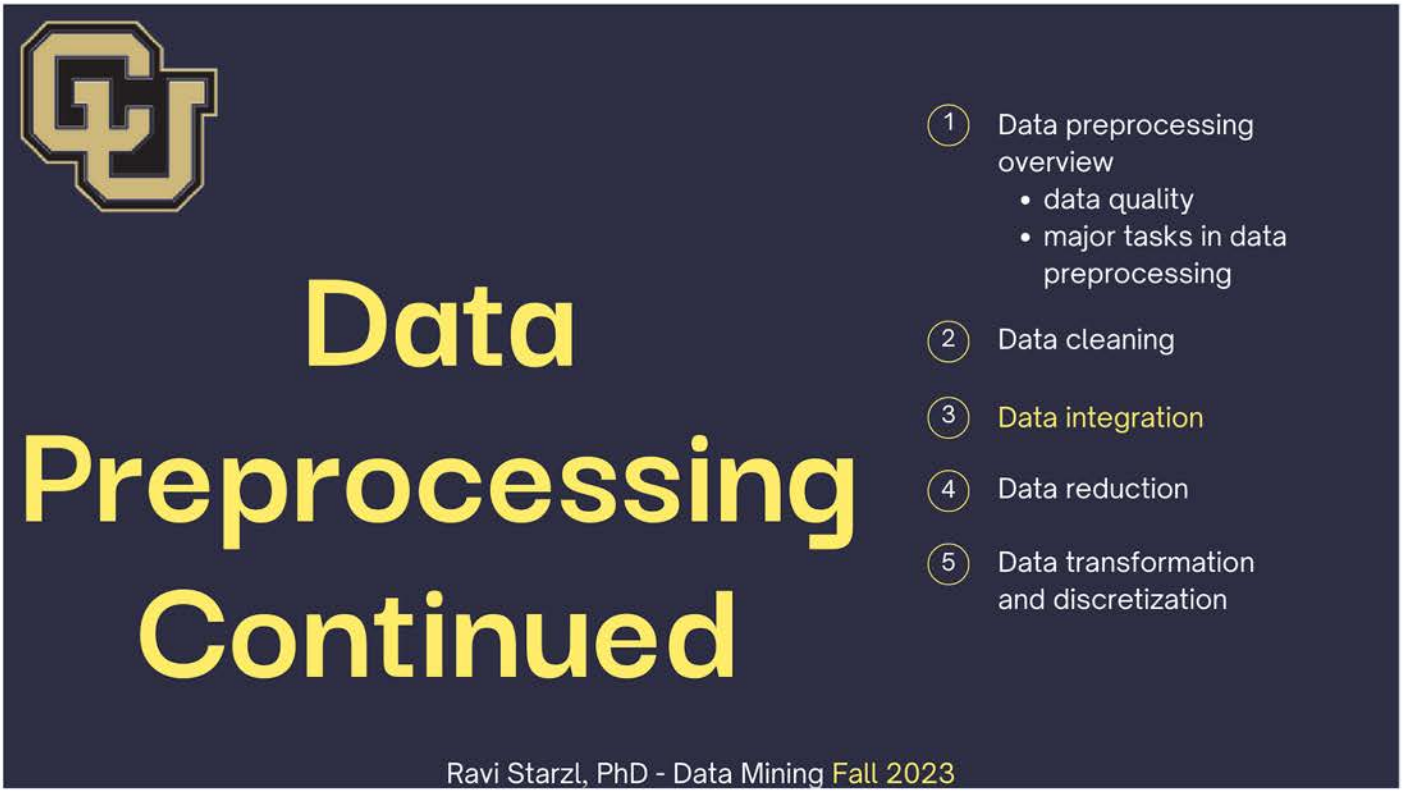
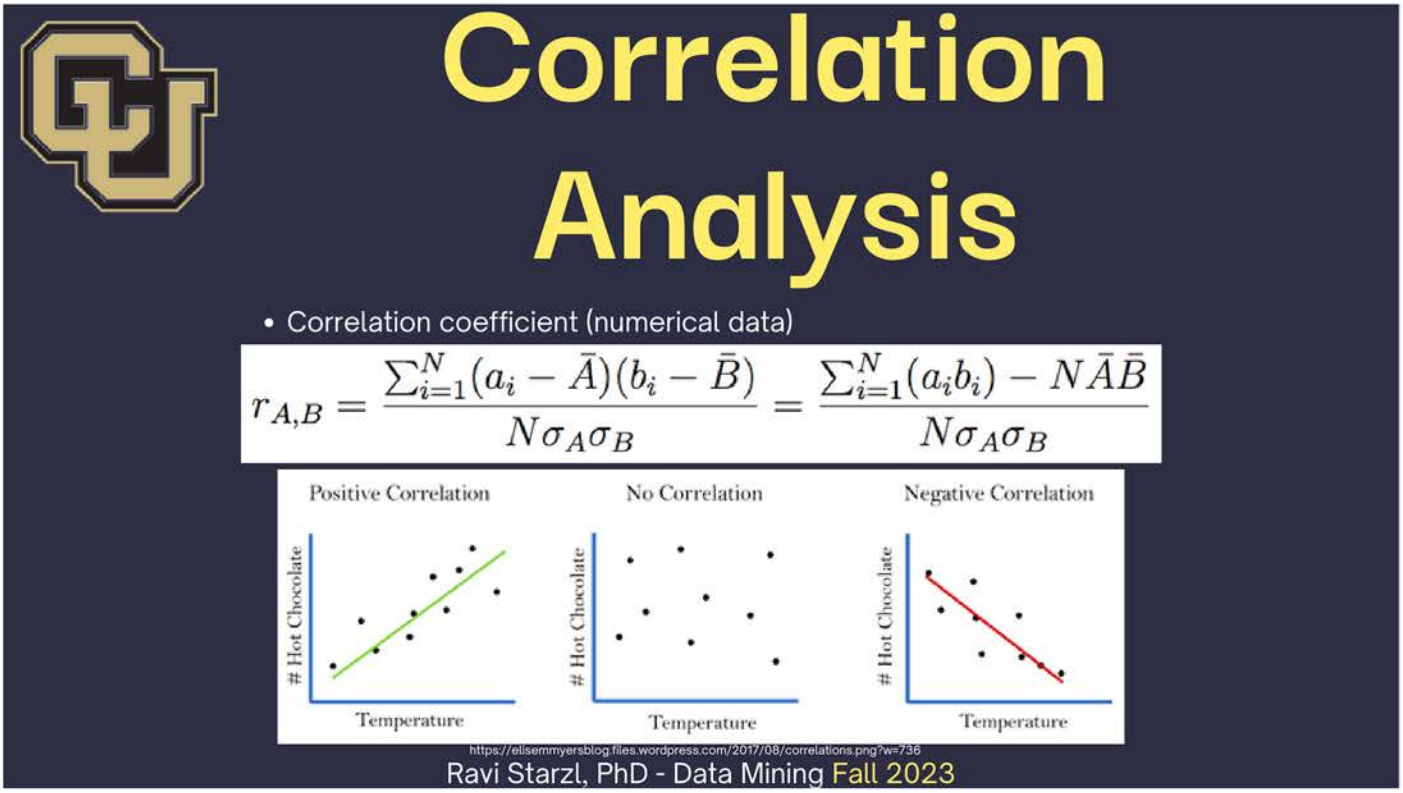
CSCI 4502/5502

Data Mining - Fall 2023 - Lecture 5

Ravi Starzl, PhD


Data Preprocessing Continued
1 Data preprocessing overview
• data quality
• major tasks in data preprocessing
2 Data cleaning
3 Data integration
4 Data reduction
5 Data transformation and discretization
Ravi Starzl, PhD - Data Mining Fall 2023



Correlation analysis studies the statistical relationship between two variables, also known as the association between them. This relationship can be quantified using correlation coefficients. The Pearson correlation coefficient, denoted r, is one of the most commonly used correlation coefficients for numerical data.

The Pearson correlation coefficient is calculated as the covariance of two variables divided by the product of their standard deviations. One mathematical formulation is shown here.

Where ai and bi are individual data points, Amean and Bmean are the mean values of each variable, and N is the number of data points.

This coefficient ranges from -1 to 1. A positive value indicates a positive correlation, meaning as one variable increases, the other also tends to increase. A negative value indicates a negative correlation, meaning as one variable increases, the other tends to decrease. A value of 0 indicates no correlation between the two variables.

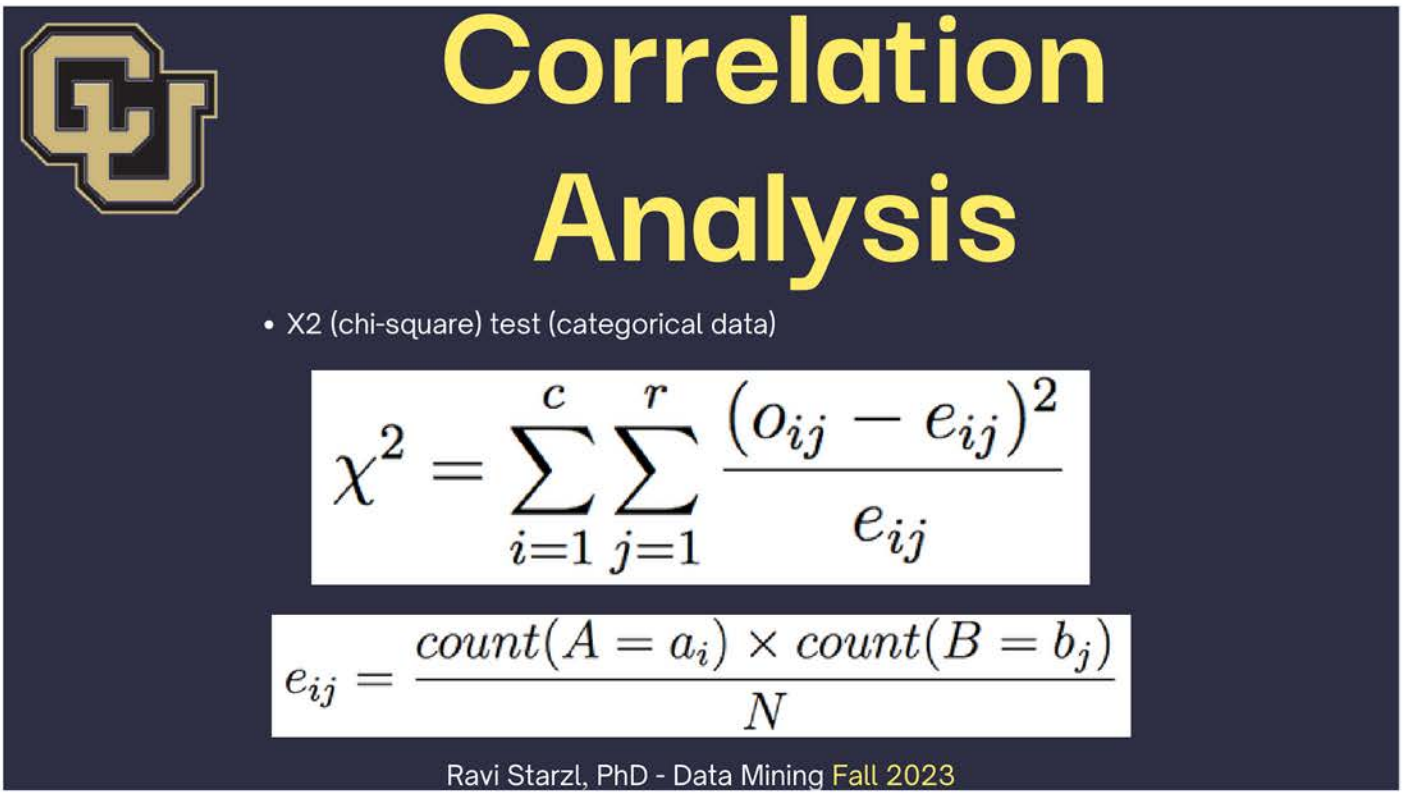
Some examples help illustrate correlation concepts:

- A positive correlation exists between height and weight - as height increases, weight also tends to increase. On a scatter plot, the data points would lean from bottom left to top right.

- No correlation exists between shoe size and intelligence - there is no reason to expect these variables are related. The data points on a scatter plot would show no clear trend.

- A negative correlation exists between time spent watching TV and time spent exercising - as TV time increases, exercise time tends to decrease. The data points would lean from top left to bottom right.

Importantly, correlation does not imply causation. Just because two variables are correlated does not mean changes in one cause changes in the other. This distinction is crucial to avoid invalid conclusions.



With categorical data, correlation can be analyzed using the Chi-Square Test instead of the Pearson correlation coefficient used for numerical data. The Chi-Square Test determines if there is a statistically significant association between two categorical variables. It is commonly used in fields like medical research and social sciences.

The analysis begins by setting up a contingency table tabulating the frequencies of each combination of the categories of the two variables. Each row represents a category of one variable, and each column represents a category of the other variable. The cell values are the observed frequencies of each combination.

