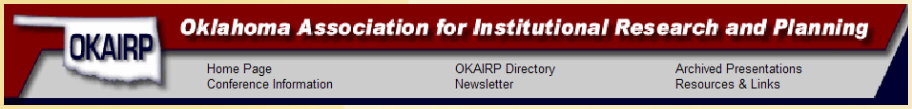


2005 Fall Conference

Data Mining, Dashboards and Data Quality

John Rome, Arizona State University

Review of "BI" Buzzwords

OLAP

De-Normalized

Data Mart

Operation Data Store (ODS)

Bit-Mapped Indexing

Drill-Down

Data Mining

ROLAP

Aggregation

MOLAP

XML

Replication

Facts/Dimensions

Data Quality

Metadata

Business Intelligence

Star Schema

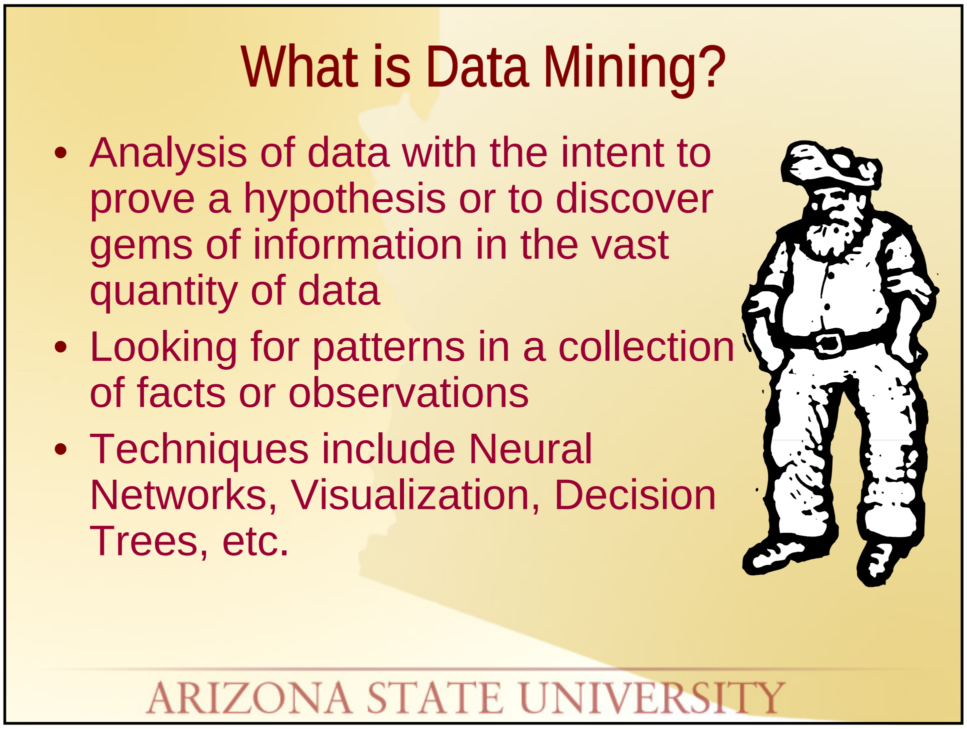
Multi-dimensional

Dashboards

Transformation Tools (ETL)

SQL

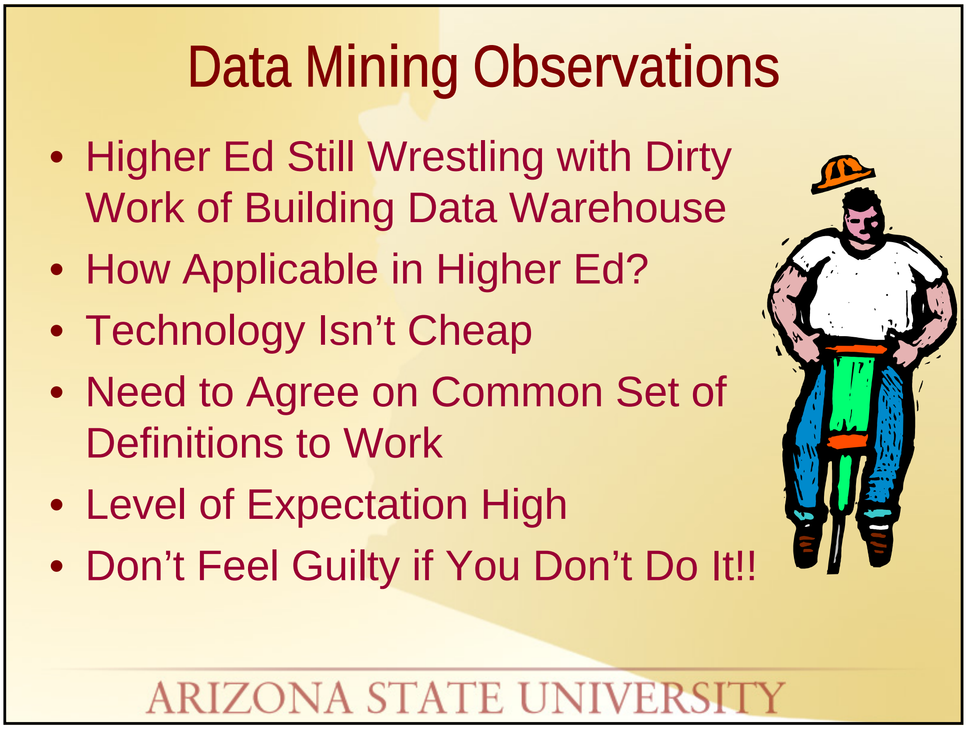
Snowflake Schema



What is Data Mining?

