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 & 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

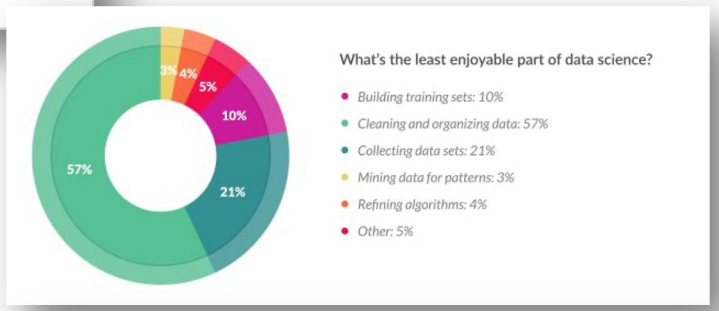
Opinions expressed by Forbes Contributors are their own.

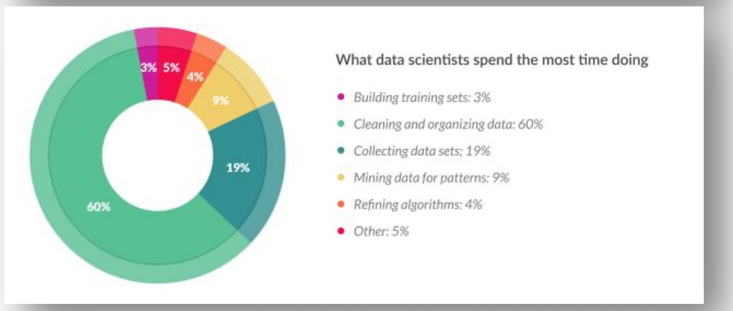
TWEET THIS

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 21st 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%**)

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

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
 BROOKLYYN
 BROOKOLYN
 BROOKYL
 BROOKYLN
 BROOKYLYN
 BROOOKKLYN
 BROOOKLN
 BROOOKLYN
```

Data Cleaning: Missing Data Imputation

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

ItemID					
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)

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

