Big Data Working Group

Identifying Scalable Techniques for Data-Centric Security and Privacy Problems in the Cloud

CSA JC cloud security alliance SM Japan Chapter

Big Dataのプライバシ

November 22, 2016
Tadashi Onodera
Big Data User WG

目次

- Big Dataの振り返り
- Big Dataのプライバシの動向
- Big Data WG策定文書のご紹介
- WG活動のご案内

資料中の意見はWG見解であり、講演者の属する法人等の公式見解ではありません。

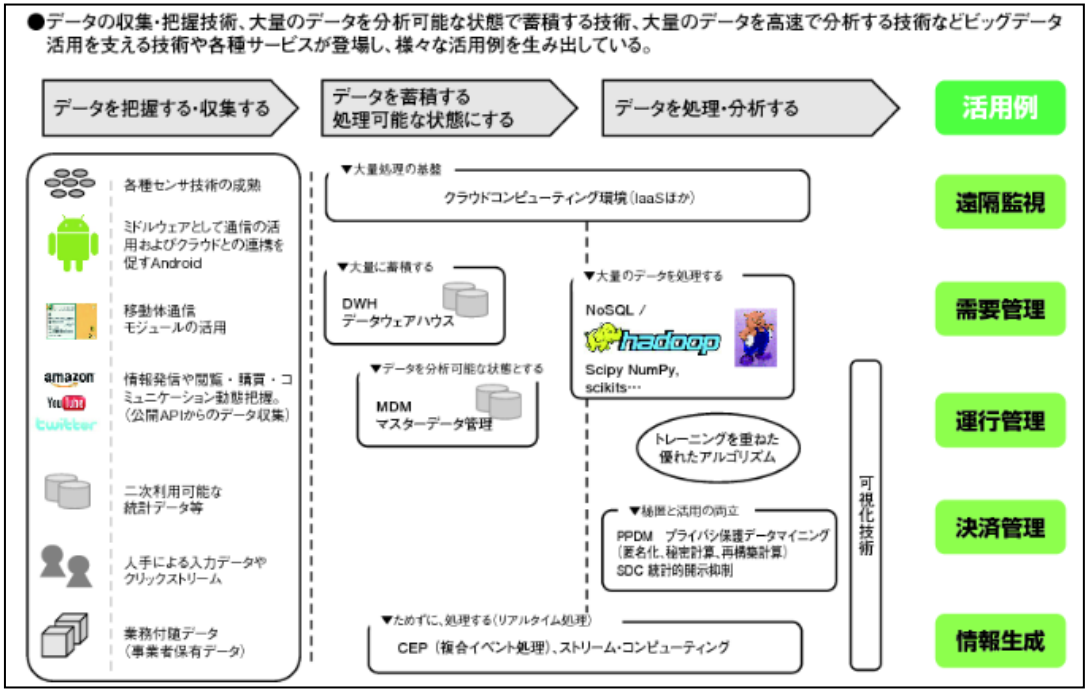
Big Dataの振り返り

ビッグデータの経済効果は大きい

- Big data:The next frontier for innovation,competition,and productivity(2011)
  - ビッグデータの活用で特に成長が見込まれる5つの部門を対象に、ビッグデータ活用により発現する経済効果・便益について推計を行った。
    - 米国のヘルスケア産業は3,333億ドルの経済効果
    - 欧州の公共事業では最大3,000億ドルの経済効果
    - 米国の小売業では生産性0.5%増加、売上純利益は60%以上増加
    - 製造業では開発コストが25%減少、製品の市場投入までの期間が20%～50%短縮化、利益マージンが2%～3%増加、オペレーションコストが10%～25%削減、そして7%の収入増
    - 位置情報サービス分野では2020年までに累計7,000億ドルから8,200億ドルの経済効果
- ビッグデータへの注目が高まった

