




ビッグデータ講座

目次

- ビッグデータ概要
- ビッグデータとセキュリティ
- ビッグデータを支える技術 (クラウド、IoT)



ビッグデータ概要

ビッグデータとは

■ ビッグデータとは

一般的なデータ管理・処理ソフトウェアで扱うことが困難なほど巨大で複雑なデータの集合のこと。

ビッグデータを取り巻く課題の範囲は、情報の収集、取捨選択、保管、検索、共有、転送、解析、可視化等多岐にわたる。これら課題を克服しビッグデータの傾向をつかむことで「ビジネスに使える発見、疾病予防、犯罪防止、リアルタイムの道路交通状況判断」に繋がる可能性がある。

つまり、ビッグデータとは、

■ Wikipediaより

- ・従来のシステムでは処理できないほど巨大なデータ
- ・定型を持たない複雑なデータ
- ・発見、予防といった新たな価値をもたらす得る、2次元的情報をもたらすデータ

であることが分かります。

「ビッグデータ」を調べると、曖昧な説明しか見つかりませんが、それらを注意して読むとエッセンスがみえてきます。以下はビッグデータの特徴です。

- ・普通の処理では対応できない量
- ・定型なデータではない
- ・本来の目的とは異なる使い方により二次効果を得る

このようなデータが身近にあるか、想像してみてください。

ビッグデータの量

■ 巨大なデータとはどれくらい？

ビッグデータの例を見てみましょう。

データ量を表す単位は、以下の順に1024倍となります。

キロ(KB)<メガ(MB)<ギガ(GB)<テラ(TB)<ペタ(PB)<エクサ(EB)<ゼタ(ZB)

全世界で生成・消費されるデジタルデータの総量

IDC (International Data Corporation) の発表： 59ゼタバイトを超える

出典：<https://www.idc.com/getdoc.jsp?containerId=prUS46286020>

ビッグデータの量はどれくらいでしょうか。その規模は、もはや1日TB級のデータを扱うレベルであるということをお覚えておいてください。

ビッグデータの質

- 蓄積されているデータはどのようなデータでしょうか？空欄を埋めてみましょう。

ビッグデータ

ブログデータ	車の位置情報	Web 操作ログ	気象情報
ドライブレコーダー	コールセンター音声	水質データ	防犯カメラ映像
人口統計情報	電力データ	SNS 写真	メール・チャット
ネット検索履歴	ツイートデータ	自販機前の動作映像	

従来型

販売 POS データ	EC 売上データ	販売・生産実績
EXCEL のデータ	基幹データベース	会計システムデータ

ビッグデータといえどどのような種類のデータを思い浮かべますか？空欄に思い浮かぶものを考えてみましょう。また、それらと比較し、従来型のデータと決定的に異なる点は何でしょうか。ビッグデータでは、「定型業務で発生したデータ以外も扱い、それらを組み合わせて分析する」ということを覚えておいてください。

ビッグデータの種類

ビッグデータの分類	構造化データ	準構造化データ	非構造化データ
分類の意味	データベースに格納される行列の二次元テーブルで表現されるデータ。 それほど増加しない見込み。 例・顧客テーブルデータ ・受注テーブルデータ ・CSV データ ・Excel データ	完全な構造定義を持たないデータ。 例・ログデータ ・センサーデータ ・SNS に書き込まれたデータ	データ部に構造定義を全く持たないデータ。準構造化データと合わせてデータ総量の 80% を占め、5年で800%の増加傾向。 例・文書 ・音声 ・動画 ・画像
前のページの例を当てはめると…	販売 POS データ 販売・生産実績	ツイートデータ Web 操作ログ	防犯カメラ映像 コールセンター音声

ビッグデータには構造化データ、準構造化データ、非構造化データがあります。特に非構造データの増加率が爆発的です。それぞれの違いをしっかりと覚えておいてください。また、前ページのデータをこのページの定義をもとに分類してみてください。

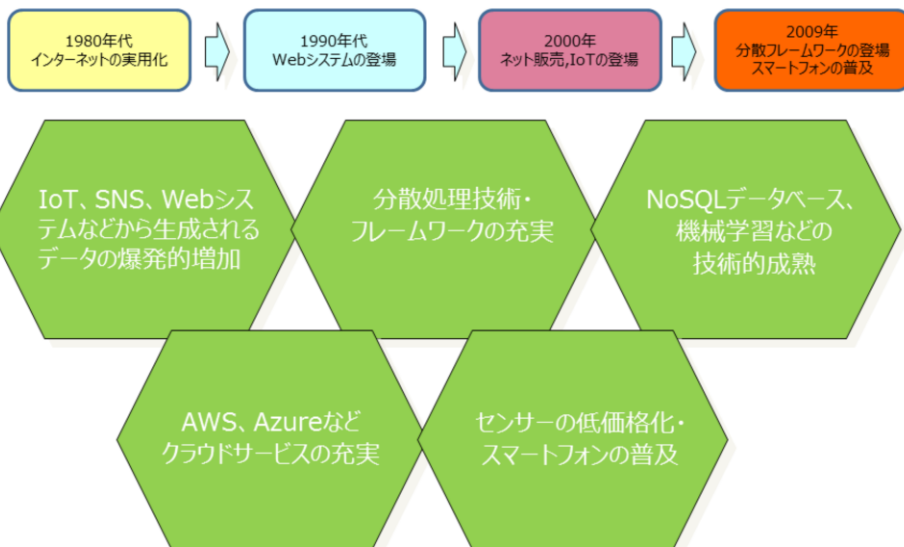
ビッグデータの3V



ビッグデータの5V(3V + Value + Veracity) ... 最近提唱されている。

量と多様性（質）のほかに、ビッグデータが持つ特性として、「リアルタイムにいつでも発生するデータ生成頻度」が挙げられます。センサーデータやTwitterデータなどは常に発生し、処理をしなければなりません。これらを合わせビッグデータの3Vと呼びますが、最近では、Veracity、Valueなども合わせて、4V、5Vといわれることがあります。

