

Kaggle: Nuts and Bolts

By Noah Gundotra

Brought to you by SUSA x Cal Kaggle Team!





Kaggle: What is Kaggle?

- Host data science competitions
 - Datasets
 - Leaderboards
 - Community (like Piazza)
- Recruitment
- Research



Kaggle

[Demo]



Kaggle: What is Kaggle?

- Hosts Jupyter Notebooks for data exploration
- Supports Python, R, Julia
- Somewhat complicated to use properly @ first



Kaggle: Why Kaggle?



- Recognition, \$, Learning
- Facebook, SAP, BNP Paribas, J.P. Morgan
- Even Shell(?)





Got a tip? [Let us know.](#)

Follow Us [f](#) [g](#) [t](#) [y](#) [in](#) [g+](#) [r](#)

News ▾ Video ▾ Events ▾ Crunchbase

[Message Us](#)

[Search](#)



DISRUPT BERLIN Early Bird sale ends this Wednesday 22 November [Get your tickets today & save ▶](#)

kaggle

Google

acquisition

Google is acquiring data science community Kaggle

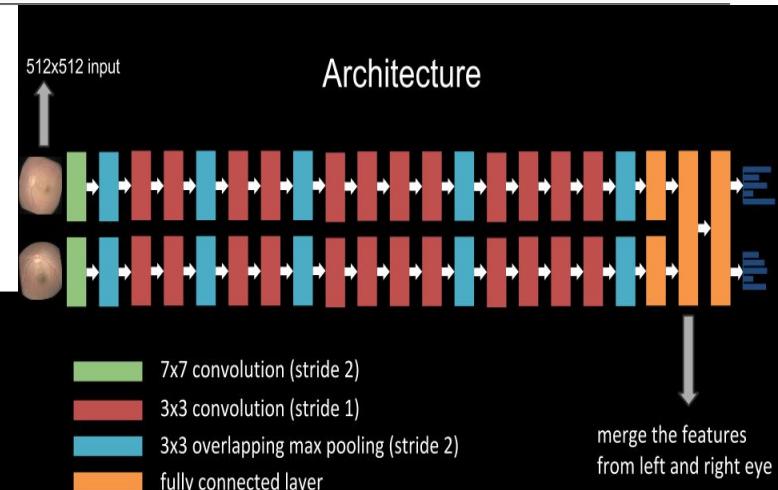
Posted Mar 7, 2017 by [Frederic Lardinois \(@fredericl\)](#), [Matthew Lynley \(@mattlynley\)](#), [John Mannes \(@JohnMannes\)](#)



[Next Story ▶](#)

From Kaggle to Google DeepMind: An interview with Jeffrey De Fauw

Megan Risdal | 07.11.2016



(Thanks to Sander for letting me use his visualisation.)



anokas

Mikel Bober-Irizar

Guildford, England, United Kingdom

Joined 2 years ago · last seen 10 days ago



<http://mxbi.net>

Followers 663

Following 32



Competitions
Master

[Home](#)

[Competitions \(43\)](#)

[Kernels \(138\)](#)

[Discussion \(426\)](#)

...

[Contact User](#)

[Follow User](#)

Competitions Master



Current Rank

101

of 66,214

Highest Rank

74

Kernels Master



Current Rank

4

of 109,136

Highest Rank

1

Discussion Master



Current Rank

16

of 40,791

Highest Rank

5

[\(Help | Advanced search\)](#)

Computer Science > Computer Vision and Pattern Recognition

Cultivating DNN Diversity for Large Scale Video Labelling

Mikel Bober-Irizar, Sameed Husain, Eng-Jon Ong, Miroslaw Bober

(Submitted on 13 Jul 2017)

We investigate factors controlling DNN diversity in the context of the Google Cloud and YouTube-8M Video Understanding Challenge. While it is well-known that ensemble methods improve prediction performance, and that combining accurate but diverse predictors helps, there is little knowledge on how to best promote & measure DNN diversity. We show that diversity can be cultivated by some unexpected means, such as model over-fitting or dropout variations. We also present details of our solution to the video understanding problem, which ranked #7 in the Kaggle competition (competing as the Yeti team).

1,111
of 66,214

1,111

1,111
of 109,136

1,111

1,111
of 40,791

1,111

\$1.25 Million

Will be awarded to Kaggle Competitors as
the TSA Screening Competition ends in 4
weeks

[Active](#)[All](#)[Entered](#)

Sort by

Prize

19 active competitions

All Categories

Search competitions



Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

\$1,500,000

401 teams

Featured · a month to go · 🛡️ terrorism, image, object detection



Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

\$1,200,000

3,780 teams

Featured · 2 months to go · 🏠 housing, real estate



Statoil/C-CORE Iceberg Classifier Challenge

Ship or iceberg, can you decide from space?

\$50,000

1,104 teams

Featured · 2 months to go · 🌡️ weather, shipping, binary classification



Competitions



ImageNet Object Detection from Video Challenge

Identify and label ordinary objects in videos

[Research](#) · 12 years to go · 📸 image, object detection



Spooky Author Identification

Share code and discuss insights to identify horror authors from their writings

\$25,000

645 teams

[Playground](#) · a month to go · 📖 literature, linguistics, multiclass classification



Dog Breed Identification

Determine the breed of a dog in an image

Kudos

326 teams

[Playground](#) · 3 months to go · 🐶 animals, image, multiclass classification, object identification



Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

9,219 teams

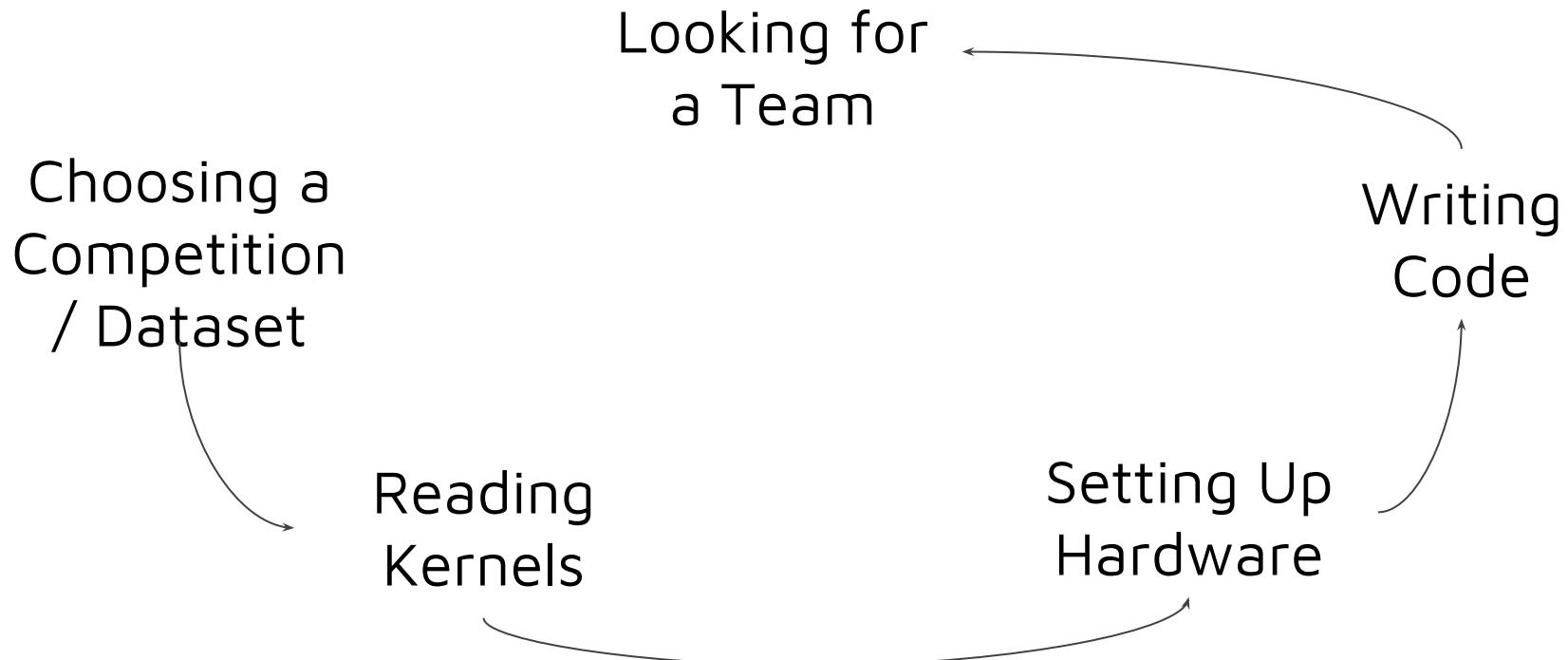
[Getting Started](#) · 2 years to go · 📊 tabular, binary classification



Competitions

Approaching Kaggle Competitions

Kaggle Workflow





How to Choose a Competition

- Money
 - High profile competitions, featured
- Recognition
 - Recruitment, Featured, Research, Playgrounds
- Learning
 - Research, Playgrounds, Getting Started “competitions”



How to Choose a Dataset

If you have something you want to study: look for it on Kaggle first

- Stock Market Data
- Iowa Liquor Sales
- Fashion MNIST
- Data on Kagglers & Data Scientists
- Video Game Sales & Pricings



How to Read Kernels

- Some code you will need to understand
- Some code you will not

We will walk through the Dog Breed Identification Competition



JERU
LUKE

Jeru666

Dog-eat-dog world! (EDA & useful scripts)

0
voters

last run 16 hours ago · Python notebook · 46 views
using data from [Dog Breed Identification](#) · Public

Notebook

Code

Data (1)

Comments (0)

Log

Versions (2)

Fork Notebook

Notebook

What is in this kernel?



Reading Kernels



**JERU
LUKE**

Jeru666

Dog-eat-dog world! (EDA & useful scripts)

0
voters

last run 16 hours ago · Python notebook · 46 views
using data from [Dog Breed Identification](#) · Public

Notebook

Code

Data (1)

Comments (0)

Log

Versions (2)

Fork Notebook

Code

This script has been released under the [Apache 2.0](#) open source license.

[Download Code](#)

```
1 {"metadata": {"kernelspec": {"name": "python3", "display_name": "Python 3", "language": "python"},  
"language_info": {"name": "python", "pygments_lexer": "ipython3", "file_extension": ".py",  
"version": "3.6.3", "codemirror_mode": {"name": "ipython", "version": 3}, "nbconvert_exporter":  
"python", "mimetype": "text/x-python"}}, "nbformat": 4, "nbformat_minor": 1, "cells":  
[{"metadata": {"uuid": "9064a37581dee01c4405b9d3a7f20ceedf38c01f", "cell_guid": "fbaac637-6aa1-  
41ac-91cf-4d3b97f1924b"}, "cell_type": "markdown", "source": ["# What is in this kernel?\n", "##  
1. Data Exploration\n", "* Checking for missing values\n", "* Visualizing distribution of various  
dog breeds\n", "\n", "## 2. Useful Scripts\n", "* Script to segregate images into their
```



Reading Kernels



**JERU
LUKE**

Jeru666

Dog-eat-dog world! (EDA & useful scripts)

0
voters

last run 16 hours ago · Python notebook · 46 views
using data from [Dog Breed Identification](#) · Public

Notebook

Code

Data (1)

Comments (0)

Log

Versions (2)

Fork Notebook

Data

Dog Breed Identification

[labels.csv](#)

[sample_submission....](#)

[test.zip](#)

[train.zip](#)

Dog Breed Identification



Determine the breed of a dog in an image

Last Updated 2 months ago

About this Dataset



Reading Kernels

▼ Dog Breed Identification

labels.csv

sample_submission....

test.zip

train.zip

train.zip

train.zip

Download

.../input/train.zip

344.54 MB

Data source

**Dog Breed Identification**

Determine the breed of a dog in an image Last Updated 2 months ago

This file is an ZIP archive. ZIP archive files will be uncompressed and their contents available at the root folder when this dataset is used in Kernels.



Reading Kernels



Kernels: How They Work

- Built on Jupyter Notebook API
 - Just different CSS style
- Run on Docker, which is virtual system on Kaggle's Cloud
- Provide real-time access to data, indefinite storage of your data results (I believe)
- **Max 1 hour of continuous usage of Kaggle kernels**

[Code](#)[Issues 1](#)[Pull requests 0](#)[Projects 0](#)[Wiki](#)[Insights](#)

Branch: master ▾

[docker-python / Dockerfile](#)[Find file](#) [Copy path](#)

```
52 RUN apt-get install -y libfreetype6-dev && \
53     apt-get install -y libgl1-mesa-glx libxext6 libsm6 libxrender1 libfontconfig1 --fix-missing && \
54 # textblob
55 pip install textblob && \
56 #word cloud
57 conda install -y -c https://conda.anaconda.org/amueller wordcloud && \
58 #igraph
59 conda install -y -c conda-forge python-igraph && \
60 #xgboost
61 cd /usr/local/src && mkdir xgboost && cd xgboost && \
62 git clone --depth 1 --recursive https://github.com/dmlc/xgboost.git && cd xgboost && \
63 make && cd python-package && python setup.py install && \
64 pip install lightgbm && \
65 #lasagne
66 cd /usr/local/src && mkdir Lasagne && cd Lasagne && \
67 git clone --depth 1 https://github.com/Lasagne/Lasagne.git && cd Lasagne && \
68 pip install -r requirements.txt && python setup.py install && \
69 #keras
```

XGBoost

Lasagne

What is in this kernel?

1. Data Exploration

- Checking for missing values
- Visualizing distribution of various dog breeds

2. Useful Scripts

- Script to segregate images into their respective breeds (to help with augmentation later on)

[2]:

```
% ls
```

```
__notebook_source__.ipynb
```

What is in this kernel?

