

データサイエンス教育と 情報教育

萩谷昌己

情報処理学会 情報処理教育委員会 委員長

東京大学数理・情報教育研究センター 基盤情報部門 部門長

数理及びデータサイエンス

- 文部科学省「数理及びデータサイエンスに係る教育強化」
 - 拠点6校：北大・東大・滋賀大・京大・阪大・九大
 - 協力20校
 - 予算：6億円 ⇒ 9億円
- 東大 --- 数理・情報教育研究センター
 - 数理情報部門
 - 数学基礎教育部門
 - 基盤情報部門
 - 応用展開部門

情報処理学会

- 情報処理教育委員会
 - J17WGデータサイエンス委員会
 - 委員長：中野美由紀（津田塾大学）
 - データサイエンス教育委員会（設立予定）
 - 委員長：加藤浩（筑波大学）
- CITP：Certified IT Professional --- IFIP IP3の認定を取得
 - データサイエンティスト資格検討WG
 - 委員長：掛下哲郎（佐賀大学）
- J17 --- 情報処理学会が2017年度に公開したカリキュラム標準

統計と情報の悩ましい関係

- データサイエンス教育は誰がやるべきか
- 大学の場合（特に全学教育）
 - 統計学者がやるべきか
 - 数学者がやるべきか
 - 情報学者がやるべきか
- 高等学校の場合
 - 数学科でやるべきか
 - 情報科でやるべきか

データサイエンス教育

- ACM (Association for Computing Machinery) も困っている
 - 多様なカリキュラム
 - (1) 人工知能を基盤とするカリキュラム
 - (2) データベースを基盤とするカリキュラム
 - (3) 統計を基盤とするカリキュラム
 - 統一カリキュラムを出せていない
- CC2020 (Computing Curricula 2020) でも検討されている
- 情報処理学会J17でも報告されている
 - しかしカリキュラムは作られず
 - 米国統計学会の学部向けカリキュラム・ガイドラインを参照するのみ
(次のスライド)

データサイエンスの6つの主要課題分野

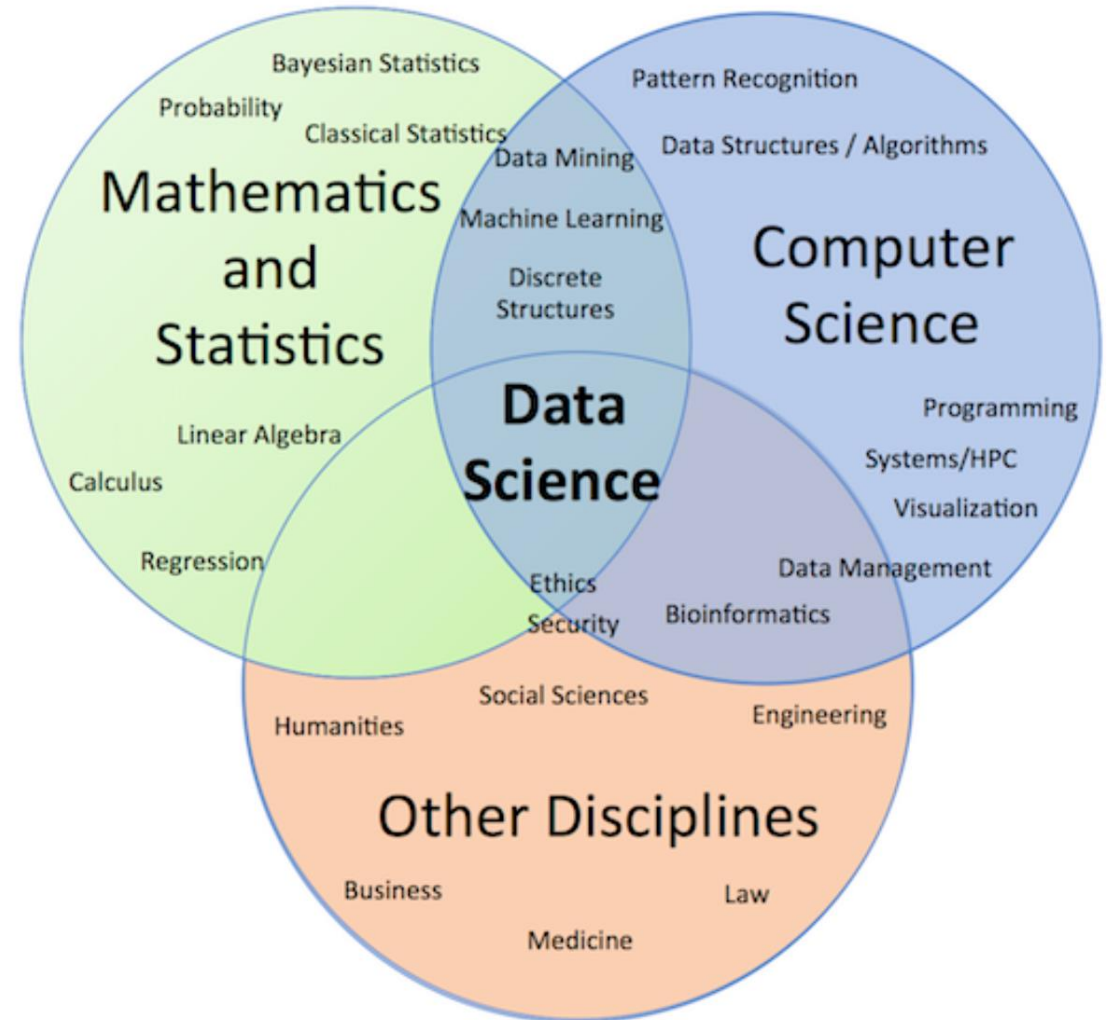
(J17報告書より)

