数据挖掘导论

Introduction to Data Mining

Getting to Know Your Data

数据智能实验室
DATA INTELLIGENCE LABORATORY
DI-LAB
浙江大学
Zhejiang University

Agenda

- Data Representation
- Data Statistics
- Data Visualization
- Data Similarity
- Summary

Data Types

Structured Data

Often numbers or labels, stored in a structured framework of columns and rows relating to pre-set parameters.

- ID CODES IN DATABASES
- NUMERICAL DATA GOOGLE SHEETS
- STAR RATINGS

Semi-unstructured Data

Loosely organized into categories using meta tags

- EMAILS BY INBOX, SENT, DRAFT
- TWEETS ORGANIZED BY HASHTAGS
- FOLDERS ORGANIZED BY TOPIC

Unstructured Data

Text-heavy information that's not organized in a clearly defined framework or model.

- MEDIA POSTS, EMAILS, ONLINE REVIEWS
- VIDEOS, IMAGES
- SPEECH, SOUNDS

Image source: https://monkeylearn.com/blog/semi-structured-data/

Data Types

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Image source: Jorge Cardoso

Structured Data

Relational Record

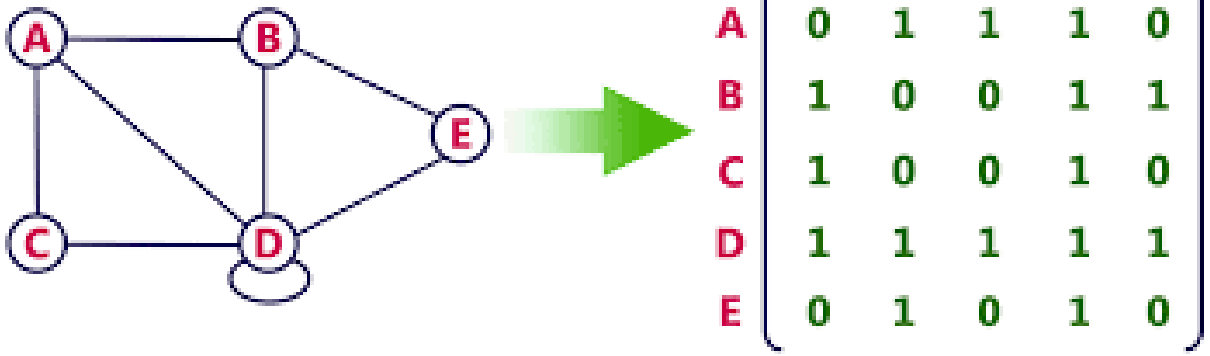
- ✓ Each relational table contains multiple attributes
- ✓ Records in the same table follow the same schema

Avg. Income	Avg. House Age	Avg. Number of Rooms	Avg. Number of Bedrooms	Avg. Population	Avg. Price	Address
79545	5.682	7.009	4.09	23086	1059033	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
79248	6.002	6.730	3.09	40173	1,505,890	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
61287	5.865	8.512	5.13	36882	1,058,987	9127 Elizabeth Stravenue Danieltown, WI 06482-3489

Structured Data

- **Matrix Data**
  - ✓ Scientific Data
  - ✓ Image Features
  - ✓ Graphs

Feature Vector Number	Features					
	Peak Value	Front time	Tail Time	….	Bandwidth	Spectral Entropy
1	0.75	0.34	0.45	….	0.71	0.39
2	0.79	0.33	0.45	….	0.72	0.41
3	0.74	0.33	0.46	….	0.67	0.41
:	:	:	:	:	:	:
N	0.78	0.35	0.48	….	0.70	0.38



Attribute Types

- **Numerical**
  - ✓ feature values, medical indicators, weather temperature, stock price
- **String**
  - ✓ address, product description

Avg. Income	Avg. House Age	Avg. Number of Rooms	Avg. Number of Bedrooms	Avg. Population	Avg. Price	Address
79545	5.682	7.009	4.09	23086	1059033	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
79248	6.002	6.730	3.09	40173	1,505,890	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
61287	5.865	8.512	5.13	36882	1,058,987	9127 Elizabeth Stravenue Danieltown, WI 06482-3489

Attribute Types

- **Nominal: categories, states, or "names of things"**
  - ✓ hair_color = {auburn, black, blond, brown, grey, red, white}
  - ✓ marital status, occupation, ID numbers, zip codes
- **Binary**
  - ✓ Nominal attribute with only 2 states (0 and 1)
    - e.g., medical test (positive vs. negative)
- **Ordinal**
  - ✓ Values have a meaningful order (ranking) but magnitude between successive values is not known.
    - e.g., size = {small, medium, large}, grades, army rankings

Discrete vs. Continuous Attributes

Discrete Attribute

- ✓ Has only a finite or countably infinite set of values
  - e.g., zip codes, profession, or the set of words in a collection of documents
- ✓ Sometimes, represented as integer variables

Continuous Attribute

- ✓ Has real numbers as attribute values
  - e.g., 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

Semi-Structured Data

Sources of semi-structured data

- ✓ Emails
- ✓ HTML
- ✓ Json

```
1  <!DOCTYPE html PUBLIC "-//W3C//DTD
   XHTML 1.0 Transitional//EN"
2  "http://www.w3.org/TR/xhtml1/DTD/
   xhtml1-transitional.dtd">
3
4  <html xmlns="http://www.w3.org/1999/
   xhtml">
5      <head>
6          <meta http-equiv="Content-
   Type" content=
7          "text/html; charset=us-
   ascii" />
8          <script type="text/
   javascript">
9              function reDo() {top.
   location.reload();}
10             if (navigator.appName ==
   'Netscape') {top.onresize = reDo;}
11             dom=document.
   getElementById;
12         </script>
13     </head>
14     <body>
15     </body>
16 </html>
```

```
//restapi.amap.com/v3/geocode/geo?key=您的key&address=方恒国际中心A座&city=北京
```


```
{
    "status" : "1",
    "info" : "OK",
    "infocode" : "10000",
    "count" : "1",
  ⊟ "geocodes" : [
      ⊟ "0" : {
            "formatted_address" : "北京市朝阳区方恒国际中心|A座",
            "country" : "中国",
            "province" : "北京市",
            "citycode" : "010",
            "city" : "北京市",
            "district" : "朝阳区",
            "township" : [ ],
          ⊞ "neighborhood" : { ... },
          ⊞ "building" : { ... },
```

Unstructured Data

- Textual Data
- Audio Data
- Image Data
- Video Data

★★★★★

终于在年底下定决心把P30pro升级为50pro，选择白色是因为黑色落土有手印等感觉会非常明显，金色又感觉有些小富即安，白色雅致干净。今天ROOT CO的壳到了装上非常漂亮，配上背扣指环不怕手滑脱落了，和防撞壳一起"双保险"。

Sequence Data

Reference Genome AATCATGTGTGGCTACTTACTGTCACT

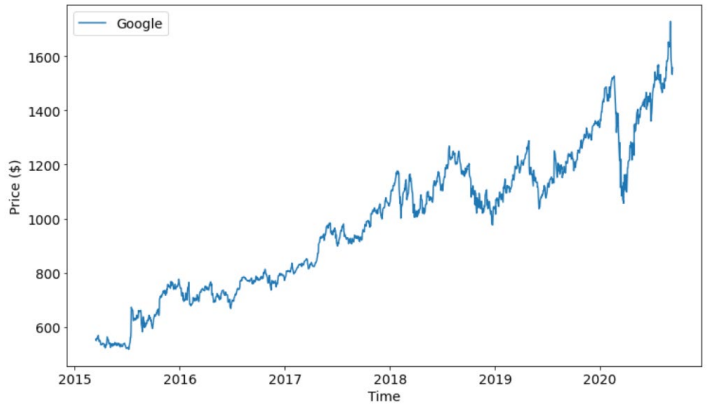
Person's sequenced DNA AATCATGTGTGGCTACTTACTGTCACT

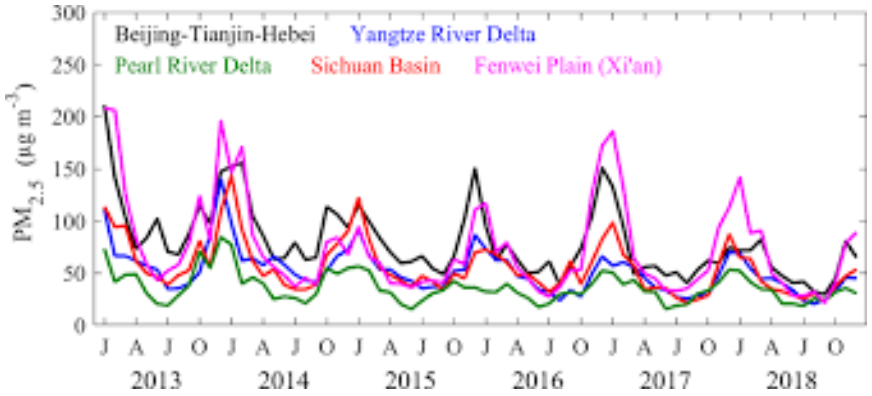
AATCATGTGTAGCTACTTACTGTCACT

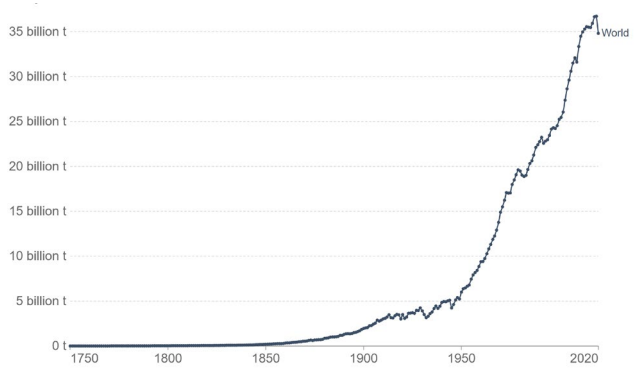
G|A

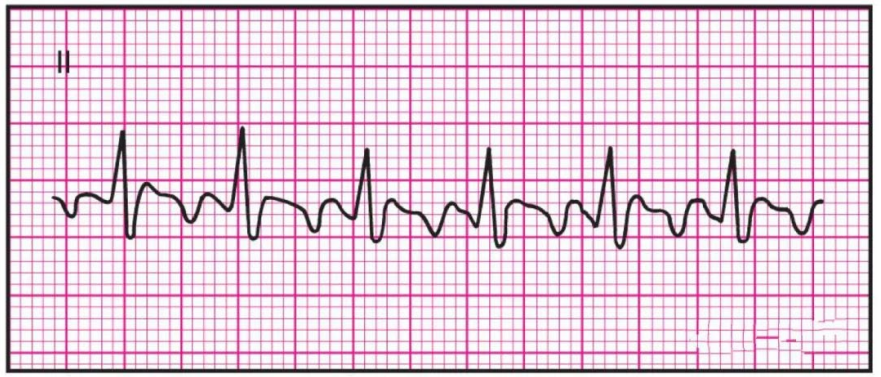
Person's genotype for this variant

Time Series Data

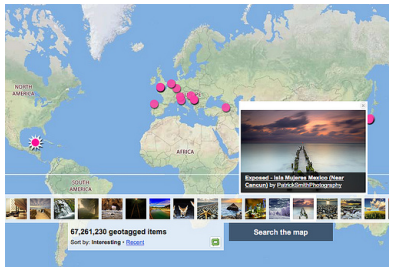








Spatial-Temporal Data



滴滴数据资产

5. 5亿+
用户

1000+座
全球城市

700亿
日ETA请求

150亿
日定位数据

日新增106TB＋数据
日均处理处理4875+TB数据

人
- 司机
- 乘客

路
- 静态信息：全球、全国、城市等各粒度
- 动态信息：行程轨迹

车
- 静态信息：归属人/公司、车牌号...
- 动态信息：里程数、维保记录

Video Database

在大规模视频数据管理和查询优化领域，**缺乏标准化商用数据库产品**

时序数据库 Time Series Database			图数据库 Graph Database			视频数据库 Video Database
influxdb D轮融资	2019年3月	$ 6千万美元	neo4j D轮融资	2021年6月	$ 3.25亿美元	
Timescale C轮融资	2022年2月	$ 1.8亿美元	TigerGraph C轮融资	2021年2月	$ 1.05亿美元	

