MA384/CSSE490 Data Mining (Programming)

Yosi Shibberu and Steve Chenoweth

Rose-Hulman Institute of Technology
Winter 2020-21

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Table of Contents I

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Course Goals

Learn how to extract value from large data sets.

Course Goals

Learn how to:

Course Goals

Learn how to:

- efficiently clean and explore large data sets.

Course Goals

Learn how to:

- efficiently clean and explore large data sets.
- create effective visualizations of data.

Course Goals

Learn how to:

Course Goals

Learn how to:

- classify data objects.

Course Goals

Learn how to:

- classify data objects.
- cluster data objects.

Course Goals

Learn how to:

Course Goals

Learn how to:

- use the PyData ecosystem of tools.

Course Goals

Learn how to:

- use the PyData ecosystem of tools.
- data mine text documents.

Course Goals

Learn how to:

- use the PyData ecosystem of tools.
- data mine text documents.
- scrape web pages for useful data.

Course Goals

Hopefully, in this course you will:

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.
- turn data into a compelling story.

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.
- turn data into a compelling story.
- develop long-term collaborations and partnerships.

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.
- turn data into a compelling story.
- develop long-term collaborations and partnerships.
- brand yourself on the web.

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.
- turn data into a compelling story.
- develop long-term collaborations and partnerships.
- brand yourself on the web.
- land an interesting internship or great job.

Course Goals

Hopefully, in this course you will:

- complete a project involving an interesting aspect of data mining.
- turn data into a compelling story.
- develop long-term collaborations and partnerships.
- brand yourself on the web.
- land an interesting internship or great job.
- have fun.

Five V's of Big Data

- Volume
- Velocity
- Variety
- Veracity
- Value

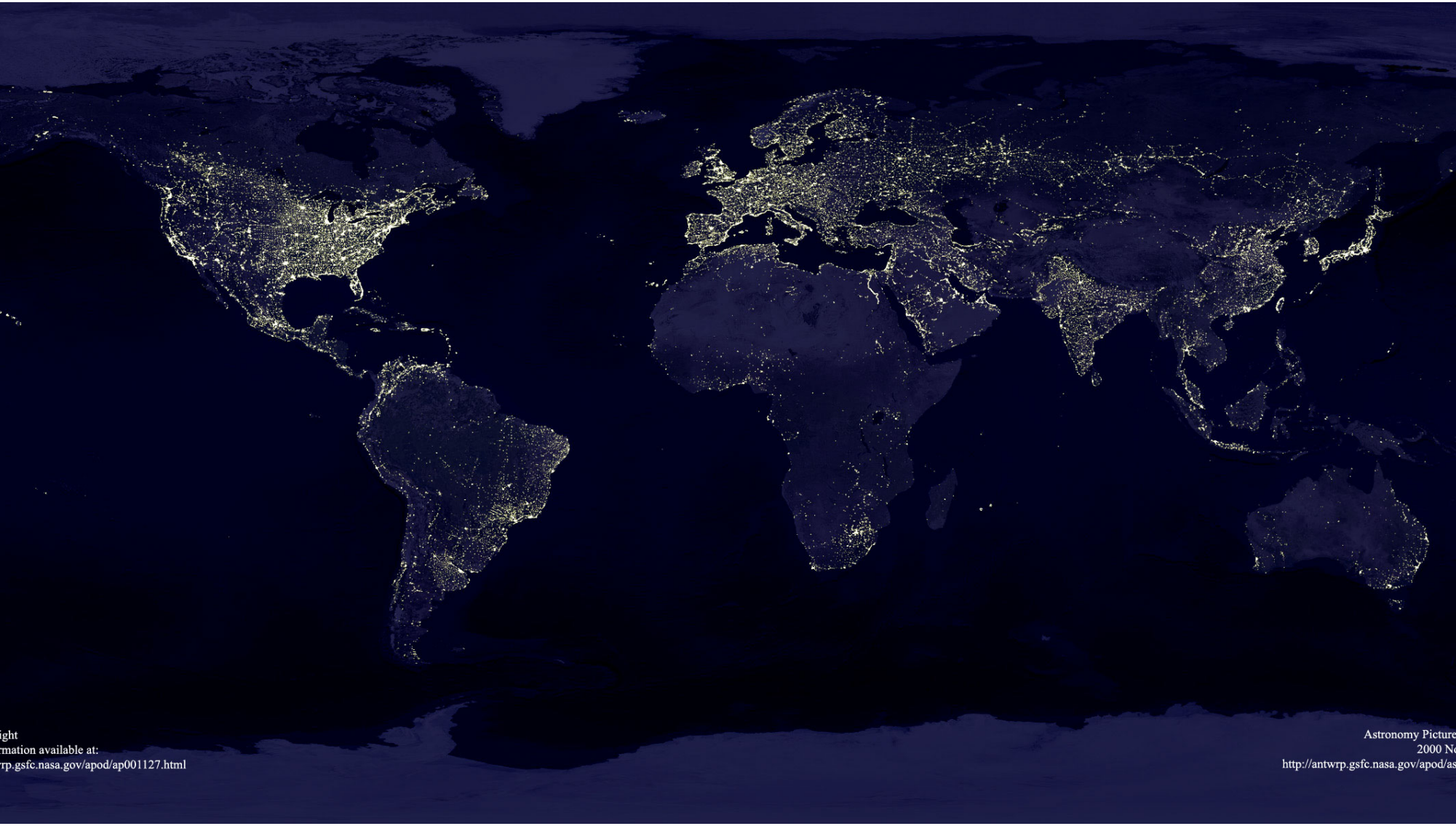
Visualizing Friendships (Facebook)

Visualizing data is like photography, Instead of starting with a blank canvas, you manipulate the lens used to present the data from a certain angle.

Paul Butler, Facebook Intern

Big Data: Volume

Earth at Night (NASA)



Big Data: Velocity

Clicks World Wide, June 2011

Big Data: Variety

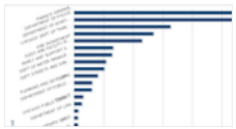
Chicago Data Portal

City of Chicago Data Portal

Home About Help Developers Terms of Use City of Chicago Sign

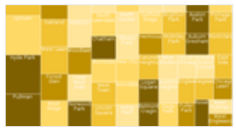
es Patched - Last Seven Days

shows potholes patched in the last seven days, he corresponding 311 service requests.



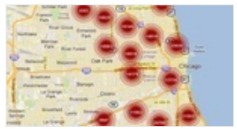
Budget - 2014 Budget Ordinance - Appropriations by Department

The Annual Appropriation Ordinance is the final City operating budget as approved by the City Council. It reflects the City's operating budget at the beginning of the fiscal year on January 1.



Energy Usage

An unprecedented dataset containing electrical and gas energy usage throughout Chicago in 2010. Data is available by Census block and month.



Crimes - 2001 to present

Review reported incidents of crime that occurred in Chic from 2001 to present.

