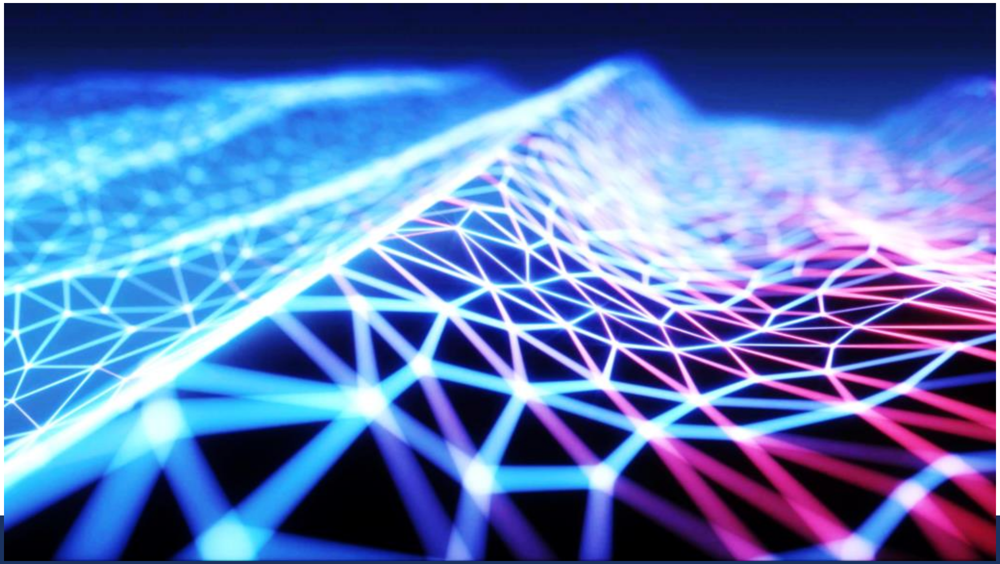


DEFINITION OF DATA

ARRIEL BENIS, PHD, HOLON INSTITUTE OF TECHNOLOGY, ISRAEL

Marginally modified by

Prof. Stefan Darmoni, Rouen University hospital & limics INSERM U1142, Sorbonne université, France

Credits: Introduction to Data Mining by Tan, Steinbach, Kumar (2004)

ATTRIBUTES AND OBJECTS
WHAT IS DATA?

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

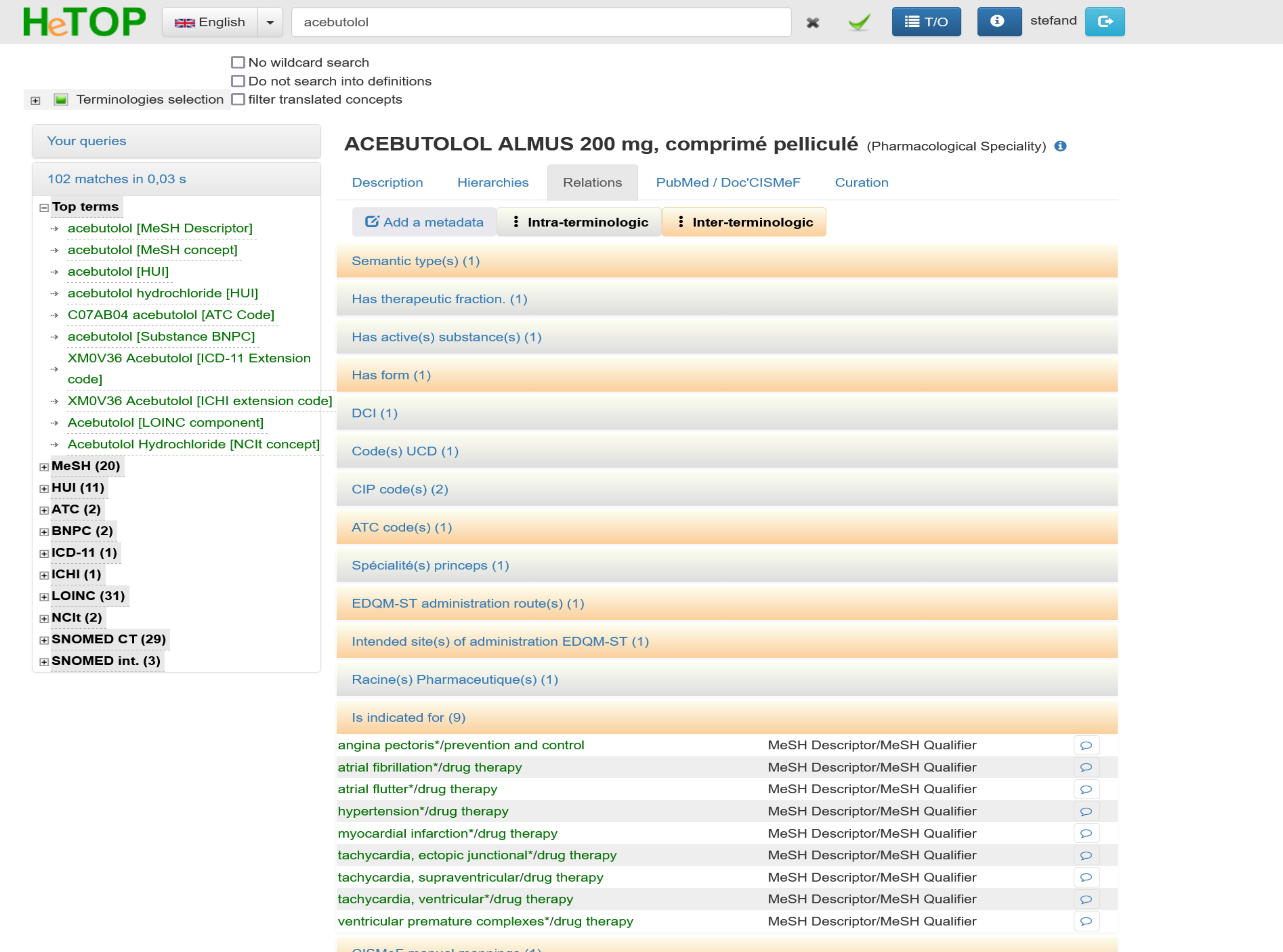
Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

