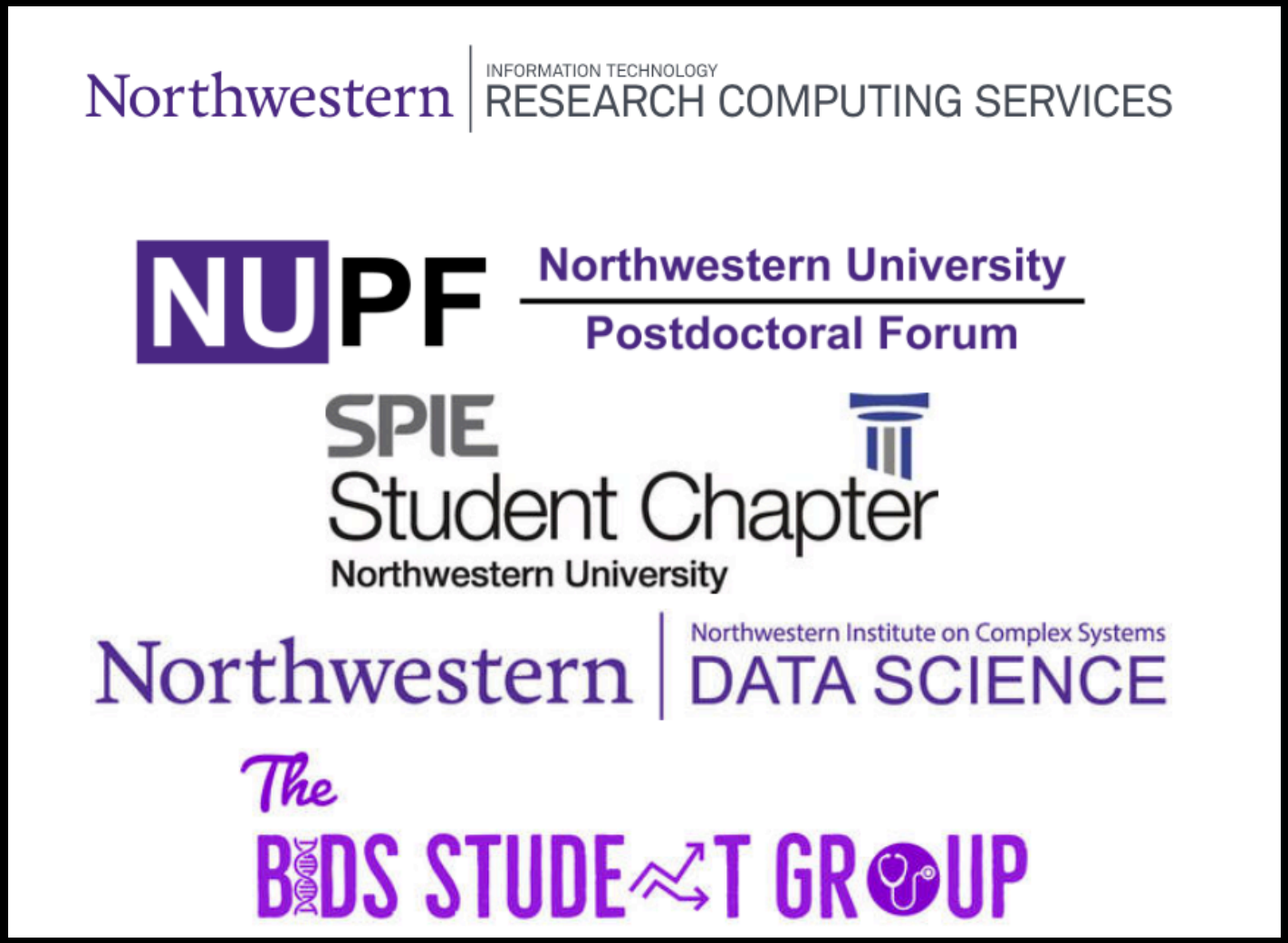
Data Strategies

+ Data Tactics

Thomas Stoeger
thomas.stoeger@northwestern.edu


Northwestern
INFORMATION TECHNOLOGY
RESEARCH COMPUTING SERVICES
NUPF
Northwestern University
Postdoctoral Forum
SPIE
Student Chapter
Northwestern University
Northwestern
Northwestern Institute on Complex Systems
DATA SCIENCE
The
BIDS STUDENT GROUP

CURRENT RESEARCH &
FUTURE CAREERS 2018

Home | Program | Registration | Abstract Submission | Contact

www.current-research-future-careers-2018.org

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment
(very short)

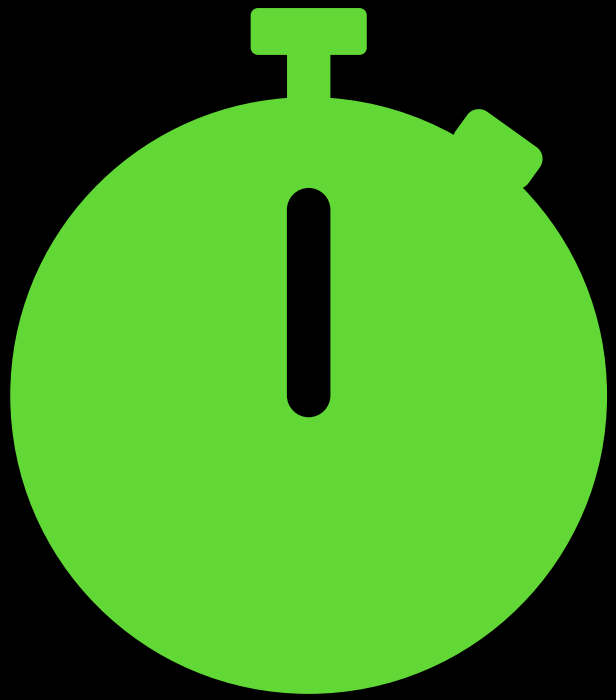
- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Summary
What did we miss?

After lecture (extra): getting your environment ready



Two minute exercise.

Find a neighbor, or neighbors, whom you do not know.

Then talk about the worst data you have encountered.

How can we anticipate problems?

“just”, “easy”, “simple”

Can not share data or code.

Understand the past history of the data.

External:

Read paper
or documentation

Internal:

Ask

"is generated by program"

Read documentation

Custom

Communication

Write consistency checks! (programmatically)

Expect to spend days or weeks

Tips for getting all information:

- be decent person
- remove fear that you will scoop

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Here

Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Summary

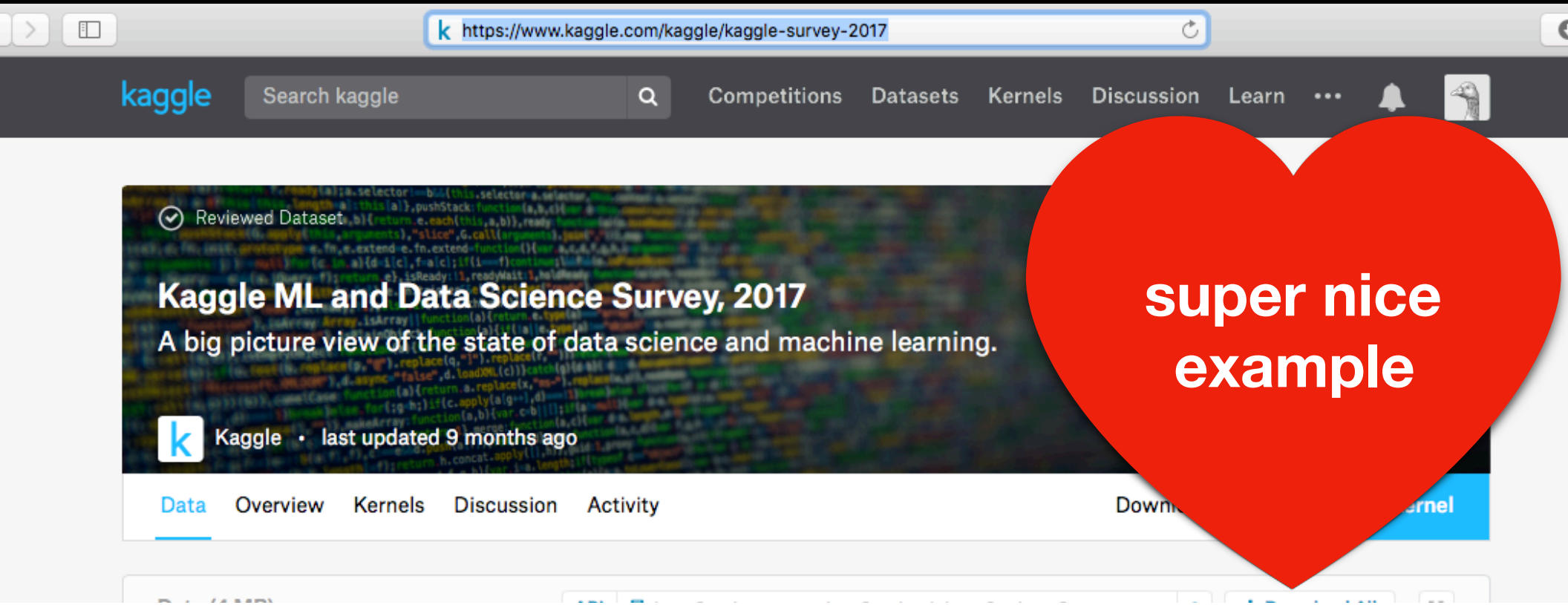
What did we miss?

After lecture (extra): getting your environment ready

Introducing

Kaggle ML and Data Science Survey, 2017

An example data set, which we will be using



Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science diffe s

Here

Will already learn some lessons

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Summary

What did we miss?

