Data Mining Concepts and Data Statistics

P. Krishna Reddy
IIIT Hyderabad

Documents Uploaded

- Data, Information, Knowledge, and Wisdom by Gene Bellinger, Durval Castro, Anthony Mills
- From Data to Wisdom: A Note by Russell Ackoff
- Data Science— A Systematic Treatment by M. TAMER ÖZSU, Communications of ACM, 2023
- Theoretical frameworks for data mining, Heikki Mannila, 2000.
- Data Science: A Comprehensive Overview, LONGBING CAO, ACM Computing Surveys, 2017

Reference

- Chapter 1 and Chapter 2 of the text book.

Detailed Syllabus

- **Unit1: Introduction, data and data preprocessing,** data summarization through characterization, discrimination and data cube techniques (9 hours)
- Unit 2: Concepts and algorithms for mining patterns and associations (10 hours)
- Unit 3: Concepts and algorithms related to classification and regression (10 hours)
- Unit 4: Concepts and algorithms for clustering the data (10 hours)
- Unit 5: Outlier analysis and future trends. (3 hours)

Presentation Outline

- **Data mining functionalities**
- Research Issues in Data mining
- Sources of data
- Data Objects and Attribute Types
- About Big Data
- Basic Statistical Descriptions of Data
- Data Visualization
- Case study
- Summary

Data Mining Functionalities

- Summarization
- Pattern mining, Association mining, correlation
- Classification
- Clustering
- Outlier analysis
- Sequential, trend and evolution analysis
- Structure and network analysis

Data Mining Function (1): Summarization

- Input: Large data
- Output: summarize/characterize an interested set of data and compare it with the contrasting sets at some high levels.
  - Data characterization: Summarizing the data of the class under study.
  - Data discrimination: Comparison of target class with one or a set of target class.
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region
- Output can be represented with pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs

Concept description: Characterization and discrimination

- Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions.
- Example

Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

Tabulation of Data

- Tabulation facilitates the presentation of large information into concise way under different titles and sub-titles.

Table 1. Students

Class	Rural				Urban			
	Male		Female		Male		Female	
	Hostellers	Day Scholars	Hostellers	Day scholars	Hostellers	Day Scholars	Hostellers	Day scholars
Intermediate								
Graduate								
Post Graduate								

Summarization techniques

- Statistical measures
- Attribute oriented induction
- Data cube based OLAP methods
  - Multidimensional summarization

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Approaches to mine Association Rules

- Apriori approach
- FP Tree approach
- Hash-based itemset counting: A *k*-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent
- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness
- Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

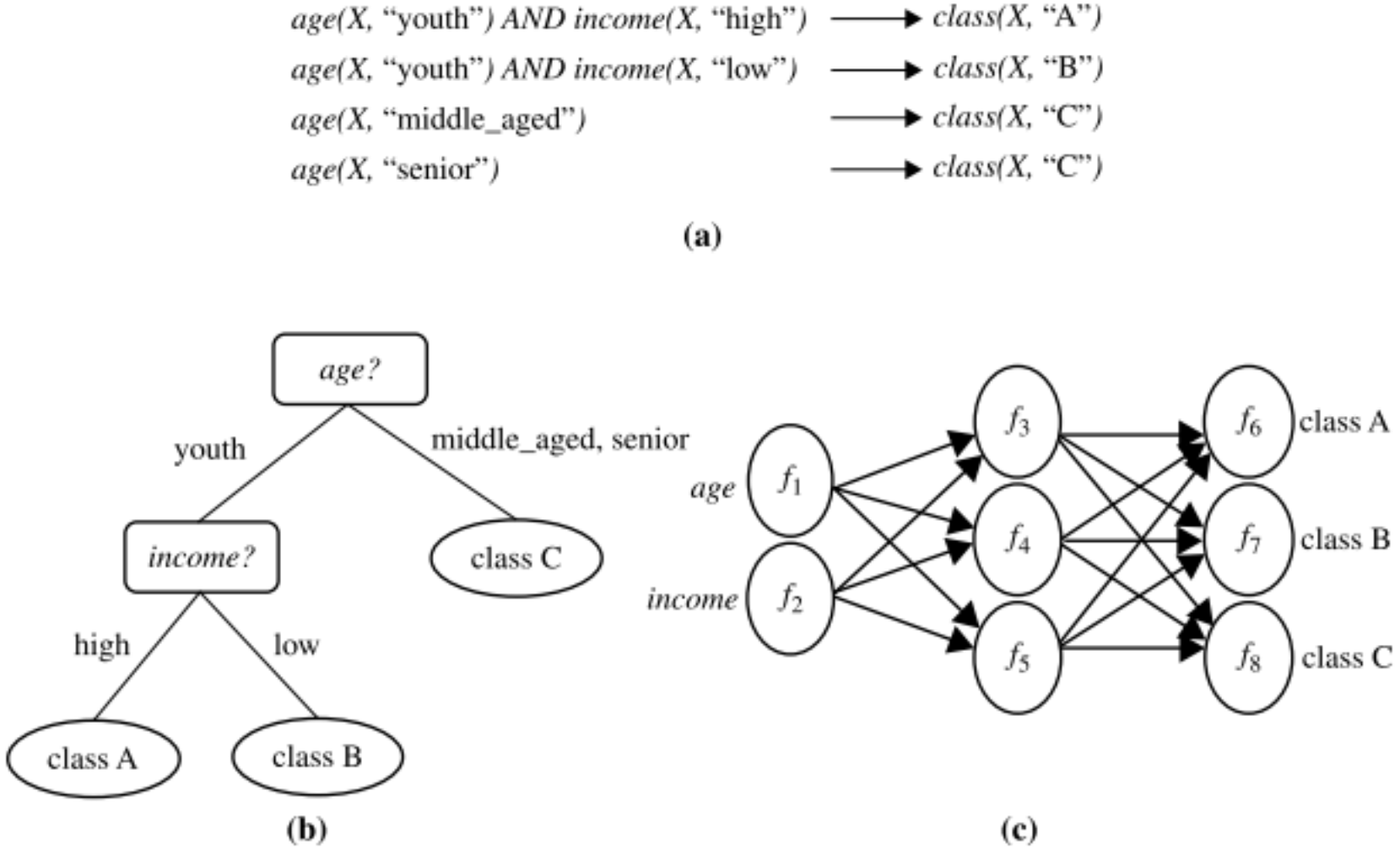


FIGURE 1.2

A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Approaches to classification

- Decision tree
- Bayesian Classification
- Neural networks
- Association based classification
- k-nearest neighbor classifier
- Ensemble classification

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

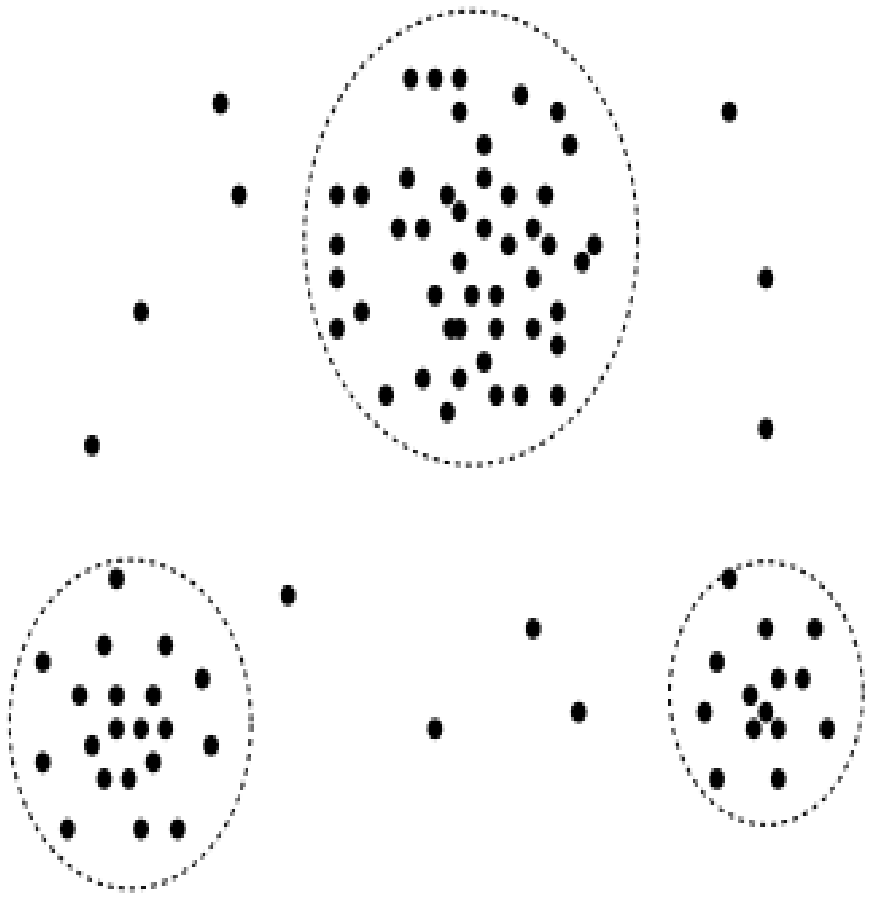


FIGURE 1.3

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

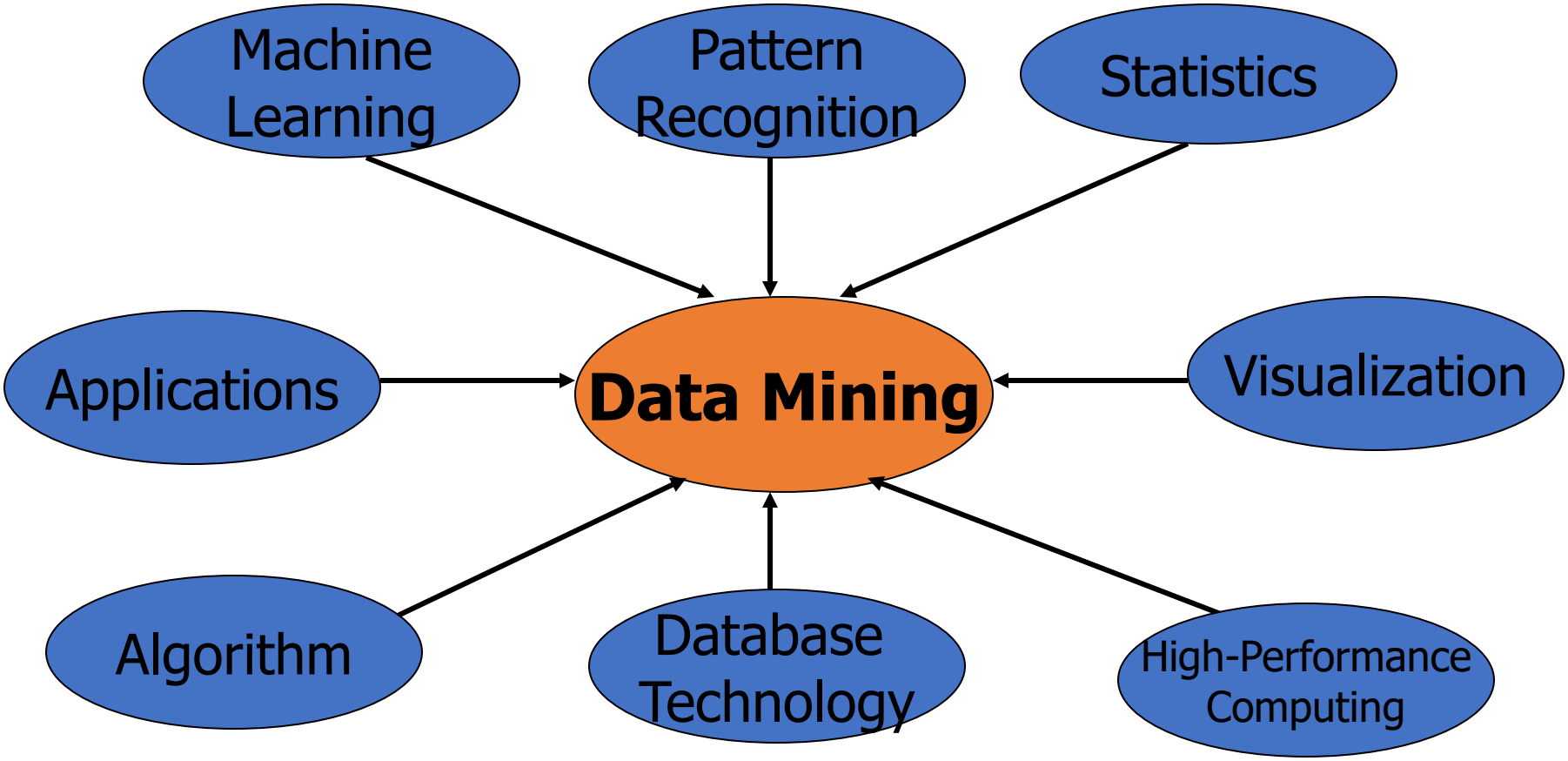
Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Data Mining: Confluence of Multiple Disciplines



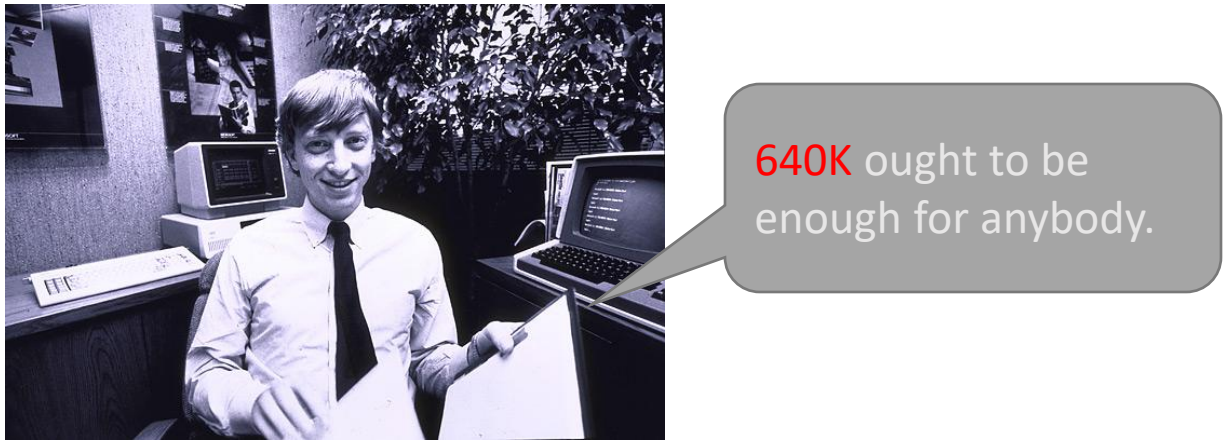
Presentation Outline

- Data mining functionalities
- **Research Issues in Data mining**
- Sources of data
- Data Objects and Attribute Types
- About Big Data
- Basic Statistical Descriptions of Data
- Data Visualization
- Case study
- Summary

About Big Data

How much data?

- Google processes 20 PB a day (2008)
- 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

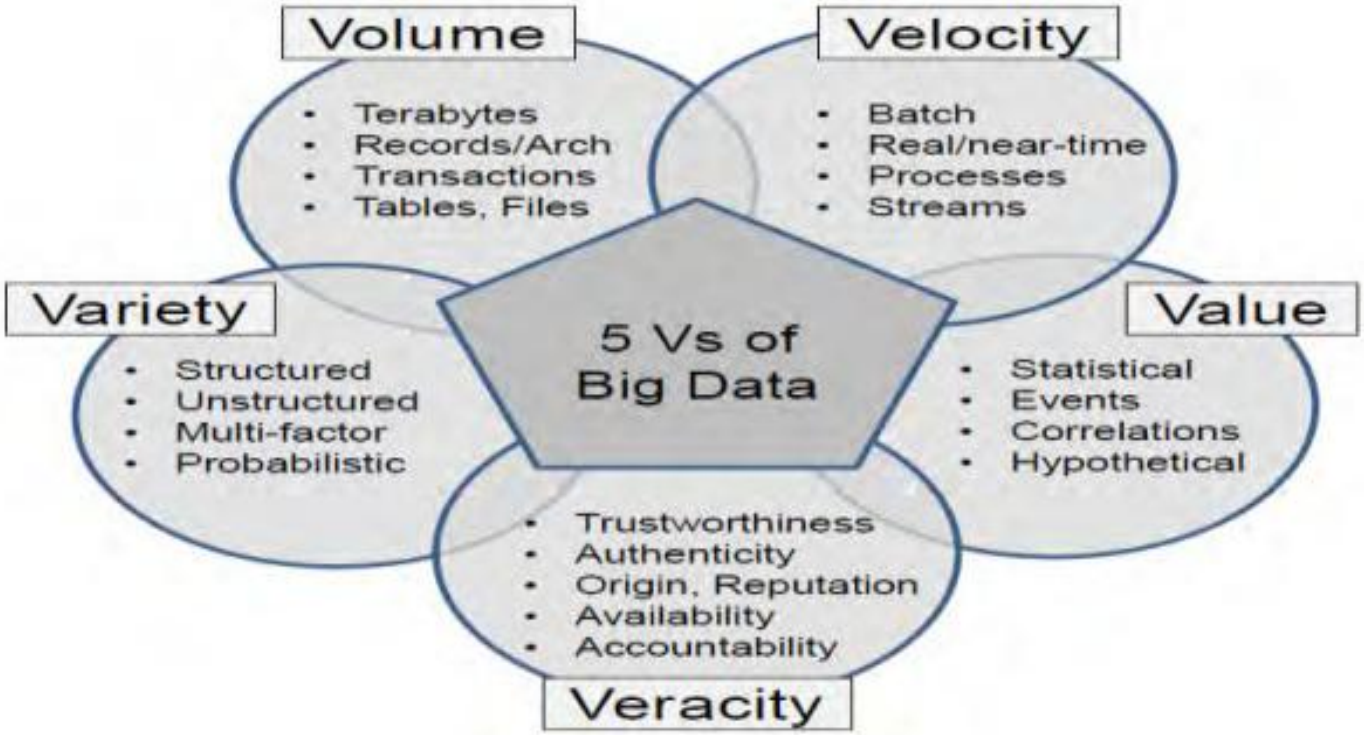


Big Data Definition

- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

Characteristics of Big data



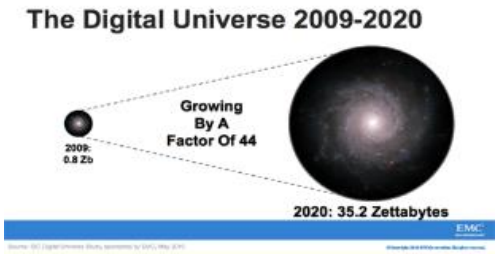
Volume

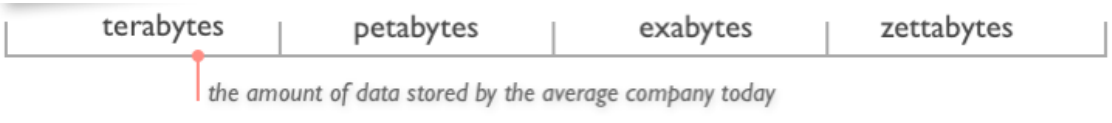
- The costs of computing, storage, and connectivity resources are plunging, and new technologies like scanners, smartphones, ubiquitous video, and other data-collectors mean we are awash in volumes of data that dwarf what was available even five to 10 years ago.

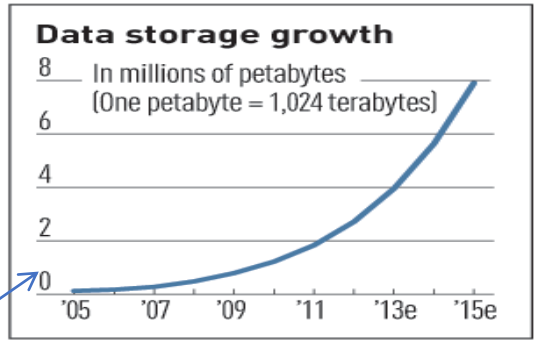
- We capture every mouse click, phone call, text message, Web search, transaction, and more. As the volume of data grows, we can learn more – but only if we uncover meaningful relationships and patterns.

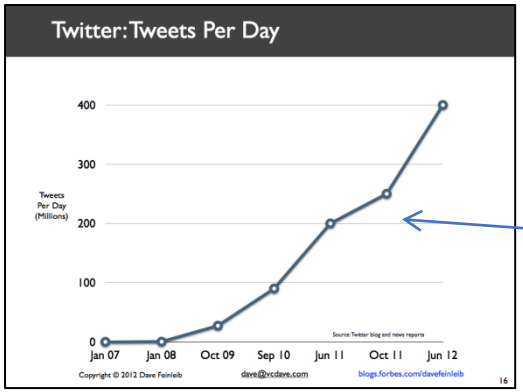
Volume..

