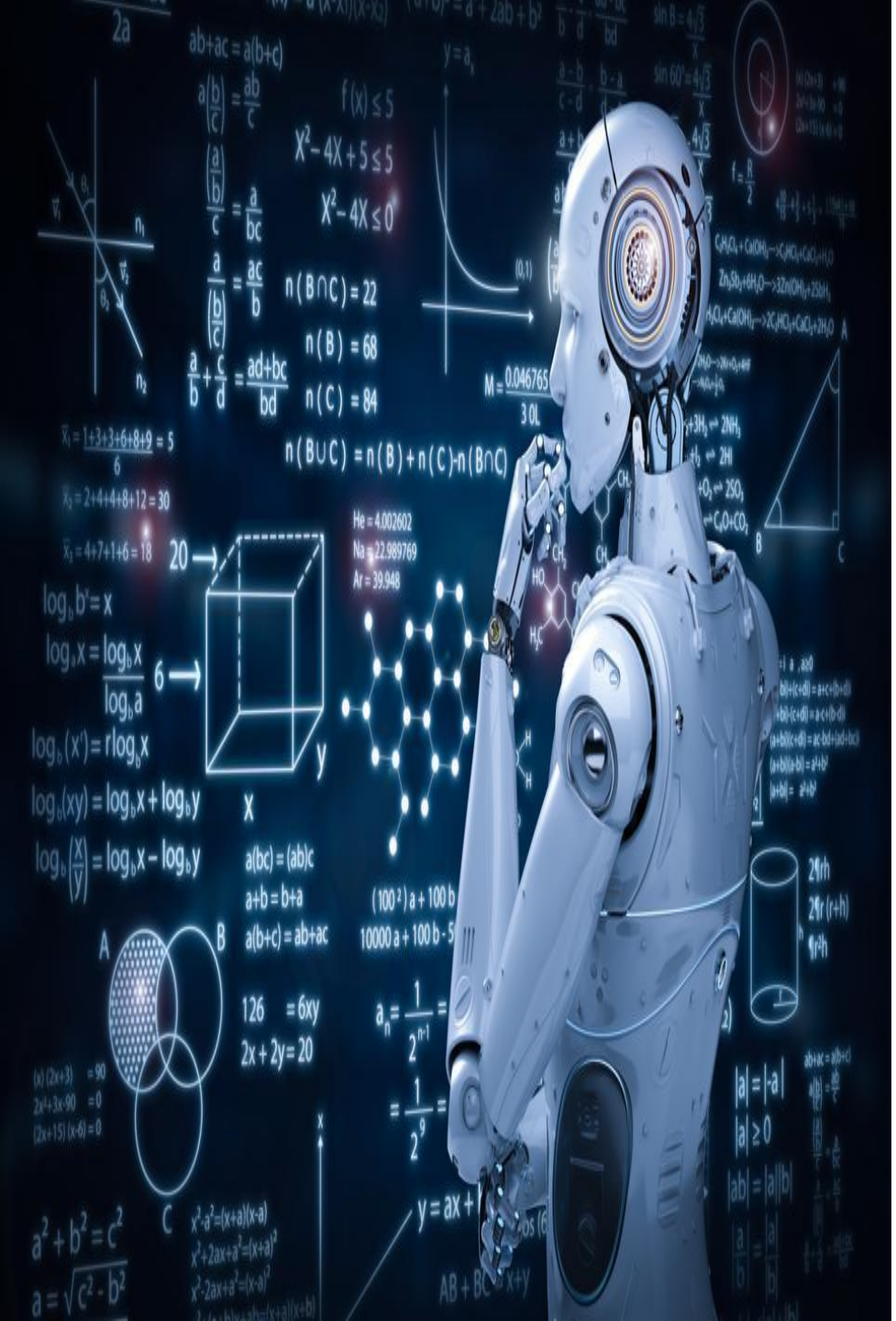


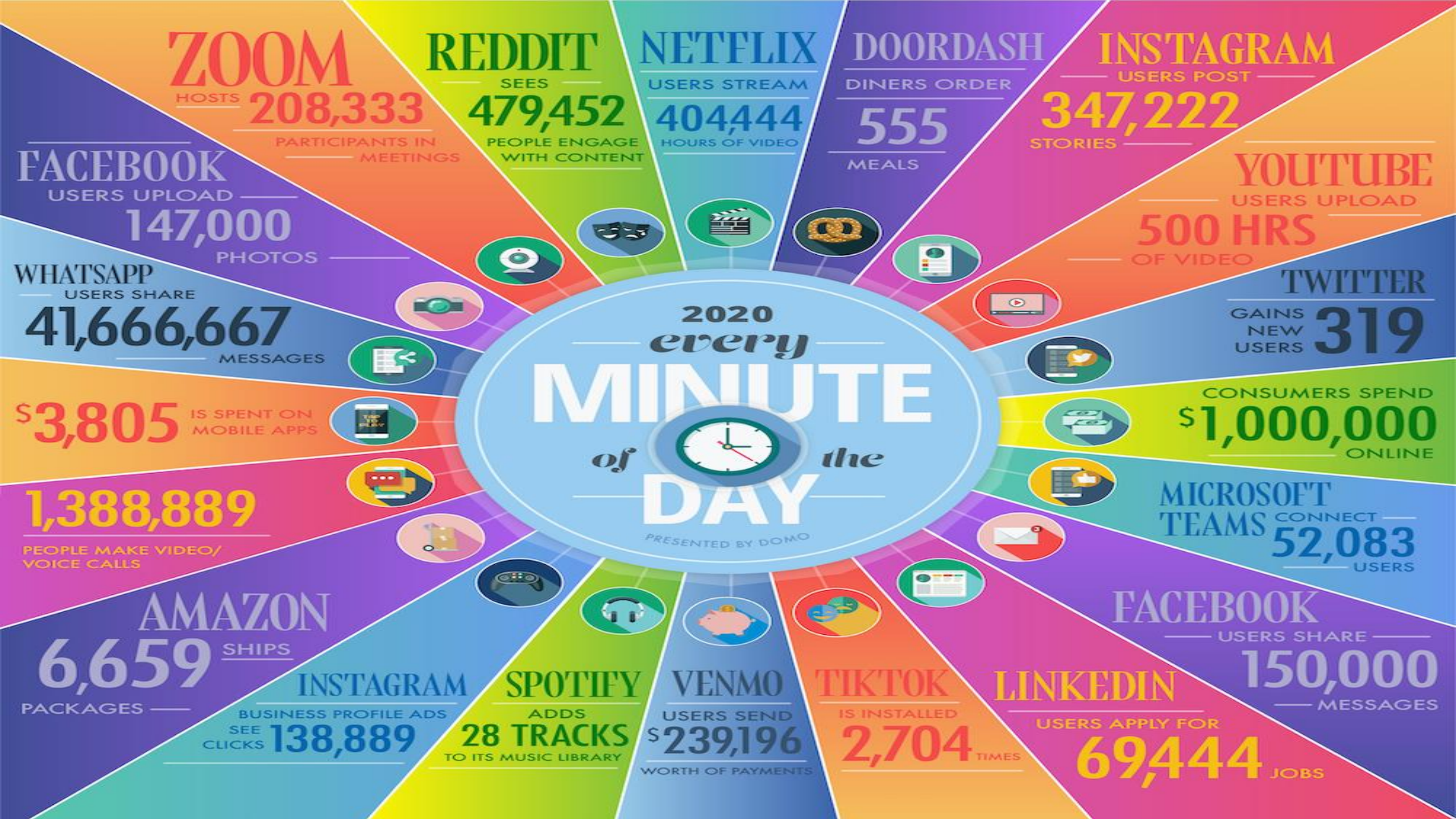
# Practical Machine Learning

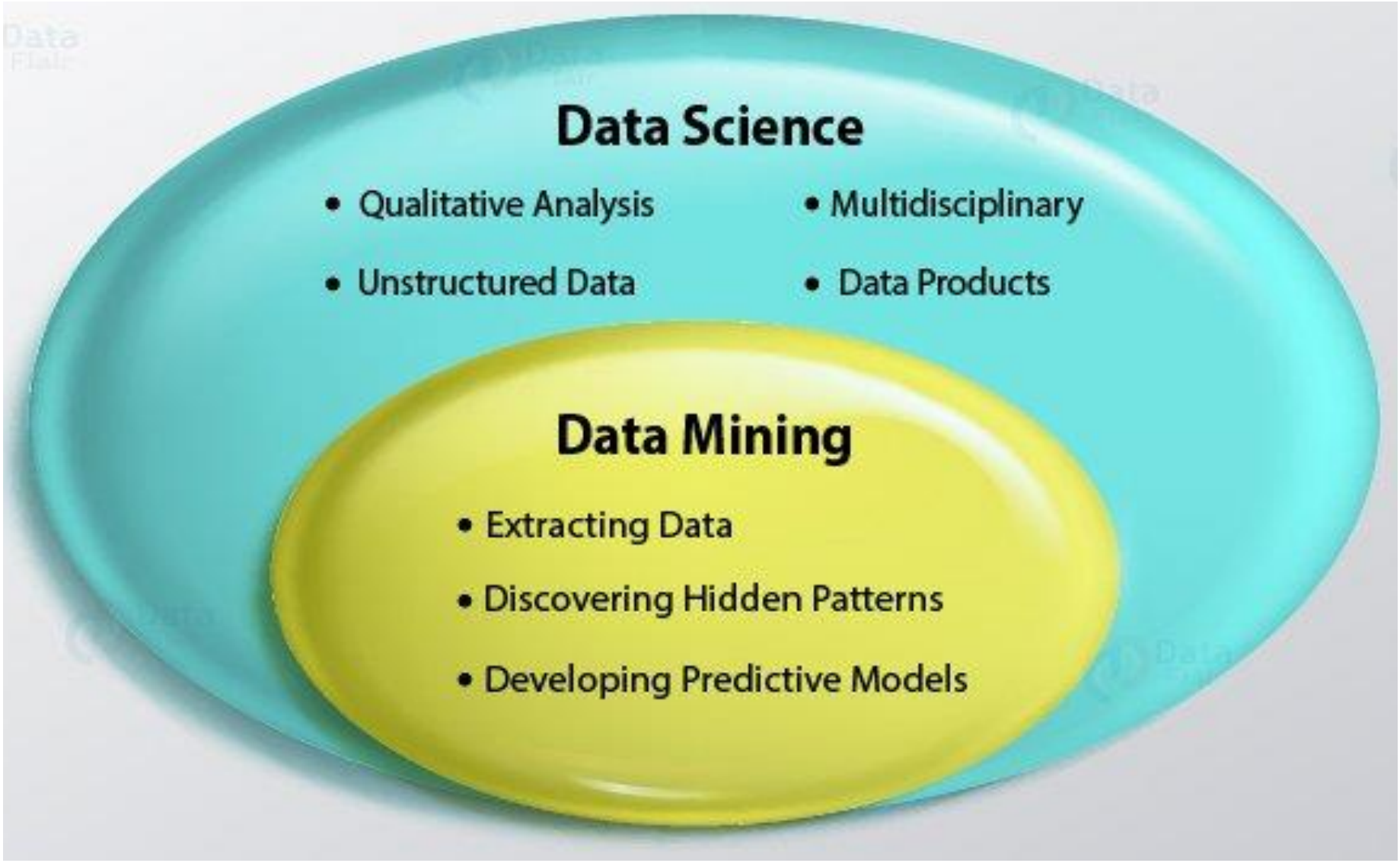
## Day 3: SEP23 DBDA

Kiran Waghmare

# Agenda

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


2020
every
MINUTE
of the
DAY
PRESENTED BY DOMO
ZOOM
HOSTS 208,333
PARTICIPANTS IN MEETINGS
REDDIT
SEES 479,452
PEOPLE ENGAGE WITH CONTENT
NETFLIX
USERS STREAM 404,444
HOURS OF VIDEO
DOORDASH
DINERS ORDER 555
MEALS
INSTAGRAM
USERS POST 347,222
STORIES
FACEBOOK
USERS UPLOAD 147,000
PHOTOS
YOUTUBE
USERS UPLOAD 500 HRS
OF VIDEO
WHATSAPP
USERS SHARE 41,666,667
MESSAGES
TWITTER
GAINS NEW USERS 319
$3,805 IS SPENT ON MOBILE APPS
CONSUMERS SPEND $1,000,000 ONLINE
1,388,889
PEOPLE MAKE VIDEO/ VOICE CALLS
MICROSOFT TEAMS CONNECT 52,083
USERS
AMAZON
6,659 SHIPS
PACKAGES
FACEBOOK
USERS SHARE 150,000
MESSAGES
INSTAGRAM
BUSINESS PROFILE ADS SEE CLICKS 138,889
SPOTIFY
