

ソーシャルメディアアナリティクス(1) ー笑いのツボを分析するー

2016年11月11日
ニフティ株式会社
松井くにお

• 職歴

- 1980: (株)富士通研究所入社
- 1980–1985: 機械翻訳システムの研究開発
- 1986–1989: 大規模電子化辞書の作成(EDR)
- 1989–1992: オンライン辞書システムの研究開発
- 1993–1994: 次世代ワープロの企画(開発事業部)
- 1994–1998: 超高速情報検索エンジンの研究開発
- 1999–2000: 情報検索支援システムの研究開発(中国R&D兼務)
- 2001–2006: ナレッジマネジメントシステムの研究開発
- 2007–2009: Web 2.0 beyondプロジェクト(米国富士通研究所に勤務)
- 2009.7– : サービスビジネスの推進(ニフティ株式会社)
- 2016.11– : 富士通研究所人工知能研究センター兼務

• 社外委員

- 2004–2010: 言語資源協会副会長
- 2006–2008: 情報処理学会理事
- 2011–2016: 静岡大学客員教授
- 2012– : 北陸先端科学技術大学院大学客員教授
- 2012–2016: JST特任フェロー
- 2014– : 内閣府SIP「レジリエントな防災・減災機能の強化」課題に設置されるプログラム会議委員(国プロの審査員・運営委員)
- 2016– : 静岡大学創造科学技術大学院特任教授

- 岩倉友哉（富士通研究所）
 - 2003年 株式会社富士通研究所入社。
 - 2011年 東京工業大学大学院総合理工学研究科物理情報システム専攻博士課程修了。博士（工学）。
 - 現在，株式会社富士通研究所 知識情報処理研究所 人工知能研究センター ナレッジシステムPJ 主任研究員
 - 入社以来、自然言語処理の研究開発に従事。
 - 機械学習に基づく固有表現抽出，文書分類など

- 11月11日
 - 分析とは？
 - 目に見えているもの／見えていないもの
 - なぜビックデータか
 - 昨年までの分析
 - ニフティサーブフォーラム＋デイリーポータルZ
 - 今までの分析結果
 - 今年の分析対象：デイリーポータルZの紹介
 - 内容、コンセプト、統計情報など
 - 授業で使うフォーマット
 - 分析実習に向けて
 - 分析方法と分析例
 - 宿題（自由課題）

分析とは？

白いシャツを着ているヒトが何回パスをしたか数えなさい

- 目に見えている(と思っている)ものだけが真実ではない
- 目に見えていない(と思っている)ものを読み取ることが必要

では、これは？



オリンピックのメダル獲得は難しい

- リオデジャネイロオリンピックの種目数
 - 306種目
- メダルのカウント
 - 金:307、銀:307、銅:370
- メダルの総数(団体競技も含めて)
 - オリンピック協会は2,488(812, 812, 864)個用意
- 各国の選手数
 - 11,320人



