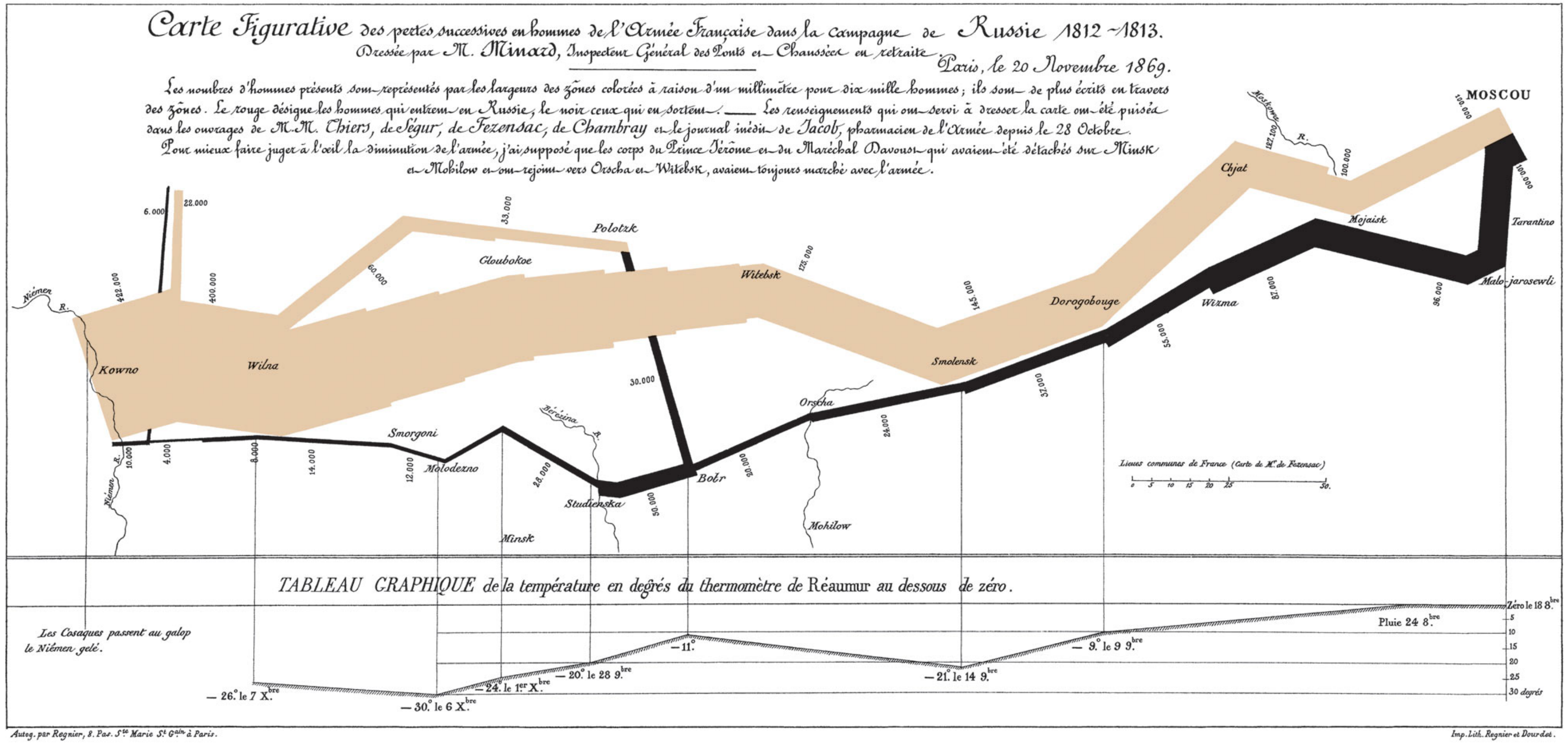
CS3319 Foundations of Data Science

2.Data Fundamentals

Jiaxin Ding
John Hopcroft Center

上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

Understanding Data



Charles Minard's map of Napoleon's Russian campaign of 1812

Content

- Data Attributes
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Probability Inequalities

Content

- Data Attributes
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Probability Inequalities

Data Attributes

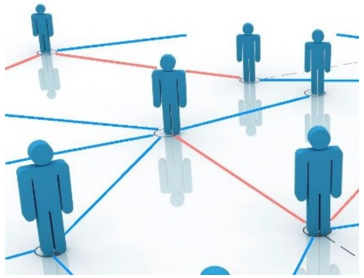
- **Data object**: an entity in the dataset
- A **data attribute** is a particular data field, representing a characteristic or feature of a data object (Feature)