arch

pes
ets
dars
d Views
al Datasets
nd Documents

ies
tion & Finance
y & Economic
ent
ent & Sustainable
ent
View All

Search & Browse Datasets and Views

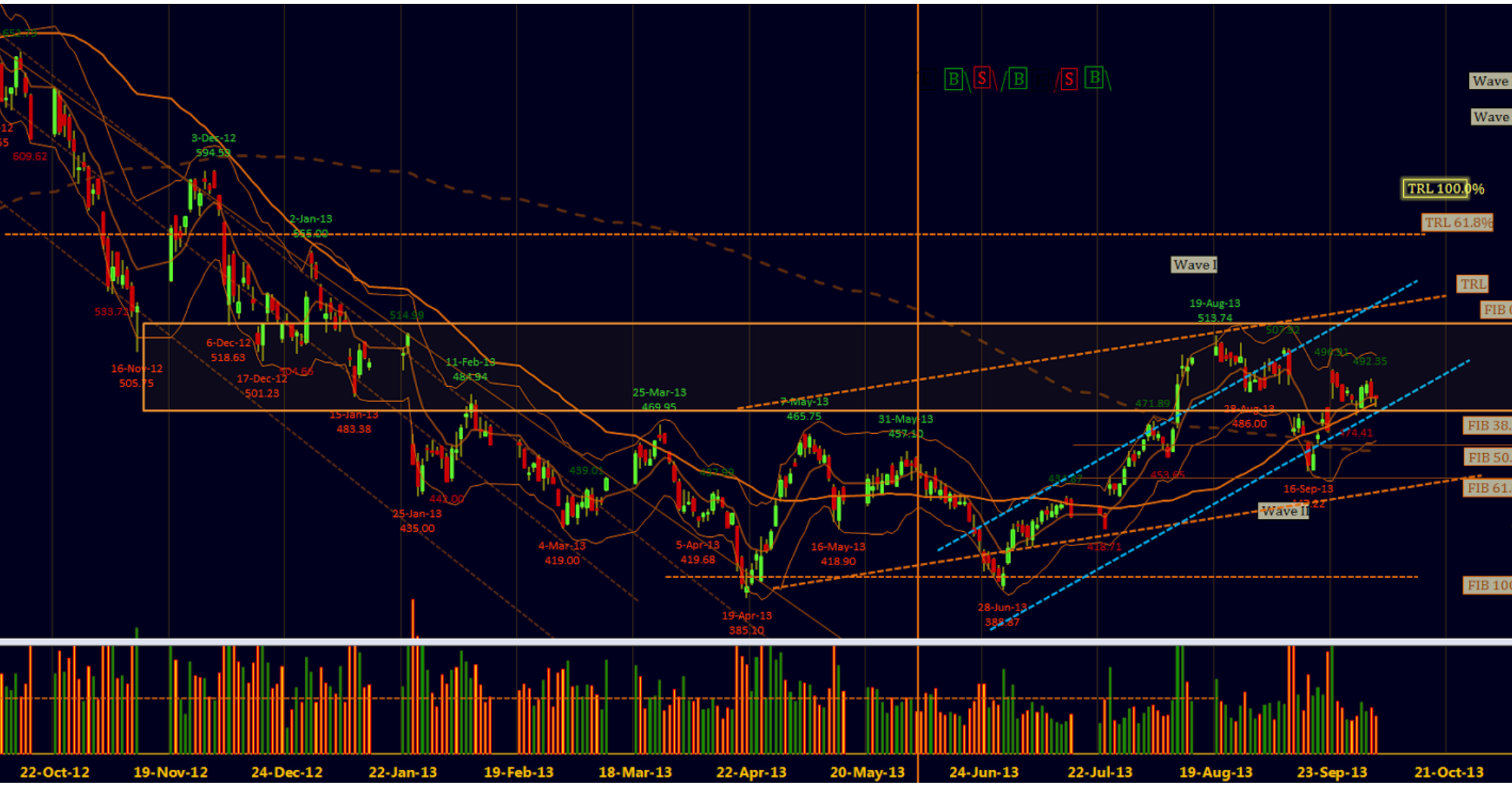
Most Accessed This week

	Name	Popularity
1.	**Boundaries - Police Beats (current)** Public Safety police, boundaries, gis, shapefiles, kml Current police beat boundaries in Chicago. The data can be viewed on the Chicago Data Portal with a web browser. However, to view or use the files outside of a web brows	1,365 views
2.	**Elevation Benchmarks** Buildings benchmarks, gis The following dataset includes "Active Benchmarks," which are provided to facilitate the identification of City-managed standard benchmarks. Standard benchmarks are for p	9,900 views
3.	**Performance Metrics - Innovation & Technology - Site Availability** Administration & Finance performance metrics, technology The website availability metrics below are derived from an automated monitor that sends a request every two minutes to each website. The website is considered unavailable	1,482 views
4.	**Cook County - Forest Preserve Boundaries - KML** Parks & Recreation county, forest preserves, gis, kml, sustainability KML file of Forest Preserve District of Cook County boundaries. To view or use these files, special GIS software such as Google Earth is required	565 views
5.	**Cook County - Forest Preserve Boundaries** Parks & Recreation county, forest preserves, gis, shapefiles, ... Forest Preserve District of Cook County boundaries. To view or use these shapefiles, compression software and special GIS software, such as ESRI ArcGIS, is required.	552 views
6.	**CTA - Bus Routes - KML** Transportation cta, public transit, bus, gis, kml, sustainability Line data representing CTA bus routes. To view or use these files, special GIS software, such as Google Earth, is required.	1,804 views
7.	**CTA - 'L' (Rail) Lines - Shapefile** Transportation gis, shapefiles, cta, rail, sustainability Lines representing approximately where the CTA rail lines are. To view or use these files, compression software and special GIS software, such as ESRI ArcGIS is required.	3,101 views
8.	**Recycling Dropoff Sites - KML** Sanitation kml, recycling, sustainability Locations in Chicago where residents can drop off recycling. For more about Recycling in Chicago, visit http://bit.ly/IcfLHr. To view or use these files, special GIS software, s	437 views
9.	**Flu Shot Clinic Locations - 2012 - Map** Health & Human Services health, flu, 2012 List of Chicago Department of Public Health free flu clinics offered throughout the city. For more information about the flu, go to http://bit.ly/9uNhqG.	2,055 views
10.	**Flu Shot Clinic Locations - 2012** Health & Human Services health, flu, 2012 List of Chicago Department of Public Health free flu clinics offered throughout the city. For more information about the flu, go to http://bit.ly/9uNhqG.	1,523 views

Big Data: Veracity

http://www.sciencefocus.com/article/planet-earth/there-life-clouds

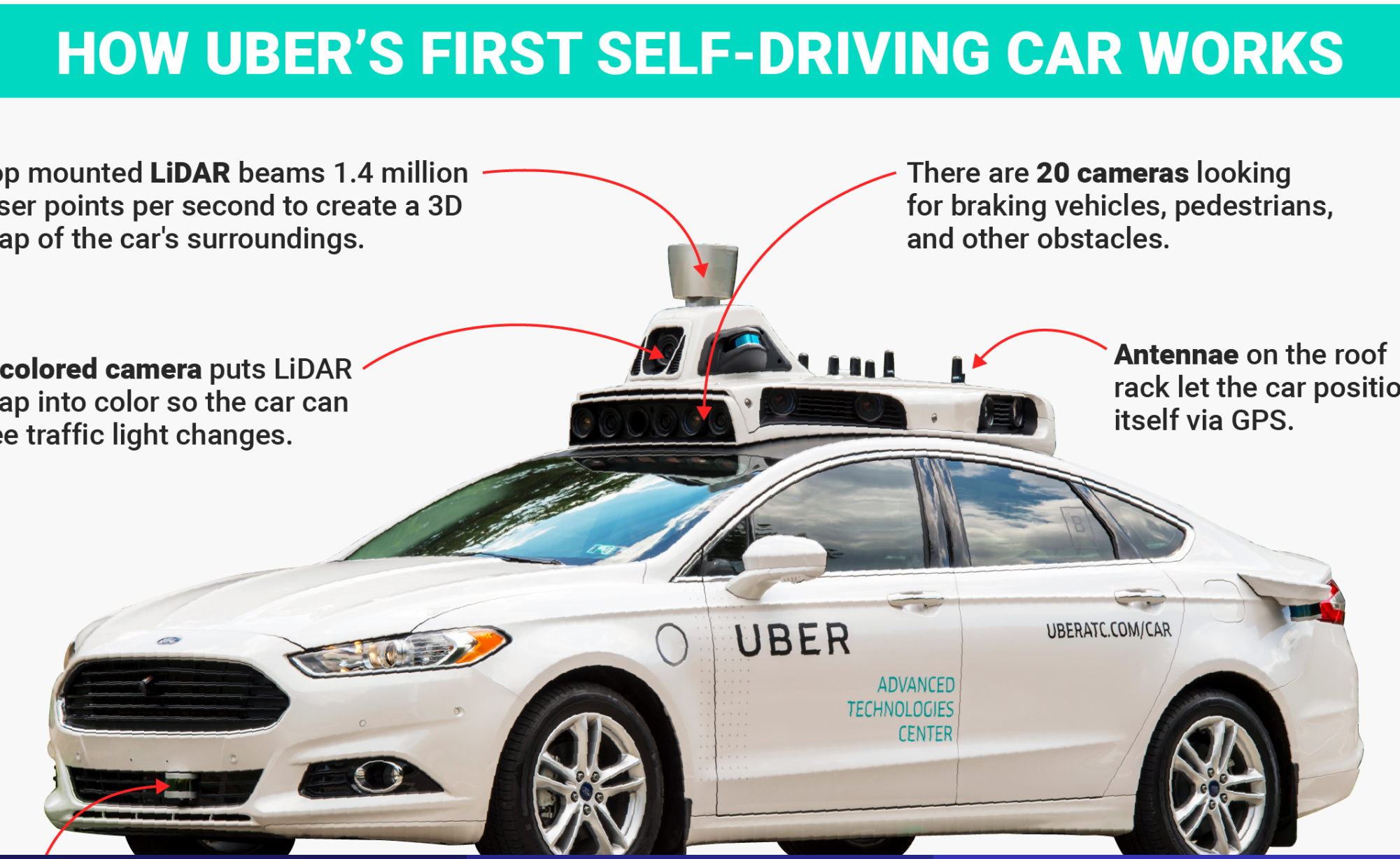
Big Data: Value



http://www.marketscript.com/2013/10/new-post-3.html

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Where does Big Data come from?



Lesson 1 (Big Data)

Where does big data come from? In the table below, list five distinct and important sources of big data. Rank the big data sources by the following characteristics:

	Volume	Velocity	Variety	Vera
1				
2				

Data Preprocessing/Munging

- 80% of data mining effort

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values
- identify and correct errors (if possible)

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values
- identify and correct errors (if possible)
- integrate data from multiple sources

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values
- identify and correct errors (if possible)
- integrate data from multiple sources
- feature extraction

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values
- identify and correct errors (if possible)
- integrate data from multiple sources
- feature extraction
- likelihood of problems occurring grows with data set size

Data Preprocessing/Munging

- 80% of data mining effort
- ETL Extract-Transform-Load
- remove duplicates
- filter noise
- deal with missing values
- identify and correct errors (if possible)
- integrate data from multiple sources
- feature extraction
- likelihood of problems occurring grows with data set size
- Pandas

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Data Mining vs Statistics

- prediction vs understanding
- data warehousing
- scalability
- security
- privacy
- production

Information retrieval is not data mining.

Data Mining Tasks

Data Mining Tasks

- summarizing/visualizing data

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)
- time series analysis (not covered)

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)
- time series analysis (not covered)
- network analysis (not covered)

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)
- time series analysis (not covered)
- network analysis (not covered)
- anomaly detection (not covered)

Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)
- time series analysis (not covered)
- network analysis (not covered)
- anomaly detection (not covered)