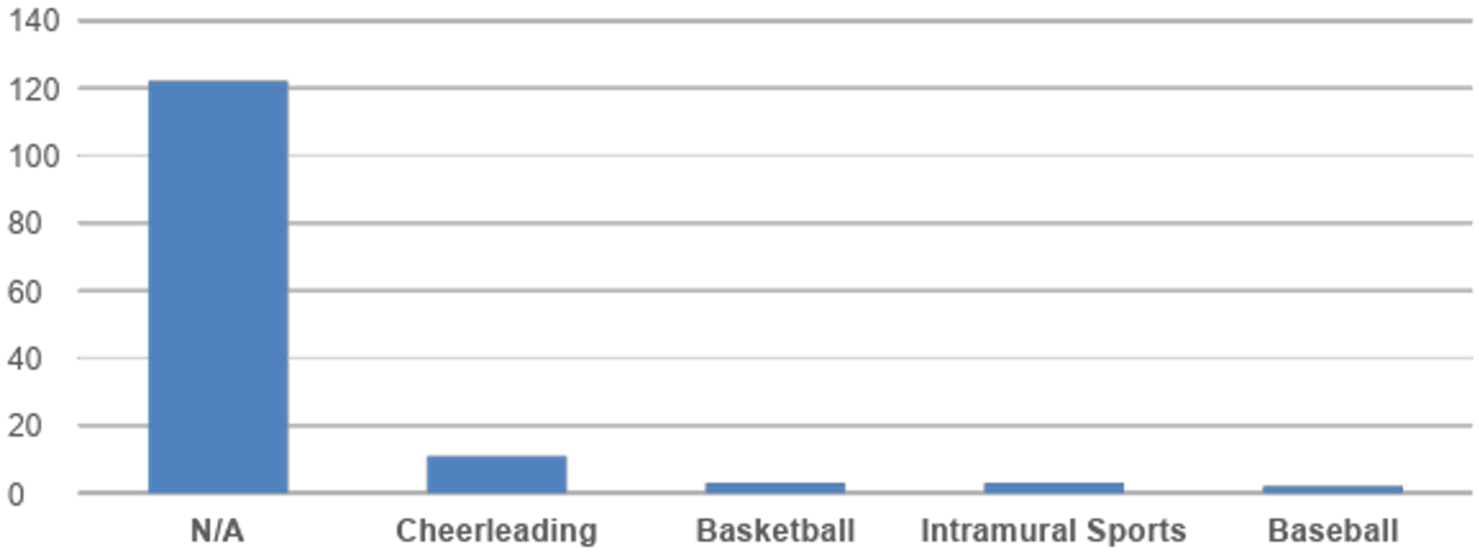
- Sometimes different text representations of NULL encode different semantics
  - unknown: there is a value, but I do not know it (e.g., unknown date-of-birth)
  - not applicable: there is no meaningful value (e.g., spouse for singles)
  - withheld: there is a value, but we are not authorized to see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

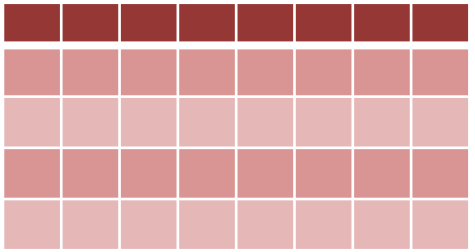
Outliers that are Data Quality Issues

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	5th Ave	10011	Manhattan	351-344-5671
Jane Smith	6th Ave	10011	Manhattan	351-244-4674
Smith, J.	Fifth Avenue	10011	NYC	+1 351.344.5671

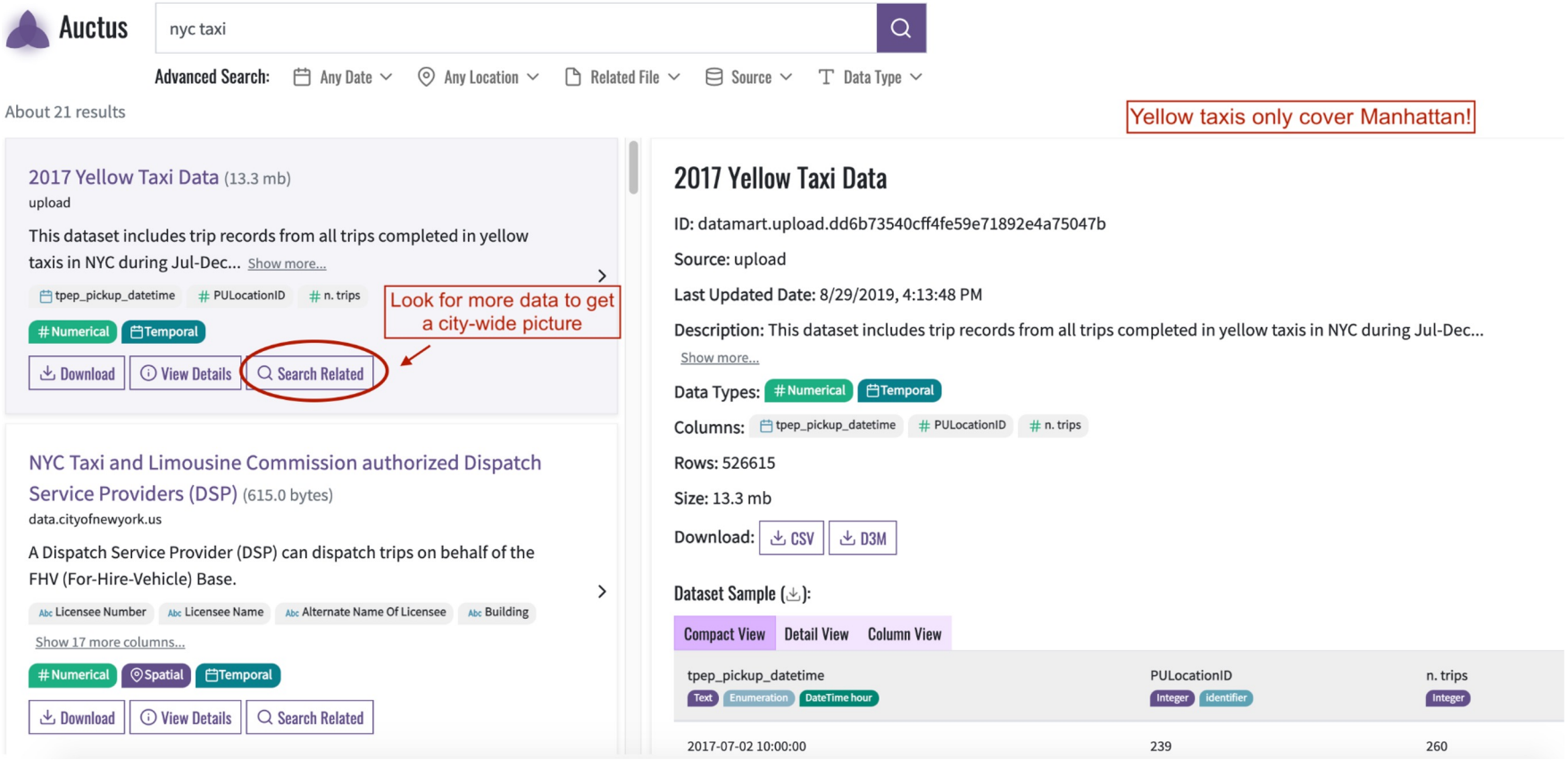
- **Standardize data**
- **Apply different similarity measures**
  - Similarity measures – Levenshtein, Soundex, Jaccard, etc.
