3-2　AIと社会

東京大学　数理・情報教育研究センター
2021年5月1日

概要

- データやAIは，強力な道具であるだけに，使い方を誤ると人間や社会に大きなダメージを与えるおそれがあります．この節では，データやAIを使うにあたり最低限気をつけるべきことについて学びます．
- データやAIにまつわる基本的な倫理，合意事項について学びます．
- データを守ること，およびそれが破られて起こった事例について学びます．

本教材の目次

データ・AIを扱う上での留意事項

概要

- 本節では，現代科学・技術が社会に対して果たすべき役割について考え，特にデータやデータサイエンス・AIを利活用する際に求められるモラルや倫理について理解することを目指します.
- いくつかの具体的事案を見ながら，データ駆動型社会における脅威（リスク）についても学びます.
- さらに，それらの流れを踏まえて，個人情報保護法やEU一般データ保護規則（GDPR）など，最先端の，データを取り巻く考え方や指針・法についても学びます.

3-1-1. ELSI：すべての科学・技術に関する普遍的考え方

“ELSI” = “Ethical (倫理的), Legal (法的), Social (社会的)
Implications (含意) あるいは Issues (事柄)”
= 「（科学における）倫理的・法的・社会的含意（事柄）」
= 「（科学技術を開発・展開した結果）
起こりうる倫理的問題・法的問題・社会的問題は何か」

1980年代に生命科学分野で，科学技術のみならずその社会的責任を考える必要を認め，提唱された概念です.
いまでは，あらゆる科学分野で必要とされています.

(参考；復習）従来の研究倫理の概念

従前より研究倫理として定着していたこと：

- 不公正・不適切な研究行為・発表の禁止：
  捏造・改竄・剽窃
  不適切な著者表示（ギフトオーサー，ゴーストオーサー）
  不適切なプレゼンテーション（チャンピオンデータ等）
- ヒト・生物に対する不適切な研究の禁止（→研究倫理審査）
- 違法な研究の禁止
- 利益相反

これらは**科学研究単体としての「健全性」「妥当性」.**
ELSI が考えるのは，より積極的な,
「社会の構成要素としての科学研究のあり方」

なぜELSIの考え方が必要なのか

[事例１：ヒトゲノムプロジェクト] 1990年頃，アメリカ

- ヒトゲノムの解読は，研究者や医療行為のみならず，すべての人，社会全体に影響が及ぶ.
  （遺伝情報＝個人情報の保護，遺伝情報差別の防止，…）
- そのため，研究者や医師・患者だけでなく，広い意味で社会がどこまでを「受容できるか」を議論する必要がある.

[事例２：原子力発電] 2011年，日本

- 東日本大震災による福島第一原子力発電所事故.
  日本でも科学技術・研究者に対する信頼が低下.
- 原子力発電技術はどこまで「受容できるか」の議論が必要.

社会は，科学技術をどこまで・どのように受け容れられるのか
科学は，社会にどこまで責任を持つのか
これらを議論する必要があります.

科学の責任：古典的な見方・より広い見方

【考え方１】科学は「直接意図した結果」にだけ責任を持つ.
→「狭い」「古典的」な見方
→想定外の事態まで責任は取れない
→科学研究は「善・悪」から切り離された客観的存在

【考え方２】「予想外だが引き起こされてしまった結果」にも責任を持つ.
→「より広い」「フォージの」見方
→過去の悲劇を振り返り，合理的に予想できる結果には責任を
→客観的事実の研究で含意がないから許されるとは言えない

（参考）ジョン・フォージ（オーストラリアの科学哲学者）

軽々にどちらとは言えません.
ただ，前ページの例を見ると，
「古典的な見方」だけでよいと言い切れないのは確かです.

あらためてELSIの３要素を見ると

- 倫理的問題（Ethical）：
  その科学（技術）研究がどのような倫理的問題をはらむか
  そもそも各科学領域における「倫理」とはなにか
  既存の倫理学の体系に収まるのか

- 法的問題（Legal）：
  その科学（技術）はどの法の枠組で捉えられるか
  現行法で不十分な場合，どのような指針・法の形成が必要か
  （関連して，どのような行政規制が必要か）

- 社会的問題（Social）：
  その科学（技術）は社会に受容されるか
  どのような形であれば受容されるのか
  メリットとデメリットのトレードオフ：社会の受容ラインはどこか

「より広い見方」に向けて，３つの観点に分けて，
科学と社会の関係を議論する場を提案しています．

データサイエンス・AIにおけるELSIとは

本教材でこれから議論していくこと：

- データサイエンス・AIの倫理とはなにか
- その法整備はどうなっているのか
- どのようなデータサイエンス・AIが社会に受容されるのか

3-1-2. データの倫理

まず考えるべきこと：

- データ取り扱いの健全性：　禁止事項（既出）
  捏造（ないデータを作り出すこと）
  改竄（実データを曲げて書き換えること）
  剽窃・盗用（データを不正に使い回すこと）

- データの保護（情報セキュリティ的な）
  →後述：第3-2節

- 個人情報とプライバシーの問題
  →後述：本節後半

データ取り扱いの健全性１：捏造

ないデータを作り出すこと.

実例：高温超伝導事件（米，2001年頃）

- ある若手物理学研究者が起こした事件.
- それまでの（有機物）高温超伝導記録（-240°C）を大幅に覆し，-156°Cを報告．（参考：安価な液体窒素が-196°C）
- これを含み，Nature誌7本，Science誌9本等の論文を発表
- しかし該当する実験装置は発見されず，捏造が発覚した.
- 論文は撤回され，当該研究員は解雇された.

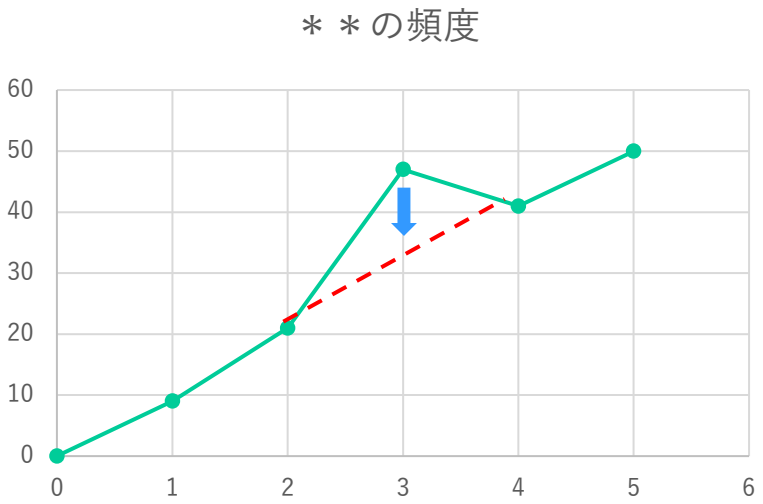
実例：STAP細胞事件（日本，2012年～2014年頃）

- ある若手生命科学研究者が起こした事件.
- iPS細胞より簡易な「初期化細胞」を作れる，とした.
- 査読（投稿論文を匿名の専門家が審査すること）で厳しい意見もついたが，結果的にNature誌に採択.
- しかしその後追試等でSTAP細胞は再現できず，調査の結果，捏造（や改竄・盗用）が発覚した.
- 結局，STAP細胞は既知の細胞の混入であることが分かった.論文は撤回された.

