Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

24-January-2023

NTNU | Norwegian University of Science and Technology

1. QA and References
   - Administrative
   - References for Today's Lecture

2. Data Warehouse Implementation
   - Indexing OLAP Data
   - OLAP Server Architectures
   - Web-Scale Indexing

3. Data
   - Data
   - Data Modeling
   - Attributes
   - Data Characteristics
   - Data Set Types

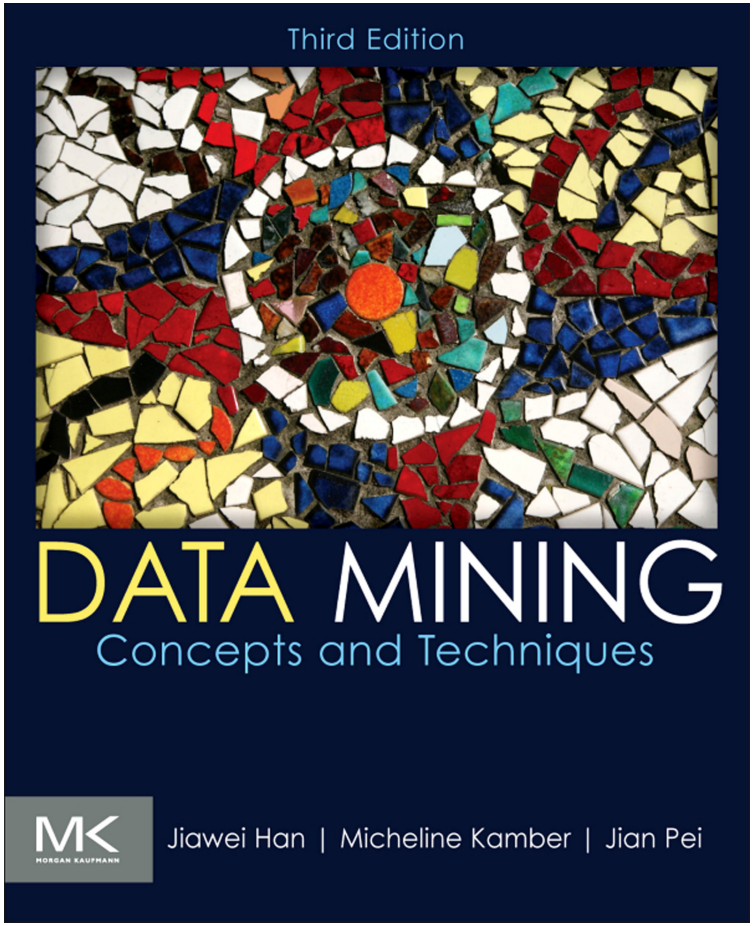
4. Summary

Administrative

1. First Assignment
   - Available this week (26.Jan.2023) and due by 09.February.2023.
   - There will be a tutorial next week on Monday (30.Jan.2023) in the lecture slot to introduce the first assingment.
2. Volunteers for feedback regarding course
   - Interested? Please contact me by email!

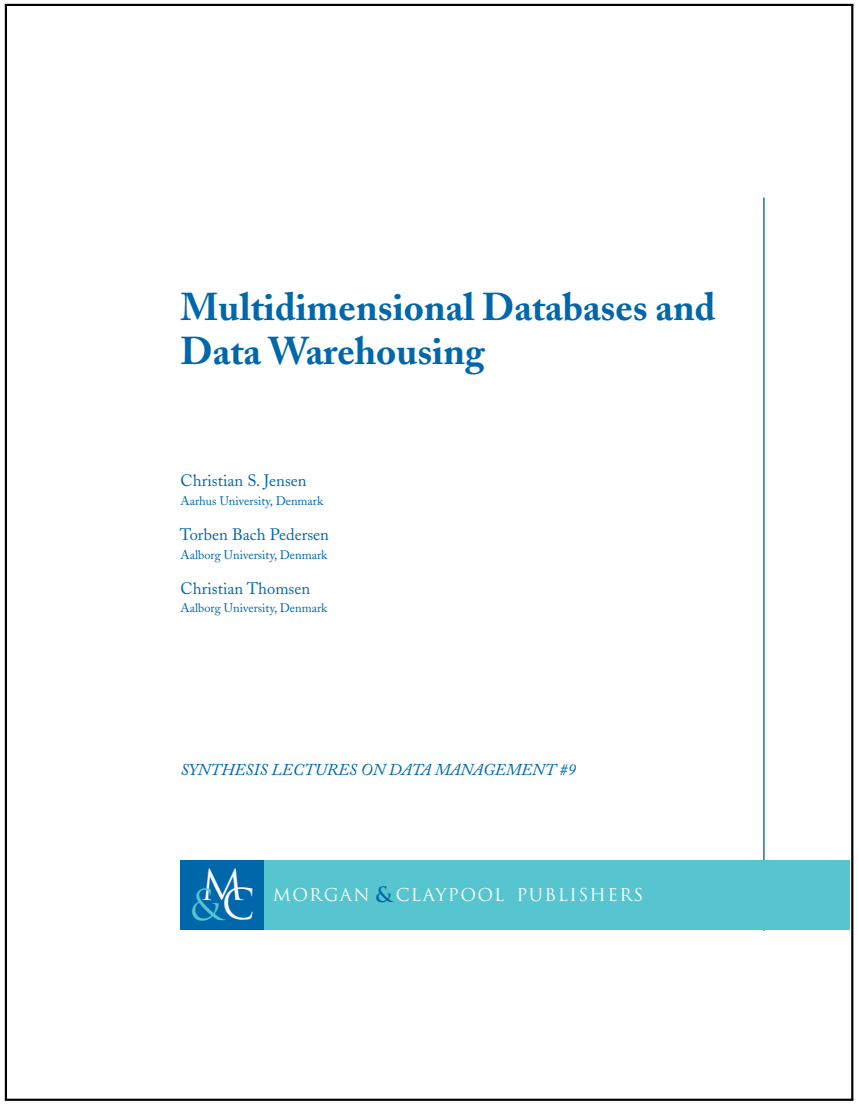
References for "Data Warehouse Concepts & Impl."

1. Book: Han et al. *"Data Mining Concepts and Techniques"*, 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. All text and images for "Data Warehouse Concepts and Implementation" are based on the book by Han et al.



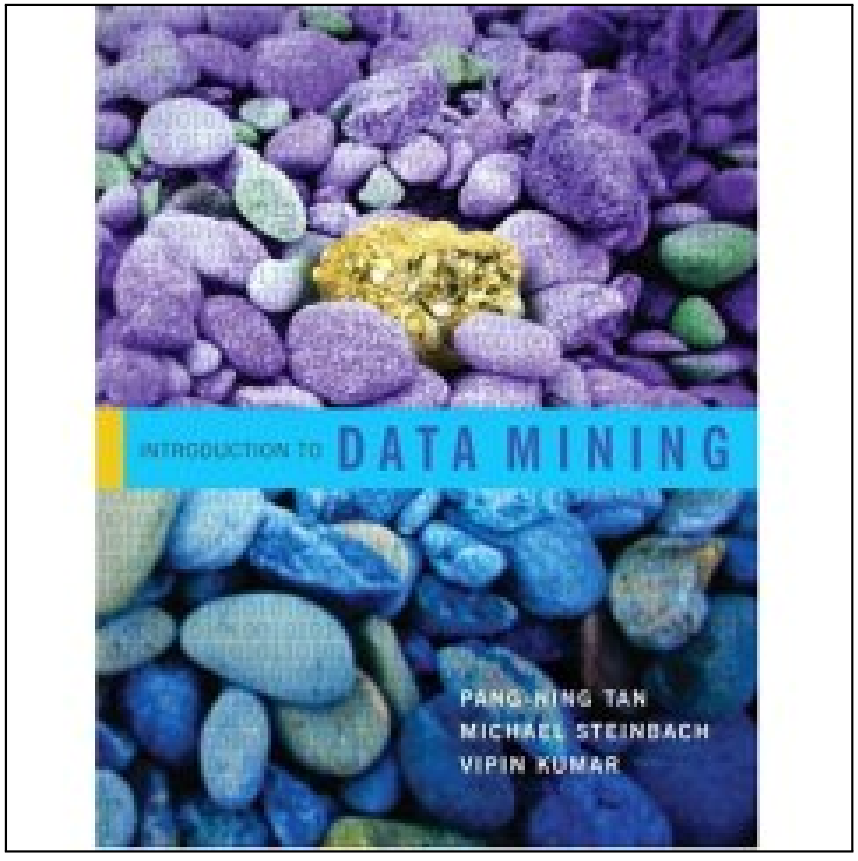
References for "Bitmap and Join Indexing"

1. Book: Jensen et al. *"Multidimensional Databases and Data Warehousing"*, 2010, Morgan & Claypool Publishers.
2. All text and images for "Bitmap and Join Indexing" are based on the book by Jensen et al.



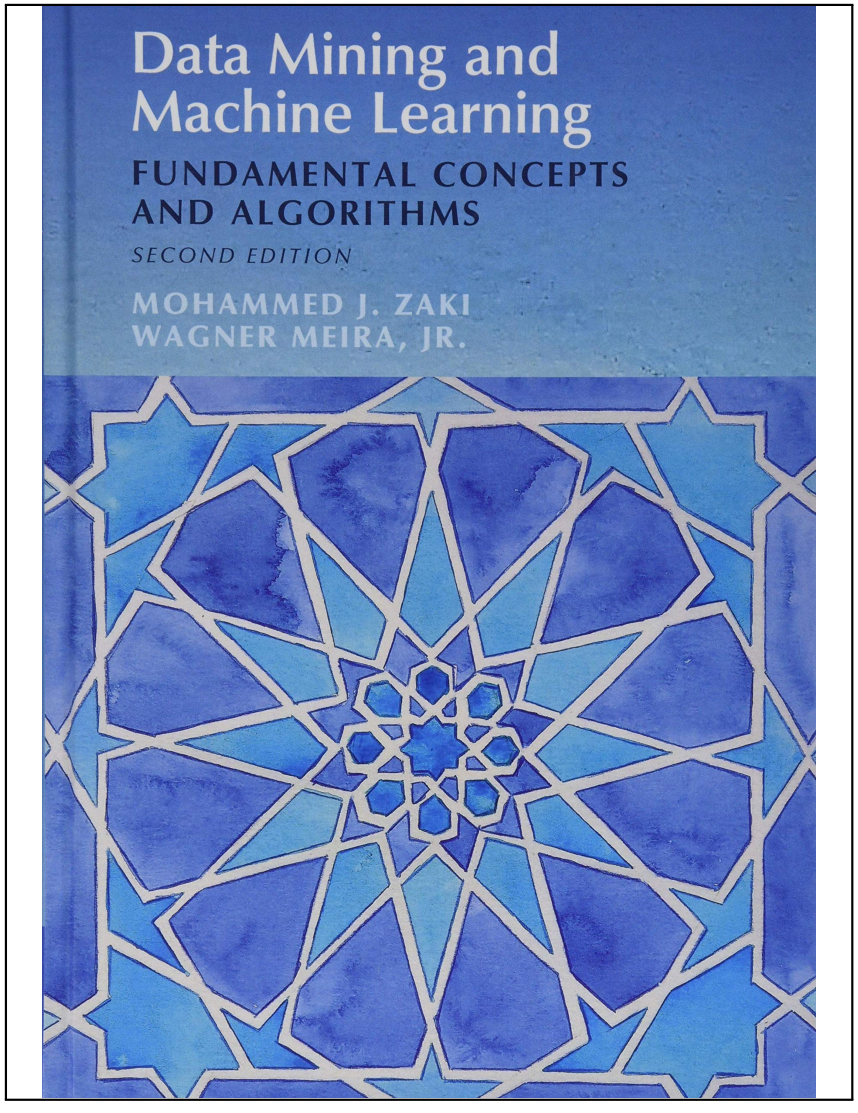
References for "Data"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for some slides in "Data" subsection are based on the book by Tan et al.