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





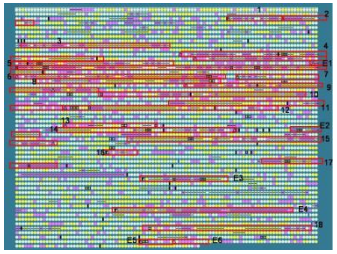


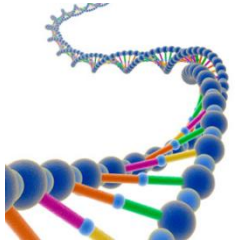


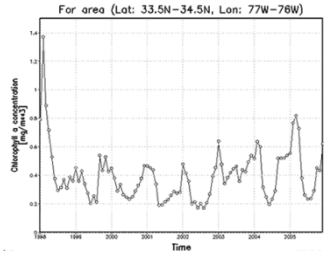
Exponential increase in collected/generated data

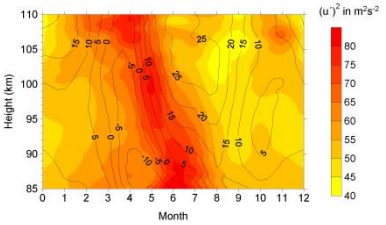
Variety

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









To extract knowledge➔ all these types of data need to linked together

Velocity

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

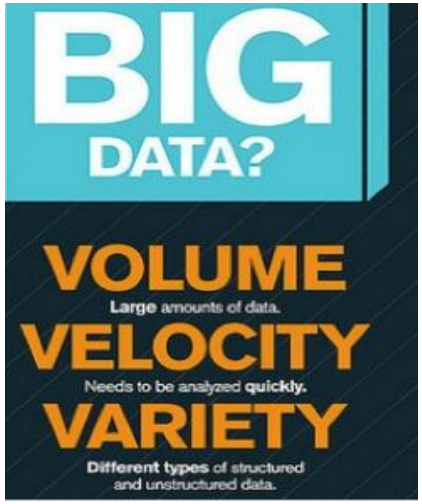
Veracity

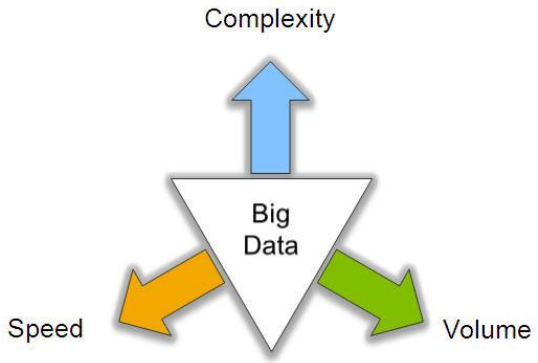
- Veracity indicates the inherent trustworthiness of data.
- The uncertainty about the consistency or completeness of data and other ambiguities can become major obstacles.
- As a result, basic principles as data quality, data cleansing, master data management, and data governance remain critical disciplines when working with Big Data.

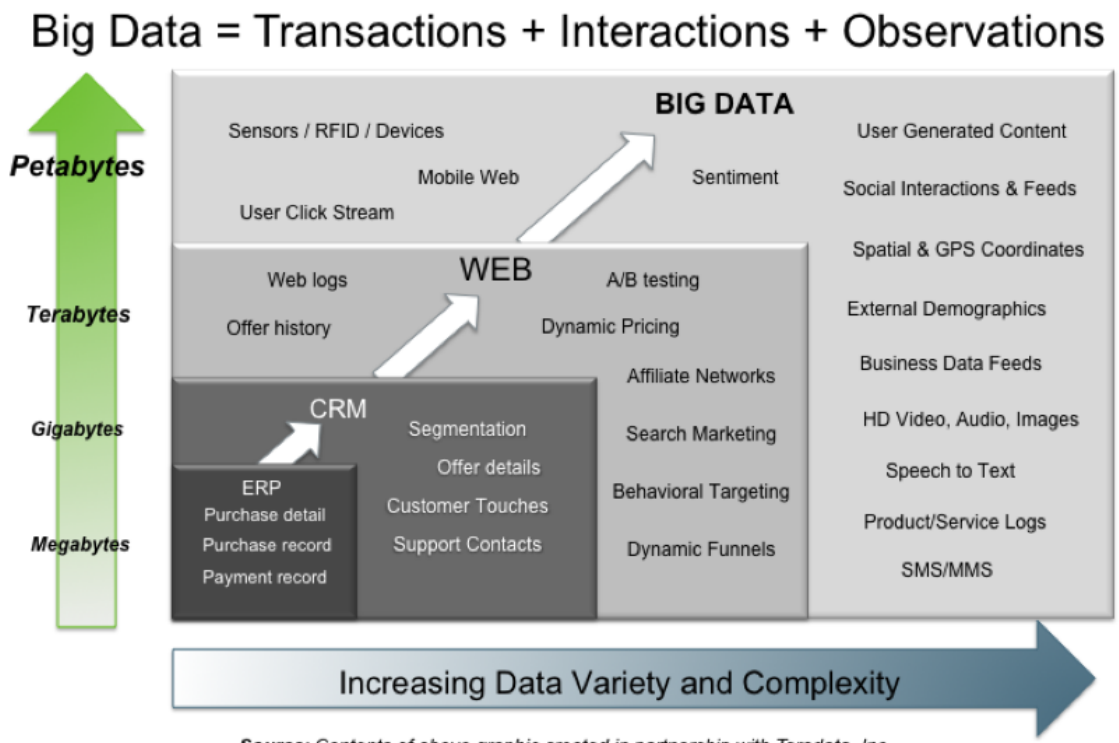
Value

- Access to big data is no good, unless it turned into a good value.
- Return on investment

Big Data: 3V's

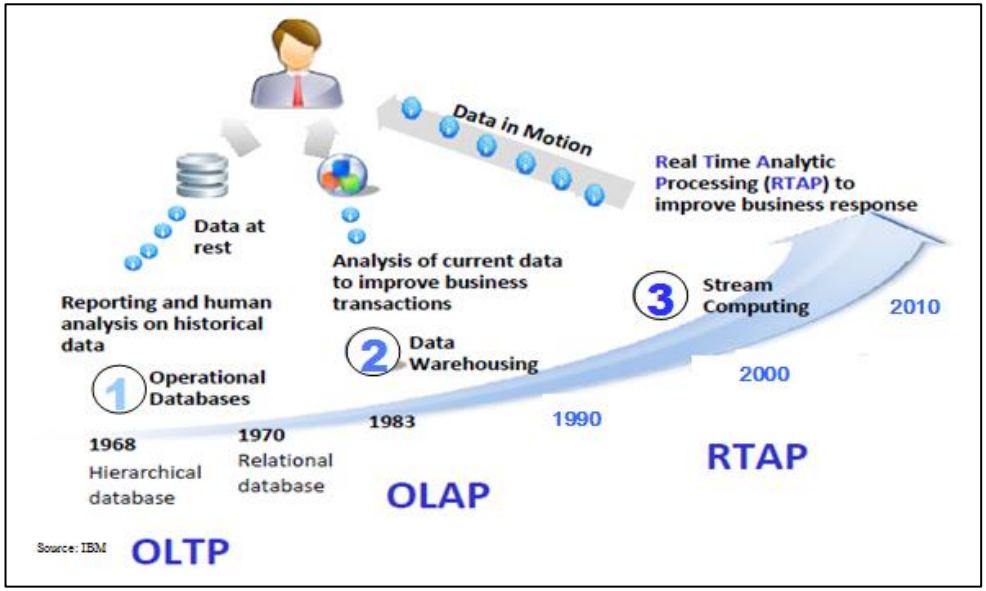






***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Harnessing Big Data



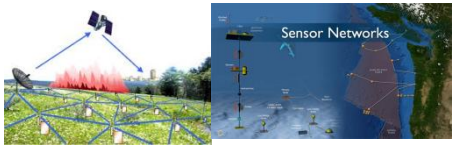
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

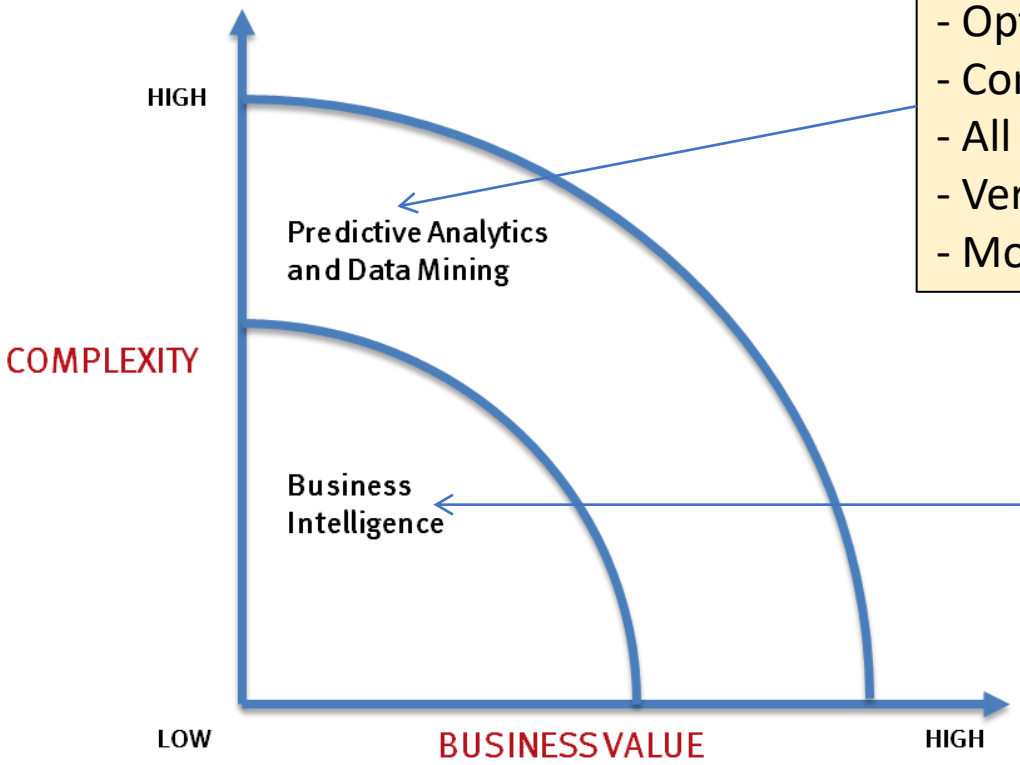
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data

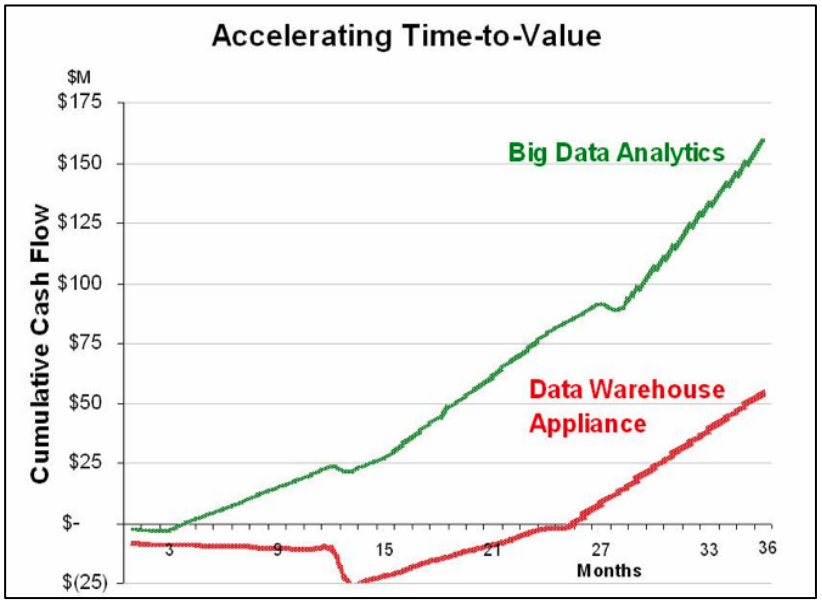


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

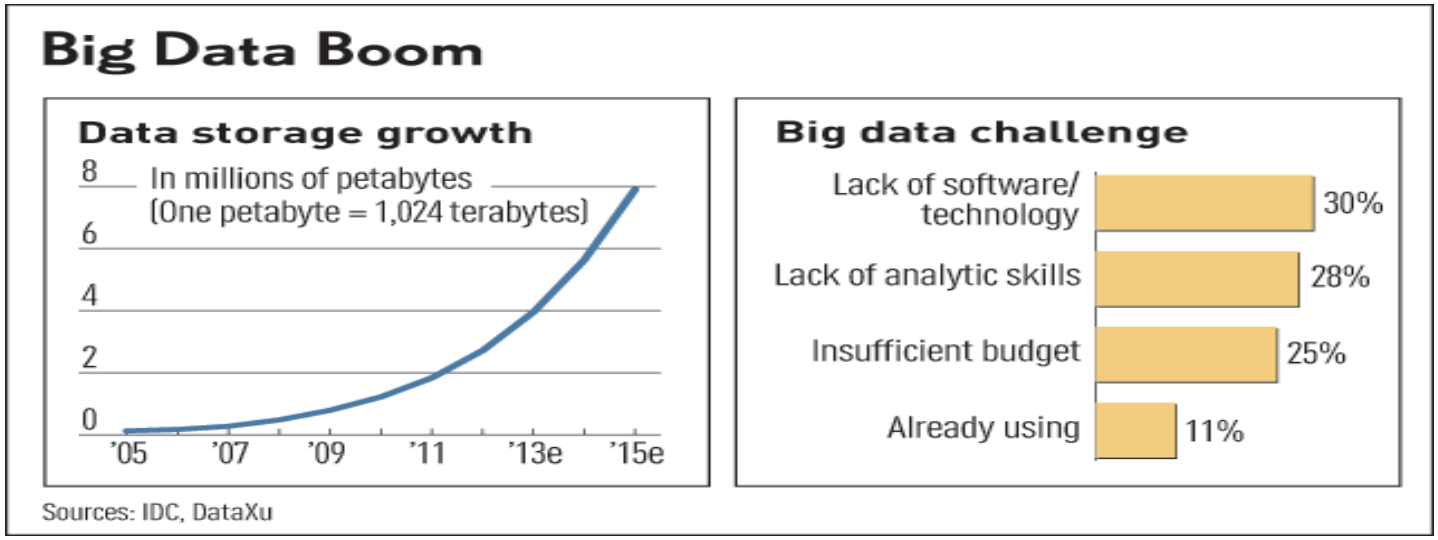
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Data As A Service

factual
kaggle
knoema beta
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Business Intelligence

ORACLE | Hyperion
SAP Business Objects
RJMetrics
Microsoft | Business Intelligence
IBM COGNOS
birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS
TIBCO
KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

COUCHBASE
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

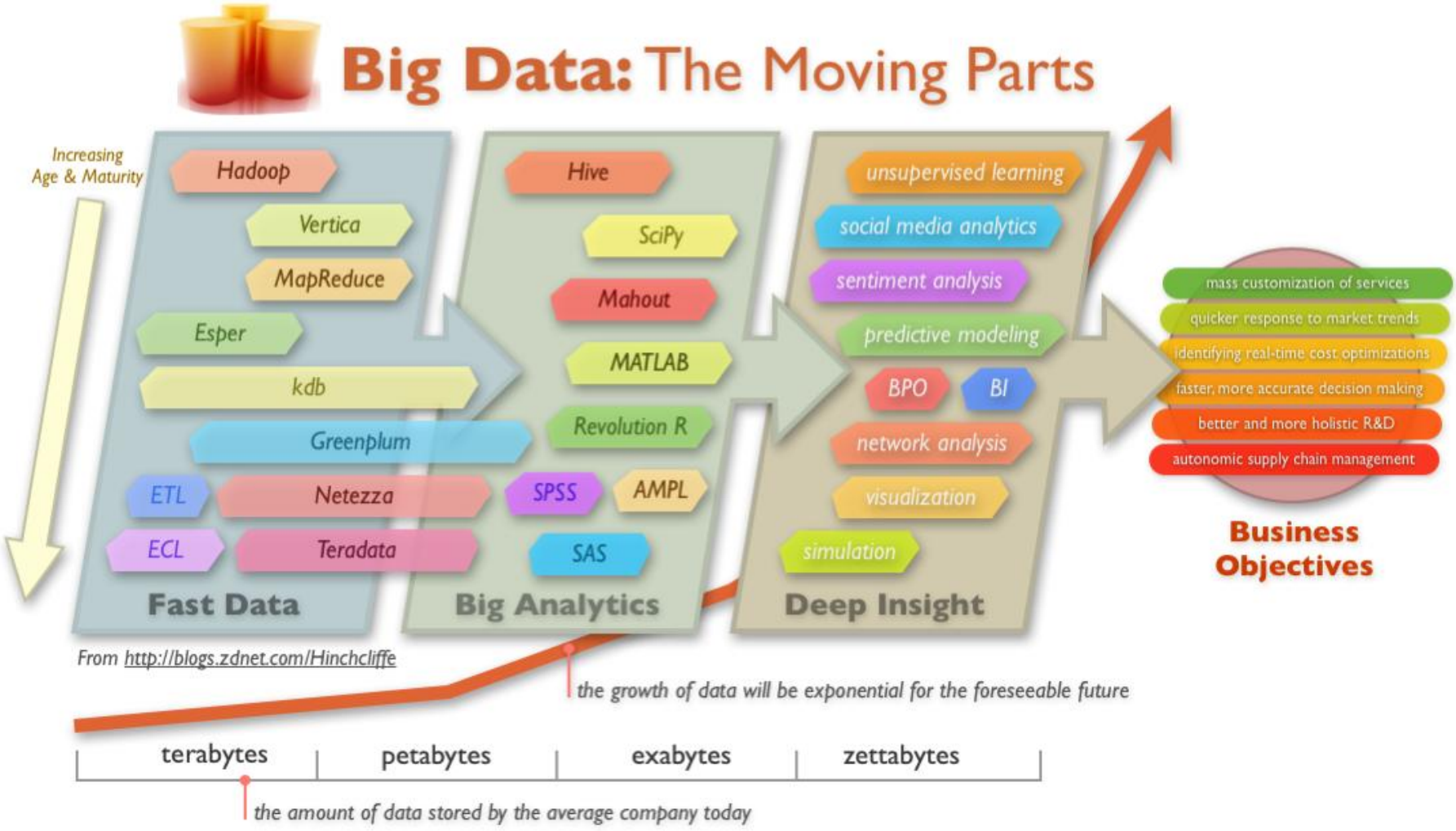
ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com
blogs.forbes.com/davefeinleib

Big Data Technology



About Big Data: Conclusion

- Beginning of big data economy.
- Big data and data science will bring a major social change.
- Companies which fail to exploit big data runs the risk of left behind.
- Should be exploited for sustainable development and equitable society

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of “patterns” and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
  - Bag-of-words
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - summarization association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

End of the 2nd Lecture

Research Issues in Data Mining

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- So far, data mining means
  - Summarization, Association rules, clustering, classification
- Research Issues
  - Finding new patterns
    - Market basket data, Complex data, stream data
  - Improving the performance of existing algorithms
    - Scalable algorithms
      - Data, features or dimensions
    - Complex data
  - Visualizing the patterns

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Presentation Outline

- Data mining functionalities
- Research Issues in Data mining
- **Sources of data**
- Data Objects and Attribute Types
- About Big Data
- Basic Statistical Descriptions of Data
- Data Visualization
- Case study
- Summary

Sources of data

- Primary data
  - Primary data or raw data is **a type of information that is obtained directly from first-hand sources through experiments, surveys, or observations**.

- Examples of primary data
  - Autobiographies and memoirs.
  - Diaries, personal letters, and correspondence.
  - Interviews, surveys, and fieldwork.
  - Internet communications on email, blogs, listservs, and newsgroups.
  - Photographs, drawings, and posters.
  - Works of art and literature.

- Secondary data
  - Secondary data means **data collected by someone else earlier**.

- Examples of secondary data
  - Tax records and social security data.
  - Census data.
  - Electoral statistics.
  - Health records.
  - Books, journals, or other print media.
  - Social media monitoring, internet searches, and other online data.
  - Sales figures or other reports from third-party companies.