ADDS 28 TRACKS
TO ITS MUSIC LIBRARY
VENMO
USERS SEND $239,196
WORTH OF PAYMENTS
TIKTOK
IS INSTALLED 2,704 TIMES
LINKEDIN
USERS APPLY FOR 69,444 JOBS


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

# 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

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 |

# 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 |

# 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 data set can be represented by an *m* x *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 |

# 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 |

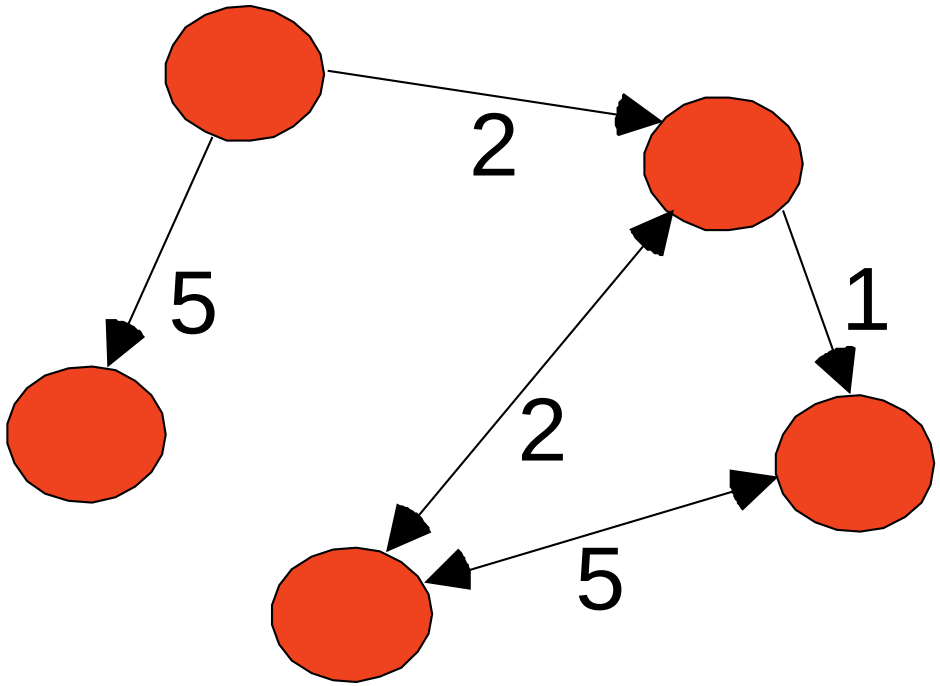
# Transaction data

- A special type of record data, where
  - Each record (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.

| *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

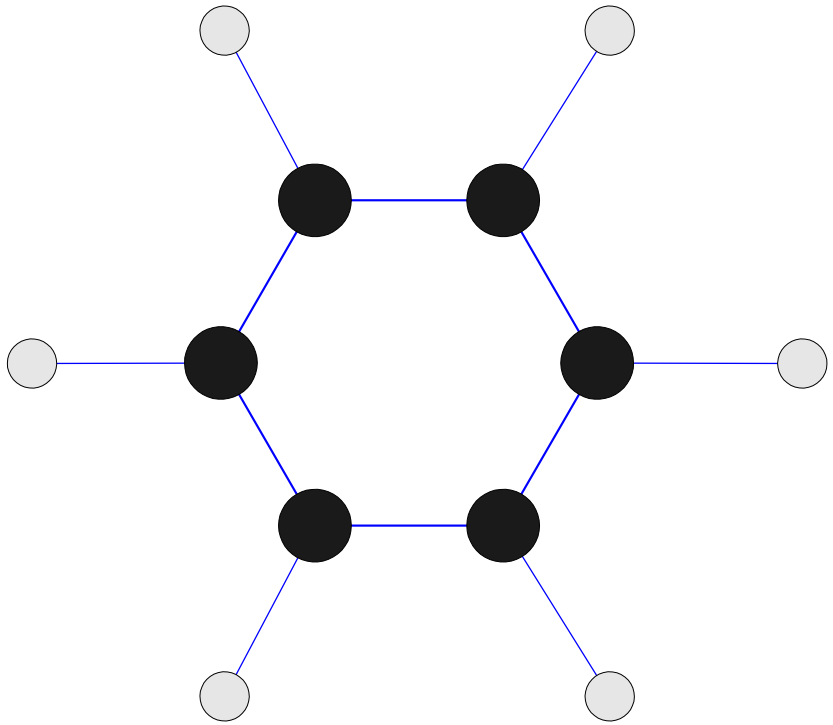
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

