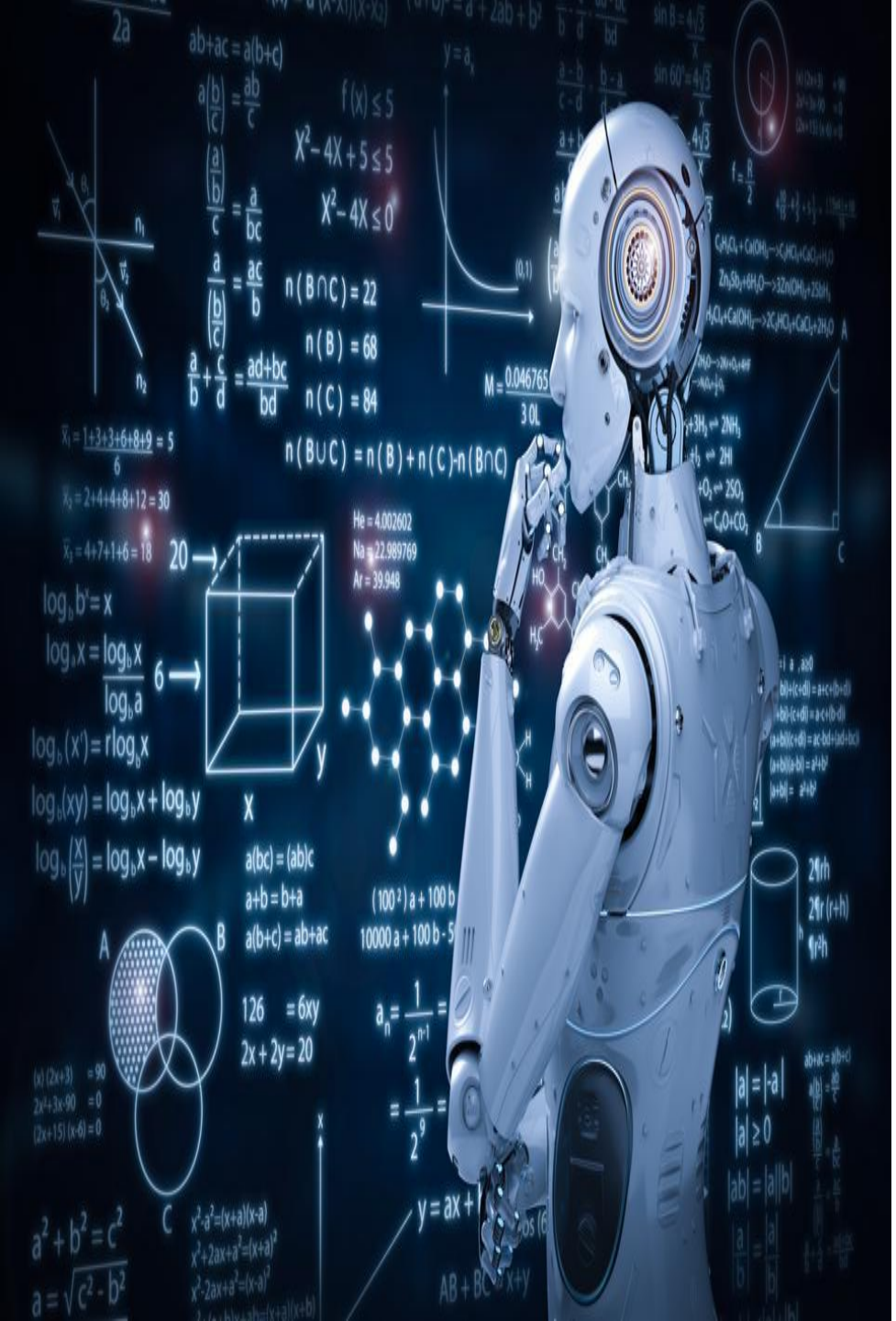


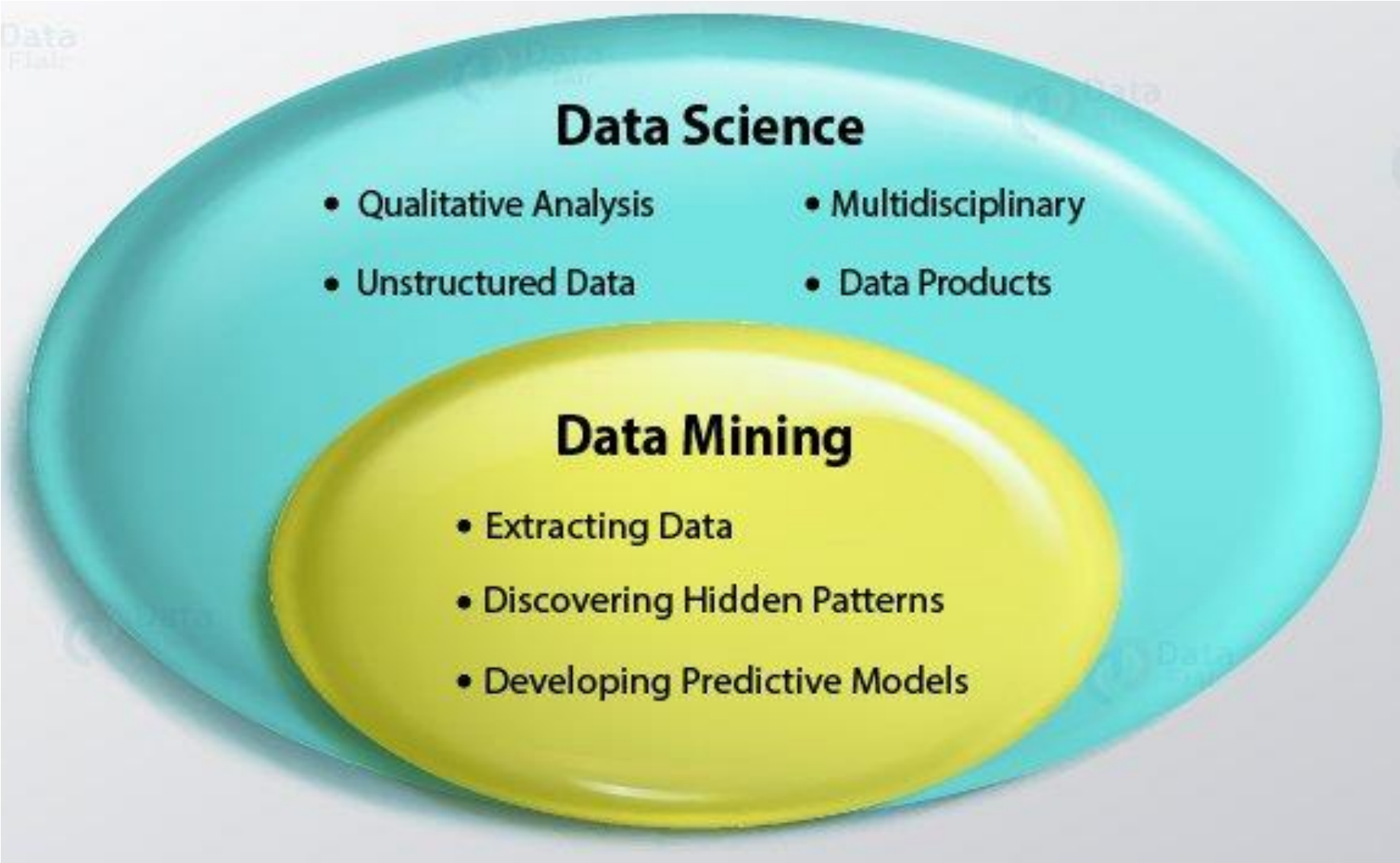
Practical Machine Learning

Day 3: Mar22 DBDA

Kiran Waghmare

Agenda

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


Data Science
• Qualitative Analysis
• Multidisciplinary
• Unstructured Data
• Data Products
Data Mining
• Extracting Data
• Discovering Hidden Patterns
• Developing Predictive Models

Develope Application

1. Task
2. Collect data
3. Clean & process that data
4. Transform
5. Algorithm
6. Data Mining
7. Evaluate, visualise & interprete → Hidden information + Noise
8. Goal: profit prediction

EX. cost: 10000/-

Error

Ac.cost 12000/-

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

Property

Features

Attributes

Object

Example

Record

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

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

Property

Features

Attributes

Object

Example

Record

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

- **Numerical**
  - Made up of number
    - **Continuous**
      - Infinite
    - **Discrete**
      - Finite
- **Categorical**
  - Made up of words
    - **Ordinal**
      - Distinct
      - Oredered
    - **Nominal**
      - Distinct
    - **Interval**
      - Distinct
      - Ordered
      - Addition
    - **Ratio**
      - All

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.

