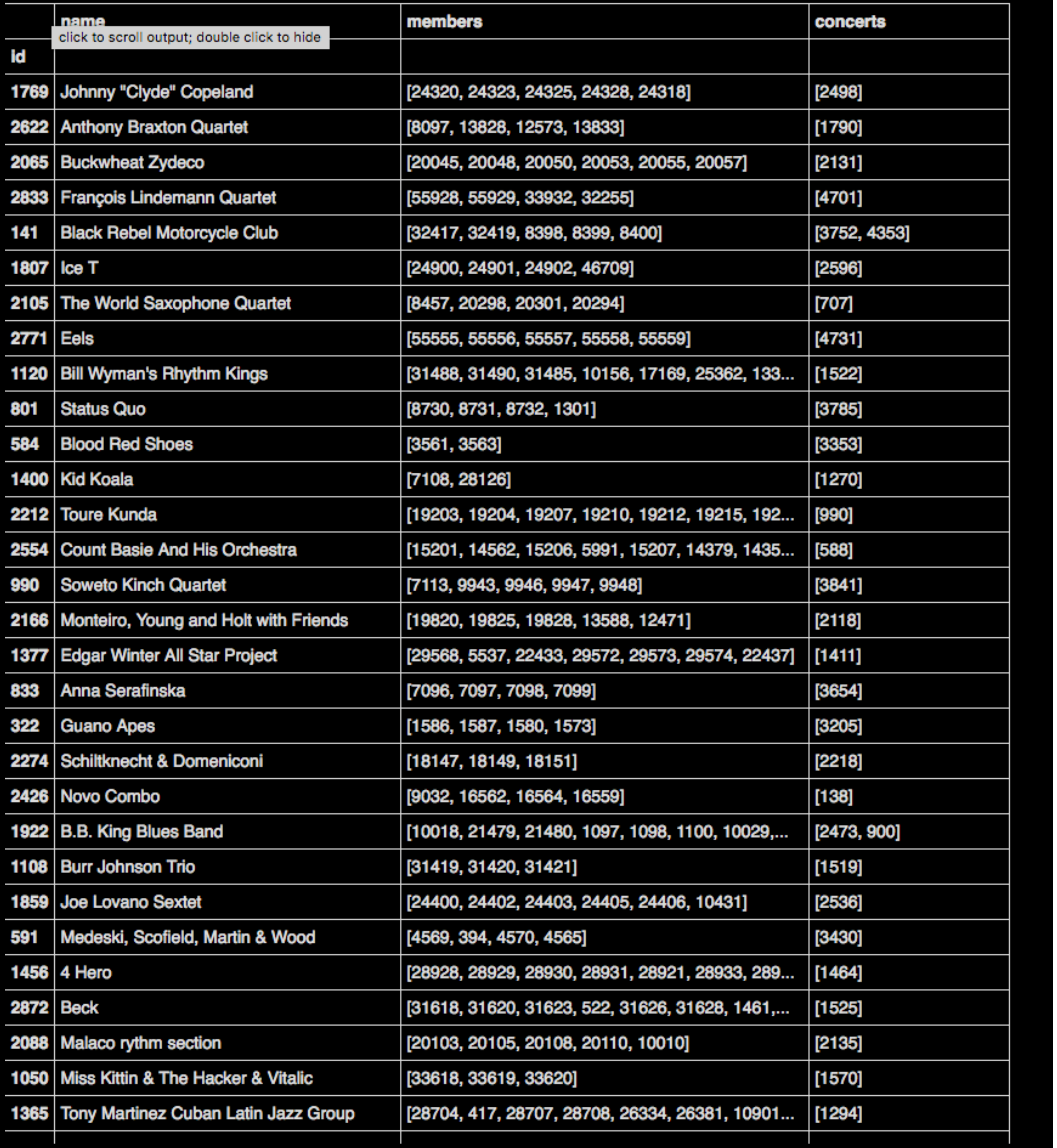
DATA

KIRELL BENZI, PH.D

@KirellBenzi

www.kirellbenzi.com



	name	members	concerts
id			
1769	Johnny "Clyde" Copeland	[24320, 24323, 24325, 24328, 24318]	[2498]
2622	Anthony Braxton Quartet	[8097, 13828, 12573, 13833]	[1790]
2065	Buckwheat Zydeco	[20045, 20048, 20050, 20053, 20055, 20057]	[2131]
2833	François Lindemann Quartet	[55928, 55929, 33932, 32255]	[4701]
141	Black Rebel Motorcycle Club	[32417, 32419, 8398, 8399, 8400]	[3752, 4353]
1807	Ice T	[24900, 24901, 24902, 46709]	[2596]
2105	The World Saxophone Quartet	[8457, 20298, 20301, 20294]	[707]
2771	Eels	[55555, 55556, 55557, 55558, 55559]	[4731]
1120	Bill Wyman's Rhythm Kings	[31488, 31490, 31485, 10156, 17169, 25362, 133...	[1522]
801	Status Quo	[8730, 8731, 8732, 1301]	[3785]
584	Blood Red Shoes	[3561, 3563]	[3353]
1400	Kid Koala	[7108, 28126]	[1270]
2212	Toure Kunda	[19203, 19204, 19207, 19210, 19212, 19215, 192...	[990]
2554	Count Basie And His Orchestra	[15201, 14562, 15206, 5991, 15207, 14379, 1435...	[588]
990	Soweto Kinch Quartet	[7113, 9943, 9946, 9947, 9948]	[3841]
2166	Monteiro, Young and Holt with Friends	[19820, 19825, 19828, 13588, 12471]	[2118]
1377	Edgar Winter All Star Project	[29568, 5537, 22433, 29572, 29573, 29574, 22437]	[1411]
833	Anna Serafinska	[7096, 7097, 7098, 7099]	[3654]
322	Guano Apes	[1586, 1587, 1580, 1573]	[3205]
2274	Schiltknecht & Domeniconi	[18147, 18149, 18151]	[2218]
2426	Novo Combo	[9032, 16562, 16564, 16559]	[138]
1922	B.B. King Blues Band	[10018, 21479, 21480, 1097, 1098, 1100, 10029,...	[2473, 900]
1108	Burr Johnson Trio	[31419, 31420, 31421]	[1519]
1859	Joe Lovano Sextet	[24400, 24402, 24403, 24405, 24406, 10431]	[2536]
591	Medeski, Scofield, Martin & Wood	[4569, 394, 4570, 4565]	[3430]
1456	4 Hero	[28928, 28929, 28930, 28931, 28921, 28933, 289...	[1464]
2872	Beck	[31618, 31620, 31623, 522, 31626, 31628, 1461,...	[1525]
2088	Malaco rythm section	[20103, 20105, 20108, 20110, 10010]	[2135]
1050	Miss Kittin & The Hacker & Vitalic	[33618, 33619, 33620]	[1570]
1365	Tony Martinez Cuban Latin Jazz Group	[28704, 417, 28707, 28708, 26334, 26381, 10901...	[1294]

DATA

Set of values of qualitative or quantitative variables.

Data (singular is datum) require interpretation to be meaningful

Data abstraction

Data-viz are depend of the kind of data we have as input.

We need two ingredients to design a good viz:

Data type - its structural or mathematical interpretation

Data semantics - its real-world meaning

Visualization Analysis and Design Ch. 2

Data semantics

1	Bob	M	Blue
2	Alice	S	Red

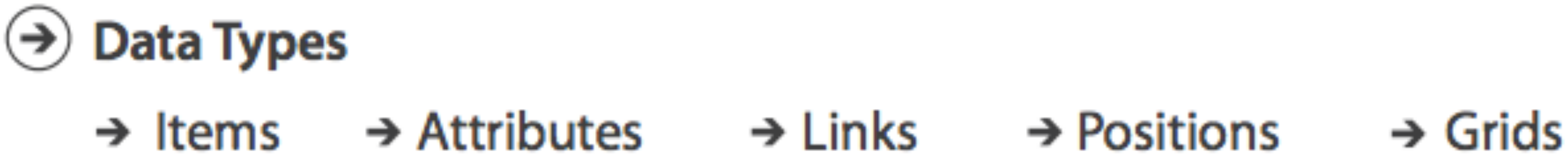
Metadata

ID	Name	Shirt Size	Favorite color
1	Bob	M	Blue
2	Alice	S	Red

Data types

Structural or mathematical interpretation of data

Different from data types in programming



➔ **Data Types**

→ Items → Attributes → Links → Positions → Grids

Items & attributes

Item: individual entity, collections of attributes.

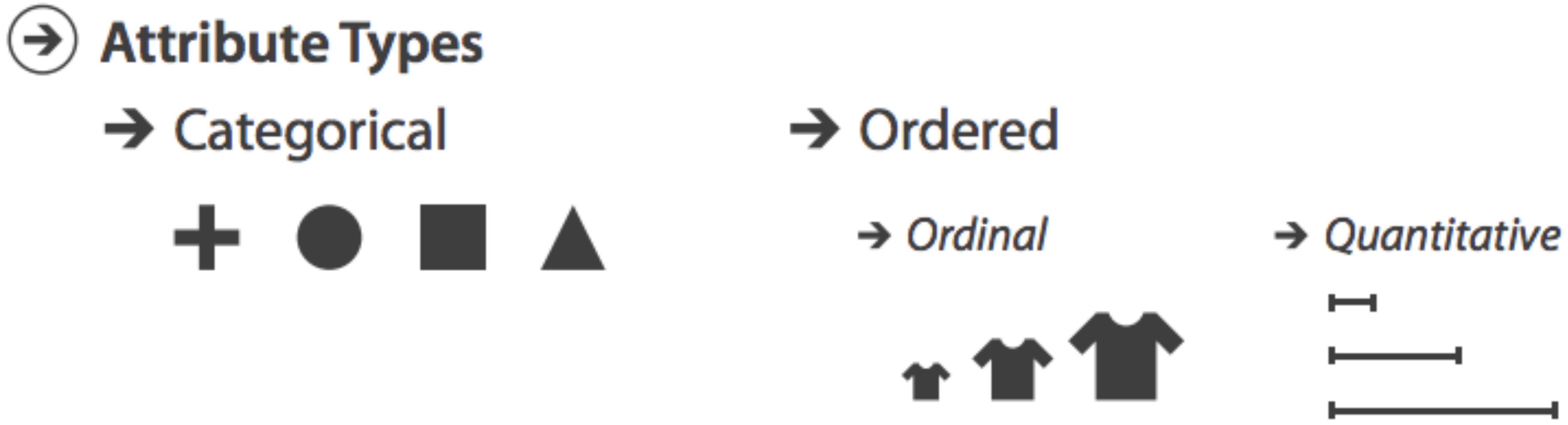
Attributes can be measured, observed and logged.

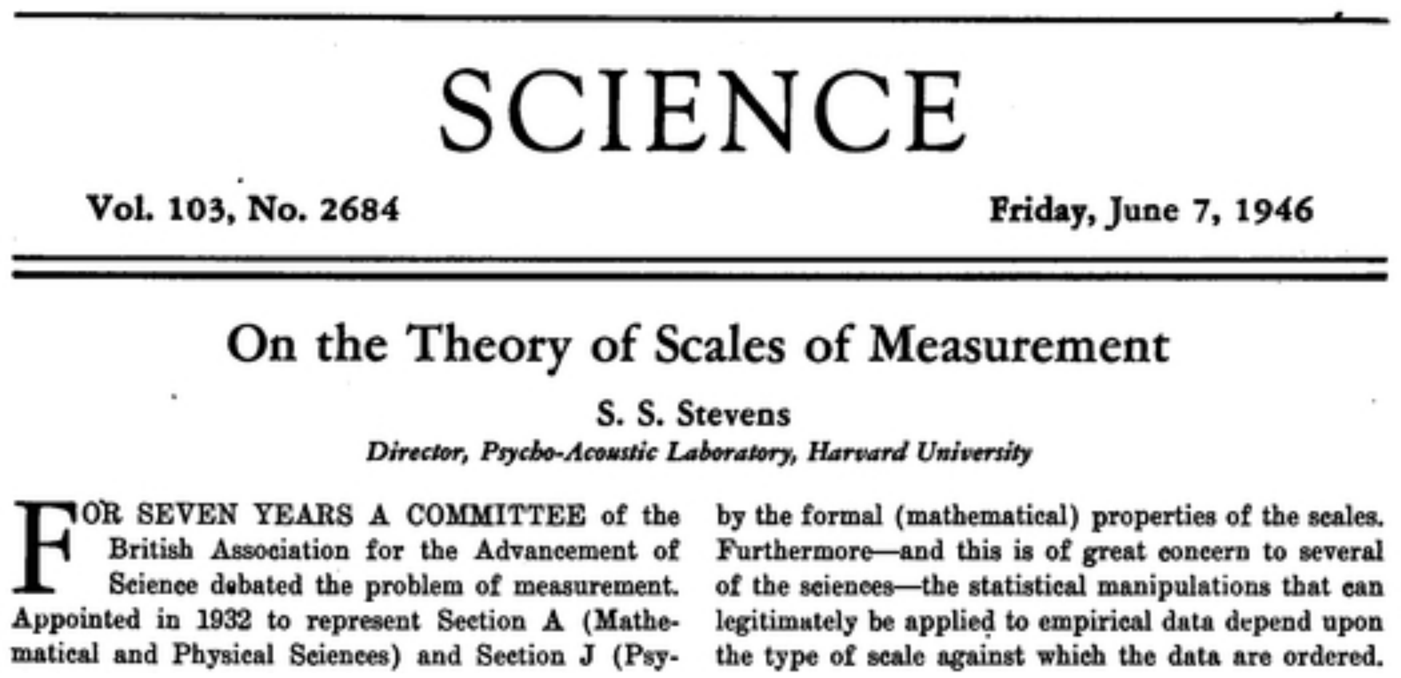
ID	Name	Shirt Size	Favorite color
1	Bob	M	Blue
2	Alice	S	Red

Item: Person

Attribute types

- What kind of measurements can we perform with attributes?
- Initiated by S.S Stevens in 1946




