Data Cleaning

June 29, 2023

Nancy Rausch

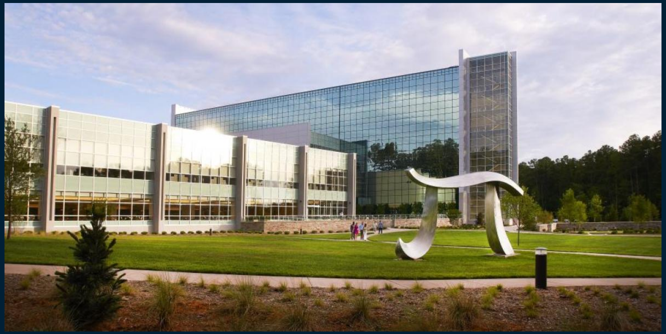
Director, Research and Development, SAS Institute

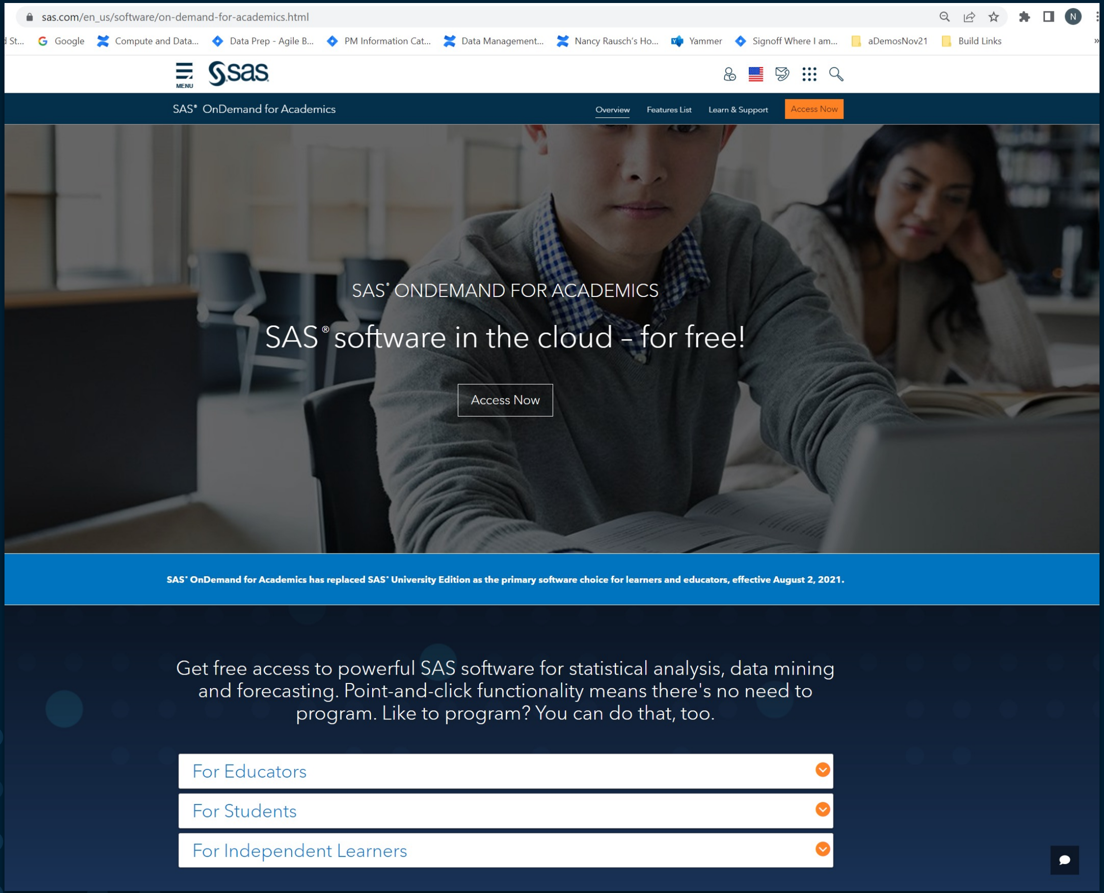
About me…

- I am a Director of Research and Development and a Data Scientist at SAS Institute in Cary, North Carolina.
- I am the Chairperson of the Linux Foundation AI & Data Technology Advisory Council.
- Industry Advisor to the NSF LASER Institute for Learning Analytics
- BS in Electrical Engineering, MS in Computer Engineering & Statistics, MS in Data Analytics
- Areas of Specialization: Data Management, Data Governance, Data Quality, AI for Data Management, Renewable Energy Forecasting, Health Care analytics, Social Science Analytics

About SAS Institute

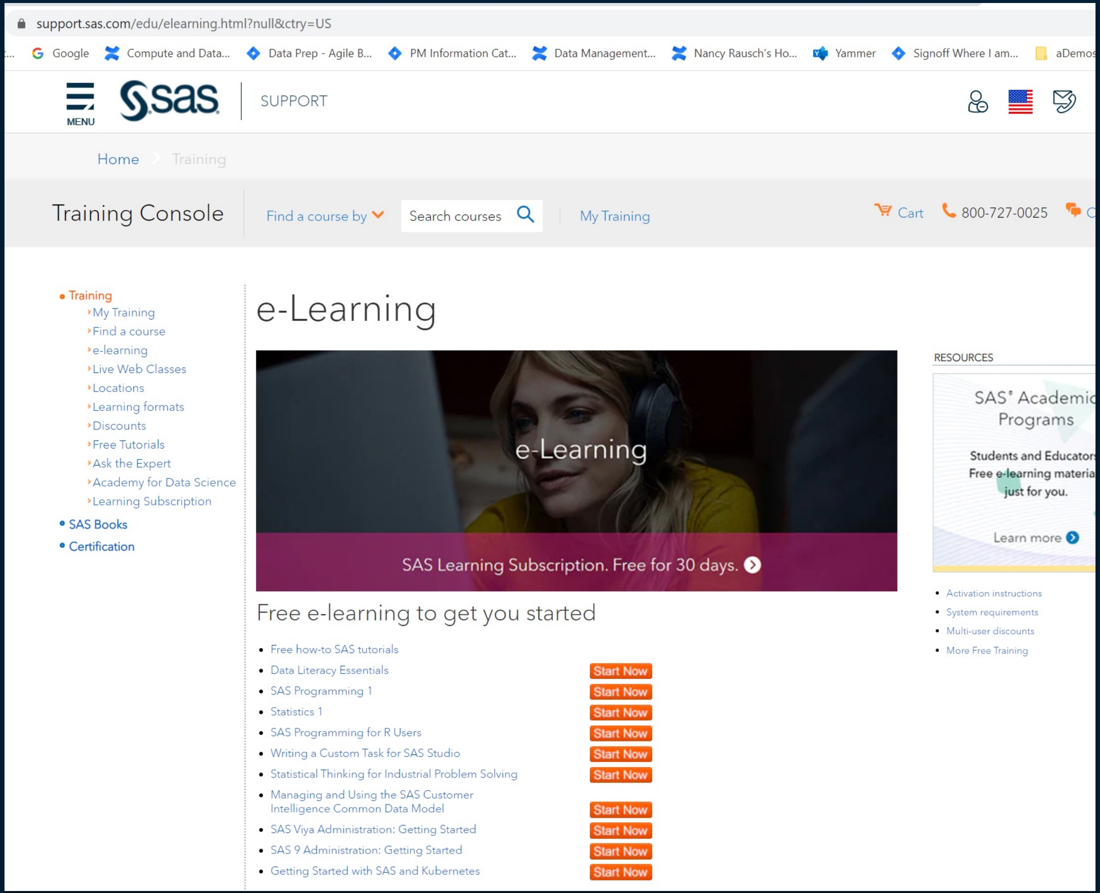
- Statistical software company
  - ~$4B in revenue
  - Worldwide sales
  - Employs about 10,000 people world-wide
  - Engineers, Data Scientists, Machine Learning Developers, Statisticians, Business Analysts, Sales, Marketing, many other roles
- We develop applications, solutions and a statistical language, also called *SAS,* to perform analytics
- We use statistics and business analytics to extract knowledge from data





Students and Educators have free access to SAS Learning Edition

Free online courses in statistics and analytics



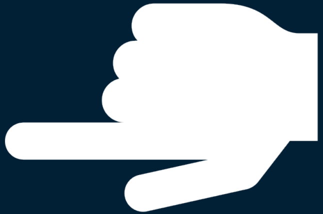
Data Cleaning for Data Quality

Poll: In a modeling or reporting project, about how much time do people spend cleaning and preparing the data?

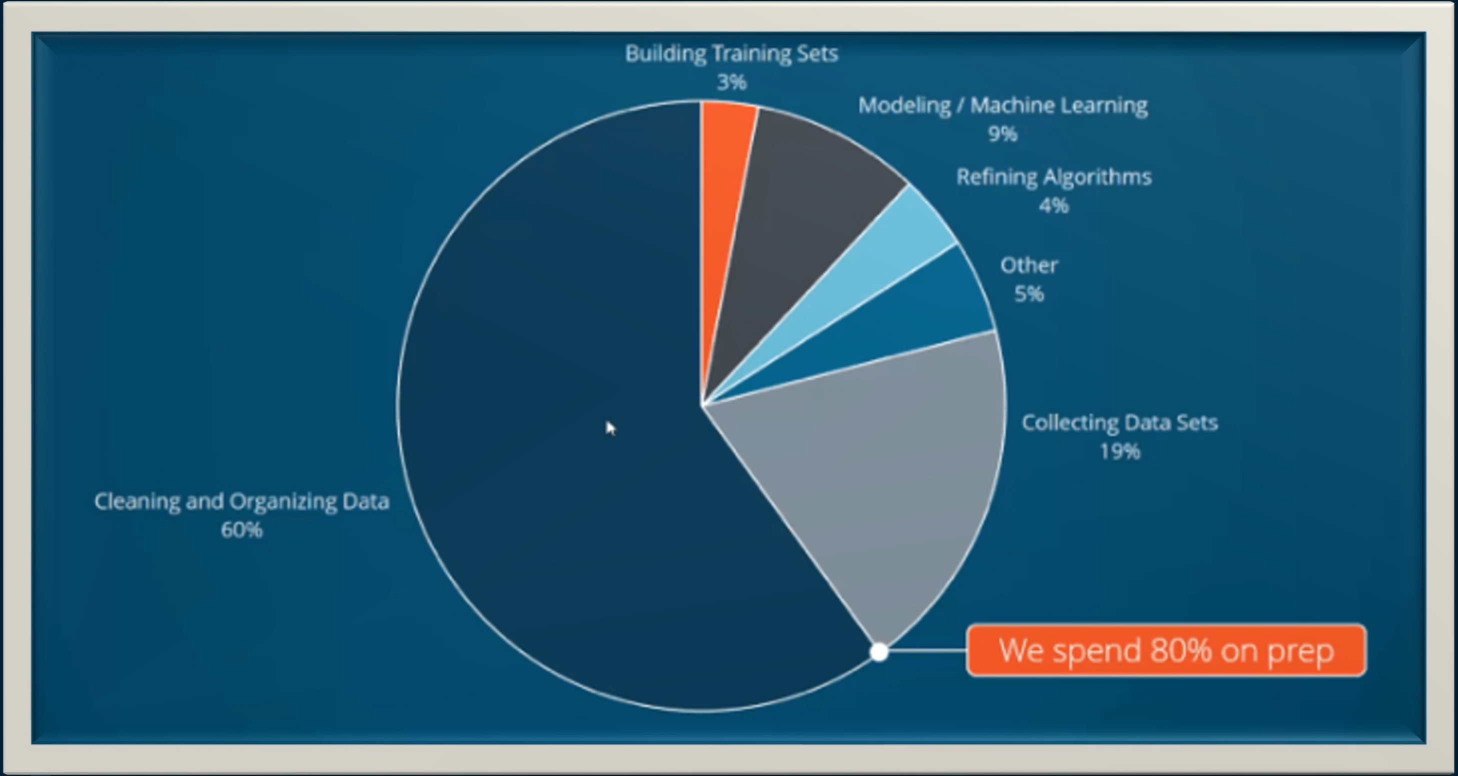
1. 20%
2. 50%
3. 80%

Answer: In a modeling or reporting project, about how much time do people spend cleaning and preparing the data? 80%

1. 20%
2. 50%
3. 80%



Preparing data is hard!



Agenda

1. Why data quality?
2. Data cleaning techniques: Dimensions of data quality
3. Design considerations for analytical and reporting use cases
4. Advanced topics
5. Wrap up and summary

Why Data Quality?

Common causes of poor Data Quality

- Human error
- Algorithmic error
- Misinterpretation error

Simpson's Paradox

UC Berkeley gender bias study: Is this bias?

