『オープンサイエンス』とAI ～オープン化は人工知能研究をどう変えるか？～

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

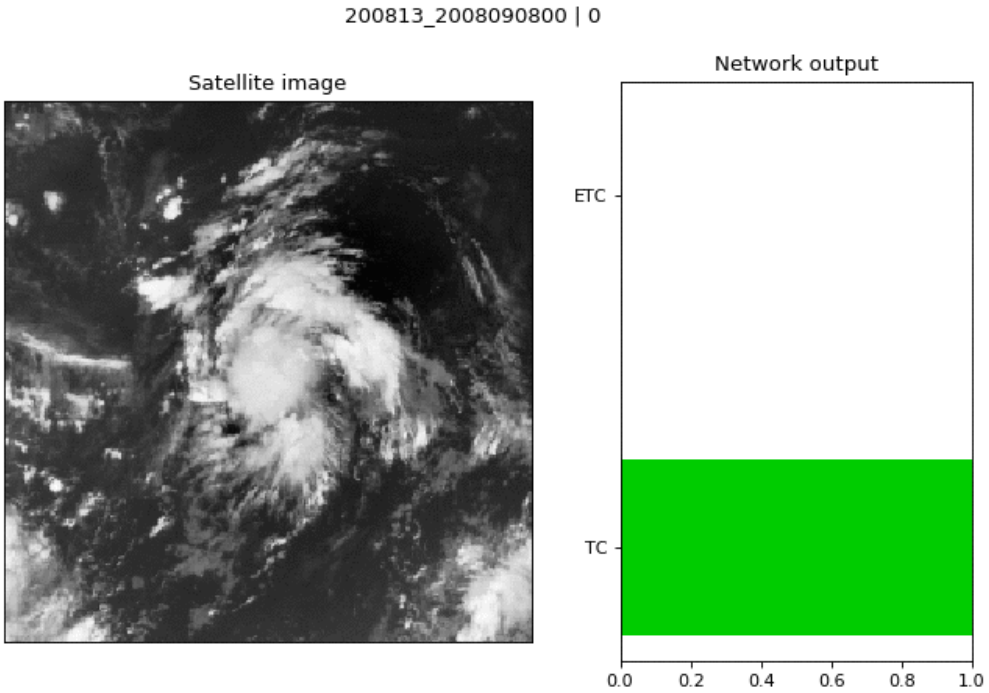
- 情報学が研究分野。他分野と協働するデータ駆動型プロジェクトが多い。
- 気象、地球環境、人文科学などの分野でデータ駆動型サイエンスを推進。
- 最近はオープンサイエンスの概念化や実践にも関わる。

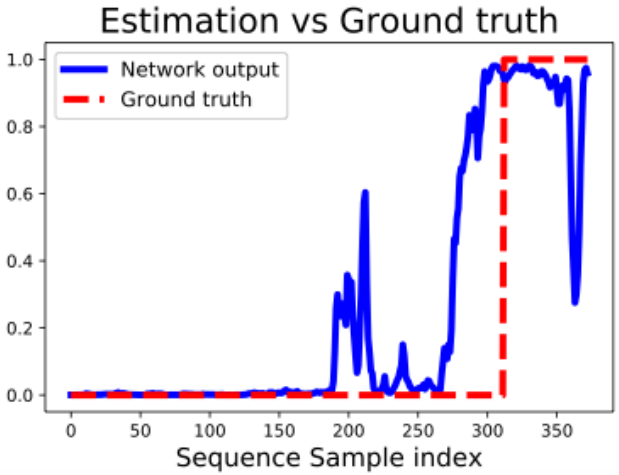
デジタル台風とは？

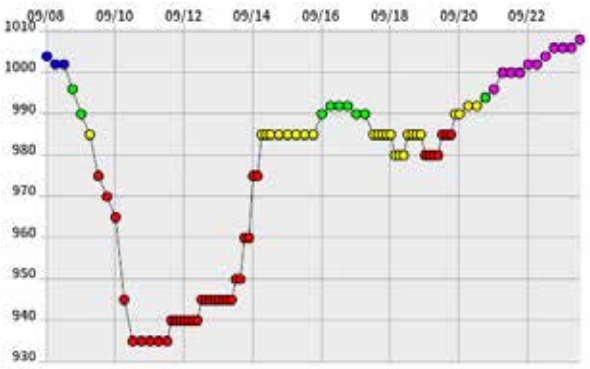
http://agora.ex.nii.ac.jp/digital-typhoon/

- 1999年から続く、台風ビッグデータ解析プロジェクト。
- 現在から過去を検索する機能＋長期データアーカイブ。
- 年間約2000万PV。多様な目的に利用されている。

台風から温帯低気圧への遷移





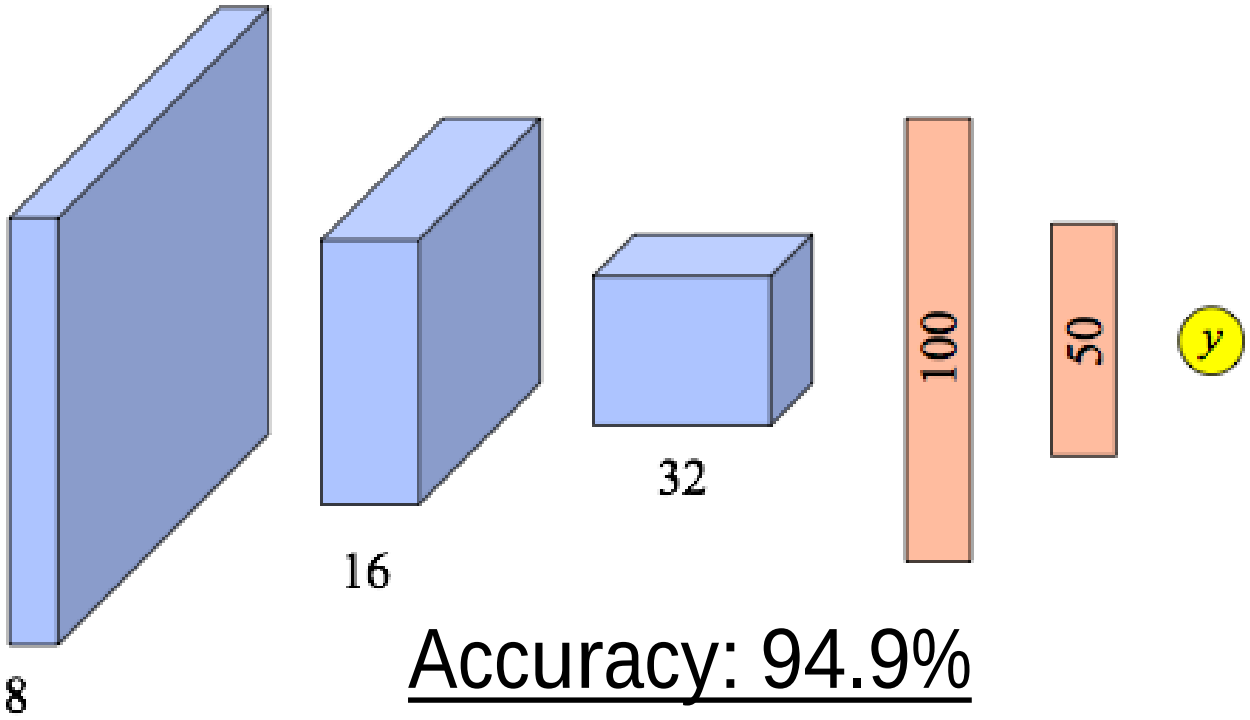


台風200813号を対象に、台風から温帯低気圧への遷移を出力。0が熱帯低気圧、1が温帯低気圧。

Collaboration with Lucas RODES GUIRAO.

台風から温帯低気圧への遷移

Model architecture (2,891,707 parameters)



Conv Layer (3x3 kernels)
ReLU
Batch Norm
Max-pooling 2x2
Conv Layer (3x3 kernels)
ReLU
Batch Norm
Max-pooling 2x2
Conv Layer (3x3 kernels)
ReLU
Batch Norm
Max-pooling 2x2
Dense Layer
ReLU
Batch Norm
Dropout 0.2
Dense Layer
ReLU
Batch Norm
Output

Accuracy: 94.9%

Collaboration with Lucas RODES GUIRAO.

