

2-1 ビッグデータとデータエンジニアリング

東京大学 数理・情報教育研究センター
2021年5月26日

概要

- ICT（情報通信技術）の進展とビッグデータについて学びます
- ビッグデータとデータエンジニアリングを実現可能とした背景やビッグデータの活用事例など2章全体のイントロにあたる節です

本教材の目次

1. ビッグデータとは	4
2. データ量の増大	5
3. ビッグデータ利活用のための課題	6
4. ビッグデータとデータエンジニアリングを実現可能とした背景 ：ICT（情報通信技術）の進展	7
1. ビッグデータの活用事例：POSの活用	17
2. ビッグデータの活用事例：ECでの事例	18
3. ビッグデータの活用事例：推薦行為事例	19
4. ビッグデータの活用事例：人流測定	24
5. ビッグデータの活用事例：機械の稼働ログ	25
6. ビッグデータの活用事例：情報通信システムログ	26
7. ビッグデータの活用事例：活用データ別	27
8. ビッグデータの活用事例：産業分野別	30

ビッグデータとは

大量の(デジタル)データのこと。多くの情報がデジタル化され、それがインターネットを介して流通することが可能となって来ています。ソーシャルメディアデータであるソーシャル・ネットワーキング・サービス(SNS)でやりとりされる大量のデータ自体、その利用履歴、インターネット上での購買履歴などは典型的ビッグデータですが、実社会での活動、例えば、コンビニでの購買行動はPOS (point of sale)などの形でビッグデータになっています。気象情報や地震波の情報などの自然現象、加速器による実験データもビッグデータになり得ます。

こうしたデータの中から必要な情報を探し出し、さらには、有効に活用し付加価値をつけることができるようになると期待されています。このような社会的な動きに関連する大量のデータという点を強調して、ここ10年ほど、ビッグデータと呼ばれることが多くなってきました[1]。

[1] <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>
(Access: 2021/2/24)

データ量の増大

我々は多くのデータに囲まれています。[1]によれば、毎分、450万 YouTubeビデオが視聴され、1億8800万のメール送信がなされ、39万のアプリがダウンロードがされ、instagramでは約28万のストーリー、5.5万の写真がポストされています(2019年)。

こうした行為がビッグデータの生成を加速しています。

[1]<https://web-assets.domo.com/blog/wp-content/uploads/2019/07/data-never-sleeps-7-896kb.jpg>
Accessed 2021/2/26

ビッグデータ利活用のための課題

ビッグデータ利活用のための課題は以下の4点とされています。

- ① 膨大なデータの蓄積
- ② 膨大なデータの効率的処理
- ③ 膨大なデータから高付加価値な生成物の作成
- ④ セキュリティ・プライバシー保護

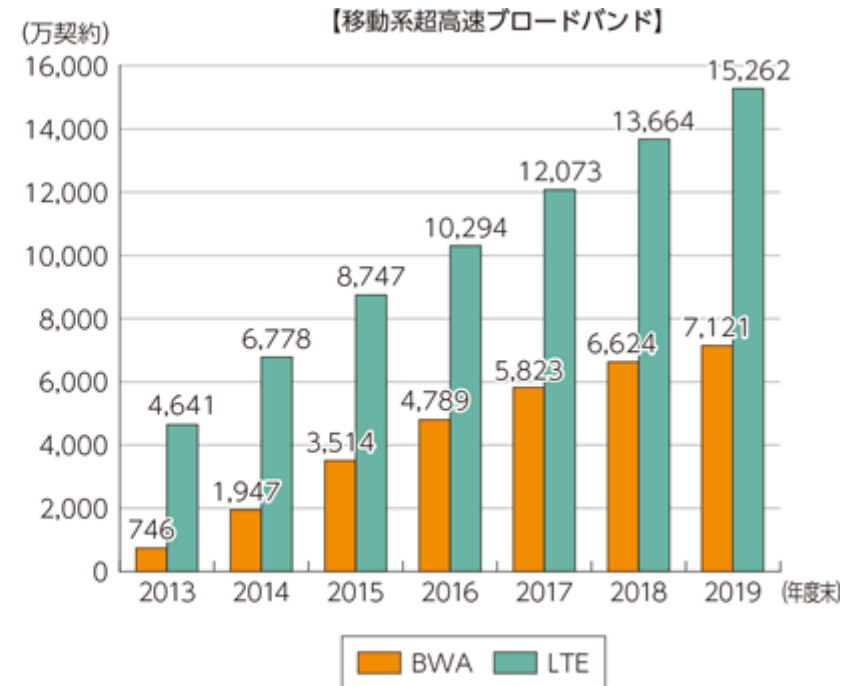
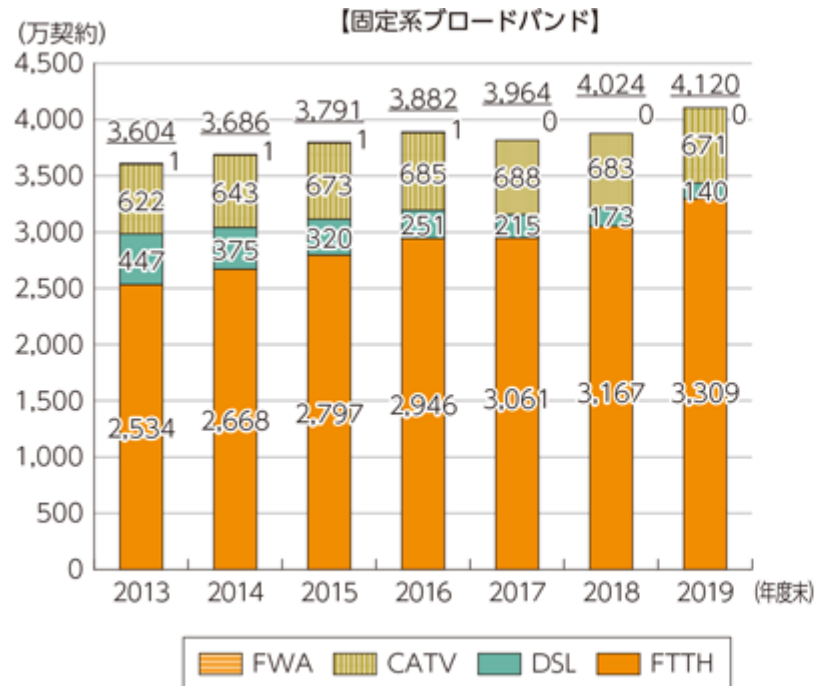
以降のページで、①、②を克服しビッグデータの利活用を実現し得る状況になりつつあることを勉強します。また、その後のページでは③に関わる具体的事例を述べます。④については、2-6で勉強してください。