学号	姓名	入学年份
1001	张三	2018
1003	李四	2019
1099	王二	2020

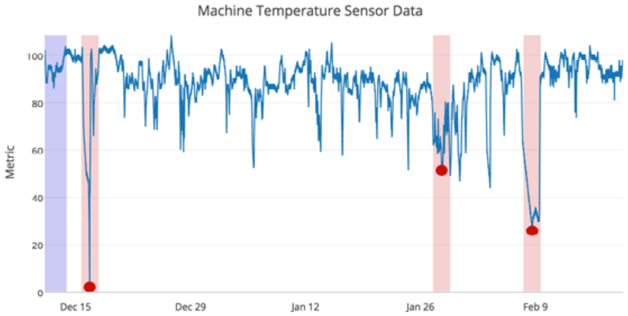
Name in the database

RGB value of a pixel

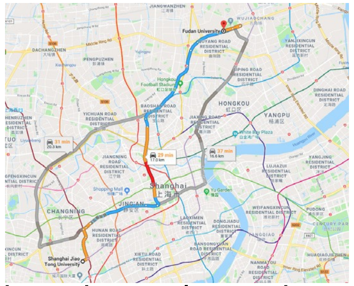
The frequency of a word



The friends of a user



The reading at time t



The time-location of a trajectory point

Record Data

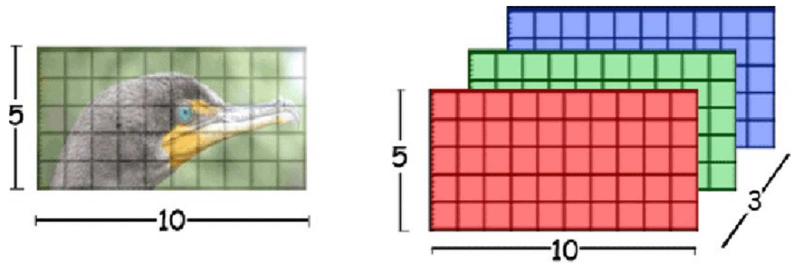
- Relational databases
  - Each row represents a data object
  - Each column represents a data attribute

WEEKDAY	GENDER	AGE	CITY
Tuesday	Male	28	London
Monday	Female	24	New York
Tuesday	Female	36	Hong Kong
Thursday	Male	17	Tokyo

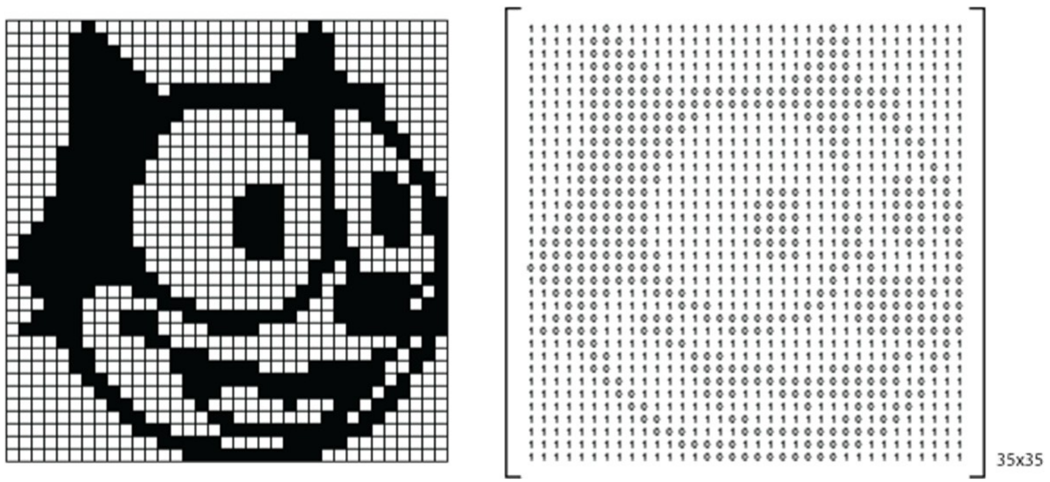
JSON Format:
{
WEEKDAY: Monday;
GENDER: Female;
AGE: 24;
CITY: New York;
}

Image Data

- A 3-layer matrix (3*height*width) of [0,255] real value



Original Color Image   Matlab RGB Matrix



Text Data

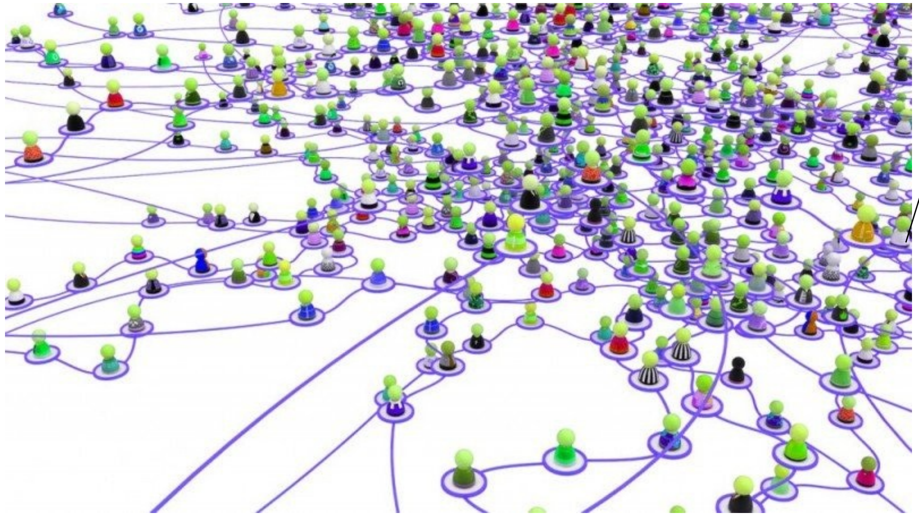
- A sequence of words/tokens that represents semantic meanings.