- Analysis of data with the intent to prove a hypothesis or to discover gems of information in the vast quantity of data
- Looking for patterns in a collection of facts or observations
- Techniques include Neural Networks, Visualization, Decision Trees, etc.

ARIZONA STATE UNIVERSITY

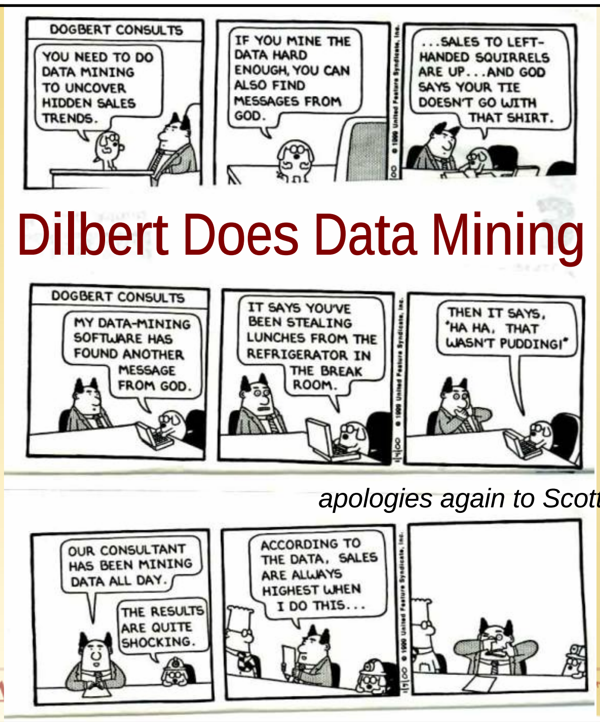


Data Mining Observations

- Higher Ed Still Wrestling with Dirty Work of Building Data Warehouse
- How Applicable in Higher Ed?
- Technology Isn’t Cheap
- Need to Agree on Common Set of Definitions to Work
- Level of Expectation High
- Don’t Feel Guilty if You Don’t Do It!!

ARIZONA STATE UNIVERSITY

Dilbert Does Data Mining


DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.
DOGBERT CONSULTS
MY DATA-MINING SOFTWARE HAS FOUND ANOTHER MESSAGE FROM GOD.
IT SAYS YOU'VE BEEN STEALING LUNCHES FROM THE REFRIGERATOR IN THE BREAK ROOM.
THEN IT SAYS, "HA HA, THAT WASN'T PUDDING!"
OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...

apologies again to Scott Adams!!

Santa Claus Uses Data Mining…

Here's some news, and brace yourselves
There's trouble hiring Santa's elves
The little guys put down their hammers
And took new job as C programmers

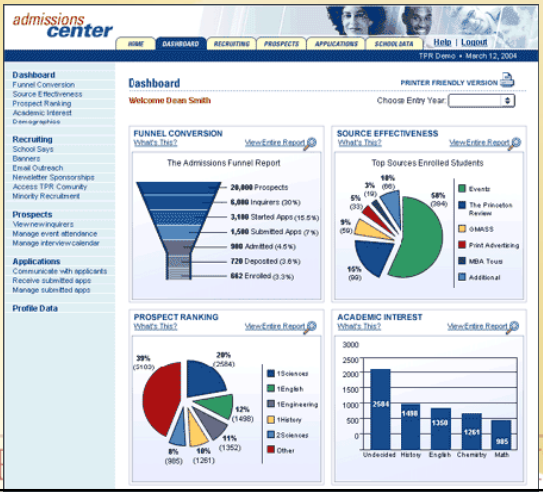
But don't you worry, Santa's finding
Salvation in **data mining**
To prep his list, his query would be…

```
SELECT * FROM KIDS, STATUS = 'GOOD'
```

-Anonymous

What are Dashboards?

A dashboard is a graphical display that compares performance against predefined goals.



Types of Dashboards?

