

Data Transformations

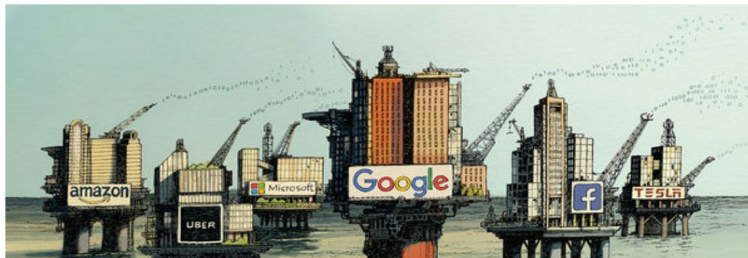
IFT6758

Fall 2019

Data are not Oil

The world's most valuable resource is
no longer oil, but data

The data economy demands a new approach to antitrust rules



Data is the New Oil

February 9th 2019

[TWEET THIS](#)

“Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”

— Clive Humby

TECH • BRAINSTORM TECH

Why Data Is The New Oil

By [Jonathan Vanian](#) July 11, 2016

Data are Desserts!


1. **Data are the result of deliberate human intervention**
2. Data are varied across domains
3. Data are varied within domains



Data are Desserts!

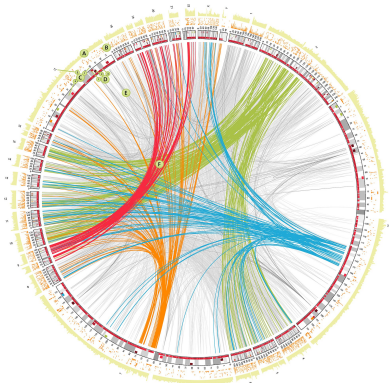
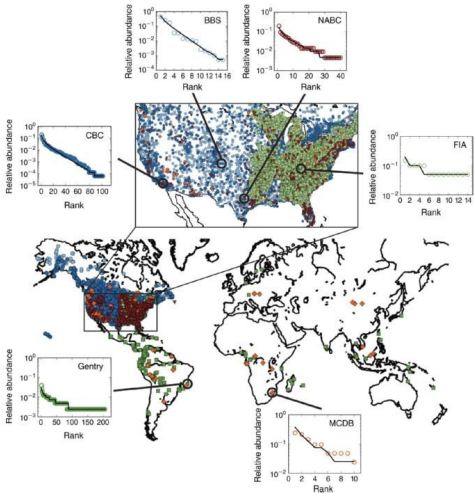
- 1. Data are the result of deliberate human intervention
- 2. Data are varied across domains
- 3. Data are varied within domains

Big Data & Smart Cities: How can we prepare for them?

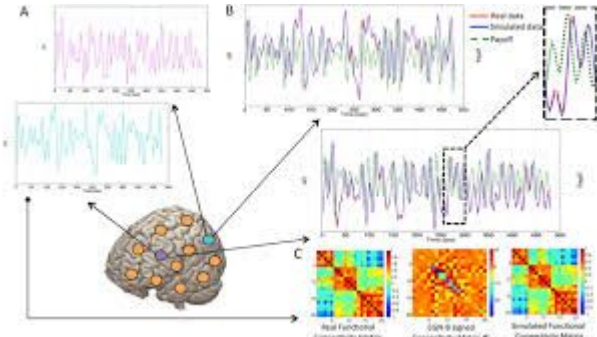
 Alexandre Gonfalonieri [Follow](#)
Dec 13, 2018 · 8 min read ★

Artnome

[Art Analytics](#) [Art Authentication](#) [Gallery](#) [About](#)