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text.

Bag-of-Words Format:
{
text: 4;
mining: 2;
also: 1;
referred: 1;
to: 2;
as: 1;
data: 1;
roughly: 1;
equivalent: 1;
analytics: 1;
is: 1;
the: 1;
process: 1;
of: 1;
deriving: 1;
high-quality: 1;
information: 1;
from: 1;
}

Graph Data

- A directed/undirected graph
  - Possibly with additional information for nodes and edges



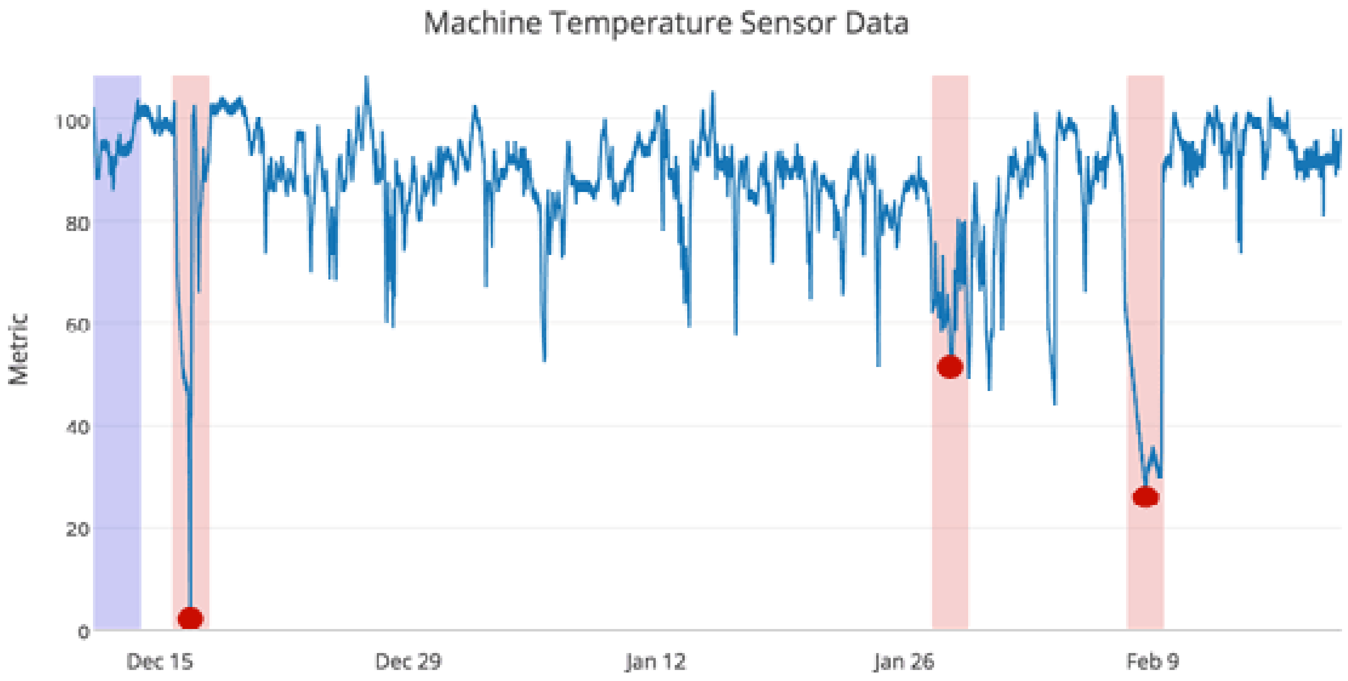
Friendship Format:

Alice	Bob
Bob	Carl
Carl	Victor
Bob	Victor
Alice	Victor
…	

Stanford network dataset collection: https://snap.stanford.edu/data/

Streaming Data

- A sequence of readings



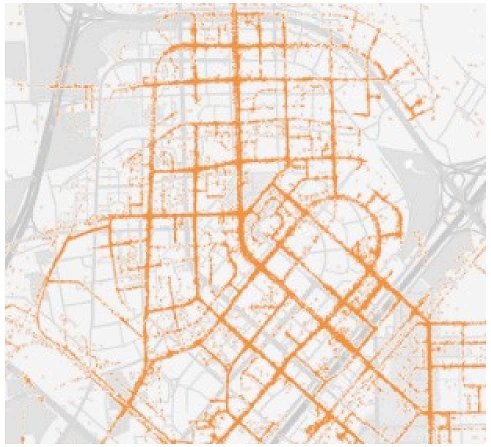
. . . 1, 5, 2, 7, 0, 9, 3

. . . a, r, v, t, y, h, b

. . . 0, 0, 1, 0, 1, 1, 0

Spatio-Temporal Data

- A sequence of (time, location, info) tuples



Content

- Data Attributes
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Probability Inequalities

Basic Statistical Descriptions of Data

- **How to capture the properties of a given data set?**
  - **Central tendency:** describes the center around the data is distributed
  - **Dispersion:** describes the data spread

Measuring the Central Tendency

- **Mean** (algebraic measure)

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

  - Weighted arithmetic mean:

$$\mu = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Geometric mean： $\mu = \sqrt[n]{\Pi x_i}$

    - The geometric mean is always <= arithmetic mean, and more sensitive to values near zero.
    - Geometric means make sense with ratios: 1/2 and 2/1 *should* average to 1.