データ取り扱いの健全性２：改竄

実データを曲げて書き換えること．

例：実験データの一部を，つい曲げてしまいたくなる．



「このデータさえなければ
まっすぐなのに・・・」

実例：

- 鉄鋼素材品質検査データ改竄（日本，2017，製鉄会社）
- 免震装置データ改竄（日本，2018，産業装置製造会社）
- 電気泳動画像データ改竄（日本，2012，大学（生物系））

データ取り扱いの健全性３：盗用

データを不正に使い回すこと.

他人のデータの盗用は論外，自分のデータの「使い回し」も危険.

例：38本の論文で同一データ使い回し（日本，2010年，大学（医学系））
　　→「二重投稿（self-plagiarism）」に該当する

【参考】実際に起こる剽窃・盗用には，むしろ，
- アイデアの盗用
- 文章の剽窃　（自分の文章の自己剽窃を含む）

が多いです.
図表等を断りなく借用することも「剽窃」にあたります.
→ 適切な引用表示が必要です.

～余談～　個人のレベルでも

現状は「個人情報」「個人データ」にかかるリテラシーが低く，
情報暴露が盛んに起きているように思われます．

例：SNSに家族の画像を無断で載せるケース．
　　→例えば子の画像の場合，子が望んでいるかどうか不明．
　　（後述「忘れられる権利」が確立するまで消せない．）

例：SNSに友人の画像を無断で載せるケース
　　→食事・飲み会の画像など．日時・場所にかかる個人情報を暴露．

例：仮名でSNS活動している人にかかる個人情報を書き込むケース
　　→うっかり友人の本名・勤務先等を含めた投稿をする．

すべて，個人情報の暴露であり侵害です．気をつけましょう．

個人情報とプライバシー

個人情報とは：

- 個人ID（個人を特定できるデータ）
- その他の個人の情報・データ

プライバシーとは：

- 上記の個人情報・データを含み，それを「守ること」自体
- 個人情報の中で，特に「機微データ」「要配慮個人情報」とその保護

個人情報とは何か

- 個人ID（個人を特定できるデータ）
  - 名前，住所
  - マイナンバー，学籍番号，職員番号
  - 免許証番号，パスポート番号，など

- その他の個人の情報・データ
  - 関連属性：本籍，人種，年齢，家族構成，勤務先，所得など
  - 健康情報（病歴）
  - 成績・評価データ：試験，勤務評価など
  - 文化的側面：宗教，性的嗜好など
  - 信用履歴（クレジットカード）
  - メールアドレス
  - 購買履歴
  - 機械等の操作履歴（ウェブ閲覧をふくむ）
  - オンライン識別子（IPアドレス，端末識別子など），など

比較的新しい

注意：この「個人情報」が誰のものか，は実は曖昧になりえます．
→あとで改めて議論します．

プライバシーとは何か

- 個人情報・データそのもの，に加えて，
  それらを保護する行為，をもふくみます．

- また，個人情報・データのうち，
  　特に注意を要するデータ：「機微データ」「要配慮個人情報」
  を特に指すこともあります．

  前項の個人情報・データの中では，例えば以下のものです．
  - 人種
  - 健康情報（病歴）
  - 文化的側面：宗教，性的嗜好など

3-1-3. データサイエンス・AIで起こりうる論点

データサイエンス，あるいはAIを用いる際，次のような問題が起こりえます.

- 統計的差別
- データバイアス・アルゴリズムバイアス
- 個人情報の暴露，プライバシーの侵害

統計的差別

統計的処理が妥当であり，その処理結果を用いる人が（偏見なく）合理的に判断している場合でも，結果として差別・不平等が肯定され，継続されうること.

典型的な場合：
雇用主が，採用段階において，個々の応募者が属する属性グループごとの「統計的平均値（あるいは平均的描像）」に基づいて能力を推測し，採用の判断をする場合.

例：「女性」は「短期勤続」（長く勤めないこと）の傾向がある
→そうしなくてよい社会になっていないことが原因
例：「黒人」は「生産能力」が低い傾向がある
→歴史的に平等な機会が与えられてこなかったことが原因
例：「外国人」は「日本語能力」が低い傾向がある
→日本語教育支援，および業務の多言語化がないことが原因

どれも，これまでのデータを統計的に処理すれば（相関関係としては）正しくても，「現状」は既に行われた不公平・差別の結果でもあり，直ちにそれに基づいて判断することは差別の延長になります.

データバイアス

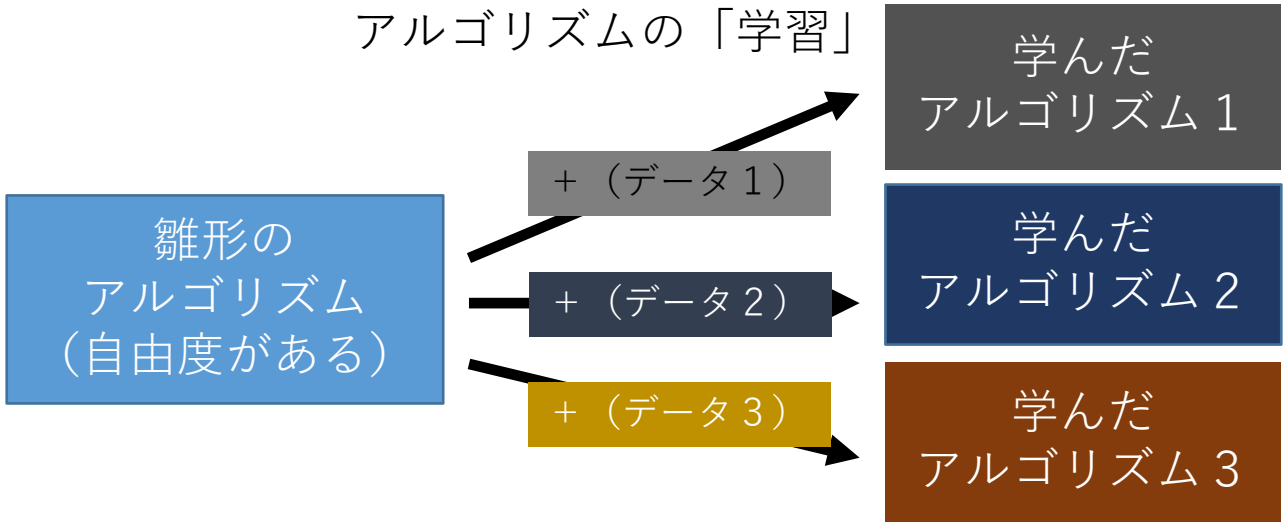
統計処理，あるいはデータサイエンス的処理を行う際扱うデータに，
そもそも　バイアス（bias)=偏り　があること，
あるいは，それに起因して起こる（よくない）こと．