The chi-square statistic is calculated as:

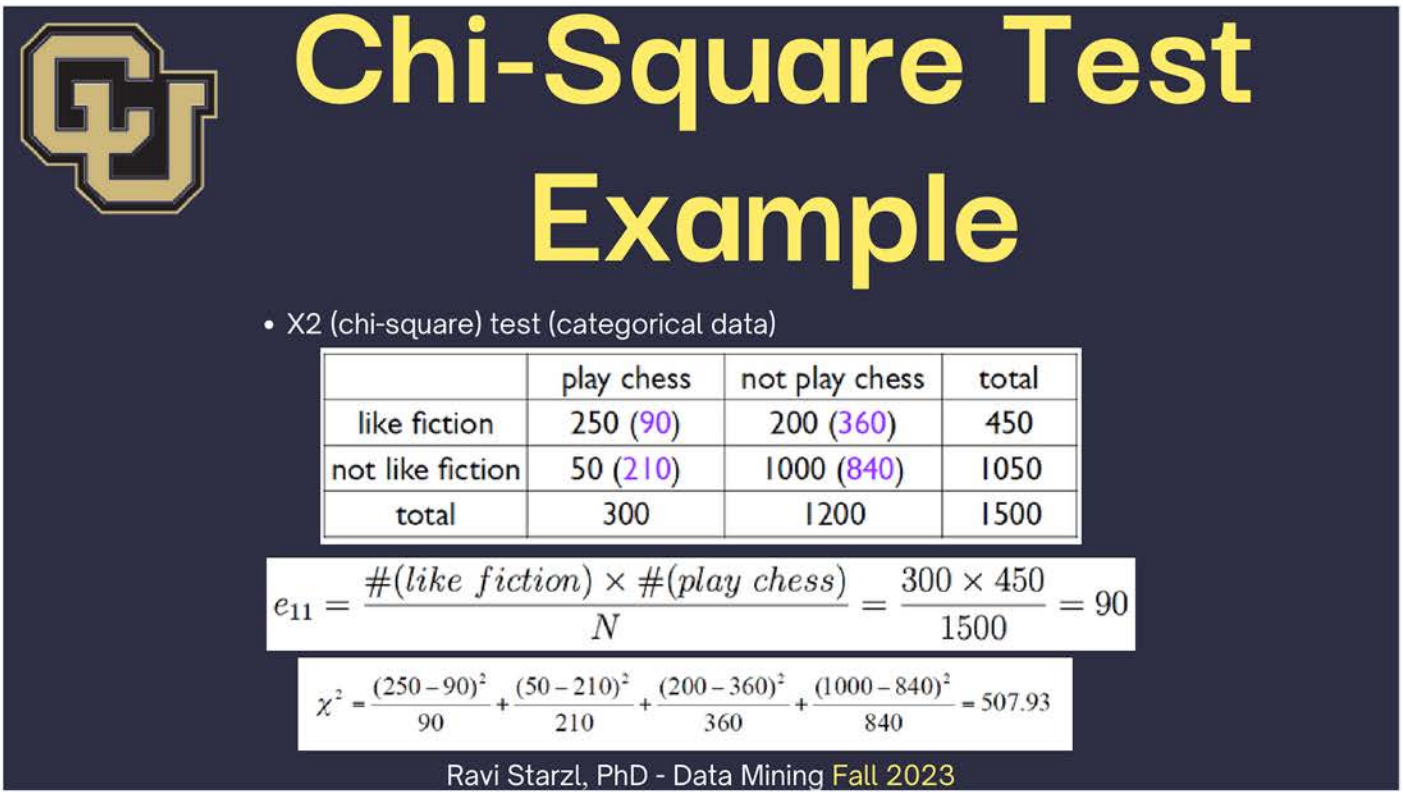
X^2 = ∑ [(Oij - Eij)^2 / Eij]

Where Oij is the observed frequency of a cell and Eij is the expected frequency, calculated as (Row Total * Column Total) / Grand Total.

This compares the observed frequencies to the expected frequencies if the variables were independent. The X^2 value quantifies the divergence between observed and expected.

The computed X^2 is compared to a critical value from the chi-square distribution with (r-1)(c-1) degrees of freedom, where r and c are the number of rows and columns. If X^2 exceeds the critical value, the null hypothesis of no association is rejected, and an association is concluded.

Importantly, as with the correlation coefficient, an association does not necessarily imply causation. Further investigation would be required to determine if one variable causes changes in the other. The chi-square test quantitatively assesses the association between categorical variables.



The chi-square test determines if there is a statistically significant difference between observed frequencies and expected frequencies in categorical data.

For example, let's observe whether individuals who like fiction are more likely to play chess. In a sample of 1500 people, 450 like fiction and 1050 do not. Of the 1500, 300 play chess. We can tabulate this in a 2x2 contingency table:

	Like Fiction	Don't Like Fiction	Total
Play Chess	250 (O)	50 (O)	300
Don't Play Chess	200 (O)	1000 (O)	1200
Total	450	1050	1500

Expected frequencies (E) are calculated assuming independence:
(Row Total x Column Total)/Overall Total

For example, expected who like fiction and play chess:
(450 x 300)/1500 = 90

The chi-square statistic is calculated as:
X^2 = Σ [(O - E)^2/E]

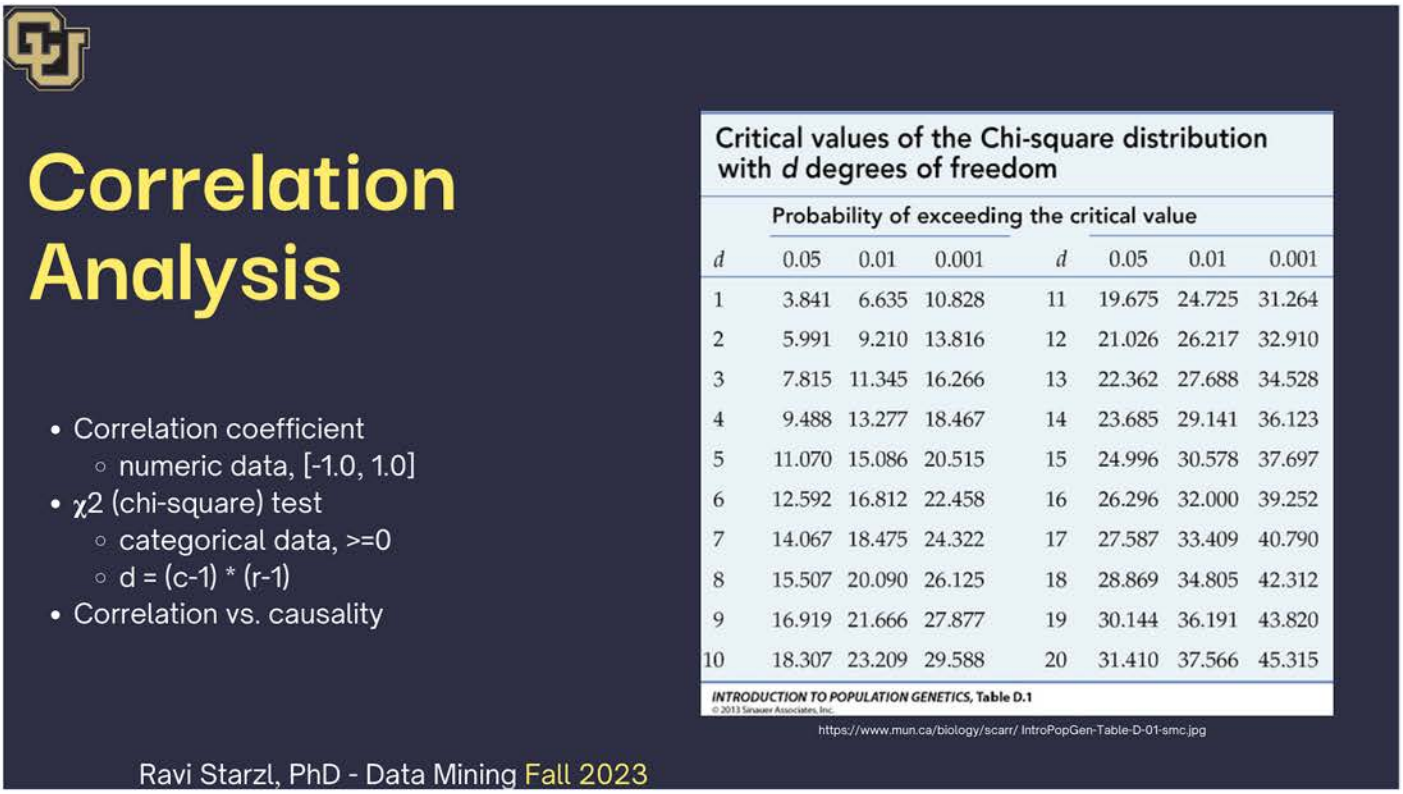
Plugging in the values:

$X^2 = (250 - 90)^2/90 + (200 - 360)^2/360 + ...$
$X^2 = 507.93$

This value is compared to a critical value from the chi-square distribution with (rows-1) (columns-1) degrees of freedom.

If X^2 exceeds the critical value, we reject the null hypothesis that the variables are independent.

A high X^2 statistic implies the observed frequencies differ significantly from expected frequencies, suggesting a relationship between the variables. However, it does not definitively establish causality. The chi-square test assesses the statistical independence of categorical variables.



The correlation coefficient and chi-square test are two main techniques used in correlation analysis of numeric and categorical data respectively.

The Pearson correlation coefficient quantifies the linear relationship between two continuous numeric variables. It ranges from -1 to +1, with values closer to the extremes indicating stronger positive or negative correlations. The correlation coefficient is computed as the covariance of the two variables divided by the product of their standard deviations.

The chi-square test evaluates the association between two categorical variables. It compares observed and expected frequencies within contingency table cells. The expected frequencies are calculated under a null hypothesis of independence. Higher chi-square values indicate a greater divergence of observed frequencies from expected frequencies.

For a chi-square test, the degrees of freedom equal (number of columns - 1) x (number of rows - 1). This affects the critical value from the chi-square distribution used to determine statistical signifi cance.

Importantly, correlation does not necessarily imply causation between variables, even when correlations are strong. Causality requires further statistical modeling and analysis.

In a chi-square test, the computed statistic is compared against the critical value based on the degrees of freedom. If the test statistic exceeds the critical value, the null hypothesis of independence can be rejected at the chosen significance level.

Additional relevant techniques in correlation analysis include logistic regression for predicting categorical outcomes, and various multivariate correlation methods for exploring relationships between larger sets of variables.

Techniques like the correlation coefficient and chi-square test quantify statistical relationships in data mining and analytics. But correlation-based methods alone cannot definitively determine causative relationships between variables. A variety of more advanced modeling techniques are required for causal inference.

The correlation coefficient and chi-square test quantify statistical relationships between variables, but correlation does not necessarily imply causation.

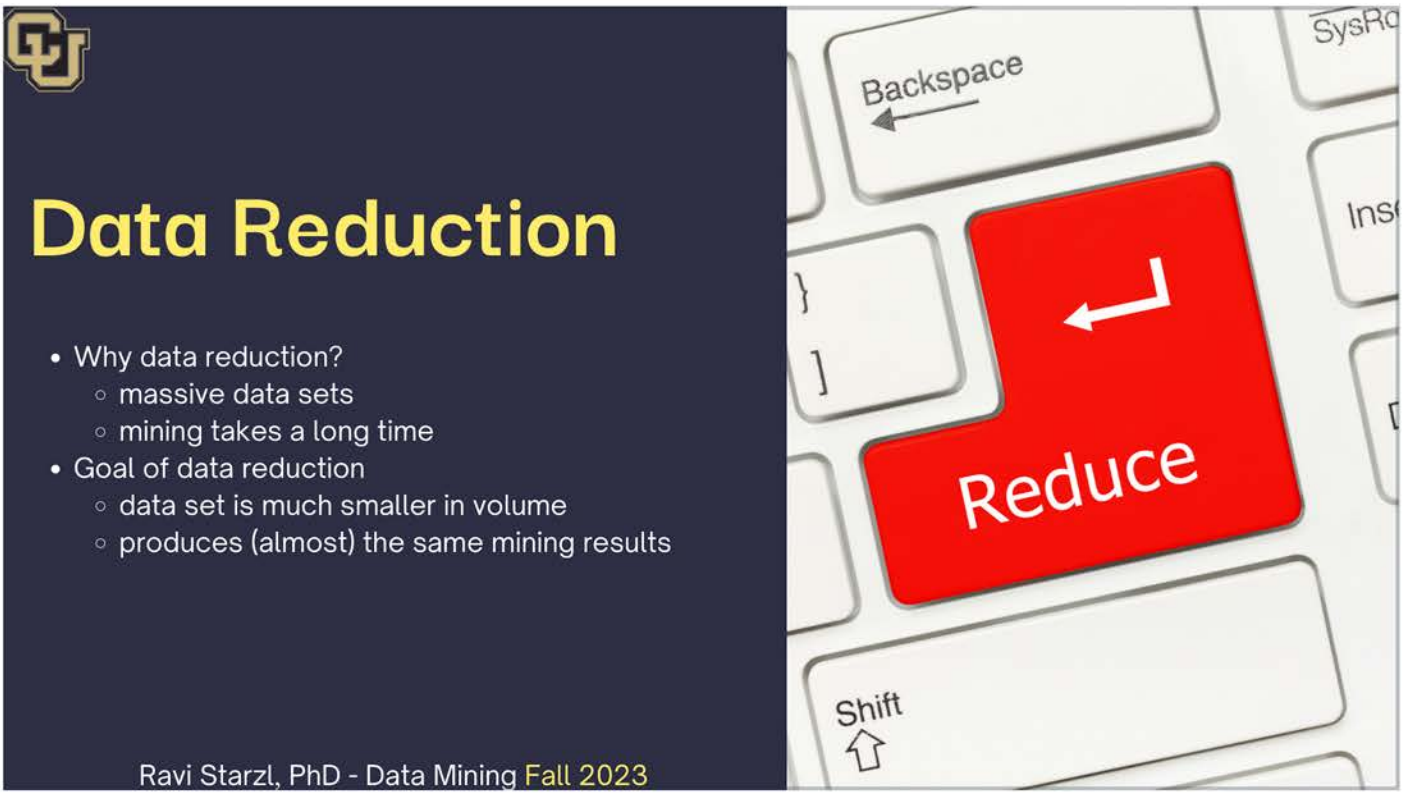
For example, variables may be correlated because of a lurking or confounding variable affecting both. Firefighter presence correlates with fire damage, but does not cause it - the underlying fire intensity drives both. Ice cream sales and drowning deaths correlate, but are both driven by hot weather rather than causing each other.

