DataLab: Introducing Software Engineering Thinking into Data Science Education at Scale

清華大學
Tsinghua University

交叉信息研究院
Institute for Interdisciplinary Information Sciences

Yang Zhang, Tingjian Zhang, Yongzheng Jia, Jiao Sun, Fangzhou Xu, and Wei Xu

Institute of Interdisciplinary Information Sciences, Tsinghua University

Department of Computer Science and Technology, Shandong University

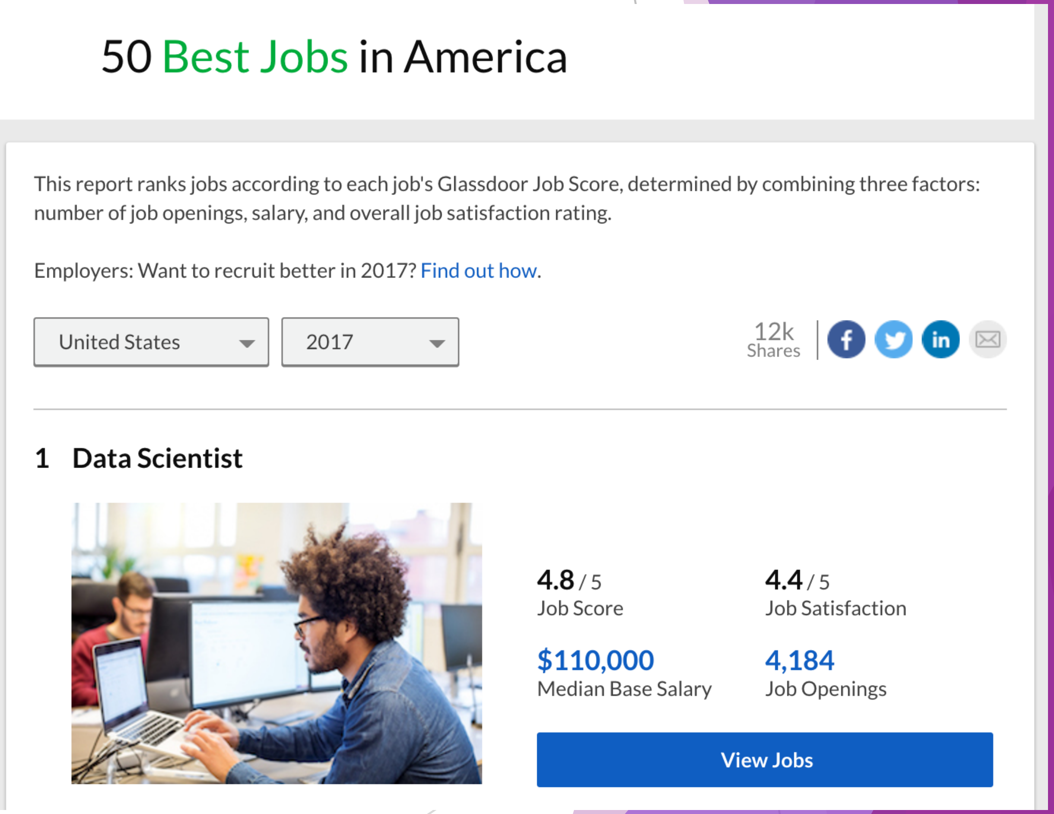
Overview - Backgrounds

- Data science

Data scientist became the best job in the US in 2016

- Data Science Education

Ubiquitous in Universities and Online Education



50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? Find out how.

United States

2017

1 Data Scientist

4.8 / 5 Job Score	4.4 / 5 Job Satisfaction
$110,000 Median Base Salary	4,184 Job Openings

[1]: 25 Best Jobs in America. https://www.glassdoor.com/List/Best-Jobs-in-America-LST KQ0,20.htm

Overview - Challenges

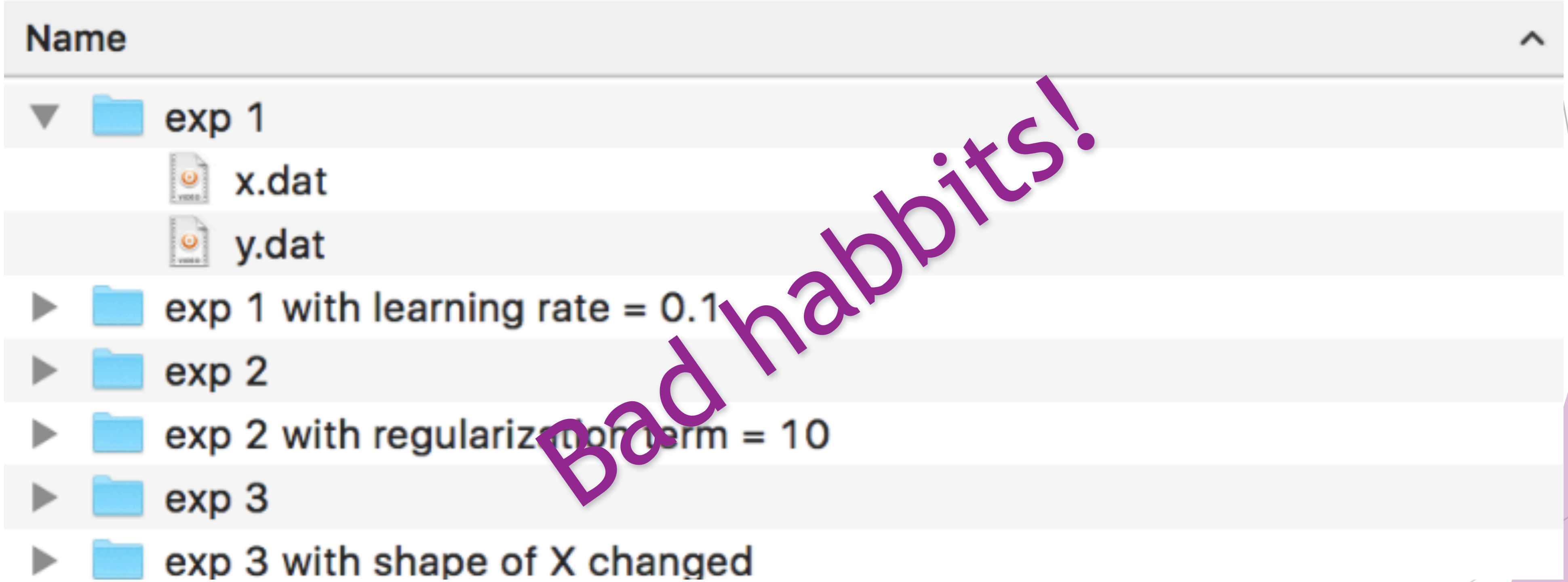
Students

- Lack formal computer science training
- Hard to set up coding tools
- Confused with data/code versions

