How to
Read & Represent Data

ISE:4172 Big Data Analytics

Stephen Baek

Vocabulary

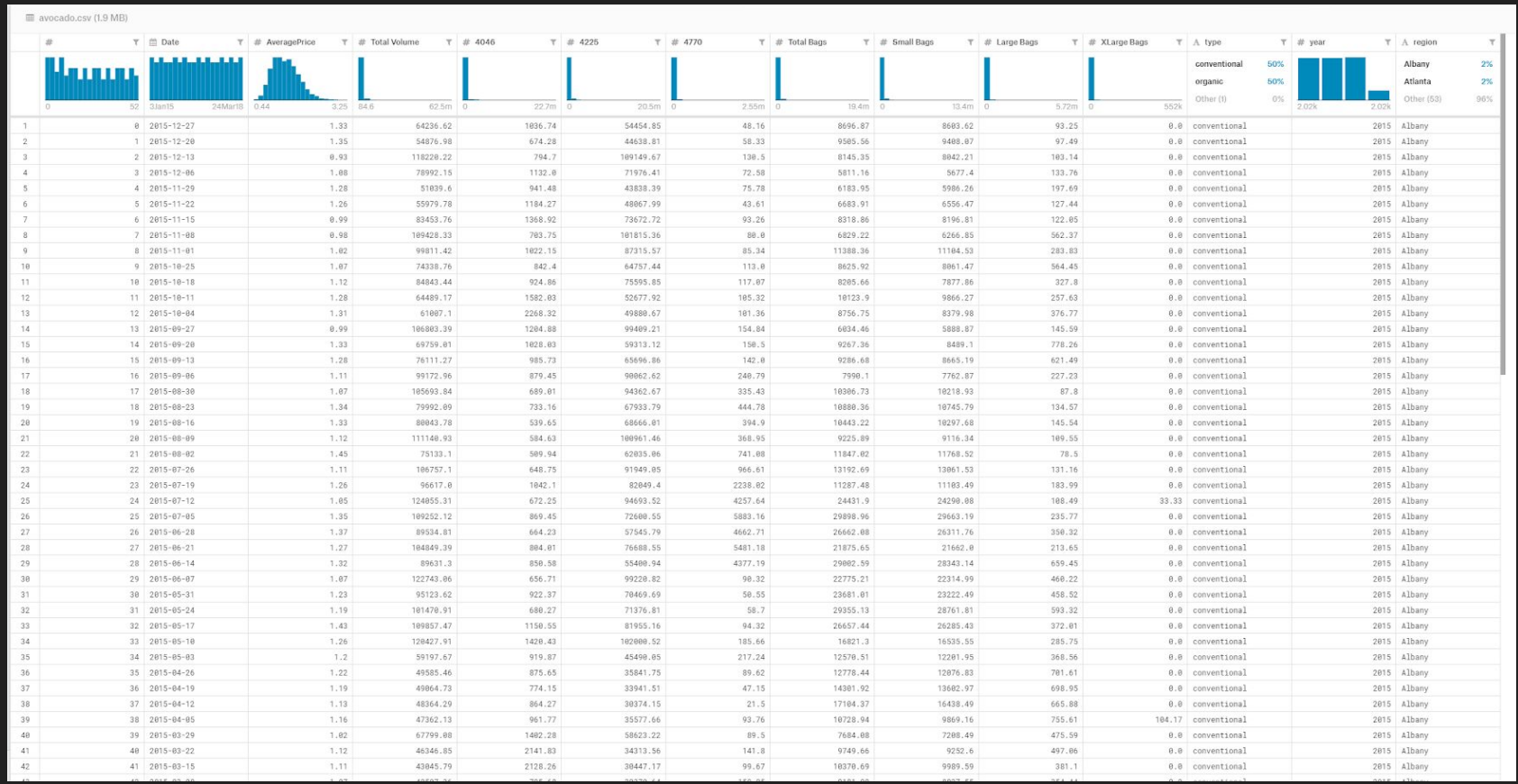
- Data: noun, *plural (singular: datum)* (dā-tə; dä-)
  - Collection of entities and attributes

Vocabulary

- Data: noun, *plural (singular: datum)* (dā-tə; dä-)
  - Collection of entities and attributes
- Object:
  - Also known as entity, sample, instance, data point, record, etc.
  - “Row” of the table
  - e.g. a person, a school, a tweet

Vocabulary

- Data: noun, *plural (singular: datum)* (dā-tə; dä-)
  - Collection of entities and attributes
- Object:
  - Also known as entity, sample, instance, data point, record, etc.
  - “Row” of the table
  - e.g. a person, a school, a tweet
- Attribute:
  - Also known as field, feature, parameter, variable, code, encoding, etc.
  - “Column” of the table
  - e.g. BMI of a person, student enrollment of a school, number of words in a tweet



avocado.csv (1.9 MB)

