

さわって・わかつう・オープンデータ ～PCとInternet回線さえあれば、分析できるんです～

2015年01月27日

先端IT活用推進コンソーシアム

クラウド・テクノロジー活用部会 リーダー

アドソル日進株式会社 荒本道隆

このセッションの目的

- 成功条件

- 持っているデータの新しい活用方法のヒントが見つかった
- 明日、会社に行ったら、オープンデータにアクセスしてみようと思った
- RStudioをインストールしてみようと思った

- 説明の途中でもガンガン発言してください

- アイデアを閃いた、素朴な疑問、こんな事できないか？

- こんな話はしない予定

- 最新の事例

- 日本国内でのオープンデータの活用事例はまだ少ない

- アメリカと日本の、オープンデータの差

- アメリカ: 実データが沢山公開済み、爆発的に増加中
- 日本: 「やるやる」って言うだけで、まだ未公開のところも多い

自己紹介

- 荒本 道隆
 - AITC:クラウド・テクノロジー活用部会 リーダー
 - 所属:アドソル日進株式会社 先端IT技術部(調査研究部門)
- 最近の成果
 - 気象庁からの配信データをRDFで蓄積・公開
 - オープンデータの可視化ツールを作成、LODチャレンジに応募

AITC SPARQL Query

気象庁防災情報XML

配信feed+Areaに対してSPARQLを実行

蓄積データ: 2013/01/01~

SPARQL Query

```

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX atom: <http://www.w3.org/2005/Atom#>
PREFIX jma: <http://cloud.projectla.jp/jma/>
PREFIX area: <http://cloud.projectla.jp/jma/area#>
PREFIX link2: <http://cloud.projectla.jp/jma/link2/>

## Tripleを取得するサンプル
# SELECT * WHERE { ?s ?p ?o } LIMIT 100
## 述語のリスト
# SELECT DISTINCT ?p WHERE { ?s ?p ?o } ORDER BY ?p
## エリア名とコードのリスト
# SELECT DISTINCT ?name ?code WHERE { ?o area:name ?name . ?o area:code ?code } ORDER BY ?code
## タイトルで集計
SELECT ?title (COUNT(?id) as ?c) WHERE { ?id atom:title ?title } GROUP BY ?title ORDER BY DESC(?c)
## 日付で集計
# SELECT ?date (COUNT(?date) as ?c) WHERE { SELECT ?id (SUBSTR(?updated,1,10) as ?date) WHERE { ?id atom:updated ?updated
  
```

出力形式: Text

出力形式にXMLを指定した場合、レスポンスの先頭にXSLTスタイルシートの指定を追加できます:

☐ HTTPレスポンスヘッダにtext/plainを追加し、そのままブラウザで表示します。

エンドポイント: <http://api.aipc.jp/ds/sparql>

クエリ対象データサンプル [rdf triple](#)

AITC SPARQL Query

RDF Explorer

RDFエクスプローラ

エンドポイント選択:

エンドポイント: <http://sparql.odp.jie.jp/api/v1/sparql>

データセット名:

始点ノード追加:

ノード操作:

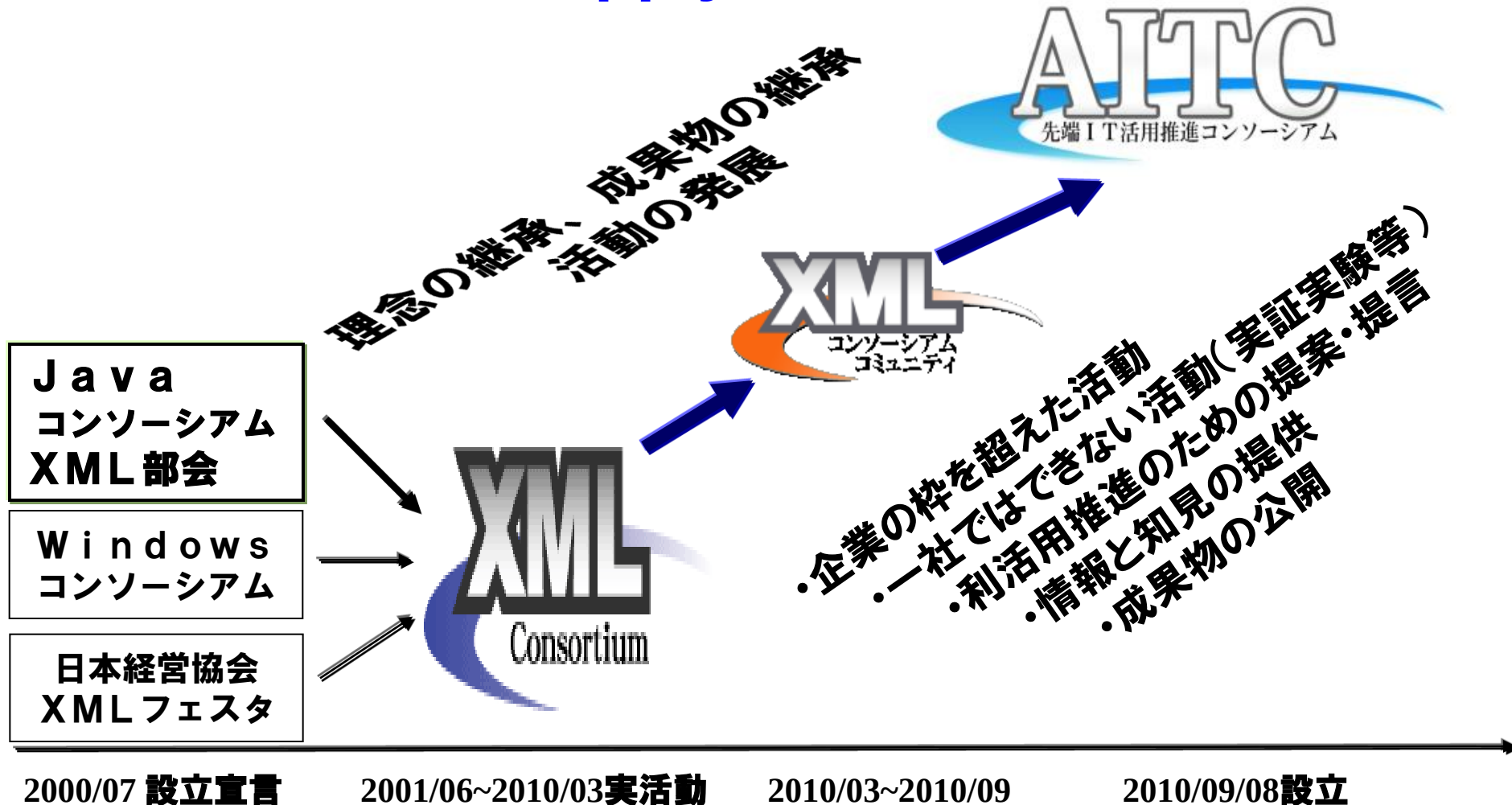
ステータス: 1つ前に戻る

Class: Equipment, AED, 設備, 112馬場医院, 112ウラセ株式会社, 112ひまわり歯科クリニック, 112市民ホール, 112靖江市立待公民館, 112靖江市立待小学校, 112中河保育所, 112靖江市公民館, 112靖江市公民館

The background of the slide is a light blue network diagram. It consists of numerous small circular nodes connected by thin, light blue lines, forming a complex web-like structure that fills the entire background.