例：　ある病気の治療方法について研究するとき，
特定の病院の患者のデータだけを集める
→　患者の地域性，富裕層・貧困層の偏り

例：　国内世論を推し量るのに，twitter 上の呟きから抽出する
→　そもそも twitter をやっている人の層，
その中でも頻繁に呟く人の層，の偏り

アルゴリズムバイアス

機械学習により，アルゴリズムにデータから学習させたとき，
データにバイアス（偏り）があったがゆえに，
学習結果のアルゴリズムにもバイアスが生じてしまうこと．



アルゴリズムは
「悪気」なく
与えられたデータを
素直に学ぶ

データ・アルゴリズムバイアスの実例

AmazonのAI人事アルゴリズム（米，2014年頃）

- Amazonは2014年頃から，新規採用に関して，AIによる自動書類審査の導入を始めた．大量の応募を迅速に処理可能．
- このAIアルゴリズムは，過去10年間の雇用パターンで「学習」したもので，アルゴリズムで応募書類をランク付けする．
- だが2015年頃，すぐに，Amazonはこのアルゴリズムが「性別に対して中立的」でないことに気がついた．
  「過去10年間の雇用パターン」は男性雇用に偏っていて，アルゴリズムはそれを素直に学習してしまっていた．
- アルゴリズムは“women’s” など，「女性」に関するキーワードにペナルティを科し，一方で男性が用いがちな表現に高い点を与えていた
- Amazonは，結局，このプロジェクトは停止した．

3-1-4. 社会的合意の形成に向けて

この節では，以下について見ていきます.

- データサイエンス時代の諸概念
  - 忘れられる権利
  - 説明に基づく同意
  - オプトイン・オプトアウト
- 形成されつつある合意の例
  - GDPR（欧州一般データ保護規則）
  - 人間中心のAI社会原則
- データサイエンスやAIの責任は誰が負うのか

諸概念１：忘れられる権利

インターネット時代のデジタルデータに関して，削除・アクセス遮断によりプライバシーが保護されることを求める権利.

[考えられるようになったきっかけ]
- デジタルデータがネット上に出回ると，消す手段がない
- 検索エンジンに登録された情報も，消す手段がない（たとえ犯罪歴のような情報であったとしても）
- 従って，通常のアナログ情報と異なり，永遠に「忘れられる」ことがない

[現状]
- EUなど一部の国では法制度化済みです.
- 日本では「インターネット法」に関係して議論が始まり，検討中.
- 「プライバシー保護」と「表現の自由・知る権利」の両立がなかなか難しいのも議論が長引く一因です.
- 法整備だけでは不十分で，技術的に消せるようになる必要があります.

諸概念２：説明に基づく同意

個人情報・データの提供を求める際は，「説明に基づく同意（informed consent)」に基づいて，提供を求める必要があります．

- どのようなデータを提供してもらうのか
- 何に使うか，いつまで使うか
- 誰と共有するか
- データを提供することのメリット・デメリットは何か

（参考）インターネット，あるいはデータサイエンスの黎明期は，様々なデータが「勝手に集まって」いましたが，いまはそのような考え方は許されなくなっています．
例：ウェブの閲覧履歴

（参考）「説明に基づく同意」に基づいて集めたデータを，当初宣言した目的以外の目的に使用する場合は，同意をとり直す必要があります．

注意：研究に使う画像・データを友人，研究室仲間からもらう場合も同様です．必ず「説明に基づく同意」書をとります（→研究倫理審査）．

諸概念３：オプトイン・オプトアウト

何らかのサービス・手続き等に「参加することを希望する」，
あるいは反対に「参加しないことを希望する」ことを表明してもらう手続き.

- 「オプトイン」(opt in)　（英単語の“opt”＝「～を望む」）
  参加することを希望することを言います.
- 「オプトアウト」（opt out)
  参加しないことを希望する（すでに参加している状態から抜ける，あるいは黙っていると参加してしまうことを拒否する）ことを言います.

例：JR東日本による，SUICA利用データの販売

- SUICAサービスに加入した最初の状態では，各利用者のSUICA利用データが，「匿名化（→後述）」を施した状態で，外部企業に販売されます.
- これを希望しない利用者は，JR東日本のホームページから「オプトアウト」手続きをとると，販売データから取り除いてもらえます.

形成中の合意１：GDPR

“EU General Data Protection Regulation (GDPR)”
＝「欧州一般データ保護規則」

- EU内28カ国でバラバラであった個人情報保護関連規則を一元化したものです.
- それと同時に，ここまで述べてきたような最先端のデータ保護の考え方を具現化したものでもあり，最先端の考え方を表したものでもあります.
- 本教材と特に関係するところ：
  - 個人情報・データは誰のものか？
    第3-1-2節で挙げたものはすべて「データ主体」（＝そのデータを発生させた自然人）のもの，とみなされます.
  - 説明に基づく同意は，目的・期間を必要な範囲に限り，一般人に分かりやすく，いつでも同意を撤回できることが求められます.
  - 手続きは「オプトアウト」ではなく，「オプトイン」で設計することが求められます.
  - AI等により機械だけで不利な判定が出る場合は，異議申し立ての権利を確保することが求められます.

形成中の合意２：人間中心のAI社会原則

内閣府 統合イノベーション戦略推進会議 決定（2019年3月29日）

- 人間中心の原則
- プライバシー確保の原則
- セキュリティ確保の原則
- 公平性・説明責任・透明性の原則，など

いずれも，本教材で議論している内容を，国レベルで方針づけたもの.

類似のものは，他にも多数提案されています：

- 人工知能学会
  倫理指針（2017年2月28日）
  機械学習と公平性に関する声明（2019年12月10日）
- 人間中心の機械学習（https://www.fatml.org/）
- 総務省 AIネットワーク社会推進会議 報告書（各年）
- IEEE Ethically Aligned Design, first edition (2019年3月)，など

データサイエンスやAIの責任は誰が負うのか

データサイエンスやAIは，万能でも完全でもありません.

- 思ったよりできないこと，反対に，人間が想像する以上にできてしまうことがある（→匿名性が剥がれ機微データが流出する事故）
- 深層学習等，いわゆる「ブラックボックス」で，科学的に明解な説明ができていない技術もある.

責任者が分かりにくい：開発者，販売者，購入者，….
どのひとりも完全なる責任者にはなりにくいし，完全なる無関係者とも言えません.