Instructors

- Time-consuming to setup tools
- Hard to scale teaching methodologies

Differences between a DS and SE project



Name

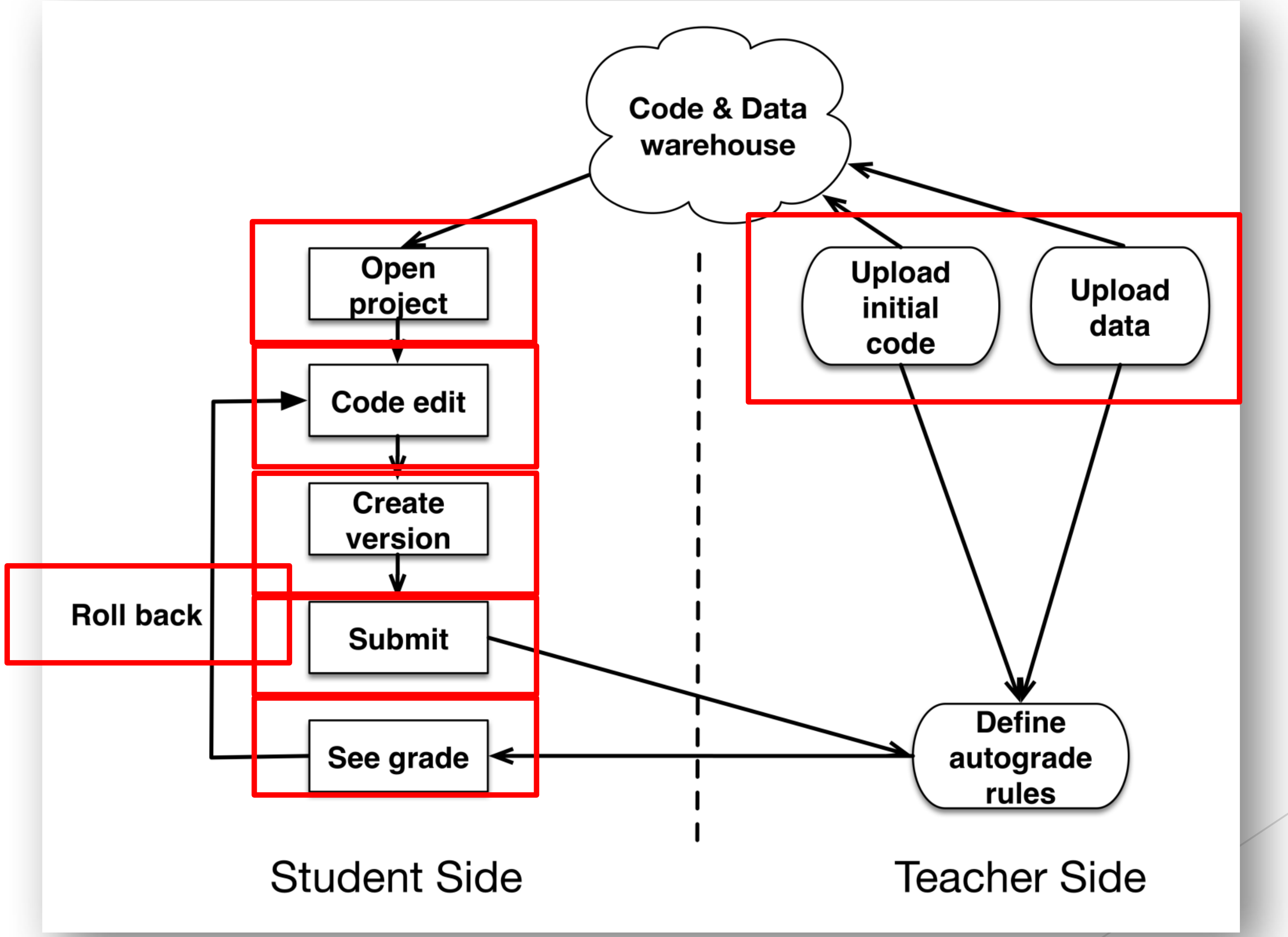
- exp 1
  - x.dat
  - y.dat
- exp 1 with learning rate = 0.1
- exp 2
- exp 2 with regularization term = 10
- exp 3
- exp 3 with shape of X changed

Bad habbits!

Our solution

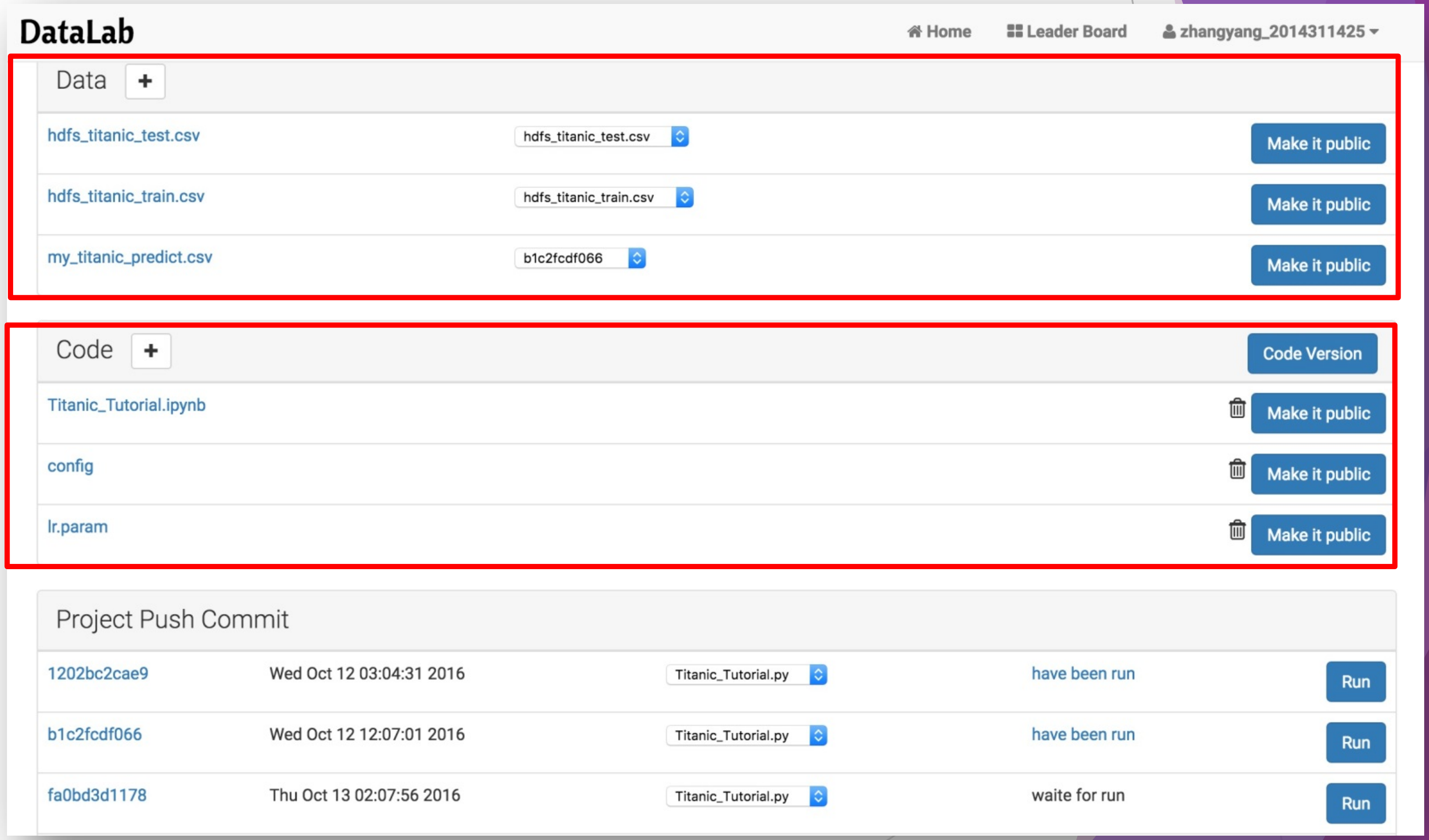
- DataLab
  - Integrates code, data and execution management into a single system
  - Creates links among code, data, parameters and their revisions
  - Provides a scalable system
  - Allows students to share their code, data, results with any versions.

