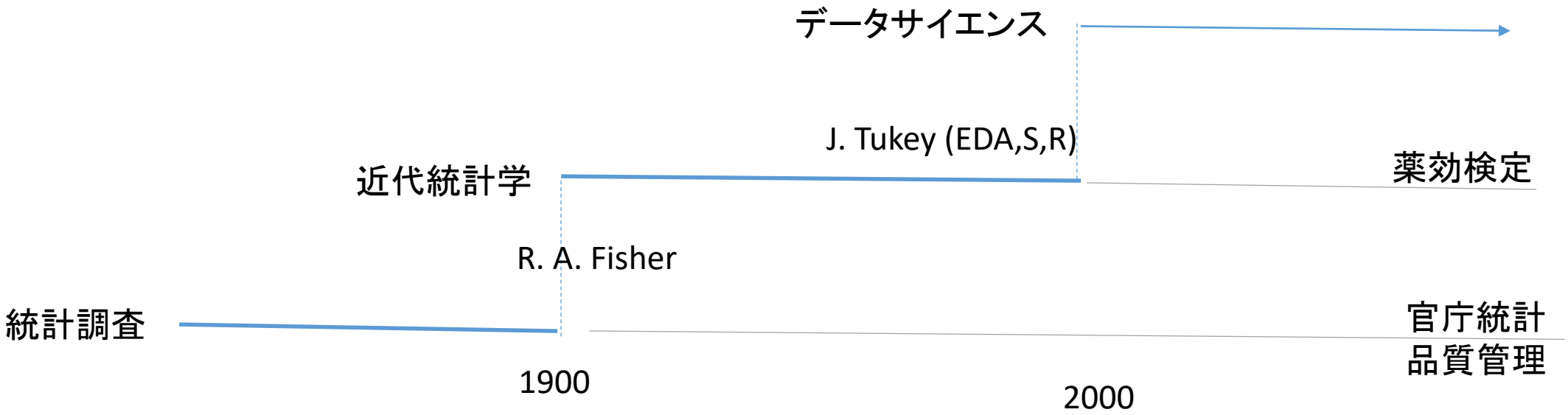
統計学のレガシー

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
北里大学　力丸佑紀

統計学からデータサイエンスへ

- 背景
  - 交通，通信，コンピュータの発展による，データ取得，蓄積の容易化
  - 計算能力，グラフィックス表示能力の画期的な向上
- データサイエンスの課題
  - 平均，分散のような単純な要約から発見へ
  - 蓄積された大規模データの活用
  - データ公開の壁打破
  - 各応用分野との連携体制
- データサイエンスの研究
  - データのサイエンス（データモデル）
  - データの理解を加速させるヒューマンインタフェース
  - データからの発見プロセス

