

# Advanced Techniques for Working With Big Data in JMP

Visualization, Analysis, Data Preparation, and More

**Julian L. Parris, Ph.D.**  
JMP Learning Strategy Manager  
SAS Institute Inc.



# Today

- The challenge of big data
- Methods for working with big data in JMP
- JiEL – JMP in Everyday Life

# Defining Big Data

**Volume** [ Number of records (tall)  
Number of variables (wide)

**Velocity** [ Frequency of Update (up to real time)  
Frequency of Analysis

**Variety** [ Expanded Types of Data  
Character, Numeric, Free Text,  
Audio, Video, Genomic

**Veracity** [ Data Quality  
Accuracy (data), Validity (method)

# Limitations

Processing speed

Memory

**Your Time**

**Your Patience**

# Big Data Methods in JMP

## Complete Data

- Tall Data
  - Virtual joins
- Wide Data (Big Statistics)
  - Data Filtering/Switching
  - Predictor/Response Screening
  - Data table graphs
  - Dimension reduction
  - Automated variable selection

## Proxy Methods (Sampling)

- The case for subsets
- Modeling with subsets
- Data preparation with subsets
  - Missing value pattern
  - Outlier screening
  - Formula columns

# Tall and Wide Data

(In Memory)



World Development Indicators  
**16.51 GB    ~305 million data points    1402 variables**



# Tall Data

via Proxy Methods



Airline Flight Data 1987-2008  
**33.60 GB** *3.58 billion data points*




00:00.00

Reset Start



1987\_2008.jmp

1987\_2008.jmp Info

 **1987\_2008.jmp** 33.6 GB  
Modified: Yesterday, 9:00 PM

Add Tags...

▼ General:

Kind: JMP data table  
Size: 33,601,516,059 bytes (33.6 GB on disk)

1987\_2008.jmp

- Source
- Files

Columns (29/0)

- Year
- Month
- DayofMonth
- DayOfWeek
- DepTime
- CRSDepTime
- ArrTime
- CRSArrTime
- UniqueCarrier
- FlightNum
- TailNum
- ActualElapsedTime
- CRSElapsedTime
- AirTime

Rows

All rows	123,534,969
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime
1	1987	10	14	3	741	730	912	849	PS					91
2	1987	10	15	4	729	730	903	849	PS	1451	NA			94
3	1987	10	17	6	741	730	918	849	PS	1451	NA			97
4	1987	10	18	7	729	730	847	849	PS	1451	NA			78
5	1987	10	19	1	749	730	922	849	PS	1451	NA			93
6	1987	10	21	3	728	730	848	849	PS	1451	NA			80
7	1987	10	22	4	728	730	852	849	PS	1451	NA			84
8	1987	10	23	5	731	730	902	849	PS	1451	NA			91
9	1987	10	24	6	744	730	908	849	PS	1451	NA			84
10	1987	10	25	7	729	730	851	849	PS	1451	NA			82
11	1987	10	26	1	735	730	904	849	PS	1451	NA			89
12	1987	10	28	3	741	725	919	855	PS	1451	NA			98
13	1987	10	29	4	742	725	906	855	PS	1451	NA			84
14	1987	10	31	6	726	725	848	855	PS	1451	NA			82
15	1987	10	1	4	936	915	1035	1001	PS	1451	NA			59
16	1987	10	2	5	918	915	1017	1001	PS	1451	NA			59
17	1987	10	3	6	928	915	1037	1001	PS	1451	NA			69
18	1987	10	4	7	914	915	1003	1001	PS	1451	NA			49
19	1987	10	5	1	1042	915	1129	1001	PS	1451	NA			47
20	1987	10	6	2	934	915	1024	1001	PS	1451	NA			50
21	1987	10	7	3	946	915	1037	1001	PS	1451	NA			51
22	1987	10	8	4	932	915	1033	1001	PS	1451	NA			61
23	1987	10	9	5	947	915	1036	1001	PS	1451	NA			49
24	1987	10	10	6	915	915	1022	1001	PS	1451	NA			67
25	1987	10	11	7	916	915	1006	1001	PS	1451	NA			50
26	1987	10	12	1	944	915	1027	1001	PS	1451	NA			43
27	1987	10	13	2	941	915	1036	1001	PS	1451	NA			55
28	1987	10	14	3	930	915	1029	1001	PS	1451	NA			59