	#	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
	0 – 52	3Jan15 – 24Mar18	0.44 – 3.25	84.6 – 62.5m	0 – 22.7m	0 – 20.5m	0 – 2.55m	0 – 19.4m	0 – 13.4m	0 – 5.72m	0 – 552k	conventional 50% organic 50% Other (1) 0%	2.02k – 2.02k	Albany 2% Atlanta 2% Other (53) 96%
1	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
2	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
3	2	2015-12-13	0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
4	3	2015-12-06	1.08	78992.15	1132.0	71976.41	72.58	5811.16	5677.4	133.76	0.0	conventional	2015	Albany
5	4	2015-11-29	1.28	51039.6	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
6	5	2015-11-22	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0.0	conventional	2015	Albany
7	6	2015-11-15	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0.0	conventional	2015	Albany
8	7	2015-11-08	0.98	109428.33	703.75	101815.36	80.0	6829.22	6266.85	562.37	0.0	conventional	2015	Albany
9	8	2015-11-01	1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0.0	conventional	2015	Albany
10	9	2015-10-25	1.07	74338.76	842.4	64757.44	113.0	8625.92	8061.47	564.45	0.0	conventional	2015	Albany
11	10	2015-10-18	1.12	84843.44	924.86	75595.85	117.07	8285.66	7877.86	327.8	0.0	conventional	2015	Albany
12	11	2015-10-11	1.28	64489.17	1582.03	52677.92	105.32	10123.9	9866.27	257.63	0.0	conventional	2015	Albany
13	12	2015-10-04	1.31	61007.1	2268.32	49880.67	101.36	8756.75	8379.98	376.77	0.0	conventional	2015	Albany
14	13	2015-09-27	0.99	106803.39	1204.88	99409.21	154.84	6034.46	5888.87	145.59	0.0	conventional	2015	Albany
15	14	2015-09-20	1.33	69759.01	1028.03	59313.12	150.5	9267.36	8489.1	778.26	0.0	conventional	2015	Albany
16	15	2015-09-13	1.28	76111.27	985.73	65696.86	142.0	9286.68	8665.19	621.49	0.0	conventional	2015	Albany
17	16	2015-09-06	1.11	99172.96	879.45	90062.62	240.79	7990.1	7762.87	227.23	0.0	conventional	2015	Albany
18	17	2015-08-30	1.07	105693.84	689.01	94362.67	335.43	10306.73	10218.93	87.8	0.0	conventional	2015	Albany
19	18	2015-08-23	1.34	79992.09	733.16	67933.79	444.78	10880.36	10745.79	134.57	0.0	conventional	2015	Albany
20	19	2015-08-16	1.33	80043.78	539.65	68666.01	394.9	10443.22	10297.68	145.54	0.0	conventional	2015	Albany
21	20	2015-08-09	1.12	111140.93	584.63	100961.46	368.95	9225.89	9116.34	109.55	0.0	conventional	2015	Albany
22	21	2015-08-02	1.45	75133.1	589.94	62035.06	741.08	11847.02	11768.52	78.5	0.0	conventional	2015	Albany
23	22	2015-07-26	1.11	106757.1	648.75	91949.05	966.61	13192.69	13061.53	131.16	0.0	conventional	2015	Albany
24	23	2015-07-19	1.26	96617.0	1042.1	82049.4	2238.02	11287.48	11103.49	183.99	0.0	conventional	2015	Albany
25	24	2015-07-12	1.05	124055.31	672.25	94693.52	4257.64	24431.9	24290.08	108.49	33.33	conventional	2015	Albany
26	25	2015-07-05	1.35	109252.12	869.45	72600.55	5883.16	29898.96	29663.19	235.77	0.0	conventional	2015	Albany
27	26	2015-06-28	1.37	89534.81	664.23	57545.79	4662.71	26662.88	26311.76	350.32	0.0	conventional	2015	Albany
28	27	2015-06-21	1.27	104849.39	804.01	76688.55	5481.18	21875.65	21662.0	213.65	0.0	conventional	2015	Albany
29	28	2015-06-14	1.32	89631.3	850.58	55400.94	4377.19	29002.59	28343.14	659.45	0.0	conventional	2015	Albany
30	29	2015-06-07	1.07	122743.86	656.71	99220.82	98.32	22775.21	22314.99	460.22	0.0	conventional	2015	Albany
31	30	2015-05-31	1.23	95123.62	922.37	70469.69	58.55	23681.01	23222.49	458.52	0.0	conventional	2015	Albany
32	31	2015-05-24	1.19	101470.91	680.27	71376.81	58.7	29355.13	28761.81	593.32	0.0	conventional	2015	Albany
33	32	2015-05-17	1.43	109857.47	1150.55	81955.16	94.32	26657.44	26285.43	372.01	0.0	conventional	2015	Albany
34	33	2015-05-10	1.26	120427.91	1420.43	102000.52	185.66	16821.3	16535.55	285.75	0.0	conventional	2015	Albany
35	34	2015-05-03	1.2	59197.67	919.87	45490.05	217.24	12570.51	12201.95	368.56	0.0	conventional	2015	Albany
36	35	2015-04-26	1.22	49585.46	875.65	35841.75	89.62	12778.44	12076.83	701.61	0.0	conventional	2015	Albany
37	36	2015-04-19	1.19	49064.73	774.15	33941.51	47.15	14301.92	13602.97	698.95	0.0	conventional	2015	Albany
38	37	2015-04-12	1.13	48364.29	864.27	30374.15	21.5	17104.37	16438.49	665.88	0.0	conventional	2015	Albany
39	38	2015-04-05	1.16	47362.13	961.77	35577.66	93.76	10728.94	9869.16	755.61	104.17	conventional	2015	Albany
40	39	2015-03-29	1.02	67799.08	1402.28	58623.22	89.5	7684.08	7208.49	475.59	0.0	conventional	2015	Albany
41	40	2015-03-22	1.12	46346.85	2141.83	34313.56	141.8	9749.66	9252.6	497.06	0.0	conventional	2015	Albany
42	41	2015-03-15	1.11	43845.79	2128.26	30447.17	99.67	10370.69	9989.59	381.1	0.0	conventional	2015	Albany