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

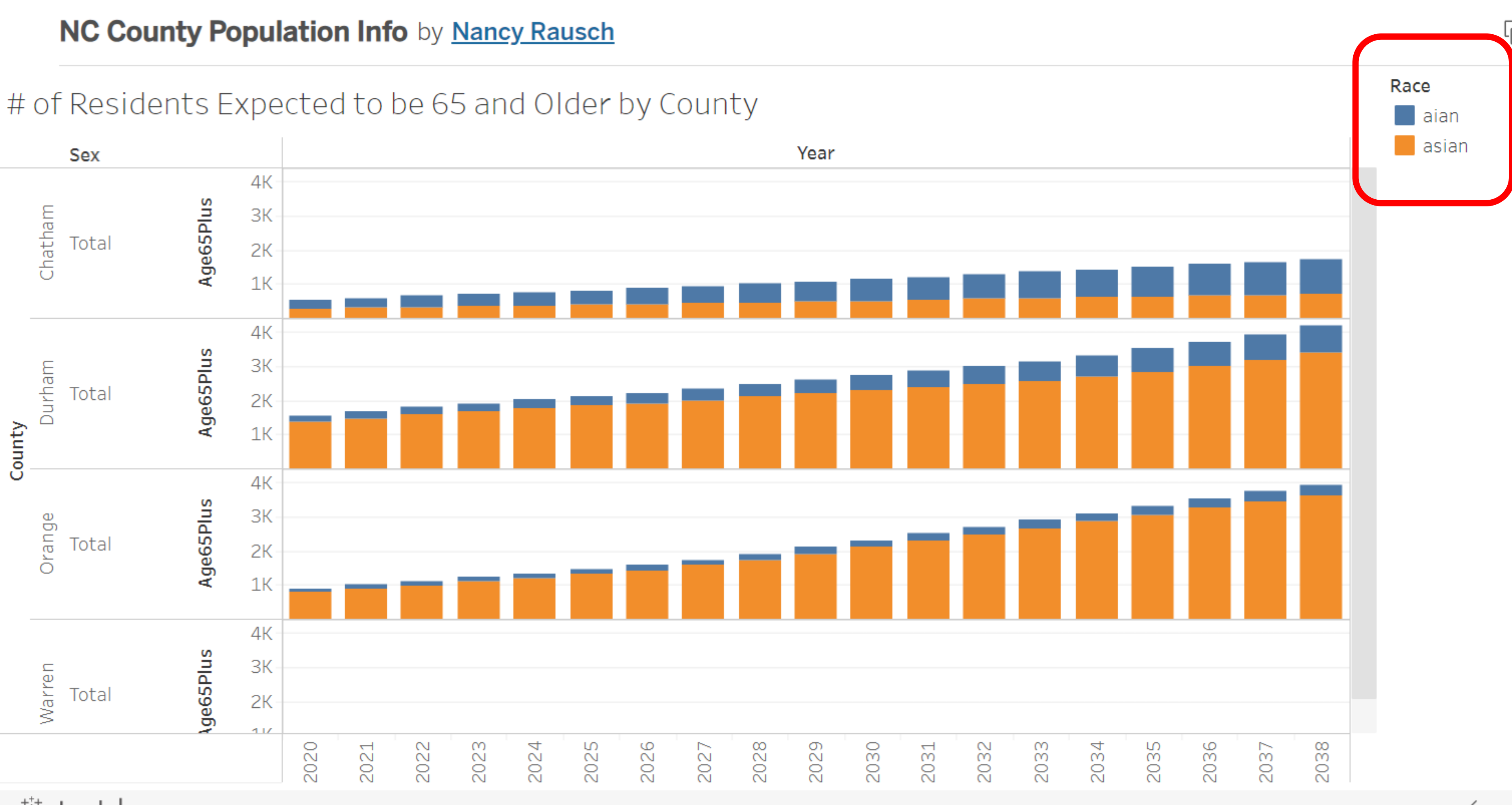
Simpson's Paradox

Not if you look closer….

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

- greater percentage of successful applicants than the other gender
- greater number of applicants than the other gender


NC County Population Info by Nancy Rausch
# of Residents Expected to be 65 and Older by County
Race
aian
asian
Sex
Year
County
Chatham
Durham
Orange
Warren
Total
Age65Plus
4K
3K
2K
1K
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038

More examples of common data quality issues..

- Missing a time period in a timeseries dataset
- Missing data in a row, data misplaced into some other cell in a row
- Invalid values
- Highly correlated variables
- Mismatched ratio of features to rows
- Difference in scale
- Using scale to represent nominal variables
- Unary or Binary variables will low information value
- Incompatible distributions or violation of assumptions for the model being used

Data cleaning techniques: Dimensions of data quality

Dimensions

6 characteristics of clean data

1. Completeness
2. Uniqueness
3. Accuracy
4. Validity
5. Consistency
6. Timeliness

1. Completeness

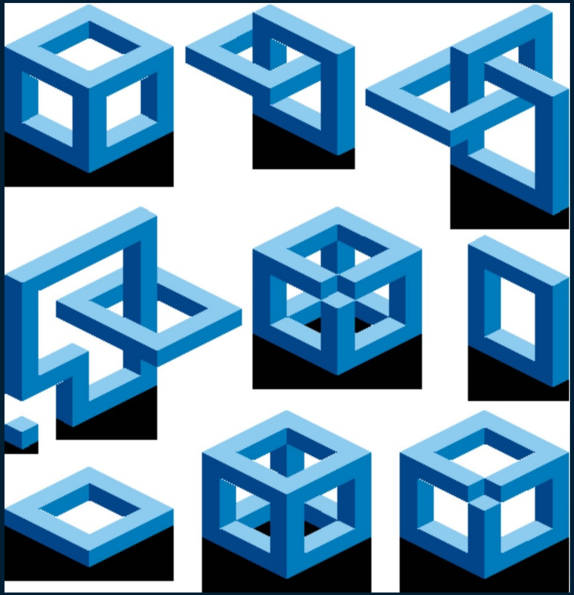
1. Completeness

Definition:

- The degree to which all required data is known.

Quality issues in this dimension:

- Missing or truncated data

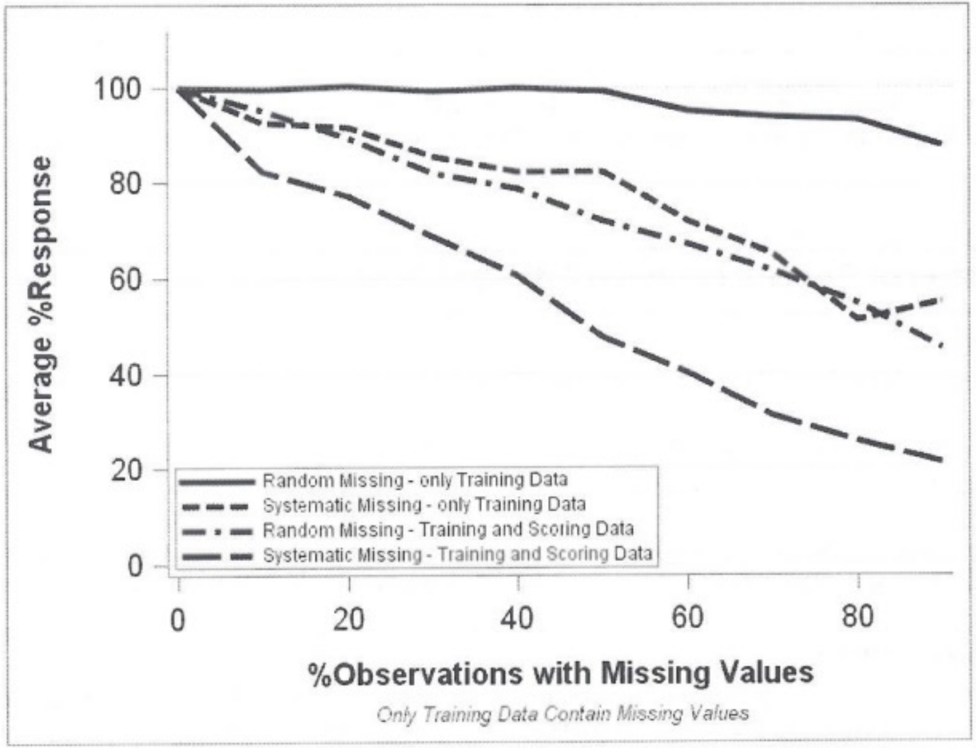


Example

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Overall AQI Value	Main Pollutant	Site Name (of Overall AQI)	Site ID (of Overall AQI)	Source (of Overall AQI)	CO	Ozone	SO2	PM10	PM25
2	1/1/2008	38	PM2.5	Millbrook School	37-183-0014	AQS	17	36	6	9	38
3	1/2/2008	28	Ozone	Millbrook School	37-183-0014	AQS	8	28	9	.	.
4	1/3/2008		PM2.5		37-183-0014	AQS	15	23	23	.	42
5	1/4/2008	75	PM2.5	Millbrook School	37-183-0014	AQS	22	20	10		75
6	1/5/2008	70	PM2.5	Millbrook School	37-183-0014	AQS	23	24	21	19	70
7	1/6/2008	51	PM2.5	Millbrook School		AQS	20	.	.	14	51
8	1/7/2008		PM2.5	Millbrook School	37-183-0014	AQS	25	28	.	15	58
9	1/8/2008	43	PM2.5	Millbrook School	37-183-0014	AQS	25	22	1	11	43
10	1/9/2008	34	PM2.5	Millbrook School	37-183-0014	AQS	18	25	4	10	34
11	1/10/2008	40	PM2.5	Finley Farm	37-183-0020	AQS	13	22	4	11	40

Impact

Figure 18.6: Line plot for the average %Response for different percentages for all the different scenarios



Correction methods

Check the variables importance

- Are there other variables that have similar information value and are better quality?
- Is the variable highly correlated to some other variable?

Remove variables that have lower information value, choose those that are less complete first

SAS Data Mining makes this easy: SAS provides packaged solutions to identify and automatically remove poor quality or redundant variables

Correction methods

SAS/STAT User's Guide

The MI Procedure

1. Impute a new value
   - SAS offers many methods for imputing values; can be calculated, adjusted to the mean, regression analysis predicted, generative data
   - Tip: check the distribution and performance of the data after imputation

2. Calculate missing value from other values
   - Examples:
     - For date-of-birth is provided, calculate age
     - Use data quality standardization techniques to impute the value

Correction methods

SAS/STAT User's Guide

The MI Procedure

1. Impute a new value
   - SAS offers many methods for imputing values; can be calculated, adjusted to the mean, regression analysis predicted, generative data
   - Tip: check the distribution and performance of the data after imputation

2. Calculate missing value from other values
   - Examples:
     - For date-of-birth is provided, calculate age
     - Use data quality standardization techniques to impute the value

	ADDRESS	StandardizedAddress
444	4861 ROSALIA DRIVE	4861 Rosalia Dr
445	2001 CONSTANCE STREET	2001 Constance St
446	300 COLONIAL CLUB DRIVE	300 Colonial Club Dr
447	2323 S. GALVEZ ST	2323 S Galvez St
448	1020 N PRIEUR STREET	1020 N Prieur St
449	2035 TOLEDANO STREET	2035 Toledano St
450	2437 JENA STREET	2437 Jena St
451	1415 TECHE STREET	1415 Teche St
452	315 CIVIC DRIVE	315 Civic Dr
453	215 BETZ PLACE	215 Betz Pl
454	75 E CHALMETTE CIRCLE	75 E Chalmette Cir
455	5701 VETERAN'S MEMORIAL BLVD.	5701 Veteran'S Memorial Blvd
456	726 JOHN HILL TAYLOR DR	726 John Hl Taylor Dr
457	1207 EAST BROADWAY	1207 E Broadway
458	170 E GAP HILL RD	170 E Gap Hl Rd
459	200 ENTERPRISE DRIVE	200 Enterprise Dr
460	RT 4 BOX 90	RR 4 PO Box 90
461	170 W A JENKINS ROAD	170 W A Jenkins Rd
462	357 WEST ARCH ST	357 W Arch St
463	10362 ST RTE 138	10362 St RR 138
464	MAIN CROSS ST BOX 69	Main Cross St PO Box 69

Tips

- Incomplete data is not necessarily missing data!
  - A valid value may not exist
  - Missing data may be useful: Example: *Does not apply* or *Unchecked*
    - Replacing with 0 is a good indicator for these situations
- Always remember to check correlation between variables and variable importance after adjustment
  - It can save you a lot of time if two variables are highly correlated. You don't need to correct data quality problems for every variable!
- Consider multiple imputation; produces less biased estimates

SAS/STAT User's Guide

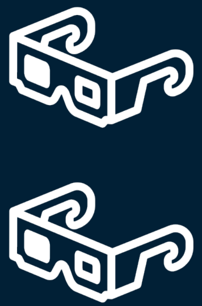
The MI Procedure

2. Uniqueness

2. Uniqueness

Definition:

- Data is unique, with only one instance of data values in the expected records.
- High uniqueness is a good indicator that you can trust the data.



Quality issues in this dimension:

- Duplicate values in expected records

Example

Simple use case

	A	B	C	D
1	Date	Overall AQI Value	Main Pollutant	Site Name (of Overall AQI)
2	1/1/2008	38	PM2.5	Millbrook School
3	1/2/2008	28	Ozone	Millbrook School
4	1/3/2008	42	PM2.5	Millbrook School
5	1/3/2008	42	PM2.5	Millbrook School
6	1/4/2008	75	PM2.5	Millbrook School
7	1/5/2008	70	PM2.5	Millbrook School
8	1/6/2008	51	PM2.5	Millbrook School
9	1/7/2008	58	PM2.5	Millbrook School

Example

Duplicate values across multiple variables

name	address
Kathy Woods	789 Belle Ln
Susan Woodward	152 Blackberry Ln
Sue Woodward	152 Blackberry Lane
Susan Woodward	152 Blackberry Ln
Donny Williams	1034 Skyview Rd
Donald Williams	1034 Skyview Rd
Don Williams	1034 Skyview Road
Colin Ware	1324 S Buchanan St
James Brigs	1507 Bear Springs Rd
Jim Briggs	1507 Bear Springs Rd
James Briggs	1507 Bear Springs Road
James Briggs	1507 Bear Springs Rd
April Lasser	5367 Rustic Elk Limits
April Lasser	5367 Rustic Elk Limits
David Lester	2910 Weisman Rd

Correction methods

Single record duplicates

Detect and remove the duplicate records

Python

```
>>> df.drop_duplicates()
    brand style  rating
0  Yum Yum   cup     4.0
2  Indomie   cup     3.5
3  Indomie  pack    15.0
4  Indomie  pack     5.0
```

To remove duplicates on specific column(s), use `subset`.

```
>>> df.drop_duplicates(subset=['brand'])
    brand style  rating
0  Yum Yum   cup     4.0
2  Indomie   cup     3.5
```

SAS

```
proc sort data=original_data out=no_dups_data nodupkey;
    by _all_;
run;
```

Impact

Multiple record duplicates

Data combined from different silos can cause the same analysis subjects to occur more than once in the database.

Integrating these records can be challenging; different databases may not have the same identifier for the records.

Standardization and records matching on significant, distinguishing attributes of the data such as names, phone numbers, or bank account numbers are helpful .

