Day1
データの集合の形

技術部データ基盤チーム 財津大夏 / GMO PEPABO inc.

2022.07.12 データエンジニアリング研修 基礎編 Day1

GMOペパボ

技術部 データ基盤チーム データサイエンティスト

財津 大夏　Hiroka Zaitsu

2012年 入社

- ペパボのデータ基盤「Bigfoot」*1の開発/運用
- Bigfoot を使ったデータ分析/活用
- Twitter : @HirokaZaitsu

#データ基盤 #DataOps #MLOps #Python
#SQL #統計学 #機械学習 #スバル #Fallout

*1: GMOペパボのサービスと研究開発を支えるデータ基盤の裏側
https://speakerdeck.com/zaimy/inside-story-of-data-infrastructure-supporting-gmo-pepabos-services-and-r-and-d

カリキュラム目標と概要

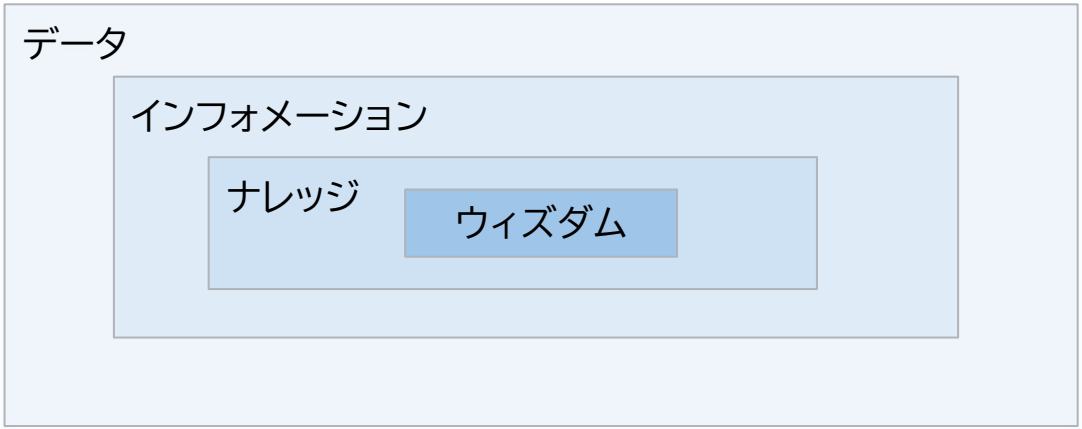
- **Day1: 扱いやすいデータの集合の形を理解できる**
  - データを構造化するための知識の導入
- **Day2: 初歩的な SQL を使ってデータベースからデータを参照できる**
  - データを参照するために必要な基礎的な知識の導入
- **Day3: 複数テーブルのデータを組み合わせて参照できる**
  - リレーショナルデータベースからデータを参照するための知識の導入
- **Day4: データを要約・可視化して情報や知識を取り出すことができる**
  - データを実際の施策や判断に利用するために必要な知識の導入

➡ 各日のハンズオンを通して手を動かしながら知識の解釈を高める

データとは？

データとは

- データ ⊃ インフォメーション
  - データは「インフォメーションの原材料」
  - インフォメーションは「コンテキストを持ったデータ」



DAMA International（2018）「データマネジメント知識体系ガイド 第二版」日経BP社より引用

```
date-id-name-num0-num1-num2-num3
2021-07-15-0123456789-test0-12-34-567-890
2021-07-15-9876543210-test1-98-76-543-210
```

```
date-id-name-num0-num1-num2-num3
2021-07-15-0123456789-test0-12-34-567-890
2021-07-15-9876543210-test1-98-76-543-210
```

```
date,id,name,num0,num1,num2,num3
2021-07-15,0123456789,test0,12,34,567,890
2021-07-15,9876543210,test1,98,76,543,210
```

データからインフォメーションを抽出する手段

- 手段1: インフォメーションになりうるデータを作る
- 手段2: 既存のデータをインフォメーションに変換する

手段1: インフォメーションになりうるデータを作る

- そもそもデータを集めるのはなぜか？
  - 何かを知りたいから
  - 適切な理解は適切なデータ作成から
- データを作るにはナレッジ(ドメイン知識)が必要
  - 目的に沿ったデータを作ることが大事
  - システムだけでなく人もデータを作る
    - スプレッドシート
    - Notion etc…

