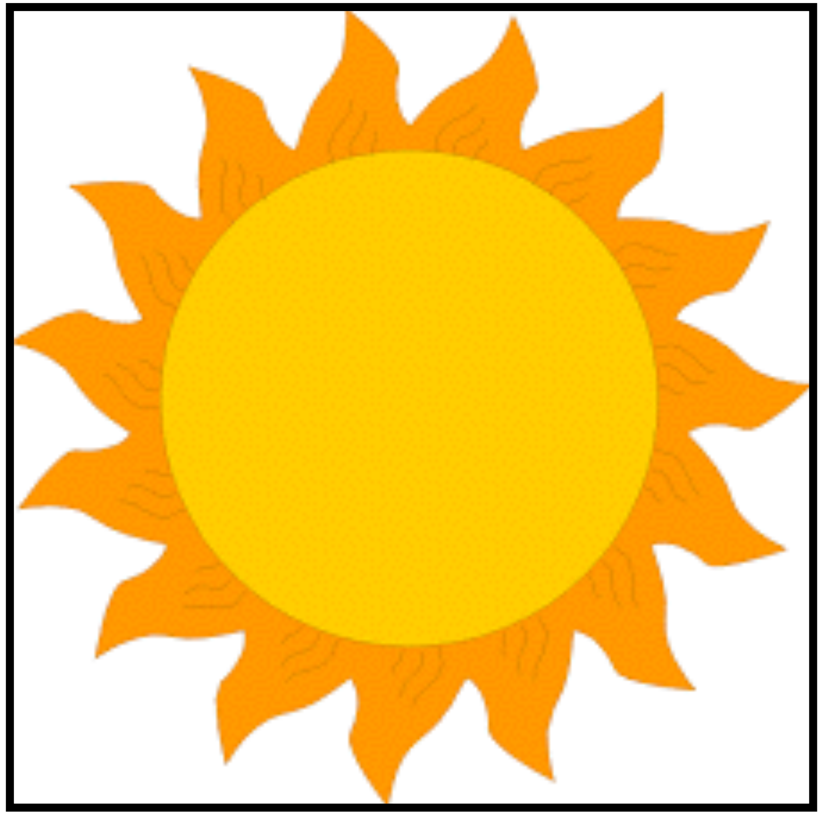
GENERAL ASSEMBLY

INTRODUCTION TO PANDAS

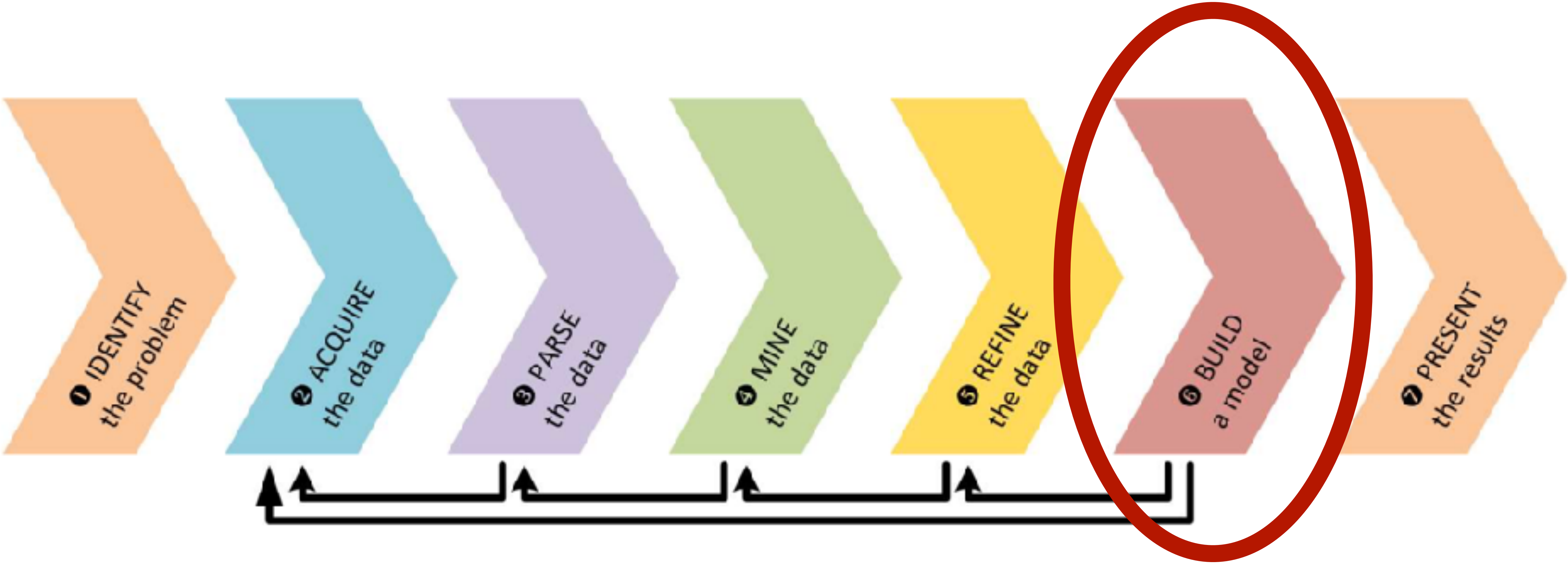
Gus Ostow

PANDAS OBJECTIVES

- Justify why we use Pandas instead of vanilla Python
- Explore data with DataFrames
- Perform rudimentary data cleaning



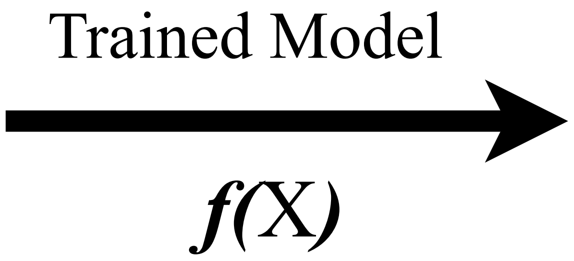
BACK TO THE WORKFLOW



TRAINING A MODEL TO MAKE PREDICTIONS

Feature Matrix, X

	Sqrft	# Bathrooms	Year Built
House #1	10,000	5	1988
House #2	6,200	2	2003
House #3	12,450	10	2014
House #4	850	0	2002



