Foundations of Data Science

DS 3001
Data Science Program
Department of Computer Science
Worcester Polytechnic Institute
Instructor: Prof. Kyumin Lee

Project Teams

- Clay Oshiro-Leavitt, Hunter Caouette, Nick Alescio
- Danielle Angelini, Elijah Ellis, Ryan Candy, Rob Wondolowski
- Eva (Yingbing) Lu, Manasi Danke, Erica Lee, Jonathan Dang
- Danielle Angelini, Rob Wondolowski, Elijah Ellis, Ryan Candy
- Arianna Kan, Yihan Lin, Margaret Goodwin, Ken Snoddy
- Yang Gao, Jose Li, Sarah Burns, Daniel McDonough
- Noah Puchovsky, Katherine Handy, Alex Tavares, Angelica Puchovsky
- Armando Zubillaga, Gabriel Rodirgues, Humberto Leon, Joao Omena de Lucena
- Jessie White, Lindsay MacInnis, ? , ?

- So far, 33 students expressed their preferences

Previous Class…

Data Science Process (Loop)

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

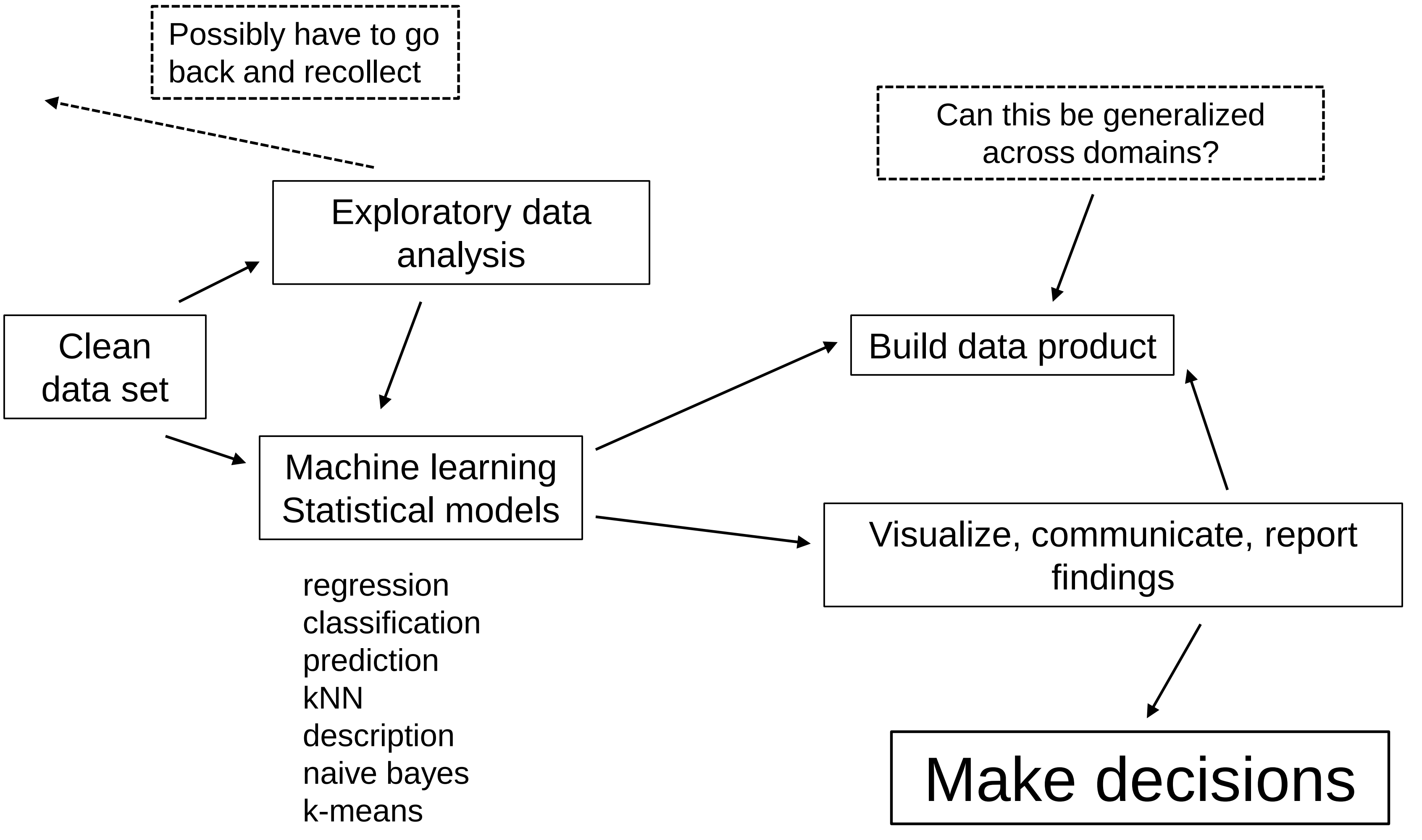
email
logs
medical records
surveys
blood drawn (microarray)
olympic records
NYT web pages

Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data Set

“clean” table

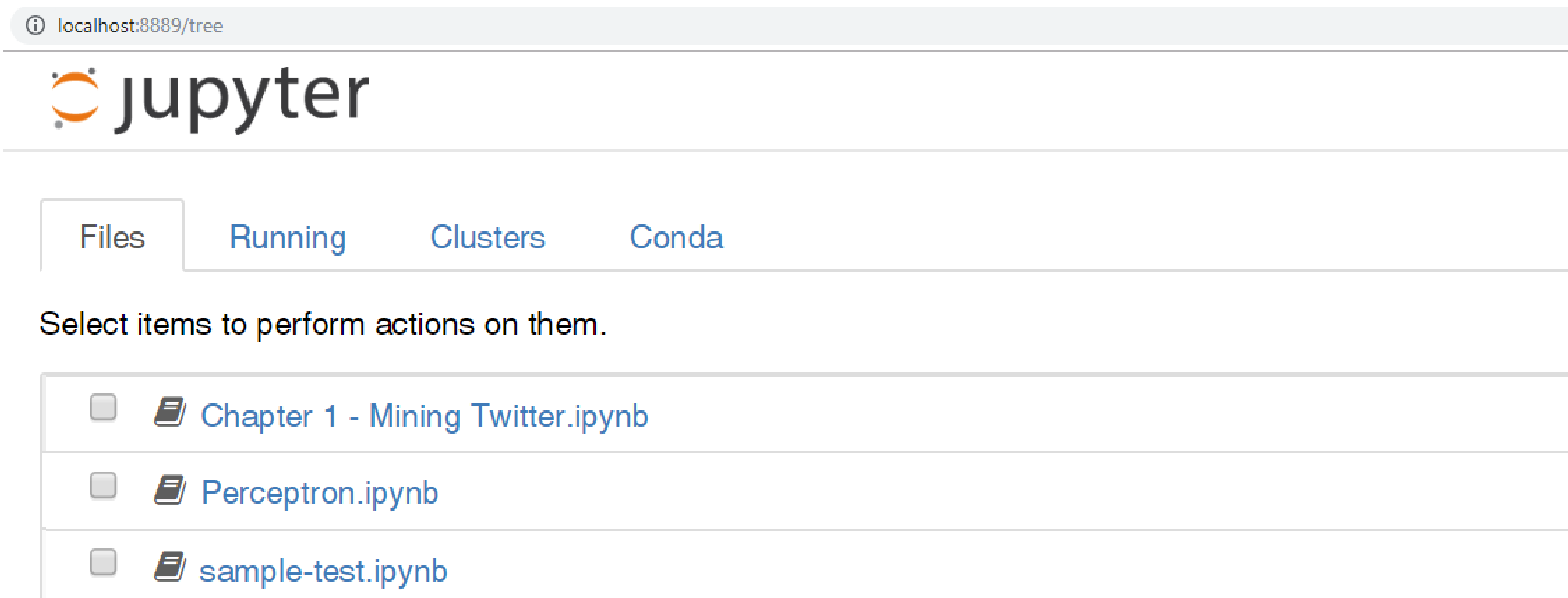

Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Data Collection

Overview

- Jupyter(IPython) Notebook
- Twitter Data
- Twitter API
- Collecting Tweets

Jupyter Notebook



https://jupyter.org/install

Learn about the Data

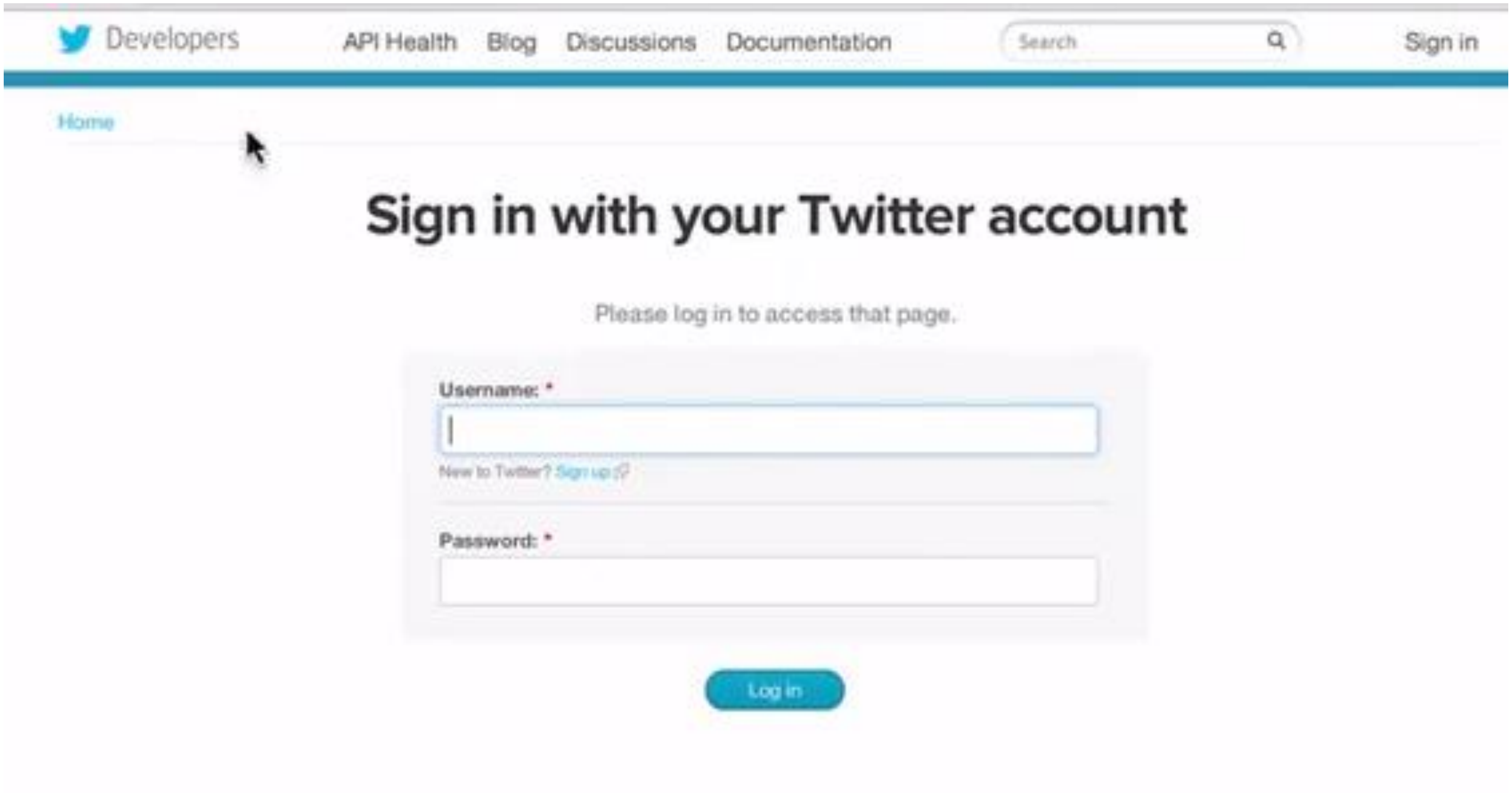
- Twitter Entities:
  - Hashtags, User mentions, URLs, Image Objects

https://twitter.com/search?q=wpi

Information in a Tweet

- User: Name, Screen Name
- Statistics: #retweets

Creating an Application



- https://dev.twitter.com/apps

How to Login: OAuth

- OAuth is an open standard for authorization
- Short for Open Authorization (OAuth)
- A standard protocol in social webs

See details: http://en.wikipedia.org/wiki/OAuth

Accessing Twitter Data from Jupyter Notebook

- Get Connected: Authorizing an application to access Twitter account data

- Download Data: Retrieving trends

- Examine the Data: Displaying API responses as pretty-printed JSON

Data Format: JSON

- JavaScript Object Notation (JSON)