After lecture (extra): getting your environment ready

Name

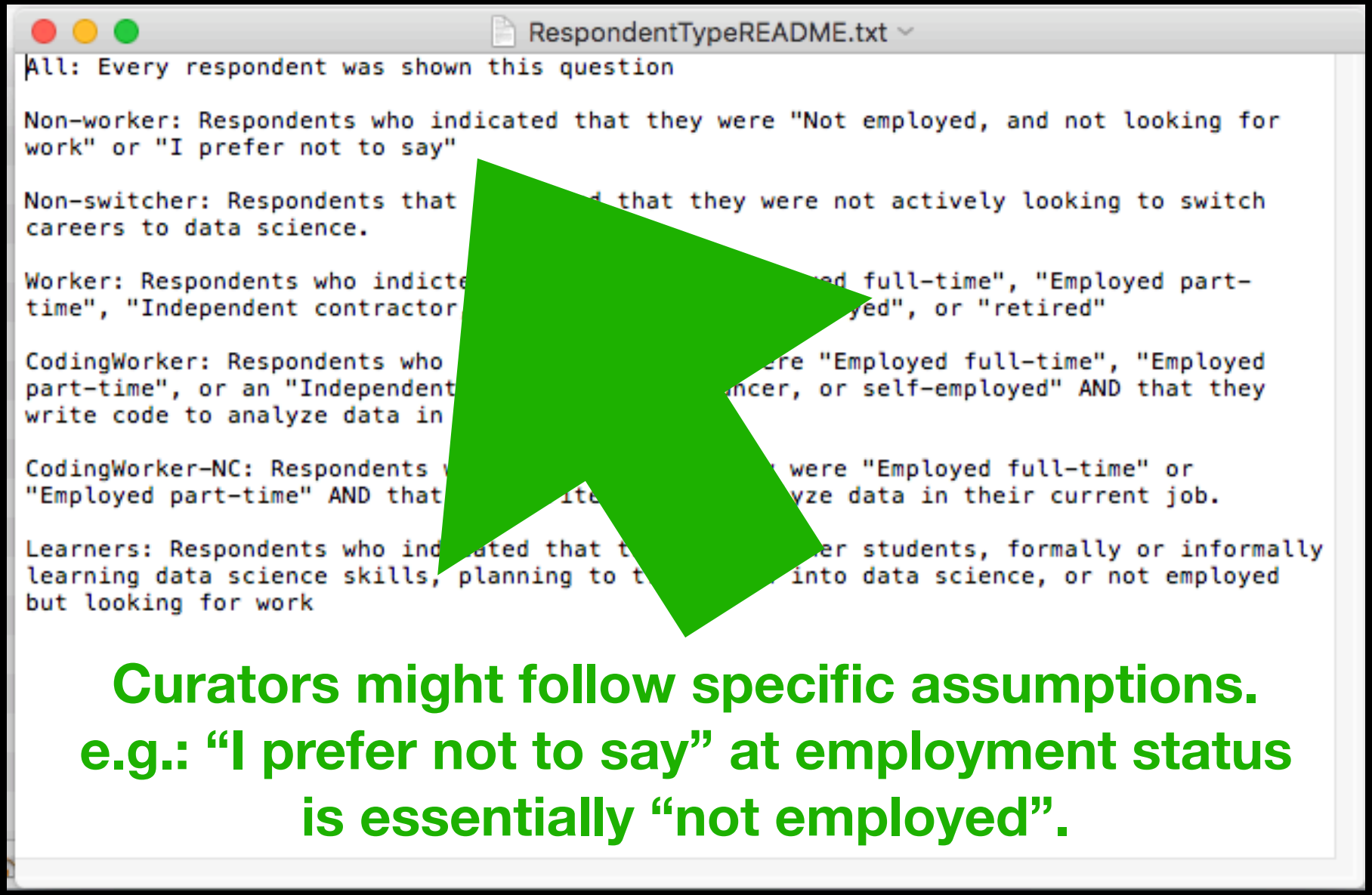
conversionRates.csv

freeformResponses.csv

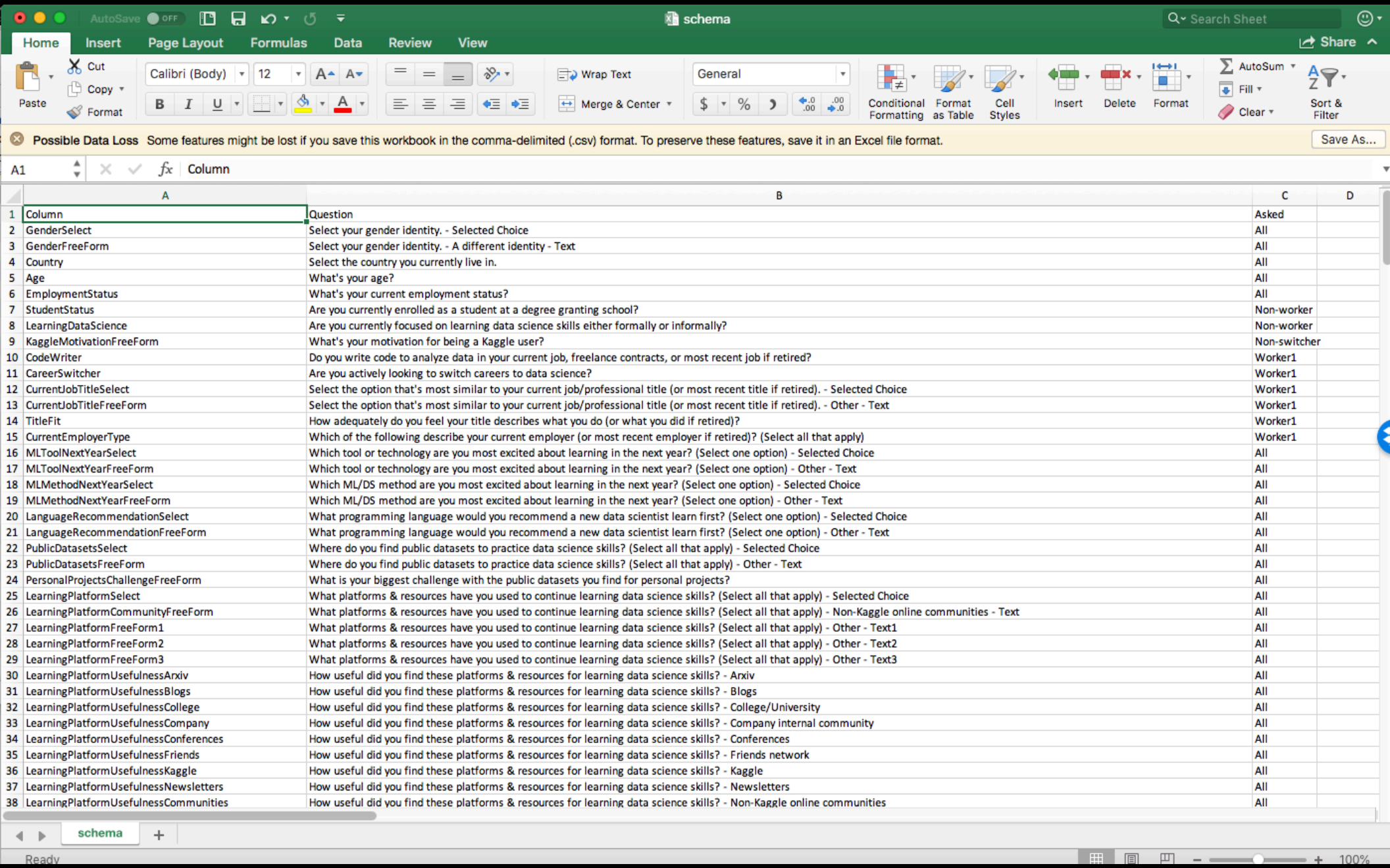
multipleChoiceResponses.csv

RespondentTypeREADME.txt

schema.csv

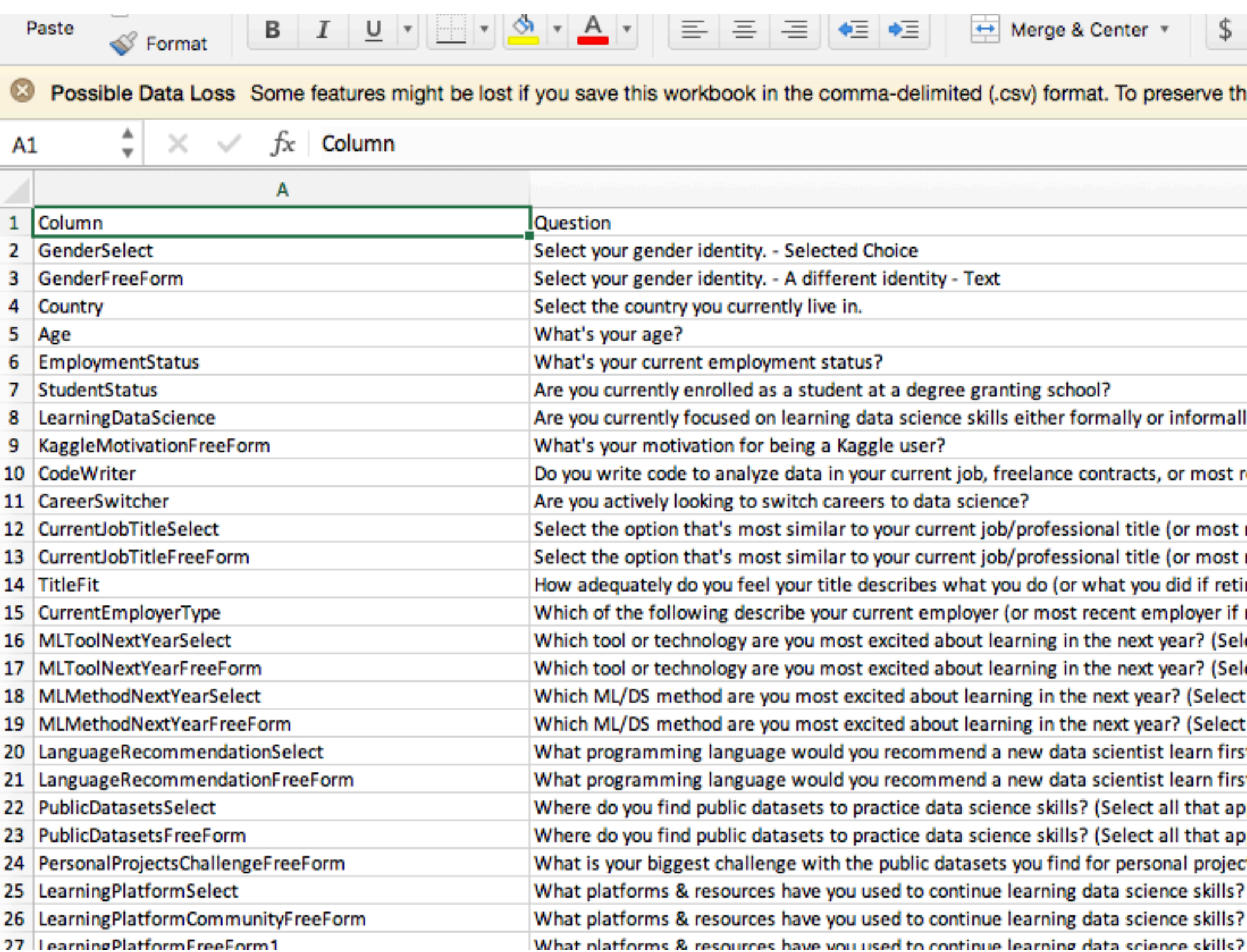


Understand terminology!

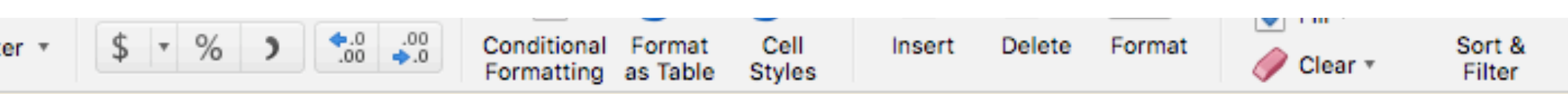


	A	B	C	D
1	Column	Question	Asked	
2	GenderSelect	Select your gender identity. - Selected Choice	All	
3	GenderFreeForm	Select your gender identity. - A different identity - Text	All	
4	Country	Select the country you currently live in.	All	
5	Age	What's your age?	All	
6	EmploymentStatus	What's your current employment status?	All	
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?	Non-worker	
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informally?	Non-worker	
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?	Non-switcher	
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most recent job if retired?	Worker1	
11	CareerSwitcher	Are you actively looking to switch careers to data science?	Worker1	
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Selected Choice	Worker1	
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most recent title if retired). - Other - Text	Worker1	
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if retired)?	Worker1	
15	CurrentEmployerType	Which of the following describe your current employer (or most recent employer if retired)? (Select all that apply)	Worker1	
16	MLToolNextYearSelect	Which tool or technology are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
17	MLToolNextYearFreeForm	Which tool or technology are you most excited about learning in the next year? (Select one option) - Other - Text	All	
18	MLMethodNextYearSelect	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Selected Choice	All	
19	MLMethodNextYearFreeForm	Which ML/DS method are you most excited about learning in the next year? (Select one option) - Other - Text	All	
20	LanguageRecommendationSelect	What programming language would you recommend a new data scientist learn first? (Select one option) - Selected Choice	All	
21	LanguageRecommendationFreeForm	What programming language would you recommend a new data scientist learn first? (Select one option) - Other - Text	All	
22	PublicDatasetsSelect	Where do you find public datasets to practice data science skills? (Select all that apply) - Selected Choice	All	
23	PublicDatasetsFreeForm	Where do you find public datasets to practice data science skills? (Select all that apply) - Other - Text	All	
24	PersonalProjectsChallengeFreeForm	What is your biggest challenge with the public datasets you find for personal projects?	All	
25	LearningPlatformSelect	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Selected Choice	All	
26	LearningPlatformCommunityFreeForm	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Non-Kaggle online communities - Text	All	
27	LearningPlatformFreeForm1	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text1	All	
28	LearningPlatformFreeForm2	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text2	All	
29	LearningPlatformFreeForm3	What platforms & resources have you used to continue learning data science skills? (Select all that apply) - Other - Text3	All	
30	LearningPlatformUsefulnessArxiv	How useful did you find these platforms & resources for learning data science skills? - Arxiv	All	
31	LearningPlatformUsefulnessBlogs	How useful did you find these platforms & resources for learning data science skills? - Blogs	All	
32	LearningPlatformUsefulnessCollege	How useful did you find these platforms & resources for learning data science skills? - College/University	All	
33	LearningPlatformUsefulnessCompany	How useful did you find these platforms & resources for learning data science skills? - Company internal community	All	
34	LearningPlatformUsefulnessConferences	How useful did you find these platforms & resources for learning data science skills? - Conferences	All	
35	LearningPlatformUsefulnessFriends	How useful did you find these platforms & resources for learning data science skills? - Friends network	All	
36	LearningPlatformUsefulnessKaggle	How useful did you find these platforms & resources for learning data science skills? - Kaggle	All	
37	LearningPlatformUsefulnessNewsletters	How useful did you find these platforms & resources for learning data science skills? - Newsletters	All	
38	LearningPlatformUsefulnessCommunities	How useful did you find these platforms & resources for learning data science skills? - Non-Kaggle online communities	All	

Understand possibilities!



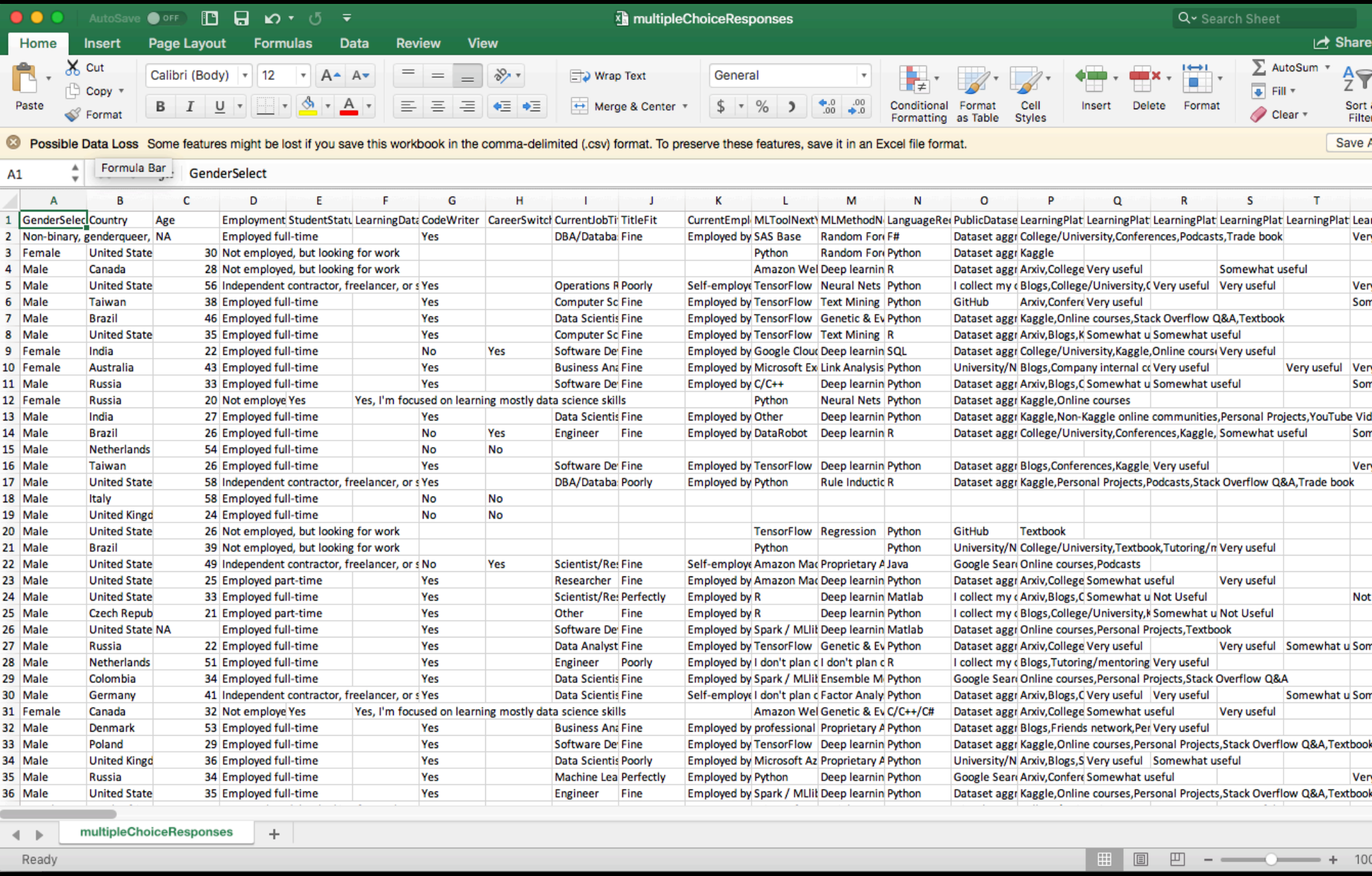
	A	
1	Column	Question
2	GenderSelect	Select your gender identity. - Selected Choice
3	GenderFreeForm	Select your gender identity. - A different identity - Text
4	Country	Select the country you currently live in.
5	Age	What's your age?
6	EmploymentStatus	What's your current employment status?
7	StudentStatus	Are you currently enrolled as a student at a degree granting school?
8	LearningDataScience	Are you currently focused on learning data science skills either formally or informall
9	KaggleMotivationFreeForm	What's your motivation for being a Kaggle user?
10	CodeWriter	Do you write code to analyze data in your current job, freelance contracts, or most r
11	CareerSwitcher	Are you actively looking to switch careers to data science?
12	CurrentJobTitleSelect	Select the option that's most similar to your current job/professional title (or most
13	CurrentJobTitleFreeForm	Select the option that's most similar to your current job/professional title (or most
14	TitleFit	How adequately do you feel your title describes what you do (or what you did if reti
15	CurrentEmployerType	Which of the following describe your current employer (or most recent employer if
16	MLToolNextYearSelect	Which tool or technology are you most excited about learning in the next year? (Sel
17	MLToolNextYearFreeForm	Which tool or technology are you most excited about learning in the next year? (Sel
18	MLMethodNextYearSelect	Which ML/DS method are you most excited about learning in the next year? (Select
19	MLMethodNextYearFreeForm	Which ML/DS method are you most excited about learning in the next year? (Select
20	LanguageRecommendationSelect	What programming language would you recommend a new data scientist learn first
21	LanguageRecommendationFreeForm	What programming language would you recommend a new data scientist learn first
22	PublicDatasetsSelect	Where do you find public datasets to practice data science skills? (Select all that ap
23	PublicDatasetsFreeForm	Where do you find public datasets to practice data science skills? (Select all that ap
24	PersonalProjectsChallengeFreeForm	What is your biggest challenge with the public datasets you find for personal projec
25	LearningPlatformSelect	What platforms & resources have you used to continue learning data science skills?
26	LearningPlatformCommunityFreeForm	What platforms & resources have you used to continue learning data science skills?
27	LearningPlatformFreeForm1	What platforms & resources have you used to continue learning data science skills?



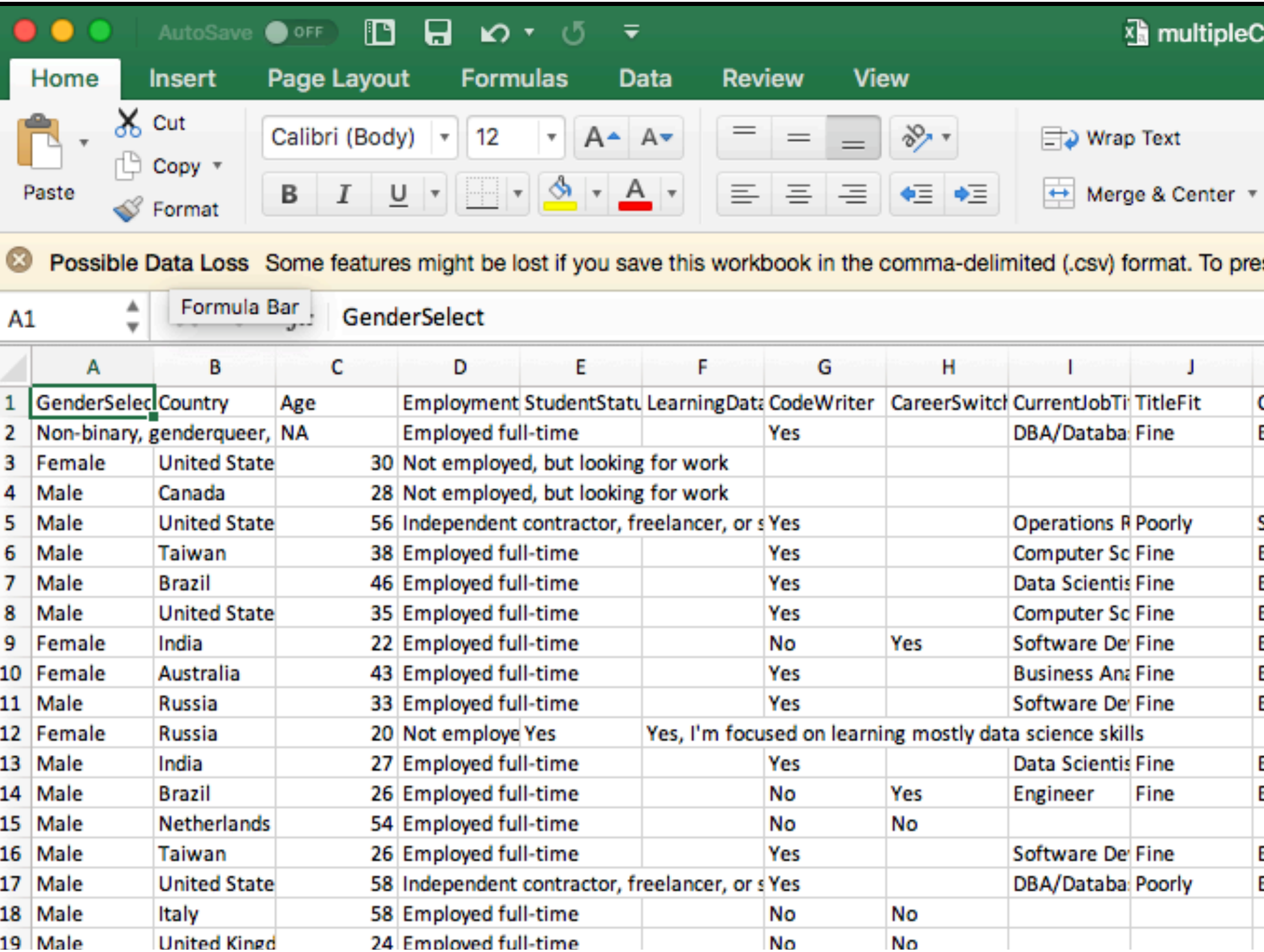
o preserve these features, save it in an Excel file format.

