オリエンテーション(前座)
データサイエンスにおける産学連携の重要性とシーズ

情報・システム研究機構　統計数理研究所
椿　広計

2021/11/16

情報・システム研究機構産学連携セミナー

オリエンテーションの内容

- オリエンテーション前半
  - 産官学一体となった人財育成の必要性
    - *統計エキスパート育成*
- *オリエンテーション後半*
  - *産官一体となった技術開発研究構築の必要性*
    - 吉田亮ものづくりデータ科学センター長講演のオリエンテーション
      - 調査統計科学⇔ビッグデータサイエンスではなくて
  - ***実験統計科学⇔創られたデータによるデータサイエンス***
    - ***日本のお家芸だった技術の将来像をシーズとして提供***

アメリカ合衆国労働統計局雇用統計2019/05

https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htmなどより作成

- 米国政府職業小分類867職種中：統計家は第5位、情報セキュリティアナリストが第10位の成長率予測
  - 第1位は風力タービンサービス技術者（7000名：62%成長）
  - 統計家も2018年から職業分類に追加されたデータサイエンティストも数理科学職（主として産官で活躍）
  - アメリカ統計学会(ASA)会長は、産官学の持ち回り

アメリカ合衆国標準職業分類	分類コード	雇用者数	10 年成長率	トップ産業雇用者数
統計家	15-2041	42,700	34.6%	科学研究開発, 6190
データサイエンティスト/その他	15-2051/99	33,200	30.9%	計算機システム設計, 9100
オペレーションズリサーチアナリスト	15-2031	105,100	24.8%	企業マネジメント, 9930
アクチュアリー	15-2011	27,700	17.6%	保険会社, 11,140
数理科学（中分類）	15-20XX	211,700	26.5%	
情報セキュリティアナリスト	15-1212	131,000	31.2%	計算機システム設計, 36,280
計算機科学（中分類）	15-12XX	4,633,400	11.5%	

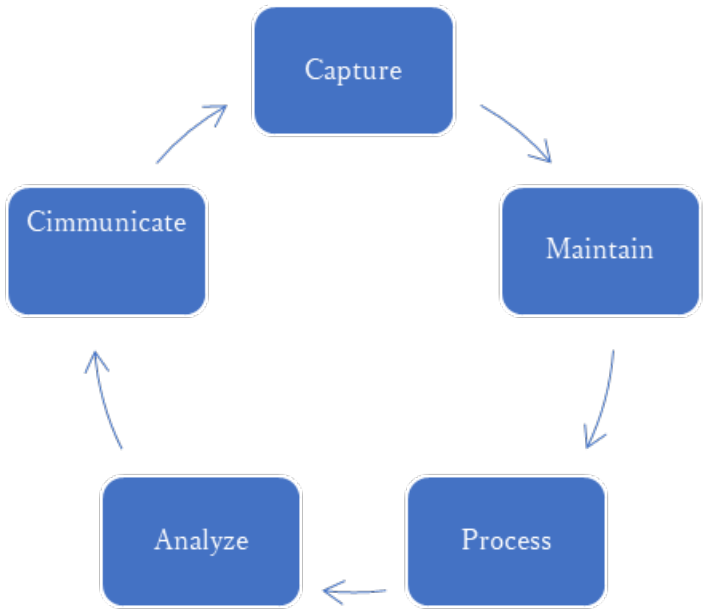
米国労働統計局が示す社会における「統計家(15-2041)」の役割

- 特定の質問や問題に答えるために必要なデータを決定
- ビジネス、工学、科学、その他の分野における実用的な問題を解決するために、数学理論と数理技術を適用
- データ収集のための調査・アアンケート（設問の設計、ターゲットとなる集団からの適切な標本の決定。サンプリング、意思決定に必要なサンプルの大きさの決定）、実験を計画
- データ分析のための数理モデル・統計モデルの開発
- データを解釈し、専門家および専門家ではない者に分析をレポート
- 統計ソフトウェアを利用したデータを分析し、トレンドや関係性を明らかにし、データの妥当性や限界も検討
- レポート（ビジュアライゼーション、集計表）を作成して、ビジネス等の意思決定を支援
  - 42,700名中：全米146統計学専攻の修士統計等取得者が65%、博士取得者が20%、学部卒業が15%

