Introduction to Data Exploration

What is Visualization?

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective 1

Define visualization

Objective 2

Describe key purposes of visualization

What is Visualization?

Definition

"The use of computer-supported, interactive visual representations of data to amplify cognition."[1]

1-Readings in Information Visualization: Using Vision to Think, SK Card, J Mackinlay and B.Shneiderman, 1999

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.

What is Visualization?

We want to help people form a mental image of something and internalize their own understanding

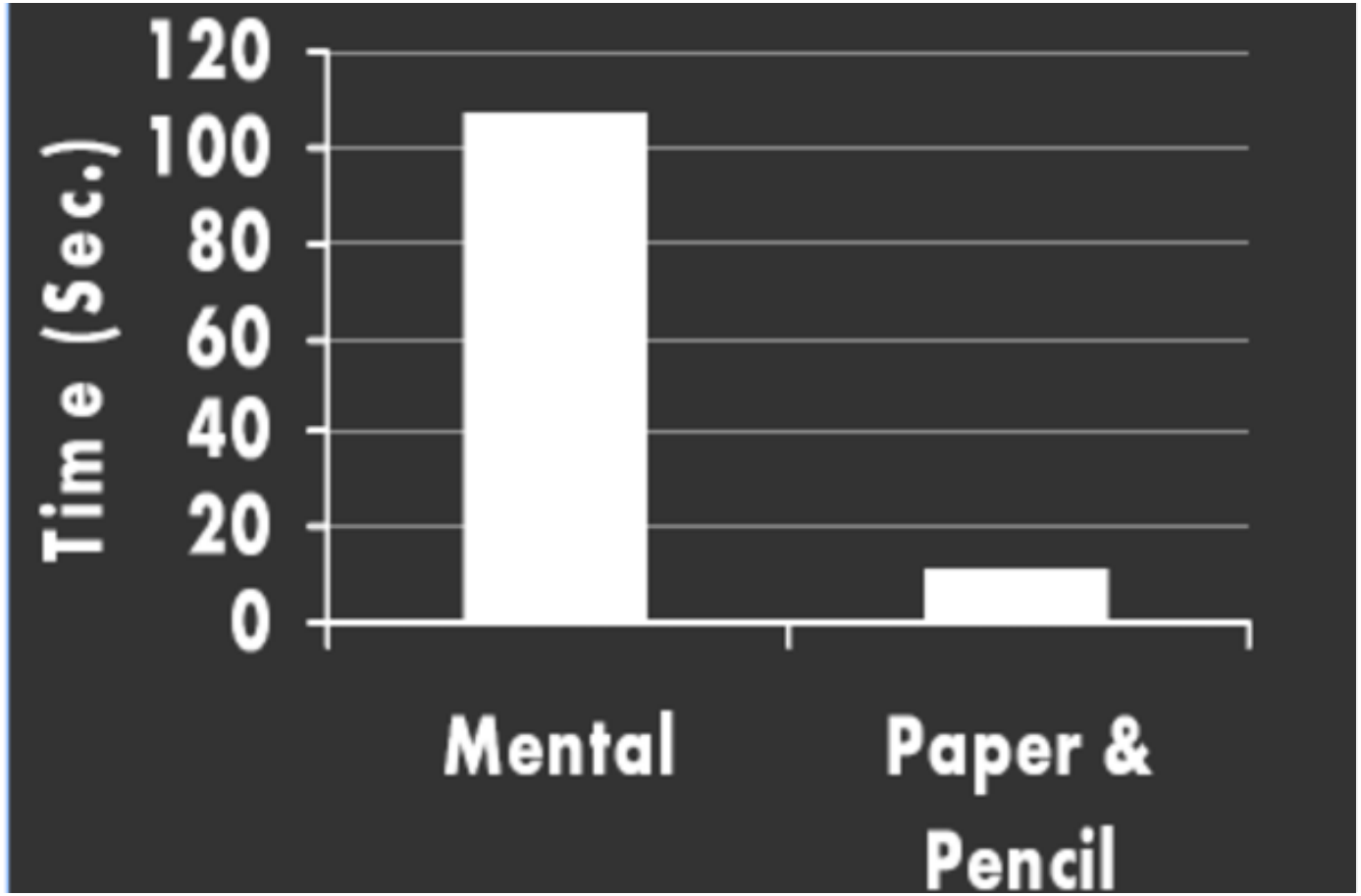
We want to promote discovery, decision making and explanations

We want to find and utilize cognitive and perceptual principles

We want to optimize our visualizations and our interactions with the visualization according to these principles

Why is Visualization Helpful?

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



Purpose of Visualization

1. **Analysis** – Understand your data better and act upon that understanding
2. Given a data set, compare, contrast, assess, evaluate
3. Solve a problem!
4. **Presentation** – Communicate and inform others more effectively
5. Visualization is most useful in **exploratory data analysis**[1]

> Information visualization is ideal for exploratory data analysis. Our eyes are naturally drawn to trends, patterns, and exceptions that would be difficult or impossible to find using more traditional approaches, such as tables or text, including pivot tables. When exploring data, even the best statisticians often set their calculations aside for a while and let their eyes take the lead.
>
> - S. Few
> *Now You See It*

[1] J. W. Turkey. Exploratory Data Analysis

Introduction to Data Exploration

Data Processing vs. Querying vs. Exploration

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Explain why we need data exploration

We are living in a data-rich world…

rehabilitation
energy
training
security
smart-offices
production
life-sciences
smart-rooms
defense
sports
elderly-care
robotics
retail
supply-chain
child-care
personal-data management
entertainment
arts
pet care
education
space exploration
advertisement
sciences
business/enterprise

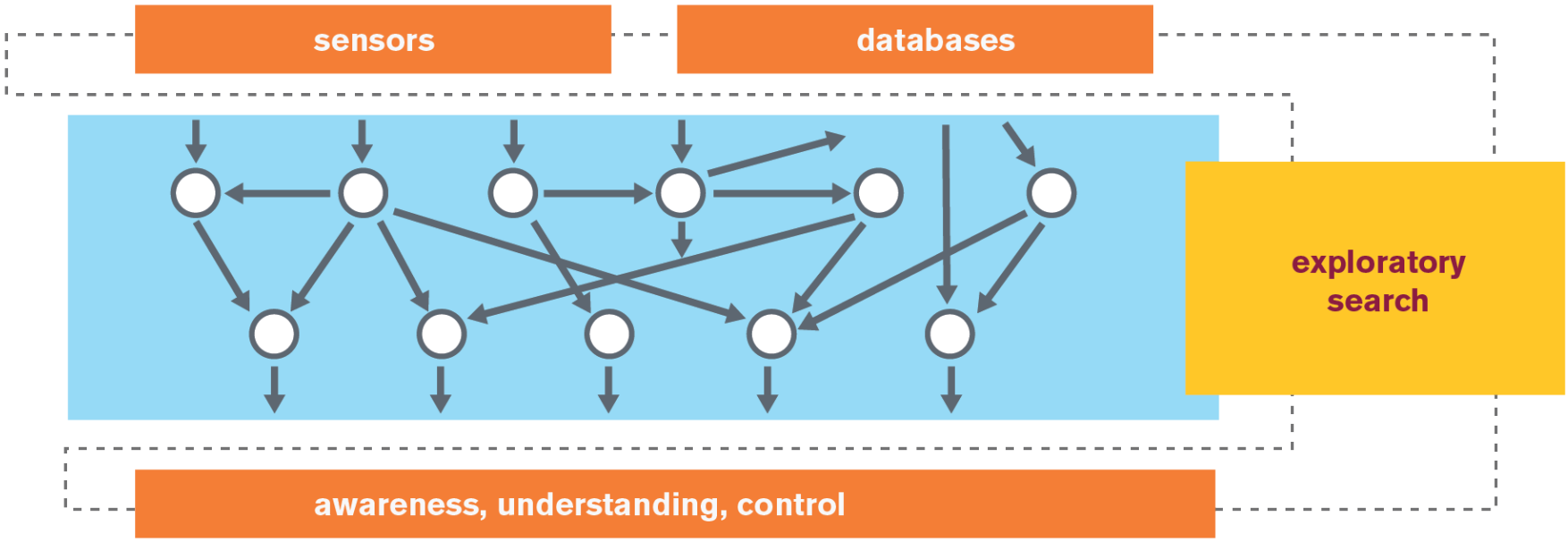
How can we make SENSE from the **REAL WORLD data**

“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 interferences 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 not 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/differences)
- Aggregation (create groups, clusters)
- Transformation (use a more convenient representation)
- Visualization

Introduction to Data Exploration

Introduction to Data Exploration

K. Selcuk Candan, Ph.D

Professor of Computer Science and Engineering

Arizona State University

Ira A. Fulton Schools of Engineering

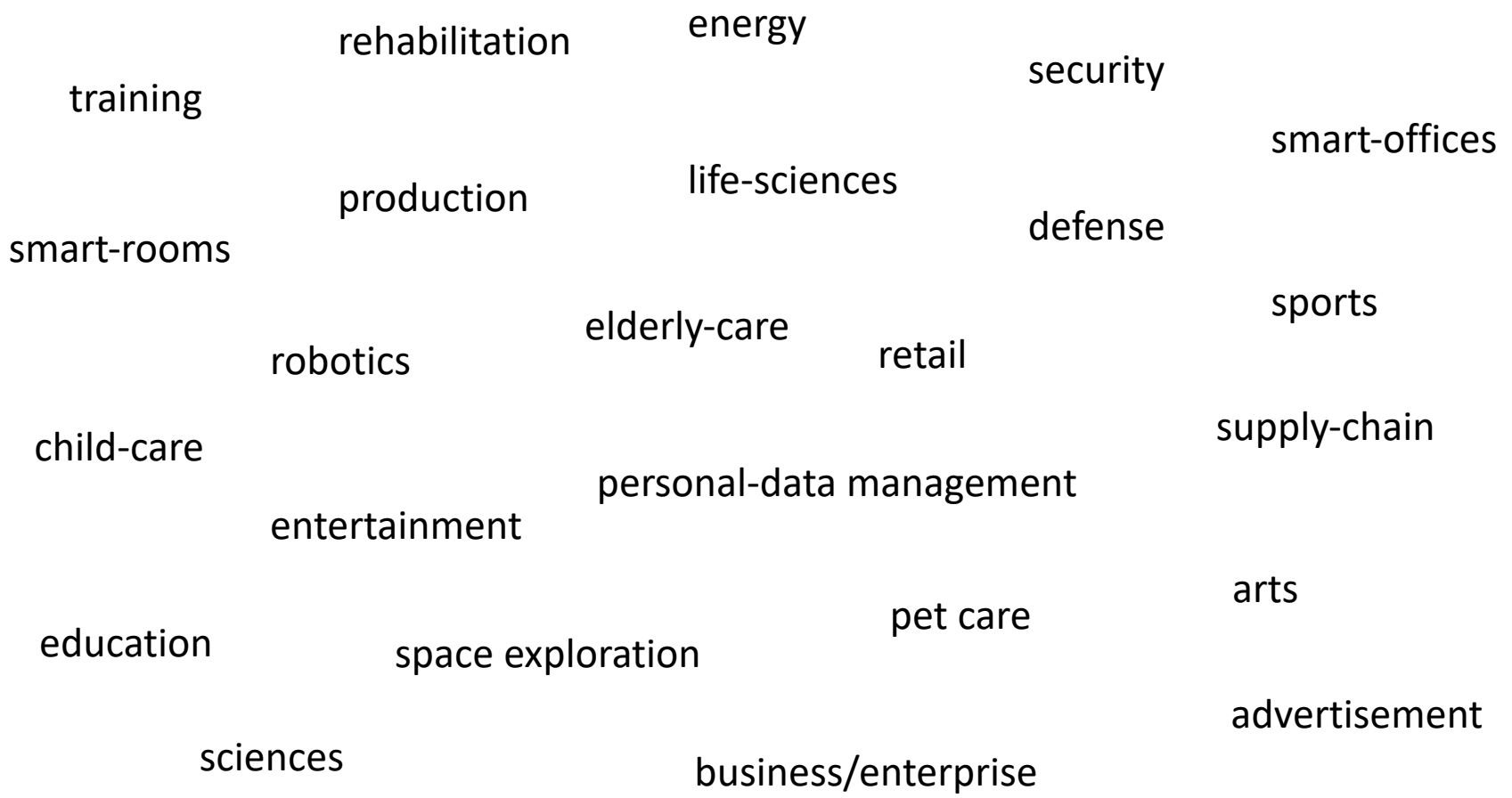
Arizona State University

Objectives

Objective

Explain what are the data challenges

We are living in a data-rich world…



How can we make SENSE from the REAL WORLD data

Data Challenges

INS

(**I**)mprecision
(**N**)oise
(**S**)parsity

3Vs

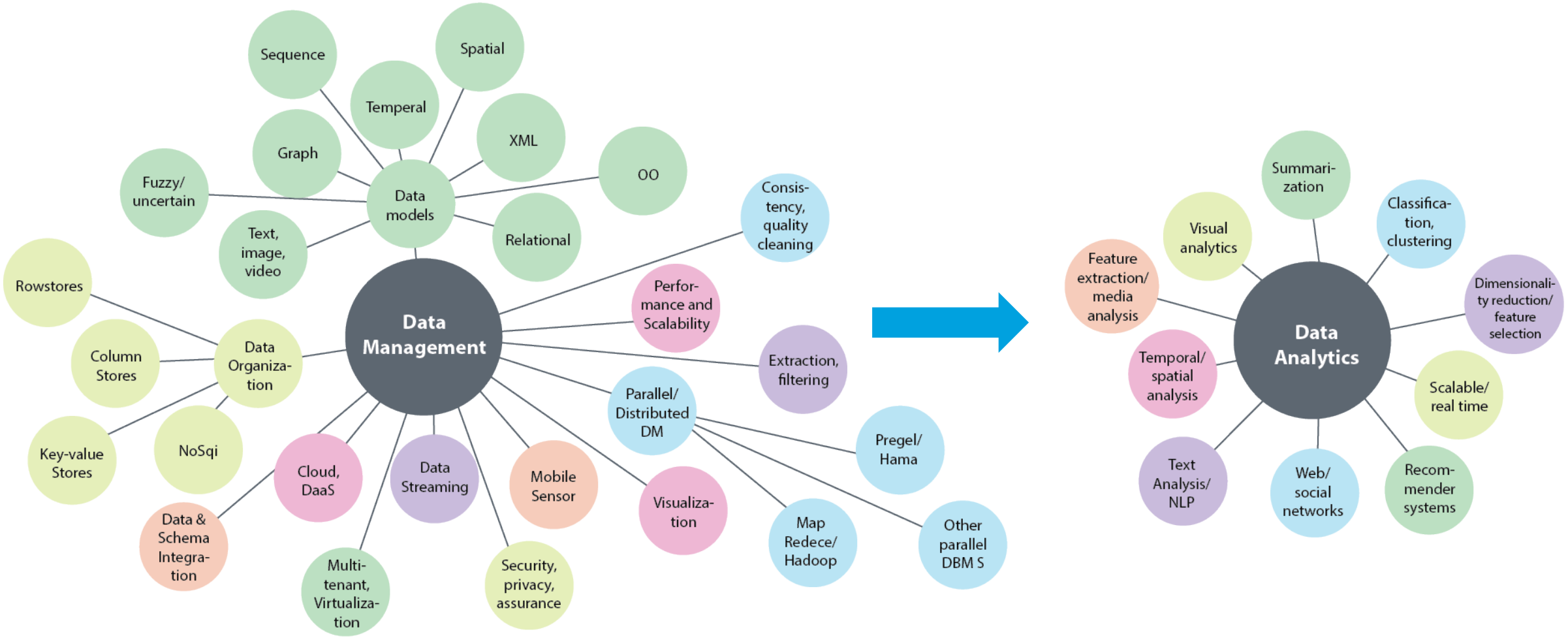
(**V**)olume
(**V**)Elocity
(**V**)ariety

HMLE

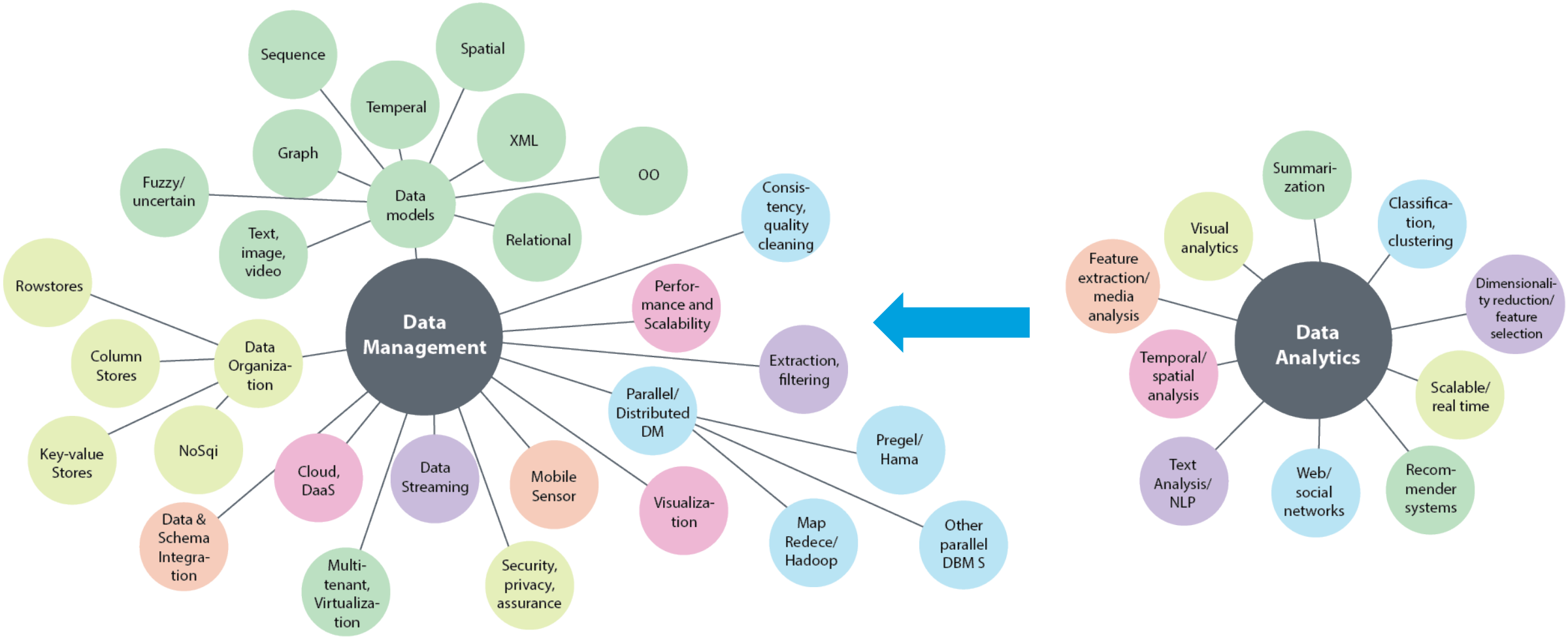
(**H**)igh-dimensional
(**M**)ulti-modal
Inter-(**L**)inked
(**E**)volving

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

Introduction to Data Exploration

Data Organization

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Explain data, data models, and data organization

What is Data?

da·ta

/ˈdadə, ˈdādə/

noun
noun: data

facts and statistics collected together for reference or analysis.
synonyms: facts, figures, statistics, details, particulars, specifics; More

- COMPUTING
  the quantities, characters, or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- PHILOSOPHY
  things known or assumed as facts, making the basis of reasoning or calculation.

How is Data Organized?

What is a database?

- Collection of data, 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

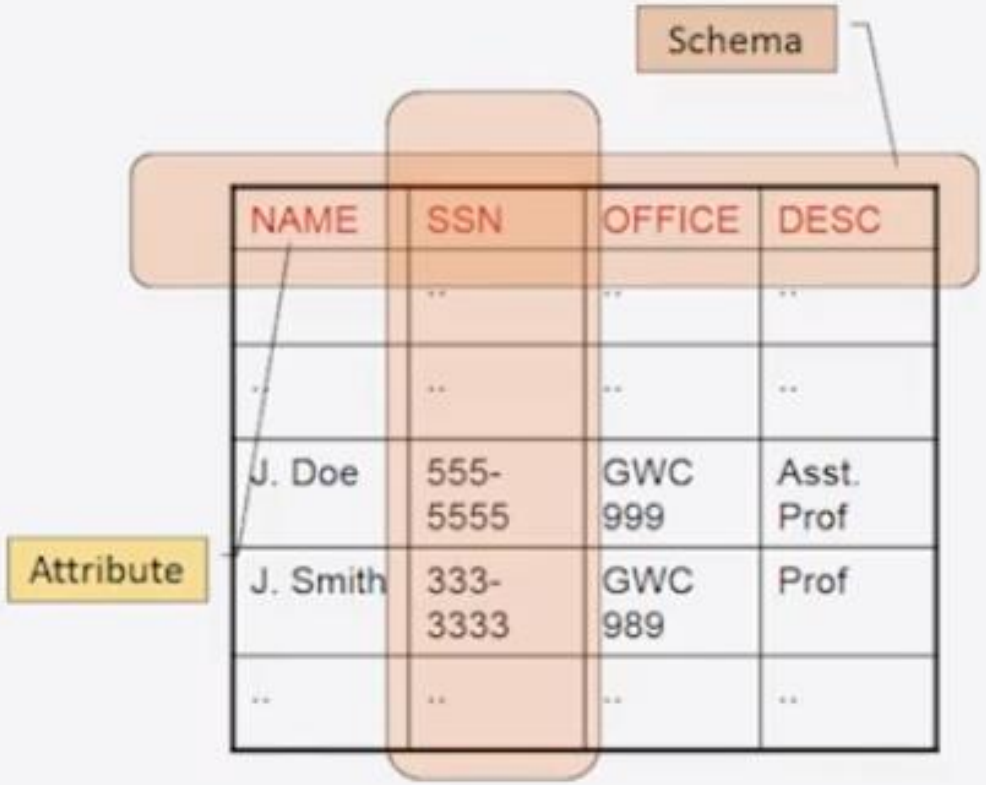
- Structured Data/Databases
- Semi-Structures Data/Databases
- Unstructured Data/Databases

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

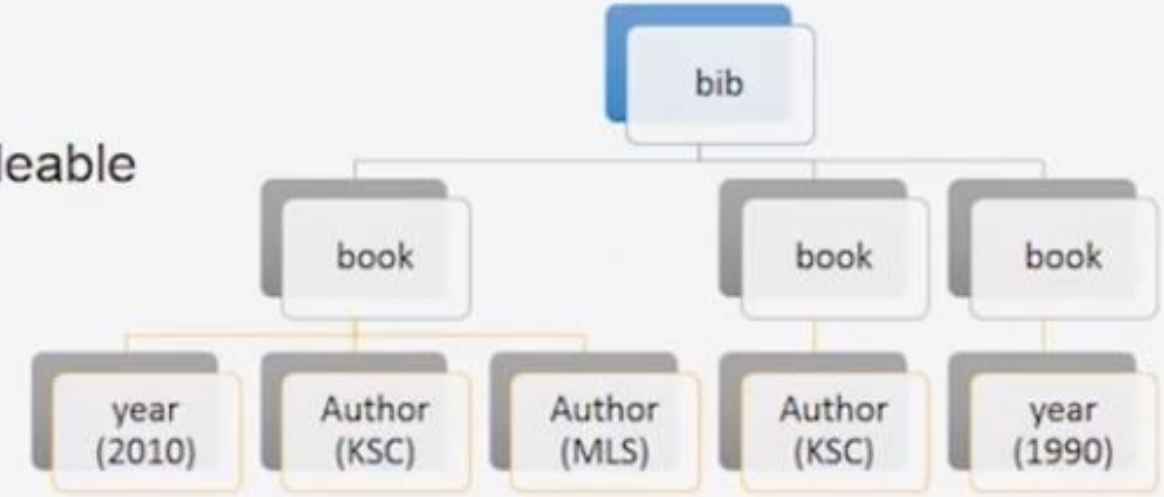
- **Informally, data is organized in tabular form**
  - Example: Data about an employee
- **Schema for each table consists of attributes**
  - Each attribute has a domain
- **Functional dependencies**
  - Describe the relationships among the attributes in the schema
  - Example: A key uniquely identifies 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 (multivalued attributes)
  - Data is self-describing: Each item in the database describes its own schema

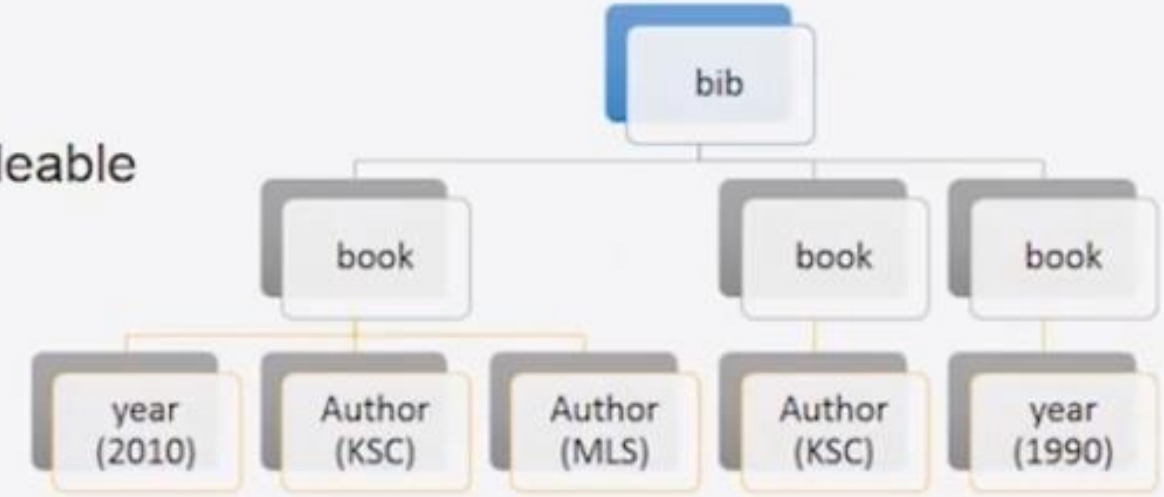
- **Advantages**
  - Data organization is flexible/malleable
    - easier to integrate
    - easier to exchange