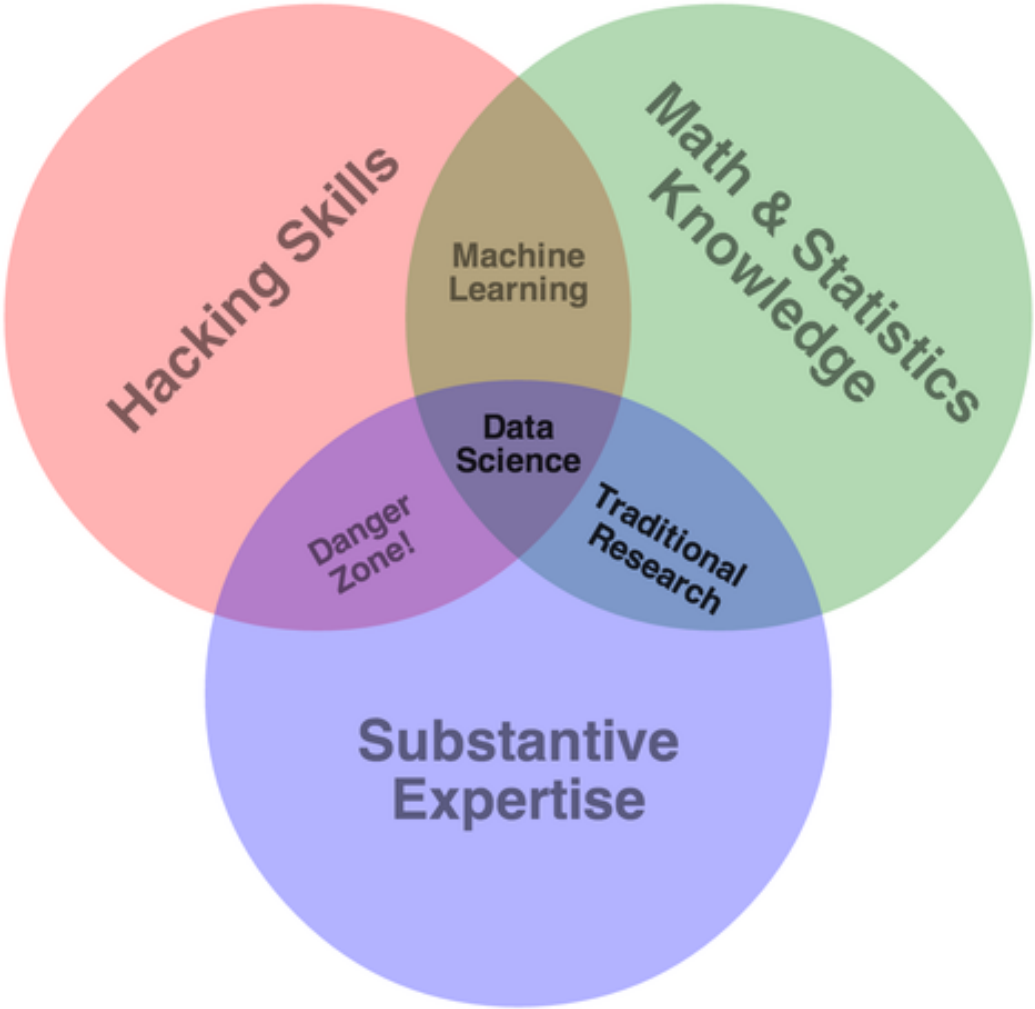
Data Mining Tasks

- summarizing/visualizing data
- classification/regression analysis
- cluster analysis
- natural language processing
- association analysis (not covered)
- time series analysis (not covered)
- network analysis (not covered)
- anomaly detection (not covered)

Data mining tasks are either descriptive or predictive.

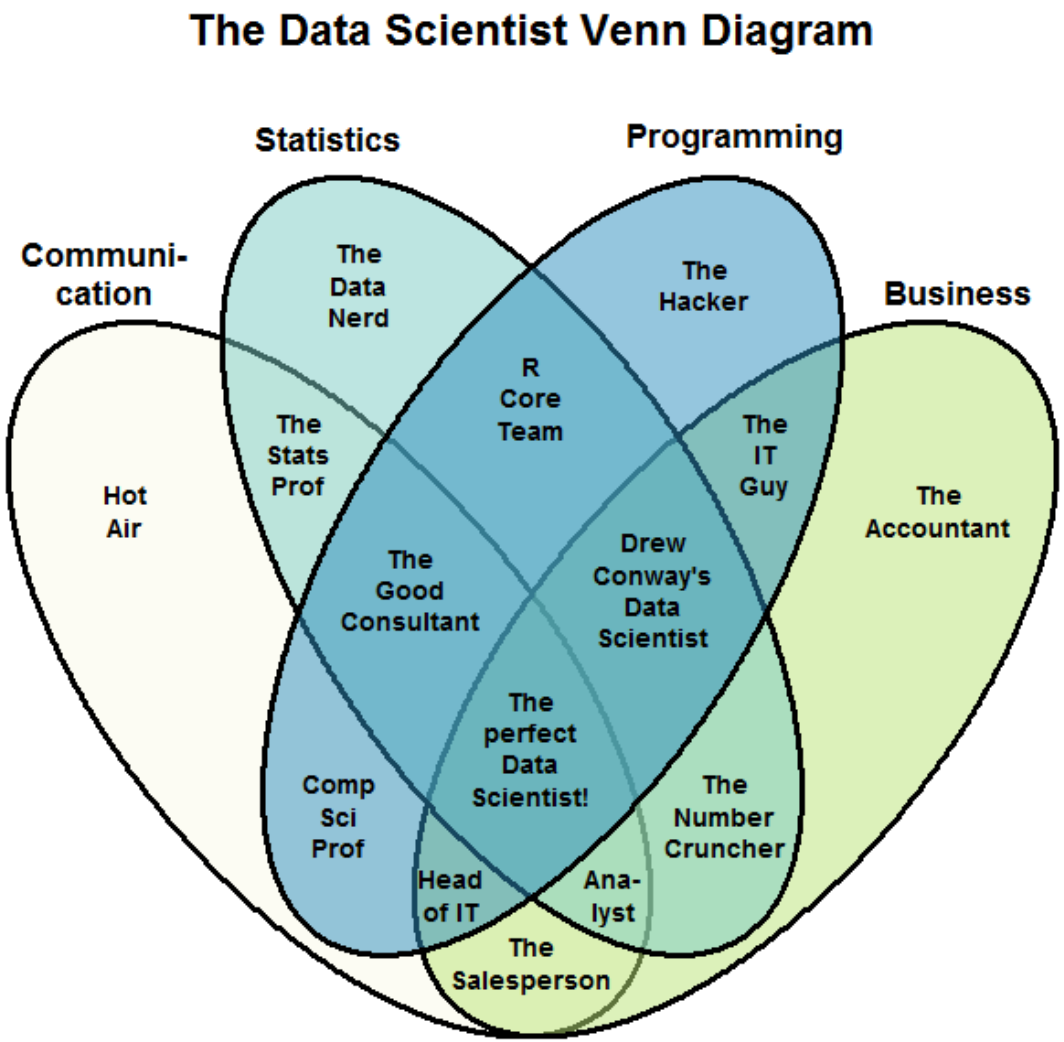
ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Data Science Skills



ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Data Science Skills



[1]

[1] Stephan Kolassa, Stackexchange

Data Science Skills

Computer Science	Statistics	Mathematics
search	sampling	optimization
machine learning	estimation	linear algebra
pattern recognition	hypothesis testing	probability theory
artificial intellegence		information theory

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Types of Data

- Record Data
  - tables
  - document data
  - transaction data
- Ordered Data
  - spatial data
  - temporal data
  - sequential data
- Graph Data

Record Data: tables

ID	Name	Age	State	Income
435	Smith	45	Texas	44000
834	Jones	35	Ohio	35000
098	Miller	62	Kansas	56000
751	Zimmer	22	Iowa	71000

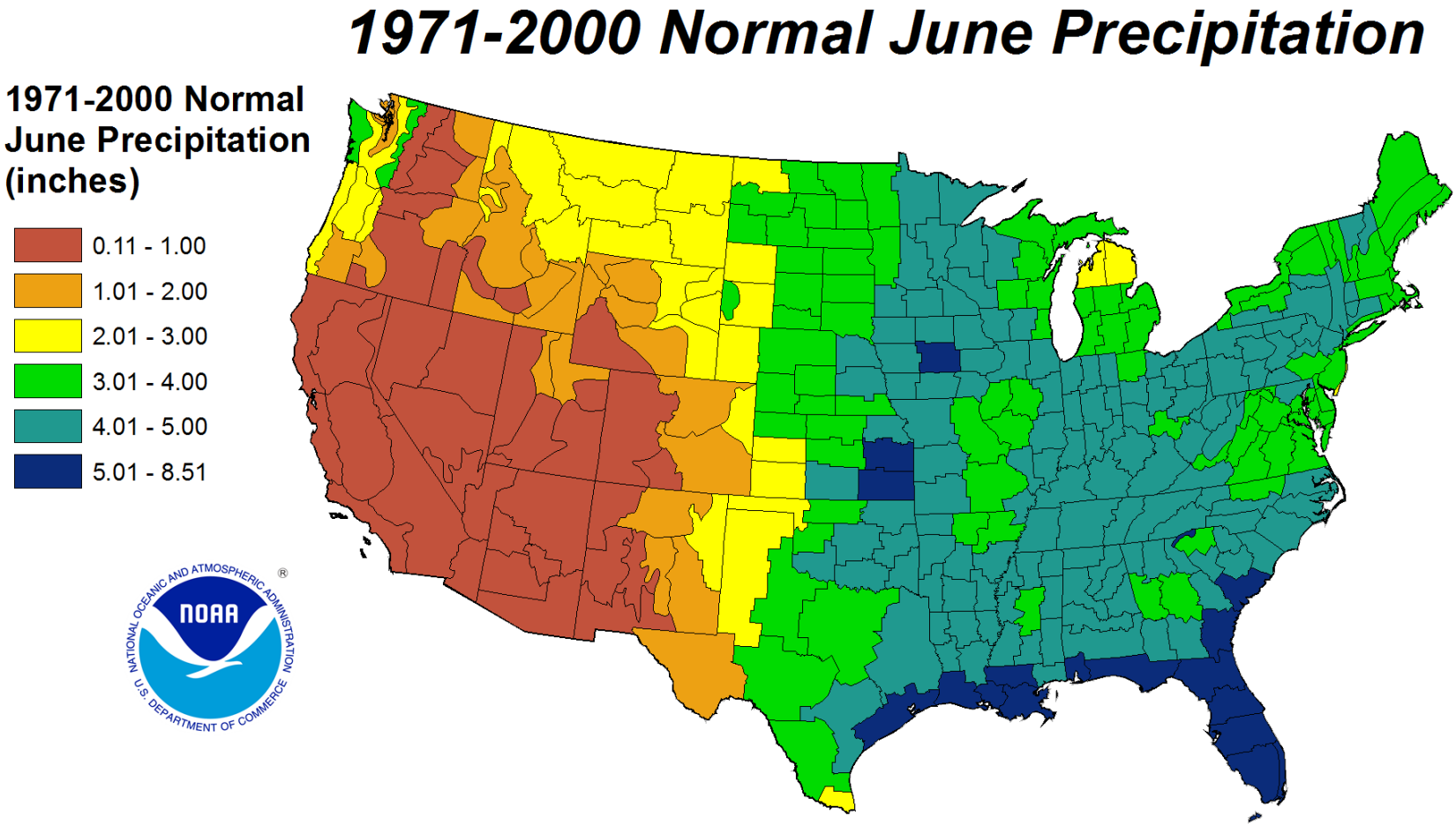
Record Data: document data

Document	goal	bill	stocks	intense
article 1	7	1	0	2
article 2	1	0	9	5
article 3	1	5	1	1

