university of groningen

umcg

A python pipeline for data analysis in psychology

Haiyang Geng

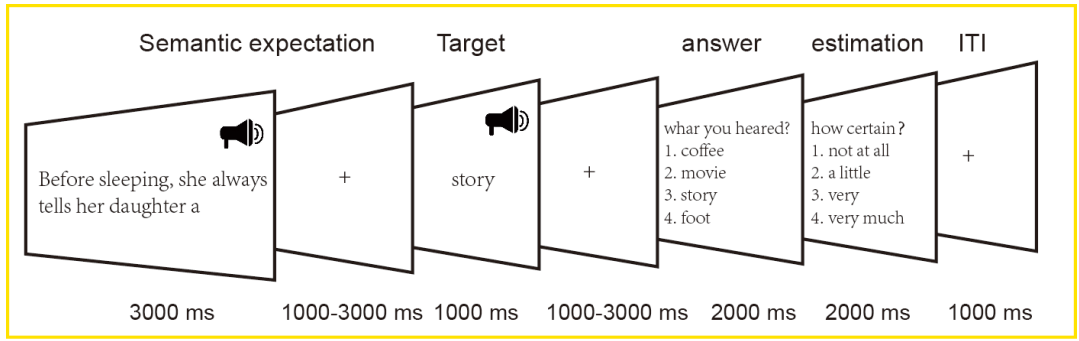
15th Jan 2020

@CNC, UMCG, Groningen

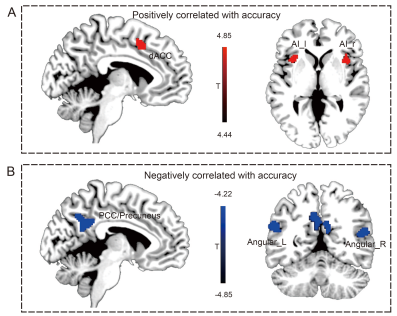
What types of data we analyze by using this python pipeline

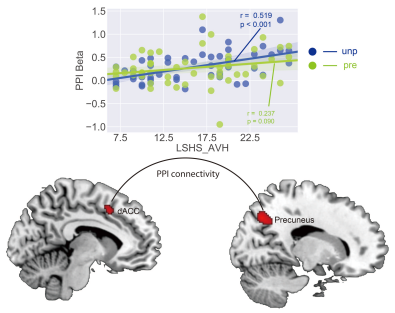
python pipeline

1. Reaction times and response



2. Beta values of brain activation and connectivity





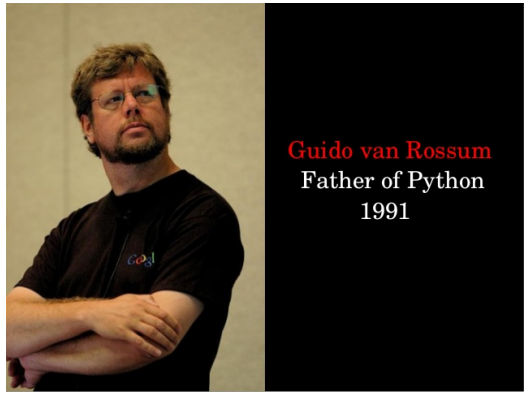
3. Questionnaires and demographic data

(PhD) Life is short,

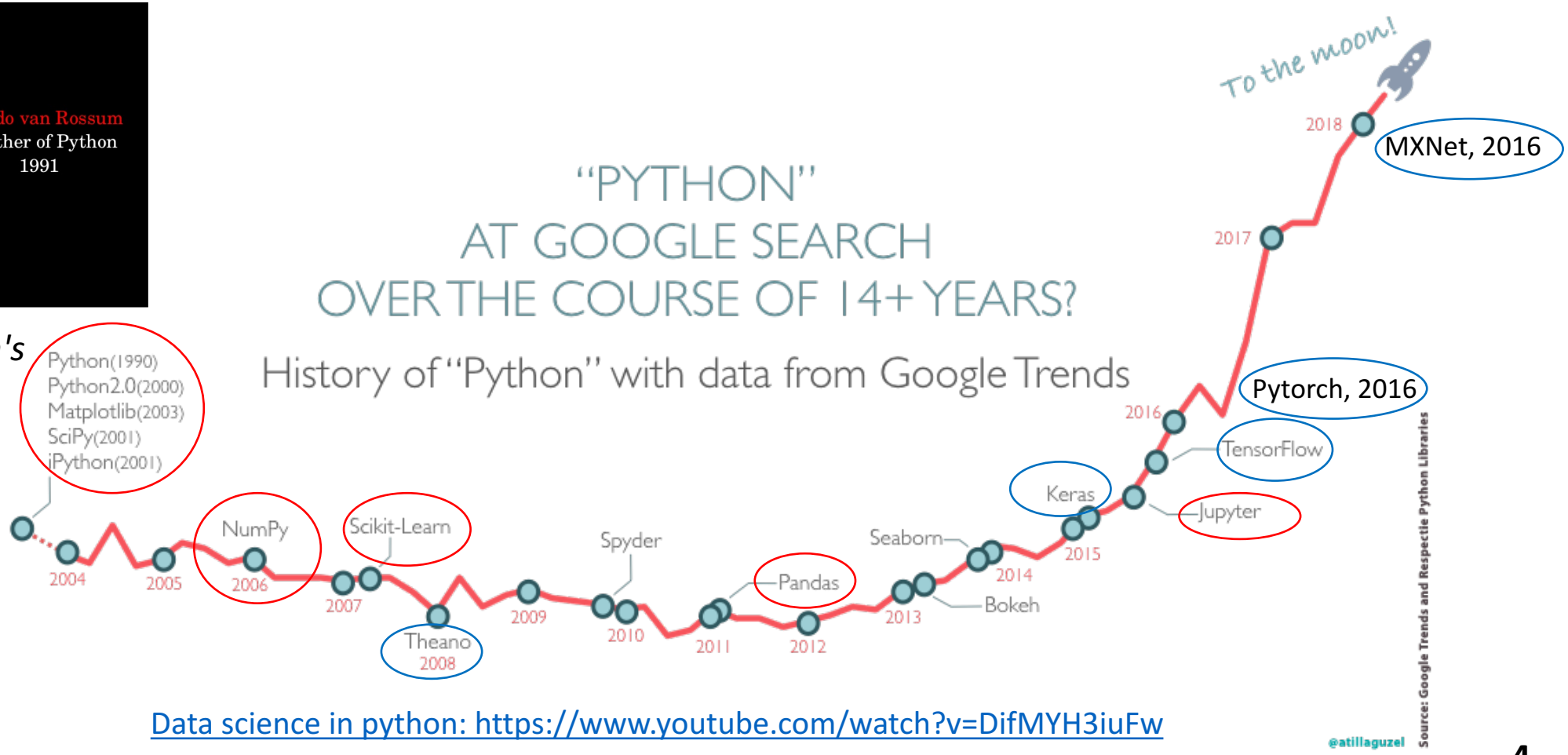
you need python (in psychology)