DATA, INFORMATION, KNOWLEDGE

- Affectation of an attribute for any specific data generates **information**
  - E.g. Temperature = 38°… Celsius, not F, (US) not K (International)
- Conditional rule on an information generates **knowledge**
  - E.g. if Temperature > 38°C then Fever
- Semantic triplet also generates **knowledge**
  - Two concepts linked by a specific relation
  - Cat IS-A a mammal



☐ No wildcard search
☐ Do not search into definitions
☐ filter translated concepts

⊞ Terminologies selection

Your queries

102 matches in 0,03 s

⊟ **Top terms**
- acebutolol [MeSH Descriptor]
- acebutolol [MeSH concept]
- acebutolol [HUI]
- acebutolol hydrochloride [HUI]
- C07AB04 acebutolol [ATC Code]
- acebutolol [Substance BNPC]
- XM0V36 Acebutolol [ICD-11 Extension code]
- XM0V36 Acebutolol [ICHI extension code]
- Acebutolol [LOINC component]
- Acebutolol Hydrochloride [NCIt concept]

⊞ **MeSH (20)**
⊞ **HUI (11)**
⊞ **ATC (2)**
⊞ **BNPC (2)**
⊞ **ICD-11 (1)**
⊞ **ICHI (1)**
⊞ **LOINC (31)**
⊞ **NCIt (2)**
⊞ **SNOMED CT (29)**
⊞ **SNOMED int. (3)**

ACEBUTOLOL ALMUS 200 mg, comprimé pelliculé (Pharmacological Speciality)

Description | Hierarchies | Relations | PubMed / Doc'CISMeF | Curation

Add a metadata | **Intra-terminologic** | **Inter-terminologic**

Semantic type(s) (1)

Has therapeutic fraction. (1)

Has active(s) substance(s) (1)

Has form (1)

DCI (1)

Code(s) UCD (1)

CIP code(s) (2)

ATC code(s) (1)

Spécialité(s) princeps (1)

EDQM-ST administration route(s) (1)

Intended site(s) of administration EDQM-ST (1)

Racine(s) Pharmaceutique(s) (1)

Is indicated for (9)

angina pectoris*/prevention and control	MeSH Descriptor/MeSH Qualifier
atrial fibrillation*/drug therapy	MeSH Descriptor/MeSH Qualifier
atrial flutter*/drug therapy	MeSH Descriptor/MeSH Qualifier
hypertension*/drug therapy	MeSH Descriptor/MeSH Qualifier
myocardial infarction*/drug therapy	MeSH Descriptor/MeSH Qualifier
tachycardia, ectopic junctional*/drug therapy	MeSH Descriptor/MeSH Qualifier
tachycardia, supraventricular/drug therapy	MeSH Descriptor/MeSH Qualifier
tachycardia, ventricular*/drug therapy	MeSH Descriptor/MeSH Qualifier
ventricular premature complexes*/drug therapy	MeSH Descriptor/MeSH Qualifier

