ビッグデータ講座

ビッグデータ概要

目次

第１章　ビッグデータ概要

目次

第２章 ビッグデータとセキュリティ

第１章

ビッグデータ概要

Computer Software Association of Japan

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・処理ソフトウエ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.

Wikipediaより

つまり、ビッグデータとは、

・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し得る、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。

CSAJ

・説明の流れ

ビッグデータとは何か、辞書的な説明ではあいまいとしていてはっきりとしていませんが、注意して読むとエッセンスが見えてきます。

・ポイント（絶対に覚えてほしいこと、など）

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2次効果を得る。

・質問（問いかけ）

このようなデータは身近にありますか。

・補足説明

・ファシリテーションテクニック
　身近な例を考えさせ、それがビッグデータかどうか、隣同士で指摘してもらう

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

・説明の流れ

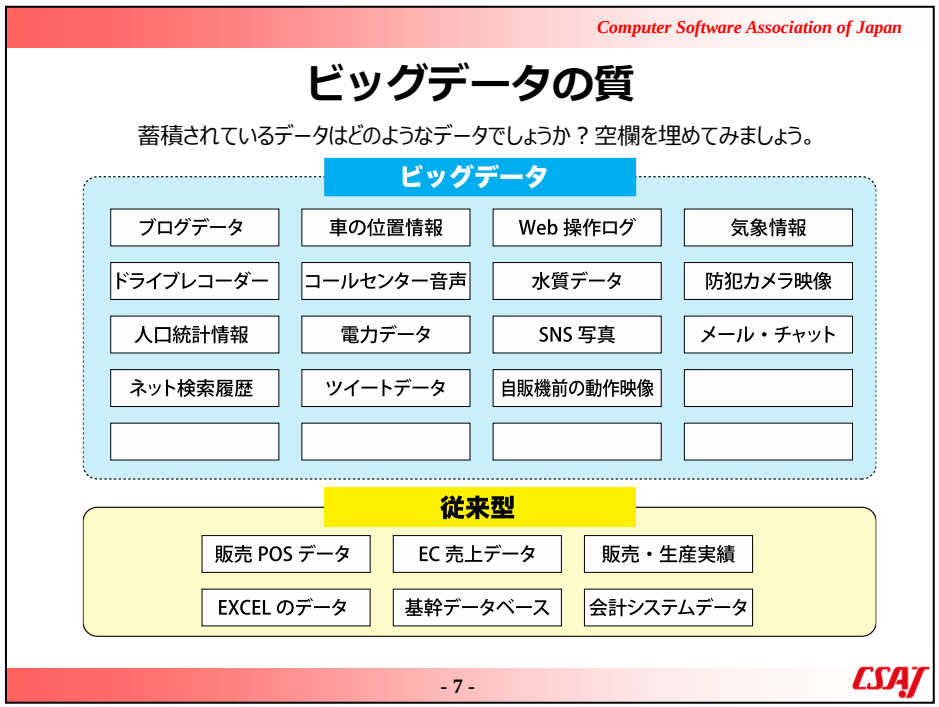
