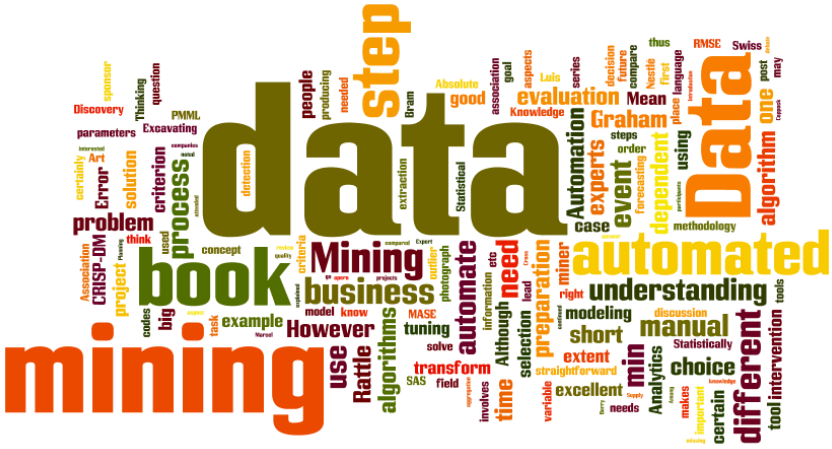
IT-533: Data Mining

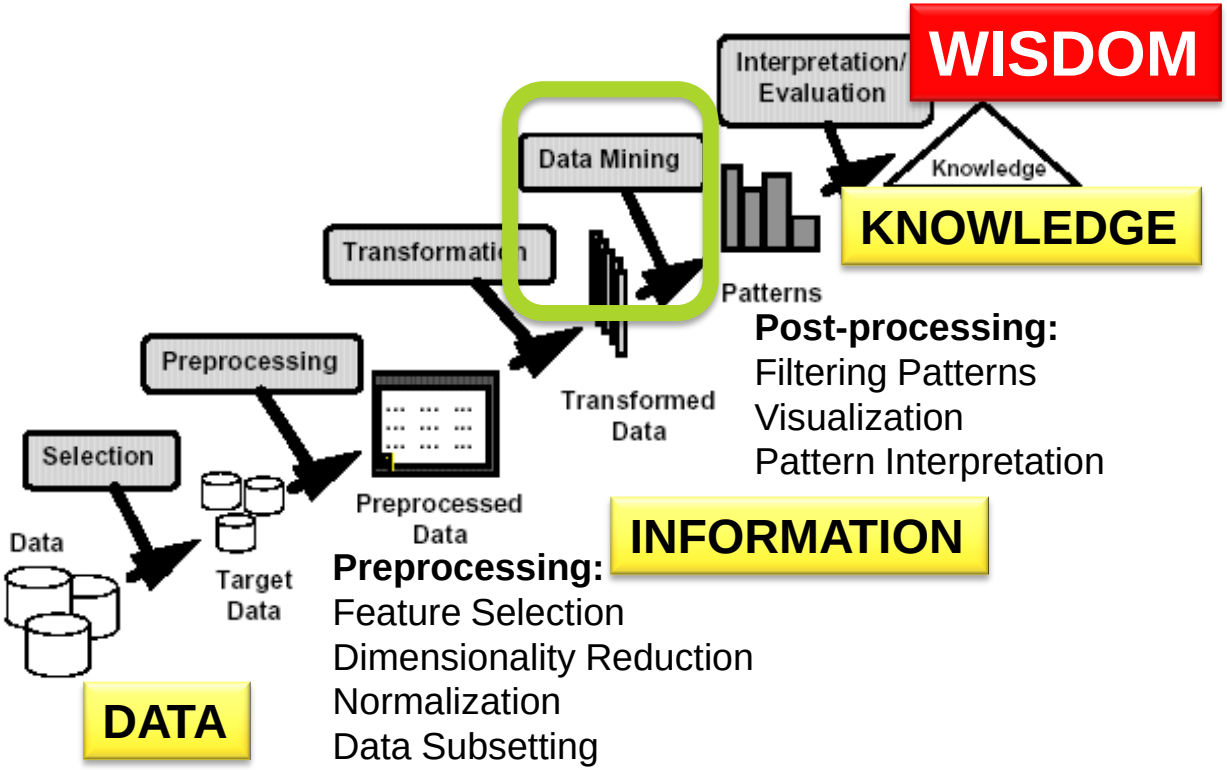


Week 2

Sonja Streuber

Welcome Back!

- Data Mining produces **WISDOM:**
  - Knowledge discovery (mining) in databases (KDD),
  - Business **INTELLIGENCE**



What decisions should executives make based on what the data mining results tell them?

- Military Generals
- Doctors
- City Planners
- Salespeople
- Meteorologists
- **YOU**

PM Reboot

What Can Data Mining Do?

Describe

- Find human-interpretable PATTERNS that describe the data.
  - Clustering
  - Association Rule Discovery (Link Analysis, Dependency Modeling)
  - Sequential Pattern Discovery (Time-series analysis, Sequential associations)

Predict

- Use existing variables to PREDICT unknown or future values of target variables.
  - Classification
  - Regression
  - Deviation Detection

PM Reboot

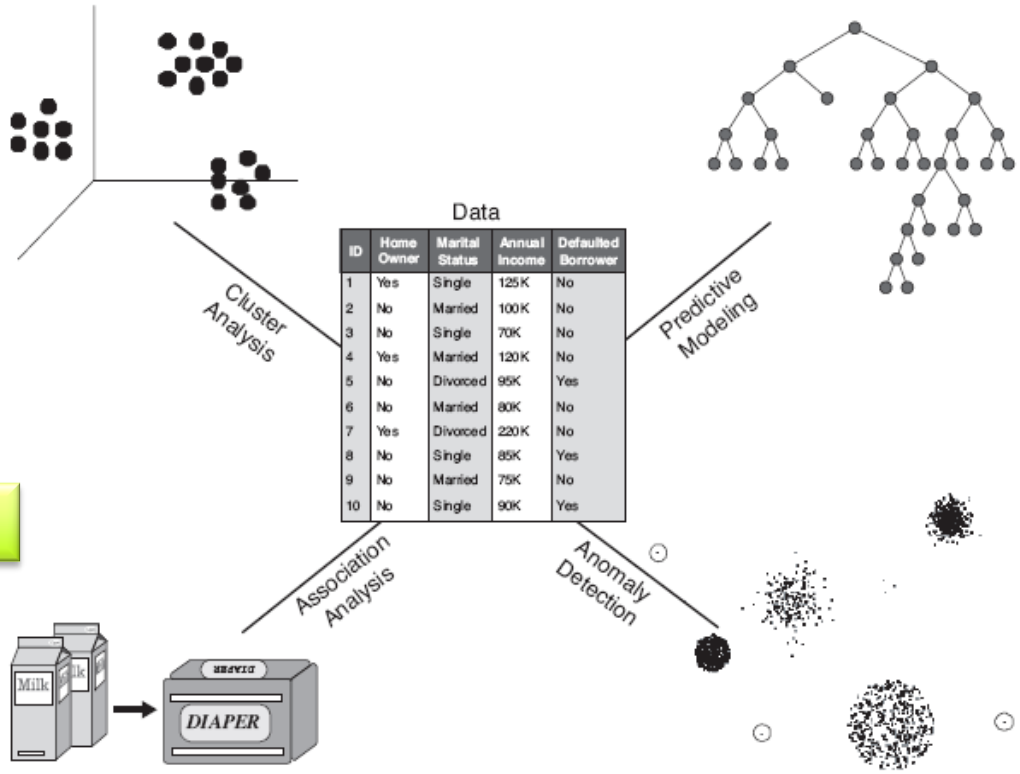
Core Tasks in Data Mining

2. Cluster Analysis

1. Classification, Prediction

3. Association

4. Anomaly Detection, Sequence



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 1.3. Four of the core data mining tasks.

Tan, Steinbach, Kumar (2005), *Introduction to Data Mining*, p. 7

Major Categories of Data Mining

Classification

- Goal: Find ways to separate data items into pre-defined classes (predict class)
  - We know X and Y belong together; find other things in same class
- Requires "training data": Data items where class is known
- Technologies:
  - Generate decision trees (results are human understandable)
  - Neural Nets

Prediction

- Predicting a numeric quantity

Clustering

- Goal: Find groups of similar data items
- Statistical techniques require some definition of "distance" (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions
- Uses:
  - Demographic analysis
  - Technologies:
  - Self-Organizing Maps
  - Probability Densities
  - Conceptual Clustering

Association

- Goal: Identify dependencies in the data:
  - X makes Y likely
- Indicate significance of each dependency
- Bayesian methods
- Uses:
  - Targeted marketing
- Technologies:
  - AIS, SETM, Hugin, TETRAD II

Sequence

- Find event sequences that are unusually (un)likely
- Requires "training" event list, known "interesting" events
- Must be robust in the face of additional "noise" events
- Uses:
  - Failure analysis and prediction
- Technologies:
  - Dynamic programming (Dynamic time warping)
  - "Custom" algorithms

Goal: Learn about Concepts and Patterns

PM Reboot

Machine Learning: Unsupervised

SUPERVISED (=output datasets are provided which are used to train the machine and get the desired outputs)

- **Classification learning** is *supervised, i.e. outcome* is provided with training data
  - Outcome is called the ***class*** *of the example*
  - Measure success on fresh data with known class labels (*test data)*
  - WEKA example problems: weather data, contact lenses, irises, labor negotiations
  - Also used in search engines
- **Association and Sequence learning** can be applied if no class is specified and any kind of structure is considered "interesting"
  - Difference to classification learning:
    - Can predict any attribute's value, not just the class, and more than one attribute's value at a time
    - Hence: far more association rules than classification rules
    - Thus: constraints are necessary (minimum coverage and minimum accuracy)

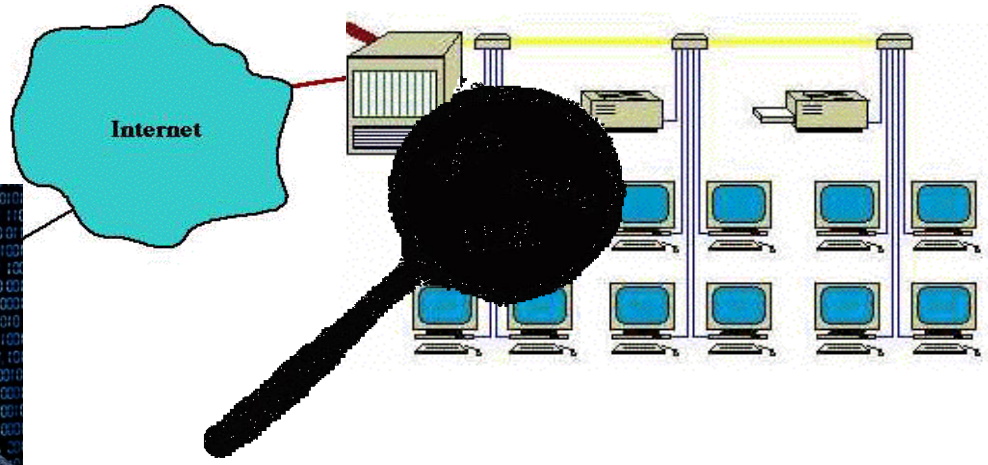
UNSUPERVISED (= no datasets are provided, instead the data is clustered into different classes)

- **Clustering learning** is UN*supervised (i.e. class of an example is NOT KNOWN*)
  - Outcome may point towards a class
  - In practice, success is often measured subjectively

Special Case Anomaly Detection

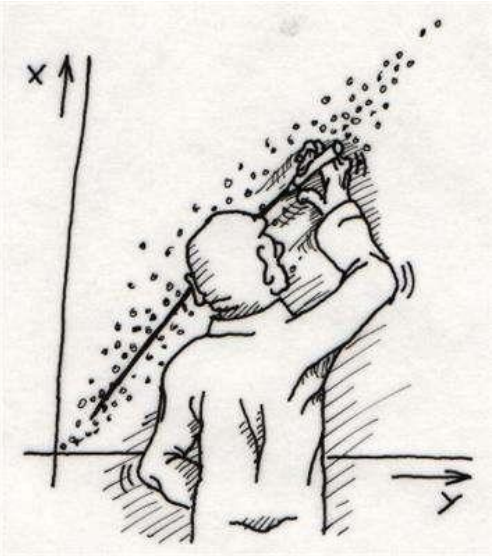
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Special Classification: Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.



Test Yourself!

Problem	Appropriate Method	Supervised or UNsupervised	Possible Attributes?
Do patients with the measles get the flu more easily?			
What weather patterns do we see in Chicago in Jan?			
How much organic milk will we sell next month?			
What political candidate is most likely to win the election?			
Should diapers and milk be in the same store location?			
What system error messages can predict a nuclear reactor meltdown?			

Basic Statistical Descriptions of Data

Metric	What question does it answer?	For what do we use it?
Frequency	Do we have even value distributions?	Do we need to consider all categories?
Minimum	What is the smallest value?	Is this an unusually low outlier?
Maximum	What is the largest value?	Is this a unusually high outlier?
Range	How far is the smallest from the largest value?	Does the range look probable?
Sum	Can/ should the values be summed?	
Mean	What is the average?	Use as missing data replacement
Median	What is the middle value?	Use as missing data replacement
Mode	What is the most frequent value?	Use as missing data replacement
Standard Dev	What is the average distance of attribute values from the attribute average?	Is the data distribution “narrow” and high or “wide” and flat? How unified does the data look?
Quartiles (1,3)	What does the middle 50% look like?	

Agenda

#	Topic	Reading	Due Sunday 11:59 pm CST
2	Data, Databases, and Data Warehouses and Basic Statistics with R	• Zhao 3 • Han/ Kamber 2,4,5 • Witten/Eibe 2 • Videos as posted on Blackboard • Lynda Course " R Statistics Essential Training" mod. 1-4 • *Install MS Excel*	**MW**: **Quiz 2**: Basic Stats question in R **S**: **FINAL_PROJECT_2**: What are the attribute data types in your data set? Display the attributes in R and generate a 5 number summary. How are the attribute values distributed?

PM Reboot

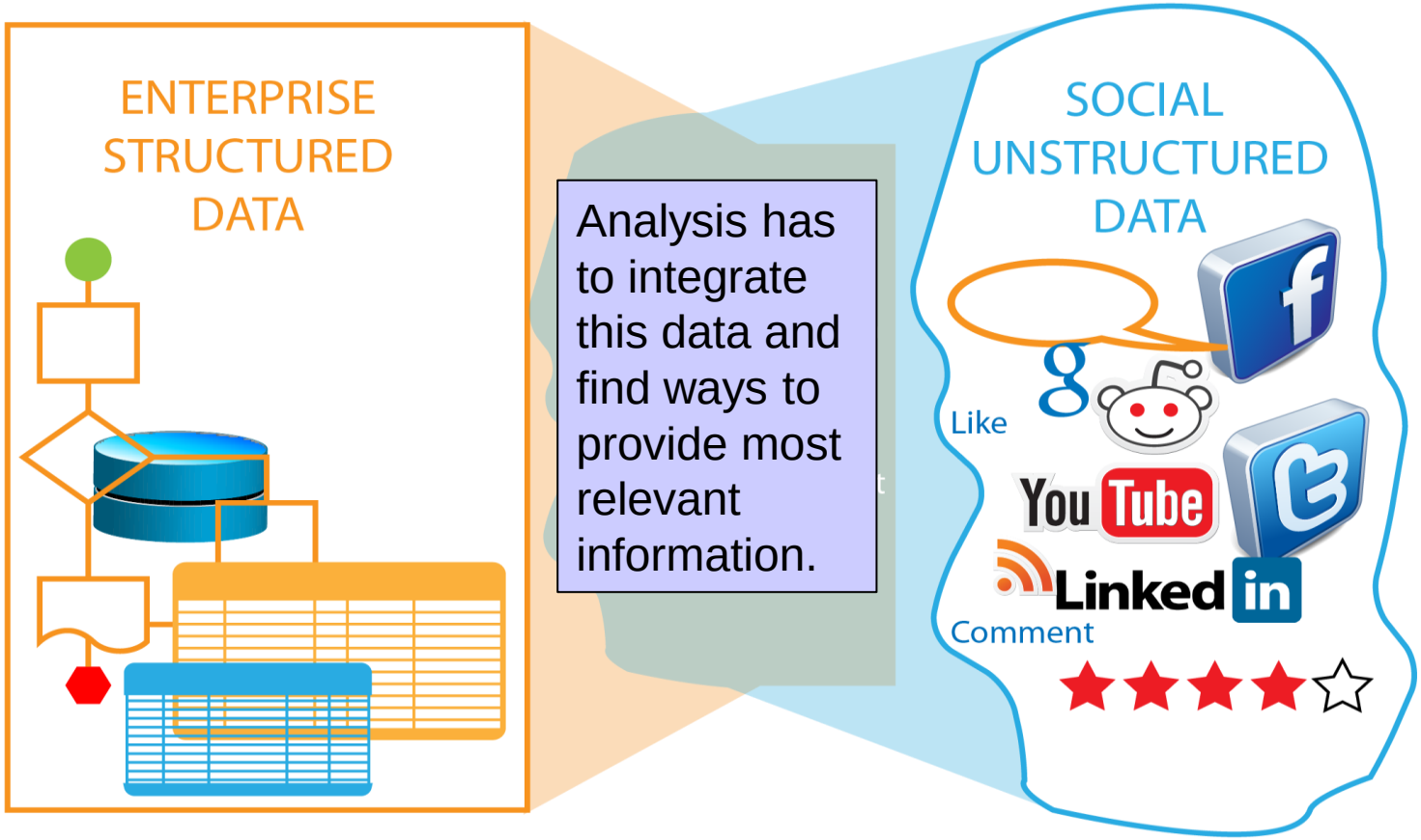
UNDERSTANDING DATA

Learning Objectives

- Identify structured or unstructured data?
- Describe and apply data types to a data set
- Determine what mathematical operations are possible
- Analyze data distribution, both mathematically and graphically
- Analyze data relationships and applying considerations of distance
- Identify, and plan mitigation of, data quality issues

PM Reboot

1. Structured vs. Unstructured Data



1. Test Yourself!

1. Structured Data

Characterized by

Dimensionality	Curse of Dimensionality
Sparsity	Only presence counts
Resolution	Patterns depend on the scale
Distribution	Centrality and Dispersion

1. **Record**
   - Relational Records
   - Data Matrix, numerical matrix, crosstabs
   - Document data: text documents: term-frequency vector
   - Transaction Data
2. **Graph**
   - World Wide Web
   - Social Networks
   - Molecular Structures
3. **Ordered**
   - Video data: sequence of images
   - Temporal data: time-series
   - Sequential Data: transaction sequences
   - Genetic sequence data

PM Reboot

1.1 Record Data

A collection of records, each of which consists of a fixed set of attributes

Data Matrix

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Document 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

Each document becomes a `term' vector,

- each term is a component (attribute)
- the value of each component is the number of times the corresponding term occurs in the document.

