What is Visualization?

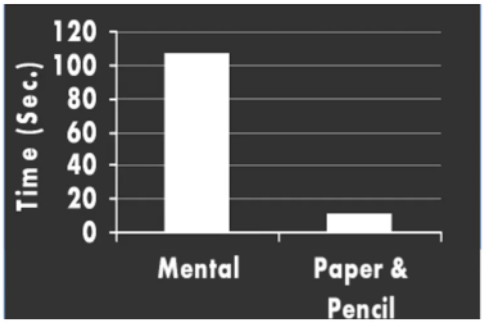
What is Visualization?

- Definition: The use of computer-supposed, interactive visual representations of data to amplify cognition
- 

What is NOT Visualization?

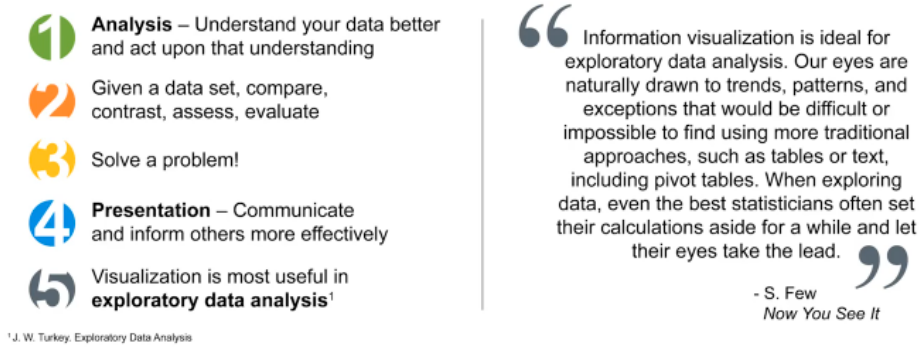
- This is not simply the process of making a graphic or an image, the goal is to create insight, not pretty pictures

Why is Visualization Helpful?

- Amplifies cognition
- Expands working memory
- Reduces search time
- Improves pattern detection and recognition
- Controls attention
- 

example 1

Purpose of Visualization:

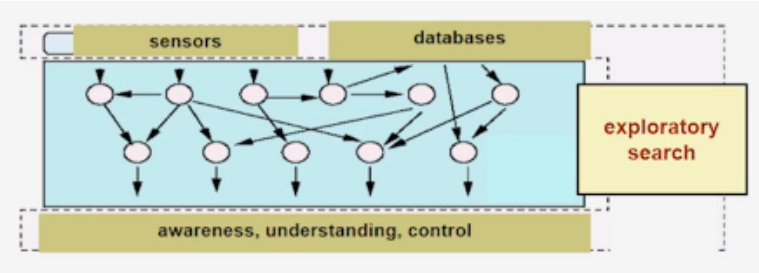
- 

Introduction to Data Exploration

“Sense” making... What Does it Mean?

- 1st sense: from latin “sentire” or “to perceive”
    - any of the faculties, as sight, hearing, smell, taste, or touch, by which humans and animals perceive stimuli originating from outside or inside the body
- 2nd sense: “to attain awareness or understanding of...”
    - “awareness” implies vigilance in observing or alertness in drawing inferences from what one experiences
    - “understanding” is the power to make experience intelligible by applying concepts and categories

Did you notice the gap?

- ... there is a gap between the first meaning (feel, measurement) and the second (awareness, understanding)
    - 

Data Processing vs. Querying vs. Exploration

- Data Processing
    - user knows what she wants
    - user has a function/procedure/workflow to compute what she wants
- Querying
    - user knows what she wants
    - user can describe what she wants
- Navigation
    - user knows what he wants
    - user does now know how to describe/locate what she wants
- Exploration
    - user does not precisely know what she wants
    - user wants to get an idea about the available data

Exploratory Search

- Acquiring new knowledge and revealing new facts
    - Analysis (identify common patterns or outliers)
    - Comparison (quantify similarity/difference)
    - Aggregation (create groups, clusters)
    - Transformation (use a more convenient representation)
    - Visualization

Data Challenges

- INS
    - Imprecision, Noise, Sparsity
- 3Vs
    - Volume, Velocity, Variety
- HMLE
    - High-Dimensional, Multi-Modal, Inter-Linked, Evolving
- Human Challenges: for many applications the final consumer is HUMAN
- Data management/mining techniques for supporting scalable, real-time, analysis and exploration
- Most data in the real world are imprecise, multi-modal, and subjective

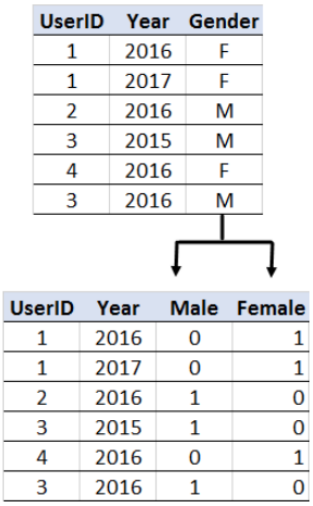
Therefore, data exploration systems need to support both...

- Effective Data Manipulation
    - Filtering: projection, selection
    - Integration: join, nearest neighbor joins
    - Set Operations: union and intersection
- Effective Data Analysis/Retrieval
    - feature extraction
    - similarity search, top-k, range, skyline, nearest-neighbor
    - clustering, partitioning
    - aggregation, summarization
    - classification, latent analysis
    - preference-driven, retrieval

Knowledge Check

- True or false? The goal of Data Visualization is mainly to create appealing images.
    - False
- True or false? Tools can improve our cognition.
    - True
        - The example (Why is Visualization Helpful?) we just discussed shows how tools can improve cognition through cognitive offloading.
- Which technique is best suited for a user who knows what he or she wants and can describe it?
    - Data Querying
        - Data querying is used when the user knows and can describe the information they seek, while data exploration is for a user who is unsure of what he or she wants and needs a general idea of the available data.
- True or false? The V's in data challenges are volume, velocity, and vulnerability.
    - False
        - The V's in data challenges are volume, velocity, and variety.

- True or false? Having less data can be a challenge.
    - True
        - That's right. With smaller amounts of data, it can be difficult to extract patterns.
- Which technique is preferred when the user has no clear idea about the data or what can be done with it?
    - Exploration
        - In the extreme situation when users do not precisely know what they want, they can use exploration to find out more about the available data and what they want to do with it. This is also known as exploratory access.
- *Figure: Gender Information*



UserID	Year	Gender
1	2016	F
1	2017	F
2	2016	M
3	2015	M
4	2016	F
3	2016	M

UserID	Year	Male	Female
1	2016	0	1
1	2017	0	1
2	2016	1	0
3	2015	1	0
4	2016	0	1
3	2016	1	0

Review Figure: Gender Information. The figure shows two different tables that represent the same data. In the first table, a single column labeled "Gender" records gender information as either "M" or "F". In the second table, however, the "Gender" column is replaced with two columns, one labeled "Male" and the other "Female", and gender information is recorded using 0's and 1's. Which data exploratory process would be used to accomplish this change?

- Transformation
    - The Gender column is transformed into Male and Female columns, represented with a numeric value of 0 or 1. Transformation is used for representing the data in a different format that is more convenient for exploration. By transforming the data we can make informed decisions and compare information in one source with another.
- Which term defines the mental action of processing information into thoughts, knowledge, and insights both efficiently and effectively?
    - Cognition
        - Cognition refers to the process of understanding the information presented effectively, which is converted to knowledge. Data Visualization is a form of external cognition that aids in understanding the meaning and patterns in data.
- What is visualization?

    - Representing data graphically to assist the processing of information into knowledge, insights, and thoughts
        - Visualization is the representation of data to amplify cognition.

Data Models and Data Organization

What is data?

- Facts and statistics collected together for reference or analysis

How is Data Organized?

- What is a database?
    - Collection of sata, organized in some fashion
- What is a data model?
    - A formalism to describe "constraints" that describe "properties" of data
        - Hierarchical
        - Relational
        - Object Oriented
        - Spatial
        - Fuzzy

What is a "Data Schema"?

- A set of constraints that
    - Describe the "properties" of data
    - Describe the structure of the data
    - Enable validation and efficient storage of the data
    - Enable querying and retrieval of data
        - Comparison
        - Indexing
        - Query optimization
        - Query processing
- "Schema" is described within the formalism corresponding to the underlying data model

Levels of Data Organization

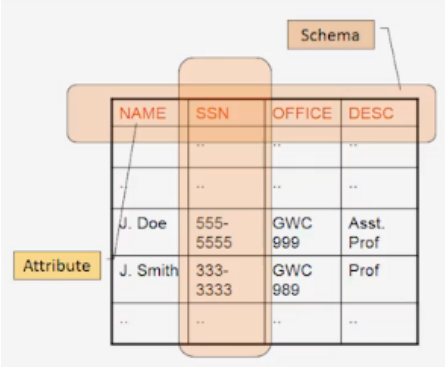
- Structured Data/Databases
- Semi-Structured Data/Databases
- Unstructured Data/Databases

Structured Data/Databases

- The data are well-structured and organized
    - A "schema" describes this structure

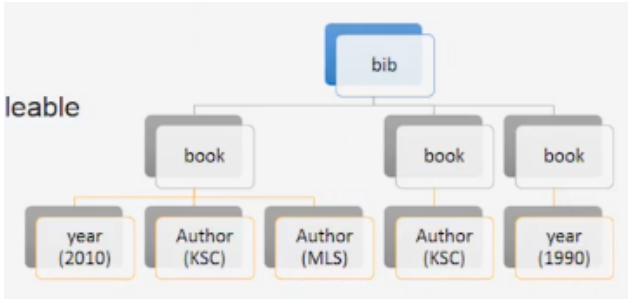
    - A Database Management System (DBMS) enforces this structure
- Advantages
    - Data organization is predictable
        - Easier to query
        - Easier to optimize
        - Easier to explore

Example: Relational Data Models

- Informally, data is organized in tabular form
    - Example: Data about an employee
- Schema for each table consists of attributes
    - Each attribute has a domain
- Functional Dependencies
    - Describes the relationships among the attributes in the schema
    - Example: A key uniquely identifies a given tuple in the table
- 

Semi-Structured Data

- The "constraints" that reflect the structure of the data are flexible
        - Ability to say "or" in the schema
        - Missing attributes (null values" or attributes which repeat itself (multi values attributes)
    - Data is self-describing: Each item in the database describes its own schema
- Advantages
    - Data organization is flexible/malleable
        - Easier to integrate
        - Easier to exchange

- 

Vector Data

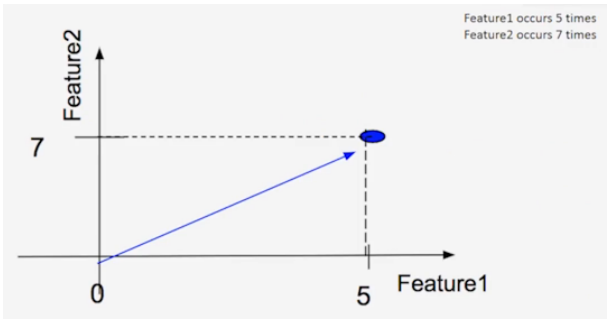
Common Data Representations

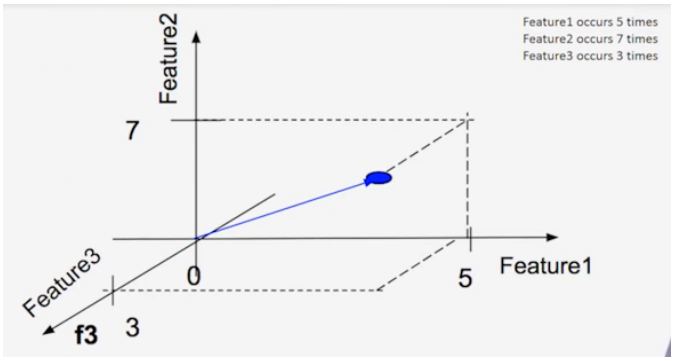
- Relational/Object Oriented Data
- Vector Space (spatial or high dimensional) Dat
- Strings, sequences, and time series data
- Trees and Graphs
- Fuzzy and Probabilistic Data

Vector Data Examples (Count!)

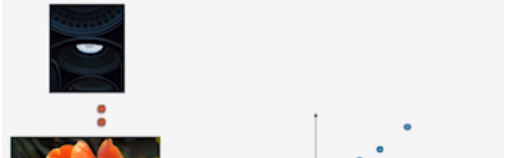
- Images
    - Colors, Textures, Shapes
- Videos
    - Acors, Ratings, Directors
- Social Networks
    - Connections, Likes, Interactions
- Books
    - Words, Authors, Publishers
- Sensor Reading
    - Sensor Values, Patterns

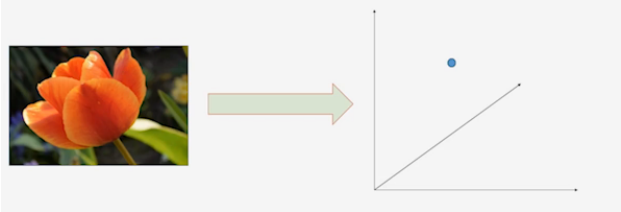
How are counts represented?



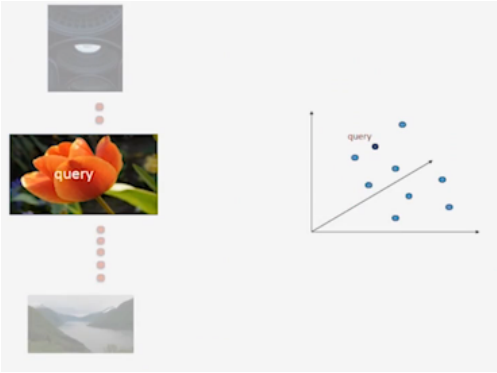


Vector Representation of a Given Object





Vector Querying

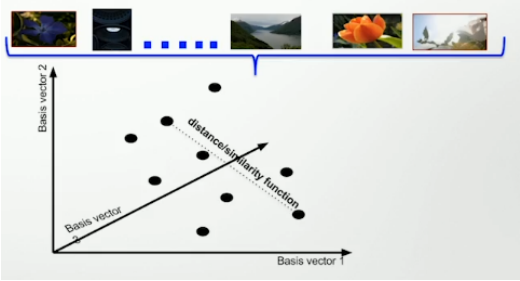


Similarity based (nearest neighbor) search

Vector Classification

Basis of a Vector Space

Vector Spaces



- What good features to use as basis vectors?
- How many features do we need as basis vectors?

- What is a good distance/similarity function?

Definition

Definition (Vector space): *The set $\mathbb{S}$ is a vector space iff for all $\vec{v}_i$, $\vec{v}_j$, $\vec{v}_k \in \mathbb{S}$ and for all $c, d \in \mathbb{R}$, the following axioms hold:*

- $\vec{v}_i + \vec{v}_j = \vec{v}_j + \vec{v}_i$
- $(\vec{v}_i + \vec{v}_j) + \vec{v}_k = \vec{v}_j + (\vec{v}_i + \vec{v}_k)$
- $\vec{v}_i + \vec{0} = \vec{v}_i$ *(for some* $\vec{0} \in \mathbb{S}$*)*
- $\vec{v}_i + (-\vec{v}_i) = \vec{0}$ *(for some* $-\vec{v}_i \in \mathbb{S}$*)*
- $(c + d)\vec{v}_i = (c\vec{v}_i) + (d\vec{v}_i)$
- $c(\vec{v}_i + \vec{v}_j) = c\vec{v}_i + c\vec{v}_j$
- $(cd)\vec{v}_i = c(d\vec{v}_i)$
- $1.\vec{v}_i = \vec{v}_i$

The elements of $\mathbb{S}$ are called vectors.

Definition (Linear independence and basis): *Let $V = \{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n\}$ be a set of vectors in a vector space $\mathbb{S}$. The vectors in V are said to be* linearly independent *if*

$$\left(\sum_{i=1}^{n} c_i\vec{v}_i = \vec{0}\right) \longleftrightarrow c_1 = c_2 = \cdots = c_n = 0.$$

non-redundant

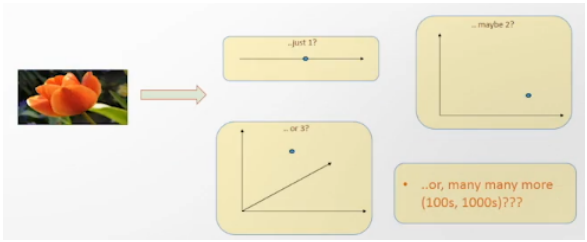
The linearly independent set V is said to be a basis for $\mathbb{S}$ if for every vector, $\vec{u} \in \mathbb{S}$, there exist constants c_1 through c_n such that

$$\vec{u} = \sum_{i=1}^{n} c_i\vec{v}_i.$$

complete

Vector Features

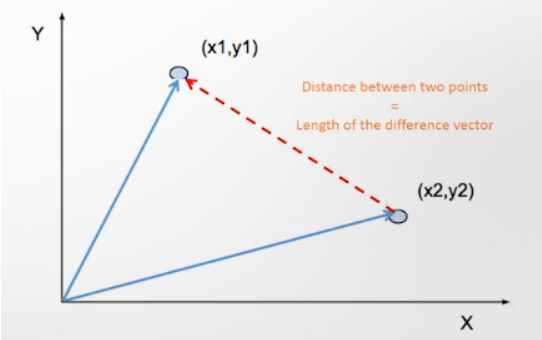
How many features do we need?



Dimensionality Curse!

- Dimensionality curse: the mor dimensions we have, the less efficient and effective search and analysis becomes
- Efficiency: search data structures are not very efficient at high dimensions
- Effectiveness: the more dimensions we have, the more data we need to discover patterns (prevent overfitting)

Distance

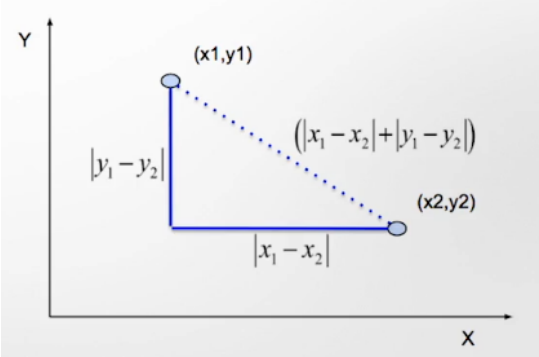


Vector Norms

Norms

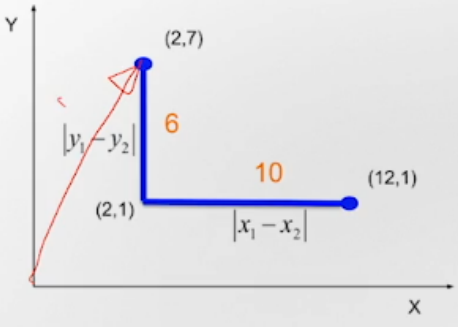
- Most commonly used family of length measurements are the p-norms
- $$\|\vec{v}\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{\frac{1}{p}}$$

1-Norm (Manhattan Distance, L1 Distance)

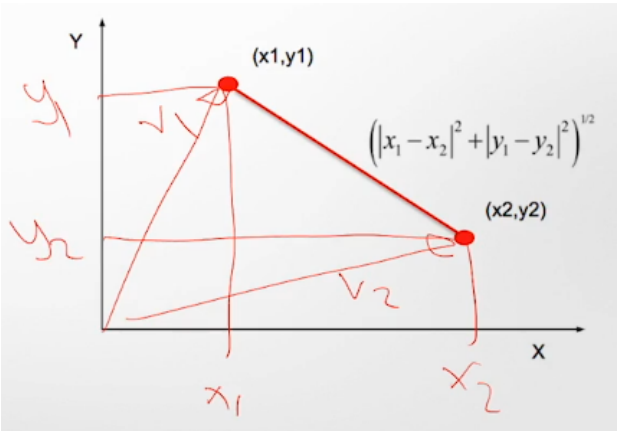


P-Norms

- 1-norm $(6+10)=16$

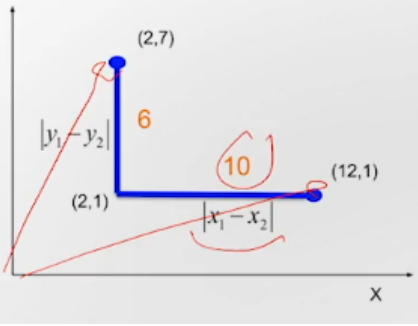


2-Norm (Euclidean Distance, L2 Distance)



P-Norms

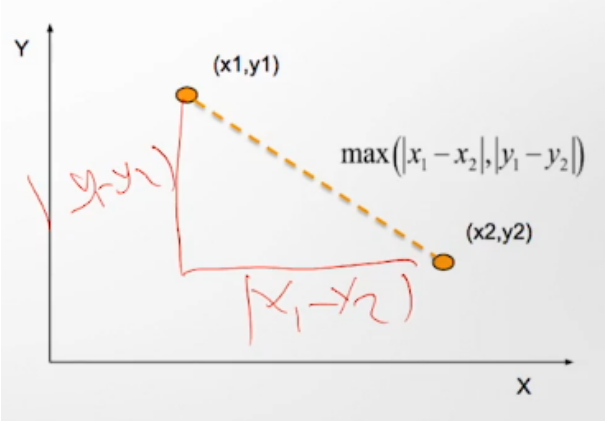
- 1-norm $(6+10)=16$
- 2-norm $(6^2+10^2)^{1/2}=136^{1/2}=11.66$



∞-Norm (L ∞ Distance)

∞-Norm

- 1-norm $(6+10)=16$



Vector Distance Measures

What is a Good Distance Measure?

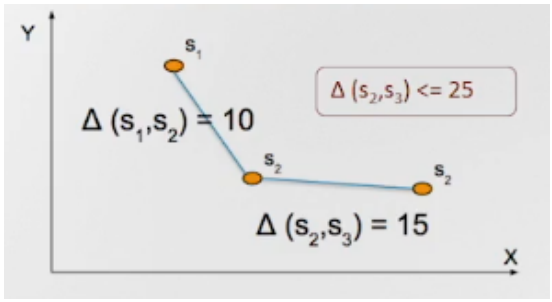
- Application dependent but, metric properties help indexing, search, and retrieval
- A metric distance, Δ, must satisfy the following conditions:
  - self-miniality: $\Delta(s,s) = 0$
  - minimality $\Delta(s_1, s_2) >= \Delta(s_1, s_2)$
  - symmetry $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
  - triangular inequality $\Delta(s_1, s_2) + \Delta(s_s, s_3) >= \Delta(s_1, s_3)$

Self-Minimality and Minimality

- Self-Minimality:
  - $\Delta(s, s) = 0$
  - Ensures that a given object matches itself perfectly
- Minimality
  - $\Delta(s_1, s_2) >= \Delta(s_1, s_1)$
  - Ensures that no other object can match the given object better than itself

Symmetry

- Symmetry:
  - $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
  - Ensures that if a given object s, is matching another object s_2, then s_2 is equally matching s_1
- Triangular Inequality
  - $\Delta(s_1\ s_2) + \Delta(s_2, s_3) >= \Delta(s_1, s_3)$
  - Enables effective printing of the search space during retrieval

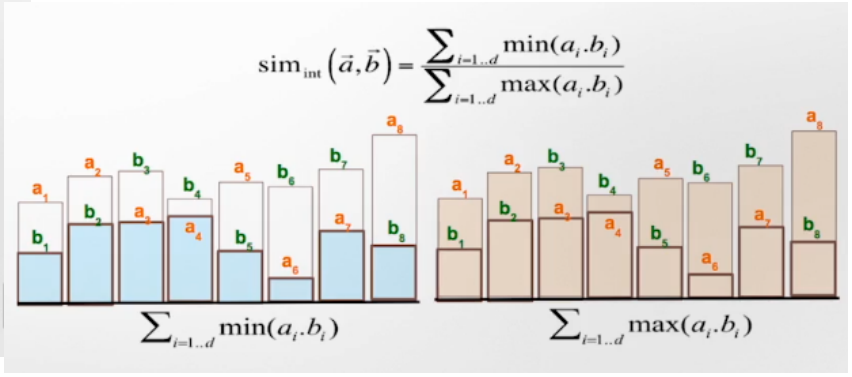


P-Norms are Metric

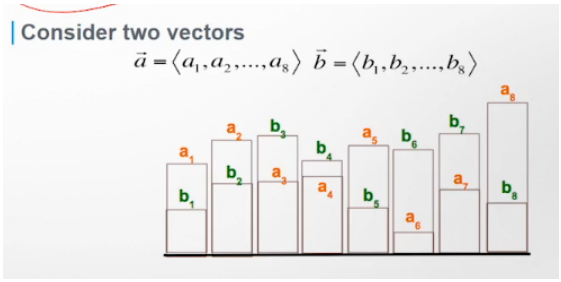
- 1-norm, L1-metric $\left(\sum_{i=1..d} |v_{1,i} - v_{2,i}|\right)$
- 2-norm, L2-metric $\left(\sum_{i=1..d} |v_{1,i} - v_{2,i}|^2\right)^{1/2}$

.....

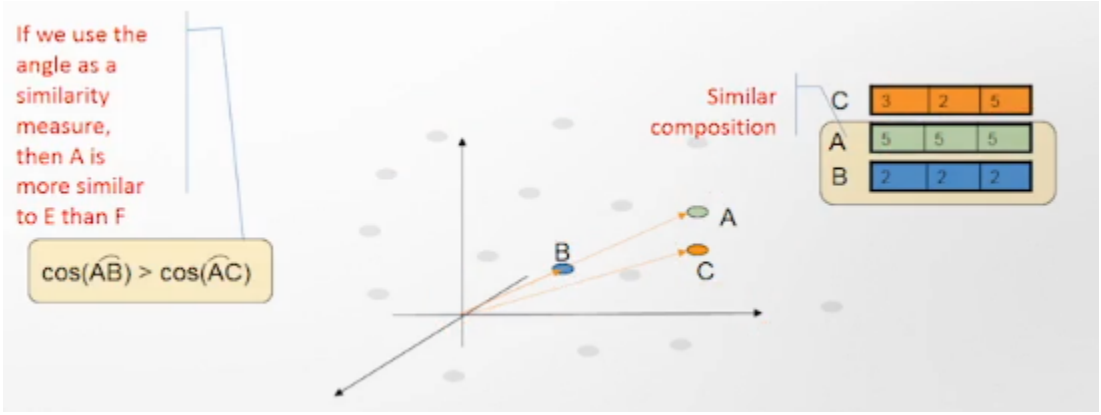
- ∞-norm, L∞-metric $\max_{i=1..d} |v_{1,i} - v_{2,i}|$



Intersection Similarity



Angle Based Similarity Measures



Angle-Based Measures

- Given $\vec{a} = \langle a_1, a_2, \ldots, a_n \rangle$ $\vec{b} = \langle b_1, b_2, \ldots, b_n \rangle$
  - Cosine similarity

$$sim_{\cos}(\vec{a}, \vec{b}) = \cos(\vec{a}, \vec{b}) = \frac{\vec{a}.\vec{b}}{|\vec{a}||\vec{b}|}$$

  - Dot product similarity

$$sim_{dot}(\vec{a}, \vec{b}) = \vec{a}.\vec{b} = \sum_{i=1}^{n} a_i b_i$$

- Cosine and dot product are the same if the vectors are unit length

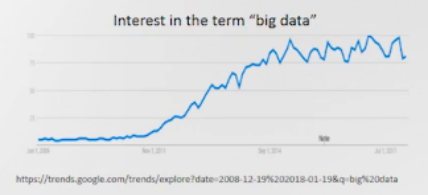
Other Commonly used Similarity/Distance Measures

- Pearson's Correlation (similarity measure)
  - linear correlation (the strength of linear association) among the corresponding components of two vectors)
- KL-Divergence (distance measure)
  - how one vector (interpreted as a probability distribution) diverges from the other
- Earth-Movers Distance (distance measure)
  - how one vector (interpreted as a probability distribution) diverges from the other

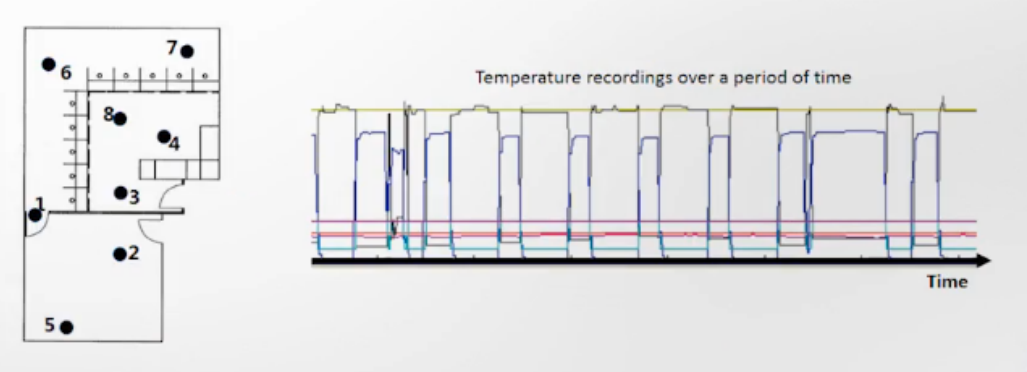
Strings and Sequences

Strings, Sequences, Time Series

- A string or sequence, $S = (c_1, c_2, \ldots c_N)$ is a finite sequence of symbols. Here, N denotes the length of the string or sequence and c_i are from an alphabet of symbols
  - Example: abcbbbaabbaabcbbbaaabbc
- A times series, $T = (d_1, d_2, \ldots, d_N)$ is a finite sequence of data values. Here, N denotes the length of the time series and $d_i \in R$
  - Example:



Multi-Variate Time Series



String/Sequence Matching and Search

- Prefix search:
    - Find all strings that starts with “tab”
        - “table”, “tabular”, “tablet”, ...
- Subsequence search:
    - Find all strings that contain the subsequence “ark”
        - “marketing”, “spark”, “quark”, ...
- Subsequence match:
    - Find the longest matching subsequence between “plasticity” and “scholastic”
    - Find the most frequently repeating 3 character subsequence
        - “abcbbbaabbaabcbbbaaabbc”
- How similar are two strings?
    - “table” vs. “cable”?
    - “table” vs. “tale”?
    - “table” vs. “tackle”?

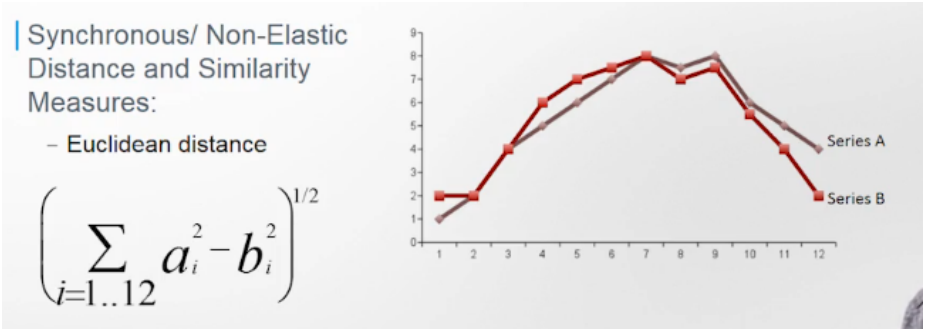
Edit Distance

- Edit distance between two sequences is the minimum number of edit operations needed to convert one sequence to the other
    - “table” vs. “cable”
        - 1 replacement (“t” with “c”)
    - “table” vs. “tale”
        - 1 deletion (“b”)
    - “table” vs. “tackle”:
        - 1 deletion (“b”) and 2 insertions (“c” and “k”)
        - 1 replacement (“b” with “c”) and 1 insertion (“k”)

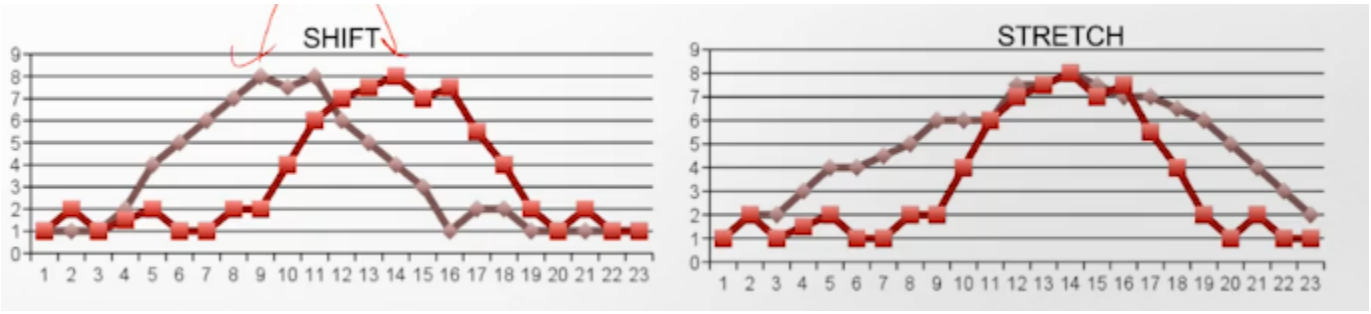
Time Series Matching

- Synchronous/Non-Elastic Distance and Similarity Measures:

- Euclidean distance

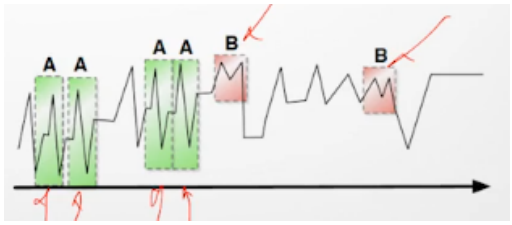


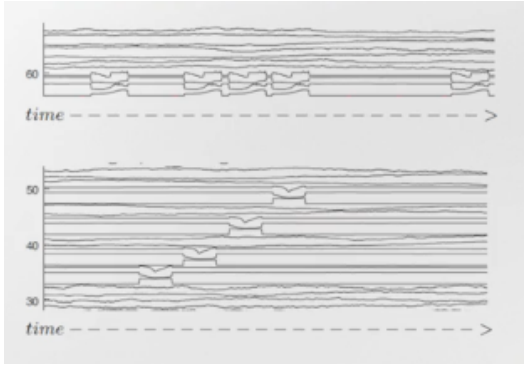
Asynchrony in Time Series



- Asynchronous/Elastic Distance and Similarity Measures
  - Edit Distance, ED
  - Dynamic Time Warping, DTW
  - Feature-based Alignment, RMT

Motifs

- Frequently repeating patterns in time series
  - 

- Motifs can also occur in multi-variate time series
  - 

Knowledge Check

- True or false? High dimensionality makes it easier to analyze data.
    - False
        - More dimensions add calls for overfitting, which makes it difficult to find the separation lines. More examples are preferable to more dimensions.
- True or false? The difference along the X-axis is given by X1 - X2.
    - True
        - The difference along the X-axis is given by X1 - X2, because that is the difference travelled by the vector from (x1, y1) to (x2, y2).
- True or false? The two main components in a sequence are a starting point and an ending point.
    - False
        - Sequences have a starting point, and ending point, and a length.
- True or false? The distance between "table" and "tale" is 1 unit.
    - True
        - The operation performed to achieve "tale" from "table" is 1 deletion.
- What are the challenges faced in data collection methods?
    - Noise
        - Separation of noisy data from the actual data is difficult and noisy data can skew the output results.
    - Imprecision
        - Imprecise records or values need to be removed while exploring the real world data for accurate analysis.
    - Sparsity
        - Effective data exploration is hard to achieve with very few data elements.
- Suppose that a sensor device attached to a vehicle is recording the speed of that vehicle at every second. The data are saved in a database table using one column per speed value. If the vehicle is traveling for an average of 8 hours per day, there will be approximately 10,512,000 columns in the table. What data challenge makes the exploration of this data difficult?
    - High Dimension
        - High dimensional data is the data that contains many different aspects. While there is only one observation, speed, the amount of columns added to the table would make it difficult to explore and visualize.
- Which aspects of data should be handled by a scalable data exploratory system? Select all that apply.
    - The diversity of the data types
        - This is an aspect that should be handled by a scalable data exploratory system. The incoming data can be heterogeneous and it needs to be filtered or integrated to remove the underlying complexity before exploration and visualization.
    - The speed of new data generated

        - This is an aspect that should be handled by a scalable data exploratory system. As the incoming velocity of the data increases we need effective and efficient algorithms to make sense out of the data to support data exploration.
    - The amount of the data
        - This is an aspect that should be handled by a scalable data exploratory system. As the volume of the data increases the exploration becomes harder.

