SASによる
データマネジメント入門

集積用データから解析用データへ

㈱スーザック　横堀　真

1．データとは何か

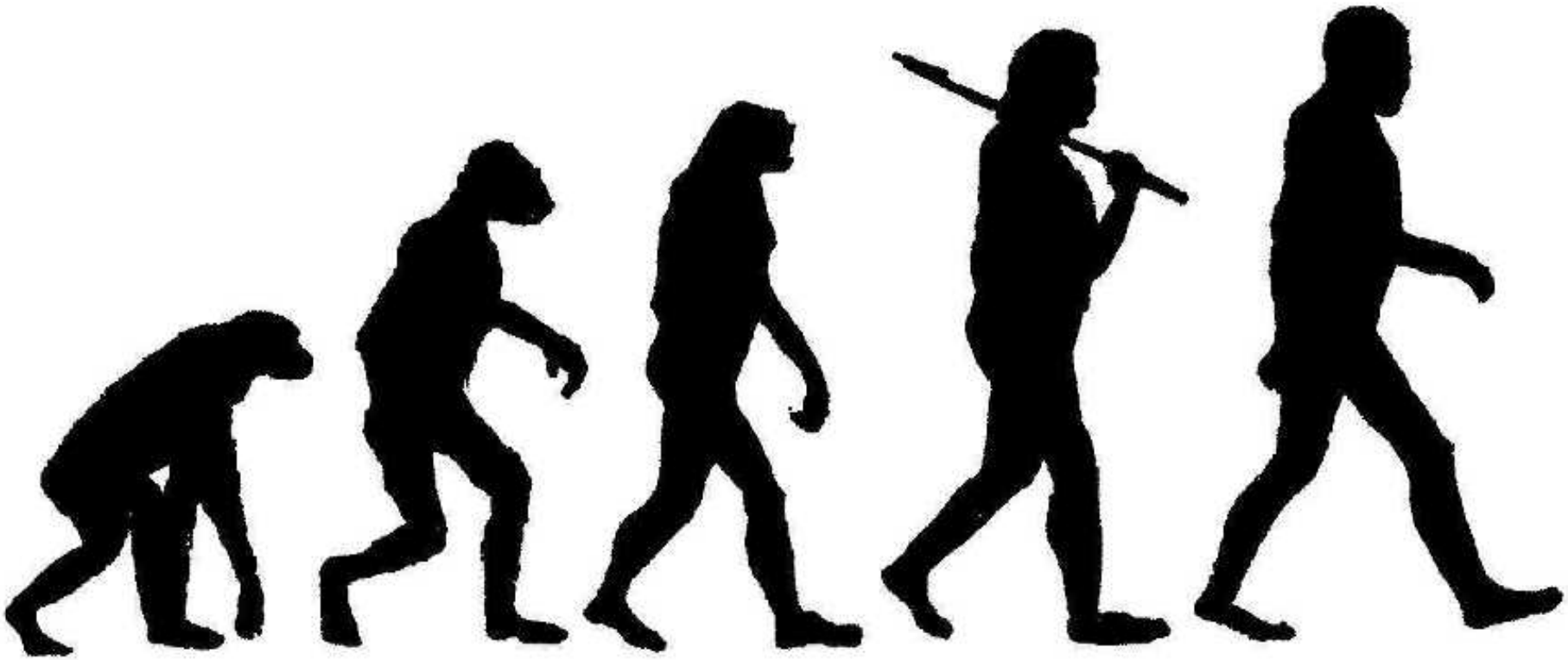
分析・解析する目的は？

散在する情報を分析して
知見を得たい！

なぜデータが必要か？

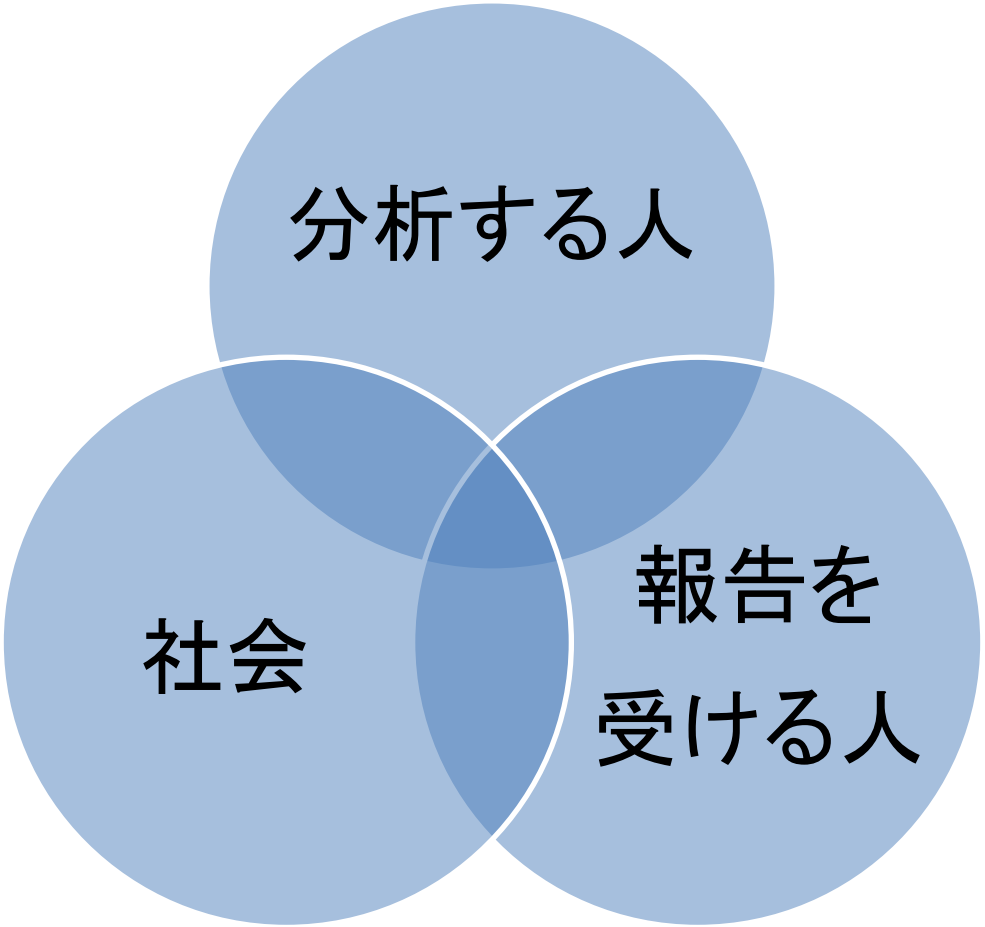
- 何かを分析しようとするとき、
- 蓄積されたデータが必要になる
- でも、そのデータは使い物になるのか・・・？
- きちんと整備された形で保管されていれば良いのですが・・・
- 中身は信頼できるのか・・・

情報は進化する



Fact **Data** **Information** **Knowledge** **Intelligence**

Intelligenceは誰のため？



データが整備されていないと

Fact

Guess!

2. データベースの基礎知識

データの正規化

なぜ正規化？

正規化とは？

臨床検査一覧表の例

伝票コード	採血日	患者コード	患者名	検査コード	検査名	基準値上限	検査値
1101	11/06/01	01	高橋	101	AST	45	20
				102	ALT	40	15
1102	11/06/02	02	山田	103	ALP	250	180
1103	11/06/03	03	鈴木	104	GTP	60	23
1104	11/06/04	01	高橋	101	AST	45	14
1105	11/06/05	03	鈴木	103	ALP	250	127
				104	GTP	60	42

第一正規化①

伝票コード	採血日	患者コード	患者名
1101	11/06/01	01	高橋
1102	11/06/02	02	山田
1103	11/06/03	03	鈴木
1104	11/06/04	01	高橋
1105	11/06/05	03	鈴木

第一正規化②

伝票コード	検査コード	検査名	基準値上限	検査値
1101	101	AST	45	20
1101	102	ALT	40	15
1102	103	ALP	250	180
1103	104	GTP	60	23
1104	101	AST	45	14
1105	103	ALP	250	127
1105	104	GTP	60	42

LDHの基準値上限が入力できない。

ASTの基準値上限が50に変更された。

第二正規化①

検査コード	検査名	基準値上限
101	AST	45
102	ALT	40
103	ALP	250
104	GTP	60

第二正規化②

伝票コード	検査コード	検査値
1101	101	20
1101	102	15
1102	103	180
1103	104	23
1104	101	14
1105	103	127
1105	104	42

第一正規化①

伝票コード	採血日	患者コード	患者名
1101	11/06/01	01	高橋
1102	11/06/02	02	山田
1103	11/06/03	03	鈴木
1104	11/06/04	01	高橋
1105	11/06/05	03	鈴木

