

# Practical Machine Learning

## Day 3: Sep22 DBDA

Kiran Waghmare

# Agenda

- Data
- Types of Attributes
- Preprocessing
- Transformations
- Measures
- Visualization

## Life cycle of Machine Learning:

=====

### 1. Collection of data : sources

- a. Identify the data sources
- b. Collect data
- c. Integration of the data

### 2. Data preparation :

- a. Data Exploration (EDA)
- b. Data preprocessing
  - 1. missing values
  - 2. Duplicate values

### 3. Data Wrangling

- process of cleaning and converting raw data into suitable format.
  - 1. Noise filtration

### 4. Data analysis

- steps
  - a. analytical techniques
  - b. build model
  - c. apply algorithms (classification, regression, association, clustering)

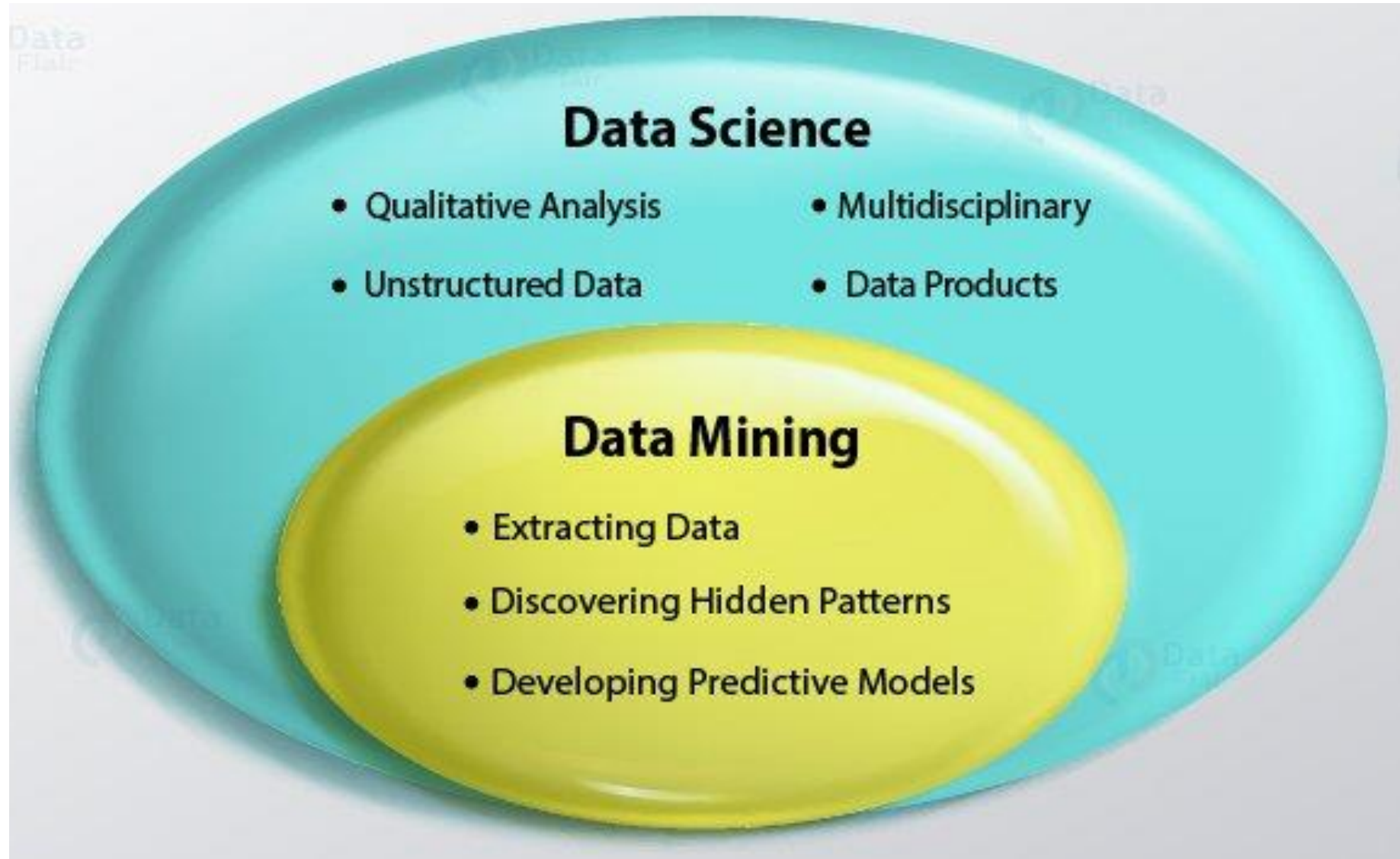
### 5. Train model

### 6. Test model

### 7. Deployment of model


Data

DataMining



# Record data

attribute/Feature/field/characteristics  
**Columns**



<i>Tid</i>	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



# 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, or feature**
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Object  
Example