Easy to set up a project



A project summary page

- Data
- Code
- Project push commit



Data

hdfs_titanic_test.csv	hdfs_titanic_test.csv	Make it public
hdfs_titanic_train.csv	hdfs_titanic_train.csv	Make it public
my_titanic_predict.csv	b1c2fcdf066	Make it public

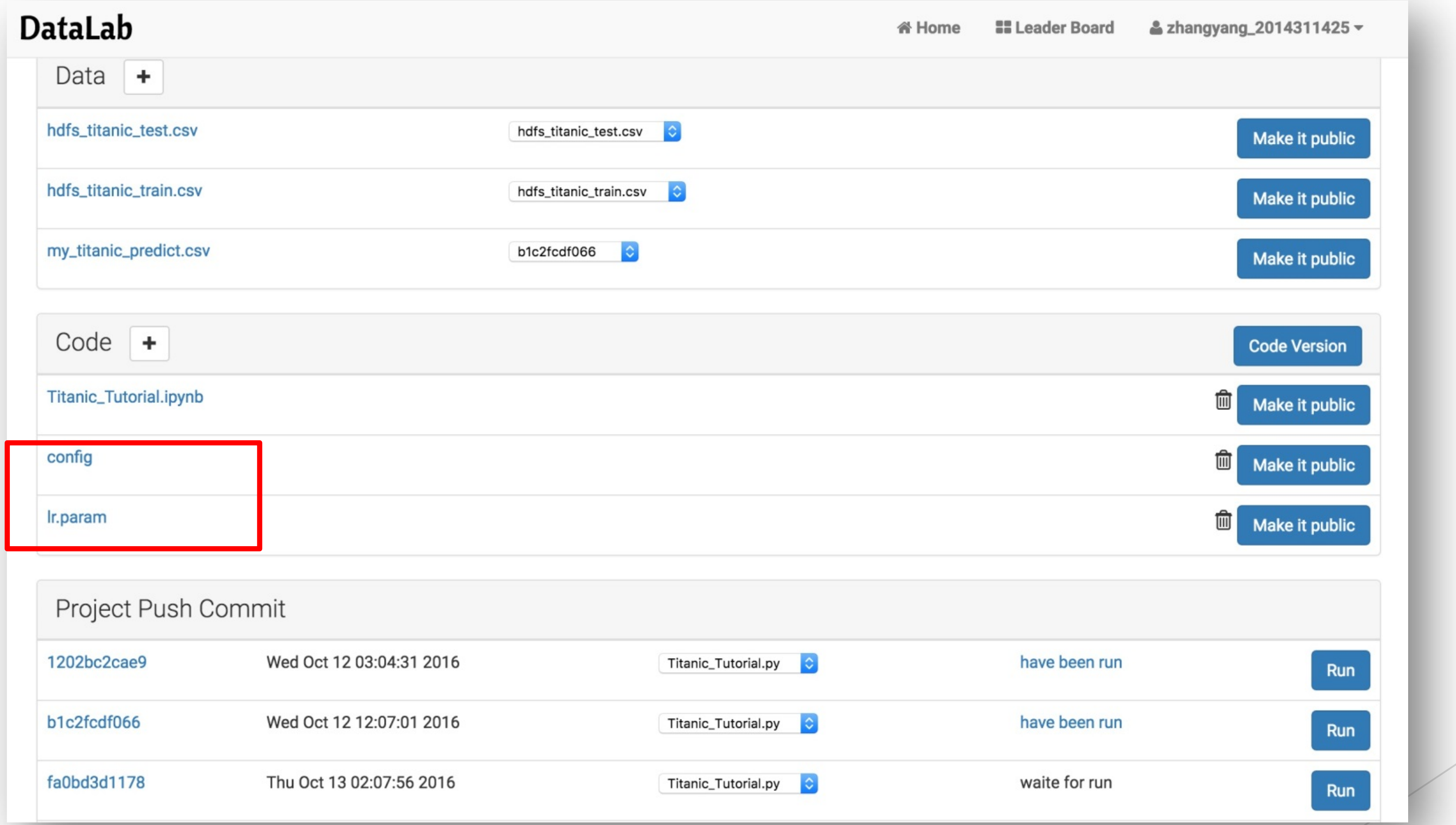
Code

Titanic_Tutorial.ipynb	Make it public
config	Make it public
lr.param	Make it public

Project Push Commit

1202bc2cae9	Wed Oct 12 03:04:31 2016	Titanic_Tutorial.py	have been run	Run
b1c2fcdf066	Wed Oct 12 12:07:01 2016	Titanic_Tutorial.py	have been run	Run
fa0bd3d1178	Thu Oct 13 02:07:56 2016	Titanic_Tutorial.py	waite for run	Run

Separate config and parameters from code



Data

hdfs_titanic_test.csv	hdfs_titanic_test.csv	Make it public
hdfs_titanic_train.csv	hdfs_titanic_train.csv	Make it public
my_titanic_predict.csv	b1c2fcdf066	Make it public