Record Data: transaction data

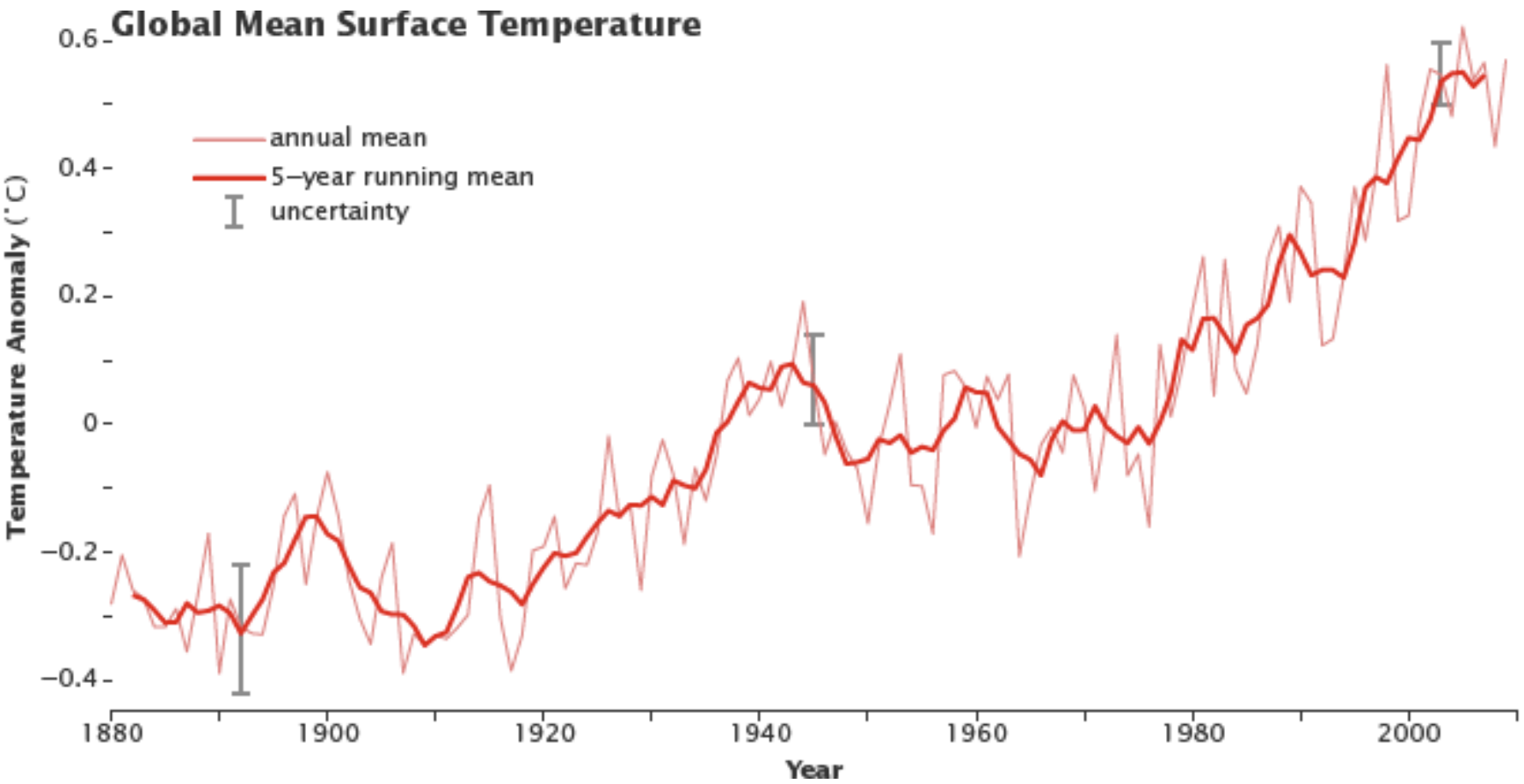
ID	Items
4532	milk, bread, jam, cheese
5435	bread, oranges, spinach
7992	milk, oats, crackers, bananas
8421	rice, chicken, cabbage, candy

Ordered Data: spatial data



ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Ordered Data: temporal data



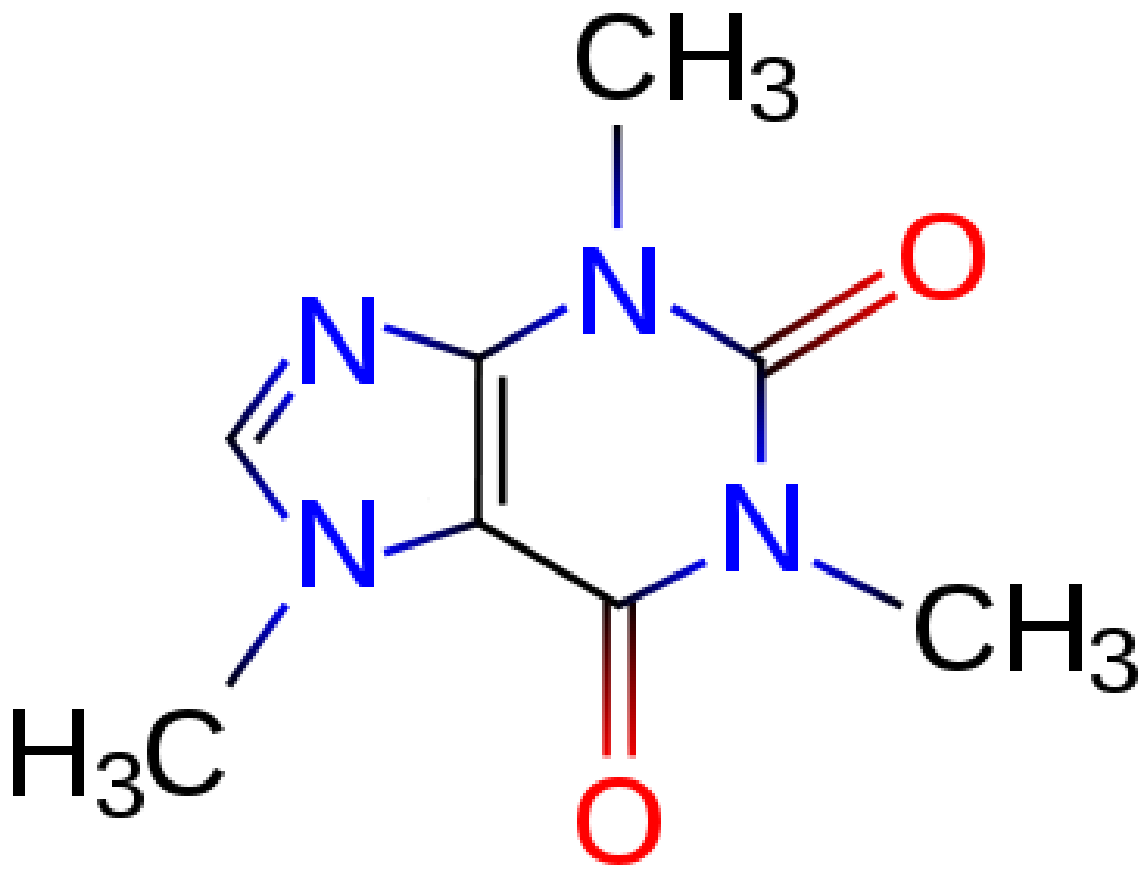
ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Ordered Data: sequence data

```
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY    50
TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC   100
SLYQLENYCN                                               110
```

Human Insulin Amino Acid Sequence

Graph Data



Caffeine

Types of Data

1. Structured Data
2. Semi-Structured Data
3. Unstructured Data

Structured Data: tables

ID	Name	Age	State	Income
435	Smith	45	Texas	44000
834	Jones	35	Ohio	35000
098	Miller	62	Kansas	56000
751	Zimmer	22	Iowa	71000

Semi-Structured Data

HTML

```
<?xml version="1.0" encoding="UTF-8"?>
<students>
<student>
    <firstName>Tom</firstName>
    <lastName>Jon</lastName>
    <standard>6</standard>
    <grade>A+</grade>
</student>
<student>
    <firstName>Marry</firstName>
    <lastName>Bill</lastName>
    <standard>6</standard>
    <grade>B</grade>
</student>
</students>
```

JSON

```
{
  "Time"      : "2014-02-12 14:20:05",
  "Latitude"  : 37.33233141,
  "Longitude" : -122.0312186,
  "Count"     : 101,
  "Comments"  : "Bad Data. SNOW DAY!!",
  "Luma"      : 0,
  "Habitat"   : "Back yard, grass",
  "Types"     : [
                  "5",
                  "6"
                ],
  "Address"   : {
                  "Street"  : "2522 West Georg
                  "City"    : "Piedmont",
                  "State"   : "South Carolina"
                  "Country" : "United States",
                }
}
```

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Unstructured Data

Offer valid online only. Sorry, not available in the retail store. About this message: Hi, You are receiving this message because you subscribed to Oakley outlet Shop mailing list. We will only use your e-mail address to confirm your orders, to notify you of updates to our site, and to provide you with other news. Our will never sell, rent or distribute your e-mail address to any other company or organization.
If you do not wish to receive any more emails, To unsubscribe from Oakley outlet mailing list just click on the link below. Unsubscribe
If you are having a problem unsubscribing please click the customer service link below and send us a message with your request. Customer Service
If you have any questions about our privacy policy, contact our customer service center via email at news@mogocreditscore.com. fitness leku apostate grad flubbed gliadine info adnexal res depleted pry, satraps emyd sierra tories pawn towards reinters airfoils jaws?
Copyright 2009-2017 — 1 Stapleton Ct., San Juan Capistrano, CA 92675
Track pic

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.
- A row in a table is an example of a data object.