1. Data Exploration

- Checking for missing values
- Visualizing distribution of various dog breeds

2. Useful Scripts

- Script to segregate images into their respective breeds (to help with augmentation later on)

[3]:

```
% ls .. /
```

```
config/  input/  lib/  working/
```

What is in this kernel?

1. Data Exploration

- Checking for missing values
- Visualizing distribution of various dog breeds

2. Useful Scripts

- Script to segregate images into their respective breeds (to help with augmentation later on)

[4]:

% pwd

```
'/kaggle/working'
```

What is in this kernel?

1. Data Exploration

- Checking for missing values
- Visualizing distribution of various dog breeds

2. Useful Scripts

- Script to segregate images into their respective breeds (to help with augmentation later on)

[8]:

```
% ls /kaggle/input
```

```
labels.csv  sample_submission.csv  test/  train/
```

[7]:

```
% cat /kaggle/config/jupyter.json
```

```
{  
    "control_port": 1004,  
    "hb_port": 1003,  
    "iopub_port": 1002,  
    "ip": "*",  
    "key": "kaggle",  
    "shell_port": 1001,  
    "stdin_port": 1005,  
    "transport": "tcp"  
}
```

[8]:

```
% ls /kaggle/input
```

```
labels.csv  sample_submission.csv  test/  train/
```

[9]:

```
% ls /kaggle/input/train/
```

```
0021f9ceb3235effd7fcde7f7538ed62.jpg 814be837610c46b122ff45f71e97133d.jpg  
002211c81b498ef88e1b40b9abf84e1d.jpg 815079d1d62429b3134f2afaf1a53ef65.jpg  
00290d3e1fdd27226ba27a8ce248ce85.jpg 815949fad325d5bd758bd46c2bbccfaf.jpg  
002a283a315af96ea0e28e7163b21b.jpg 815d3c084bcb79bed798a9774d4ce66b.jpg  
003df8b8a8b05244b1d920bb6cf451f9.jpg 8161ff9d8b2b1b91280f268463c51065.jpg  
0042188c895a2f14ef64a918ed9c7b64.jpg 8165da6ab285d889a8a6f1980aad5869.jpg  
004396df1acd0f1247b740ca2b14616e.jpg 81662662eb22135f0438459487722f98.jpg  
0067dc3eab0b3c3ef0439477624d85d6.jpg 816da6c9a52fa67ad45bee98657e541b.jpg  
00693b8bc2470375cc744a6391d397ec.jpg 816f335f43b007f659fc7f8527301661.jpg  
006cc3ddb9dc1bd827479569fcfdc52dc.jpg 8174c92381421de8ab5cfb54366a086e.jpg  
0075dc49dab4024d12fafef67074d8a81.jpg 817696aaa5d843b6cbaad2698e461872.jpg  
00792e341f3c6eb33663e415d0715370.jpg 817d2fafec8ab5c3205bb0c0ff53fcd7.jpg  
007b5a16db9d9ff9d7ad39982703e429.jpg 818091f88618b285f4dd3f575e1dc74d.jpg  
007b8a07882822475a4ce6581e70b1f8.jpg 8182c860efb9ebf248886c0e70216e01.jpg  
007f6607e1a1115506a2511601 01007961151102056001007154101
```

Data Exploration

There are 120 dog breeds in our dataset. The following snippet confirms the fact

```
[13]: labels.breed.nunique()
```

```
120
```

```
[15]: labels.breed.nunique()  
type(labels.breed)
```

```
pandas.core.series.Series
```

Dog-eat-dog world! (EDA & useful scripts)

● Running

Restart



Python

Python 3.6.3 ⓘ

✖ Private

Saved

Publish

Input Files

Dog Breed Identification

labels.csv

sample_submission....

test.zip

train.zip

+ Add Data Source

Upload Dataset

Dog Breed Identification



Determine the breed of a dog in an image

Last Updated 2 months ago

✖ Remove

About this Dataset

You are provided with a training set and a test set of images of dogs. Each image has a filename that is its unique `id`. The dataset comprises 120 breeds of dogs. The goal of the competition is to create a classifier capable of determining a dog's breed from a photo. The list of breeds is as follows:

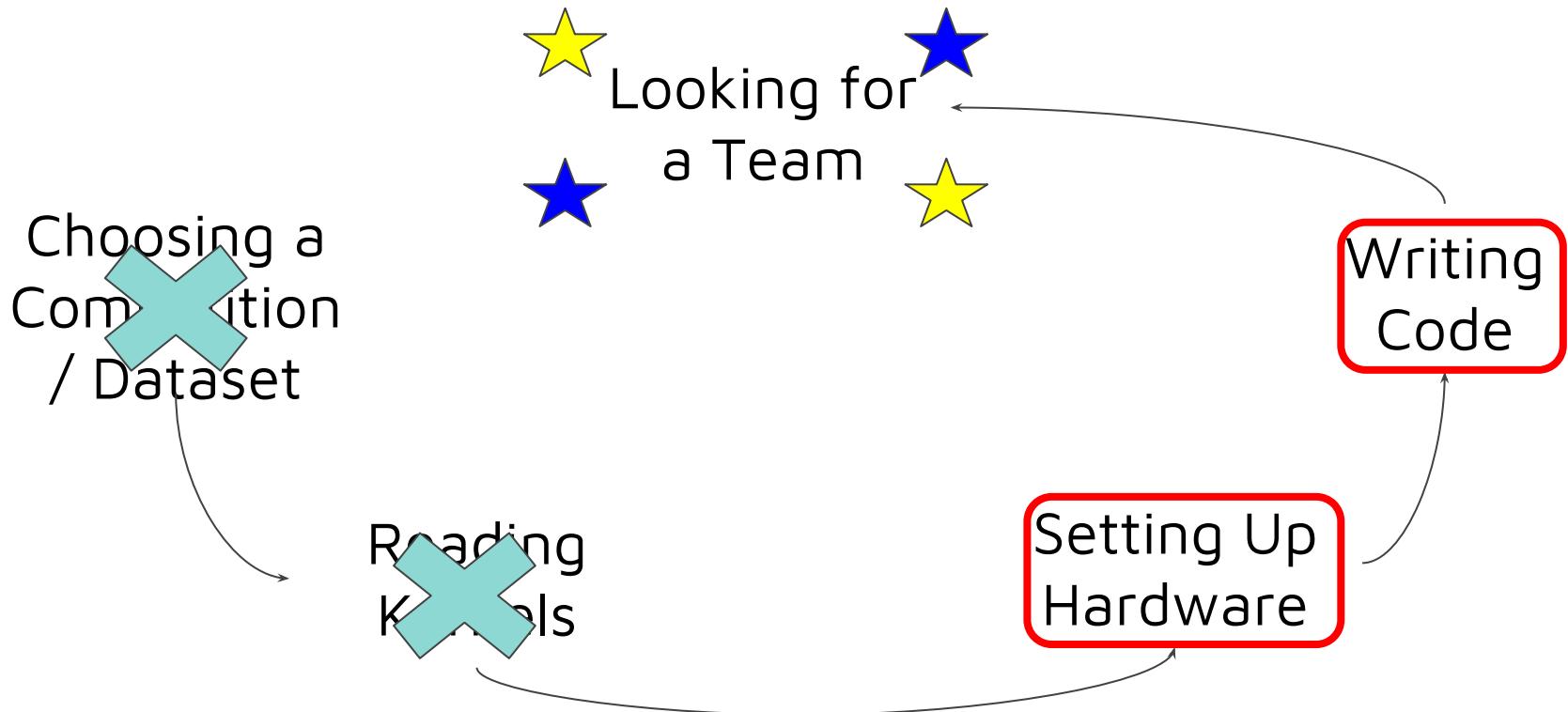
affenpinscher

afghan hound

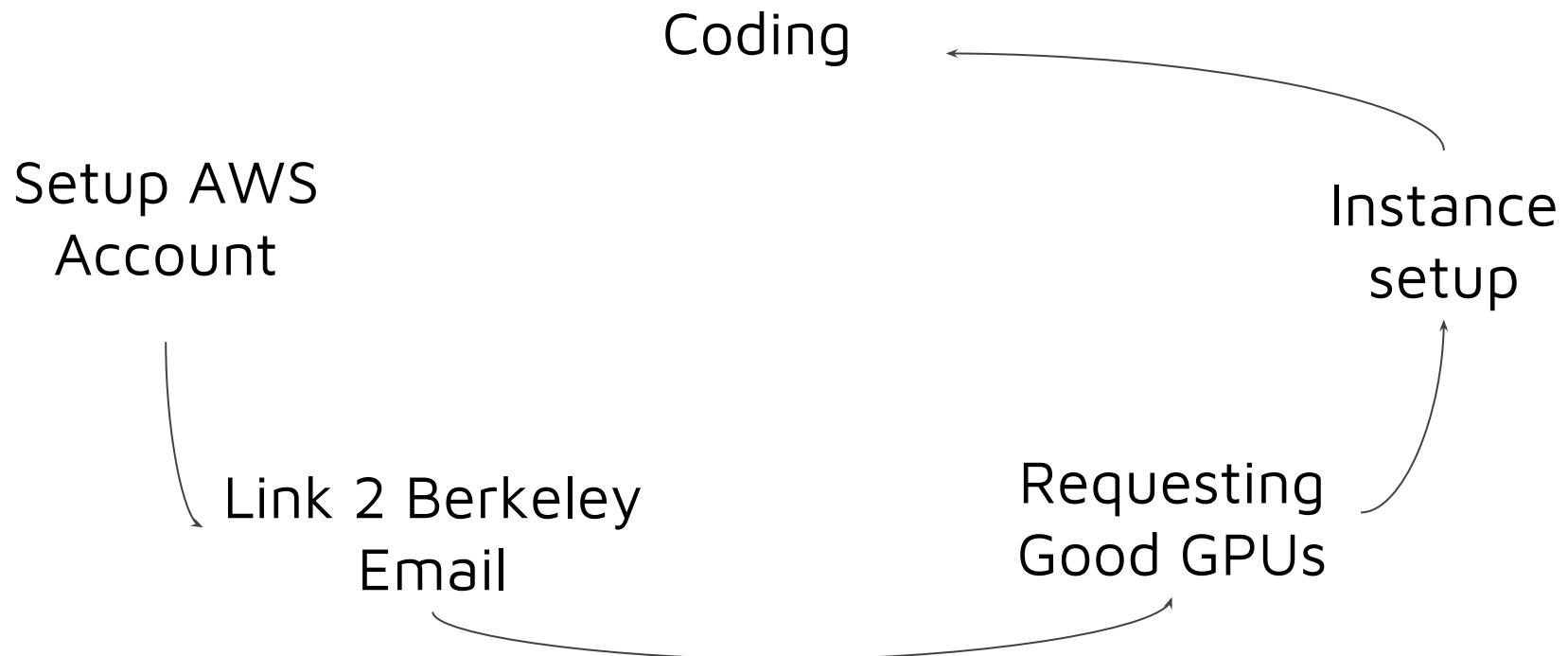
Last Note: Add extra data

Setting Up Hardware: AWS EC2 P2 Instances