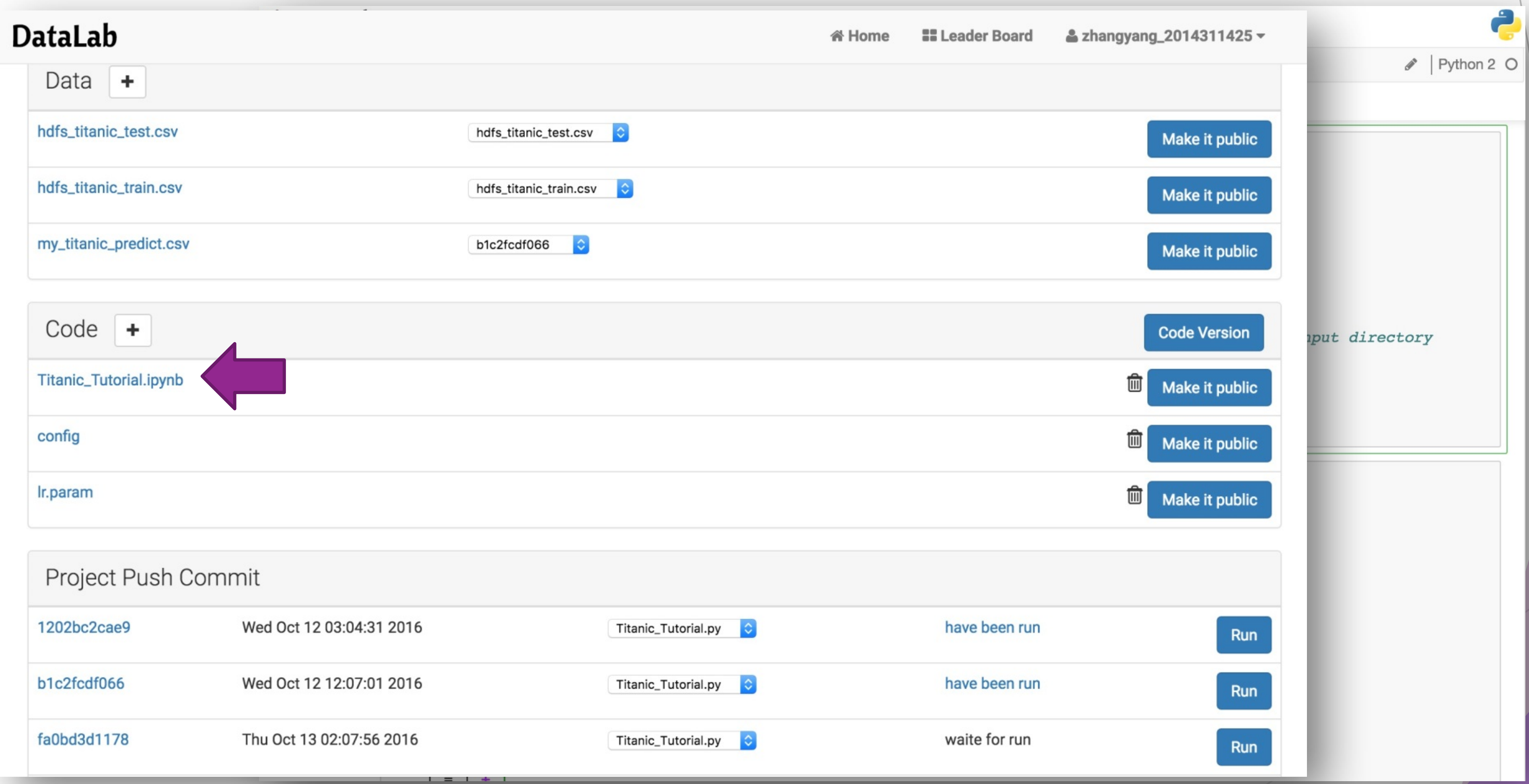
Code

Titanic_Tutorial.ipynb	Make it public
config	Make it public
lr.param	Make it public

Project Push Commit

1202bc2cae9	Wed Oct 12 03:04:31 2016	Titanic_Tutorial.py	have been run	Run
b1c2fcdf066	Wed Oct 12 12:07:01 2016	Titanic_Tutorial.py	have been run	Run
fa0bd3d1178	Thu Oct 13 02:07:56 2016	Titanic_Tutorial.py	waite for run	Run

Online development environment



Data +

hdfs_titanic_test.csv	hdfs_titanic_test.csv	Make it public
hdfs_titanic_train.csv	hdfs_titanic_train.csv	Make it public
my_titanic_predict.csv	b1c2fcdf066	Make it public

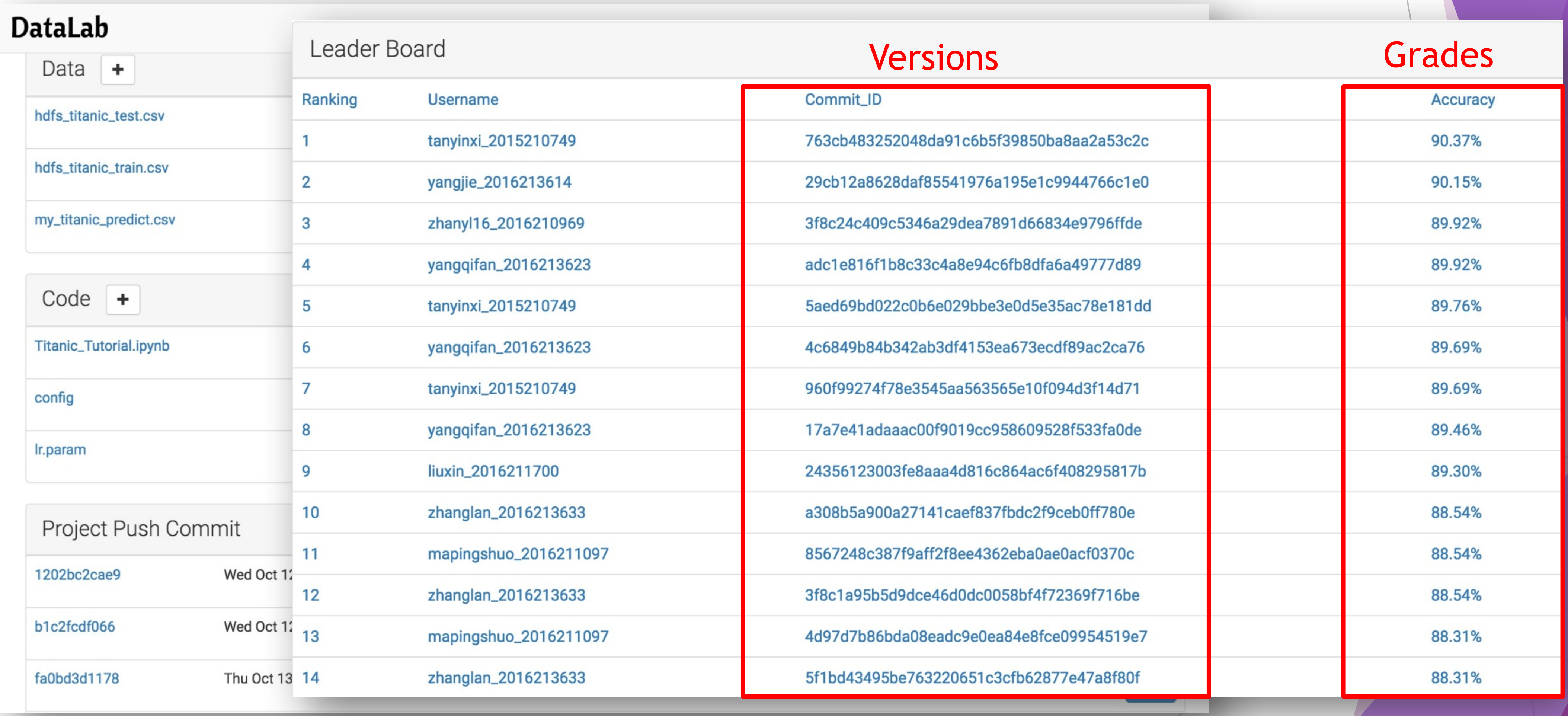
Code + Code Version

Titanic_Tutorial.ipynb	Make it public
config	Make it public
lr.param	Make it public

Project Push Commit

1202bc2cae9	Wed Oct 12 03:04:31 2016	Titanic_Tutorial.py	have been run	Run
b1c2fcdf066	Wed Oct 12 12:07:01 2016	Titanic_Tutorial.py	have been run	Run
fa0bd3d1178	Thu Oct 13 02:07:56 2016	Titanic_Tutorial.py	waite for run	Run

