データサイエンスの展望

滋賀大学データサイエンス教育研究センター
竹村彰通

２０１６年７月２２日
エジソンの会

項目

1. 自己紹介
2. 滋賀大学データサイエンス学部新設申請の話
3. 統計学の流れとデータサイエンスのこれから

今回のスライドには、一部以下の
内容が含まれています

1. 滋賀大学データサイエンス学部に関する説明スライド
2. 東京大学学術俯瞰講義 「数学―革新の歴史と伝統の力」(坪井俊　他)　2014年夏学期
   第7,8回 「統計学の過去と未来」 (竹村)
3. 司法研修所での講演「統計・データ分析のこれから」 2016年7月7日

1. 自己紹介

- 研究は「数理統計学」
- 統計学にかかわる数学的な理論を研究している
- 最近では「数学」が「計算機」におきかわっている部分も多い
- 計算機を前提とした新しいタイプの数学にも興味がある
- 最近の研究のキーワード:「計算代数統計」
- 研究以上に、学会活動、統計教育などに重点
- 現在の仕事は「データサイエンス学部」の立ち上げ

学会活動等

- 2004年9月～ 2006年9月
  日本統計学会理事長
- 2011年4月～ 2013年6月
  日本統計学会会長
- 一昨年まで10年ほど
  日本学術会議数理科学委員会数理統計学分科会長
- 2010年より「統計検定」の立ち上げ作業

「統計教育」に関する執筆・編集

- 数学セミナー2013年5月号

  「統計計算から統計的思考へ／数学教育の中での統計教育の新たな方向」

- 数学セミナー特集 2013年10月号「時代が求める統計学」

  「統計的な考え方と結果の見方」

- 青戸社『現代思想』特集2014年6月号「ポスト・ビッグデータと統計学の時代」

  「統計学にとって情報とは何か」

「統計教育」に関する執筆・編集

- 日本学術会議数理科学委員会数理統計学分科会の提言(2014年8月、委員長)

  「ビッグデータ時代における統計科学教育・研究の推進について」

- 日本学術会議数理科学委員会統計学分野の参照基準検討分科会委員長(2015年12月、委員長)

  「大学教育の分野別質保証のための教育課程編成上の参照基準 ― 統計学分野」

- 統計検定の問題策定全般
- mooc 講座(インターネット無料講座)の作成

gacco.org/courses/course-v1:gacco+ga047+2016_10/about

検索

gacco The Japan mooc Certified by JMOOC

ログイン　新規会員登録

ga047: 統計学Ⅱ:推測統計の方法

ga047を受講登録

講座概要

この講座には「通常コース（無料）」「対面学習コース（有料）」の2種類のコースがあります。＊対面学習コース（有料）では、対面授業を11/12（土） 東京・京都で開催します。
対面学習とは http://gacco.org/face-to-face.html

通常コース(無料)	対面学習コース(有料)	
お申込み方法 講座詳細ページ（本ページよりすぐにお申込みいただけます）	**お申込み方法** 外部サイト「Peatix」よりお申込みいただけます。	**お申込み方法** 外部サイト「Peatix」よりお申込みいただけます。
	コース内容	コース内容

※受講登録するとお客様の利用者情報

2. 滋賀大学データサイエンス学部新設申請の話

- 滋賀大学では日本初の「データサイエンス学部」設立準備中(来年4月開始)
- この分野でトップブランドを目指す

データサイエンスなら滋賀大に行こう！

- 社会人や企業からも強い再教育の需要

⇒ 大学院設置も早期に目指す

注: 以下は文部科学省申請中の内容です

- 募集定員は1学年100名
- 逆T型(逆Π型)人材の育成
- データサイエンスの3要素
  - データエンジニアリング　(情報学)
  - データアナリシス　(統計学)
  - 価値創造 (ビジネス分野でのデータ分析を経験)
- すでに入学意向、採用意向とも、定員を大きく超える反応あり (昨年9月のアンケート調査)
- 多くの企業と教育・研究に関して連携交渉中

文部科学省平成27年度
「国立大学改革強化推進補助金」選定事業

新学部設立より早く、改革補助金に採択

滋賀大学

（連携大学：滋賀医科大学　京都大学　大阪大学）

人文社会系大学から文理融合への転換
－データサイエンス教育研究拠点形成のための大学間連携の推進－

大胆なガバナンス改革と学内資源の再分配等による日本初の「データサイエンス学部（仮称）」を設置。データサイエンスを含む自然科学分野の多様な領域の英知を大学間連携により結集し、先行事例のない最先端の教育プログラム・教材・教授法の開発や教育の質保証システムを確立。人文社会系大学から文理融合型大学への転換に向けた先行モデルを提起。

滋賀大学
SHIGA UNIVERSITY

項目

- データサイエンスとは？
- 海外の大学の動向
- 滋賀大の改革構想
- 育成する人材像
- カリキュラムの考え方
- 高等学校からの期待 (アンケート調査の報告)
- 産業界からの期待(〃)

以下、文字数の多いスライドが続きます

データサイエンス ― 価値創造のための新たな科学

客観的な存在としてのビッグデータを対象として、そこから新たな知見を引き出し、価値を創造するための科学

様々な領域で集積されるビッグデータ

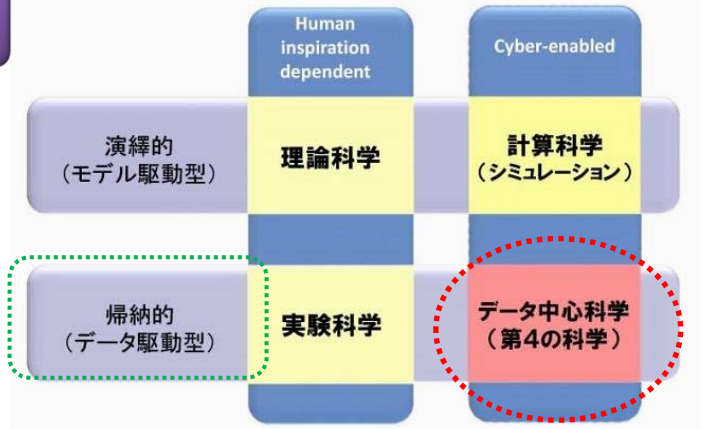
データアナリシス
大規模データを分析・解析するための専門知識とスキル
（統計学）