ビッグデータ登場の背景



処理能力の向上とWebシステムの登場が大きいといえます。例えば、googleなどの検索内容から顧客の購買動向を把握することにビッグデータが活用されていきました。今日に至るまで、様々な技術により、巨大なデータ処理技術が支えられています。「データを保持し続ける状況とそれを分析するニーズ、技術の進歩がマッチして初めて、ビッグデータが登場する契機が生まれた」ということをしっかりと覚えておいてください。

ビッグデータの所在

社内 (ローカル)	自社基幹システム 販売実績や、生産実績データ、会計データなど		Web、SNSサービス等 ECサイト、社内ポータルサイト、アプリ操作ログ	
	顧客・ユーザー スマホや家電、メーター上のデータ	グループ企業 企業間で共有される情報	取引企業 サプライチェーンの情報	
社外	政府・自治体等 統計データや地図情報など公開されている情報	提携企業 SNSデータ、位置情報空間統計、交通機関乗降情報等	データ提供事業者 地図、統計情報など目的に合わせて整備したデータ	
	一般			

ビッグデータはどこに保管されているのでしょうか。取引先から渡ってくるデータや今まで保存していたログデータ、社外Webサイトなど、自社内はもちろん、自社ローカル以外にも眠っている場合があります。また公共団体が提供するデータや、事業者が提供する商用データなどもビッグデータである場合もあり、ビッグデータはあらゆるところに散在しています。

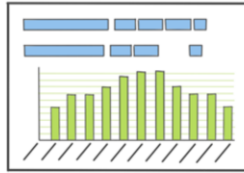
オープンデータとは

特徴	<ul style="list-style-type: none">■ 誰でも入手可能で、自由に利用・再配布できる状態で存在する■ 特許・著作権に制限がない■ コンピューターから利用できる状態となっている
公開主体	<ul style="list-style-type: none">■ 政府■ 地方自治体■ 研究機関・大学■ 民間企業
具体例	<ul style="list-style-type: none">■ 国勢調査データ（政府統計の総合窓口「e-Stat」）■ 公共施設やAEDの位置データ■ 気象データ■ 有志により作られた地図データ（OpenStreetMap など） 「行政と市民によるオープンデータ共創支援プラットフォーム（LinkData）」

官公庁が公開しているデータなどをオープンデータと呼ぶことがありますが、これもビッグデータの1つです。オープンデータは誰でも、いつでも、利用・再配布できる状態にあるデータです。「身近なオープンデータ」にはどんなものがあるか、考えてみてください。

オープンデータ×ビッグデータ活用例

- 過去のTwitterなどのSNS上の書き込み + 販売データ
→ 相関を調べ、売上の増減や欠品可能性を予測する。
- 自社の販売データ + 気象データ
→ 気象変化と売上推移の相関を見出し、予測を行う。
- 医療施設の位置データ + 患者の郵便番号のデータ
→ 来院マップを作成し、診療費ごとの外来状況を分析することで、地域医療に関して重点的な連携、促進を図る。
- 国勢調査などの人口統計情報 + 将来の人口推計
+ ターゲット層の世帯が多数存在する地域の売上相関
→ 重点的に販売を行う地域を模索する。



オープンデータ単体では価値を生み出さない場合がありますが、オープンデータとビッグデータを掛け合わせると、それを単体で使うより大きな価値を生み出す場合があります。もちろん、公共施設マップなど、オープンデータそのもので価値があるデータも存在します。

- ①気象データ単体を利用するケース
- ②ビッグデータを掛け合わせるケース

を考え、価値の違いがあるかを考えてみましょう。

ビッグデータが活用される分野

マーケティング

Webやカメラから顧客の行動を分析
→レコメンドシステム



製品開発

センサーからのフィードバック・顧客の声
→開発の指針

コンプライアンス

文書の全文検索とトピック抽出
→不適切な行動を察知



セキュリティ

サイバー攻撃のパターン検知
→予防策



メンテナンス

センサーデータのパターン検知
→故障予測

社会インフラ

気象などのデータ、地図情報
→災害の予測、避難経路



ビッグデータはどのような分野で活用されているでしょうか。ビッグデータ適用分野は多岐に渡り、これまでのアプリケーションの枠を超え、全く新しいデータ活用をもたらします。今まで耳にしたビッグデータの活用分野を挙げてみてください。

ビッグデータが活用される業界

	運輸	金融	医療・健康	製造
利活用事例	<ol style="list-style-type: none"> 1. 航空機チケットの割引サービスの改善：Web販売サイトでの購入傾向を分析 2. 渋滞予測：スマホGPSの情報を分析 3. トラックの最適な輸送ルート：過去の情報から算出 	<ol style="list-style-type: none"> 1. 金融商品の提案・開発：数千万件の顧客情報から、決済や資産運用の動きを分析 2. 保険サービスの開発：車の走行距離の情報を元に保険料を定めるサービスを開始 	<ol style="list-style-type: none"> 1. インフルエンザ対策：SNS上のコメントや検索キーワードから、広がりを検知 2. メディケア（アメリカの公的保険）のデータ公開：公的機関が分析して公開 	<ol style="list-style-type: none"> 1. 品質改善のための最善策の策定：生産工程に関する多数の情報を収集、分析 2. 製品開発：建機にセンサーを設置し故障時の稼働環境を分析してフィードバック。
効果	販売促進 人流動態分析	新サービス開発	兆候把握・ 情報提供	品質向上

ビッグデータ活用の具体的な例です。新聞などで日々掲載される技術革新が、ビッグデータによって支えられているケースも増えています。



ビッグデータとセキュリティ

■ 個人情報とは

個人情報の保護に関する法律 第二条 第1項 第一号において、次のように定義されています。

「生存する個人に関する情報であつて、」「当該情報に含まれる氏名、生年月日その他の記述等」「により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）」