Types of Datasets

- Record
  - Data matrix
  - Document data
  - Trasaction
- Graph
  - Web
  - Genome
  - DNA
- Ordered
  - spatial data
  - Temoporal data
  - Sequential data

Record data

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 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

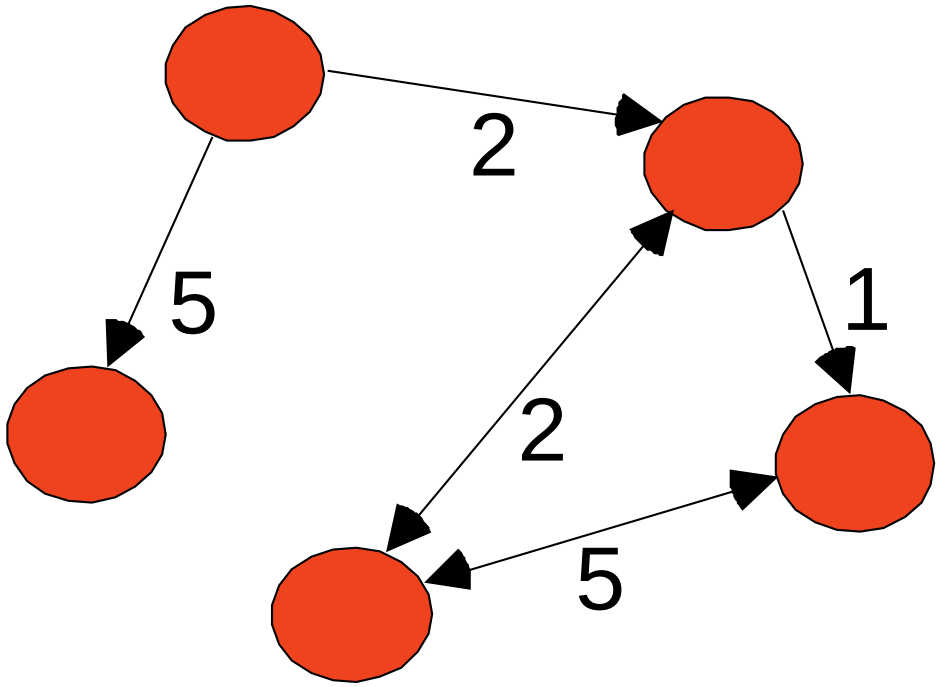
Document data

	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

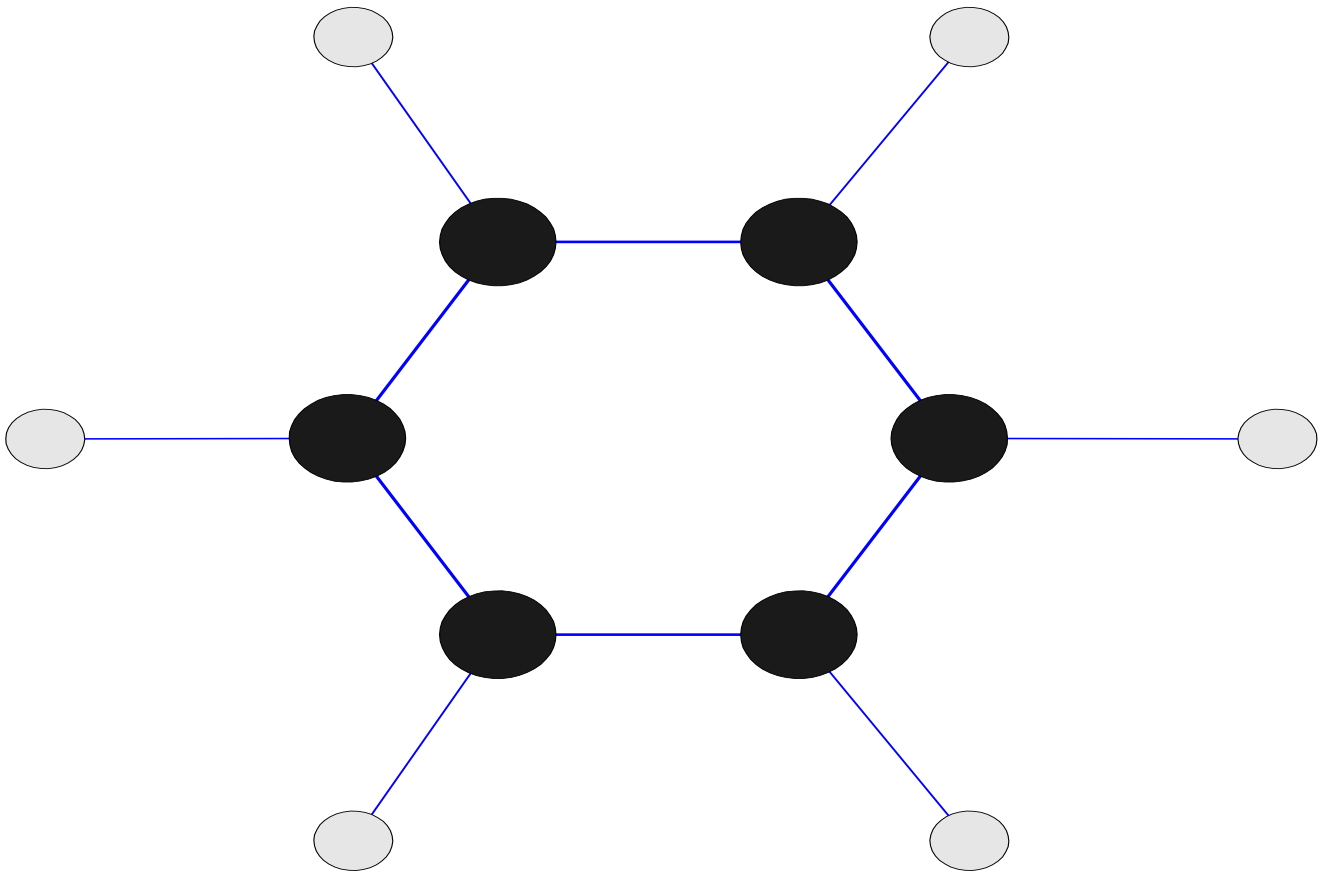
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph data



