How to deal with data?

IFT6758, Fall 2020; Lecture 1

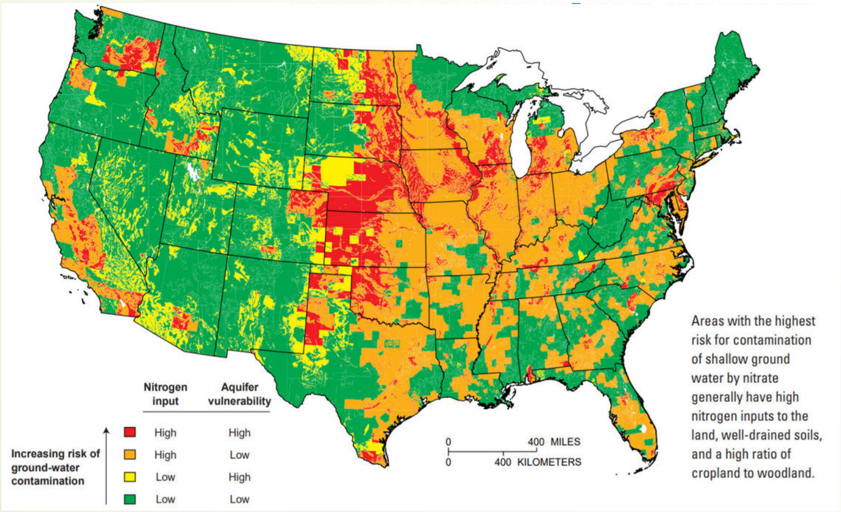
Data is variable

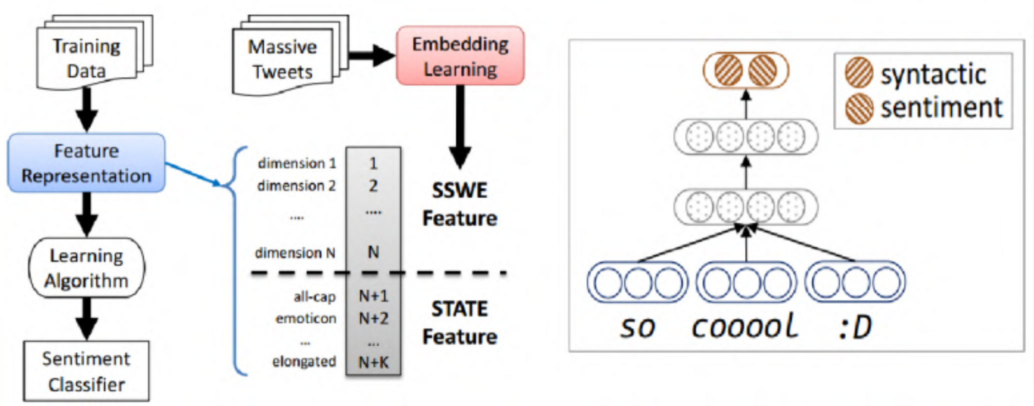
Data is variable

- Data are the result of deliberate human intervention

Data is variable

- Data are the result of deliberate human intervention
- Data is varied across domains and within domains





Data wrangling

Data wrangling

- Data (+ people who collect them) are varied

Data wrangling

- Data (+ people who collect them) are varied
  - Some amount of preparation is always needed.

Data wrangling

- Data (+ people who collect them) are varied
  - Some amount of preparation is always needed.
  - Example: tidying data, a small part of Data Cleaning process

Data wrangling

- Data (+ people who collect them) are varied
  - Some amount of preparation is always needed.
  - Example: tidying data, a small part of Data Cleaning process
    - Reading: How to share data with a statistician
      Tidy data

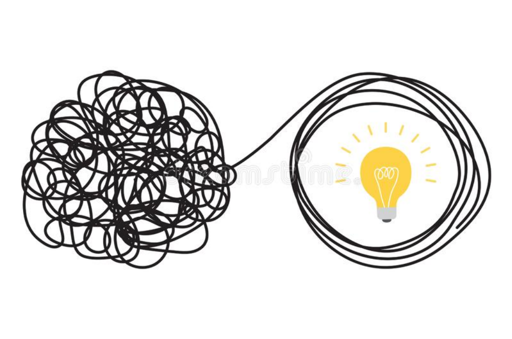
Data wrangling

- Data (+ people who collect them) are varied
  - Some amount of preparation is always needed.
  - Example: tidying data, a small part of Data Cleaning process
    - Reading: How to share data with a statistician

