



Google AI ML Winter Camp  
谷歌 AI 机器学习应用冬令营

# Project Announcement

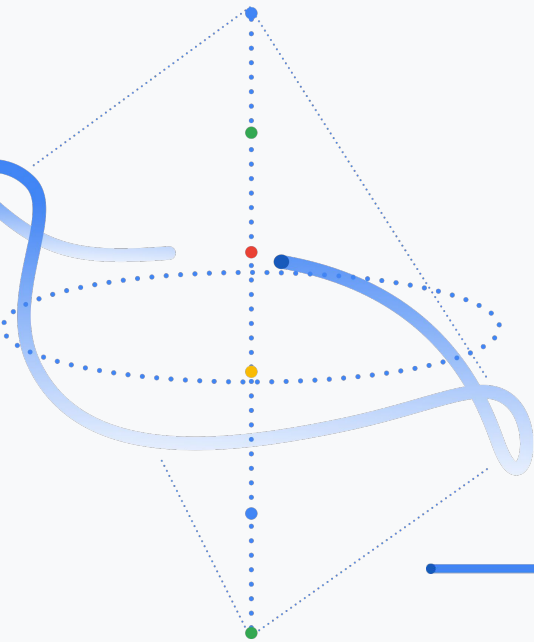
Google AI ML Winter Camp, SHA  
Jan 14 -- 18, 2019

# ML Projects in Action

Practice real ML application

ML requires practice with real problems.

- Form: Group project ( $\leq 3$  ppl)
- Complete an innovative ML application.
  - Pick a project from candidate list, and formulate an ML problem (CV, NLP, etc.).
  - Design an ML algorithm.
  - Train models and iterate to tune parameters.
  - Evaluation and application demo.
- Github repo (you own your project, but open for peer review)
- Friday: Presentation and poster.



# Computational Resources

Each one of you is provided with 1 GCP Linux.

Google Cloud Platform ML Winter Camp SHA

Compute Engine VM instances

Filter VM instances

Name	Zone	Recommendation	Internal IP	External IP	Connect
<input checked="" type="checkbox"/> admin-instance	asia-east1-a		10.140.0.57 (nic0)	34.80.106.39	SSH
<input type="checkbox"/> instance-1	us-east1-b		10.142.0.6 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-archonshen	asia-east1-a		10.140.0.55 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-chuangchuan	asia-east1-a		10.140.0.18 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-cyx2010	asia-east1-a		10.140.0.46 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-dongjiali1994	asia-east1-a		10.140.0.35 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-dongming-sun95	asia-east1-a		10.140.0.40 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-forwchen	asia-east1-a		10.140.0.45 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-gongjiyang5885	asia-east1-a		10.140.0.10 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-gothic2ai	asia-east1-a		10.140.0.56 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-guifuqiansha	asia-east1-a		10.140.0.13 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-hatsuyukiw	asia-east1-a		10.140.0.38 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-hongge831	asia-east1-a		10.140.0.28 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-jzqz17	asia-east1-a		10.140.0.11 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-kylechenkc	asia-east1-a		10.140.0.44 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-layla-laisy	asia-east1-a		10.140.0.53 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-lchn-guo	asia-east1-a		10.140.0.20 (nic0)	None	SSH
<input type="checkbox"/> winter-camp-lhc19941010	asia-east1-a		10.140.0.27 (nic0)	None	SSH

<https://console.cloud.google.com/home>

```
Linux admin-instance 4.9.0-8-amd64 #1 SMP Debian 4.9.110-3+deb9u6 (2018-10-08) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

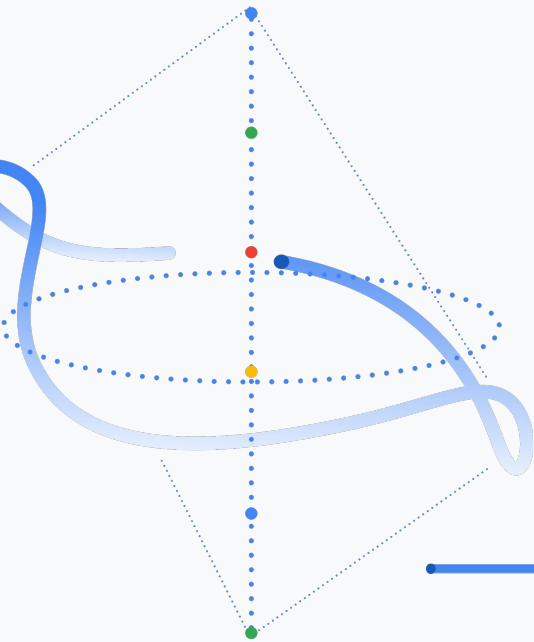
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Jan 13 14:13:54 2019 from 74.125.41.97
tianlin@admin-instance:~$
```

# Computational Resources

Each one is provided with 1 GCP Linux machine:

- 16 CPUs
- 60 GB memory
- 1 NVIDIA Tesla P100 GPU
- 500 GB SSD persistent disk

During Winter Camp. (Mon. -- Fri.; Backup Sat.)