Chemical data

- Benzene molecule: C_6H_6



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

Ordered data

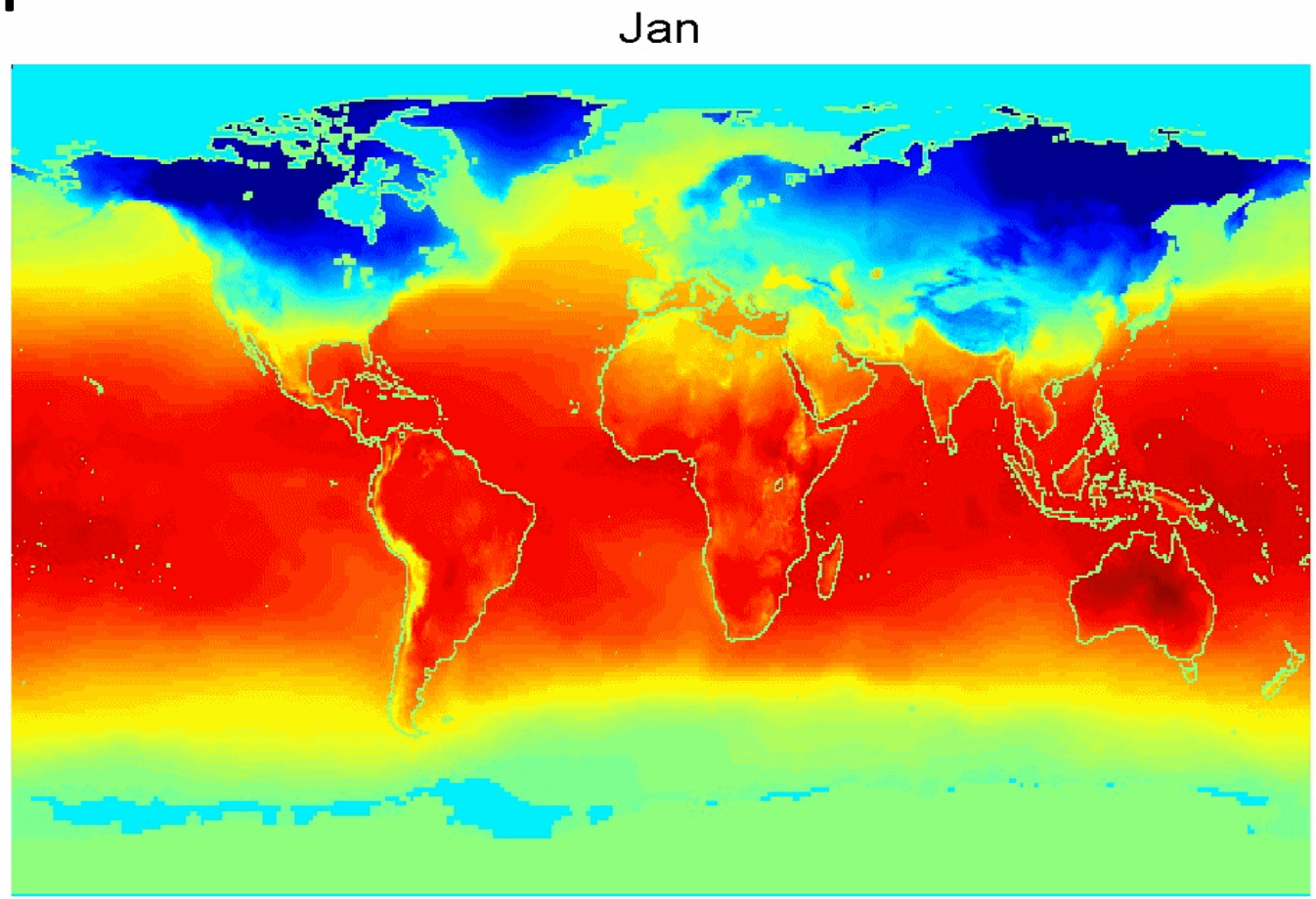
- Genomic sequence data

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

Ordered data

- Spatio-temporal data

Average monthly temperature of land and ocean



Converting to Numerical Features

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

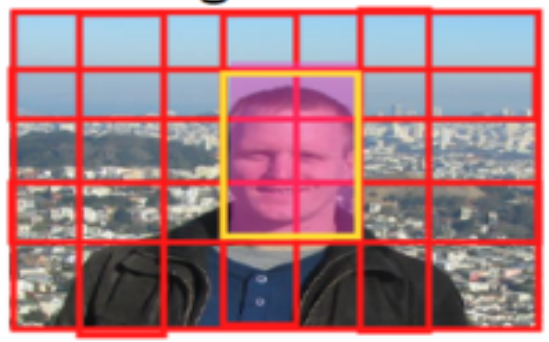
- Bag of words replaces document by word counts:

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

Approximating Images and Graphs

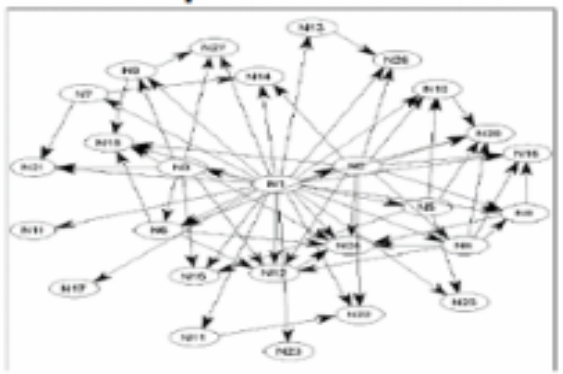
– Images:



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

Data Cleaning

- ML+DM typically assume 'clean' data.
- Ways that data might not be 'clean':
  - Noise (e.g., distortion on phone).
  - Outliers (e.g., data entry or instrument error).
  - Missing values (no value available or not applicable)
  - Duplicated data (repetitions, or different storage formats).
- Any of these can lead to problems in analyses.
  - Want to fix these issues, if possible.
  - Some ML methods are robust to these.
  - Often, ML is the best way to detect/fix these.

Feature Aggregation

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

Van	Bur	Sur	Edm	Cal
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1
0	0	1	0	0

BC	AB
1	0
1	0
1	0
0	1
0	1
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.

