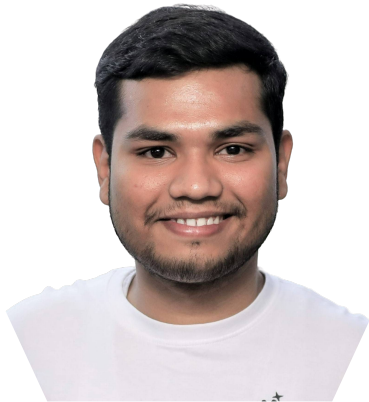


Unleashing the Data Journey

Date : 28-02-2024 | Speaker : Ayon Roy |
Event : Data Science Bootcamp | Venue : USAR, GGSIPU

Hello World!



I am **Ayon Roy**

Executive Data Scientist @ NielsenIQ

Z by HP Global Data Science Ambassador

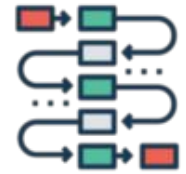
- Mentored/Judged **100+** Hackathons
- Delivered **100+** Technical Talks
- Brought **Kaggle Days Meetup** Community in India for the 1st time

If you haven't heard about me yet, you might have been living under the rocks. Wake up !!

Agenda

- Key Concepts in today's Data Science Industry
- Things to focus on while making a Data project
- A Primer to Competitive Data Science & Kaggle
- A Primer to Data Science Internships

Data Science Process



OBTAIN

SCRUB

EXPLORE

MODEL

INTERPRET

O

Gather data from relevant sources

S

Clean data to formats that machine understands

E

Find significant patterns and trends using statistical methods

M

Construct models to predict and forecast

N

Put the results into good use

Originally by Hillary Mason and Chris Wiggins

Visit - <https://ayon-roy.netlify.app>

Key Concepts

- What is Microservices & Docker
- Where Microservices & Docker fits in a Data Science process
- Using Microservices & Docker for Data Science

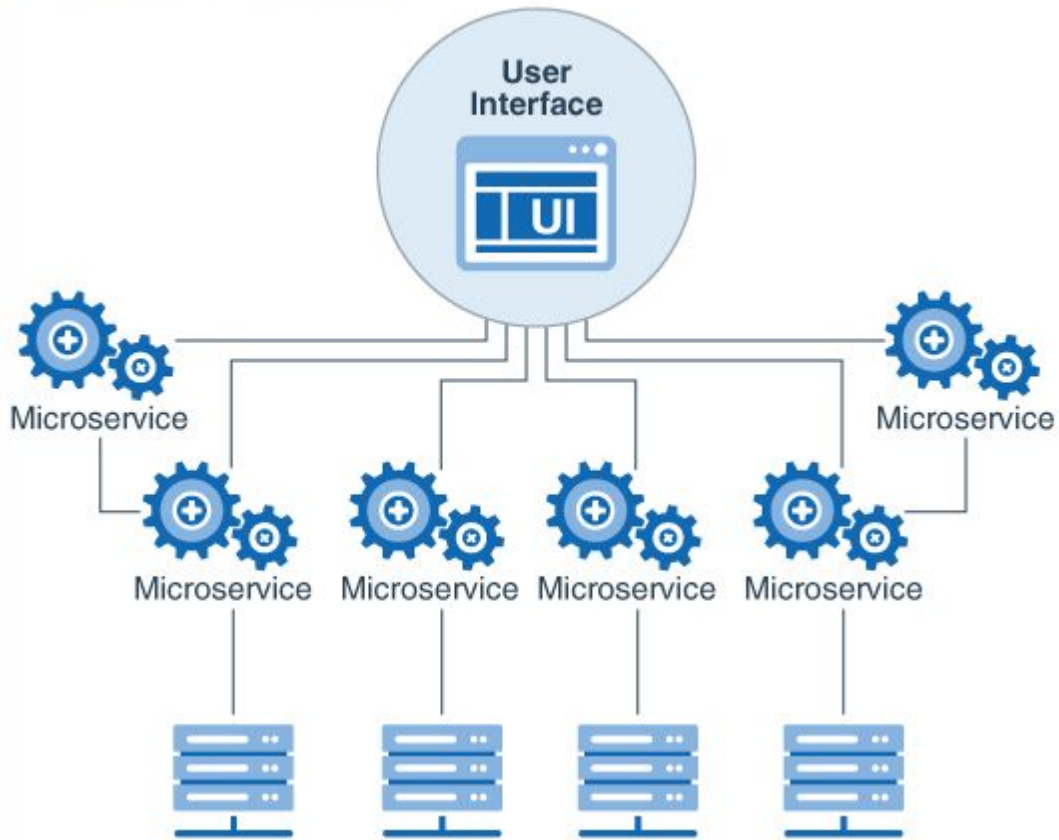
What is Microservices ?

