オープンキャンパス2013工学部コース講義

Are you Sexy?
～データサイエンス入門～

東京大学大学院工学系研究科
鳥海不二夫

自己紹介

- 鳥海不二夫
- 東京大学大学院工学系研究科システム創成学専攻准教授
- 研究テーマ
  - ソーシャルメディア，金融情報学，集合知
  - データマイニング，社会シミュレーション

問題1

- 「サバの缶詰を食べるとダイエットに効く」
- テレビ番組での実験
  - 5人の被験者が一週間毎日サバ缶を食べた
  - 5人中4人の体重が減少！
  - サバ缶すごい！
  - サバ缶売り切れ続出

- 果たしてサバ缶は本当にダイエットに効くのか？

サバ缶がダイエットに効いている確率

- 5人中4人の体重が「偶然」減る可能性はどのくらいあるのか？
  - 50%の確率で体重が増減するとする
- 5人中4人の体重が偶然減少する確率
  - $P = 18.75\%$
  - 実はそんなに低い確率でもない
- 偶然における確率が5%以上の場合はあまり信用できない

問題2

- センター試験の結果から東京大学理科一類に合格するかどうかを判定
  - センター試験の受験者は57万人
  - 定員1108人
- 精度は99%
  - 合格者を合格と判断し，不合格者を不合格と判断する割合
- 合格判定が出たらほぼ確実に受かる？

合格と判断される条件

- 合格と判定される人はどんな人か？
  - 合格者に合格と判断する数
    - $N_s = 1108 \times$ 精度
  - 不合格者に合格と判断する数
    - $N_f = 569900 \times (1 -$ 精度$)$
- 合格と判断されたときの本当の合格率
  - $p = \frac{N_s}{N_s + N_f} = \frac{1096.92}{1096.92 + 5699} =$ **16.1%**

なお，この問題は説明用に作成したもので
実際の判定とは異なります

データサイエンティストは21世紀もっともSEXYな職業である

魅力的な

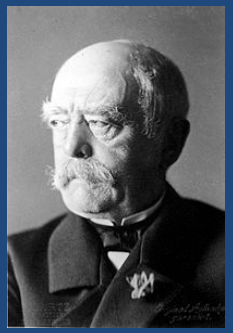
Thomas H. Davenport(1954-)

目次

- なぜデータサイエンスなのか
- データ分析の例
  - マーケティング
  - 観光業界
  - スポーツ科学
- データサイエンティストを目指して

愚者は経験に学び
賢者は歴史に学ぶ

オットー・フォン・ビスマルク(1815-1898)



経験vsデータ

- 経験：深いがその範囲は狭い
  - 一部の情報のみ取得可能
  - 主観性が強い
- データ：カバーする範囲が広い
  - 一般的な情報を取得可能
  - 客観性が強い