      Tidy data

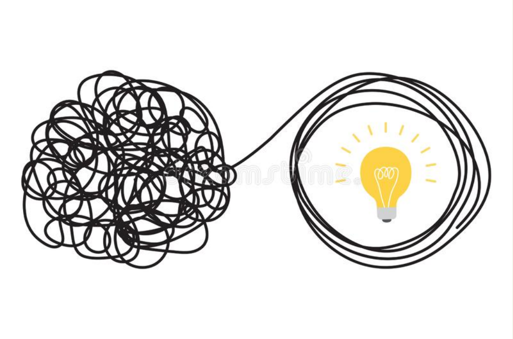
      Tidy data in Python

You just got some data



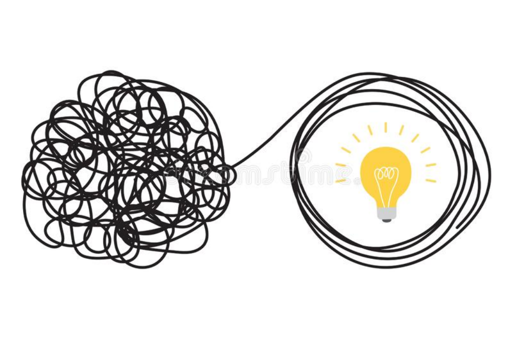
You just got some data

- Understand what the variables are



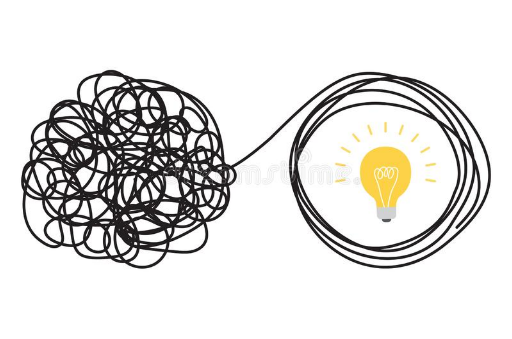
You just got some data

- Understand what the variables are
- Manage column types



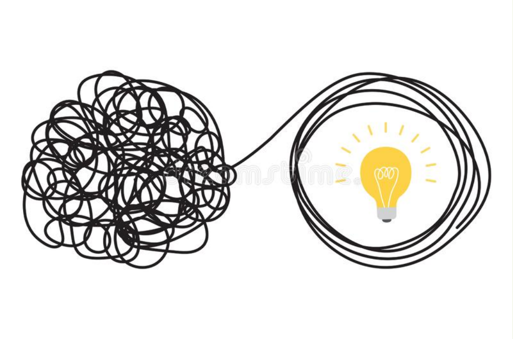
You just got some data

- Understand what the variables are
- Manage column types
- Handle missing values



You just got some data

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



Understand the data: Metadata

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

Understand the data: Metadata

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

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Cornelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	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	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Data

Metadata

Column	Data Type	Description
emlployee_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 postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
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. Null if employee sti

Managing types

Managing types

- Data come in different “types”

Managing types

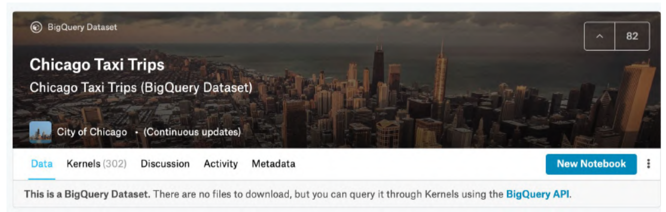
- Data come in different “types”
  - Numeric, (ordered) categorical, dates, (positive) integers