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

ビッグデータの活用の機運が盛り上がった理由の1つがICT技術の進展であることは間違いありません。特に、通信技術の発展によりビッグデータ収集能力が飛躍的に向上しました。また、SNSなどの形でデータそれ自体がネットワーク上に出現するようになりました。データ収集に関わる具体的な話は2-3で勉強するとして、ここでは、高速な通信ネットワークと携帯電話ネットワークを中心とする無線ネットワークがいかに普及したか、を見てみましょう。

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

固定、移動ともに高速化が急速に進展



※過去の数値については、事業者報告の修正があったため、昨年の公表値とは異なる。

FWA: fixed wireless access
CATV: cable television
DSL: digital subscriber line
FTTH: fiber to the home

BWA: broadband wireless access
LTE: long term evolution

出典：総務省ホームページ

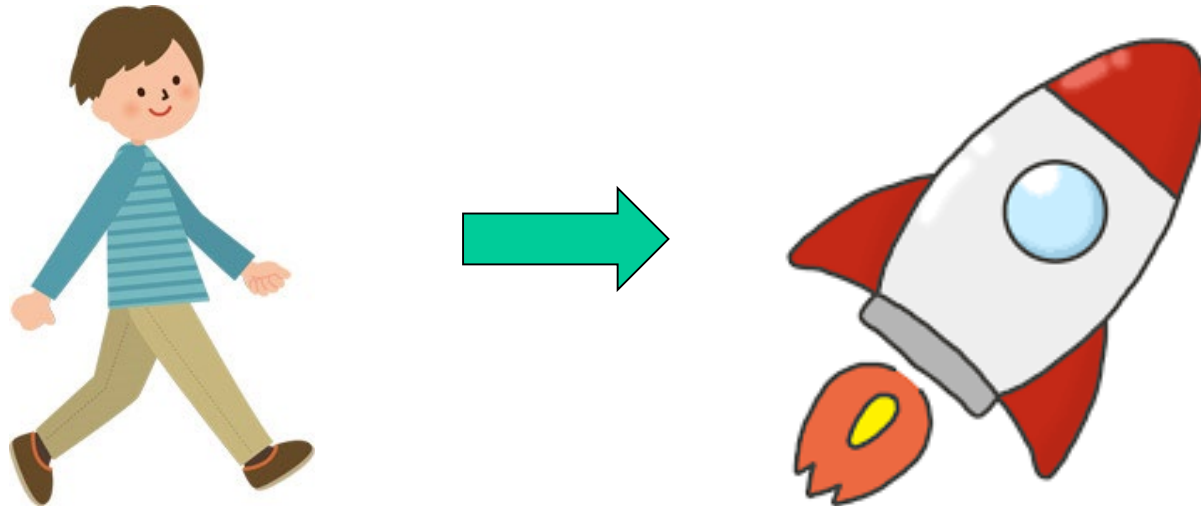
図表5-2-2-2 ブロードバンド契約数の推移