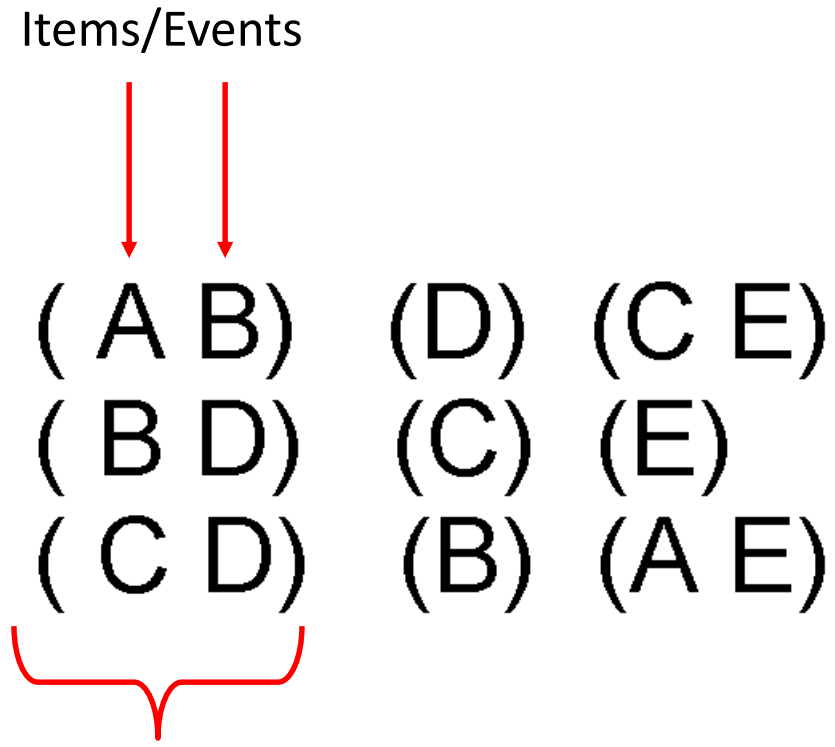
# Chemical data

- Benzene molecule: $C_6H_6$



# Ordered data

- Sequences of transactions



# Ordered data

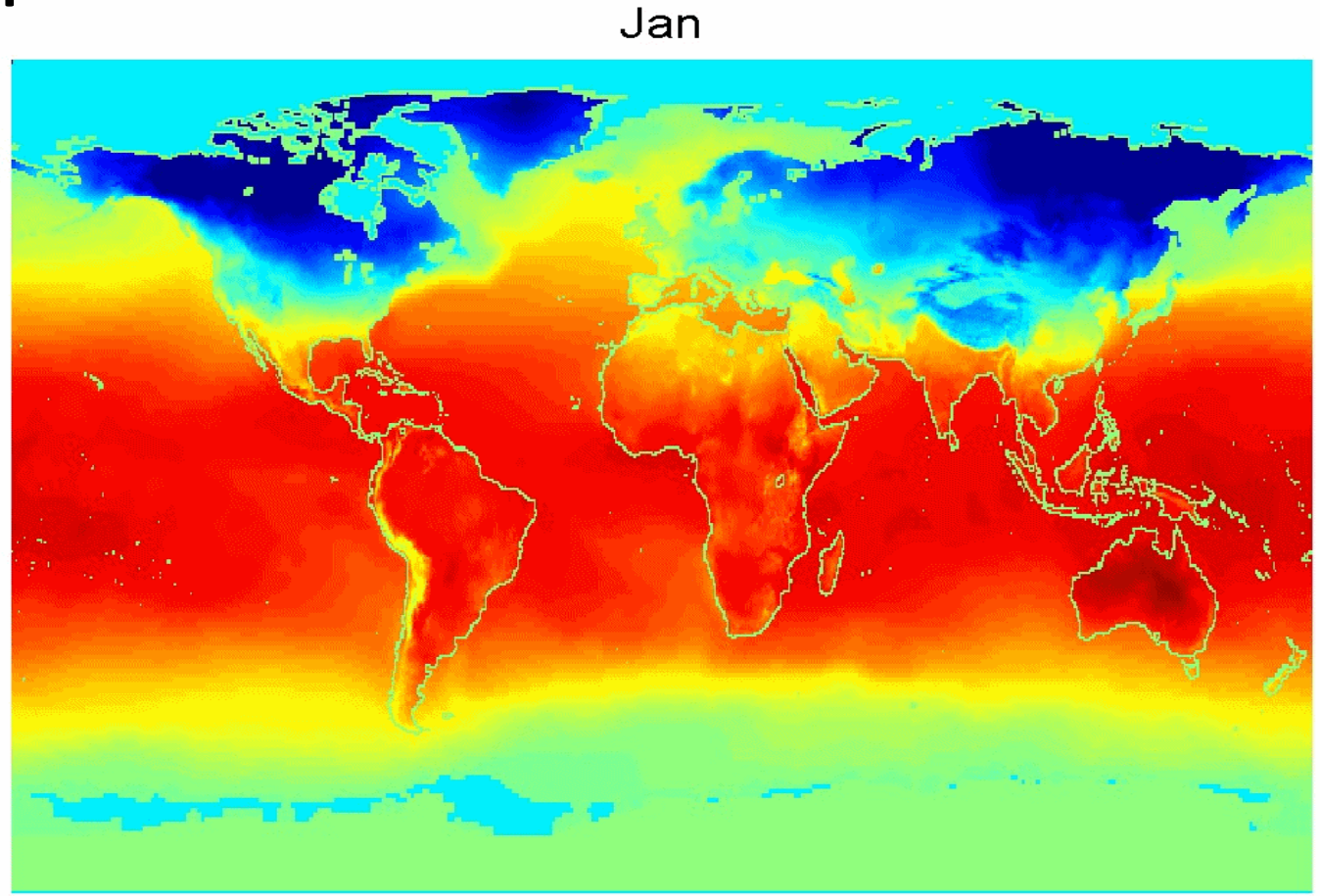
- Genomic sequence data

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

# Ordered data

- Spatio-temporal data

Average monthly temperature of land and ocean



# 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.

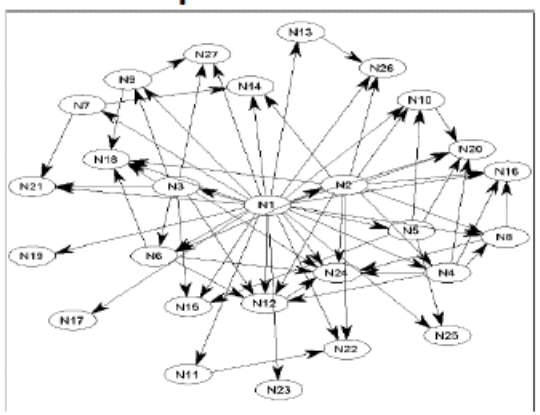
# Approximating Images and Graphs

- We can think of other data types in this way:
  - Images:

graycale intensity

| (1,1) | (2,1) | (3,1) | … | (m,1) | … | (m,n) |
|---|---|---|---|---|---|---|
| 45 | 44 | 43 | … | 12 | … | 35 |

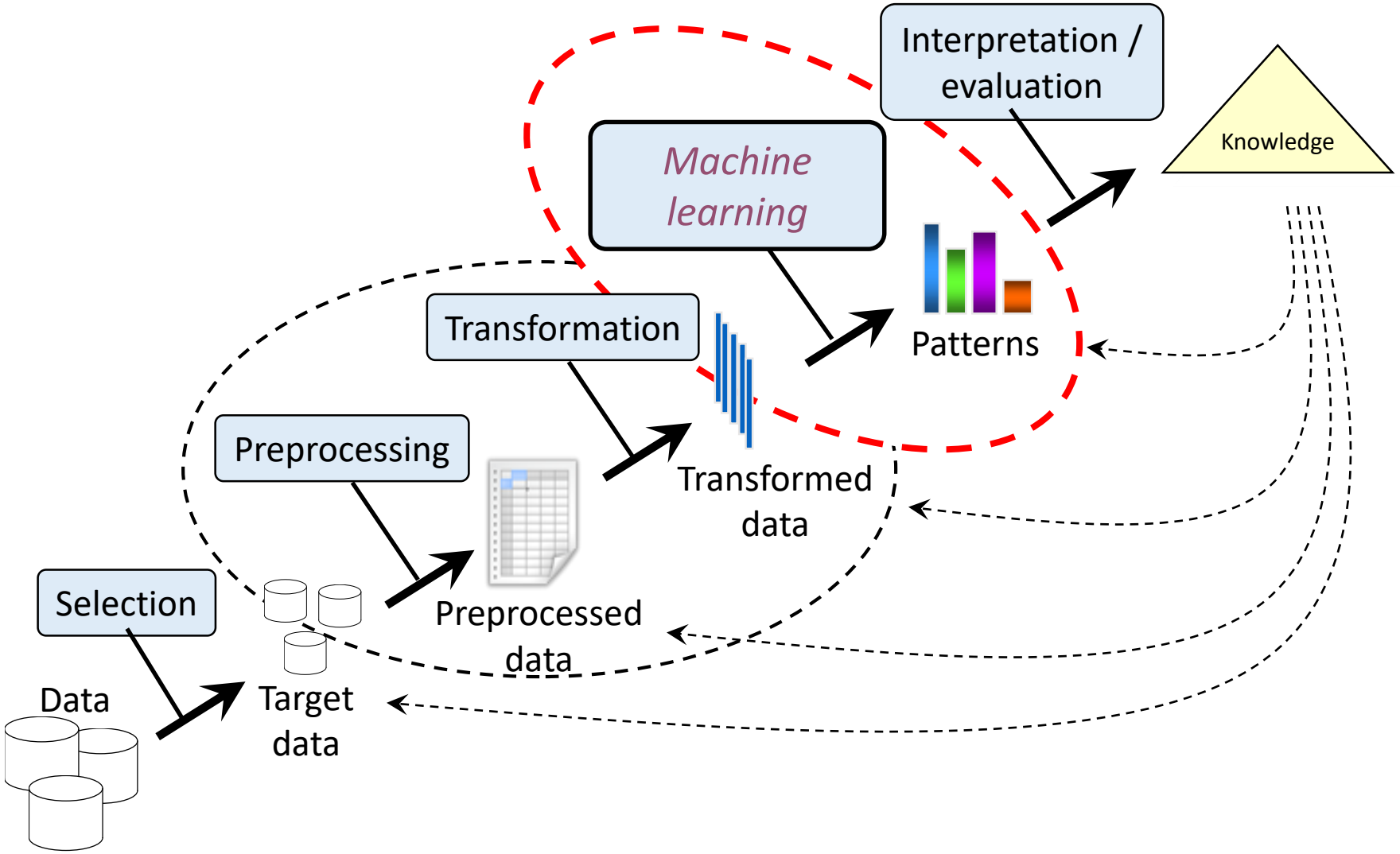
  - Graphs:

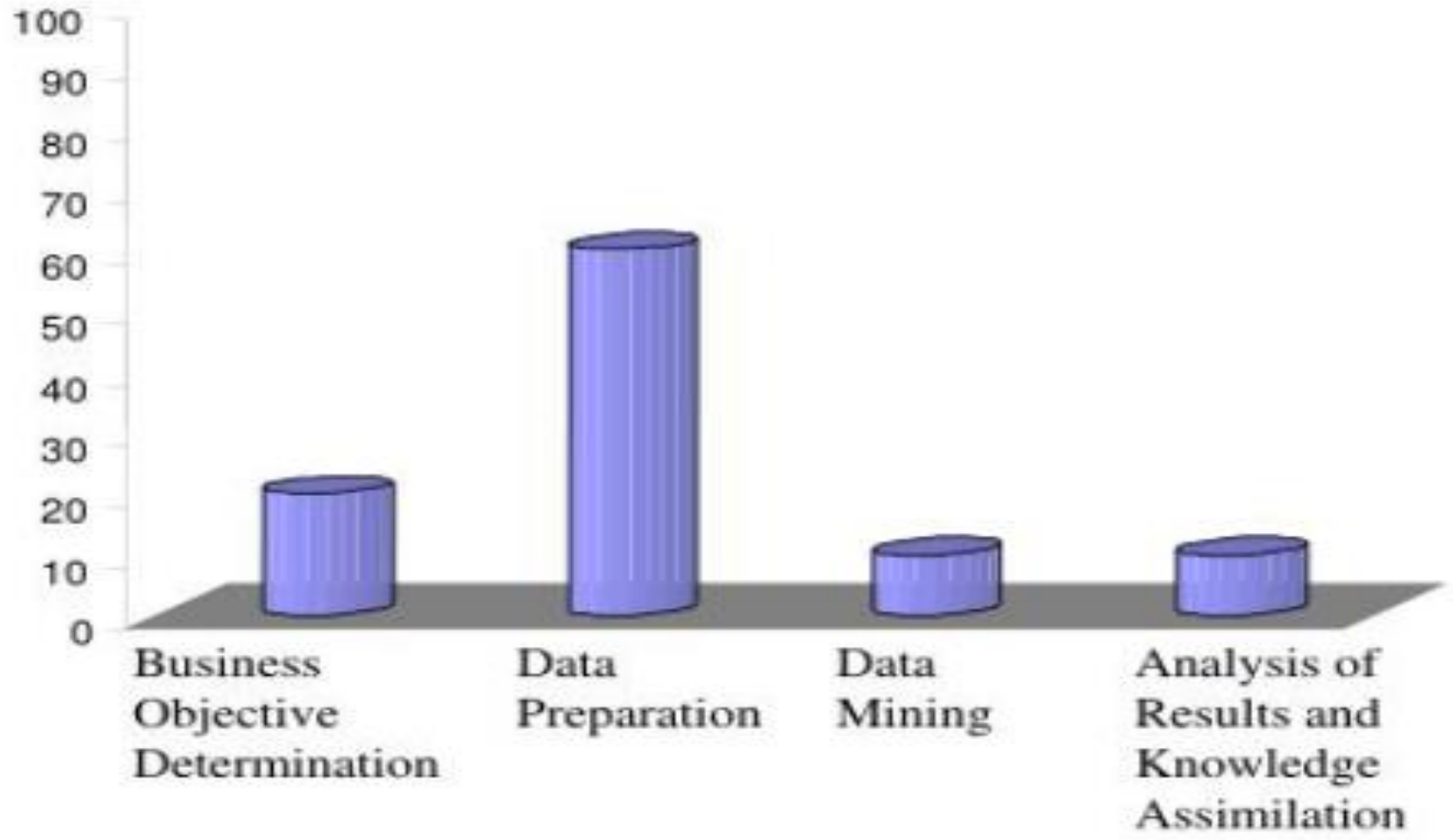



adjacency matrix

| 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 |

# Stages of knowledge extraction






Effort for each data-mining process step

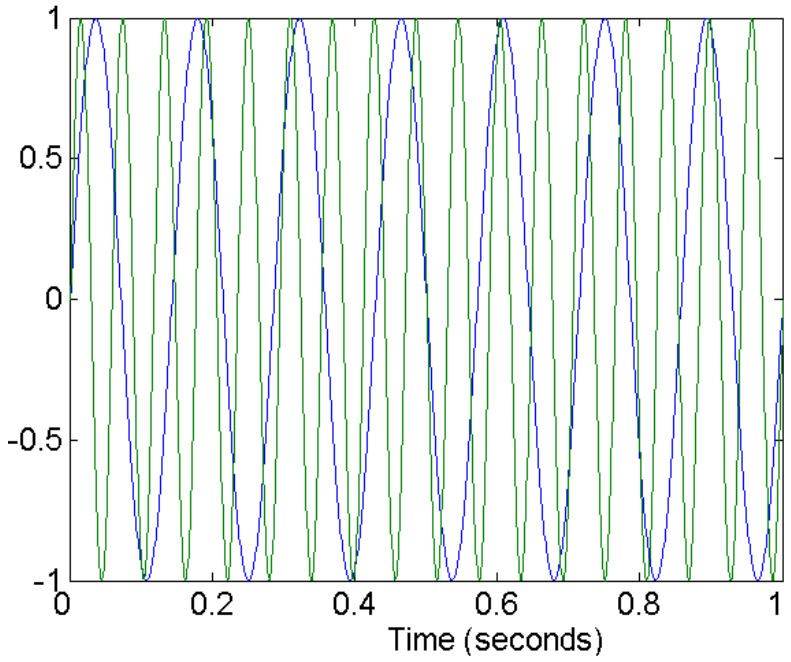
# Data quality

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

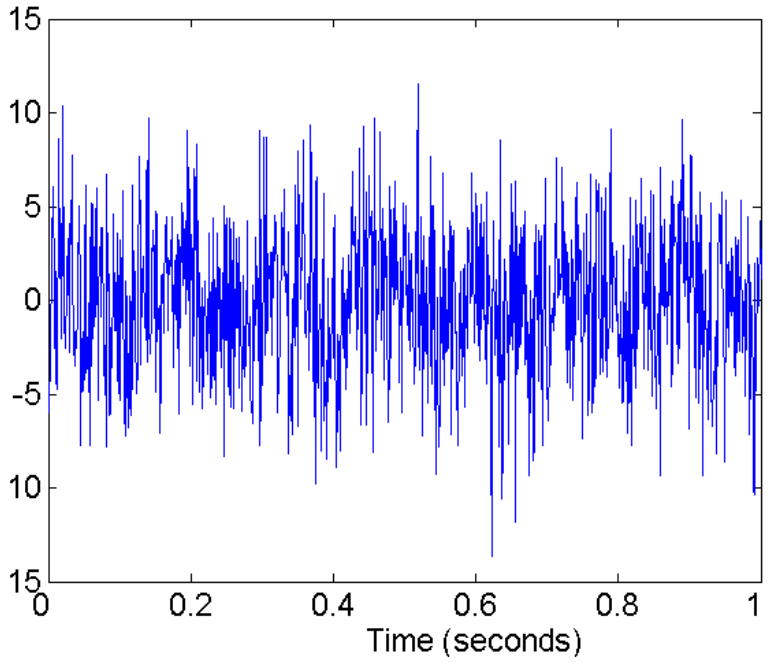
- Examples of data quality problems:
  - noise and outliers
  - missing values
  - duplicate data