CISMeF manual mappings (1)

ATTRIBUTES AND OBJECTS
ATTRIBUTE VALUES

- **Attribute values** are **numbers** or **symbols** assigned to an **attribute** for a **particular object**

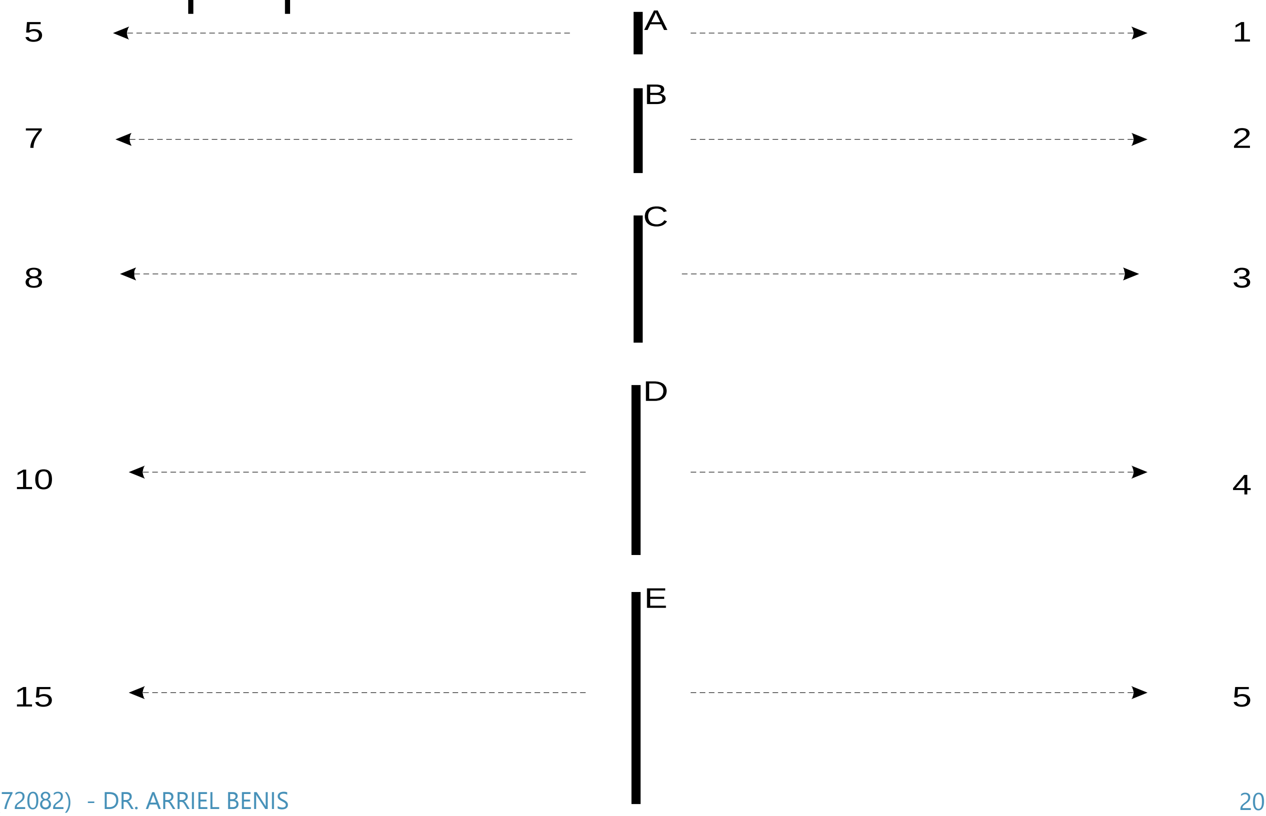
- Distinction between **attributes** and **attribute values**
  - Same attribute can be mapped to **different attribute values**
    - Example: height can be measured in feet or meters

  - Different attributes can be **mapped to the same set of values**
    - Example: Attribute values for ID and age are **integers**
  - But **properties of attribute** can be different than the **properties of the values** used to represent the attribute *(e.g., string-numbers, string-character)*

ATTRIBUTES AND OBJECTS
MEASUREMENT OF LENGTH

- The way you measure an attribute may not match the attributes properties.

This scale preserves only the ordering property of length.



This scale preserves the ordering and additivity properties of length.