<20
21-22
>22

Teen
young
Adult

Exploratory Data Analysis

Statistics-->Mean, Max, Std dev,

1. Summary
2. Visualize
3. ML(Algorithms apply)

Data Quality:

-Noise

-Missing values

-duplicate vlues

smoothing, averaing, filters
(signal processing)

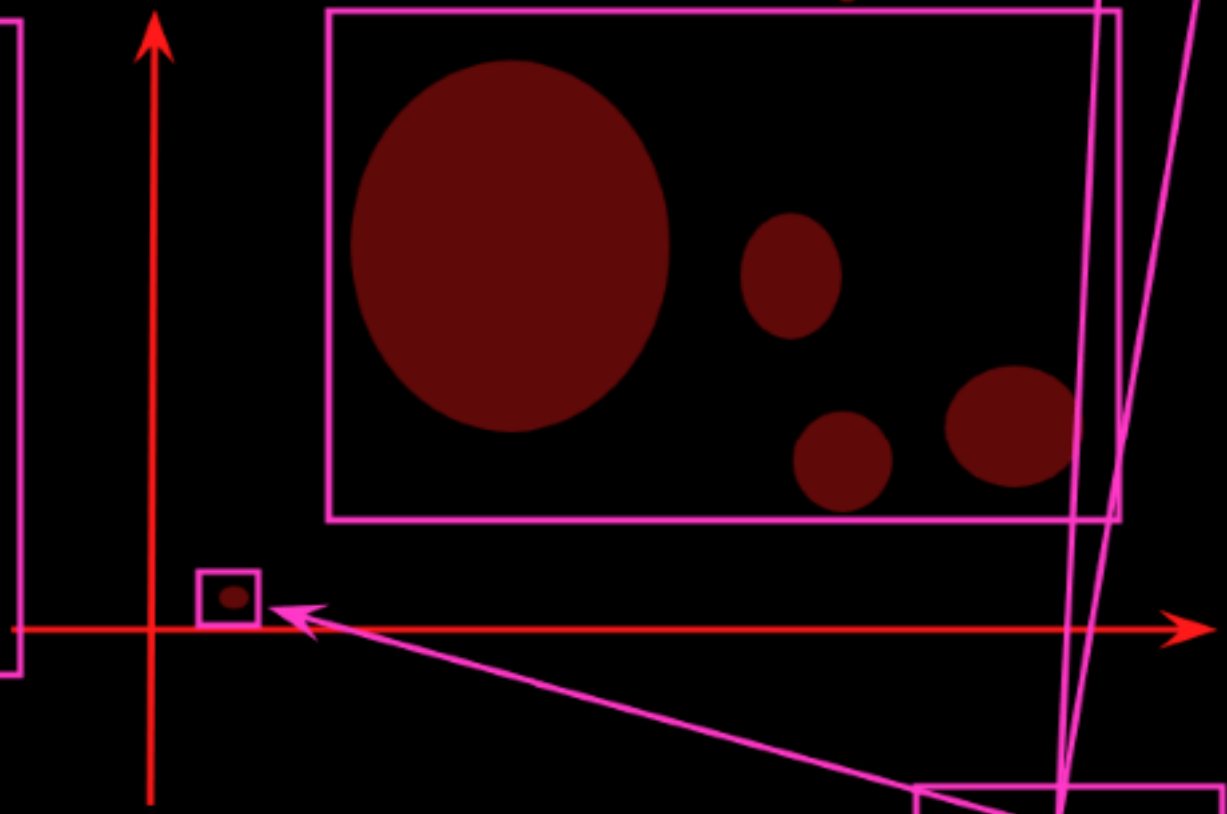
Handling missing values

-Eliminate data records

-Imputation(Estimating the missing value-mean,min,max)

-Ignore

-Replace by default value



Outlier

```
In [32]: dataset.head(10)
```

Classifier

Out[32]:

	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

```
In [30]: dataset.shape
Out[30]: (10, 4)
```

```
In [31]: dataset.describe()
```

```
In [28]: dataset=pd.read_csv('D:\Test\Data.csv')
```

```
In [32]: dataset.head(10)
```

Out[32]:

	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

Classifier

Y

Independent → X

Dependent → Y

Task:purchase?

yes/no

```
In [30]: dataset.shape
Out[30]: (10, 4)
```

```
In [31]: dataset.describe()
```

Run Code Commit

```
In [74]: import seaborn as sns

In [75]: dataset=pd.read_csv('D:\Test\Data.csv')

In [76]: dataset
```

Out[76]:

	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