- an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.
- A list of Dictionaries

```json
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1239"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

Calling Twitter APIs from Python

- Either directly call the API link, or use 3rd party library (e.g., Tweepy and python-twitter).

- In the sample code, register your own Twitter account, and Twitter app. Fill in the blanks in the code for the oauth authorization keys and secrets.

- https://developer.twitter.com/en/docs.html

Most APIs have Rate Limits

- Twitter has rate limits on their APIs
  - https://developer.twitter.com/en/docs/basics/rate-limits

In general: Look at Documentation

- API Documentation
- How can we authenticate with a token
- Most modern APIs use something like oauth

Data Science: The Context

Goal of Data Science

- Discovery of patterns and models that are:
  - Valid: hold on new data with some certainty
  - Useful: should be possible to act on the item
  - Unexpected: non-obvious to the system
  - Understandable: humans should be able to interpret the pattern

Two Major Tasks

- **Predictive** Methods (supervised learning methods)
  - Use some variables to predict unknown or future values of other variables

- **Descriptive** Methods (unsupervised learning methods)
  - Find human-interpretable patterns that describe the data
    - e.g., categorize customers by their product preferences (clustering) or understand relations (association)

Meaningfulness of Answers

- A big data mining risk is that you will "discover" patterns that are meaningless
- **Bonferroni's principle** (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

