Feb – Jun 2023
IT258

Data & Datasets

What Is a Dataset ?

- Dataset
  - A collection of **data objects**
- Objects (*have* **attributes**)
  - E.g. record, point, case, sample, entity, item, instance.. ..
- Attributes and variables (*describe* **objects**)
  - E.g. field, characteristic, feature or observation …

Types of Dataset Attributes

- ***Type of an attribute*** depends on ***properties*** it possesses:

Property	Application
Distinctness	= ≠
Order	< >
Addition	+ -
Multiplication	* /

Types of Dataset Attributes

Property	Application
Distinctness	= ≠
Order	< >
Addition	+ -
Multiplication	* /

- **Nominal** – exhibits ‘distinctness’
- **Ordinal** – exhibits ‘distinctness’ and ‘order’
- **Interval** – exhibits ‘distinctness’ , ‘order’ and ‘addition’
- **Ratio** – exhibits all four properties

Types of Dataset Attributes

- **Nominal attributes**
  - provide only enough information to distinguish one object from another
    - Each value represents a label.
  - Typical comparisons between two values are limited to “similar” or “dissimilar”
    - E.g. ID numbers, eye color, pincode …

Types of Dataset Attributes

- **Ordinal attributes**
  - values of an ordinal attribute provide enough information to order the given objects.
  - Typical comparisons between two values are “equal” or “greater” or “lesser”.
    - E.g. grades, income, rank, age ..

Types of Dataset Attributes

- **Interval attributes**
  - Differences between values are meaningful, i.e., a unit of measurement exists. (+, -)
    - E.g. Calendar dates, temperature, marks …

Types of Dataset Attributes

- **Ratio attributes**
  - Both differences and ratios are meaningful in the context of the given objects.
    - E.g. Money, counts, age, mass, length …

Attribute Type	Description	Examples
Nominal	Each value represents a label. (Typical comparisons between two values are limited to "equal" or "no equal")	Flower color, gender, zip code
Ordinal	The values can be ordered. (Typical comparisons between two values are "equal" or "greater" or "less")	grades, street numbers, rank, age
Interval	The differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	Calendar dates, temperature in Celsius or Fahrenheit
Ratio	Differences and ratios are meaningful. (*, /)	Monetary quantities, counts, age, mass, length, electrical current

Other types of Dataset Attributes

- **Discrete attributes**
  - have only a finite or countably infinite set of values.
  - often represented as integer variables.
    - E.g: Pincode, counts, the set of words in a document….
- **Continuous Attributes**
  - have real number values.
  - are typically represented as floating point variables.
    - E.g.: Temperature, height or weight, year …

Other types of Dataset Attributes

- **Binary attributes**
  - a special case of discrete attributes that assume only two values.
  - are often represented as Boolean variables, or as integer variables that take on the values 0 or 1
    - E.g.Yes/no, true/false, pass/fail . . .

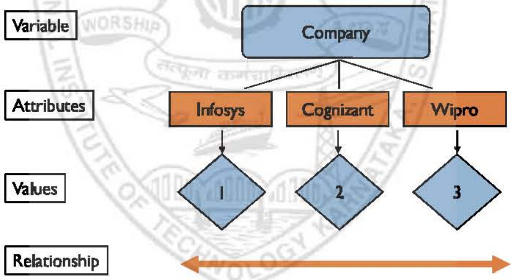
Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Variables

- A **variable** is a logical set of attributes.
  - Where as an **attribute** is a characteristic of an object.
- Variables can "**vary**"
  - E.g. Value of a variable can be high or low.
- A **domain** is a set of all possible values that a variable is allowed to have.
  - E.g. How high or how low, is determined by the domain of the attribute value.

Variables vs. Attributes

Relationship between the attributes and a variable



Variable

Company

Attributes

Infosys

Cognizant

Wipro

Values

1

2

3

Relationship

Types of Variables

- **Dichotomous Variables**
  - Can have precisely two distinct values.
    - E.g. Gender – Male / Female, Employed/Unemployed …
  - Special case: Binary variable
    - E.g. 0/1 ;Yes/ No; True/ False

Types of Variables

- **Categorical Variables**
  - variables on which calculations are not meaningful.
    - E.g. Pincode, telephone number.

