The 80% of Data Mining

Yabebal Fantaye

Credit

- Many of the slides are from
  - Sanjay Ranka lecture: https://cise.ufl.edu/class/cis6930sp14ids/15.%20Data%20Preprocessing.pdf
  - Anita Wasilewska lecture: http://www3.cs.stonybrook.edu/~cse634/ch2preprocess.pdf
  - Hanspeter Pfister & Joe Blitzstein lectures http://cs109.github.io/2015/index.html

"In our experience, the tasks of exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of theultimate data."

T. Dasu and T. Johnson
Authors of Exploratory Data Mining and Data Cleaning

Data Set

- Attributes (describe objects)
  - Variable, field, characteristic, feature or observation
- Objects (have attributes)
  - Record, point, case, sample, entity or item
- Data Set
  - Collection of objects

Data Attributes/Features

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads

We may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules.

Type of Attributes

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
Nominal Categorical Qualitative	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
Interval	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

On the theory of scales and measurements [S. Stevens, 46]

Data Types

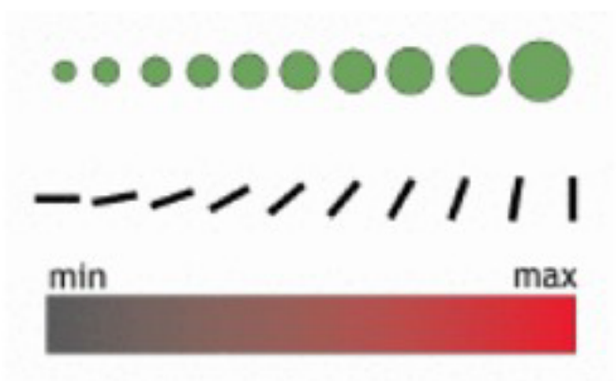
- N - Nominal (labels)
  - Operations: =, ≠
- O - Ordinal (ordered)
  - Operations: =, ≠, >, <
- Q - Interval (location of zero arbitrary)
  - Operations: =, ≠, >, <, +, − (distance)
- Q - Ratio (zero fixed)
  - Operations: =, ≠, >, <, +, −,x, ÷ (proportions)

Data Types

- Nominal (Categorical) (N)
  - Are = or ≠ to other values
  - Apples, Oranges, Bananas,...

- Ordinal (O)
  - Obey a < relationship
  - Small, medium, large

- Quantitative (Q)
  - Can do arithmetic on them
  - 10 inches, 23 inches, etc.



Data Types

- Q - Interval (location of zero arbitrary)
  - Dates: Jan 19; Location: (Lat, Long)
  - Like a geometric point. Cannot compare directly.
  - Only differences (i.e., intervals) can be compared
- Q - Ratio (zero fixed)
  - Measurements: Length, Mass, Temp, ...
  - Origin is meaningful, can measure ratios & proportions
  - Like a geometric vector, origin is meaningful

Data - Conceptual Model

- Data Model: Low-level description of the data
  - Set with operations, e.g., floats with +, -, /, *
- Conceptual Model: Mental construction
  - Includes semantics, supports reasoning

Data	Conceptual
1D floats	temperature
3D vector of floats	space

Data - Conceptual Model

- From data model...
  - 32.5, 54.0, -17.3, … (floats)
- using conceptual model...
  - Temperature
- to data type
  - Continuous to 4 significant figures (Q)
  - Hot, warm, cold (O)
  - Burned vs. Not burned (N)

Data sets

Data Formats

- Delimited values
  - Comma Separated Values (CSV)
  - Tab Separated Values (TSV)
- Markup languages
  - Hypertext Markup Language (HTML5 / XML)
  - JavaScript Object Notation (JSON)
  - Hierarchical Data Format (HDF5)
- Ad hoc formats
  - Graph edge lists, voting records, fixed width files, ...

CSV

Country	Indicator	Year	Value
Cameroon	Population (thousands)	2014	22818.63
Cameroon	Land area (thousands of km2)	2014	475.44
Cameroon	Population density (pop / km2)	2014	47.99
Cameroon	GDP based on PPP valuation (USD million)	2014	67224.93
Cameroon	GDP per Capita (PPP valuation, USD)	2014	2946.05
Cameroon	Annual real GDP growth (2006-2014 average)	2014	3.8
Algeria	Population (thousands)	2014	39928.95
Algeria	Land area (thousands of km2)	2014	2381.74
Algeria	Population density (pop / km2)	2014	16.76
Algeria	GDP based on PPP valuation (USD million)	2014	551720.2
Algeria	GDP per Capita (PPP valuation, USD)	2014	13817.55
Algeria	Annual real GDP growth (2006-2014 average)	2014	2.8
Angola	Population (thousands)	2014	22137.26
Angola	Land area (thousands of km2)	2014	1246.7
Angola	Population density (pop / km2)	2014	17.76
Angola	GDP based on PPP valuation (USD million)	2014	175540.07
Angola	GDP per Capita (PPP valuation, USD)	2014	7929.62
Angola	Annual real GDP growth (2006-2014 average)	2014	7
Egypt	Population (thousands)	2014	83386.74
Egypt	Land area (thousands of km2)	2014	1001.45
Egypt	Population density (pop / km2)	2014	83.27
Egypt	GDP based on PPP valuation (USD million)	2014	945387.82
Egypt	GDP per Capita (PPP valuation, USD)	2014	11337.39
Egypt	Annual real GDP growth (2006-2014 average)	2014	4.3
Chad	Population (thousands)	2014	13211.15
Chad	Land area (thousands of km2)	2014	1284
Chad	Population density (pop / km2)	2014	10.29
Chad	GDP based on PPP valuation (USD million)	2014	29850.78
Chad	GDP per Capita (PPP valuation, USD)	2014	2259.51
Chad	Annual real GDP growth (2006-2014 average)	2014	4.7
Congo	Population (thousands)	2014	4558.59
Congo	Land area (thousands of km2)	2014	342
Congo	Population density (pop / km2)	2014	13.33
Congo	GDP based on PPP valuation (USD million)	2014	28090.01
Congo	GDP per Capita (PPP valuation, USD)	2014	6161.99
Congo	Annual real GDP growth (2006-2014 average)	2014	4.8

https://stats.oecd.org/Index.aspx?DataSetCode=RS_AFR#

XML and JSON

- XML
  - Self-describing

```
<order>
   <crust>original</crust>
   <toppings>
      <topping>cheese</topping>
      <topping>pepperoni</topping>
      <topping>garlic</topping>
   </toppings>
   <status>cooking</status>
<order>
```

- JSON
  - Dictionary [of dictionaries]

```
{
   “crust”: “original”,
   “toppings”: [ “cheese”,
                 ”pepperoni”,
                 ”garlic” ],
   “status”: “cooking”,
   “customer”: {“name”: ”Brian”,
                “phone”: 0824540028
                }
}
```

AEO Statistical annex, 2015 : *Table 1. Basic Indicators, 2014*

Customise | Export | Draw chart | My Queries

Indicator		ANN1_POP	ANN1_AREA	ANN1_POPDENS	ANN1_GDPPPP
Year					2014
Country					
Oil producers	Cameroon	22 818.63		47.99	67 224.93
	Algeria	39 928.95		16.76	551 720.2
	Angola	22 137.26		17.76	175 540.07
	Egypt	83 386.74	1 001.45	83.27	945 387.82
	Chad	13 211.15	1 284	10.29	29 850.78
	Congo	4 558.59	342	13.33	28 090.01
	Nigeria	178 516.9	923.77	193.25	1 057 830.61
	Equatorial Guinea	778.06	28.05	27.74	25 331.33
	Gabon	1 711.29	267.67	6.39	34 280.07
	Libyan Arab Jamahiriya	6 253.45	1 759.54	3.55	103 266.72
	South Sudan	11 738.72	644.33	18.22	23 306.43
Non-Oil producers	Democratic Republic of the Congo	69 360.12	2 344.86	29.58	55 731.34
	Côte d'Ivoire	20 804.77	322.46	64.52	71 951.62
	Benin	10 599.51	114.76	92.36	19 846.76
	Ghana	26 442.18	238.54	110.85	109 391.52
	Botswana	2 038.59	581.73	3.5	33 622.48
	Ethiopia	96 506.03	1 104.3	87.39	139 434.04
	Burkina Faso	17 419.62	274.22	63.52	30 080.58
	Burundi	10 482.75	27.83	376.67	8 395.84
	Cape Verde	503.64	4.03	124.97	3 286.12
	Eritrea	6 536.18	117.6	55.58	7 855.11
	Central African Republic	4 709.2	622.98	7.56	2 860.64
	Kenya	45 545.98	580.37	78.48	134 710.65
	Comoros	752.44	1.86	404.32	1 210.96
	Mauritius	1 249.15	2.04	612.33	23 422.22
	Morocco	33 492.91	446.55	75	254 362.12
	Mozambique	26 472.08	799.38	33.12	29 757.17

Semantics

Data extracted on 21 May 2017 22:49 UTC (GMT) from OECD.Stat

AEO Statistical annex, 2015 : *Table 1. Basic Indicators, 2014*

Customise | Export | Draw chart | My Queries

Indicator		ANN1_POP	ANN1_AREA	ANN1_POPDENS	ANN1_GDPPPP
Year					2014
Country					
Oil producers	Cameroon	22 818.63	475.44	47.99	67 224.93
	Algeria	39 928.95	2 381.74	16.76	551 720.2
	Angola	22 137.26	1 246.7	17.76	175 540.07
	Egypt	83 386.74	1 001.45	83.27	945 387.82
	Chad	13 211.15	1 284	10.29	29 850.78
	Congo	4 558.59		13.33	28 090.01
	Nigeria	178 516.9		193.25	1 057 830.61
	Equatorial Guinea	778.06		27.74	25 331.33
	Gabon	1 711.29	267.67	6.39	34 280.07
	Libyan Arab Jamahiriya	6 253.45	1 759.54	3.55	103 266.72
	South Sudan	11 738.72	644.33	18.22	23 306.43
Non-Oil producers	Democratic Republic of the Congo	69 360.12	2 344.86	29.58	55 731.34
	Côte d'Ivoire	20 804.77	322.46	64.52	71 951.62
	Benin	10 599.51	114.76	92.36	19 846.76
	Ghana	26 442.18	238.54	110.85	109 391.52
	Botswana	2 038.59	581.73	3.5	33 622.48
	Ethiopia	96 506.03	1 104.3	87.39	139 434.04
	Burkina Faso	17 419.62	274.22	63.52	30 080.58
	Burundi	10 482.75	27.83	376.67	8 395.84
	Cape Verde	503.64	4.03	124.97	3 286.12
	Eritrea	6 536.18	117.6	55.58	7 855.11
	Central African Republic	4 709.2	622.98	7.56	2 860.64
	Kenya	45 545.98	580.37	78.48	134 710.65
	Comoros	752.44	1.86	404.32	1 210.96
	Mauritius	1 249.15	2.04	612.33	23 422.22
	Morocco	33 492.91	446.55	75	254 362.12
	Mozambique	26 472.09	799.38	33.12	29 757.17

Item

Data extracted on 21 May 2017 22:49 UTC (GMT) from OECD.Stat

AEO Statistical annex, 2015 : *Table 1. Basic Indicators, 2014*

Customise | Export | Draw chart | My Queries

Indicator		ANN1_POP	ANN1_AREA	ANN1_POPDENS	ANN1_GDPPPP
Year					2014
Country					
Oil producers	Cameroon	22 818.63	475.44	47.99	67 224.93
	Algeria	39 928.95	2 381.74	16.76	551 720.2
	Angola	22 137.26	1 246.7	17.76	175 540.07
	Egypt	83 386.74	1 001.45	83.27	945 387.82
	Chad	13 211.15	1 284	10.29	29 850.78
	Congo	4 558.59	342	13.33	28 090.01
	Nigeria	178 516.9		193.25	1 057 830.61
	Equatorial Guinea	778.06		27.74	25 331.33
	Gabon	1 711.29		6.39	34 280.07
	Libyan Arab Jamahiriya	6 253.45		3.55	103 266.72
	South Sudan	11 738.72		18.22	23 306.43
Non-Oil producers	Democratic Republic of the Congo	69 360.12		29.58	55 731.34
	Côte d'Ivoire	20 804.77		64.52	71 951.62
	Benin	10 599.51	114.76	92.36	19 846.76
	Ghana	26 442.18	238.54	110.85	109 391.52
	Botswana	2 038.59	581.73	3.5	33 622.48
	Ethiopia	96 506.03	1 104.3	87.39	139 434.04
	Burkina Faso	17 419.62	274.22	63.52	30 080.58
	Burundi	10 482.75	27.83	376.67	8 395.84
	Cape Verde	503.64	4.03	124.97	3 286.12
	Eritrea	6 536.18	117.6	55.58	7 855.11
	Central African Republic	4 709.2	622.98	7.56	2 860.64
	Kenya	45 545.98	580.37	78.48	134 710.65
	Comoros	752.44	1.86	404.32	1 210.96
	Mauritius	1 249.15	2.04	612.33	23 422.22
	Morocco	33 492.91	446.55	75	254 362.12
	Mozambique	26 472.08	799.38	33.12	29 757.17

