データサイエンス普及の隘路

データサイエンスコンソーシアム，慶應義塾大学

柴田　里程

バズワード（英: buzzword）

日本語での意味は、もっともらしいけれど実際には定義や意味があいまいな用語のことで、英語では、特定の期間や分野の中でとても人気となった言葉のことである。権威付けされたり、専門用語や印象付けるような技術用語。コンピュータの分野でよく使われるが、政治など広い分野で使われる。1940年代半ばのアメリカのスラングが起源。

(Wikipedia)

情報，IT，ICT，ビックデータ，AI, …

データサイエンス (data science)

- バズワード？　新しい科学分野？

- バズワードの副作用
  - データサイエンス==統計学
  - データサイエンス==データエンジニアリング
  - データサイエンス==最高の学問
  - データサイエンス==Nothing
  - データサイエンス==自分には無関係

データ (Data)

- 数字，記号で表した推論(action)の根拠となるもの
  - ただ存在するだけではデータではない．「情報」でしかない
  - 整理されていないとデータとはいえない．

- データベースのデータ(data)
  - 業務の根拠，必ずしも推論の根拠ではない
  - 組織化され，運用されている
- 地図，書籍(raw data)
  - 整理はされているが，推論の意図は存在しない
- 眠っているデータ (pre-data)
  - クラウド上に集まってしまった，調査の結果としてたまっている
  - どう活用したらよいかわからない

データ学(dataology)

- データサイエンス(data science)
  - データに関して生まれる「なぜ」,「どうして」に答えを追求する科学

- データエンジニアリング(data engineering)
  - 目標に対して最適なデータ取り扱い技術を探索する研究分野

統計学

データの海に浮かぶ孤島

海面上昇で消え去るかもしれない？

Dr D.R. Cox
“Statistics survives but not necessarily by statistician”

データサイエンス普及の隘路

- 日本では「データにもとづく客観的な判断」が未熟
- 「データはだれでも扱える，サイエンスなど必要ない」という思い込み
- 扱ってみると，その入り口で厄介さに音を上げ，放り出す
  - 理論の不備
  - ソフトウエアの不備
  - 退屈

理論が存在しないサイエンスなどありえない

理論を実装したソフトウエア環境は必ず存在する

データサイエンスの基礎理論

- データモデル
- データの属性
- データフロー
- データの視覚表現
- データの意味と質

データモデル

- 標準形
  - 関係形式 (Relational Scheme), ER(Entity Relation)モデル: データテーブル
- 派生形
  - 配列
  - 分割表
  - API
  - リスト

- 研究課題:
  - RDB の正規形をどのように乗り越えるか
  - 依存関係の排除と関係の探索はバッティングする
  - 冗長性の排除と効率化
  - 欠損値の扱い
  - 整合性の確保

データの属性

- 階層別
  - 値
  - ベクトル（カラム）
  - データテーブル
  - データベース
- 種類
  - 型
  - クラス
  - ShortName, LongName
  - 単位　　．．．
- 研究課題:
  - 必要十分な属性？
  - その表現？

- TRADにおける型
  - Measurement
    - Measurement
    - Cardinal
    - Ordinal
    - Frequency
  - Mark
    - Mark
    - Ordered Mark
    - Logical
  - Time
    - Time
    - Elapsed Time

データフロー

- データの変容
  - Raw, 1^{st}, 2^{nd} , 3^{rd} ,…
  - 正規化
    - 一つの変量の値の列としてデータベクトルを組織化
      - 複数回答
      - 変量の値ごとに度数の列となっている場合
      - 基数系
  - 併合，部分抽出
  - 値の変換
  - 変量の役割
    - ID
    - 補助
- 研究課題: データ変容の定式化

データの視覚表現

- データフローに沿った視覚表現
  - Raw data
  - Cooked data
  - Analysed data
  - Communication
  - Presentation

- 研究課題：各段階での最適な視覚表現，汎用な視覚表現

データの時代

- 第4次産業: データを「材」とする産業
- いまの統計学の枠組みでは時代の変化に追いつけない
  - 官庁統計
    - 記述統計
  - 農業実験
    - 実験計画
  - 品質管理
    - 推測統計
  - 薬効検定
    - 統計的検定
- **新しい酒は新しい革袋に盛れ**(『新約聖書』マタイ伝第九章)