Types of Variables

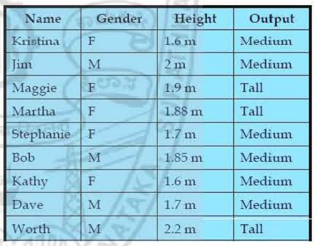
- **Metric variables**
  - variables on which calculations are meaningful.
    - E.g. interval and ratio variables are metric variables.
  - Analyzed using descriptive statistics techniques like mean, standard deviation and skewness.
    - E.g. marks, price …

Types of Datasets

- Record
  - Data Matrix, Document Data, Transaction Data …
- Graph
  - World Wide Web, Molecular Structures, Networks …
- Ordered
  - Spatial Data, Temporal Data, Sequential Data, Genetic Sequence Data…
- Structured Data
  - Relational DBs, RDF, Linked Data…

Dataset types - *Record Data*

- a collection of records (*data objects*), each of which consists of fixed set of data fields (*variables-attributes*)
- Very commonly used in data mining applications

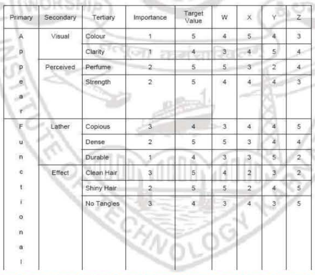


Name	Gender	Height	Output
Kristina	F	1.6 m	Medium
Jim	M	2 m	Medium
Maggie	F	1.9 m	Tall
Martha	F	1.88 m	Tall
Stephanie	F	1.7 m	Medium
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Medium
Dave	M	1.7 m	Medium
Worth	M	2.2 m	Tall

Dataset types - *Record Data*

Data Matrix

- Datasets with fixed set of numeric attributes
  - the data objects can be treated as points in a multi-dimensional space, where each dimension represents a distinct attribute.



Primary	Secondary	Tertiary	Importance	Target Value	W	X	Y	Z
Appear	Visual	Colour	1	5	4	5	4	3
		Clarity	1	4	3	4	5	4
	Perceived	Perfume	2	5	5	3	2	4
		Strength	2	5	4	4	4	3
Functional	Lather	Copious	3	4	3	4	4	5
		Dense	2	5	5	3	4	4
		Durable	1	4	3	3	5	2
	Effect	Clean Hair	3	5	4	2	3	2
		Shiny Hair	2	5	5	2	4	5
		No Tangles	3	4	3	4	3	5

Dataset types – Record Data

Document Data

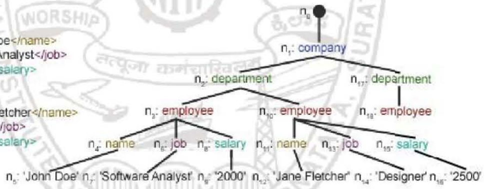
- Each document becomes a 'term' vector, where,
  - Each term is an *attribute* of the vector,
  - *Value* of each attribute of the vector = no.of times the corresponding term occurs in the document.

	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

Dataset types – *Graph Data*

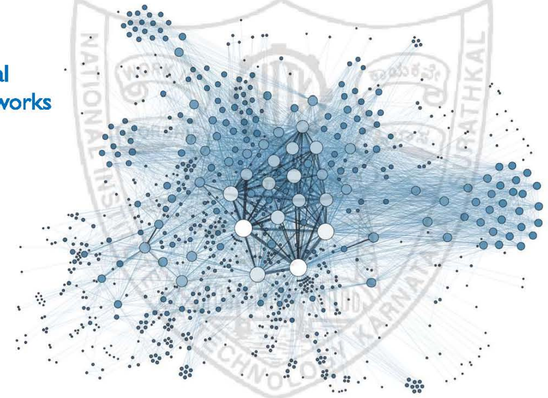
XHTML, XML documents

```
<company>
 <department>
  <employee>
   <name>John Doe</name>
   <job>Software Analyst</job>
   <salary>2000</salary>
  </employee>
  <employee>
   <name>Jane Fletcher</name>
   <job>Designer</job>
   <salary>2500</salary>
  </employee>
 </department>
 <department>
  <employee>
  </employee>
 </department>
</company>
```

