



QuintilesIMS™

PhUSE Annual Conference, Edinburgh 2017
Paper DH08

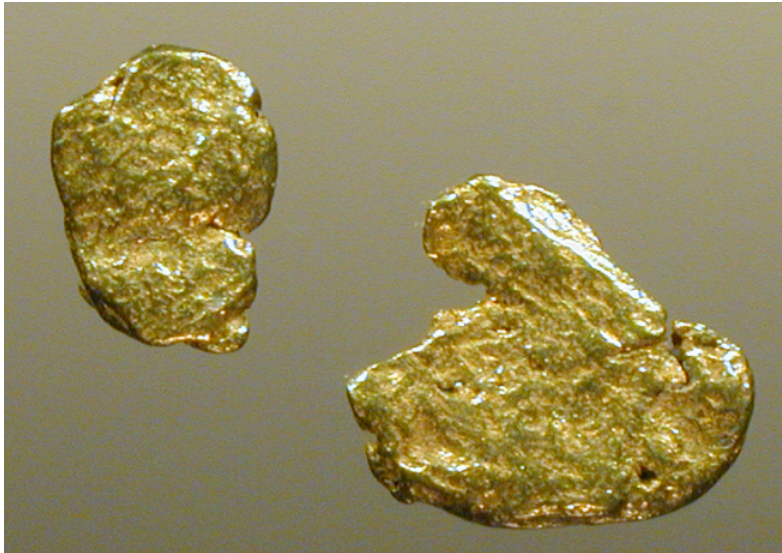
Data Cleaning Technique for Business Analytics
What Stats Can Do for DM?

Adam Korczyński
10th October 2017

Meaning of the data in business analytics

Gold from a Neolithic site vs. Bitcoin

Reality



Source: <http://ancientafricah.wikispaces.com/Ghana+Enviroment>

I receive my revenue from goldsmith on a monthly basis but it comes in irregular shapes, has missing parts... I have questions:

- How to estimate the funds I have?
- Can I make proper financial decisions?

Expectation
to work
towards ideal
world



Ideal world



Source: <https://www.forbes.com/sites/naeemaslam/2017/08/29/is-it-time-to-own-gold-or-bitcoin/#758dc5887f5d>

I know what I want and to get closer to the ideal world I need to inform the goldsmith on what needs to be done prior to the transfer. That will save our time and effort and will make us both happier!

Working with data in practice – an overview

On the cooperation between DM and Stats

Real Data

Issues with characters:

Patient?took the IO

Not standardized entries:

Haematology
Hematology

Issues with dates:

2017-10-01
2017-09
2916

Issues with numeric results:

Age: 180 years
SYSBP: 12 millimetres of mercury

Expectation
to work
towards clean
data



Ideal Data

Remove unexpected characters,
correct spelling:

Patient took the IP

Use one standard:

Hematology
Hematology

Query and fill-in dates:

2017-10-01
2017-09-01
2016-01-01

Query and amend numeric data:

Age: 180 years
SYSBP: 120 millimetres of mercury

Data management (DM)

- (i) as clean as possible...sufficient?
- (ii) a broader perspective needed

Analyst (Stats)