データ分析の重要性

1.網羅性による説得力

2.主観に左右されない

3.新しい事実の発見

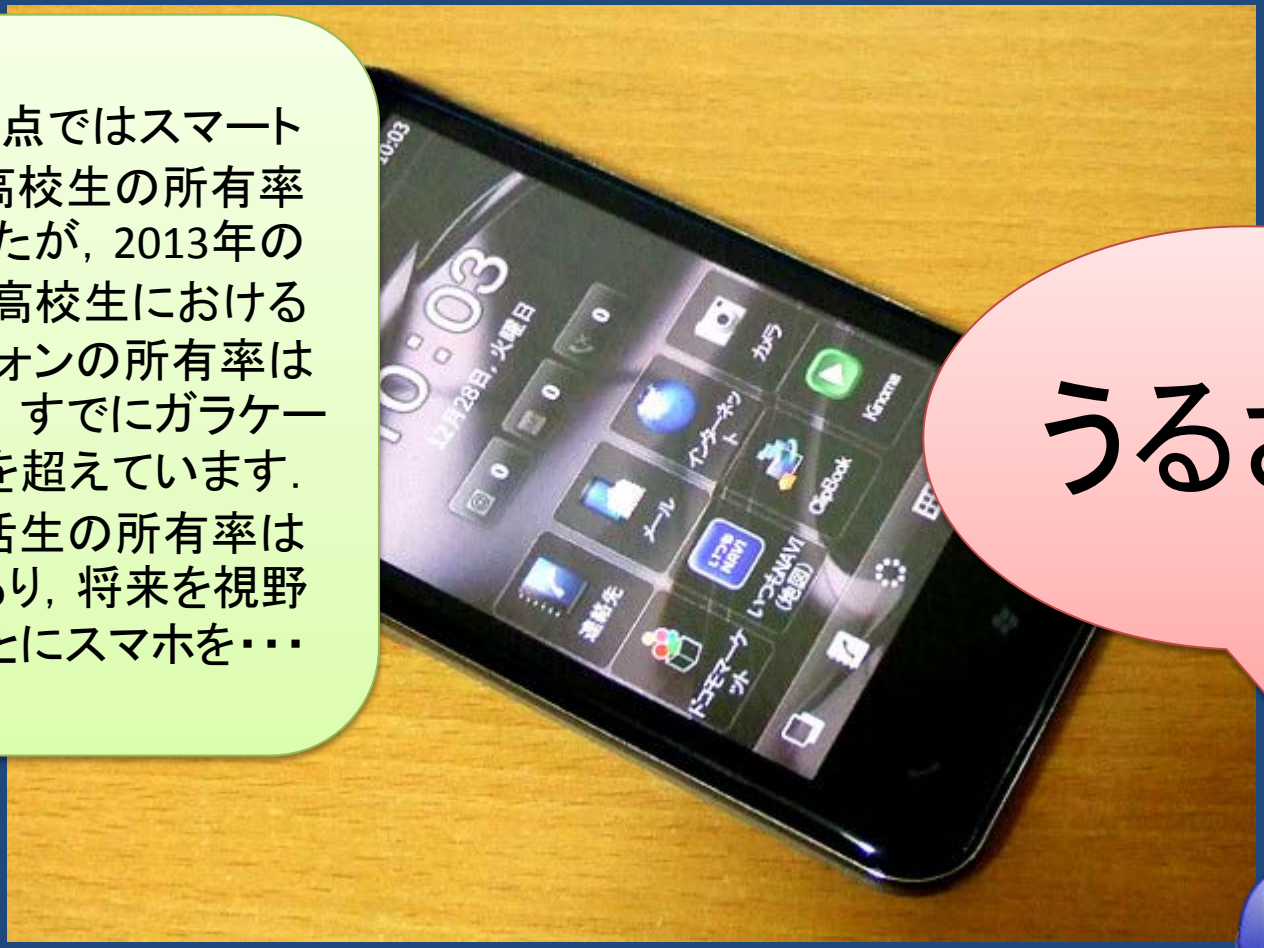
経験による分析

皆持ってるからスマホ買って！

みんなって誰？

データによる分析

2012年時点ではスマートフォンの高校生の所有率が7%でしたが，2013年の調査では高校生におけるスマートフォンの所有率は56%であり，すでにガラケーの所有率を超えています．さらに就活生の所有率は83.9%であり，将来を視野に入れるとにスマホを・・・



うるさい

データ分析の重要性

1.網羅性による説得力

2.主観に左右されない

3.新しい事実の発見

どちらの死亡数が多い？

• 脳卒中	• （あらゆる）事故
• 竜巻	• 喘息
• 落雷	• 食中毒
• 病死	• 事故死
• 事故死	• 糖尿病

2倍

20倍

52倍

18倍

4倍

データ分析の重要性

1.網羅性による説得力

2.主観に左右されない

3.新しい事実の発見

野菜ジュースはどこへ置く？

- スーパーで野菜ジュースはどこに置くと一番売れるか？
  1. 飲み物売り場
  2. お弁当売り場
  3. カップラーメン売り場
  4. 野菜売り場
  5. レジの前