Example: Rhine Paradox

- Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had Extra-Sensory Perception (ESP).
- He devised (something like) an experiment where subjects were asked to guess 10 hidden cards – red or blue.
- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right!

Example: Rhine Paradox

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he **conclude**?
- He concluded that you shouldn’t tell people they have ESP; it causes them to lose it.

Back to Basics:

Getting to Know Our Data

Types of Datasets

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

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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

Example: Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

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**

Example: Data Matrix

- 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

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

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

Example: Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	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

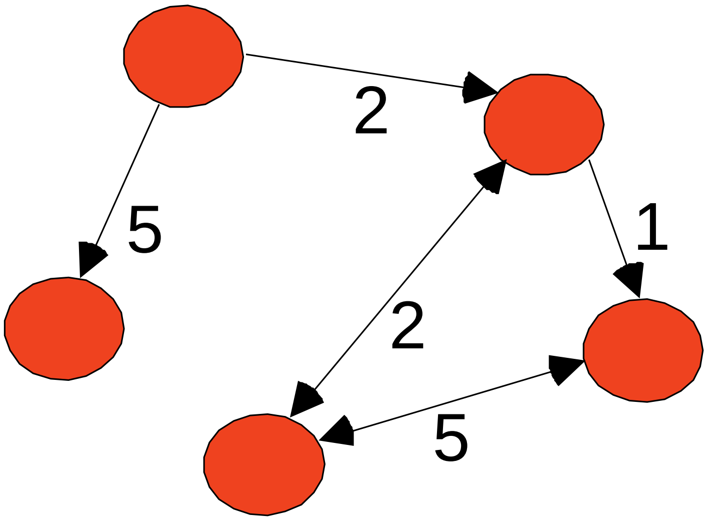
Example: Transaction Data

- A special type of record data, where each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

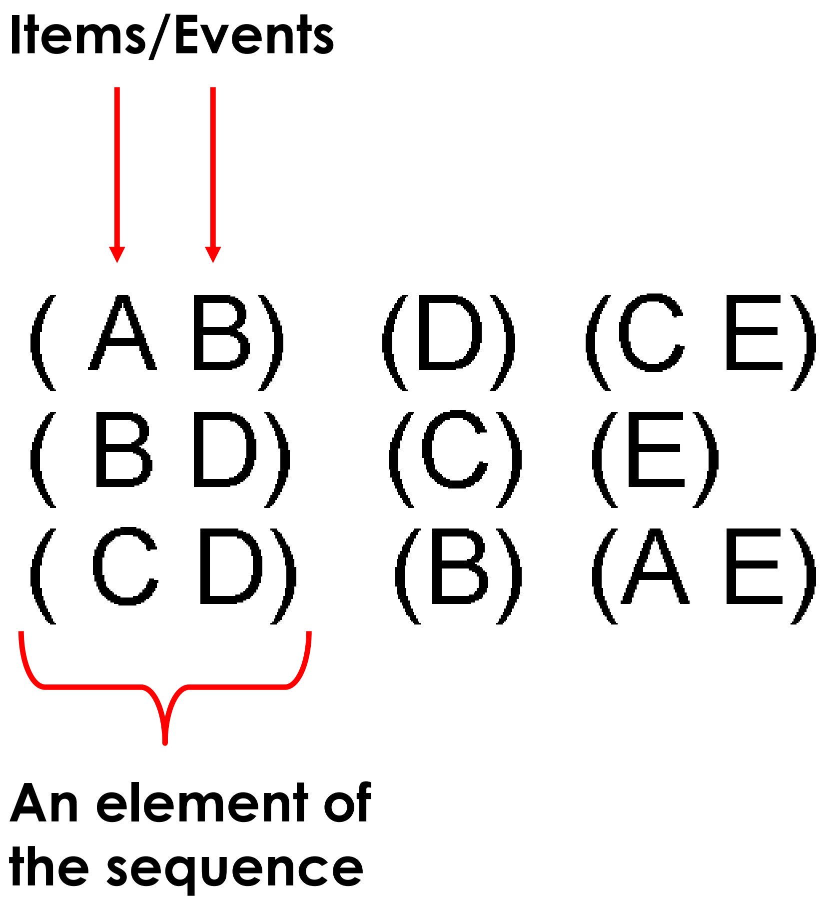
Example: Graph Data

- Examples: Generic graph and HTML Links



Example: Ordered Data

- Sequences of transactions



Example: Ordered Data

- Genomic sequence data

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

Data Objects

- Data sets are made up of data objects.
- A **data object (instance)** represents an entity.
  - Examples:
    - sales database: customers, store items, sales