SCIENCE

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

On the Theory of Scales of Measurement

S. S. Stevens

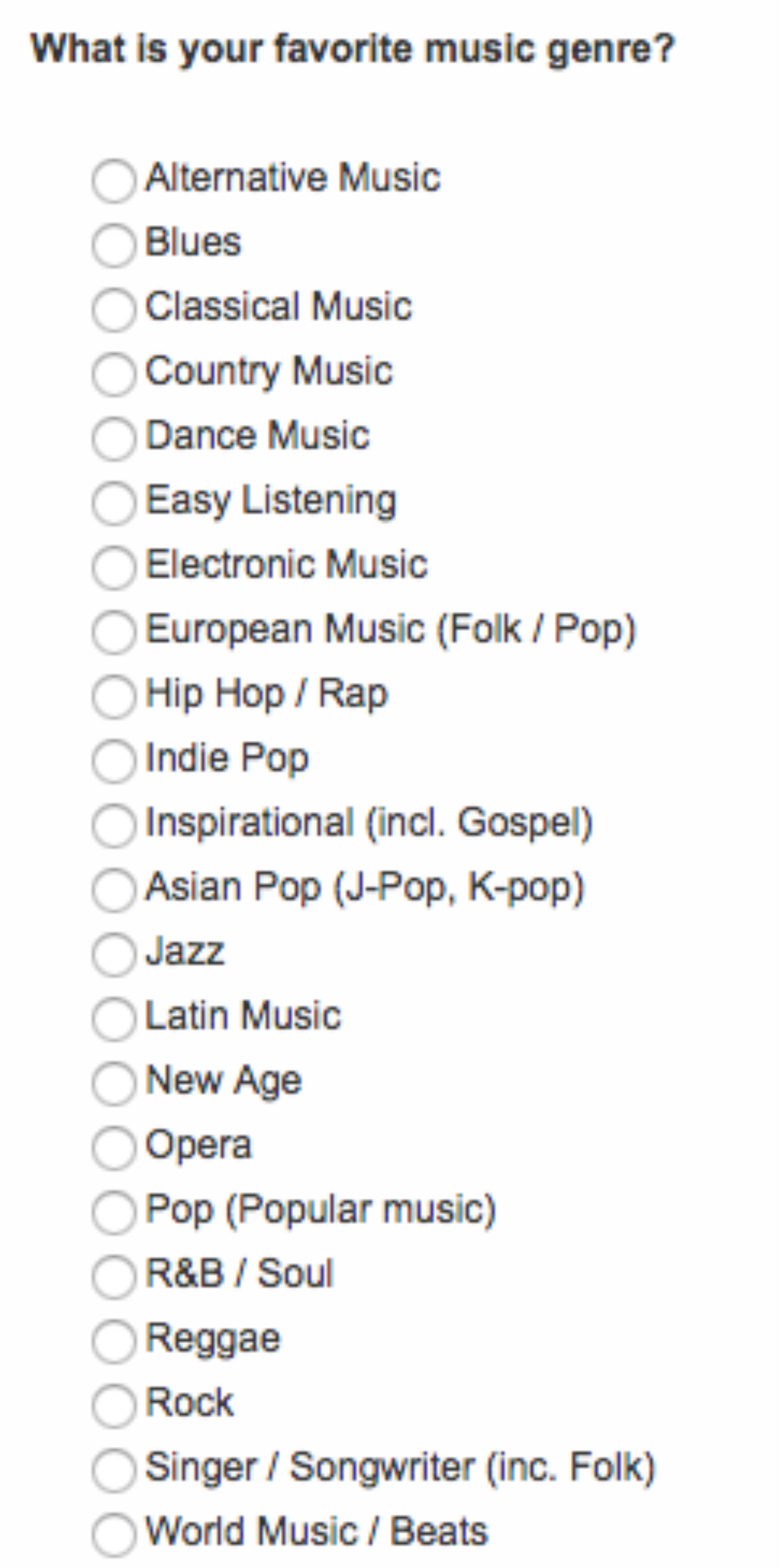
Director, Psycho-Acoustic Laboratory, Harvard University

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 (Psy-

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.

Categorial (Nominal) type

- Could be simply called labels
- No quantitative value
- List of choices in survey
- Classes in machine learning



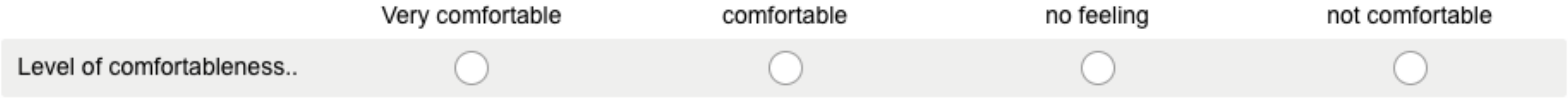
What is your favorite music genre?

○ Alternative Music
○ Blues
○ Classical Music
○ Country Music
○ Dance Music
○ Easy Listening
○ Electronic Music
○ European Music (Folk / Pop)
○ Hip Hop / Rap
○ Indie Pop
○ Inspirational (incl. Gospel)
○ Asian Pop (J-Pop, K-pop)
○ Jazz
○ Latin Music
○ New Age
○ Opera
○ Pop (Popular music)
○ R&B / Soul
○ Reggae
○ Rock
○ Singer / Songwriter (inc. Folk)
○ World Music / Beats

http://www.musicgenreslist.com/

Ordered: Ordinal type

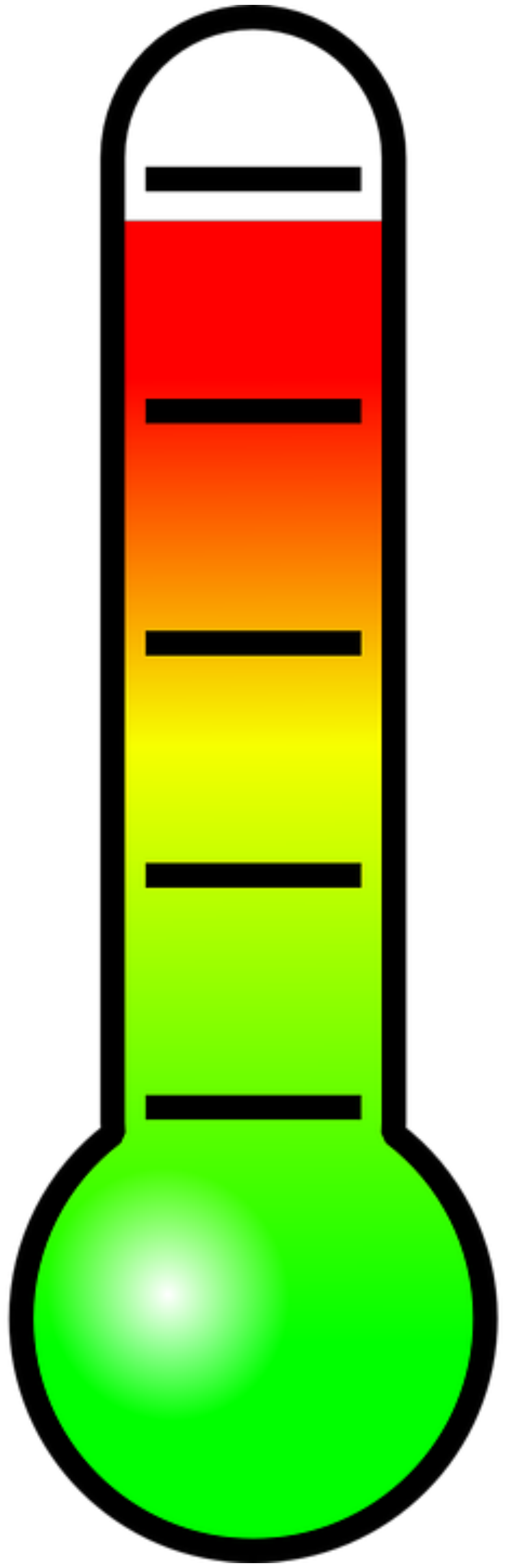
- The order is important <, >
- Difference between elements is not really known but we can rank them
- Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.



	Very comfortable	comfortable	no feeling	not comfortable
Level of comfortableness..	○	○	○	○

Quantitative: Interval

- Variables are classfied into ordered categories
- Direct measure of comparison between values
- Problem: no true zero



1985 Super Mario Bros.

1989 Super Mario Bros. 2

1991 Super Mario Bros. 3

1992 Super Mario World

1997 Super Mario 64

2002 Super Mario Sunshine

2006 New Super Mario Bros.

2007 Super Mario Galaxy

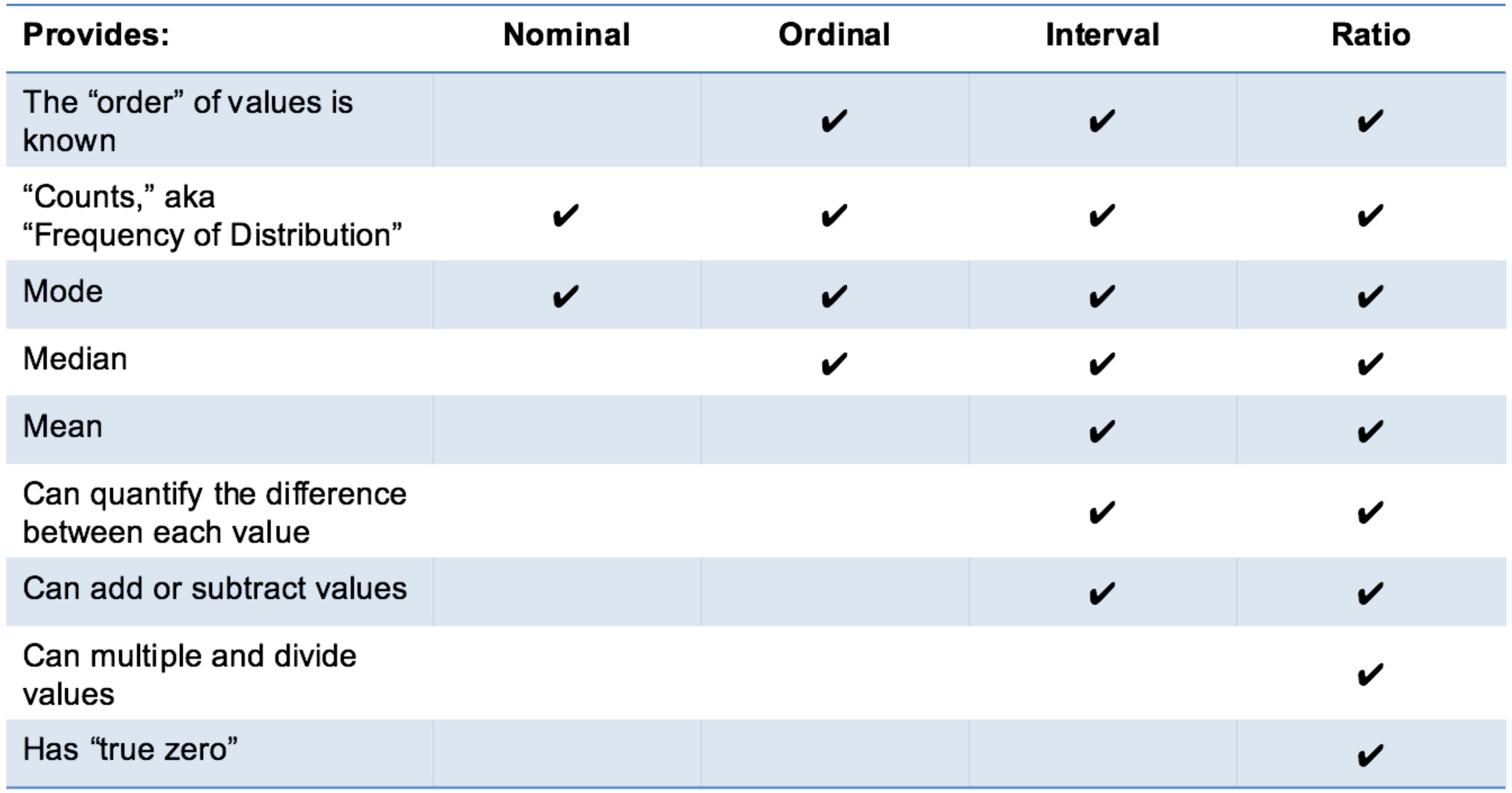
2009 New Super Mario Bros. Wii

2010 Super Mario Galaxy 2

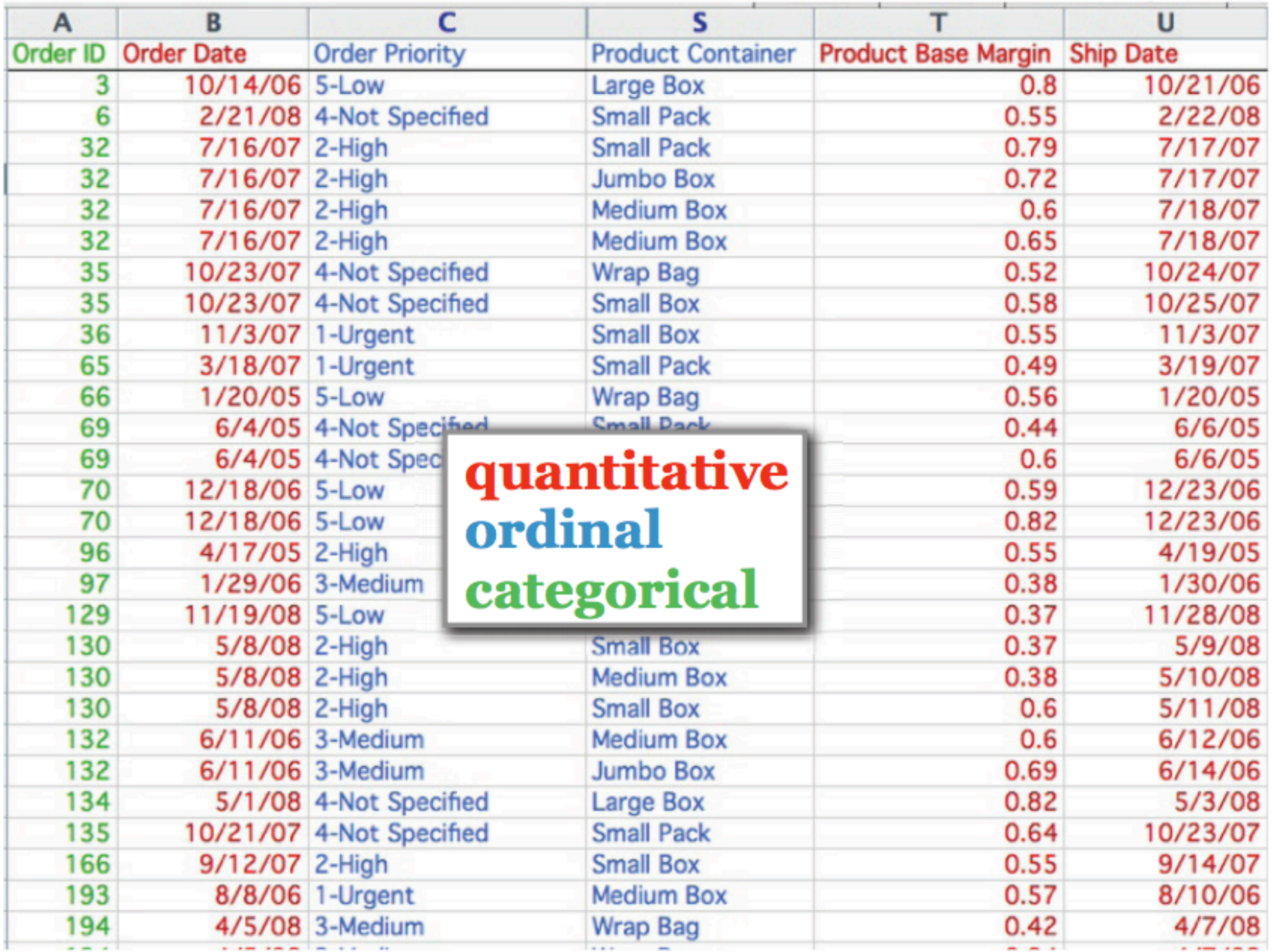
[IGN]