データ・サイエンティスト(15-2051)のミッション

- データ・オリエンテッド・プログラミング言語と視覚化ソフトウェアを使用して、原データを意味のある情報に変換するための一連の手法または分析アプリケーションを開発および実装
- データマイニング、データモデリング、自然言語処理、機械学習を適用して、大規模な構造化データセットと非構造化データセットから情報を抽出して分析
- データの調査結果を視覚化、解釈、および報告
- 動的データレポートを作成
  - ただし、『統計家（15-2041）』、『地図製作者および写真製作者（17-1021）』および『健康情報技術者および医療登録者（29-9021）』を除外
  - 現時点では学部卒業生が主体
- データサイエンティスト協会のスキルセットよりは狭い
  - データサイエンス力×データエンジニアリング力×ビジネス力
  - https://www.datascientist.or.jp/common/docs/skillcheck_ver3.00.pdf

米国データサイエンス大学院ランキング
1位：UC Berkley:　オンライン大学院



データサイエンスのライフサイクル
これができるエキスパートを育成

- Capture
  - データ取得、データ入力、信号入力とデータの抽出
- Maintain
  - データウェアハウスの作成、分析可能なデータの編成（クレンジング、ステージング）、データ処理、データアーキテクチャー
- Process
  - データマイニング、クラスタリング／分類、データモデリング、データの要約
- Analyze
  - 探索的分析と検証的分析、予測分析、回帰分析、テキストマイニング、質的分析
- Communicate
  - レポーティング、可視化、ビジネスインテリジェンス、意思決定

修士課程（産業界への人材輩出）そのカリキュラム

DS修士課程基礎コース

2.5(統計学)： 2（計算機科学）： 0.5(応用AI)

科目名	キーワード
DSプログラミング入門	オブジェクト指向、Pythonによるデータ分析、モジュール等の開発、Jupyter Notebookによるプレゼン、GitとGithub
研究計画とデータや分析への適用	研究計画、研究の問いかけの定式化、データと意思決定、認知バイアスの理解、説得とアクションのためのデータ、データと固有知識の統合、データで物を語る
DSのための統計学	研究計画、様々な量的研究方法と統計的分析技法、記述統計、推測統計、サンプリング、実験計画、仮説検定、最小二乗法と回帰分析、ロジスティック回帰、Rによる実社会データ分析実践
データ工学の基礎	データセットの保存、管理、処理システムの基礎、分析によるソリューションのアーキテクチャー、分散データ処理、リレーショナルデータベース、グラフデータベース、ストリーミング分析、クラウドコンピューティング
応用機械学習	実験計画、機械学習アルゴリズム、特徴量エンジニアリング、予測か説明か、ネットワーク解析、協調フィルタリング、

DS修士課程発展コース

2: 1: 2: 2(コミュニケーション・倫理・マネジメント)