<https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/html/nd252210.html>

Access: 2021/2/27

ビッグデータとデータエンジニアリングを実現 可能とした背景：ICT（情報通信技術）の進展

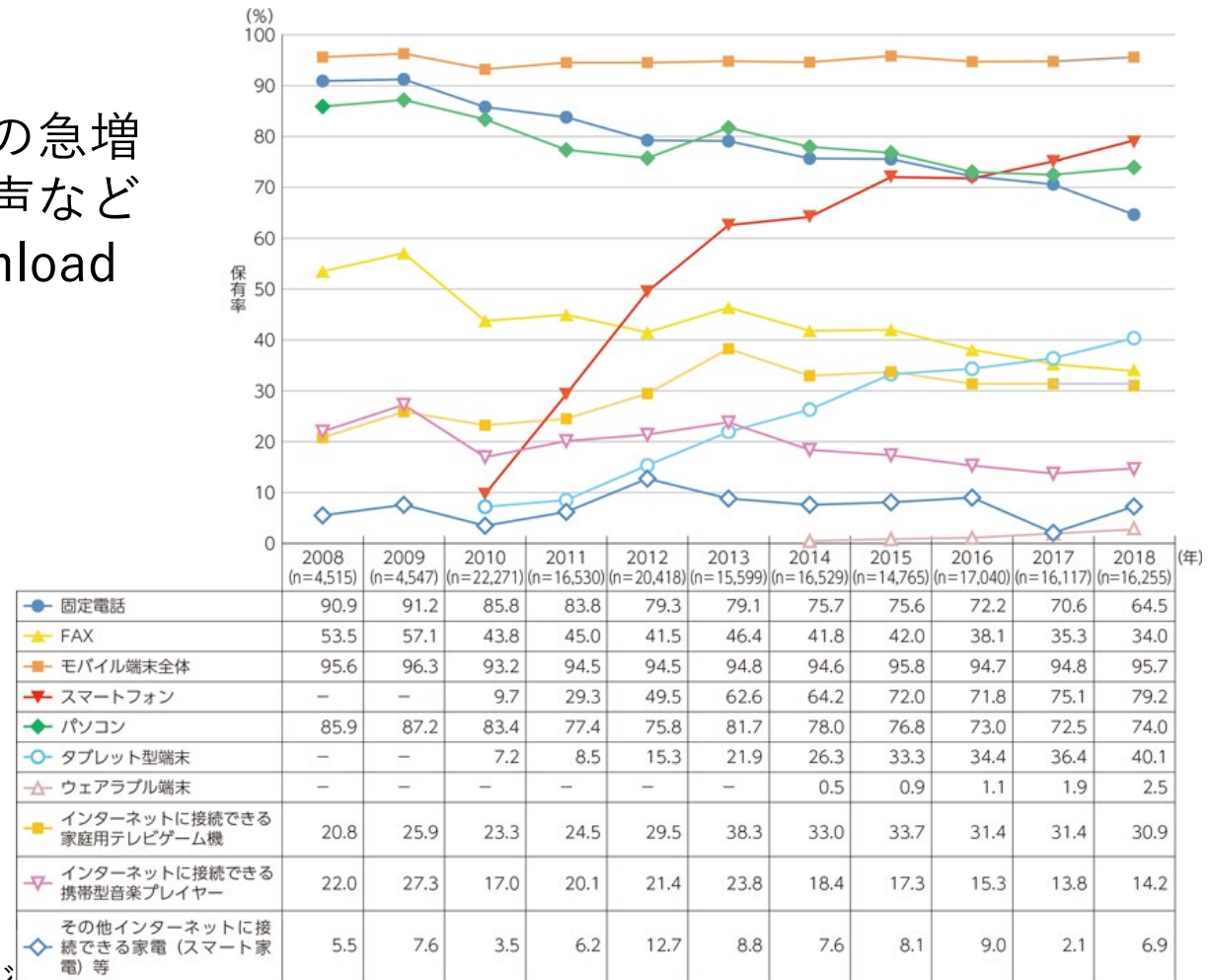
あるインターネット接続サービス事業者の例[1]:
約20年で7800倍の高速化。
これは徒歩とロケットほどの速度差ということです。



[1] <https://www.ocn.ne.jp/ocn20th/history/column1/>
Access: 2021/2/27

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

- スマホの保有率の急増
- 画像・動画・音声などのupload、downloadの容易化



出典：総務省ホームページ

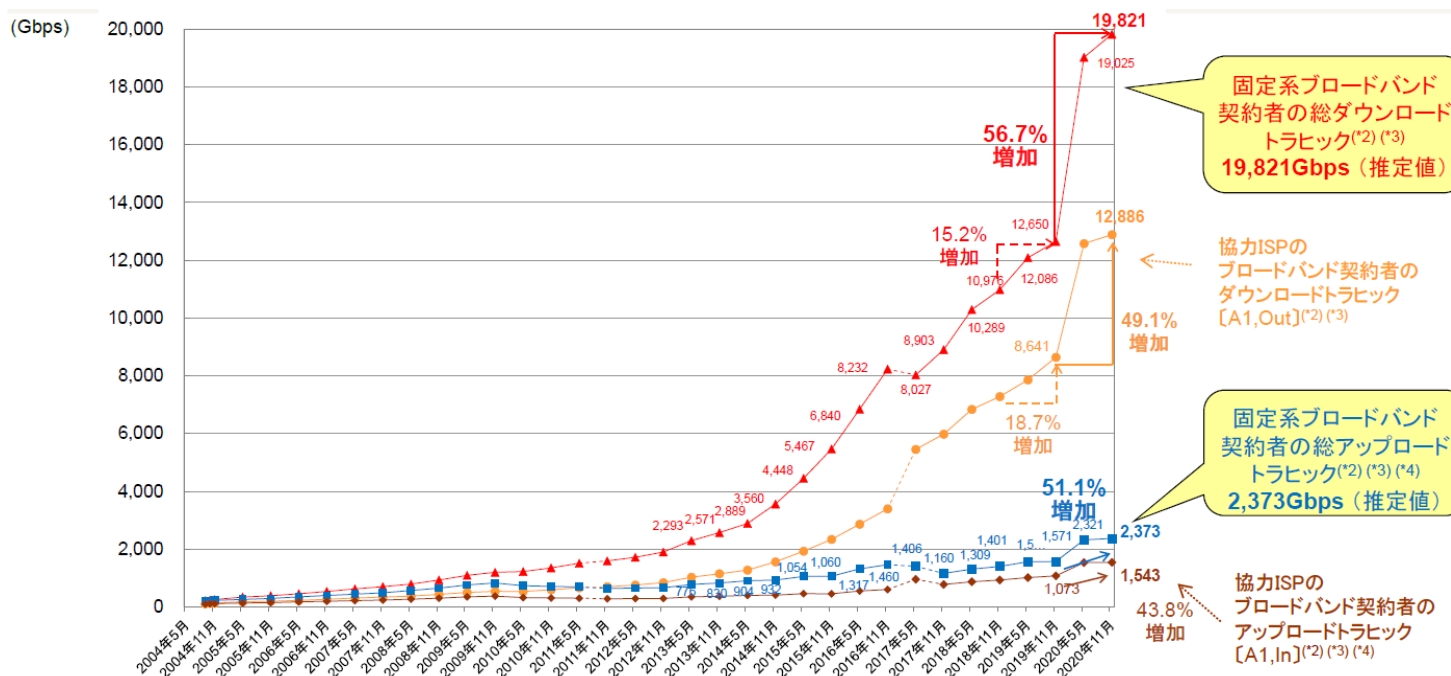
図表3-2-1-1 情報通信機器の世帯保有率の推移

<https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/image/n3201010.png>

Access: 2021/2/27

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

急増が続く固定系ブロードバンド契約者のトラフィック(通信量)