世に言うビッグデータの量はどれくらいでしょうか。

・ポイント（絶対に覚えてほしいこと、など）

ビッグデータの規模はもはや1日TB級のデータを扱うレベル。

・質問（問いかけ）

この規模のデータで思いつくものは？



・説明の流れ

ビッグデータと言えばどのような種類のデータを思い浮かべるでしょうか。

空欄に今まで上げたビッグデータを書いてみましょう。

これと比べて、従来型のデータと決定的に異なる点は何でしょう。

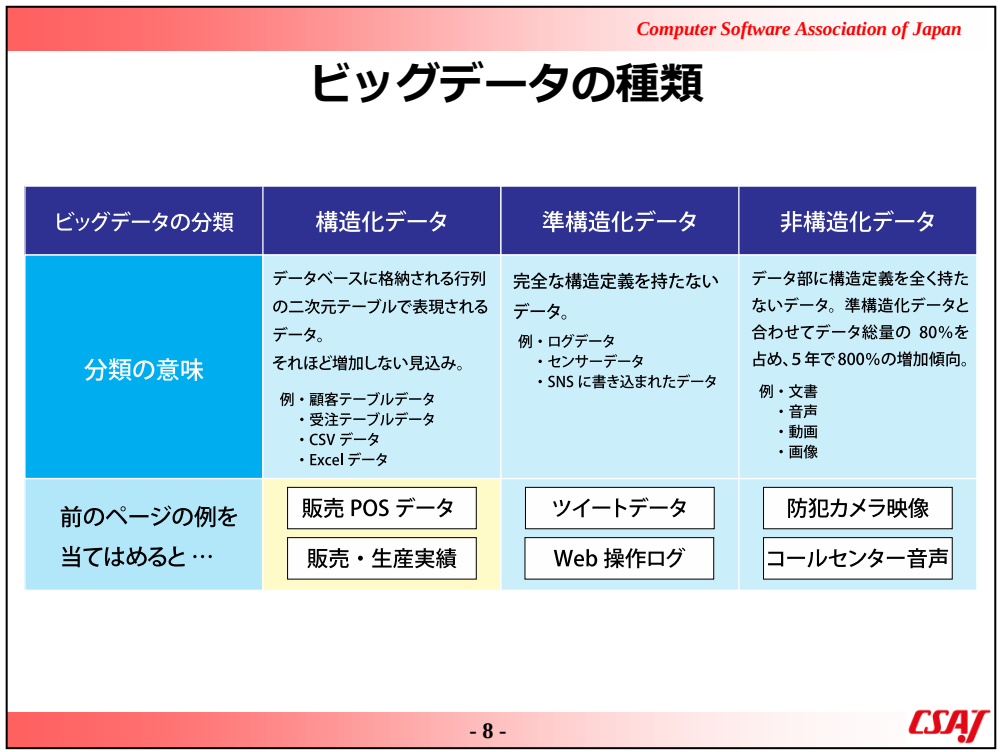
・ポイント（絶対に覚えてほしいこと、など）

ビッグデータでは、定型業務で発生したデータ以外も扱う。

またはそれらを組み合わせて分析する。

・質問（問いかけ）

従来型のデータと決定的に異なる点は何でしょう。



・説明の流れ

ビッグデータには構造化データと準構造化データ、非構造化データがあることを説明します。

特に非構造データの増加率が爆発的であることを強調します。

・ポイント（絶対に覚えてほしいこと、など）

構造化データと準構造化データ、非構造化データの違い

・質問（問いかけ）

前ページのデータを分類してみましょう。

・補足説明
・ファシリテーションテクニック
　分類した内容を隣の方と意見交換する。

・説明の流れ

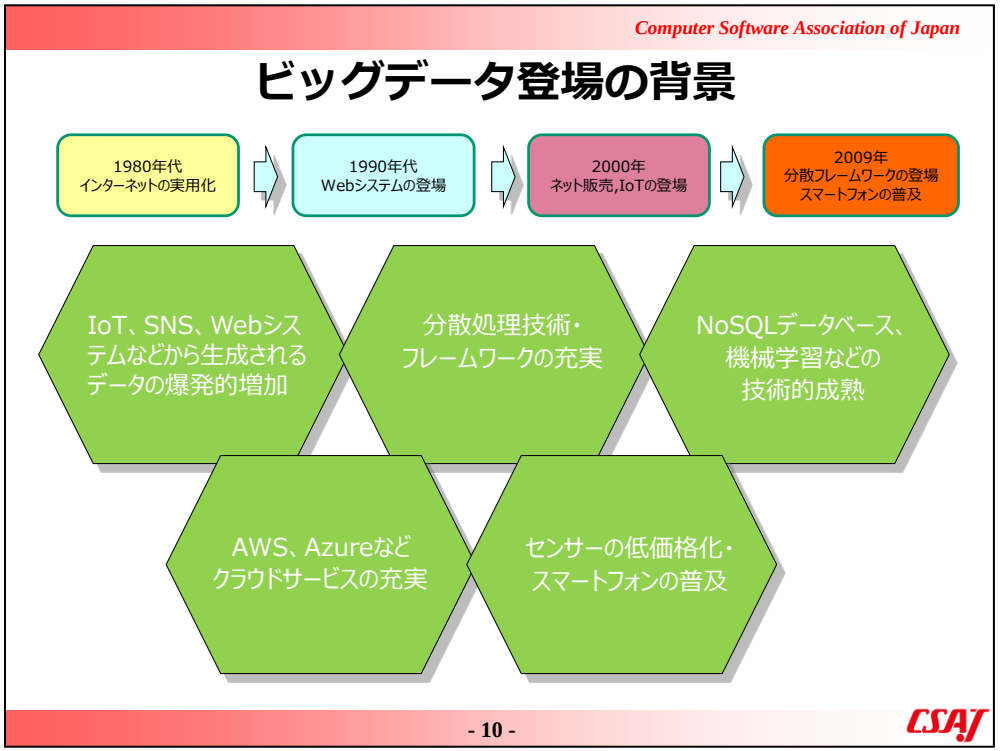
ビッグデータの特性はこれまで述べてきたように、量と多様性（質）があげられますが、もう一つ重要な要素として、リアルタイムにいつでも発生する　データ生成頻度があげられます。