ああ、
なるほど！

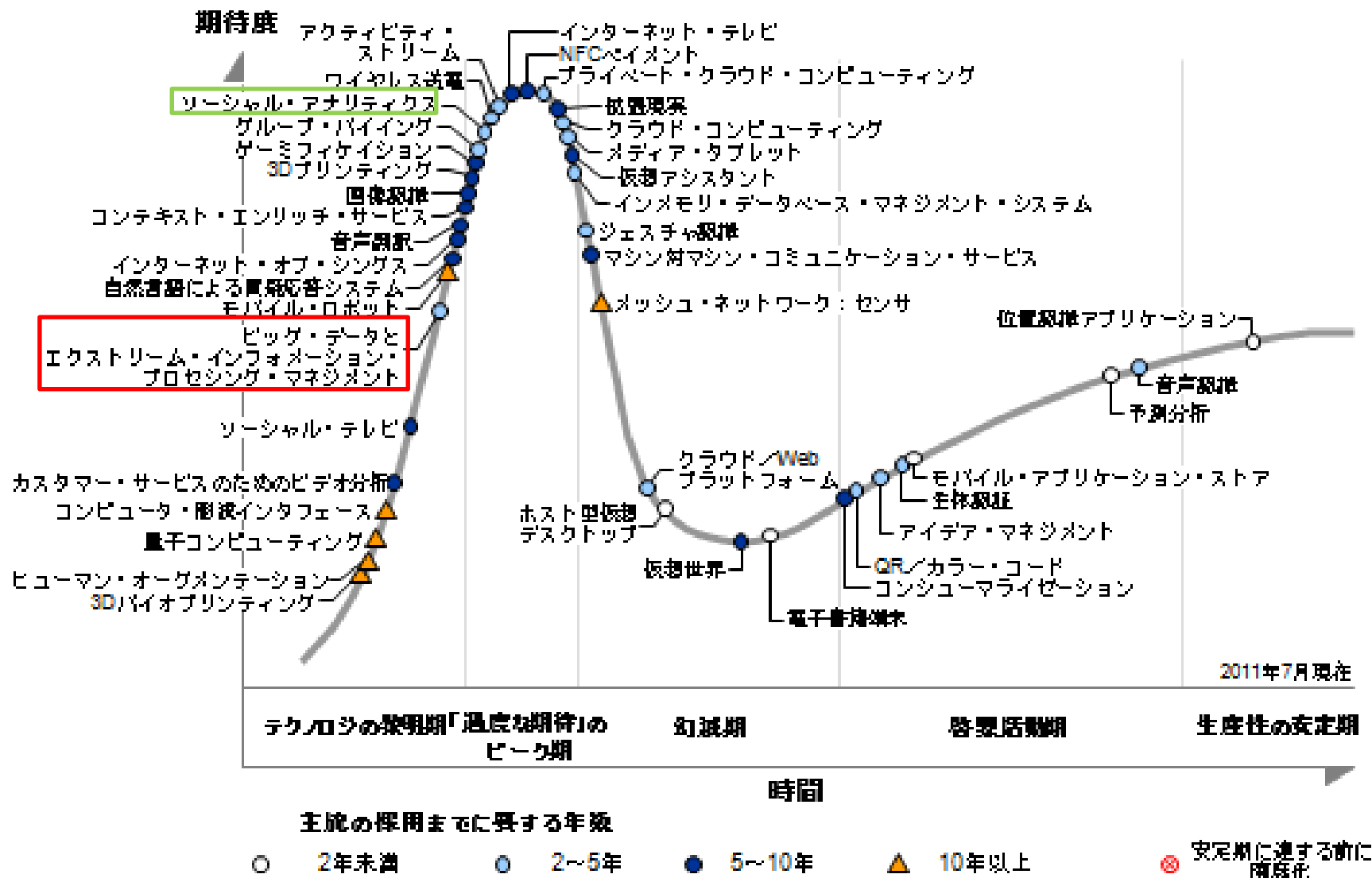
4.5人に1人はメダル獲得可能の計算

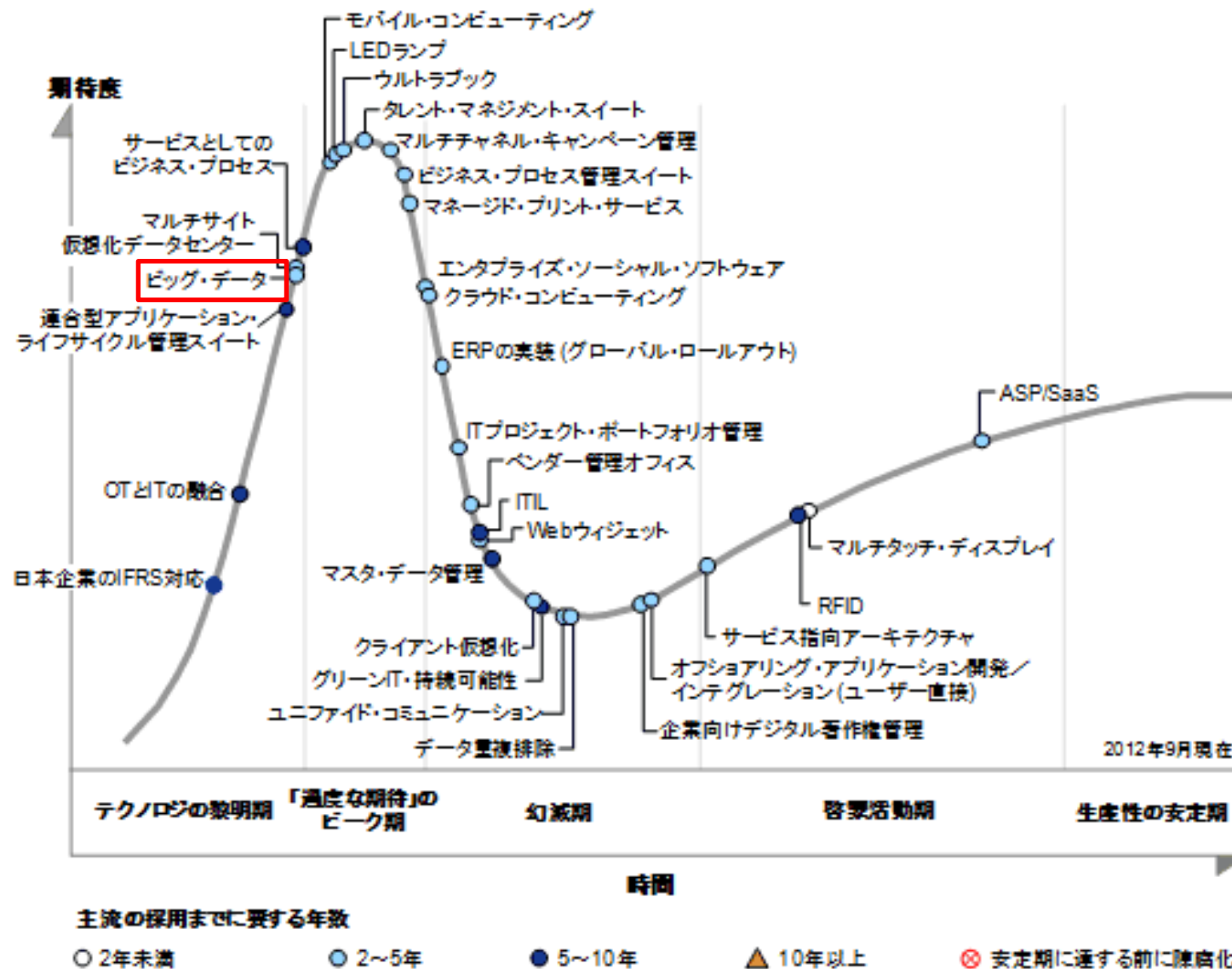
- 見つけたいことを決める
 - オリンピックでメダルを取るのはどんなに難しいのだろう？
- 必要な情報は分散している
 - 「リオデジャネイロオリンピック 28競技 306種目」
 - 「リオデジャネイロオリンピック (2016年) での国・地域別メダル受賞数一覧」
 - 「2016年リオデジャネイロ五輪のために用意されたメダルの数は2,488個」
- 分析手法を決める
 - 今回は足し算と割り算

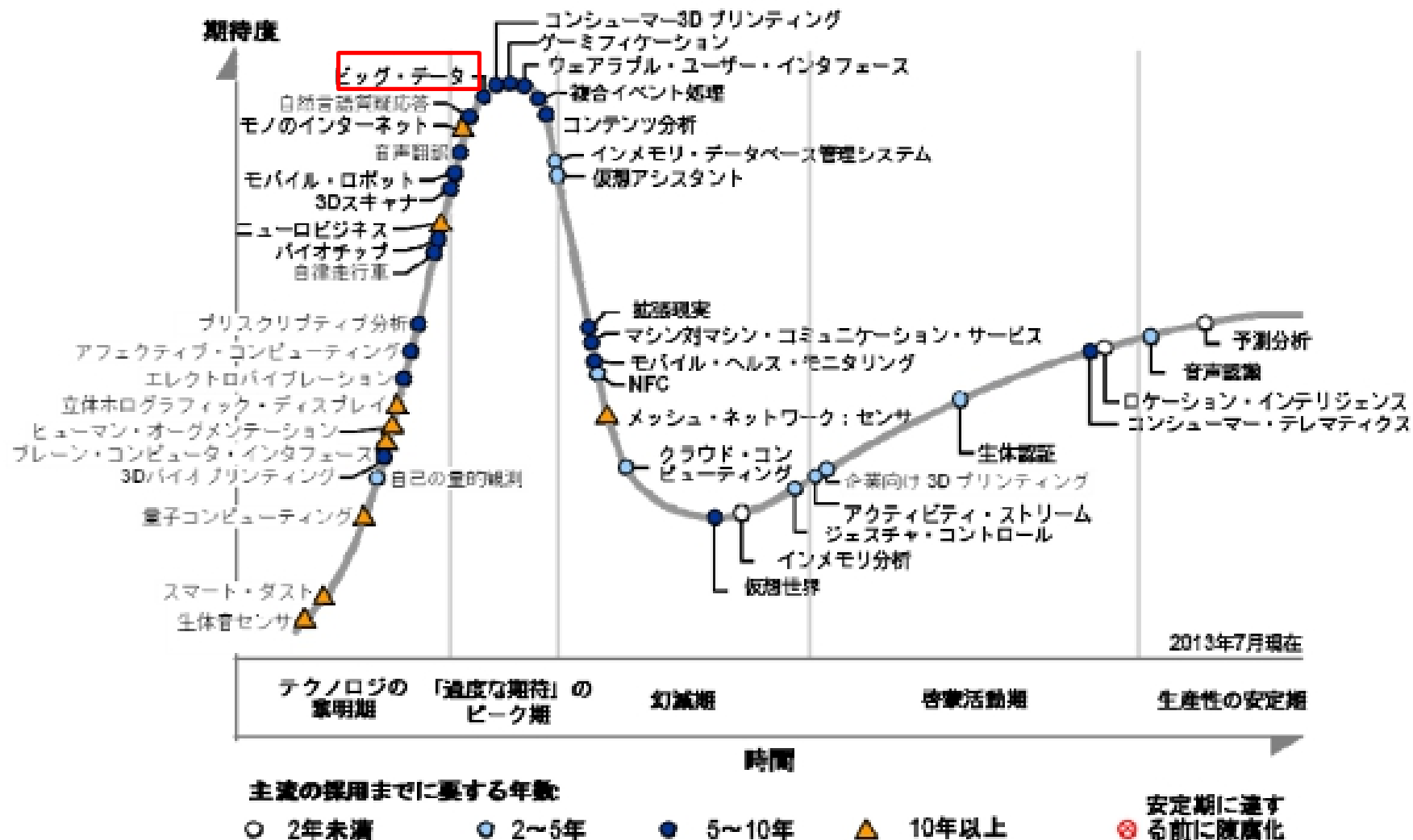
- 構成要素を明らかにするため
 - 化学実験
- 未来を予測するため
 - 天気予報、経済予測
- 新たな気づきを得るため
 - ナレッジマネジメント、見える化