Monolith vs

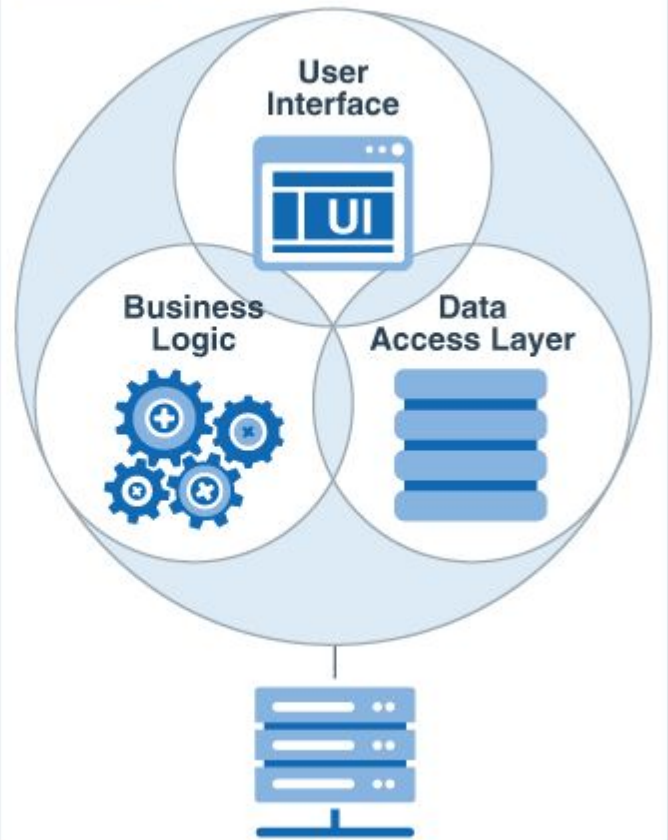
Microservices



Microservice Architecture

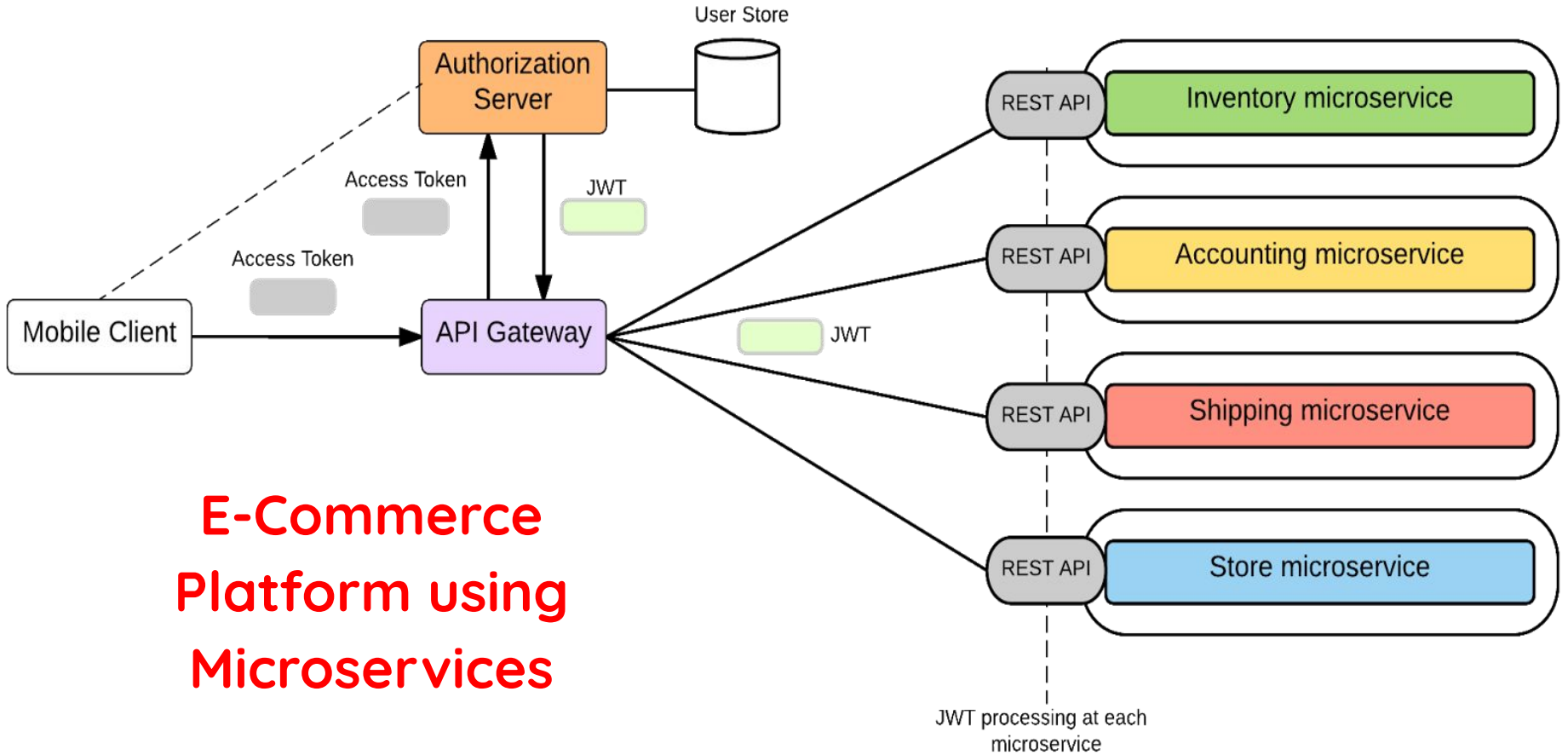


Monolithic Architecture



Microservices - also known as the microservice architecture - is **an architectural style that structures an application as a collection of services** that are :

- Highly maintainable and testable
- Loosely coupled
- Independently deployable
- Organized around business capabilities
- Owned by a small team