多模数据库

Video Database

高通量实时视频数据库

标准化SQL查询语言

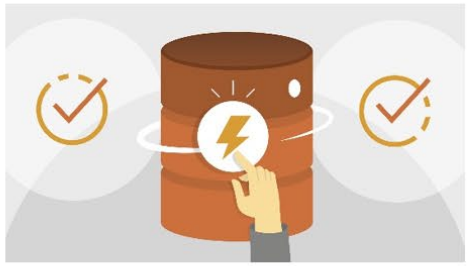
- 内置选择、聚合、连接等常见数据库查询因子
- 降低使用门槛，无需具备AI算法能力

海量结构化和非结构化数据管理

- 实现10万路摄像头产生的高通量视频数据的自动化管理
- 有效集成1000+视觉模型

查询优化引擎与实时检索

- 提供复杂视频查询的优化引擎
- 实现秒级别的实时查询响应能力



视频数据管理的Oracle！

Agenda

- Data Representation
- Data Statistics
- Data Visualization
- Data Similarity
- Summary

Measuring the Central Tendency

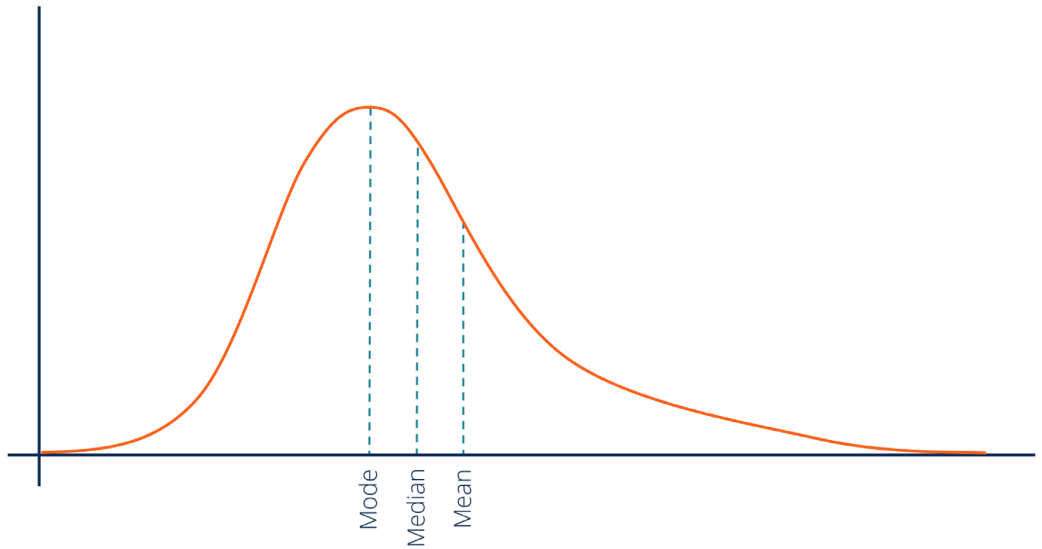
- **Mean**
  - ✓ n is sample size and N is population size

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - ✓ Weighted arithmetic mean:

$$\bar{x} = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i$$

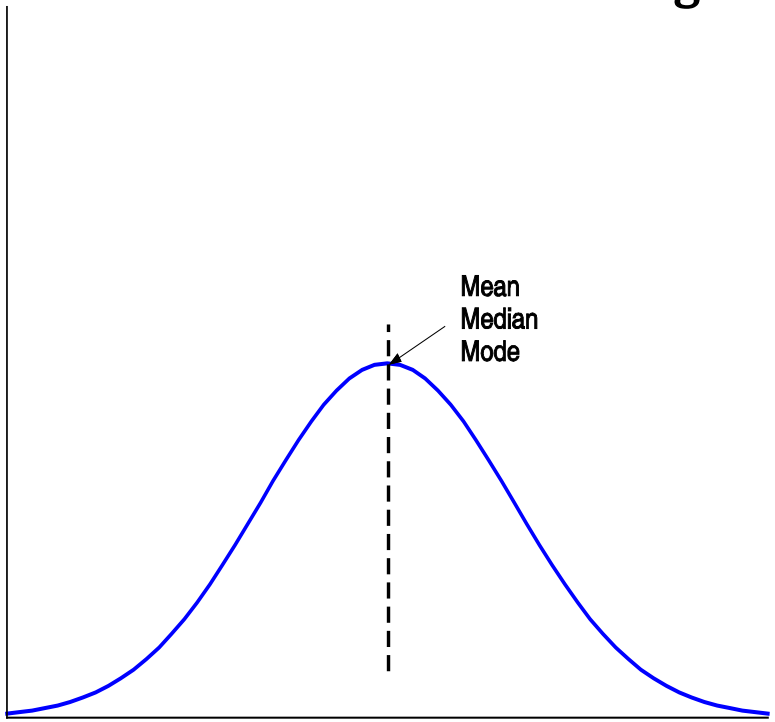
- **Median**
  - ✓ Middle value if odd number of values, or average of the middle two values otherwise
- **Mode**
  - ✓ Value that occurs most frequently in the data



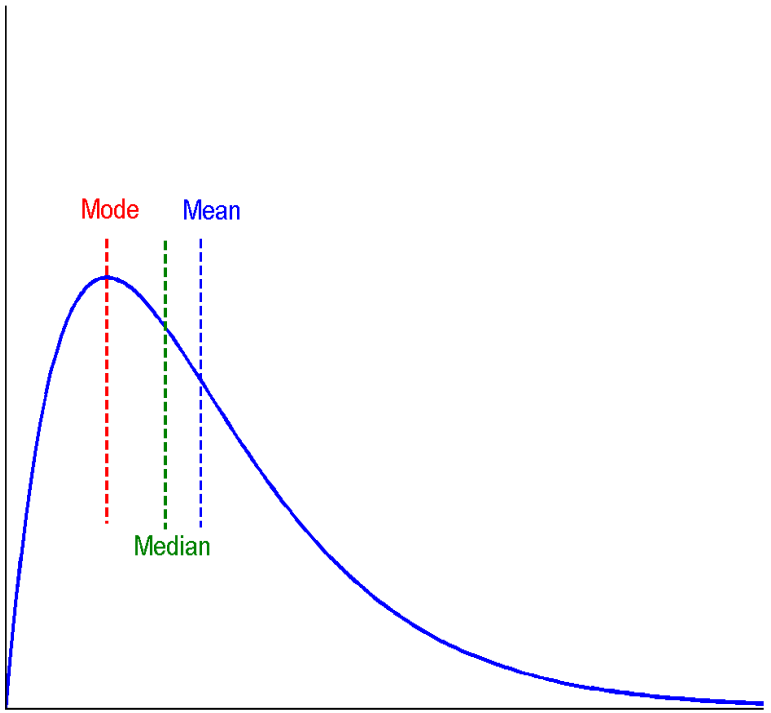
Symmetric vs. Skewed Data

Symmetric, Positively and Negatively Skewed data

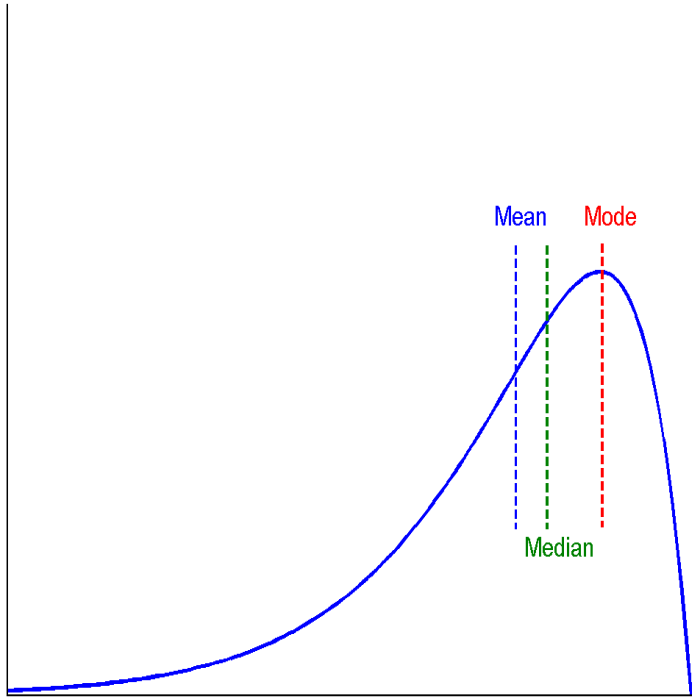
- ✓ A distribution is **positively** skewed if it has a "tail" on the **right** side of the distribution
- ✓ A distribution is **negatively** skewed if it has a "tail" on the **left** side of the distribution