ATTRIBUTES AND OBJECTS
TYPES OF ATTRIBUTES

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

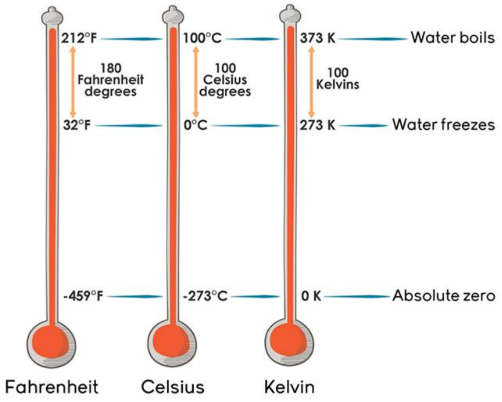
ATTRIBUTES AND OBJECTS
PROPERTIES OF ATTRIBUTE VALUES

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal (נומינלי)	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ2 test
	Ordinal (סידורי)	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval (טווח)	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio (יחסי)	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length	geometric mean, harmonic mean, percent variation

Developed by Stanley Smith Stevens (Psychologist)

ATTRIBUTES AND OBJECTS
DIFFERENCE BETWEEN RATIO AND INTERVAL

- Is it physically **meaningful to say that a temperature of 10 ° is twice that of 5°** on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?



- Consider **measuring** the height above average
  - If Bill's height is three centimeters above average and Bob's height is six centimeters above average, **then would we say** that Bob is twice as tall as Bill?
  - **Is this situation analogous to** that of temperature?

ATTRIBUTES AND OBJECTS
TRANSFORMATION OF ATTRIBUTE VALUES

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	**Any permutation** of values	If all employee ID numbers were reassigned, would it 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* where a and b are **constants**	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value =* ***a*** ** old_value*	Length can be measured in meters or feet.

Developed by Stanley Smith Stevens (Psychologist)

ATTRIBUTES AND OBJECTS
DISCRETE AND CONTINUOUS ATTRIBUTES

Discrete Attribute

- 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

- **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**.
- Continuous attributes are typically represented as **floating-point variables**.

ATTRIBUTES AND OBJECTS
CRITIQUES OF THE ATTRIBUTE CATEGORIZATION

Incomplete

- Asymmetric binary
  (e.g., sex: symmetric, gender, temperature, pain…: asymmetric)
- Cyclical
  (e.g., seasonality)
- Partially ordered
- Partial membership
  (e.g., fuzzy logic)
- Multivariate and Relationships between the data

Real data is approximate and noisy

- Can **complicate recognition** of the proper attribute type
  (e.g., distance as a category)
- Treating one attribute type as another may be **approximately correct**
  (e.g., age, height, distance)