References for "Data"

1. Book: Zaki and Meira. *"Data Mining and Machine Learning: Fundamental Concepts and Algorithms"*, 2nd Edition, 2020, Cambridge University Press.
2. Text and images for some slides in "Data" subsection are based on the book by Zaki and Meira.



1. QA and References
   - Administrative
   - References for Today's Lecture

2. Data Warehouse Implementation
   - Indexing OLAP Data
   - OLAP Server Architectures
   - Web-Scale Indexing

3. Data
   - Data
   - Data Modeling
   - Attributes
   - Data Characteristics
   - Data Set Types

4. Summary

Indexing Overview

- Fact tables are typically very large and are used in the majority of queries.
- Therefore, a fact table often has many indices such that most queries can benefit from one or more of them.
- Typically, a separate index is built on each dimension key.

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Indexing Overview

- If the DBMS supports index intersection, the indices can then be used in combination when answering a query.
- It can also be a good idea, however, to create indices on combinations of dimension keys for those combinations that are often used together by queries.
- The combinations for which to create indices thus have to be chosen carefully based on the typical usage.
- In general, the key referencing the Date dimension should be put first in combinations since most queries refer to dates.

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Indexing Overview

- Virtually any DBMS supports the B-tree index which is a tree structure where the leaves contain lists of row IDs (where a row ID can be a physical location of the row or something else the system can use to identify and find a certain row).
- To find rows that have a given value for the indexed attribute, the system traverses the tree and obtains a list of row IDs. The row IDs are then used to retrieve the rows.
- The B-tree is an efficient index.
- However, when the indexed attribute has low cardinality (i.e., holds few different values), another representation can be a better choice.

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

- When there are few values in a column, we can create a position bitmap for each value instead of maintaining lists of row IDs for each value (as in the B-tree).
- We call such a collection of position bitmaps a bitmap index.

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

- When there are few values in a column, we can create a position bitmap for each value instead of maintaining lists of row IDs for each value (as in the B-tree).
- We call such a collection of position bitmaps a bitmap index.

Base table

RID	*item*	*city*
R1	H	V
R2	C	V
R3	P	V
R4	S	V
R5	H	T
R6	C	T
R7	P	T
R8	S	T

item bitmap index table

RID	H	C	P	S
R1	1	0	0	0
R2	0	1	0	0
R3	0	0	1	0
R4	0	0	0	1
R5	1	0	0	0
R6	0	1	0	0
R7	0	0	1	0
R8	0	0	0	1

city bitmap index table

RID	V	T
R1	1	0
R2	1	0
R3	1	0
R4	1	0
R5	0	1
R6	0	1
R7	0	1
R8	0	1

Note: H for "home entertainment," C for "computer," P for "phone," S for "security," V for "Vancouver," T for "Toronto."

Bitmap Vector

Figure 4.15 Indexing OLAP data using bitmap indices.

All from: 1. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

RowID	BookID	Title	Binding	Language	...
1	9436	Winnie the Pooh	Hardcover	English	...
2	1029	Le Petit Prince	Paperback	French	...
3	8733	Alice in Wonderland	Paperback	English	...
4	2059	Wind in the Willows	Hardcover	English	...
5	5995	A Bear Called Paddington	Paperback	English	...
6	1031	Pierre Lapin	Hardcover	French	...
7	3984	Le avventure di Pinocchio	Hardcover	Italian	...

Figure 4.2: Book dimension table also showing the special system-maintained attribute RowID, which normally cannot be seen by the user

Bitmap for `Binding`

```
Hardcover: 1001011
Paperback: 0110100
```

Bitmap for `Language`

```
English:  1011100
French:   0100010
Italian:  0000001
```

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

- A bitmap index makes it very fast to locate rows with a certain value.
- It is also possible to combine bitmap indices.
- Example 1: To find all hardcover books in English can be done by computing the logical AND of the position bitmaps for hardcover in Binding and for English in Language. In other words, we compute:

```
        1001011
AND     1011100
=       1001000
```

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

- A bitmap index makes it very fast to locate rows with a certain value.
- It is also possible to combine bitmap indices.
- Example 2: To find books written in French or Italian, we compute logical OR of the position bitmaps for French and Italian in Language:

```
          0100010
OR        0000001
 =        0100011
```

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing

- A bitmap index makes it very fast to locate rows with a certain value.
- It is also possible to combine bitmap indices.
- Example 2: To find books written in French or Italian, we compute logical OR of the position bitmaps for French and Italian in Language:

```
          0100010
OR        0000001
=         0100011
```

- The logical AND can be computed very efficiently as a single CPU operation can compare 32 or 64 bits on a modern CPU.

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Bitmap Indexing — Compression

- Run Length Encoding (RLE) compresses long runs of identical values: it replaces any repetition by the number of repetitions followed by the value being repeated.
- Current microprocessors perform operations over words of 32 or 64 bits and not individual bits.
- CPU cost of RLE might be large.
- Byte-Aligned Bitmap Compression (BBC): By trading some compression for more speed, a RLE variant working over bytes instead of bits (BBC).
- Word-Aligned Hybrid (WAH): Trading even more compression for even more speed.

All from: Lemire et al. "Sorting improves word-aligned bitmap indexes", Data Knowl. Eng. 69(1): 3-28 (2010).

Join Indexing

- Traditional indexing maps the value in a given column to a list of rows having that value.
- In contrast, join indexing registers the joinable rows of two relations from a relational database.
- For example, if two relations `R(RID, A)` and `S(B, SID)` join on the attributes A and B, then the join index record contains the pair (RID, SID), where RID and SID are record identifiers from the R and S relations, respectively.
- Join index records can identify joinable tuples without performing costly join operations.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Join Indexing

- Consider star schema for AllElectronics of the form sales_star [time, item, branch, location]: dollars_sold = SUM(sales_in_dollars)"

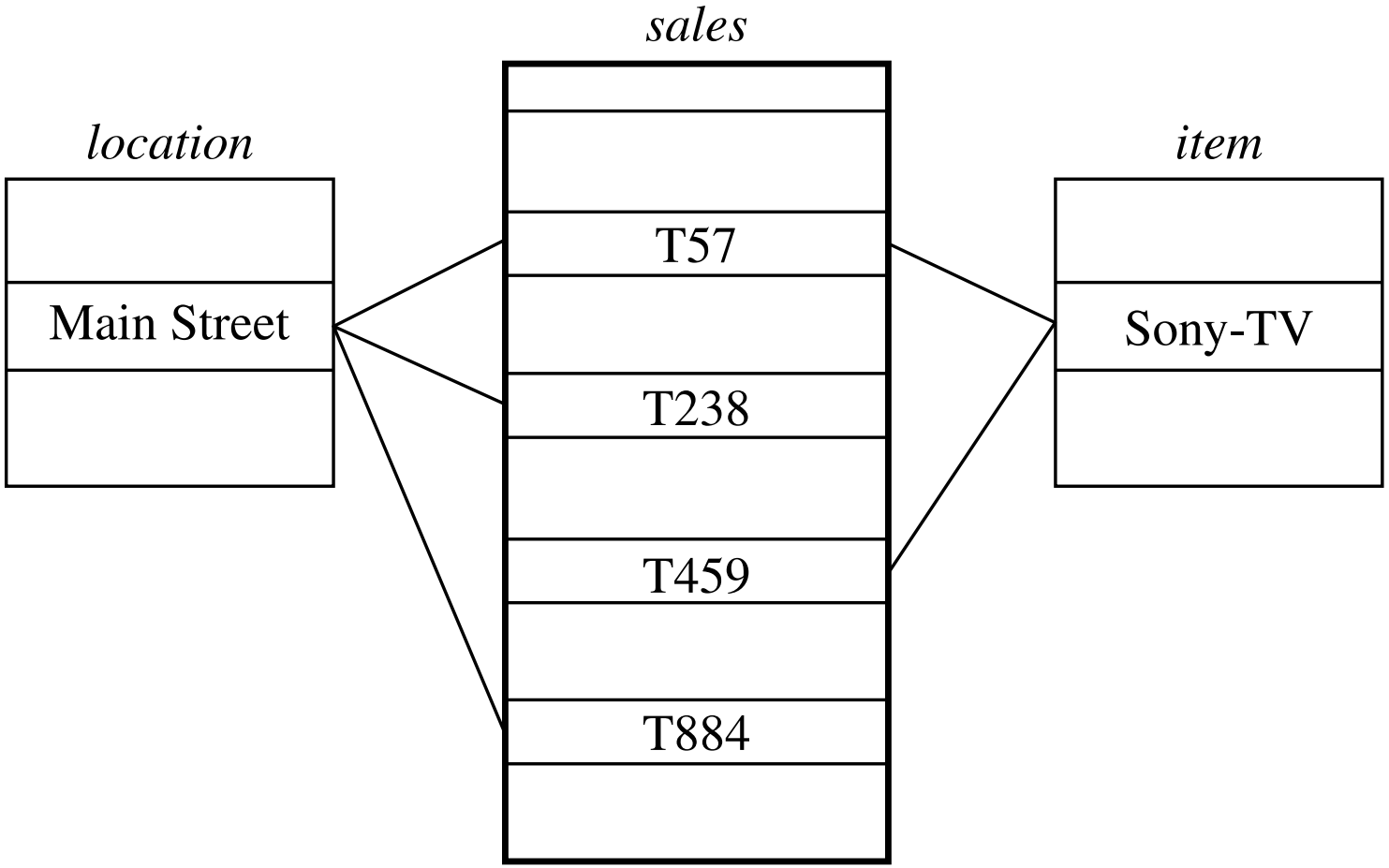


Figure 4.16 Linkages between a *sales* fact table and *location* and *item* dimension tables.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Join Indexing

- Consider star schema for AllElectronics of the form sales_star [time, item, branch, location]: dollars_sold = SUM(sales_in_dollars)

Join index table for *location/sales*

location	*sales_key*
. . .	. . .
Main Street	T57
Main Street	T238
Main Street	T884
. . .	. . .