それでは，技術の提供者は，何をどこまで行うべきでしょうか？

- Accountability（アカウンタビリティ）：技術や商品そのものに加え，誰が責任を負うのかまで説明できること.
（ただしこれは現状では難しいです．）
- Trust（トラスト）：（アカウンタビリティを必ずしも完全に果たせない場合に）過去の類似例を示して，現状の技術や商品の妥当性，公平性，正当性に納得してもらう方法.

データを守る上での留意事項

概要.

- この節では，前節までに述べた個人情報・データを中心に，
  それらをどのように守るかについて理解します.

- より具体的には，下記について学びます.
  - ✓ データの守り方
  - ✓ 悪意のある攻撃と既に起こった事例

3-2-1. データの守り方

守り方１：情報管理三原則

外部の脅威から守りながら，情報（データ）をうまく活用するための
３つの原則を「情報管理三原則（情報セキュリティ三原則）」と呼びます.

- 機密性
  情報（データ）に，正当な権限を持つ者だけがアクセスできること.
  このためには情報の適切な分類，アクセス権限の設定，情報の暗号化
  などが必要です.
- 完全性・整合性
  情報が完全であり，誤りのない状態であること.
  不正なアクセスの検知，誤り検出などの技術が必要です.
- 可用性
  （正当な権限を持った者が）情報に適切にアクセスできること.
  可用性を阻害する攻撃（DoS攻撃など）への防御,
  ハードウェア障害のための冗長保存性などが必要です.

データの守り方２：匿名化（1/2）

データから「個人ID」（→第3-1-2節）を適切に削除して，
データを扱いながらも当該個人の情報であることが分からないように処理することを「匿名化」と言います．

これまでよく行われてきた匿名化：

1. 連結可能匿名化：仮IDを発行して個人IDを隠す方法．

No.1, 身長170cm, 体重60kg；	No.1＝山田太郎
No.2, 身長160cm, 体重45kg；	No.2＝佐藤花子．．．
	↑「連結表」と呼ばれる

   データ処理は左の表（連結可能匿名化を施したデータの表）を用い，
   個人特定が必要なときに限り，別途厳重保存した「連結表」を使う．
2. 連結不可能匿名化：そもそも連結表を持たない方法．
   連結表が存在すると危ないほどの機微データの場合，
   連結表を持たない，あるいはさらに仮IDすら持たないデータのみを扱う．

しかしながら，これで「本当に個人特定は不可能なのか」には慎重な議論が必要で，上記の匿名化の考え方は，2015年の個人情報保護法改正で破棄されています．

データの守り方２：匿名化（2/2）

匿名化は本当に可能でしょうか？

- 匿名化が剥がれた例：AOL検索履歴公開案件
  - 2006年8月4日，AOL（米インターネットサービス）は，65万人の3ヶ月に渡る検索履歴（2,000万キーワード）を研究目的で公開した．
  - データには仮IDしか付けられていなかったが，検索データから，一部ユーザが特定され問題になった．
  - AOLは8月7日にデータを取り下げたが，すでにネット上に広まり手遅れであった．
  - 同9月，訴訟が起こされた．
- 社会学者 Sweeneyの研究（2000年）：
  ZIPコード（郵便番号）＋生年月日＋性別 で，アメリカ人の87%が数学的にひとりまで絞り込み可能

データは，組み合わせることで匿名性が剥がれ，機微データが漏れ出すことがあります．データサイエンスの進化に伴いこの危険性は上昇します．
データは，使う人も，提供する人も，これを前提にする必要があります．

データの守り方３：暗号化とパスワード

- 暗号化：
  元のデータに対して，特別な処理を施して，そのままでは読めない特殊なデータに変換すること.
  元に戻すことを「復号化」と言います.
  暗号化には様々な種類があり，「安全性（強度）」，「処理速度」が異なります.
  電話やインターネットの通信などは，原則として暗号化され，
  通信内容が保護されるようになっています.

- パスワード：
  データやサービスにアクセス権限を持つ人間であることを証明するための文字列.
  銀行のキャッシュカードやクレジットカードのPINコードを始め，
  日常で幅広く用いられていますが，短いパスワード，単純なパスワードは簡単に見破られるため注意が必要です.

3-2-2. 悪意のある攻撃と既に起こった事例

データは，暗号化やパスワードで原則として保護されますが，
それでも，過失や悪意を持った人の攻撃で，データが漏洩したり，
それによってプライバシーが侵害されることがあり，注意が必要です．

以下，いくつか類例を挙げます．

- データの持ち出し，あるいは紛失による流出
- 攻撃による情報漏洩
- スパイウェア，マルウェアによる情報搾取

データの持ち出し・紛失による流出

厳重に保管しているはずのデータでも，内部の人間による（悪意のある）持ち出し，あるいは正当な権限者が持ち出した際の過失による紛失で，データは漏洩しえます.

実例（持ち出し）：

- 市議会に立候補した市職員が，職務上の地位を利用して市民個人情報（住所等）を持ち出し，選挙応援依頼書を発送（平塚市，2020年）
- 県庁で使用していたパソコンのハードディスクの処分を依頼された企業の職員が，それを持ち出してオークションサイトで売却；ハードディスクには各種納税記録なども入っていた（神奈川県，2019年）

実例（紛失による流出）：

- 学校等において，成績データなどが入ったUSBメモリ，パソコンなどを紛失（複数案件あり）
- メール誤送信によるメールアドレス流出（複数案件あり）

（注意）メールアドレスも個人情報です.
本人に断りなく他人に暴露してはいけません.
（複数人に同時にメールを発送する際注意）

攻撃による流出

内部の人間は気をつけていても，外部からの（しばしばインターネット経由での）攻撃により，情報は漏洩しえます．

実例：

- 化粧品会社の決済サーバーの脆弱性（セキュリティ上の「穴」が開いていること）を攻撃され，購入者のクレジットカード情報が流出（日本，2020年3月；他，類似案件多数あり）
- QRコード決済システムのひとつで不正利用が発覚，多数のユーザの電子マネーが不正に使用された．後に，認証システムの設計が甘く，攻撃者が容易に他人のアカウントを乗っ取れることが明らかになった．そのQRコード決済システムはサービスを終了．（日本，2019年）

スパイウェア，マルウェアによる情報搾取

通常のアプリ，あるいはシステムの一部のように見せかけてパソコン等に取り込ませ，ユーザに分からないように内部情報・データを外部送信するプログラムを「スパイウェア」と呼びます．
「マルウェア」（“malicious”＝「攻撃的な」から来た名称）とも呼びますが，この場合，「ウィルス」（パソコン等に入り込んで動作に障害を起こすもの）なども含み，悪意あるソフトウェア全体も指します．

実例：

- Androidスマホに入り込み写真・動画・音声記録・連絡先等の情報にアクセスして外部送信しうる“Exodus”が発見され，その後iOS版も発見された（世界的，2019年）※現在は対処済み
- 仮想通貨取引所が攻撃され，ある仮想通貨580億円相当（当時レートのもの）が盗まれた．取引所内のパソコンがマルウェアに汚染され，情報搾取・外部からの悪意を持ったアクセスを許した，とされている（日本，2018年）