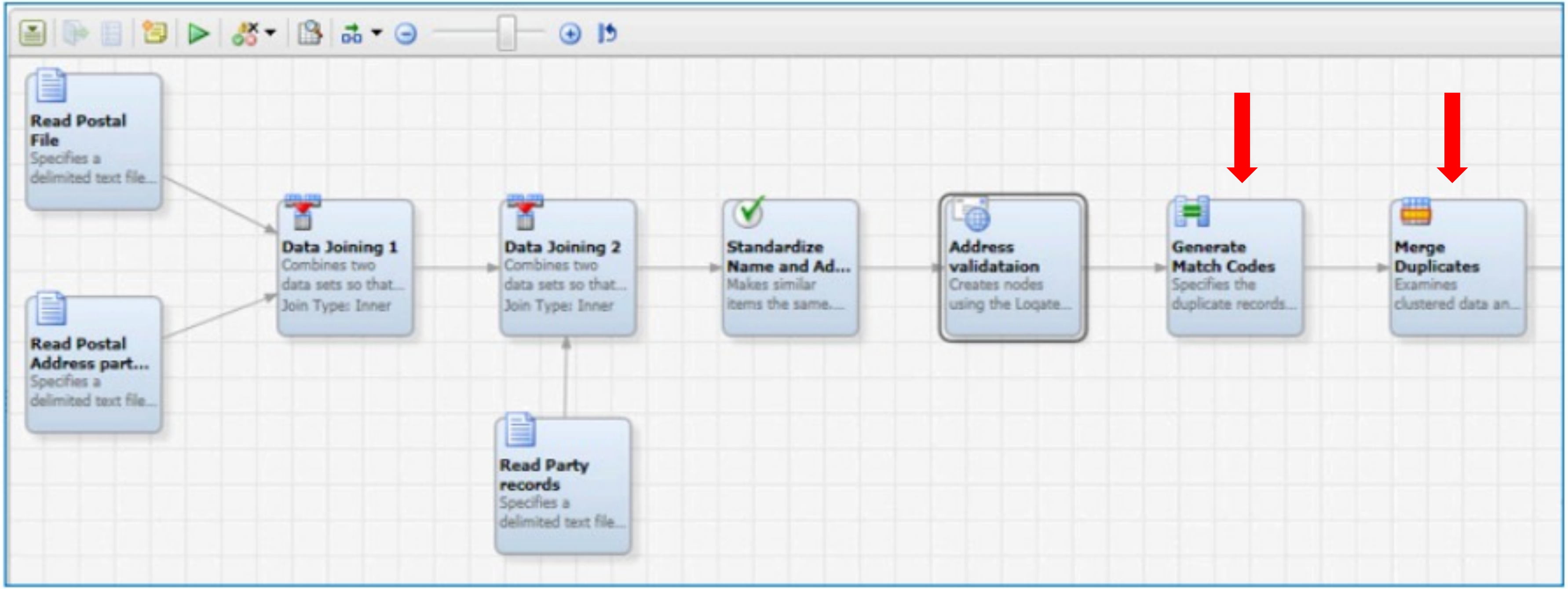


Figure 3. Join, standardize and remove duplicates from your data with the powerful and intuitive SAS Data Quality user interface.

Correction Techniques

Use clustering techniques – SAS Example

#	ID	COMPANY	CONTACT	CLUSTER	MATCH_CODE
3025	215	Allied Signal Inc.	Rich Temple	10	$$$$$$$$$$$$$$$!&W~4FPW$$$$$...
3026	219	Allied Signal Inc.	Todd Kotte	10	$$$$$$$$$$$$$$$!&W~4FPW$$$$$...
3027	483	Salt River Project	Susan Bradshaw	11	ZK00LYP$$$$$$$$!4W~YVYNYC$$$...
3028	486	Salt River Project	John Reiss	11	ZK00LYP$$$$$$$$!4W~YVYNYC$$$...
3029	490	San Diego County	Curt Delarosa	12	$$$$$$$5HZ$$$$$!4P8F3P~$$$$$$...
3030	488	San Diego County	Benjamin Mccorkle	12	$$$$$$$5HZ$$$$$!4P8F3P~$$$$$$...
3031	03	Jaxson Data Corporation	DONALD WILLIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3032	04	The Jackson Data Corp.	DONALD F. WILLIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3033	02	Jackson Data Co.	DON WILLIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3034	06	Jackson Data Inc.	DONNY WILLIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3035	07	The Jacksen Data Co...	DON WILIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3036	05	Jackson Data	MR DON F WILLIAMS	13	H56I2CL$$$$$$$$!CXP8~$$$$$$$$...
3037	735	Northrop Corporation	Anthony Lutzker	14	D0Z$CGY$$$$$$$$!PY~2YN$$$$$$$...
3038	737	Northrop Corporation	W. Binns	14	D0Z$CGY$$$$$$$$!PY~2YN$$$$$$$...
3039	495	Glendale Advenist Me...	Elijah Mellor	15	ZK00LYP$$$$$$$$!FWP8W8VP4$$$...
3040	496	Glendale Advenist Me...	Debbie Cochrane	15	ZK00LYP$$$$$$$$!FWP8W8VP4$$$...
3041	977	Kings County	Bev Johanson	16	$$$$$$$$$$$$$$$!3PF3P~$$$$$$$...

FINAL_MATCHES ▾

Filter and Sort | Query Builder | Data ▾ | Describe ▾ | Graph ▾ | Analyze ▾ | Export ▾ | Send To ▾

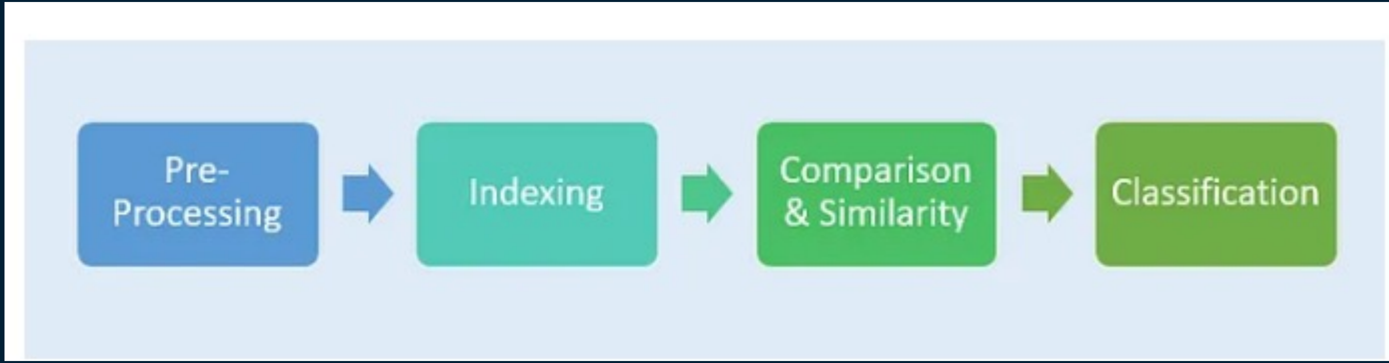
	NAME	ADDRESS	CITY	STATE	ZIP	PHONE	DOB	Gender	Education	Income_Level	Household	CAR_MAKE	CAR_MODEL	CAR_YEAR
1	Abigail Sargent	4211 S Rushford St	Palo Alto	CA	94303	860-952-3496	16APR1965	F	High School gradu...	Less than $25,000	Single, never marr...			.
2	Andre Hulbert	5024 Fairbanks W...	Sunnyvale	CA	94089	618-121-6649	30SEP1971	M	Bachelor's degree	$100,000 to $149,...	Separated			.
3	Ashley Bey	P.O. Box 6239	N. Ridgeville	OH	44039	271-476-5054	14MAR1972	F	Bachelor's degree	$100,000 to $149,...	Single, never marr...	BMW	M2	2014
4	Cathy Lapp	4400 NC Highway...	ST. LOUIS	MO	63134	718-922-0353	28JAN1982	F	Doctorate degree	$75,000 to $99,999	Widowed	Acura	RDX	2009
5	Cindy Prentiss	515 E. Broad St.,...	St Louis	MO	63146	949-161-1908	26JUN1974	F	Doctorate degree	$50,000 to $74,999	Divorced	Mercedes-Benz	C300	2007
6	David Grassi	824 Valerie Dr Uni...	Pomona	CA	91768	806-296-2544	20OCT1985	M	Bachelor's degree	Less than $25,000	Single, never marr...	Toyota	Camry	2012
7	Denise Nath	1215 N Caldwell St	New Haven	CT	06516	660-469-0704	12NOV1977	F	Some high school,...	Less than $25,000	Separated			.
8	Douglas Doty	406 Mcclure Cir	Minneapolis	MN	55431	850-323-7265	14FEB1970	M	Doctorate degree	$75,000 to $99,999	Divorced	Mercedes-Benz	E350	2014
9	E Fusco	2617 Ramsey Rd...	Herndon	VA	22070	500-881-3331	08JUN1983	M	Bachelor's degree	$75,000 to $99,999	Single, never marr...	Toyota	Tundra	2013
10	Eric Einhorn	23 S Saunders Rd	Hutchins	KS	67504-5282	530-406-2382	12FEB1973	M	Bachelor's degree	$75,000 to $99,999	Single, never marr...	Hyundai	Tuscon	2008
11	Gary Stratman	5153 Camino Ruiz	Sunnyvale	CA	94086	360-964-1700	12FEB1974	M	Bachelor's degree	$100,000 to $149,...	Married or domest...			.
12	Gwen Story	2120 Raven Glass...	PHILADELPHIA	PA	19178-4955	284-565-7463	17OCT1976	F	Some high school,...	$25,000 to $34,999	Married or domest...			.
13	Jessica Macias	5230 Walnut Grov...	Sherman	TX	75090-4440	738-818-2196	20JUN1977	F	Bachelor's degree	$100,000 to $149,...	Widowed	Acura	MDX	2010
14	Johnathon Soon	239 N Edgeworth...	St Louis	MO	63104	615-005-6993	17JUN1976	F	Master's degree	$100,000 to $149,...	Married or domest...	Hyundai	Santa Fe	2013
15	Jonathan Mc Gee	4900 Rivergrade...	Bannockburn	IL	60015	843-042-5960	19OCT1974	F	Bachelor's degree	$150,000 to $199,...	Married or domest...	Honda	Accord	2015
16	Jose Rochford	2800 Woodlawn D...	Ballwin	MO	63011	721-144-7436	28OCT1970	M	High School gradu...	$150,000 to $199,...	Divorced			.
17	K Sekeres	5541 Central Ave...	Marlow	NH	03456	991-312-5527	06AUG1971	F	Master's degree	$50,000 to $74,999	Widowed	Acura	RDX	2009
18	Kristina Radley	4430 E Greensbor...	Kansas City	MO	64141 6267	979-842-2568	18NOV1979	F	Doctorate degree	$200,000 or more	Married or domest...	Acura	TLX	2015
19	Lou Voss	777 S. Harbor Blvd.	Bellevue	WA	98006-1800	357-989-4735	02FEB1967	M	Some high school,...	$75,000 to $99,999	Widowed	Hyundai	Elantra	2010
20	Margaret Muench	3001 W Mission R...	Dallas	TX	75247	883-271-9095	15AUG1969	F	Doctorate degree	$75,000 to $99,999	Separated			.
21	Michelle Wan	4000 East Sky Ha...	Chesterfield	MO	63005	426-758-4180	06MAR1974	F	Some high school,...	$75,000 to $99,999	Single, never marr...			.
22	Miranda Andre	510 LightHouse A...	Cupertino	CA	95014	620-959-5034	13JAN1976	F	Doctorate degree	$200,000 or more	Divorced			.
23	Ms. Shannon Mazo	1515 Lord Ashley...	St. Louis	MO	63128	966-586-4147	04SEP1981	F	Bachelor's degree	Less than $25,000	Single, never marr...	[illegible]	[illegible]	2006
24	Pat Pietron	4305 Central Ave...	Orem	UT	84058	456-854-5248	10JUL1977	M	Some high school,...	$75,000 to $99,999	Divorced	[illegible]	[illegible]	2012
25	Philip Gerstle	PO Box 13507	Skokie	IL	60076-2999	360-681-2934	09JUL1967	M	Master's degree	$150,000 to $199,...	Divorced	[illegible]	[illegible]	2011
26	Rob Rubenstein	3939 Ruffin Rd	St Louis	MO	63021	297-073-4204	14OCT1968	M	Some high school,...	$35,000 to $49,999	Divorced	[illegible]	[illegible]	2013
27	Samuel Harleen	102 Echo Glen Dr	St. Charles	MO	63301	731-887-4557	29JUL1992	M	Master's degree	$200,000 or more	Married or domest...	[illegible]	[illegible]	2011
28	Sean Nugent	1993 Emsford Dr	St. Louis	MO	63114	495-764-8564	09NOV1982	M	High School gradu...	$25,000 to $34,999	Divorced	[illegible]	[illegible]	2011
29	Sharon Mandelba...	3540 Wilshire Blvd	Redwood Shores	CA	94065	902-861-5137	28SEP1960	F	Bachelor's degree	$200,000 or more	Married or domest...	[illegible]	[illegible]	2005
30	Sidney Trotter	161 Northfork Rd	Lima	OH	45801	419-566-4321	15NOV1972	F	Master's degree	$50,000 to $74,999	Widowed	[illegible]	[illegible]	2009

All unique

Figure 12. Final Scoring Results

Correction Techniques

Record Linkage in Python



3. Accuracy & 4. Validity

3. Accuracy & 4. Validity

Definition:

- How well the data *verifiably* represents the real-world scenario.

Quality issues in this dimension:

- Inaccurate or invalid data

Very important in some domains, such as health care or finance.

