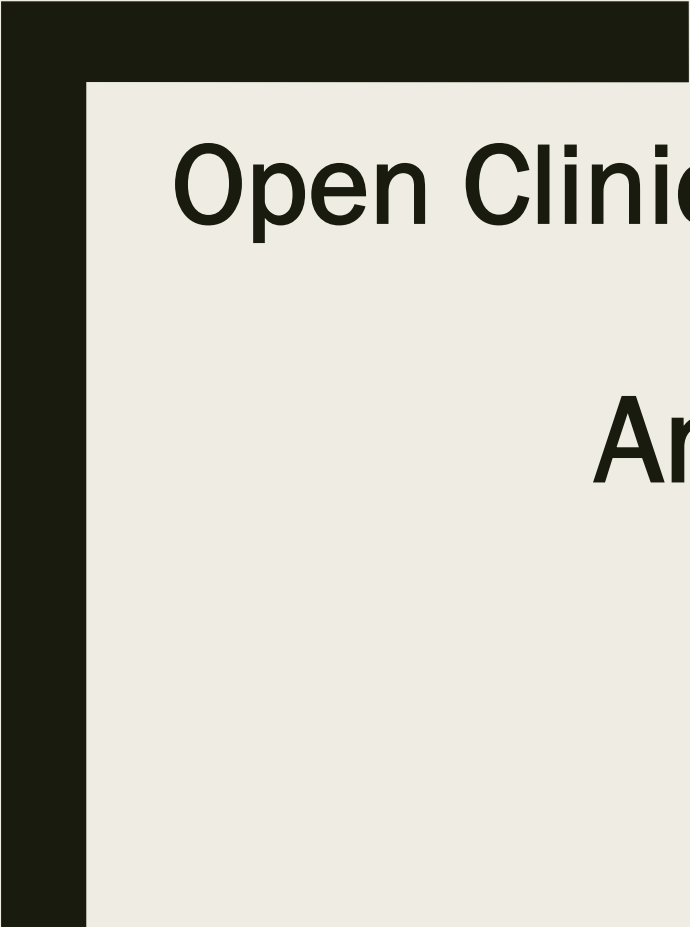
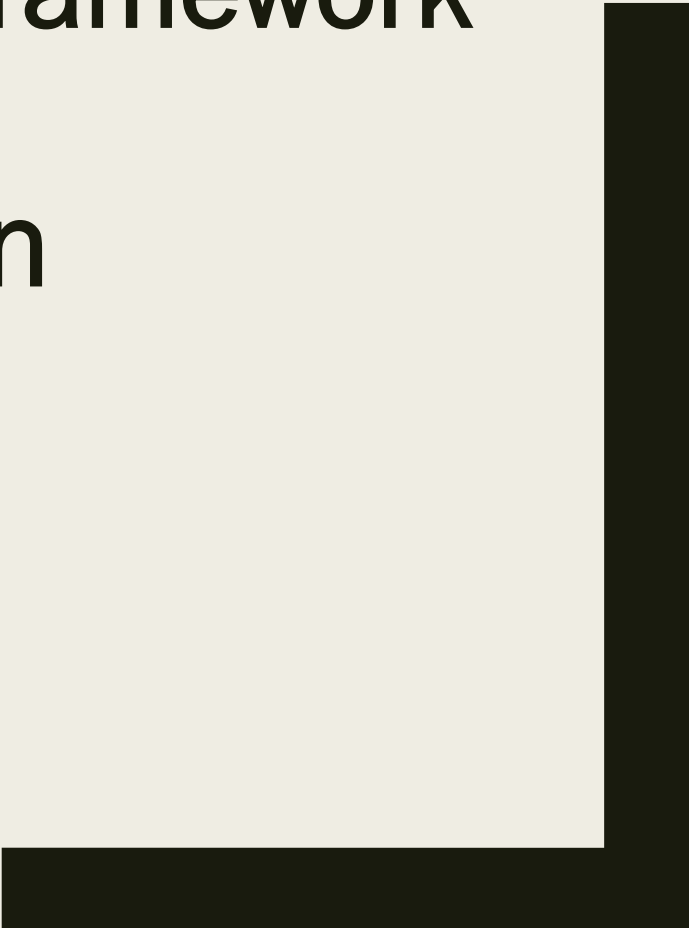