Observing a correlation should be the starting point for investigating relationships between variables, not the ending point. Further analysis through experiments, statistical modeling, or identification of confounders is required to prove causation. Correlation suggests connections between variables that warrant further examination, but does not definitively determine causality on its own.

Additional relevant techniques in causal analysis include regression models incorporating instrumental variables, interrupting time series analysis, and sophisticated causal modeling algorithms. But a key first step is always recognizing that correlation alone is insufficient to determine causation.

In data science, it is crucial to avoid making unsupported predictions or conclusions based solely on observed correlations. While correlation analysis provides value in discovering variable relationships, caution must be taken to not mistake correlation with causation. This distinction remains fundamental in statistics and data analysis, allowing for deeper understanding of data.

Now shifting to data reduction techniques...



Data reduction techniques are crucial when working with massive datasets, in order to improve efficiency without losing critical information.

The goal is to produce a reduced dataset that is much smaller in volume, yet yields similar analytical results when mined. This is extremely valuable in modern data mining, given the vast volumes of data available today.

Methods of data reduction include:

- Principal component analysis (PCA) - identifies the most meaningful basis to re-express data at lower dimensionality
- Data sampling - extracting a subset representative of the full dataset
- Feature selection - selecting the most relevant variables
- Feature extraction - creating new features that summarize key information from the original varia bles

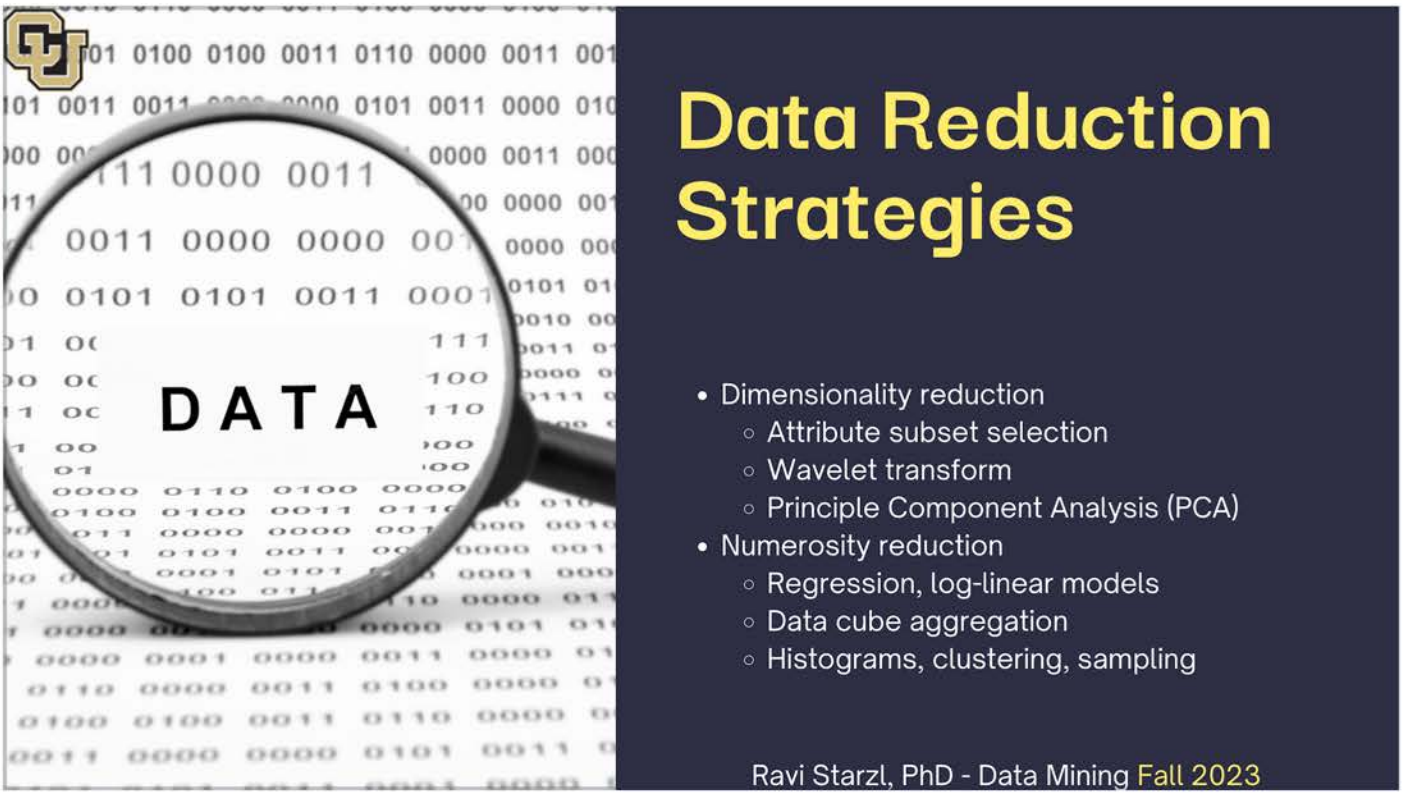
For example, in a dataset of millions of call records, dimensionality could be reduced by extracting features like 'peak hour usage' rather than analyzing time, duration and location separately.

However, care must be taken to avoid losing critical information or introducing bias during reduction. There is a tradeoff between efficiency and result accuracy that must be evaluated.

Additional relevant techniques include discretization to reduce continuous variables to categorical

levels, regression and clustering to identify representative data points, and online algorithms that process data sequentially.

In summary, data reduction techniques like PCA, sampling, feature selection and extraction enable efficient analysis of massive datasets. When applied judiciously, they reduce data volume while maintaining analytical integrity and usefulness for data mining.



Several main approaches for data reduction include dimensionality reduction, numerosity reduction, and data transformation techniques.

Dimensionality reduction removes redundant or irrelevant attributes. Methods include attribute subset selection to identify and retain the most significant attributes, wavelet transform to analyze data at different scales, and principal component analysis (PCA) to identify the most meaningful basis vectors capturing variance.

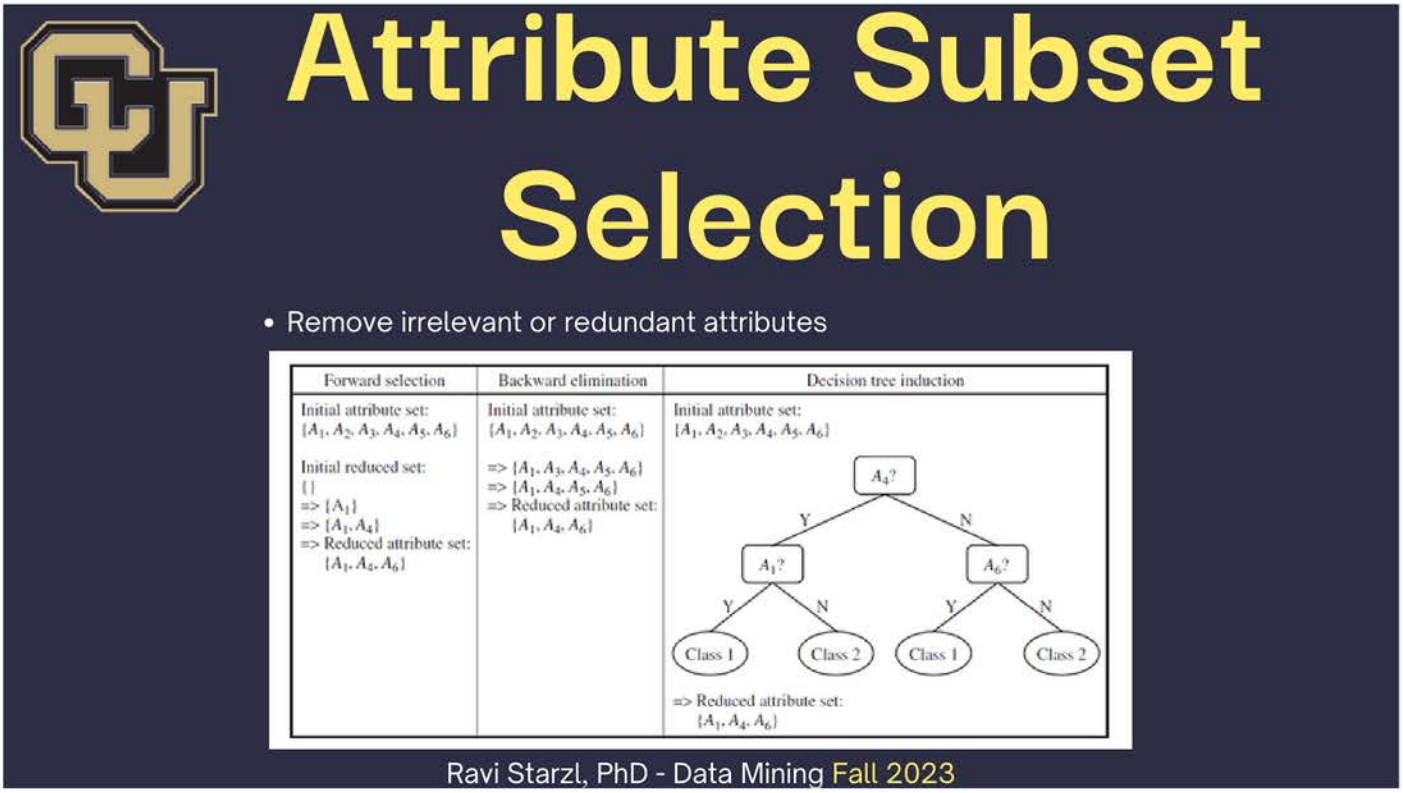
Numerosity reduction replaces data with more compact representations. Regression models the relationship between variables with an equation. Log-linear models represent associations between categorical variables compactly. Data cube aggregation summarizes data points into fewer aggregated points.

Data transformation techniques include histograms to represent data distributions graphically, clustering algorithms to group similar data points, and sampling to extract representative data subsets.

When applying these techniques, the goal is reducing volume while retaining the essence of the information. Tradeoffs between efficiency and accuracy must be evaluated.

Additional relevant methods include discretization to reduce continuous attributes to intervals, online algorithms that process data sequentially, and compression techniques like wavelet compression that reduce storage needs.

The optimal data reduction approach depends on the dataset characteristics, intended uses, and available computational resources. A combination of techniques is often required as part of an overall strategy to distill data to its key components, enabling efficient yet effective analysis.



Attribute subset selection removes irrelevant or redundant attributes to reduce dimensionality. Common techniques include:

- Forward selection: Starts with no attributes, iteratively adds attributes that improve model performance. Continues until additions no longer significantly improve performance.

- Backward selection: Starts with all attributes, iteratively removes attributes that least degrade or even improve performance. Stops when removals cause unacceptable performance loss.

- Decision tree induction: Builds a decision tree predictive model that implicitly performs attribute selection by choosing optimal attributes to split on at each node. Attributes never selected can be discarded.

These are examples of greedy algorithms that make locally optimal choices iteratively in hopes of finding a globally optimal attribute subset. However, they do not consider all possible subsets exhaustively.

Decision trees have advantages in avoiding overfitting via mechanisms like tree pruning and identifying important attributes directly from the final model structure.

For example, in a medical dataset with many factors to predict disease likelihood, forward/backward selection could incrementally add/remove attributes by measuring model performance at each step. Or decision tree induction could build a model that selects relevant

attributes like age, family history, smoking habits based on the splits chosen.

Attribute subset selection is a key data preprocessing technique for handling high-dimensional data by removing irrelevant and redundant attributes. This improves model performance, generalizability, and interpretability by focusing on the most important explanatory attributes.

The Discrete Wavelet Transform (DWT) is a technique that decomposes a signal into wavelet coefficients that can be reduced to lower dimensionality.

DWT provides both time and frequency information, giving a multi-resolution perspective. This is advantageous compared to Fourier Transform which provides only frequency information.

DWT breaks down a signal into wavelets - mathematical functions localized in time and frequency. This wavelet decomposition provides a way to represent the signal efficiently using fewer coefficients.