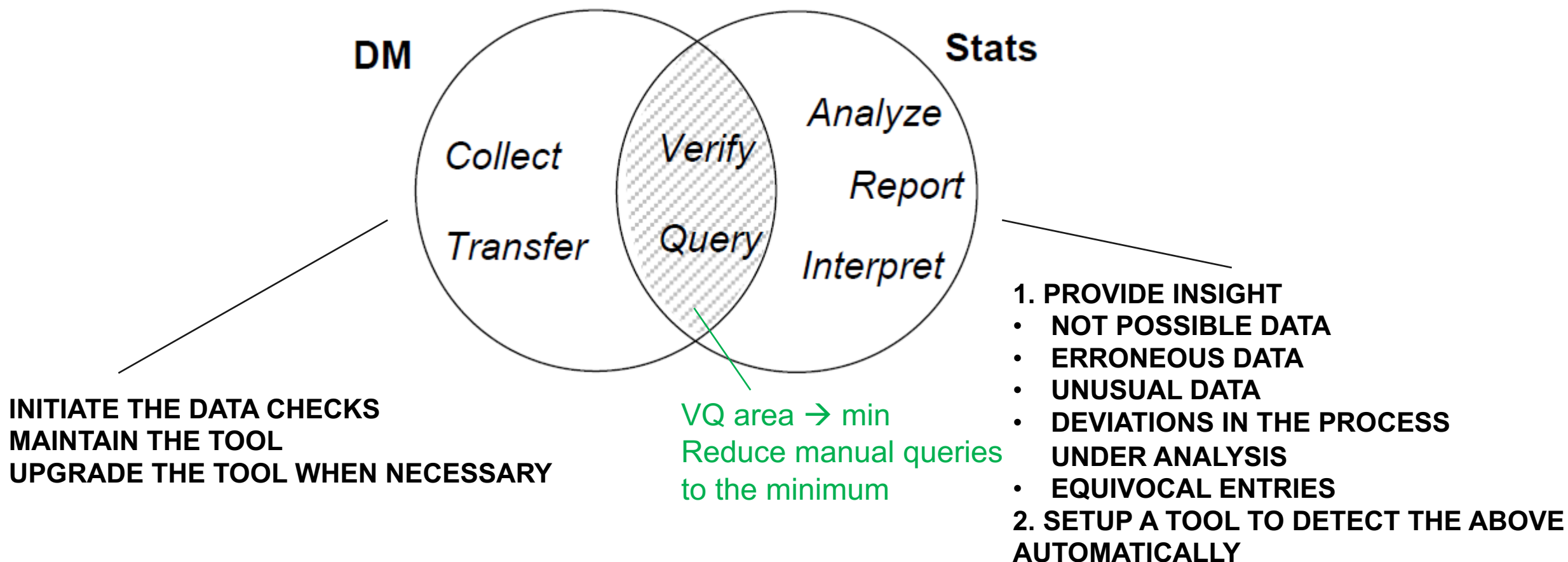
statistical requirements:

- (i) what do we need?
- (ii) what we should avoid or amend?

any chance to work together smoother and faster?

DM and Stats – roles, responsibilities and synergy

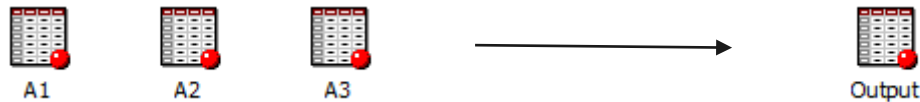
How to make our lives easier



Data cleaning scheme for structured databases

All in one

- Multiple datasets → where are the issues?
- Let us create a standardized dataset with all entries from the database:



What do we need to do practically?

1. Transpose a single dataset to get a new row for each individual value.
2. Make a use of metadata to locate the issues across datasets. The values are kept along with descriptive characteristics enabling us to identify each entry.

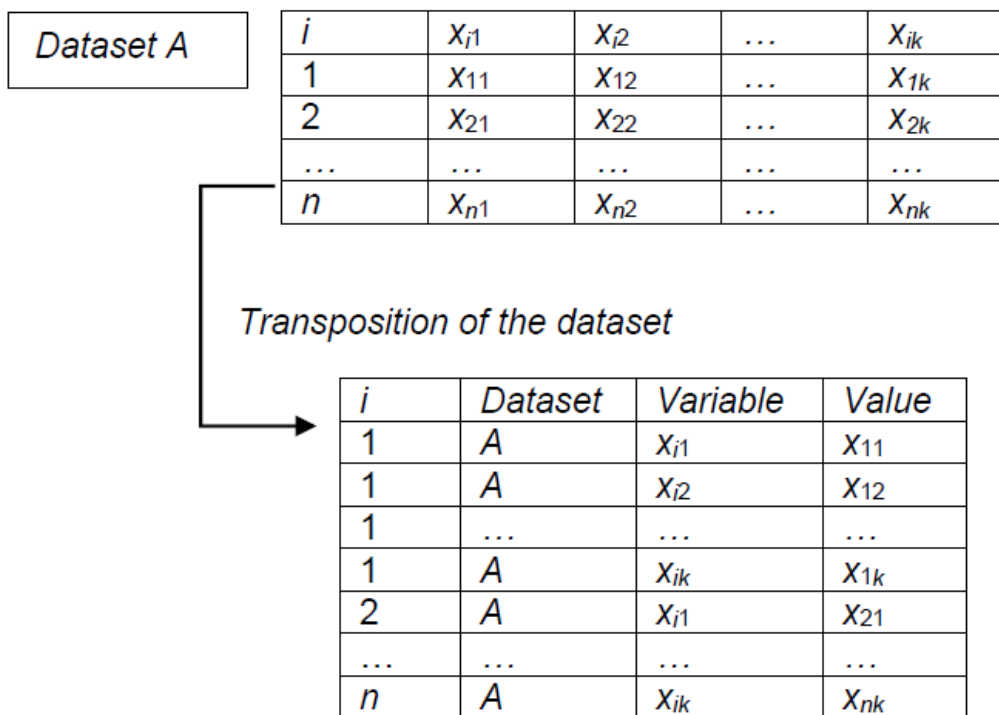
The key message here is that no prior knowledge on the content is needed by using metadata!

