DETECTING TEMPORAL DEPENDENCIES IN DATA

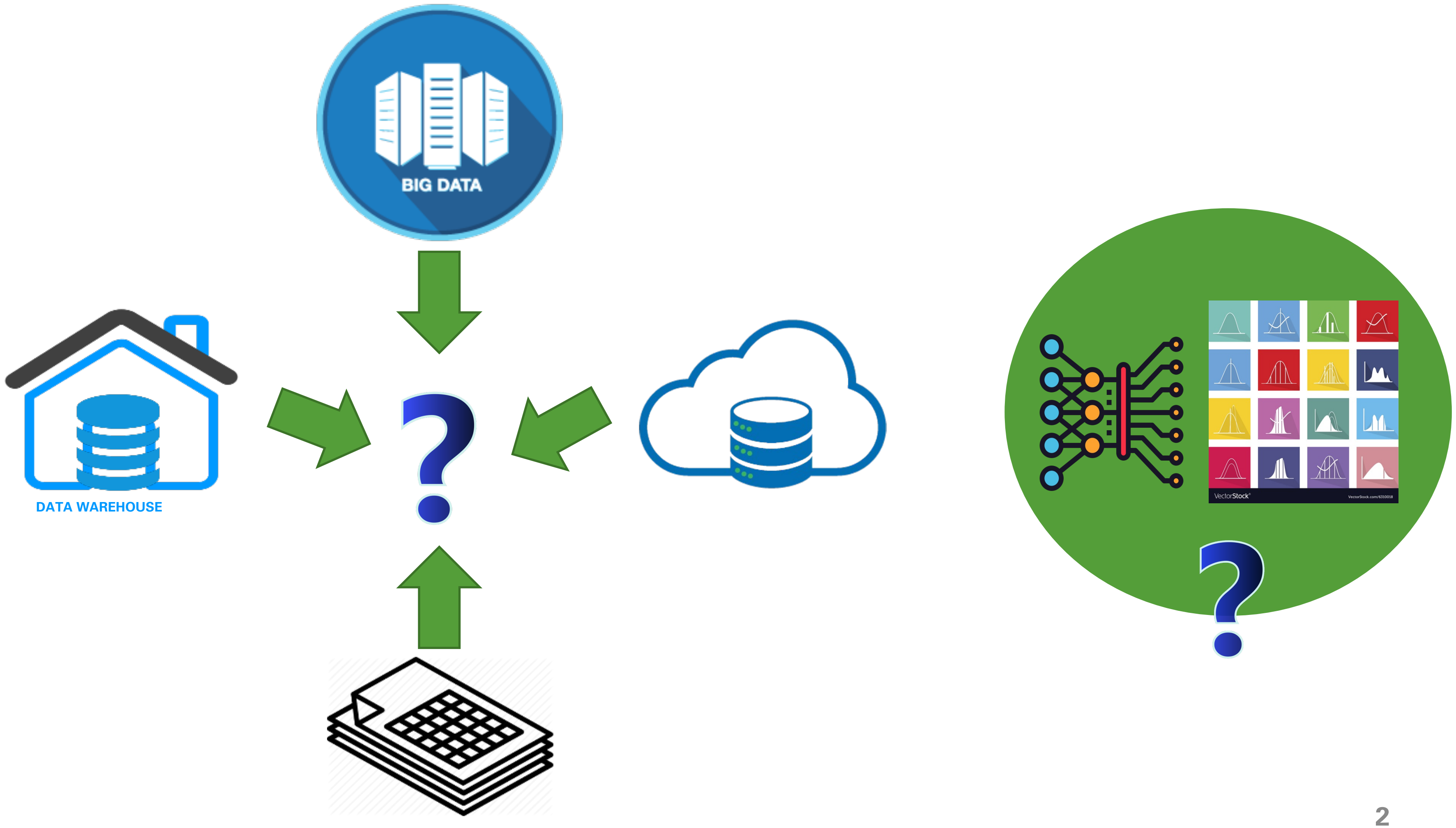
Joaquin Cuomo, Hajar Homayouni, Indrakshi Ray, Sudipto Ghosh

hhomayouni@sdsu.edu

San Diego State University

Colorado State University

Datasets with Unkown Characteristics



Attribute Characteristics

- Non-temporal
- Temporal
- Hidden Temporal

Non-Temporal Attribute

A set of observations for which the order does not matter

US Election 2020 by County

state	county	current_votes	total_votes	percent
Delaware	Kent County	85415	87025	100
Delaware	New Castle County	280039	287633	100
Delaware	Sussex County	127181	129352	100
Indiana	Adams County	14154	14209	100
Indiana	Allen County	168312	169082	100

