Working with Sparse Matrices in SAS®

Andrew T. Kuligowski, HSN

Andrew has been a SAS user for over 25 years – WELL over. Currently the Manager of CRM Data Infrastructure at HSN in St. Petersburg, he has augmented his professional coding experiences in the retail, media, insurance, and petrochemical fields by speaking at various SAS user events. Andrew was conference chair of SAS Global Forum 2012 here in Orlando, and co-chair for SESUG'97 in Jacksonville and Tennessee SAS Users Day in Knoxville. In his spare time, Andrew volunteers at the Florida Aquarium in Tampa.

Lisa Mendez Ph.D., QuintilesIMS

Lisa has been a SAS programmer for over 15 years. She has experience in various industries such as student achievement testing, clinical trials, medical equipment sales, retail grocery, and Military Health Systems (including workload & expense, business planning, and pharmacy).

Working with Sparse Matrices in SAS®

INTRODUCTION

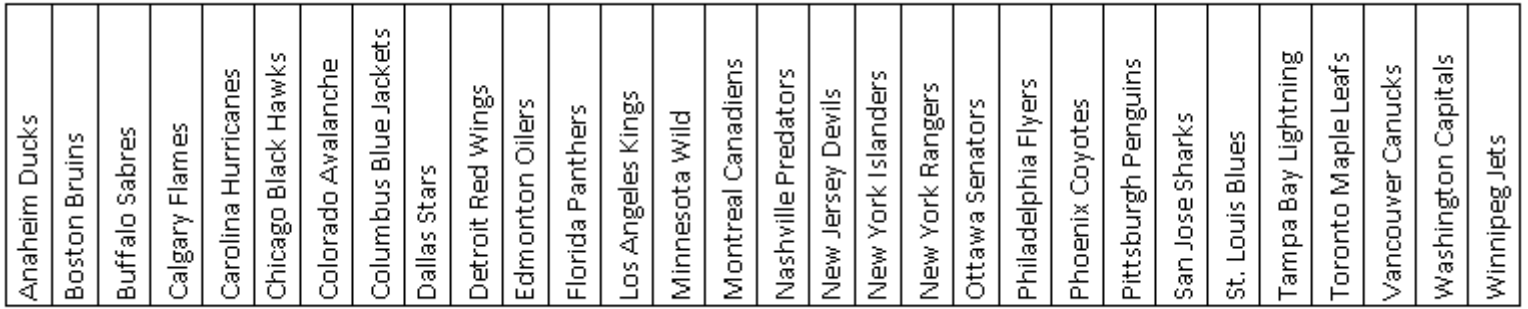
An ***Introduction*** to Sparse Matrices. (Matrixes?)

- Aimed at the BASE SAS programmer.
  (More advanced options available when SAS/STAT and/or SAS/IML are brought into the mix)

Background Information - Arrays

INTRODUCTION

Background: Arrays

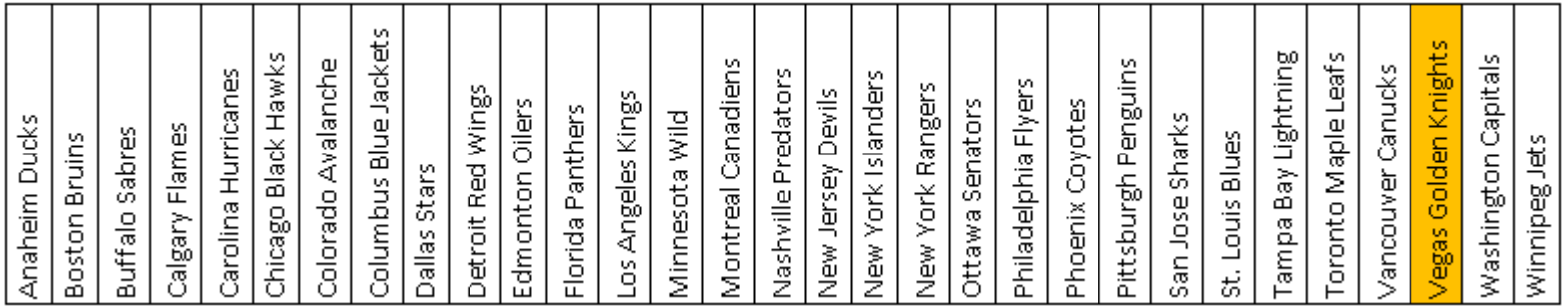



```
ARRAY NHL_TEAMS (30) Team01-Team30;
```