データエンジニアリング
大規模データを加工・研磨・処理するための専門知識とスキル
（情報学・コンピュータ科学）

新たな知見

価値創造
ビジネスや政策などの領域で課題を読み取り
データエンジニアリングとデータアナリシスにより得られた知見を
現場の意思決定に生かして価値を創造する
（領域分野での成功体験＝PBLの繰り返し）

データサイエンス



（左図の出典） 北川源四郎氏（情報・システム研究機構長）
「ビッグデータ時代に対応する人材の育成」
2014年11月21日、滋賀大学外部有識者会議、報告資料

データサイエンス分野の重要性

- 「ビッグデータ時代を迎え、データの利活用により付加価値を生み出す新事業・新サービスの創出が重要」（骨太の方針2015、日本再興戦略改訂2015）
- 「欧米等と比較し、データ分析のスキルを有する人材や統計科学を専攻する人材が極めて少ないという危機的状況」（「科学技術基本計画」（第５期））
- 現実にはこの分野で日本は著しく立ち遅れ

  ←一つの要因として、統計学部がなかった

アメリカの学部・大学院教育の動向

- 海外の大学には、もともと**統計学の独立した学部・学科**が存在する（アメリカ、イギリス、欧州はもちろん、中国、韓国などのアジア諸国も同様）
- 独立した統計学のセクションがあると、情報工学、コンピュータ科学、ときには数学も加えて、**データサイエンス・プログラムが柔軟に生成されやすい**
- アメリカの学部教育では、**統計学専攻のプログラムがデータサイエンスを意識した内容**に変更されている
- 統計学専攻は近年の理系学位（STEM）の**一番人気で、卒業後の給与も一番高い**。

アメリカの学部・大学院教育の動向

アメリカ統計学会ニュースレターから（amstat news, February & April, 2015）

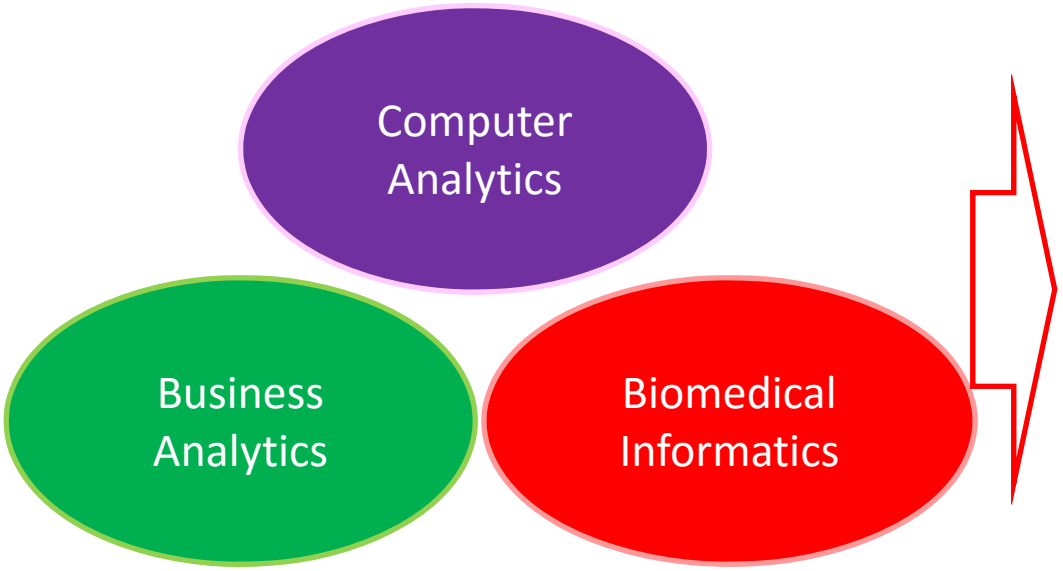
統計学・学士号	取得者比率 2011-13／2003-05	データサイエンスを意識したカリキュラム
Purdue University	875%	Big Data course programming language (C, Python, Java, etc.) プログラミング言語
University of California, Berkeley	224%	the upper division electives are almost all centered on data analysis (statistical learning theory)　上級選択科目としての統計的学習理論
University of Illinois, Urbana-Champaign	452%	new analytics courses emphasizing data management and statistical analysis of databases, Big Data methods　ビッグデータ手法に重点化したコース a new statistical programming course　統計的プログラミング
Carnegie Mellon University	191%	experiential learning through the use of real data sets　実際のデータ利用 programming and software engineering in R, and databases and data management visualization, data mining　プログラミング、 R、可視化、マイニング

オハイオ州立大学の例 (学士)

大学独自で150億円投資。教員50名採用

2013年からData Analytics majorを開始

3つの専門分野



応用分野

遺伝子、医療、健康、製薬、
セキュリティ、マーケット、
ファイナンス、機械学習、
検索技術、スポーツ、航空、
保険分野等

実データを使った解析、各種企業（Nation wide、JPMorgan Chase、Microsoft、IBM、Google、TATA Group、Infosys等）との共同研究、様々なプロジェクト研究が行われている。

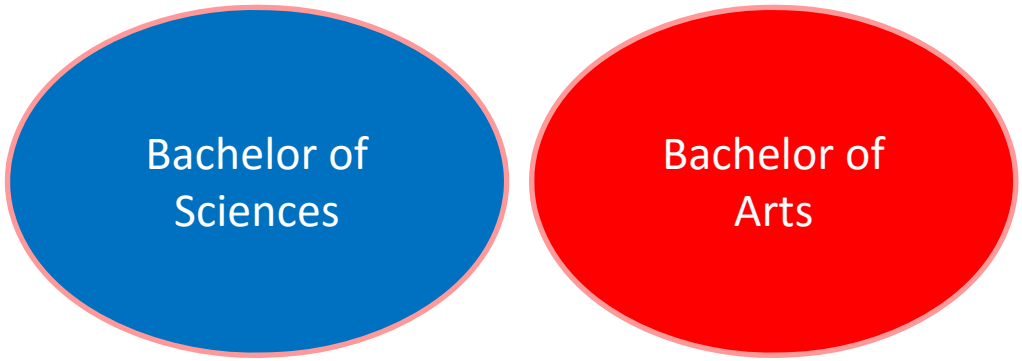
DSのための建物改修(40億円)