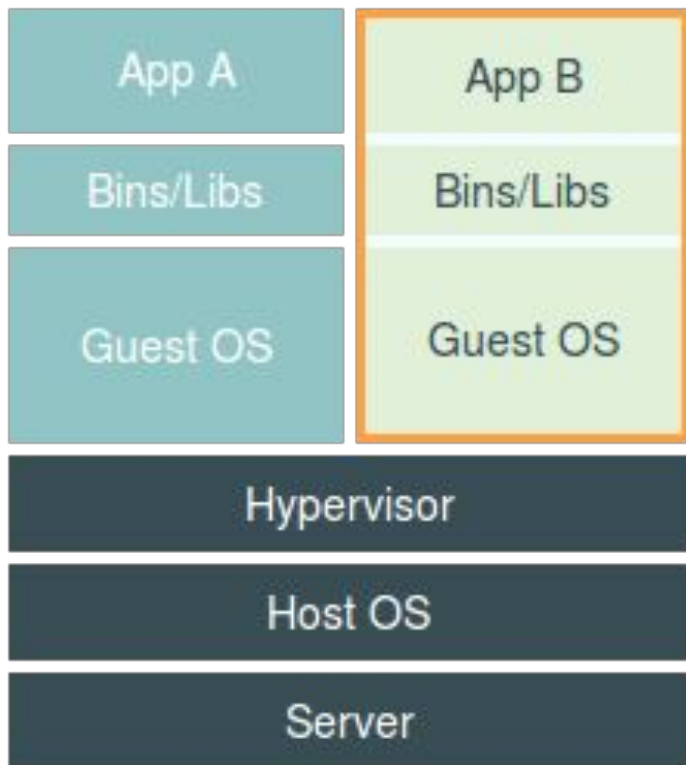


E-Commerce Platform using Microservices

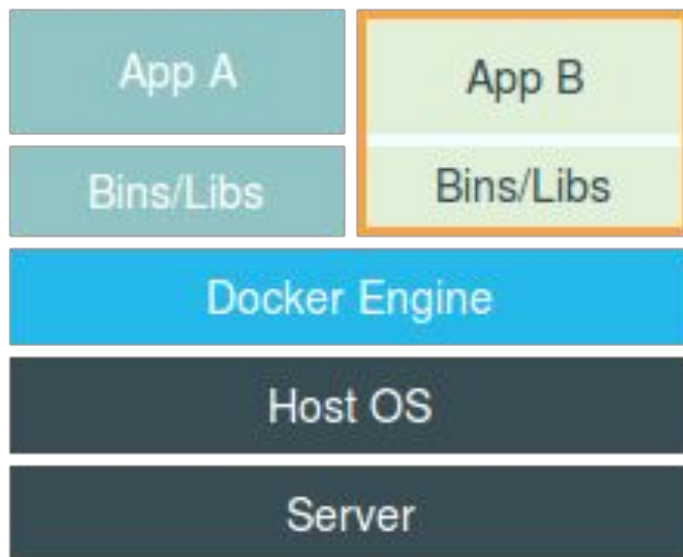
What is Docker ?

Docker is an open-source project for automating the deployment of applications as **portable, self-sufficient containers that can run anywhere on the cloud or on-premises.**

Virtual Machine



Docker

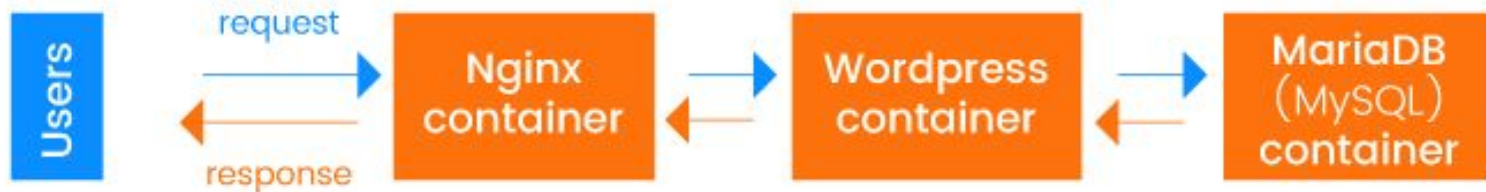


Virtual Machines	Docker
Each VM runs its own OS	All containers share the same Kernel of the host
Boot up time is in minutes	Containers instantiate in seconds
VMs snapshots are used sparingly	Images are built incrementally on top of another like layers. Lots of images/snapshots
Not effective diffs. Not version controlled	Images can be diffed and can be version controlled. Dockerhub is like GITHUB
Cannot run more than couple of VMs on an average laptop	Can run many Docker containers in a laptop.
Only one VM can be started from one set of VMX and VMDK files	Multiple Docker containers can be started from one Docker image