- Operational
- Strategic
- Tactical

-Wayne Eckerson

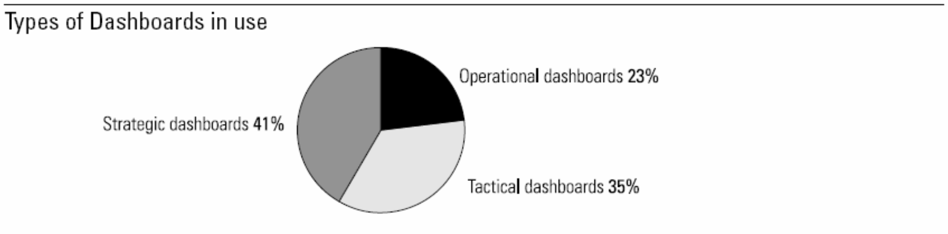


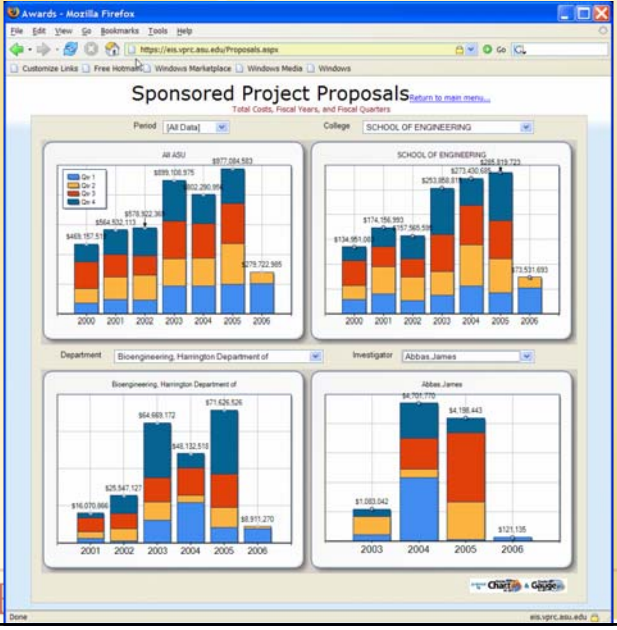
Illustration 32. A larger percentage of organizations are implementing strategic dashboards than any other type of dashboard. Based on 240 respondents who said their group has deployed a dashboard.

Dashboard Tools…

How Organizations Are Building:

- 41% - BI Tools
- 22% - Custom Code
- 13% - Microsoft
- 17% - Purchased from Vendor

Example of Dashboard



Sponsored Project Proposals

Total Costs, Fiscal Years, and Fiscal Quarters

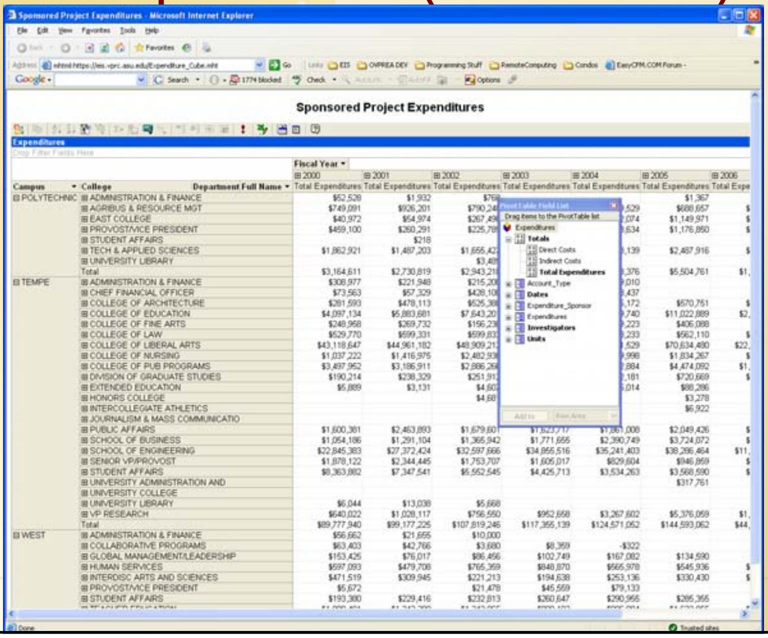
Period: [All Data]

College: SCHOOL OF ENGINEERING

Department: Bioengineering, Harrington Department of

Investigator: Abbas James

Example of MDD (Dashboard?)


Sponsored Project Expenditures - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Favorites
Address mhtml:https://eis.vprc.asu.edu/Expenditure_Cube.mht
Go
Links EIS OVPREA DEV Programming Stuff RemoteComputing Condos EasyCFM.COM Forum -
Google Search 1774 blocked Check Options
Expenditures
Drop Filter Fields Here
PivotTable Field List
Drag items to the PivotTable list
Expenditures
Totals
Direct Costs
Indirect Costs
Total Expenditures
Account_Type
Dates
Expenditure_Sponsor
Expenditures
Investigators
Units
Add to
Row Area
Done
Trusted sites