Save As

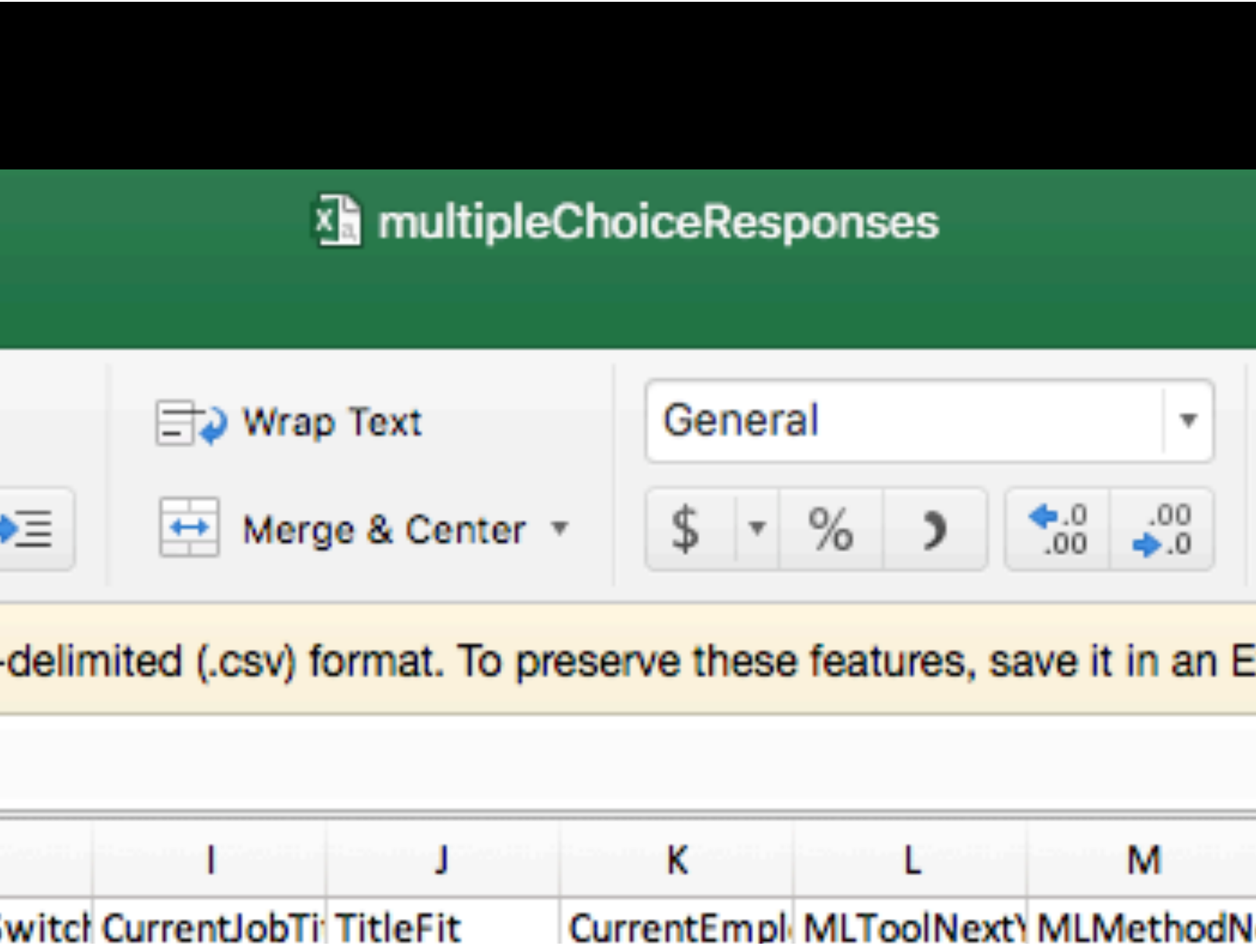
B	C	D
	Asked	
	All	
	All	
	All	
	All	
	All	
	Non-worker	
ly or informally?	Non-worker	
	Non-switcher	
cts, or most recent job if retired?	Worker1	
	Worker1	
title (or most recent title if retired). - Selected Choice	Worker1	
title (or most recent title if retired). - Other - Text	Worker1	
you did if retired)?	Worker1	
t employer if retired)? (Select all that apply)	Worker1	
ext year? (Select one option) - Selected Choice	All	
ext year? (Select one option) - Other - Text	All	
year? (Select one option) - Selected Choice	All	
year? (Select one option) - Other - Text	All	
ntist learn first? (Select one option) - Selected Choice	All	
ntist learn first? (Select one option) - Other - Text	All	
ect all that apply) - Selected Choice	All	
ect all that apply) - Other - Text	All	
ersonal projects?	All	
science skills? (Select all that apply) - Selected Choice	All	
science skills? (Select all that apply) - Non-Kaggle online communities - Text	All	

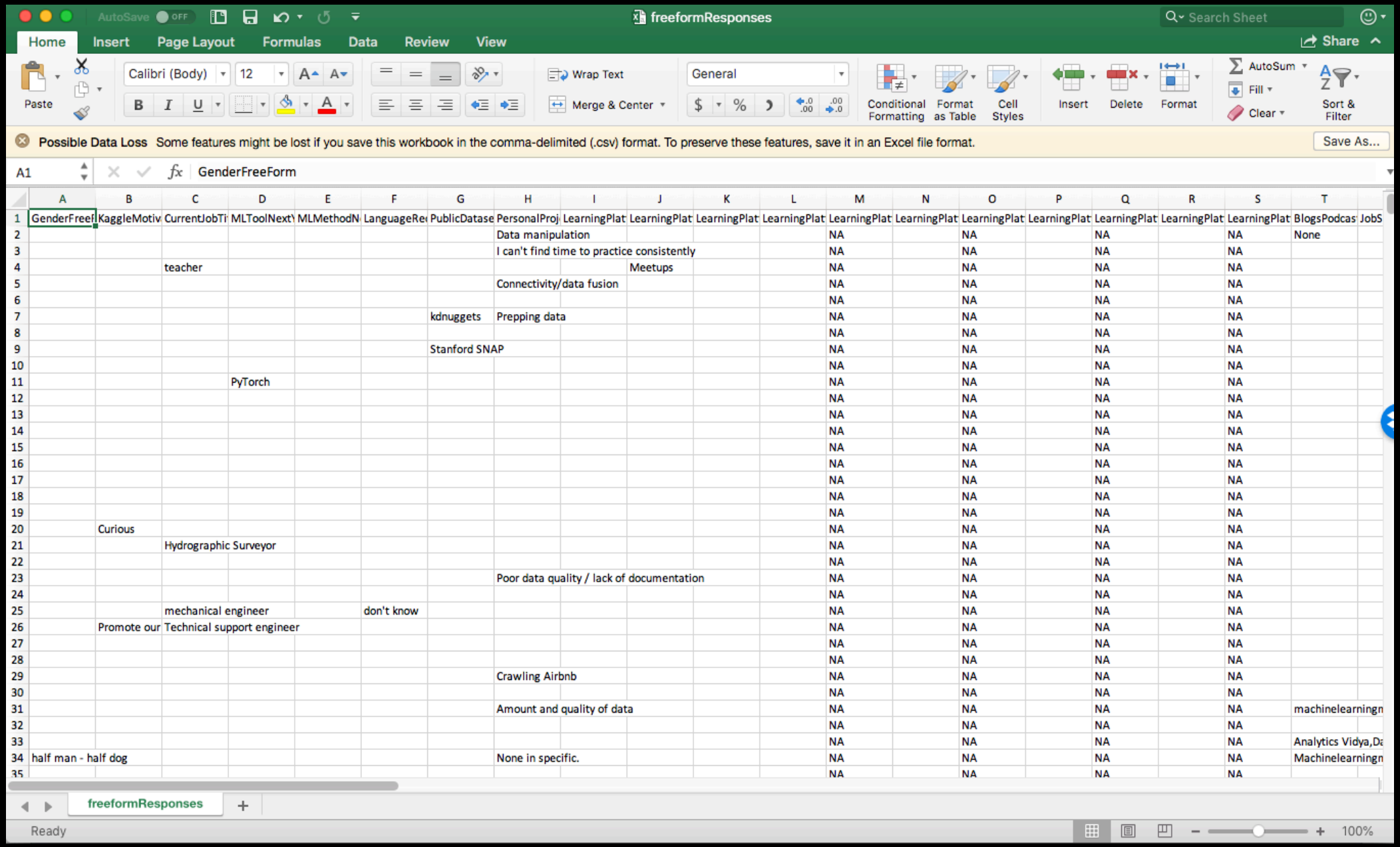


Some answers are well defined (multiple choice).



	A	B	C	D	E	F	G	H	I	J
1	GenderSelec	Country	Age	Employment	StudentStatu	LearningData	CodeWriter	CareerSwitcl	CurrentJobTi	TitleFit
2	Non-binary, genderqueer,		NA	Employed full-time			Yes		DBA/Databa	Fine
3	Female	United State	30	Not employed, but looking for work						
4	Male	Canada	28	Not employed, but looking for work						
5	Male	United State	56	Independent contractor, freelancer, or s			Yes		Operations R	Poorly
6	Male	Taiwan	38	Employed full-time			Yes		Computer Sc	Fine
7	Male	Brazil	46	Employed full-time			Yes		Data Scientis	Fine
8	Male	United State	35	Employed full-time			Yes		Computer Sc	Fine
9	Female	India	22	Employed full-time			No	Yes	Software De	Fine
10	Female	Australia	43	Employed full-time			Yes		Business Ana	Fine
11	Male	Russia	33	Employed full-time			Yes		Software De	Fine
12	Female	Russia	20	Not employe	Yes	Yes, I'm focused on learning mostly data science skills				
13	Male	India	27	Employed full-time			Yes		Data Scientis	Fine
14	Male	Brazil	26	Employed full-time			No	Yes	Engineer	Fine
15	Male	Netherlands	54	Employed full-time			No	No		
16	Male	Taiwan	26	Employed full-time			Yes		Software De	Fine
17	Male	United State	58	Independent contractor, freelancer, or s			Yes		DBA/Databa	Poorly
18	Male	Italy	58	Employed full-time			No	No		
19	Male	United Kingd	24	Employed full-time			No	No		


multipleChoiceResponses
Wrap Text
General
Merge & Center
$
%
.0
.00
.00
.0
-delimited (.csv) format. To preserve these features, save it in an E
I
J
K
L
M
witch
CurrentJobTi
TitleFit
CurrentEmpl
MLToolNext
MLMethodN



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	GenderFreef	KaggleMotiv	CurrentJobTi	MLToolNext\	MLMethodN	LanguageRe	PublicDatase	PersonalProj	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	LearningPlat	BlogsPodcas	JobS
2								Data manipulation					NA		NA		NA		NA	None	
3								I can't find time to practice consistently					NA		NA		NA		NA		
4			teacher							Meetups			NA		NA		NA		NA		
5								Connectivity/data fusion					NA		NA		NA		NA		
6													NA		NA		NA		NA		
7							kdnuggets	Prepping data					NA		NA		NA		NA		
8													NA		NA		NA		NA		
9							Stanford SNAP						NA		NA		NA		NA		
10													NA		NA		NA		NA		
11				PyTorch									NA		NA		NA		NA		
12													NA		NA		NA		NA		
13													NA		NA		NA		NA		
14													NA		NA		NA		NA		
15													NA		NA		NA		NA		
16													NA		NA		NA		NA		
17													NA		NA		NA		NA		
18													NA		NA		NA		NA		
19													NA		NA		NA		NA		
20		Curious											NA		NA		NA		NA		
21			Hydrographic Surveyor										NA		NA		NA		NA		
22													NA		NA		NA		NA		
23								Poor data quality / lack of documentation					NA		NA		NA		NA		
24													NA		NA		NA		NA		
25			mechanical engineer			don't know							NA		NA		NA		NA		
26		Promote our	Technical support engineer										NA		NA		NA		NA		
27													NA		NA		NA		NA		
28													NA		NA		NA		NA		
29								Crawling Airbnb					NA		NA		NA		NA		
30													NA		NA		NA		NA		
31								Amount and quality of data					NA		NA		NA		NA	machinelearningn	
32													NA		NA		NA		NA		
33													NA		NA		NA		NA	Analytics Vidya,Da	
34	half man - half dog							None in specific.					NA		NA		NA		NA	Machinelearningn	
35													NA		NA		NA		NA		

Some answers are not well defined (free form).

24				
25			mechanical engineer	
26		Promote our	Technical support engineer	
27				
28				
29				
30				
31				
32	Gender (free form).			
33				
34	half man - half dog			
35				

freeformResponses

Ready

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

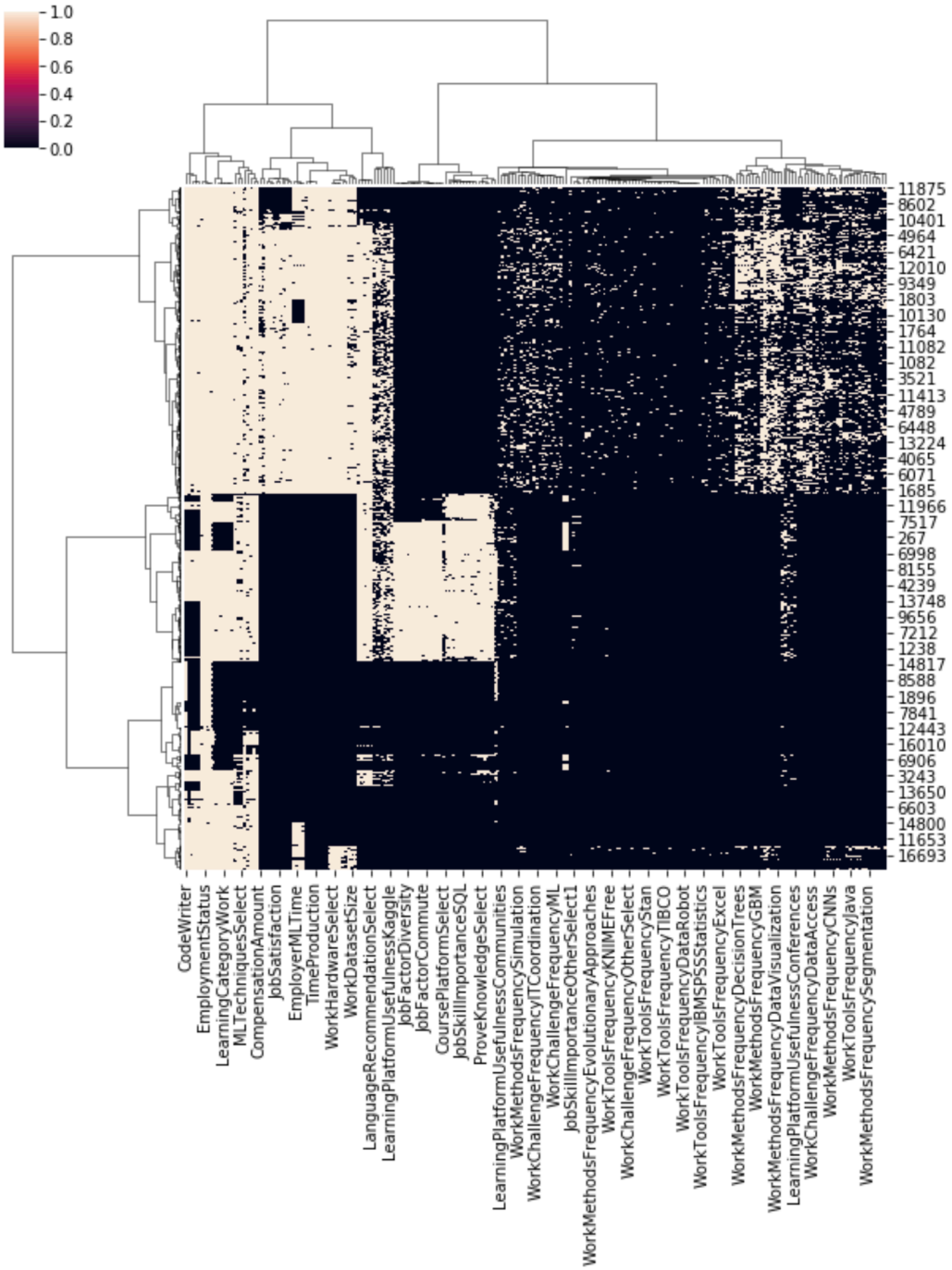
In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

In [6]:
```
sns.clustermap(                    # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

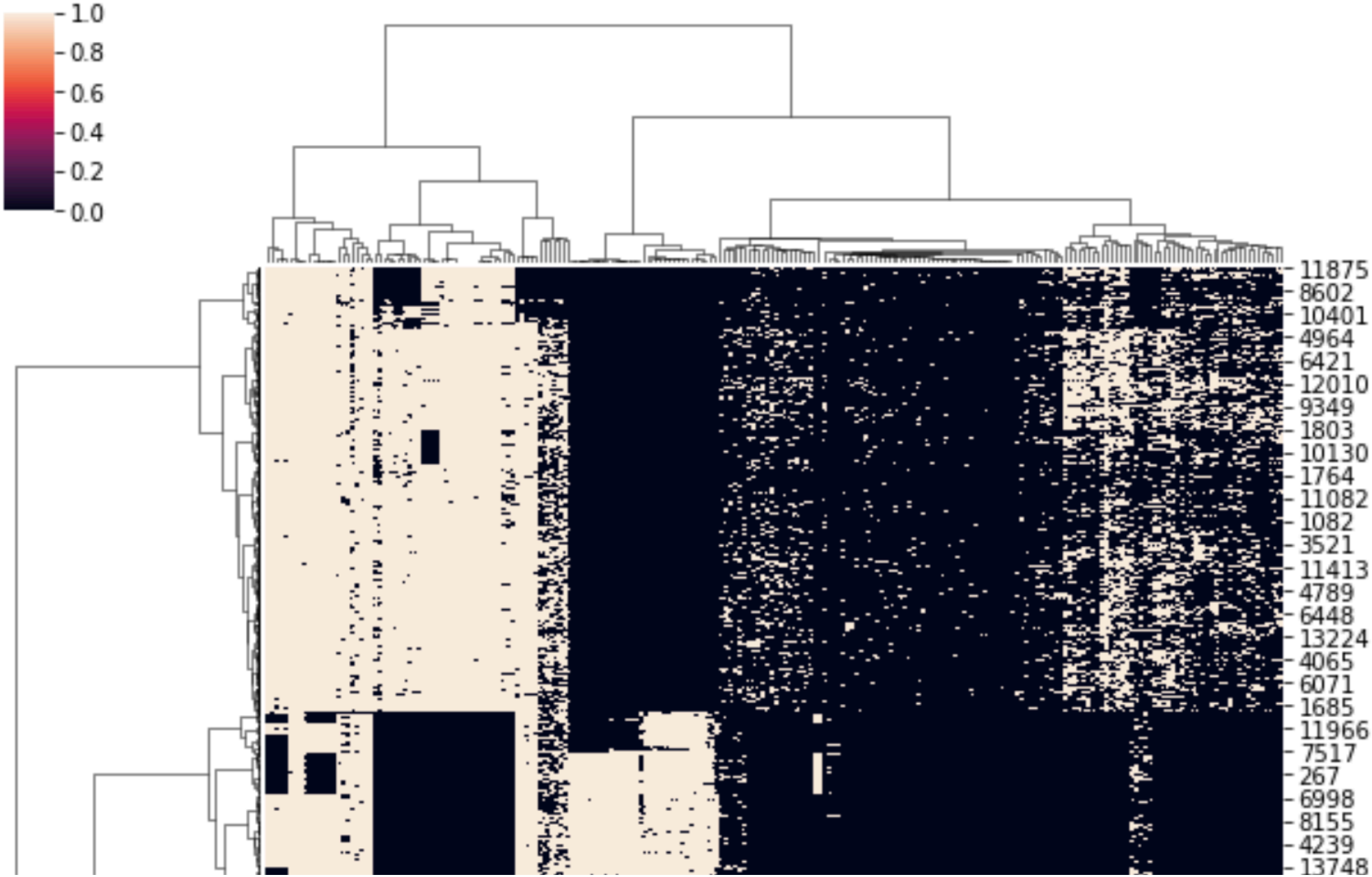
In [6]:
```
sns.clustermap(             # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