Background Information - Arrays

INTRODUCTION

Background: Arrays

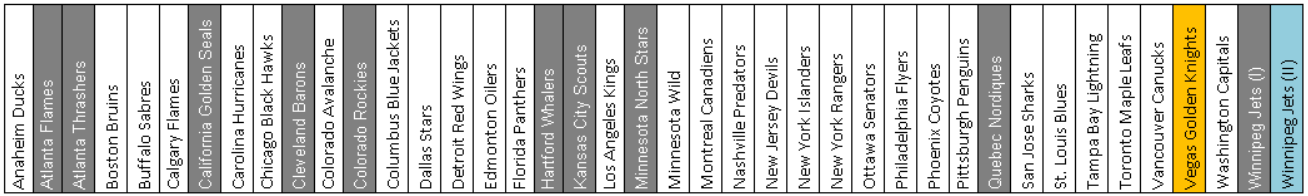



```
ARRAY NHL_TEAMS (31) Team01-Team31;
```

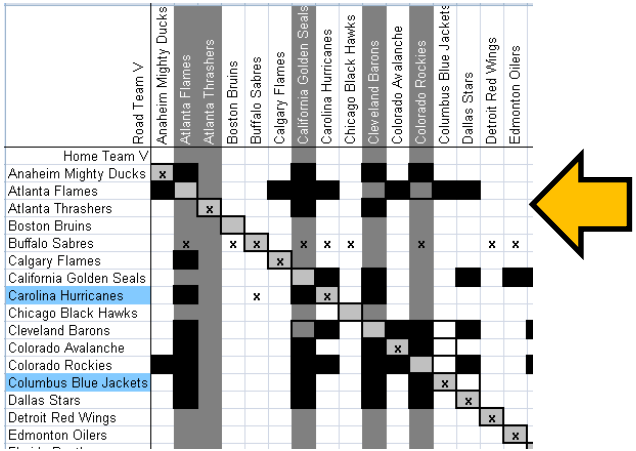
New Team
2017-18 Season

Background Information - Arrays

INTRODUCTION



Teams that have moved or folded (How far back do we go?)



Tracking which teams I've seen play vs. which opponents.

Then, factor in teams that have changed arenas. Or played outdoors. And neutral site games. And …

Then, factor in the year/season?

Then, factor in minor leagues , juniors, colleges…

What is exactly is a sparse matrix?

INTRODUCTION

… and we eventually have a large matrix. Very few of the cells actually contain a positive number denoting attendance at 1 or more games between two opponents in a given arena in a given season.

You may ask, “SO WHAT?”

According to Wikipedia …

*“In numerical analysis, a **sparse matrix** is a matrix in which most of the elements are zero.“*

What is exactly is a sparse matrix?

INTRODUCTION

- **Sparsity** = # of empty cells / total # of cells
- **Density** = # of populated cells / total # of cells
- What exactly *is* “sparse”? What exactly *is* “large”?
- THAT, my friends, is a matter of personal opinion and experience

What is exactly is a sparse matrix?