  - Tokenize values
- **Different weights for attributes (apply domain knowledge)**
  - Two records with the same telephone number are more likely to be duplicates than records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

Major Tasks in Data Preprocessing

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

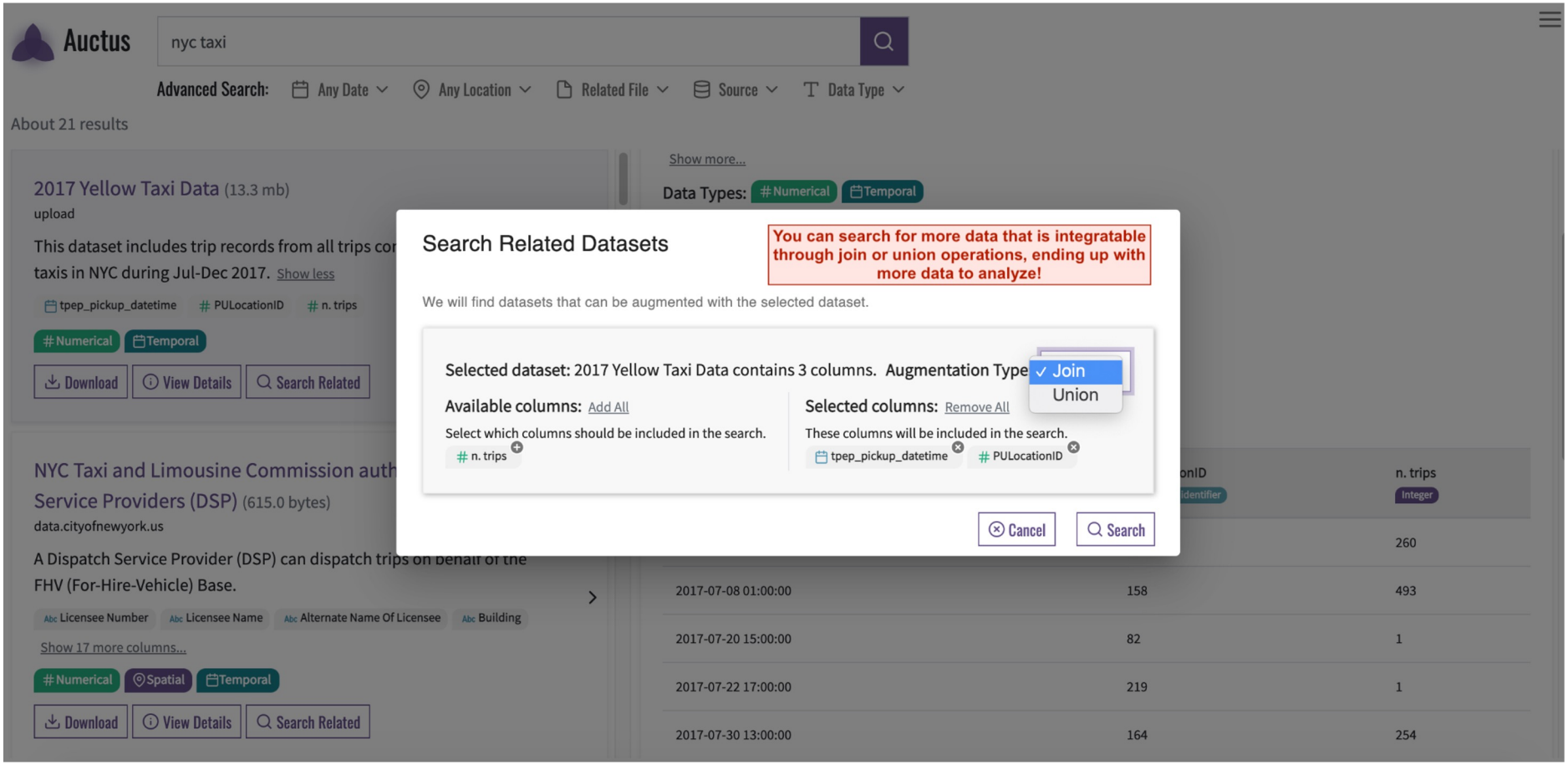
Data Integration

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



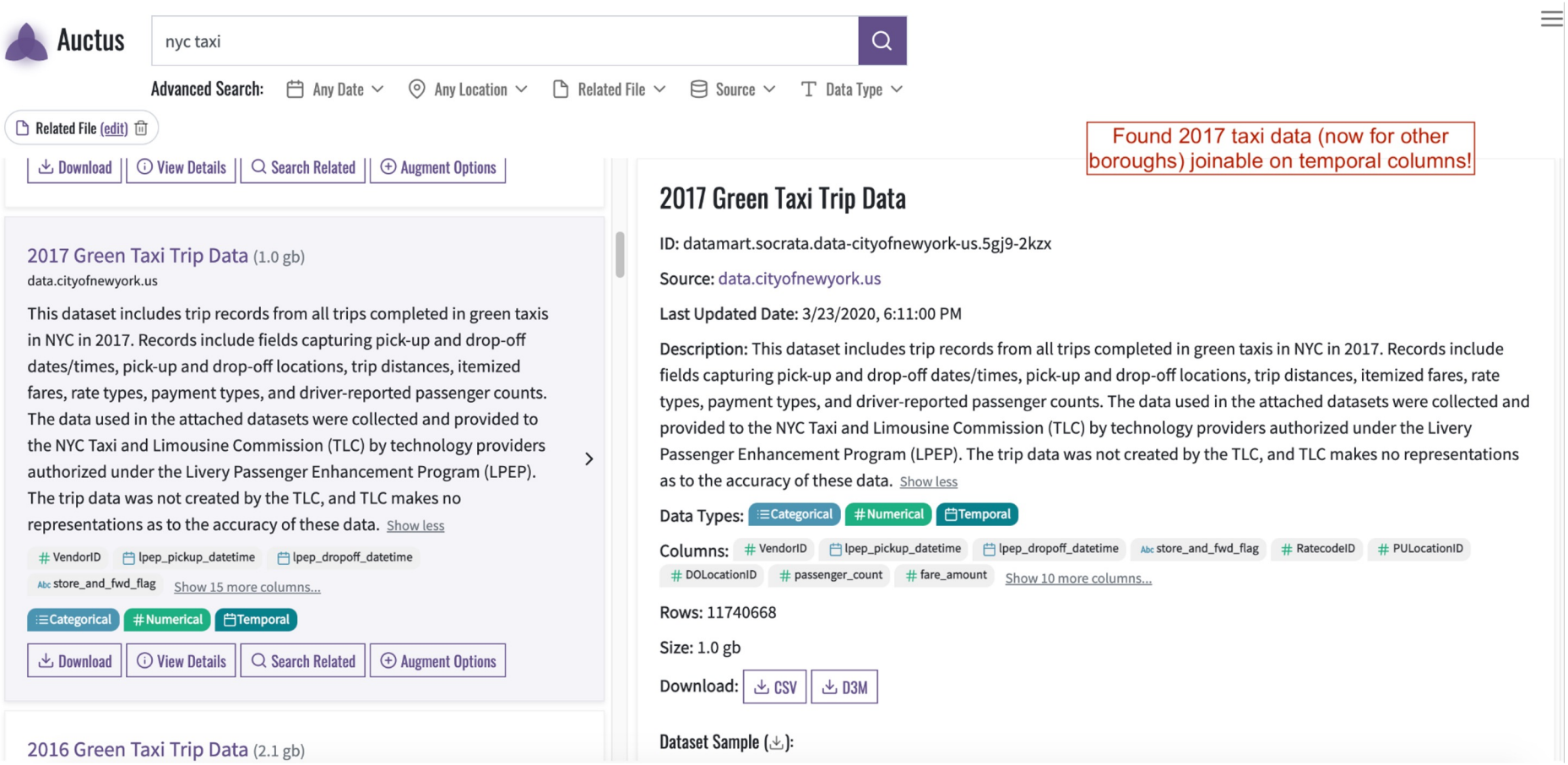
Data Integration

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



Data Integration

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