AIの責任

- 第４次産業革命の核心：データ駆動型社会
  - それまではデータの分析は自動化されていたが、分析結果に基づき行動する際の意思決定は人間が行っていた
  - 機械学習の発展によって、意思決定までもAIに委ねることができるようになってきた
  - データに内在する様々な偏りが意思決定に影響を及ぼす可能性
    → 「データがそうなっている」「アルゴリズムがそうさせた」という主張により、偏った意思決定が正当化されることに対する懸念が高まっている
- AIによる意思決定は人間の心理や生活に影響を及ぼしている
  - Webニュースのフィルタリング
  - 検索エンジンにおける検索結果のソート
  - ECサイトにおける書籍や商品の推薦
  - クレジットカードの不正利用判定
  - 生命保険の保険料設定
  - 医療における画像診断や疾病診断
  - 仮釈放の判断（米国ウィスコンシン州再犯予測プログラムCOMPAS）

AIの信頼性

機械学習モデルは信頼できるのか？

- 差別的なアルゴリズム
  - 英国内務省が、人種的なバイアスがあると批判されているアルゴリズムを用いたビザ申請の審査を中止すると発表（2020年８月）
  - 米国犯罪予測アルゴリズムCOMPAS「再犯の可能性が高いと判断したにもかかわらず、実際には再犯しなかった黒人は白人のおよそ2倍」

  「データがそうなっているから」という言い訳は許されない
- 説明可能AI（XAI： Explainable AI）：機械学習モデルの出力に加えて、その出力を補助する追加の情報（モデルの解釈、判断根拠の説明など）を出力する技術
  - ブラックボックスのままで使用するとGDPRやAI開発ガイドラインに抵触する可能性
  - XAIはアメリカの国防高等研究計画局（DARPA）が名付けた呼称

参考）　原 聡, “私のブックマーク「説明可能AI」（Explainable AI）”, 人工知能学会, vol.34, No. 4, 2019, https://www.ai-gakkai.or.jp/resource/my-bookmark/my-bookmark_vol34-no4/

説明可能AI (XAI： Explainable AI)の挑戦

1. 複雑なモデルを可読性の高いモデルで近似することで判断過程を解釈
   - 似た答えを出すモデルは似たように判断しているという発想
   - 可読性の高い決定木やルールモデルで近似
2. 説明可能なモデルの設計
   - 深層学習のような可読性の低いモデルではなく、初めから可読性の高い決定木やルールを使えばいいという発想
   - 訓練データの分析結果（典型的なデータや例外的なデータ）を提供する手法もある
3. 入力データのどの部分・処理過程のどの特徴が結果に影響を与えたかを推定
   - 画像分類の深層学習モデルであれば、結果を大きく変化させる画素領域を特定
   - 文書分類であれば影響を与えた単語に、画像分類であれば影響を与えた領域に色付けなどでハイライトすることで可視化
4. 説明の方法そのものを学習
   - ３の手法で「結果に影響を与えた領域をハイライトする」そのやり方自体を学習

AIの知的財産権

東京大学 数理・情報教育研究センター 山肩洋子

知的財産権

人間の幅広い知的創造活動の成果について、その創作者に一定期間の独占権を与える制度

- 知的財産
  - 発明、考案、植物の新品種、意匠、著作物その他の人間の創造的活動により生み出されるもの(発見又は解明がされた自然の法則又は現象であって、産業上の利用可能性があるものを含む。)
  - 商標、商号その他事業活動に用いられる商品又は役務を表示するもの
  - 営業秘密その他の事業活動に有用な技術上又は営業上の情報
- 知的財産権
  - 特許権、実用新案権、育成者権、意匠権、著作権、商標権その他の知的財産に関して法令により定められた権利
  - 法律上保護される利益に係る権利
- 「もの」ではなく「情報」
  - 容易に模倣され、消費されることがなく、多くの人が同時に利用できる
  - 社会が必要とする限度で利用の自由を制限することで創作者の権利を保護