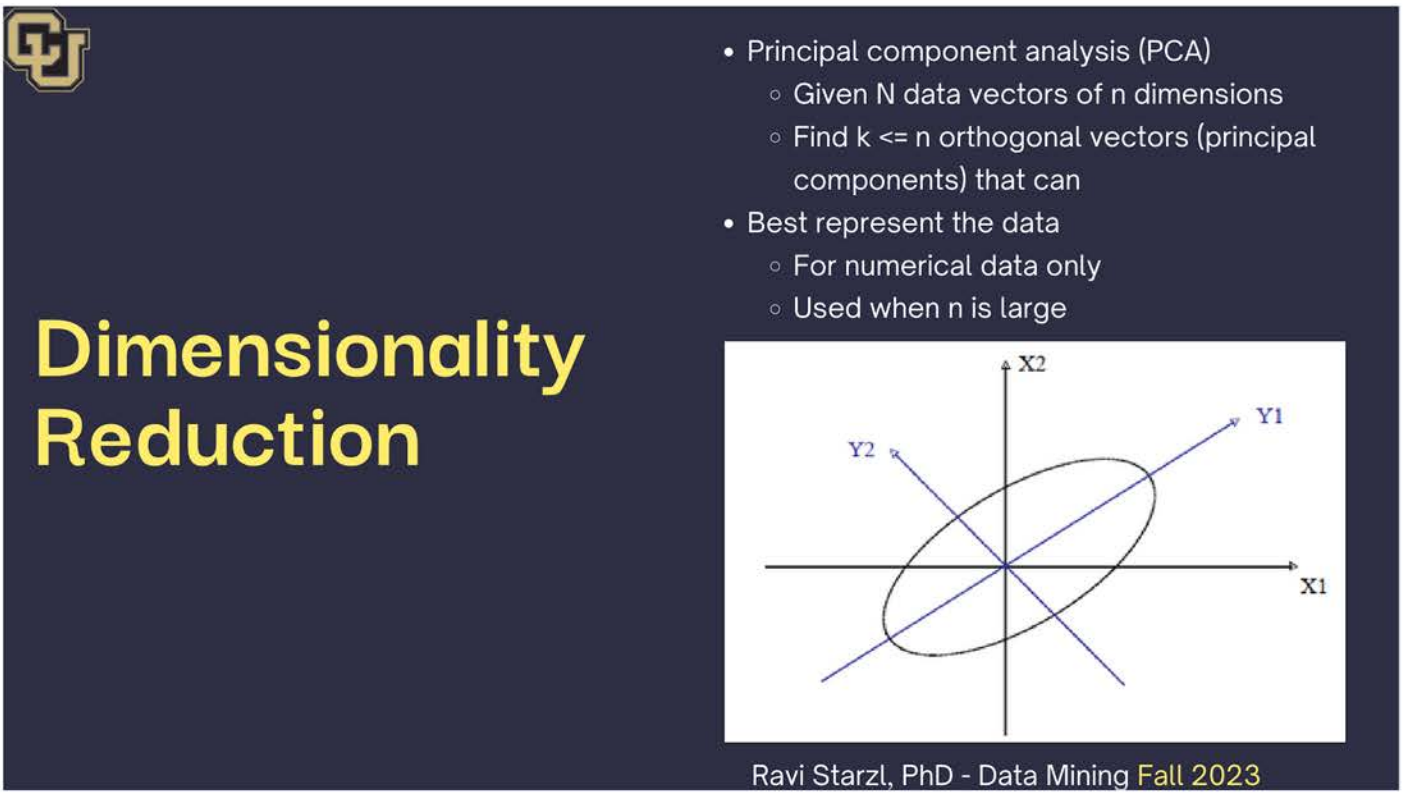
The strongest, most significant wavelet coefficients can be retained, while weaker ones are discarded. This allows approximation of the original signal with far fewer data points.

For example, an image can be decomposed into thousands of wavelet coefficients. Storing only a fraction of the strongest coefficients enables compact representation, data compression, and reconstruction of the image at lower resolution.

More coefficients retained means higher accuracy and detail, but even with just a subset, the essence of the original is preserved. DWT is thus a powerful dimensionality reduction technique.

Additional relevant mathematical details include the use of scaling and wavelet functions as basis vectors, multi-resolution analysis to examine data at different levels of detail, and entropy-based methods to identify optimal wavelet coefficients to retain.

DWT provides an efficient way to reduce dimensionality for data compression and analysis by transforming a signal into its salient wavelet components. This has major advantages for storage, transmission, and computational efficiency while minimizing information loss.



Principal Component Analysis (PCA) identifies new variables called principal components that better explain variance in multidimensional data.

Given a dataset of N vectors with n dimensions, PCA finds k ≤ n orthogonal vectors that represent the data's most significant patterns. This is useful when n is large, like in genetics, image processing, etc.

PCA rotates the data to align with axes of maximum variance. The first principal component is the direction of greatest variability. Further components are orthogonal and capture successive variances.

Fewer components can be retained to simplify the data. As components are linear combinations of original features, no information is discarded.

For example, with 2D data, PCA rotates the axes to align with the direction of maximum spread. Analyzing data per these new axes can reveal insights.

PCA assumes principal components are linear combinations of features. This may not hold in complex real-world data. Nonlinear methods like kernel PCA can then be advantageous.

In PCA leverages orthogonal linear transformations to uncover dominant patterns in multidimensional data. This enables simplification and analysis of key relationships otherwise obscured in high-dimensional spaces. PCA is thus a widely used technique for dimensionality redu

ction.

Additional mathematical details include the use of eigenvalue decomposition of the covariance matrix or singular value decomposition of the data matrix to derive the principal component vectors. Also, the proportion of total variance explained by each component can be examined to determine how many components to retain.

Numerosity reduction replaces data with more compact representations to improve efficiency and simplify analysis. Methods are categorized as parametric or non-parametric.

Parametric methods assume the data follows a model. The model's parameters succinctly summarize the data. For example, a normal distribution can be represented by just its mean and standard deviation. Parametric methods are efficient when model assumptions hold.

Non-parametric methods do not make assumptions about underlying data distribution. Examples i nclude:

- Histograms - Divide data into bins, store bin counts
- Clustering - Group similar data points, store cluster centers and assignments
- Sampling - Select representative data subset randomly or via stratified sampling

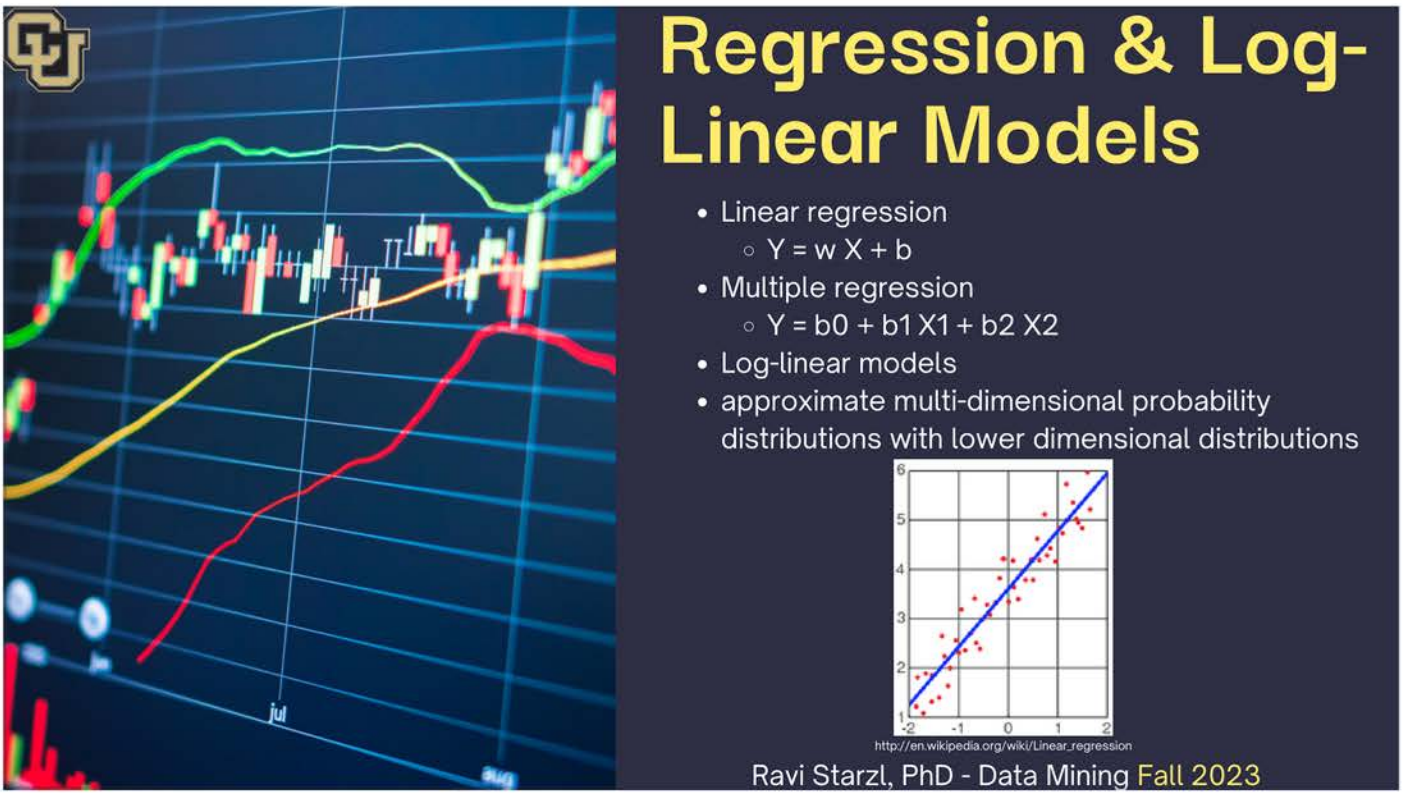
Non-parametric methods are flexible for diverse data types, but may be less efficient than parametric methods when model assumptions are valid.

Choosing between parametric and non-parametric approaches depends on:

- Data characteristics and distributions
- Validity of model assumptions
- Need to balance reduction size and preserving information

Additional relevant methods include regression models, log-linear models for categorical data, online aggregation algorithms, wavelet compression techniques, and sketching methods based on random projections.

Parametric and non-parametric numerosity reduction provides compact data representations to improve efficiency and combat the curse of dimensionality. The appropriate approach depends on the specific data mining task and resources available.



Regression models predict a dependent variable from independent variables. Linear regression expresses the relationship as Y = wX + b, where X is the independent variable, Y is the dependent variable, w is the coefficient, and b is the intercept.

Multiple linear regression incorporates multiple independent variables as Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn. Each b represents the coefficient of that variable.

For example, test scores (Y) could be predicted from study hours (X) in simple linear regression. Adding variables like health and stress levels leads to multiple regression.

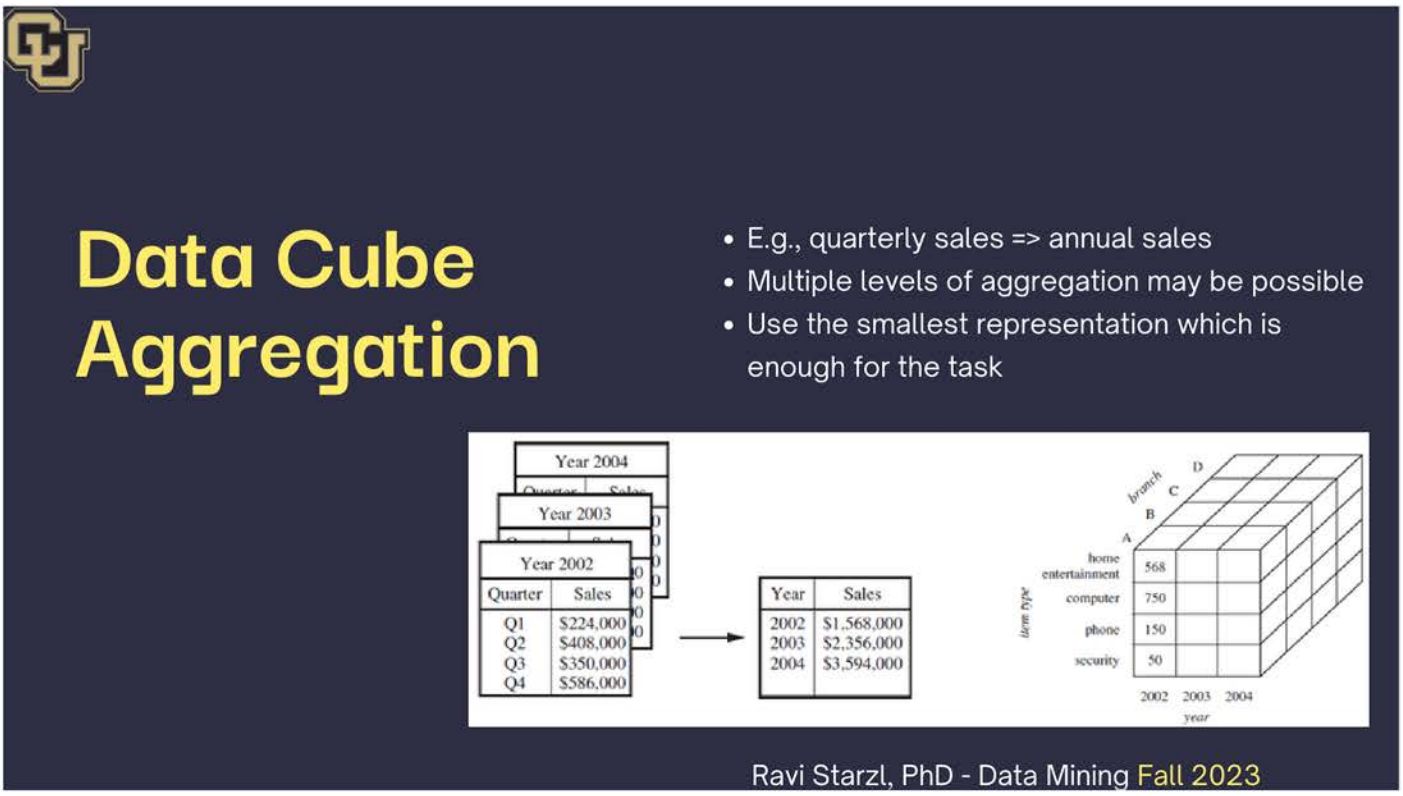
Log-linear models approximate multidimensional probability distributions of categorical data using lower-dimensional representations. They are commonly applied to model contingency tables and frequency data.