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)

- X^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	**1500**

  - 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)

- X^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	**1500**

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

Correlation Analysis (Categorical Data)

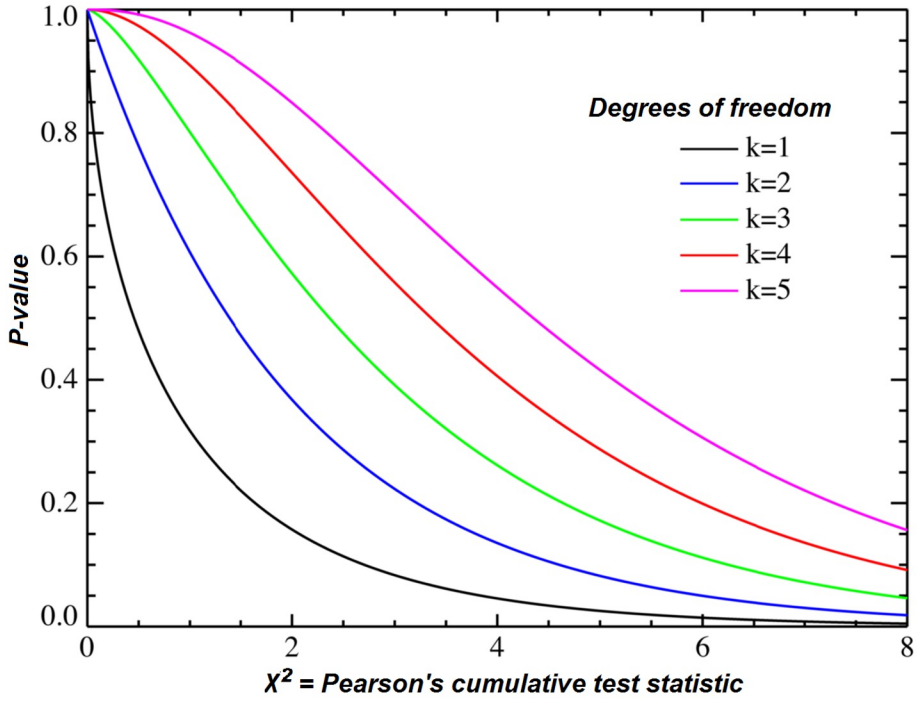
- X^2 (chi-square) statistic: tells you how much difference exists between observed counts *for two categorical variables* and the counts you would expect if *there were no relationship at all* between them.
  - Example:

	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	**1500**