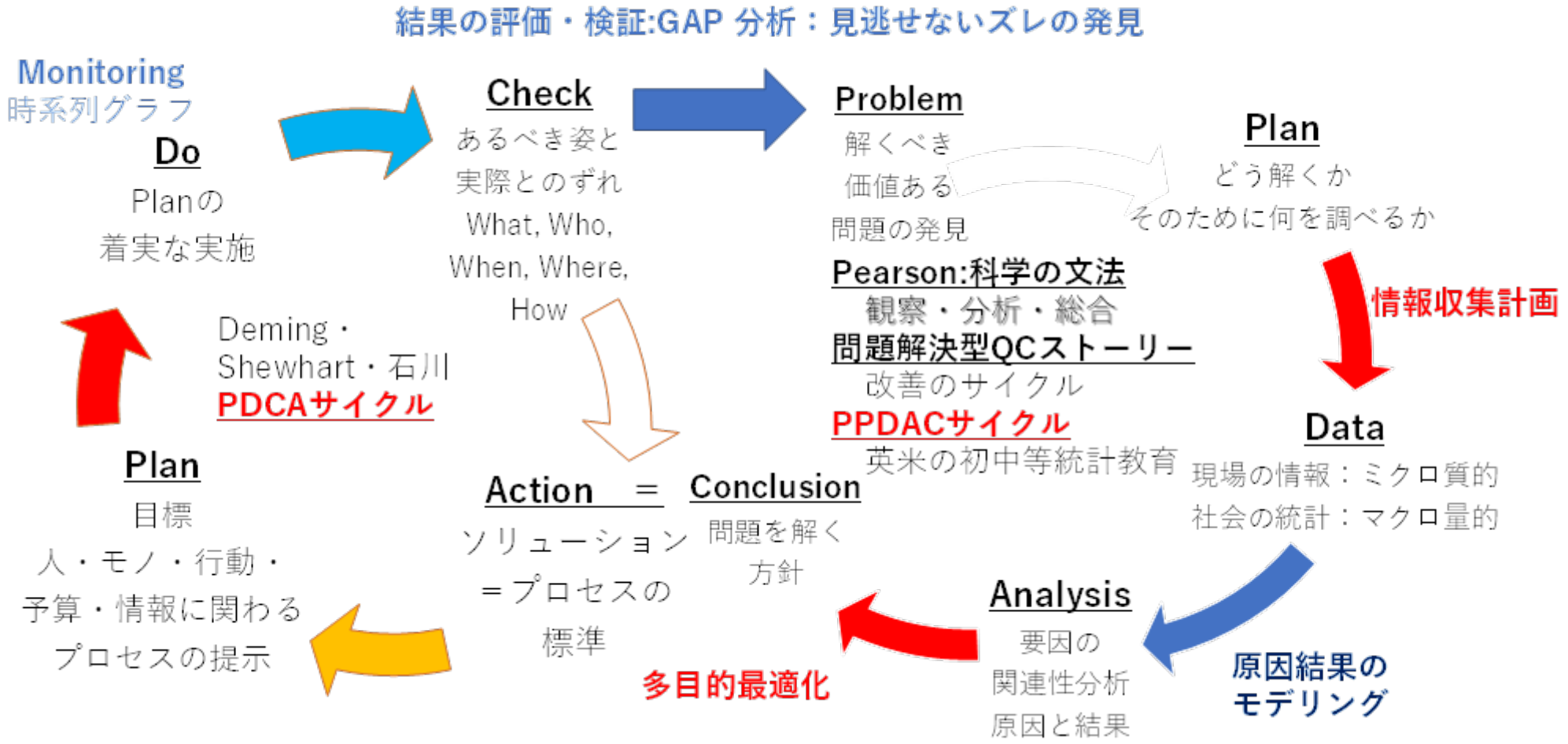
科目名	キーワード
実験と因果推論	実験計画に基づく因果関係の推論。無作為化、統計分析と結果の伝達、データのクリーニング、データマイニングと探索
データの背景：人間と価値	倫理と法的枠組み、政策分析、DSライフサイクル全体での倫理問題、データの収集やタイプなどへの倫理・政策的制約、ケーススタディ
クラウドとエッジでの深層学習	画像やビデオなど大規模データの深層学習などの処理環境の実践、クラウド、分散ストレージ、イーサリアムブロックチェイン、Keras,IBM Watson,利用可能なApacheのソフトウェア
質的回答・時系列・パネルデータのための統計的方法	断面的データと時系列データの可視化、確率と数理統計の重要概念、古典的回帰分析、変数変換、モデルの規定、因果推論、操作変数法、時系列モデル（ARIMA、GARCH、VARモデル）、統計的予測、時系列データを伴う回帰分析
大規模機械学習	単一あるいは複数計算機上での機械学習アルゴリズムの実装、AWS,テラバイト級データの作業における問題点、ペタバイト級データの機械学習パイプライン、アルゴリズムの設計（決定樹、グラフ処理、最急降下法）、並列処理
深層学習を用いた自然言語処理	言語現象の紹介と機械学習による分析、情報抽出、機械翻訳、センチメント分析、要約
データの可視化	探索的データ分析、効果的な文書コミュニケーション、効果的なデータの可視化プレゼンテーション、人間の知覚のデザイン