For instance, a log-linear model could capture interactions between the categorical variables vegetarian status, exercise habits, and smoking in a three-way contingency table. It predicts outcome probabilities.

Regression and log-linear models uncover relationships between variables, enabling prediction and probabilistic modeling from compact model representations rather than full datasets.

Assumptions should be evaluated, like linearity and normality for linear regression. Regularization methods and basis expansions can improve model flexibility. Validating on test data is key.

Regression and log-linear models provide effective numerosity reduction through predictive modeling of relationships between variables in statistical data analysis.



Data cube aggregation involves summarizing multidimensional data along different dimensions, similar to creating summary tables in a spreadsheet.

In a data cube, dimensions like product, location, time period are organized with measurements like sales. Aggregation sums up data points across chosen dimensions.

For example, a sales cube may contain monthly data. Aggregating this to a quarterly level sums the monthly sales into quarterly totals. Further aggregation could produce annual sales.

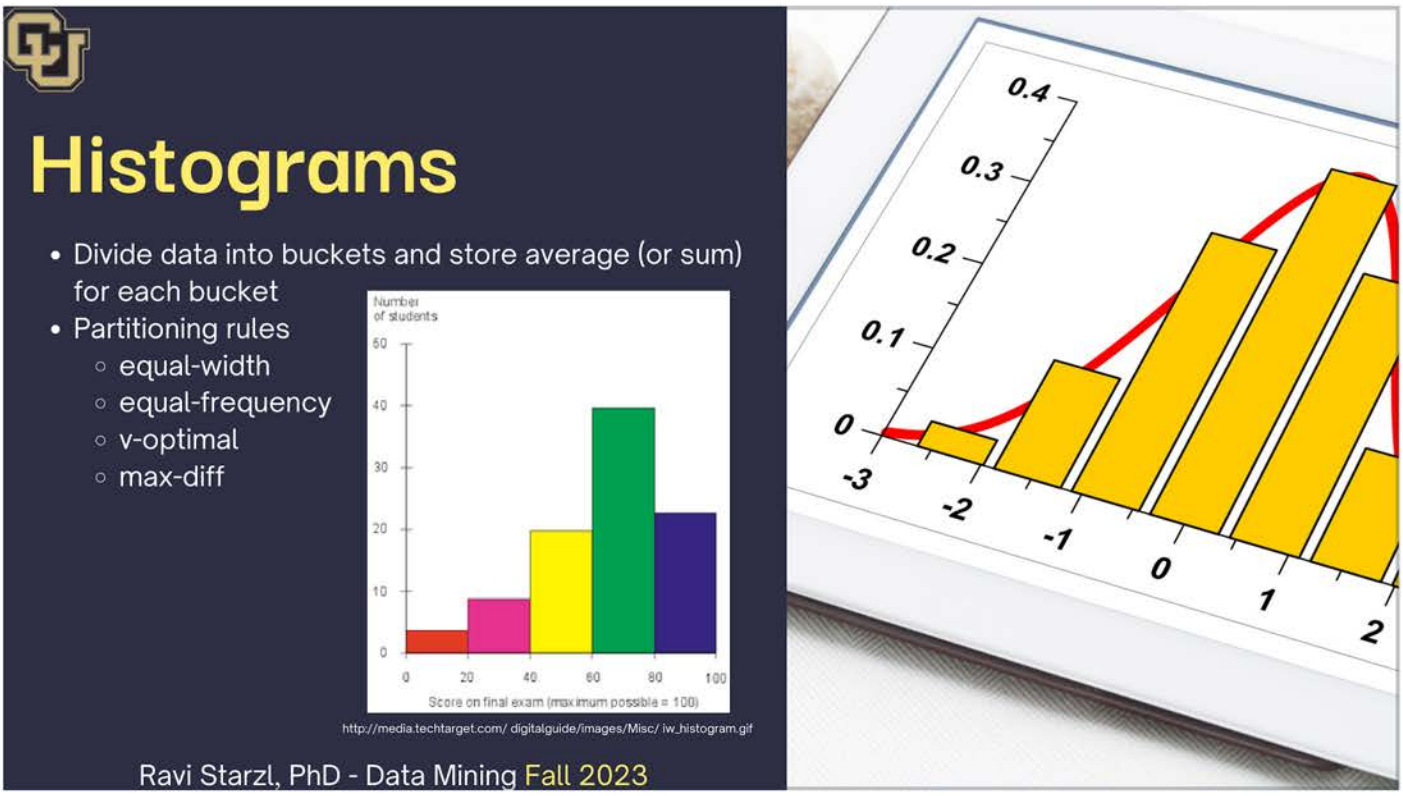
The goal is to use the most granular data necessary for the intended analysis. Aggregating quarterly to annual data discard details but allows focusing on broader yearly trends. More aggregation leads to more simplification but less fine-grained information.

Choosing the right aggregation level depends on the tradeoff between condensing the size and retaining interpretability. Higher levels like annual are appropriate for identifying macro trends while lower levels like monthly preserve details.

Caution is needed when aggregating to avoid incorrectly summarizing the data. Common aggregation techniques include sums, averages, counts, minimums, maximums. These must be applied carefully based on the metric.

Data cube aggregation reduces volume through selective summarization across dimensions. This improves efficiency but loses granularity. The optimal aggregation level provides a balance

between summarization and retention of details relevant for the analysis task.



Histograms provide a graphical summary of the distribution of data by partitioning values into "bins" and tallying frequencies. This serves as an effective data reduction technique.

There are various strategies for defining histogram bins:

- Equal-width bins have identical sizes based on range
- Equal-frequency bins contain an equal number of data points
- V-optimal bins minimize variance within each bin
- Max-diff bins are created at points of maximum change

The bins store summary statistics like the count or average of data points falling into that interval. This compactly represents the distribution while discarding individual data values.

Histograms enable data reduction by replacing raw data with bin counts or averages. They also inform sampling strategies by revealing the shape of the distribution. For skewed data, stratified sampling may be advised based on bin frequencies.

As an example, a histogram of normally distributed heights data may have equal-width bins showing the characteristic bell curve shape. The tallest frequencies are in the middle bins around the mean.

Histograms provide an intuitive visualization of data variability. However, choices like bin size and alignment can impact interpretation. Caution is required in their creation and analysis.

Histograms are a fundamental exploratory technique for concisely summarizing the distribution of data. Their ability to reduce dataset size while retaining distribution information makes them invaluable in statistics and data science.

Additional relevant details include cumulative histograms, 2D/3D histograms, histogram equalization in image processing, and kernel density estimation as a related non-parametric method.

Clustering involves partitioning data points into clusters based on a measure of similarity. It is an unsupervised learning technique used to uncover inherent groupings within a dataset.

A key step is selecting an appropriate similarity measure, such as Euclidean distance for numeric data. Points are grouped into clusters such that similarity between points within a cluster is maximi zed.

Once clusters are formed, only compact statistical representations like cluster centroids and diameters need to be stored, rather than all data points. This provides significant data reduction.

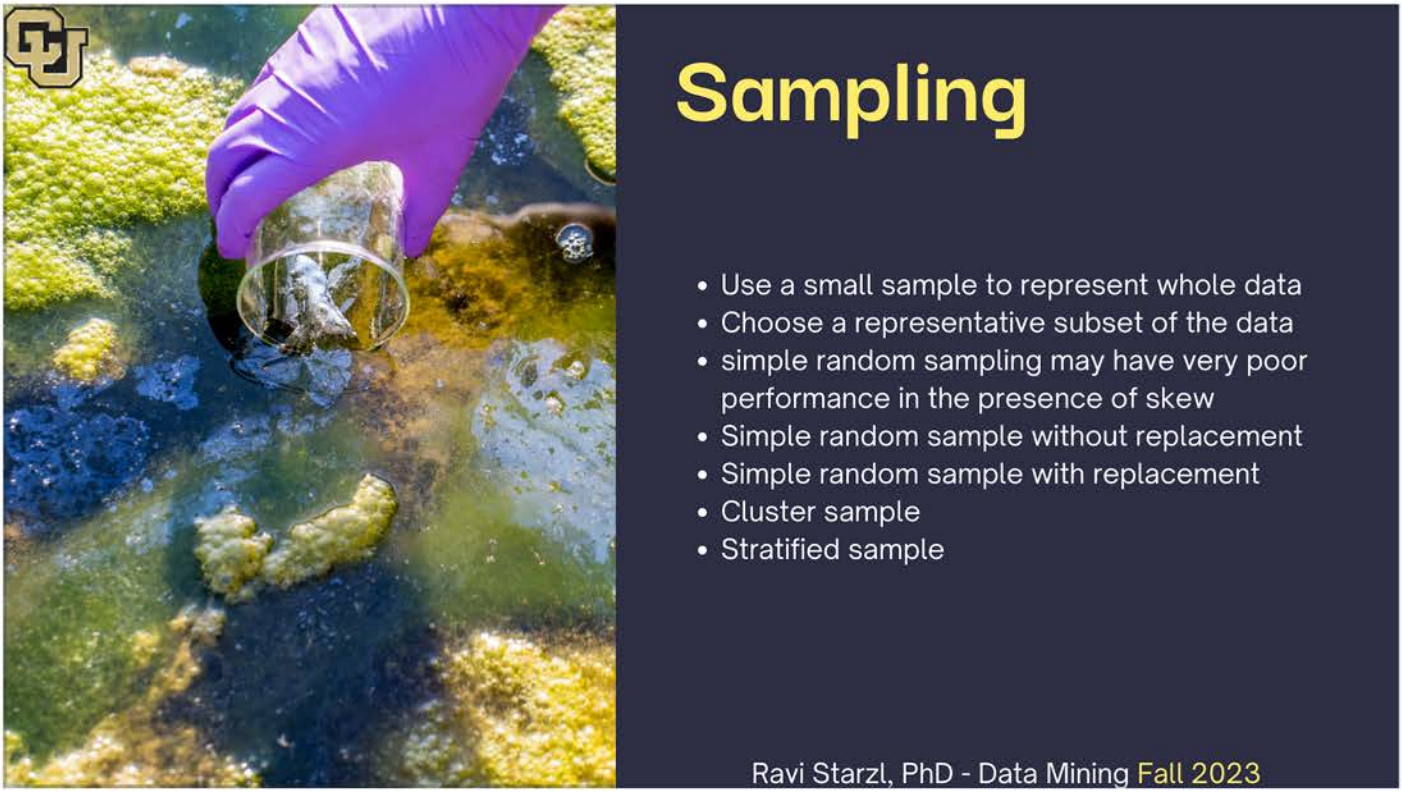
For instance, in a large customer dataset, clustering could identify groups of customers with similar demographics and purchase behaviors. Storing a summarized profile for each cluster reduces the data volume substantially compared to the full detailed records.

An advantage of clustering is the ability to create hierarchical clusterings, which produce a tree of merged or split clusters. This reveals relationships between clusters and can help determine the appropriate number of clusters.

There are many clustering algorithms to choose from, each with pros and cons. K-means clustering partitions data into k non-overlapping clusters. DBSCAN expands high density areas into clusters, useful when density varies. Hierarchical clustering builds a hierarchy of clusters through merging or splitting.

The success of clustering depends on the choice of similarity measure, clustering algorithm, and careful interpretation. Clustering uncovers inherent data groupings that can be leveraged for summarization and data reduction. But caution is required in application and analysis.

Clustering is an essential unsupervised learning technique for data reduction. By summarizing inherent clusters in the data through compact statistical representations, it condenses datasets for more efficient analysis and storage.



Sampling involves taking a subset of data points from a larger population in order to identify patterns or trends. There are various sampling methods, each with their own pros and cons.

Simple random sampling (SRS) randomly selects data points, either with or without replacement. SRS without replacement ensures independence of points but may underrepresent skewed data. SRS with replacement can improve representation in skewed data but allows duplicate points.

Cluster sampling divides the population into groups or clusters, then randomly selects some clusters. All observations in chosen clusters form the sample. This is efficient for large or widespread populations where surveying the entire population would be costly.

For example, a retail company surveying customer satisfaction nationally could cluster by state and survey all customers in sampled states.

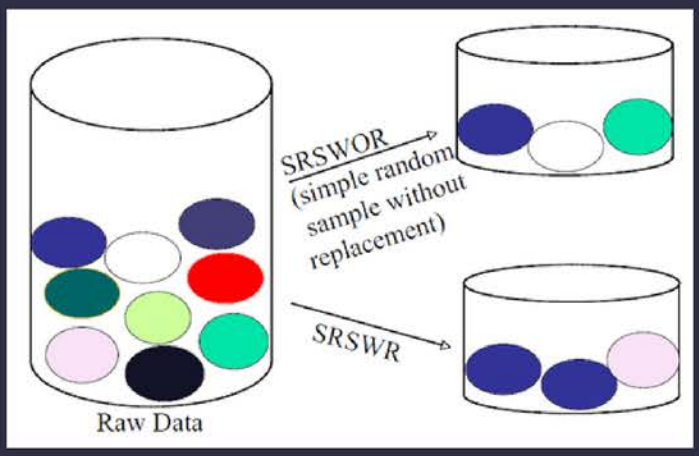
Stratified sampling divides the population into non-overlapping subgroups called strata, then samples randomly from each stratum. This aims to ensure representation of key subgroups.

A university surveying students may stratify by year – freshman, sophomore, etc. – then sample randomly within each stratum, capturing all years in the final sample.