Bruce Eckel (Haiyang Geng)

A brief history of python



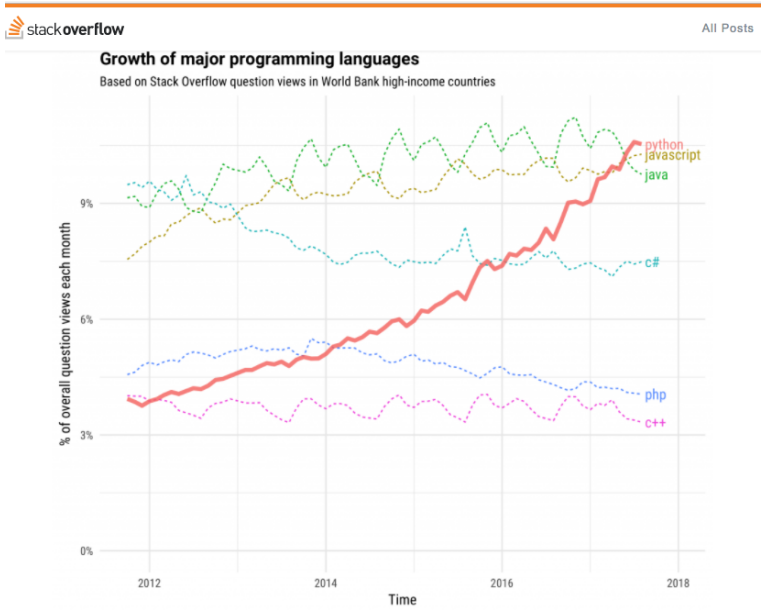
Monty Python's Flying Circus



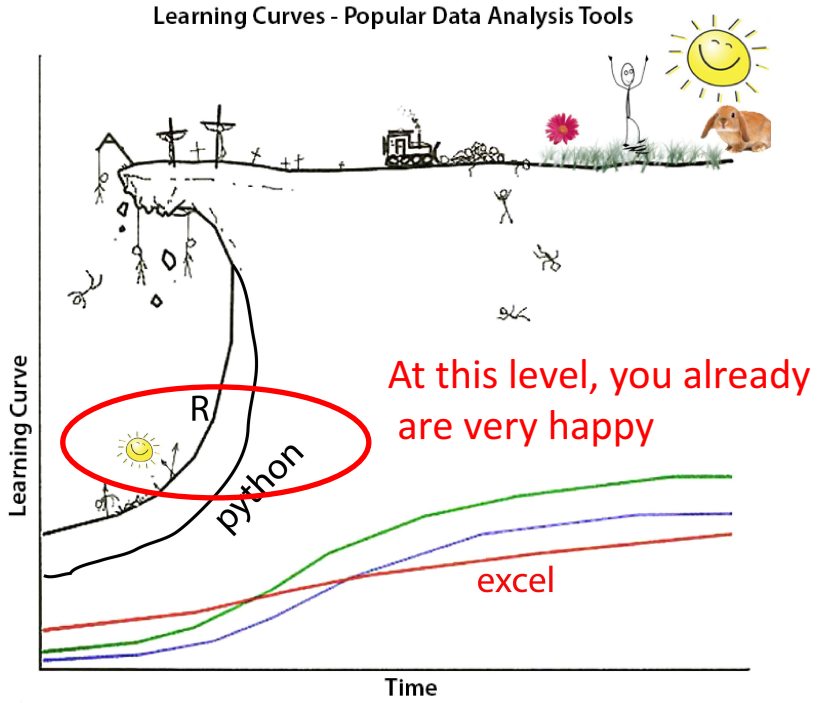
Data science in python: https://www.youtube.com/watch?v=DifMYH3iuFw

Why do we need python?

1. Big and active community:
- Abundant open source codes
- Easy to solve problems

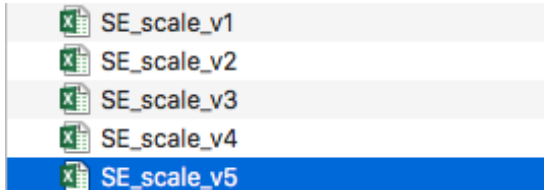


2. Python is easier to start:
- Not spend much time to learn
- Save you lots of time



https://nceas.github.io/oss-lessons/spatial-data-gis-law/1-mon-spatial-data-intro.html

3. Python is friendly for version control:
- Excel is a mess for recording history
- Everything can be written in python code
  - (tracked by Git)