Avocado prices dataset (https://www.kaggle.com/neuromusic/avocado-prices)

Attributes

Objects

avocado.csv (1.9 MB)

	#	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
	0 – 52	3Jan15 – 24Mar18	0.44 – 3.25	84.6 – 62.5m	0 – 22.7m	0 – 20.5m	0 – 2.55m	0 – 19.4m	0 – 13.4m	0 – 5.72m	0 – 552k	conventional 50% organic 50% Other (1) 0%	2.02k – 2.02k	Albany 2% Atlanta 2% Other (53) 96%
1	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
2	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
3	2	2015-12-13	0.93	118220.22	794.7	109149.67	130.5	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
4	3	2015-12-06	1.08	78992.15	1132.0	71976.41	72.58	5811.16	5677.4	133.76	0.0	conventional	2015	Albany
5	4	2015-11-29	1.28	51039.6	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
6	5	2015-11-22	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0.0	conventional	2015	Albany
7	6	2015-11-15	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0.0	conventional	2015	Albany
8	7	2015-11-08	0.98	109428.33	703.75	101815.36	80.0	6829.22	6266.85	562.37	0.0	conventional	2015	Albany
9	8	2015-11-01	1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0.0	conventional	2015	Albany
10	9	2015-10-25	1.07	74338.76	842.4	64757.44	113.0	8625.92	8061.47	564.45	0.0	conventional	2015	Albany
11	10	2015-10-18	1.12	84843.44	924.86	75595.85	117.07	8285.66	7877.86	327.8	0.0	conventional	2015	Albany
12	11	2015-10-11	1.28	64489.17	1582.03	52677.92	105.32	10123.9	9866.27	257.63	0.0	conventional	2015	Albany
13	12	2015-10-04	1.31	61007.1	2268.32	49880.67	101.36	8756.75	8379.98	376.77	0.0	conventional	2015	Albany
14	13	2015-09-27	0.99	106803.39	1204.88	99409.21	154.84	6034.46	5888.87	145.59	0.0	conventional	2015	Albany
15	14	2015-09-20	1.33	69759.01	1028.03	59313.12	150.5	9267.36	8489.1	778.26	0.0	conventional	2015	Albany
16	15	2015-09-13	1.28	76111.27	985.73	65696.86	142.0	9286.68	8665.19	621.49	0.0	conventional	2015	Albany
17	16	2015-09-06	1.11	99172.96	879.45	90062.62	240.79	7990.1	7762.87	227.23	0.0	conventional	2015	Albany
18	17	2015-08-30	1.07	105693.84	689.01	94362.67	335.43	10306.73	10218.93	87.8	0.0	conventional	2015	Albany
19	18	2015-08-23	1.34	79992.09	733.16	67933.79	444.78	10880.36	10745.79	134.57	0.0	conventional	2015	Albany
20	19	2015-08-16	1.33	80043.78	539.65	68666.01	394.9	10443.22	10297.68	145.54	0.0	conventional	2015	Albany
21	20	2015-08-09	1.12	111140.93	584.63	100961.46	368.95	9225.89	9116.34	109.55	0.0	conventional	2015	Albany
22	21	2015-08-02	1.45	75133.1	589.94	62035.06	741.08	11847.02	11768.52	78.5	0.0	conventional	2015	Albany
23	22	2015-07-26	1.11	106757.1	648.75	91949.05	966.61	13192.69	13061.53	131.16	0.0	conventional	2015	Albany
24	23	2015-07-19	1.26	96617.0	1042.1	82049.4	2238.02	11287.48	11103.49	183.99	0.0	conventional	2015	Albany
25	24	2015-07-12	1.05	124055.31	672.25	94693.52	4257.64	24431.9	24290.08	108.49	33.33	conventional	2015	Albany
26	25	2015-07-05	1.35	109252.12	869.45	72600.55	5883.16	29898.96	29663.19	235.77	0.0	conventional	2015	Albany
27	26	2015-06-28	1.37	89534.81	664.23	57545.79	4662.71	26662.88	26311.76	350.32	0.0	conventional	2015	Albany
28	27	2015-06-21	1.27	104849.39	804.01	76688.55	5481.18	21875.65	21662.0	213.65	0.0	conventional	2015	Albany
29	28	2015-06-14	1.32	89631.3	850.58	55400.94	4377.19	29002.59	28343.14	659.45	0.0	conventional	2015	Albany
30	29	2015-06-07	1.07	122743.86	656.71	99220.82	98.32	22775.21	22314.99	460.22	0.0	conventional	2015	Albany
31	30	2015-05-31	1.23	95123.62	922.37	70469.69	58.55	23681.01	23222.49	458.52	0.0	conventional	2015	Albany
32	31	2015-05-24	1.19	101470.91	680.27	71376.81	58.7	29355.13	28761.81	593.32	0.0	conventional	2015	Albany
33	32	2015-05-17	1.43	109857.47	1150.55	81955.16	94.32	26657.44	26285.43	372.01	0.0	conventional	2015	Albany
34	33	2015-05-10	1.26	120427.91	1420.43	102000.52	185.66	16821.3	16535.55	285.75	0.0	conventional	2015	Albany
35	34	2015-05-03	1.2	59197.67	919.87	45490.05	217.24	12570.51	12201.95	368.56	0.0	conventional	2015	Albany
36	35	2015-04-26	1.22	49585.46	875.65	35841.75	89.62	12778.44	12076.83	701.61	0.0	conventional	2015	Albany
37	36	2015-04-19	1.19	49064.73	774.15	33941.51	47.15	14301.92	13602.97	698.95	0.0	conventional	2015	Albany
38	37	2015-04-12	1.13	48364.29	864.27	30374.15	21.5	17104.37	16438.49	665.88	0.0	conventional	2015	Albany
39	38	2015-04-05	1.16	47362.13	961.77	35577.66	93.76	10728.94	9869.16	755.61	104.17	conventional	2015	Albany
40	39	2015-03-29	1.02	67799.88	1402.28	58623.22	89.5	7684.08	7208.49	475.59	0.0	conventional	2015	Albany
41	40	2015-03-22	1.12	46346.85	2141.83	34313.56	141.8	9749.66	9252.6	497.06	0.0	conventional	2015	Albany
42	41	2015-03-15	1.11	43845.79	2128.26	30447.17	99.67	10370.69	9989.59	381.1	0.0	conventional	2015	Albany