Measuring the Central Tendency

- **Median**
  - Middle value if odd number of values, or average of the middle two values otherwise.
- Example:
  - Five data points **{1.2, 1.4, 1.5, 1.8, 10.2}**
  - Mean: 3.22 Median: **1.5**

- Mean is meaningful for symmetric distributions without outliers: e.g. height and weight.
- Median is better for skewed distributions or data with outliers: e.g. wealth and income.

Measuring the Central Tendency

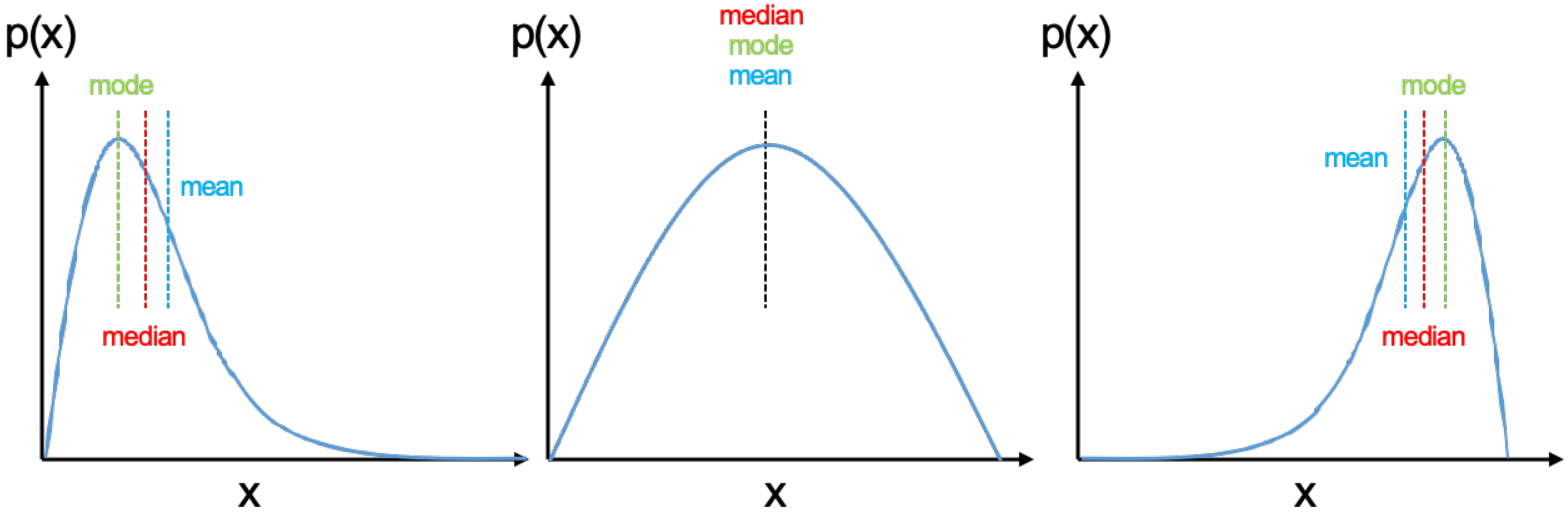
- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula (moderately skewed distribution):

$$\text{mean} - \text{mode} \simeq 3 \times (\text{mean} - \text{median})$$

- Example:
  - Five data points **{1, 1, 1, 1, 1, 2, 2, 2, 3, 3}**
  - Mean: 1.7 Median: 1.5 Mode: **1**