(*)1 個人向けサービス(FTTH, DSL, CATV, FWA) (ただし、一部法人を含む)

(*)2 2011年5月以前は、一部の協力ISPとブロードバンドサービス契約者との間のトラフィックに携帯電話網との間の移动通信トラフィックの一部が含まれていたが、当該トラフィックを区別することが可能となったため、2011年11月より当該トラフィックを除く形でトラフィックの集計・試算を行うこととした。

(*)3 2017年5月より協力ISPが5社から9社に増加し、9社からの情報による集計値及び推定値としたため、不連続が生じている。

(*)4 2017年5月から11月までの期間に、協力事業者の一部において計測方法を見直したため、不連続が生じている。

出典：総務省ホームページ

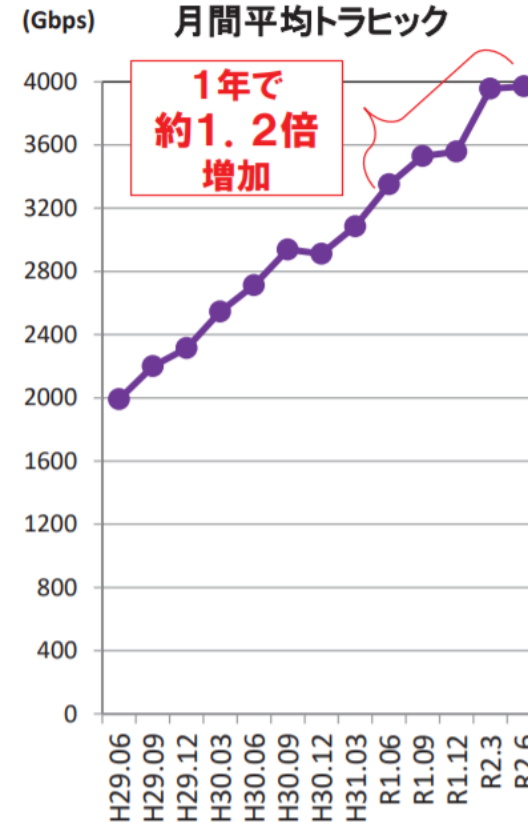
我が国の固定系ブロードバンド契約者の総トラフィック

https://www.soumu.go.jp/main_content/000731585.pdf

Access: 2021/2/27

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

移動体通信トラヒックも大幅に増加



出典：総務省ホームページ

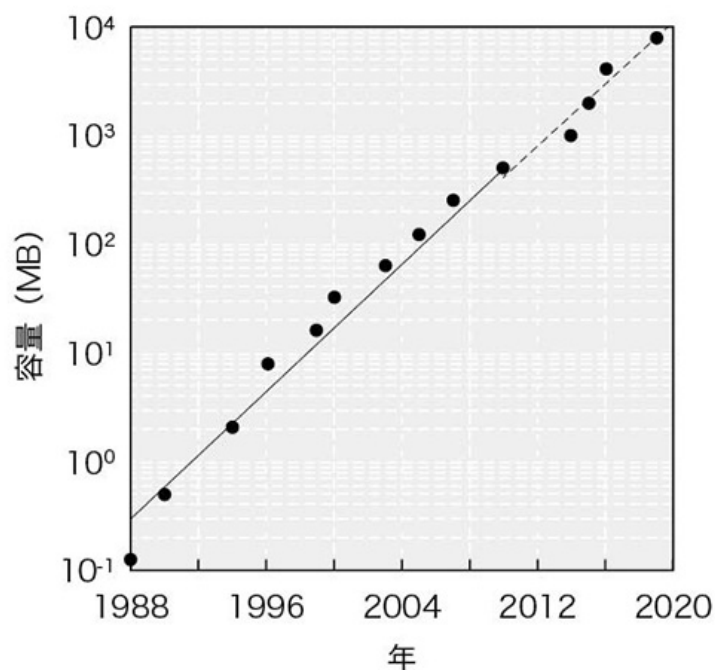
移動体通信トラヒックの推移

<https://www.soumu.go.jp/johotsusintokei/field/data/gt010602.pdf>

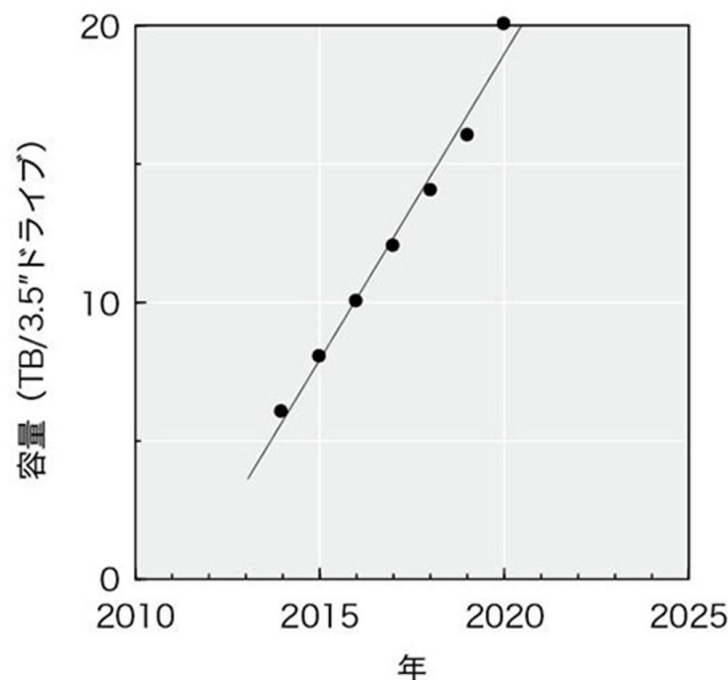
Access: 2021/2/27

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

ビッグデータ利活用のための課題の1つである膨大なデータの蓄積をICT技術が解決しつつある証左として蓄積容量の拡大のデータを示します。また、これらを用いたデータベース技術については、2-4で勉強します。



5 図 DRAMの容量変化のトレンド[1]



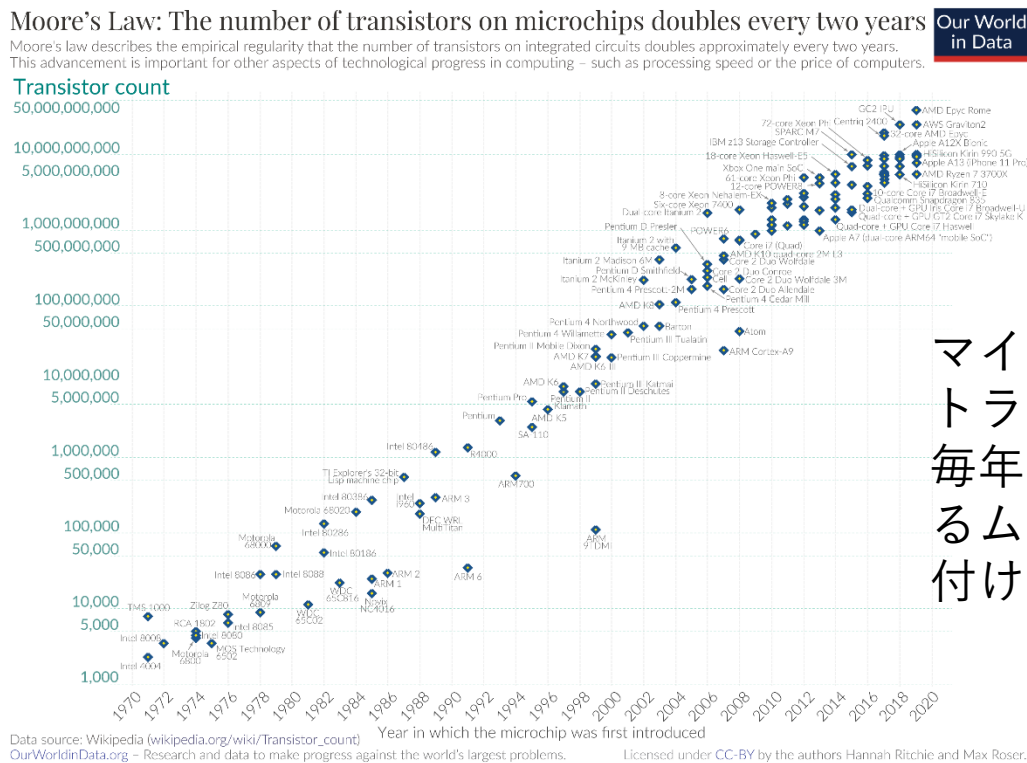
8 図 HDDの容量変化のトレンド[1]

出典 NHK [1] <https://www.nhk.or.jp/strl/publica/rd/181/2.html>