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

Correlation Analysis (Categorical Data)

- From X^2 statistic to X^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 X^2 statistic

Degrees of freedom (df):
(#rows - 1)(#columns - 1)

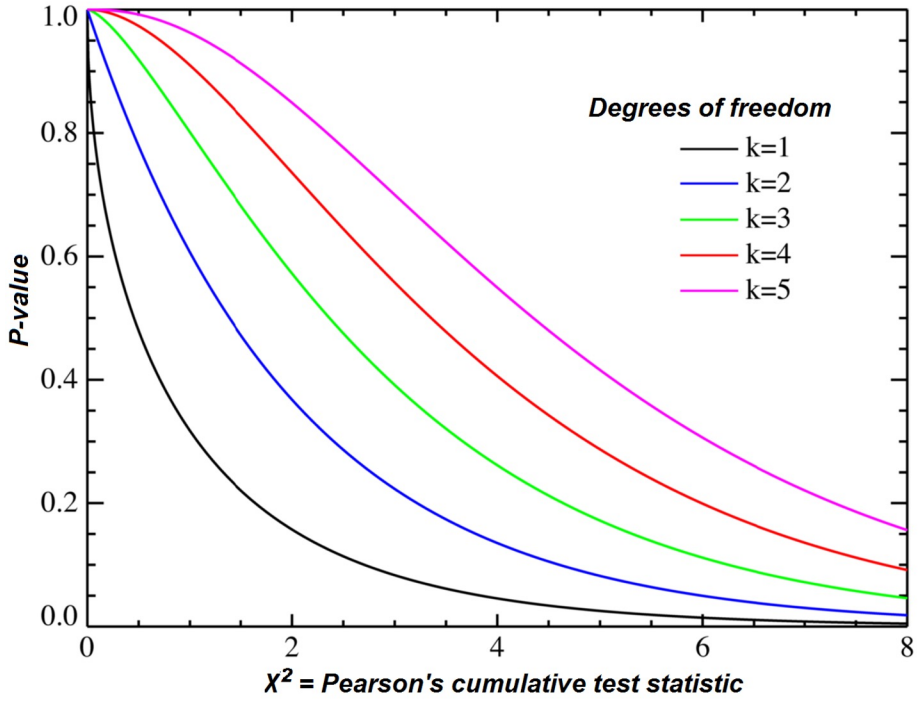
In the example, df = 1

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 X^2 statistic to X^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 X^2 statistic

Degrees of freedom (df):
(#rows - 1)(#columns - 1)

In the example, df = 1

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

Bottom line: likes_scifi and plays_chess are likely related!