人文学オープンデータ共同利用センター

CODH http://codh.rois.ac.jp/

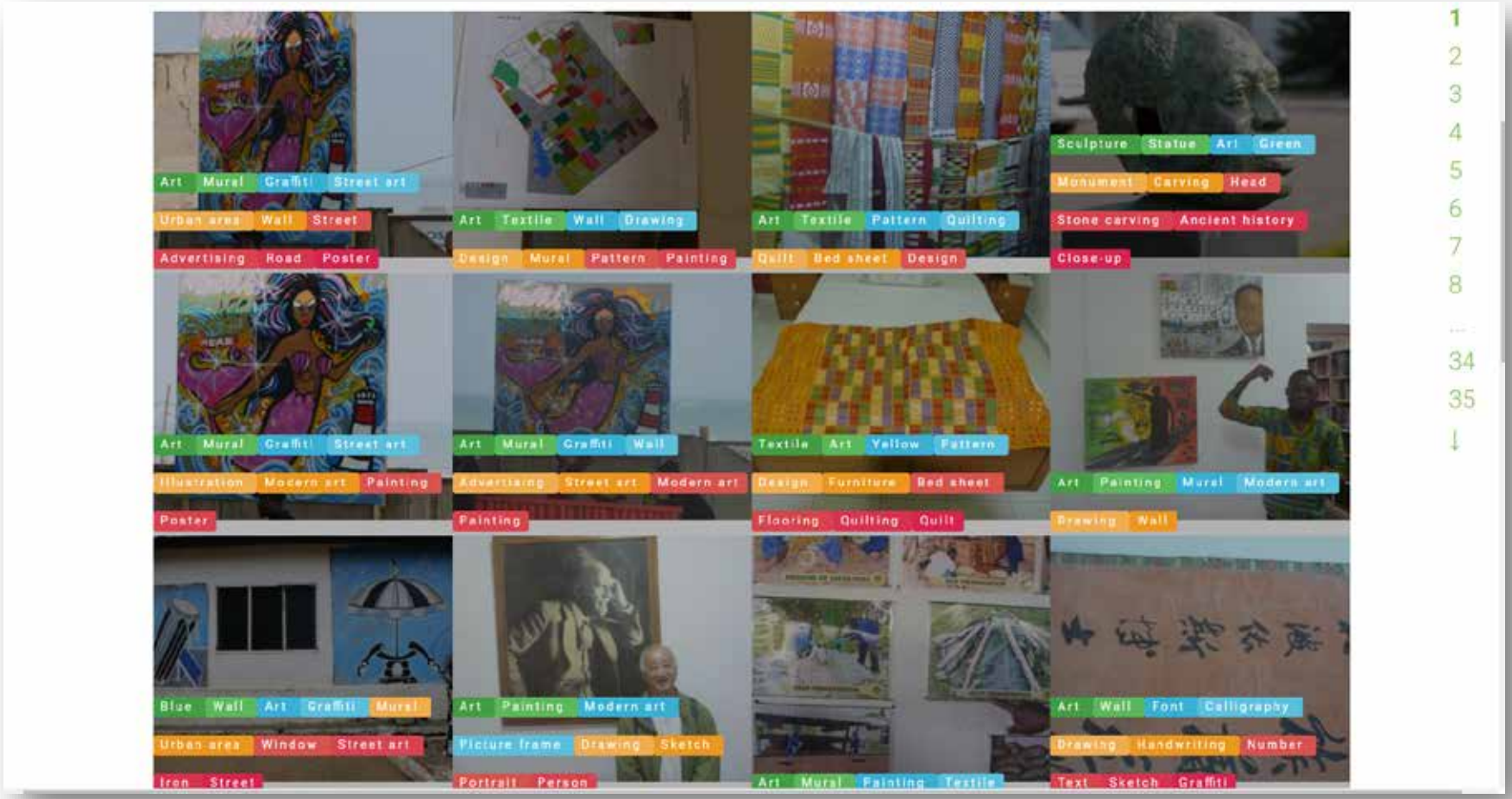
- **情報・システム研究機構 データサイエンス共同利用基盤施設内に、2017年4月1日に正式に発足。センター長：北本 朝展。**

1. **情報学・統計学の技術を用いて、人文学の研究を革新する。**
2. **人文学のデータを用いて、情報学・統計学の研究を革新する。**

古典籍くずし字の文字認識

日本古典籍データセット・日本古典籍字形データセット（国文研所蔵、CODH配信）

人類学調査写真自動タグ付け



国立民族学博物館との共同研究

今日お話ししないこと

1. AIでどんなビジネスが生まれるか？
2. AIにより人間は職を失うのか？
3. AI時代はベーシックインカムなのか？
4. AI**は将来的に人間を越えるのか（シン**ギュラリティ仮説）？

- **これらのトピックは、他の調査プロジェクトですでに議論されているため。**

今日お話しすること

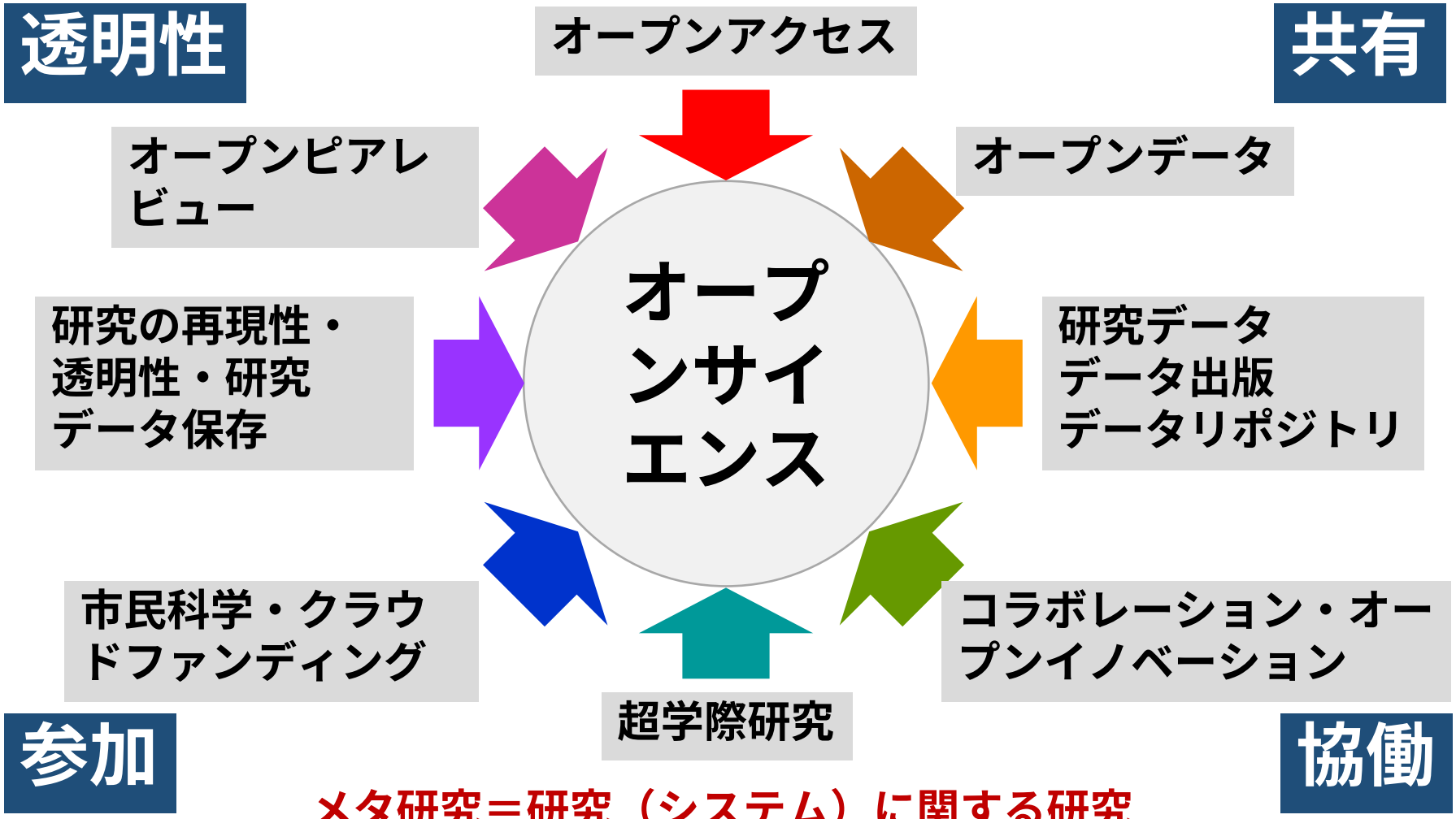
1. **オープンサイエンスの背景**
2. **AIとオープンソース**
3. **AIとオープンアクセス**
4. **AIと透明性**
5. **オープンサイエンスの推進力**