Avocado prices dataset (https://www.kaggle.com/neuromusic/avocado-prices)

Set, Sequence, & Space


Ordered
Set, Sequence, & Space
Unordered
Geometric-
Structured

Unordered Data

- A set of attributes
- Ordering of attributes doesn’t really matter {a, b} = {b, a}

- Examples
  - Documents, tweets, web contents
  - Demographic data
  - Employee records
  - Student records
  - Bank transaction records
  - Product inventory
  - ...

Unordered Data

- e.g. Student record table

qryStudentInformation : Select Query

strStudentID	strFirstName	strLastName	strAddress1	strPostCode
BJ25555	LEE	FORD	28 SIMPSON CLOSE	L21 5ER
BJ25567	JANE	PIPER	8 LUBERNUM ROAD	CH45 5RT
BJ26667	ADELE	CUSHING	12 SEDDEN ROAD	L19 2LJ
BJ26887	MATTHEW	FOGGERTY	15 WOODMEDOW COURT	0L5 0BQ
BJ29991	DARREN	PHILLIPS	101 HERSCHELL STREET	L5 1XE
BJ34684	IAN	FORSTER	20 ALT ROAD	L20 5ES
BJ45126	PAUL	LAWSON	29 BROAD LANE	L11 8LY
CY21147	PENNY	HARES	75 WIDCOMBE RD	SE9 4HY

Record: 12 of 52

http://www.databasedev.co.uk/delete_query.html

Unordered Data

- e.g. Documents (bag of words)

Article ID	biolog	biopsi	biolab	biotin	almost	cancer-surviv	cancer-stage	Article Class
00001	12	1	2	10	0	1	4	breast-cancer
00002	10	1	0	3	0	6	1	breast-cancer
00014	4	1	1	1	0	28	0	breast-cancer
00063	4	0	0	0	0	18	7	breast-cancer
00319	0	1	0	9	0	20	1	breast-cancer
00847	7	2	0	14	0	11	5	breast-cancer
03042	3	1	3	1	0	19	8	lung-cancer
05267	4	4	2	6	0	14	11	lung-cancer
05970	8	0	4	9	0	9	17	lung-cancer
30261	1	0	0	11	0	21	1	prostate-cancer
41191	9	0	5	14	0	11	1	prostate-cancer
52038	6	1	1	17	0	19	0	prostate-cancer
73851	1	1	8	17	0	17	3	prostate-cancer

Ye et al. (2016) - https://doi.org/10.1371/journal.pone.0162721.t001

Ordered Data

- Ordered set of attributes
- Ordering matters! (a, b) != (b, a)

- Examples
  - Time series
  - Sequence
  - ...

Ordered Data

- e.g. Genetic Sequence

https://en.wikipedia.org/wiki/Nucleic_acid_sequence#/media/File:AMY1gene.png

Ordered Data

- e.g. Stock price (candlestick chart)