ビッグデータに個人情報が含まれている理由で、活用を断念するケースがあります。ここでは、個人情報を取り上げます。

重要なのは、データ単体では個人を特定できない情報であっても、他のデータと組み合わせることによって、個人を特定できる場合があるということです。「他の情報と容易に照合」することにより、「特定の個人を識別」ができてしまう例として、下記が挙げられます。

- ・ 過疎地の郵便番号
- ・ 希少な病名
- ・ 顧客IDと顧客マスター
- ・ 写真とGPS情報 など

個人情報保護法

■ 個人情報保護法

- ・ 成立：2003年（平成15年）5月23日
- ・ 施行：即日（但し、一般企業に直接関わり罰則を含む第4～6章を除く）
- ・ 全面施行：2005年（平成17年）4月1日 … 成立の2年後

■ 個人情報取扱事業者

- ・ 個人情報を個人情報データベース等として所持し事業に用いている事業者のことをいう。
- ・ 個人情報保護法および同施行令により、取扱件数に関わらず、個人情報取扱事業者とされるようになった。
- ・ 主務大臣への報告や、それに伴う改善措置に従うなどの適切な対処を行わなかった個人情報取扱事業者に対しては、刑事罰が科される。

個人情報保護法は2015年に改正され、より範囲が明確になり、運用方法も明確に定義されました。ここで重要なのは、「個人情報を取り扱う場合には、何らかの形で個人の同意が必要」ということです。

■ 個人情報保護法 2015年の改正内容

- ・これまで対象外だった、5,000人以下の個人情報を取り扱う小規模な事業者に対しても、改正法が適用されるようになった。
- ・個人情報を取得する場合、予め本人に利用目的を明示することが必要となった。
- ・個人情報を他企業などの第三者に提供する場合、予め本人から同意を得ることが必要となった。
- ・オプトアウトには、個人情報保護委員会への届出が必須となった。
更に、第三者提供の事実、その対象項目、提供方法、望まない場合の停止方法などを、全て予め本人に示さなければならなくなった。
※オプトアウト：本人の同意を得ないで個人情報を提供できる特例のこと。
- ・「人種」、「信条」、「病歴」といった「要配慮個人情報」は、オプトアウトでは提供できないこととされた。

個人情報保護法は2015年に改正され、より範囲が明確になり、運用方法も明確に定義されました。ここで重要なのは、「個人情報を取り扱う場合には、何らかの形で個人の同意が必要」ということです。

個人情報保護の情勢

- 1980年 プライバシー保護と個人データの国際流通についてのガイドラインに関するOECD理事会勧告（OECDプライバシーガイドライン）
（OECDの34加盟国）
①収集制限 ②データ内容 ③目的明確化 ④利用制限
⑤安全保護措置 ⑥公開 ⑦個人参加 ⑧責任
の8原則からなる。
- 1995年 EUデータ保護指令（EUの28構成国）
EUおよび英国においては、十分なデータ保護レベルが確保されていない第三国への個人データの移動を禁止する。
- 2003年 個人情報保護法（日本）
個人情報を扱う事業者に対し、個人情報の適切な取り扱いを求める。
- 2012年 消費者プライバシー権利章典（アメリカ）
①個人によるコントロール ②透明性 ③背景情報の尊重
④セキュリティ ⑤アクセスと正確性 ⑥適切な範囲の収集 ⑦説明責任
の7つの権利を定める。

海外の個人情報に対する規制はどのようになって
いるかみてみましょう。

海外では個人の意思表示により、データの削除が
できる仕組みなど様々な規制が存在します。

個人情報保護法の情勢

- EU一般データ保護規定（GDPR）が可決（2016年4月 EU）
データポータビリティ権が提唱される。
→ 域外適応につき、日本の事業者に影響が出る。
- EU - USプライバシーシールドに米国と欧州委員会が合意（2016年2月 米国）
スノーデン事件を受けて無効化されていたセーフハーバーの後継。
商務省とFTCに強い権限が与えられ、企業に対して自主規制を求める機運が高まった。
- APEC 越境プライバシールールシステム（CBPRs）への参加（アジア）
日本に関しては、2014年にJIPDECがCBPR認証機関に認定された。
→ 2016年6月1日から申請受付開始。
- 個人情報保護委員会が発足（2016年1月 日本）
個人情報保護法の改正を受け、政府の第三者機関として設立した。
- 一般財団法人情報法制研究所（JILIS）が設立（2016年5月 日本）

海外の個人情報保護の情勢をもう少し詳しくみてみましょう。

EU法のGDPR（General Data Protection Regulation：一般データ保護規則）では、

- ①IDなども個人情報扱い
- ②EU域外に個人情報を持ち出せない
- ③規定に違反した場合などは制裁金が科せられる場合もある

などの点に注意が必要です。

オプトインとオプトアウト

■ オプトイン（事前承認）

明示的な同意が無い限り、個人情報やプライバシー情報は収集されないような仕組みのことを言います。

- 例・ショッピングサイトからのセール情報に関するメールの送付を許可する。
- ・個人情報の収集・利用を含むサービスの利用規約に同意する。

■ オプトアウト（事後承諾）

オプトインとは反対に、明示的に拒否していない限りは同意したものとみなし、明示的な拒否があった場合に個人情報やプライバシー情報の利用が停止されるような仕組みのことを言います。