Tradeoffs exist between sample size, cost, and representation. The optimal sampling method depends on the data structure, analysis objectives, and available resources.

Other techniques like systematic, multistage, and reservoir sampling have their own strengths and limitations. Overall, sampling reduces large populations down to more manageable representative datasets to enable efficient analysis.

Sample With or Without Replacement



Ravi Starzl, PhD - Data Mining Fall 2023

Sampling without replacement involves drawing data points from a population without returning them. The composition changes after each draw, affecting future probabilities.

For example, a jar with 20 beans - 5 red, 5 blue, 5 green, 5 yellow. Initially the chance of drawing a red bean is 25%. If a red bean is drawn and not replaced, the next draw has a 19 bean jar with only 4 reds, lowering the red probability to 21%.

In contrast, sampling with replacement returns each drawn data point to the population. Probabilities remain constant since the population composition is unchanged.

In the bean example with replacement, no matter how many draws, the chance of red stays at 25% - the jar always rebounds to 5 red beans out of 20 total.

Without replacement reflects real-world finite sampling but has complex changing probabilities. With replacement has simplifying independence assumptions with fixed probabilities.

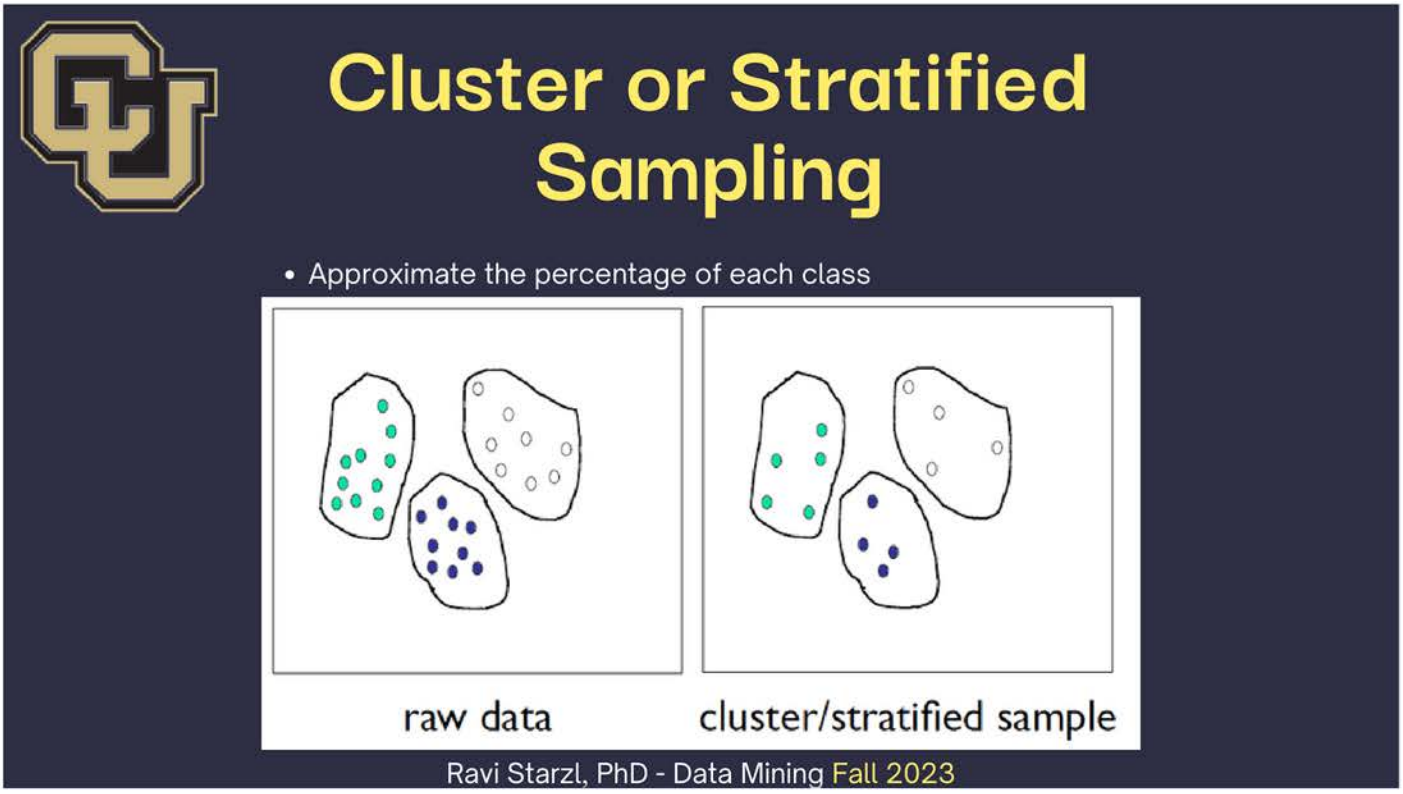
Without replacement suits studying proportions and large sampling fractions. With replacement suits independent events and near-infinite populations.

Tradeoffs exist between representativeness, computation complexity, and principles being modeled. The method should suit the study goals and population characteristics.

Additional details include sample size determination, calculating standard errors, and more

advanced methods like Markov Chain Monte Carlo sampling.

Key differences between sampling with and without replacement underlie many sampling strategy decisions in statistics.



Cluster sampling is used when the population is widespread, making simple random sampling impractical. The population is divided into clusters based on geography or other factors. A random sample of these clusters is selected and all observations within those clusters make up the sample.

For example, a national survey on diet could cluster by county. A subset of counties is randomly chosen, and all individuals in those counties surveyed. This simplifies data collection across a dispersed population but risks bias if the clusters are unrepresentative.

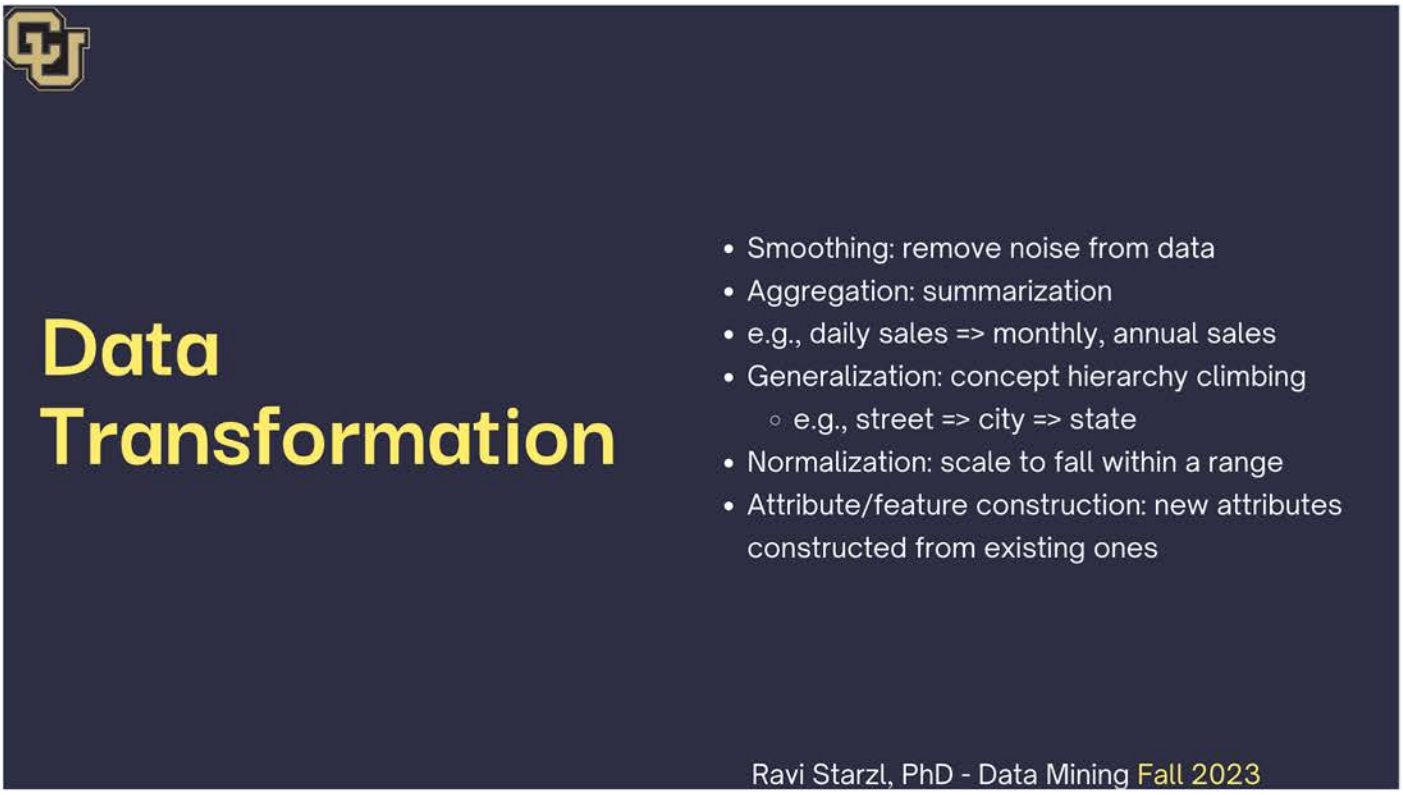
Stratified sampling divides the population into non-overlapping strata or subgroups based on characteristics like age, gender, income etc. A random sample is then drawn from each stratum.

A study on smoking habits may stratify by age group - 18-25, 26-35 etc. Random sampling is done within each age strata to ensure representation in the final sample. This requires knowing the population structure upfront.

Cluster sampling reduces logistical issues for large populations by consolidating data collection into selected clusters. But omitted clusters may skew results if not reflective of the overall populati on.

Stratified sampling aims to improve representation of distinct subgroups by guaranteeing their inclusion through stratified random sampling. It requires understanding the population composition to create appropriate strata.

The choice depends on the population distribution, costs, and the need to accurately represent certain segments. Both approaches improve simple random sampling for large heterogeneous populations.



Data transformation converts raw data into appropriate formats for analysis through various techniques:

- Smoothing removes noise and fluctuations to reveal fundamental patterns. For example, a moving average can smooth daily temperature data to emphasize broader seasonal trends.

- Aggregation combines detailed data into summary views, like aggregating daily sales data to monthly totals. This reduces volume while still providing key insights.

- Generalization moves up to a higher conceptual level, such as from street address to city to state. This reveals broader relationships not visible at lower levels.

- Normalization scales data to a standard range so variables on different scales contribute equally. Helps comparison and analysis.

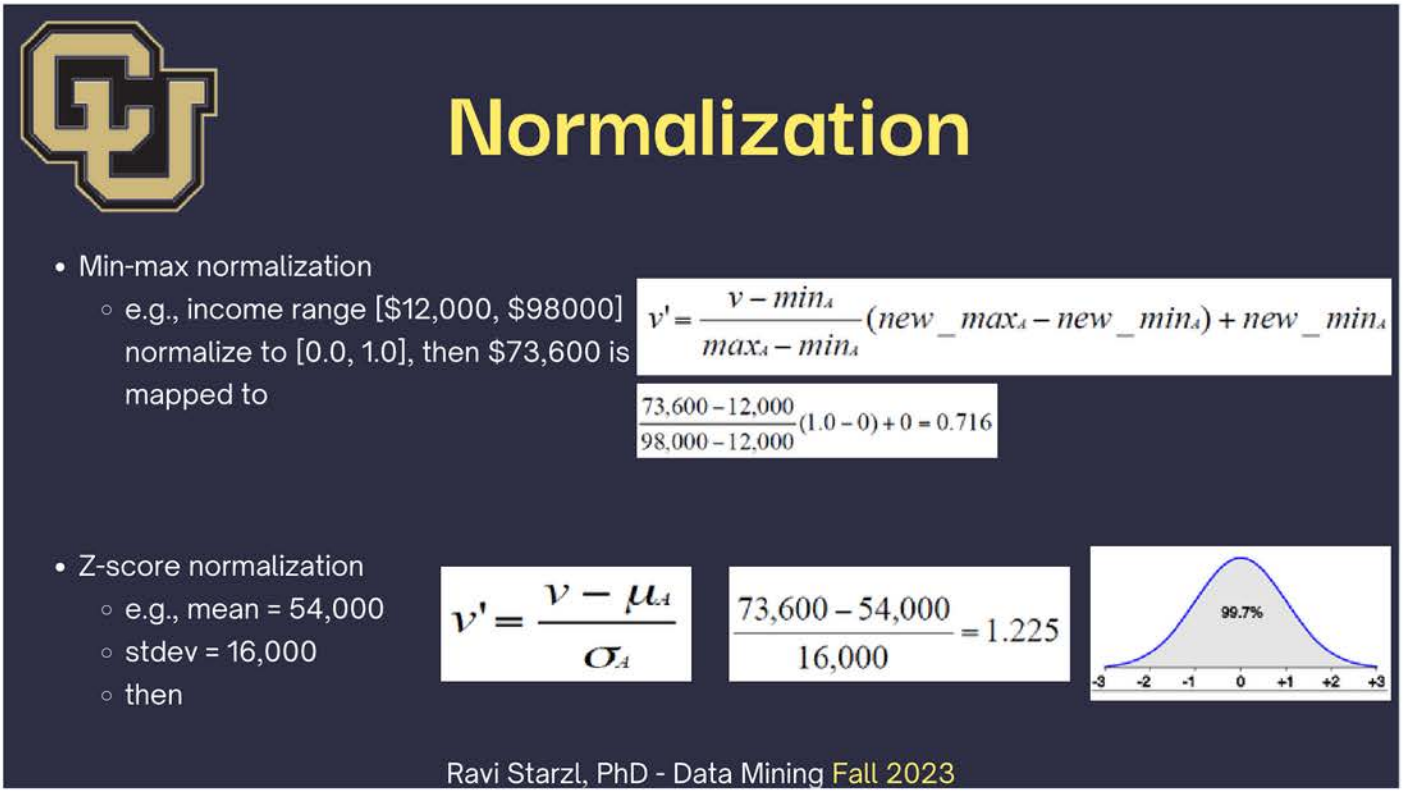
- Attribute/feature construction creates new attributes from existing ones that may be more useful for analysis. For instance, deriving price per square foot from price and size variables.