Data cleaning preliminary steps - transposition of a dataset



Technically speaking...

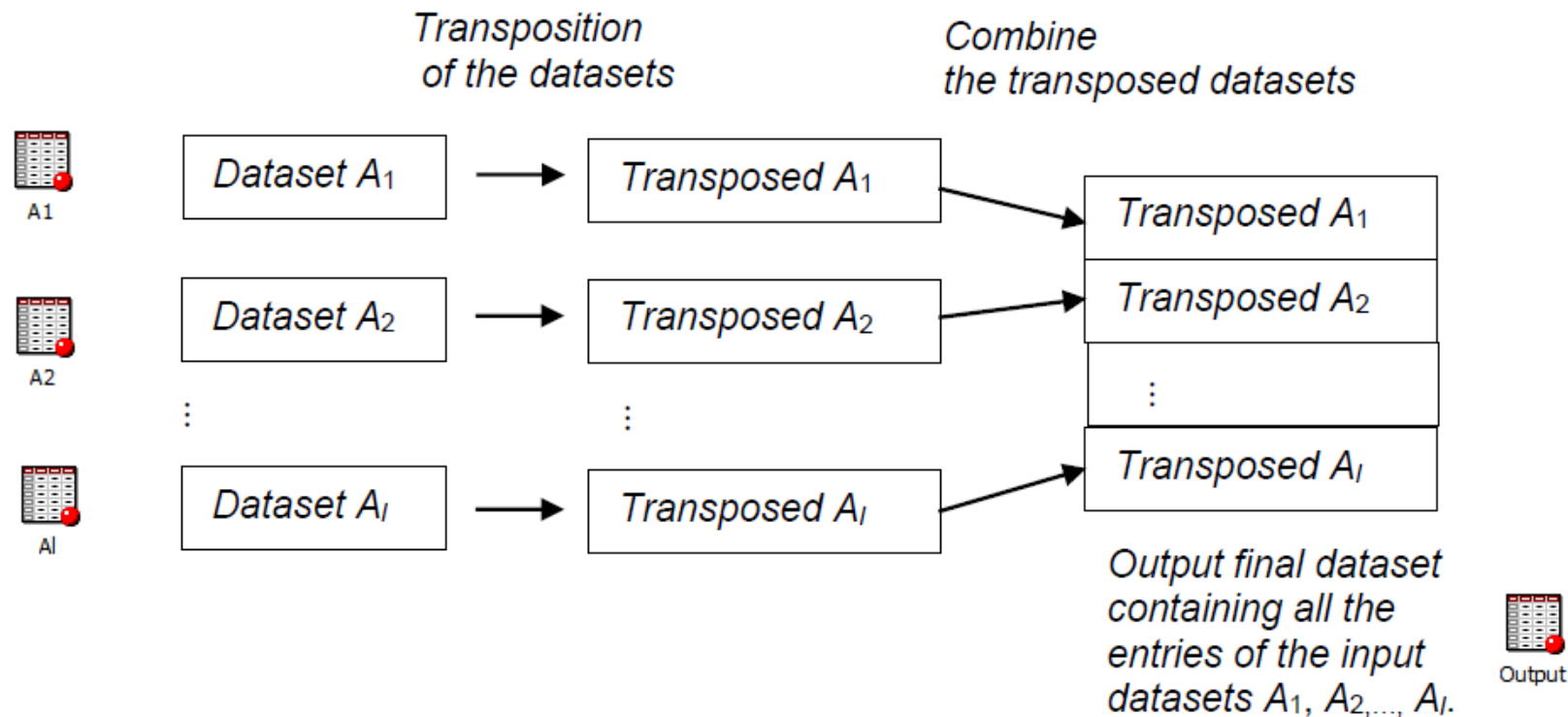
- Let us assume that in dataset A we have identification variable i - where $i = 1, 2, \dots, n$ - and k explanatory variables $x_{i1}, x_{i2}, \dots, x_{ik}$. The transposition of the dataset A would mean to create a column based on each row of values for variables $x_1, x_2, \dots, x_k$ for each unique i :



Data cleaning preliminary steps – combining the transposed datasets

Further on the programmatic perspective

- Let us further assume that we have I datasets $A_1, A_2, \dots, A_I$. The steps of the data cleaning technique are respectively to transpose each dataset in a directory and combine these datasets afterwards to get one final dataset including all entries from the database



Example of a transposed dataset – AEA dataset from NIDA-CPU-0016 (2014)

What does it look like?