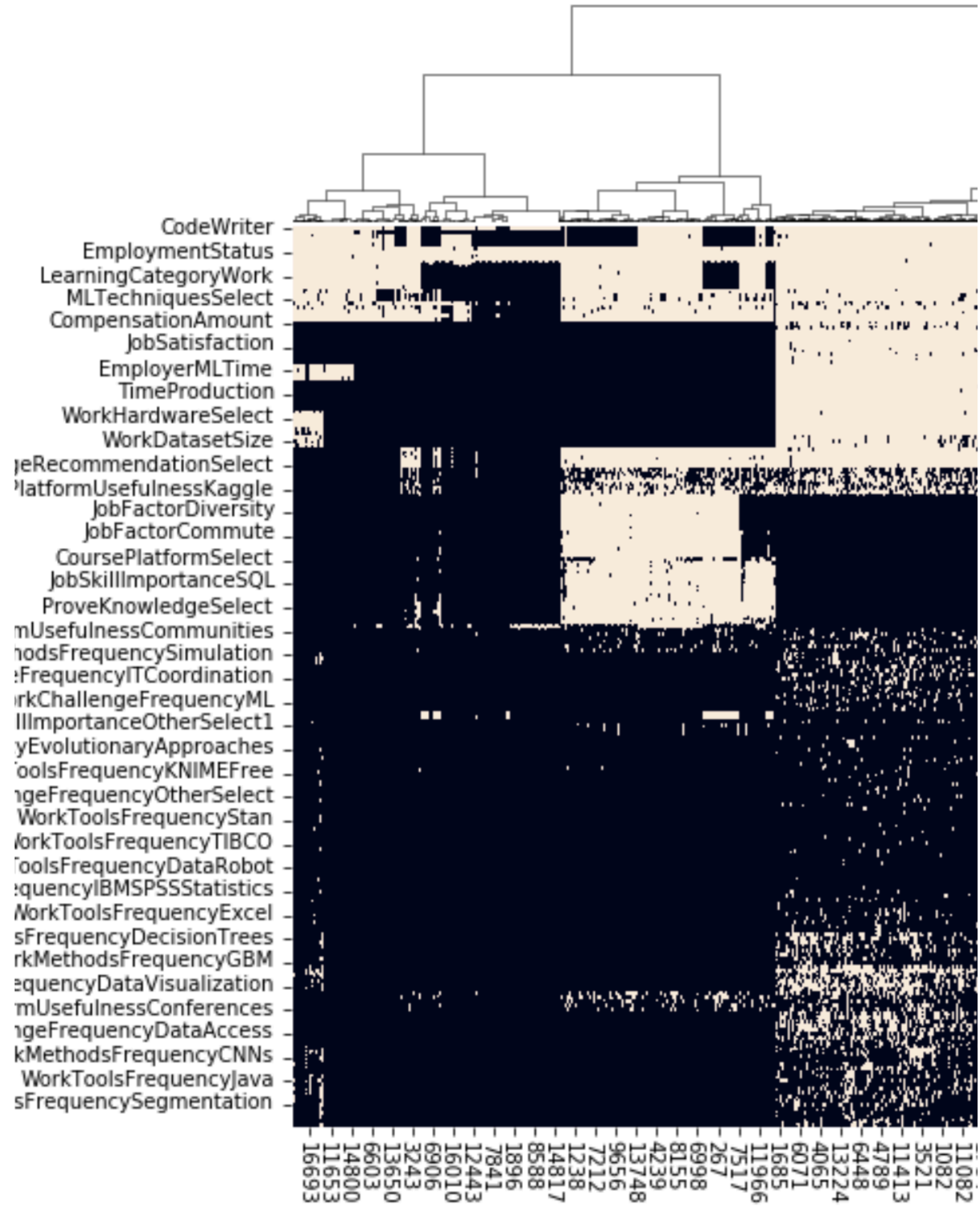
Out[6]: <seaborn.matrix.ClusterGrid at 0x10cd02a58>



```
sns.clustermap(                # visualize presence of data
    df_multiple_choice.notnull(),
    method='ward')
```

```
<seaborn.matrix.ClusterGrid at 0x10cd02a58>
```




CodeWriter
EmploymentStatus
LearningCategoryWork
MLTechniquesSelect
CompensationAmount
JobSatisfaction
EmployerMLTime
TimeProduction
WorkHardwareSelect
WorkDatasetSize
geRecommendationSelect
PlatformUsefulnessKaggle
JobFactorDiversity
JobFactorCommute
CoursePlatformSelect
JobSkillImportanceSQL
ProveKnowledgeSelect
mUsefulnessCommunities
hodsFrequencySimulation
eFrequencyITCoordination
rkChallengeFrequencyML
llImportanceOtherSelect1
yEvolutionaryApproaches
oolsFrequencyKNIMEFree
ngeFrequencyOtherSelect
WorkToolsFrequencyStan
WorkToolsFrequencyTIBCO
ToolsFrequencyDataRobot
equencyIBMSPSSStatistics
WorkToolsFrequencyExcel
sFrequencyDecisionTrees
rkMethodsFrequencyGBM
equencyDataVisualization
mUsefulnessConferences
ngeFrequencyDataAccess
kMethodsFrequencyCNNs
WorkToolsFrequencyJava
sFrequencySegmentation
16693
11653
14800
6603
13650
3243
6906
16010
12443
7841
1896
8588
14817
12238
7212
9656
13748
4239
8155
6998
267
7517
11966
1685
6071
4065
13224
6448
4789
11413
3521
1082

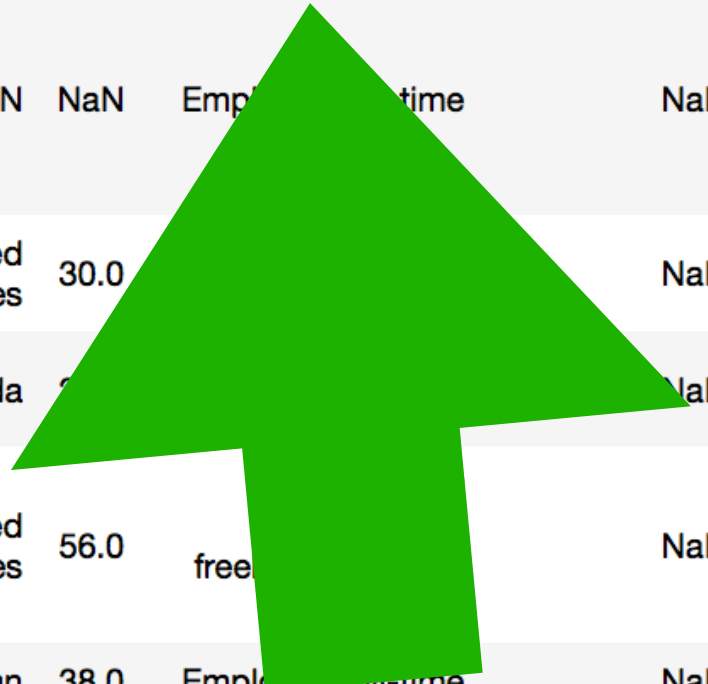
Attention. There will be a small test in three slides.

Always look at data!

```
In [7]: df_multiple_choice
```

Out[7]:

	GenderSelect	Country	Age	EmploymentStatus	StudentStatus	LearningDataScience
0	Non-binary, genderqueer, or gender non-conforming	NaN	NaN	Emp time	NaN	NaN
1	Female	United States	30.0		NaN	NaN
2	Male	Canada			NaN	NaN
3	Male	United States	56.0	free	NaN	NaN
4	Male	Taiwan	38.0	Employed full-time	NaN	NaN
						NaN
						NaN
7	Female	India	22.0	Employed full-time	NaN	NaN
8	Female	Australia	43.0	Employed full-time	NaN	NaN
9	Male	Russia	33.0	Employed full-time	NaN	NaN



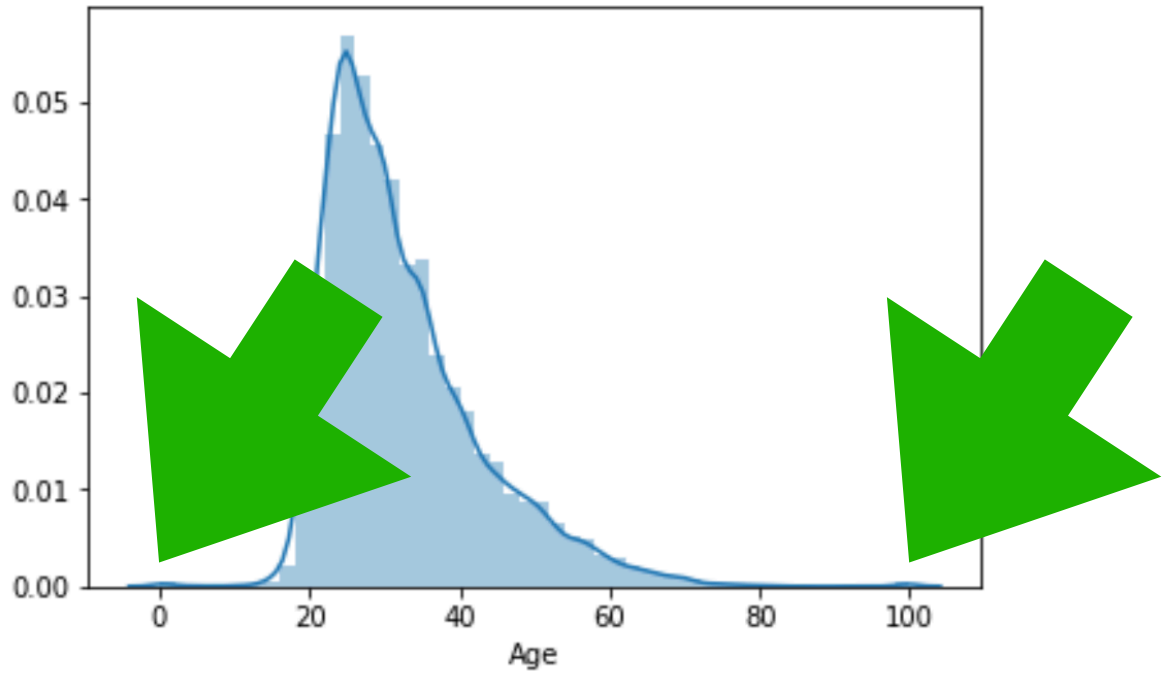
In a nicely organized data set, columns are organized in a meaningful way

In [8]:
```
sns.distplot(
    df_multiple_choice['Age'].dropna(),
)
```

Sanity-check numbers, and test for unusual values.

```
In [8]: sns.distplot(
            df_multiple_choice['Age'].dropna(),
        )
```

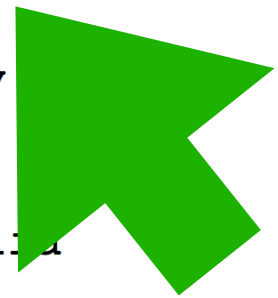
```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1a124307f0>
```



Sanity-check numbers, and test for unusual values.

```
In [11]: df_multiple_choice['Country'].value_counts()

Out[11]: United States                      4197
         India                              2704
         Other                              1023
         Russia                              578
         United Kingdom                      535
         People 's Republic of China         471
         Brazil                              465
         Germany                             460
         France                              442
         Canada                              440
         Australia                           421
         Spain                               320
```



Always get aware of spelling conventions!

Sanity-check occurrences of categorial data.

```
In [10]: df_multiple_choice['GenderSelect'].value_counts()

Out[10]: Male                                                      13610
         Female                                                     2778
         A different identity                                        159
         Non-binary, genderqueer, or gender non-conforming            74
         Name: GenderSelect, dtype: int64
```

Test: Given the observed gender frequency, why do we have to be suspicious about some of the data seen in one of the recent slides?

Sanity-check occurrences of categorial data.

Workflow

- Anticipate problems
- Get Overview
- Clean data ← Here
- Find connections
- How science differs

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

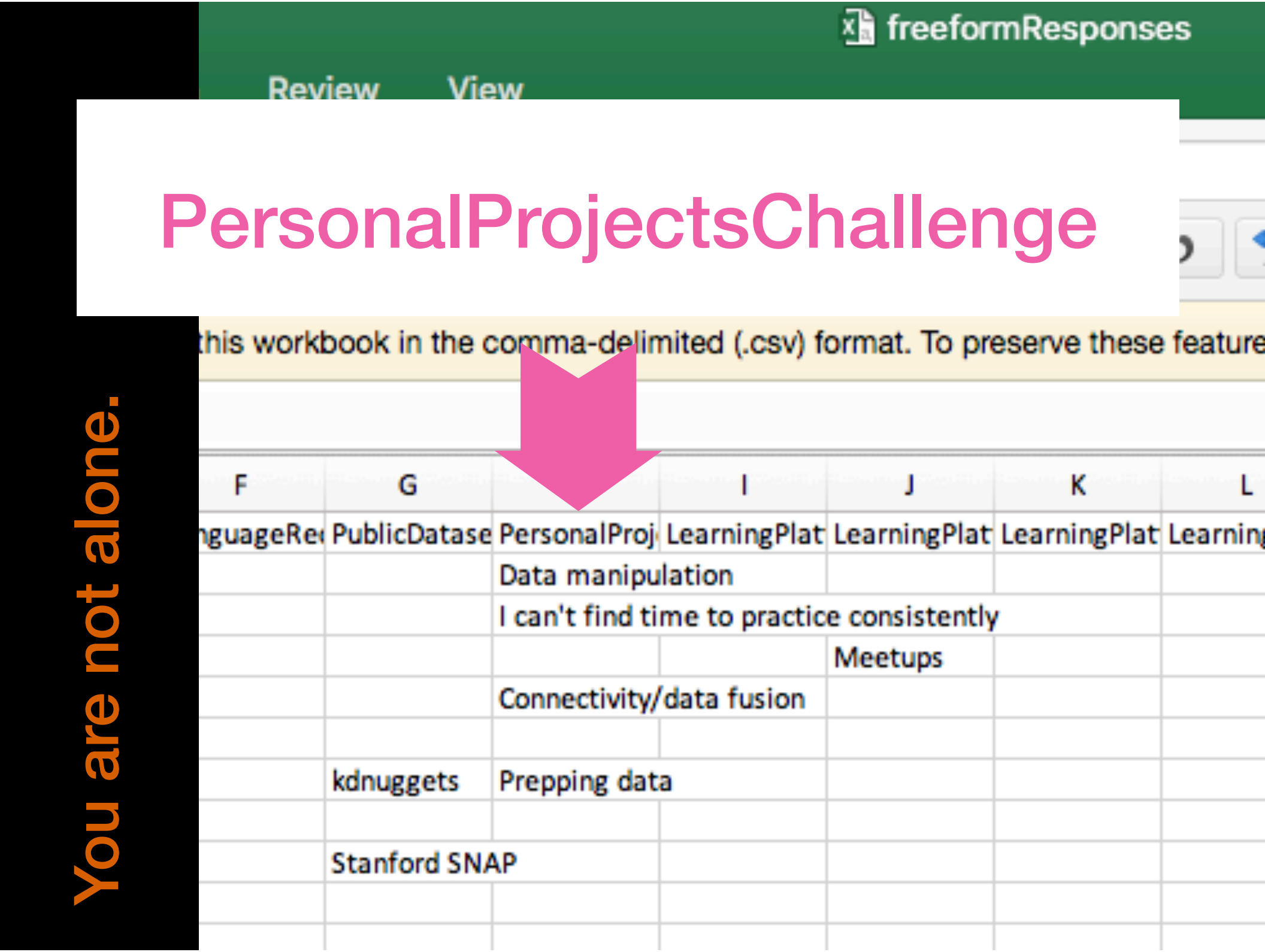
- Data
- Code
- Computation

Summary

What did we miss?