まだ検査を受けていない田中さんの患者コードが入力できない。

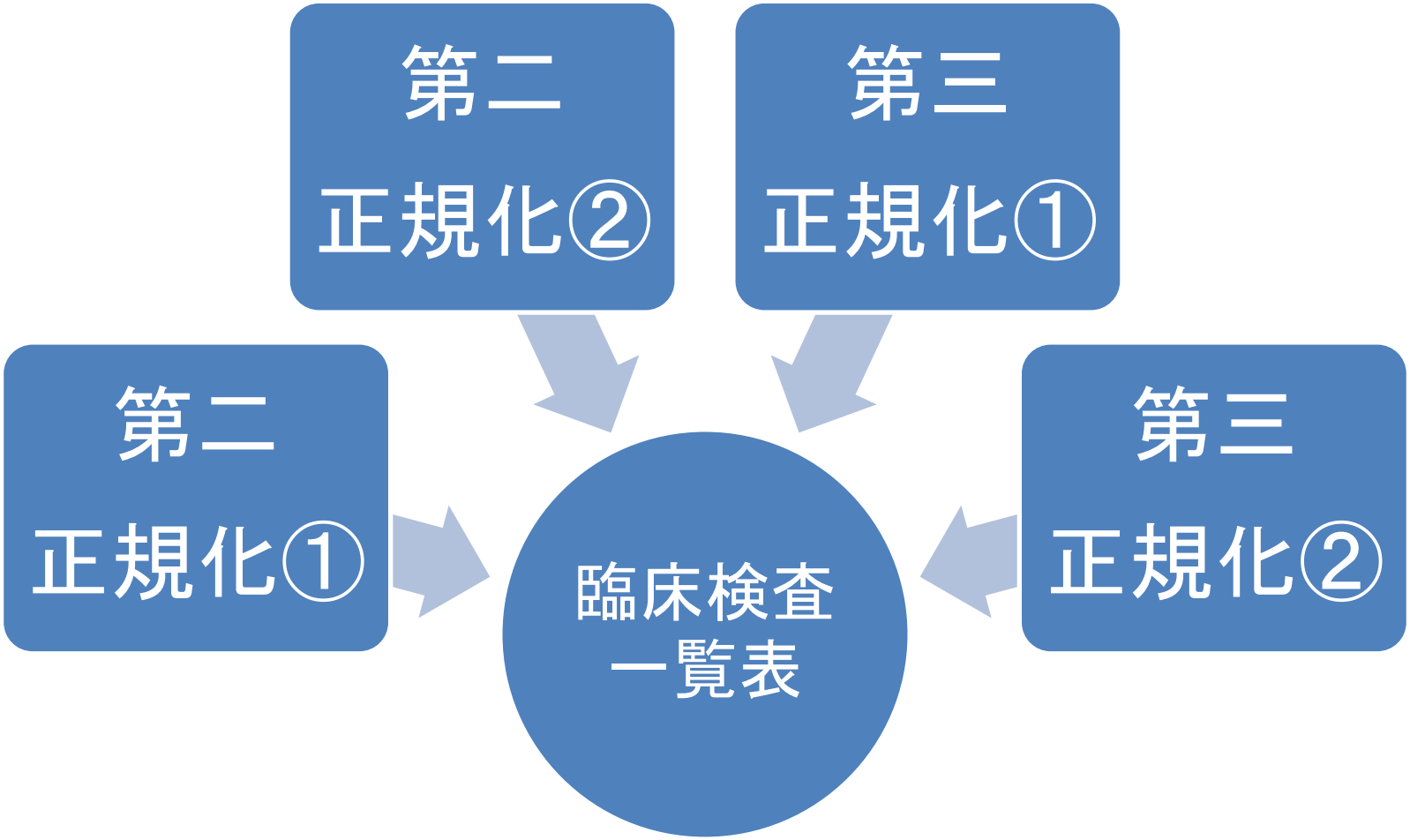
第三正規化①

伝票コード	採血日
1101	11/06/01
1102	11/06/02
1103	11/06/03
1104	11/06/04
1105	11/06/05

第三正規化②

患者コード	患者名
01	高橋
02	山田
03	鈴木

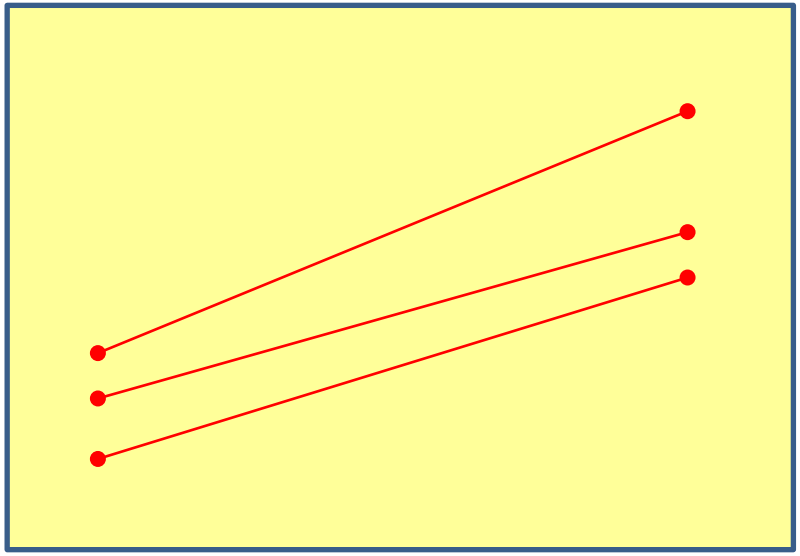
リレーショナルデータベース

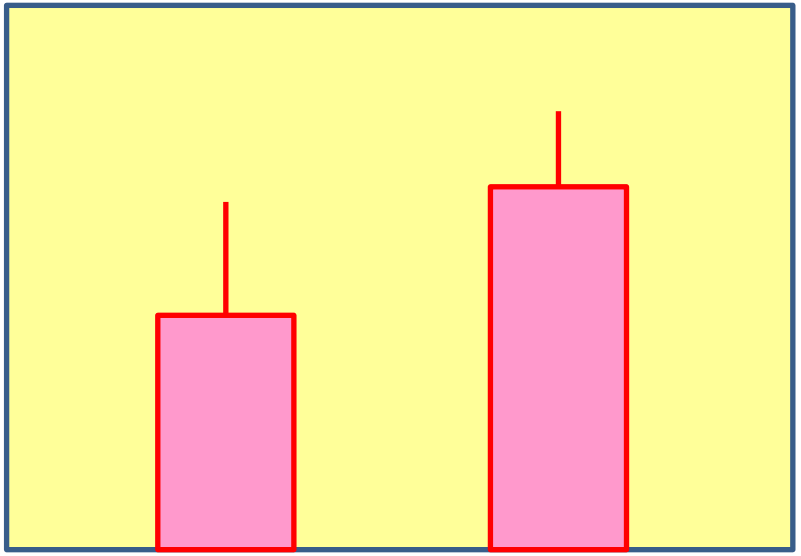


集積用データから解析用データへ

解析用のデータセット

患者コード	患者名	検査コード	検査名	飲酒前	飲酒後
01	高橋	101	AST	20	37
02	山田	101	AST	31	43
03	鈴木	101	AST	26	41





集積用データ(データベース)

SUBJ	VISIT	AST
01	1	20
01	2	37
02	1	31
02	2	43
03	1	26
03	2	41

縦横変換 : DATA STEP①

```
data AST_1;
  set AST;
  by SUBJ VISIT;
  array AST_V{2};
  retain AST_V1 - AST_V2;
  if first.SUBJ then call missing(of AST_V1 - AST_V2);
  AST_V{VISIT} = AST;
  if last.SUBJ then output;
  keep SUBJ AST_V1 - AST_V2;
run;
```