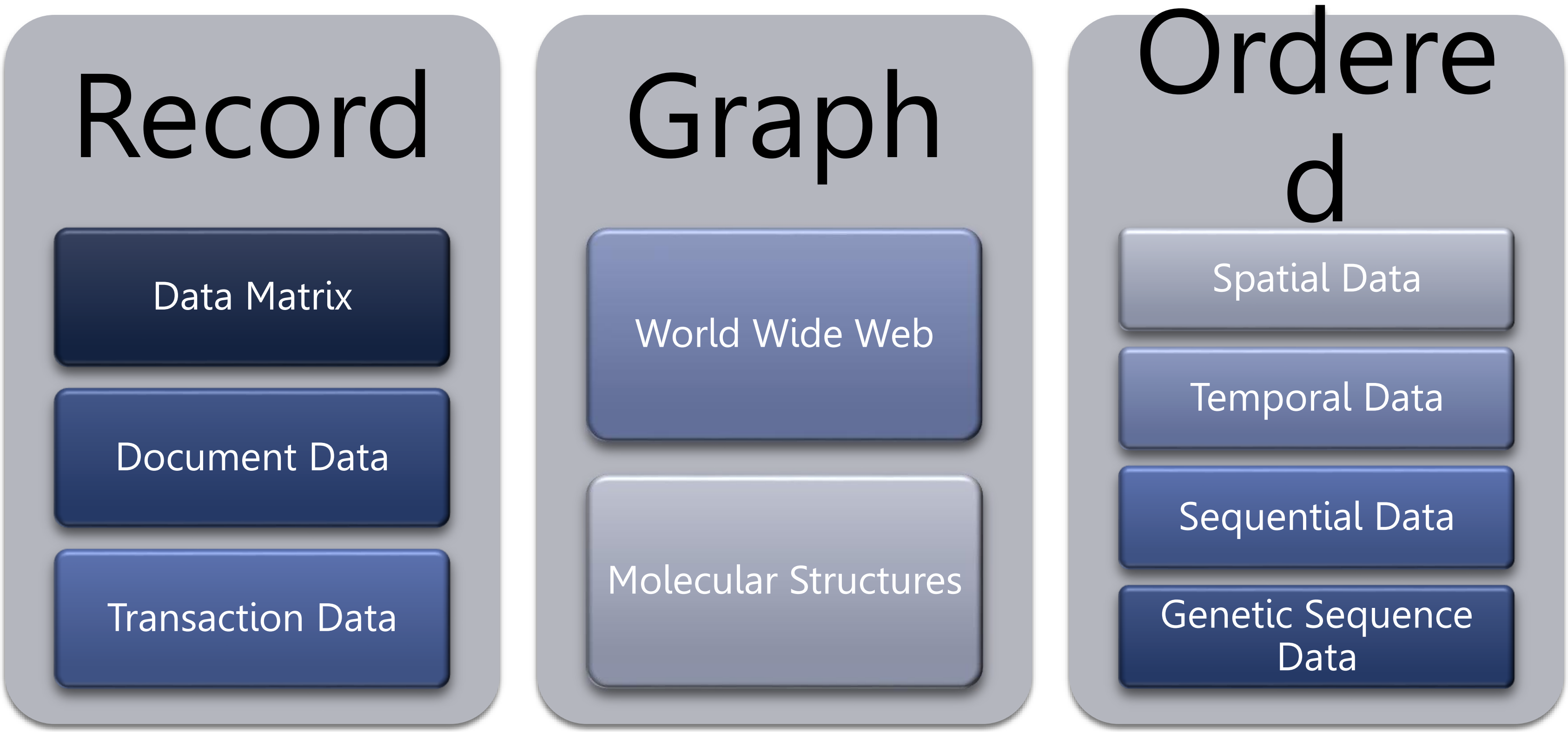
ATTRIBUTES AND OBJECTS
KEY MESSAGES FOR ATTRIBUTE TYPES

- The types of operations you choose should be "**meaningful**" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are **only four (among many possible) properties of data**
  - The data type you see – often numbers or strings **–** **may not capture all the properties** or **may suggest properties that are not present**
  - Analysis may **depend on these other properties of the data**
    - Many statistical analyses depend only on the **distribution**
  - In the end, **what is meaningful can be specific to domain**

TYPES OF DATA
IMPORTANT CHARACTERISTICS OF DATA

- **Dimensionality** (number of attributes)
  - High dimensional data brings a **number of challenges** ***(Curse of dimensionality)***
- **Sparsity**
  - ***Only presence counts***
- **Resolution**
  - Patterns ***depend on the scale***
- **Size**
  - Type of analysis may depend on ***size of data***

TYPES OF DATA
TYPES OF DATA SETS



Record

- Data Matrix
- Document Data
- Transaction Data

Graph

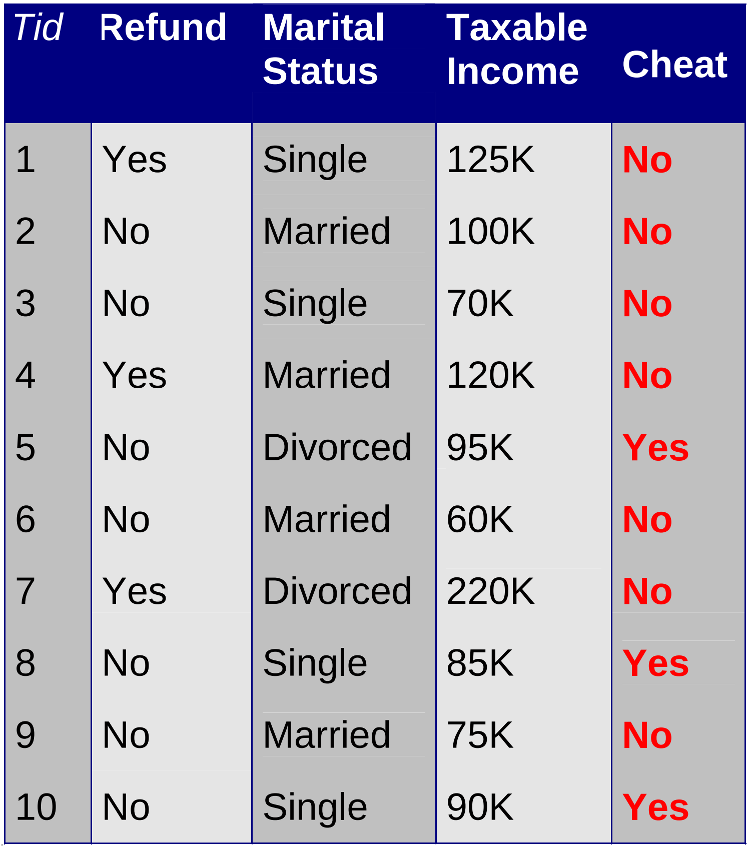
- World Wide Web
- Molecular Structures

Ordere
d

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

TYPES OF DATA
RECORD DATA

- Data that consists of a **collection of records**, each of which consists of **a fixed set of attributes**



Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

TYPES OF DATA
DATA MATRIX

- If data objects have the same **fixed set of numeric attributes,** then the **data objects** can be thought of as points in a **multi-dimensional space**, where **each dimension represents a distinct attribute**

- Such a data set can be represented by an ***m* by *n* matrix**, where there are ***m* rows**, one for **each object**, and *n* **columns**, one for **each attribute**

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

TYPES OF DATA
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.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

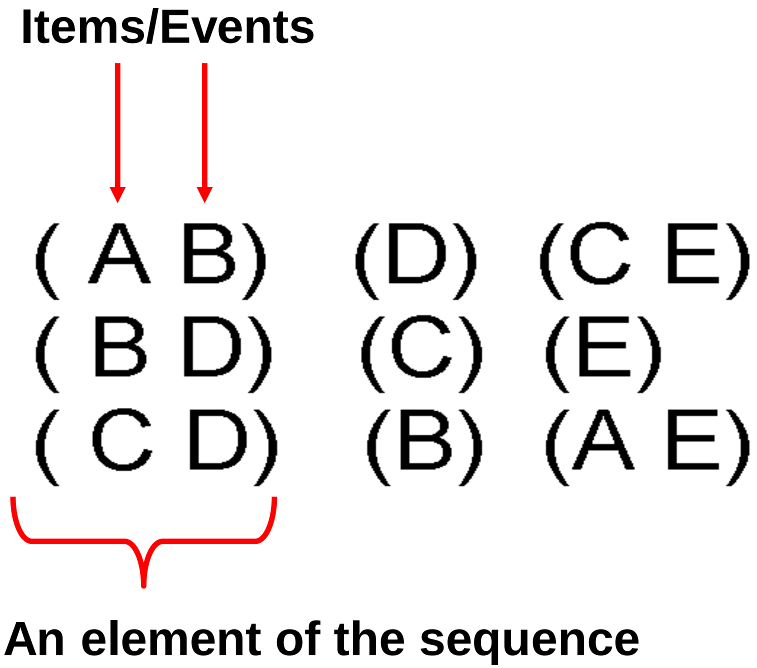
TYPES OF DATA
TRANSCRIPTION DATA

- A **special type of data**, where
  - Each **transaction involves a set of items**.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
  - **We can represent transaction data as record data**

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

TYPES OF DATA
ORDERED DATA

- Sequences of transactions



Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

TYPES OF DATA
ORDERED DATA

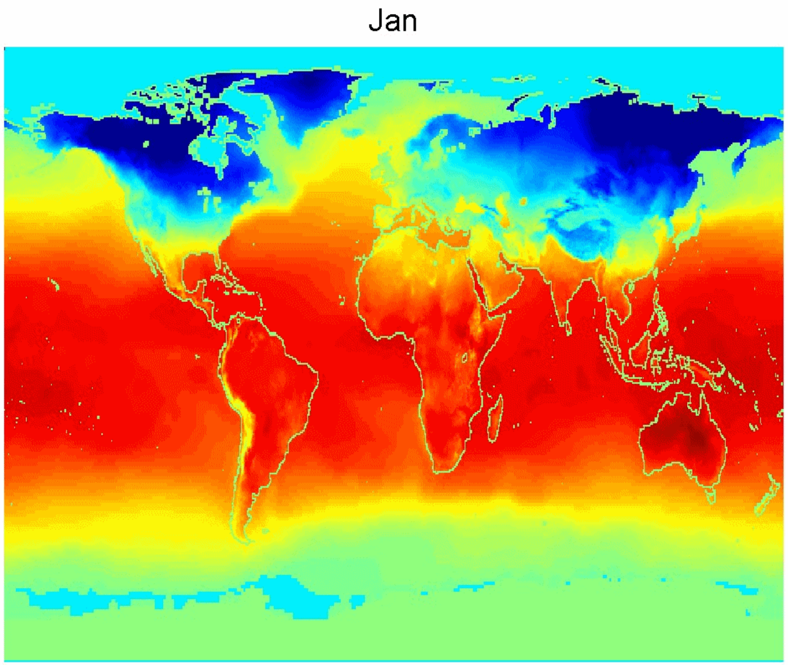
- **Genomic sequence data**

What's happen if we change the order?!?