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 (multivalued 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



Introduction to Data Exploration

Vector Data

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Understand vector representation of data

Common Data Representations

Relational/Object Oriented data

Vector Space (spatial or high-dimensional) data

Strings, sequences, and time series data

Trees and graphs

Fuzzy and probabilistic data

Vector Data

Images

Videos

Social Networks

Books

Sensor Reading

Vector Data

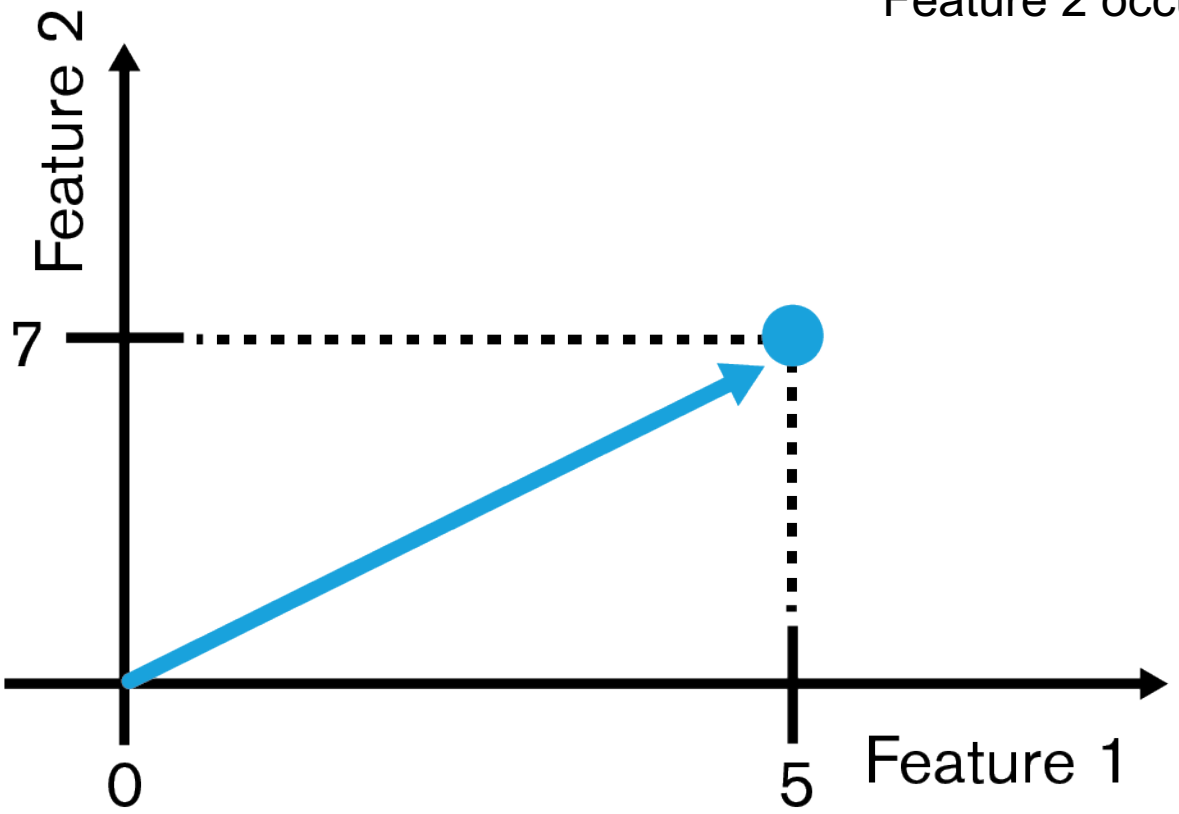
Images	Videos	Social Networks	Books	Sensor Reading
				
• colors • textures • shapes	• actors • ratings • directors	• connections • likes • interactions	• words • authors • publishers	• sensor values • patterns

How are counts represented?

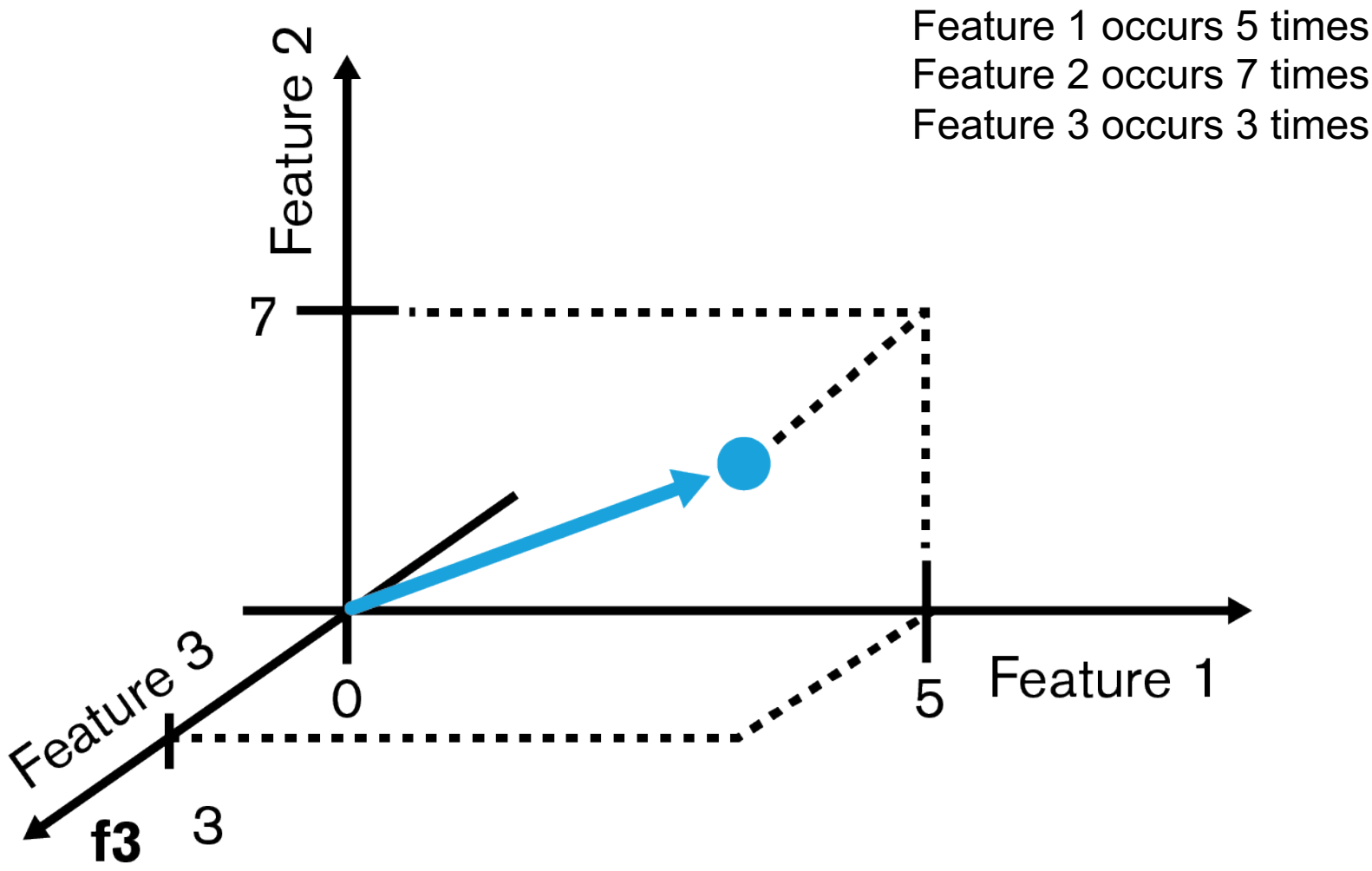
Feature 1 occurs 5 times

How are counts represented?

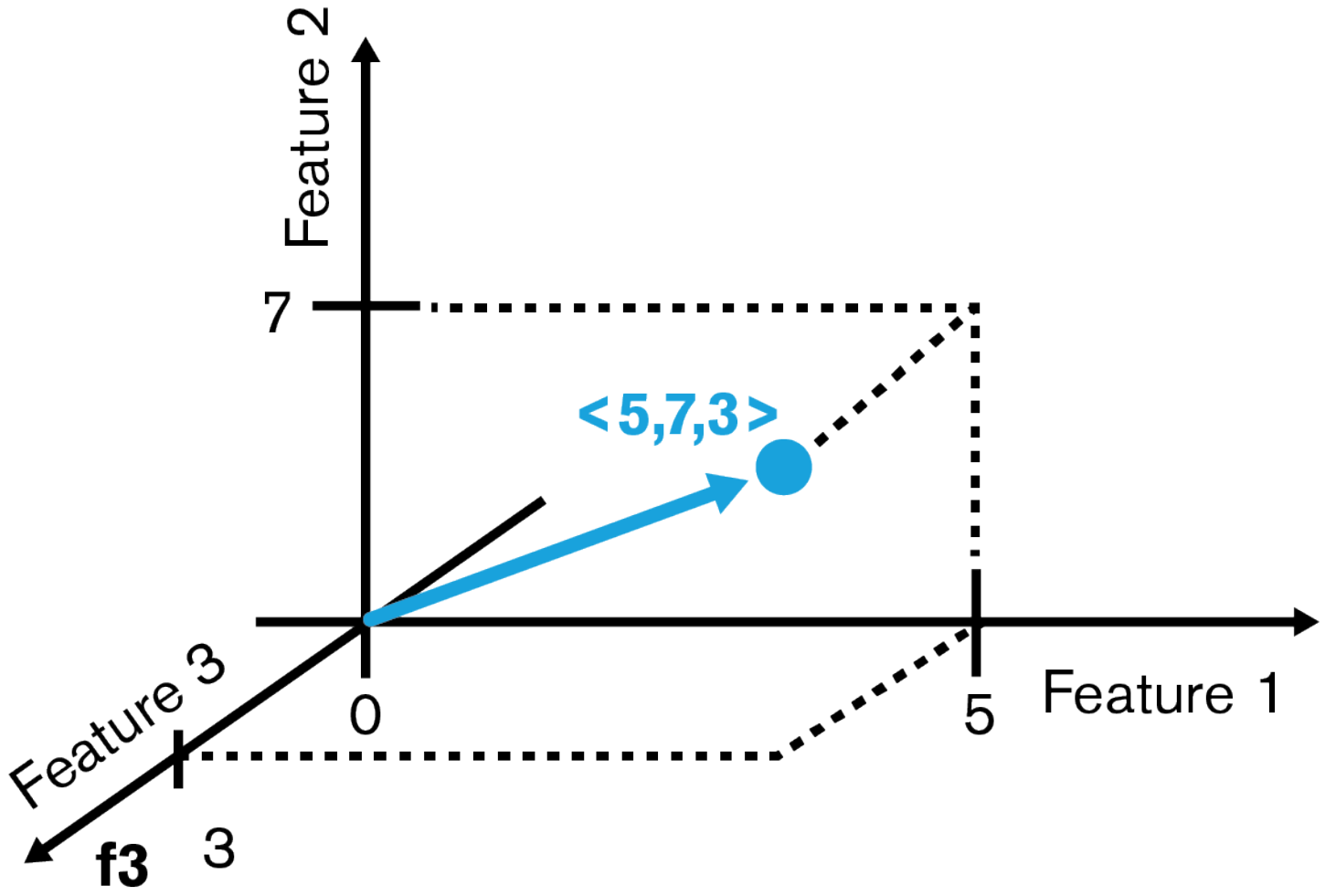
Feature 1 occurs 5 times
Feature 2 occurs 7 times



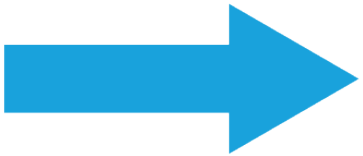
How are counts represented?

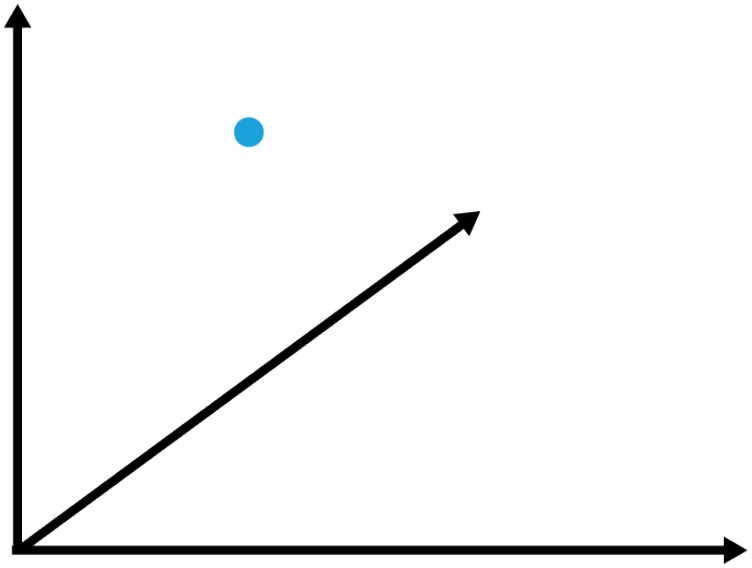


Vector Representation

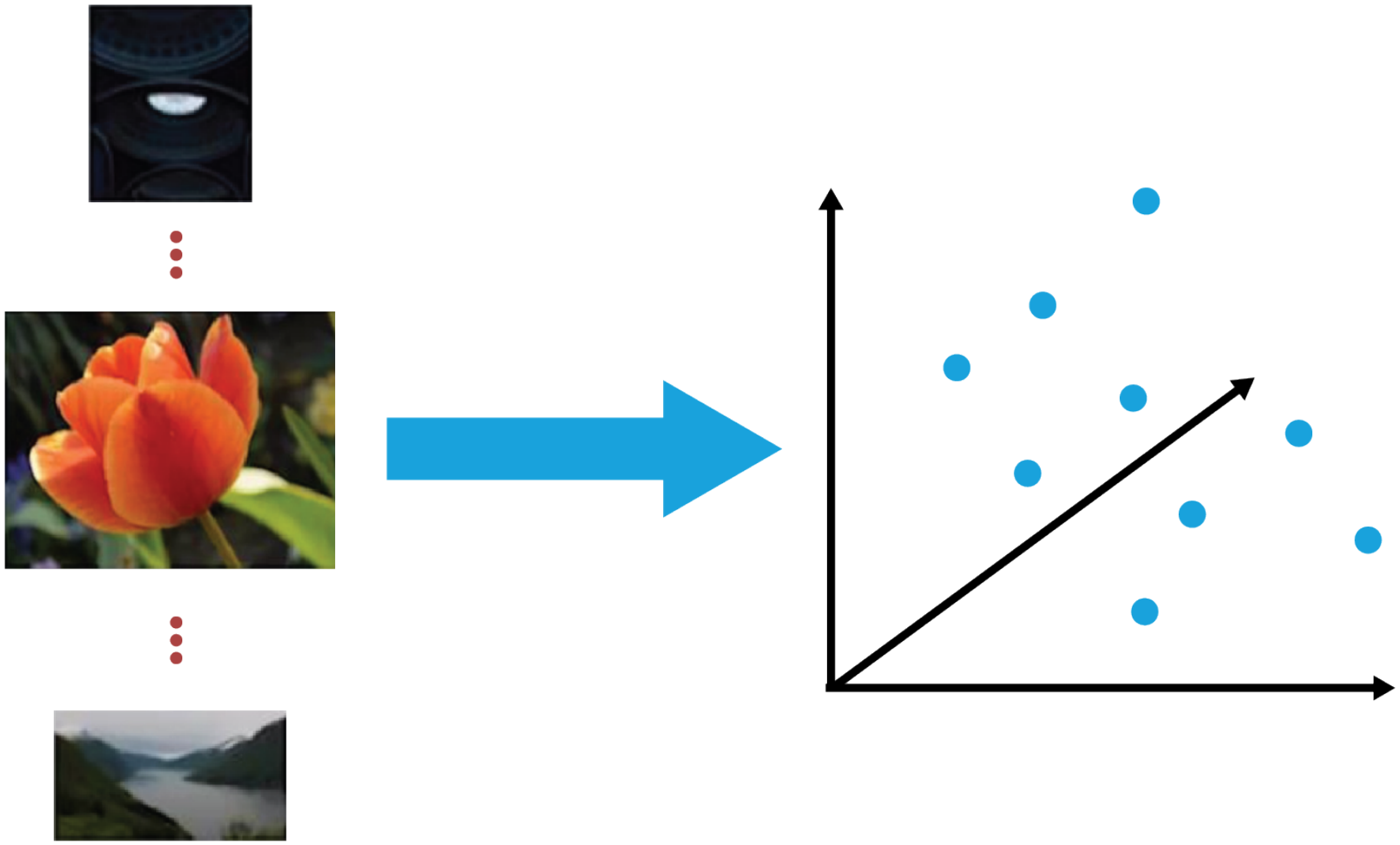


Vector Representation of a Given Object

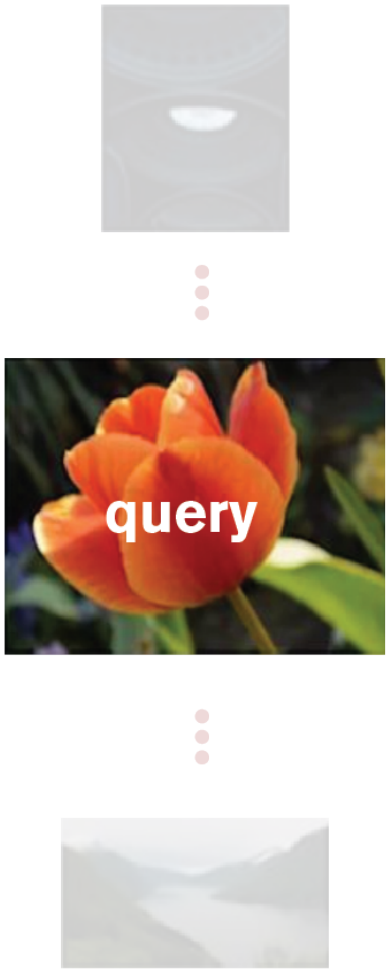


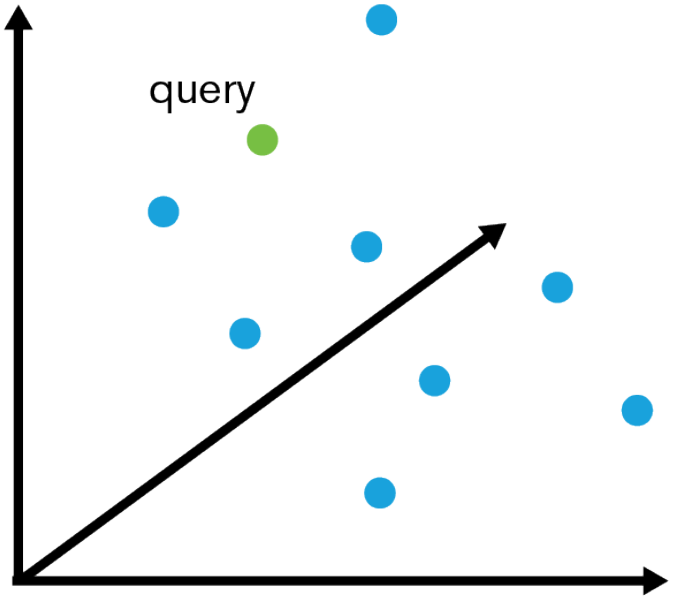


Vector Representation

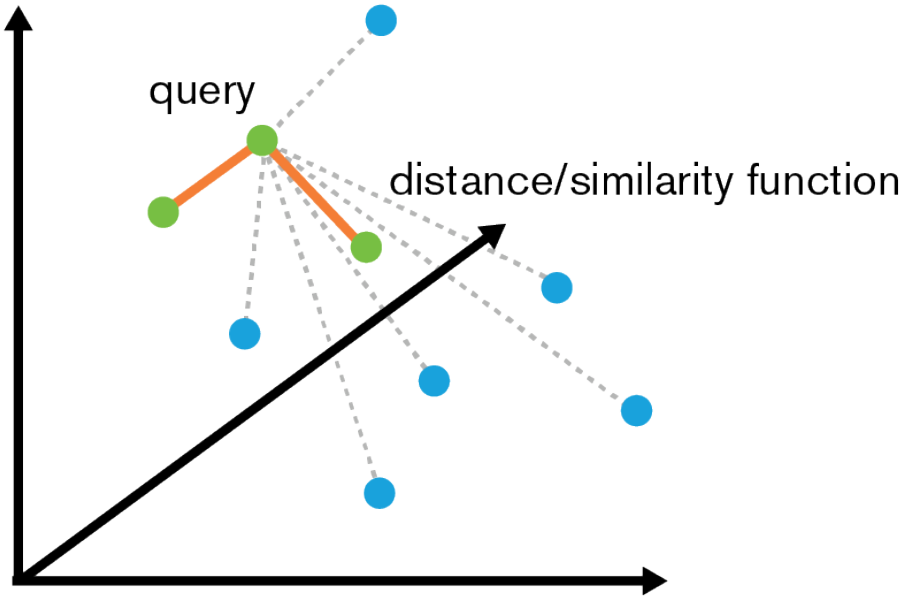


Querying

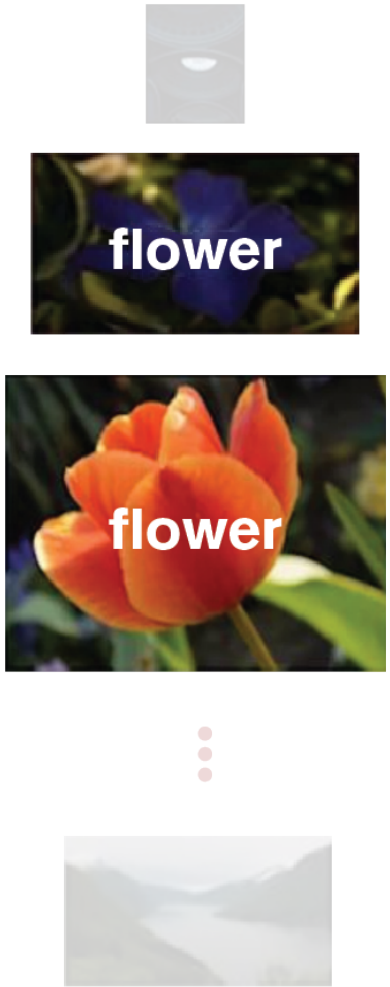


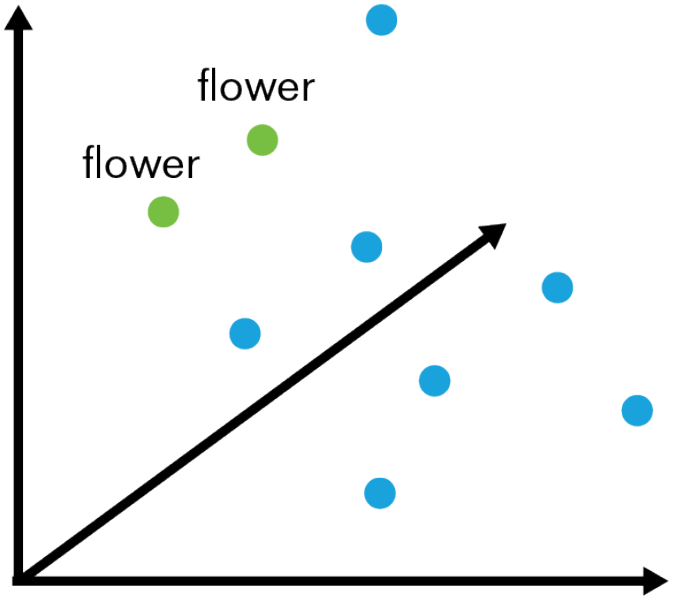


Similarity based (nearest neighbor) search

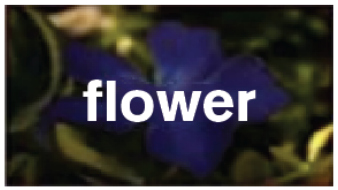


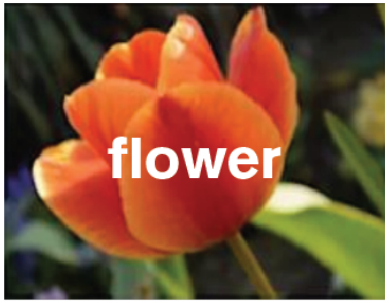
Classification

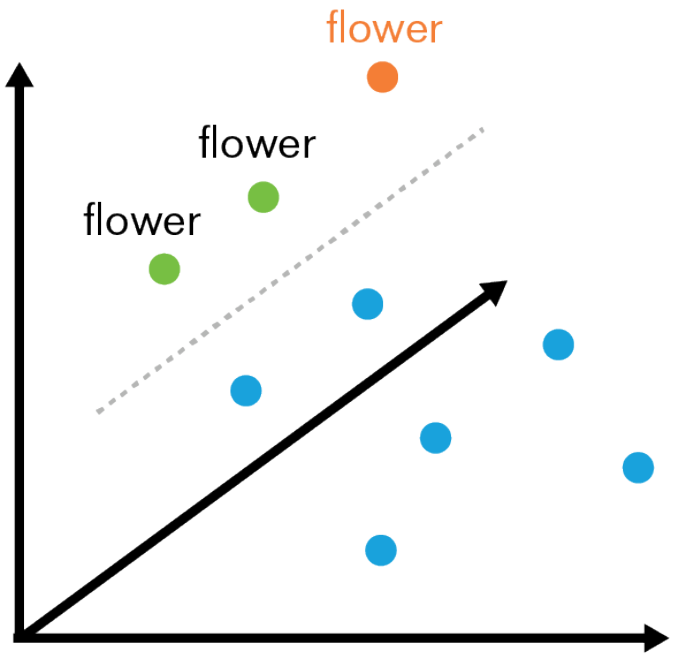




Classification







Introduction to Data Exploration

Vector Spaces

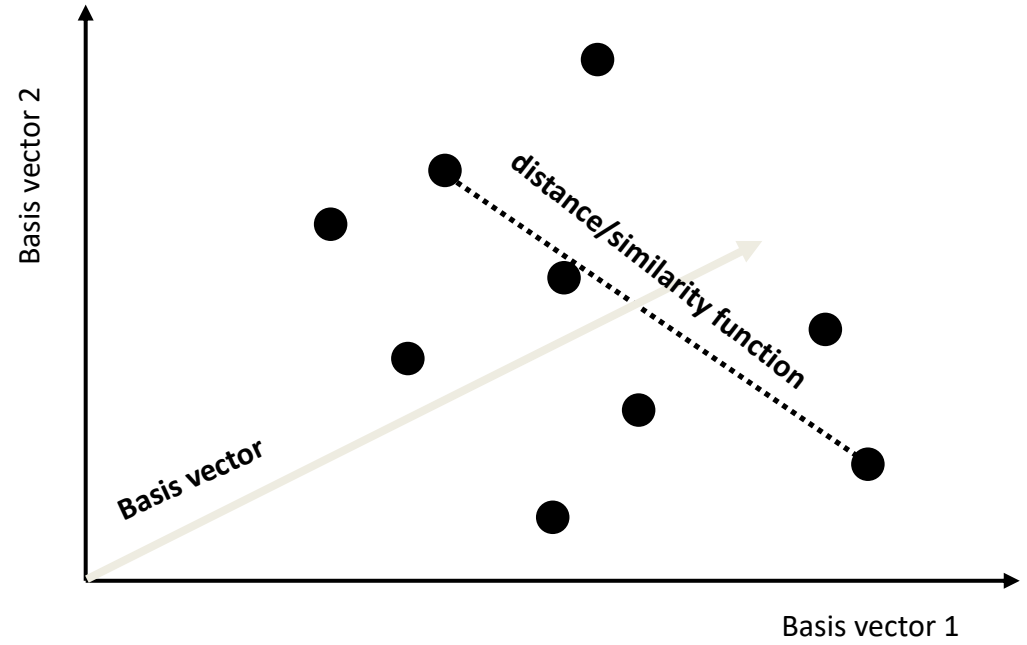
Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Identify the data types of elements within a dataset's data

Vector Spaces



- What are good features to use as basis vectors?