Join index table for *item/sales*

item	*sales_key*
. . .	. . .
Sony-TV	T57
Sony-TV	T459
. . .	. . .

Join index table linking *location* and *item* to *sales*

location	*item*	*sales_key*
. . .	. . .	. . .
Main Street	Sony-TV	T57
. . .	. . .	. . .

Figure 4.17 Join index tables based on the linkages between the *sales* fact table and the *location* and *item* dimension tables shown in Figure 4.16.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Join Indexing

RowID	BookID	Title	Genre	. . .
1	7493	Tropical Food	Cooking	. . .
2	9436	Winnie the Pooh	Childrens' books	. . .
3	9948	Gone With the Wind	Fiction	. . .
4	9967	Italian Food	Cooking	. . .

Book (dimension table)

RowID	BookID	ShopID	SalesID	DayID	Count	Price
1	9436	854	1021	2475	2	30
2	7493	854	1021	2475	1	20
3	9948	876	2098	3456	1	20
4	7493	876	2231	3456	2	40
5	7493	876	3049	2475	1	20
6	9436	854	3362	3569	2	30
7	9967	731	3460	3569	1	35
8	7493	731	3460	3569	1	15
9	9948	731	3460	3569	1	15

Sales (fact table)

Book_RowID	*Sales_RowID*
1	⟨*2, 4, 5, 8*⟩
2	⟨*1, 6*⟩
3	⟨*3, 9*⟩
4	⟨*7*⟩

Join index for Book and Sales

Figure 4.3: A fact table, a dimension table, and a join index with a list of pointers

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Join Indexing

- Join indexing is especially useful for maintaining the relationship between a foreign key and its matching primary keys, from the joinable relation.
- Star schema model of data warehouses makes join indexing attractive for cross-table search, because the linkage between a fact table and its corresponding dimension tables comprises the fact table's foreign key and the dimension table's primary key.
- Join indices may span multiple dimensions to form composite join indices. We can use join indices to identify subcubes that are of interest.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Efficient Processing of OLAP Queries

- The purpose of materializing cuboids and constructing OLAP index structures is to speed up query processing in data cubes.
- Given materialized views, query processing should proceed as follows:
  1. Determine which operations should be performed on the available cuboids:
     - Transform any selection, projection, roll-up (group-by), and drill-down operations specified in the query into corresponding SQL and/or OLAP operations.
     - For example, slicing and dicing a data cube corresponds to selection and/or projection operations on a materialized cuboid.
  2. Determine to which materialized cuboid(s) the relevant operations should be applied:
     - Identifying materialized cuboids with query answering potential and selecting the cuboid with the least cost.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Efficient Processing of OLAP Queries

- Example OLAP query processing: Suppose a data cube for AllElectronics of the form sales cube [time, item, location]: sum (sales in dollars).
- Dimension hierarchies:
  - time: day < month < quarter < year
  - item: item name < brand < type
  - location: street < city < province_or_state < country
- Query needs to be processed on {brand, province_or_state} with the selection constant year = 2010
- Four materialized cuboids:
  - cuboid 1: {year, item name, city}
  - cuboid 2: {year, brand, country}
  - cuboid 3: {year, brand, province_or_state}
  - cuboid 4: {item name, province_or_state}, where year = 2010
- Which of the four cuboids should be selected for query processing?

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

ROLAP

- Relational OLAP (ROLAP) servers:
  - Intermediate servers that stand in between a relational back-end server and client front-end tools.
  - Use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces.
  - ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.
  - ROLAP technology tends to have greater scalability than MOLAP technology.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

MOLAP

- Multidimensional OLAP (MOLAP) servers:
  - These servers support multidimensional data views through array-based multidimensional storage engines.
  - They map multi-dimensional views directly to data cube array structures.
  - The advantage of using a data cube is that it allows fast indexing to precomputed summarized data.
  - Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse.
  - In such cases, sparse matrix compression techniques should be explored.
  - MOLAP servers adopt a two-level storage representation to handle dense and sparse data sets: Denser subcubes are identified and stored as array structures, whereas sparse subcubes employ compression technology for efficient storage. utilization.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

HOLAP

- Hybrid OLAP (HOLAP) servers:
  - The hybrid OLAP approach combines ROLAP and MOLAP technology.
  - Benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.
  - For example, a HOLAP server may allow large volumes of detailed data to be stored in a relational database, while aggregations are kept in a separate MOLAP store.
- Specialized SQL servers:
  - Provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Hadoop Distributed File System — HDFS

- HDFS is a filesystem designed for storing very large files with streaming data access patterns, running on clusters of commodity hardware.
- Very large files:
  - Files that can be upto petabytes of data.
- Streaming data access:
  - Data processing pattern is write-once, read-many-times.
  - Time to read the whole dataset is more important than the latency in reading the first record.
- Commodity hardware:
  - Designed to run on clusters of commodity hardware.
  - Chance of node failure across the cluster is high, at least for large clusters.
  - HDFS is designed to carry on working without a noticeable interruption to the user in the face of such failure.

All from: Tom White. "Hadoop: The Definitive Guide", 4th Edition, 2015, O'Reilly Media, Inc.

HBase (BigTable)

- HBase is a distributed column-oriented database built on top of HDFS.
- HBase is the Hadoop application to use when you require real-time read/write random access to very large datasets.

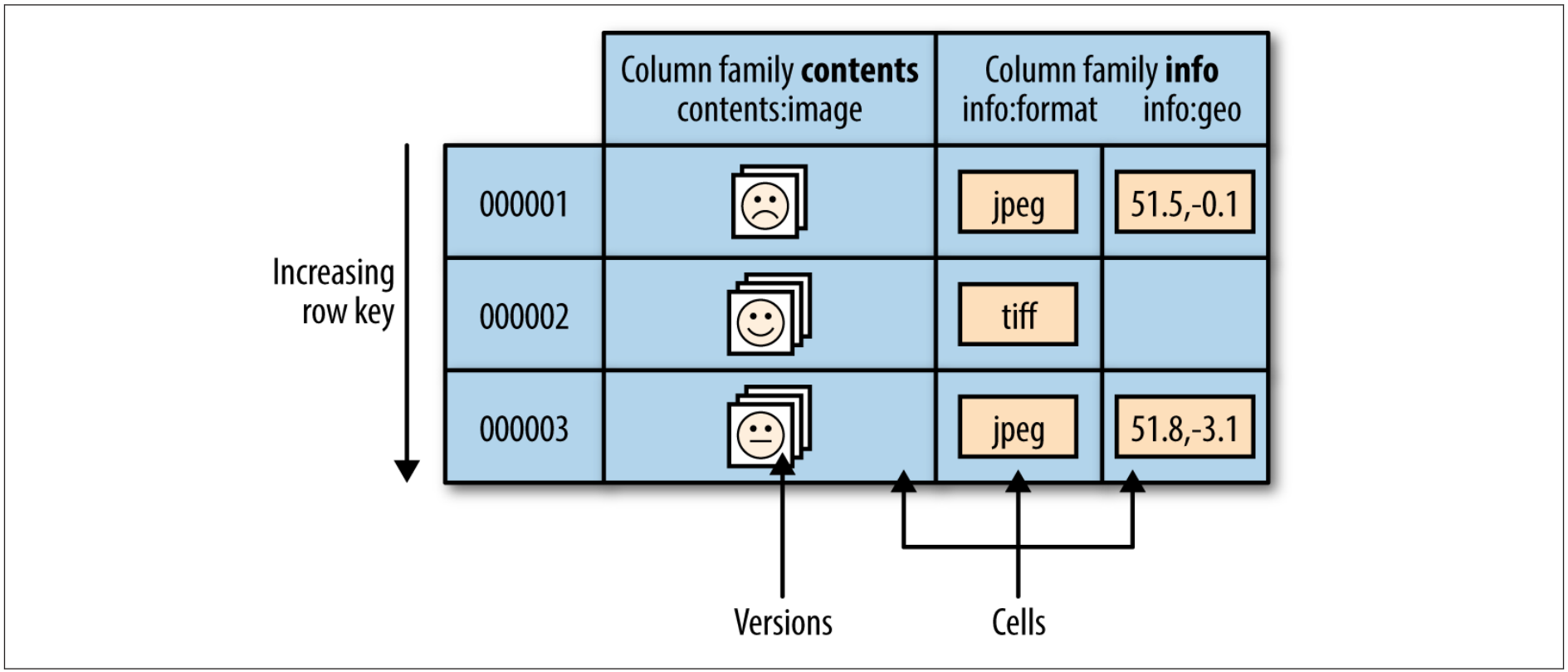


Figure 20-1. The HBase data model, illustrated for a table storing photos

All from: Tom White. "Hadoop: The Definitive Guide", 4th Edition, 2015, O'Reilly Media, Inc.

Good Talk by Prof. Micheal Stonebreaker

- Michael Stonebreaker — Big Data is (at least) Four Different Problems.
- 'At least 90% of time spent by data scientists is done finding and cleaning the data!'.
- Link: `https://www.youtube.com/watch?v=KRcecxdGxvQ`