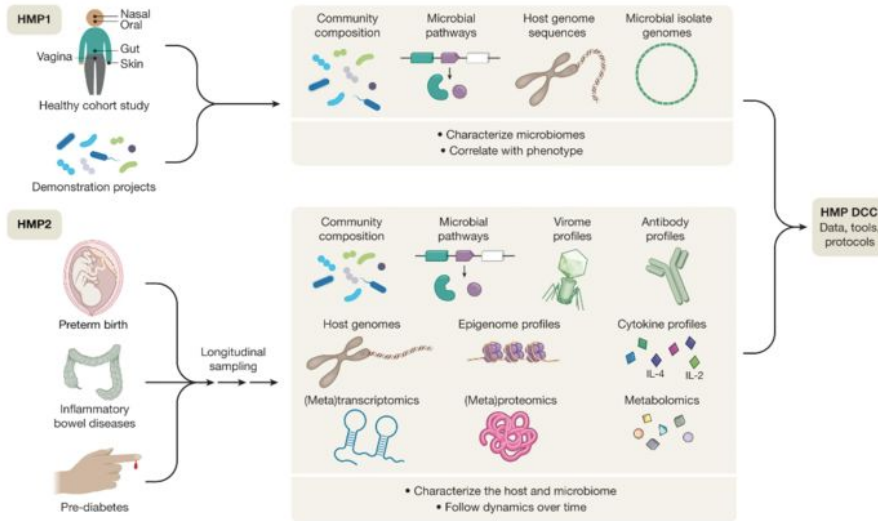


Exploring Art Through Data

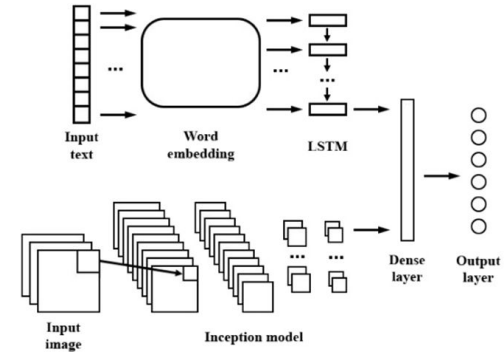


Data are Desserts!

1. Data are the result of deliberate human intervention
2. Data are varied across domains
3. **Data are varied within domains**



DeepSentiment: a multimodal neural network (2)



Data Wrangling

Data (+ people who collect them) are varied

→ Some amount of preparation is always needed.



Readings

[How to share data with a statistician](#)

[Tidy Data](#)

[Tidy Data in Python](#)

Data Wrangling

Data (+ people who collect them) are varied

→ Some amount of preparation is always needed.

This is *before* you prepare your deliverables,

- Prediction tools
- Model summaries
- Figures and reports

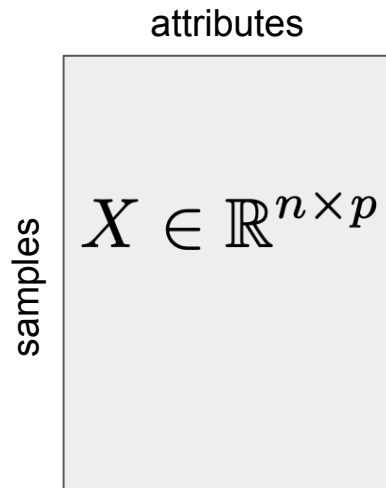
which inform **understanding** and **decision making**.

Walkthrough

You just got a dataset! You should

- Understand what the variables are
- Manage column types
- Handle missing values
- Join, reorganize, and tidy

The data don't arrive on our doorstep as "X".



Data Dictionary + Code Book

- What do the tables mean?
- What do the columns mean?
- How were the data collected?

DATA

employee_id	first_name	last_name	nin	dept_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Bary	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Bemdt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

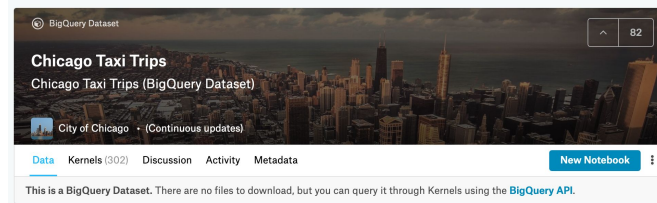
Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
dept_id	int	Employee department. Ref: Departments
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Managing types

- Data come in different “types”
 - Numeric, (ordered) categorical, dates, (positive) integers
- It's valuable to ensure consistency with what you were expecting. (why?)

```
In [6]:
taxi.dtypes
Out [6]:
unique_key          object
taxi_id             object
trip_start_timestamp  object
trip_end_timestamp  object
trip_seconds        float64
trip_miles          float64
pickup_census_tract float64
dropoff_census_tract float64
pickup_community_area float64
dropoff_community_area float64
fare                float64
tips                float64
tolls               float64
extras              float64
trip_total           float64
payment_type        object
company             object
pickup_latitude     float64
pickup_longitude    float64
pickup_location     float64
dropoff_latitude    float64
dropoff_longitude   float64
dropoff_location    float64
dtype: object
```

What do you need to change?