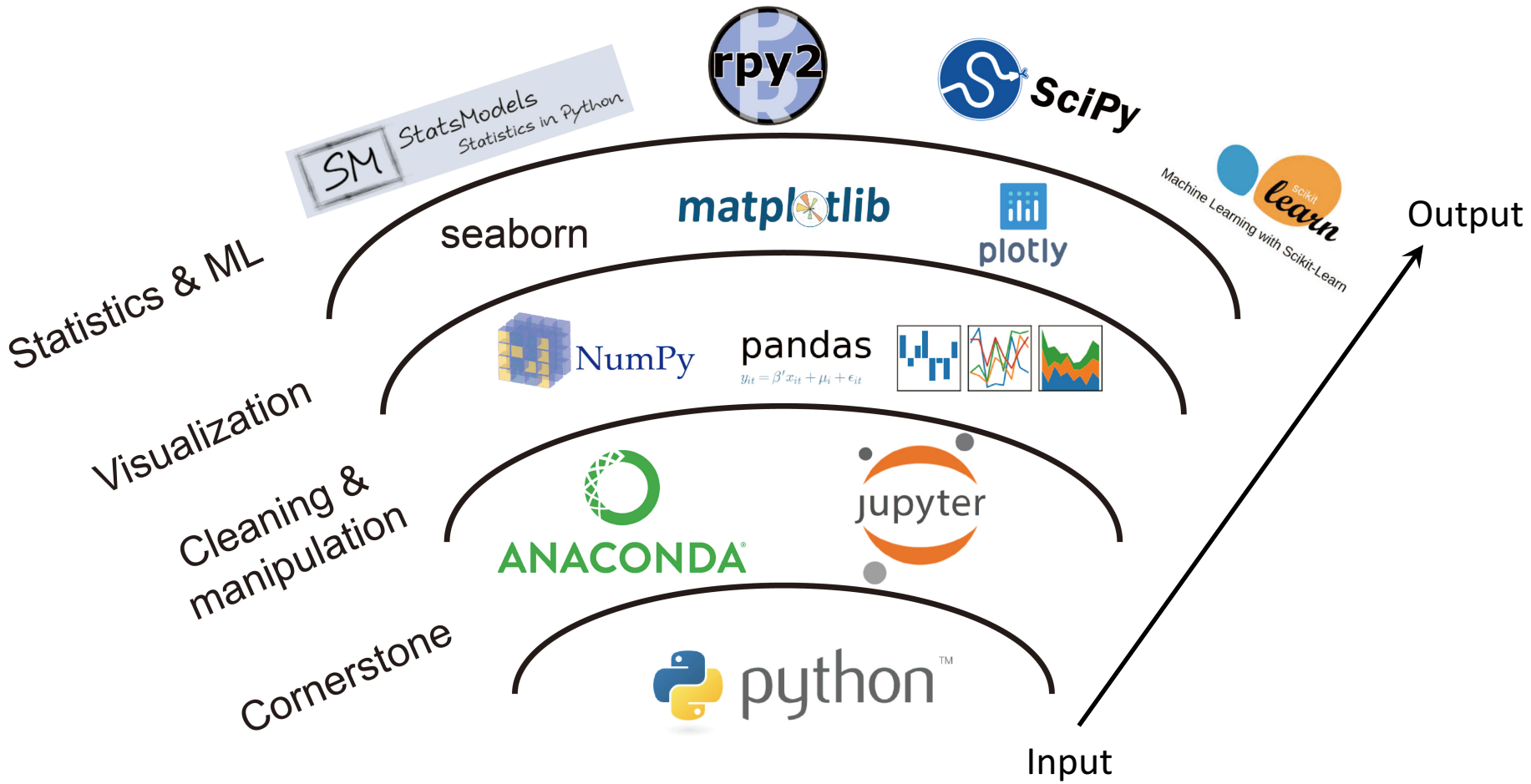



```
# combine with questionaire data
se_scale = pd.read_excel('SE_scale.xlsx')
se_scale  = se_scale.set_index('Subject')
#print(se_scale)
#se_scale = se_scale.set_index(['Subject'])
se_scale = se_scale.iloc[:,7:76]
opt_all = opt_all.join(se_scale)
se_qpe = pd.read_excel('SE-QPE_40.xlsx')
se_qpe = se_qpe.set_index('Subject')
QPE_severity = pd.DataFrame(se_qpe.iloc[:,12:16].sum(axis = 1),columns = ['QPE_severity'])
opt_all = opt_all.join(QPE_severity)
```

4. Excellent ecosystem (packages) for data science
- Numpy, pandas, scipy and scikit-learn…

Why do we need python?

- Excellent ecosystem (packages) for data science

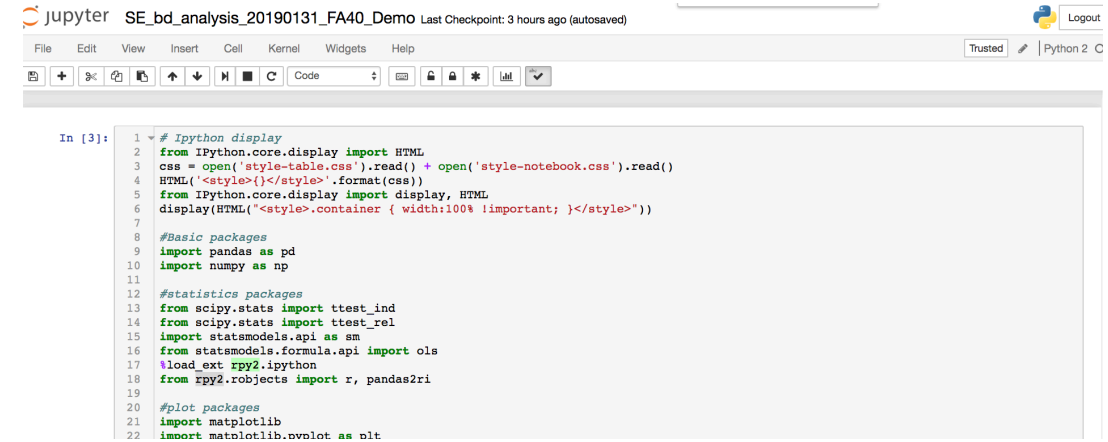


Cornerstone in python ecosystem

conda create
conda install

ANACONDA®

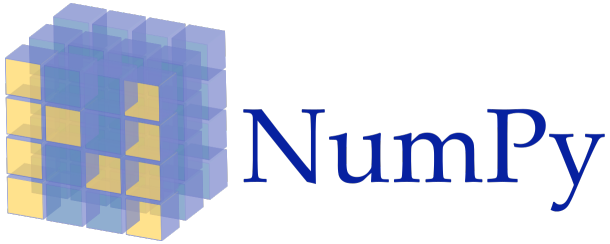
The World's Most Popular Python/R Data Science Platform