縦横変換：DATA STEP②

SUBJ	AST_V1	AST_V2
01	20	37
02	31	43
03	26	41

縦横変換：PROC TRANSPOSE①

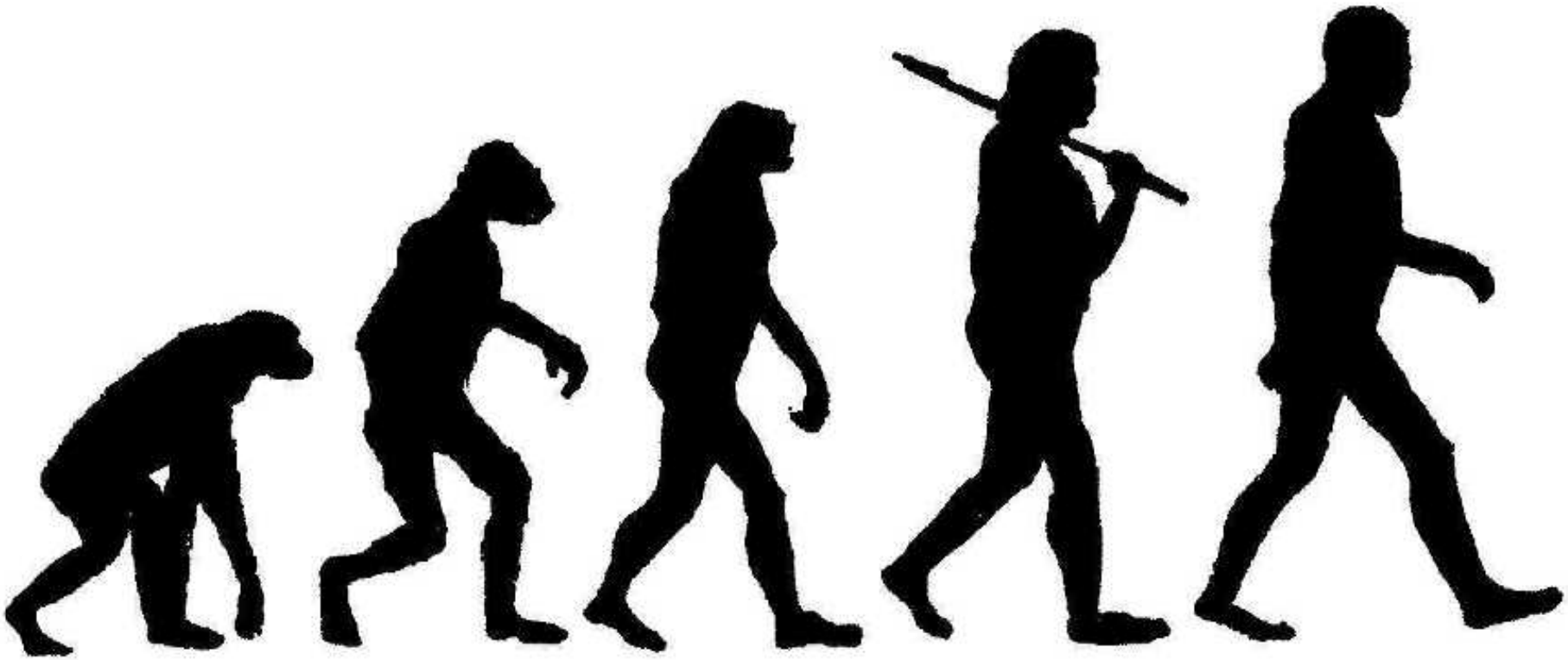
```
proc transpose data=AST prefix=AST_V
                        out=AST_2(drop=_NAME_);
  by   SUBJ;
  id    VISIT;
  var  AST;
run;
```