Access: 2021/3/26

ビッグデータとデータエンジニアリングを実現 可能とした背景：ICT（情報通信技術）の進展

ビッグデータ利活用のための課題のもう1つは、膨大なデータの効率的処理です。プロセッサ速度の向上に関するデータから、この課題を克服しつつあるということがうかがえます。



マイクロチップ上の
トランジスタ数は
毎年2倍になるとす
るムーアの法則を裏
付けるグラフ

<https://ourworldindata.org/technological-progress>
Access: 2021/2/26

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

さらに、以下のような技術シーズがビッグデータの活用を後押ししています。

1 つは、GPU(graphics processing unit)です。GPUとは、画像処理用のプロセッサのことで、ゲーム機などに搭載されてきました。その後、それがCADや科学計算にも使われるようになりました。これをGPGPU(General-purpose computing on graphics processing units)と言います。データサイエンスやAIで使われるニューラルネットワークなどの機械学習アルゴリズムでもGPUの利用を想定して高速化を図っているものもあり、それらはGPUがないコンピュータでは動作しません。最近では、多くのコンピュータにGPUが搭載されています。

ビッグデータとデータエンジニアリングを実現可能とした背景：ICT（情報通信技術）の進展

さらに、クラウド（コンピュータクラウド）と呼ばれる技術もビッグデータの活用を後押ししています。

クラウド（コンピュータクラウド）とは、コンピュータ群をネットワークの先のネットワークセンターに用意し、ユーザは、そのコンピュータ群（の一部）を必要に応じて利用するような環境のことです。GPUを具備するようなコンピュータなどを専用に用意して、データサイエンスやAIを行うことは、初期投資などの点から、ハードルが高いですが、クラウドの形態であれば、容易に実現できます。これも、ビッグデータを用いて機械学習が身近になって行われる理由の1つです。国内のIT企業やGoogleやAmazon, Microsoftなどの海外の有力企業がクラウドサービスを提供しています。