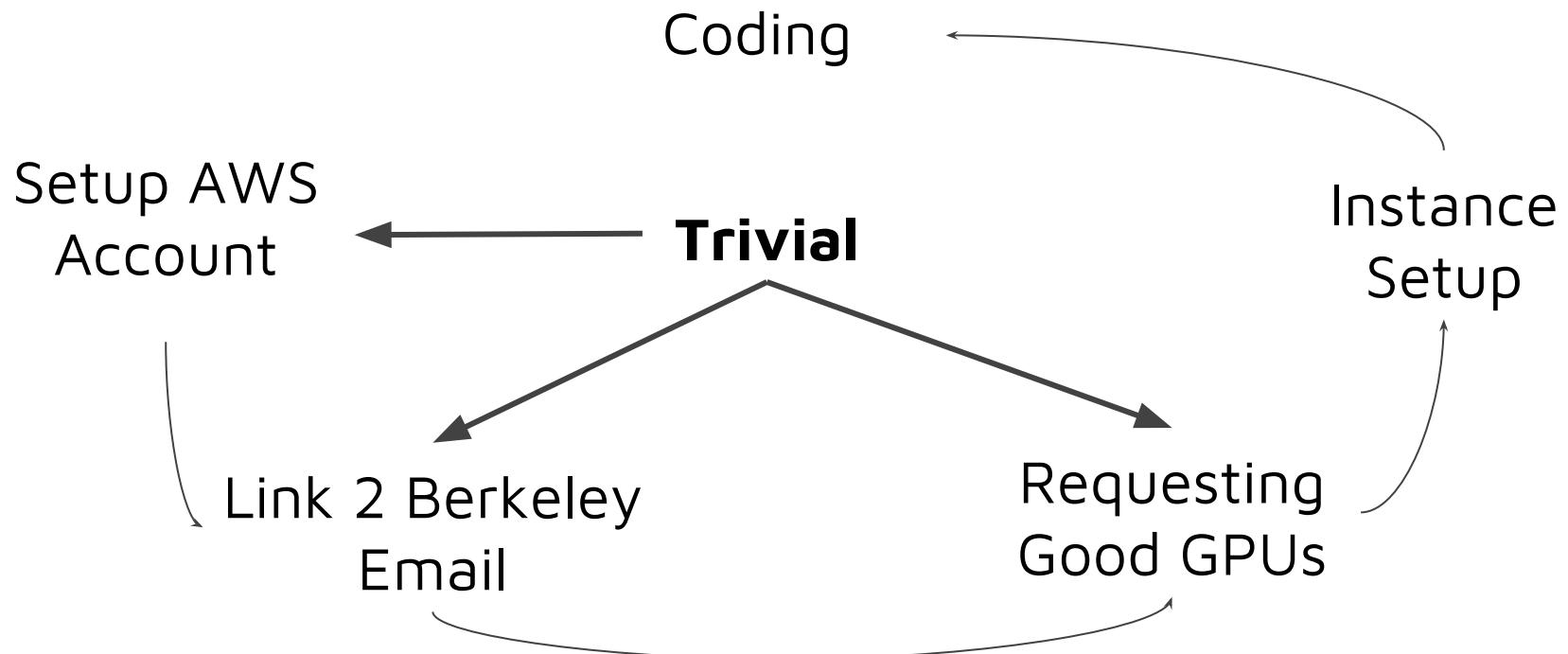
Kaggle Workflow



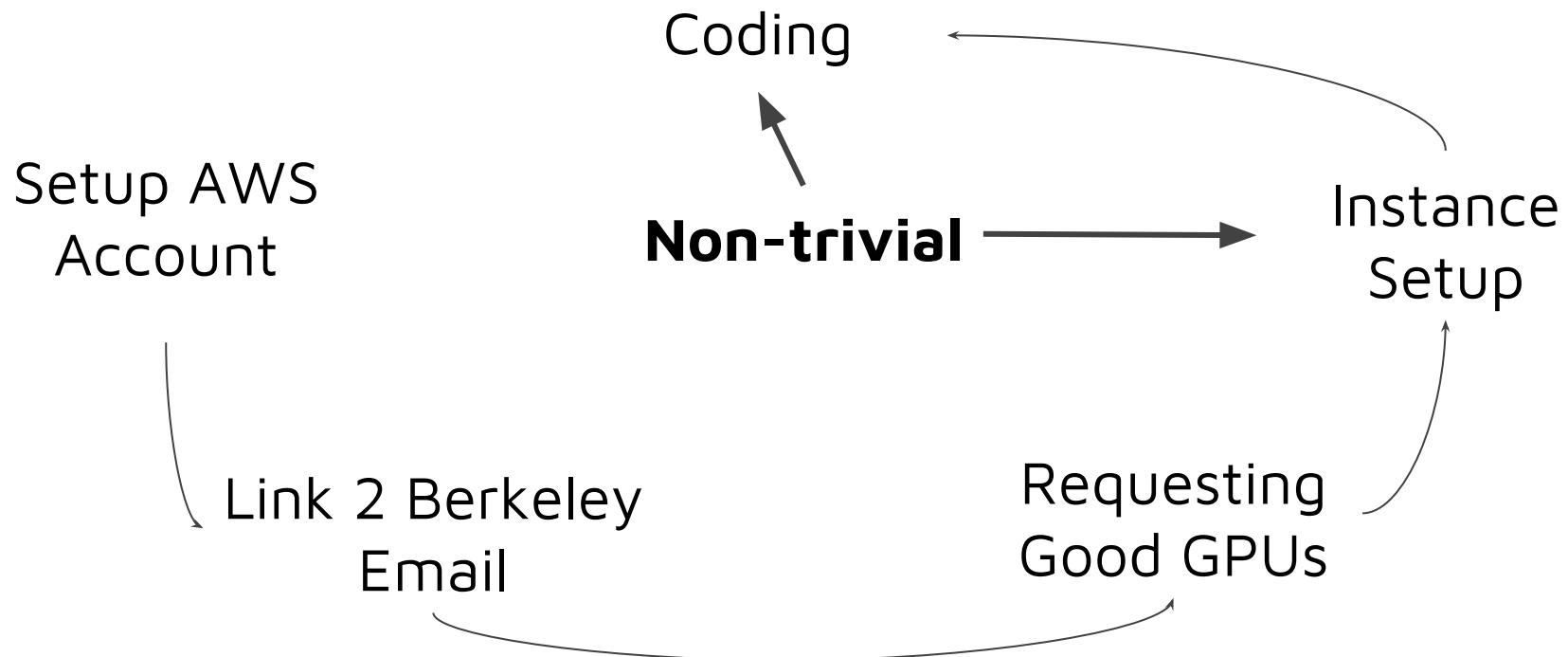
Hardware



Hardware



Hardware



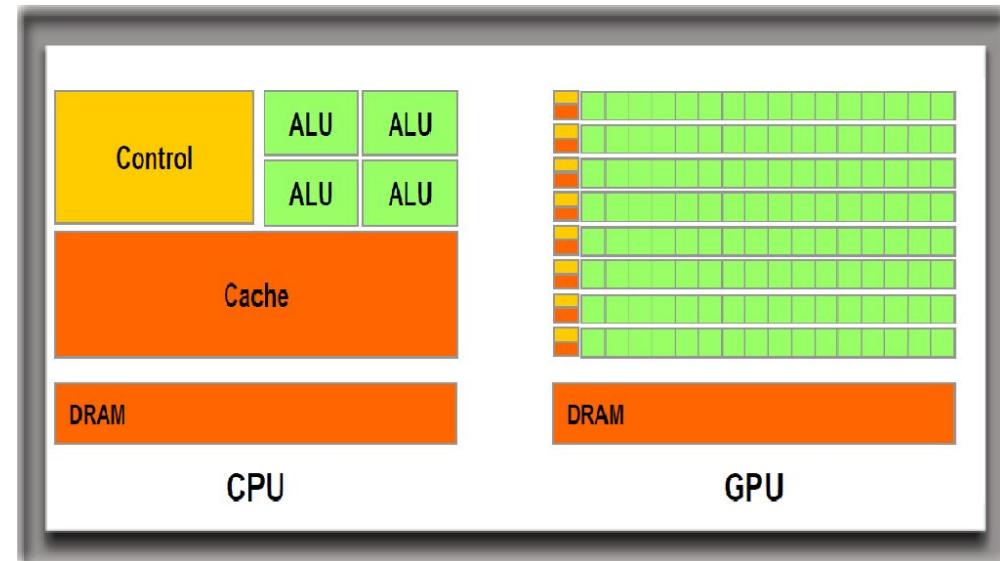
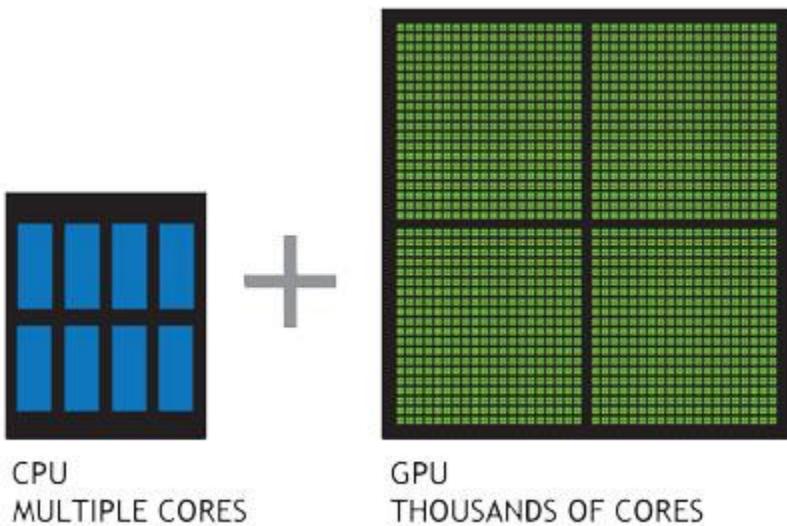