1. 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



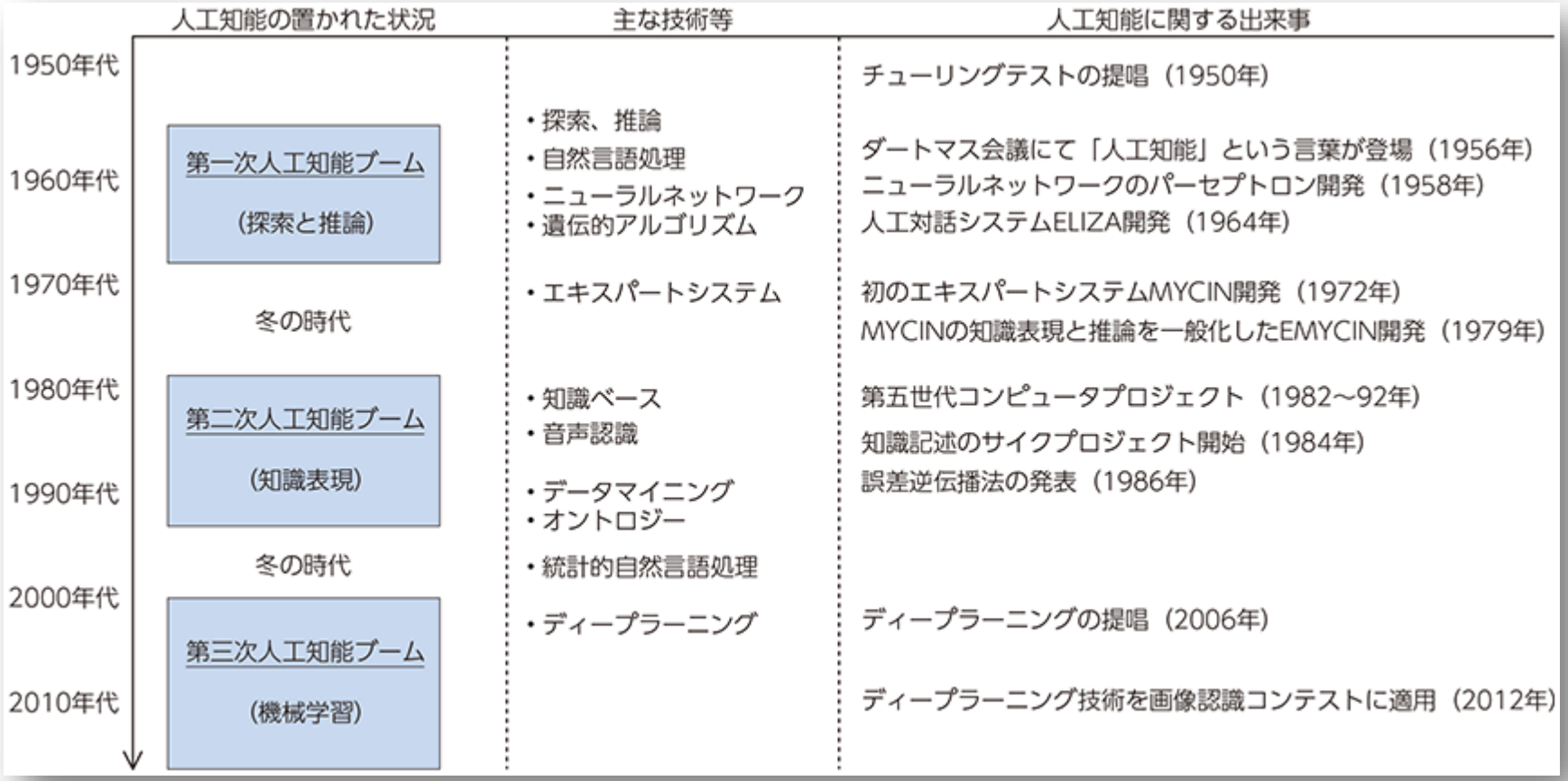
メタ研究＝研究（システム）に関する研究

オープン化の4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。
4. **摩擦を減らしてスムーズに協働（スピード）**
   - オープンなコラボレーションをスムーズに進めるために、障壁や摩擦になる部分を減らしていく。

2. AIとオープンソース

人工知能 Artificial Intelligence



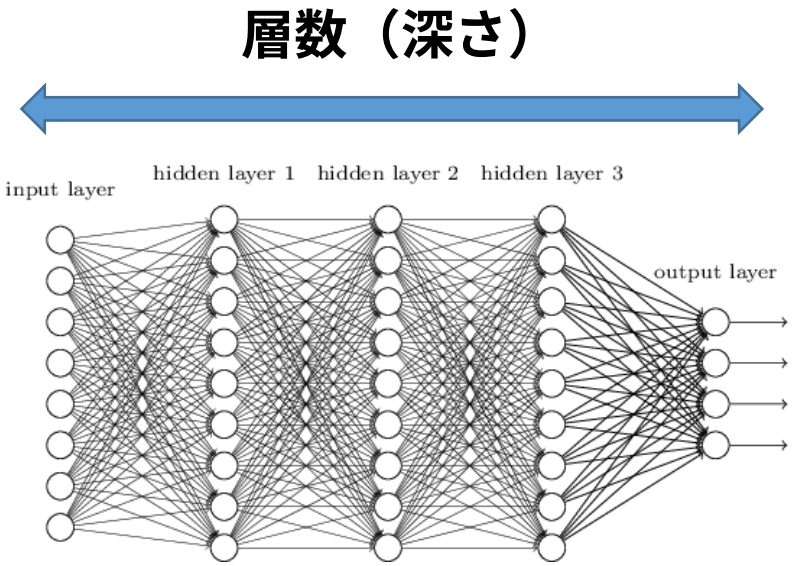
出典：平成28**年度** 総務省通信白書：
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc142120.html

人工知能の様々な手法

1. 第一次ブーム＝探索と推論
2. 第二次ブーム＝知識表現
3. 第三次ブーム＝機械学習

- 機械学習とは、問題と正解のセットから、自動的に問題の答え方を学習する方法。
- 見たことのある問題は答えられるが、見たことない問題への答えは簡単ではない。

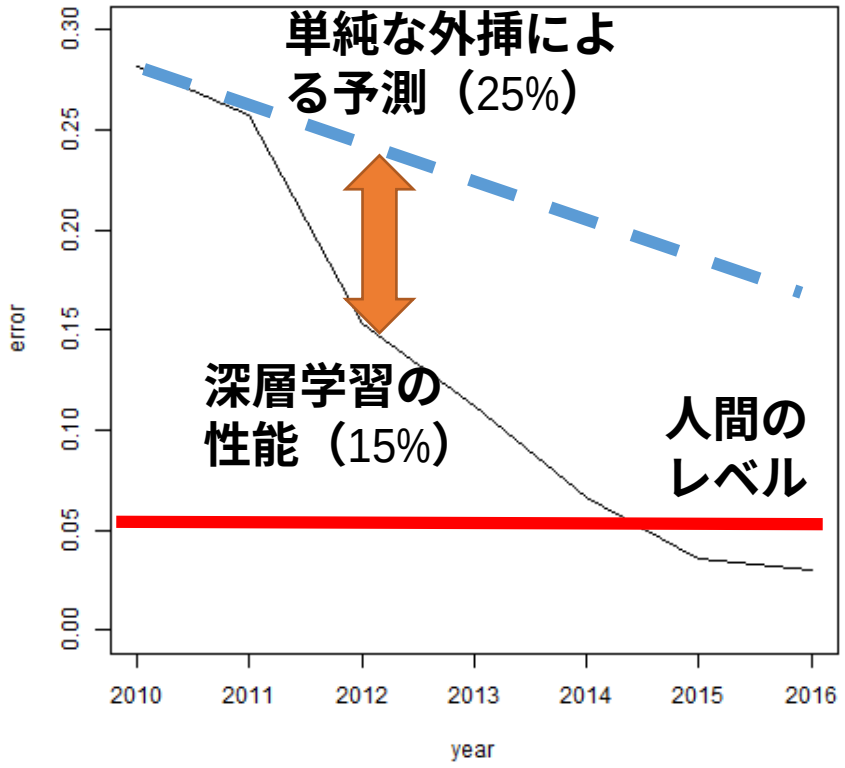
ディープラーニング登場



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

- **機械学習の一手法**であり、**ニューラルネットワーク**の中でも特に層が多いもの（深層）を指す。
- **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。