Trained Model

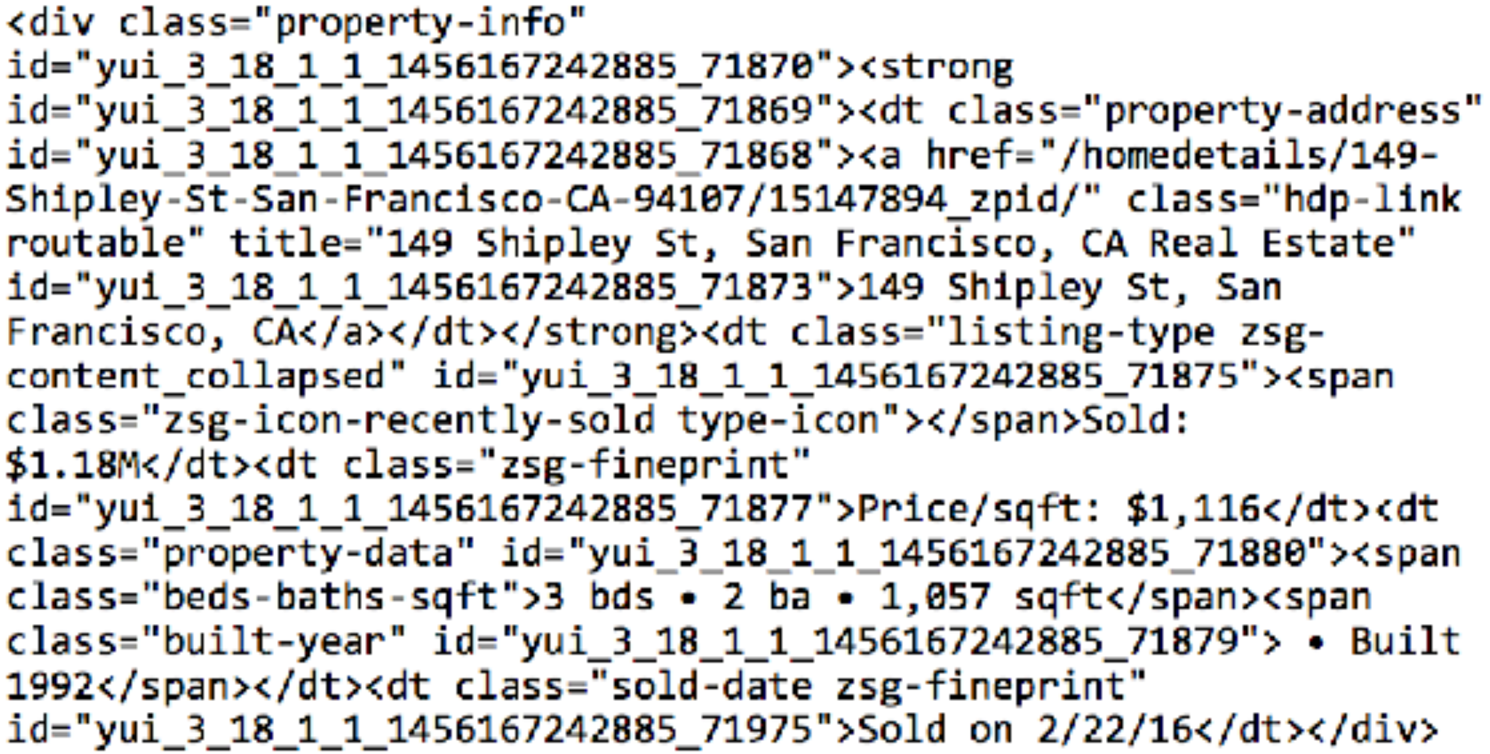
$f(X)$

Response Vector y

Sale Price
525K
384K
1.2M
74K

DATA ARRIVES UGLY

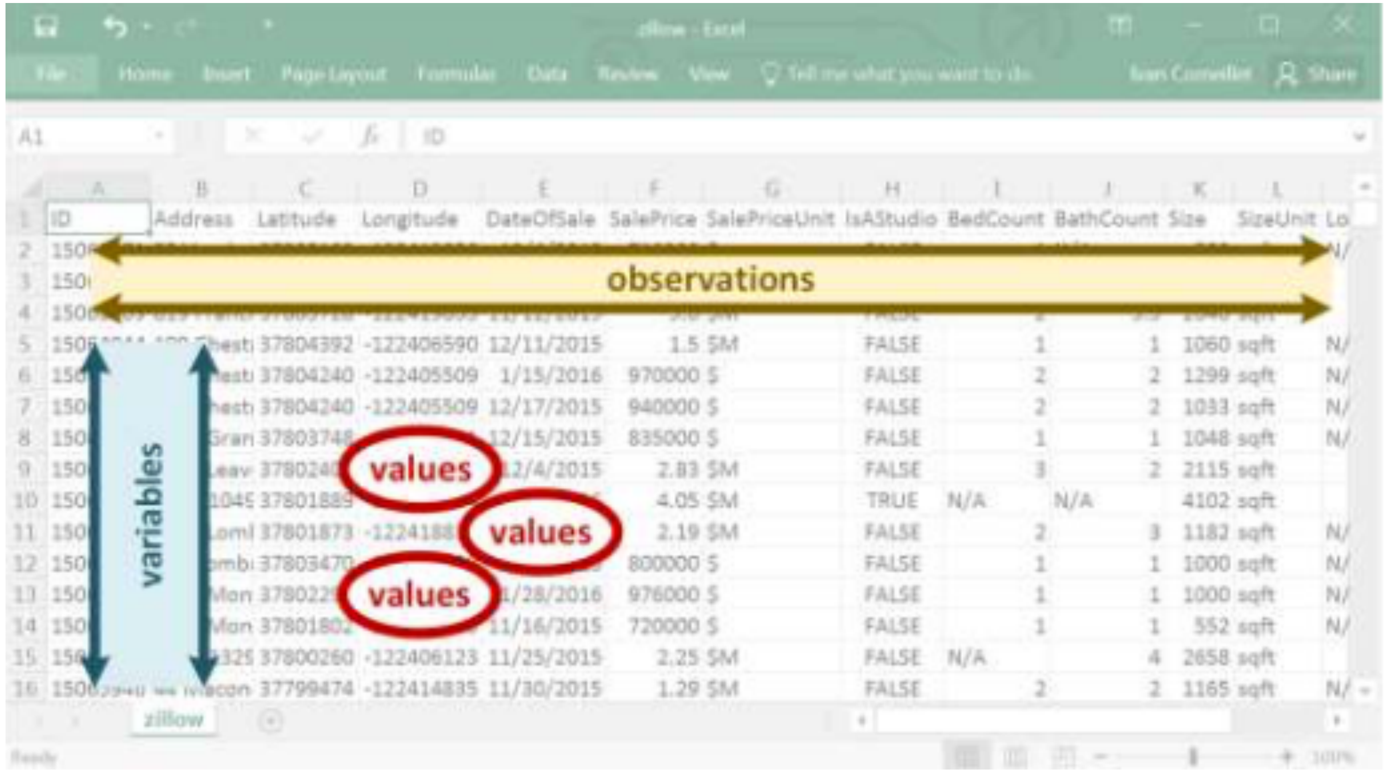
- Missing values
- Wrong data types
- Bad symbols
- Ambiguity
- Anything else you can imagine



```
<div class="property-info"
id="yui_3_18_1_1_1456167242885_71870"><strong
id="yui_3_18_1_1_1456167242885_71869"><dt class="property-address"
id="yui_3_18_1_1_1456167242885_71868"><a href="/homedetails/149-
Shipley-St-San-Francisco-CA-94107/15147894_zpid/" class="hdp-link
routable" title="149 Shipley St, San Francisco, CA Real Estate"
id="yui_3_18_1_1_1456167242885_71873">149 Shipley St, San
Francisco, CA</a></dt></strong><dt class="listing-type zsg-
content_collapsed" id="yui_3_18_1_1_1456167242885_71875"><span
class="zsg-icon-recently-sold type-icon"></span>Sold:
$1.18M</dt><dt class="zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71877">Price/sqft: $1,116</dt><dt
class="property-data" id="yui_3_18_1_1_1456167242885_71880"><span
class="beds-baths-sqft">3 bds • 2 ba • 1,057 sqft</span><span
class="built-year" id="yui_3_18_1_1_1456167242885_71879"> • Built
1992</span></dt><dt class="sold-date zsg-fineprint"
id="yui_3_18_1_1_1456167242885_71975">Sold on 2/22/16</dt></div>
```