ビッグデータの活用事例：POSの活用

お店に行って商品を購入した際の記録は、POS (point of sale) などと呼ばれ、マーケティングにおいて重要なデータとなります。例えば、どの商品とどの商品は同時に購入される可能性が高いかが分かれば、それらを近くに並べる商品陳列にすることにつながります。どの曜日にどんな商品が売れるかが分かれば、曜日に応じた仕入れの量を変えることができます。天候やイベント情報と組み合わせて分析できれば、それらに対応した仕入れ量の決定も可能です。[1]にある外食産業の事例では、POSデータからキャンペーン効果を分析することにより、広告費を削減しながら売り上げの増加を達成しました。これらは実市場でのPOSですが、インターネット上での購買行動はより詳細なデータの取得が可能となります。それらを用いることが、より増加すると想定されます。

[1]<https://liskul.com/one-to-onemarketing-13202>

ビッグデータの活用事例：ECでの事例

インターネット上での行動は、一人一人の行動ログ（記録）という形で記録に残ります。特に、自社サイト利用者のCookieなどの情報や電子商取引市場(EC)での購買行動履歴を利用できる立場にある事業者は、それを用いて顧客ごとのマーケティングを実行できるようになりました。これを従来のマスマーケティングとの対照でOne-to-oneマーケティングと呼ぶ人もいます。具体的には以下のようなことが可能になりました。

●推薦行為

いわゆる「おすすめ商品」の提示です。

●関連情報提供行為

関心がありそうな行動をするユーザに、興味を持ちそうな、購買につながりそうな情報を再提示する、あるいは、新たに提示することです。

ビッグデータの活用事例：推薦行為事例

ビッグデータ活用の成功例としてもよく取り上げられる推薦行為（レコメンド、リコメンドなどと呼ばれることが多い）について具体的に見てみましょう。用いるデータの種類から、2つに大別されます。

1) 多数のユーザの過去の購買履歴による方法

ユーザ嗜好を多数のユーザの過去の購買履歴に基づきモデル化するものです。ソーシャルフィルタリング、協調フィルタリングなどとも呼ばれています。似た購買履歴ユーザの購入商品から推薦を行えるため、年齢や性別などのユーザ属性や種別などの商品属性が不要な反面、新しい商品など購買履歴がないものに対する推薦ができません。属性に直接リンクしない要因も反映できます。

2) 属性ベースの方法

ユーザ属性や商品の属性を利用する方法。属性情報を得られない場合には適用できませんが、履歴データのない新商品などにも使えます。

ビッグデータの活用事例：推薦行為事例

推薦行為の具体例について見てみましょう。まずは、推薦行為の発祥ともいえるAmazon.comの事業例です。



ビッグデータの活用事例：推薦行為事例

推薦行為の具体例について見てみましょう。まずは、推薦行為の発祥ともいえるAmazon.comの事業例です。

Amazon.comでは、推薦システムが、単に商品を推薦するのみならずWebサイトをパーソナライズするために用いられています[1]。Item-to-Item Collaborative Filteringと呼ばれる推薦アルゴリズムが1998年に導入されました[2]。購入商品と類似すると判定された商品を推薦します。類似商品表作成など、計算の大部分をオフラインで実施しておき、推薦対象ユーザの購入物をもとに表検索を行う部分だけオンラインで計算することで高速に推薦をおこなっています[1]。その結果Amazon.comのサイトの視聴の30%は、推薦に基づくものと推定されています[2]。

[1] Greg Linden, Brent Smith, and Jeremy York, "Amazon.com Recommendations Item-to-Item Collaborative Filtering," IEEE Internet Computing, 7, 1, pp. 76-80, 2003.

[2] Brent Smith and Greg Linden, "Two Decades of Recommender Systems at Amazon.com," IEEE Internet Computing, 21, 3, pp. 12-18, 2017.

ビッグデータの活用事例：推薦行為事例

推薦行為の別の具体例としてYouTubeについて見てみましょう。

YouTube の推薦アルゴリズムは以下のようなものです[1]。あるビデオAを見た人が見る別のビデオ（共通視聴数）に基づく関連性（relatedness）を尺度にビデオAの関連性Top-Nビデオ(Aの関連ビデオ集合 $R(A)$)をあらかじめ計算しておきます。そしてユーザの視聴履歴などこの計算結果の表を組み合わせる推薦ビデオを提示しています。例えば、Aを見たユーザには「 $R(A)$ に含まれるビデオ」と「 $R(A)$ に含まれるビデオの関連ビデオ集合のビデオ」、などを提示するという具合です。

[1] J. Davidson, et al., "The YouTube Video Recommendation System," Proc. 4th ACM Conf. Recommender Systems, pp. 293–296, 2010.

ビッグデータの活用事例：推薦行為事例

Netflixの推薦はどうなっているのでしょうか。

Netflixの80%は推薦によると言われています[1]。それほど推薦の影響は大きいわけです。現在Netflixの推薦は、多くのアルゴリズムの集合体になっており、画面の下部にマトリクス状に配置された小画面の各行が1つのアルゴリズム（例えば、特定のカテゴリの全ビデオについて、そのユーザの属性から予想される評価点順に提示する）に対応しています[2]。

[1] Brent Smith and Greg Linden, "Two Decades of Recommender Systems at Amazon.com," IEEE Internet Computing, 21, 3, pp. 12-18, 2017.