野菜を買った主婦が
ついでにジュースも買う

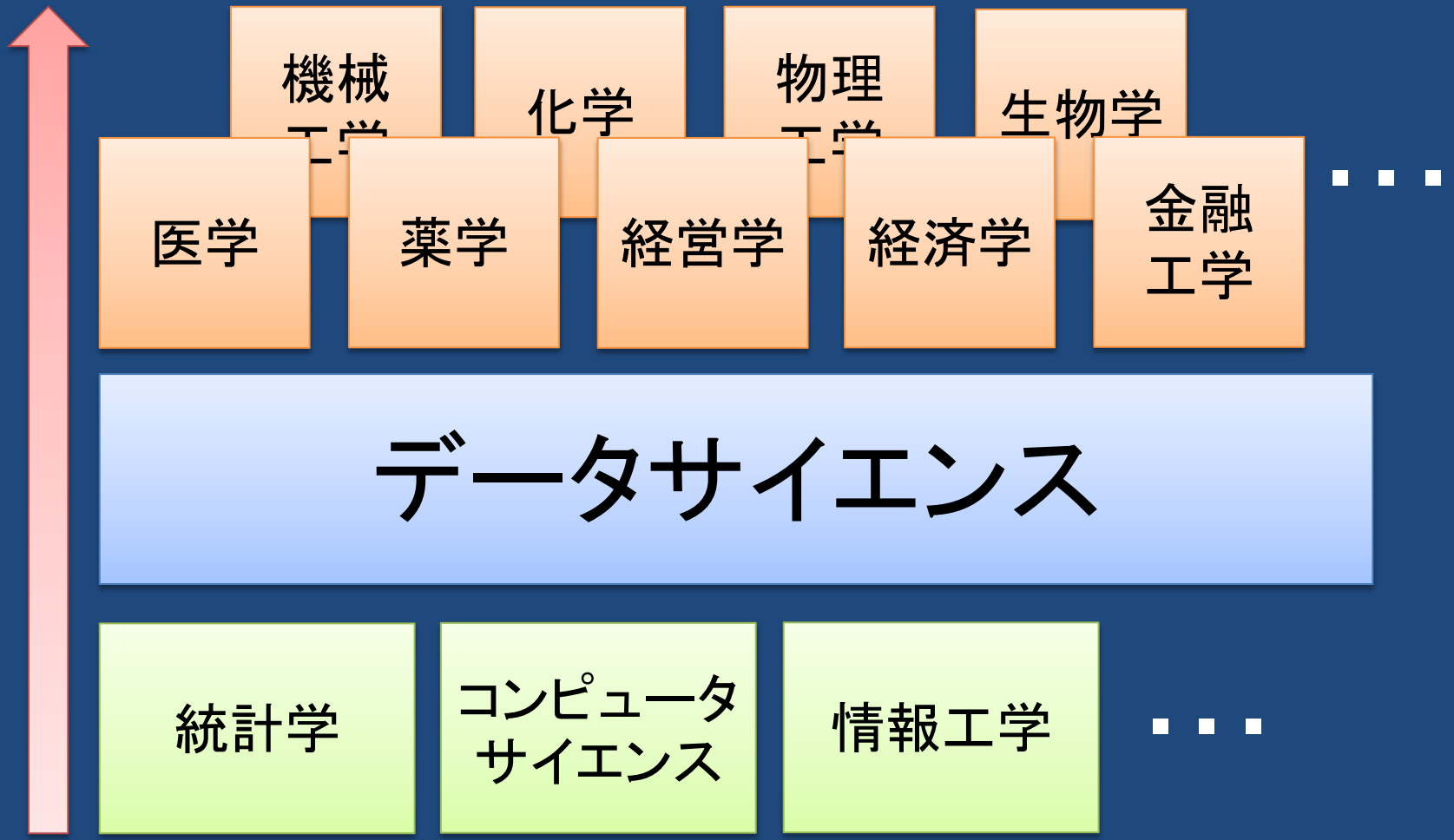
Data Science is Everywhere

データサイエンスが使われる分野

- 自然科学
  - 医学，薬学，化学，地球科学，宇宙科学
- 人文科学
  - 心理学，言語学，教育学，考古学，スポーツ科学
- 社会科学
  - 経営学，経済学，商学，金融
- 応用化学
  - 工学，図書館情報学，交通科学，メディア科学

などなど・・・

データサイエンスから見た科学



目次

- なぜデータサイエンスなのか
- データ分析の例
  - マーケティング
  - 観光業界
  - スポーツ科学
- データサイエンティストを目指して

データ分析の例

- マーケティングにおける分析
  - POSデータの場合
- 観光業界の例
  - じゃらんの場合
  - ソーシャルメディアを利用した場合
- スポーツ科学における例
  - セイバーメトリクス
  - 大相撲における八百長問題

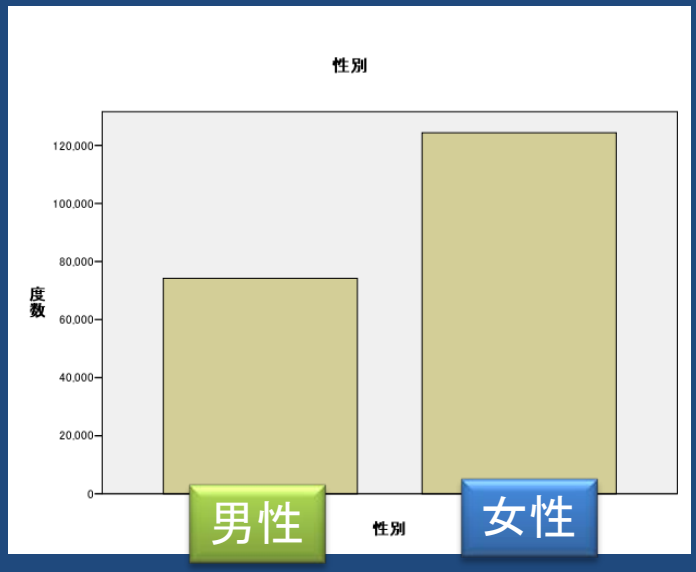
マーケティング
における
データ分析

POSデータの分析

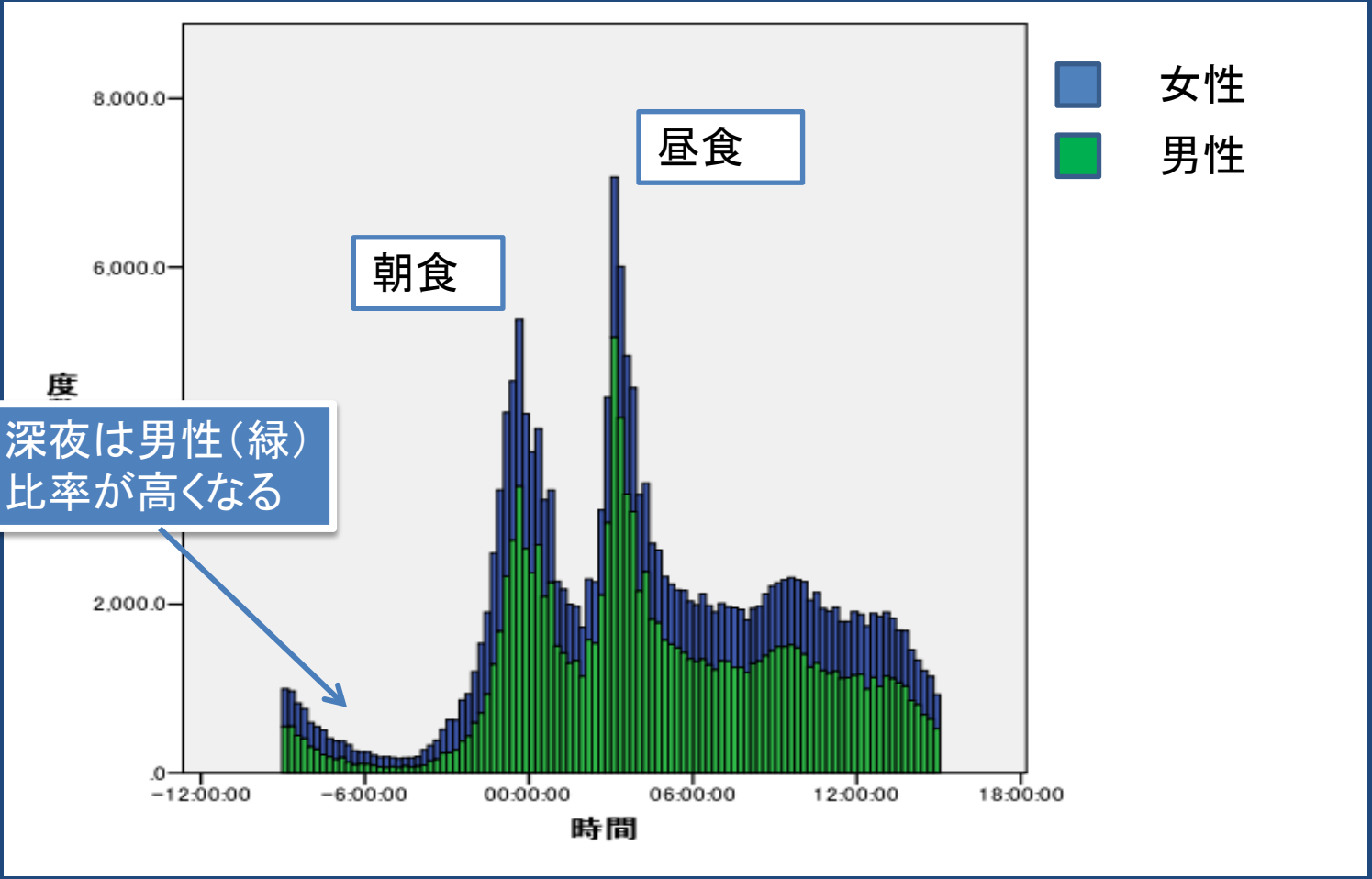
- 電子マネーやポイントカード等の普及
  - T-Pointカード，Edy，Suica
- 消費者毎の購買履歴を追跡
  - 曜日（平日/休日）・時間帯（早朝/昼間/夜間）
  - 共購買パターン
  - 消費者の継続した購買履歴