- 例・Webサイトにおけるクッキーを用いた行動追跡
- ・ショッピングサイトにおける購買履歴の削除

個人情報の許諾を個人から得る方法として、オプトインとオプトアウトがあります。オプトインは事前承認、オプトアウトは事後承認と覚えておいてください。

匿名加工処理の手法

以下のそれぞれの手法を組み合わせることで、より強固な匿名化が実現されます。

技法大部類	No.	技法例	概要
摂動法	1	K-匿名化	同じグループ内に、同じ属性のユーザが「K人以上いる」状態を作り出す。
	2	L-多様性	漏えいさせたくない属性が同じグループ内で「L種類以上ある」状態を作り出す。
	3	T-近接性	マイナー属性を持つグループが生まれるなど、属性値の分布に偏りが出てしまう場合に、グループの分割や一般化を行う。
	4	差分プライバシー	2006年に提案された新しい手法。元のデータベースにノイズを足した別のデータベースを用意し、守りたいレコードを特定しづらくする。
暗号法	5	質問監査	データベースへのアクセス者に質問を投げかけ、答えられれば、アクセスに対する回答を返す。
	6	秘密計算	関係者全員が、自社データを他人が読めないように変換し、秘密計算のシステムへ投入する。そのシステムの管理者が、秘密計算の結果を求め、関係者に回答する。
	7	準同型性公開鍵暗号を用いた暗号プロトコル	遺伝子データなど、加工してしまうと、そもそも分析できなくなるデータを処理するときに活用。検索者の検索クエリ、データベース、その回答それぞれを暗号化する。分析者が元データにふれずとも、望む解析結果が得られる。

出典：中川裕志『プライバシー保護入門：法制度と数理的基礎』（2015年）

個人を特定できないようにデータを加工することを匿名加工処理と呼びます。具体的にどのような手法で匿名化を実現するのかをみてみましょう。

K-匿名化や、L-多様化はデータそのものを加工し、もしくはレコードを増やし、個人を特定できないようにする技術です。あまり、K=XXなどの数値を大きくし過ぎると、安全性は高まりますが、統計上や機械学習上のノイズになる場合があり、正しい結果が得られなくなる場合もあるので注意です。

識別子の削除（仮名化）

個人の識別・特定に直結するカラムを削除して、仮名化を行います。

No.	ZIPコード	年齢	職業	病状
1	13068	28	ダンサー	心臓病
2	13068	29	技術者	心臓病
3	13053	21	法律家	感染症
4	13053	23	技術者	感染症
5	14853	31	技術者	風邪
6	14853	37	作家	風邪
7	14850	36	法律家	がん
8	14850	35	技術者	がん

← 準識別子

→ 漏えいさせたくない属性

出典：「情報処理学会」(Vol.54 No.11 Nov.2013) より

識別子の削除（仮名化）は、個人の氏名や住所などの情報を仮名と置き換える、もしくは削除することによって、データから直接個人を特定できなくすることをいいます。この表の場合は、個人を特定できる氏名を削除し、住所をZIPコードに置き換えています。これによって、個人を特定することが容易ではなくなります。

K-匿名化したテーブル

次に、再特定・識別につながる「職業」を秘匿した上で、「年齢」、「病状」の列に「同じ値が少なくとも2つ以上は存在する状態」のテーブルを作ります。

No.	ZIPコード	年齢	職業	病状
1	13068	28-29	*	心臓病
2	13068	28-29	*	心臓病
3	13053	21-23	*	感染症
4	13053	21-23	*	感染症
5	14853	31-37	*	風邪
6	14853	31-37	*	風邪
7	14850	35-36	*	がん
8	14850	35-36	*	がん

出典：「情報処理学会」(Vol.54 No.11 Nov.2013) より

K-匿名化は必ず、K=数値で表された数以上レコード（行）が存在するようにデータを加工します。同じ保護属性（漏洩させたくない情報）の組み合わせを持つレコードが、少なくともK個存在し、保護属性からの識別がK人未満に絞り込めない状態になります。この場合はK=2で、同じ識別子をもったレコードが少なくとも2個以上存在するようにしています。そのために、例えば年齢を28という数値ではなく28-29という幅で表すなどの加工をしています。

L-多様化したテーブル

「ZIPコード」と「年齢」を曖昧にして、「どのレコードを取り出しても、2種類の「病状」が存在する状態」になるようにします。

No.	ZIPコード	年齢	職業	病状
1	130**	21-29	*	心臓病
2	130**	21-29	*	心臓病
3	130**	21-29	*	感染症
4	130**	21-29	*	感染症
5	148**	31-37	*	風邪
6	148**	31-37	*	風邪
7	148**	31-37	*	がん
8	148**	31-37	*	がん

出典：「情報処理学会」(Vol.54 No.11 Nov.2013) より

L-多様性は、同じ保護属性の組み合わせを持つレコードが、複数個存在し、かつ対応する非保護の属性情報の値が少なくともL種の多様性を持つことで、属性推定が起らない状態です。具体的には似たようなレコードを追加して、個人を推定できない状況を作り出します。この場合は、年齢帯とIPコードの上3桁の組み合わせからレコードを抽出しても、一つの保護属性（病状）には絞れないようになっています。例えば、130**かつ21-29でも、心臓病と感染症の2つの保護属性が抽出されます。