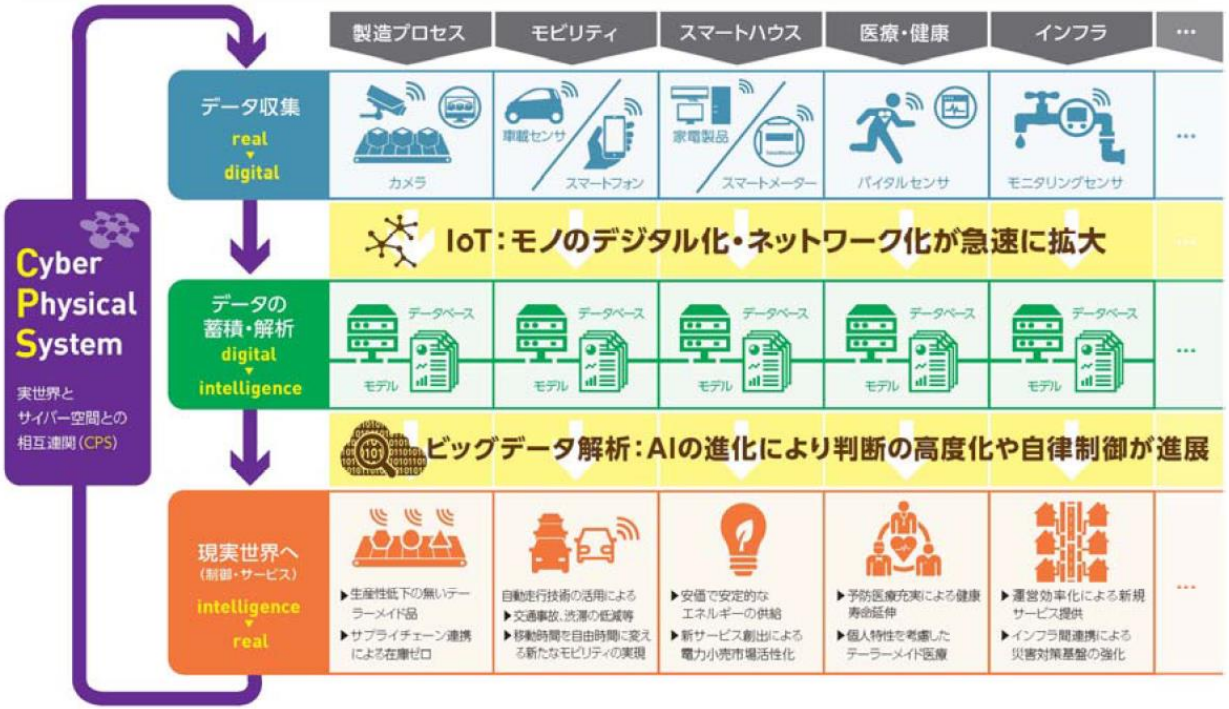
ビッグデータに多様なデータが集まる

- 平成24年度　情報通信白書
  - ビッグデータの定義：利用者の商品・デジタルコンテンツ等の購買履歴や決済情報、コミュニケーションの発信履歴など膨大なデータを活用し、サービス革新を行う
  - 主要プレーヤー：米国のネット系プラットフォーマー
  - 今後はM2M等のセンサネットワークなどの多様なデータが使われる事が予測されていた



IoT化によりBig Dataに集まるデータソースが増える

- 平成27年4月　産業構造審議会
  - 「CPSによるデータ駆動型社会の到来を見据えた変革」と称し、リアルとデジタル空間が相互に連動するようになってきたことを示す

