WHY CHOOSE BETWEEN SAS DATASTEP & PROC SQL WHEN YOU CAN HAVE BOTH

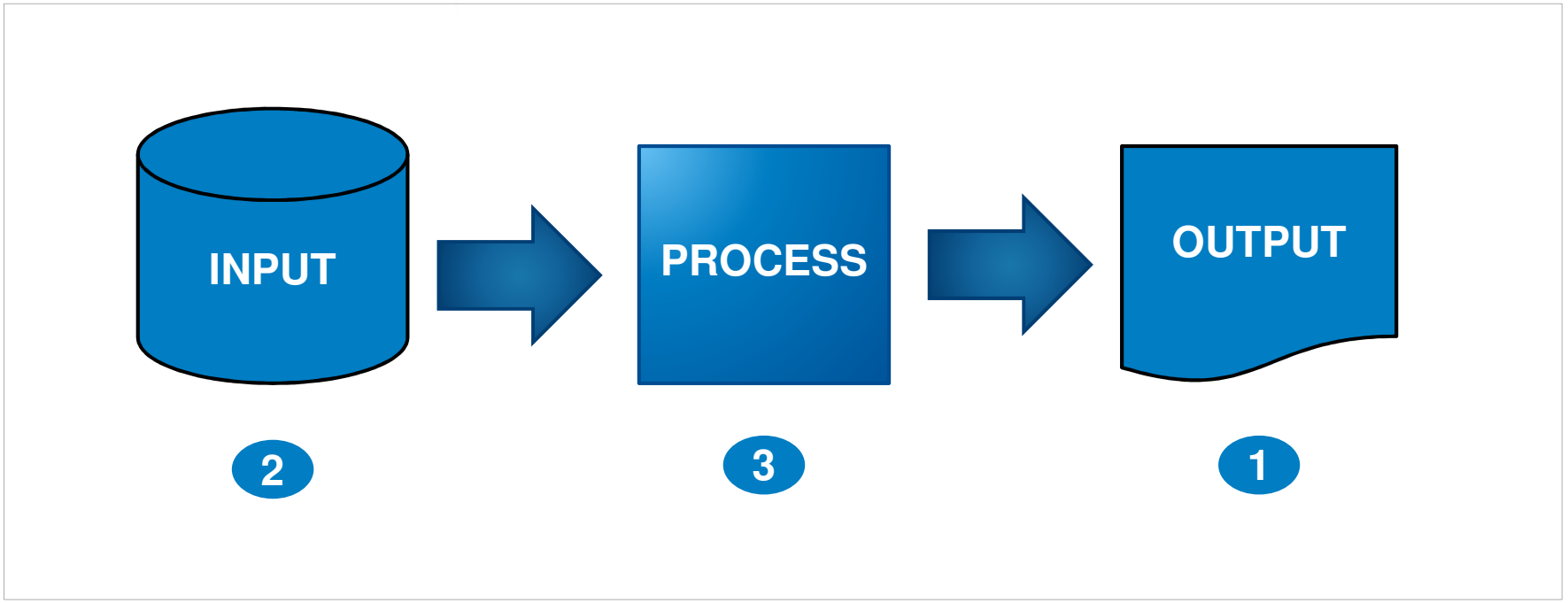
Wisconsin Illinois SAS Users Conference
Milwaukee

Charu Shankar
Technical Training Specialist
SAS institute

AGENDA

1. 3 primary questions
2. **INPUT** Data Access
3. **PROCESS**
   - 3.1 Data Joining
   - 3.2 Data Analysis
   - 3.3 Data Querying
   - 3.4 Data Validation
   - 3.5 Data Cleansing
   - 3.6 Data Manipulation
   - 3.7 Data Management– Satisfying Programmer's #1 rule
4. **OUTPUT** Data Presentation
5. In Sum
6. Q&A
7. Resources