Creating code/data versions and autograding



DataLab

Data +

- hdfs_titanic_test.csv
- hdfs_titanic_train.csv
- my_titanic_predict.csv

Code +

- Titanic_Tutorial.ipynb
- config
- lr.param

Project Push Commit

1202bc2cae9	Wed Oct 1
b1c2fcdf066	Wed Oct 1
fa0bd3d1178	Thu Oct 13

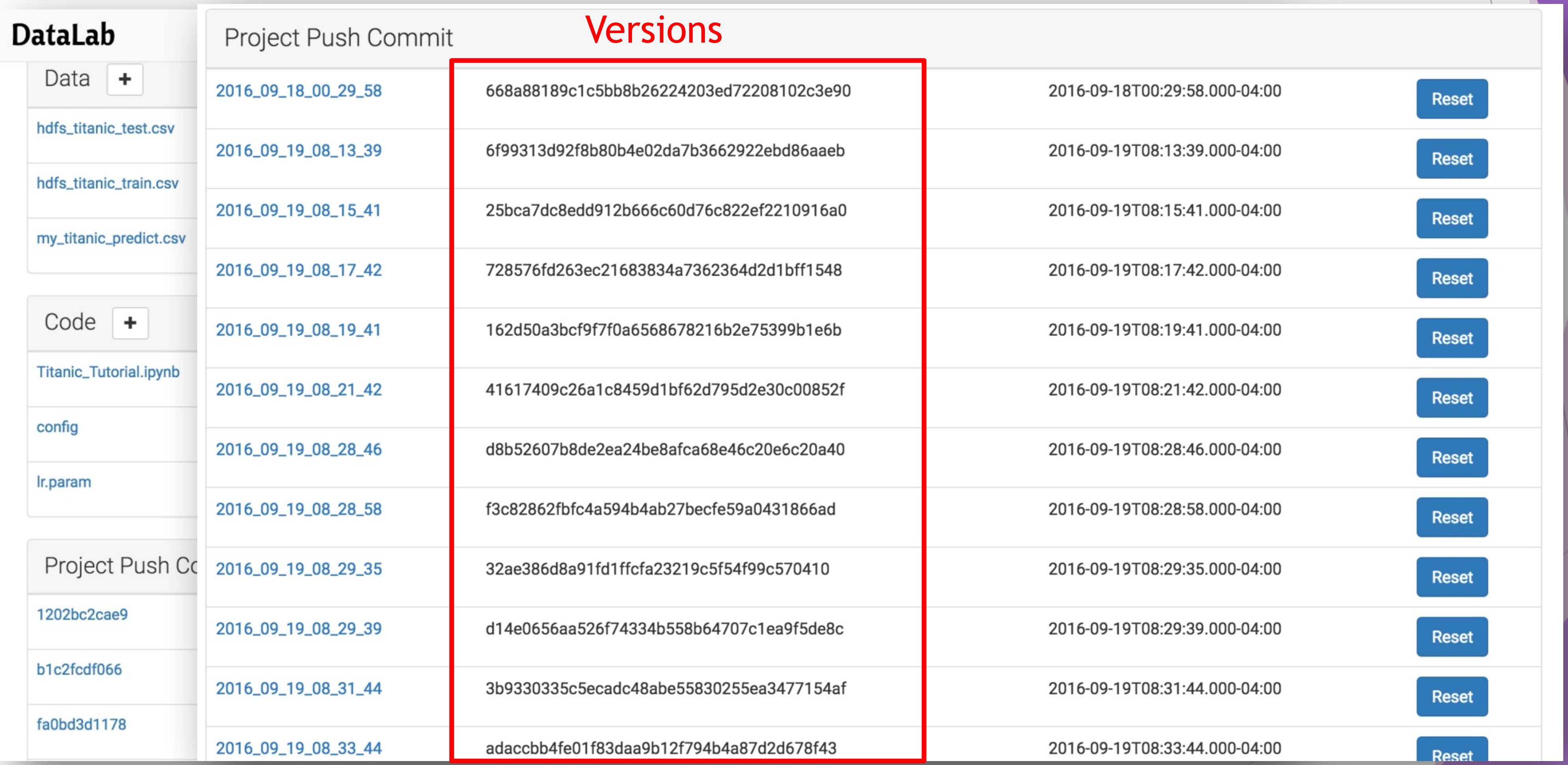
Leader Board

Versions

Grades

Ranking	Username	Commit_ID	Accuracy
1	tanyinxi_2015210749	763cb483252048da91c6b5f39850ba8aa2a53c2c	90.37%
2	yangjie_2016213614	29cb12a8628daf85541976a195e1c9944766c1e0	90.15%
3	zhanyl16_2016210969	3f8c24c409c5346a29dea7891d66834e9796ffde	89.92%
4	yangqifan_2016213623	adc1e816f1b8c33c4a8e94c6fb8dfa6a49777d89	89.92%
5	tanyinxi_2015210749	5aed69bd022c0b6e029bbe3e0d5e35ac78e181dd	89.76%
6	yangqifan_2016213623	4c6849b84b342ab3df4153ea673ecdf89ac2ca76	89.69%
7	tanyinxi_2015210749	960f99274f78e3545aa563565e10f094d3f14d71	89.69%
8	yangqifan_2016213623	17a7e41adaaac00f9019cc958609528f533fa0de	89.46%
9	liuxin_2016211700	24356123003fe8aaa4d816c864ac6f408295817b	89.30%
10	zhanglan_2016213633	a308b5a900a27141caef837fbdc2f9ceb0ff780e	88.54%
11	mapingshuo_2016211097	8567248c387f9aff2f8ee4362eba0ae0acf0370c	88.54%
12	zhanglan_2016213633	3f8c1a95b5d9dce46d0dc0058bf4f72369f716be	88.54%
13	mapingshuo_2016211097	4d97d7b86bda08eadc9e0ea84e8fce09954519e7	88.31%
14	zhanglan_2016213633	5f1bd43495be763220651c3cfb62877e47a8f80f	88.31%