The screenshot shows a BigQuery Dataset page for "Chicago Taxi Trips". The page title is "Chicago Taxi Trips (BigQuery Dataset)". Below the title, there is a description: "Chicago Taxi Trips (BigQuery Dataset)". There is a small icon of the City of Chicago and the text "City of Chicago • (Continuous updates)". At the bottom of the page, there is a navigation bar with links for "Data", "Kernels (302)", "Discussion", "Activity", and "Metadata". A "New Notebook" button is also visible. A message at the bottom states: "This is a BigQuery Dataset. There are no files to download, but you can query it through Kernels using the BigQuery API."

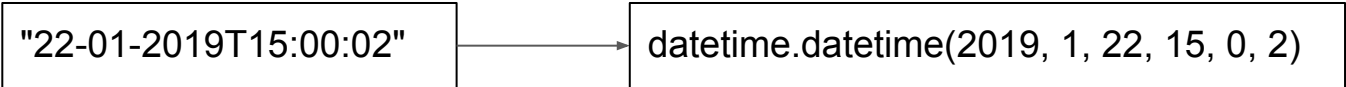
You'll see many more examples in practicals + HW + project.

Managing types: Dates

- You can use the `datetime` package and `pandas`' `to_datetime`
- Lets you convert arbitrary strings into `datetime` objects

"22-01-2019T15:00:02"

`datetime.datetime(2019, 1, 22, 15, 0, 2)`

A diagram illustrating the conversion of a string to a datetime object. On the left, a rectangular box contains the ISO 8601 string "22-01-2019T15:00:02". A horizontal arrow points from this box to a second rectangular box on the right. The second box contains the Python datetime object representation: datetime.datetime(2019, 1, 22, 15, 0, 2).

Managing types: Dates

- You can use the datetime package and pandas' `to_datetime`
- Lets you convert arbitrary strings into datetime objects

"22-01-2019T15:00:02"

`datetime.datetime(2019, 1, 22, 15, 0, 2)`

```
In [13]:
taxi = taxi.assign(
    day=lambda df: df.trip_start_timestamp.map(lambda x: x.day),
    weekday=lambda df: df.trip_start_timestamp.map(lambda x: x.weekday),
    hour=lambda df: df.trip_start_timestamp.map(lambda x: x.hour)
)
taxi[["day", "weekday"]]
Out [13]:
```

	day	weekday
0	7	6
1	7	6
2	7	6
3	20	5
4	30	5
...
19995	7	6
19996	9	1
19997	7	6
19998	16	1
19999	5	4

```
[20000 rows x 2 columns]
```

Once it's a datetime, you can derive new features.