Exploratory Querying

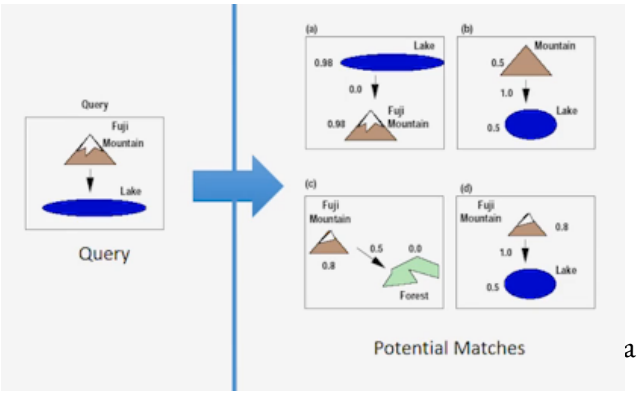
Exploratory Search

- Acquiring new knowledge and revealing new facts
    - Analysis (identify common patterns or outliers)
    - Comparison (quantify similarity/difference)
    - Aggregation (create groups, clusters)
    - Transformation (use a more convenient representation)
    - Visualization

Exploratory Querying

- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Query by Example / Similarity Search

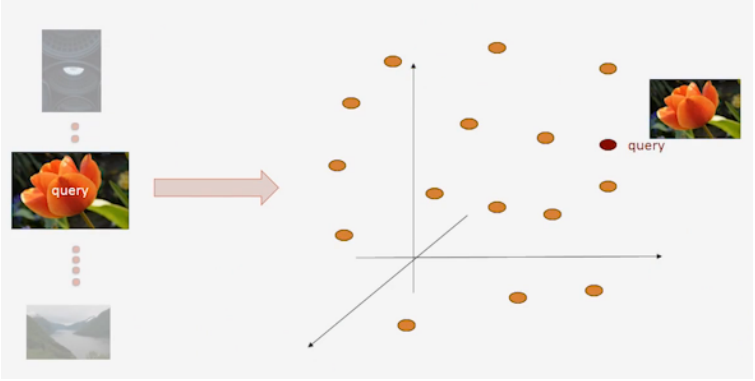


Ranked Retrieval

- When not all sub-goals need to be satisfied a element in the database is a potential match
- Hence the query results need to be ranked according to some objective subjective criteria

Top-K Search

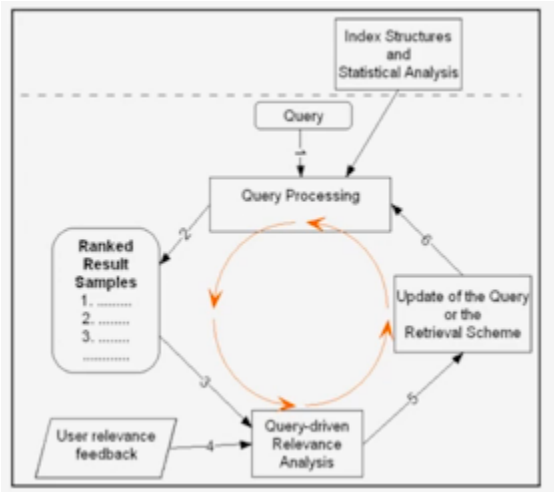
“Find K=2 most similar images”



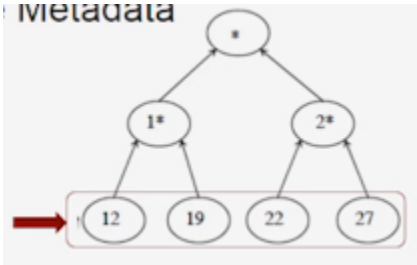
Nearest-Neighbor Search

Semantic Gap/Subjectivity

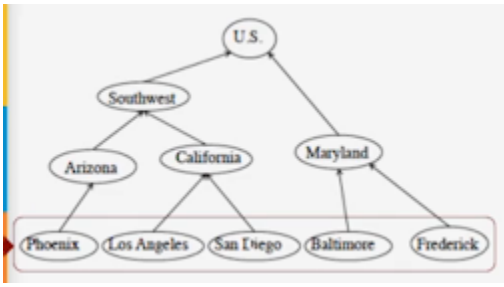
- Relevance Feedback
    - in q request to identify images “similar” to an example, the visual features of the images that are relevant for the user’s query must be inferred from feedback to identify the most relevant images



- Age Metadata



Location Metadata

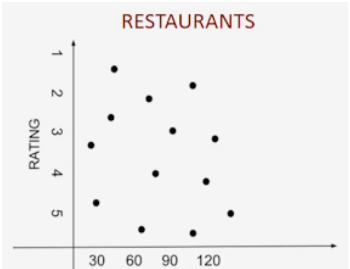


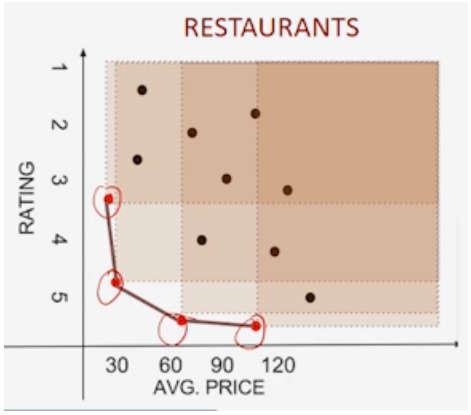
Data Table

NAME	AGE	LOCATION
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Skylines

- Question:
    - The higher the rating the better
    - The cheaper the price the better
    - Which restaurants would you consider?



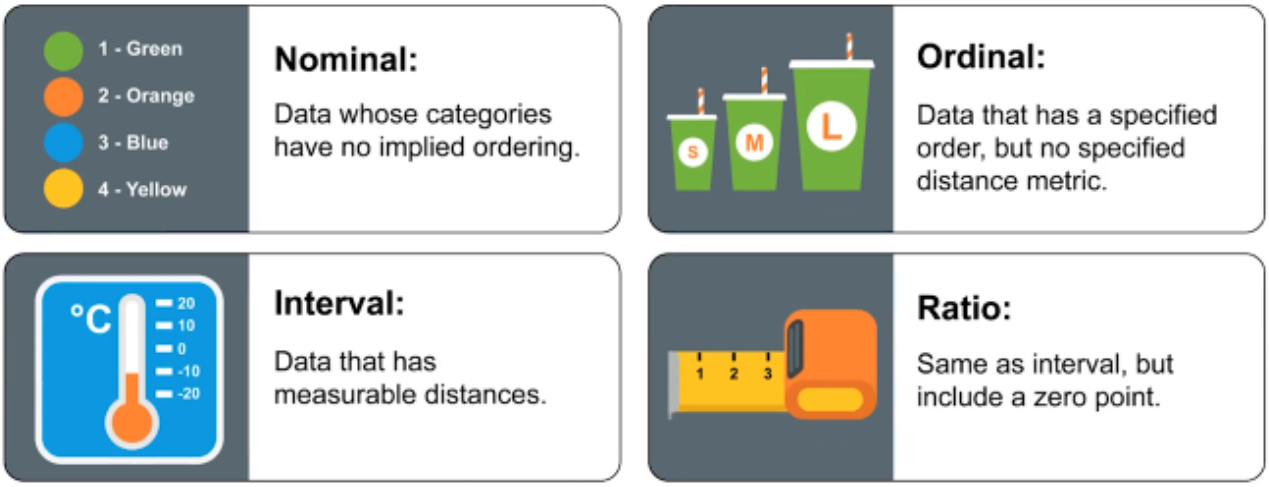
- Objects in the “skyline” are not dominated by any other objects in the database
    - Also known as the Maximum Vector Problem [Kung75]
    - Coined as “Skylines” in [Borzsonyi01]
- 

Data Sketches – Example: Tag Clouds

Visual Variables

Data Types

- Nominal: Data whose categories have no implied ordering
- Ordinal: Data that has a specified order, but no specified distance metric
- Interval: data that has measurable distances
- Ratio: same as interval, but include a zero point

- 

Visualization Pipeline

- Data Analysis: data are prepared for visualization (smooth, interpolate, transform)
- Filtering: a subset of the data is selected for visualization
- Mapping: data are mapped to geometric primitives and their attributes
- Rendering: geometric data are transformed to image data

- We want to take these different data types and map them to an appropriate visual representation

Mapping Data: Aesthetic Attributes

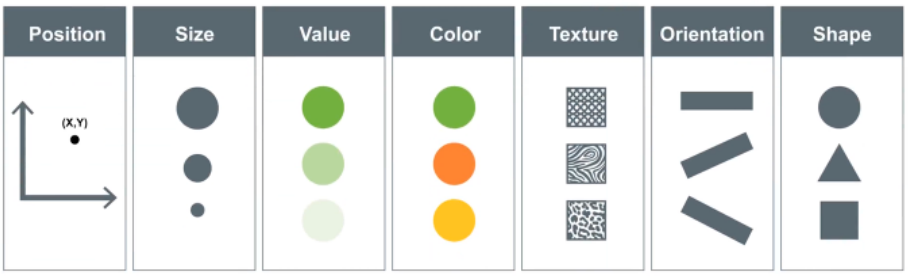
- Form
- Surface
- Motion
- Sound
- Text

Aesthetic Attributes

- Must be capable of representing both continuous and categorical variables
    - Continuous variable: an attribute must vary primarily on **one** psychophysical dimension
    - Multidimensional attributes: must scale them on a single dimension
- Does not imply a linear perceptual scale

Bertin's Visual Variables

- Visualization is concerned primarily with a mapping to visual form

- 

Position

- A location in a multi-dimensional space
- Continuous variables map to densely distributed locations
- Categorical variables map to a lattice
- Ordering may or may not have meaning in terms of what is being measured
- Best way to represent a quantitative dimension visually
- Points or line lengths places adjacent to a common axis enable judgements with the least bias or error

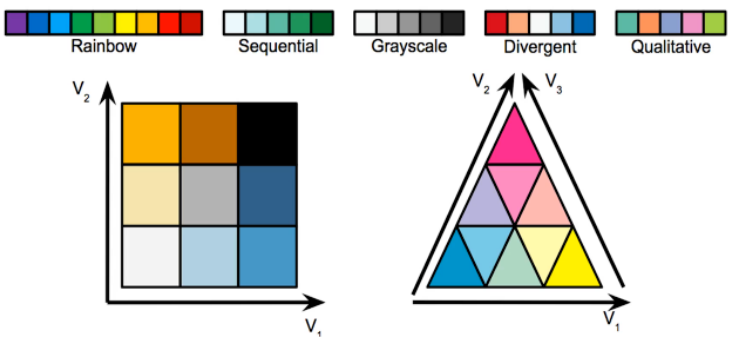
Size

- The variation in terms of length or area
- In three dimensions, includes volume
- Area and volume representations among the worst attributes to use for graphing data
- Size for lines is usually equivalent to thickness
    - Less likely to induce perceptual distortion
- Size can be used to great effect with path
- For objects with rotational symmetry, map size to the diameter rather than area
- Representing data through area or volume should probably be confined to positively skewed data that can benefit from the perceptual equivalent of the square root transformation

Rotation

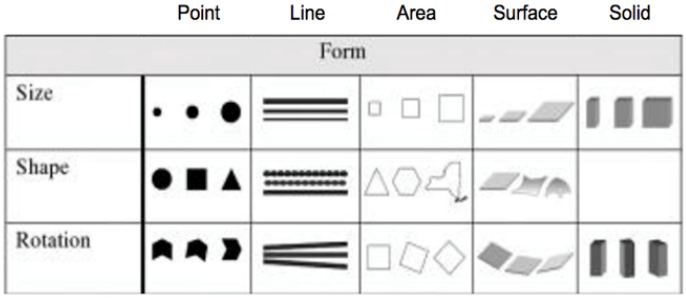
- Rotational angle of the graphic primitive
- Lines, areas and surfaces can only rotate if they are positionally unconstrained

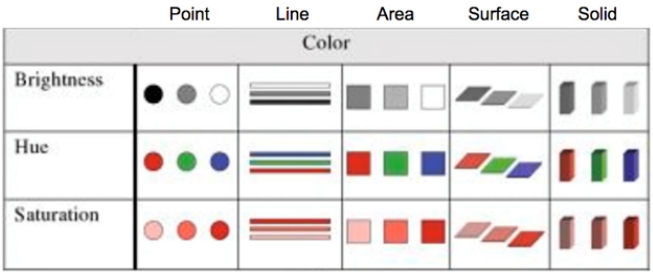
Color

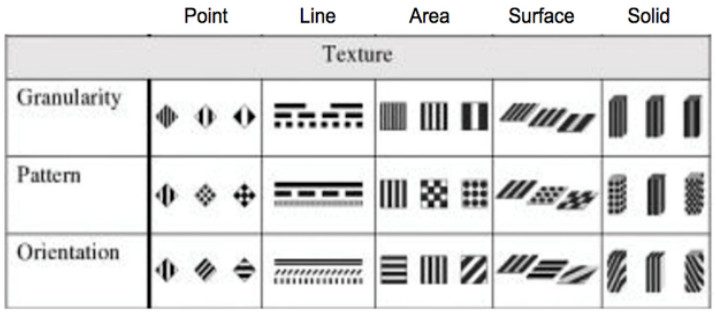


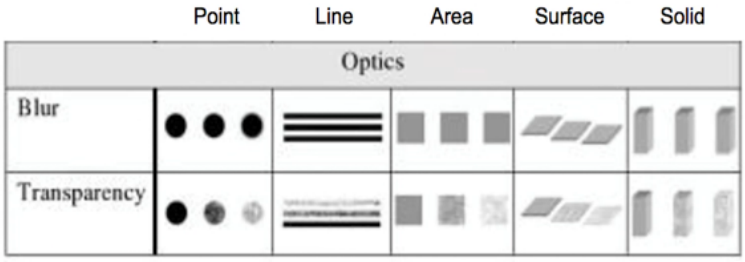
Texture

- Includes pattern, granularity and orientation
  - Granularity
    - Repetition of a pattern per unit of area
  - Orientation
    - Angle of the pattern
- Texture alone can be a basis for perception
- Textures can be described in a variety of ways
  - Fourier Transform: decomposes a grid brightness values into sums of trigonometric components
  - Auto-Correlogram: characterize the spatial moments of a texture









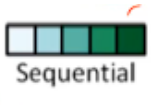
should be intuitive

Univariate Color Schemes

- Rainbow COlor Scheme
  - Rainbow color scale is one of the most commonly used
  - It is a poor color map in a large variety of domain problems
  - Ordering of the hues is unintuitive
  - Nominal data types can use this scale as no ordering is implied

- Qualitative Color Scheme

- Sequential Color Scheme
    - Sequential maps represent ordered data
    - Dark colors typically represent high ranges, bright, low
    - Benefits are that the scale is intuitive
    - Weakness is that limited number of distinguishable colors can be represented

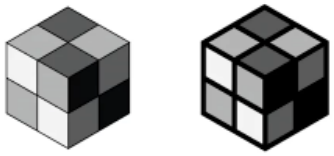


- Grayscale Color Scheme
    - Simplest is the gray scale map where variable is mapped to brightness

Illusions in Grayscale

- The eye sees size different shades of gray, but actually there are only four

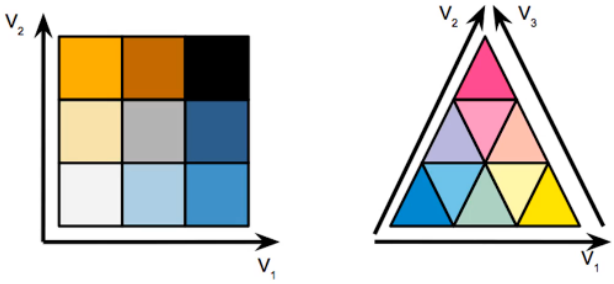
- Adding a thin border has minimal effect on the illusion, but having a thick border is able to neutralize the effect



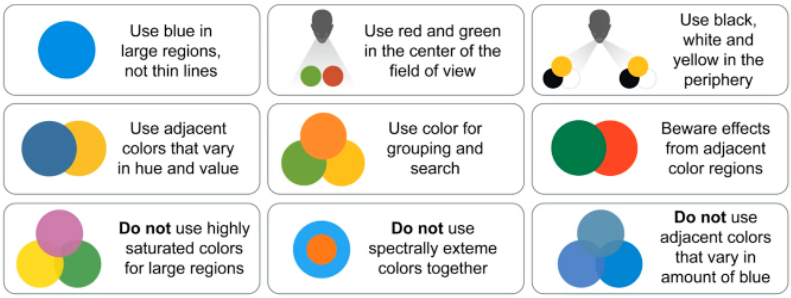
Divergent Color Scheme

- Provides means for variable comparisons
- Best suited for ratio data where there is some meaningful zero point
- Scale lacks a natural ordering of colors
- Careful choices must be made in choosing high and low ends
- Can use concept of cool (clues) and warm (reds and yellow) colors)

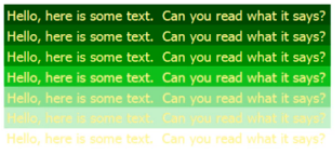
Multivariate Color Schemes

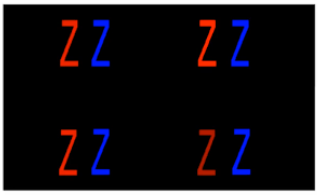


Mapping Color



- Use blue in large regions, not thin lines
- Use red and green in the center of field of view
- Use black, white, and yellow in the periphery
- Use adjacent colors that vary in hue and value
- Use color for grouping and search
- Beware effects from adjacent color regions
- **Do not** use highly saturated colors for large regions
- **Do not** use spectrally extreme colors together
- **Do not** use adjacent colors that vary in amount of blue

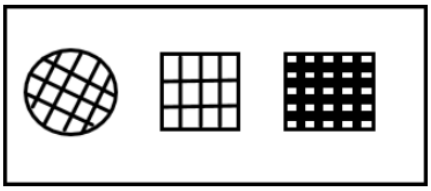




Knowledge Check

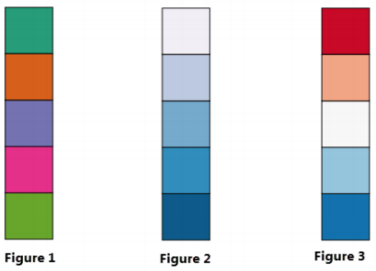
- True or false? Query results are ranked only by objective criteria.
  - False
    - Query results need to be ranked by some objective criteria and some subjective criteria.
- True or false? Roll-up data has less detail.
  - True
    - As it moves up the hierarchy, roll-up data contains fewer details.
- True or false? We do not reduce data in Skyline queries.
  - False
    - As with summarizing, we also reduce data in skyline queries.
- True or false? Intervals are data that have measurable distances.
  - True
- True or false? Lines, areas, and surfaces may only rotate if they are constrained.
  - False
    - Things like lines, areas, and surfaces can only rotate if they are positionally unconstrained.
- True or false? In nominal data types, there is an order.
  - False
    - In nominal data types, there is no implied ordering.
- True or false? Multivariate color schemes are harder to understand.

    - True
        - Multivariate color schemes can increase the cognitive load.
- How many visual variables did Bertin identify?
    - 7
        - Bertin identified seven (7) visual variables. These include position, size, value, color, texture, orientation, and shape.
- *Figure Visual Variable Example*



Review Figure: Visual Variable Example. Which visual variable, as defined by Bertin, is best demonstrated in this figure?

    - Texture
        - When considering only the texture, the three objects in this figure can be distinguished from one another.
- *Image: Color Scheme Example*

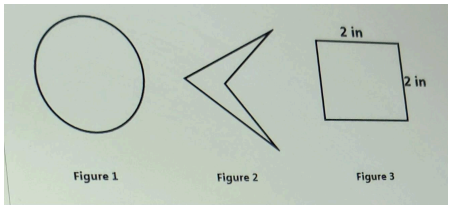


Review Image: Color Scheme Example. This image depicts three figures, each with a different color scheme. Which figure represents the sequential color scheme?

    - Figure 2
        - Figure 2 uses different hues of a single color. Sequential color schemes can be used to represent ordered data. Typically, darker colors represent high ranges while lighter (or brighter) colors represent low ranges.

Module 1 Quiz Questions

1. According to Sephan Few, what is the purpose of visualization?
    a. Visualization is used for data exploration and draws patterns from data.
2. Which step in the visualization pipeline involves preparing the data (interpolation, smoothening, etc.)?
    a. Data Analysis
3. Which statement describes the best practices for effective use of color?
    a. Use adjacent colors that vary in hue and value

4. Suppose that an online shopping website wants to visualize the behavior of its clients while they navigate through the site pages. In which step of the visualization pipeline would customers between the ages of 20 and 50 be selected?
    a. Filtering
5. 

Which shape should be used when illustrating Bertin's orientation variable?
    a. Figure 2
6. Which term defines the user of computer-supported, interactive, visual representation stations of data to amplify cognition?
    a. Visualization
7. Which visual variables is *least* suitable for representing nominal data?
    a. Size
8. Suppose you are creating a visualization using this data:
   1cm, 100cm, 1km, 1mm
   What type of data is this?
    a. Ratio
9. What is the condition required to rotate shapes such as lines, areas, and surfaces?
    a. The shape must be positionally unconstrained

What is Exploratory Data Analysis?

Exploratory Data Analysis

- Approach to analyzing data sets to summarize main characteristics +
- Elements Include:
    - Data Visualization
    - Residual Analysis
    - Data transformation/re-expression
    - Resistance Procedures
- = Exploratory Data Analysis

Data Visualization

- Data visualization facilitates advanced data analysis
- Checks distributional and other assumptions
- Observes time-based processing
- Spots outliers
- Examines relationships

- Discriminates clusters
- Compares mean differences

Data Distributions

- The type of data distribution affects
  - How it should be analyzed
  - How it should be visualized
- Key step is pre-conditioning data

The Normal Distribution

- Normal (Gaussian) Distribution
  - Popular
  - Fully characterizes with two parameters
  - Probability is determined knowing distance from mean
  - Many measures and tests are designed for this
- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Mean and Standard Deviation

- For sample population $X = \{x_1, ..., x_n\}$ the mean is defined as:

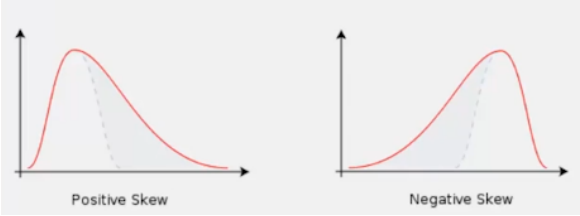
  $\mu = \frac{1}{N} \sum_{i=0}^{N} x_i$

- The standard deviation is defined as:

  $\sqrt{\frac{1}{N} \sum_{i=0}^{N} (x_i - \mu)^2}$

Skewness

- Measure of the asymmetry of the probability distribution



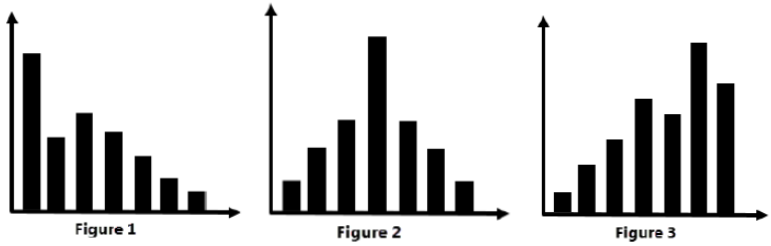
Skewed Data

- For a sample of N values, the sample skewness is:

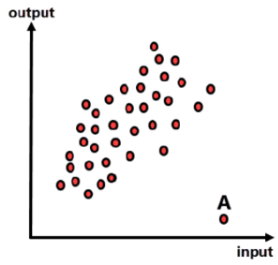
$$\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right)^{3/2}}$$

Knowledge Check

- True or false? Skewness is the measure of the symmetry of the probability distribution.
    - False
- Review Figure 1, Figure 2, and Figure 3. Each figure depicts a data distribution. Which figures depict data with a skewed distribution? Select all that apply.



    - Figures 1 and 3
- Review Figure: Output Data. The figure depicts a graph that plots the outputs for certain inputs as discrete points. What does the point labeled "A" represent?



    - An outlier
        - The point labeled with A is distant from most of the other observations.
- In a positively skewed distribution, which measure of central tendency will have the largest value?
    - Mean
        - In positively skewed data, the mean has the highest value. In positive- or right-skewed data, most of the values are clustered around the left tail of the distribution. Thus, the mean value will be greater than the values for the median and the mode.

Introduction to Pie Charts

When to use a pie chart

- Categorical data
    - Each slice can represent a different category
- How many categories do you have?
    - Good rule of thumb is ~7 categories maximum

Pie Charts

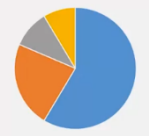
- Use to encode values:
    - Angle
    - Area
    - Arc Length

Interpreting Pie Charts

- Pie charts lead to an overestimation of small values and underestimation of large ones
- Perceptual research shows that error is high when estimating values
- Pie Charts should be presented with values as pie slice labels

Visual Variables

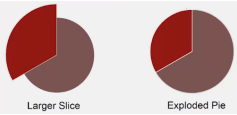
- Angle is not a key visual clue
- Arc Length and Area are important



- Doughnut charts are just as effective

Interpreting Pie Chart Variants

- Larger Slice
- Exploded Pie



Knowledge Check

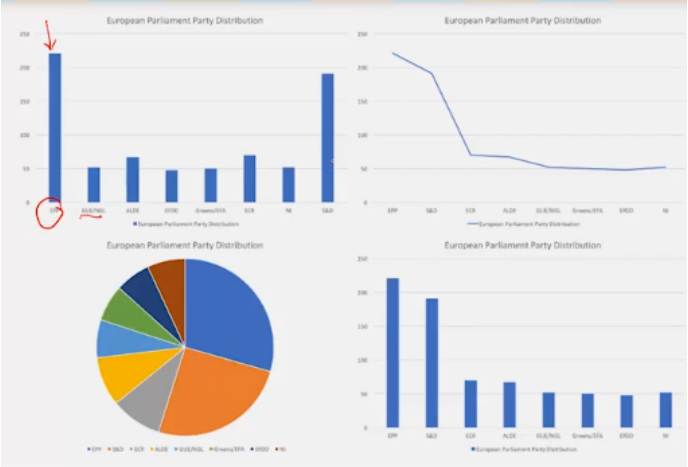
- True or false? Pie charts with a maximum of 7 to 9 categories are best for human perception.
    - True
        - Too many categories make the visualization cluttered and sometimes incomprehensible.

Bar and Line Charts

Which Type of Graph Should I Use?

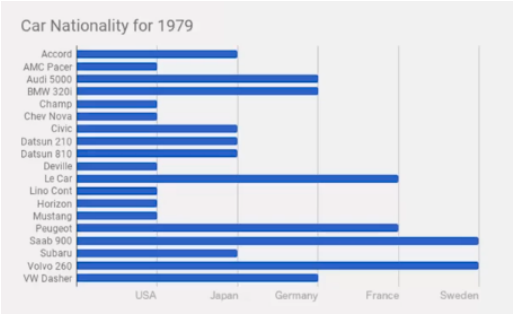
- Pie Charts
    - Comparing parts of a whole
- Bar Charts
    - Comparing between groups over time
- Line Chart
    - Changes over time

Non-Time Series Data



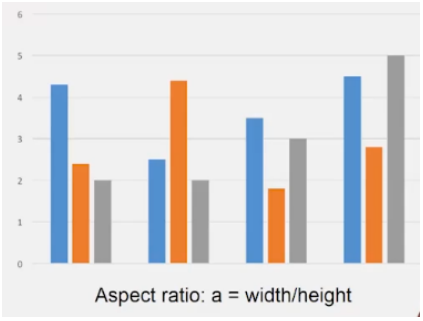
Time Series Data

When to NOt Use Bar Charts



Graph Aspect Ratios

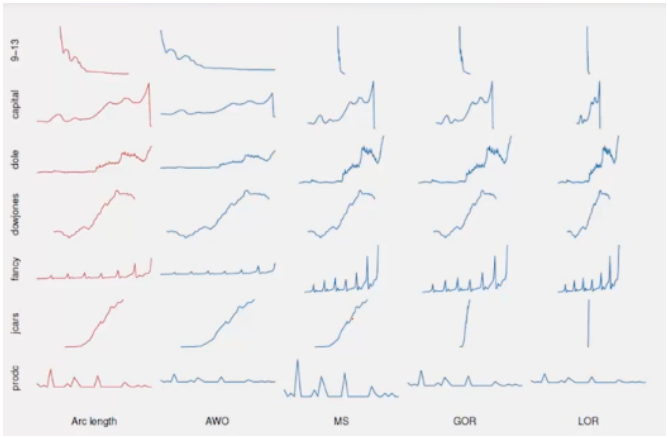
- Perception of trends and patterns is heavily influenced by the aspect ratio
- Aspect ratios affects:
    - Densities
    - Relative distances orientations



Arc Length-Based Aspect Ratio

$$max_a \frac{\Sigma_i \Sigma_j |sin\theta_{ij}| l_i(a) l_j(a)}{\Sigma_i \Sigma_j l_i(a) l_j(a)}$$

Arc Length-Based Aspect Ratio Selection

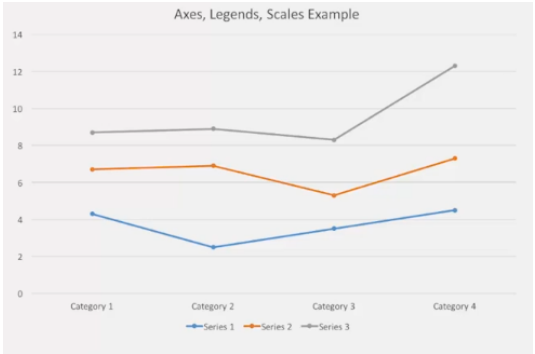


Knowledge Check

- True or false? Bar charts are best for showing trends over a period of time.
    - False
        - With gaps between the bars in a bar chart, it is difficult to comprehend a trend. Bar charts are not ideal for showing trends over a period of time.

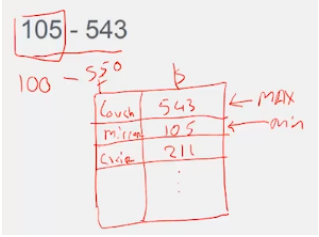
Design Considerations for Non-Data Components of Graphs

Axes, Legends and Scales

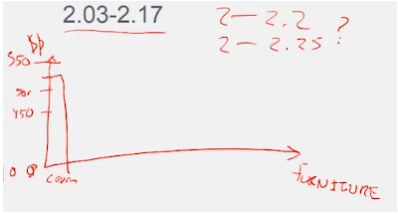


Heckbert's Labeling Algorithm

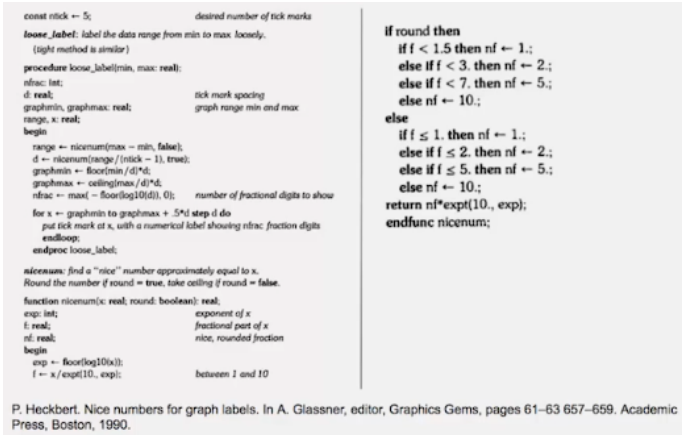
- Data Range
  - 105-543



- Data Range
  - 2.03-2.17



Nice Numbers



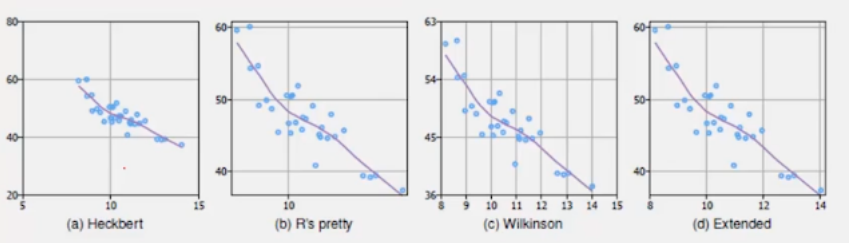
```
const ntick ← 5;                        desired number of tick marks

loose_label: label the data range from min to max loosely.
  (tight method is similar)
procedure loose_label(min, max: real);
nfrac: int;
d: real;                                tick mark spacing
graphmin, graphmax: real;               graph range min and max
range, x: real;
begin
  range ← nicenum(max − min, false);
  d ← nicenum(range/(ntick − 1), true);
  graphmin ← floor(min/d)*d;
  graphmax ← ceiling(max/d)*d;
  nfrac ← max( − floor(log10(d)), 0);   number of fractional digits to show
  for x ← graphmin to graphmax + .5*d step d do
    put tick mark at x, with a numerical label showing nfrac fraction digits
  endloop;
endproc loose_label;

nicenum: find a "nice" number approximately equal to x.
Round the number if round = true, take ceiling if round = false.
function nicenum(x: real; round: boolean): real;
exp: int;                               exponent of x
f: real;                                fractional part of x
nf: real;                               nice, rounded fraction
begin
  exp ← floor(log10(x));
  f ← x/expt(10., exp);                 between 1 and 10
if round then
  if f < 1.5 then nf ← 1.;
  else if f < 3. then nf ← 2.;
  else if f < 7. then nf ← 5.;
  else nf ← 10.;
else
  if f ≤ 1. then nf ← 1.;
  else if f ≤ 2. then nf ← 2.;
  else if f ≤ 5. then nf ← 5.;
  else nf ← 10.;
return nf*expt(10., exp);
endfunc nicenum;
```

P. Heckbert. Nice numbers for graph labels. In A. Glassner, editor, Graphics Gems, pages 61–63 657–659. Academic Press, Boston, 1990.

Heckbert's Labeling Algorithm

- Problem
  - For small numbers, the range of labels can be much larger than the data range
- Solution
  - Drop labels which overlap or fall outside the data range
  - This leads to unevenly spaced labels or axes with only one label

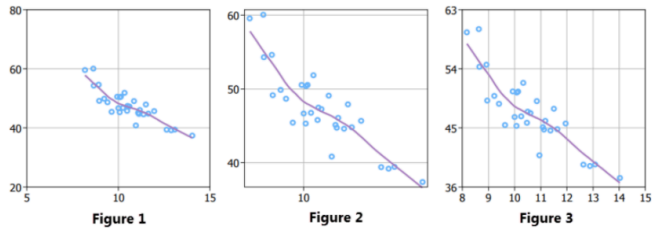
Extension of Wilkinson's Algorithm



- Coverage = $1 - \frac{1}{2}\frac{(d_{max}-l_{max})^2+(d_{min}-l_{min})^2}{[.1((d_{max}-d_{min})]^2}$
- Legibility = $\frac{format + font_{size} + orientation + overlap}{4}$

Knowledge Check

- *Image: Charts*



Review Image: Charts. The image shows three charts. Which chart best illustrates the problem with the Heckbert's Labeling Algorithm?

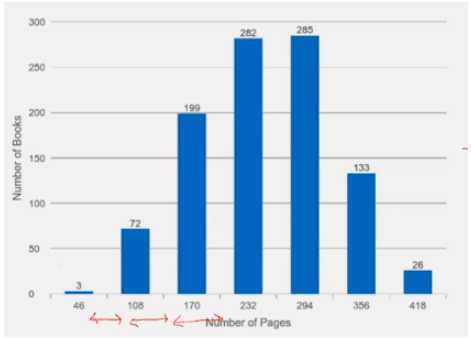
  - Figure 1
    - Heckbert's algorithm uses optimization to represent graph labels. In Figure 1, as the numbers have very small difference, the range of labels is a lot larger than the data range.
- Which description provides an accurate summary of the extension of Wilkinson's algorithm?
  - Utilize graph components to optimize tick placement
    - Wilkinson's algorithm finds the optimal number of tick marks and labels for a given graph. It will make sure that the label range does not exceed the data range and reduces the white spaces.
- *Figure: Orange Tree Growth Chart*

  - Review Figure: Orange Tree Growth Chart. This line chart displays the growth of five different orange trees. Which component of the graph is missing?
    - Legend
      - This figure should have a legend to show which tree each line corresponds to.
- Which quantities are used in the optimization of the Heckbert's nice numbers algorithm? Select all that apply.
  - Max and min of graph range

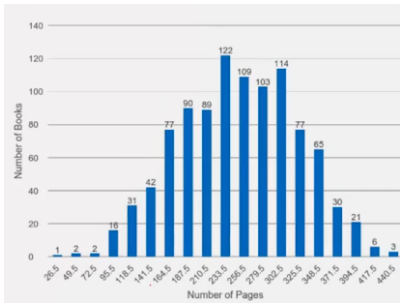
        - This is one of the quantities used to optimize Heckbert's nice number's algorithm. Optimization algorithm uses Min and Max to label the data range. Please review "Design Considerations for Non-Data Components of Graphs" and answer this question again.
    - Tick marks spacing
        - This is one of the quantities used to optimize Heckbert's nice number's algorithm. Tick marks spacing is used by optimization algorithm for evenly spacing the number of tick marks. Please review "Design Considerations for Non-Data Components of Graphs" and answer this question again.
    - Number of tick marks
        - This is one of the quantities used to optimize Heckbert's nice number's algorithm. The optimization algorithm uses a number of tick marks to organize them exactly for a given graph. Please review "Design Considerations for Non-Data Components of Graphs" and answer this question again.

Creating Histograms

Histogram Binning

- Number of bins (k) can be user-specified or chosen from a suggested bin width (h) such that:
    - $k = \lceil \frac{max\ x - min\ x}{h} \rceil$
- Common choices for k include the square-root choice where:
    - $k = \sqrt{N}$
- Sturge's Formula:
    - $k = \lceil logN + 1 \rceil$
    - 

- Scott's Choice:
    - $h = \frac{3.5\sigma}{V^{1/3}}$
- Freedman-Diaconis Rule:
    - $h = 2IQR(x)N^{-\frac{1}{3}}$

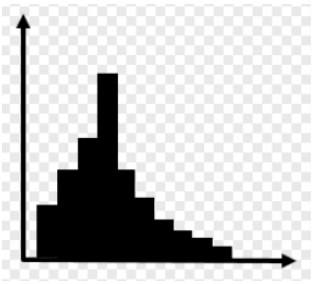
- 

Histogram Example

- Plot a histogram of 1000 block lengths
- Use all four common choices for k or h
- All $x - axis$ labels indicate the center of the histogram bin

Knowledge Check

- Histograms are specific type of which type of graph
  - Bar chart
- *Figure: Histogram*



Review Figure: Histogram. Which kind of distribution do the data in this histogram have?

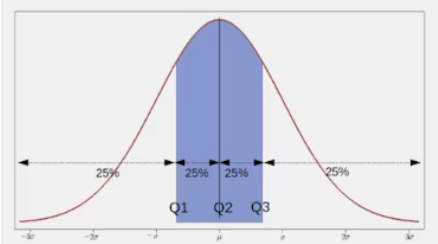
- Right skewed
  - This distribution has a tail that goes to the right, so it is right skewed (also known as a positive skew).

- Suppose that you have a dataset of 10 points. Using Sturge's formula, how many bins should there be for this dataset? (Note: Use log base 2.)
  - 5
    - Sturge's formula is $K = 1 + log_2 N$, where K is number of class intervals (bins), N is the number of observations in the set, and log_2 is the logarithm of the number. Once we calculate the answer, we apply the ceiling function and round up. Sturge's formula is useful in situations where we want to make the data fit a normal distribution pattern. It transforms the data on a logarithmic scale.