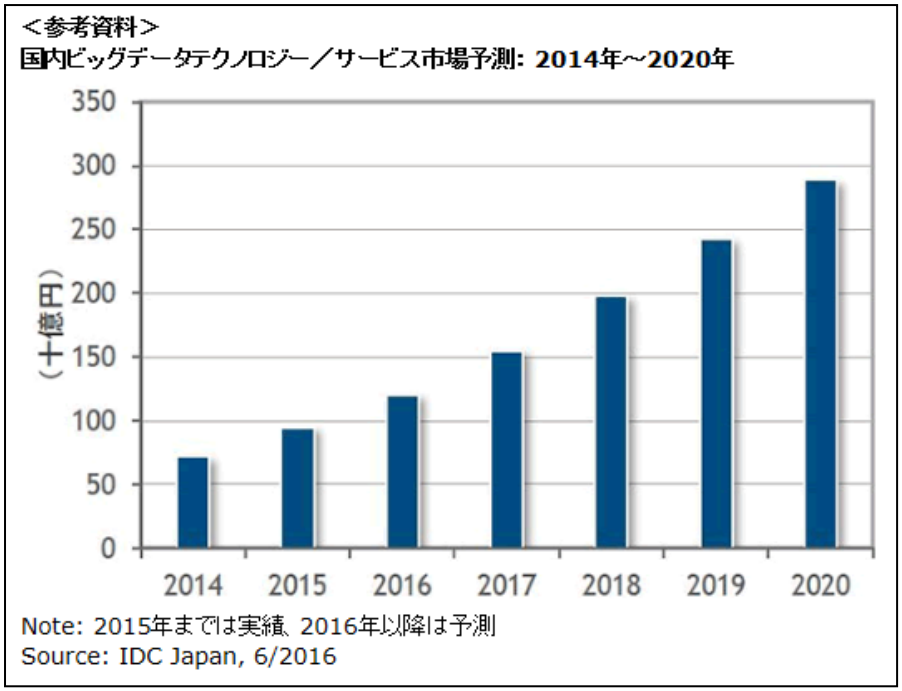


データ総量・トラフィック・クラウド保管データ量も増える

- The Digital Universe of Opportunities（2014年）
  - デジタルデータ総量
    - 2013年：4.4 ゼタバイト
    - 2020年：44ゼタバイト（44兆ギガバイト）
- Cisco Global Cloud Index（2015）
  - クラウドのネットワークトラフィック量
    - 2014年：2.1 ゼタバイト
    - 2019年：8.6ゼタバイト
  - パーソナル／ビジネス向けクラウド、M2M接続の増加
  - 2019年には全データ総量の51%がPC以外のデバイスで管理されると予想

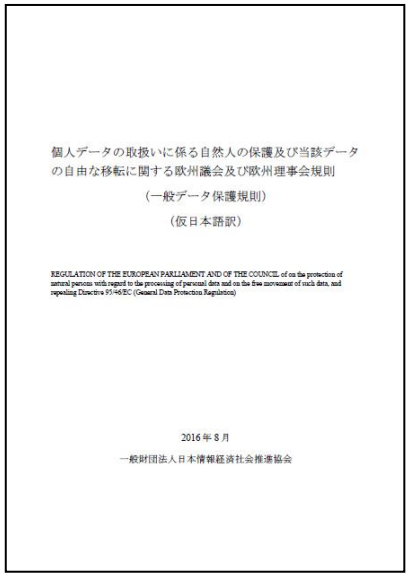
ビッグデータを支えるテクノロジー市場も成長

- 2015年 国内ビッグデータテクノロジー／サービス市場規模(2016)
  - インフラストラクチャ、ソフトウェア、サービスの3つの市場セグメントに分類し、分析
  - 2020年に約2,889億円に到達、2015-2020年のCAGRは25.0%を試算



Big Dataの
プライバシの動向

集めた情報のプライバシ配慮が必要である



個人データの取扱いに係る自然人の保護及び当該データの自由な移転に関する欧州議会及び欧州理事会規則

（一般データ保護規則）

（仮日本語訳）

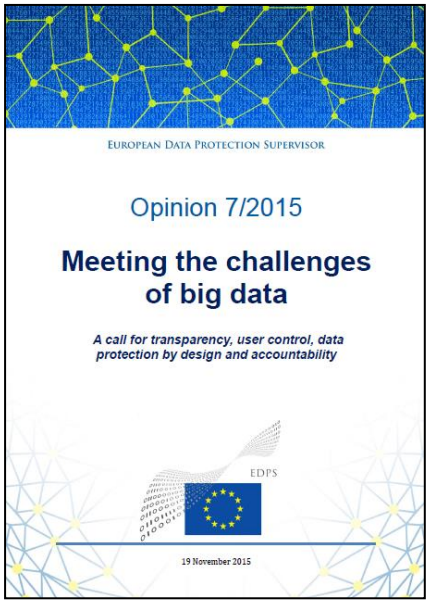
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