https://www.investopedia.com/trading/candlestick-charting-what-is-it/

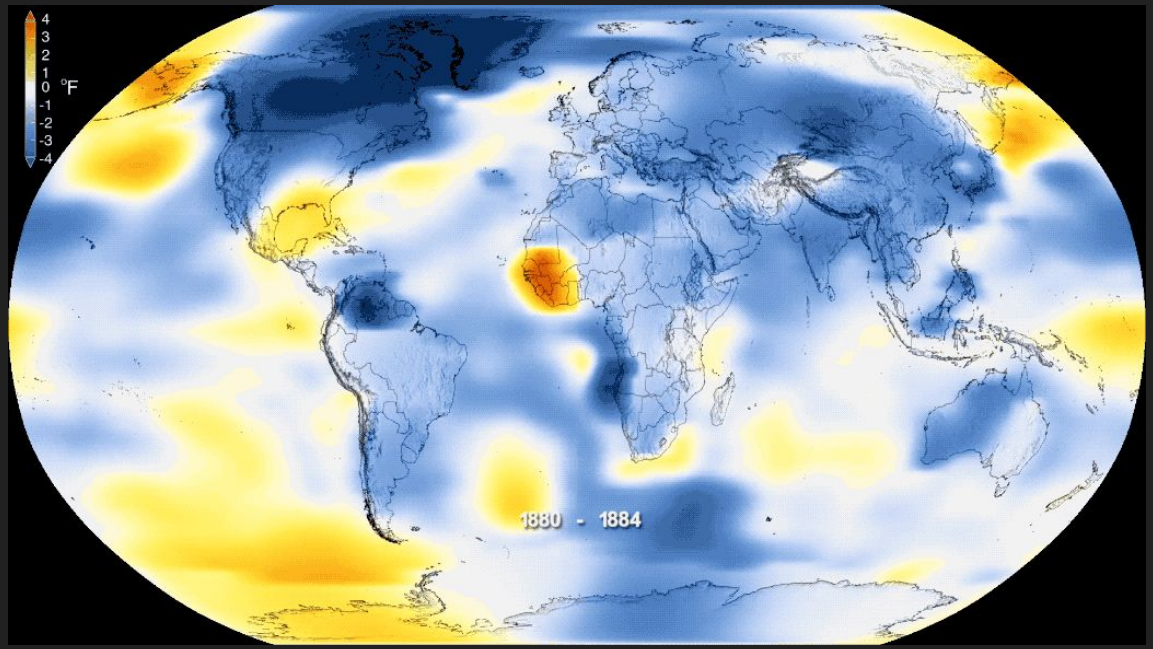
Geometric/Structured Data

- Data sets that have geometric/topological/geographical structures.
- Spatial location of an object comes into play.

- Example
  - Spatio-temporal data
  - Image pixels, points in LiDAR, computer graphics models
  - Graph data

Geometric/Structured Data

- e.g. Spatio-temporal data



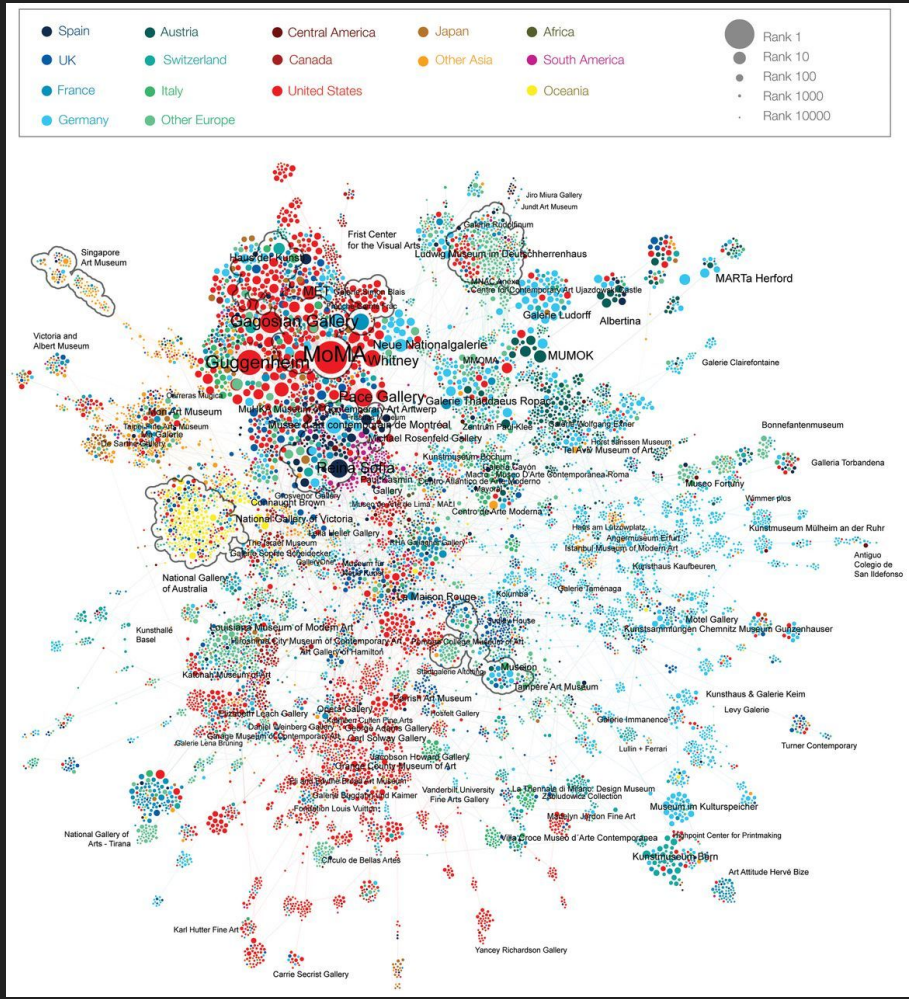
https://climate.nasa.gov/news/2876/new-studies-increase-confidence-in-nasas-measure-of-earths-temperature/

Geometric/Structured Data

- e.g. Graph data (social network)

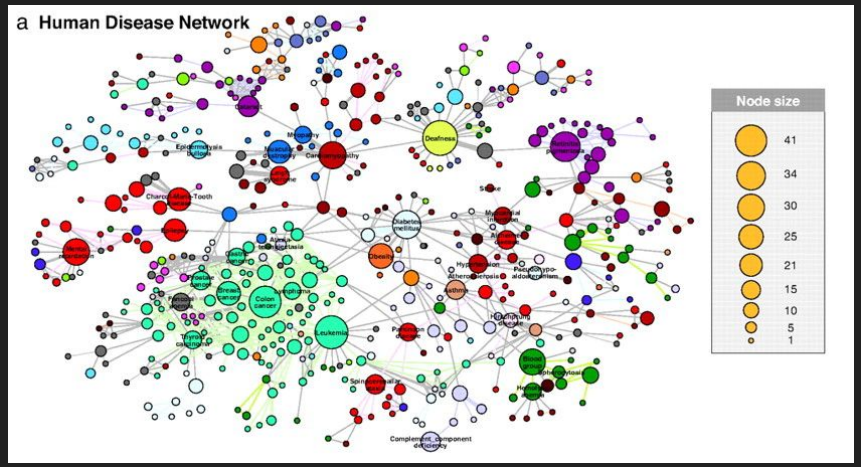
“Quantifying reputation and success in art”
Fraiberger et al. (Science, 2018)

