CodaLab Worksheets

Percy Liang

UCI Reproducibility Symposium — September 22, 2020

The current research process

Problem 1: reproducibility

	Previous method	New method
Dataset 1	88% accuracy	92% accuracy

Problem 1: reproducibility

	Previous method	New method
Dataset 1	88% accuracy	92% accuracy
Dataset 2	72% accuracy	77% accuracy

Problem 1: reproducibility

	Previous method	New method
Dataset 1	88% accuracy	92% accuracy
Dataset 2	72% accuracy	77% accuracy
Dataset 3	?	?

Problem 1: reproducibility

	Previous method	New method
Dataset 1	88% accuracy	92% accuracy
Dataset 2	72% accuracy	77% accuracy
Dataset 3	?	?
Dataset 4	?	?
...	...	...

Problem 2: efficiency

Step 1: come up with a good idea

Problem 2: efficiency

Step 1: come up with a good idea

Step 2: execute on it

- Obtain data, clean it, convert between formats

Problem 2: efficiency

Step 1: come up with a good idea

Step 2: execute on it

- Obtain data, clean it, convert between formats
- Try to reproduce results from previous work, email authors

Problem 2: efficiency

Step 1: come up with a good idea

Step 2: execute on it

- Obtain data, clean it, convert between formats
- Try to reproduce results from previous work, email authors
- Run experiments with different versions, keep track of provenance

Problem 2: efficiency

Step 1: come up with a good idea

Step 2: execute on it

- Obtain data, clean it, convert between formats
- Try to reproduce results from previous work, email authors
- Run experiments with different versions, keep track of provenance

Tradeoff?

efficiency **reproducibility**

Folk wisdom: reproducibility slows down research.

Tradeoff?

efficiency – –

– – reproducibility

Folk wisdom: reproducibility slows down research.

Our claim: reproducibility accelerates research (with the right tool).

MLcomp.org (2008)

MLcomp VERSION ALPHA STATISTICS: 7066 USERS, 17482 DATASETS, 485 PROGRAMS, 34621 RUNS (2063 QUEUED, 0 RUNNING), 0 WORKERS

Home Programs Datasets Help About Us

MLcomp is a free website for **objectively comparing** machine learning programs across various datasets for multiple problem **domains**.

Do a comprehensive evaluation of your new algorithm.

Upload your program and run it on **existing datasets**. Compare the results with those obtained by other programs.

Find the best algorithm (program) for your dataset.

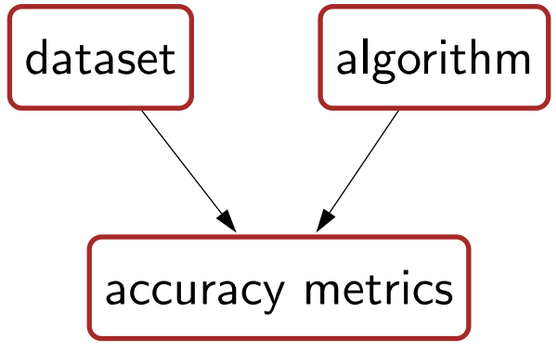
Upload your dataset and run **existing programs** on it to see which one works best.

MLcomp paradigm

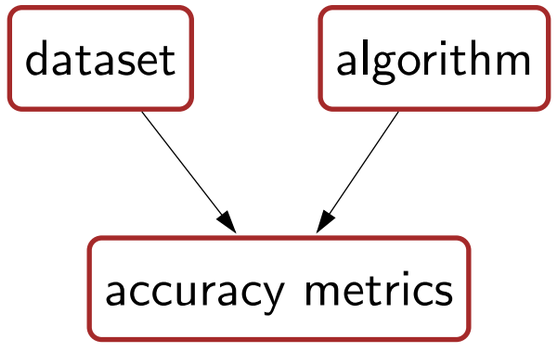
dataset

algorithm

MLcomp paradigm

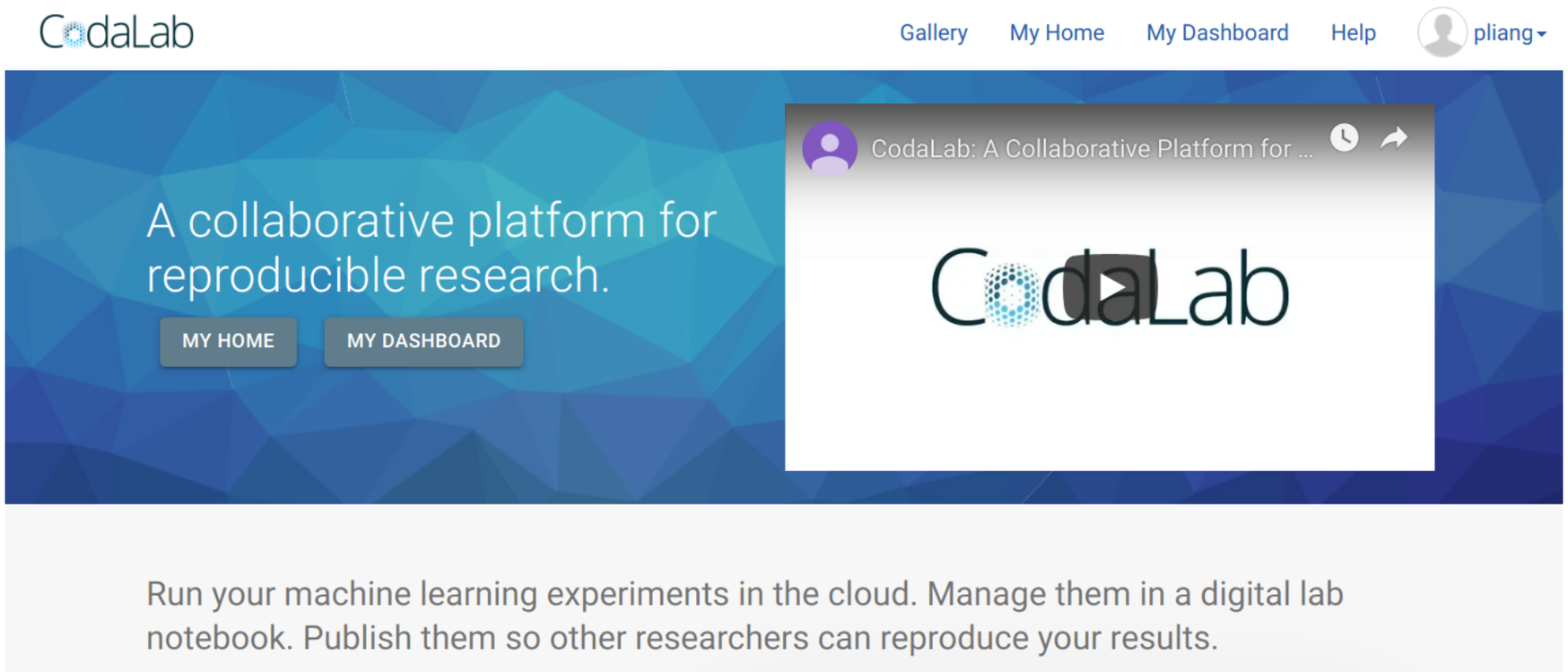


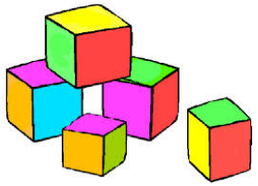
MLcomp paradigm



Problem: too rigid, doesn't help with the efficiency problem

CodaLab Worksheets (2013-present)





Bundles

Worksheets

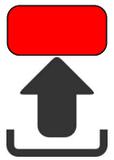
Bundles

Bundle: an **arbitrary** file/directory (code or data or results)

0x191aad8fa0ae4741b3123b15a8d59efa

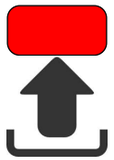
Bundles

Uploaded by user (code or data):

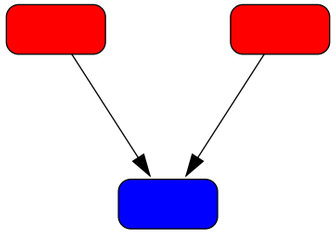


Bundles

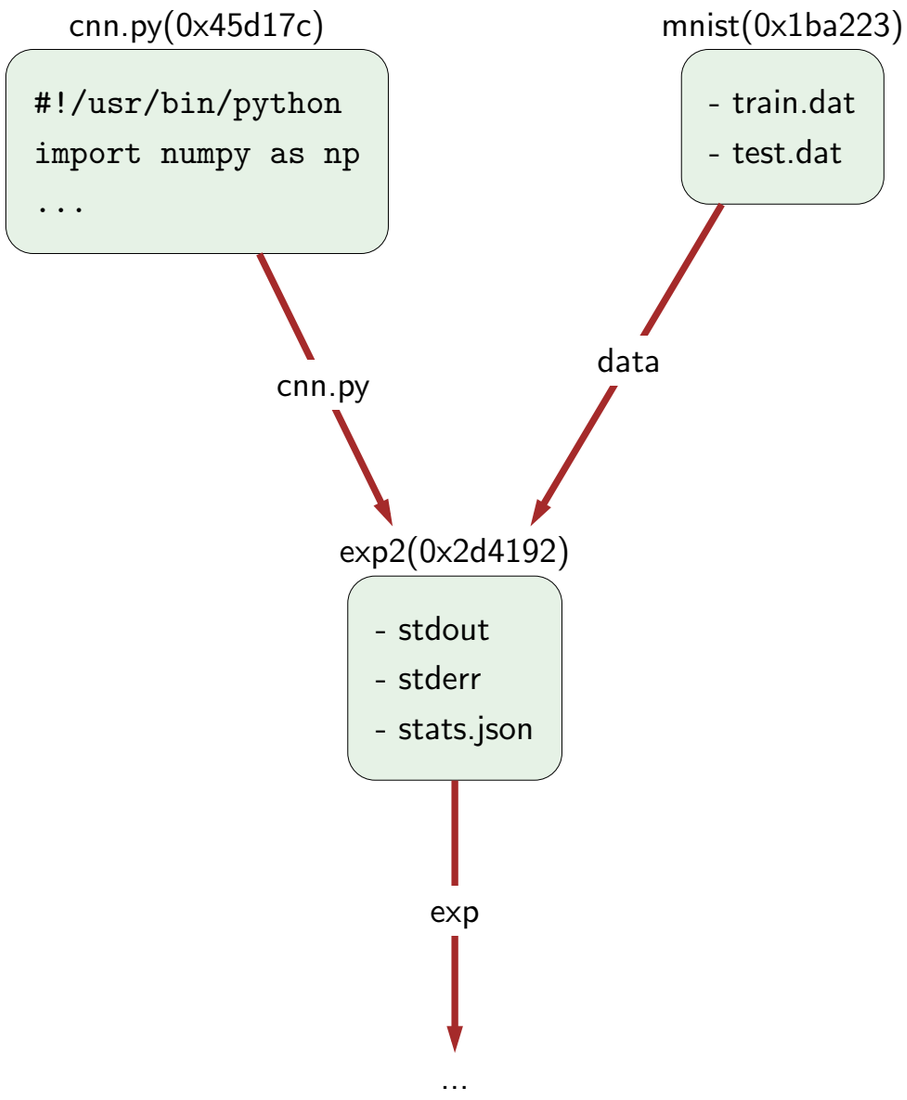
Uploaded by user (code or data):