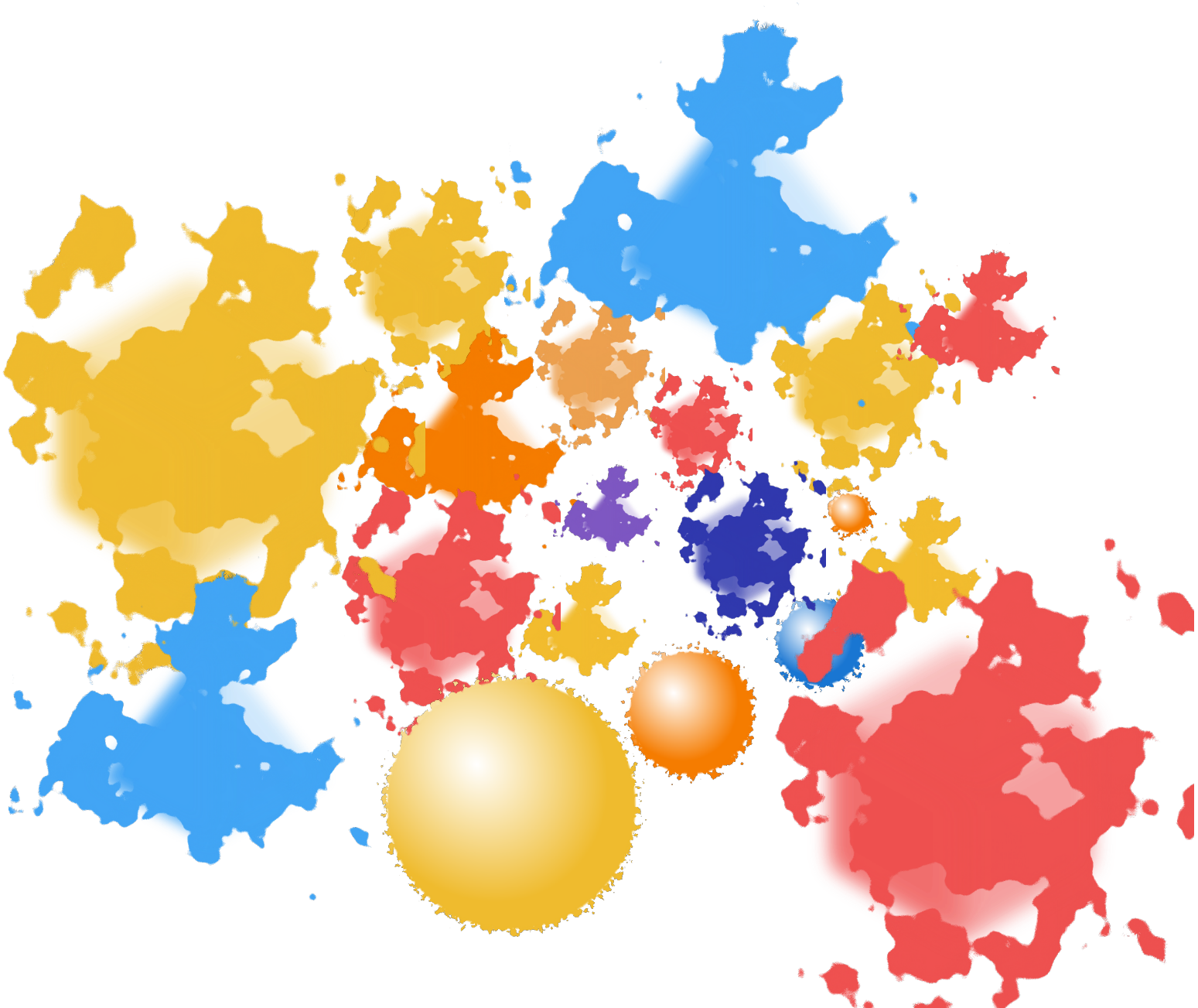
1. QA and References
   - Administrative
   - References for Today's Lecture

2. Data Warehouse Implementation
   - Indexing OLAP Data
   - OLAP Server Architectures
   - Web-Scale Indexing

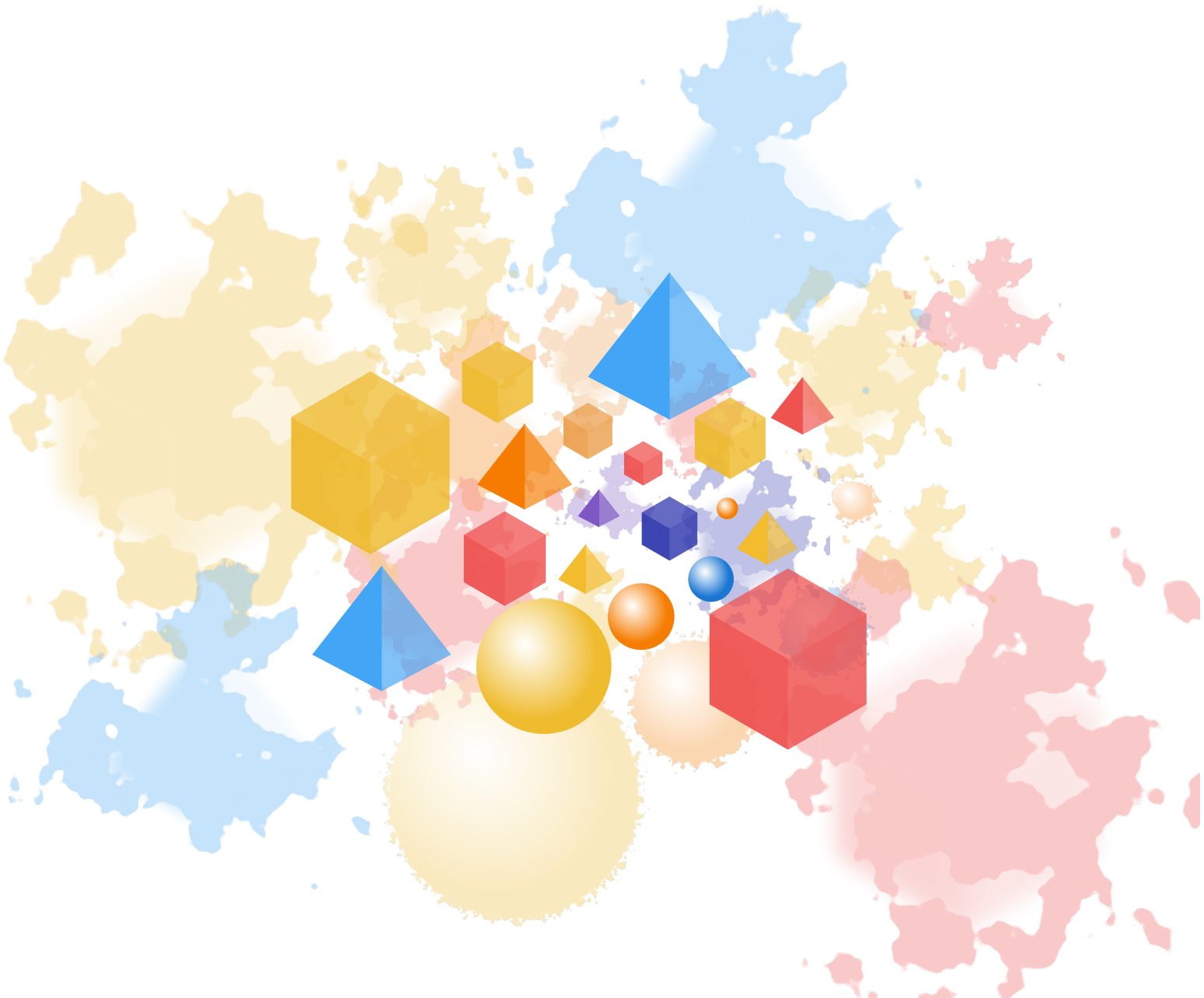
3. Data
   - Data
   - Data Modeling
   - Attributes
   - Data Characteristics
   - Data Set Types

4. Summary

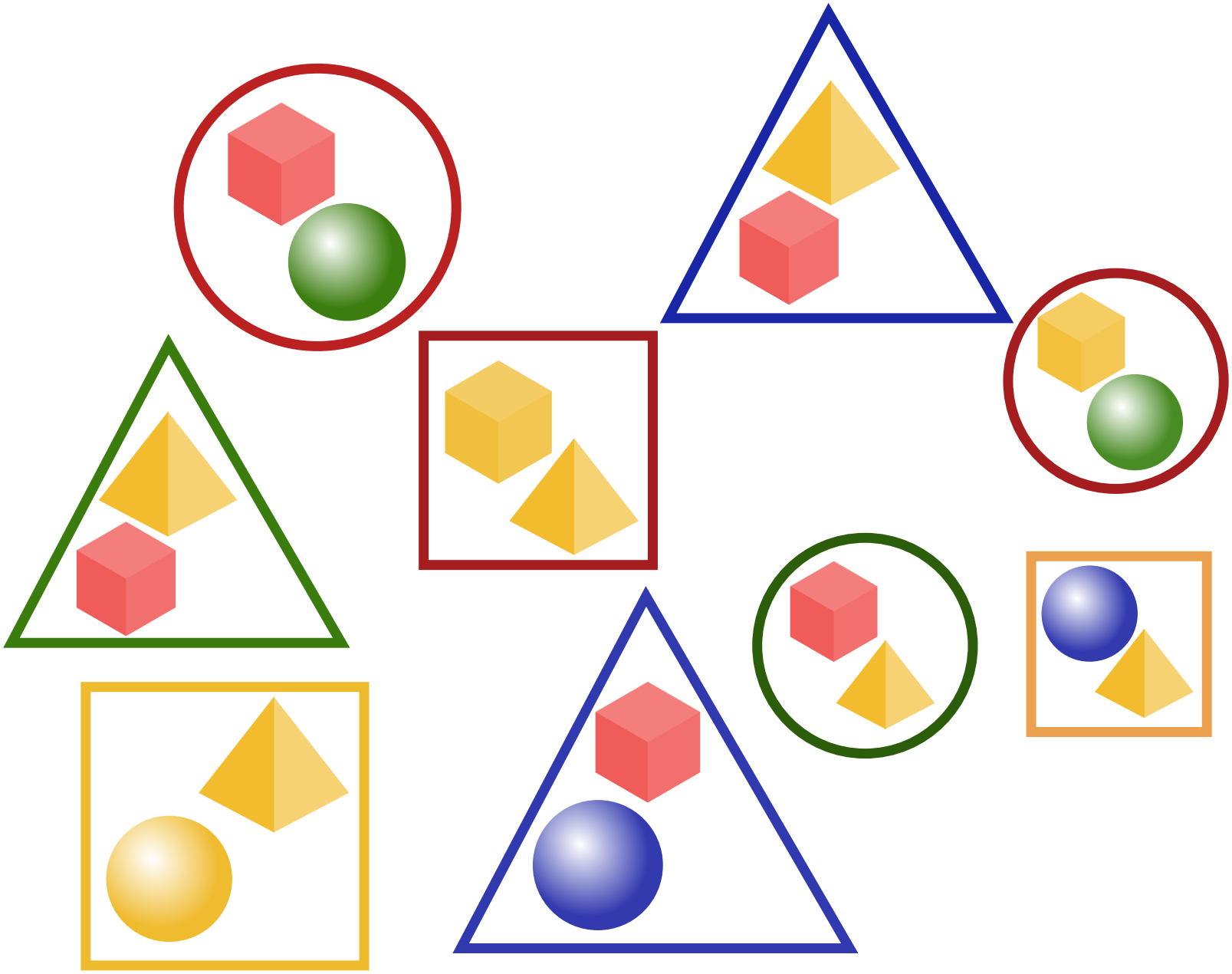
Real World (Abstract)