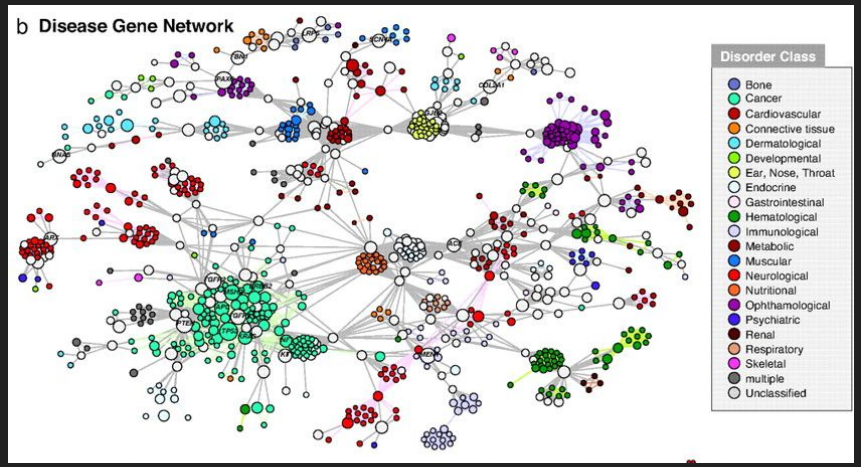
Force-directed layout of the order $\tau = \infty$ coexhibition network, whose nodes are institutions (galleries, museums). Node size is proportional to each institution’s eigenvector centrality. Nodes are connected if they both exhibited the same artist, with link weights being equal to the number of artists’ coexhibitions. Node colors encode the region in which institutions are located. Links are of the same colors as their end nodes, or gray when end nodes have different colors. For visualization purposes, we only show the 12,238 nodes corresponding to institutions with more than 10 exhibits; we pruned the links by keeping the most statistically significant links (20) (supplementary text S2.2). We implemented community detection on the pruned network (21), identifying 122 communities (supplementary text S2.3). We highlighted five of them, the full community breakdown being shown in fig. S3. We also show the names of the most prestigious institution for each community.



Geometric/Structured Data

- e.g. Graph data (disease network)

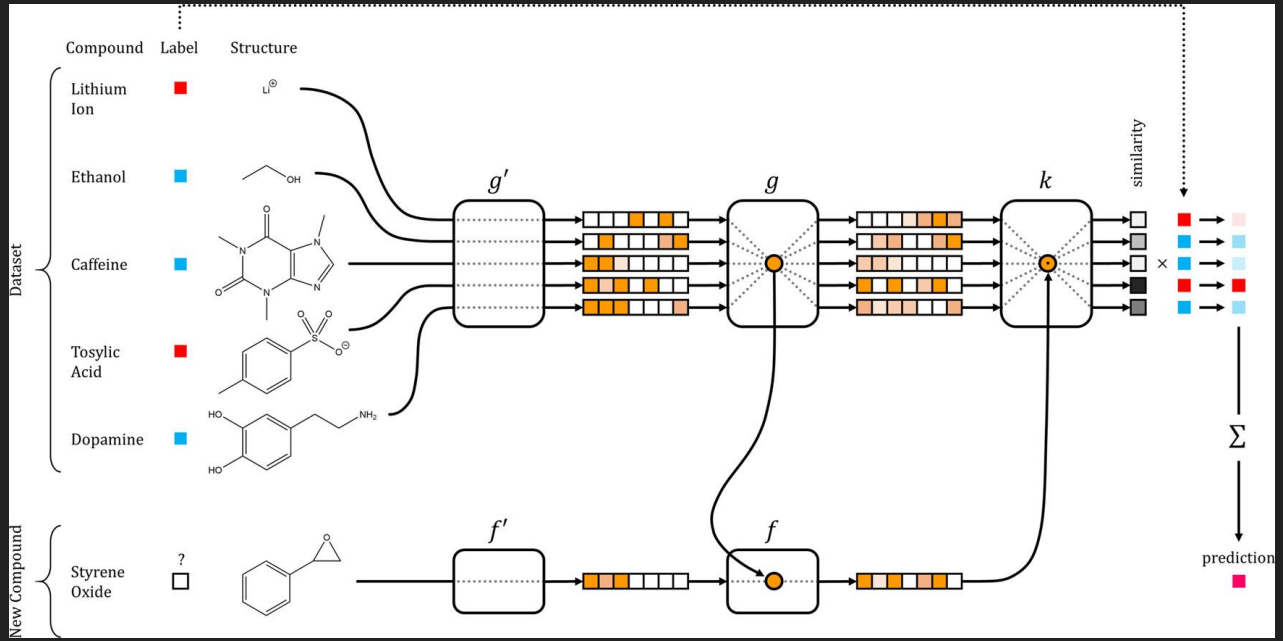




"The Human Disease Network"
Goh et al. (PNAS, 2007)

Geometric/Structured Data

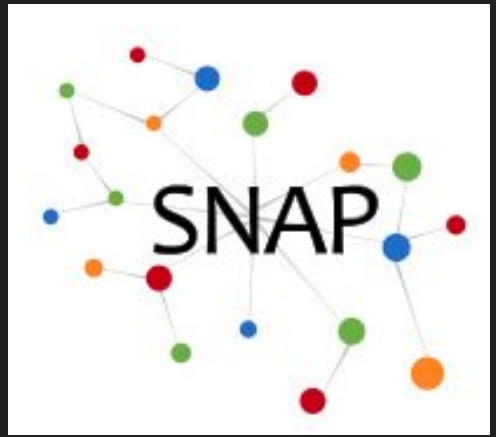
- e.g. Graph data (chemical compounds)



"Low Data Drug Discovery with One-Shot Learning"
Altae-Tran et al. (ACS Central Science, 2017)

Geometric/Structured Data

- e.g. Graph data (Stanford Large Network Dataset Collection)
  - http://snap.stanford.edu/
  - Social Networks
  - Communication Networks
  - Citation Networks
  - Collaboration Networks
  - Road Networks
  - Temporal Networks
  - ...