- How many features do we need as basis vectors?

Vector Spaces

Definition 3.1.1 (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.*

Basics of a Vector Space

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*

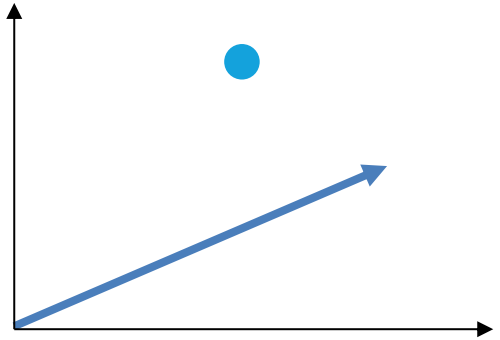
$$\vec{u} = \sum_{i=1}^{n} c_i \vec{v}_i.$$

complete

How many features we need?

…just 1?

…or 3?



…maybe 2?

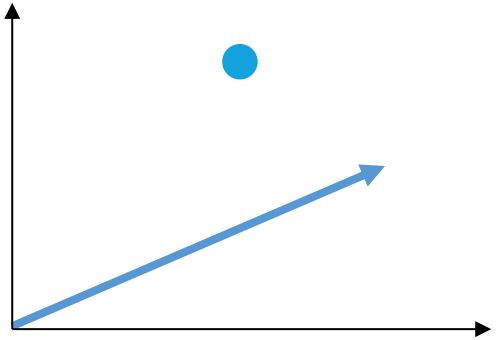
…or, many many more (100s, 1000s) ???

How many features we need?

…just 1?

…maybe 2?

…or 3?



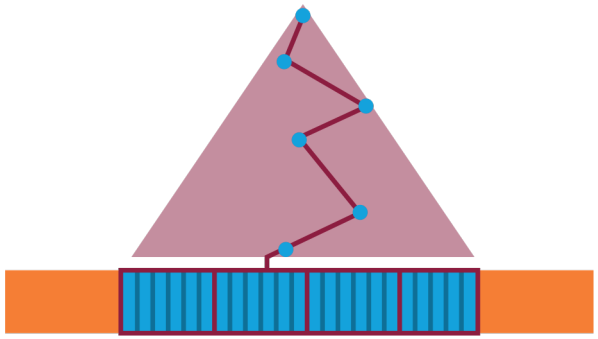
…or, many many more (100s, 1000s) ???

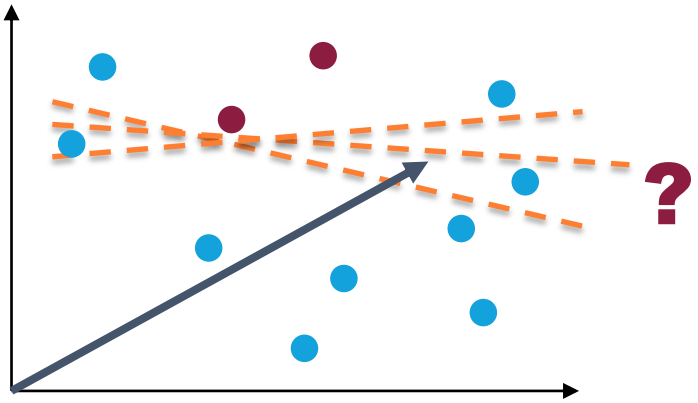
Dimensionality curse!!!

Dimensionality curse: The more dimensions we have, the less efficient and effective search and analysis becomes

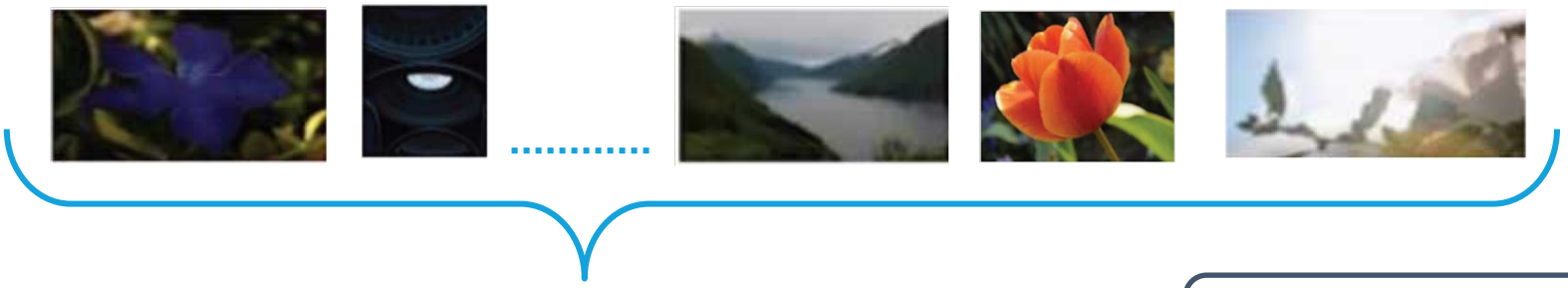
Efficiency: Search data structures are not very efficient at high dimensions

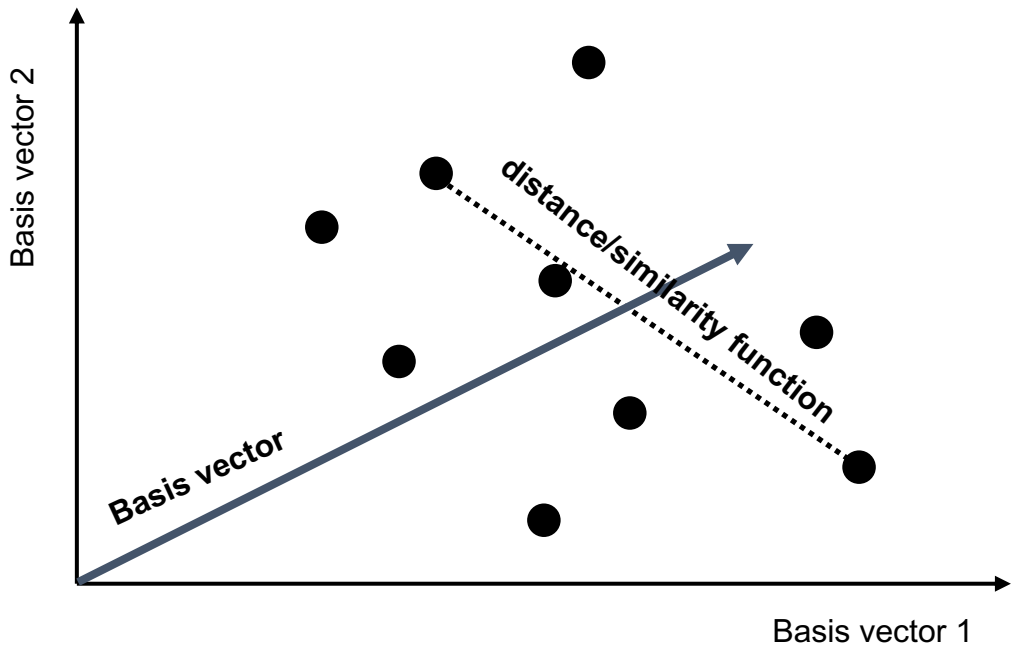
Effectiveness: The more dimensions we have, the more data we need to discover patterns (prevent overfitting)





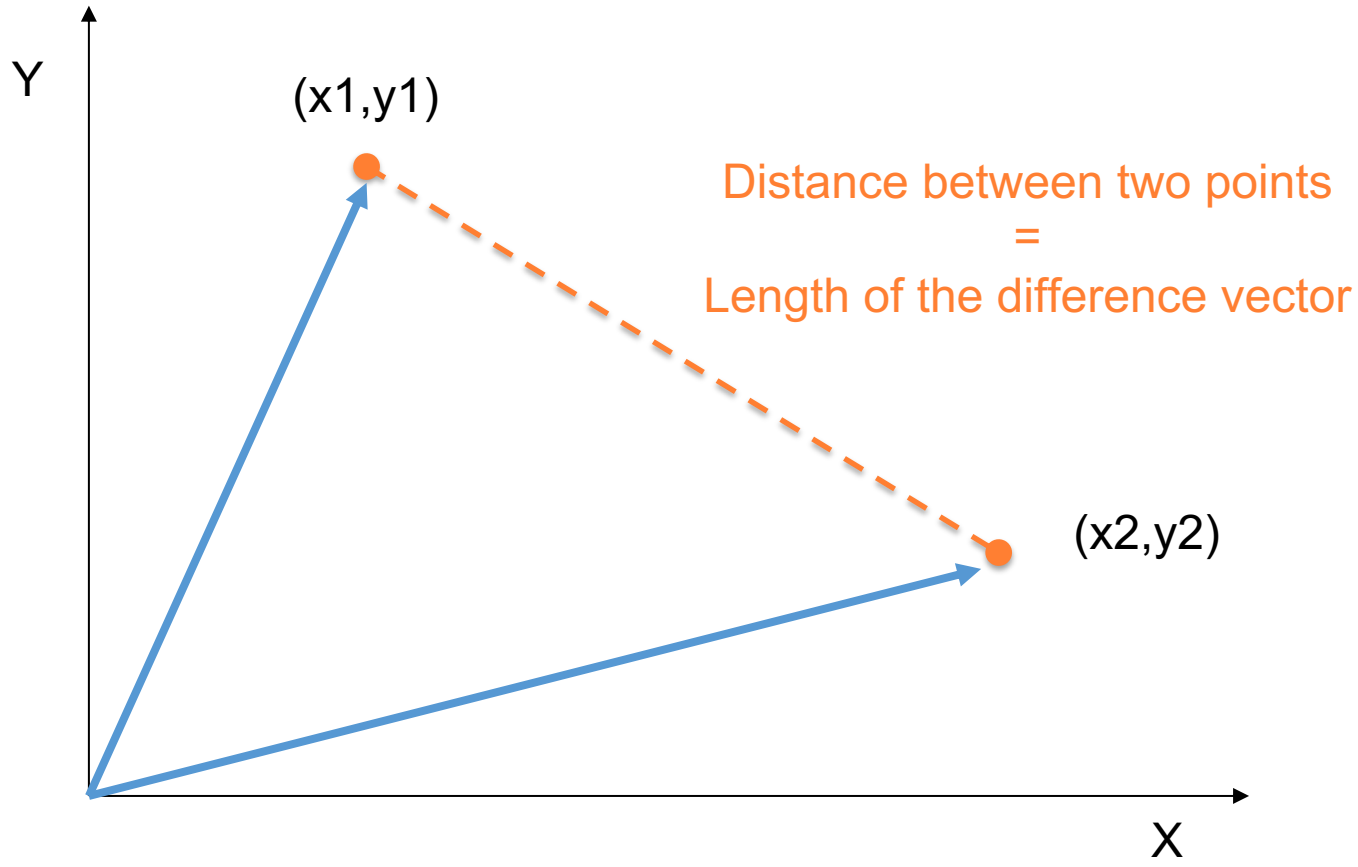
Vector Spaces





- What are good features to use as basis vectors?
- How many features do we need as basis vectors?
- What is a good distance/similarity function?

Distance

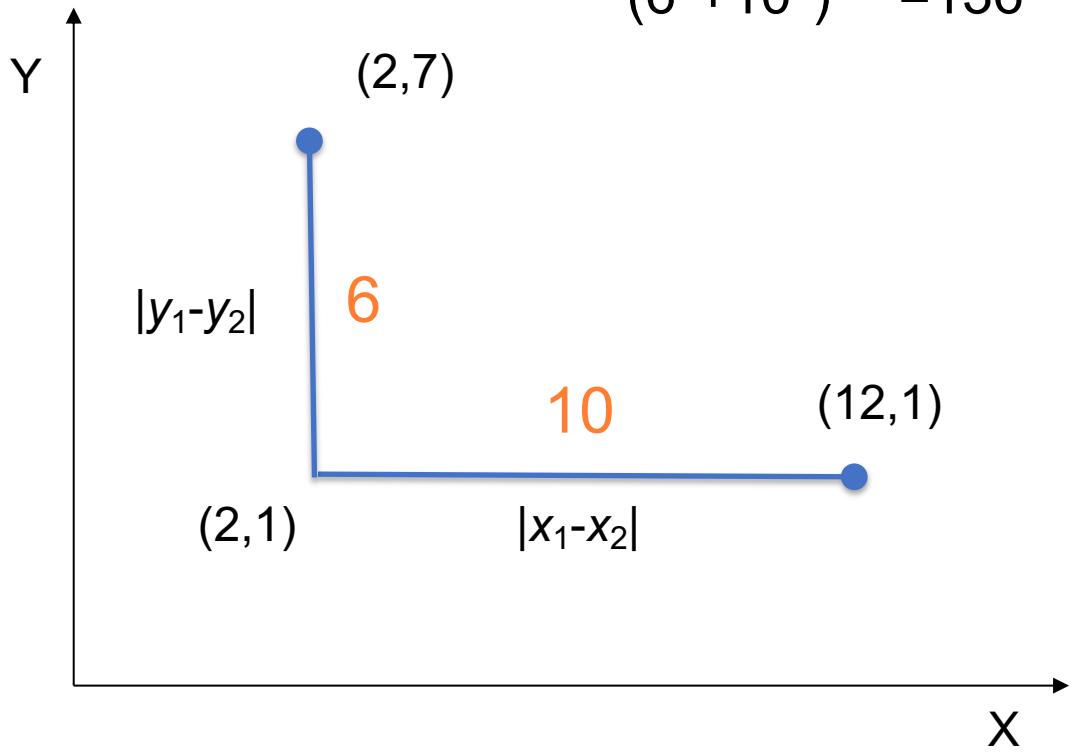

Y
(x1,y1)
Distance between two points
=
Length of the difference vector
(x2,y2)
X

P-Norms

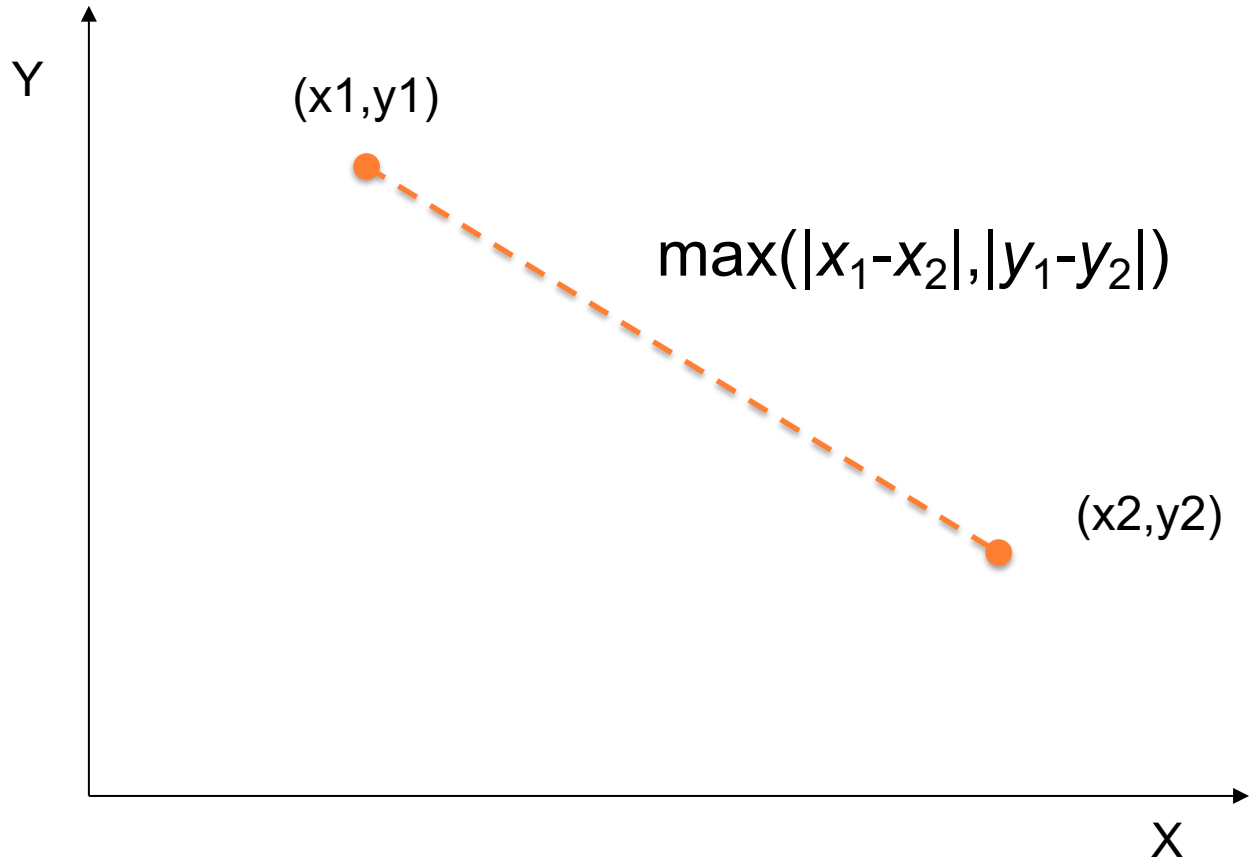
- 1-norm
- 2-norm

(6+10)=16

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



∞-norm (L∞ distance)

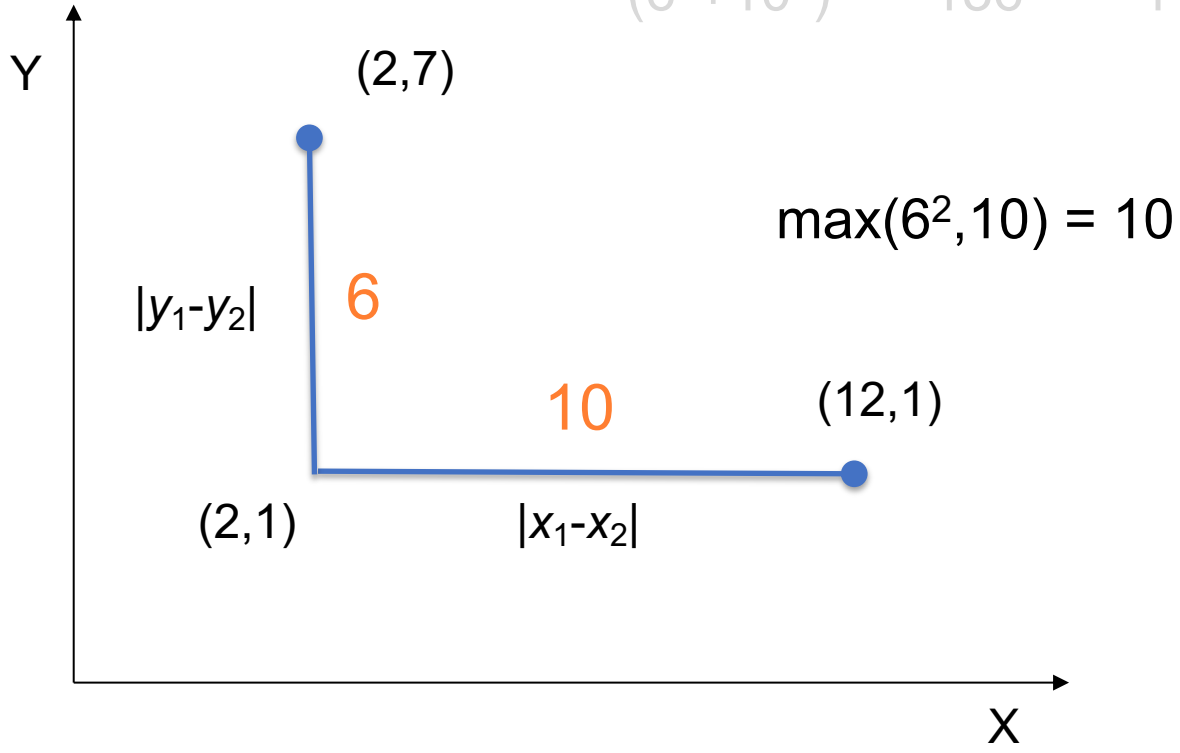


P-Norms

- 1-norm
- 2-norm
- …
- ∞-norm

$(6+10)=16$

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

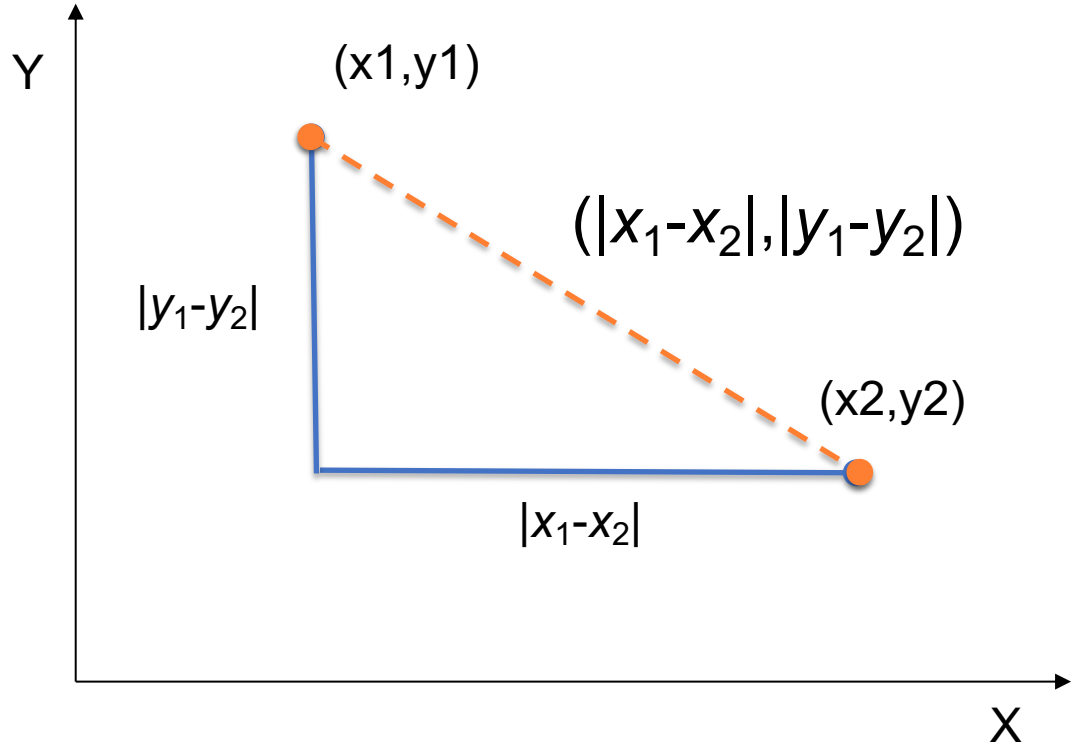


Norms

- Most commonly used family of length measurements are the p-norms

$$\| \vec{\mathbf{v}} \|_{\mathrm{p}} = \left(\sum_{i=1}^{n} |w_i|^p \right)^{\frac{1}{\mathrm{p}}}$$

