

# Foundations of Data Science

## Lecture 3, Module 1

### Fall 2022

Rumi Chunara, PhD

*Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.*

# Data Preprocessing

# Major Tasks in Data Preprocessing

- Data sampling
- Data cleaning
- Data integration
- Data reduction

# Major Tasks in Data Preprocessing

- Data sampling covered in our last class
- Data cleaning
- Data integration
- Data reduction

# Major Tasks in Data Preprocessing

- Data sampling
- **Data cleaning**
- Data integration
- Data reduction

# What is Data Cleaning?

*“Data cleansing or data cleaning is the process of detecting and repairing corrupt or inaccurate records from a data set in order to improve the quality of data.”*

[<https://en.wikipedia.org/wiki/Data\\_cleansing>](https://en.wikipedia.org/wiki/Data_cleansing) & Erhard Rahm, Hong Hai Do: Data Cleaning: Problems and Current Approaches, IEEE Data Engineering Bulletin, 2000.

*“[...] data is generally considered high quality if it is fit for [its] intended uses in operations, decision making and planning”.*

Thomas C. Redman, Data Driven: Profiting from Your Most Important Business Asset. 2013

*“Even though quality cannot be defined, you know what it is.”*

Robert M. Prisig, Zen and the Art of Motorcycle Maintenance, 1975

*“Data of poor quality is lacking rich metadata.”*

Divesh Srivastava, AT&T Research

# Data Cleaning

MAR 23, 2016 @ 09:33 AM 15,078 VIEWS

## Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says



**Gil Press, CONTRIBUTOR**

I write about technology, entrepreneurs and innovation. [FULL BIO](#)

Opinions expressed by Forbes Contributors are their own.

TWEET THIS

data scientists found that they spend most of their time massaging rather than mining or modeling data.

76% of data scientists view data preparation as the least enjoyable part of their work

A new survey of data scientists found that they spend most of their time massaging rather than mining or modeling data. Still, most are happy with having [the sexiest job of the 21<sup>st</sup> century](#). The survey of about 80 data scientists was conducted for the second year in a row by CrowdFlower, provider of a “data enrichment” platform for data scientists. Here are the highlights:

### ***Least enjoyable part of Data Science?***

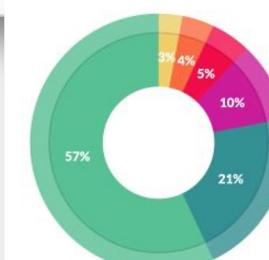
Collecting data (**21%**)

Cleaning and organizing data (**57%**)

### ***Spend most time doing***

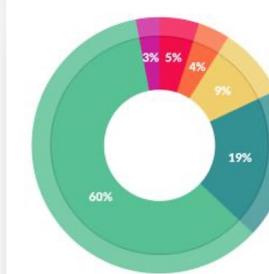
Collecting data (**19%**)

Cleaning and organizing data (**60%**)



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

*Why is Data Cleaning so  
Difficult and Time Consuming?*

# Many reasons...

1. Variety of **data formats and sources**
  - Requires specific scripts for individual sources
2. Different **data quality issues** to treat
  - Wrong types, outliers or missing values, duplicated data
3. Cleaning the data requires **domain knowledge**
4. If using large volumes of data, **automated methods that encode the notion of high quality data** may be required

# Things to Keep in Mind

- Purpose
  - What is the intended use of the data?
- Domain Knowledge
  - Learn as much as possible about the data  
(generation process, exploratory data analysis)
- Attention
  - Be mindful and pay attention to details

And remember: If you don't need to, **don't clean it!**

# Example: Jobs in NY



## DOB Job Application Fillings

**Challenge:** identify and replace all incorrect spellings.

- String similarity algorithms:
  - Edit-distance based (e.g. Levenshtein Distance)
  - Token based
  - Sequence based
- Soundex (<https://en.wikipedia.org/wiki/Soundex>)

```
name
-----
-----
B2ROOKLYN
B4ROOKLYN
BBBROOKLYN
BBROOKLYN
BERKELEY
BERKELEY
HEIGHT
BERKELEY
HTS
BERKELEY
HTS.
BERKLEY
BERKLEY
HEIGHTS
BERKLEY
```

```
BRKKLYN
BRKLLYN
BRKLN
BRKLY
BROOKLYNTCHEN
BROOKLYNY
BROOKLYON
BROOKLYTN
BROOKLYU
BROOKLYNN
BROOKOLYN
BROOKYL
BROOKYLN
BROOKYLYN
BROOOKLYN
BROOOKLN
BROOKLYN
```

# Data Cleaning: Missing Data Imputation

- Imputation is the process of **replacing missing data** with substituted values.

