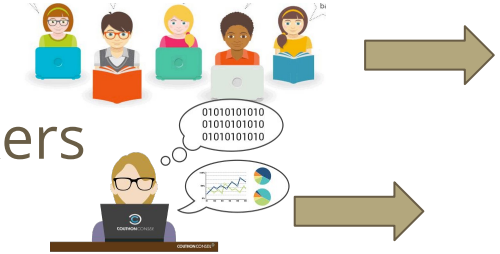
# PISAnalysisTool

Nixi Wang, Hongbin Qu, Luyu Xu, Shenghao Xie

# User cases

- Teachers, students, and parents
- Education researchers, policy makers and data analysts





| | | | | | | |
|---|---|---|---|---|---|---|
| 11219 | PV6SSES | Plausible Value 6 in System Subscale of Science - Earth | NUM | 8.3 | 915 | 0 - 934.21 |
| 11220 | | | | | | |
| 11221 | PV7SSES | Plausible Value 7 in System Subscale of Science - Earth | NUM | 8.3 | 916 | 34.017 - 922.825 |
| 11222 | | | | | | |
| 11223 | PV8SSES | Plausible Value 8 in System Subscale of Science - Earth | NUM | 8.3 | 917 | 34.746 - 935.05 |
| 11224 | | | | | | |
| 11225 | PV9SSES | Plausible Value 9 in System Subscale of Science - Earth | NUM | 8.3 | 918 | 12.619 - 928.349 |
| 11226 | | | | | | |
| 11227 | PV10SSES | Plausible Value 10 in System Subscale of Science - Eartl | NUM | 8.3 | 919 | 16.907 - 898.335 |
| 11228 | | | | | | |
| 11229 | SENWT | Senate Weight (sum of 5000 per country) | NUM | 12.5 | 920 | 0.00375 - 12.89587 |
| 11230 | | | | | | |
| 11231 | VER_DAT | Date of the database creation | CHAR | $20. | 921 | |
| 11232 | | | | | | |
| 11233 | | | | | | |

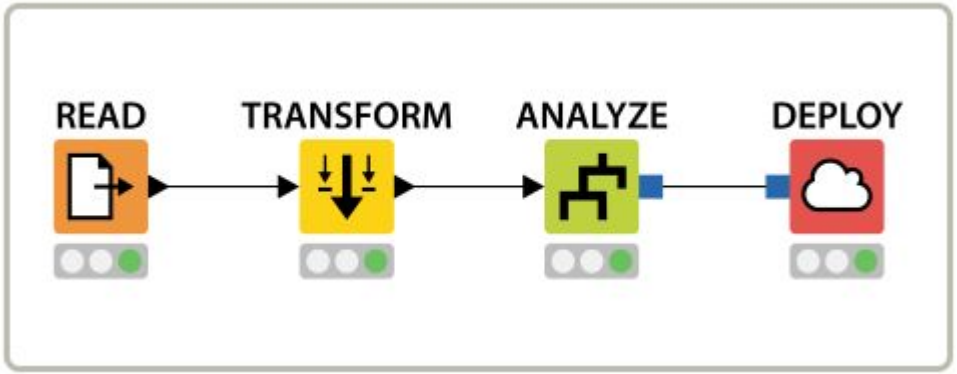
Student - QQQ | Student - QQ2 | School | Cognitive | Timing | Teacher | Financial Literacy | Coll. Pb. Solving

# Multiple data sources

**PISA: 2015**

- students.csv
- schools.csv
- teachers.csv

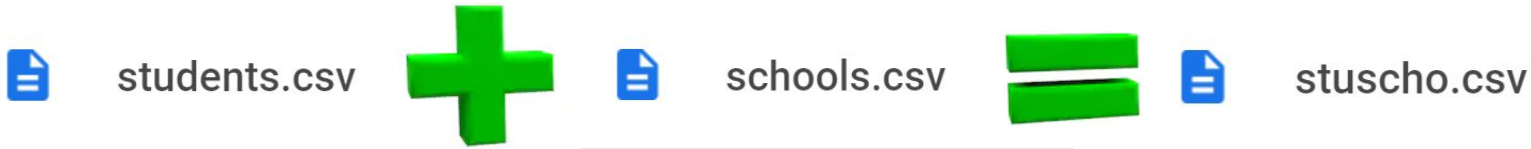
**WB: import wbdata (demo)**




**Four components:**

# Merge Data -- merge_csv.py

- Takes a student/teacher file as a parameter
- Merges the file with the school file based on:
  - ‘CNTSCHID’: School ID
  - 'SUBNATIO': Nation ID



# PISA Data -- get_students_info.py

- Takes merged school-student file and user-specified attributes as inputs
- Extracts user-specified attributes for modeling and visualization

http://35.247.97.57:8888/notebooks/Notebooks/get_student_info.ipynb

# PISA Data -- students_info_helper

1. Built nations.csv and IDs_sorted_by_student.csv

| | A | B | C | D |
|---|---|---|---|---|
| 1 | nationID | nationCode | continent | nationName |
| 2 | 80000 | ALB | EU | Albania |
| 3 | 120000 | DZA | AF | Algeria |
| 4 | 320100 | QAR | SA | Argentina - Ciudad A |
| 5 | 360000 | AUS | AP | Australia |
| 6 | 400000 | AUT | EU | Austria |
| 7 | 560000 | BEL | EU | Belgium |
| 8 | 760000 | BRA | SA | Brazil |
| 9 | 1000000 | BGR | EU | Bulgaria |
| 10 | 1240000 | CAN | NA | Canada |
| 11 | 1520000 | CHL | SA | Chile |
| 12 | 1580000 | TWN | AP | Taiwan |
| 13 | 1700000 | COL | SA | Colombia |
| 14 | 1880000 | CRI | SA | Costa Rica |
| 15 | 1910000 | HRV | EU | Croatia |
| 16 | 2030000 | CZE | EU | Czech Republic |
| 17 | 2080000 | DNK | EU | Denmark |
| 18 | 2140000 | DOM | NA | Dominican Republic |