画像認識の画期的な性能向上



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。

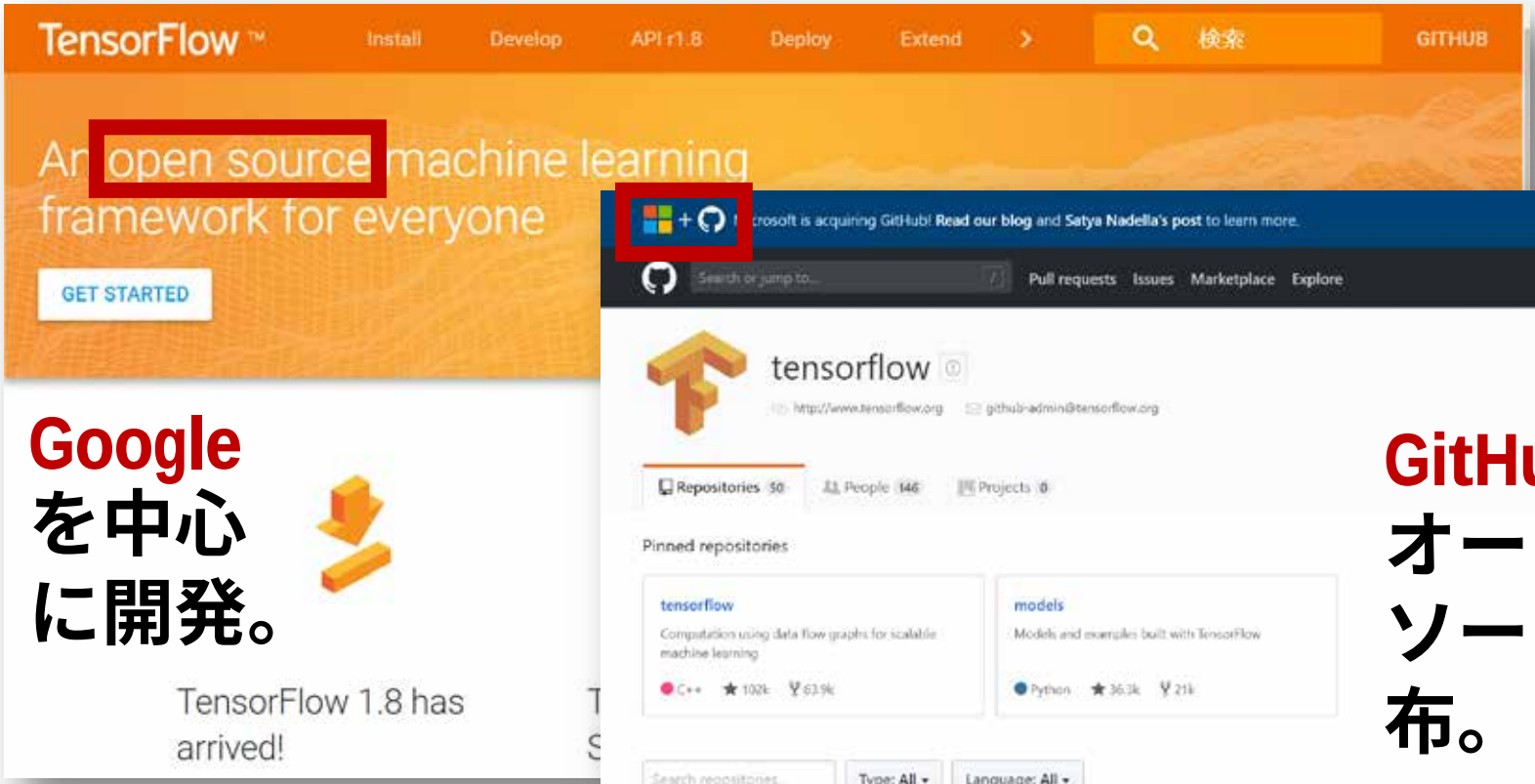
AlphaGoの衝撃

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）

TensorFlowとオープンソース



Googleを中心に開発。

GitHubでオープンソース配布。

https://www.tensorflow.org/

オープンソースが台風の目

International

日本経済新聞

2018年6月6日（水）

トップ　経済・政治　ビジネス　マーケット　テクノロジー　国際・アジア　スポーツ　社会

マイクロソフト、開発者向け共有サイト買収　8200億円

ネット・IT 北米

2018/6/5 0:45

保存　共有　印刷　その他

【シリコンバレー＝白石武志】米マイクロソフト（MS）は4日、ソフト開発者が設計図（ソースコード）を公開・共有できるサイトを運営する米ギットハブを75億ドル（約8200億円）で買収すると発表した。MSは世界で約2800万人が利用するギットハブを取り込むことで、ソフト開発者向けのクラウドサービス事業を強化する。

ギットハブは2008年の設立。スマートフォンの普及などとともに、無償公開し自由に改良できる「オープンソースソフトウエア」の開発や普及の基盤になってきた。開発者同士が情報をやりとりするソーシャル・ネットワーキング・サービス（SNS）の機能も果たしている。

買収手続きは米国や欧州の独禁当局の認可を前提に18年中に完了する予定で、ギットハブの新たな最高経営責任者（CEO）にはMS副社長のナット・フリードマン氏が就任する。MSは買収後もギットハブの利用者が米アマゾンウェブサービスや米グーグルなどのクラウドサービスも利用できるようにするなど、ギットハブの経営の独立性は維持するという。

日経新聞, 2018年6月5日
https://www.nikkei.com/article/DGXMZO31366610V00C18A6FF8000/

- **マイクロソフトがGitHubを8200億円で買収。**
- **優れた研究者、技術者はGitHubで成果を共有している。**
- **ソフトウェアの利用だけでなく、人材発掘にも有効？**

オープンソースとは何か？

```
Status InvGrad(const Scope& scope, const Operation& op,
               const std::vector<Output>& grad_inputs,
               std::vector<Output>* grad_outputs) {
  // Use the built-in operator.
  grad_outputs->push_back(
      internal::ReciprocalGrad(scope, op.output(0), grad_inputs[0]));
  return scope.status();
}
REGISTER_GRADIENT_OP("Inv", InvGrad);
REGISTER_GRADIENT_OP("Reciprocal", InvGrad);

Status SquareGrad(const Scope& scope, const Operation& op,
                  const std::vector<Output>& grad_inputs,
                  std::vector<Output>* grad_outputs) {
  // dy/dx = (2 * x)
  auto two = Cast(scope, Const(scope, 2), op.input(0).type());
  auto dydx = Mul(scope, two, op.input(0));
  // grad(x) = grad(y) * conj(dy/dx)
  grad_outputs->push_back(
      Mul(scope, grad_inputs[0], ConjugateHelper(scope, dydx)));
  return scope.status();
}
REGISTER_GRADIENT_OP("Square", SquareGrad);

Status SqrtGrad(const Scope& scope, const Operation& op,
                const std::vector<Output>& grad_inputs,
                std::vector<Output>* grad_outputs) {
  // Use the built-in operator.
  grad_outputs->push_back(
      internal::SqrtGrad(scope, op.output(0), grad_inputs[0]));
  return scope.status();
}
REGISTER_GRADIENT_OP("Sqrt", SqrtGrad);
Status RsqrtGrad(const Scope& scope, const Operation& op,
```

https://github.com/tensorflow/tensorflow/blob/master/tensorflow/cc/gradients/math_grad.cc

- **ソースコード**：コンピュータへの命令を文字列で書いたもの。
- **オープンソース**：その内容が誰でも見られる＝コピーできる。
- **知的財産がコピーできる？**GitHubを買収したマイクロソフトは、当初この考えを敵視した。

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。

市民もSNS等で簡単に情報共有

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。。

研究者の焦り

- **オープンソース**として各種ライブラリがダウンロード可能。各種の実験コードもオープンソース。誰でもいつでも試せる。
- **共通基盤データ**（例ImageNet）もオープン化。誰でもいつでも試せる。
- **応用分野（囲碁その他）が急速に広がり**、多くの分野の研究者や技術者が大挙参入。
- **一刻も早く成果を世界に公表せねば！！**

研究の爆速化と摩擦の低減

- **研究のスピードが極限まで高速化**すると、情報流通も同時に高速化する必要がある。
- **研究プロセス上の摩擦を減らそうとすると、研究は結果的にオープン化する。**
- 毎日新しい結果が出る状況では、**査読を待てず即時オープン化**せざるを得ない。
- **ディープラーニング（深層学習）**の分野では、特にこの傾向が顕著。

3. AIとオープンアクセス

研究成果の公表と共有

1. **学術論文**：査読を通れば出版でき、最も伝統的かつ権威がある。
2. **学会発表**：情報系では、有力国際会議での（査読有）発表にも権威がある。
3. **プレプリント**：正式に出版する前の原稿。査読がないので素早く公表できる。
4. **その他**：ネットの誕生により、公表ルートが非常に多様化した。

シリアルズ・クライシス

スウェーデン・Bibsamコンソーシアム、Elsevier社との契約を更新しないと発表

Posted 2018年5月17日

スウェーデン・Bibsamコンソーシアムが、Elsevier社との契約を更新しないと発表しています。

Bibsamコンソーシアムを代表してライセンス契約の交渉を行っているスウェーデン王立図書館（NLS）が2018年5月16日に発表したもので、同国政府が設定した2026年までのオープンアクセス（OA）を達成するための要件を満たすモデルをElsevier社が提示できなかったことによるものです。

参加館の研究者は現在の契約条件に基づき、引き続き1995年から2017年にかけて発行された論文にはアクセスできますが、2018年6月30日以降にElsevier社のプラットフォームで公開された論文は利用できなくなります。

Sweden stands up for open access - cancels agreement with Elsevier（NLS OpenAcccess.se，2018/5/16）
http://openaccess.blogg.kb.se/2018/05/16/sweden-stands-up-for-open-access-cancels-agreement-with-elsevier/

http://current.ndl.go.jp/node/36014

学術雑誌の購読料は毎年値上げ。出版社は足元を見つつ、利益を増やしている。

貴重な研究費が購読料として流出している。出版社と戦いつつも、代替案を考えるべきではないか？

ドイツ・DEALプロジェクト、2017年末をもってElsevier社との契約を延長しない国内機関のリストを公開

Posted 2017年8月28日

2017年8月21日、ドイツにおけるナショナルライセンス契約を目的とするプロジェクトDEAL（Projekt DEAL）が、同プロジェクトのウェブサイトにおいて、2017年末をもってElsevier社との契約を延長しない国内機関のリストを公開しています。

Vertragskündigungen Elsevier 2017（Projekt DEAL，2017/8/21）
https://www.projekt-deal.de/vertragskundigungen-elsevier-2017/

via;
Project Deal: Germany's Largest Scientific Organization Announces It Will Terminate Elsevier Contract at the End of 2017(indoDOCKET，2017/8/25)
http://www.infodocket.com/2017/08/25/largest-german-research-organization-announces-it-will-terminate-elsevier-contract-at-the-end-of-2017/

http://current.ndl.go.jp/node/34579

権威ある学術雑誌

Nature, Volume 557 Issue 7707, 31 May 2018

1. 読者数が多く、歴史的な蓄積もある。
2. 良い論文がこれまで多数掲載。自分も載りたい。
3. **インパクトファクター**が高い。みな引用する。
4. **日本の研究力指標**にもよく使われる。

Zero-cost digital publishing options could help cope with an explosion of artificial intelligence papers. GURZZZA/ISTOCK

Why are AI researchers boycotting a new *Nature* journal—and shunning others?