Managing types: Categoricals

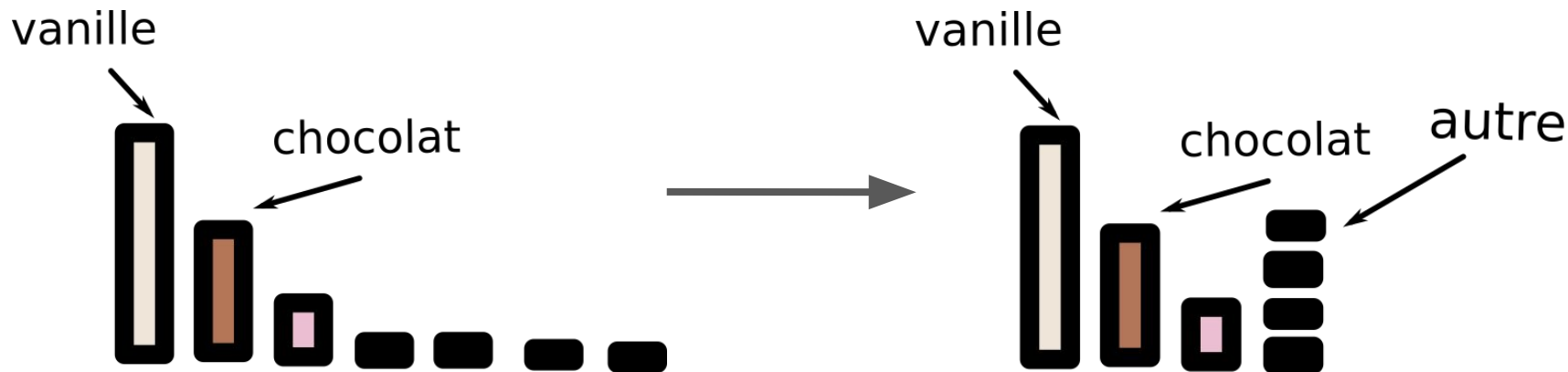
There are three common issues,

- The number of levels is overwhelming
- A single categorical might encode multiple pieces of information
- The levels might not be consolidated
- You might want to convert into numerical vectors

Managing types: Categoricals

There are three common issues,

- **The number of levels is overwhelming**
- A single categorical might encode multiple pieces of information
- The levels might not be consolidated
- You might want to convert into numerical vectors



Managing types: Categoricals

There are three common issues,

- The number of levels is overwhelming
- **A single categorical might encode multiple pieces of information**
- The levels might not be consolidated
- You might want to convert into numerical vectors

name
Kris Sankaran

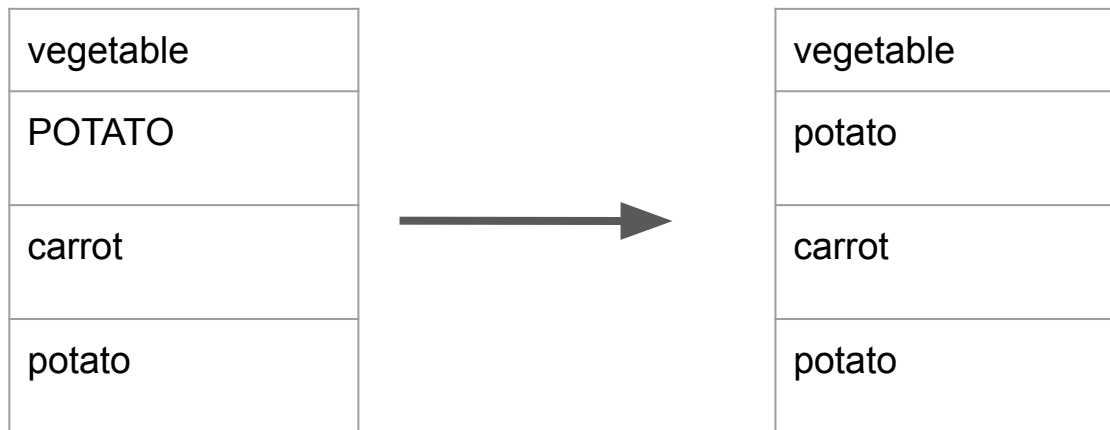


first_name	last_name
Kris	Sankaran

Managing types: Categoricals

There are three common issues,

- The number of levels is overwhelming
- A single categorical might encode multiple pieces of information
- **The levels might not be consolidated**
- You might want to convert into numerical vectors



Managing types: Categoricals

There are three common issues,

- The number of levels is overwhelming
- A single categorical might encode multiple pieces of information
- The levels might not be consolidated
- **You might want to convert into numerical vectors**

Happy?				
yes		yes	no	maybe
yes	1	0	0	
no	0	1	0	
maybe	0	0	1	
no	0	1	0	

Missing Values

- Not always properly read in
- Difference between structural and stochastic missingness
- Necessary for proper inference downstream

gsod1929		20 of 32 columns		Views	
A mo	A da	# temp	# count_temp	# dewp	# count_dev
The month	The day	Mean temperature for the day in degrees Fahrenheit to tenths. Missing = 9999.9	Number of observations used in calculating mean temperature	Mean dew point for the day in degrees Fahrenheit to tenths. Missing = 9999.9	Number of observations used in calculating dew point
10	03	49	4	41.7	
10	04	45.7	4	38.5	
10	07	48.2	4	44.8	
10	17	49.7	4	48.7	