内装（コンセプト)

ロチェスター大学の例 (学士)
大学独自で100億円投資

ロチェスター大学では、データサイエンスプログラムで、以下の学位を取得可能

学位（undergraduate）



これらの学位を取得するには
・mathematics
・computer science
・statistics
の3つのコア科目を取得する必要がある。

生命情報学、医用生体工学、脳科学、認知科学、ビジネス、経済学、コンピューター科学、物理学、数学、統計学、政治学等の多分野の研究者を採用し、様々な応用研究を可能とする。

建設中の Wegmans Hall (コンセプト)

コンファレンスルール　(コンセプト)

建設中の現場（2017冬完成予定）

滋賀大学の改革構想 ― 文理融合大学への変革

改革方針

人文社会系大学から文理融合型大学への大転換

- 教育・経済2学部＆分離キャンパスの大学からの脱却
- 社会的要請の強い新学部創設とそれによる教育組織の改組
- 専門職業人育成機能の強化

本学の強み・特色

ビジネススクール型教育での情報管理学科と統計分析の強み
教育学部における情報教育課程
経済・教育における伸びしろ豊かな専門職業人の養成

日本初のデータサイエンス学部創設
データ駆動型価値創造人材の育成
先行事例のない教育プログラムの確立

新学部設置を契機に

経済学部のビジネススクール型教育の強化　←DS副専攻
データ分析に明るい教員養成　←全学教養としてのDS導入

データサイエンス教育拠点形成

必要性･緊急性

欧米等と比較し、データ分析のスキルを有する人材や統計科学を専攻する人材が極めて少ないという危機的状況
・日本学術会議『ビッグデータ時代に対応する人材の育成』
・第5期「科学技術基本計画」

ガバナンス改革

- 教育研究組織と教員組織の分離
  教員人事の全学管理、全学的な視点から人的資源の再配分
- 人事・給与システム改革（年俸制の導入等）
  年俸制やクロスアポイントメント制

データサイエンス教育拠点

データサイエンス学系（H28）
データサイエンス教育研究センター（H28）
データサイエンス学部（H29）
データサイエン研究科(H33)

世界的に先進的で質の高いDS教育

カリキュラムの標準化
DS基礎教育
多様な領域でのデータ駆動型PBL演習

現実の成功体験から理論学習を誘う反転カリキュラム
反転教育とアクティブラーニングによる基礎教育
1年次から4年次までのデータ駆動型PBL演習
現実の問題解決に挑む4年次卒業研究

人材育成像の継続的検証
教育内容の標準化・質保証のための
PDCAサイクル
↓
外部連携による組織体制

DS基盤研究
機械学習、最適化等の研究
国際シンポジウム、教材開発

DS価値創造プロジェクト研究
各領域でのデータ利活用

DS教育開発運営
・価値創造プロジェクトによるデータ駆動型PBL演習教材の開発
・MOOC教材義

DS調査・広報
海外動向視察・調査
育成人材像調査

海外の連携大学

総務省
統計研修所

滋賀医科大学
滋賀県立大学

長浜バイオ大学

企業や自治体等

統計教育大学間連携ネットワーク
JINSE
大阪大学、同志社大学など

環びわ湖地域コンソーシアム

データサイエンティスト協会
経団連、関経連など

教育研究アドバイザリーボード

後発大学への教育ノウハウの提供

データサイエンス学部における育成人材像

データ駆動型価値創造　逆T（またはΠ）型人材

《多様な価値創造のフィールド》
- マーケティング
- ファイナンス、保険
- 企業会計
- ビジネスエコノミクス
- 医療・健康・福祉
- バイオ、製薬
- 環境、防災、気象
- 教育
- 公的統計
- 社会心理
- 地域文化情報

逆T型

文系的

価値創造の経験とノウハウ & 領域における専門知識

理系的

データサイエンスの専門知識とスキル

データアナリシス	データエンジニアリング
大規模データの分析・解析 専門知識とスキル（統計学）	大規模データを加工・研磨・処理 専門知識とスキル（情報学・コンピュータ科学）

逆Π型

領域が複数あれば

領域①価値創造の経験とノウハウ

領域②価値創造の経験とノウハウ

データサイエンスの専門知識とスキル

データサイエンス学部のディプロマポリシー

1. データエンジニアリングとデータアナリシスの専門知識とスキルを修得し、データサイエンスの基礎的力量を備えている。
2. データサイエンスの基礎を応用して、多様な領域でのデータ駆動型価値創造を導くための実装力を備えている。
3. 多様なコミュニケーションの力量を備え、データ利活用の現場で相互補完的な専門性を有する仲間と協力して、組織目標を追求できる。
4. データ駆動型価値創造社会の哲学・倫理・政治等について、バランスのとれた見識を有している。
5. 上記のようなデータサイエンティストの専門的力量とイノベーティブな心の習慣を背景に、卒業後の現場での課題に対応して、自律的な学習を進めることができ、多様な領域における価値創造のための創造的イノベーションにも貢献できる。

履修モデル、想定される就職先、キャリアプラン

Bachelor of Data Science

3つの履修モデル＆人材像	想定される就職先
データエンジニア	• IT系企業、シンクタンク、製薬企業や医療機関等 • 企業のデータ分析部門 • 大学院
データアナリスト	• 企業のデータ分析部門 • 経営コンサルタント、シンクタンク、製薬企業や医療機関等 • 企業の財務経理、経営企画、マーケティング、生産管理等の諸部門 • 大学院
データコンサルタント	• 企業の財務経理、経営企画、マーケティング、生産管理等の諸部門 • 国または地方自治体 • 企業のデータ分析部門

就職後の発展・成長キャリアプラン

独り立ち

Associate Data Scientist

企業や官公庁の現場の先輩たちの助力を得ながら、データサイエンティストとしての基本的な仕事を、概ね一人で遂行できる

プロフェッショナル

Data Scientist
(Master of Science in Data Science)