symmetric



positively skewed

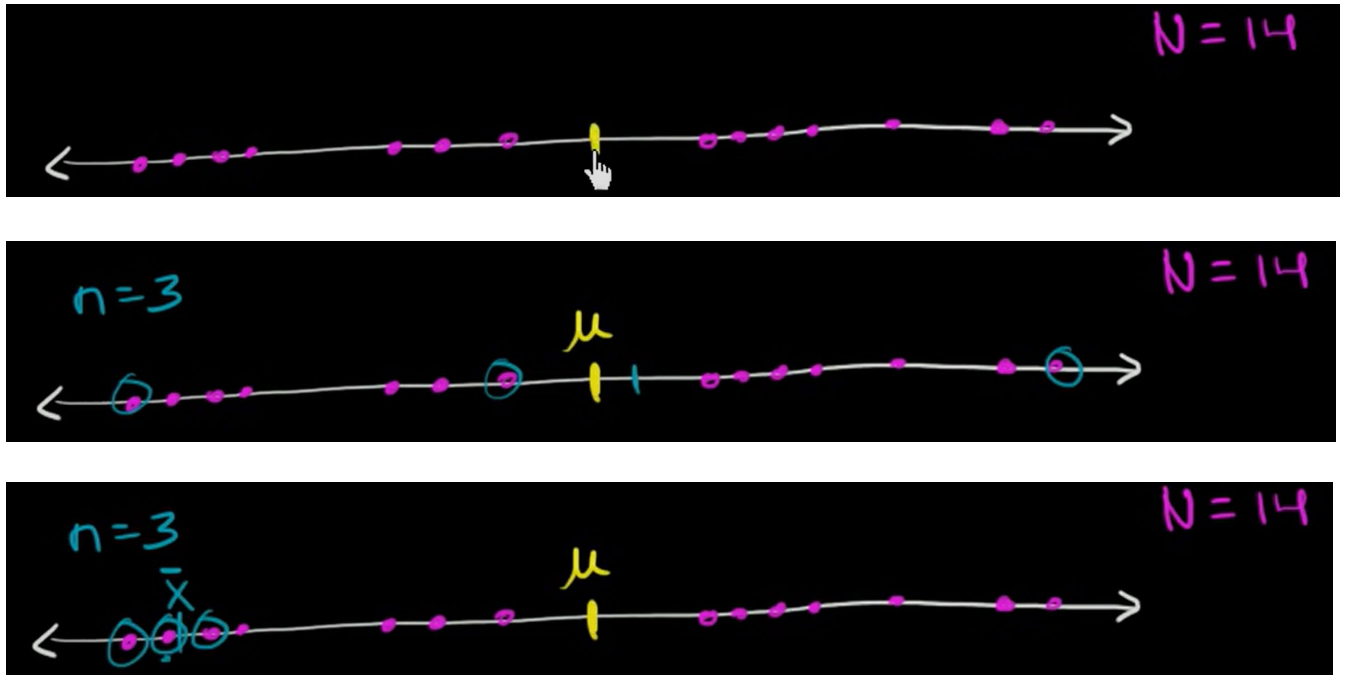


negatively skewed

Measuring the Dispersion of Data

- **Variance and standard deviation (sample: s, population: σ)**

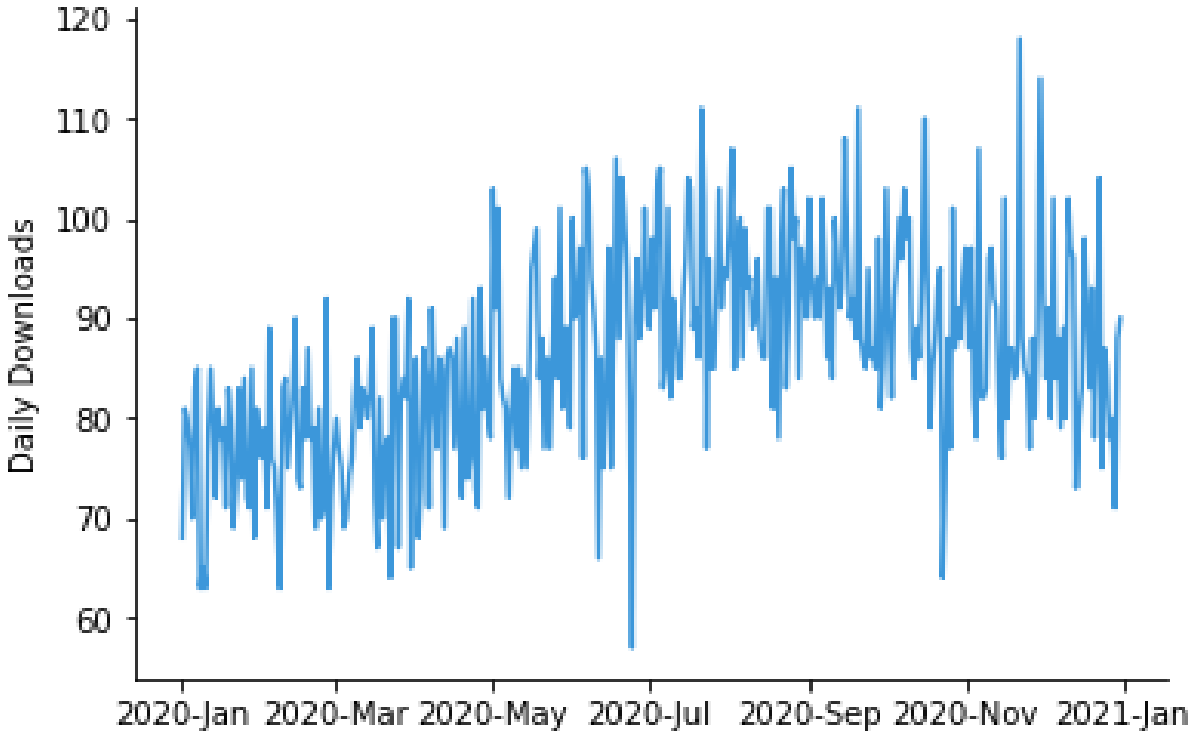
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\,2} - \mu^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$ **Unbiased Estimation**

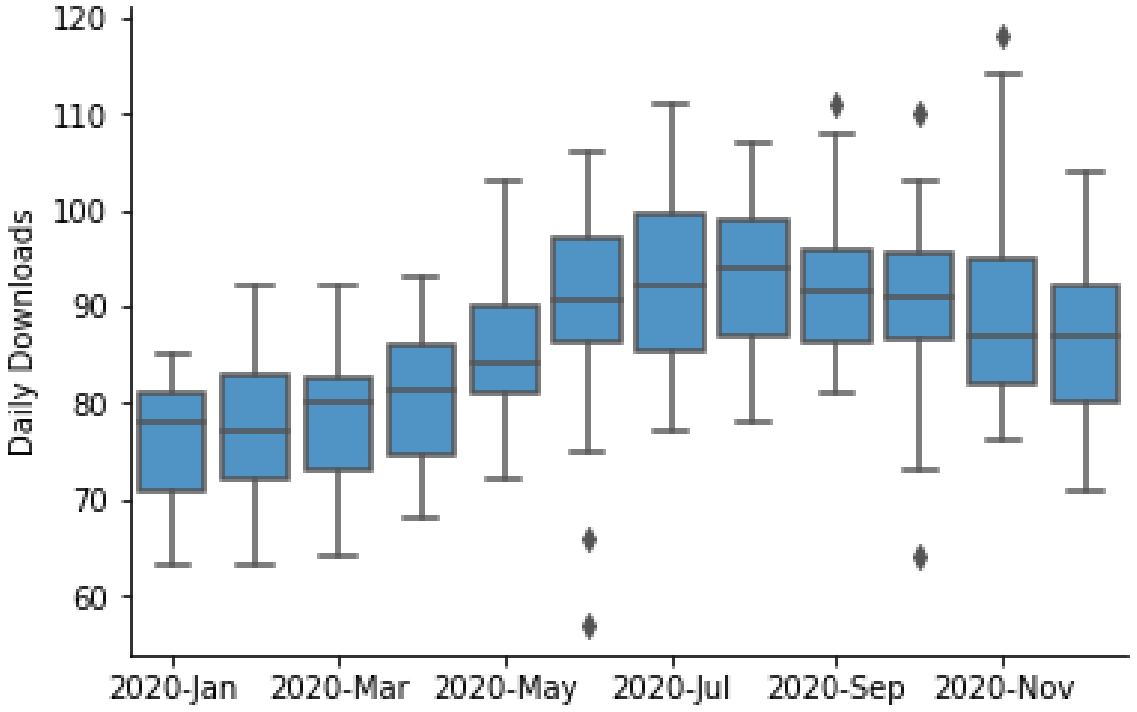


- **Standard deviation**: s (or σ) is the square root of variance s^2 (or σ^2)

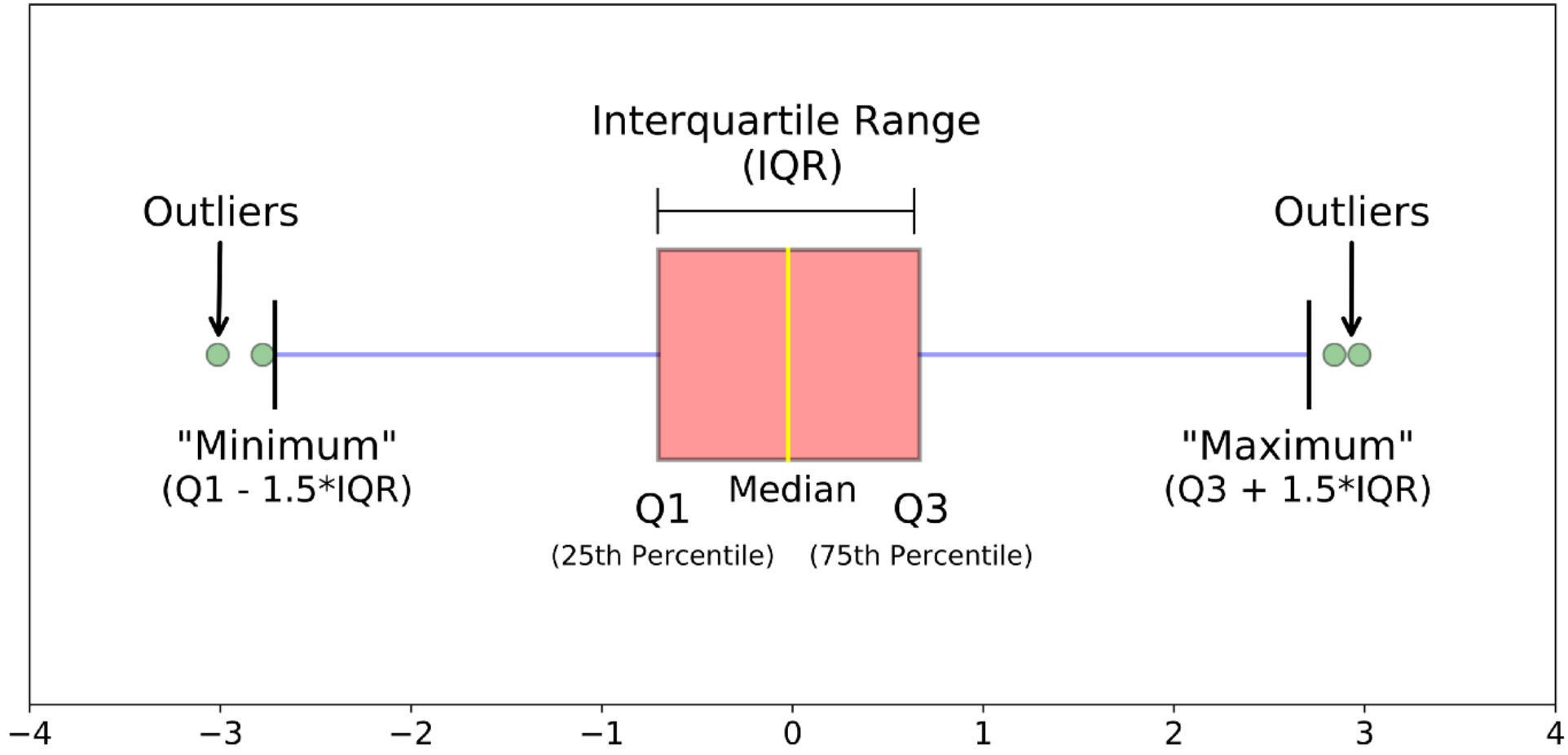
Reference: https://www.youtube.com/watch?v=KkaU2ur3Ymw