Managing types

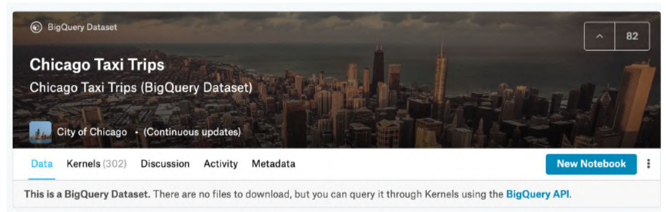
- Data come in different “types”
  - Numeric, (ordered) categorical, dates, (positive) integers
- Type should be (made) consistent with the purpose

Managing types

- Data come in different "types"
  - Numeric, (ordered) categorical, dates, (positive) integers
- Type should be (made) consistent with the purpose
  - Chicago Taxi Trips (BigQuery Dataset)
  - includes taxi trips (7000 licensed taxicabs) from 2013 to the present



Managing types




```
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
```

Managing types: Dates

- Use **Python datetime package** and pandas' timestamp and to_datetime
- Lets you convert **arbitrary strings into datetime objects**

Managing types: Dates

- Use **Python datetime package** and pandas' timestamp and to_datetime
- Lets you convert **arbitrary strings into datetime objects**

"22-01-2019T15:00:02" → datetime.datetime(2019, 1, 22, 15, 0, 2)

Managing types: Dates

- Use **Python datetime package** and pandas' timestamp and to_datetime
- Lets you convert **arbitrary strings into datetime objects**

"22-01-2019T15:00:02" → datetime.datetime(2019, 1, 22, 15, 0, 2)

Once it is in datetime format, new attributes can be derived

Managing types: Dates

- Use **Python datetime package** and pandas' timestamp and to_datetime
- Lets you convert **arbitrary strings into datetime objects**

"22-01-2019T15:00:02" → datetime.datetime(2019, 1, 22, 15, 0, 2)

Once it is in datetime format, new attributes can be derived

```
import datetime

x = datetime.datetime(2018, 6, 1)

print(x.strftime("%B"))
```

June

Managing types: Dates

- Use **Python datetime package** and pandas' timestamp and to_datetime
- Lets you convert **arbitrary strings into datetime objects**

"22-01-2019T15:00:02" → datetime.datetime(2019, 1, 22, 15, 0, 2)

Once it is in datetime format, new attributes can be derived

```
import datetime

x = datetime.datetime.now()

print(x.year)
print(x.strftime("%A"))
```

```
2020
Friday
```