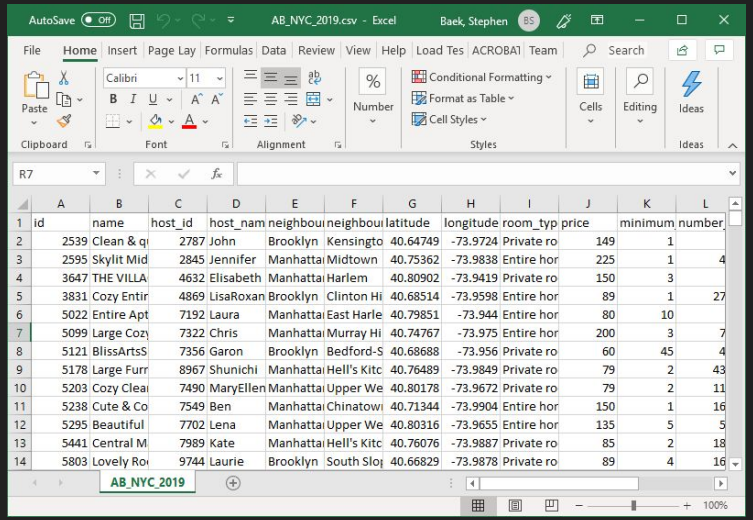


Data Formats

- There are MANY data representations in computers!
  - Comma Separated Values (CSV)
  - JavaScript Object Notation (JSON)
  - eXtensible Markup Language (XML)
  - ...

Data Formats

- Comma Separated Values (CSV)
  - Delimited text file that uses comma (,) to separate values
  - Some ~~weird~~ people use something else, like semicolon (;), instead
  - Tabular data is stored in plain text → large file size

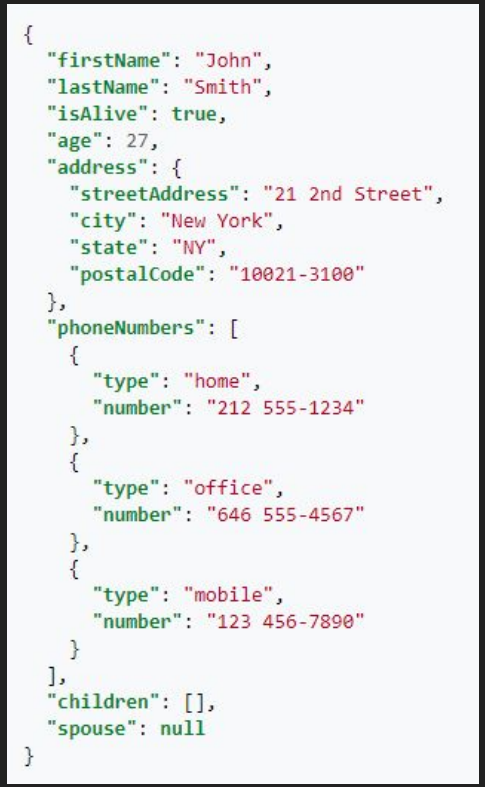


	A	B	C	D	E	F	G	H	I	J	K	L
1	id	name	host_id	host_nam	neighbou	neighbou	latitude	longitude	room_typ	price	minimum	number
2	2539	Clean & q	2787	John	Brooklyn	Kensingto	40.64749	-73.9724	Private ro	149	1	
3	2595	Skylit Mid	2845	Jennifer	Manhattai	Midtown	40.75362	-73.9838	Entire hor	225	1	4
4	3647	THE VILLA	4632	Elisabeth	Manhattai	Harlem	40.80902	-73.9419	Private ro	150	3	
5	3831	Cozy Entir	4869	LisaRoxan	Brooklyn	Clinton Hi	40.68514	-73.9598	Entire hor	89	1	27
6	5022	Entire Apt	7192	Laura	Manhattai	East Harle	40.79851	-73.944	Entire hor	80	10	
7	5099	Large Cozy	7322	Chris	Manhattai	Murray Hi	40.74767	-73.975	Entire hor	200	3	7
8	5121	BlissArtsS	7356	Garon	Brooklyn	Bedford-S	40.68688	-73.956	Private ro	60	45	4
9	5178	Large Furr	8967	Shunichi	Manhattai	Hell's Kitc	40.76489	-73.9849	Private ro	79	2	43
10	5203	Cozy Clear	7490	MaryEllen	Manhattai	Upper We	40.80178	-73.9672	Private ro	79	2	11
11	5238	Cute & Co	7549	Ben	Manhattai	Chinatow	40.71344	-73.9904	Entire hor	150	1	16
12	5295	Beautiful	7702	Lena	Manhattai	Upper We	40.80316	-73.9655	Entire hor	135	5	5
13	5441	Central M	7989	Kate	Manhattai	Hell's Kitc	40.76076	-73.9887	Private ro	85	2	18
14	5803	Lovely Roo	9744	Laurie	Brooklyn	South Slop	40.66829	-73.9878	Private ro	89	4	16