Sponsored Project Expenditures

Fiscal Year

Campus	College / Department Full Name	2000 Total Expenditures	2001 Total Expenditures	2002 Total Expenditures	2003 Total Expenditures	2004 Total Expenditures	2005 Total Expenditures	2006 Total Expe
POLYTECHNIC	ADMINISTRATION & FINANCE	$52,528	$1,932	$768			$1,367	
	AGRIBUS & RESOURCE MGT	$749,091	$926,201	$790,24		,529	$688,657	$
	EAST COLLEGE	$40,972	$54,974	$267,49		,074	$1,149,971	$
	PROVOST/VICE PRESIDENT	$459,100	$260,291	$225,78		,634	$1,176,850	$
	STUDENT AFFAIRS		$218					
	TECH & APPLIED SCIENCES	$1,862,921	$1,487,203	$1,655,42		,139	$2,487,916	$
	UNIVERSITY LIBRARY			$3,48				
	Total	$3,164,611	$2,730,819	$2,943,21		,376	$5,504,761	$1,
TEMPE	ADMINISTRATION & FINANCE	$308,977	$221,948	$215,20		,010		
	CHIEF FINANCIAL OFFICER	$73,563	$57,329	$428,10		,437		
	COLLEGE OF ARCHITECTURE	$281,593	$478,113	$525,38		,172	$570,751	$
	COLLEGE OF EDUCATION	$4,097,134	$5,883,681	$7,643,20		,740	$11,022,889	$2,
	COLLEGE OF FINE ARTS	$248,958	$269,732	$156,23		,223	$406,088	
	COLLEGE OF LAW	$529,770	$599,331	$599,83		,233	$562,110	$
	COLLEGE OF LIBERAL ARTS	$43,118,647	$44,961,182	$48,909,21		,529	$70,634,480	$22,
	COLLEGE OF NURSING	$1,037,222	$1,416,975	$2,482,93		,998	$1,834,267	$
	COLLEGE OF PUB PROGRAMS	$3,497,952	$3,186,911	$2,886,26		,884	$4,474,092	$1,
	DIVISION OF GRADUATE STUDIES	$190,214	$238,329	$251,91		,181	$720,669	$
	EXTENDED EDUCATION	$5,889	$3,131	$4,60		,014	$88,286	
	HONORS COLLEGE			$4,68			$3,278	
	INTERCOLLEGIATE ATHLETICS						$6,922	
	JOURNALISM & MASS COMMUNICATIO							
	PUBLIC AFFAIRS	$1,600,381	$2,463,893	$1,679,601	$1,623,717	$1,861,008	$2,049,426	$
	SCHOOL OF BUSINESS	$1,054,186	$1,291,104	$1,365,942	$1,771,655	$2,390,749	$3,724,072	$
	SCHOOL OF ENGINEERING	$22,845,383	$27,372,424	$32,597,666	$34,855,516	$35,241,403	$38,286,464	$11,
	SENIOR VP/PROVOST	$1,878,122	$2,344,445	$1,753,707	$1,606,017	$829,604	$946,859	$
	STUDENT AFFAIRS	$8,363,882	$7,347,541	$5,552,545	$4,425,713	$3,534,263	$3,568,590	$
	UNIVERSITY ADMINISTRATION AND						$317,761	
	UNIVERSITY COLLEGE							
	UNIVERSITY LIBRARY	$6,044	$13,038	$5,668				
	VP RESEARCH	$640,022	$1,028,117	$756,550	$952,658	$3,267,602	$5,376,059	$1,
	Total	$89,777,940	$99,177,225	$107,819,246	$117,355,139	$124,571,052	$144,593,062	$44,
WEST	ADMINISTRATION & FINANCE	$56,662	$21,655	$10,000				
	COLLABORATIVE PROGRAMS	$63,403	$42,766	$3,680	$8,359	-$322		
	GLOBAL MANAGEMENT/LEADERSHIP	$153,425	$76,017	$86,456	$102,749	$167,082	$134,590	
	HUMAN SERVICES	$597,093	$479,708	$765,359	$848,870	$565,978	$545,936	$
	INTERDISC ARTS AND SCIENCES	$471,519	$309,945	$221,213	$194,638	$253,136	$330,430	$
	PROVOST/VICE PRESIDENT	$5,672		$21,478	$45,559	$79,133		
	STUDENT AFFAIRS	$193,380	$229,416	$232,813	$260,647	$290,966	$285,355	

What is Data Quality (DQ)?

Data Having:

- Accuracy
- Integrity
- Consistency
- Completeness
- Validity
- Timeliness
- Accessibility

Data Quality Impact

“The Business costs of nonquality data, including irrecoverable costs, rework of products or services, workarounds, and lost and missed revenue may be as high as ***<u>10 to 25 percent of revenue or total budget</u>*** *of an organization.”*

-Larry English

“Father of Data Quality”