参考： 特許庁「制度・手続き」（https://www.jpo.go.jp/system/index.html）

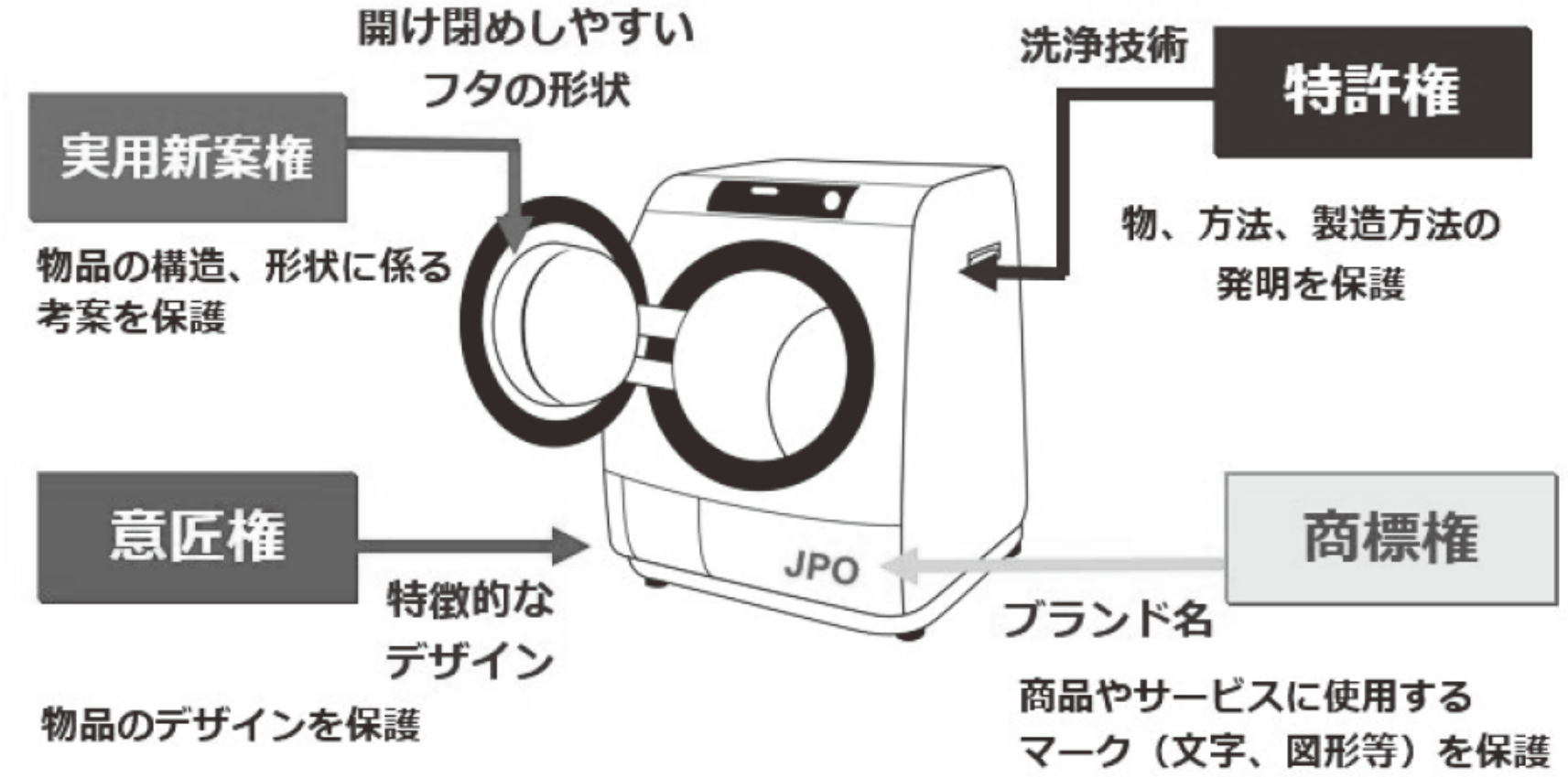
知的財産の種類

「産業財産権」と呼ばれ、特許庁が管理

- 特許権：発明を保護
- 実用新案権：物品の形状等の考案を保護
- 意匠権：物品、建築物、画像のデザインを保護
- 商標権：商品・サービスに使用するマークを保護
- 著作権：文芸、学術、美術、音楽、プログラム等の精神的作品を保護
- その他
  - 回路配置利用権：半導体集積回路の回路配置の利用を保護
  - 育成者権：植物の新品種を保護
  - 営業秘密：ノウハウや顧客リストの盗用など不正競争行為を規制
  - 商号、商品等表示、地理的表示

参照： 2019年度説明会テキスト「知的財産権制度入門」, 特許庁
https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_syosinsya/all.pdf

産業財産権



出典： 2019年度説明会テキスト「知的財産権制度入門」, 特許庁
https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_syosinsya/all.pdf

特許法上の「発明」とは

- 自然法則を利用している
  - 経済法則など、自然法則以外の法則 → No
  - ゲームのルールなど人為的取り決め→ No
  - エネルギー保存の法則、万有引力の法則など、自然法則自体→ No
- 技術的思想である
  - 「フォークボールの投げ方」などの技能→ No
  - デジカメで撮影された画像データなど、単なる情報の提示→ No
  - 絵画、彫刻などの美術的創作物→ No
  - プログラム言語又はプログラムそのもの→ No
- 創作である
  - 天然物から人為的に分離した化学物質→ Yes
  - 天然物の単なる発見→ No
  - 未完成発明→ No
- 高度である

参照： 2019年度説明会テキスト「知的財産権制度入門」, 特許庁
https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_syosinsya/all.pdf

AI分野における日本の審査基準

- 発明該当性：AI関連発明はコンピュータソフトウェアを利用
  →次の2ステップで判断
  1. 全体として自然法則を利用しており、「自然法則を利用した技術的思想の創作」と認められるか？
  2. ソフトウェアの観点に基づく考え方により、「自然法則を利用した技術的思想の創作」と認められるか？
- 進歩性：当業者が先行技術に基づいて、請求項に係る発明を容易に想到できないものであること
- 記載要件：
  - 実施可能要件：発明の詳細な説明に、請求項に係る発明を実施できるように発明が記載されていること
  - サポート要件：
    請求項に係る発明が、発明の詳細な説明に記載されたものであること

参照： 2019年度説明会テキスト「知的財産権制度入門」, 特許庁
https://www.jpo.go.jp/news/shinchaku/event/seminer/text/document/2019_syosinsya/all.pdf

AI関連技術等に関する事例集

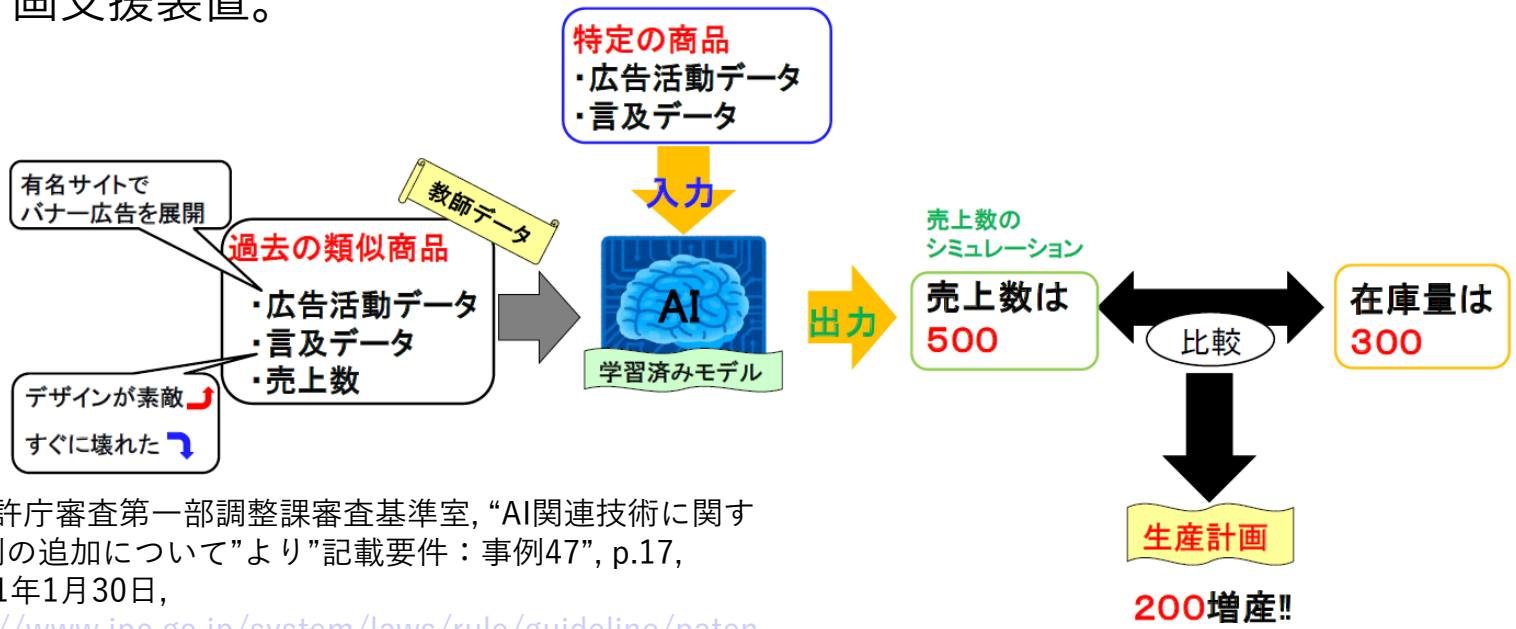
AI関連技術が様々な技術分野に発展していることに伴い、特許庁からAI関連技術に関する事例が公表されている。

