総務省　ICTスキル総合習得教材

【概要版】eラーニング用

［コース3］データ分析

3-1：ビッグデータの活用と分析に至るプロセス

	1	2	3	4	5
［コース1］データ収集					
［コース2］データ蓄積					
［コース3］データ分析	◆				
［コース4］データ利活用					

本講座の学習内容（3-1：ビッグデータの活用と分析に至るプロセス）

【講座概要】

- 日本企業におけるデータ分析の実態に加えて、より良いデータ分析の設計を示します。
- ビッグデータが持ち得る特性として3つのVを紹介します。
- データの形態としての構造化データ、半構造化データ、非構造化データを説明します。
- データの品質の説明に加えて、データの標準化、データクレンジングの重要性を示します。

【講座構成】

座学

- [1]データ分析の実態と設計
- [2]ビッグデータの特徴と利用事例
- [3]データの品質と標準化・クレンジングの重要性

【学習のゴール】

- ✓ 日本企業におけるデータ分析の実態、より良いデータ分析の設計を把握する。
- ✓ ビッグデータが持ち得る3つのVを理解する。
- ✓ 構造化データ、半構造化データ、非構造化データを区別できるようになる。
- ✓ データの標準化、データクレンジングの重要性が説明できるようになる。

蓄積されたデータの分析

◆このコースでは、蓄積されたデータを利用する方法の一つとして「データ分析」を学びます。

●クラウド等のサーバに蓄積されたデータの利用方法には大別して2種類あり、一つは「データベースとしての利用」、もう一つは「分析用データとしての利用」です。

- 講座2-1で示したように「データベース」の要件として、個々のデータレコードを「検索ができること」が挙げられます。

●データベースとしての利用では、検索によって抽出された「個々のデータレコード」に注目する一方で、分析用データとしての利用では、「データ全体または一部の傾向や特徴」に注目します。

- 「データベースとしての利用」は、個々のデータレコードを抽出できることで、「カタログ、データレコード別の情報サービス」として利用できます。

●データを分析することで、データ全体または一部に関する傾向や特徴の発見することができます。

- データの特徴や傾向を発見、把握することで、未知の情報を予測できるケースもあります。

「データベースとしての利用」と「分析用データとしての利用」

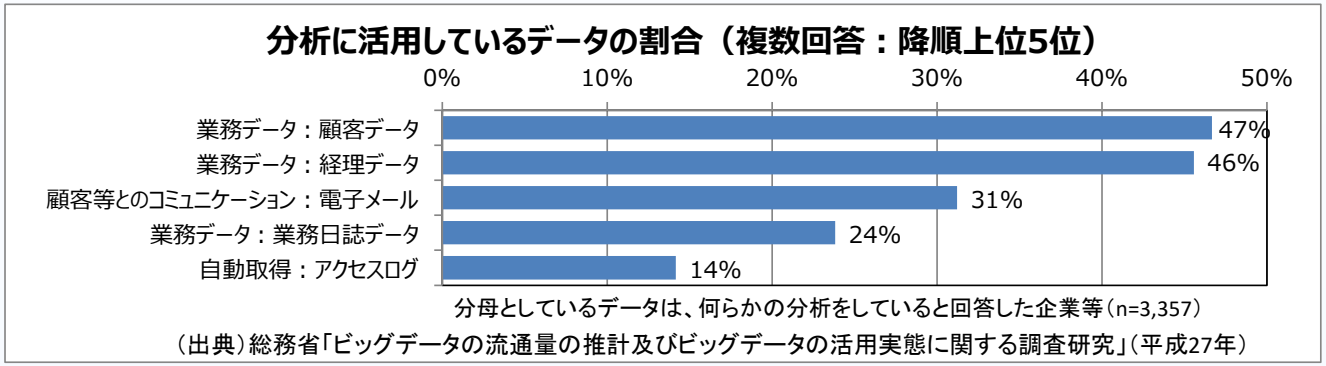
利用方法	注目対象	利用事例	天気データでの利用例
[1] データベースとしての利用（検索による抽出）	個々のデータレコード	カタログ、データレコードの情報利用	特定の場所、時点に関する天気情報の検索と抽出
[2] 分析用データとしての利用	データ全体または一部の傾向・特徴	傾向・特徴の発見、未知の情報の予測	天気の地域性・季節性の発見、天気予報