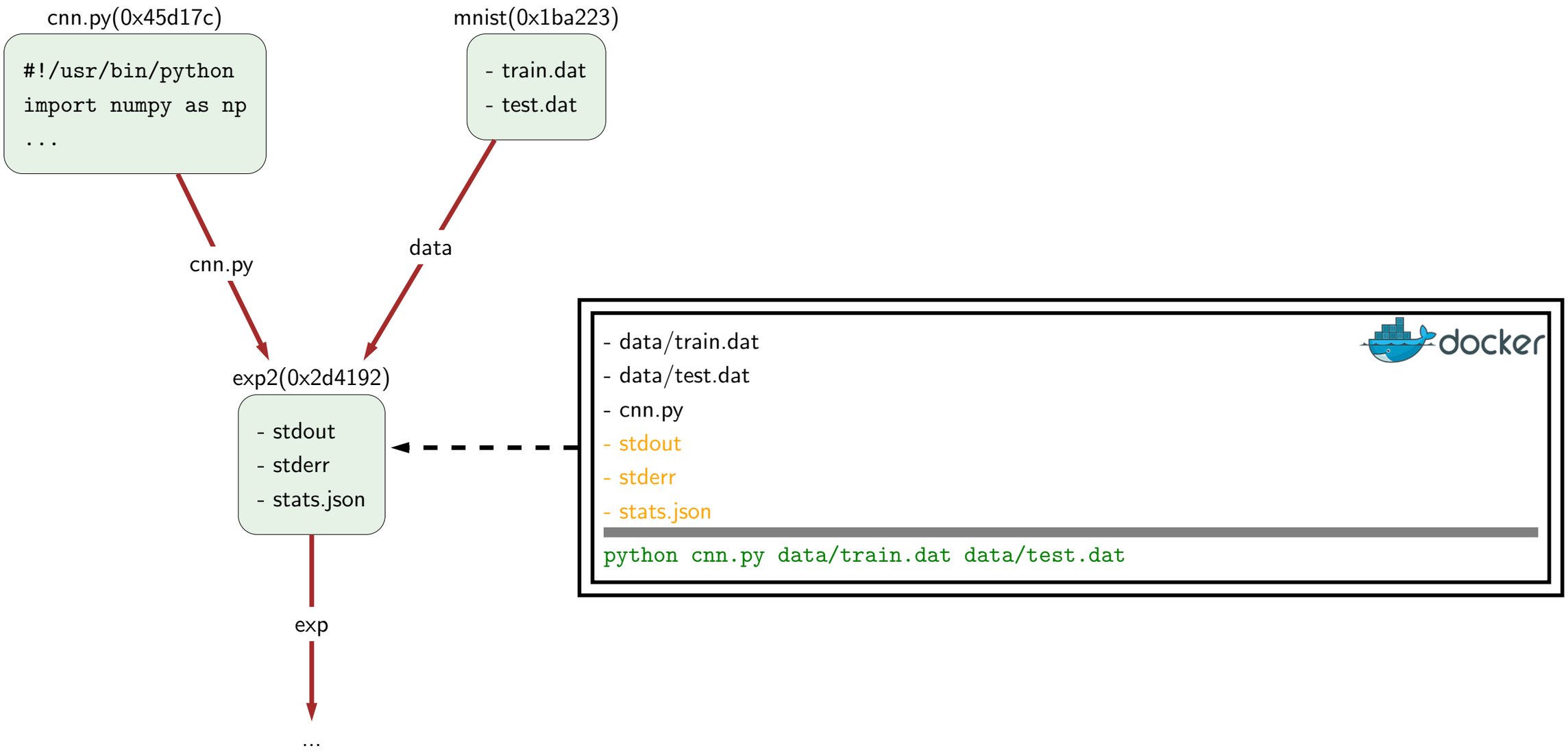
Derived by running an **arbitrary** command:



Bundles


cnn.py(0x45d17c)
#!/usr/bin/python
import numpy as np
...
mnist(0x1ba223)
- train.dat
- test.dat
cnn.py
data
exp2(0x2d4192)
- stdout
- stderr
- stats.json
exp
...

Bundles


cnn.py(0x45d17c)
#!/usr/bin/python
import numpy as np
...
mnist(0x1ba223)
- train.dat
- test.dat
cnn.py
data
exp2(0x2d4192)
- stdout
- stderr
- stats.json
exp
...
- data/train.dat
- data/test.dat
- cnn.py
- stdout
- stderr
- stats.json
python cnn.py data/train.dat data/test.dat
docker

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Run experiments with arbitrary commands:

```
$ cl run :cnn.py data:mnist "python cnn.py data/train.dat data/test.dat"
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Run experiments with arbitrary commands:

```
$ cl run :cnn.py data:mnist "python cnn.py data/train.dat data/test.dat"
```

Look at output of runs:

```
$ cl cat exp2/stdout
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Run experiments with arbitrary commands:

```
$ cl run :cnn.py data:mnist "python cnn.py data/train.dat data/test.dat"
```

Look at output of runs:

```
$ cl cat exp2/stdout
```

Manage runs:

```
$ cl kill exp2; cl rm exp2
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Run experiments with arbitrary commands:

```
$ cl run :cnn.py data:mnist "python cnn.py data/train.dat data/test.dat"
```

Look at output of runs:

```
$ cl cat exp2/stdout
```

Manage runs:

```
$ cl kill exp2; cl rm exp2
```

Run an entire pipeline with a different dataset or newer version of your code:

```
$ cl mimic mnist exp2 cifar -n exp3
```

Command-line Interface (CLI)

Search for existing code and data:

```
$ cl search mnist
```

Upload new code or data:

```
$ cl upload cnn.py
```

Run experiments with arbitrary commands:

```
$ cl run :cnn.py data:mnist "python cnn.py data/train.dat data/test.dat"
```

Look at output of runs:

```
$ cl cat exp2/stdout
```

Manage runs:

```
$ cl kill exp2; cl rm exp2
```

Run an entire pipeline with a different dataset or newer version of your code:

```
$ cl mimic mnist exp2 cifar -n exp3
```

Copy from one CodaLab instance to another:

```
$ cl add bundle mnist stanford::pliang-demo main::pliang-demo
```

Modularity

Real-world problems require efforts of entire community

Modularity

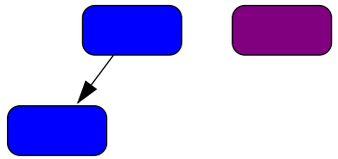
Real-world problems require efforts of entire community

People specialize, contribute in decentralized way

Modularity

Real-world problems require efforts of entire community

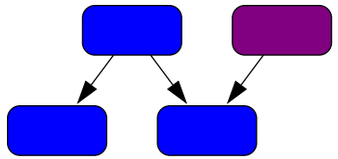
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

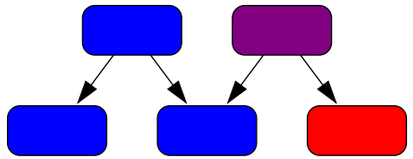
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

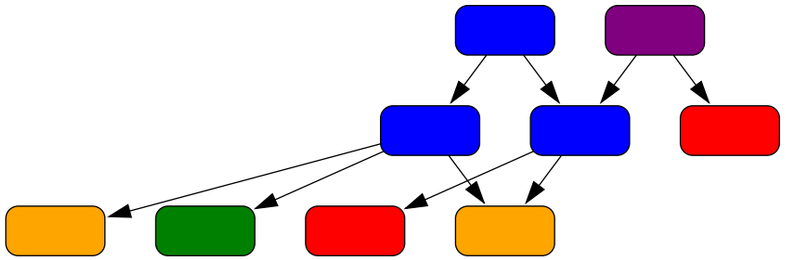
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

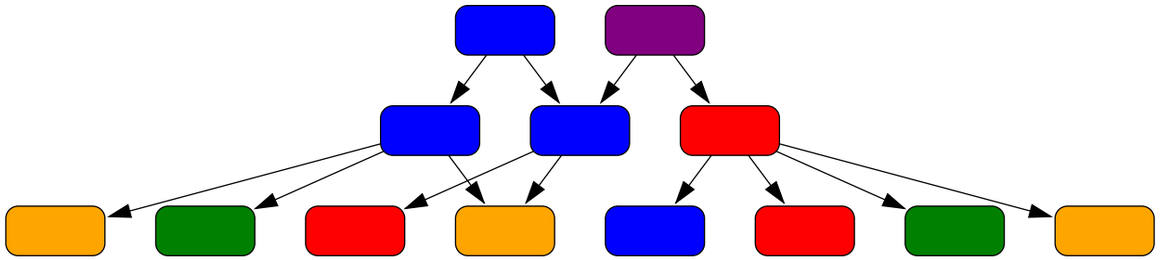
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

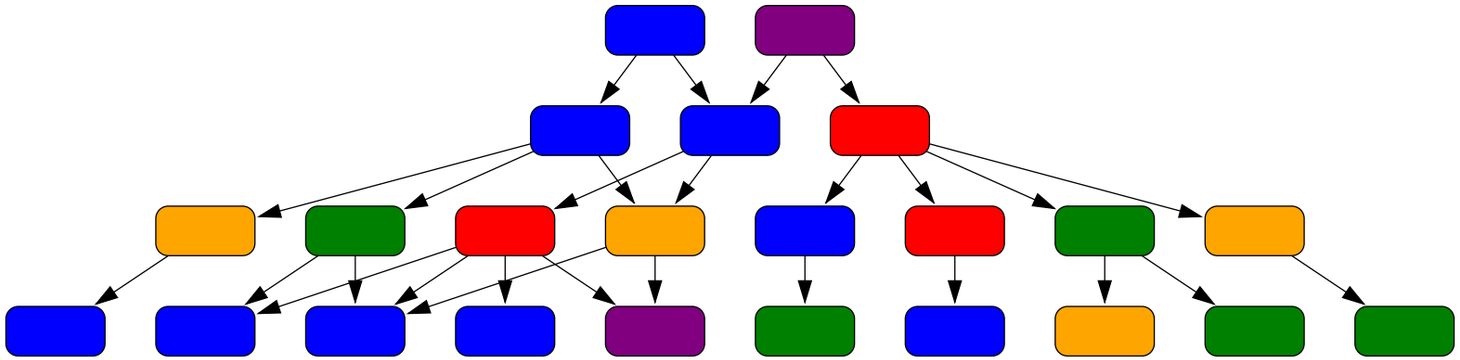
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

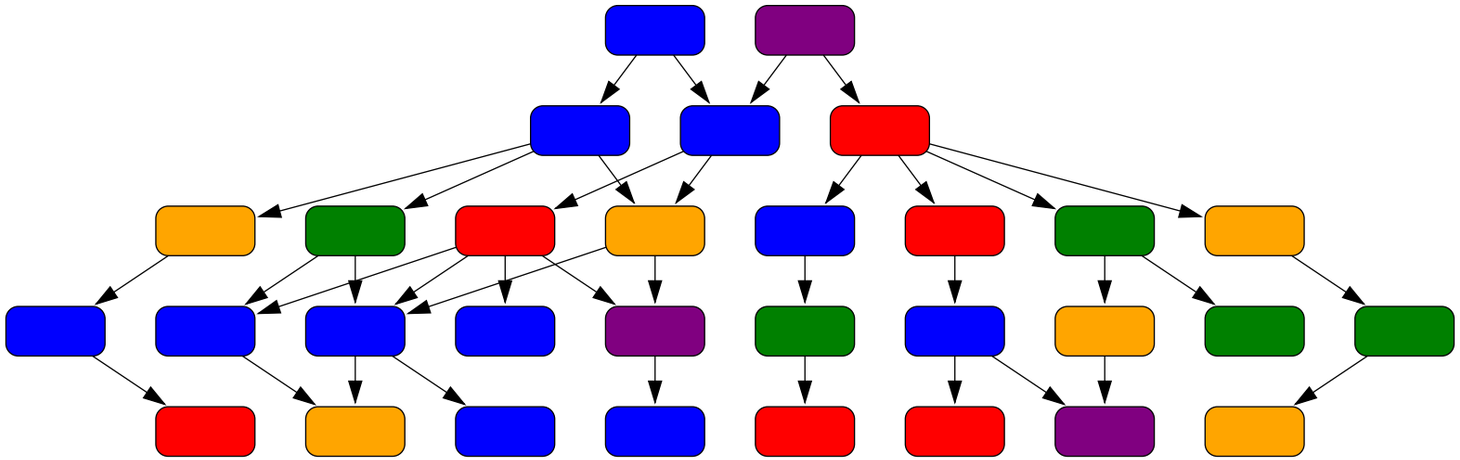
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

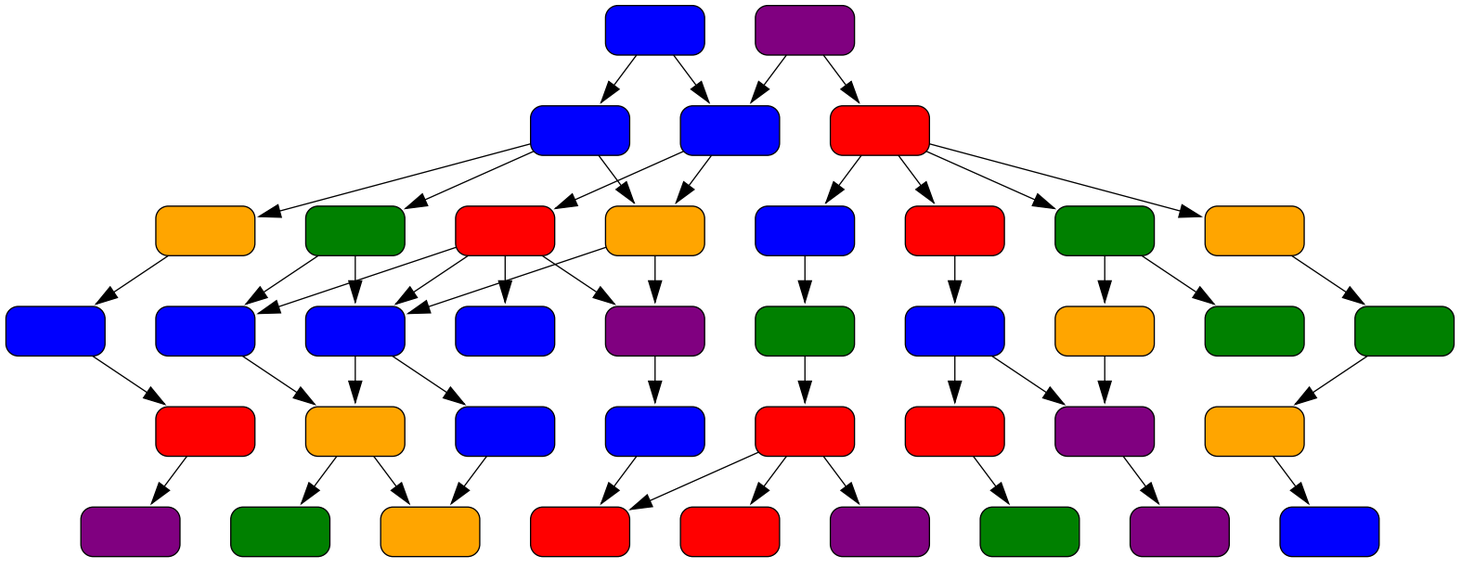
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

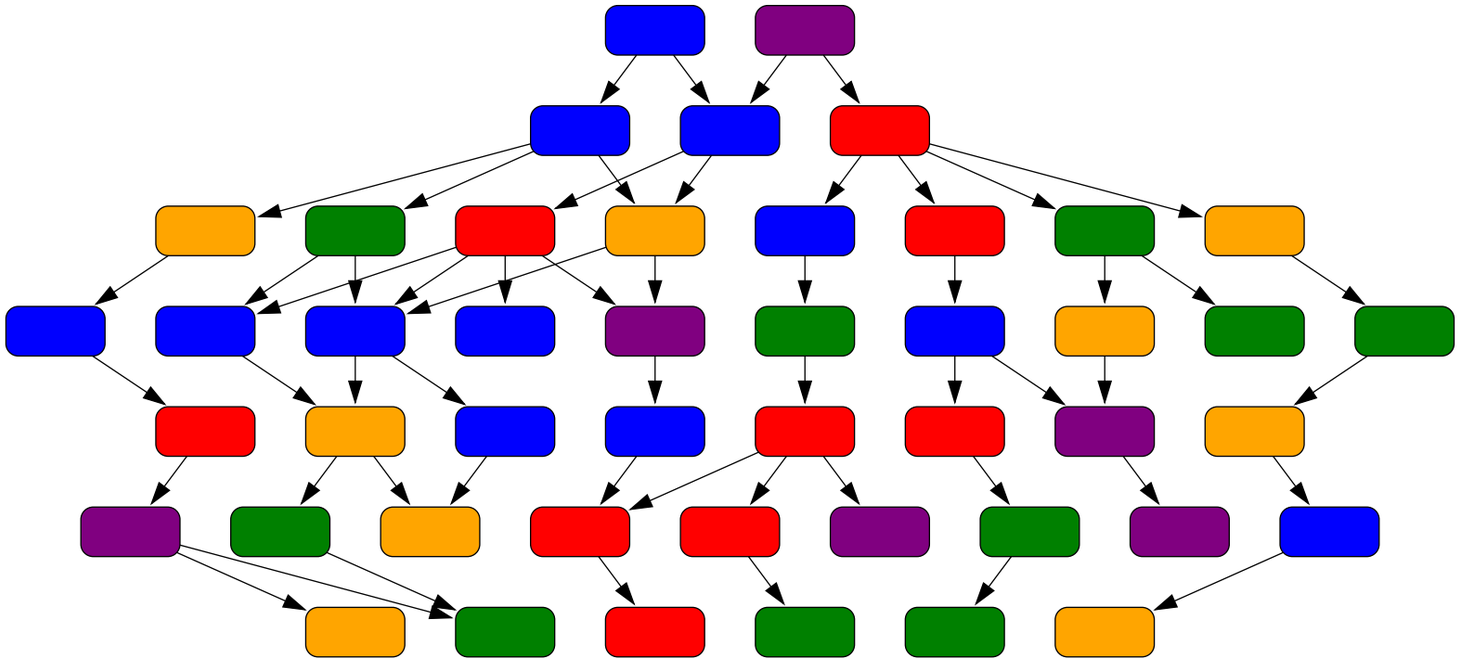
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

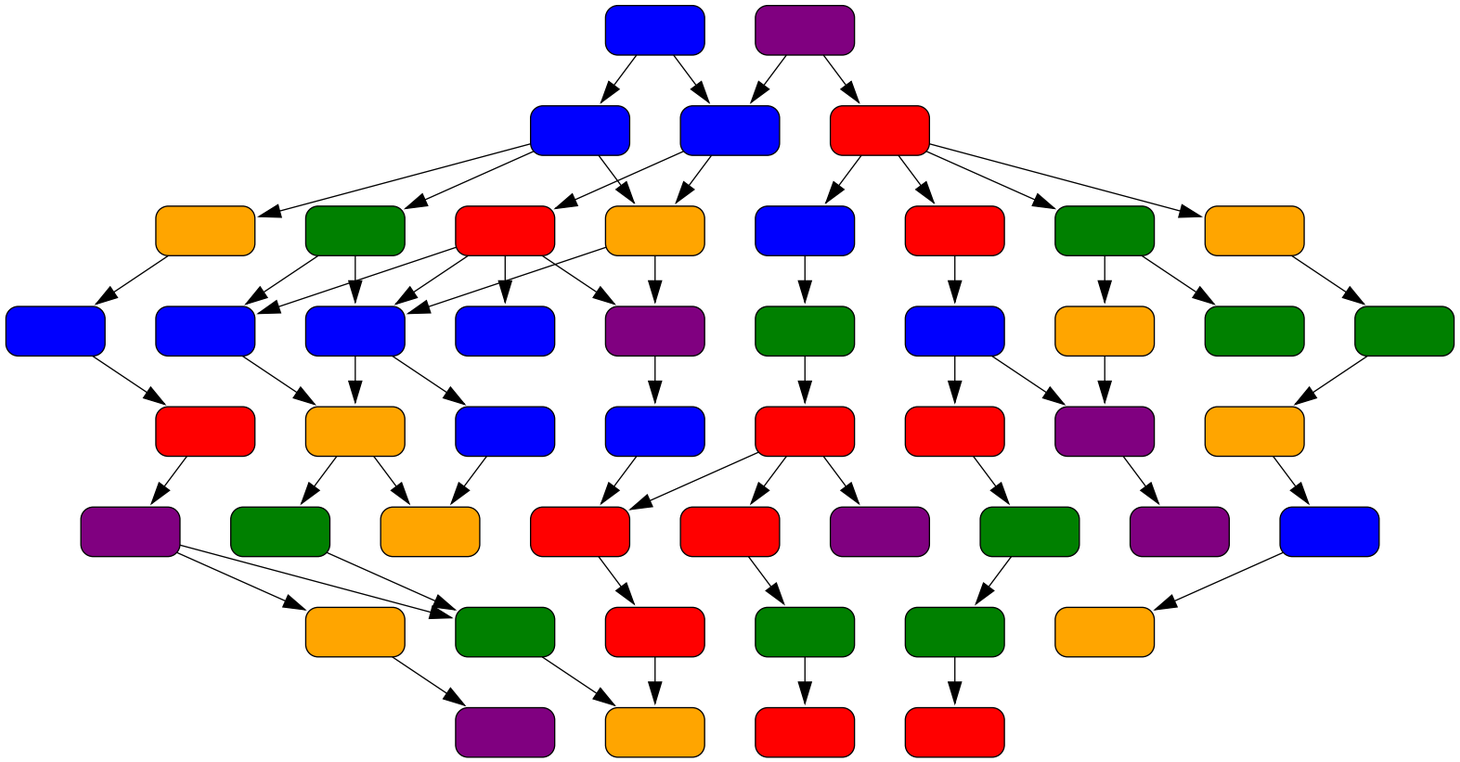
People specialize, contribute in decentralized way



Modularity

Real-world problems require efforts of entire community

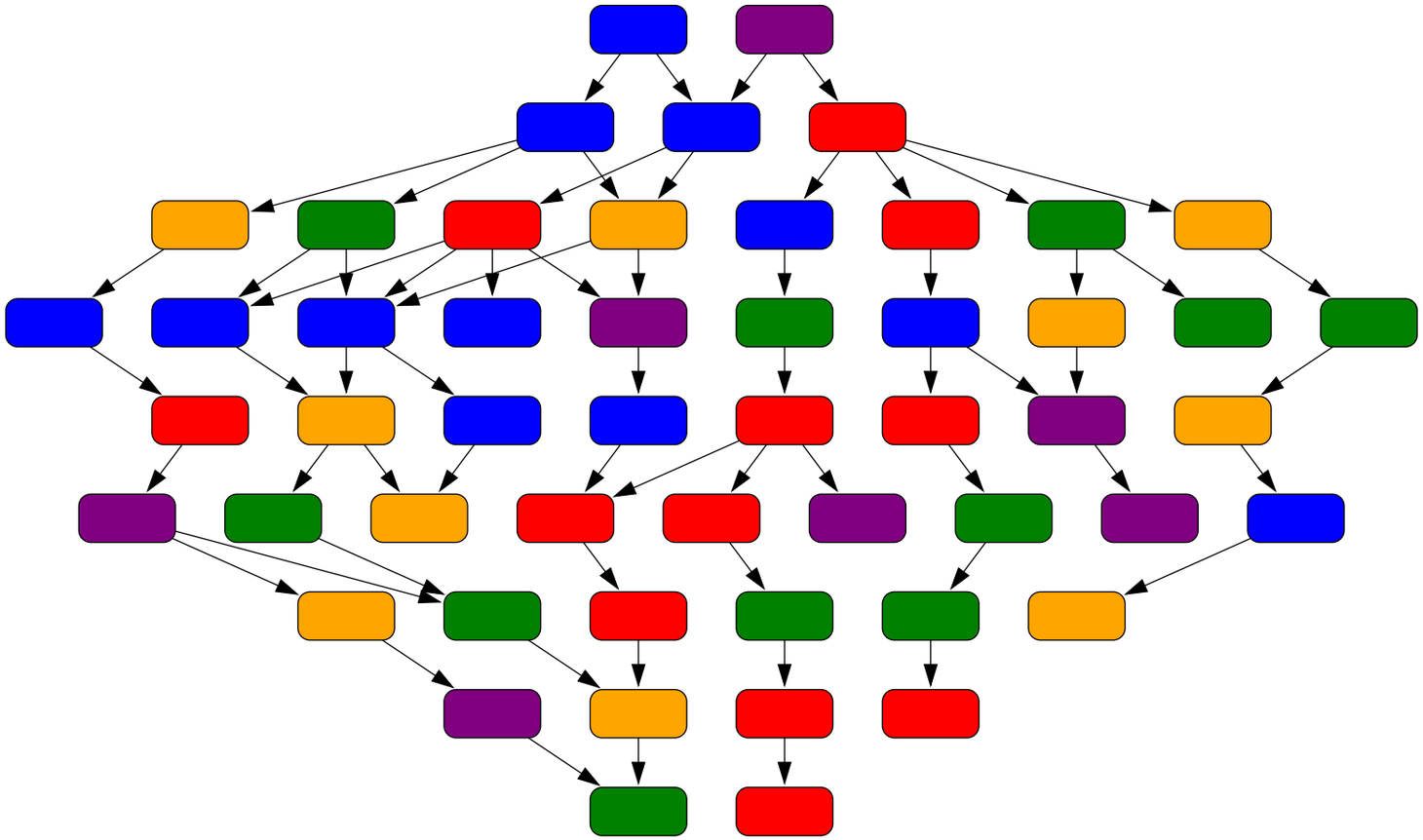
People specialize, contribute in decentralized way



Modularity

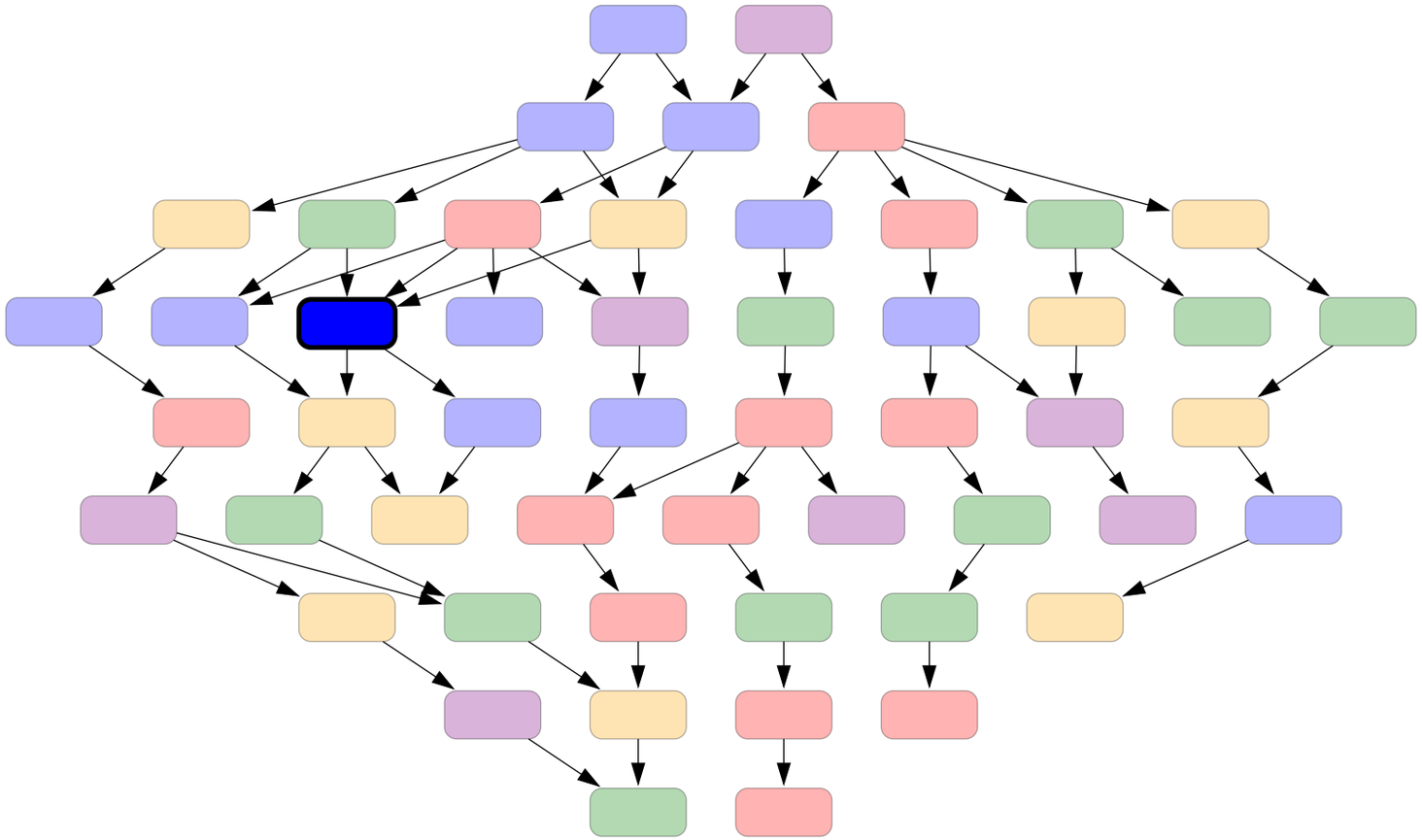
Real-world problems require efforts of entire community

People specialize, contribute in decentralized way



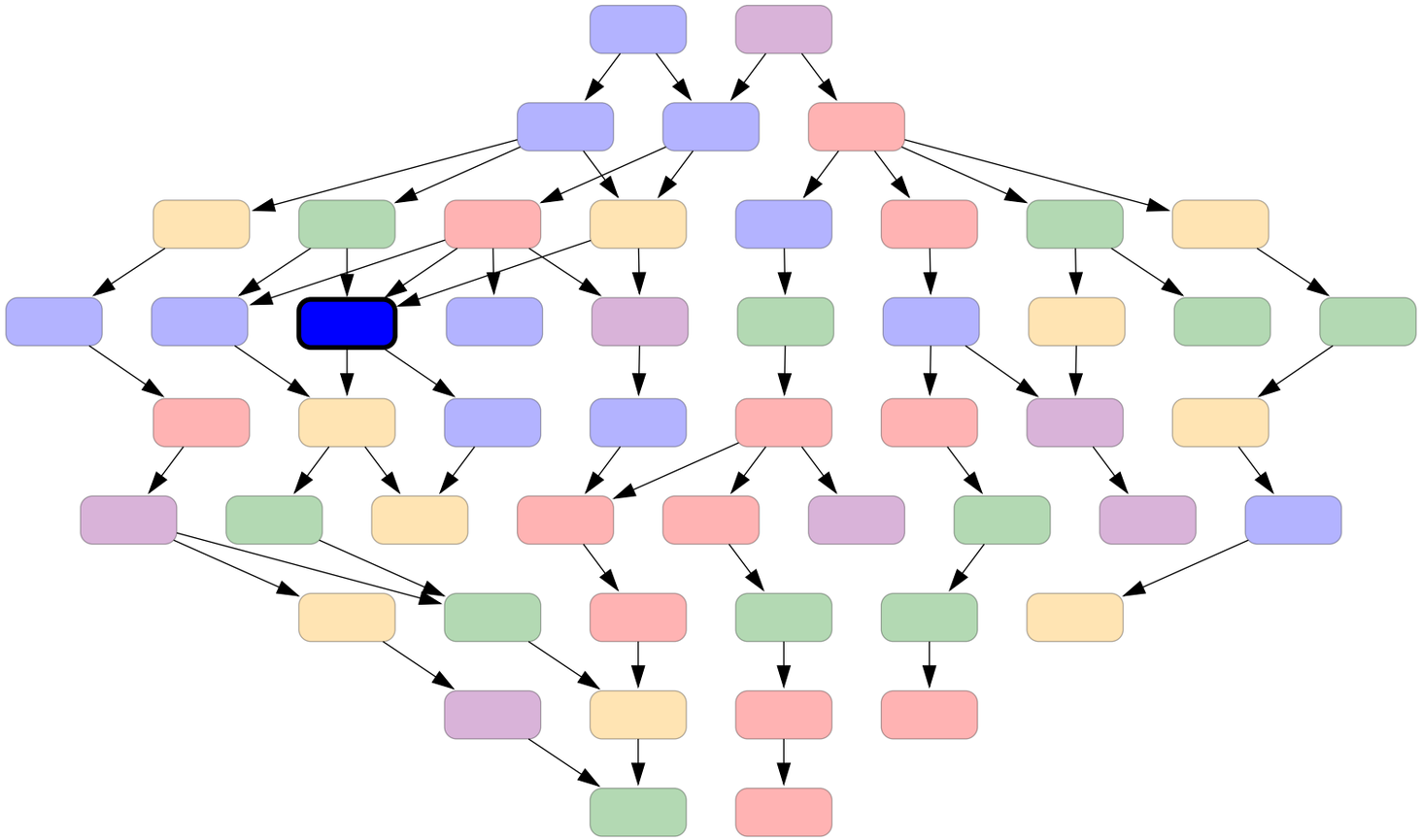
Intermediate tasks

- Old way: use intermediate metrics, rhetoric



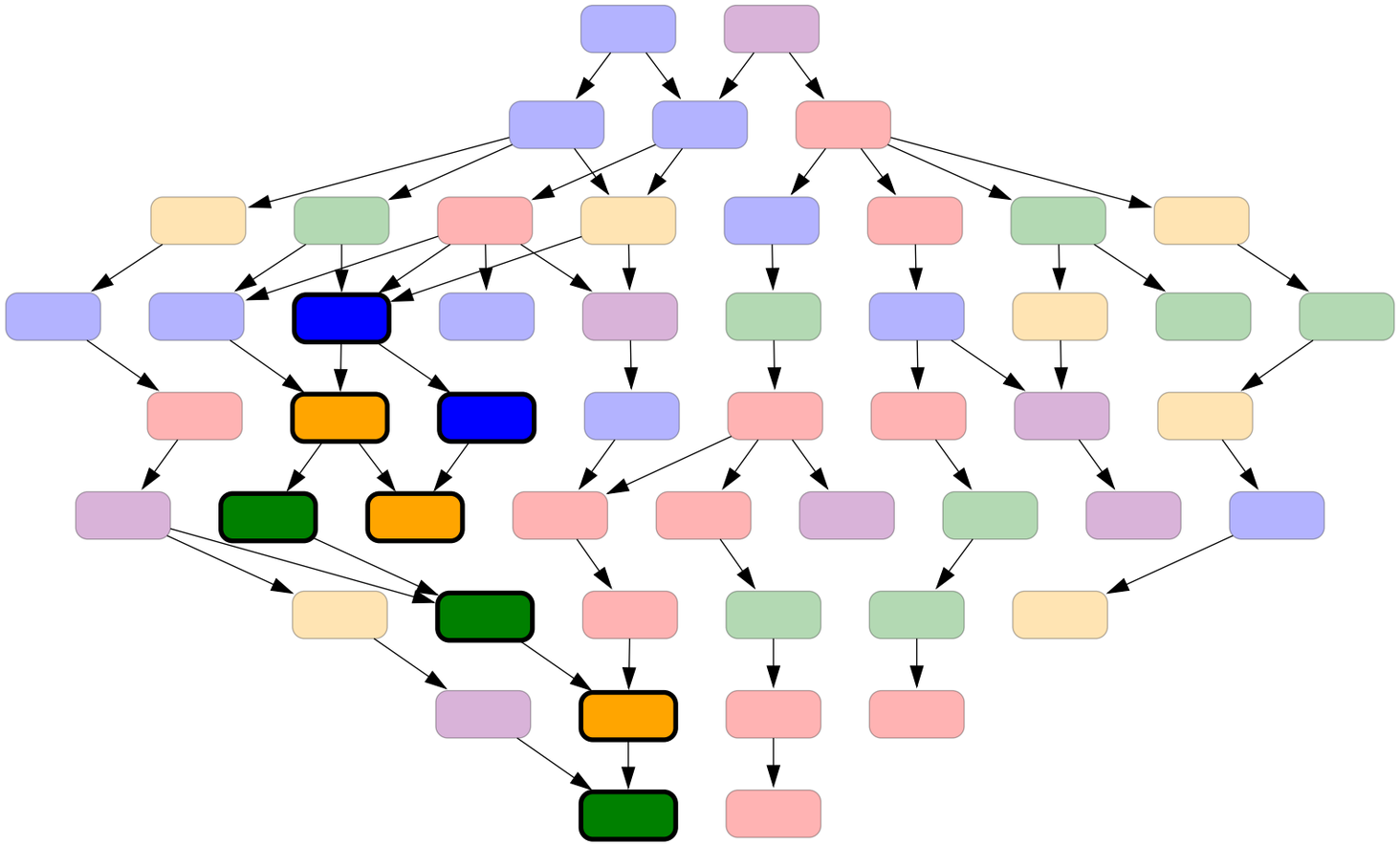
Intermediate tasks

- Old way: use intermediate metrics, rhetoric
- New way: plug in and see ramifications **automatically**



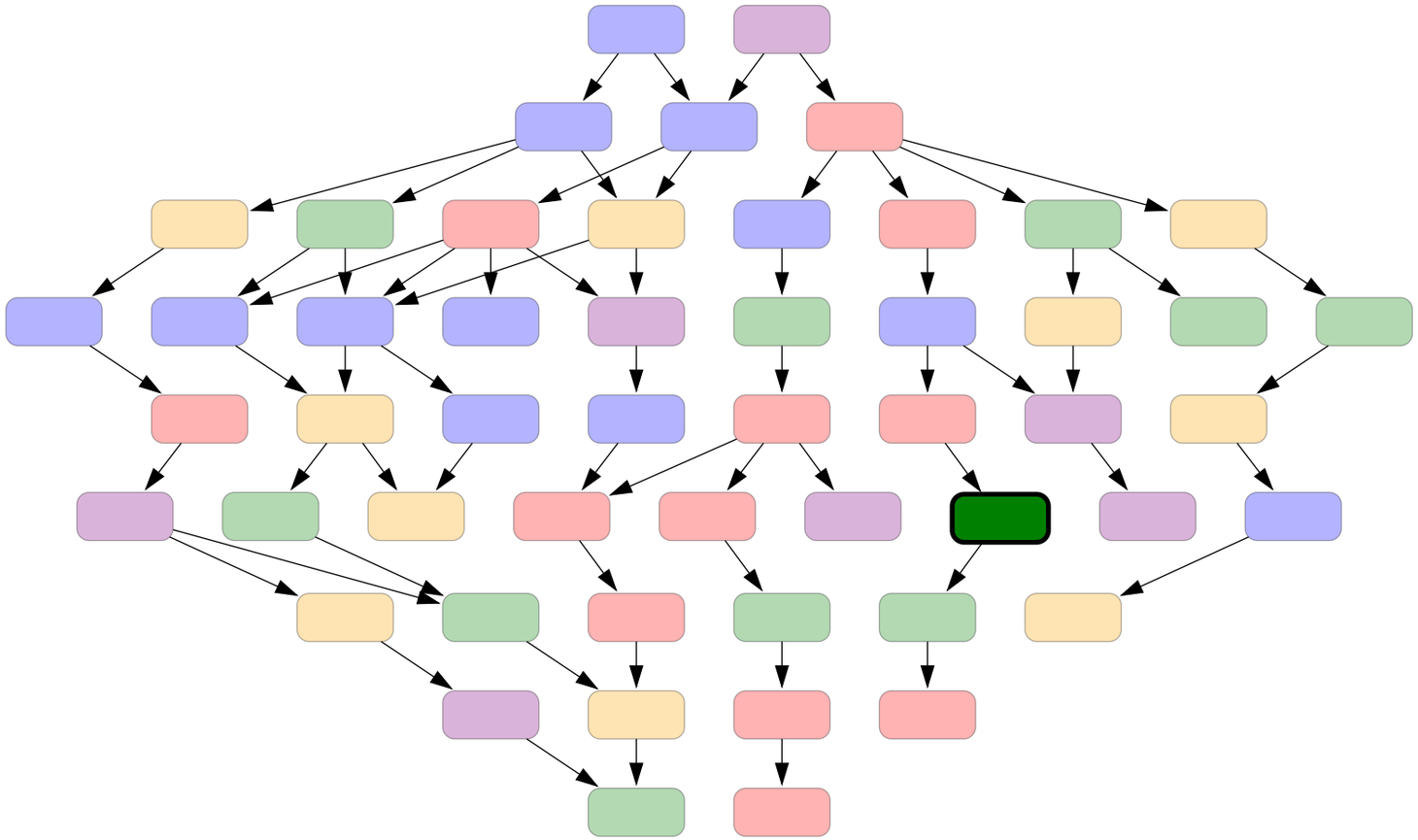
Intermediate tasks

- Old way: use intermediate metrics, rhetoric
