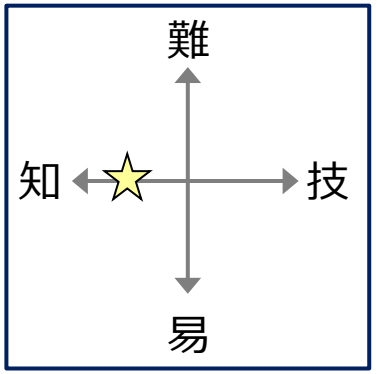
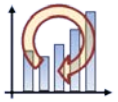
総務省　ICTスキル総合習得教材



［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

http://www.soumu.go.jp/ict_skill/pdf/ict_skill_3_1.pdf

	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】

- ビッグデータの特性の「3つのV」を説明し、それぞれの特性によって可能になる分析を示します。
- データの品質のいくつかの観点から紹介し、品質の悪いデータがもたらす社会的費用を紹介します。
- データ形式の標準化およびデータクレンジングの重要性を示します。
- 国内企業におけるデータ分析の実態、効率的なデータ分析の設計、本格的なデータ分析に至るプロセス（工程）を紹介します。

【講座構成】

座学

- [1] ビッグデータの特性と分析
- [2] データの品質と標準化・クレンジング
- [3] データ分析の設計と分析に至るプロセス

【学習のゴール】

- ✓ ビッグデータの「3つのＶ」と、各特性によって可能になる分析事例を紹介できる。
- ✓ 品質の悪いデータの社会的費用とその軽減策としてのデータ形式の標準化、データクレンジングの重要性を理解する。
- ✓ 効率的なデータ分析の設計と本格的なデータ分析に至るプロセスを説明できる。

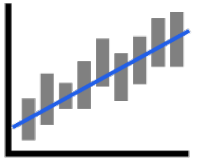
データ利用方法としての分析

◆この講座では、データの利用方法の一つとしての「データ分析」と関連事項を概説します。

- 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」に注目します。
  - 「データベースとしての利用」は、例えば、個々のデータレコードを抽出して「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。

➢ **「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。



2種類のデータの利用方法の要点・天気データでの利用例

利用方法	注目対象	利用事例	天気データでの利用例
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
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

□ この講座では、データ分析の序論として、データの種類、品質、望ましい分析の設計を紹介します。

ビッグデータ

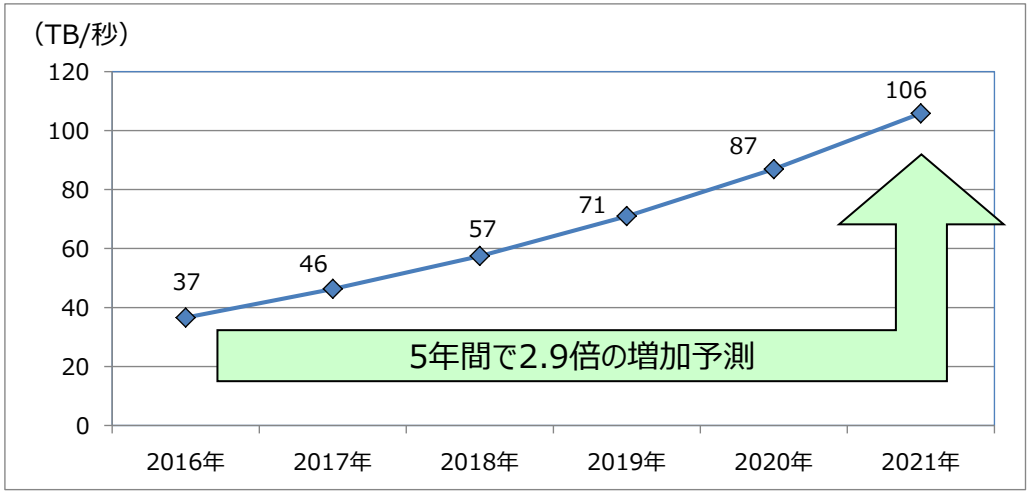
◆インターネット関連のデータは、その特性によって「ビッグデータ」と呼ばれることもあります。

- PC・スマートフォンをはじめとして、インターネットにつながる機器であるIoT機器が増加しています。
- SNS等の普及によって、一般利用者がプログラム不要で様々なデータをインターネット上に保存できるようになりました。
  - 数値データ・テキストデータのみならず、画像や動画といった容量の大きいマルチメディアデータの送受信が拡大してきています。
- 様々な機能・活動によって蓄積された膨大なデータは、その特性に応じて**ビッグデータ**と呼ばれることがあります。
- 2017年6月にシスコ社から公表された資料によれば、全世界のインターネットにおいて送受信されたデータ量は、2016年において年間1.2ZB（ゼタバイト）でしたが、2021年には3.3ZBに達すると予測されています。
  - データのサイズは、byte（バイト）から1000倍ごとにkB、MB、GB、TB、PB、EB、ZBへ単位が変わります。
  - 1年間を365日（31,536,000秒）と見なせば、年間1.2ZBのデータ通信量は、1秒間に約37TB（36,550GB）に相当します。

データサイズの単位

単位	読み方	数値表記（バイト）	
B	バイト	1	バイト
kB	キロバイト	1,000	千バイト
MB	メガバイト	1,000,000	百万バイト
GB	ギガバイト	1,000,000,000	十億バイト
TB	テラバイト	1,000,000,000,000	兆バイト
PB	ペタバイト	1,000,000,000,000,000	千兆バイト
EB	エクサバイト	1,000,000,000,000,000,000	百京バイト
ZB	ゼタバイト	1,000,000,000,000,000,000,000	十垓バイト

インターネット上で1秒間に送受信されるデータ量（予測値）



【出所】Cisco Visual Networking Index：予測と方法論［Cisco］に基づき作成
https://www.cisco.com/c/ja_jp/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

ビッグデータの特性

◆ビッグデータが持ち得る特性として「3つのV」が挙げられます。

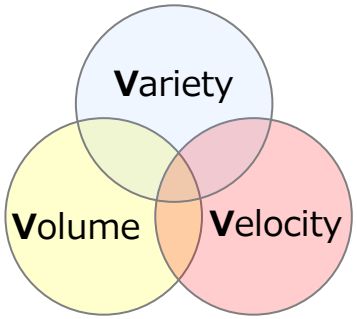
- ビッグデータの持ち得る標準的な特性としての**「3つのV」**は、2001年にアメリカのデータ分析者によって提示され、現在でもビッグデータに関する標準的な考え方となっています。

【出所】Deja VVVu: Others Claiming Gartner's Construct for Big Data［Gartner | Doug Laney］
https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