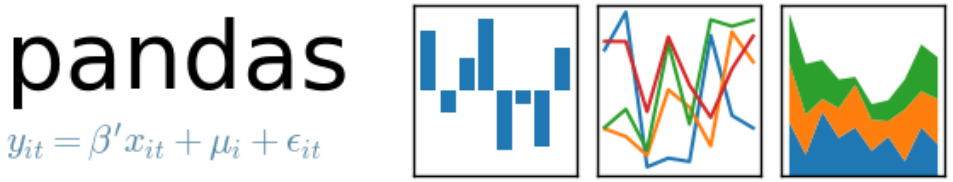
An open-source, interactive, light and web browser-based python GUI

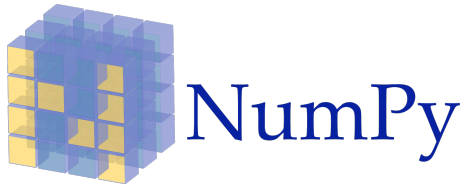
Data manipulation and cleaning

For matrix (Matlab)



For table (R)





For matrix data
Neuroimaging data

Python For Data Science *Cheat Sheet*

NumPy Basics

Learn Python for Data Science Interactively at www.DataCamp.com

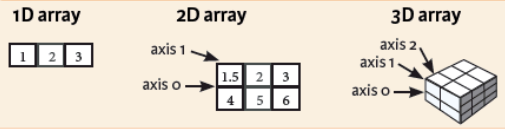
NumPy

The **NumPy** library is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Use the following import convention:

```
>>> import numpy as np
```

NumPy Arrays



Creating Arrays

```
>>> a = np.array([1,2,3])
>>> b = np.array([(1.5,2,3), (4,5,6)], dtype = float)
>>> c = np.array([[(1.5,2,3), (4,5,6)], [(3,2,1), (4,5,6)]],
                 dtype = float)
```

Initial Placeholders

```
>>> np.zeros((3,4))                  Create an array of zeros
>>> np.ones((2,3,4),dtype=np.int16)  Create an array of ones
>>> d = np.arange(10,25,5)           Create an array of evenly
                                     spaced values (step value)
>>> np.linspace(0,2,9)               Create an array of evenly
                                     spaced values (number of samples)
>>> e = np.full((2,2),7)             Create a constant array
>>> f = np.eye(2)                    Create a 2X2 identity matrix
>>> np.random.random((2,2))          Create an array with random values
>>> np.empty((3,2))                  Create an empty array
```

I/O

Saving & Loading On Disk

```
>>> np.save('my_array', a)
>>> np.savez('array.npz', a, b)
>>> np.load('my_array.npy')
```

Saving & Loading Text Files

```
>>> np.loadtxt("myfile.txt")
>>> np.genfromtxt("my_file.csv", delimiter=',')
>>> np.savetxt("myarray.txt", a, delimiter=" ")
```

Data Types

```
>>> np.int64       Signed 64-bit integer types
>>> np.float32     Standard double-precision floating point
>>> np.complex     Complex numbers represented by 128 floats
>>> np.bool        Boolean type storing TRUE and FALSE values
>>> np.object      Python object type
>>> np.string_     Fixed-length string type
>>> np.unicode_    Fixed-length unicode type
```

Inspecting Your Array

```
>>> a.shape          Array dimensions
>>> len(a)           Length of array
>>> b.ndim           Number of array dimensions
>>> e.size           Number of array elements
>>> b.dtype          Data type of array elements
>>> b.dtype.name     Name of data type
>>> b.astype(int)    Convert an array to a different type
```

Asking For Help

```
>>> np.info(np.ndarray.dtype)
```

Array Mathematics

Arithmetic Operations

```
>>> g = a - b                        Subtraction
  array([[-0.5,  0. ,  0. ],
         [-3. , -3. , -3. ]])
>>> np.subtract(a,b)                 Subtraction
>>> b + a                            Addition
  array([[ 2.5,  4. ,  6. ],
         [ 5. ,  7. ,  9. ]])
>>> np.add(b,a)                      Addition
>>> a / b                            Division
  array([[ 0.66666667,  1.        ,  1.        ],
         [ 0.25      ,  0.4       ,  0.5       ]])
>>> np.divide(a,b)                   Division
>>> a * b                            Multiplication
  array([[  1.5,   4. ,   9. ],
         [  4. ,  10. ,  18. ]])
>>> np.multiply(a,b)                 Multiplication
>>> np.exp(b)                        Exponentiation
>>> np.sqrt(b)                       Square root
>>> np.sin(a)                        Print sines of an array
>>> np.cos(b)                        Element-wise cosine
>>> np.log(a)                        Element-wise natural logarithm
>>> e.dot(f)                         Dot product
  array([[ 7.,  7.],
         [ 7.,  7.]])
```

Comparison

```
>>> a == b                           Element-wise comparison
  array([[False,  True,  True],
         [False, False, False]], dtype=bool)
>>> a < 2                            Element-wise comparison
  array([True, False, False], dtype=bool)
>>> np.array_equal(a, b)             Array-wise comparison
```