2016年5月

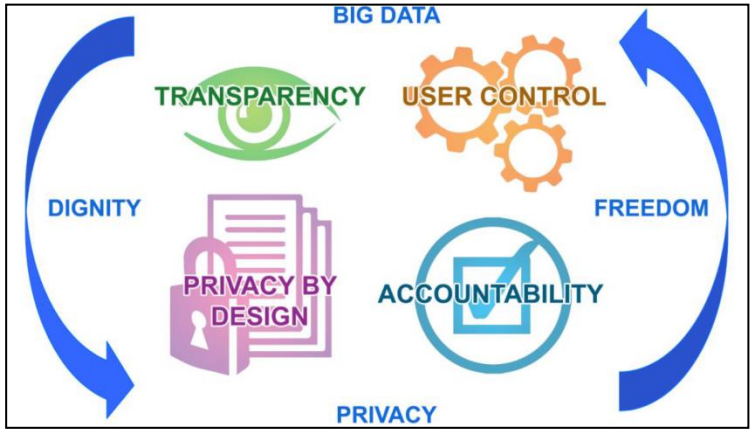
一般財団法人日本情報経済社会推進協会

- EUデータ保護規則(2016)
  - EUデータ保護指令から規則に改正
  - 欧州議会、閣僚理事会、欧州委員会が合意、採択
  - 2018年5月までに各国で適用
- 日本企業に求められる対応
  - 処理対象の個人データおよび処理過程を特定
  - 適切な安全対策を実施
  - EU域外へのデータ移転にあたり、適切な方法を選択し運用を実施
  - インシデント発生時には、データ主体および監督機関に通知
  - データ保護影響評価を実施し、必要に応じて監督機関に通知
  - など
- EU市場で事業を展開する日本企業にも影響
- 違反した場合、最大で2,000万ユーロまたは前年度の全世界連結売上の4%のいずれか高い方を制裁金とする

データ活用とプライバシのバランスが必要



- European Data Protection Supervisor 「Meeting the challenges of big data」(2015)
  - ビッグデータによるイノベーションの促進とプライバシーに係る基本的な権利の保護はバランスが必要
  - ビッグデータの責任ある持続的発展に不可欠な要素
    - 透明性：組織は、個人データをどのように処理しているかについて、透明性を高めるべきである
    - ユーザコントロール：自己のデータがどのように利用されるかについて、ユーザに高いレベルのコントロールを与える
    - プライバシー・バイ・デザイン：製品やサービスの中にユーザ本位のデータ保護を設計する
    - 責任：何をするかについて一層の責任を持つ



データ活用とプライバシのバランスが必要

- Privacy by Design in Big Data(2015)
  - Big Data対PrivacyからBig Data with Privacyへのシフト
  - ビッグデータのプライバシー課題
    - 制御と透明性の欠如
    - データの再利用性
    - データの推定と再特定
    - プロファイリングと自動化された意思決定
  - ビッグデータのバリューチェーンのフェーズ毎に具体的な推奨事項を整理

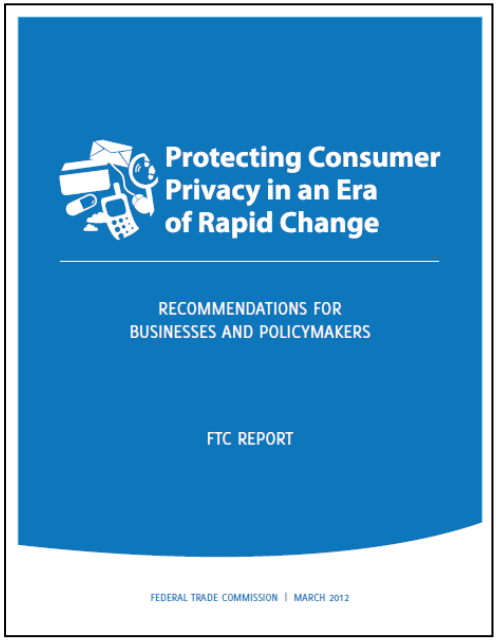
ビッグデータのバリューチェーン	プライバシー・バイ・デザイン戦略	実行内容（例）
データ取得／収集	最小化する	収集前に何のデータが必要かを定義して、収集前に選択し（例：データフィールドの削減、適切な制御の定義、不必要な情報の削除）、プライバシー影響度を評価する
	集約する	ローカルの匿名化（ソース元で）
	隠す	プライバシーを強化するエンドユーザーツール（例：追跡防止ツール、暗号化ツール、アイデンティティマスキングツール、セキュアなファイル共有）
	知らせる	個人への適切な通知の提供――透明性のメカニズム
	制御する	同意取得を示す適切なメカニズム、プライバシー選択を示すメカニズム、粘り強いポリシー、個人データ保存
データ分析&データキュレーション	集約する	匿名化技術（例：k-匿名性、差分プライバシー）
	隠す	検索可能な暗号化、プライバシー保護計算処理
データストレージ	隠す	停止状態のデータ暗号化、認証とアクセス制御のメカニズム、その他セキュアなデータストレージのための手法
	分離する	分散／非集中型ストレージと分析機能
データ利用	集約する	匿名化技術、データ品質、データ来歴
全フェーズ	執行／証明する	自動化されたポリシー定義、執行、責任追跡／コンプライアンスツール