Data

It’s good,

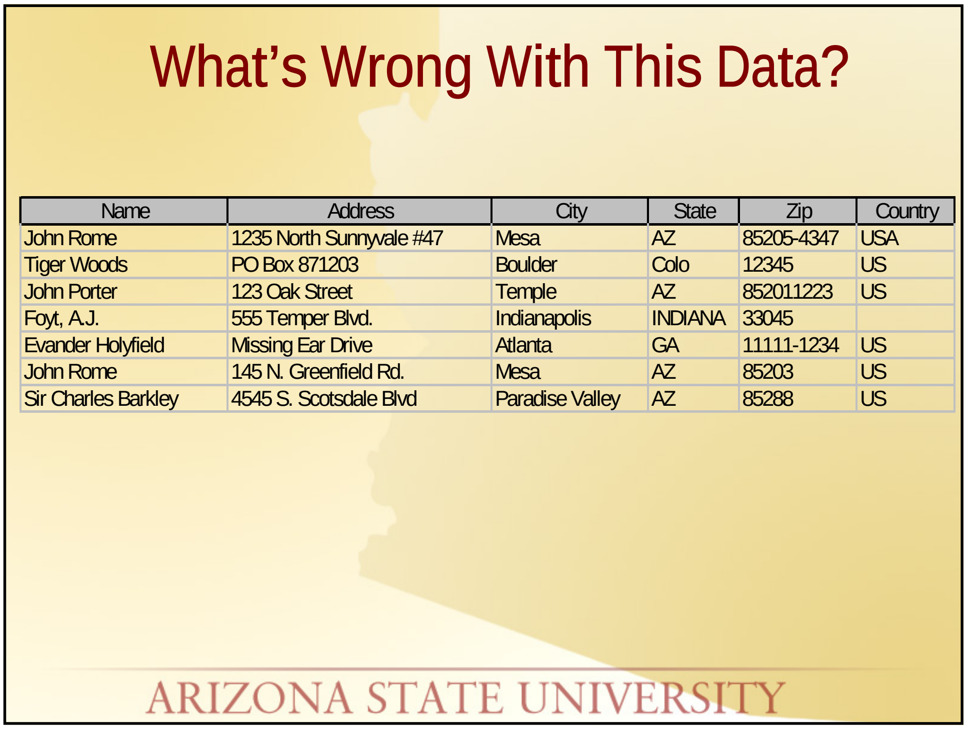
it’s bad,

and it’s ugly!

Mainly Ugly

Problems I’ve Seen…

- Building named “China”
- High number of +95 year old students
- Is Pat “M” or “F”?
- Negative SQFT
- Addresses
- Gender Identity
- Macintosh worth $6



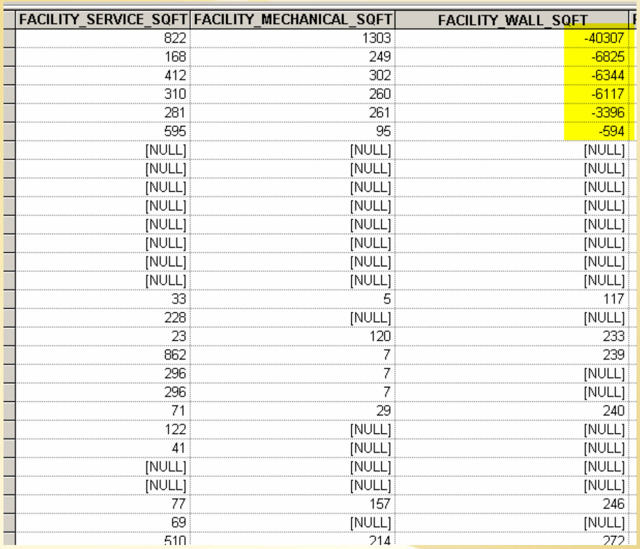
What’s Wrong With This Data?

Name	Address	City	State	Zip	Country
John Rome	1235 North Sunnyvale #47	Mesa	AZ	85205-4347	USA
Tiger Woods	PO Box 871203	Boulder	Colo	12345	US
John Porter	123 Oak Street	Temple	AZ	852011223	US
Foyt, A.J.	555 Temper Blvd.	Indianapolis	INDIANA	33045	
Evander Holyfield	Missing Ear Drive	Atlanta	GA	11111-1234	US
John Rome	145 N. Greenfield Rd.	Mesa	AZ	85203	US
Sir Charles Barkley	4545 S. Scotsdale Blvd	Paradise Valley	AZ	85288	US

Data Warehouse Helps Identify…

1. Objectionable Outliers
2. Nuisance Nulls
3. Disorderly Domains
4. Downright Bad Data

Objectionable Outlier Example

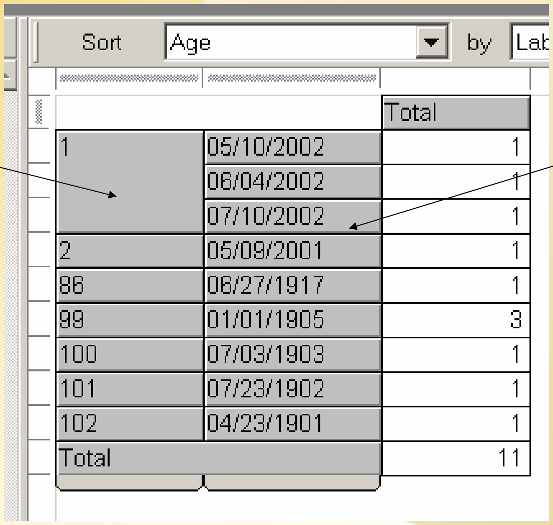


FACILITY_SERVICE_SQFT	FACILITY_MECHANICAL_SQFT	FACILITY_WALL_SQFT
822	1303	-40307
168	249	-6825
412	302	-6344
310	260	-6117
281	261	-3396
595	95	-594
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
33	5	117
228	[NULL]	[NULL]
23	120	233
862	7	239
296	7	[NULL]
296	7	[NULL]
71	29	240
122	[NULL]	[NULL]
41	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
[NULL]	[NULL]	[NULL]
77	157	246
69	[NULL]	[NULL]
510	214	272