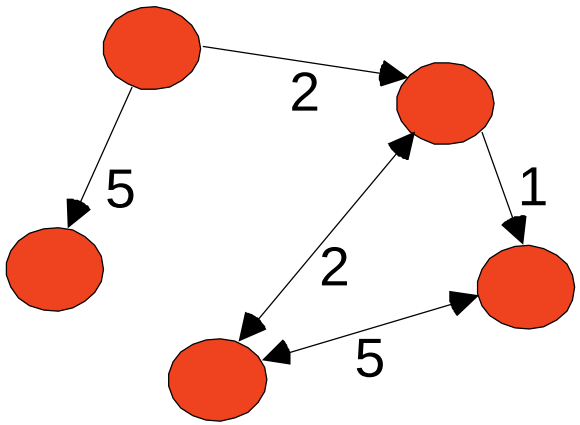
Transaction Data

A special type of record data, where each record (transaction) involves a set of items.

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

1.2 Graph Data

- Examples: Generic graph and HTML Links

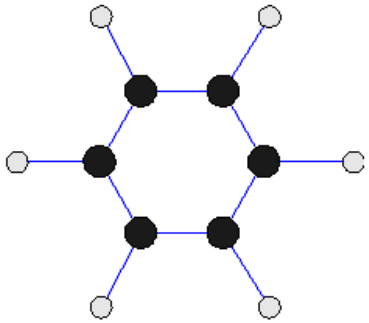


2.1 WWW

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

2.2 Molecular Structures

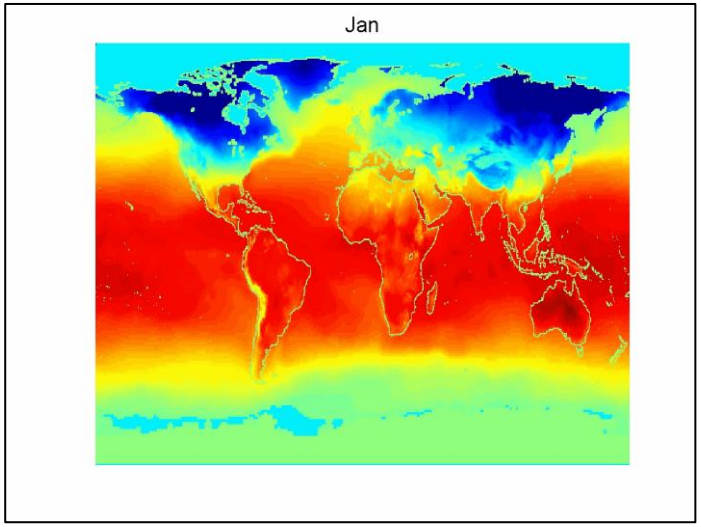
- Benzene Molecule: C_6H_6



1.3 Ordered Data

- Sequences of transactions

3.1 and 3.2 Spatio-Temporal



3.3 Sequence

Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

3.4 Genetic

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

2. Datasets/ Instances and Attributes

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance
- Related problem: existence of an attribute may depend of value of another one

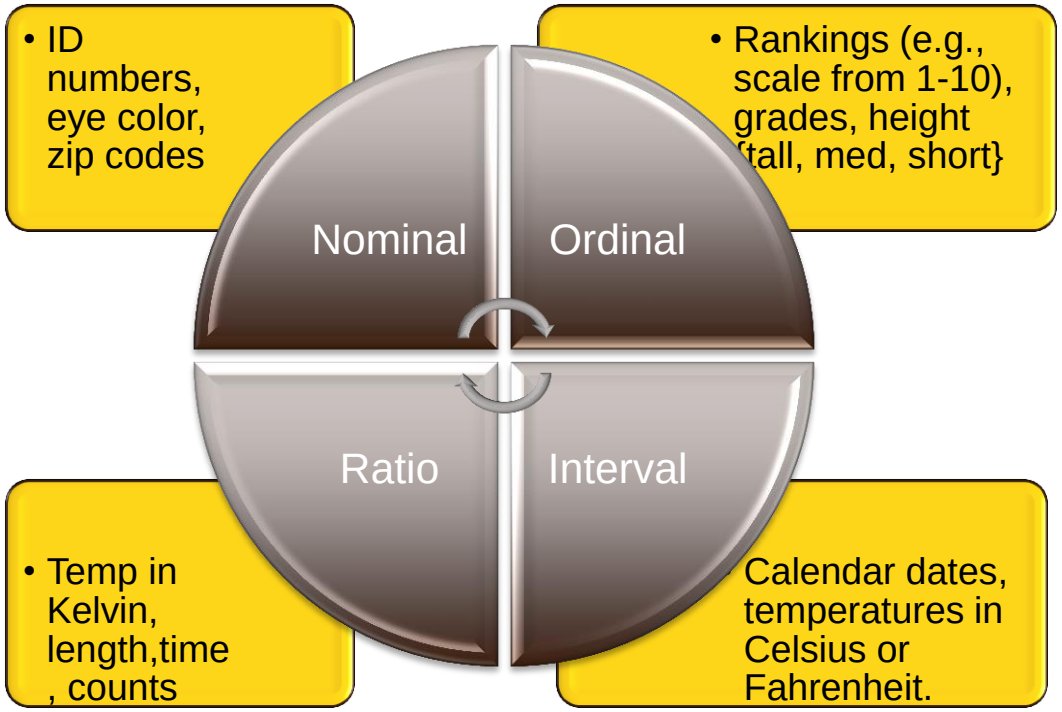
Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

PM Reboot

2. Attribute Types: Examples



The type of an attribute depends on which of the following properties it possesses:

- Distinctness: = ≠
- Order: < >
- Addition: + -
- Multiplication: * /

- Nominal attribute: Distinctness
- Ordinal attribute: Distinctness & Order
- Interval attribute: Distinctness, Order & Addition
- Ratio attribute: all 4 properties

PM Reboot

2. Attribute Types—Overview

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

PM REBOOT

3. Attribute Types—Transformations

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

PM Reboot

2 & 3: Test Yourself!

Attribute	Quantitative or Qualitative?	Type?	Discrete or Continuous?
Patients seen in June			
Product Rating (1 to 5 stars)			
% of hack attacks averted per week			
Days per month			
Avg temperature in Spring, Summer, Fall, Winter			
Currency exchange rates to US$.			
Table length in metres.			
Customer address			
Zip code			
Distance between Valparaiso and Chicago			

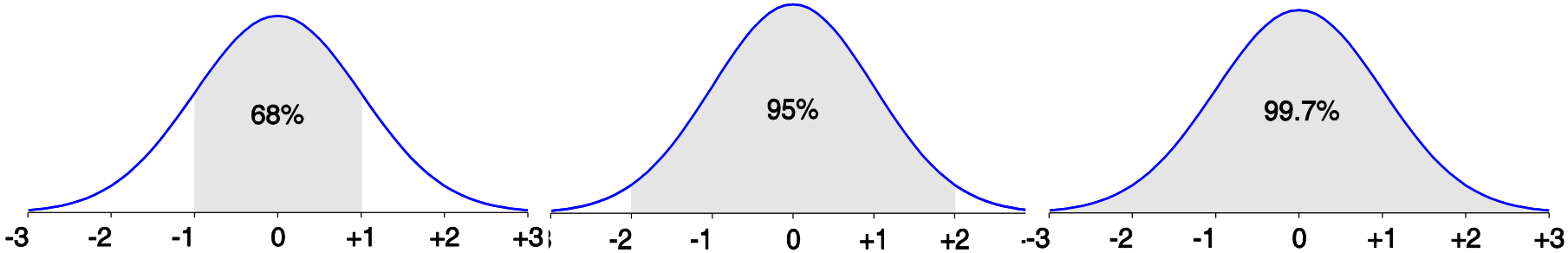
4. Graphic Displays

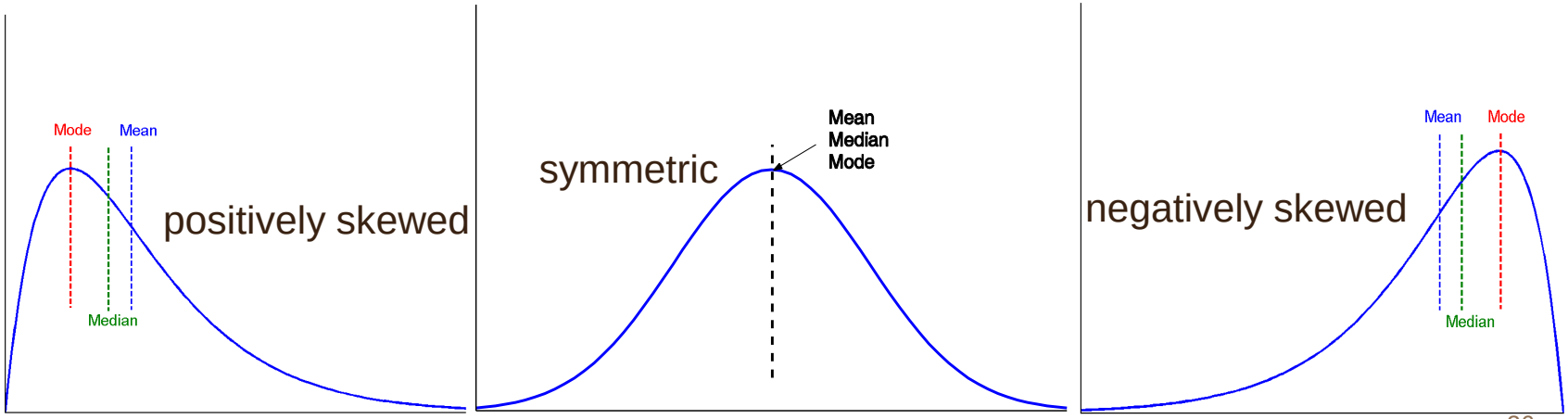
- **Normal Curve:** graphic display of data distribution
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (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

4. Dispersion: Normal Distribution

μ: mean, σ: standard deviation

- From μ–σ to μ+σ: contains about 68% of all data
- From μ–2σ to μ+2σ: contains about 95% of all data
- From μ–3σ to μ+3σ: contains about 99.7% of all data





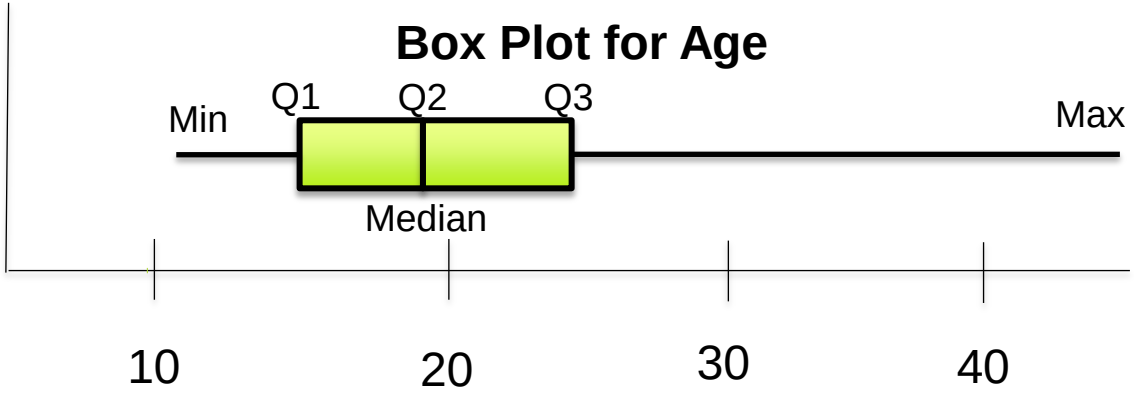
4. Boxplots—5 Number Summary

#	Sentiment	Customer Age
1	Strongly Agree	32
2	Agree	45
3	Neutral	17
4	Disagree	12
5	Strongly Agree	22
6	Neutral	25
7	Agree	27
8	Agree	18
9	Strongly Disagree	13
10	Strongly Disagree	14
11	Strongly Agree	24
12	Neutral	29
13	Agree	19
14	Neutral	11
15	Strongly Agree	23
16	Strongly Disagree	18
17	Neutral	25
18	Disagree	17
19	Strongly Agree	22
20	Strongly Disagree	11
21	Strongly Disagree	15
22	Strongly Agree	28
23	Strongly Disagree	14
24	Disagree	18
25	Agree	27

A boxplot shows the attribute values as quartiles. The middle 50% of the data, from Q1 to Q3, are displayed as boxes, with the median (Q2) marked. The lower 25% and the upper 25% appear as "whiskers."

5 Number Summary for Age

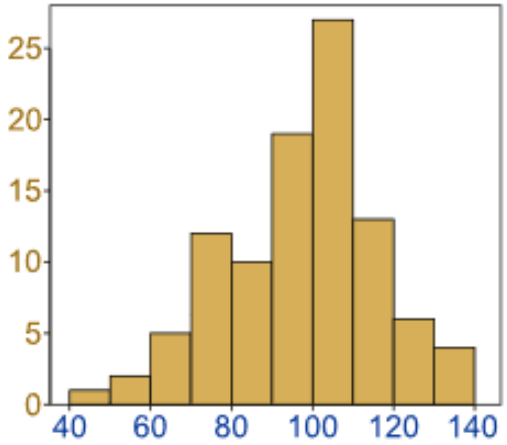
Age	Min	Q1	Median	Q3	Max
	11	15	19	25	45