コンビニエンスストアのID-POS

- 7ヶ月間分のデータを分析
  - 秒単位で購買日時を記録
  - 継続して同じ店を利用
- 人数：199,491人
  - 女性が若干多い
- レシート数：2,339,701枚
  - 平均購買金額：約550円



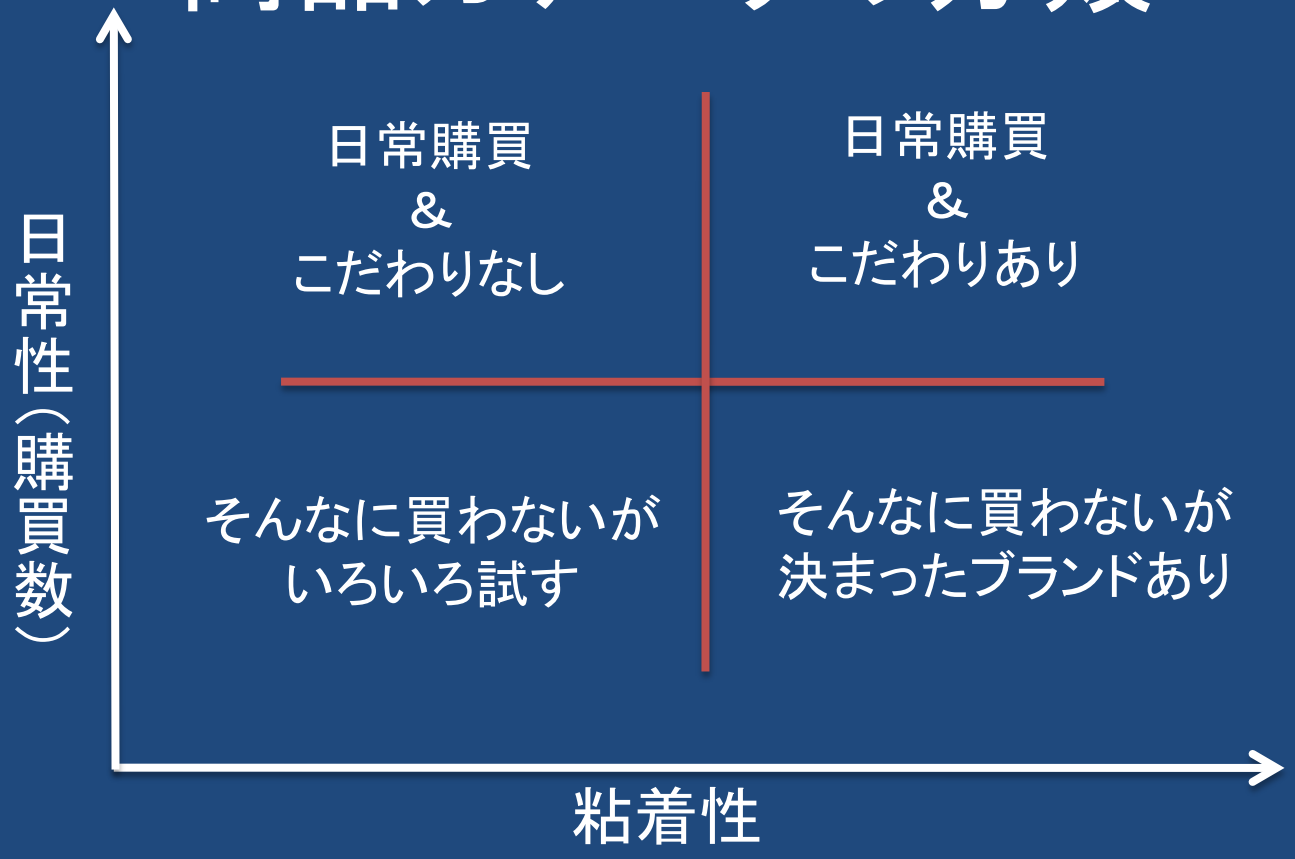
来店時間（性別）


8,000.0
6,000.0
2,000.0
.0
度
−12:00:00
−6:00:00
00:00:00
06:00:00
12:00:00
18:00:00
時間
女性
男性
昼食
朝食
深夜は男性（緑）
比率が高くなる