00:00.00

Reset Start

# Why Subsets?

A review of the properties of statistical estimators:  
*asymptotic consistency*

$$\mathbf{\text{plim}}_{n \rightarrow \infty} T_n = \theta.$$

a sample of 1 million  
is nearly as good as 100 million\*

## The Value of Analytic Immediacy


00:00.00

Reset Start



1m Row Subset of 1987\_2008.jmp

1m Row Subset of 1987\_2008.jmp Info

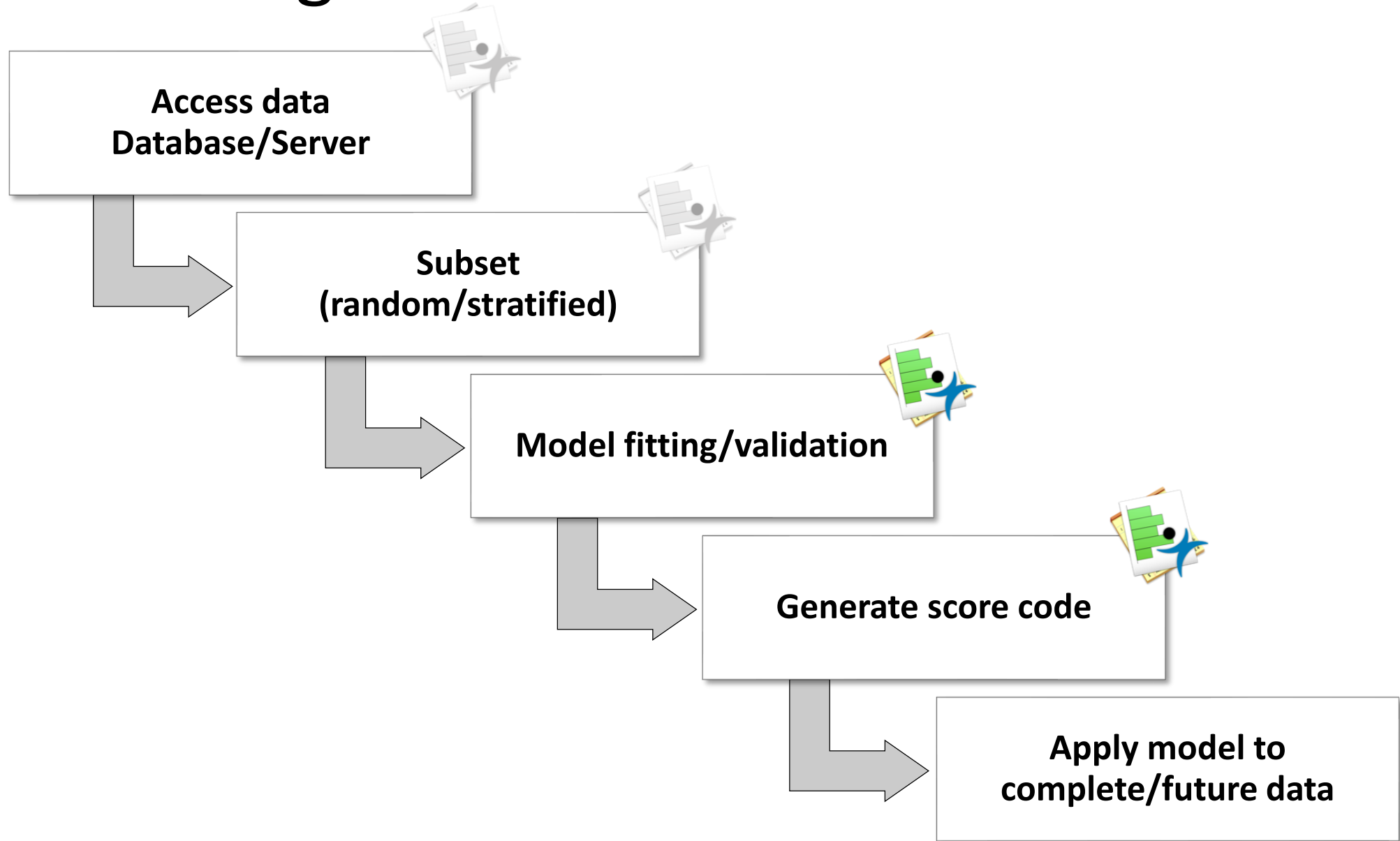
 **1m Row Subset of 1987\_2008.jmp** 272 MB  
Modified: Today, 3:59 PM

Add Tags...

▼ General:

Kind: JMP data table  
Size: 272,003,365 bytes (272 MB on disk)

# Proxy Modeling Process



# Today

- The challenge of big data
- Methods for working with big data in JMP
- JiEL – JMP in Everyday Life

# JMP in Everyday Life



Fitbit Steps Data, 2015-Present  
*2.18 million rows @ 1 Minute Resolution*