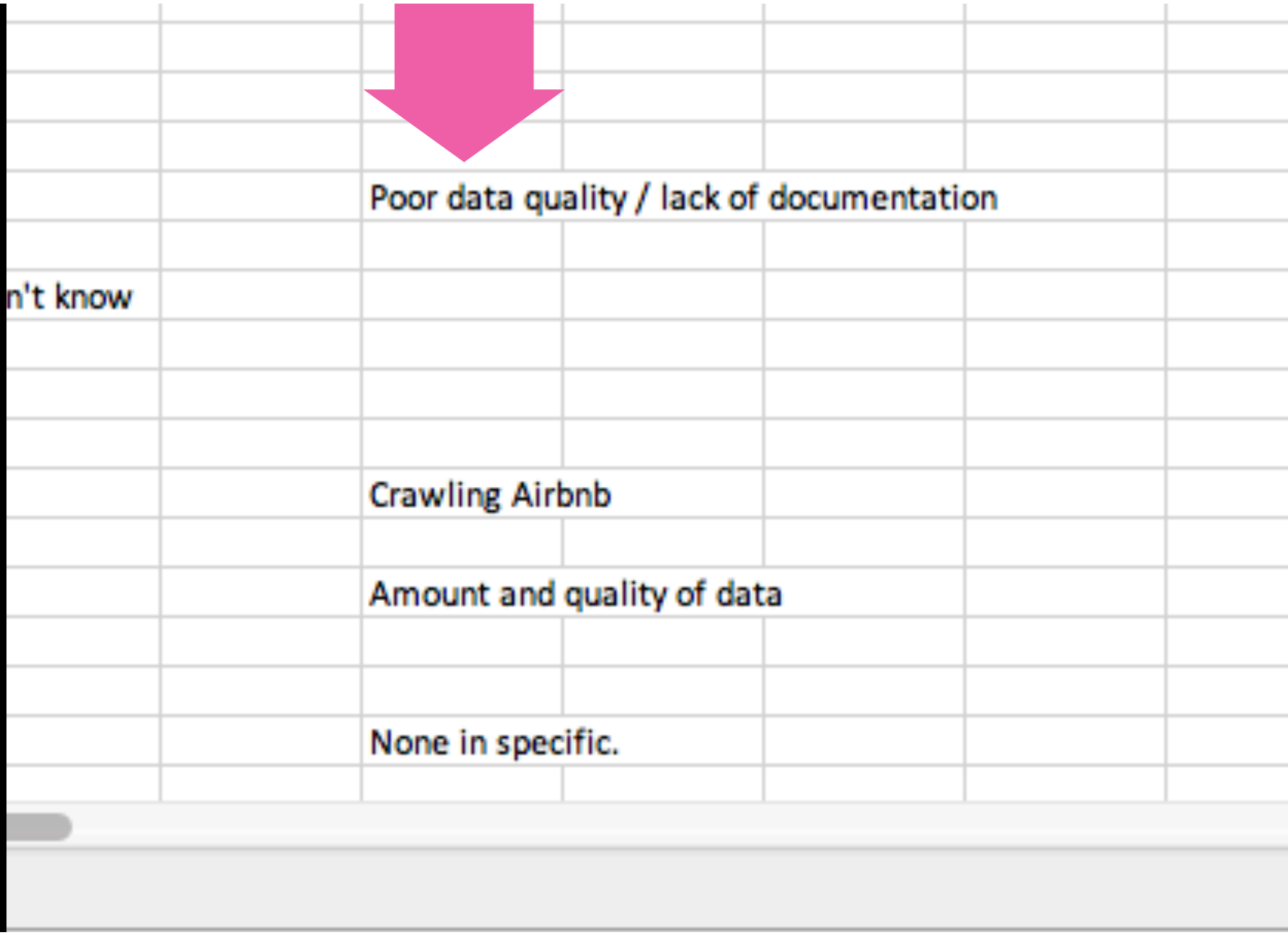
After lecture (extra): getting your environment ready

Data science is 90% *data cleaning* and 10% complaining about *data cleaning*.


freeformResponses
Review
View
PersonalProjectsChallenge
this workbook in the comma-delimited (.csv) format. To preserve these feature
F
G
I
J
K
L
nguageRe
PublicDatase
PersonalProj
LearningPlat
LearningPlat
LearningPlat
Learning
Data manipulation
I can't find time to practice consistently
Meetups
Connectivity/data fusion
kdnuggets
Prepping data
Stanford SNAP
You are not alone.

PersonalProjectsChallenge

You are not alone.



Main tactics for data cleaning.

- Regular expressions are your best friend to only clean what you want to clean.
- Use errors (codified sanity checks) to your advantage.
- Fuzzy matching and entity matching help you to combine related data sources, which do not have a shared identifier.

“Let us explain the salary of data scientists!”

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

In [14]:
```
# For the sake of this example, let us only consider
# records, where CompensationAmount is defined
# (we still keep records, if something else, e.g.
# StudentStatus, whould be not defined)
df_multiple_choice = df_multiple_choice.dropna(
    subset=['CompensationAmount']
)
```

In [13]:
```
# For the sake of this example, let us only consider
# salaries that have been converted to USD
f = df_multiple_choice['CompensationCurrency'] == 'USD'
df_multiple_choice = df_multiple_choice[f]
```

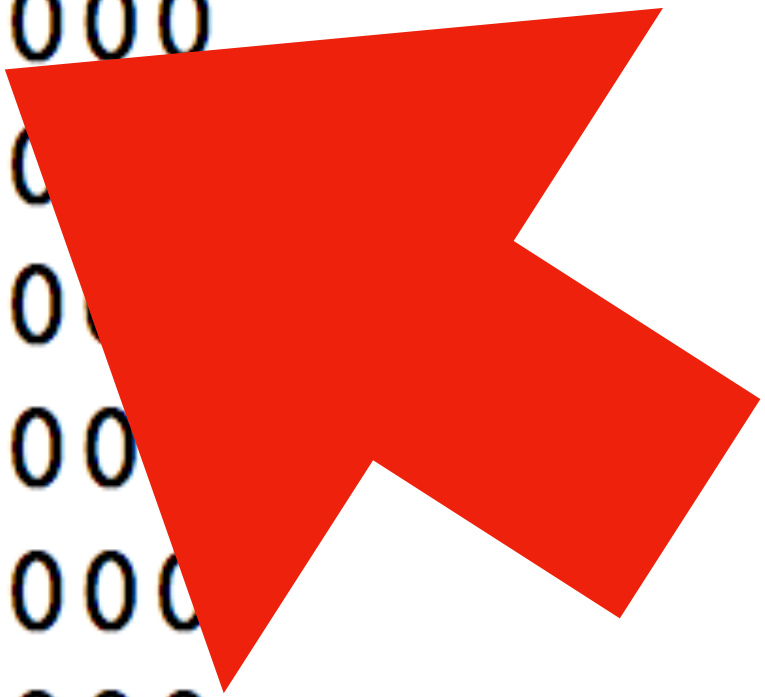
In [14]:
```
# For the sake of this example, let us only consider
# records, where CompensationAmount is defined
# (we still keep records, if something else, e.g.
# StudentStatus, whould be not defined)
df_multiple_choice = df_multiple_choice.dropna(
    subset=['CompensationAmount']
)
```

In [15]:
```
# Manually inspect the data in CompensationAmount
df_multiple_choice['CompensationAmount']
```

Out[15]:
```
3         250,000
21          20000
22         100000
34         133000
37          80000
61          15000
75         215000
86          83500
```

```
# Manually inspect the data
df_multiple_choice['Compensa
```

```
3          250,000
21             200
22            1000
34           13300
37           80000
61           15000
75          215000
86           83500
```



In [16]:
```
# Let us convert these values to a number
df_multiple_choice['CompensationAmount'] = df_multiple_choice[
    'CompensationAmount'].astype(float)
```

Error

```
# Let us convert these values to a number
df_multiple_choice['CompensationAmount'] = df_multiple_choice[
    'CompensationAmount'].astype(float)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-16-f163d2412d49> in <module>()
      1 df_multiple_choice['CompensationAmount'] = df_multiple_choice[
----> 2     'CompensationAmount'].astype(float)

~/anaconda3/lib/python3.6/site-packages/pandas/util/_decorators.py in wrapper(*args, **kwargs)
    176                 else:
    177                     kwargs[new_arg_name] = new_arg_value
--> 178             return func(*args, **kwargs)
    179         return wrapper
    180     return _deprecate_kwarg

~/anaconda3/lib/python3.6/site-packages/pandas/core/generic.py in astype(self, dtype, copy, errors, **kwargs)
   4995             # else, only a single dtype is given
   4996             new_data = self._data.astype(dtype=dtype, copy=copy, errors=errors,
-> 4997                                          **kwargs)
   4998             return self._constructor(new_data).__finalize__(self)
   4999

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, **kwargs)
   3712
   3713     def astype(self, dtype, **kwargs):
-> 3714         return self.apply('astype', dtype=dtype, **kwargs)
   3715
   3716     def convert(self, **kwargs):

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in apply(self, f, axes, filter, do_integrity_check,
consolidate, **kwargs)
   3579
   3580             kwargs['mgr'] = self
-> 3581             applied = getattr(b, f)(**kwargs)
   3582             result_blocks = _extend_blocks(applied, result_blocks)
   3583

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, copy, errors, values, **kwar
gs)
    573     def astype(self, dtype, copy=False, errors='raise', values=None, **kwargs):
    574         return self._astype(dtype, copy=copy, errors=errors, values=values,
--> 575                             **kwargs)
    576
    577     def _astype(self, dtype, copy=False, errors='raise', values=None,

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in _astype(self, dtype, copy, errors, values, klass
, mgr, **kwargs)
    662
    663                 # _astype_nansafe works fine with 1-d only
--> 664                 values = astype_nansafe(values.ravel(), dtype, copy=True)
    665                 values = values.reshape(self.shape)
    666

~/anaconda3/lib/python3.6/site-packages/pandas/core/dtypes/cast.py in astype_nansafe(arr, dtype, copy)
    728
    729     if copy:
--> 730         return arr.astype(dtype, copy=True)
    731     return arr.view(dtype)
    732

ValueError: could not convert string to float: '85,000'
```

Could not convert string to float: ’85,000’

Only fix what you want to fix. (85,000 and similar)

```
ValueError: could not convert string to float: '85,000'
```

```
In [17]: f = df_multiple_choice['CompensationAmount'].str.contains(
             '[0-9]*,[0-9]{3}$')   # create a highly specific regular expression
```

```
In [18]: df_multiple_choice.loc[f, 'CompensationAmount'] = df_multiple_choice.loc[
             f, 'CompensationAmount'].str.replace(',', '')
```

Can someone translate this?

[0-9]*,[0-9]{3}$

Trick question: Could this be shorter to solve our problem?

In [19]:
```
df_multiple_choice['CompensationAmount'] = df_multiple_choice[
    'CompensationAmount'].astype(float)
```

Cleaning data is an interactive process.

```
In [19]: df_multiple_choice['CompensationAmount'] = df_multiple_choice[
             'CompensationAmount'].astype(float)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-19-f163d2412d49> in <module>()
      1 df_multiple_choice['CompensationAmount'] = df_multiple_choice[
----> 2     'CompensationAmount'].astype(float)

~/anaconda3/lib/python3.6/site-packages/pandas/util/_decorators.py in wrapper(*args, **kwargs)
    176                 else:
    177                     kwargs[new_arg_name] = new_arg_value
--> 178             return func(*args, **kwargs)
    179         return wrapper
    180     return _deprecate_kwarg

~/anaconda3/lib/python3.6/site-packages/pandas/core/generic.py in astype(self, dtype, copy, errors, **kwargs)
   4995             # else, only a single dtype is given
   4996             new_data = self._data.astype(dtype=dtype, copy=copy, errors=errors,
-> 4997                                          **kwargs)
   4998             return self._constructor(new_data).__finalize__(self)
   4999

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, **kwargs)
   3712
   3713     def astype(self, dtype, **kwargs):
-> 3714         return self.apply('astype', dtype=dtype, **kwargs)
   3715
   3716     def convert(self, **kwargs):

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in apply(self, f, axes, filter, do_integrity_check,
consolidate, **kwargs)
   3579
   3580             kwargs['mgr'] = self
-> 3581             applied = getattr(b, f)(**kwargs)
   3582             result_blocks = _extend_blocks(applied, result_blocks)
   3583

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in astype(self, dtype, copy, errors, values, **kwar
gs)
    573     def astype(self, dtype, copy=False, errors='raise', values=None, **kwargs):
    574         return self._astype(dtype, copy=copy, errors=errors, values=values,
--> 575                             **kwargs)
    576
    577     def _astype(self, dtype, copy=False, errors='raise', values=None,

~/anaconda3/lib/python3.6/site-packages/pandas/core/internals.py in _astype(self, dtype, copy, errors, values, klass
, mgr, **kwargs)
    662
    663                 # _astype_nansafe works fine with 1-d only
--> 664                 values = astype_nansafe(values.ravel(), dtype, copy=True)
    665                 values = values.reshape(self.shape)
    666

~/anaconda3/lib/python3.6/site-packages/pandas/core/dtypes/cast.py in astype_nansafe(arr, dtype, copy)
    728
    729     if copy:
--> 730         return arr.astype(dtype, copy=True)
    731     return arr.view(dtype)
    732

ValueError: could not convert string to float: '-'
```

Main tactics for data cleaning.

- Regular expressions are your best friend to only clean what you want to clean.
- Use errors (codified sanity checks) to your advantage.
- Fuzzy matching and entity matching help you to combine related data sources, which do not have a shared identifier.

e.g: match Thomas Stoeger to Thomas Stoger

Missing data can be tricky.

Advance Access publication August 22, 2016

Political Analysis (2016) 24:414–433
doi:10.1093/pan/mpw020

How Multiple Imputation Makes a Difference

Ranjit Lall

Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: ranjitlall@fas.harvard.edu (corresponding author)

Edited by R. Michael Alvarez

Political scientists increasingly recognize that multiple imputation represents a superior strategy for analyzing missing data to the widely used method of listwise deletion. However, there has been little systematic investigation of *how* multiple imputation affects existing empirical knowledge in the discipline. This article presents the first large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science. The examination focuses on research in the major subfield of comparative and international political economy (CIPE) as an illustrative example. Specifically, I use multiple imputation to reanalyze the results of almost every quantitative CIPE study published during a recent five-year period in *International Organization* and *World Politics*, two of the leading subfield journals in CIPE. The outcome is striking: in almost half of the studies, key results "disappear" (by conventional statistical standards) when reanalyzed.

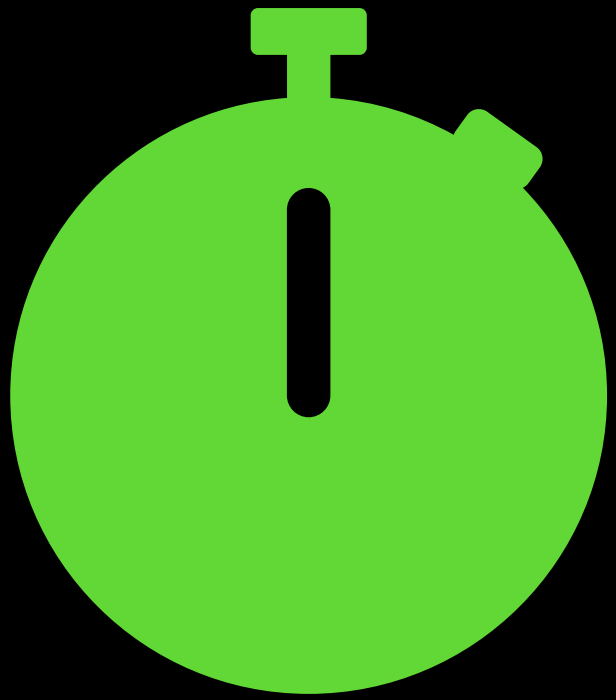
A rough guide to missing data:

Random

- Discard records
- Impute based on similar records

Non-random

- Use as information
- Consider as "below measurement limit"
- Advanced (often domain-specific) imputation



Three minute exercise.

Talk to your neighbor.

Thinking of your worst encounter with data. What would have been, or was, a good strategy to clean the data or avoid unwanted surprises.

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Here is the fun

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Summary

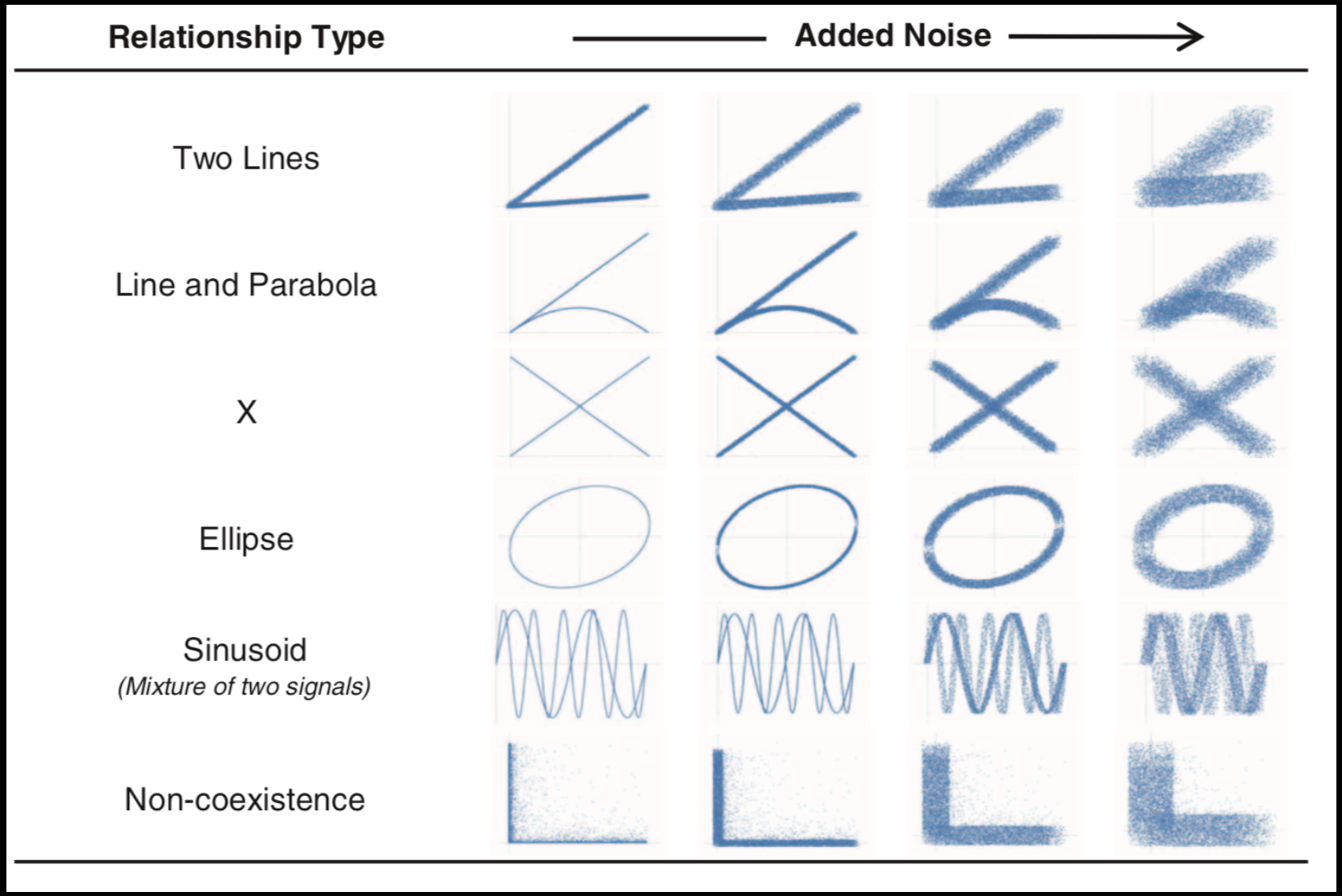
What did we miss?