購買パターンの分析

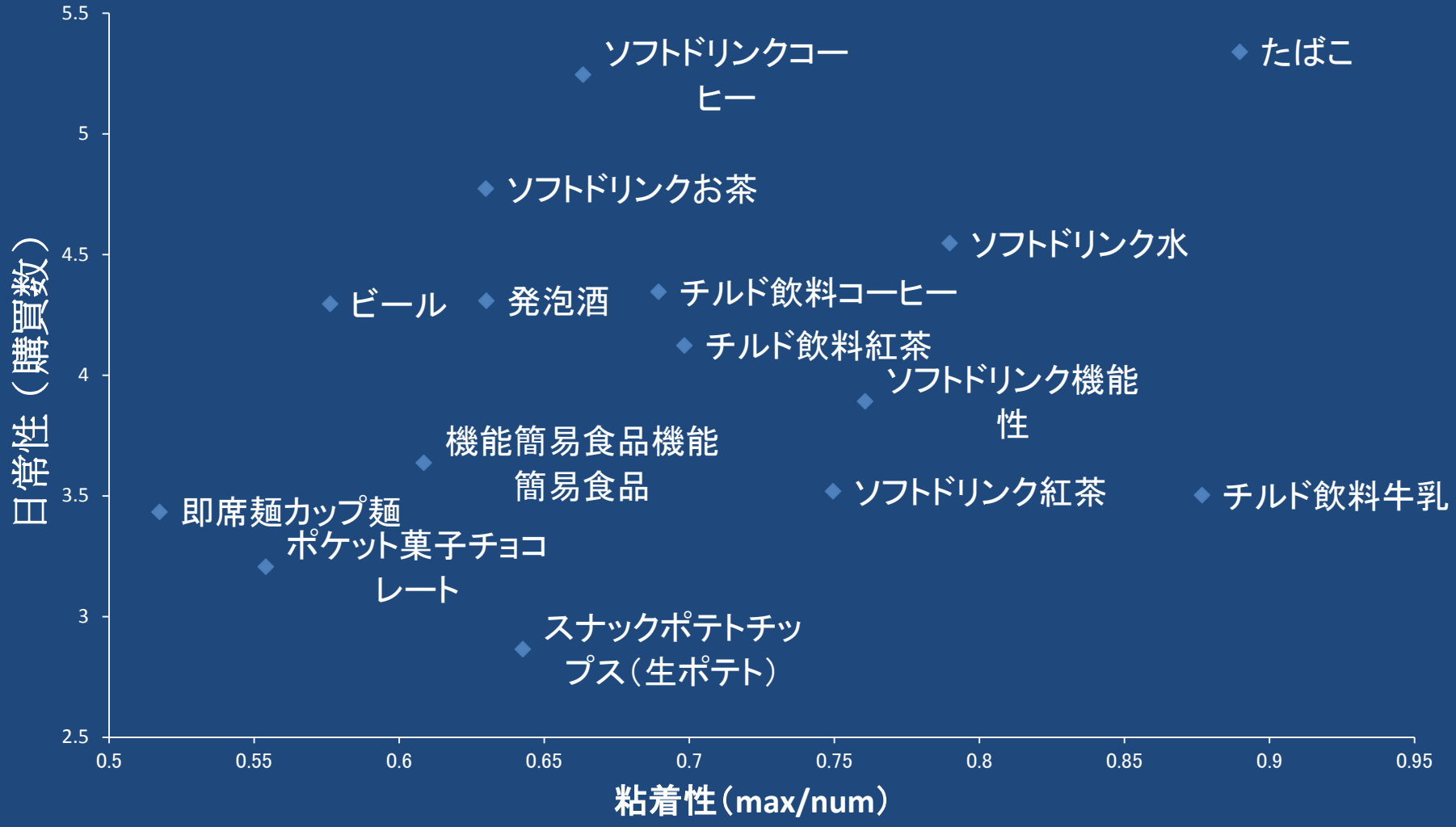
- 商品ごとの購買パターンを分析
  - 商品の買われ方にパターンを見いだす
- 特定の銘柄を好むような商品 → **粘着性**
  - こだわりのある商品
- 日常的に購入する商品 → **日常性**
  - よく売れる商品