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



Positively skewed data

mode < median

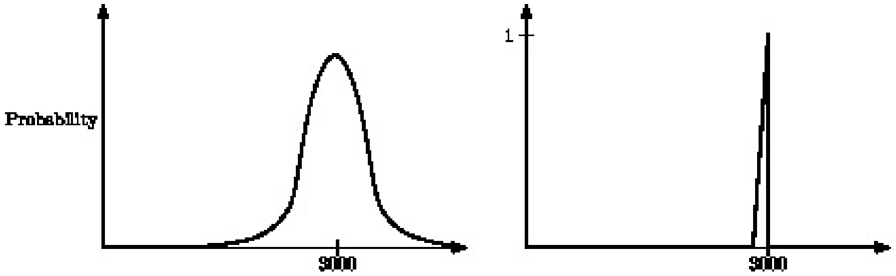
Symmetric data

mode = median

Negatively skewed data

mode > median

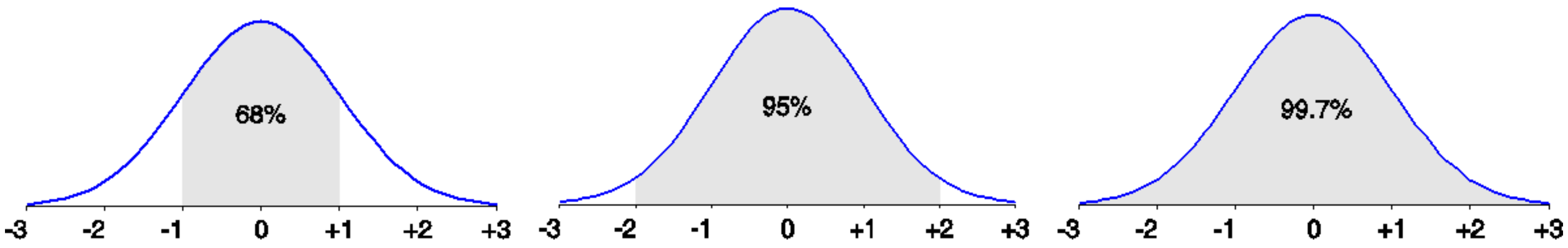
Measuring the Dispersion of Data



- Variance and standard deviation
  - Variance

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathbb{E}[x] \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

  - Standard deviation σ is the square root of variance σ^2
- **The normal distribution curve**
  - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements
  - From $\mu-2\sigma$ to $\mu+2\sigma$: contains about 95% of it
  - From $\mu-3\sigma$ to $\mu+3\sigma$: contains about 99.7% of it

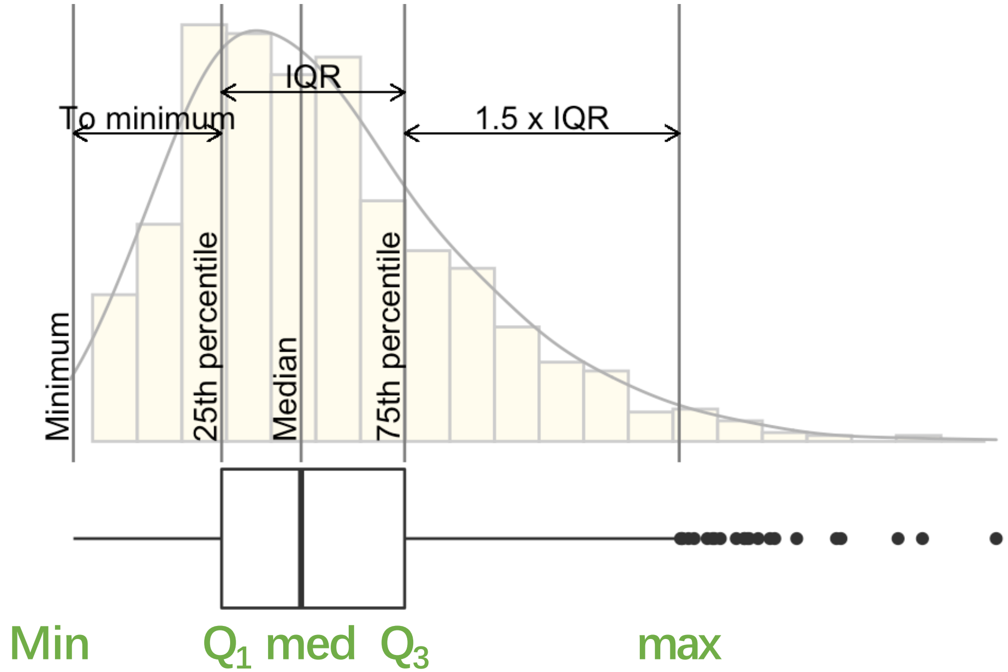


Measuring the Dispersion of Data

- Regardless of how data is distributed, at least $\left(1-\frac{1}{k^2}\right)$ of the points must lie within $k\sigma$ of the mean.
  - Thus at least 75% must lie within two sigma of the mean.
  - The normal distribution can achieve tighter bound.