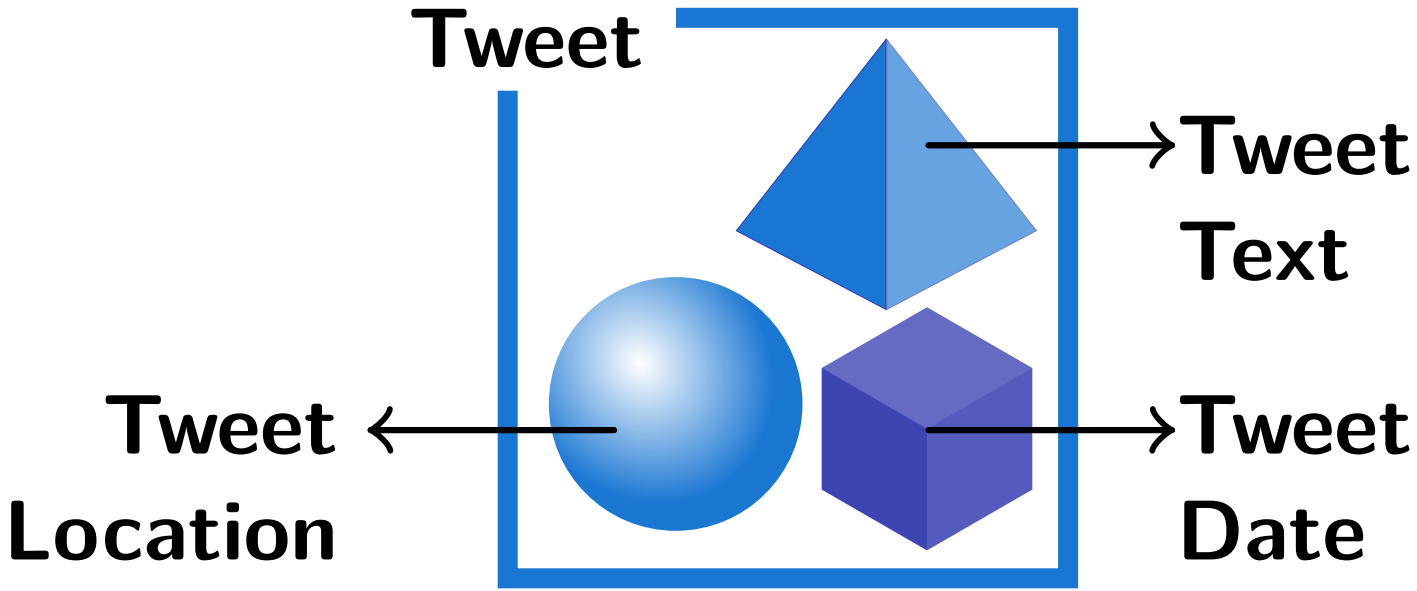
Data in the Real World (Abstract)



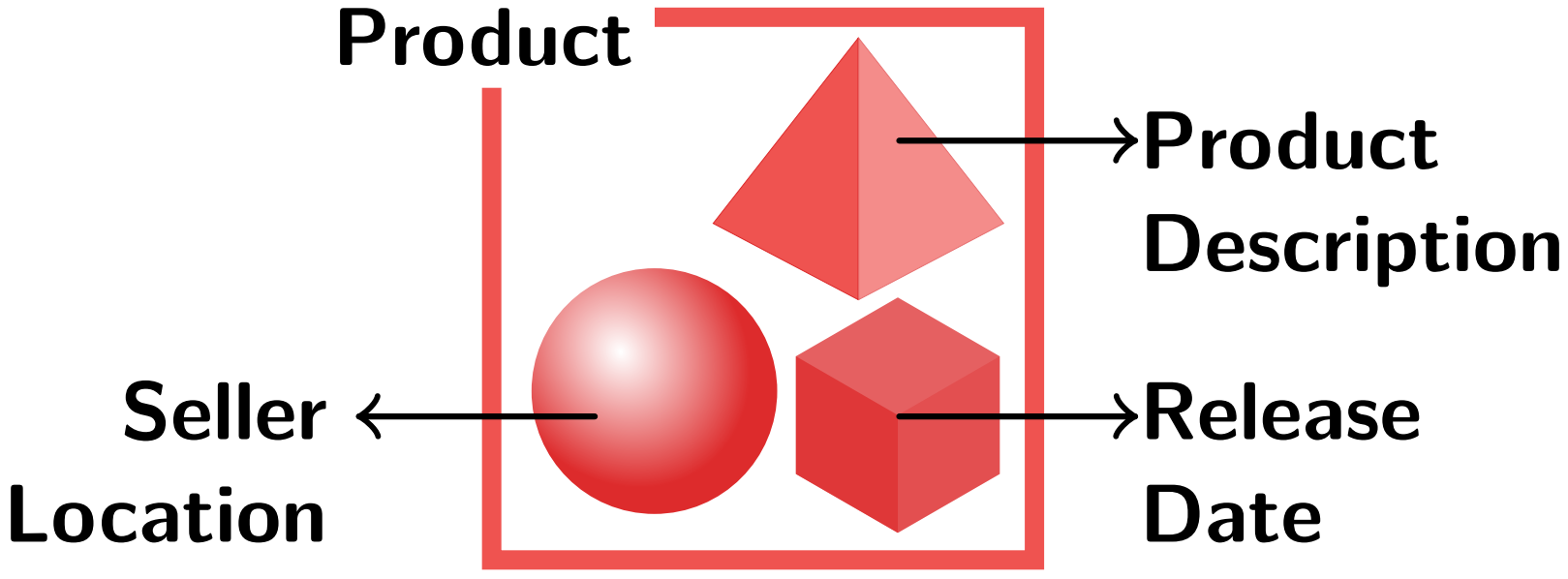
Data — Observations using Measurements (Abstract)



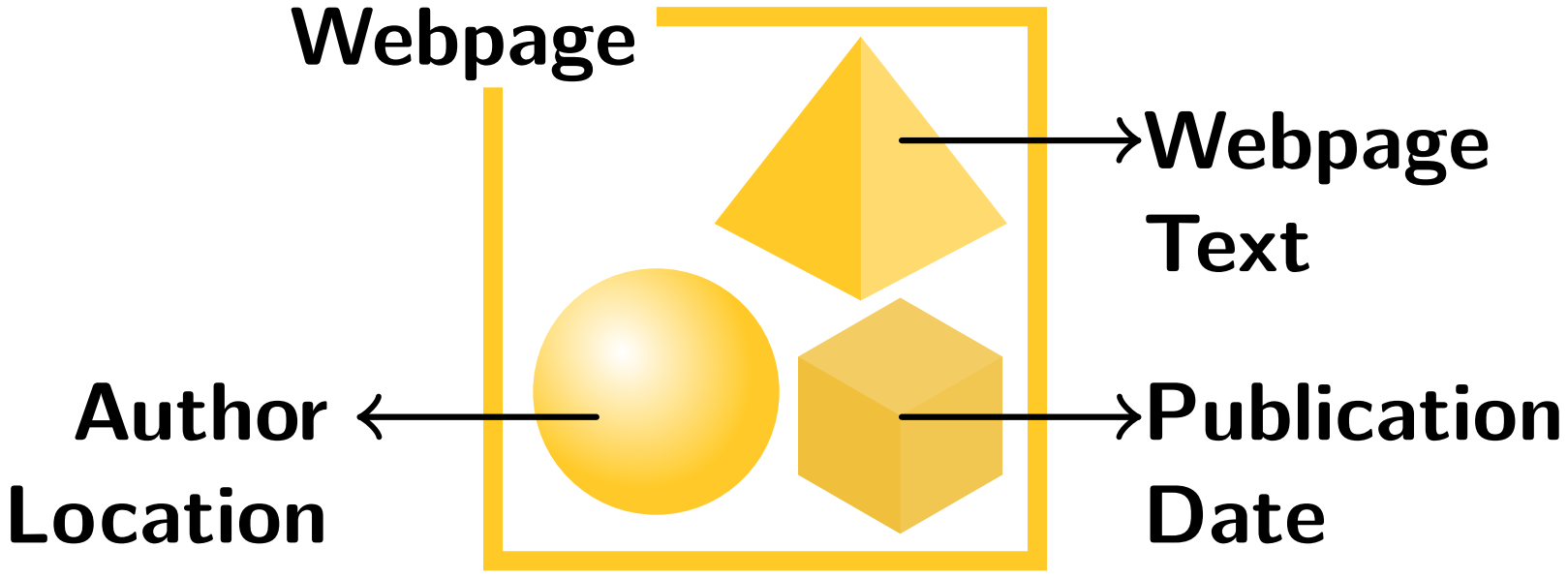
Data — Tweets Example



Data — Amazon Products Example

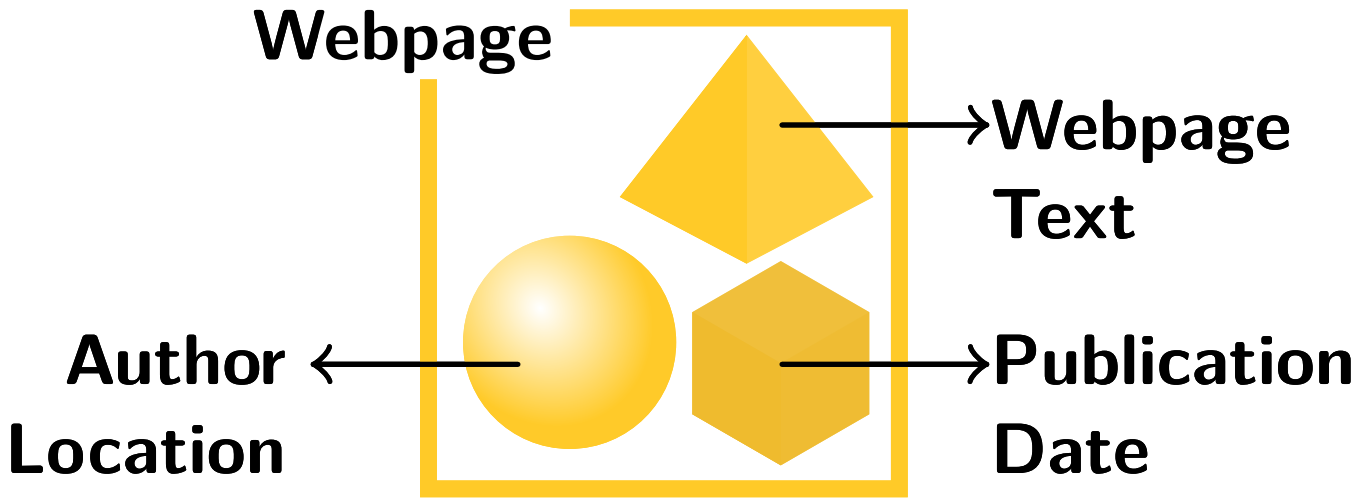


Data — Google Webpages Example



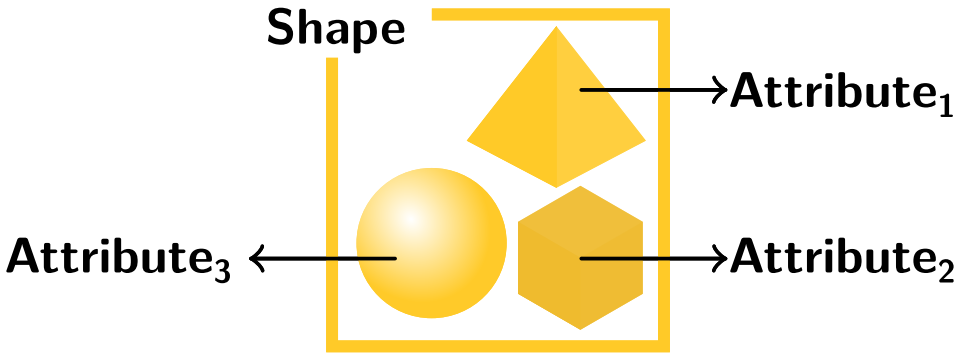
Data — Modeling and Storage Example

```
Class Webpage{
 String text; // Webpage Text
 Date   pubDate; // Publication Date
 Geo    location; // Publisher Location
}
```



