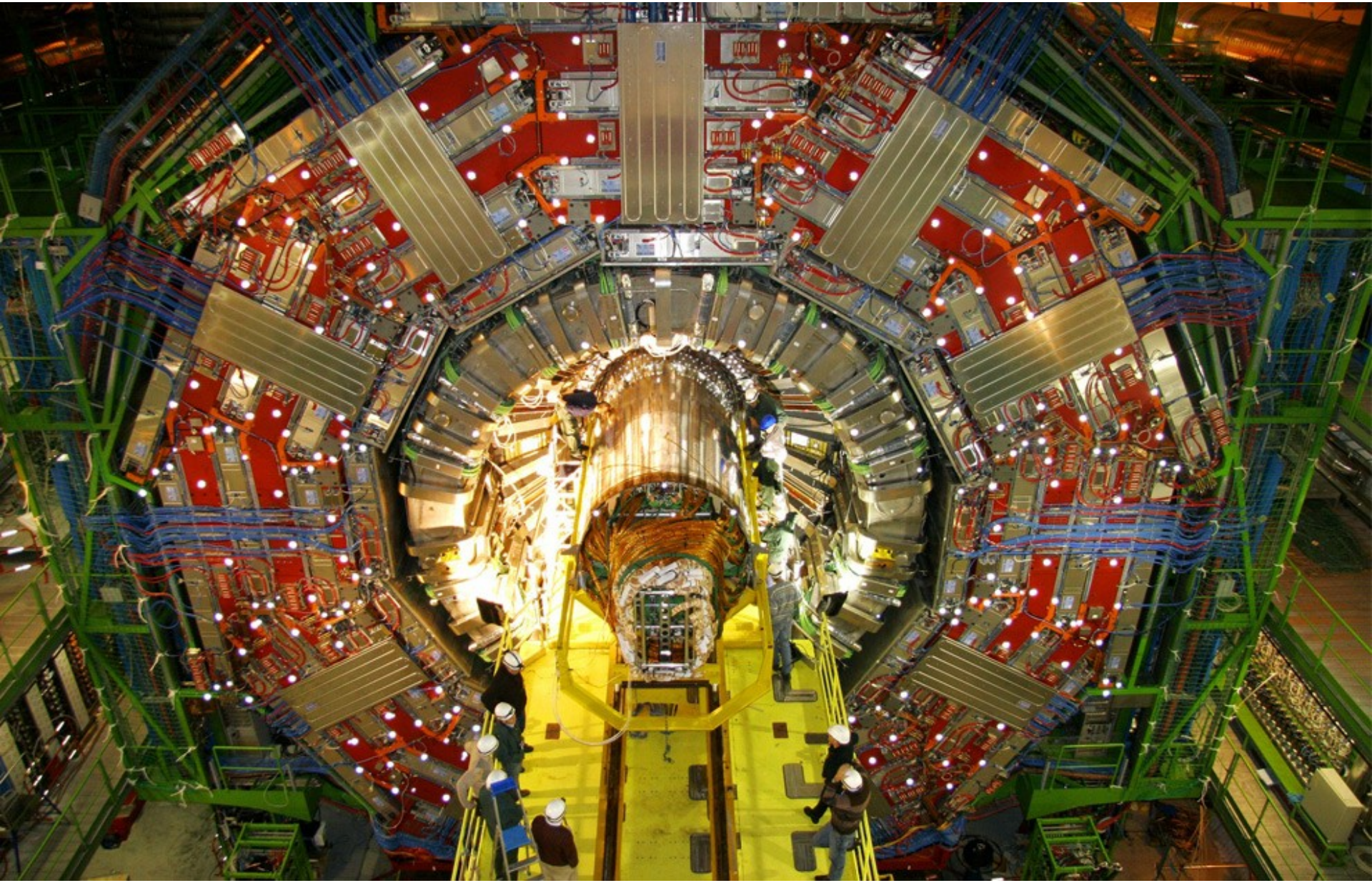
Introduction to Data & Basic Data Analysis

Data Every Where!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network

How much data?

- Google processes 20 PB a day (2008)
- Way back Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's Large Hydron Collider (LHC) generates 15 PB a year



The Earth scope

- The Earth scope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Statistics 101

Random Sample and Statistics

- *Population: is used to refer to the set or universe of all* entities under study.
- However, looking at the entire population may not be feasible, or may be too expensive.
- Instead, we draw a random sample from the population, and compute appropriate *statistics* from the sample, that give estimates of the corresponding population parameters of interest.

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

Table 1.2: Iris Dataset: `sepal length`

Statistic

- Let Si denote the random variable corresponding to data point xi , then a *statistic* $\hat{\theta}$ is a function $\hat{\theta} : (S1, S2, \cdots, Sn) \rightarrow R$.

- If we use the value of a statistic to estimate a population parameter, this value is called a *point estimate of the parameter, and the statistic is called as an estimator of the* parameter.