Graphics Cards

- You cannot do anything useful on big datasets without them
- 100 free credits (per year) from AWS if you create an account with Berkeley email
- Running on a GPU-accelerated instance < \$1/hr
- CDiscount (\$35k) has train files 58GB... GL processing on CPU



GPU: Graphics Processing Unit





AWS: Your Hardware Lifeline

- Amazon Web Services
- Elastic Compute Cloud (EC2)
- Provide on demand creation of servers you can ssh into and run processes on



AWS: Your Hardware Lifeline

This usually translates to

- Training models
 - Write code locally
 - Test code (mostly) locally
 - Then scp code onto server & run in special process (tmux &/ nohup)
- Data exploration
 - Run Jupyter Notebook on GPU
 - Port forward access to that notebook to your computer & run explorative code on the notebook



AWS 1: Sign up for account

AWS 2: Sign up for AWS educate & link account

The screenshot shows the AWS Educate homepage with a blue background. At the top, there's a navigation bar with links for Menu, Contact Sales, Products, Solutions, Pricing, More, English, My Account, and a yellow 'Sign In to the Console' button. Below the navigation, the AWS Educate logo is displayed, featuring the word 'aws' in white and 'educate' in a serif font with a graduation cap icon integrated into the letter 'e'. A large, bold, white text 'Teach Tomorrow's Cloud Workforce Today' is centered on the page. Below this text, a paragraph explains the initiative: 'With the increasing demand for cloud employees, AWS Educate provides an academic gateway for the next generation of IT and cloud professionals. AWS Educate is Amazon's global initiative to provide students and educators with the resources needed to accelerate cloud-related learning endeavors.' At the bottom center, a yellow call-to-action button contains the text 'Join AWS Educate Today'.

Menu  Contact Sales Products Solutions Pricing More English My Account Sign In to the Console

aws  educate

Teach Tomorrow's Cloud Workforce Today

With the increasing demand for cloud employees, AWS Educate provides an academic gateway for the next generation of IT and cloud professionals. AWS Educate is Amazon's global initiative to provide students and educators with the resources needed to accelerate cloud-related learning endeavors.

Join AWS Educate Today



AWS services

Find a service by name or feature (for example, EC2, S3 or VM, storage).



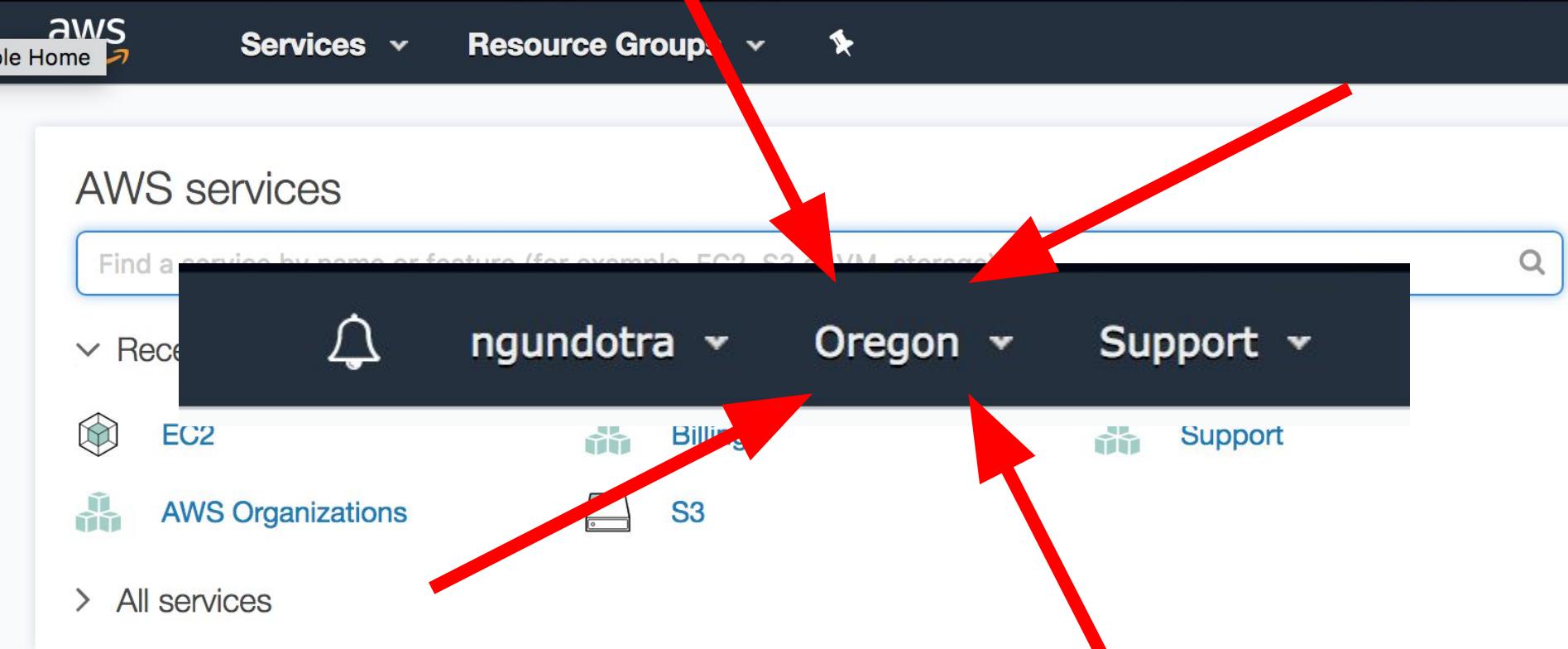
Recently visited services

[EC2](#)[Billing](#)[Support](#)[AWS Organizations](#)[S3](#)

All services



AWS 3: Requesting Instances



AWS 3: Requesting Instances



Services ▾

Resource Groups ▾



ngundotra ▾

Oregon ▾

Support ▾

EC2 Dashboard

Events

Tags

Reports

Limits

INSTANCES

Instances

Spot Requests

Reserved Instances

Scheduled Instances

Dedicated Hosts

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Snapshots

Running On-Demand m3.large instances	20	Request limit increase
Running On-Demand m3.medium instances	20	Request limit increase
Running On-Demand m3.xlarge instances	20	Request limit increase
Running On-Demand m4.10xlarge instances	1	Request limit increase
Running On-Demand m4.16xlarge instances	1	Request limit increase
Running On-Demand m4.2xlarge instances	5	Request limit increase
Running On-Demand m4.4xlarge instances	2	Request limit increase
Running On-Demand m4.large instances	20	Request limit increase
Running On-Demand m4.xlarge instances	10	Request limit increase
Running On-Demand p2.16xlarge instances	0	Request limit increase
Running On-Demand p2.8xlarge instances	0	Request limit increase
Running On-Demand p2.xlarge instances	1	Request limit increase

AWS 3: Requesting Instances

[Cancel and Exit](#)

Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.