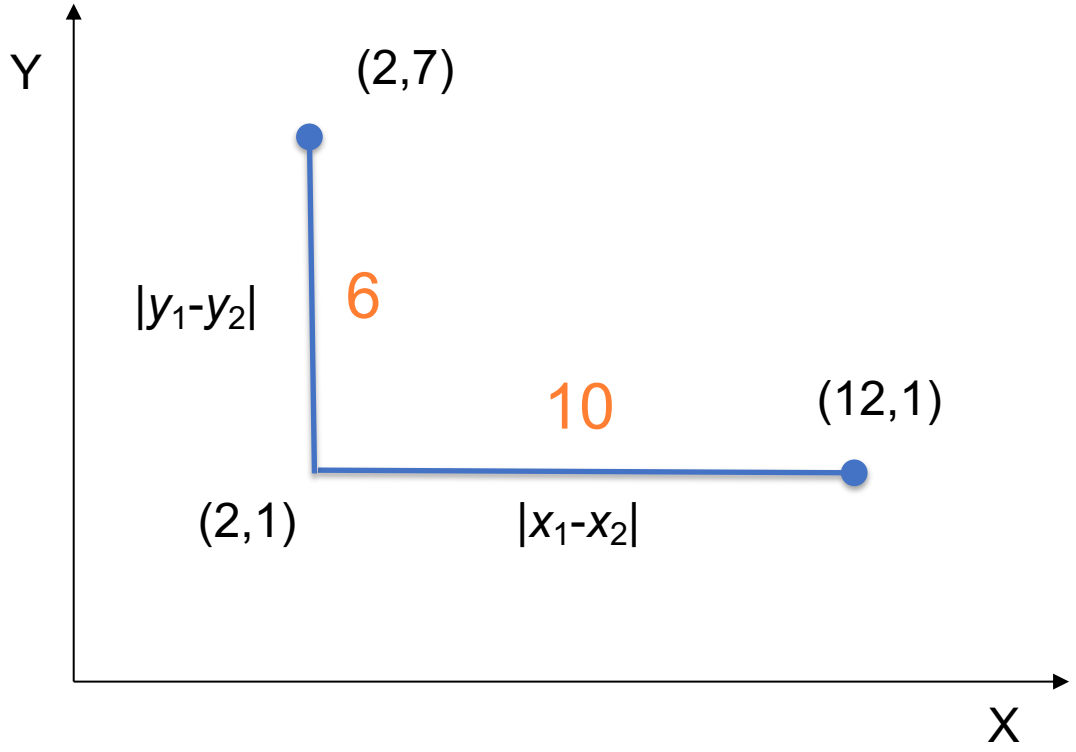
1-norm (Manhattan Distance, L1 distance)



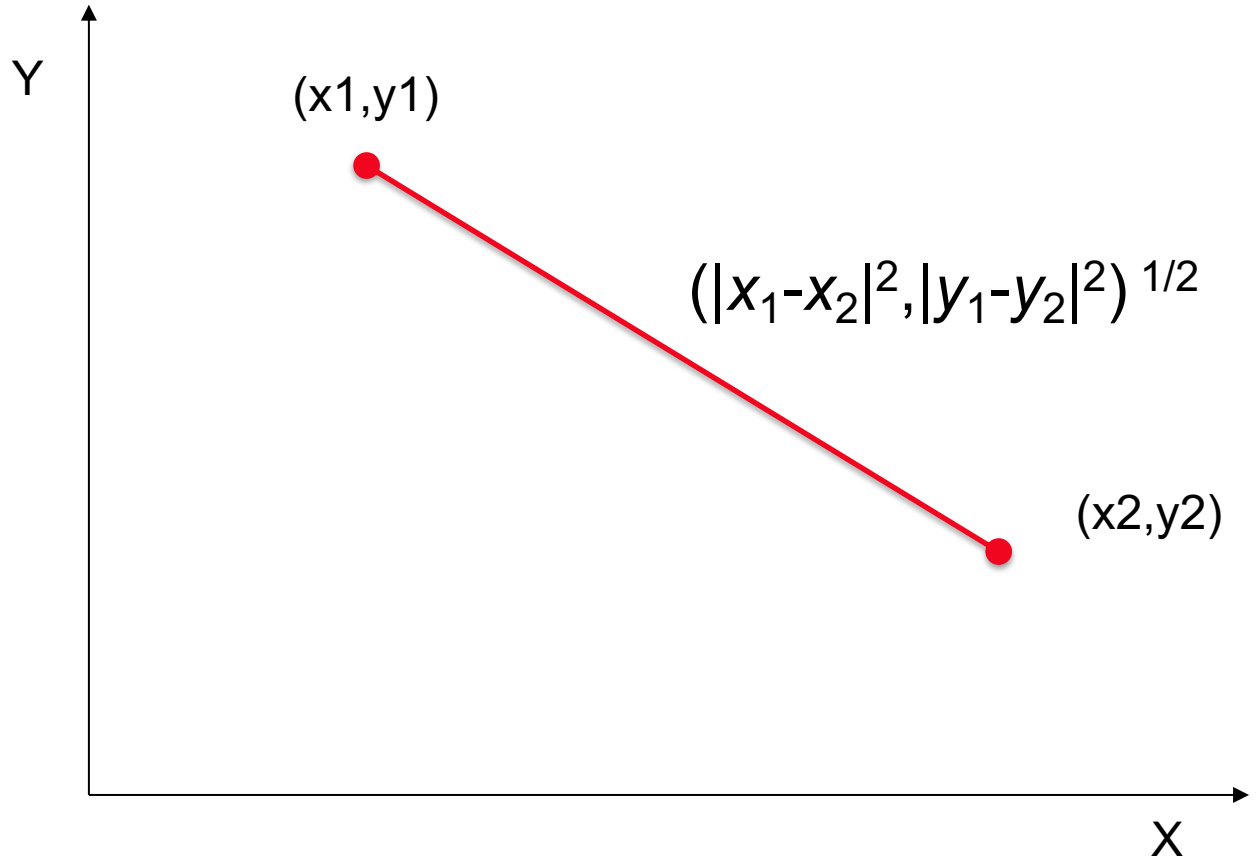
P-Norms

- 1-norm

(6+10)=16



2-norm (Euclidean Distance, L2 distance)

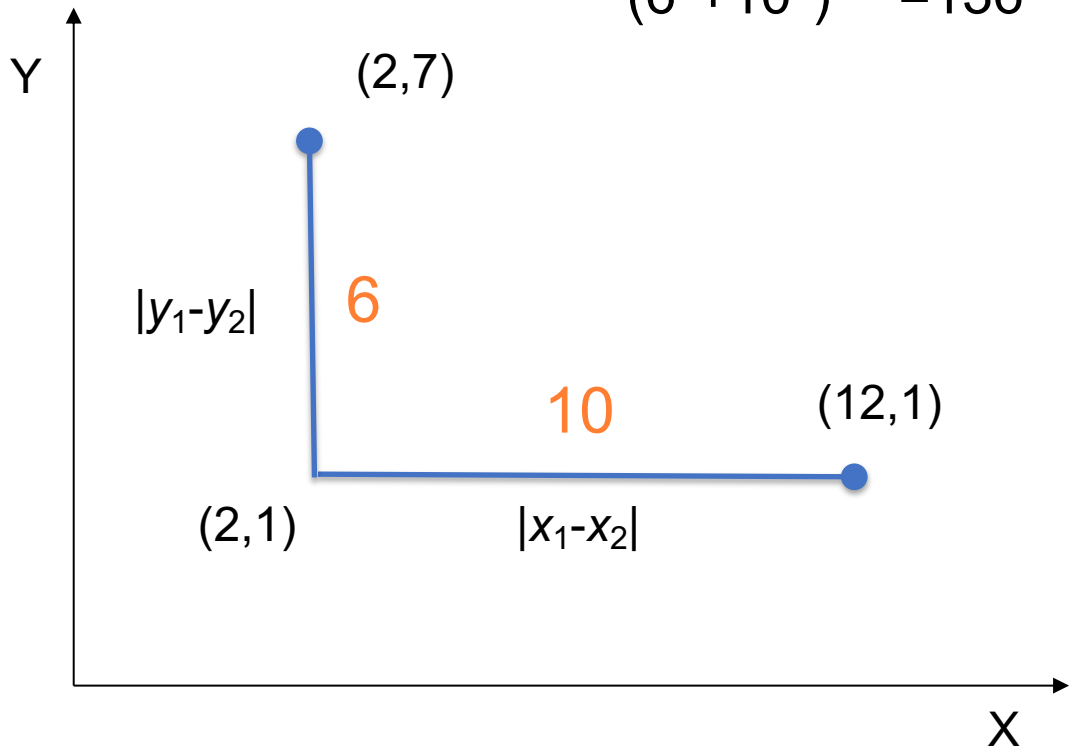


P-Norms

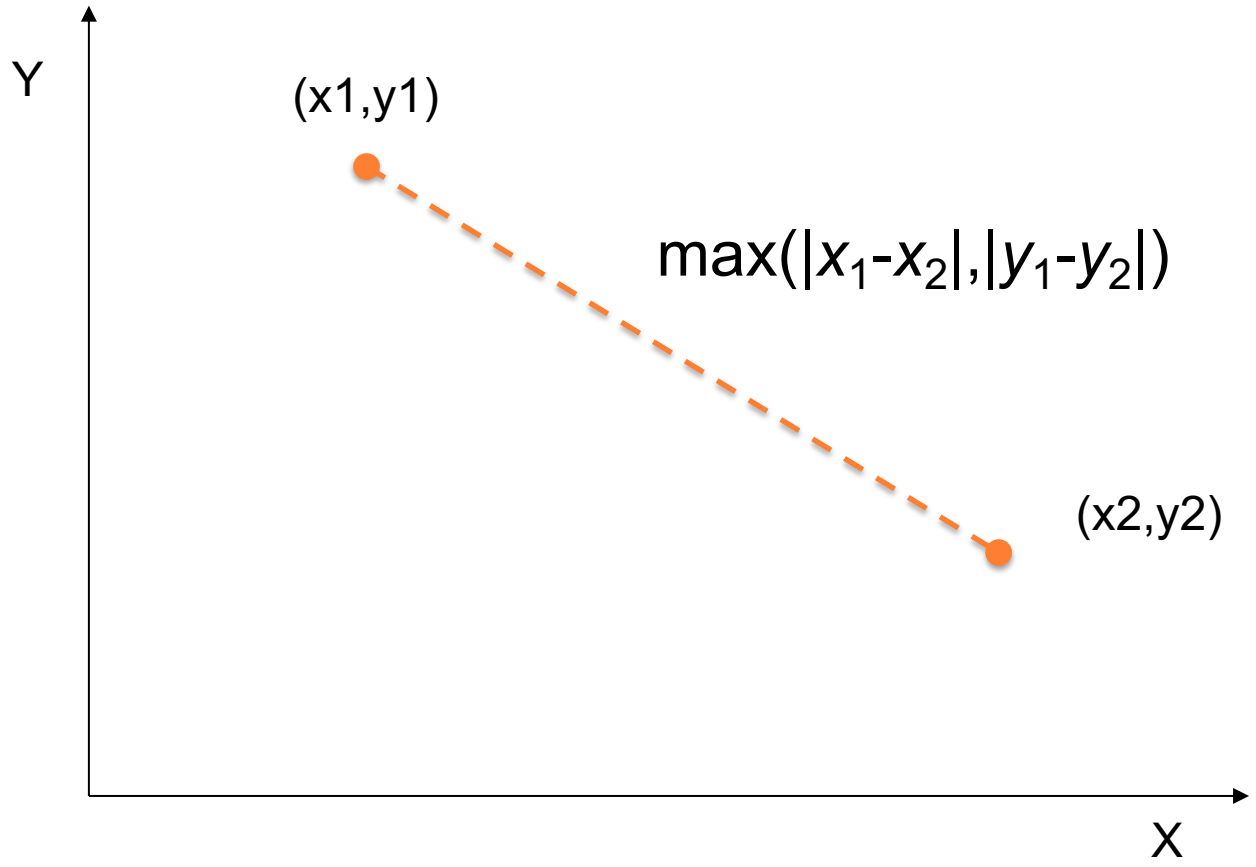
- 1-norm
- 2-norm

(6+10)=16

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



∞-norm (L∞ distance)

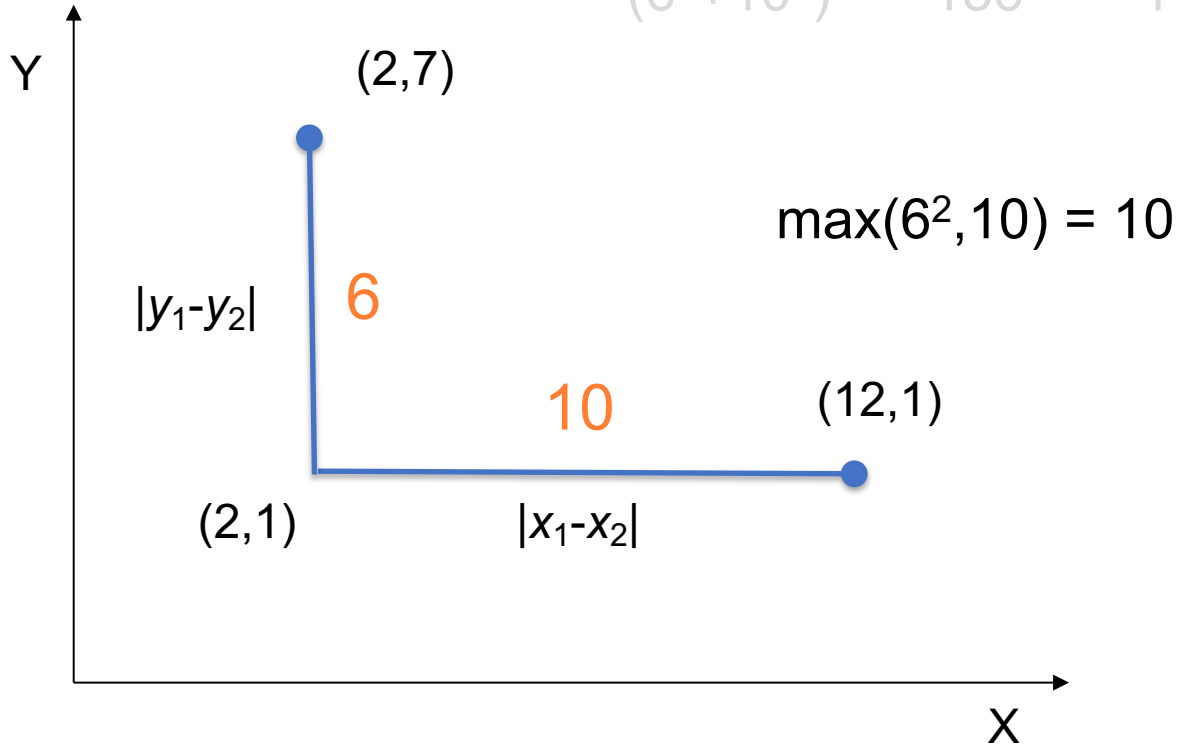


P-Norms

- 1-norm
- 2-norm
- …
- ∞-norm

$(6+10)=16$

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



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-minimality: $\Delta(s,s) = 0$
- minimality $\Delta(s_1,s_2) >= \Delta(s_1,s_1)$
- symmetry $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
- triangular inequality $\Delta(s_1,s_2) + \Delta(s_2,s_3) >= \Delta(s_1,s_3)$

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 a no other object can match the given object better than itself

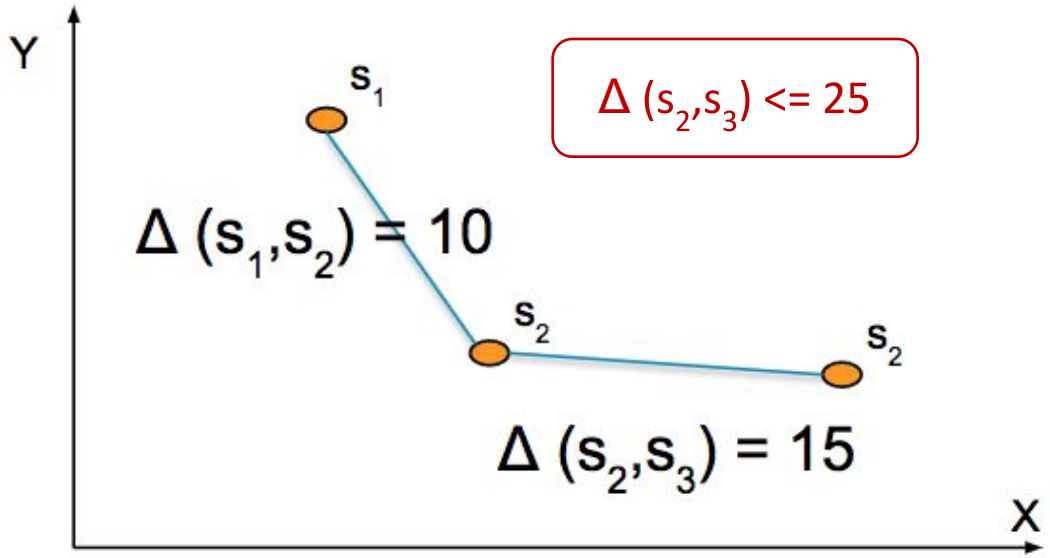
Symmetry

Symmetry:

- $\Delta(s_1, s_2) = \Delta(s_2, s_1)$
- Ensures that if a given object s_1 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 pruning of the search space during retrieval



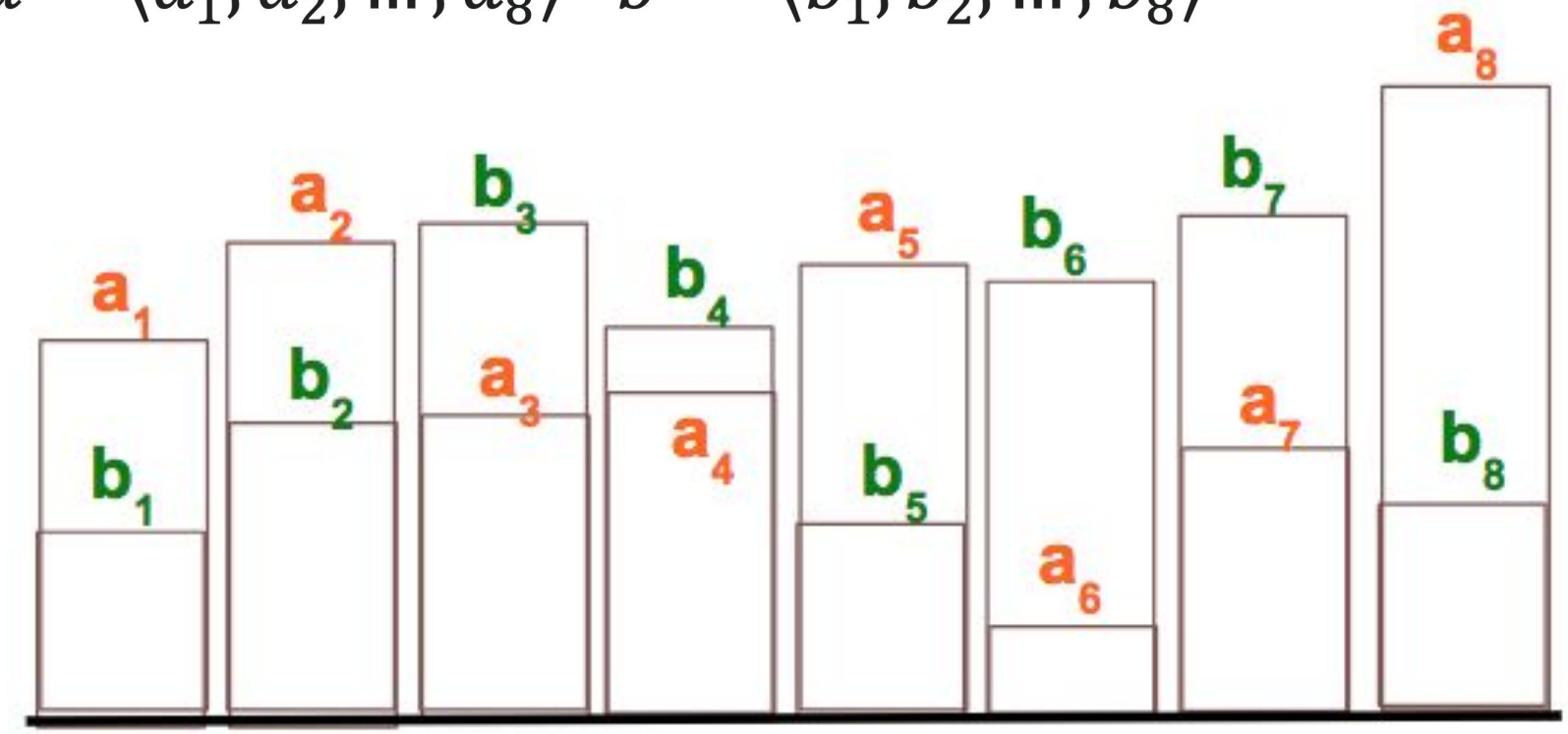
P-norms are Metric

- 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 $\max_{i=1..d} |v_{1,i} - v_{2,i}|$

Are there other Similarity Measures?

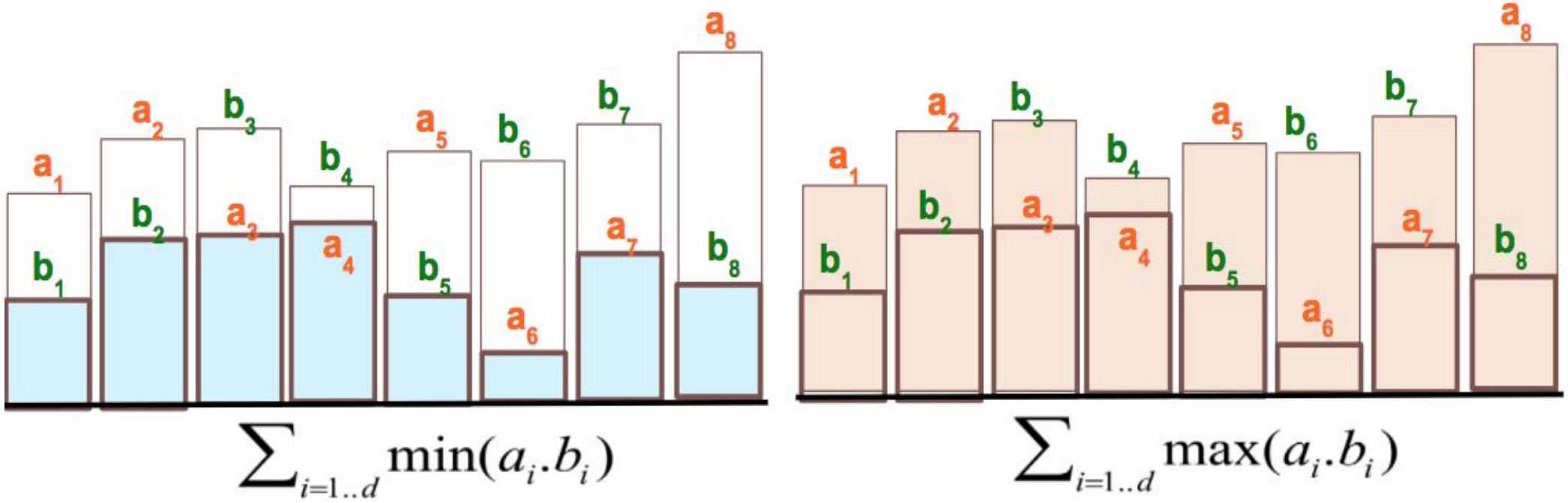
Consider two vectors

$$\vec{a} = \langle a_1, a_2, \ldots, a_8 \rangle \quad \vec{b} = \langle b_1, b_2, \ldots, b_8 \rangle$$



Intersection Similarity

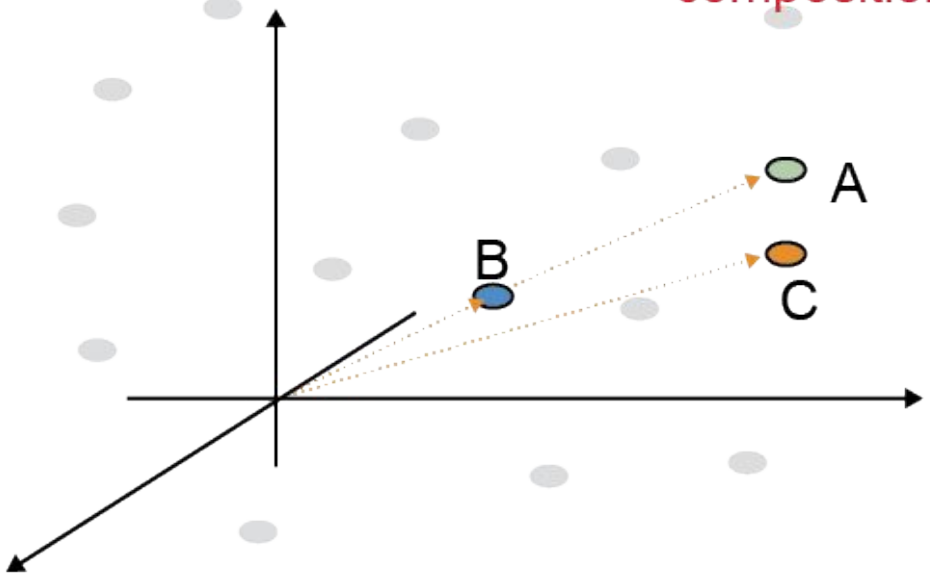
$$\mathrm{sim}_{\mathrm{int}}\left(\vec{a},\vec{b}\right)=\frac{\sum_{i=1..d}\min(a_i.b_i)}{\sum_{i=1..d}\max(a_i.b_i)}$$



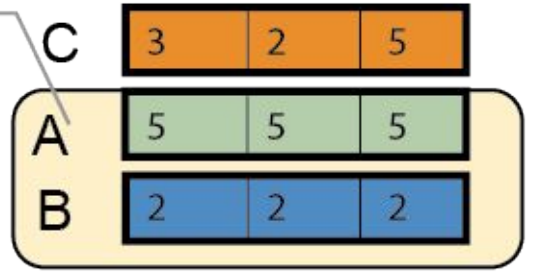
Angle based similarity measures

If we use the angle as a similarity measure, then A is more similar to E than F

$\cos(\widehat{AB}) > \cos(\widehat{AC})$



Similar composition



Angle-based Measures

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

- 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

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

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

Introduction to Data Exploration

Strings and Sequences

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Understand string, sequence, and time series data

Common Data Representations

Relational/Object Oriented data

Vector Space (spatial or high-dimensional) data

Strings, sequences, and time series data

Trees and graphs

Fuzzy and probabilistic data

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.