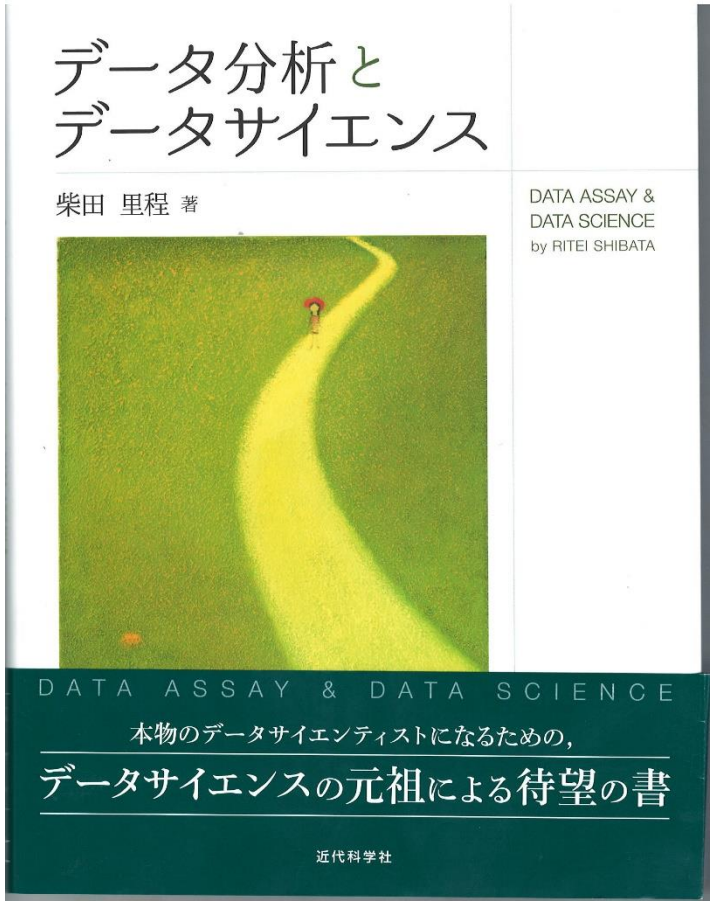

データサイエンス
J. Tukey (EDA,S,R)
近代統計学
薬効検定
R. A. Fisher
統計調査
官庁統計
品質管理
1900
2000

データサイエンスへのパラダイムシフト

- 課題
  - 時代背景をどう反映させるか
  - 旧パラダイムの遺産をどう継承するか，あるいはしないのか
  - 刷新感が本質ではない
- しかし
  - 既存権益の侵害
    - これまでの単なる延長線上での研究では済まない
    - 勉強しなおしが必要
  - 追い付けない
    - 置き去りにされる
  - 感情的な反発
    - これまで通りで何がいけない？

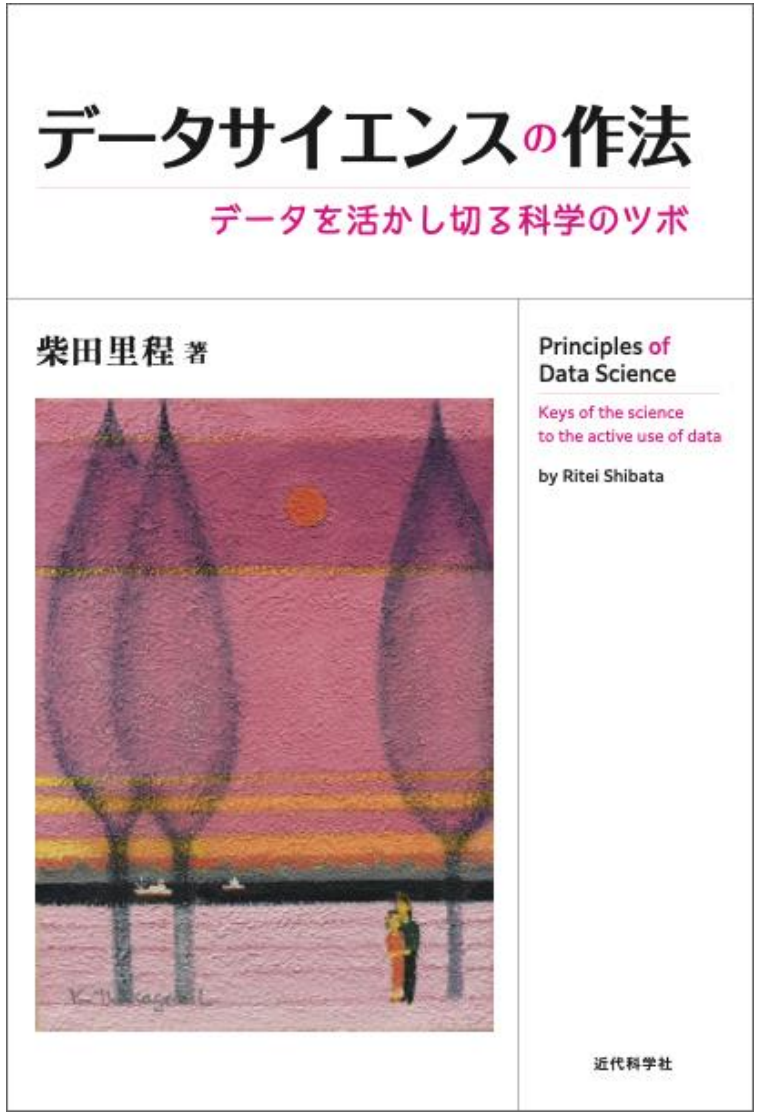
データサイエンスが引き継ぐべき統計学のレガシー

- データ分布，代表値
- 個体(記録)の雲の探索（クラスタリング）
- 変量間の関係の探索
- 変量の相関・偏相関
- 確率モデル
  - 大数法則，中心極限定理，正規分布
  - ポアッソン現象（小数法則）
  - 極値分布