Berkleyの企業幹部向けDSリテラシーコース 10週間70時間オンライン2850$:この程度

- 確率的意思決定
  - データ分析の前に離散・連続データの比較等DSの背後にある基礎概念等
- 標本データの生成
  - 用語、抽出方法の種類、標本データの知る、第1種の過誤と第2種の過誤等
- 仮説検定
  - 仮説検定、信頼区間、実験の基本原理、4M（動機、方法、メカニズム、メッセージ）モデルによる問題の定式化
- 標本データからの情報推論
  - 直線パターンや曲線パターンの探索とデータに線形モデルを当てはめる様々な方法の理解と産業界での応用
- 基本的回帰モデル
  - より精緻なビジネス意思決定の理論の機関となる単回帰分析の使い方、
  - その利用上の注意とビジネスの意思決定がどう改善するのか
- 発展的回帰モデル:重回帰分析の様々な利用方法
- 予測と機械学習
  - 機械学習の基礎原理と様々な応用を分かりやすく解説、教師付き学習と教師無学習、時系列回帰予測等
- A/BテストとのDSチームの効果的構築
  - 組織のデータ主導型文化構築、データサイエンティストとの効果的連携戦略と陥りやすい問題

日本の製造業界の統計的管理・問題解決・人財育成

「Deming-石川の「統計哲学」は統計学の産業界に対する最大の貢献である」, V. Nair(国際統計学会会長講演, 2015)



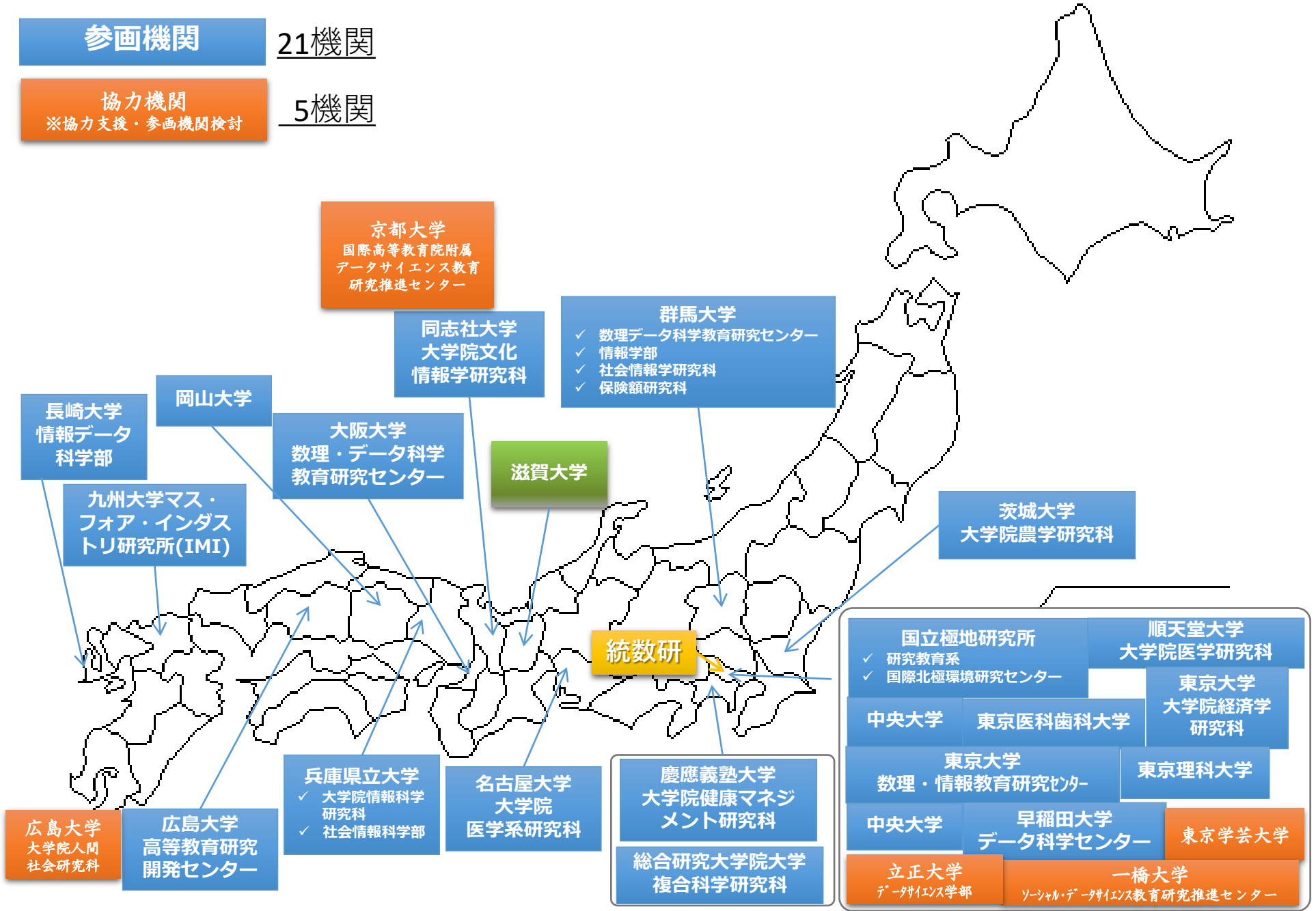
大学共同利用機関情報・システム研究機構
統計数理研究所(ISM)のヒトつくりとコトつくり

