

データサイエンス作法

慶應義塾大学 柴田里程

データサイエンス

- バズワード(はやり言葉, 受けのよい言葉)
 - コンピュータサイエンスの言い換え
 - コンピュータサイエンスではコンピュータで扱うものは何でも「データ」
 - 統計学の言い換え
 - 統計学はデータの科学
- 何か主張するにはデータの裏付けが必要
 - うまい裏付けとなるデータを探す, 加工する (data driven science)
 - 「データサイエンスによれば」ということで説得力が増す
- AI (深層学習) はデータサイエンス？
 - 大量のデータを扱っているという意味でデータサイエンス
 - しかし, サイエンスとは真逆の精神

データのサイエンス

- サイエンス

- 「なぜ？」から出発して謎を解く
- 「科学的には」という言葉が象徴しているように、サイエンスは客観性を担保するものと思われる
- その理由は、好奇心だけにもとづいて謎を解くから万人が納得できる
- 汎用, 抽象化 → 他の分野に展開可能

- エンジニアリング

- 「なぜ？」はとりあえず置いといて目的達成に全力を挙げる
- It works.

サイエンスでないものをサイエンスと呼んでよいのか？

- データサイエンティスト/データ・サイエンティスト
 - データ・サイエンスはサイエンスではない
- AI（深層学習）
 - なんでも「まねる」ことから始まる
 - 世の中を便利にし、退屈な仕事をなくせる
- エンジニアリング？
 - サイエンスではなくビジネス

サイエンスとしての「データサイエンス」 を発展させ、世の中に定着させよう！

- ブームは去る. しかしサイエンスはブームではない.
- サイエンスは人間の営み
- 人間から「好奇心」がなくなる限りサイエンスは永遠.
- データサイエンスの存在価値
 - データにどう対峙したらよいか, その基本的な枠組みを提示
 - データから発見へ, その道筋を明らかにする
 - データと人間の間に必要なインターフェースを明らかにする.
- データサイエンス作法

TRAD プロジェクト

(<https://datascience.jp/TRAD.html>)