Open Clinical Data Quality Framework for Anomaly Detection

By
Vineet Jain,
Nimble Clinical Research



Data Abstraction and Review

Subject Level

- Edit Checks
- Monitoring, SDV
- Data Listings
- Other Misc.

Site/Country/Cohort level

- Custom Reports, Visualization
- Risk Based Monitoring, KRIs, KPIs
- **Central Statistical Monitoring**

Central Statistical Monitoring



Advantages

- Find deviations from study procedures, data entry errors, miscalibrated equipment, sloppiness, or fraud
- Issues found sooner than later
- Most of the data can be analyzed



Challenges

- Costly and study specific (i.e. Lack of standard approach)
- May not be appropriate for smaller trials
- Findings may be hard to interpret
- Selecting Outliers from extreme values

Features of our Anomaly Detection Framework

Unsupervised

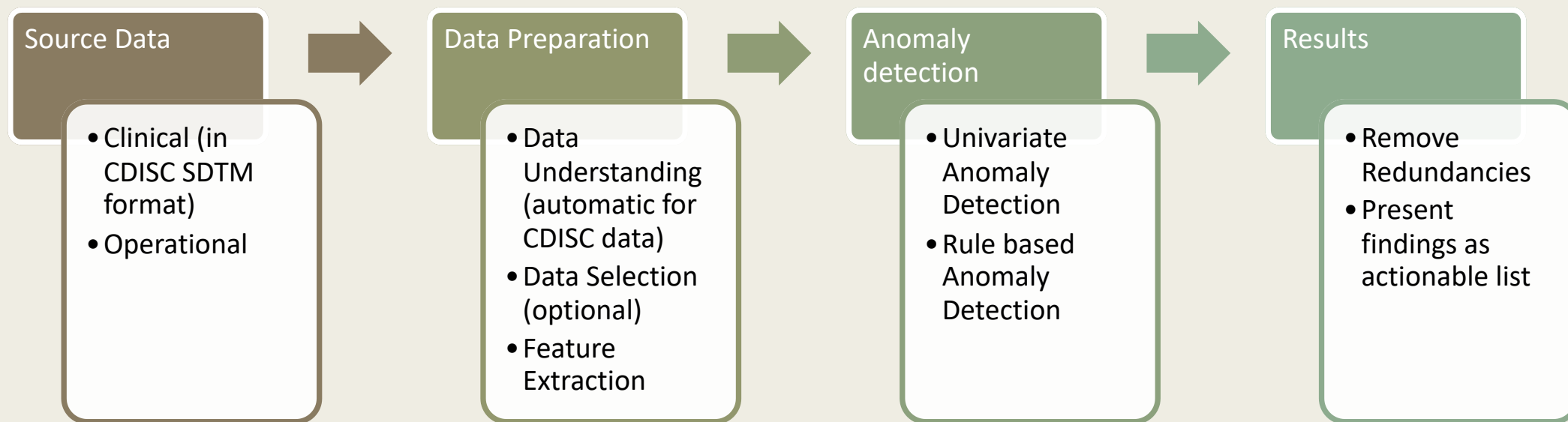
Deeper learning

Simple & sound Statistical Approaches

Open & Free Actionable Findings

Works for small to large Trials

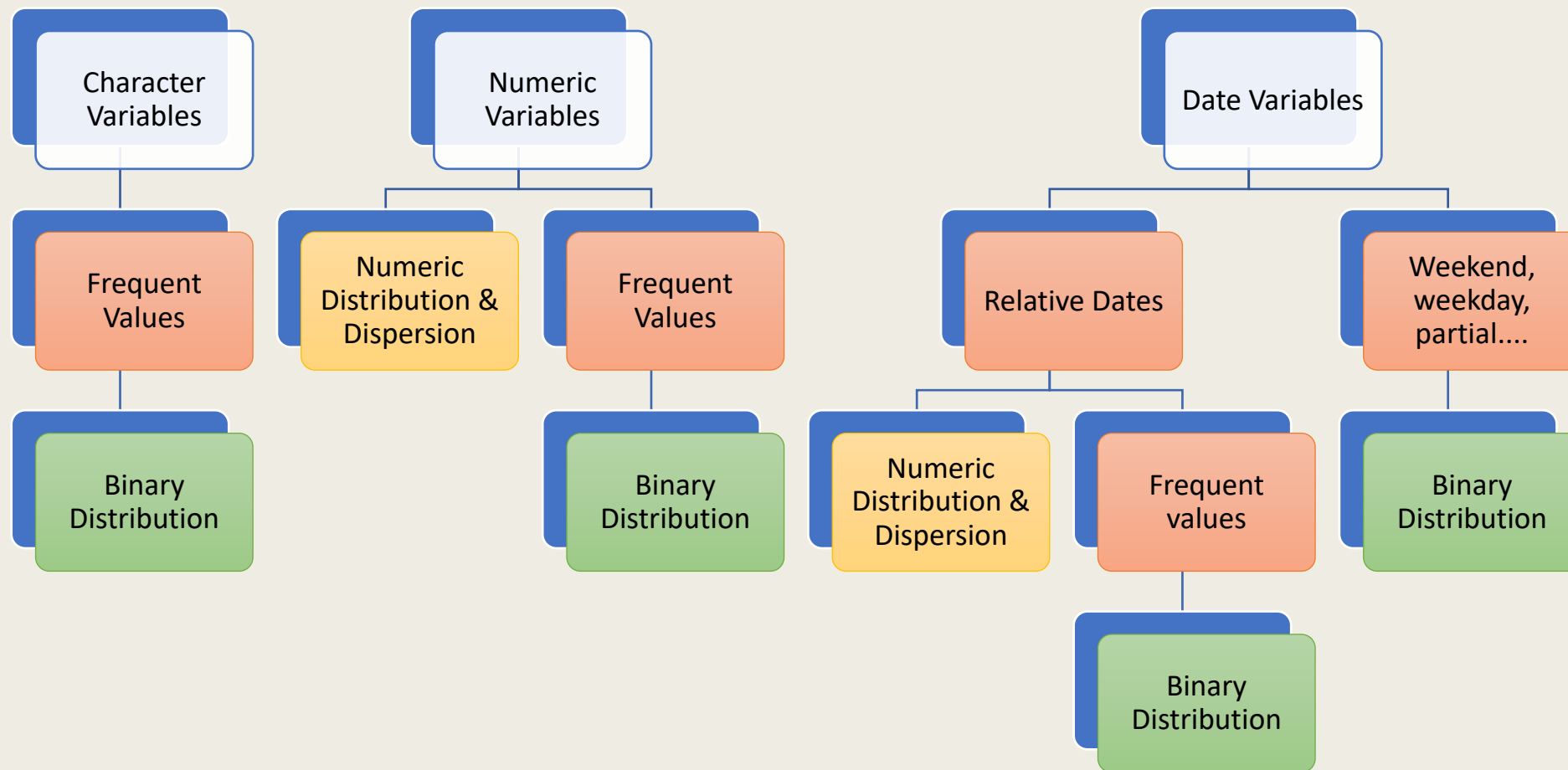
Anomaly Detection Process



Unsupervised Data Understanding via CDISC data standard

SDTM domain Type	Variable Name	Similar to	Role
All	ARMCD	ARM	
All	--CAT		Category
All	EPOCH		Skip
Findings	--LLOQ		Skip
Events, Interventions	--MODIFY		Skip
Findings	--NAM		Skip
All	--SEQ		Skip
Events	--SOC	--BODSYS	
Findings	--TESTCD	--TEST	
DM	RFXSTDTC	RFSTDTC	
All	USUBJID		Subject ID