These techniques are often combined and tailored to the specific analytical needs and data characteristics. Key goals are removing noise, highlighting meaningful patterns, reducing data volume, and improving suitability for analysis.

Additional relevant techniques include discretizing/binning continuous variables, identifying and

removing outliers, imputing missing values, and integrating data from disparate sources.

Careful data transformation is a critical step in extracting insight from raw data. These methods form an essential part of the initial data preprocessing phase before analysis.



Data normalization rescales features to a common range to enable comparison and prepare data for analysis. Two main techniques are min-max normalization and Z-score normalization.

Min-max normalization transforms data to a fixed range such as [0,1] using the formula:

normalized_x = (x - min_x) / (max_x - min_x)

Where x is the value being normalized, min_x and max_x are the minimum and maximum of the dataset.

For example, normalizing an income value of $73,600 within a range of $12,000 to $98,000 results in approximately 0.72 on the 0-1 scale. Min-max normalization bounds all values to the same preset range.

Z-score normalization centers data around 0 with a standard deviation of 1 using:

$z = (x - \mu) / \sigma$

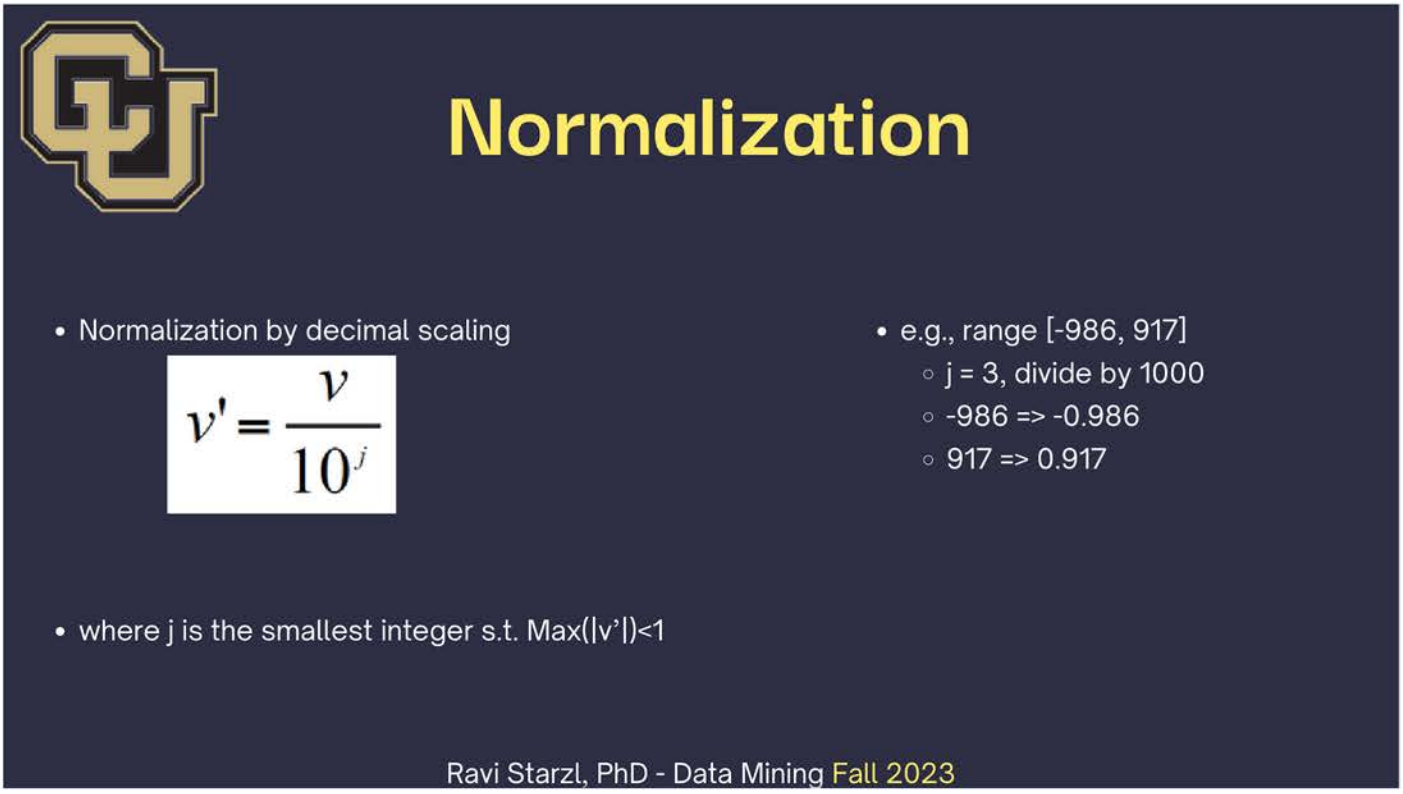
Where μ is the mean and σ is the standard deviation.

For instance, normalizing an income of $73,600 with mean $54,000 and standard deviation $16,000 yields a Z-score of approximately 1.23. This indicates the value is 1.23 standard deviations above the mean.

The choice of technique depends on whether a bounded range like 0-1 or standard deviation relationship is more useful for the analysis. Both transform differing scales to a common, standardized range.

Normalizing data is a crucial preprocessing step for many machine learning algorithms. Additional considerations include handling new data points and normalization of multivariate data.

min-max and Z-score normalization are useful techniques to standardize features to a common scale, enabling analytical comparability.



Decimal scaling is a normalization technique that scales data values such that the maximum absolute value is less than 1. This is done by dividing all values by 10 raised to the power of j, where j is the smallest integer such that the maximum absolute value is less than 1 after division.

The decimal scaling formula is:

v' = v / 10^j

Where v is the original data value, v' is the normalized value, and j is chosen such that Max(|v'|) < 1.

For example, given a temperature dataset with values [32, 45, 67, 90, 103]:

1) Determine j where 103/10^j < 1. In this case, j=3 since 103/10^3 = 0.103 < 1.

2) Apply decimal scaling formula:
- 32 -> 32/10^3 = 0.032
- 45 -> 45/10^3 = 0.045
- 67 -> 67/10^3 = 0.067
- 90 -> 90/10^3 = 0.09
- 103 -> 103/10^3 = 0.103

This scales the temperatures from 0.032 to 0.103.

Advantages of decimal scaling are preserving relative distances between data points and handling unknown data ranges. Drawbacks include difficulty handling outliers.

Choice of technique depends on data characteristics, range, distribution, and analysis needs. Decimal scaling is one option when range is unknown or data has multiple decimal places.

Other considerations include sparsity, skewness, dealing with new data points, and whether exact values or relative differences are more important.

In summary, decimal scaling is a normalization technique that divides by a power of 10 to scale values under 1. This transforms data to a common scale.

Discretization

- Three types of attributes
  - nominal: unordered set (e.g., profession)
  - ordinal: ordered set (e.g., military rank)
  - continuous: e.g., integer or real numbers
- Discretization
  - divide continuous range into intervals
  - interval labels used to replace data values
- supervised vs. unsupervised, split vs. merge

Ravi Starzl, PhD - Data Mining Fall 2023

Discretization is the process of converting continuous numeric attributes, like integers or real numbers, into discrete categorical intervals. It is useful when the raw data contains many unique values that can overwhelm machine learning models. Discretization transforms these into a smaller number of discrete bins or intervals, enabling more efficient analysis.

Discretization can be performed in a supervised or unsupervised manner. Supervised discretization utilizes the output variable to determine appropriate cut points for the intervals. This can help when the relationship between the attribute and output is complex. Unsupervised discretization only looks at the distribution of the attribute values themselves when creating intervals.

Another consideration is using a splitting versus merging approach. Splitting starts with the entire range as one interval, then splits it into smaller intervals based on certain criteria. Merging starts with each value as its own interval, and merges values into broader intervals according to some cri teria.

For example, discretizing income data from $20k to $120k:
- Splitting may start as one range, splitting it based on changes in income distribution
- Merging may start with each income as its own interval, merging based on similarity

Additional strategies include top-down versus bottom-up splitting/merging. Top-down works from the full range down to intervals, bottom-up starts from unique values up to intervals.

Benefits of discretization include simplifying complex continuous data and aiding model interpretability. Risks include potential information loss if intervals are improperly defined. The approach should suit the data characteristics and analysis objectives.

Discretization is an important data transformation technique that converts continuous data into discrete intervals or categories for improved machine learning performance.

Discretization simplifies analysis of data where continuous attributes have a large or infinite number of possible values. Common discretization techniques include:

Binning - An unsupervised splitting method that divides the range of a continuous attribute into a predefined number of equal-width bins or intervals. For example, ages from 1-100 could be binned into groups of 0-20, 21-40, 41-60, etc.

Histogram analysis - An unsupervised splitting method similar to binning, but uses the distribution of the data itself to determine the optimal bin ranges rather than equal widths. It analyzes histogram frequencies to identify natural cut points for binning.

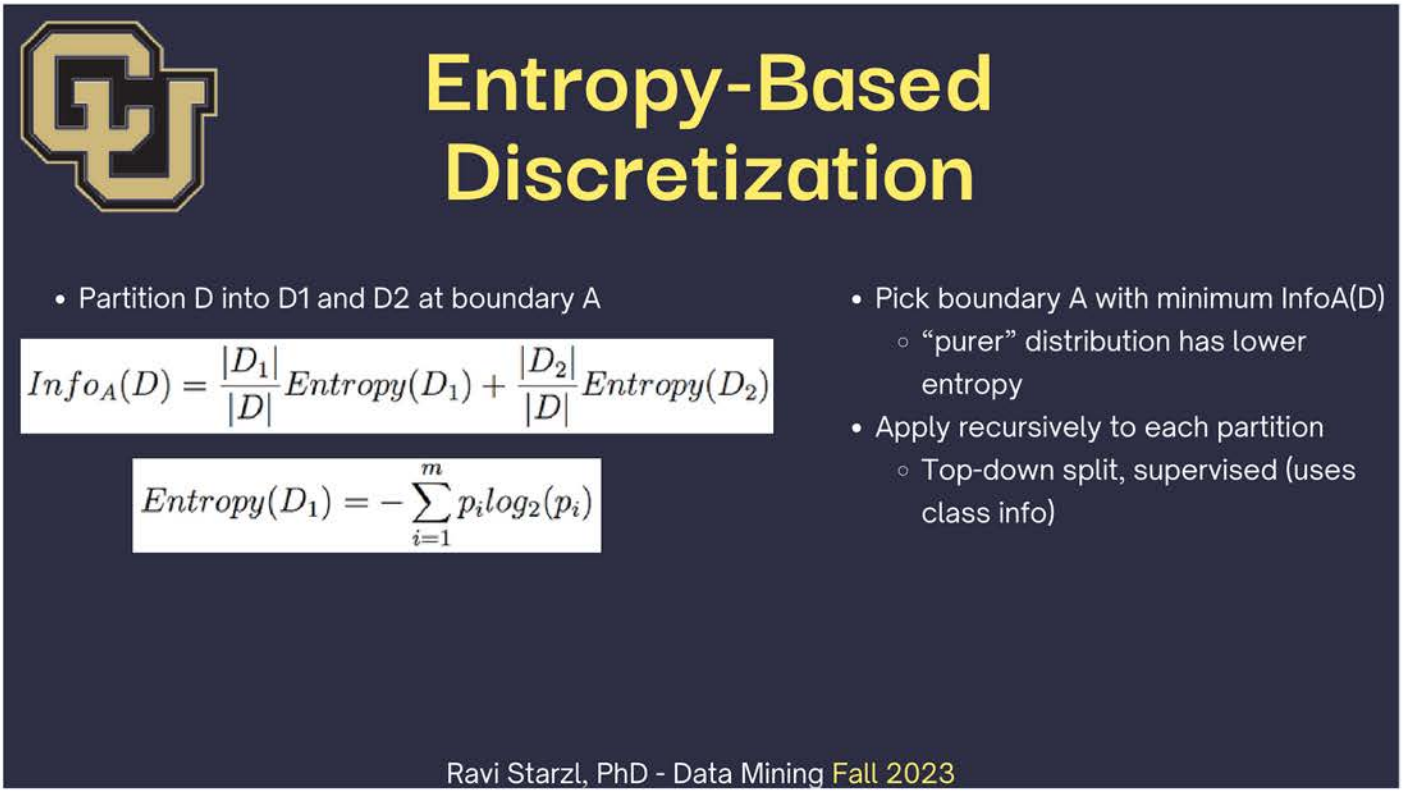
Clustering analysis - An unsupervised method that can split or merge based on inherent clusters in the data. For example, expenditure data may be clustered based on natural groupings of low, medium, and high spending levels rather than predefined splits.

Entropy-based discretization - A supervised splitting method that uses class entropy, a measure of disorder, to determine optimal split points that minimize class impurity. For instance, this could identify cholesterol level cut-offs for "high", "medium", and "low" that best distinguish diseased patients from healthy.