Quantitative: Ratio

- All characteristics of nominal, ordinal and interval variables
- Meaninful zero point
- add, subtract, divide and multiply two ratios
- Most useful type of variable for statistics
- Example: Height or weight



Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

quantitative
ordinal
categorical

Other data types

Links

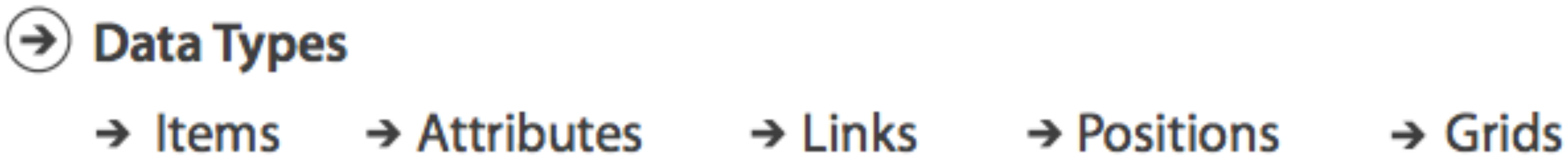
Express relationship between two items: friendship on Facebook, followers on Twitter

Positions

Location in 2D or 3D for spatial data: geolocalization of best restaurants

Grids

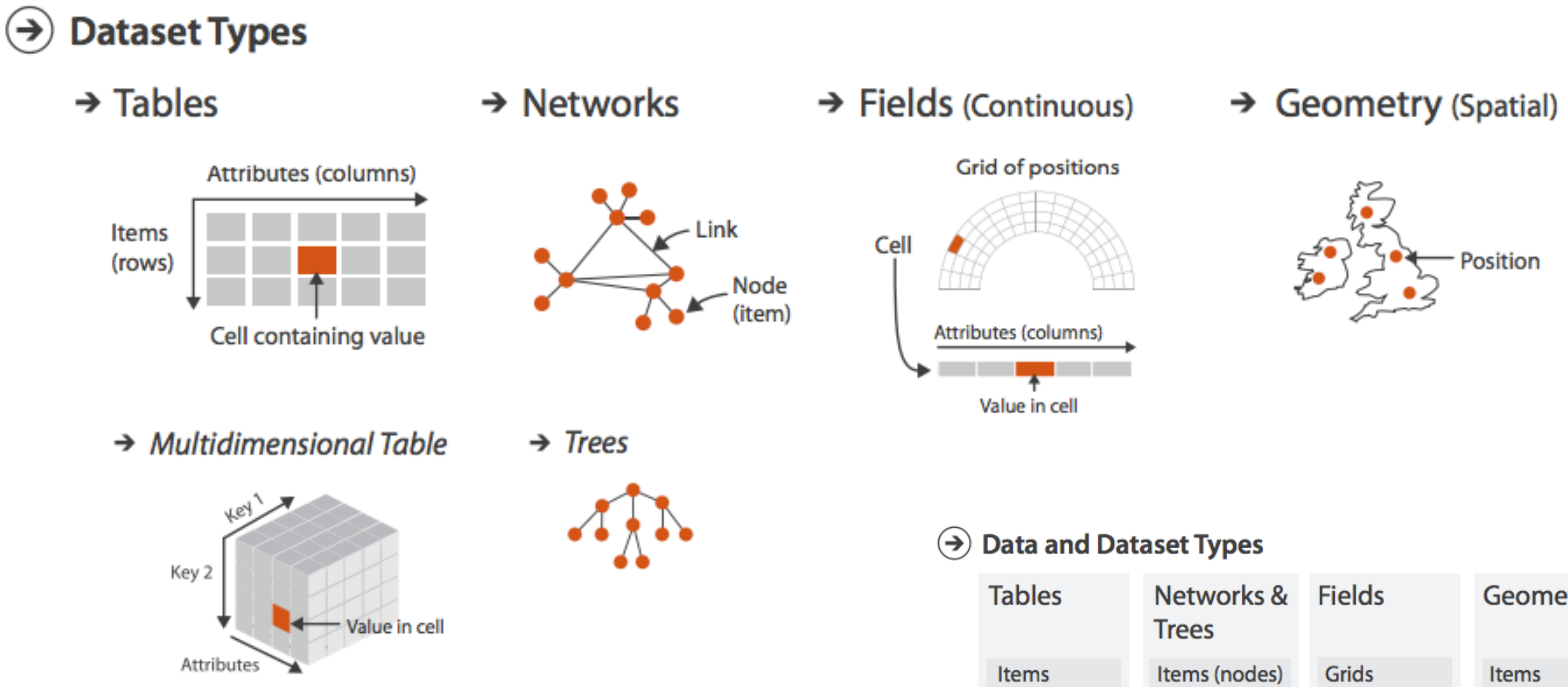
Sampling strategy for continuous data: voxels in MRI scan, sensors in a city



Data Types

→ Items → Attributes → Links → Positions → Grids

Dataset types



Dataset Types

→ Tables

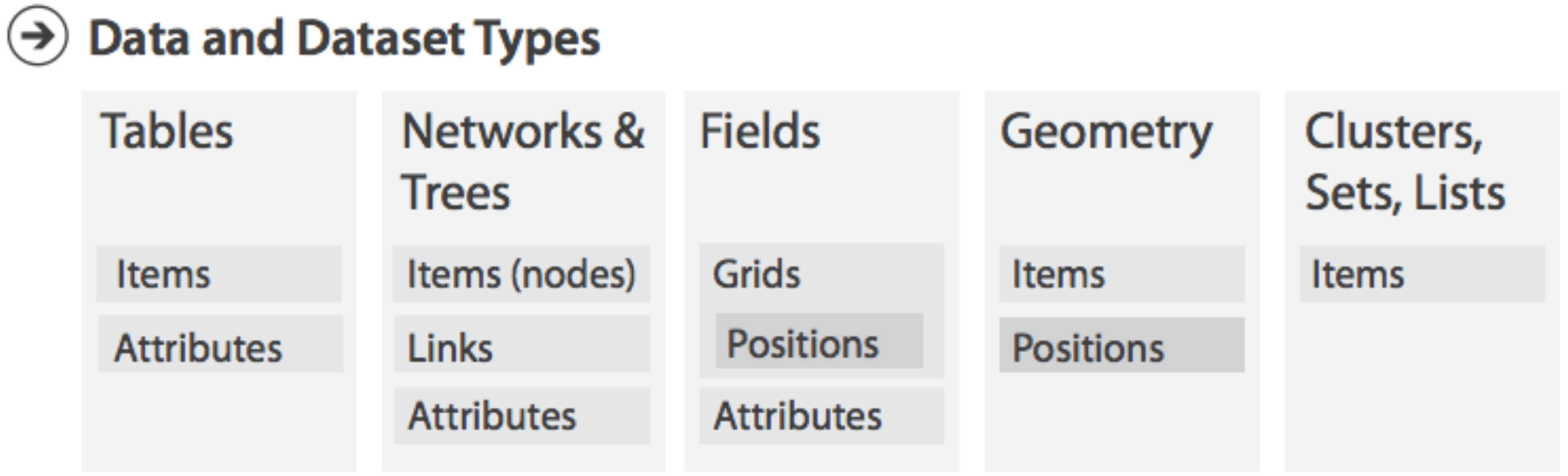
→ Networks

→ Fields (Continuous)

→ Geometry (Spatial)

→ *Multidimensional Table*

→ *Trees*



Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Tables

Flat Table

one item per row

attributes are stored in columns

Mental image of Relational database (e.g. MySQL)

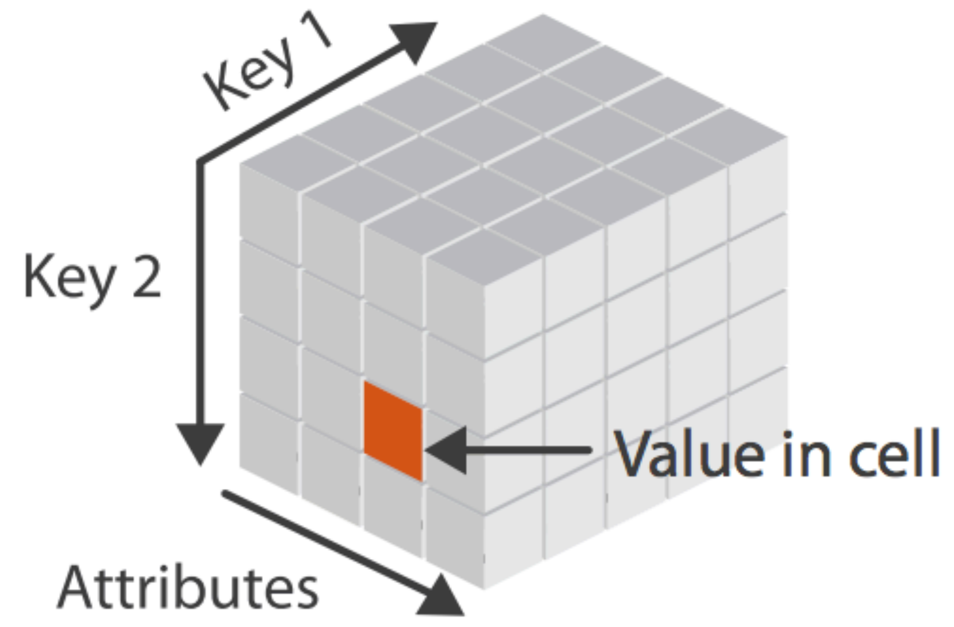
Multidimensional table

indexing on multiple keys

ID	Name	Shirt Size	Favorite color
1	Bob	M	Blue
2	Alice	S	Red

unique key

→ *Multidimensional Table*



Example: Parallel Coordinates

protein (g)
calcium (g)
sodium (g)
fiber (g)
vitaminc (g)
potassium (g)
carbohydrate (g)
sugars (g)
fat (g)
water (g)
calories
saturated (g)
monounsat (g)
polyunsa
0 10 20 30 40 50 60 70 80
0 1 2 3 4 5 6 7
0 5 10 15 20 25 30 35
0 10 20 30 40 50 60 70
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2
0 2 4 6 8 10 12 14 16
0 10 20 30 40 50 60 70 80 90 100
0 10 20 30 40 50 60 70 80 90
0 10 20 30 40 50 60 70 80 90 100
0 10 20 30 40 50 60 70 80 90 100
0 100 200 300 400 500 600 700 800 900
0 10 20 30 40 50 60 70 80 90
0 10 20 30 40 50 60 70 80
0 10 20 30 40 50 60 70
Hide Ticks
Show Ticks
Dark
Light

Rendered: 7637
Selected: 7637
Opacity: 0.0731

Drag along a vertical axis to brush
Tap the axis to remove its brush

■ Pasta with meatballs in tomato sauce, canned entree
■ Fish, tuna, fresh, bluefin, cooked, dry heat
■ Fish, tilapia, cooked, dry heat
■ Gravy, beef, canned, ready-to-serve
■ Beef, chuck eye Country-Style ribs, boneless, separable lean and fat, trimmed to 0" fat, choice, raw
■ CAMPBELL Soup Company, CAMPBELL'S SELECT Soup, Harvest Tomato with Basil Soup
■ Sherbet, orange
■ CAMPBELL Soup Company, SUPPER BAKES MEAL KITS, Lemon Chicken with herb rice (chicken not included)
■ Seeds, safflower seed kernels, dried
■ Alcoholic beverage, wine, table, white, Pinot Blanc

Networks / Trees

A graph G(V,E,W) consists of:

a set of **vertices** V (nodes)

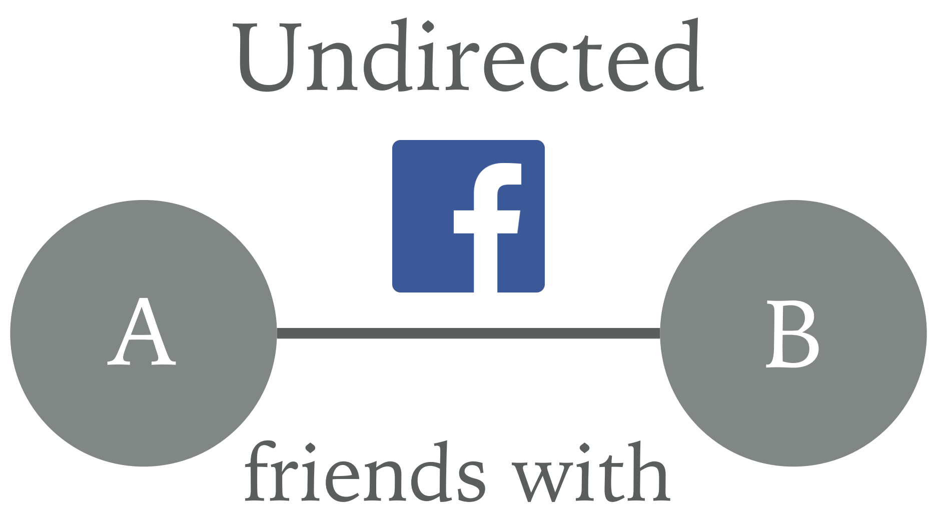
a set of **edges** E (links)