# Noise

- Noise refers to random modification of original values
- Examples:
  - distortion of a person's voice when talking on a poor phone
  - "snow" on television screen




Two sine waves



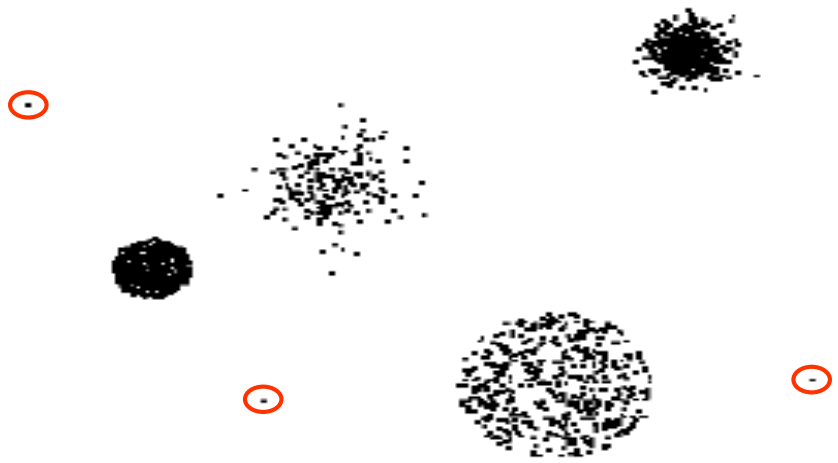

Two sine waves + noise

# Noise

- Dealing with noise
  - Mostly you have to live with it
  - Certain kinds of smoothing or averaging can be helpful
  - In the right domain (e.g. signal processing), transformation to a different space can get rid of majority of noise

# Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



# Outliers

- **Dealing with outliers**
  - There are robust statistical methods for detecting outliers
  - In some situations, you want to get rid of outliers
    - but be judicious – they may carry useful, even important information
  - In other situations, the outliers are the objects of interest
    - anomaly detection

# 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
  - Estimate missing values (imputation)
  - Ignore the missing value during analysis
  - Replace with all possible values (weighted by their probabilities)

# Duplicate data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources
- Example:
  - Same person with multiple email addresses
- Data cleaning
  - Includes process of dealing with duplicate data issues

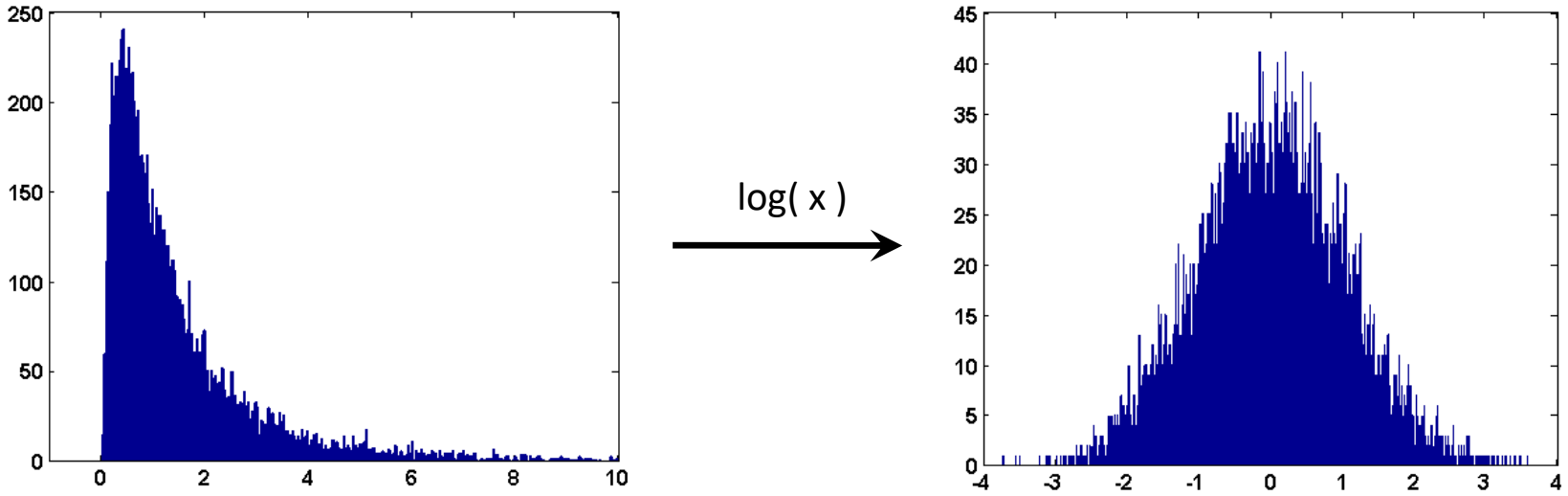
# Attribute transformation

Definition:

A function that maps the entire set of values of a given attribute to a new set of replacement values, such that each old value can be identified with one of the new values.

# Attribute transformation

- Simple functions
  - Examples of transform functions:
    $x^k$ $\log(x)$ $e^x$ $|x|$
  - Often used to make the data more like some standard distribution, to better satisfy assumptions of a particular algorithm.
    - Example: discriminant analysis explicitly models each class distribution as a multivariate Gaussian



# Attribute transformation

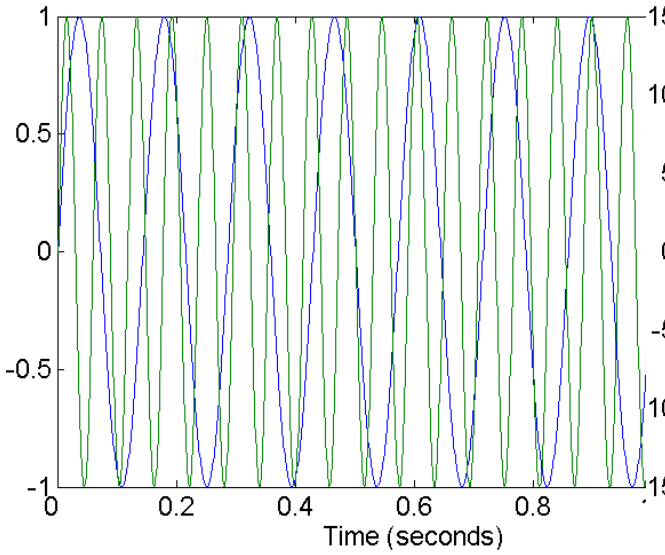
- Standardization or normalization
  - Usually involves making attribute:

    mean = 0
    standard deviation = 1

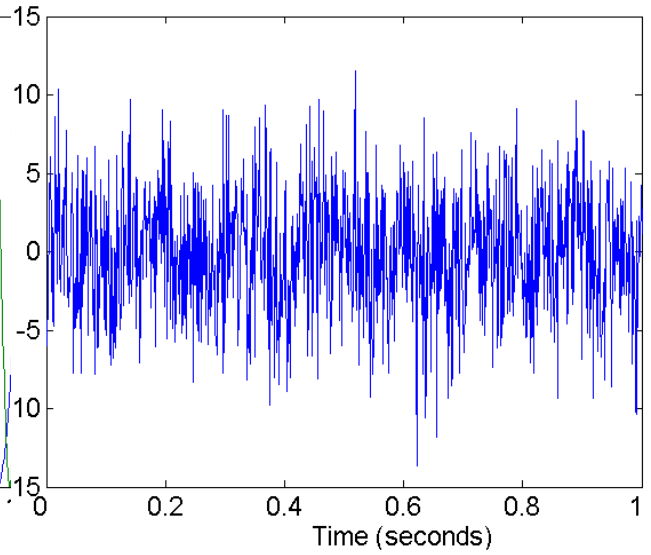
    - in MATLAB, use zscore() function
  - Important when working in Euclidean space and attributes have very different numeric scales.
  - Also necessary to satisfy assumptions of certain algorithms.
    - Example: principal component analysis (PCA) requires each attribute to be mean-centered (i.e. have mean subtracted from each value)

# Transform data to a new space

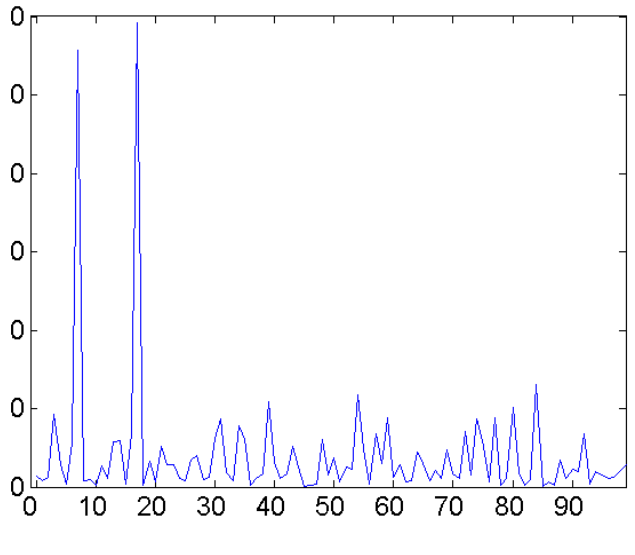
- Fourier transform
  - Eliminates noise present in time domain




Two sine waves




Two sine waves + noise




Frequency

# Feature Transformation

# Converting to Numerical Features

- Often want a real-valued example representation:

| 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 |

- This is called a "1 of k" encoding.
- We can now interpret examples as points in space:
  - E.g., first example is at (23,1,0,0,22000).