Temporal Attribute (Time Series)

- A sequence of observations equally spaced and ordered by time
- Temporal dependence implies that future values are influenced by past values

Daily Average Energy Consumed by Residents

id	Time	Temperature	DailyDelivered
1	1/1/2019	5.541666667	182207.1887
2	1/2/2019	19.23791667	205679.3273
3	1/3/2019	32.9275	197726.4854
4	1/4/2019	39.55958333	192919.3705
5	1/5/2019	41.17291667	173748.8355

Hidden-Temporal Attribute

A hidden (grouping) temporal attribute can only be treated as temporal if its values are categorized in groups

Patient ID	Date	Weight
1001	6/1/2020	125.2
1001	7/1/2020	125.6
1005	7/1/2020	26.5
1001	8/1/2020	126.1
1005	8/1/2020	27

Patient ID	Date	Weight
1001	6/1/2020	125.2
1001	7/1/2020	125.6
1001	8/1/2020	126.1

Patient ID	Date	Weight
1005	7/1/2020	26.5
1005	8/1/2020	27

Finding proper grouping attribute is main challenge in this case

Existing Approaches

Rely on domain experts to:

- Identify type of data
- Choose appropriate techniques to model data

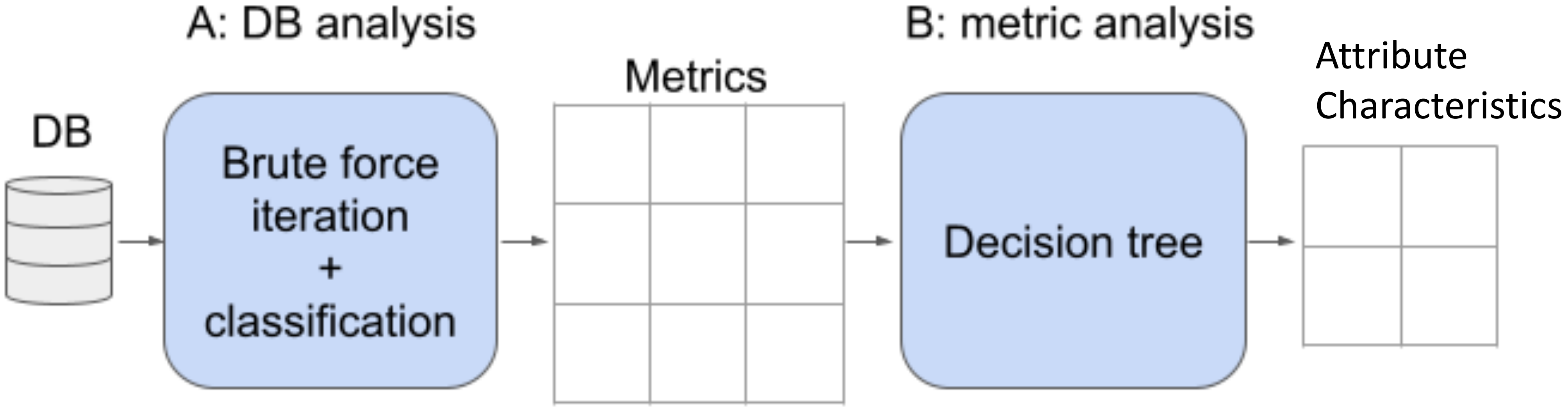
Limitations

- Domain experts cannot analyze all attributes in big datasets
- Domain experts may not be aware of temporal dependencies among a subset of attributes in big datasets
- Data transformations can make temporal nature of target attributes unknown to domain experts

Research Goals

- Automatically determine presence of temporal data in a dataset given no prior knowledge about its attributes
- Automatically identify grouping attributes by which we can group dataset records and obtain intergroup temporal attributes but not intragroup
- Classify an attribute as temporal, non-temporal, or hidden temporal