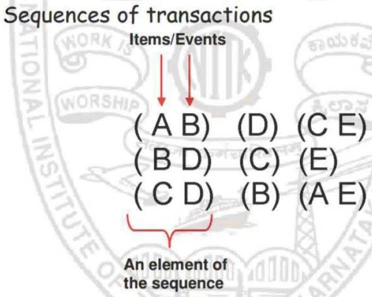


Dataset types – *Graph Data*

Social Networks

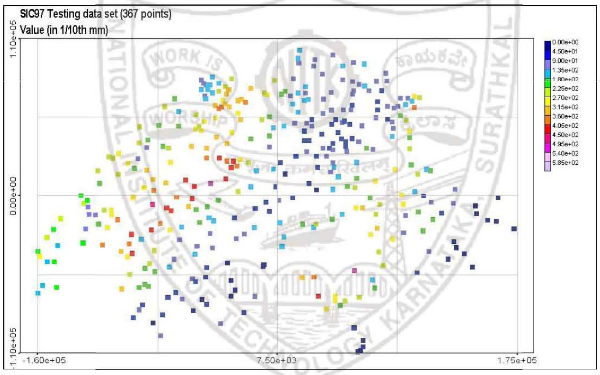


Dataset types – *Ordered Data*



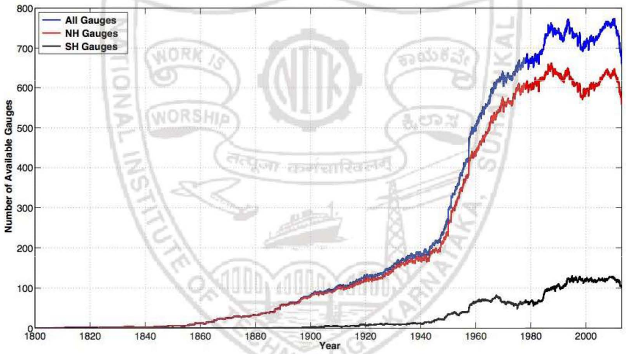
Dataset types – *Ordered Data*

Spatial Data



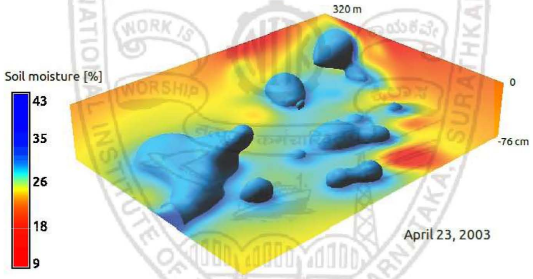
Dataset types – *Ordered Data*

Temporal Data



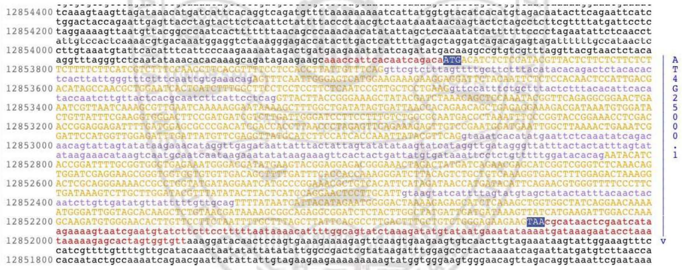
Dataset types – *Ordered Data*

Spatio-temporal Data



Dataset types – *Ordered Data*

Genomic Sequences



```
12854400 tcaaagtaagttagataaacatgatcattcacaggtcagatgttttaaaaaaaaatcattatggtgtacatcacatgtagacaatacttcagaattcatc
         tggactaccagaattgagttacctagtacttctcaattctatttaccctaacgtctaataaataacaagtactctagcctcttcgttttatgattcctc
12854200 taggaaaagttaatgttacggcccaatcactttttttaacagcccaaacaacatatattagctccaaatatcatttttttcccctagaatattctcaacct
         attgtccactcaaaacgtgacaaatggaggtctaaagggagaccatacttgactcattttagagctaggatcagacagagtagattttttgccataactc
12854000 cttgtaaatgtattcacatttcattcccaagaaaaatagactgatgaagaaatatatcagatatgacaaggccgtgtcgtttaggttacgtaactctaca
         aggtttagggtctcaatataaacacacaaagcagatagaagaagcaaaccattcacaatcagacaATGACATCTCTCCATACGTTACTCTTCTCTTCTCT   A
12853800 TCTTTTCTTCATCGTCTTTCCAACCTTCACGTTTTCCTCCACCTTATTGTTTCAGgttcgtctttagttttgcttctttacatacacagactctacacac   T
         tcacttattgggtttctttcaattgtgaaacagAGTTTCAATTGGAGTCATGGAAGAAAGAAGGAGGATTCTACAATTCTCTCCACAACTCCATTGACG   4
12853600 ACATAGCCAACGCTGGAATCACTCATCTTTGGCTTCCTCCTCCTTCTCAATCCGTTGCTCCTGAAGgttccatttctgctttactctttacacattcaca   G
         taccaatcttgttactcacgcaatcttcattcctcagGTTACTTACCGGGAAAGCTATACGATCTAAACAGCTCCAAATACGGTTCAGAGGCGGAACTGA   2
12853400 AATCGTTAATCAAAGCGTTGAATCAAAAAGGAATAAAAGCTTTGGCTGATATAGTGATTAACCACAGAACAGCTGAGAGGAAAGACGATAAATGTGGATA   5
         CTGTTATTTCGAAGGTGGACTTCCGATGATCGTCTTGATTGGATCCTTCCTTTGTCTGGCGCAATGACGCTAAATTTCCCGGTACCGGAAACCTCGAC   0
12853200 ACCGGAGGAGATTTTGATGGAGCGCCCGACATGGACCACGTTAACCCTAGAGTTCAGAAAGAGTTGTCCGAATGGATGAATTGGCTTAAAACTGAAATCG   0
         GATTCCATGGTTGGAGATTTGATTATGTTCGAGGTTATGCATCTTCCATCACCAAATTATACGTTCAGgtaaatcacatatgaattctcaaatatcagac   0
12853000 aacagtattagtatataagaaacataggttgagataattatttactattagtatatataagtatcataggttgatagggttatttactactatttagtat   .