Managing types: Categoricals

Managing types: Categoricals

Common issues

- Overwhelming number of levels

Managing types: Categoricals

Common issues

- Overwhelming number of levels

Managing types: Categoricals

Managing types: Categoricals

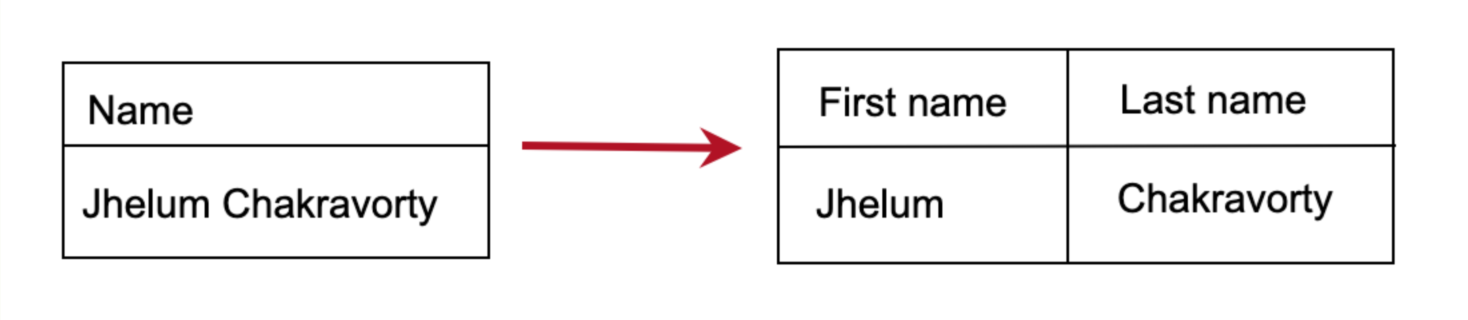
Common issues

- A single categorical might encode multiple pieces of information

Managing types: Categoricals

Common issues

- A single categorical might encode multiple pieces of information



Managing types: Categoricals

Managing types: Categoricals

Common Issues

- The levels might not be consolidated

Managing types: Categoricals

Common Issues

- The levels might not be consolidated

vegetable
POTATO
carrot
potato

vegetable
potato
carrot
potato

Managing types: Categoricals

Managing types: Categoricals

Common Issues

- You might want to convert into numerical vectors

Managing types: Categoricals

Common Issues

- You might want to convert into numerical vectors

Happy?
yes
yes
no
maybe
no

yes	no	maybe
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Missing values

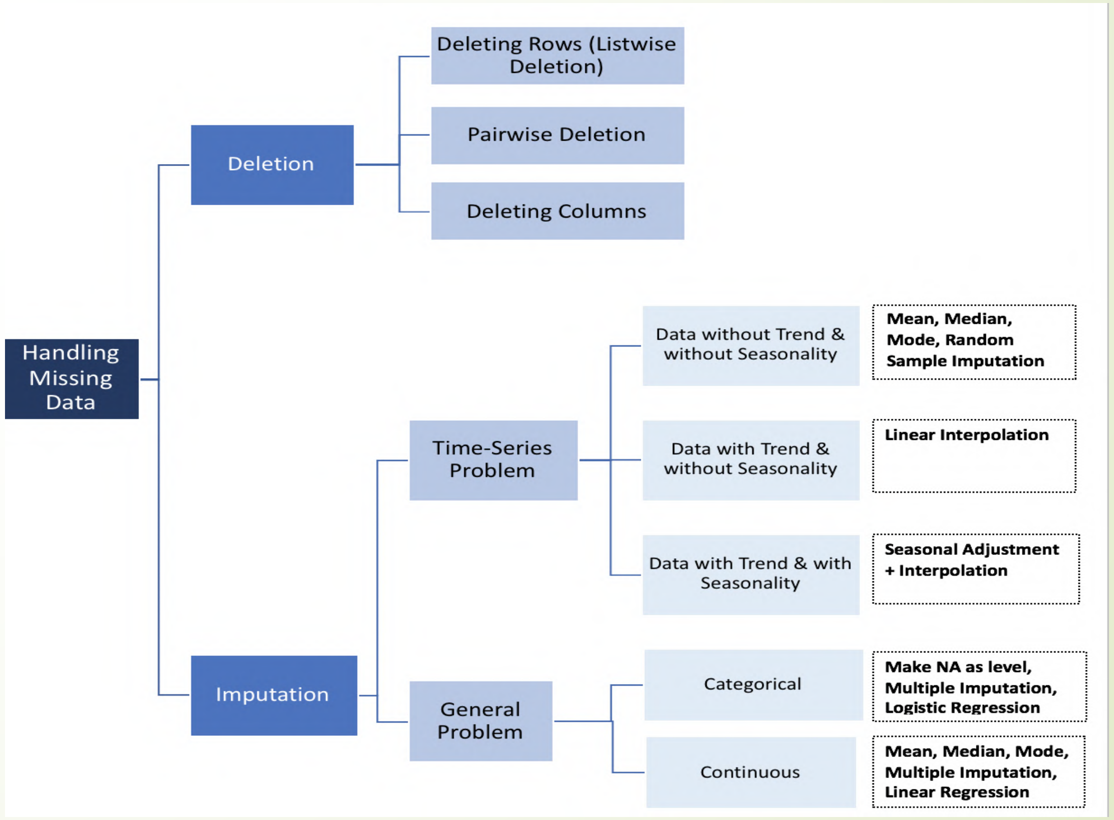
- Real-world data can be missing due to various reasons: e.g., observations that were not recorded and data corruption.
- **Handling missing values is important**. Many machine learning algorithms do not support data with missing values.
- Many ways.