なぜビッグデータか？

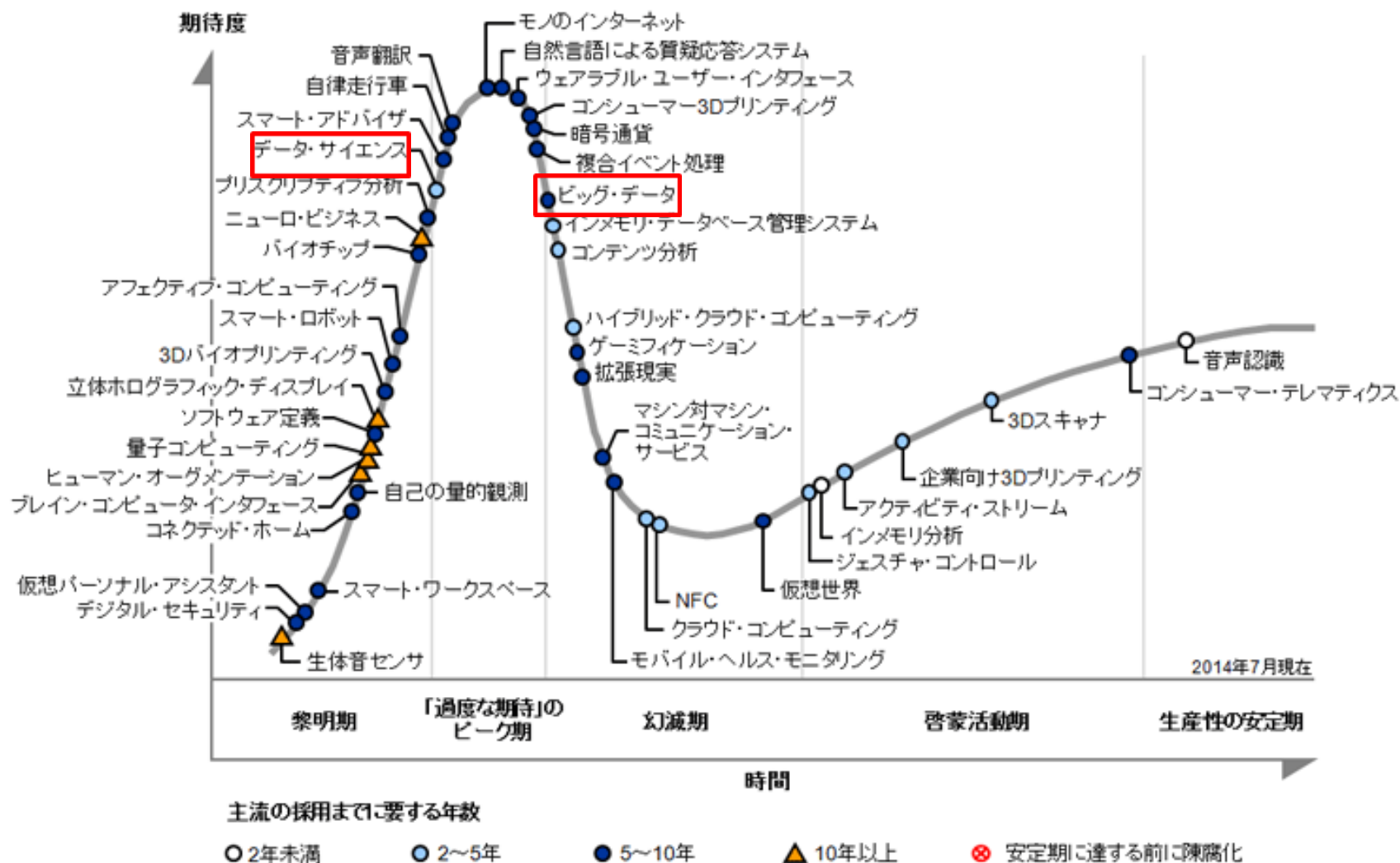
2011年のハイブサイクル(ガートナー資料から)