Objectionable Outlier Example

AGE

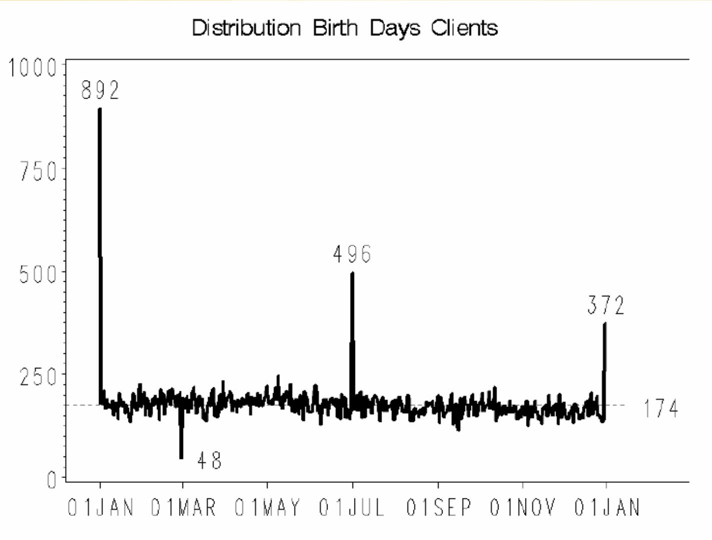
BIRTHDATE



Sort Age by Lab

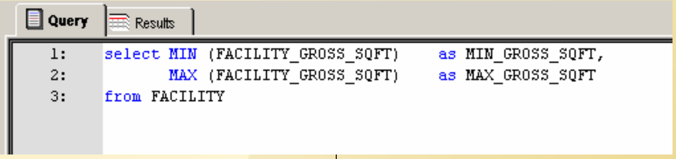
		Total
1	05/10/2002	1
	06/04/2002	1
	07/10/2002	1
2	05/09/2001	1
86	06/27/1917	1
99	01/01/1905	3
100	07/03/1903	1
101	07/23/1902	1
102	04/23/1901	1
Total		11

What Wrong with These Birthdates?

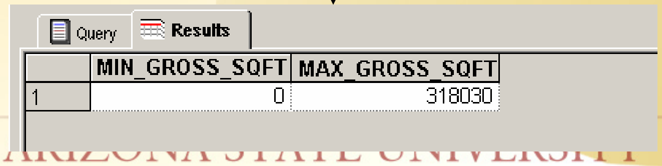


Objectionable Outlier Techniques

- Min/Max Functions (SQL)
- Standard Deviation
- Data Visualization (after sorting results)

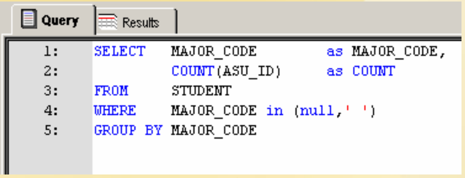


```
select MIN (FACILITY_GROSS_SQFT)     as MIN_GROSS_SQFT,
       MAX (FACILITY_GROSS_SQFT)     as MAX_GROSS_SQFT
from FACILITY
```

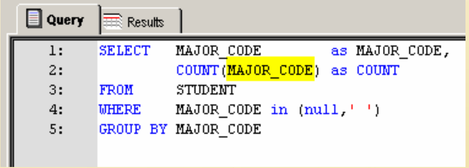


	MIN_GROSS_SQFT	MAX_GROSS_SQFT
1	0	318030

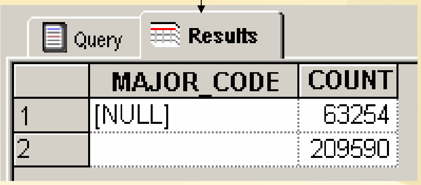
Nuisance Nulls Example



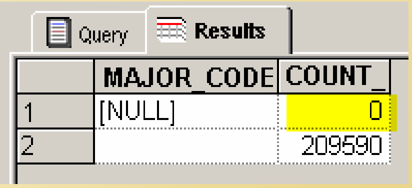
```
SELECT   MAJOR_CODE          as MAJOR_CODE,
         COUNT(ASU_ID)       as COUNT
FROM     STUDENT
WHERE    MAJOR_CODE in (null,' ')
GROUP BY MAJOR_CODE
```



```
SELECT   MAJOR_CODE          as MAJOR_CODE,
         COUNT(MAJOR_CODE)   as COUNT
FROM     STUDENT
WHERE    MAJOR_CODE in (null,' ')
GROUP BY MAJOR_CODE
```