nations.csv

| | A | B | C |
|---|---|---|---|
| 1 | CNTSCHID | CNTSTUID | SUBNATIO |
| 2 | 800029 | 800001 | 80000 |
| 3 | 800005 | 800002 | 80000 |
| 4 | 800013 | 800003 | 80000 |
| 5 | 800107 | 800004 | 80000 |
| 6 | 800054 | 800005 | 80000 |
| 7 | 800051 | 800006 | 80000 |
| 8 | 800097 | 800007 | 80000 |
| 9 | 800093 | 800008 | 80000 |
| 10 | 800095 | 800009 | 80000 |
| 11 | 800020 | 800010 | 80000 |
| 12 | 800231 | 800011 | 80000 |
| 13 | 800057 | 800012 | 80000 |
| 14 | 800065 | 800013 | 80000 |
| 15 | 800208 | 800014 | 80000 |
| 16 | 800123 | 800015 | 80000 |
| 17 | 800236 | 800016 | 80000 |
| 18 | 800075 | 800017 | 80000 |

IDs_sorted_by_student.csv

# PISA Data -- students_info_helper

2. Wrote student_info_helper.py

   Including support functions to get_students_info.py

   `sav_to_dataframe, student2school, school2nation, student2nation, infos`

# Hierarchical generalized linear model

**Exemplar inquiry:**

*** Gender difference in science assessment internationally***

- What school and country-level factors are contributing to science performance?
- What sociocultural factors are mediating the gender effect?

Jupyter notebook demo

http://35.247.97.57:8888/tree/Notebooks

# Folium visualization

http://35.247.97.57:8888/view/Notebooks/map_data.html

# Testing

```
df = pd.read_csv('sample data.csv', header=0)
colnames = ('Science', 'IBTEACH', 'WEALTH', 'ESCS', 'School_type',
            'Sch_science_resource', 'log_science', 'female')

def smoke_test_calculate_correlation(self):
    ''' Smoke test'''
```

```
def test_sav_to_datafram
    ''' Test for file no
    with self.assertRais
        sav_to_dataframe
    ''' Test for file wi
    with self.assertRais
        sav_to_dataframe
    ''' Test for a valid
    file = 'sample.sav'
    df = sav_to_datafram
    self.assertTrue(len(

def test_student2school(
    df = pd.read_csv('ID
    ''' Test for invalid
    with self.assertRais
        student2school(d
    with self.assertRais
        student2school(d
    ''' Test for valid s
    nID = student2school
    self.assertTrue(nID

def test_school2nation(s
    ''' Test for id less
    nID1 = school2nation
```

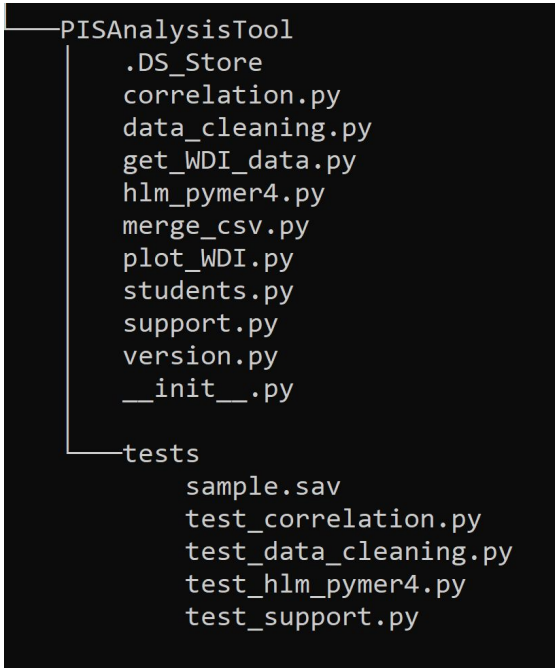
```
def test_school_type(self):
    ''' Test for valid variable'''
    series1 = pd.Series(['.I', '.M', 8, 9, 99, 1, 2])
    res1 = school_type(series1)
    self.assertTrue(res1 == pd.Series(['NaN', 'NaN', 'NaN', 'NaN', 'NaN', 1, -1]))
    ''' Test for invalid variable'''
    series2 = pd.Series(['invalid', 10])
    with self.assertRaises(KeyError):
        res2 = school_type(series2)

def smoke_test_data_cleaning(self):
    ''' Smoke test'''
    df = pd.read_csv("student_info.csv", encoding='latin-1', na_values=['', ' '])
    try:
        data_cleaning(df)
    except:
        print('Smoke test failed!')
    print('Smoke test passed!')

def smoke_test_small_samle(self):
    ''' Smoke test'''
    df = pd.read_csv('data.csv', header=0)
    try:
        small_sample(df)
    except:
        print('Smoke test failed!')
    print('Smoke test passed!')
```

# Project structure

- README.md
- LICENSE
- requirements.txt
- Unit tests
- A setup.py for PyPi (i.e. pip) installation
  - https://pypi.org/project/PISAnalysisTool/
- An autodoc generation system (Sphinx)
  - https://github.com/nixiwang/PISAnalysisTool



# Moving forward

- Refine component and function specifications
- Refine documentation, PEP 8 compliance
- Develop more generic functions
    - ICCs
    - Comparing model fits
- Continuous integration
- Test coverage