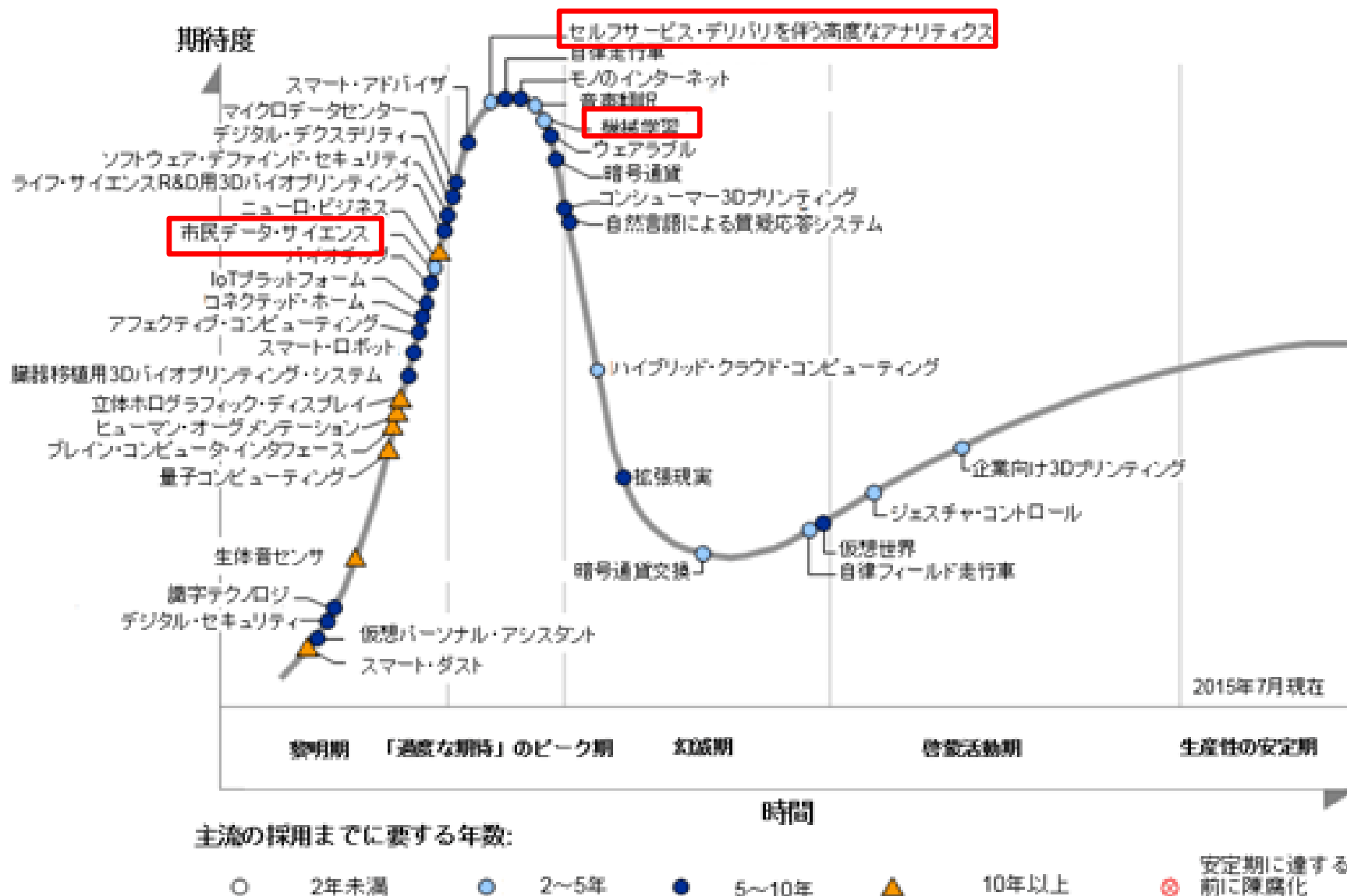




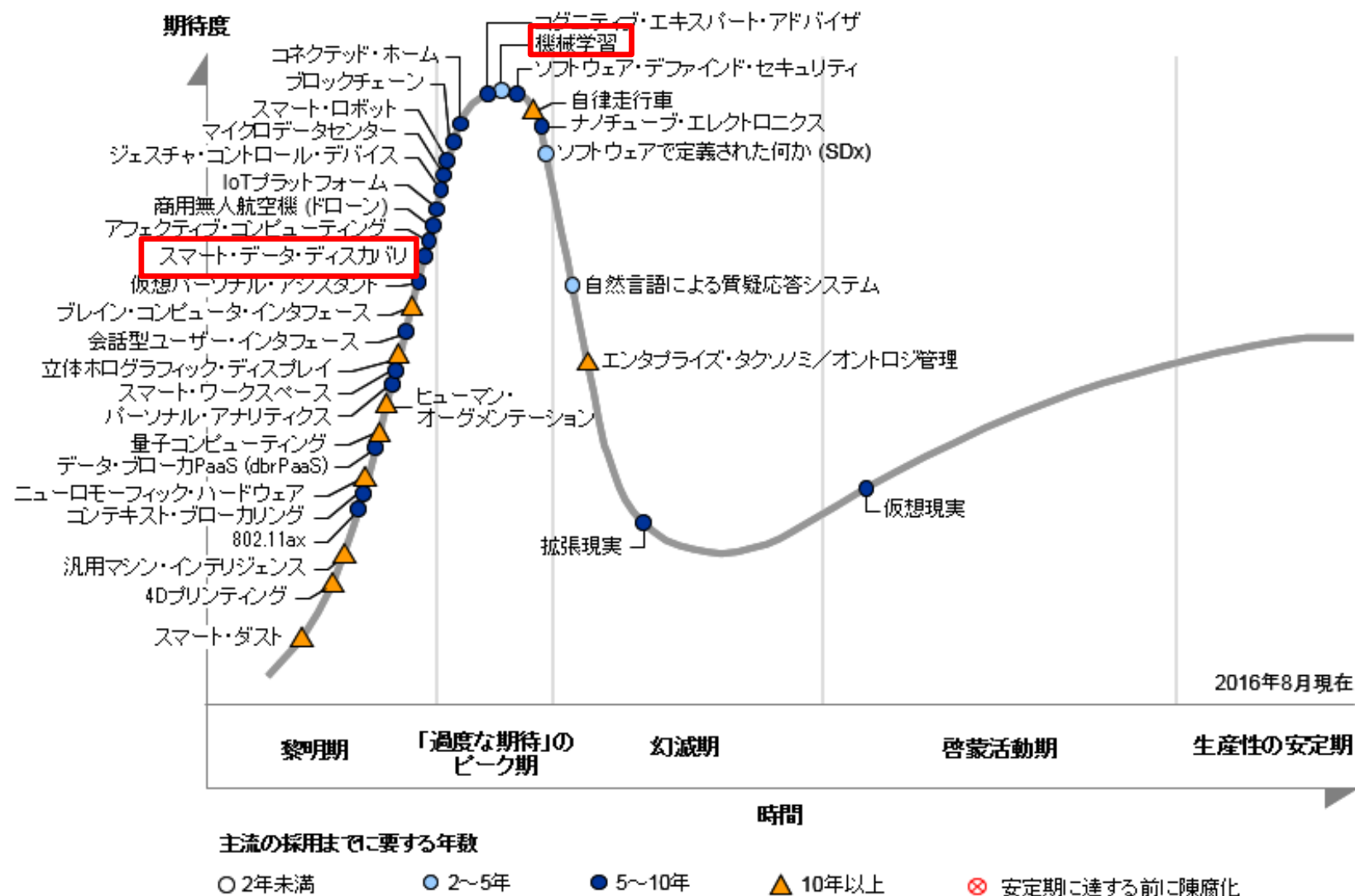
2016年のハイプサイクル



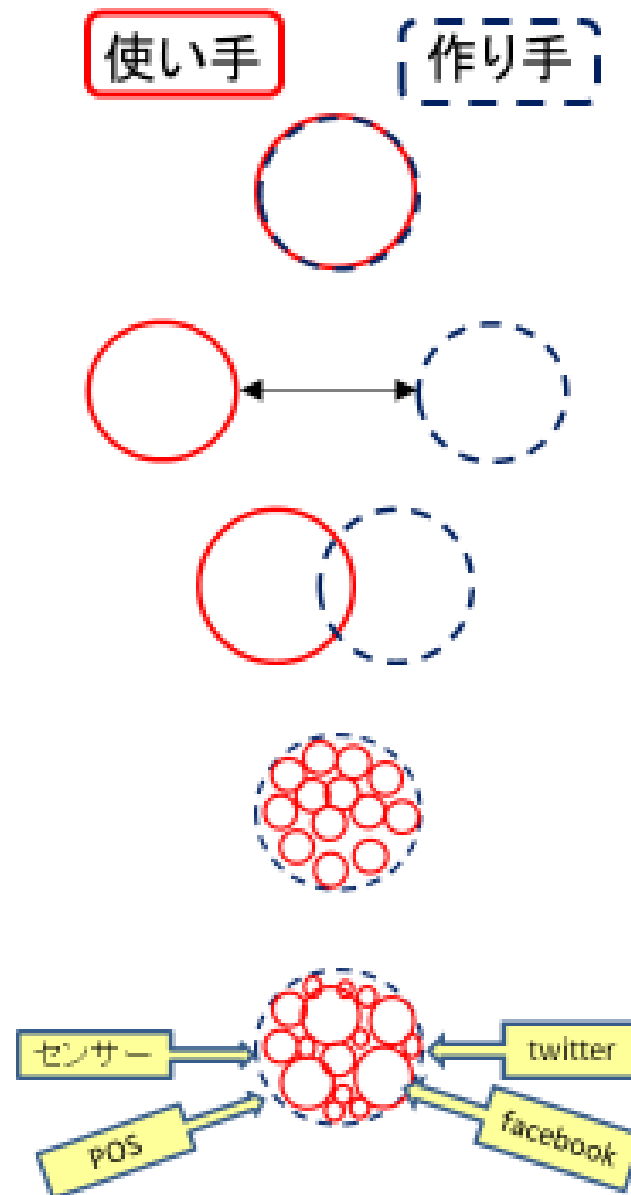
2016年のハイブサイクル



2016年のハイプサイクル



- 社会の変化
- 自給自足社会
- ↓
- 工業化社会
- ↓
- 情報化社会
- ↓
- 関係性社会
 - 使い手は同様な属性を持つtribeを形成
- ↓
- コンテキスト社会
 - tribeが時間や場所などの状況に応じて多様に変化