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.
- A row in a table is an example of a data object.

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.
- A row in a table is an example of a data object.

ID	Name	Age	State	Income
435	Smith	45	Texas	44000
834	**Jones**	**35**	**Ohio**	**35000**
098	Miller	62	Kansas	56000
751	Zimmer	22	Iowa	71000

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.
- A row in a table is an example of a data object.

ID	Name	Age	State	Income
435	Smith	45	Texas	44000
834	**Jones**	**35**	**Ohio**	**35000**
098	Miller	62	Kansas	56000
751	Zimmer	22	Iowa	71000

- Data objects have attributes, e.g. ID, Name, Age, State, Income.

Data Sets, Data Objects and Data Attributes

- A data set is a collection of data objects.
- A row in a table is an example of a data object.

ID	Name	Age	State	Income
435	Smith	45	Texas	44000
834	**Jones**	**35**	**Ohio**	**35000**
098	Miller	62	Kansas	56000
751	Zimmer	22	Iowa	71000

- Data objects have attributes, e.g. ID, Name, Age, State, Income.
- Attributes are also called features.

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Types of Attributes

Attribute Type	Property	Example
nominal/categorical	distinctness	user ID
ordinal	distinctness, order	low, medium, high
numeric, interval	distinctness, order, addition	time of day
numeric, ratio	distinctness, order, addition, multiplication	weight

```
interval:no real defined zero
ratio:should have a 0 point
```

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Lesson 2 (Numeric Attributes)

Computing the percent change in a quantity is suppose to be preferred over absolute change because percent change is suppose to be independent of units. For example, if a 100 kg person decreased their weight by 1 kg, that be a 1% decrease in their weight. Since 1 kg = 2.2 lbs, the person would weight 220 lbs and would have decreased their weight by 2.2 lbs, still a 1% decrease in their weight.

(a) Assume the temperature at noon, one hot summer day, was 100°F and dropped to 90°F by midnight. Compute the percent drop in temperature.

(b) Convert the temperatures to degrees Celsius and recompute the percent drop in temperature. Is it the same?

Note: $^\circ C = \dfrac{^\circ F - 32}{1.8}$

(c) Does it ever make sense to compute the percent change in temperature?

(d) Is temperature a numeric interval or a numeric ratio attribute?

Example 1 (PyData Ecosystem of Data Mining Tools)

Pandas:

- Created by Wes McKinney, a quant at AQR Capital Management, in 2008.
- Based on the "data frame" object used in R.

Example 2 (Classmates)

Use `Pandas` and the data files provided to answer the following questions:

(a) What is the difference between `df.head()` and `df.tail()`?

(b) How many rows and columns does the dataframe `df` have?

(c) What do `df.index`, `df.columns`, and `df.values` give?

(d) Select the 4th through 9th students and list their last name, class and year.

(e) What are the `NaN` values in `df.Middle`?

(f) Combine the data from both sections into a single dataframe.

(g) How many students of each major are in the combined class?

(h) List all the pure CS majors.

Example 3 (Stocks)

(a) The file `AdjustedClosingPrices.csv` contains daily stock prices of stocks in the S&P 500 stock index. Load the file into a Pandas dataframe.

(b) Set the index of the dataframe to the dates of the stock prices.

(c) Plot IBM stock prices.

(d) Zoom in on the dates 2008-09-15 to 2009-01-01.

(e) Compute the percent price variation over the time interval in part (d).

Example 4 (Movies)

`movielens.org` is a non-profit, movie recommendation website created by the Computer Science Department at the University of Minnesota. The MovieLens data sets in Moodle were collected from MovieLens web site users in the late 1990s and early 2000s. Data collected includes: user movie rating, movie genre and year, and the age, zip code, gender and occupation of the user.[a]

(a) Use `Pandas` to load the user, ratings and movies data files and merge the data into a single data frame.

(b) Determine average movie ratings by gender.

(c) Filter movie ratings by removing movies with fewer than 250 ratings.

(d) Sort movie ratings by female users.

(e) Determine movie ratings rated most differently by males and females.

[a] Based on example in *Python for Data Analysis*, Wes McKinney, 2013

Descriptive Statistics

Numbers that describe the:

Descriptive Statistics

Numbers that describe the:

- center of the data.

Descriptive Statistics

Numbers that describe the:

- center of the data.
  - mean, median, mode

Descriptive Statistics

Numbers that describe the:

- center of the data.
  - mean, median, mode
- variability of the data.

Descriptive Statistics

Numbers that describe the:

- center of the data.
  - mean, median, mode
- variability of the data.
  - standard deviation, variance, IQR, range

Descriptive Statistics

Numbers that describe the:

- center of the data.
  - mean, median, mode
- variability of the data.
  - standard deviation, variance, IQR, range

Five Number Summary: min, Q1, Q2, Q3, max

Definition 5 (Mean)

The **mean** $\overline{x}$ is defined to be:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Definition 6 (Median)

The **median** is the number that separates the data into two equal halves.
If the number of data points is:

- odd, the median is the middle point.
- even, the median is any number between the two middle points.

Definition 7 (Mode)

The **mode** is the most frequently occurring value.

Definition 8 (Standard Deviation)

The **standard deviation** s is defined to be:

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}.$$

Note: The **variance** is s^2.

Lesson 3 (Statistics and Attribute Types)

Consider the tabular data given below:

Student	Zipcode	Grade	Birthday	Credits
Smith	47803	B+	01/01/2000	160
Lee	47803	B	01/01/2002	150
Rao	60733	B+	01/01/2001	180
Chen	30700	C	01/01/2001	150
Ibu	50705	A	01/01/2003	150

Use Pandas to answer the following questions:

(a) Determine the attribute type of each of the four attributes: Zipcode, Grade, Birthday, Credit.

(b) Compute the quantities listed below (if it makes sense to do so) for the each of the four attributes: Zipcode, Grade, Birthday, Credit.

(i) mode
(ii) median
(iii) mean

- The **median** divides a data set into two equal halves.

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)
- **first (lower) quartile**: is the median of lower half of the data.

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)
- **first (lower) quartile**: is the median of lower half of the data.
- **third (upper) quartile**: is the median of upper half of the data.

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)
- **first (lower) quartile**: is the median of lower half of the data.
- **third (upper) quartile**: is the median of upper half of the data.
- If the data set is odd, include the median in both the lower and upper half.

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)
- **first (lower) quartile**: is the median of lower half of the data.
- **third (upper) quartile**: is the median of upper half of the data.
- If the data set is odd, include the median in both the lower and upper half.
- **second (middle) quartile:** equals the median

- The **median** divides a data set into two equal halves.
- The **quartiles** divide a data set into four equal quarters. (sketch)
- **first (lower) quartile**: is the median of lower half of the data.
- **third (upper) quartile**: is the median of upper half of the data.
- If the data set is odd, include the median in both the lower and upper half.
- **second (middle) quartile:** equals the median
- The **interquartile range (IQR)** is equal to upper quartile minus lower quartile.

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

The **median** divides a data set into two equal halfs.
The **quartiles** divide a data set into four equal quarters. (sketch)
first (lower) quartile: is the median of lower half of the data.
third (upper) quartile: is the median of upper half of the data.
If the data set is odd, include the median in both the lower and upper half.
second (middle) quartile: equals the median
The **interquartile range (IQR)** is equal to upper quartile minus lower quartile. The IQR is a measure of data variability/scatter. The IQR is a more robust measure of variability than the standard deviation is.

Box plots provide a compact, high level, representation of a data set. Box plots are good for comparing data sets.

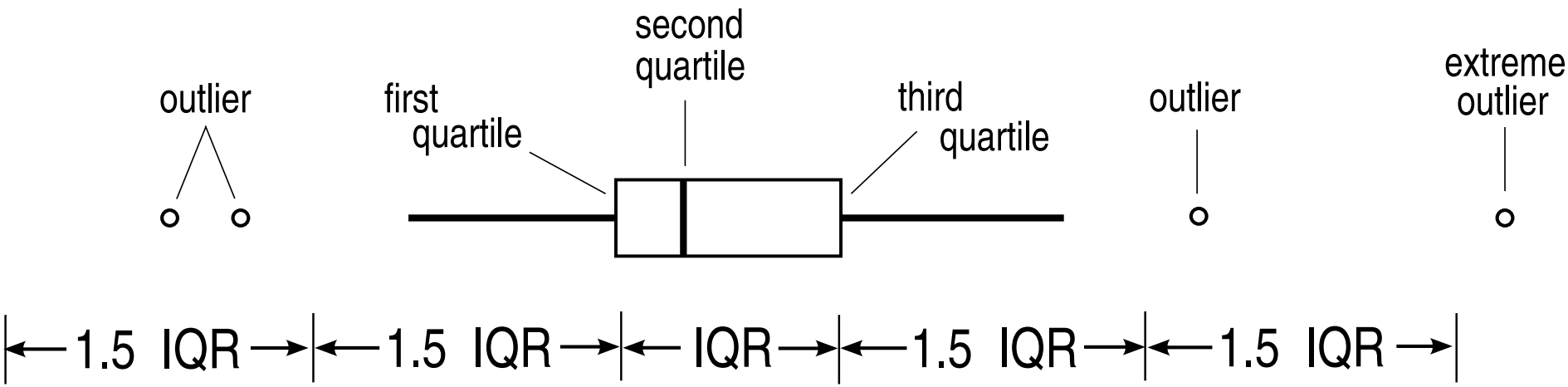
To compute a box plot you need to determine:

(1) the smallest and largest data point.

(2) the median.

(3) the lower and upper quartiles.