ItemID	1	2	3	4	5	6
1	NaN	4.0	2.0	5.0	1.0	
2	4.0	NaN	2.0	4.0	1.0	
3	5.0	NaN	4.0	4.0	5.0	
4	NaN	4.0	NaN	NaN	NaN	
5	3.0	3.0	3.0	4.0	2.0	
6	NaN	5.0	4.0	NaN	NaN	

- **Delete rows:** “... *the most common means of dealing with missing data is deletion.*”
- **Hot-deck:** “... *a missing value is imputed from a randomly selected similar record.*”
- **Mean substitution:** “...*replace a missing value with the mean of that variable for all other cases.*”
- **Regression:** “*A regression model is estimated to predict observed values of a variable based on other variables.*”

**Exploratory Data Analysis** can give you insights on what strategy to use!

[https://en.wikipedia.org/wiki/Imputation\\_\(statistics\)](https://en.wikipedia.org/wiki/Imputation_(statistics))

F. Brieseman et al., "Deep" Learning for Missing Value Imputation in Tables with Non-Numerical Data, CIKM 2018.

- Sometimes different text representations of NULL encode different semantics
  - **unknown**: there is a value, but I do not know it (e.g., unknown date-of-birth)
  - **not applicable**: there is no meaningful value (e.g., spouse for singles)
  - **withheld**: there is a value, but we are not authorized to see it (e.g., private phone line)

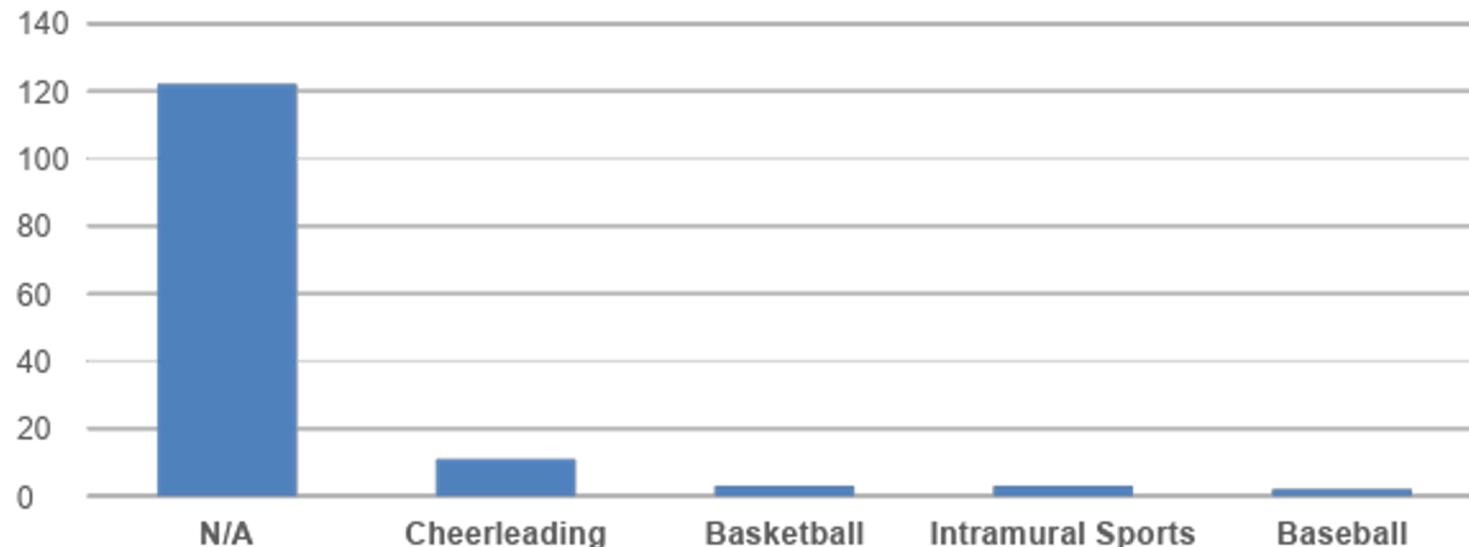
Copied from: Data Fusion - VLDB 2009 Tutorial - Luna Dong & Felix Naumann

# Data Cleaning: Outliers



DOE High School Directory 2013-2014

## school\_sports



# Frequent Outliers in Datasets

- Values that occur as high frequency outliers
  - Values that occur with frequency >50% in + 15,000 columns of NYC Open Data datasets (in Nov. 2016).