次につながる
重要な要因

立地^{*}

情報^{*}
(マーケティングの必要性)

関係性^{*}
(ソーシャルメディアの影響)

^{*}: 上原征彦教授の講演から

コンテキスト

ビッグデータ
リアルタイム

企業	事例	データ種類	データ規模	継続性	高精度化	必要性
成城石井	値引き決定	POS	3万アイテム	○	△	△
クックパッド	注目食材の分析	検索キーワードログ	3.5億件	×	△	×
VISA	不正利用発見	トランザクション	1.3億／日	×	×	○
楽天	商品のレコメンデーション	コマース購入履歴		○	○	○
Cyber Agent	退会兆候の行動分析	コミュニティの行動履歴	1000万人	△	○	○
がんこフードサービス	業務低下の行動分析	センサ、POS、クレーム		○	△	×
Vestas社	風力タービンの最適位置導出	気象、GPS、衛星写真	2.8PB相当	×	○	△
カカドットコム証券	株価予測配信	twitter情報	2億／月	○	△	△
ウェザーニュース	ゲリラ雷雨の予測	GPS、カメラ情報	30万人、1万／日	○	△	○
AIT航空機整備サービス	整備業務効率化	航空機整備情報	6,300万／日	△	△	△
富士通WisReed	電力需給の安定化・最適化	スマートメーター情報	960億／日	○	△	△
富士通	糖尿病予備軍の予測	レセプト、健康診断、バイタル		×	○	△

・ データを活用するポイント

- － 継続性：何回も使うか
- － 高精度：ビッグデータだから精度が上がるか
- － 必要性：そもそもビッグデータが必要か

昨年の分析対象 デイリーポータルZの紹介

Daily Portal Z オルタナ系ポータルサイト
| 検索 | 天気 | 地図 | 路線 | このサイトについて | 無題地獄 | ツールバー |

@nifty 2012.09.08 (THU)

サイト内検索 インターネット検索

今日の目玉

- 09/05 19:11 ナックルをそっくり作ればお土産になるのではありませんか。結論おもしろい。もしくは息子の偽装工作に付き合ってくれた両親の優しさか。
- 09/05 12:20 木になった状態のナスをそのまま漬物に。まじろしの漬物の木です。もいで食べるとみずみずしい。まじろしと漬物になっていて不思議な感覚。

Daily Portal Z 友の会 はじまったよ
10日まで!松コース入会でオリジナルカードもらえます

エレガンスイブ10月号「嫁姑の拳Z」との連動記事はこちらです!

Headline はっけんの水曜日

RSS リロードするたびに写真がかわります

コンビニのお菓子を東京土産と偽ってもバレないか
コンビニのお菓子でも東京土産らしい表紙を作れば東京土産としてみられなくてもいい。 (小柳健次郎) (2012.09.05 16:00)

木になったままナスの漬物をつくる
木になった状態のナスに味噌味噌を塗りそのまま漬物にする「ナスの活き漬物」を作った(圧置) (2012.09.05 11:00)

手裏剣大会で忍者を目指す!
忍者の定番武器・手裏剣の技術を磨くこの大会に出場で、ボクも忍者になれるのでは? (北村 淳) (2012.09.05 11:00)

全国で当地レトルトカレーを温ぜたら日本カレーと呼べるか
北海道のカニ、水戸の納豆など各地のカレーをひとつにしたら「日本カレー」ではなからうか(四方静香) (2012.09.04 16:00)

デパートダイエツト
涼しくて快適なウオーキングはデパートが最適。清潔なトイレもあるし、休憩用のいす完備 (林 雄司) (2012.09.04 11:00)

知らない人に話を聞く~小堀丸子これまでの記事から見知らぬ人にいろいろ教えてもらっているライターになにかコツはあるのか聞いてみました。 (小堀丸子) (2012.09.04 11:00)

遊安?フェリーで行く太平洋の旅
東京から四国までフェリーの旅。海が見える風景や四角いまくらなど、旅気分いっぱい。 (小野 浩二) (2012.09.03 16:00)

買ってその場で焼いて食べられるお肉屋さん
買った肉を軒先でバーベキューして食べられる肉屋。しかもお代は肉の代金のみ。 (尾張由見) (2012.09.03 11:00)

再放送:カステラゲーム (萩原 光明)
(初出 2009.09.27)

iPhoneをショルダーフォンにする
80年代に使われていた肩掛けの携帯電話。iPhoneケースにつけてはいないか。 (石川大樹) (2012.08.29 11:00)

インドにインド式数学を開きに行く

食卓コソバット (2012.09.05)
フープテレビ

TOKYO CULTURE CULTURE (by NIFTY)

Club クラブ活動:ダムー!!知恵袋
彼がフープバダムに行くのが楽しみです。 (09.05)

生放送:
いかにいかに

よりめきDPZ

「[金曜日食]列島の空、金のリング 広範囲、4分間の英文ショー」 MSN産経ニュース より

>>よりめきDPZ (Tumblr)

れんらくシート

お知らせ:北村さんついに珍スポットの一日店長に(9/9)

お知らせ:10周年月間やくも9月からスタート(9/8友の会、9/16エキスポ)