Measuring the Dispersion of Data

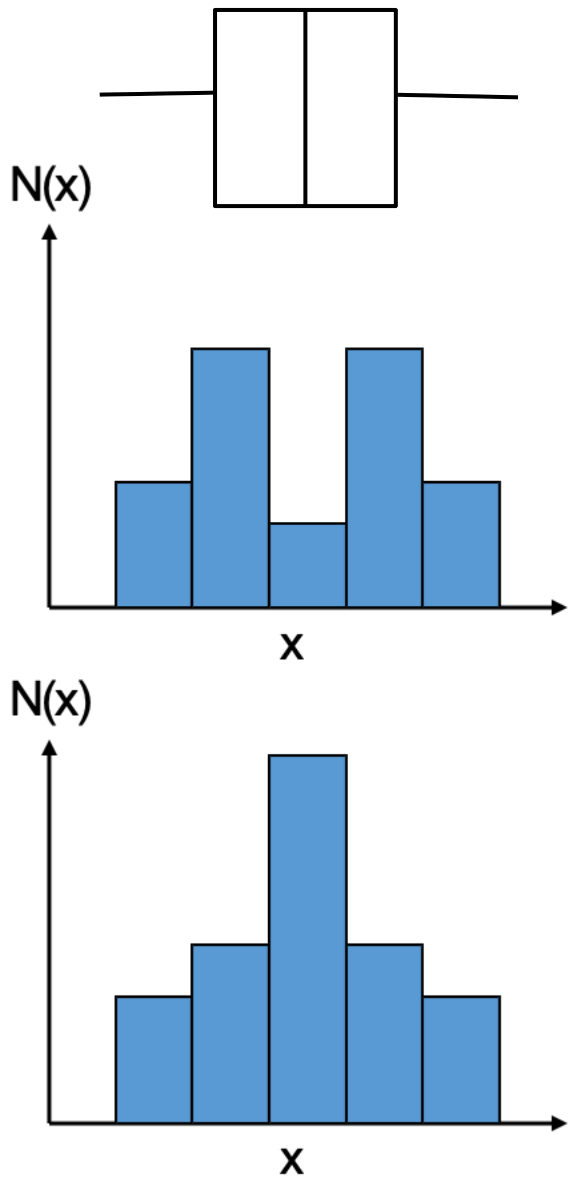
- Quartiles, outliers and boxplots
  - **Quartiles:** Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range:** IQR = $Q_3 - Q_1$
  - **Five number summary:** min, Q_1, median, Q_3, max
  - **Outlier:** usually, a value higher(lower) than 1.5 x IQR than Q_3 (Q_1)



Histograms

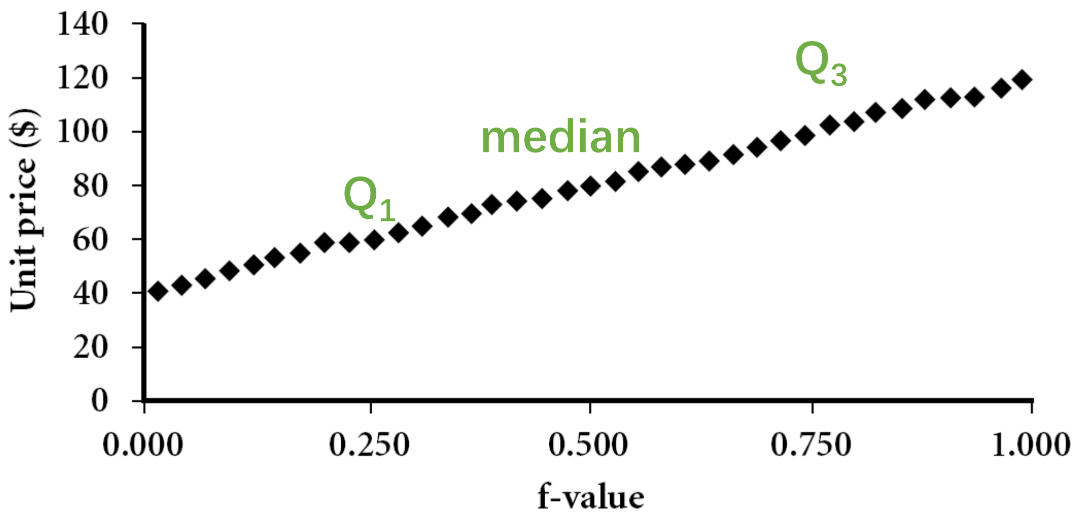
- **Histogram:** Graph display of tabulated frequencies, shown as bars. It shows what proportion of cases fall into each of several categories

- The two histograms shown may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
  - But they have rather different data distributions