- New way: plug in and see ramifications **automatically**



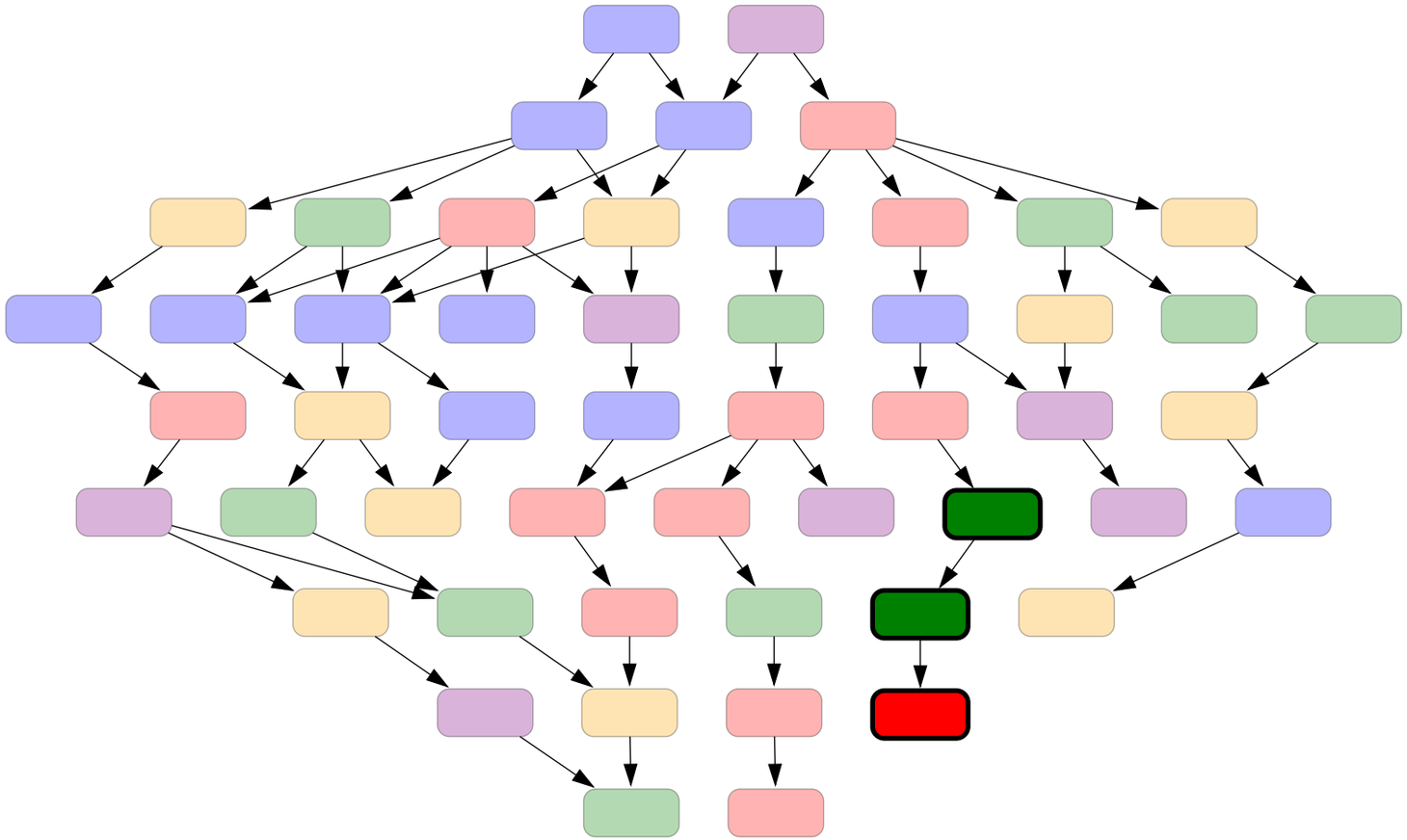
Intermediate tasks

- Old way: use intermediate metrics, rhetoric
- New way: plug in and see ramifications **automatically**

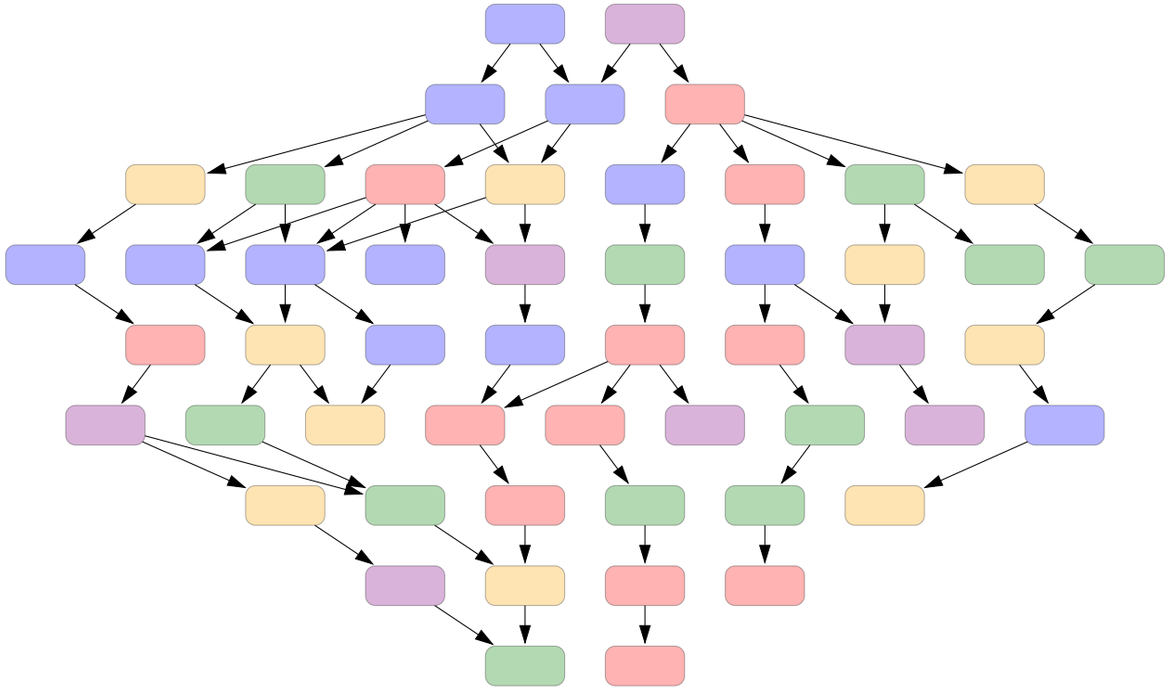


Intermediate tasks

- Old way: use intermediate metrics, rhetoric
- New way: plug in and see ramifications **automatically**

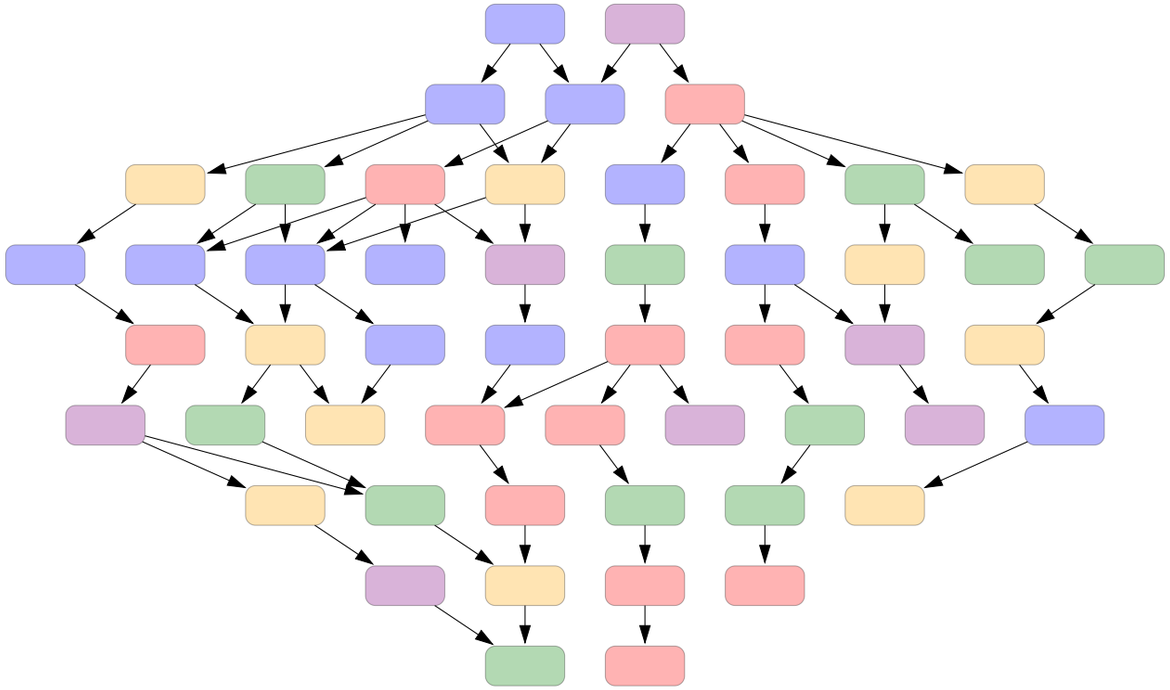


Immutability



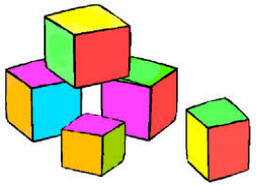
Inspiration: Git version control system

Immutability



Inspiration: Git version control system

- All programs/datasets/runs are write-once
- Enable collaboration without chaos
- Capture the research process in a **reproducible** way



Bundles

Worksheets

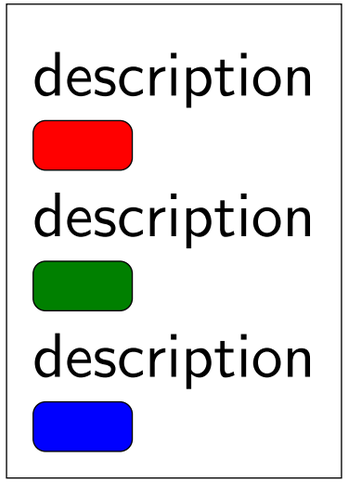
Literacy

Bundle graphs are about **truth**; what about **interpretation**?

Literacy

Bundle graphs are about **truth**; what about **interpretation**?

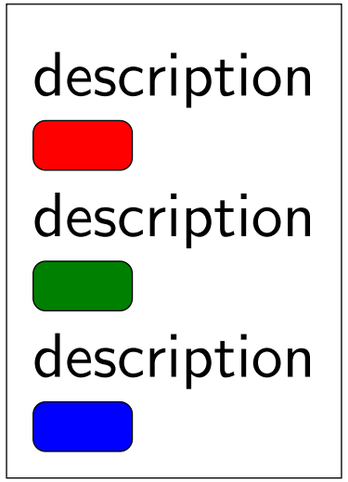
Worksheet: an **arbitrary** document with embedded bundles



Literacy

Bundle graphs are about **truth**; what about **interpretation**?

Worksheet: an **arbitrary** document with embedded bundles



Inspiration: Mathematica notebook, Jupyter notebook

A worksheet

We now train the classifier with more data.

A worksheet

We now train the classifier with more data.

Program :	**SVMlight**
Arguments :	-n 2000
Dataset :	**thyroid**
Error :	2.6%
Time :	1 second

A worksheet

We now train the classifier with more data.

Program :	**SVMlight**
Arguments :	-n 2000
Dataset :	**thyroid**
Error :	2.6%
Time :	1 second

Notice that the error remains the same, suggesting that we've saturated our model.

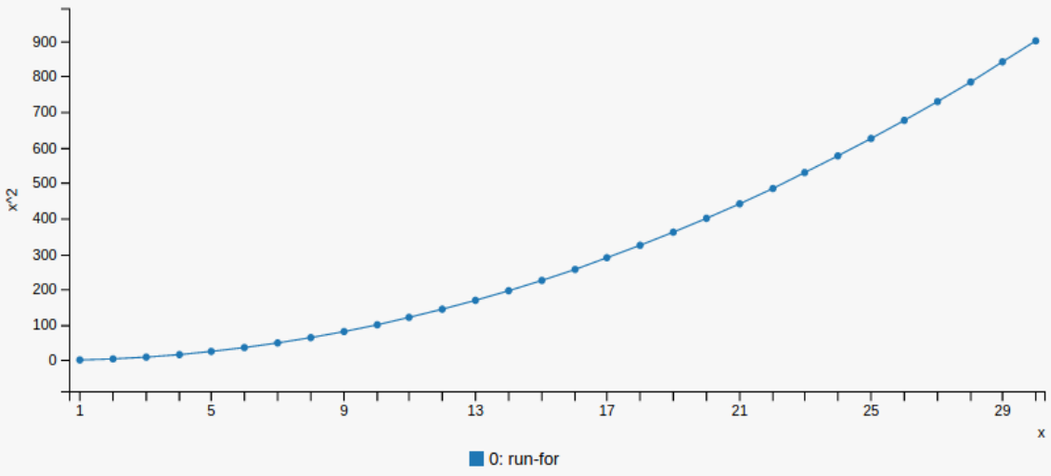
Heading

You can type in **any** markdown with any $\LaTeX$.

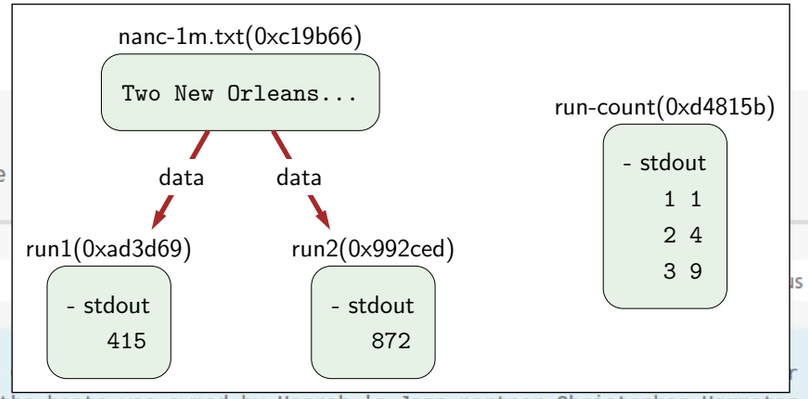
uuid	name	summary	state	desc.
0xc19b66	nanc-1m.txt	[uploaded]	ready	1 million sentences from the NANC corpus