- 1943/11　学術研究会議で統計数学を中心とする研究所設立建議
- **1944/06**　勅令第385号統計数理研究所官制交付：文部省直轄研
  確率数理とその応用研究，研究連絡・統一促進
- 1955/04　広尾移転　（上野帝国学士院⇒小石川細川亭⇒三軒茶屋）
- 1985/04　（国立）大学共同利用機関
- 1988/10　**総合研究大学院大学**設置・**統計科学専攻**として参画
- 2004/04　**大学共同利用機関法人情報・システム研究機構・**
  統計数理研究所
  （国立遺伝学研究所，**国立極地研究所**，国立情報学研究所），
- 2005/04　**リスク解析戦略研究センター(椿センター長,筑波大)**設置
- 2009/10 立川(アカデミック・プラザ）への移転
  （国立極地研究所，国文学研究資料館，国立国語研究所）
- 2011/11　**統計思考院設置（統計思考力育成事業開始)**
- **2017/07**　**ものづくりデータ科学研究センター：Materials Informatics**
- ***2021/06　統計エキスパート人材育成事業(5年間）受託***
- ***2021/10　８大学・1研究所から11名の助教授・准教授研修1期生***
- ***2022/01　大学統計教員育成センター設置（千野雅人先生：センター長）***

参画機関および協力機関

参画機関 21機関

協力機関
※協力支援・参画機関検討
5機関



統計エキスパート人材育成コンソーシアムの概要

【日本の課題】 **統計エキスパート育成が大学統計教員育成機関が僅少のため困難**

【統計学博士号年間取得者数】
年間取得者数 米国600名超 ⇔ 日本５名程度
（統計数理研究所/総合研究大学院大学）

【教員候補数】
アメリカ統計学会19,000名 ⇔ 日本統計学会1,400名
（うち経済系700名）

【統計学部の数】
・アメリカ：138学部（大学院研究科＝146科）
一部の大学はデータサエインス学部化
また，統計学科・生物統計学科の2学科設置も有
・イギリス・韓国：50学部程度
・中国：300学部超（米中では統計学部が急増）
☆アジア・中近東・アフリカ諸国でも統計学科は存在
・日本：専攻(博士後期課程)レベル⇒総研大統計科学専攻のみ（2019年まで）（定員：5名）

【事業の背景】
AIを支える現代統計学の教育研究指導可能な大学統計教員と大学院レベルのエキスパート育成システムが欠如

【事業の概要】
大学統計教員育成システムと大学院での統計エキスパート育成システムを同時開発し、大学統計教員育成研修を実施
⇒大学統計教員の育成
質保証された大学統計教員を大学の統計エキスパート育成システムに投入
⇒統計エキスパートの育成

データ社会における国際競争力低下

実施事業
統計エキスパート育成システム
大学統計教員育成システムからなる
統計エキスパート育成エコシステムの開発と運用

コンソーシアムが目指す統計エキスパート育成エコシステム

情報・システム研究機構が主として行う事業範囲

大学統計教員としての教育を受ける前

PD・助教

半年の研修と1.5年のFD

中核機関
大学統計教員育成研修

シニア教員

統計コミュニティトップレベルの教員

参画機関が主として行う事業範囲

参画機関
統計エキスパート育成
統計エキスパート育成システム運用

大学統計教員

大学に必要な統計の講義・共同研究ができるレベル

学生

DS的素養は全くないか低い

【現状】統計エキスパートでもあるDS・AI技術者を育てたいが、大学統計教員が不足
統計が専門でない教員が授業担当する弊害

統計エキスパート

社会へ輩出

実社会で正確な業務を遂行できる統計知識を有する

はレベルインジケータ：上に行くほど専門的でレベルが高いが人口は少なくなる

大学統計教員育成の仕組み
・大学統計教員育成システム構築とその改善のPDCAサイクル確立
・大学統計教員としての研究力向上
・質保証された大学統計教員の認定
・統計エキスパート育成システムの開発