商品カテゴリの分類



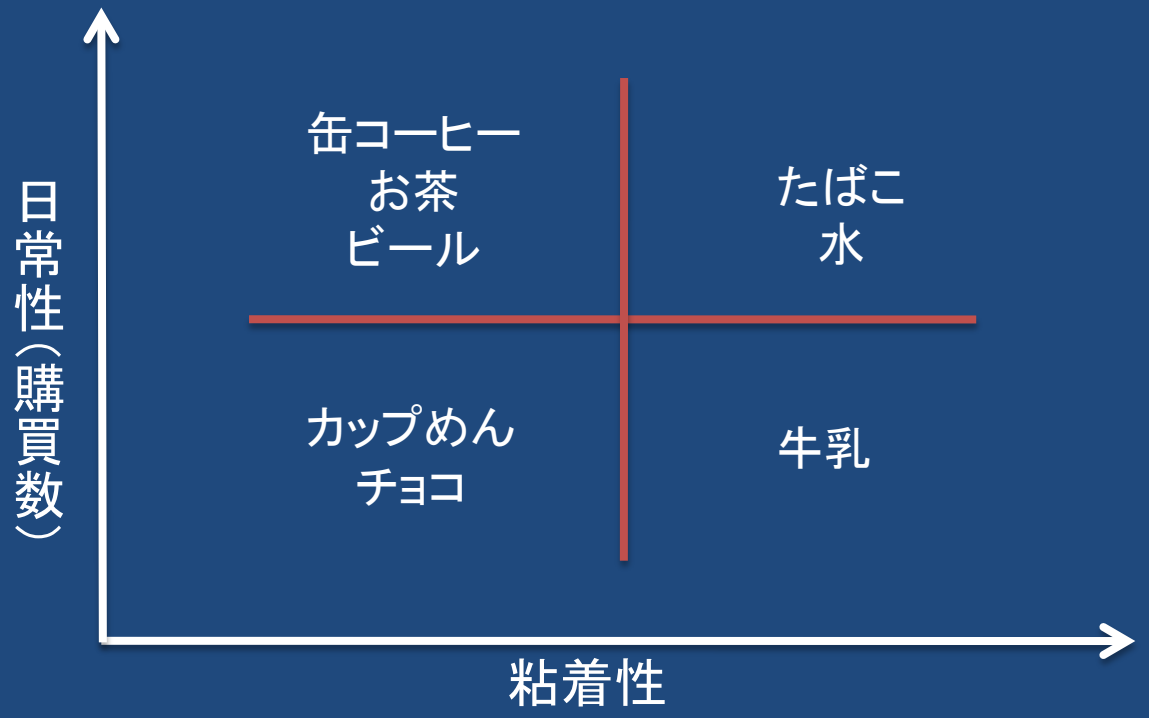
どんな商品カテゴリがどのタイプに分類されるのか

粘着性×日常性


日常性（購買数）
粘着性（max/num）
5.5
5
4.5
4
3.5
3
2.5
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
ソフトドリンクコーヒー
たばこ
ソフトドリンクお茶
ソフトドリンク水
ビール
発泡酒
チルド飲料コーヒー
チルド飲料紅茶
ソフトドリンク機能性
機能簡易食品機能簡易食品
ソフトドリンク紅茶
チルド飲料牛乳
即席麺カップ麺
ポケット菓子チョコレート
スナックポテトチップス（生ポテト）

分析結果の考察

- こだわりが高い商品
  - たばこ・水・牛乳
- お茶と水の粘着性に違い
  - お茶は色々試すが，水は特定のものしか買わない！
  - お茶の新商品は売りやすい (よく売れるし，色々試される)

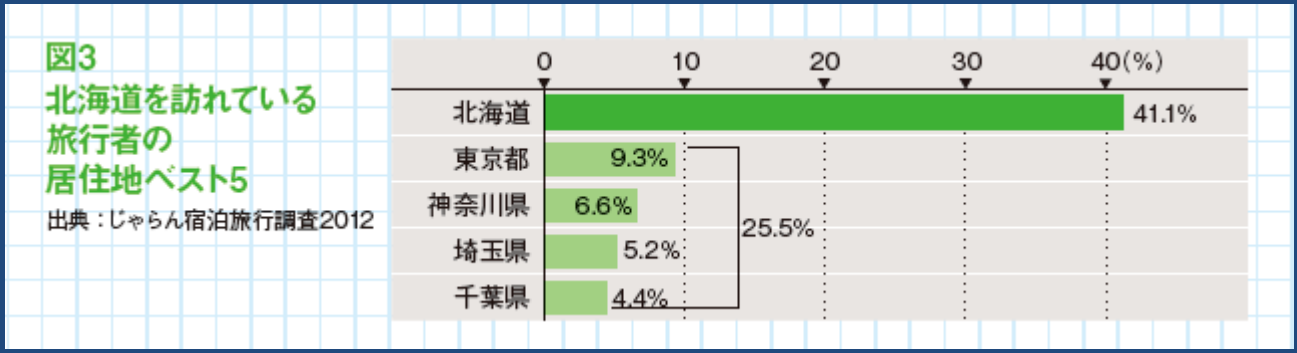


観光業界
における
データ分析

じゃらんによる旅行者分析

- 利用データ
  - 個人旅行者の宿泊地
  - コロプラ(位置登録アプリの利用)
- 北海道旅行を分析
  - 誰が北海道を旅行するのか？
  - どこを旅行するのか？
  - あらたな観光スポットは発見できるか？