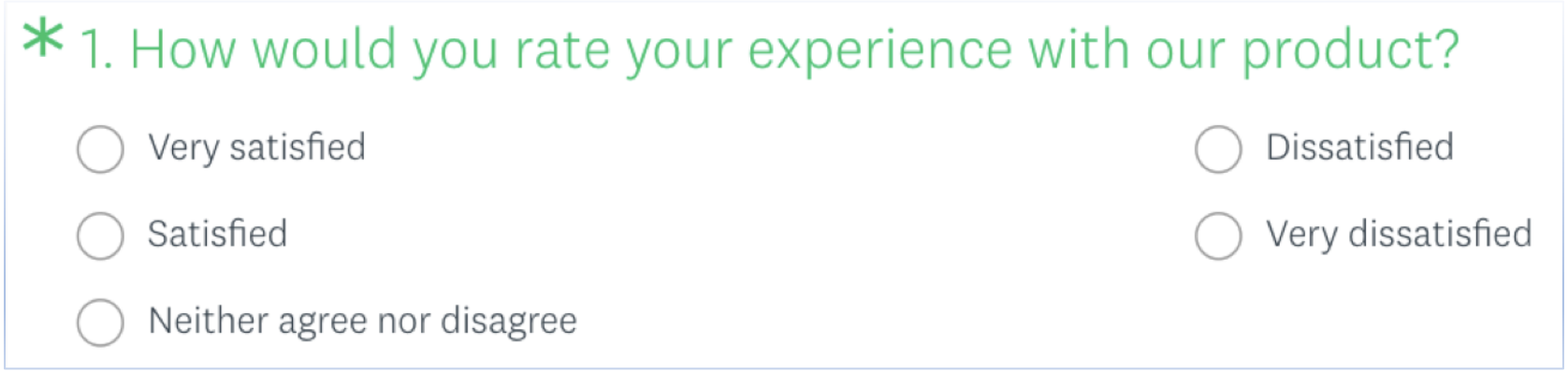
Example of primary data collection

- Quantitative methods
  - Quantitative research is the process of collecting and analyzing numerical data. It can be used to find patterns and averages, make predictions, test causal relationships, and generalize results to wider populations.
- Qualitative methods
  - Examples
  - One-on-one interviews.
    - Interviews are one of the most common qualitative data-collection methods, and they're a great approach when you need to gather highly personalized information
  - Open-ended surveys and questionnaires.
  - Focused groups
  - Observation
  - Case studies

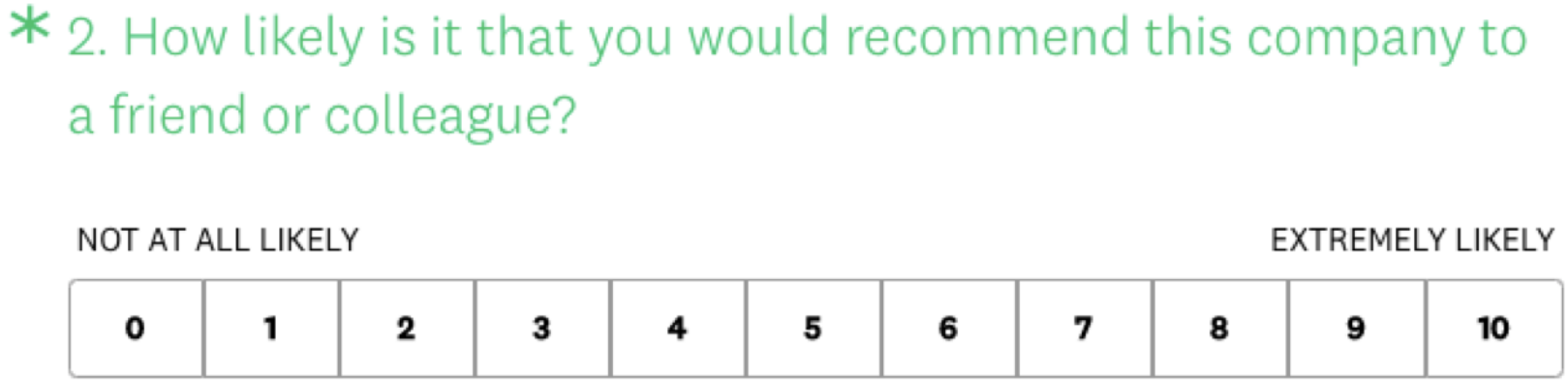
A few sample questions

- Next few slides contain a few sample survey questions

Single-answer multiple choice question.



Rating scales questions



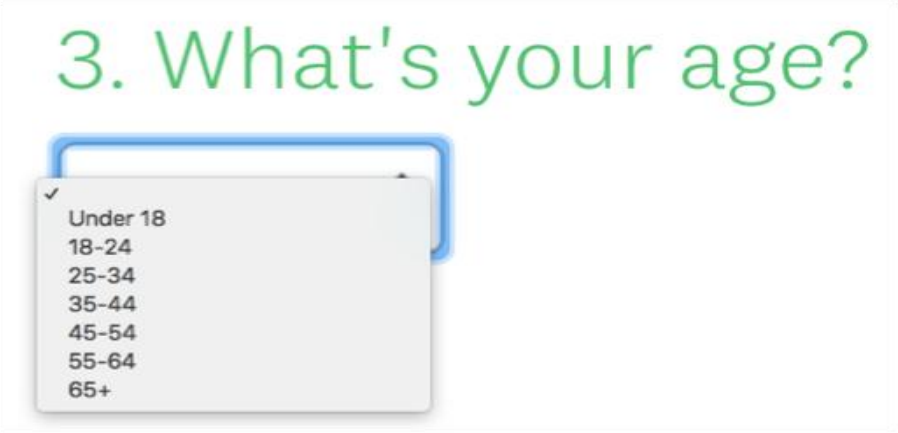
Source: https://www.surveymonkey.com/mp/survey-question-types/

Likert scales: Do you agree or disagree

8. I'm satisfied with the investment my organization makes in education:

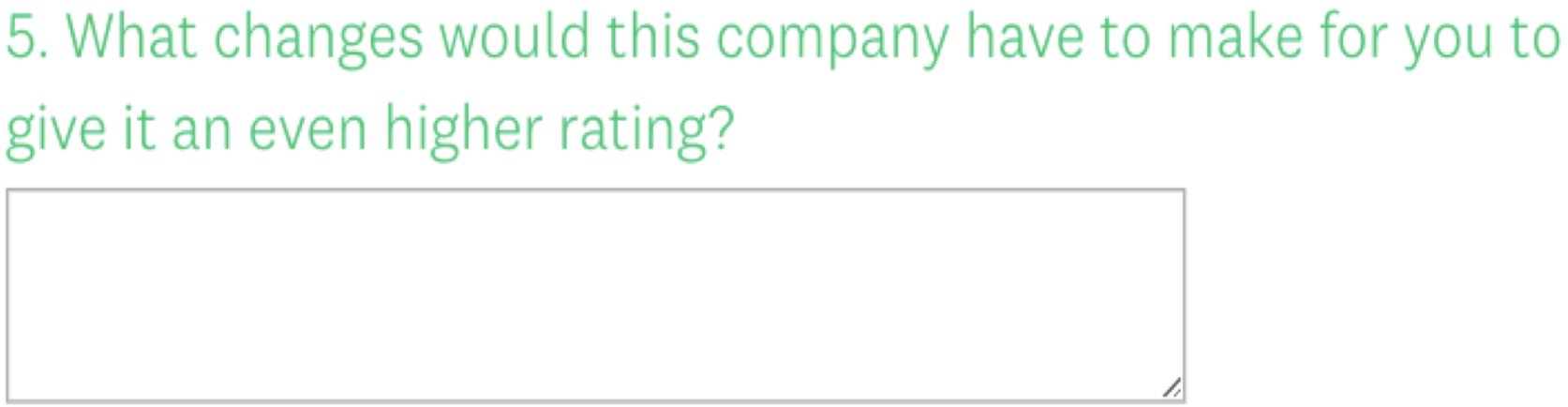
- ◯ Strongly agree
- ◯ Agree
- ◯ Neither agree nor disagree
- ◯ Disagree
- ◯ Strongly disagree

Dropdown questions



Open-ended questions

- Open-ended survey questions require respondents to type their answer into a comment box and don't provide specific pre-set answer options. Responses are then viewed individually or by text analysis tools.



Demographic questions

14. Which of the following best describes your current relationship status?

- ◯ Married
- ◯ Widowed
- ◯ Divorced
- ◯ Separated
- ◯ In a domestic partnership or civil union
- ◯ Single, but cohabiting with a significant other
- ◯ Single, never married

Ranking questions

5. Rank the following shows in order of preference—1 being your favorite and 5 being your least favorite.

- ⠿ ⬍ The office
- ⠿ ⬍ Parks and Recreation
- ⠿ ⬍ Arrested Development
- ⠿ ⬍ Orange is the New Black
- ⠿ ⬍ New Girl

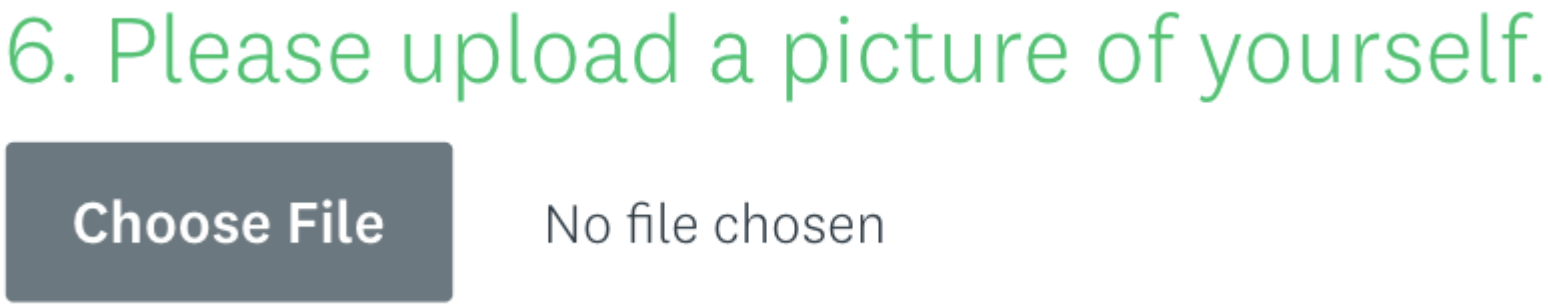
Image choice questions

7. Now that you've reviewed the logos, please pick your favorite.

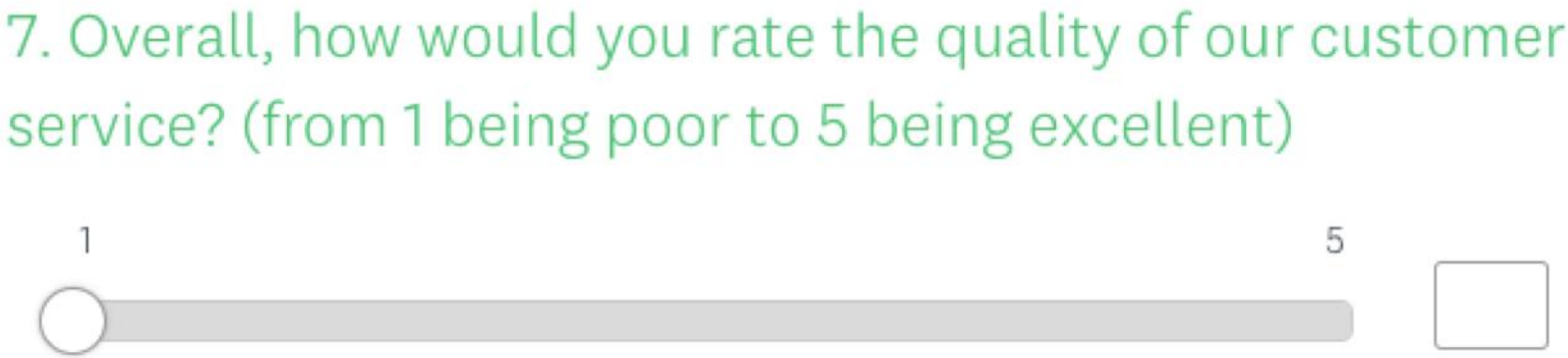
Click map questions

Click the part of the packaging that is the most appealing to you.

File upload questions: Uploading of resume or image



Slider questions



Benchmarkable questions

How likely is it that you would recommend

Acme ▾ to a friend or colleague?

Service Feedback | Product Feedback | Insurance | Brand Research | Show More

Form Number 2802 3739

Census of India 2021 Houselisting and Housing Census Schedule

Confidential when filled

Page No. SIDE- A

State/UT | District | Tahsil/Taluk/ P.S./Dev. Block Circle/Mandal | Town/ Village | Ward Code No. (only for Town) | Houselisting Block No.

Fill if the census house is used wholly or partly as a residence (If '1' or '2' in col. 7)

Information relating to the head of the household (Do not fill columns 12 and 13 for institutional households)

Fill columns 14 to 34 for normal households. Put dash (-) in case of institutional households and non-residential census houses

Amenities available to the household

Assets possessed by the household

2	3	4	5	6	7		8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Building number (Municipal / local authority / census number)	Census house number	Predominant material of floor, wall and roof of the census house (Give code from the respective lists below)			Ascertain use of Census house (Write the actual use and then choose the appropriate code from the list below and record the same in the box at the left hand side of the column)		If '1' or '2' in col. 7, condition of this census house: Good-1 / Livable- 2 / Dilapidated-3	Household number (Give separate serial number to each household and write '999' for every institutional household)	Total number of persons normally residing in this household	Name of the head of the household	Sex: Male-1/ Female-2/ Transgender person-3	If SC* or ST* or Other: SC - 1/ ST - 2 / Other - 3	Ownership status of this house: Owned-1/ Rented but has own house elsewhere-2/ Rented and doesn't own any house-3/ Any other-4	Number of dwelling rooms # exclusively in possession of this household (Record 0, 1,2,3...)	Number of married couple(s) living in this household (Record 0, 1,2,3....)	Main source of drinking water (Give code from the list below)	Availability of drinking water source : $ Within premises-1/ Near the premises-2/ Away-3	Main source of lighting: Electricity-1/ Kerosene-2 Solar-3/ Other oil-4/ Any other-5/ No lighting-6	Access to latrine: Yes: Exclusively for household use only -1/ Shared with other household-2/ Public latrine-3/ No: Open-4	If '1' or '2' in col. 20, then type of latrine (Give code from the list below)	Waste water outlet connected to: Closed drainage-1/Open drainage-2/ No drainage-3	Bathing facility available within the premises: Yes: Bathroom-1/ Enclosure without roof-2/ No-3	Availability of kitchen and LPG/PNG Connection: (Give code from the list below)	Main fuel used for cooking: (If '1' to '6' in col. 24) (Give code from the list below)	Radio/ Transistor: Yes: Traditional radio set-1/ On mobile/smartphone-2/ On any other device-3 /No-4	Television: Yes: Doordarshan free dish-1/ Other DTH/Dish-2/ Cable connection-3/ Any other-4/ No -5	Access to internet: Yes: On laptop/Computer-1/ On mobile/smartphone-2/ On any other device-3/ No-4	Laptop/ Computer: Yes: 1/ No-2	Telephone and Mobile phone/Smartphone (Give code from the list below)	Bicycle and Scooter/ Motorcycle/ Moped (Give code from list below)	Car/ Jeep/ Van: Yes: 1/ No-2	Main cereal consumed in the household: Rice-1/ Wheat-2/ Jowar-3/ Bajra-4/ Maize-5/ Any other-6	Mobile Number (For Census related communications only)
		Floor	Wall	Roof	Code No.	Actual use																										

Presentation Outline

- Data mining functionalities
- Research Issues in Data mining
- Sources of data
- **Data Objects and Attribute Types**
- Basic Statistical Descriptions of Data
- Data Visualization
- Case study
- Summary

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples, records, vector, patten, event, case, observation or entity*
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Types of Data Sets

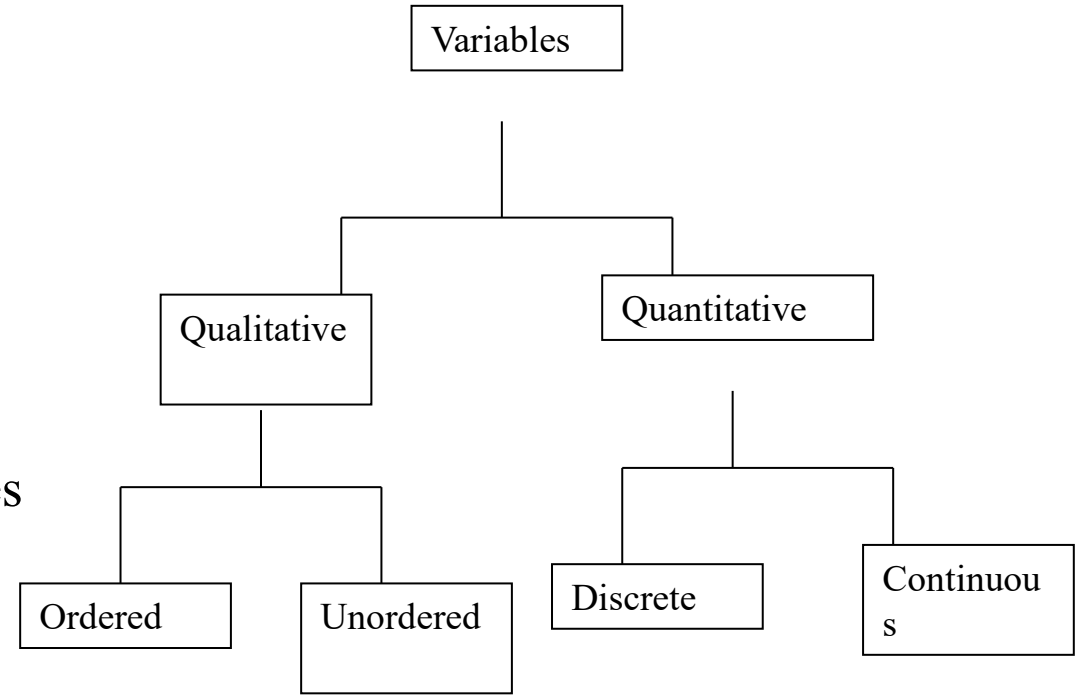
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video 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

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

Attributes

- **Attribute (or dimensions, features, variables):**
  - a data field, representing a characteristic or feature of a data object.
    - *E.g., customer _ID, name, address*
- **A measurement scale:**
  - Associates a numerical or symbolic value with an attribute of an object.
- Type of attribute: Type of measurement of scale
- Attribute Types:
  - Qualitative/categorical attributes
    - Nominal attributes
    - Binary attributes
    - Ordinal attributes
  - Quantitative attributes or numeric attributes
    - Interval-scaled attributes
    - Ratio-scaled attributes



Qualitative Attributes

- **Nominal attributes :** categories, states, or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
  - **No order**
  - **Mean/median can not be calculated. Mode is the option.**
- **Binary attributes:** Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary attribute: both outcomes are equally important
    - e.g., gender
  - Asymmetric binary attribute: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

- **Ordinal**
  - Values have a meaningful order (ranking), but the magnitude between successive values is unknown.
  - *Example*
    - *Size = {small, medium, large},* grades, army rankings
    - Customer satisfaction: 0: very satisfied, 1: some dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied
    - Central tendency: Mode or median is the measure
  - New value=f(old value)

Quantitative Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point. New value= a*old value+b
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., the *temperature in Kelvin, length, counts, monetary quantities*
    - *New value= a*old-value*

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
  - Note: Binary attributes are a special case of discrete attributes
- **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

Presentation Outline

- Data mining functionalities
- Research Issues in Data mining
- Sources of data
- Data Objects and Attribute Types
- **Basic Statistical Descriptions of Data**
- Data Visualization
- Summary

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Basic Statistical Descriptions of Data…

- Measures of Central Tendency (middle or center of the distribution)
  - Mean, Median, Mode, Geometric Mean, Harmonic Mean
- Variability measures or measuring the dispersion. How the data is spread out?
  - Range, Quartile Deviation, Mean Deviation, Standard Deviation, Coefficient of Variation
- Skewness
- Constructing a boxplot

Measures of Central Tendency

- A group of data is represented with a single number
- It brings very important information from it

Arithmetic Mean

- Arithmetic mean: Sum of observations divided by its number.

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

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:

$$\mu = \frac{\sum x}{N}$$

  - Trimmed mean: chopping extreme values

- It should be computable.
  - Mean from categorical data is not calculable

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

- Advantages: Can carry out algebraic manipulations.
- Demerits:
  - Gives more weight to extreme items (outliers).

Median

- Median: Middle value if odd number of values, or average of the middle two values otherwise
- Merits
  - It can be calculated even if extreme classes are not defined.
- Property
  - Sum of absolute deviations from is least when it is taken from median.
- Demerits
  - Not based on all observations
  - Not widely used in practice

Mode

- Mode: It is the value which occurs most frequently.
- Merits
  - It can be easily located
  - It can be calculated when extreme classes are not defined
  - It is used widely in the business.
  - It can be calculated easily except if maximum frequency occurs more than once.
- Demerits
  - Not based on the all observations
  - It is not having algebraic properties.
  - It is not stable
    - Different class intervals results into different mode value.
- The relationship between mean, median and mode
  - Mean - mode=3(mean-median)

Selecting among the mean, median and mode

- Common mistake: specifying the wrong index of central tendency
  - It is common to specify the mean
- If data is categorical, if yes use “mode”.
  - If no
    - If the total is of interest, use “mean”
    - If no
      - If the distribution is skewed, use “median”
      - Otherwise, use “mean”

- Type of micro processor → use mode
- Total CPU time → mean

- Skew: ratio of minimum and maximum is large, the data is skewed.