Examples

- Inaccurate or invalid phone numbers:
  - XX-335-32 vs. XXX-335-3232 (too short!)
  - XXX-385-3232
- Invalid birth details
- Invalid credit card charges

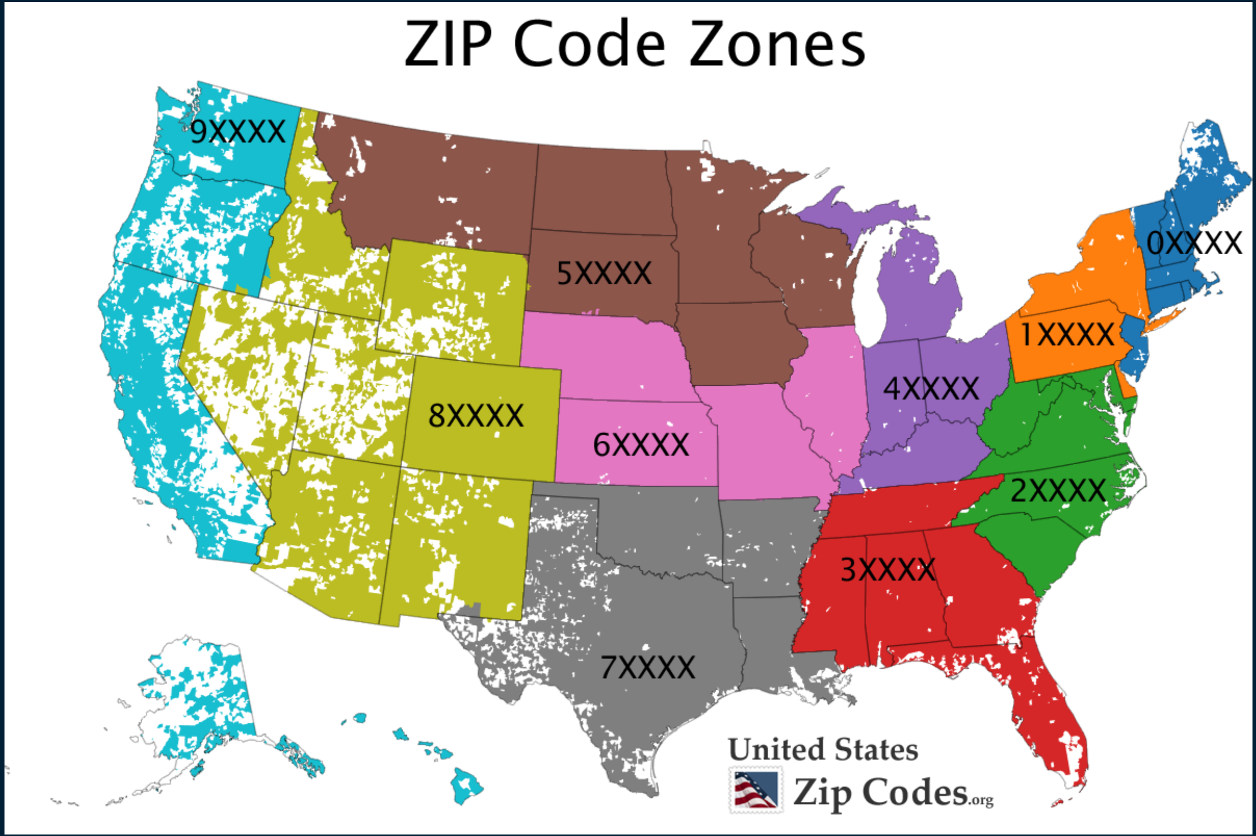
Tip: Data in this dimension frequently requires verification using domain expertise.

- Examples:
  - a certificate from a bank
  - a medical coding expert

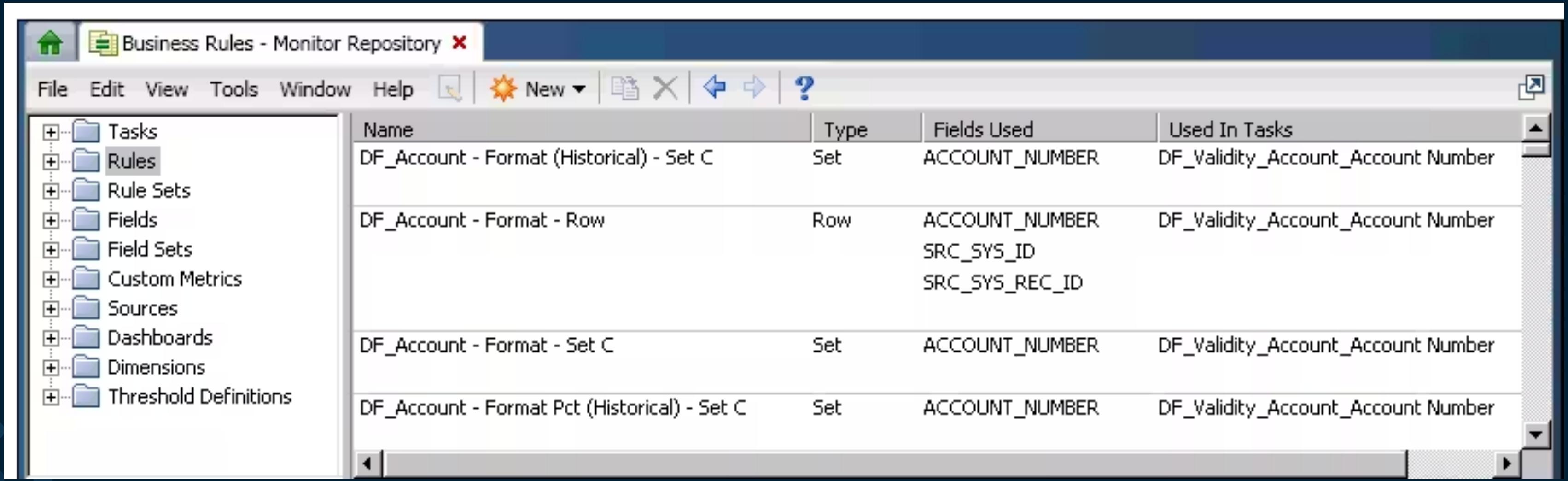
Handling Techniques – Apply Business Rules

Business Rules are frequently used to determine Validity and Accuracy, when data requires domain expertise to detect and correct

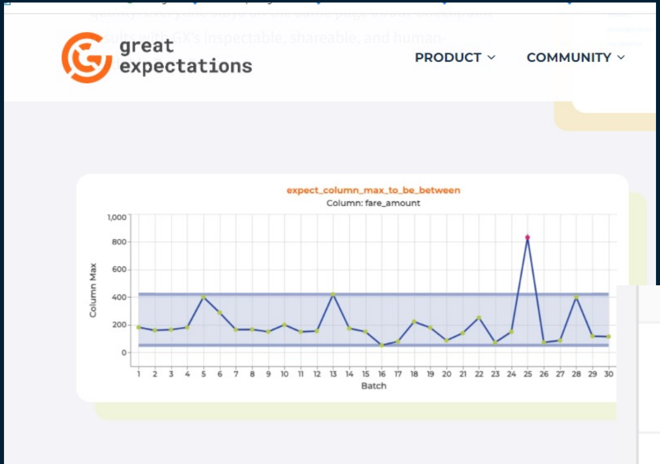
Handling techniques – Use known data (lookup tables)



SAS Business Rules Manager



Python – Great Expectations



- **expect_batch_row_count_to_match_prophet_date_model** (Contrib BatchExpectation)
- **expect_column_average_lat_lon_pairwise_distance_to_be_less_than** (Contrib ColumnAggregateExpectation)
- **expect_column_average_to_be_within_range_of_given_point** (Contrib ColumnAggregateExpectation)
- **expect_column_discrete_entropy_to_be_between** (Contrib ColumnAggregateExpectation)
- **expect_column_distinct_values_to_be_continuous** (Contrib ColumnAggregateExpectation)

5. Consistency

5. Consistency

Definition:

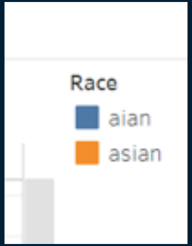
- Data is stored in a similar way across dimensions and records.

Quality issues in this dimension:

- formatting problems
- data stored at different levels of summarization
- data mismatch

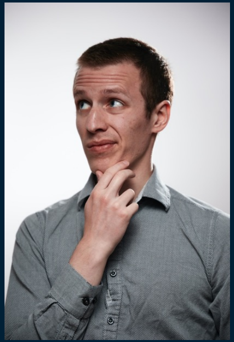
Examples

Data errors



Inconsistent calculations

- Monthly revenue: $70
- Monthly cost: $10
- -----------
- Monthly profit: <u>$567 </u> - ??



Handling Techniques

- Use database constraints:
  - Example: If I have a PERSON I must also have an ADDRESS
- Use calculated columns
  - Example: Calculate AGE instead given a birth date
- Create Business Rules
  - Example: If VALUE > 500 then RAISE ERROR
- Use Anomaly detection
  - SAS supports multiple techniques
  - Regression analysis

SAS Anomaly Detection

Outliers of Amount Paid

Are There Outlier Values of Amount Paid?

There are 64 outlier values of Amount Paid. These outliers change the overall sum and average by more than 5%.

What Are the Details of These Outliers?

Amount Paid	Product Name	Product SKU	Product Price	Product Category	Purchases Amount
$1,887.31	puffer jacket	49164579	$1,887.31	Men Leisure	$11,363
$1,838.80	puffer jacket	49164579	$1,887.31	Men Leisure	$9,290
$1,736.33	puffer jacket	49164579	$1,736.33	Men Leisure	$8,798
$1,736.33	puffer vest	3941183	$1,736.33	Men Leisure	.
$1,691.70	puffer jacket	49164579	$1,736.33	Men Leisure	$9,258
$1,690.48	puffer vest	3941183	$1,736.33	Men Leisure	$9,258

What Is the Effect of Outliers on Amount Paid?

Metric	Including Outliers	Excluding Outliers	Outlier Impact	Difference
Sum	$855,494.20	$722,440.06	15.55%	133054.1332
Average	$719.51	$642.17	10.75%	77.338368903
Median	$581.95	$572.87	1.56%	9.08265

Python

using PyOD

ID	Name	Reference
abod	Angle-base Outlier Detection	pyod.models.abod.ABOD
cluster	Clustering-Based Local Outlier	pyod.models.cblof.CBLOF
cof	Connectivity-Based Local Outlier	pyod.models.cof.COF
iforest	Isolation Forest	pyod.models.iforest.IForest
histogram	Histogram-based Outlier Detection	pyod.models.hbos.HBOS
knn	K-Nearest Neighbors Detector	pyod.models.knn.KNN
lof	Local Outlier Factor	pyod.models.lof.LOF
svm	One-class SVM detector	pyod.models.ocsvm.OCSVM
pca	Principal Component Analysis	pyod.models.pca.PCA
mcd	Minimum Covariance Determinant	pyod.models.mcd.MCD
sod	Subspace Outlier Detection	pyod.models.sod.SOD
sos	Stochastic Outlier Selection	pyod.models.sos.SOS

Image by Author

6. Timeliness

6. Timeliness

Definition:

- The data is available with the expected timeframe.

Quality errors in this dimension:

- Stale or old data
- Data is not received at the expected pace
- Missing time periods

Examples

- Data is expected quarterly, but it is collected yearly
  - Systems produce data at different times, and they need to be consolidated

- Data is collected at different or fluctuating intervals

al-data-collection

DATA POINT	MACHINE A	MACHINE B	STANDARDIZED DATA
Date	December 27, 2015	12/27/2015	12/27/15
Part Count	part_ct	Part:Count	PartCount
Machine Alarm	estop	Alarm:EStop	EmergencyStop

Impact

1

Inability to tie a problem back to the root cause

2

Difficult to consolidate the data across different systems

3

Impossible to impute missing time periods for values

Handling Techniques

- Clean out old/stale data
- Use date/timestamps in your data
- Standardize on data/time collection periods and apply those across systems
- Use business rules to watch for errors
- Use SAS procedures to detect and correct timeseries errors

Interlude

Question:

Why has Roman concrete survived the centuries?

1. Special ingredients only found in Rome

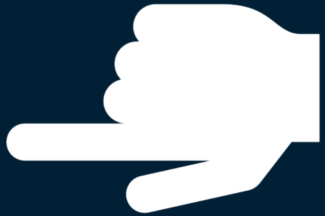
2. Tiny impurities in the concrete

3. Aliens helped

Question:

Why has Roman concrete survived the centuries?

1. Special ingredients only found in Rome
2. Tiny impurities in the concrete
3. Aliens helped



Tiny impurities in Roman concrete make it self-healing

A large-area elemental map (Calcium: red, Silicon: blue, Aluminum: green) of a 2 cm fragment of ancient Roman concrete (right) collected from the archaeological site of Privernum, Italy (left). A calcium-rich lime clast (in red), which is responsible for the unique self-healing properties in this ancient material, is clearly visible in the lower region of the image.

Courtesy of the researchers

Part 2:

Good data quality starts with good design

Tips For Tidy Data Part 1

- Column headers are variable names, not values
  - Every column is a variable
  - Make variable names easy to understand

```
relig_income
#> # A tibble: 18 × 11
#>   religion           `<$10k` $10-2…¹ $20-3…² $30-4
#>   <chr>                <dbl>   <dbl>   <dbl>   <db
#> 1 Agnostic                27      34      60
#> 2 Atheist                 12      27      37
#> 3 Buddhist                27      21      30
#> 4 Catholic               418     617     732     6
#> 5 Don't know/re…          15      14      15
```

- Don't store variables in one column unless you need the data that way
  - Ex: Smith, Peter

- Every row should be an observation

Tips For Tidy Data Part 2

- Use database constraints to ensure data integrity
  - Ex: If you add a NAME, then must have ZIPCODE
- Use a single date column that represents the record date
  - 5 date columns make it difficult to understand which date actually represents the record
- Don’t use data ‘as-is’. Use processes to help prepare it, preferably repeatable
  - Examples: Rescale and Center Numeric values before using them in models
    - 827433.33 vs. 0.8274333 have very different information value!
    - Center around a 0 or a 1 mean

Tips for Tidy Data Part 3

- Have or generate a primary key for each record
  - Don’t rely on row order, because databases may mix them up
  - Sequential values do not work in parallel systems
  - A UUID is better in these cases (for big data)

- ---

- Add descriptive information to the Dataset itself. Use labels, descriptions, and other good information. We all forget!