# Project List

## Quick, Draw! Doodle Recognition Challenge (CV)

<https://www.kaggle.com/c/quickdraw-doodle-recognition/data>

### Data

50 million drawings across 345 categories.

### Potential Tasks

Image classification, similarity search.



# Project List

## Face Attribute Transfer (CV)

<http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

### Data

10,177 identities, 202,599 face images, and 5 landmark locations, 40 binary attributes annotations

### Potential Tasks

Image Transfer, Face Recognition

Eyeglasses



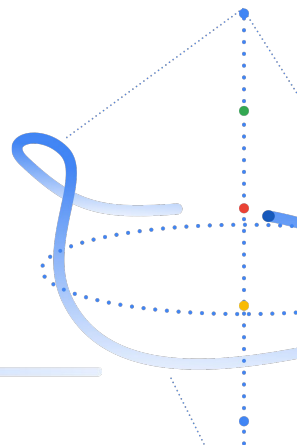
Wearing Hat



Bangs



Wavy Hair



# Project List

## Cartoon Face (CV)

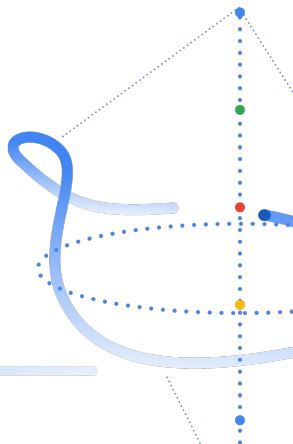
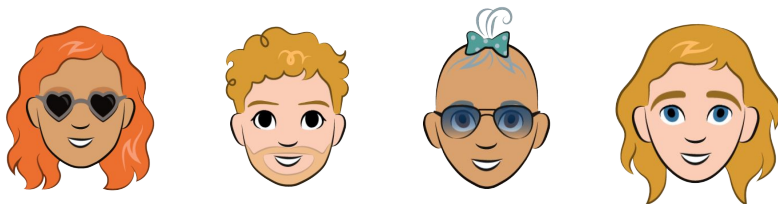
<https://google.github.io/cartoonset/download.html>

### Data

10k/100k images. 16 components that vary in 10 artwork attributes (eye, chin, face, hair, etc.), 4 color attributes, and 4 proportion attributes.

### Potential Tasks

Face recognition, classification, transfer



# Project List

## Speech Synthesis (Speech)

<https://keithito.com/LJ-Speech-Dataset/>

### Data

Public domain speech dataset. 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. 1~10 seconds each, 24 hours in total.

### Potential Tasks

Text To Speech (Char2Feats, WaveRNN)

### Instructions

Use <https://github.com/tensorflow/lingvo>

See: *Project Instructions - Build a Text-To-Speech System (attached doc)*

### Sample Audio:

Many animals of even complex structure which live parasitically within others are wholly devoid of an alimentary cavity.



# Project List

## Humpback Whale Identification (CV)

<https://www.kaggle.com/c/humpback-whale-identification>

### Data

25,000+ images with 3,000+ whale IDs gathered from research institutions and public contributors. Real world problem for animal protection.

### Potential Tasks

Image classification, similarity search



# Project List

## Douban Movies Short Comments (NLP)

<https://www.kaggle.com/utmhikari/doubanmovieshortcomments>

### Data

2,131,887 examples, 5 stars rating. Chinese data.

### Potential Tasks

Text classification, regression, multi-tasks

MovieEnName	MovieCn Name	Crawl Date	#Number	Username	Date	#Stars	Comment	#Like
Avengers Age of Ultron	复仇者联盟2	2017-01-22	71	Nina	2015-05-12	2	超长版运动产品广告	10
The Ghouls	寻龙诀	2017-01-25	75005	layliet	2015-12-18	5	电影院刚看完, 给九分, 还有一分不给是怕中国电影骄傲。	0

# Project List

## Chinese Daily New Text (NLP)









<https://www.kaggle.com/noxmoon/chinese-official-daily-news-since-2016>

### Data

Chinese news data, Xinwen Lianbo. 20,738 piece of news since 1 January 1978.