Data — Modeling and Storage (Abstract)

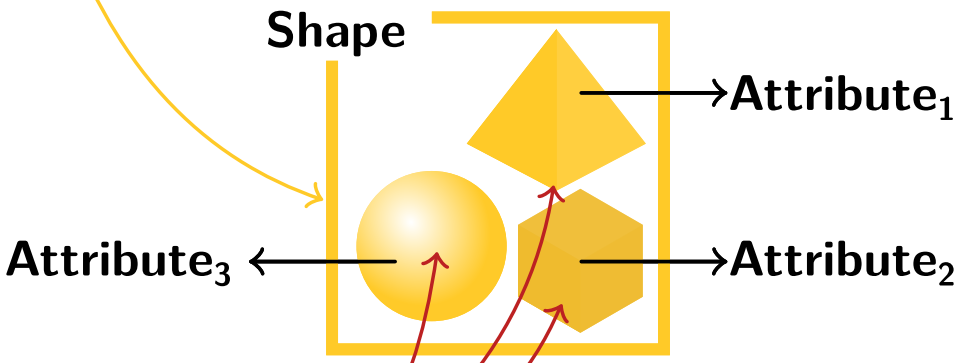
```
Class Data {
 String shape;
 Color  shapeColor;
 String[] objects;
 Color[] objectColors;
}
```



Data — Modeling and Storage (Abstract)

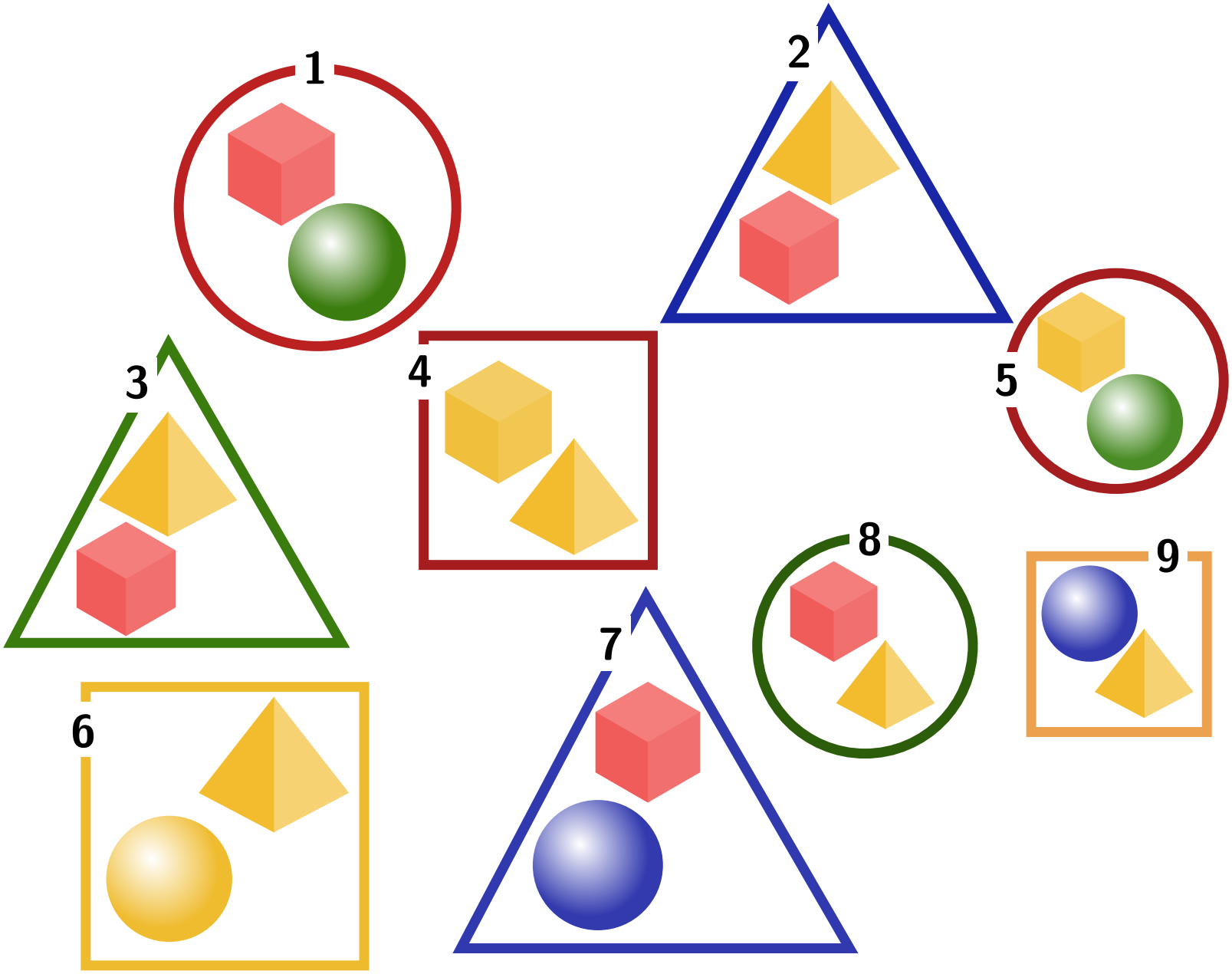
Object

```
Class Data {
  String shape;
  Color  shapeColor;
  String[] objects;
  Color[] objectColors;
}
```



Attributes

Data — Representation



Data — Representation

ID	Shape	Shape Color	Object_1	Color_1	Object_2	Color_2
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

Data — Representation

Attributes

Objects

ID	Shape	Shape Color	$Object_1$	$Color_1$	$Object_2$	$Color_2$
1	Circle	Red	Cube	Red	Sphere	Green
2	Triangle	Blue	Pyramid	Yellow	Cube	Red
3	Triangle	Green	Pyramid	Yellow	Cube	Red
4	Square	Red	Cube	Yellow	Pyramid	Yellow
5	Circle	Red	Cube	Yellow	Sphere	Green
6	Square	Yellow	Pyramid	Yellow	Sphere	Yellow
7	Triangle	Blue	Cube	Red	Sphere	Blue
8	Circle	Green	Cube	Red	Pyramid	Yellow
9	Square	Yellow	Sphere	Blue	Pyramid	Yellow

Data

1. Collection of observations (that are data objects and their attributes).
2. An attribute is a property or characteristic of an object.
   - Attribute is also known as variable, field, characteristic, dimension, or feature.
   - Examples: eye color of a person, temperature, etc.
3. A collection of attributes describe an object.
   - Object is also known as record, point, case, sample, entity, or instance.

Attributes

Objects

ID	Shape	Shape Color
1	Circle	Red
2	Triangle	Blue
3	Triangle	Green
4	Square	Red
5	Circle	Red
6	Square	Yellow
7	Triangle	Blue
8	Circle	Green
9	Square	Yellow

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Measurements

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness ($=, \neq$).
  - Order ($>, <$).
  - Differences are meaningful ($+, -$).
  - Ratios are meaningful ($\times, \div$).
- Nominal attribute: distinctness.
- Ordinal attribute: distinctness and order.
- Interval attribute: distinctness, order and meaningful differences.
- Ratio attribute: all 4 properties/operations.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Measurements

	Type	Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values.	If all employee ID numbers are reassigned, it will not make any difference.
	Ordinal	An order-preserving change of values, i.e., $new\_value = f(old\_value)$ where f is a monotonic function.	An attribute encompassing the notion of {good, better, best} can be represented equally well by the values {1,2,3} or by {0.5,1,10}.
Numeric (Quantitative)	Interval	$new\_value = a * old\_value + b$, a and b are constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree (unit).
	Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Measurements

	Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$.	ZIP codes, employee ID numbers, eye color, gender.	Mode, entropy, contingency correlation, χ^2 test.
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(>, <)$.	Hardness of minerals, {good,better,best}, grades, street numbers.	Median, percentiles, rank correlation, run tests, sign tests.
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$.	Calendar dates, temperature in Celsius or Fahrenheit.	Mean, standard deviation, Pearson's correlation, t and F tests.
	Ratio	For ratio variables, both differences and ratios are meaningful. $(\times, \div)$.	Temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current.	Geometric mean, harmonic mean, percent variation.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Number of Values

- Discrete Attribute:
  - Has only a finite or countably infinite set of values.
  - Examples: zip codes, counts, or the set of words in a collection of documents.
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes.
- Continuous Attribute:
  - Has real numbers as attribute values.
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Often represented as floating-point variables.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Importance

- Binary variable has either two outcomes: 1 (positive/present) or 0 (negative/absent).
- If there is no preference for which outcome should be coded as 0 and which as 1, the binary variable is called symmetric.
- Example: binary variable for "is evergreen?". A plant has the possible states "loses leaves in winter" and "does not lose leaves in winter."
- Both are equally valuable and carry the same weight when similarity is computed.

All from:
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_distance_sect003.htm

Attribute Types — Based on Importance

- If the outcomes of a binary variable are not equally important, the binary variable is called asymmetric.
- Presence or absence of a relatively rare attribute.
- Example: binary variable for "is color-blind?". Two people being color-blind is an important observation versus two people who are not color-blind.
- The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).
- Agreement of two 1's is more significant than the agreement of two 0's.
- Usually, the negative match is treated as irrelevant.

All from:
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_distance_sect003.htm

Data Characteristics

1. Dimensionality
2. Sparsity
3. Resolution

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Characteristics — Dimensionality

- Dimensionality (number of attributes):
  - Data with a small number of dimensions tends to be qualitatively different than moderate or high-dimensional data.
  - Curse of Dimensionality: Difficulties associated with analyzing high-dimensional data.
  - Curse of Dimensionality can be overcome with the help of dimensionality reduction techniques.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Characteristics — Sparsity

- Sparsity:
  - For some data sets (e.g., with asymmetric features) fewer than 1% of the entries are non-zero.
  - Sparsity is an advantage: only non-zero values need to be stored and manipulated.
  - Results in significant savings with respect to computation time and storage.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Characteristics — Resolution

- Resolution:
  - Patterns in the data depend on the level of resolution.
  - Properties of data are different at different resolutions.
  - Fine Resolution: a pattern may not be visible or may be buried in noise.
  - Coarse Resolution: pattern may disappear.
  - Example: Variations in atmospheric pressure on a scale of hours reflect the movement of storms and other weather systems. On a scale of months, such phenomena are not detectable.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Set Types

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Record Data

- Data mining algorithms assume the data set is a collection of records (data objects).
- Each record consists of a fixed set of data fields (attributes).
- Often there is no explicit relationship among records or data fields.
- Every record (object) has the same set of attributes.
- Record data is usually stored either in flat files or in relational databases.
- Database serves as a convenient place to find records.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Record Data Examples — Transaction Data

- Each record (transaction) involves a set of items.
- Can be viewed as a set of records whose fields are asymmetric attributes.
- Often attributes are binary (item was purchased / not purchased), can also be discrete or continuous (# items purchased or $ spent on items).
- Grocery Store Example:
  - Set of products purchased by a customer in one transaction.
  - Individual products are items.
  - Also known as market basket data.