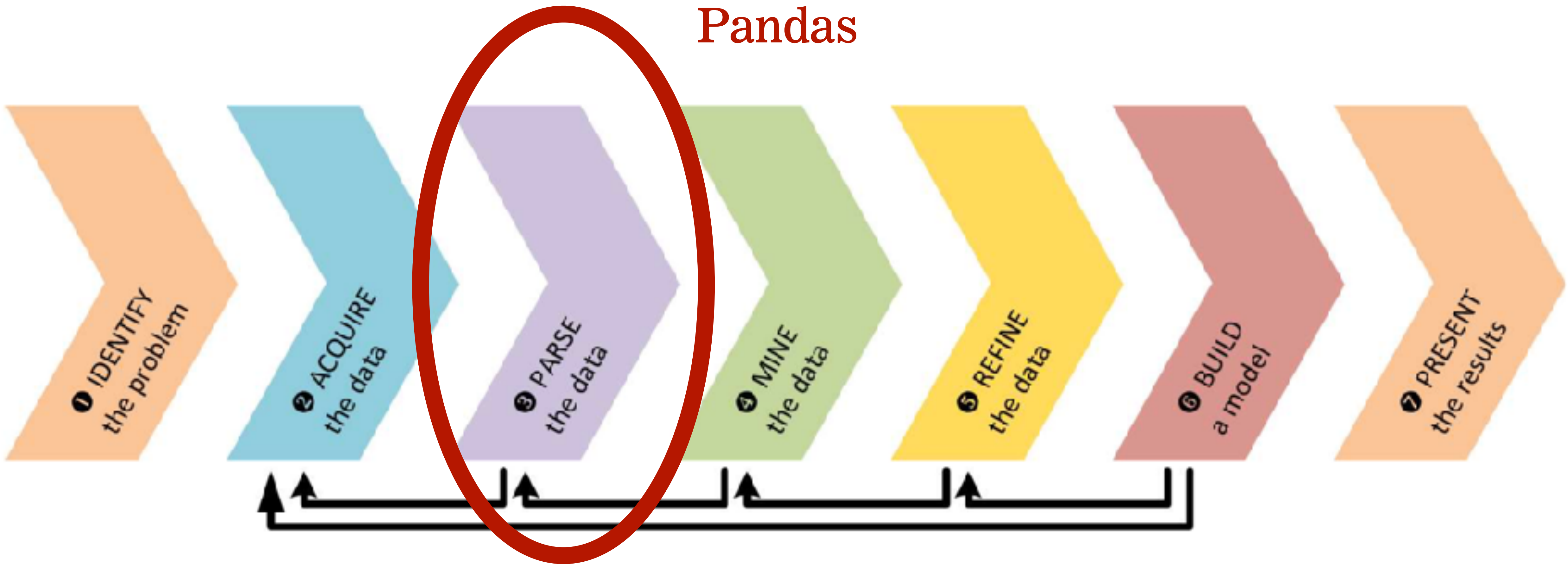
WHAT DOES CLEAN DATA LOOK LIKE?