Aggregate Functions

```
>>> a.sum()              Array-wise sum
>>> a.min()              Array-wise minimum value
>>> b.max(axis=0)        Maximum value of an array row
>>> b.cumsum(axis=1)     Cumulative sum of the elements
>>> a.mean()             Mean
>>> b.median()           Median
>>> a.corrcoef()         Correlation coefficient
>>> np.std(b)            Standard deviation
```

Copying Arrays

```
>>> h = a.view()         Create a view of the array with the same data
>>> np.copy(a)           Create a copy of the array
>>> h = a.copy()         Create a deep copy of the array
```

Sorting Arrays

```
>>> a.sort()             Sort an array
>>> c.sort(axis=0)       Sort the elements of an array's axis
```

Subsetting, Slicing, Indexing

Also see Lists

Subsetting

```
>>> a[2]                 Select the element at the 2nd index
  3
>>> b[1,2]               Select the element at row 0 column 2
  6.0                    (equivalent to b[1][2])
```

Slicing

```
>>> a[0:2]               Select items at index 0 and 1
  array([1, 2])
>>> b[0:2,1]             Select items at rows 0 and 1 in column 1
  array([ 2.,  5.])
>>> b[:1]                Select all items at row 0
  array([[1.5, 2., 3.]])  (equivalent to b[0:1, :])
>>> c[1,...]             Same as [1,:,:]
  array([[[ 3.,  2.,  1.],
          [ 4.,  5.,  6.]]])
>>> a[ : :-1]            Reversed array a
  array([3, 2, 1])
```

Boolean Indexing

```
>>> a[a<2]               Select elements from a less than 2
  array([1])
```

Fancy Indexing

```
>>> b[[1, 0, 1, 0],[0, 1, 2, 0]]   Select elements (1,0),(0,1),(1,2) and (0,0)
  array([ 4. ,  2. ,  6. ,  1.5])
>>> b[[1, 0, 1, 0]][:,[0,1,2,0]]   Select a subset of the matrix's rows
  array([[ 4. ,  5. ,  6. ,  4. ],  and columns
         [ 1.5,  2. ,  3. ,  1.5],
         [ 4. ,  5. ,  6. ,  4. ],
         [ 1.5,  2. ,  3. ,  1.5]])
```

Array Manipulation

Transposing Array

```
>>> i = np.transpose(b)     Permute array dimensions
>>> i.T                     Permute array dimensions
```

Changing Array Shape

```
>>> b.ravel()               Flatten the array
>>> g.reshape(3,-2)         Reshape, but don't change data
```

Adding/Removing Elements

```
>>> h.resize((2,6))         Return a new array with shape (2,6)
>>> np.append(h,g)          Append items to an array
>>> np.insert(a, 1, 5)      Insert items in an array
>>> np.delete(a,[1])        Delete items from an array
```

Combining Arrays

```
>>> np.concatenate((a,d),axis=0)     Concatenate arrays
  array([ 1,  2,  3, 10, 15, 20])
>>> np.vstack((a,b))                 Stack arrays vertically (row-wise)
  array([[ 1. ,  2. ,  3. ],
         [ 1.5,  2. ,  3. ],
         [ 4. ,  5. ,  6. ]])
>>> np.r_[e,f]                       Stack arrays vertically (row-wise)
>>> np.hstack((e,f))                 Stack arrays horizontally (column-wise)
  array([[ 7.,  7.,  1.,  0.],
         [ 7.,  7.,  0.,  1.]])
>>> np.column_stack((a,d))           Create stacked column-wise arrays
  array([[ 1, 10],
         [ 2, 15],
         [ 3, 20]])
>>> np.c_[a,d]                       Create stacked column-wise arrays
```

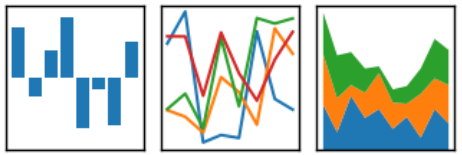
Splitting Arrays

```
>>> np.hsplit(a,3)                   Split the array horizontally at the 3rd
  [array([1]),array([2]),array([3])]  index
>>> np.vsplit(c,2)                   Split the array vertically at the 2nd index
[array([[[ 1.5,  2. ,  1. ],
         [ 4. ,  5. ,  6. ]]]),
 array([[[ 3.,  2.,  3.],
         [ 4.,  5.,  6.]]])]
```

DataCamp
Learn Python for Data Science Interactively

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



http://pandas.pydata.org

For table data

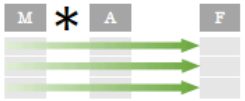
read_csv
df.head, describe
df.join
df[‘clm’],df.iloc
pd.concat
groupy

Tidy Data – A foundation for wrangling in pandas

Each **variable** is saved in its own **column**

Each **observation** is saved in its own **row**

Tidy data complements pandas's vectorized operations. pandas will automatically preserve observations as you manipulate variables. No other format works as intuitively with pandas.