BoxPlot

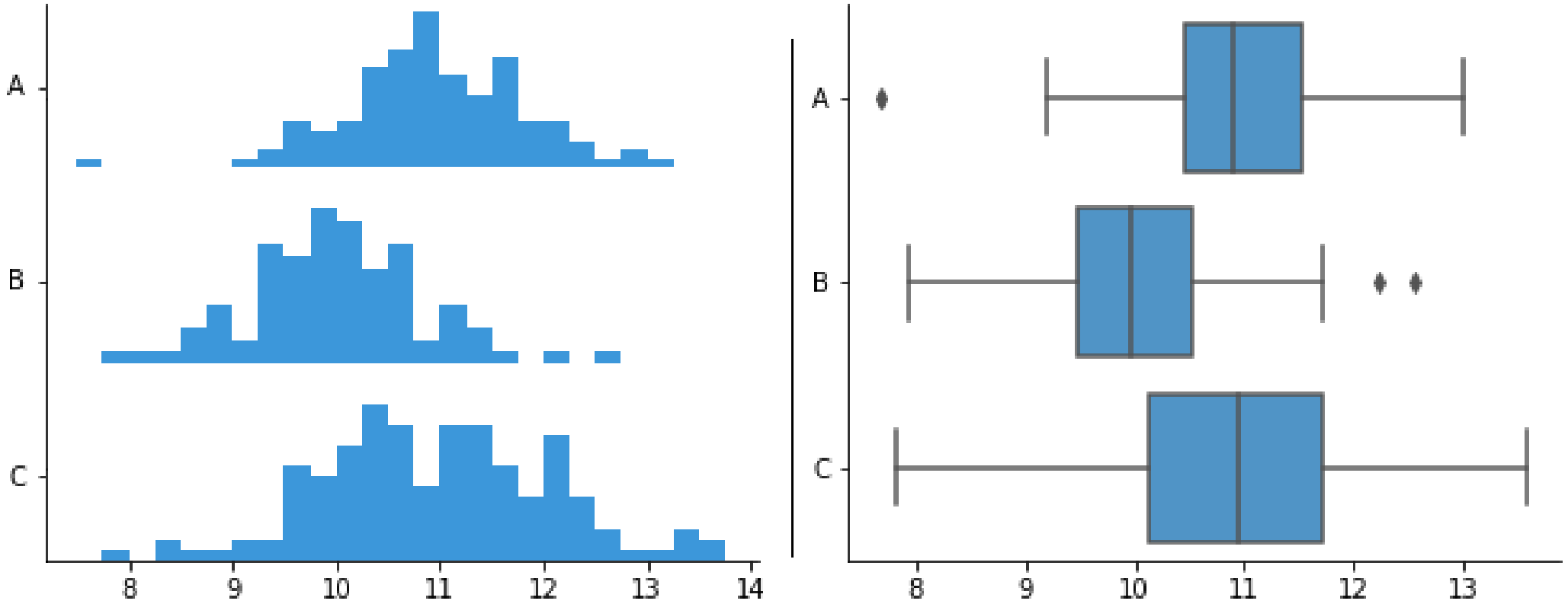




BoxPlot

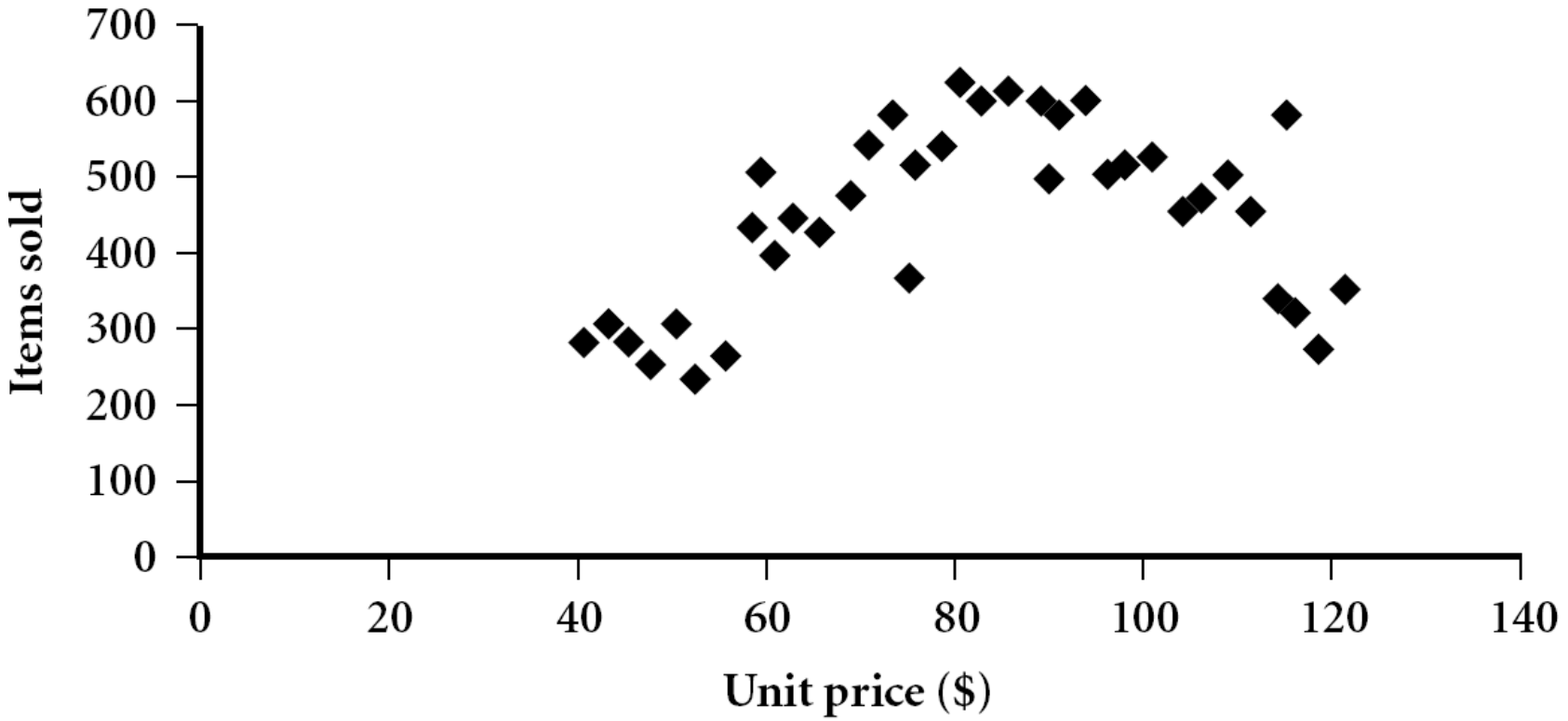


Histogram

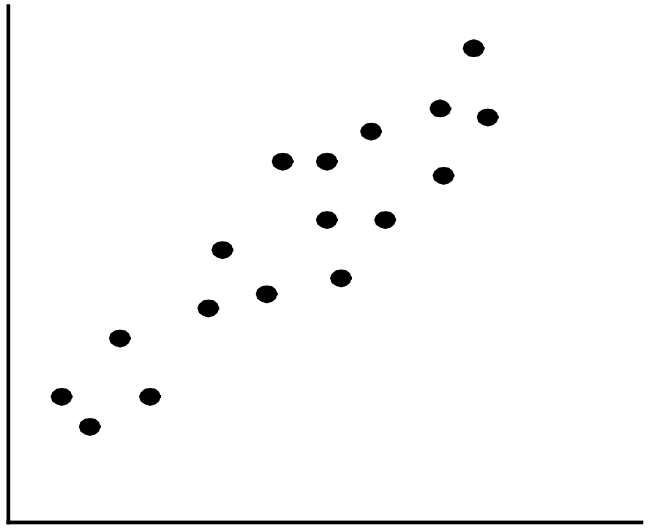


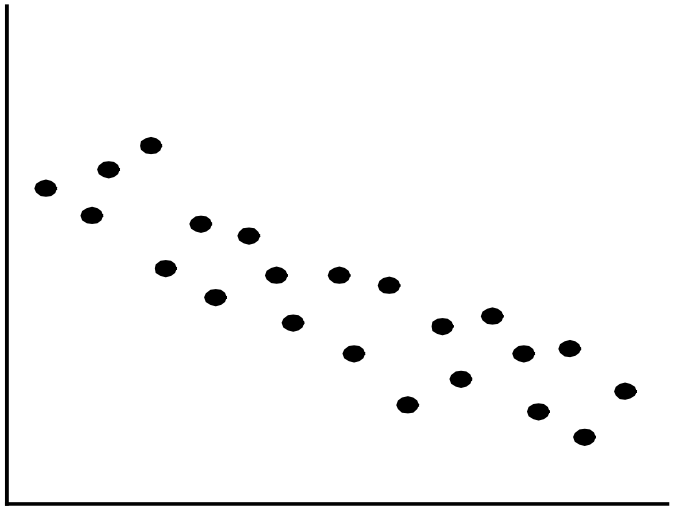
Scatter plot

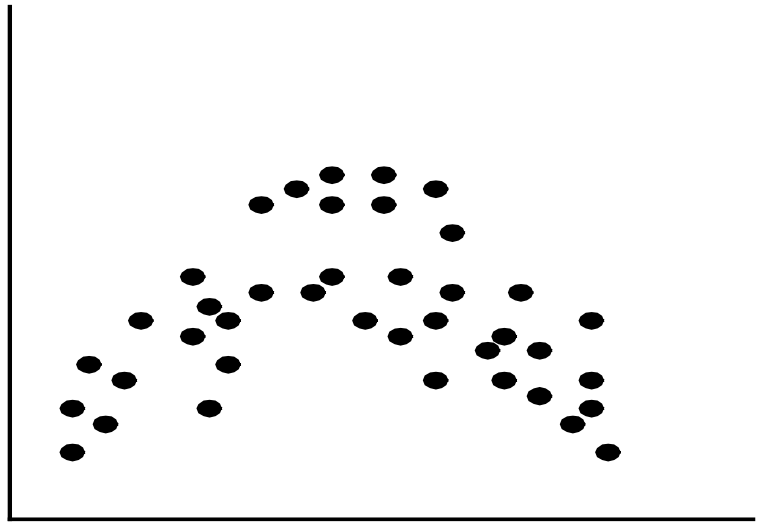
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