Recoding variables

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Recode variables where the numeric value has no relevance to the data

Examples to watch out for:

- Increasing dates
- Categories coded as numbers
  - Gender
  - Demographic variables
  - Race

Use Surviving Records

Deduplication by selecting the best record

GroupID	_Frequency_	_Position_	id	name	address	updatedate
23	2	2	105345	Kathy Woods	789 Belle Ln	08JAN2018
24	3	1	11004	Susan Woodward	152 Blackberry Ln	01JAN2018
24	3	2	58786	Sue Woodward	152 Blackberry Lane	13JAN2018
24	3	3	12004	Susan Woodward	152 Blackberry Ln	02JAN2018
25	3	1	66252	Donny Williams	1034 Skyview Rd	16JAN2018
25	3	2	99307	Donald Williams	1034 Skyview Rd	23JAN2018
25	3	3	36247	Don Williams	1034 Skyview Road	09JAN2018
26	1	1	92333	Colin Ware	1324 S Buchanan St	27JAN2018
27	4	1	79521	James Brigs	1507 Bear Springs Rd	02JAN2018
27	4	2	88367	Jim Briggs	1507 Bear Springs Rd	25JAN2018
27	4	3	11345	James Briggs	1507 Bear Springs Road	01JAN2018
27	4	4	16206	James Briggs	1507 Bear Springs Rd	05JAN2018
28	2	1	95948	April Lasser	5367 Rustic Elk Limits	19JAN2018
28	2	2	85948	April Lasser	5367 Rustic Elk Limits	05JAN2018
29	1	1	30245	David Lester	2910 Weisman Rd	02JAN2018

Dataset design Principals

One row-per-subject: Transpose is your friend!

Table 7.2: Content of CUSTOMER table

CustID	Birthdate	Gender
1	16.05.1970	Male
2	19.04.1964	Female

Table 7.3: Content of ACCOUNT table

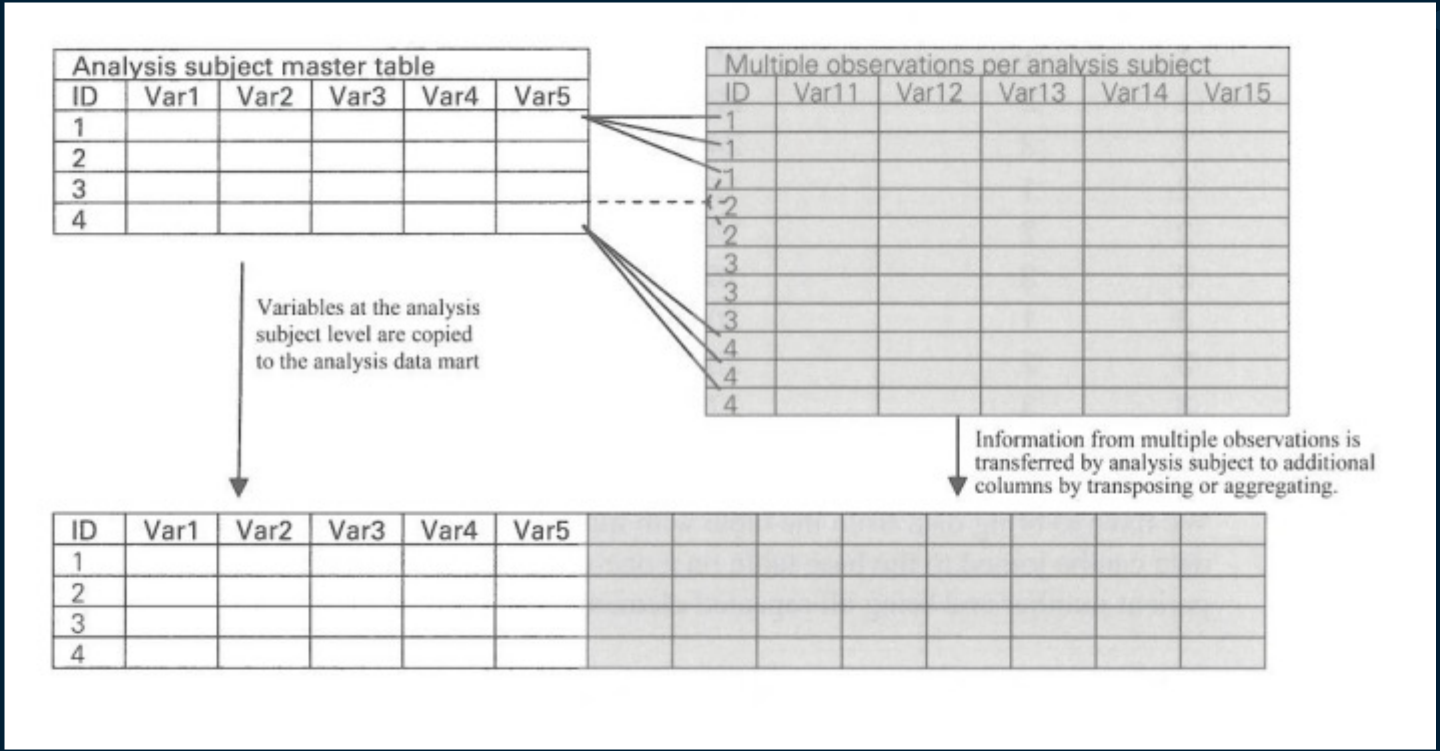
AccountID	CustID	Type	OpenDate
1	1	Checking	05.12.1999
2	1	Savings	12.02.2001
3	2	Savings	01.01.2002
4	2	Checking	20.10.2003
5	2	Savings	30.0[illegible]2004

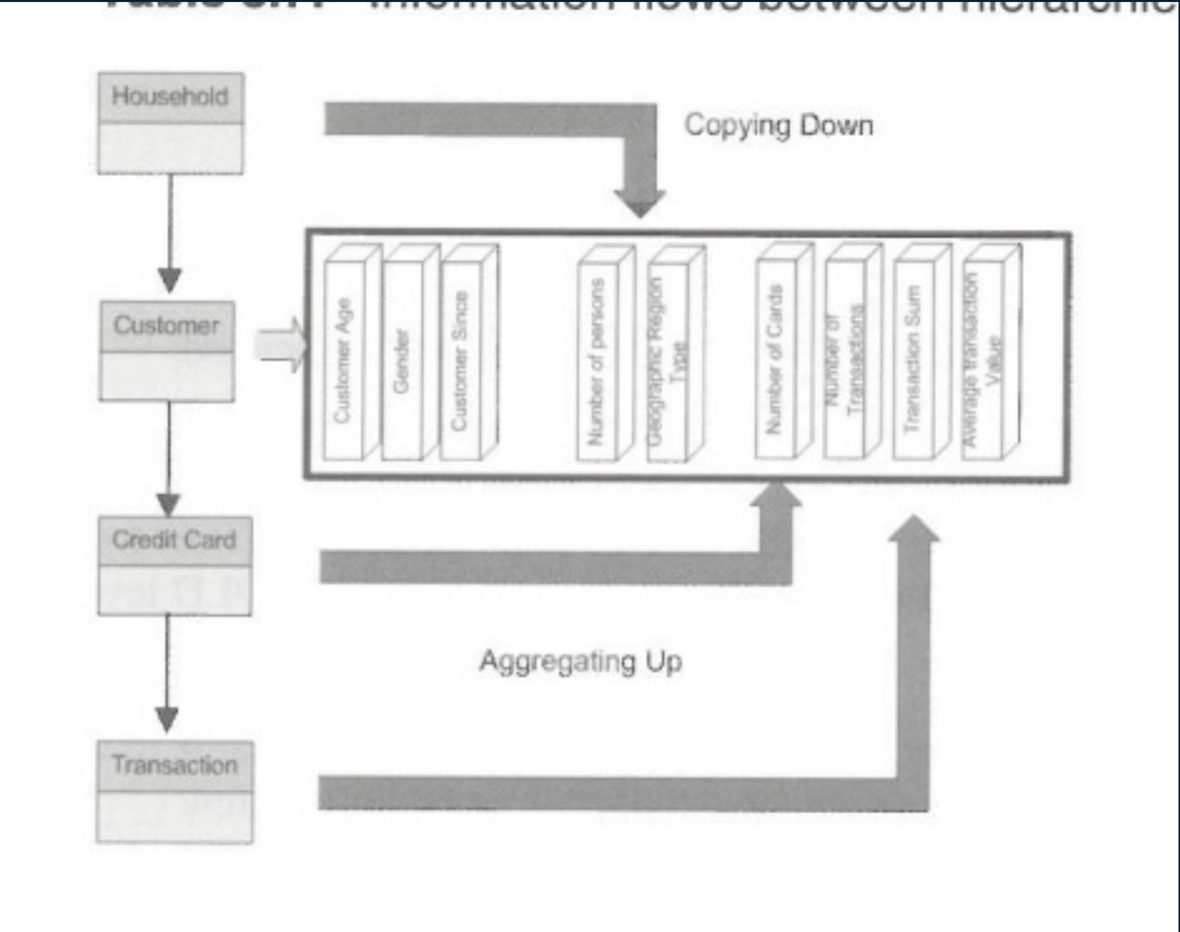
CustID	Birthdate	Gender	Number of Accounts	Proportion of Checking Accounts	Opendate of oldest account
1	16.05.1970	Male	2	50 %	05.12.1999
2	19.04.1964	Female	3	33 %	01.01.2002

One row-per-subject designs

When to use

- Prediction of events like a Campaign
- Time series models for Prediction for single dimensions
- Cluster segmentation: patients, customers, text documents
- Reporting where you want to provide detail and summary report views into the data (just easier this way!)


Analysis subject master table
ID Var1 Var2 Var3 Var4 Var5
Multiple observations per analysis subject
ID Var11 Var12 Var13 Var14 Var15
Variables at the analysis subject level are copied to the analysis data mart
Information from multiple observations is transferred by analysis subject to additional columns by transposing or aggregating.



Aggregation may be required

One row-per-subject designs

Analytical models that do well with this design

- Regression models
- Time series forecasting
- ANOVA
- Survival analysis
- PCA

Dataset design

Multiple rows-per-subject

ACCOUNT tables

CustID	Birthdate	Gender	AccountID	Type	OpenDate
1	16.05.1970	Male	1	Checking	05.12.1999
1	16.05.1970	Male	2	Savings	12.02.2001
2	19.04.1964	Female	3	Savings	01.01.2002
2	19.04.1964	Female	4	Checking	20.10.2003
2	19.04.1964	Female	5	Savings	30.09.2004

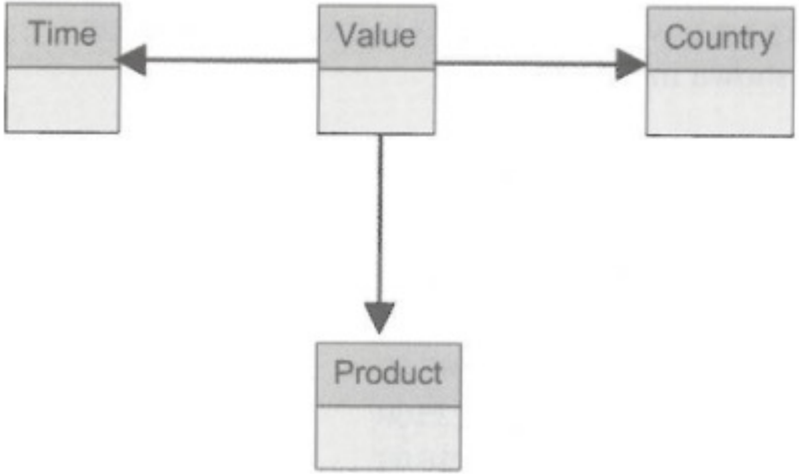
Examples

Table 9.2: Market basket data for two customers with additional data on CUSTOMER level

	CUSTOMER	PRODUCT	Segment
1	213	baguette	SILVER
2	213	hering	SILVER
3	213	avocado	SILVER
4	213	artichok	SILVER
5	213	heineken	SILVER
6	213	chicken	SILVER
7	213	coke	SILVER
8	217	baguette	GOLD
9	217	hering	GOLD
10	217	avocado	GOLD
11	217	artichok	GOLD
12	217	heineken	GOLD
13	217	apples	GOLD
14	217	peppers	GOLD
15	221	soda	SILVER
16	221	olives	SILVER
17	221	bourbon	SILVER
18	221	cracker	SILVER
19	221	heineken	SILVER
20	221	turkey	SILVER
21	221	steak	SILVER

Table 9.5: Web log data with a session sequence variable

	Session Identifier	requested_file	session_sequence
1	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp	1
2	43d0a4da826149b5 2002-02-17 08:38:12	/Cookie_Check.jsp	2
3	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp	3
4	43d0a4da826149b5 2002-02-17 08:38:12	/Corporate_Relations.jsp	4
5	43d0a4da826149b5 2002-02-17 08:38:12	/Retail_Store.jsp	5
6	43d0a4da826149b5 2002-02-17 08:38:12	/Store/Store_Locations.jsp	6
7	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp	1
8	43d639ebce6c73d8 2002-02-17 23:43:16	/Cookie_Check.jsp	2
9	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp	3
10	43d639ebce6c73d8 2002-02-17 23:43:16	/Department.jsp	4



Month	Country	Product	Actual Sales
1993.01	CANADA	BED	$856.00
1993.02	CANADA	BED	$1,581.00
1993.03	CANADA	BED	$1,900.00
1993.01	CANADA	SOFA	$1,953.00
1993.02	CANADA	SOFA	$2,483.00
1993.03	CANADA	SOFA	$2,495.00
1993.01	GERMANY	BED	$1,875.00
1993.02	GERMANY	BED	$1,929.00
1993.03	GERMANY	BED	$1,222.00
1993.01	GERMANY	SOFA	$3,723.00
1993.02	GERMANY	SOFA	$2,393.00
1993.03	GERMANY	SOFA	$2,799.00

Multiple row-per-subject designs

When to use

- Products frequently bought together (aka Market Basket Analysis)
- Association analysis
- Time series analysis with different analysis subjects
- Longitudinal analysis
- Sequence analysis

Summary: Good design saves time

Advanced topics

Catalogs, Lineage, Maintenance, data drift, curation, bias

Who will you call when you need some new data?

- Does this data exist?
- Where is it?
- What is the source of truth of that data?
- Do I have access to it?
- Who is the owner?
- Who are the common users?
- Is there existing work I can re-use?
- Can I trust this data?