Empirical Cumulative Distribution Function

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(S_i \leq x)}{n}$$

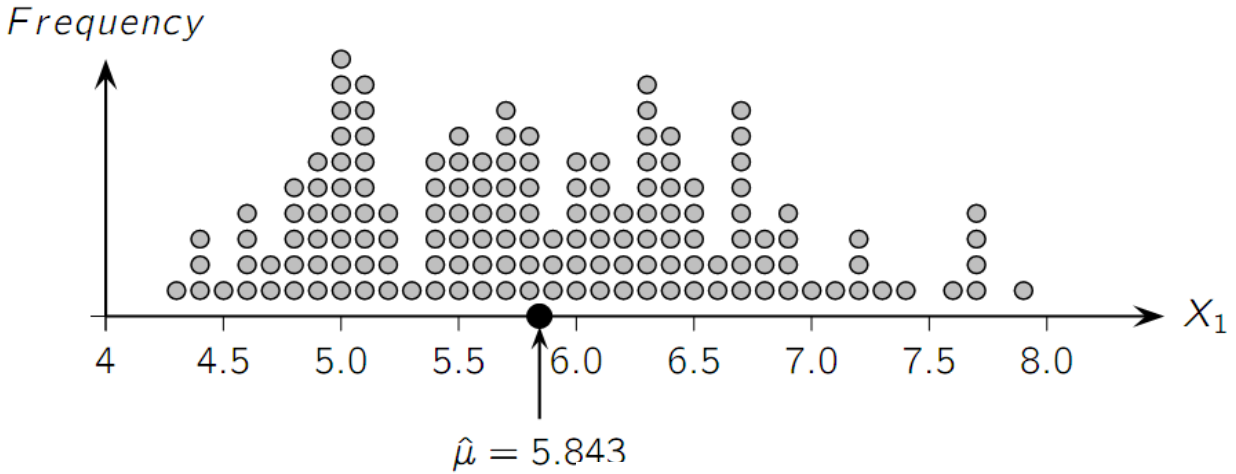
Where

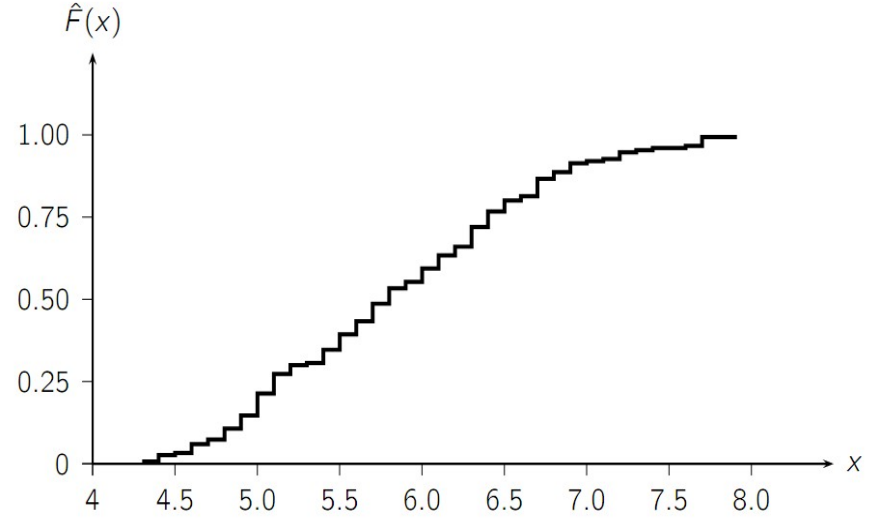
$$I(S_i \leq x) = \begin{cases} 1 & \text{if } S_i \leq x \\ 0 & \text{if } S_i > x \end{cases}$$

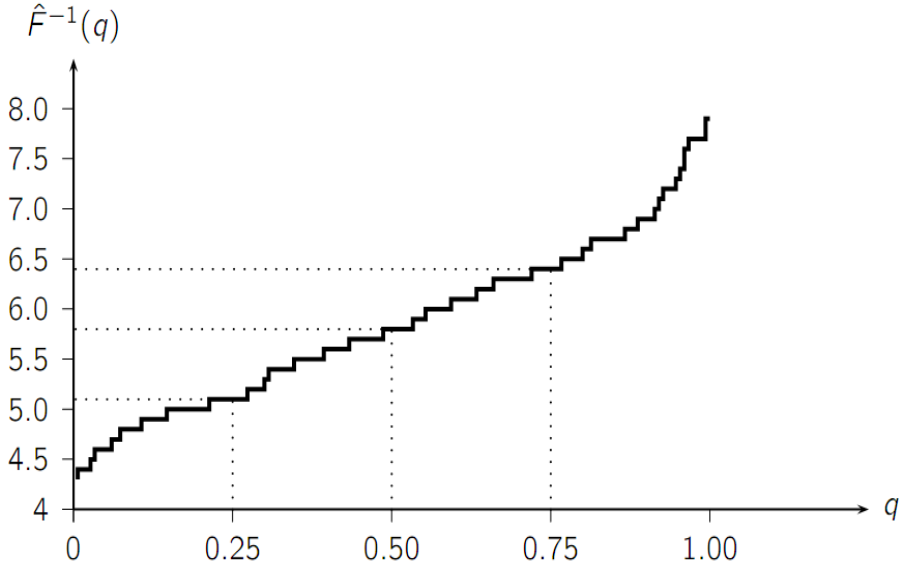
Inverse Cumulative Distribution Function

$$F^{-1}(q) = \min\{x : F(x) > q\} \qquad \text{for } q \in [0, 1]$$

Example


Frequency
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
X_1
$\hat{\mu} = 5.843$


$\hat{F}(x)$
1.00
0.75
0.50
0.25
0
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
x


$\hat{F}^{-1}(q)$
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4
0
0.25
0.50
0.75
1.00
q

Measures of Central Tendency (Mean)

Population Mean:

$$\mu = E[X] = \sum_x x\, f(x)$$

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

Sample Mean (Unbiased, not robust):

$$\hat{\mu} = \sum_x x \hat{f}(x) = \sum_x x \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n S_i}{n}$$

$$E[\hat{\mu}] = E\left[\frac{\sum_{i=1}^n S_i}{n} \right] = \frac{1}{n} \sum_{i=1}^n E[S_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

Measures of Central Tendency (Median)

Population Median:

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

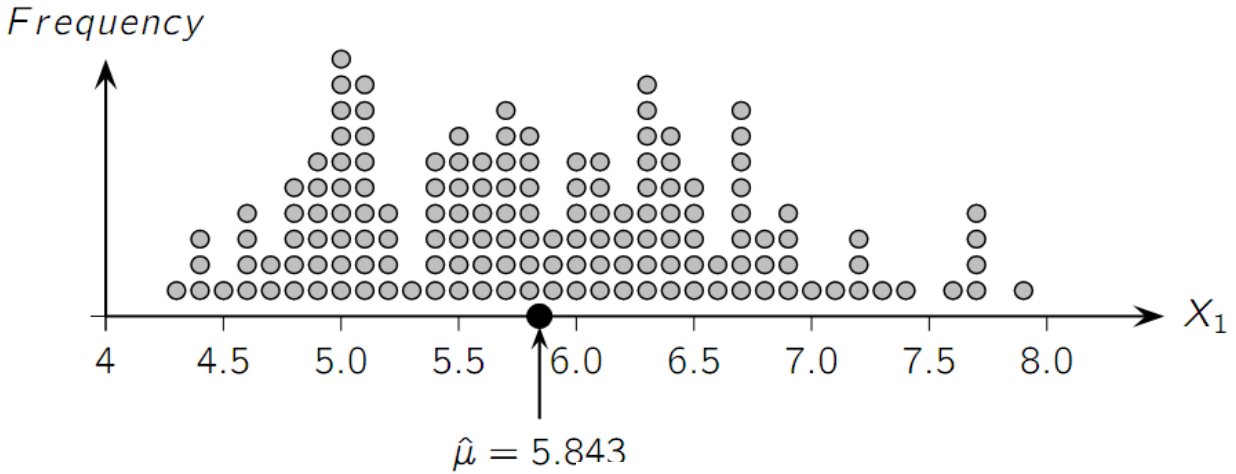
or

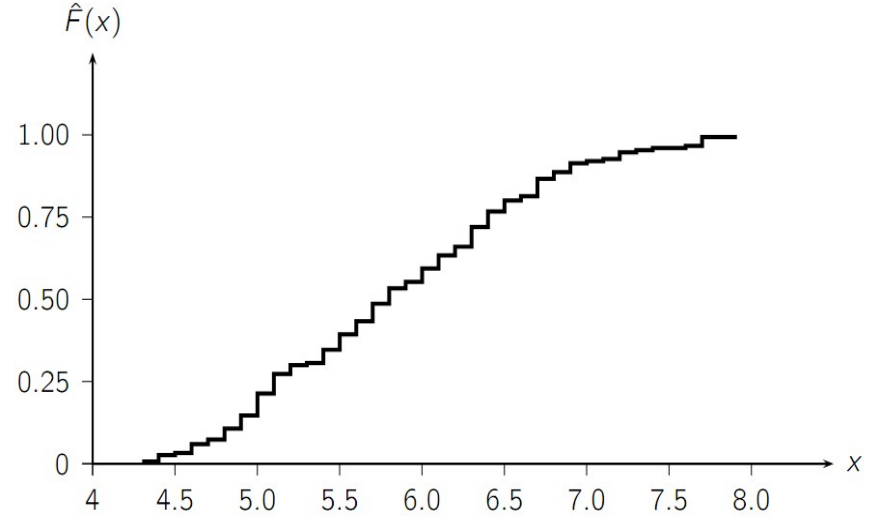
$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

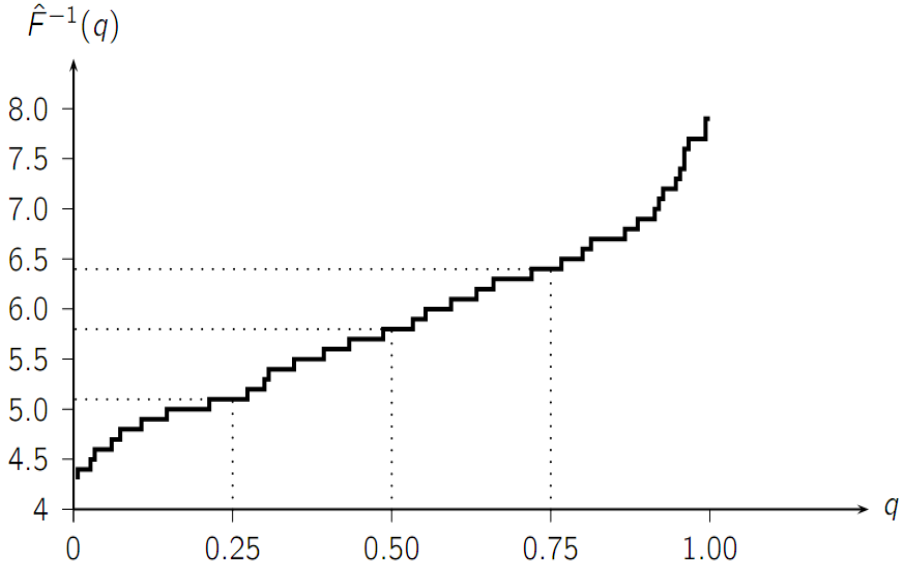
Sample Median:

$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

Example


Frequency
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
X_1
$\hat{\mu} = 5.843$


$\hat{F}(x)$
1.00
0.75
0.50
0.25
0
4
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
x


$\hat{F}^{-1}(q)$
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4
0
0.25
0.50
0.75
1.00
q

Measures of Dispersion (Range)

Range:

$$r = \max_x\{x\} - \min_x\{x\}$$

Sample Range:

$$\hat{r} = \max_i\{S_i\} - \min_i\{S_i\} = \max_i\{x_i\} - \min_i\{x_i\}$$

- Not robust, sensitive to extreme values

Measures of Dispersion (Inter-Quartile Range)

Inter-Quartile Range (IQR):

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

Sample IQR:

$$\widehat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

- More robust

Measures of Dispersion
(Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 \, f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Measures of Dispersion (Variance and Standard Deviation)

Variance:

$$var(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

Standard Deviation:

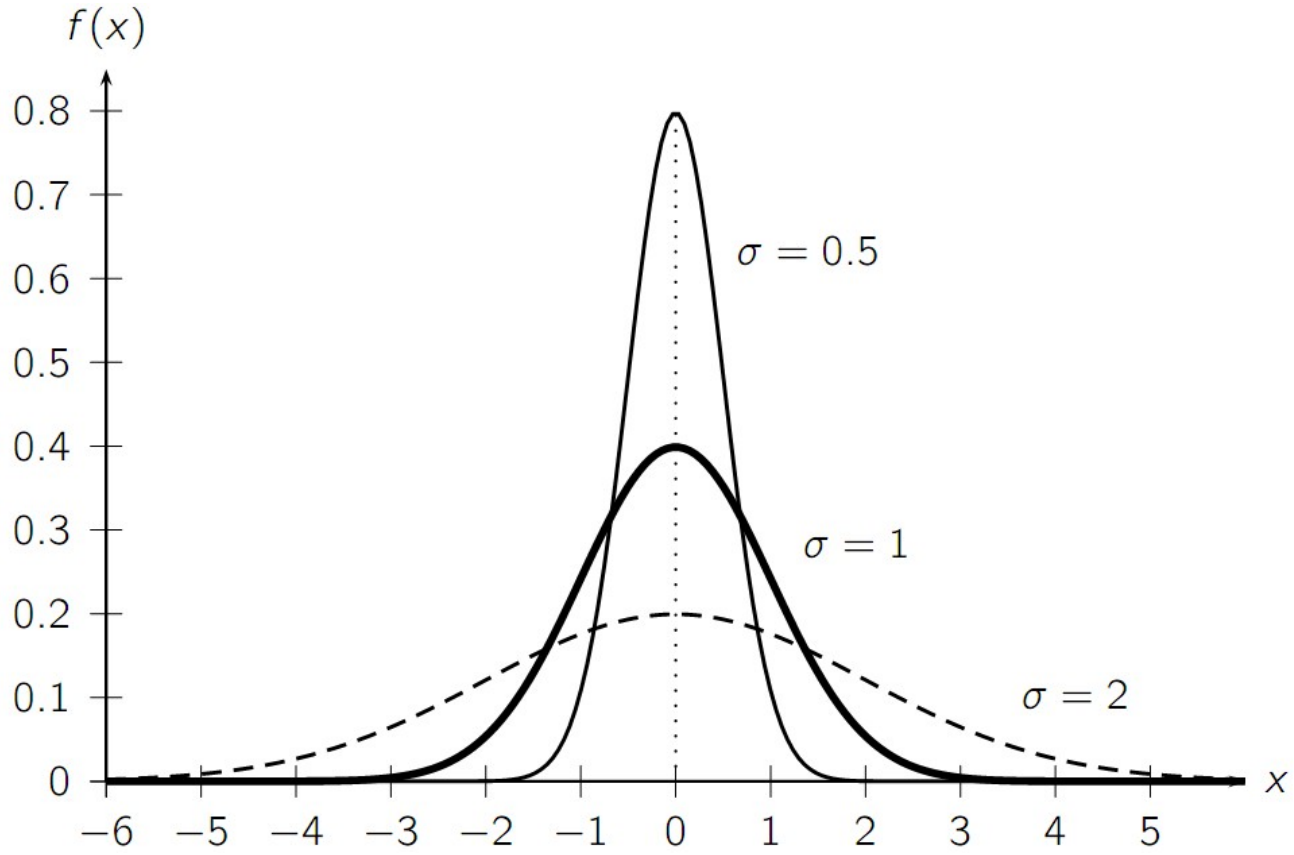
$$\begin{aligned} \sigma^2 = var(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Sample Variance & Standard Deviation:

$$\hat{\sigma}^2 = \sum_x (x-\hat{\mu})^2 \hat{f}(x) = \sum_x (x-\hat{\mu})^2 \left(\frac{\sum_{i=1}^n I(S_i = x)}{n} \right) = \frac{\sum_{i=1}^n (S_i - \hat{\mu})^2}{n}$$

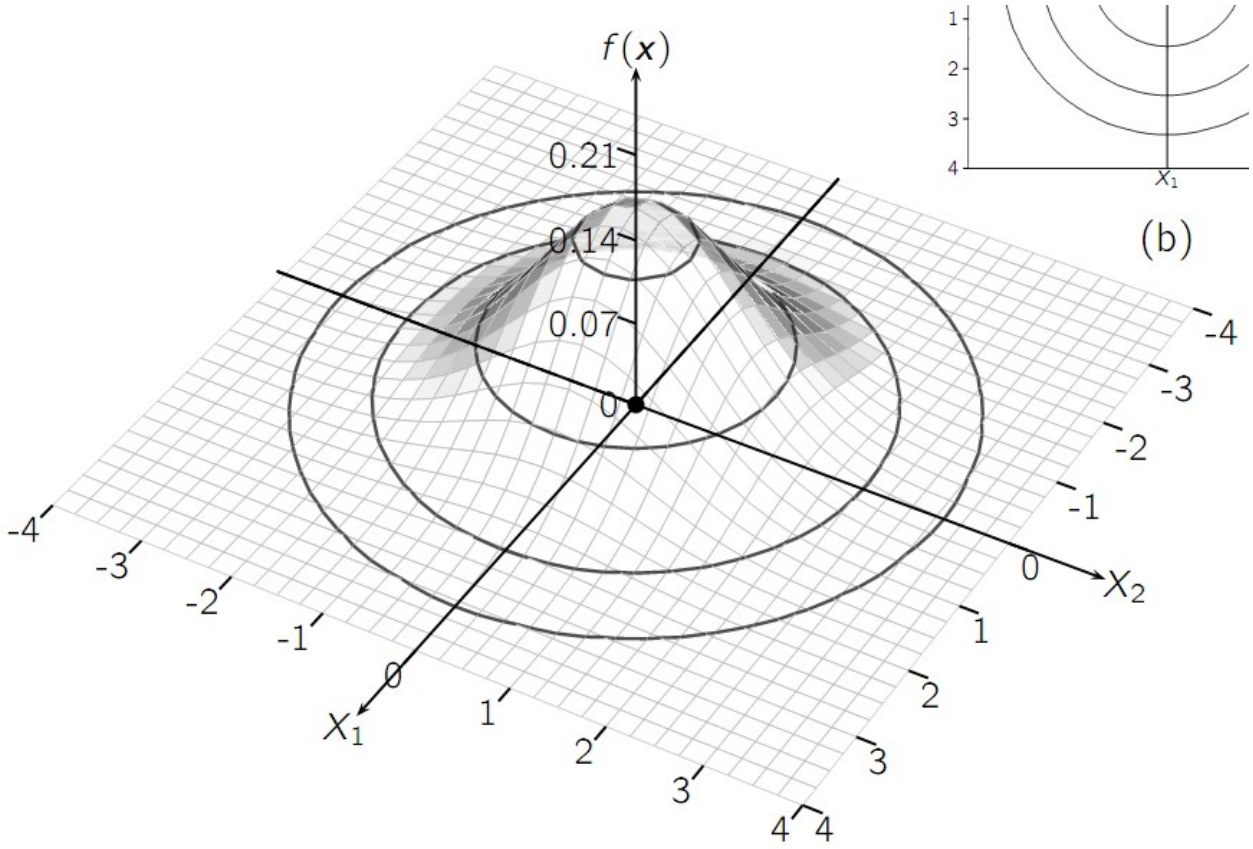
Univariate Normal Distribution

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



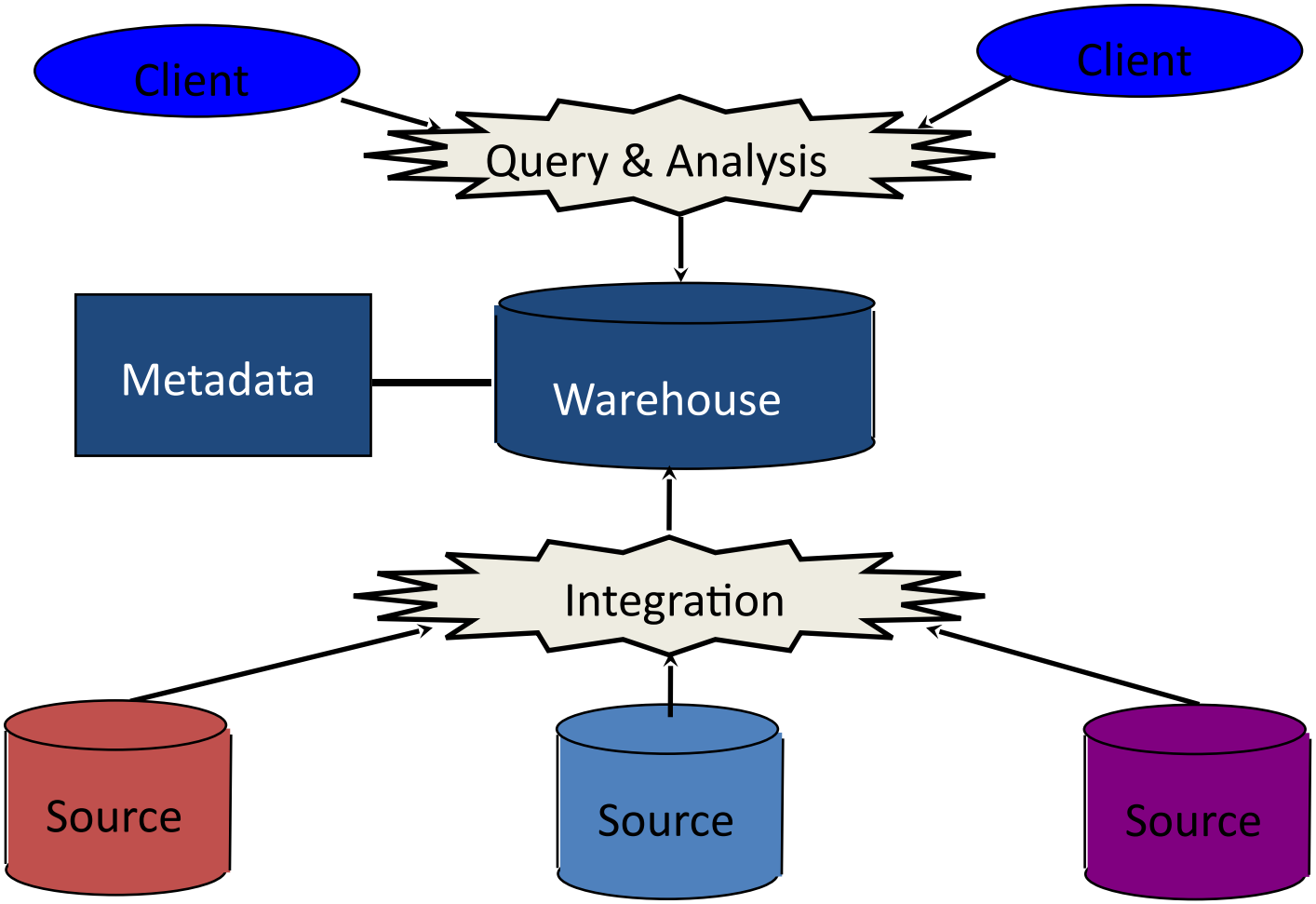
Multivariate Normal Distribution

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2} \right\}$$



OLAP and Data Mining

Warehouse Architecture

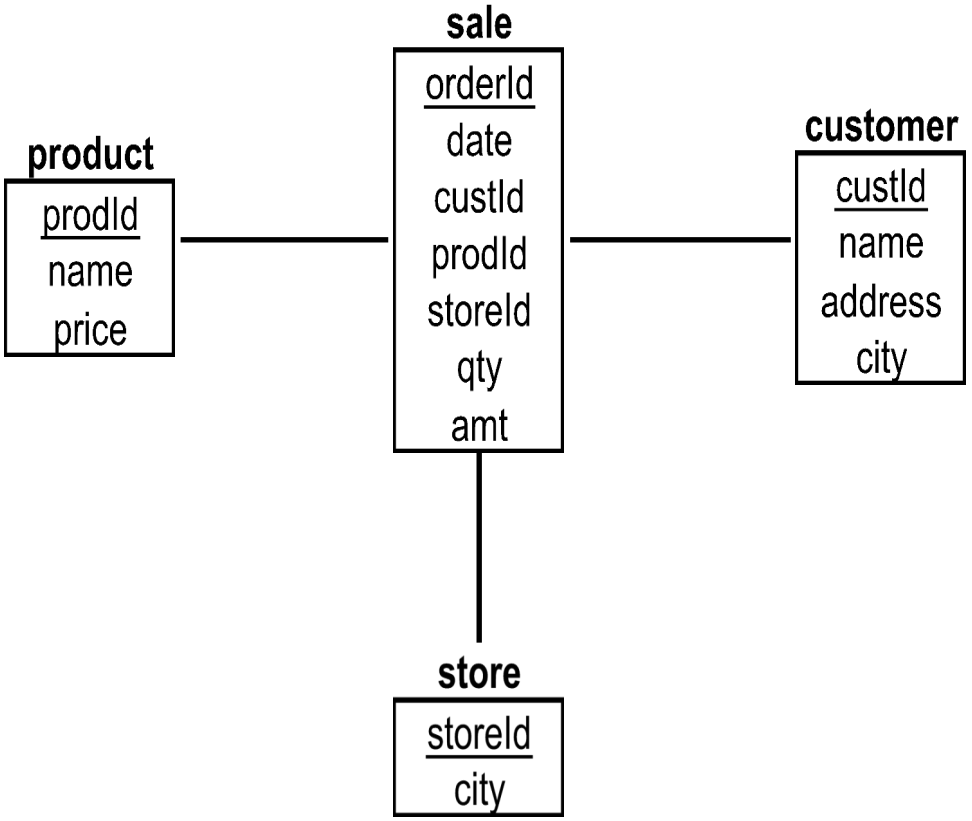


Star Schemas

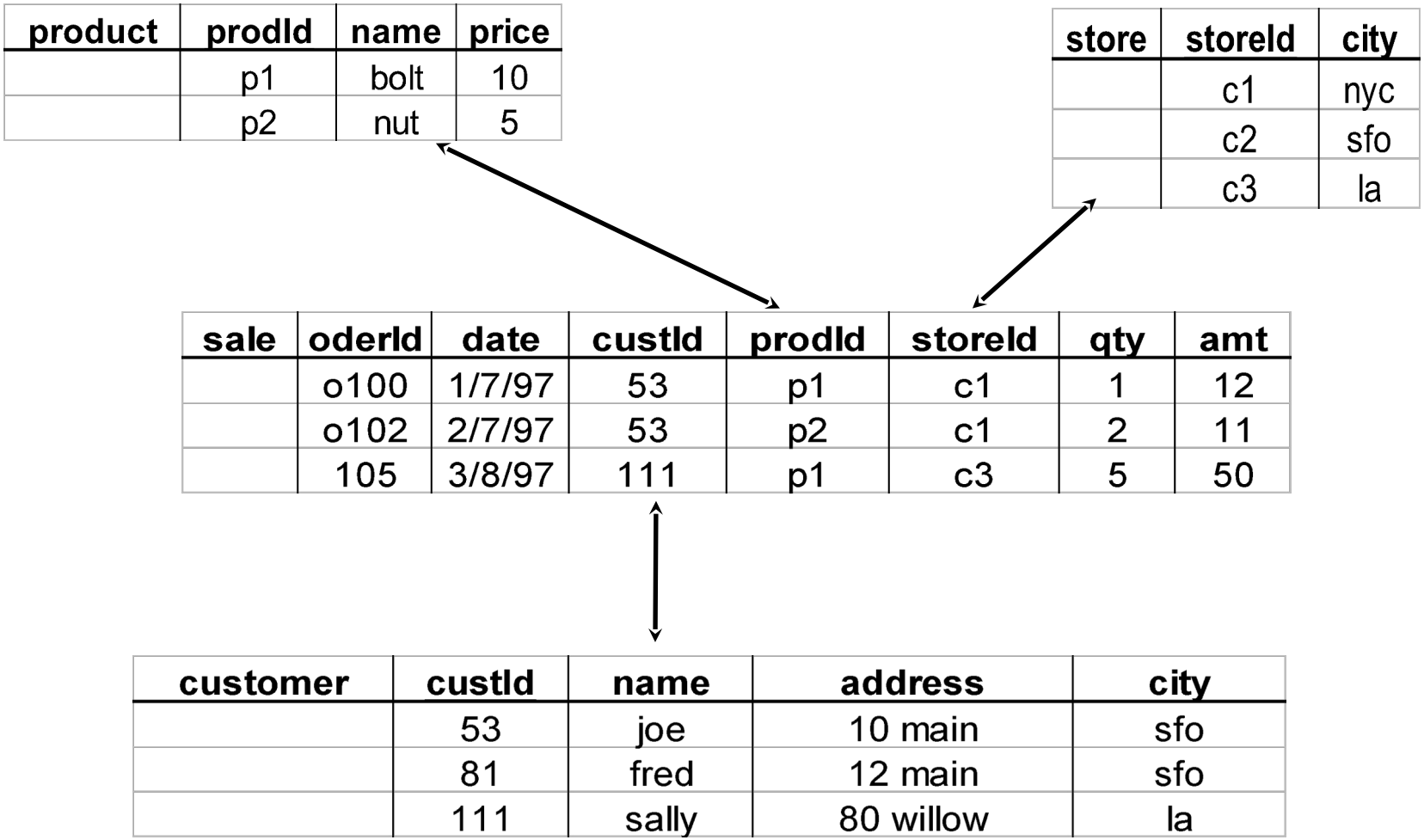
- A *star schema* is a common organization for data at a warehouse. It consists of:
  1. *Fact table* : a very large accumulation of facts such as sales.
     - Often “insert-only.”
  2. *Dimension tables* : smaller, generally static information about the entities involved in the facts.