Record

Property  
Features  
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

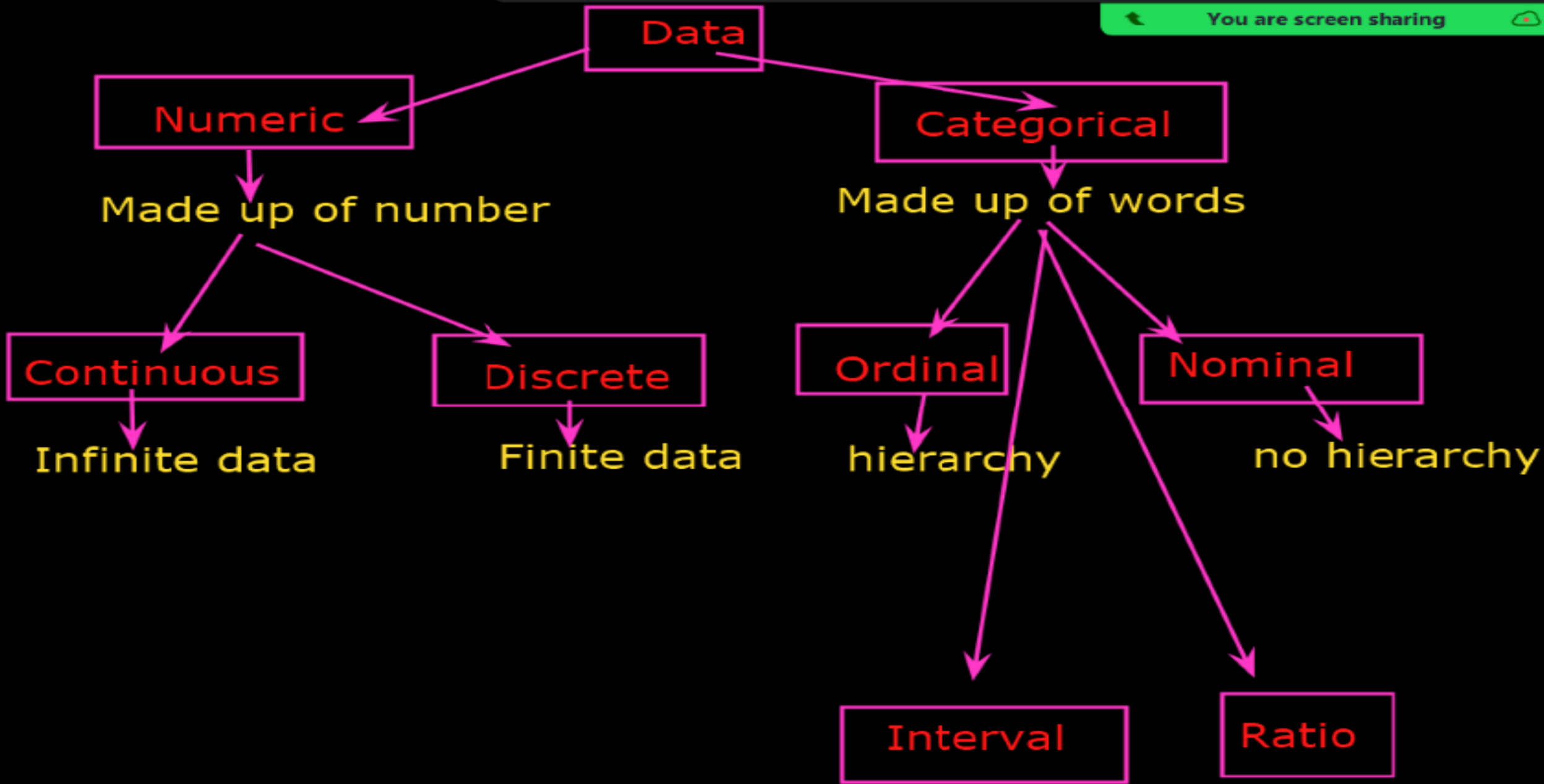
# Types of Data

- **Categorical features** come from an unordered set:
  - Binary: job?
  - Nominal: city.
- **Numerical features** come from ordered sets:
  - Discrete counts: age.
  - Ordinal: rating.
  - **Continuous**/real-valued: height.

# Discrete and continuous attributes

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





# Record data

Binary classifier

classifier