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などというようになってきています。

・ポイント（絶対に覚えてほしいこと、など）

ビッグデータの３Ｖ



・説明の流れ

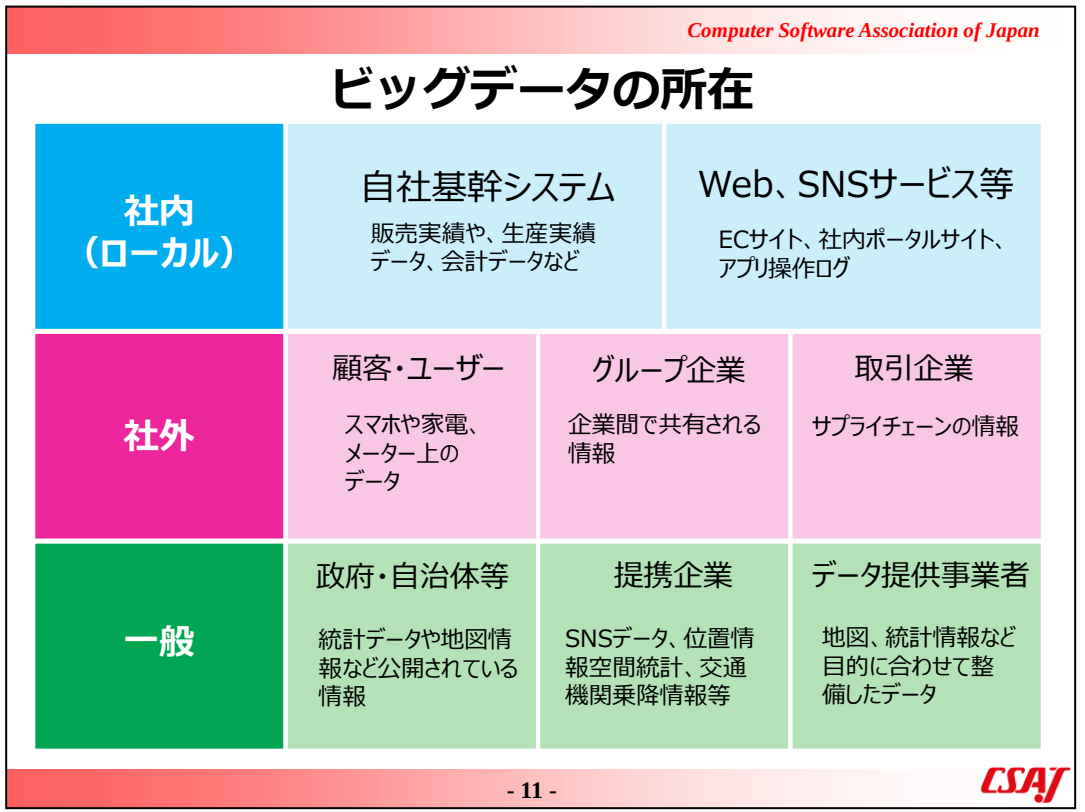
ビッグデータ登場の背景としては、コンピューターの処理能力の向上と、Webシステムの登場が大きいと言えます。

これにより顧客の購買動向や、googleなどの検索内容を補完できるようになり、ビッグデータの道が開けました。

その後の発展は様々な技術により、巨大なデータ処理技術が支えられ続けています。

・ポイント（絶対に覚えてほしいこと、など）

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。



・説明の流れ

ビッグデータはどこにあるものでしょうか。取引先から渡ってくるデータや今まで保存しかしていなかったログデータ、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ます。

・ポイント（絶対に覚えてほしいこと、など）

ビッグデータはあらゆるところに散在している。

・質問（問いかけ）

身の回りや企業内で眠っているビッグデータを上げてみましょう。

Computer Software Association of Japan

オープンデータとは

特徴	■ 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	■ 政府 ■ 地方自治体 ■ 研究機関・大学 ■ 民間企業
具体例	■ 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）