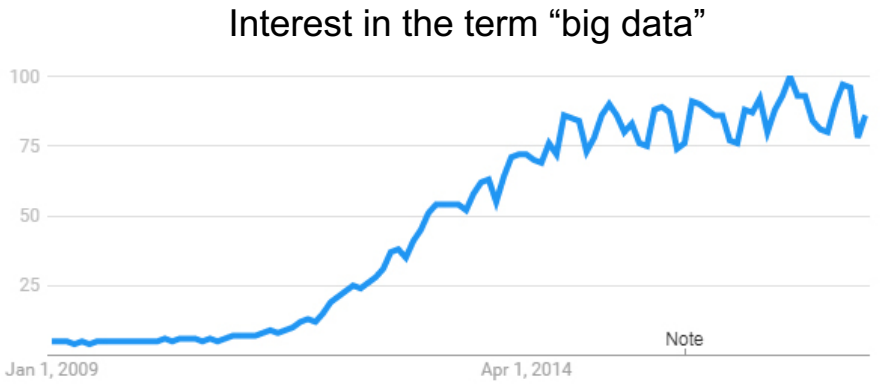
abcbbbaabbaabcbbbaaabbc

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.

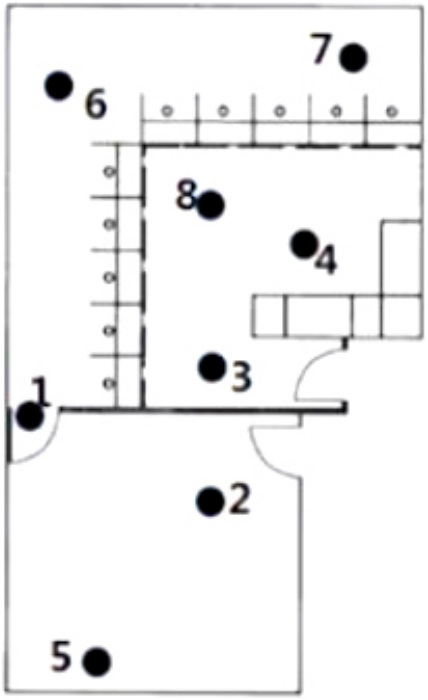
abcbbbaabbaabcbbbaaabbc

- A *time 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 $d_i \in R$

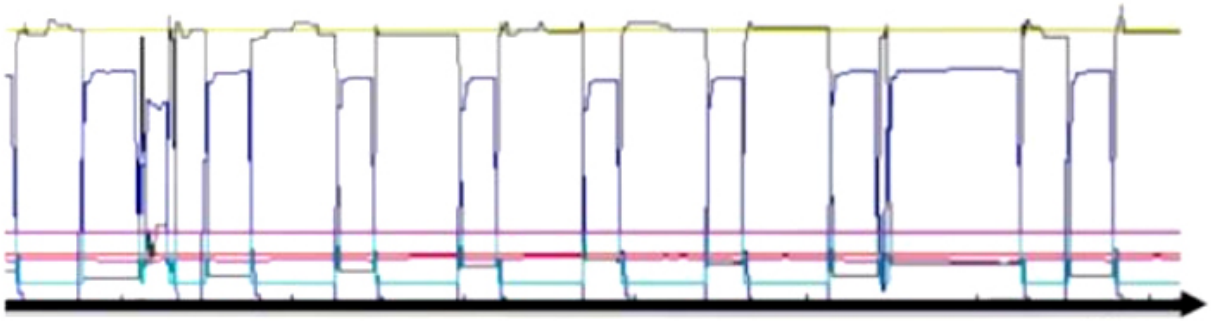


https://trends.google.com/trends/explore?date=2008-12-19%202018-01-19&q=big%20data

Multi-variate time series



Temperature recordings over a period of time



Time

Strings/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”;…

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

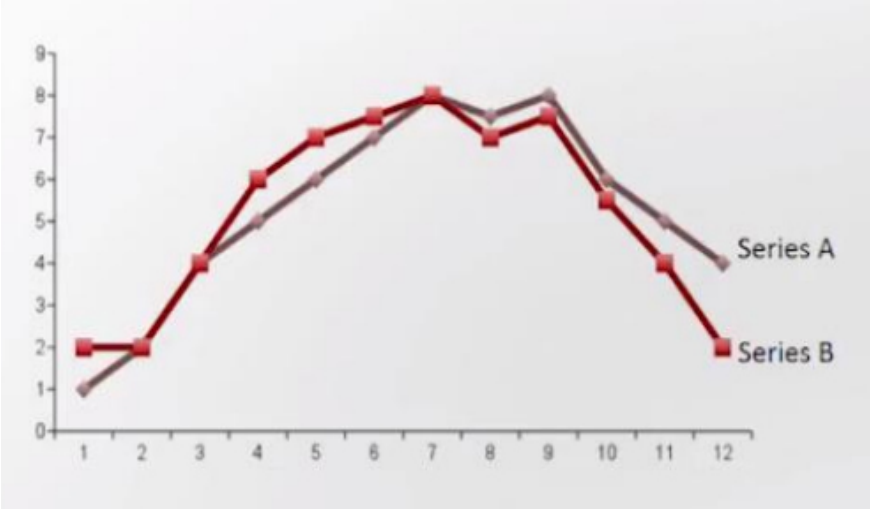
Edit distance between two sequences is the minimum number of edit operations needed to convert one sequence to the other:

- "**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")

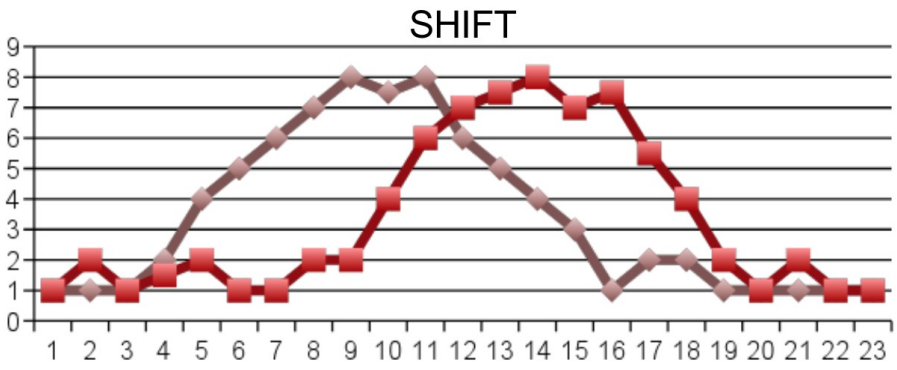
Time Series Matching

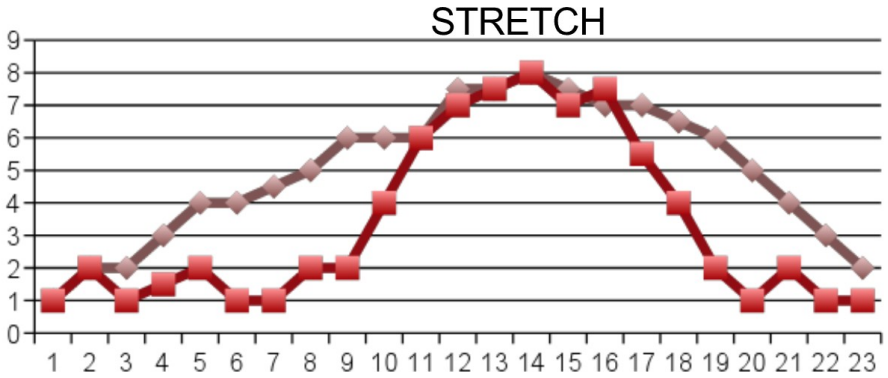
- **Synchronous/Non-Elastic Distance and Similarity Measures:**
  - Euclidean distance

$$\left(\sum_{i=1\ldots12} a_i^2 - b_i^2 \right)^{1/2}$$



Asynchrony in Time Series

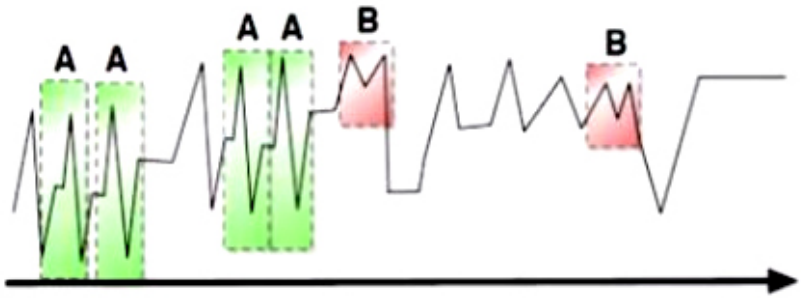


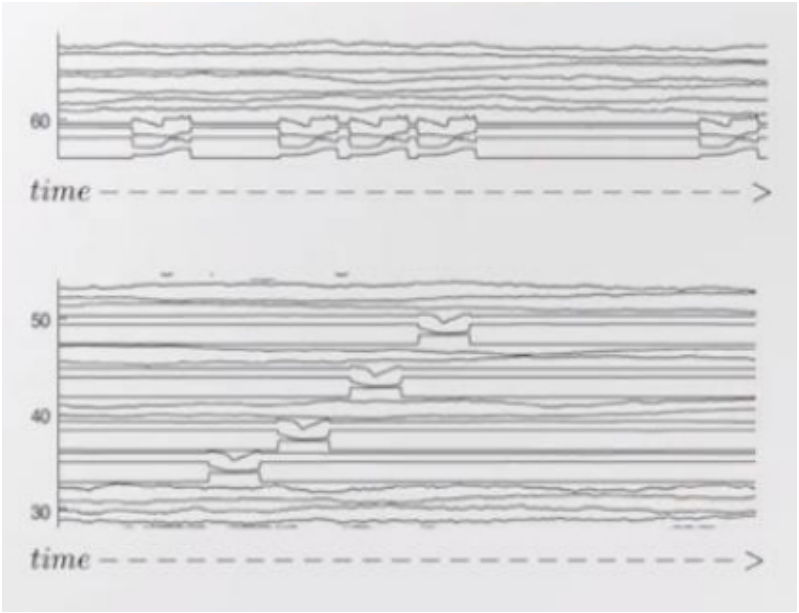


Asynchronous/Elastic Distance and Similarity Measures:

- 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

Introduction to Data Exploration

Data Processing vs. Querying vs. Exploration

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Explain modalities for data exploration

Exploratory Search

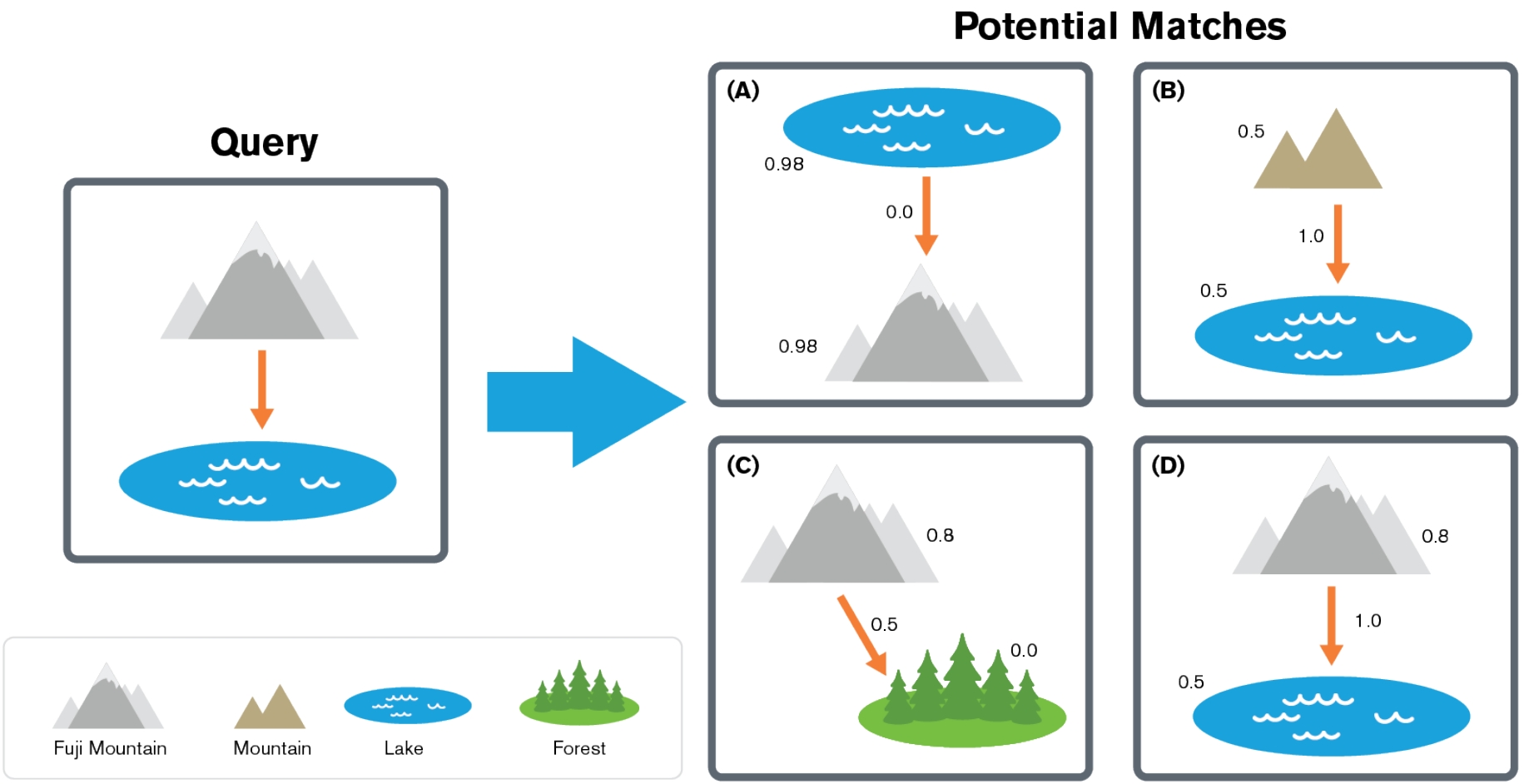
Acquiring new knowledge and revealing new facts

- Analysis (identify common patterns or outliers)
- Comparison (quantify similarity/differences)
- 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 (i.e., partial matches are allowed) each and every data element in the database is a potential match
- Hence the query results need to be ranked according to some objective or subjective criteria

Top-K Search

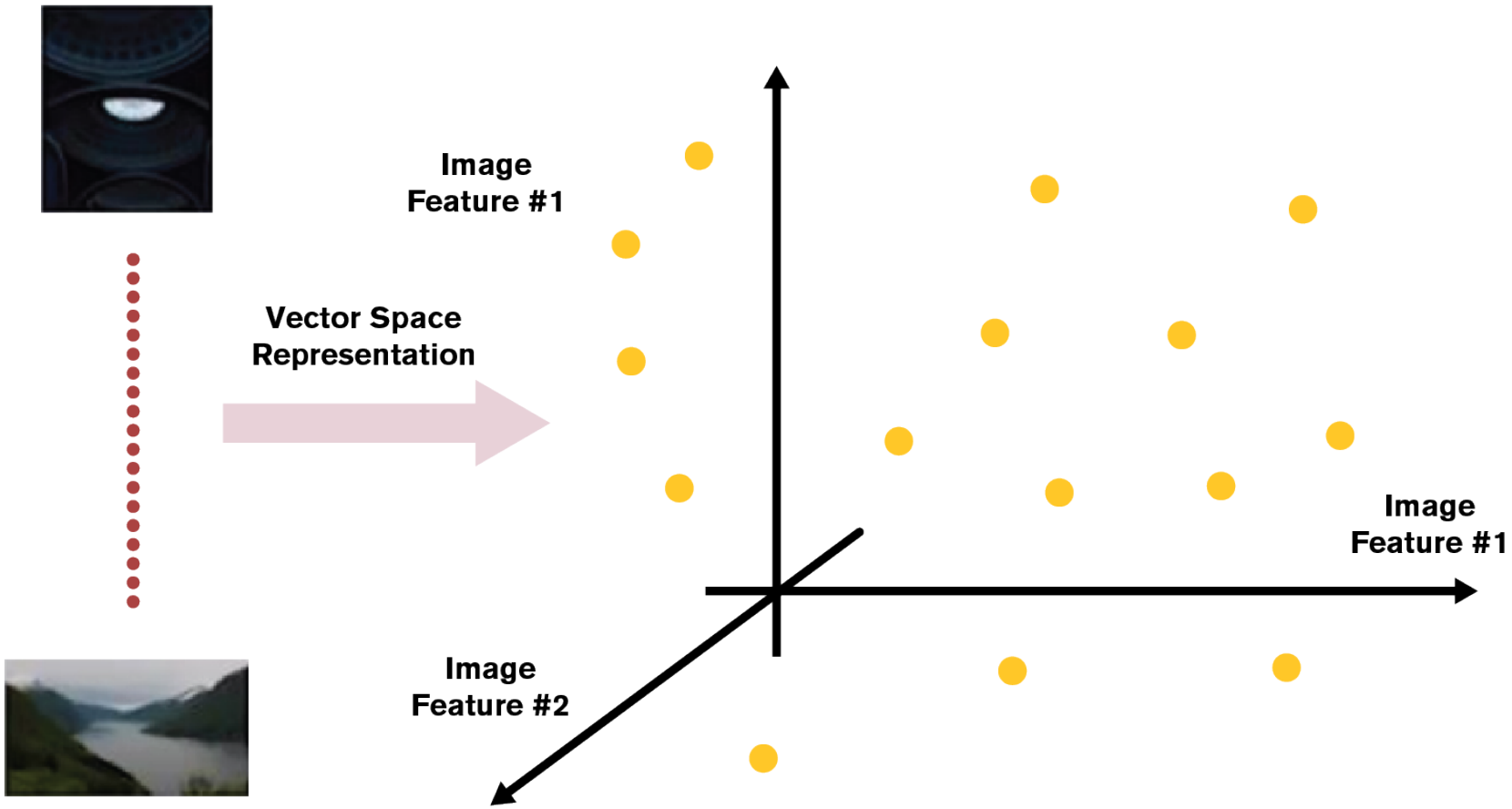
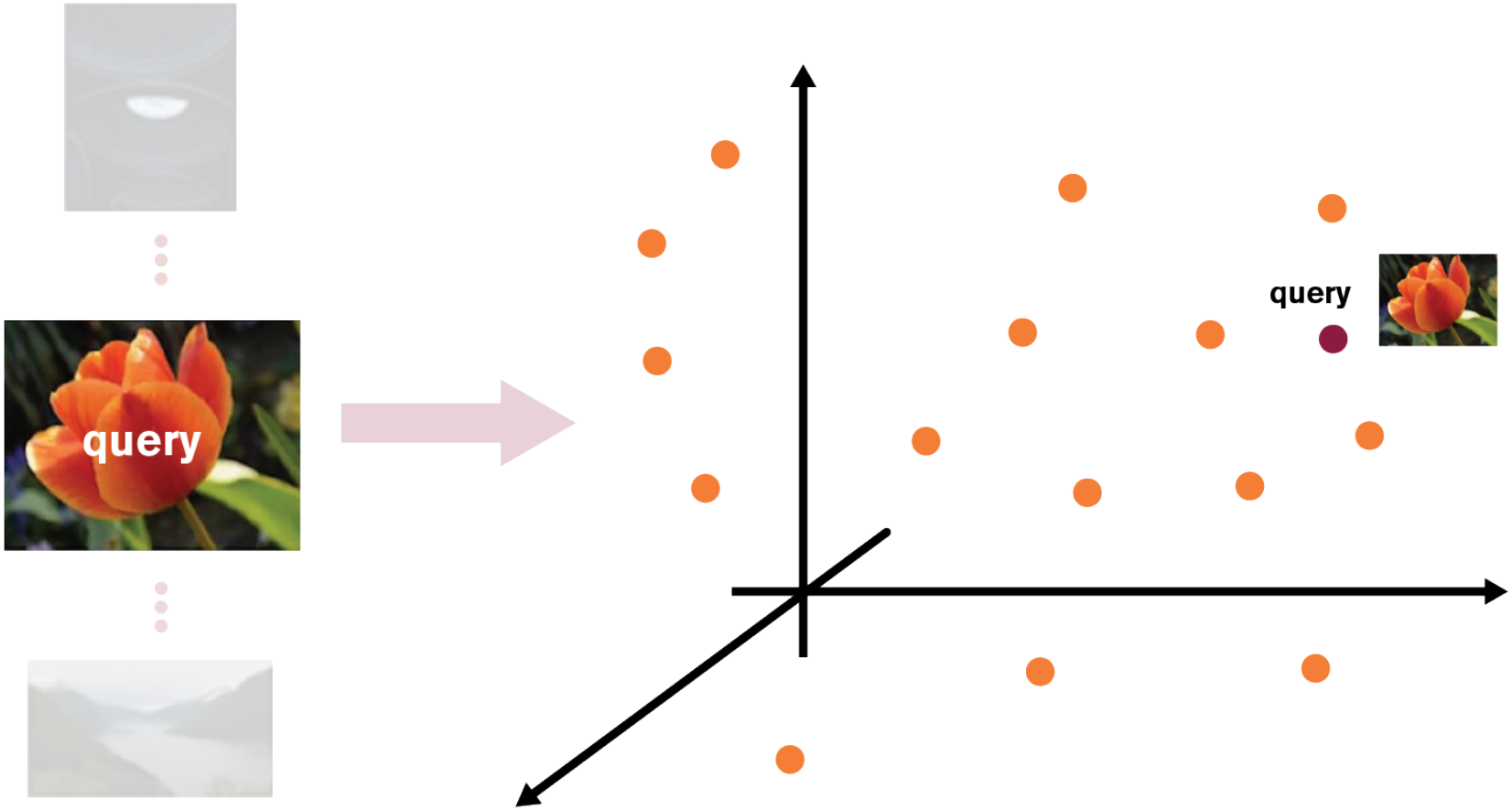
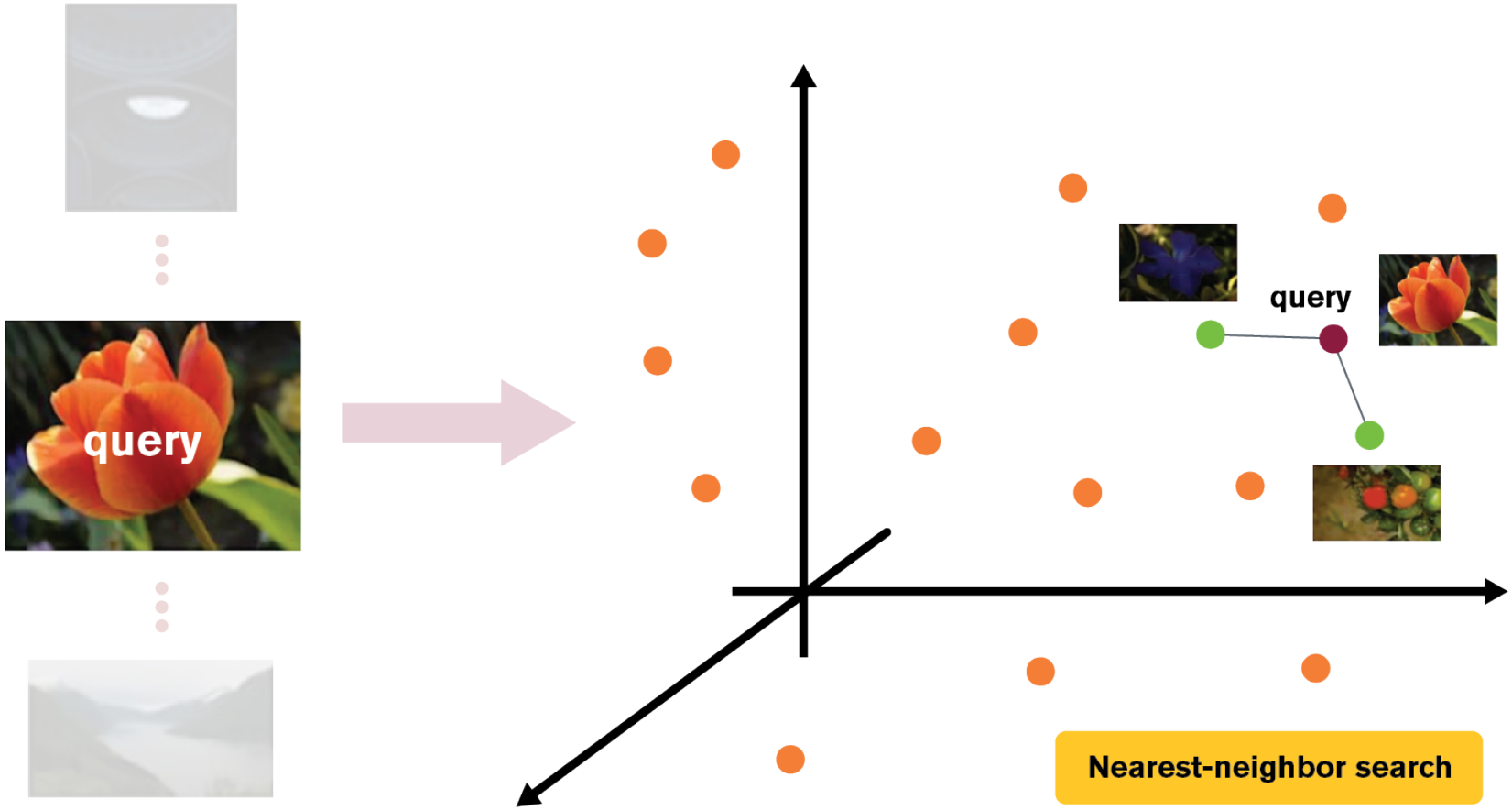