Summarizing the Variability

- It is a measure that can give the widespread or scattering of observations among themselves or from a center point.
  - Range, Quartile deviation, Mean deviation, and standard deviation (variance)
- Range:
  - The difference between the highest and lowest values in a series of observations.
  - Problem: it depends on two extreme values.
- Quartile deviation
  - Quartile deviation (Q.D)=$(Q_3-Q_1)/2$, where Q1 is first quartile, Q3 is third quartile. The first and third quartiles are called lower and upper quartiles, respectively.
  - First quartile: The value of the variate below which one-fourth of the values lie and above which three-fourths of the values lie, when the values are arranged in ascending order of magnitude.
  - Third quartile: The value of the variate below which three-fourths of the values lie and above which the remaining one-fourth of the values lie, when the values are arranged in ascending order of magnitude.
  - Merits:
    - The presence of abnormal values does not affect quartile deviation.

Mean deviation, Standard Deviation, Variance

- Mean Deviation: The mean deviation is the mean of the absolute values of the deviations taken from the average.

- Standard deviation: It is defined as the square root of the mean of the squares of the deviations taken from the arithmetic mean.

  - σ =square root($1/n \sum (x1 - X)^{2)}$ where X is arithmetic mean

- Variance: Square of standard deviation.

Coefficient of Variation

- CV is the percentage ratio of S.D to mean.
- CV= (SD/Mean)*100

- For a player scores: if CV is high he/she is inconsistent,

Statistical Population

- Population: All the values/objects
- Sample: A part of population.
  - A sample should be a representative of population
  - Should be taken in a random manner.
  - Population is specified with parameters
  - Sample is specified with statistics
- Population error: Difference between the sample mean and population mean.

Covariance and correlation analysis

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

  where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1; similarly for μ_2

- Sample covariance between X_1 and X_2: $\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \widehat{\mu_1})(x_{i2} - \widehat{\mu_2})$
- Sample covariance is a generalization of the sample variance:

$$\hat{\sigma}_{11} = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \widehat{\mu_1})(x_{i1} - \widehat{\mu_1})$$

- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance**: If $\sigma_{12} < 0$

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  - $E(X_1) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
  - $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14)/5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

- **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

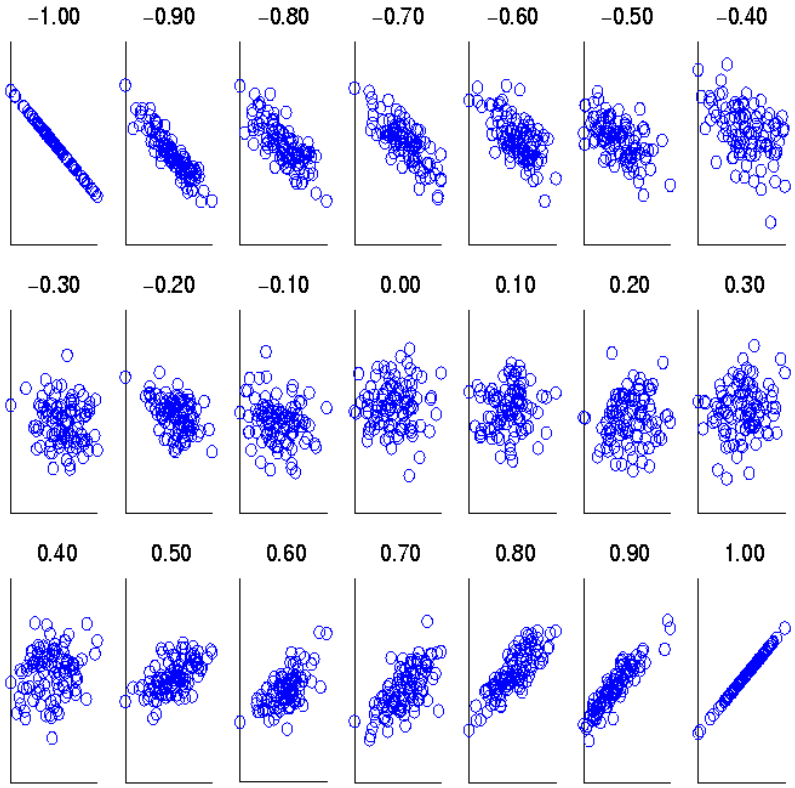
$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

- **Sample correlation** for two attributes X_1 and X_2: $\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$

  where n is the number of tuples, μ_1 and μ_2 are the respective means of X_1 and X_2 , σ_1 and σ_2 are the respective standard deviation of X_1 and X_2

- If $\rho_{12} > 0$: A and B are positively correlated (X_1's values increase as X_2's)
  - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: [–1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1

Correlation Analysis (for Categorical Data)

- **X^2 (chi-square) test:**

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

(observed: O_i; expected: E_i)

- Oi= Observed value, Ei= Expected value
- Null hypothesis: The two distributions are independent
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
  - The larger the X^2 value, the more likely the variables are related
- Note: Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

χ^2 correlation test for nominal data

For nominal data, a correlation relationship between two attributes, A and B, can be discovered by a χ^2 (**chi-square**) test. Suppose A has c distinct values, namely, $a_1, a_2, \ldots a_c$, and B has r distinct values, namely, $b_1, b_2, \ldots b_r$. The data tuples described by A and B can be shown as a **contingency table**, with the c values of A making up the columns and the r values of B making up the rows. Let (A_i, B_j) denote the joint event that attribute A takes on value a_i and attribute B takes on value b_j, that is, where $(A = a_i, B = b_j)$. Each and every possible (A_i, B_j) joint event has its own cell (or slot) in the table. The χ^2 value (also known as the *Pearson* χ^2 *statistic*) is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{2.10}$$

where o_{ij} is the *observed frequency* (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the *expected frequency* of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}, \tag{2.11}$$

How to use Chisquare test?

- Use the formula to compute the chi-square value.
- Find critical value using the table (use p=0.05)
- df (degrees of freedom=n-1)
- If chi-square< critical value, accept the null hypothesis, i.e., differences are due to chance
- If chi-square value >critical value, differences in the data are not due to chance, i.e., two variables are correlated.

Using Chi-square test

Two columns: <Gender, preferred reading>

Table 2.2 Example 2.1's 2 × 2 contingency table data.

	Male	Female	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are gender and preferred_reading correlated?

Using Eq. (2.11), we can verify the expected frequencies for each cell. For example, the expected frequency for the cell (*male, fiction*) is

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90,$$

Using Eq. (2.10) for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

For this 2×2 table, the degrees of freedom are $(2-1) \times (2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution, typically available from any textbook on statistics).

We reject the hypothesis that Gender and Preferred reading is independent

Chi-square distribution table

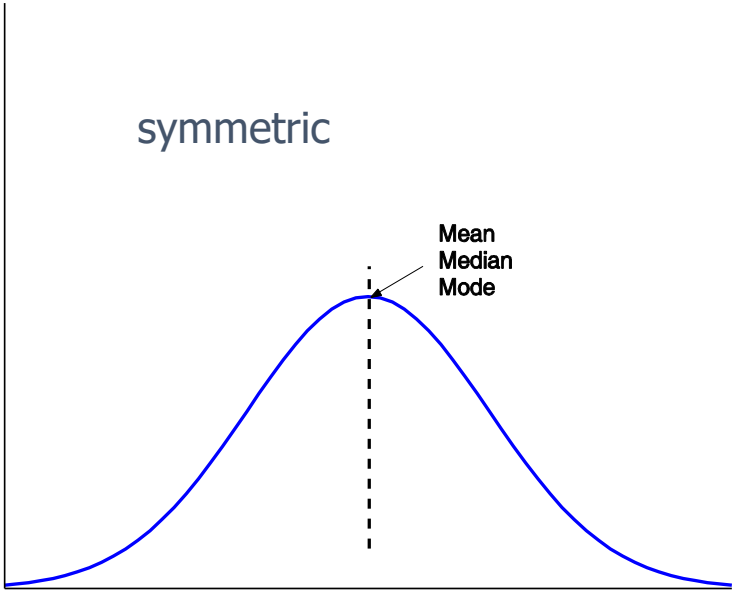
	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

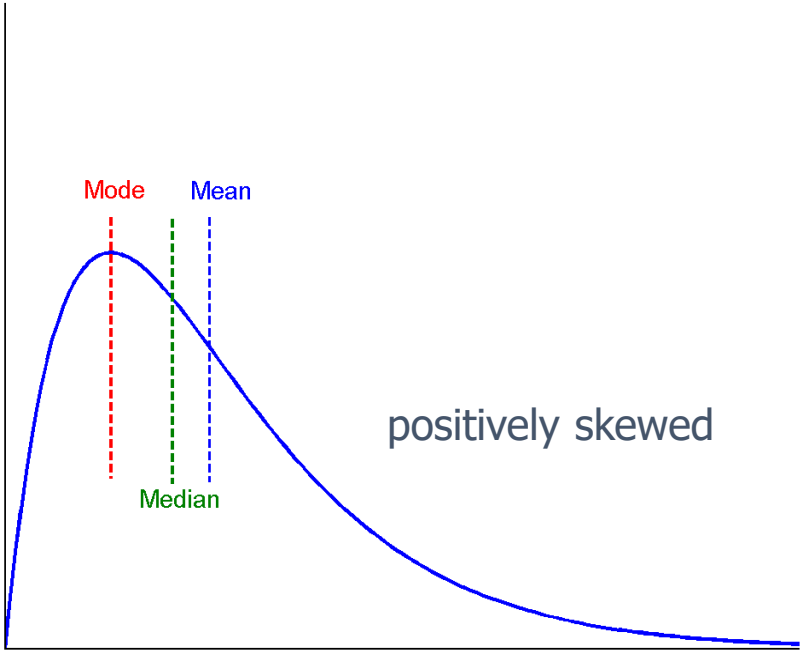
Determining the Distribution of Data

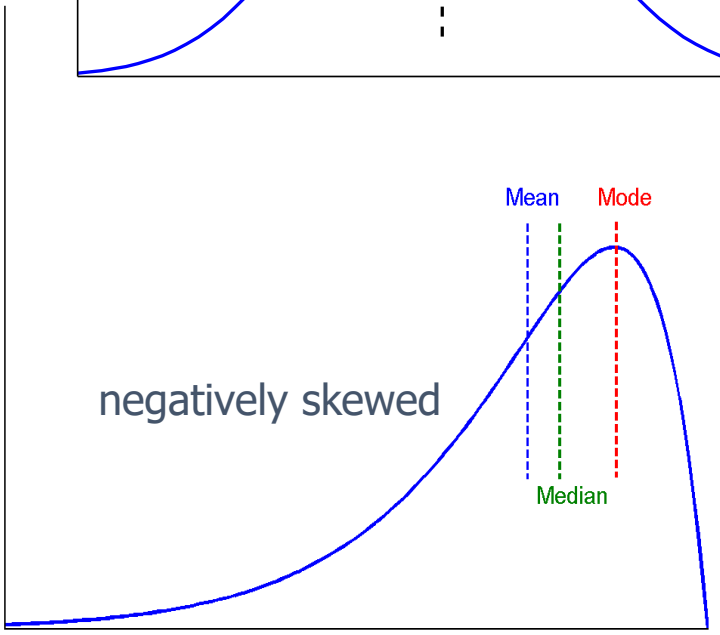
- The simplest way is to plot the histogram of data
  - Requires dividing the range into a number of sub-ranges called cells or buckets.
- The count of observations that fall into each cell are determined.
- The counts are normalized to frequencies by dividing by the total number of observations. The cell frequencies are plotted as a column chart
- Key problem in determining the cell size.
  - Small cells lead to very few observations per cell and large variations in the number of observations.
  - Large cells result in fewer variations, but the details of observations are completely lost.
  - Guideline: if a cell has fewer than five observations, the cell size should be increased, or a variable cell histogram should be used.

Symmetric vs. Skewed Data

- Even if two measures mean and standard deviations are same for the distributions, still the shape of two curves may differ.







Measures of Skewness

- Pearson's Coefficient of skewness
  - (Mean-mode)/SD
- Quartile coefficient of skewness
  - ((Q3-Q2)-(Q2-Q1))/(Q3-Q1)

Selecting Index of Dispersion

- If the variable is bounded, use “range”.
- If there are no natural bounds, and the distribution is symmetric use either SD, variance or CV.
- If the distribution is non-symmetric, percentiles are best choices.

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Box Plot
or The 5 Number Summary

- A **box plot** or **boxplot** (also known as a **box-and-whisker diagram** or **plot**) is a convenient way of graphically depicting groups of numerical data through their five-number summaries:
  - the smallest observation (sample minimum),
  - lower quartile (Q1),
  - median (Q2),
  - upper quartile (Q3), and
  - largest observation (sample maximum).
  - A boxplot may also indicate which observations, if any, might be considered outliers.
    - usually, a value higher/lower than 1.5 x IQR

Constructing a box and whisker plot

- Step 1 - Find the median.
- Remember, the median is the middle value in a data set.

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100

68 is the median of this data set.

Constructing a box and whisker plot

- Step 2 – Find the lower quartile.
- The lower quartile is the median of the data set to the left of 68.

(18, 27, 34, 52, 54, 59, 61,) 68, 78, 82, 85, 87, 91, 93, 100

52 is the lower quartile

Constructing a box and whisker plot

- Step 3 – Find the upper quartile.
- The upper quartile is the median of the data set to the right of 68.

18, 27, 34, 52, 54, 59, 61, 68, (78, 82, 85, 87, 91, 93, 100)

87 is the upper quartile

Constructing a box and whisker plot

- Step 4 – Find the maximum and minimum values in the set.
- The maximum is the greatest value in the data set.
- The minimum is the least value in the data set.

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100

18 is the minimum and 100 is the maximum.

Constructing a box and whisker plot

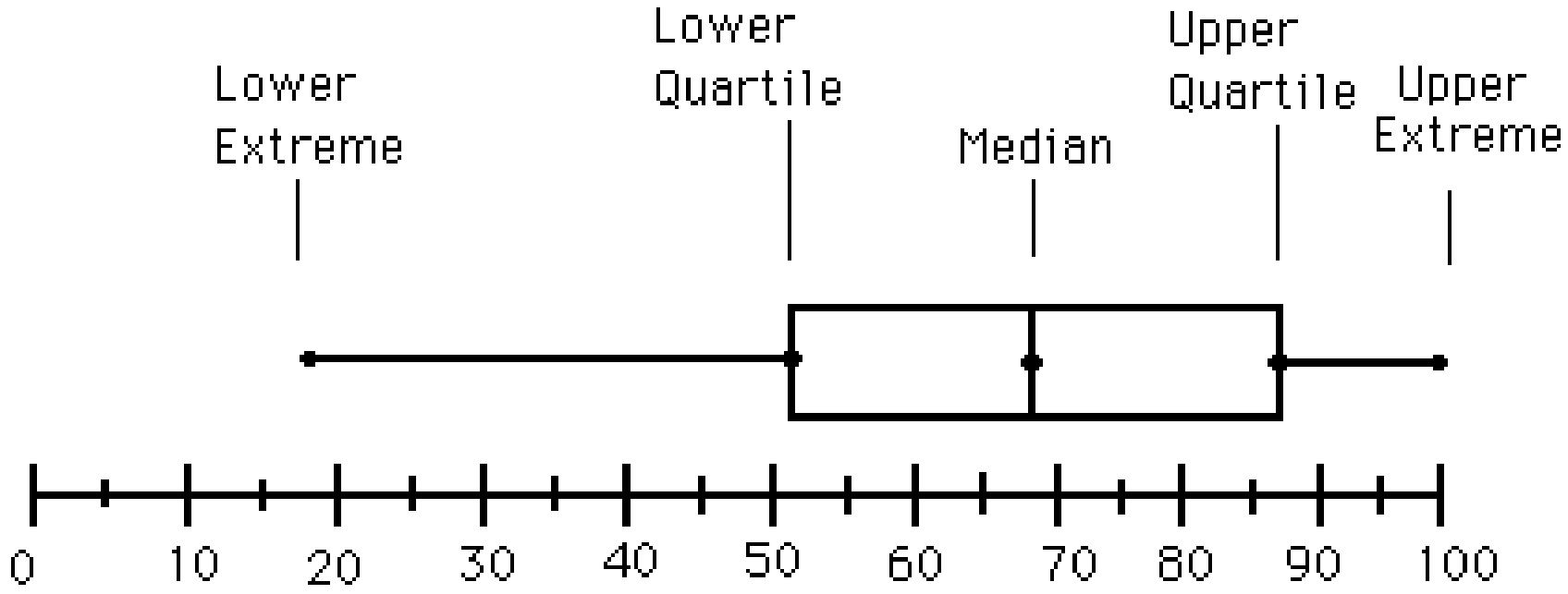
- Step 5 – Find the inter-quartile range (IQR).
- The inter-quartile (IQR) range is the difference between the upper and lower quartiles.
  - Upper Quartile = 87
  - Lower Quartile = 52
  - 87 – 52 = 35
  - 35 = IQR

The 5 Number Summary

- Organize the 5 number summary
  - Median – 68
  - Lower Quartile – 52
  - Upper Quartile – 87
  - Max – 100
  - Min – 18

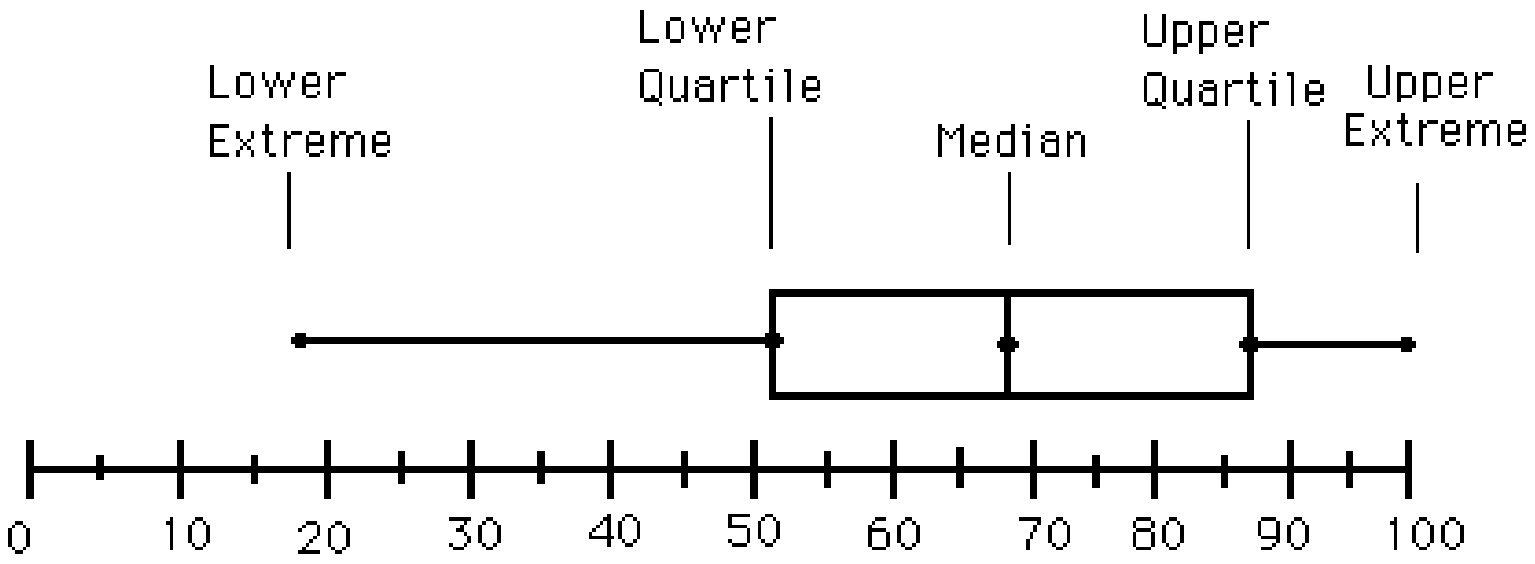
Graphing The Data

- Notice, the Box includes the lower quartile, median, and upper quartile.
- The Whiskers extend from the Box to the max and min.