Understanding Quantiles

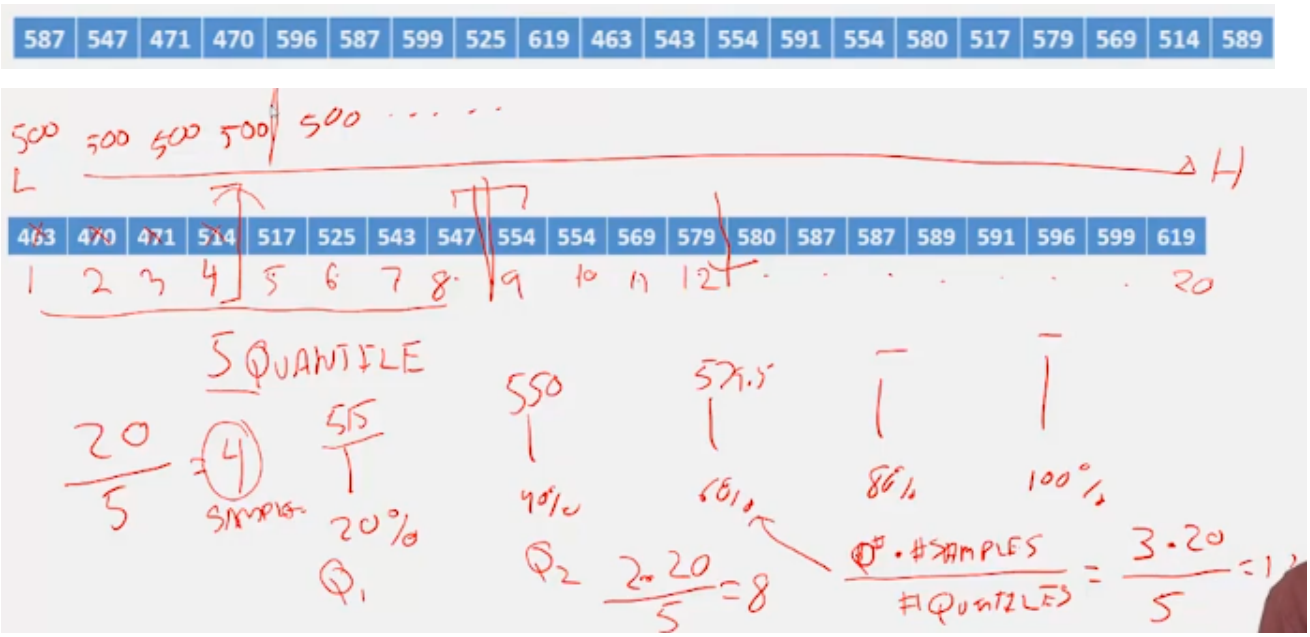
Quantiles

- Quantiles: points taken at regular intervals from the cumulative distribution function of a random variable



Calculating Quantiles

- Distribution of at-bats

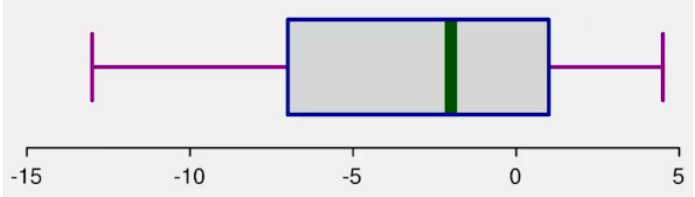


Quantiles

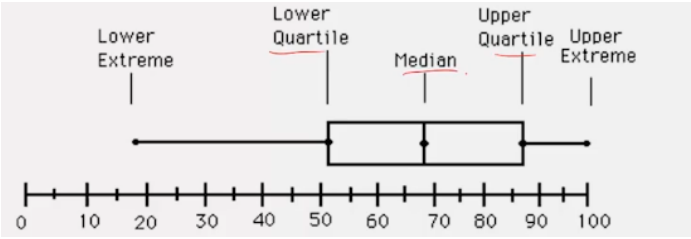
- Useful measures because they are less susceptible to long-tailed distributions and outliers
- May be more descriptive statistics than means and other moment-related statistics
- Quantiles of a random value are preserved under increasing transformations
- Can be used where only original data are available

Box and Whisker Plots

Box and Whisker Plot

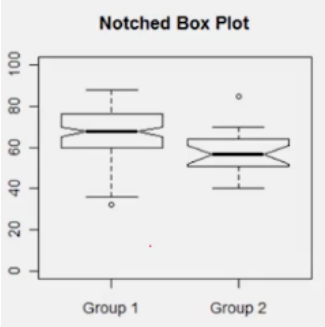


Summaries

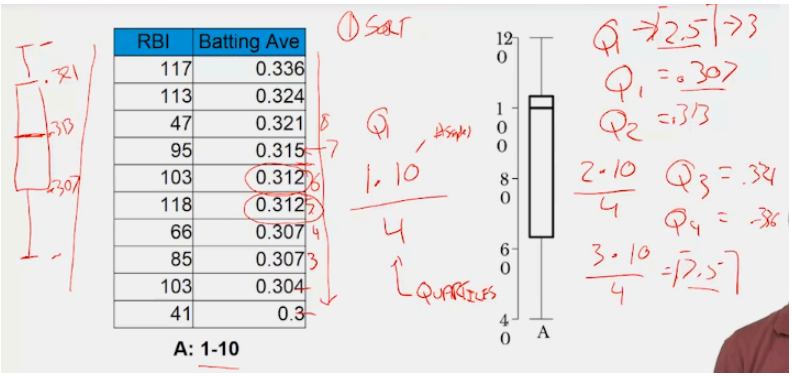


Alternate forms of Box and Whisker Plots

- Width of the box
  - mapped to the size of the group
  - can make width proportional to the square root of the group size
- Notched box plot
  - width of notches is proportional to IQR

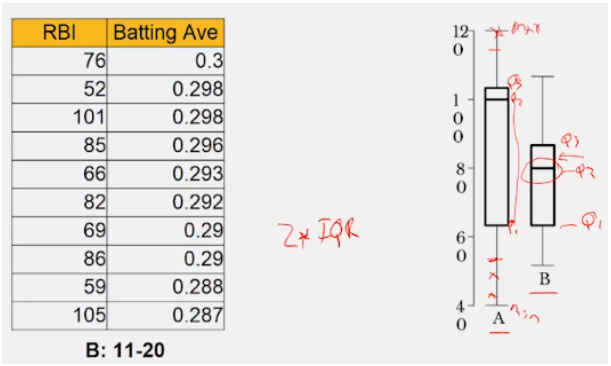


Box and Whisker Plots: Baseball Example



RBI	Batting Ave
117	0.336
113	0.324
47	0.321
95	0.315
103	0.312
118	0.312
66	0.307
85	0.307
103	0.304
41	0.3

A: 1-10

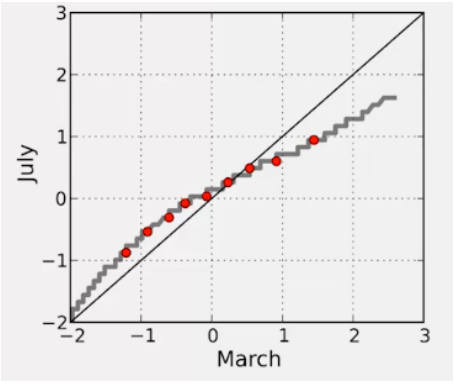


RBI	Batting Ave
76	0.3
52	0.298
101	0.298
85	0.296
66	0.293
82	0.292
69	0.29
86	0.29
59	0.288
105	0.287

B: 11-20

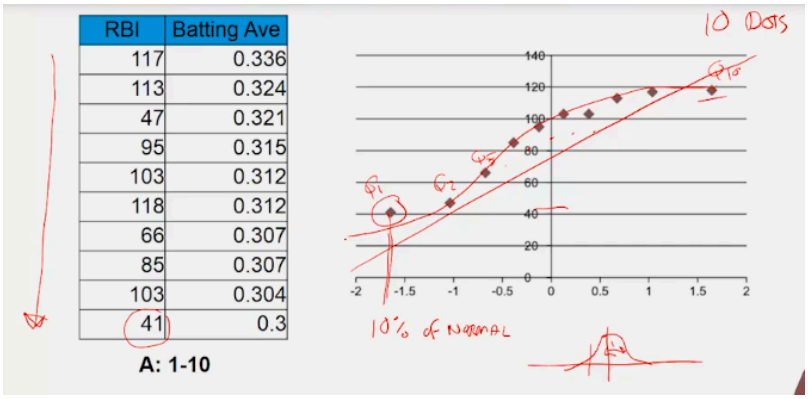
Q-Q Plots

Q-Q Plots



- More powerful comparing two distributions than histograms
- Sample sizes do not need to be equal
- Alternative is a probability plot

Creating a Q-Q Plot



Knowledge Check

- Which measurement requires the data to be ordered?
  - Quantiles
    - The first step to finding the quantiles of a dataset is to sort the data.
- Suppose that you are analyzing this dataset:
  19 23 26 30 33 35 38 38 40 42 45 45 47 56
  What is the value of the first quartile in this dataset?
  - 30
    - After the data are ordered, we can locate the first quartile, which covers the lower 25% of the data.
- Suppose that a dataset containing 36 data points is divided into 9 quantiles. What is the maximum number of data points that falls below the third quantile?
  - 12
    - This number is calculated by dividing the number of data points by the number of quantiles, and the obtained value is multiplied by the quantile number.

Module 2 Quiz Questions

- A data set containing 36 points is divided into 9 quantiles. What is the maximum number of data points that falls below the second quantile?
  - 8
    - 2/9 * 36 = 8
- Which type of graph should be used to visualize stock prices over a given period of time?
  - Line graph
- Suppose you are given a dataset, and you are required to visualize its median, lower extreme, and higher extreme, and outliers. Which type of visualization would be most suitable for this task?
  - Box plot
- What are the advantages of quantiles over other techniques? (Select all that apply)
  - Quantiles can be used in cases where only ordinal data are available
  - Quantiles are preserved under increasing transformations

- Quantiles are less susceptible to outliers

- For a data set of N points, the number of bins suggested by Sturge's formula is 5. What is the value of N? (Hint: use Log base 2)
  - 16
    - $k = 1 + log_2(N)$

      $5 = 1 + log_2(N)$

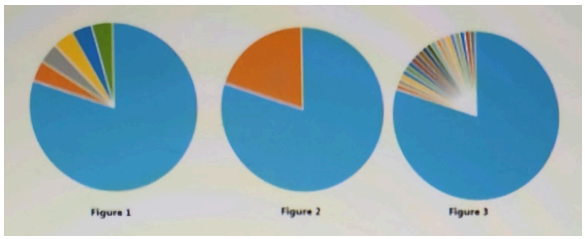
      $log_2(N) = 4$

      $N = 2^4 = 16$

- Which components are used in the visual representation of a graph? Select all that apply:
  - Scales reflecting the range of numbers
  - Labels describing what is being represented on the axis
  - Legends describing what makes up the graph

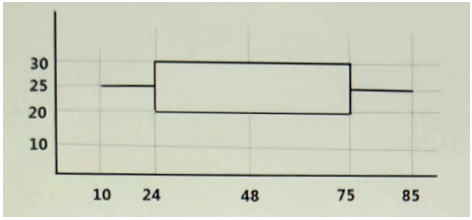
- Suppose that you have a sample of categorical data, and you want to create a visualization for it. Which chart would be best to choose for this data?
  - Pie chart

- Which pie chart above displays a categorical variable for which 80 of 100 rows are in one category and the remaining 20 rows are distributed evenly among five different categories?



  - Figure 1

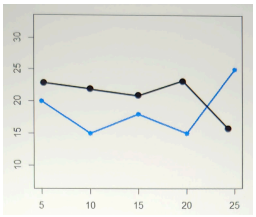
- What is the IQR in this box and whiskers plot?



  - 51
    - $IQR = Q3 - Q1$

      $IQR = 75 - 24 = 51$

- This chart displays the temperature changes in a month of a year. Which components of the graph are missing? Select all that apply:



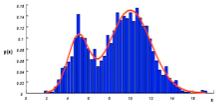
  - Axes labels
  - Units of measurements
  - Legend

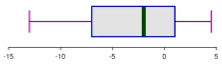
Introduction to Multivariate Analysis

Terms

- Multivariate Data
    - Any statistical technique used to analyze data from more than one variable
    - Used to process information in a meaningful way
- Curse of Dimensionality
    - The more dimensions in a visualization, the less effective standard computational and statistical techniques

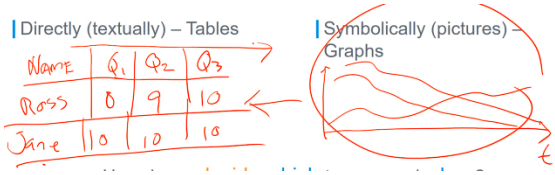
Univariate Visualization





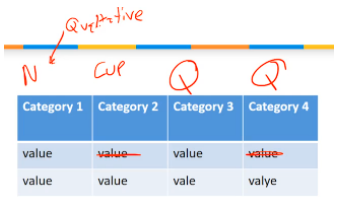
Representation

- What are the two main ways of presenting multivariate data sets?
    - Directly (textually) - Tables
    - Symbolically (pictures) - Graphs
- How do we decide which to use, and when?



Tables

- Use tables when:
    - The document will be used to look individual value
    - The document will be used to compare individual values
    - Precise values are required
    - The quantitative info to be communicated involves more than one unit of measure



Graphs

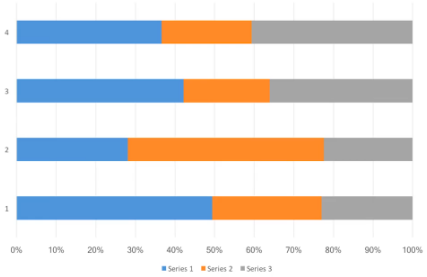
- Graph:
    - Visual display that illustrates one or more relationships among entities
    - Shorthand way to present information
    - Allows a trend, pattern or comparison to be easily comprehended

Task-Centric Graphing

- Why do you need a graph?

- What questions are being answered?
- What data is needed to answer those questions?
- Who is the audience?

Bivariate Case - Stacked Bars



Knowledge Check

- True or false? The more dimensions that a data set has, the less effective its standard computational techniques become.
    - True
        - With more dimensions in data, we end up with problems like P fishing, where you can find correlations between lots of random things. Most of them result in non-meaningful correlation
- Which scenario has the "curse of dimensionality" problem?
    - Recording the speed of a vehicle at every second for a year and saving data in a table where each speed value is a column (the vehicle is driven 10 hours per day)
        - The "Curse of Dimensionality" refers to the phenomenon where the standard computational and statistical techniques become less effective as the number of dimensions increases.
- Which tools can be used for multivariate data visualization? Select all that apply.
    - Stacked Bar Charts
        - Stacked bar charts are useful for representing multiple types of data on top of each other. It is mainly used for representing the parts of overall data.
    - Tables
        - Tables are useful for lookup and compare individual values. Multiple data types can be represented and communicated efficiently using tables.
- For which cases can tables be used for data visualization? Select all that apply.
    - Requiring precise data values for the records
        - Each record of the table holds the precise values of all attributes. These values can be used for lookup.
    - Communicating quantitative info for multiple units of measure
        - The values in the table can be both qualitative and quantitative. The data can be nominal, ordinal, etc. This quantitative information needs to be processed for communicating in different ways.
    - Comparing individual values

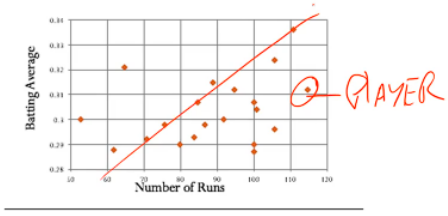
- Attributes are represented as columns, and each row of the attribute holds the individual values, which can be used for comparison.

Introduction to Scatterplots

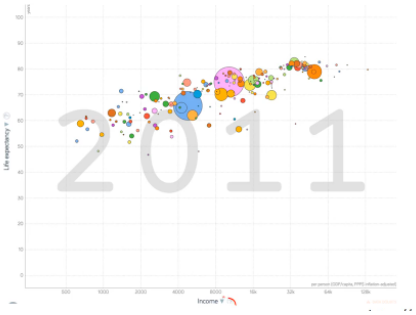
Bivariate Case - Scatterplots

- Visualizes discrete data values along two axes
- Used as a means of analyzing bivariate relationships
- Quick means of assessing outliers, clusters and distributions
- Putting a line through the data can help assess trends, but can also mislead viewer

Scatterplots



Multivariate Case - Scatterplot



Scagnostics: Scatterplot Diagnostics

- Graph-theoretic measures for detecting a variety of structural anomalies in a geometric graph representation of scatterplot data
- Ratings can be used to pick views that show particular structures that are of interest to the user
- Coined by Tukey, it is an exploratory graphical technique to help determine notable relationships between two variables

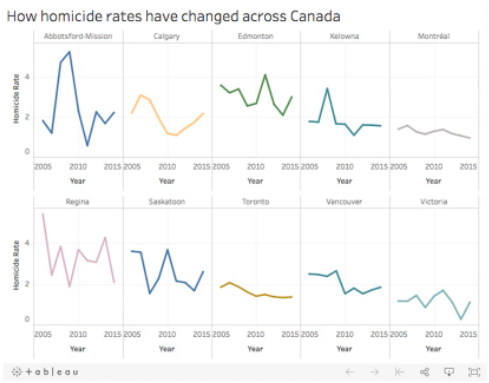
Scagnostics

- Wilkinson et al. propose nine agnostic measures to characterize the scatterplots:
    - Outlying
    - Sparse
    - Stiated
    - Skinny
    - Monotonic
    - Skewed
    - Clumpy
    - Convex
    - Stringy

Interaction

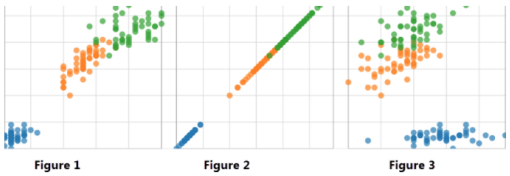
- Adding interaction allows viewers to visualize other combinations of variables

Creating Small Multiples



Knowledge Check

- True or false? Scatter plots help identify outliers.
    - True
        - The presentation of scatter plots are such that an outlying data point can be identified at a glance.
- Which statements are true about Scagnostics? Select all that apply.
    - They are used to detect anomalies in scatter plot data
    - It is used to help determine notable relationships between two variables
    - It is a graph-theoretic measure
    - THe name is coined by Tukey
- *Image: Three Scatterplots*



Review Image: Three Scatterplots. The diagram shows three separate scatter plots. Which figure represents the scatter plot in which X is plotted against itself?

- Figure 2
    - The straight line in Figure 2 indicates that the values of X are plotted against itself. This scatterplot shows a high correlation and a linear relationship between the values.
- What can be used to rearrange graph representations of scatterplots in a scatterplot matrix?
    - Scagnostics
        - Scagnostics (scatterplot diagnostics) is an exploratory graphical technique that can help determine notable relationships between two variables.
- Which statements accurately describe what small multiples can be used for? Select all that apply.
    - Small multiples make it possible to spot patterns easily
        - By displaying multiple variables on a single screen with trends and changes over time, it is possible to spot intriguing patterns in the data.
    - Small multiples show snapshots of events that change over time

        - By organizing the same data over different time periods, small multiples can show changes in the events over time.
    - Small multiples make it possible to scan rapidly across a trellis of small similar charts
        - Small multiples are used to show relationships between multiple variables in a single screen with information about trends and changes over time, which aids in scanning across multiple graphs quickly.

Mosaic Plots

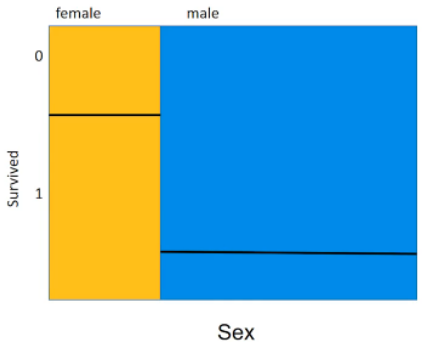
Introduction to Mosaic Plots

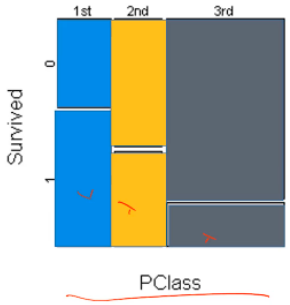
- Graphical display that allows you to examine the relationship among two or more categorical variables
- To create:
    - Start as a square with length one
    - Divide first into horizontal bars who widths are proportional to the probabilities associated with the first categorical variable
    - Next each bar is split vertically by the conditional probability of the second categorical variable

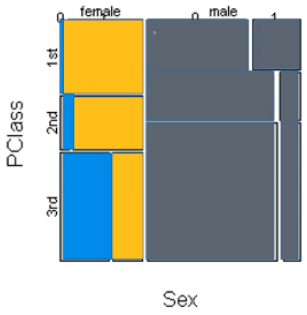
Example: Mortality Rates

Adults	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	57	140	118	4
2nd Class	14	80	154	13
3rd Class	75	76	387	89
Crew	192	20	670	3

Children	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	5	1	0	0
2nd Class	11	13	0	0
3rd Class	13	14	35	17
Crew	0	0	0	0





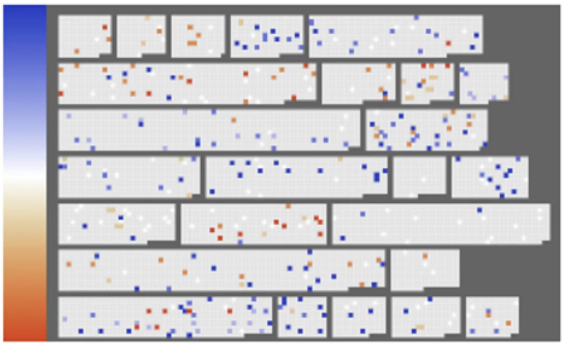


Mosaic Plots

- It is tempting to dismiss mosaic plots because they represent counts as rectangular areas and so provide a distorted perceptual encoding
- In fact, the important encoding is the length
- At each stage, the comparison of interest is of the length of the sides

Pixel Based Displays

Pixel Based Display Example

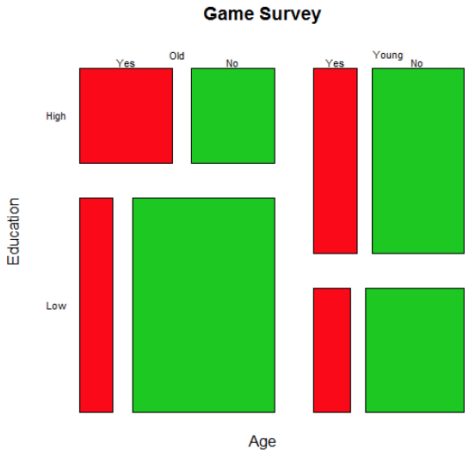


Designing Pixel Based Displays

- Could modify the pixel based display to incorporate components that will draw attention to the salient aspects of the data
    - Halo
    - Color
    - Distortion
    - Hatching

Knowledge Check

- True or false? The concept behind pixel-based display is that the most data we can represent is dependent on screen real estate.
    - True
        - Each pixel can represent an element of the data.
- *Figure: Game Survey Mosaic Plot*

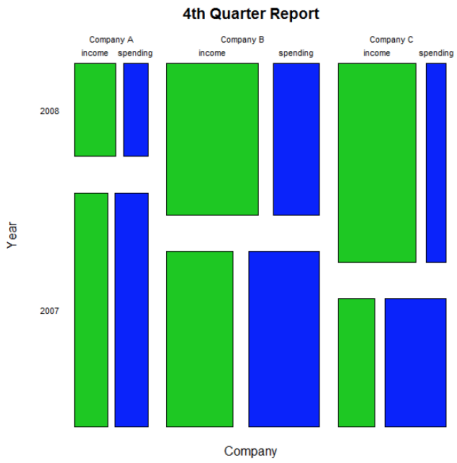


Review: *Figure: Game Survey Mosaic Plot*. A TV station sent a survey to a group of people to get data about who had watched the finals of a popular sports game on a certain channel. After receiving the responses, the TV station compiled the data in a mosaic plot that classifies the respondents by age, education, and whether they watched the game. Based on this plot, which conclusions can we make? Select all that apply.

- Younger respondents are more highly educated than older ones
    - The lengths of the mosaic plot representing the age and education level is higher for young respondents
- The game-watching habits of older respondents seem to be dependent on their education level
    - The length of highly educated older respondents that have game watching habits is greater compared to the length of the older respondents who are less educated.

- *Figure: 4th Quarter Report Mosaic Plot*

Review *Figure: 4th Quarter Report Mosaic Plot*. This plot shows financial information about three different companies. Which company had the largest income in 2008?



- Company C
  - The length of the mosaic plot in green, representing the income is higher for company C in 2008 compared to other companies.

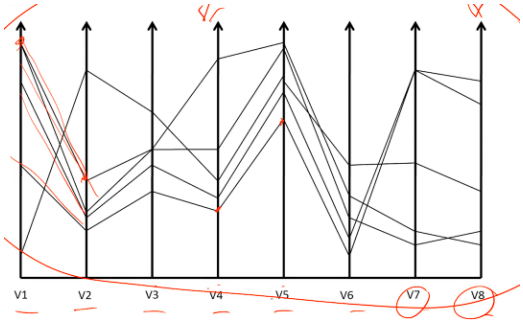
- Which statements about pixel-based displays are most accurate? Select all that apply.
  - Pixel-based displays are capable of displaying a large amount of data
    - One aspect about these displays is that they are capable of displaying a large amount of data.
  - Pixel-based displays can provide many details about the data
    - One aspect about these displays is that they can provide many details about the data.
  - Pixel-based displays can be augmented with techniques such as halos, distortion, and background coloring
    - Pixel-based displays can be boosted using techniques like halos for data elements, background coloring, distortion, and hatching.

Parallel Coordinate Plots

Parallel Coordinate Plots

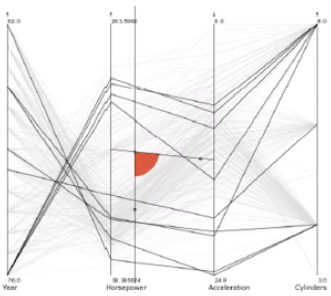
- Different variables can take different values with very different ranges
- Need to normalize data ranges
- Order of the parallel coordinate plots has a major impact on the resultant visualization
- The more variables we plot, the more lines we get and the more clutter that we get

Example

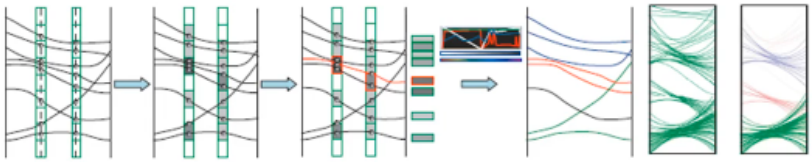


Angular Brushing

- Angle between axes indicates level of correlation
- Select subsets which exhibit a correlation along two axes by specifying angle of interest

- 

Visual Clustering

- Apply color and opacity based on line density
- Compute local density for each line by averaging the density values of all control points
- Apply color and opacity based on user specification
- 

Screen Space Metrics

- Creates lower-dimensional projections that provide maximum insight into the data and optimizes the parameter space for pixel-oriented visualizations
- Metrics based on a particular view of parallel coordinate plots
- Screen space metrics - depends on the size of the display
- Space between the azes is where interesting patterns occur

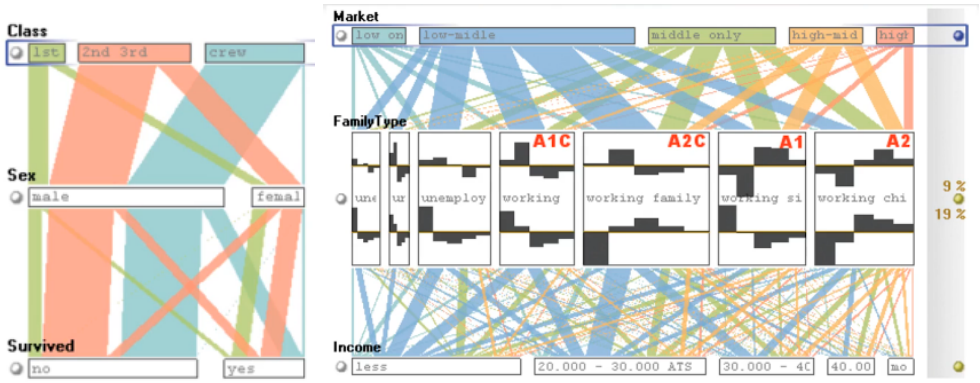
Screen Sizes Metrics

- Use a variety of metrics to try and optimize the use of the screen space
    - One-Dimensional Histogram Distance
        - records the slope of the lines between the axes
    - Two-Dimensional Axis pair Histogram
        - histogram of all the lines covering both axes
    - Line Crossings
        - interpret each line between a pair of axes as a directed interval
    - Angles of Crossing
        - determine angle between line crossings

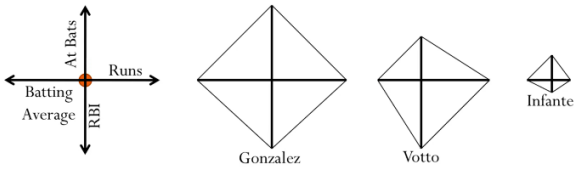
Parallel Sets

- Visualization method adopting parallel coordinate layout but uses frequency based representation
- Layout similar to parallel coordinate plots
- Continuous azes replaced with boxes

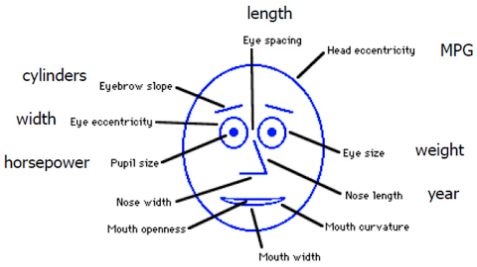
- Used for categorical data



Star Plot



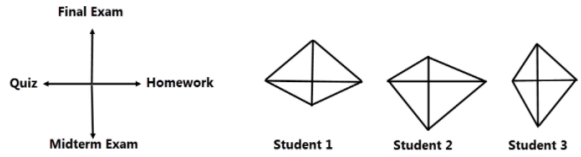
Multivariate Case: CHernoff Faces



Knowledge Check

- True or false? Parallel coordinate plots can get cluttered when more and more variables are plotted on one visualization.
    - True
        - More lines are drawn, so the visualization gets cluttered.
- True or false? The order of the axes on parallel coordinate plots does not have any impact on the visualization.
    - False
        - The order of the axes on a parallel coordinate plot greatly influences how the visualization looks.
- Which statements about properties of parallel coordinate plots are accurate? Select all that apply.
    - The correlation between axes can be used to create mathematical models for predictive analytics
        - Correlation can help in creating mathematical formulas like X=MY. These formulas can help in predicting the value of X for a given value of Y.
    - The order of axes influences the visualization
        - The order of axes would influence how the visualization looks. In a graph, downward trend, pairwise correlations, etc., can be totally misinterpreted by changing the order of the axes.
    - The angles between axes represent the level of correlation
        - The angle between axes represents the level of correlation, which can be used for creating mathematical models.
- What is the purpose of using angular brushing in parallel coordinate plots?

    - To select subsets of data that exhibit a correlation along two axes
        - By using the angles between the axes, it allows us to create chunks of data between axes as subsets and explore the correlation between them.
- Which metrics can be used to optimize the use of screen space in parallel coordinate plots? Select all that apply.
    - Histogram Distance
        - Histogram distance is used to measure the slope of lines between the axes, which determines how far two axes should be kept.
    - Line Crossings
        - Line crossings interpret each line between a pair of axes as a directed interval, which can be used to count how many times the lines crossed.
    - Angles of Crossings
        - Correct! Angles of crossing is used to measure the angle between each pair of crossed lines.
- Which dataset would be best to represent with a parallel set plot instead of a parallel coordinate plot?
    - A dataset of the number of “buy” or “no-buy” visitors to an online shopping website selling a T-shirt of five different sizes for men, women, and kids
        - The parallel set plot is preferred in this case because this is categorical data. The parallel set plot is a visualization method that adopts the parallel coordinate layout but uses frequency-based representation.
- *Figure: Student Score Star Plot*



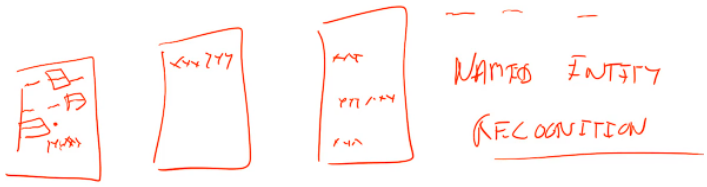
Review *Figure: Student Score Star Plot*. Based on this plot, which student had the lowest score on the final exam?

- Student 2
    - Student 2 scored the lowest on the final exam. It also appears that the final exam was this student's worst score.

Text Visualization

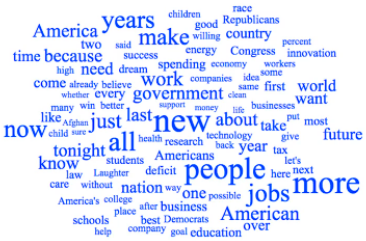
Text Visualization

- Visual representation for text data where words are placed and scaled based on some statistical measures
- Font size is typically determined by the number of instances a word is used



Example: Word Clouds

2002 State of the Union Address by U.S. President Bush



2011 State of the Union Address by President Obama

Knowledge Check

- Which graphical representation uses font size to correspond to the relative frequency of words?
    - Word Cloud
        - A word cloud is a visual representation of text data, usually used to give a summary overview. The frequency of the words represented in a word cloud is distinguishable by font size and color.
- *Diagram: Text Visualization*

Review *Diagram: Text Visualization*. The diagram gives a visualization of the text in a document. According to this visualization, which word occurs least frequently in the document?

- Network
    - Because this is a word cloud, we can assume that relative frequency is depicted by font size. Therefore, because the word "Network" has the smallest font size, we know that it is the word that occurs the least in the document.

- What are characteristics of word clouds? Select all that apply.
    - A word cloud can be used to study the differences and similarities between two documents
        - By identifying the relative frequency of words used in each document we can find similarities or differences between documents using word cloud.
    - The font size of a word in the cloud is typically determined by the relative frequency of that word in the document
        - Font size of a word determines the relative frequency of the word that appears in the document.

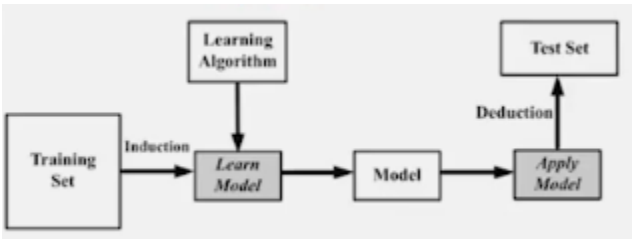
Supervised Learning

A Twitter Example

ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes

Supervised Learning: The Process

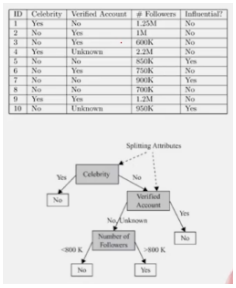
- We are given a set of labeled records/instances
  - In the format (X, y)
  - X is a vector of features
  - y is the class attribute (commonly a scalar)
- [Training] supervised learning task is to build a model that maps X to y
  - Find a mapping, m, such that $m(x) = y$'
- [Testing] Given an unlabeled instance (X'.?), we compute $m(X')$
  - E.g., spam/non-spam prediction



Supervised Learning Algorithms

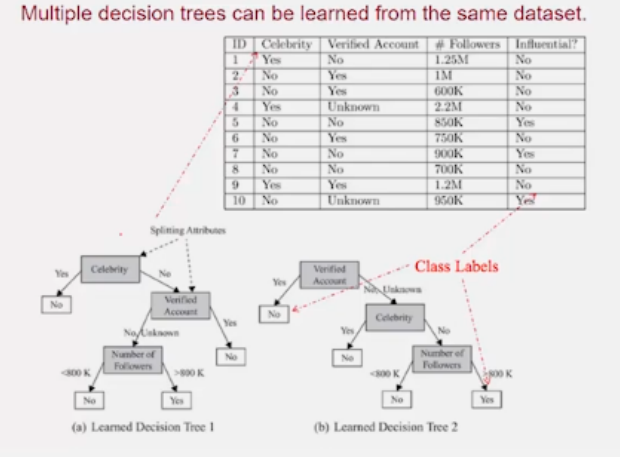
- Classification
  - Decision Tree Learning
  - k-Nearest Neighbor Classifier
- Regression
  - Linear Regression

Decision Trees



- A decision tree is learned from the dataset
  - Training data with known classes
- The learned tree is later applied to predict class attribute value of new data
  - Test data with unknown classes
  - Only feature values are known

Decision Tree Example



ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes

Decision Tree Construction

- Decision Trees are constructed recursively
- After selecting a feature for each node, different branches are created
- Training set is then partitioned into subsets based on feature values
- When selecting features, we prefer features that partition the set of instances into subsets that are more *pure*
- A *pure* subset has instances that all have the same class attribute value

Stopping Criteria for Decision Tree Induction

When reaching pure (or highly pure) subsets under a brach

- Decision tree construction process no longer partitions the subset
- Creates a leaf node under the branch
- Assigns class attribute value (or the majority class attribute value) for subset instances as the leaf's predicted class attribute value

Measuring Purity

- To measure purity we can use/minimize entropy
- Over a subset of training instances, T, with a binary class attribute (values in {+,-}), the entropy of T is defined as: $entropy(T) = \quad - p_{+} log(p_{+}) - p\_log(p\_)$

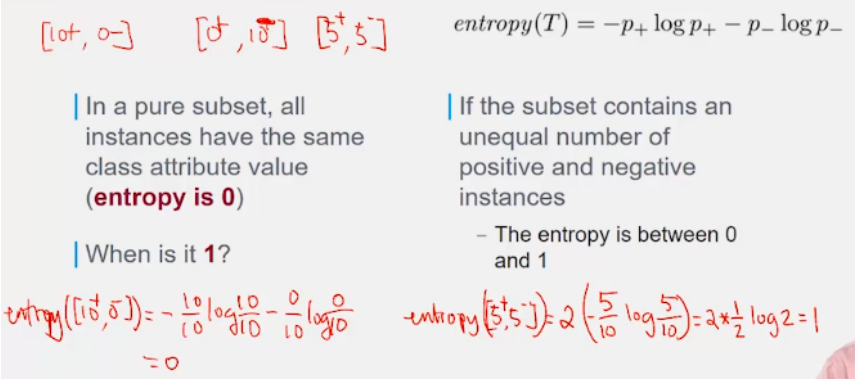
Entropy Example

Assume there is a subset T that has 10 instances:

- Seven instances have a positive class attribute value
- Three have a negative class attribute value
- Denote T as [7+, 3-]
- The entropy for subset T is: $entropy(T) = \quad - \frac{7}{10} log(\frac{7}{10}) - \frac{3}{10} log(\frac{3}{10}) = 0.881$

Entropy Values vs. Purity

- In a pure subset, all instances have the same class attribute value (entropy is 0)
- When is it 1?
- If the subset contains an unequal number of positive and negative instances
    - The entropy is between 0 and 1

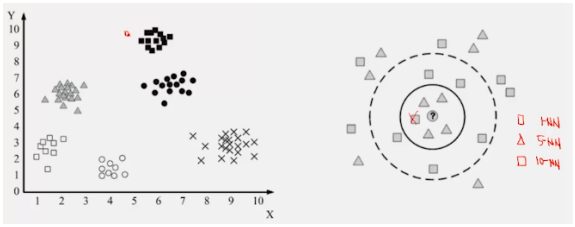


Supervised Learning: Nearest Neighbor

An Intuitive Illustration of *k*-NN

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

k-NN: Example



k-Nearest Neighbors

- *K*-nearest neighbors of *k*-NN
- Uses *k* nearest instances, called neighbors, to perform classification
- Instances being classified is assigned label (class attribute values) that majority of its *k* neighbors are assigned
- When *k*=1, the closest neighbor's label is used as predicted label for instance being classified
- To determine the neighbors of an instance, we need to measure its distance to all other instances based on some distance metric

k-NN: Example

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

Similarity between row 8 and other data instances;
(Similarity = 1 if attributes have the same value, otherwise similarity = 0)

Data instance	Outlook	Temperature	Humidity	Similarity	Label	K	Prediction
2	1	1	1	3	N	1	N
1	1	0	1	2	N	2	N
4	0	1	1	2	Y	3	N
3	0	0	1	1	Y	4	?
5	1	0	0	1	Y	5	Y
6	0	0	0	0	N	6	?
7	0	0	0	0	Y	7	Y

k-NN: Algorithm

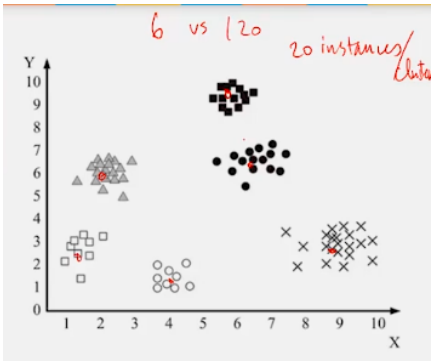
Algorithm 5.1 *k*-Nearest Neighbor Classifier

Require: Instance i, A Dataset of Real-Value Attributes, k (number of neighbors), distance measure d

1: **return** Class label for instance i
2: Compute k nearest neighbors of instance i based on distance mea- sure d.
3: l = the majority class label among neighbors of instance i. If more than one majority label, select one randomly.
4: Classify instance i as class l

k-NN: A Lazy Learning Algorithm

- Does *k*-NN learn?
- How fast is *k*-NN?