>>もっと細かいお知らせ

投稿で未とからおう! デイリーヘルズ

みんなの力をかしてくれ投稿コーナーが6種類 デイリーヘルズとは

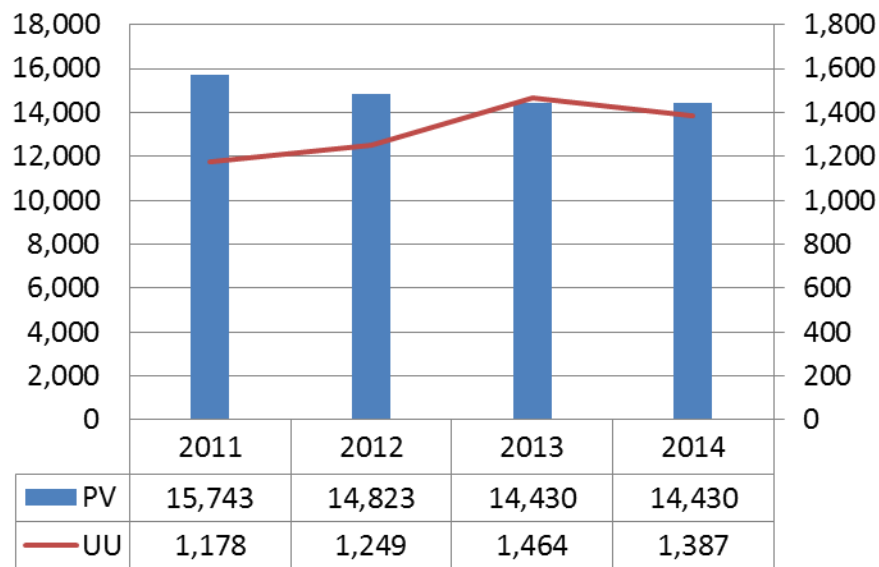
必要なのはほんのり文章ヘルズ

たまたま休載中 いらないけど捨てられないもの

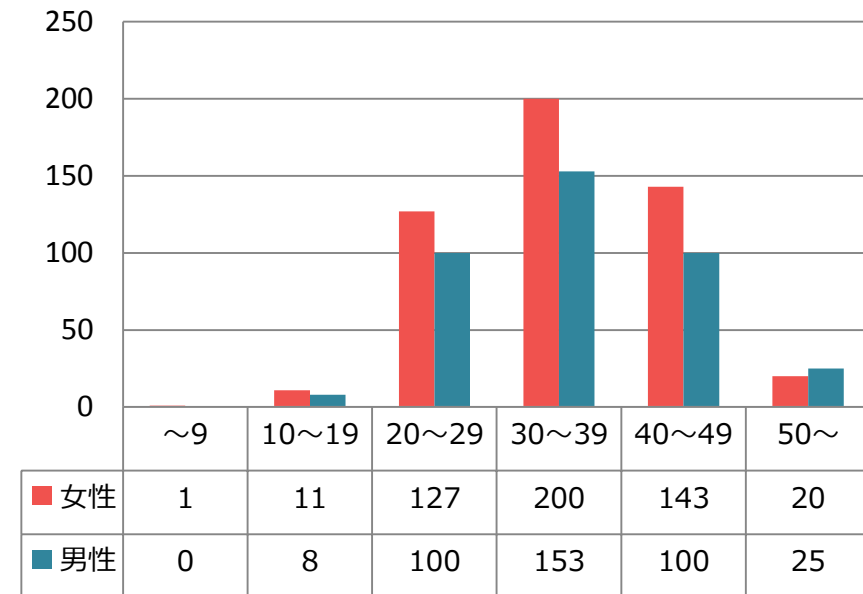
写真募集 ちょっと見たいわ 盛り上がりつつあるテーマ 伝像など

- 1日3本記事を掲載
- 珍しい場所・人
- 工作・料理
- 実験
- まじめに片足残しつつ、笑いを目指す
- 2002年10月7日開設
- 累計記事数10,000本以上

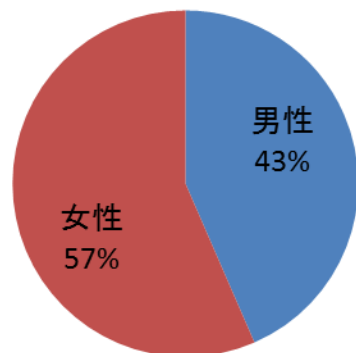
PV/UU (月平均/単位・千)



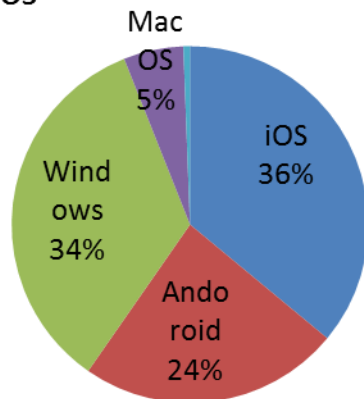
読者年齢/性別



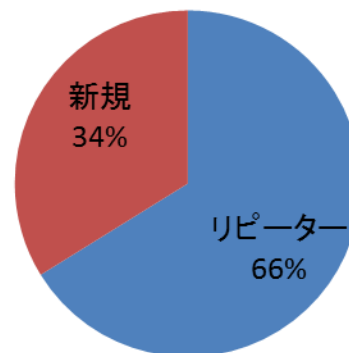
男女比



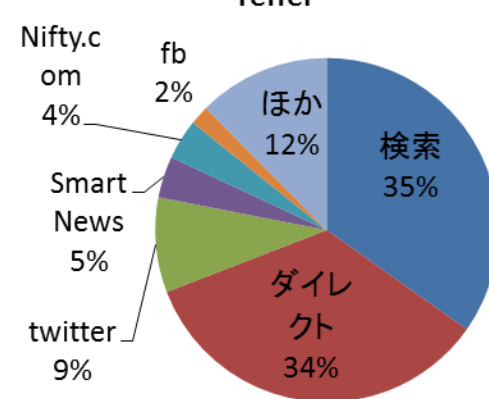
OS



リピーター/新規



referrer



コネタ508

納豆を一万回混ぜる

納豆、どれくらいかき混ぜますか？

私はほとんど練らずにサラサラで食べる方が好きなのだが、祖母など習慣的に毎回200回と決めて混ぜている。聞くところによると一般的にも年輩の方ほどよく混ぜる傾向があるそうだ。どうやら科学的にはよく混ぜた方が美味しいことが分かっているらしい。

よく混ぜるってどの程度混ぜるといいんだろう。せっかくだし、やるならコトンやってみよう。そんなわけで、1万回混ぜました。手加減ナシです。

納豆のネクストステージを見てきました。海の幸の味がしました。

(text by 古賀 及子)



フェティッシュの火曜日

知らない人の結婚式の二次会に混ざる

休日だけとスーツにネクタイで出かけた。なぜなら結婚式に出るからだ。

知らない人の。

まったく知らない人の結婚式の二次会に潜入し、どこまでなじむことができるだろうか。自分の社会性をかけてチャレンジします。

ともあれ、結婚おめでとう！知らない人。

(林 謹司)



おまえ、誰？

チャレンジの日曜日

ペリーがパワポで提案書を持ってきたら

条約締結のご提案

- 通商条約を結んでWin-Winの関係を

いまならキャンペーン中(3月末まで)

フィルモア大統領の親書つき

ぜひ我が国との交易を

※数量限定につき予定期日前に終了する場合がございます

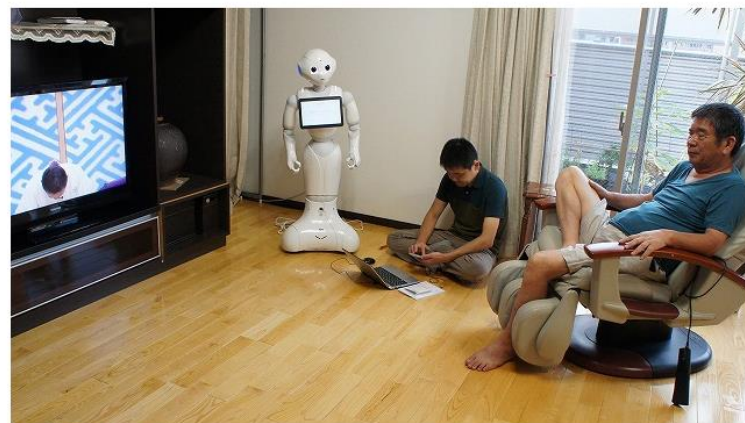
1854 Perry strictly confidential

提案の半端

ひらめきの月曜日

2015年9月14日

義父がペッパーを買った



この状況がキツい

自分が楽しいと思うものを紹介する

- 個人発信
- 好きなことなので批判にはならない
- 世間の流行は気にしない

わかりやすくサービスする

- 好きなことを書くとただマニアックになってしまう
- 興味がない人も振り向くようにアレンジする



路線図だと鉄道ファンしか見てくれないので料理のレシピ、というアレンジをした例 ↑

外に出る

- 現実で起きていることのほうが分かる人が多い。
- 趣味性の高い分野（音楽、ネットなど）はわかりにくい

失敗してもOK

- 器用な人はおもしろくない
- たいていの人は工作を失敗しているので共感される

知識よりもエピソード

- 情報よりも経験したことを書く
- 情報がなくても共感があればいい

ちしきの金曜日 2012年8月31日

サンバやったことないのに浅草サンバカーニバルに出た



本当に出了！

浅草サンバカーニバルに出場することになった。サンバはもちろん、ダンスなんて今までやったことがない。

というか、あの有名な浅草サンバカーニバル、である。出てしまっただけにいいのか。



高橋 克博
(さいとうみつひろ)
1982年、栃木県生まれ。国家資格
保持の推定師です。口の中に入れた
飲み物の温度を言い当てることがで
きる。現在はフリーランスで活動中。