         ataagaaacataagtcaatgcaatcaataagaaatatataagaaagttcactactgattatgtgataaattcctctgttttggatacacagAATACATC   1
12852800 ACCGGATTTTGCGGTGGGTGAGAAATGGGACGATATGAAGTACGGAGGAGACGGGAAACTAGACTATGATCAGAACGAGCATCGGTCGGGTCTCAAACAG
         TGGATCGAGGAAGCGGGTGGTGGTGTTGACAGCTTTTGATTTCACCACCAAAGGGATCTTACAGTCTGCTGTCAAAGGTGAGCTTTGGAGACTAAAGG
12852600 ACTCGCAGGGAAAACCGCCTGGTATGATAGGAATCATGCCCGGAAACGCTGTCACATTCATAGATAACCATGATACATTCAGAACGTGGGTTTTCCCTTC
         TGATAAAGTCTTGCTTGGATACGTTTATATACTTACTCATCCAGGAACTCCTTGCATTgtaagtatcatttagtatgtagctatactatttacaactac
12852400 aatcttgttgatatgttattttgttgcagTTTTATAATCATTACATAGAATGGGGACTAAAAGAGAGCATCTCAAAGCTGGTGGCTATCAGGAACAAAA
         ATGGGATTGGTAGCACAAGCTCTGTAACGATAAAAGCGGCAGAGGCGGATCTCTACTTGGCTATGATTGATGATAAAGTTATCATGAAGATTGGACCAAA
12852200 GCAAGATGTGGGAACACTTGTTCCTTCTAATTTTGCTTTAGCTTATTCAGGCCTTGACTTTGCTGTCTGGGAGAAGAAGTAAcgcataactcgaatcata
         agaaaagtaatcgaatgtatcttcttccttttaataaaacatttggcagtatctaaagatatgtataatgaaatataaaatgataaagaatacctaaa
12852000 taaaaagagcactagtggtgttaaaggatacaactccagtgaaagaaaagagttcaagtgaagaagtgtcaacttgtagaaataagtattggaaagtttc  v
         catcgttttgttttgttgcatacaactaatatattatatattggccgactcgtataagatttggagccctactaaaatcagaattatgatgtcttaacca
12851800 cacaatactgccaaaatcagaacgaattatattattgtagaagaagaaaaaaaaaagtatggtgggaagtgggaacagttagacaggtaaattcgaataaa
```

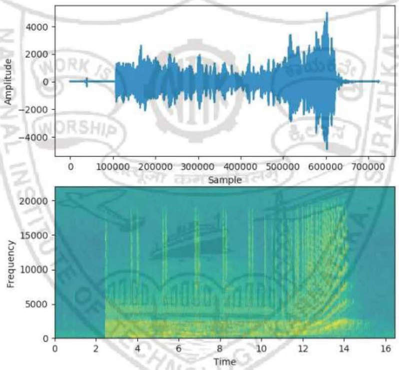
Dataset types – *Ordered Data*

Video data: sequence of images

Video frames

Dataset types – *Ordered Data*

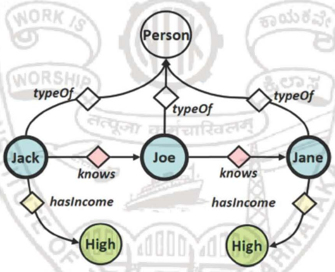
Audio data



Sound signal and its Spectrogram

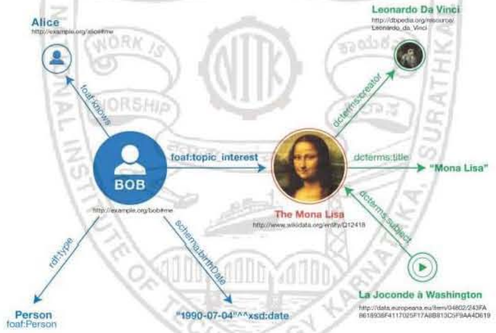
Dataset types – *Structured Data*

RDF data



Dataset types – *Structured Data*

Linked Data


Alice
http://example.org/alice#me
foaf:knows
BOB
http://example.org/bob#me
foaf:topic_interest
rdf:type
Person
foaf:Person
schema:birthDate
"1990-07-04"^^xsd:date
Leonardo Da Vinci
http://dbpedia.org/resource/
Leonardo_da_Vinci
dcterms:creator
dcterms:title
"Mona Lisa"
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:subject
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA
8618938F4117025F17A8B813C5F9AA4D619