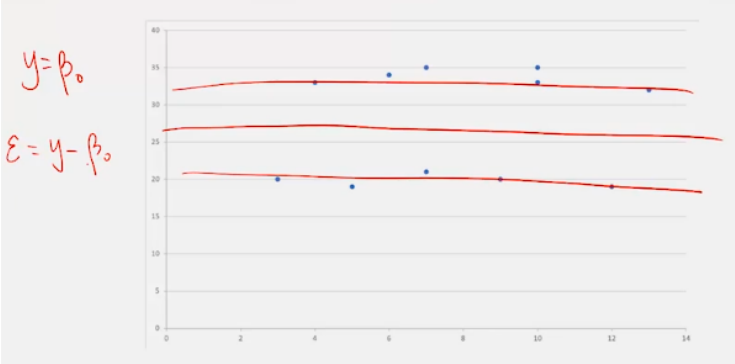


Supervised Learning: Regression

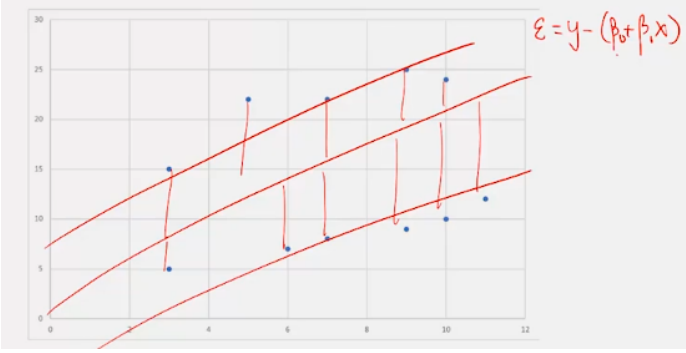
Regression

- $y = c$
- $y = \beta_0 + \beta_1 x$
- $\varepsilon = y - (\beta_0 + \beta_1 x)$

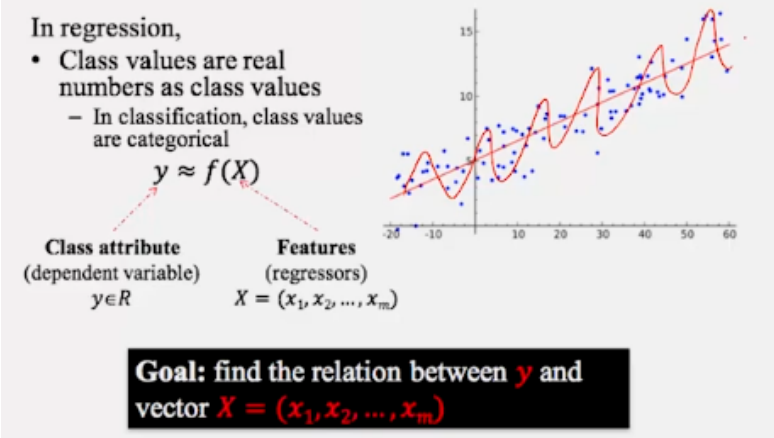
Approximation



Linear Approximation



Regression



Linear Regression

- Linear Regression
  - $Y = XW + \varepsilon$
  - we assume the relation between the class attribute Y and feature set X is linear
  - W represents the vector of regression coefficients
- Regression
  - $\varepsilon^2 = ||\varepsilon^2|| = ||Y - XW||^2$
  - can be solved by estimating W and epsilon using the provided dataset and the labels Y

    - "Least squares" is a popular method to solve regression

Knowledge Check

- True or false? In regression, class values are categorical.
    - False
        - In classification, class values are categorical. In regression, class values are real numbers.

Supervised Learning: Evaluation

Evaluating Supervised Learning: Why and How

- Training/Testing framework: A training dataset is used to train a model. The model is evaluated on a test dataset
- The correct labels of a test dataset are unknown
- When testing, the labels from this test set are removed

Some Basic Methods of Evaluation

Dividing the training st into train/test sets:

- Leave-one-out training
    - Use all instances but one to train and the on left out for testing
- k-fold cross validation training
    - Divide training set into k equally sized sets
    - Run algorithm k times
    - In round i, use all folds but fold i for training and fold i for testing
    - Average performance of algorithm over k rounds measures performance of algorithm

Measures used in Evaluation

- Class labels are discrete, measure accuracy by dividing number of correctly predicted labels (C) by total number of instances (N)
- More sophisticated approaches of evaluation:
    - AUC
    - F-measure
- $accuracy = \frac{C}{N}$
- $error\ rate = 1 - accuracy$

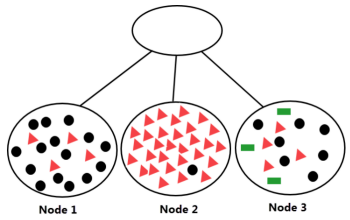
Beyond Accuracy Measure

- Labels cannot be predicted precisely
- Set a margin to accept or reject the predictions
    - Example: When the observed temperature is 71, any prediction range of 71 ± 0.5 can be considered as a correct prediction

- Use correlation between predicted labels and ground

Knowledge Check

- What kind of dataset would be used to evaluate a model?
    - Training datasets are used to train models; test datasets are used to evaluate the model.
- What does K-fold cross-validation training do? Select all that apply.
    - Run algorithm the same number of times as the number of sets
    - Divide training sets into some number of equally-sized sets
- Which of the below statements are valid about supervised learning? Select all that apply.
    - The goal in supervised learning is to fit a model that relates the responses (i.e.m classes) to the features
        - The objective of a supervised learning model is to predict the correct label for a newly presented input data.
    - For each observation of the feature measurement(s), x_i, $i = 1, ..., n$ there is an associated response measurement y_i
        - In supervised learning each input variable is mapped to an output or response variable. The mapping function is given as y=f(x) which is used for predicting the output variable for a newly created input.
- *Diagam: Tree Nodes*



Review *Diagram: Tree Nodes*. Which node of this tree has the smallest entropy?

- Node 2
    - A pure cluster or node will have instances that all have the same class attribute values. As Node 2 is more pure compared to Nodes 1 and 3 its entropy is less.

- *Table: Distance Between Data Points*

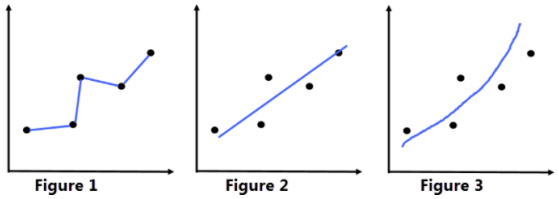
Y	X_1	X_2
Red	0	1
Green	-1	0
Red	1	0
Green	1	-1
Green	1	-1

Blue	1	1

Review Table: Distance Between Data Points. In this table, Euclidean distance is used to measure the distance between data points. What is the K-nearest neighbors prediction for Y, when $X_1 = 0, X_2 = 0$, and $K = 4$?

- Green or Red
  - K-NN would predict Red or Green when $K = 4$ as we can observe two instances of Red and Green occurring for $X_1 = 0,\ X_2 = 0$

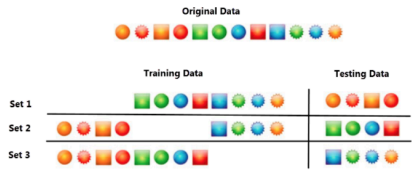
- What is one of the methods to speed up the K-NN algorithm when it is classifying a new item added to the dataset?
  - Using cluster representations for comparison
    - Cluster representations would represent the entire cluster as a datapoint, which would reduce the number of computations required to classify the new items added to the dataset.
- Which statements are correct for regression in supervised learning? Select all that apply.
  - "Least squares" is a popular method to solve regression
    - As mentioned and demonstrated in the lecture on regression, the "least squares" is a popular method for solving regression.
  - It can be solved by estimating the vector of regression coefficients (W) and epsilon, using the provided data set and labels Y
    - As mentioned and demonstrated in the lecture on regression, regression can be solved by estimating the vector of regression coefficients (W) and epsilon, using the provided data set and labels Y.
- *Diagram: Three Models*



Review *Diagram: Three Models*. Each figure represents a different model of a dataset. Which model is overfit to the data?

- Figure 1
  - Figure 1 demonstrates that the model is able to predict the output with extreme accuracy. Overfitting is a scenario where your model performs too well on the training data and poorly on test data.
- In supervised learning, how is testing data used? Select all that apply.
  - Testing data is used to determine whether the model is overfitted
    - Testing data is used to identify whether a model is overfit or generalized. An overfit model would perform too well on training data and poorly on test data.
  - Testing data is used to evaluate the model

        - Testing data is a subset used to represent the entire dataset. The trained model is executed on this data to evaluate its accuracy and performance.
- *Diagram: K-Fold Cross Validation*



Review *Diagram: K-Fold Cross Validation*. The diagram shows a dataset where k-fold cross validation is used for model construction. What is the value of k?

- 3
    - K-fold cross validation is a resampling strategy used to evaluate the model. The parameter k refers to the number of groups into which the given data is split. In the figure, the data are resampled into three groups.

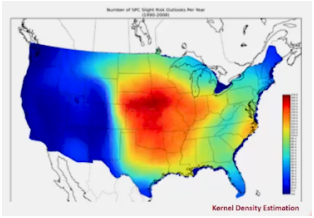
- Suppose you have a dataset of 100 points. If the leave-one-out validation technique is used, how many times does a model need to be fit?
    - 100
        - In the leave-one-out technique, we use all the instances but one for training. The one instance left is used for testing. If we have N instances, we use N-1 for training and 1 for testing.
- *Table: Disease Prediction Model Classes*

		Predicted Class	
		Disease	No Disease
True Class	Disease	30	40
	No Disease	50	60

Review *Table: Disease Prediction Model*. A model was built to predict if a patient has a disease or not, and this table shows the true classes and the model's predicted classes. What is the accuracy of the model?

- 50%
    - The model predicted the "disease" and "no disease" classes correctly for 90 records out of 180. Therefore, accuracy is 50%.

Introduction to Unsupervised Learning



Unsupervised Learning

- Clustering is a form of unsupervised learning
- Clustering algorithms group together similar items

Measuring Distance/Similarity in Clustering

- Clustering Goal: group together similar items
- Instances are put into different clusters based on distance to other instances
- Any clustering algorithm requires a distance measure

The most popular (dis)similarity measure for continuous features are ***Euclidean Distance*** and ***Pearson Linear Correlation***

$$d(X,Y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Similarity Measures

X and Y are n-dimensional vectors

$$X = (x_1, x_2, \ldots, x_n)$$
$$Y = (y_1, y_2, \ldots, y_n)$$

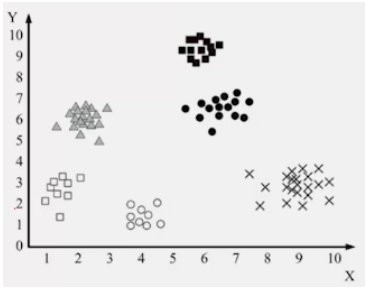
Measure Name	Formula	Description
Mahalanobis	$d(X,Y) = \sqrt{(X-Y)^T \Sigma^{-1}(X-Y)}$	X, Y are features vectors and Σ is the covariance matrix of the dataset
Manhattan (L_1 norm)	$d(X,Y) = \sum_i \lvert x_i - y_i \rvert$	X, Y are features vectors
L_p-norm	$d(X,Y) = (\sum_i \lvert x_i - y_i \rvert^n)^{\frac{1}{n}}$	X, Y are features vectors

Once a distance measure is selected, instances are grouped using it.

Clustering

- Clusters are usually represented by compact and abstract notations
- "Cluster centroids" are one common example of this abstract notation
- Partitional Algorithms (most common type):
    - Partition dataset into a set of clusters
    - Ech instance is assigned to a cluster exactly once
    - No instance remains unassigned to clusters

A 2-D Data with 6 Clusters



k-Means: An intuitive and common algorithm

Given data points x_i and an initial set of k centroids $m_1^1,\ m_2^1,\ \ldots,\ m_k^1$

- Assignment step:
    - Assign each data point to cluster S_i^t with closest centroid

        - Each data point goes into exactly one cluster
- Update step:
    - Calculate new mean to be the centroid of data points in cluster
- $S_i^t = \left\{x_p : ||x_p = m_i^t|| \leq ||x_p - m_i^t|| \forall 1 \leq j \leq k\right\}$

k-Means - the most commonly used

Algorithm 5.2 k-Means Algorithm

Require: A Dataset of Real-Value Attributes, k (number of Clusters)

1: **return** A Clustering of Data into k Clusters
2: Consider k random instances in the data space as the initial cluster centroids.
3: **while** centroids have not converged **do**
4: Assign each instance to the cluster that has the closest cluster centroid.
5: If all instances have been assigned then recalculate the cluster centroids by averaging instances inside each cluster
6: **end while**

- Also often used as a baseline algorithm for empirical comparison

When do we stop?

The procedure is repeated until convergence:

- Convergence:
    - Whether centroids are no longer changing
    - Equivalent to clustering assignments not changing
- Convergence:
    - Algorithm can be stopped when Euclidean distance between centroids in two consecutive steps is less than some small positive value

k-Means (alternative!)

As an alternative, k-means can be implemented to minimize an objective function.

- Example: squared distance error
    - x, is the jth instance of the cluster i
    - $n(j)$ is the number of instances in the cluster i
    - c, is the centroid of cluster i
    - $\sum_{i=1}^{k} \sum_{j=1}^{n(i)} ||x_j^i - c_i||^2$
- Stopping Criterion
    - When the difference between objective function values of two consecutive iterations of k-means algorithm is less than some small value

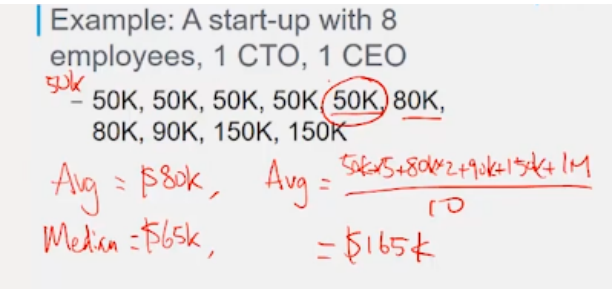
k-Means Discussion

- Finding global optimum of k partitions is computationally expensive (NP-hard)

- This is equivalent to finding optimal centroids that minimize objective function
- Solution: efficient heuristics
- Outcome: converge quickly to a local optimum that might not be global
    - Example: running *k*-means multiple times

Means vs. Medians

- Means is the average
- Median is the middle value
- Example: A start-up with 8 employees, 1 CTO, 1 CEO
    - 50K, 50K, 50K, 50K, 50K, 80K, 80K, 90K, 150K, 150K
    - Average = $80K
    - Median = $65K
- Should we use means or medians?
- Which one is easier to update?

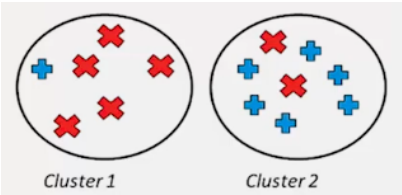


Knowledge Check

- True or false? Grouping together similar items is a form of unsupervised learning.
    - True
        - Clustering algorithms group similar items, and clustering is a form of unsupervised learning.
- Which of the following statements is true for clustering? Select all that apply.
    - Clustering algorithm requires a dissimilarity measure
    - Instances are put into different clusters based on the distance to other instances
    - Euclidean distance is one of the most popular dissimilarity mearues for continuous features
    - The closer instances are, the more similar they are
- True or false? Euclidean distance is used to calculate the co-variance matrix.
    - False
        - Mahalanobis distance is used to calculate co-variance matrix.

Unsupervised Learning: Evaluation

Evaluating the CLusterings

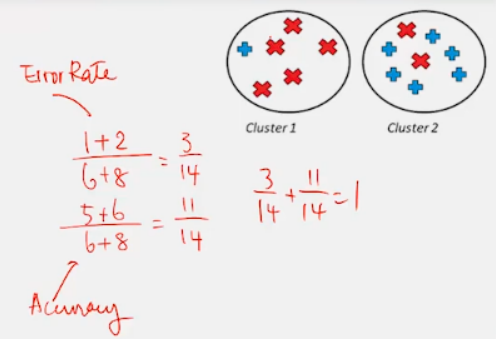


- We are given two types of objects
    - In perfect clustering, objects of the same type are clustered together

Evaluation with Ground Truth

When ground truth is available

- We have prior knowledge on what the clustering should be, or the correct clustering
    - We can use Accuracy to measure
- But, what is the use of clustering?

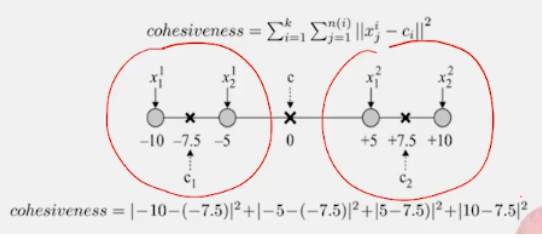


Evaluation without Ground Truth

- Cohesiveness
    - In clustering, we are interested in clusters that exhibit cohesiveness
    - In cohesive clusters, instances inside the clusters are close to each other
- Separateness
    - We are interested in clusters that are well separated from one another

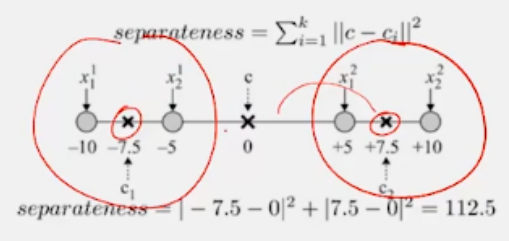
Cohesiveness

- Being close to the centroid of the cluster



Separateness

- Cluster centroids being far from the man of entire dataset



Silhouette Index

We are interested in clusters that are both cohesive and separate

- Silhouette Index

- It compares:
  - Average distance value between instances in the the same cluster
- To:
  - Average distance value between instances in the same cluster
- In a well-clustered dataset
  - Average distance between instances in the same cluster is small (cohesiveness)
  - And average distance between instances in different clusters is large (separateness)

For any instance x that is a member of cluster C:

- $a(x) = \frac{1}{|C|-1} \sum_{y \in C,\, y \neq x} ||x - y||^2$
- Compute the within-cluster average distance

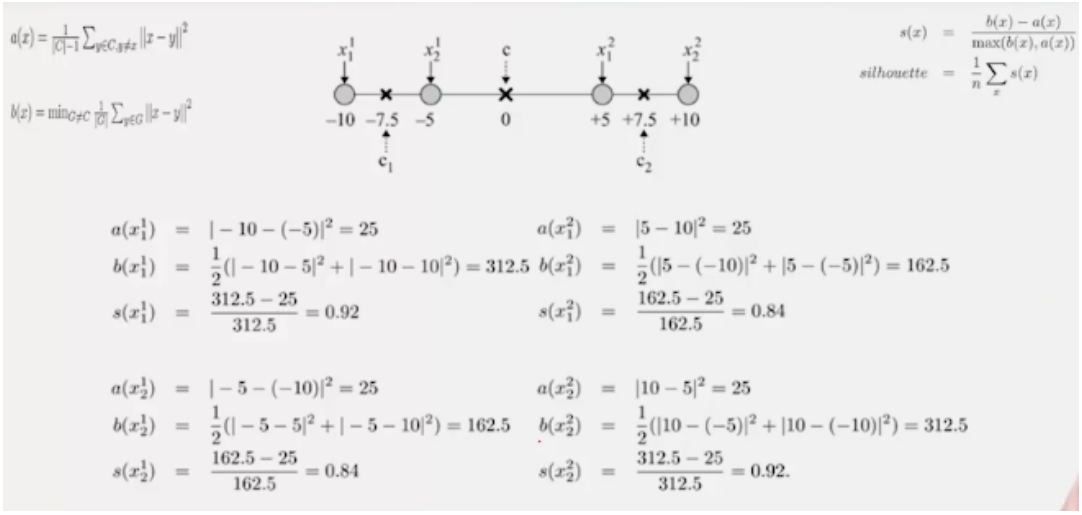
Compute the average distance between x in C and instances in cluster G

- $b(x) = min_{G \neq C} \frac{1}{|G|} \sum_{y \in G} ||x - y||^2$
- G is closest to x in terms of the average distance between x in C and members of G

Our interest: clusterings where $a(x) < b(x)$

- Silhouette can take values between $[-1, 1]$
- The best case happens when for all $x, -a(x) = 0, b(x) > a(x)$
- $s(x) = \frac{b(x)-a(x)}{max(b(x), a(x))}$
- $silhouette = \frac{1}{n} \sum_{x} s(x)$

Silhouette Index: Example



Knowledge Check

- True or false? We can use accuracy when ground truth is available.
    - True
        - In order to calculate the accuracy, we need to have the ground truth available.
- True or false? In clustering, we are only interested in cohesiveness, not separateness.
    - False
        - We are interested in cohesiveness among the instances within a cluster and separateness between different clusters.
- True or false? Silhouette index can take values between [-1, 1].
    - True
        - Silhouette is an average of the distances in the clusters. The best case happens when silhouette is 1, meaning the distance with the cluster is 0 and the distance between the clusters is higher.
- *Table A: X and Y Coordinates*

(x,y)
(1, 1)
(2, 4)
(3, 3)
(3, 4)
(4, 2)
(4, 4)
(5, 3)

Table B: Cluster 1

Cluster 1
(1, 1)
(2, 4)
(3, 4)

Table C: Cluster 2

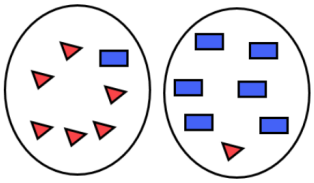
Cluster 2
(3, 3)
(4, 2)
(4, 4)
(5, 3)

Review *Table A: X and Y Coordinates, Table B: Cluster 1,* and *Table C: Cluster 2*. The x-y coordinates given in Table A are classified using k-means. The assignment step of the k-means algorithm creates Cluster 1 (in Table B) and Cluster 2 (in Table C). What are the new centroids of Cluster 1 and Cluster 2?

- (2, 3) and (4, 3)
    - The new centroids (2,3) and (4,3) are formed by taking the mean of all the points in Cluster 1 and Cluster 2, respectively.

- Which criteria can be used as stopping criteria for k-means classification? Select all that apply.
    - The centroids have stopped changing

        - If the centroids remain unchanged after multiple iterations, we can say that the algorithm is not learning any new patterns, so it can be used as a stopping condition.
    - Few than 0.00001% of the points are shifting from one cluster to another
        - If the points remain in the same cluster even after training the algorithm for multiple iterations, it can be used as a stopping condition.
    - No points are shifting from one cluster to another
        - If the points remain in the same cluster even after training the algorithm for multiple iterations, it can be used as a stopping condition.
- In unsupervised learning, one particular task involves dividing a dataset into homogenous groups. What is the name of this particular task?
    - Clustering
        - Clustering is the process of grouping similar data points together such that elements within the group are more similar to each other than to elements in other group.
- *Diagram: K-Means Algorithm Classification*



Review *Diagram: K-Means Algorithm Classification*. The diagram shows a dataset that was classified into two groups by the K-means algorithm. What is the accuracy of the algorithm?

- $\frac{6}{7}$
    - In the figure, 12 out of the 14 data points are classified accurately.
- What is the relationship between the number of clusters and cohesiveness in a clustering algorithm?
    - As the number of clusters increases, the value of cohesiveness decreases
        - As the number of clusters increases, the value of cohesiveness decreases.
- Which value would be the silhouette index of a dense and well-separated cluster?
    - 0.9
        - The silhouette index lies between [-1,1], with an index closer to 1 meaning that the clusters are dense and nicely separated.

Module 3 Quiz Questions

- What is the name of the unsupervised learning task of dividing a dataset into homogeneous groups?
    - Clustering
- What are examples of a supervised learning algorithm? Select all that apply:
    - Linear Regression
    - Decision Trees
    - K-Nearest Neighbors
- What are disadvantages of scatter plots? Select all that apply:

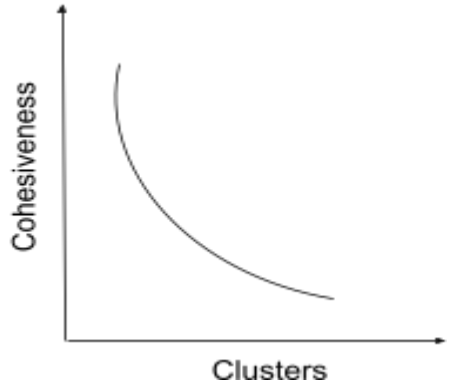
    - Scatter plots take up a lot of real estate
    - Scatter plots are not a great technique to observe trends in time series data
    - Scatter plots get congested and hard to read when the data is large
- Which measures are used in unsupervised model evaluation? Select all that apply:
    - Separateness
    - Cohesiveness
- *Figure 1: Model 1 Performance*

	Predicted: Disease	Predicted: No Disease
True: Disease	30	50
True: No Disease	60	60

Figure 2: Model 2 Performance

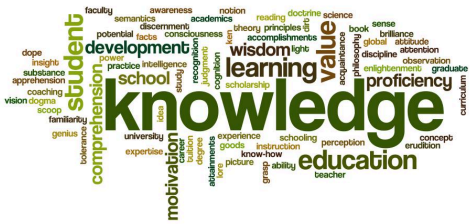
	Predicted: Disease	Predicted: No Disease
True: Disease	100	90
True: No Disease	0	10

Review *Figure 1: Model 1 Performance,* and *Figure 2: Model 2 Performance*. These two models were built to predict whether a patient has a disease or not. The tables show the true classes and the predicted classes of these two models. Following the performance measure of "accuracy" introduced in this course, what is the accuracy of Model 1?

- 45%
    - $accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$
    - Total = 30 + 50 + 60 + 60 = 200
    - $accuracy = \frac{30+60}{200} = \frac{90}{200} = 0.45 = 45\%$
- Of the figures, which one describes the relationship between the number of clusters and cohesiveness in a clustering algorithm?
    - 

        - As the number of clusters increase, the value of cohesiveness decreases

- Of this word cloud, which word appears the most?



  - knowledge

Temporal Analysis

Temporal Analysis

- “Time is an outstanding dimension reflected by Schneiderman’s Task by Data Type Taxonomy.”

Time-Oriented Data

- TIme oriented data is ubiquitous
  - Stock markets
  - Movie trends
  - Business
  - Medicine
- Each data case is likely an event of some kind, with one variable being the date and time

Time Series

- “A ransome selection of 4000 graphs from 15 newspapers and magazines worldwise showed that between 1974 and 1980, 75% of these graphs were time series.” - E. Tufte 1983
- What questions can we ask of these visuals?
  - Does a data object exist at a certain time?
  - When does a certain data object exist?
  - How long does a data object exist?
  - How fast and how much does the data object change?
  - What order do objects appear/disappear?
  - Is there a cyclical pattern to appearances?
  - WHich objects exist simultaneously?

Time is...

- Ordered
- Continuous
- Cyclical
- Independent of location

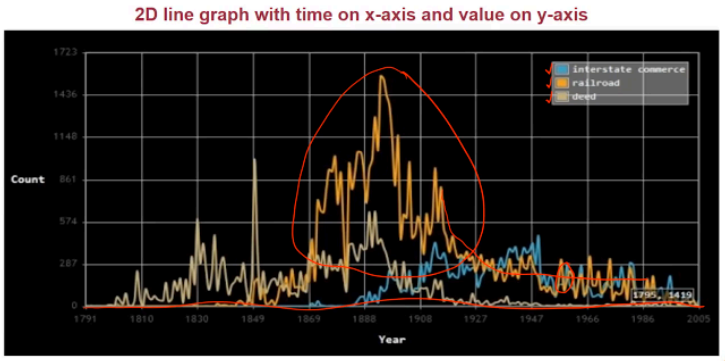
Linear vs. Cyclical Time

- Linear Time

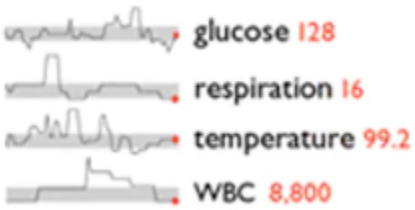
    - One time point precedes another
    - Time being ordered is locally bound to notion of causality
- Cyclical TIme
    - The ordering of points in a cyclic time domain would be meaningless
    - Winter comes before summer, but also after summer

Temporal Analysis and Visualization

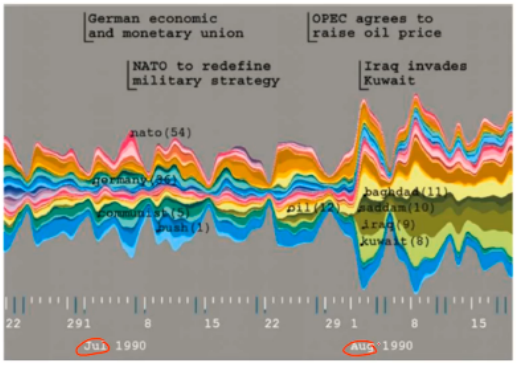
Linear Time



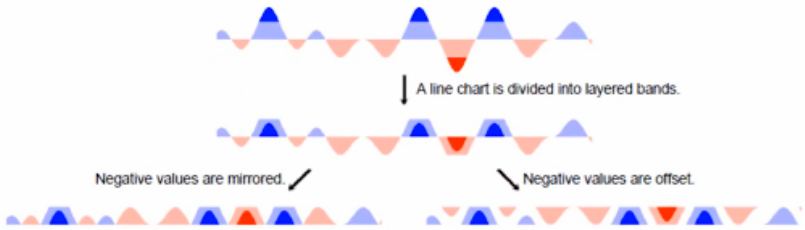
"Intense, simple, wordlike graphics"



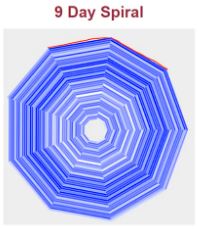
Theme River

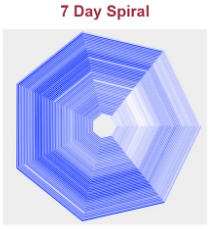


Sizing the Horizon

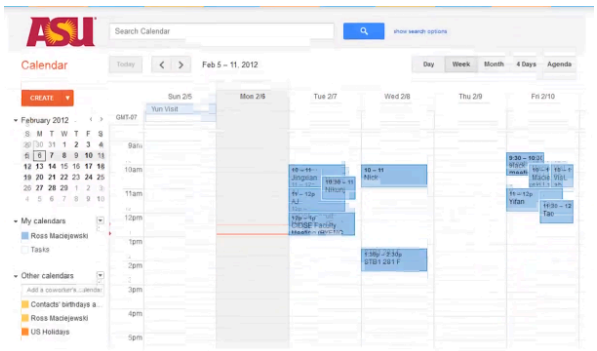


Spiral Graph



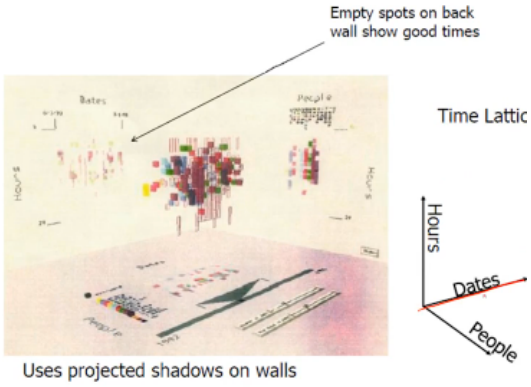


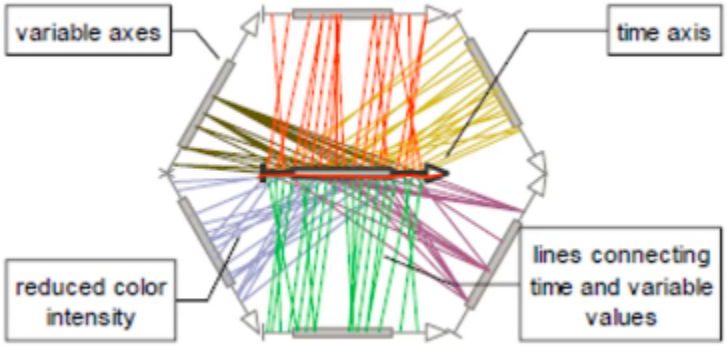
Calendar Visualization



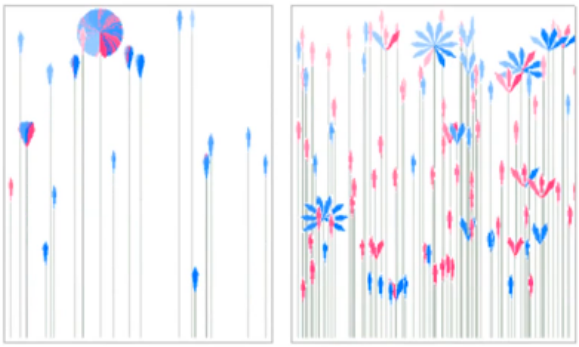
Spiral Calendar

Time Wheel

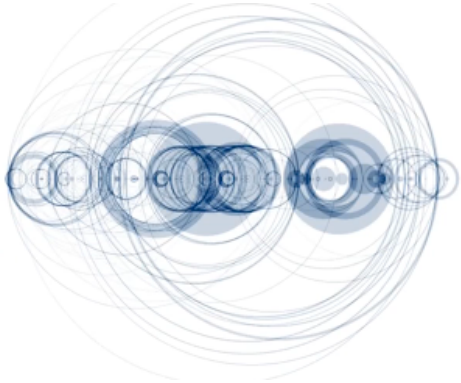




People Gardens



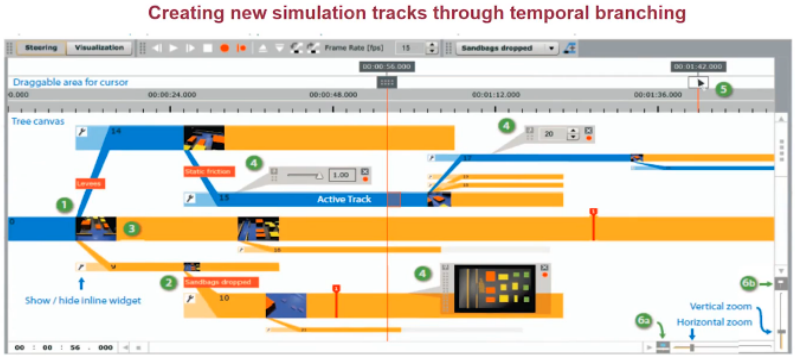
Arc Diagrams



Ordered Time vs. Branching Time

- Ordered time domains consider things that happen one after another
- Branching time considers multiple what-if scenarios, allowing comparison of alternate scenarios

World Lines



Design Principles

- Show familiar visual representations whenever possible
- Provide side-by-side comparisons of small multiple views
- Spatial position is strongest visual cue
- Multiple views are more effective when coordinated through explicit linking
- Avoid abrupt visual change

Control Chart Analysis

Data Mining

- Data Mining domain has techniques for examining time series. Looking for
    - patterns
    - anomalies
- Enhance the visualizations
    - show what is important
- Used in exploratory analysis
    - “I think this looks interesting, show me similar trends.”