# 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.

# Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.

| Age |
|---|
| 23 |
| 23 |
| 22 |
| 25 |
| 19 |
| 22 |

→

| < 20 | >= 20, < 25 | >= 25 |
|---|---|---|
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |

- Only need consider 3 values.

# Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, logarithm, and so on.


# Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.
  - Square, exponentiation, or take logarithm.
  - Scaling: convert variables to comparable scales
    (E.g., convert kilograms to grams.)

# Exploratory Data Analysis

- You should always ‘look’ at the data first.

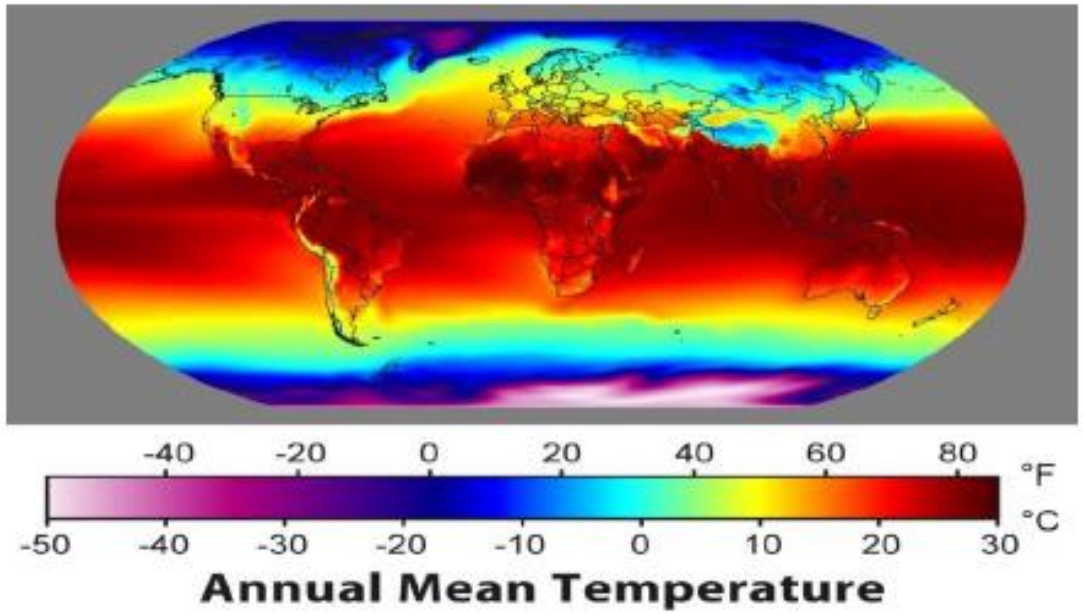
- But how do you ‘look’ at features and high-dimensional examples?
  - Summary statistics.
  - Visualization.
  - ML + DM (later in course).

# Visualization

- You can learn a lot from 2D plots of the data:
  - Patterns, trends, outliers, unusual patterns.

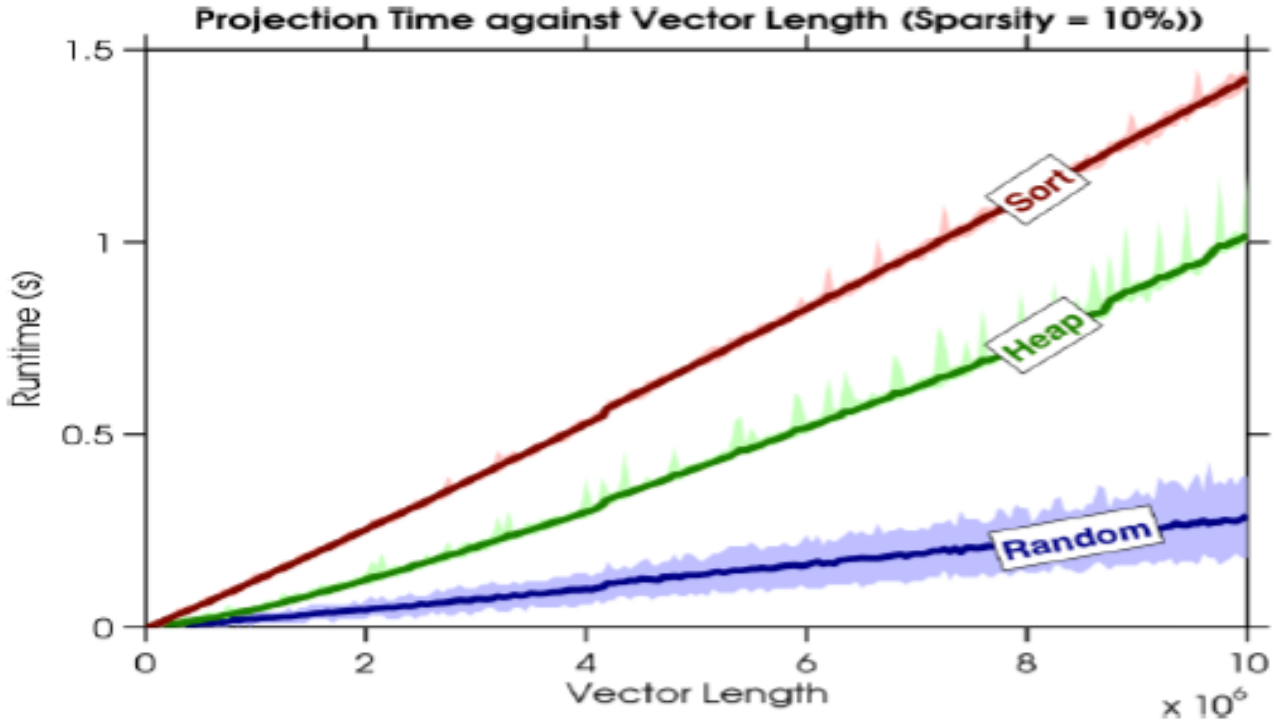
| Lat | Long | Temp |
|---|---|---|
| 0 | 0 | 30.1 |
| 0 | 1 | 29.8 |
| 0 | 2 | 29.9 |
| 0 | 3 | 30.1 |
| 0 | 4 | 29.9 |
| … | … | … |

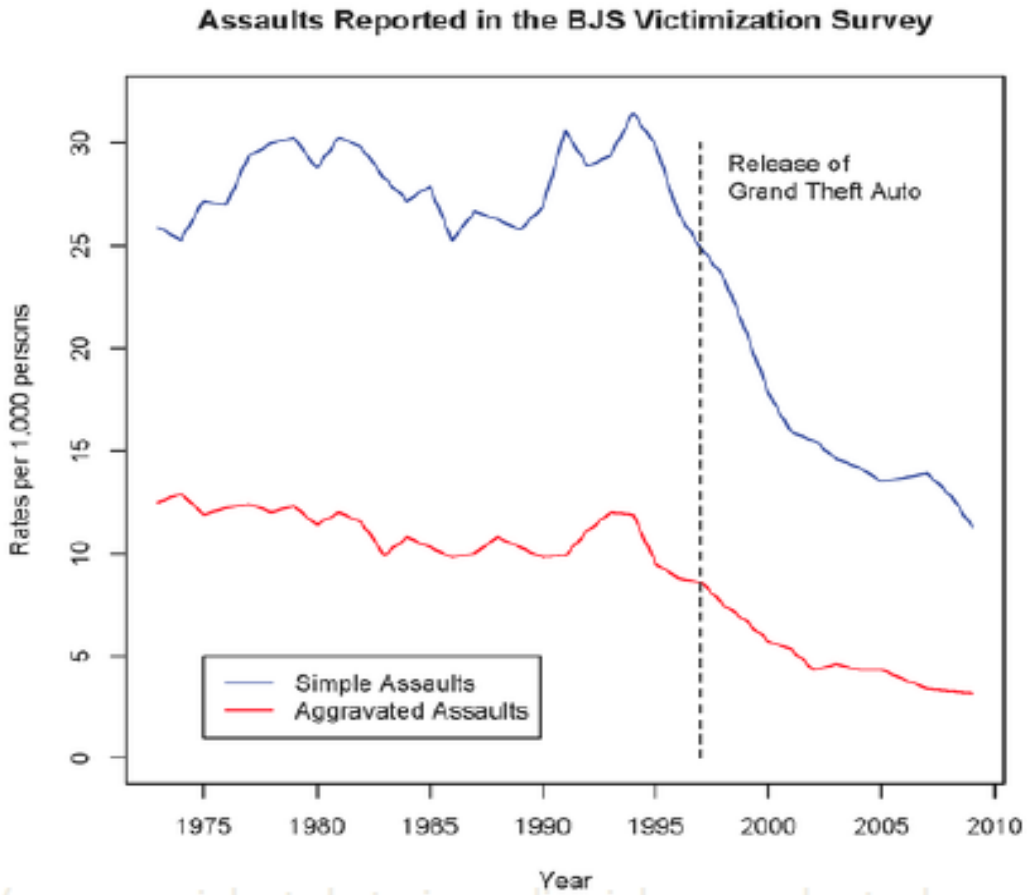
vs.