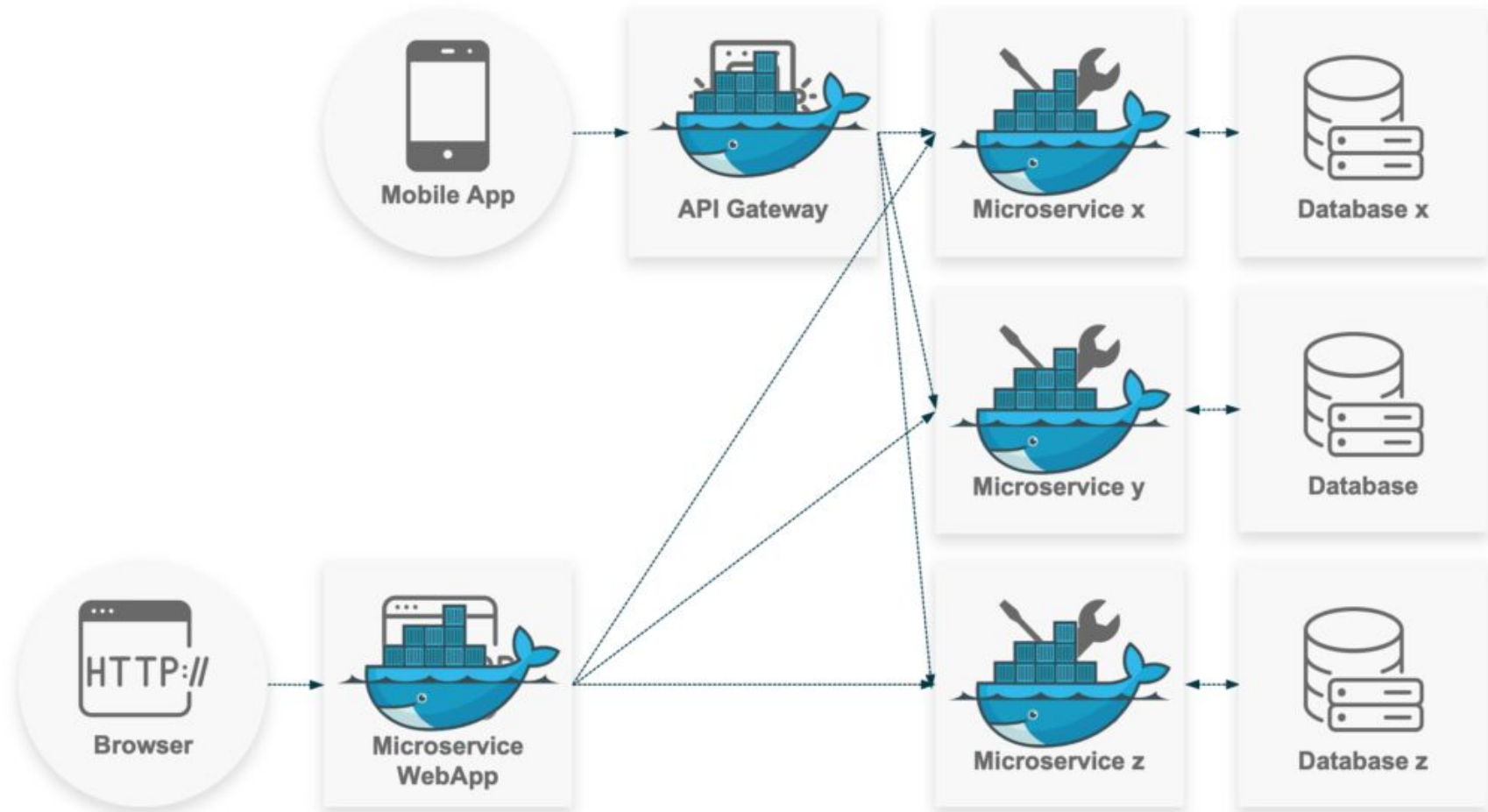
How Docker starts running?