縦横変換：PROC TRANSPOSE ②

SUBJ	AST_V1	AST_V2
01	20	37
02	31	43
03	26	41

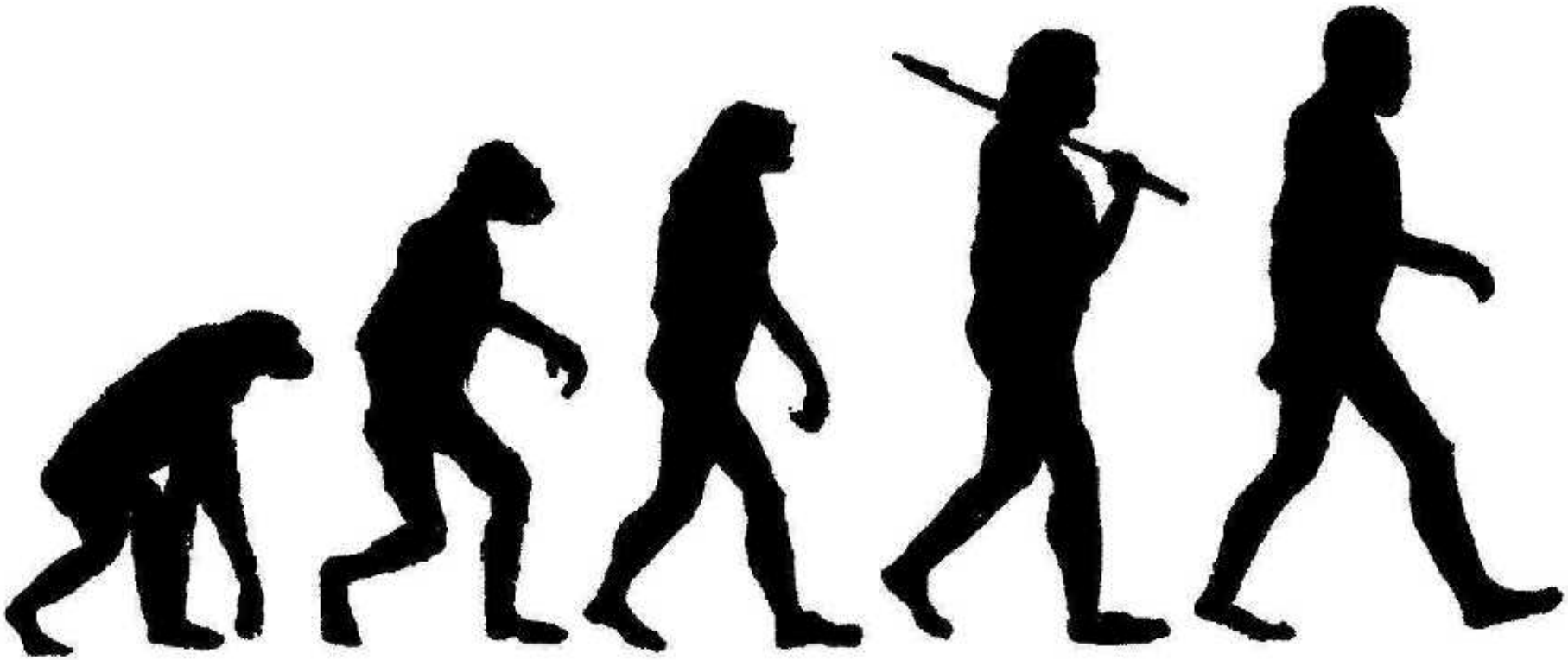
３．データクリーニング

Factは誰にも変えられない



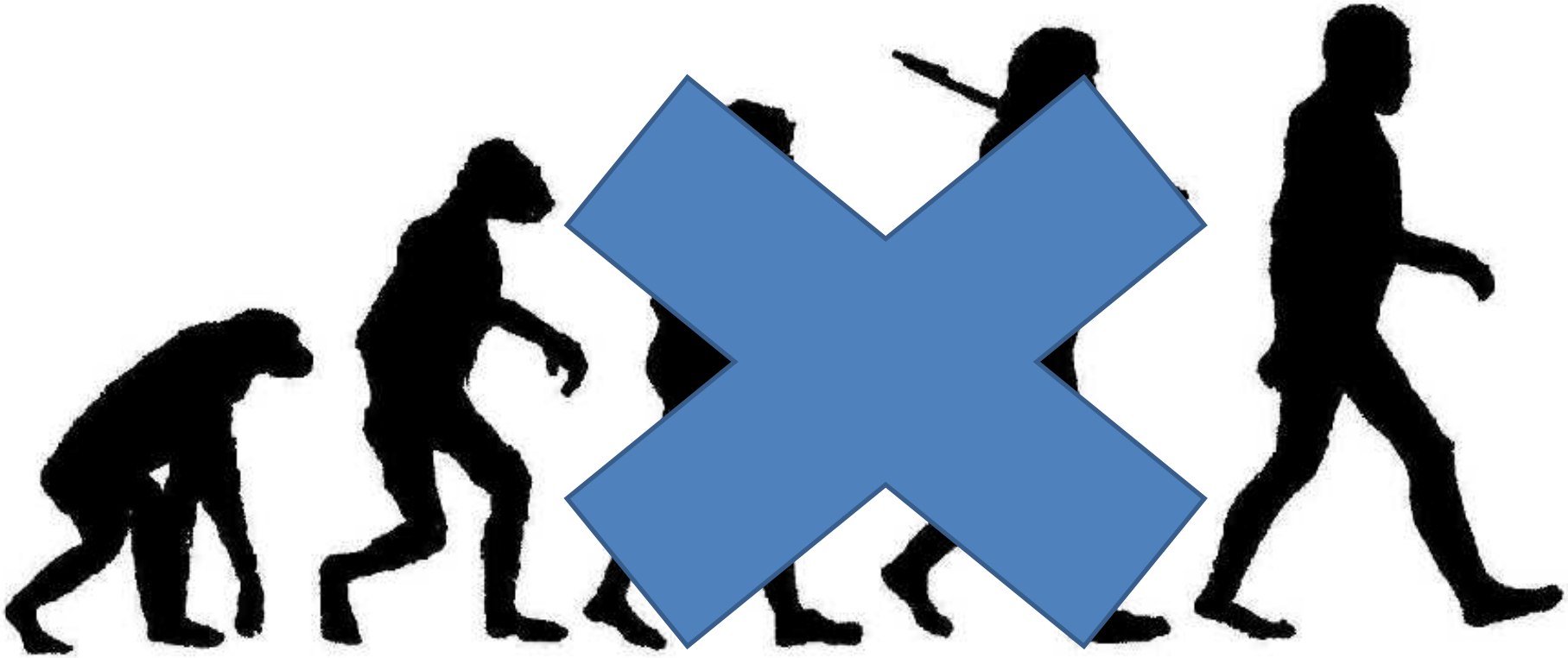
Fact **Data** **Information** **Knowledge** **Intelligence**