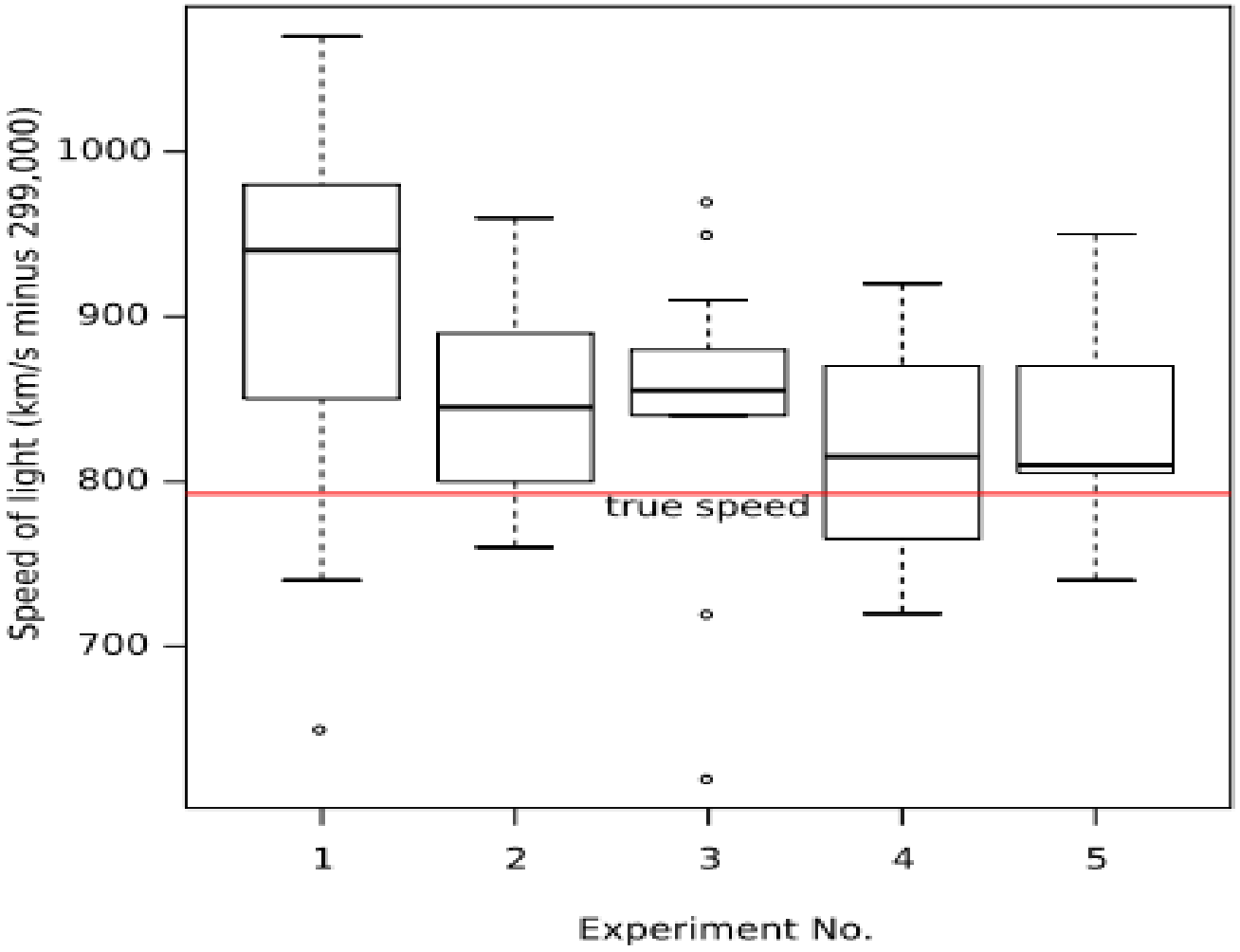


Analyzing The Graph

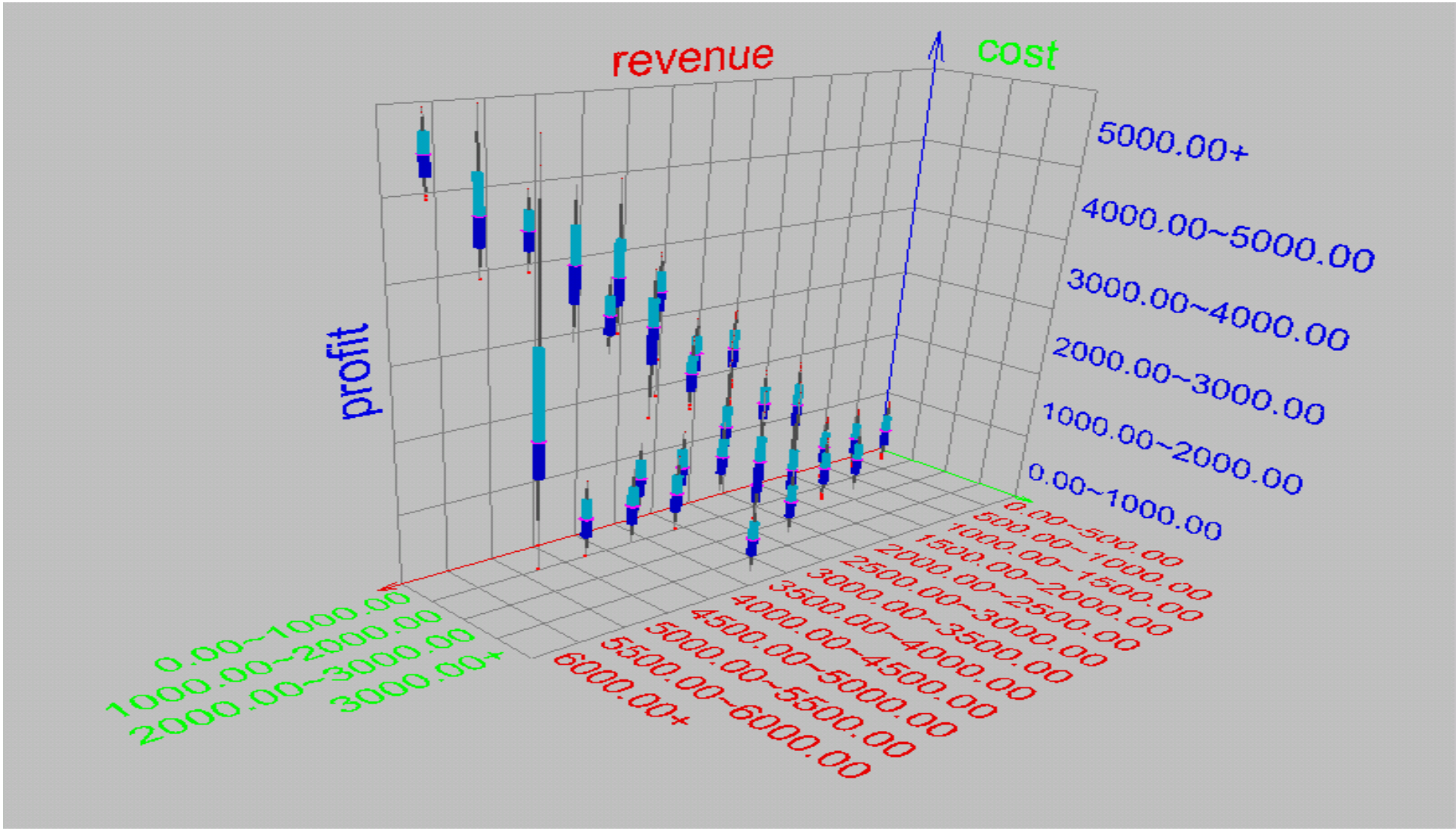
- The data values found inside the box represent the middle half (50%) of the data.
- The line segment inside the box represents the median



Example Box Plot

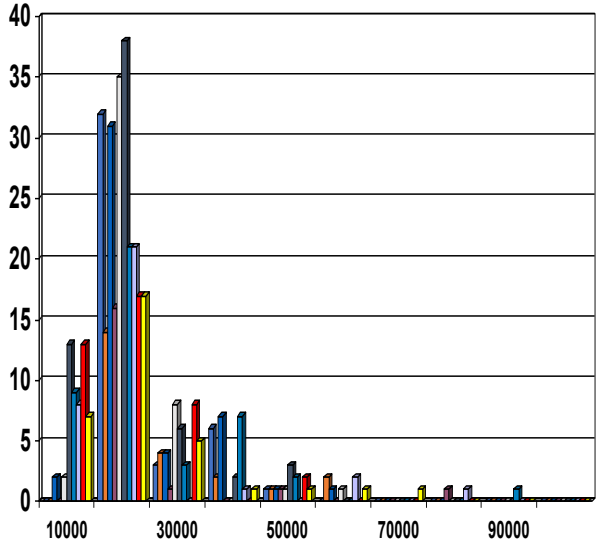


Visualization of Data Dispersion: 3-D Boxplots



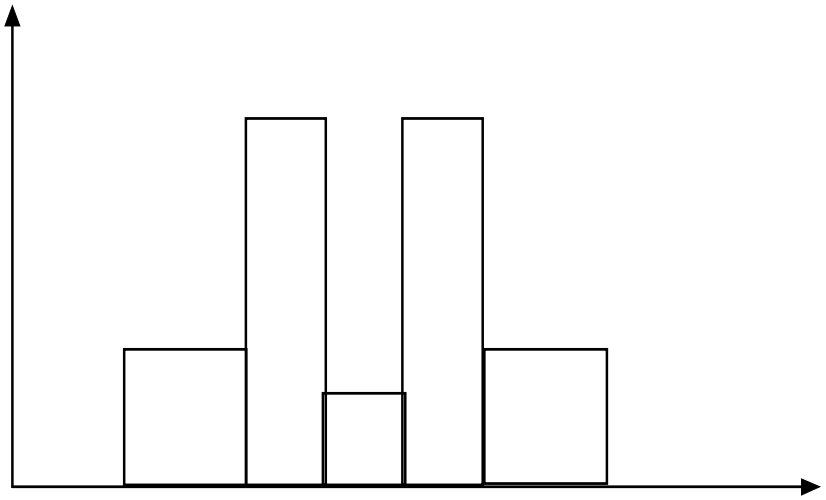
Histogram Analysis

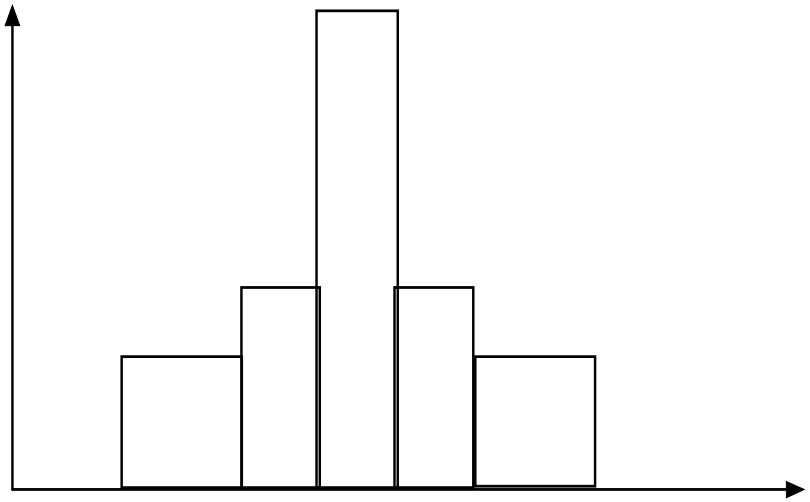
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart
  - Bar chart uses a set of bars (often separated with space) with *X* representing a set of categorical data, such as *automobile_model* or *item_type*, and the height of the bar (column) indicates the size of the group defined by the categories



- On the other hand, histogram plots quantitative data with a range of *X* values grouped into bins or intervals.
- Histograms are used to show distributions (along *X* axis) while bar charts are used to compare categories

Histograms Often Tell More than Boxplots

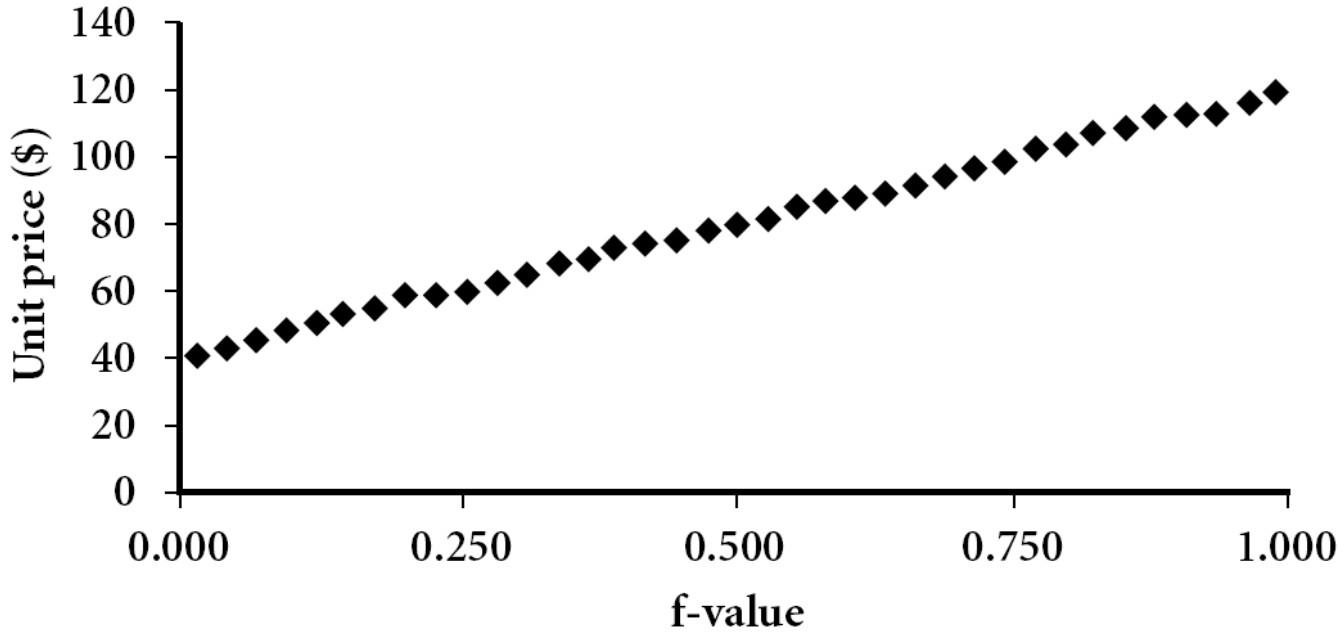




- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

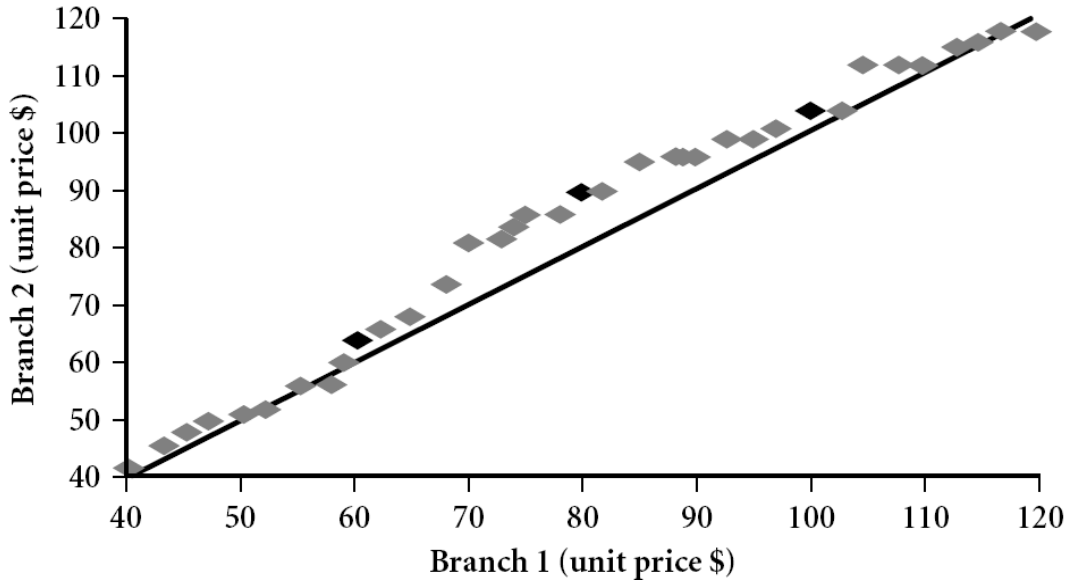
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



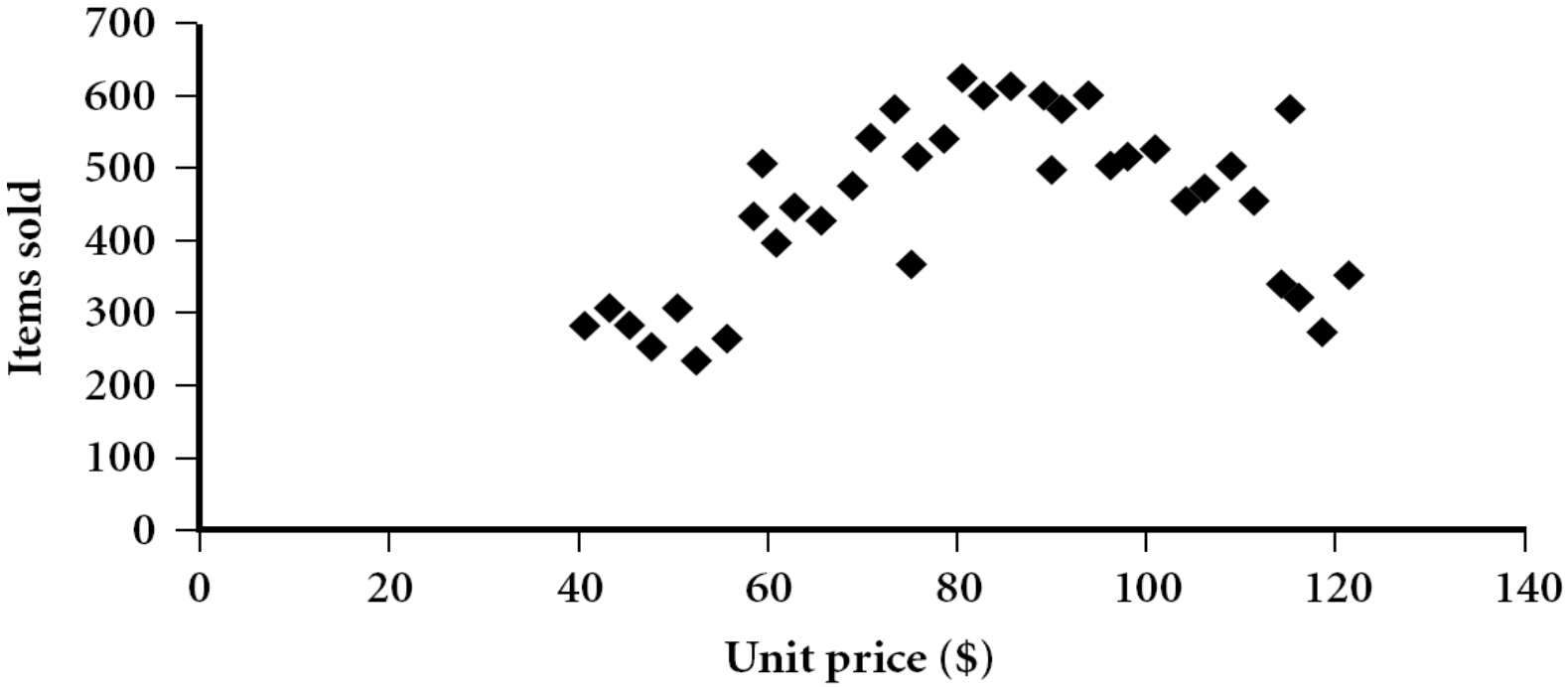
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- 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.