0	262	Columns
N/A	71	- " -
UNSPECIFIED	67	- " -
S	57	- " -
-	50	- " -
0.00	47	- " -
NY	38	- " Not dirty! Not all outliers need to be cleaned.
1	25	- " -
0.0	20	- " -
IND	12	- " -
CLOSED	10	- " -
100	8	- " -
NOT AVAILABLE	8	- " -

# Outliers that are Data Quality Issues

Dataset	Statistic	Trip Duration (min)	Trip Distance (mi)	Fare Amount (US\$)	Tip Amount (US\$)
2008	Min	0.00	0.00	0.00	0.00
	Avg	16.74	2.71	0.09	0.10
	Max	1440.00	50.00	10.00	8.75
2009	Min	0.00	0.00	2.50	0.00
	Avg	7.75	6.22	6.04	0.38
	Max	180.00	180.00	200.00	200.00
2010	Min	-1,760.00	-21,474,834.00	-21,474,808.00	-1,677,720.10
	Avg	6.76	5.89	9.84	2.11
	Max	1,322.00	16,201,631.40	93,960.07	938.02
2011	Min	0.00	0.00	2.50	0.00
	Avg	12.35	2.80	10.25	2.22
	Max	180.00	100.00	500.00	200.00
2012	Min	0.00	0.00	2.50	0.00
	Avg	12.32	2.88	10.96	2.32
	Max	180.00	100.00	500.00	200.00

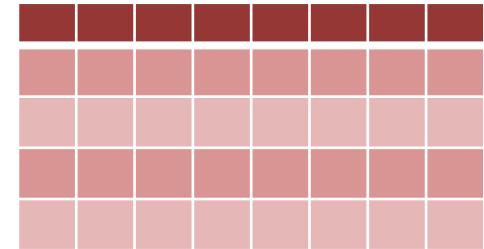
NYC Taxi Data quality issues [Freire et al., IEEE DEB 2016]

# Data Cleaning: Duplicate Records

- **Problem:** *Given one or more data sets, find all sets of records that represent the same real-world entity.*

- **Main Difficulties**

- Duplicates are not identical
- Large volume, cannot compare all sets of records



# Solutions to Duplicate Records

***Decide if two records represent the same entity***

Name	Street	Zip	City	Phone
James Smith	5 <sup>th</sup> Ave	10011	Manhattan	351-344-5671
Jane Smith	6 <sup>th</sup> Ave	10011	Manhattan	351-244-4674
Smith, J.	Fifth Avenue	10011	NYC	+1 351.344.5671

- **Standardize data**
- **Apply different similarity measures**
  - Similarity measures – Levenshtein, Soundex, Jaccard, etc.
  - Tokenize values
- **Different weights for attributes (apply domain knowledge)**
  - Two records with the same telephone number are more likely to be duplicates than records with the same (ZIP, City).

# Major Tasks in Data Preprocessing

- Data sampling
- Data cleaning
- **Data integration**
- Data reduction

# Data Integration

- Combines **data from multiple sources** into a coherent story
- Example: Auctus <https://auctus.vida-nyu.org/>

The screenshot shows the Auctus interface with a search bar for "nyc taxi". Below the search bar are advanced search filters: Any Date, Any Location, Related File, Source, and Data Type. A message in a red box states: "Yellow taxis only cover Manhattan!".

**2017 Yellow Taxi Data (13.3 mb)**  
upload  
This dataset includes trip records from all trips completed in yellow taxis in NYC during Jul-Dec... [Show more...](#)

Columns: `tpep_pickup_datetime` # PULocationID # n. trips

Data Types: # Numerical # Temporal

Download View Details Search Related

**NYC Taxi and Limousine Commission authorized Dispatch Service Providers (DSP) (615.0 bytes)**  
data.cityofnewyork.us  
A Dispatch Service Provider (DSP) can dispatch trips on behalf of the FHV (For-Hire-Vehicle) Base.

Columns: # Licensee Number # Licensee Name # Alternate Name Of Licensee # Building

Show 17 more columns...