ビッグデータを支える技術

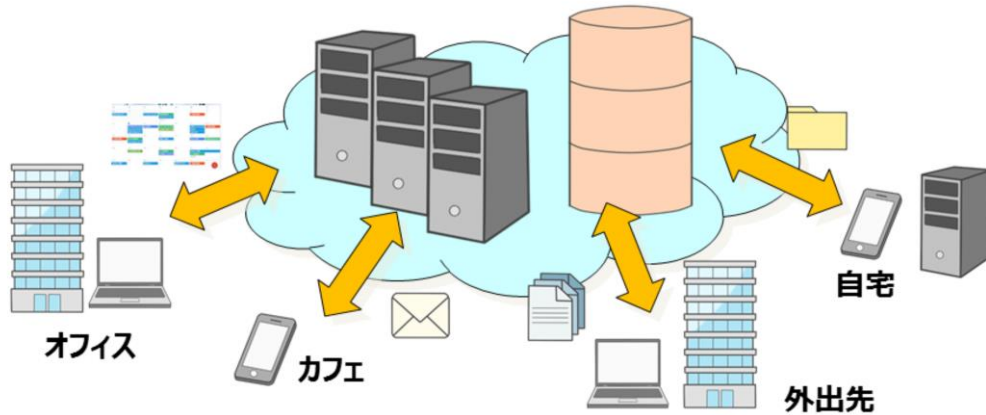
- クラウド、IoT

ビッグデータ周辺技術 クラウド

■ クラウドコンピューティングとは

メールやグループウェア、その他様々なサービスをインターネット上で提供し、インターネット上にデータまでも保存するようなサービス形態のことをいいます。

アプリケーションの実行場所が、雲（クラウド）のようにどこにあるのかが分からないモヤモヤとした場所にあることから、このように呼ばれています。



ここからは、ビッグデータを支える周辺技術を学んでいきます。

クラウドサービスはビッグデータ登場の背景にも密接にかかわっており、ビッグデータの置き場所などに活用されています。クラウドはどこに存在するか分からないデータセンターのことです。データの共有や業務アプリなどの様々なサービスが提供されています。

クラウドサービスの種類

■ クラウドサービスの種類

■ SaaS (Software as a Service)

インターネットを経由してソフトウェアパッケージを提供するサービスのこと。アプリをPCにダウンロードしなくても、Webなどのブラウザ上で利用することができる。

例・メール ・カレンダー ・チャット

■ PaaS (Platform as a Service)

インターネットを経由してアプリの開発・運用環境全体を提供するサービスのこと。システム管理者、開発者向けである。

■ HaaS / IaaS (Hardware as a Service / Infrastructure as a Service)

インターネットを経由してハードウェアや回線などのインフラを提供するサービスのこと。ユーザーはハードウェア資産を所有することなく、仮想サーバーやストレージ（外部記憶装置）を利用することができる。

クラウドには様々な利用形態の種類があります。代表的な分け方は上記のようにSaaS（ソース）、PaaS（パース）、IaaS（イアース）の分類です。これらは提供するもの（レイヤ）の違いであり、

IaaS（イアース）は、ハードウェアやネットワーク、OSなどの基盤を

PaaS（パース）は、基盤の上で動くミドルウェアなども含む開発・運用環境を

SaaS（ソース）は、さらにその上で動くアプリケーションを

提供します。

クラウドサービスのメリット・デメリット

メリット

・サーバー・ソフトウェアを購入する必要がない

・システム構築期間を短縮できる

・効率的なIT投資やリソース配分を実現できる

・メンテナンスが不要である

・IT部門の負担が軽減される

・カスタマイズが基本的にできない、もしくは難しい

・不特定多数が利用するため、安定して稼働できないリスクがある

・セキュリティ面のリスクがある

・利用するデータ量、時間によっては費用が増加するリスクがある

デメリット

クラウドサービスのメリットとデメリットには、どのようなものがあるかみていきましょう。

サーバーの調達や、メンテナンスから解放される一方、カスタマイズの難しさや、不特定多数のユーザーの目にさらされる危険もあるため、個人情報の扱いがあるデータの運用には注意が必要です。

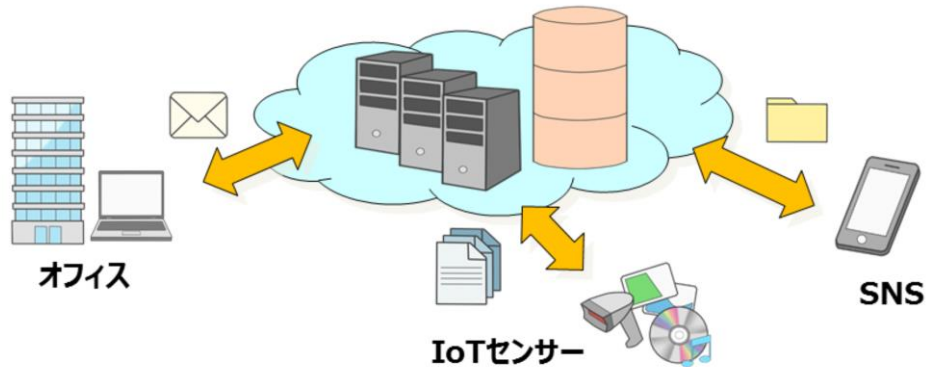
クラウドコンピューティングとビッグデータ

■ クラウドサービスに蓄積されるビッグデータ

クラウドサービスは、ビッグデータの蓄積場所に有力な選択肢です。

実際、以下はビッグデータがクラウドに保存されています。

- ・ GREE、mixi、Facebook、Twitterなどに代表されるSNSのユーザーデータ
- ・ IoTのようにセンサーなどで発生したデータ



ビッグデータの保存先として、サーバーサイジングを自由に変更できるクラウドサービスは有力な選択肢の1つとなります。IoTセンサーデータのデータ収集もクラウドサービスベンダーが担う状況も整いつつあります。

■ IoTとは

IoT = Internet of Things 「モノのインターネット」

ここで、「モノ」とは、ネットワークに繋がるあらゆる物のことです。
従来はインターネットとは関係のなかったもの、例えば、

- ・眼鏡 ・服 ・時計 ・冷蔵庫 ・電力メーター ・自動車
- ・太陽光パネル ・家 ・スマートフォン

もすべて「モノ」です。

IoTとは、モノがネットワークに繋がれることによって、モノとインターネットが相互に情報交換をできるようになった状態のことをいいます。

センサーデータを分析する際に耳にするIoTという言葉は何を意味するか確認しましょう。モノが双方向でインターネットにつながりデータのやり取りができる状態をIoTと呼びます。