職場での経験や教育訓練を経て、専門的知識やスキル、価値創造の現場力に磨きがかかり、データサイエンティストとして、かなり高度な業務を遂行できる

統括レベル

Senior Data Scientist

データサイエンスの特定の専門的知識とスキルについて顕著なレベルにあり、他の専門家と協力・協働して、高度なプロジェクトを遂行できるし、全体を統括することができる

日本の雇用慣行とデータサイエンティスト

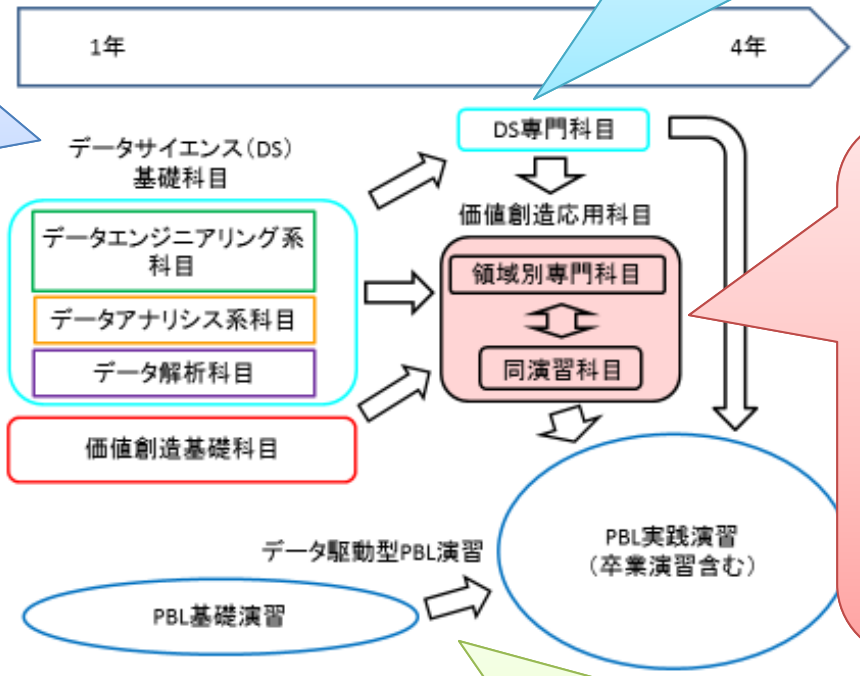
- 企業は実際に「データサイエンティスト」をやとってくれるか
- 学生が自信を持って就職活動に臨めるか？
- 専門性を生かしたキャリアアップが図れるか？
- 企業が大学に求めるものは何か
  ー 「基礎力」「適応力」「専門性」？
- 日本の大企業と、web 系のスタートアップではかなり異なる
- 企業は本当にデータを生かそうとしているのか?

カリキュラムマップ

欧米の大学の場合と基本構造は同じだが、
データサイエンスの専門知識とスキルだけでなく、
それを活用した価値創造の実践を特に重視！

ビッグデータ解析のための発展的科目
データマイニング、テキストマイニング、
機械学習、ベイズ理論、最適化など

- 統計学、コンピュータ科学・情報工学から構成
- 相対的に統計分析によるアナリシスを重視
- 汎用解析ソフトの利用・訓練



- 文理融合の実践を重視
- データサイエンス基礎を応用する多様な領域を準備
  - マーケティング
  - ファイナンス
  - 会計
  - 医療・健康・福祉
  - ビジネスエコノミクス
  - 環境
  - 教育
  - 保険・リスク
  - 公的統計
  - 心理
  - 地域文化情報

本学DSプログラムの真骨頂―現場のデータを利用した価値創造PBL演習での成功体験

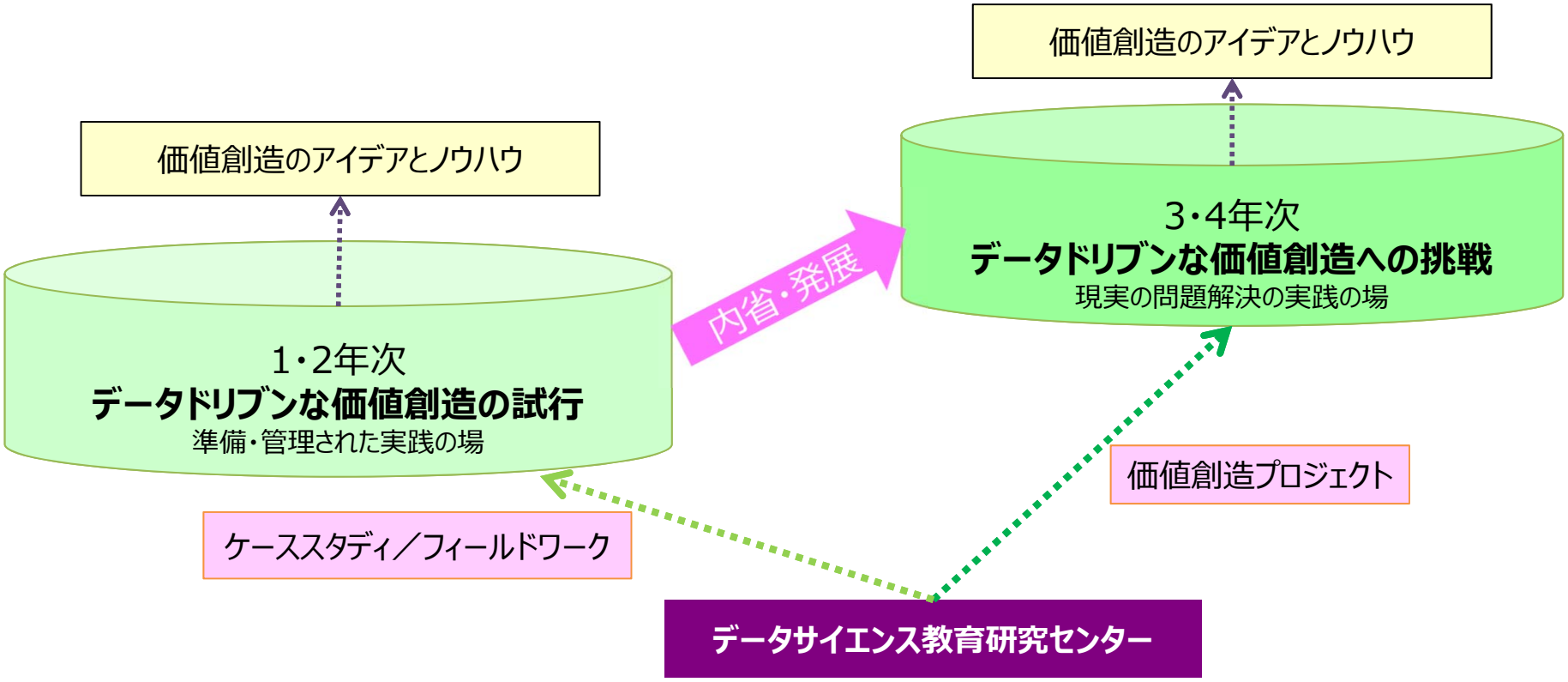
- ナレッジマネジメント理論に基づくPBL演習の設計
- DS教育研究センターにおける価値創造プロジェクトが企業、自治体、非営利団体等の現場とデータを提供
- 外部に開かれた実践の場でのコミュニケーション力やチームワーク形成力の鍛錬