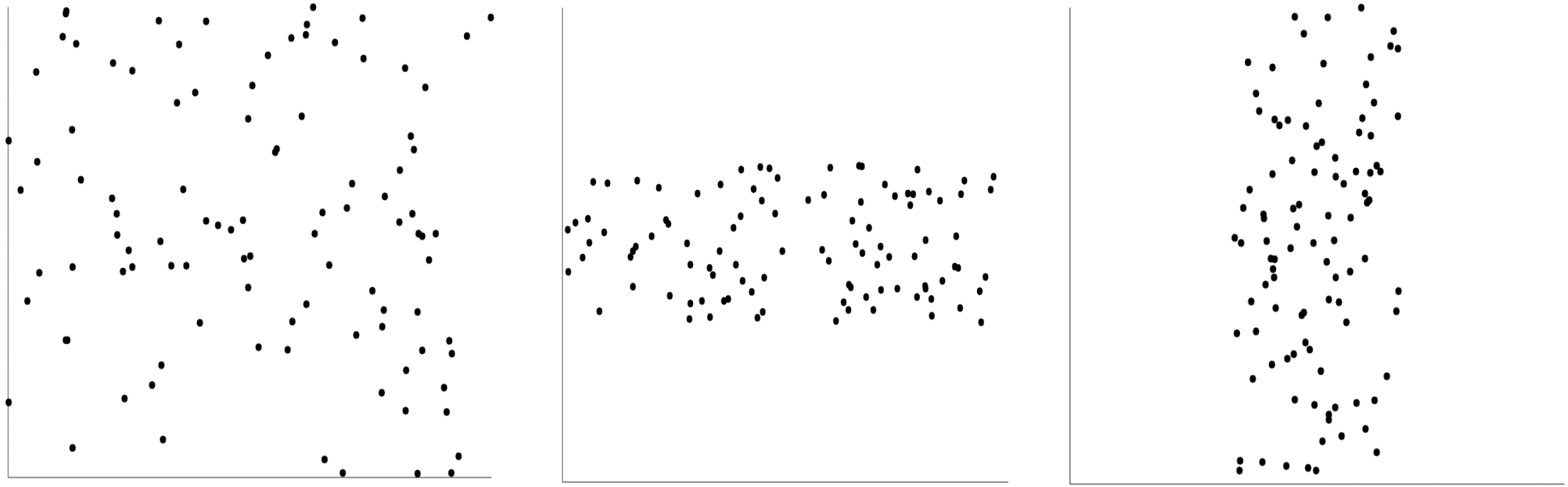






- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data



Pearson correlation coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

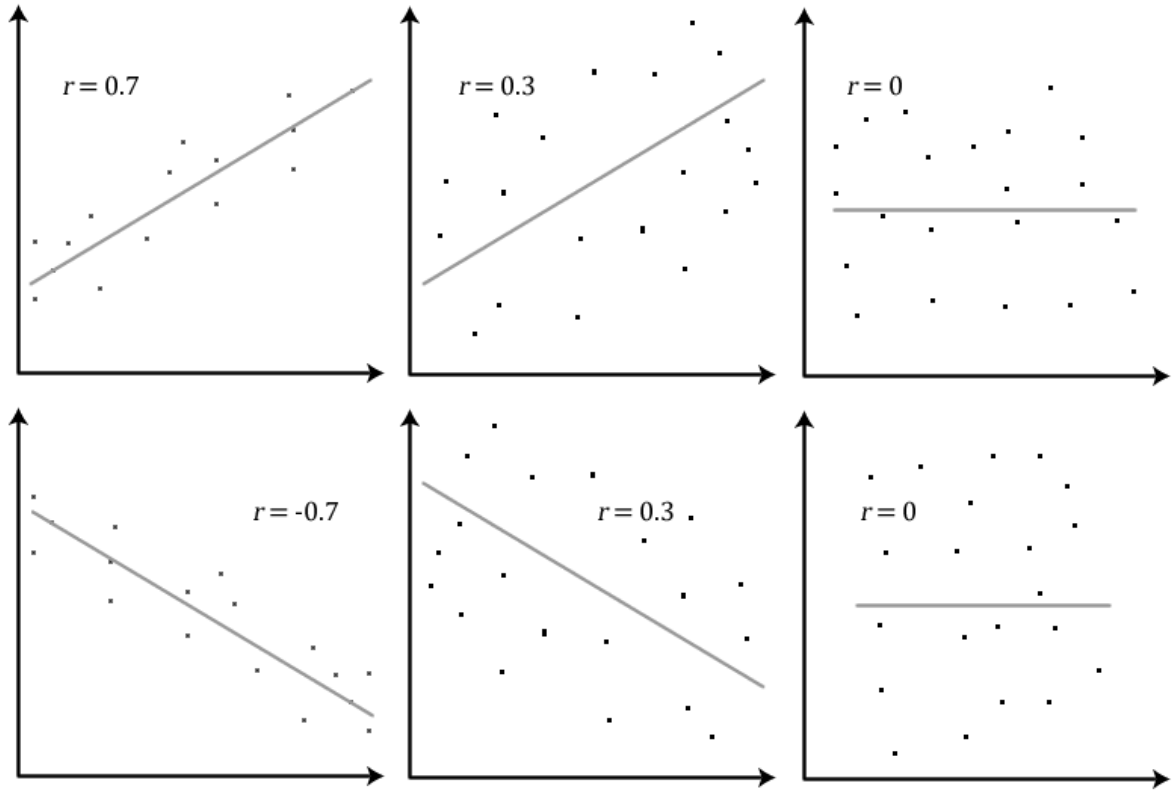
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

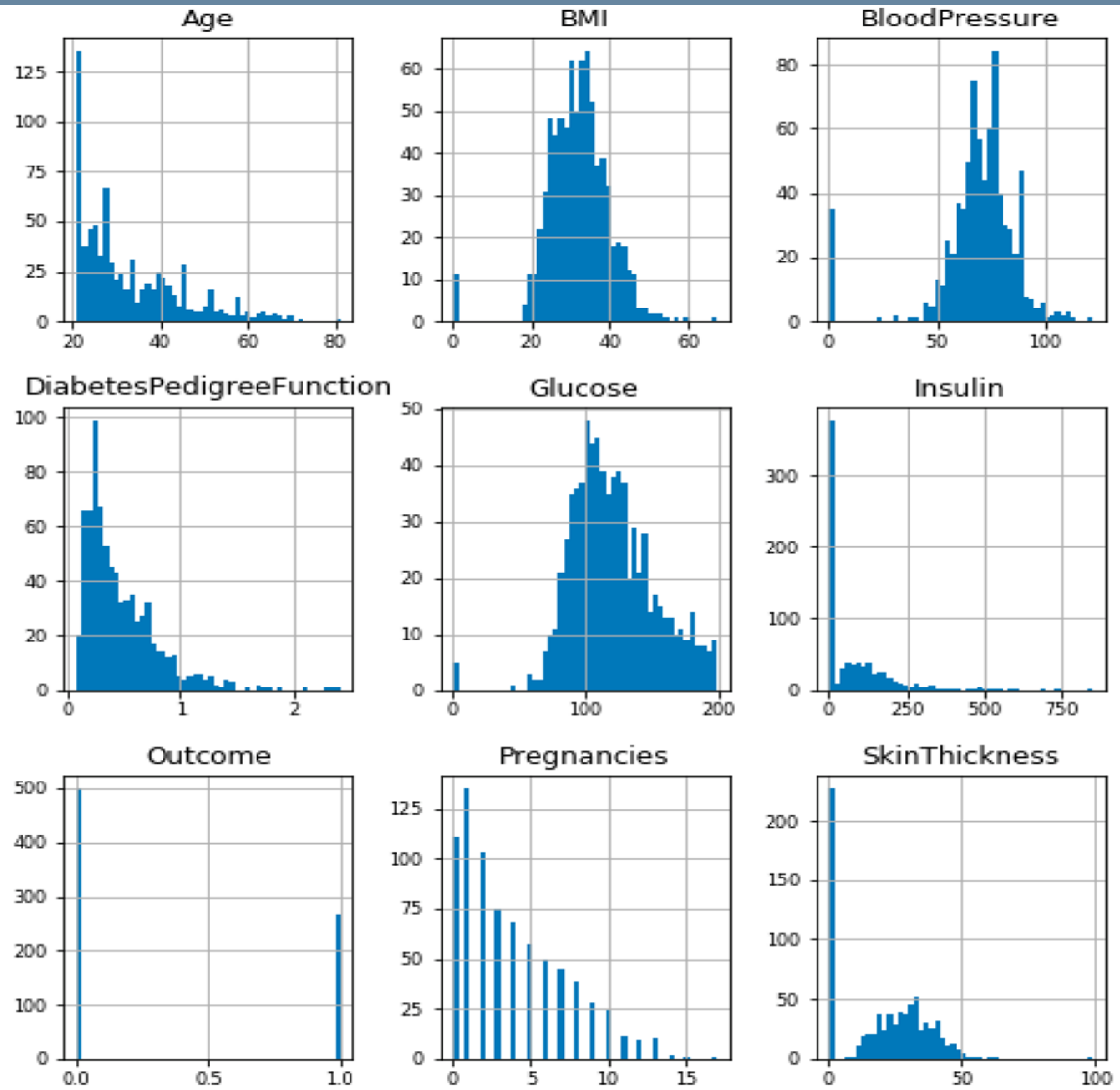
$\bar{y}$ = mean of the values of the y-variable



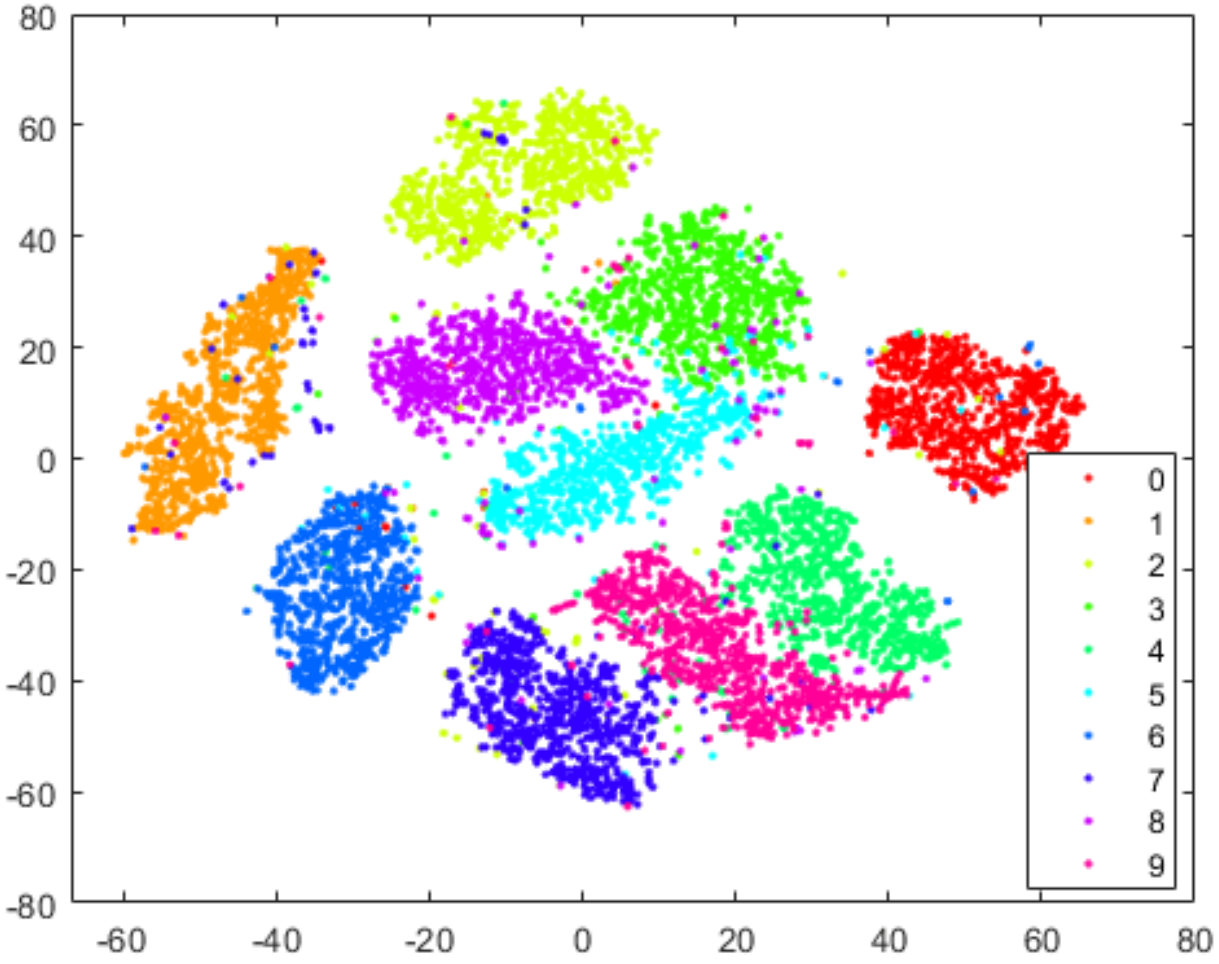
Agenda

- Data Representation
- Data Statistics
- Data Visualization
- Data Similarity
- Summary