1. THE WORLD OF DATA, 1, 2, 3

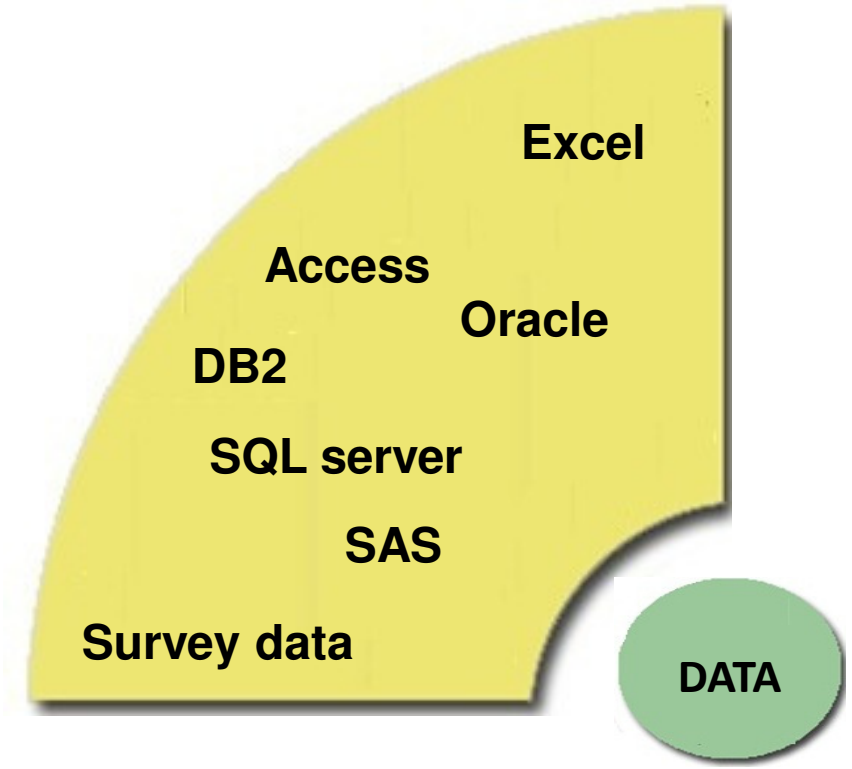


- Data Access

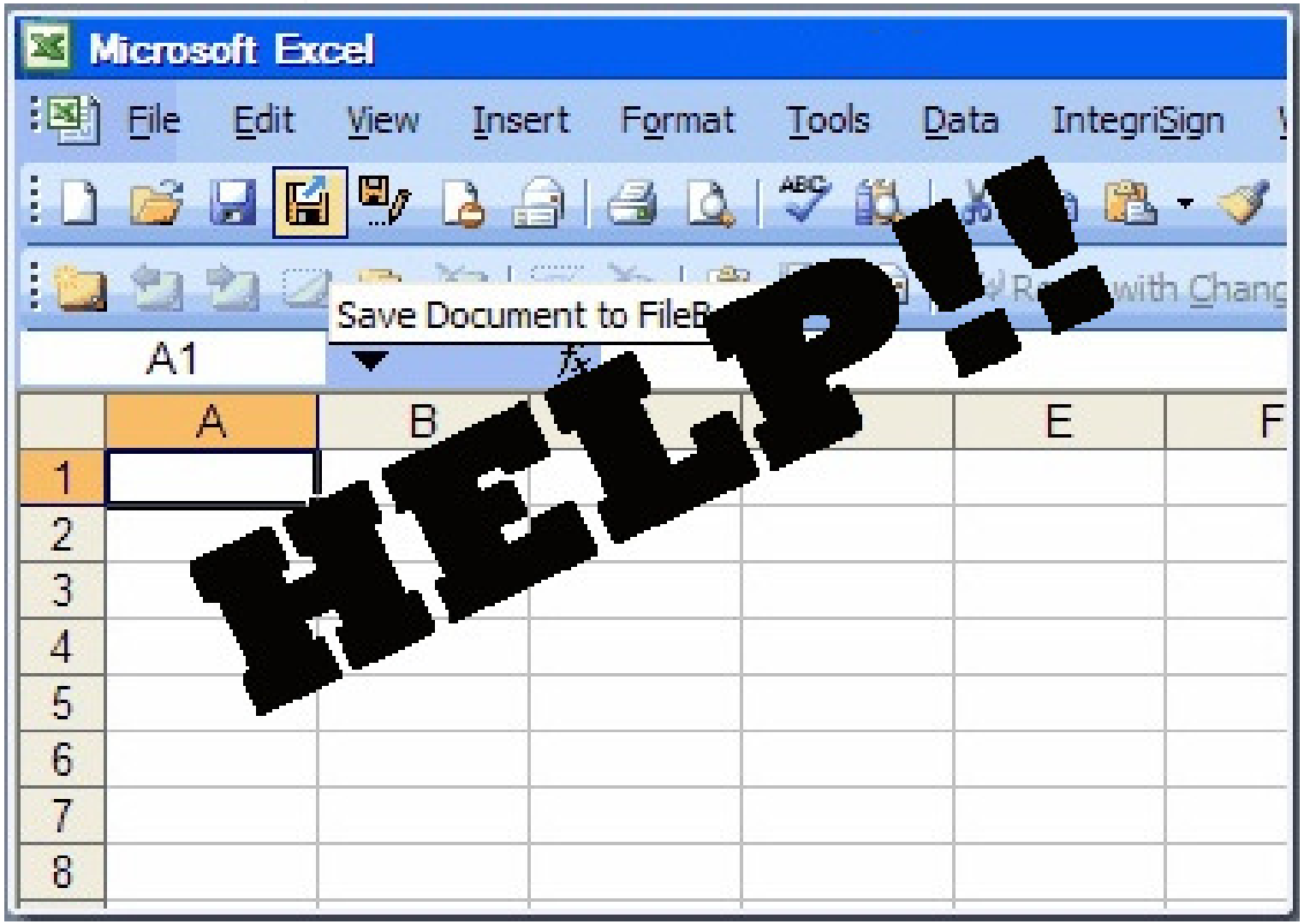
- Data Joining
- Data Analysis
- Data Querying
- Data Validation
- Data Cleansing
- Data Manipulation
- Data Management