Typical Time Series Analysis

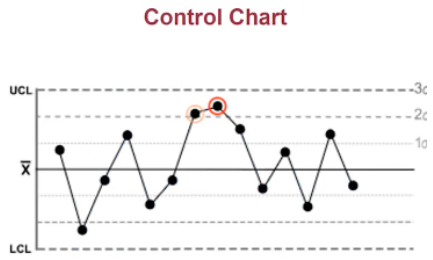
- Trend analysis
    - A company’s linear growth in sales over the years
- Seasonality
    - Sales are higher in summer than winter
- Forecasting
    - What is expected sales next quarter

Control Chart Overview

- Control Chart Components
    - For temporal data, we can find anomalies using control chart methods
    - Control charts consist of a statistic representing some measurement in time
- Calculations
    - The mean and standard deviation of the statistic is calculated given all the available samples
    - If the current value is greater than some pre-set number of standard deviations from the mean, then an alert is generated

What is a control chart?

- A graph used to study how a process changes over time
- Data are plotted in time order
- Always has a central line for average, an upper line for upper control limit and a lower line for lower control limit
- Lines are determined from historical data

- 

When to use a control chart?

- Controlling ongoing processes by finding and correcting problems as they occur
- Predicting the expected range of outcomes from a process
- Determining whether a process is stable (in statistical control).
- Analyzing patterns of process variation from special causes or common causes
- Determining whether the quality improvement project should aim to prevent specific problems or to make fundamental changes to the process

Control Chart Model

- Upper Control Limit: $\mu + k\sigma$
- Center Line: μ
- Lower Control Limit: $\mu - k\sigma$

Moving Average/Range Charts

- Moving Average Chart
  - monitors the process location over time
  - generally used for detecting small shifts in the process mean
  - control limits are derived from average range on Range Chart
- Range Chart
  - monitors the process variation over time
  - should reviewed before Moving Average Chart

Moving Average: Stock Market Closing Example

- Daily Closing Prices: $11,\ 12\ 13\ 14\ 15\ 16\ 17$
- First day of 5-day SMA: $(11 + 12 + 13 + 14 + 15)/5 = 13$
- Second Day of 5-day SMA: $(12 + 13 + 14 + 15 + 16)/5 = 14$
- Third Day of 5-day SMA: $(13 + 14 + 15 + 16 + 17)/5 = 15$

Exponentially Weighted Moving Average

- SMA: $10\ period\ sum/10$
- Multiplier: $(2/(Time\ periods + 1)) = (2/(10 + 1)) = 0.1818 = 18.18\%$
- EMA: $\{Close - EMA(previous\ day)\} \times multiplier + EMA(previous\ day)$

The Log Factor

- Shorter MOving
    - nimble and quick to change
- Longer Lag
    - Longer the moving average, more the lag
- Longer Moving
    - Longer moving - slow to change
- Differences between simple moving averages and exponential moving averages, one is not necessarily better than the other
- Length of your moving average depends on your analytical goal

Knowledge Check

- True or false? Repeating patterns can be easily seen in linear plots.
    - False
        - Trends are easily seen on linear plots, but not repeating patterns. For repeating patterns, spirals can be used, as they can easily represent the idea of repetition.
- True or false? Ordered time allows comparison of alternate scenarios.
    - False
        - Branching time considers multiple what-if scenarios, allowing comparison of alternate scenarios.
- Select all that apply. How does data mining aid visualizations?
    - Data mining domain has techniques for examining time series to look for anomalies
    - Data mining domain has techniques for examining time series to look for patterns
    - Data mining techniques enhance the visualization to show what is important
- True or false? Control chart is a graph used to study how a process changes over time.
    - True
        - Control charts can show a number of patients coming to a hospital over time, the quality of parts coming out of a manufacturing facility over time, etc.

Introduction to Strings and Sequences

Strings, sequences, time series

- A string of sequence, $S = (c_1, c_2, ..., c_n)$, is a finite sequence of symbols

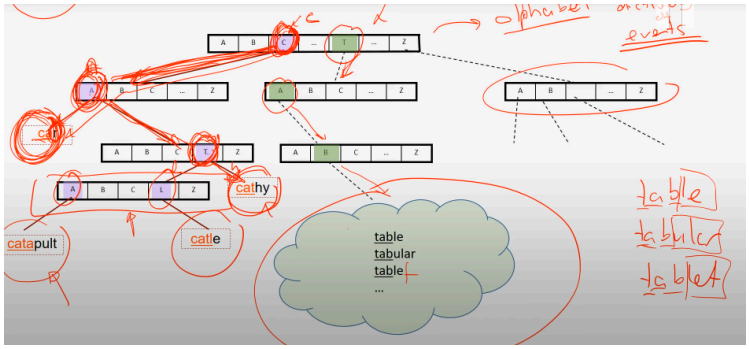
```
abcbbbaabbaabcbbbaaabbc
```

String/sequence Matching and Search

- Prefix Search:
    - Find all strings that start with “tab”:
        - “Table”; “tabular”; “tablet”; ...
- Subsequence Search:
    - Find all strings that contain the subsequence “ark”,
        - “Marketing”; “spark”; “quark”
    - Find all occurrences of “acd”;
        - “Aabacdcdabdcababdacddcab”
- Sequence Similarity
    - “table” vs. “cable”?
    - “table” vs. “tale”?
    - “table” vs. “tackle”?

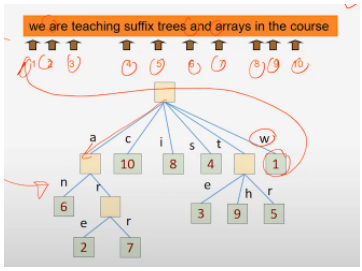
Strings and Sequences: Prefix Search

Trie Data Structure



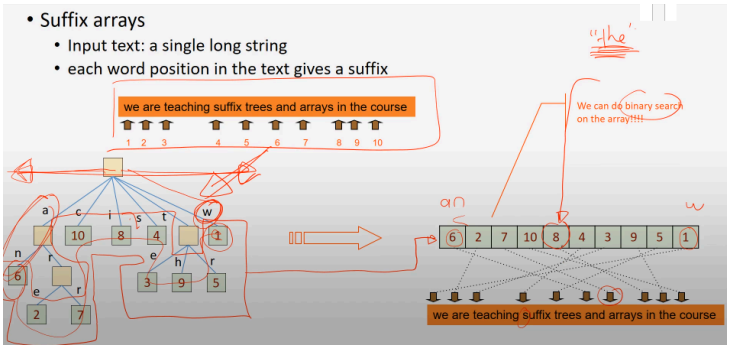
Strings and Sequences: Subsequence Search and Suffix Trees

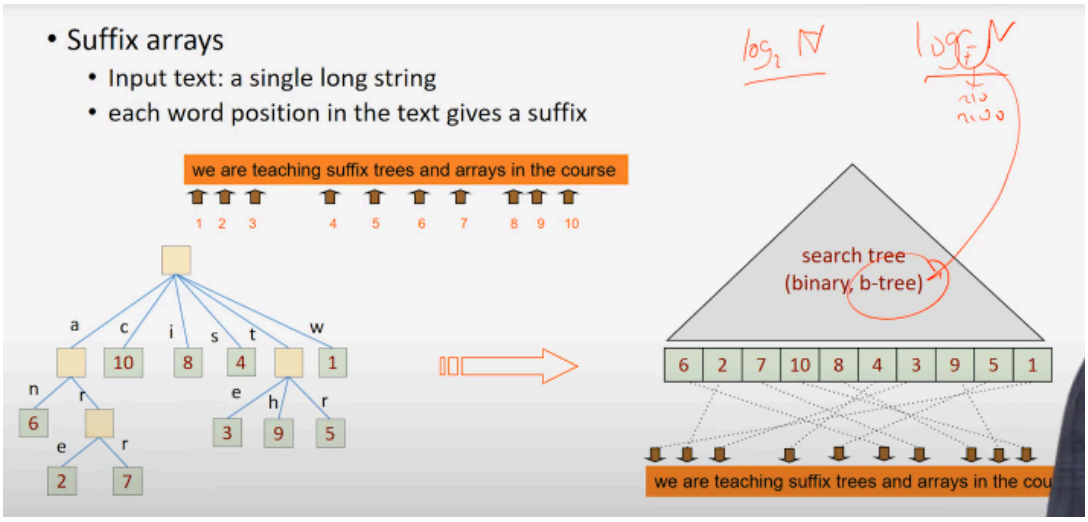
Suffix Trees

- Input text: a single long string
- each word position in the text gives a suffix
    - 

Strings and Sequences: Suffix Arrays

Suffix Arrays



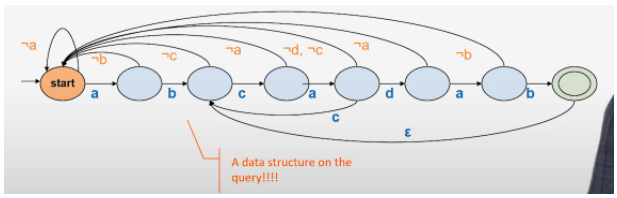


Strings and Sequences: Subsequence/Pattern Search

Subsequence/Pattern Search

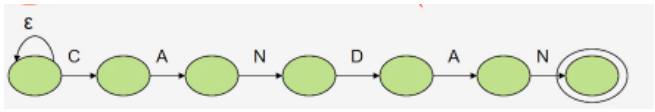
- What if we are not given the sequence in advance; can we do search without a data structure on the sequence?
  - Yes, scan the sequence…
  - …but, we have seen that this is expensive
    - Given a sequence of length N, and pattern of length M
    - Cost: $O(NxM)$
- If we are given the pattern in advance, can we create a data structure on the pattern, instead

Knuth-Morris-Pratt (KMP)

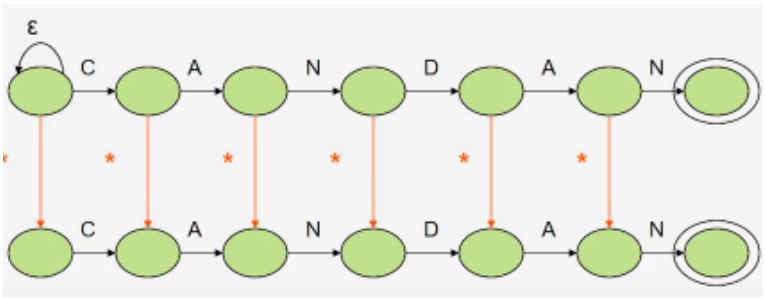
- Given a sequence of length N, and pattern of length M
- Knuth-Morris-Pratt: $O(N)$
- Example
  - Pattern: abcadab
  - 

Strings and Sequences: Approximate Matches

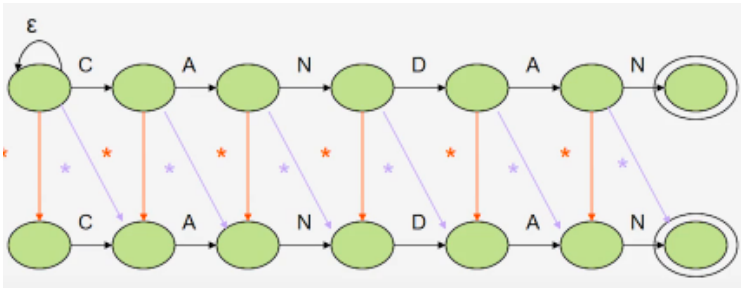
What about approximate matches?

- Example
  - Pattern: CANDAN
  - 

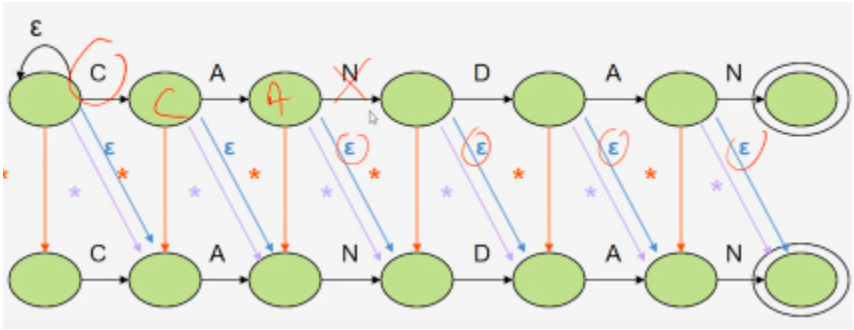
NFA... upto 1 insertion



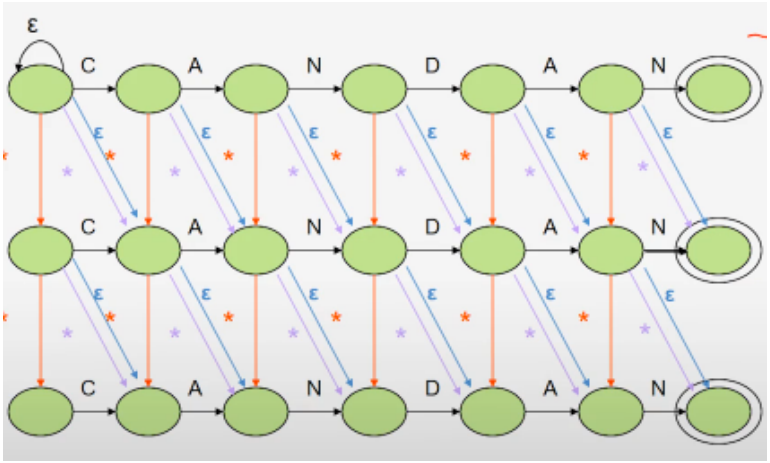
NFA... upto 1 insertion\replacement



NFA... upto 1 insertion\replacement\deletion



NFA... upto 2 insertion\replacement\deletions



Summary

- Prefix based sequence exploration
    - Trie data structure helps prune the candidate set
- Subsequence search and exploration
    - Suffix trees and suffix arrays helps focus on the part of a long sequence
- Pattern matching
    - Non-deterministic finite can be used to support exact and approximate pattern matching

Knowledge Check

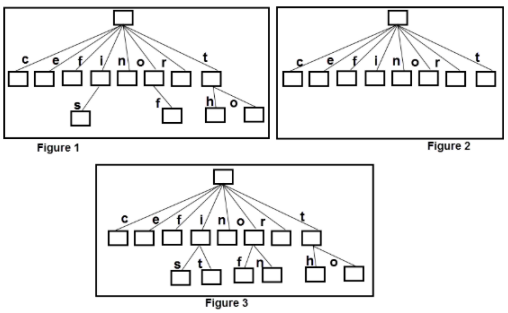
- True or false? In prefix search, we try to find all the matching strings in the database which starts with the given pattern.
    - True
        - In prefix search we are given a database, and we are given a pattern and we try to match the pattern by finding them in the prefixes.
- Select all that apply. Which of the following statements is true about Trie data structure?
    - It has edges
        - Trie data structure essentially is a tree that is available to implement prefix search efficiently.
    - It is a tool to implement prefix search efficiently
        - Trie data structure essentially is a tree that is available to implement prefix search efficiently.
    - It has nodes
        - Trie data structure essentially is a tree that is available to implement prefix search efficiently.
- True or false? Using trie data structure, given a pattern, the rest of the database can be pruned efficiently so that there is no need to consider those strings that are not relevant.
    - True
        - This is one of the advantages of the Trie data structure.
- Select all that apply. Which are true for subsequence search?
    - Finds the query pattern anywhere in the database

        - Subsequence search is similar to prefix search, but looks for the query patterns anywhere in the sequence and the database.
- True or false? Brute force approach for subsequence search is to align the pattern for each position in the sequence.
    - False
        - Brute force approach for subsequence search is to scan the sequence and align the pattern for each position in the sequence.
- True or false? In order to have an efficient subsequence search, we can employ the core approach in the two data structures of suffix trees and suffix arrays.
    - True
        - Converting the trie data structure idea for prefix search to be used for the subsequence search, the core approaches in those two data structures are used.
- True or false? Suffix array is merely a new representation of the suffix tree data structure.
    - True
- Select all that apply. Which of the following is true for suffix arrays?
    - A single long string is the input text
        - A single long string is put as input text of the suffix array and each word position in the text gives a suffix and binary search can be performed on the array.
    - Binary search can be performed on the array
        - A single long string is put as input text of the suffix array and each word position in the text gives a suffix and binary search can be performed on the array.
- True or false? Constant checking if the old searches match any of the query patterns is called triggering.
    - False
        - Constant checking if the most recent history matches any of the query patterns is called triggering.
- Select all that apply. Why should patterns or sequences of interest be brought quickly to the user?
    - It could be an indicator of a car engine failure
        - Patterns and sequences of interest should be brought very quickly to the user because they may indicate an emergency situation.
    - It might indicate a fire that need to be found quickly
        - Patterns and sequences of interest should be brought very quickly to the user because they may indicate an emergency situation.
- True or false? Using Aho-Corasick Trie, given any sequence of patterns the redundancies between the different patterns within a single pattern and across patterns may be leveraged to efficiently identify trigger conditions.
    - Using Aho-Corasick Trie, you can essentially create a combined finite automata where the redundancies between the different patterns within a single pattern and across patterns can be identified.
- Consider this sample of street addresses:
  1212 N CLARK ST

2360 W ADDISON ST
1239 W GRANVILLE AVE
2722 N CLARK ST
8902 N BROAD LANE
Which technique would be best for extracting all addresses that have a "W" after the house number?

- Subsequence
  - Subsequences indicate the string formed by removing some symbols from the original string. Subsequence search is useful for finding the subsequences in a larger regular expression.

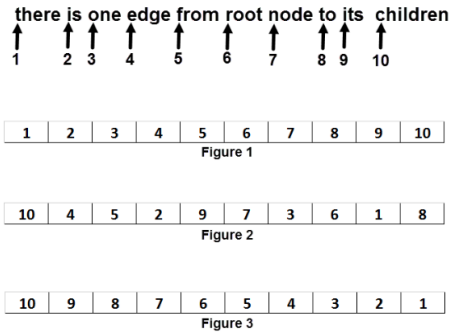
- Consider this sample of street addresses:
  1221 N CLARK ST
  2360 W ADDISON ST
  1239 W GRANVILLE AVE
  2712 N CLARK ST
  8902 N BROADWAY
  Which technique would be best for extracting addresses where the street number starts with 12?
  - Prefix
    - Prefix search is the best technique for identifying and extracting the patterns that match the beginning of a string.
- Let S be a set of s strings from alphabet Σ such that no string S is prefix of another string. If T is the trie for S, then how many leaves does T have?
  - S
    - The number of unique strings in the given S are s
- Which of the following is not a subsequence of 'GCFITQSPPN'?
  - IST is not a subsequence of 'GCFITQSPPN' because it is not contiguous.
- With a brute-force approach, what is the worst case cost of finding a 2-character substring in a string of 10 characters?
  - The minimum number of comparisons made to find a substring of length M in a string of length N is M*(N – M + 1).
- *Image: Suffix Trees*



Review *Image: Suffix Trees*. Also consider this text: "there is one edge from the root node to one of its children". Which figure in the image is the suffix tree for the provided text?

- Figure 3

        - Suffix tree is a trie representation of a string, with suffixes of given text as key and position in the text as value. Figure 3 accurately represents the given text.
- *Image: Suffix Arrays for Text*



  Review *Image: Suffix Arrays for Text*. Which figure is the suffix array for the text provided ("there is one edge from root node to its children")?
    - Suffix array is the lexicographically sorted array of suffixes of the given string. Figure 2 accurately represents the sorted suffixes from left to right.
- Recall that a "proper prefix" is defined as all characters in a string with one or more characters cut off the end of the string. For example, "T", "Th", "Thi" are proper prefixes of "This". Recall also that a "proper suffix" is all characters in a string with one or more characters cut off the beginning of the string.
  Now consider this string: CGTACGTTTCGTACG
  What is the longest proper prefix of the provided string that is also a proper suffix of the string?
    - CGTACG
        - CGTACG is the longest prefix which is also a suffix.
- Suppose you are using the KMP algorithm to search a pattern of length M in a string of length N. Which statements accurately identify characteristics of the KMP algorithm? Select all that apply.
    - The KMP algorithm never needs to move backward in the input text
        - By identifying how much of a last failed comparison of a string can be reused, KMP avoids backward iteration in the input text.
    - The running time of the KMP algorithm is linear
        - KMP has a linear complexity, O(M+N) for finding the occurrence of a pattern in string.
    - The KMP algorithm minimizes the total number of comparisons of the pattern against the input string
        - By preprocessing the pattern and storing the mismatches we reduce the number of comparisons done.
- *Diagram A: String and Pattern*

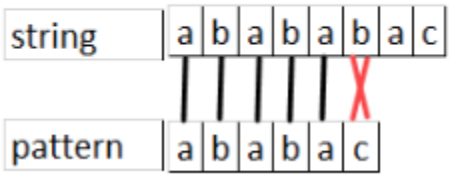
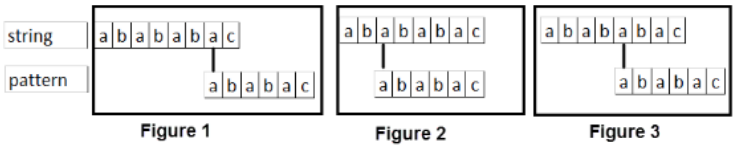


Diagram B: Next Iteration



Review *Diagram A: String and Pattern* and *Diagram B: Next Iteration*. Diagram A shows that the KMP algorithm is used to find the pattern "ababac" in the string "abababac". It is the first iteration step of the algorithm where it fails to match on the 6th character. Which figure in Diagram B correctly illustrates where the KMP resumes the search for the next iteration?

- Figure 2
  - The KMP algorithm skips comparison of the first two characters, as we have already computed the prefix and suffix match for them in an earlier iteration.

Sequence and Time Series: Edit Distance

Approximate String Match

- How can we quantify distance/similarity between pairs of strings?:
  - “table” vs. “cable”
  - “table” vs. “bale”
- Edit distance:
  - “table” vs. “cable”: 1 (replace “t” with “c”)
  - “table” vs. “bale”: 3 (delete “t”; replace “a” with “b”; replace “b” with “a”)
- Common edit operations:
  - Replacement:
    a ->b
  - Deletion:
    a ->λ
  - Insertion:
  - λ ->a

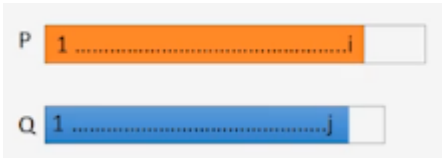
Edit Cost

- Let E be a sequence of edit operations to convert one string to another
- Let us associate a cost, C, to each edit operation
  - Cost of edit operations can be different from each other
    - Type of the operation (replace, delete, insert)
    - Symbols involved in the operation
    - Position of the edit operation
- Given a sequence of edit operations, E

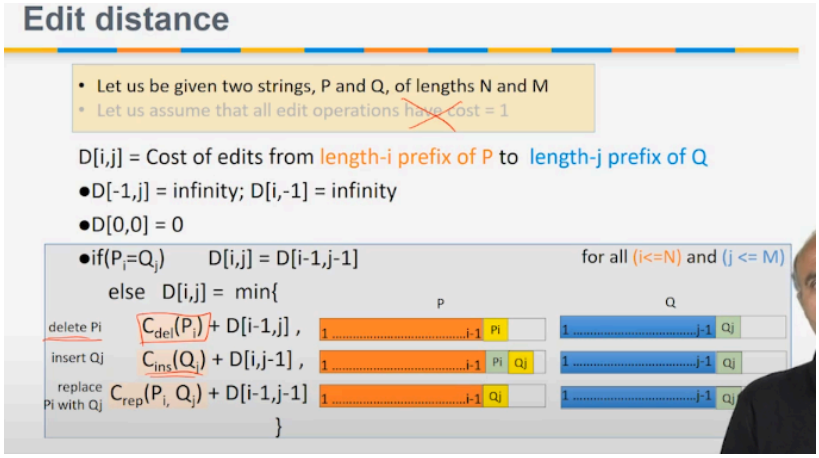
- $C(E) = \sum_{e_i \epsilon E} C(e_i)$

- Edit Distance:
  - $D(String_1, String_2) = min\{C(E)\}$

Edit Distance

- Let us be given two strings, P and Q, of lengths N and M
- Let us assume that all edit operations have cost = 1
- $D[i, j] = \#of\ edits\ from\ length = i\ prefix\ of\ P\ to\ length - j\ prefix\ of\ Q$
  - 

- $D[N, M] = edit\ distance\ from\ P\ to\ Q$



Summary

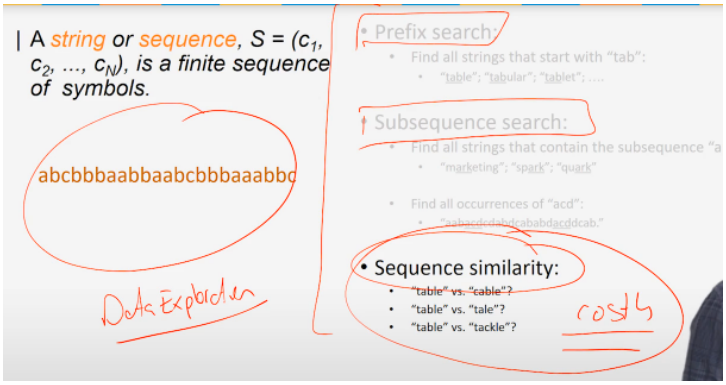
- Edit distance can be used to assess how similar or different two strings are
  - Recursive [Dynamic Programming]
- Problem: Edit distance can be costly for matching long strings
  - $O(N \times M)$

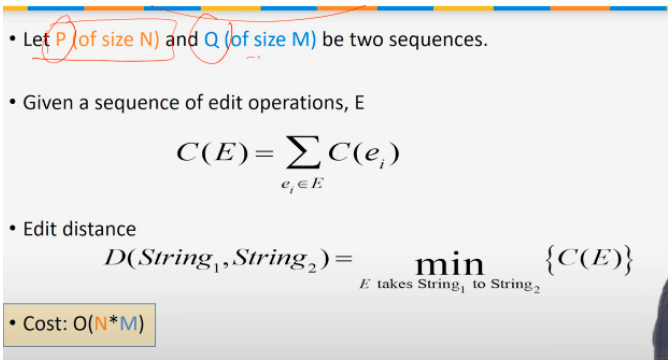
Knowledge Check

- What is the edit distance between "algorithm" and "rhythm"?
  - 6
    - Edit distance counts the number of edits we would need to perform to convert a query string to a data string. We would need to perform six edits to convert string "algorithm" to "rhythm".
- Let X and Y be strings of length n and, respectively. WHat is the cost of computing the edit distance between X and Y?
  - O(nm)

- For computing the edit distance between two strings, we need to iterate both of the strings. Therefore, the cost of computing the edit distance is O(nm).

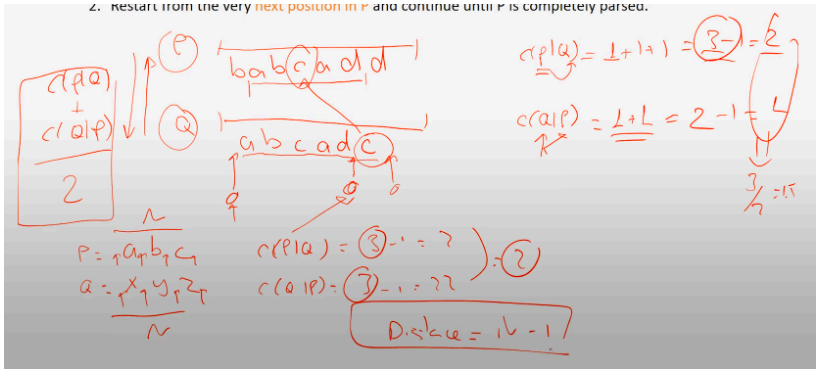
Data Review





Cross-Parsing Distance

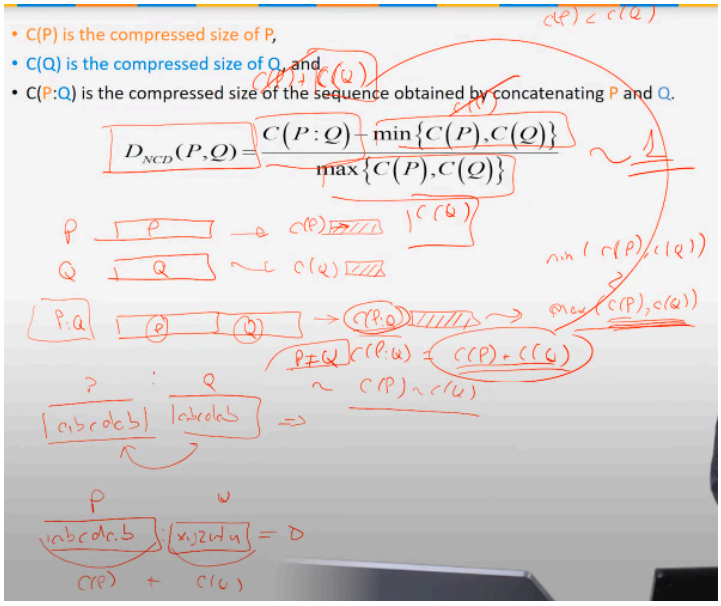
- Let P (of size N) and Q (of size M) be two sequences
- $c(P|Q)$: cost of parsing P with respect to Q
  - Find the longest (possible emptyy_ prefix of P that appears as a string somewhere in Q
  - Restart from the very next position in P and continue until P is completely parsed



- $D_{crossparse}(P, Q) = \frac{(C(P|Q)-1)+(c(Q|P)-1)}{2}$
- Linear time if the strings are indexed using a suffix tree

Compression Distance

- $C(P)$ is the compressed size of P
- $C(Q)$ is the compressed size of Q, and
- $C(P{:}\,Q)$ is the compressed size of the sequence obtained by concatenating P and Q
  - $D_{NCD}(P, Q) = \frac{C(P:Q)-min\{C(P), C(Q)\}}{max\{C(P), C(Q)\}}$

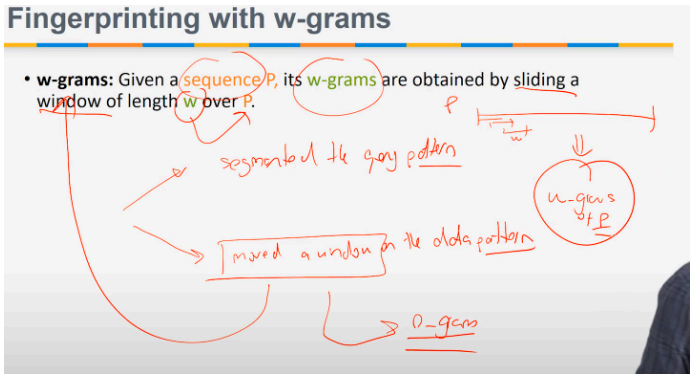
- 

Filtering

- Given a string P of (length N) and a pattern Q (of length M), determine whether the string P may contain an approximate match to Q with at most k errors
- Approach 1: Given a maximum error rate, k,
  - cut the pattern Q into $k + 1$ pieces
  - verify that at least one piece of Q exists in P exactly
  - This is because k errors cannot affect more than k pieces
- Approach 2: Given a maximum error rate, k,
  - slide a window of length M over the string P and count the number of symbols that are included in the pattern Q
  - only windows that have at least $M - k$ matching symbols need to be considered

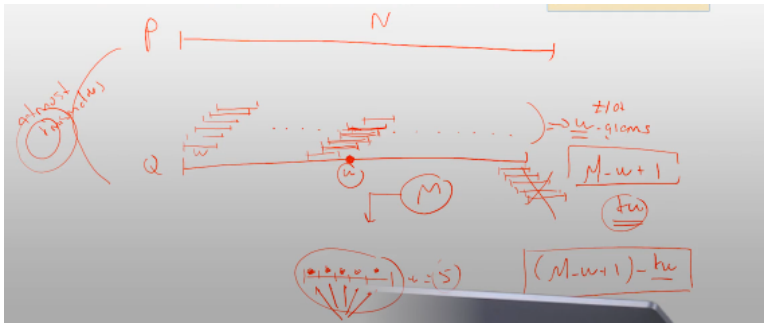
W-Grams

Fingerprinting with W-Grams

- w-grams: given a sequence P, its w-grams are obtained by sliding a window of length w over P
- 

Common W-Gram Counting

- Given a string P of (length N) and a string Q (of length M), determine whether the two strings may match each other with at most k errors
    - w-grams: given a sequence P, its w-grams are obtained by sliding a window of length w over P
- Approach (common w-gram counting):
    - Identify $(M - w + 1)$ w-grams of the query string Q
    - Each mismatch between Q and P can affect w many w-grams
        - given an upper bound of k errors, at least $(M - w + 1 - kw)$ w-grams must match
    - Search for these matches using a suffix tree (in linear time)



String Kernels and Min-Sampling Similarity Algorithms

String Kernels

- Given a string P of (length N) and a string Q (of length M), determine whether the strings may approximately match each other
    - w-grams: given a sequence P, its w-grams are obtained by sliding a window of length w over P
- Approach (string kernels):
    - Identify all w-grams of the query sequence Q; create a counting vector, q
    - Identify all w-grams of the data sequence P; create a counting vector, p
    - Measure the (dot product) similarity of the two counting vectors
        - If the dot product similarity is low, then P, is not likely to match Q

Min-Sampling Similarity

- Given a string P of (length N) and a string Q (of length M), determine whether the strings may approximately match each other
    - w-grams: given a sequence P, its w-grams are obtained by sliding a window of length w over P
- Approach (min-sampling similarity):
    - Consider r random hash orders of the w-grams of P and Q
    - For each order o_i
        - Find the smallest B w-grams of strings P based on the chosen order
        - Find the smallest B w-grams of string Q based on the chosen order
        - If they agree, $match,\ (P, Q) = 1$
    - After computing the match for all r orders

- $sim(P,\ Q) = \frac{\sum_{o_i} match_i(P, Q)}{r}$

Summary

- Edit distance can be costly for matching long strings
  - Cross-Parsing and COmpression-Distance can be used to approximate the edits distance comparison
- W-grams (more commonly known as n-grams) can be used to help filter unpromising candidates before more costly distance computations

Knowledge Check

- For strings X and Y of length N, which statements are true for cross-parsing distance between X and Y? Select all that apply:
  - Cross-parsing distance is computed by using the cost of parsing X with respect to Y and vice versa
    - The cost of computing cross-parsing distance between two strings is linear.
  - Cross-parsing distance is an approximation to edit distance
    - Cross-parsing distance is an approximation to edit distance that uses suffix tree data structure for computation.
- Let X and Y be strings of length n and m, respectively. The cost of parsing X with respect to Y is 6, and the cost of parsing Y with respect to X is 8. What is the cross-parsing distance between X and Y?
  - 6
    - Cross parsing distance is calculated as (C(X/Y)-1 + C(Y/X)-1)/2.
- Let $X = abcdef$. and $Y = zytrspm$. What is the compression distance between X and Y?
  - 1
    - Since the two strings are very dissimilar, the compression distance is closer to 1.
- How many subsegments will be created for each search of a pattern Y (with 10 characters) from a string X (with 100 characters) with a maximum error rate of 7?
  - 8
    - The given string is divided into (maximum error rate +1) segments for search.
- Which search windows are candidates for finding the pattern " ababc" in a string of 100 characters if a maximum error rate of 2 is allowed? Select all that apply.
  - abccd
    - In this string we have one a, one b, two c's, and one d. There are two mismatch symbols. Since the maximum error rate allowed is 2, this window is a valid candidate.
  - acccb
    - In this string we have one a, one b and three c's. There are two mismatch symbols. Since the maximum error rate allowed is 2, this window is a valid candidate.
  - abadb

        - In this string we have two a's, two b's, and one d. There is only one mismatch symbol. Since the maximum error rate allowed is 2, this window is a valid candidate.
- How many w-grams will contain a mismatch for each search of a pattern Y (with 10 characters) from a string X (with 100 characters) with a maximum error rate of 7?
    - 7
        - The number of mismatched w-grams is at most equal to the maximum error rate.
- *Table: Dot Product Similarities*

	V1	V2	V3	V4
V1	1	0.45	0.55	0.01
V2	0.45	1	0.6	0.39
V3	0.55	0.6	1	0.3
V4	0.01	0.39	0.3	1

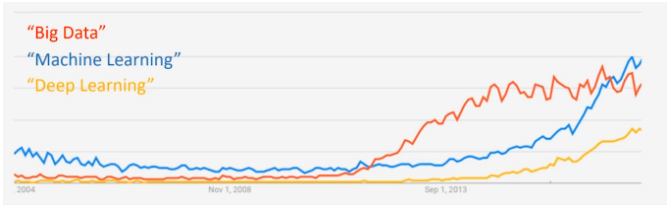
  Review *Table: Dot Product Similarities*. The table shows the dot product similarities between four vectors of sequences of strings. Which vectors contain the most similar strings? Select all that apply.
    - V3
        - The dot product value is used to determine if two strings are similar and not. The Vector V2 and V3 have higher dot product values compared to V1 and V4.
    - V2
        - The dot product value is used to determine if two strings are similar and not. The Vector V2 and V3 have higher dot product values compared to V1 and V4.
- For which min-sampling similarity value is it worth applying edit distance to find the pattern Q in a string P?
    - 1
        - A higher min-sampling similarity value implies that the strings match each other.