Business Challenges

Cope with the digital transformation disruption

Complex Ecosystems

Democratization

Privacy & Ethics

"The two biggest challenges in data management are centered around data catalogs—finding and identifying data that delivers value, and supporting data governance, data privacy and data security." (Gartner)

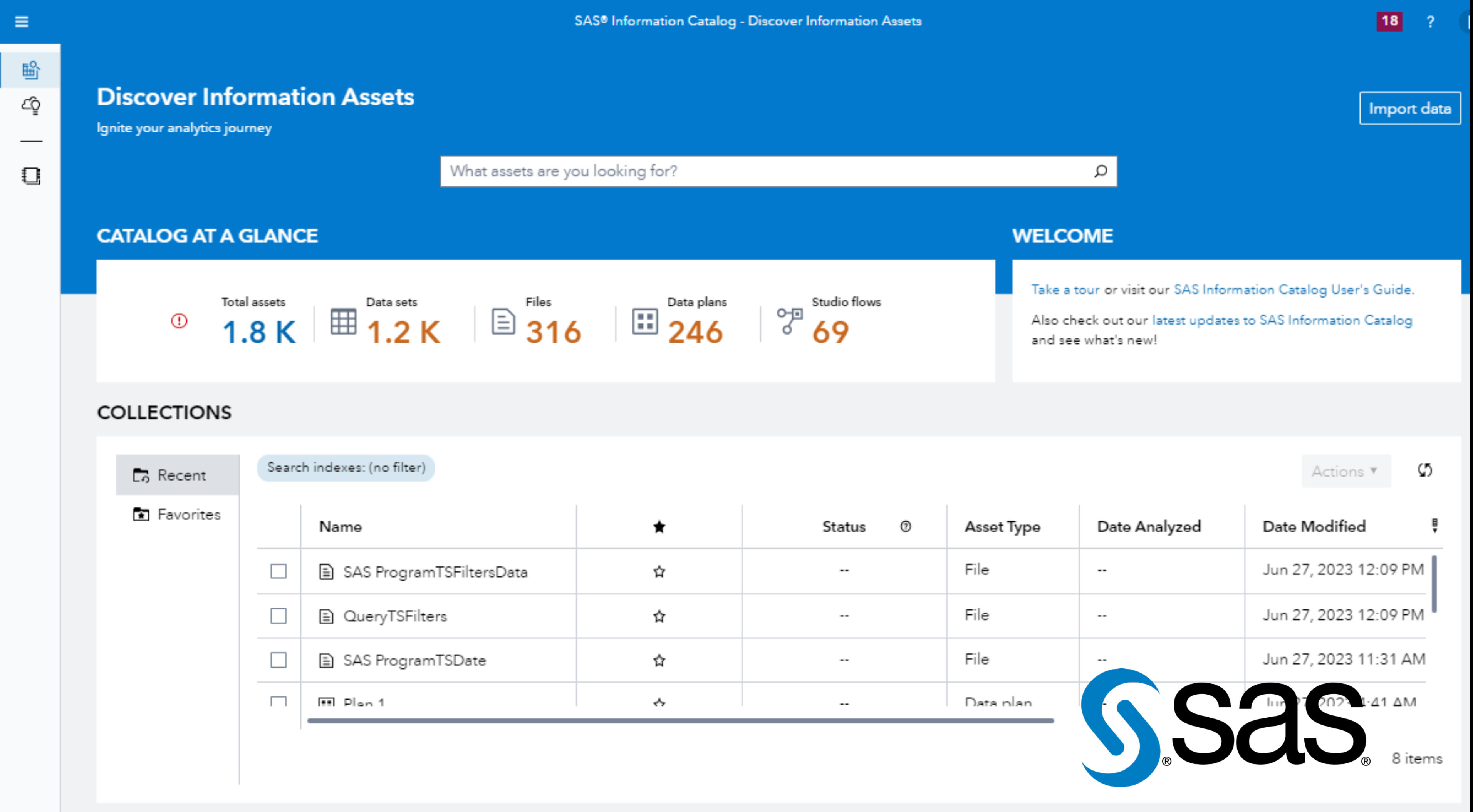
Data Catalogs capture Data about Data, aka Metadata

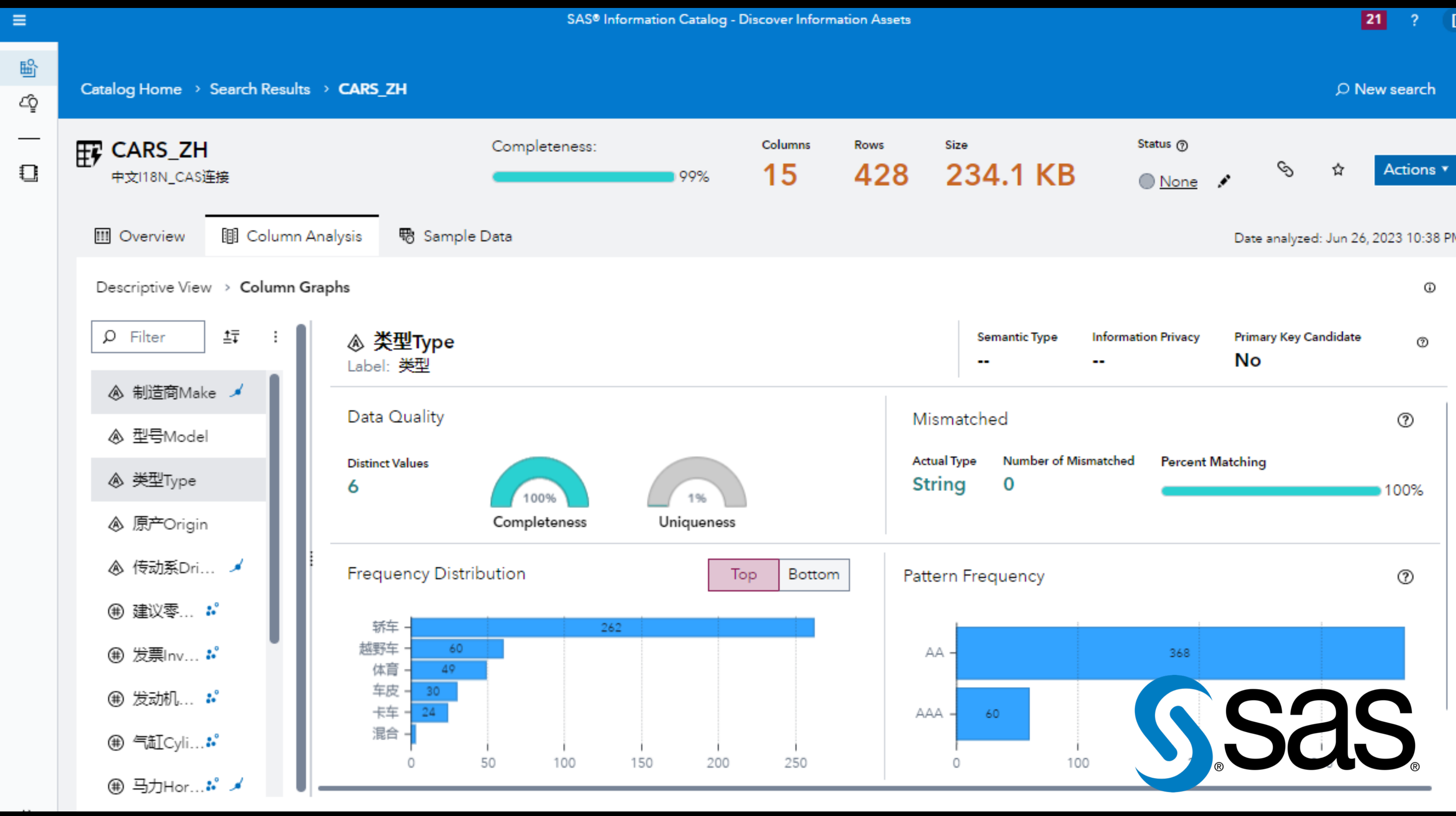
The ABC's of metadata

- Application Context — information needed to understand the data, its context, description, semantics

- Behavior — information about how the data was created, how it is maintained, who owns it, how is it provisioned?

- Change — how the data changes over time, and the processes that manage it


SAS® Information Catalog - Discover Information Assets
18
Discover Information Assets
Ignite your analytics journey
Import data
What assets are you looking for?
CATALOG AT A GLANCE
Total assets 1.8 K
Data sets 1.2 K
Files 316
Data plans 246
Studio flows 69
WELCOME
Take a tour or visit our SAS Information Catalog User's Guide.
Also check out our latest updates to SAS Information Catalog and see what's new!
COLLECTIONS
Recent
Favorites
Search indexes: (no filter)
Actions
Name
Status
Asset Type
Date Analyzed
Date Modified
SAS ProgramTSFiltersData -- File -- Jun 27, 2023 12:09 PM
QueryTSFilters -- File -- Jun 27, 2023 12:09 PM
SAS ProgramTSDate -- File -- Jun 27, 2023 11:31 AM
8 items


SAS® Information Catalog - Discover Information Assets
21
Catalog Home › Search Results › CARS_ZH
New search
CARS_ZH
中文I18N_CAS连接
Completeness:
99%
Columns
15
Rows
428
Size
234.1 KB
Status
None
Actions
Overview
Column Analysis
Sample Data
Date analyzed: Jun 26, 2023 10:38 PM
Descriptive View › Column Graphs
Filter
制造商Make
型号Model
类型Type
原产Origin
传动系Dri...
建议零...
发票Inv...
发动机...
气缸Cyli...
马力Hor...
类型Type
Label: 类型
Semantic Type
--
Information Privacy
--
Primary Key Candidate
No
Data Quality
Distinct Values
6
100%
Completeness
1%
Uniqueness
Mismatched
Actual Type
String
Number of Mismatched
0
Percent Matching
100%
Frequency Distribution
Top
Bottom
轿车
越野车
体育
车皮
卡车
混合
262
60
49
30
24
0
50
100
150
200
250
Pattern Frequency
AA
AAA
368
60
0
100
sas

Amundsen

Stars 4k | license Apache 2 | PRs welcome | commit activity 0/week | contributors 209 | Follow | Slack join chat

Amundsen is a *data discovery and metadata engine* for improving the productivity of data analysts, data scientists and engineers when interacting with data. It does that today by indexing data resources (tables, dashboards, streams, etc.) and powering a page-rank style search based on usage patterns (e.g. highly queried tables show up earlier than less queried tables). Think of it as **Google search for data**. The project is named after Norwegian explorer Roald Amundsen, the first person to discover the South Pole.

LF AI & DATA

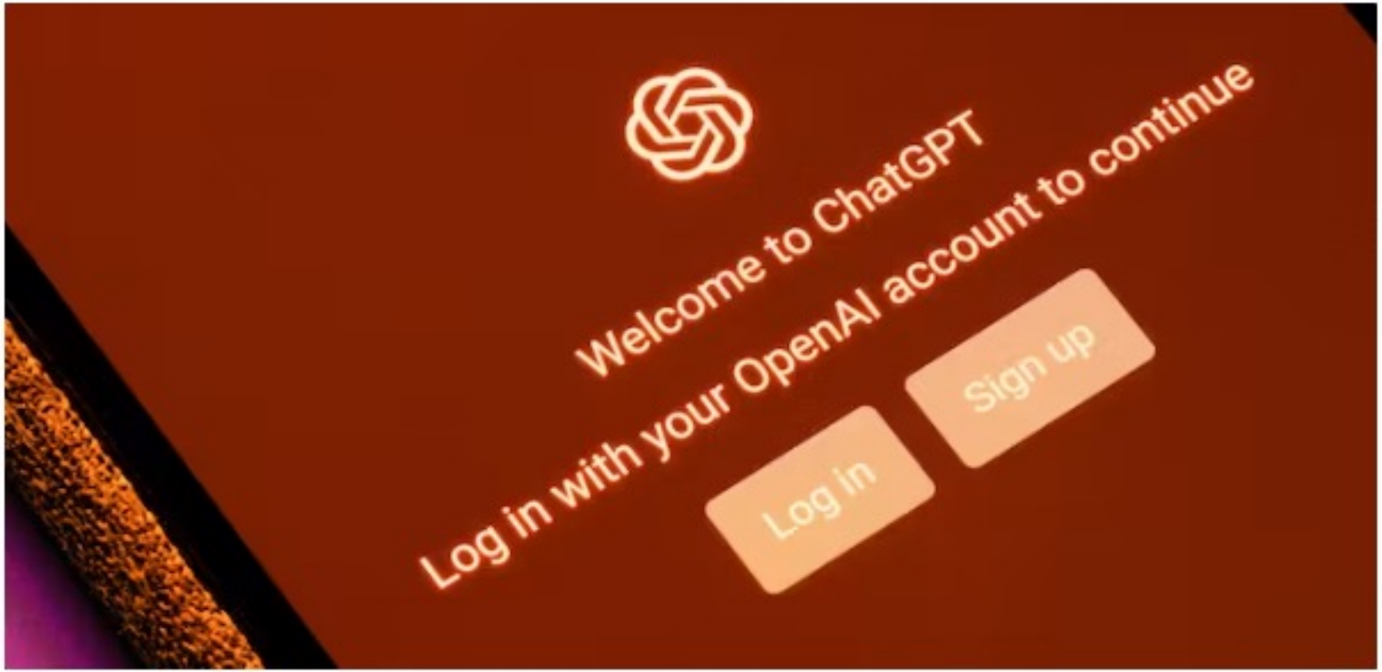
Amundsen is hosted by the LF AI & Data Foundation. It includes three microservices, one data ingestion library and one common library.

open source

Data Provenance

Why is it important?