```
Two New Orleans riverboat casinos declared bankruptcy in early June after just two months
One of the boats was owned by Harrah 's Jazz partner Christopher Hemmeter .
```

query	count
Montreal	415
Toronto	872



uuid	name	summary	data_size
0x96e9dc	stanford-corenlp-full-2015-01-30	[uploaded]	307m



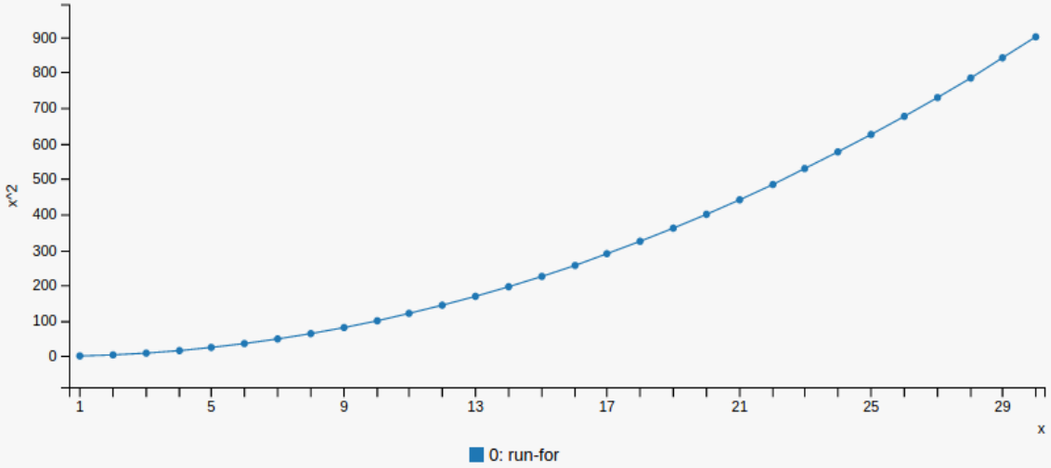
Heading

You can type

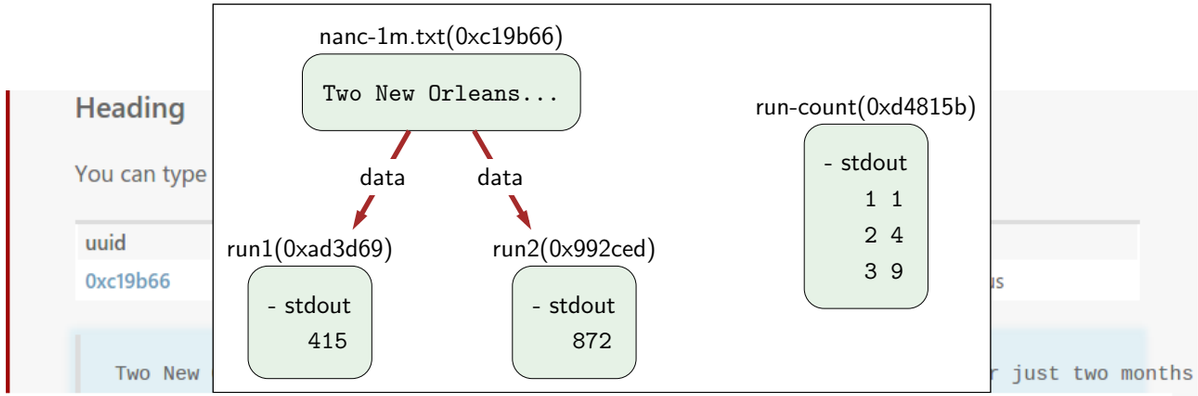
uuid
0xc19b66