    - medical database: patients, treatments
    - university database: students, professors, courses
  - Also called samples, examples, instances, data points, objects, tuples.
- Data objects are described by **attributes (features)**.
- In database… rows → data objects; columns→ attributes.

What is Data?

- 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

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

Attributes

- **Attribute** (or dimension, feature, variable): a data field, representing a characteristic or feature of a data object.
  - E.g., customer_ID, name, address
- Types:
  - Nominal / Binary (a part of Nominal)
  - Ordinal
  - Quantitative (Numeric)
    - Interval-scaled
    - Ratio-scaled

Attributes: Nominal, Ordinal, and Quantitative

- **Nominal** (categories, states, labels :: “names of things”)
  - Ex. Fruits: Apples, oranges, ...
  - Special case of Nominal: **Binary**
- **Ordinal** (Ordered)
  - Values have a meaningful order (rank), but magnitude between successive values is unknown
  - Quality of meat: Grade A, AA, AAA
- **(Q) Interval** (No true zero-point)
  - Calendar dates: Jan 24, 2012; Location (Lat/Long)
  - Only differences (intervals) may be compared
- **(Q) Ratio** (Inherent zero-point)
  - Physical measurements: Length, Mass, ...
  - Counts and amounts

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countable set of values
    - E.g., zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as floating-point variables.

Quiz! Census Data

- **People**: # of people in group
- **Year**: 1850 – 2000 (every decade)
- **Age**: 0 – 90+
- **Sex**: Male, Female
- **Marital Status**: Single, Married, Divorced