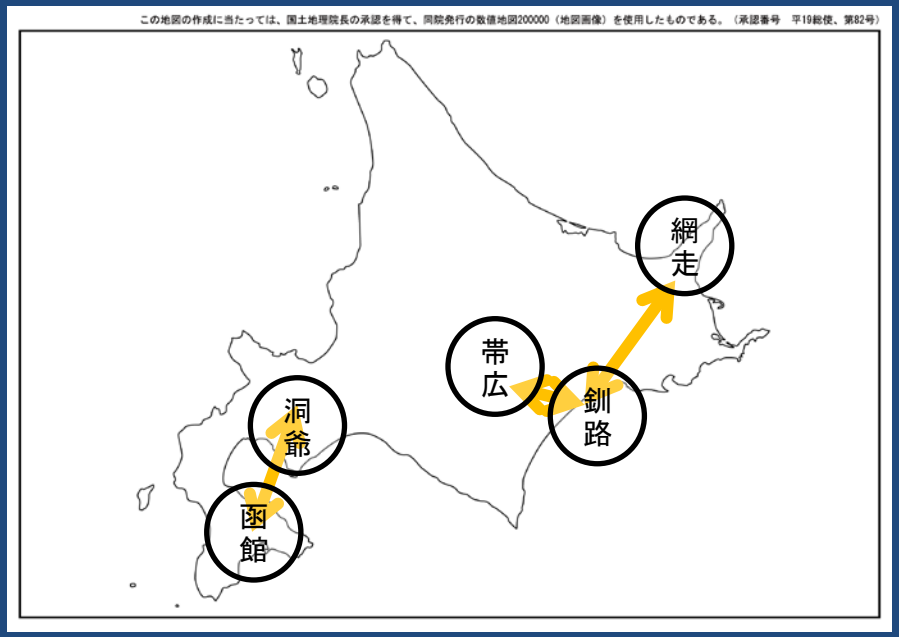
誰がどこに旅行しているのか



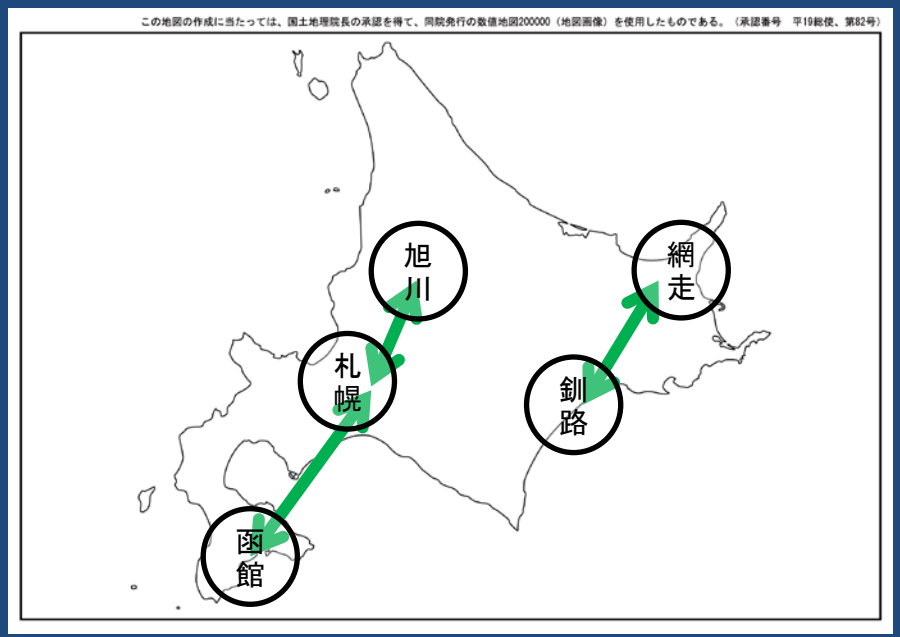
- 北海道旅行者の40%以上が北海道在住者
  - 南関東からの旅行者は25.5%
- 北海道の観光産業は北海道在住者をターゲットにするべき
  - かもしれない

主な旅行エリア

北海道在住者



南関東在住者



どこに在住しているかで勧めるべき観光スポットが異なる

穴場ポイント

表4　釧路・阿寒・根室・川湯・屈斜路エリアで冬に位置登録された場所

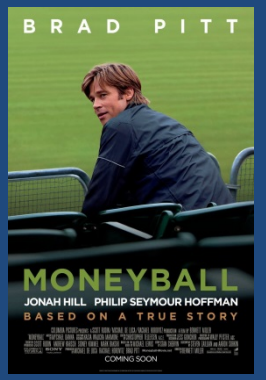
釧路・阿寒・根室・川湯・屈斜路		
	南関東居住者	北海道居住者
1	阿寒温泉	幣舞橋
2	釧路フィッシャーマンズワーフMOO	阿寒温泉
3	和商市場	和商市場
4	阿寒湖遊覧船	阿寒湖遊覧船
5	幣舞橋	川湯温泉
6		丹頂鶴自然温泉
7		美幌峠
8		硫黄山麓の噴煙
9		釧路市こども遊学館

- 北海道の人はよく訪れるが関東の人はほとんど訪れない
- まだ知られていない穴場かも

スポーツ業界におけるデータ分析

セイバーメトリクス

- スポーツデータを統計的・客観的に評価
  - お金を使わずに強いチームを作る
- オーランド・アスレチクスが導入
  - 出塁率，選球眼などを重視
  - バイト，盗塁は重要視しない
  - 安い年俸総額で全球団で最高の勝率
- 映画「マネーボール」で有名



大相撲の八百長の発見

- 7勝7敗の力士が8勝6敗の力士に勝つ割合
  - 79.6% (確率的には48.7%)
  - 八百長報道があった直後はほぼ50%
  - 「負けてあげた」力士の次の場所の勝率：60%

シカゴ大学経済学部教授スティーブン・レベット博士

大相撲の八百長の発見

- 1980年に出たとある論文の記述
  - 7勝7敗で千秋楽を迎えた幕内力士は35人いたが28人が勝ち越した.
  - 瀬戸際の力士は並外れて「強く」なり，簡単なモデルでは表現しきれない

宮川 雅巳, 鳩山 由起夫
強さと試合形式の合理性,
Operations research as a management science Research, Vol.25 No.10 pp.649-657 (1980)

目次

- なぜデータサイエンスなのか
- データ分析の例
  - マーケティング
  - 観光業界
  - スポーツ科学
- データサイエンティストを目指して

なぜ今データサイエンスなのか

- ビッグデータ
  - 多くの分野でデータが大量に存在
    - 数億にも及ぶYoutube上に存在する動画
    - 数百万人分のSUICAの利用記録
    - 何年にもわたるスポーツデータの蓄積
    - 何万人分の臨床データ
- IT技術の発展
  - データ分析手法の開発
  - 大量データを処理可能な技術