□ このコースでは、蓄積されたデータから「傾向・特徴の発見する」ためのデータ分析を学びます。

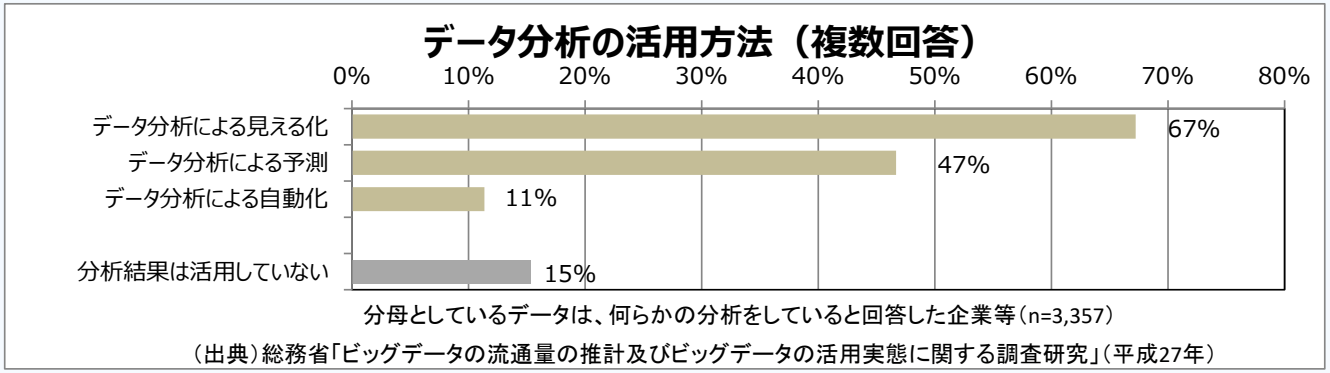
日本企業におけるデータの分析の実態

◆日本企業では「業務データ」を「見える化」する形のデータ分析が、最も多くなっています。

- 総務省の委託調査によれば、日本企業4,672社のうち72%の3,357社がデータを分析しています。
- 日本企業において、分析に活用しているデータとして「顧客データ」、「経理データ」の割合が高くなっています。



- データ分析の活用方法として、最も割合が高いのは「データ分析による見える化（可視化）」の67%です。
  - 「見える化」や「可視化」とは、図表作成などを行うことでデータを分かりやすく示すことを指しています。



□ データ分析と言っても、特別に入手したデータを使って高度な分析を行う必要はなく、 自然に集まる業務データを活用し、見える化（可視化）することで分かりやすく表すことが第一歩です。

より良いデータ分析の設計

◆より良いデータ分析の設計として、まず「何をしたいのか？」を明確にすることが挙げられます。

- ●私達はビジネスにおいても、私生活においても、様々な目的があり、それに対する意思決定（選択）をしています。
- ●データ分析を行うことで、目的に対して、より効果的な意思決定（選択）を行うことができます。
  - 必ずしも自分自身でデータ分析を行う必要はなく、データ分析を依頼することも、公表されている分析結果のみを確認することもあります。
  - データ分析を行わない人や場合においても、まずは定量的なデータや指標を確認する姿勢が重要です。