T-ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Record Data Examples — Data Matrix

- If all attributes are numeric, then the records can be thought of as points (vectors) in a multidimensional space.
- That is, a *m* by *n* matrix, where there are *m* rows (data objects) and *n* columns (numeric attributes).
- Data matrix is the standard data format for most statistical data.

x-Load	y-Load	Distance	Load	Thickness
10.26	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Record Data Examples — Document Data

- Each document becomes a term vector.
- Each term is a component (attribute) of the vector.
- The value of each component is the number of times the corresponding term occurs in the document.
- Example of sparse data matrix.

	$\mathbf{word}_1$	$\mathbf{word}_2$	$\mathbf{word}_3$	...	$\mathbf{word}_{\lvert\mathcal{V}\rvert}$
$\mathbf{document}_1$	10	0	0	...	5
$\mathbf{document}_2$	1	2	0	...	0
⋮	⋮	⋮	⋮	...	⋮
$\mathbf{document}_n$	0	1	0	...	0

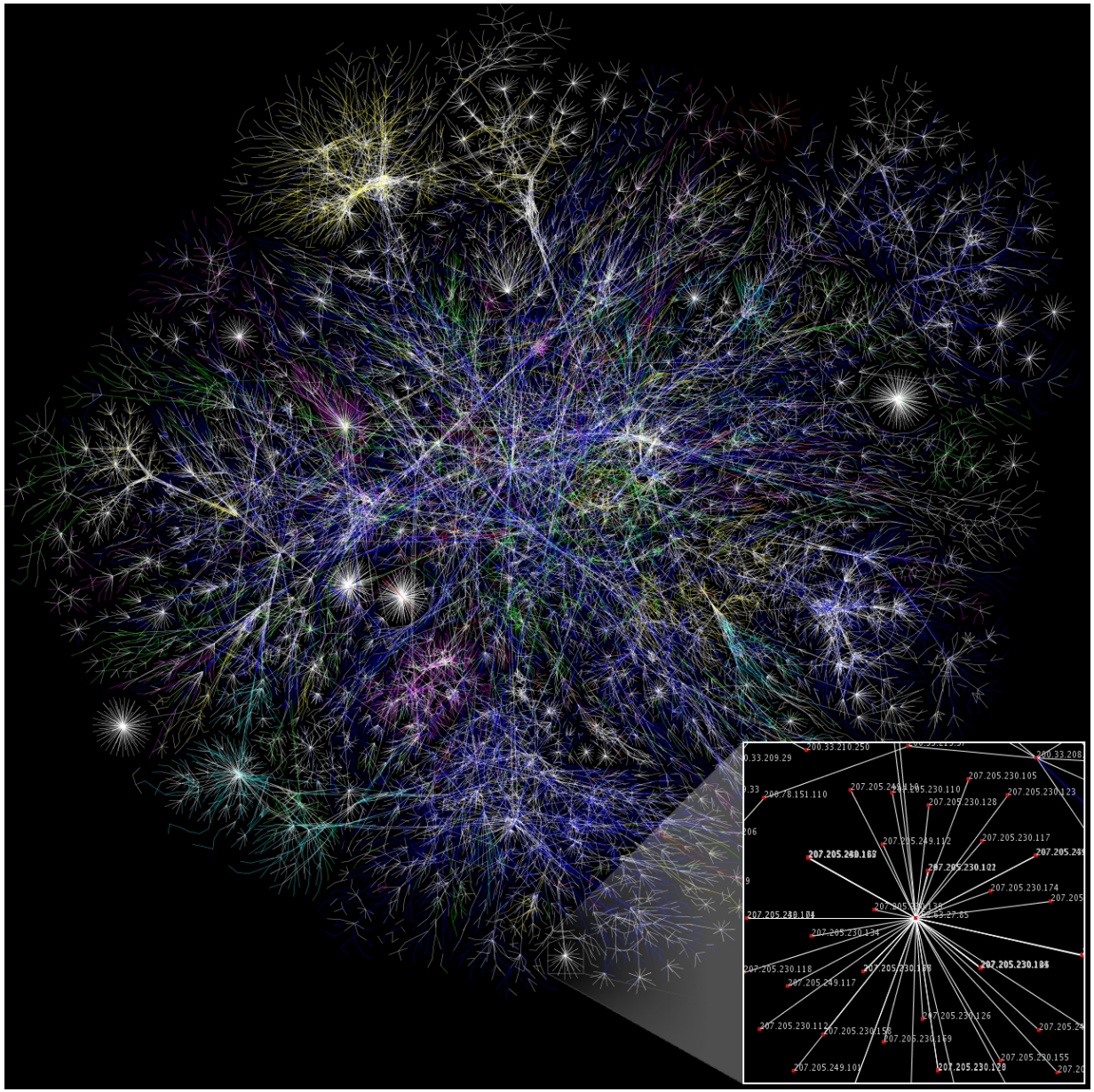
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Graph Data

- Traditional paradigm in data analysis assumes independence between data instances.
- Oftendata instances (nodes) may be connected or linked together via relationships (edges).
- What emerges is a network or graph.
- Both the nodes and edges in the graph may have several attributes that may be numerical or categorical, or even more complex (e.g., time series data).
- Examples:
  1. World Wide Web (with its Web pages and hyperlinks)
  2. Social Networks (wikis, blogs, tweets etc.)
  3. Semantic Networks (ontologies)
  4. Citation Networks for scientific literature
  5. Biological Networks (protein interactions, gene regulation networks, metabolic pathways)

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

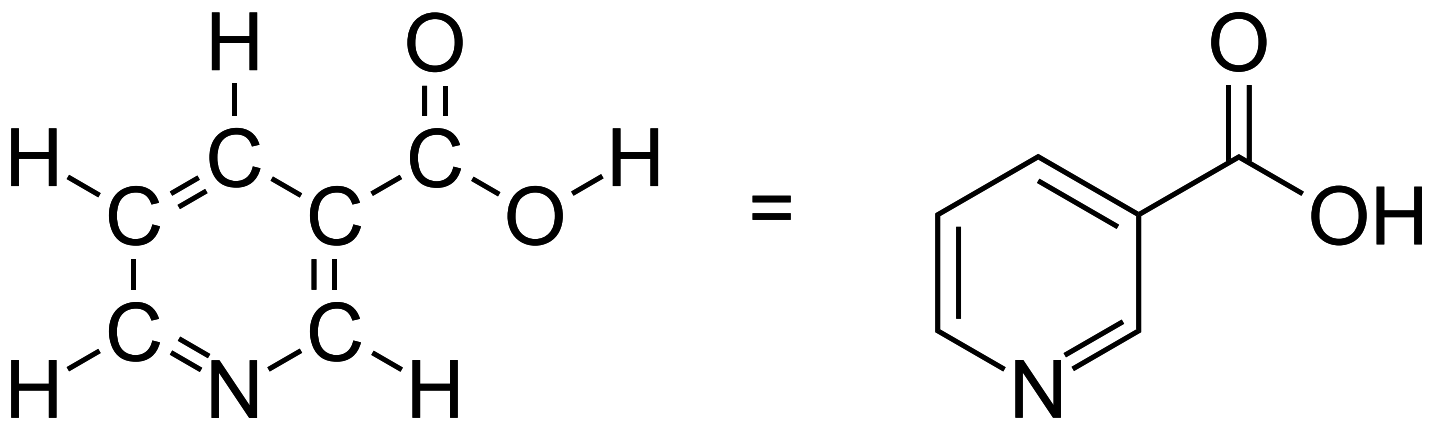
Graph Data



Source: https://commons.wikimedia.org/wiki/File:Internet_map_1024.jpg

Graph Data

- According to Tan et al. graph data can be of two types:
  1. Data with Relationships among Objects.
  2. Data with Objects that are Graphs.
     - Example: chemical compound structures can be represented by a graph.
     - Substructure Mining: determine which substructures occur frequently in a set of compounds ascertain chemical properties (e.g., melting point).



C_5H_4N-3-CO_2H = pyridine-3-carboxylic acid = niacin = vitamin B_3

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.
Image source: https://commons.wikimedia.org/wiki/File:OrgNom.svg

Ordered Data

- For some types of data, the attributes have relationships that involve order in time or space.
  - Sequential Data
  - Sequence Data
  - Time Series Data
  - Spatial Data

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Ordered Data — Sequential Data

- Sequential data, also referred to as temporal data.
- Each record has a time associated with it.

Time	Customer	Items Purchased
t1	C1	{A,B}
t2	C3	{A,C}
t2	C1	{C,D}
t3	C2	{A,D}
t4	C2	{E}
t5	C1	{A,E}

Customer	Time and Items Purchased
C1	⟨ (t1: {A,B}), (t2:{C,D}), (t5:{A,E})⟩
C2	⟨ (t3: {A,D}), (t4: {E}) ⟩
C3	⟨ (t2: {A,C}) ⟩

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Ordered Data — Sequence Data

- Sequence of individual entities, such as a sequence of words or letters.
- Similar to sequential data, except that there are no time stamps; instead, there are positions in an ordered sequence.

```
   1 ATTAAAGGTT TATACCTTCC CAGGTAACAA ACCAACCAAC TTTCGATCTC TTGTAGATCT
  61 GTTCTCTAAA CGAAATTTAA AATCTGTGTG GCTGTCACTC GGCTGCATGC TTAGTGCACT
 121 CACGCAGTAT AATTAATAAC TAATTACTGT CGTTGACAGG ACACGAGTAA CTCGTCTATC
 181 TTCTGCAGGC TGCTTACGGT TTCGTCCGTG TTGCAGCCGA TCATCAGCAC ATCTAGGTTT
 241 CGTCCGGGTG TGACCGAAAG GTAAGATGGA GAGCCTTGTC CCTGGTTTCA ACGAGAAAAC
 301 ACACGTCCAA CTCAGTTTGC CTGTTTTACA GGTTCGCGAC GTGCTCGTAC GTGGCTTTGG
 361 AGACTCCGTG GAGGAGGTCT TATCAGAGGC ACGTCAACAT CTTAAAGATG GCACTTGTGG
 421 CTTAGTAGAA GTTGAAAAAG GCGTTTTGCC TCAACTTGAA CAGCCCTATG TGTTCATCAA
 481 ACGTTCGGAT GCTCGAACTG CACCTCATGG TCATGTTATG GTTGAGCTGG TAGCAGAACT
 541 CGAAGGCATT CAGTACGGTC GTAGTGGTGA GACACTTGGT GTCCTTGTCC CTCATGTGGG
 601 CGAAATACCA GTGGCTTACC GCAAGGTTCT TCTTCGTAAG AACGGTAATA AAGGAGCTGG
 661 TGGCCATAGT TACGGCGCCG ATCTAAAGTC ATTTGACTTA GGCGACGAGC TTGGCACTGA
 721 TCCTTATGAA GATTTTCAAG AAAACTGGAA CACTAAACAT AGCAGTGGTG TTACCCGTGA
 781 ACTCATGCGT GAGCTTAACG GAGGGGCATA CACTCGCTAT GTCGATAACA ACTTCTGTGG
 841 CCCTGATGGC TACCCTCTTG AGTGCATTAA AGACCTTCTA GCACGTGCTG GTAAAGCTTC
 901 ATGCACTTTG TCCGAACAAC TGGACTTTAT TGACACTAAG AGGGGTGTAT ACTGCTGCCG
 961 TGAACATGAG CATGAAATTG CTTGGTACAC GGAACGTTCT GAAAAGAGCT ATGAATTGCA
1021 GACACCTTTT GAAATTAAAT TGGCAAAGAA ATTTGACACC TTCAATGGGG AATGTCCAAA
 ...        ...        ...        ...        ...        ...        ...
```

Figure: SARS-CoV-2 Complete Genome https://www.ncbi.nlm.nih.gov/nuccore/MT049951.1/

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Ordered Data — Time Series Data

- Time series data is a special type of sequential data in which each record is a time series, i.e., a series of measurements taken over time.
- Temporal Autocorrelation: if two measurements are close in time, then the values of those measurements are often very similar.