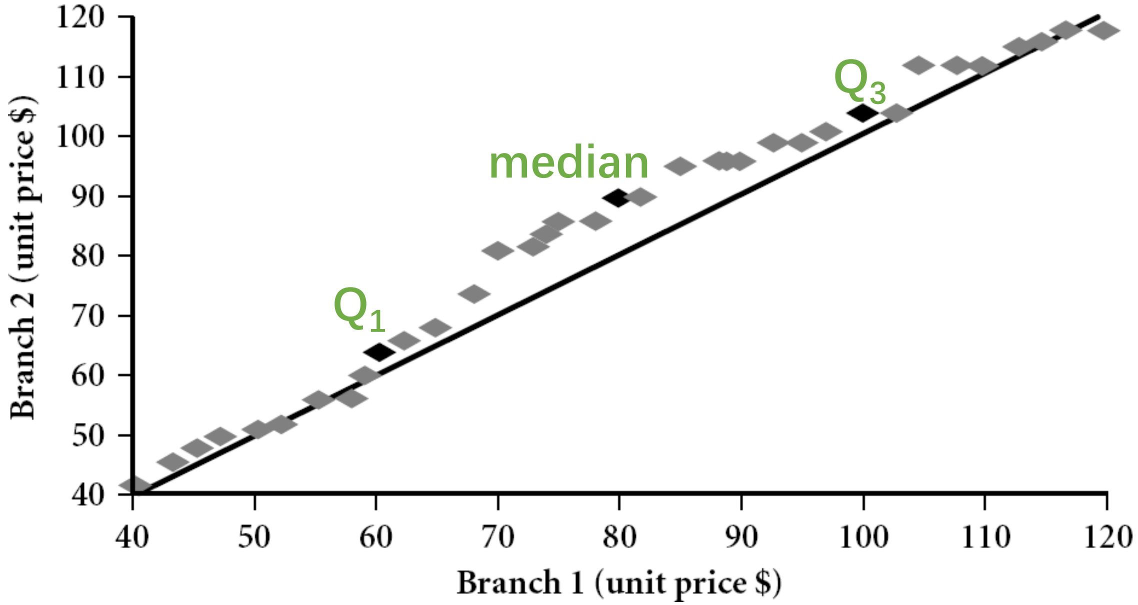
Quantile Plot

- **Quantile Plot**: Each value x_i is paired with f_i indicating that approximately $100 f_i\%$ of data $\leqslant x_i$



Quantile-Quantile (Q-Q) Plot

- **Quantile-Quantile (Q-Q) Plot:** graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Which branch has a lower price?
  - Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



Content

- Data Attributes and Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Probability Inequalities

Proximity Measure

- Proximity refers to a similarity or dissimilarity of two data objects
- 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
- Applications: clustering, anomaly detection, and nearest neighbor search

Proximity Measure for Binary Attributes

- A contingency table for binary data
  - E.g. (1,0,1,0,1,0,⋯)

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables:

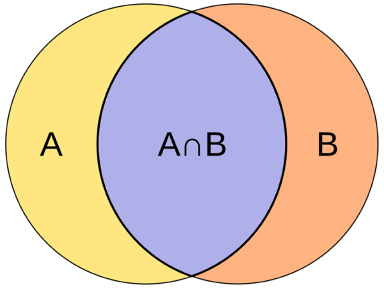
$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables(if t is too large):

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient:

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$



  - Note: Jaccard coefficient is the ratio of intersection over union of two sets.

Minkowski Distance

- Minkowski distance:

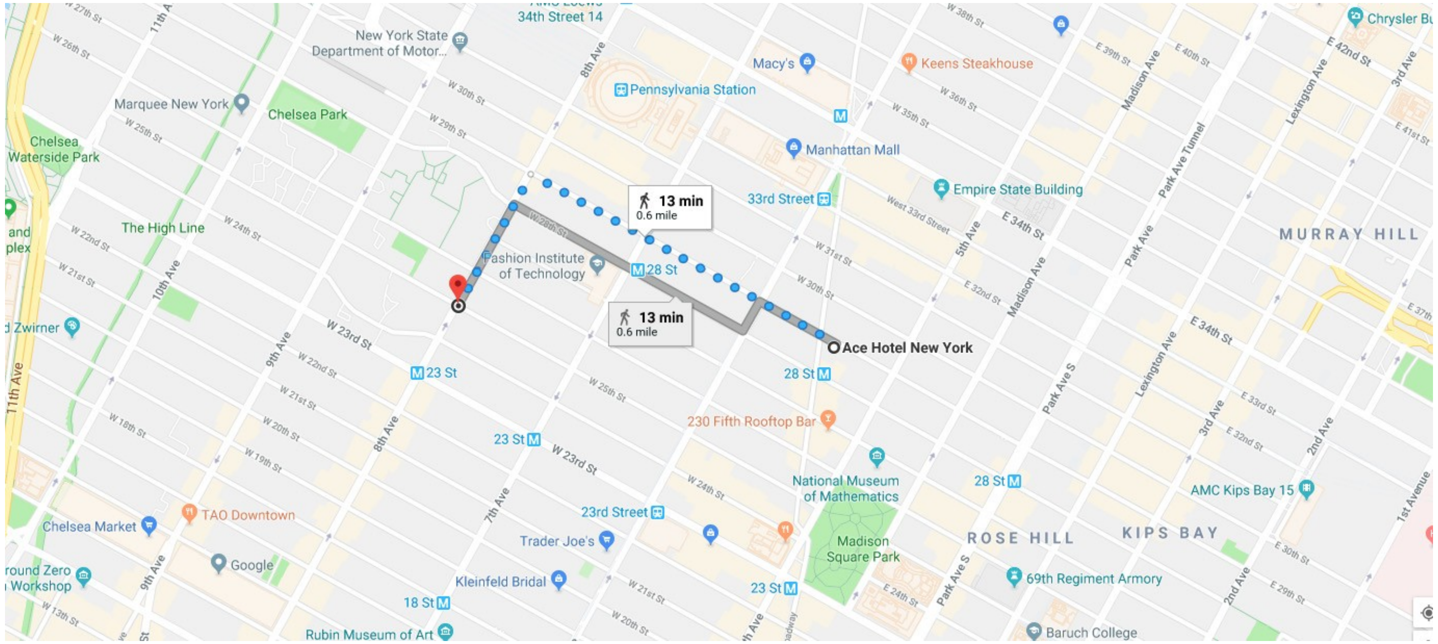
$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$
$$x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$$
$$d(i,j) = \left(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h\right)^{\frac{1}{h}}$$