Missing values

- Real-world data can be missing due to various reasons: e.g., observations that were not recorded and data corruption.
- **Handling missing values is important**. Many machine learning algorithms do not support data with missing values.
- Many ways.
  - Imputation and deletion

Missing values

- Real-world data can be missing due to various reasons: e.g., observations that were not recorded and data corruption.
- **Handling missing values is important**. Many machine learning algorithms do not support data with missing values.
- Many ways.
  - Imputation and deletion



Missing values

- Real-world data can be missing due to various reasons: e.g., observations that were not recorded and data corruption.
- **Handling missing values is important**. Many machine learning algorithms do not support data with missing values.
- Many ways.
  - Imputation and deletion
  - A useful tutorial

Joining, Reorganizing and Tidying

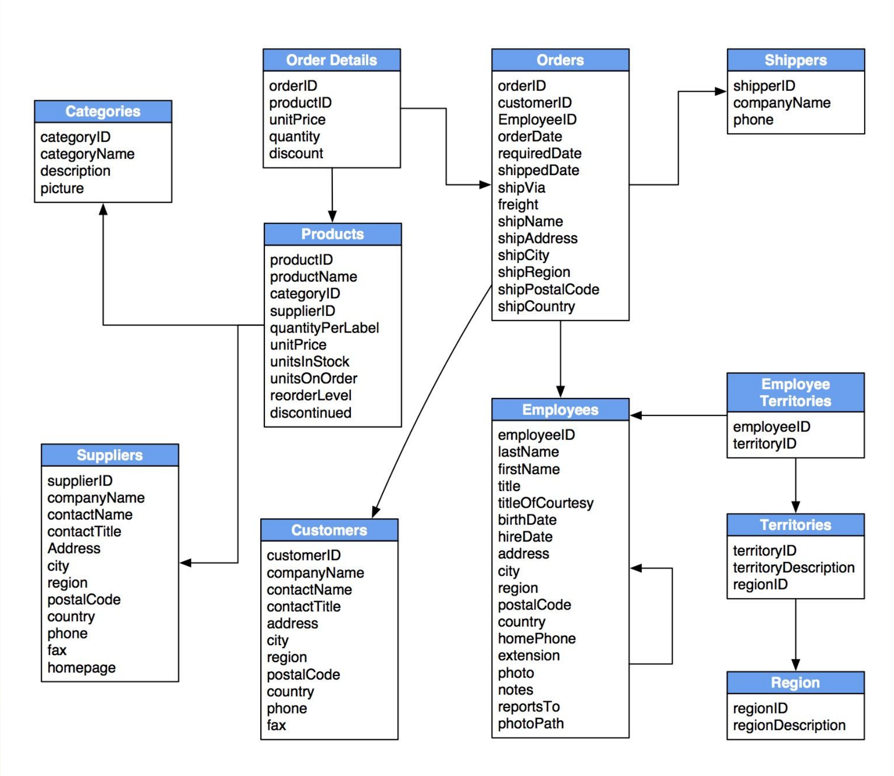
- Data might be available in messy forms

Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data

Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data

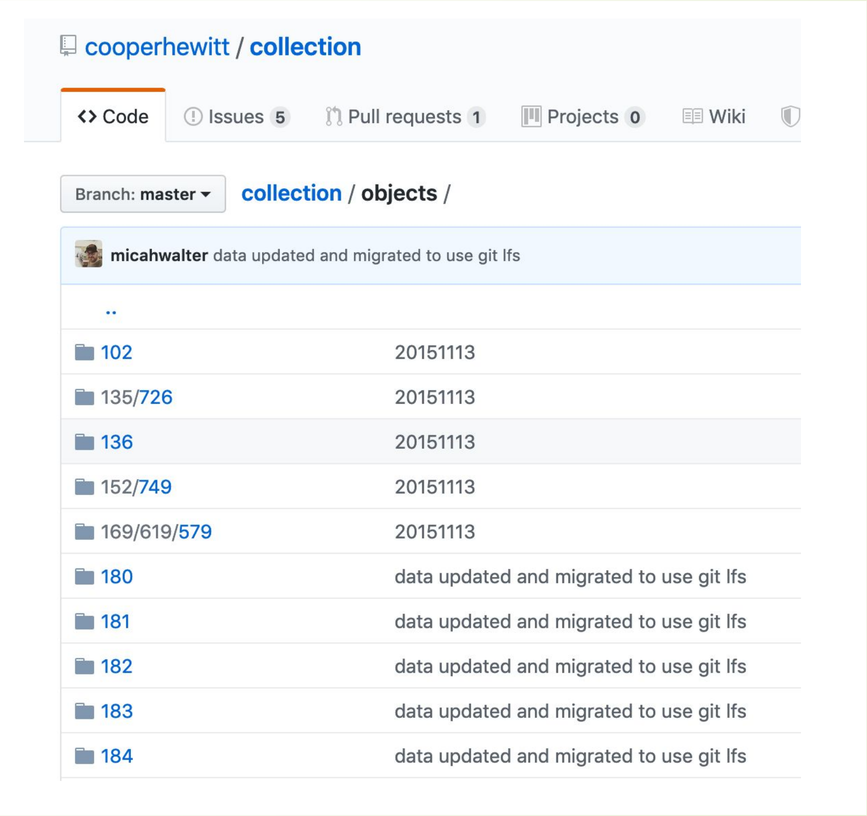


Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data
  - Rows are written to different files

Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data
  - Rows are written to different files



Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data
  - Rows are written to different files
  - May need to link to nontabular signals

Joining, Reorganizing and Tidying

- Data might be available in messy forms
  - Columns are stored across tables, relational data
  - Rows are written to different files
  - May need to link to nontabular signals

```
,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
```

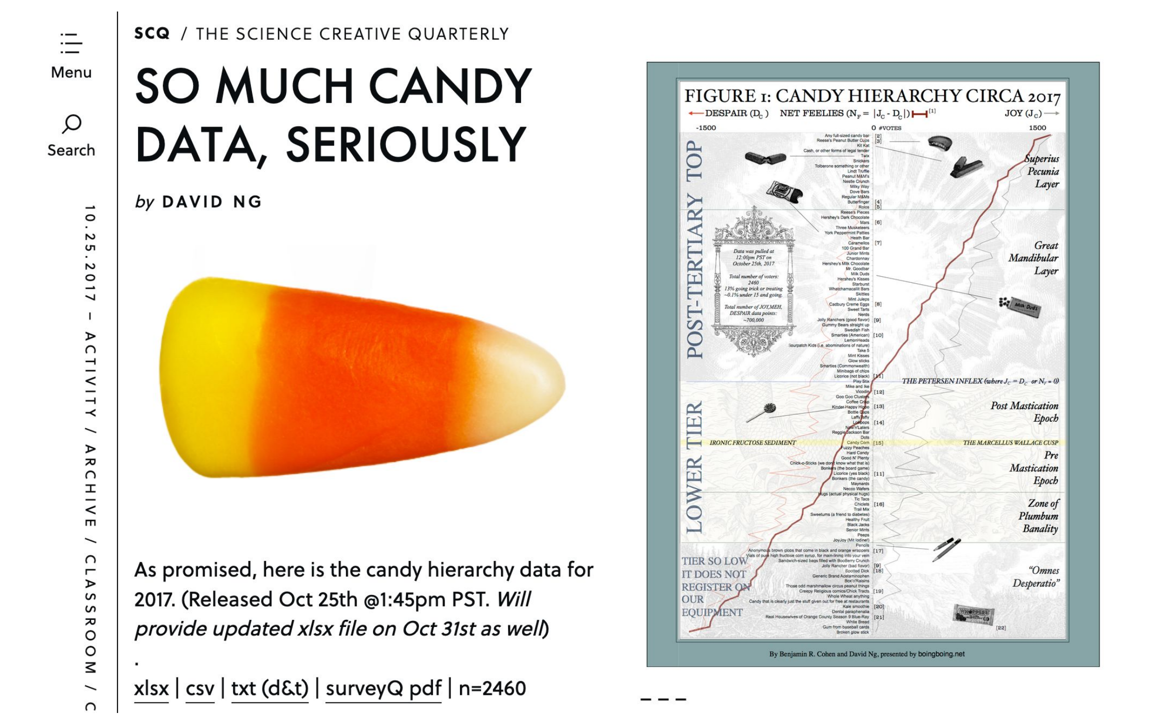
Painful? Intriguing?

Painful? Intriguing?

- Persist. There are so many datesets to have fun with.

Painful? Intriguing?

- Persist. There are so many datesets to have fun with.



Painful? Intriguing?

- Persist. There are so many datesets to have fun with.
- Embrace complexity and move forward