コラボ企画

ローソンの空き箱でエヴァンゲリオンを作る



全く同じ

昨年の分析例

- ・ 別指標の「はてブ」を用いる
- ・ 面白い記事进行分析

まとめ

成 果

- ・ 記事が面白い $\neq$ 総PVが多い
- ・ はてブ回数が面白さを計るある程度の指標
- ・ はてブ回数から人気記事の傾向がつかめた
 - 食べ物系が拡散されやすく、
 - 固定客はアイデア、突飛な記事が好き

課 題

- ・ 面白い記事の他指標：Twitter
 - 拡散する意図があるのではてブより有効
 - 仕様変更のため断念

- 提供、家電、ストアの単語で手がかりを発見
- ビジネスに誘導する証拠はあったか

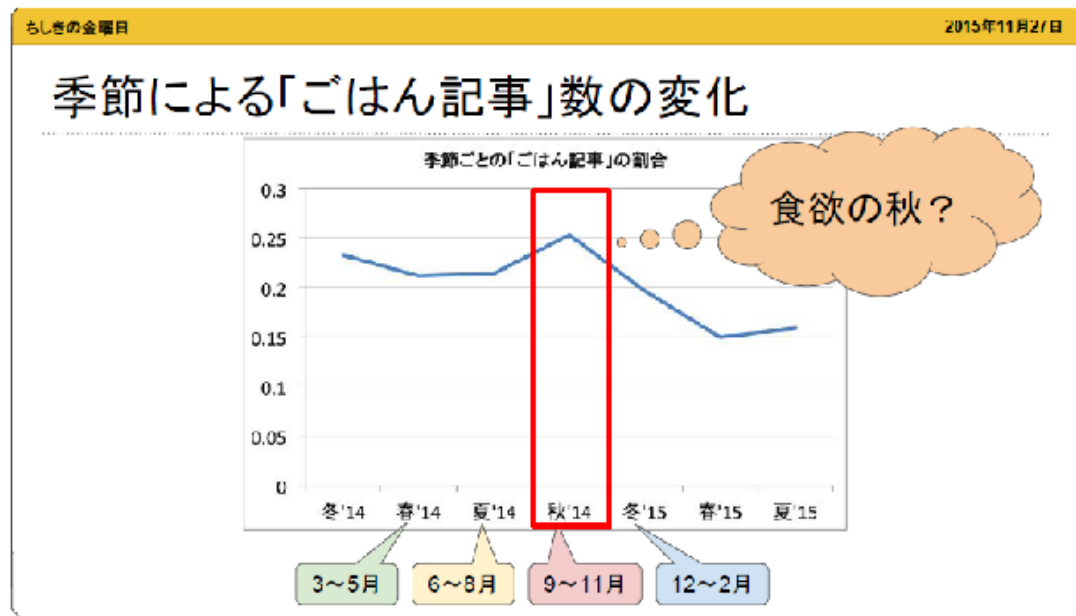
結果

- 景気悪化に伴う株価の下落
→ 宣伝、広告を目的とした単語("ストア")の増加
- 景気の回復に伴い、ワードの出現回数減少
→ デイリーポータルZストア閉店(2011/11/30)|

今後の予想

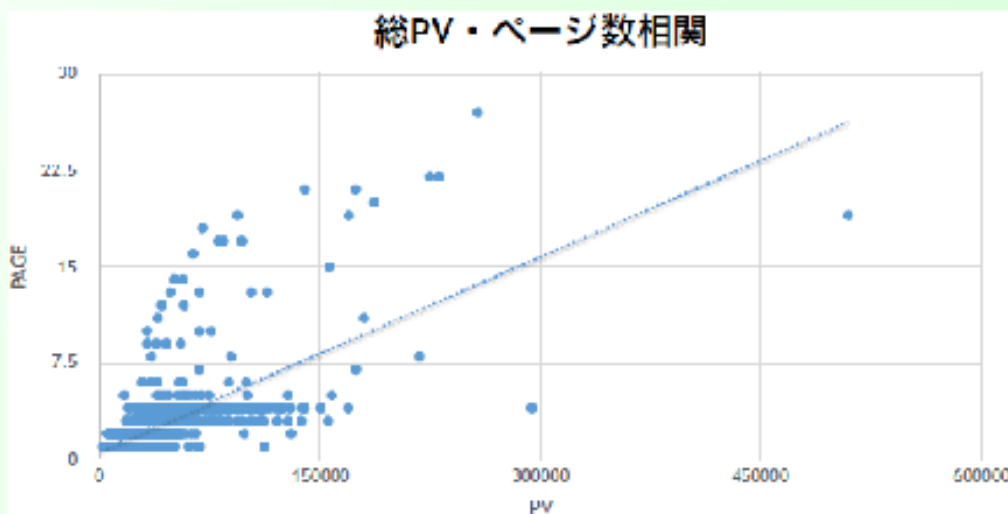
- ふたたび景気が悪化すると...
→ デイリーポータルZストア復活と予想

- 仮説1: ごはんを連想させるタイトルの記事はPV数が多い
 - 結論1: 「確かに」ごはんを連想させるタイトルの記事はPV数が多い
- 仮説2: ごはん記事を書けば人気記事になる
 - 結論2: 「残念ながら」ごはん記事を書いても必ずしも人気記事になるとは限らない
- 仮説3: ごはん記事人気はライター人気とは別
 - 結論3: 「やはり」ごはん記事人気はライター人気とは別



- PVとページ数は強い相関がある

分析：PVとページ数



強い

PV



ページ数

授業で使うデータフォーマット

例

CONTENT=BEGIN

記事のはじまり

PUBDATE=yyyy/mm/dd

公開日

TITLE=title

記事のタイトル

WRITER=writer

著者

BASEURL=url

記事のトップページurl

PV=dddd

ページビュー(公開後3日間の値)

PAGENUM=n

ページ数

PAGEURL\$=url

\$には現在のページ数を示す数字が入ります

1<= \$ <= PAGENUM以下になります

PAGE\$=xxxxx

\$ページ目の内容(htmlが保存されております)

CONTENT=END

記事の終了

CONTENT=BEGIN

PUBDATE=2008/10/1

TITLE=忍者になって人にまぎれる

WRITER=安藤昌教

BASEURL=http://portal.nifty.com/2008/10/01/b/

PV=53,929

PAGENUM=3

PAGEURL1=http://portal.nifty.com/2008/10/01/b/

PAGE1=<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">

PAGE1=<html lang="ja">

:

PAGE1=</html>

PAGEURL2=http://portal.nifty.com/2008/10/01/b//2.htm

PAGE2=<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">

PAGE2=<html lang="ja">

:

PAGE3=</body>

PAGE3=</html>

CONTENT=END

分析実習に向けて

- **デイリーポータルZを用いた仮説検証**
 - 開発用と評価用のデータを二種類用意
 - 最初に開発用データを用いて仮説を立てて、検証手法を検討
 - その後、評価用データに検証手法を適用し、仮説が成り立つかを検証
- **次の仮説を例に紹介**
 - ライターの人気は期間によらず一定である

- 検証方法：人気が一定であることを調べるために、年別の人気順の変化のあり／なしを調査
- 検証対象
 - 「2008年」を基準に2009年～2016年で比較
- 検証手法
 - 2008年のライターの順位とその他の年の順位の相関係数を計算
 - 2008年の順位と高い正の相関があれば、ライターの人気は一定と言えそうである

1. データを年別にわけると
 - 今回は年別に配布するのでここは省略
2. データの加工
 - 各年のライターの人気順を決める
3. 集計・分析処理
 - Rやプログラミング言語を利用
4. 結果の分析

- 各年の平均PV数でWRITERを順位づけ
 - WRITER名は空白を削除したり、全角の「Z」を半角の「Z」に変換するなどの処理を導入

TITLE=ファミレスのノンアルコールビール飲み比べ

WRITER=梅田カズヒコ

BASEURL=http://portal.nifty.com/2008/10/14/c/

PV=19,570

TITLE=やろうと思ったけどやんなかった記事(2008年9月)

WRITER=デイリーポータルZ編集部

BASEURL=http://portal.nifty.com/2008/10/04/c/

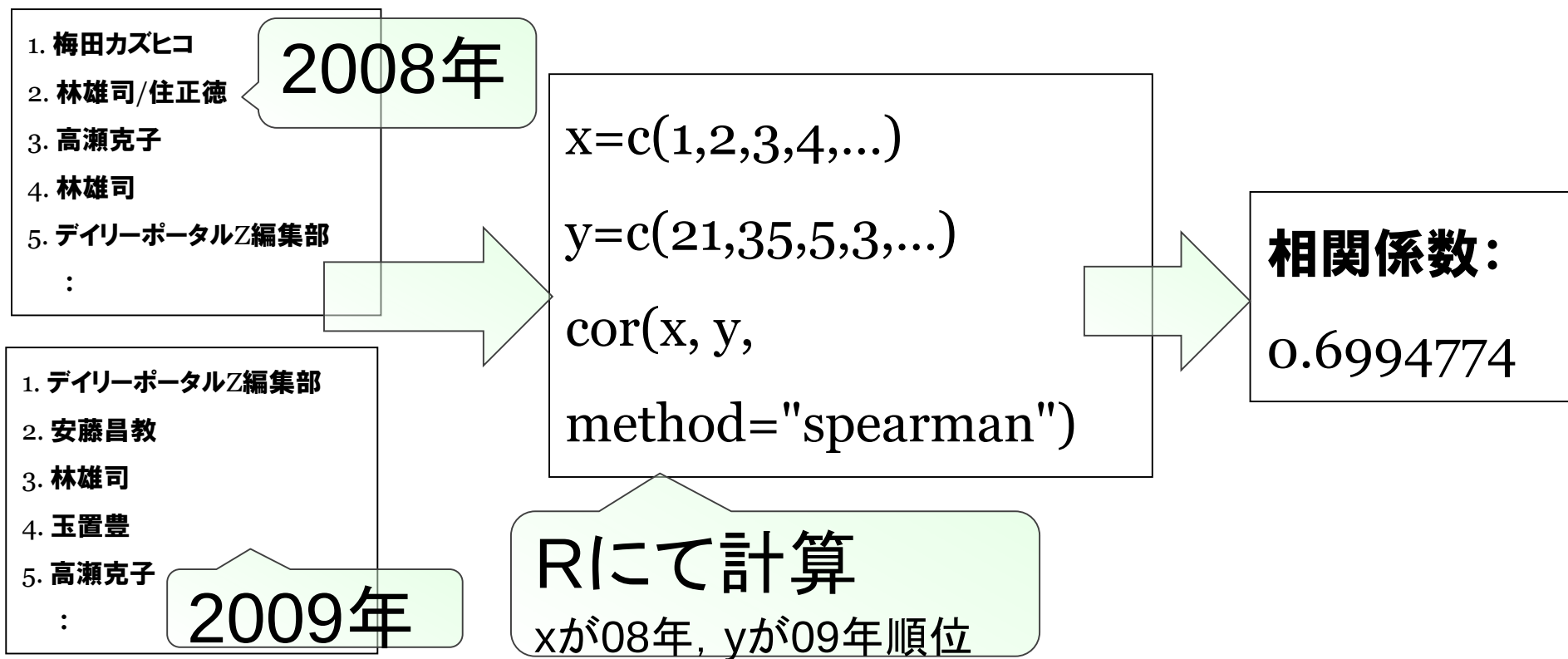
PV=93,944

平均PVで
順位づけ

1. 梅田カズヒコ
2. 林雄司/住正徳
3. 高瀬克子
4. 林雄司
5. デイリーポータルZ編集部

ライター
の順位

- 二つの年の順位の相関係数を計算
 - 2つの年で記事を書いているWRITERを対象
 - 今回は、Rを用いて計算



- 表は08年と15年までの順位の相関係数
 - 08年は当然ながら1。年が近いほど相関が高い。
 - 09年は0.7に近く、かなり相関が高い
 - 14年から負の相関となっており、続くわけではなさそう

比較対象	相関係数
08	1
09	0.699477
10	0.697193
11	0.596257
12	0.125807
13	0.31746
14	-0.16783
15	-0.1087

- 仮説：記事をページビュー数に並び替えるために要素はXXX
- 仮説：ページビューを予測するための要素はYYY

次回(11月25日)までやること

- 自由課題（8チームに分けます）
 - 各チームそれぞれが、仮説作成・検証手法の検討までの一連の分析を行なう
 - 分析の結果を通して知見をまとめ、発表資料を作成
- データセットは二種類用意：開発用で仮説検証のための手法を開発し、評価用データに適用し、評価
 - 開発用データ（11月11日公開）
 - 8,358記事（2008年10月1日から2016年8月30日まで）
 - 評価用データ（11月18日公開）
 - 1,285記事（2016年8月31日から2016年9月11日まで）

- ピッチ
 - 投資家に行なうプレゼン
 - 5ー6分間程度
 - 厳しい質問後、その場で投資決定
- 内容
 - 困っている問題
 - 今、何が起きているか
 - Disruptすべき既存の仕組み
 - 解決策
 - 誰がユーザか？
 - どんな方法で？
 - 特徴
 - Better, Faster, Cheaper
 - 主観／客観で見た市場、ビジネスモデル
 - チーム紹介
 - 役割、実績
 - 解決可能なチーム



- 自由課題のプレゼンテーション
 - データ・サイエンティストのつもりで
 - 5分間
- 内容
 - 仮説
 - 仮説立案
 - 仮説の検証方法
 - 分析手法
 - 対象とするデータ決め
 - 集計・分析・可視化
 - 比較・深掘り
 - いくつかの分析の比較、工夫点
 - 結果
 - できたこと、できなかったことの考察・検証
- 優秀賞
 - 優秀な発表を行なったグループに特製グッズを差し上げます