Dataは変えられる



Fact **Data** **Information** **Knowledge** **Intelligence**

Dataが変えられていると・・・



Fact **Data** **Error!**

Digression

なぜ間違えるのか？

- いろいろな間違い
  - 転記・入力ミス
  - 検査・測定ミス
  - 定義の誤解釈

- 結果を変えてしまう力が大きいものは？

Digression

間違いの例

- ある解熱剤の試験にて・・・
  - 転記・入力ミス (Case 1)
    - 体温「36.5度」を「35.6度」と記録してしまった。
  - 検査・測定ミス (Case 2)
    - 体温を測定するのを忘れてしまった。
  - 定義の誤解釈 (Case 3)
    - 服薬「30分後」の体温を「30秒後」と勘違いしていた。

Digression

「転記・入力ミス」の結果
(Case 1)

被験者	服薬前	服薬後
Aさん	38.5	36.9
Bさん	39.2	35.6
Cさん	38.9	37.1

Digression

「検査・測定ミス」の結果
(Case 2)

被験者	服薬前	服薬後
Aさん	38.5	36.9
Bさん	39.2	-
Cさん	38.9	37.1

Digression

「定義の誤解釈」の結果
(Case 3)

被験者	服薬前	服薬後
Aさん	38.5	38.2
Bさん	39.2	39.3
Cさん	38.9	38.5

Digression

集計してしまうと・・・

```
title “Pre vs. Post";
proc means data=case1 MEAN CLM;
 var pre post;
run;
```

Digression

何が起こっていますか？

Case 1

MEANS プロシジャ

変数	平均	平均の下側 95% 信頼限界	平均の上側 95% 信頼限界
pre	38.8666667	37.9942662	39.7390672
post	36.5333333	34.5101205	38.5565462

Case 2

変数	平均	平均の下側 95% 信頼限界	平均の上側 95% 信頼限界
pre	38.8666667	37.9942662	39.7390672
post	37.0000000	35.7293795	38.2706205

Case 3

変数	平均	平均の下側 95% 信頼限界	平均の上側 95% 信頼限界
pre	38.8666667	37.9942662	39.7390672
post	38.6666667	37.2541262	40.0792072

間違った結果を避けたい

- 結果を“ねじ曲げる”パワーがあるものには注意が必要。
  - 日本人の平均貯蓄額＝1,400万円　？？？

Outlier

文字変数：PROC FREQ①

```
title "Frequency of SEX variables";
proc freq data=test;
 tables SEX / nocum nopercent;
run;
```

文字変数：PROC FREQ②

Frequency of SEX variables

FREQ プロシジャ

SEX	度数
1	1
F	1
M	2

欠損値の度数 = 2

数値変数：PROC UNIVARIATE①

```
title "Using PROC UNIVARIATE to look for outliers";
proc univariate data=test plot;
 id SUBJ;
 var WEIGHT;
run;
```

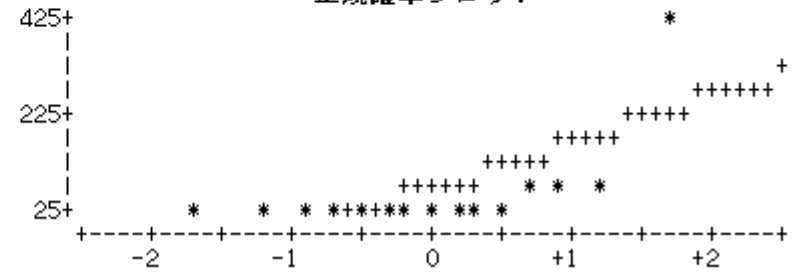
数値変数：PROC UNIVARIATE②

極値

---------最小値---------			---------最大値---------		
値	SUBJ	Obs	値	SUBJ	Obs
32	06	6	48	08	8
36	13	13	50	04	4
40	09	9	60	07	7
40	02	2	74	12	12
42	05	5	410	11	11

幹	葉	#	箱ひげ図
4	1	1	*
3			
3			
2			
2			
1			
1			
0	55555567	8	+--0--+
0	344444	6	+-----+