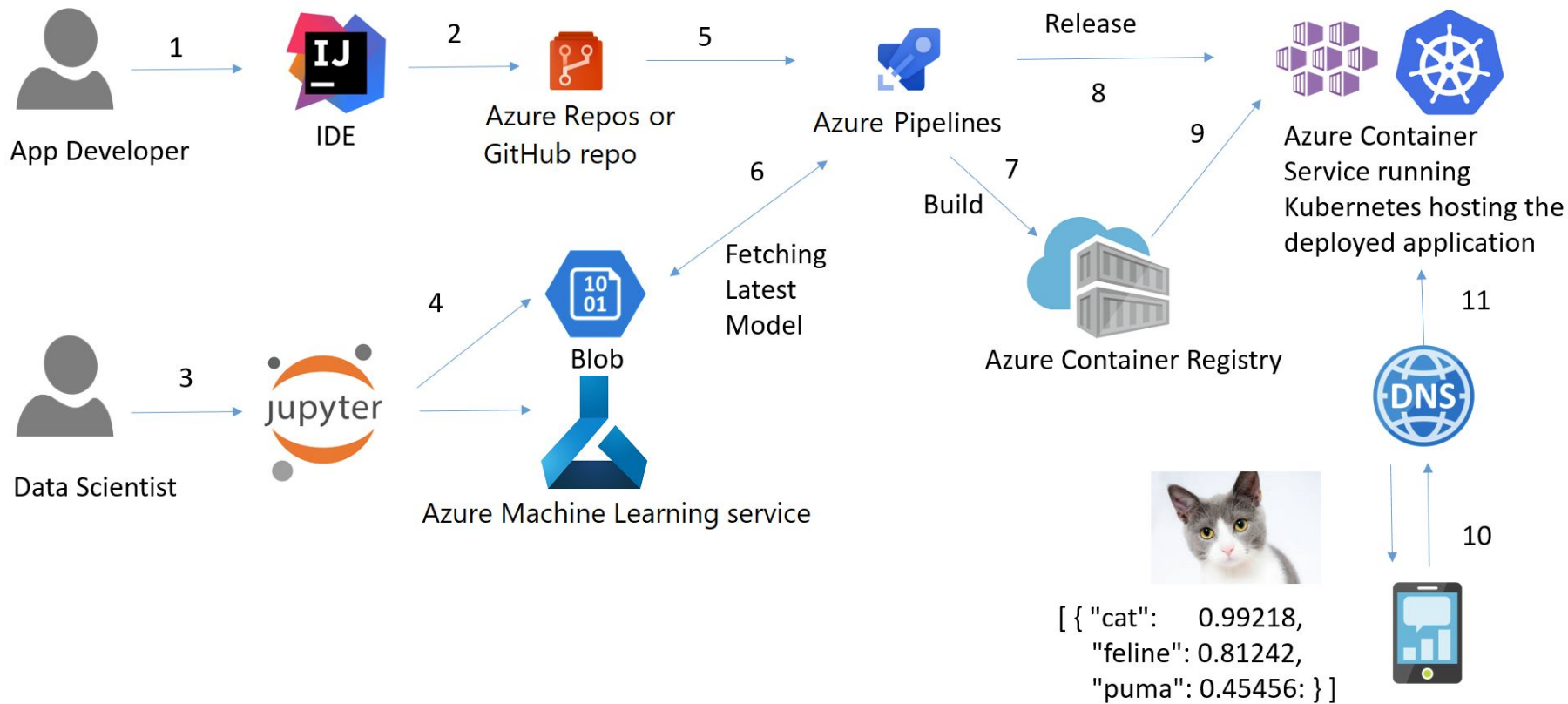


Dockerized App (microservice)