- ビッグデータには、その特性とされる「3つのV」の**Variety（バラエティ）**、**Volume（ボリューム）**、**Velocity（ベロシティ）**のいずれかを持っていることが挙げられます。
  - 「Variety」「Volume」「Velocity」のそれぞれの頭文字の「V」から「3つのV」と呼ばれます。

ビッグデータの「3つのV」の意味

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



- IBM社の資料では、「3つのV」に加えて、4つめのVとして「データの正確さ」の（Veracity）を挙げています。
  - 様々な組織が公表する資料によっては、「価値あるデータ」（Value）を加えて「5つのV」としているケース、10以上のVではじまる英単語を列挙しているケースもありますが、4つ以上のVの中には「3つのV」の「Variety」「Volume」「Velocity」が含まれることが標準的です。

【出所】IBM Data Engine for Hadoop and Spark（P4）［IBM］ http://www.redbooks.ibm.com/abstracts/sg248359.html

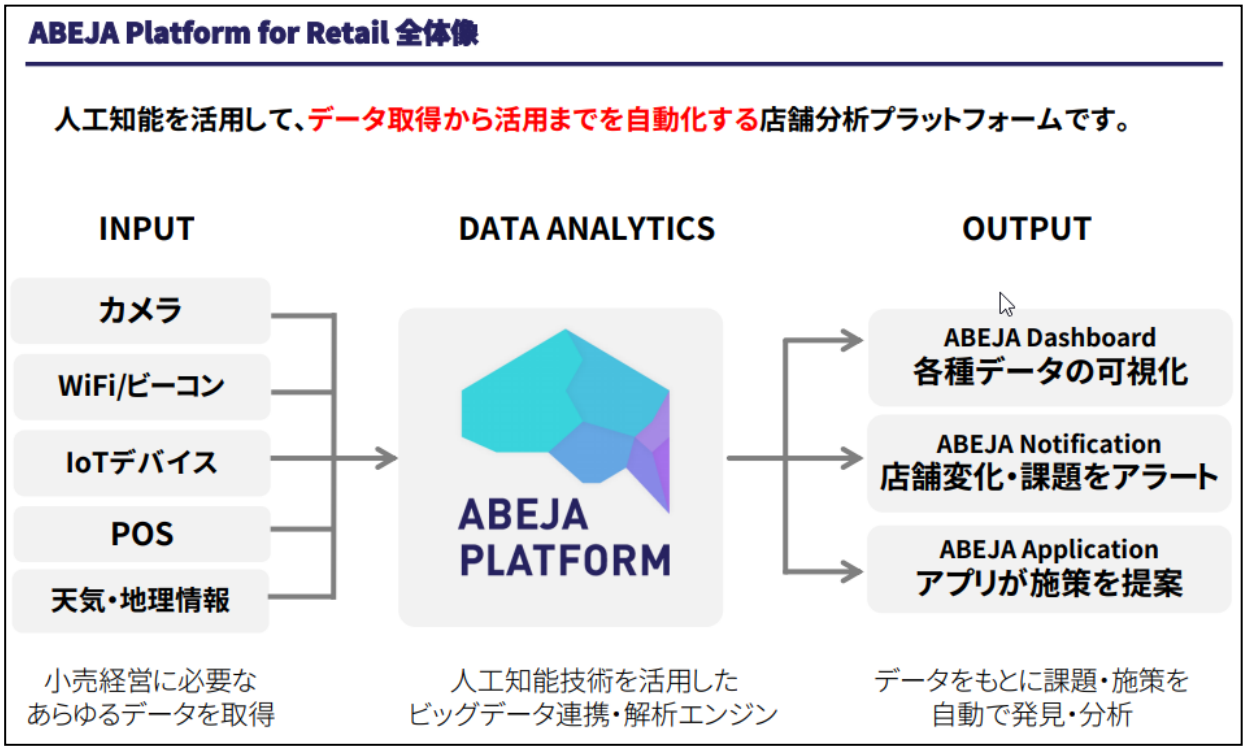
□ 3つのVはビッグデータが持ち得る特性であるため、ビッグデータであっても「対象情報やファイル形式が固定しているケース」「データ量が小さいケース」「データの生成速度が遅い、低頻度のケース」があり得ます。

Variety（多様性）により可能となる分析

◆ビッグデータのVariety（多様性）から様々なデータを統合した分析が可能となります。

- 株式会社ABEJA（アベジャ）が提供しているABEJA Platformでは小売店の店舗にカメラを設置して、来客人数をカウントし、来客者の年齢層・性別を人工知能で判定します。
- 「カメラによる画像情報」「Wi-Fi/ビーコンによる顧客の移動」「IoTデバイスによる扉の開閉状況」「POSによる売上データ」「インターネットから得られた天候情報」を組み合わせて、販売状況の管理・分析が可能となります。
  - POS（ポス）は［point of sales （system） ：販売時点情報管理］の頭文字に由来するレジの販売情報管理です。