- Each observation (aka sample) is represented by a single row
- Features are represented by a column
- One value per cell



	A	B	C	D	E	F	G	H	I	J	K	L	
1	ID	Address	Latitude	Longitude	DateOfSale	SalePrice	SalePriceUnit	IsAStudio	BedCount	BathCount	Size	SizeUnit	Lo
2	150	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	N/
3	150	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	150	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	150	hesti	37804392	-122406590	12/11/2015	1.5	$M	FALSE	1	1	1060	sqft	N/
6	150	esti	37804240	-122405509	1/15/2016	970000	$	FALSE	2	2	1299	sqft	N/
7	150	hesti	37804240	-122405509	12/17/2015	940000	$	FALSE	2	2	1033	sqft	N/
8	150	Gran	37803748	[illegible]	12/15/2015	835000	$	FALSE	1	1	1048	sqft	N/
9	150	Leav	3780240	[illegible]	12/4/2015	2.83	$M	FALSE	3	2	2115	sqft	
10	150	1045	37801889	[illegible]	[illegible]	4.05	$M	TRUE	N/A	N/A	4102	sqft	
11	150	Loml	37801873	-1224188	[illegible]	2.19	$M	FALSE	2	3	1182	sqft	N/
12	150	mbi	37803470	[illegible]	[illegible]	800000	$	FALSE	1	1	1000	sqft	N/
13	150	Mon	3780229	[illegible]	1/28/2016	976000	$	FALSE	1	1	1000	sqft	N/
14	150	Mon	37801802	[illegible]	11/16/2015	720000	$	FALSE	1	1	552	sqft	N/
15	150	325	37800260	-122406123	11/25/2015	2.25	$M	FALSE	N/A		4	2658	sqft	
16	150	Macon	37799474	-122414835	11/30/2015	1.29	$M	FALSE	2	2	1165	sqft	N/