# Numerical # Spatial # Temporal

Download View Details Search Related

**2017 Yellow Taxi Data**  
ID: datamart.upload.dd6b73540cff4fe59e71892e4a75047b  
Source: upload  
Last Updated Date: 8/29/2019, 4:13:48 PM  
Description: This dataset includes trip records from all trips completed in yellow taxis in NYC during Jul-Dec...  
[Show more...](#)

Data Types: # Numerical # Temporal

Columns: tpep\_pickup\_datetime # PULocationID # n. trips

Rows: 526615  
Size: 13.3 mb

Download CSV D3M

**Dataset Sample (>):**

Compact View Detail View Column View

tpep_pickup_datetime	PULocationID	n. trips
Text Enumeration Datetime hour	Integer Identifier	Integer
2017-07-02 10:00:00	239	260

# Data Integration

- Combines **data from multiple sources** into a coherent story
- Example: Auctus <https://auctus.vida-nyu.org/>

The screenshot shows the Auctus platform interface for searching datasets. The search bar at the top contains "nyc taxi". Below the search bar are advanced search filters: Any Date, Any Location, Related File, Source, and Data Type. The results section indicates "About 21 results". The first result is "2017 Yellow Taxi Data (13.3 mb)" with an "upload" link. A detailed description follows, mentioning trip records from July-Dec 2017. Below the description are download and view details buttons. A modal window titled "Search Related Datasets" is open over the results. It displays the selected dataset ("2017 Yellow Taxi Data") and its columns ("# Numerical", "# Temporal"). It also shows the available columns ("# n. trips") and the selected columns ("# tpep\_pickup\_datetime", "# PUlocationID"). A tooltip in the modal states: "You can search for more data that is integrable through join or union operations, ending up with more data to analyze!" At the bottom of the modal are "Join" and "Union" buttons, with "Join" being highlighted. The main results table below the modal lists various dates and trip counts.

Date	Trips
2017-07-08 01:00:00	158
2017-07-20 15:00:00	82
2017-07-22 17:00:00	219
2017-07-30 13:00:00	164

# Data Integration

- Combines **data from multiple sources** into a coherent story
- Example: Auctus <https://auctus.vida-nyu.org/>

The screenshot shows the Auctus interface with a search bar containing "nyc taxi". Below the search bar are advanced search filters: Any Date, Any Location, Related File, Source, and Data Type. There are also buttons for "Related File (edit)", "Download", "View Details", "Search Related", and "Augment Options".

The main search results are displayed in cards:

- 2017 Green Taxi Trip Data (1.0 gb)**  
data.cityofnewyork.us  
This dataset includes trip records from all trips completed in green taxis in NYC in 2017. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Livery Passenger Enhancement Program (LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data. [Show less](#)  
# VendorID    lpep\_pickup\_datetime    lpep\_dropoff\_datetime  
Abc store\_and\_fwd\_flag    [Show 15 more columns...](#)  
Categorical    Numerical    Temporal  
Download    View Details    Search Related    Augment Options
- 2016 Green Taxi Trip Data (2.1 gb)**  
This dataset includes trip records from all trips completed in green taxis in NYC in 2016. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Livery Passenger Enhancement Program (LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data. [Show less](#)  
# VendorID    lpep\_pickup\_datetime    lpep\_dropoff\_datetime  
Abc store\_and\_fwd\_flag    [Show 15 more columns...](#)  
Categorical    Numerical    Temporal  
Download    View Details    Search Related    Augment Options

A red callout box highlights a note about the 2017 dataset: "Found 2017 taxi data (now for other boroughs) joinable on temporal columns!"

**2017 Green Taxi Trip Data**

ID: datamart.socrata.data-cityofnewyork-us.5gj9-2kzx  
Source: data.cityofnewyork.us  
Last Updated Date: 3/23/2020, 6:11:00 PM

Description: This dataset includes trip records from all trips completed in green taxis in NYC in 2017. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Livery Passenger Enhancement Program (LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data. [Show less](#)

Data Types: Categorical    Numerical    Temporal

Columns: VendorID    lpep\_pickup\_datetime    lpep\_dropoff\_datetime    Abc store\_and\_fwd\_flag    RatecodeID    PUlocationID  
DOlocationID    passenger\_count    fare\_amount    [Show 10 more columns...](#)

Rows: 11740668  
Size: 1.0 gb  
Download: CSV    D3M

Dataset Sample (↓):

# Data Integration: Entity Identification

- Entity identification problem:
  - Identify real world entities from **multiple data sources**, e.g., **Bill Clinton = William Clinton**
  - For the same real-world entity, attribute values from different sources can be different
    - Example: a same object can have different weight values depending on whether it was weighed in kilograms or pounds!
  - **Classification techniques** and **statistical pattern recognition** can be used to tackle this problem

# Handling Redundancy in Data Integration

- Redundant data occurs often during integration
  - *Object identification*: the same attribute or object may have different names in different databases
  - *Derivable data*: one attribute may be a “derived” attribute in another table through, e.g., aggregation
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve processing speed, quality and interpretability