「匿名化」でプライバシを守る

- 個人が再識別できず、個人に関する情報を知ることができないような方法で、個人データを修正するプロセス
- データセットの有用性を損なうこと無しに利用できる、完璧な匿名化技術を実現する事は困難
  - Privacy by Design in Big Dataで整理された匿名化技術

視点	匿名化技術
制御されたリンク可能性	一定のリンク可能性を許容しながら、再識別化と属性開示を防止するための技術
構成可能性	複数ソースのリンクにより構築されたデータセットのプライバシーを保証する匿名化プライバシーモデル（例：k-匿名化、差分プライバシー）
動的／ストリーミングデータの匿名化	有用性と開示のリスクの制御
大容量データの計算処理能力	選択したプライバシーモデルや匿名化手法に対応する計算処理効率
分散型匿名化	個々の計算処理デバイスを利用した、ソースレベルでの匿名化（例：局所的匿名化、コラボレーティブ匿名化）

米国での匿名化の要求（FTC）



- Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers(2016)
  - 「プライバシー・バイ・デザイン」「消費者への簡潔な選択肢の提供」「透明性の確保」を提唱
    - 事業者は、そのデータの非識別化を確保するために合理的な措置を講ずるべきである
    - 事業者は、そのデータを非識別化された形態で保有および利用し、そのデータの再識別化を試みないことを、公に約束すべきである
    - 事業者がかかる非識別化されたデータを他の事業者に提供する場合には、それがサービス提供事業者であろうとその他の第三者であろうと、その事業者がデータの再識別化を試みることを契約で禁止すべきである
  - コンテンツ配信会社が、非識別化したユーザの視聴履歴データを提供し、映画推薦アルゴリズムの開発コンテストを行ったところ、他社が運営するユーザレビューと紐付けることでユーザ情報の一部を再識別化出来てしまったため、FTCが指摘、その結果コンテストは中止

非構造化データの非識別化・匿名化とAI


NISTIR 8053

De-Identification of Personal Information

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

Research and applications

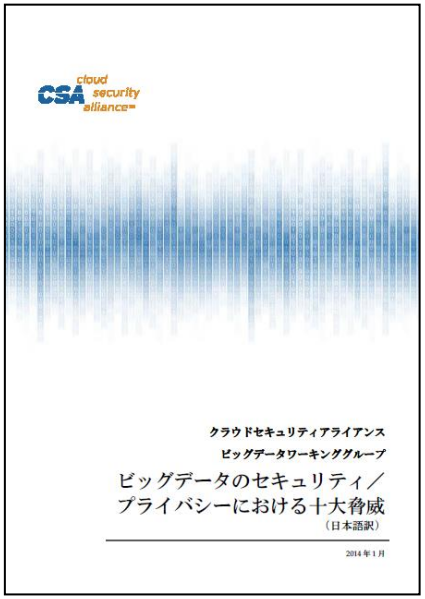
OPEN ACCESS

Large-scale evaluation of automated clinical note de-identification and its impact on information extraction