By Matthew Hutson | May. 17, 2018 , 12:10 PM

2018年5月17日

Computer science was born of a rebellious, hacker culture, a spirit that lives on in the publishing culture of artificial intelligence (AI). The burgeoning field is increasingly turning to conference publications and free, open-review websites while shunning traditional outlets—sentiments dramatically expressed in a growing boycott of a high-profile AI journal. As of 15 May, about 3000 people, mostly academic computer scientists, had signed a petition promising not to submit, review, or edit articles for *Nature Machine Intelligence* (*NMI*), a new journal from the publisher Springer Nature set to begin publication in January 2019.

http://doi.org/10.1126/science.aau2005

時代に逆行？

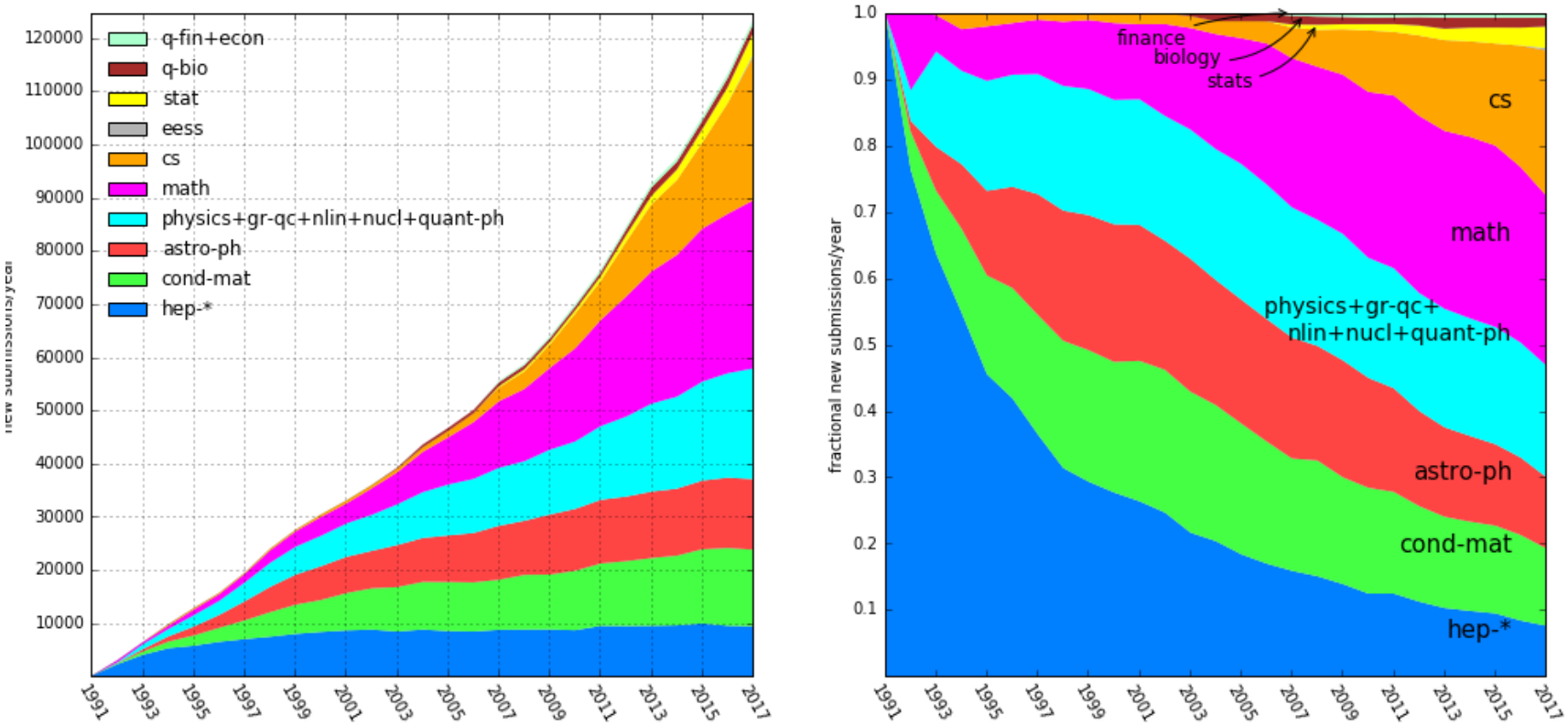
- Natureが新たに機械学習の有料雑誌の創刊を計画。
- AI研究者が、投稿、査読、編集等のボイコットを呼びかける事態に。
- なぜ既存の権威が通用しないのか？

プレプリントサーバarXiv

https://arxiv.org/

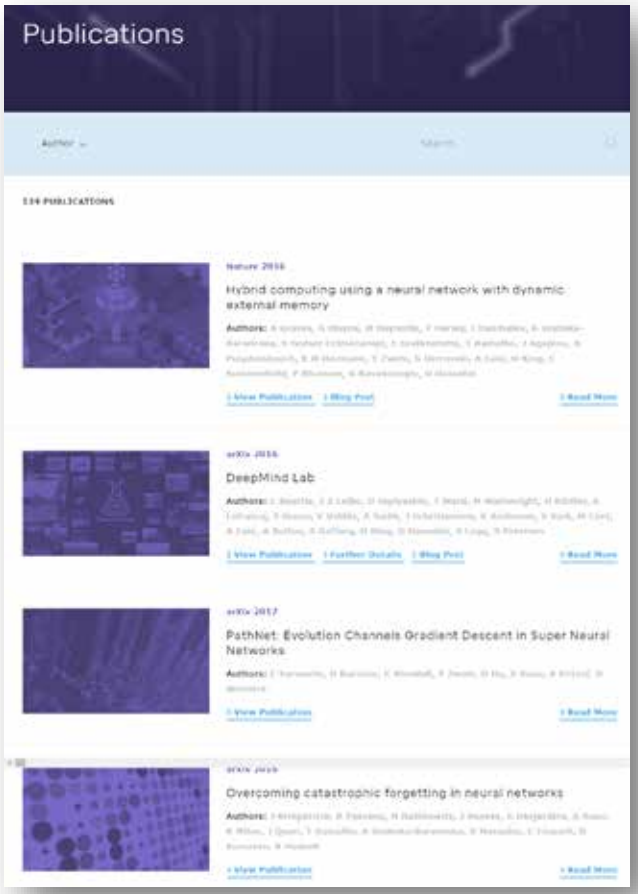
- 1991年登場の元祖プレプリントサーバ。現在はコーネル大学運営。
- 元々は物理学論文対象、後に他分野に拡大。
- 査読前論文をオープンアクセス化。よほど不適格な論文以外は掲載。

arXivへの投稿推移



Left: number of new submissions/year as a function of calendar year. Right: ubmission rates divided by the total for each year, giving the fractional submission rates for each of the domains. https://arxiv.org/help/stats/2017_by_area/index

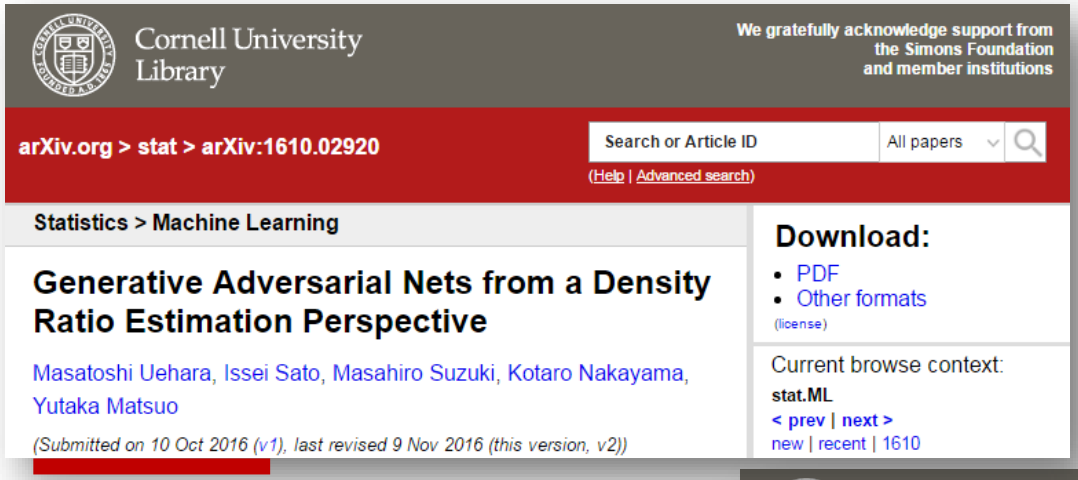
AI研究はarXivが主戦場



- AI研究の中心の一つDeepMindでは、**権威ある学術雑誌とarXivが同格に並んでいる。**
- arXivにまず成果を公表し、査読は後で必要に応じて受ける。
- **研究成果公表が「即時オープン」にシフト。**

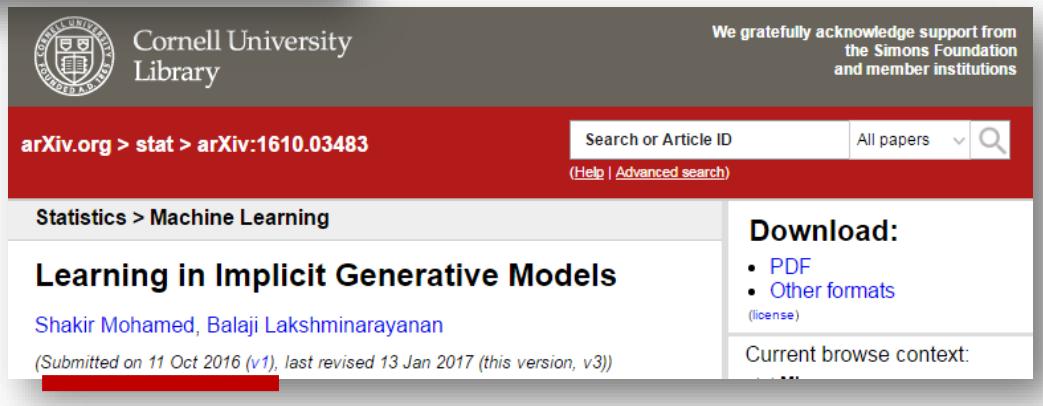
https://deepmind.com/research/publications/