# Basic Plot

- Visualize one variable as a function of another.



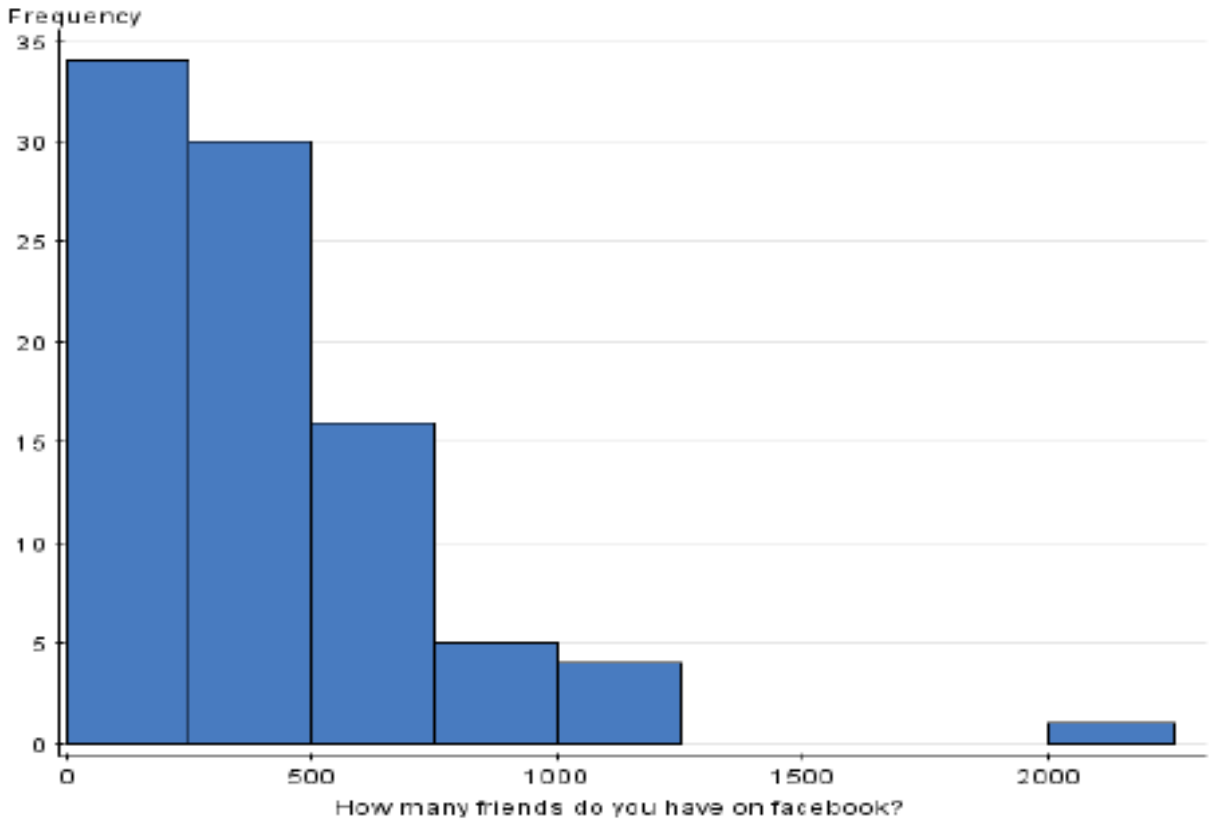




- Fun with plots.

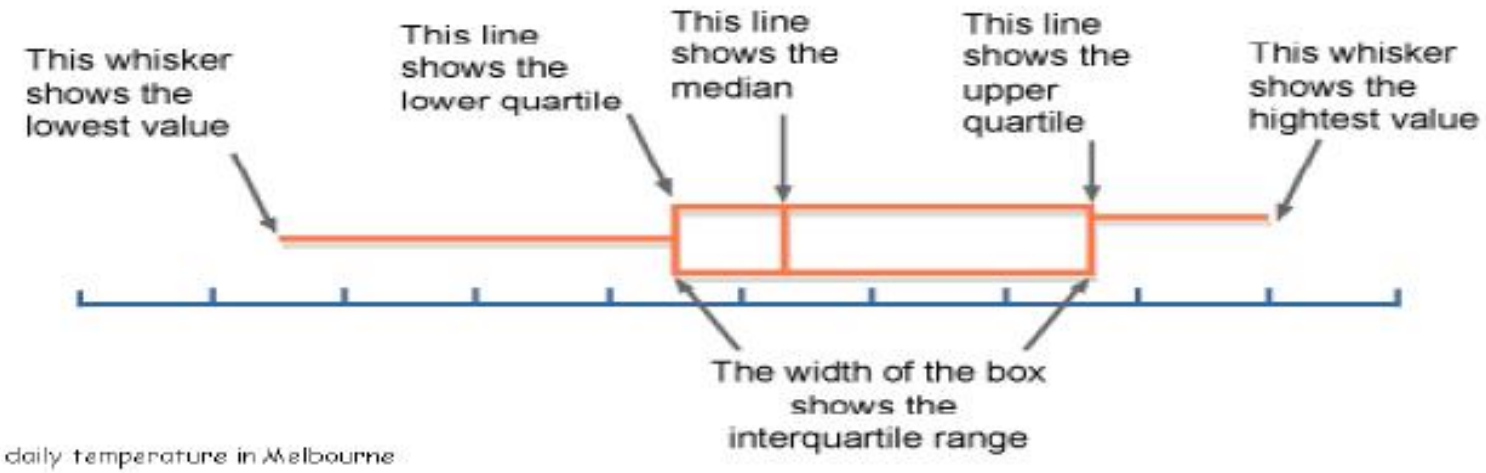
http://notunlikeresearch.typepad.com/something-not-unlike-rese/2011/01/more-on-violent-rhetoric-media-violence-and-actual-

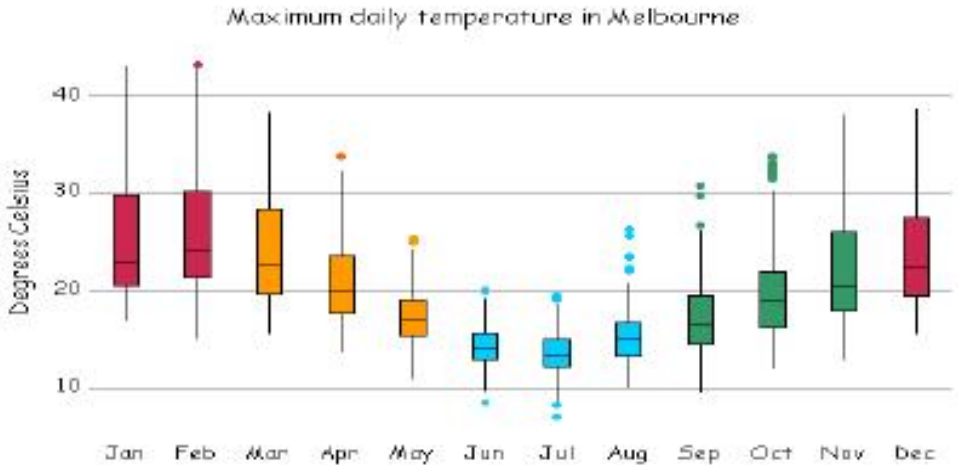
# Histogram

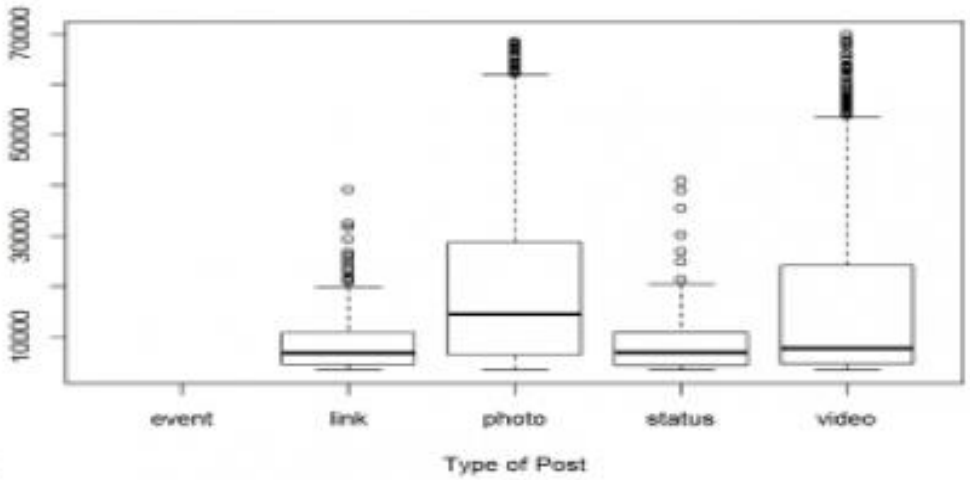
- Histograms display distribution of a variable.