<i>Tid</i>	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

yes:3

no:7

Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00



Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

# Record data

<i>Tid</i>	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 matrix

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

# Document data

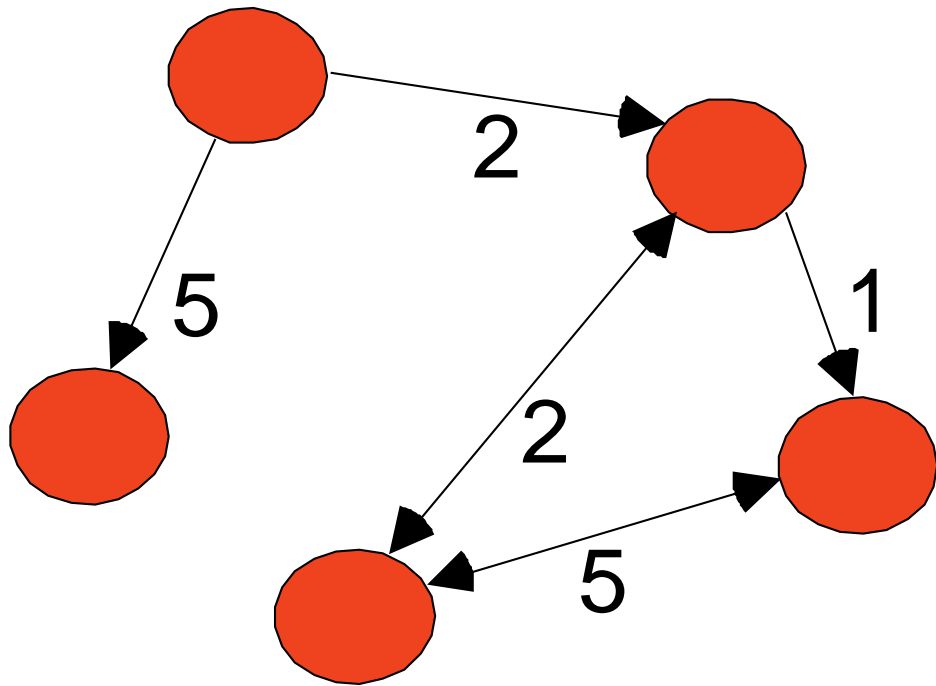
	team	coach	play	ball	score	game	win	lost	timeout	season
document 1	3	0	5	0	2	6	0	2	0	2
document 2	0	7	0	2	1	0	0	3	0	0
document 3	0	1	0	0	1	2	2	0	3	0