Quiz! Census Data

- **People**
- **Year**
- **Age**
- **Sex**
- **Marital Status**

- 2,348 data points

	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976
27	1850	60	0	2	162236

Census: N, O, Q?

- **People**
- **Year**
- **Age**
- **Sex (M/F)**
- **Marital Status**

Census: N, O, Q?

• **People**	**Q-Ratio**
• **Year**	**Q-Interval (O)**
• **Age**	**Q-Ratio (O)**
• **Sex (M/F)**	**N**
• **Marital Status**	**N**

Basic Statistical Descriptions of Data*

*(These are mainly for understanding individual attributes)

Basic Statistical Descriptions of Data

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

Measuring the Central Tendency

- **Mean** (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  Note: *n* is sample size and *N* is population size. $\mu = \frac{\sum x}{N}$

  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

  - Trimmed mean: chopping extreme values

Measuring the Central Tendency

- **Median**:
  - Middle value if odd number of values, or average of the middle two values otherwise

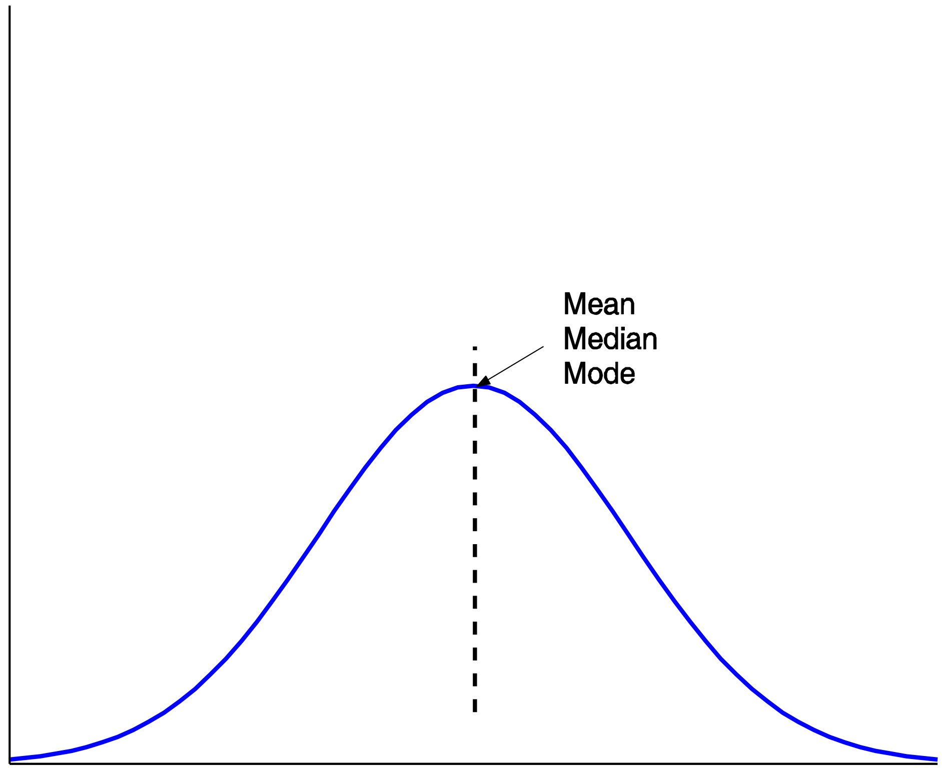
age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Measuring the Central Tendency

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

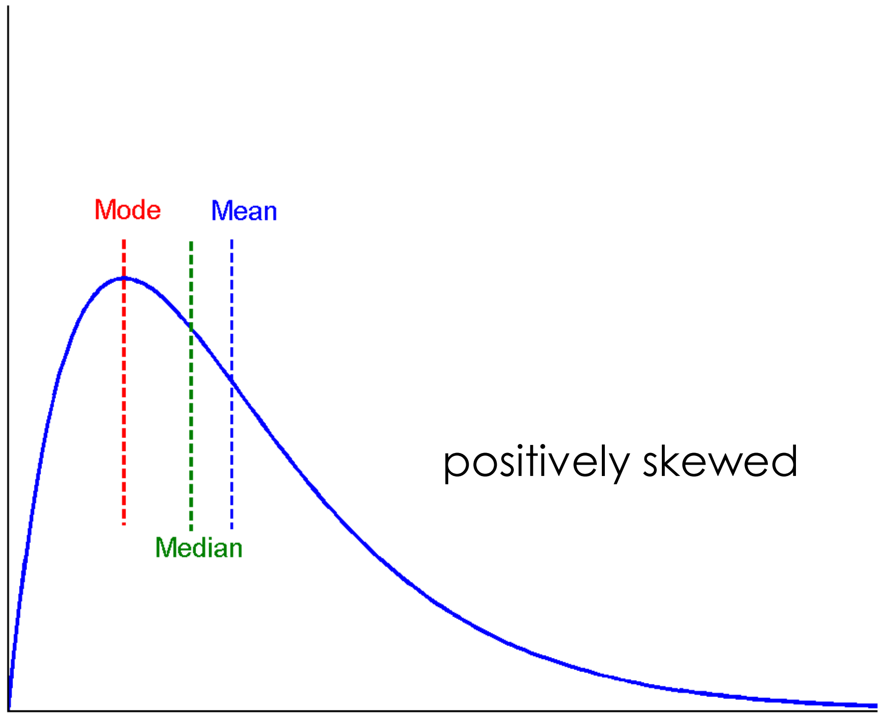
Symmetric vs. Skewed Data

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



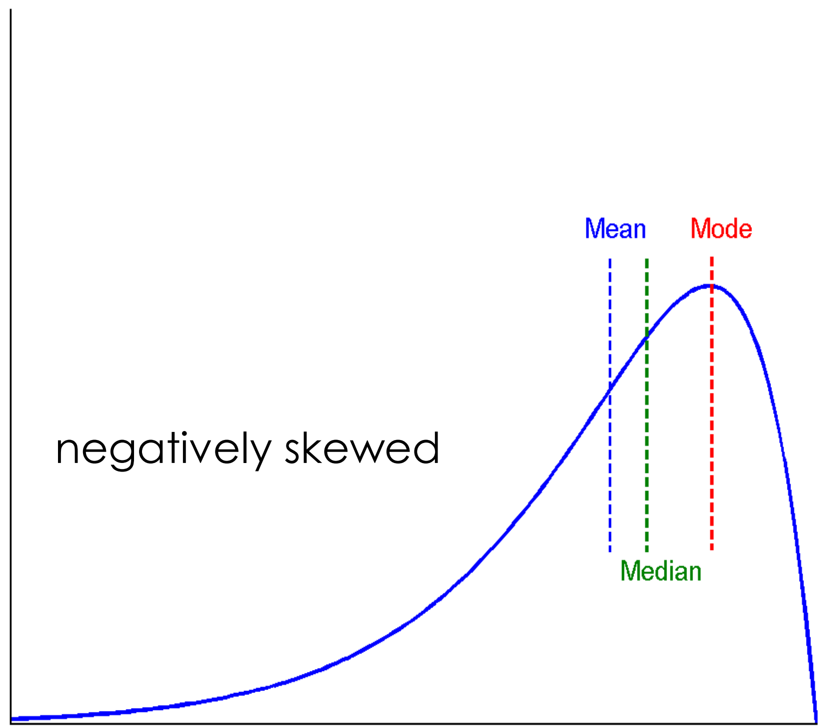
Symmetric vs. Skewed Data

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



Symmetric vs. Skewed Data

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

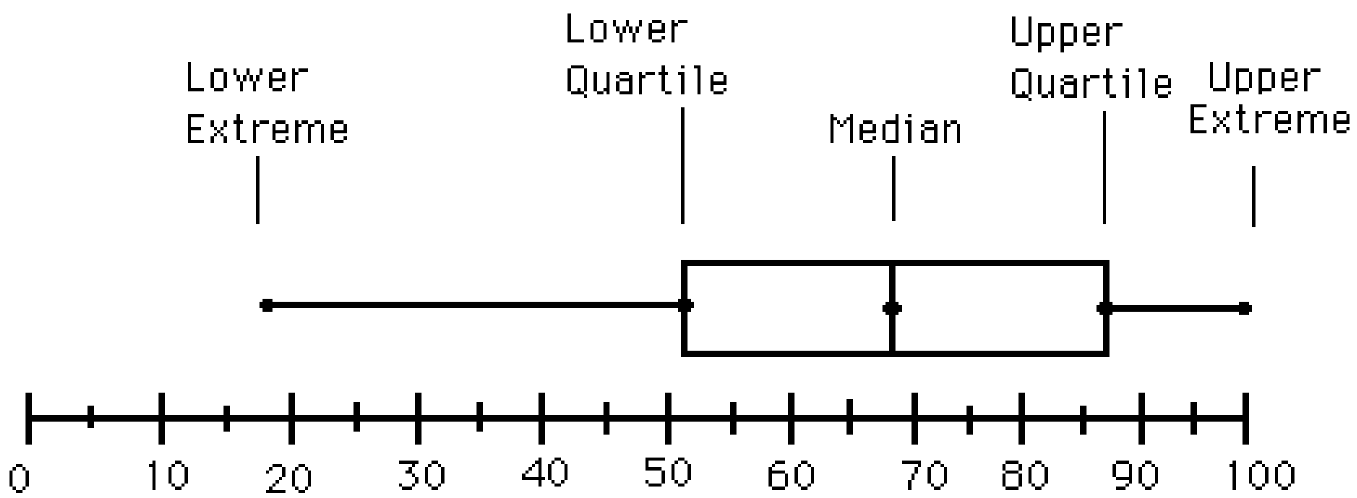


Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Five number summary: min, Q_1, median, Q_3, max