Introduction to Time Series

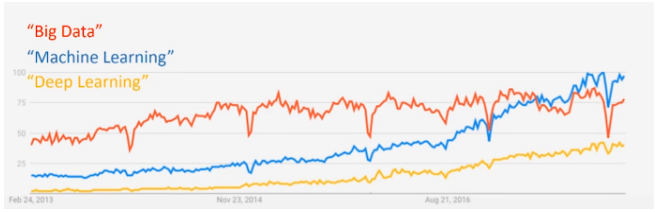
Time Series

- A time series, $T = (d_1, d_2, ..., d_N)$, is a finite sequence of data values

Comparing Time Series



Motifs



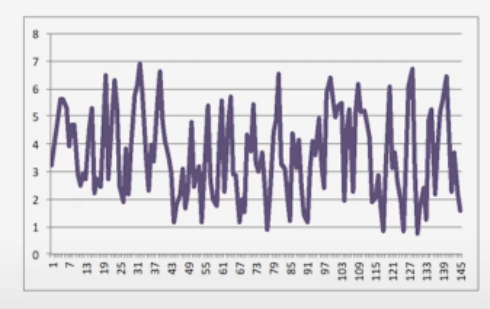
Classification

Time Series Models

- Question: Can we discover a closed form formula (or a model) that describes a given time series?
- Simpler QUestion: Can we characterize high-level properties of a given time series?

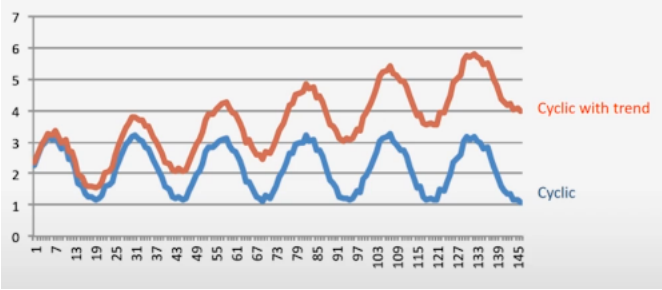
Stationary Series

- Stationary Series
    - statistical properties, such as mean and variance, are constant over time



Non-Stationary Series

- statistical properties change over time



Why do these matter?

- Stationary Series
    - statistical properties, such as mean and variance, are constant over time
- Non-Stationary Series
    - statistical properties change over time
- Why important?
    - most statistical forecasting methods assume that the series can be rendered (approximately) stationary through mathematical transformations

Time Series Modeling

Random Time Series Model

- Current observations only reflect current (random) error:
    - $X_t = E_t$
- Intuitively, the "random error" represents a stochastic process that does not depend on the past (e.g. an external input to the system)

Autoregressive Time Series Model

- Current observations only reflect current (random) error and has no dependence on the past observations
  - $X_t = E_t$
  - $E_i \sim N(0, \sigma^2)$
- AR(1): the current observation depends only the previous time instance and the current error
  - $X_t = \alpha X_{t-1} + E_t + \lambda$
- AR(2): the current observation depends only the previous 2 time instances and the current error
  - $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-1} + E_t + \lambda$

Moving Average Time Series Model

- MA(1): the current observation depends only to the error in the previous time instance and the current error
  - $X_t = \beta E_{t-1} + E_t + \lambda$
- MA(2): the current observation depends only to the error in the previous 2 time instances and the current error
  - $X_t = \beta_1 E_{t-1} + \beta_2 E_{t-2} + E_t + \lambda$

ARMA Model

- ARMA(a,m):
  - AR(a): the model has a auto-regressive terms
  - MA(m): the model has m moving-average terms
  - $X_t = (\alpha_1 X_{t-1} + \ldots + \alpha_\alpha X_{t-\alpha}) + (\beta_1 E_{t-1} + \ldots + \beta_m E_{t-m}) + E_t + \lambda$
- Shortcoming of ARMA Models:
  - These models cannot model where the current value is determined by taking into account the speed of change or degree of acceleration observed in the past

Models with Differencing

- "differencing" enables models to also consider speed of change and degree of acceleration in determining the current value:
  - Order-1 Differencing:
    - $X_t^{(1)} = X_t - X_{t-1}$ speed of change
  - Order-2 Differencing:
    - $X_t^{(2)} = X_t^{(1)} - X_{t-1}^{(1)}$ degree of acceleration
  - ...
  - Order-d Differencing:
    - $X_t^{(d)} = X_t^{(d-1)} - X_{t-1}^{(d-1)}$

ARIMA Model

- ARIMA(a,d,m)
  - AR(a): the model has a auto-regressive terms
  - MA(m): the model has m moving-average terms
  - Integrated with order-d differencing

$$X_t = (\alpha_1 X_{t-1} + \ldots + \alpha_\alpha X_{t-\alpha}) + (\beta_1 E_{t-1} + \ldots + \beta_m E_{t-m}) + E_t + \lambda + (\Theta_1 X_{t-1}^{(d)} + \ldots + \Theta_m X_{t-m}^{(d)}$$

Time Series Seasonal Differencing

Models with Seasonal Differencing

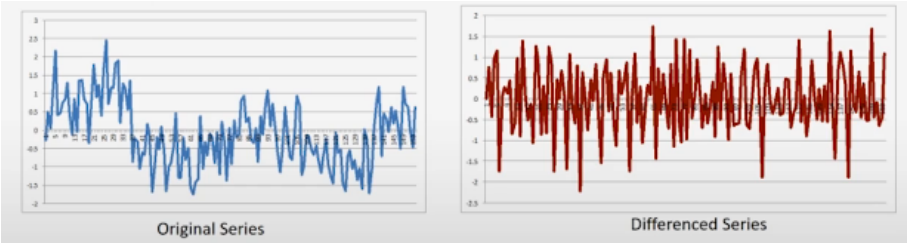
- Seasonal "differencing" enables models to also consider speed of changes of values across a gap (or a lag):
  - lag "S" seasonal differencing
  - $S_t^{(s)} = X_t - X_{t-s}$
- Seasonal ARIMA models also incorporate seasonal terms
  - additive seasonal models
    - $X_t = ARIMA(a,d,m) + SEASONAL_s(A,D,M)$
  - Multiplicative seasonal models
    - $X_t = ARIMA(a,d,m) \times SEASONAL_s(A,D,M)$
- A = # seasonal autoregressive terms, D = # seasonal differences, M = # seasonal moving average terms

Model based Time Series Analysis

- Find the parameters of the model
  - Model Fitting
    - the model should be as simple as possible (Contain as few terms as possible)
    - the fit to historic data should be as good as possible
  - Plot Analysis
    - Autocorrelation Function (ACF) helps observe linear relationships between lagged values of a time series
      - time series
      - $$\text{ACF}(X, lag) = \frac{E\left[\left(X_t - \mu\right)\left(X_{t+lag} - \mu\right)\right]}{\sigma^2} = \frac{\text{Covar}(X_t, X_{t+lag})}{\sigma^2}$$
    - Partial Autocorrelation Function (PACF) adjusts for the presence of intermediate values
      - $$\text{PACF}(X, lag) = \frac{\text{Covar}(X_t, X_{t+lag} \mid X_{t+1}, \ldots, X_{t+lag-1})}{\sigma^2}$$

Box-Jenkins Procedure

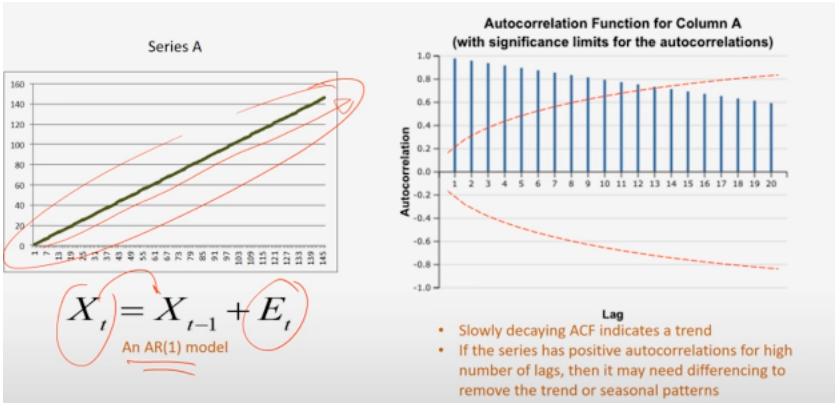
- Remove any seasonal patterns and deterministic trends that may hide valuable information and patterns through differencing

- When the mean trend is stochastic, differencing the series may yield a stationary stochastic process and, thus, may help convert a non-stationary series to a stationary one
- 

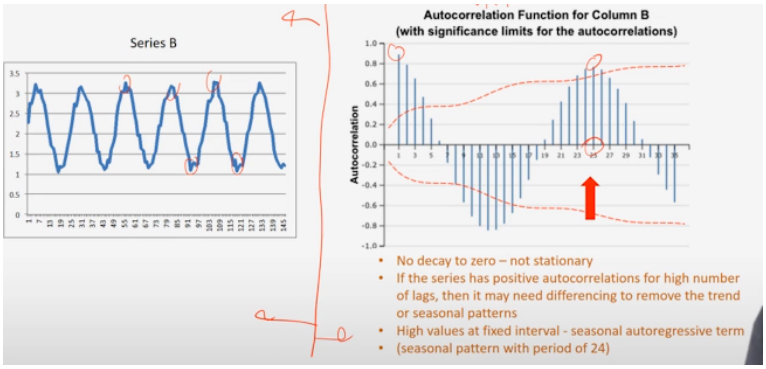
Key Properties of ACF and PACF

- For trend series autocorrelation function (ACF) slowly decays
- For an autoregressive, AR(a), series, the partial autocorrelation function (PACF) gives 0 at lag >= a + 1
- For a moving average, MA(m), series,
    - the partial autocorrelation function (PACF) does not "shut off" at a fixed lag, but moves toward 0
    - the autocorrelation function (ACF) has non-zero autocorrelation only at the lags of the model
- For a seasonal series autocorrelation function (ACF) at the seasonal lag will be large and positive

Autoregressive Model



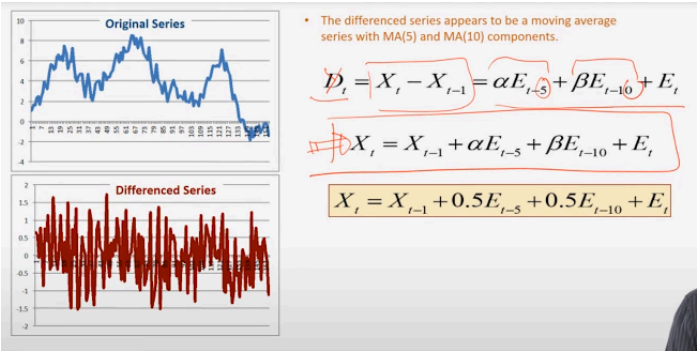
Non-Stationary (Cyclic) Model



Seasonal Series Using ACF and PACF 5

Complex Model

- The differenced series appease to be a moving average series with MA(5) and MA(10) components



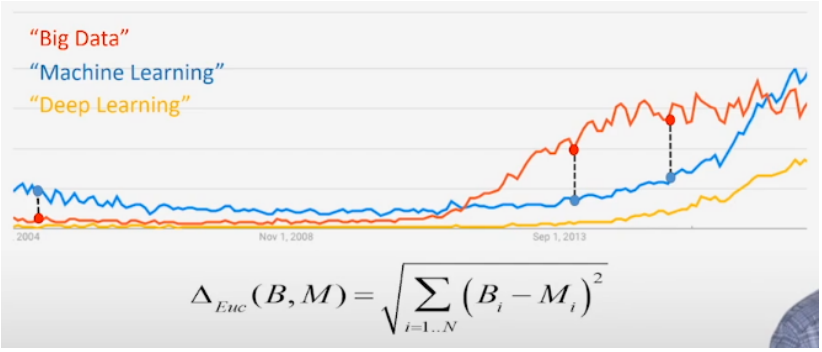
Knowledge Check

- Which are characteristics of the "Time" dimension? Select all that apply.
  - Time is continuous
    - The sequence of data in a time series is continuously measured and collected.
  - Time is cyclical
    - Time is non-stationary, and the mean of the time series changes over time.
  - Time is ordered
    - It is a finite sequence of data values with an order dependency between the data points.
- When comparing linear time data and cyclical time data, which aspect is different between the two?
  - The ordering of the data
    - In cyclical data, the current observation is only dependent on the previous 1-2 observations or events. In linear time data, there is a fixed frequency of change.
- What kind of visualization is a compressed, intense, wordlike graph that shows the trends of data?
  - Sparkline
    - Sparklines are tiny wordlike graphs used for visualizing trends in a series of data.
- Which of these options are time series data operations? Select all that apply.
  - Comparison
    - Time series data can be compared by identifying the common frequency of measurements. The ARIMA model can be used for comparing the observations.
  - Motif Search
    - Motifs are repeated patterns in a time series. Motif search is used for analyzing these repeated patterns to understand the occurrence of the similar pattern in future.
  - Predictive Analytics
    - By analyzing the sequence of data collected over an interval of time, predictive analytics can be performed to forecast future values.
- Which kind of time series data has constant statistical properties?
  - Stationary
    - Stationary time series data show similar patterns over time. The statistical properties like mean, variance, etc., remain constant over time.
- Which statements are true about the Moving Average Time Series Model MA(1)? Select all that apply.

    - The model has a closed-form formula
        - Closed-form model has a finite set of operations. Moving average model is represented by $X_t = \beta E_{t-1} + E_t + \lambda$
    - The model depends on the immediate random external event in the past
        - The moving average model depends on the previous external event and current external event.
    - The model depends on the current random external event
        - The moving average model depends on the previous external event and current external event.
- What happens when a time series is lagged for N observations?
    - The time series is shifted back by N observations
        - A lag in a time series is a cyclic behavior in which an event that is S instances past is able to describe the current event. In order to account for this lag, the time series needs to be shifted N observations back.
- What is the goal when adding seasonal differencing to the ARIMA?
    - To achieve the stationarity of the time series data
        - Seasonal differencing is performed by subtracting the event that is S units in the past from the current event in order to convert a non-stationary cyclic series into a stationary series.
- Suppose that we have a set of time series data, and the value of PACF is 0 for lags greater than 5. What is the value of the model parameter p in AR(p)?
    - 5
        - The partial correlation function takes the model parameter value as lag ≥ a+1. Therefore, for the above time series data, we consider a p-value as 5.
- Which statement is true about autocorrelation function (ACF) in trend series?
    - ACF slowly decays

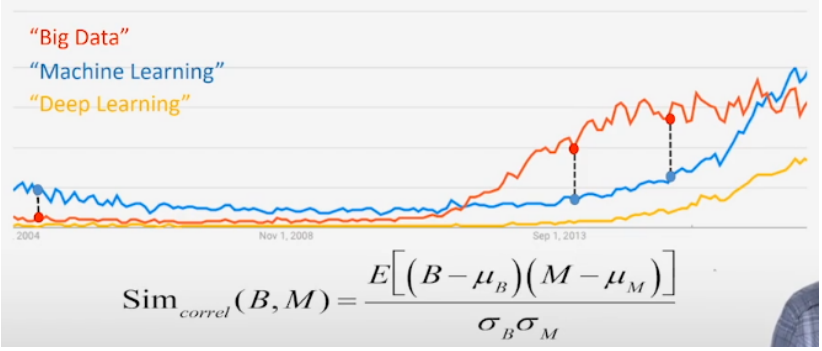
Time Series Matching

Euclidean Distance

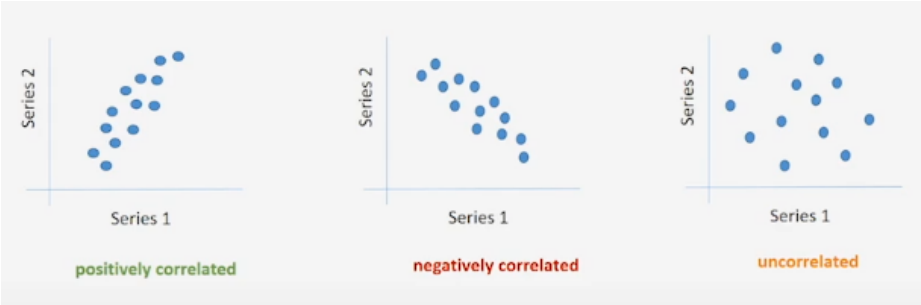


Euclidean Distance and Correlation Similarity

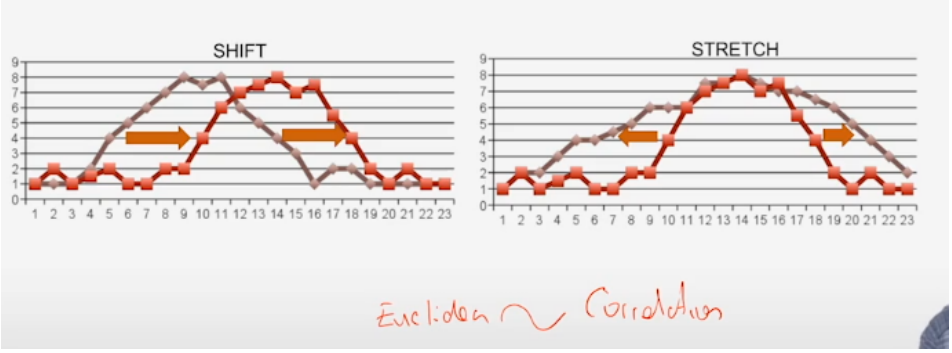
Correlation Similarity



Correlation



Issues with Synchronized Measures



Reminder: Edit Cost

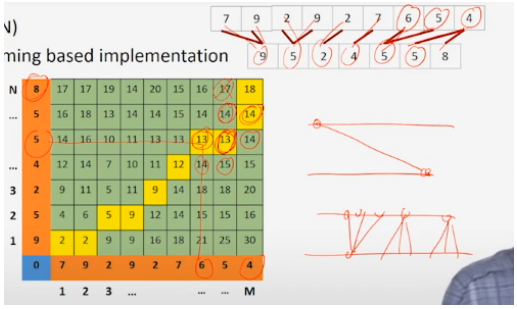
- Let E be a sequence of edit operations to convert one string to another
- Let us associate a cost, C, to each edit operation
    - Costs of edit operations can be different from each other
        - Type of the operation (replace, delete, inset)
        - Symbols involved in the operation
        - Position of the edit operation
- Given a sequence of edit operations, E,
    - $C(E) = \sum_{e_i \epsilon E} C(e_i)$

Edit Distance and Dynamic Time Warping

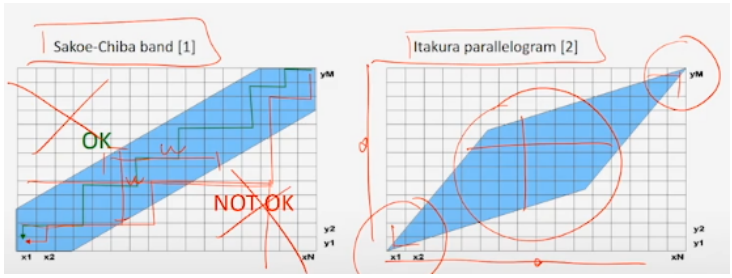
Dynamic Time Warping

- Let us be given two time series, P and Q, of lengths N and M
- $D[i,\ j]$ = # of edits from length-i prefix of$P$ to length-j prefix of Q
    - $D[0,\ j]$ = infinity; $D[i,\ 0]$ = infinity
    - $D[0,\ 0]\ =\ 0$
        - else $D[i,\ j]$ = abs(P_1, -Q_{jj}) + min{D[i-1, j], D{i, j-1], D[i-1, j-1]}

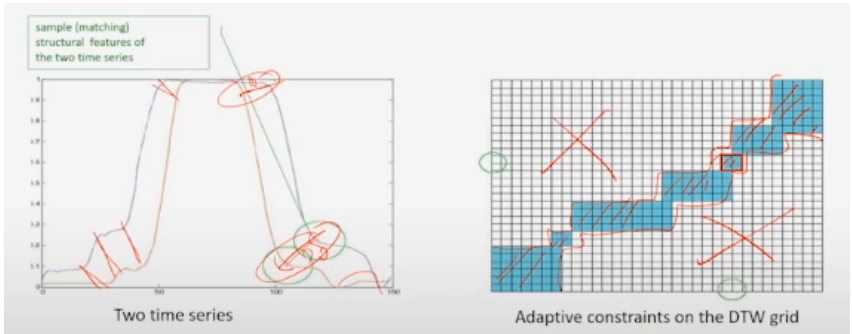
Dynamic Time Warping (DTW)

- Complexity: $O(M, N)$
- Dynamic Programming based implementation
- 

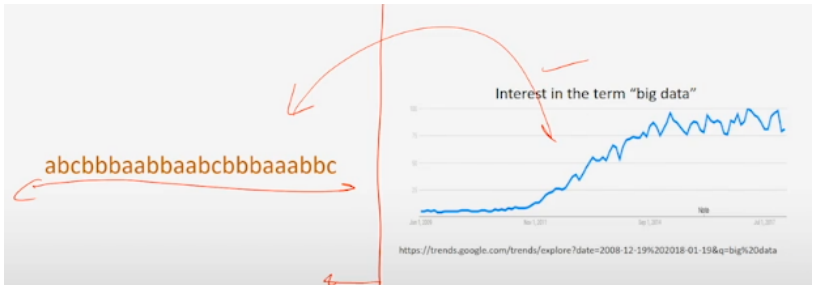
Reducing the Cost of DTW

- To reduce the $O(NM)$ cost of filling the grid versus heuristics import constraints on the grid regions through which the warp paths can pass
- 

sDTW

- Time series often carry temporal features that can be used for identifying locally relevant constraints to eliminate redundant work in an adaptive manner
- 

Time Series are Similar to Sequences

- A string or sequence, $S = (c_1, c_2, ..., c_n)$, is a finite sequence of symbols
- A time series, $T = (d_1, d_2, ..., d_n)$, is a finite sequence of data values
- 

SAX (Symbolic Aggregate Approximation)

- Time series are similar to sequences
- Can we transform a time series into a compact sequence representation?
- Transform a time series into a compact sequence representation:
  - Divide the time series into w-length (non-overlapping) windows
  - For each window
    - compute the average amplitude
    - 

- Note that, SAX reduces
  - temporal resolution, by dividing the string into windows (length ~ N/w)
  - amplitude resolution, by using only one of the s symbols per window

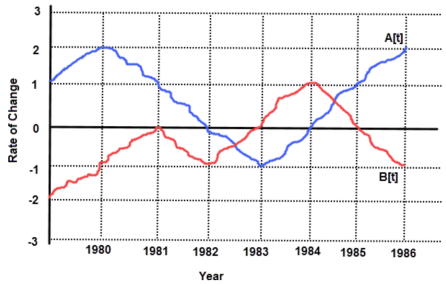
Time Series Summary

Summary

- Synchronized measures (Euclidean distance, correlation) are relatively cheap to compute, but may not account for shifts, scratches, and other temporal misalignments
- Asynchronous approaches (edit distance, DTW) can account for misalignments, but are expensive. Solutions include
  - constraining the search space
  - creation of reduced representations of the time series

Knowledge Check

- *Diagram: Time Series*



Review *Diagram: Time Series*. There are two time series in the diagram: A[t] and B[t]. What is the Euclidean difference between A[t] and B[t]?

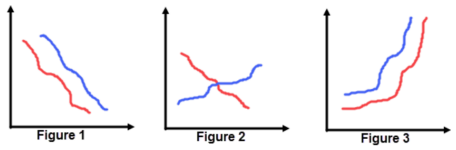
- The mathematical value $\sqrt{32}$
  - The square root of summation of distance between A[t] and B[t] at each point is calculated to get Euclidean difference.

- *Diagram: Data Set Correlation*

Review *Diagram: Data Set Correlation*. The diagram shows three different data sets. Which data set has the weakest correlation?

- Figure 3
  - Unlike Figure 1 and Figure 2, which are negatively and positively correlated, respectively, Figure 3 has no strong correlation between the data points.

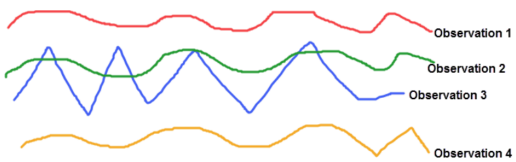
- *Diagram: Time Series Correlation*



Review *Diagram: Time Series Correlation*. Which set of time series shows a negative correlation?

- Figure 2
  - Negative correlation is a relationship between two variables in which one variable increases while the other decreases.

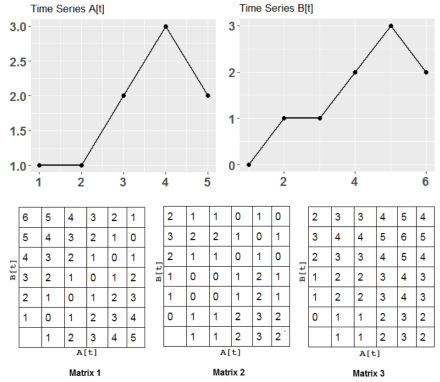
- *Figure: Observations*



Review *Figure: Observations.* This figure shows four different observations, each denoted by a different line. Which two observations have a large correlation-based distance between them?

- Observation 1 and Observation 3

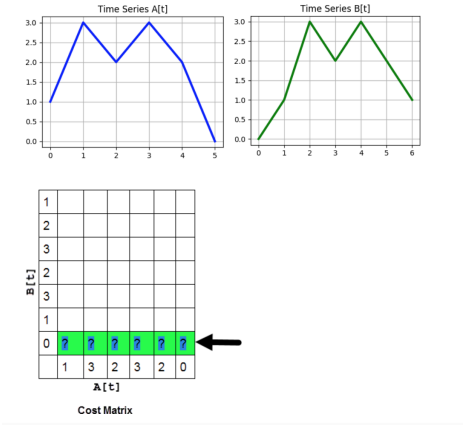
        - The distance between observations is zero when they are highly correlated. From the figure we can discern that Observation 1 and 3 are not highly correlated.
- When comparing time series data, what is a problem with using Euclidean or correlation-based distances?
    - The assumption of time synchronicity between time series data
        - When two synchronized time series are compared, even though the shape of the time series are the same, they might be shifted in position. This will make them uncorrelated and will have a large Euclidean distance.
- *Image: Time Series and Matrices*



Review *Image: Time Series and Matrices*. There are two time series, A[t] and B[t], and the distance between them is defined by the absolute value of the difference between each pair: abs(A[i] - B[j]). Which matrix represents the distance matrix between the two time series?

- Maxtrix 2
    - Matrix 2 accurately represents the distance matrix between two time series by following the dynamic time warping approach.

- *Images Time Series and Cost Matrix*



Review *Image: Time Series and Cost Matrix*. There are two time series, A[t] and B[t], and the distance between them is defined by the absolute value of the difference between each pair: abs(A[i] - B[j]). What are the values in the first row of the given cost matrix?

- 1, 4, 6, 9, 11, 11
    - The values are calculated as per the dynamic time warping formula D[i,j]=abs(A[i]-B[j]) + Min(D[i-1,j], D[i-1,j-1], D[i,j-1]).

- *Figure A: Distance Matrix*

1	0	2	0	1	0
2	1	1	1	0	1
3	2	0	2	1	2
1	0	2	0	1	0
2	1	1	1	0	1
2	1	1	1	0	1
1	0	2	0	1	0

Figure B: Different Cost Matrices

4	6	6	4	5	12
5	6	4	4	4	11
5	5	3	5	4	9
3	3	5	3	4	6
4	3	3	3	3	5
4	2	2	3	3	3
4	1	3	3	4	1

Cost Matrix 1

12	6	6	4	5	4
11	6	4	4	4	5
9	5	3	5	4	5
6	3	5	3	4	3
5	3	3	3	3	4
3	2	2	3	3	4
1	1	3	3	4	4

Cost Matrix 2

12	6	6	4	1	4
11	6	4	1	4	5
9	1	3	5	4	5
6	3	1	3	4	3
5	3	3	3	3	4
3	2	1	1	3	4
1	1	3	3	4	4

Cost Matrix 3

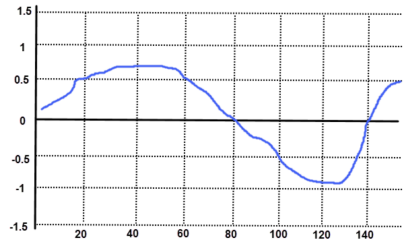
Review *Figure: Distance Matrix* and *Figure: Different Cost Matrices.* Figure A gives the distance matrix between two time series, A[t] and B[t]. The distance between the series is defined by the absolute value of the difference between each pair: abs(A[i] - B[j]). Figure B provides three different cost matrices. Which cost matrix corresponds to the distance matrix?

  - Cost Matrix 2
    - The values are calculated as per the dynamic programming approach D[i,j]=abs(A[i]-B[j]) + Min(D[i-1,j], D[i-1,j-1], D[i,j-1]).

- *Table: Symbol Map*

Range	Symbol
[0, 0.5)	A
[0.5, 1)	B
[1, 1.5)	C
(0, -0.5)	D
[-0.5, -1)	E
[-1, -1.5)	F

Figure: Time Series



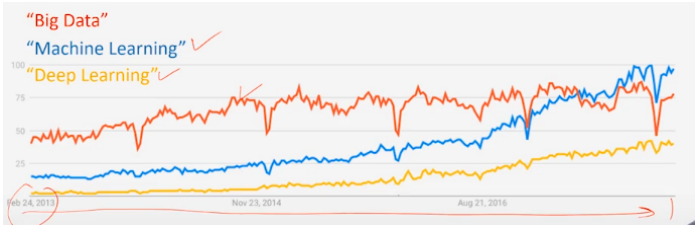
Review *Table: Symbol Map* and *Figure: Time Series*. The table gives a symbol map for a range of values. What is the symbolic representation of the time series provided in the figure?

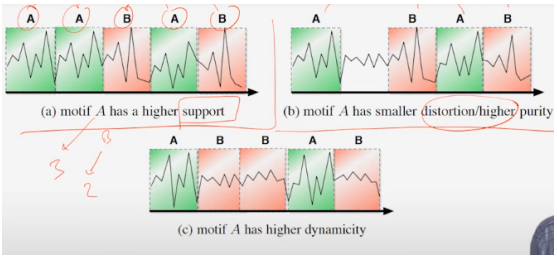
- ABBADEEA
  - Symbols are used for more compact sequence representation. According to the figure, each datapoint is compared with the window size and assigned a symbol from symbol map.

- Which statements are correct for creating a distance matrix and calculating distances between time series? Select all that apply.
  - Edit Distance and Dynamic Time Warping are expensive operations
    - Edit Distance and Dynamic Time Warping require quadratic runtime complexity.
  - Compared to Dynamic Time Warping, Euclidean and correlation-based distances are relatively cheap to compute
    - Euclidean and correlation-based distances are synchronized and assume two time series are aligned, making them easy for computation.
  - Edit Distance and Dynamic Time Warping account for misalignments
    - Edit Distance and Dynamic Time Warping are asynchronous and do not consider the assumption that time series are aligned.

Time Series Motifs

Motifs



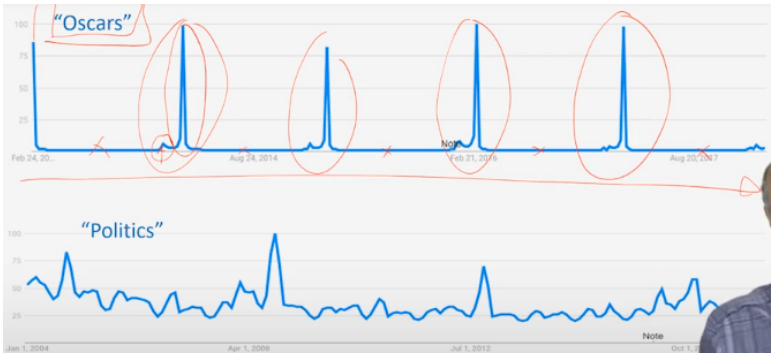


Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with a varying length)
    - all subsequences
    - subsequences that are sufficiently different from their immediate neighborhood
    - subsequences that have high dynamicity
    - subsequences that have at least one another similar subsequence in the series
  - Apply a clustering algorithm to identify groups of similar subsequences
    - allow for length differences
  - Eliminate clusters with
    - too few subsequences (not enough support)
    - too imperfect matches (not well defined motifs)

        - too low dynamicity (patterns that do not carry much useful information or are not interesting)

Motif Quality



Other Questions

- What is the motif length?
- Do all instances of a motif need to have the same length?

Summary

- Motif Search
    - divide series into sub-sequences
    - eliminate un-interesting sub-sequences
    - apply clustering algorithm
    - eliminate un-interesting clusters
- Challenge:
    - There are many sub-sequences
    - How to define what is an interesting sub-sequence?
    - How to define what is an interesting cluster?

Knowledge Check

- What are repeated patterns in time series data called?
    - Motifs
        - Motifs are the repeated patterns in a time series. These are useful in understanding the occurrence of a repeating pattern over time.
- Which metric is a quality measure for quantifying motifs? Select all that apply.
    - Purity
        - The motifs that are similar in structure have lesser distortion and higher purity.
    - Support
        - The motif that is repeated more often is considered to have higher support.
    - Distortion
        - The motifs that are similar will have lesser distortion and are preferred over others.

- What gets eliminated in the pruning step of motif search algorithm? Select all that apply.
    - Motifs with no dynamicity
        - Motifs with low dynamicity are removed as they point to regions in the graph that have no substantial information.
    - Motifs with low support
        - Motifs with low support have only few subsequences and are not very similar to each other.
    - Imperfect Matches
        - Imperfect matches are not similar to each other and have low support.
- Which conditions do motif search algorithms account for? Select all that apply.
    - Differences in motifs lengths
        - Motif search accounts for varying lengths by creating more subsequences and applying pruning methods early in the process.
    - Differences in motif amplitudes
        - Motif search accounts for amplitudes by considering the clustering algorithm where the amplitude differences are not strong.
    - Subsequences of varying lengths
        - Motif search is useful to solve subsequences of varying length by performing a pruning step before clustering.

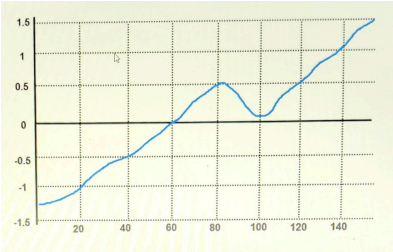
Module 4 Quiz Questions

- What are the characteristics of Moving Average Time Series Model, MA(2)? Select all that apply. (Hint: Angel external event brings in external input or random error to the outcome.)
    - The model has a closed form formula
    - The model depends on the previous 2 time instances of external events in the past
    - The model depends on the immediate random external event in the past
    - The model depends on the current random external event
        - Selected all
- Which models can be used to smooth an analyze time series? Select all that apply:
    - Suffix Tree and Suffix Array
    - Autoregressive model
    - Autoregressive integrated moving average model (ARIMA)
    - Tried Data Structure
    - Selected 2 and 3
- Which graph (or set of graphs) can be used to detect seasonality in time series data?
    - Multiple box and autocorrelation
    - Autocorrelation only
    - Multiple box only
    - Line graph only
    - Selected 1

- In terms of the dynamic programming implementation of the DTW algorithm presented in this course, which moves are the kinds that the dynamic time warping algorithm assumes when moving through the cells of the matrix in finding the best path? Select all that apply:
    - Up-and-Right
    - Right
    - Down
    - Up
    - Selected 2, 3, 4
- What type of analysis would be most effective for predicting temperature?
    - Classification
    - Time Series Analysis
    - Clustering
    - Decision Tree Analysis
    - Selected 2
- *Table: Symbol Map*

  Range Symbol
  [0, 0.5) A
  [0.5, 1) B
  [1, 1.5) C
  [0, -0.5) D
  [-.5, -1) E
  [-1, -1.5) F

  Figure: Time Series

  Review the Table: Symbol Map, and Figure: Time Series. The table is a Symbol Map of a range of values. In terms of Symbolic Aggregate Approximation (SAX) algorithm, what is the symbolic representation of the provided Time series?
    - FEDAAABC
- In forecasting with time series analysis, suppose that the demand for a component is 100 in October 2016, 200 in November 2016, 300 in December 2016, and 400 in January 2017. What is the first 4-month simple moving average?
    - 250
- *Figure 1: Time Series Distance Matrix*

2	0	0	2
0	2	2	0
1	1	1	1
1	1	1	1
0	2	2	0

Figure 2: Cost Matrices

4	2	2	4
2	4	4	2
2	2	2	3
1	1	2	3
0	2	4	4

Cost Matrix 1

4	2	2	4
2	1	2	2
2	2	2	3
2	4	4	2
0	2	4	4

Cost Matrix 2

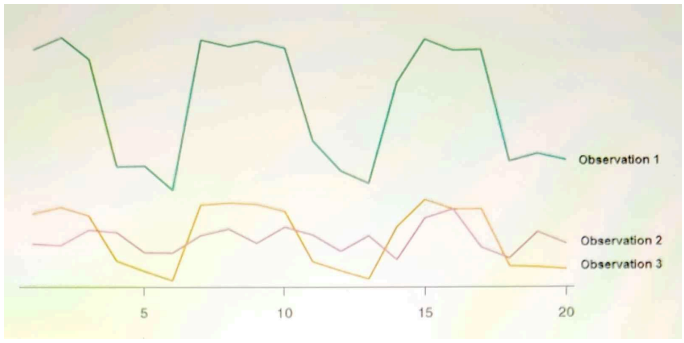
0	2	4	4
2	4	4	2
2	2	2	3
1	1	2	3
4	2	4	4

Cost Matrix 3

Review *Figure 1: Time Series Distance Matrix*, and *Figure 2: Cost Matrices*. For two time series , the distance between them is defined by the absolute value of the difference between each pair: . Given the distance matrix between and in Figure 1, which matrix in Figure 2 is the cost matrix?

- Cost Matrix 1
  - The values should be calculated using the dynamic time warping formula:
    $D[i,\ j] = abs(A[i] - B[j]) + Min(D[i-1, j],\ D[i-1,\ j-1,\ j-1],\ D[i,\ j-1])$

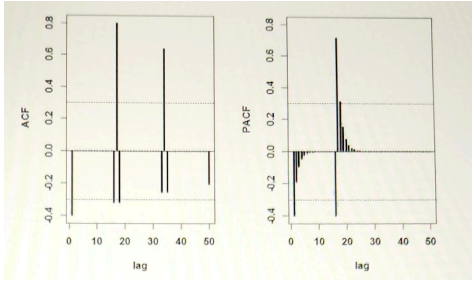
- *Figure: Observations*



Review *Figure: Observations*. Which observations have a small correlation-based distance between them?