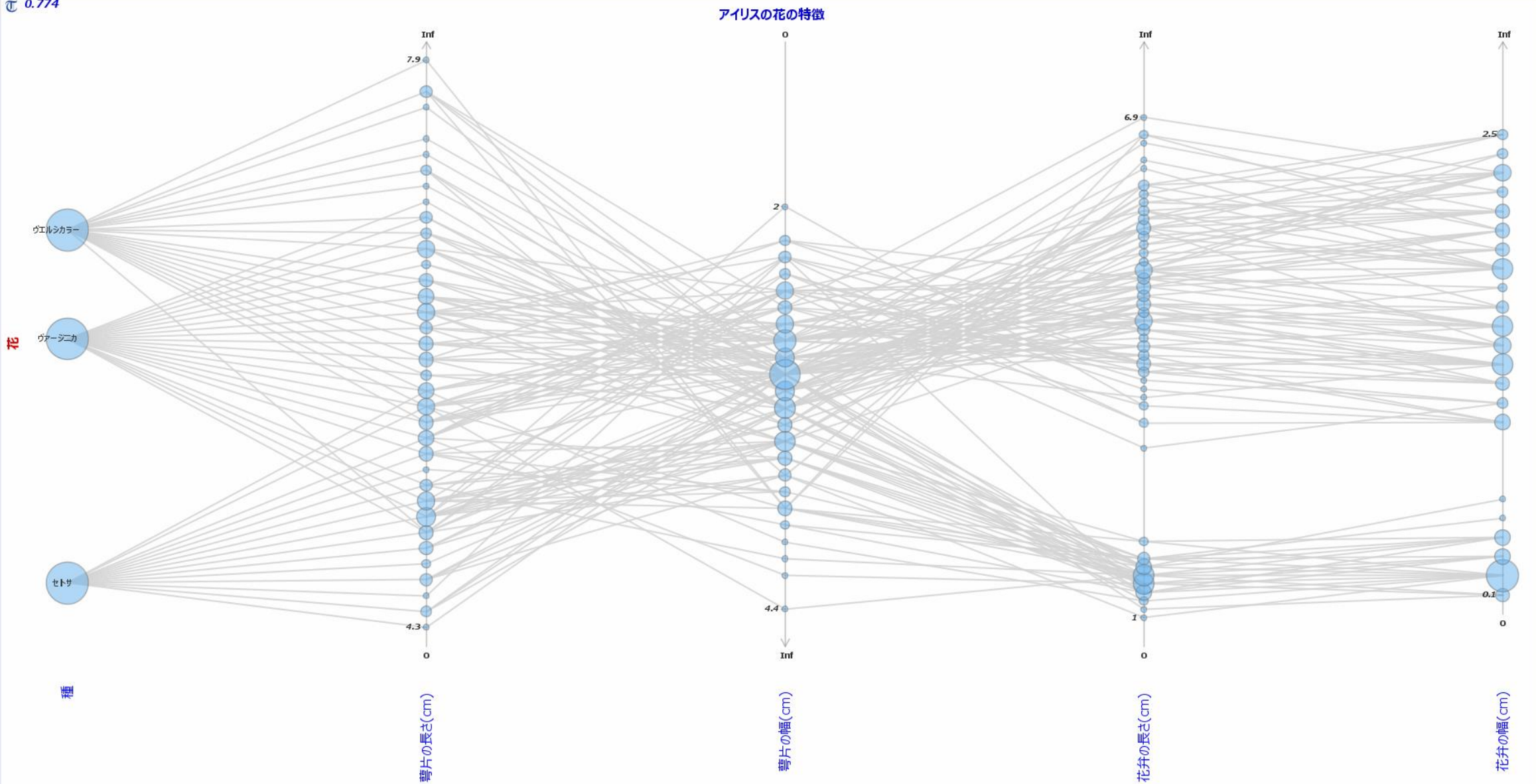
- 次世代のデータ解析環境
 - R: キーボードからのコマンド入力
 - TRAD: マウス操作(GUI) による感覚的なデータ探索
 - Java
 - DandD (データ保持形式)
 - TextilePlot, ParallelCoordinatePlot (データ表示形式)
 - R への GUI の提供と連携
 - RDB へのインタフェース
 - データサイエンスの実践環境