After lecture (extra): getting your environment ready

Main tactics for finding connections.

- Correlations
- Exploratory visualization
- Machine learning

Distinct types of correlations can embody distinct assumptions on the data.



Reshef et al. 2011

Distinct types of correlations can embody distinct assumptions on the data.

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

Reshef et al. 2011

Exploratory visualization can give an idea of magnitudes.

In [108]:
```
f = (
    df_multiple_choice['GenderSelect'].isin(['Male', 'Female'])) & (
    df_multiple_choice['Country'].isin(['United States'])) & (
    df_multiple_choice['EmploymentStatus'].isin(['Employed full-time']) & (
    df_multiple_choice['Age'].isin(range(20, 35))) & (
    df_multiple_choice['CurrentJobTitleSelect'].isin(['Data Scientist']))
)
```

Exploratory visualization can give an idea of magnitudes.

In [108]:
```
f = (
    df_multiple_choice['GenderSelect'].isin(['Male', 'Female'])) & (
    df_multiple_choice['Country'].isin(['United States'])) & (
    df_multiple_choice['EmploymentStatus'].isin(['Employed full-time']) & (
    df_multiple_choice['Age'].isin(range(20, 35))) & (
    df_multiple_choice['CurrentJobTitleSelect'].isin(['Data Scientist']))
)
```

In [109]:
```
import numpy as np
```

In [110]:
```
sns.pointplot(
    x='Age',
    y='CompensationAmount',
    data=df_multiple_choice[f],
    estimator=np.median,
    hue='GenderSelect',
    dodge=True
)
```

Exploratory visualization can give an idea of magnitudes.

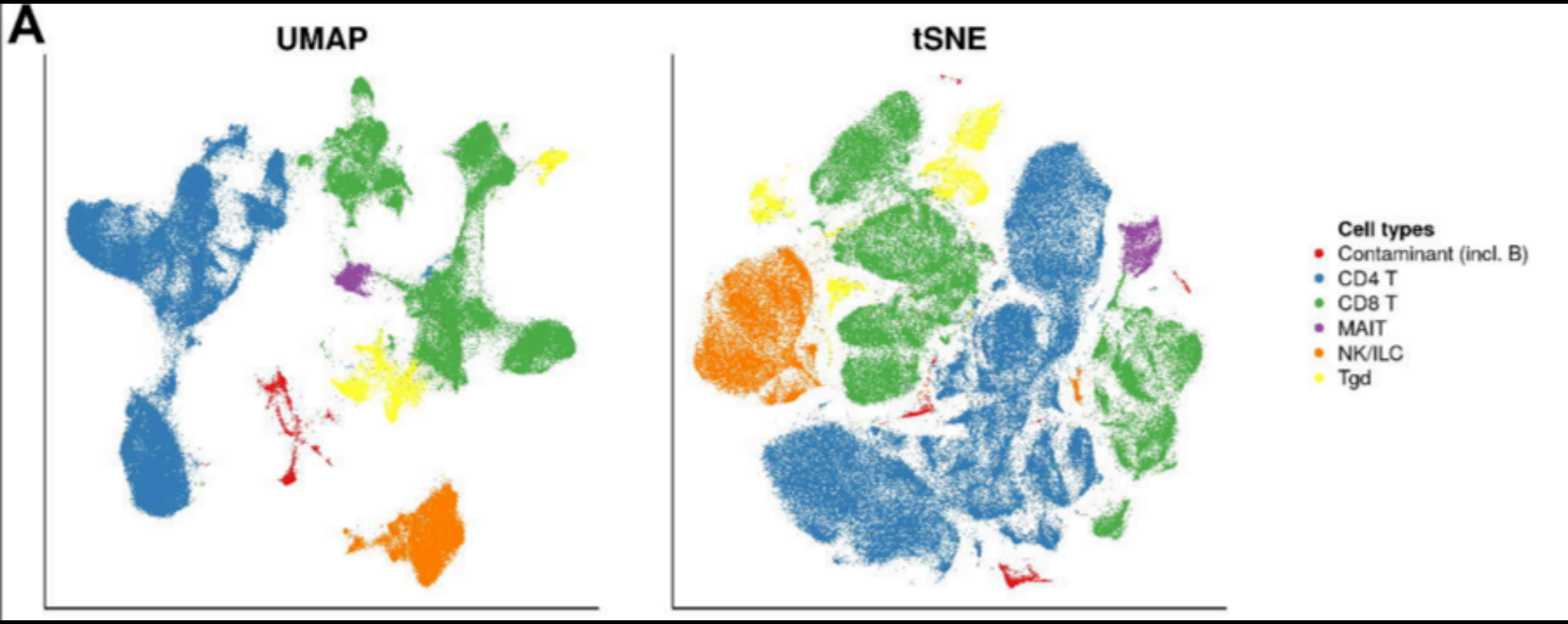
In [108]:
```
f = (
    df_multiple_choice['GenderSelect'].isin(['Male', 'Female'])) & (
    df_multiple_choice['Country'].isin(['United States'])) & (
    df_multiple_choice['EmploymentStatus'].isin(['Employed full-time']) & (
    df_multiple_choice['Age'].isin(range(20, 35))) & (
    df_multiple_choice['CurrentJobTitleSelect'].isin(['Data Scientist']))
)
```

In [109]:
```
import numpy as np
```

In [110]:
```
sns.pointplot(
    x='Age',
    y='CompensationAmount',
    data=df_multiple_choice[f],
    estimator=np.median,
    hue='GenderSelect',
    dodge=True
)
```

Median salary of data scientists at university ages.

Extra tactic: visualize upon dimensionality reduction.



Dimensionality reduction
Fast

Only visualization
Slow

Becht et al., bioRxiv 2018

Main tactics for finding connections.

- Correlations
- Exploratory visualization
- **Machine learning**

```
In [113]: df_multiple_choice['MLMethodNextYearSelect'].value_counts()

Out[113]: Deep learning                                          535
          Neural Nets                                            207
          Bayesian Methods                                       101
          Time Series Analysis                                    90
          Genetic & Evolutionary Algorithms                       66
          Anomaly Detection                                       61
          Text Mining                                             52
          Social Network Analysis                                 48
          Other                                                   46
          Ensemble Methods (e.g. boosting, bagging)               42
          I don't plan on learning a new ML/DS method             33
          Cluster Analysis                                        31
          Monte Carlo Methods                                     28
          Support Vector Machines (SVM)                           26
          Proprietary Algorithms                                  21
          Random Forests                                          18
          Survival Analysis                                       18
          Regression                                              15
          Rule Induction                                           8
          Link Analysis                                            8
          Decision Trees                                           8
          Uplift Modeling                                          6
          MARS                                                     5
          Factor Analysis                                          5
          Association Rules                                        3
          Name: MLMethodNextYearSelect, dtype: int64
```

Reminder.

Any question?

How to learn about relations through predictions?

1) **Through contribution to model. E.g.: linear regression, or random forest, or neural nets.**

How to learn about relations through predictions?

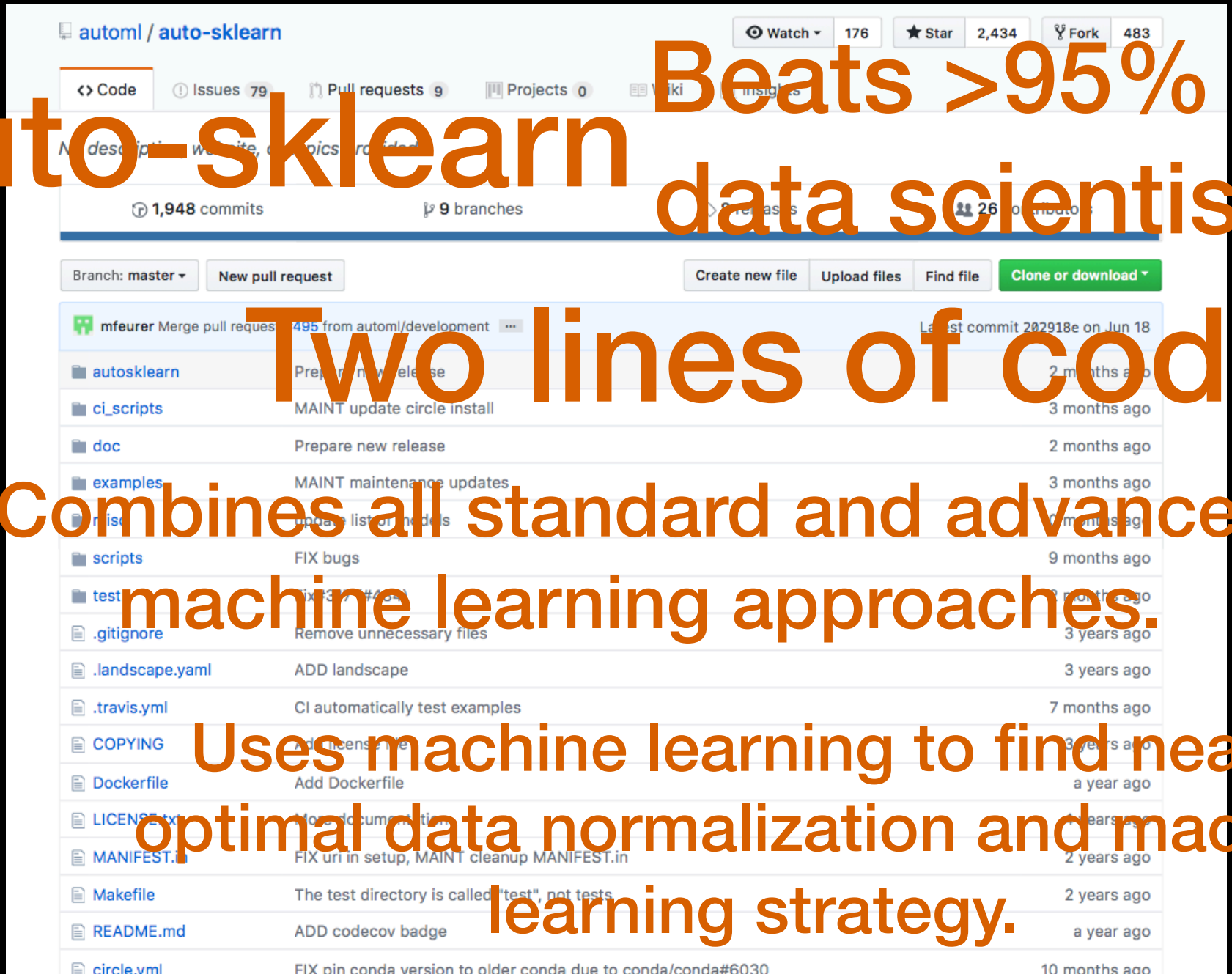
1) **Through contribution to model.** E.g.: linear regression, or random forest, or neural nets.

2) **Through ability to predict.**

How to find best modeling strategy?

Generalist **Domain-specific**

Finding strategies through generalist approaches can be automated.



auto-sklearn

Beats >95% of data scientists

Two lines of code.

Combines all standard and advanced machine learning approaches.

Uses machine learning to find nearly optimal data normalization and machine learning strategy.

Some science-specific approaches have very interesting properties.

TIME SERIES ANALYSIS

Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality

Turn high dimensionality from a curse to a blessing.

Hao Ye and George Sugihara*

In ecological analysis, complexity has been regarded as an obstacle to overcome. Here we present a straightforward approach for addressing complexity in dynamic interconnected systems. We show that complexity, in the form of multiple interacting components, can actually be an asset for studying natural systems from temporal data. The central idea is that multidimensional time series enable system dynamics to be reconstructed from multiple viewpoints, and these viewpoints can be combined into a single model. We show how our approach, multiview embedding (MVE), can improve forecasts for simulated ecosystems and a mesocosm experiment. By leveraging complexity, MVE is particularly effective for overcoming the limitations of short and noisy time series and should be highly relevant for many areas of science.

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Summary

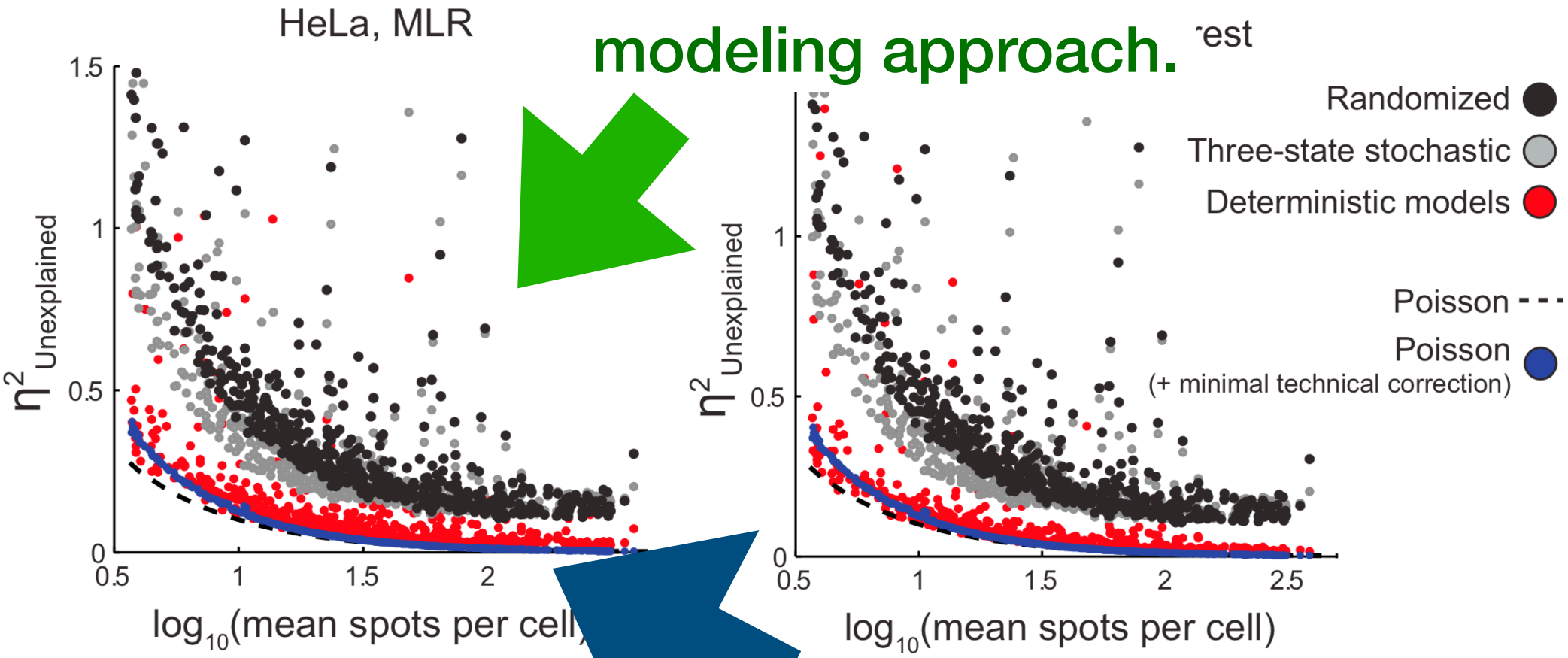
What did we miss?

After lecture (extra): getting your environment ready

Pass reviewers.

We wanted to say that the red dots are as low as possible (little remained unexplained by our models).

Choose the "worse" modeling approach.

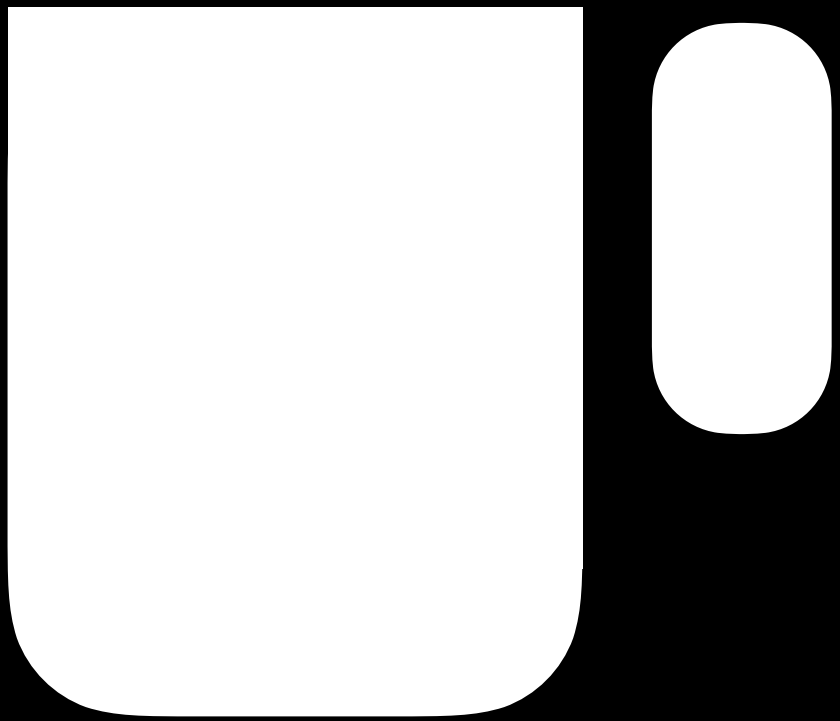


If some data isn't explained yet, test your performance against a problem-specific theoretical model.