Overview of Proposed Approach

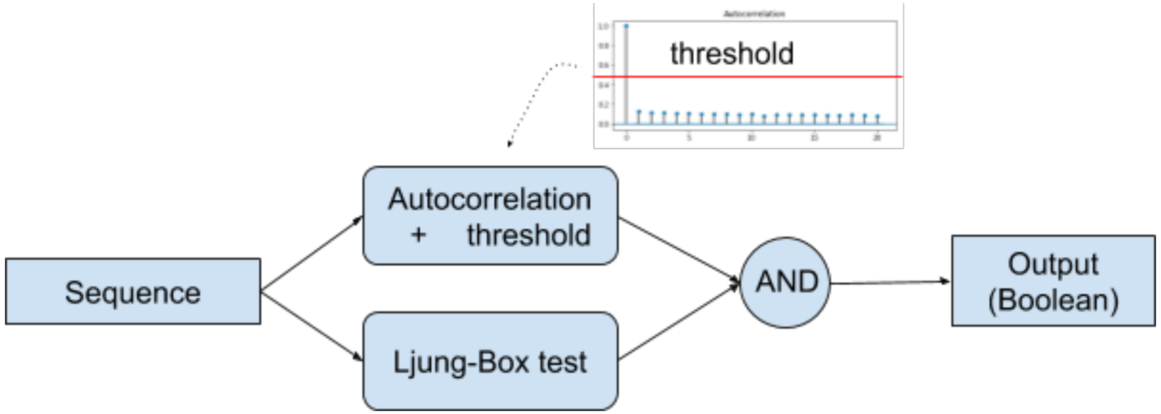


Stage A: DB Analysis

Objective: Calculate a set of metrics that help identify attribute characteristics

Approach

- Iterates over all numeric attributes
- Groups dataset by those attributes
- Classifies resulting subsequencess as time-dependent or not



Classifier: analyzes a single attribute and determining if it has autocorrelation

Stage A: Ljung-Box Test

Objective: Detect statistically significant autocorrelation

Null hypotheses: Data is independently distributed

Alternative hypotheses: Data exhibits serial correlation up to any lag

Test statistic:

$$Q(m) = n(n+2) \cdot \sum_{k=1}^{m} \frac{r_k^2}{n-k}$$

where n is the sample size, *r* is sample autocorrelation at lag *k*, and *m* is number of lags being tested

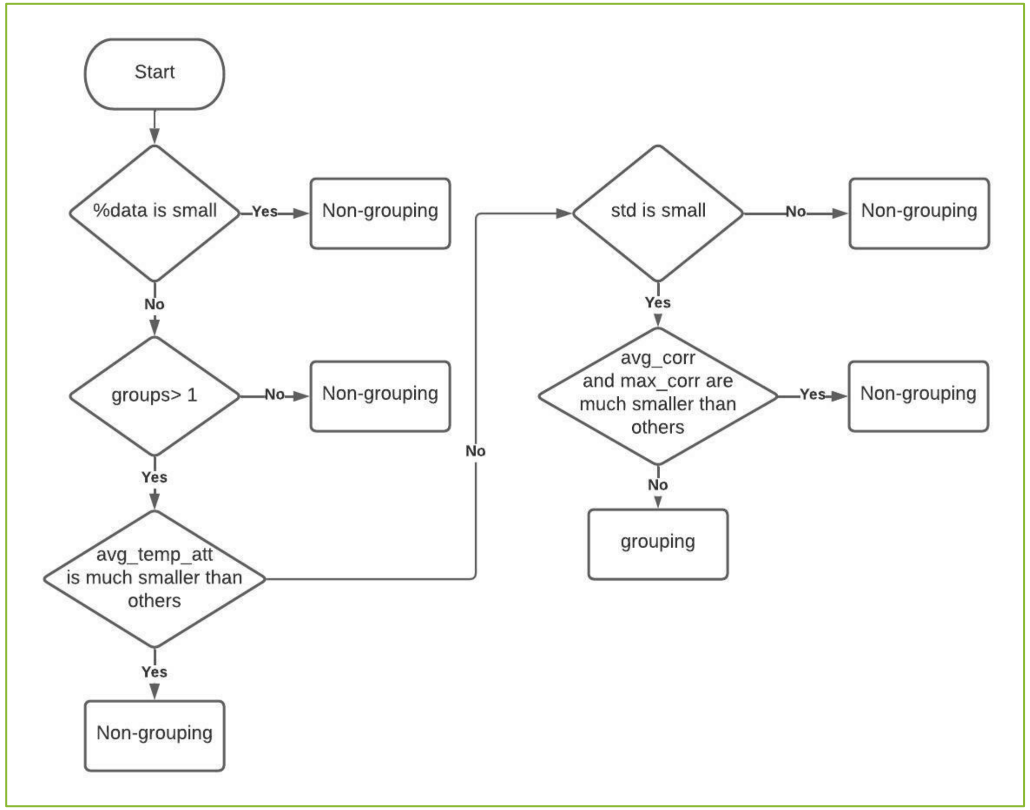
Under null hypotheses statistic Q asymptotically follows a χ^2

Rejection of null hypotheses indicates that there is autocorrelation in input sequence