Image
Feature #1
Vector Space
Representation
Image
Feature #1
Image
Feature #2

“Find K=2 most similar images”



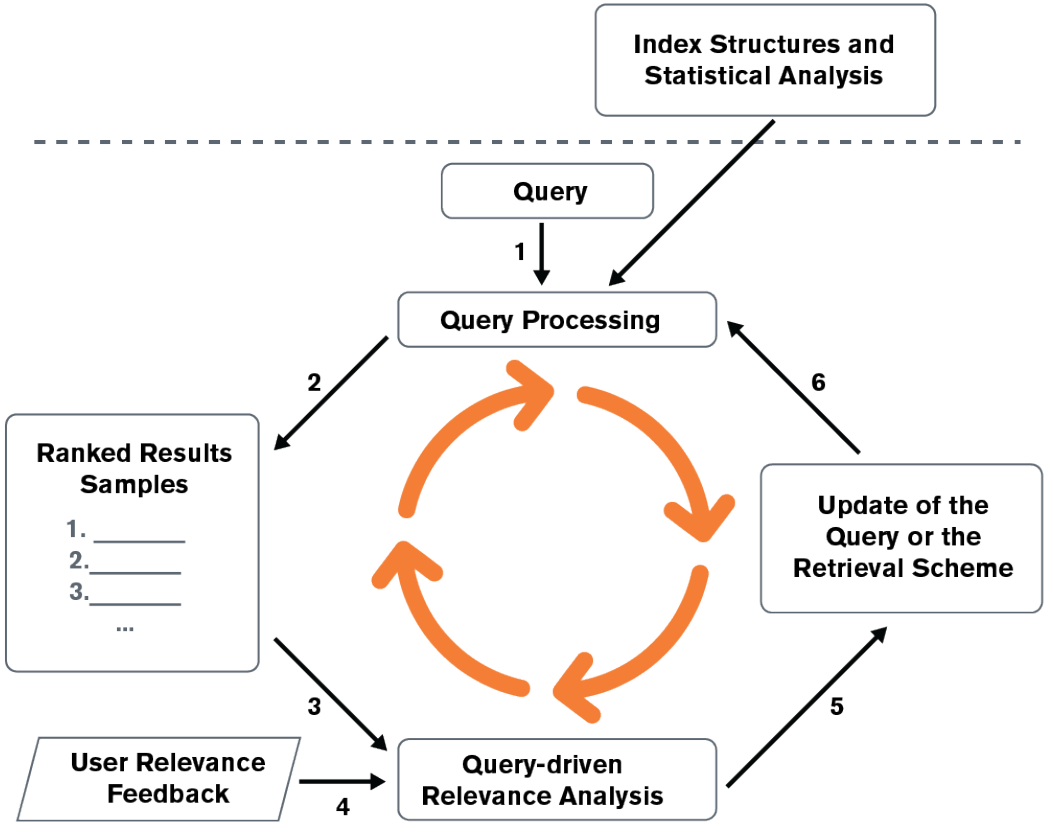
“Find K=2 most similar images”



Sematic gap/subjectivity

Relevance feedback

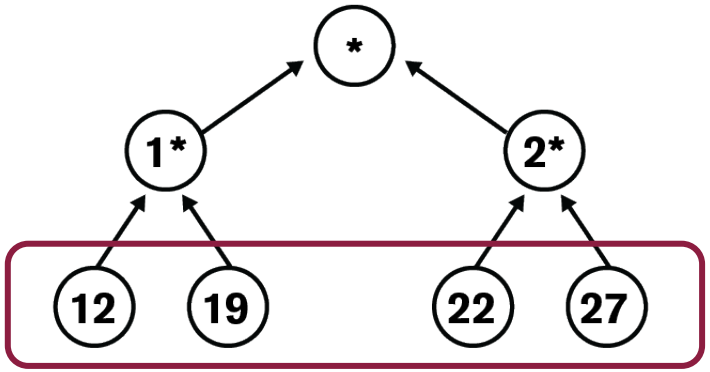
- In a request to identify images “similar” to an example, the visual features of the images are relevant for the user’s query must be inferred from feedback to identify the most relevant images.



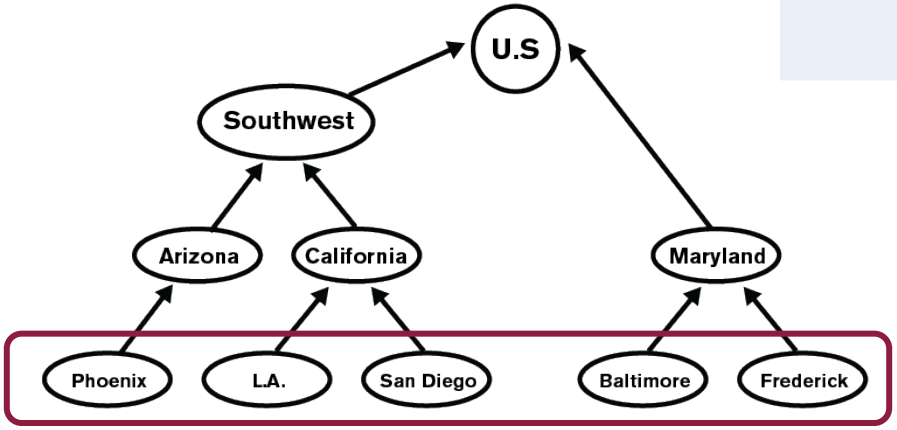
Exploratory Querying

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

Age Metadata



Location Metadata

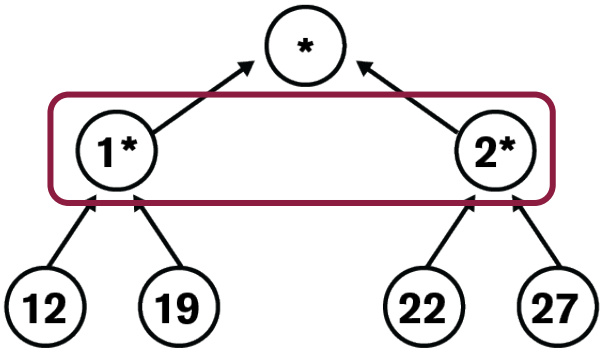


Data Table

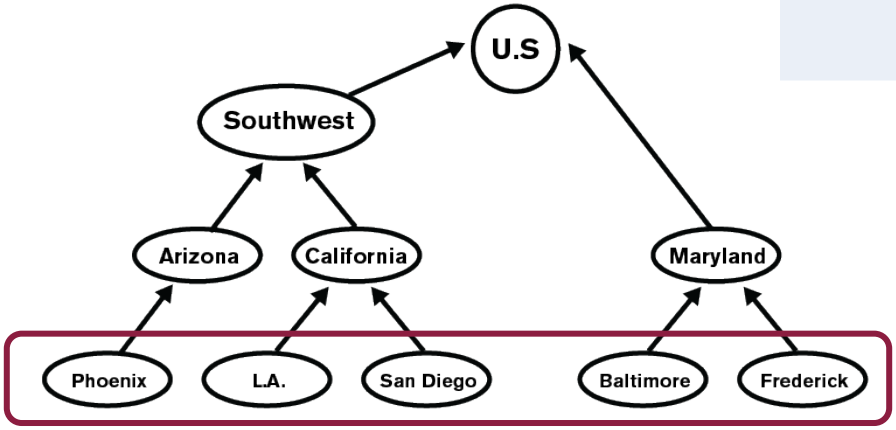
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Roll-Up on Age

Age Metadata



Location Metadata

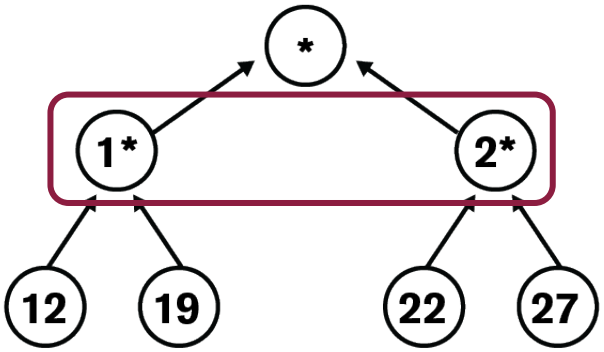


Data Table

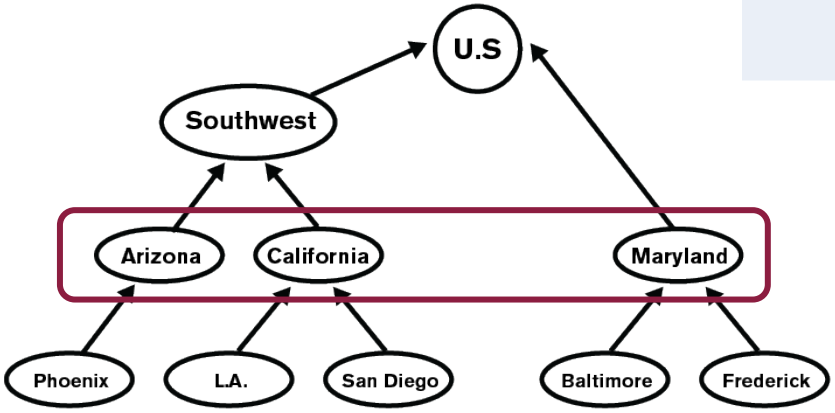
Name	Age	Location
John	1*	Phoenix
Sharon	1*	Los Angeles
Mary	1*	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Roll-Up on Location

Age Metadata



Location Metadata

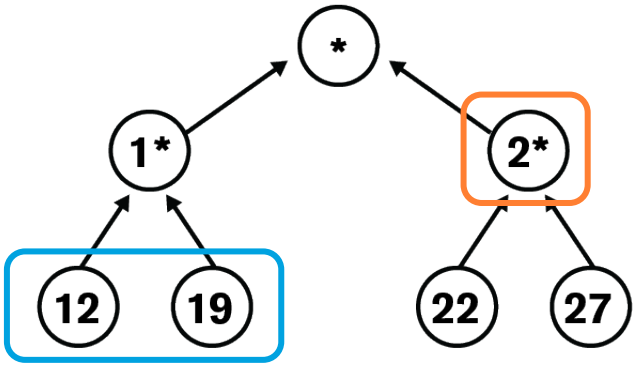


Data Table

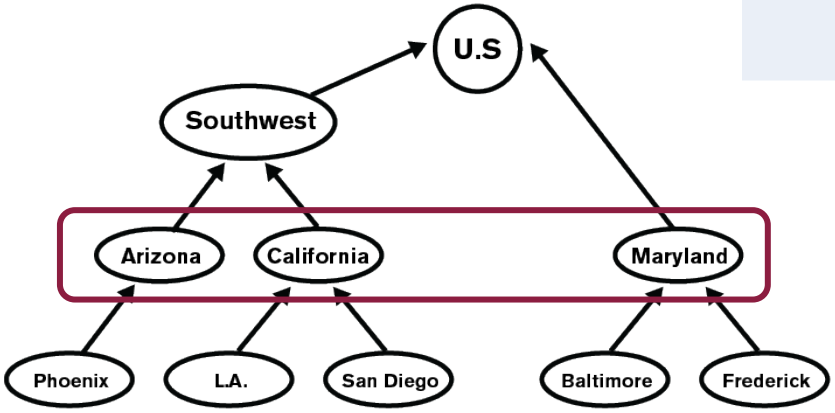
Name	Age	Location
John	1*	Phoenix
Sharon	1*	Los Angeles
Mary	1*	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Drill Down on Age (1*)

Age Metadata



Location Metadata

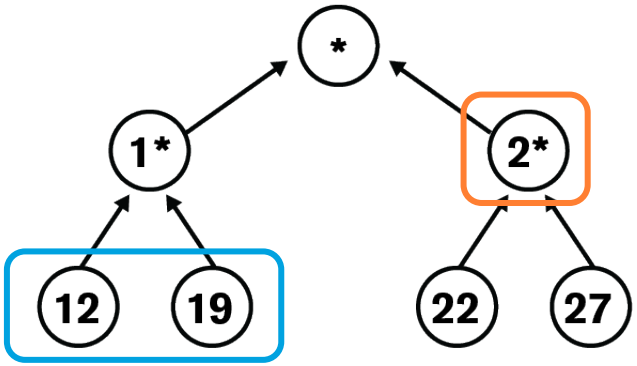


Data Table

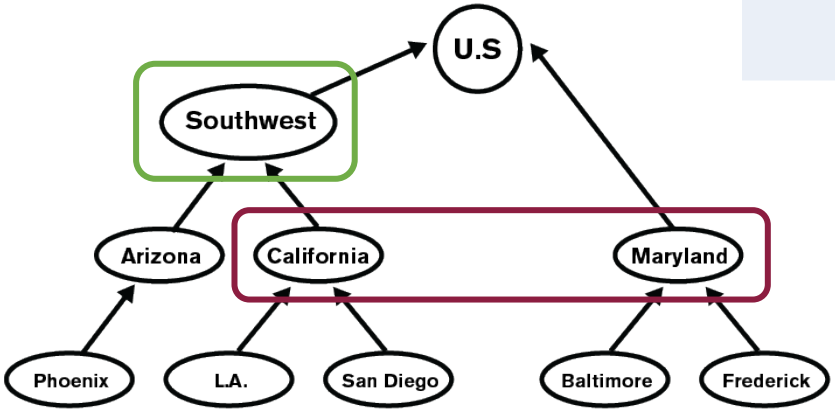
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	2*	Baltimore
James	2*	Frederick
Alice	2*	Baltimore

Roll-Up on Location (Arizona)

Age Metadata



Location Metadata



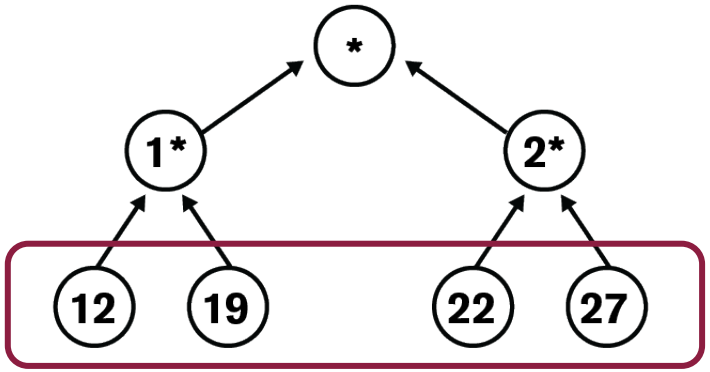
Data Table

Name	Age	Location
John	12	Southwest
Sharon	19	California
Mary	19	California
Peter	2*	Maryland
James	2*	Maryland
Alice	2*	Maryland

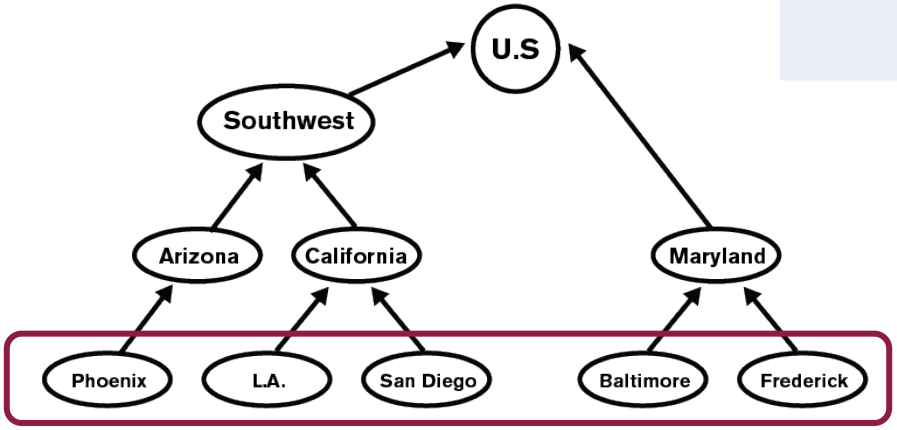
Exploratory Querying

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

Age Metadata



Location Metadata

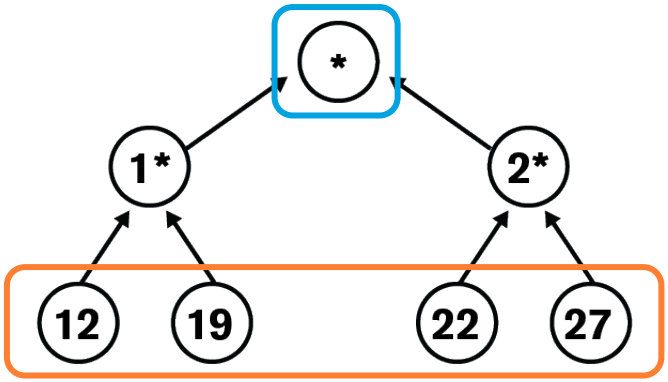


Data Table

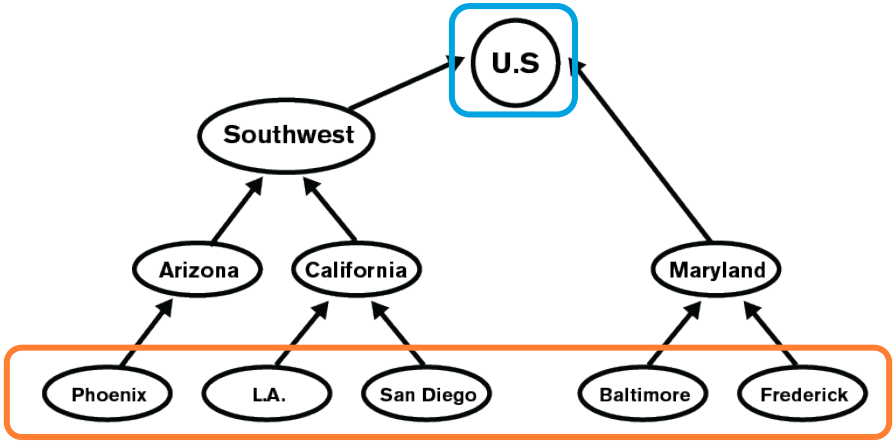
Name	Age	Location
John	12	Phoenix
Sharon	19	Los Angeles
Mary	19	San Diego
Peter	22	Baltimore
James	22	Frederick
Alice	27	Baltimore

Summarization (target # rows = 2)

- Age Metadata



- Location Metadata



Summarized Data Table

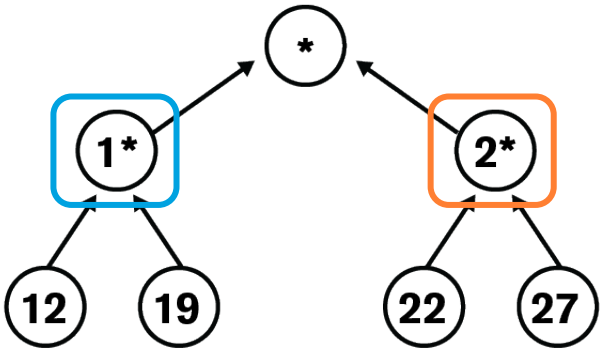
Name	Age	Location	Aggregate (count)
-	1*	Southwest	3
-	2*	Maryland	3

Exploratory Querying

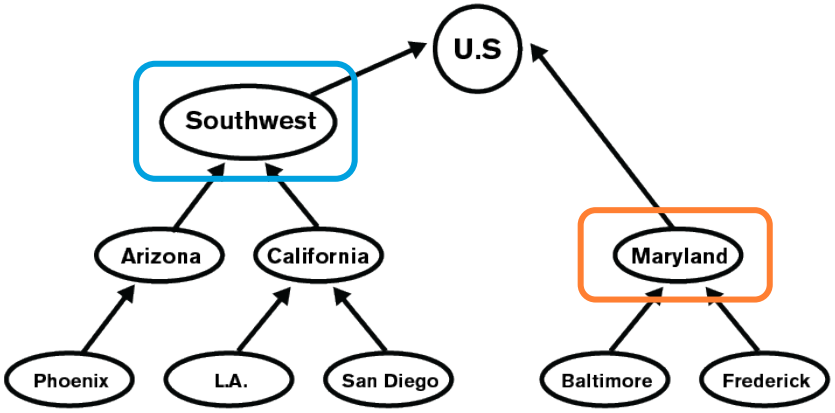
- Similarity queries/Ranked queries
- Drill-down/Roll-up
- Frequent itemsets; sketches; summaries
- Aggregate/iceberg queries
- Skyline queries

Alternative Summarization (target # rows = 2)

Age Metadata



Location Metadata

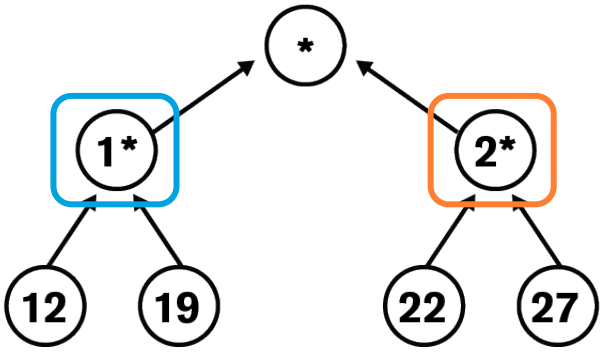


Summarized Data Table

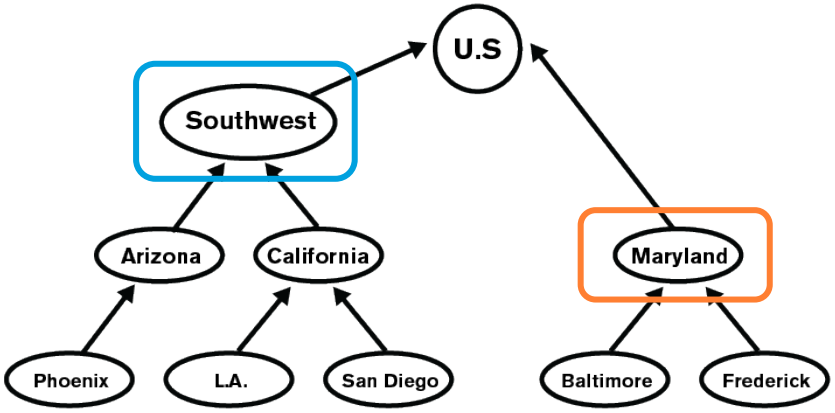
Name	Age	Location	Aggregate (count)
-	1*	Southwest	3
-	2*	Maryland	3

Summarization + Aggregation
(target # rows = 2; max(age))

- Age Metadata



- Location Metadata

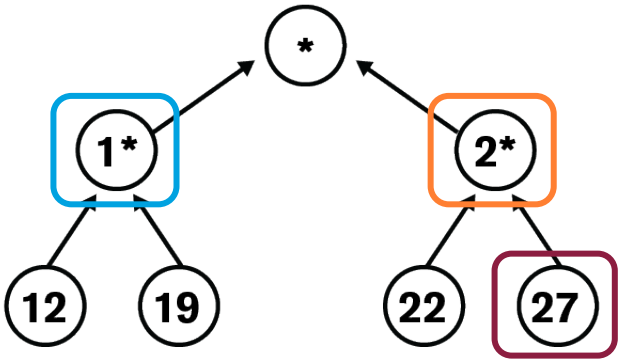


Summarized/Aggregated Data Table

Name	Max(AGE)	Location	Aggregate (count)
-	19	Southwest	3
-	27	Maryland	3

Summarization + Iceberg
(target # rows = 2; max(AGE)>20)

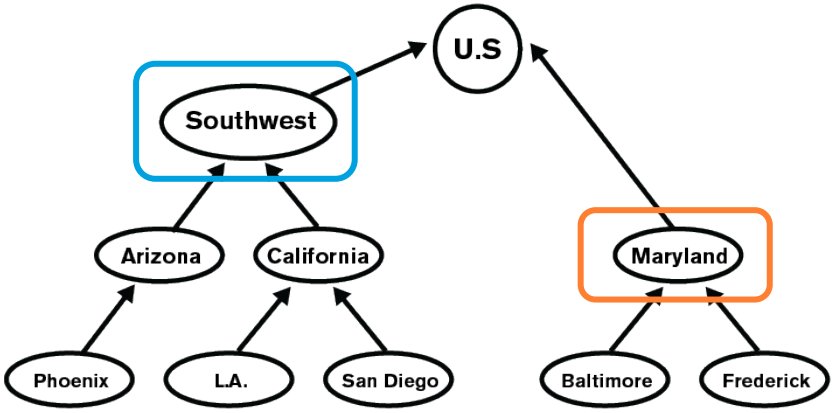
- Age Metadata



Summarized/Aggregated Data Table

Name	Age	Location	Aggregate (count)
-	27	Maryland	3

- Location Metadata



Exploratory Querying

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

Skylines

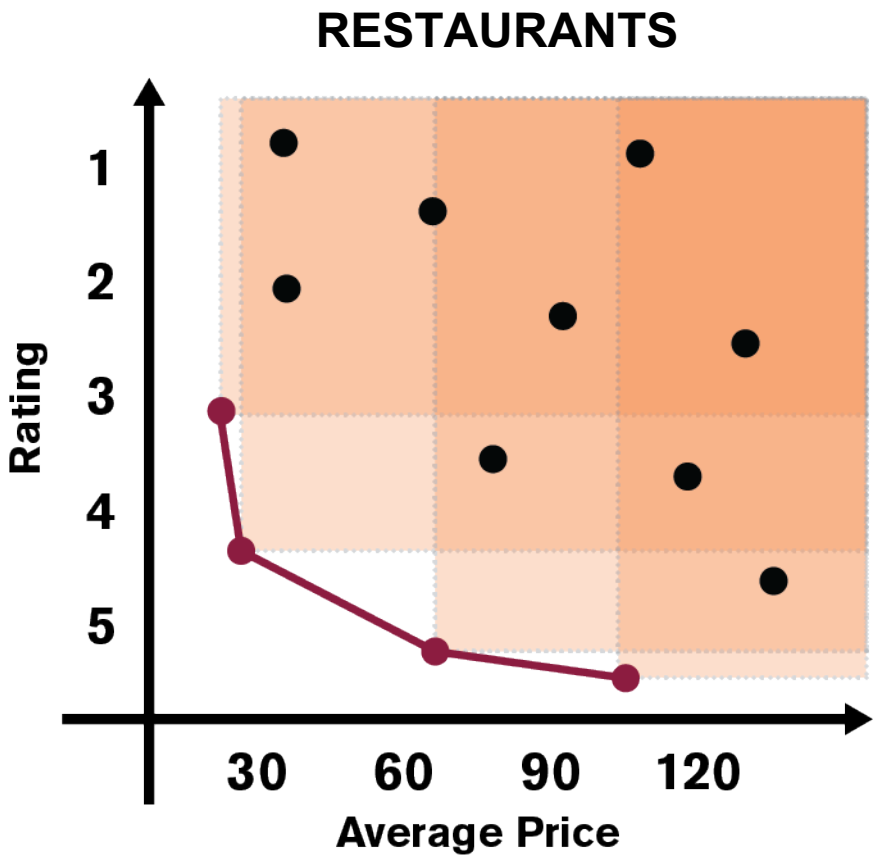
Question

- The higher the rating, the better
- The cheaper the price the better

Which restaurants would you consider?