### Potential Tasks

Text summarization/abstraction, transfer learning

 date		 tag		 headline		 content	
date of the news		“详细全文” some relatively long news, “国内” domestic short news, “国际” international short news. These tags may not		summary for each piece of news		news text	

# Project List

## MBTI personality type with posts (NLP)

<https://www.kaggle.com/datasnaek/mbti-type>

### Data

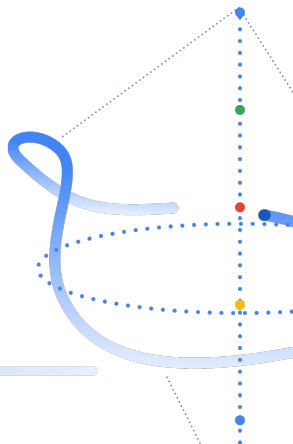
The Myers Briggs Type Indicator (or MBTI for short). 16 personality types across 4 axis. Text Classification 8675 examples

### Potential Tasks

Text classification, sentimental analysis

**I (Introversion) N (Intuition) F (Feeling) P(Perceiving)**

*... Hey there stranger :P I remember you from when I first joined PerC several years ago. I always felt you were cheerful, intelligent, and wise beyond your years. Look at the amount of thanks you...*



# Project List

## Relatively Easy Kaggle Tasks

### Data

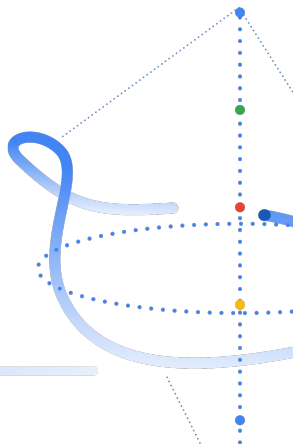
Titanic: <https://www.kaggle.com/c/titanic>

House Prices: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Predict Future Sales: <https://www.kaggle.com/c/competitive-data-science-predict-future-sales>

### Potential Tasks






Structural data prediction, Multi-tasks.



# Open Datasets

## Google AI competitions on Kaggle (CV)

<https://www.kaggle.com/googleai/competitions>







5 competitions		
	<b>Quick, Draw! Doodle Recognition Challenge</b> How accurately can you identify a doodle? <i>Featured</i> · a month ago	<b>\$25,000</b> 1,316 teams
	<b>Inclusive Images Challenge</b> Stress test image classifiers across new geographic distributions <i>Research</i> · 2 months ago	<b>\$25,000</b> 468 teams
	<b>Google AI Open Images - Object Detection Track</b> Detect objects in varied and complex images. <i>Featured</i> · 4 months ago	<b>\$30,000</b> 454 teams
	<b>Google AI Open Images - Visual Relationship Track</b> Detect pairs of objects in particular relationships. <i>Featured</i> · 4 months ago	<b>\$20,000</b> 232 teams
	<b>The 2nd YouTube-8M Video Understanding Challenge</b> Can you create a constrained-size model to predict video labels? <i>Featured</i> · 5 months ago	<b>\$25,000</b> 312 teams

# Open Datasets

## Google BigQuery

<https://www.kaggle.com/bigquery/datasets>

27 datasets

5		<b>USPTO OCE Patent Claims Research Data</b> US patents granted claims and published applications (BigQuery) Google BigQuery (Continuous updates)	research law bigquery	BigQuery 119.3 GB CC4	</> 2 0 967
22		<b>World Development Indicators (WDI) Data</b> World Bank collection of global development indicators (BigQuery) Google BigQuery (Continuous updates)	bigquery economics world finance	BigQuery 2.2 GB CC3	</> 2 0 7k
2		<b>USPTO Patent Trial and Appeal Board (PTAB) Data</b> Trials conducted by PTAB for issues of patentability (BigQuery) Google BigQuery (Continuous updates)	law bigquery	BigQuery 137 MB CC4	</> 2 0 780
2		<b>USPTO Patent Examiner Data System (PEDS) Data</b> Data from the examination process of USPTO patent applications (BigQuery) Google BigQuery (Continuous updates)	bigquery law research	BigQuery 40.2 GB CC4	</> 2 0 877
10		<b>USPTO Patent Examination Research Data (PatEx)</b> Millions of publicly viewable patent applications filed with USPTO (BigQuery) Google BigQuery (Continuous updates)	research law bigquery	BigQuery 30.2 GB CC4	</> 2 0 897
6		<b>The Office Action Research Dataset for Patents</b> Actions taken by patent examiners to patent applicants (BigQuery) Google BigQuery (Continuous updates)	bigquery research law	BigQuery 3.9 GB CC4	</> 2 0 1k

## Important Tips:

- Innovative applications are beyond a single model. Find a good problem, create a demo and prepare for presentation.
- End2End asap, and then polish details. Appreciate original ML model. Using API is OK but not major innovation.
- (If based on existing work) we mainly evaluate your work and innovation during Winter Camp. Your reference should be clear.