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
    - Below $Q_1 - 1.5\times IQR$, or Above $Q_3 + 1.5\times IQR$

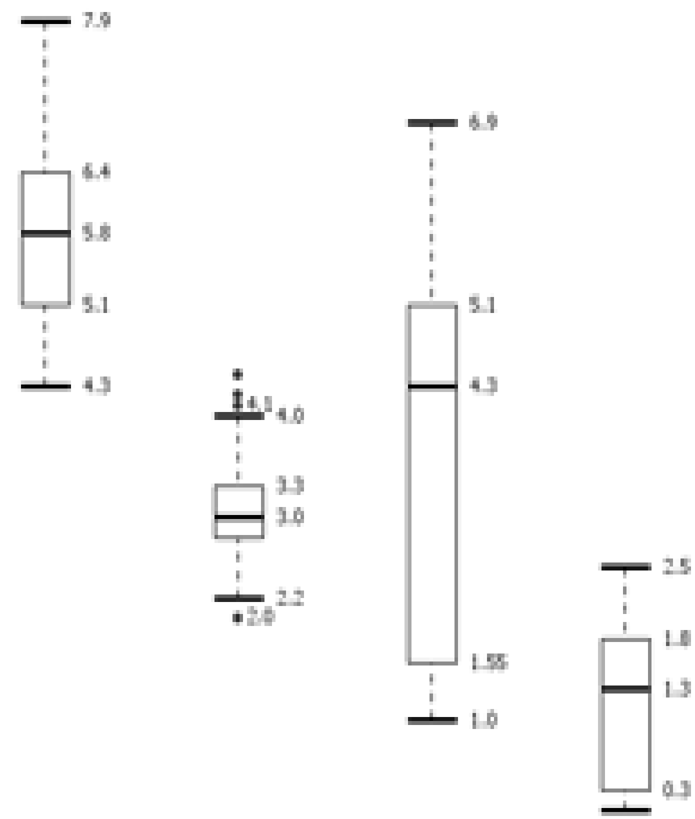
Boxplot

- Five-number summary of a distribution
  - Minimum, Q1, Median, Q3, Maximum



Boxplot Analysis

- Data is represented with a box
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extended to Minimum and Maximum
- Outliers: points beyond a specified outlier threshold, plotted individually



Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^{\ 2} - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

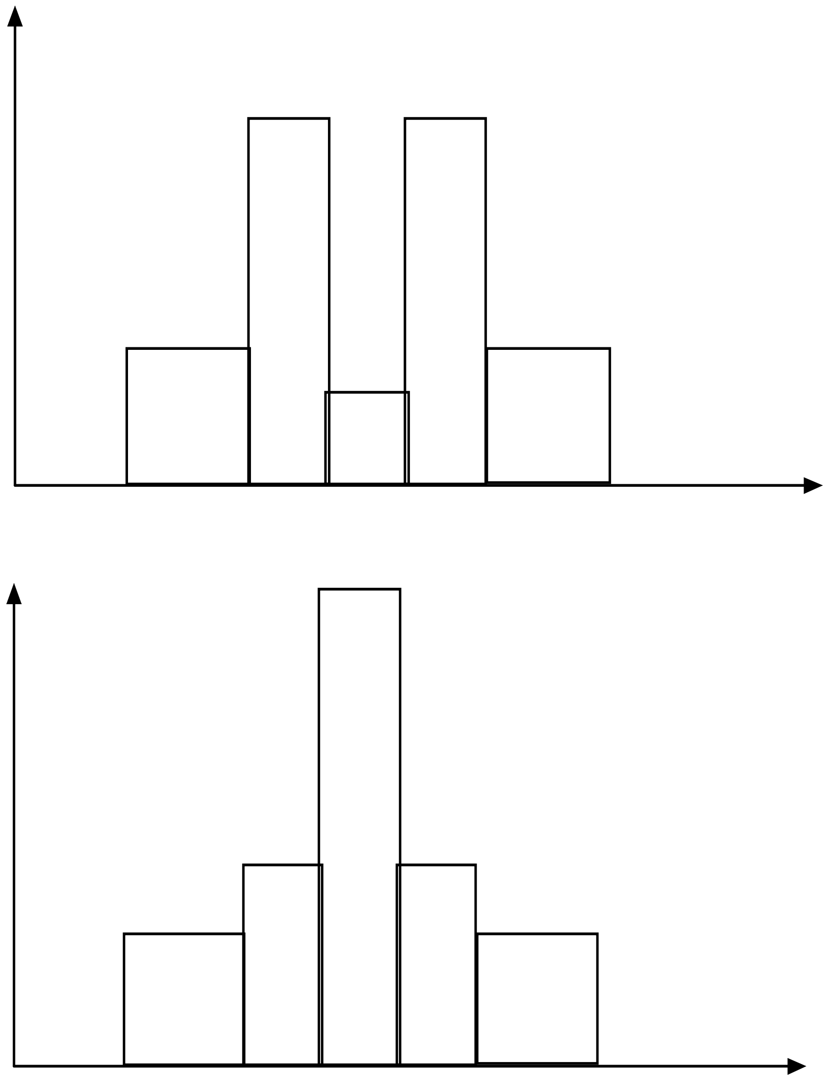
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\ 2} - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance s^2 *or* σ^2

Histogram

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

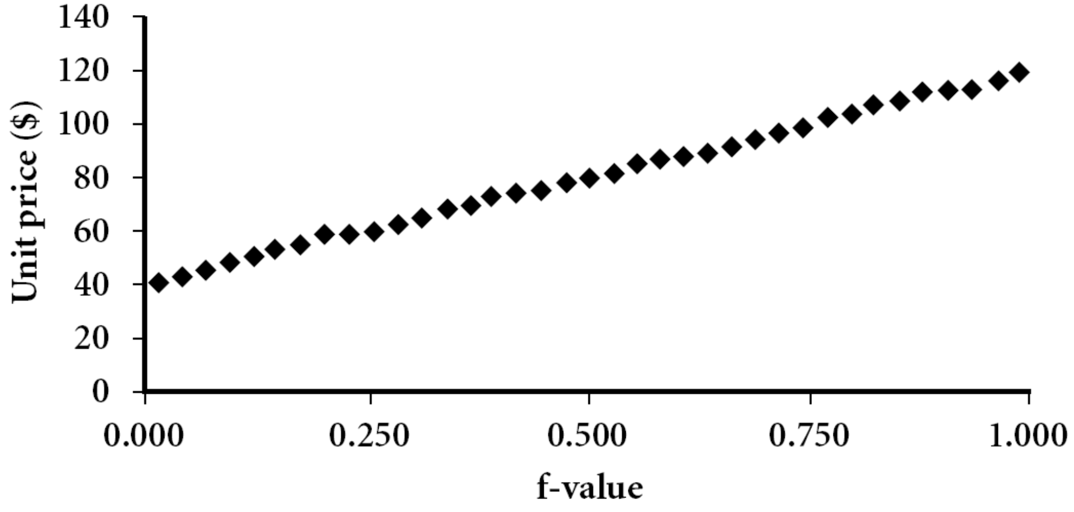
Histograms Often Tell More than Boxplots



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

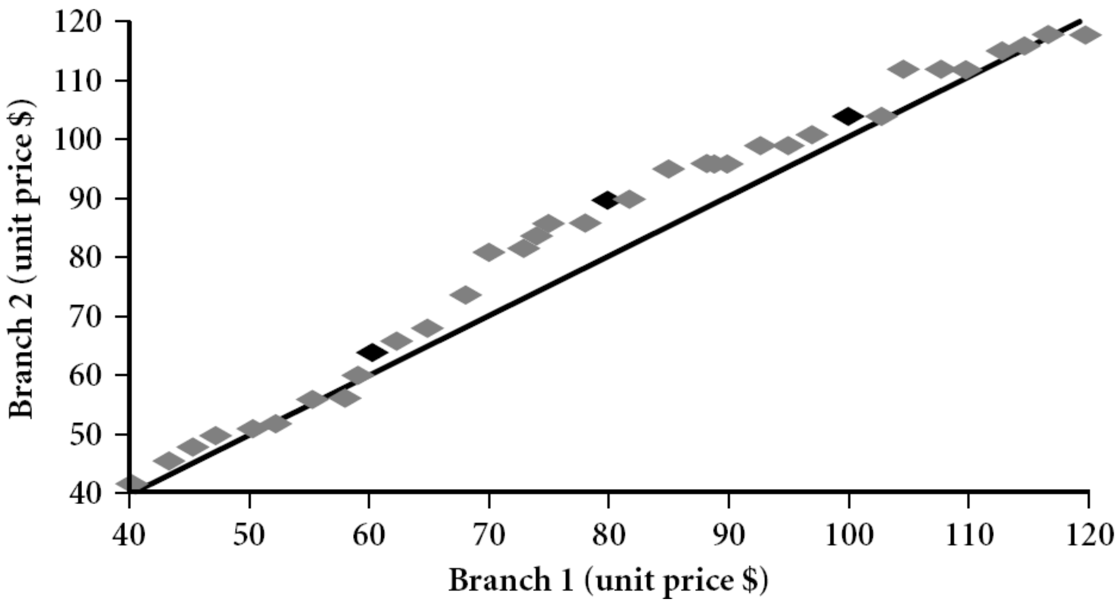
Quantile Plot

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



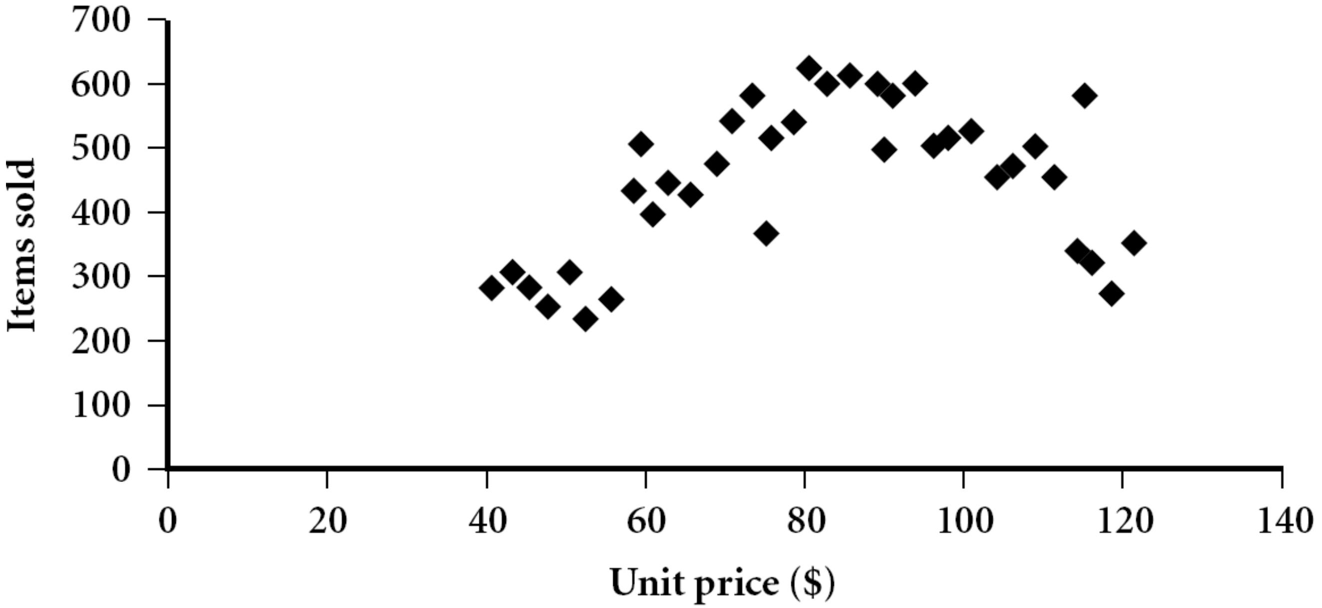
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

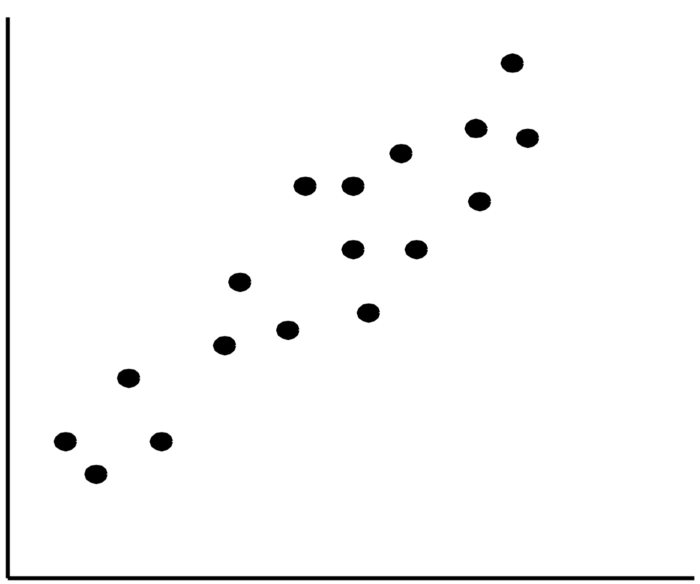


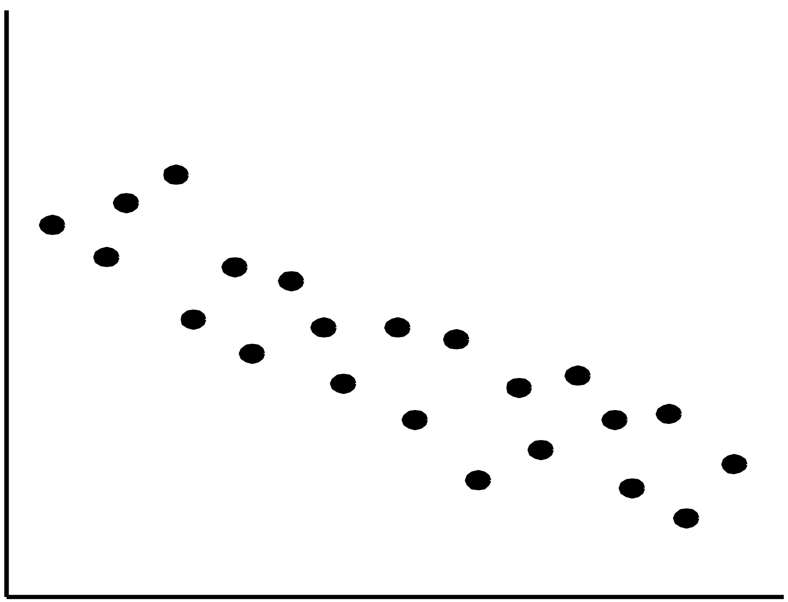
Scatter plot

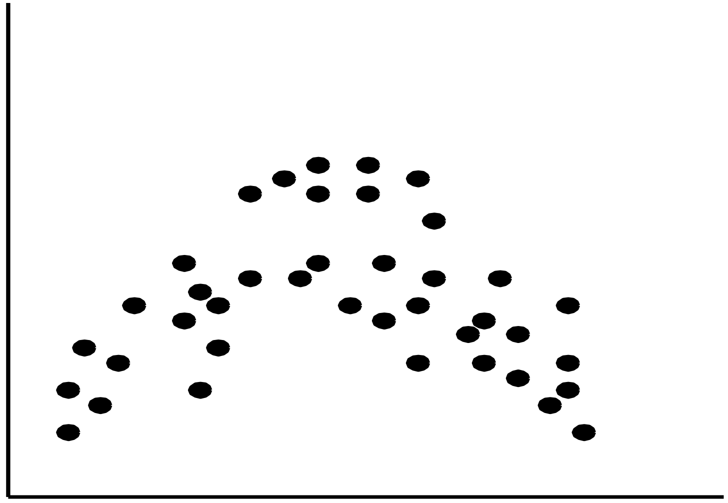
- Each pair of values (of two numeric attributes) is treated as a pair of coordinates and plotted as points in the plane
- Provides a first look at bivariate data to see clusters of points, outliers, etc



Positively and Negatively Correlated Data

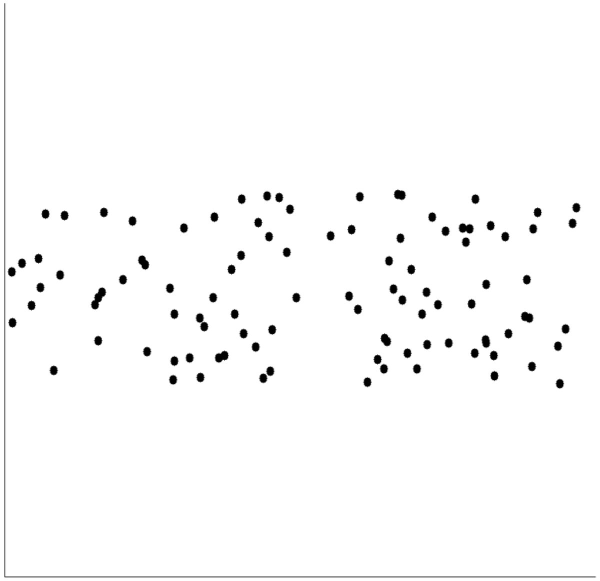


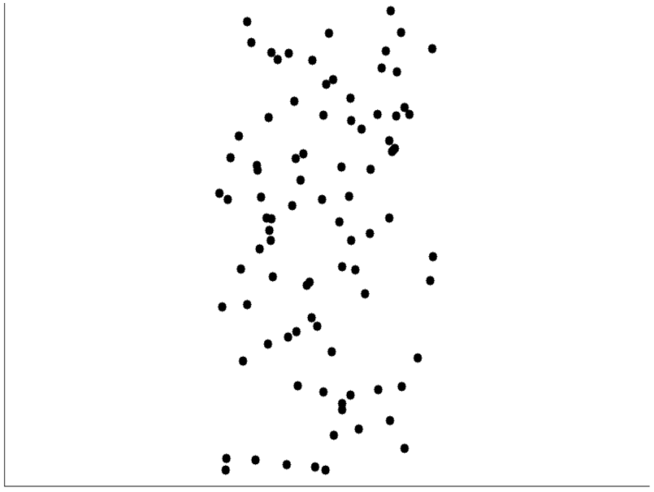




- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data