ビジネスにおける目的例：売上総額を上げたい

- ◆ 売上総額は［販売単価］×［販売個数］で構成されている。
- ◆ ［販売単価］は企業が決められるが、［販売単価］を上げれば［販売個数］は下がる関係にある。

⇨ **データ分析をすることで、売上総額を最大化するための［販売単価］が分かる。**

私生活における目的例：ダイエット（減量）したい

- ◆ ダイエットには［食事制限］と［運動］の両方に効果があるとされている。
- ◆ ［食事制限］と［運動］をどのように組み合わせることが、ダイエットに効果的かがはっきりと分からない。

⇨ **データ分析をすることで、ダイエットに効果的な［食事制限］と［運動］の組み合わせが分かる。**

- ●あらかじめ「何をしたいのか？」や「何を知りたいのか？」を明確にすることで、意思決定（選択）に反映できるデータ分析の方針を定められるとともに、効率的に分析作業ができます。
  - データが手元にありつつも、データ分析の目的を明確にしにくいケースにおいては、見える化（可視化）によってデータを分かりやすく図表に表すことによって定期的に実態を把握できたり、より高度な分析へのヒントが得られるケースもあります。

ビッグデータの3つのV

◆ビッグデータが持つ特徴として「3つのV」が挙げられることが一般的です。

●IoT等によって記録され、インターネット等を通じて収集された多様かつ膨大なデータは、「ビッグデータ」と呼ばれることがあります。

●ビッグデータには、その特徴とされる3つのVの「Variety（バラエティ）」「Volume（ボリューム）」「Velocity（ベロシティ）」のいずれかをも持っていることが挙げられます。

- 「3つのV」は2001年にアメリカのデータ分析者によって提示され、現在でも一般的な考え方となっています。
- いずれかの特徴を持っていれば、ビッグデータと言えるため、これらの特徴を併せ持つ必要はありません。

ビッグデータの「3つのV」

V	日本語訳	意味
Variety	データの多様性	テキスト、画像、音声といった多様な情報とファイル形式
Volume	データ量	膨大なデータ量
Velocity	データ生成速度・頻度	リアルタイムで収集できるデータ・秒単位など高頻度のデータ

●ビッグデータの持つ特性（3つのV）を活かすことで、新たな分析、サービスができるようになりました。

- Yahoo! Japanでは、ビッグデータのデータ量（Volume）を活かし、時期別、都道府県別のインフルエンザに関する検索数から、感染数の予測値を示しています。
- ネットショッピングのサイトにおいては、ビッグデータのデータ生成速度（Velocity）を活かし、購入した商品から即座に関連商品を推薦します。

構造化データ、非構造化データ

◆ビッグデータは、データ整理がしやすい構造化データではないケースもあります。

- ビッグデータはその特性である多様性（Variety）から「半構造化データ」「非構造化データ」のケースもあります。

「構造化データ」「半構造化データ」「非構造化データ」の説明

データ種別	説明	データ形式の例
構造化データ	二次元の表形式（Excel形式）になっているか、データの一部を見ただけで二次元の表形式への変換可能性、変換方法が分かるデータ	CSV、固定長、Excel（リレーショナルデータベース形式）
半構造化データ	データ内に規則性に関する区切りはあるものの、データの一部を見ただけでは、二次元の表形式（Excel形式）への変換可能性、変換方法が分からないデータ	XML、json
非構造化データ	データ内に規則性に関する区切りがなく、データ（の一部）を見ただけで、二次元の表形式に変換できないことが分かるデータ	規則性に関する区切りのないテキスト、PDF、音声、画像、動画

二次元の表形式の構造化データ

世帯名	大人1	大人2	子供1
山田家	世帯主	妻	長女

XML形式の半構造化データ

```
<世帯>
  <世帯名>山田家</世帯名>
  <大人>世帯主</大人>
  <大人>妻</大人>
  <子供>長女</子供>
</世帯>
```

画像形式の非構造化データ

- 一般に半構造化データ、非構造化データは分析を行う前にデータの整理や変換が必要です。

データの品質と質の悪いデータによるコスト

◆データには品質があり、データの品質が悪ければ社会的、経済的なコストを生みます。

- ビッグデータには、重複するデータレコード、表記揺れがあるなど、データの品質が悪いケースもあります。
- 国際データマネジメント協会の英国支部によれば、データの品質には6つの主要基準があるとしています。

