

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

1-1.ビッグデータとは	5
1-2.ビッグデータの量	6
1-3.ビッグデータの質	7
1-4.ビッグデータの種類	8
1-5.ビッグデータが持つ特性	9
1-6.ビッグデータ登場の背景	10
1-7.ビッグデータの所在	11
1-8.オープンデータとは	12
1-9.オープンデータ×ビッグデータ活用例	13
1-10.ビッグデータが活用される分野	14
1-11.ビッグデータが活用される業界	15

目次

第2章 ビッグデータとセキュリティ

2-1. ビッグデータとセキュリティ（個人情報保護）	17
2-2. 個人情報保護法	18
2-3. 個人情報保護法改正	19
2-4. 個人情報保護の情勢	20
2-5. 個人情報保護法の情勢	21
2-6. オプトインとオプトアウト	22
2-7. 匿名加工処理の手法	23
2-8. 識別子の削除（仮名化）	24
2-9. K-匿名化したテーブル	25
2-10. L-多様化したテーブル	26

第1章

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。

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。

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
<p>分類の意味</p>	<p>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。</p> <p>それほど増加しない見込み。</p> <p>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</p>	<p>完全な構造定義を持たないデータ。</p> <p>例・ログデータ ・センサーデータ ・SNS に書き込まれたデータ</p>	<p>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。</p> <p>例・文書 ・音声 ・動画 ・画像</p>
<p>前のページの例を当てはめると…</p>	<p>販売 POS データ</p> <p>販売・生産実績</p>	<p>ツイートデータ</p> <p>Web 操作ログ</p>	<p>防犯カメラ映像</p> <p>コールセンター音声</p>