- The validity, authenticity and integrity of experiments hinges on ability to reproduce the results consistently
- Ethical data provenance is important to ensure that we do no harm
- Lineage
  - Where data comes from
  - where it goes
  - know when to update for changes
  - how to handle change
  - who is using the data we have



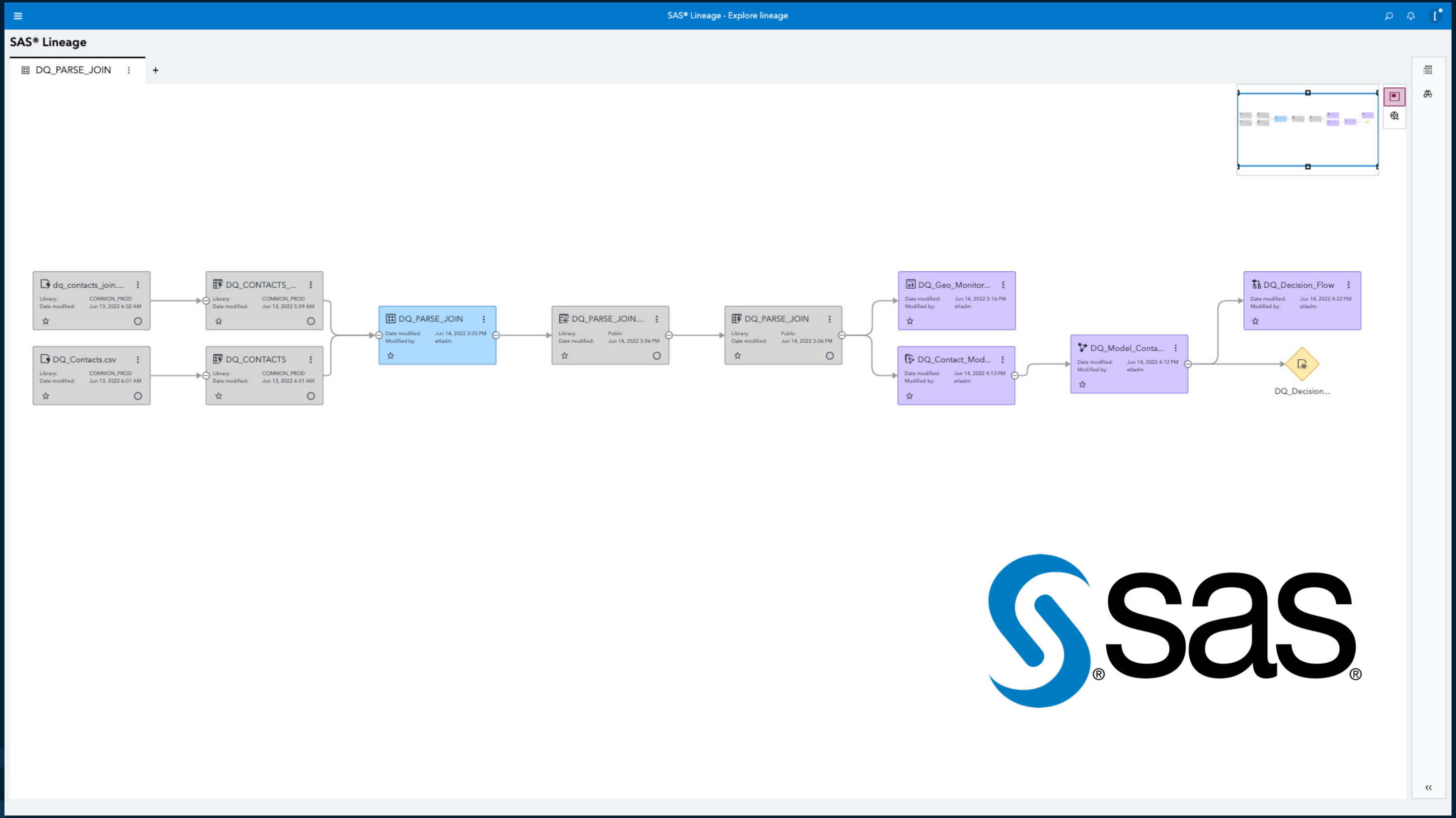
February 9, 2023

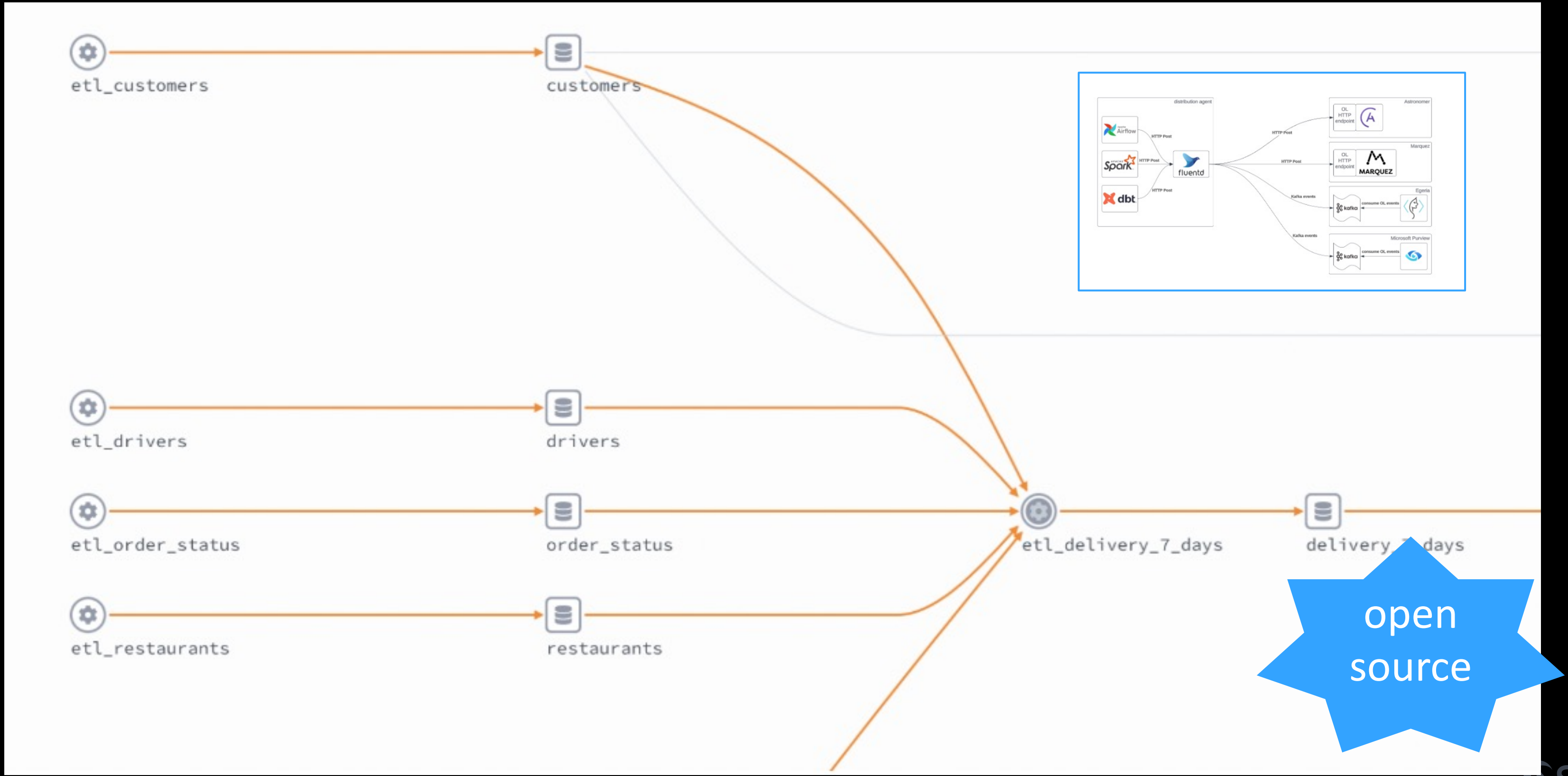
ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned

Uri Gal, *University of Sydney*

ChatGPT is fuelled by our intimate online histories. It's trained on 300 billion words, yet users have no way of knowing which of their data it contains.

Lineage/Data Provenance



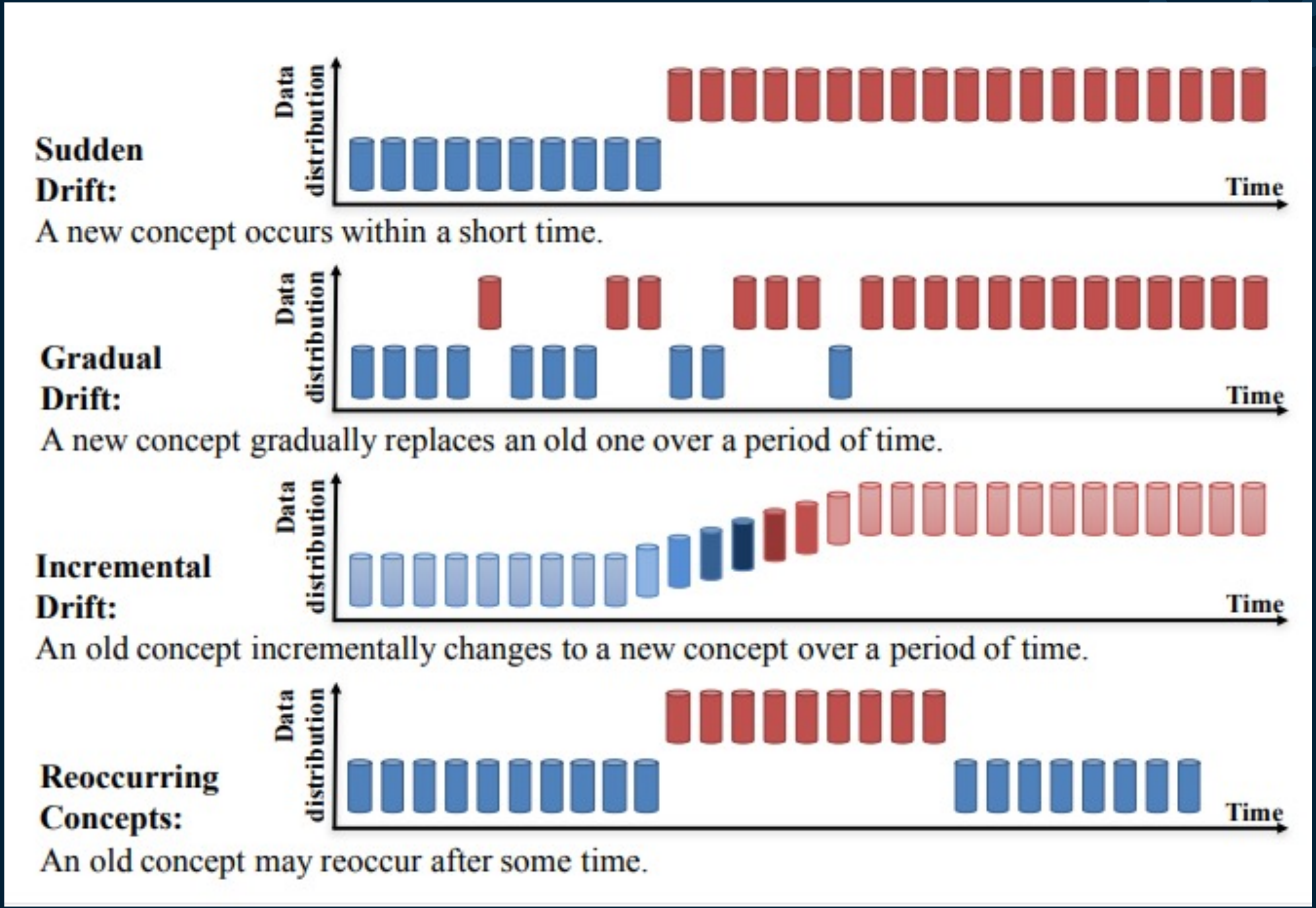

etl_customers
customers
etl_drivers
drivers
etl_order_status
order_status
etl_restaurants
restaurants
etl_delivery_7_days
delivery
days
distribution agent
Airflow
Spark
dbt
fluentd
HTTP Post
Astronomer
Marquez
MARQUEZ
Egeria
Microsoft Purview
kafka
open
source

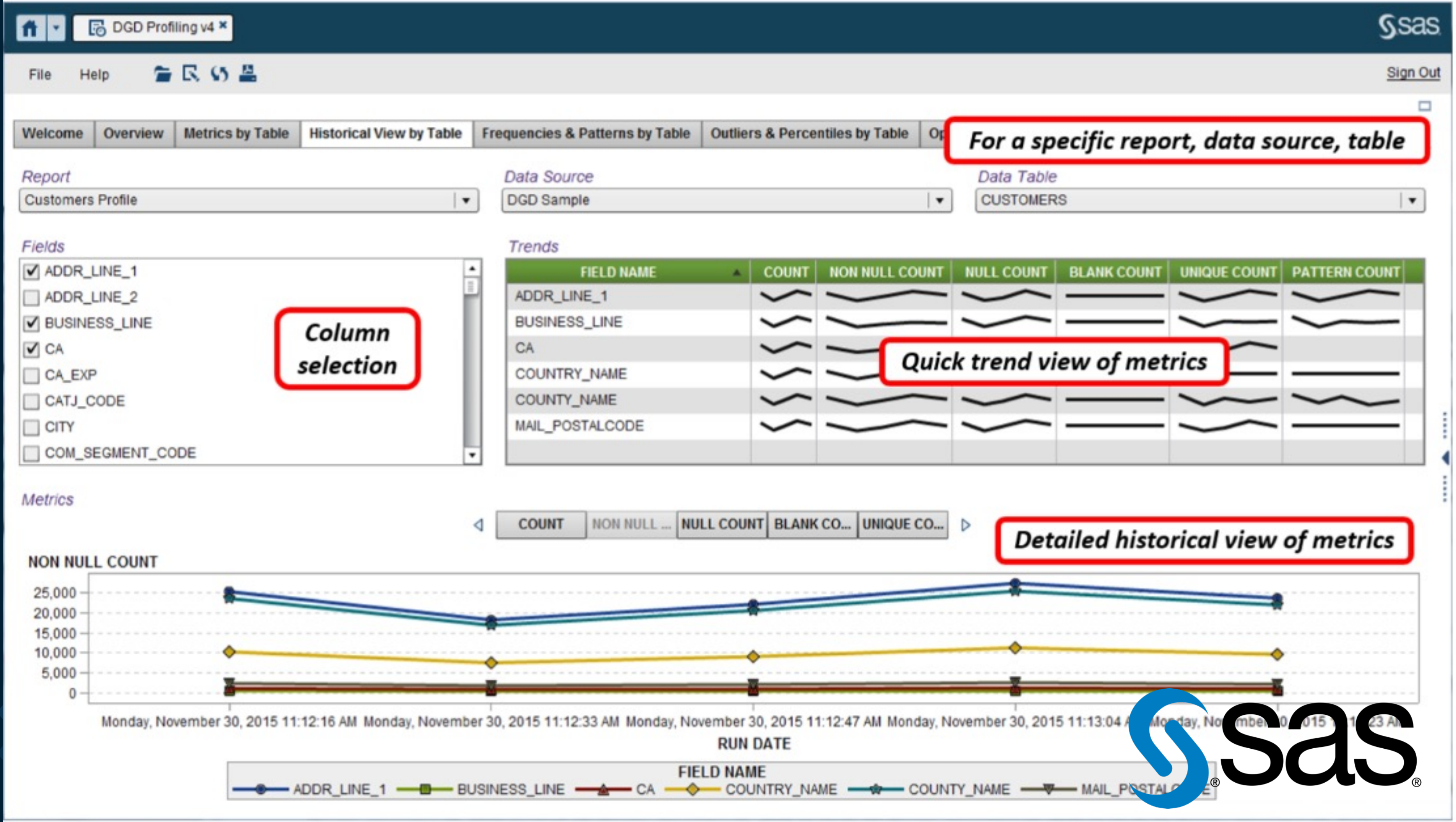
An update from the ML Workflow & Interop Committee dataset license compliance initiative