Stage B: Decision Tree

Objective: analyze metrics to determine if grouping by attributes generates temporal sequences

Name	Description
%data	Percentage of records from groups with at least one attribute classified as temporal over the entire dataset
groups	Count of groups with at least one attribute classified as temporal
avg_temp_att	Average of the count of attributes classified as temporal over the groups
std	Standard deviation of avg_temp_att.
avg_corr	Average of maximums autocorrelations over all groups. Maximum values are calculated within a group, over all attributes classified as temporal
Max_corr	Maximum autocorrelation over all attributes and groups



Output: attributes with temporal dependence along with percentage of times it was detected as temporal over all groups

Evaluation

Objectives

- Demonstrate that our approach can correctly identify attributes with temporal dependence in datasets
- Demonstrate that our approach can correctly identify grouping attributes to form multiple temporally dependent sequences

Subjects: 15 datasets with attributes given a priori classification by domain experts

Metrics: F1 score and Accuracy

$$F1 = \frac{TP}{TP + 1/2(FP + FN)}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Experimental Results

	case	FP	TP	FN	TN	ACC	F1	# temp att detected	grouping	grouping detected
elections	0	0	0	0	3	1	N/A	0/0	False	False
incomes	0	0	0	0	3	1	N/A	0/0	False	False
countries	0	2	0	0	16	0.88	0	2/0	False	False
biomechanical	0	2	0	0	4	0.66	0	2/0	False	False
crime	0	0	0	0	13	1	N/A	0/0	False	False
covid1	1	0	2	0	0	1	1	2/2	False	False
energy1	1	0	4	0	0	1	1	4/4	False	False
yahoo	1	0	101	0	0	1	1	101/101	False	False
india	1	0	2	0	0	1	1	2/2	False	False
exchange	1	0	8	0	0	1	1	8/8	False	False
covid2	2	0	2	0	1	1	1	2/2	True	True
wage	2	0	12	0	0	1	1	12/12	True	True
market	2	0	5	0	0	1	1	5/5	True	True
avocado	2	0	11	0	0	1	1	11/11	True	True
suicides	2	0	6	0	0	1	1	6/6	True	True

Conclusions

- We proposed an approach to classify datasets based on whether they contain temporally dependent data
- Our approach could identify temporal sequences when sequence corresponds to the entire dataset, and also when grouping by attributes was needed

Future Work

- We will investigate whether different types of correlations can be used within the Ljung-Box or Box-Pierce test
- We will conduct a deep analysis on which autocorrelation function to use when no prior information on data is known

THANK YOU!

hhomayouni@sdsu.edu

Backup Slides

Autocorrelation

A metric to determine whether a dataset has temporal dependence

- A measure of similarity of observations at certain lag
- Correlation of series with a delayed copy of itself

Analysis of Metrics: An Example

	date	county	cases	deaths
	2020-01-21	Snohomish	1	0.0
	2020-01-22	Snohomish	1	0.0
	2020-01-23	Snohomish	1	0.0
	2020-01-24	Cook	1	0.0
	2020-01-24	Snohomish	1	0.0
	...	...	...	...
	2021-02-02	Sweetwater	3510	33.0
	2021-02-02	Teton	3151	7.0
	2021-02-02	Uinta	1975	12.0
	2021-02-02	Washakie	867	26.0
	2021-02-02	Weston	611	5.0

date	county	cases	deaths
2021-01-29	Larimer	17914	192.0
2021-01-30	Larimer	17914	192.0
2021-01-31	Larimer	17914	192.0
2021-02-01	Larimer	18115	196.0
2021-02-02	Larimer	18160	198.0

date	county	cases	deaths
2021-01-29	Boulder	17225	232.0
2021-01-30	Boulder	17279	232.0
2021-01-31	Boulder	17329	232.0
2021-02-01	Boulder	17376	232.0
2021-02-02	Boulder	17433	232.0

In average it detected 1.94 attributes with autocorrelation when grouping by 'county'. That means that for some counties autocorrelation was not found in both numerical attributes.

	% data	groups	avg_temp_att	std	avg_corr	max_corr
date	0	0	NaN	NaN	0.000000	0.000000
county	95	1923	1.939158	0.239041	15.814377	18.813475
cases	0	196	1.000000	0.000000	0.616551	1.703790
deaths	1	314	1.000000	0.000000	0.833849	3.475165
no-grouping	100	1	0.000000	0.000000	0.000000	0.000000

The 95% indicates that for some counties (corresponding to 5% of the data) no autocorrelation was found in any attribute.

When grouping by 'county' there were 1923 different groups. When the database is not grouped, only one group is found.