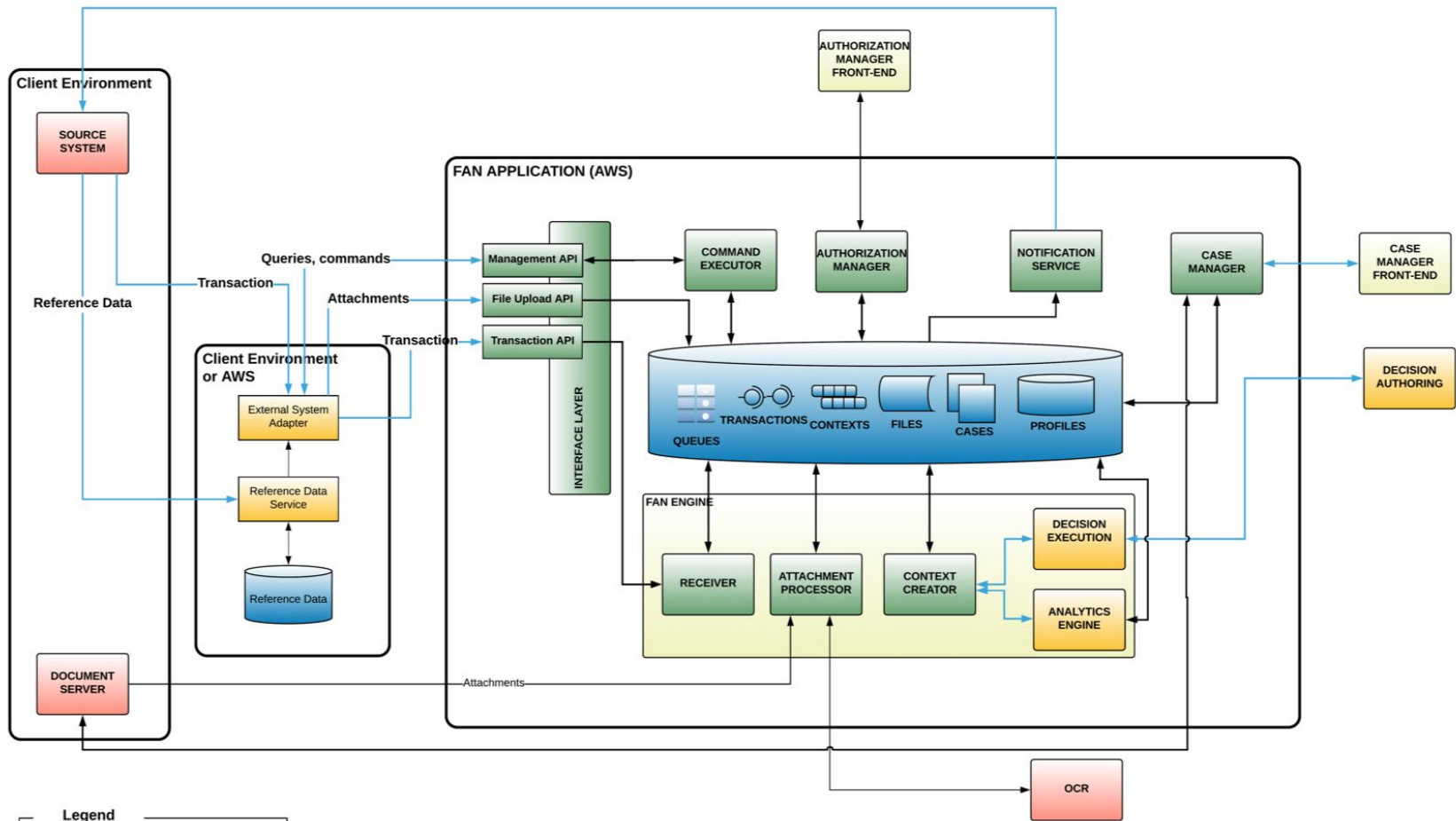


Where
Microservices & Docker
fits in a Data Science
Process?










Using Microservices for Data Science




← → ↻ 🏠 github.com/melofred/FraudDetection-Microservices 🔍 ☆ 🧑‍🤖 📄 👁 ⚙️ 👤 ⋮


 Search or jump to... / Pulls Issues Codespaces Marketplace Explore 🔔 + ▾ 👤 ▾



 **melofred / FraudDetection-Microservices**  Watch ▾ 11  Star 81  Fork 50


<> Code ! Issues 🔗 Pull requests 1 ▶ Actions 📁 Projects 📖 Wiki 🛡 Security ⋮