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Data Formats

- JavaScript Object Notation (JSON)
  - Easy for humans to read and write, easy for machines to parse and generate
  - Attribute-value pairs + arrays
  - Good for structured data



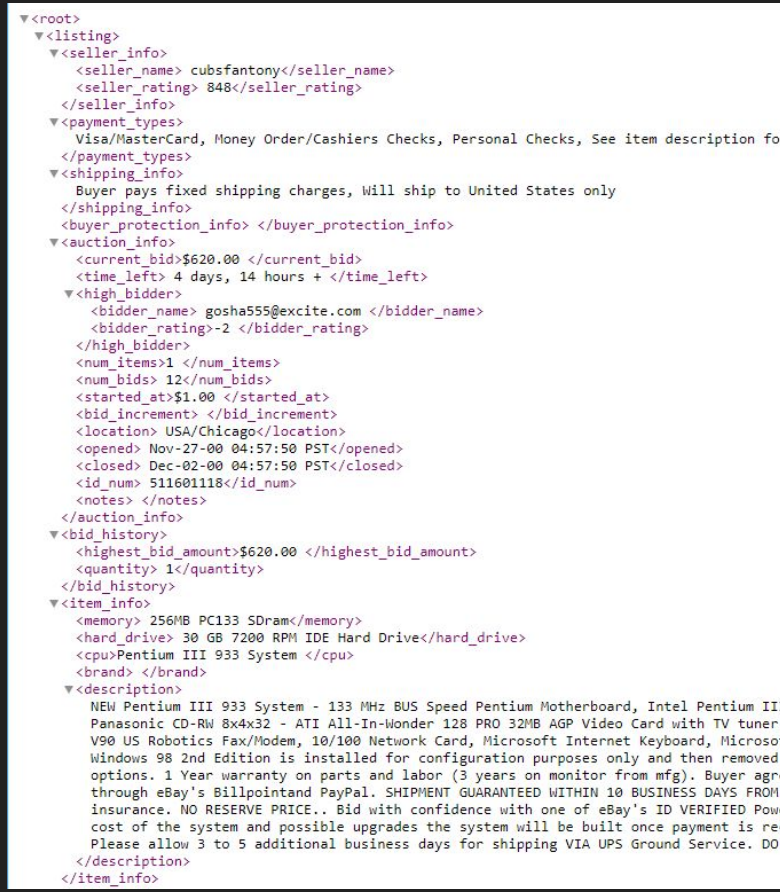
```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 27,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

Wikipedia

Data Formats

- eXtensible Markup Language (XML)
  - Easy for humans, easy for machines
  - Made of tags
    - Start-tag: <tagname>
    - End-tag: </tagname>
    - Empty-element-tag: <tagname />
  - Attributes
    - Name-value pair that exists in a tag.
    - For example,

      - Tag: img (an empty tag)
      - Attributes: src, alt



*In-class Assignment

(ICA) Let's Play with Examples!

- Python & Pandas

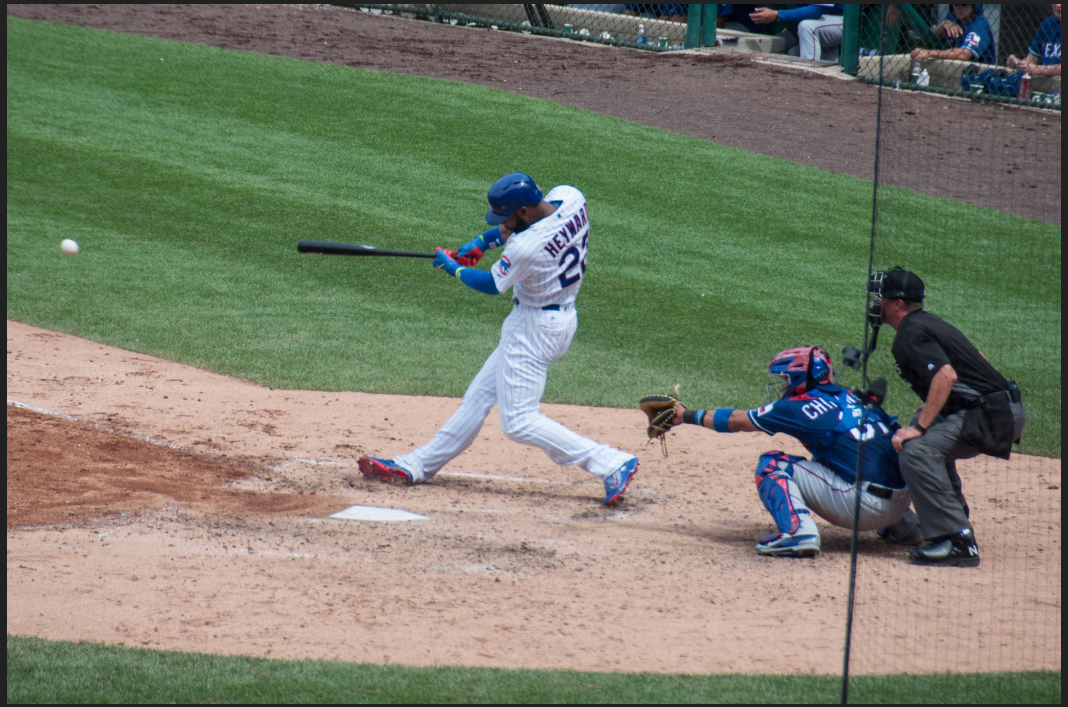


Image Source: Wikipedia