ビッグデータが持つ特性

ビッグデータの3V

Velocity
生成頻度

データの生成、処理の速度が速い。
リアルタイム性が求められ、
センサーや Web などから
常にデータが発生している。

Variety
多様性

データの質が多岐に渡っている。
構造化データのみならず、
音声、動画といった
非構造化データも含まれる。

Volume
量

データ量が圧倒的に多い。
大量に爆発的に増加し、
予測は不可能である。

Veracity
正確性

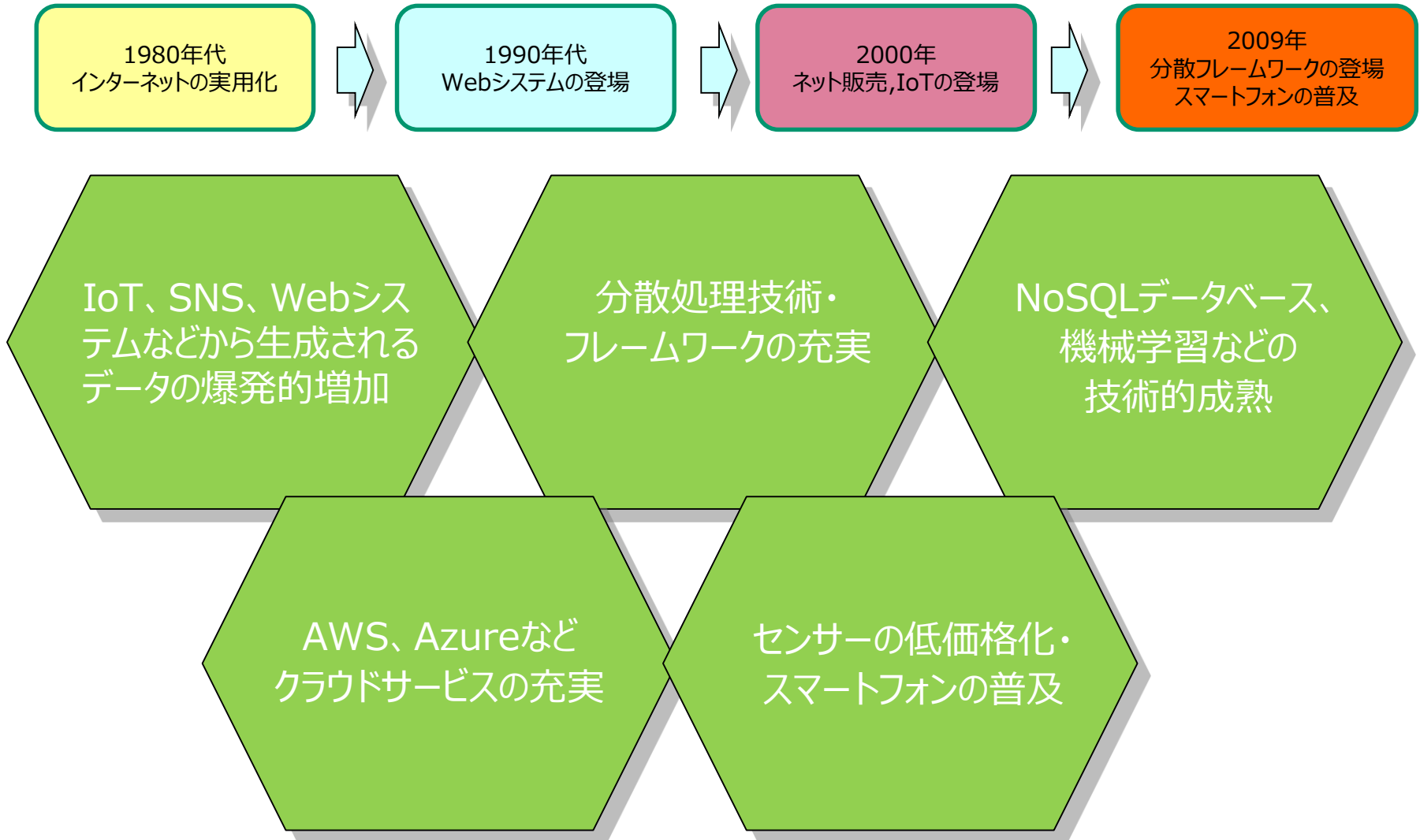
センサーデータやユーザー
コンテンツにより正確性が
求められるようになった。

Value
価値

大量にあるデータを
組み合わせて活用することで、
新たな価値が生まれる。

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

ビッグデータ登場の背景



ビッグデータの所在

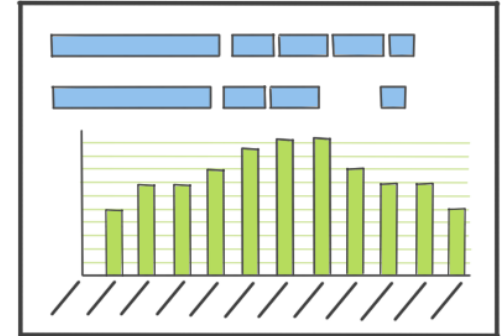
<p>社内 (ローカル)</p>	<p>自社基幹システム 販売実績や、生産実績データ、会計データなど</p>		<p>Web、SNSサービス等 ECサイト、社内ポータルサイト、アプリ操作ログ</p>
<p>社外</p>	<p>顧客・ユーザー スマホや家電、メーター上のデータ</p>	<p>グループ企業 企業間で共有される情報</p>	<p>取引企業 サプライチェーンの情報</p>
<p>一般</p>	<p>政府・自治体等 統計データや地図情報など公開されている情報</p>	<p>提携企業 SNSデータ、位置情報空間統計、交通機関乗降情報等</p>	<p>データ提供事業者 地図、統計情報など目的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

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）」

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。



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

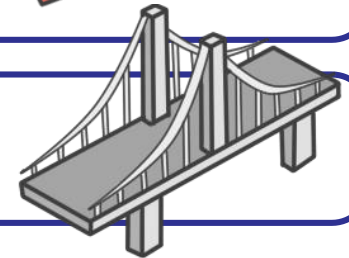
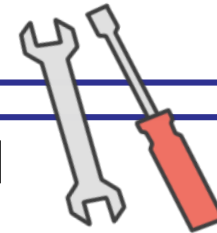
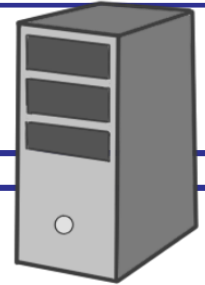
サイバー攻撃のパターン検知
→予防策

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

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。）」

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。

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。

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）

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

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）

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

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

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