➡ 扱いやすいデータの形を知っておく必要がある

手段2: 既存のデータをインフォメーションに変換する

- 誰もが分かる表現に加工する
  - 例）0～6 で曜日を表現している
- 別のデータとつなぎ合わせる
  - 例）ユーザー情報 * 注文情報
- メタデータ（データを説明するデータ）を追加する
  - 例）5W1H

➡ データを扱う方法を知っておく必要がある

表形式のデータの集合の構造化

行と列(1)

- 表形式のデータの集合は、行と列から構成される
  - 横方向が行
  - 縦方向が列

	列 サイズ	身丈 (cm)	身幅 (cm)	肩幅 (cm)	袖丈 (cm)
行	S	65	49	42	19
	M	69	52	46	20
	L	73	55	50	22

	日付	曜日	勤務条件	休暇申請内容		警告	就業時間	打刻訂正理由	休憩	実働時間
☐	6/1	水	FLX				05時47分 ～ 17時22分		1:00	8:13
☐	6/2	木	FLX				09時43分 ～ 18時03分		1:00	7:20
☐	6/3	金	FLX				06時16分 ～ 18時39分		1:00	9:24

行と列(2)

- 同じデータの集合を構造化する方法は数多くある
  - 例)人に対して処置 a, b を施した際の結果を表す Table 1, 2
- 行と列は表の外観を表す言葉なので
  構造化する方法によって表すものが変わる
  - 例)Table 1 の行は処置, Table 2 の行は人

➡ 構造化する方法によって意味が変わる表を
機械的に読み取ることは困難

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

画像は Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10 より引用

データの意味に着目する

- データの集合は「値」を集めたもの
- どの値も 1 つの「変数」と 1 つの「観測」に属する
  - 変数とは、色々な値を取りうるデータの入れ物
  - 観測とは、事象を観察し測定すること
- 例）3 つの変数と値
  - 「人」: 取りうる値は John Smith, Jane Doe, Mary Johnson
  - 「処置」: 取りうる値は a, b
  - 「結果」: 取りうる値は -, 1, 2, 3, 11, 16

➡ 意味と構造（行と列）を一致させることができないか？

GMOペパボ

tidy data

- 次のように整理すると意味と構造を一致させられる
  - 1 つの観測が 1 つの行に対応する
  - 1 つの変数が 1 つの列に対応する
  - 1 つの観測の単位の類型が 1 つの表に対応する

➡ 統計学の「tidy*1 data」

➡ データベースの「正規形」「テーブル」

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

画像は Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10 より引用

*1: 「tidy」は「きれいな」「整然とした」という意味

tidy data と messy data*1 で観測を検索する

- 例)Jane Doe の処置 a の結果を参照する
  - tidy data
    - 各変数を列方向に探して観測を特定できる
      - 異なる変数を参照するには列を変更すれば済む
      - 変数(列)の数にも制限がない
  - messy data
    - 行と列をそれぞれ探す必要がある
      - 異なる変数を参照するには行列の変更が必要
      - 行と列の意味は変わりうる

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

画像は Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10 より引用
*1: tidy data でないデータの集合。「messy」は「雑然とした」「汚い」という意味

データとビューの分離

- 「tidy data が便利なのは分かったけど、人間が読みづらいんですが…？」
- 人間が読みやすいビュー(集計表)は tidy data から作成する
  - 例)ピボットテーブル
- 集計表から tidy data には戻せないのでデータとビューは分離する

地域	商品名	商品数	商品単価
東日本	A	14	¥1,000
東日本	B	15	¥1,100
西日本	A	11	¥1,000
西日本	B	21	¥900
東日本	C	16	¥800
西日本	C	18	¥1,200
東日本	D	11	¥900
東日本	A	10	¥900
東日本	B	9	¥1,300
西日本	A	12	¥1,000
西日本	B	15	¥1,000

地域	*商品名*	SUM of 商品数	AVERAGE of 商品単価
⊟ 東日本	A	24	¥950
	B	24	¥1,200
	C	16	¥800
	D	11	¥900
東日本 Total		75	¥1,000
⊟ 西日本	A	23	¥1,000
	B	36	¥950
	C	18	¥1,200
西日本 Total		77	¥1,020
Grand Total		**152**	**¥1,009**