IoTとは

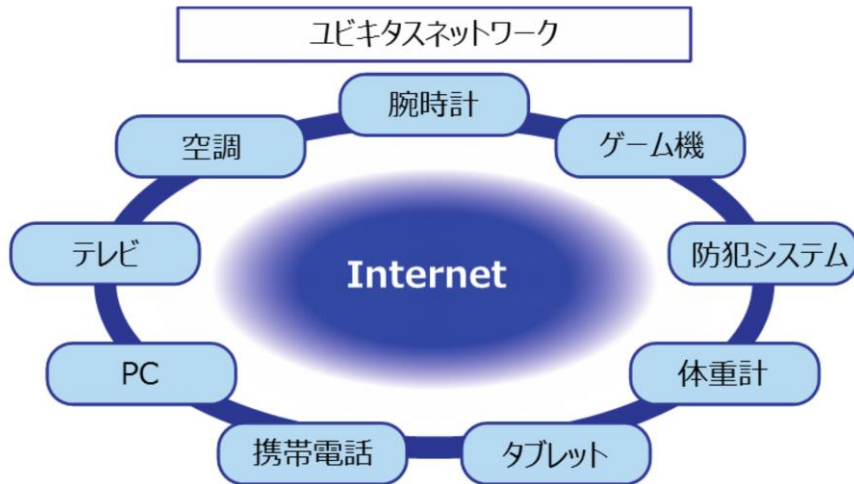
- Internet of Thingsという用語は1999年、ケビン・アシュトン（イギリス）によって初めて提唱されました。
- 当初はRFIDによる商品管理システムをインターネットに例えたものでした。
 - 徐々にスマートフォンやクラウドコンピューティングが普及。
 - IoTはモノ自体がインターネットを形作るという環境全体のことを表す概念として捉えられるようになる。
- IDC（ICT市場調査会社）による定義
「IP接続による通信を人の介在なしにローカルまたはグローバルに行うことができる識別可能なエッジデバイスから成るネットワークのネットワーク」

ここでは、IoTという言葉はいつ頃から使われ始め、定着していったのかをみていきましょう。

IoTは2000年頃から使われ始めましたが、当初は現在よりも狭義な使われ方でした。現在では、モノ自体がインターネットを形作る（モノがインターネットで情報交換する）という広い概念として捉えられています。

■ ユビキタスネットワークとは

いつでもどこでもインターネットを利用できるという概念のことをいいます。
1988年ゼロックス社パロアルト研究所のマーク・ワイザーにより提唱されました。



ここからは、IoTと似た概念であるユビキタスとM2Mについて確認し、その違いをみていきましょう。

まず、ユビキタスについてです。ユビキタスは、2000年頃に登場した概念で、人がどこにいてもネットワークにつなげる（つながる）状態を指します。ユビキタスネットワークの中心は人なので、IoTのモノとモノとの相互制御とはその点で異なります。

IoTとユビキタス社会

■ IoTとユビキタスの違い

- ・ IoT：モノとモノが相互に制御し合っている状態を表す。
…「モノ」を中心とした概念
- ・ ユビキタス：ユーザーが時間や場所にとらわれずインターネットに繋がって様々なサービスを受けられる状態を表す。
…ユーザーという「人」を中心とした概念

参考

ユビキタス (ubiquitous) は、遍在 (いつでもどこでも存在すること) をあらわす言葉。

パロアルト研究所のマーク・ワイザーが、1991年の論文『The Computer for the 21st Century』にて、コンピュータやネットワークなどの遍在を表す意味合いで用いた。以来、ユビキタスコンピューティングやユビキタスネットワーク、さらにはそれらが当たり前になった社会を指す「ユビキタス社会」の意味で用いられるようになった。

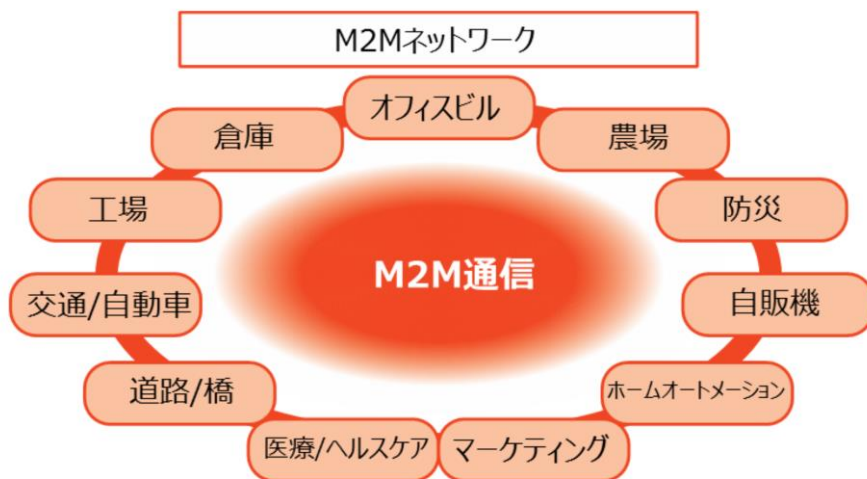
Wikipedia

ここからは、IoTと似た概念であるユビキタスとM2Mについて確認し、その違いをみていきましょう。

まず、ユビキタスについてです。ユビキタスは、2000年頃に登場した概念で、人がどこにいてもネットワークにつなげる（つながる）状態を指します。ユビキタスネットワークの中心は人なので、IoTのモノとモノとの相互制御とはその点で異なります。