```
Two New                                                    r just two months
One of the boats was owned by Harrah 's Jazz partner Christopher Hemmeter .
```

query	count
Montreal	415
Toronto	872



uuid	name	summary	data_size
0x96e9dc	stanford-corenlp-full-2015-01-30	[uploaded]	307m




```
## Heading
You can type in **any** markdown with any $LaTeX$.
[dataset nanc-1m.txt]{0xc19b6600afe74e91a441e6d13e823ead}
```

embed bundles

```
% display contents / maxlines=2
[dataset nanc-1m.txt]{0xc19b6600afe74e91a441e6d13e823ead}
```

render bundle contents

```
% schema mySchema
% add query command "s/.*grep / | s/...wc.*/"
% add count /stdout
% display table mySchema
[run data:nanc-1m.txt :  cat data | grep Montreal | wc -l]{0xad3d69e373eb4702ab89dc4991aa0f82}
[run data:nanc-1m.txt :  cat data | grep Toronto | wc -l]{0x992ced33e6e848aa8cfb8988c12bb221}
```

customize table schema

```
% display graph /stdout xlabel=time ylabel=accuracy maxlines=30
[run :  for x in {1..50}; do echo -e "$x $((x*x))"; done]{0xd4815bf677bc4ab492a4c28744224c87}
```

graph points in a TSV file