Attribute / Features

Data extracted on 21 May 2017 22:49 UTC (GMT) from OECD.Stat

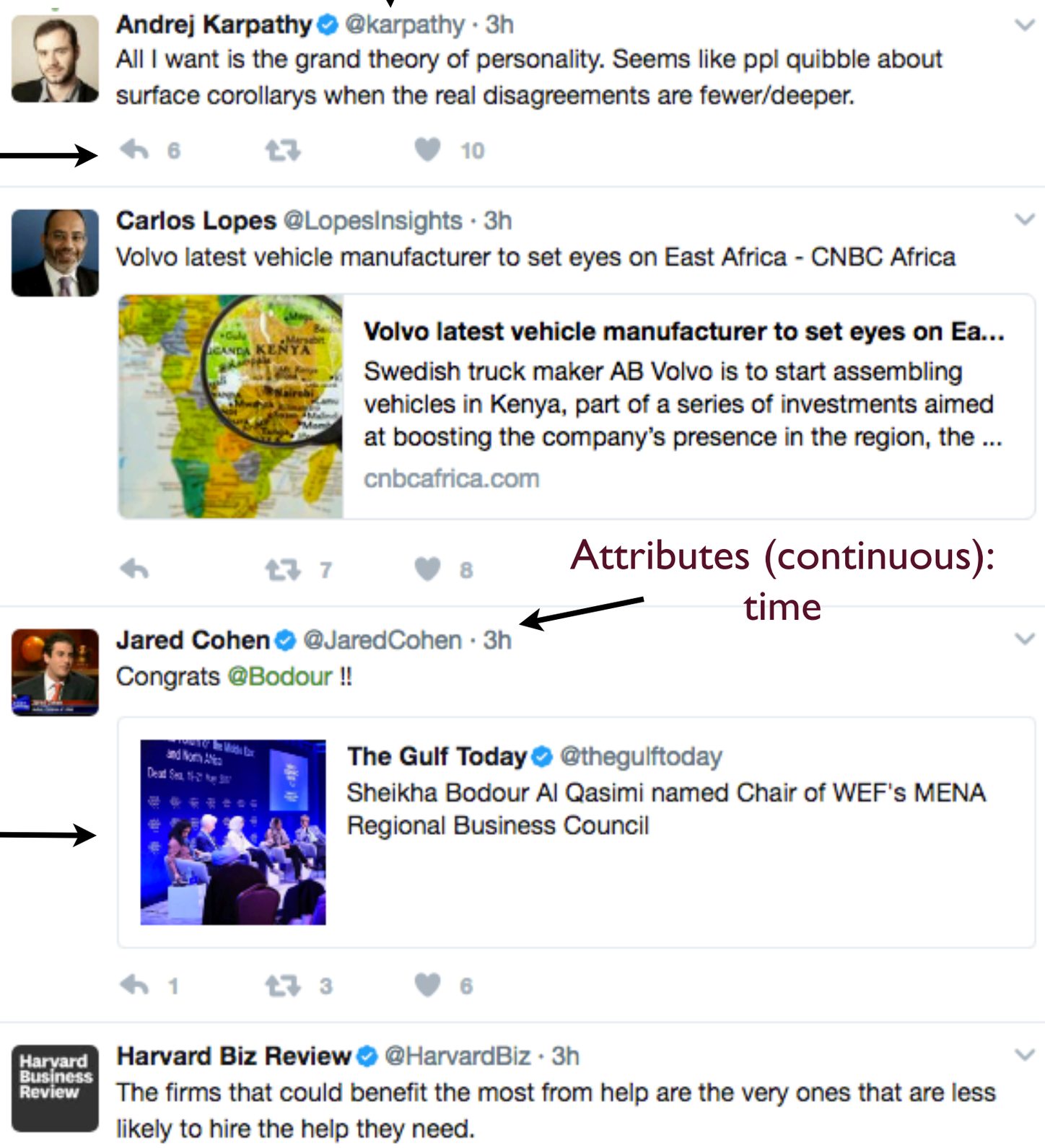
Data Set

Attributes (discrete):
Likes, Retweets, Replies

Attribute (Nominal):
Name, Twitter ID

Object:
Tweet, Media

Attributes (continuous):
time

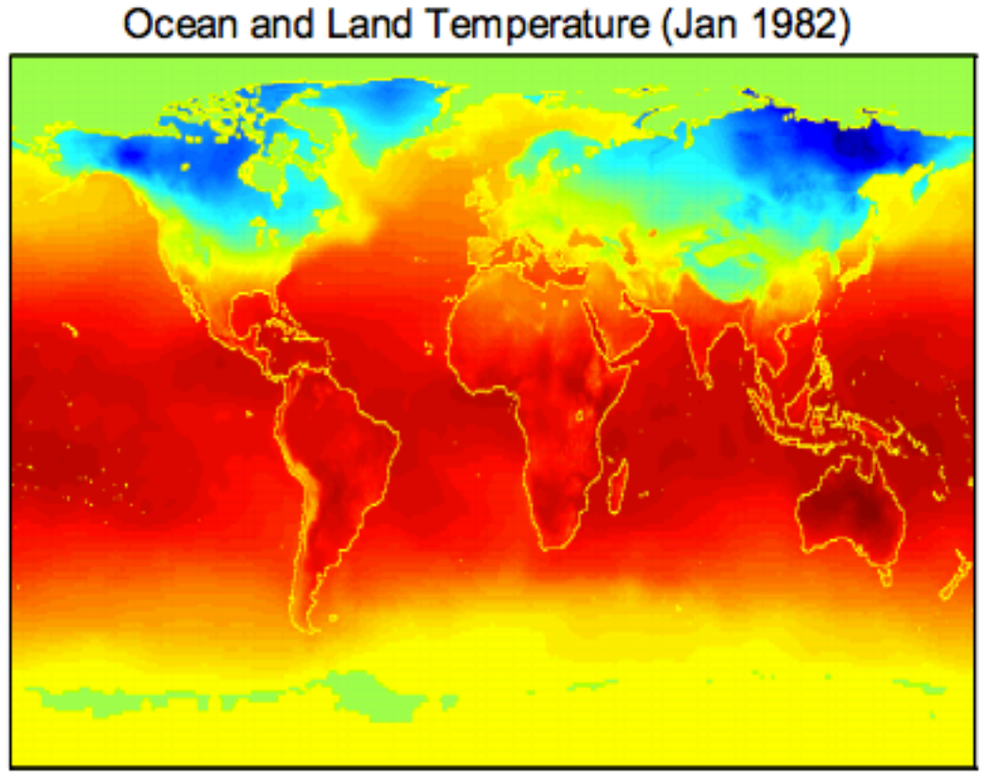


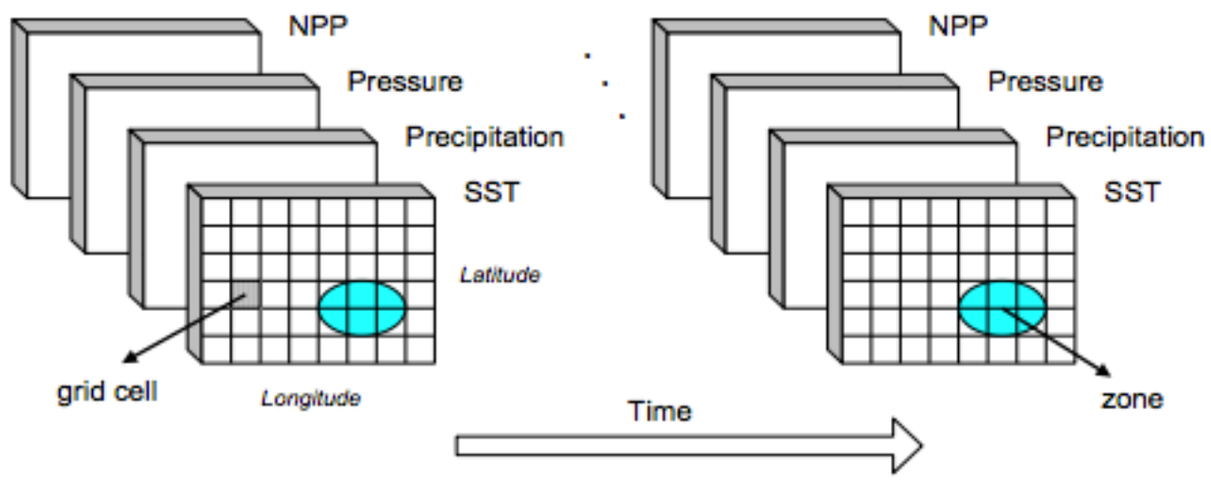
Twitter JSON data example


```
{
  "contributors": null,
  "truncated": false,
  "text": "@MGen_Shemsu There is no
Oromo,Amhara,Gurage, kembata,Tigre you name
it ..right. there is The people of Ethiopia right.I don't
buy this bud.",
  "is_quote_status": false,
  "in_reply_to_status_id": 791811322585444352,
  "id": 791812444918284288,
  "favorite_count": 0,
  "entities": {
    "symbols": [

    ],
    "user_mentions": [
      {
        "id": 334095331,
        "indices": [
          0,
          12
        ],
        "id_str": "334095331",
        "screen_name": "MGen_Shemsu",
        "name": "\u121c\u1300\u122d
\u1300\u1228\u1293\u120d \u1238\u121d\u1231
\u1262\u1228\u12f3 PhD"
      }
    ],
    "hashtags": [

    ],
    "urls": [

    ]
  },
  "retweeted": false,
  "coordinates": null,
  "source": "<a href=\"http://twitter.com/download/
iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "in_reply_to_screen_name": "MGen_Shemsu",
  "in_reply_to_user_id": 334095331,
  "retweet_count": 0,
  "id_str": "791812444918284288",
  "favorited": false,
  "user": {
    "follow_request_sent": false,
    "has_extended_profile": false,
    "profile_use_background_image": true,
    "default_profile_image": false,
    "id": 2311110587,
    "profile_background_image_url_https": "https://
abs.twimg.com/images/themes/theme1/bg.png",
    "verified": false,
    "translator_type": "none",
    "profile_text_color": "333333",
    "profile_image_url_https": "https://pbs.twimg.com/
profile_images/565152812511014912/
VtHOvMSw_normal.jpeg",
    "profile_sidebar_fill_color": "DDEEF6",
    "entities": {
      "description": {
        "urls": [

        ]
      }
    },
    "followers_count": 69,
    "profile_sidebar_border_color": "C0DEED",
    "id_str": "2311110587",
    "profile_background_color": "C0DEED",
    "listed_count": 2,
    "is_translation_enabled": false,
    "utc_offset": null,
    "statuses_count": 3183,
    "description": "",
    "friends_count": 252,
    "location": "",
    "profile_link_color": "0084B4",
    "profile_image_url": "http://pbs.twimg.com/
profile_images/565152812511014912/
VtHOvMSw_normal.jpeg",
    "following": false,
    "geo_enabled": false,
    "profile_background_image_url": "http://
abs.twimg.com/images/themes/theme1/bg.png",
    "screen_name": "d_jote",
    "lang": "en",
    "profile_background_tile": false,
    "favourites_count": 321,
    "name": "dereje jote",
    "notifications": false,
    "url": null,
    "created_at": "Wed Jan 29 03:55:28 +0000 2014",
    "contributors_enabled": false,
    "time_zone": null,
    "protected": false,
    "default_profile": true,
    "is_translator": false
  },
  "geo": null,
  "in_reply_to_user_id_str": "334095331",
  "lang": "en",
  "created_at": "Fri Oct 28 01:22:52 +0000 2016",
  "in_reply_to_status_id_str":
```

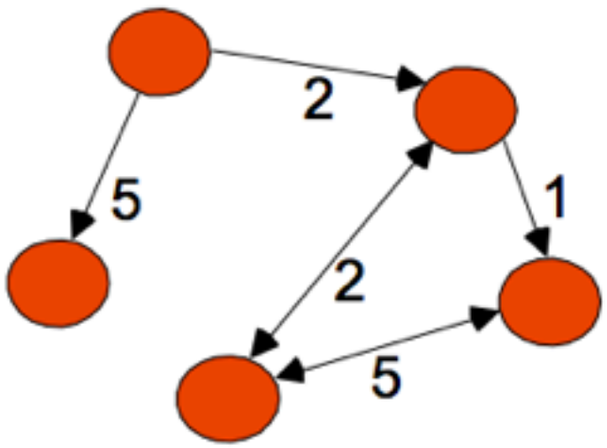
Ordered data

- Maps from satellite observations.
- Global snapshots of values for a number of variables on land surfaces and water surfaces
- Monthly over a range of 10 to 50 years





Graph data




```
<tr>
<td>
<b>Publications:<br>
          </b>
<a href="books.htm">Books</a><br>
     
<a href="journal.htm">Journals</a><br>
     
 <a href="refconf.htm">Conferences</a><br>
         
<a href="workshop.htm">Workshops</a>
</td>
</tr>
```

Publications:

Books
Journals
Conferences
Workshops

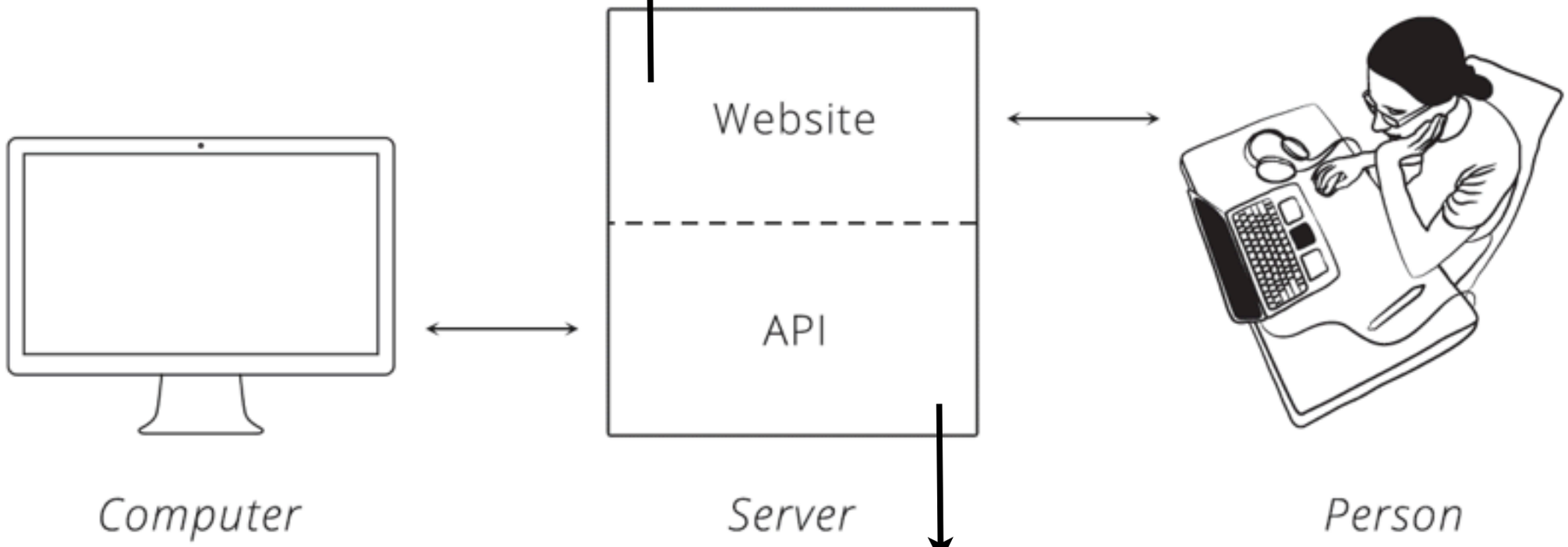
Data Sources

Data Sources

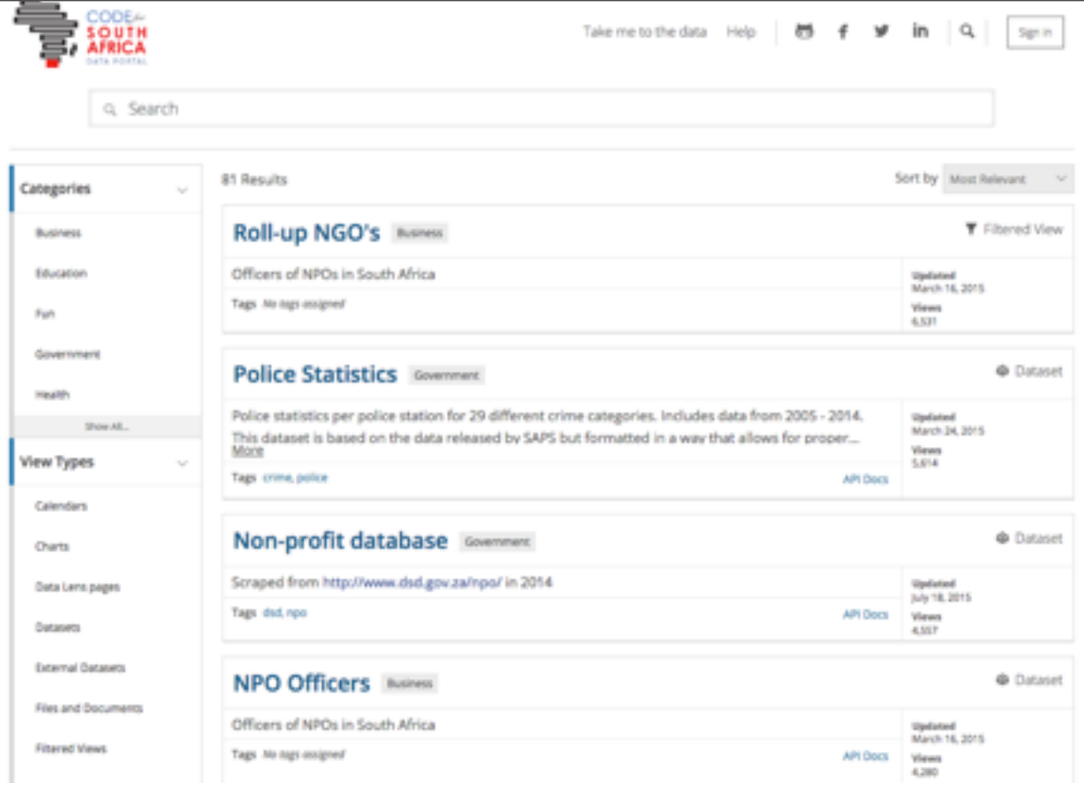
- Bulk downloads
  - Wikipedia, IMDB, Million Song Database, etc.
  - See list of data web sites on the Resources page
- API access
  - Twitter, Landsat, Facebook, Google, ...
- Web scraping

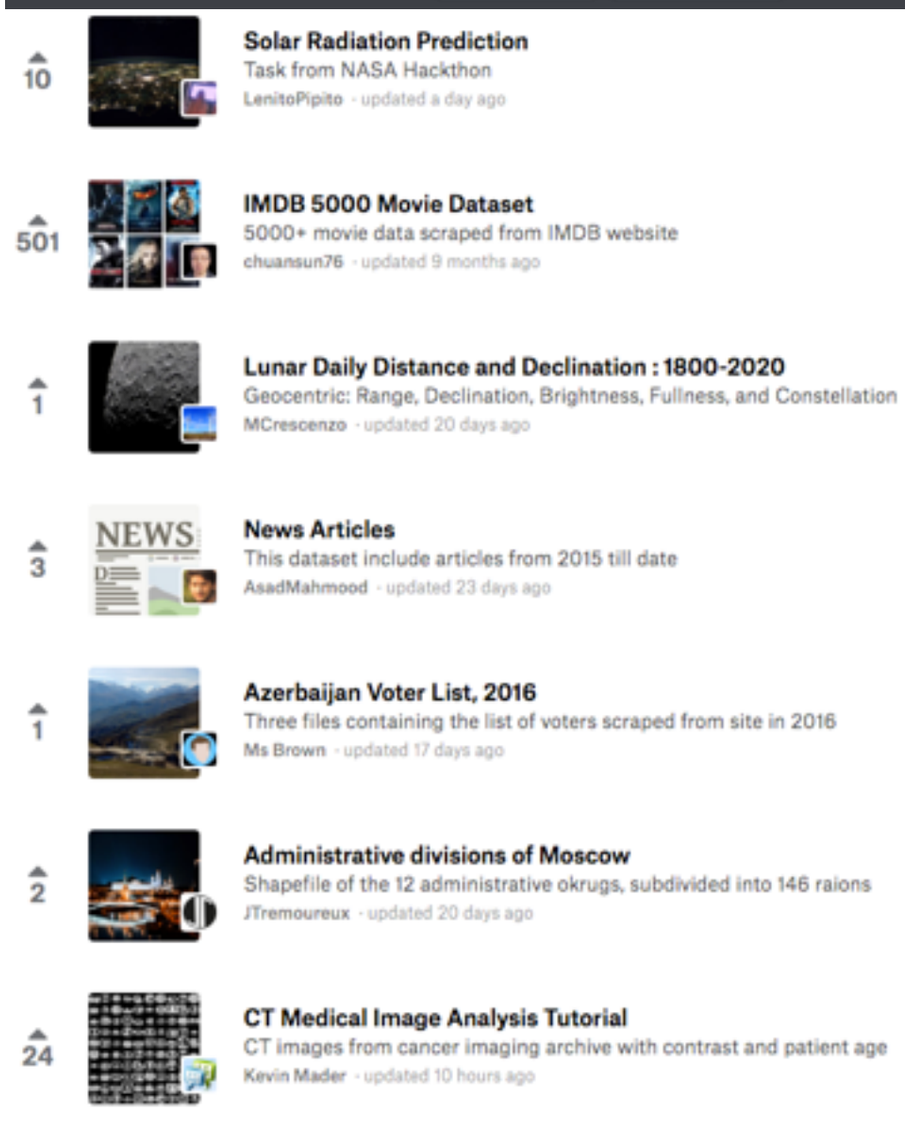
APIs

Optimal for humans
difficult for computers



digestible for computers





https://data.code4sa.org/browse

https://www.kaggle.com/datasets

List of lists:
https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/
http://www.sciencemag.org/site/feature/data/compsci/machine_learning.xhtml

https://www.google.com/publicdata/directory

Google Search public data

Public Data

Datasets

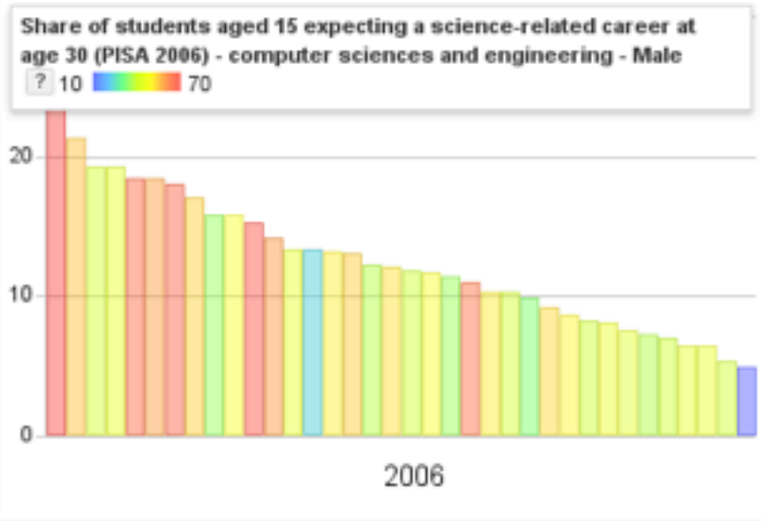
Metrics

Any data provider (106)

Eurostat (9)
Statistics Iceland (6)
U.S. Census Bureau (5)
Central Statistics Office, Ireland (4)
Data.gov.uk (4)

My Datasets

Share of students aged 15 expecting a science-related career at age 30 (PISA 2006) - computer sciences and engineering - Female ? Countries ?



Where will the next generation of software engineers come from?

In 2006 the OECD asked 15-year-olds whether they aspired to work in computer science or engineering by age 30. The graph ranks the results by country for female students, with the colors showing the corresponding values for males. The dataset shows that the next generation of software engineers may be coming from Eastern and Southern Europe; Estonian students, in particular, are the most open to a technical career.

Explore the data

Dataset: OECD Factbook 2013
Source: OECD Factbook 2013

World Development Indicators
World Bank
This dataset contains the World Development Indicators (WDI).

Global Competitiveness Report
World Economic Forum
Global Competitiveness Report

International Monetary Fund, World Economic Outlook
IMF, October 2014 WEO
International Monetary Fund, World Economic Outlook

WTO, International trade
WTO
WTO, International trade

Popular APIs

Google Cloud APIs
Compute Engine API
BigQuery API
Cloud Storage Service
Cloud Datastore API
Cloud Deployment Manager API
Cloud DNS API
More

Google Cloud Machine Learning
Vision API
Natural Language API
Speech API
Translation API
Machine Learning Engine API

Google Maps APIs
Google Maps Android API
Google Maps SDK for iOS
Google Maps JavaScript API
Google Places API for Android
Google Places API for iOS
Google Maps Roads API
More

Google Apps APIs
Drive API
Calendar API
Gmail API
Sheets API
Google Apps Marketplace SDK
Admin SDK
More

Mobile APIs
Google Cloud Messaging
Google Play Game Services
Google Play Developer API
Google Places API for Android

Social APIs
Google+ API
Blogger API
Google+ Pages API
Google+ Domains API

YouTube APIs
YouTube Data API
YouTube Analytics API
YouTube Reporting API

Advertising APIs
AdSense Management API
DCM/DFA Reporting And Trafficking API
Ad Exchange Seller API
Ad Exchange Buyer API
DoubleClick Search API
DoubleClick Bid Manager API

Other popular APIs
Analytics API
Custom Search API
URL Shortener API
PageSpeed Insights API
Fusion Tables API
Web Fonts Developer API

KDnuggets™ Subscribe to KDnuggets News | | Contact

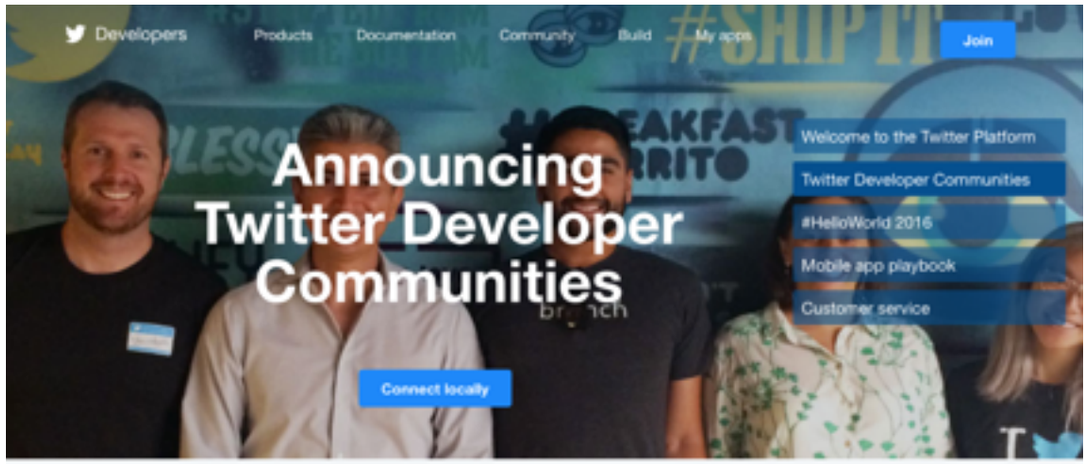
SOFTWARE | NEWS | Top stories | Opinions | Tutorials | JOBS | Companies | Courses | Datasets |

KDnuggets Home » Datasets

Datasets for Data Mining and Data Science

See also

- Government, State, City, Local, public data sites and portals
- Data APIs, Hubs, Marketplaces, Platforms, and Search Engines.
- Data Mining and Data Science Competitions



Explore our products

Publisher platform
The power of Twitter in your website or app

Enterprise data
Real-time and historical Twitter data to power your business

APIs
Programmatic access to read and write Twitter data

Advertising
Drive more mobile ad revenue and measure mobile campaigns

Data Exploration

Raw Data is Dirty!

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”
- Noisy: containing errors or outliers
  - e.g., Salary=“-10”
- Inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why raw data is dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission

Why raw data is dirty?

- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
  - Duplicate records with different attribute values

Preprocessing

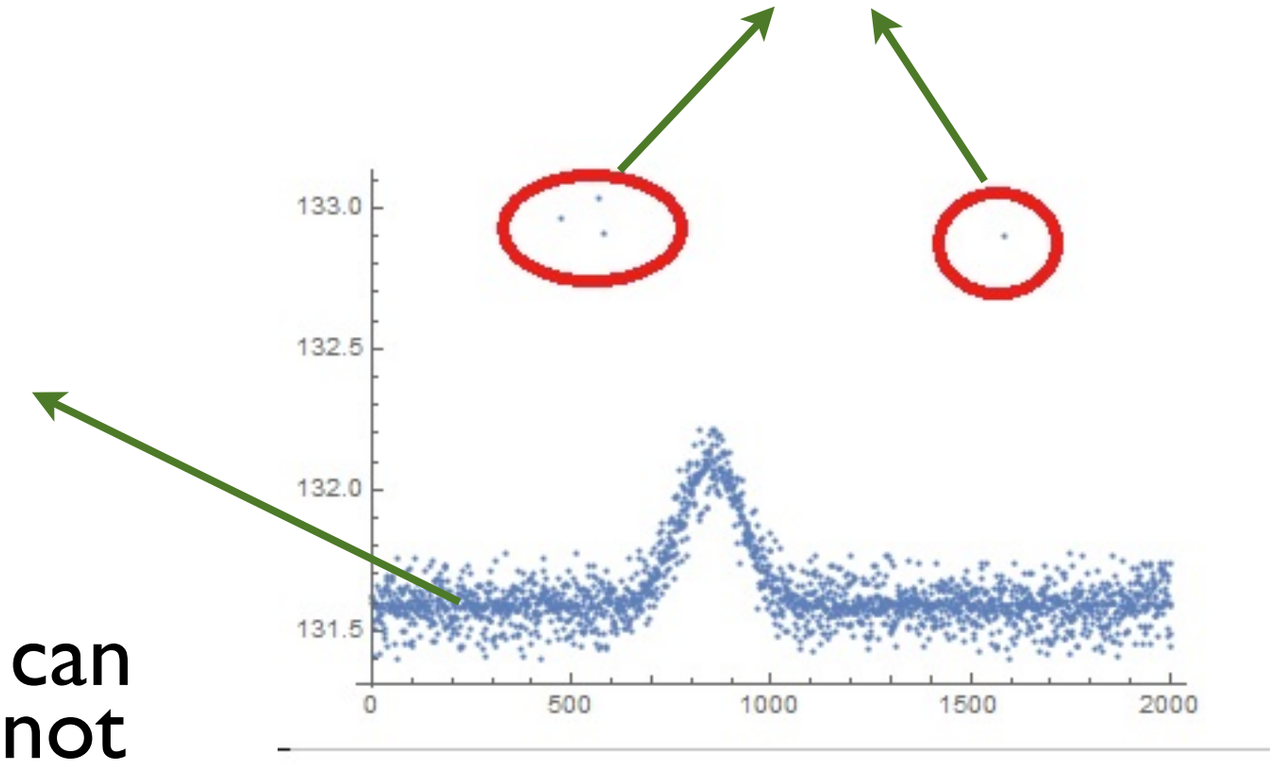
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data transformation
  - Normalization and aggregation reduces the number of values of the attributes; – particular importance especially for numerical data
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but reduces the number of values of the attributes;
  - particular importance especially for numerical data

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Noise and Outliers

- Noise – Modification of original value
  - random
  - non-random (artifact of measurement)
- Noise can be
  - temporal
  - spatial
- Signal processing can reduce (generally not eliminate) noise

- Outliers: Small number of points with characteristics different from rest of the data



Missing Values

- Eliminate row and/or columns with NA
  - Drawback: you may end up removing a large number of objects
- Fill by most common (N) or average over nearest neighbors (Q)
  - Drawback: may introduce bias

User	Device	OS	Transactions
A	Mobile	NA	5
B	Mobile	Android	3
C	NA	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4
F	NA	Android	2
G	Tablet	NA	4

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Y. Fantaye= Yabebal Fantaye
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Redundancy in data integration

- Redundant data occur often when integration of multiple databases
  - 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, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

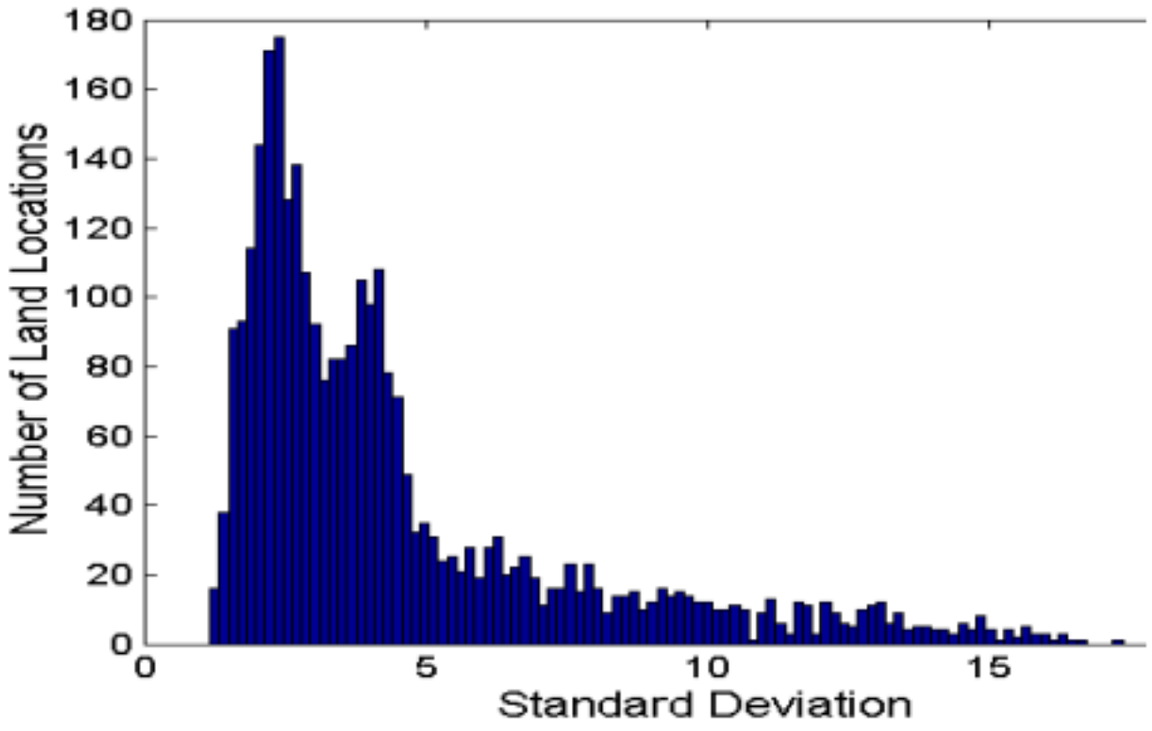
- Data transformation
  - Normalization and aggregation reduces the number of values of the attributes; – particular importance especially for numerical data

Aggregation

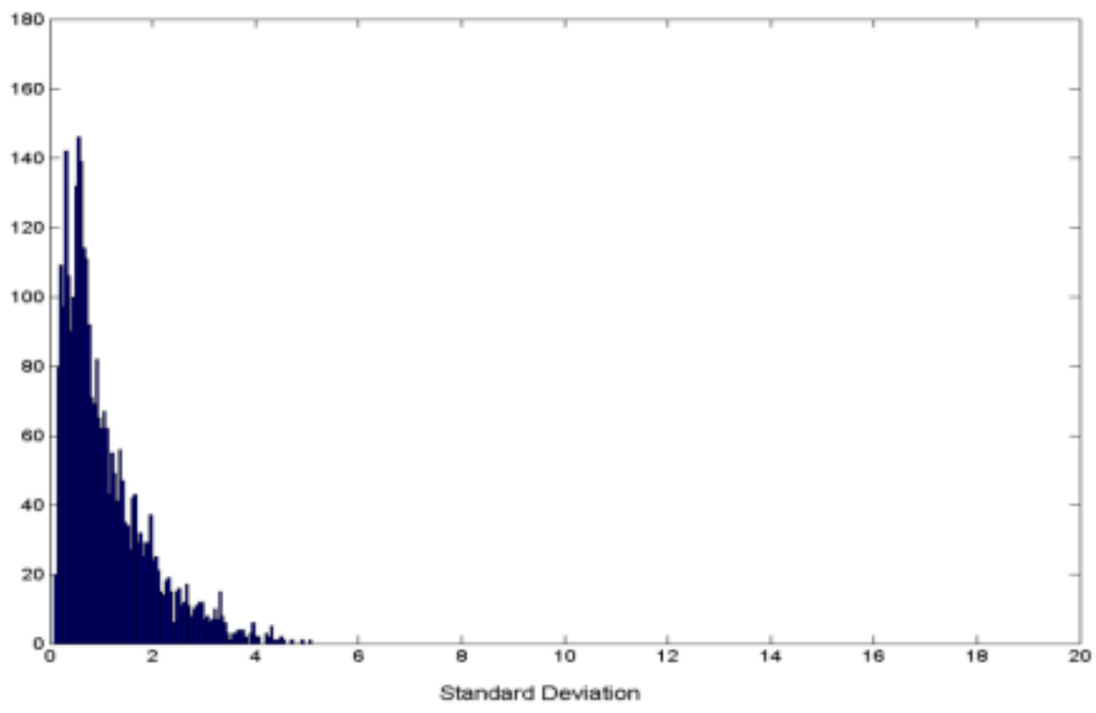
- Aggregation refers to combing two or more attributes (or objects) into a single attribute (or object)
- For example, merging daily sales figures to obtain monthly sales figures
- Why aggregation?
  - Data reduction
    - Allows use of more expensive algorithms
  - If done properly, aggregation can act as scope or scale, providing a high level view of data instead of a low level view

Aggregation

- Behavior of group of objects in more stable than that of individual objects
  - The aggregate quantities have less variability than the individual objects being aggregated



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Feature Creation

- Sometimes, a small number of new attributes can capture the important information in a data set much more efficiently than the original attributes
- Also, the number of new attributes can be often smaller than the number of original attributes. Hence, we get benefits of dimensionality reduction
- Three general methodologies:
  - Feature Extraction
  - Mapping the Data to a New Space
  - Feature Construction

Feature construction

- Sometimes features have the necessary information, but not in the form necessary for the data mining algorithm. In this case, one or more new features constructed out of the original features may be useful
- Example, there are two attributes that record volume and mass of a set of objects
- Suppose there exists a classification model based on material of which the objects are constructed
- Then a density feature constructed from the original two features would help classification

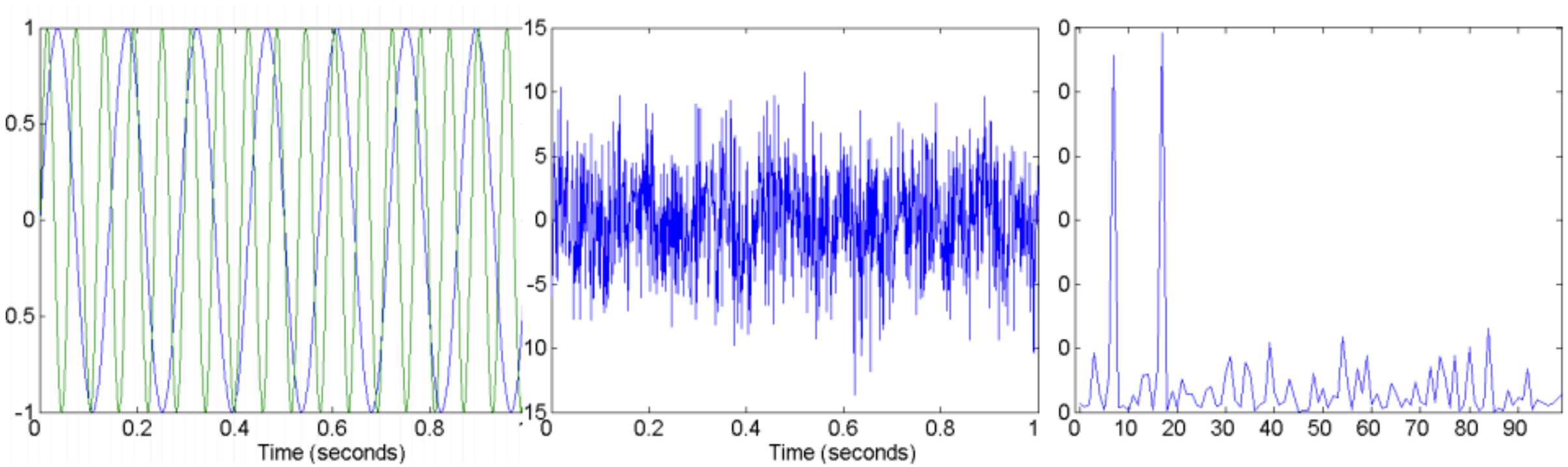
Attribute transformation

- An attribute transformation refers to a transformation that is applied to all values of an attribute, i.e., for each object, the transformation is applied to the value of the attribute for that object

- There are two important types of attribute transformations
  - Simple function transformations
    - Example: x^k , log x, e^x , sqrt x, etc.
  - Standardization or normalization

Mapping data to new space

- Sometimes, a totally different view of the data can reveal important and interesting features
- Example: Applying Fourier transformation to data to detect time series patterns



Original Time Series | Time Series with noise | Frequency plot

- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

Sampling

- Sampling is the process of understanding characteristics of data or models based on a subset of the original data. It is used extensively in all aspects of data exploration and mining.

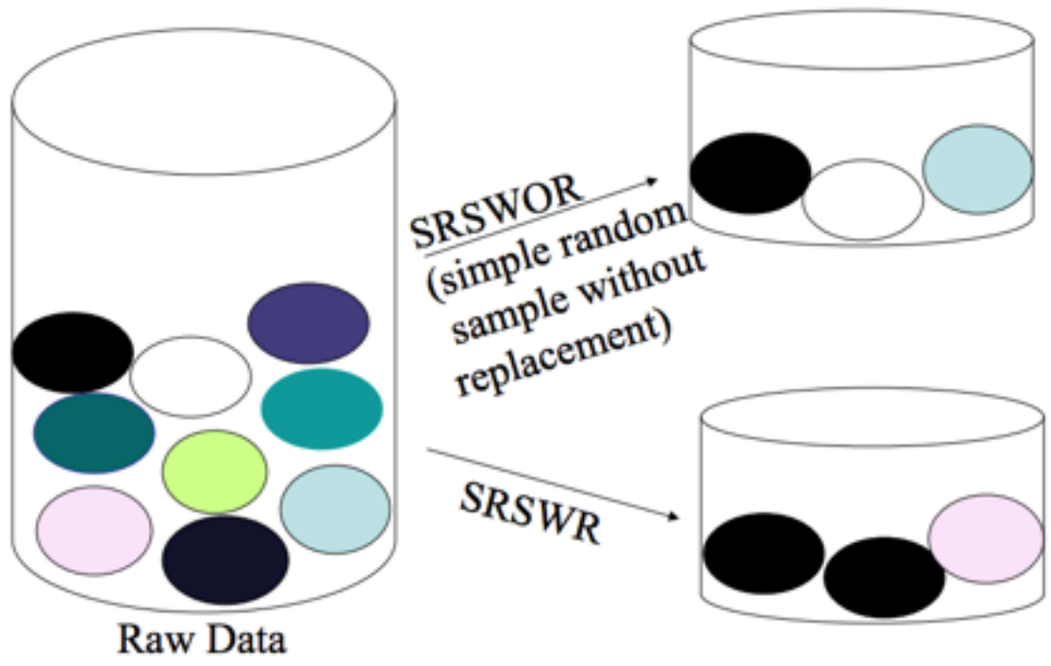
- Why sample
  - Obtaining the entire set of "data of interest" is too expensive or time consuming
  - Obtaining the entire set of data may not be necessary (and hence a waste of resources)

Representative Sample

- A sample is representative for a particular operation if it results in *approximately* the same outcome as if the entire data set was used
- A sample that may be representative for one operation, may not be representative for another operation
  - For example, a sample may be representative for histogram along one dimension but may not be good enough for correlation between two dimensions