■ **M2M = Machine to Machine**

機械同士が、人を介在しないで情報をやり取りするシステムのことをいいます。



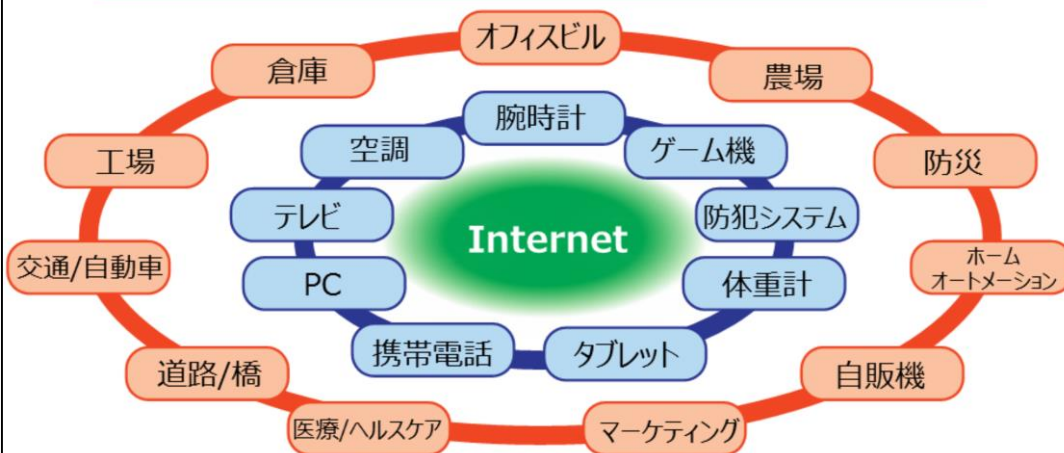
次にM2Mをみていきましょう。

M2Mは人が介在しない機械同士の連携です。機会が情報を集め、他の機械にその情報を提供しますが、人が分かる形でデータを表示することはありません。IoTは人が介在するセンサーも対象となりますし、結果をみるのも人である場合もあるため、M2Mの概念だけではIoTを説明することはできません。

IoTとM2M

IoTはユビキタスとM2Mを包括する概念といえます

IoT社会
Machine to Machine、Human to Machine、Human to Human



ここまで学んできたように、モノ自体がインターネットを形作るIoTは、ユビキタスとIoTを合わせたような概念といえます。人とモノ、モノとモノ、人と人をつなぐ総合的な社会インフラになる可能性を秘めています。

IoTを構成する技術要素

- センサー：物理的な現象を検知し、電気信号として出力する装置。
- デバイス：センサーが組み込まれることによって、ネットワークに接続された装置やモノ。
例 スマホ、時計、メガネ
- ネットワーク：デバイスをIoTサービスに繋ぐ、あるいはデバイス同士を繋ぐことでデータを共有、処理するシステム。
- IoTサービス：①デバイスとのデータの送受信 … IoT
②データの処理と保存 … ビッグデータの技術的守備範囲を行うサービス。
- データ分析：蓄積したデータについて統計分析や機械学習を行う。
→最適な判断や行動方針を導き出す。

IoTには基本的に、センサー、デバイス、ネットワーク、サービス、データ分析がセットになって、IoTサービス全体を形作っています。それぞれの技術要素は独立していることが多く、技術要素の組み合わせによって、IoTを実現します。IoT構成要素として欠かせない、データの処理と保存はまさにビッグデータの技術的守備範囲となります。IoTの実現にはビッグデータ技術が必要となります。逆にIoTを知ることでビッグデータの現実を知ることになります。

■ センサー

物理的な現象を検知し、電気信号として出力する装置のことをいいます。
多くの場合、一つのデバイスに対して複数のセンサーが埋め込まれています。

- ・ 画像センサー：光を捉えて処理することで、画像や動画を撮影する。
赤外線を検知して画像処理するものもある。
- ・ 光センサー：光の強度を測定する。
- ・ 温度センサー：温度を測定する。
- ・ 湿度センサー：湿度を測定する。
- ・ 振動/速度/加速度センサー：機器の振動や速度、加速度を測定する。
- ・ 地磁気センサー：地磁気を検出することで、方角を計測する。
- ・ ジャイロセンサー：デバイスの傾きを検知する。
- ・ 音声マイク：機器が発する音や、人の声などの音声を収集する。

IoTのセンサーにはどのようなものがあるかみてみましょう。データの中には、画像データや、音声データのような非構造データや、準構造化データ、バイナリーデータも存在します。

IoTを構成する技術要素 センサー

センサーの代表的なデータフォーマットとしては、
・XML ・JSON ・MessagePack
があります。

```
XML
<xml>
  <info>
    <id>12996</id>
    <name>RoomSensor</name>
    <date>20170123112255</date>
  </info>
  <data>
    <temperature>27.8</temperature>
    <humid>72</humid>
  </data>
</xml>
```

人が読んで分かりやすい
データ量が多い

```
JSON
{"info":{"id":123,
        "name":"RoomSensor",
        "date":20170123112255
      },
 "data":{"temperature":27.8,
        "humid":72
      }
}
```

データ量が少ない

いずれも各言語のライブラリが充実していますが、文字データであることから、パース（解析）をしないとプログラムで利用できません。
MessagePackは、バイナリデータをそのまま扱いたい場合、有利です。

センサーデータの基本的なフォーマットを見てみましょう。タグ構造のXMLや、Webサイトでよく使われるJSON形式もありますが、MessagePackのようなバイナリーデータも存在します。「センサーデータは、構造化データでないことが多い」ということを覚えておいてください。

■ MessagePack

- ・センサーの代表的なデータフォーマットの一つ。
- ・JSONと似た形式だが、値はバイナリのままである。
- ・軽量でプログラム間処理に向いている。

■ MessagePackの特徴

- ・シリアライズ（データの直列化）、デシリアライズ：非常に高速
- ・シリアライズされたデータのサイズ：小さい
- ・フォーマット定義：不要
- ・ストリーム処理：可能

■ JSONとMessagePackの比較

JSON `{"a":null,"b":10,"c":[20],"d":"30"}` … 35byte

→ MessagePack に変換すると…

`84 a2 61 c0 a2 62 0a a2 63 91 14 a2 64 a2 33 30` … 16byte

MessagePackの形式をみてみましょう。人がみて理解できる形式ではありませんが、データ量が小さく、変換が高速であることが魅力です。「IoTデータにはバイナリーデータも存在する」ということを覚えておいてください。