X^2 test: Considerations

- The larger the X^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)

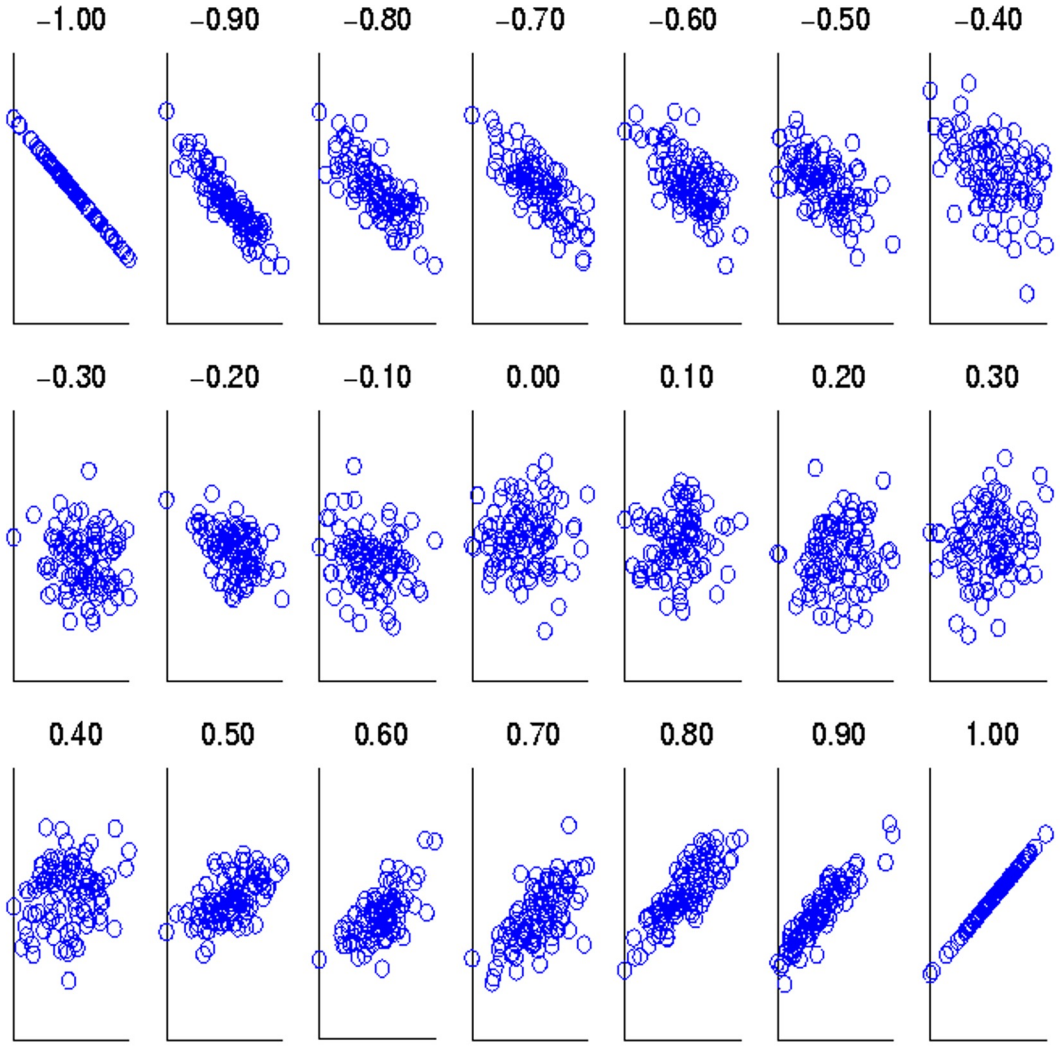
- Correlation coefficient (also called Pearson's correlation)

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

  - where *n* is the number of tuples, $\overline{A}$ and $\overline{B}$ are the respective means of numerical variables *A* and *B*, and σ_A and σ_B are the respective standard deviations of *A* and *B*.