# Correlation Analysis (Categorical Data)

- $\chi^2$  (chi-square) statistic tells you how much difference exists between observed counts for two categorical variables and the counts you would expect if there were no relationship at all between them.
  - Example:

	plays_chess (yes)	plays_chess (no)	Sum
likes_scifi (yes)	250 (90)	200 (360)	450
likes_scifi (no)	50 (210)	1000 (840)	1050
Sum (col.)	300	1200	<b>1500</b>

- Variables: *plays\_chess* and *likes\_scifi*
- Numbers in parenthesis are expected counts (based on the data distribution of the two variables)
- **Hypothesis:** *plays\_chess* and *likes\_scifi* are independent

# Correlation Analysis (Categorical Data)

- **$\chi^2$  (chi-square) statistic:** tells you how much difference exists between **observed** counts *for two categorical variables* and the counts you would **expect** if *there were no relationship at all* between them.
  - Example:

	plays_chess (yes)	plays_chess (no)	Sum
likes_scifi (yes)	250 (90)	200 (360)	450
likes_scifi (no)	50 (210)	1000 (840)	1050
Sum (col.)	300	1200	<b>1500</b>

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

# Correlation Analysis (Categorical Data)

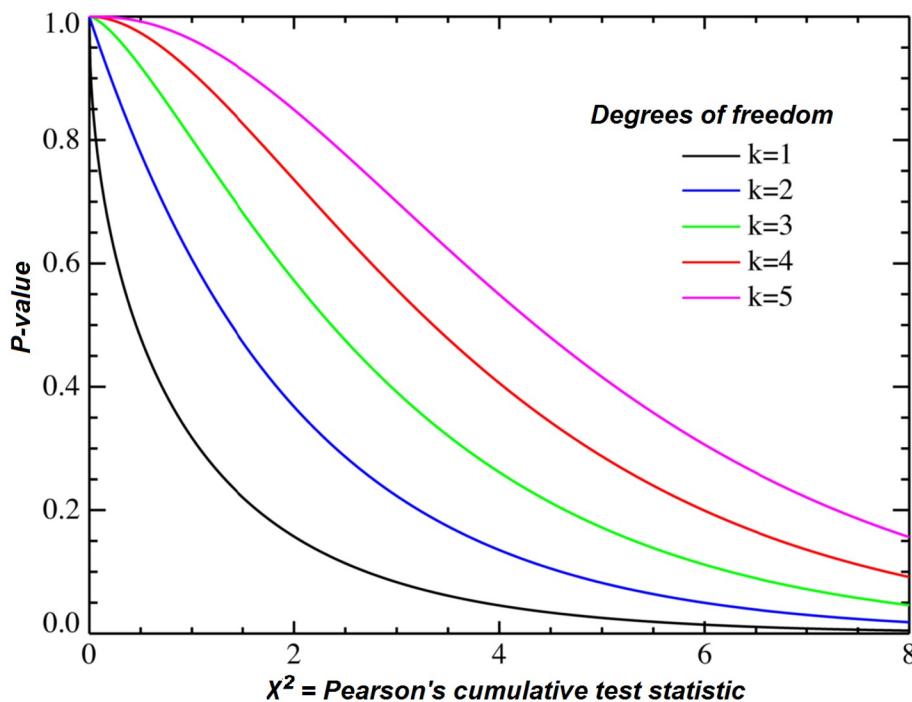
- **$\chi^2$  (chi-square) statistic:** tells you how much difference exists between **observed** counts *for two categorical variables* and the counts you would **expect** if *there were no relationship at all* between them.
  - Example:

	plays_chess (yes)	plays_chess (no)	Sum
likes_scifi (yes)	250 (90)	200 (360)	450
likes_scifi (no)	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	<b>1500</b>

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = \boxed{507.93}$$

# Correlation Analysis (Categorical Data)

- From  $\chi^2$  statistic to  $\chi^2$  test: used to determine whether there is a **statistically significant difference** between expected and observed frequencies in one or more variables.