CSAJ

・説明の流れ

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一つです。

具体的な定義は何でしょうか。

・ポイント（絶対に覚えてほしいこと、など）

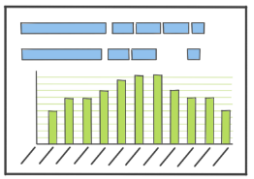
オープンデータは誰でもいつでも利用・再配布できる状態にあるデータ

・質問（問いかけ）

身近なオープンデータをあげてみましょう。

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。

・説明の流れ

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。

・ポイント（絶対に覚えてほしいこと、など）

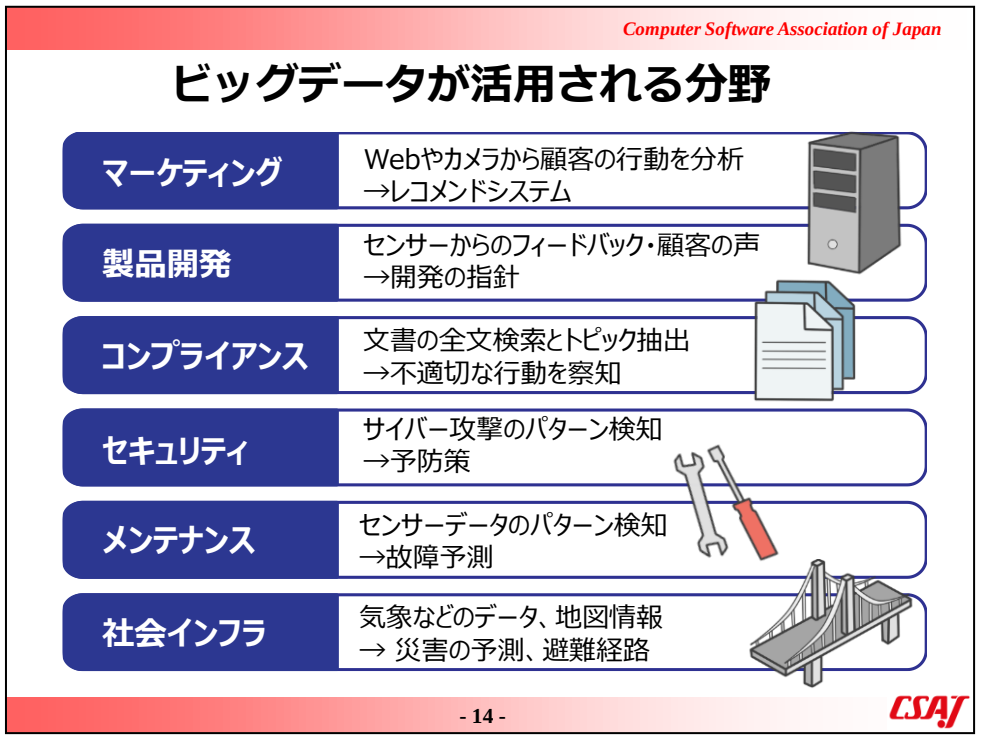
オープンデータ単体では価値を生み出さない場合がある。

・質問（問いかけ）

気象データ単体で利用できるケースとビッグデータを利用した場合の利用ケースをそれぞれ考え、価値に違

いがあるかを考えてみましょう。

・補足説明
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。



・説明の流れ

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。

・ポイント（絶対に覚えてほしいこと、など）

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。

・質問（問いかけ）

今まで耳にしたビッグデータの活用分野を上げてみましょう。

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	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	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	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	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

・説明の流れ

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。

第2章
ビッグデータとセキュリティ

ビッグデータとセキュリティ（個人情報保護）

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って、」「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。）」

・説明の流れ

ビッグデータを扱う際に、個人情報が含まれているため、利用を断念するケースがあります。

それでは個人情報とはそもそもどのような定義でしょうか。

・ポイント（絶対に覚えてほしいこと、など）

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。

・たとえ話、小ネタ

過疎地の郵便番号や、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。

・説明の流れ

日本では個人情報保護法はいつ、どのような内容で施行されたのでしょうか。

・ポイント（絶対に覚えてほしいこと、など）

個人情報の改善措置に従わない場合は、事業者に対して罰則も科される。

Computer Software Association of Japan

個人情報保護法改正

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人分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。

CSAJ

・説明の流れ

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。

・ポイント（絶対に覚えてほしいこと、など）

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。

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。

・説明の流れ

海外の個人情報に対する規制どのようになっているか見てみましょう。

・ポイント（絶対に覚えてほしいこと、など）

海外では個人の意思表示によりデータの削除もできる仕組みであるなど様々な規制が存在する。

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）

・説明の流れ

　海外の個人情報保護の情勢をもう少し詳しく見てみましょう。

・ポイント（絶対に覚えてほしいこと、など）

　EU法のGDPR（General Data Protection Regulation：一般データ保護規則）では、IDなども個人情報扱いであったり、EU域外に個人情報を持ち出せない、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り、注意が必要。

　特にEU国内からアクセスが発生するWebサイトを運営している場合も規定に反する状況が発生している場合も考えられる。

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

・説明の流れ

個人情報の許諾を個人から得る方法にはどのようなものがあるでしょうか。

・ポイント（絶対に覚えてほしいこと、など）

オプトインは事前承認、オプトアウトは事後承認

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）

・説明の流れ

個人を特定できないようにデータを加工する事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か見てみましょう。

・ポイント（絶対に覚えてほしいこと、など）

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ある。

K=XXなどの数値を大きくすると安全性は高まるが、大きくしすぎると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ため注意が必要。

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

・説明の流れ

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することができないようにすることを言います。

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

・説明の流れ

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する事です。

同じ保護属性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。

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

・説明の流れ

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、少なくともk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”良い”多様性を持つことで、属性推定が起こらない状態です。

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。