4. Histogram and Pareto Chart

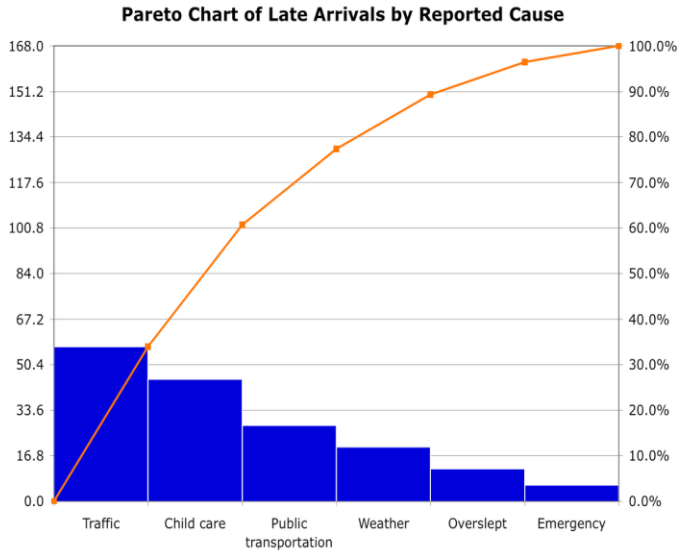
Histogram

- Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

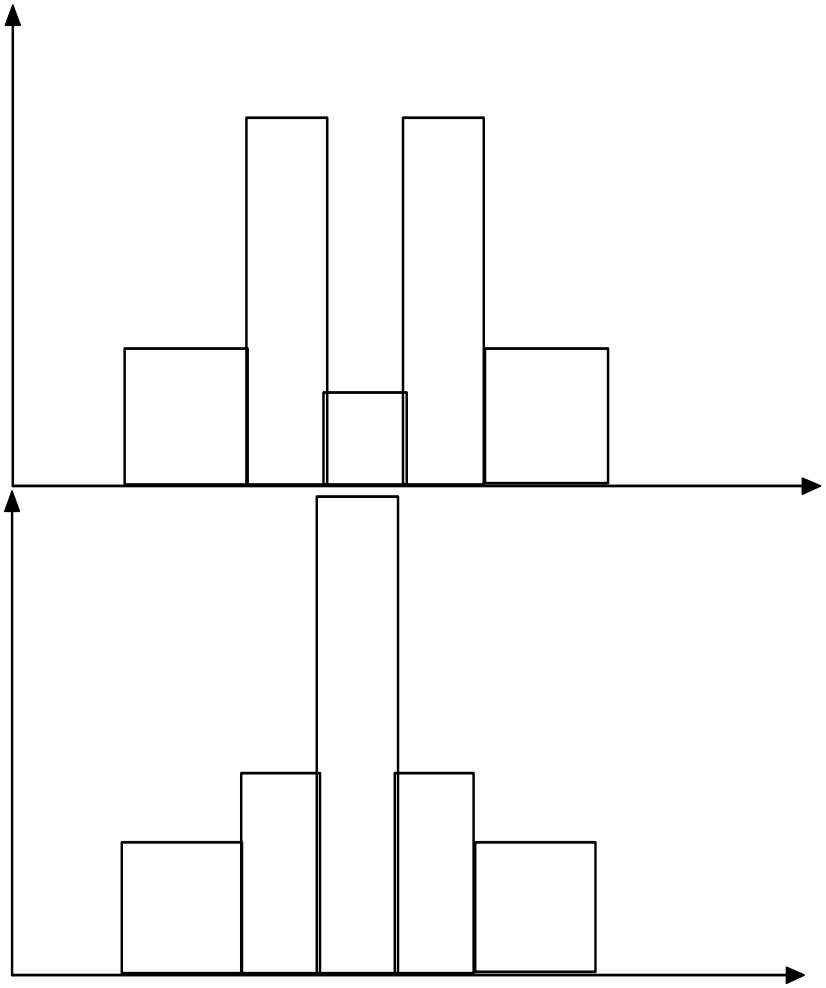


Pareto Chart

- Contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line



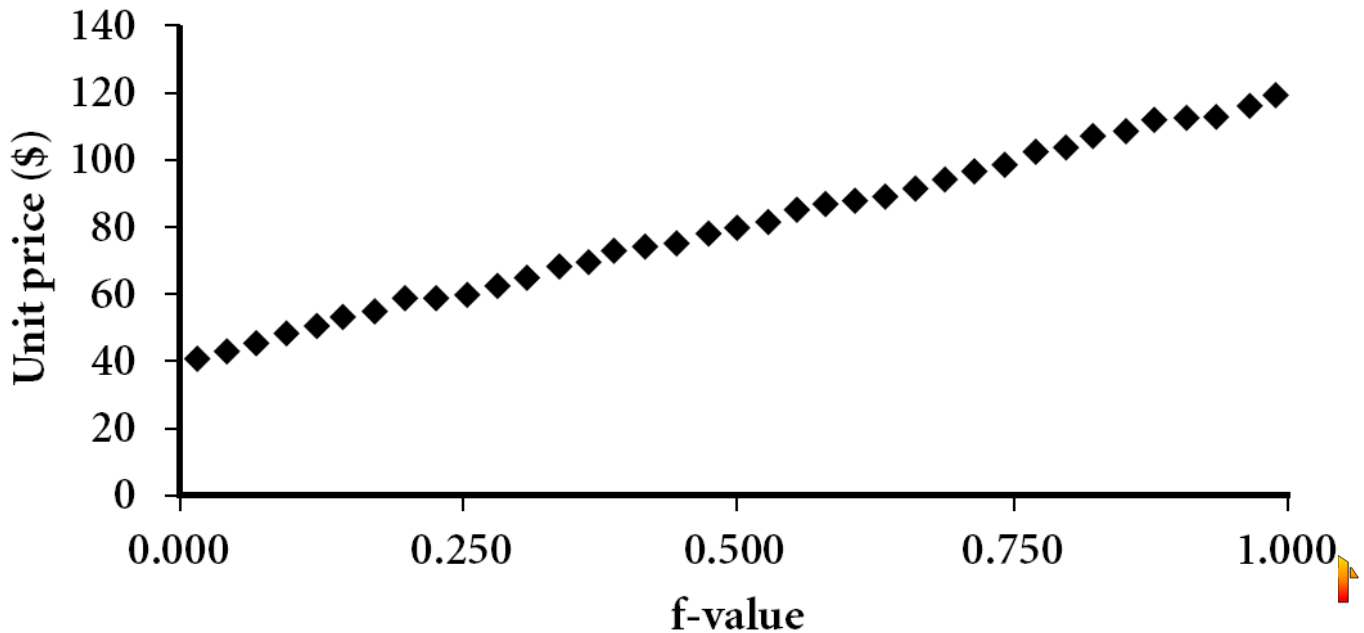
4. Histograms vs.Boxplots



- 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

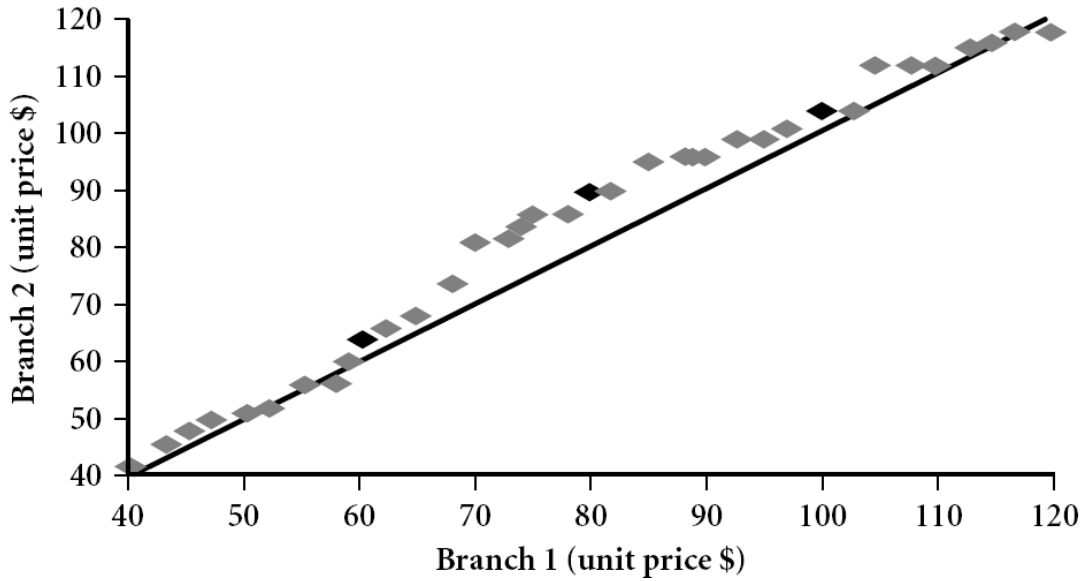
4. Quantile Plot

- 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



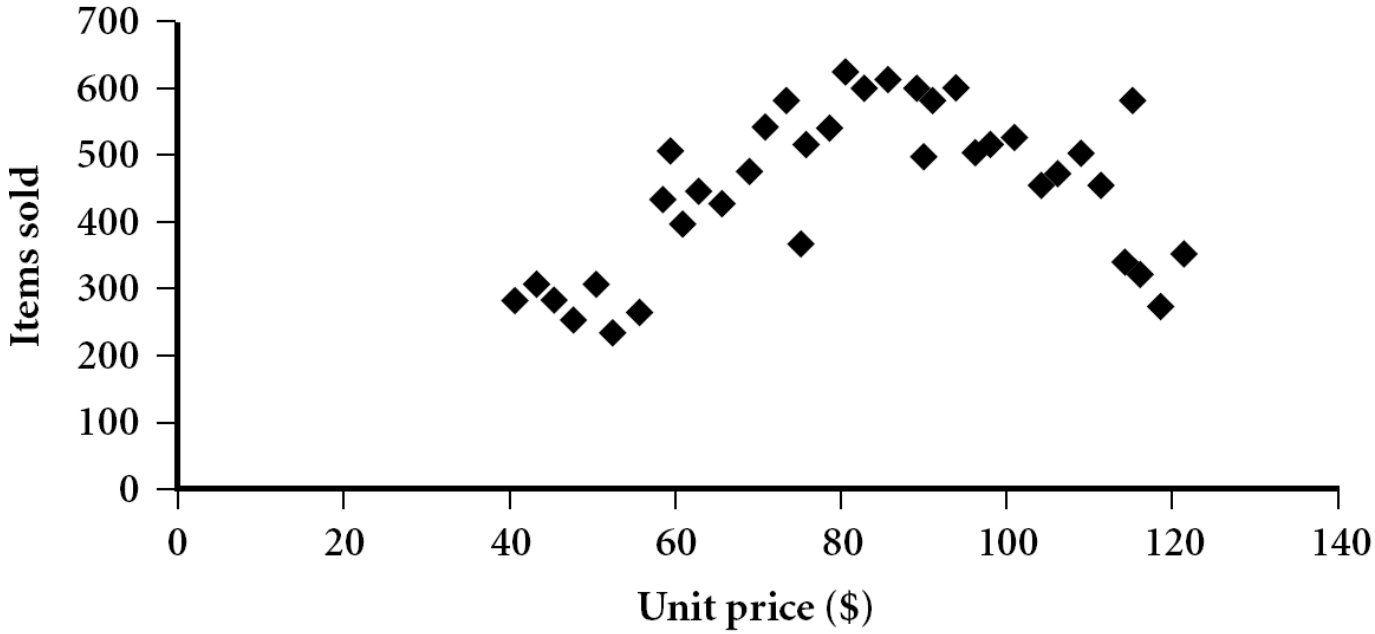
4. Quantile-Quantile (Q-Q) Plot

- 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.

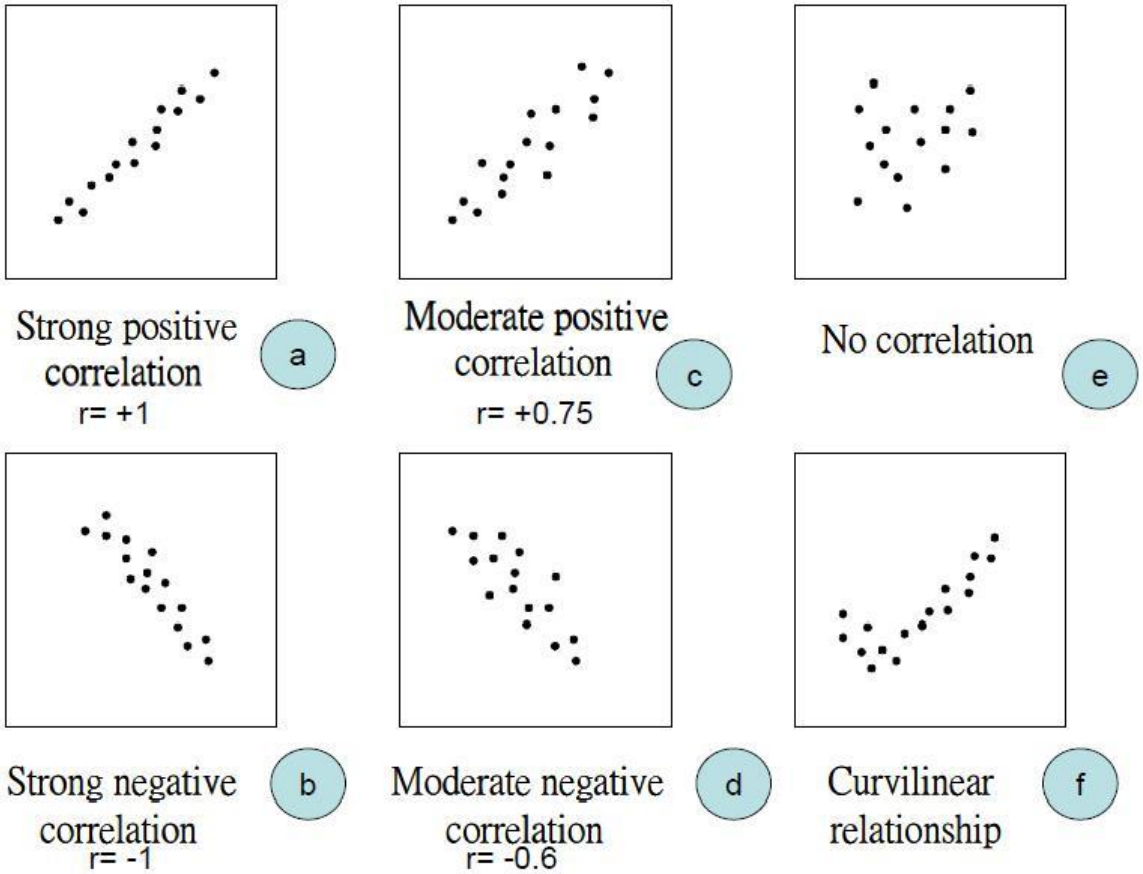


4. Scatter plot

- 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



4. Correlated Data



5. Data Relationships

- Data Similarity and Dissimilarity
- Distance between data points

5 .Data Similarity and Dissimilarity

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

Can be explained by different measures, like

- Spatial distance (distance **matrix**),
- Correlation,
- Comparison of local histograms
- Spectral properties
- Etc.

PM Reboot

5. Data Matrix and Dissimilarity Matrix

For instances with several attributes, we need to look at all attributes/ dimensions at the same time → change in notation

- **Data matrix**
  - n data points with p dimensions
- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - d(i, j) is the measured dissimilarity or "difference" between objects i and j. In general, d(i, j) is a non-negative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

5. Dissimilarity: Proximity Measure

Nominal attributes (i.j)

- Simple matching
- *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

Binary attributes (i,j)

- Set up contingency table

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

- Measure simple distance

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

PM Reboot

5. Dissimilarity between Binary Variables—Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

5. Distance between Data Points

Euclidian Distance

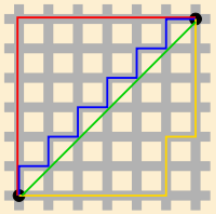
- The "ordinary" (i.e. straight-line) **distance** between two points
- The **Euclidean distance** between points **i** and **j** is the length of the line connecting them

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

Manhattan Distance

- Also called “city block” or taxicab distance
- Calculates the grid lengths between points I and j

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



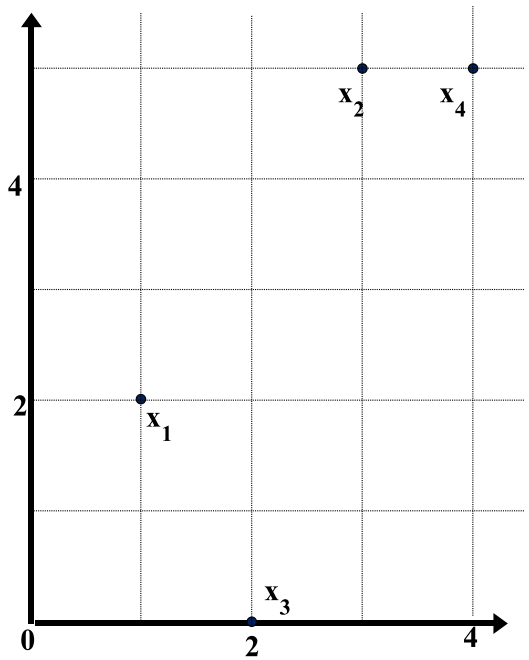
Minkowski Distance

Where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two *p*-dimensional data objects, and *h* is the order)

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

5. Example: Distance Measurements

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Most frequently used Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

5. Dissimilarity for Numeric Data—Example

6. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

   (a) Compute the *Euclidean distance* between the two objects.

   (b) Compute the *Manhattan distance* between the two objects.

   (c) Compute the *Minkowski distance* between the two objects, using $h = 3$.

   (d) Compute the *supremum distance* between the two objects.

Answer:

(a) Compute the *Euclidean distance* between the two objects.
The Euclidean distance is computed using Equation (2.6).
Therefore, we have $\sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45} = 6.7082$.

(b) Compute the *Manhattan distance* between the two objects.
The Manhattan distance is computed using Equation (2.7). Therefore, we have $|22-20| + |1-0| + |42-36| + |10-8| = 11$.

(c) Compute the *Minkowski distance* between the two objects, using $h = 3$.
The Minkowski disance is

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdot + |x_{ip} - x_{jp}|^h} \tag{2.10}$$

where h is a real number such that $h \geq 1$.
Therefore, with $h = 3$, we have $\sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233} = 6.1534$.

(d) Compute the *supremum distance* between the two objects.
The supremum distance is computed using Equation (2.8). Therefore, we have a supremum distance of 6.

5. Cosine Similarity

Often used in search engines to determine if documents have enough words in common to be ranked closely:

How similar are these two sentences?

1. Chris loves me more than Linda loves me
2. Jane likes me more than Chris loves me

We will decide how close these two texts are to each other by calculating the cosine of their angle.

The two vectors are:

s1: [2, 0, 1, 1, 0, 2, 1, 1] s2: [2, 1, 1, 0, 1, 1, 1, 1]

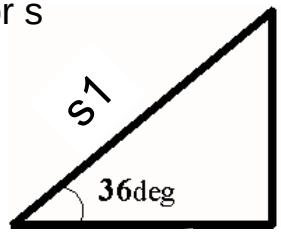
Word	s1	s2
me	2	2
Jane	0	1
Chris	1	1
Linda	1	0
likes	0	1
loves	2	1
more	1	1
than	1	1

8-dimensional

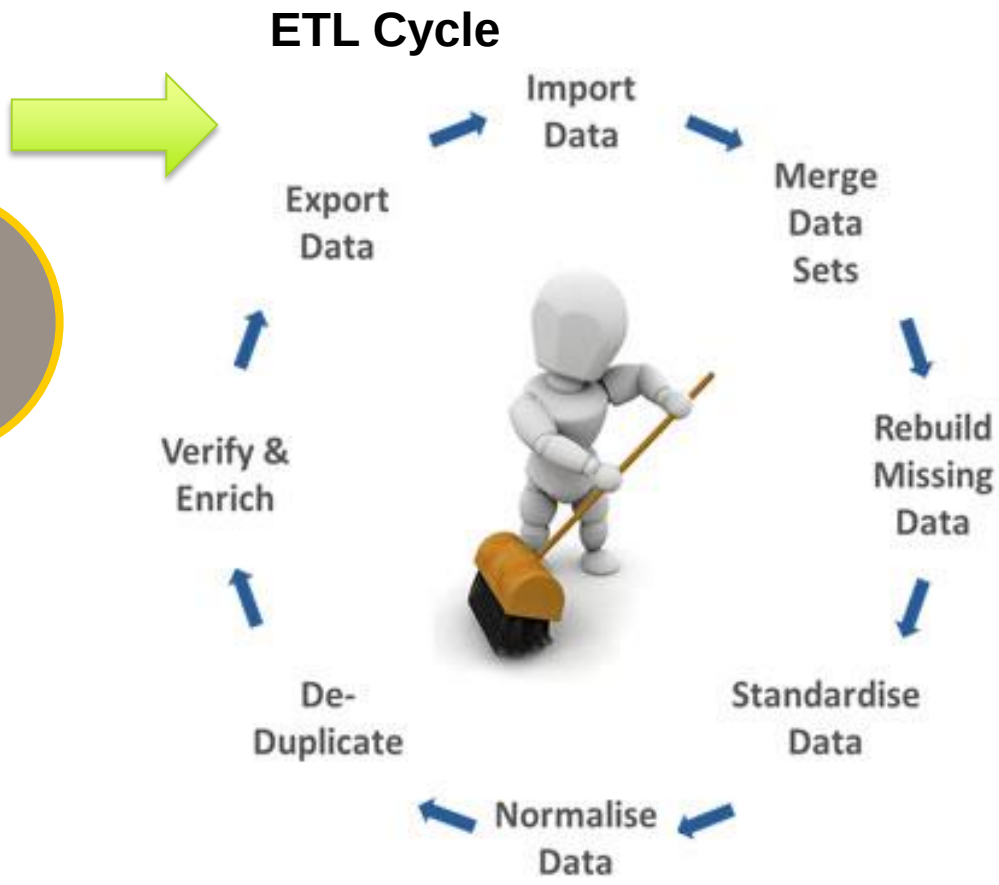
$$\cos(s1, s2) = (s1 \bullet s2) / ||s1||\ ||s2||$$

where ● indicates vector dot product,
||s||: the length of vector s

s1●s2 = 2*2+0*1+1*1+1*0+0*1+2*1+1*1+1*1= 9
||s1||= (2*2+0*0+1*1+1*1+0*0+2*2+1*1+1*1)^0.5=(12)0.5 = 3.46
||s2||= (2*2+1*1+1*1+0*0+1*1+1*1+1*1+1*1)^0.5=(10)0.5 = 3.16
cos(s1, s2) = 9/(3.46*3.16) = 0.823 ➔ 34.6 deg



Data Quality Problems



PM Reboot

Solving Data Quality Problems

Missing Data

- Evaluate if the record is needed at all
- Replace with Mean, Median, or Mode
- Replace with estimated value

Duplicate Data

- Determine original record
- Merge data where possible
- Delete duplicates

Non-standard data

- Determine most logical measurement unit (e.g. pounds or kilograms?)
- Transform data into standardized units (e.g. convert pounds to kilograms
- Often also noted like this:

$$x_{new} = \frac{x - \mu}{\sigma}$$

Unnormalized data

- Scale all numeric variables in the range [0,1]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Summary

Know your data!