- Data Reporting

2. INPUT - DATA ACCESS



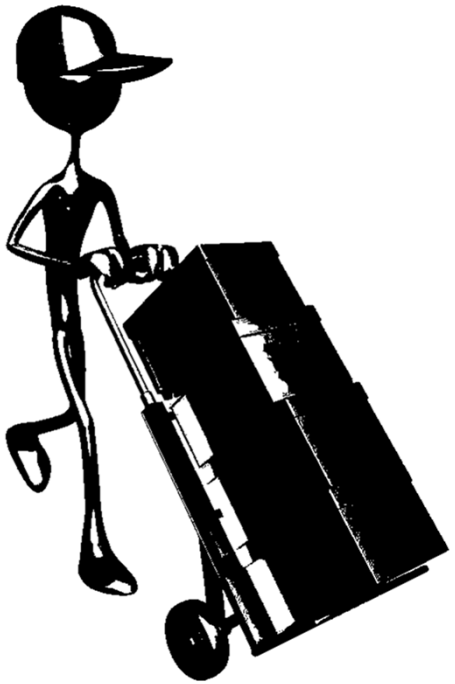
2.1 INPUT - DATA ACCESS - EXCEL



DEMO

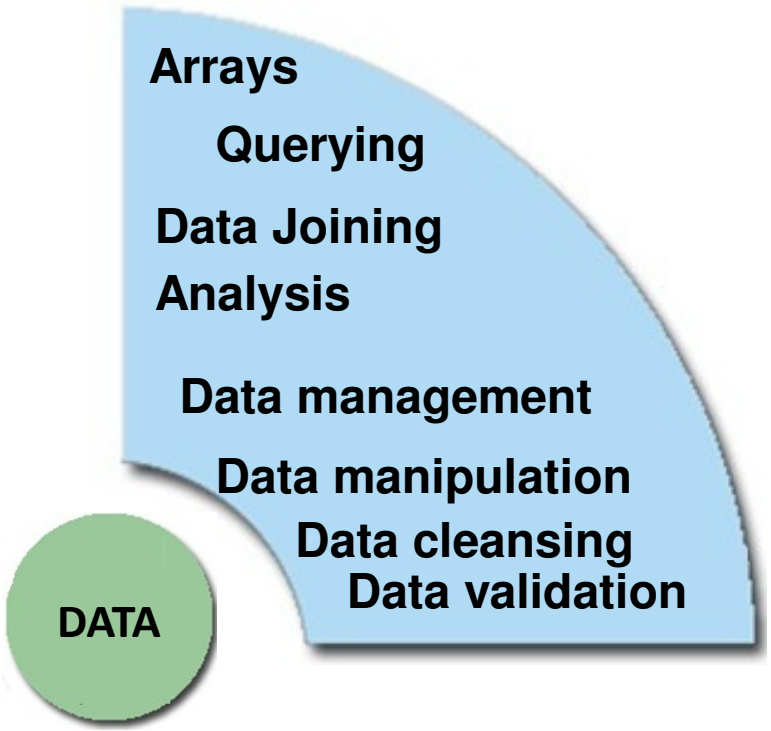
2.2 INPUT DATA ACCESS – RAW DATA

Can you read raw data with PROC SQL?