データ駆動型価値創造PBL演習

データエンジニアリング及びデータアナリシスの専門知識とスキルを活用した**データドリブンな価値創造の実践**
成功体験から価値創造のアイデアとノウハウを身に付ける

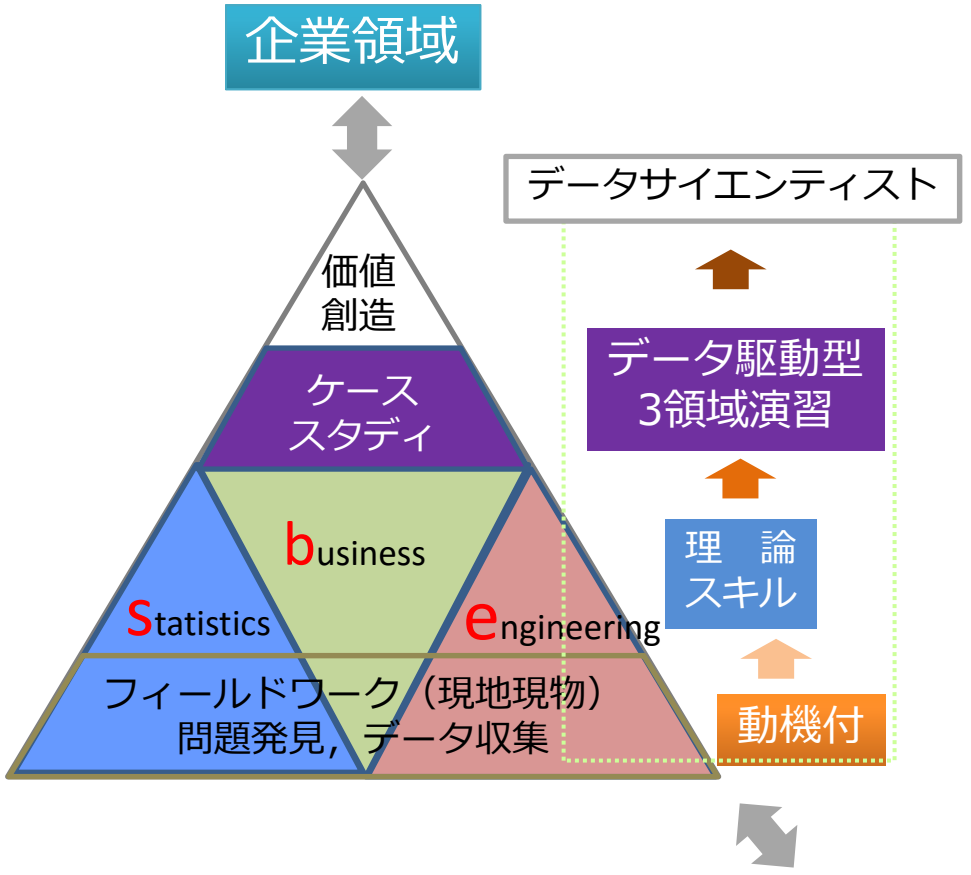
PPDACサイクルの繰り返し

Problem	…	現場の体感と問題意識による分析課題の設定
Plan	…	必要なデータ、データの処理方法、統計分析手法等の選択
Data	…	実際のデータの収集、データの整理・加工・抽出
Analysis	…	分析ソフト等を利用した分析結果の導出
Conclusion	…	分析結果を活用した価値創造の提案



DS PBL演習：フィールドワーク&ケーススタディ

課題：データ解析力，KKD，グローバル，etc.
分析：ビッグ&スモールデータ，Web，センシング，etc.
成果：企業の最適化，戦略と価値創造，etc.



課題：福祉，健康・教育，地域志向人材， etc.
分析：センシング，医療，Web・SNS，環境，etc.
成果：教育科学，社会福祉，地方創生，etc.

課題：行財政，少子高齢化，子育て支援，地方衰退，農業，etc.
分析：オープンデータ，センシング(IoT)，Web・SNS，etc.
成果：行政・社会の最適化，事業と雇用創出，共同体，防災，観光，etc.

データ駆動型演習における企業との連携

- 演習では、実際のビジネスの現場で得られるデータを扱うことにより、実際のデータから価値創造を経験することが必要
- 地元の企業や地方公共団体との連携を依頼中
- 地域のデータを扱うことにより、地域に対する理解も深まる
- 地元で新たに起業する学生が現れることを期待

連携検討先

滋賀大学

・滋賀県（総合政策部統計課など）

- 県民の統計リテラシーの向上
- 地元企業等の統計分析、統計相談
- 滋賀県のデータの活用

・滋賀銀行、京都銀行

- 企業紹介、人材活用、学生支援

・大学生協（滋賀大学、京都大学）

- 売り上げデータの分析

今後も様々な協力・
連携の強化を目指す

・データサイエンティスト協会、日本IBM、野村総合研究所:

- 実務家による講義の提供
- 共同研究

・個別企業との共同研究・連携（一部）

- 日本ソフト開発、平和堂
- 村田製作所、玉田工業
- アイディーズ、オプトホールディング

卒業までに取得可能な資格・検定

- 社会調査士
- 統計検定2級、準1級
- 基本情報処理技術者試験（レベル1、2）
- 品質管理検定2級

アドミッションポリシー、入学者選抜方法

アドミッションポリシー

◆高校の様々な教科・科目の学習を通して、バランスよく、文・理の基礎的知識を身につけてきた、潜在性ゆたかな人

◆コミュニケーション力を有し、多様な人々と協働して、理想の未来に向けた価値創造に貢献したい人

◆ものごとを筋道立てて考えることができ、人間社会や自然の現象を数理的に分析することに関心のある人

◆情報ネットワーク、プログラミング、コンピュータグラフィックス（視覚化）などに関心がある人

学習内容は理系的ではあるが、文系タイプの学生でも十分受験可能

入学者選抜方法

高大連携・接続の工夫・改善を考慮

（1）アドミッション・ポリシーの観点からみた学外コンペや資格試験の利用

（2）大学における講義・プロジェクト演習への参加＋レポート作成

（3）新テストにおける「合教科・科目型」「総合型」に対応して、統計と情報を組合せた出題を検討

アンケート調査結果の概要 (高等学校調査)

滋賀大学

1. **調査対象**　滋賀県，岐阜県，京都府，愛知県の高校22校
2. **調査方法**　高校留置き調査
3. **依頼数**　5,290人　　**回収数** 4,328人
4. **調査時期** 8/24～10/2

- データサイエンス学部を受験したいと思う: 13%(562人)
- 受験意向のうち入学したい思う: 91%(512人)
- 文系･理系別入学意向
  理系クラス: 321人　文系クラス: 178人

100名の定員をすでに大幅に超過

アンケート調査結果の概要 (企業対象調査)

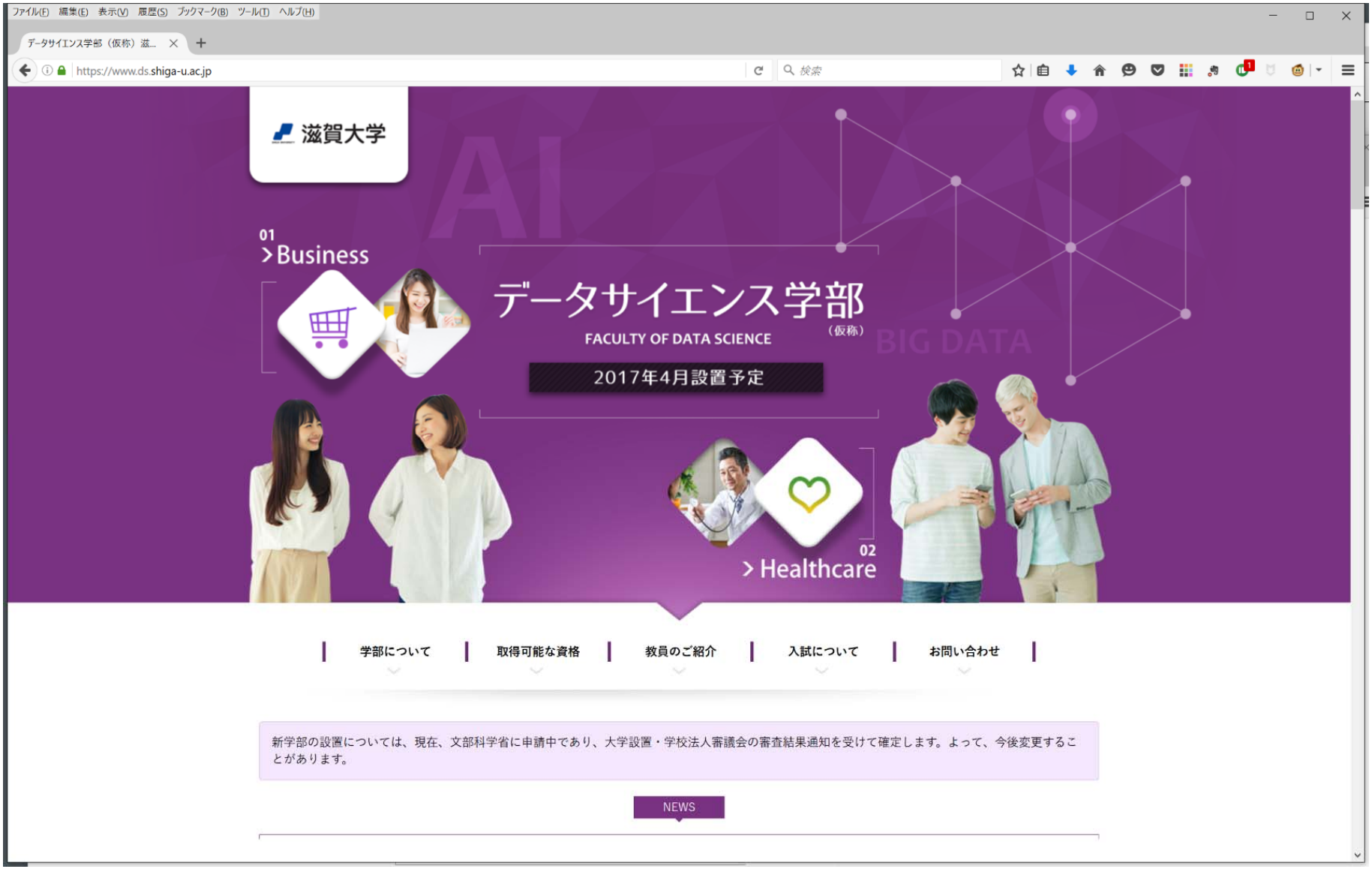
滋賀大学

1. **調査対象**　滋賀大学の卒業生を受け入れた実績のある企業の採用担当者
2. **調査方法**　郵送調査
3. **依頼数**　1132社　　**回収数** 320社
4. **調査時期** 8/31～9/18

- データサイエンス学部の社会的必要性: 97%の企業が評価
- データサイエンス学部の卒業生を採用したい: 90%の企業
- アンケートに回答した企業のみで300名以上の採用意向
  ← 100名の定員をすでに大幅に超過
- 特に規模の大きい企業の採用意向が強い