- Observation 1 and Observation 3

- *Figure: ACF and PACF*



Review *Figure: ACF and PACF*. The figure provides the ACF and PACF of a time series. Which model would best represent the time series?

- MA

Introduction to Geographical Analysis and Visualization

Geovisualization

- Primarily denotes tools and techniques that are designed to support analyses that focus on datasets with a geographic component
- Visual representations are designed and built utilizing cartographic principles
- Look for trends over geographic regions

Geographic Visualization

- Utilizes sophisticated, interactive maps to explore information
- Recently, focuses on incorporating temporal components into analysis → Moving towards a combination of interactive maps and statistical analysis methods

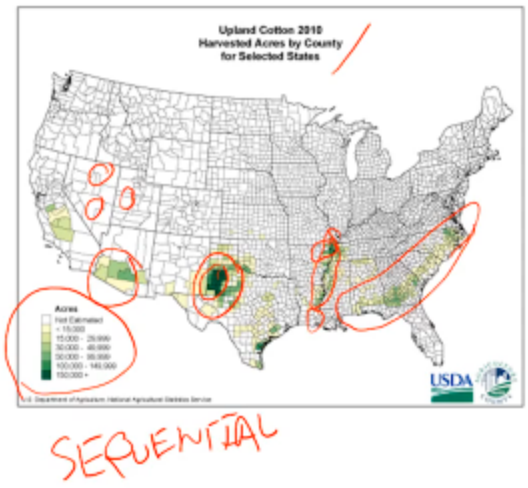
Tobler's First Law of Geography

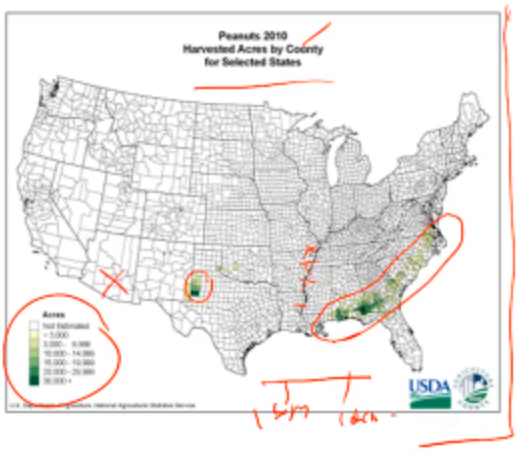
- "Everything is related to everything else, but near things are more related than distant things."

Thematic Maps

- A thematic map (or statistical map) is used to display the spatial pattern of a theme or attribute
- Focus on spatial pattern as opposed to a general-reference map which focuses on location
- Goal of a thematic map is to emphasize spatial patterns of geographic attributes (e.g., population density)

Examples





How are Thematic Maps Used?

- Locations: to provide specific information about particular locations

- Patterns: to provide general information about spatial patterns
- Comparisons: to compare patterns on multiple maps

Basic Steps for Communicating Map Information

1. Consider what the real-world distribution of the phenomenon might look like
2. Determine the purpose of the map and the intended audience
3. Collect data appropriate for the map's purpose
4. Design and construct the map
5. Determine whether users find the map useful/informative

5 Types of Map Phenomena

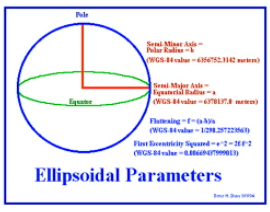
- Spatial Dimension
    - Point Phenomena
    - Linear Phenomena
    - Aerial Phenomena
    - 2.5D Phenomena
    - 3D Phenomena

Types of Map Coordinate Systems

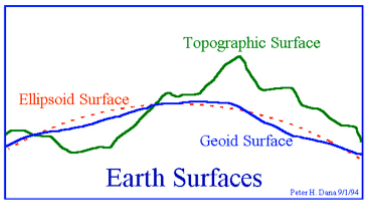
- Latitude/Longitude
- Universal Transverse Mercator
- State Plane
- Metes and Bounds

Geodetic Datums

- Datus define the size and shape of the earth
- The initial point of the coordinate system is determined by the projection, ellipse model and datum



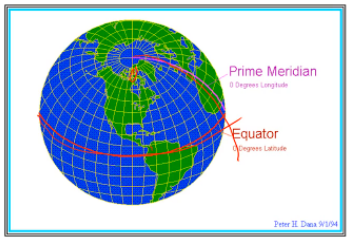
Common datums in the US:
North American Datum 1927
North American Datum 1983
World Geodetic System



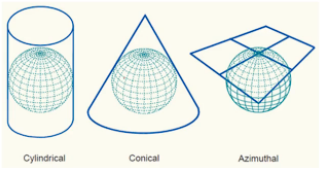
Topographical Surface Sea Level Gravity Models

Coordinate Systems: Latitude and Longitude

- Most commonly used coordinate system
- No transformations necessary between areas
- Scale, shape and direction distortion all increase with increase area of interest

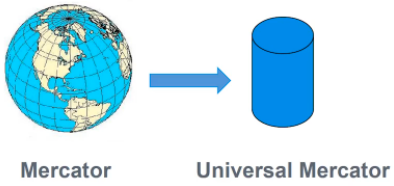


Map Projections



- Area/Shape Distortion
  - Equivalent: area is similar on globe and flat map, but shape is not
  - Conformal: shape is similar on globe and flat map, but the area is not
- Shape of Projections
  - Cylindrical: projection of sphere onto a cylinder
  - Conic: projection of sphere onto a cone
  - Azimuthal: projection of a sphere onto a plane
  - Miscellaneous

Cylindrical Projections

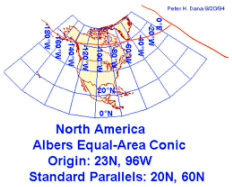


- Characteristics:
  - Straight meridians and parallels
  - Meridians equally space
  - Parallels unequally spaces

Conic Projections

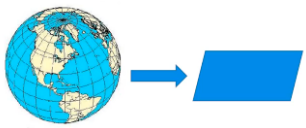
- Characteristics:
  - Straight meridians, curved parallels
  - Meridians radiate from poles
  - Parallels may be equally spaces
- Common conic projections:
  - Albers
  - Lambert
  - Prolyconic

Conic Projections - Albers Equal Area



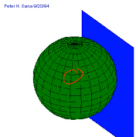
- Direction, area and shape are distorted away from the standard parallels
- Area and directions are true only in limited portions of a map

Azimuthal Projections



- Characteristics:
  - Straight or curved meridians, curved parallels
  - Meridians radiate from poles
  - Parallels may be equally spaced

Azimuthal Projections - Orthographic



- Simplest form is orthogonal projection

- Adequate only for very small areas
- Scale and area distortion increases as distance from the tangent center increases

Map Elements and Typography

- Map Elements
    -

Frame line and neat line	Mapped area	Inset	Title and subtitle
Legend	Date source	Scale	Orientation

Legend

- Map element that defines all of the thematic symbols of the map
- Symbols
    - if self explanatory or not directly related to the map's theme can be omitted
    - should be vertically centered with their definition
- Textual definitions should be horizontally centered
- Legend heading can be used to further explain the maps theme
- Scale is added to indicate the distances

Topography

- Use
    - use bold and italic type sparingly
    - a realistic lower limit for all type size
    - type size should correspond with the size or importance of map features
- Do not use
    - decorative type families
    - script, cursive and ornate style
    - two type families on a given map
- Limited Use
    - reserve italics for water features or to identify the data source

Orientation

- Orient type horizontally
- Avoid overprinting
- Ensure type is placed so that it is clearly associated with the feature it is representing

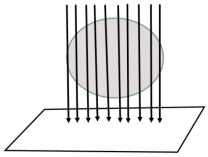
Knowledge Check

- In addition to exploration of geographic data, which tools and techniques are utilized by geographic visualization? Select all that apply.
    - Interactive maps
    - Statistical analysis
    - Temporal analysis
- *Image: Geographic Map*

Review Image: Geographic Map. In this map, three separate regions have been identified. Which region best exemplifies Tobler's First Law of Geography?

    - Region 2
        - As distance increases in this region, similarity decreases.
- What makes thematic maps more useful compared to general-reference maps? Select all that apply.
    - Thematic maps show statistics about places
        - Thematic maps use location-specific information to show spatial patterns and statistics about a place.
    - Thematic maps emphasize spatial patterns of geographic attributes
        - Unlike general reference maps, thematic maps emphasize spatial patterns of geographic attributes to provide specific information about a location.
- Which components are important in thematic map design? Select all that apply.
    - Source of data
        - The geographical source of data is useful for representing different features in the map.
    - Title
        - A title describes the subject of the map and gives the user an idea about what it depicts.
    - Legend
        - A legend gives visual explanation of colors or symbols used in the map.
- Which category of visualization does a terrain map belong to?
    - 2.5D
        - Terrain maps are used to visualize the height component in the maps.
- What is the most commonly used map coordinate system?
    - Latitude and Longitude
        - Latitude and longitude provide scale, shape, and direction distortions, and it does not need any transformations.
- Which type of map projection is the orthogonal projection?
    - Azimuthal
        - Azimuthal projection has proportional distance and direction from all the points to the center.
- Which statement defines a key feature of map projections?

    - The shape and area ratios are not preserved when a globe is mapped to a flat map
        - In equal area maps, the shapes of most features are distorted. No map projection can preserve all the spatial properties without compromise or distortion.
- *Diagram: Map Projection*



Review *Diagram: Map Projection*. The diagram depicts a projection that maps a globe to a planar surface. What is the name of this kind of projection?

    - Azimuthal
        - Azimuthal projections essentially try to project onto a flat plane in order to get curved meridians and curved parallels.
- Which practices are used in the design and construction of legends on a map? Select all that apply.
    - Align symbols vertically
        - The symbols should be self-explanatory and need to be vertically centered for better readability.
    - Use scales to indicate distances
        - Scales are used to visual perception and understanding of distances used in the map.
    - Align textual definitions of symbols horizontally
        - Textual definitions need to be aligned horizontally for better readability.

Introduction to Choropleth Maps and Color Schemes

Introduction to Choropleth Maps

- Earliest known choropleth map was created in 1826 by Baron Pierre Charles
- Term choropleth map was coined in 1938 by John Ketland Wright
- Choropleth maps are based on statistical data aggregated over defined geographical regions

Choropleth Maps

- Areas of the map are shaded in proportion to a measured variable
- Coloring is based on a classification (histogram binning) of the distribution of the measured variable

Coloring Choropleth Maps

- Colors
    - Relates to number of classes
    - Color Schemes: sequential, divergent, qualitative
    - Cartographic rule = 5-7 classes
    - Choose carefully to allow viewers to see trends

Color Schemes: Sequential

- Suited for ordered data

- Lightness steps dominate the look of the scheme
- Light values are low data values, dark are high
- Good for Original, interval and ratio data types

Color Schemes: Diverging

- Puts and emphasis on critical midrange values
- Color change represents deviation from a meaningful midrange critical value
- Good for ratio data types where looking at data above and below a 'zero' point

Color Schemes: Qualitative

- Does not imply magnitude difference
- Used to show differences between classes
- Good for nominal data types

Class Interval Selection

- Choices for optimizing the class interval selection are highly dependent on the underlying data distribution
- Similar to the concept of histogram binning
- Popular choices for class interval selection include
    - Equal interval selection
    - Jenks' Natural Breaks
    - Minimum boundary error

Equal Interval

- Classified data such that each case occupies an equal interval along the number line
- Advantage: easy to compute
- Disadvantage: fails to consider how data are distributed
- $\frac{range}{NumClasses} = \frac{High-Low}{NumClasses}$

Quantiles

- The number of color bins (classes) will determine the number of quantiles
- Advantages:
    - Easy to compute
    - Percentage of observations in each class will be the same
    - Class assignment is based on rank order
- Disadvantages:
    - Fails to consider data distribution
    - Dissimilar data can be placed into same class
- $Number\ in\ Class = \frac{TotalSamples}{NumClasses}$

Mean-Standard Deviation

- Classes are formed by adding or subtracting some number of standard deviations from the mean
- Advantages:
    - If data normally (or near normally) distributed, the mean serves as a useful dividing point
    - Legend will contain no gaps
- Disadvantage:
    - Only works well only with data that are normally distributed

Maximum Breaks

- Goal is to consider individual data values and group those that are similar
    - Order data for low to high and differences between adjacent values are computed
    - The largest differences ("breaks") are used for the class divisions
- Advantages: easy to compute
- Disadvantage: may miss natural clusters

Natural Breaks

- Data values are examined visually to determine logical breaks within the data
- Goal is to minimize differences between data values in the same class and maximize differences between different classes
- Advantage: tries to take into account natural underlying structure of the data
- Disadvantage: subjective, different mapmakers may choose other values

Optimal Classification

- Same as natural breaks, but minimizes an objective function
- Goal is to place values into groups
- Measure an "error" - common is to use the sum of absolute deviations about class medians
- Calculate each class median and then add the resulting sums of absolute deviations
- Many computer based algorithms have been developed to do this:
    - Jenks-Caspall
    - Fisher-Jenks

Optimal Classification

- Advantages:
    - Good empirical method for grouping data
    - Can assist in determining appropriate number of classes
- Disadvantages:
    - Hard to explain in novice users
    - May leave gaps in the map legend

Utilizing Spatial Context

Optimal Map (with spatial constraint)

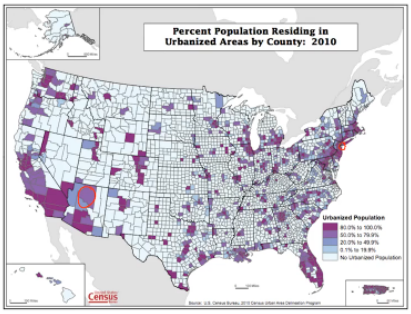
Raw Data
1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20

1	9	5	6
20	3	16	12
19	4	13	11
18	10	17	2

GADF = 0.70
C_F = 6/16 = 0.38

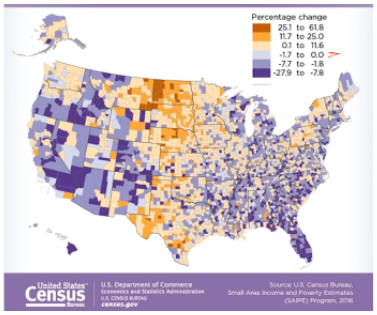
Choropleth Maps

- Need to standardize the statistics
    - Divide by population
    - Divide by area
- Modifiable areal unit problem - source of statistical bias occurring when data is aggregated into districts



Ecological Fallacy

- Inferences about individuals area based solely upon aggregate statistics collected for the group to which those individuals belong
    - Assumes that individual members of a group have the average characteristics of the group at large
    - Group characteristics do not necessarily apply to individuals within that group



Knowledge Check

- What is the main visual variable in choropleth maps?

    - Color
        - Color is the main visual variable in choropleth maps as it can be modified as per the user's needs without changing the meaning of the visualization.
- In general, what is the preferred number of colors used in the design of a choropleth map?
    - $4 < X < 8$
        - A typical choropleth map uses 5 to 7 colors to effectively represent the shades and territorial boundaries in a map.
- What color schemes can be used in choropleth maps? Select all that apply.
    - Sequential
        - Sequential color scheme is used if the data is numeric and ordered.
    - Qualitative
        - Qualitative color scheme is used for representing categorical data.
    - Divergent
        - Divergent color scheme is used when a choropleth map has meaningful center value with components on both positive and negative sides.
- *Table: Fruits*

Fruits
Limes
Grapes
Grapefruit
Banana
Apricot

  Review *Table: Fruits*. Suppose that we want to create a choropleth map for the fruits listed in this table. While constructing the map, we want to use equal interval classification based on the number of letters in the fruit name. What will be the size of the intervals if the fruits are classified into three groups?
    - 2
        - The number of equal intervals can be computed by finding the difference between the maximum number of letters in a fruit name and the minimum number of letters in a fruit, and the result is divided by the number of groups (three in this case).
- Suppose we want to create a choropleth map with the requirement that each class will have approximately the same number of quantities. Which classification method would be best in this situation?
    - Quantile
        - In the quantile method, the number of color bins determines the number of quantiles.
- Which statement describes the classification process of the optimal classification method?
    - The optimal classification method minimizes an objective function while placing values into groups
        - Optimal classification is used for determining the appropriate number of classes needed, and it can also be used for adding spatial constraints.

Spatial Statistics

Tobler's First Law of Geography

- "Everything is related to everything else, but near things are more related than distant things"

Spatial Statistics

- Spatial dependency is the co-variation of properties within a geo-space
- If correlation exists (either positively or negatively), there are at least three possible explanations:
    - There is a simple spatial correlation relationship
        - Spatial causality
        - Spatial interaction

Covariance

$$Cov(x,\ y)\ = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Covariance: the pattern of common variation observed in collection of two (or more) datasets, or partitions of a single dataset

Person's Correlation Coefficient

$$r\ = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson's COrrelation Coefficient: a measure of similarity between two or more paired datasets

Entropy

$$I\ = -\sum_{i=1}^{k} p_i log(p_i)$$

- Entropy: a measure of the amount of pattern, disorder, or information in a set of $\{x_i\}$ where p_i is the proportion of events of values occurring in the ith class or range

Diversity

$$Div\ = \frac{-\sum_{i=1}^{k} p_i log(p_i)}{log(k)}$$

- Diversity: entropy standardized by the number of classes, k

Spatial Statistics

- For spatial data, we want to find related regions

- Can do analysis prior to the visualization to ding areas that are statistically correlated and focus the visual representation on these areas
- One method of doing this is using spatial autocorrelation

Issues in Spatial Statistics

- Scaling
- Sampling
- Logical Fallacy
- Ecological Fallacy

Distance and Direction

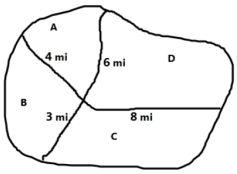
- Knowledge of location also allows the analyst to determine the distance and direction between objects
- Many types of spatial analysis require the calculation of a table expressing the relative proximity of paris of places

Elements of Matrix W

- Elements of Matrix W:
    - 1 is the places share a common boundary, else 0
    - The length of any common boundary between places, else 0
    - A decreasing function of the distance between the places, or between their representative points

Knowledge Check

- *Diagram: Map and Matrices*



Matrix 1				
	A	B	C	D
A	0	4	0	6
B	4	0	3	0
C	0	3	0	8
D	6	0	8	0

Matrix 2				
	A	B	C	D
A	0	1	0	1
B	1	0	1	0
C	0	1	0	1
D	1	0	1	0

Matrix 3				
	A	B	C	D
A	1	4	0	6
B	4	1	3	0
C	0	3	1	8
D	6	0	8	1

Matrix 4				
	A	B	C	D
A	0	4	12	6
B	4	0	11	9
C	12	3	0	8
D	6	9	8	0

Review *Diagram: Map and Matrices*. The map in the diagram shows four regions: A, B, C, and D. The map also labels the length of the shared borders between regions (e.g., the shared border between regions B and C is 3 miles long). Which matrix correctly depicts the information in the map in matrix form?

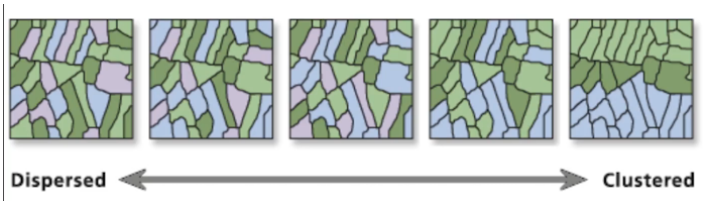
- Matrix 1
    - Matrix 1 accurately captures the length of common boundaries between the places.

- Suppose that a certain school district is known for having high math scores. Also suppose that a researcher using the aggregated math scores for the district infers that each student in a math class is a math genius. Which issue in spatial statistics can this inference be attributed to?
    - Ecological Fallacy

- Ecological Fallacy is the interpretation of a single entity based on the interpretation of the group to which it belongs. Please review "Spatial Statistics" and answer this question again.

Spatial Autocorrelation

Spatial Autocorrelation



Moran's I

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

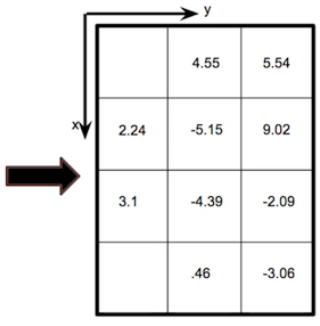
- Given a set of features and an associated attribute, Global Moran's *I* evaluates whether the pattern expressed is clustered, dispersed or random
- Values near +1.0 indicate clustering while values near -1.0 indicate dispersion
- Global Moran's *I* function also calculates a *Z* score value that indicates whether we can reject the null hypothesis of "there is no spatial clustering"

What is a Z Score?

- Most statistical tests begin by identifying a null hypothesis
- The Z score is a measure of standard deviations
- The Z score is associated with a normal distribution
- Critical Z score values when using a 95% confidence level are -1.96 and +1.96 standard deviations

Calculating Moran's *I*

X	Y	Z
1	2	4.55
1	3	5.54
2	1	2.24
2	2	-5.15
2	3	9.02
3	1	3.1
3	2	-4.39
3	3	-2.09
4	2	.46
4	3	-3.06



Geary's C

$$C = \frac{(N-1)\sum_i\sum_j w_{ij}(X_i - X_j)^2}{2W\sum_i (X_i - \bar{X})^2}$$

- Value of Geary's C lies between 0 and 2
  - Values of 1 means no spatial autocorrelation
  - Smaller than one means positive spatial autocorrelation
  - Larger than one means negative spatial autocorrelation
- Geary's C is more sensitive to local spatial autocorrelation

Local Indicators of Spatial Association

- Moran's I is a global measure of correlation
- The individual components can be mapped and tested for significance to provide an indication of clustering patterns within the study region

Getis-Ord Statistic

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij}x_j - \bar{x}\sum_{j=1}^{N} w_{ij}}{s\sqrt{\frac{N\sum_{j=1}^{N} w_{ij}^2 - \left(\sum_{j=1}^{N} w_{ij}\right)^2}{N-1}}}$$

- Unlike Moran's I, the Getis and Ord statistic identifies the degree to which high or low values cluster together

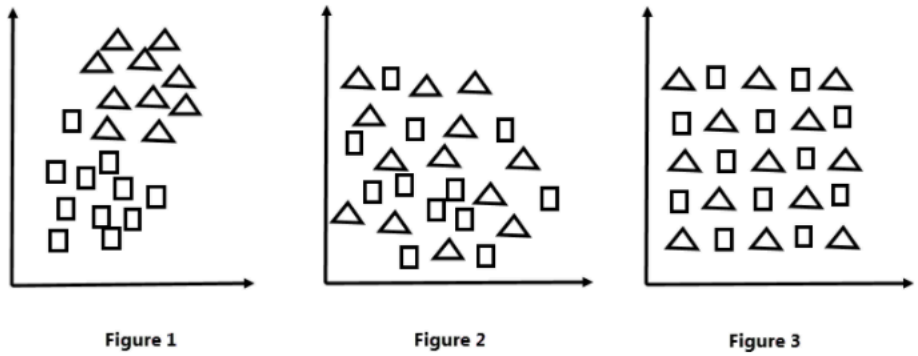
Significance Tests for Autocorrelation Indices

- Autocorrelation coefficients can be tested for statistical significance under two different model assumptions
  - Classical statistical assumption of Normality
  - Assume that values are independent and identically distributed drawings from a Normal distribution
  - Observed pattern of a set of clause is assumed to be just one realization from all possible random permutations
  - Utilizes Monte Carlo testing

Knowledge Check

- What differentiates spatial autocorrelation from standard statistical techniques? Select all that apply.
  - The assumption that data from nearby areas are statistically independent
    - Standard statistics assume observations to be independent of each other, whereas autocorrelation is a tendency to group the similar values together.
  - The space factor when computing statistical quantities
    - Unlike normal time series, spatial autocorrelation uses two dimensions.

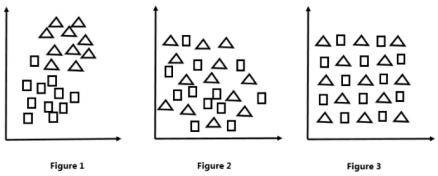
- *Diagram: Data Patterns*



Review *Diagram: Data Patterns*. In which data pattern would Moran's I value be closest to 1?

  - Figure 1
    - If Moran's I value is 1, this indicates that the pattern is clustered. Figure 1 has two separate clusters.

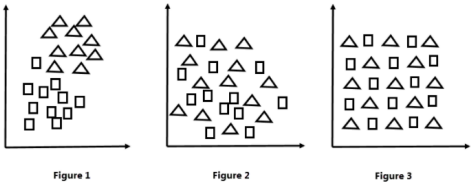
- *Diagram: Data Patterns*



Review *Diagram: Data Patterns*. Which data patterns would be associated with a z-score that is greater than 3.5?

  - Figure 1
    - Z-score is associated with a normal distribution, and as Figure 1 demonstrates proper clustering, it has a higher Z-score.

- *Data: Patterns*



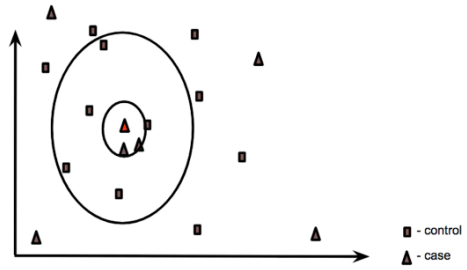
Review *Diagram: Data Patterns*. For Moran's I value, the null hypothesis states that the attribute being analyzed is randomly distributed among the features. For which data patterns can the null hypothesis be rejected? Select all that apply.

  - Figure 1
    - Moran's I function calculates the z-score, which is used to reject the null hypothesis based on spatial clustering. Figure 1 has perfect clustering, hence the null hypothesis can be rejected.
  - Figure 3

- Moran's I function calculates the z-score, which is used to reject the null hypothesis based on spatial clustering. Even though Figure 3 does not have perfect clustering, it is not showing a completely random pattern, hence the null hypothesis can be rejected.

Spatial Scan Statistics

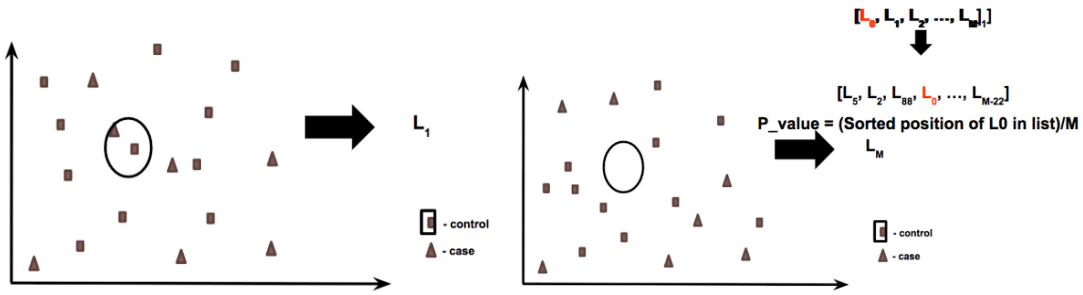
Scan Statistics



- For each circle (window) compute the likelihood function
- For Bernoulli distributions, the function is:

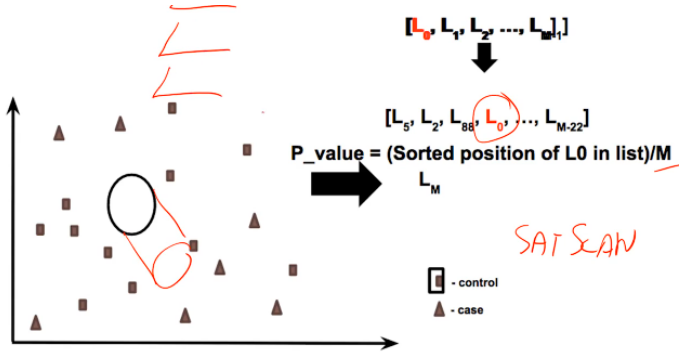
$$\left(\frac{c}{n}\right)^{c}\left(\frac{n-c}{n}\right)^{n-c}\left(\frac{C-c}{N-n}\right)^{C-c}\left(\frac{(N-n)-(C-c)}{N-n}\right)^{(N-n)-(C-c)} I()$$

$[L_0]$



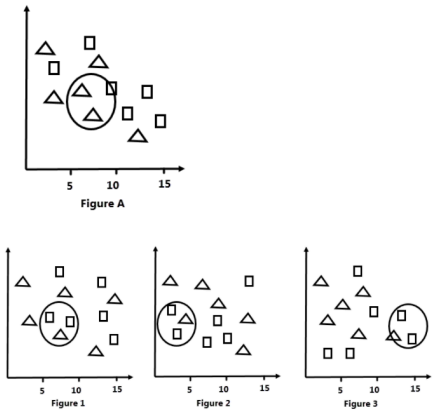
Scan Statistics

- Keeping the window in the same location, randomly redistribute the cases and controls
- Calculate the likelihood function for the new distribution
- Repeat 99, 999, xxx times and determine a p-value for the original distribution by sorting the likelihood functions
- Related work includes the Geographical Analysis Machine

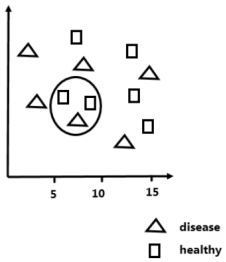


Knowledge Check

- *Image: Data Sets*



  Review *Image: Data Sets*. In this image, Figure A shows the data set and the fixed selection window used to compute the first likelihood value. Which other figure can be used to compute the second likelihood value?
    - Figure 1
      - Scan statistics capture possible windows in the dataset and evaluate whether the distribution pattern of points in that window occurred by chance or not by keeping the window at the same location.
- *Diagram: Spread of Disease*



- Review Diagram: Spread of Disease. This diagram shows the spread of a disease in a region. If the fixed-window technique is applied to determine if the patterns of the distribution or points occurred by chance or not, what is the value of the expression (C-c)/(N-n) used in the Bernoulli distribution?
    - 5/9
      - Using the figure, we can discern that C (the total population with disease) = 6, c (the population with disease in window) =1, N (total population) = 12, and n (the population in window) = 3.
- *Table: Likelihood Values*

Likelihood Distribution	Likelihood Value
L_0	0.4
L_1	0.1
L_2	0.7
L_3	0.5
L_4	0.8

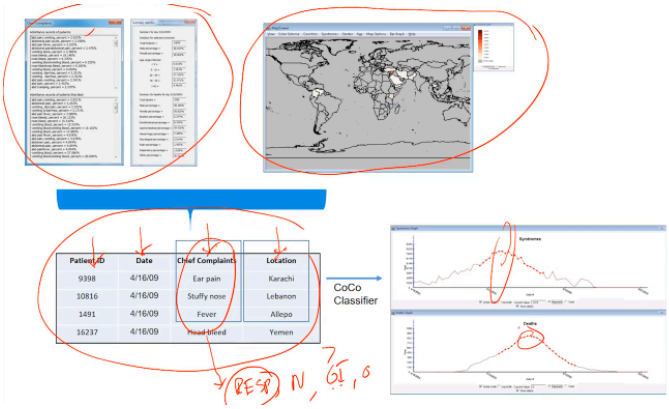
L_5	0.9
L_6	0.6
L_7	0.7
L_8	0.3
L_9	0.9

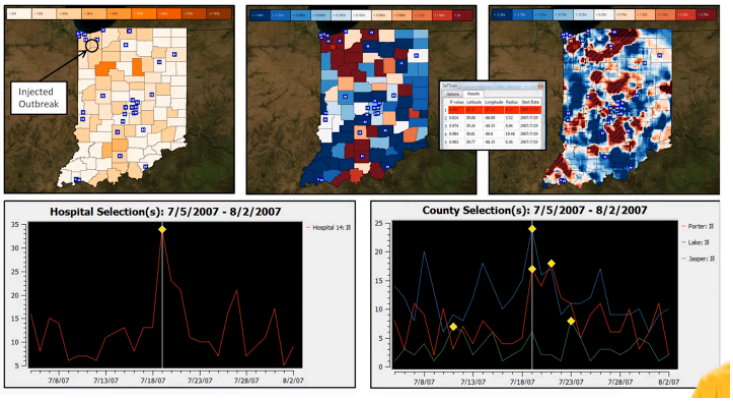
Review Table: Likelihood Values. The table gives the likelihood values for a dataset. The likelihood values were computed for ten trials using the fixed-window technique. What is the p-value of L_0?

- 0.3
  - The p-value is calculated as p = (sorted position of L_0 in the list) / (length of the list).

Geovisual Analytics Systems

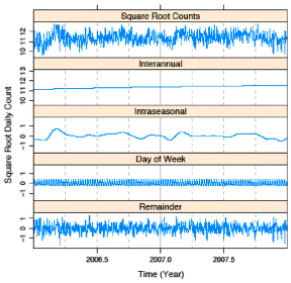
Examples



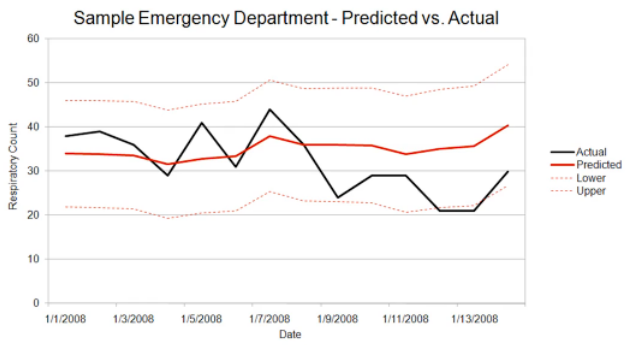


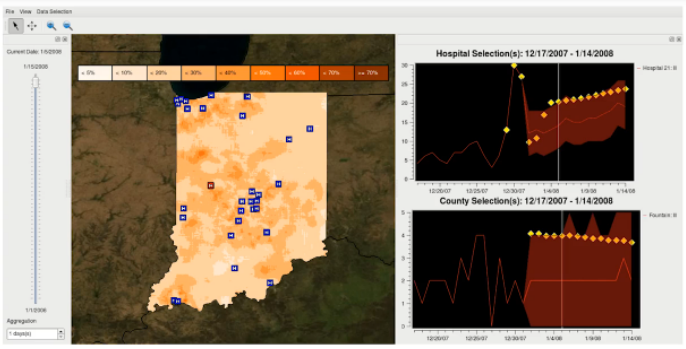
Time Series Modeling

- Seasonal-Trend Decomposition Based on Loess
  - Time series can be viewed as the sim multiple trend components
  - For each data signal, components are extracted
  - Can then analyze correlation between components



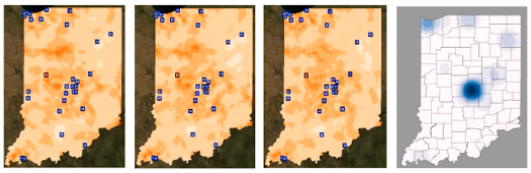
Predictive Visual Analytics





Spatiotemporal

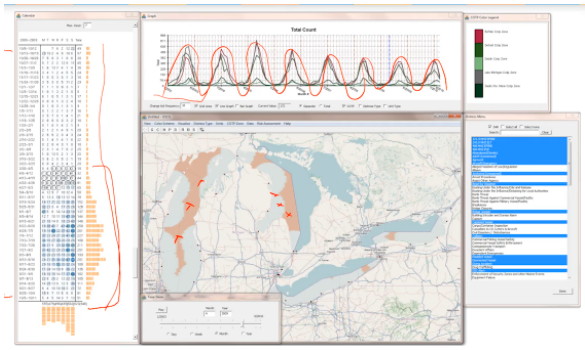
- The probability distribution of tomorrow can based on past probability distributions



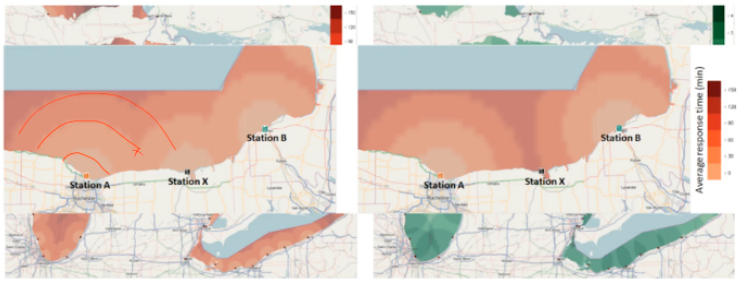
Crime Mapping

- The problem is analogous to syndromic surveillance
- Instead of patient addresses, now there is criminal incident reports
- More data (who, what, when, where, how)

Coast Guard Search and Rescue Visual Analytics (CGSARVA)



Assessing Risk in the Great Lakes



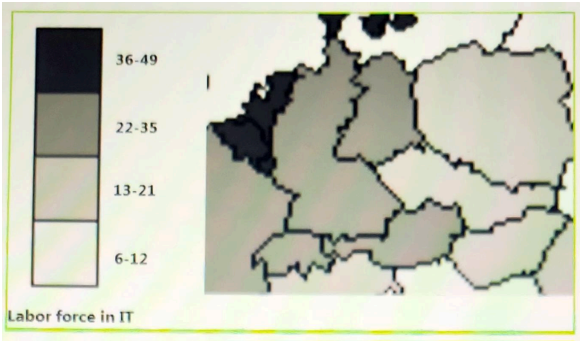
Knowledge Check

- What is the name of the system that combines geovisualization, spatial analysis, temporal analysis, and data mining?
    - Geovisual Analytics
        - Geovisual analytics is a combination of geovisualization, spatial analysis, data exploration and temporal analysis. It is generally used for performing analysis on spatial data.
- Which statements accurately describe time series? Select all that apply.

    - Time series can be decomposed into individual trend components
        - Decomposition of time series is useful for time series analysis and forecasting.
    - Time series can fit a model for predictive analytics
        - Historical time series data can be used for performing time series prediction and forecasting.
    - Time series can account for periodic fluctuations in the data
        - Time series can fit periodic fluctuations. It also accounts for the data that exhibits a rise and fall cyclic pattern with no fixed period.
- Which examples of data exploration can be accomplished at geographic scales of measurement by using geovisual analytics systems? Select all that apply.
    - Predicting the future
        - Combining different methods into a single interactive system helps us to calculate the past probability distributions to predict future probabilities.
    - Spatiotemporal Analysis
        - Since the interactive system can combine both the temporal elements and time series data, it helps in performing data exploration with both space and time.
    - Finding temporal anomalies
        - Spatial scan statistics can be used for finding temporal anomalies.