論文の即時引用



https://arxiv.org/abs/1610.02920

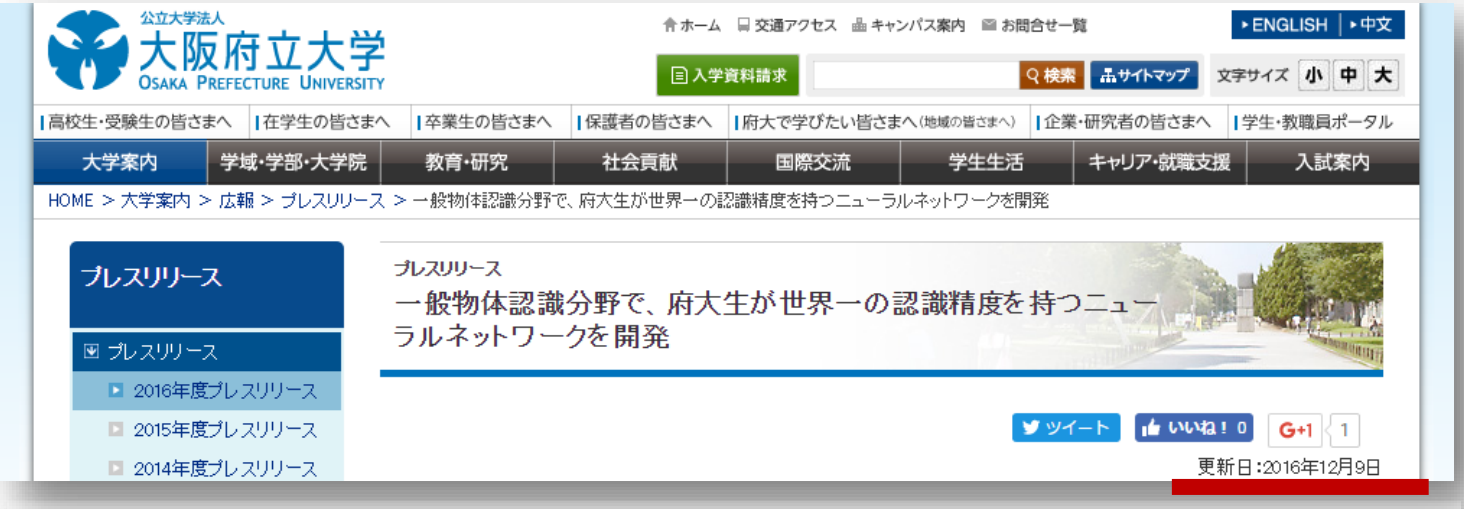
https://arxiv.org/abs/1610.03483



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！

成果の即時プレスリリース



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

30年前の大フィーバー

2008/08/18 09:00

日経エレが目撃した電子産業・歴史の現場

【電子産業史】1986年：高温超電導の発見

世紀の発見から20年，ブレークスルーの先には

日経エレクトロニクス

1986年，高温超電導体が発見された。

スイスのIBM Zürich研究所が発見したLaBaCuOの30K付近の抵抗異常を，東京大学が超電導だと確認したのが1986年11月。東大はこの結果を12月5日に米国ボストンで開かれていた材料関連の学会「MRS（Material Research Society）」で緊急発表した。高温超電導フィーバーと呼ばれた時期は，このときから翌1987年の秋ごろまでの間だと思う。

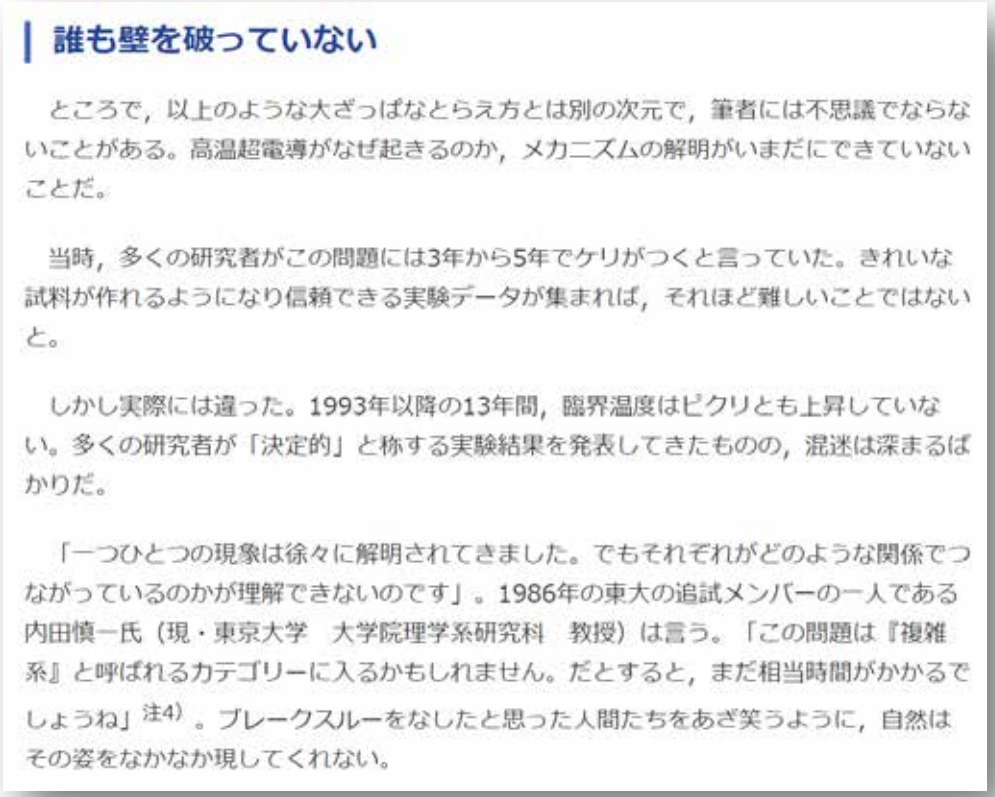
MRSでの東大の発表を受けて世界中で新しい超電導物質の探索が始まり，2カ月後の1987年2月には，米University of HoustonがYBaCuO化合物で液体窒素温度（77K）を超える84Kの臨界温度を確認したと発表した。そして翌3月18日に，歴史的な米国物理学会「APS（American Physical Society）」がニューヨークで開かれる。臨時シンポジウムに集まった2000人を超す参加者が夜7時半から翌未明まで8時間近く議論を続けた。米Business Week誌はこれを「物理学者のウッドストック」と呼んだ。

この時期をピークとする半年間ほど，学会は機能を停止する。研究者はファクスでプレプリントを配布し，実験結果を報道機関へリークした。超電導の臨界温度はいつの間にか数百Kも上昇し，室温でも超電導は起こり得るという期待感が広まった[注1)]。高名な学者が「トランジスタの発明よりもはるかに波及効果は大きい」と語り，マスコミは21世紀の初めに数十兆円の市場が生まれると予測した[注2)]。

http://tech.nikkeibp.co.jp/dm/article/COLUMN/20080807/156207/

- 1986年高温超電導体発見で大フィーバー発生。
- 臨時シンポジウムは、夜から翌朝まで会議。
- 学会は機能を停止。
- **研究者はファクスでプレプリントを配布。**
- 実験結果は報道機関にリーク。

高温超電導研究のその後



誰も壁を破っていない

ところで，以上のような大ざっぱなとらえ方とは別の次元で，筆者には不思議でならないことがある。高温超電導がなぜ起きるのか，メカニズムの解明がいまだにできていないことだ。

当時，多くの研究者がこの問題には3年から5年でケリがつくと言っていた。きれいな試料が作れるようになり信頼できる実験データが集まれば，それほど難しいことではないと。

しかし実際には違った。1993年以降の13年間，臨界温度はピクリとも上昇していない。多くの研究者が「決定的」と称する実験結果を発表してきたものの，混迷は深まるばかりだ。

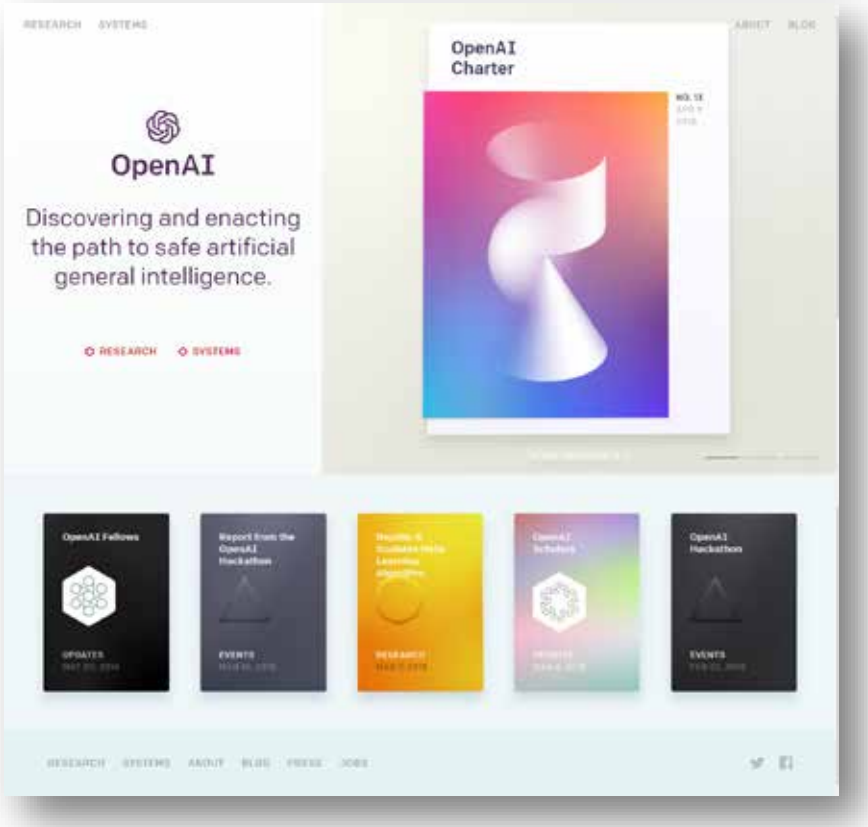
「一つひとつの現象は徐々に解明されてきました。でもそれぞれがどのような関係でつながっているのかが理解できないのです」。1986年の東大の追試メンバーの一人である内田慎一氏（現・東京大学　大学院理学系研究科　教授）は言う。「この問題は『複雑系』と呼ばれるカテゴリーに入るかもしれません。だとすると，まだ相当時間がかかるでしょうね」[注4]。ブレークスルーをなしたと思った人間たちをあざ笑うように，自然はその姿をなかなか現してくれない。

http://tech.nikkeibp.co.jp/dm/article/COLUMN/20080807/156207/?P=2

- メカニズムの解明は3年から5年でケリがつく＝楽観的。
- **実際は今も混迷が深まるばかり。**
- **本当に難しい問題はまだ解けてない。**
- AIも数年後にこうならないか心配。