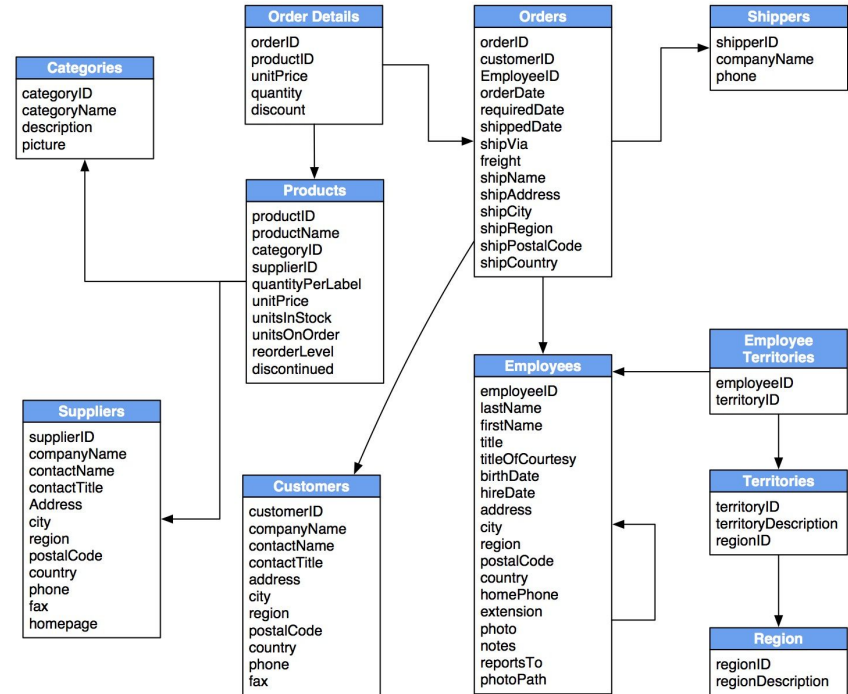
Joining, Reorganizing and Tidying

- We want a unified **X**
- May not happen because
 - The columns are stored across tables
 - The rows are written to different files
- May need to link to nontabular signals
 - Images, polygons, audio, ...

Joining, Reorganizing and Tidying

- We want a unified X
- May not happen because
 - **The columns are stored across tables**
 - The rows are written to different files
- May need to link to nontabular signals
 - Images, polygons, audio, ...

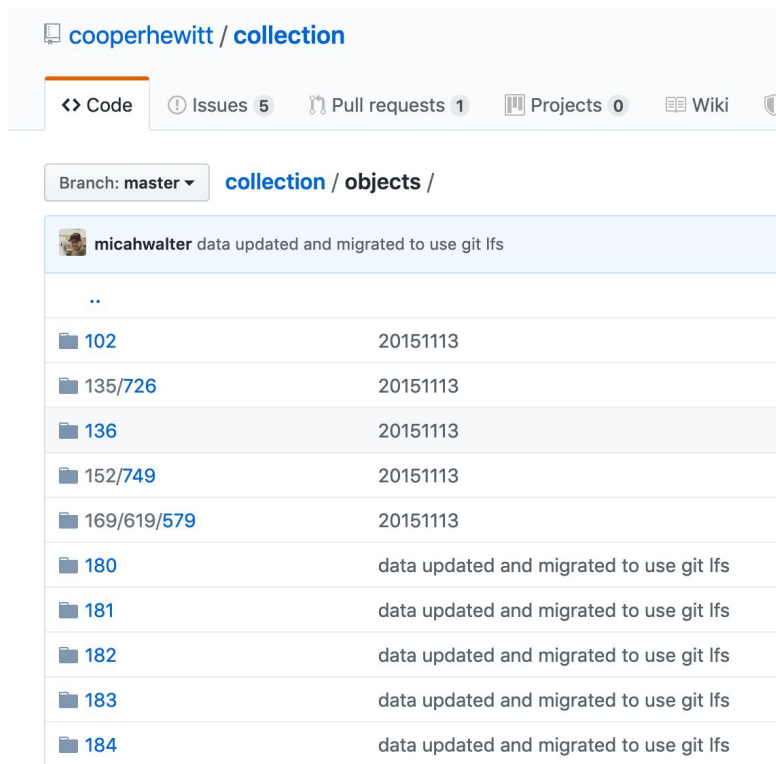
I sometimes have nightmares about relational databases.