# Box Plot




http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hi

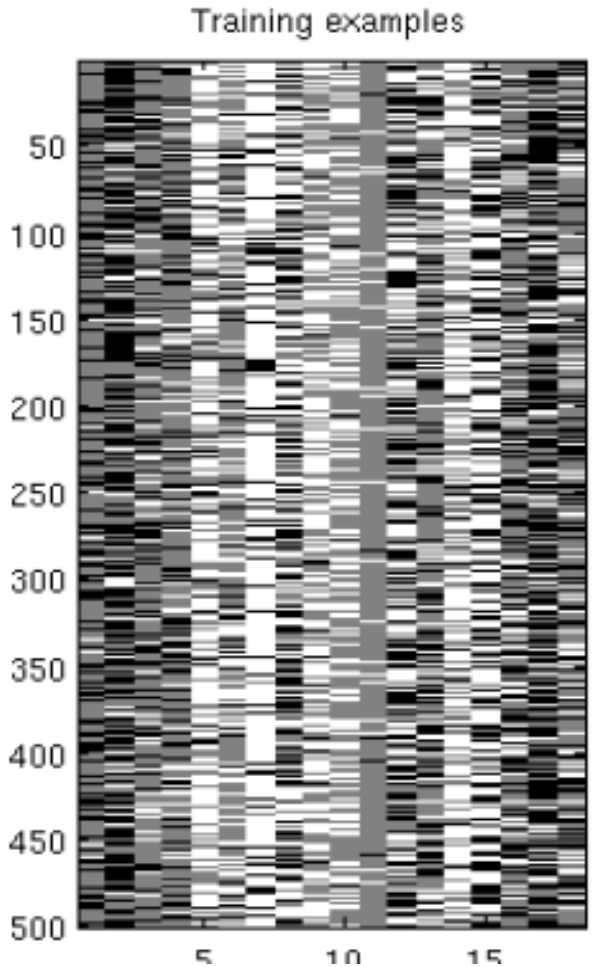
http://www.ccc.ms.unimelb.edu.au/whatisstatistics/weather.html

# Box Plot

- Photo from CTV Olympic coverage in 2010:

# Matrix Plot

- We can view (examples) x (features) data table as a picture:
  - "Matrix plot".
  - May be able to see trends in features.



# Matrix Plot

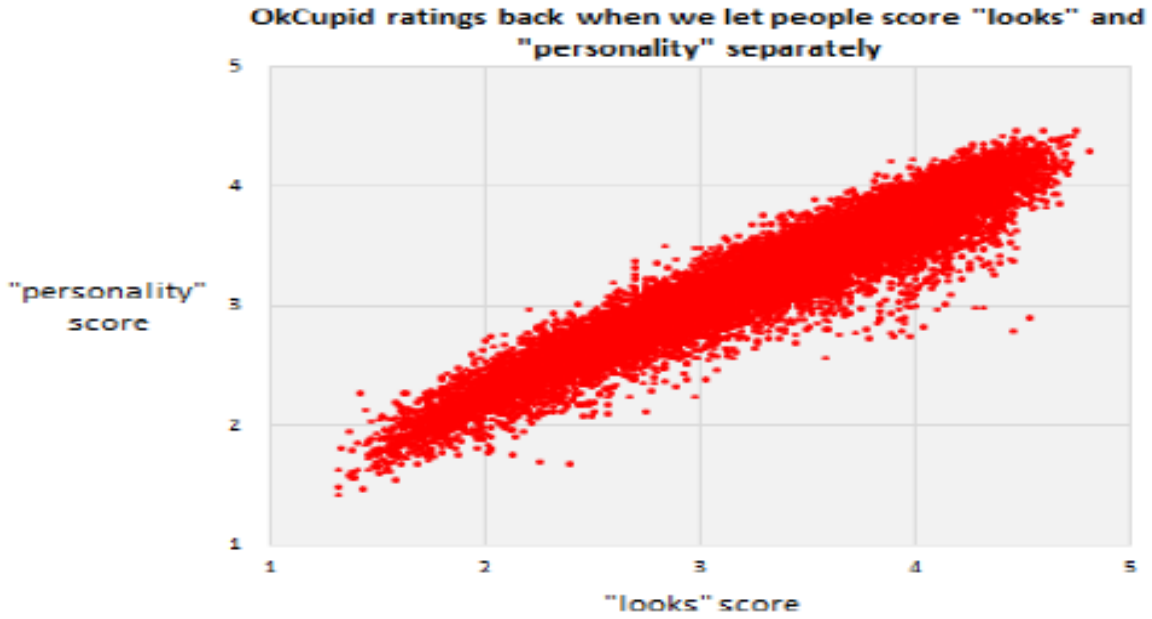
- A matrix plot of all similarities (or distances) between features:
  - Colour used to catch attention.

| | BTC | ETH | XRP | XEM | ETC | LTC | DASH | XMR |
|---|---|---|---|---|---|---|---|---|
| BTC | 1.00 | 0.61 | 0.36 | 0.51 | 0.60 | 0.56 | 0.55 | 0.66 |
| ETH | 0.61 | 1.00 | 0.28 | 0.49 | 0.68 | 0.43 | 0.70 | 0.64 |
| XRP | 0.36 | 0.28 | 1.00 | 0.48 | 0.08 | 0.35 | 0.40 | 0.44 |
| XEM | 0.51 | 0.49 | 0.48 | 1.00 | 0.40 | 0.43 | 0.47 | 0.52 |
| ETC | 0.60 | 0.68 | 0.08 | 0.40 | 1.00 | 0.47 | 0.56 | 0.53 |
| LTC | 0.56 | 0.43 | 0.35 | 0.43 | 0.47 | 1.00 | 0.59 | 0.67 |
| DASH | 0.55 | 0.70 | 0.40 | 0.47 | 0.56 | 0.59 | 1.00 | 0.74 |
| XMR | 0.66 | 0.64 | 0.44 | 0.52 | 0.53 | 0.67 | 0.74 | 1.00 |

"Correlation plot"

# Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.

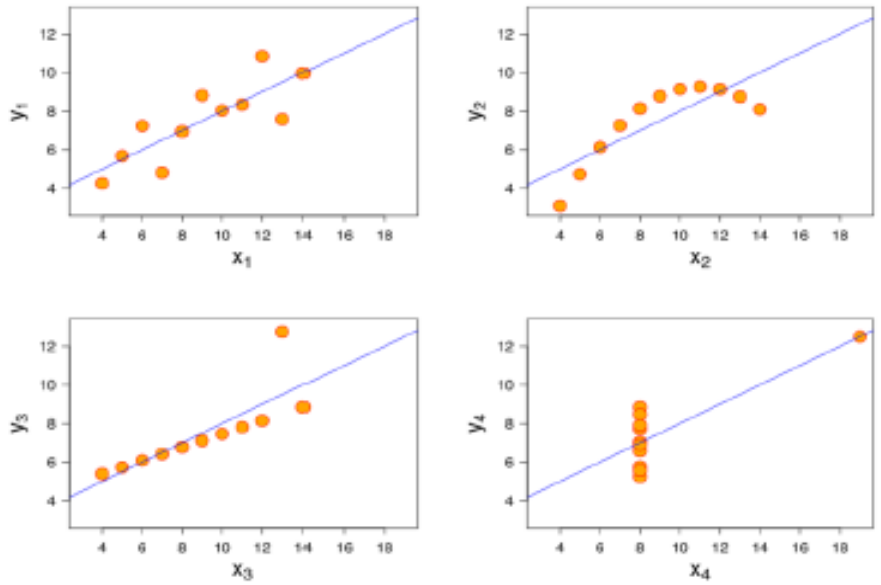



http://cdn.okccdn.com/blog/humanexperiments/looks-v-personality.png

- Shows correlation between “personality” score and “looks” score.

# Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.




https://en.wikipedia.org/wiki/Anscombe%27s_quartet

- Shows correlation between “personality” score and “looks” score.
- But scatterplots let you see more complicated patterns.