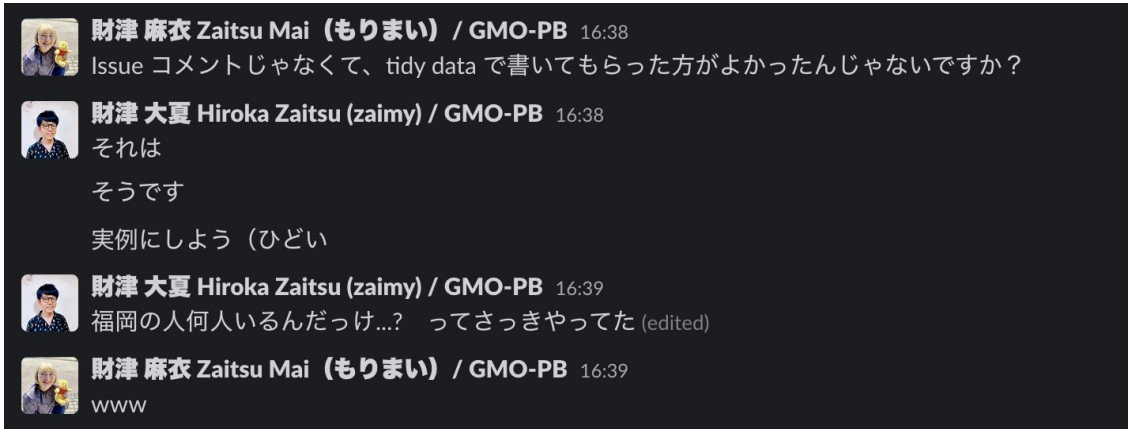
画像は Google Workspace ラーニングセンター https://support.google.com/a/users/answer/9308944?hl=ja よりピボットテーブルの例を改変

失敗談


データエンジニアリング研修基礎編の参加者のみなさんが[redacted]データセットにアクセスできるようにする #1009
Edit
Merged
zaimy merged 3 commits into master from data-engineering-training on Jul 12
Conversation 38
Commits 3
Checks 9
Files changed 1
+35 −1
zaimy commented on Jun 28 • edited
レビューして欲しい人
データ基盤チーム: @bigfoot/data-team
情報セキュリティ規定における責任者・管理者: @hsbt, @tnmt
任意のレビュアー: 技術部のデータセットなので @hsbt
何を解決したいのか
データエンジニアリング研修基礎編の参加者のみなさんが tech データセットにアクセスできるようにします。
レビューポイント
新卒のみなさんは別途 #1008 で追加します。
権限追加・変更の詳細
ユーザーアカウント
Reviewers
hsbt
y-suzuki
kozee
tnmt
Assignees
No one—assign yourself
Labels
needs review
Projects
None yet
Milestone

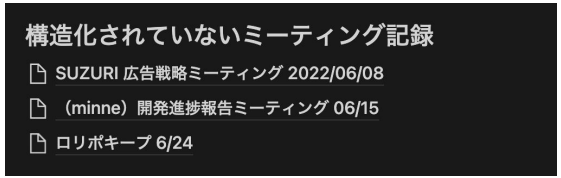
失敗談

- 「Slack じゃなくて情報がストックされる P/R コメントで書いてもらえばいいかな～」
  - 後からコメントを見て人数を数え上げる必要がある
  - 「福岡の人は何人いるんだっけ...」
  - BigQuery の権限の申請に使うためにお名前とご所属とメールアドレスをひたすらコピペ



他の例を見てみる

- 整然データとは何か｜Colorless Green Ideas をみんなで見てみましょう
  - https://id.fnshr.info/2017/01/09/tidy-data-intro/
- minne のデータベースのテーブルをみんなで見てみましょう
- Notion のドキュメントの構造化
  - データベース機能を適切に活用する
    - 変数ごとに Sort や Filter できる



まとめ

- データからインフォメーションを抽出するために構造化する
- データの集合は tidy data にして構造化する
  - 1 つの観測が 1 つの行に対応する
  - 1 つの変数が 1 つの列に対応する
  - 1 つの観測の単位の類型が 1 つの表に対応する
- 意味と構造を一致させることで機械的に扱いやすくなる
- データとビューは分離する

明日の準備

- BigQuery の権限が付いているか確認します

ここには画面キャプチャがありました
公開資料からは削除しています