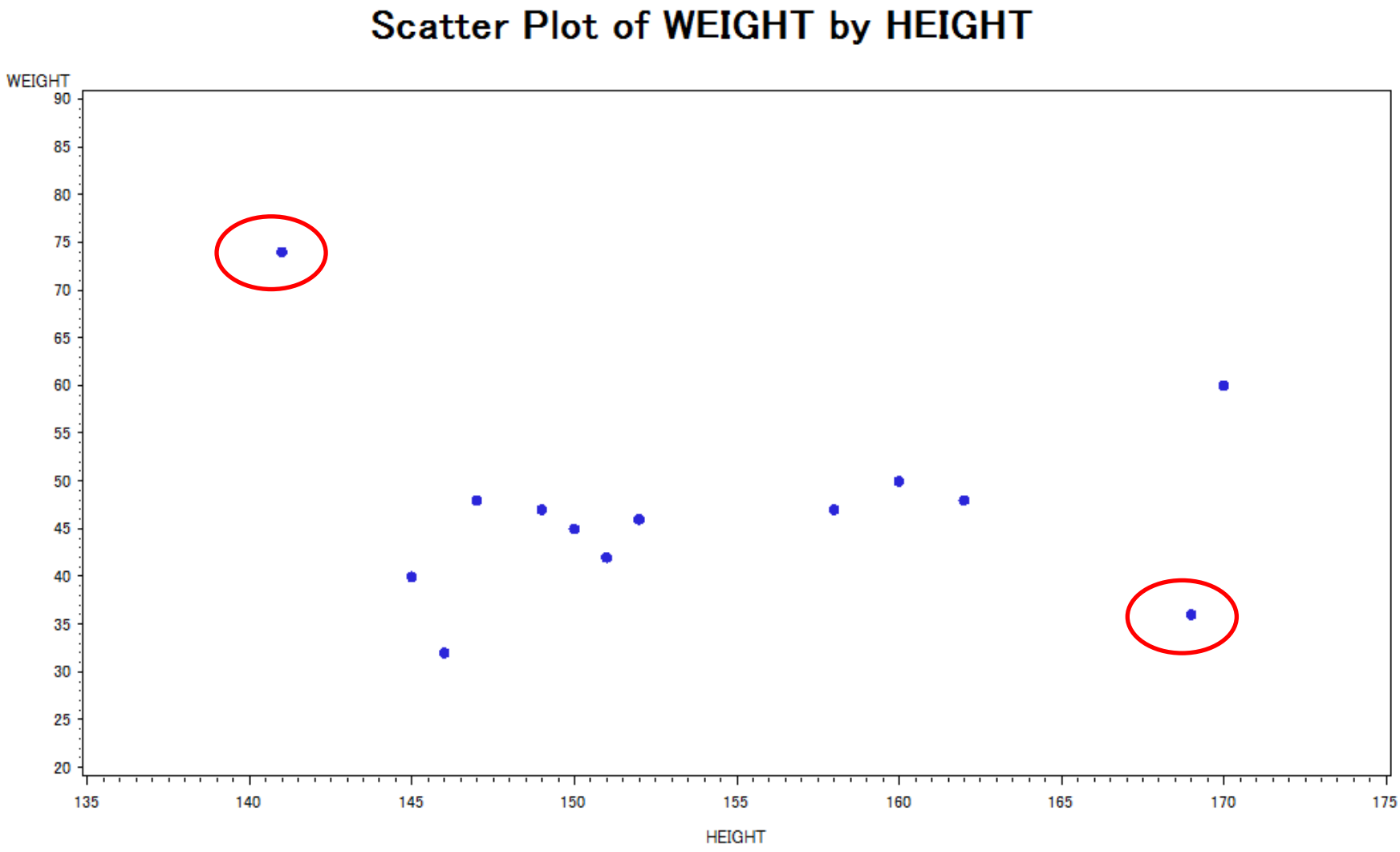
幹.葉の単位 : 10**+2



数値変数：Scatter Plot①

```
title "Scatter Plot of WEIGHT by HEIGHT";
symbol value=dot;
proc gplot data=test;
  plot WEIGHT * HEIGHT /  vaxis=20 to 90 by 5
                          haxis=135 to 175 by 5;
run;
```

数値変数：Scatter Plot②



最後に

- 『データマネジメント』の重要性
  - 事実はそこに存在する
  - 欲しい情報は得られているか？
- 『目的』や『定義』の重要性
  - その情報は妥当か？

- 以上のことを十分に踏まえたうえで・・・
  - SASは強力なデータマネジメントのツールとなり得ます

THANK YOU