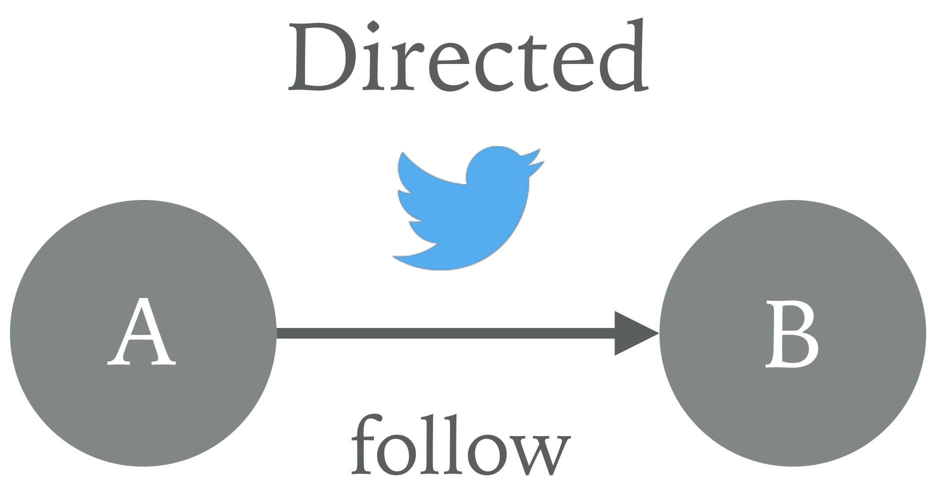
a set of weights W associated to the links

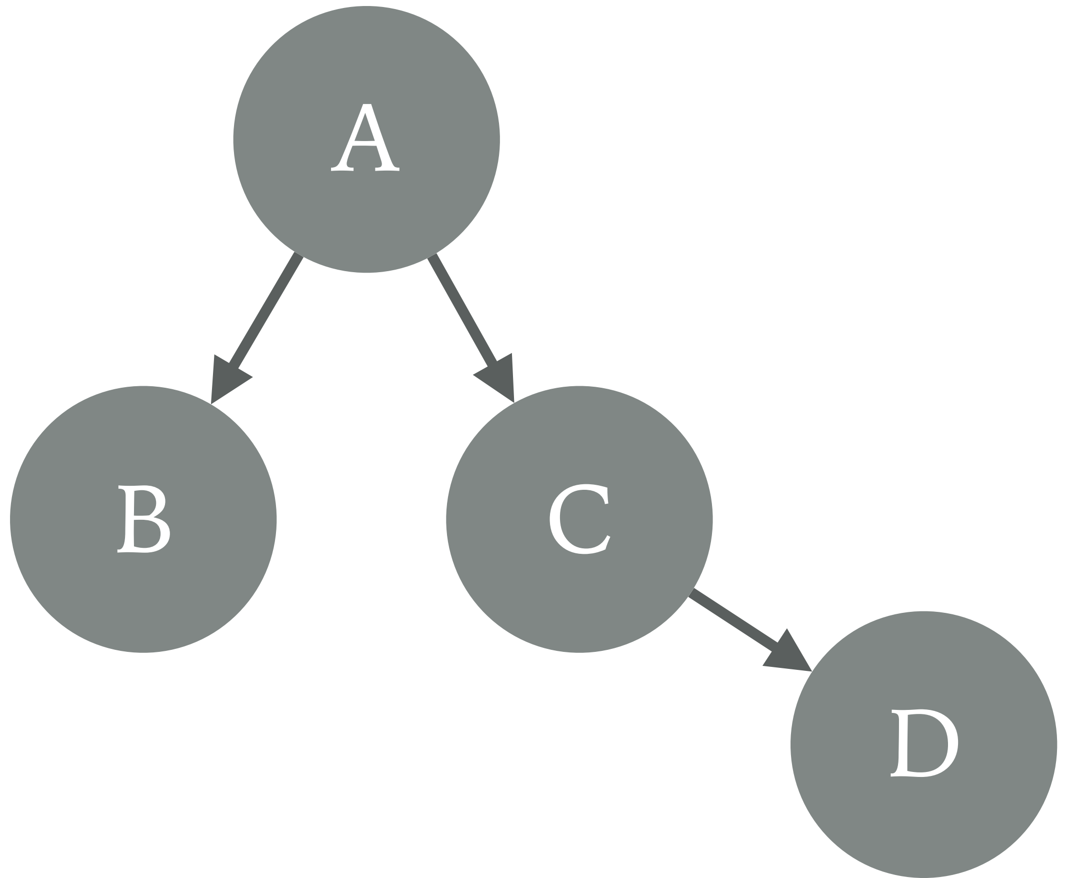
Networks with hierarchical structure are called trees

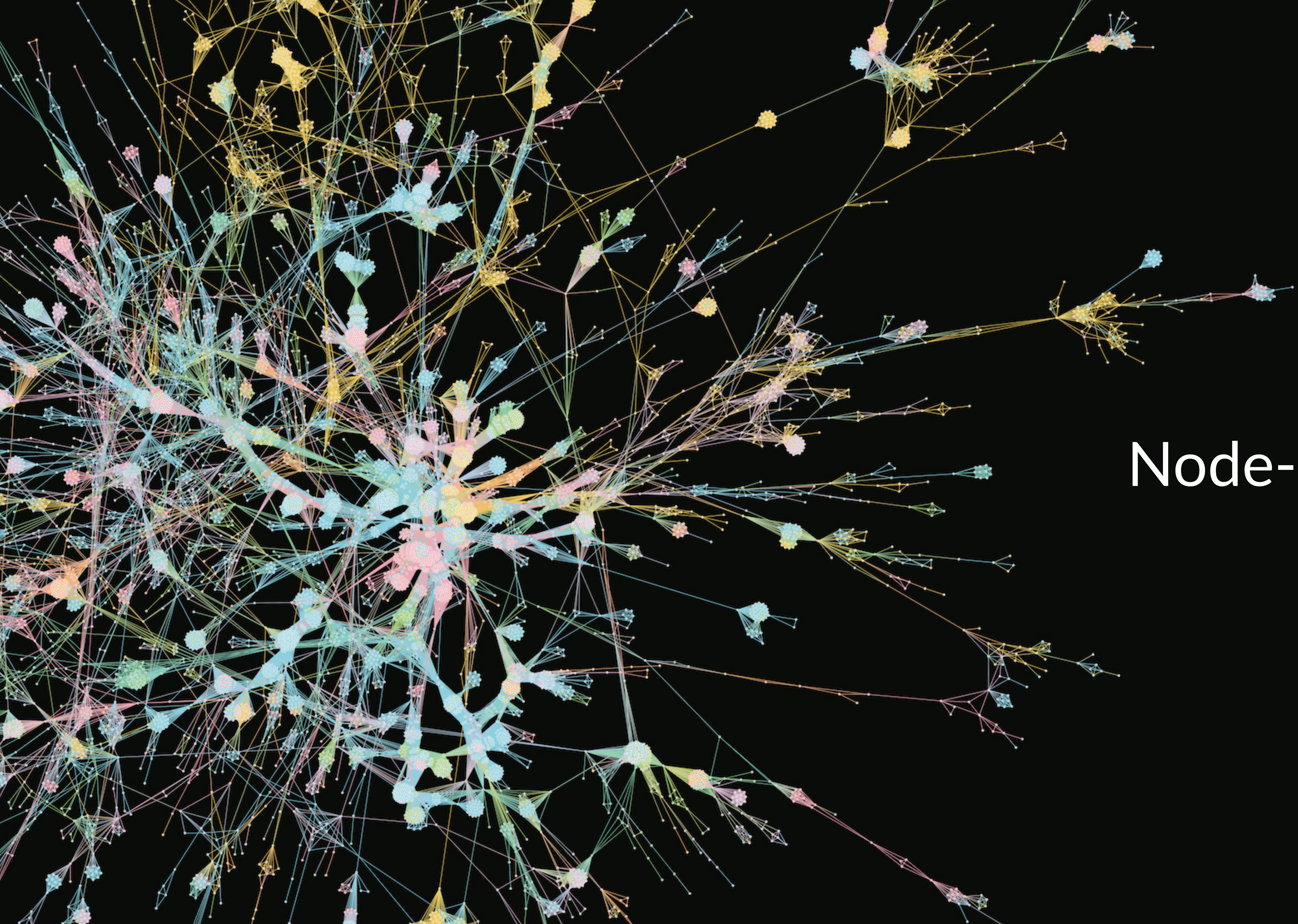
Each child one has only one parent node.

No cycles





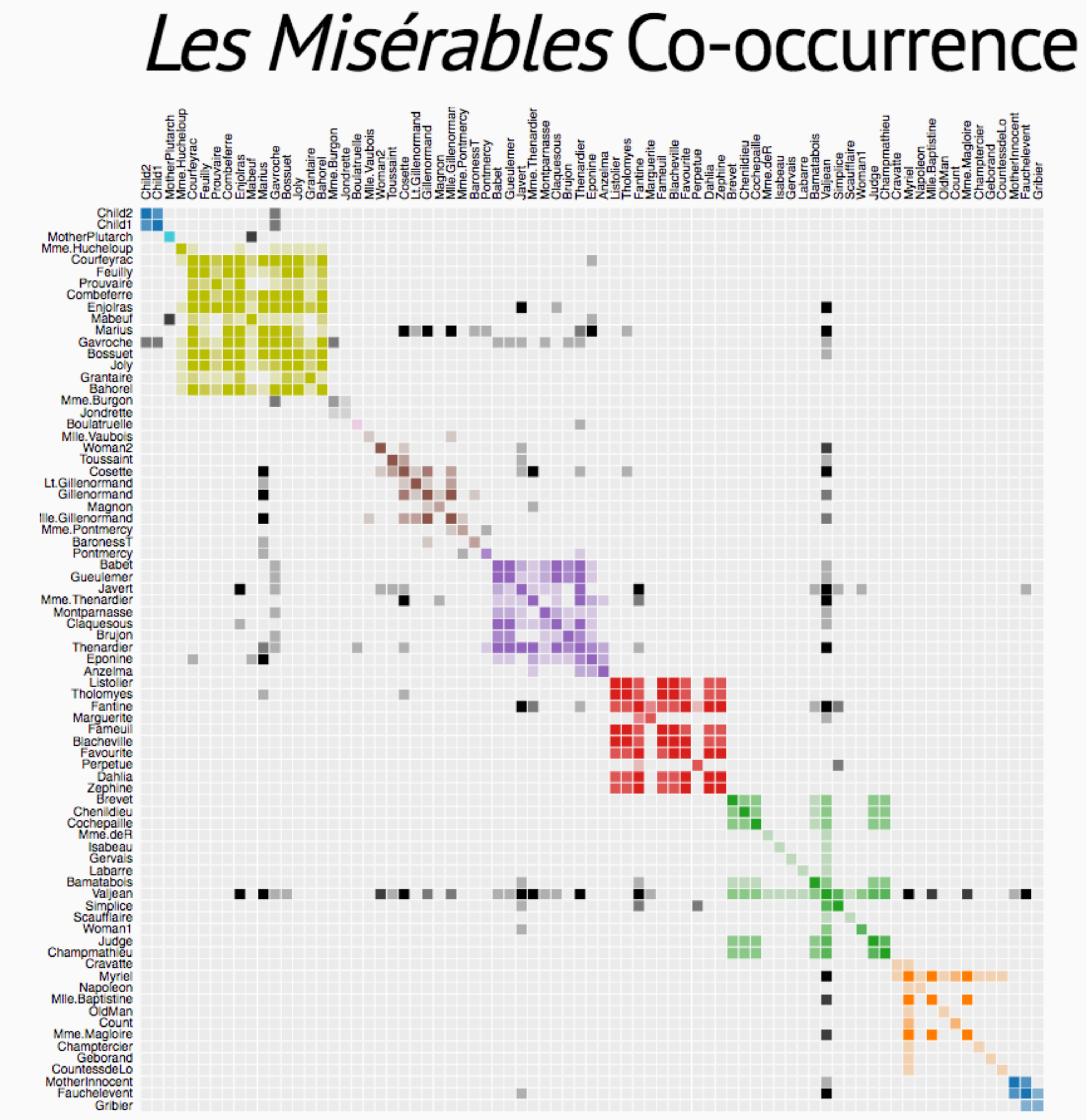




Node-link diagram

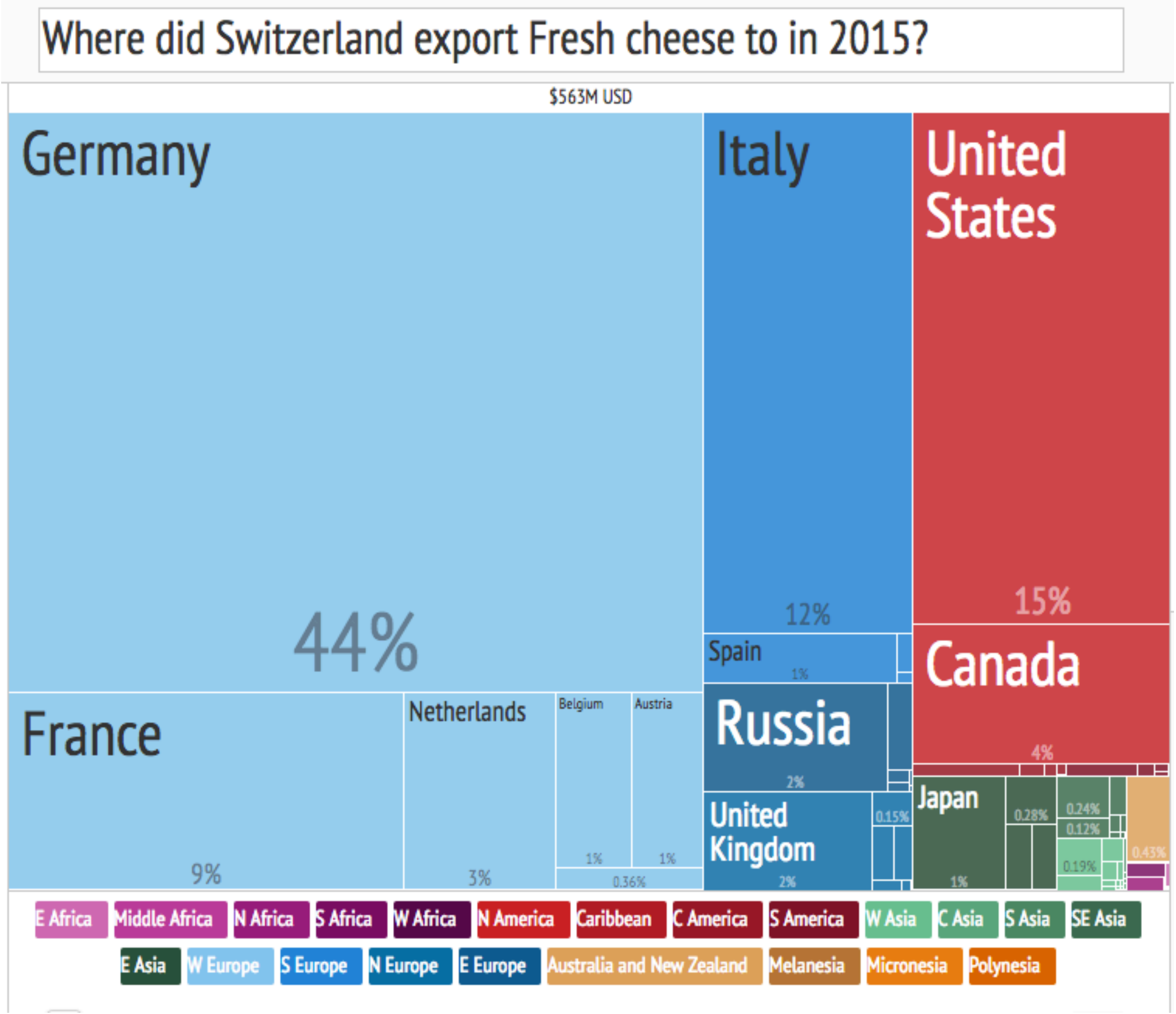
Kirell Benzi

Other graph viz



Les Misérables Co-occurrence

Adjacency matrix



Where did Switzerland export Fresh cheese to in 2015?