Distribution of the  $\chi^2$  statistic

Degrees of freedom (df):  
 $(\#\text{rows} - 1)(\#\text{columns} - 1)$

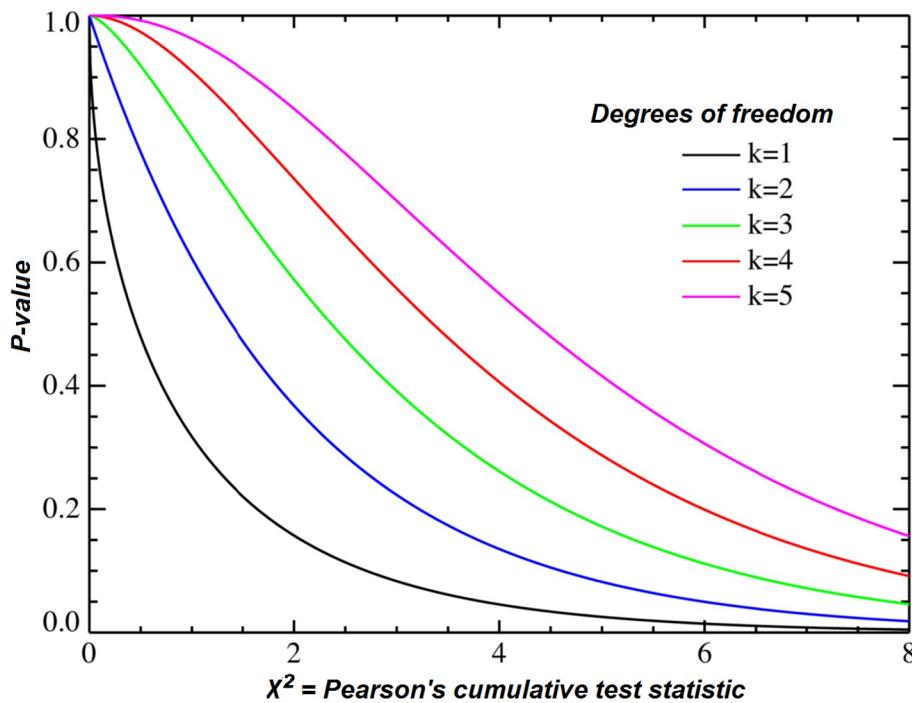
In the example,  $\text{df} = 1$

Note that the p-value is going to be very small!

If the p-value is **very small** ( $< 0.05$ ), reject the hypothesis!

# Correlation Analysis (Categorical Data)

- From  $\chi^2$  statistic to  $\chi^2$  test: used to determine whether there is a **statistically significant difference** between expected and observed frequencies in one or more variables.



Distribution of the  $\chi^2$  statistic

Degrees of freedom (df):  
 $(\#\text{rows} - 1)(\#\text{columns} - 1)$

In the example,  $\text{df} = 1$

Note that the p-value is going to be very small!

**Bottom line: likes\_scifi and plays\_chess are likely related!**

# $\chi^2$ test: Considerations

- The larger the  $\chi^2$  value, the more likely the variables are related
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