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

TYPES OF DATA
ORDERED DATA

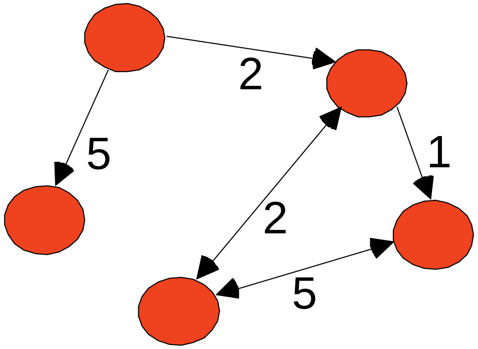
- **Spatio-Temporal Data**

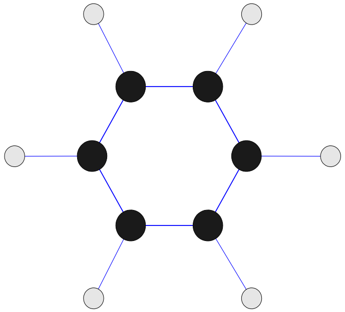


Average Monthly Temperature of land and ocean

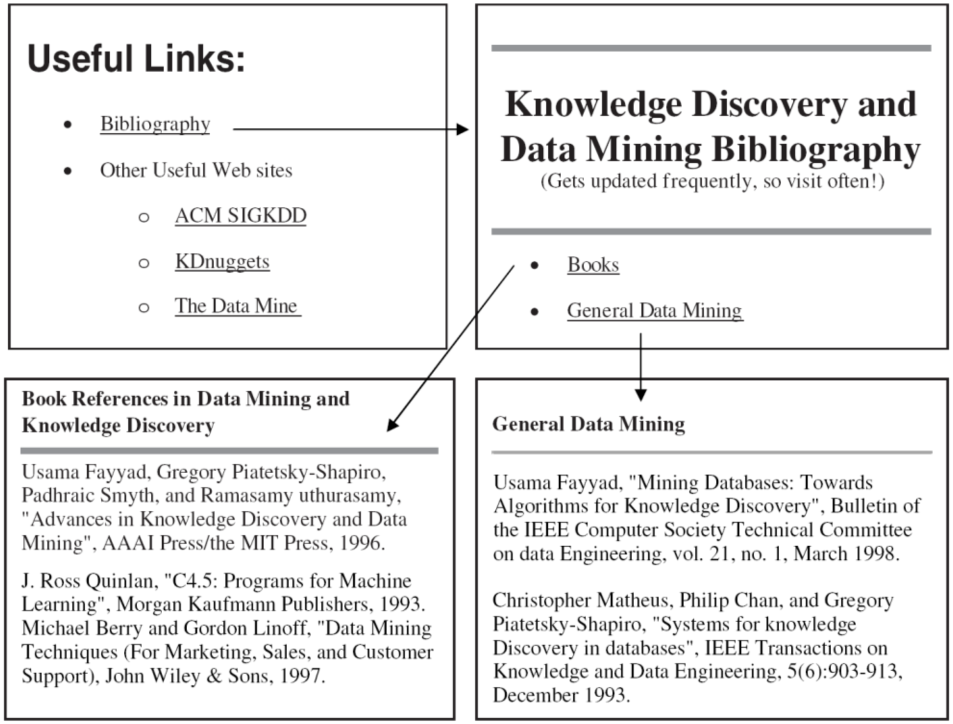
TYPES OF DATA
GRAPH DATA

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6



Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

DATA QUALITY
DEFINITION(S)

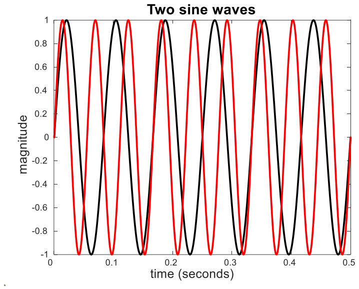
- **Poor** data quality **negatively affects** many data **processing** efforts
- Data mining ***example***:
  - a ***classification model*** for detecting people who are **loan risks** is built using poor data
    - Some credit-worthy candidates are denied loans
    - **More loans are given to individuals that default!!!**