	MAJOR_CODE	COUNT
1	[NULL]	63254
2		209590

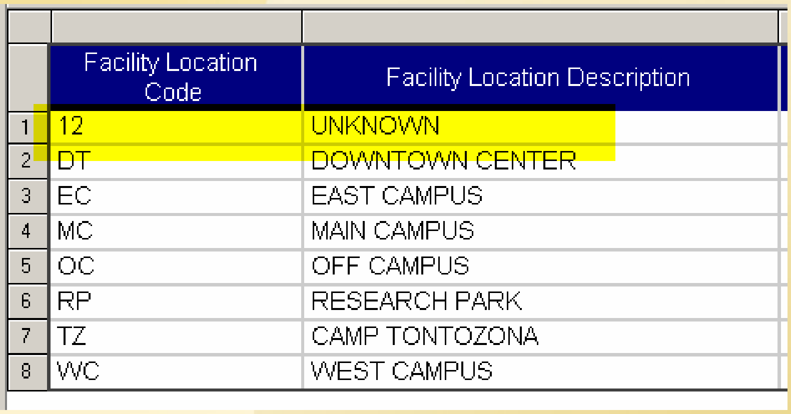


	MAJOR_CODE	COUNT_
1	[NULL]	0
2		209590

Nuisance Nulls Techniques

- WHERE Column_Name in (null, ' ')
- Data Visualization (after sorting results)

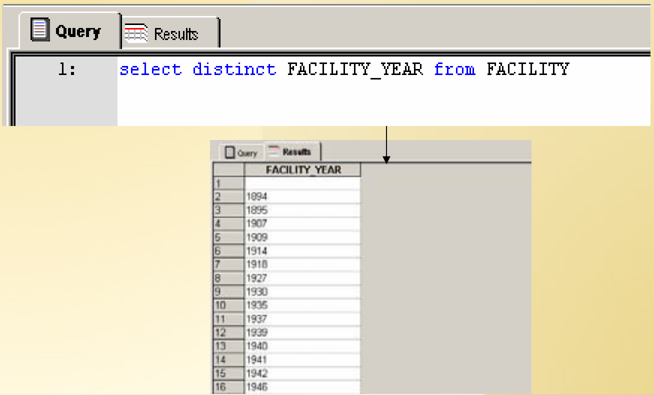
Disorderly Domains Example



	Facility Location Code	Facility Location Description
1	12	UNKNOWN
2	DT	DOWNTOWN CENTER
3	EC	EAST CAMPUS
4	MC	MAIN CAMPUS
5	OC	OFF CAMPUS
6	RP	RESEARCH PARK
7	TZ	CAMP TONTOZONA
8	WC	WEST CAMPUS

Disorderly Domains Techniques

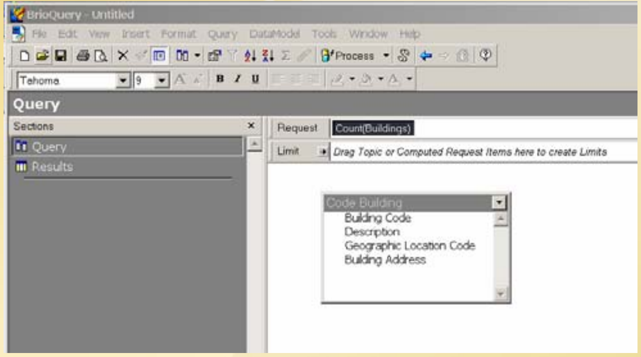
- Select Distinct (SQL)
- Outer joins to find missing values



```
select distinct FACILITY_YEAR from FACILITY
```

	FACILITY_YEAR
1	
2	1894
3	1895
4	1907
5	1909
6	1914
7	1918
8	1927
9	1930
10	1935
11	1937
12	1939
13	1940
14	1941
15	1942
16	1946

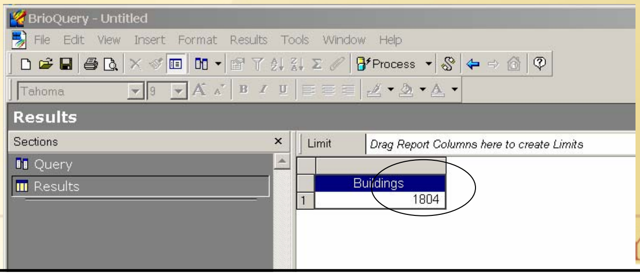
Downright Bad Data Example


BrioQuery - Untitled
File Edit View Insert Format Query DataModel Tools Window Help
Process
Tahoma
9
Query
Sections
Query
Results

Request: Count(Buildings)