Quick Start

◀ ▶ 1 to 35 of 35 AMIs

My AMIs
AWS Marketplace
Community AMIs

Free tier only ⓘ

 Amazon Linux Free tier eligible	Amazon Linux AMI 2017.09.1 (HVM), SSD Volume Type - ami-32d8124a The Amazon Linux AMI is an EBS-backed, AWS-supported image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Docker, PHP, MySQL, PostgreSQL, and other packages. Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	Select 64-bit
 Red Hat Free tier eligible	Red Hat Enterprise Linux 7.4 (HVM), SSD Volume Type - ami-9fa343e7 Red Hat Enterprise Linux version 7.4 (HVM), EBS General Purpose (SSD) Volume Type Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	Select 64-bit
 SUSE Linux Free tier eligible	SUSE Linux Enterprise Server 12 SP3 (HVM), SSD Volume Type - ami-2c8f5b54 SUSE Linux Enterprise Server 12 Service Pack 3 (HVM), EBS General Purpose (SSD) Volume Type. Public Cloud, Advanced Systems Management, Web and Scripting, and Legacy modules enabled. Root device type: ebs Virtualization type: hvm ENA Enabled: Yes	Select 64-bit



AWS 3: Requesting Instances



Deep Learning AMI (Ubuntu) Version 1.0 - ami-f1e73689

Select

Free tier eligible

Deep Learning AMI with Conda-based virtual environments for Apache MXNet, TensorFlow, Caffe2, PyTorch, Theano, CNTK and Keras

64-bit

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes



AWS 3: Requesting Instances

[1. Choose AMI](#)[2. Choose Instance Type](#)[3. Configure Instance](#)[4. Add Storage](#)[5. Add Tags](#)[6. Configure Security Group](#)[7. Review](#)

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by:

[All instance types](#)[Current generation](#)[Show/Hide Columns](#)

Currently selected: t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

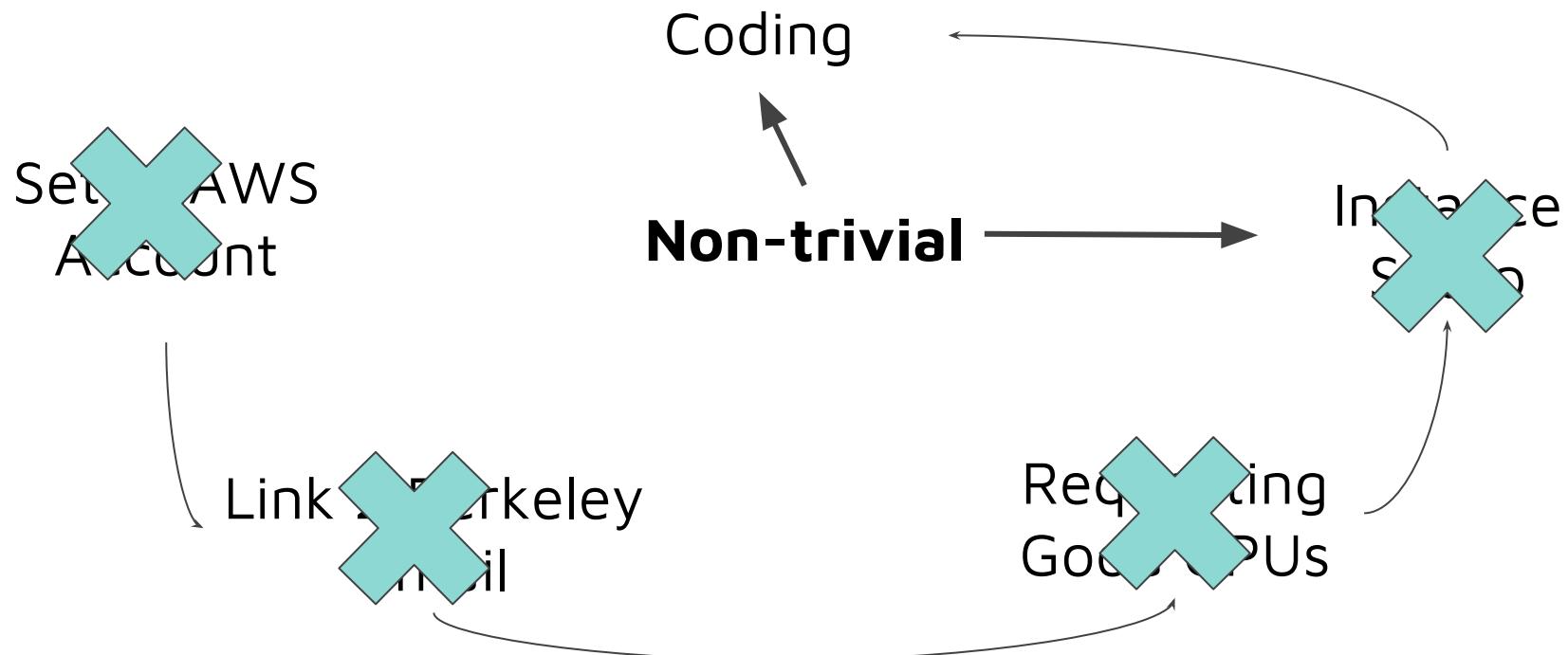
	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
<input type="checkbox"/>	General purpose	t2.nano	1	0.5	EBS only	-	Low to Moderate	Yes
<input checked="" type="checkbox"/>	General purpose	t2.micro <small>Free tier eligible</small>	1	1	EBS only	-	Low to Moderate	Yes
<input type="checkbox"/>	GPU compute	p2.xlarge	4	61	EBS only	Yes	High	Yes