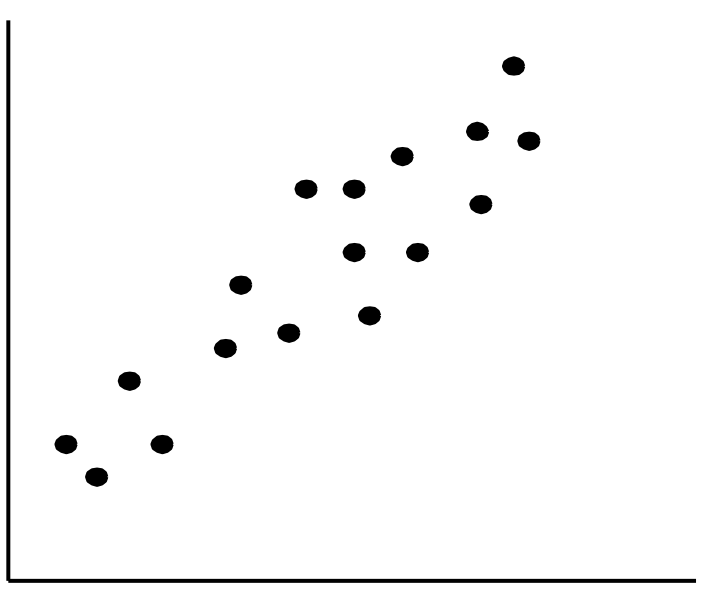


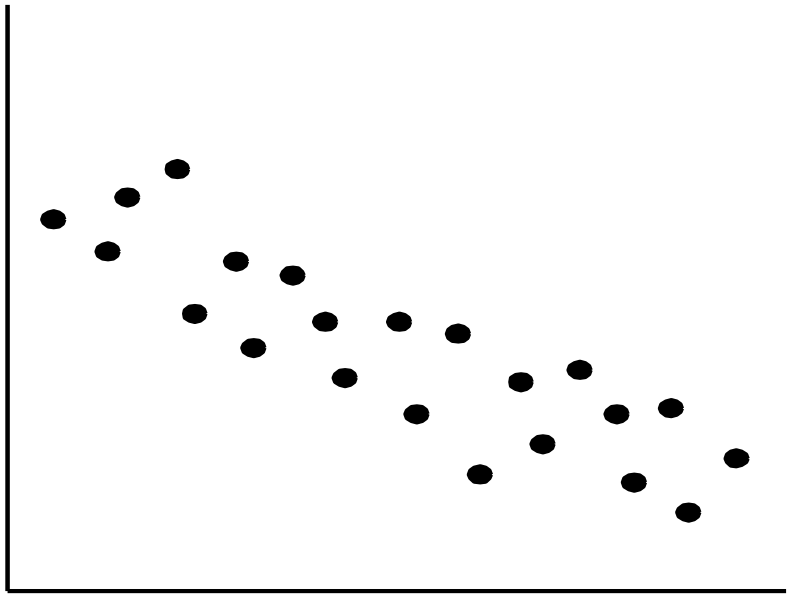
Scatter plot

- 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



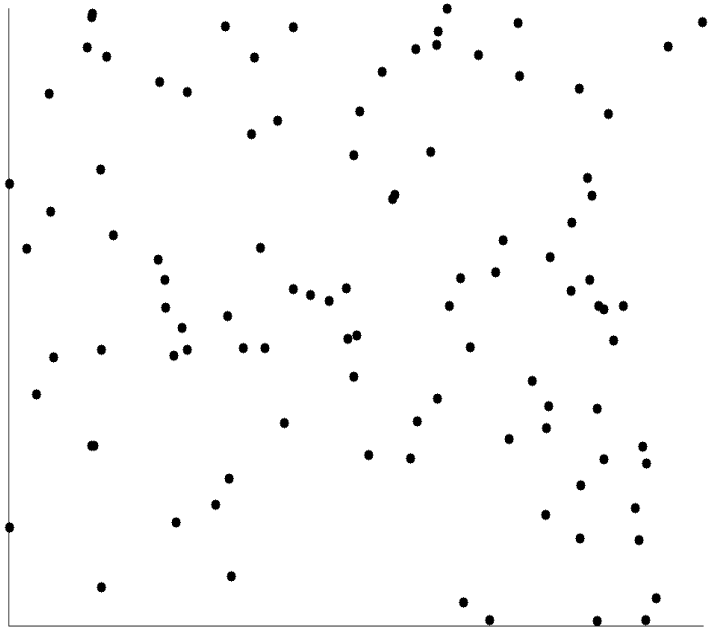
Positively and Negatively Correlated Data

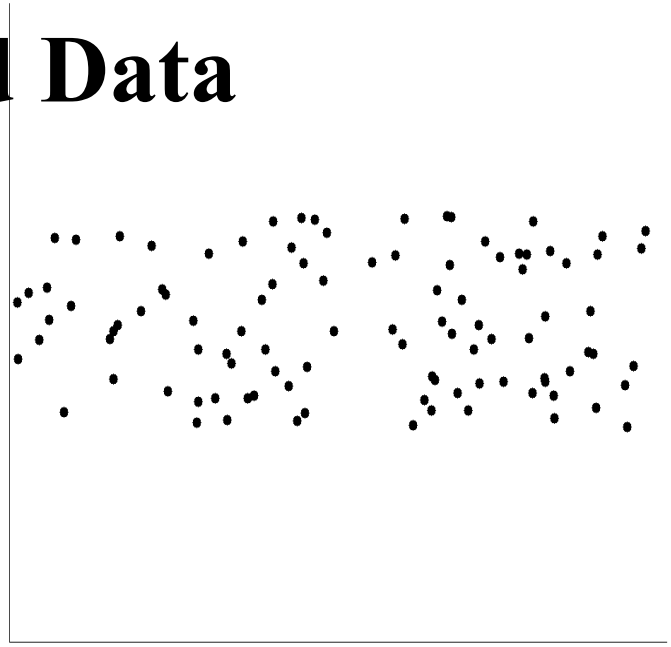


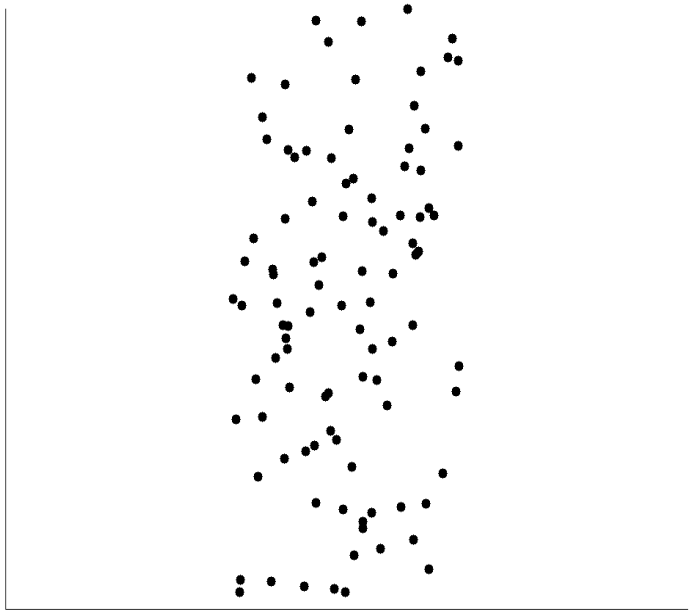


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







Presentation Outline

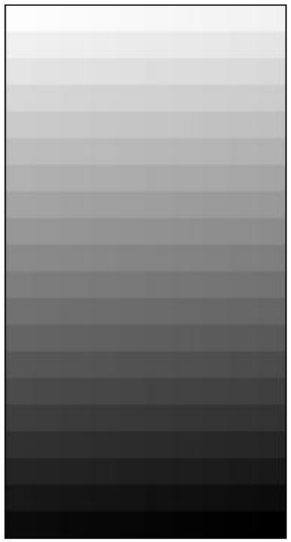
- Data mining functionalities
- Research Issues in Data mining
- Sources of data
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- **Data Visualization**
- Case study
- Summary

Data Visualization

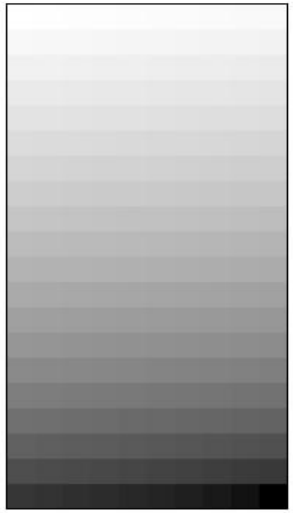
- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

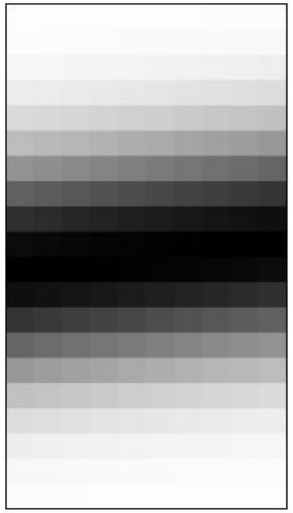
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values
- Example: Income ascending order of customer information table



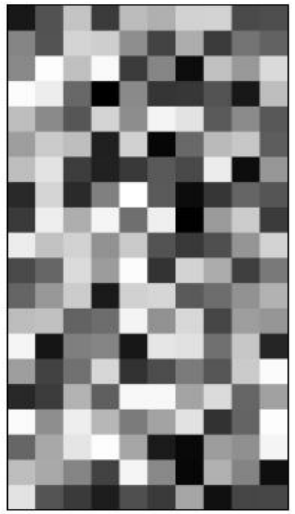
(a) Income



(b) Credit Limit



(c) transaction volume



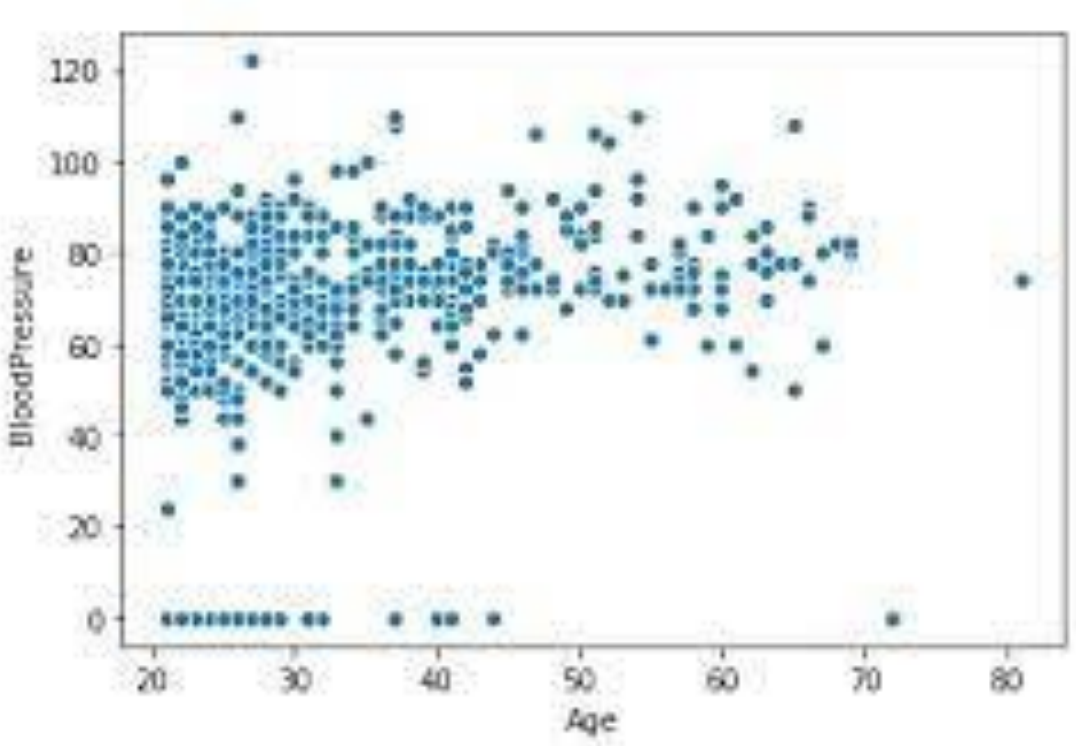
(d) age

Geometric Projection Visualization Techniques

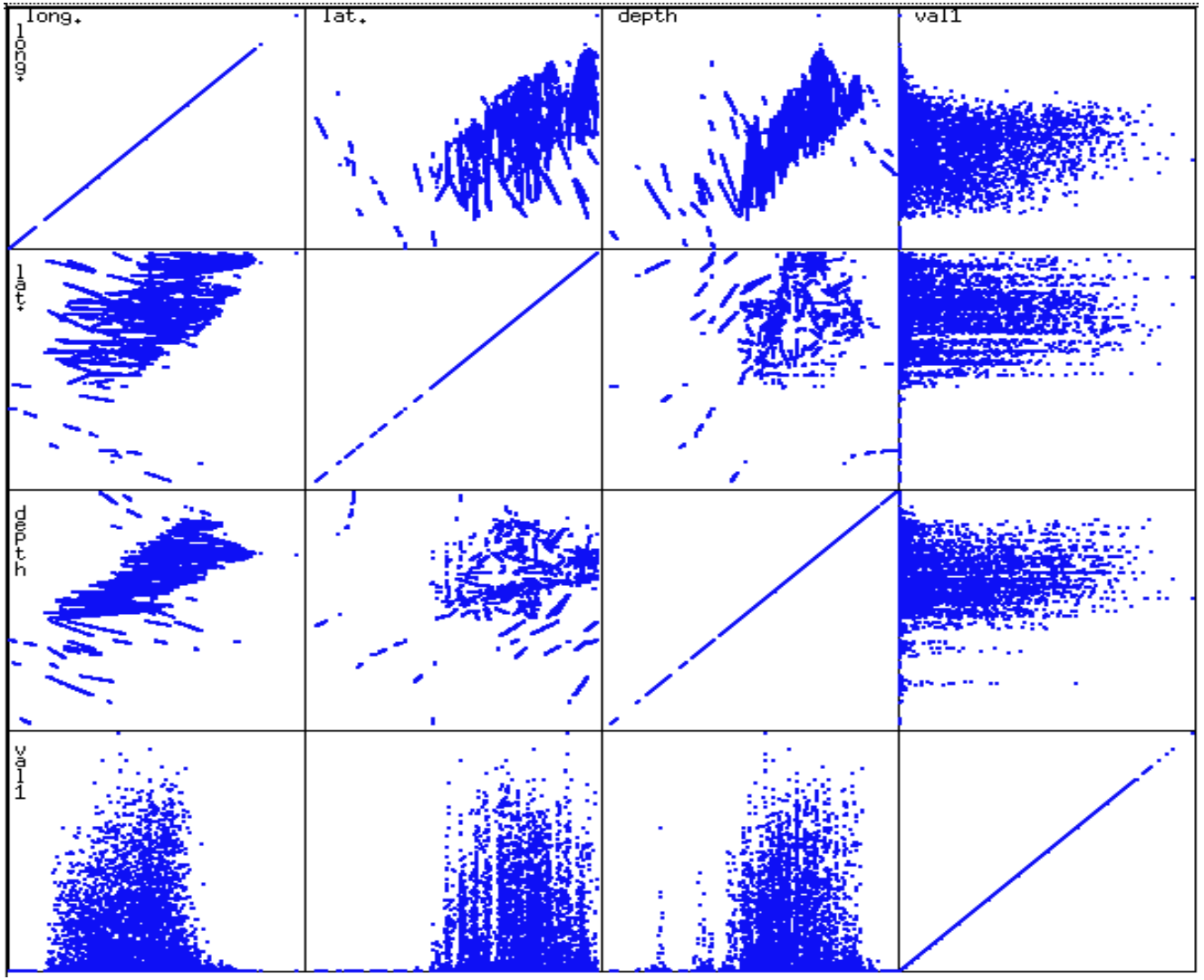
- Visualization of geometric transformations and projections of the data
- Methods
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Parallel coordinates

Scatter plot

A scatter plot displays 2D data points using Cartesian coordinates.

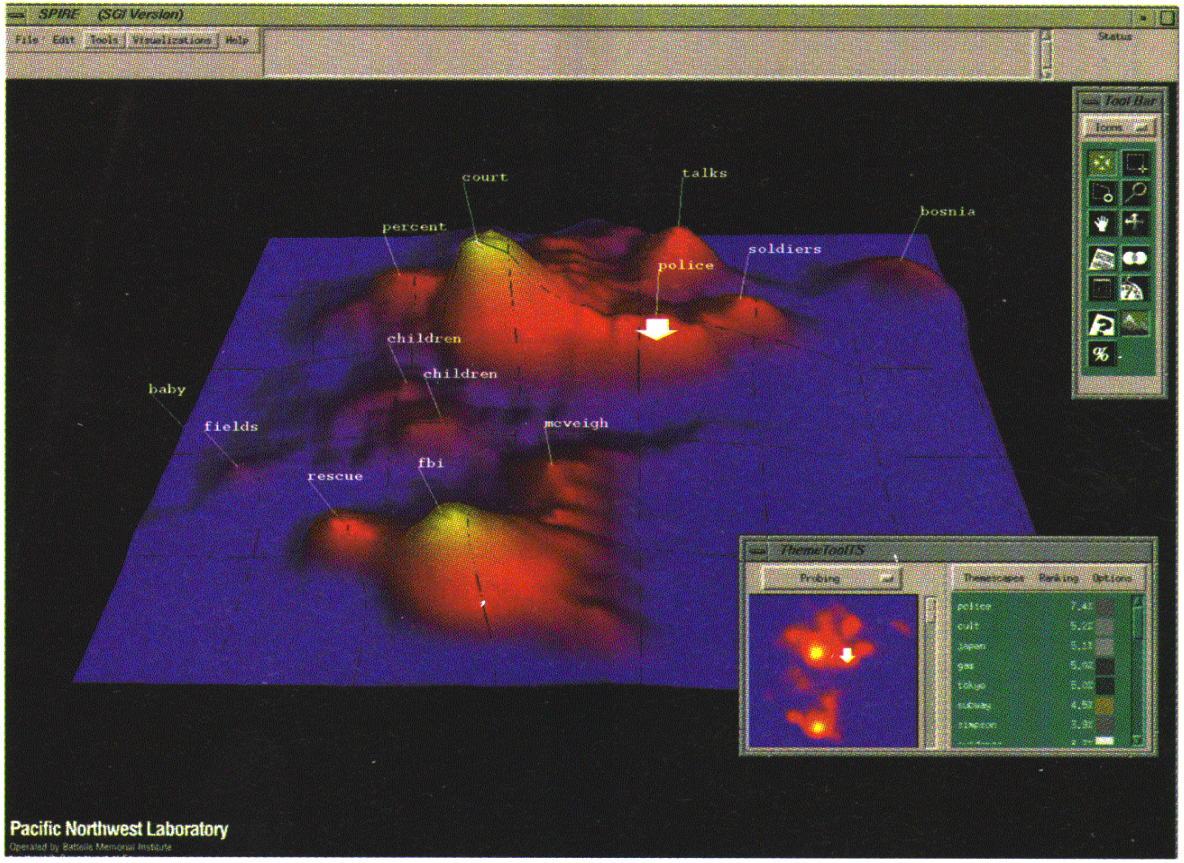


Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Landscapes



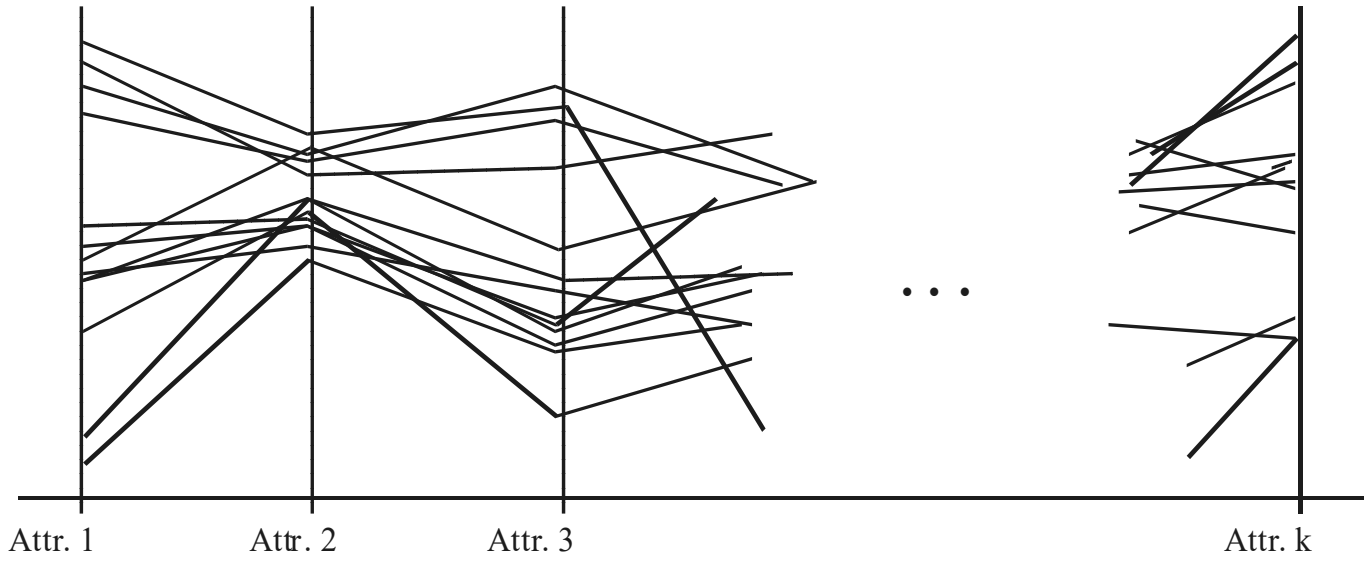
Used by permission of B. Wright, Visible Decisions Inc.