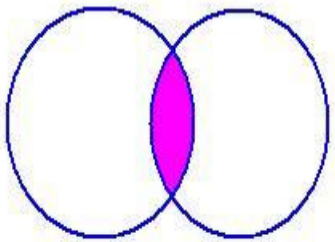
DEMO

3. PROCESS

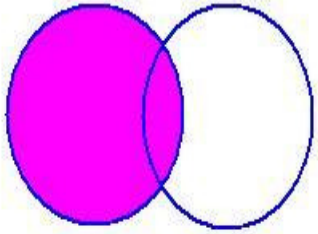


3.1. DATA JOINING

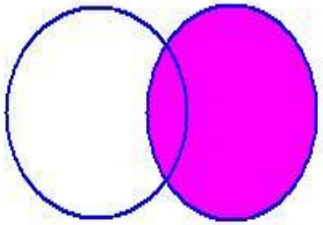
JOINS AND SET OPERATIONS IN RELATIONAL DATABASES



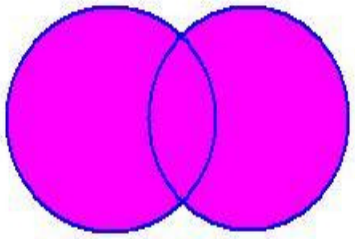
Inner join (result similar to Intersect)