**ちょっとだけ
先端IT活用推進コンソーシアム**
<http://aitc.jp>
ご紹介

Advanced IT Consortium to Evaluate, Apply and Drive



企業における先端ITの活用および

先端ITエキスパート技術者の育成を目的とし、

もって、社会に貢献することを目指す非営利団体

設 立： 2010年9月8日（会期： ～2016年8月31日）

会 長： 鶴保 征城（IPA顧問、HAL校長）

会 員： 法人会員＆個人事業主、個人会員、学会会員

特別会員（産業技術総合研究所、気象庁、
消防研究センター、防災科学技術研究所）

顧 問： 稲見 昌彦（慶応義塾大学大学院 教授）

和泉 憲明（産業技術総合研究所 上級主任研究員）

萩野 達也（慶応義塾大学 教授）

橋田 浩一（東京大学大学院 情報理工学系研究科 教授）

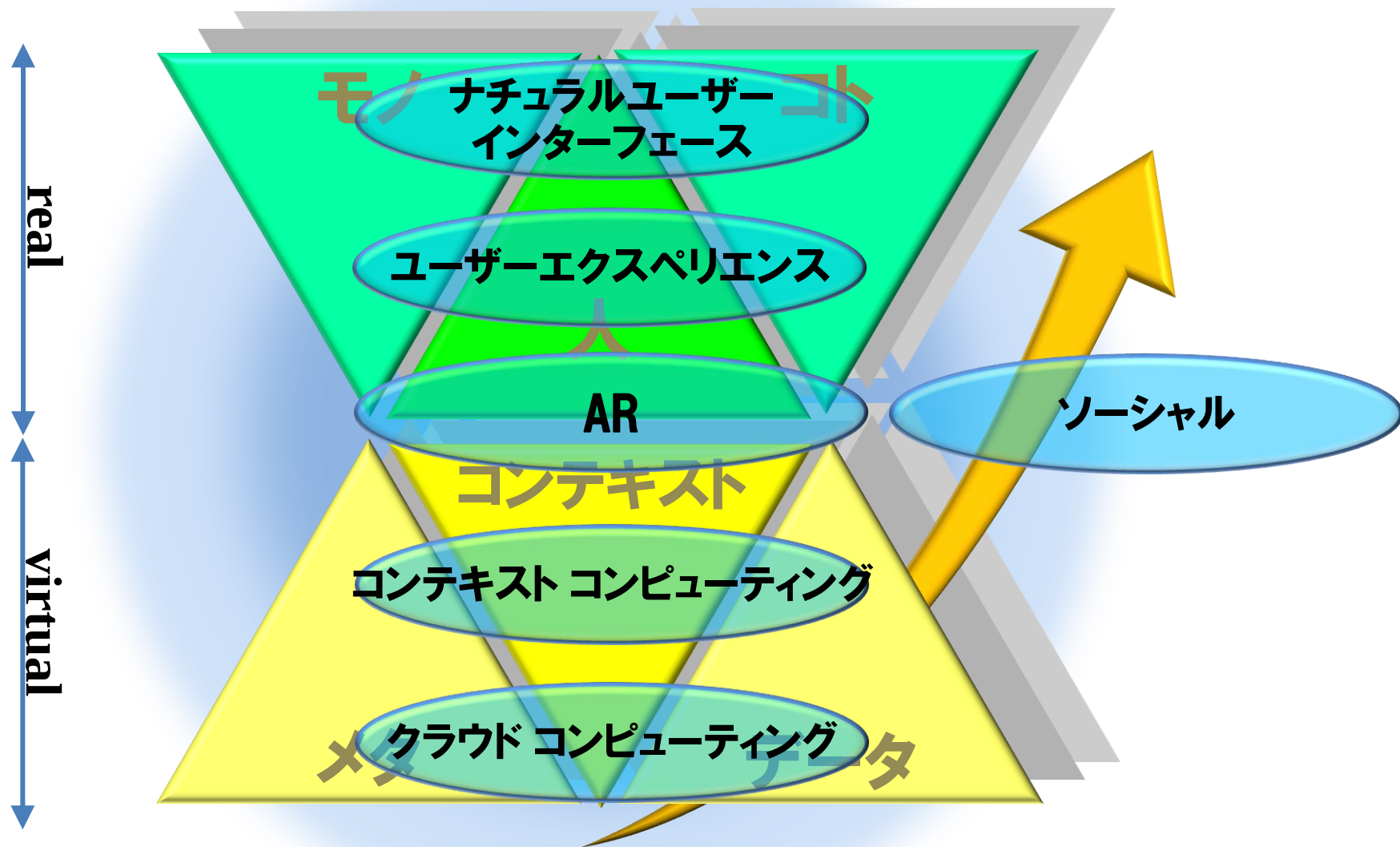
丸山 不二夫（早稲田大学大学院 客員教授）

山本 修一郎（名古屋大学大学院 教授）

BizAR顧問：三淵 啓自（デジタルハリウッド大学大学院 教授）

川田 十夢（AR三兄弟 長男）

第5期活動対象分野



活動の種類と関係性

部会のシーズ

セミナー・勉強会

部会へ
フィードバック

部会

AITC
オープンラボ

部会の
ノウハウを展開

月1回程度の開催
スキルと知見の伝播

プロジェクト

部会横断・外部との連携

発表会・シンポジウム

部会・プロジェクトの成果を発信

The background of the slide is a complex network diagram. It consists of numerous small blue circular nodes connected by thin, light blue lines. The network is dense and spans the entire width and height of the slide, creating a sense of interconnectedness and digital infrastructure.

ここからメインです

- 気象庁防災情報XMLフォーマットの策定に協力
 - 気象庁が発信しているテキストデータをXML化したもの
 - <http://xml.kishou.go.jp/>
 - <http://xml.kishou.go.jp/specifications.html>
 - http://xml.kishou.go.jp/jmaxml_20141024_manual_list.pdf
- 気象庁防災情報XMLを受信して蓄積するサーバを公開
 - 気象庁は、XML配信を試行的に実施中
 - http://xml.kishou.go.jp/open_trial/index.html
 - 申請さえすれば、無償で受信可能
 - 過去のXMLは参照できない
 - 受信にはWebサーバが必要なので、ちょっと敷居が高い
 - REST, SPARQL, WebSocket で参照可能なサーバを公開
 - <http://api.aitc.jp/>
 - 2013年1月1日以降の気象庁防災情報XMLを参照できる
 - 気象データは一番親しみのあるデータだし、色々な場面で使えるはず

世の中のデータを分類してみた

- 社内にあるデータ
 - 顧客に関するもの → 漏えいするとインパクト大
 - 社員に関するもの → 社内の様々なシステムに点在
 - センサーデータ → **IoT**
 - セキュリティ要件で、プライベートクラウドとパブリッククラウドを使い分ける
- 購入可能なデータ
 - データ収集業者
 - Tポイント
- 研究用に入手可能なデータ
 - 統計局 <http://www.stat.go.jp/info/tokumei/>
 - 医療DB <http://headlines.yahoo.co.jp/hl?a=20150104-00000020-san-pol>
 - 楽天 <http://rit.rakuten.co.jp/opendataj.html>
 - 利用制限や漏えい時のペナルティがある
- 公開されているデータ
 - Web上で公開されているデータ
 - オープンデータ

オープンデータとは

- Web上で公開されているデータ
 - 二次利用や二次配布が制限されている場合が多い
 - 掲載されているWebサイトのライセンスに従う
 - ライセンスが明記されていない場合も多い＝使う場合は慎重に
- オープンデータ
 - 一切の制限無しに使える
 - セキュリティの心配が無い
 - 悪意のある改ざん、誘導はダメ

概要 [編集]

オープンデータの概念は新しいものではない。しかし、形式的な定義は新しい。[Open Definition](#)による形式化は、「オープンデータとは、自由に使えて再利用もでき、かつ誰でも再配布できるようなデータのことだ。従うべき決まりは、せいぜい「作者のクレジットを残す」あるいは「同じ条件で配布する」程度である」という命題として要約できる[1]。