Feature Extraction



Types of Anomalies detected

■ Univariate

- *Distribution & Dispersion of number of records per subject*
- *Distribution and Dispersion of numeric features/variables*
- *Binary Distribution for frequent values*

■ Rule based Bivariate

- *Binary Distribution of paired associated frequent values*

Site	Subject Identifier	Vital Signs Test	Result	Units	Visit	Date/Time	Clinically Significant
01	2301	Diastolic BP	81	mmHg	Screen	2012-01-26	
01	2301	Diastolic BP	106	mmHg	Day 1	2012-02-06	Y
01	2301	Diastolic BP	58	mmHg	Week 1	2012-02-13	
01	2301	Diastolic BP	66	mmHg	Week 2	2012-02-20	
01	2301	Height	191	cm	Screen	2012-01-26	
01	2301	Systolic BP	89	mmHg	Screen	2012-01-26	
01	2301	Systolic BP	170	mmHg	Day 1	2012-02-06	Y
01	2301	Systolic BP	77	mmHg	Week 1	2012-02-13	
01	2301	Systolic BP	85	mmHg	Week 2	2012-02-20	

Rule based Anomalies

Overall AE Data			
		Severe	
		Yes	No
Serious	Yes	30	10
	No	10	30

Site 101 AE Data	
Severe	
Yes	No
5	2
8	5

		Serious = Yes & Severe = Yes	
		Yes	No
Sites	101	5	15
	Remaining	25	35

Serious = Yes & Severe = No	
Yes	No
2	18
8	52

Serious = No & Severe = Yes	
Yes	No
8	12
2	58

Serious = No & Severe = No	
Yes	No
5	15
25	35

P-values:

0.28

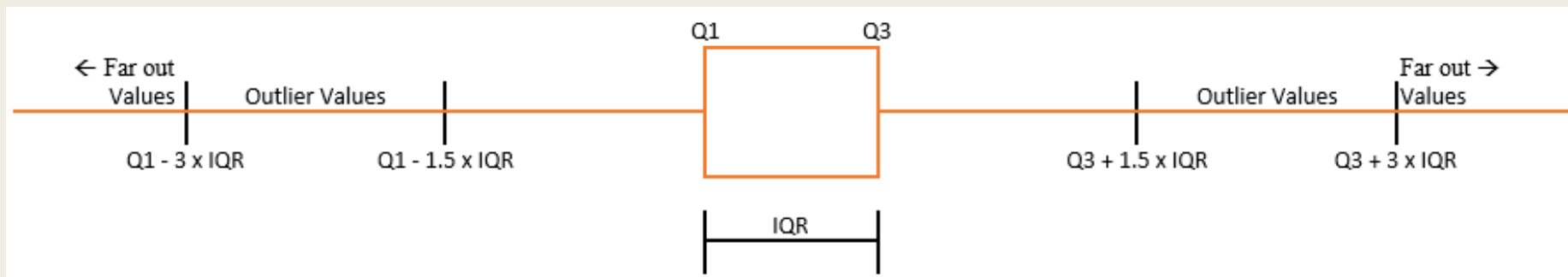
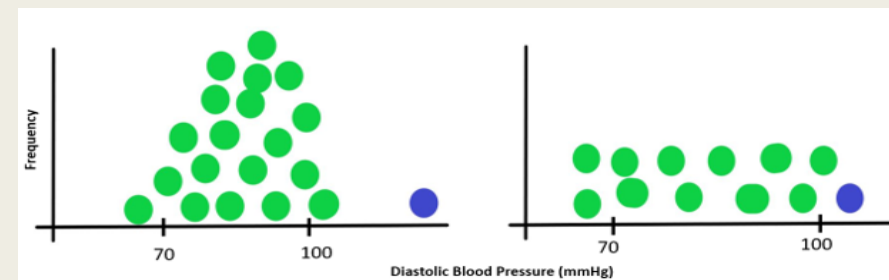
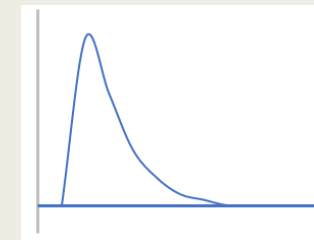
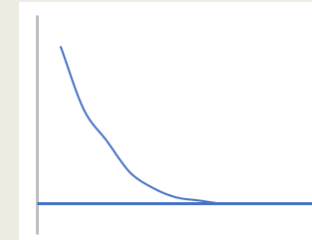
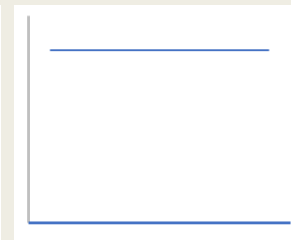
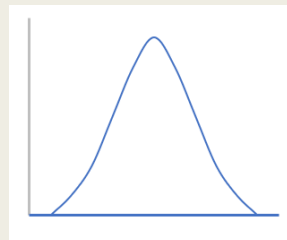
1