VIEWTABLE: Work.Ae

	study	patient	a_v	a_event	FORM_NO	CENTER	VST_DATE	RAT_PER
1	999A	290	2	INSOMNIA	16	324	7	1
2	999A	290	2	HEADACHE	16	324	16	2
3	999A	290	2	HEADACHE	16	324	21	3
4	999A	290	2	HEADACHE	16	324	28	4
5	999A	290	2	HEADACHE	16	324	35	5
6	999A	290	2	HEADACHE	16	324	43	6
7	999A	290	1	DIARRHEA	16	324	48	7
8	999A	290	2	DIARRHEA	16	324	55	8
9	999A	290	2	DIARRHEA	16	324	64	9
10	999A	290	.		16	324	69	10
11	999A	290	.		16	324	76	11

patient	study	line	memname	libname	ParamCd	ParamLb	ParamPt	ParamTp	ParamVIC	ParamVCD	ParamVI	ParamVO
290	999A	2	AEA	RAW	A_V	V. Outcome - a	BEST12.	N	2	2	2	2
290	999A	2	AEA	RAW	A_EVENT	Not on CRF (text code) - a	\$19.	C	HEADACHE	HEADACHE	1	.

ID variables

Location
within
database

Variable
description

Actual values (numeric vs
character, formatting)

Nature of data as a source of data issues typology

Data scales and typical data issues

- Nominal data:

Woman or Man

- Ordinal scale:

Primary education, secondary education or tertiary education

- Interval data:

2017-10-10, 10°C

- Ratio scale:

83.9 kg, 30301 GBP

- The traditional scales → main data types: characters, numbers and a special type of number, which is date and time. What are the data issues to expect then?

Data issues to consider

Examples

ABC

- hidden or mistaken characters.
- 'Neg' 'neg' and 'negative'

123

- how plausible a given result is?
- 3-sigma rule and quantiles rule: $|x| > 3\sigma$ or $[Q1 - 1.5IQR, Q3 + 1.5IQR]$, where $IQR = Q3 - Q1$
- Use a goodness-of-fit test to compare empirical and theoretical distribution

Let us see how does it work in practice...

Hidden characters

```
%macro FindHidChars (InpSet= /*Input dataset*/  
                    ,Var=/*Variable to check*/  
                    ,Whr=/*Optional filter*/  
                    ,OutSet= /*Output dataset*/);  
  
data &OutSet. (where=(sum > 0));  
  set &InpSet.;  
  array _tmpsc {*} _tmpsc1-_tmpsc31;  
  do i=1 to dim(_tmpsc);  
    _tmpsc{i} = index(&Var.,byte(i));  
  end;  
  sum=sum(of _tmpsc1-_tmpsc31);  
  %if %length(&whr.)>0 %then where &Whr.;;  
  
run;  
%mend FindHidChars;
```

Symbols 1-31 are not necessarily welcome...

```
%FindHidChars (InpSet=o  
  ,Var=paramv1c  
  ,Whr=  
  ,OutSet=o1);
```

3-sigma in action – prepare the dataset

*Select variables representing systolic blood pressure;

```
data a01;  
  set o;  
  std=ParamV1;  
  where index(paramcd, 'SYS')>0;  
run;
```

	line	memname	libname	ParamCd	ParamLb	ParamVIO
1	1	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	120
2	2	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	100
3	3	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	100

*Standardize the values;

```
proc standard data=a01 out=a02 mean=0 std=1;  
  var std;  
run;
```

	line	memname	libname	ParamCd	ParamLb	ParamVIO	std
1	1	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	120	0.1150785265
2	2	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	100	-1.13400569
3	3	PHYSEXA	RAW	Q4_SYS	Blood pressure - sitting	100	-1.13400569