Version management



Versions

DataLab

Data +

- hdfs_titanic_test.csv
- hdfs_titanic_train.csv
- my_titanic_predict.csv

Code +

- Titanic_Tutorial.ipynb
- config
- lr.param

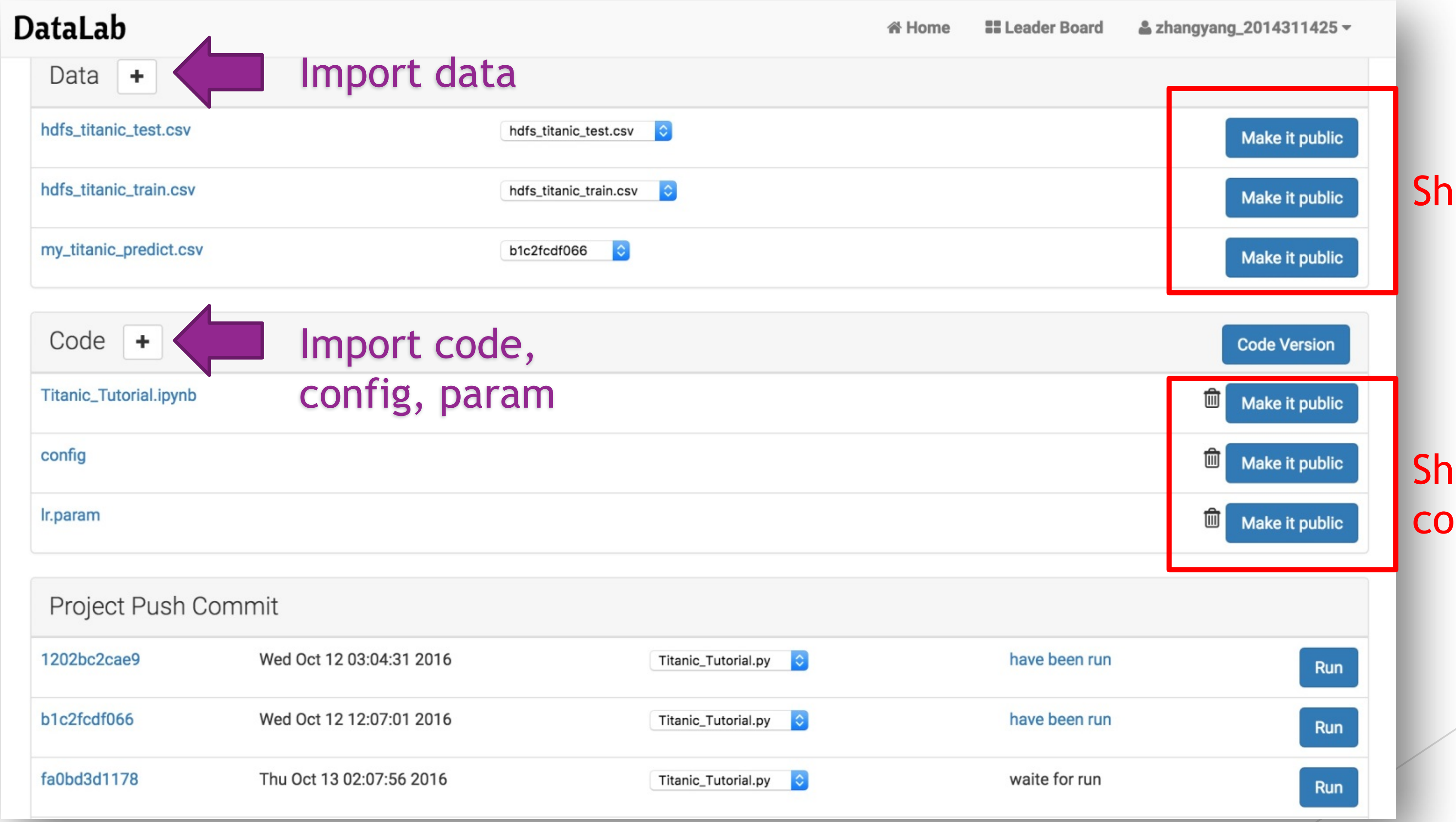
Project Push Co

- 1202bc2cae9
- b1c2fcdf066
- fa0bd3d1178

Project Push Commit

2016_09_18_00_29_58	668a88189c1c5bb8b26224203ed72208102c3e90	2016-09-18T00:29:58.000-04:00	Reset
2016_09_19_08_13_39	6f99313d92f8b80b4e02da7b3662922ebd86aaeb	2016-09-19T08:13:39.000-04:00	Reset
2016_09_19_08_15_41	25bca7dc8edd912b666c60d76c822ef2210916a0	2016-09-19T08:15:41.000-04:00	Reset
2016_09_19_08_17_42	728576fd263ec21683834a7362364d2d1bff1548	2016-09-19T08:17:42.000-04:00	Reset
2016_09_19_08_19_41	162d50a3bcf9f7f0a6568678216b2e75399b1e6b	2016-09-19T08:19:41.000-04:00	Reset
2016_09_19_08_21_42	41617409c26a1c8459d1bf62d795d2e30c00852f	2016-09-19T08:21:42.000-04:00	Reset
2016_09_19_08_28_46	d8b52607b8de2ea24be8afca68e46c20e6c20a40	2016-09-19T08:28:46.000-04:00	Reset
2016_09_19_08_28_58	f3c82862fbfc4a594b4ab27becfe59a0431866ad	2016-09-19T08:28:58.000-04:00	Reset
2016_09_19_08_29_35	32ae386d8a91fd1ffcfa23219c5f54f99c570410	2016-09-19T08:29:35.000-04:00	Reset
2016_09_19_08_29_39	d14e0656aa526f74334b558b64707c1ea9f5de8c	2016-09-19T08:29:39.000-04:00	Reset
2016_09_19_08_31_44	3b9330335c5ecadc48abe55830255ea3477154af	2016-09-19T08:31:44.000-04:00	Reset
2016_09_19_08_33_44	adaccbb4fe01f83daa9b12f794b4a87d2d678f43	2016-09-19T08:33:44.000-04:00	Reset

Team collaboration



Data +

hdfs_titanic_test.csv	hdfs_titanic_test.csv	Make it public
hdfs_titanic_train.csv	hdfs_titanic_train.csv	Make it public
my_titanic_predict.csv	b1c2fcdf066	Make it public

Code + | Code Version

Titanic_Tutorial.ipynb	Make it public
config	Make it public
lr.param	Make it public

Project Push Commit

1202bc2cae9	Wed Oct 12 03:04:31 2016	Titanic_Tutorial.py	have been run	Run
b1c2fcdf066	Wed Oct 12 12:07:01 2016	Titanic_Tutorial.py	have been run	Run
fa0bd3d1178	Thu Oct 13 02:07:56 2016	Titanic_Tutorial.py	waite for run	Run