Wikipediaより

注意：今日はオープンデータ以外の話も出てきます

オープンデータの5つの段階

「オープンデータの5つの段階(出典:★)」と、データ形式

段階	公開の状態	データ形式 例	参考) Linked Open Data 5star	
1段階	オープンライセンスの元、データを公開	PDF、JPG	OL – Open License (計算機により参照できる (可読))	人が理解 するための 公開文書 (編集不可)
2段階	1段階に加え、コンピュータで処理可能なデータで公開	xls、doc	RE – Readable (Human & Machine) (コンピュータでデータが編集可能)	公開文書 (編集可)
3段階	2段階に加え、オープンに利用できるフォーマットでデータ公開	XML、CSV	OF – Open Format (アプリケーションに依存しない形式)	
4段階	Web標準 (RDF等) のフォーマットでデータ公開	RDF、XML	URI – Universal Resource Identifier (リソースのユニーク化、Webリンク)	機械判読 可能な 公開データ
5段階	4段階が外部連携可能な状態でデータを公開	LoD、RDF スキーマ	LD – Linked Data (データ間の融合情報が規定。検索可能)	

オープンデータの5つの段階

出典：★ Open Dataのサイト (<http://5stardata.info/>) およびTim Berners-Lee氏のLinked Dataに関する提言ページ (<http://www.w3.org/DesignIssues/LinkedData.html>) を参考に作成。