DAMA UKによるデータの品質に関する6つの主要基準

基準	説明	品質が損なわれている例
Completeness（網羅性）	保存されているデータの割合は、潜在的な全データに対して「100%網羅」していること	部分的なデータ
Uniqueness（唯一性）	特定された対象が、2行以上にわたって記録されていないこと	重複するデータレコード
Timeliness（適時性）	要求する時点の現実を表している程度	速報性がない調査データ、低頻度の調査データ【利用者のニーズに依存】
Validity（正当性）	定義されている構文規則（フォーマット、型、範囲）に正しく準拠していること	表記揺れ、誤記入、数値が入るべきデータ項目へのテキストの記入
Accuracy（正確性）	記述している現実世界の対象やイベントを正確に表している程度	測定誤差の大きいレコード
Consistency（一貫性）	データセット内、データセット間で一つの定義に対して、複数の表現等の相異がないこと	データセット間の「西暦と和暦」の混在【他のデータセットとの関係に依存】

【出典】the six primary dimensions for data quality assessment（DAMA UK）

- このデータの品質基準には、客観的でデータ固有の基準のみではなく、利用者の主観的な有用度合いに依存する「Timeliness（適時性）」、他のデータとの照合しやすさとして「Consistency（一貫性）」が含まれていることが特徴的です。

- データの品質が悪ければ、分析の実施が不可能になったり、分析前に大きな時間や費用がかかる等の様々なコストを生みます。
- 2016年にIBM 社より公刊された書籍では、「質の悪いデータが、アメリカ経済に与えているコストの推定値は、年間3.1兆ドル」と紹介しています。【出典】Data Engine for Hadoop and Spark（IBM）

データの標準化とデータクレンジングの重要性

◆ビッグデータの活用には「データの標準化」や「データクレンジング」が必要です。

- 2015年に総務省統計委員会から公表された報告書では、ビッグデータ活用における課題として、「データクレンジング技術の高度化、企業・業界横断的にデータ形式の標準化」を挙げています。

【出典】公的統計におけるビッグ・データの活用に関する調査研究（総務省）
http://www.soumu.go.jp/main_content/000422923.pdf

- データ形式を標準化して一貫性を高めることで、データ分析をはじめとする利活用をスムーズに行うことができます。

住所の表記揺れ	
霞ヶ関1丁目1番地	霞が関1丁目1番地
霞ヶ関1丁目1	霞が関1丁目1
霞ヶ関1-1	霞が関1-1

住所表記の標準化の取り組み例

- 標準記載法の策定と公表
- 標準記載名データベースの公表
- 表記揺れの統一エンジンの公開

住所表記の標準化例

霞が関１丁目１
1-1 Kasumigaseki
〒100-0013
緯度: 35.675836 経度: 139.754734

- 住所の表記においては、「ヶ」と「が」の混在、丁番地の表記が不統一となっていることで、一貫性が損なわれています。
- データ形式を標準化しておくことで、様々なIoT機器等で収集したデータを統合して分析することができます。

- 品質が一部損なわれたデータであっても、「データクレンジング」することで分析をはじめとする利活用が可能です。
  - 収集したデータを利用しやすい形に整えたり、表記揺れの統一、異常値を取り除いたりして、分析をはじめとする利活用に適した状態にすることをデータクレンジング（データの前処理）といいます。
  - 根本療法としての「データの標準化」の推進が重要である一方で、対処療法としての「データクレンジング」技術を身につけておくことも必要です。
  - 品質の良いデータであっても、利用者の利活用に適する形への整理や変換として、「データクレンジング」が必要となるケースがあります。

□ 次の講座3-2では、Microsoft Excelを使ったデータクレンジングの基本技術、可視化の手順を紹介します。