[2] C.A. Gomez-Urbe and N. Hunt, "The Netflix Recommender System: Algorithms, Business Value, and Innovation," ACM Trans. Management Information Systems, 6, 4, pp. 1-19, 2016.

ビッグデータの活用事例：人流測定

人流測定による行動把握と予測、制御もビッグデータの有力な活用分野です。例えば、

- 博物館内の人の流れを測定し、見学ルートを変更する[1]
 - 店内の商品レイアウトを変更する
 - 道案内標識の適切な設定位置の決定
 - 電子表示装置と連動させた避難ルート指示[2]
- などがあります。

[1] <https://xtech.nikkei.com/it/atcl/column/14/122600137/122600001/> [Access 2021/2/26]

[2] <https://www.airc.aist.go.jp/achievements/ja/p-028.html>

通常時と災害時で電子表示装置の表示内容を変化させる



[2] (産業技術総合研究所人工知能研究センター社会知能研究チーム)

ビッグデータの活用事例：機械の稼働ログ

人の行動ログもビッグデータとなったように機械の稼働ログもビッグデータとなり得ます。その先駆けとも言える事例がコマツの建設機械に標準装備され稼働状況を監視するKOMTRAXです。インターネットなどのネットワークを通じ、機械のデータがサーバに送信されます。これにより、世界中のコマツの機械の状況が一元的に把握できます[1]。

また、IHIのガスタービン発電プラント運用支援ではプラントの運転データを収集し蓄積し、集まったビッグデータをもとに故障の予兆を発見します[2]。

[1] <https://sanki.komatsu/komtrax/>

[2] https://www.ihi.co.jp/powersystems/lifecycle/operation_support/index.html

ビッグデータの活用事例：情報通信システムログ

情報通信システムのログもビッグデータです。これをもとに、故障診断などを行う試みがなされています[1]。

もともと大規模ネットワークには、事前設定されたルールに従いアラームが発生する仕組みが備わっています。しかし、ハードウェアの致命的故障など単体で異常と分かるものの以外のもの、一時的な故障や、致命的な故障に至る段階での軽微な故障には、十分には機能せず、ベテランオペレータの経験によるところも多いのが実際です。そこで、故障に関わるメッセージからコンソール履歴まで含むログデータの活用が期待されます。

故障などの1つのイベントが発生すると関連する多くの機器で大量のメッセージが生成されます。こうした共起関係を利用して何が起きているかを把握することができるようになり故障診断精度が上がります。

[1]木村達明、Syslog分析による大規模ネットワーク故障診断、電子情報通信学会誌、98、9、pp.823-828、2015.

ビッグデータの活用事例：活用データ別

これ以外にもビッグデータを機械学習・AIなどで処理する事例は数多く見受けられます。各分野について見ていきましょう。

1. 画像データ活用

画像データは非常に活用が進んでいるようです。

- 自動運転車両プログラム[1]
- 医療画像データからのガンなどの検出[2][5]
- 食品原料検査[3]
- 車両画像識別[4]

[1]<https://jp.techcrunch.com/tag/mobileye/>

[2]<https://pubmed.ncbi.nlm.nih.gov/21452728/>

[3] <https://monoist.atmarkit.co.jp/mn/articles/2004/03/news080.html>

[4] <https://cloudplatform->

[jp.googleblog.com/2017/05/googlecloudplatform-aucnet-ibs.html](https://cloudplatform-jp.googleblog.com/2017/05/googlecloudplatform-aucnet-ibs.html)

[5] <https://aitimes.media/2019/06/21/3013/>

[1] –[4] Access 2021/3/6, [5] Access 2021/3/7

ビッグデータの活用事例：活用データ別

2.音声データ

音声は認識に加えて、最近では、合成も盛んです。

●音声認識

Apple 「Siri」

Google 「Google assistant」

NTTドコモ 「しゃべってコンシェル」

Amazon 「Amazon Echo」

Microsoft 「Cortana」, 「Azure cognitive services」

など

●音声合成

ニュース読み、AIアナウンサー

[1]<https://www.itmedia.co.jp/news/articles/1803/26/news127.html>

[2]<https://www.nhk.or.jp/voice/yomiko/>

[3]<https://www.sony.jp/professional/ai-announcer/>

[4]<https://www.nikkei.com/theme/?dw=19102801>

[1]-[4]Access 2021/3/6

ビッグデータの活用事例：活用データ別

3.文書データ、自然言語データ

翻訳に加えて、最近では、文章の合成や原稿作成が行われるようになってきました。

●機械翻訳

Google「Google natural machine translation」
Microsoft「Microsoft translation」,「Azure」
など

●文書合成、原稿作成

Articoolo
PSW
ライターズボックス
EasyWriter
automatedinsights
など

ビッグデータの活用事例：産業分野別

産業分野別で見てみましょう。

1. 農業

●圃場の状況を撮影したり、センサーで計測したりして集めたビッグデータを解析し、効率的に栽培管理し、ブドウの生産に用いる[1]-[3]

●ビッグデータを使って農業の生産性の向上や経営の改善を目指す農業データ連携基盤協議会 -WAGRI協議会-の発足[4]