Data Science: The Context

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

→ Raw Data is Collected / Recorded

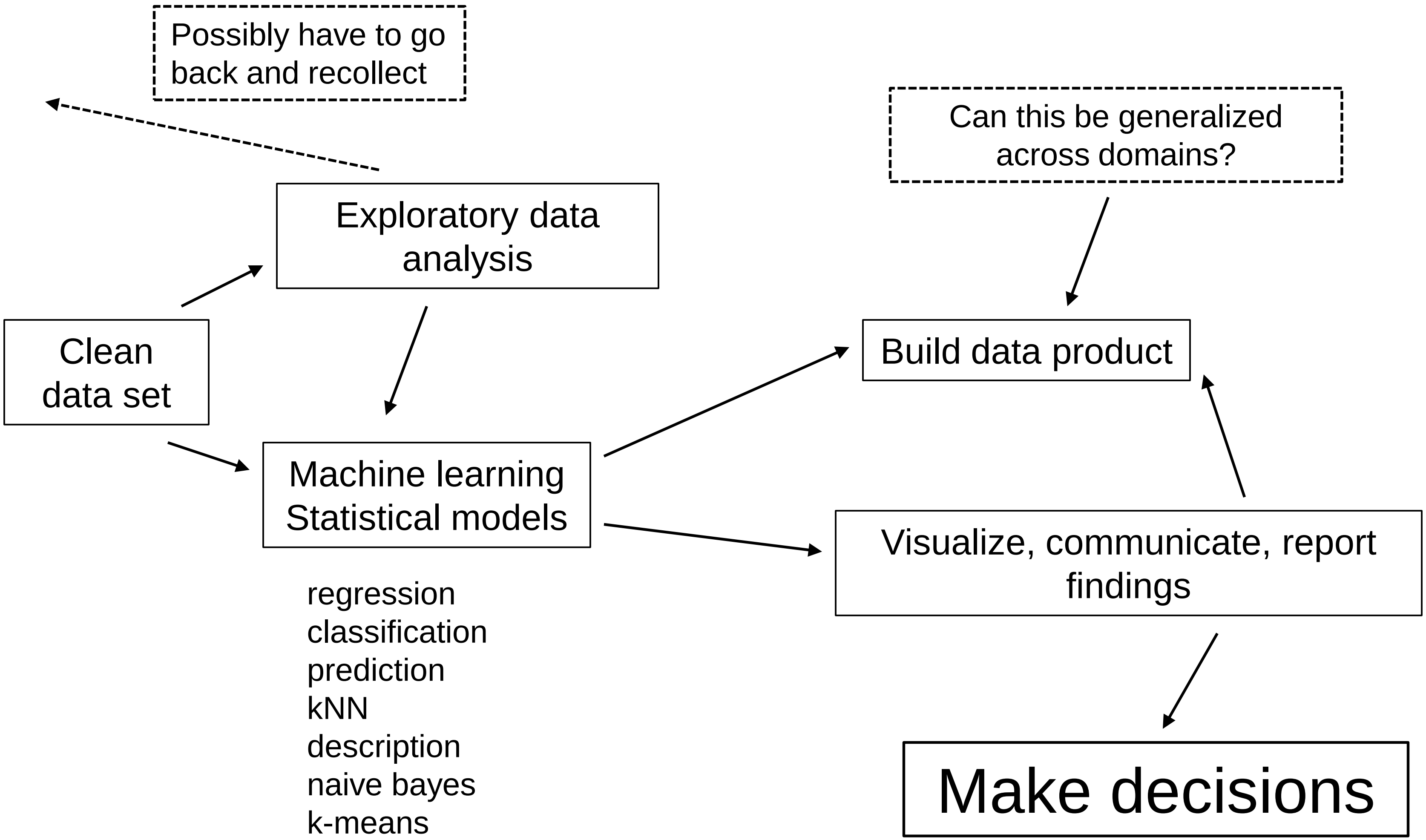
email
logs
medical records
surveys
blood drawn (microarray)
olympic records
NYT web pages

→ Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

→ Data Set

“clean” table


Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Data Preprocessing: Overview

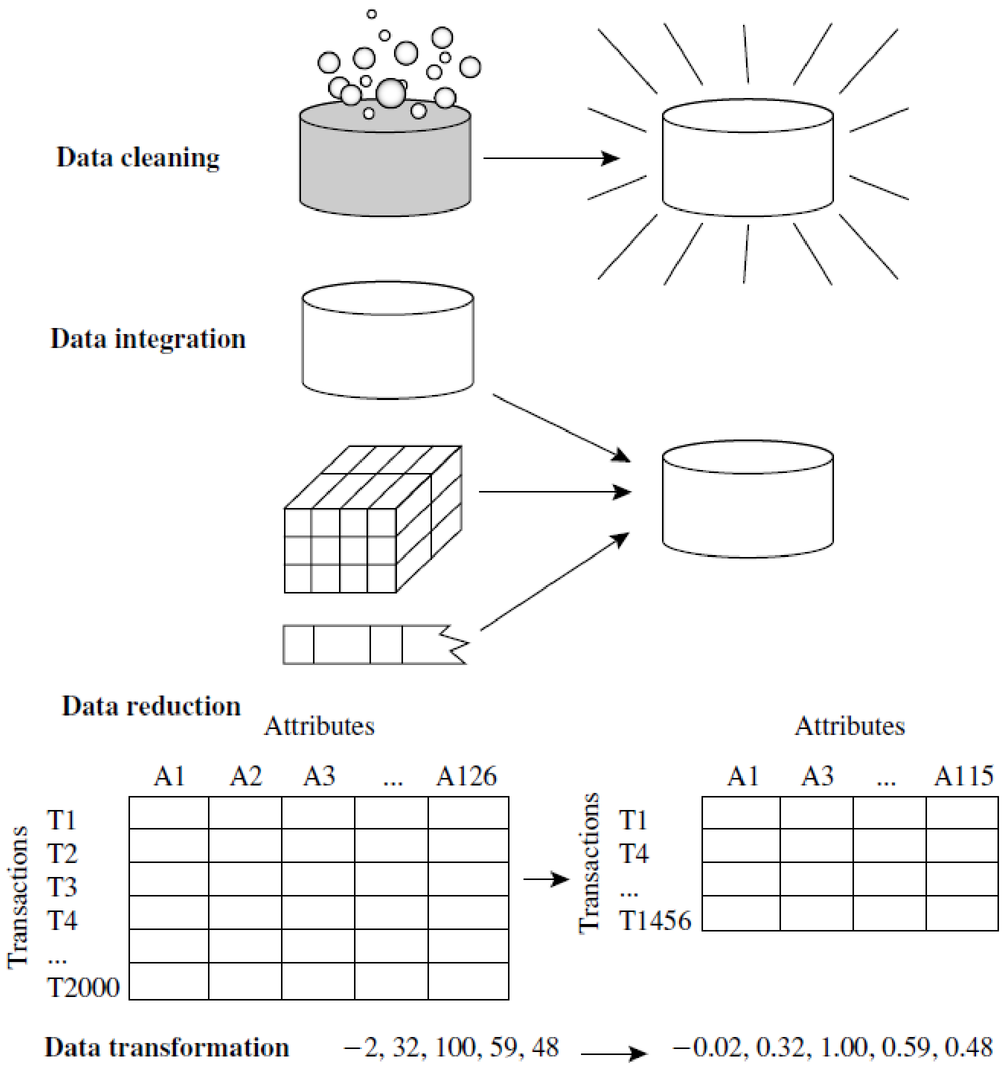
Why Preprocess the Data?

- **Measures for data quality: A multidimensional view**
  - **Accuracy**: correct or wrong, accurate or not
  - **Completeness**: not recorded, unavailable
  - **Consistency**: some modified but some not
  - **Timeliness**: timely update?
  - **Believability**: how trustable the data are correct?
  - **Interpretability**: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases/data sources, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization

Forms of Data Preprocessing



Data Cleaning

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - **noisy**: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - **inconsistent**: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - Discrepancy between duplicate records
  - **Intentional** (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples/instances have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

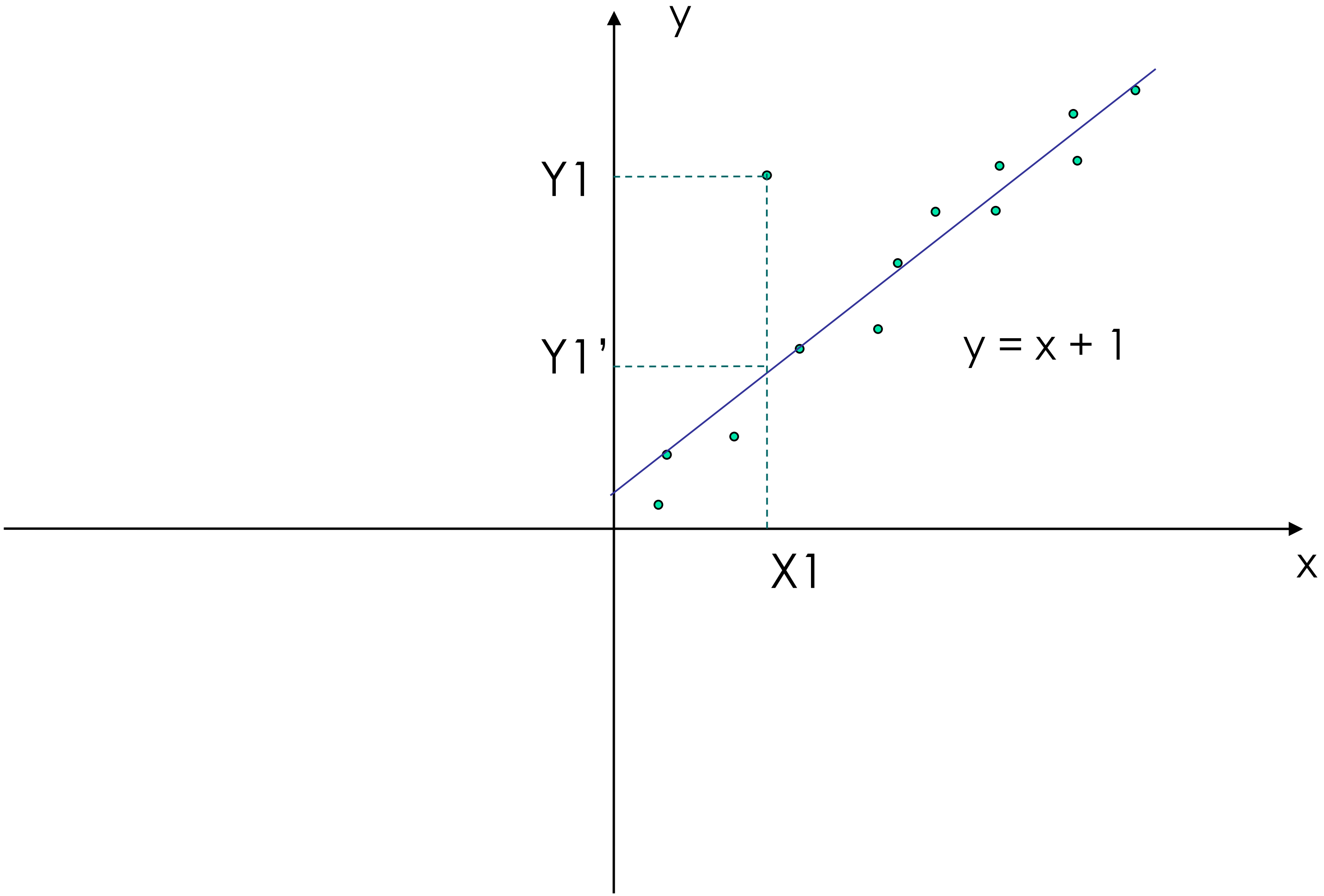
Noisy Data

- **Noise**: random error or variance in a measured variable.
- **Incorrect attribute values** may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - etc
- **Other data problems** which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Regression



Cluster Analysis