- Data Context
- Data Structure
- Attributes and Data Types
- Measures of Central Tendency and Distribution: Mean, Median, Mode, Standard Deviation, Skewedness, Normal Distribution Curve, Proximity of Datapoints

Handle Missing and Incorrect Attribute Values

- Remove tuples
- Substitute with mean, median, or mode

Handle Outliers and Special Cause Data

- Remove tuples
- Substitute with mean, median, or mode

Only if you know your data really well will you be able to generate a good analysis!

QUIZ # 2

Week 2 Quiz

Look at the adult dataset again in R, then pick ONE numeric and ONE nominal attribute.

1. How are the attribute values distributed? Build a boxplot and a histogram for each. You will have to transform the nominal attribute so you can plot its values.
2. How does EACH OF the two attributes relate to the “income” attribute? Build a scatter diagram for EACH relationship and determine which attribute is more closely related..
3. Calculate the Manhattan distance between the FIRST FIVE points of each attribute and the “income” attribute. Which attribute is more closely related to “income”?
4. In MS Excel, build a Pivot chart showing aggregated data for your two attributes and income by country. What can we learn from this roll-up?

Learning Objectives

- ✓ Identify structured or unstructured data?
- ✓ Describe and apply data types to a data set
- ✓ Determine what mathematical operations are possible
- ✓ Analyze data distribution, both mathematically and graphically
- ✓ Analyze data relationships and applying considerations of distance
- ✓ Identify, and plan mitigation of, data quality issues

PM Reboot

DATA WAREHOUSING AND OLAP VS OLTP

PM Reboot

Learning Objectives

- Define data warehouse, OLAP, and OLTP
- Describe data aggregation with data cubes
- Identify basic data warehouse schema architectures
- Build a pivot table in MS Excel

PM Reboot

What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- Data warehousing:
  - The process of constructing and using data warehouses

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

Characteristics of a Data Warehouse

Subject-Oriented

- Organized around major subjects, such as customer, product, sales
- Supports modeling and analysis of data for decision makers, not on transaction processing
- Data selected based on subject and purpose

Integrated

- Constructed by integrating multiple, data sources (relational databases, flat files, on-line transaction records)
- Data cleaning and data integration techniques are applied (consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources)

Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

Nonvolatile

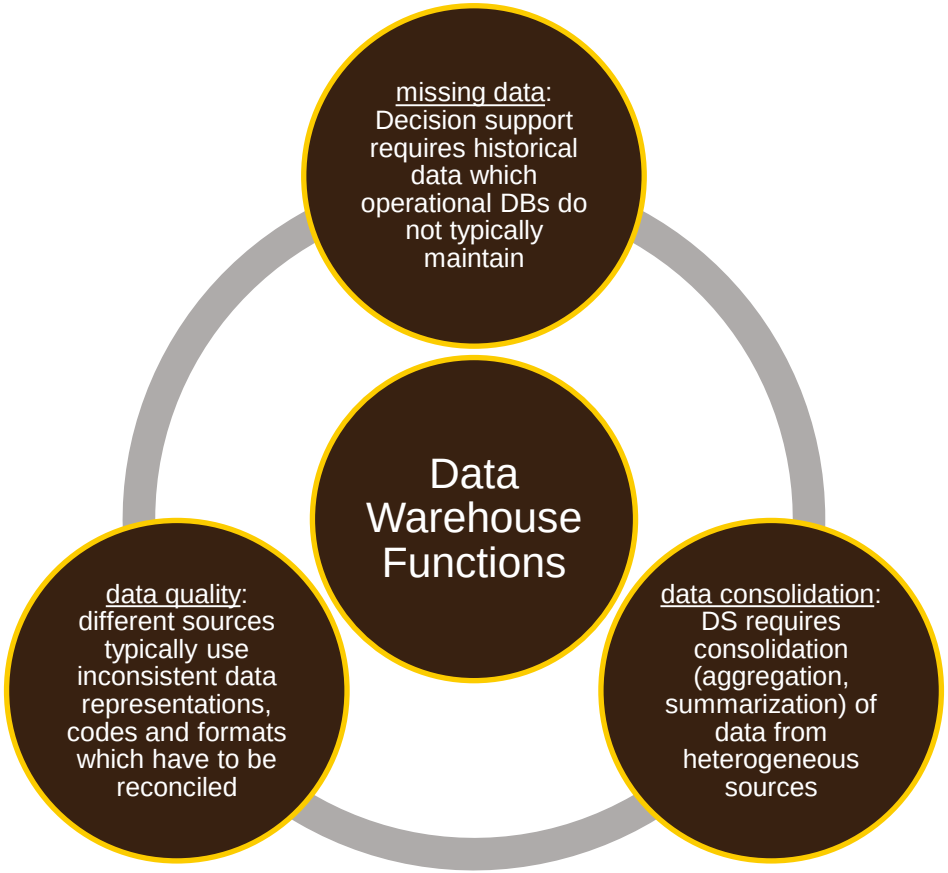
- A physically separate store of data transformed from the operational environment
- No operational data updates (no transaction processing, recovery, and concurrency control mechanisms)—only *initial loading of* and *access to data*

OLTP vs. OLAP

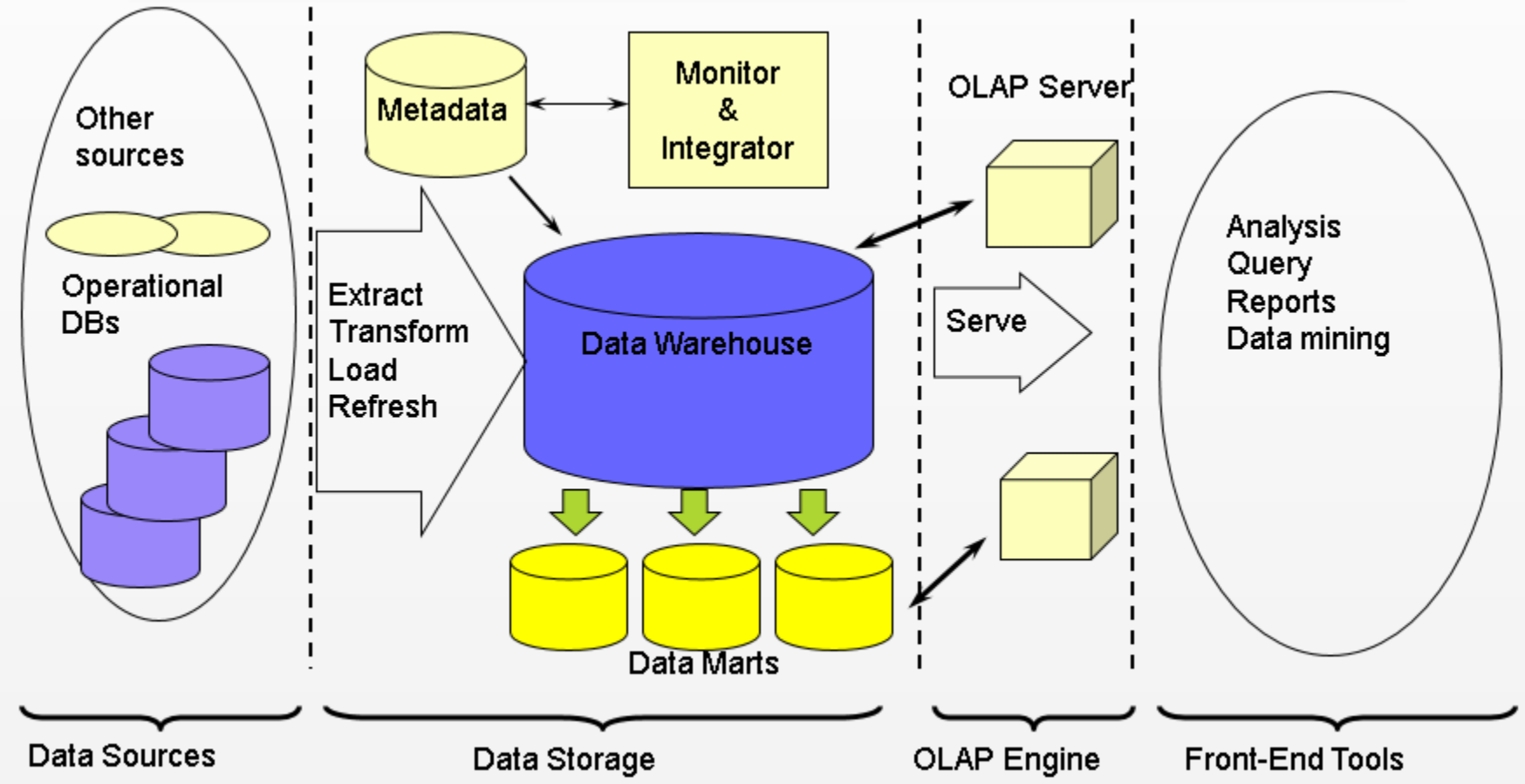
	OLTP (database)	OLAP (data warehouse)
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on primary key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why a Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Note: There are more and more systems which perform OLAP analysis directly on relational databases



Data Warehouse Architecture



Data Warehouse Models

Enterprise Warehouse

- All of the information about subjects spanning the entire organization
- Corporate-wide data integration, detailed & summarized
- From a few GB to exabytes
- On mainframes, superservers, or cluster architectures.
- Requires extensive business modeling and may take years to design and build.

Data Mart

- Contains a subset of corporate-wide data for a specific group of users.
- Scope confined to specific selected subjects (e.g. marketing);
- Usually on LINUX- or Windows servers
- Implementation takes weeks, but may involve complex integration in the long run

Virtual Warehouse

- A set of views over operational databases.
- For efficient query processing, only some of the possible summary views may be materialized.
- Easy to build but requires excess capacity on operational database servers.

All warehouses contain METADATA:

- **Descriptive:** Data warehouse structure
- **Operational:** Data historyusage statistics, error reports, audit trails
- **Algorithms** used for summarization
- **Mapping** from operational environment to the data warehouse
- **Business: T**erms and definitions, ownership of data, charging policies

Building a Data Warehouse: Extraction, Transformation, Loading (ETL)

Data Extraction

- Get data from multiple, heterogeneous, and external sources

Data Cleaning

- Detect errors in the data and rectify them when possible

Data Transformation

- Convert data from legacy or host format to warehouse format

Load

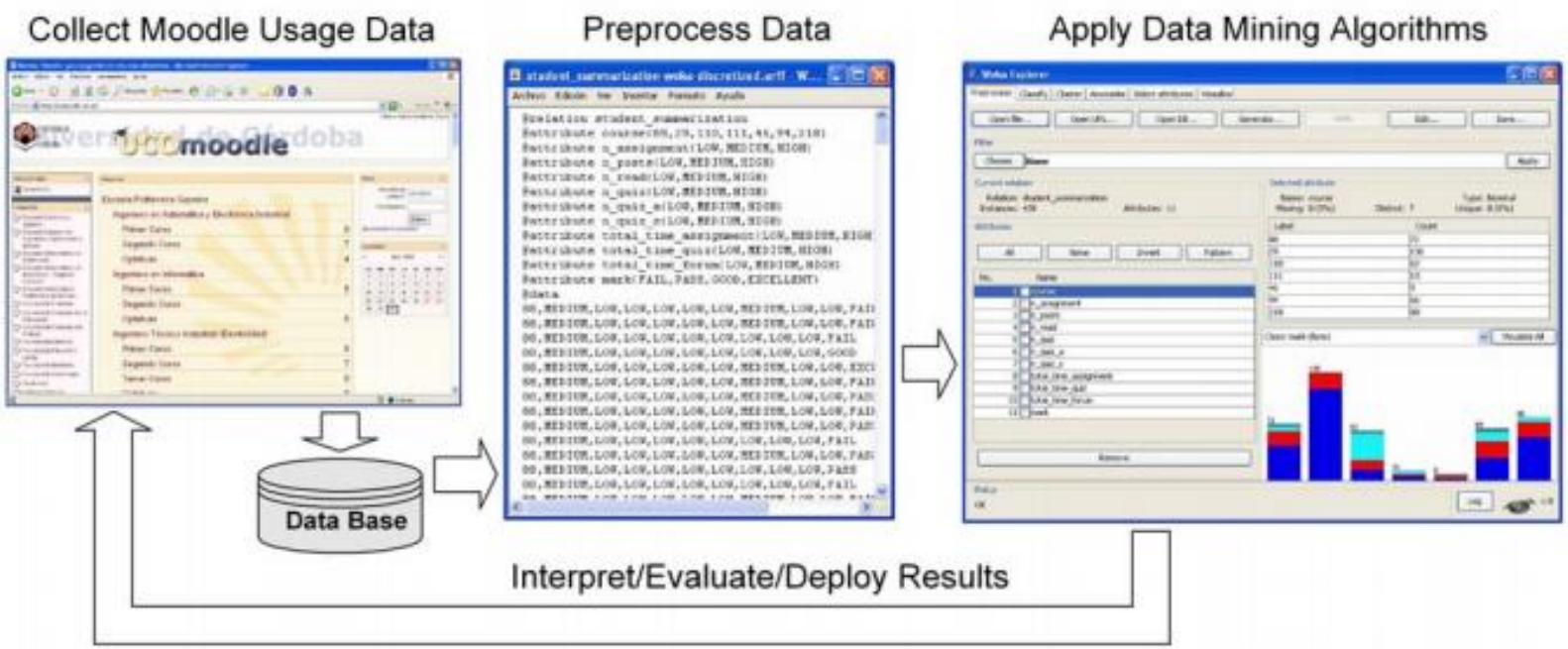
- Sort, summarize, consolidate, compute views, check integrity, and build indicies and partitions

Refresh

- Propagate the updates from the data sources to the warehouse

DATA WAREHOUSING Basics

Example: Education Data



Romero ,Cristóbal, Sebastián Ventura, Enrique García (2007) "Data mining in course management systems: Moodle case study and Tutorial." Elsevier Science.

From Tables and Spreadsheets to Data Cubes

- A **data warehouse** is based on a **multidimensional data model** which views data in the form of a data cube
- We typically think of a cube as 3-dimensional; in data mining, cubes are n-dimensional (or n-D)
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Example: From 2D Data File to 3D Data Cube—Sales Data

Table 4.2 2-D View of Sales Data for *AllElectronics* According to *time* and *item*

	location = *"Vancouver"*			
	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

Note: The sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

2 dimensions = TABLE:

- Item Type
- Time (Sales Quarter)

+ same tables for Chicago, New York, Toronto

3 dimensions = CUBOID:

- Location
- Item Type
- Time (Sales Quarter)

Table 4.3 3-D View of Sales Data for *AllElectronics* According to *time, item, location*

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).

Han, Kamber 3rd ed. 137

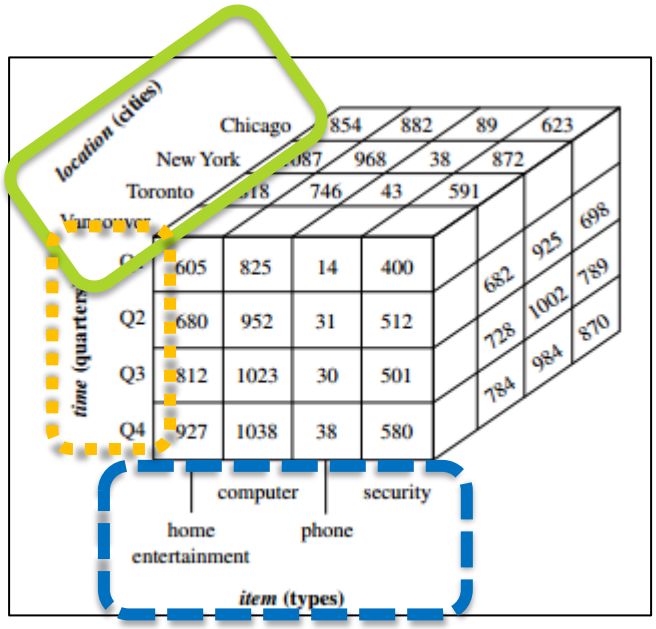
PM Reboot

Example: 3D Data Cube—Sales Data

Table 4.3 3-D View of Sales Data for *AllElectronics* According to *time*, *item*, and *location*

	location = "Chicago"				*location* = "New York"				*location* = "Toronto"				*location* = "Vancouver"			
	item				*item*				*item*				*item*			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

Note: The measure displayed is *dollars_sold* (in thousands).