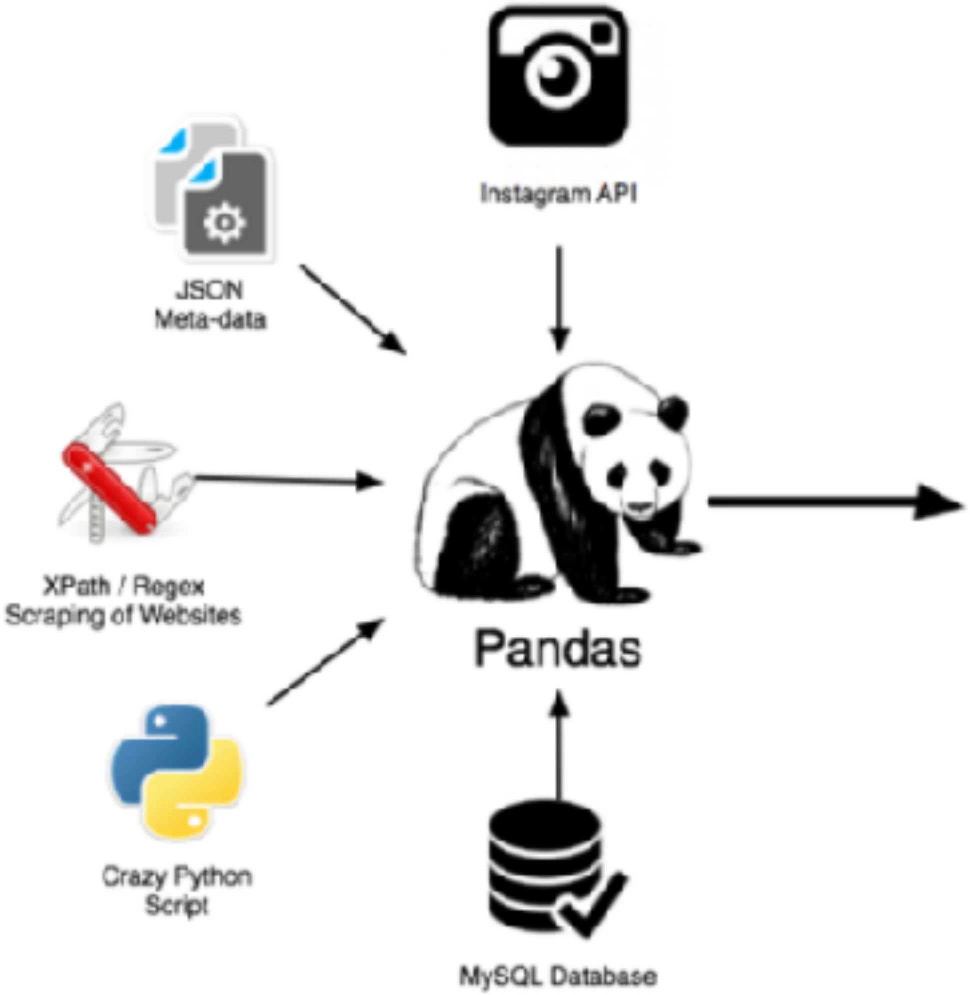
BACK TO THE WORKFLOW

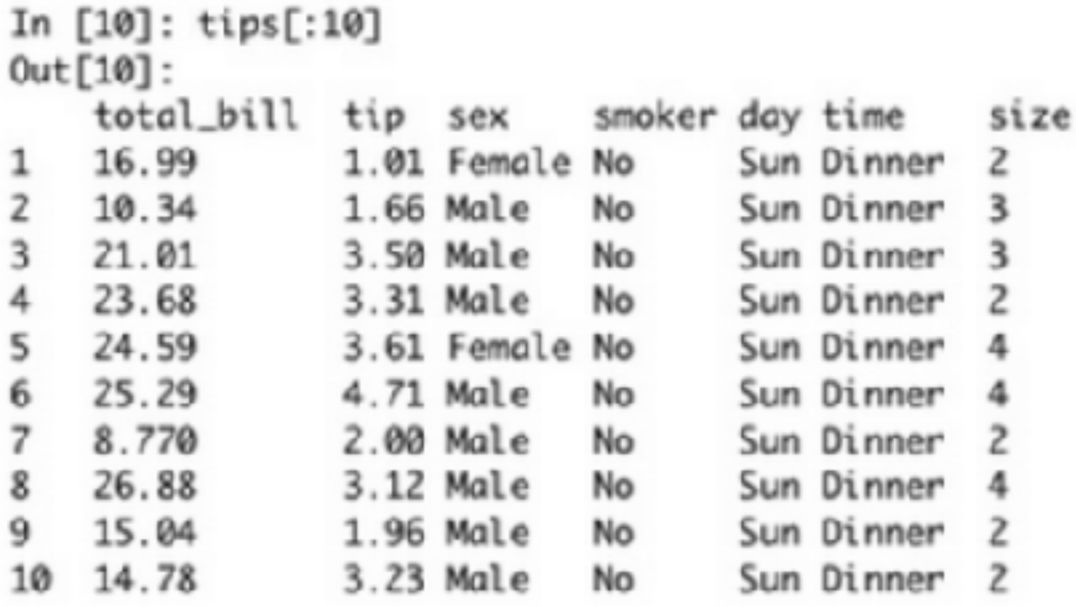


PANDAS

- Use cases
  - Exploration
  - Data cleaning
  - Transforming data
  - Joins
  - Filtering