Howard <huangzhipeng@huawei.com>
Liza lizi4@huawei.com
Gopi Krishnan Rajbahadur <gopi.krishnan.rajbahadur1@huawei.com>

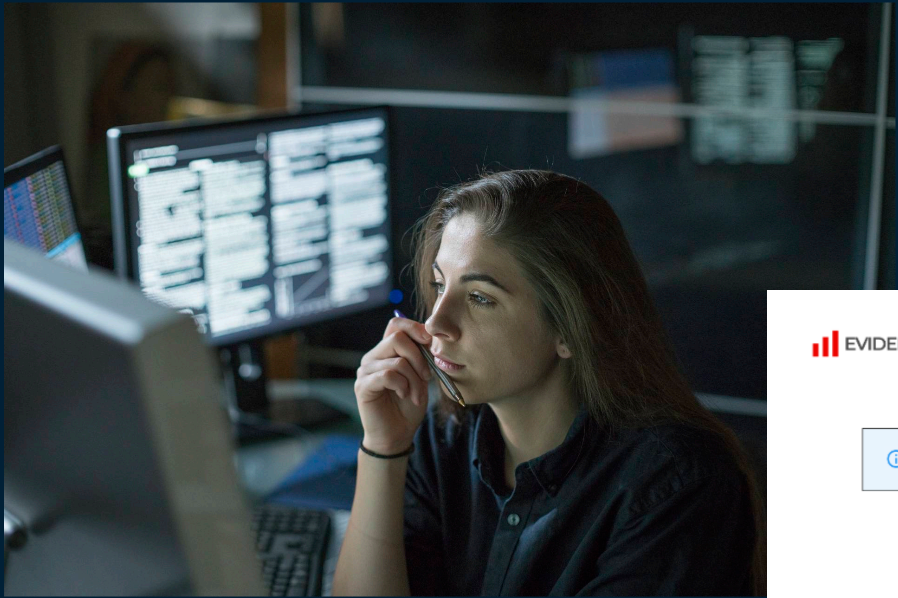
LF AI & DATA

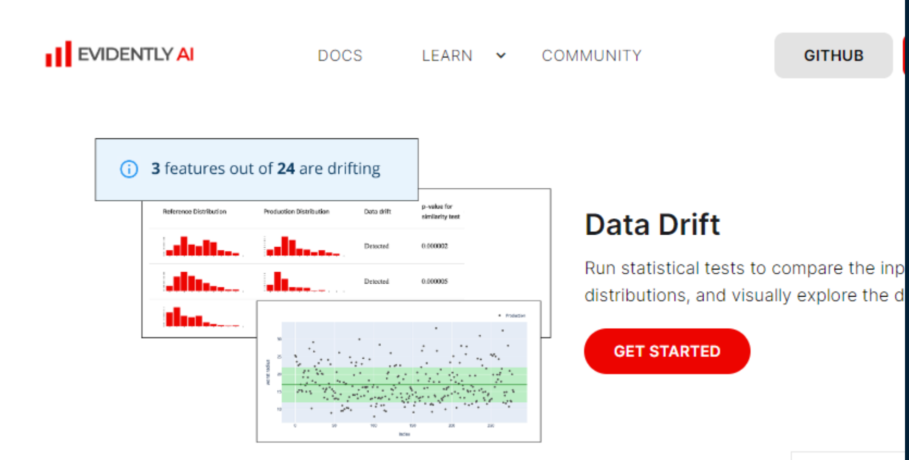
Data drift

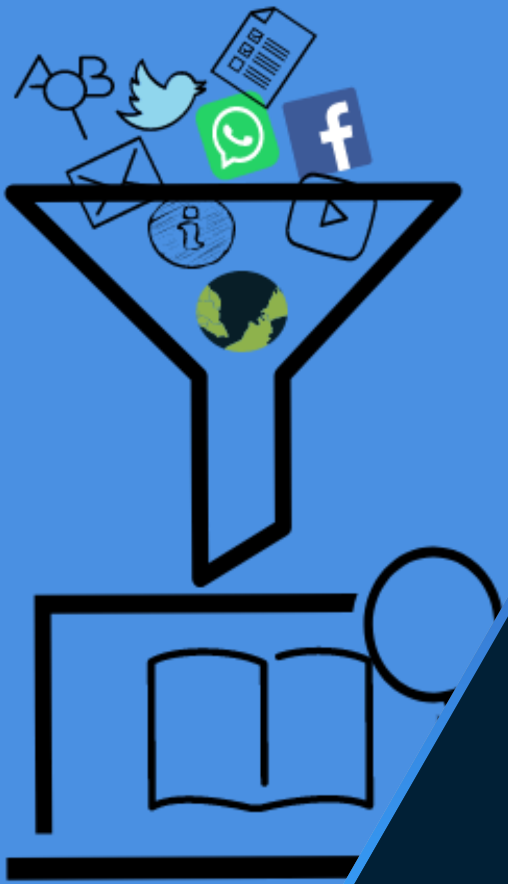

Sudden Drift:
Data distribution
Time
A new concept occurs within a short time.
Gradual Drift:
Data distribution
Time
A new concept gradually replaces an old one over a period of time.
Incremental Drift:
Data distribution
Time
An old concept incrementally changes to a new concept over a period of time.
Reoccurring Concepts:
Data distribution
Time
An old concept may reoccur after some time.


DGD Profiling v4
File
Help
Sign Out
Welcome
Overview
Metrics by Table
Historical View by Table
Frequencies & Patterns by Table
Outliers & Percentiles by Table
For a specific report, data source, table
Report
Customers Profile
Data Source
DGD Sample
Data Table
CUSTOMERS
Fields
ADDR_LINE_1
ADDR_LINE_2
BUSINESS_LINE
CA
CA_EXP
CATJ_CODE
CITY
COM_SEGMENT_CODE
Column selection
Trends
FIELD NAME
COUNT
NON NULL COUNT
NULL COUNT
BLANK COUNT
UNIQUE COUNT
PATTERN COUNT
ADDR_LINE_1
BUSINESS_LINE
CA
COUNTRY_NAME
COUNTY_NAME
MAIL_POSTALCODE
Quick trend view of metrics
Metrics
COUNT
NON NULL ...
NULL COUNT
BLANK CO...
UNIQUE CO...
Detailed historical view of metrics
NON NULL COUNT
25,000
20,000
15,000
10,000
5,000
0
Monday, November 30, 2015 11:12:16 AM
Monday, November 30, 2015 11:12:33 AM
Monday, November 30, 2015 11:12:47 AM
Monday, November 30, 2015 11:13:04
RUN DATE
FIELD NAME
ADDR_LINE_1
BUSINESS_LINE
CA
COUNTRY_NAME
COUNTY_NAME
sas

open
source

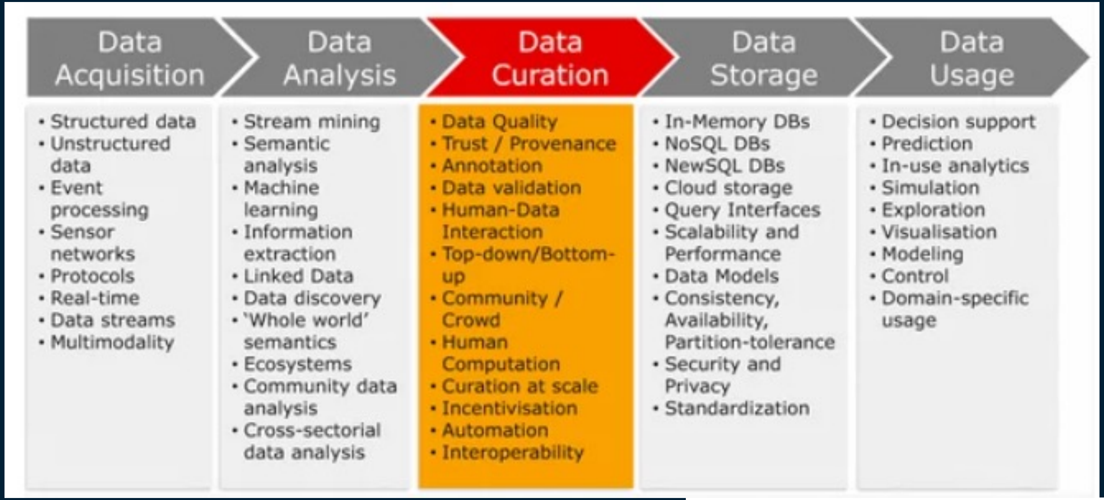



EVIDENTLY AI
DOCS
LEARN
COMMUNITY
GITHUB
3 features out of 24 are drifting
Reference Distribution
Production Distribution
Data drift
p-value for similarity test
Detected
0.000002
Detected
0.000005
Data Drift
Run statistical tests to compare the inp
distributions, and visually explore the d
GET STARTED



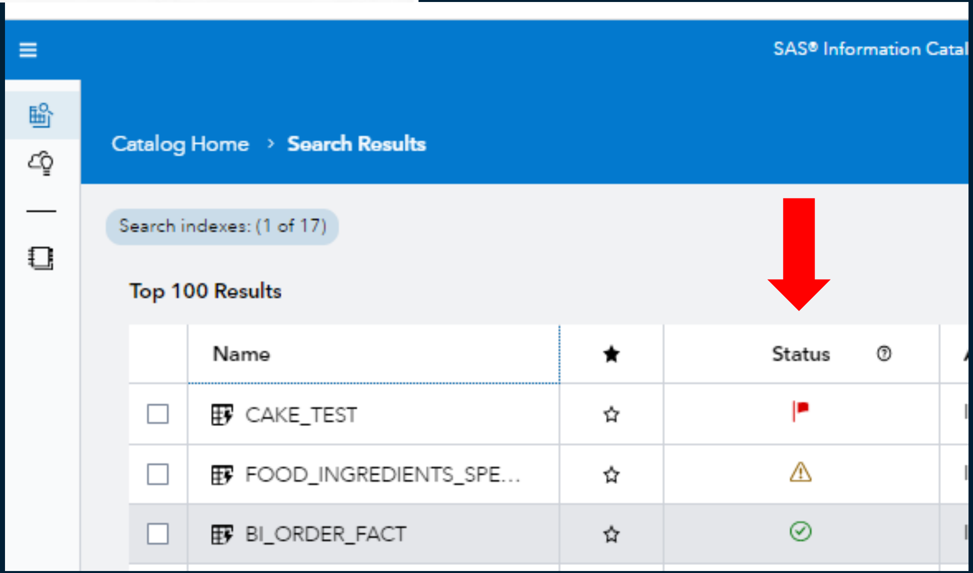
Curation

The process of gathering relevant information to add value through the process of selecting, organizing, and looking after the items in a collection.



Data curation

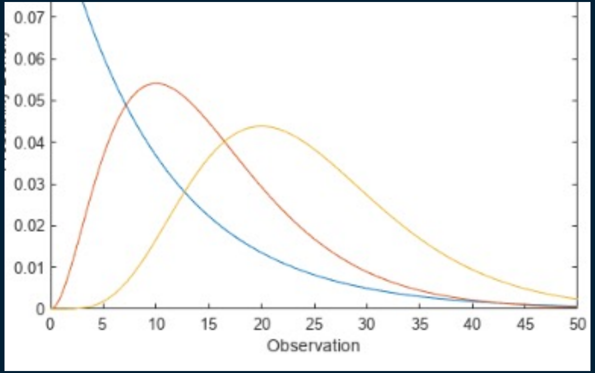
- Domain expert knowledge
- Shared information about usage

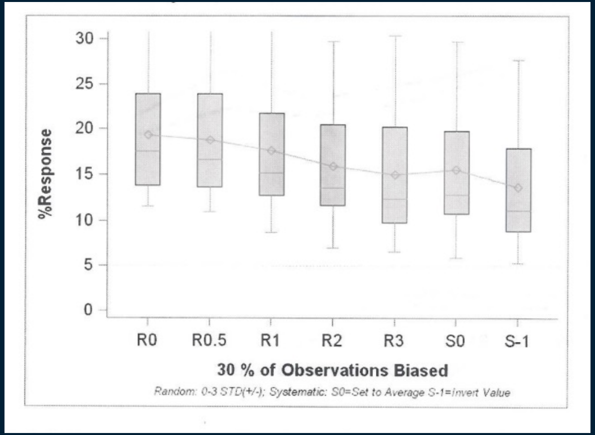


Bias

Early detection is very useful

- Watch for protected groups
- Expert domain knowledge in feature engineering
- Ensure fair and unbiased data collection strategies
- Watch for types of data bias; random is less intrusive than systemic
- Mitigate using techniques such as resampling, augmentation, cross-validation, improved feature engineering





Agenda

1. Why data quality?
2. Data cleaning techniques: Dimensions of data quality
3. Design considerations for analytical and reporting use cases
4. Advanced topics
5. Wrap up and summary

Want to learn more?

Check out these links!

- https://www.coursera.org/sas - Coursera SAS classes
- https://www.sas.com/en_us/training/overview.html - SAS Training website
- Statistics 1: ANOVA, Regression, Logistic Regression - https://support.sas.com/ecst1
- SAS Programming Essentials https://support.sas.com/ecprg1
- Over 200 free tutorials: https://video.sas.com/detail/videos/how-to-tutorials
- SAS OnDemand for Academics https://www.sas.com/en_us/software/on-demand-for-academics.html#section=5

Questions?

Thank you!