大学統計教員に要求される力量・技能の獲得
(1)分野共通の統計学的知識→不足知識体系を習得
(2)統計教育力→教材開発や教育実習などによる教育力育成
(3)専門分野における統計的共同研究→論文執筆能力および研究指導力の育成

【大学に必要な大学統計教員の育成】
1. 全学教育統計科目の教員
2. 工学部、医学部、経済学部など統計的研究が必要不可欠な学部・大学院の教員
3. 既設・新設のデータサイエンス学部等の教員
4. 1～3の教員を再生産する教員

事業目標：事業年度内に認定された大学統計教員30名輩出
統計エキスパート育成システム（教材・共同研究体制等）で、大学統計教員１名当り３名以上創成可能なエコシステム確立

米国
自動車殿堂WEB

Genichi Taguchi

Inducted 1997

Served as a consultant to both Japanese and American auto manufacturers, leading numerous companies to the adoption of Taguchi quality control methods

In 1960, received the Deming Prize, Japan's highest honor, for his contributions to the field of quality engineering; also received the Deming Literature Award three times for books on quality control methodologies and industrial design

Dr. Genichi Taguchi brought a new sense of understanding to the automotive industry by unlocking the mystery of statistics. Born in Japan's Niigata prefecture, Taguchi attended Kiryu Technical College before joining the Imperial Japanese Navy. After World War II, Naguchi's scientific career began with employment at the Japanese Ministry of Education's Institute of Statistical Mathematics. Later, focusing on automotive engineering with emphasis on the use of statistics in quality control, Taguchi created books and programs that served auto manufacturers around the world. Beginning in the 1950s, seminars promoting Taguchi methodologies conducted by both Japanese and international quality control and statistical standard organizations trained thousands of experts in Taguchi methods. In addition to stimulating discussion at universities and educational institutions throughout the world, Taguchi methods have been instrumental in the development and continuing improvement of quality control programs of auto manufacturers including Toyota Motor Company, Mitsubishi Motors, and Ford Motor Company.

最適なデータを創る数理：実験計画

- 1919年：R. A. Fisher(1890-1962)英国ロザムステッド農事試験場に就職
- ギネスビール：Gossett（ペンネームStudent）
  - 最先端の農事試験場がベストと考えて推奨した品種
  - スコットランドでは全滅
  - 実験データが示す結果の「一般化可能性(Generalizability)」を担保
- 基礎研究が、製品になるまでには越えることが困難な「死の谷」
- Fisherのデータ創成(実験計画）3原則
  - 無作為化原則：品種や処置の割り付けはできるだけランダムに決める
  - 繰り返しの原則：実験を繰り返し、ノイズの影響を平均化によって減少
  - 局所管理の原則：繰り返し実験を行うにしても、同一環境中で、比較したい一そろいの実験を行う
- 一要因実験より多要因同時実験が技術情報取得効率が良い

最適実験計画としての直交計画
米国の30倍以上のスピードで技術開発

- 農業実験計画から工業技術開発実験へ
  - 日本の貢献と奇跡
  - 1950年：日本で直交表実験による技術開発開始
    - ペニシリンの製造効率・電電公社交換機リレー・伊那製陶・建機自由化に向けた最適化等々
      - 増山元三郎→**田口玄一**
        （統数研→電電公社西堀特別研究室）
- Kiefer：　1959以降
  - 最適実験計画：応答予測精度最適化計画
    - **D-Optimal Design; 近年機械学習分野でも利活用**
    - 直交表計画」単純な統計モデルが正しければ D-Optimal
- 21世紀：数値実験（シミュレーション）
  - 複雑な応答曲面最適化：数値実験計画法
  - Kai-Tai Fangの「一様実験」：より効率的な実験