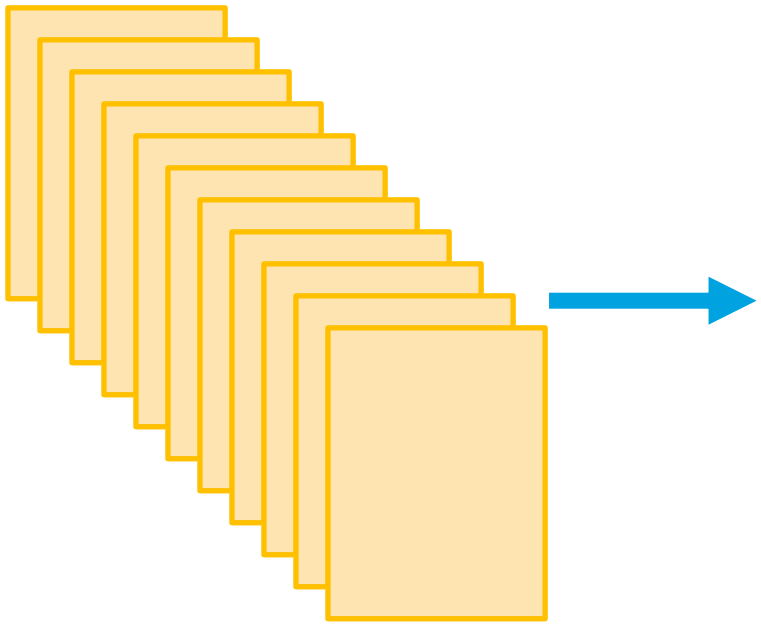
Skylines

- **Objects in the "skyline" are not dominated by any other objects in the database**
  - Also known as the Maximum Vector Problem [Kung 75]
  - Coined as "Skylines" in [Börzsönyi01]



Data sketches – example: tag clouds

Document Collection

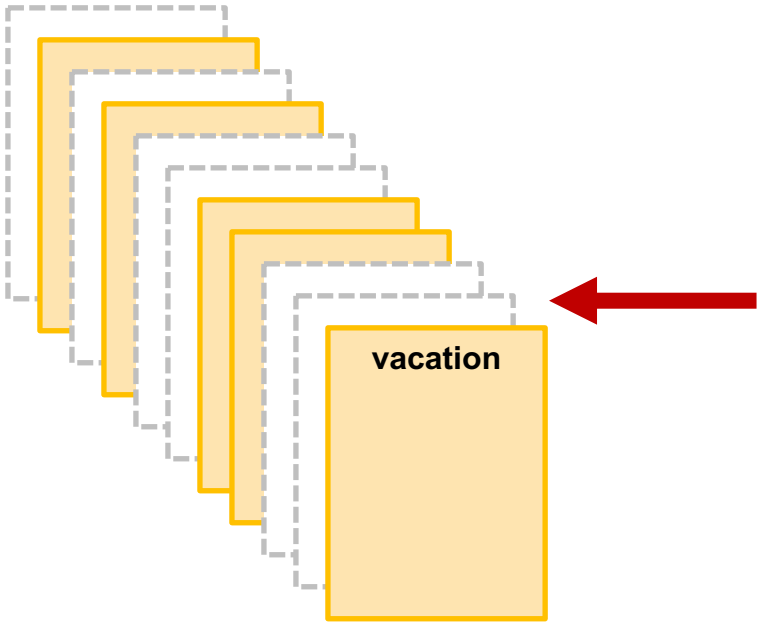


Tag/term cloud

amsterdam animal animals april architecture art australia baby barcelona beach berlin bird birthday black blackandwhite blue boston building bw california cameraphone camping canada canon car cat cats chicago china christmas church city clouds color concert day dc dog dogs england europe family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon hongkong house india ireland italy japan july kids lake landscape light live london losangeles macro march may me mexico moblog mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo portrait red river roadtrip rock rome san sanfrancisco school scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo

Data sketches – example: tag clouds

Document Collection

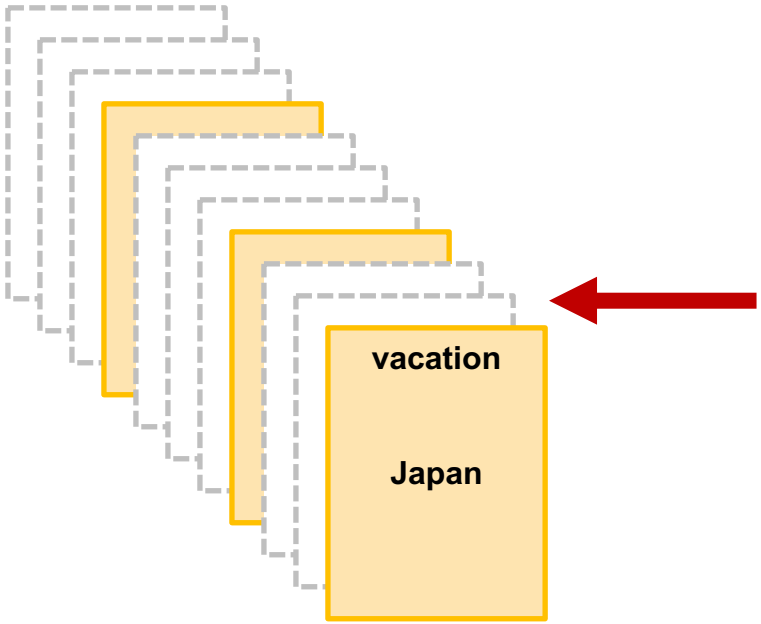


Tag/term cloud

amsterdam animal animals april architecture art australia baby barcelona beach berlin bird birthday black blackandwhite blue boston building bw california cameraphone camping canada canon car cat cats chicago china christmas church city clouds color concert day dc dog dogs england europe family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon hongkong house india ireland italy japan july kids lake landscape light live london losangeles macro march may me mexico moblog mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo portrait red river roadtrip rock rome san sanfrancisco school scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo

Data sketches – example: tag clouds

Document Collection



Tag/term cloud

Introduction to Data Exploration

Visual Variables

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Define the properties of Bertin’s visual variables

Data Types

Nominal:

Data whose categories have no implied ordering.

Ordinal:

Data that has a specified order, but no specified distance metric.

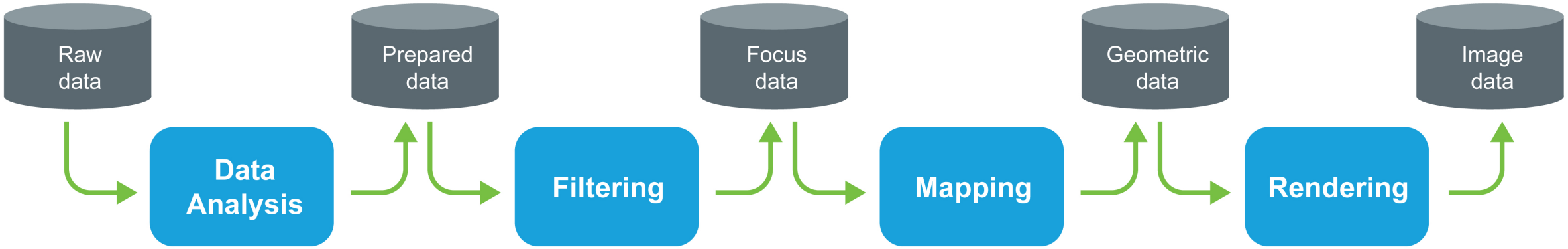
Interval:

Data that has measurable distances.

Ratio:

Same as interval, but include a zero point.

Visualization Pipeline



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

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

Mapping Data: Aesthetic Attributes

Form

Surface

Motion

Sound

Text

L Wilkinson (2005) *The Grammar of Graphics*

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

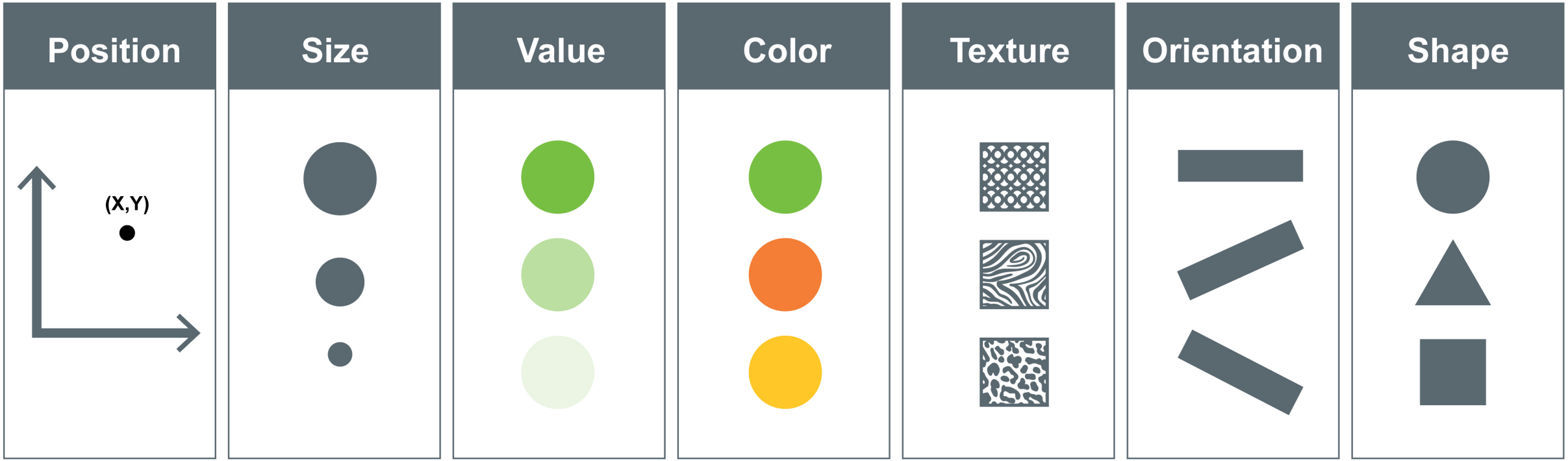
Graphic Design

Much of the skill in graphic design is knowing what combination of attributes should be avoided.

1-SM Kosslyn (1994), *The Elements of Graph Design*

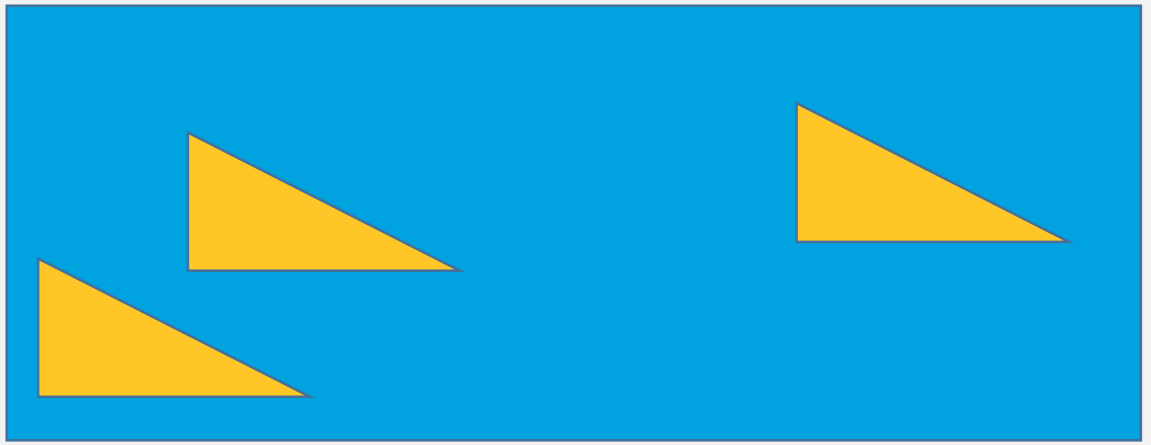
Bertin's Visual Variables

Visualization is concerned primarily with a mapping to visual form



Position

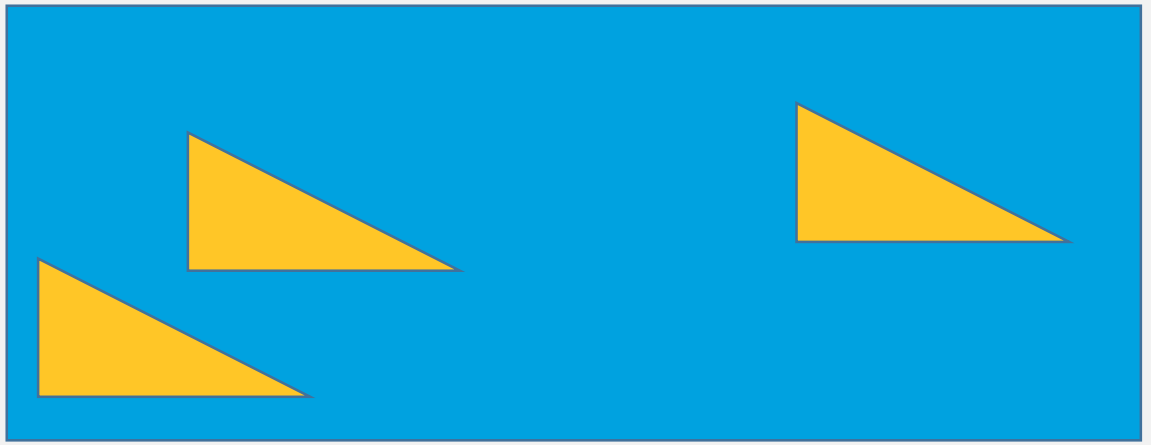
- A location in a multi-dimensional space
- **Continuous variables** map to densely distributed locations
- **Categorical variables** map to a lattice
- Ordering may or may not have meaning in terms of what is being measured



J Bertin (1967), *The Semiology of Graphics*

Position

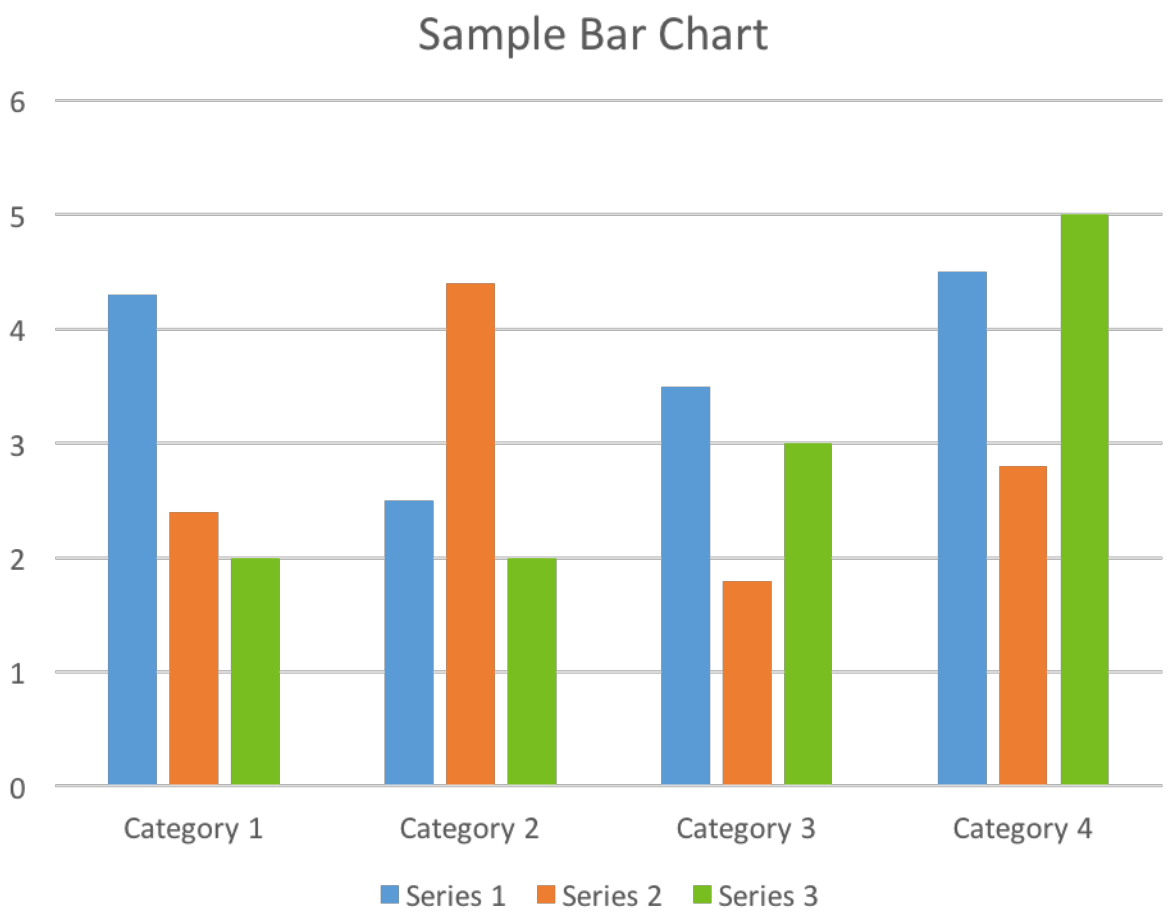
- Best way to represent a quantitative dimension visually
- Points or line lengths placed adjacent to a common axis enable judgments with the least bias or error



J Bertin (1967), *The Semiology of Graphics*

Size

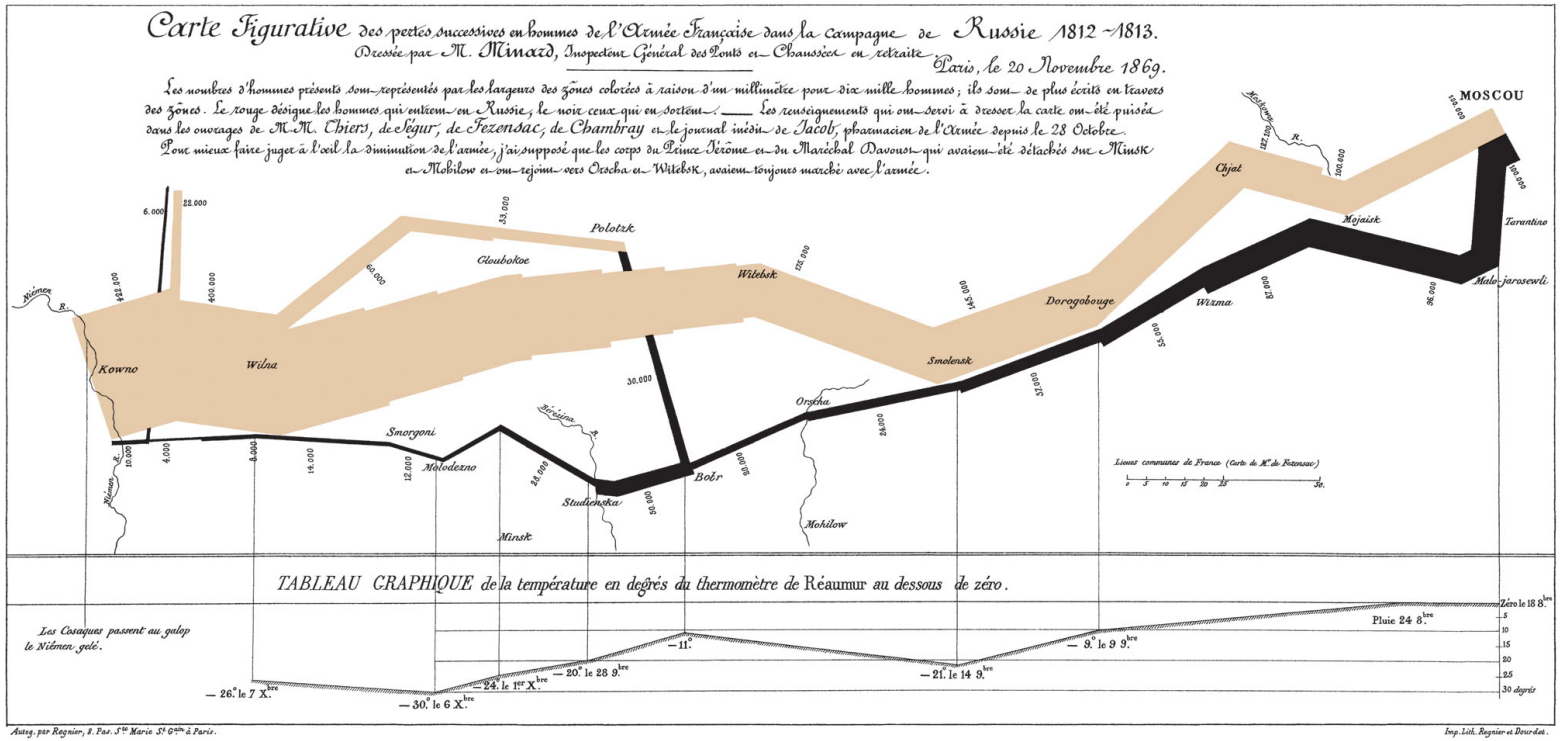
- The variation in terms of length or area
- In three dimensions, includes volume
- Area and volume representations among the **worst attributes** to use for graphing data



J Bertin (1967), *The Semiology of Graphics*

Size

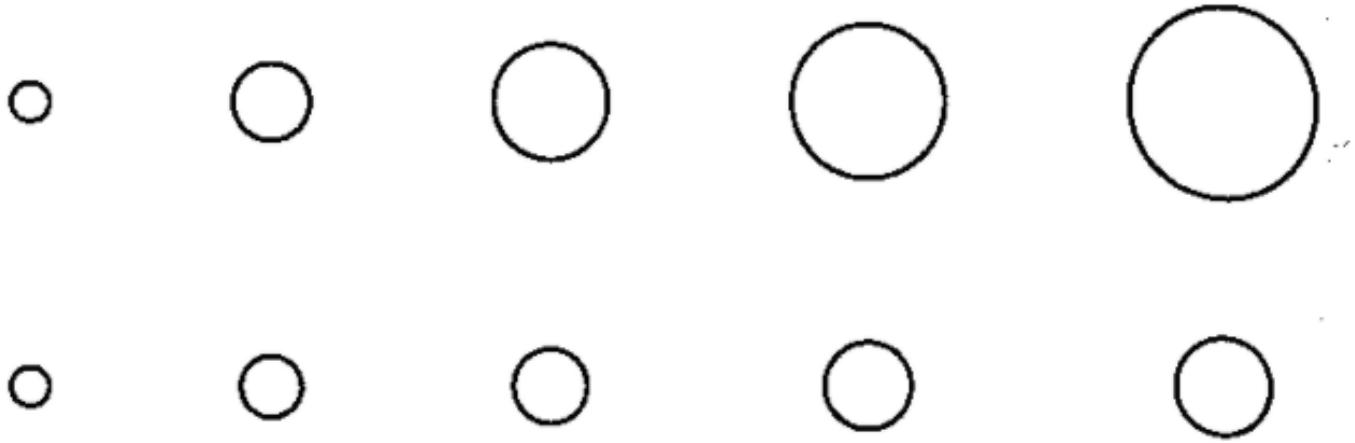
- Size for lines is usually equivalent to thickness
  - less likely to induce perceptual distortion
- Size can be used to great effect with path



1 – Charles Joseph Minard: Mapping Napoleon's March, 1861 by John Corbett, Center for Spatially Integrated Social Science