M * A

Syntax – Creating DataFrames

	a	b	c
1	4	7	10
2	5	8	11
3	6	9	12

```
df = pd.DataFrame(
        {"a" : [4 ,5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
        index = [1, 2, 3])
```
Specify values for each column.

```
df = pd.DataFrame(
    [[4, 7, 10],
     [5, 8, 11],
     [6, 9, 12]],
    index=[1, 2, 3],
    columns=['a', 'b', 'c'])
```
Specify values for each row.

		a	b	c
n	v			
d	1	4	7	10
	2	5	8	11
e	2	6	9	12

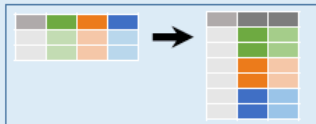
```
df = pd.DataFrame(
        {"a" : [4 ,5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
index = pd.MultiIndex.from_tuples(
        [('d',1),('d',2),('e',2)],
          names=['n','v']))
```
Create DataFrame with a MultiIndex

Method Chaining

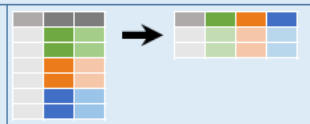
Most pandas methods return a DataFrame so that another pandas method can be applied to the result. This improves readability of code.

```
df = (pd.melt(df)
        .rename(columns={
                'variable' : 'var',
                'value' : 'val'})
        .query('val >= 200')
    )
```

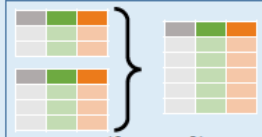
Reshaping Data – Change the layout of a data set



`pd.melt(df)`
Gather columns into rows.



`df.pivot(columns='var', values='val')`
Spread rows into columns.



`pd.concat([df1,df2])`
Append rows of DataFrames

`pd.concat([df1,df2], axis=1)`
Append columns of DataFrames

`df.sort_values('mpg')`
Order rows by values of a column (low to high).

`df.sort_values('mpg',ascending=False)`
Order rows by values of a column (high to low).

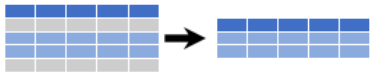
`df.rename(columns = {'y':'year'})`
Rename the columns of a DataFrame

`df.sort_index()`
Sort the index of a DataFrame

`df.reset_index()`
Reset index of DataFrame to row numbers, moving index to columns.

`df.drop(columns=['Length','Height'])`
Drop columns from DataFrame

Subset Observations (Rows)



`df[df.Length > 7]`
Extract rows that meet logical criteria.

`df.drop_duplicates()`
Remove duplicate rows (only considers columns).

`df.head(n)`
Select first n rows.

`df.tail(n)`
Select last n rows.

`df.sample(frac=0.5)`
Randomly select fraction of rows.

`df.sample(n=10)`
Randomly select n rows.

`df.iloc[10:20]`
Select rows by position.

`df.nlargest(n, 'value')`
Select and order top n entries.

`df.nsmallest(n, 'value')`
Select and order bottom n entries.

Logic in Python (and pandas)			
<	Less than	!=	Not equal to
>	Greater than	`df.column.isin(values)`	Group membership
==	Equals	`pd.isnull(obj)`	Is NaN
<=	Less than or equals	`pd.notnull(obj)`	Is not NaN
>=	Greater than or equals	&,|,~,^,df.any(),df.all()	Logical and, or, not, xor, any, all

Subset Variables (Columns)

`df[['width','length','species']]`
Select multiple columns with specific names.

`df['width']` *or* `df.width`
Select single column with specific name.

`df.filter(regex='regex')`
Select columns whose name matches regular expression *regex*.

regex (Regular Expressions) Examples	
'\.'	Matches strings containing a period '.'
'Length$'	Matches strings ending with word 'Length'
'^Sepal'	Matches strings beginning with the word 'Sepal'
'^x[1-5]$'	Matches strings beginning with 'x' and ending with 1,2,3,4,5
'^(?!Species$).*'	Matches strings except the string 'Species'

`df.loc[:,'x2':'x4']`
Select all columns between x2 and x4 (inclusive).

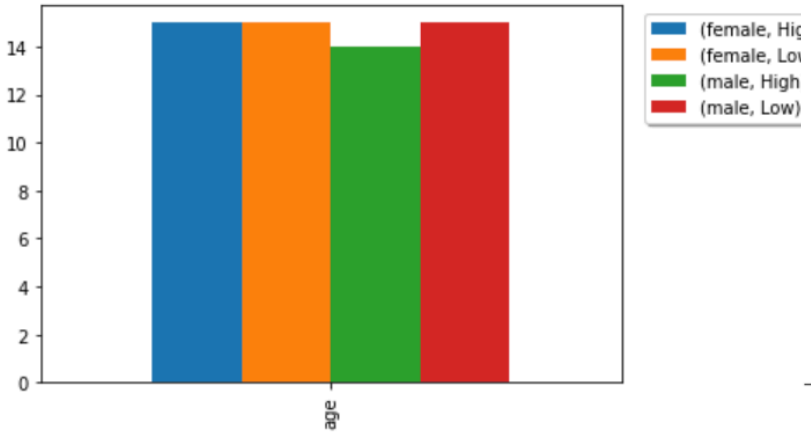
`df.iloc[:,[1,2,5]]`
Select columns in positions 1, 2 and 5 (first column is 0).

`df.loc[df['a'] > 10, ['a','c']]`
Select rows meeting logical condition, and only the specific columns .

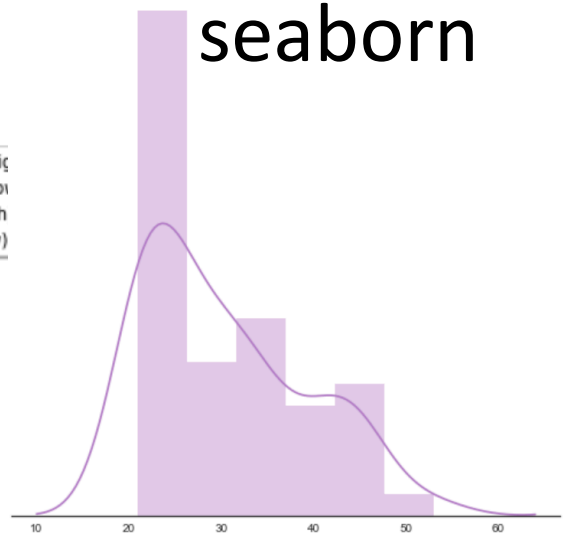
http://pandas.pydata.org/ This cheat sheet inspired by Rstudio Data Wrangling Cheatsheet (https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf) Written by Irv Lustig, Princeton Consultants