Figure: GameStop Stock Price

https://www.nasdaq.com/market-activity/stocks/gme/advanced-charting?timeframe=5D

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Ordered Data — Spatial Data

- Some objects have spatial attributes, such as positions or areas, as well as other types of attributes.
- Spatial Autocorrelation: objects that are physically close tend to be similar in other ways as well.

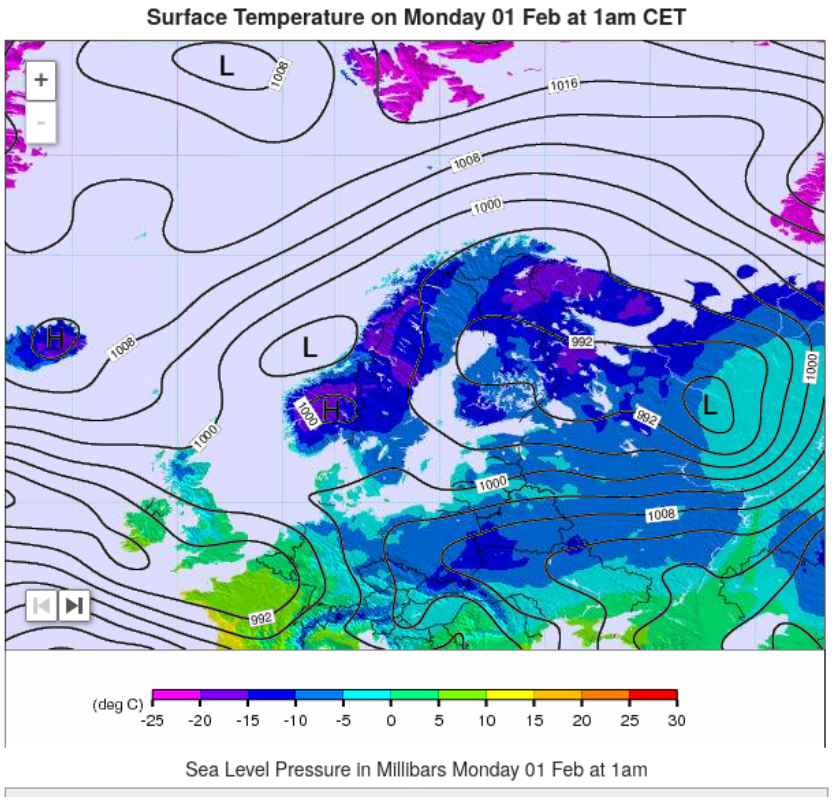


Figure: Trondheim Weather Map https://www.weather-forecast.com/maps/Norway?symbols=none&type=lapse

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1. QA and References
   - Administrative
   - References for Today's Lecture

2. Data Warehouse Implementation
   - Indexing OLAP Data
   - OLAP Server Architectures
   - Web-Scale Indexing

3. Data
   - Data
   - Data Modeling
   - Attributes
   - Data Characteristics
   - Data Set Types

4. Summary

Summary

RowID	BookID	Title	Binding	Language	...
1	9436	Winnie the Pooh	Hardcover	English	...
2	1029	Le Petit Prince	Paperback	French	...
3	8733	Alice in Wonderland	Paperback	English	...
4	2059	Wind in the Willows	Hardcover	English	...
5	5995	A Bear Called Paddington	Paperback	English	...
6	1031	Pierre Lapin	Hardcover	French	...
7	3984	Le avventure di Pinocchio	Hardcover	Italian	...

Figure 4.2: Book dimension table also showing the special system-maintained attribute RowID, which normally cannot be seen by the user

Bitmap for `Binding`

```
Hardcover: 1001011
Paperback: 0110100
```

Bitmap for `Language`

```
English:  1011100
French:   0100010
Italian:  0000001
```

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Summary

RowID	BookID	Title	Genre	. . .
1	7493	Tropical Food	Cooking	. . .
2	9436	Winnie the Pooh	Childrens' books	. . .
3	9948	Gone With the Wind	Fiction	. . .
4	9967	Italian Food	Cooking	. . .

Book (dimension table)

RowID	BookID	ShopID	SalesID	DayID	Count	Price
1	9436	854	1021	2475	2	30
2	7493	854	1021	2475	1	20
3	9948	876	2098	3456	1	20
4	7493	876	2231	3456	2	40
5	7493	876	3049	2475	1	20
6	9436	854	3362	3569	2	30
7	9967	731	3460	3569	1	35
8	7493	731	3460	3569	1	15
9	9948	731	3460	3569	1	15

Sales (fact table)

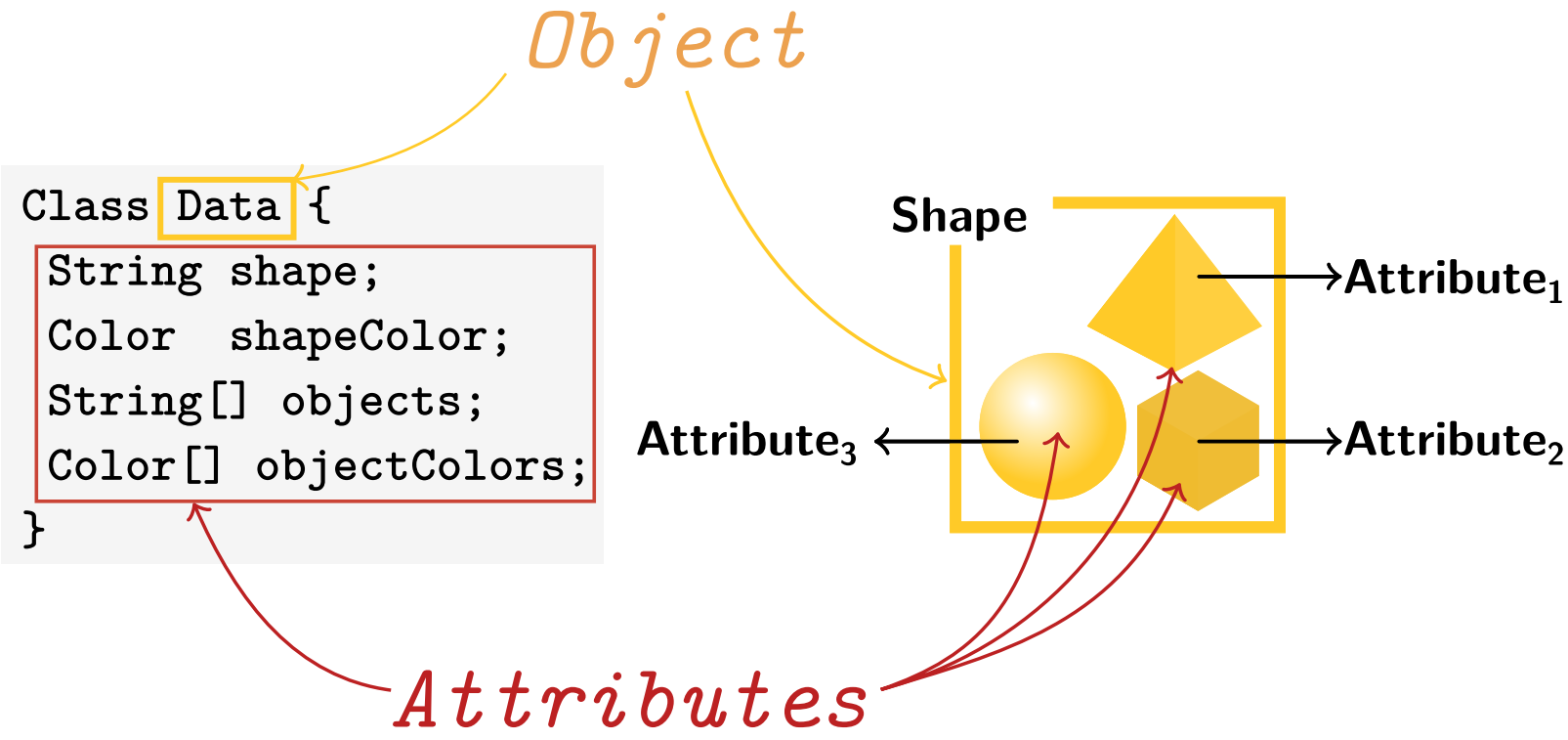
Book_RowID	*Sales_RowID*
1	⟨*2, 4, 5, 8*⟩
2	⟨*1, 6*⟩
3	⟨*3, 9*⟩
4	⟨*7*⟩

Join index for Book and Sales

Figure 4.3: A fact table, a dimension table, and a join index with a list of pointers

All from: Jensen et al. "Multidimensional Databases and Data Warehousing", 2010, Morgan Claypool Publishers.

Summary


Object
Class Data {
String shape;
Color shapeColor;
String[] objects;
Color[] objectColors;
}
Shape
Attribute1
Attribute2
Attribute3
Attributes

Summary

	Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$.	ZIP codes, employee ID numbers, eye color, gender.	Mode, entropy, contingency correlation, χ^2 test.
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(>, <)$.	Hardness of minerals, {good,better,best}, grades, street numbers.	Median, percentiles, rank correlation, run tests, sign tests.
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$.	Calendar dates, temperature in Celsius or Fahrenheit.	Mean, standard deviation, Pearson's correlation, t and F tests.
	Ratio	For ratio variables, both differences and ratios are meaningful. $(\times, \div)$.	Temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current.	Geometric mean, harmonic mean, percent variation.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Summary

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Summary

1. First Assignment
   - Available this week (26.Jan.2023) and due by 09.February.2023.
   - There will be a tutorial next week on Monday (30.Jan.2023) in the lecture slot to introduce the first assingment.
2. Volunteers for feedback regarding course
   - Interested? Please contact me by email!