Treemap

Discrete vs continous

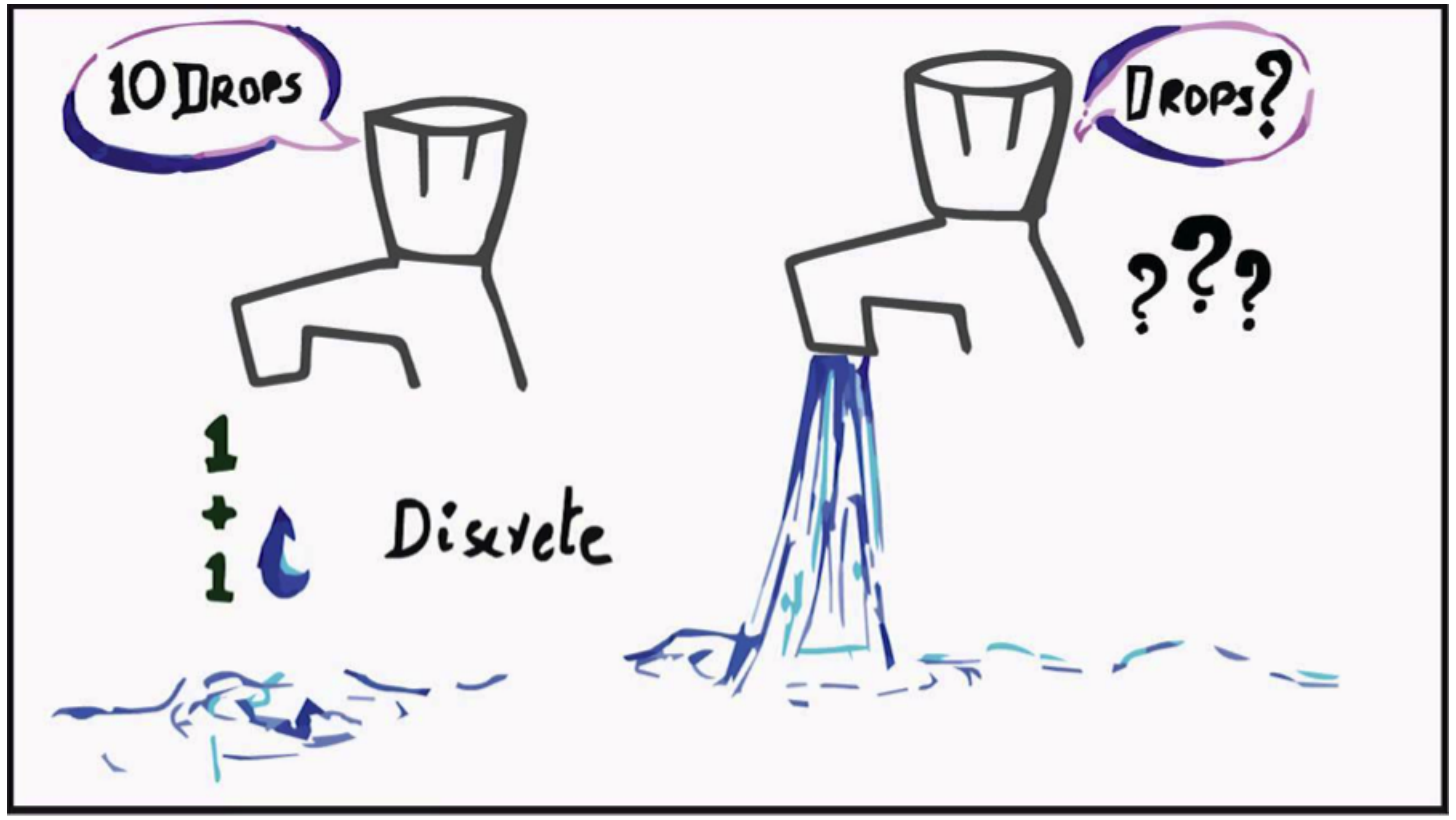
In mathematics, a variable may be **continuous** or **discrete.**

Continuous variable can take on infinitely uncountable values

Discrete variable has a finite set of possible values

Sampling - how frequently do you take the measurements?

Interpolation - how to show values in between sampled points?

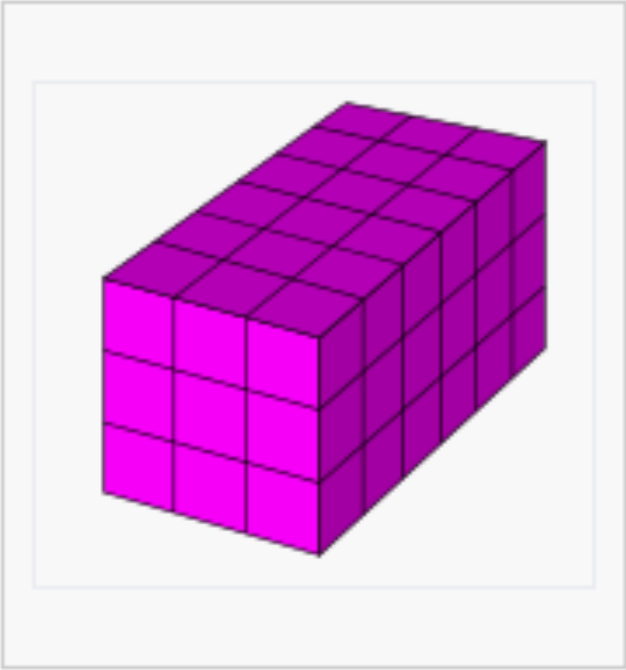


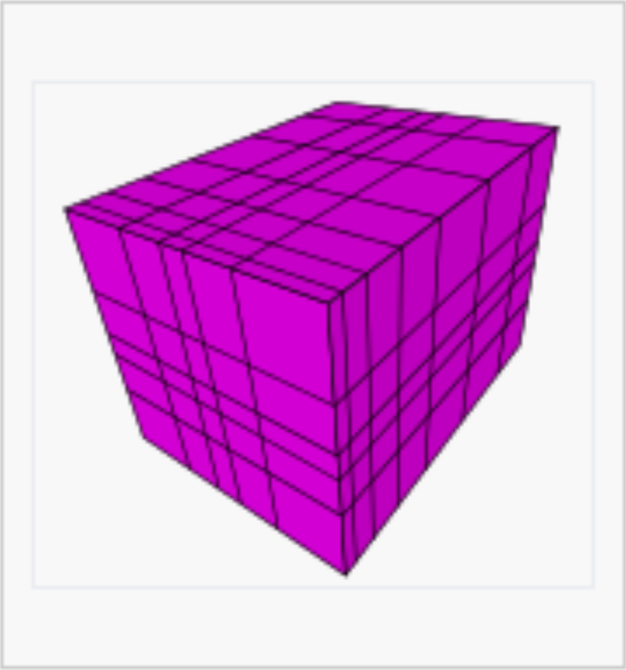
Fields and grid types

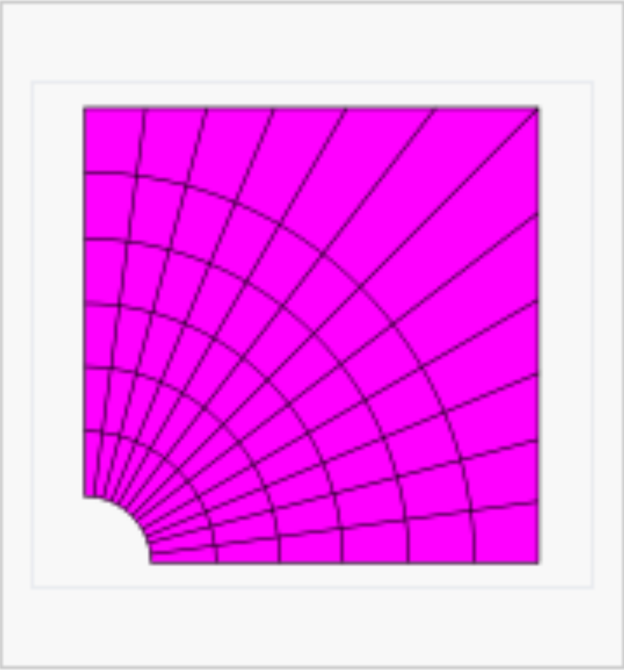
Data is sampled and interpolated from continuous domain into cells (tessellation)

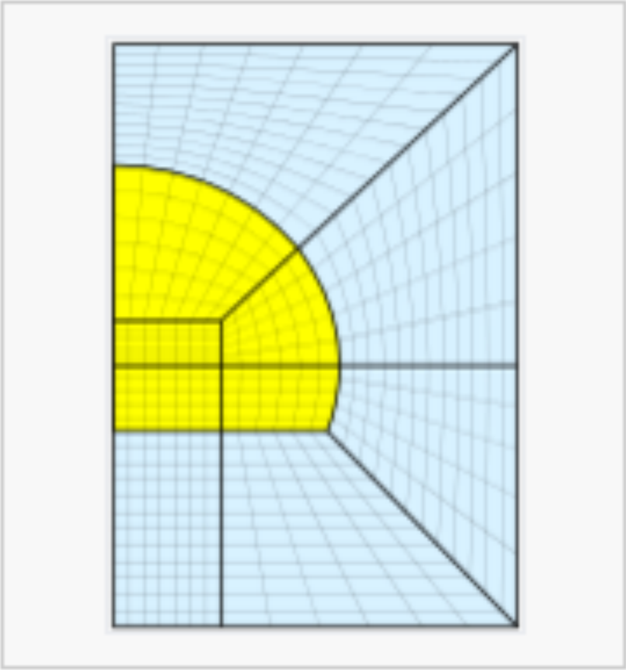
Examples: temperature, pressure, voxels (3D pixels)

Measured or simulated

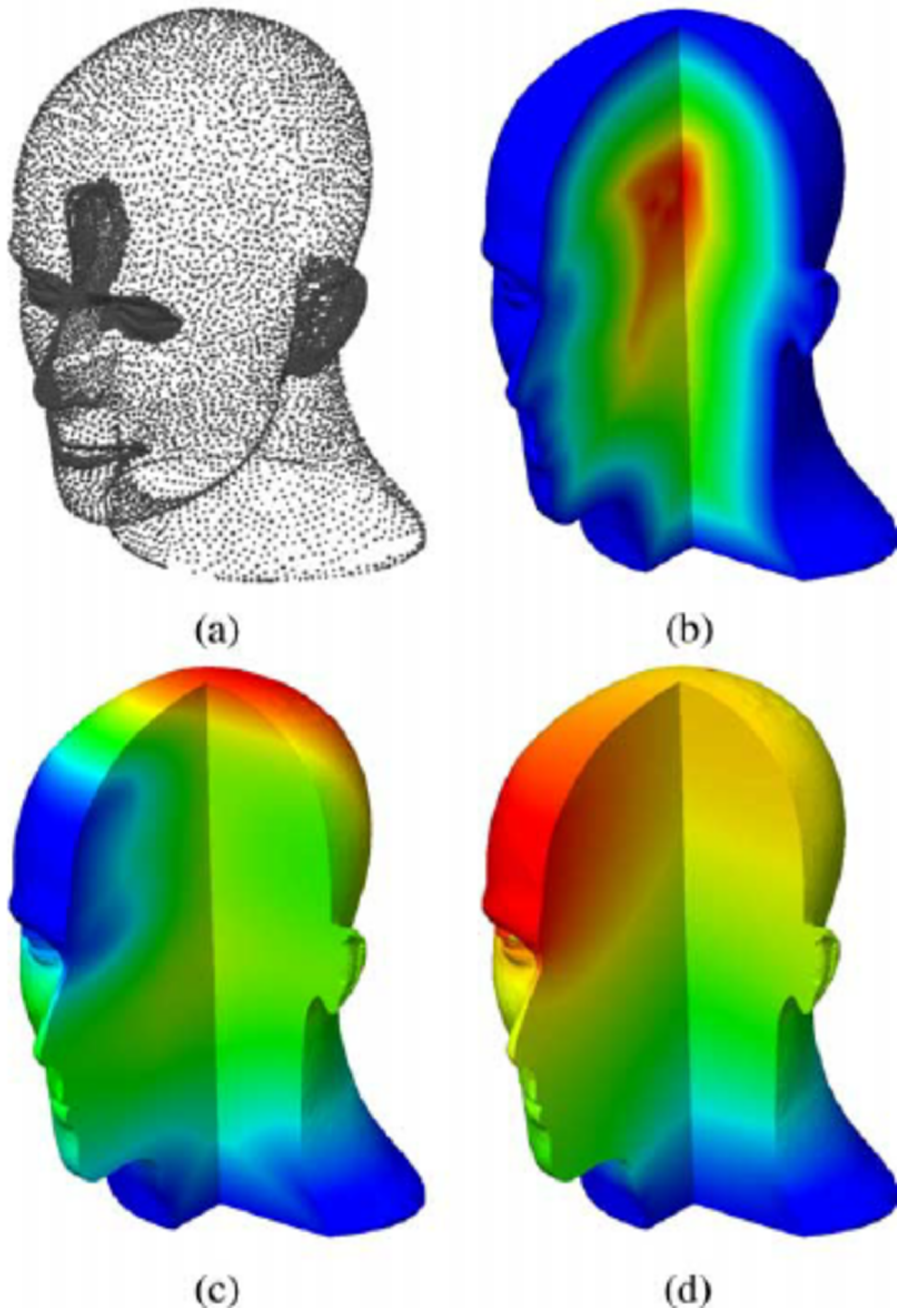

Cartesian grid


Rectilinear grid


Curvilinear grid


Curvilinear grid

[Wikipedia]