INTRODUCTION

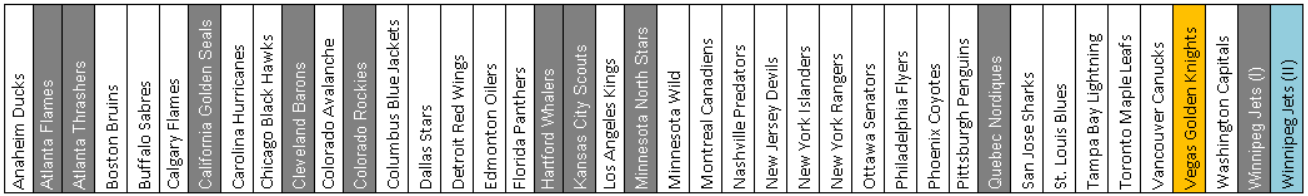
- Here is a very small example of a sparse matrix:

15	20	0	0	0	0	0	0
0	25	30	0	0	0	0	0
0	0	50	23	12	0	0	0
0	0	0	50	45	0	0	0
0	0	0	0	0	35	65	0
0	0	0	0	0	0	0	41

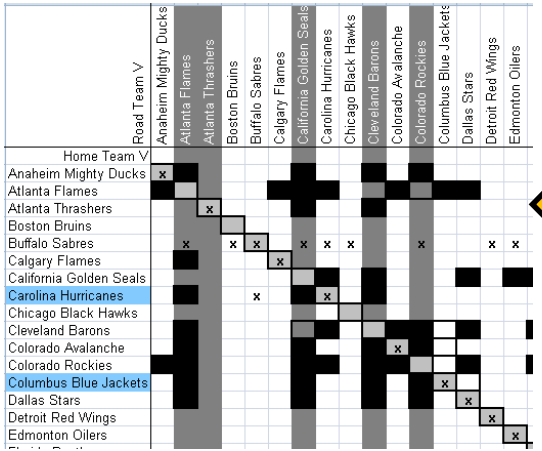
- Sparcity = (36/48) = 75% sparse
- Density = (12/48) = 25% dense

What is exactly is a sparse matrix?

"BRUTE FORCE" RESOLUTION



Teams that have moved or folded (How far back do we go?)



Tracking which tea… Then, factor in pl… Then, factor

```
ARRAY NHL_TEAMS (many, many)
                Team1-Team<many-squared>;
```

teams. Or played

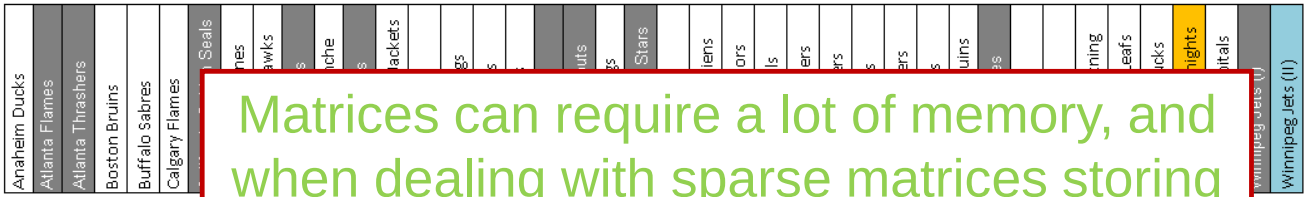
Then, factor in

```
ERROR: The SAS System stopped processing
this step because of insufficient memory.
```

colleges…

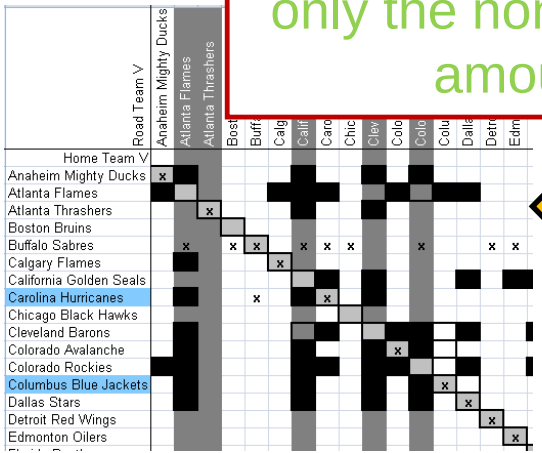
What is exactly is a sparse matrix?

"BRUTE FORCE" RESOLUTION



Teams that have moved or folded (How far back do we go?)

Matrices can require a lot of memory, and when dealing with sparse matrices storing only the non-zero values may reduce the amount of memory needed



Then, factor

```
ny, many)
Team1-Team<many-squared>;
```

teams. Or played

Then, factor in

```
ERROR: The SAS System stopped processing
this step because of insufficient memory.
```

colleges…

What is exactly is a sparse matrix?