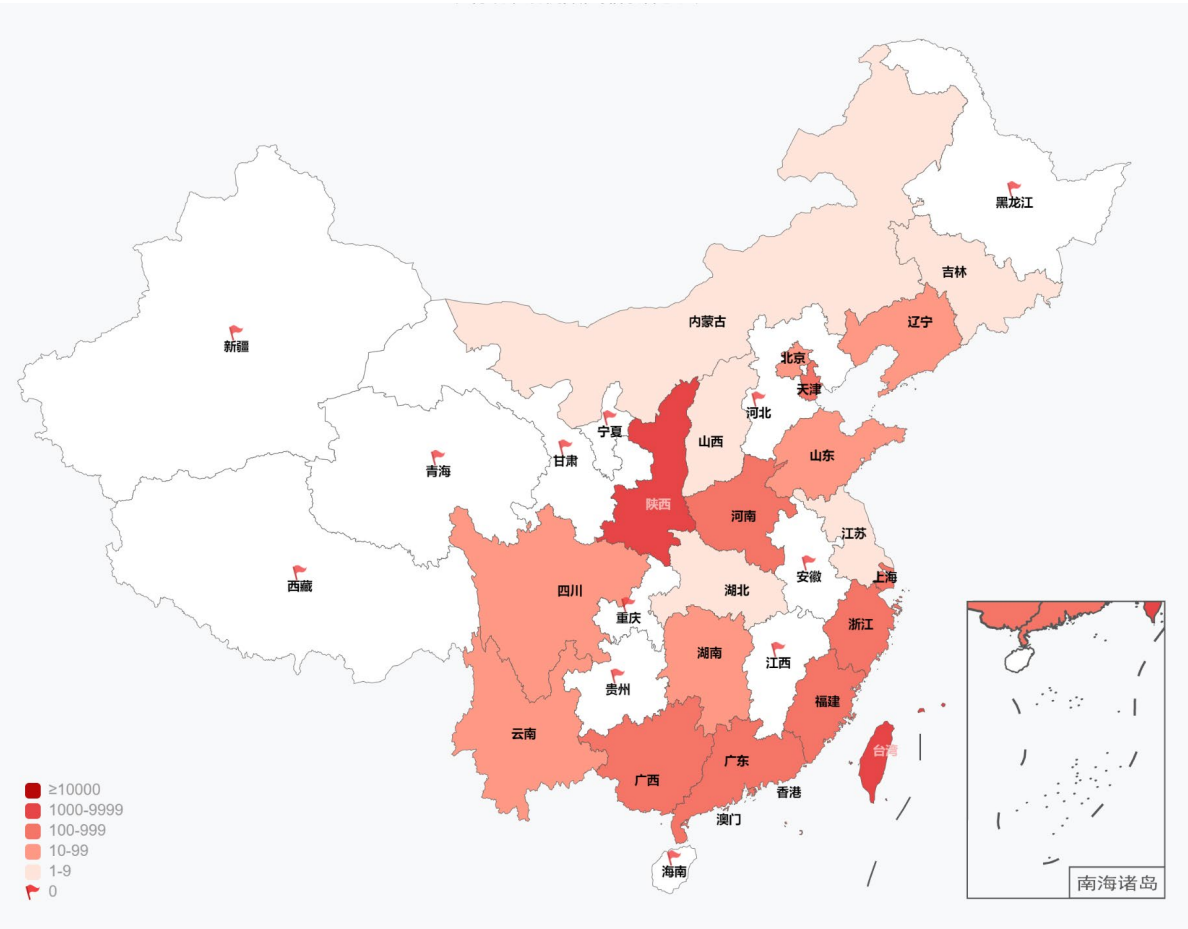
Feature Data Visualization in Kaggle

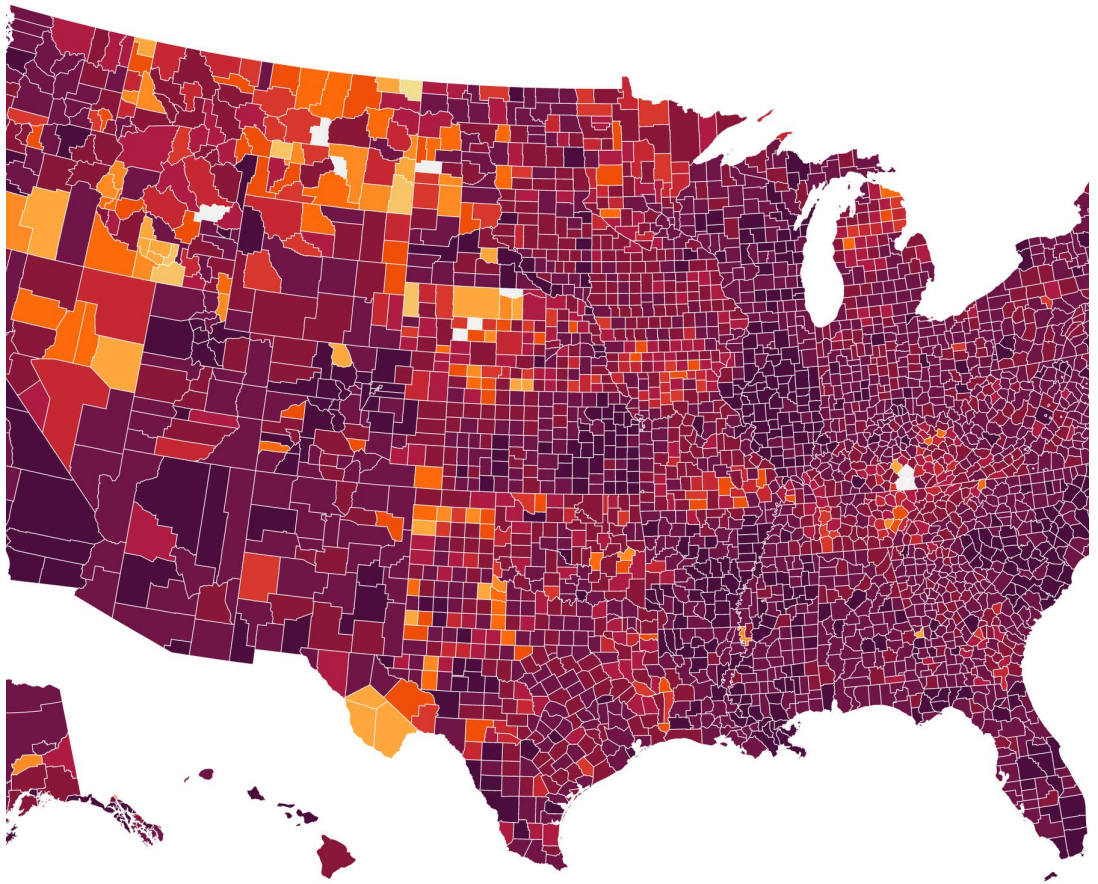


High-Dimensional Data Visualization with t-SNE



Map Mesh





Hybrid Visualization

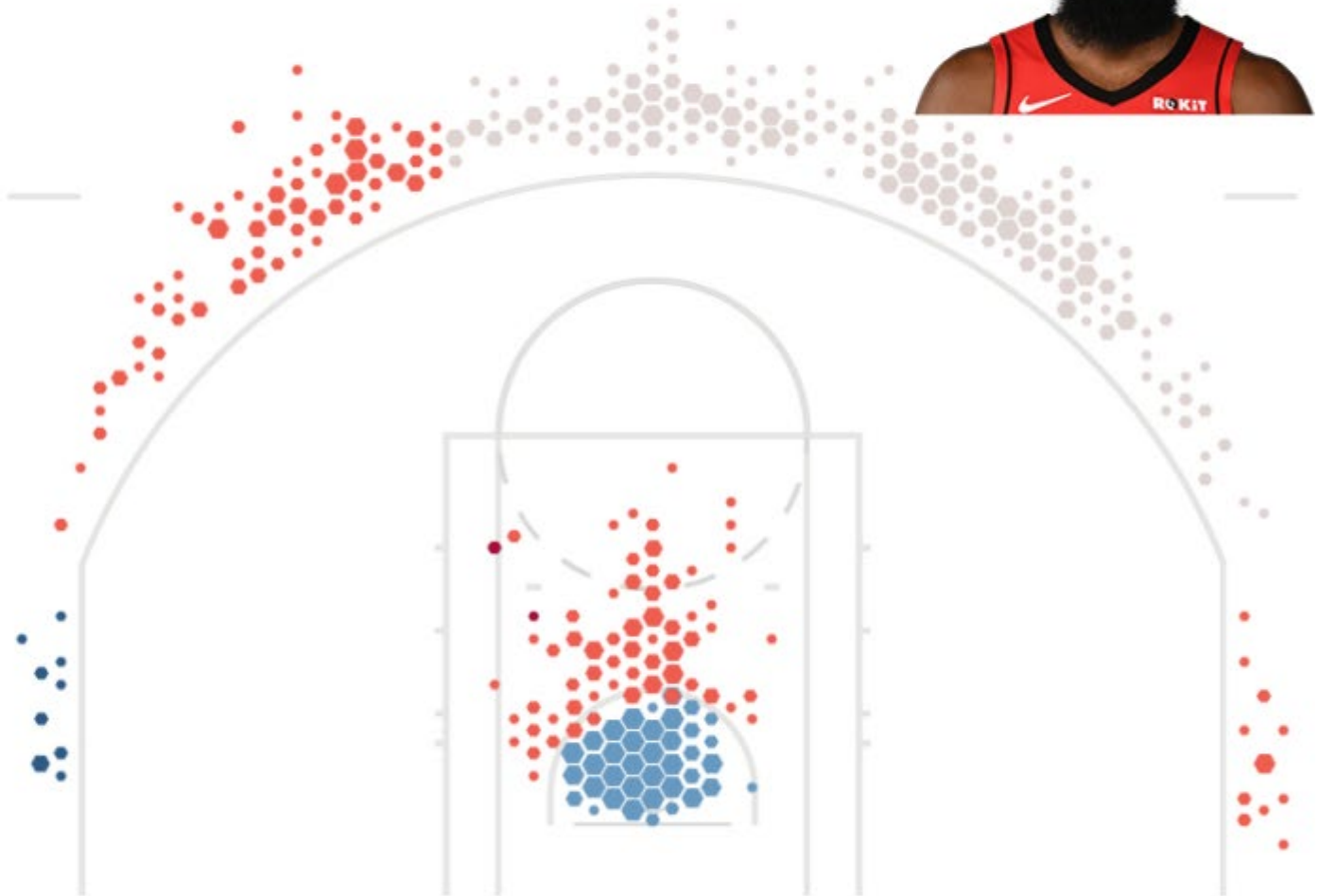
NBA Shot Charts - Regular Season 2018-19

Top 25 Players (Shot Attempts)

Click on a player to see his shot chart

James Harden	1,909
Kemba Walker	1,684
Paul George	1,614
Bradley Beal	1,609
Damian Lillard	1,533
Donovan Mitchell	1,530
D'Angelo Russell	1,517
Russell Westbrook	1,473
Klay Thompson	1,402
Kevin Durant	1,383
Buddy Hield	1,360
Nikola Vucevic	1,354
Blake Griffin	1,341
Stephen Curry	1,340
LaMarcus Aldridge	1,319
Karl-Anthony Towns	1,314
DeMar DeRozan	1,313
Trae Young	1,256
Devin Booker	1,255
Tobias Harris	1,254
Giannis Antetokounmpo	1,247
CJ McCollum	1,243
Kyrie Irving	1,241
Andrew Wiggins	1,209
Nikola Jokic	1,206

James Harden



DATAVIZARDRY

Created by Daniel Teo
Inspired by Kirk Goldsberry (https://www.instagram.com/kirkgoldsberry/)