ABEJA platform for Retailにおけるカメラによる情報収集と分析概要



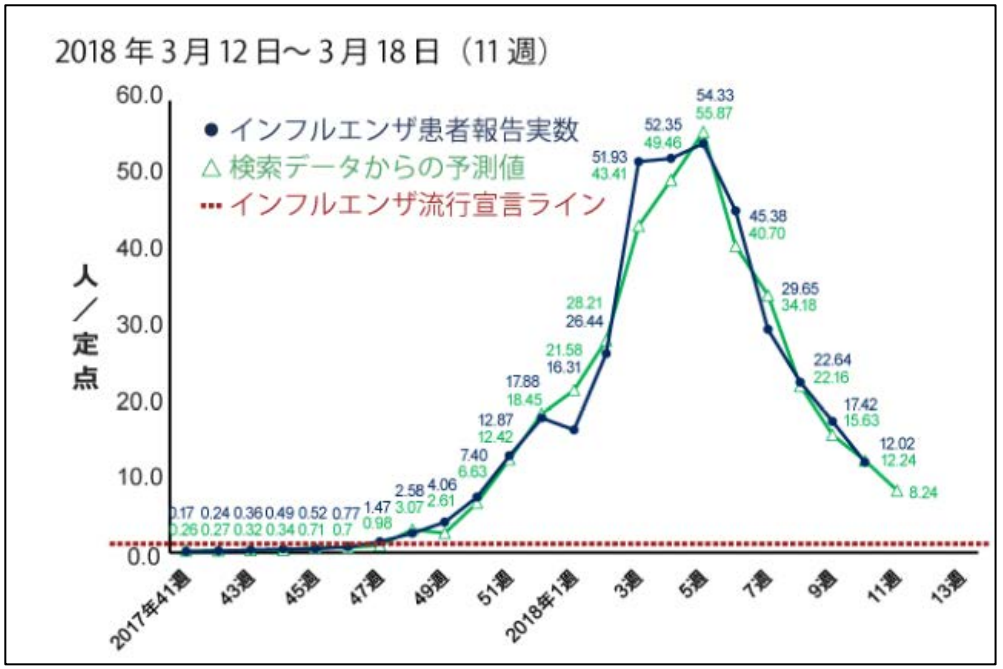
【出所】株式会社ABEJA https://abejainc.com/ja/

Volume（データ量）により可能となる分析

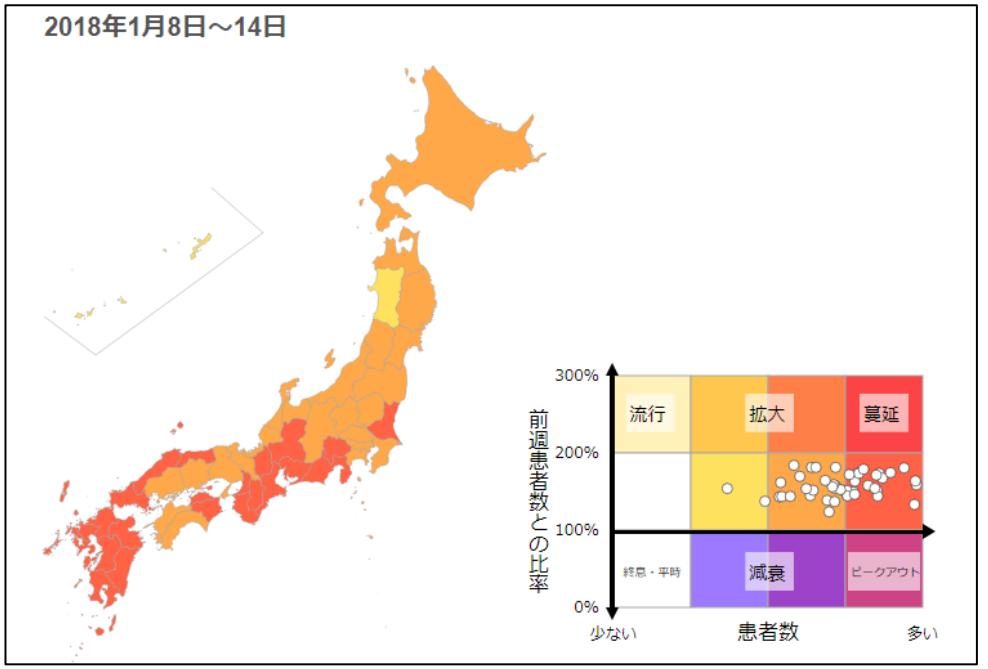
◆ビッグデータのVolume（データ量）から、膨大なデータに基づく分析が可能となります。

- Yahoo! Japanでは、時期別・都道府県別のインフルエンザの患者数と相関の高いキーワード検索数から、インフルエンザの感染数の予測値を示しています。
  - 「インフルエンザ」「発熱」「寒気」等のキーワードでの検索数を時期別・都道府県別に集計して分析用データとして活用しています。
  - 検索のキーワード、検索数と実際の患者数の対応関係を分析することで、予測の精度を一層高めることができます。
- 週単位、都道府県別にインフルエンザの「流行期」「拡大期」「蔓延期」「減衰期」の推移を確認できます。
  - ウェブサイトにおいて、データの対象期間を動かす矢印のボタンをクリックすることで、インフルエンザの流行と減衰が動的に把握できます。

検索データに基づくインフルエンザの感染予測値と患者数



インフルエンザ感染状況マップ



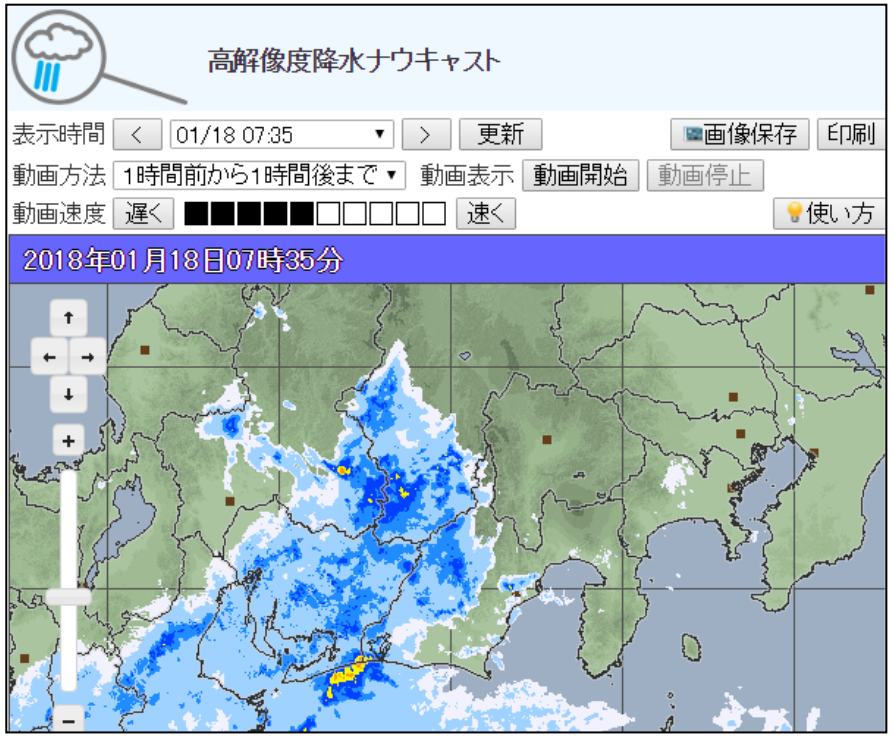
【出所】ビッグデータ分析でみるインフルエンザ感染状況：2017－2018［ヤフー株式会社］
https://about.yahoo.co.jp/info/bigdata/influenza/2017/01/

Velocity（データ生成速度・頻度）により可能となる分析

◆ビッグデータのVelocity（データ生成速度・頻度）からリアルタイムでの活用ができます。

- 気象庁が提供する「高解像度降水ナウキャスト」では、5分単位での降水状況および、1時間後までの降水予想を地図上に示します。
  - 詳細な地図で表示できるため、ゲリラ豪雨の予測にも利用できます。