"BRUTE FORCE" RESOLUTION

```
ARRAY NHL_TEAMS (many, many)
          Team1-Team<many-squared>;
ERROR: The SAS System stopped processing this step
  because of insufficient memory.
```

MEMSIZE option

- Must specify when starting SAS
- Default = 2G
- Can specify in bytes, kilobytes, megabytes, gigabytes, terabytes, or MAX
- A value of 0 is the same as specifying MAX

What is exactly is a sparse matrix

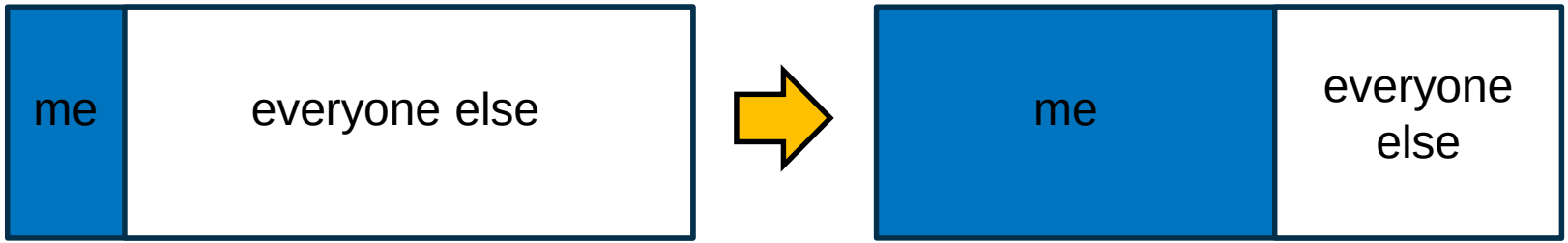
"BRUTE FORCE" RESOLUTION

MEMSIZE option - Potential issues:

Must specify when starting SAS. (Cannot change on the fly.)

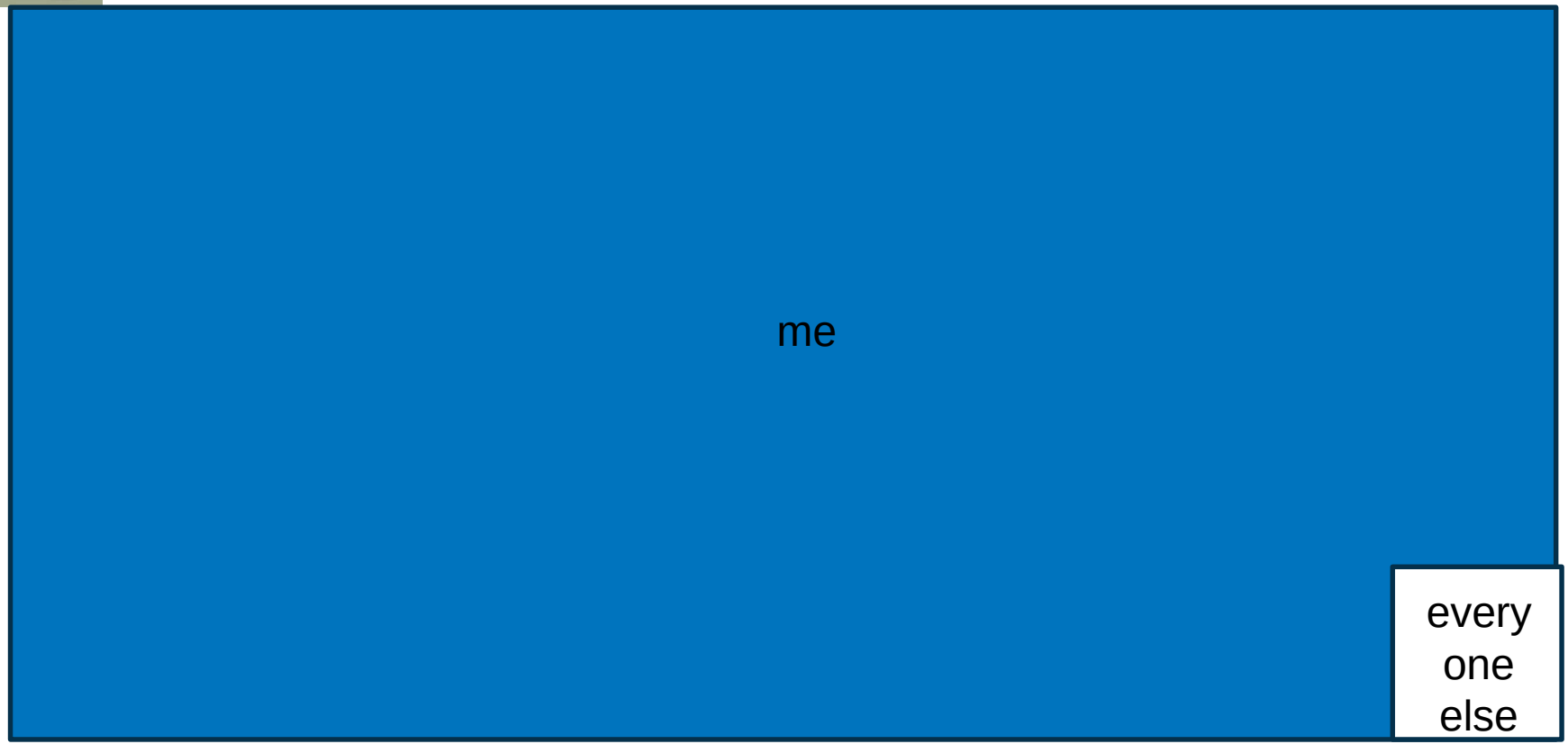
```
WARNING 30-12: SAS option MEMSIZE is valid only at startup
               of the SAS System. The SAS option is ignored.
```