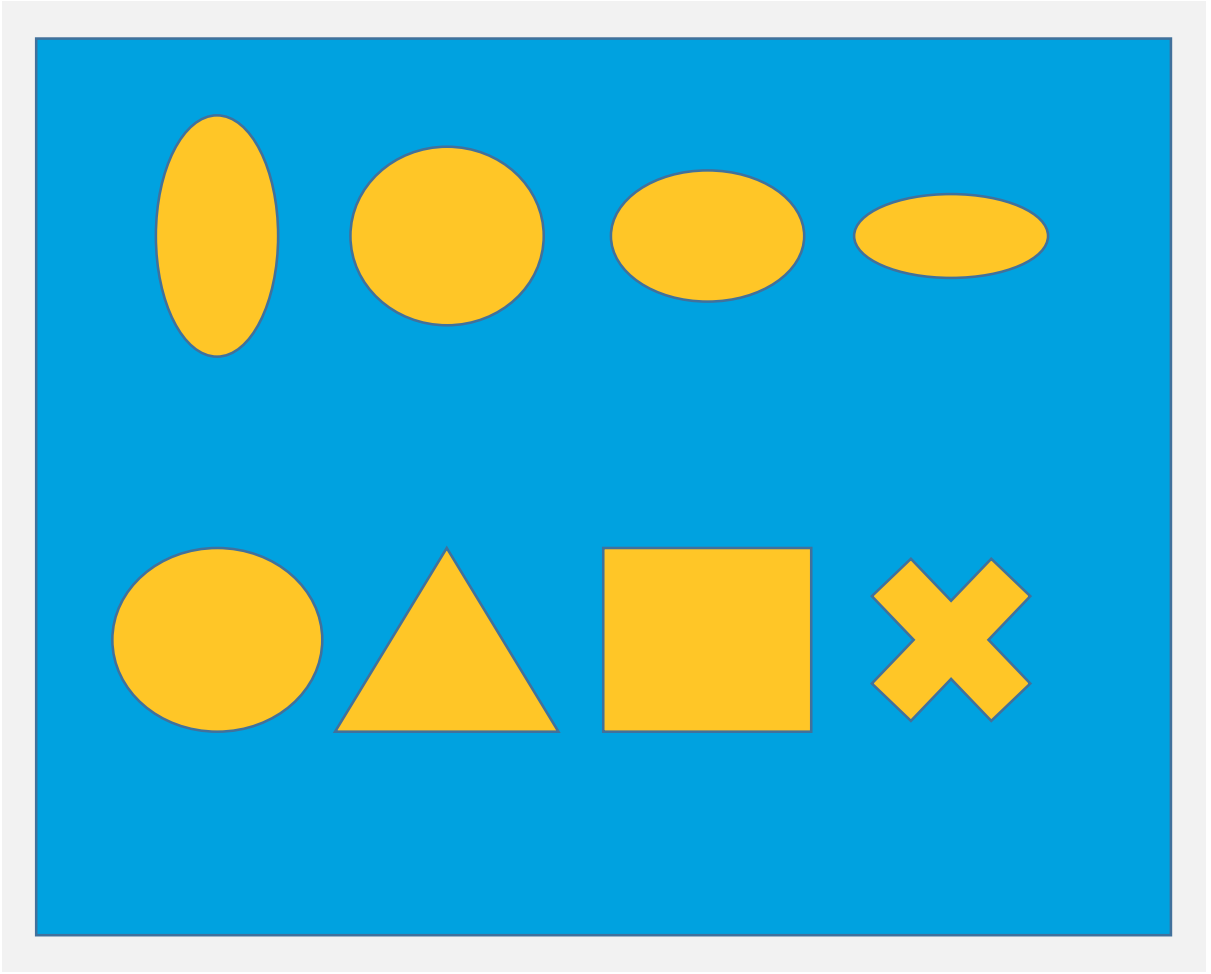
- For objects with rotational symmetry, map size to the diameter rather than area

- 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



Shape

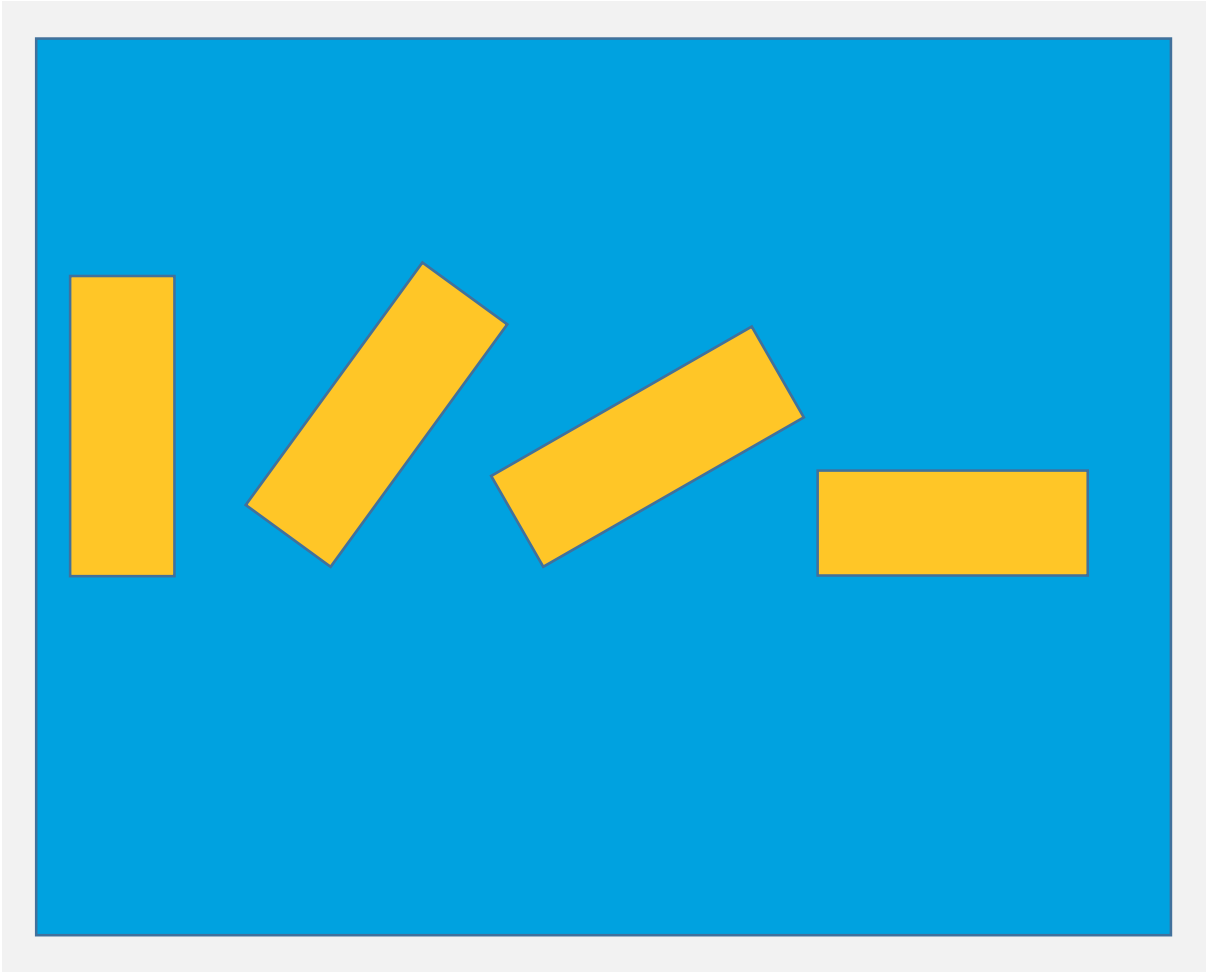
- The shape or boundary of an object
- Shape must vary without affecting size, rotation and other attributes
- **Example:**
  - Map symbols



J Bertin (1967), *The Semiology of Graphics*

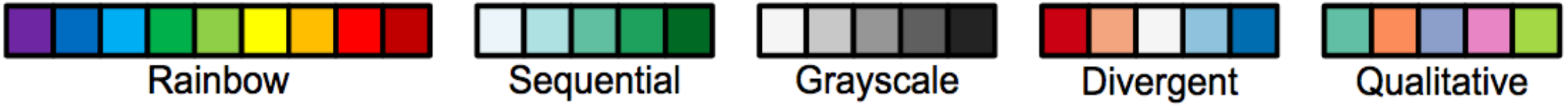
Rotation

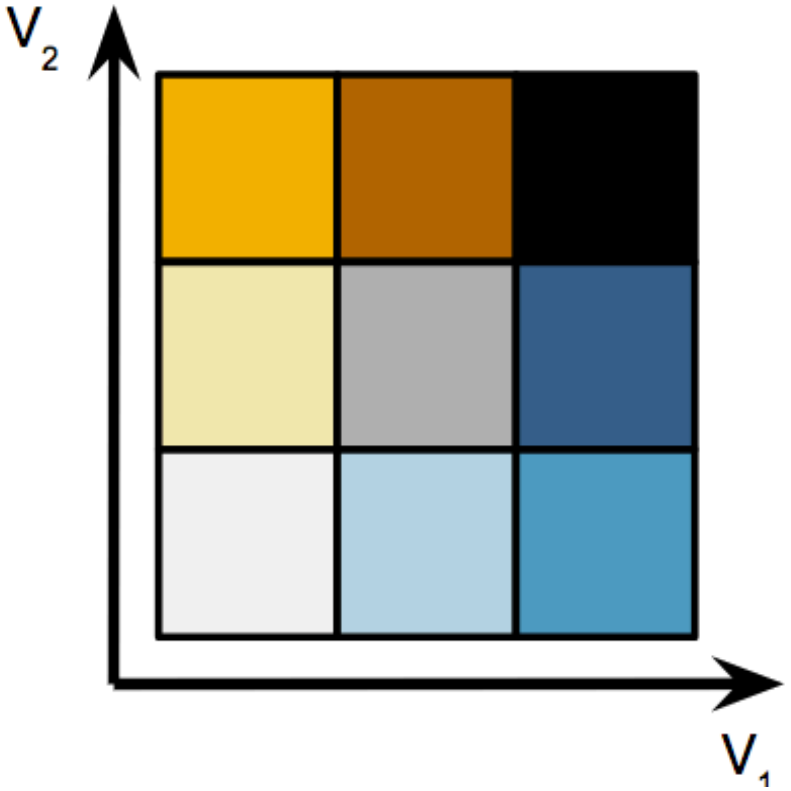
- Rotational angle of the graphic primitive
- Lines, areas and surfaces can only rotate if they are positionally unconstrained

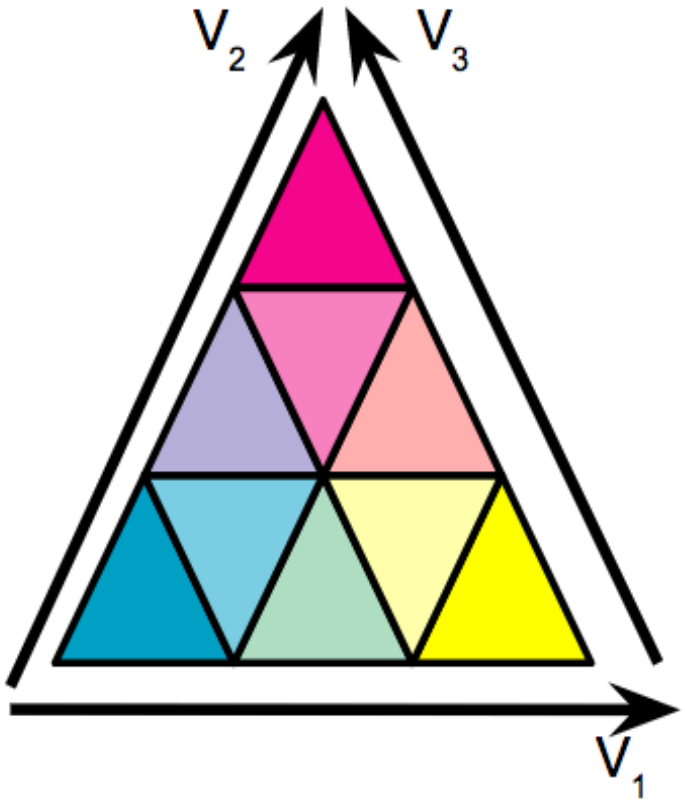


J Bertin (1967), *The Semiology of Graphics*

Color


Rainbow
Sequential
Grayscale
Divergent
Qualitative


V2
V1


V2
V3
V1

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

J Bertin (1967), *The Semiology of Graphics*

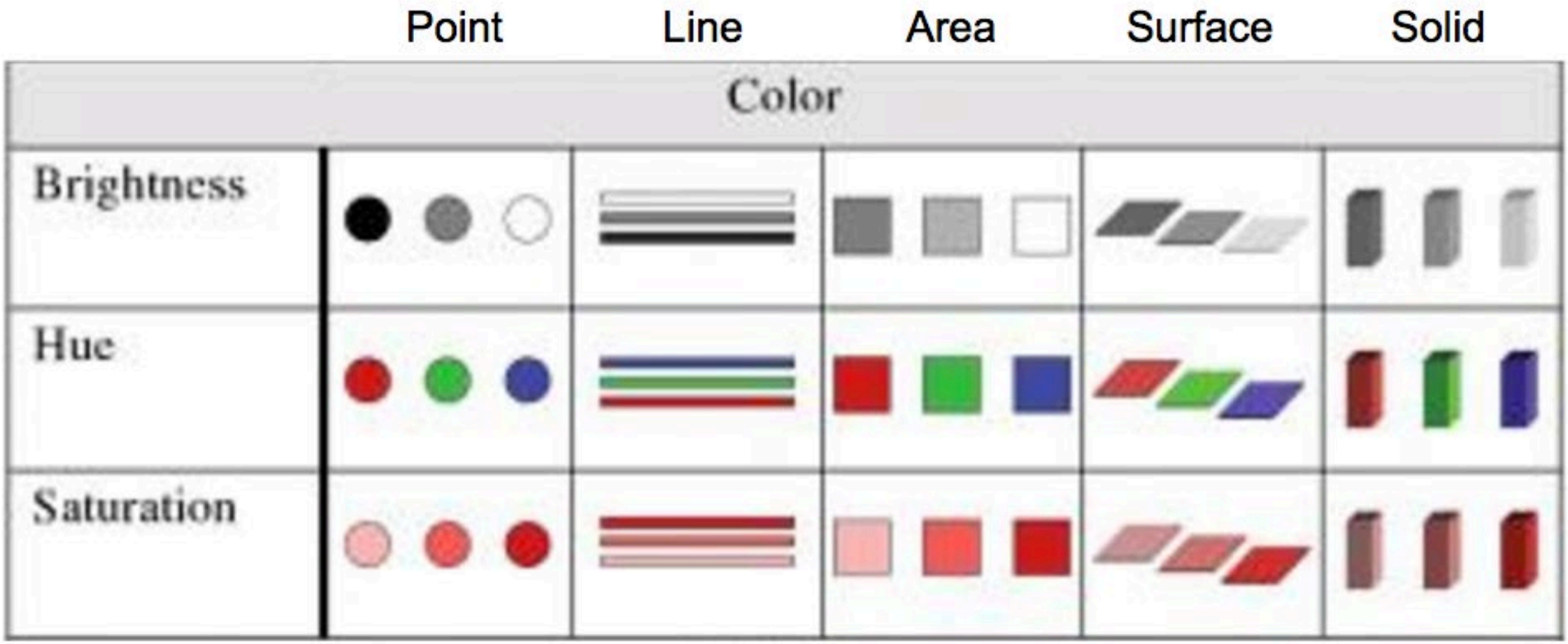
Texture

Textures can be described in a variety of ways

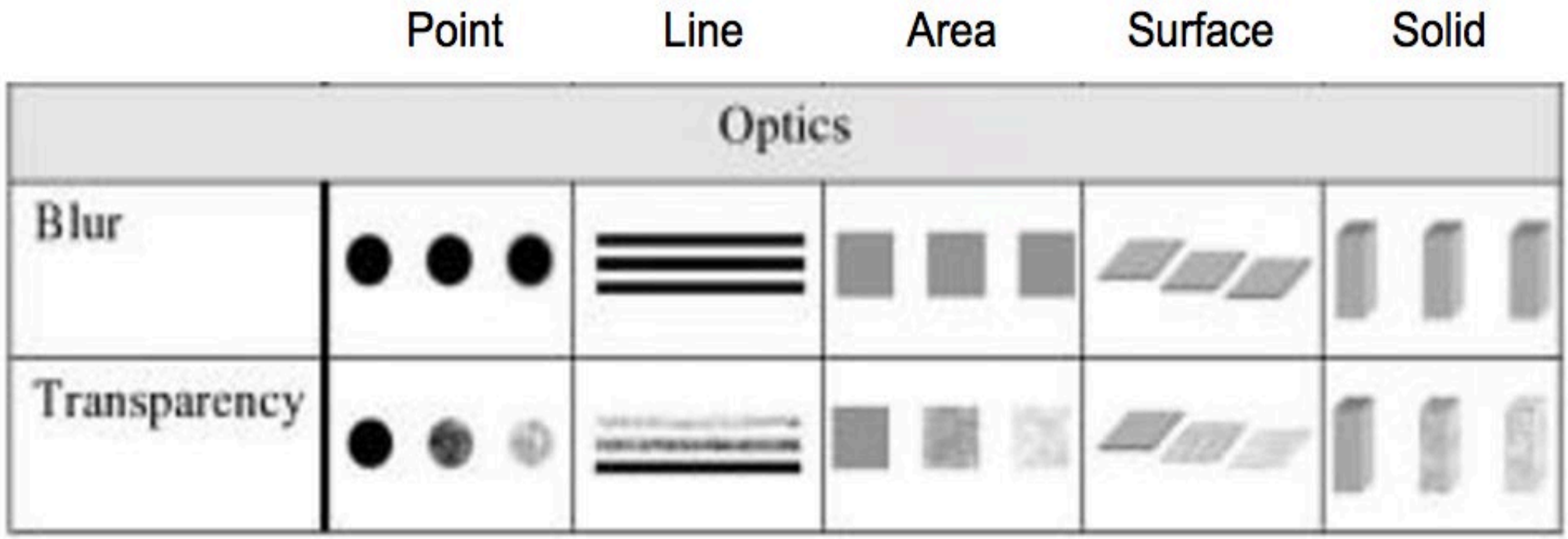
- **Fourier transform** – decomposes a grid of brightness values into sums of trigonometric components
- **Auto-correlogram** – characterize the spatial moments of a texture

J Bertin (1967), *The Semiology of Graphics*

	Point	Line	Area	Surface	Solid
Form					
Size					
Shape					
Rotation					


Point
Line
Area
Surface
Solid
Color
Brightness
Hue
Saturation

	Point	Line	Area	Surface	Solid
Texture					
Granularity					
Pattern					
Orientation					


Point
Line
Area
Surface
Solid
Optics
Blur
Transparency

Introduction to Data Exploration

Color Schemes and Design

Ira A. Fulton Schools of Engineering
Arizona State University

Objectives

Objective

Identify appropriate color schemes for different data types

Design Principles

Given a univariate data type,

Order

- the color scale that is chosen to map the data must represent a perceived ordering

Separation

- the color scale that is chosen to map the data must represent a perceived ordering

Aesthetics

- color map should be aesthetically pleasing, contain a maximum perceptual resolution, and ordering 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

Univariate Color Schemes

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

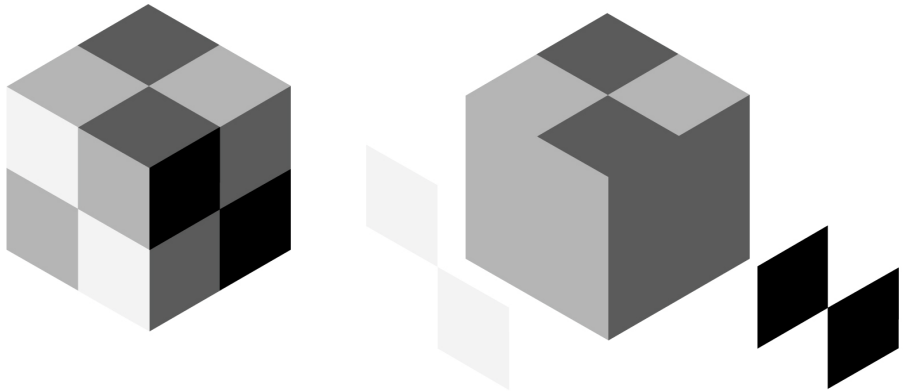
Sequential

Grayscale Color Scheme

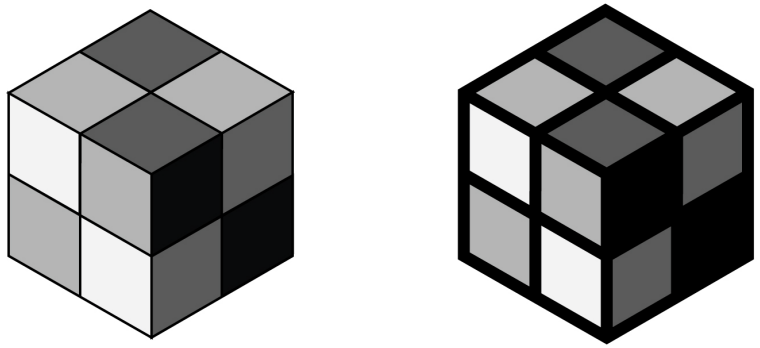
Simplest is the gray scale map where variable is mapped to brightness

Grayscale

Illusions in Grayscale



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



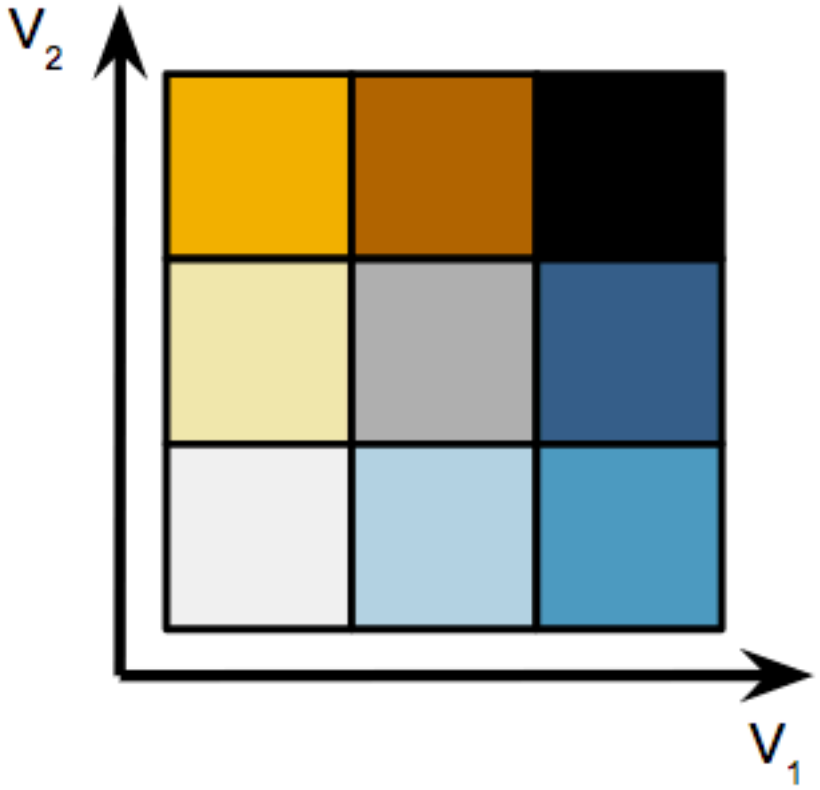
Adding a thin border has minimal effect on the illusion, but having a thick border is able to neutralize the effect

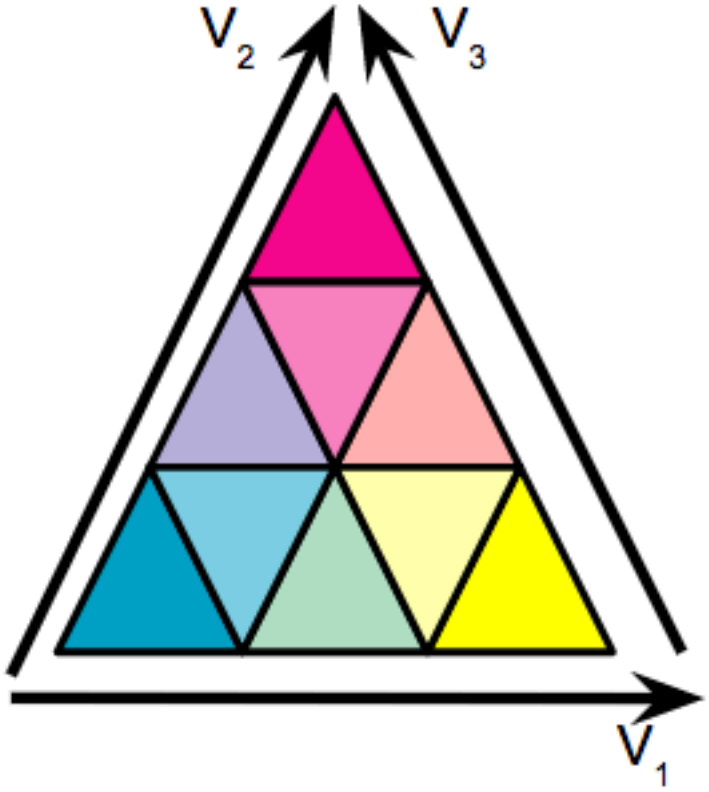
Divergent Color Scheme

- 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 (blues) and warm (reds and yellow) colors

Divergent

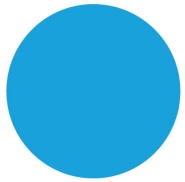
Multivariate Color Schemes



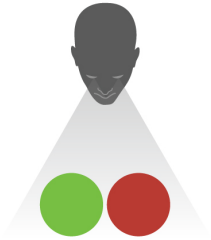


M. A. Harrower and C. A. Brewer, “ColorBrewer.org: An online tool for selecting color schemes for maps," *The Catrtographic Journal*, vol. 40, no. 1, pp. 27-37, 2003.

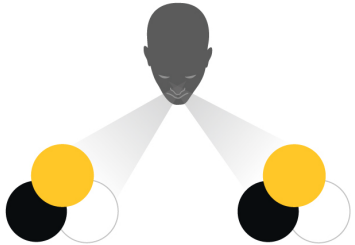
Mapping Color



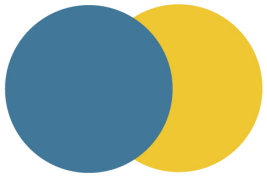
Use blue in large regions, not thin lines



Use red and green in the center of the field of view

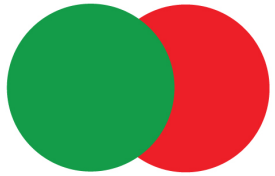


Use black, white and yellow in the periphery



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 exteme colors together

Do not use adjacent colors that vary in amount of blue

Mapping Color

Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?
Hello, here is some text. Can you read what it says?