高解像度降水ナウキャスト（気象庁）



【出所】高解像度降水ナウキャスト［気象庁］
https://www.jma.go.jp/jp/highresorad/

- ネットショッピングサイトのAmazon.co.jpでは、各ユーザーの購入予定の商品に合わせて、即座にお勧め商品を表示します。
  - 講座3-5に示す「アソシエーション分析」に基づいて表示します。

購入予定商品に合わせた商品推薦（Amazon.co.jp）

【出所】Amazon.co.jp　https://www.amazon.co.jp/

構造化データ、半構造化データ、非構造化データ

◆ビッグデータは、人間にとって読みやすく、分析しやすい構造化データだけではありません。

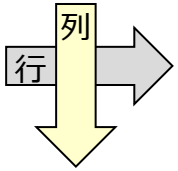
- ビッグデータはその特性である**多様性（Variety）**から**構造化データ**、**非構造化データ**のケースもあります。
  - 講座2-1のデータベースの説明においても、「構造化データ」「半構造化データ」「非構造化データ」を紹介しました。
  - 半構造化データの「XML」「JSON」に関しては、講座1-4のAPIで利用されるファイル形式として説明しました。

「構造化データ」「半構造化データ」「非構造化データ」に関する説明表

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型）
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整理や変換が必要です。

日本政府の構造化・非構造化データの提供サイト

◆日本政府のウェブサイトには、公的統計の構造化データを提供するe-Stat、非構造化データを含めて幅広く提供するDATA.GO.JPがあります。

- 公的統計の調査結果データを提供しているe-StatではExcel形式、CSV形式のデータをダウンロードでき、構造化データを提供しているウェブサイトと言えます。
  - 講座4-3の参考2にて、Rによる利用方法を紹介するe-Stat APIにおいては、e-Statが蓄積するデータを半構造化データ（XML、JSON）の形式でも提供しています。
- DATA.GO.JP(データカタログサイト)は、「政府の報告書などのPDF」「政府ウェブサイトのHTML」「報告書内の画像JPEG」といった非構造化データを含めて幅広く提供しています。
  - DATA.GO.JP(データカタログサイト)は、講座4-1でも紹介する日本政府のオープンデータの提供サイトでもあります。

公的統計の構造化データを提供するe-Stat

【出所】e-Stat［総務省］https://www.e-stat.go.jp/

非構造化データを含めて提供するDATA.GO.JP

【出所】DATA.GO.JP［総務省］http://www.data.go.jp/

データの品質

◆データには品質があり、データの品質が悪ければ、利用や分析における障害となります。

- 構造化データに限っても、重複するデータ、表記揺れ等があり、**データの品質が悪いケース**があります。
- 国際データマネジメント協会の英国支部の資料では、**データの品質**には6つの主要基準があると示しています。
  - 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。

DAMA UKのレポート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】

【出所】THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT［DAMA UK］
http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193

- データの品質が悪ければ、データ利用・データ分析ができなかったり、誤った出力が得られたりします。
  - ある側面でデータの品質が悪かったとしても、利用目的によっては不都合がないケース、データクレンジングによって修正ができるケースもあります。

品質の悪いデータによる社会的費用

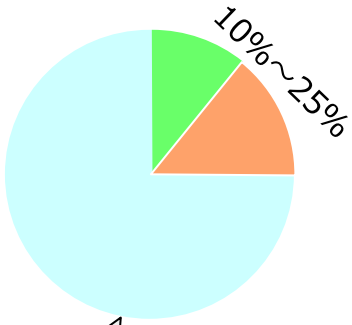
◆品質の悪いデータは、大きな社会的費用を生んでいます。

- 2016年にIBM社より公刊された書籍では、「品質の悪いデータがアメリカ経済に与えているコスト推定値は年間3.1兆ドル」と紹介しています。

【出所】Data Engine for Hadoop and Spark（P4）［IBM］
Http://www.redbooks.ibm.com/abstracts/sg248359.html

- 品質の悪いデータが生み出す社会的費用は、「正しいデータが確認できないことによる機会損失」「データの廃棄や追加的な作業によるコスト」「不正確なデータ利用に基づくコスト」が挙げられます。
  - 社会的費用には実際に支出した費用のみならず、データの品質が悪いことによって得られなかった潜在的な利益も含まれます。

アメリカにおける「品質が悪いデータが生み出すコスト」に関するインフォグラフィック（翻訳）



品質の悪いデータがもたらす経営上のコストは、組織の収益の10～25%にのぼる。

ヘルスケア部門における質の悪いデータによる費用は3140億ドル（→31兆4000億円）となっている。

質の悪いデータによるアメリカ経済へのコストは3兆ドル（→300兆円）であり、2011年の財政赤字の2倍以上となっている。

正しいデータが分からないことで、ある主要な小売業者が被った被害は年間300万ドル（→3億円）となっている。

典型的なIT予算の50%は、使えない情報とその補修に費やされる。

平均的な会社は、不正確なデータのために想定した受取人に届かない郵送料に毎年18万ドル（1800万円）を費やしている。

＊ オリジナルのインフォグラフィックの値から「1ドル＝100円」にて換算

【出所】SOFTWARE AGのインフォグラフィックに基づき作成　https://lemonly.com/work/the-cost-of-bad-data

データ形式の標準化とデータクレンジングの重要性

◆「データ形式の標準化」や「データクレンジング」によってデータの品質を高めることができます。

- 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。
  - 「データ形式の標準化」や「データのクレンジング」によって、品質の悪いデータによる社会的費用を軽減することができます。

【出所】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（調査委託先：株式会社 NTTデータ経営研究所）］
http://www.soumu.go.jp/main_content/000422923.pdf

- **データ形式の標準化**は、公的機関や業界等のコンソーシアムが形式を定め、データの提供者が実施する根本療法に相当し、**データクレンジング**は、一般に分析者・利用者自身が行う対処療法に相当します。
  - 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」の技術が必要なケースもあります。

ビッグデータ活用における課題（品質の悪いデータに対する対応策）

対応策	主な実施主体	意味
データ形式の標準化	データ形式の決定：公的機関・業界等のコンソーシアム 標準化の実施：データ提供者	定められた基準によって、データのファイル形式や変数名を統一し、利用可能なデータレコードを抽出することによって、データの利用やデータセット同士の連結を容易にすること
データクレンジング	分析者・利用者	データレコードの重複、データ内の誤記、表記の揺れなどを修正・統一することでデータの品質を高めること

- 品質の良いデータであっても、利活用に適する形への「データ整理・抽出」や「データ加工・結合」は必要であり、「データクレンジング」「データ抽出・加工・結合」との技術は、データ分析者・利用者にとって重要です。
  - 「データクレンジング」「データ抽出・加工・結合」といった分析前の一連の作業は、「データの前処理（まえしょり）」とも言われます。