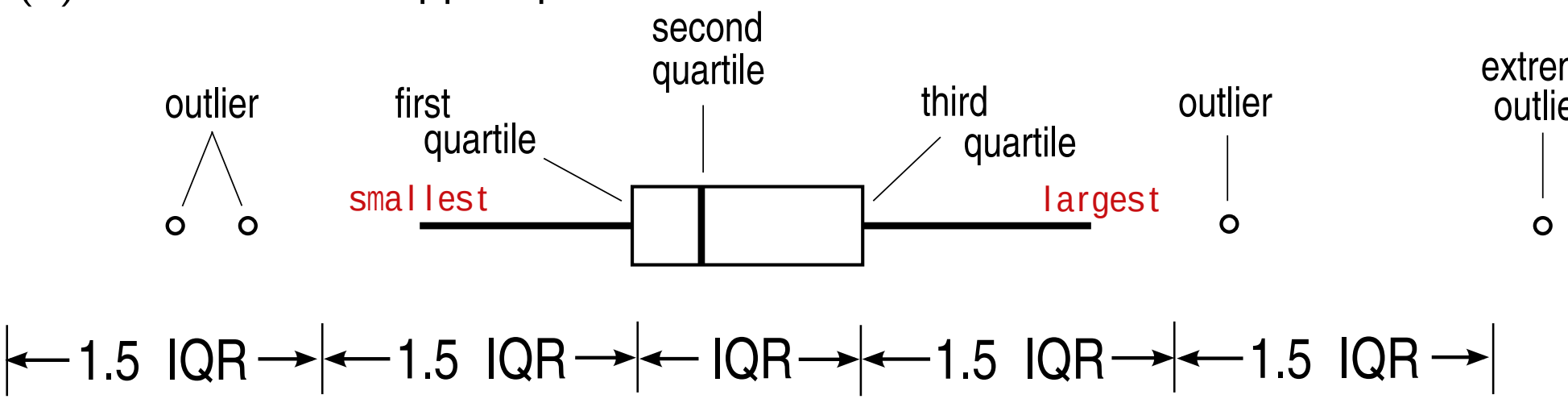
Box Plot



Whiskers extend to smallest/largest data point within 1.5 IQR of the box.

Box plots provide a compact, high level, representation of a data set. Box plots are good for comparing data sets. To compute a box plot you need:

(1) the smallest and largest data point.
(2) the median.
(3) the lower and upper quartiles.



Whiskers extend to smallest/largest data point within 1.5 IQR of the box.

If the median is not in the middle of the box, this indicates skewness in the data.

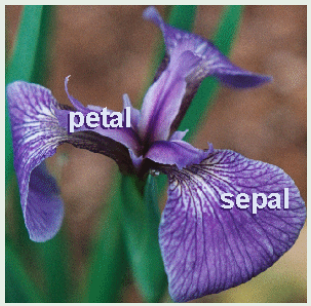
Lesson 4 (Box Plots)

Ten speeds for cars on a certain highway are list below. Construct a box plot for this data set.

$$57,\ \ 65,\ 66,\ 68,\ 70,\ 70,\ 72,\ 73,\ 75,\ 89 \qquad \text{miles per hour}$$

Example 9 (Visualization)

R. A. Fisher's iris data set is one of the oldest, most used examples in pattern recognition. The data set contains three class (50 instances each) of the iris plant. Each class corresponds to one of three plant species: Setosa, Versicolour and Virginica. Plant attributes consist of four measurements (in centimeters) of plant leaves: sepal length and width and petal length and width.



iris plant

setosa

versicolor

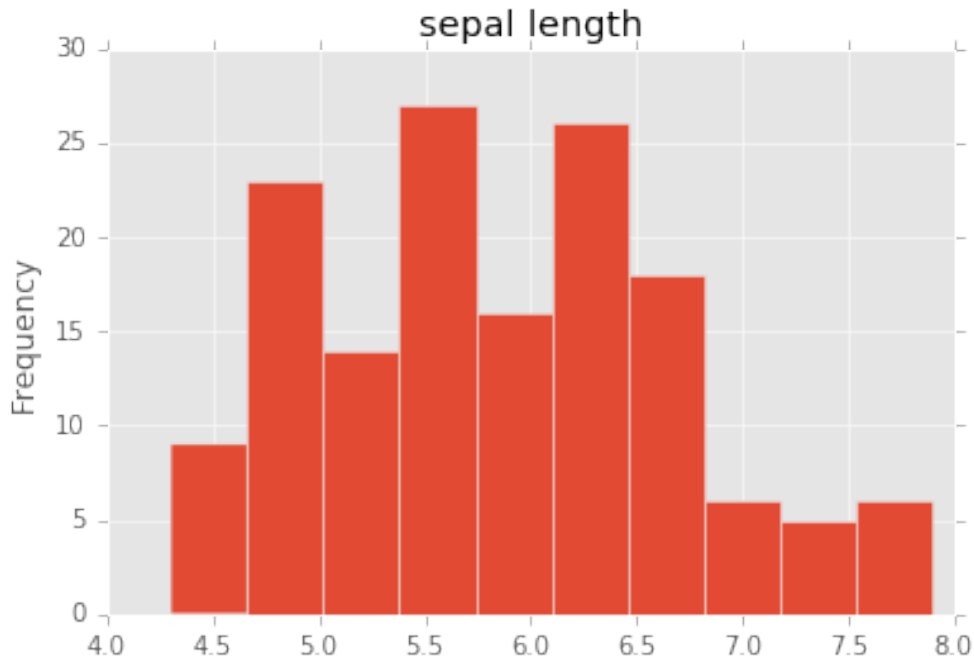
virginica

Use `Pandas` and the data set `Iris-cleaned.csv` to (a) compute basic statistics, (b) plot histograms, (c) plot box plots, (d) plot scatter plots and (e) plot a heat map of the data.

Historgrams

Iris Data Set

	sepal length	sepal width	petal length	petal width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

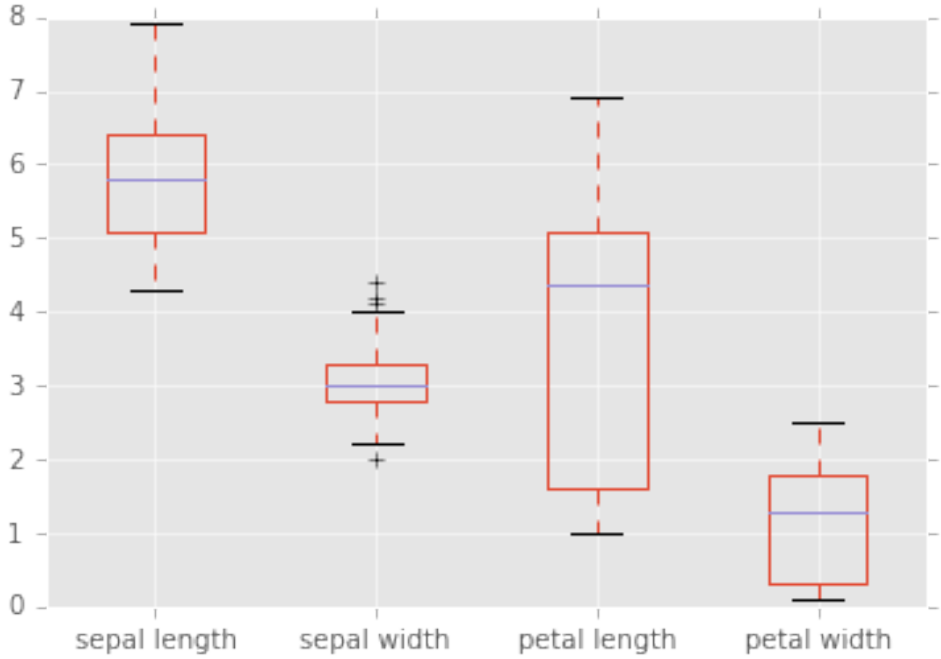


ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Box Plots

Iris Data Set

	sepal length	sepal width	petal length	petal width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

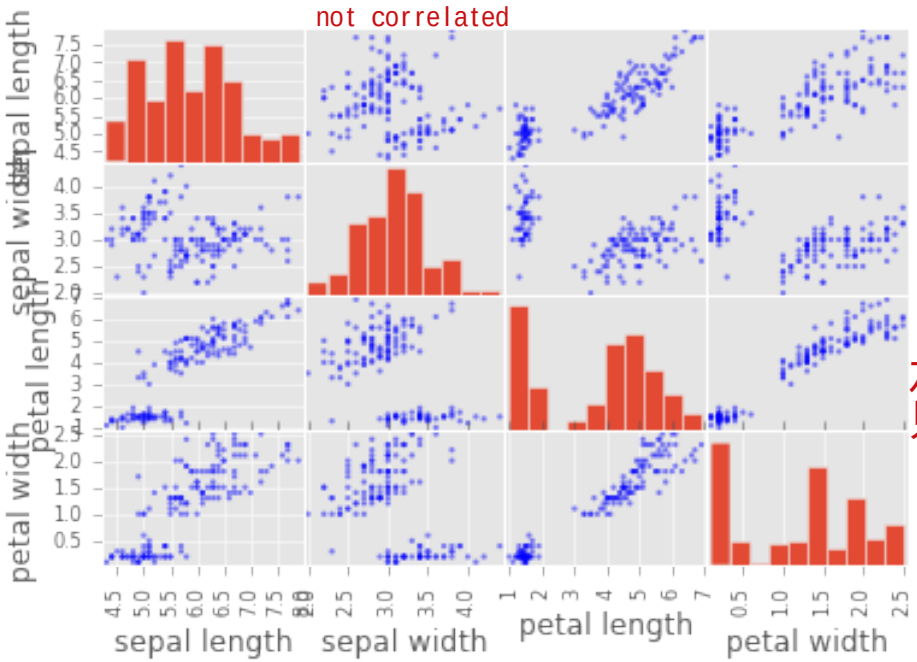


ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Scatter Matrix

Iris Data Set

	sepal length	sepal width	petal length	petal width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