Sampling Approaches

- Simple Random Sampling
  - equal probability of selecting any particular item
  - Sampling without replacement: Once an item is selected, it is removed from the population for obtaining future samples
  - Sampling with replacement: Selected item is not removed from the population for obtaining future samples



Sampling Approaches

- Stratified Sampling – When subpopulations vary considerably, it is advantageous to sample each subpopulation (stratum) independently
  - Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling
  - The strata should be mutually exclusive : every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive : no population element can be excluded
  - Then random sampling is applied within each stratum. This often improves the representative-ness of the sample by reducing sampling error

Sample Size

- Even if proper sampling technique is known, it is important to choose proper sample size
- Larger sample sizes increase the probability that a sample will be representative, but also eliminate much of the advantage of sampling
- With smaller sample size, patterns may be missed or erroneous patters detected

Sample Size / Fraction

8000 points

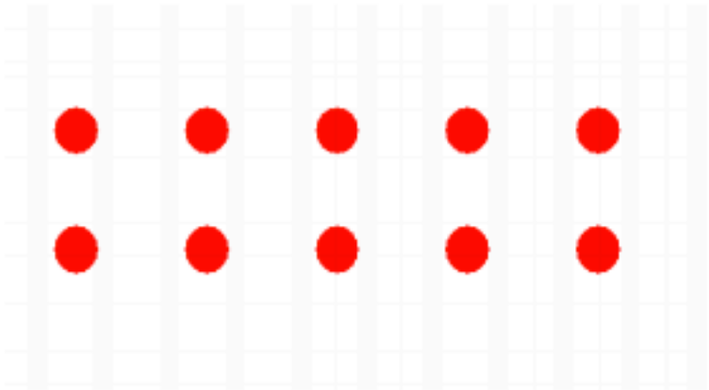
2000 points

500 points

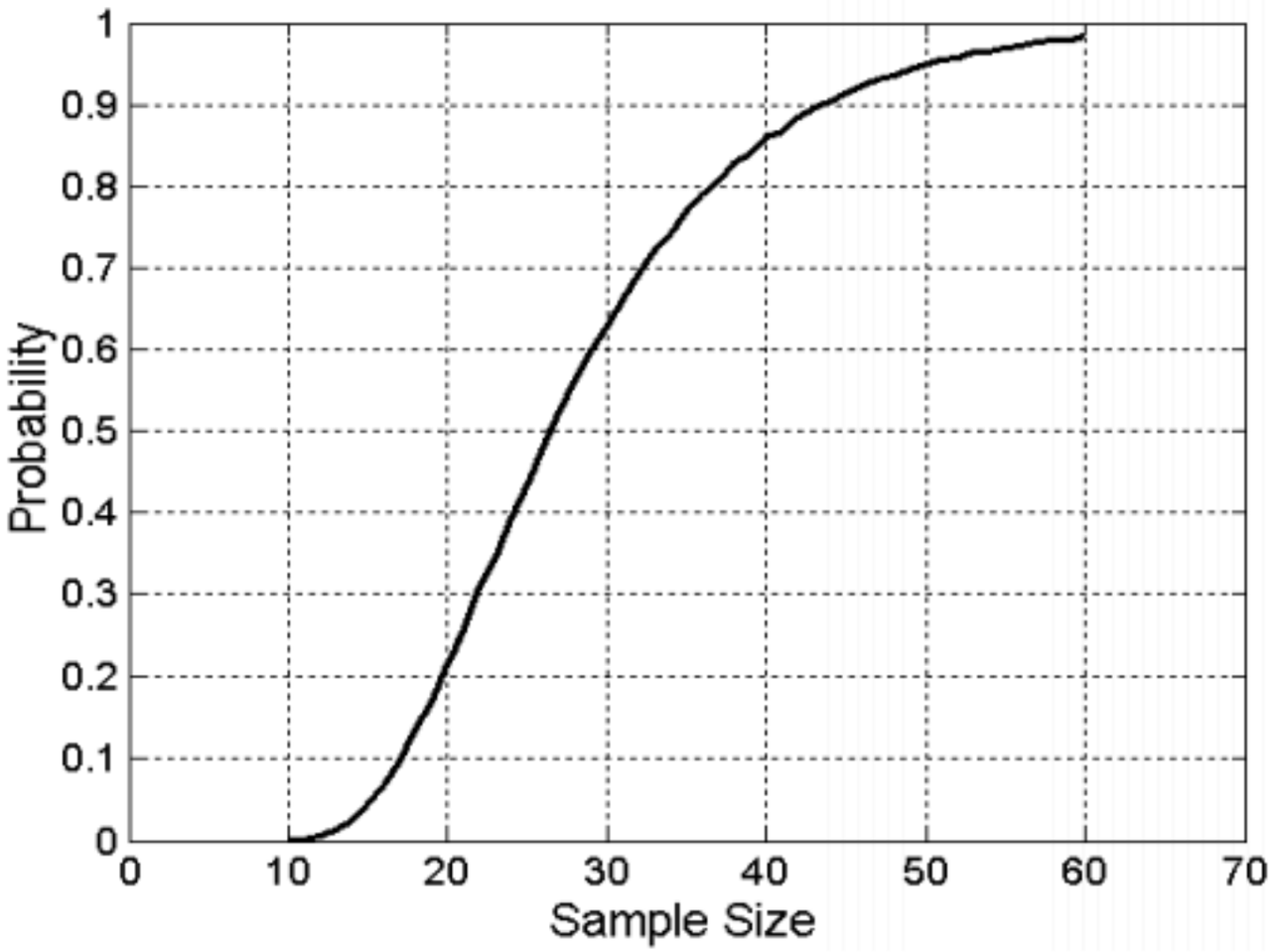
Example of the Loss of Structure with Sampling

Sample Size

- Sample size required to obtain at least one sample from each group



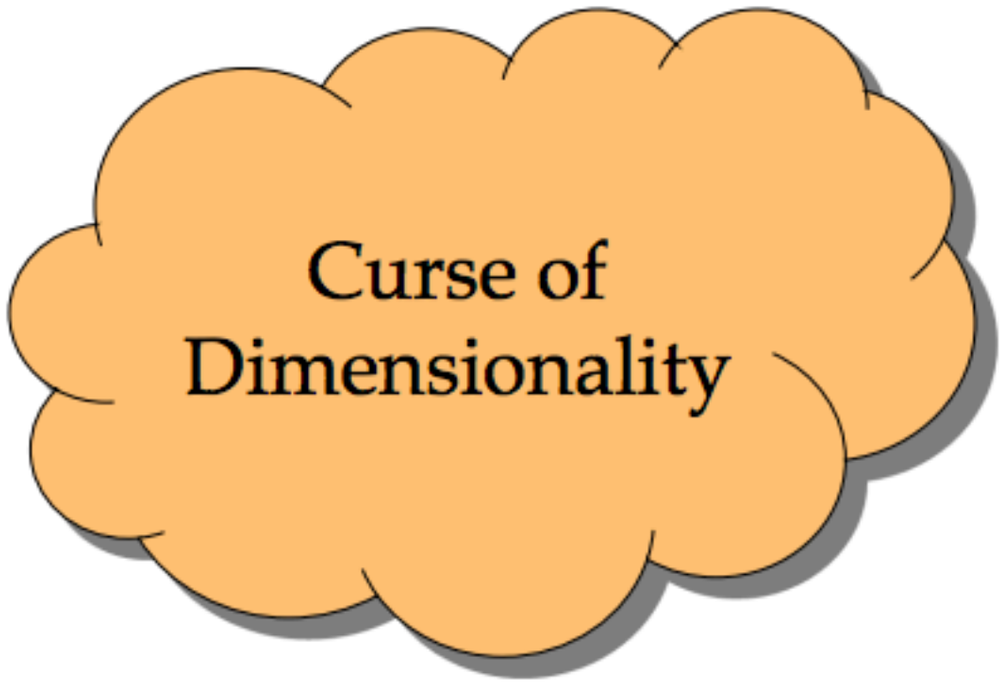
Ten Clusters

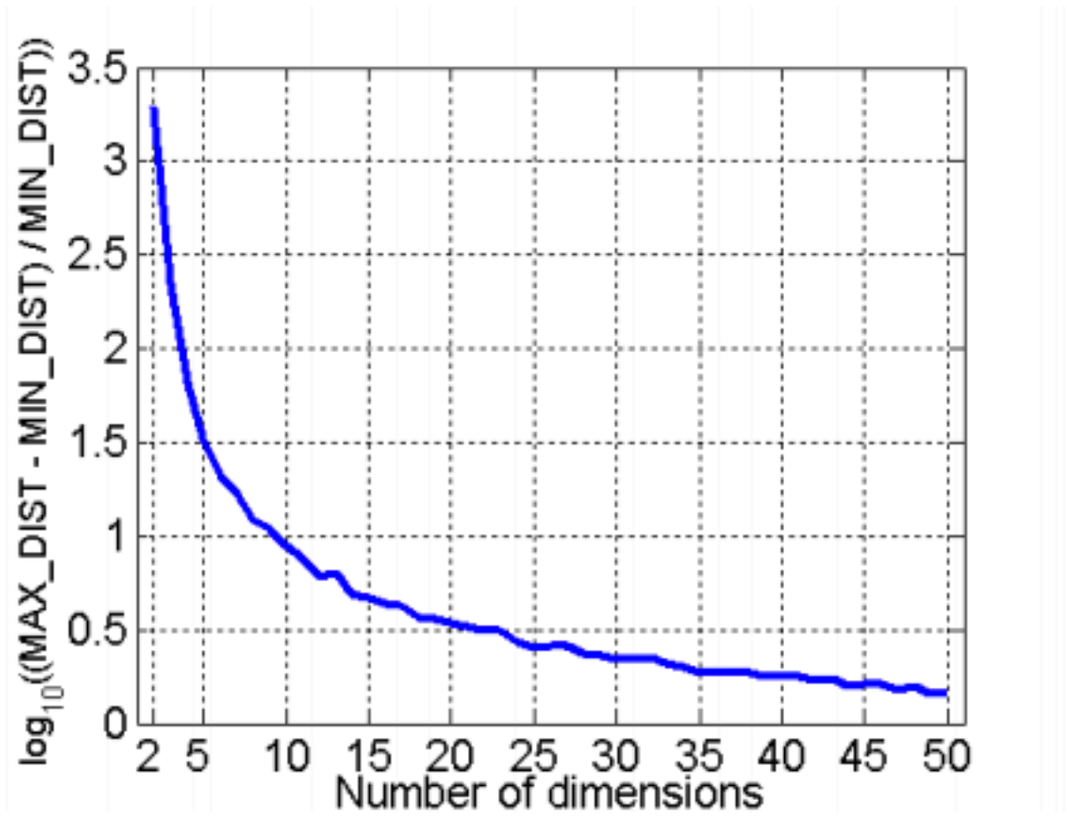


Probability a sample contains points from each of ten groups

Dimension reduction

Curse of dimensionality: Data analysis becomes significantly harder as the dimensionality of the data increases





Decrease in the relative distance between points
As dimensionality increases

Dimension reduction

- Determining dimensions (or combinations of dimensions) that are important for modeling
- Why dimensionality reduction?
  - Many data mining algorithms work better if the dimensionality of data (i.e. number of attributes) is lower
  - Allows the data to be more easily visualized
  - If dimensionality reduction eliminates irrelevant features or reduces noise, then quality of results may improve
  - Can lead to a more understandable model

Dimension reduction

- Redundant features duplicate much or all of the information contained in one or more attributes
  - The purchase price of product and the sales tax paid contain the same information
- Irrelevant features contain no information that is useful for data mining task at hand
  - Student ID numbers would be irrelevant to the task of predicting their GPA

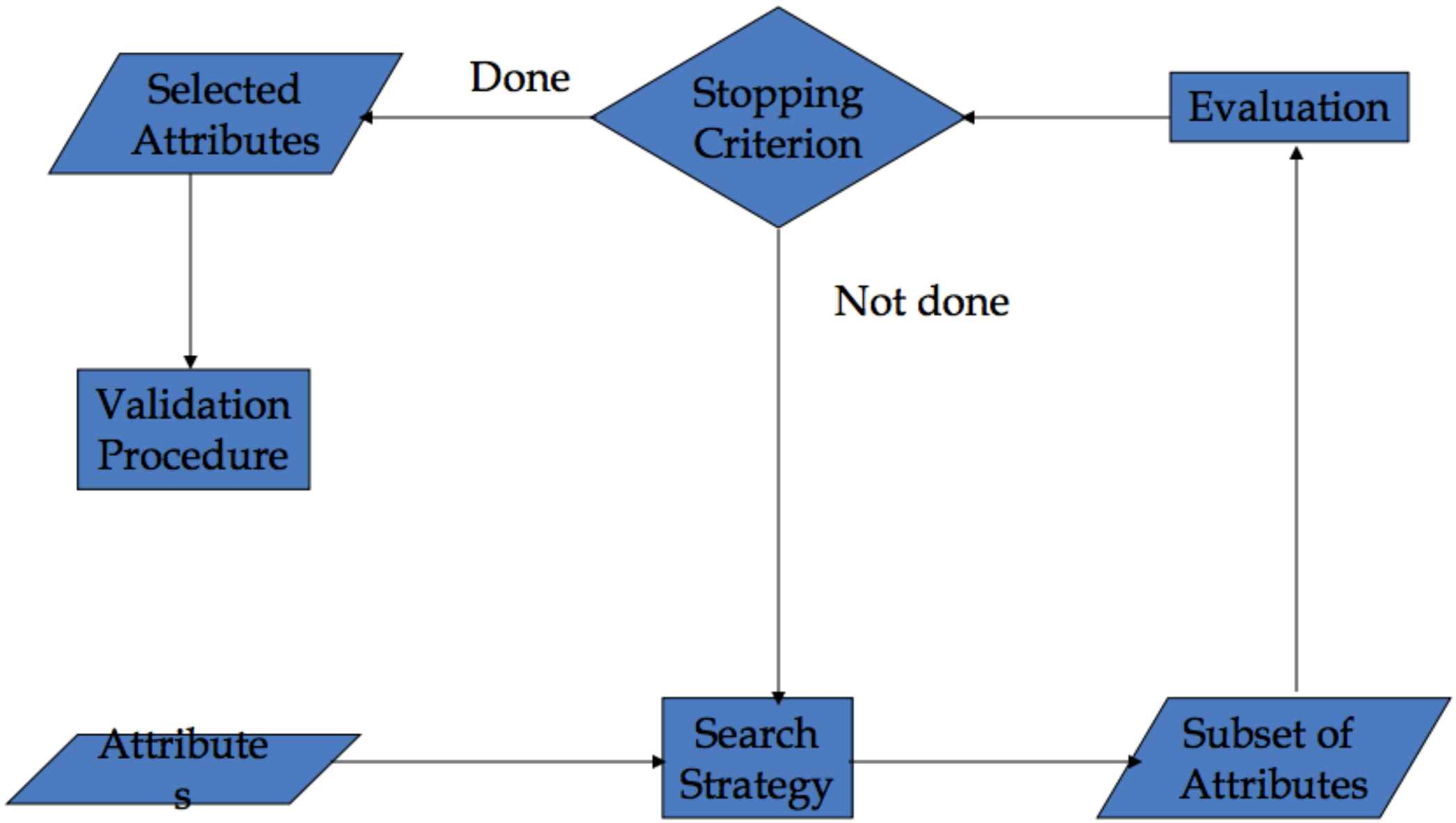
Principal Component Analysis

- PCA is a linear algebra technique for continuous attributes that finds new attributes (principal components) that
  - Are linear combinations of original attributes
  - Are orthogonal to each other
  - Capture the maximum amount of variation in data