L8直交表実験：8回の実験で
1因子実験64回を行ったと同じ情報量を獲得

測定値	測定条件						
	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Y_1	$x_{20}+\delta_2$	$x_{30}+\delta_3$	$x_{40}+\delta_4$	$x_{50}+\delta_5$	$x_{60}+\delta_6$	$x_{70}+\delta_7$	$x_{80}+\delta_8$
Y_2	$x_{20}+\delta_2$	$x_{30}+\delta_3$	$x_{40}+\delta_4$	$x_{50}-\delta_5$	$x_{60}-\delta_6$	$x_{70}-\delta_7$	$x_{80}-\delta_8$
Y_3	$x_{20}+\delta_2$	$x_{30}-\delta_3$	$x_{40}-\delta_4$	$x_{50}+\delta_5$	$x_{60}+\delta_6$	$x_{70}-\delta_7$	$x_{80}-\delta_8$
Y_4	$x_{20}+\delta_2$	$x_{30}-\delta_3$	$x_{40}-\delta_4$	$x_{50}-\delta_5$	$x_{60}-\delta_6$	$x_{70}+\delta_7$	$x_{80}+\delta_8$
Y_5	$x_{20}-\delta_2$	$x_{30}+\delta_3$	$x_{40}-\delta_4$	$x_{50}+\delta_5$	$x_{60}-\delta_6$	$x_{70}-\delta_7$	$x_{80}+\delta_8$
Y_6	$x_{20}-\delta_2$	$x_{30}+\delta_3$	$x_{40}-\delta_4$	$x_{50}-\delta_5$	$x_{60}+\delta_6$	$x_{70}+\delta_7$	$x_{80}-\delta_8$
Y7	$x_{20}-\delta_2$	$x_{30}-\delta_3$	$x_{40}+\delta_4$	$x_{50}+\delta_5$	$x_{60}-\delta_6$	$x_{70}+\delta_7$	$x_{80}-\delta_8$
Y8	$x_{20}-\delta_2$	$x_{30}-\delta_3$	$x_{40}+\delta_4$	$x_{50}-\delta_5$	$x_{60}+\delta_6$	$x_{70}-\delta_7$	$x_{80}+\delta_8$

逐次実験計画法から強化学習へ
欧米の強み

- 逐次推論：Wald, A. (逐次抜取検査：米国軍事機密）：戦時中
  - 抜取検査による合否判定を最小の抜取検査サイズで実現
- 1950年代：Box, G.E.P., Evolutional Operation, Response Surface Design
  - 英国ICH, 化学工業における最適小実験の逐次利用による品質の山登り
- Armitage, P. 逐次臨床試験
  - 患者さんが一人ずつ臨床試験に参加
  - できるだけ多くの患者さんに有効率の高い薬を投与
  - A薬の方が有効率の高いことが統計的に検証出来たら、後は全ての患者さんにA薬を投与
- Bandit Problemとして逐次最適化実験の枠組み慣性
  - Berry & Fristedt (1985): bandit Problems:Sequential Allocation of Experiments, Chapman and Hall.
  - 推論と最適データ収集の一体化
- 強化学習近年の注目

ロバストパラメータ設計から敵対的学習へ
日本の産業界の強みなのだが

- 田口玄一：1980年代後半：統計学からの決別
  - ロバスト・パラメータ設計（RPD, ISO 16336）
    - 制御因子×ノイズ因子（制御因子の制御を邪魔する要因）：直積型直交実験
      - 制御因子とノイズ因子のゲーム論的構造を実験計画に導入：1950年代後半
    - SN比：ノイズ因子に寄与する機能のバラツキの尺度：1980年頃
    - 制御因子側（技術者側）：SN比最適化（安定化）設計
    - ノイズ因子の調合：一番制御因子にとって不都合なノイズ因子の組み合わせ条件
  - ノイズ因子に対してロバストな技術開発：膨大な産業界適用
- Goodfellow et al.(2014)：GAN
  - 敵対的学習・敵対的生成ネットワーク（GAN）
  - 機能を実現するAIとそれを邪魔するAIのゲームの均衡点で
ロバストなAI実現

統計数理研究所
ものづくりデータ科学研究センター

- 新時代のデータ創造型データサイエンスのトップランナー
  - 実験計画法・RPDを超えて新たな産学連携の拠点
  - ***マテリアルズ・インフォマティックス拠点***
- 情報・システム研究機構が提供するシーズ
- オリエンテーションはこれ位にして
  - 本番をお楽しみください
    - なお、この種のデータサイエンスに関する放談は
    - 月間「アイソス ISOS」：2021年10月号~12月号に書きましたのでご参照下さい