Data visualization

matplotlib



seaborn



plotly

Easy → Hard (Hacking)

Low → High (Beauty)

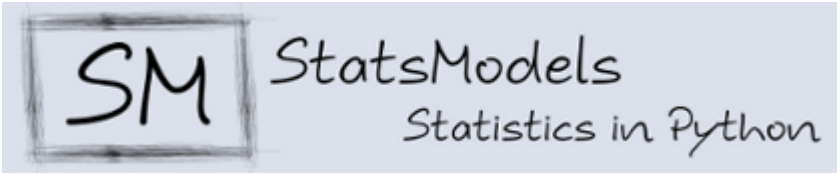
Hard ← Easy (Easy to use)

Statistics

Basic statistics (t test, Pearson correlation): scipy.stats


```
ttest_ind(opt_all_fa_High_unp,opt_all_fa_Low_unp).statistic
```

Advanced statistics (GLM and AVOVA): statsmodels

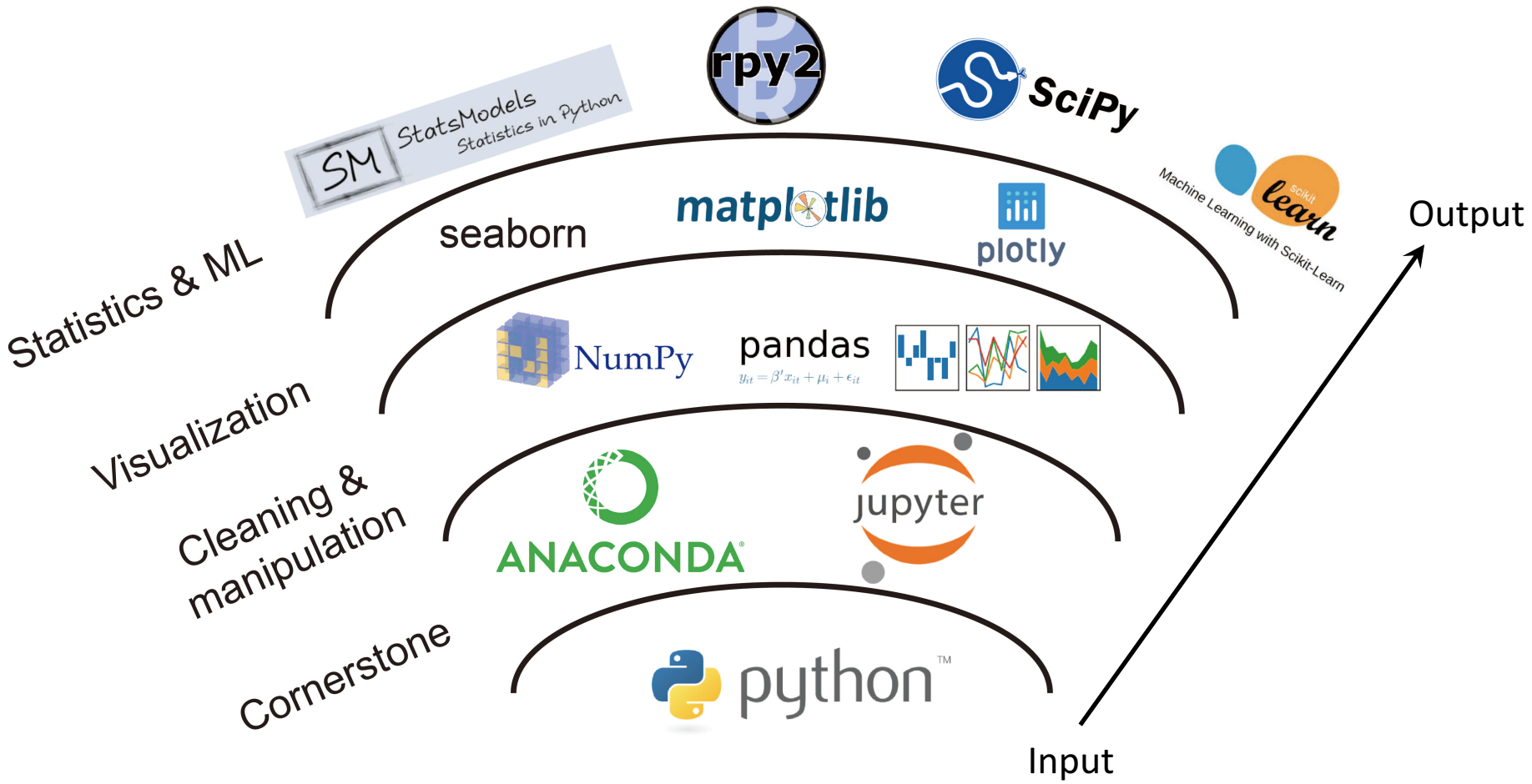



```
results = smf.ols('Lottery ~ Literacy + np.log(Pop1831)',
data=dat).fit()
```

More advanced statistics (everything from R): rpy2


```
%R -i r_df_bd_val r_df_bd_val$Subject <-
factor(r_df_bd_val$Subject);aov_group_conditon <- aov(conf ~
group*condition + Error(Subject/condition),
data=r_df_bd_val);print(summary(aov_group_conditon));print(model.t
ables(aov_group_conditon, "means"))
```