特許庁調整課審査基準室, “AI関連技術に関する特許審査事例について”, https://www.jpo.go.jp/system/laws/rule/guideline/patent/ai_jirei.html

- 記載要件及び進歩性についての判断のポイントを、分かりやすく示すことを目的とする
- 各事例は記載要件または進歩性に着目した事例であり、着目した要件以外の拒絶理由に関して例示するものではないことに留意が必要

記載要件を満たす事例１：事業計画支援装置

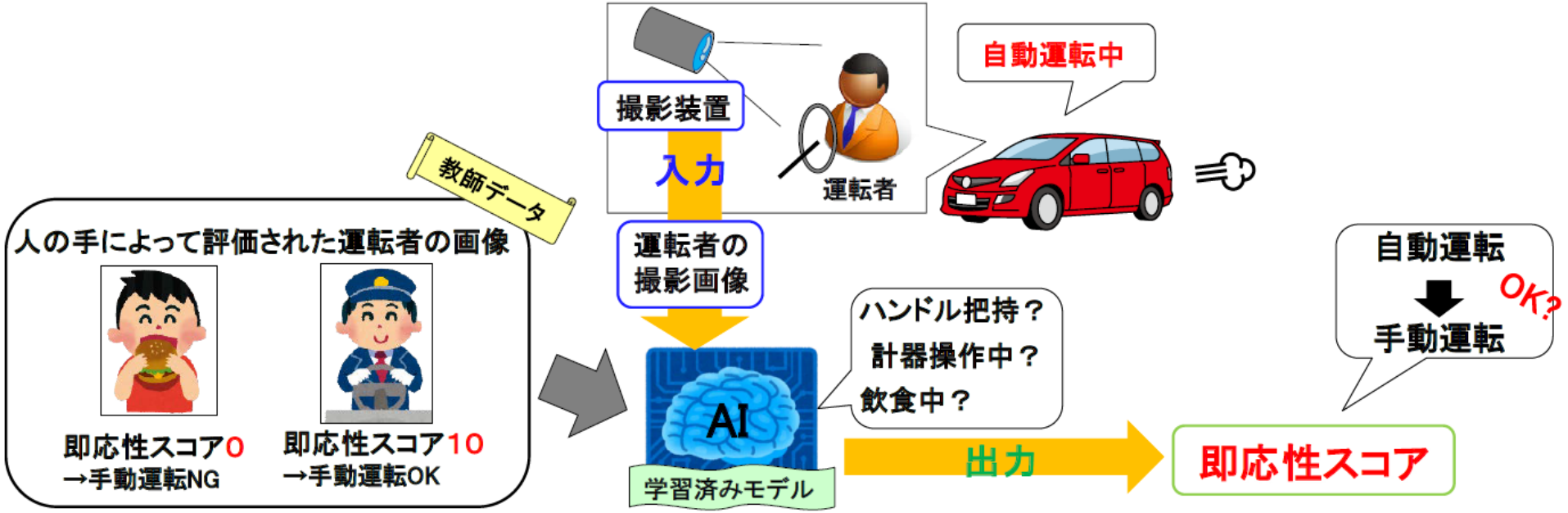
特定の商品の在庫量を記憶する手段と、前記特定の商品のウェブ上での広告活動データ及び言及データを受け付ける手段と、過去に販売された類似商品に関するウェブ上での広告活動データ及び言及データと、前記類似商品の売上数とを教師データとして機械学習された予測モデルを用いて、前記特定の商品の広告活動データ及び言及データから予測される今後の前記特定の商品の売上数をシミュレーションして出力する手段と、前記記憶された在庫量及び前記出力された売上数に基づいて、前記特定の商品の今後の生産量を含む生産計画を策定する手段と、前記出力された売上数と、前記策定した生産計画を出力する手段と、を備える事業計画支援装置。



ref.特許庁審査第一部調整課審査基準室, “AI関連技術に関する事例の追加について”より”記載要件：事例47”, p.17, 平成31年1月30日, https://www.jpo.go.jp/system/laws/rule/guideline/patent/document/ai_jirei/jirei_tsuika.pdf

記載要件を満たす事例２：自動運転車両

運転者監視装置を備える自動運転車両であって、前記運転者監視装置は、車両の運転席に着いた運転者を撮影可能に配置された撮影装置から撮影画像を取得する画像取得部と、前記運転者の運転に対する即応性の程度を推定するための機械学習を行った学習済みの学習モデルに前記撮影画像を入力することで、前記運転者の運転に対する即応性の程度を示す即応性スコアを当該学習モデルから取得する即応性推定部と、を備え、取得した即応性スコアが所定の条件を満たさない場合に、自動的に運転操作を行う自動運転モードから運転者の手動により運転操作を行う手動運転モードへの切り替えを禁止する自動運転車両。

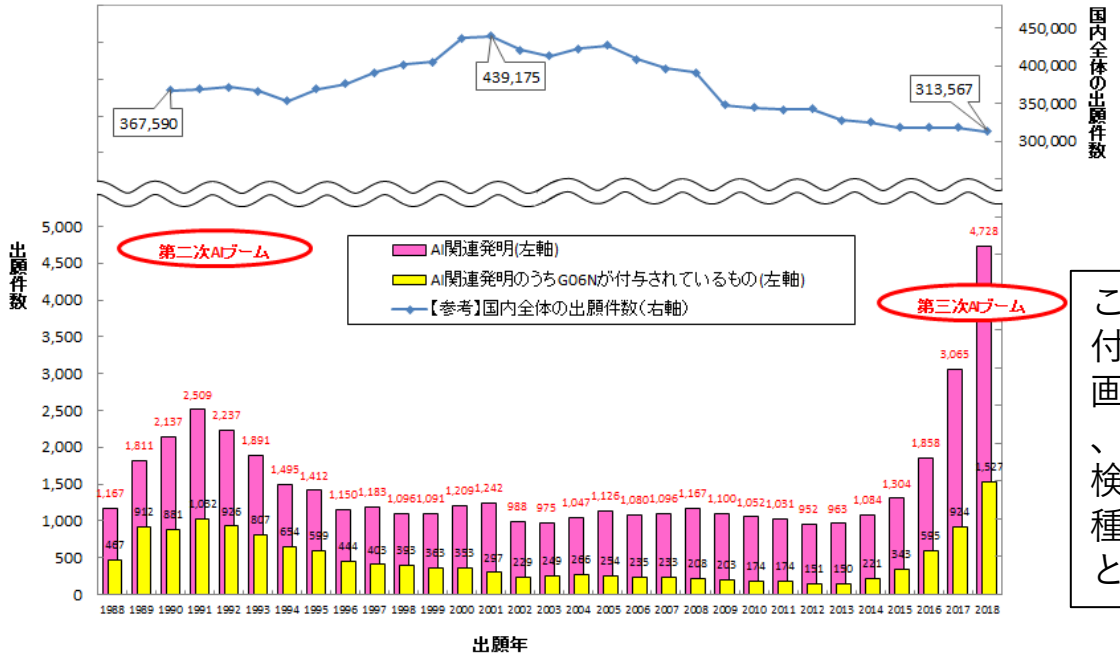


ref.特許庁審査第一部調整課審査基準室, “AI関連技術に関する事例の追加について”より”記載要件：事例48”, p.19, 平成31年1月30日, https://www.jpo.go.jp/system/laws/rule/guideline/patent/document/ai_jirei/jirei_tsuika.pdf