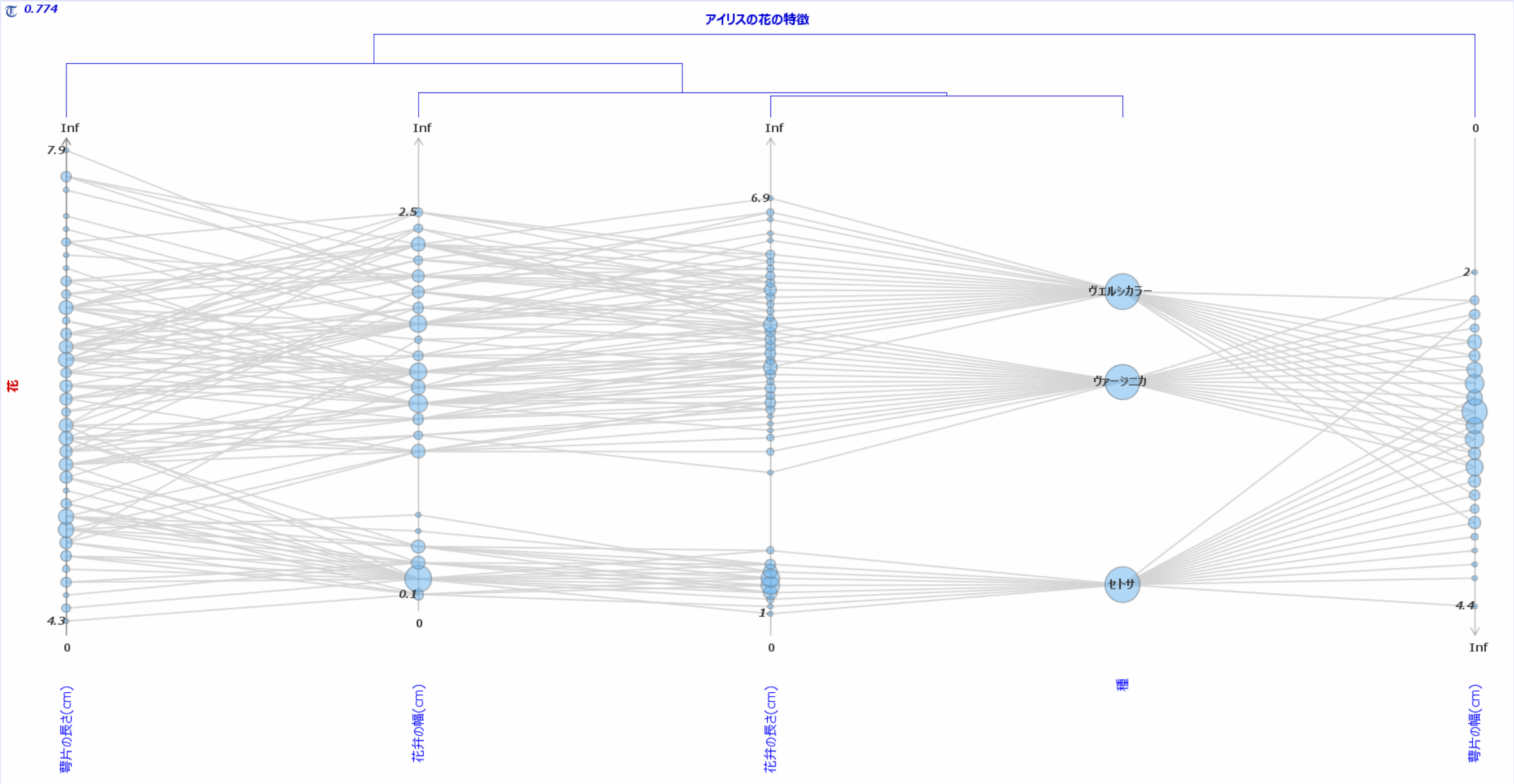
アイリスデータ



0.774



0.774



原子型, 型, クラス

- 目的によって必要な型は変化する
 - コンピュータでの原子型(ハードでの扱いの差異を反映)
 - 論理, 整数, ロング整数, 浮動小数, 倍精度, 文字, 文字列
 - 統計ソフトウェア R での型
 - 数値, ファクター(Categorical)
 - データサイエンスで必要な型？
- クラス
 - 型以外の様々な属性を加えることでクラスを構成できる
 - サブクラス, スーパークラス
 - メソッド

データベクトルの型

- 数値

- | | | |
|------|----------------------|---------------|
| • 計量 | 実数あるいはその一部 | 正規分布 |
| • 計数 | 非負整数 | ポアソン分布 |
| • 序数 | 自然数 | |
| • 時間 | 経過時間 | |
| • 日時 | 特定の日時を起点とする経過時間 | 1990年初頭からのミリ秒 |
| • 度数 | 自然数: その記録の1次データでの重複数 | 分割表 |

- 非数値

- | | |
|--------|--------------------|
| • マーク | マーク値 |
| • 順マーク | 順序のあるマーク値 |
| • 論理 | TRUE, FALSE の2値マーク |

単位と型

- 数値と非数値ではありかが違う

- 数値

- 単位という別の属性が必要

1, “cm”

- 非数値

- 単位はマーク値に含まれる

“1 cm”

- 「前置単位」と「後置単位」

- 表示

”¥1,000”

- 非数値型と数値型の間の変換

1000, “¥”, 前置

- 「形式的な単位」と「束縛的な単位」

- 形式的な単位: 人間の理解を助けるための単位で, その変更は値に影響しない

- 束縛的な単位: その変更が値の変更を強制する. “cm” -> “m”

- 設計: 限られた単位から選ばせ, 変更させない or そのたびに値を書き換える
マッピングで対応する

型	単位
• 計量	束縛的
• 計数	形式的
• 序数	形式的
• 時間	束縛的
• 日時	形式的*
• 記録度数	なし

*表示形式は単位に左右されるがそれだけでは定まらない. 単位から定まる選択肢のなかから選ぶ必要がある.