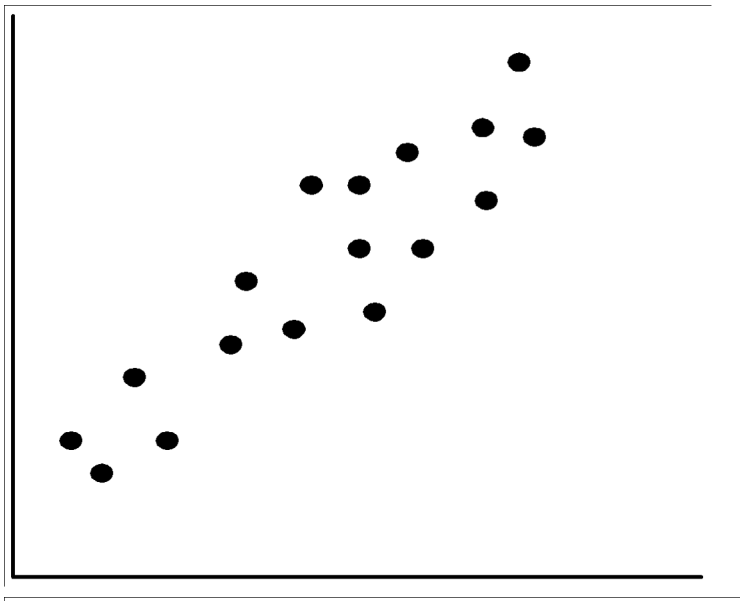
- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's do).
  - The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

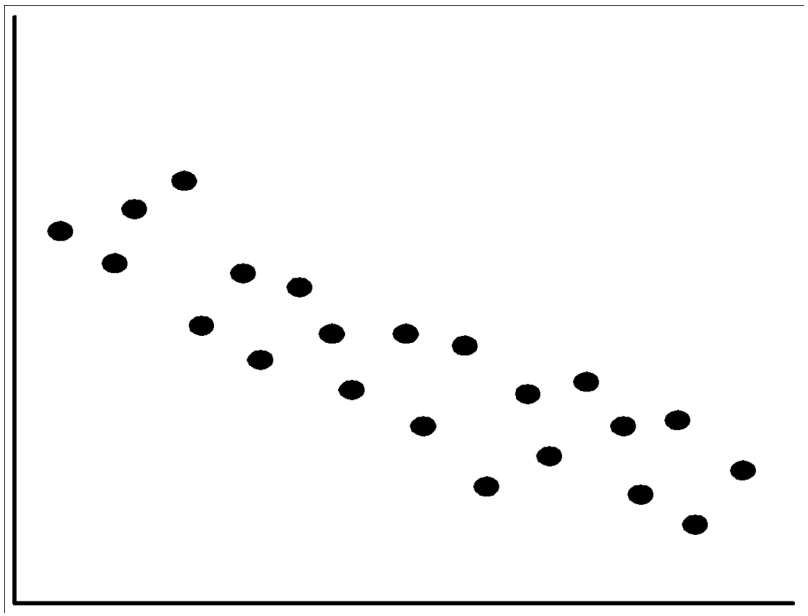
Visually Evaluating Correlation

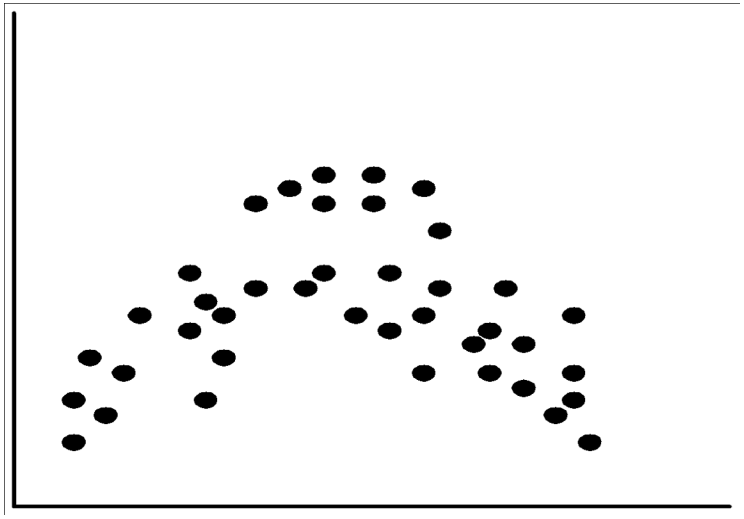


Scatter plots showing correlations from –1 to 1.