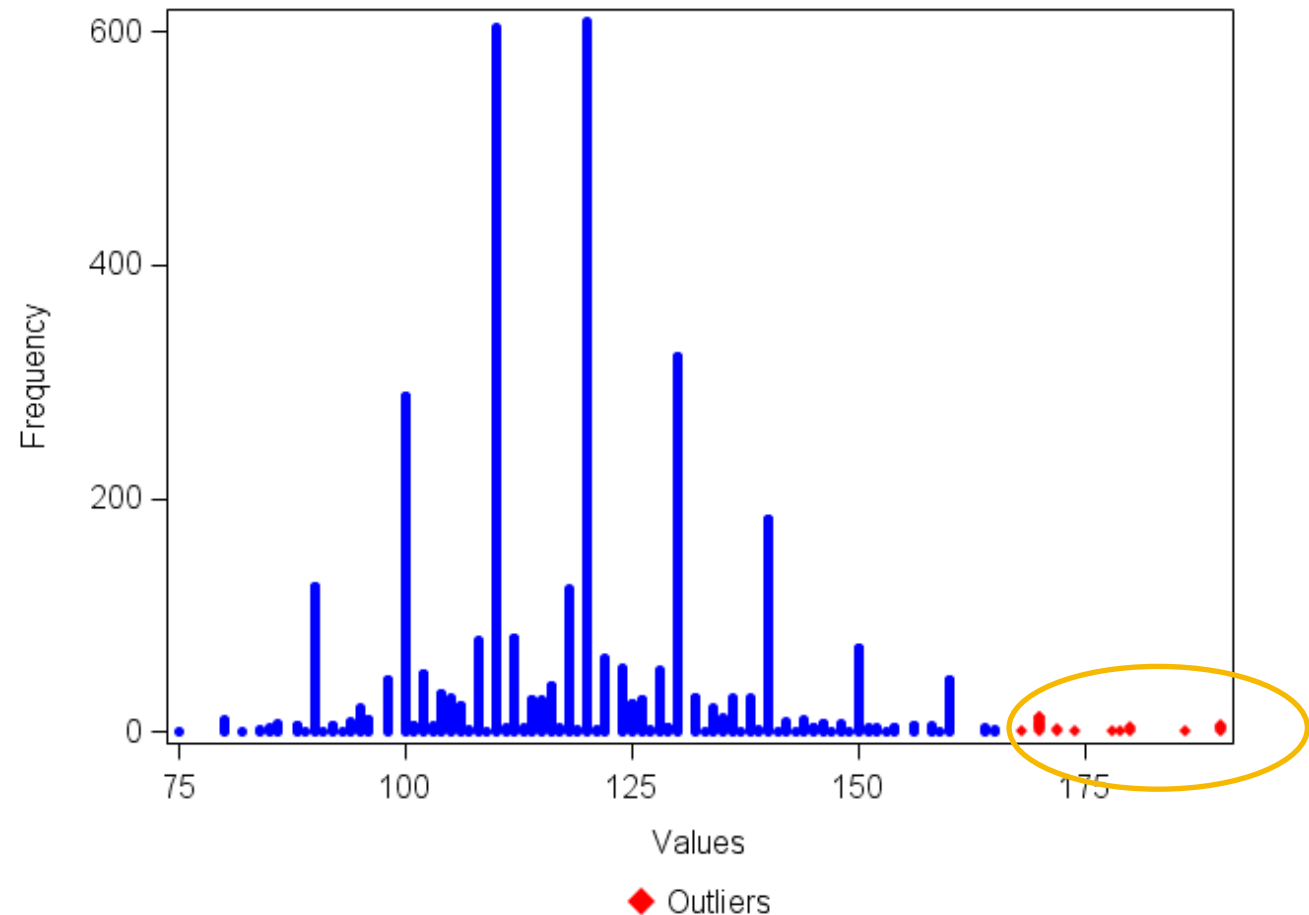
*Calculate the frequencies;

```
proc freq data=a02 noprint;  
  tables ParamV1*std / out=a03;  
run;
```

	ParamV1	std	COUNT	PERCENT
1	75	-2.695360961	1	0.0297530497
2	80	-2.383089907	10	0.2975304969
3	82	-2.258181485	1	0.0297530497

3-sigma in action – draw the histogram and identify outliers

```
proc sgplot data=a04;  
  scatter x=Val      y=i / markerattrs=(symbol=CircleFilled size=5 color=blue);  
  scatter x=OutVal   y=i / markerattrs=(symbol=DiamondFilled size=6 color=red);  
run;
```



Further on the check for potentially fraudulent numeric data

Statistics to help out!