A pipeline for psychological data analysis



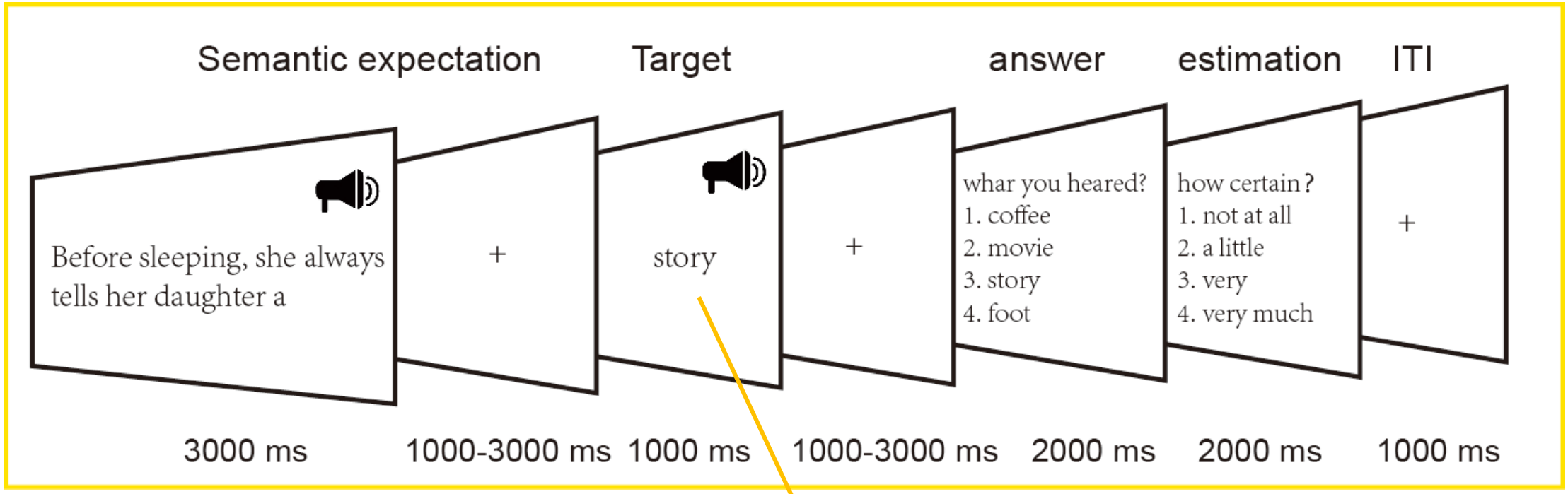
Programming saves your time to do more important things

- Standardize your pipeline by programming, make your and other's life easy
- Put all the data and codes into one box, avoid confusion, data missing and mistakes
- Look into the details of your data by visualizing, remove bad (noisy) data to make your data clean

An example: get your hand dirty

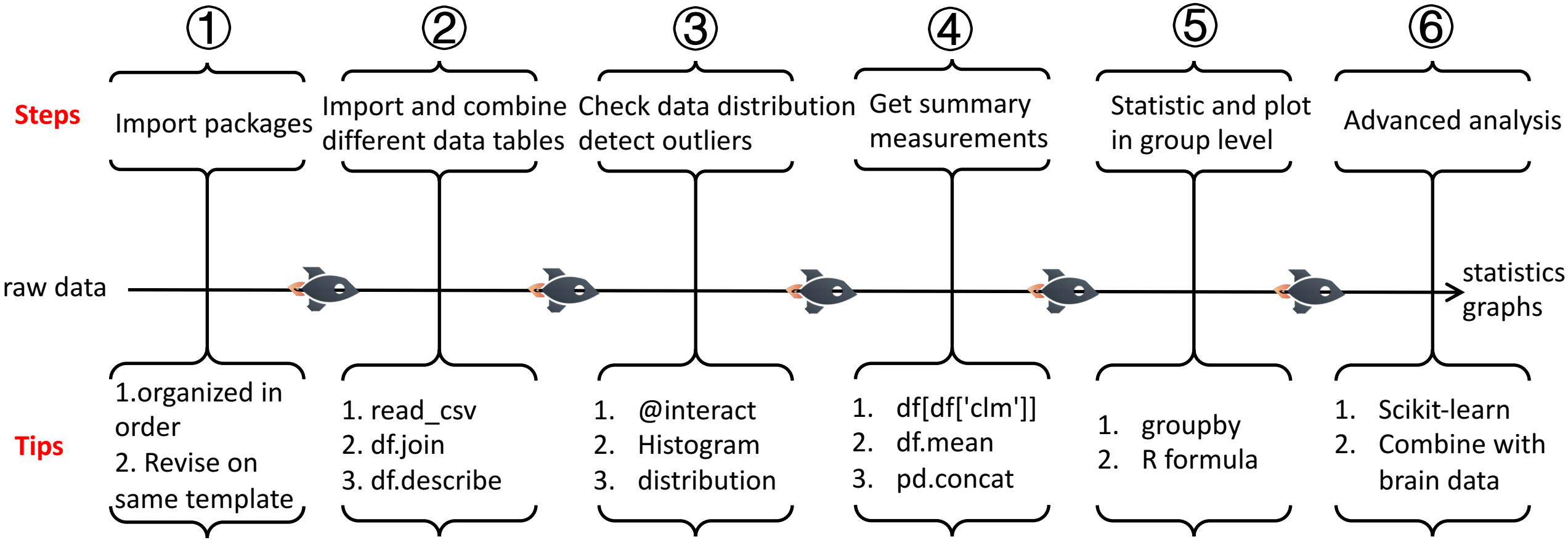
Task design

Semantic Expectation Task



Predictable condition: story, 60 trials in four sessions
Unpredictable condition：coffee, 60 trials in four sessions

Summary for the demo pipeline



Several useful tips (suggestions) for using this pipeline

1. During programming
   - Use Dash and Stack Overflow, which can almost solve all the (programing) questions.
   - Use Anaconda to manage different python environments
   - Collect optimal tools (packages) to polish your pipeline regularly
2. During data analysis
   - Check data before and after manipulation by using visualization and summary statistic
   - Validate your code by using 'benchmark' software (tool) (spss, jasp, R) when doing complex statistic
3. About algorithms
   - Implementation is easy, understanding is hard, put more time on figuring out underlying algorithms

Thanks for your attention