滋賀大学データサイエンス学部に関する現時点のまとめ

- 滋賀大学のデータサイエンス学部構想は先進的な大学改革として評価されている
- 高校生や企業からの期待も高い
- 多くの企業・団体と連携検討中
  - 実務家による講義
  - 学生のインターンシップ、実習
  - 共同研究


滋賀大学
01
>Business
データサイエンス学部
（仮称）
FACULTY OF DATA SCIENCE
2017年4月設置予定
02
>Healthcare
学部について
取得可能な資格
教員のご紹介
入試について
お問い合わせ
新学部の設置については、現在、文部科学省に申請中であり、大学設置・学校法人審議会の審査結果通知を受けて確定します。よって、今後変更することがあります。
NEWS

３．統計学の流れとデータサイエンスのこれから

最近の私自身の問題意識

1. 不確実性の扱い
2. 取りやすいデータと取りにくいデータ
3. 個人情報保護と社会的に有用なデータ
4. 機械と人間の関係
   人間が理解できない場合の扱い
   人間の動機づけ

統計学の流れ (数年前まで)

戦後の統計学

- アメリカ中心
- 「数理統計学」の成立
  （Wald の統計的決定理論等）
- 数学の形式化の流れ
- 独自な理論体系と内的発展の重視

応用分野の広がり

- 計量経済学 (大規模計量モデルの功罪)
- 心理学，社会学 (アメリカ的な数量的な研究の重視)
  [我が国で言う「文系」とはやや異なる]
- 医学統計
  大規模臨床試験、新薬の認可、副作用、Evidence Based Medicine
- 実験計画法の工業への応用
- 統計的品質管理
  - 日本からの貢献も大きい (因子の性質の区別等)
  - 「田口メソッド」(田口玄一)
  - 品質のばらつきに影響を与える要因の分析の重視
- 製造業・大量生産に非常に有効だった

応用分野の広がり

- 金融工学，数理ファイナンス
  (1980年代より，主に確率論)
- ブラック･ショールズ式などが注目を集めた
- 株価変動などの金融データを確率過程としてモデル化
- 現実とモデルには乖離が見られた
  ← リーマンショックのような現象を想定できなかった

確率モデルと統計手法の変化

- 背景： 計算機の発達により、実際に複雑な計算ができるようになった
- 統計モデルの大規模化：

  データ自身が大量に得られるため、統計モデル(データを説明する確率分布)も大規模化してきた
- 大規模モデルの計算にともなう面が手法的には重要となってきた

モデル選択の方法　(最尤法ベース)

- モデルの説明力(データへのフィット)とモデルの簡潔さのバランス
- 複雑なモデルは手もとのデータへのフィットは高いが、新たなデータの 予測が不安定となる場合がある (overfitting,「過学習」)
- (赤池情報量規準) などの手法によるモデル選択. 我が国の貢献が 大きい

ベイズ法

- システムのパラメータに関する不確実性や、分析者の主観的な 情報などをすべて確率変数としてモデル化する(事前分布)
- システムの未知変数は観測されていない確率変数であると考える
- 統計的推論は、データに基づいて事前分布を事後分布でおきかえる機械的な操作となる
- 機械的な操作で統計的推論をおこなえるという点が、計算機との相性がよい
- 私自身は「確率計算万能主義」とよぶ

(伝統的な)統計的方法の性格

均一性を持つ母集団の想定し、母集団におけるバラツキの存在を前提

- 生命保険の例
  - 性別、年齢等で分ければ均一の集団と考えられ、そこでリスクをプールする
  - 「均一性」を信じるには情報が少ないほうがいいこともある
  - 詳しい遺伝情報が得られた時に、生命保険の性格が変わるのか?
  - アンジェリーナ・ジョリーの手術の例

統計的方法の性格: 薬の例

平均的に効く薬がいいのか、「特効薬」がいいのか

- とりあえずの答: (劇的な効果はなくても) どのような条件の人にも平均的に効く薬がいい
- 「交互作用」のないものがよい
- 「ロバスト性」、「副作用と特効薬」
- 「特効薬」といっても、一定の大きさの母集団は仮定されている
- ただし、現在の流れとしては “personalized medicine” に進んでいる

統計的方法の性格

- 統計モデル = 固定効果 + 誤差
- 誤差はブラックボックス
- 固定効果は未知母数を含む
- 誤差の分布について平均的な性能で評価する

- 数年前から「ビッグデータ」の時代となり、様相が変化してきた
- ビッグデータ時代で、統計学の性格はどのように変わっていくのか

ビッグデータの時代

データ量　(wikipedia等から)

- メガバイト (10^6): 3.5 インチフロッピー
- ギガバイト (10^9): cd-rom, dvd
- テラバイト (10^{12}): 現在の普通のハードディスク
- ペタバイト (10^{15}): アメリカにある図書館全体の情報
- Exabyte (10^{18}): 人類がこれまで発した言葉の全体
- Zettabyte (10^{21}): 2012年中のインターネットトラフィック

- ただし量だけではない
- 扱うデータの種類が変わってきている (テキスト、画像、音声等)
- 特に動画のデータのサイズが大きい

ビッグデータ分析のニーズ
(客観的な存在としてのビッグデータ)

ビッグデータをなぜ分析するのか？

- ビジネス: 利益に通じるなら
- 製造業: 品質の向上や新しい製品開発につながるなら
- 科学: 新しい知見が得られるなら

ビッグデータからの価値創造

誰が分析するのか？

- 「データサイエンティスト」の世界的不足
- コンピュータとアルゴリズムがある程度自動的に分析「機械学習」

→ 新たな分析手法の開発

→ でも解釈するのは人間

結果の解釈、信頼性の評価の問題

- 「データマイニング」の成功例

相関と因果

- 相関と因果は異なる
- ビールの売り上げとおむつの売り上げに相関があるからと言って、ビールのそばにおむつをおけばより多く売れるとは限らない
- 朝ごはんを食べる児童が成績が良いからと言って、児童に朝ごはんを食べさせれば成績が上がるとは限らない
- 大量に観察しても、観察データだけからでは相関と因果を簡単には区別できない（かなり気をつけて数理モデリングする必要がある）
- 予測のモデリング ⇔　構造を記述するモデリング
- 実験計画法