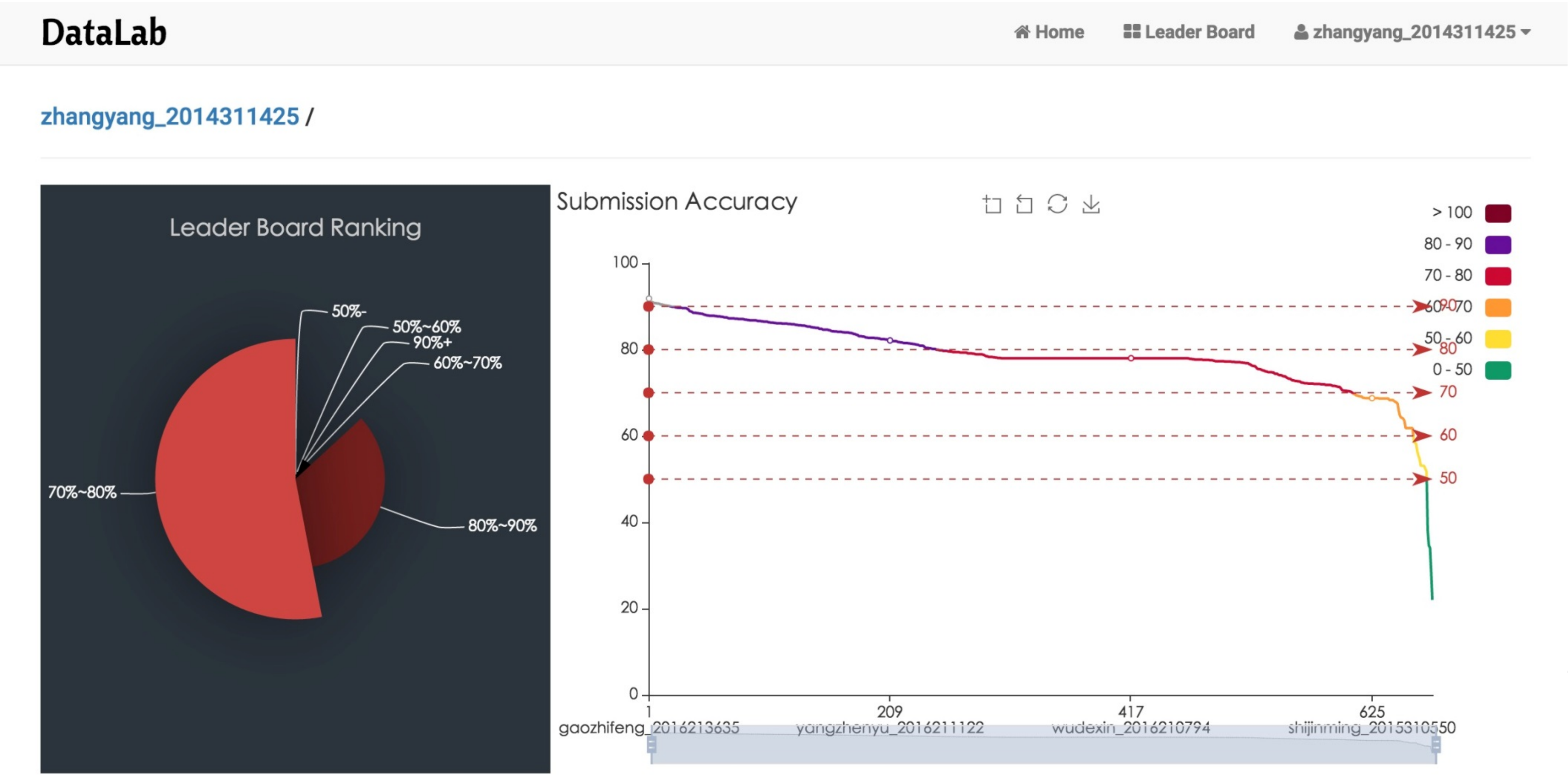
Import data

Import code, config, param

Share data

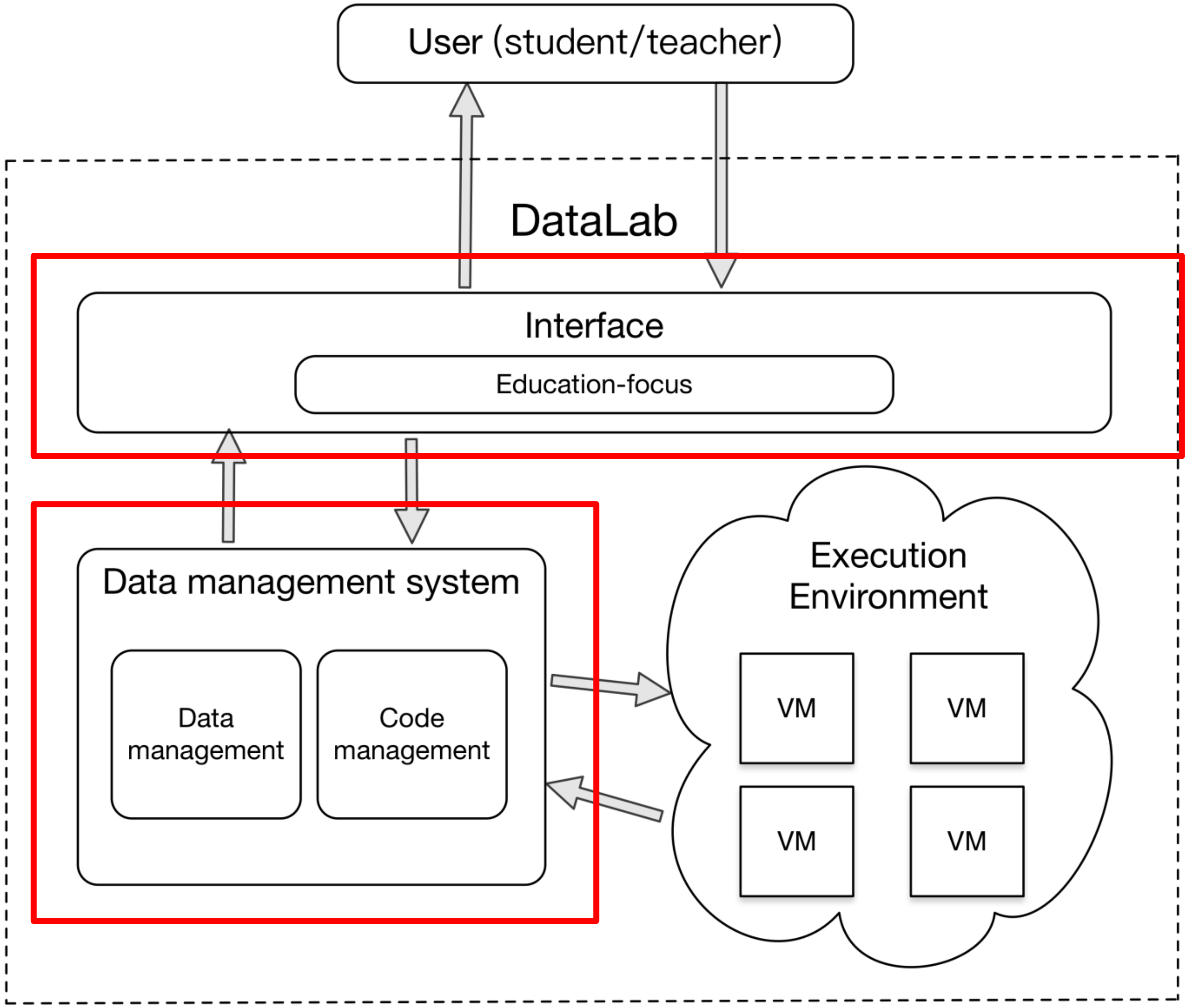
Share code, config, param

Instructor tools



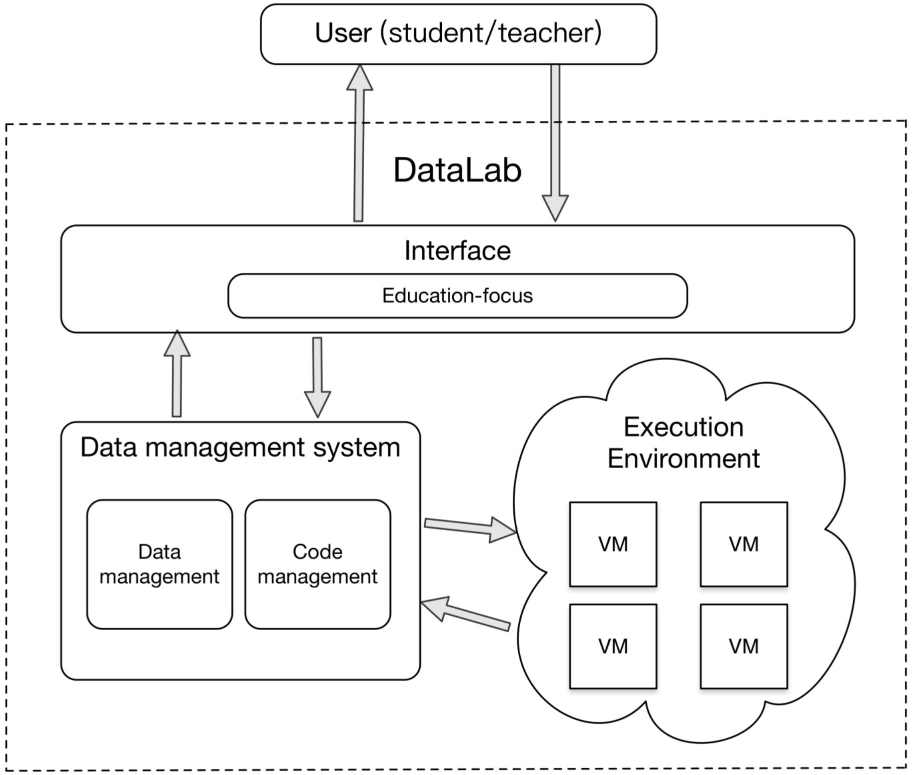
DataLab is scalable

- Data management system
- Scalable execution environment
- Extensible APIs



Evaluation - Deployment

- DataLab: 3 machines
  - 8 cores
  - 16 GB memory
  - 80GB of hard disk storage



[1]: Kaggle. https://www.kaggle.com

Evaluation : in-classroom experiment

- A graduate-level introductory data science course with **81** students and **20** volunteers
- Classical Kaggle[1] competition project: *Titanic Machine Learning from Disaster*
  - Predict survivors from gender, age, cabin class, and other information
  - **1,979** different versions of code submissions

[1]: Kaggle. https://www.kaggle.com

Log analysis

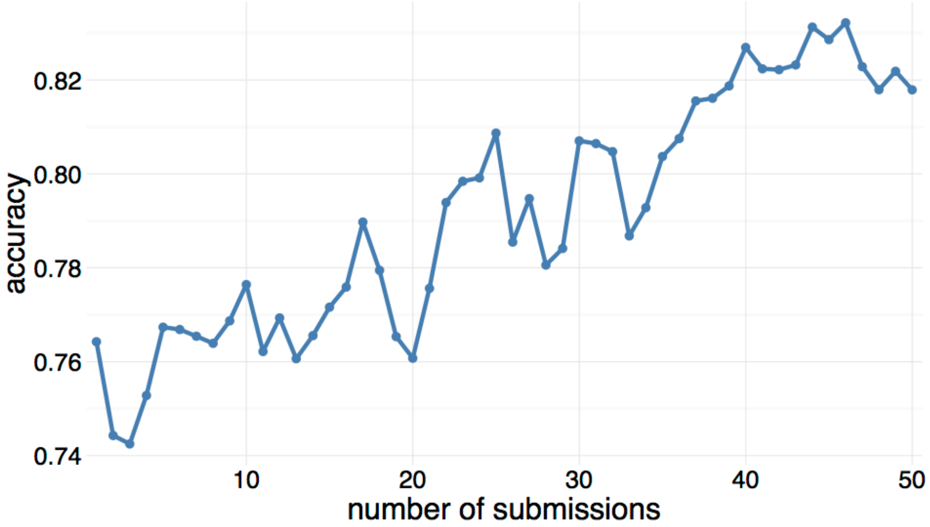