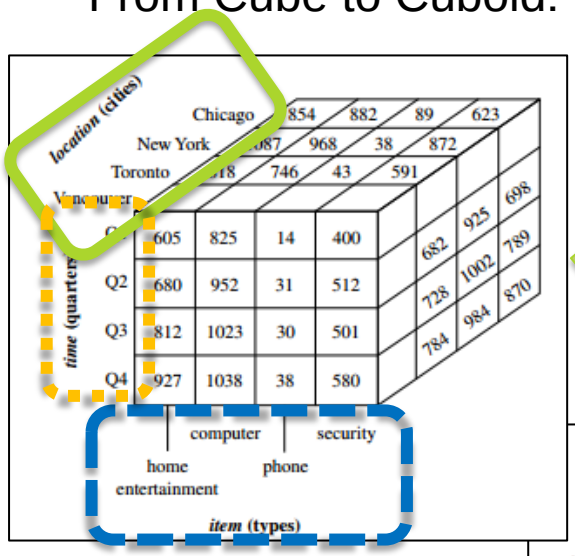
3 dimensions = CUBOID:

- Location
- Item Type
- Time (Sales Quarter)

Han, Kamber 3rd ed. 137

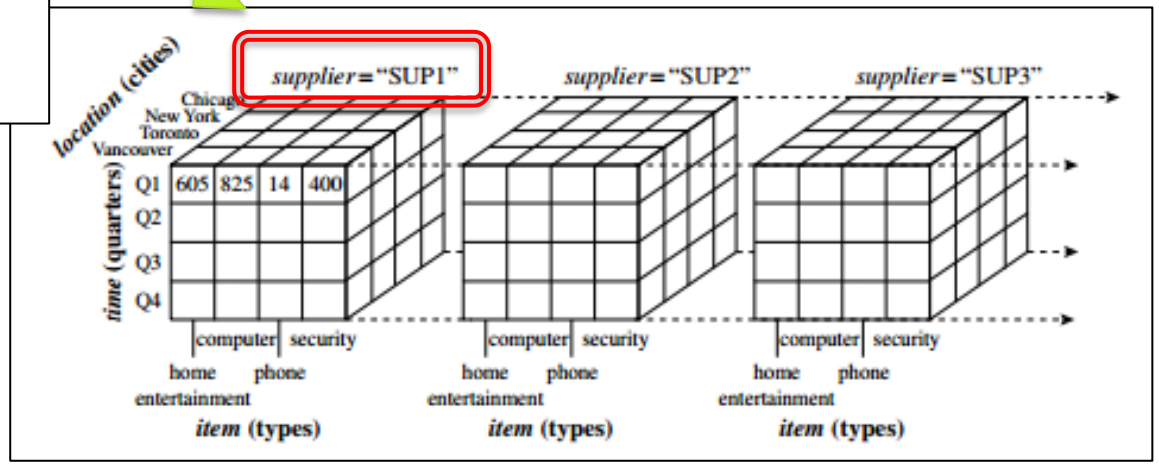
From 3D Cube To 4D Lattice

From Cube to Cuboid: Adding “Supplier” as a 4th dimension



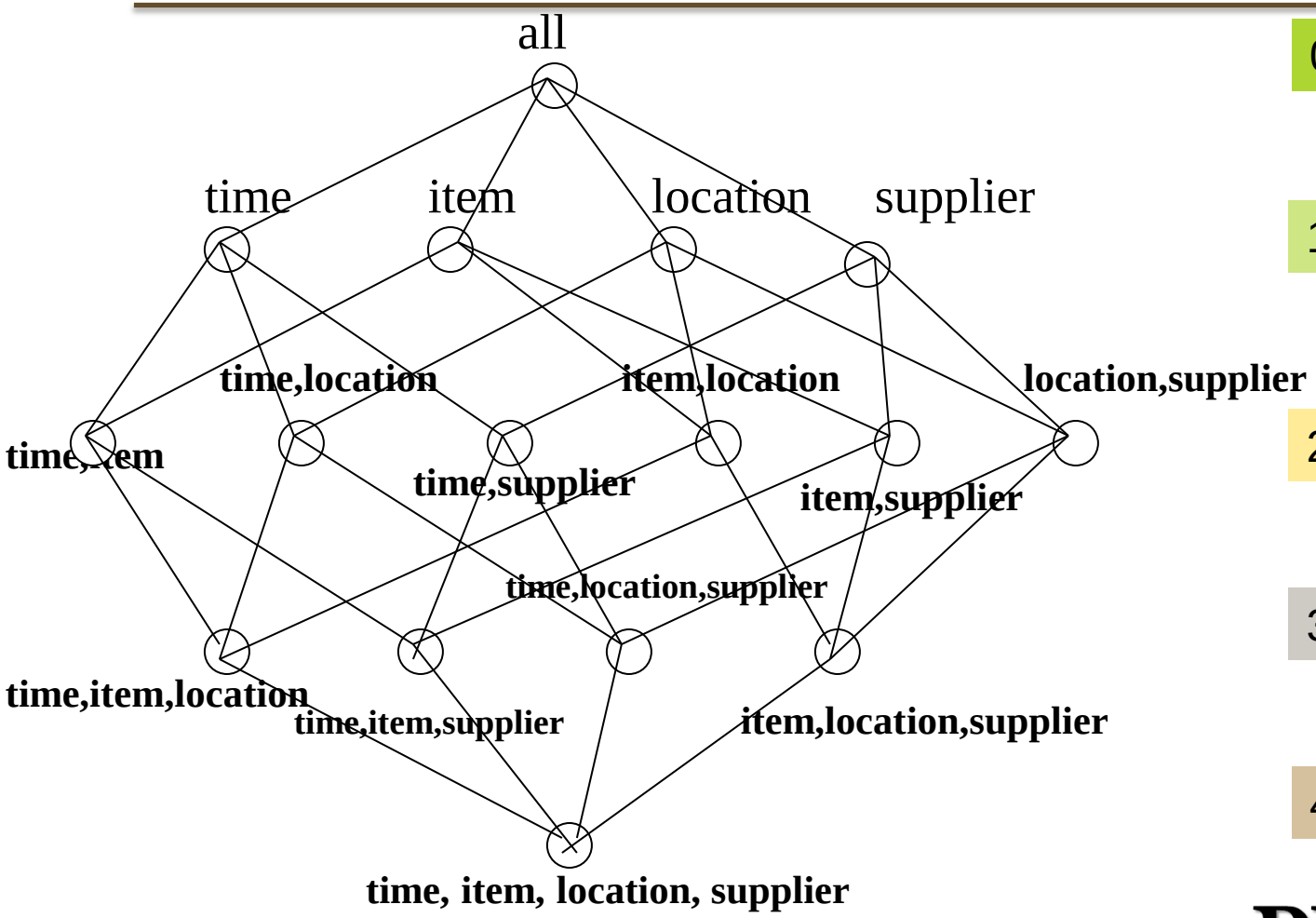
Now we need to draw several cubes—but they are all connected!

➔ We need to draw a lattice of **cuboid**s!



Han, Kamber 3rd ed. 139

Cube: A Lattice of Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

Typical OLAP Operations for Multidimensional Data

Roll up (drill-up): summarize data ("generalize")

- *by climbing up hierarchy or by dimension reduction*

Other operations

- *drill across: involving (across) more than one fact table*
- *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

Drill down (roll down): reverse of roll-up ("specialize")

- *from higher level summary to lower level summary or detailed data, or introducing new dimensions*

Pivot (rotate):

- *reorient the cube, visualization, 3D to series of 2D planes*

Slice and dice:

- *project and select*

EXAMPLE: Data Summarization

A pivot table is a data summarization tool found in data visualization programs such as spreadsheets or business intelligence software.

Example Pivot Table

Example Pivot Tables 2

Online Pivot Table 1

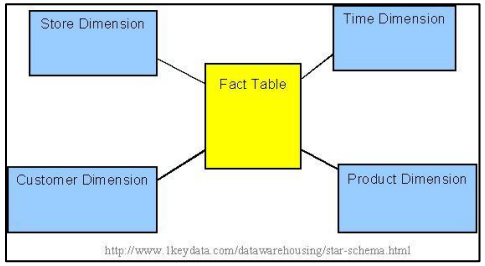
Online Pivot Table 2

Open any of these files and investigate the data.

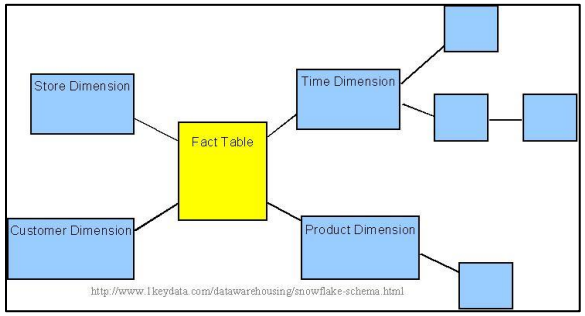
1. What kinds of data do you see?
2. How is the data structured?
3. Which attributes can be aggregated to provide a summary?
4. How do you aggregate the attributes?

Conceptual Modeling of Data Warehouses

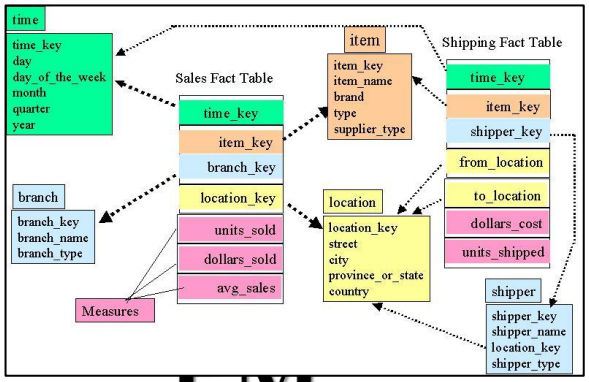
- **Star schema**: A fact table in the middle connected to a set of dimension tables



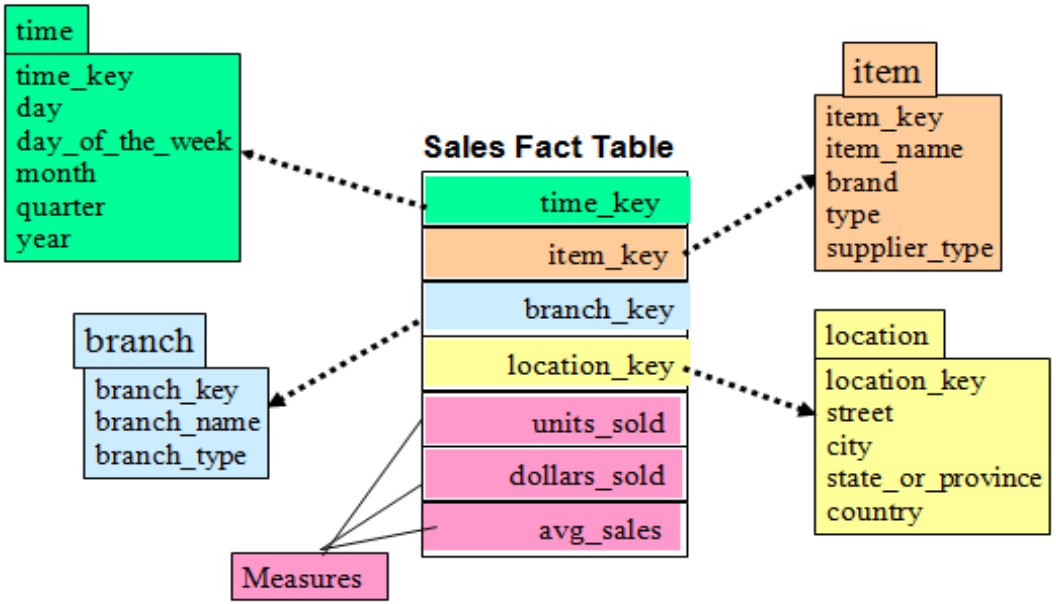
- **Snowflake schema**: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake



- **Fact constellation schema**: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation



Example of Star Schema



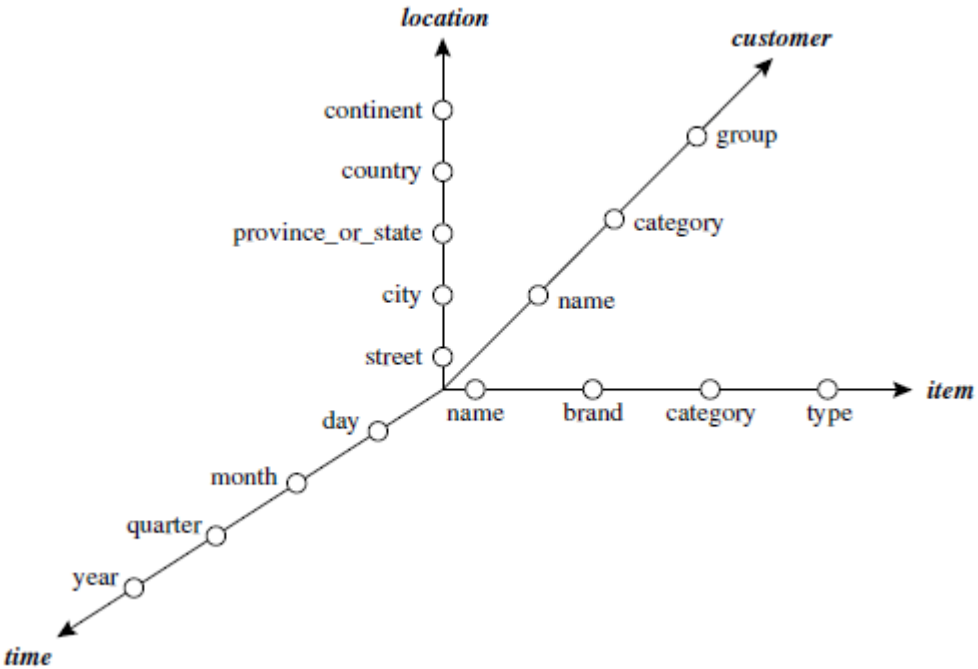
- Four dimensions: time, item, branch, and location
- Each dimension is represented by only one table, and each table contains a set of attributes. May introduce redundancy (e.g., where two cities are in the same state)
- Could use further normalization

Cube query

```
select s.time key, s.item key, s.branch key, s.location
key, sum(s.number of units sold  s.price), sum(s.number
of units sold)
from time t, item i, branch b, location l, sales s,
where s.time key = t.time key and s.item key = i.item key
and s.branch key = b.branch key and s.location key =
l.location key
group by s.time key, s.item key, s.branch key, s.location
key
```

A Star-Net Query Model

A star net model consists of radial lines emanating from a central point, where each line represents a concept hierarchy for a dimension.

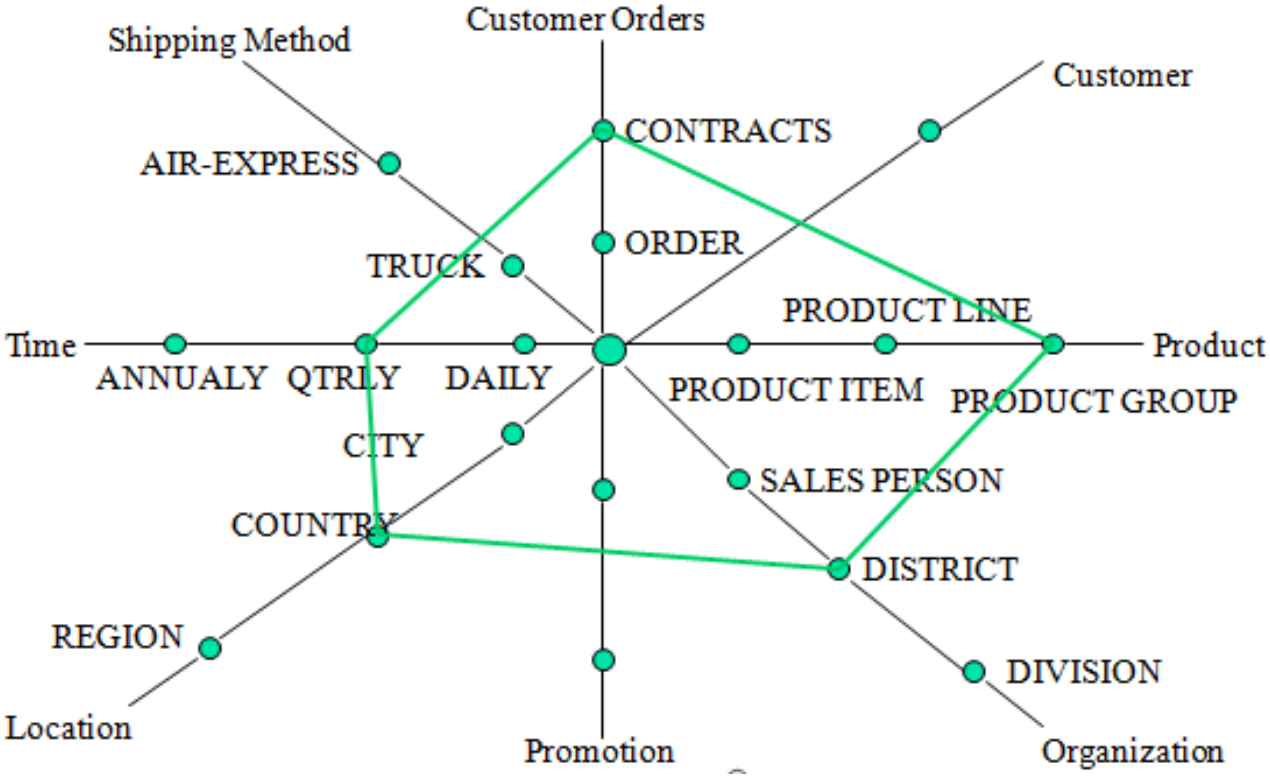


Each abstraction level in the hierarchy is called a footprint. These represent the granularities available for use by OLAP operations such as drill-down and roll-up.