データと情報

- データとしては大量でも，情報はほとんどないものもある．
- 例: サイコロを投げ続けて出た目を記録して行く: 1,5,2,4,3,…
- ハードディスク一杯になるまで記録したら「ビッグデータ」
- このデータは有用？
  - 常識的な答: 単にデタラメを記録しただけだから無意味．
  - 「サイコロを投げて出た目でディスクをうめただけ」と言えば済む．情報とは言えない．
  - 「データの生成過程」の記述のほうが重要．

    → 統計学の考え方
  - 1がたまたま27回続けて出たら珍しいので記録の価値があるか？

データと情報

- デタラメなデータほど情報量が多いとする考え方もある

  → コルモゴロフ複雑度

- でたらめなデータ (1,5,2,4,3,…) はそのまま記録するよりない．簡単に記述できない
- パスワードはでたらめでないと困る(12345678とかは駄目)
- たんさくのパスワードは覚えられないので，皆困っている
- 1,1,1,1,…. とハードディスクを1で埋め尽くすのはもちろん無意味
- どちらにせよ、多量のデータ自体が情報を含むわけではない
- でも何が有用な情報かはわからない

監視カメラや衛星が記録する世界

- 監視カメラが人々の出入りを記録している
- 人々の移動はランダム
- ランダムな現象をそのまま記録しても意味があるのか
- 通常の監視カメラは，２週間程度で画像を消去
- しかし犯罪がおきるととたんに有用となる
- 画像解析技術が進歩して，個人が自動的に識別されるようになったとして、プライバシーの問題はどうなるか

監視カメラや衛星が記録する世界

- 天気予報も解像度があがり，局所的な予報が可能になってきた.
- 短期的な予報の精度もあがってきた
- それでも，いくらデータあっても，今年の夏が暑いかどうかは予測できない　← 研究者によって見解の相違がある
- データそのものというより、数理モデルが重要

SNSのデータから何がわかるのか

- ブログや twitter のデータはほとんどがノイズ
  「おなかがすいた」「今日の講義つまらない」等
- 発言している人々の属性が把握できない.
- 発言数の分布の裾が重い.
- 動向をとらえることはできるが，確定的なことを言うのは難しい
- 差分には情報があるが，水準には意味がない

- 「母集団」が設定できない.
- 統計的手法の根拠: 母集団，あるいはより仮想的な統計モデルの安定性

統計モデル＝固定効果＋確率的な誤差

取りやすいデータと取りにくいデータ

- 取りやすいデータ
  - 電子的なデバイスにより得られるデータ
    スマホ、ポイントカード、交通系データ、ネットでのオンライン取り引
    き、　遺伝子解析機器
- 取りにくいデータ
  - 測定が困難なデータ
    福島第一原子力発電所の原子炉の中の状況
    地下10kmより深い地球の様子
  - ネットを使わない人の考え
  - 記録されなかった行動（店で選ばなかった商品）

- 取りにくくても社会的に重要なデータについては「統計調査」が必要

個人情報の扱い

- 最近では、商品購入の際にIDつきポイントカードを用いることが多い
  ← ポイント還元があるため
- ポイントカード作成の時に、年齢、性別などの背景情報もとられている
- IDつきポイントカードのデータは、売り上げのみのPOSデータよりデータの「価値」が高い
- SUICAなどの交通系カードも同様
- データを囲いこんだ企業が強くなる
- 社会的な観点からも有用なデータであり、一定の公共性がある(電話会社の通話データの裁判での利用の例)

ビッグデータ時代における
個人の情報と集団の情報

- ビッグデータ時代は「個人情報」の時代
- ネット経由で買い物をすると履歴が残り、また他のユーザを参考にして商品が推薦される
- 利便性のために，消費者が個人情報を提供している
- 一方で，国のおこなう統計調査への非協力が増えており、統計調査の環境が悪化している.
- 国勢調査などの国の基本的な調査で調べる情報は、利便性と引き換えに得られる情報ではない。
- 統計: 集団に関する情報
  - 個人の情報 ↔ 小集団の情報 ↔ 集団の情報

最近の人工知能ブームについて

- 多量のデータに対して、多層ニューラルネットワークモデルなどの複雑な数理モデルを適用(フィット)した成功事例が報告して、大きな注目をあびている。
- 画像認識やパターン認識での成功例
- モデルのフィットがある程度自動化されているところも魅力的
- 人間の判断を「教師」として、「ほとんどの場合に」人間と「同様」の判断が高速にできればよい
- 人間自身の判断のための前処理として有用

最近の人工知能ブームについて

- 問題点: モデルが複雑すぎて、なぜ機械がそのような答えを出しているかが、人間にはわからない
- 不確実性の観点からは、答えがどのくらいの確信に基づいているのかがわからない
- 「なぜそのような答になるのか」「どのくらい確かな答なのか」を同時に示すことが必要
- 不確実性のモデリングとしての統計学の手法を、多層ニューラルネットワークモデルなどと融合していく必要がある

最近の人工知能ブームについて

- 人々の行動に影響をあたえるには、「理由」が理解できることが重要
- 理由が正当だったとしても、必ずしも人は動かない
- そもそも同じ情報を得ても、人々の判断は異なる
- 人工知能の判断が誤っていて事故につながった場合、法的な責任は誰にあるのか?

最近の人工知能ブームについて

- 日本の競争力の観点からは、アルゴリズム開発等では全く勝負にならない現状がある
- 日本もあきらめてはいけないが、やはり相対的に強みのある部分を発展させて行くのがよい

  → 物作りやきめ細かいサービス

- IoTは重要
- 品質管理分野でも新たな流れが必要

データサイエンティストの育成と統計リテラシーの向上

- 人工知能もビッグデータ解析に基づいている
- 人工知能の出す答についても、人間が判断して利用
- データ解析に専門性を有する人材は重要

  データサイエンティスト
- 数学もアルゴリズムも統計もビジネスもできる人材
- 「サイエンス」には違和感あるが、他にいいことばがない
- 専門性を有する人材のみならず、一般的な統計的なリテラシーの向上も必要

まとめ

- 滋賀大学データサイエンス学部について紹介した
- データサイエンスの今後について論点をあげた
  - 統計学の流れと考え方
  - データサイエンス：客観的な存在としてのビッグデータの解析
  - 最近の人工知能ブームについて

ご清聴ありがとうございました