Limit: Drag Topic or Computed Request Items here to create Limits

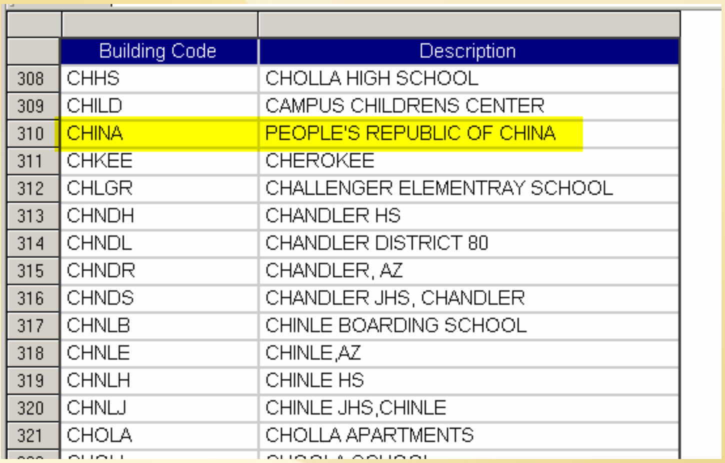
Code Building
Building Code
Description
Geographic Location Code
Building Address



Limit: Drag Report Columns here to create Limits

	Buildings
1	1804

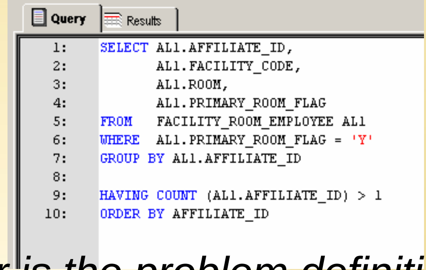
Some of the Buildings We Found…



	Building Code	Description
308	CHHS	CHOLLA HIGH SCHOOL
309	CHILD	CAMPUS CHILDRENS CENTER
310	CHINA	PEOPLE'S REPUBLIC OF CHINA
311	CHKEE	CHEROKEE
312	CHLGR	CHALLENGER ELEMENTRAY SCHOOL
313	CHNDH	CHANDLER HS
314	CHNDL	CHANDLER DISTRICT 80
315	CHNDR	CHANDLER, AZ
316	CHNDS	CHANDLER JHS, CHANDLER
317	CHNLB	CHINLE BOARDING SCHOOL
318	CHNLE	CHINLE,AZ
319	CHNLH	CHINLE HS
320	CHNLJ	CHINLE JHS,CHINLE
321	CHOLA	CHOLLA APARTMENTS

Downright Bad Data Techniques

- Min/Max/Avg/Count/Sum,etc. (Aggregate)
- Data Visualization
- Numeric, String, Date functions
- Subselects, other SQL syntax, etc



```
SELECT AL1.AFFILIATE_ID,
       AL1.FACILITY_CODE,
       AL1.ROOM,
       AL1.PRIMARY_ROOM_FLAG
FROM   FACILITY_ROOM_EMPLOYEE AL1
WHERE  AL1.PRIMARY_ROOM_FLAG = 'Y'
GROUP BY AL1.AFFILIATE_ID

HAVING COUNT (AL1.AFFILIATE_ID) > 1
ORDER BY AFFILIATE_ID
```

Or is the problem definitional?

Data Cleansing Techniques

AKA - Data Hygiene

- Focus on high-payoff data elements
- Interrogate data elements individually and collectively
- Standardization on national codes
- Conduct data audit for conformity of domain
- Document transformation rules and test
- Go back to the source if necessary

Case For Data Warehousing

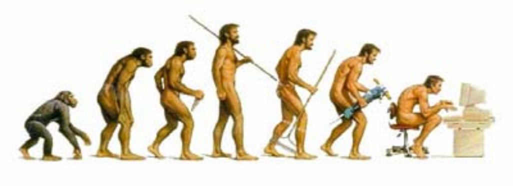
- Perform queries/reporting on servers not used by transaction processing system
- Use data models or server technologies that speed up query and reporting not appropriate for transaction processing
- Having an environment where you don't need a programmer to get your information
- Having an environment that contain history or where you can generate data "as of"
- To provide data that is "cleaned up"

Case Against Data Warehousing

- Historical data often has limited value
- Data warehouses complicate the business processes
- All the data you may need is in your operational system
- Data warehouses required a great deal of maintenance which many can't support
- Might not "take" in the user community
- Cost might be too great

Data Warehouse Evolution

- Access to Data
- Reporting (Query Tools)
- Analysis (OLAP)
- "Operationalize" (Applications)
- Prediction (Data Mining)



Warehousing Biggest Obstacles

- Political Issues
- Poorly defined goals
- Lack of resources
- Technical limits
- Poor understanding of legacy data
- Poor data quality
- Lack of end-user support

ARIZONA STATE UNIVERSITY

IR Challenges and Next Steps?

- Learn more about your Institution’s DW or start discussion
- Learn to Leverage your DW and get data you need (have a a section for you!!)
- Become More Technical (modeling/DBA,etc.)
- Educate Yourself on DW (books/Web/blogs)
- Visit a couple of organizations that have had warehousing systems in production

- http://dheise.andrews.edu/dw/DWData.htm

ARIZONA STATE UNIVERSITY

Questions