- How much IS enough?
  Next time, will you want even more? Can you get THAT much?
- Personal computer vs. server
  If you're sharing the machine, how will this affect other users?



What is exactly is a sparse matrix

“BRUTE FORCE” RESOLUTION



What is exactly is a sparse matrix?

"BRUTE FORCE" RESOLUTION

MEMSIZE option - Potential issues:

```
65   PROC OPTIONS GROUP=MEMORY; RUN;

Group=MEMORY
```

SORTSIZE=1073741824 Specifies the amount of memory that is available to the SORT procedure.

SUMSIZE=0 Specifies a limit on the amount of memory that is available for data summarization procedures when class variables are active.

MAXMEMQUERY=268435456 For certain procedures, specifies the maximum amount of memory that can be allocated per request.

LOADMEMSIZE=0 Specifies a suggested amount of memory that is needed for executable programs loaded by SAS.

MEMSIZE=42949672960 Specifies the limit on the amount of virtual memory that can be used during a SAS session.

REALMEMSIZE=0 Specifies the amount of real memory SAS can expect to allocate

What is exactly is a sparse matrix?

"ITTY BITTY" RESOLUTION

Pinto: "Okay. That means that our whole solar system could be, like one tiny atom in the fingernail of some other giant being. … That means one tiny atom in my fingernail could be—"

Jennings: "Could be one little tiny universe."

-- from "Animal House"

8 bit byte: There are 8 "subatomic particles" that can be set to True (1) or False (0)

What is exactly is a sparse matrix?

"ITTY BITTY" RESOLUTION

Bit processing in SAS

Comparison:

```
IF   CHARVAR_LEN1 = '00001000'b   THEN ...
```

5th bit is 1, 1st-4th and 6th-8th bits are 0.

```
IF   CHARVAR_LEN1 = '....1...'b   THEN ...
```

5th bit is 1, 1st-4th and 6th-8th bits are ignored.

What is exactly is a sparse matrix?