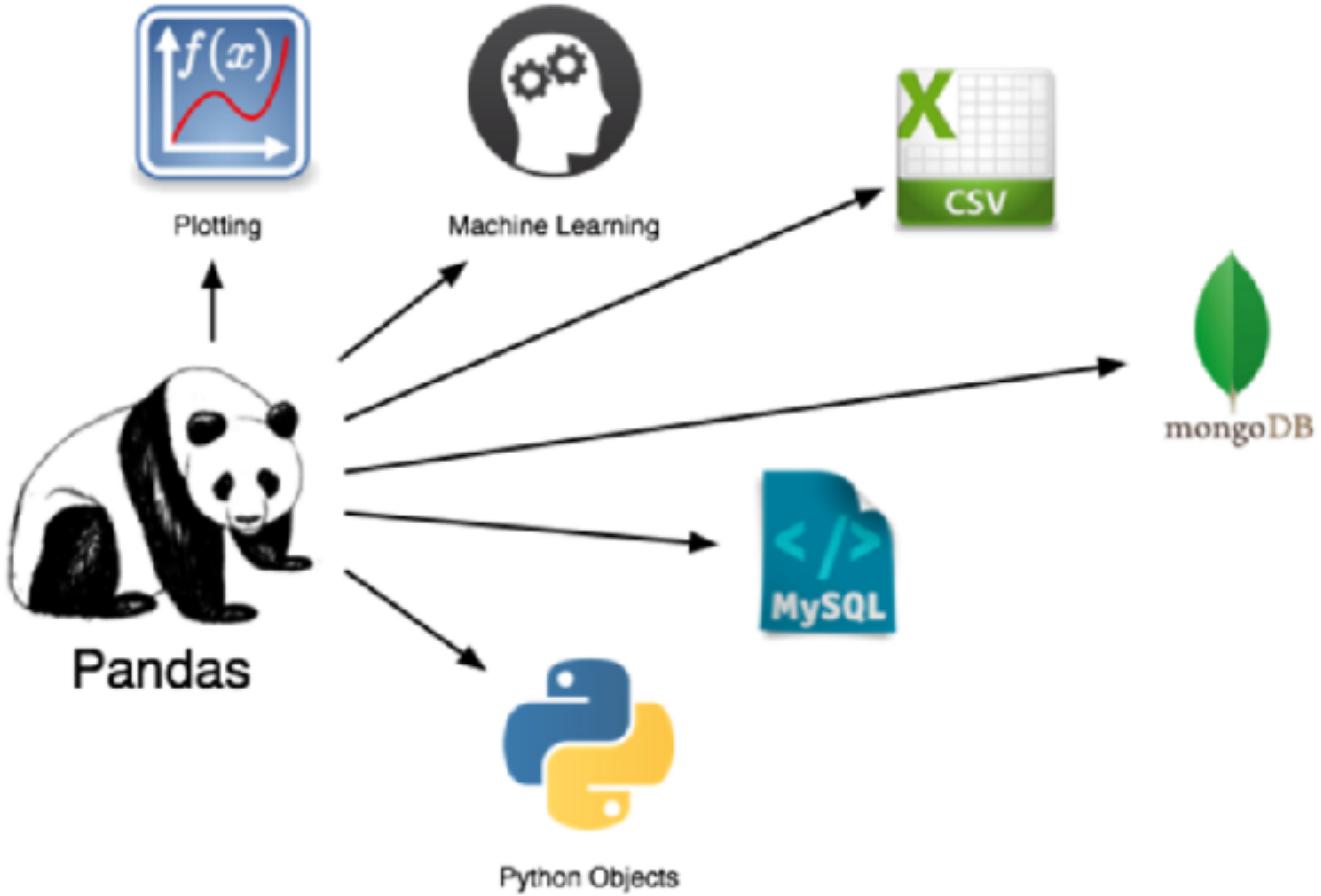
PANDAS





```
In [10]: tips[:10]
Out[10]:
     total_bill   tip  sex    smoker day time   size
1    16.99        1.01 Female No     Sun Dinner 2
2    10.34        1.66 Male   No     Sun Dinner 3
3    21.01        3.50 Male   No     Sun Dinner 3
4    23.68        3.31 Male   No     Sun Dinner 2
5    24.59        3.61 Female No     Sun Dinner 4
6    25.29        4.71 Male   No     Sun Dinner 4
7    8.770        2.00 Male   No     Sun Dinner 2
8    26.88        3.12 Male   No     Sun Dinner 4
9    15.04        1.96 Male   No     Sun Dinner 2
10   14.78        3.23 Male   No     Sun Dinner 2
```

PANDAS


f(x)
Plotting
Machine Learning
CSV
mongoDB
MySQL
Pandas
Python Objects

QUESTIONS?