[Freytag 2006]

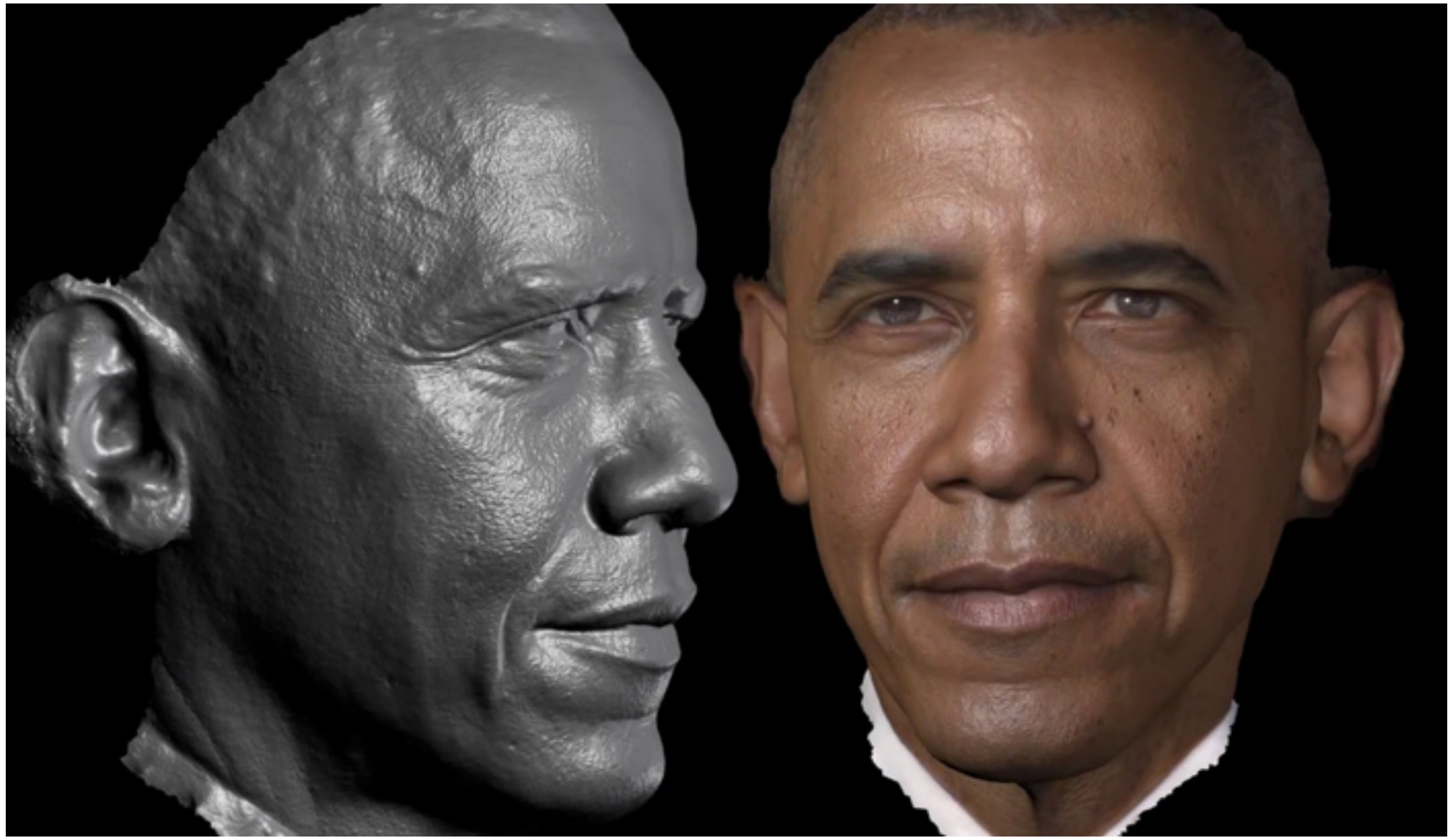
Spatial datasets

Explicit spatial positions

Fixed shapes

Cannot choose to where to draw points, lines, curves, regions, etc.

Not a data viz topic

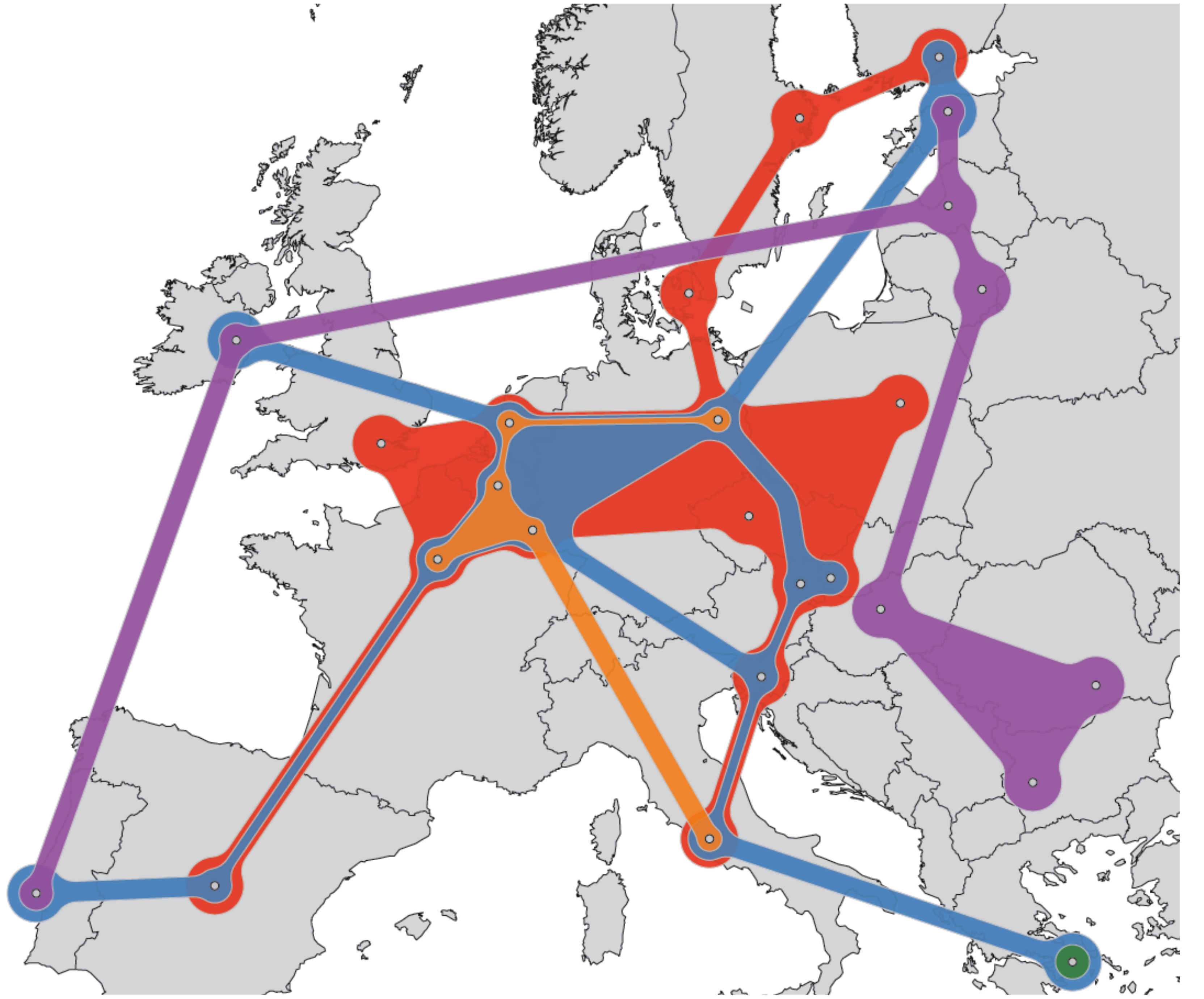


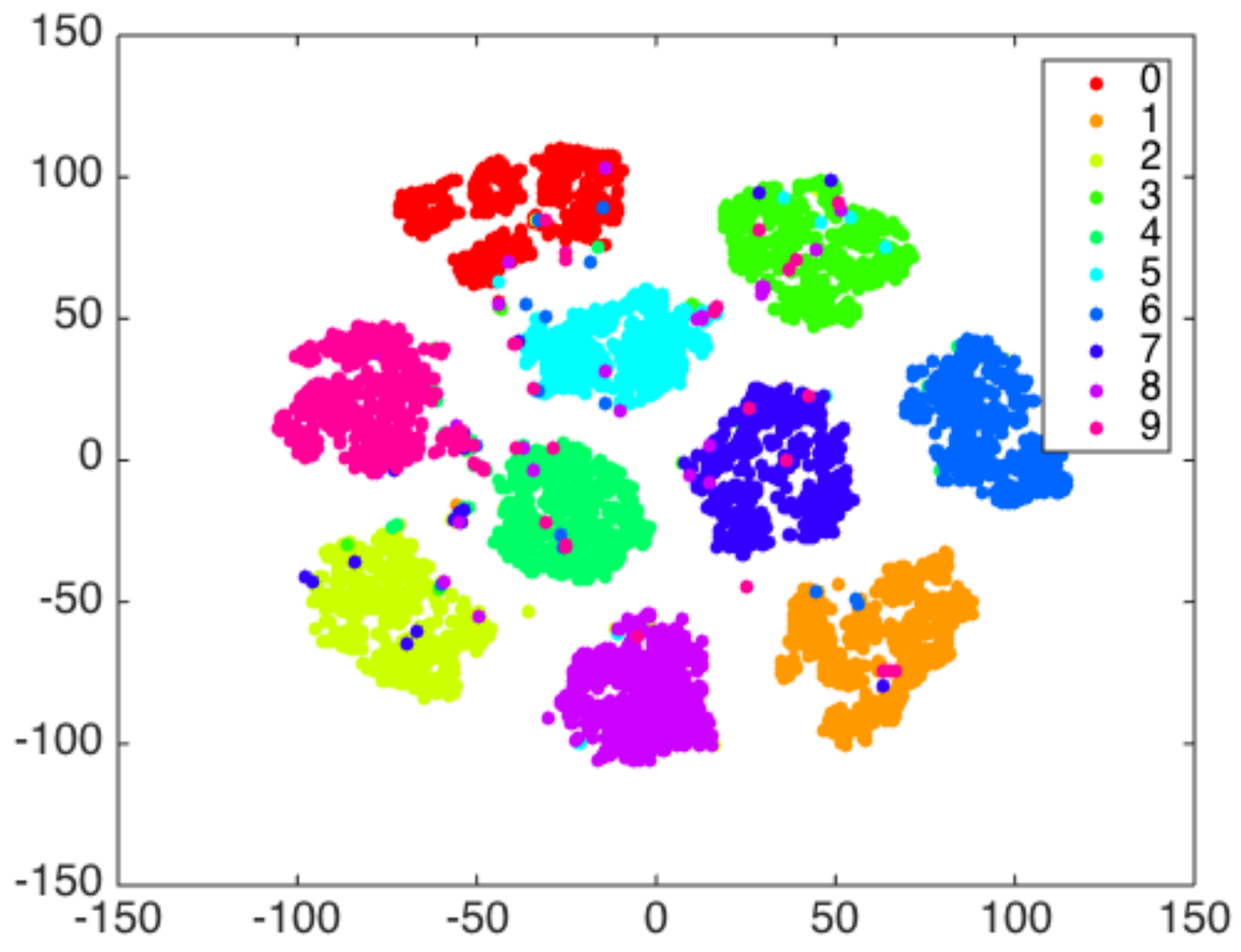
Sets, Lists, Clusters

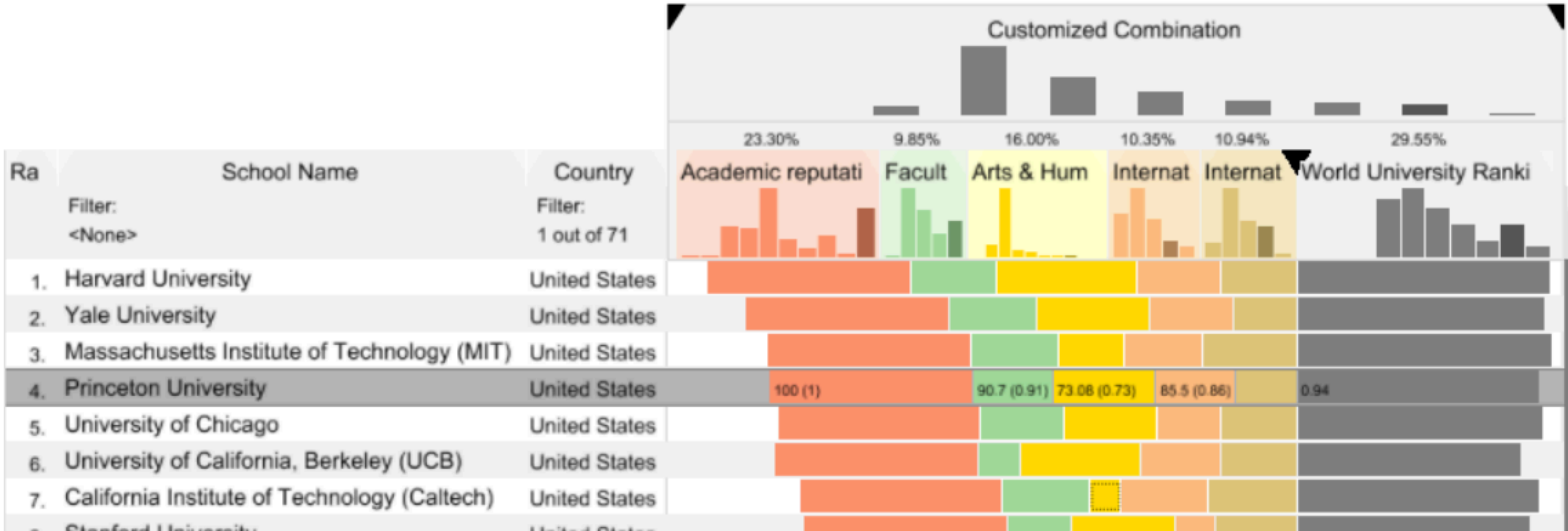
Set: unique items unordered

List: ordered with possible duplicates

Cluster: group of similar items







Customized Combination

Ra	School Name	Country	Academic reputati	Facult	Arts & Hum	Internat	Internat	World University Ranki
			23.30%	9.85%	16.00%	10.35%	10.94%	29.55%
	Filter: <None>	Filter: 1 out of 71						
1.	Harvard University	United States						
2.	Yale University	United States						
3.	Massachusetts Institute of Technology (MIT)	United States						
4.	Princeton University	United States	100 (1)	90.7 (0.91)	73.08 (0.73)	85.5 (0.86)		0.94
5.	University of Chicago	United States						
6.	University of California, Berkeley (UCB)	United States						
7.	California Institute of Technology (Caltech)	United States						
8.	Stanford University	United States						