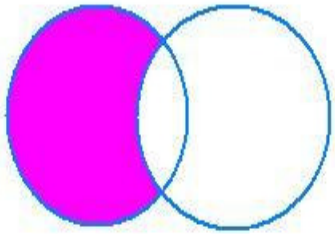
Left outer join



Right outer join



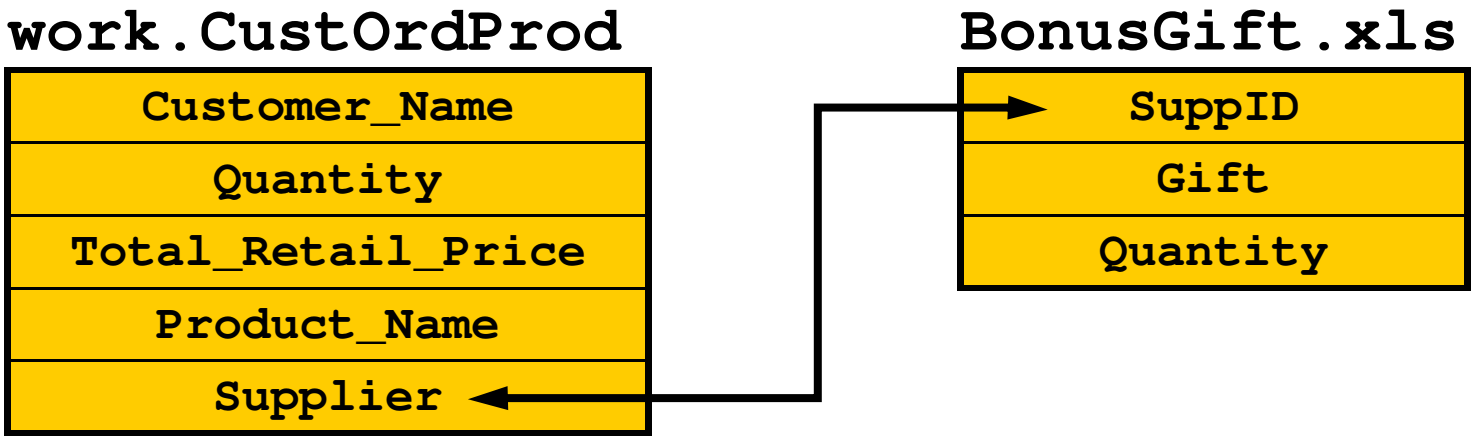
Full outer join



Minus

BUSINESS SCENARIO

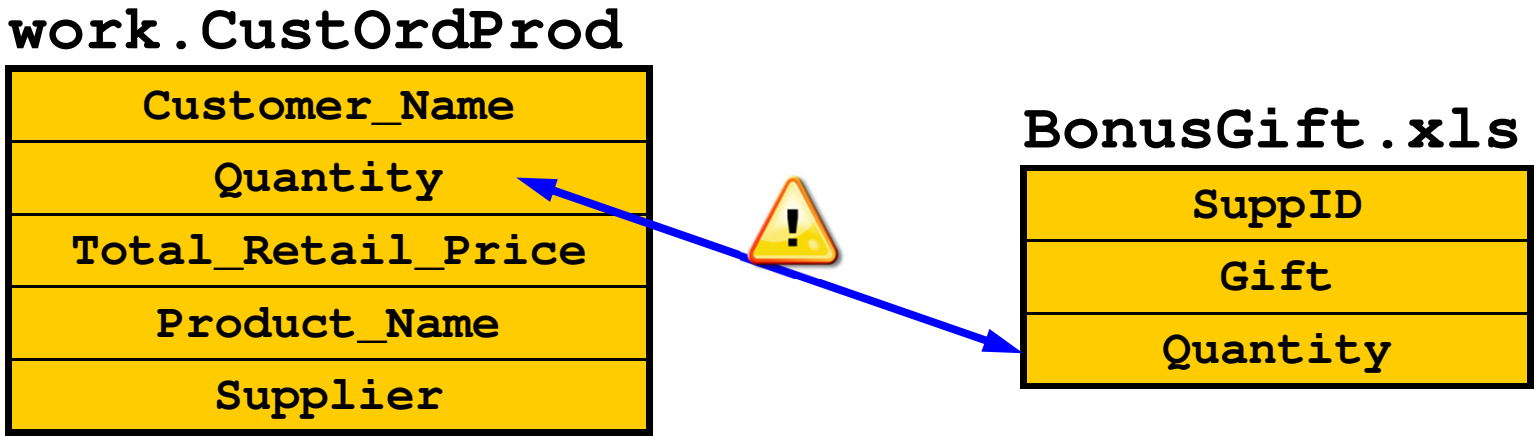
The data sets **work.CustOrdProd** and **BonusGift.xls** must be merged on values that are in two differently named variables.



The variables must have the same name for the match-merge to work correctly.

BUSINESS SCENARIO

You want to keep merged observations where the value of **Quantity** in **work.CustOrdProd** is more than the value of **Quantity** in **BonusGift.xls**.



The variables must have different names so that you can use a subsetting IF statement to compare them.

DEMO

COMPARING MERGING AND SQL

Match-Merge	SQL Inner Join
There is no limit to the number of data sets nor the size of the data sets other than disk space.	The maximum number of tables that can be joined at one time is 256.
Data is processed sequentially so that observations with duplicate BY values are joined one-to-one.	Data is processed using a Cartesian product for duplicate BY values.
Multiple data sets can be created.	Only one data set can be created with one CREATE TABLE statement.
Complex business logic can be incorporated using IF-THEN or SELECT/WHEN logic.	CASE logic can be used for business logic; however, it is not as flexible as DATA step syntax.
The data sets being merged must be sorted or indexed on the BY variable(s).	The data sets being joined do not have to be sorted nor indexed.
An exact match on the BY-variable(s) value(s) must be found.	Inequality joins can be performed.
Like-named BY variables must be available in all data sets.	Common variables do not have to be in all data sets.