情報収集方法

- LODチャレンジ

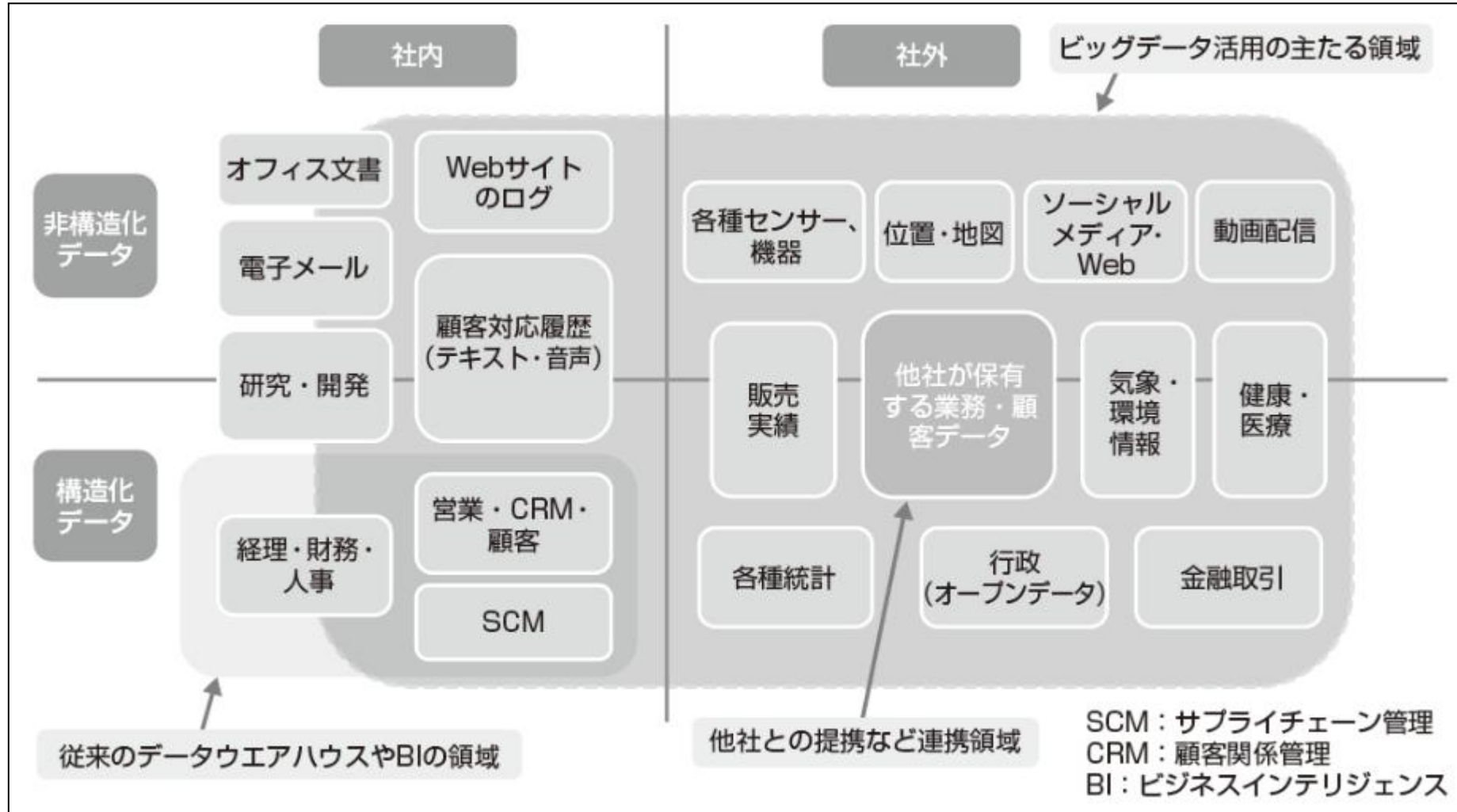


<http://lod.sfc.keio.ac.jp/challenge2014/>

- オススメの習得方法

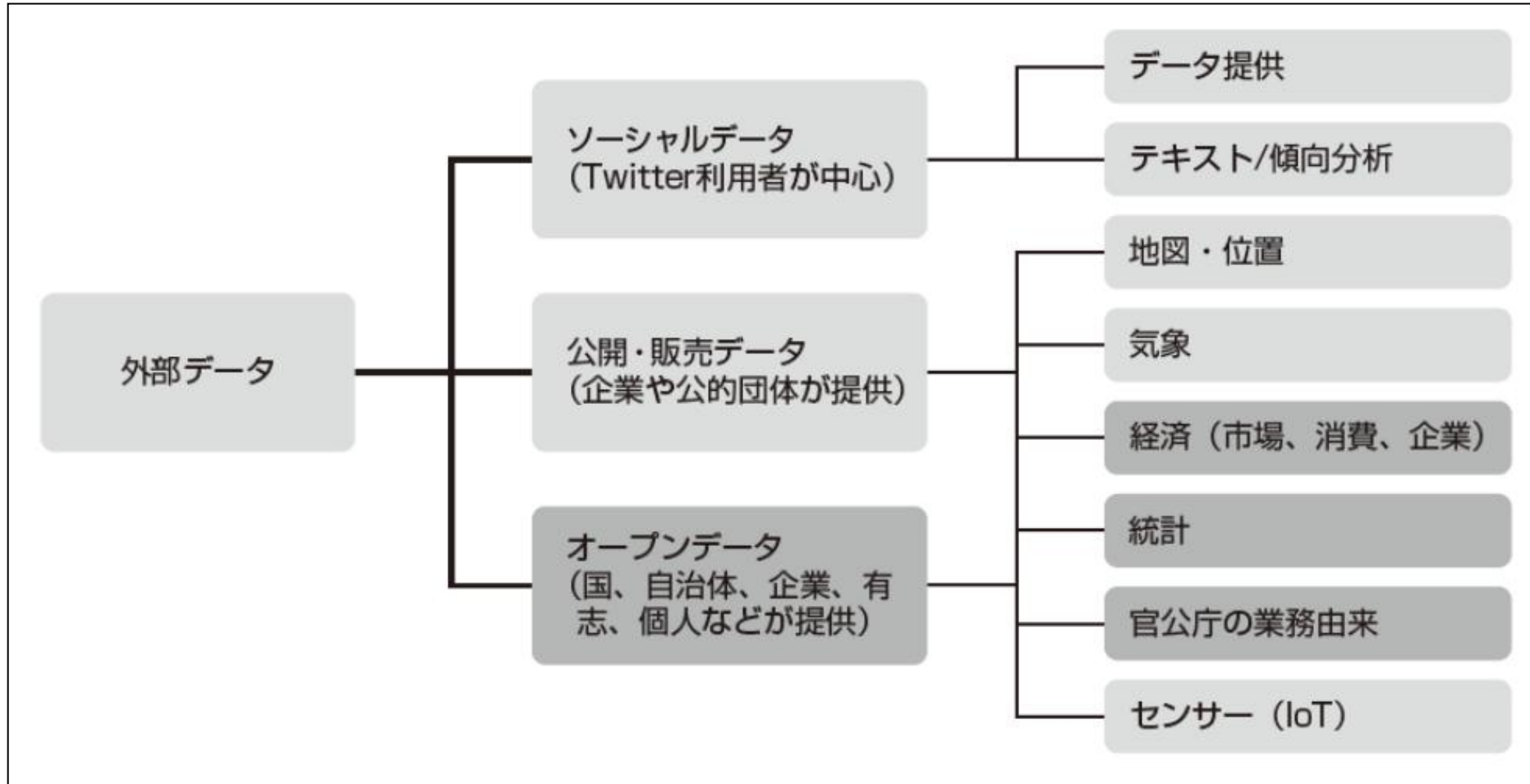
- 土日開催されているアイデアソン／ハッカソンに参加

活用の対象となっているデータ



出典: 日経BPビックデータ総覧2014-2015

主な外部データとその分類



出典：日経BPビックデータ総覧2014-2015

オープンデータをどう使うか

- 興味のある数字と、オープンデータとの相関を見つける

- 例：売上額と、最高気温もしくは最低気温

http://www.data.jma.go.jp/gmd/risk/hyoka_jiturei.html

- 例：ワインの品質と、雨量

<http://jibun.atmarkit.co.jp/ljibun01/rensai02/book/17/01.html>

- 例：景気指標と、Yahooの検索キーワード

<http://event.yahoo.co.jp/bigdata/keiki/>

- 例：ぜんそくの発症と、地図データと風向き

<http://www3.nhk.or.jp/news/iryoubd/>



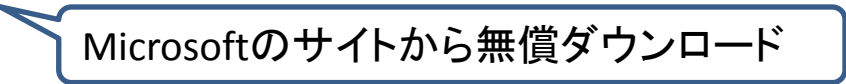
- オープンデータから、興味のある数字を予測

- 時間のかかる集計結果より、すぐ出る予測値の方がいい場合も多い

- オープンデータの予測値から、未来の興味のある数字を予測

- 売上額と、最高気温との相関があることが分かれば、
 - 気象庁ではスパコンで明日の最高気温、最低気温を予測している
 - ということは、「スパコンで明日の売上額を予測」と同じこと

相関を見つける方法

- 可視化して、相関のありそうな項目を見つける
 - 相関係数というものもあるけど、最終的には人間のセンス
 - → 最近流行りの「データサイエンティスト」
 - 総務省統計局の「社会人のためのデータサイエンス入門」を「gacco」上で開講
 - http://nttks.co.jp/news/news_20141219.html
 - 我が国の国際競争力を維持し、経済成長を加速化させるためには、ビジネスの現場においても、データに基づいて課題を解決する能力の高い人材、いわゆるデータサイエンスを身に着けた人材が不可欠となっています。
 - 2015年3月 開講予定
 - 2045年にCPUが人間の脳を超えるが、それまでは機械より人の方が優秀
- 可視化ツール
 - **R** 
 - **Excel (+ Power BI for Excel)** 
 - JavaScriptで自作
 - データを投入すると、可視化してくれるWebサイト
 - 注: ビジューアライズ専用ツールを使うと、発想が限られてしまう

可視化までの手順例

1. データ集め
2. フィルタリング
3. 結合
4. 統計処理
5. 表示
6. 絞り込む
7. 相関を発見！

ここまでで、全体の工数の
80～90%の時間がかかる

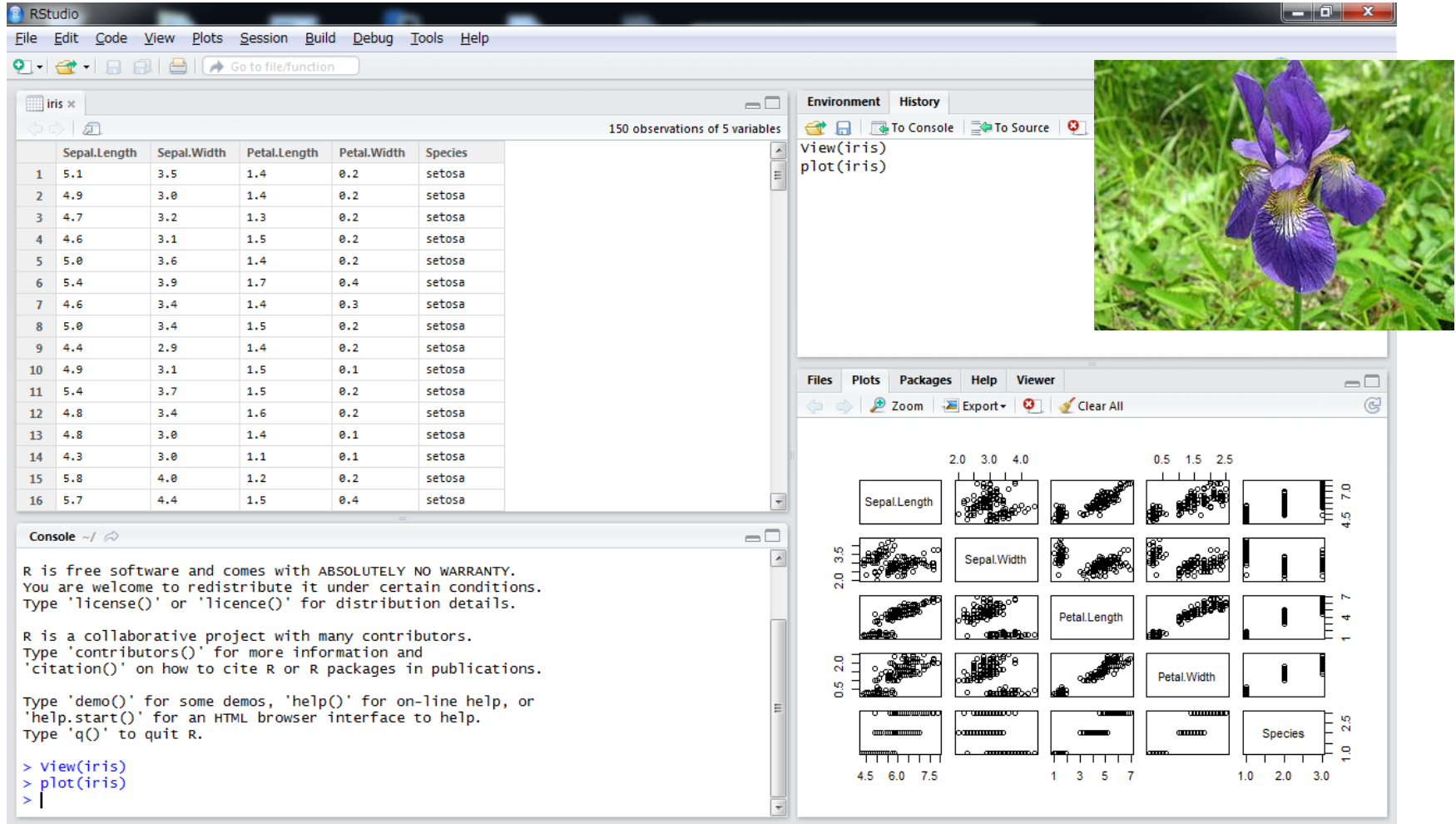
試行錯誤している途中に、
データをついつい暗記し、
新しいアイデアを閃く事
が多い