# Transaction data

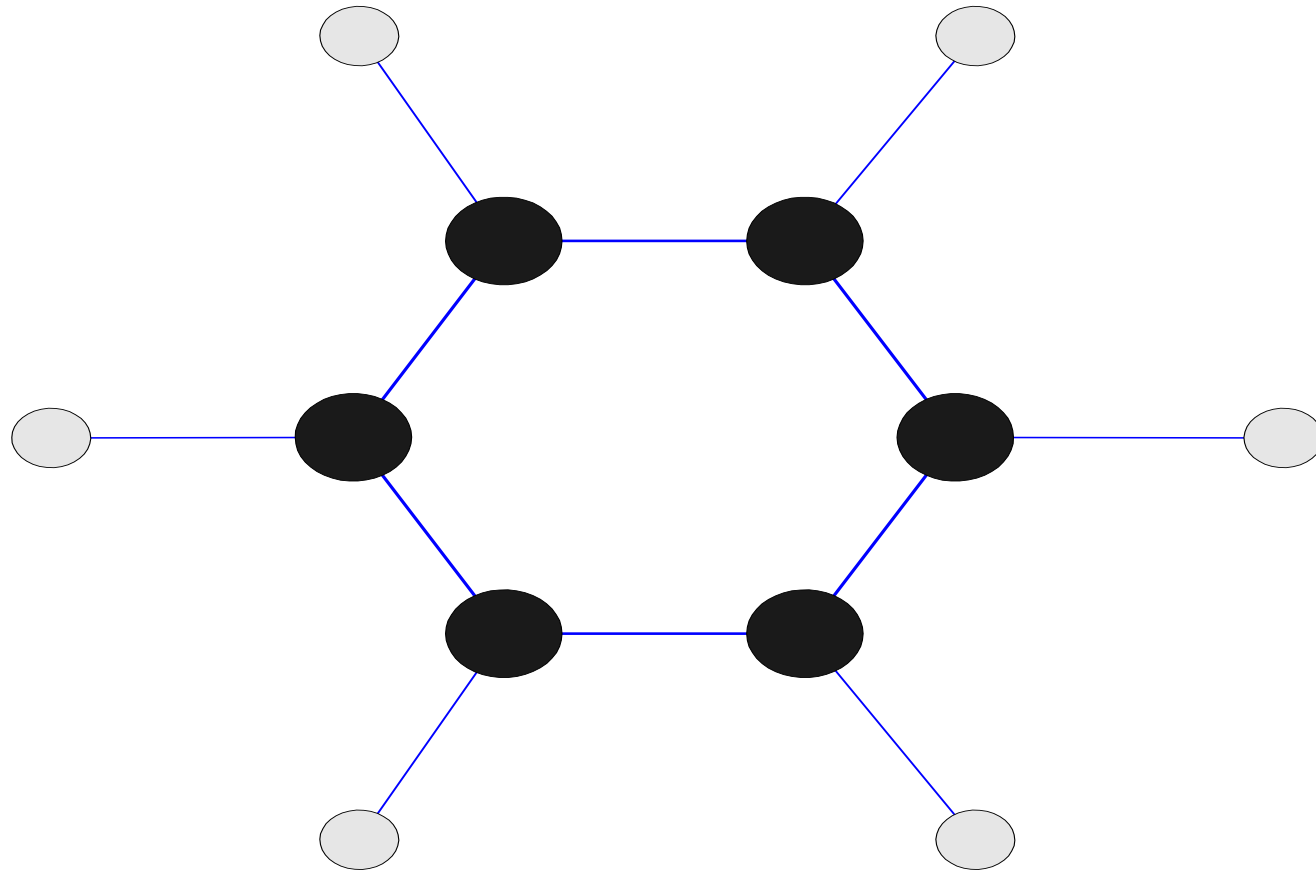
<i><b>TID</b></i>	<i><b>Items</b></i>
<b>1</b>	<b>Bread, Coke, Milk</b>
<b>2</b>	<b>Beer, Bread</b>
<b>3</b>	<b>Beer, Coke, Diaper, Milk</b>
<b>4</b>	<b>Beer, Bread, Diaper, Milk</b>
<b>5</b>	<b>Coke, Diaper, Milk</b>

