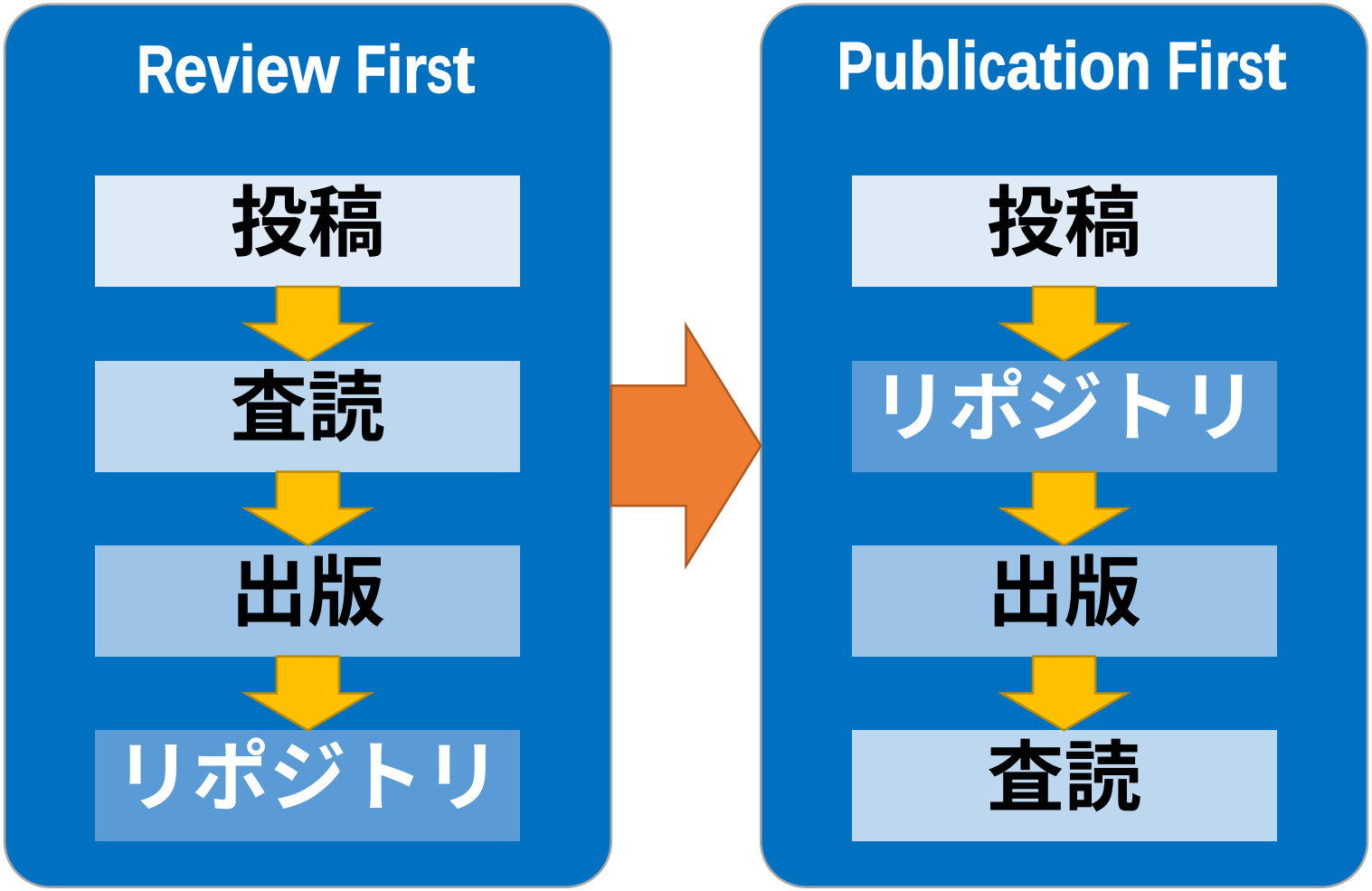
ゲートキーパーの変質

- 査読の役割は、優れた論文だけを選択するゲートキーパー（門番）。
- スペースの有限性 → ゲートキーパー：提供者側の紙面や時間枠の制約を解決する「アーキテクチャ」。
- スペースの無限化 → キュレーション：利用者側のアテンション（注意力）の制約を解決する「アーキテクチャ」。

Subscription Model 2.0

- **アテンションは有限なので「読むべき論文」をまとめてくれるサービスが欲しい。**
- **キュレーションという付加価値サービスは、人間でなくAIが行う可能性もある。**
- **出版物へのアクセスは、著者支払いモデルから「新」購読モデルに回帰？**
- **リポジトリの評価が「流量」勝負となる中、機関リポジトリは意義を保てるか？**

“Publication” First



学術会議の反転

- すでにarXivやSNSで知っている話を、なぜ学術会議で聞かねばならないの？
- **反転授業化**：研究成果の中身について著者に会って議論する場になる。
- **コンサート化**：何度でも聞きたい有名人のストーリーを楽しむ場になる。
- **ブランド化**：コミュニティが重要と考える業績に「焼印」を押す場になる。

出版と評価の分離

- **出版は人類の知の共有が目的**。成果を速やかに共有すれば、知の成長は加速する。
- **評価は人類の知の継承が目的**。有益な知識を選択しながら、知を体系化する。
- 多様な評価に基づき残すべき知を選択するには、**オープンな「出版」**が不可欠。
- 人々が本当に欲しいのは、**厳格な門番よりも信頼のブランド**ではないか？

まとめ

オープンサイエンス再考

- オープンサイエンスの原動力として、スピードは第四の軸となる。
- ディープラーニングは特殊な事例？ただし孤立した事例ではないと考える。
- 臨界点を過ぎると急速な変化が起きる。その時期を迎えた分野も出てきた。
- この未来に図書館が果たせる役割は何か？次のアクションを始めましょう！

続きはウェブで！

http://agora.ex.nii.ac.jp/~kitamoto/research/open-science/