Han/ Kamber, 3rd edition, 149

A Star-Net Query Model (2)

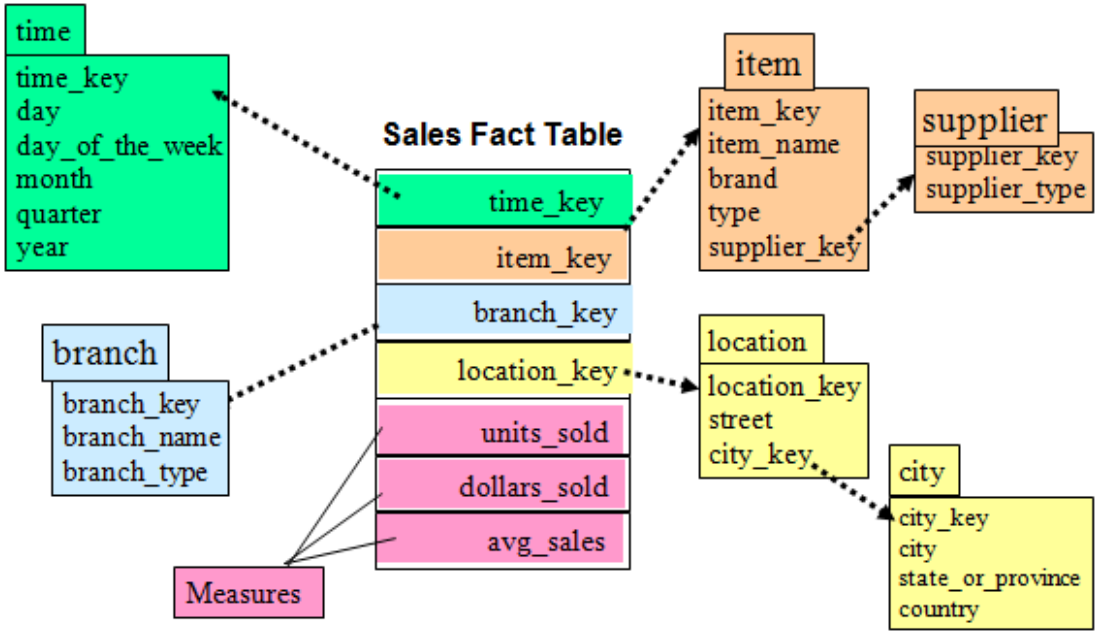
Star net models can be complex.



Example of Snowflake Schema

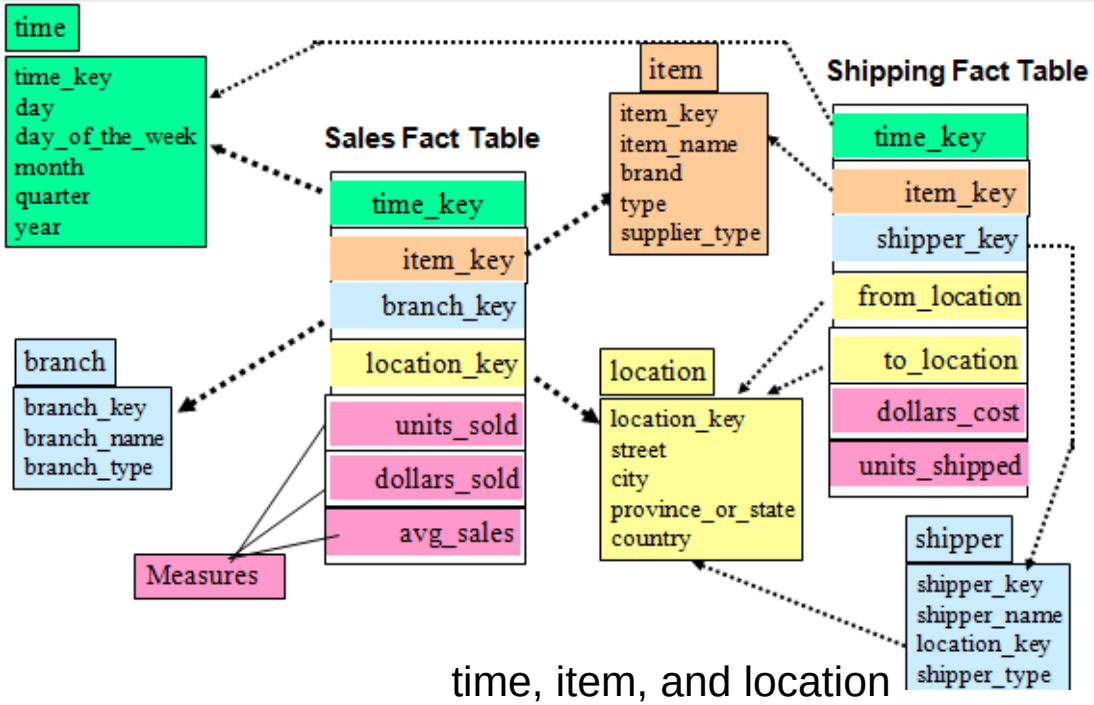
Normalized dimension tables.

- Formerly single dimension table for “item” is normalized ➔ new item and supplier tables.
- “item” dimension table now contains item key, item name, brand, type, and supplier key, where supplier key is linked to supplier dimension table, containing supplier key and supplier type information.
- Single dimension table for location in star schema normalized into two new tables: location and city. The city key in the new location table links to the city dimension.



Example of Fact Constellation

A fact constellation schema allows dimension tables to be shared between fact tables.



5 dimensions, or keys: item key, time key, shipper key, from location, and to location, and two measures: dollars cost and units shipped

time, item, and location are shared between both the sales and shipping fact tables.

Design of Data Warehouse: A Business Analysis Framework

- Four views regarding the design of a data warehouse

Top-down view	Data source view	Data warehouse view	Business query view
• allows selection of the relevant information necessary for the data warehouse	• exposes the information being captured, stored, and managed by operational systems	• consists of fact tables and dimension tables	• sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse Design Processes

- **Top-down, bottom-up approaches or a combination** of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- **From software engineering point of view**
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- **Typical data warehouse design process**
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

From (OLAP) to On Line Analytical Mining (OLAM)

- Why online analytical mining?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks

Efficient Data Cube Computation

- Data cube can be viewed as a lattice of cuboids
  - The bottom-most cuboid is the base cuboid
  - The top-most cuboid (apex) contains only one cell
  - How many cuboids in an n-dimensional cube with L levels?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

L_i is the number of levels associated with dimension i. 1 is added to L_i to include the *virtual top level, all. E.g.,* If the cube has 10 dimensions and each dimension has 5 levels (including all), the total number of cuboids that can be generated is $5^{10} \approx 9.8 * 10^6$.

- Materialization of data cube
  - Materialize <u>every</u> (cuboid) (**full materialization**), <u>none</u> (**no materialization**), or <u>some (**partial materialization**)</u>
  - Selection of which cuboids to materialize
    - Based on size, sharing, access frequency, etc.

The “Compute Cube” Operator

- Cube definition and computation in DMQL

  define cube sales [item, city, year]: sum (sales_in_dollars)

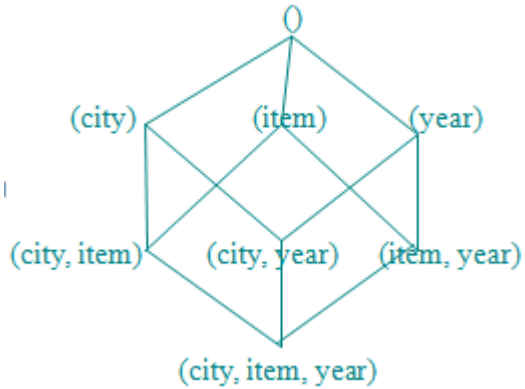
  compute cube sales

- Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al.’96)

  SELECT item, city, year, SUM (amount)

  FROM SALES

  CUBE BY item, city, year



- Need to compute the following Group-Bys

  (date, product, customer),
  (date,product),(date, customer), (product, customer),
  (date), (product), (customer)
  ()

Efficient Processing OLAP Queries

- **Determine which operations** should be performed on the available cuboids
  - Transform drill, roll, etc. into corresponding SQL and/or OLAP operations, e.g., dice = selection + projection
- **Determine which materialized cuboid(s)** should be selected for OLAP op.
  - Let the query to be processed be on {*brand, province_or_state*} with the condition "*year = 2004*", and there are 4 materialized cuboids available:

    1) {*year, item_name, city*}

    2) {*year, brand, country*}

    3) {*year, brand, province_or_state*}

    4) {*item_name, province_or_state*} where *year = 2004*

    Which should be selected to process the query?
- Explore indexing structures and compressed vs. dense array structures in MOLAP

OLAP Server Architectures

- Relational OLAP (ROLAP)
  - Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware
  - Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services
  - Greater scalability
- Multidimensional OLAP (MOLAP)
  - Sparse array-based multidimensional storage engine
  - Fast indexing to pre-computed summarized data
- Hybrid OLAP (HOLAP) (e.g., Microsoft SQLServer)
  - Flexibility, e.g., low level: relational, high-level: array
- Specialized SQL servers (e.g., Redbricks)
  - Specialized support for SQL queries over star/snowflake schemas

Summary

- Data warehousing: A multi-dimensional model of a data warehouse
  - A data cube consists of *dimensions* & *measures*
  - Star schema, snowflake schema, fact constellations
  - OLAP operations: drilling, rolling, slicing, dicing and pivoting
- Data Warehouse Architecture, Design, and Usage
  - Multi-tiered architecture
  - Business analysis design framework
  - Information processing, analytical processing, data mining, OLAM (Online Analytical Mining)
- Implementation: Efficient computation of data cubes
  - Partial vs. full vs. no materialization
  - Indexing OALP data: Bitmap index and join index
  - OLAP query processing
  - OLAP servers: ROLAP, MOLAP, HOLAP

QUIZ # 2

Week 2 Quiz

Look at the adult dataset again in R, then pick ONE numeric and ONE nominal attribute.

1. How are the attribute values distributed? Build a boxplot and a histogram for each. You will have to transform the nominal attribute so you can plot its values.
2. How does EACH OF the two attributes relate to the “income” attribute? Build a scatter diagram for EACH relationship and determine which attribute is more closely related..
3. Calculate the Manhattan distance between the FIRST FIVE points of each attribute and the “income” attribute. Which attribute is more closely related to “income”?
4. In MS Excel, build a Pivot chart showing aggregated data for your two attributes and income by country. What can we learn from this roll-up?

Learning Objectives

- ✓ Define data warehouse, OLAP, and OLTP
- ✓ Describe data aggregation with data cubes
- ✓ Identify basic data warehouse schema architectures
- ✓ Build a pivot table in MS Excel

PM Reboot

BASIC STATISTICS WITH R AND R STUDIO

PM Reboot

Basic Statistics with R and R Studio

Learning Objectives:

Upon completion of the module, you will be able to :

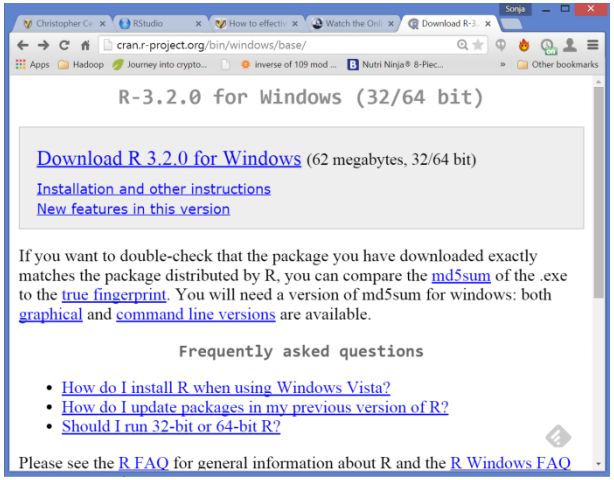
- ❑ Install R and R Studio
- ❑ Review the bundled datasets
- ❑ Learn how to view the Frequency Distribution
- ❑ Learn the commands to produce the Five Number Summary and other important metrics
- ❑ Learn how to build a histogram and a boxplot in R

Required:

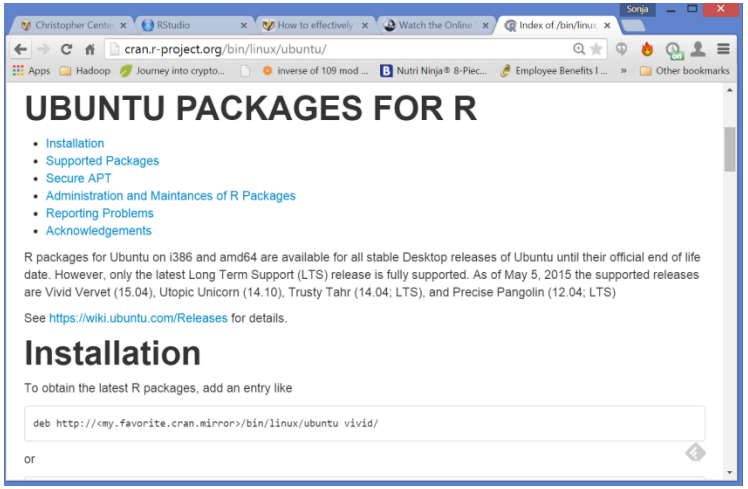
- R and R Studio, downloadable for FREE

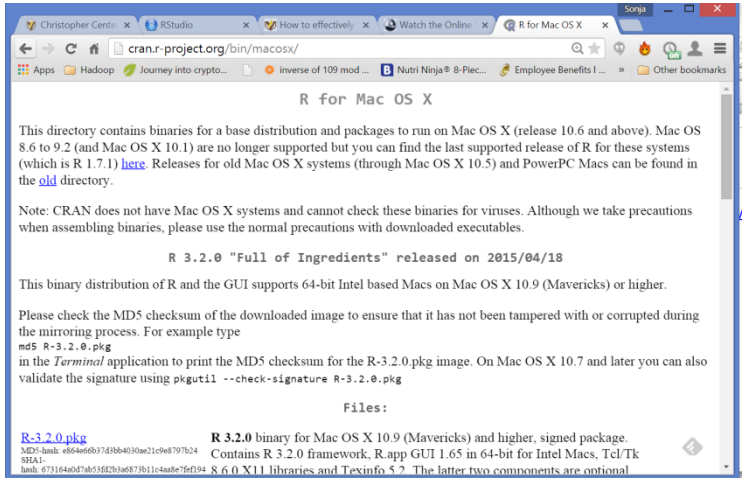
Installing R

http://cran.r-project.org/bin/windows/base/



http://cran.r-project.org/bin/linux/ubuntu/

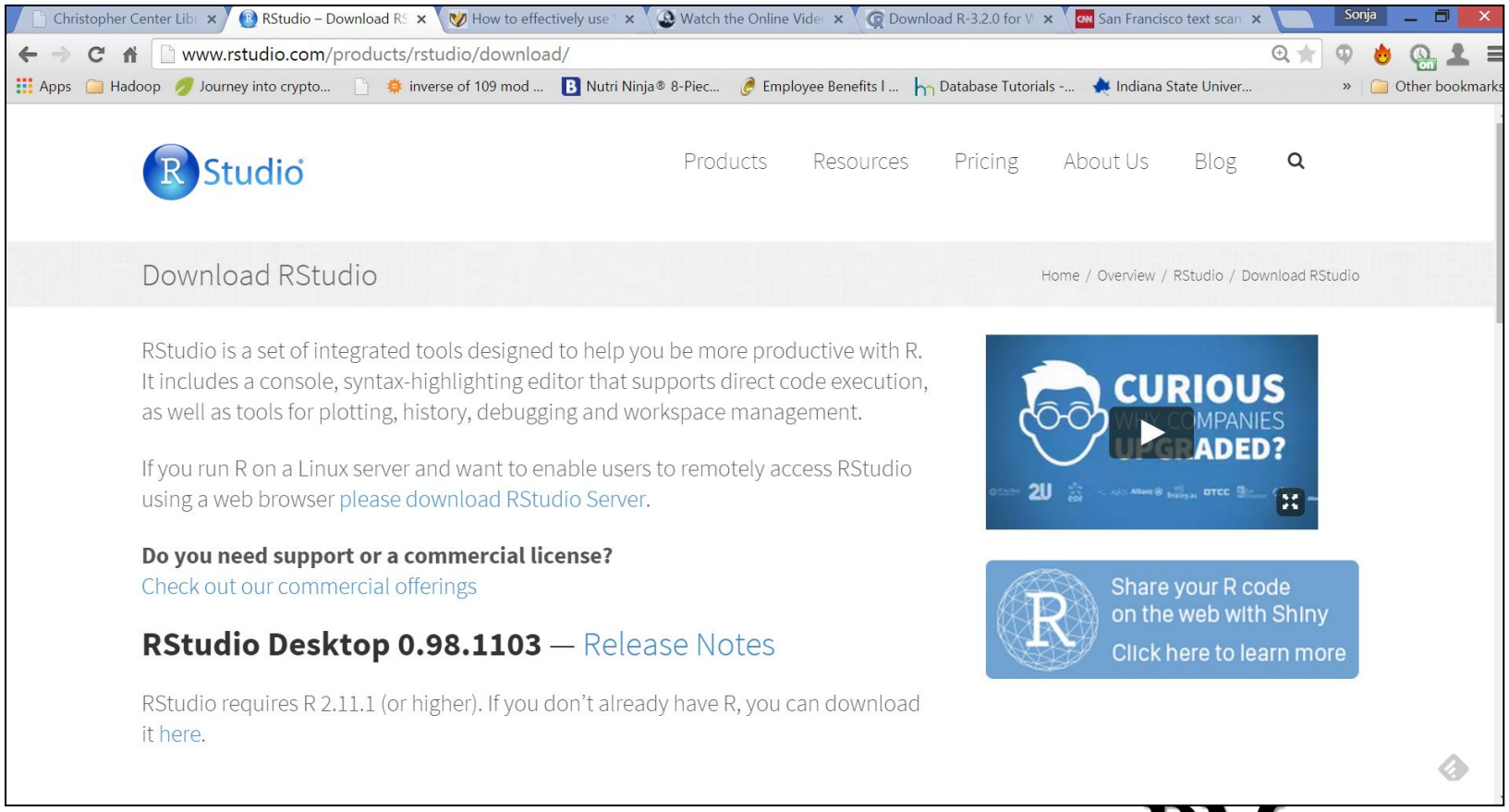




http://cran.r-project.org/bin/macosx/

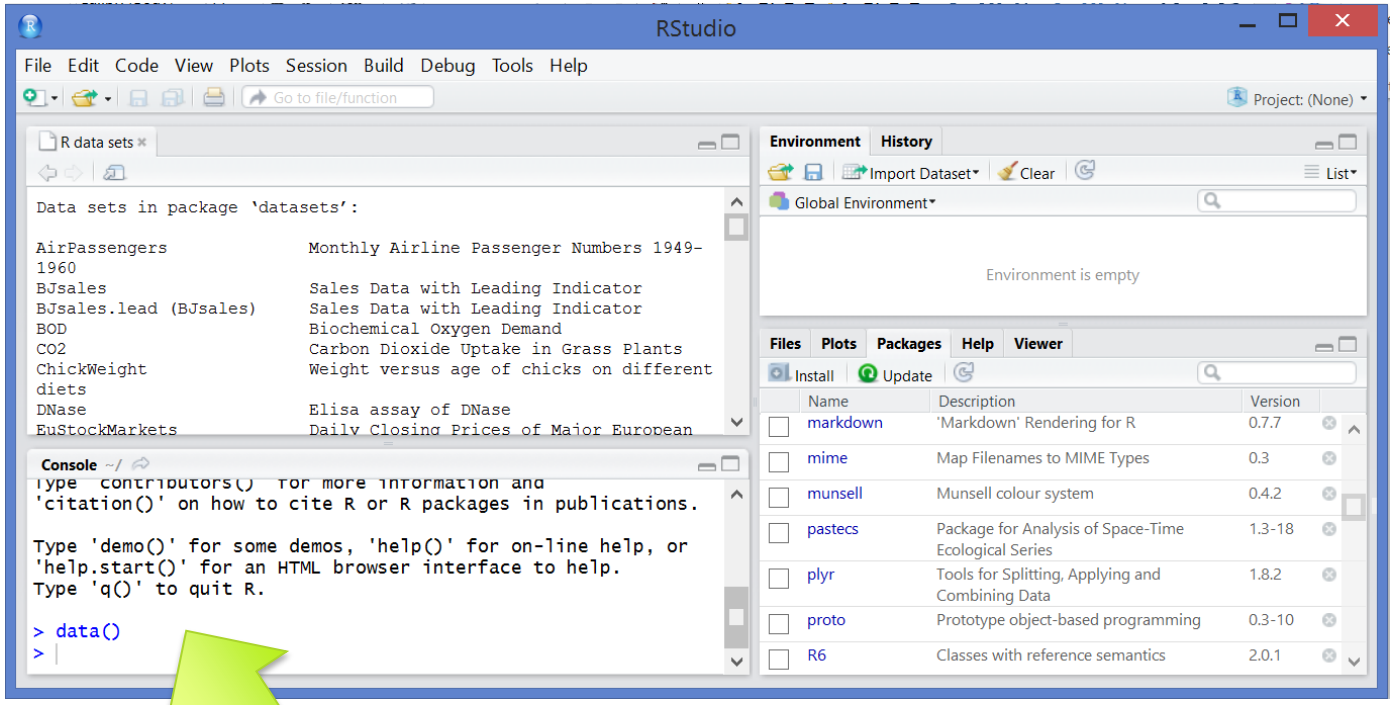
Installing R Studio

http://www.rstudio.com/products/rstudio/download/



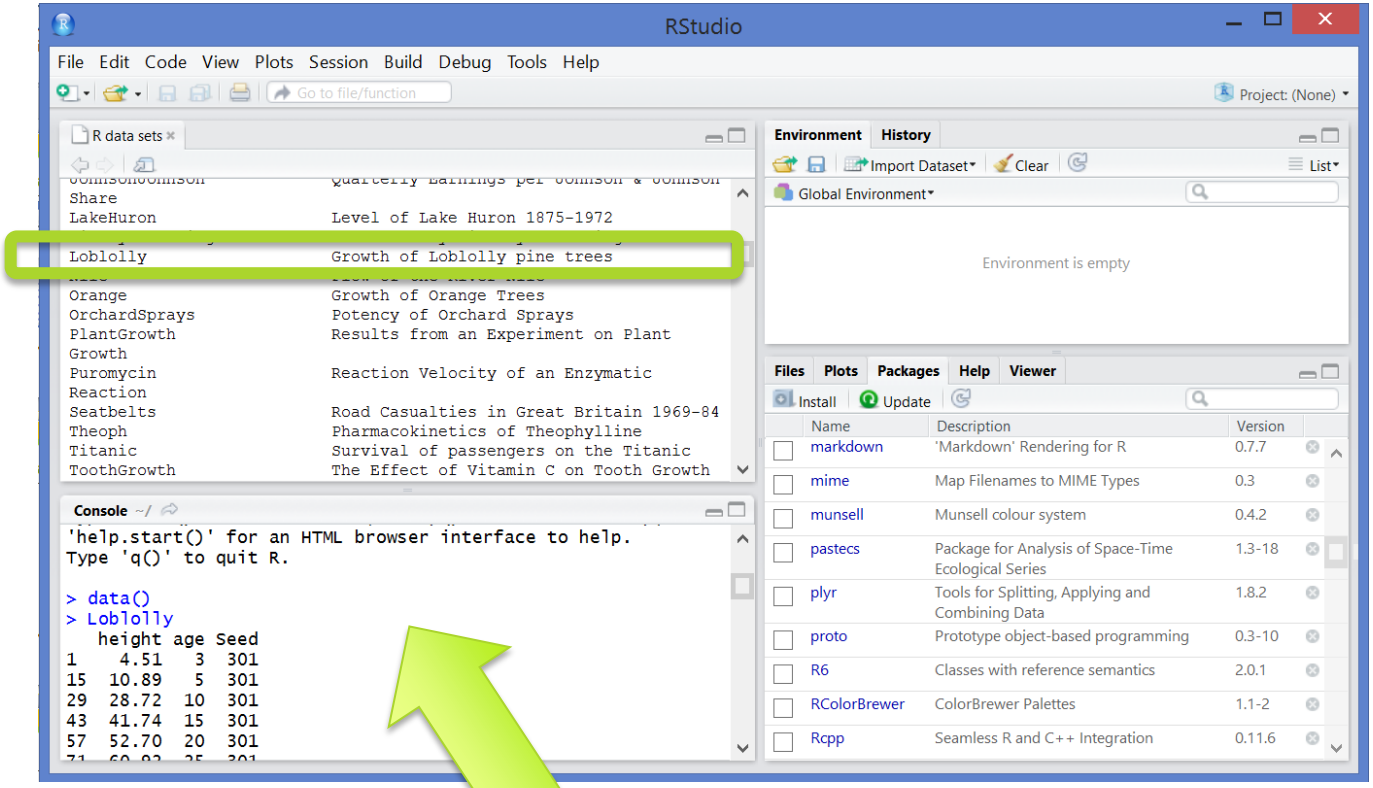
PM Reboot

Reviewing the Datasets



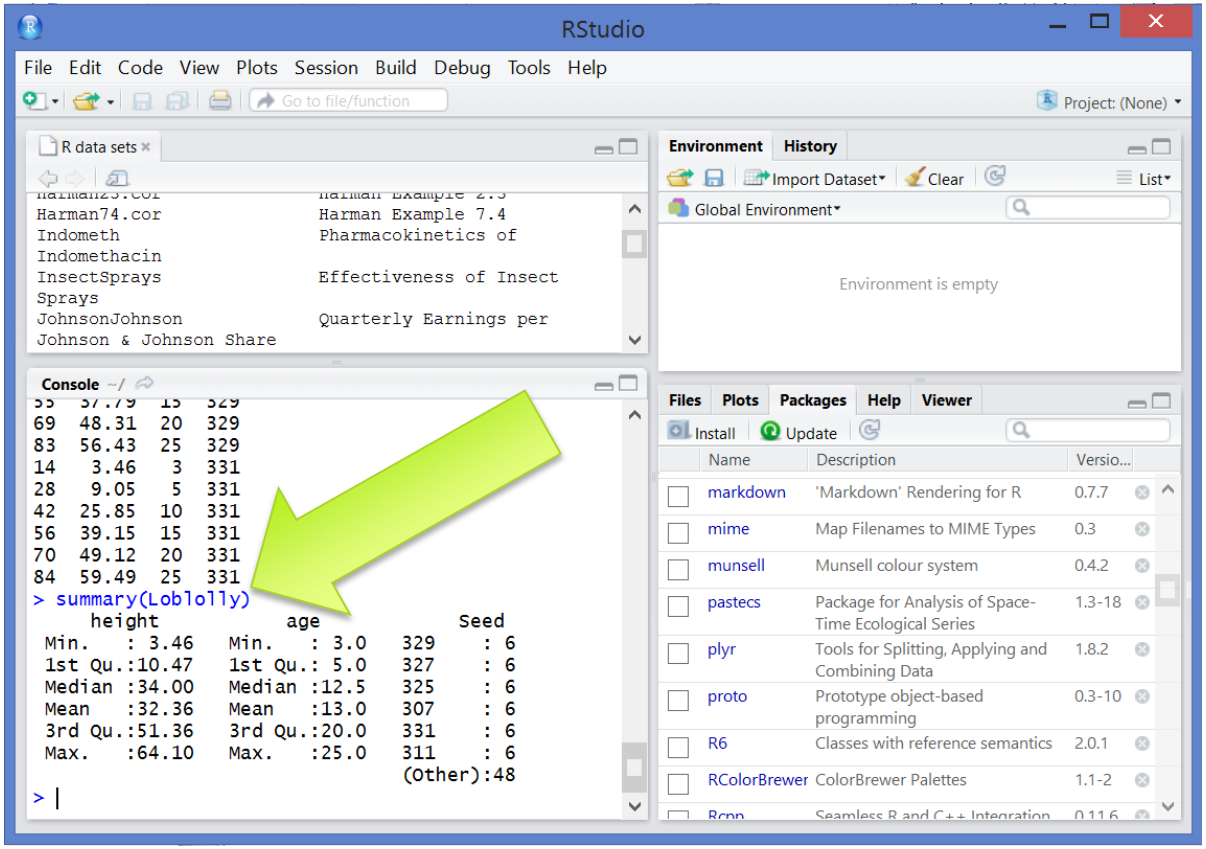
At the prompt, type `data()`

Reviewing the Datasets



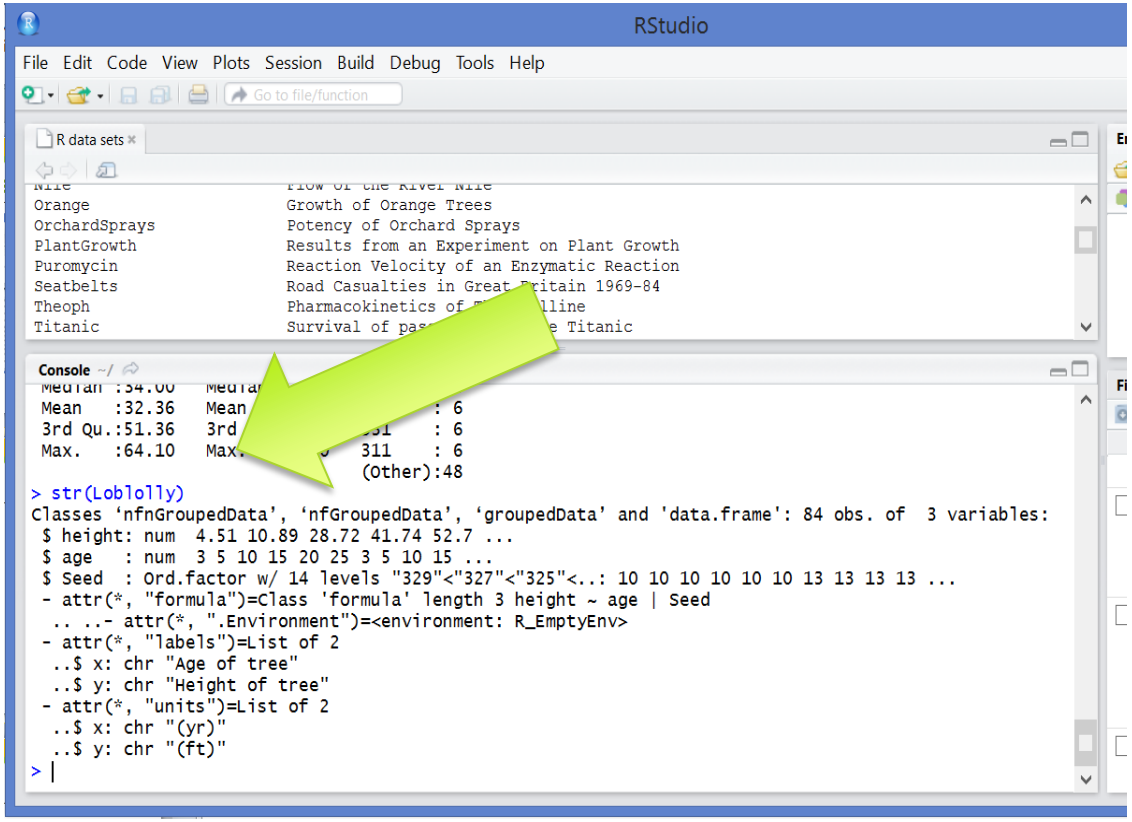
At the prompt, type the name of the built-in dataset with which you want to work.

The Five Number Summary



To produce the Five Number Summary, type `summary(<dataset>)` at the command prompt.

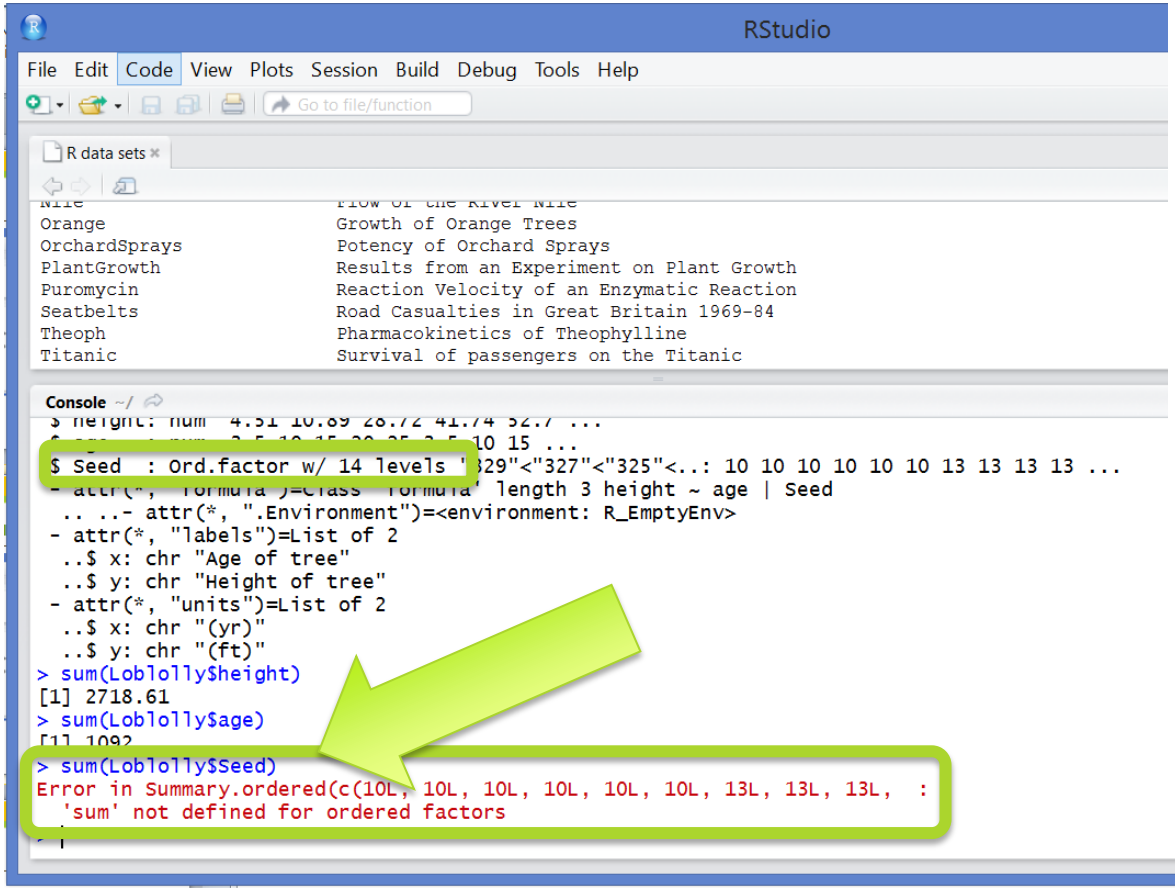
The Structure of a Dataset



To inspect the structure of a dataset, type `str(<dataset>)` at the command prompt.