4. AIと透明性

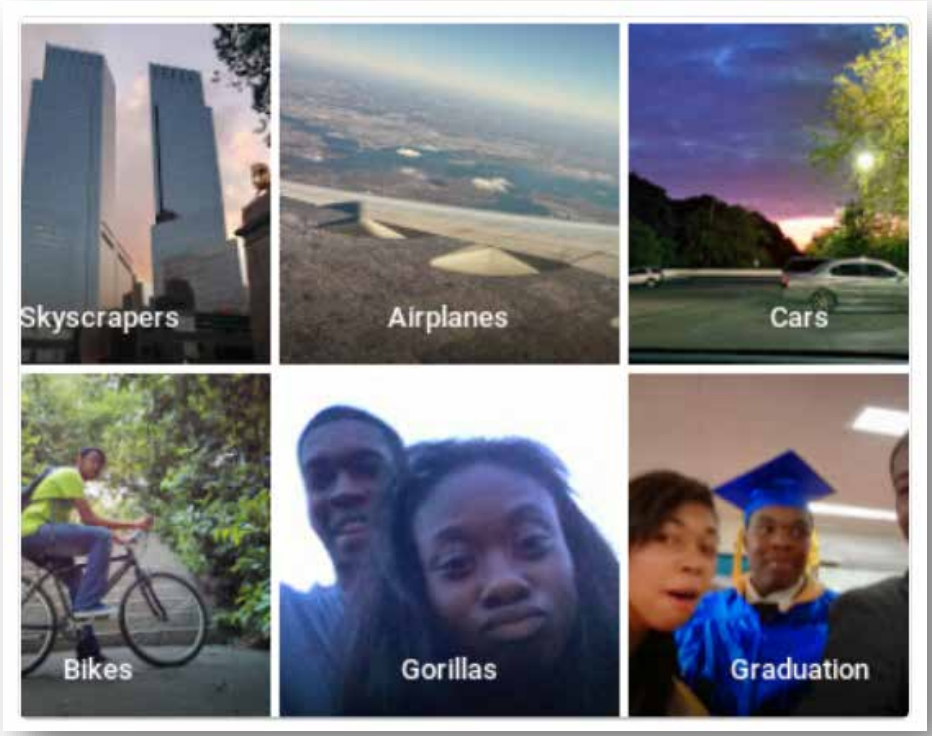
非営利団体によるAI



https://openai.com/

1. **一部の企業が技術を独り占めするのは人類にマイナス。**
2. **すべての成果をオープンにし、人類に貢献する。**
3. **安全なAIの構築に向けて、オープンな研究を探る。**

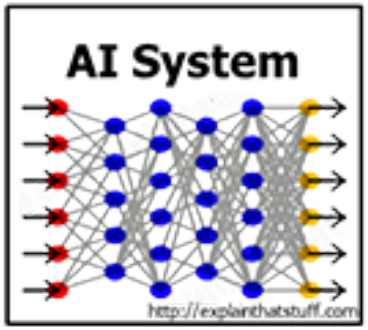
AIの倫理的問題



https://twitter.com/jackyalcine/status/615329515909156865/

http://www.itmedia.co.jp/news/articles/1603/25/news069.html

ブラックボックスの問題



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

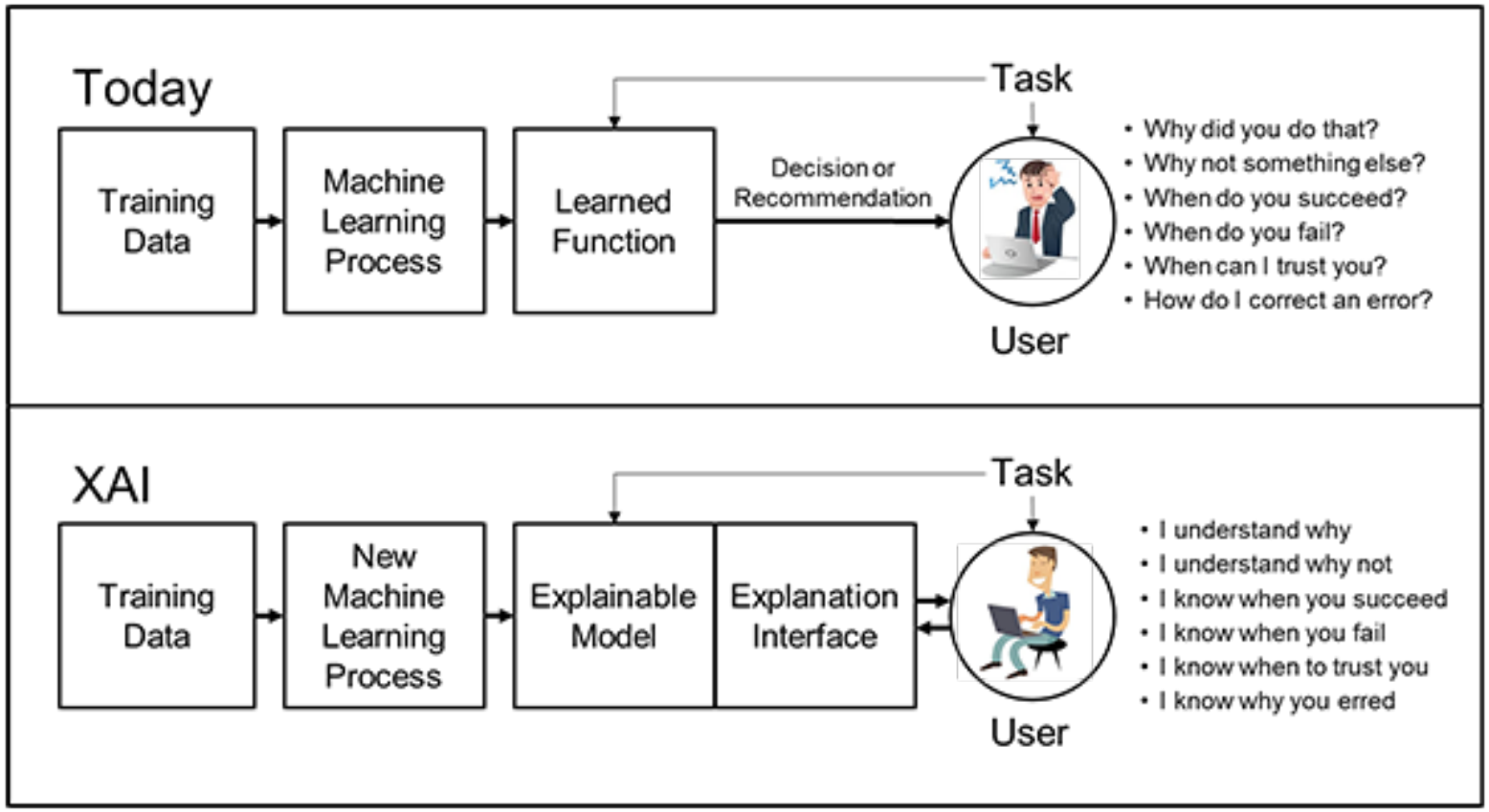
DoD and non-DoD Applications

- Transportation
- Security
- Medicine
- Finance
- Legal
- Military

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

The Need for Explainable AI : https://www.darpa.mil/program/explainable-artificial-intelligence

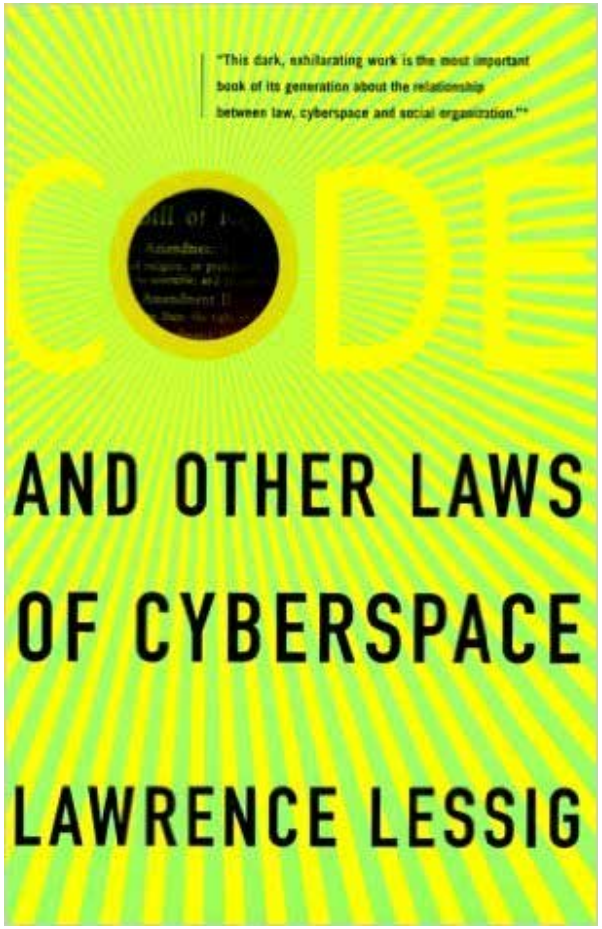
説明責任を果たすAI



The Need for Explainable AI : https://www.darpa.mil/program/explainable-artificial-intelligence