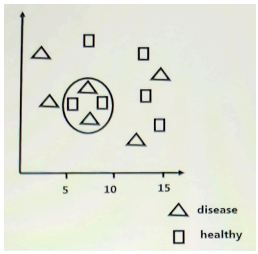
Module 5 Quiz Questions

- What are the characteristics of a thematic map? Select all that apply:
    - Provides specific information about important locations
    - Can use multiple maps to compare patterns
        - Thematic maps are a good tool to show multiple maps simultaneously to compare the patterns.
- *Figure: Labor Force Map*



Review *Figure: Labor Force Map*. The map illustrates the labor force in Information Technology (IT). What type of color scheme is represented on this map?

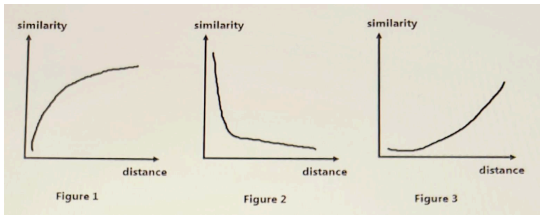
    - Sequential
- Figure: *Disease Spread*



Review *Figure: Disease Spread*. The figure shows the spread of a disease in a region. What is the value of the expression (C-c)/(N-n) used in Bernoulli distribution when the fixed windows technique is applied to determine whether the patterns of the distribution of the points occurred by chance?

- 1/2
  - $\frac{(C-c)}{(N-n)}$
  - $= \frac{Total\ number\ of\ disease\ cases\ in\ entire\ region - number\ of\ disease\ cases\ inside\ fixed\ window}{Total\ number\ of\ individuals\ (disease\ and\ healthy)\ in\ entire\ region - TOtal\ Number\ of\ individuals\ inside\ fixed\ window}$
  - $= \frac{6-2}{12-4} = \frac{1}{2}$

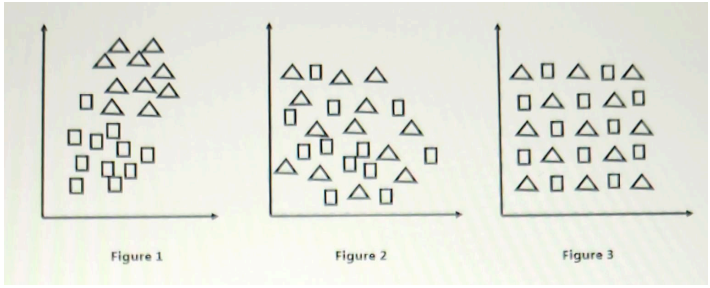
- *Image: Graphs*



Review the image: Graphs. Which graph can be used to explain Tobler's First Law of Geography?

- Figure 2
  - As distance increases in this region, similarity decreases

- *Image: Data Patterns*



Review *Image: Data Patterns*. For which data pattern would Geary's C value be closest to 0?

- Figure 1

- Which kind of coloring scheme puts emphasis on critical mid-range values and hence is suitable for ratio Data types when looking at above and below point zero points?
  - Diverging
- Which kind of coloring scheme is suited for ordered, ordinal, interval, and ratio data types?
  - Sequential
- What is the name of the projection that maps a globe to a cylinder with equally spaced longitudes and unequally spaced latitudes where it enlarges the countries along the equator?

    - Peters
- Which statements about Z-score are correct? Select all that apply:
    - Z-score is a measure of standard deviation
    - Z-score is a statistical test to identify a null hypothesis
- Suppose we have a county dataset for a state, where some counties share some common attributes. For this dataset, spatial autocorrelation is found to be -1. What does -1 indicate about the counties relating to each other?
    - The counties are dispersed

Introduction to Trees and Hierarchies

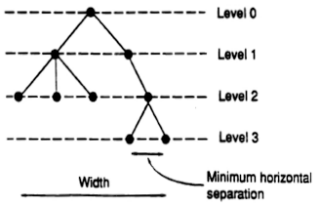
Hierarchies

- Definition
    - Data repository in which cases are related to subcases
    - Can be thought of as imposing an ordering in which cases are parents or ancestors of other cases

Trees

- Hierarchies are often represented as trees
    - Directed
    - Acyclic
- Two main representation schemes
    - Node-link
    - Space-Filling

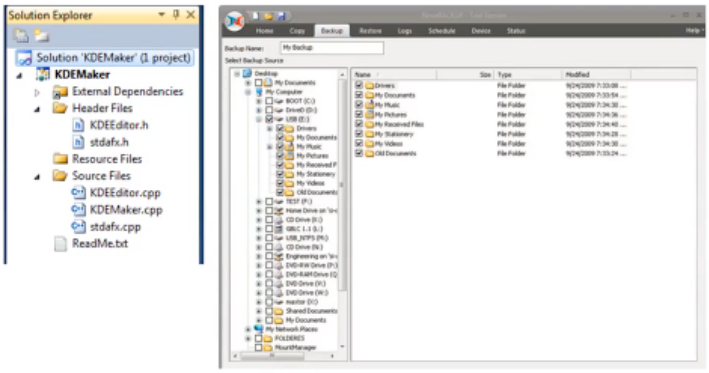
Rooted Trees



- A graph might be used to represent some hierarchy, so we often utilize a tree metaphor
- Typically these utilize the following aesthetics
    - Vertices are placed along horizontal lines according to their level
    - Minimum separation distance between two consecutive vertices on the same level
    - Width of the drawing is as small as possible

Using Rooted Trees

What are such sorts of structures useful for?



Top-Down Approach

- Width of fan-out uses up horizontal real estate very quickly
    - At level n there are 2^n notes
- Trees may grow very long in one branch
- Essentially uo can wind up leaving a lot of screen real estate empty

Space Tree

- Visualization techniques try to overcome some of these issues in node link tree diagrams
- *Space Tree* by Plaiant et al.
    - Dynamic rescaling of branches to best fit available screen space
    - Utilized preview icons to summarize branch topology
- Don't have to constrain to top-down geometry approach
- Apply to a hyperbolic transformation to the space
- Distance between parent and child decreases as you move farther from the center
- Children go in a wedge rather than a circle

Cone Trees

- Add a third dimension for the layout
- Children of a node are laid out in a cylinder below the parent
    - Siblings live in one of the 2D planes

Node-Link Shortcomings

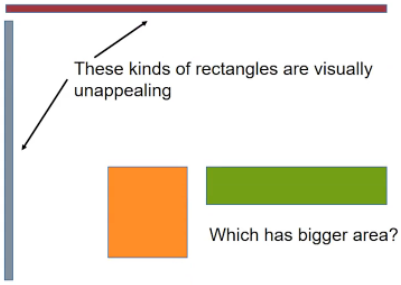
- Difficult to encode more variables of data cases
    - Shape, Size, Color
- All of these can clash with the basic node-link structure

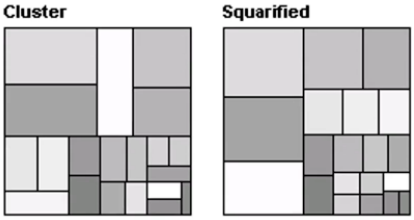
Tree Map Algorithms

Aspect Ratios

- Cleveland's "Banking to 45"

- Here, the aspect ratio will drastically affect the visualization





Clustered and Squarified Treemaps

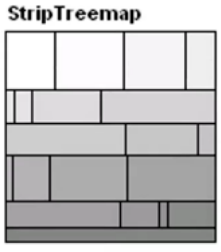
- Simple recursive algorithm to reduce overall aspect ratio
- Bruls et al. introduced squarified treemap
- Methods has two major drawbacks
    - Changes in the set can cause discontinuities in the layout
        - If treemap data is dynamic large visual changes make data hard to track
    - If the data has ordering information this is not preserved

Ordered Treemap Algorithm

- Starting with a rectangle R to be subdivided, first algorithm pivot-by-size, the pivot is item with largest area
    - Let P, the pivot be the item with the largest area in list of items
    - If width R is greater than or equal to the height, divide R into four rectangles
    - Place P in the rectable R_p
    - Divide the items in the list, other than P, into three lists, $L_1,\ L_2,\ L_3$ to be laid out in the $R_1,\ R_2$ and R_3
    - Recursively lay out $L_1,\ L_2$ and L_3 in $R_1,\ R_2$ and R_3

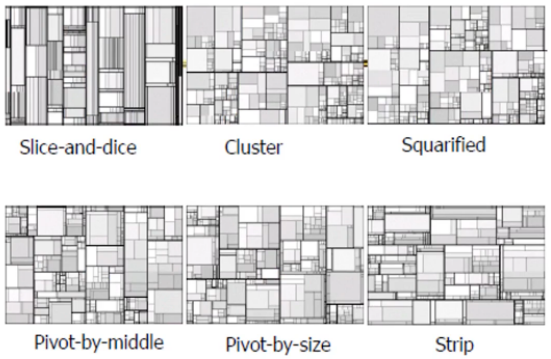
Strip Treemaps

- Scale the area of al rectangles so total area of input rectangles equals that of layout rectangle
- Create a new empty strip, the current strip
- Add next rectangle to current strip, recomputing height of strip based on area of all rectangles within the strip and recomputing width of each rectangle
- If average aspect ratio of the current strip has increased, as a result of adding rectangle in step 3, remove rectangle pushing it back onto list of rectangles and go to step 2
- If all rectangles have been processed stop, else step 3



Metrics for Treemaps

- In order to assess all these different treemap algorithms, we need metrics to define how "good" they are
- Use two metrics:
    - Average aspect ratio of treemap layout
    - Layout distance change function
- Average aspect ratio is the unweighted average



Cushion Treemap

- Using chasing and texture to help convey structure of hierarchy

Another Problem

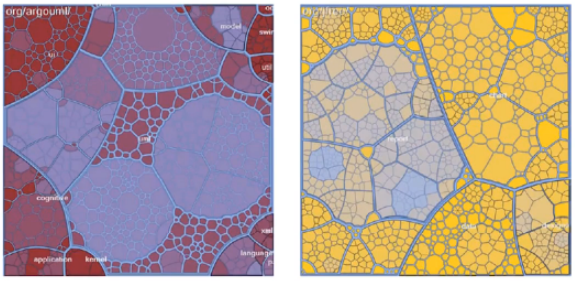
- What if nodes with zero value are very important?
- If we're mapping areas, how do we map to zero?
    - Example: stocks portfolios

Context Treemap

- One way to overcome this is distort classic treemap visualization
- Distorted treemap can show one more attribute than a classic treemap
    - node area is no longer proportional to attribute being visualized
- Several different implementation strategies for this
    1. Use a regular tree map but add some epsilon to zero value item
    2. Use an exponential mapping area(node)=2^(value(node))
    3. Assign some minimal screen space size to zero nodes
- Final solution was to calculate intermediate values
    1. Calculate the total (in this paper it was total invest money)
        a. Value(total)
    2. Create an additional total with respect to the context
        a. Value(total)*v, where v can be modified as a scale vector
    3. Split context screen real estate amon all empty nodes
        a. $Value_c$=value(total)*v#empty

$$value'(node) = \begin{cases} value_c \text{ if } value(node) = 0 \\ value(node) \text{ otherwise} \end{cases}$$

Voronoi Treemaps

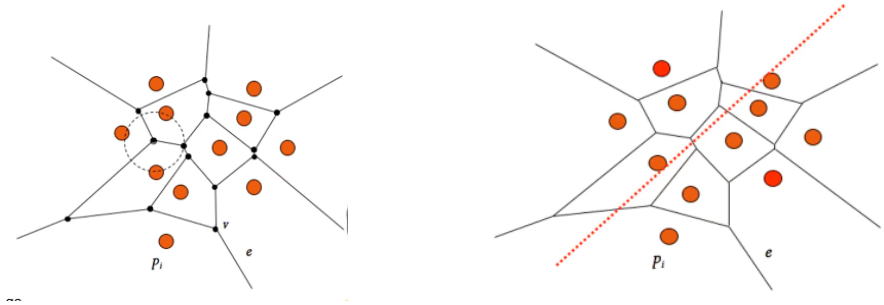


Definition of Voronoi Diagram

- Let P be a set n distinct points(sites) in the plane
- The Voronoi diagram of P is the subdivision of the plan into n cells, one for each site
- A point q lies in the cell corresponding to a site $p_i \in P \; iff$
- Main Algorithm is Fortune's Algorithm
- Euclidean_Distance(q, p_i) < Euclidean_distance(q, p_j), for each $p_i \in P$, $j \neq i$

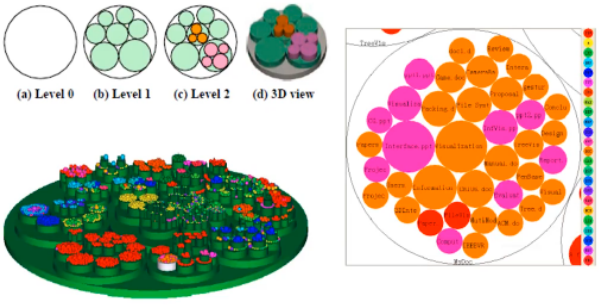
Summary of Voronoi Properties

- A point q is a vertex iff the largest empty circle centered at q touches at least 3 sites
    - A Voronoi vertex is an intersection of 3 or more segments, each equidistant from a pair of sites
        - p_i : site points
        - e : Voronoi edge
        - v : Voronoi vertex



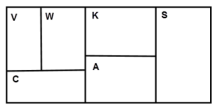
- Voronoi diagrams have linear complexity $\{|v|, \; |e| \; = O(n)\}$
    - Intuition: Not all bisectors are Voronoi edges!

Circle Packing



Knowledge Check

- Which descriptions correctly identify what space trees do? Select all that apply.
    - Space trees overcome screen space issues in node-link
        - Space trees overcome the problem of white spaces by dynamically re-scaling their branches to best fit the available screen space.
    - Space trees use dynamic rescaling to fit the available screen space
        - Space trees overcome the problem of white spaces by dynamically re-scaling their branches to best fit the available screen space.
    - Space trees summarize branch topology by utlizing preview icons
        - Space trees use preview icons to summarize branch topology and keep it compact.
- What are examples of a tree representation scheme? Select all that apply.
    - Space-filling
        - Space-filling is a scheme that looks like mosaic plots and uses space trees for dynamic re-scaling to fit the screen space.
    - Node-link
        - The node-link scheme is a hierarchical structure where nodes are placed in a horizontal line at fixed distances until they fit the screen space.
- *Diagram: Treemap and Trees*



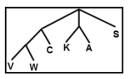


Figure 1

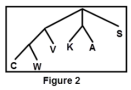


Figure 2

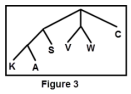
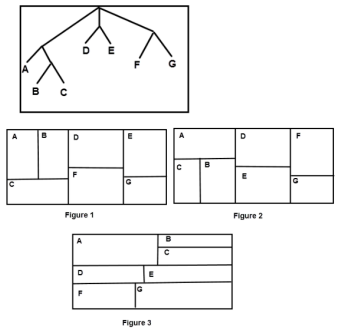


Figure 3

Review *Diagram: Treemap and Trees*. The diagram provides a treemap and three trees. Which figure represents the tree of the given treemap?

- Figure 1 accurately represents the space-filling diagram. As per the space filling representation, the children are placed inside the parent, and the horizontal and vertical slices depict the successive levels in a tree.

- *Diagram: Tree and Treemaps*



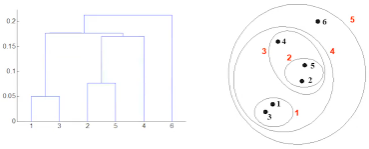
Review *Diagram: Tree and Treemaps*. The diagram provides a tree and three treemaps. If the initial cut is vertical, which figure represents the treemap for the given tree?

- Figure 2
  - Figure 2 accurately depicts the treemap. According to the Treemap algorithm, the horizontal lines represent the levels, and the vertical lines are used to denote the children under the parent nodes. The rectangles are scaled to size.

- What is the average line slope in a line chart proposed by Cleveland?
  - 45 degrees
    - The average line slope technique proposed by Cleveland is known as banking to 45 degrees, and it is designed to maximize the discriminability of the orientations of the line segments in the chart.
- Which algorithms are used in the construction of treemaps? Select all that apply.
  - Ordered
    - Ordered treemap is used to overcome the limitations of squared treemap. In this layout, the items that are next to each other are adjacent in the treemap.
  - Squarified
    - Squarified treemaps have overcome the limitation of ordering of direction and reduced the overall aspect ratios.
  - Clustered
    - Clustered treemap is a recursive algorithm used to reduce the overall aspect ratio.

What is Hierarchical Data Clustering

Hierarchical Clustering



- Produces a set of nesters clusters organized as a hierarchical tree
- Can be visualized as a dendrogram (along with other options)
  - A tree-like diagram that records the sequences of merges or splits
- Agglomerative:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or *k* clusters) left
- Divisive:
  - Start with one, all-inclusive cluster

- At each step, split a cluster until each cluster contains a point)or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Strengths of Hierarchical Clustering

- No assumptions on the number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level
- Hierarchical clusterings may correspond to meaningful taxonomies
  - Example in biological sciences (e.g. phylogeny reconstruction, etc), web (e.g., product catalogs) etc.

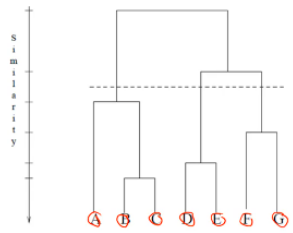
Complexity of Hierarchical Clustering

- Distance matrix is used for deciding which clusters to merge/split
- Not usable for large datasets
- At least quadratic in the number of data points

Agglomerative Clustering

Agglomerative Clustering Algorithm

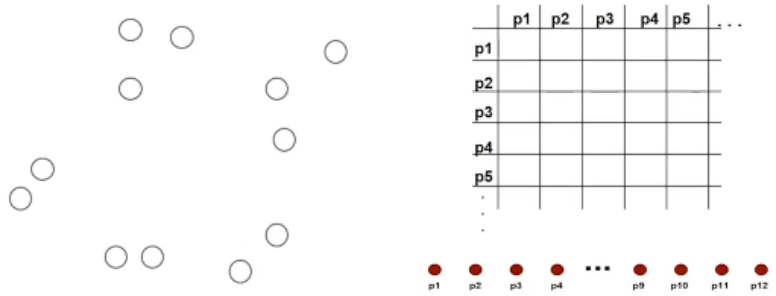
- Most popular hierarchical clustering technique



- Key operation is the computation of distance between two clusters
  - Different definitions of the distance between clusters lead to different algorithms

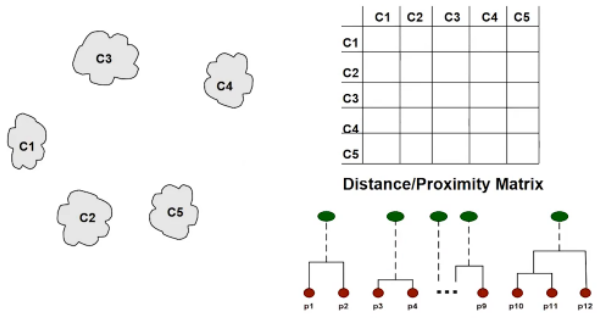
Hierarchical Clustering: Input/Initial Setting

- Start with clusters of individual points and a distance/proximity matrix



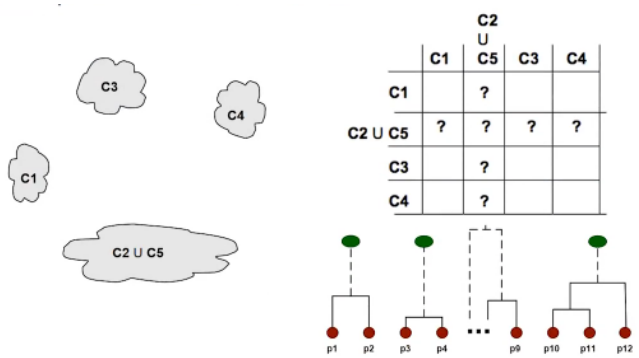
Intermediate State

- After some merging steps, we have come clusters



After Merging

- "How do we update the distance matrix?"



Distance Between Two Clusters

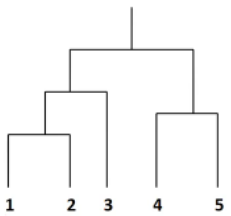
- Each cluster is a set of points
- How do we define distance between two sets of points
  - Lots of alternatives
  - Not an easy task
- Single-link distance between clusters C_i and C_j is the minimum distance between any object in C_i and any object in C_j
- The distance is defined by the two most similar objects
  - $D_{si}(C_i, C_j) = min_{x,y}\{d(x, y) | x \epsilon C_i, y \epsilon C_j\}$

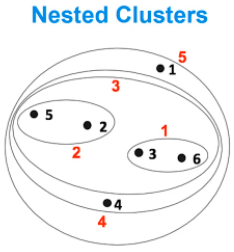
Distance Metrics in Hierarchical Clustering

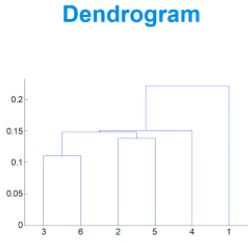
Single-Link Clustering: Example

- Determined by one pair of points, i.e., by one link in proximity graph

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

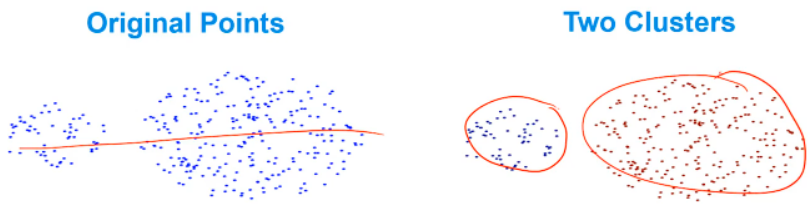




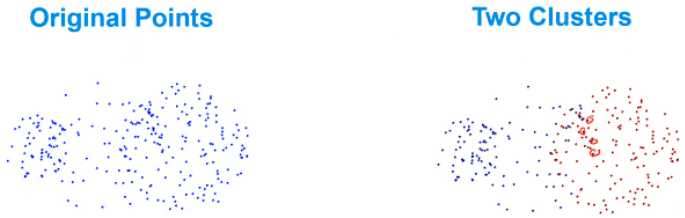


Strengths of Single-Link Clustering

- Can handle non-elliptical shapes



Limitations of SIngle-Link Clustering

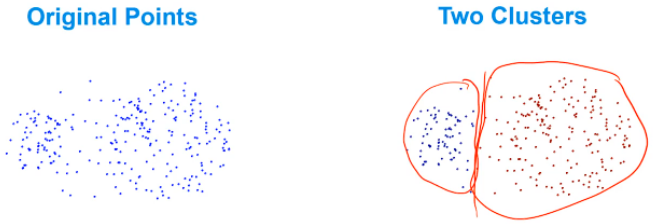


- Sensitive to noise and outliers
- It produces lone, elongated clusters

Complete-Link CLustering: Example

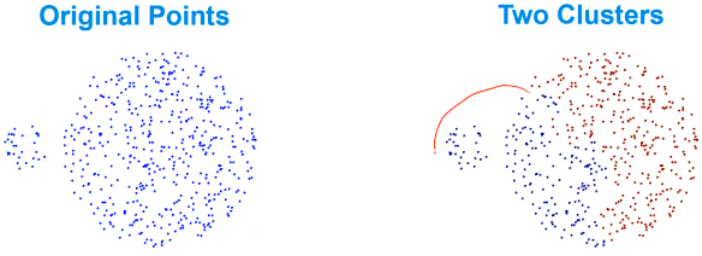
- Distance between clusters in determined by two most distant points in different clusters

Strengths of Complete-Link Clustering



- More balanced clusters (with equal diameter)
- Less susceptible to noise

Limitations of Complete-Link Clustering

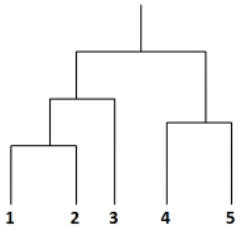


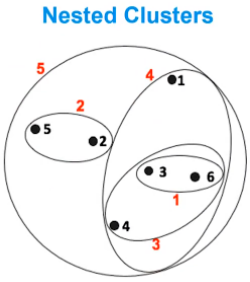
- Tends to break large clusters
- All clusters tend to have same diameter - small clusters are merged with larger clusters

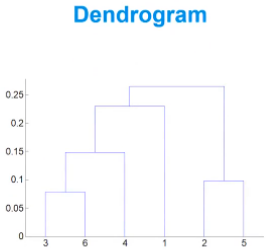
Average-Link Clustering: Example

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00







Average-Link Clustering: Discussion

- Compromise between Single and Complete Link
  - Strengths:
    - Less susceptible to noise and outliers
  - Limitations:
    - Biased toward globular clusters

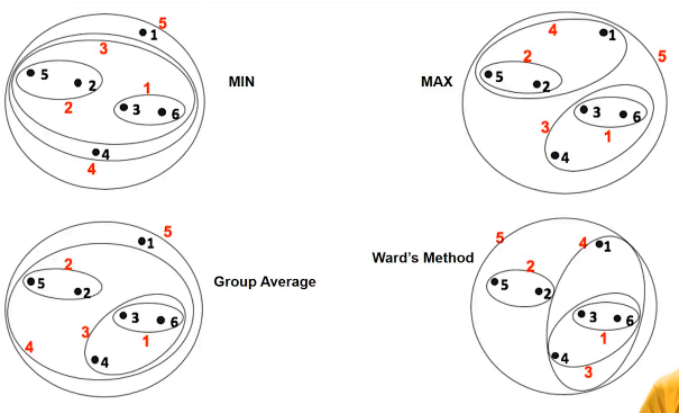
Distance Between Two Clusters

- Centroid distance between clusters C_i and C_j is the distance between the centroid r_i of C_i and the centroid r_j of C_j
  - $D_{centroids}(C_i, C_j) = d(r_i, r_j)$

Ward's Distance for Clusters

- Similar to group average and centroid distance
- Less susceptible to noise and outliers
- Biased toward globular clusters
- Hierarchical analogue of k-means
  - Can be used to initialize k-means

Hierarchical Clustering: Comparison



Hierarchical Clustering: Time and Space Requirements

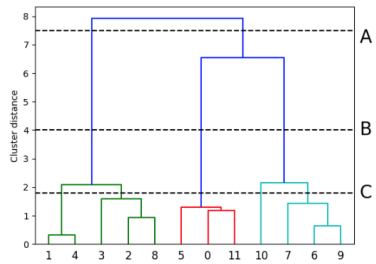
- For a distance X consisting of n points
- $O(n^2)$ space; it requires storing distance matrix
- $O(n^3)$ time in most of the cases
    - There are n steps and at each step the size n^2 distance matrix must be updated and searched
    - Complexity can be reduced to $O(n^2 log(n))$ time for some approaches by using appropriate data structures

Hierarchical Clustering Issues

- Distinct clusters are not produced
- Methods for producing distinct clusters but involve specifying somewhat arbitrary cutoff values
- What if data doesn't have a hierarchical structure?
- Is HC appropriate?

Knowledge Check

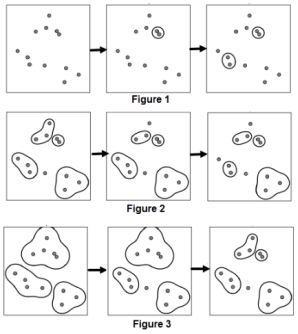
- *Figure: Dendrogram*



Review *Figure: Dendrogram*. If the dendrogram in the figure is cut by line C, how many clusters will be created?

- 5
    - Hierarchical clustering uses dendrograms to represent the clusters. In the figure, we obtain 5 clusters by cutting the dendrogram at level C.

- *Image: Iterative Steps*



Review *Diagram: Iterative Steps*. The image gives three figures that represent the iterative steps in hierarchical clustering for a dataset. Which figure belongs to agglomerative clustering?

  - Figure 1
    - In agglomerative clustering, we start with points as clusters and merge the closest points together until only one cluster is left. Figure 3 represents the agglomerative clustering.
- Which statement accurately describes how agglomerative and divisive clustering work?
  - Agglomerative clustering starts with the points as individual clusters and divisive clustering starts with one all-inclusive cluster
    - In agglomerative clustering, we start with points as clusters and merge the closest points together until only one cluster is left. In divisive clustering, the clusters are split until a single point remains.
- *Figure: Distance Matrix*

	P	Q	R	S	T	U
P	0					
Q	13	0				
R	12	5	0			
S	27	10	5	0		
T	24	4	18	19	0	
U	13	15	3	12	29	0

Review *Figure: Distance Matrix*. The figure gives the distance matrix between six data points. When single-link distance is used, which two points will be included in the first cluster constructed by hierarchical agglomerative clustering?

  - (R, U)
    - The single linkage distance between two clusters is the minimum distance between any object in cluster1 and any object in cluster2. The smallest distance, 3, is between R and U, which will form the first cluster.
- *Figure: Similarity Matrix*

	A	B	C	D	E
A	1				
B	0.5	1			
C	0.6	0.3	1		

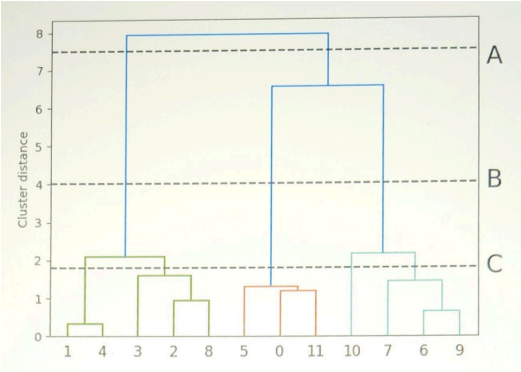
D	0.8	0.1	0.95	1	
E	0.9	0.2	0.3	0.5	1

Review *Figure: Similarity Matrix*. The figure gives the similarity matrix between five objects. Which two objects are farthest from each other?

- B and D
  - The similarity between two objects is closer to 0 if the points are farther from each other. Using the figure, we can discern that points B and D have the smallest similarity of 0.1.

Module 6 Quiz Questions

- Which statements about space trees are the most accurate? Select all that apply:
  - Space trees apply hyperbolic transformation to the space
  - Space trees move children in a wedge of space rather than a circle
  - Space trees are not contained to a top-down geometry approach
- In the context of Cleveland's ratio for the line slope on a line chart, what does "baking to 45 degrees" mean?
  - The average line slope in a line chart is 45 degrees
- *Figure: Dendrogram*



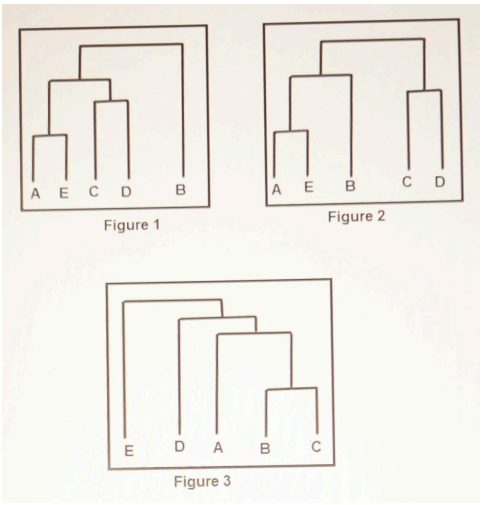
Review *Figure: Dendrogram.* How many clusters will be created if the dendrogram is cut by line B?

- 3 clusters

- Which algorithm assigned all of the observations to a single cluster and then partitions the cluster to two least similar clusters?
  - Divisive Clustering Algorithm
- What are the characteristics of a divisive clustering algorithm? Select all that apply:
  - Top Down Clustering
  - Splits one cluster at a time
- In which applications can treemaps be used? Select all that apply:
  - Stock market data
  - Sports analysis
  - File directory structures
- What is a drawback of treemaps? (Select one)
  - Treemaps do not show the hierarchical levels as clearly as other charts that visualize hierarchical data
- *Image 1: Distance Matrix*

  Image 2: Dendrograms

	A	B	C	D	E
A	0				
B	5	0			
C	6	3	0		
D	18	11	12	0	
E	19	12	13	15	0



Review the images. Image 1 gives the distance matrix between five data points. Which figure in Image 2 is the dendrogram constructed by hierarchical agglomeration clustering when single-link distance is used?

- Figure 3

- *Image: Figure: Similarity Matrix*

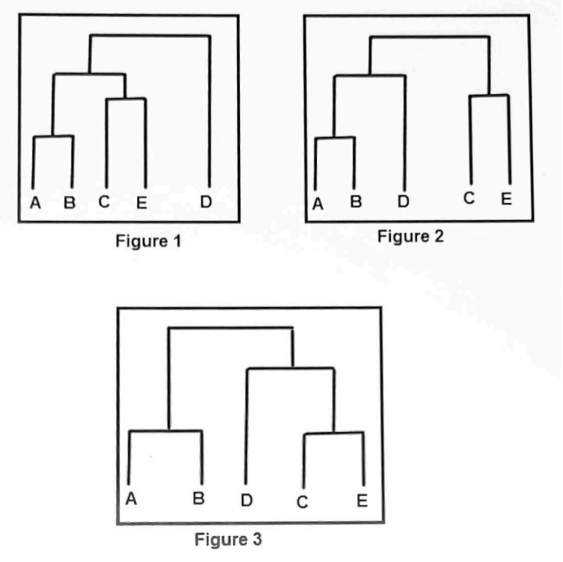
	A	B	C	D	E
A	1				
B	0.5	1			
C	0.6	0.3	1		
D	0.8	0.1	0.95	1	
E	0.9	0.2	0.3	0.5	1

Review the images. The figure gives the similarity matrix between five objects. Which two objects are farthest from each other?

- B and D

- *Image 1: Similarity Matrix*
  Image 2: Dendrograms

	A	B	C	D	E
A	1				
B	0.9	1			
C	0.3	0.7	1		
D	0.5	0.6	0.2	1	
E	0.54	0.81	0.8	0.65	1



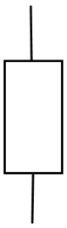
Review the images. For the provided Similarity Matrix between five objects, which figure is the dendrogram for complete-link clustering?

- Figure 2

Loading Data Into Tableau

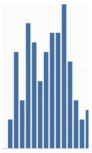
Knowledge Check

- What is the section in Tableau that is used to select data items to place in the "Columns" shelf?
  - Dimensions
    - Dimensions can be dragged into the column shelf to select data items and create quick visualization.
- *ImageL Tableau Tool Example*



Review *Image: Tableau Tool Example*. Which visualization tool in Tableau is this image an example of?

- Box-and-Whisker Chart
  - The box-and-whisker plot is a way of displaying the data distribution through quartiles.
- *Image: Tableau Tool Example*



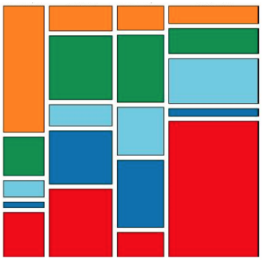
Review *Image: Tableau Tool Example*. Which visualization tool in Tableau is this image an example of?

- Bar Chart
  - A bar chart presents data in the form of rectangular bars with heights that are proportional to the values.

Advanced Tableau

Knowledge Check

- *Image: Tableau Tool Example*



Review *Image: Tableau Tool Example*. Which visualization tool in Tableau is this image an example of?

- Mosaic Plot

- *Image A: Data Plot*

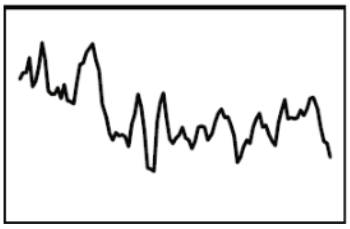
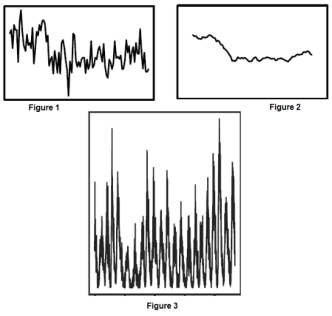


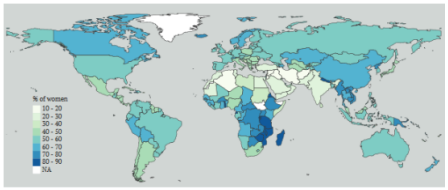
Image B: Plots



Review *Image A: Data Plot* and *Image B: Plots.* Which figure in Image B would be the "average moving" plot of the data plot in Image A?

- Figure 8
  - Moving average is used as a technical indicator to identify the trend direction. Figure 1 accurately depicts the trend followed in Image A.

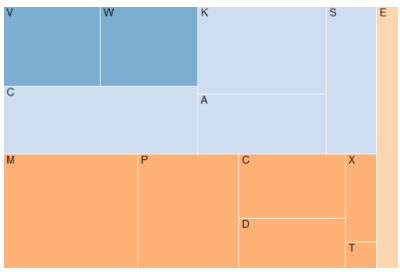
- *Image: Tableau Tool Example*



Review *Image: Tableau Tool Example*. Which visualization tool in Tableau is this image an example of?

- Choropleth Map
  - Choropleth maps display divided geographical areas or regions that are colored, shaded, or patterned in relation to a data variable.

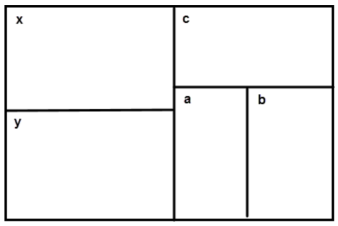
- *Image: Tableau Tool Example*



Review *Image: Tableau Tool Example.* Which visualization tool in Tableau is this image an example of?

- Tree Map

        - Tree map is used to display data in the form of nested rectangles.
- *Diagram: Tableau Tool Example*



Review *Diagram: Tableau Tool Example.* What is the name of the visualization tool that displays the expression "(x+y)*((a+b)/c)" by the plot in the diagram?

    - Tree Map
        - TreeMap is used to display data in the form of nested rectangles.
-

-

-

Knowledge Check

- —

Module x Quiz Questions

- —