Feature (subset) selection

- There are three standard approaches to feature selection:
  - Embedded approaches: Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches: Features are selected before the data mining algorithm is run
  - Wrapper approaches: Use the target data mining algorithm as a black box to find the best subset of attributes (typically without enumerating all subsets)

Feature (subset) selection



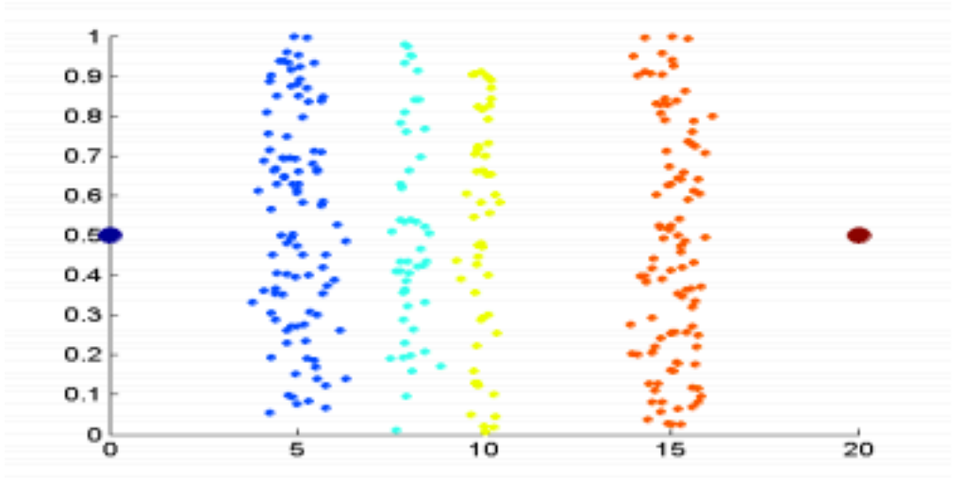
Flowchart of a feature subset selection process

- Data discretization
  - Part of data reduction but reduces the number of values of the attributes;
  - particular importance especially for numerical data

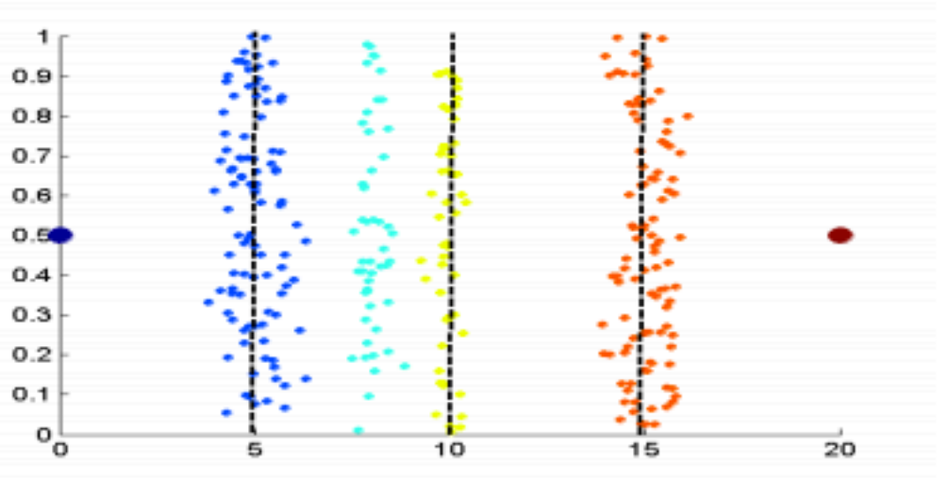
Discretization and Binarization

- Discretization is the process of converting a continuous attribute to a discrete attribute
- A common example is rounding off real numbers to integers
- Some data mining algorithms require that the data be in the form of categorical or binary attributes. Thus, it is often necessary to convert continuous attributes in to categorical attributes and / or binary attributes
- Its pretty straightforward to convert categorical attributes in to discrete or binary attributes

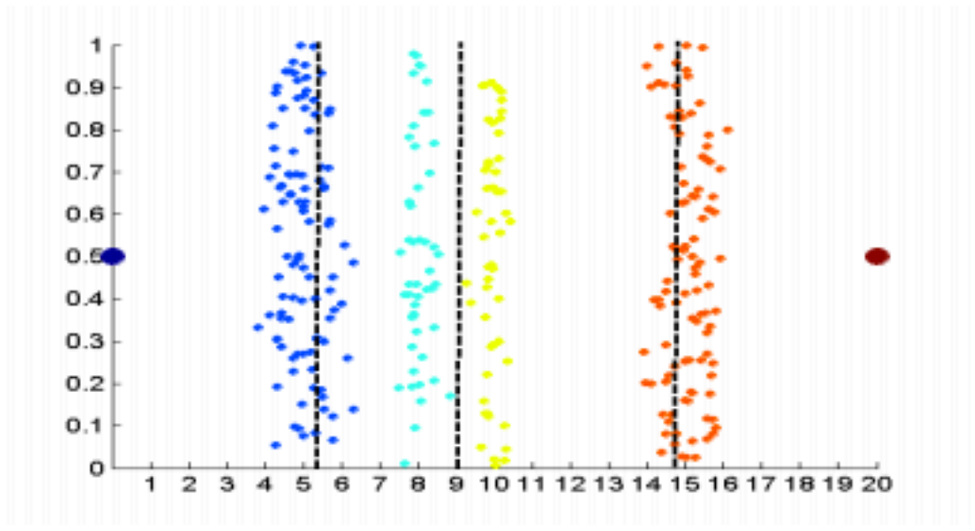
Discretization technique



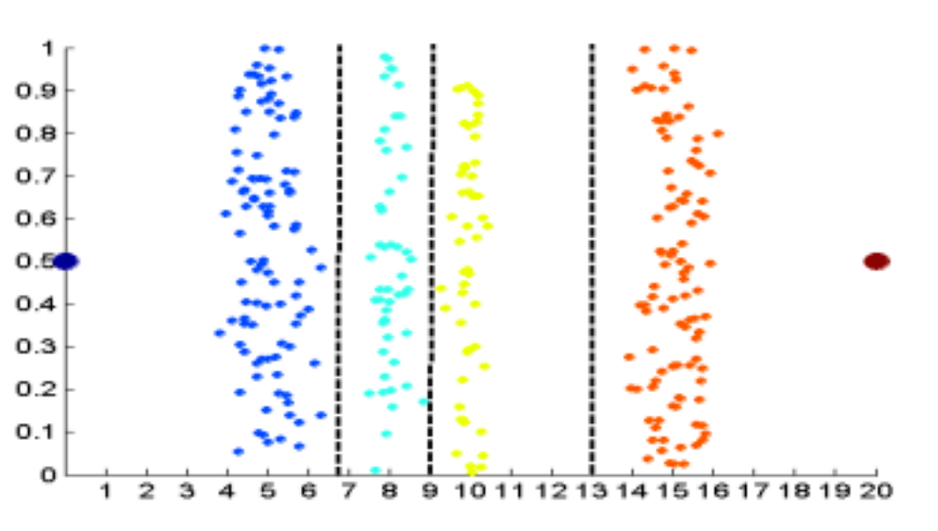
Data



Equal interval width



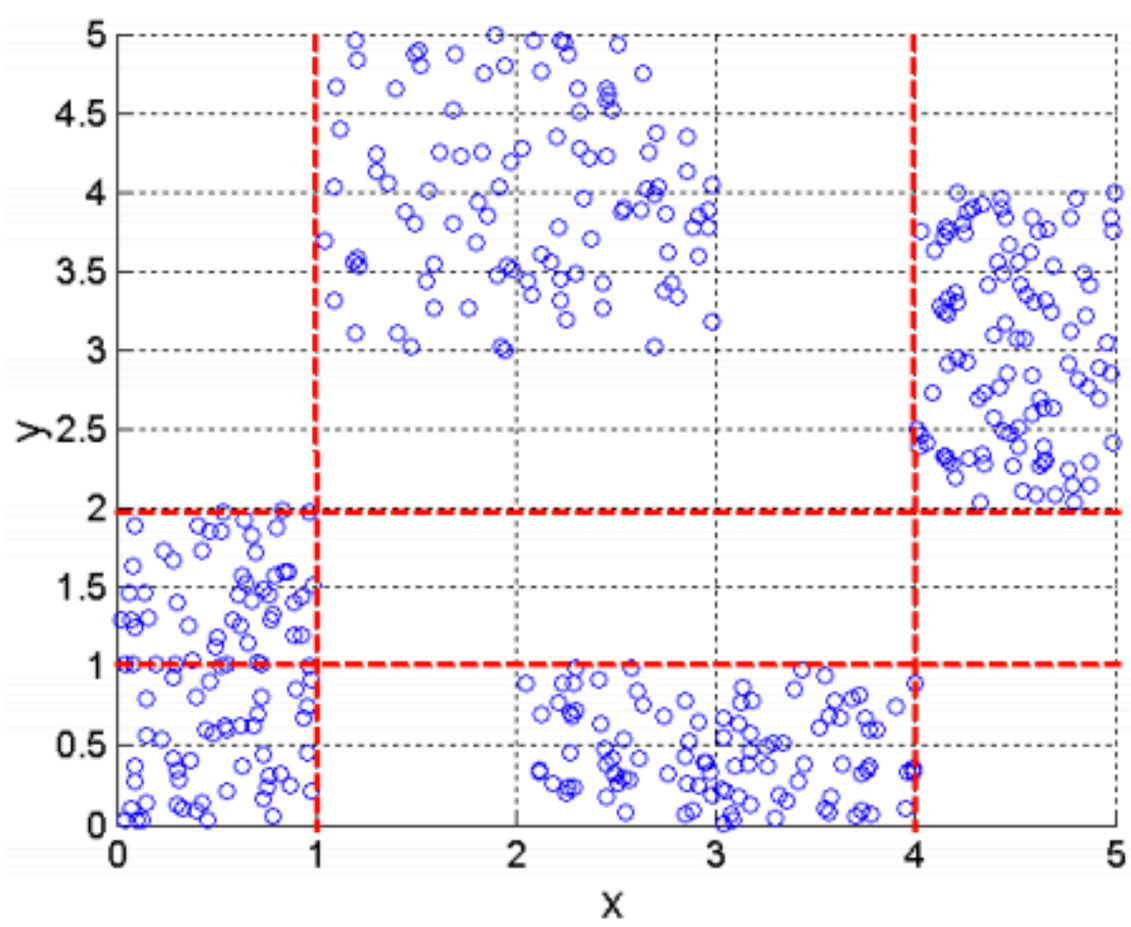
Equal frequency



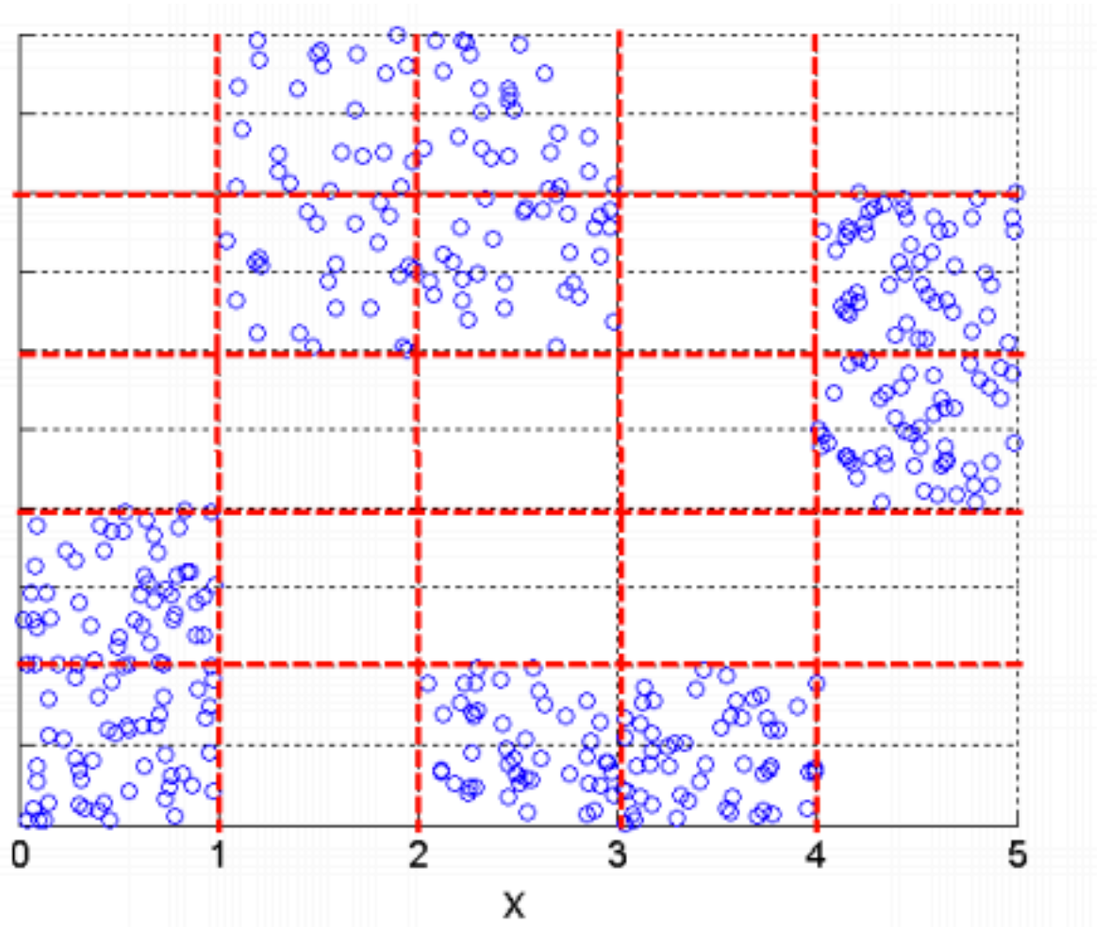
K-means

Discretization technique

Entropy based approach



3 categories for both x and y



5 categories for both x and y

Case Study

Expedia

https://www.dataquest.io/blog/kaggle-tutorial/

https://www.kaggle.com/c/expedia-hotel-recommendations

Expedia Hotel Recommendations

Which hotel type will an Expedia customer book?