  - *h* is the order (the distance so defined is also called L_{h} norm)
- Properties
  - Positive definiteness: $d(i,j) > 0 \; if \; i \neq j, and \; d(i,i) = 0$
  - Symmetry: $d(i,j) = d(j,i)$
  - Triangle Inequality: $d(i,j) \leq d(i,k) + d(k,j)$
- A distance that satisfies these properties is a metric

Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm) distance
  - $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$

    E.g., the Hamming distance: the number of bits that are different between two binary vectors



Cosine Similarity

- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
d1	5	0	3	0	2	0	0	2	0	0
d2	3	0	2	0	1	1	0	1	0	1
d3	0	7	0	2	1	0	0	3	0	0
d4	0	1	0	0	1	2	2	0	3	0

- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \cdot d_2)/(\| d_1 \| \cdot \| d_2 \|)$$

where • indicates vector dot product, $\| d \|$ is the length of vector d

Content

- Data Attributes and Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Probability Inequalities

Markov's Inequality

- If $\boldsymbol{X}$ is a **non-negative** r.v. then for every $c > 0$:

$$\Pr\left[\boldsymbol{X} \geq c\mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

- **Proof**

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}] &= \textstyle\sum_i i \cdot \Pr[\boldsymbol{X} = i] && \text{(by definition)} \\
&\geq \textstyle\sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} i \cdot \Pr[\boldsymbol{X} = i] && \text{(pick only some i's)} \\
&\geq \textstyle\sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} c\mathbb{E}[\boldsymbol{X}] \cdot \Pr[\boldsymbol{X} = i] && (i \geq c\mathbb{E}[\boldsymbol{X}]) \\
&= c\mathbb{E}[\boldsymbol{X}] \textstyle\sum_{i=c\mathbb{E}[\boldsymbol{X}]}^{\infty} \Pr[\boldsymbol{X} = i] && \text{(by linearity)} \\
&= c\mathbb{E}[\boldsymbol{X}] \Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] && \text{(same as above)}
\end{aligned}$$

$$\Rightarrow \Pr\left[\boldsymbol{X} \geq c\ \mathbb{E}[\boldsymbol{X}]\right] \leq \frac{1}{c}$$

Pro: always works!

Cons:

Not very precise

Doesn't work for the lower tail: $\Pr\left[\boldsymbol{X} \leq c\ \mathbb{E}[\boldsymbol{X}]\right]$

Chebyshev's Inequality

Measuring the Dispersion of Data

- Regardless of how data is distributed, at least $\left(1 - \frac{1}{k^2}\right)$ of the points must lie within $k\sigma$ of the mean.
  - Thus at least 75% must lie within two sigma of the mean.

- For every $c > 0$:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$$

- Proof:

$$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right]$$

$$= \Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 Var[\boldsymbol{X}]\right] \quad \text{(by squaring)}$$

$$= \Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2 \geq c^2 \mathbb{E}[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]|^2]\right] \quad \text{(def. of Var)}$$

$$\leq \frac{1}{c^2} \quad \text{(by Markov's inequality)}$$

Chernoff bound

- Let $X_1 \dots X_{\boldsymbol{t}}$ be **independent and identically distributed** random values with range [0,1] and expectation μ.
- Then if $X = \frac{1}{t}\sum_i X_i$ and $1 > \delta > 0$,

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu \boldsymbol{t} \delta^2}{3}\right)$$

Chernoff v.s Chebyshev: Example

Let $X = \frac{1}{t}\sum_i X_i, \sigma = Var[X_i]$:

- Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq c'] \leq \frac{Var[\boldsymbol{X}]}{c'^2} = \frac{\sigma}{\boldsymbol{t}\ \boldsymbol{c'^2}}$

$\Pr\left[|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]| \geq c\sqrt{Var[\boldsymbol{X}]}\right] \leq \frac{1}{c^2}$

- Chernoff: $\Pr[|X - \mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\mu \boldsymbol{t} \delta^2}{3c}\right)$

If $\boldsymbol{t}$ is very big:

- Values $\mu, \sigma, \delta, c, c'$ are all constants!
  - Chebyshev: $\Pr[|\boldsymbol{X} - \mu| \geq z] = O\left(\frac{1}{\boldsymbol{t}}\right)$
  - Chernoff: $\Pr[|\boldsymbol{X} - \mu| \geq z] = e^{-\Omega(\boldsymbol{t})}$

So is Chernoff always better for us?
Yes, if we have i.i.d. variables.

Summary

- Data Attributes
- Basic Statistical Descriptions of Data
  - Centrality/Dispersion
- Measuring Data Similarity and Dissimilarity
  - Distances for binary/numerical
- Probability Inequalities
  - Markov/Chebyshev/Chernoff