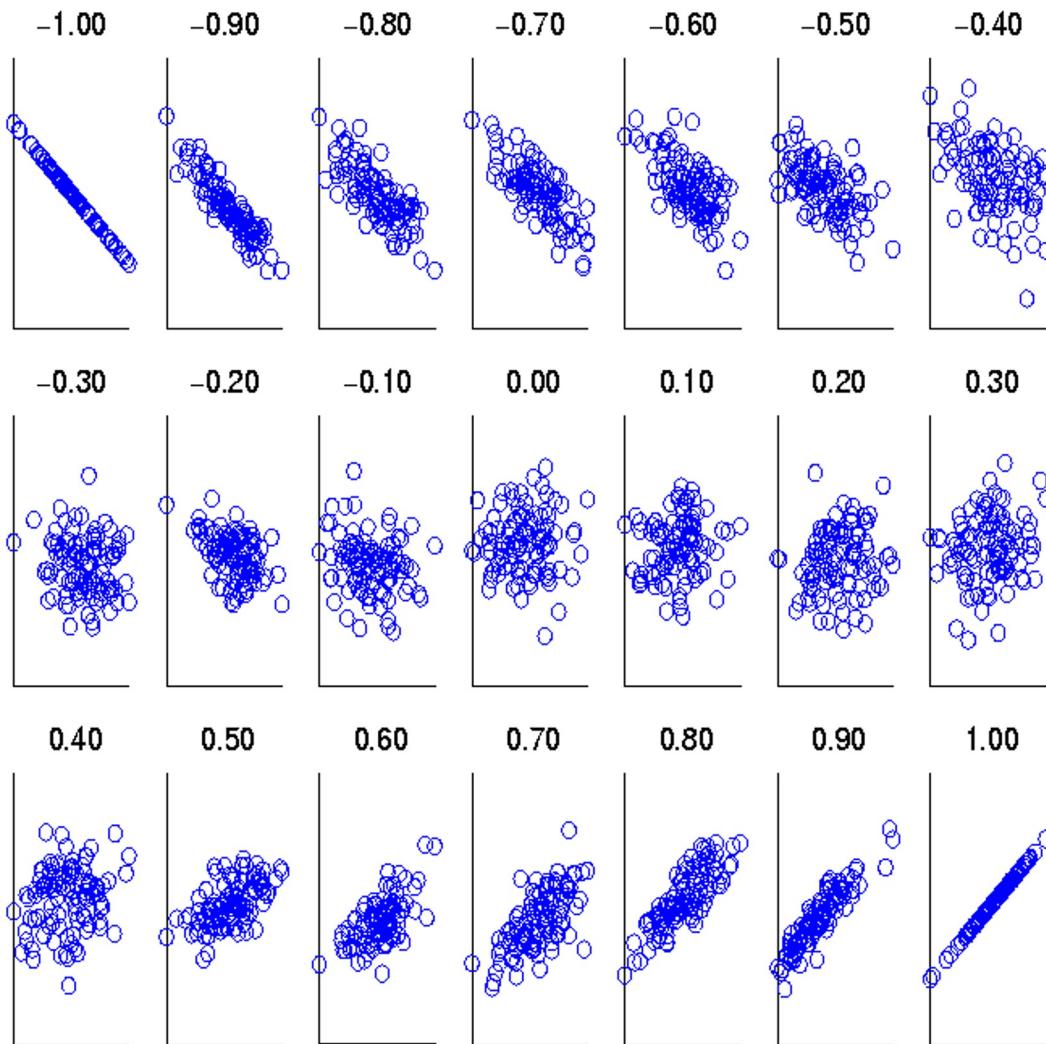
# Correlation Analysis (Continuous Data)

- Correlation coefficient (also called Pearson's correlation)

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n \bar{A} \bar{B}}{(n-1)\sigma_A \sigma_B}$$

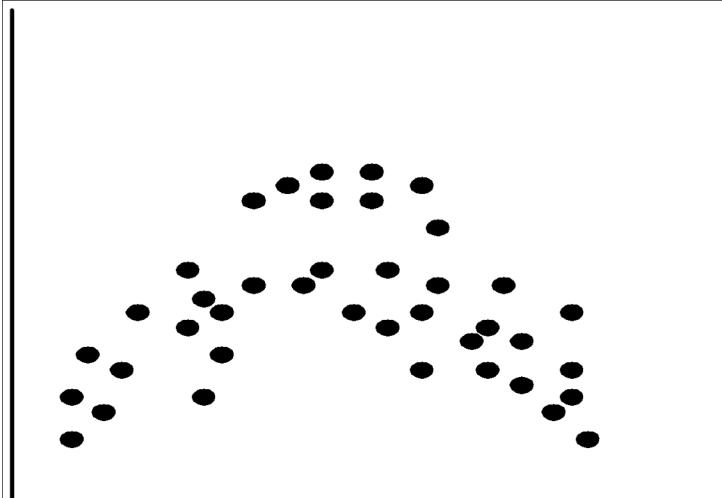
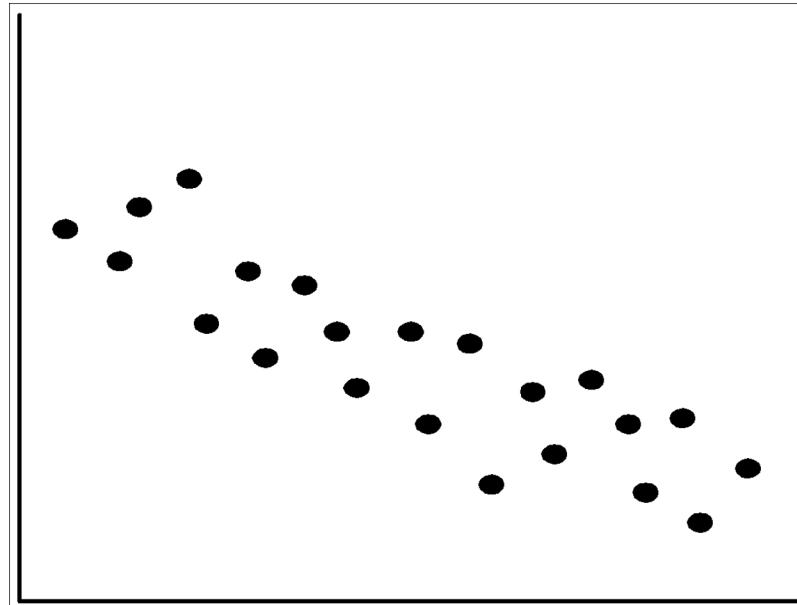
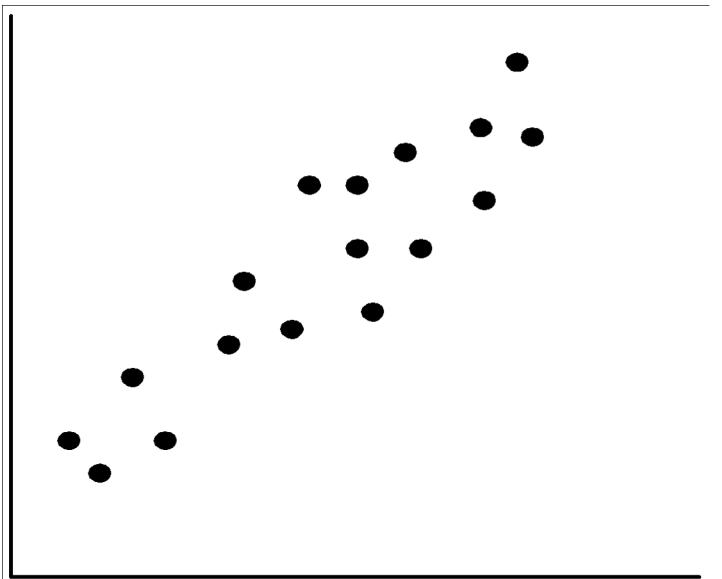
- where  $n$  is the number of tuples,  $\bar{A}$  and  $\bar{B}$  are the respective means of numerical variables  $A$  and  $B$ , and  $\sigma_A$  and  $\sigma_B$  are the respective standard deviations of  $A$  and  $B$ .
- If  $r_{A,B} > 0$ ,  $A$  and  $B$  are positively correlated ( $A$ 's values increase as  $B$ 's do).
  - The higher, the stronger correlation.
- $r_{A,B} = 0$ : independent;  $r_{AB} < 0$ : negatively correlated