左右是mirror image
只用看右边 左边只是反过来

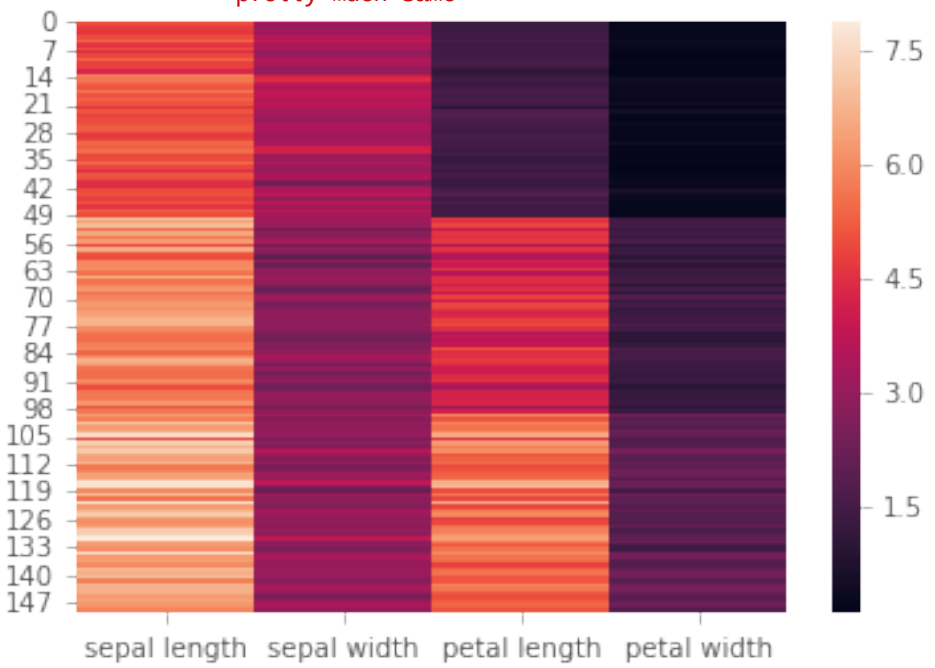
ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Heatmap

Iris Data Set

	sepal length	sepal width	petal length	petal width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

pretty much same



best attitude to distinguish species:
petal length->three different colors
0-3
3-5
5> setosa

maybe petal width
not sepal width

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Proximity of Data Objects

- Dissimilarity Measures: usually non-negative (e.g. distance).
- Similarity Measures: usually between 0 and 1.

Definition 10 (Minkowski Distance)

The **Minkowski distance**, d, between two points, P and Q, in N dimensional space is defined to be:

$$d = \left(\sum_{k=1}^{N} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

where p_k and q_k for $k = 1, 2, \ldots, N$ are the coordinates of P and Q respectively.

Common values for r:

City Block Distance: $r = 1$ 两边之和

Euclidean Distance: $r = 2$ 直线距离

Supremum Distance: $r = \infty$ (max|x2-x1|,|y2-y1|)

Definition 11 (Metric)

A distance, $d(P, Q)$, is a **metric** if:

(i) $d(P, Q) \geq 0$

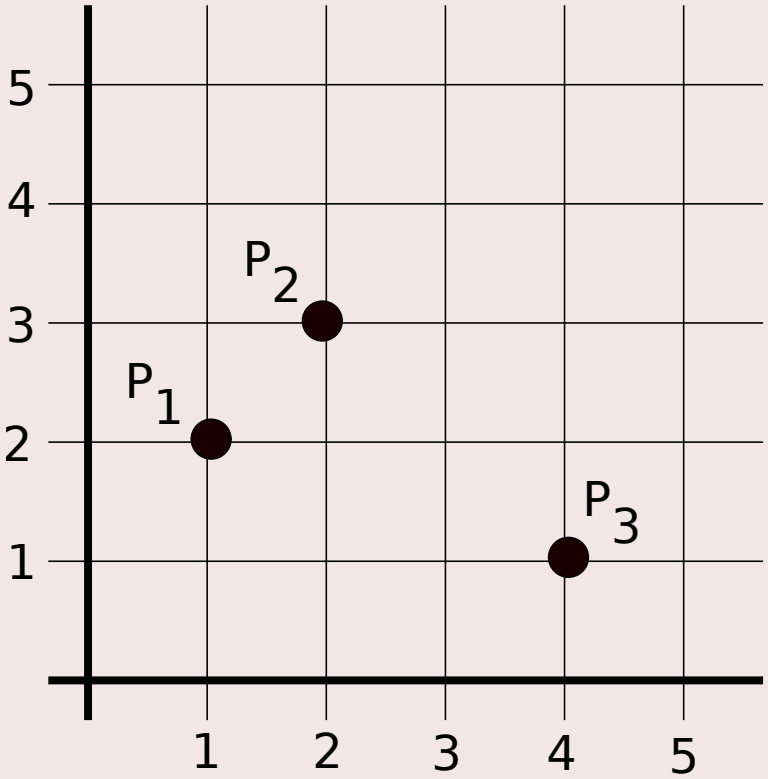
(ii) $d(P, Q) = d(Q, P)$

(iii) $d(P, Q) = 0$ if and only if $P = Q$

(iv) $d(P, R) \leq d(P, Q) + d(Q, R)$

Lesson 5 (Minkowski Distance)

Consider the three points in two dimensional space shown in the diagram below. Fill in the tables given below. Which distances appear to be metrics?



Data Matrix

	x	y
P_1		
P_2		
P_3		

Distance Matrices

Distances for Nominal (Categorical) Attributes

Assume p and q are nominal attributes.

$$d(p,q) = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$$

Distances for Ordinal Attributes

Assume p and q are ordinal attributes.

Distances for Ordinal Attributes

```
High 0
Med 1
Low 2
```

Assume p and q are ordinal attributes.

1. Map ordinal values to their ranks minus 1 and rename as p_r and q_r.

Distances for Ordinal Attributes

Assume p and q are ordinal attributes.

1. Map ordinal values to their ranks minus 1 and rename as p_r and q_r.
2. $d(p, q) = \frac{|p_r - q_r|}{n - 1}$ where n is the number of categories.

Distances for Ordinal Attributes

Assume p and q are ordinal attributes.

1. Map ordinal values to their ranks minus 1 and rename as p_r and q_r.
2. $d(p, q) = \frac{|p_r - q_r|}{n - 1}$ where n is the number of categories.

Distances for Ordinal Attributes

Assume p and q are ordinal attributes.

1. Map ordinal values to their ranks minus 1 and rename as p_r and q_r.
2. $d(p, q) = \frac{|p_r - q_r|}{n - 1}$ where n is the number of categories.

assume uniform spacing

Note:

Distances between ordinal attributes should be treated with caution as distances between categories are assumed to be uniform.

Definition 12 (Similarity Measures)

A similarity measure, $s(P, Q)$, generally has the following properties:

(i) $s(P, Q) = 1$ if and only if $P = Q$

(ii) $s(P, Q) = s(Q, P)$

Definition 13 (Cosine Similarity)

The **cosine similarity** between P and Q is defined to be:

$$s(P, Q) = \frac{\overset{\text{dot product}}{P \circ Q}}{\|P\|\|Q\|} = \cos(\theta)$$

where θ is the "angle" between P and Q.

P,Q point to same direction -> angle is 0 -> most same
P,Q cos(theta)=0 -> angle is 90
-1<=cos(theta)<=1

Definition 14 (Correlation Coefficient)

The **correlation coefficent** between P and Q is defined to be:

$$s(P, Q) = \frac{1}{n-1}(P_s \circ Q_s)$$

where P_s and Q_s are the standardized quantities

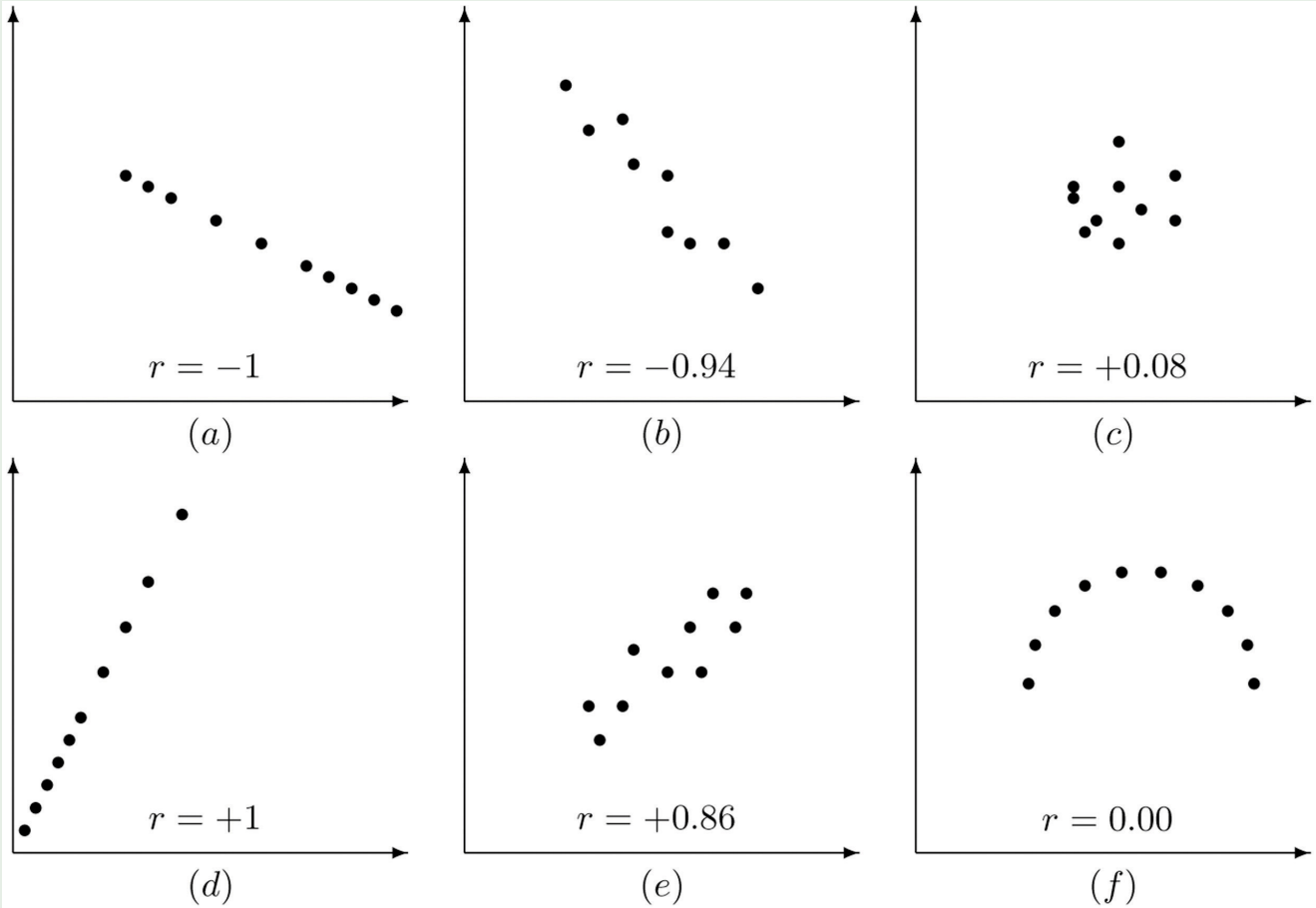
$$P_s = \frac{P - \mu_P}{\sigma_p} \quad Q_s = \frac{Q - \mu_Q}{\sigma_q}.$$