BELOW LEAGUE AVG

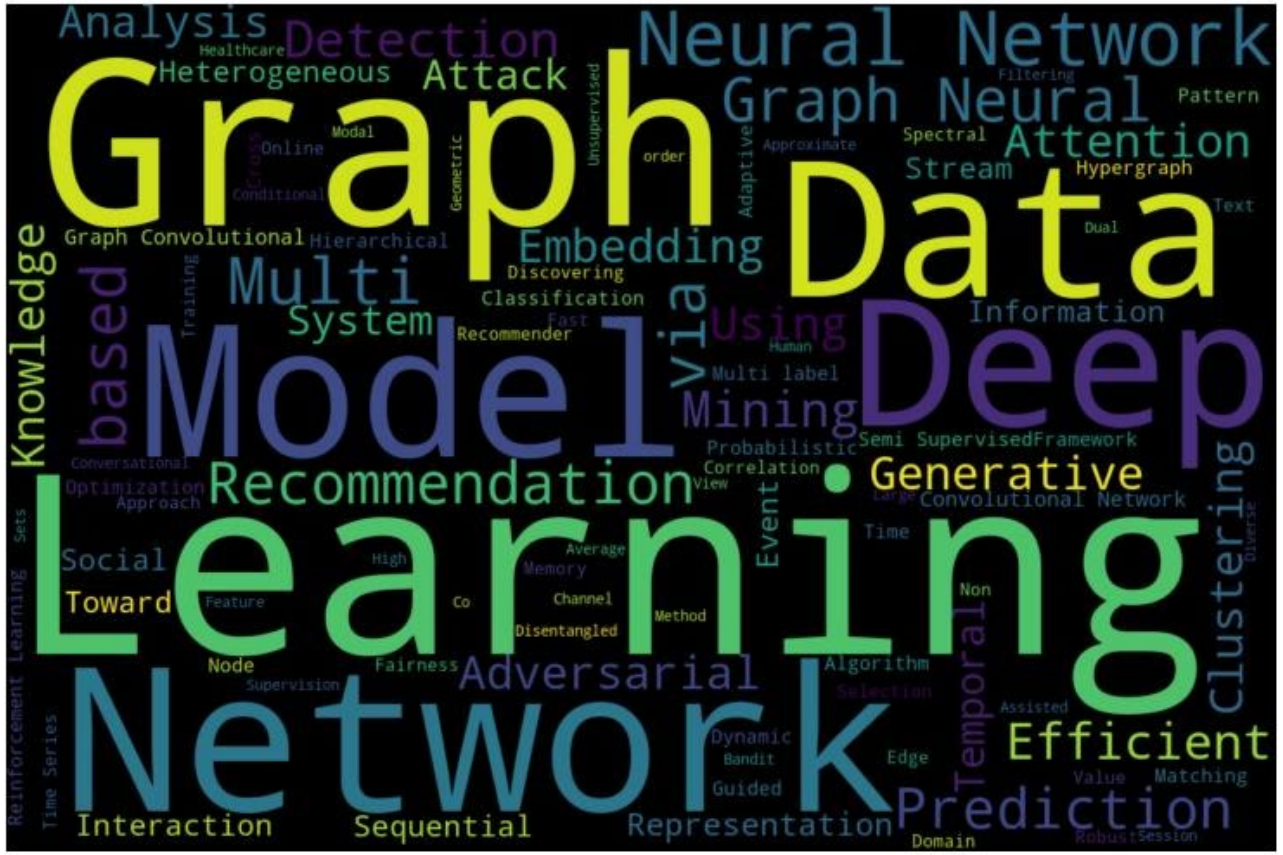
Efficiency By Location

ABOVE LEAGUE AVG

浙江大學 Zhejiang University

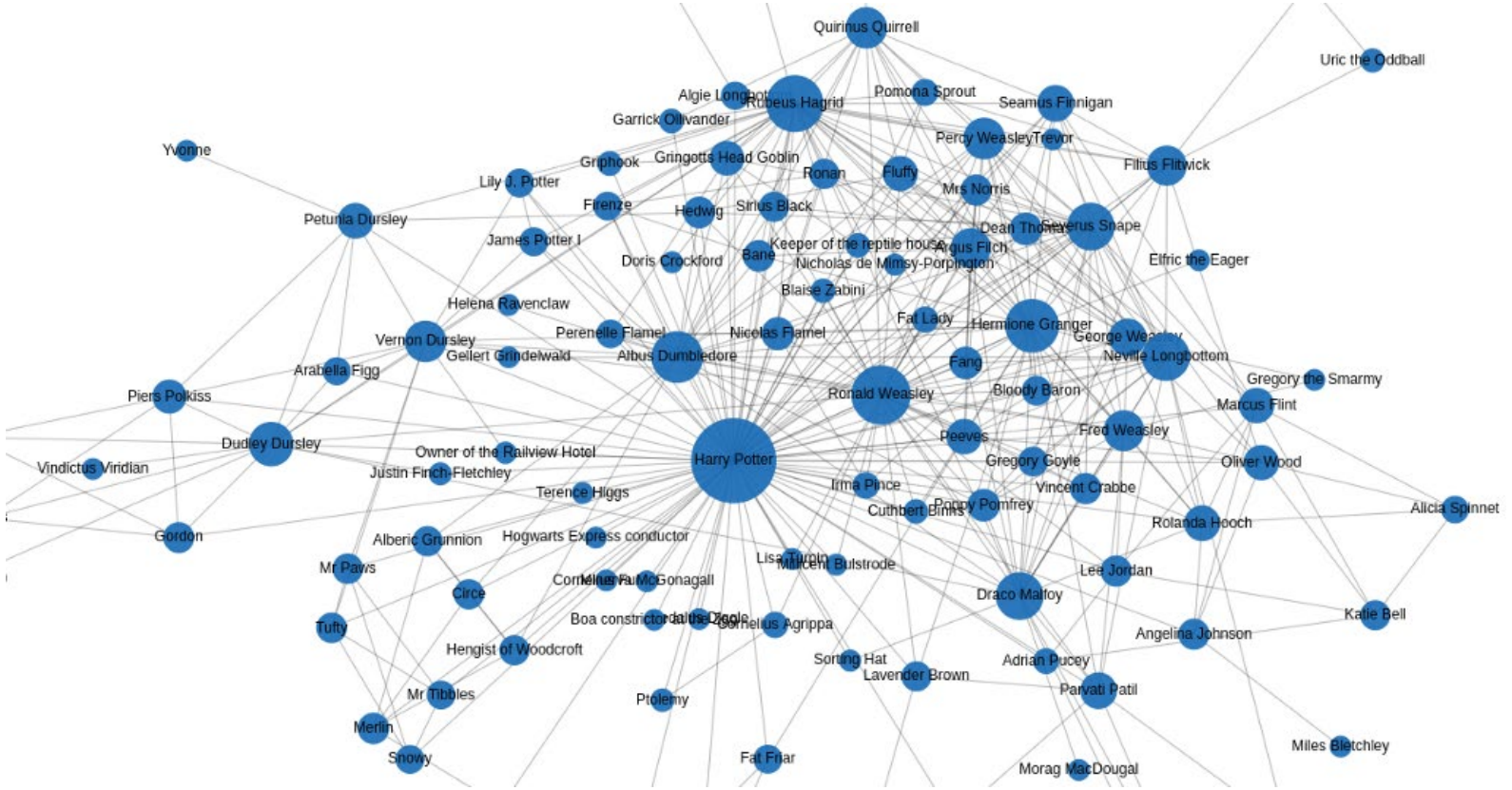
Tag Cloud

- KDD 2020 Highlight



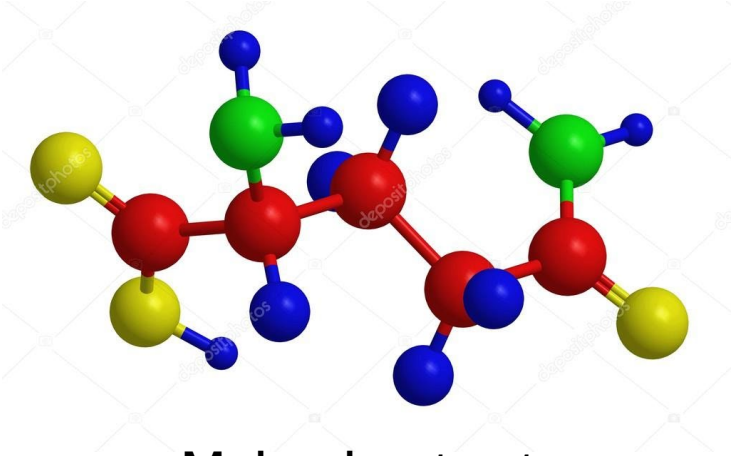
Graph Data

Social Network

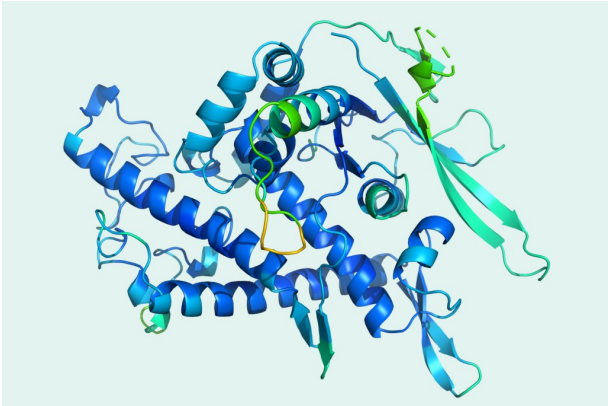


Knowledge Graph

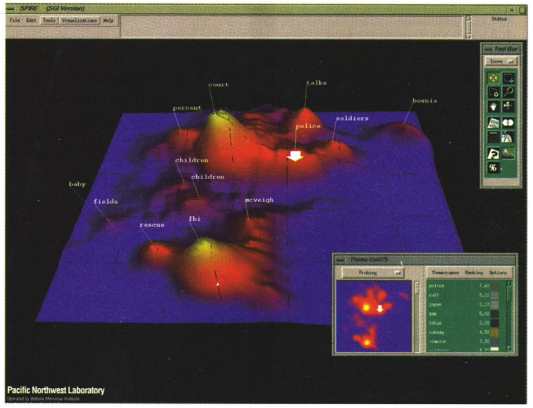
Scientific Data



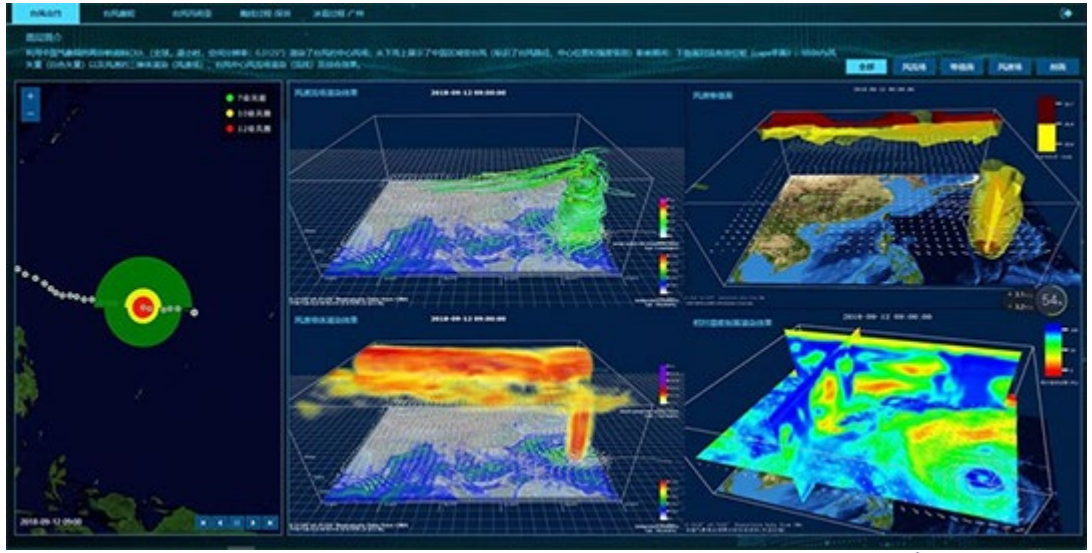
Molecular structure



Protein structure



Landscape



Typhoon

Agenda

- Data Representation
- Data Statistics
- Data Visualization
- Data Similarity
- Summary

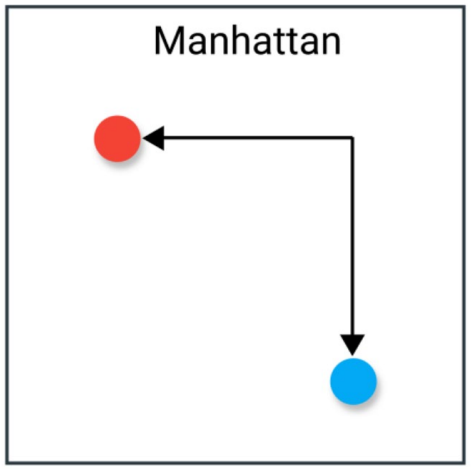
Similarity and Dissimilarity

- **Similarity**
  - ✓ Numerical measure of how alike two data objects are
  - ✓ Value is higher when objects are more alike
  - ✓ Often falls in the range [0,1]
- **Dissimilarity (e.g., distance)**
  - ✓ Numerical measure of how different two data objects are
  - ✓ Lower when objects are more alike
  - ✓ Minimum dissimilarity is often 0
  - ✓ Upper limit varies
- **Proximity refers to a similarity or dissimilarity**