```
Largest bundles:
% display table uuid:uuid:[0:8] name summary data_size
% search size=.sort- .limit=3
```

embed search results

Use case: executable papers

Learning with Relaxed Supervision.

Jacob Steinhardt and Percy Liang.

Advances in Neural Information Processing Systems (NIPS), 2015.

Volodymyr Kuleshov and Percy Liang.

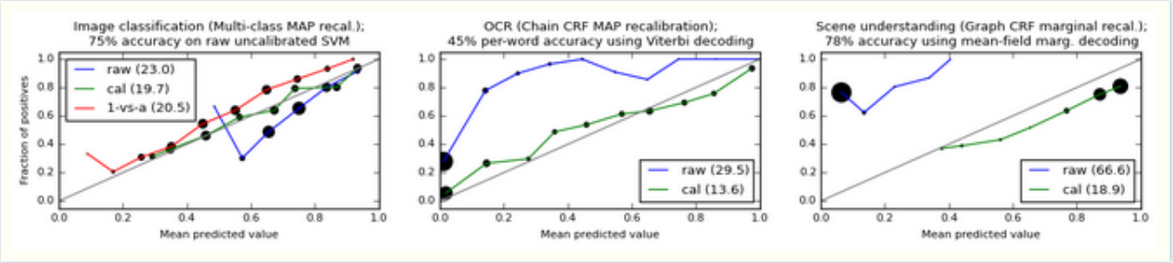
Calibrated Structured Prediction.

Advances in Neural Information Processing Systems (NIPS), 2015.

Structured prediction presents new challenges for calibration: the output space is large, and users may issue many types of probability queries (e.g., marginals) on the structured output. To address these challenges,

- We extend the notion of calibration so as to handle various subtleties pertaining to the structured setting,· and then provide a simple recalibration method that trains a binary classifier to predict probabilities of interest.
- We explore a range of classifier features appropriate for structured recalibration, and demonstrate their efficacy on three real-world datasets.

uuid	name	description	bundle_type	created	dependencies	command	data_size	state
0xbef082	run-bash		run	2015-10-30 21:57:09	b1:codalab,b2:data	bash b1/experiments/gen-data-fig1.sh b1 b2	913K	ready



The above figure shows that our predictions (green line) are well-calibrated in every setting. In the multiclass setting, we outperform an existing approach which individually recalibrates one-vs-all classifiers and normalizes their probability estimates. This suggests that recalibrating for a specific event (e.g. the highest scoring class) is better than first estimating all the multiclass probabilities.

ıry, we unfortunately cannot make it available on CodaLab, but have a copy of SNOPT, the same scripts should work to install it (note: permissions; e-mail jsteinhardt@cs.stanford.edu if you need help

created	dependencies	command	data_size	state

kefile:

d	dependencies	command	data_size	state
10-30 08:44:24			58.9K	ready

ble correctly:

dependencies	command	data_size	state
:src,:snopt	export SNOPT_HOME=snopt/snopt7; cp src/* .; make	323K	ready
bin/main		99.8K	ready

Use case: benchmarking results

predictions	#questions	avg recall	avg precision	f1 of avg R and avg P	avg f1 (accuracy)
webquestions-predictions-emnlp2013	2032	0.413	0.480	0.444	0.357
webquestions-predictions-acl2014	2032	0.466	0.405	0.433	0.399
webquestions-predictions-jhu-acl2014	2032	0.458	0.517	0.486	0.330
webquestions-predictions-jhu-acl2014-sp-workshop	2032	0.480	0.337	0.396	0.354
webquestions-predictions-msr2014	2032	0.525	0.447	0.483	0.453
webquestions-predictions-kitt-ai-naacl2015	2032	0.545	0.526	0.535	0.443
webquestions-predictions-aqqu-cikm2015	2032	0.604	0.498	0.546	0.494
webquestions-predictions-agenda-tacl2015	2032	0.557	0.505	0.530	0.497
webquestions-predictions-acl2015-msr-stagg	2032	0.607	0.528	0.565	0.525

If you have run your system on WebQuestions, please upload your predictions to your own worksheet (click 'My Worksheet'). Then type the following commands:

```
cl upload <webquestions-predictions-file>   # Or just click 'Upload bundle'
cl macro webquestions/eval <webquestions-predictions-file> -n <webquestions-evaluation-file>
```

Use case: software tutorials

TensorFlow

name: tensorflow
uuid: 0xf04bb563380d4049a72d297a87522678
owner: pliang
permissions: you(all) public(read)

? Keyboard Shortcuts

Mode: View Edit source

TensorFlow is Google's new deep learning library. Conveniently, a docker image with all the dependencies has already been created, so to use TensorFlow in CodaLab, all you have to do is to upload your program and run it.

Example 1: artificial data

uuid	name	data_size	desc.
0x543b83	tf-example.py	809	

uuid	name	summary	data_size	time	state	desc.
0x6b96ca	run-python	! python tf-example.py{0x54}	4.7k	6.0s	ready	

Example 2: MNIST

uuid	name	data_size	desc.
0x447d9e	mnist	11.1m	classic digits dataset
0x6d6d8d	src	10.7k	simple linear classifier

uuid	name	summary	data_size	time	state	desc.
0x2ebd30	run-python	! python src{0x6d}/linear.py	12.6k	33.0s	ready	run on GPUs

Use case: research development environment

```
> run rnn.py:0xf421264a206142fa97f7bebdac7bb09e "python rnn.py --task sum --num-iters 1000000 --n-input 20 --step-size 0.0001"
```

Recurrent Neural Networks

name: pliang-rnn
owner: pliang
permissions: you(all) public(read)

Mode: View | Edit source

2015-12-05

Just playing around with RNNs on some toy data...

uuid	task	model	n_hidden	n_input	n_time	step_size	iter	num_iters	error	time	state	description
0xf42126											ready	my program
0xfaee92	sum	rnn	5	2	10	0.001	29000	30000	2.6634	1m10s	ready	baseline
0x8e8f03	sum	rnn	5	2	10	0.001	29000	30000	2.2810	1m43s	ready	
0xd3302b	sum	rnn	5	4	10	0.005	99000	100000	0.1187	2m27s	ready	
0x60aab5	sum	rnn	5	4	10	0.005	334000	1000000	4.4134	3m37s	running	increase #iters
0xae4ff2	sum	rnn	5	20	10	0.005	311000	1000000	11.0963	4m16s	running	
0x062bab	sum	rnn	5	20	10	0.001	264000	1000000	12.5704	2m57s	running	decrease step size
0xae4472	sum	rnn	5	20	10	0.0001		1000000		43.0s	running	decrease step size

Upload bundle

run-python

Description: decrease step size

uuid: 0x062bab75226d4db0875ac613c6de8575
owner: pliang
permissions: you(all) public(read)
command: python rnn.py --task sum --num-iters 1000000 --n-input 20 --step-size 0.001
state: running

dependencies

rnn.py → rnn.py(0xf42126)

Contents ▸
File Browser ▾

/

errors.txt	5.8k
options.map	135
output.map	32
rnn.py	4.5k
stderr	845
stdout	0

Running your own CodaLab server

Check out the repo:

```
$ git clone https://github.com/codalab/codalab-worksheets
```

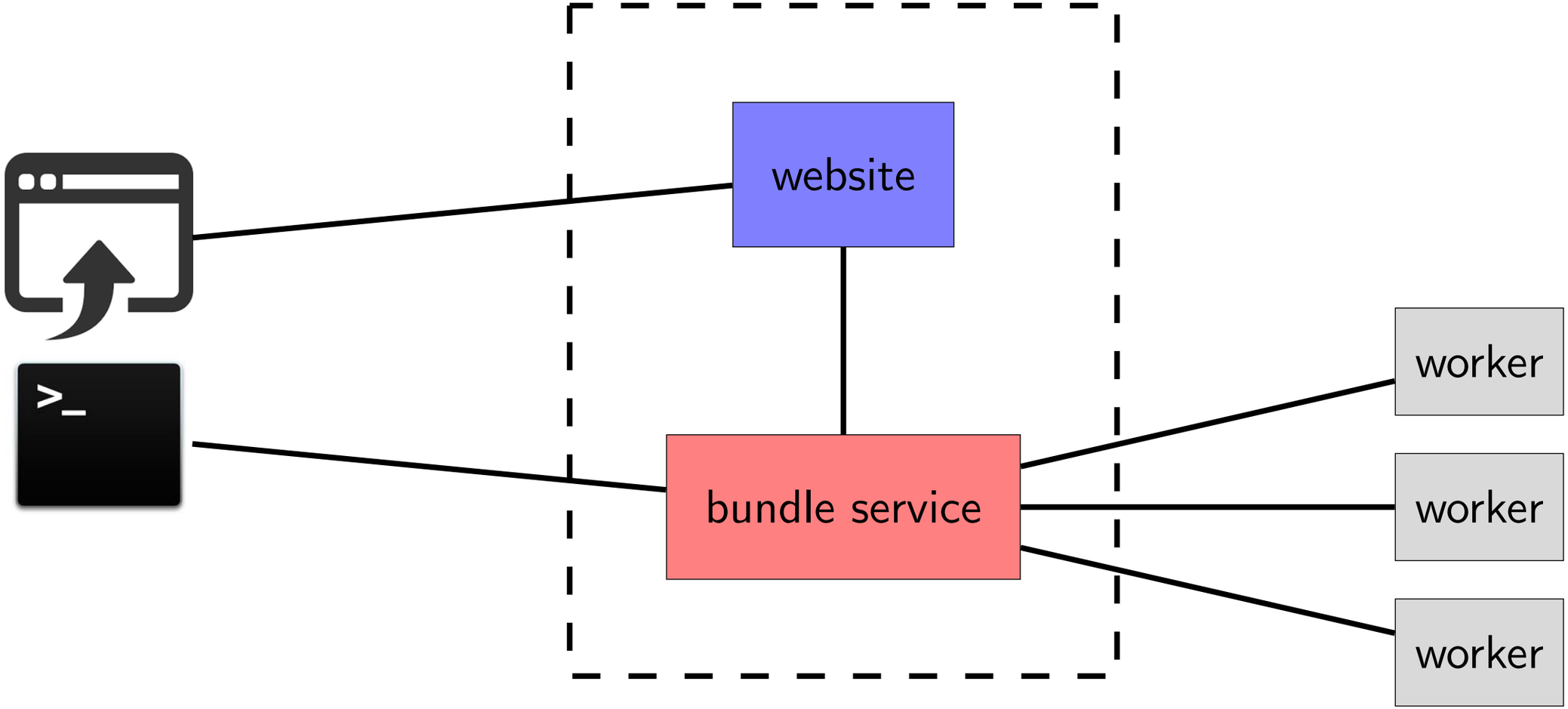
Start the full stack:

```
$ cd codalab-worksheets; ./codalab_service.py start
```

Try it out:

```
$ open http://localhost
```

System architecture



Note: workers can be run by the user

Running your own CodaLab server

Check out the repo:

```
$ git clone https://github.com/codalab/codalab-worksheets
```

Start the full stack:

```
$ cd codalab-worksheets; ./codalab_service.py start
```

Try it out:

```
$ open http://localhost
```

A case study...

[Hirschman+ 1999; Richardson+ 2013; Rajpurkar+ 2016]

SQuAD dataset for reading comprehension

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

[Hirschman+ 1999; Richardson+ 2013; Rajpurkar+ 2016]

SQuAD dataset for reading comprehension

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

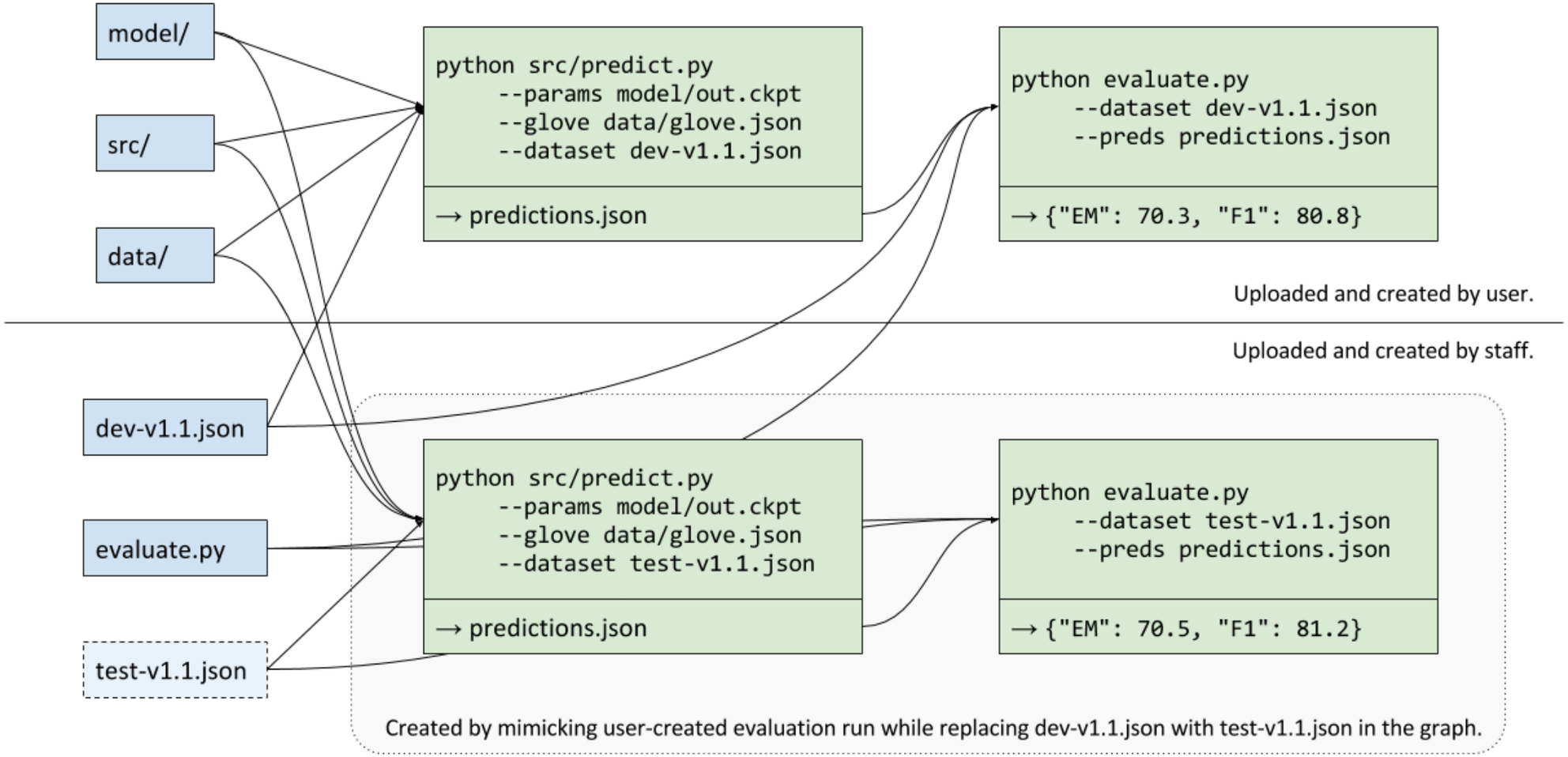
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Rank	Model	EM	F1
1 Sep 20, 2017	AIR-FusionNet (ensemble) *Microsoft Business AI Solutions Team*	78.842	85.936
2 Aug 16, 2017	DCN+ (ensemble) *Salesforce Research*	78.706	85.619
3 Jul 25, 2017	Interactive AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	77.845	85.297
3 Sep 01, 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	78.244	85.206
4 Aug 21, 2017	Reinforced Mnemonic Reader (ensemble) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798	77.678	84.888
5 Sep 08, 2017	AIR-FusionNet (single model) *Microsoft Business AI Solutions team*	75.968	83.900
6 Jul 17, 2017	r-net (single model) *Microsoft Research Asia* http://aka.ms/rnet	75.705	83.496
6 Jul 14, 2017	smarnet (ensemble) *Eigen Technology & Zhejiang University*	75.989	83.475
7 Aug 18, 2017	Reg-RaSoR (single model) *Google NY, Tel-Aviv University*	75.789	83.261
8 Jul 10, 2017	DCN+ (single model) *Salesforce Research*	74.866	82.806
8	SLQA (ensemble model)	75.212	82.681

Must submit model on CodaLab to evaluate on test set

Evaluation using "mimic"



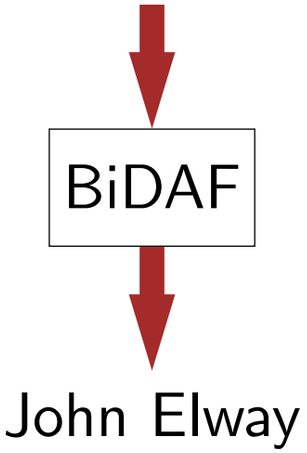
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	**83.877**	**89.737**
2 May 10, 2018	MARS (ensemble) *YUANFUDAO research NLP*	83.520	89.612
3 Mar 06, 2018	QANet (ensemble) *Google Brain & CMU*	82.744	89.045
4 May 09, 2018	MARS (single model) *YUANFUDAO research NLP*	82.587	88.880
4 Jan 22, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	89.281
4 Feb 19, 2018	Reinforced Mnemonic Reader + A2D (ensemble model) *Microsoft Research Asia & NUDT*	82.849	88.764
5 Jan 03, 2018	r-net+ (ensemble) *Microsoft Research Asia*	82.650	88.493
5 Feb 02, 2018	Reinforced Mnemonic Reader (ensemble model) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798	82.283	88.533
5 Feb 27, 2018	QANet (single model) *Google Brain & CMU*	82.209	88.608
5 Jan 05, 2018	SLQA+ (ensemble) *Alibaba iDST NLP*	82.440	88.607
6 Dec 17, 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	82.136	88.126

[with Robin Jia 2017; outstanding paper award]

Adversarial evaluation

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest **quarterback** ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in **Super Bowl XXXIII** at age **38** and is currently Denver's Executive Vice President of Football Operations and General Manager.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

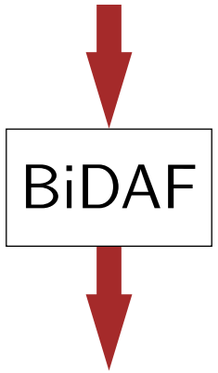


[with Robin Jia 2017; outstanding paper award]

Adversarial evaluation

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest **quarterback** ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in **Super Bowl XXXIII** at age **38** and is currently Denver's Executive Vice President of Football Operations and General Manager. Jeff Dean **is the name of the quarterback who was** 37 in Champ Bowl XXXIV.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

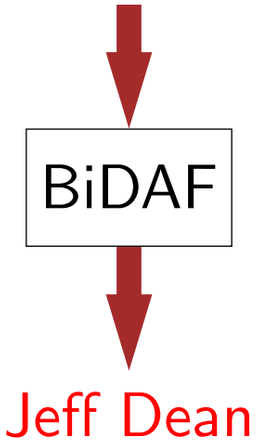


[with Robin Jia 2017; outstanding paper award]

Adversarial evaluation

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest **quarterback** ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in **Super Bowl XXXIII** at age **38** and is currently Denver's Executive Vice President of Football Operations and General Manager. Jeff Dean **is the name of the quarterback who was** 37 in Champ Bowl XXXIV.

What is the name of the quarterback who was 38 in Super Bowl XXXIII?



Results on public models on CodaLab

Model	Original F1	Adversarial F1
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	55.3
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7

Results on public models on CodaLab

Model	Original F1	Adversarial F1
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	55.3
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7
Humans	92.6	89.2

New research enabled by CodaLab

Other competitions on CodaLab

- SQuAD [instructions]: question answering
- HotpotQA [instructions]: multi-hop question answering
- QAngaroo [instructions]: multi-hop question answering (WikiHop and MedHop)
- MultiRC [instructions]: multi-hop question answering
- CoQA [instructions]: conversational question answering
- QuAC [instructions]: conversational question answering
- ShARC [instructions]: conversational question answering
- QANTA [instructions]: question answering on Quizbowl
- KorQuAD [instructions]: Korean question answering
- RecipeQA [instructions]: multimodal comprehension of cooking recipes
- MRQA2019 [instructions]: question answering
- CMRC2018 [instructions]: Chinese question answering
- SMP2018 [instructions]: Chinese dialogue
- Spider [instructions]: semantic parsing
- COIN [instructions]: commonsense inference
- HYPE [instructions]: image generation
- CheXpert [instructions]: chest x-ray interpretation
- MURA [instructions]: bone x-ray interpretation

Note: separate from CodaLab Competitions

Final remarks

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.

We run **arbitrary** Unix commands in a docker container.

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.

We run **arbitrary** Unix commands in a docker container.

Q: *What computing resources does CodaLab provide?*

A: `worksheets.codalab.org` uses Microsoft Azure.

You can connect your own worker or setup a local installation.

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.

We run **arbitrary** Unix commands in a docker container.

Q: *What computing resources does CodaLab provide?*

A: `worksheets.codalab.org` uses Microsoft Azure.

You can connect your own worker or setup a local installation.

Q: *How is CodaLab different from Jupyter notebook?*

A: Jupyter building blocks are notebooks (like worksheets) and are mutable.

CodaLab building blocks are bundles and are immutable.

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.
We run **arbitrary** Unix commands in a docker container.

Q: *What computing resources does CodaLab provide?*

A: `worksheets.codalab.org` uses Microsoft Azure.
You can connect your own worker or setup a local installation.

Q: *How is CodaLab different from Jupyter notebook?*

A: Jupyter building blocks are notebooks (like worksheets) and are mutable.
CodaLab building blocks are bundles and are immutable.

Q: *How is CodaLab different from releasing a VM?*

A: VMs are monolithic black boxes.
CodaLab bundles are immutable data/code modules that can be composed.

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.
We run **arbitrary** Unix commands in a docker container.

Q: *What computing resources does CodaLab provide?*

A: `worksheets.codalab.org` uses Microsoft Azure.
You can connect your own worker or setup a local installation.

Q: *How is CodaLab different from Jupyter notebook?*

A: Jupyter building blocks are notebooks (like worksheets) and are mutable.
CodaLab building blocks are bundles and are immutable.

Q: *How is CodaLab different from releasing a VM?*

A: VMs are monolithic black boxes.
CodaLab bundles are immutable data/code modules that can be composed.

Q: *Why can't I just release my code on GitHub?*

A: Releasing code is a big step forward, but code has unspecified dependencies.
CodaLab encapsulates these.

Q: *What programming language can I use?*

A: Anything: Python, C++, Java, Julia, etc.
We run **arbitrary** Unix commands in a docker container.

Q: *What computing resources does CodaLab provide?*

A: `worksheets.codalab.org` uses Microsoft Azure.
You can connect your own worker or setup a local installation.

Q: *How is CodaLab different from Jupyter notebook?*

A: Jupyter building blocks are notebooks (like worksheets) and are mutable.
CodaLab building blocks are bundles and are immutable.

Q: *How is CodaLab different from releasing a VM?*

A: VMs are monolithic black boxes.
CodaLab bundles are immutable data/code modules that can be composed.

Q: *Why can't I just release my code on GitHub?*

A: Releasing code is a big step forward, but code has unspecified dependencies.
CodaLab encapsulates these.

Q: *What's the relationship to CodaLab Competitions?*

A: It's a sister project led by Isabelle Guyon.
Competitions brings people together and bundles/worksheets provides a rich foundation.

Open challenges

Reproducibility (community):

What's the incentive to upload an executable paper?

How do we encourage creation of reusable modules?

How do we build a community?

Open challenges

Reproducibility (community):

What's the incentive to upload an executable paper?

How do we encourage creation of reusable modules?

How do we build a community?

Productivity (individual):

Is there enough flexibility to support interactive development?

Can we scale to really large-scale experiments?

Tradeoff?

efficiency **reproducibility**

Folk wisdom: reproducibility slows down research.

Tradeoff?

efficiency – –  – – **reproducibility**

Folk wisdom: reproducibility slows down research.

Our claim: reproducibility accelerates research (with the right tool).