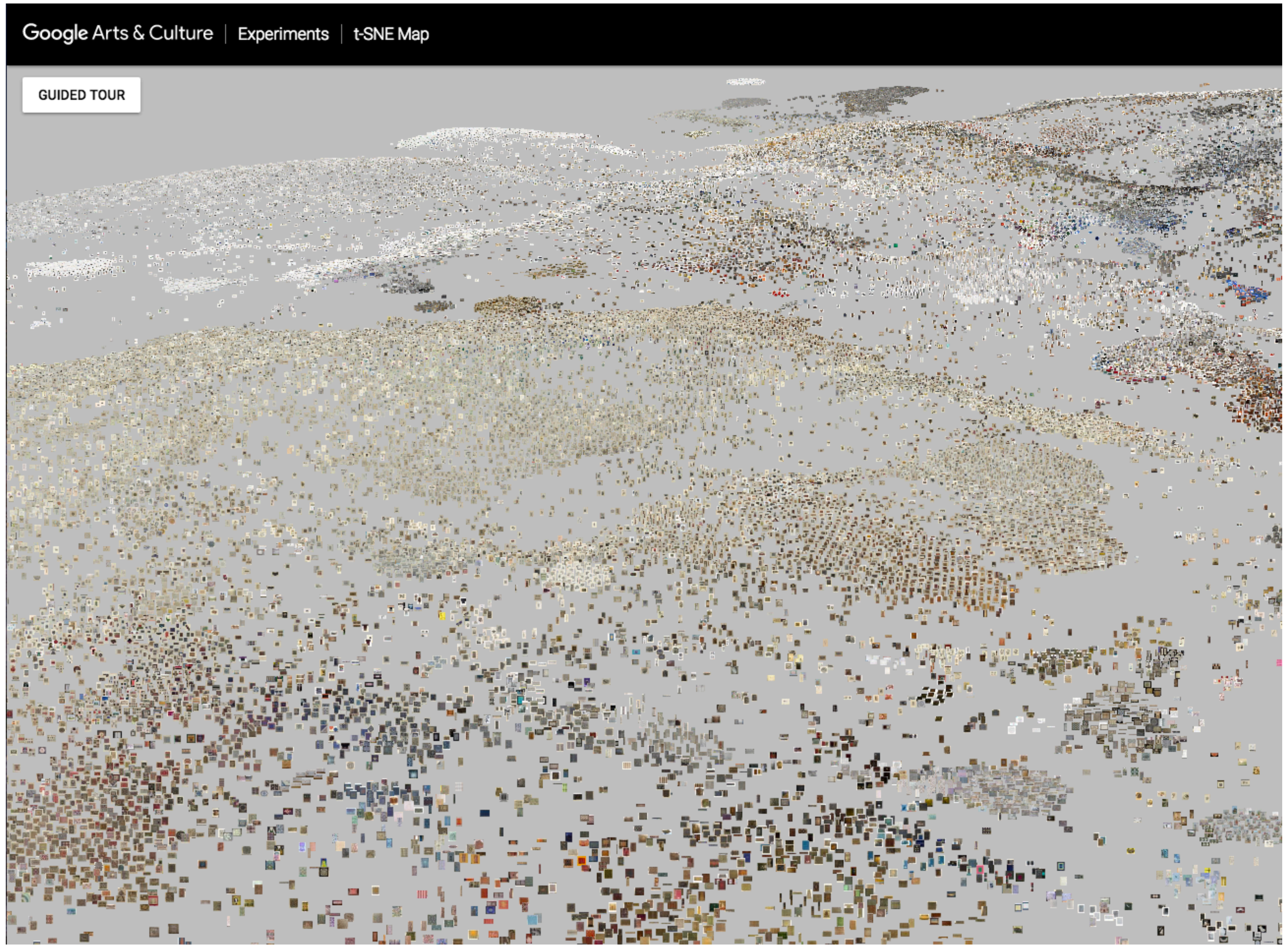
What about unstructured data?

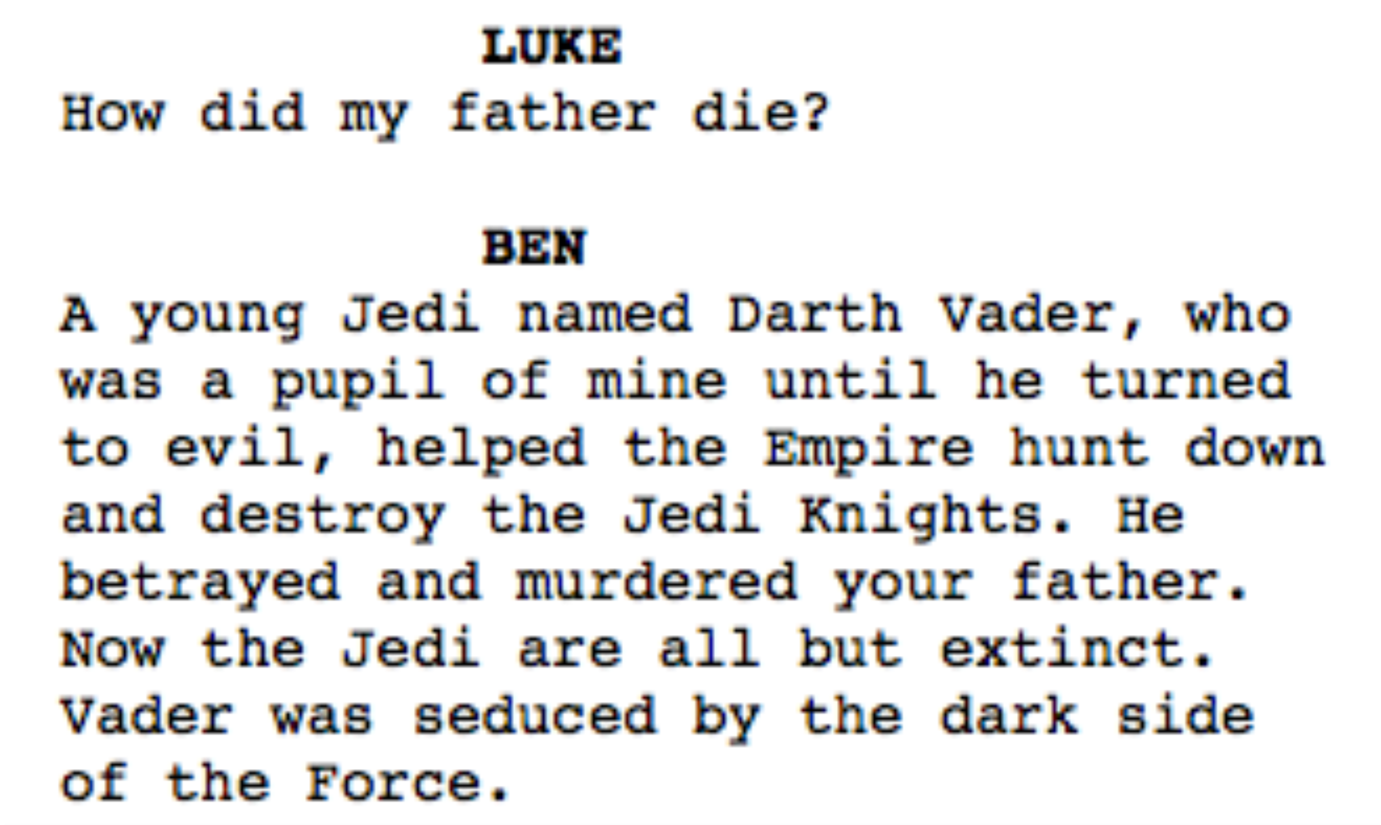
No predefined data model

Mixed media: rich-text with images, videos

We have to transform into structured data:

- Natural Language Processing
- Text mining (keywords, ontologies, categories)
- Features extraction





```
                  LUKE
How did my father die?

                  BEN
A young Jedi named Darth Vader, who
was a pupil of mine until he turned
to evil, helped the Empire hunt down
and destroy the Jedi Knights. He
betrayed and murdered your father.
Now the Jedi are all but extinct.
Vader was seduced by the dark side
of the Force.
```

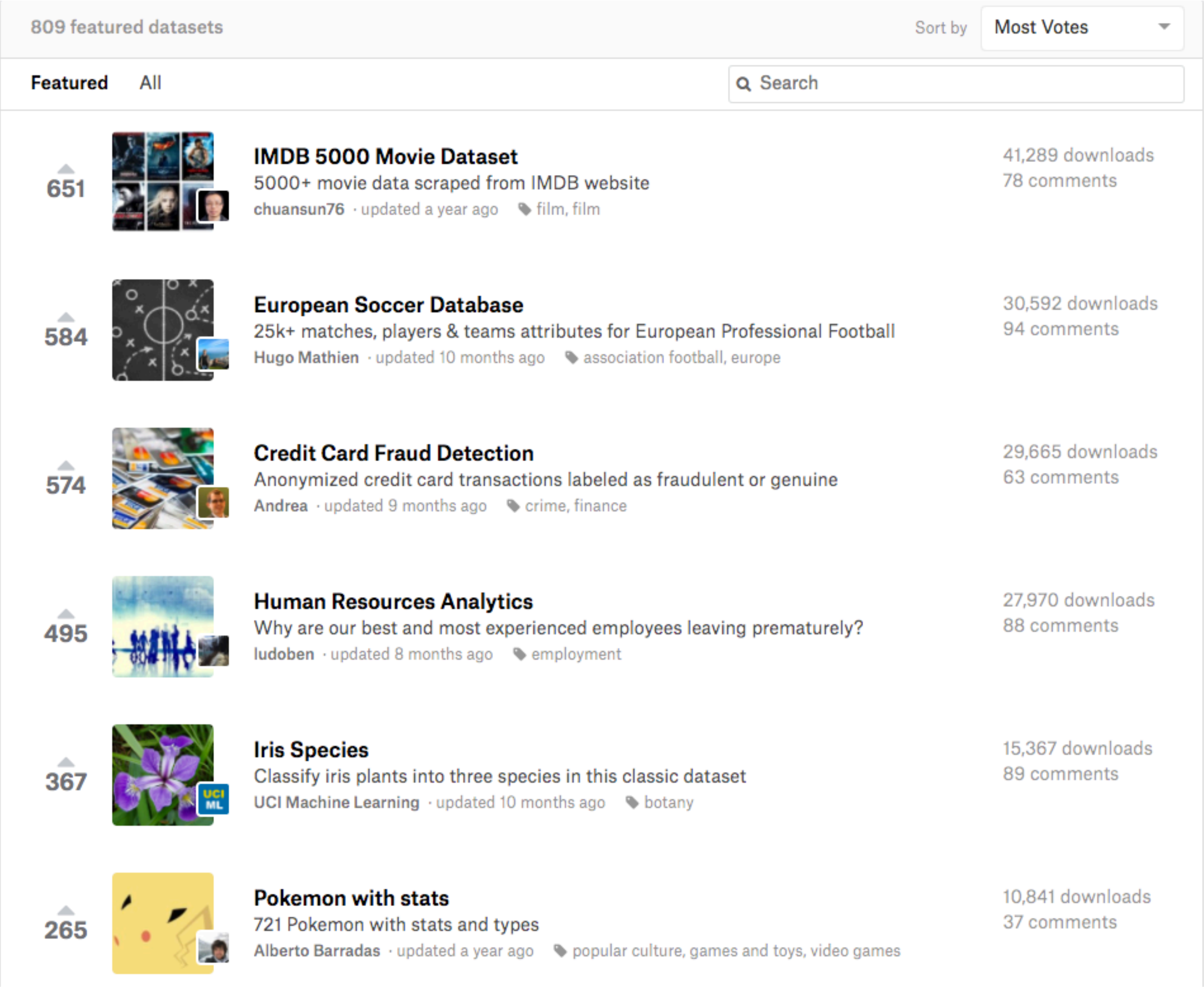
Data acquisition

Datasets curated by others

Third-party services: APIs

Web scraping

Public datasets



809 featured datasets

Sort by Most Votes

Featured All

Search

651 **IMDB 5000 Movie Dataset**
5000+ movie data scraped from IMDB website
chuansun76 · updated a year ago film, film
41,289 downloads
78 comments

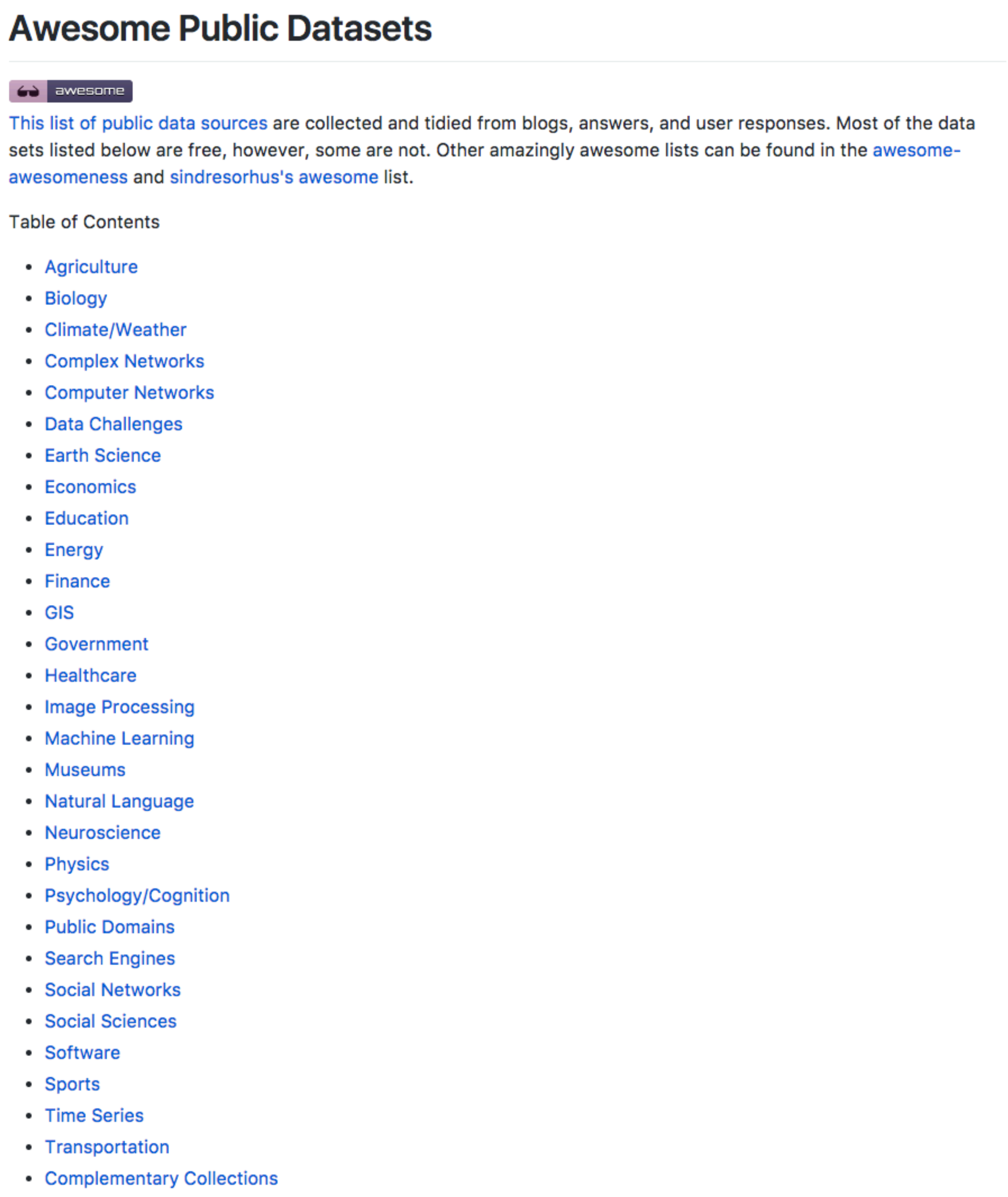
584 **European Soccer Database**
25k+ matches, players & teams attributes for European Professional Football
Hugo Mathien · updated 10 months ago association football, europe
30,592 downloads
94 comments