5. オープンサイエンスの推進力

制度を分析する4つの視点

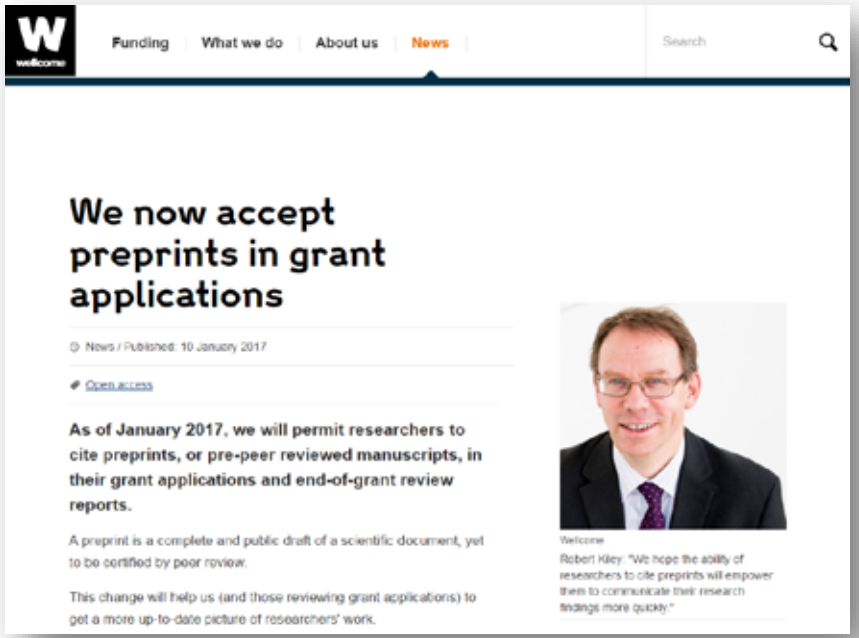


- Lawrence Lessig (Founder of Creative Commons), *Code: And Other Laws of Cyber Space* (first edition 1999)
- **法＝しなければならない**
- **規範＝すべきである**
- **市場＝した方が利益がある**
- **アーキテクチャ＝せざるを得ない**

「法」によるオープン化

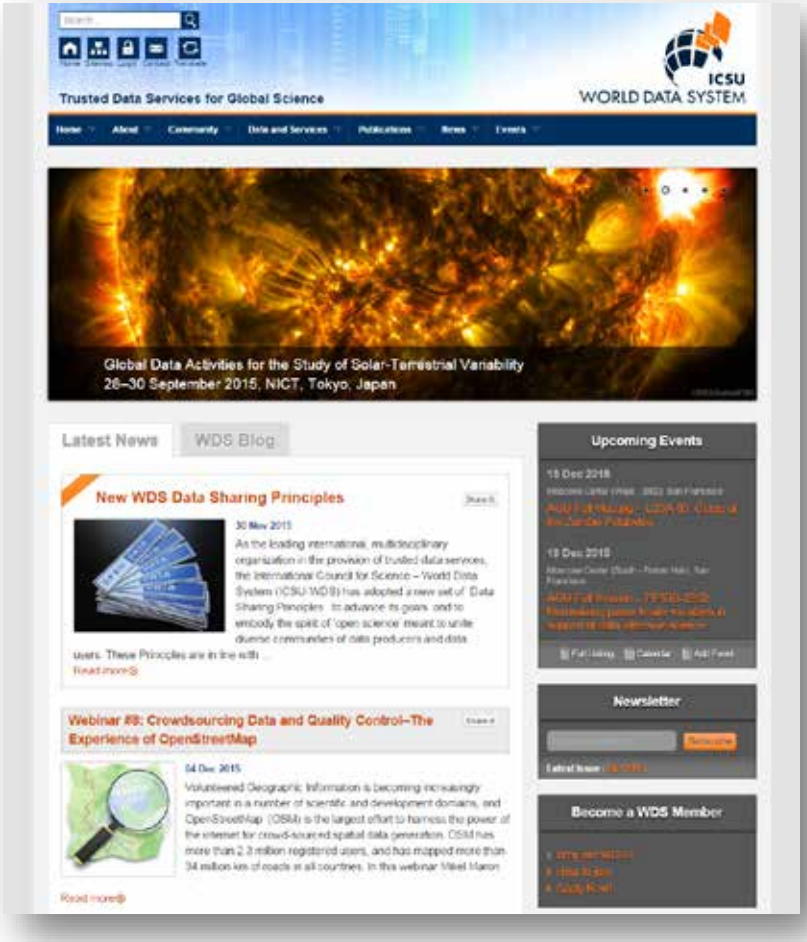
http://www.nature.com/news/gates-foundation-research-can-t-be-published-in-top-journals-1.21299

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

「規範」によるオープン化



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

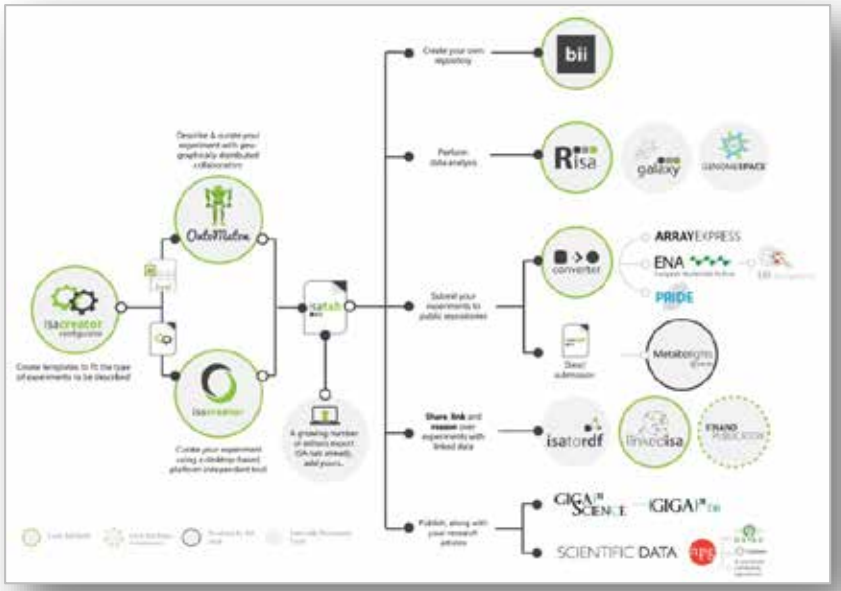
- **オープンな文化**：データ共有が不可欠な分野もある。
- **世代の差**：若い世代では共有文化の経験がより強い。
- **文化の差**：異なる文化圏に対する説得力が弱い。

「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。
- **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。

オープン化を担うインフラ

1. **計算インフラ**：スーパーコンピュータ、大規模クラウド、AI向け高速計算機（GPU/TPU）
2. **データインフラ**：大規模ディスク、長期保存システム
3. **ネットワークインフラ**：超高速ネットワークによるデータ収集・共有・配信
4. **知識インフラ**：高度処理のためのソフトウェア、研究資料、知識体系アーカイブ
5. **法制度インフラ**：プライバシー、著作権

オープン化を担う人材

1. **AI研究者**：大学・研究機関よりも民間企業の方が研究環境が充実？人材の移動も話題。
2. **データサイエンティスト**：技術、ビジネス、**システム化のバランスが取れた人材が必要だが、全く足りていない。**
3. **データライブラリアン・データキュレーター**：図書館などにおける情報整理の専門スキルを活かせないか？
4. **評価の問題**：**オープンサイエンスでは研究のやり方が変わるため、評価基準も変えるべき。**

オープン vs. クローズ

オープン	クローズ
1. ソフトウェア	1. データ
2. プレプリントサーバ	2. 有料査読付き論文
3. アカウンタビリティ	3. ブラックボックス

1. **データをクローズド**にしておけば、ソフトウェアをオープンソースにしても競争に負けることはない。
2. **一部の学術雑誌はクローズド**で高価すぎるので、プレプリントサーバというオープンな方法を使おう。
3. ディープラーニングの動作は**ブラックボックスで説明できず**、アカウンタビリティを果たせていない。

オープンサイエンスとAI

- 研究を高速化するには、**プロセスを減速させる摩擦を減らす**必要がある。
- 摩擦を減らす方向に進化した結果、**AI研究は結果的にオープン化しつつある**。
- ただし全面的にオープンではなく、**クローズな部分が利益の源泉**となる場合も。
- 高速化を妨げる障壁や摩擦を減らし、世界で戦えるインフラと人材を日本にも！

参考リンク

- 研究室ウェブサイト
  - http://agora.ex.nii.ac.jp/~kitamoto/
- Researchmap
  - http://researchmap.jp/kitamoto/
- オープンサイエンス
  - http://agora.ex.nii.ac.jp/~kitamoto/research/open-science/
- 人文学オープンデータ共同利用センター
  - http://codh.rois.ac.jp/