# Graph data



# Chemical data

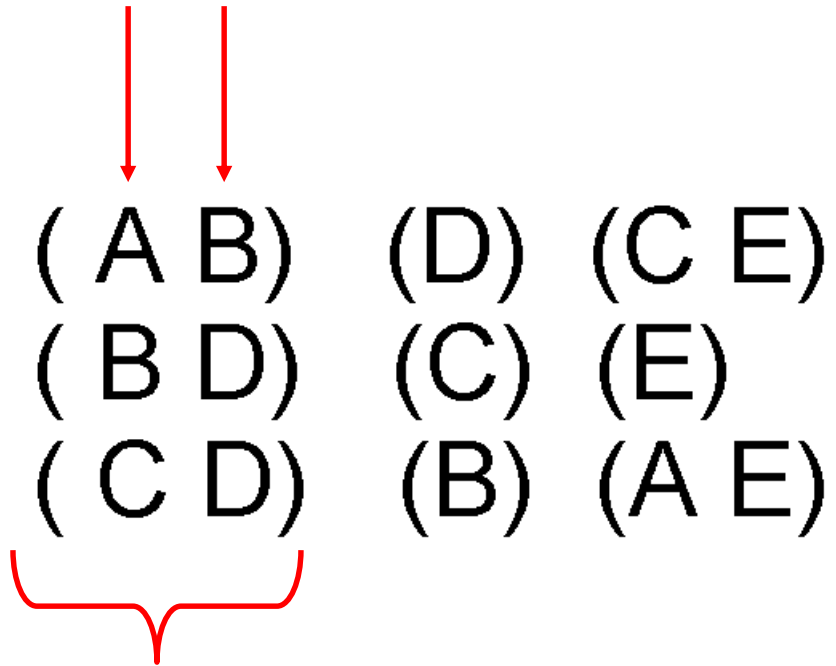
- Benzene molecule:  $\text{C}_6\text{H}_6$



# Ordered data

- Sequences of transactions

Items/Events



An element of the  
sequence

# Ordered data

- Genomic sequence data

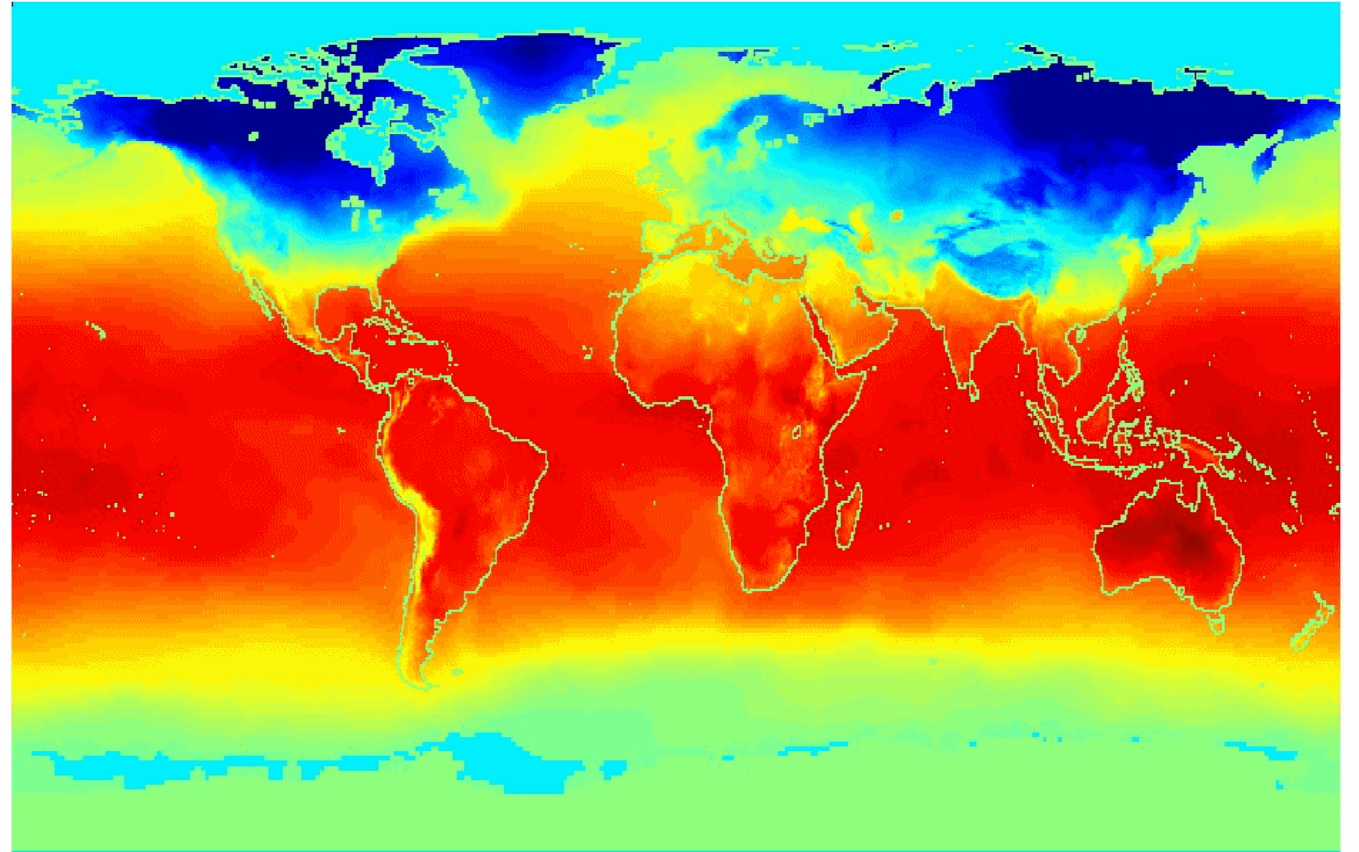
**GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCCGCCCCGCGCCGTC  
GAGAAGGGCCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAAGGTGCC  
CCCTCTGCTCGGGCCTAGACCTGA  
GCTCATTAGGCGGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAAGG**