Octotree >

 master ▾ **About**

 **melofred** Update README.adoc ... on 18 Jan 2017 🕒 48

 ClusteringService	changes for SCDF 1.0GA	4 years ago
 Enrich-processor	clean-up	4 years ago

 **Readme**

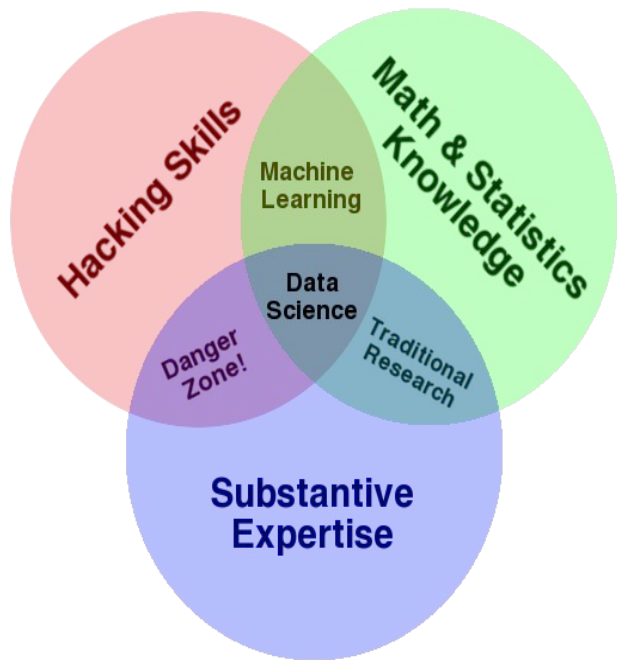
 **Apache-2.0 License**

No description, website, or topics provided.

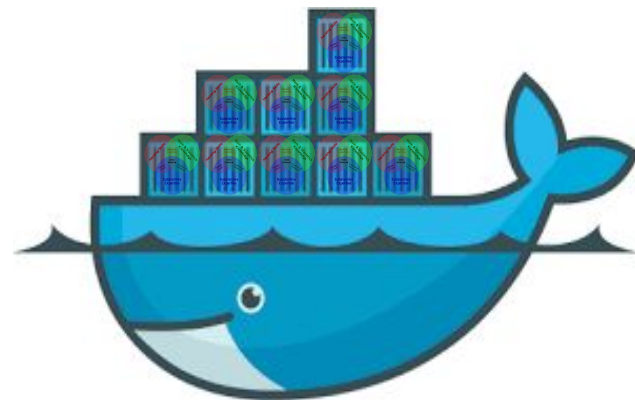
<https://github.com/melofred/FraudDetection-Microservices>