3.2 DATA ANALYSIS

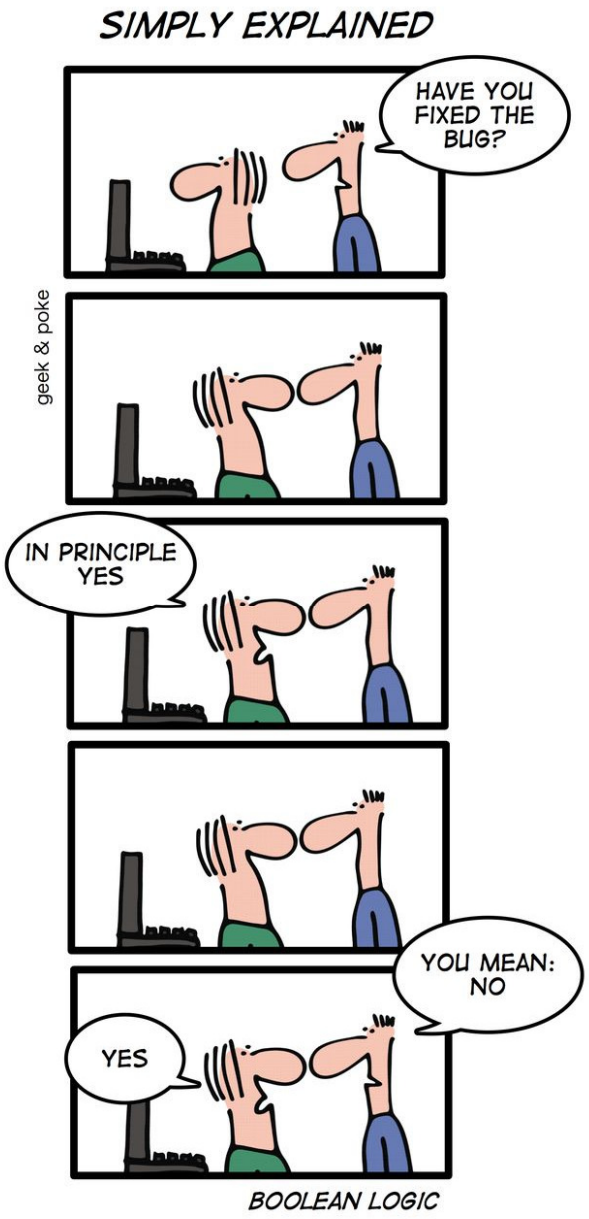
Quick – Which Base SAS technique can slice and dice your data like pivot tables?

HINT : No BI installation required
Involves some cheating

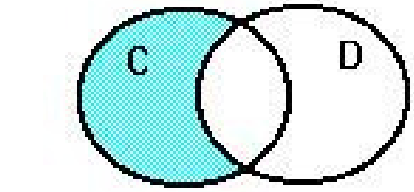
DEMO

3.3 DATA QUERYING

THE BOOLEAN GATE

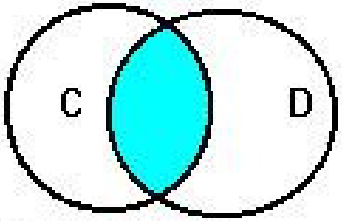


3.3 DATA QUERYING



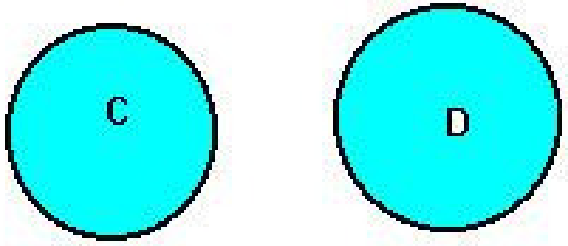
Cats NOT Dogs

Find all pages that have the word cats but don't have the word dogs.



Cats AND Dogs

Find all pages that have both the word cats and the word dogs.

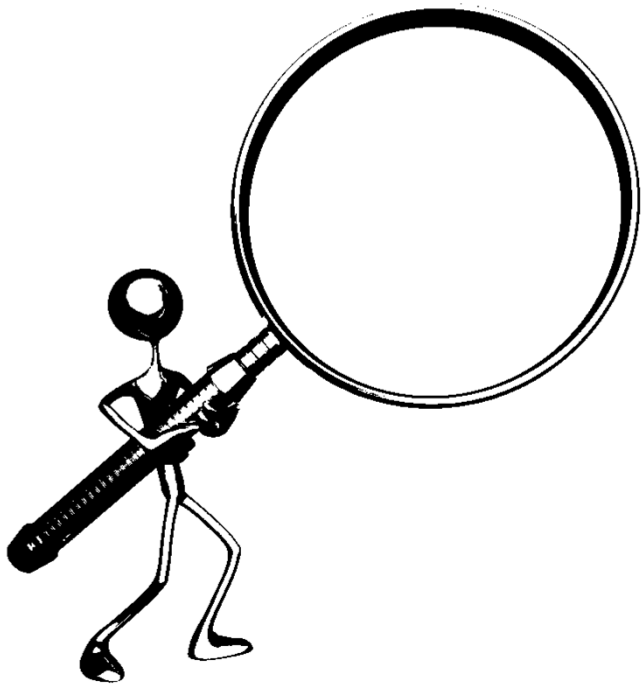


Cats OR Dogs

Find all pages that have the word cats and all pages that have the word dogs.