Positively and Negatively Correlated Data

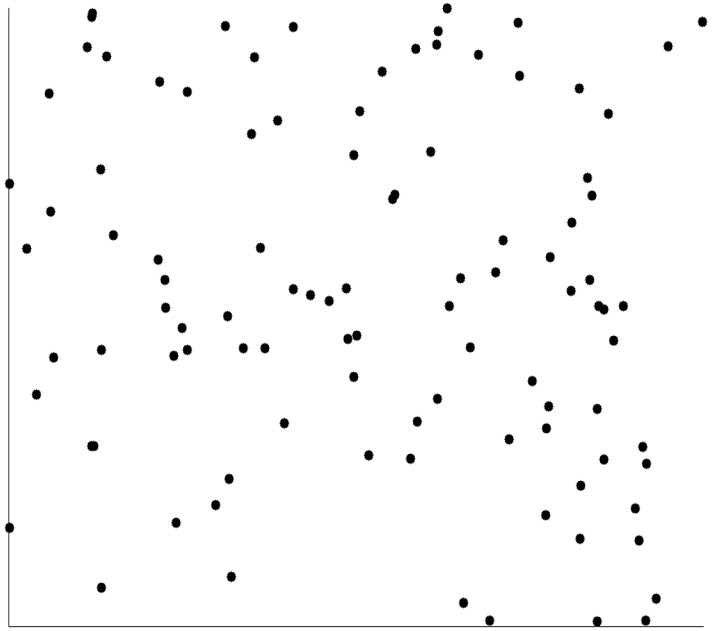


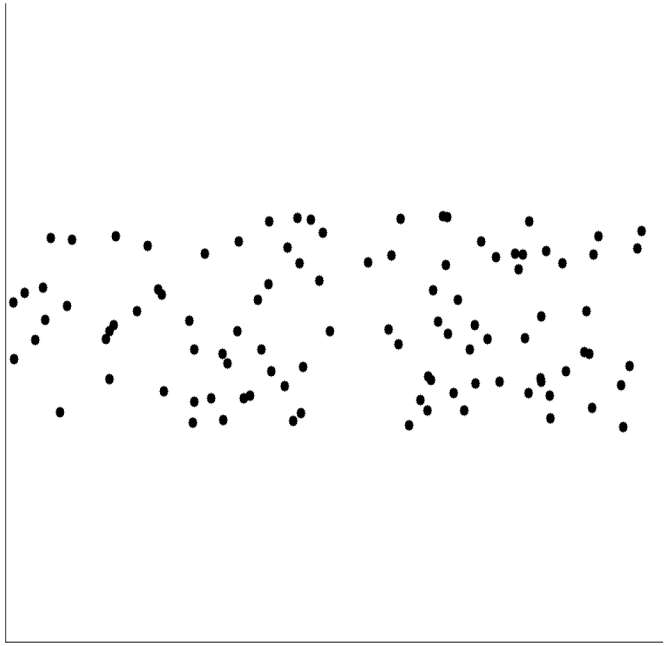


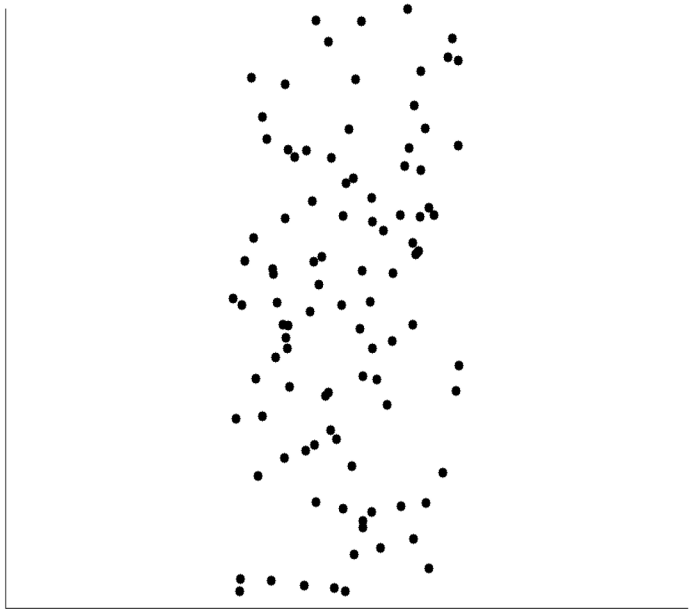


- The left half fragment is positively correlated
  - **Not always linear!**
- The right half is negative correlated

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

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$