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

What did we miss?



Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Language

- Images: Matlab
- Otherwise: Python
- Maybe: R, Excel, Tableau

Computer

- RAM bottlenecks will have strongest impact on your waiting time (if you have more RAM than needed, you can write inefficient, but possibly more readable, code)
- Consider a fast single core performance as it will mean less waiting time than multiple cores, if your code isn't optimized to multiple cores (and no workstation can compete with a cluster).
- If possible: get a desktop and a small laptop.

Space

- Avoid isolated rooms.

Constant Learning

- Books, scientific literature; e.g.: note that NU has license for huge ebook collection on data science through Safari Online
- Meetups

Please object. As some of this is scientific field specific.

Some shameless advertising.

data-science-nights.org

DATA SCIENCE NIGHTS
@NORTHWESTERN

HOME ABOUT GETTING STARTED WITH DATA SCIENCE DATA SCIENCE AROUND CAMPUS ARCHIVES

DATA SCIENCE NIGHTS

MONTHLY HACK NIGHTS ON POPULAR DATA SCIENCE TOPICS, ORGANIZED BY FELLOWS AND SCHOLARS FROM THE NORTHWESTERN DATA SCIENCE INITIATIVE. EACH NIGHT WILL FEATURE REFRESHMENTS, A **TALK ON DATA SCIENCE TECHNIQUES OR APPLICATIONS**, AND A **HACKING NIGHT** WITH DATA SCIENCE PROJECTS OR LEARNING GROUPS OF YOUR CHOICE.

ASPIRING, BEGINNING, AND ADVANCED DATA SCIENTISTS ARE WELCOME!

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Here

Summary

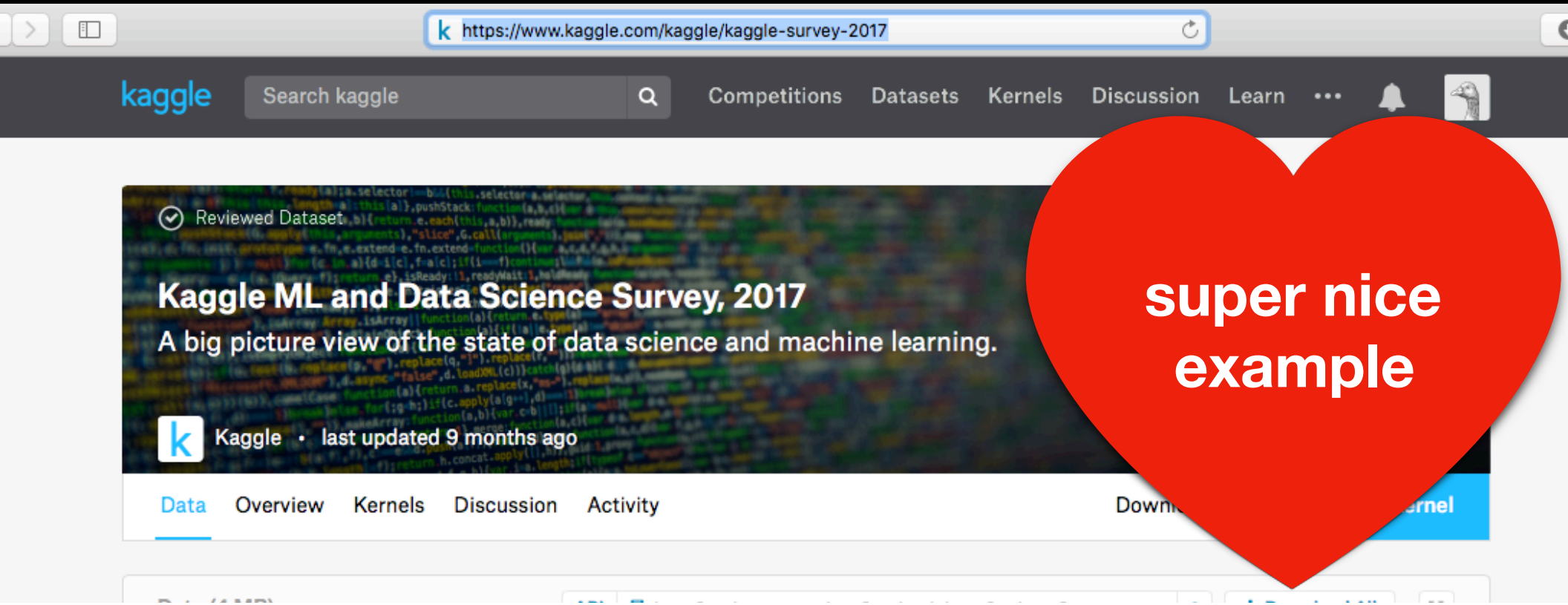
What did we miss?

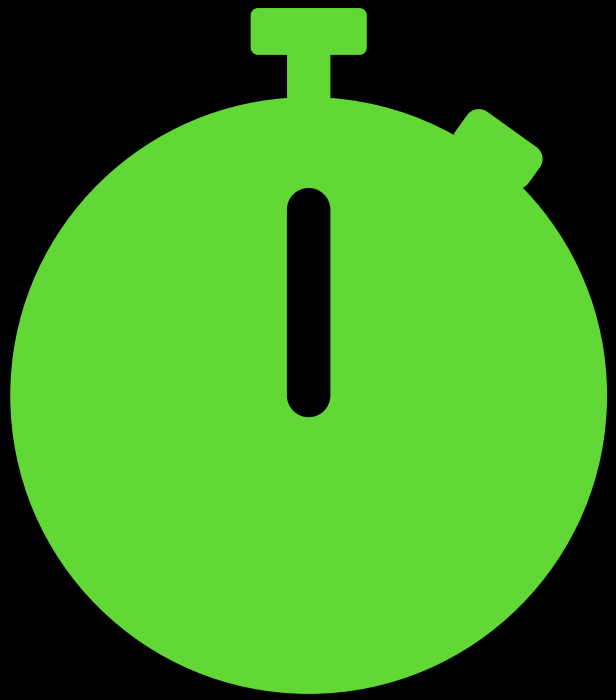
After lecture (extra): getting your environment ready

Introducing

Kaggle ML and Data Science Survey, 2017

An example data set, which we will be using





Three minute exercise.

Find a neighbor, or neighbors, whom you do not know.

Then recall tactics applied by Kaggle to create an organized data set.

Logical layout.

Portability and access.

Know your options.

Portability and access.

Do I want to work without internet access?

With whom do I collaborate?

Is data structured?

Portability and access.

Central location

- Database
- Folder with datasets
- Domain-specific online host (e.g.: NIH's Synapse)

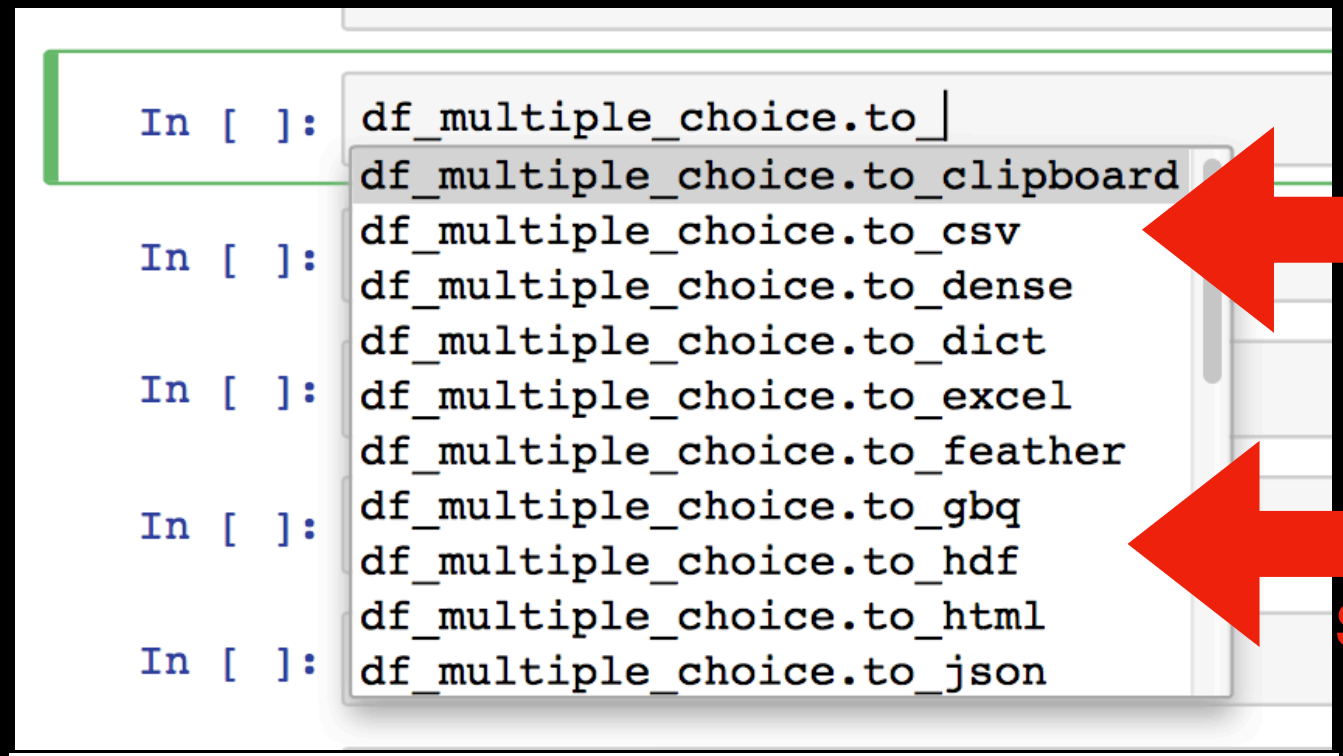
Local copies

- Have one single reference
- e.g.: have reference synced over Dropbox across Workstations, and unidirectionally copy to NU's cluster (but never copy data back from cluster)

Logical layout.

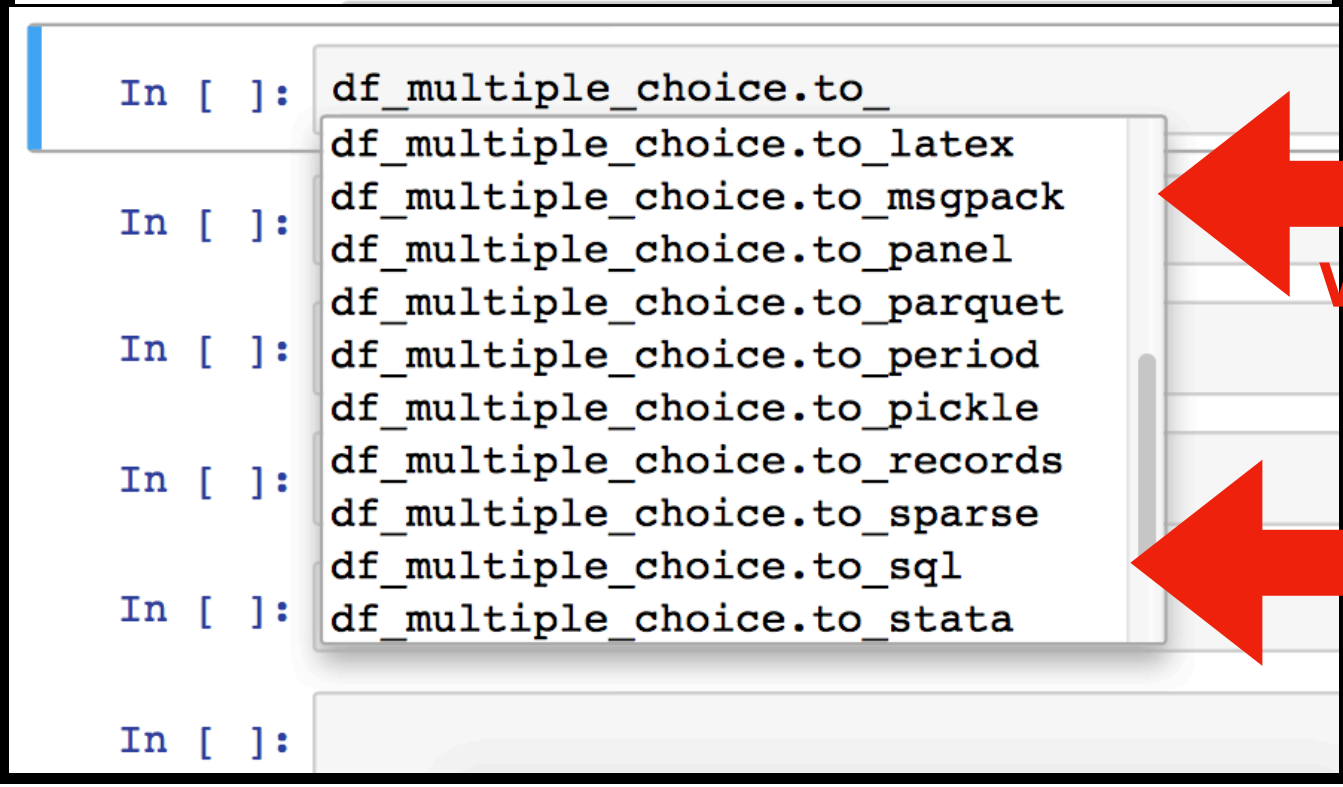
Portability and access.

Know your options.



CSV
Spreadsheets

HDF5
Selectively load Parts of data (avoid breaking cluster)



MSGPACK
Very fast reading and writing of entire tables

SQL
Elegant query language

In [172]:
```
example_data = df_multiple_choice[['Country', 'Age', 'GenderSelect']].dropna()
```

In [173]:
```
example_data.head()
```

Out[173]:

	Country	Age	GenderSelect
1	United States	30.0	Female
2	Canada	28.0	Male
3	United States	56.0	Male
4	Taiwan	38.0	Male
5	Brazil	46.0	Male

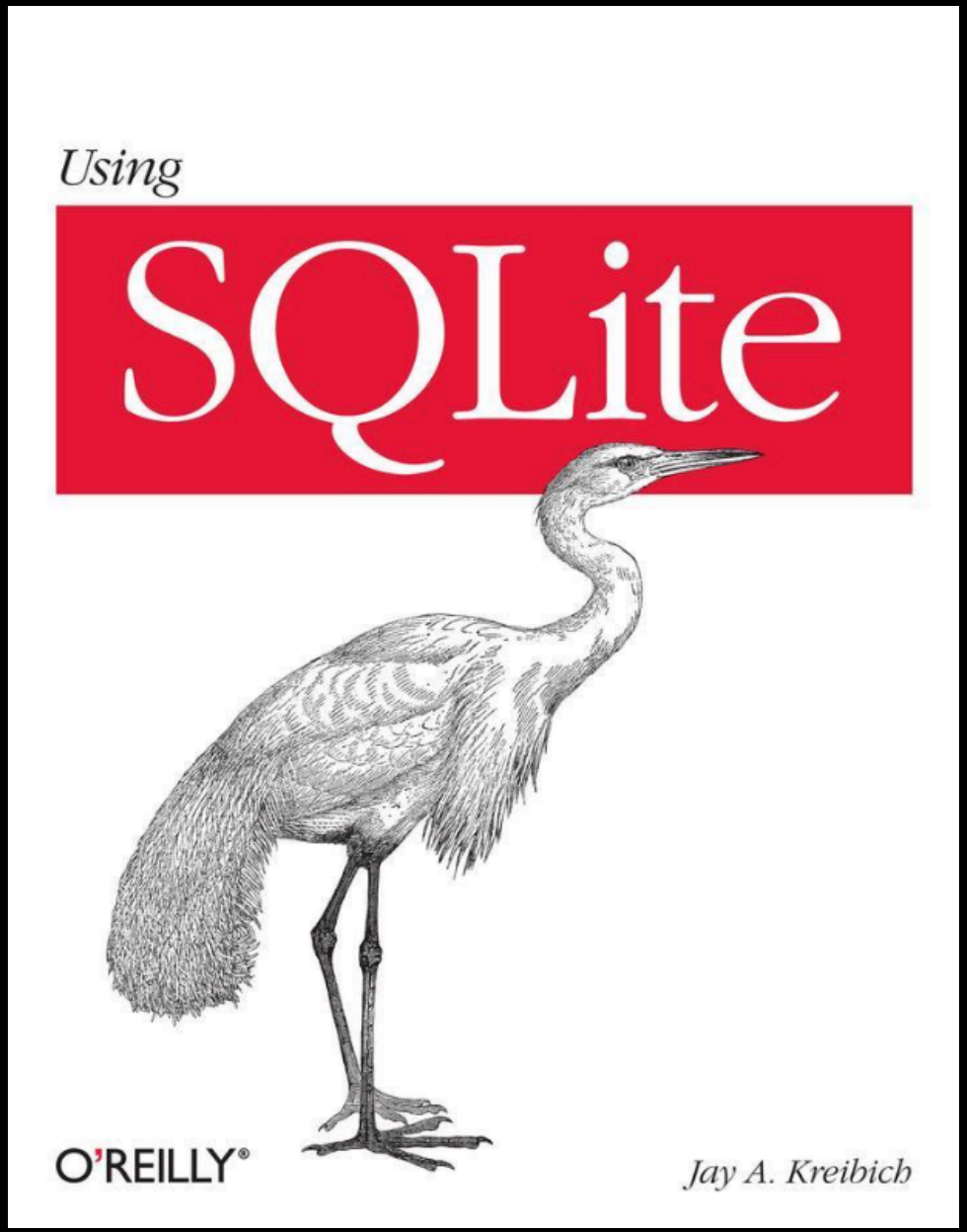
HDF5

In [174]:
```
example_data.to_hdf(
    './example.h5',
    format='table',
    key='searchable_data',
    mode='w',
    data_columns=True,
    )
```

In [175]:
```
pd.read_hdf('./example.h5', 'searchable_data', where=['Age=20'])
```

Out[175]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female
125	Turkey	20.0	Male
144	United States	20.0	Male
148	India	20.0	Male



Remember

You can have a server-less database to store and organize data.