レガシーの継承と発展

- 変量の雲の探索
  - 視覚的な探索(TextilePlot)
    - 視覚
    - 視点
    - 視野
    - 型の見直し
      - 計測，計数，序数，日時，時間
      - マーク，順マーク，論理



モデル論

- モデル(模型)は，対象が大きすぎたり，複雑すぎてそのまま扱うのが困難なときに一つの近似として用いるもの
- 目的により，同一の対象でも異なるモデル
- データサイエンスに登場する3種類のモデル
  1. 論理的に導かれたモデル
  2. データを探索するためのモデル
  3. 観測メカニズムを表現するモデル

3種類のモデル

- 演繹モデル(deduction model)
  - 論理の積み重ねで作りあげたモデル（サイエンスモデル）
    - 物理モデル
    - 宇宙モデル
      - データでこれらのモデルを検証したい
- 帰納モデル(induction model)
  - データからの帰納を助けるモデル
  - 役割：データの理解を助ける，視点・視覚・要約
    - 回帰モデル
- 観測モデル (observation model)
  - データ生成から取得に至るまでを記述したモデル
    - 平面上をスキャンして得られるデータ
    - 実験計画，標本調査
  - 一貫性，一般性

演繹モデル

- 論理的に導かれたモデルが現実と矛盾しないかどうか確かめたい
  - データにもとづく検証
    - モデルの妥当性
    - 正規性の検定
  - あらたな発見は期待していない
    - データの処理
    - 検証のための適切なツール
  - 基本的にそのモデルが妥当であることを期待している
    - 乖離を見逃す恐れがある
    - 「データに正直」でなくなる恐れがある

帰納モデル(作業モデル)

- データから何か帰納したい
  - 具体的な目標がまだ明確ではない
    - どんな現象を表しているか
    - 何が原因か
  - ある程度具体的な目標が定まっている
    - すでに基本的に満たすべきモデルが存在
      - 微分方程式
      - 関数関係
    - しかし，それ以外の要因も影響しているだろう
- 最初は近似としては粗くても扱いやすいモデルが望ましい
  - いい料理には切れ味のよい包丁
- それをブラッシュアップして理解しやすい汎用なモデルに昇華
- モデルのセマンティックス（意味）とその理解
  - データとの対話を助ける道具．それを使いこなす巧みな戦略
    - 視覚，視点，視野

観測モデル(データモデル)

- データ生成メカニズムをなるべく忠実に記述
  - 帰納モデルのようにハンディーである必要はない
  - 一貫性，独立性が重要
- データ取得過程もなるべく忠実に記述
  - スキャニング
  - サンプリング
    - 連続・離散
    - システマティック・ランダム
  - 不完全性
- 変量定義
- 帰納モデル構築とその評価の基盤となる

帰納モデルと観測モデルの例

帰納モデル:

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$

$\mathbf{y} = \rho W\mathbf{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$

$\mathbf{y} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\varepsilon}$, $\mathbf{f}$と$\boldsymbol{\varepsilon}$は独立
＋一意性のための付加条件

観測モデル:

$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_p) + \boldsymbol{\varepsilon}$

$\mathbf{y} = (I - \rho W)^{-1} X\boldsymbol{\beta} + \mathbf{u}$,
$P_\rho(T_1, T_2)u_\nu = \varepsilon_\nu$: 空間定常AR

$\mathbf{y} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \boldsymbol{\varepsilon}$, $\mathbf{f}$: 共通因子と$\boldsymbol{\varepsilon}$: 独自因子は
平均$\mathbf{0}$で, $\boldsymbol{\varepsilon}$ の各要素は無相関.
$\mathbf{y}$ のみが観測される.

頑健性，同定可能性の問題はこの帰納モデルと
観測モデルの関係で考えるとわかりやすい