- 米国NISTの個人情報非識別化ガイドライン「NISTIR 8053 De-Identification of Personal Information」(2015年12月)
  (http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf)
  - **非構造化データの非識別化・匿名化における課題**
    - 直接識別子が必ずしも明確化されていないケースがある
      （例．マルチメディア・コンテンツの動画・音声）
    - ユーザーが、プライバシーに関わる情報を追加したり、削除したりするケースがある
      （例．電子カルテの医師コメント、添付画像など）
    - 非識別化して法令要件をクリアした個人データが、後から再識別化できてしまうケースがある
      （例．医薬品副作用報告データを利用した2次解析研究）
  - （例）医療分野のAI（例．機械学習）を利用した非識別化システムに関する研究
    「Large-scale evaluation of automated clinical note de-identification and its impact on information extraction.」
    J Am Med Inform Assoc. 2013 Jan 1;20(1):84-94. doi: 10.1136/amiajnl-2012-001012. Epub 2012 Aug 2.
    (https://www.ncbi.nlm.nih.gov/pubmed/22859645)

Big Data WG策定文書のご紹介

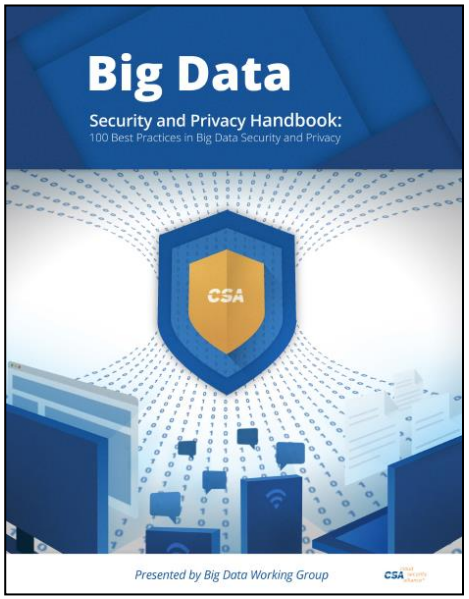
ビッグデータのセキュリティ/プライバシ課題



- Top 10 Big Data Security and Privacy Challenges(2012)
  - 1. 分散プログラミングフレームワークにおけるセキュアな計算処理
  - 2. ノンリレーショナルデータストアのセキュリティのベストプラクティス
  - 3. セキュアなデータ保存とトランザクションのログ
  - 4. エンドポイントの入力の検証／フィルタリング
  - 5. リアルタイムのセキュリティ／コンプライアンスモニタリング
  - 6. 拡張性があり構成可能なプライバシー保護データマイニング／分析
  - 7. 暗号化により強制されたアクセス制御とセキュアな通信
  - 8. 詳細なアクセス制御
  - 9. 詳細な監査
  - 10. データ来歴

日本語訳がダウンロード出来ます。

ビッグデータのセキュリティ/プライバシ対応



- Big Data Security and Privacy Handbook: 100 Best Practices in Big Data Security and Privacy(2016)
  - 「ビッグデータのセキュリティ／プライバシーにおける十大脅威」で提示した課題への対応方法を紹介
  - 例：“6. 拡張性があり構成可能なプライバシー保護データマイニング／分析”の対応
    - 6.1 Implement differential privacy
    - 6.2 Utilize homomorphic encryption
    - 6.3 Maintain Software infrastructure
    - 6.4 Use separation of duty principle
    - 6.5 Be aware of re-identification techniques
    - 6.6 Incorporate awareness training with focus on privacy regulations
    - 6.7 Use authorization mechanisms
    - 6.8 Encrypt data at rest
    - 6.9 Implement privacy-preserving data composition
    - 6.10 Design and implement linking anonymized datastores

日本語訳の作成を検討中です。

WG活動のご紹介

Big Dataのセキュリティ／プライバシの検討

- CSAグローバルのBig Data WGと協調する方針
  - Big Data Security and Privacy Handbook: 100 Best Practices in Big Data Security and Privacyの翻訳を検討
- Big Dataのセキュリティ／プライバシのベストプラクティスを整理し、グローバルに展開
  - セキュリティ・バイ・デザイン
  - プライバシ・バイ・デザイン
  - 匿名化技術、再識別化技術
  - 機械学習・深層学習
  - 利便性とプライバシのバランスが取れている必要があるため、Big Dataによるデータ活用についても調査し、事例集などを作成予定
- 他WGとの連携
  - IoT WG、SLAイノベーションWGなど

ありがとうございました