NU provides tons of free ebooks.

Some NU specific considerations:

Data transfer to Quest: Globus (also has command line option)

Box: has some file restrictions and sometimes doesn’t copy all files. Doesn’t work on all (old) Linux, Mac, Windows-> recommendation: Instead pay for Dropbox Plus with automated backup (~150 USD per year)

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation

Here

Summary

What did we miss?

After lecture (extra): getting your environment ready


1) Future
2) But, spending time
To develop something
unused, is a waste.
3) Plan for data storage to change.
What data organization problem
did we not discuss yet?

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code ← Here
- Computation

Summary

What did we miss?

After lecture (extra): getting your environment ready

Plan for changing data.

Very brief coding practices.

Plan for changing data.

Very brief coding practices.

Functions:

- organize code
- Reusability

Adaptor:

- Can change data resource without worrying in notebook

Direct loading is perfectly fine for new datasets.

In [1]:
```
%matplotlib inline
```

In [2]:
```
cd '/Users/tstoeger/Dropbox/Work/kaggle_survey'
```

```
/Users/tstoeger/Dropbox/Work/kaggle_survey
```

In [3]:
```
import pandas as pd
import seaborn as sns
```

In [4]:
```
df_multiple_choice = pd.read_csv(
    './multipleChoiceResponses.csv',
    encoding='latin-1', # files have some special characters
    low_memory=False)   # some columns have mixed type
```

In [5]:
```
df_multiple_choice.shape   # get idea of number of datasets
```

Out[5]:
```
(16716, 228)
```

… but move your loading into adaptor functions in the long run.

In [190]:
```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_csv(
        './multipleChoiceResponses.csv',
        encoding='latin-1', # files have some special characters
        low_memory=False)   # some columns have mixed type
    df = df=df[['Country', 'Age', 'GenderSelect']].dropna()
    df = df[df['Age']==age_of_interest]

    return df
```

Define function
In human digestible
Manner

… but move your loading into adaptor functions in the long run.

In [190]:
```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_csv(
        './multipleChoiceResponses.csv',
        encoding='latin-1', # files have some special characters
        low_memory=False)   # some columns have mixed type
    df = df=df[['Country', 'Age', 'GenderSelect']].dropna()
    df = df[df['Age']==age_of_interest]

    return df
```

In [191]:
```
d = load_country_gender_at_given_age(20)
```

In [192]:
```
d.head()
```

Out[192]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female

In [193]:

```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_hdf('./example.h5', 'searchable_data', where=['Age={}'.format(
        int(age_of_interest))])

    return df
```

Changed to HDF5

In [193]:
```
def load_country_gender_at_given_age(age_of_interest):
    """
    Loads the country and gender of survey takes, which
    are of a given age

    Input:
        age_of_interest  int

    Output:
        dataframe

    """

    df = pd.read_hdf('./example.h5', 'searchable_data', where=['Age={}'.format(
        int(age_of_interest))])

    return df
```

In [194]:
```
d = load_country_gender_at_given_age(20)
```

In [195]:
```
d.head()
```

Out[195]:

	Country	Age	GenderSelect
10	Russia	20.0	Female
123	United States	20.0	Female
125	Turkey	20.0	Male
144	United States	20.0	Male
148	India	20.0	Male

Plan for changing data.

Very brief coding practices.

Reduce mistakes.

Aim for: Everything important on same screen page.

Meaningful active names, unless variable
used in same paragraph, e.g.: “is_stored”

Avoid duplications.

Use GitHub to track code and collaborate (free private repositories through their educational service); every ~20 minutes. Consider GUI (GitHub desktop) rather than command line.

Any further recommendation?

Plan for changing data.

Very brief coding practices.

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment (very short)

- Language
- Computer
- Space
- Constant learning

Organization

- Data
- Code
- Computation ← Here

Summary

What did we miss?

After lecture (extra): getting your environment ready

Save time!

Reduce time coding.

Reduce time running code.

Save time!

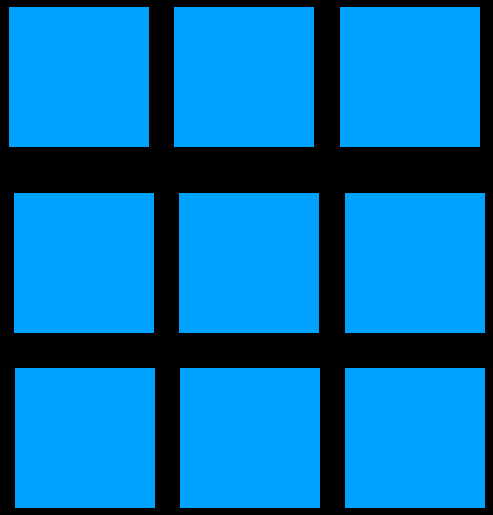
Important to many.

Reduce time coding.

Specialist applications.

Reduce time running code.

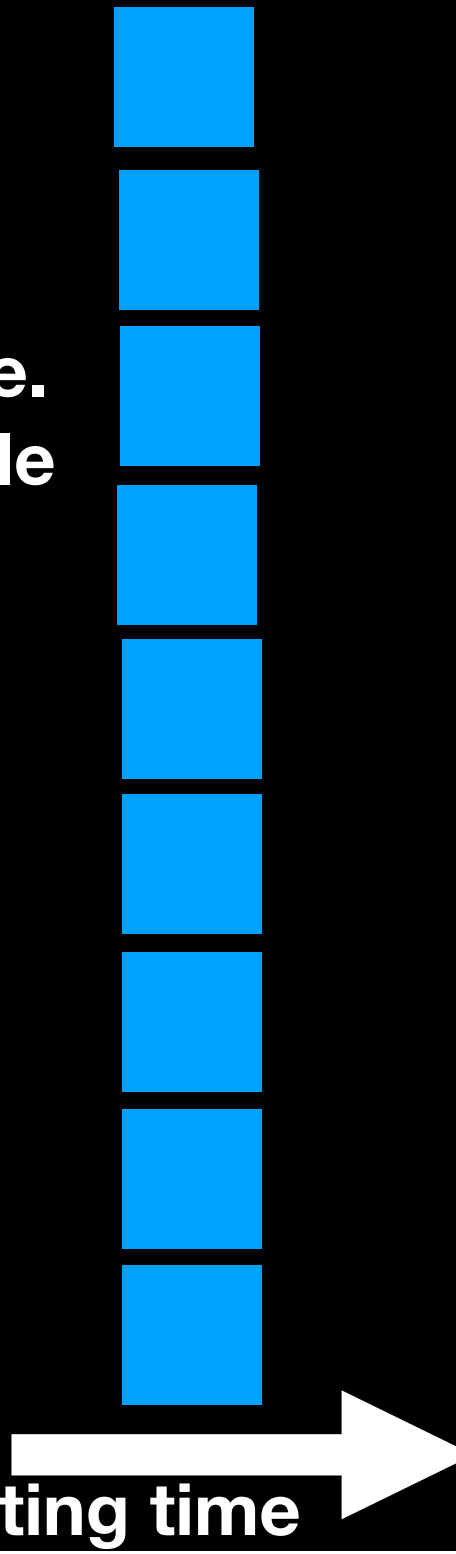
How to organize many computations?



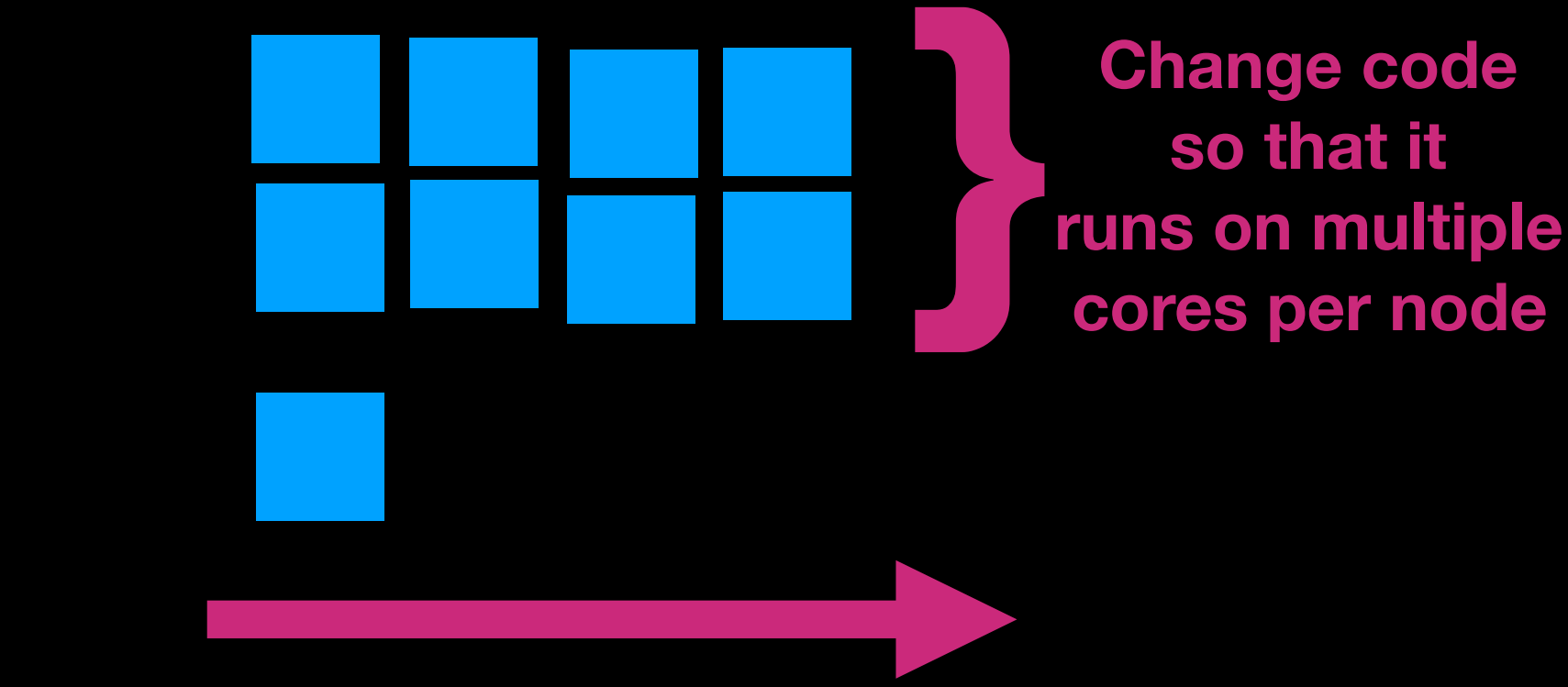
e.g.: iterations of a model

Ideal computing strategy:

Parallelize to different nodes, and use one core per node. (cluster distributes jobs over single cores and nodes)

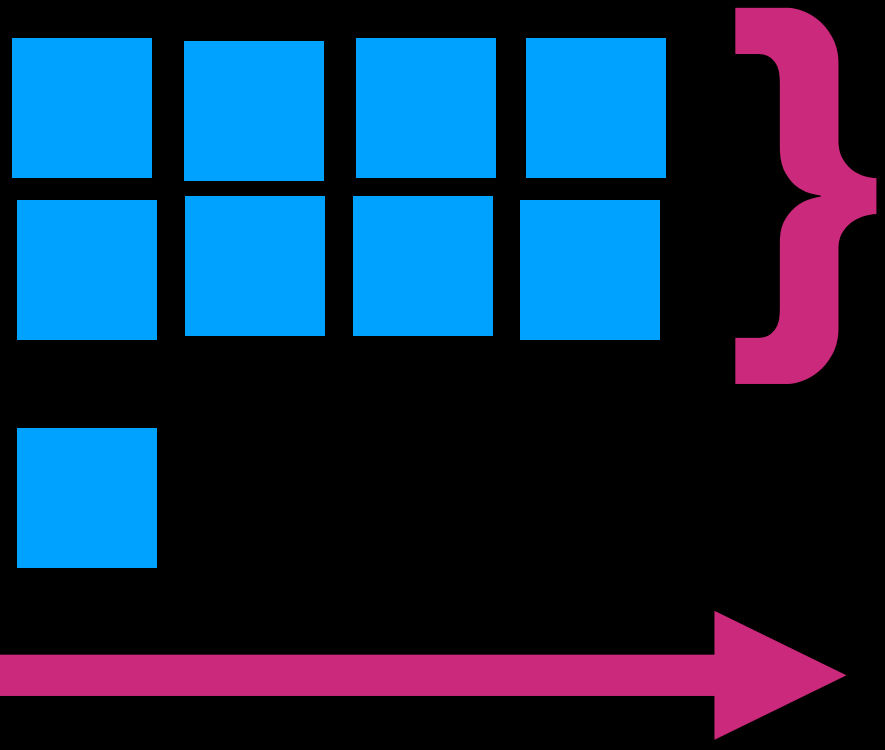


Attention: The prior strategy doesn't work on all clusters.



Attention: The prior strategy doesn't work on all clusters.

Alternatives seen:
- own clusters
- amazon cloud
- reorganize computations



Plan for changing data.

Very brief coding practices.

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Here

Organization

- Data
- Code
- Computation

Summary

What did we miss?

After lecture (extra): getting your environment ready

One last four minute exercise.

Discuss in groups of 3.

What is applicable to your research? What data science problems of your field are not covered?

What is common to your data science?

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Here

Organization

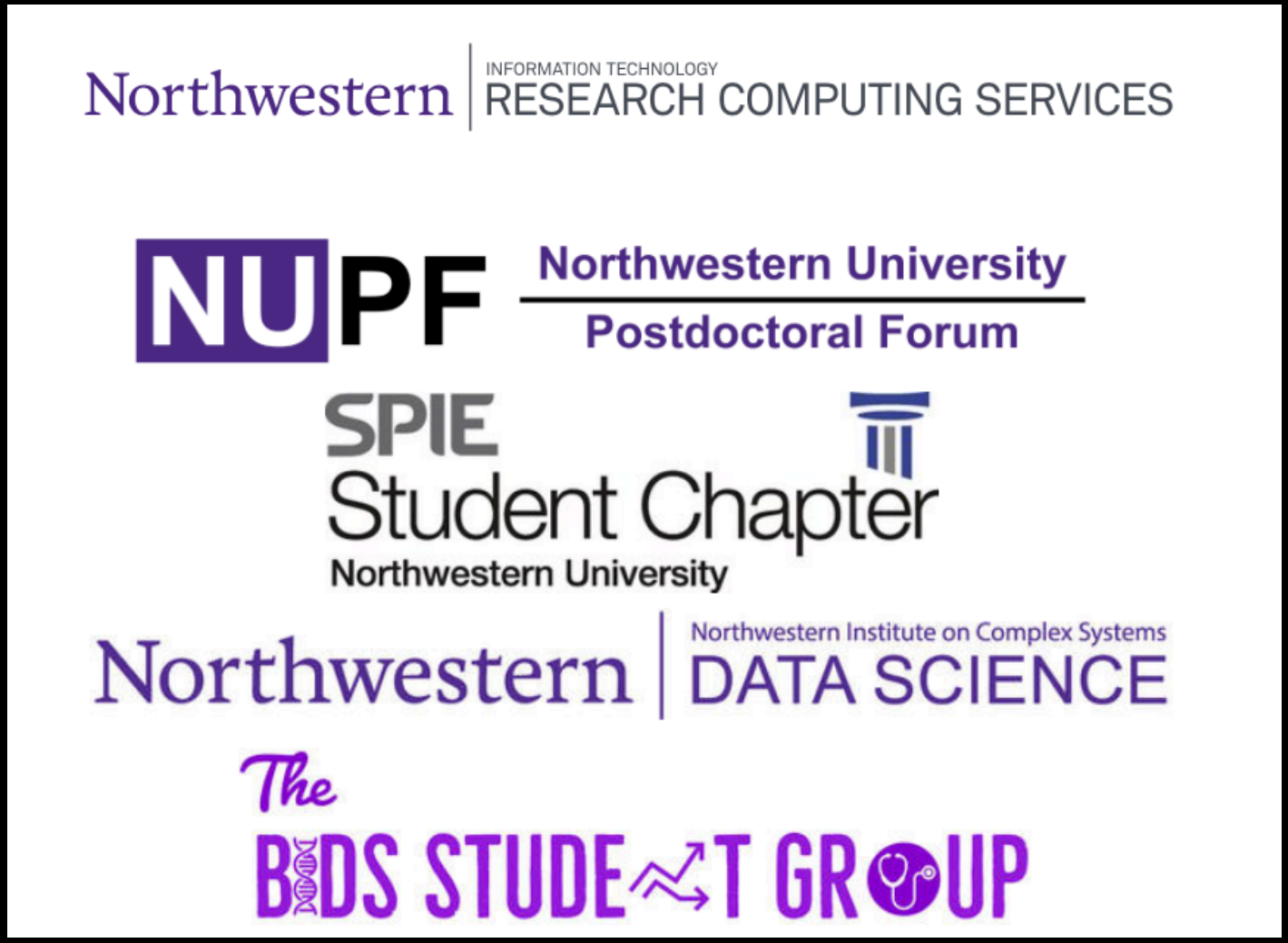
- Data
- Code
- Computation

Summary

What did we miss?

After lecture (extra): getting your environment ready

Thank you!


Northwestern
INFORMATION TECHNOLOGY
RESEARCH COMPUTING SERVICES
NUPF
Northwestern University
Postdoctoral Forum
SPIE
Student Chapter
Northwestern University
Northwestern
Northwestern Institute on Complex Systems
DATA SCIENCE
The
BIDS STUDENT GROUP

Workflow

- Anticipate problems
- Get Overview
- Clean data
- Find connections
- How science differs

Environment
(very short)

- Language
- Computer
- Space
- Constant learning

Here

Organization

- Data
- Code
- Computation

Summary

What did we miss?

After lecture (extra): getting your environment ready