データサイエンスに求められるもの

- 大量のデータの収集運用管理
  - Facebook：11600コンテンツ/秒
  - Youtube：48分の動画Upload/秒
  - WEBページの総数：1,000,000,000,000 (2008年)
- データの正しい理解
  - 商学を知らずにマーケティングデータは使えない
  - 個人情報の保護
- データからの真実の抽出=データ分析
  - 大量データを扱う技術
  - データ分析の技術

データサイエンスの分類

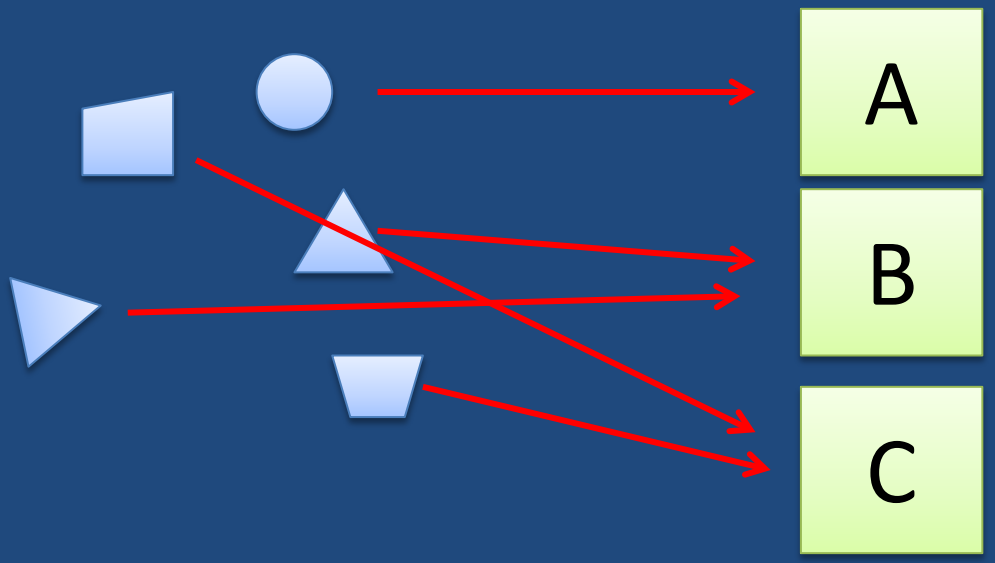
- データの収集
  - 心理学，実験計画法，感性工学，社会学
- データの保管
  - データベース工学，情報セキュリティ
- データの分析
  - 統計学，データマイニング，パターン認識，機械学習
- データの予測
  - シミュレーション，数理計画法

データ分析

- 大量データからの知識抽出
  - 知られていなかった真実の発見
  - 見えなかった情報の可視化
- 主な分析手法
  - クラス分類
  - 予測
  - パターン抽出
  - クラスタリング

クラス分類

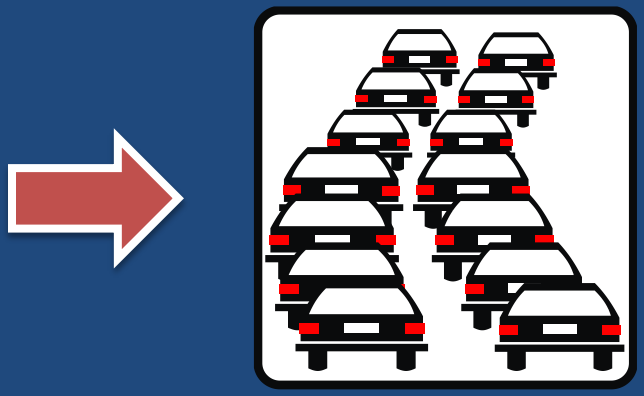
- 与えられたデータをカテゴリに分類
  - 例：迷惑メールフォルダ
    - 「迷惑メールらしさ」を学習して迷惑メールを判定
    - 判定されたメールは迷惑メールカテゴリに分類



予測

- 与えられたデータから予測する
  - 例：観光客数の予測
    - 昨年度の来客数，曜日，天気から予測
  - 例：SNSへの書き込みから株価を予測
    - 景気が良いと皆の書き込みも明るくなる！？

状態	値
天気	晴れ
去年の来客数	1000
曜日	日曜日



パターン抽出

- データ内に存在する特徴的なパターンの発見
  - 例：野菜と野菜ジュースを一緒に買う人が多い
  - 例：水は決まったものが買われる可能性が高い
- パターン抽出の利用例
  - 商品の推薦
    - 「この本を買った人はこんな本も買っています」

クラスタリング

- 類似したデータ同士をまとめる
  - 例：写真から同一人物の顔を抽出
  - 例：観光客の滞在場所
    - 観光客が興味を持つ場所を特定

目次

- なぜデータサイエンスなのか
- データ分析の例
  - マーケティング
  - 観光業界
  - スポーツ科学
- データサイエンティストを目指して

まとめ

- なぜデータサイエンスなのか
  - ビッグデータの時代
  - Data Science Everywhere
- データサイエンスの主な役割
  - データの収集，保管，分析，予測
  - 医学，工学，農学・・・あらゆる分野への応用
- 求められる技術
  - 統計，コンピュータサイエンス，データ工学・・・
  - 利用分野の知識

Wanna be sexy?
Be a Data Scientist!

DataSicence is everywhere!