データ形式の標準化政策

◆日本政府では「データ形式の標準化」に関する政策を推進しています。

- 総務省の自治体クラウドポータルサイトでは、地方自治体が保有するデータの標準化を推進するべく中間標準化レイアウト仕様を公開してます。
  - 中間標準レイアウト仕様では「住民基本台帳」「印鑑登録」「戸籍」といった行政書類の標準フォーマットを公開しています。
  - 地方自治体のデータ形式を標準化することで、広域でのデータ連携、住民の転居に伴うデータの移行をスムーズに行うことができます。

- 経済産業省では、消費・購買データの標準的なフォーマットを設定し、電子化された買物レシート（電子レシート）の標準仕様を検証する実証実験を2018年2月に実施しました。
  - レシートのデータを電子化・標準化することで、様々な商店・ネットショッピングサイトの消費・購買データを一括して取り扱うことができるようになります。

総務省（自治体クラウドポータルサイト）の中間標準レイアウト

○中間標準レイアウト仕様V24

・(表形式)一括ダウンロード(zip)(12.9MB)

・(XML形式)一括ダウンロード(zip)(8.4MB)

	一括ダウンロード(zip)	表形式					XML形式		改訂履歴
		移行ファイル構成表	移行ファイル関連図	データ項目一覧表	コード構成表	コード一覧	レイアウト仕様	サンプル	
1. 住民基本台帳									
2. 印鑑登録									
3. 住登外管理									
4. 戸籍									
5. 就学									
6. 選挙人名簿管理									

【出所】中間標準レイアウト仕様［総務省］
http://www.soumu.go.jp/main_sosiki/jichi_gyousei/c-gyousei/lg-cloud/02kiban07_03000024.html

経済産業省の電子レシート実証実験用アプリ

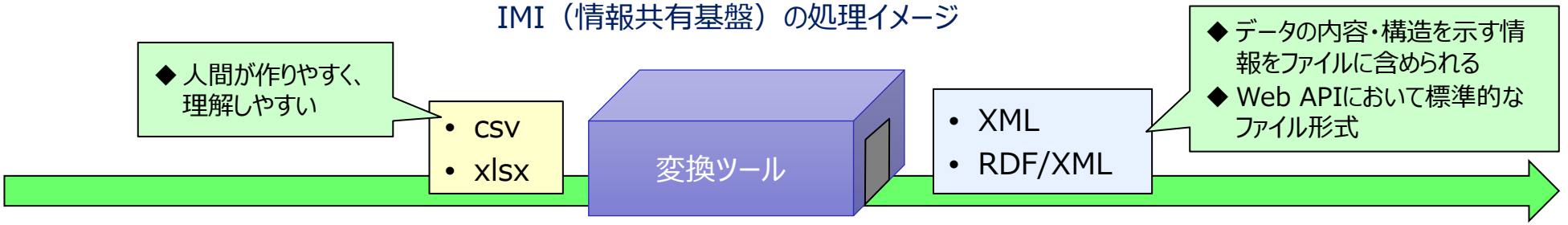
【出所】電子レシートの標準仕様を検証する実験を行います［経済産業省］
http://www.meti.go.jp/press/2017/01/20180131004/20180131004.html

データ形式の標準化ツール

◆日本政府ではデータ形式を標準化するツールの開発・公開を行っています。

- 経済産業省が設置し、情報処理推進機構（IPA）が事務局を担当する**IMI（情報共有基盤）**では**「DMD Editor」というデータ形式の標準化・変換を行うウェブツール**を提供しています。
  - IMIは［Infrastructure for Multilayer Interoperability（情報共有基盤）］の略であり、DMDは［Data Model Description（データモデル記述）］の略となっています。
  - DMD Editorはウェブサイトに「csv」や「xlsx」をアップロードすることで、自動で「RDF/XML」「JSON」といったファイルに変換できます。
  - 「RDF/XML」は講座1-5にて示したXMLに外部からの特定・リンクを可能とした規格であり、講座4-1の「機械判読への適性の5段階」でも紹介します。

IMI（情報共有基盤）の処理イメージ



【出所】IMI（情報共有基盤）https://imi.go.jp/

災害被災者支援 支援物資提供データ（二次元の表形式）

場所	提供者	支援物資	提供状況
○×小学校	NPO◆◆	飲料水	提供中
△□公民館	株式会社■■	米	提供準備中

RDF/XML

```
…<ic:場所 rdf:resource="○×小学校"/>
   <ic:状況型>
     <rdf:Description>
       <ic:説明-単純型>飲料水</ic:説明-単純型>…
```

JSON

```
…{"@id":"_:b2",
"http://imi.go.jp/ns/core/rdf#説明-単純型":
[{"@value":"飲料水"}],
"http://imi.go.jp/ns/core/rdf#名称型":
[{"@id":"_:b4"}]},...
```

- 2018年1月決定の「デジタル・ガバメント実行計画」において、**日本政府はIMIを整備・活用**する旨が示されました。

【出所】デジタル・ガバメント実行計画［eガバメント閣僚会議］https://www.kantei.go.jp/jp/singi/it2/kettei/pdf/egov_actionplan.pdf

データのクレンジングツール

◆無償利用可能なデータクレンジングツールもありますが、日本語への対応は不十分です。

- データクレンジングを行うための無償利用が可能な英語版ソフトウェアとしてOpenRefineが挙げられます。

【出所】OpenRefine http://openrefine.org/

- 日本語は英語に比べても、漢字表記や送り仮名の違い等の表記揺れが多く、標準化（名寄せ）は、より重要です。

住所表記・会社表記のデータ形式の標準化（名寄せ）例

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事だけでも、一貫性が損なわれてしまいます。

ソニー株式会社の表記揺れ				
ソニー株式会社	Sony株式会社	SONY株式会社	Ｓｏｎｙ株式会社	ＳＯＮＹ株式会社
ソニー（株）	Sony（株）	SONY（株）	Ｓｏｎｙ（株）	ＳＯＮＹ（株）
ソニー(株)	Sony(株)	SONY(株)	Ｓｏｎｙ(株)	ＳＯＮＹ(株)
ソニー㈱	Sony㈱	SONY㈱	Ｓｏｎｙ㈱	ＳＯＮＹ㈱

会社表記の標準化例

ソニー株式会社
Sony Corporation
東証一部 6758（電気機器）
設立年月日　1946年5月7日

- 日本人が見れば、上記16種の企業表記は同一の企業だと分かりますが、文字列が異なるためデータ集計時には異なる企業として扱われてしまいます。
- 法人マイナンバー（法人番号）を利用すれば、正式な企業名を確認することができ、同じ企業名が複数ある場合でも企業を特定することができます。