$25,000 · 1,974 teams · a year ago

Overview

Planning your dream vacation, or even a weekend escape, can be an overwhelming affair. With hundreds, even thousands, of hotels to choose from at every destination, it's difficult to know which will suit your personal preferences. Should you go with an old standby with those pillow mints you like, or risk a new hotel with a trendy pool bar?

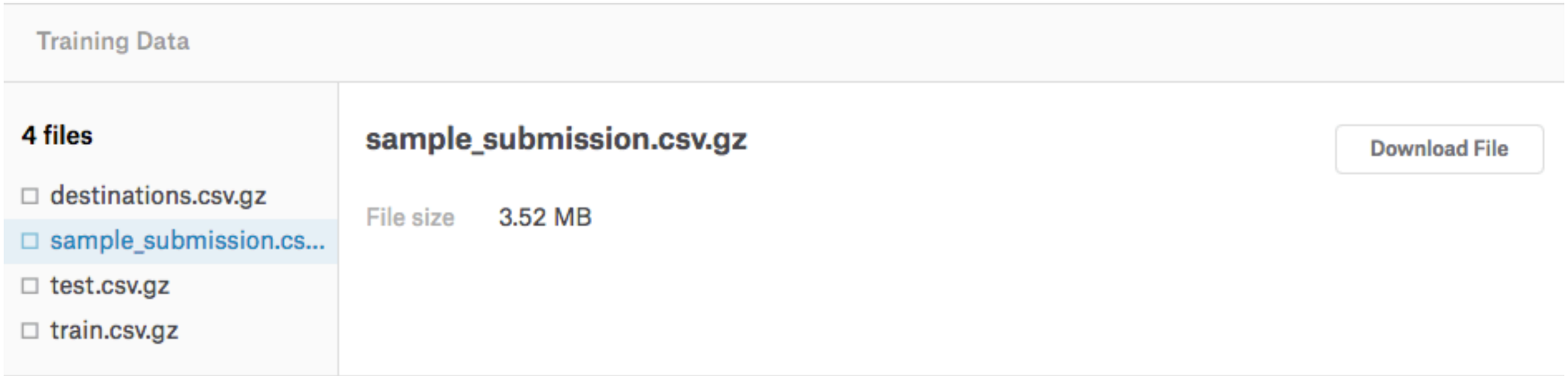
Expedia wants to take the proverbial rabbit hole out of hotel search by providing personalized hotel recommendations to their users. This is no small task for a site with hundreds of millions of visitors every month!

Currently, Expedia uses search parameters to adjust their hotel recommendations, but there aren't enough customer specific data to personalize them for each user. In this competition, Expedia is challenging Kagglers to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.

The data in this competition is a random selection from Expedia and is not representative of the overall statistics.

Expedia Kaggle competition

- Challenge: what hotel a user will book based on some attributes about the search the user is conducting on Expedia



train.csv: 37.7 million rows by 24 columns

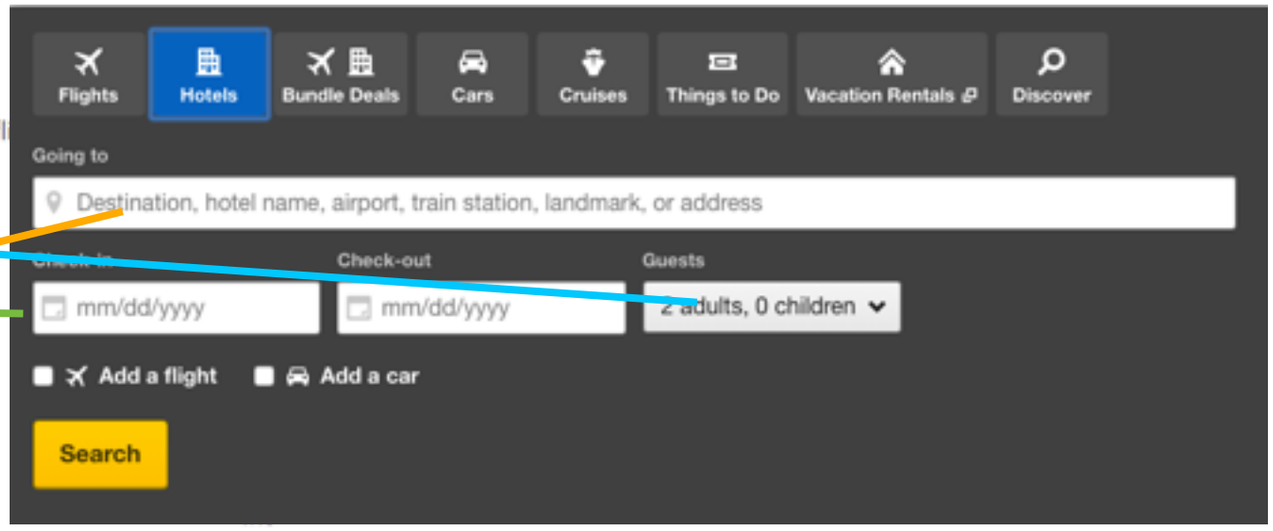
test.csv: 2.5 million rows by 22 rows

destinations.csv: 149 columns

Data fields

train/test.csv

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated	double
user_id	ID of user	
is_mobile	1 when a user connected from a mobile device, 0 otherwise	
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a fl	
channel	ID of a marketing channel	
srch_ci	Checkin date	
srch_co	Checkout date	
srch_adults_cnt	The number of adults specified in the hotel room	
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room	
srch_rm_cnt	The number of hotel rooms specified in the search	
srch_destination_id	ID of the destination where the hotel search was performed	
srch_destination_type_id	Type of destination	
hotel_continent	Hotel continent	int
hotel_country	Hotel country	
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Numer of similar events in the context of the same user session	bigint
hotel_cluster	ID of a hotel cluster	int



destinations.csv

Column name	Description	Data type
srch_destination_id	ID of the destination where the hotel search was performed	int
d1-d149	latent description of search regions	double

Case Study

Prudential Life insurance

https://www.kaggle.com/c/prudential-life-insurance-assessment

Prudential Life Insurance Assessment

Can you make buying life insurance easier?

$30,000 · 2,619 teams · a year ago

Overview

Picture this. You are a data scientist in a start-up culture with the potential to have a very large impact on the business. Oh, and you are backed up by a company with 140 years' business experience.

Curious? Great! You are the kind of person we are looking for.

Prudential, one of the largest issuers of life insurance in the USA, is hiring passionate data scientists to join a newly-formed Data Science group solving complex challenges and identifying opportunities. The results have been impressive so far but we want more.

The Challenge

In a one-click shopping world with on-demand everything, the life insurance application process is antiquated. Customers provide extensive information to identify risk classification and eligibility, including scheduling medical exams, a process that takes an average of 30 days.

The result? People are turned off. That's why only 40% of U.S. households own individual life insurance. Prudential wants to make it quicker and less labor intensive for new and existing customers to get a quote while maintaining privacy boundaries.

By developing a predictive model that accurately classifies risk using a more automated approach, you can greatly impact public perception of the industry.

Data fields

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

The following variables are all categorical (nominal):

Product_Info_1, Product_Info_2, Product_Info_3, Product_Info_5, Product_Info_6, Product_Info_7, Employment_Info_2, Employment_Info_3, Employment_Info_5, InsuredInfo_1, InsuredInfo_2, InsuredInfo_3, InsuredInfo_4, InsuredInfo_5, InsuredInfo_6, InsuredInfo_7, Insurance_History_1, Insurance_History_2, Insurance_History_3, Insurance_History_4, Insurance_History_7, Insurance_History_8, Insurance_History_9, Family_Hist_1, Medical_History_2, Medical_History_3, Medical_History_4, Medical_History_5, Medical_History_6, Medical_History_7, Medical_History_8, Medical_History_9, Medical_History_11, Medical_History_12, Medical_History_13, Medical_History_14, Medical_History_16, Medical_History_17, Medical_History_18, Medical_History_19, Medical_History_20, Medical_History_21, Medical_History_22, Medical_History_23, Medical_History_25, Medical_History_26, Medical_History_27, Medical_History_28, Medical_History_29, Medical_History_30, Medical_History_31, Medical_History_33, Medical_History_34, Medical_History_35, Medical_History_36, Medical_History_37, Medical_History_38, Medical_History_39, Medical_History_40, Medical_History_41

The following variables are continuous:

Product_Info_4, Ins_Age, Ht, Wt, BMI, Employment_Info_1, Employment_Info_4, Employment_Info_6, Insurance_History_5, Family_Hist_2, Family_Hist_3, Family_Hist_4, Family_Hist_5

The following variables are discrete:

Medical_History_1, Medical_History_10, Medical_History_15, Medical_History_24, Medical_History_32

Medical_Keyword_1-48 are dummy variables.

Case Study

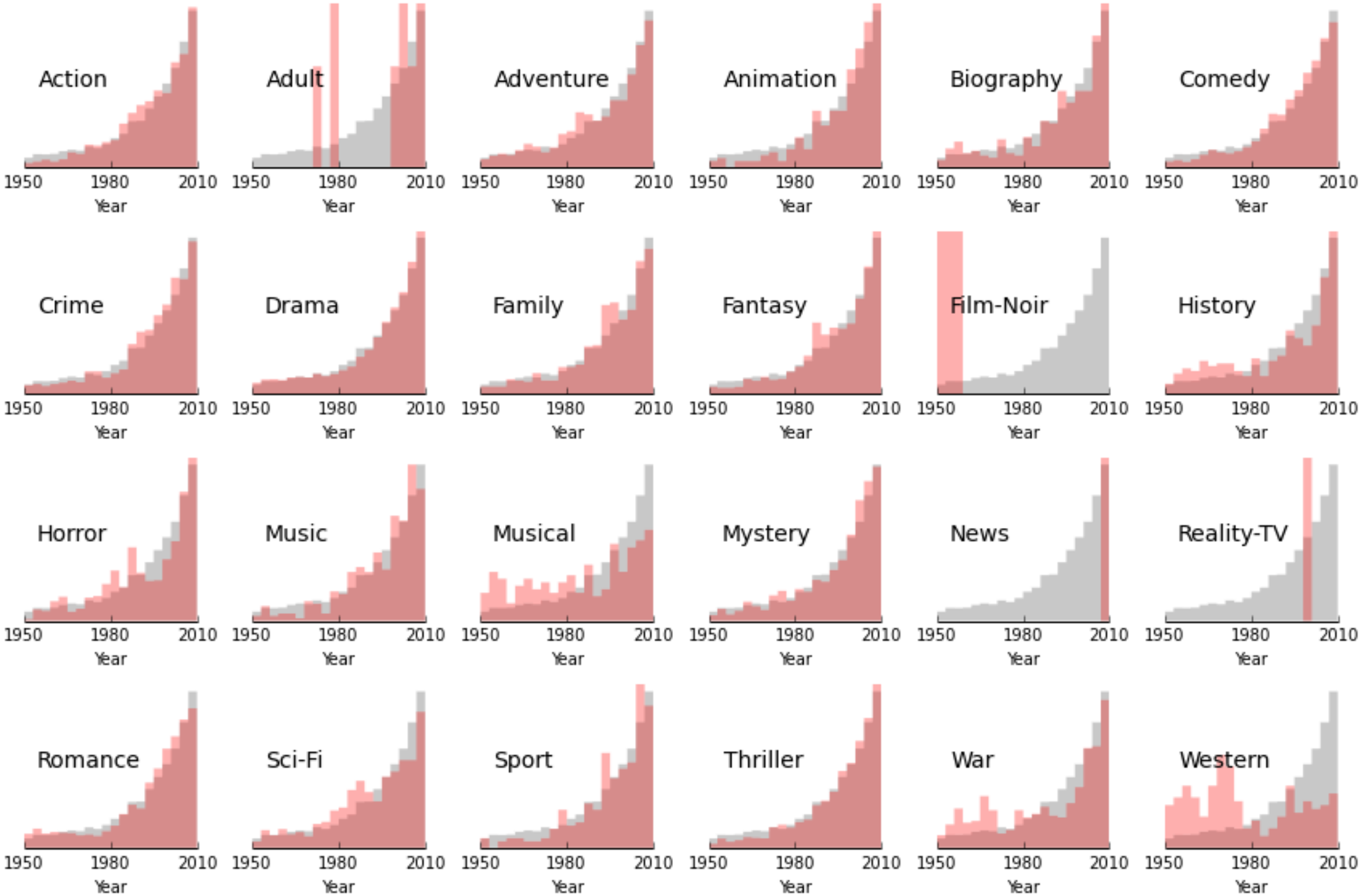
IMDB

http://nbviewer.jupyter.org/github/cs109/content/blob/master/lec_04_wrangling.ipynb