DEMO

3.4. DATA VALIDATION



DEMO

3.5 DATA CLEANSING

DEMO

3.6 DATA MANIPULATION

3.6 DATA MANIPULATION

The data set **orion.employee_payroll** contains each employee's hired date and current salary.

Partial Listing of **orion.employee_payroll**

Employee_ID	Employee_Gender	Salary	Birth_Date	Employee_Hire_Date	Employee_Term_Date	Marital_Status	Dependents
120101	M	163040	18AUG1976	01JUL2003	.	S	0
120102	M	108255	11AUG1969	01JUN1989	.	O	2
120103	M	87975	22JAN1949	01JAN1974	.	M	1
120104	F	46230	11MAY1954	01JAN1981	.	M	1
120105	F	27110	21DEC1974	01MAY1999	.	S	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

3.6 DATA MANIPULATION

The data set **orion.salary_stats** contains statistics for all Orion Star employees for the years 1974 through 2007. For example, the average salary of the employees hired in 1974 is currently $39,243.61.

Partial Listing of **orion.salary_stats**

Statistic	Yr1974	Yr1975	Yr1976	. . .	Yr2006	Yr2007
Num_of_Emps	61	4	6	. . .	97	3
Median_Salary	30025	29442.5	30020	. . .	26970	27240
Std_Salary	28551.9	9918.35	22356.91	. . .	2579.67	2922.12
Sum_Salary	2393860	132150	235030	. . .	2704720	86585
Avg_Salary	39243.61	33037.5	39171.67	. . .	27883.71	28861.67

3.6 DATA MANIPULATION

- The two data sets must be combined to calculate the difference between the average salary and the actual current salary for each employee based on the year of hire.
- Partial Listing of **compare**

Using One Dimensional Arrays

Obs	Employee_ID	Year_ Hired	Salary	Average	Salary_Dif
1	120101	2003	$163,040.00	$35,082.50	$127,957.50
2	120102	1989	$108,255.00	$88,588.75	$19,666.25
3	120103	1974	$87,975.00	$39,243.61	$48,731.39
4	120104	1981	$46,230.00	$36,436.67	$9,793.33
5	120105	1999	$27,110.00	$36,533.75	$-9,423.75
6	120106	1974	$26,960.00	$39,243.61	$-12,283.61
7	120107	1974	$30,475.00	$39,243.61	$-8,768.61
8	120108	2006	$27,660.00	$27,883.71	$-223.71

3.6 DATA MANIPULATION

Partial Listing of **orion.salary_stats**

Statistic	Yr1974	Yr1975	Yr1976	. . .	Yr2006	Yr2007
Avg_Salary	39243.61	33037.5	39171.67	. . .	27883.71	28861.67

Partial Listing of **orion.employee_payroll**

Employee_ID	Employee_Gender	Salary	Birth_Date	Employee_Hire_Date	. . .
120101	M	163040	18AUG1976	01JUL2003	. . .
120102	M	108255	11AUG1969	01JUN1989	. . .
120103	M	87975	22JAN1949	01JAN1974	. . .
120104	F	46230	11MAY1954	01JAN1981	. . .
120105	F	27110	21DEC1974	01MAY1999	. . .

3.6 DATA MANIPULATION

Can the two data sets be merged with the DATA step MERGE statement or joined with the SQL procedure without pre-processing the data?

- Yes
- No

3.6 DATA MANIPULATION

```
data compare;
   keep Employee_ID Year_Hired Salary Average
        Salary_Dif;
   format Salary Average Salary_Dif dollar12.2;
① array yr{1974:2007} Yr1974-Yr2007;
② if _N_=1 then set orion.salary_stats
          (where=(Statistic='Avg_Salary'));
   set orion.employee_payroll
       (keep=Employee_ID
             Employee_Hire_Date
             Salary);
   Year_Hired=year(Employee_Hire_Date);
③ Average=yr{Year_Hired};
   Salary_Dif=Salary-Average;
run;
```