- 日本語のデータクレンジングは、個々のケースに合わせてExcelやプログラミングで行っているケースが多くなっています。

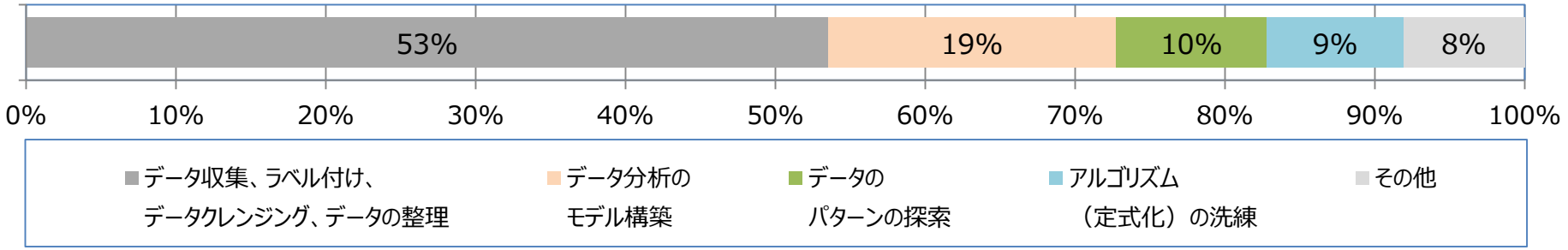
□ 講座3-2では、Excelを用いて日本語の表記揺れの統一を含むデータクレンジングの実習を行います。

データクレンジングの負担

◆データ分析において、データクレンジング・データ整理は時間がかかり、好まれない作業です。

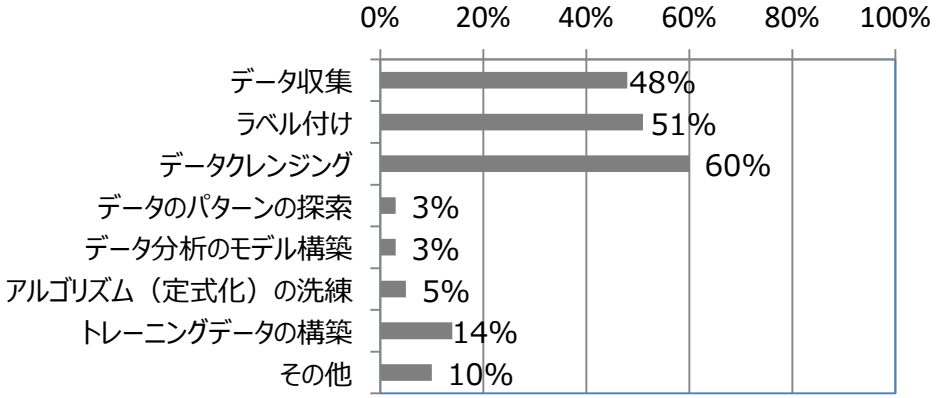
- 2017年2月における世界のデータサイエンティスト（データ分析者）197名に対する調査では、データクレンジングを含むデータの前処理が最も時間を割いている業務と回答した者が過半の53%となっています。

データサイエンティストがデータサイエンスの業務時間で最も時間を割いている業務

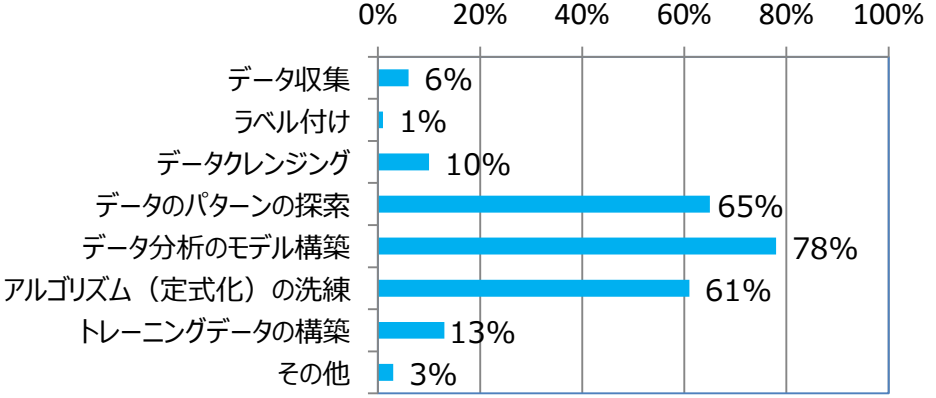


- データクレンジングは、データサイエンティストの業務の中で「最も楽しめない業務」として挙げられています。
  - データサイエンティストが最も楽しめる業務として「データ分析のモデル構築」が挙げられています。

データサイエンティストが最も楽しめない業務（最大3つ）



データサイエンティストが最も楽しめる業務（最大3つ）



【出所】2017 Data Scientist Report［CrowdFlower］に基づいて作成

□ データクレンジングは労働時間の大部分を占めるとともに心理的な負担になっており、その軽減が課題となっています。

国内企業におけるデータ分析の実態

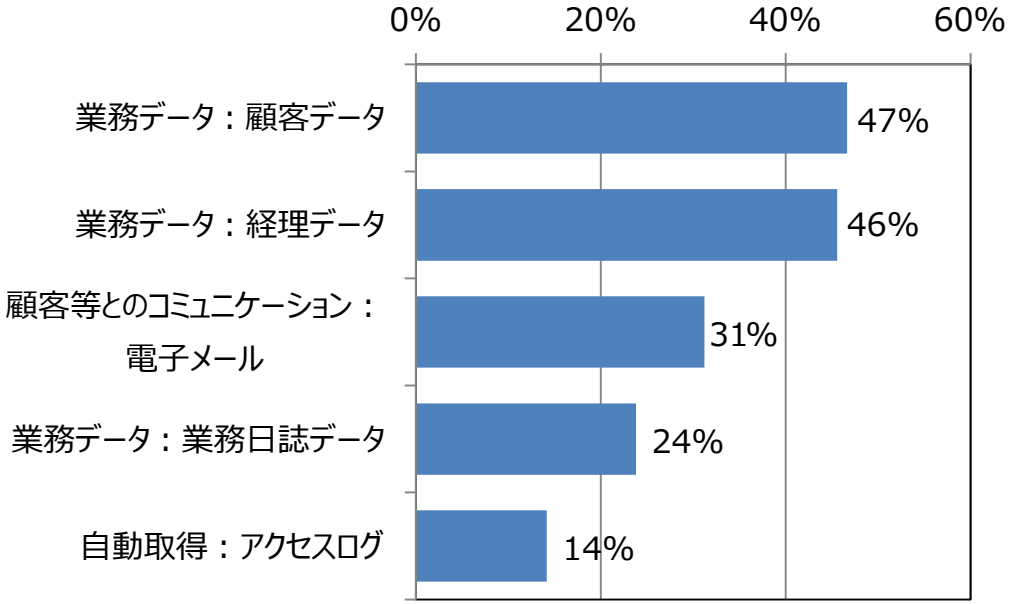
◆国内企業では「業務データ」を「見える化」するデータ分析の方法が、最も多くなっています。