Terms

- Fact table
- Dimension tables
- Measures



Star



product	prodId	name	price
	p1	bolt	10
	p2	nut	5

store	storeId	city
	c1	nyc
	c2	sfo
	c3	la

sale	oderId	date	custId	prodId	storeId	qty	amt
	o100	1/7/97	53	p1	c1	1	12
	o102	2/7/97	53	p2	c1	2	11
	105	3/8/97	111	p1	c3	5	50

customer	custId	name	address	city
	53	joe	10 main	sfo
	81	fred	12 main	sfo
	111	sally	80 willow	la

Cube

Fact table view:

sale	prodId	storeId	amt
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

Multi-dimensional cube:

	c1	c2	c3
p1	12		50
p2	11	8	

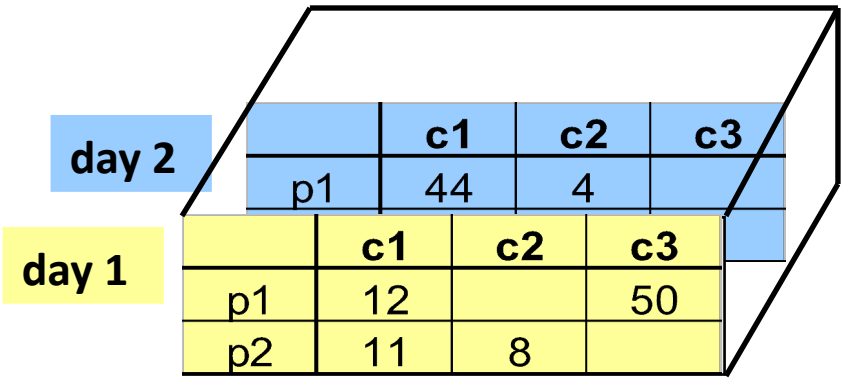
dimensions = 2

3-D Cube

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Multi-dimensional cube:



dimensions = 3

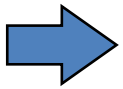
ROLAP vs. MOLAP

- ROLAP:
  Relational On-Line Analytical Processing
- MOLAP:
  Multi-Dimensional On-Line Analytical Processing

Aggregates

- Add up amounts for day 1
- In SQL: SELECT sum(amt) FROM SALE
  WHERE date = 1

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

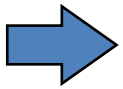


81

Aggregates

- Add up amounts by day
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

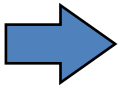


ans	date	sum
	1	81
	2	48

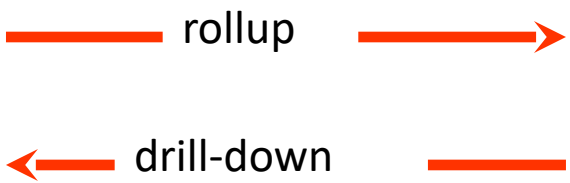
Another Example

- Add up amounts by day, product
- In SQL: SELECT date, sum(amt) FROM SALE
  GROUP BY date, prodId

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4



sale	prodId	date	amt
	p1	1	62
	p2	1	19
	p1	2	48



Aggregates

- Operators: sum, count, max, min, median, ave
- “Having” clause
- Using dimension hierarchy
  - average by region (within store)
  - maximum by month (within date)

What is Data Mining?

- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Collaborative Filter [Predictive]

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Decision Trees

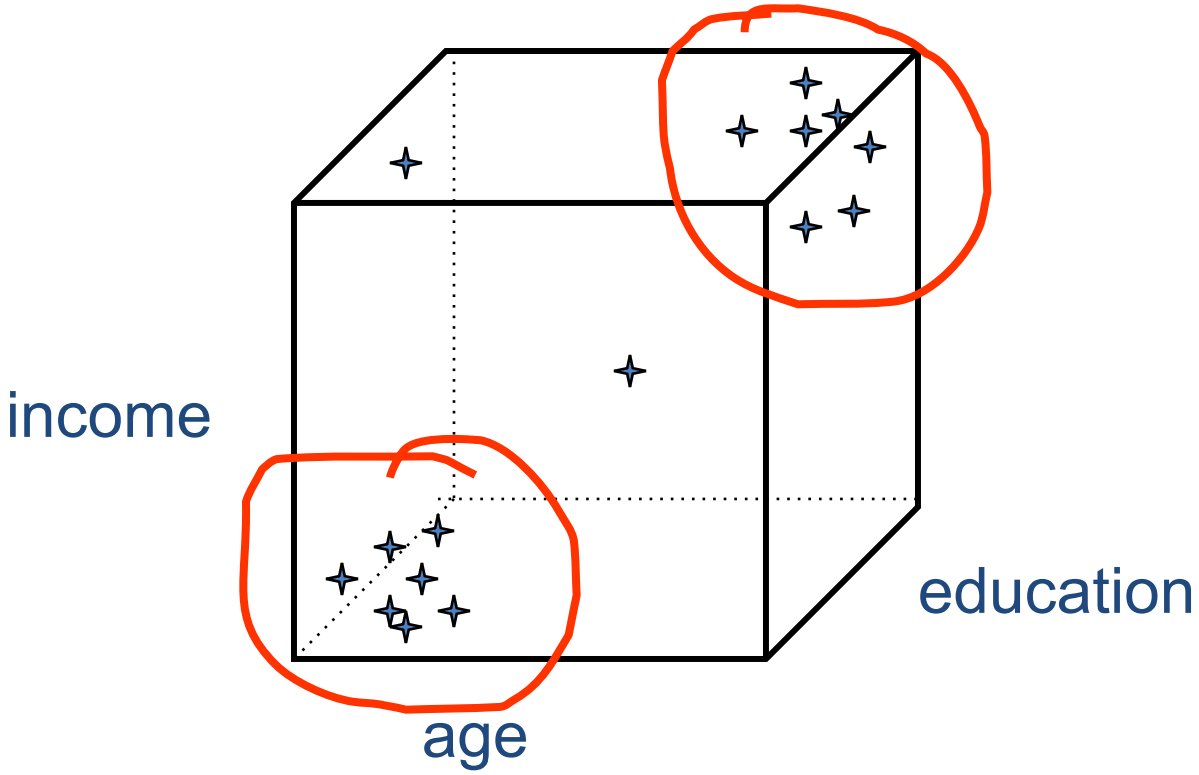
Example:

- Conducted survey to see what customers were interested in new model car
- Want to select customers for advertising campaign

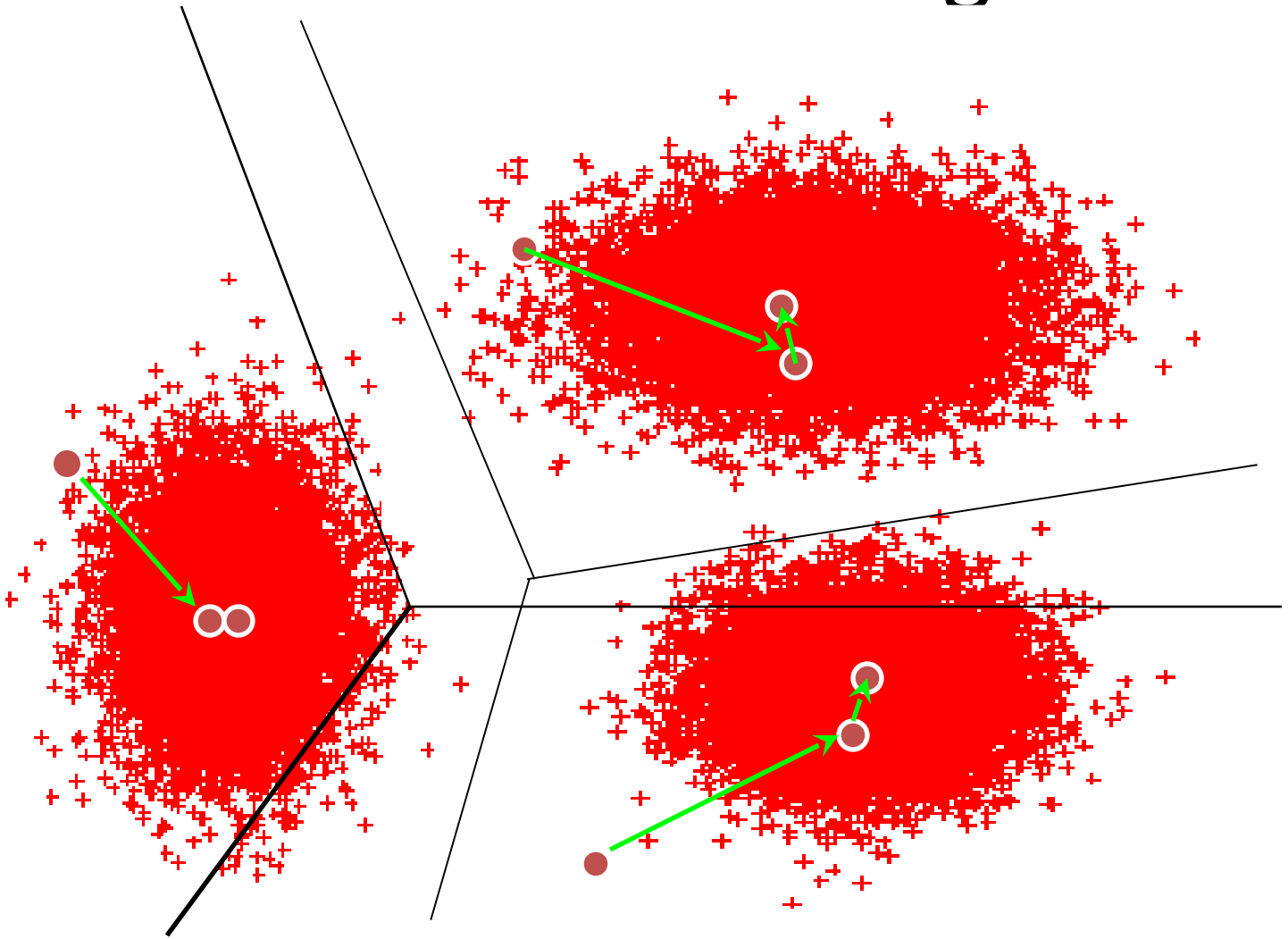
sale	custId	car	age	city	newCar
	c1	taurus	27	sf	yes
	c2	van	35	la	yes
	c3	van	40	sf	yes
	c4	taurus	22	sf	yes
	c5	merc	50	la	no
	c6	taurus	25	la	no

training set

Clustering



K-Means Clustering



Association Rule Mining

sales records:

transaction id	customer id	products bought
tran1	cust33	p2, p5, p8
tran2	cust45	p5, p8, p11
tran3	cust12	p1, p9
tran4	cust40	p5, p8, p11
tran5	cust12	p2, p9
tran6	cust12	p9

market-basket data

- Trend: Products p5, p8 often bough together
- Trend: Customer 12 likes product p9