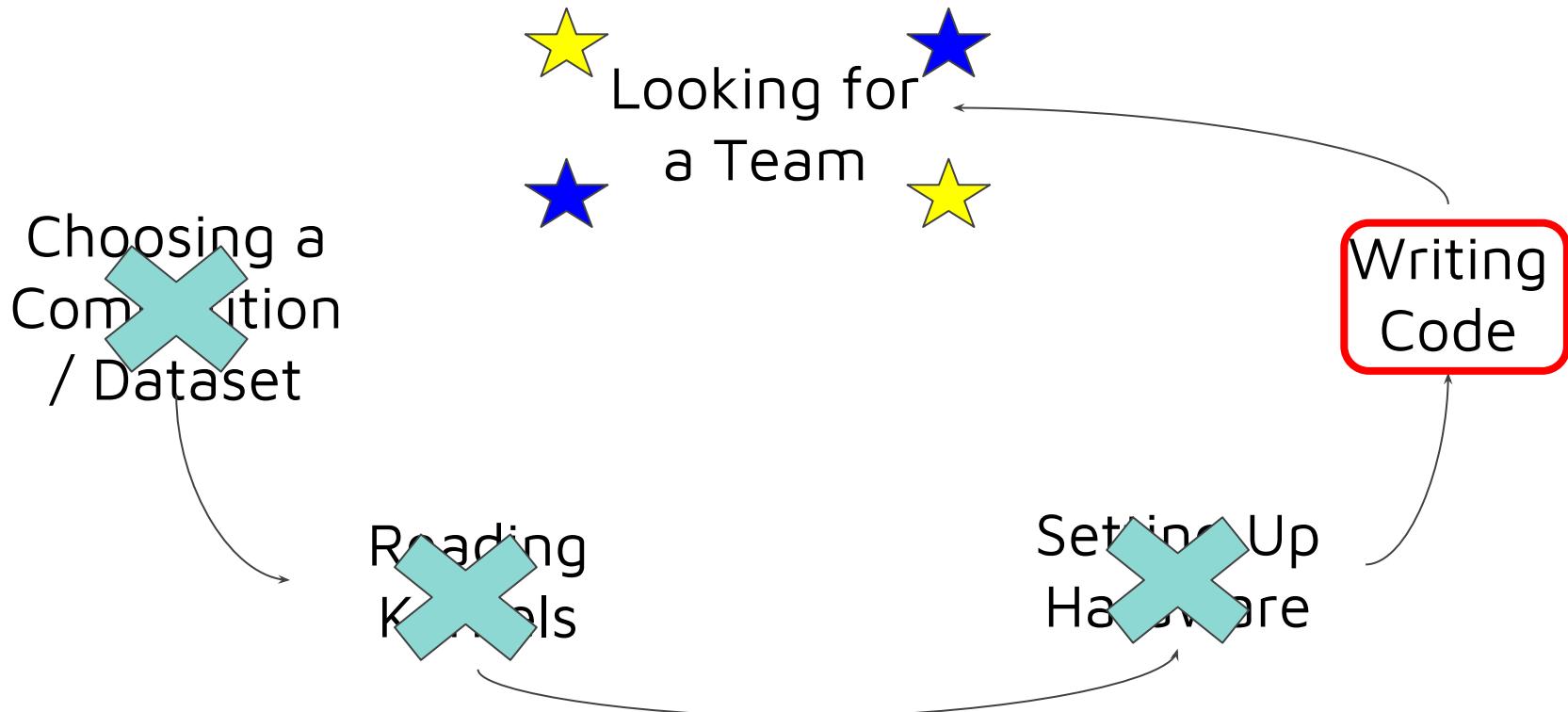
AWS 3: Requesting Instances

Momentary Pause:
Kaggle -> Hardware ->
Instance setup

Hardware



Kaggle Workflow



Kaggle: Deep
Learning/Coding :)



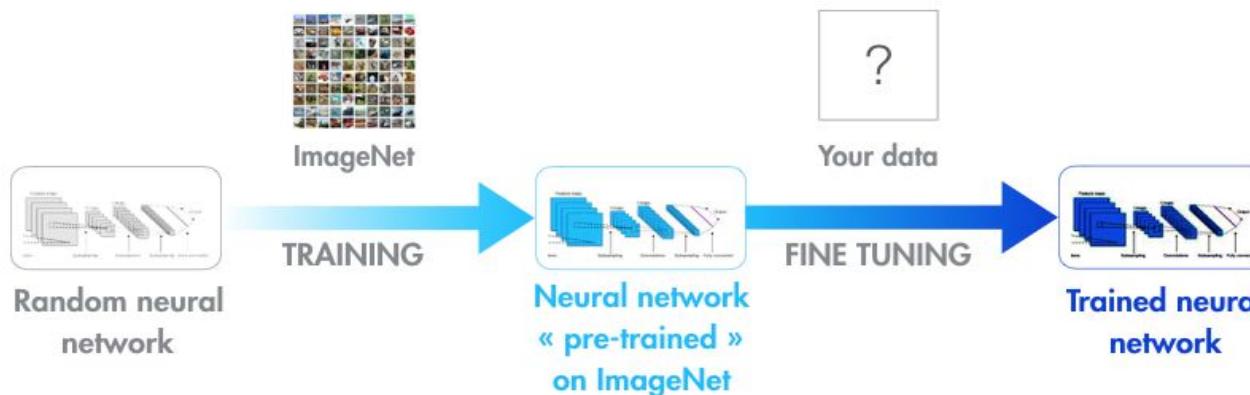
Deep Learning Approaches:

1. Transfer Learning
2. Ensemble Models

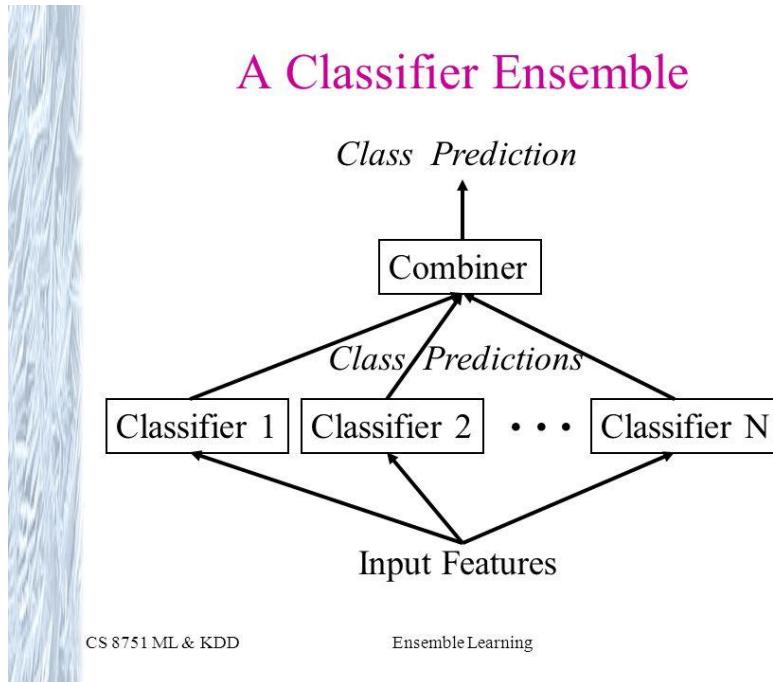


Transfer Learning (Most Popular)

TRANSFER LEARNING WITH WARM RESTART



Ensemble Learning: Mash models together





Transfer Learning (Most Popular)

[Demo]