Explore Movie ratings from imdb.com

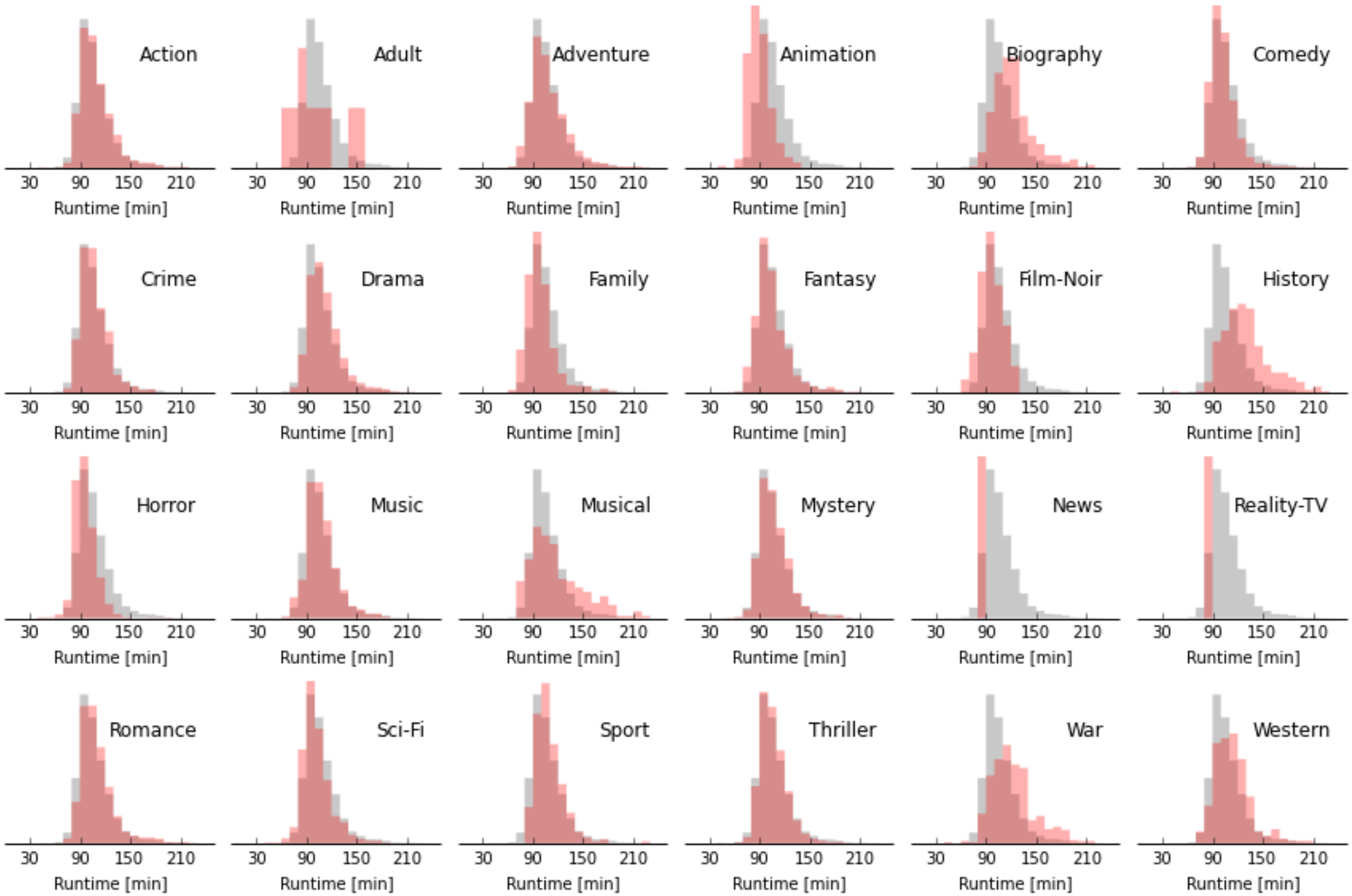
Data is obtained by Web scrapping

Grey: all movies



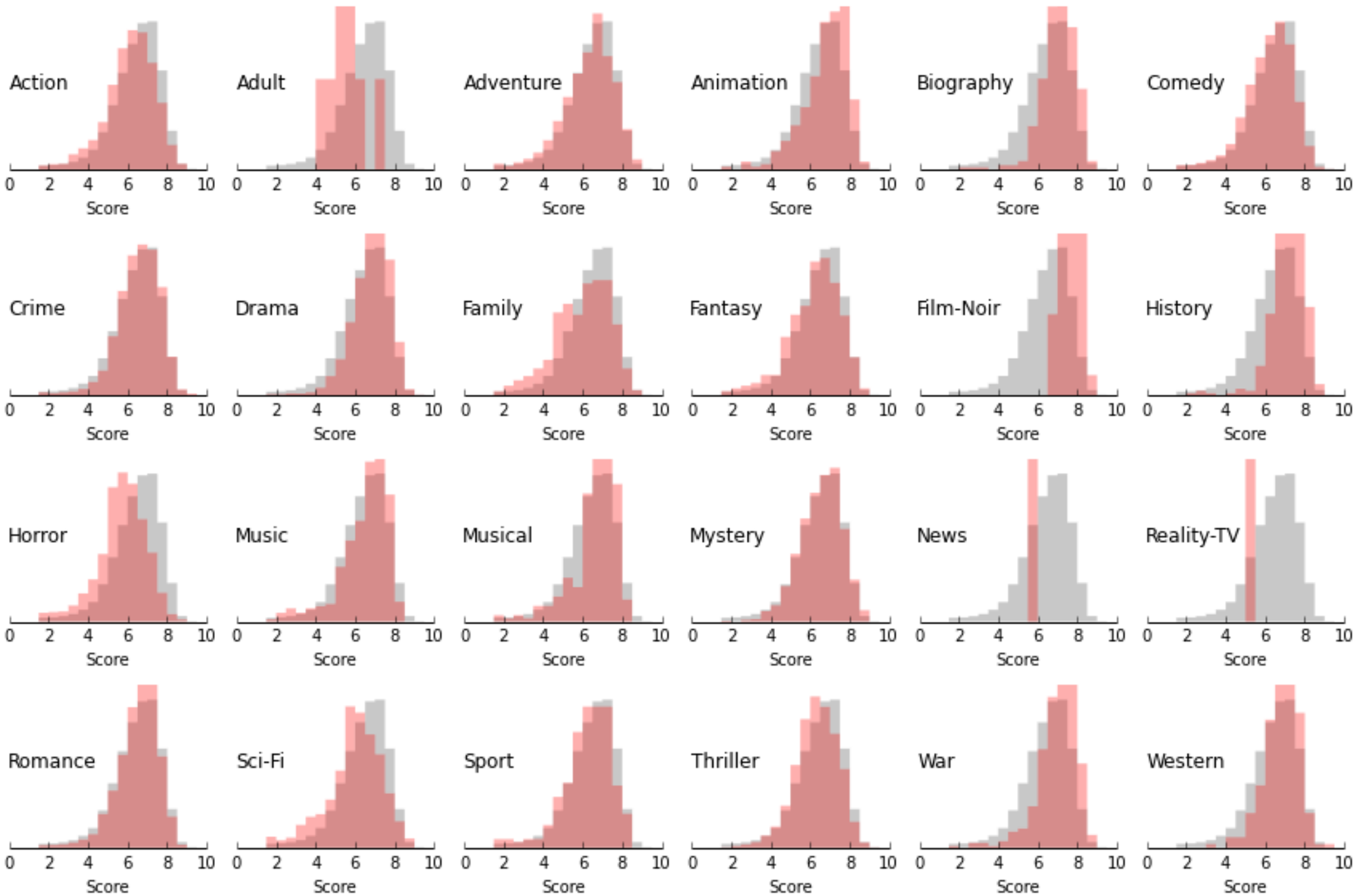
Explore Movie ratings from imdb.com

Data is obtained by Web scrapping



Explore Movie ratings from imdb.com

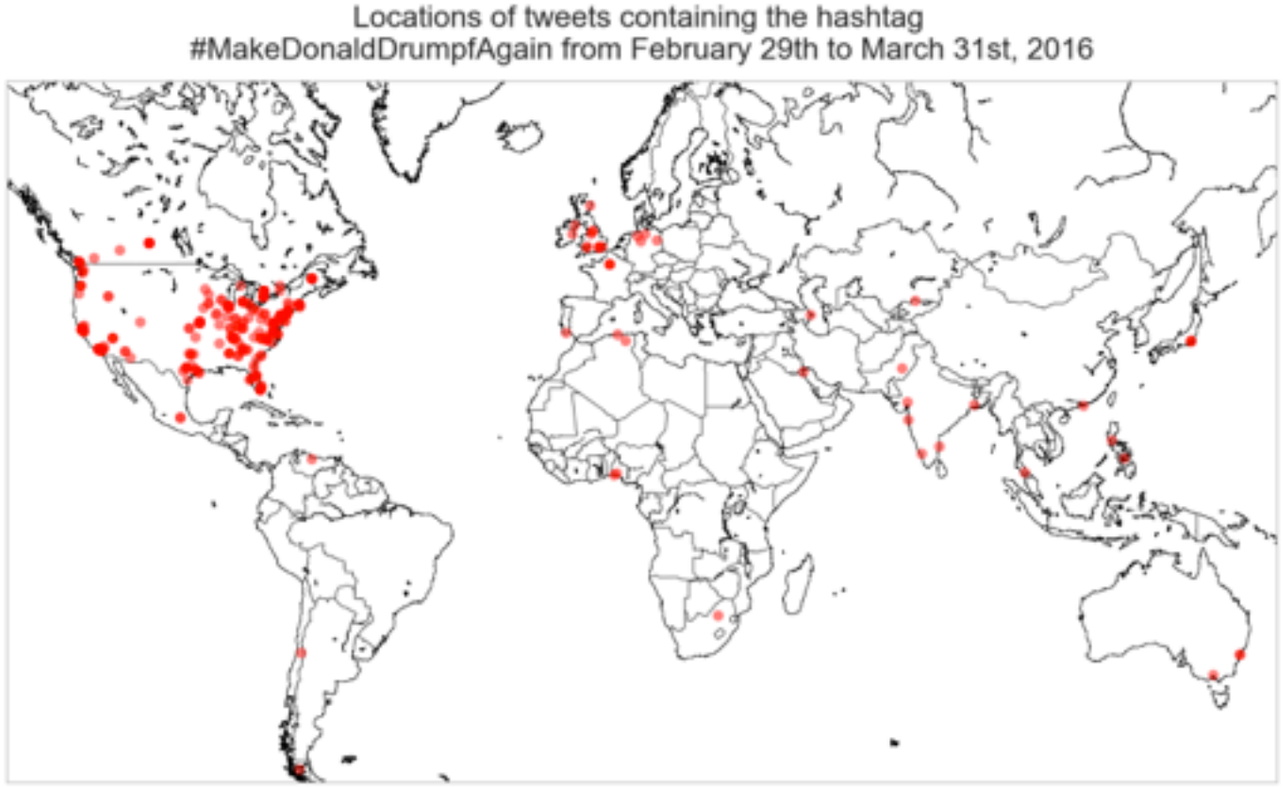
Data is obtained by Web scrapping

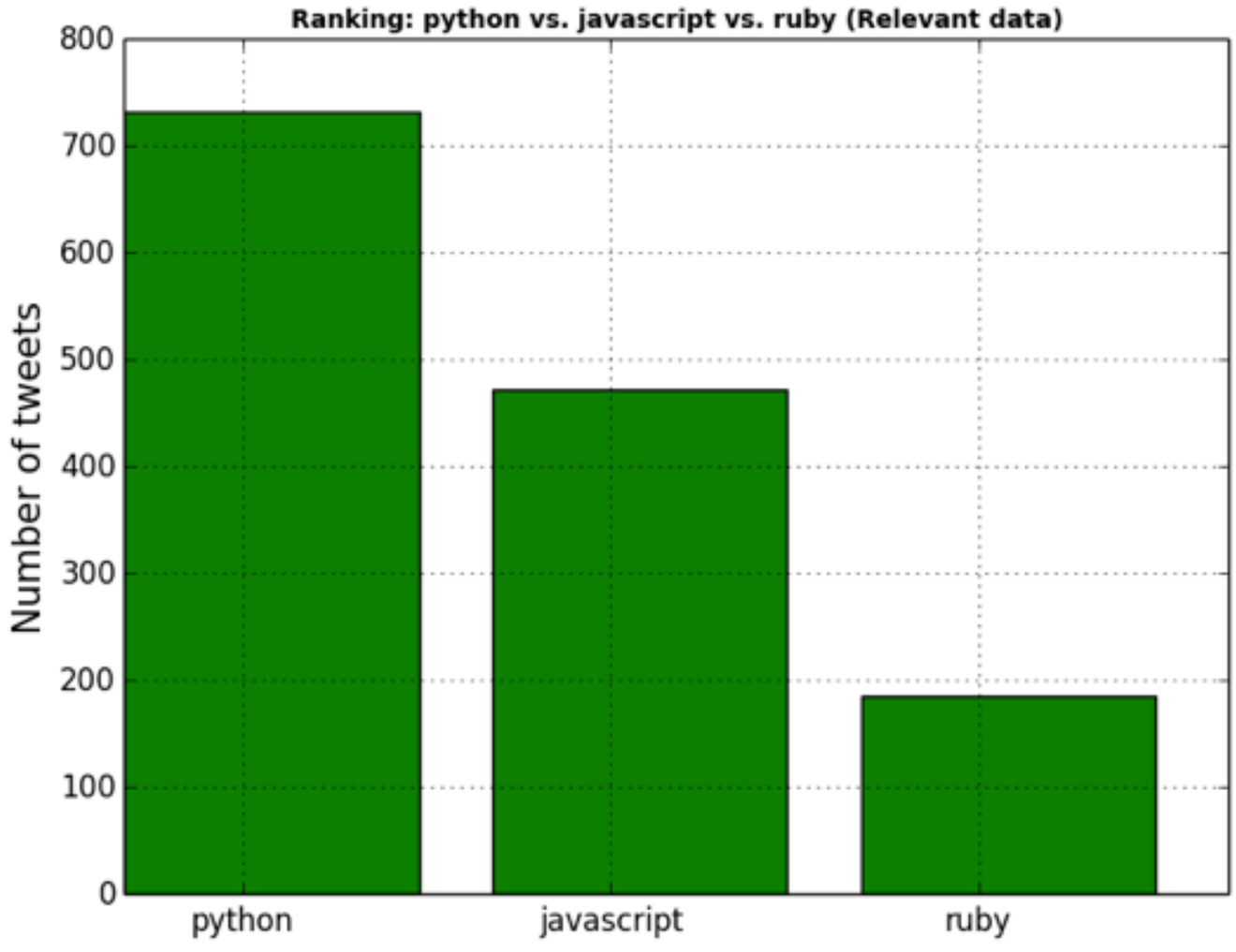


Case Study
Twitter text mining

http://adilmoujahid.com/posts/2014/07/twitter-analytics/

https://galeascience.wordpress.com/2016/03/18/collecting-twitter-data-with-python/


Locations of tweets containing the hashtag
#MakeDonaldDrumpfAgain from February 29th to March 31st, 2016


Ranking: python vs. javascript vs. ruby (Relevant data)
Number of tweets
0
100
200
300
400
500
600
700
800
python
javascript
ruby