Joining, Reorganizing and Tidying

- We want a unified **X**
- May not happen because
 - The columns are stored across tables
 - **The rows are written to different files**
- May need to link to nontabular signals
 - Images, polygons, audio, ...

Or if you're in really bad luck, across different directories.



The screenshot shows a GitHub repository page for 'cooperhewitt / collection'. The navigation bar includes 'Code', 'Issues 5', 'Pull requests 1', 'Projects 0', and 'Wiki'. The current branch is 'master' and the path is 'collection / objects /'. A commit by 'micahwalter' is shown with the message 'data updated and migrated to use git lfs'. Below this is a directory listing of files:

File Name	Commit Date	Commit Message
..		
102	20151113	
135/726	20151113	
136	20151113	
152/749	20151113	
169/619/579	20151113	
180		data updated and migrated to use git lfs
181		data updated and migrated to use git lfs
182		data updated and migrated to use git lfs
183		data updated and migrated to use git lfs
184		data updated and migrated to use git lfs

Joining, Reorganizing and Tidying

- We want a unified **X**
- May not happen because
 - The columns are stored across tables
 - The rows are written to different files
- **May need to link to nontabular signals**
 - Images, polygons, audio, ...

```
,width,height,channels,im_size,ctime,mtime,img_files img_folders,img_subfolders
0,1300,1300,1,3383838,1515131214.7275624,1511274026.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1510.tif
1,1300,1300,1,3383838,1515131210.4355597,1511274012.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img338.tif
2,1300,1300,1,3383838,1515131211.9995608,1511274005.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1468.tif
3,1300,1300,1,3383838,1515131223.5595682,1511274001.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img857.tif
4,1300,1300,1,3383838,1515131222.8075676,1511274003.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img86.tif
5,1300,1300,1,3383838,1515131215.503563,1511274006.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1276.tif
6,1300,1300,1,3383838,1515131223.479568,1511274002.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1582.tif
7,1300,1300,1,3383838,1515131223.0555677,1511274031.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img127.tif
8,1300,1300,1,3383838,1515131213.975562,1511274031.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1550.tif
9,1300,1300,1,3383838,1515131211.9235606,1511274013.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1447.tif
10,1300,1300,1,3383838,1515131213.0635614,1511274005.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img894.tif
11,1300,1300,1,3383838,1515131215.6915631,1511274005.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img741.tif
12,1300,1300,1,3383838,1515131210.0635595,1511274007.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img522.tif
13,1300,1300,1,3383838,1515131210.1115596,1511274002.0,AOI_2_Vegas_Roads_Test_Public,PAN,PAN_AOI_2_Vegas_img1151.tif
```

If this sounds painful...

- Be patient, “[the data are imperfect, as are we.](#)”

If this sounds painful...

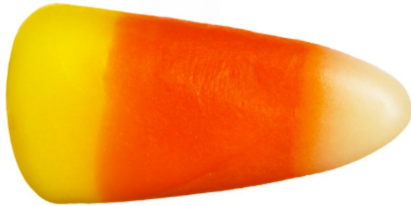
- Persist, it will get easier and you can do the fun stuff

Menu
Search

SCQ / THE SCIENCE CREATIVE QUARTERLY

SO MUCH CANDY DATA, SERIOUSLY

by DAVID NG



As promised, here is the candy hierarchy data for 2017. (Released Oct 25th @1:45pm PST. Will provide updated xlsx file on Oct 31st as well)

[xlsx](#) | [csv](#) | [txt \(d&t\)](#) | [surveyQ pdf](#) | n=2460

10.25.2017 - ACTIVITY / ARCHIVE / CLASSROOM / C