ツールに頼ったり、
外注しない方が良い

データを覚えるほど見
続ける、が一番重要

RStudioの画面紹介

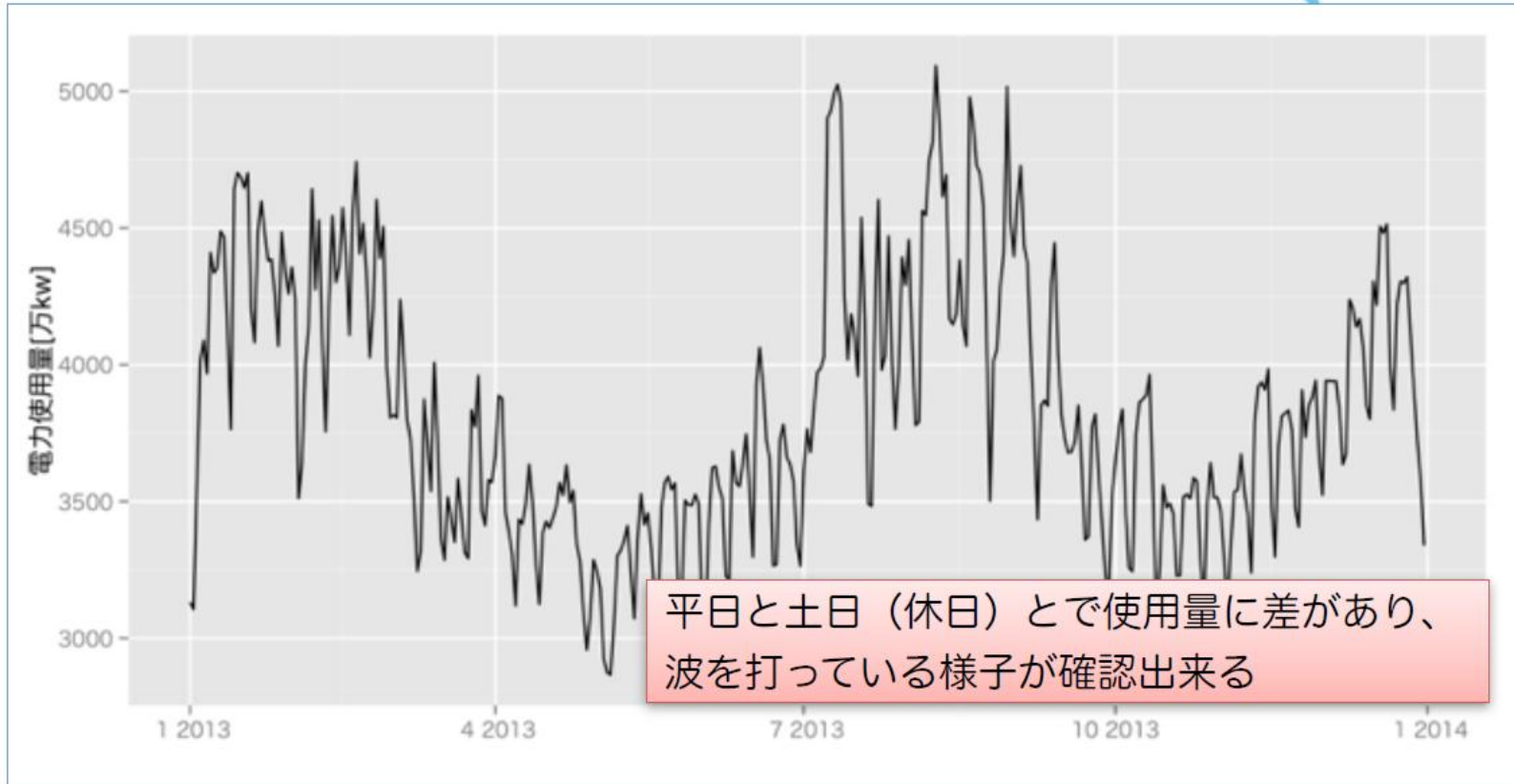
- Iris(アヤメ)データの項目間の相関を可視化した例



- 電力需要と気温との相関
 - － 電力需要：東京電力の2013年の出力量
 - 1時間単位の電力使用実績データ
 - <http://www.tepco.co.jp/forecast/html/download-j.html>
 - － 気温：気象庁が公開している過去の統計データ
 - <http://www.data.jma.go.jp/gmd/risk/obsdl/index.php>
 - － RStudioを使った実際の手順
 - <http://www.slideshare.net/yasuyukisugai/2-r-34017181>
 - データの加工部分は、Excelでやった方が早いかも

RStudioを使った分析例－2

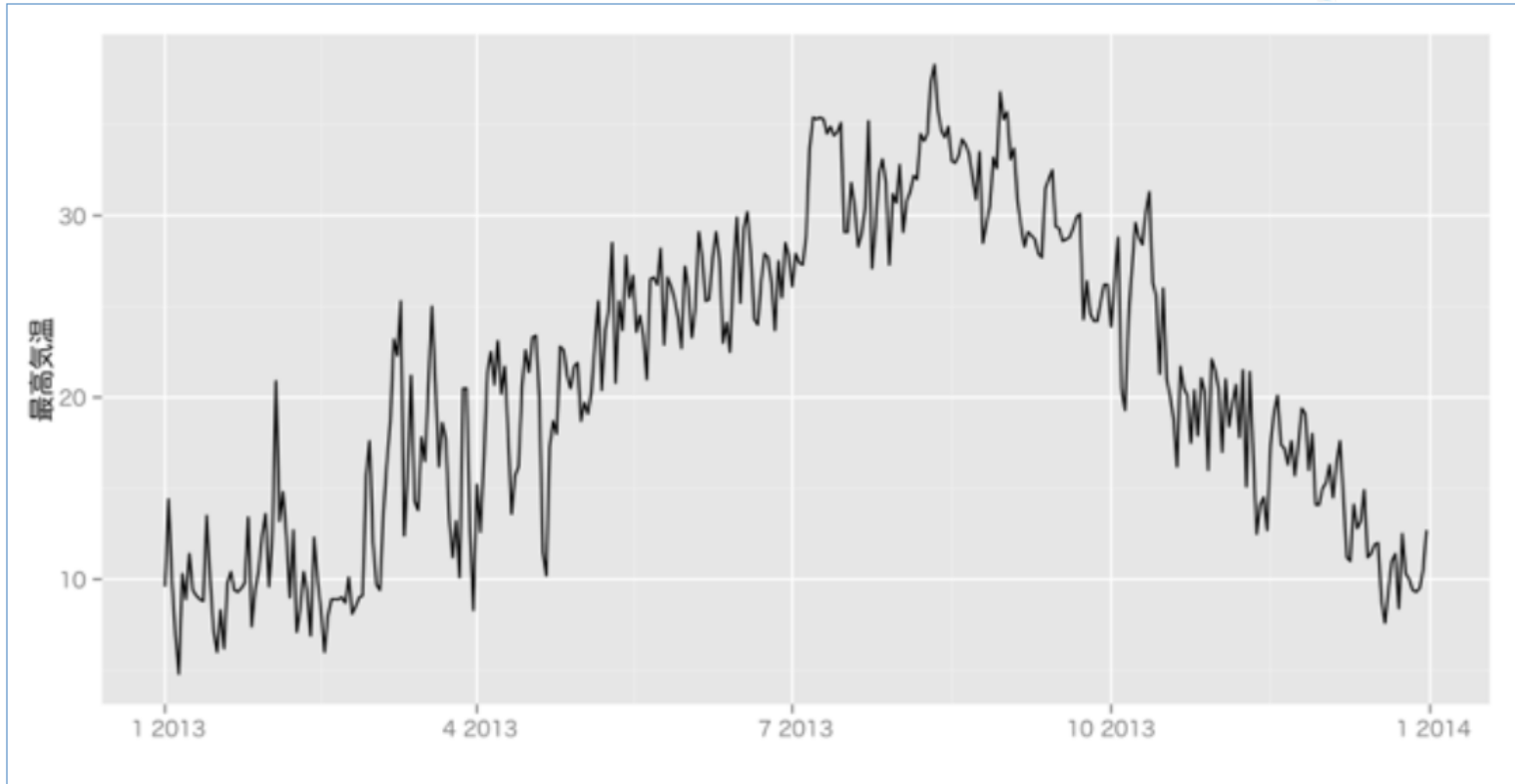
- 最大電力使用量(ピーク)