1. データデスクリプションとキュレーション
2. 数学的基礎
3. 計算論的思考
4. 統計的思考
5. データモデリング
6. コミュニケーション、生産性と倫理

- ほとんどは情報系のカリキュラムに含まれる

Computing Competencies for Undergraduate Data Science Curricula

- ACM Data Science Task Force
- Initial Draft
- January 2019



Knowledge Areas for Data Science

- Computing Fundamentals, including
 - Programming Data Structures, Algorithms, and Software Engineering
- Data Acquisition and Governance
- Data Management, Storage, and Retrieval
- Data Privacy, Security, and Integrity
- Machine Learning
- Data Mining
- Big Data, including
 - Complexity, Distributed Systems, Parallel Computing, and High Performance Computing
- Analysis and Presentation, including
 - Human-Computer Interaction and Visualization
- Professionalism

Computing Competencies for Undergraduate Data Science Curricula

- データサイエンスに必要なコンピュータサイエンスの素養をまとめたもの
 - プログラミング・データベース・情報セキュリティ
- 機械学習とデータマイニングは入っているが、
- 統計教育に踏み込んではいない（cf. 高等学校「情報科」）
- “This project looks at data science from the perspective of the computing disciplines, but recognizes that other views contribute to the full picture.”
- “This ACM Data Science report builds on the Park City work with a heavy orientation toward computer science.”

高等学校「情報科」次期学習指導要領

- 情報Ⅰ

- (1) 情報社会の問題解決
- (2) コミュニケーションと情報デザイン
- (3) コンピュータとプログラミング
- (4) 情報通信ネットワークと**データの活用**

- 情報Ⅱ

- (1) 情報社会の進展と情報技術
- (2) コミュニケーションとコンテンツ
- (3) **情報とデータサイエンス**
- (4) 情報システムとプログラミング
- (5) 情報と情報技術を活用した問題発見・解決の探究

情報科におけるデータサイエンス教育

- 情報Ⅰ(4)(ウ)
 - データを表現，蓄積するための表し方
 - データを収集，整理，分析する方法
 - 結果の表現の方法
- 情報Ⅱ(3) 情報とデータサイエンス
 - 学習指導要領の「解説」では、多変量解析や人工知能・機械学習まで取り上げられている
 - 回帰（重回帰分析）
 - 分類（条件付確率・近傍法・木構造）
 - クラスタリング（階層的方法・非階層的方法）
 - 画像・音声・Web・GIS

情報教育の中での
データサイエンス教育

情報教育の中での データサイエンス教育の可能性

- 問題解決 → モデル化とシミュレーション --- 情報教育の中核
- 統計の基本概念を確認
 - 記述統計量・帰無仮説・有意水準・p値
- それらをプログラミングにより体験する
 - モデル化とシミュレーション
 - 特にモンテカルロ法
 - データの可視化
 - 基本的な性質を再発見させることも
- 作ったコードをツールとして利用
- 培ったプログラミング能力を活用

Simulation Based Inference (SBI)

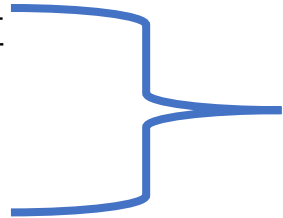
- シミュレーションを通して統計的推論の論理を理解させる授業方法
 - ジミー・ドイ: シミュレーションに基づく統計的推論とアクティブ・ラーニングの授業事例, 数学科におけるデータサイエンス(1), 日本数学教育学会誌, 101(3), pp.28-39
- トピック例
 - 乳児に対する実験 --- 二項検定
 - 研究者が乳児の前に助けてくれるキャラクターと邪魔するキャラクターを提示し、乳児がどちらのキャラクターに最初に手を出すのかを調べる
 - ベンフォードの法則
 - 2^n の最上位の桁の分布
 - 同事象の連続出現の回数 (Longest Run)
 - 公正なコインによる表と裏の連続

高等学校「情報Ⅰ」でも十分に可能

- 統計の基本概念の理解を確認
 - 確率モデルのシミュレーション（モンテカルロ法）
 - 繰り返しと配列の簡単なプログラミング
 - 簡単な視覚化（グラフ表示）
-
- 工夫により少ない授業時間で効果的に
 - 大学の数理・データサイエンス教育にも

情報教育を基盤とするデータサイエンス教育

- 情報教育の幅広さ
- 特にモデル化とシミュレーション
 - 統計の基本概念の体験・再発見
 - ここで統計教育のノウハウが重要
 - データ分析の実践
 - プログラミング教育のノウハウ
- ツールとしてのプログラミングの重要性
 - データのクレンジング・可視化・キュレーション
 - より一般的には計算論的思考の重要性
- 人工知能・機械学習への自然な接続
 - 特に最適化



情報教育と統計教育の協力