Association Rule Discovery

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!
- Supermarket shelf management.
- Inventory Managemnt

Collaborative Filtering

- Goal: predict what movies/books/… a person may be interested in, on the basis of
  - Past preferences of the person
  - Other people with similar past preferences
  - The preferences of such people for a new movie/book/…
- One approach based on repeated clustering
  - Cluster people on the basis of preferences for movies
  - Then cluster movies on the basis of being liked by the same clusters of people
  - Again cluster people based on their preferences for (the newly created clusters of) movies
  - Repeat above till equilibrium
- Above problem is an instance of **collaborative filtering**, where users collaborate in the task of filtering information to find information of interest

Other Types of Mining

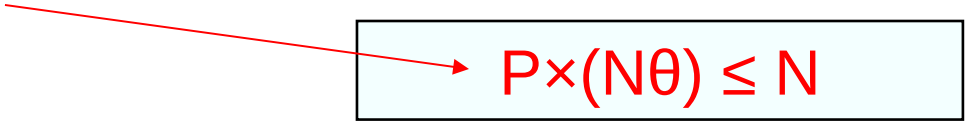
- **Text mining**: application of data mining to textual documents
  - cluster Web pages to find related pages
  - cluster pages a user has visited to organize their visit history
  - classify Web pages automatically into a Web directory
- **Graph Mining**:
  - Deal with graph data

Data Streams

- What are Data Streams?
  - Continuous streams
  - Huge, Fast, and Changing
- Why Data Streams?
  - The arriving speed of streams and the huge amount of data are beyond our capability to store them.
  - “Real-time” processing
- Window Models
  - Landscape window (Entire Data Stream)
  - Sliding Window
  - Damped Window
- Mining Data Stream

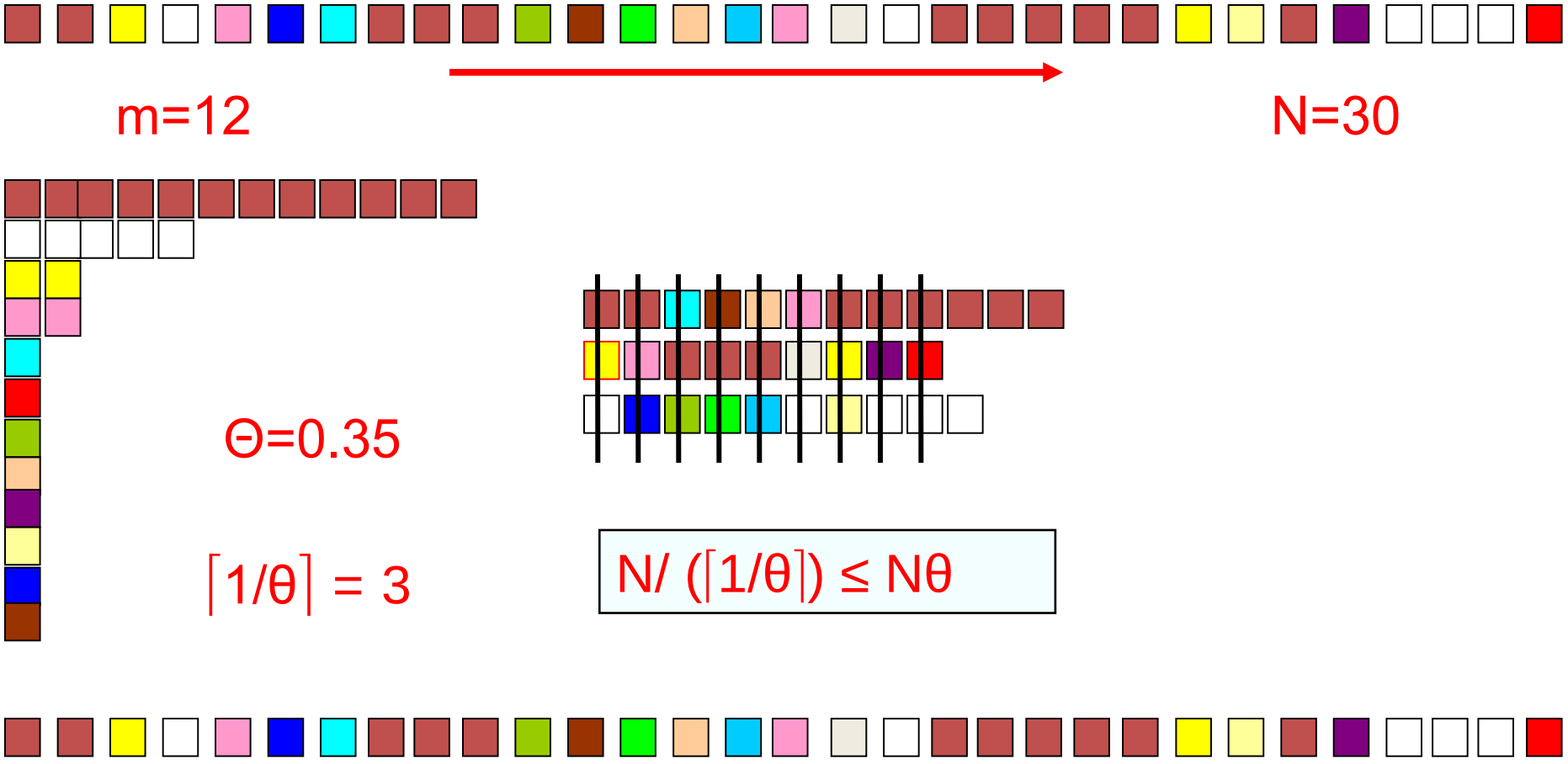
A Simple Problem

- Finding frequent items
  - Given a sequence $(x_1,\ldots x_N)$ where $x_i \in [1,m]$, and a real number θ between zero and one.
  - Looking for x_i whose frequency > θ
  - Naïve Algorithm (m counters)
- The number of frequent items ≤ 1/θ
- Problem: N>>m>>1/θ



KRP algorithm

— Karp, et. al (TODS' 03)



Streaming Sample Problem

- Scan the dataset once
- Sample K records
  - Each one has equally probability to be sampled
  - Total N record: K/N