1 時間辺りの電力使用量から、1 日の最大電力使用量（ピーク）を算出

RStudioを使った分析例－3

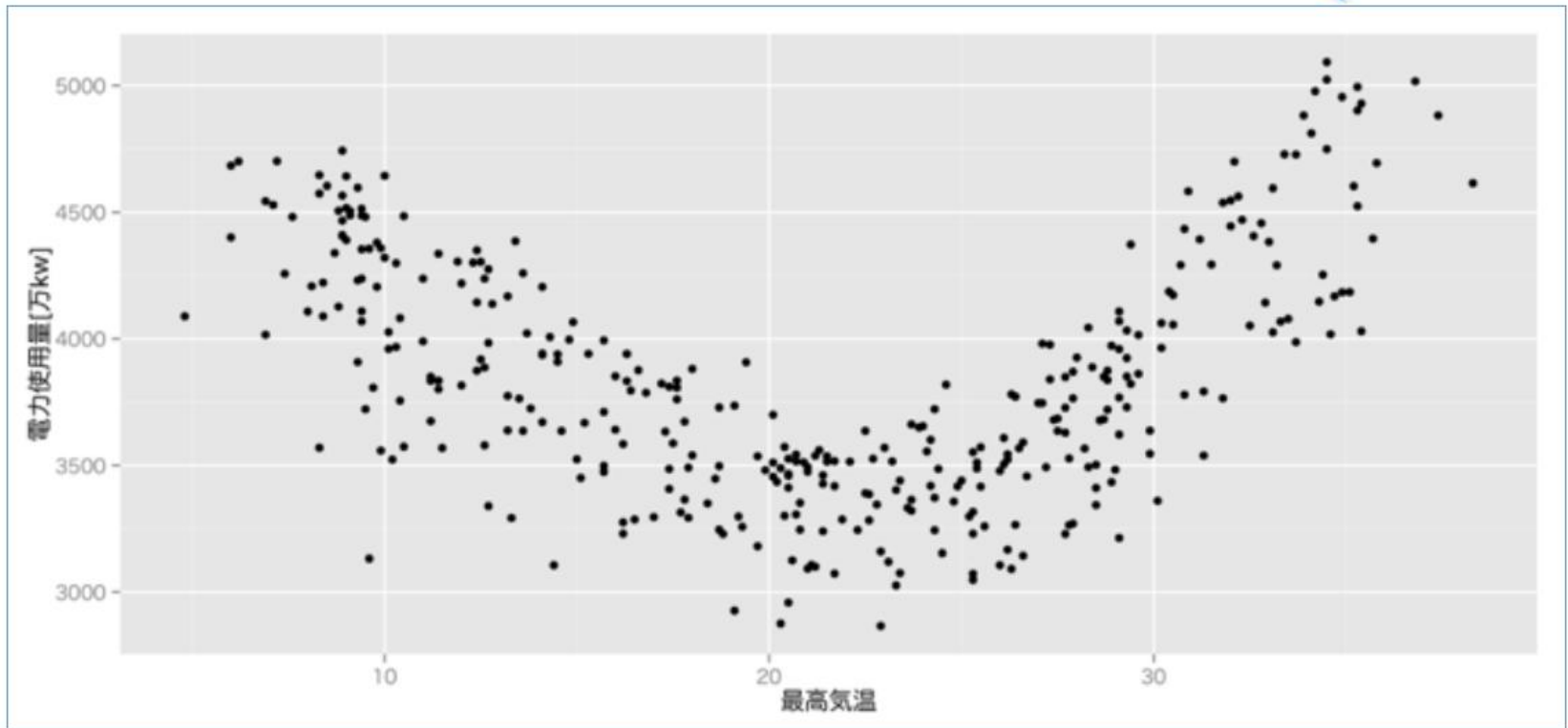
- 最高気温



季節の中でも、最高気温にバラつきがあることが確認できる

RStudioを使った分析例－4

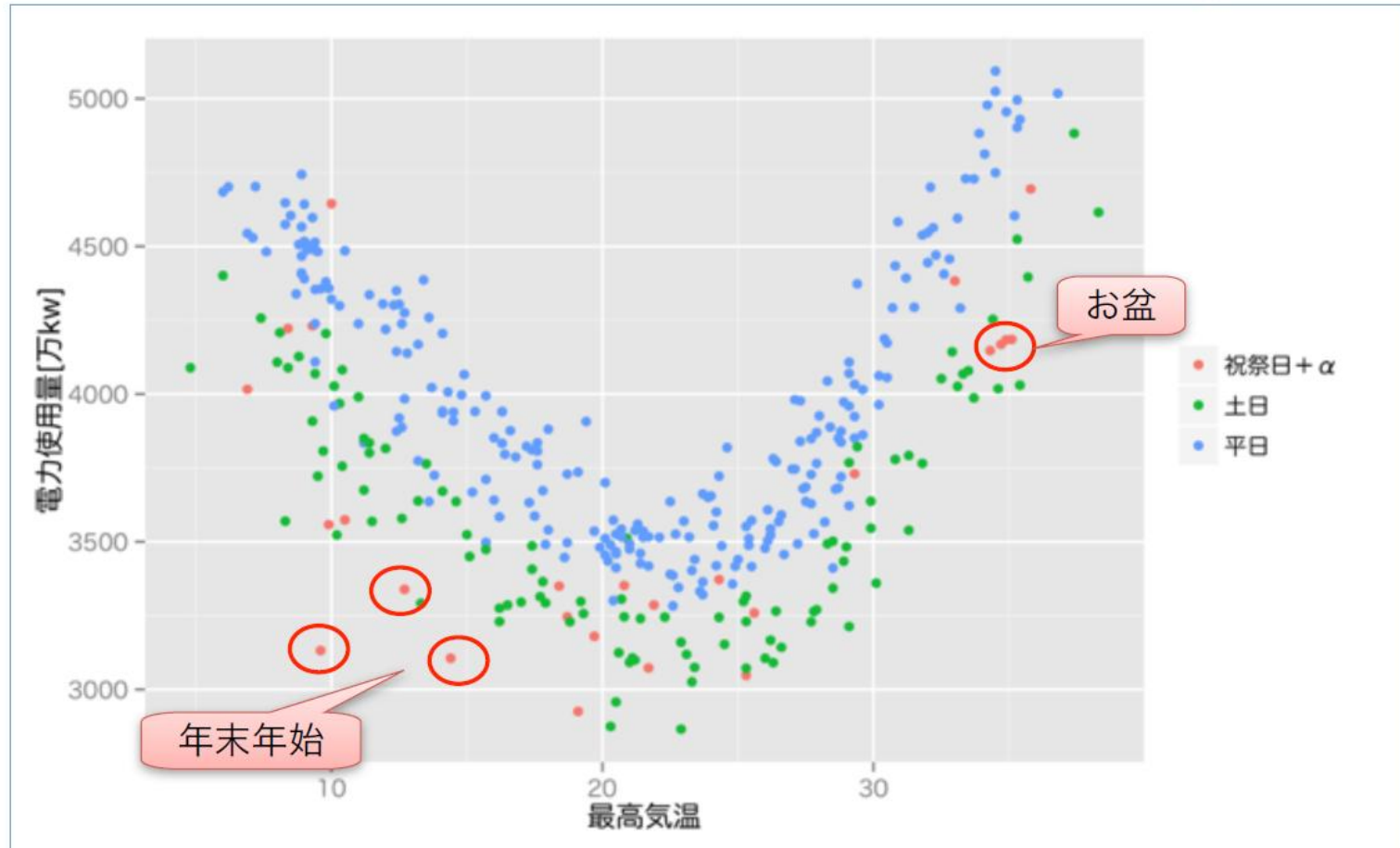
- 最大電力使用量と最高気温の相関



気温が上がるor下がる程、使用量が増加する傾向

RStudioを使った分析例－5

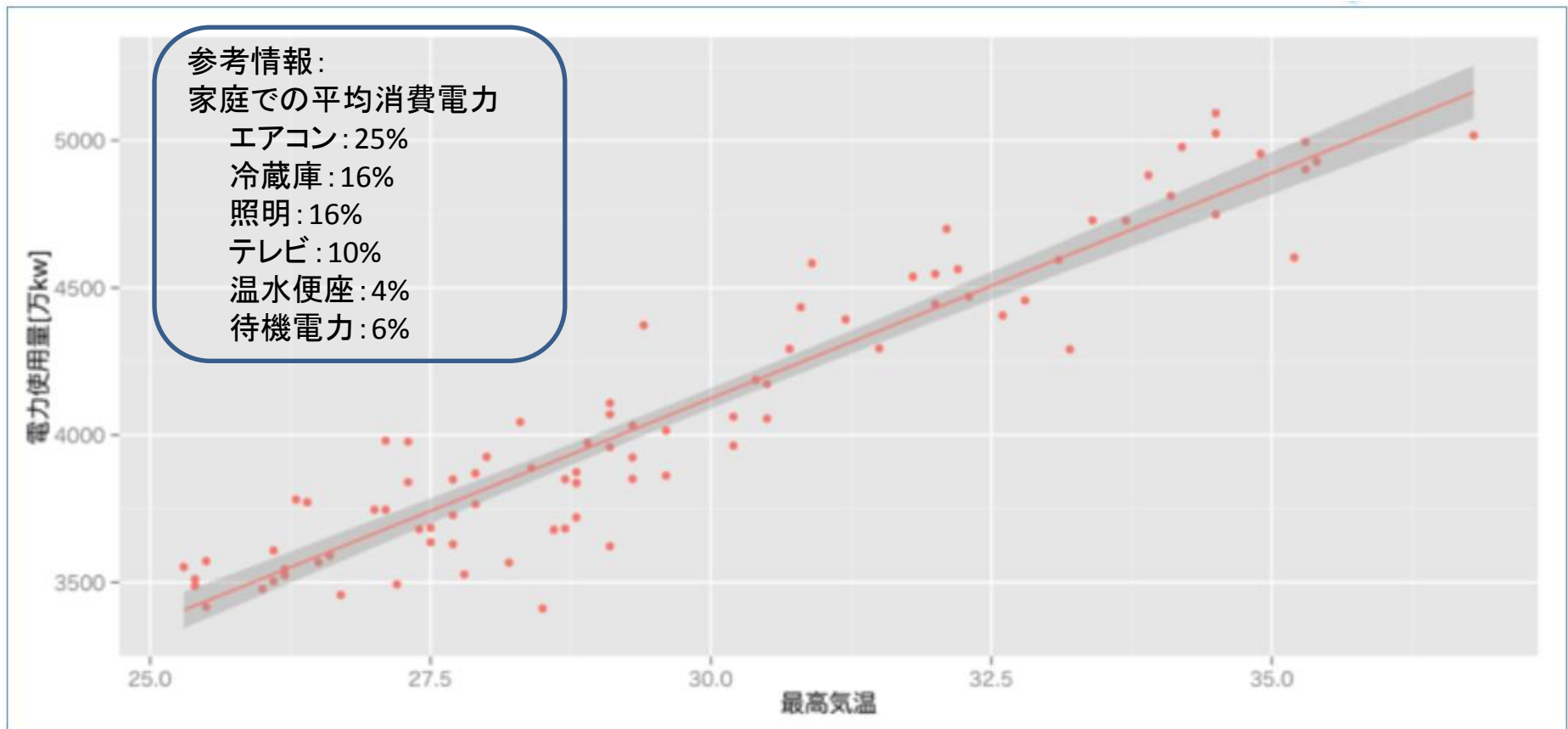
- 最大電力使用量と最高気温の相関(祝日分類)



平日、土日に加えて祝祭日やお盆、年末年始をさらに分類

RStudioを使った分析例－6

- 最大電力使用量と最高気温の相関(平日の25度超)
 - 最高気温が予測できれば、最大電力使用量(ピーク)も予測できそう



可視化時の注意

- 人間の目は、2次元より高い次元の図は、直観的に理解できない
 - 使えるのは、2軸(X軸とY軸)＝2項目まで
 - 先の例では、最高気温がX軸、最大電力使用量(ピーク)がY軸
 - もっと多くの次元を追加したい場合
 - 重回帰により、各項目ごとの定数を算出し、合算する
 - 主成分分析、次元圧縮などを行う
- さらに項目を追加したい場合
 - プロットの色、形、大きさで、情報量を追加する
 - 位置情報は、それだけでX軸とY軸を使ってしまう
 - プロットの色、形、大きさで別の情報を表現する事が多い
 - 時間軸を使用 → アニメーション
 - **ハンス・ロスリング教授のサイト**: <http://www.gapminder.org/>
- 参考文献
 - エンジニアのための データ可視化[実践]入門
 - Kindle版:1,946円、単行本:3,002円



JavaScriptを使った分析例

- 横浜市のテレビ放送受信契約数を年でアニメーションしてみる
 - 横浜市統計ポータル:テレビ放送受信契約数