Chi-squared interval merging - A supervised merging method that starts with each unique value as its own interval, then recursively merges adjacent intervals that lead to a non-significant change in chi-squared value compared to the merged categories.

Intuitive partitioning - An unsupervised splitting method that relies on human domain expertise to determine appropriate splits. For example, ages could be divided into intuitive groups like "children", "adolescents", "adults", and "seniors".

The choice of discretization technique depends on the data characteristics, whether class labels are available, the intended uses of the data, and other factors. But used properly, discretization can simplify analysis of complex continuous data.



Entropy-based discretization is a supervised, top-down method for discretizing continuous attributes using class information to minimize entropy. Key steps include:

1. Calculating entropy (E) which measures impurity in a dataset D as:

E(D) = - ∑ p(ci) * log2 p(ci)

Where p(ci) is the proportion of examples belonging to class ci in D. Higher entropy means more disorder, or 'surprise'.

2. Calculating information gain, InfoA(D), which is the expected reduction in entropy after partitioning dataset D into D1 and D2 based on threshold boundary A:

InfoA(D) = (D1/D) * E(D1) + (D2/D) * E(D2)

3. Selecting the threshold boundary A that minimizes InfoA(D). This maximizes information gain, or reduction in entropy.

4. Recursively repeating process on partitions D1 and D2 to maximize purity.

For example, given subset D with 7 Class 1 and 3 Class 2 examples:

- Boundary A splits it into D1 (4 Class 1, 2 Class 2) and D2 (3 Class 1, 1 Class 2).

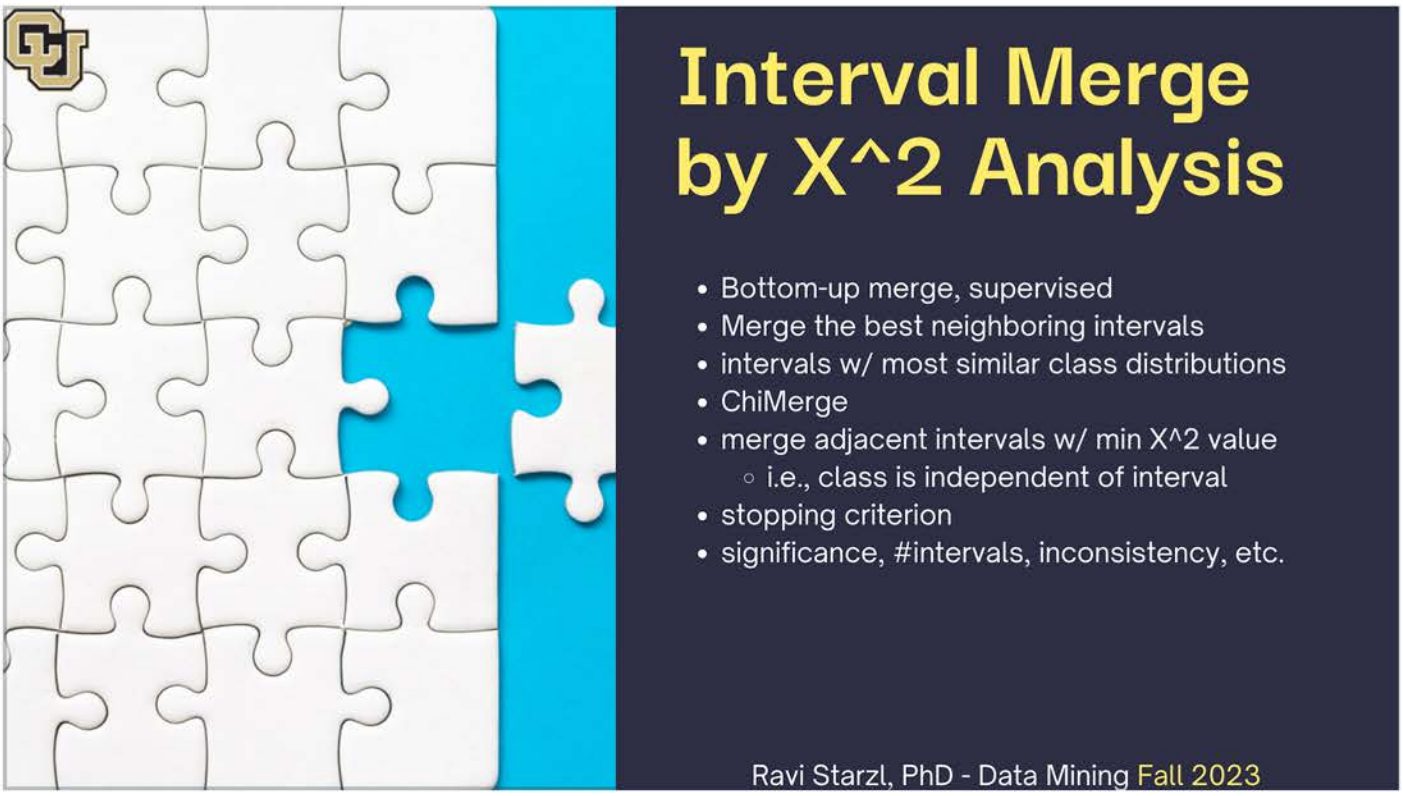
- Entropy E is calculated for D, D1, D2.

- InfoA(D) is computed as weighted average entropy after split.

- Process repeats, choosing thresholds that minimize InfoA(D) to maximize information gain.

This supervised approach leverages class labels to determine optimal splitting boundaries for discretization. Considerations include computational complexity and overfitting. Pre-pruning and stopping criteria can help.

Entropy-based discretization is a top-down supervised method that uses class entropy minimization to optimally discretize continuous attributes.



ChiMerge is a supervised, bottom-up discretization technique that uses chi-squared analysis to merge intervals of a continuous attribute based on class label distribution similarity.

It starts by initializing intervals as each unique value. The chi-square statistic is calculated for each adjacent interval pair and measures independence between the intervals and class label.

A low chi-square value means the class distribution is similar in both intervals. A high value means the distributions are very different.

The pair of intervals with lowest chi-square statistic is merged recursively. This continues until a stopping criterion is met, like maximum number of intervals, chi-square threshold, or statistical significance level.

For example, discretizing a continuous age attribute for loan default prediction:

- Start with initial intervals [25,25], [30,30], [35,35] etc. for each unique age

- Calculate chi-square statistic for each interval pair to measure class distribution differences

- Merge pair with lowest chi-square value, like [25,30] if ages 25 and 30 have similar default rates

- Recalculate chi-square values on new intervals and repeat merging process

- Stop when criteria met, resulting intervals become discretized age attribute

This uses class label information to merge adjacent intervals with similar class distributions, transforming the continuous attribute into a categorical attribute for improved analysis.

Considerations include computational complexity, handling sparse intervals, determining appropriate stopping criteria, and balancing interval granularity vs overfitting.

ChiMerge is a supervised discretization technique that leverages class information to determine optimal merging of intervals for a continuous attribute.

Concept hierarchy generation is the process of establishing levels of abstraction for categorical attributes. This enables analysis of categorical data at different granularities and simplifies identification of high-level patterns.

A concept hierarchy arranges categorical attributes from the most specific to the most general. For example, a geographic concept hierarchy might be:

Street Address < City < County < State < Country

Here, each attribute is more abstract and general than the one preceding it. This creates a hierarchy of increasing conceptual levels, allowing us to climb up to broader perspectives.

For categorical attributes that lack natural ordering, domain knowledge is required to manually define appropriate hierarchies. The geographic hierarchy relies on understanding of geographic relations to determine the logical attribute ordering.

Automatic methods can infer basic hierarchies by assuming attributes with fewer distinct values are more general levels. However, this heuristic is not foolproof.

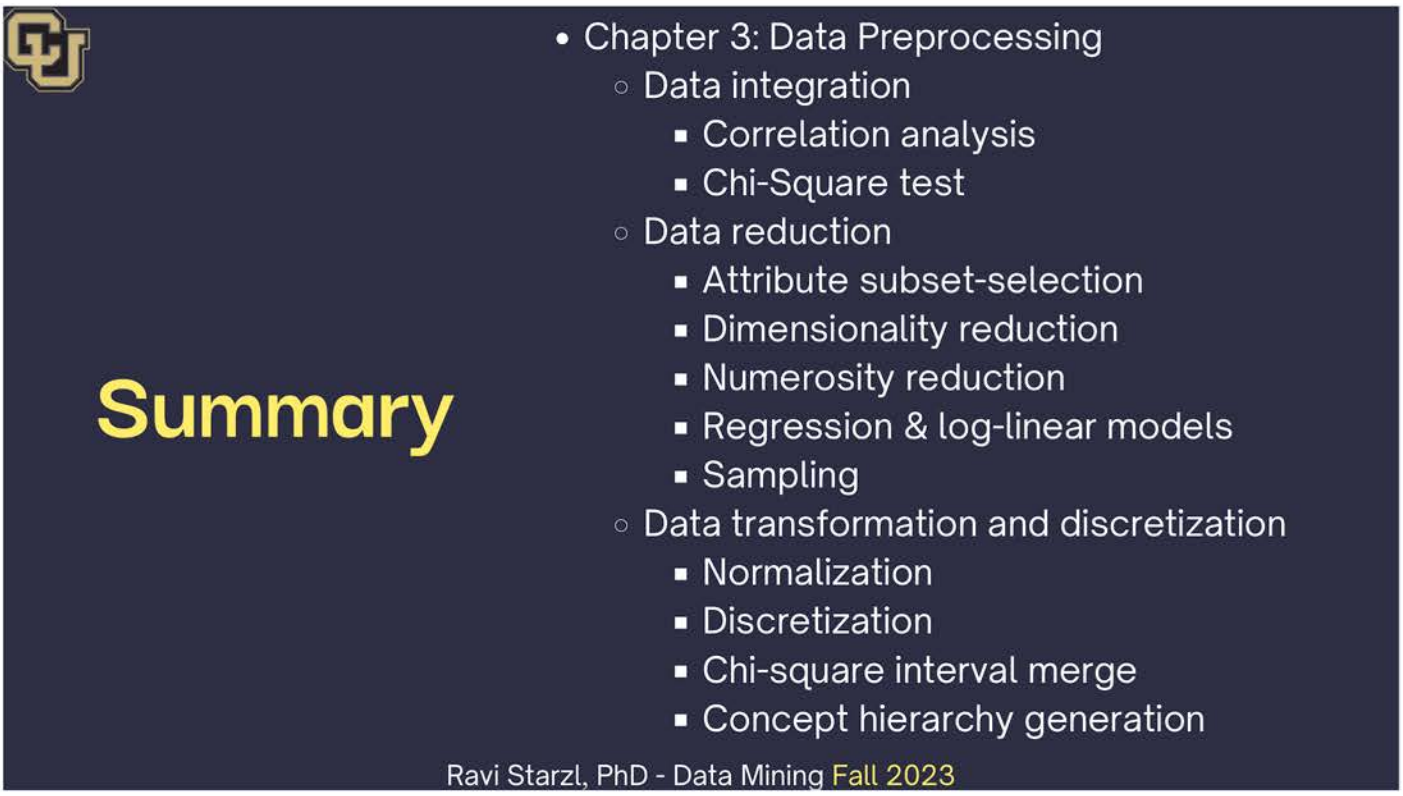
For instance, in temporal data such as "Weekday, Month, Quarter, Year", Weekday does not fit the automatic hierarchy since it has fewer distinct values than Month, but is not conceptually a broader attribute. Weekday belongs to a separate parallel hierarchy.

Benefits of hierarchical abstraction include:

- Reducing data volumes by generalization
- Enabling identification of high-level patterns
- Allowing drilling down to increasing levels of detail

Poorly conceived hierarchies can lead to incorrect conclusions and misunderstanding of patterns. Domain expertise is key to develop semantically meaningful attribute relationships.

Concept hierarchy generation is an essential technique for simplifying categorical data, revealing insights at multiple abstraction levels, and adding structural organization. Hierarchical analysis is a fundamental concept in multivariate data mining.



Data preprocessing is an essential phase that transforms raw data into a suitable format for mining. It includes critical techniques like integration, reduction, transformation, discretization, and correlation analysis.

Data Integration combines data from disparate sources like databases, texts, spreadsheets into a unified datastore. It enables a consolidated view and avoids siloed analysis. A key aspect is detecting and removing redundancies through correlation analysis between attributes from integrated sources.

Correlation Analysis quantifies the statistical relationships between variables using methods like Pearson's correlation for continuous attributes and chi-square tests for categorical attributes. Correlation helps identify redundant attributes from merged sources to avoid duplication.

Data Reduction decreases data volume through techniques like dimensionality reduction, numerosity reduction, discretization, sampling, clustering etc while retaining key information. This improves computational efficiency and model building without losing analytical value.

Data Transformation converts data into appropriate formats for applying mining algorithms, through processes like normalization, smoothing, aggregation, generalizing and constructing new attributes. This step makes the data "algorithm-ready".

Discretization transforms continuous attributes into discrete intervals or concepts to improve model interpretability and effectiveness on discrete data. Supervised and unsupervised methods

like binning, histogram analysis, entropy-based discretization are used.

Careful, thoughtful application of these techniques helps improve data quality, enhances model accuracy, reduces storage needs, and enables impactful insights through cleaning, standardizing and streamlining the data. Preprocessing is an indispensable phase and the foundation of effective data mining.

Thank you

A special thank you to Qin Lv for her slides,
on which this lecture is based

Ravi Starzl, PhD - Data Mining Fall 2023