574 **Credit Card Fraud Detection**
Anonymized credit card transactions labeled as fraudulent or genuine
Andrea · updated 9 months ago crime, finance
29,665 downloads
63 comments

495 **Human Resources Analytics**
Why are our best and most experienced employees leaving prematurely?
ludoben · updated 8 months ago employment
27,970 downloads
88 comments

367 **Iris Species**
Classify iris plants into three species in this classic dataset
UCI Machine Learning · updated 10 months ago botany
15,367 downloads
89 comments

265 **Pokemon with stats**
721 Pokemon with stats and types
Alberto Barradas · updated a year ago popular culture, games and toys, video games
10,841 downloads
37 comments



Awesome Public Datasets

awesome

This list of public data sources are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. Other amazingly awesome lists can be found in the awesome-awesomeness and sindresorhus's awesome list.

Table of Contents

- Agriculture
- Biology
- Climate/Weather
- Complex Networks
- Computer Networks
- Data Challenges
- Earth Science
- Economics
- Education
- Energy
- Finance
- GIS
- Government
- Healthcare
- Image Processing
- Machine Learning
- Museums
- Natural Language
- Neuroscience
- Physics
- Psychology/Cognition
- Public Domains
- Search Engines
- Social Networks
- Social Sciences
- Software
- Sports
- Time Series
- Transportation
- Complementary Collections

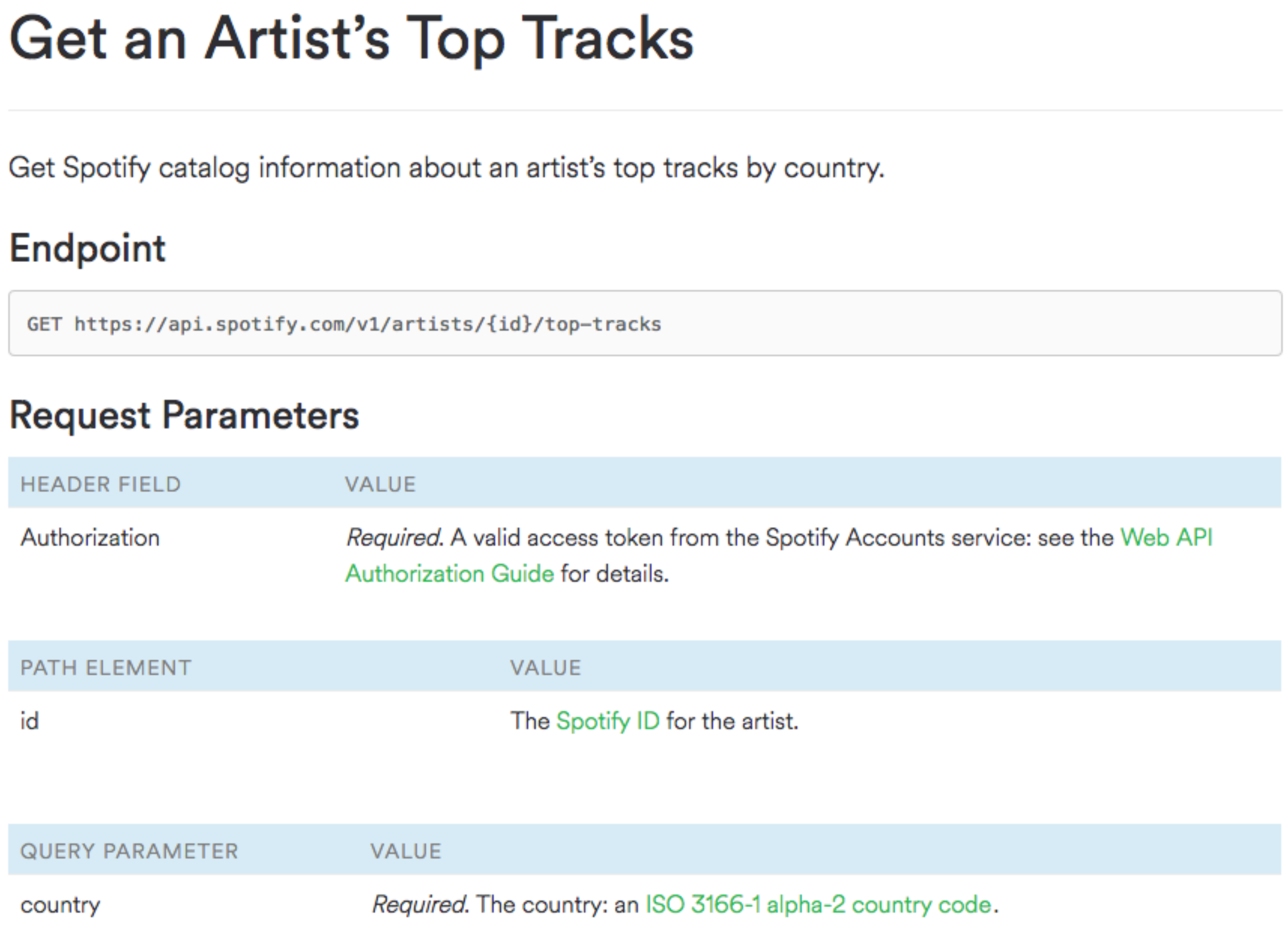
Third-party services: APIs

Search for a developer section or API on the service you want

Register for free or paid to access the service with an API key

Look for a client library in your favorite programming language

Download data



Get an Artist’s Top Tracks

Get Spotify catalog information about an artist’s top tracks by country.

Endpoint

```
GET https://api.spotify.com/v1/artists/{id}/top-tracks
```

Request Parameters

HEADER FIELD	VALUE
Authorization	*Required.* A valid access token from the Spotify Accounts service: see the Web API Authorization Guide for details.

PATH ELEMENT	VALUE
id	The Spotify ID for the artist.

QUERY PARAMETER	VALUE
country	*Required.* The country: an ISO 3166-1 alpha-2 country code.

Read the legal terms on what you can and cannot do with the data

Web scraping

If an information is visible on a browser, it can be downloaded

Extracting information from the DOM is painful and should be considered only on last resort

Depend on the layout, change can happen without notice

Use a library: Scrapy (Python)

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

Don’t be evil!

Data formats

One of the most common data format that we encounter on the web is **JSON**: JavaScript Object Notation (.json)

The other common format for datasets is **CSV**: Comma separated value (.csv)

CSV example

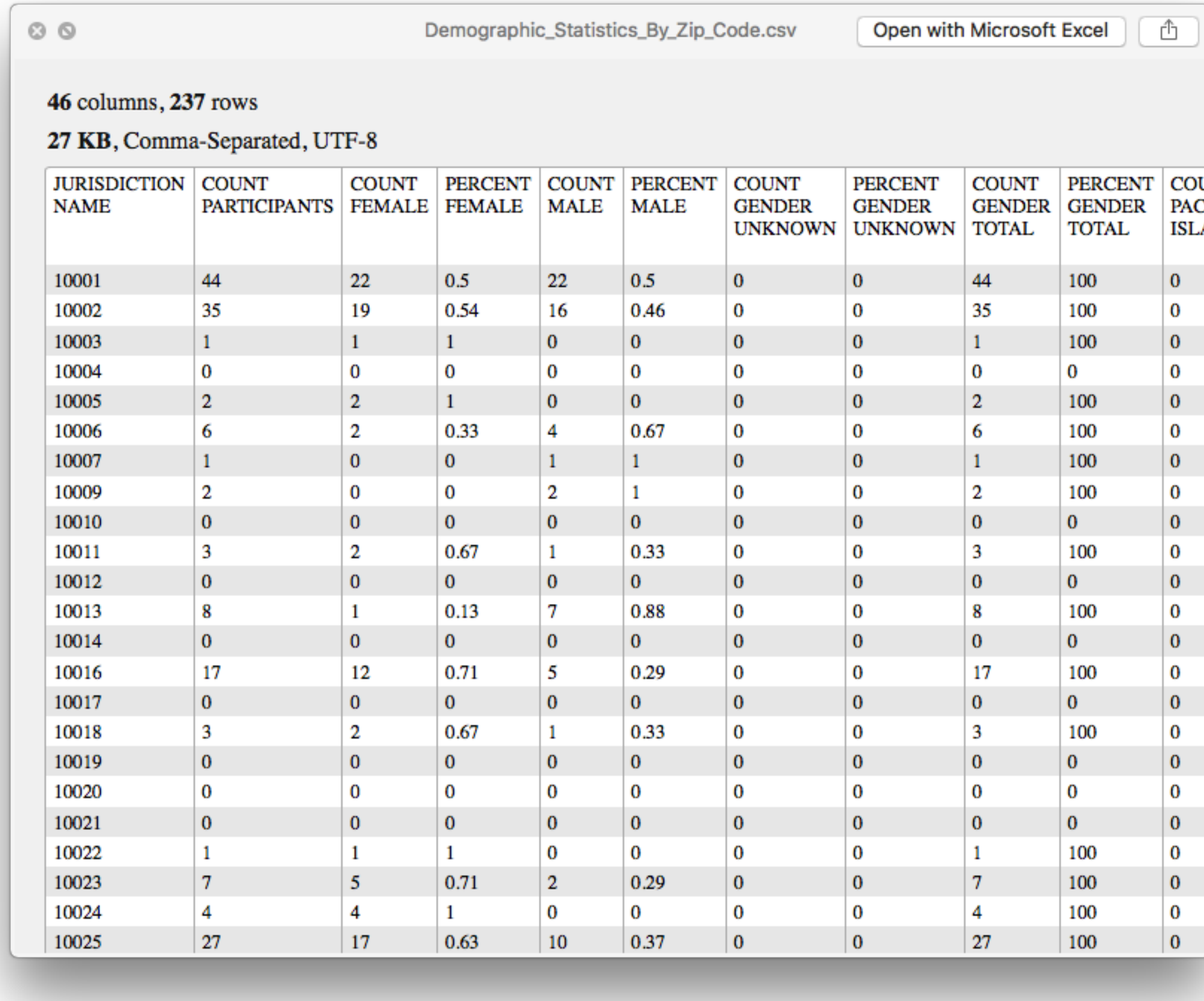
Flat table view with column names

Easy to append data

CSV has no standard encoding

No standard column separator and multiple character escaping standards.

String is the only type supported for cell values


Demographic_Statistics_By_Zip_Code.csv
Open with Microsoft Excel

46 columns, **237** rows

27 KB, Comma-Separated, UTF-8

JURISDICTION NAME	COUNT PARTICIPANTS	COUNT FEMALE	PERCENT FEMALE	COUNT MALE	PERCENT MALE	COUNT GENDER UNKNOWN	PERCENT GENDER UNKNOWN	COUNT GENDER TOTAL	PERCENT GENDER TOTAL	COU PAC ISL
10001	44	22	0.5	22	0.5	0	0	44	100	0
10002	35	19	0.54	16	0.46	0	0	35	100	0
10003	1	1	1	0	0	0	0	1	100	0
10004	0	0	0	0	0	0	0	0	0	0
10005	2	2	1	0	0	0	0	2	100	0
10006	6	2	0.33	4	0.67	0	0	6	100	0
10007	1	0	0	1	1	0	0	1	100	0
10009	2	0	0	2	1	0	0	2	100	0
10010	0	0	0	0	0	0	0	0	0	0
10011	3	2	0.67	1	0.33	0	0	3	100	0
10012	0	0	0	0	0	0	0	0	0	0
10013	8	1	0.13	7	0.88	0	0	8	100	0
10014	0	0	0	0	0	0	0	0	0	0
10016	17	12	0.71	5	0.29	0	0	17	100	0
10017	0	0	0	0	0	0	0	0	0	0
10018	3	2	0.67	1	0.33	0	0	3	100	0
10019	0	0	0	0	0	0	0	0	0	0
10020	0	0	0	0	0	0	0	0	0	0
10021	0	0	0	0	0	0	0	0	0	0
10022	1	1	1	0	0	0	0	1	100	0
10023	7	5	0.71	2	0.29	0	0	7	100	0
10024	4	4	1	0	0	0	0	4	100	0
10025	27	17	0.63	10	0.37	0	0	27	100	0

JSON example

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

Can hold complex data structure with nested fields

Parsing is difficult and should never be done by hand

Cannot append data easily

JSON Lines

```
{"name": "Gilbert", "wins": [["straight", "7♣"], ["one pair", "10♥"]]}
{"name": "Alexa", "wins": [["two pair", "4♠"], ["two pair", "9♠"]]}
{"name": "May", "wins": []}
{"name": "Deloise", "wins": [["three of a kind", "5♣"]]}
```

Valid JSON per line

Can hold complex data structure with nested fields

Append data easily

Support streaming

Tips

Don’t aim for the perfect dataset: complete, accurate, up to date. It doesn’t exist.

Collecting and cleaning data can take up to 80% of your time.

Backup and version the data you have.

Clean the data, take your time to learn the dataset.

If you don’t know about some aspects of the data: ask, don’t assume.