 - <http://www.city.yokohama.lg.jp/ex/stat/index2.html#24>
 - 住所 → 緯度経度変換サイト
 - <http://newspat.csis.u-tokyo.ac.jp/geocode/>
 - 国土交通省、国土数値情報
 - <http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N03.html>
 - 実際の手順
 - http://cloud.aitc.jp/20140717_D3js/tv/
 - http://cloud.aitc.jp/20140717_D3js/20140717_D3js.pdf
 - キモは、Excelを使ったデータ整形
- 自分でコードを書くメリット
 - GoogleMapsを利用して簡単に実現できるサイトもあるけど、
 - データを外に出さないなので、社外秘データでも同じことができる
 - 細かい調整や機能追加をいくらでもできる
 - 慣れれば、お仕着せのツールよりも短時間でできる？

オープンデータではないので、勝手に二次利用しちゃダメ

以前、横浜市に問い合わせをして、「出典を明確にすれば、勉強会ならOK」との回答アリ

どこでも使えそうなデータ

- 気象データ

- <http://www.data.jma.go.jp/gmd/risk/obsdl/index.php>
- <http://hello-apis.blogspot.jp/2013/03/webapi.html>

- 気象庁が配信している電文データ

- <http://api.aitc.jp/>

対象電文

http://xml.kishou.go.jp/jmaxml_20130628_format_v1_1_hyo1_1.pdf

- 総務省統計局、統計データ

- <http://www.stat.go.jp/data/>

もはや嫌がらせレベルのExcelファイル

- e-Stat

- <http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>
- APIで公開されているので、取り込み易い