news articles visualized as a landscape

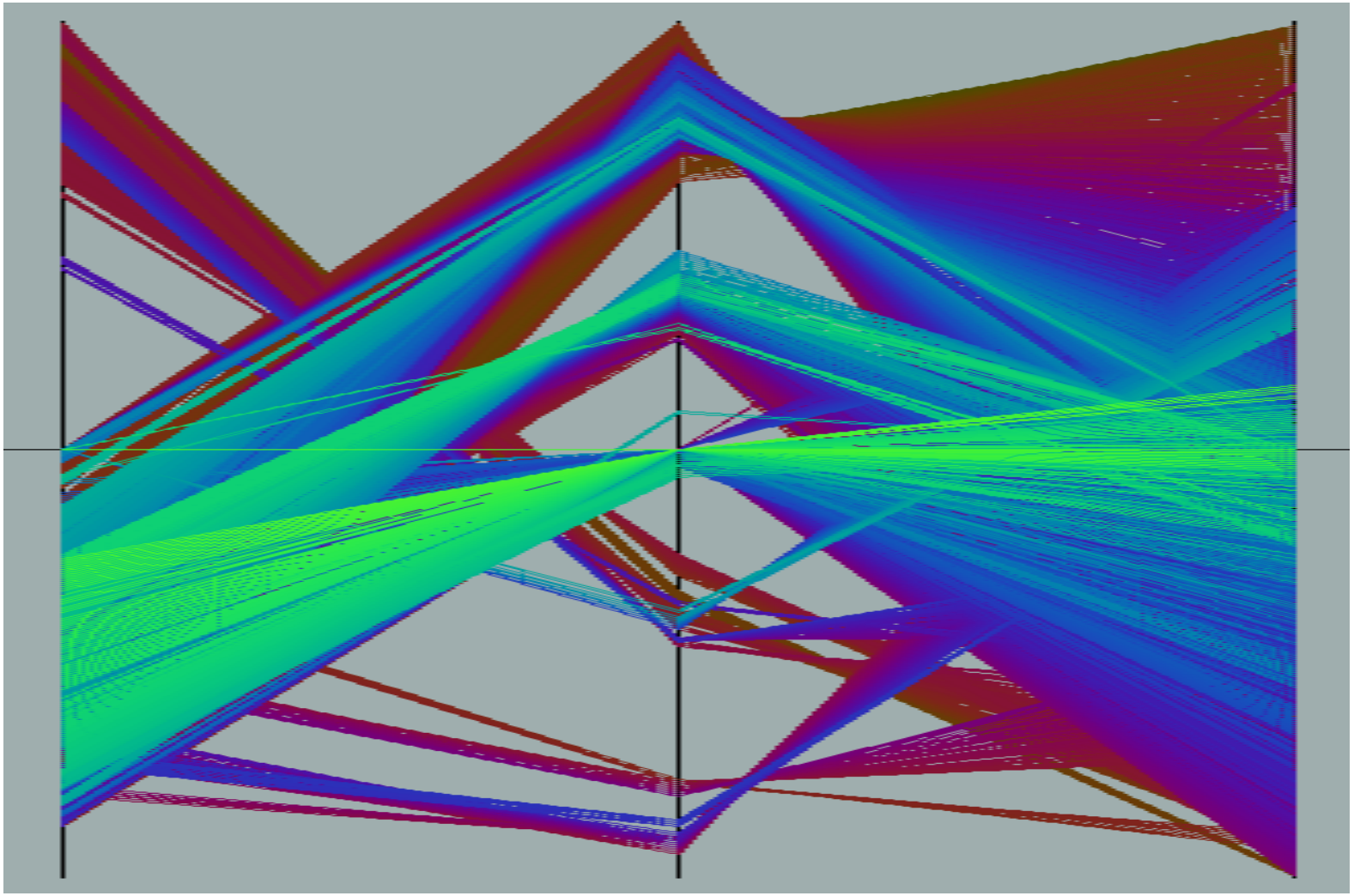
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Parallel Coordinates

- In equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



Parallel Coordinates of a Data Set

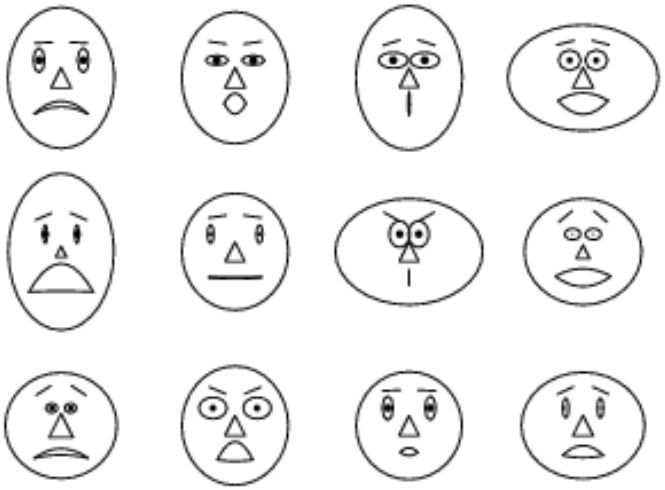


Icon-Based Visualization Techniques

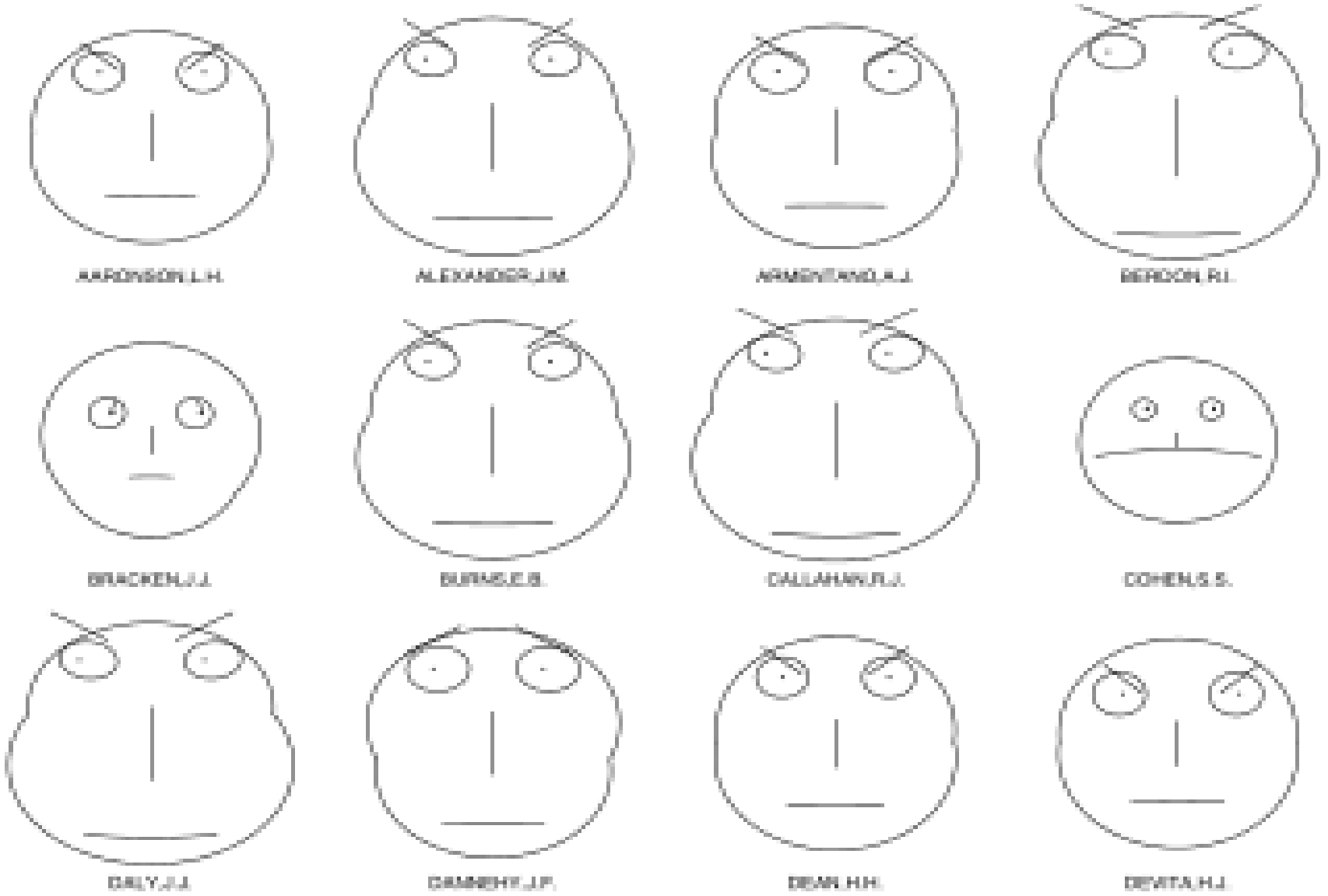
- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)

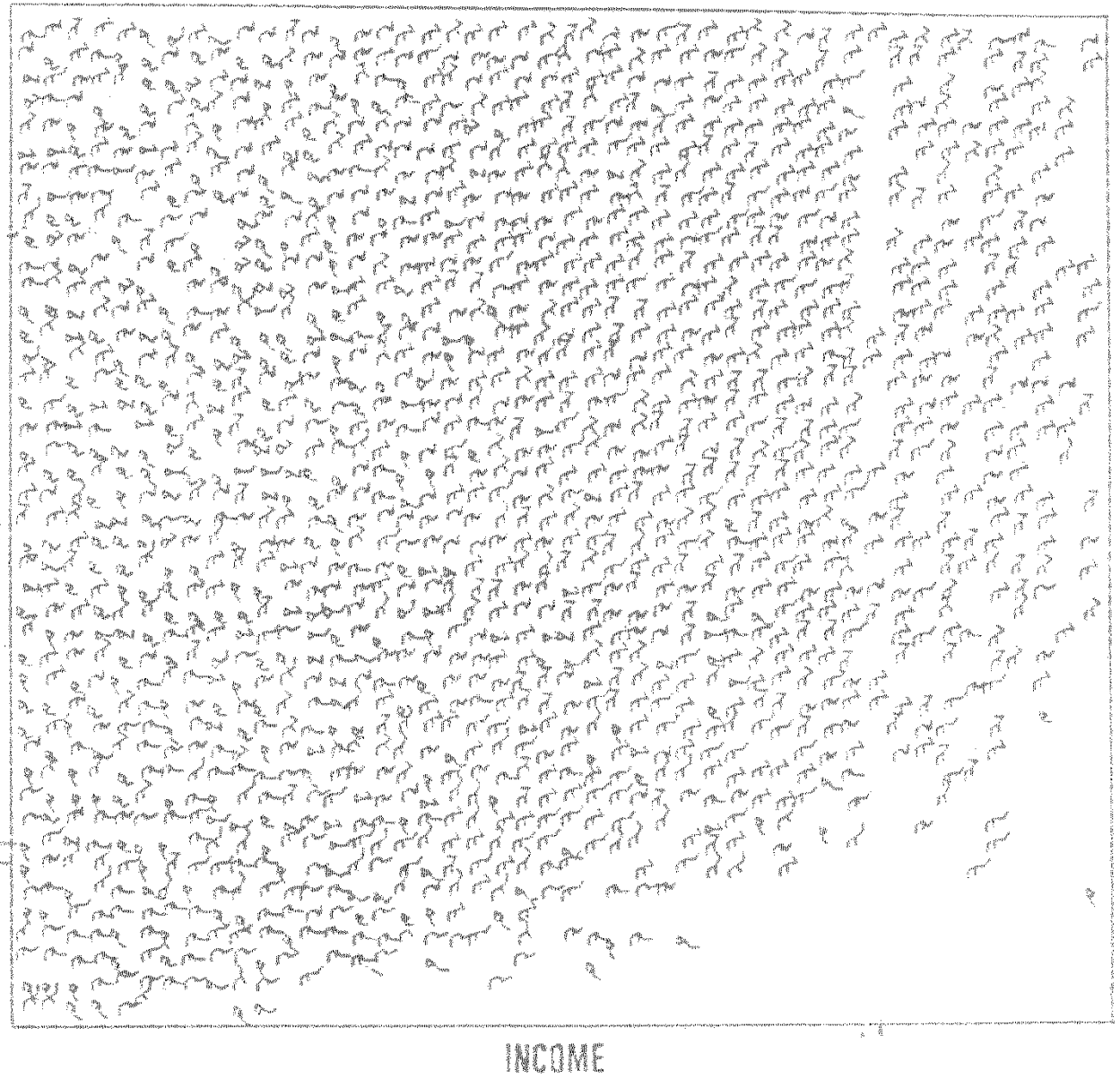


- REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics.* New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld*--A Wolfram Web Resource. mathworld.wolfram.com/ChernoffFace.html



This example shows Chernoff faces for lawyers' ratings of twelve judges

Stick Figure



A census data figure showing age, income, gender, education, etc.

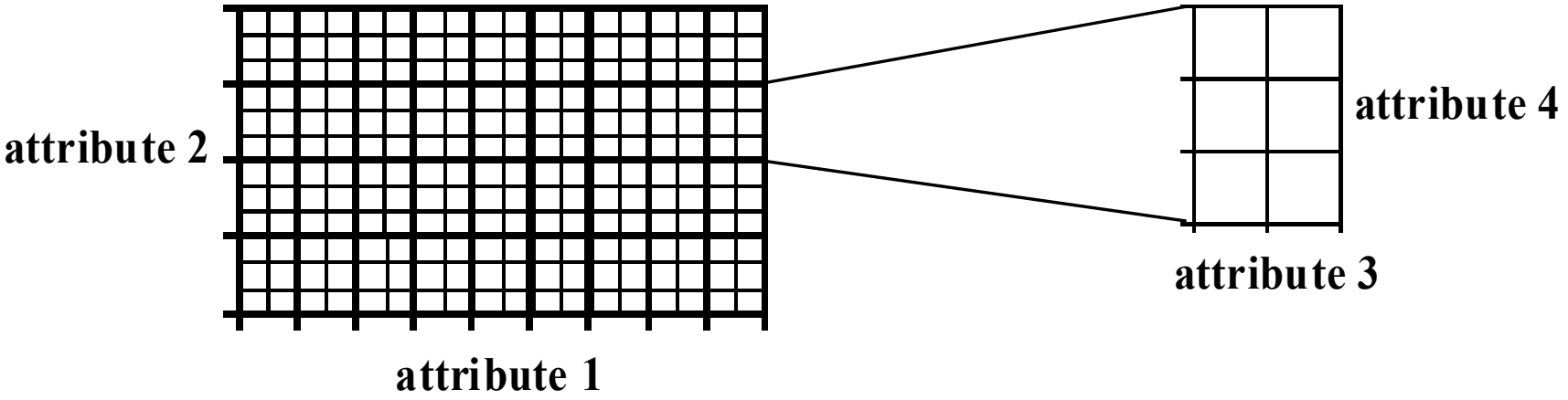
A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

Two attributes mapped to axes, remaining attributes mapped to angle or length of limbs". Look at texture pattern

Hierarchical Visualization Techniques

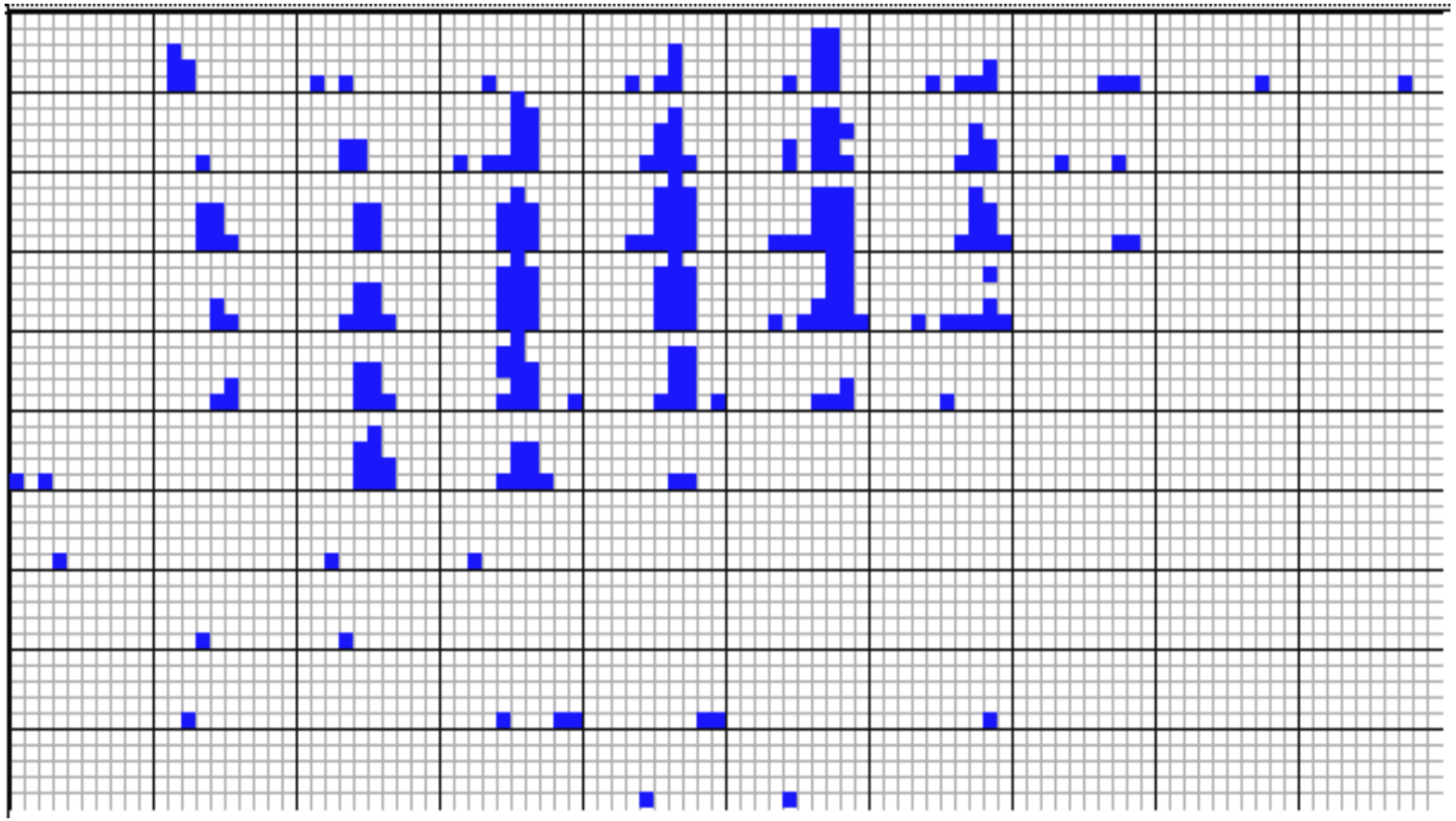
- Visualization of the data using a hierarchical partitioning into subspaces
  - Dimensionality stacking
  - Tag cloud

Dimensional Stacking



- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

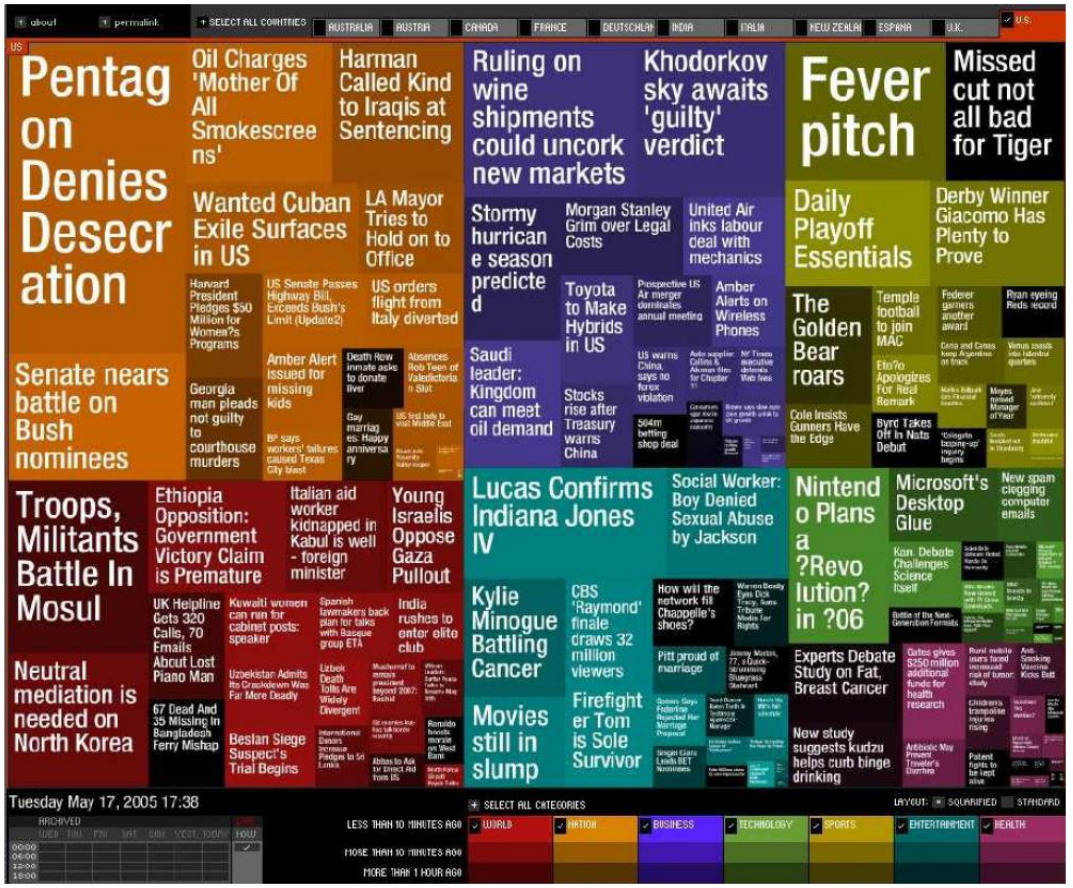
Dimensional Stacking



Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
  - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks



Newsmap: Google News Stories in 2005

Presentation Outline

- Data mining functionalities
- Research Issues in Data mining
- Sources of data
- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- **Case study**
- Summary

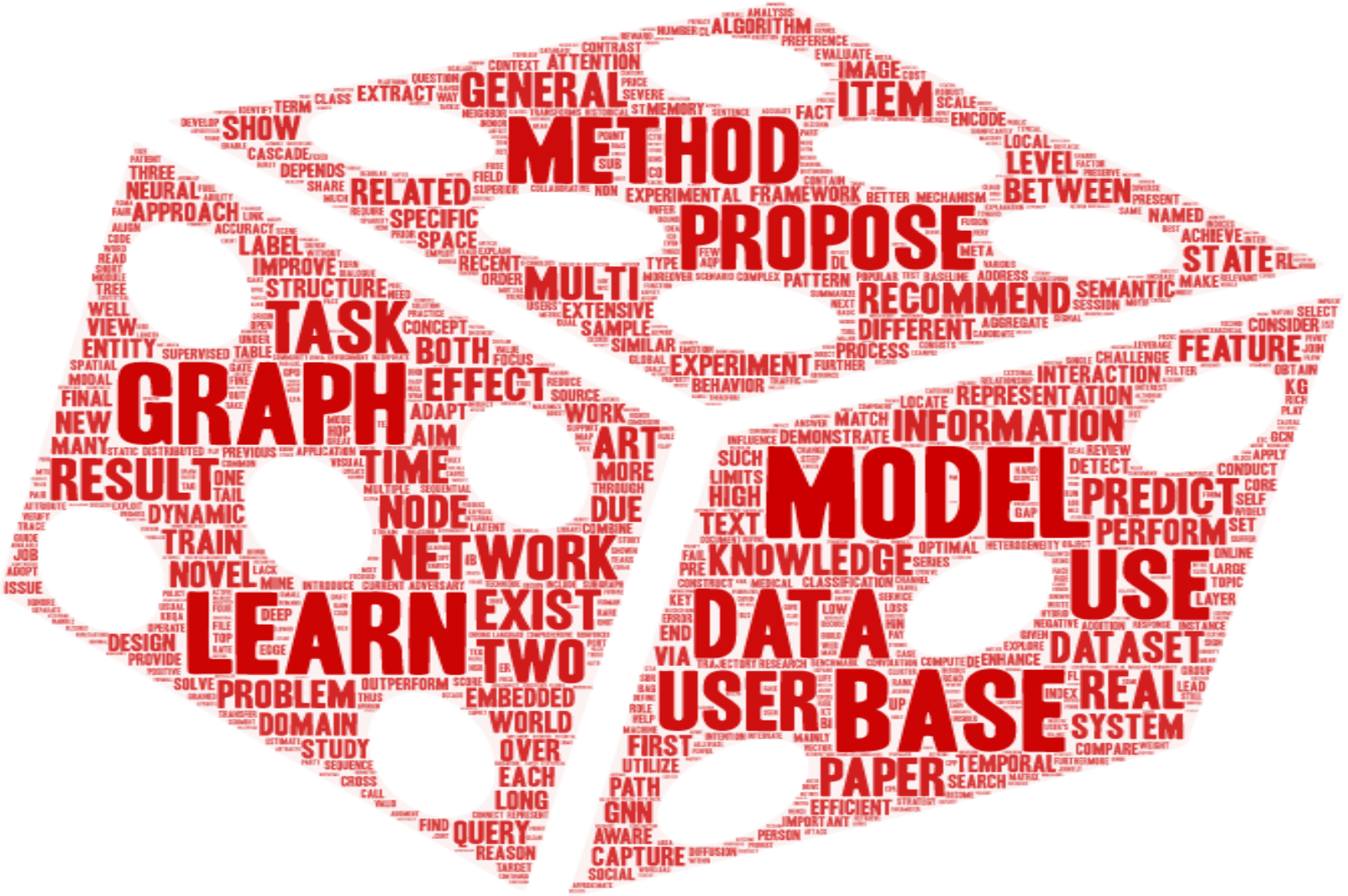
Case Study: Analysis of paper submissions to DASFAA-2022 conference

The 27th International Conference on Database Systems for Advanced Applications (DASFAA-2022), April 11-14, 2022, Hyderabad, India.

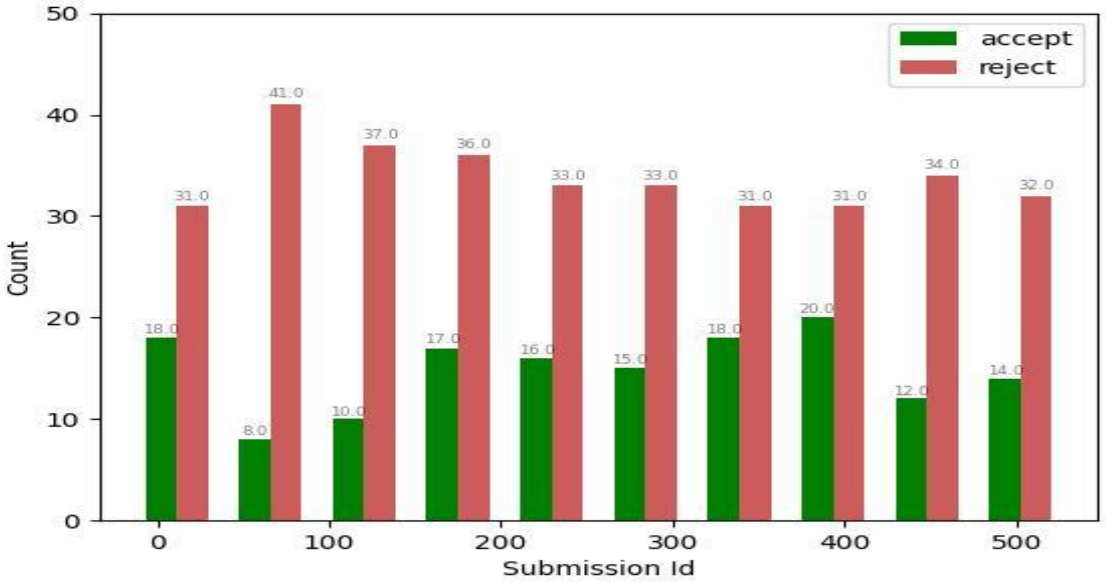
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

- Number of valid submissions (after desk-reject): 400
  - Yes, this is the number; no rounding off was done!
- Program committee
  - Reviewers: 205
  - Meta-Reviewers: 41
- Accept
  - Full: 72 (18%)
  - Short: 76 (19%)

Tag Cloud from Accepted Abstracts

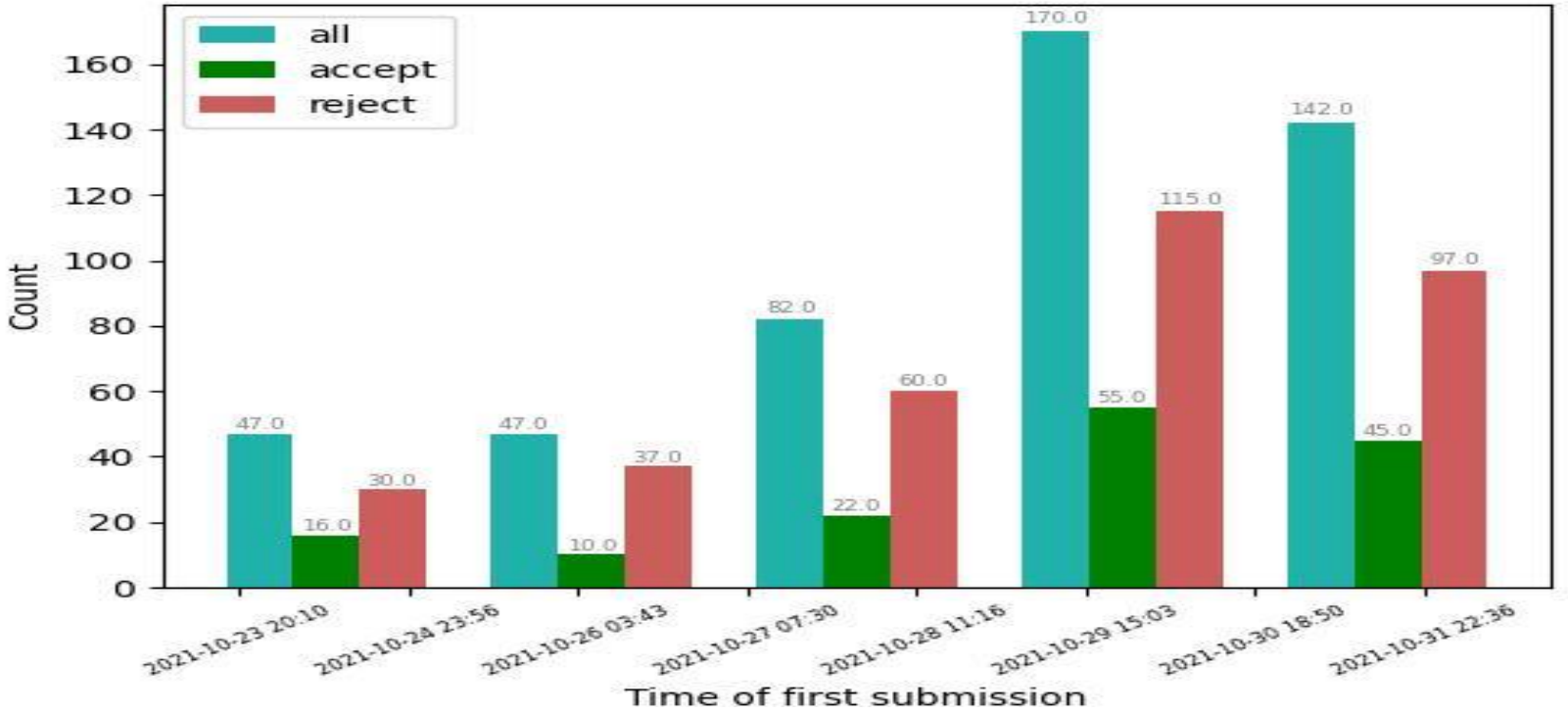


Does submission id matter?



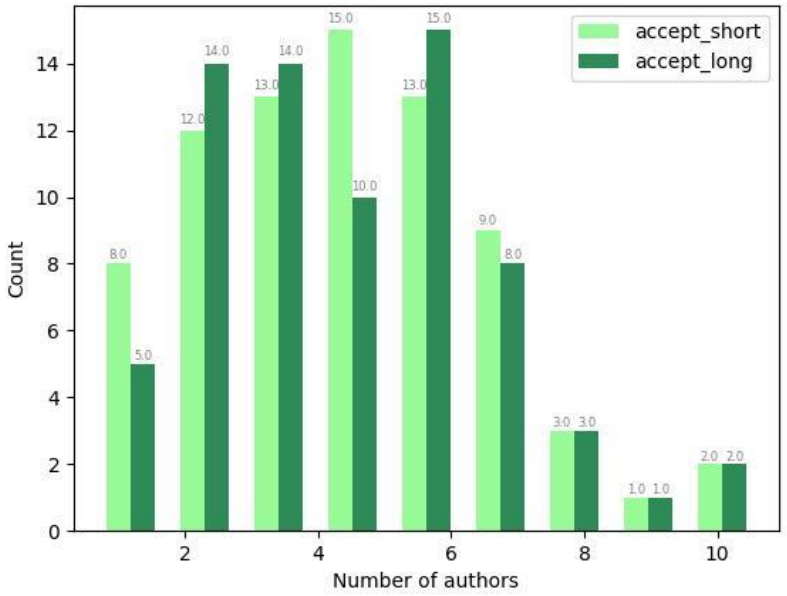
- Is there a sweet spot around 80 percentile?
- How will one know when to hit the 80 percentile?

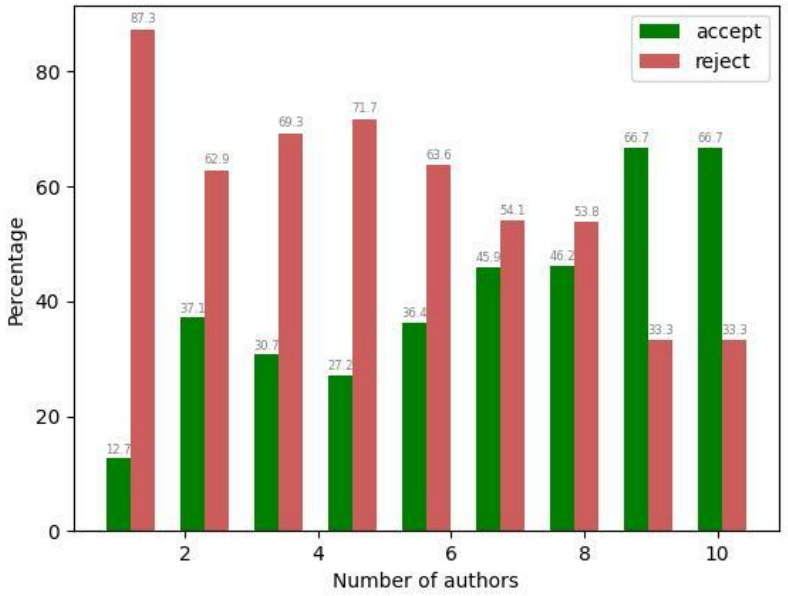
Time of first submission



- 1.5 to 2 days before the deadline!

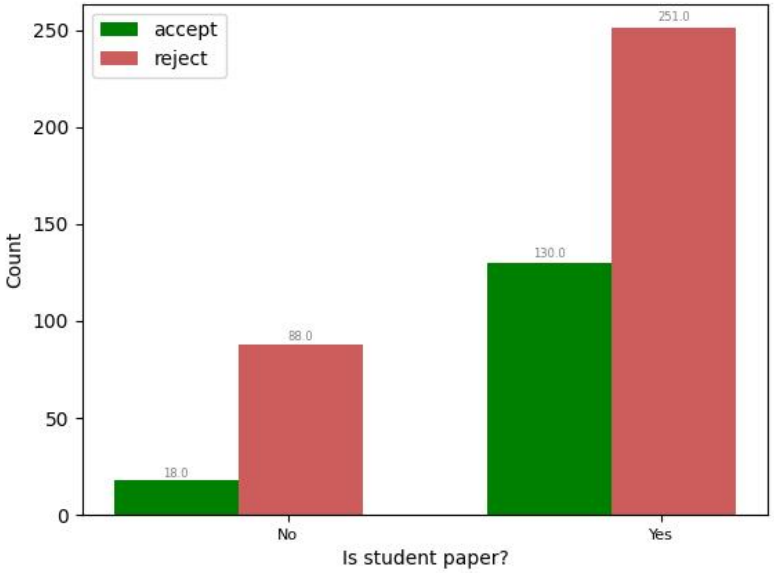
Does number of authors matter?

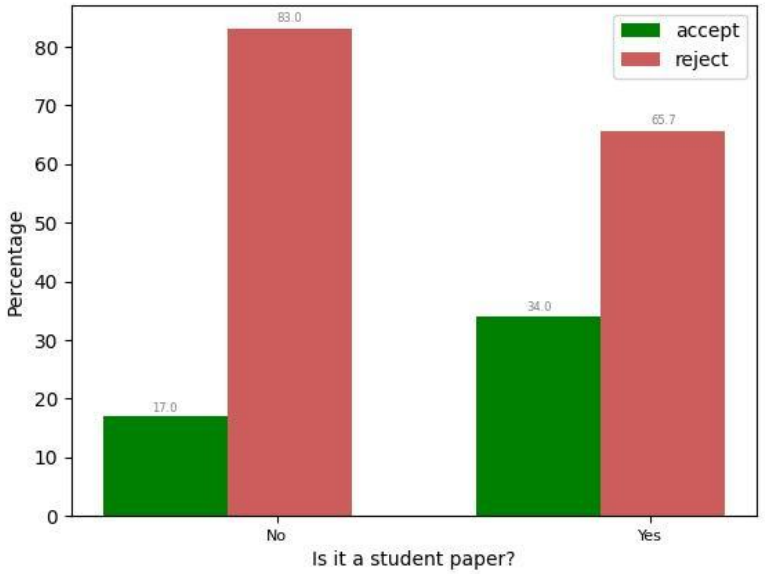




- The more the merrier – collaborations and multiple eyes matter!

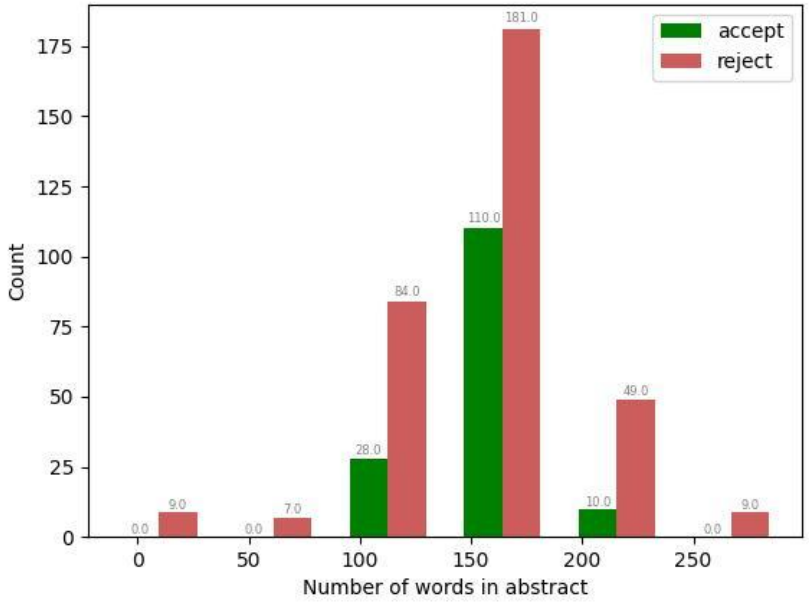
Student paper – first author is a student

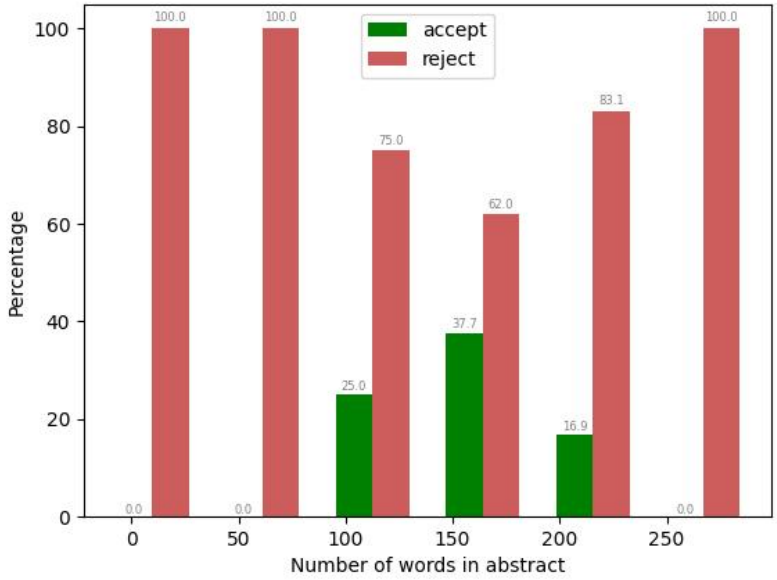




- Significant jump in acceptance percentage when a student is the first author – why?

Number of words in the abstract





- Too long or too short an abstract is not good

Other factors

- Other factors tried but no significant visual difference
- Number of words in the title
- Number of characters in the title
- ...
- ...

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009