"ITTY BITTY" RESOLUTION

Bit processing in SAS

Assignment:

```
CHARVAR_LEN1 = '00001000'b;
```

NO!

```
ERROR 216-185: The use of a BIT string constant is not
               allowed in this context.
```

Convert a character string of 1's and 0's into binary:

```
INPUT   NUMVAR_LEN1     binary8.;
```

Numvar_Len1 contains a numeric value based on 1 / 0 string passed in from input.

```
PUT     CHARVAR_LEN1    $binary8.;
```

Output contains an 8 byte character string consisting of 1 / 0 string corresponding to Charvar_Len1's internal representation.

What is exactly is a sparse matrix?

ARRAY'ZIN IN THE SUN

Hypothetical example: Grocer selling breakfast cereal

Did customers buy multiple kinds during their visit? If so, what kind(s)?

1) Set up a 2000 by 2000 matrix

(in theory; don't actually write an ARRAY statement!)

	P1	P2	P3	...	P1999	P2000
P1						
P2						
P3						
...						
P1999						
P2000						

This array has 2000*2000 (4 million) cells. Is that "large"? It depends on your perspective, BUT few would consider it "small".

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Hypothetical example: Grocer selling breakfast cereal

Did customers buy multiple kinds during their visit? If so, what kind(s)?

1) Many / most customers only by 1 kind of cereal

   (The number of boxes / bags of it that they buy is immaterial.)

	P1	P2	P3	...	P1999	P2000
P1						
P2						
P3						
...						
P1999						
P2000						

This would be known as a *diagonal matrix*.

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Hypothetical example: Grocer selling breakfast cereal

Did custo[illegible]at kind(s)?

1) Mar[illegible]

(The[illegible]erial.)

If this is all we have, it would be better represented as a single dimensional array, rather than as a 2 dimensional matrix!

	P1	P2	P3	...	P1999	P2000
P1						
P2						
P3						
...						
P1999						

This v

	P1	P2	P3	...	P1999	P2000
Customer						

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Hypothetical example: Grocer selling breakfast cereal

Did customers buy multiple kinds during their visit? If so, what kind(s)?

2) In this example, the remaining cells mirror each other across the diagonal. Customers who bought both P1 AND P3 can also be categorized as customers who bought both P3 AND P1.

	P1	P2	P3	...	P1999	P2000
P1			1			1
P2						
P3	1				1	
...						
P1999			1			
P2000	1					

This would be known as a *symmetrical matrix*.

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Hypothetical

Did customer

You may remember those kinds of charts from the old (pre-GPS) paper road atlases.

kind(s)?

1)

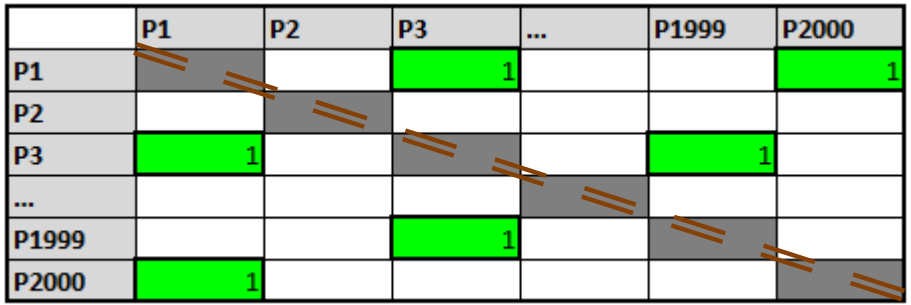
	Abilene	Amarillo	Arlington	Austin	Beaumont	Carrollton	Corpus Ch	Dallas	El Paso	Fort Wort
Amarillo	272.11									
Arlington	163.24	353.76								
Austin	227.36	489.4	192.29							
Beaumont	418.1	643.81	296.77	241.29						
Carrollton	183.23	356.15	24.97	208.43	290.44					
Corpus Ch	392.19	653.53	385.	193.22	286.24	400.74				
Dallas	181.65	365.95	19.06	194.1	276.42	14.82	388.39			
El Paso	449.17	416.78	612.11	583.99	818.54	631.04	693.53	630.38		
Fort Wort	150.59	343.89	14.2	186.64	303.14	33.96	380.93	33.	597.28	
Garland	195.56	372.37	33.91	208.07	277.73	17.87	402.37	15.04	642.25	49.56
Houston	360.74	600.12	253.08	162.68	85.19	252.45	212.62	238.41	734.1	257.31
Irving	174.42	358.68	13.92	201.15	287.44	12.49	395.45	11.34	621.11	29.14
Laredo	399.	630.29	424.13	232.35	418.38	439.87	167.66	426.59	618.69	417.42
Lubbock	163.48	120.54	315.63	375.63	586.5	328.67	530.24	333.23	368.08	306.31