Numerical Vectors

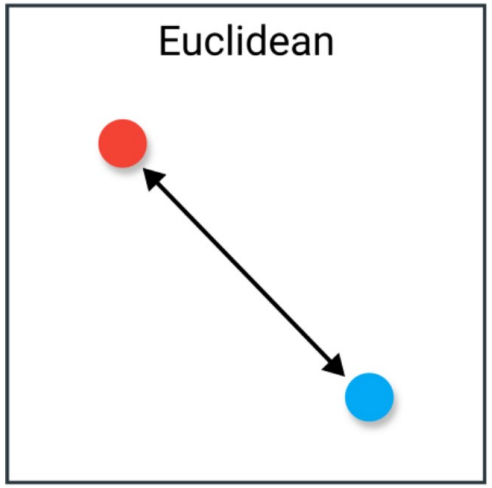
Manhattan Distance (L1-distance)

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$



Euclidean Distance (L2-distance)

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$



Minkowski Distance

- **Minkowski distance (L-h Norm): A popular distance measure**

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$

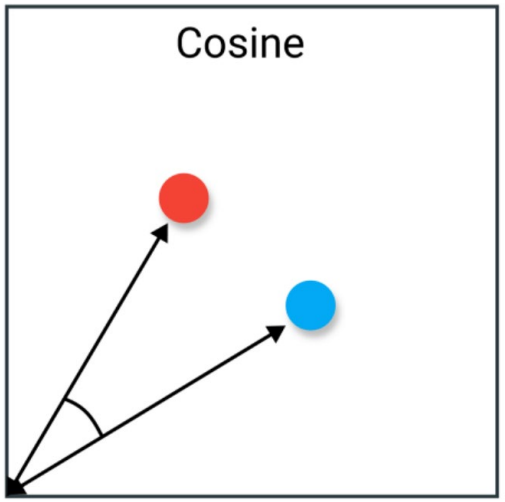
- **Properties are two p-dimensional data objects**
  - ✓ $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
  - ✓ $d(i, j) = d(j, i)$ (Symmetry)
  - ✓ $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- **A distance that satisfies these properties is a metric**

Example: Cosine Similarity

- $\cos(d1, d2) = (d1 \bullet d2) / ||d1|| \, ||d2||$
- **Example:** Find the similarity between documents 1 and 2.

$d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$

$d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$



- **Solution:**

$d1 \bullet d2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$

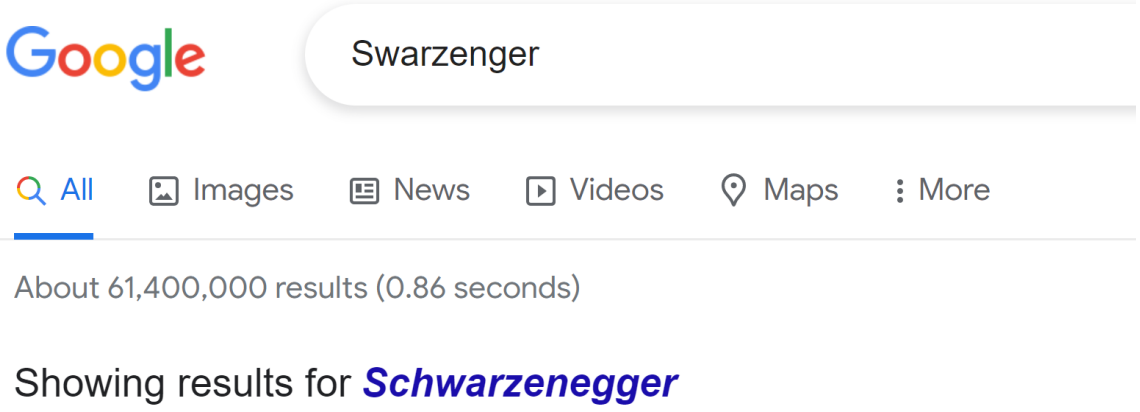
$||d1|| = (5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0)^{0.5} = 6.481$

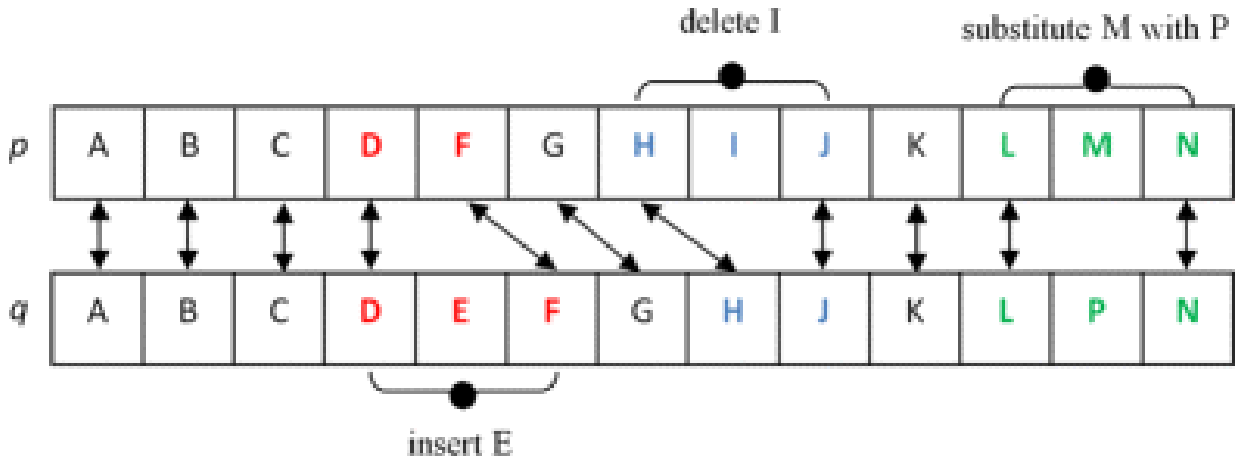
$||d2|| = (3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1)0.5 = 4.12$

$\cos(d1, d2) = 0.94$

String Similarity

Edit Distance





Set Data

Jaccard Similarity

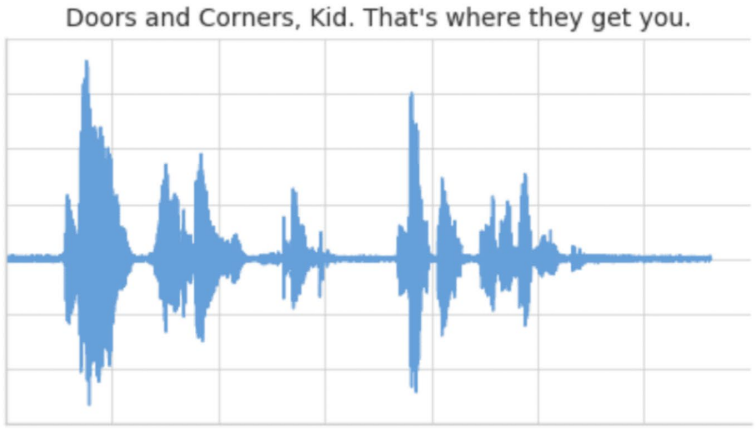
- ✓ **J(A, B) =** $|A \cap B| / |A \cup B|$

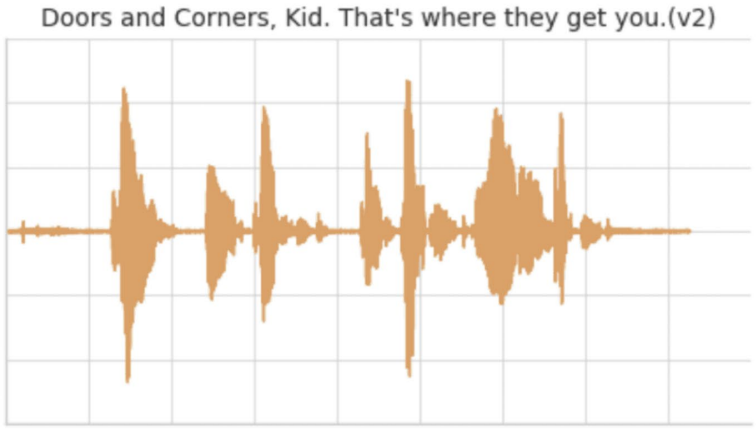
```
E = ['cat', 'dog', 'hippo', 'monkey']
F = ['monkey', 'rhino', 'ostrich', 'salmon']
```

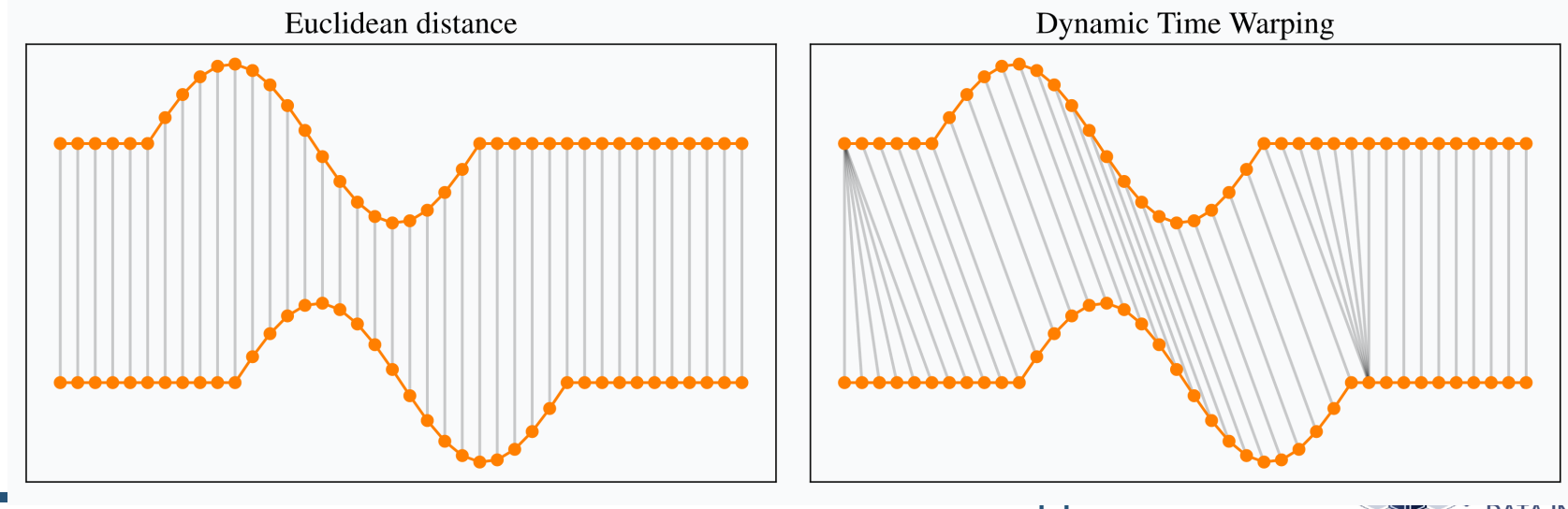
- **Number of observations in both:** {‘monkey’} = 1
- **Number of observations in either:** {‘cat’, ‘dog’, hippo’, ‘monkey’, ‘rhino’, ‘ostrich’, ‘salmon’} = 7
- **Jaccard Similarity:** 1 / 7= 0.142857

Time Series

DTW (Dynamic Time Wrapping)







Agenda

- Data Representation
- Data Statistics
- Data Visualization
- Data Similarity
- Summary

Summary

- **Data Representation**
  - ✓ Structured data, Semi-structured data, Unstructured data
- **Data Statistics**
  - ✓ Mean, mode, median, variance, quantile, boxplot, histogram
- **Data Visualization**
  - ✓ Histogram, map mesh, tag cloud, graph, scientific data
- **Data Similarity**
  - ✓ Numeric vector, set, string, time series