- 総務省の2014年度の委託調査によれば、国内企業4,672社のうち72%の3,357社がデータ分析を行っています。
  - 本スライド下部の2種類のグラフはいずれも、データ分析を行っている3,357社が分母となっています。

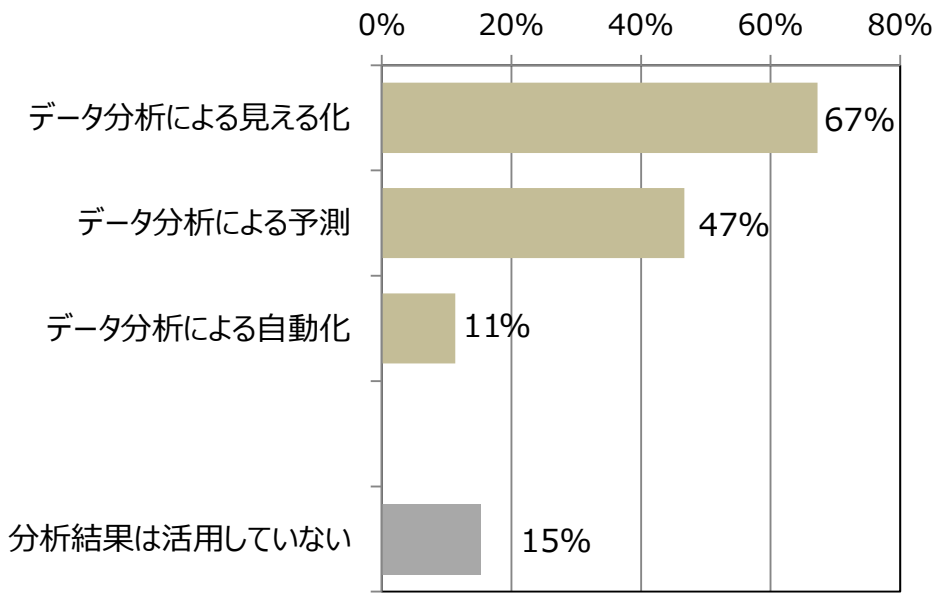
- 分析に活用しているデータとして「顧客データ」、「経理データ」の割合が高くなっています。
  - いずれも意図的に取得したデータではなく、自然に集まる業務データとなっています。

- 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（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。

分析に活用しているデータの割合（複数回答：降順上位5位）



データ分析の活用方法（複数回答）



【出所】ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究［総務省（調査委託先：株式会社 情報通信総合研究所）］に基づき作成
http://www.soumu.go.jp/johotsusintokei/linkdata/h27_03_houkoku.pdf

□ 自然に集まる業務データを活用し、見える化（可視化）して、分かりやすく表すことが分析の第一歩となっています。

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、目的や分析課題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。

ビジネスの目的例：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。

⇨ **データ分析によって、売上総額を最大化するための［販売単価］を知りたい。**

私生活の目的例：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分からない。

⇨ **データ分析によって、ダイエットに効果的な［食事制限］と［運動］の組み合わせを知りたい。**

- あらかじめ**「何をしたいのか？（⇒目的）」**や**「何を知りたいのか？（⇒分析課題）」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、データ分析の目的や分析課題を明確にしにくいケースにおいては、見える化（可視化）によってデータをく図表に表し、実態や外れ値を確認することで、高度な分析へのヒントが得られるケースもあります。

本格的なデータ分析に至るプロセス（工程）

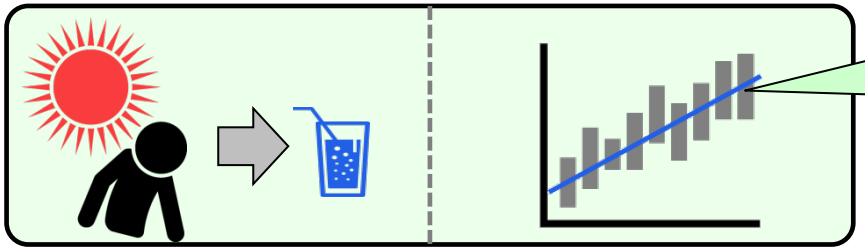
◆本格的なデータ分析に至る前には、いくつかのプロセス（工程）があります。

- データ分析を行う際の目的や分析課題には、様々なケースが考えられます。

目的・分析課題の設定例

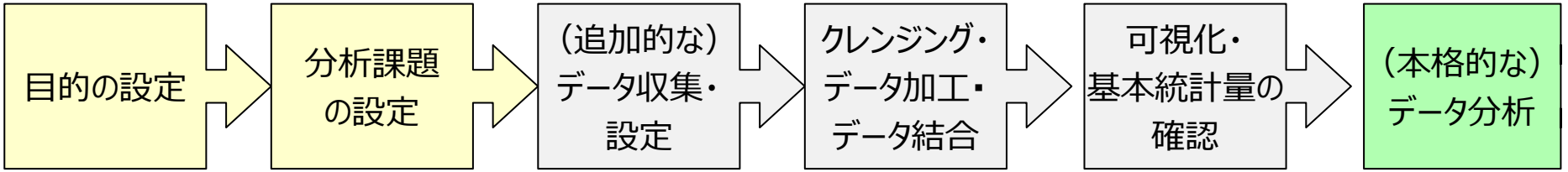
目的の設定	分析課題の設定
最適な仕入れ量の設定	環境と売上の関係を定量的に測定
購入機材の選択	各機材の費用対効果を測定
生産費用の削減	生産機械の最適なメンテナンス時期の把握
顧客満足度の向上	顧客満足度が増減する要因の特定

- 「定性的には当たり前のこと」であっても、定量的な関係な測定が分析課題となることもあります。
  - 「気温が上がれば、冷たい飲み物の販売量が増加する」ことは、感覚的・定性的に当たり前ですが、「気温1度の上昇につき平均○本の増加」「気温△度において、平均□本の販売量」という定量的な関係はデータ分析を行わないと把握できません。