Fig 1. Relation between number of submissions and accuracy

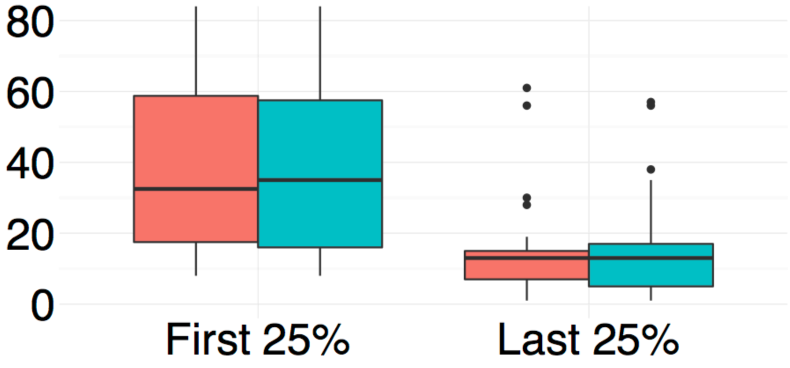


Fig 2. How many times did students push and submit their code given their ranks?

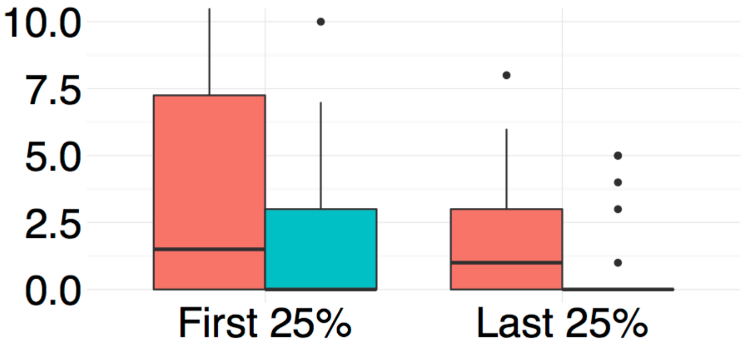


Fig 3. How many times did students check branches and reset their code given their ranks?

Survey results

- 18 subjective questions
- The survey has 3 parts
  - Students' coding experience
  - Students' opinions
  - Students suggestions

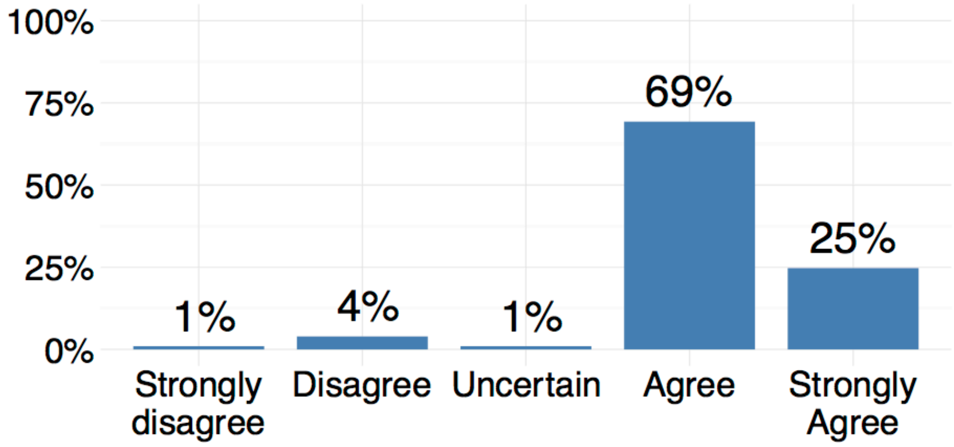


Fig 1. Is DataLab helpful for learning data analysis techniques?

- **92** out of **101** students indicate that they will continue to use DataLab for their future data science projects

Conclusion

Datalab: introducing SE Thinking to DS Education

Save instructors' time

Improve students' development efficiency

Manage data/code/execution automatically

DataLab

Can scale at low cost

Thank You!

Q&A

DataLab