DEMO

3.7 DATA MANAGEMENT

1. How can I look up Metadata?

DEMO

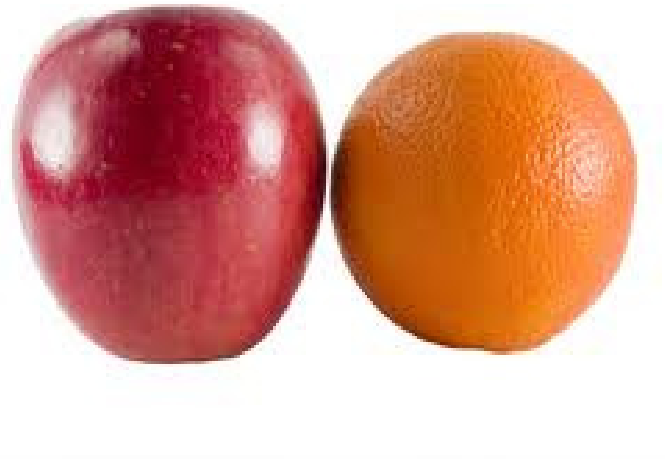
3.7 DATA MANAGEMENT

2. How can I easily investigate common columns for joins

DEMO

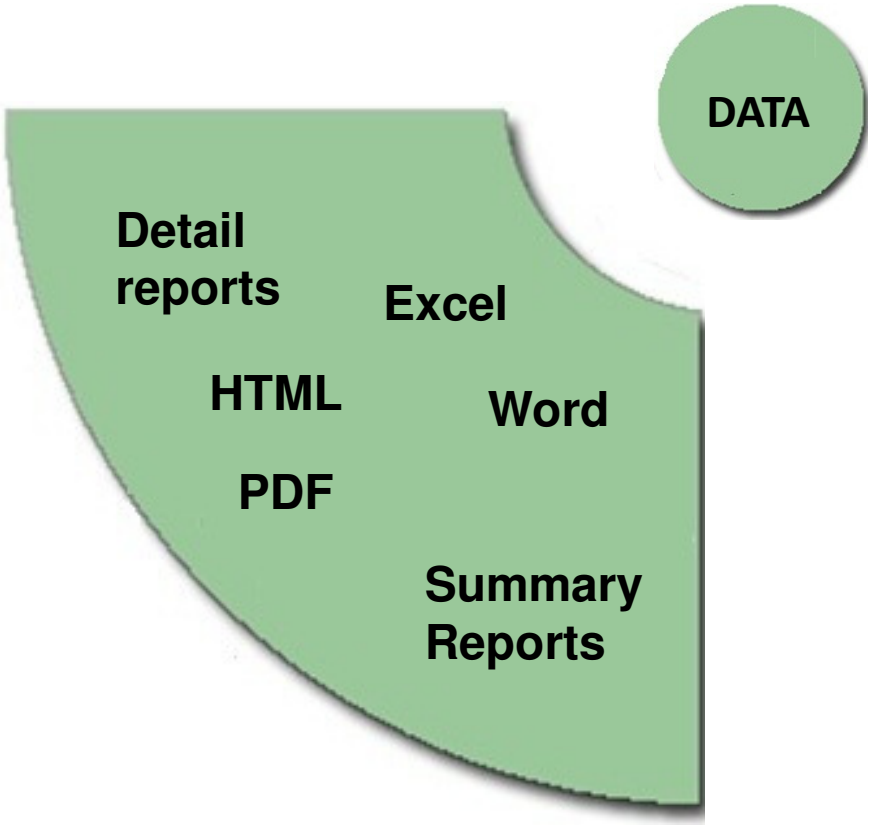
3.7 DATA MANAGEMENT

3. An efficiency question-PROC SQL or SAS datastep?



DEMO

4. DATA PRESENTATION



5. IN SUM

6. Q & A

Thank you!

Contact Charu.shankar@sas.com

Continue the conversation.
Connect with me on LinkedIn.
http://ca.linkedin.com/pub/charu-shankar/0/b42/892

7. FOR MORE....

Blog http://blogs.sas.com/content/sastraining/author/charushankar/

Boolean http://blogs.sas.com/content/sastraining/2012/05/10/1-sas-programming-tip-for-2012/

Programmer rule #1
http://blogs.sas.com/content/sastraining/2011/03/30/sas-programmer-rule-1/

BASE SAS 9.3 Procedures guide
http://support.sas.com/documentation/cdl/en/proc/65145/PDF/default/proc.pdf

PROC SQL SAS 9.3 User's guide
http://support.sas.com/documentation/cdl/en/sqlproc/63043/PDF/default/sqlproc.pdf