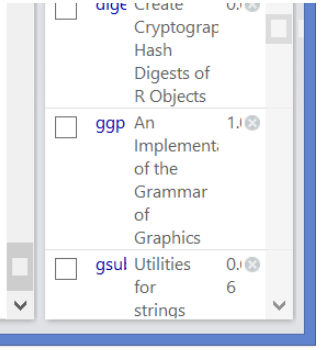
- number of attributes
- names of attributes
- number of observations
- variable type (ie. number, factor, character, etc)

The Sum of a Column

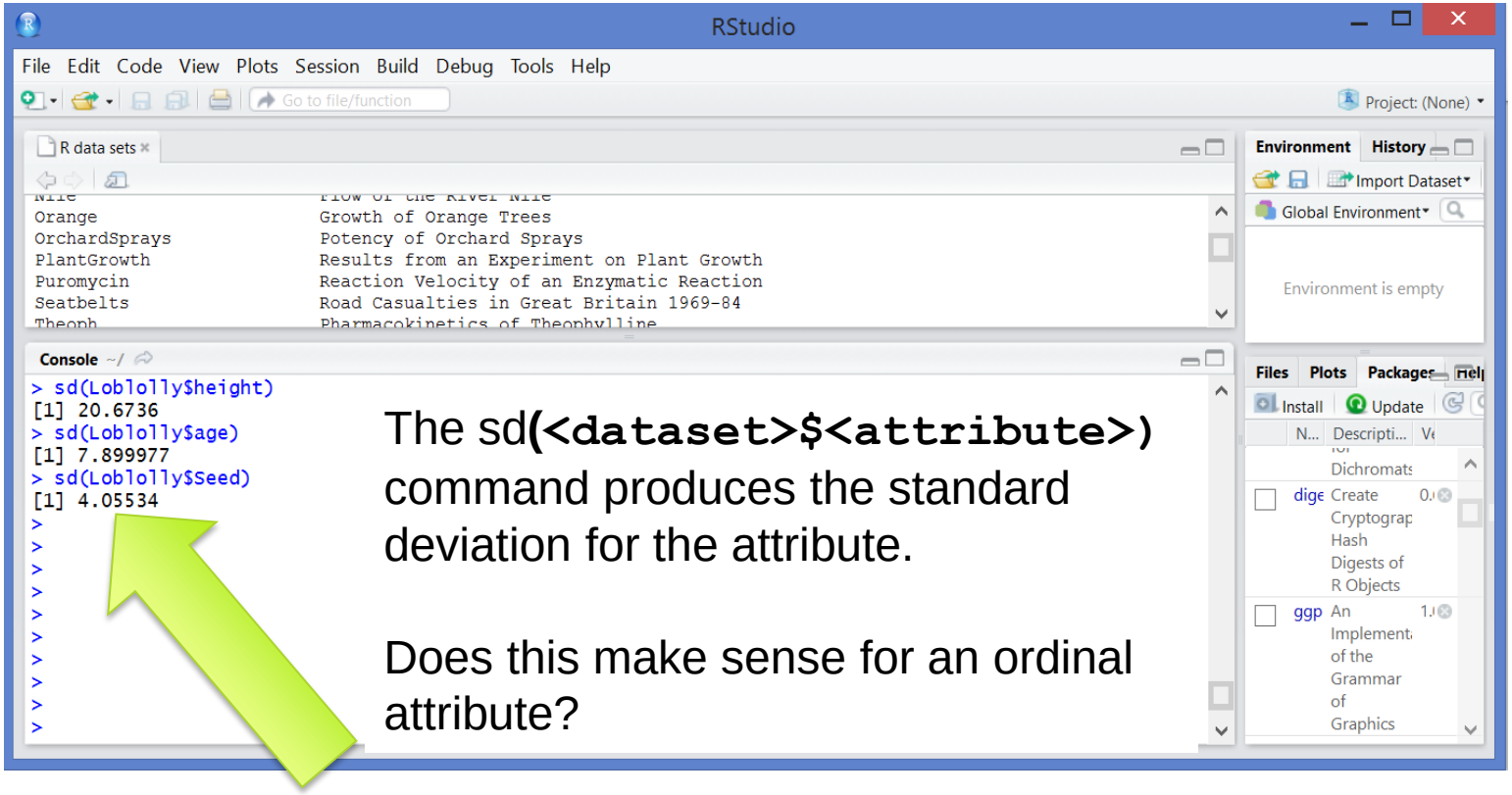


R has a command to compute the sum of a column of a data set: **`>sum(<dataset>$<attribute>)`**

This command works only with numeric attributes!



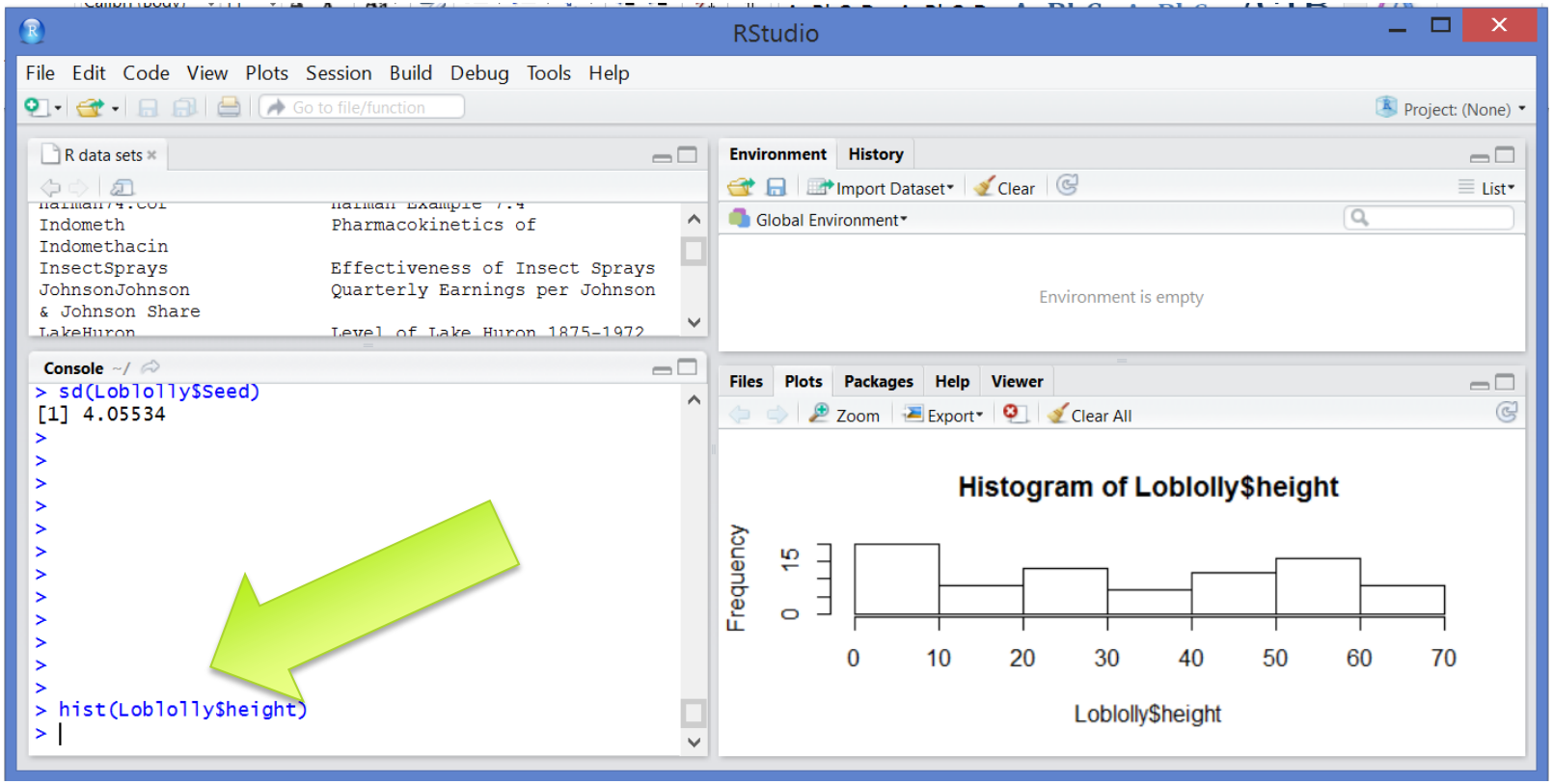
Standard Deviation of an Attribute



The sd(**<dataset>$<attribute>**) command produces the standard deviation for the attribute.

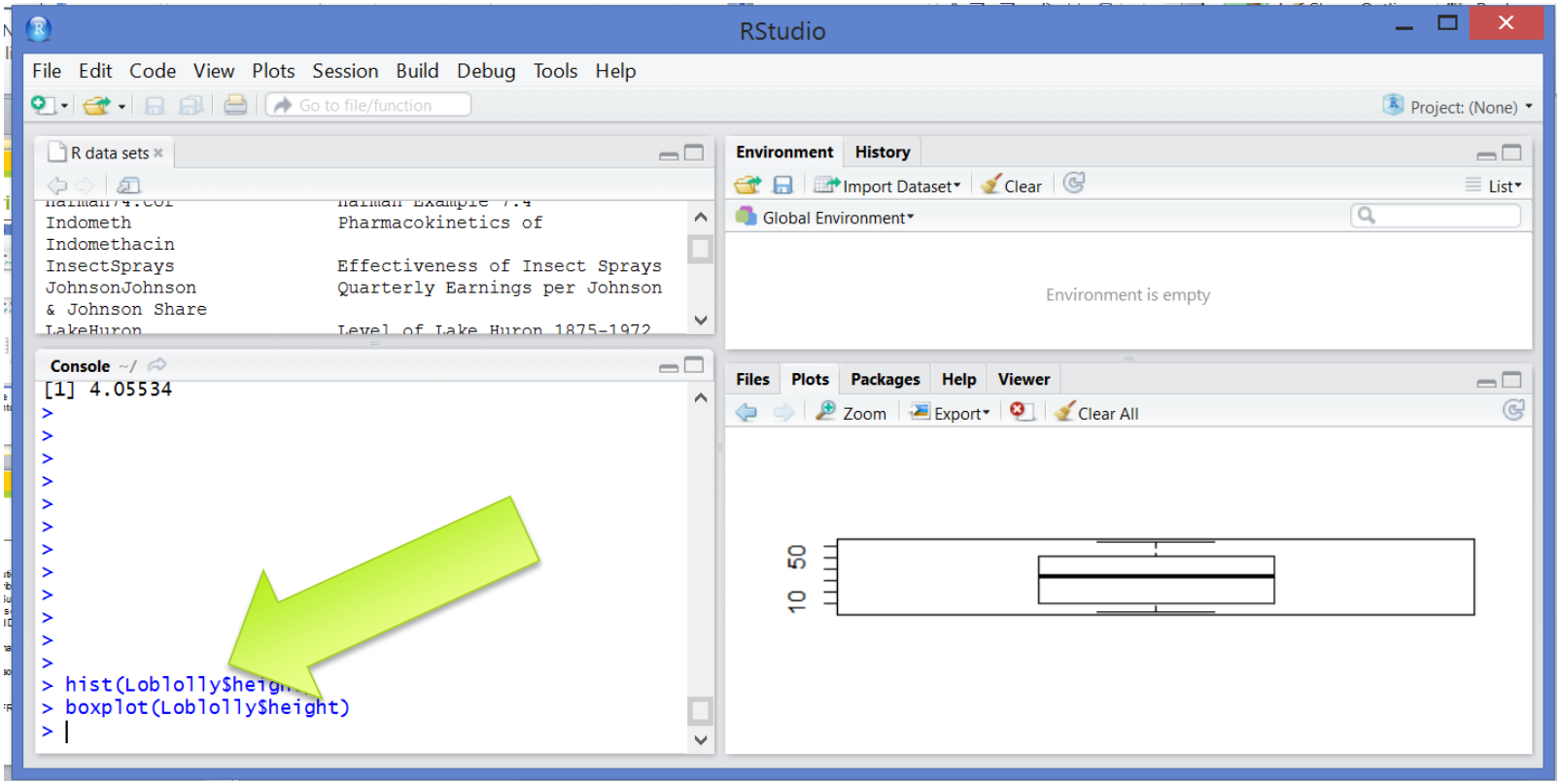
Does this make sense for an ordinal attribute?

Histogram of an Attribute



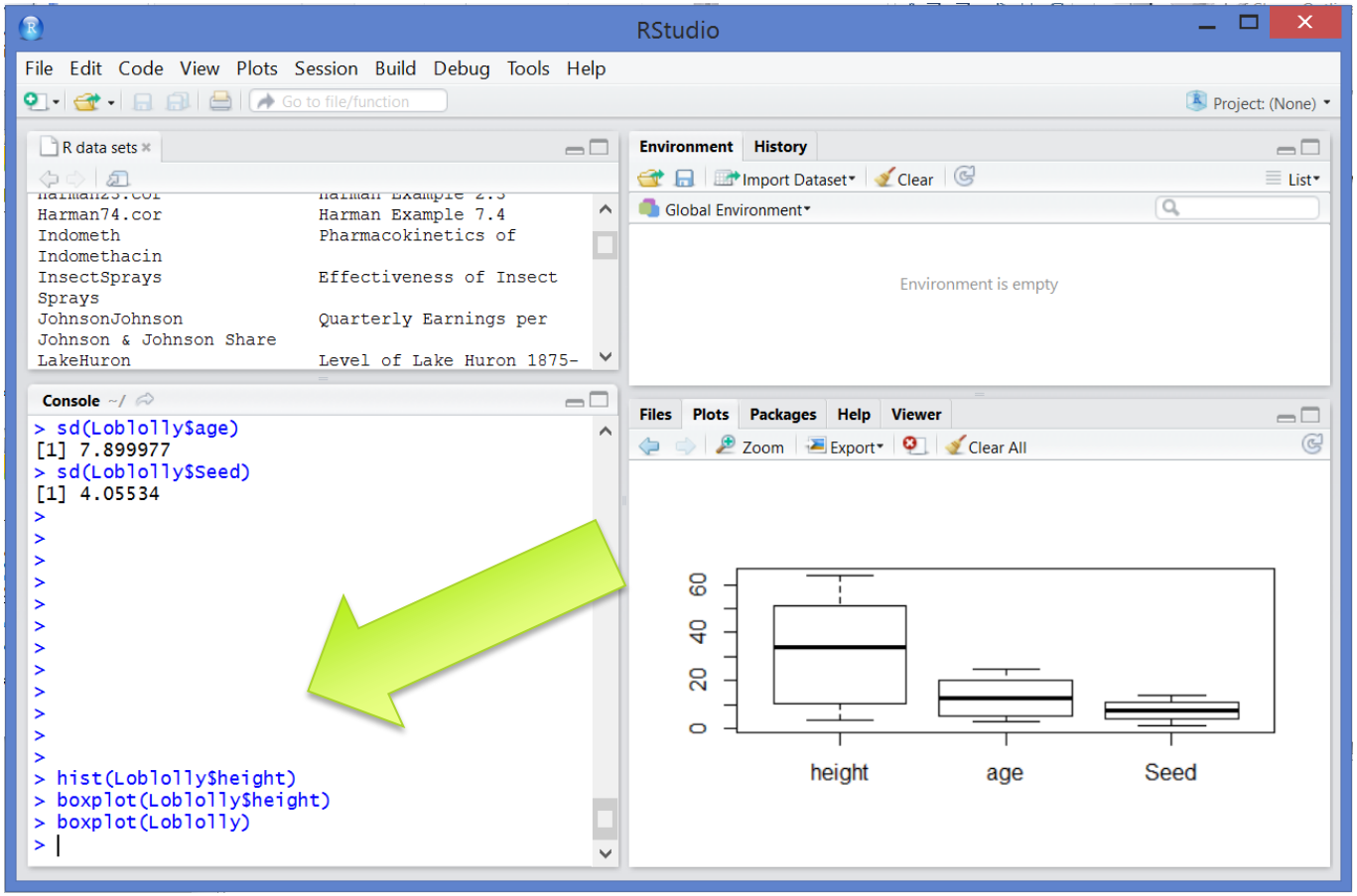
To build a histogram for an attribute, type **`hist(<dataset>$<attribute>)`** into the command line.

Boxplot of an Attribute



To build a boxplot for an attribute, type **`boxplot(<dataset>$<attribute>)`** into the command line.

Boxplot of all Attributes



To build a boxplot for all attributes, type **`boxplot(<data set>)`** into the command line.

Summary

Learning Objectives:

Upon completion of the module, you will be able to :

- ✓ Install R and R Studio
- ✓ Review the bundled datasets
- ✓ Learn how to view the Frequency Distribution
- ✓ Learn the commands to produce the Five Number Summary and other important metrics
- ✓ Learn how to build a histogram and a boxplot in R

Required:

- R and R Studio, downloadable for FREE

If you want to get fancy and add color:

- http://www.r-bloggers.com/box-plot-with-r-tutorial/

FINAL_PROJECT 2

FINAL_PROJECT 2: Basic Statistics in R

Take a look at the data in your project dataset and note the data types and their values. Then answer the following questions on 1 single-spaced page:

1. Find the five most important attributes in your dataset. What makes them important? What are their data types?
2. For these five attributes, use R to determine mean, median, mode, standard deviation, and Interquartile Range (IQR). If the attributes are nominal, convert them into numeric format to do this analysis. For what purpose are these measurements important?
3. If the attributes have missing or implausible values, what is going to be your strategy for filling them in?

Include your R code and its output in this submission.

PM Reboot

For Next Week

#	Topic	Reading	Due in Class
3	Preprocessing: Data Preparation for Data Mining	• Han/ Kamber 3 • Witten/ Eibe 7 • Ledolter 2 • An introduction to data cleaning with R • Lynda Course "R Statistics Essential Training" mod. 5-end • Videos as posted on Blackboard	**MW: Quiz 3**: Preprocessing using MS Excel and R **S**: **FINAL_PROJECT_3**: Determine 3 attributes in your dataset that present quality or analysis problems and transform and evaluate them.