# Ordered data

- Spatio-temporal data

Jan

Average monthly  
temperature of land  
and ocean





Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00



Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

# Approximating Text with Numerical Features

- **Bag of words** replaces document by word counts:

The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

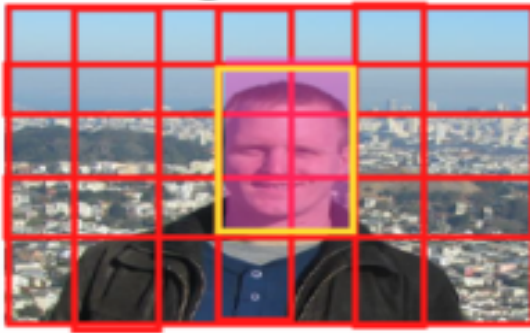


ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

- Ignores order, but often captures general theme.
- You can compute a “distance” between documents.

# Approximating Images and Graphs

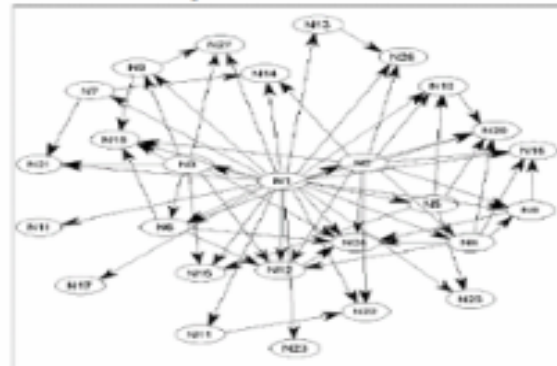
– Images:



graycale  
intensity

(1,1)	(2,1)	(3,1)	...	(m,1)	...	(m,n)
45	44	43	...	12	...	35

– Graphs:



adjacency  
matrix

N1	N2	N3	N4	N5	N6	N7
0	1	1	1	1	1	1
0	0	0	1	0	1	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0

# Feature Aggregation

- Feature aggregation:
  - Combine features to form new features:

Van	Bur	Sur	Edm	Cal		BC	AB
1	0	0	0	0		1	0
0	1	0	0	0		1	0
1	0	0	0	0	→	1	0
0	0	0	1	0		0	1
0	0	0	0	1		0	1
0	0	1	0	0		1	0

- Fewer province “coupons” to collect than city “coupons”.

# Feature Selection

- Feature Selection:
  - Remove features that are not relevant to the task.

SID:	Age	Job?	City	Rating	Income
3457	23	Yes	Van	A	22,000.00
1247	23	Yes	Bur	BBB	21,000.00
6421	22	No	Van	CC	0.00
1235	25	Yes	Sur	AAA	57,000.00
8976	19	No	Bur	BB	13,500.00
2345	22	Yes	Van	A	20,000.00

- Student ID is probably not relevant.

Out[6]:

	Independent			Dependent
	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

$X = x_1, x_2, x_3$

$X$ : Independent variable

$y$ : Dependent variable

(Classifier): Purchase

yes

no