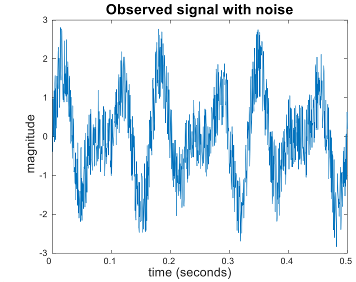
DATA QUALITY
DATA QUALITY …

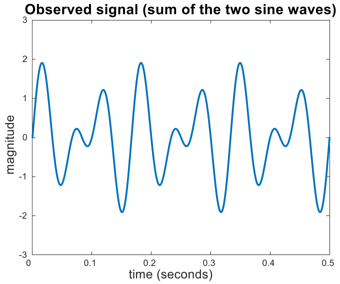
- What kinds of **data quality problems**? ………………………………………
- How can we **detect problems** with the data? ……………………………………
- **What can we do** about these problems? ………………………………………

- Examples of data quality problems:
  - **Noise** and **outliers**
  - **Wrong** data
  - **Fake** data ………………………………………………………………………………
  - **Missing** values
  - **Duplicate** data

DATA QUALITY
NOISE

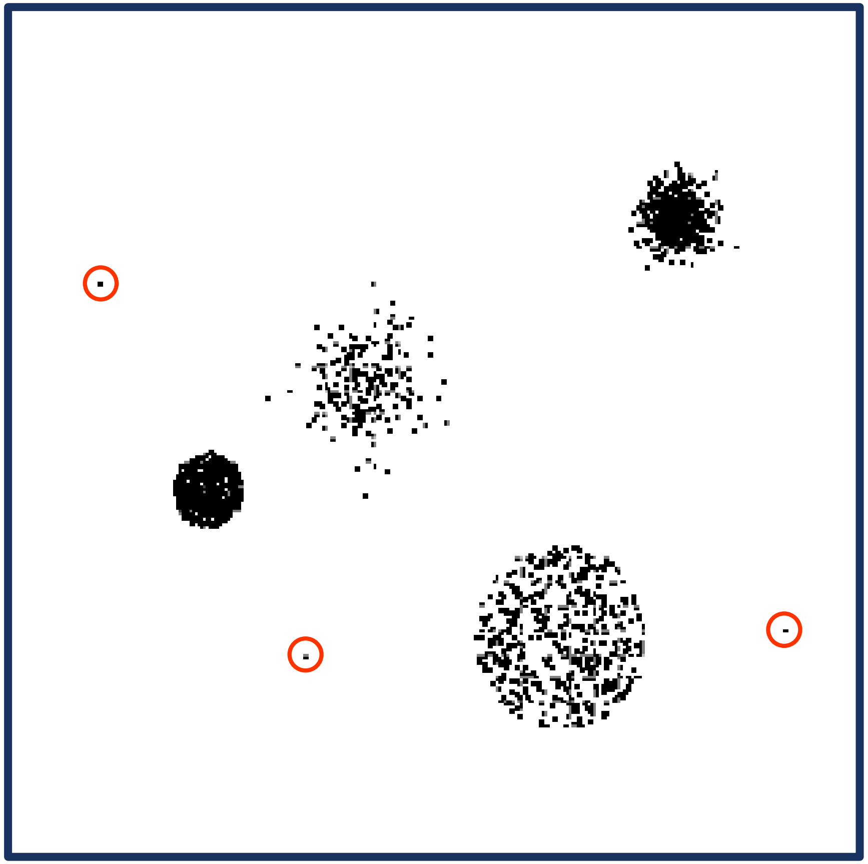

Two sine waves


Observed signal with noise


Observed signal (sum of the two sine waves)

- For objects,
  noise is an extraneous object
- For attributes, **noise refers to modification** of original values
  - Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen
  - The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
    - **The magnitude and shape of the original signal is distorted**

DATA QUALITY
OUTLIERS



- ***Outliers*** are data objects with characteristics that are **considerably different than most of the other data objects in the data set**
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection

- **Causes?**

DATA QUALITY
MISSING VALUES

- Reasons for missing values
  - **Information is not collected**
    (e.g., people decline to give their age and weight)
  - Attributes may **not be applicable to all cases**
    (e.g., annual income is not applicable to children)

- Handling missing values
  - **Eliminate data objects or variables**
  - **Estimate missing values**
    - Example: time series of temperature
    - Example: census results
  - **Ignore the missing value** during analys

DATA QUALITY
DUPLICATE DATA

- Data set may include **data objects that are duplicates**, or almost duplicates of one another
  - Major **issue when merging data** from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- **Data cleaning**
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?