[1]<https://smartagri-jp.com/smartagri/20>

[2]<https://smartagri-jp.com/smartagri/113>

[3]<https://smartagri-jp.com/smartagri/155>

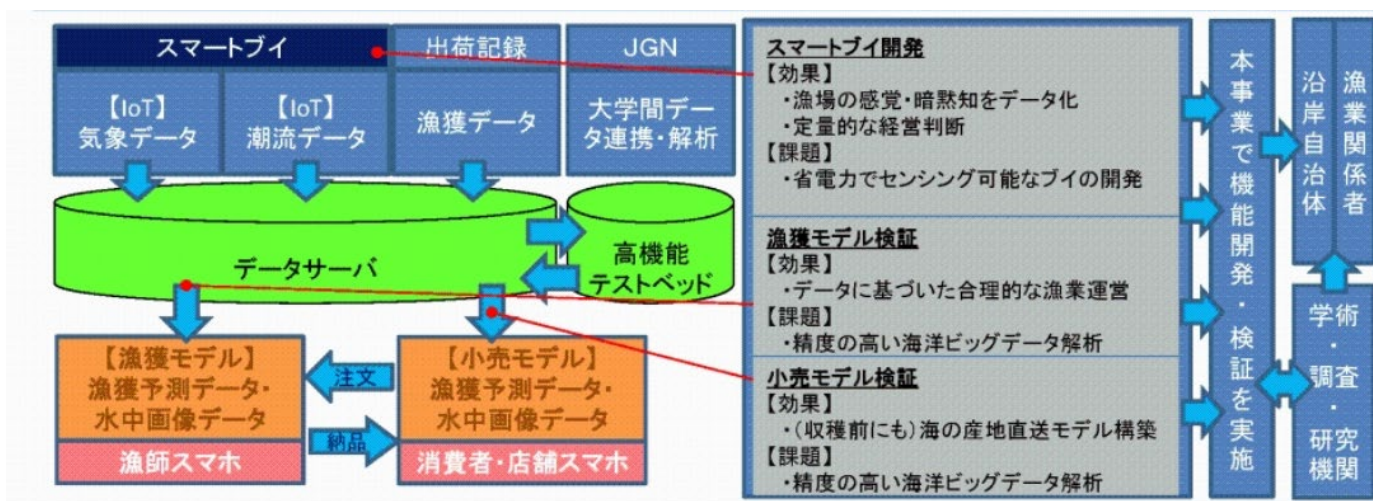
[4]<https://wagri.net/ja-jp/>

[1]-[4]Access 2021/3/6

ビッグデータの活用事例：産業分野別

2. 漁業

- 定置網漁などに対するスマート漁業モデル事業[1]
気象データ、海流データ、漁獲データに基づく漁獲モデルの構築など
- 遠洋かつお・まぐろ漁業のビッグデータ化[2]
海況予測や資源評価、魚群の来遊予測
- 養殖魚管理、作業効率化[3]



[1]より

[1]https://www.soumu.go.jp/main_sosiki/joho_tsusin/top/local_support/ict/jirei/2017_013.html

[2]https://www.ifa.maff.go.jp/j/kikaku/wpaper/h29_h/trend/1/t1_1_3_3.html

[3] <https://nissui.disclosure.site/ja/themes/144>

[1]-[3]Access 2021/3/7

ビッグデータの活用事例：産業分野別

3.製造業

●製造業のIoTプラットフォーム

ファナックの「Field System」

DMG 森精機の「ADAMOS」

Edgecross

●生産工程不良品検査、予測

[1]<https://www.automation-news.jp/2018/06/32869/>

[2]<https://www.omron.co.jp/technology/omrontechnics/2019/20190510-ogino.html>

[3]

https://www.hitachi.co.jp/products/it/lumada/usecase/case/lumada_uc_00016.html

[1]-[3] Access 2021/3/7

など

ビッグデータの活用事例：産業分野別

4. エネルギー産業

- 電力需要予測など発電事業への導入 [1][2]
- 電力メータ（スマートメータ）データを用いた他産業との連携[2]

配送業の配送効率化

家電利用状況管理

銀行口座本人確認

[1] <https://eneken.iecej.or.jp/data/8680.pdf>

[2]

https://www.toshiba.co.jp/tech/review/2019/05/74_05pdf/a06.pdf

[3] https://www.enecho.meti.go.jp/about/special/johoteikyo/digitalization2019_3.html

[1]-[3] Access 2021/3/7

ビッグデータの活用事例：産業分野別

5. 医療

- 医学論文データに基づくガン種別の特定[1]
- 創薬分野[2]
- 画像認識による診断(ビッグデータの活用事例：画像データのページ参照)
- ヘルスケアー支援[3]
- 血液データによるガン発見[4]
- ゲノムデータを用いた創薬、個別化医療

[1]<https://www.nikkei.com/article/DGXLZO05697850U6A800C1000000/>

[2] http://cbi-society.org/home/documents/seminar/2017to20/CBI391_Sawada.pdf

[3] <https://www.apple.com/jp/healthcare/>

[4] <https://grail.com/>

[1]-[4]Access 2021/3/7

ビッグデータの活用事例：産業分野別

6. インフラ管理

下水道管に関するデータに基づく道路陥没対策[1]
社会資本情報プラットフォームによるインフラ情報ビッグデータ化[2]
構造物損傷評価支援[3][4]
河川管理[5]

[1]https://www.gesui.metro.tokyo.lg.jp/business/pdf/3-1-3_2018.pdf

[2] <https://www.ipf.mlit.go.jp/ipf>

[3] https://www.scopenet.or.jp/main/scope_net/pdf/Vol71.pdf

[4] <https://www.shutoko.co.jp/efforts/safety/idreams/>

[5] https://www.ktr.mlit.go.jp/ktr_content/content/000775651.pdf

[1]-[5] Access: 2021/3/10