Th

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

```
ARRAY DIAGONAL (2000) DIAGONAL0001-DIAGONAL2000;
IF  LAST.CUSTOMER_ID THEN DO;
    IF  FIRST.CUSTOMER_ID  THEN DO;
        DIAGONAL( CEREAL_ID ) = 1;
        OUTPUT  DIAGONAL_MATRIX;  /* One cell */
    END;
END;
```



	P1	P2	P3	...	P1999	P2000
P1			1			1
P2						
P3	1				1	
...						
P1999			1			
P2000	1					

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

~~ARRAY DIAGONAL (2000) DIAGONAL0001-DIAGONAL2000;~~

```
IF  LAST.CUSTOMER_ID THEN DO;
   IF  FIRST.CUSTOMER_ID  THEN DO;
      ~~DIAGONAL( CEREAL_ID ) = 1;~~  DIAGONAL = CEREAL_ID;
      OUTPUT  DIAGONAL_MATRIX;  /* One cell */
   END;
END;
```

	P1	P2	P3	...	P1999	P2000
P1			1			1
P2						
P3	1				1	
...						
P1999			1			
P2000	1					

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Faking up some sample data

```
data Cereal_Purchases;
  DO Customer_ID = 1 TO 5000;
    MaxBuys = FLOOR( RanUni( 0 )*2000 + 1 ) /
              FLOOR( RanUni( 0 )* 125 + 1 );
    IF  ROUND( Customer_ID, 75 ) = Customer_ID  THEN
          Ceiling_Buys = MaxBuys;
    ELSE  Ceiling_Buys = 12 ;
    NumBuys = MAX( FLOOR( RanUni( 0 )*Ceiling_Buys + 1 ), 1 );
    NumBuysMax = MAX( NumBuysMax, NumBuys );
    DO J = 1 TO NumBuys;
      Cereal_ID = FLOOR( RanUni( 0 )*2000 + 1 ) ;
      OUTPUT;
    END;
  END;
  CALL  SYMPUT( "NumBuysMax", NumBuysMax );
run;
```

```
%PUT INFO: NumBuysMax =  &NumBuysMax;
INFO: NumBuysMax =              482
```

```
NOTE: … WORK.CEREAL_PURCHASES has 32964 observations …
```

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

```
DATA Cereal_Approach;
   ARRAY CEREAL_BUYS (&NumBuysMax.) Buy1-Buy&NumBuysMax.;
   RETAIN Buy1-Buy&NumBuysMax. ;
   set Cereal_Purchases(KEEP=Customer_ID Cereal_ID);
   by Customer_ID;
   IF  First.Customer_ID  Then  Indx = 1;
   ELSE                         Indx + 1;
   Cereal_Buys( Indx ) = Cereal_ID;
   IF  Last.Customer_ID  THEN DO;
      Cereal_Count = Indx;
      output;
      DO J = 1 TO &NumBuysMax.;
         Cereal_Buys(J)=.;
      end;
   END;
run;
```

Replace with `Cereal_Count`

No need to run through 492 elements each time when we know how many we actually modified on each record.