0.000095

0.28

Statistical Models

- Unknown Distribution of data
- Non-parametric
 - *Wilcoxon Rank Sum Test*
 - *Conover Squared Ranks Test*
 - *Chi-Square Test*
- Supplemented by Tukey Fences



Presentation of Results

Group	Data				Finding Type	Records		Non-Parametric Statistics		Tukey Fences		
	Dataset	Data Split	Variables	Data Selection		Numbers	%	Model	p-value	Statistic	Q1, Q3	Distance from IQR
SITEID = '103'	CE	NA	NA		No. of records per subject	78 out of 345	22.6	Wilcoxon rank-sum	<.00001	Mean: 15.6	0, 4.5	2.5
SITEID = '101'	CM	Category for Medication (CMCAT) = 'CONCOM-ITANT MEDS'	Start Date (CMSTDTC) w.r.t to Treatment Start Date (Relative)		Numeric Distribution	20 out of 523	3.8	Wilcoxon rank-sum	0.00005	Mean: 100.6	25, 50	2.0
SITEID = '103'	VS	Vital Signs Test Name (VSTEST) = 'Temperature'	Numeric Result/Finding in Standard Units (VSSTRESN)		Numeric Dispersion	210 out of 2212	9.5	Conover squared ranks	<.00001	SD: 1.13	.11, 0.27	5.4
SITEID = '105'	AE	NA	Serious Event (AESER)	AESER= 'Mild'	Univariate Binary Distribution	5 out of 34 vs. 352 out of 383	8.9	2x2 Chi-Square	<.00001	Ratio: .15	.75, 1	2.4
SITEID = '107'	CM	Category for Medication (CMCAT) = 'Rescue Medication'	Reported Name (CMTRT) & Dose per Administration (CMDOSE)	CMTRT = 'Dummy' & CMDOSE > 5	Bivariate Binary Distribution	15 out of 100 vs. 5 out of 353	22.1	2x2 Chi-Square	<.00001	Ratio: .15	0, .03	4.0

Challenges and solutions

Challenges

- *Computationally expensive*
- *Too many trivial findings*

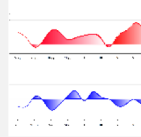
Solutions

- ***Better Data selection***
 - drop uninteresting variables/data
 - select representative data among interesting yet similar data
 - Remove repeating data, and
 - De-normalized data if needed.
- ***Appropriate thresholds, e.g.***
 - Minimum Data cut for binary data
 - P-value cut off for statistical tests

Future Improvements



Move away from current dependence on SAS to a cross platform application



Add ability to detect **temporal anomalies**



Supplement the listing based findings with **Visualization**



Use **better algorithms** to make anomaly discovery computationally efficient



THANK YOU
&
QUESTIONS