東京大学 数理・情報教育研究センター 山肩洋子

AI関連発明の国内出願件数の推移

- AI関連発明は2014年以降急激に増加しており、2018年は約4,700件（うち、G06Nが付与されているものは約1,500件）
- AI関連発明はG06Nが付与された出願に連動して増減が推移



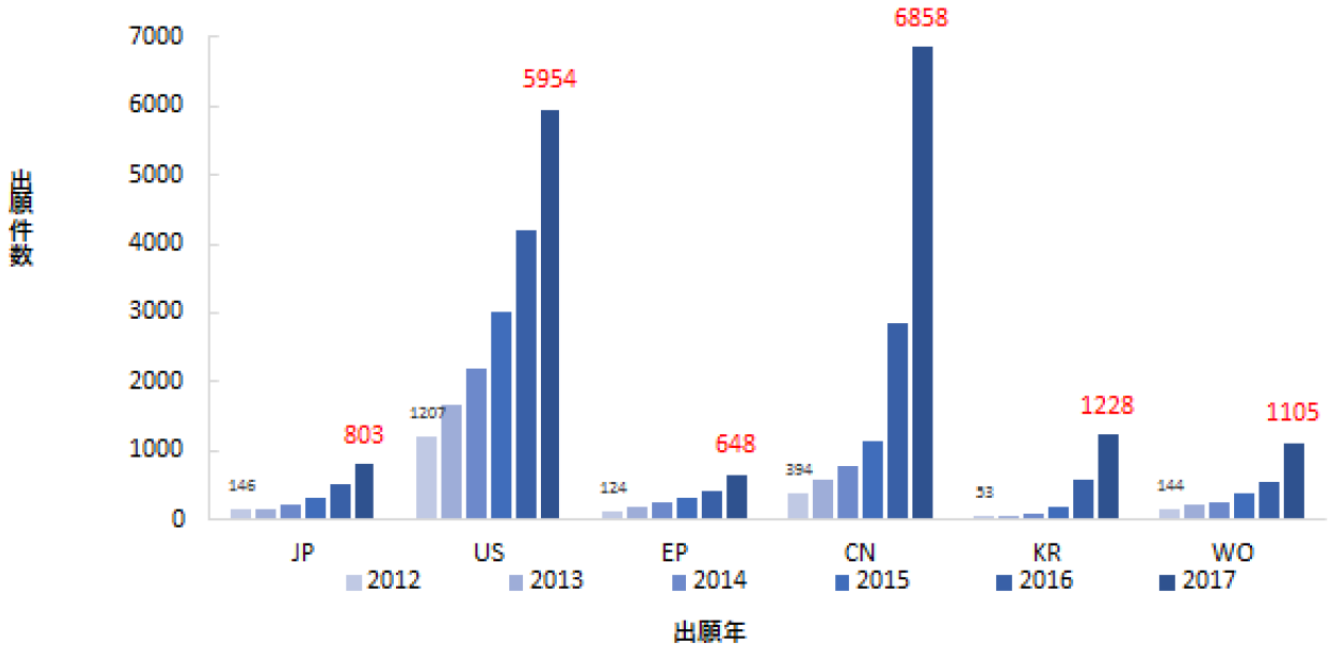
ここでAI関連発明とは、主にG06Nが付与されているAIコア発明に加え、画像処理、音声処理、自然言語処理、機器制御・ロボティクス、診断・検知・予測・最適化システム等の各種技術に、AIコア発明を適用したことに特徴を有する発明を指す。

国際特許分類G06N: ニューラルネットワーク、深層学習、サポートベクタマシン、強化学習等を含む各種機械学習技術のほか、知識ベースモデルやファジィ論理など、AIの基礎となる数学的又は統計的な情報処理技術に特徴を有する発明に付与される分類記号

ref.特許庁 審査第四部 審査調査室, “AI 関連発明の出願状況調査 報告書”, 2020年7月, https://www.jpo.go.jp/system/patent/gaiyo/sesaku/ai/ai_shutsugan_chosa.html

G06Nが付与されている各国出願件数の推移

各国にてAI関連技術の出願が増加傾向（米国と中国が突出）



JP:日本、US：米国、EP：欧州（EPO）、CN：中国、KR：韓国、WO：PCT国際出願

※PCT（Patent Cooperation Treaty: 特許協力条約）制度

ひとつの出願願書を条約に従って提出することによって、すべてのPCT加盟国に同時に出願したことと同じ効果を与える出願制度。WIPO（World Intellectual Property Organization：世界知的所有権機関）が事務局となっており、「WO」で始まる番号が付与される。

ref.特許庁 審査第四部 審査調査室, “AI 関連発明の出願状況調査 報告書”, 2020年7月, https://www.jpo.go.jp/system/patent/gaiyo/sesaku/ai/ai_shutsugan_chosa.html

オープンソース

- 学術界ではソースコードやデータセットの公開が推奨されている
  - 論文と同時に Github [1]などを通じてソースコードが公開
  - 誰もが再実験により検証可能となり、捏造を防げるのと同時に、そのソースやデータを使い技術を発展させることができる
  - GoogleやFacebook、Microsoftなど多くの企業が貢献
- ソースコードには通常、ライセンスが付与されている
  - 代表的なライセンス： MIT License
    - 許可されること：商用利用、コードの変更、再配布、個人利用
    - 利用条件：ライセンスと著作権を表示すること
    - 制限：作者や著作権者はいかなる責任も負わず保証も行わない
  - その他のライセンス：Apache License 2.0、GNU General Public License v3.0、BSD 3-Clause License、他（詳しくは[2]を参照）
- コードやデータセットはライセンスに従い使用する
  - 商用利用や再配布を許可しない場合や、ライセンスを継承しなければならない場合（コピーレフト）などがあるので使用許諾を十分確認すること
  - 自らのコードを公開する際も適切なライセンスを付与する

[1] Github,ソフトウエア開発者向けのウェブプラットフォーム, https://github.com/
[2] オープンソースガイドライン, https://opensource.guide/ja/

東京大学 数理・情報教育研究センター 山肩洋子