IoTを構成する技術要素 デバイス

■ デバイス

センサーが組み込まれることによって、ネットワークに接続された装置、モノのことをいいます。例えば、スマホ、時計、メガネはいずれもデバイスです。

■ デバイスの2つの機能

- ・ センシング：センサーを利用して、デバイス自身や周りの環境の状態を収集し、IoTシステムに通知すること。
 - 例・画像センサーによる人の有無の検知
 - ・スマホの位置情報や加速度の計測
- ・ フィードバック：システムからの通知を受け、指示や動作をもとのシステムに返すこと。次のような方法がある。
 - ・可視化：センシング結果表示、デバイスの管理、画面表示
 - ・通知：システムが判断した結果を画面に表示
 - ・制御：デバイス自身や環境の状態そのものを変更

ここでは、IoTにおけるデバイスとは何かをみていきましょう。

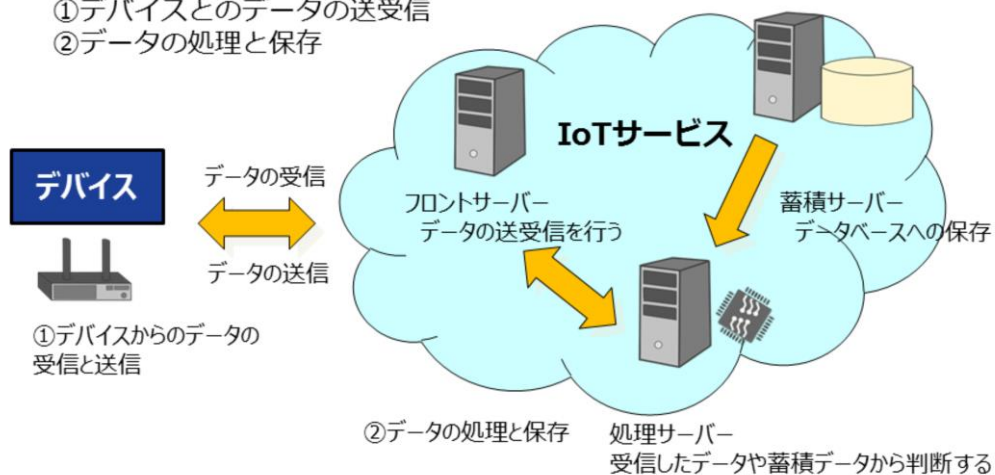
IoTにおいてデバイスとは、センサーとネットワーク接続を併せ持った機器を指します。スマートフォンやスマートウォッチに搭載されるセンサーとネットワーク通信機能もデバイスとして捉えることができます。デバイスには2つの機能が要求されており、1つは周囲の情報をシステムに伝えるセンシング、もうひとつはシステムからの指示によって動作を行うフィードバックです。システムにとっては、デバイスはインプット（センシング）とアウトプット（フィードバック）の両方を担う存在と言えるでしょう。

IoTを構成する技術要素 IoTサービス

■ IoTサービス

以下を行うサービスのことを指します。

- ①デバイスとのデータの送受信
- ②データの処理と保存



IoTサービスはデバイスから送られてきたデータを受信したり、蓄積、分析したり、サービスを行ううえで必要な処理を行うサーバーサイドの処理です。サーバーの主なものは下記です。

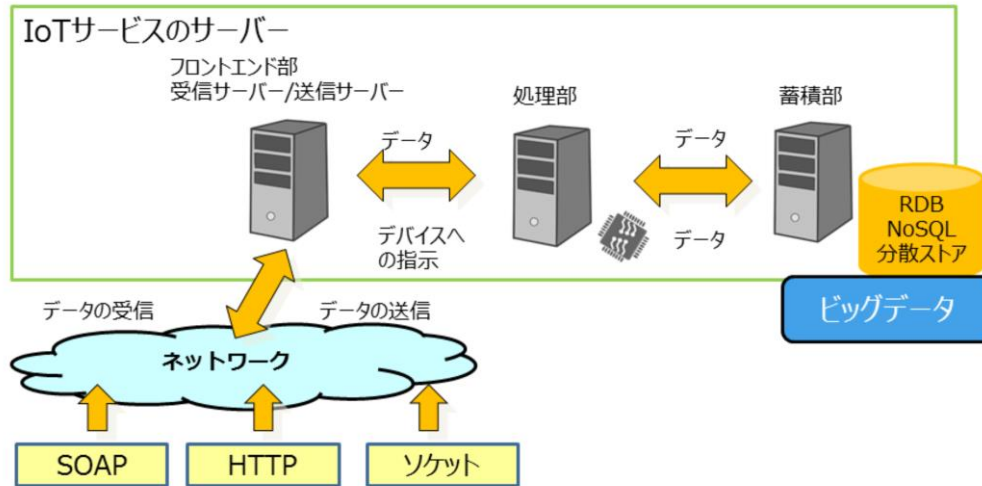
データの送受信を担当するフロントサーバー
データベースへの保存を行う蓄積サーバー
判断のためのデータ処理を行う処理サーバー

IoTで処理されるデータは、膨大であり構造化されていないものも多いため、ここで処理されるデータは、まさにビッグデータといえるでしょう。

IoTを構成する技術要素 IoTサービス

■ サーバー構成

IoTサービスの役割は、フロントエンド、処理、蓄積の大きく3つに分けられます。蓄積されるデータは膨大な量となります。



IoTのサーバー側の役割をもう少し詳しくみてみましょう。

サービスの奥にある蓄積部（蓄積サーバー）では、NoSQLや分散処理システムなど、非構造化データを大量に処理できる基盤が利用されています。大量の非構造化データは、当然ビッグデータの定義に当てはまります。つまり、IoTのサーバー構成は、ビッグデータを扱える能力が要求されているのです。