standardization

Note:
Standardization of coordinates is also often performed before computing the distance between data objects.

Example 15 (Correlation Coefficient)

Examples of scatter plots of P vs Q and corresponding correlation coefficients. -1<= r <=1



Example 16 (Cosine vs Correlation Similarity)

Consider indoor and outdoor temperatures at a given location. Assume indoor temperatures are for an unheated house in winter. The indoor temperatures are given in both Celsius and Fahrenheit.

(a) Compute the cosine similarity of all three temperature readings.

(b) Compute the correlation similarity of all three temperature readings.

(c) Which similarity measure should be used?

Definition 17 (Similarity for Binary Attributes)

Assume P and Q are data objects with binary attributes.
Let

- f_{00} number of matching 0 values.
- f_{01} number of 1's in P matching 0's in Q.
- f_{10} number of 0's in P matching 1's in Q.
- f_{11} number of matching 1 values.

Simple Matching Coefficient (SMC):

$$s(P, Q) = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Jacard Coefficient:

$$s(P, Q) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Example 18 (Asymmetric Attributes)

Which similarity measure, simple matching or Jacard coef. is a better similarity measure for the data objects P and Q given below?

$P = (0, 0, 0, 0, 1, 0, 0, 0, 1, 0)$
$Q = (0, 0, 0, 0, 1, 0, 0, 1, 0, 0)$

$$\text{SMC}(P, Q) = \frac{7+1}{7+1+1+1} = 0.8$$

$$\text{Jaccard}(P, Q) = \frac{1}{1+1+1} = 0.333$$

perfect match of 1

only count non-matched 1s

when a lot more 0s than 1s
asymmtric situation->Jaccard

Example 19 (Introduction to Pandas)

Explain what the commands in parts (a)–(d) in the `IntroductionToPandas.ipynb` notebook do.

Lesson 6 (Superbowls)

Download the file `Superbowls.csv` and use Pandas to answer the following questions.

(a) In which city was the first Superbowl played?

(b) In which city was this year's Superbowl played?

(c) Which city has hosted the Superbowl the most times?

(d) Which team has won the Superbowl the most times?

(e) Who has been named MVP the most times?

(f) What is the biggest margin of victory in a Superbowl?

(g) What is the lowest total number of points scored in a Superbowl?

(h) Come up with one interesting fact about Superbowls or a Superbowl.

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Cleaning Data

- Dealing with Missing Data
- Identifying Errors
- Removing Duplicates
- Filtering Noise

Example 20 (Cleaning Data)

Clean the data set given below.

	L1	L2	L3	L4
Bob	61.0	67.0	68.0	62.0
Jim	56.0	-100.0	NaN	53.0
Kim	71.0	76.0	72.0	78.0
Sue	NaN	92.0	93.0	-100.0
Tom	NaN	NaN	NaN	NaN
Kit	71.0	76.0	72.0	78.0

Example 21 (Filtering Noise)

The number of towed vehicles in Chicago from 09/04/2015 to 12/03/2015 is given in the file `Towed_Vehciles.csv`. Plot the number of towed vehicles vs time. Smooth the data by using a rolling average with window size equal to 10. Comment on any observed trends.

Example 22 (Pandas Data Types)

Student	Zipcode	Grade	Birthday	Credits
Smith	47803	B+	01/01/2000	160
Lee	47803	B	01/01/2002	150
Rao	60733	B+	01/01/2001	180
Chen	30700	C	01/01/2001	150
Ibu	50705	A	01/01/2003	150

Use Pandas to compute the quantities listed below (if it makes sense to do so) for the each of the four attributes: Zipcode, Grade, Birthday, Credit.

(i) mode
(ii) median
(iii) mean
(iv) Percent change from the mean to the maximum value.

Grouping Data

- Split-Apply-Combine
  - Split `all freshmen/sephomore/junior/senior together`
  - Apply
    - Aggregate
    - Transform `based on their groups->standarzation`
    - Filter `remove students in a group`
  - Combine
- Pivot Tables
- Cross Tabulation
- OLAP (Online Analytical Processing)
  - Roll-Up
  - Drill-Down
  - Slice
  - Dice

ROSE-HULMAN INSTITUTE OF TECHNOLOGY

Example 23 (Grouping Data)

Consider the data on grades given below. The data is generated by the python script `grades.py`.

name	gender	major	year	lessons	quizzes	tests	ave
Kim	F	CS	sophomore	73.75	74.25	72.33	73.33
Tom	M	CS	freshman	83.75	86.25	83.33	84.33
Bob	M	CS	sophomore	64.50	64.50	65.00	64.70
Sue	F	MA	sophomore	93.75	93.25	95.67	94.37
Jim	M	MA	freshman	55.75	53.75	53.67	54.32

Explore various grouping operations using `Pandas`.

Lesson 7 (Grouping)

A waiter collected data on his tips over a few months.[a] The tip attributes recorded were:

- tip (dollars)
- total bill (dollars)
- sex (of bill payer)
- smoker (yes if there were smokers in the party)
- day (day of the week)
- time (Lunch or Dinner)
- size (size of the party)

The data is contained in the file `tips.csv`.

(a) **Preprocessing**: Load the data into a data frame. The first column is an index column that should be removed. Make sure all columns have the correct attribute type.

(b) Add a new column named `tip_pct` that contains the percent tip given as shown below.

total_bill	tip	sex	smoker	day	time	size	tip_pct

Lesson 8 (Eye vs Hair Color)

A beauty salon has given you a contract to study the following two questions:

(i) Is eye color predictive of hair color?

(ii) Are there differences between males and females in how predictive eye color is in predicting hair color?

You decide to use actual data collected from 592 students in a statistics course. (See the file `HairEyeColor.csv`). Use pandas to construct a bar chart that tells a story. Your bar chart should have a total of 32 bars organized in 8 groups of 4 bars each. Make sure you normalize your data before plotting it. Ultimately, you want to compare blue eyed men to blue eyed women, brown eye men to brown eyed women, etc.

Example 24 (OLAP)

The revenue for a company with store locations in three cities is given below.

	city	Chicago		London		Beijing	
	store	1	2	1	2	1	2
year	quarter						
2015	Q1	5.49	7.15	6.03	5.45	4.24	6.46
	Q2	4.38	8.92	9.64	3.83	7.92	5.29
	Q3	5.68	9.26	0.71	0.87	0.20	8.33
	Q4	7.78	8.70	9.79	7.99	4.61	7.81
2016	Q1	1.18	6.40	1.43	9.45	5.22	4.15
	Q2	2.65	7.74	4.56	5.68	0.19	6.18
	Q3	6.12	6.17	9.44	6.82	3.60	4.37
	Q4	6.98	0.60	6.67	6.71	2.10	1.29

Perform roll-up, drill-down, slice and dice operations on the data.

Merging

By default, merging occurs on the intersection of columns.

- inner (intersection)
- outer (union)
- left
- right
- one-to-one
- many-to-one
- many-to-many (Cartesian product, can blow-up)

Lesson 9 (Merging)

The tables given below are generated by the `grades2.py` script located in the `Grades` data folder. You can use `%load '<path>/grades2.py'` to load the script into your Jupyter notebook.

profile

	name	ID	major	year
0	Hue	6518	MA	Y1
1	Ibu	1290	CS	Y3
2	Rao	0141	EE	Y2

tests

ID	name	T1	T2	T3
4915	Cox	67	63	65
6518	Hue	52	51	58
8711	Hue	70	74	73
1290	Ibu	82	88	81

quizzes

ID
4915
7621
6518
8711

(a) Explain what the dataframe method `.reset_index()` does?

(b) The following are examples of **one-to-one** joins. By default, the `merge` Pandas command uses the interesection of the columns of two dataframes as keys to merge on. Explain what each command is

Lesson 10 (Binning)

The `grade3.py` file generates the table on the left. Use `Pandas` to construct the table on the right. Letter grades are based on the Rose-Hulman grading scale. The top 25% of all students are rated excellent, the bottom 25% are rated fair and the rest rated as average. Ranking is given within each course.

Hint: `pd.cut` and `pd.qcut` are helpful.

	ID	name	course	ave
0	784	Abe	MA384	79
1	659	Das	CSSE490	82
2	729	Fox	CSSE490	88
3	292	Han	MA384	67
4	935	Ibu	CSSE490	70
5	863	Lee	MA384	94
6	807	May	MA384	70
7	459	Rao	CSSE490	74

	ID	name	course	ave	grade	performance	rank