Benford's Law - distribution of first digit

- According to the Benford's Law, first digit (1,2,...,9) follow a specified distribution (Judge & Schechter, 2009, p. 2):

$$P(D = d) = \log_{10} \left(1 + \frac{1}{d} \right), \quad d = 1, 2, \dots, 9 \quad (1)$$

- Any unusual observations given (1)?

$H_0: F(x) = F_0(x)$, versus the alternative that:

$H_1: F(x) \neq F_0(x)$.

- The test statistic compares the observed counts (n_i) with the counts under the null hypothesis ($n \cdot p_i$) in each of the m categories ($i = 1, 2, \dots, m$):

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i} \quad (2)$$

- where n is the sample size, n_i is the empirical count in the i th category and p_i represents the probabilities of the theoretical distribution. Statistic (2) follows χ^2 distribution with $m-1$ degrees of freedom, where m represents the number of categories of the variable (in this case $m=9$). The rejection region is defined as $P(\chi^2 \geq \chi^2_{\alpha, m-1}) = \alpha$, while the p -value is $P(\chi^2 \geq \chi_e^2) = P$, where χ_e^2 represents the value of the test statistic (2).
- The difference between the theoretical counts of first digit and the empirical counts can be inspected graphically.

Data issues to consider

Examples continued

10th October 2017

- Partial dates and times
 - Fraudulent events
 - IPDs
-
- Missing data. Not collected data vs. non-response.



- Duplicate records

Data cleaning in practice

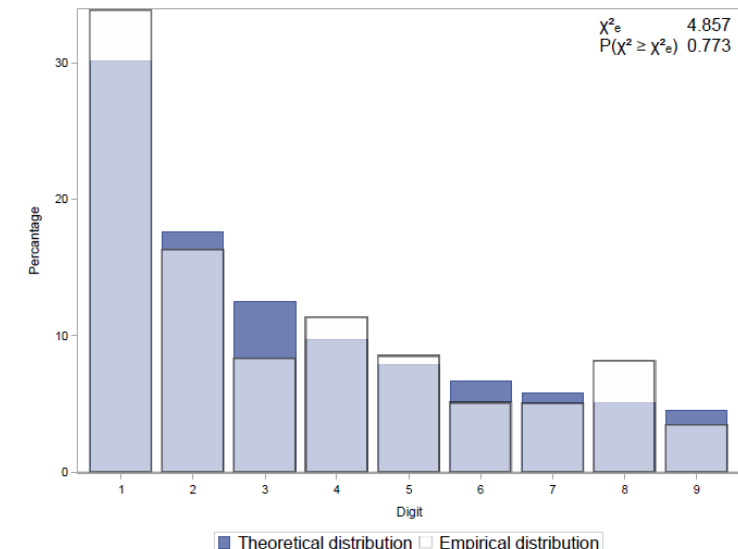
An application for clinical trial data

- The technique applied to the data from clinical trial conducted by the National Institute on Drug Abuse (NIDA). The input database contained 33 datasets with 215,053 observations and 2,306 variables in total. The input datasets have been transposed and set with the use of the *TransposeAllLib* macro program as shown and explained in the paper, Appendix A.
- The transposition of the datasets on a standard local computer took 14 minutes and 15 seconds. The output dataset composed of 2,555,301 records, has been used to apply two example checks on the data: check for hidden characters and check for first digit distribution for the numeric values from the laboratory dataset LABSA.

Results:

1. There were no hidden characters found in any of the datasets.
2. No indication of statistically significant deviation of the empirical counts from the theoretical counts under Benford's Law (see figure).

GOOD POINT FOR THE DATA CLEANING ON THE STUDY!



Concluding notes

- Successful analytics requires good quality data. Be realistic and proactive
- Discuss what can be done between DM and the analyst
- The implementation of a data cleaning technique enables us to reduce the amount of resource needed for both data manager and analyst. Data manager would need to spend less time on the queries and the analyst will have access to the cleaner data, which will boost up the analysis process!
- Are there any areas in this context that should be investigated more?
- At least two:
 - Further insight into the types of data checks;
 - Efficient methods for large database applications.

Going back to the initial example it is good to remember that data is turned into money... but only if it's reliable.



THANK YOU FOR YOUR ATTENTION!

ANY QUESTIONS ?

CONTACT INFORMATION:

Adam Korczyński

QuintilesIMS

ul. Gottlieba Daimlera 1

02-460 Warszawa

Work Phone: +48 22 430 31 57

Email: Adam.Korczynski@QuintilesIMS.com