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

First cut of the DATA step.
(HINT: It's not going to work.)

```
DATA Cereal_Combos;
   SET Cereal_Approach;
   ARRAY CEREAL_BUYS (&NumBuysMax.) Buy1-Buy&NumBuysMax.;
   DO  I = 1 TO Cereal_Count;
     Cereal_Purchase1 = Cereal_Buys(I);
     DO J = 1 TO Cereal_Count;
        Cereal_Purchase2 = Cereal_Buys(J);
        OUTPUT;
     END;
  END;
run;
```

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Fixes and Enhancements (a.k.a. "Cereal Killer")

```
DATA Cereal_Combos(KEEP=Customer_ID  Cereal_Count
                       Cereal_Purchase1  Cereal_Purchase2)
     Cereal_Single(KEEP=Customer_ID Cereal_Count Cereal_Purchase1);
   SET Cereal_Approach;
   ARRAY CEREAL_BUYS (&NumBuysMax.) Buy1-Buy&NumBuysMax.;
   IF  Cereal_Count = 1  THEN  DO;
      Cereal_Purchase1 = Cereal_Buys(1);
      OUTPUT Cereal_Single;
   END;
   ELSE  DO  I = 1 TO Cereal_Count - 1;
     Cereal_Purchase1 = Cereal_Buys(I);
     DO J = I + 1 TO Cereal_Count;
        Cereal_Purchase2 = Cereal_Buys(J);
        OUTPUT Cereal_Combos;
     END;
  END;
run;
```

What is exactly is a sparse matrix

ARRAY'ZIN IN THE SUN

Fixes and Enhancements (a.k.a. "Cereal Killer")

```
DATA Cereal_Combos(KEEP=Customer_ID  Cereal_Count
                       Cereal_Purchase1  Cereal_Purchase2)
                                                          e1);
```

Cutting to the chase to save some time …

Running this through a PROC MEANS or equivalent gets us a count of the number of customers who purchased each 2-production combination of cereal.

It will be a symmetrical matrix – or rather, ½ of a symmetrical matrix. If you need both halves, or you are splitting up the report by individual brand(s) of cereal, you may need to work both halves of the symmetry.

Unless purchases are incredibly skewed, you will probably no longer be working with a "sparse matrix" should you recreate the table using the summarized data. (It would definitely be expected that Density would go up, and Sparsity down.)

```
run;
```

What is exactly is a sparse matrix

STUFF BEYOND OUR SCOPE TODAY

We're not statisticians. (Other than a brief period of overconfidence after acing a STATS 301 test as an undergraduate, I never claimed to be.)

Those interested in topics outside of the scope of this paper may be interested in other papers – a quick search uncovered the following:

SAS/IML ®

See Kuss "A SAS/IML® Macro for Goodness-of-Fit Testing in Logistic Regression Models with Sparse Data" (SUGI 26)

SAS® TEXT MINER

See Zhao, Albright, and Cox "Processing and Storing Sparse Data in SAS® using SAS Text Miner Procedures" (SASGF 2014)

What is exactly is a sparse matrix

STUFF BEYOND OUR SCOPE TODAY

Additional papers:

PROC HPMIXED

See Wang / Tobias "All the Cows in Canada: Massive Mixed Modeling with the HPMIXED Procedure in SAS® 9.2" (SASGF 2009)

See Fenchel, McPhail, VanDyke "Using HPMIXED with Other SAS® 9.2 Procedures to Efficiently Analyze Large Dimension Registry Data" (MWSUG 2010)

These and other fine presentations can be found on: www.lexjansen.com

What is exactly is a sparse matrix

APPENDIX: TEST DATA

```
%LET  XDim = 100;
%LET  YDim = 100;
%LET  NumElem = %EVAL( &XDim. * &YDim. );
%PUT &NumElem.;

DATA temp;
   ARRAY BigDeal (&XDim., &YDim.)  Element_000001-Element_&NumElem.;
   DO I = 1 TO 5000;
      DO J = 1 TO 50;
         XDim = FLOOR( Ranuni( 0 )*&XDim. + 1 ) ;
         YDim = FLOOR( Ranuni( 0 )*&YDim. + 1 ) ;
         BigDeal(XDim, YDim) = 1;
      END;
      OUTPUT;
      DO J = 1 TO &YDim.;
         DO J1 = 1 TO &YDim.;
            BigDeal( J, J1 ) = .;
         END;
      END;
   END;
RUN;
```

Questions?