# Visually Evaluating Correlation



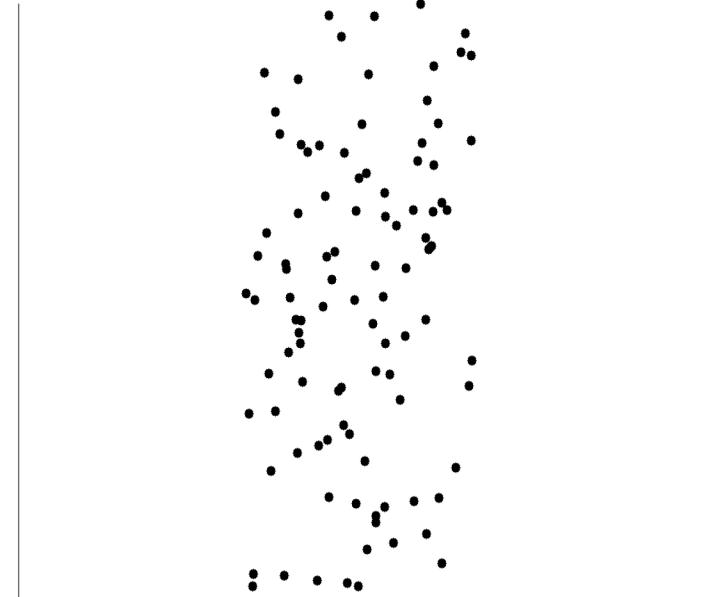
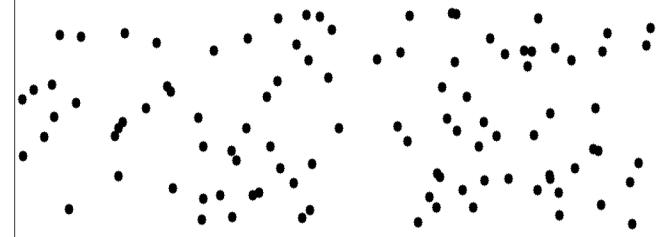
**Scatter plots showing correlations from  $-1$  to  $1$ .**

# Positively and Negatively Correlated Data



- The left half fragment is positively correlated
  - **Not always linear!**
- The right half is negative correlated

# Uncorrelated Data



# Correlation: Linear Relationship

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \bullet B'$$