【分析結果例】
◆ 気温1度の上昇につき平均12本の販売量増加
◆ 気温30度の平均販売量は123本

- 本格的なデータ分析の前には、「目的の設定」「分析課題の設定」に続く一般的なプロセス（工程）があります。



「目的・課題」に基づく「(追加的な）データ収集・設定」

◆「目的、課題ありき」の分析では、追加的なデータ収集や仮設定を検討することができます。

- 目的や分析課題が明確になっている「目的・分析課題ありきの方法」では、手元にないデータ項目があっても、追加的な収集や外部からの提供を検討することができます。
  - 追加データを入手するには、費用や労力がかかるため、データ取得自体の費用対効果も検討する必要があります。

「目的・分析課題ありきの方法」と「手元にあるデータからできることを考える方法」の比較

◆ 目的・分析課題ありきの方法 ⇒
- ○ 効率的な分析作業となりやすい
- ○ ストーリー・実益のあるレポートとなりやすい
- ○ **不足データの入手や設定を検討できる**

ゴール

◆ 手元にあるデータからできることを考える方法 ⇒
- × 効率的な分析作業となりにくい
- × ストーリー・実益のないデータ集になりやすい
- × **不足データは検討・言及できない**

スタート

- 入手できないデータ項目がある場合でも、近似値や仮定による設定を与えて分析をするケースもあります。
  - 利用可能なデータが利用したいと時点や地点と乖離しているなど、近似値のデータしか入手できないケースもあり得ます。
- 近似値のデータしか利用できないなど、データの品質が悪いケースでも、利用データの注意点を記載すれば、分析レポートとして提出・公表することができます。
  - 実際のデータ分析においては、万全の品質のデータが揃っていることは稀です。品質の悪いデータを利用しても、分析結果には大きな影響がないケース、品質の良いデータの収集のきっかけになるケースもあります。

データクレンジング・データ加工・データセットの結合

◆必要に応じて、データクレンジング、データ加工、データセットの結合を行います。

- 重複レコードや表記揺れがあれば、それらを修正・補正する**データクレンジング**を行います。
- 分析対象データの抽出や生年データから年齢データへ変換するなど、必要な**データ加工**があれば行います。
- 構造化データにおける**データセットの結合**は、同種のデータを追加し、**行（データレコード）が増加するアペンド（append）型**と外部データとの照合などによって**列（変数）を追加するマージ（merge）型**に分かれます。
  - **アペンド型のデータ結合**は、特定の県のデータに、比較対象としての他県のデータを追加するなど、**同じ変数でデータレコードを追加し、比較する範囲を広げるデータ結合**です。
  - **マージ型のデータ結合**は、特定の県の人口のデータに、可住地面積のデータを加えるなど、**新たな変数を追加し、新しい視点を与えるデータ結合**です。

データ結合の事例（アペンド型・マージ型）

都道府県	人口（万人）
神奈川県	913
静岡県	370

アペンド型のデータ結合 →

都道府県	人口（万人）
神奈川県	913
静岡県	370
愛知県	748

行（データレコード）の追加 →

マージ型のデータ結合 →

都道府県	人口（万人）	可住地面積（km^2）
神奈川県	913	1,471
静岡県	370	2,749

列（変数）の追加 ↓

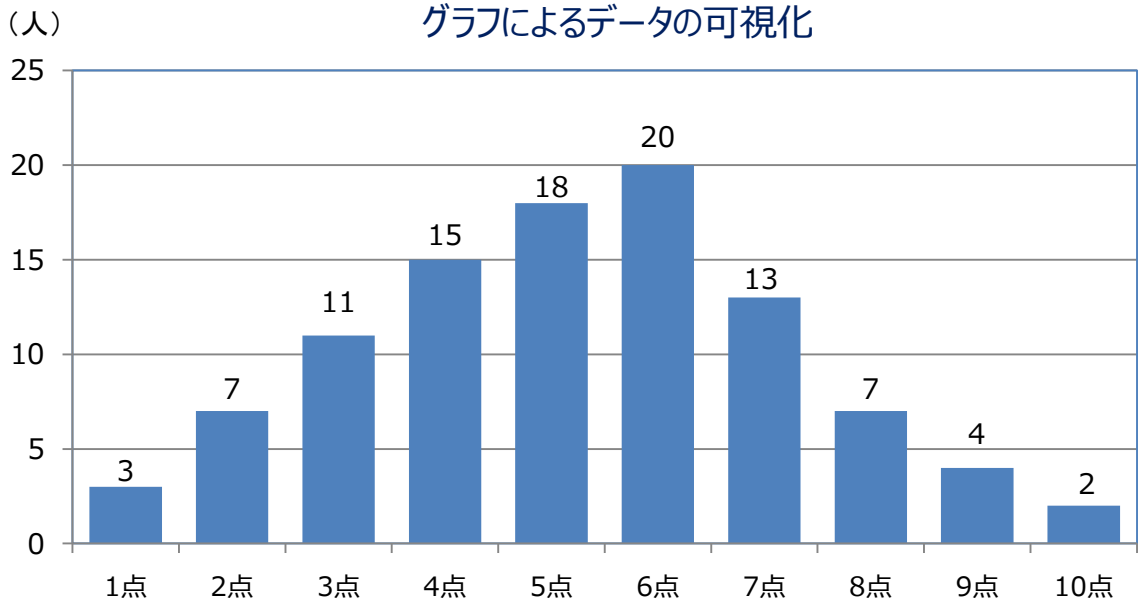
□ ExcelのVLOOKUP関数を利用したマージ型のデータセットの結合は、講座3-2にて紹介します。

データの可視化、基本統計量の導出

◆データの可視化や基本統計量を導出することで、データの全体像および外れ値を確認します。

- 本格的なデータの分析を行う前に、グラフ等で視覚的にデータの状態を確認する**データの可視化**によって、データの全体像や外れ値を把握します。
  - 突出した外れ値は、観測エラーや記入ミスとして除外すべきケースもあれば、現実の突出した値を示し、価値ある分析の糸口となるケースもあります。
- 各変数の代表的な値、バラツキに関する指標、最大値、最小値などの**基本統計量**を算出し、データの特徴を概観します。

グラフによるデータの可視化



基本統計量の導出

基本統計量	
平均値	5.23
最頻値	6
第1四分位	4
中央値（第2四分位）	5
第3四分位	7
最小値	1
最大値	10
分散	4.18
標準偏差	2.04

□ Excelにおけるデータの可視化については講座3-2、基本統計量の導出については講座3-3で説明します。