- データカタログサイト

- <http://www.data.go.jp/>

変態CSVが多い

用途に合えば使えそうなデータ

- オープンデータプラットフォーム
 - <http://odp.jig.jp/>
- LOD チャレンジ 2014 データセット部門
 - <http://lod.sfc.keio.ac.jp/challenge2014/dataset.html>
- LOD検索サイト
 - <http://lodlinker.org/>
- 自治体が公開しているオープンデータ
 - 横浜オープンデータカタログ
 - <http://www.city.yokohama.lg.jp/seisaku/seisaku/opendata/catalog.html>
 - ふじのくにオープンデータ
 - <http://open-data.pref.shizuoka.jp/>
 - → 静岡県の実タイム河川情報
 - 雨量・水位は24時間しか保持されない
 - 別の場所で、10分間隔でGET & 保管 <http://aramoto.sakura.ne.jp/shizuoka2/>

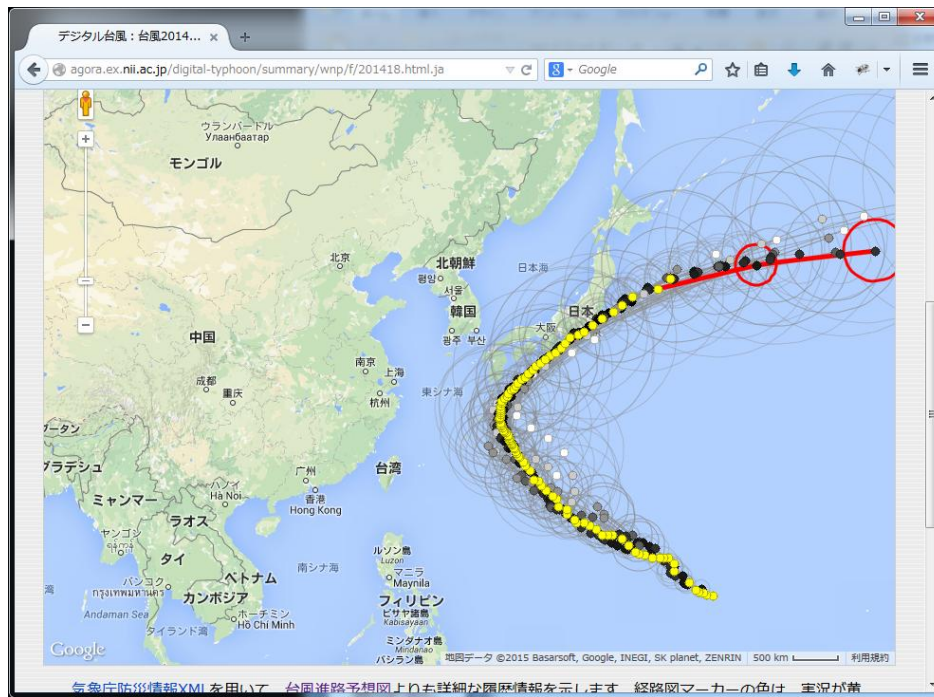
Webサイトの紹介－1

- 地図で見る統計(統計GIS)
 - <http://www.e-stat.go.jp/SG1/estat/toukeiChiri.do?method=init>
 - 例: 東京駅周辺の、5歳階級の人口



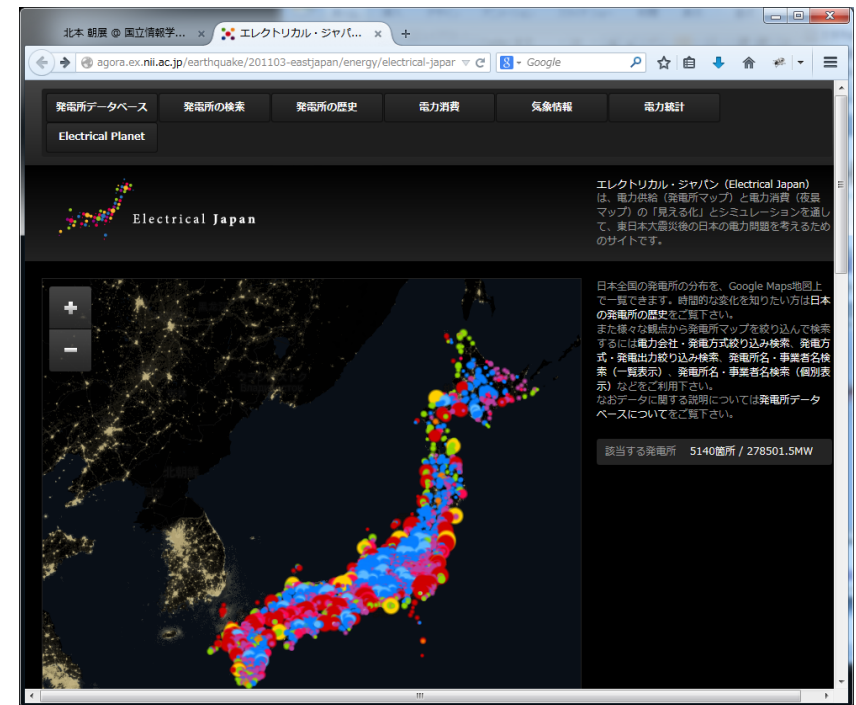
Webサイトの紹介ー2

- CPS-IIPプロジェクト
 - <http://agora.ex.nii.ac.jp/cps/>
 - 可視化するだけでなく、アートに昇華することで人の心を動かす



台風の予報履歴

<http://agora.ex.nii.ac.jp/digital-typhoon/summary/wnp/f/201418.html.ja>



発電所の分布

<http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/energy/electrical-japan/>

オープンデータを使ってみよう

- 社内データとセンサーデータとオープンデータの組み合わせ
 - まずは、手元にある社内データと組み合わせてプロットしてみる
 - オープンデータだけでは、あまり面白くない
 - センサーデータ(IoT)はビックデータ
 - AITCでも、IoT関連のイベントを月1ペースで開催中
- 例：売上情報と、センサと、気象データで何ができる？
 - 年齢、性別、表情を測定可能なセンサが購入可能
 - <http://plus-sensing.omron.co.jp/technology/detail/>
 - 事前に測定
 - 来店者の年齢、性別と、天気との相関 → 天気から来店者数を予測
 - 来店者の年齢、性別と、購入者数との相関 → 「来店した人が買う確率」が分かる
 - 来店時の表情と、購入者数との相関 → 「この人が買う確率」が個別に分かる
 - キャンペーン時に測定
 - どの年齢、性別の来店者が増えたか？ → キャンペーンに反応した年齢、性別
 - どの年齢、性別の購入者数が増えたか？ → そのキャンペーンが有効だったか
 - 表情と、購入者数との相関 → どんな表情の人が買い易いか → 声かけのヒントに？

- どれか1つでも該当しましたか？
 - 持っているデータの新しい活用方法のヒントが見つかった
 - 明日、会社に行ったら、オープンデータにアクセスしてみようと思った
 - RStudioをインストールしてみようと思った
- 今後の進め方(オススメ順)
 - LODチャレンジの過去の入賞作品をしてみる
 - 統計、R関係のイベントに参加してみる
 - アイデアソン／ハッカソンに参加してみる
 - センサーを買って、どこかに付けてみる
 - ハンス・ロスリング教授のプレゼンを視聴する
 - <http://matome.naver.jp/odai/2137432558736685501>

AITCでも、月1で公開イベントを開催中。
興味のある人は是非ご参加ください。



<http://aitc.jp>



<https://www.facebook.com/aitc.jp>



ハルミン

AITC非公式イメージキャラクター