Visit - <https://ayon-roy.netlify.app>

Using Docker for Data Science



+



Here are a few examples of applications relevant to data science where you might try out with Docker:

- ***Create an ultra-portable, custom development workflow:*** Build a personal development environment in a Dockerfile, so you can access your workflow immediately on any machine with Docker installed.
- ***Create development, testing, staging, and production environments:*** Your code will run as you expect and become able to create staging environments identical to production so you know when you push to production, you're going to be OK.
- ***Reproduce your Jupyter notebook on any machine:*** Create a container that runs everything you need for your Jupyter Notebook data analysis, so you can pass it along to other researchers / colleagues and know that it will run on their machine.

Self-Contained Container

- **Problem:** Sharing results (Jupyter notebook)
- **Workflow:**
 - Create Docker image with libraries, data and notebook

Self-Contained Container: Dockerfile

```
FROM python:3.6.3-slim
```

```
LABEL maintainer="Ayon Roy <ayon-roy@outlook.com>"
```

```
WORKDIR /app
```

```
COPY . /app
```

```
RUN pip --no-cache-dir install numpy pandas seaborn sklearn jupyter
```

```
EXPOSE 8888
```

```
# Run app.py when the container launches
```

```
CMD ["jupyter", "notebook", "--ip='*'", "--port=8888", "--no-browser", "--allow-root"]
```

A few useful resources

- <https://docs.microsoft.com/en-us/dotnet/architecture/microservices/container-docker-introduction/docker-defined>
- <https://github.com/docker-for-data-science/docker-for-data-science-tutorial>
- <https://docs.docker.com/get-started/overview/>
- <https://unsupervisedpandas.com/data-science/docker-for-data-science/>
- [Docker for Data Scientists, Strata 2016, Michaelangelo D'Agostino \(YouTube Video\)](#)
- [Data Science Workflows Using Containers, by Aly Sivji \(YouTube Video\)](#)
- [A 3 Hour Docker for Data Scientists Workshop \(YouTube Video\)](#)
- <https://www.andrewmahon.info/blog/docker-compose-data-science>
- <https://towardsdatascience.com/jupyter-data-science-stack-docker-in-under-15-minutes-19d8f822bd45>
- <https://www.dataquest.io/blog/docker-data-science/>

Things to focus on while making a Data project



How to organize your Data Project?

Local Project Directory	Github Repository
<ul style="list-style-type: none">▪ Project plans/objectives▪ Project datasets▪ Project codes<ul style="list-style-type: none">○ Jupyter notebook○ R scripts○ Python scripts▪ Output files<ul style="list-style-type: none">○ Visualizations○ Tables○ Other useful outputs▪ Project report	<ul style="list-style-type: none">▪ README file▪ Project datasets▪ Project codes<ul style="list-style-type: none">○ Jupyter notebook○ R scripts○ Python scripts▪ Output files<ul style="list-style-type: none">○ Visualizations○ Tables○ Other useful outputs▪ Project report

<https://gist.github.com/ericmjl/27e50331f24db3e8f957d1fe7bbbe510>

**But why organize
your
Data Project?**

- **Organization increases productivity** as avoid wasting time searching for project files such as datasets, codes, output files, and so on.
- A well-organized project helps you to keep and **maintain a record of your ongoing and completed data science projects.**
- Completed data science projects could be **used for building future models.**
- A well-organized project **can easily be understood by other data science professionals** when shared on platforms such as Github.

What is Competitive Data Science ?

A great opportunity to

- **Sharpen your programming & analytical skills**
- **Enhance domain knowledge**
- **Learn more about practical applications of data science & machine learning algorithms**

by participating in some real world Data Science Competitions hosted by organizations on various platforms.

But why
Competitive Data Science
is gaining traction in 2024?

It's possibly due to the



Organizations are having hard time to solve so many data science problems while their data science team is busy with other projects. So hosting a data science competition on certain platform can help & is helping.

Data science competitions help organizations solve complex business problems while enabling data scientists to learn from the experience and win awards.

Organizations need to define the problem, provide data and put a prize on the challenge. Competing data scientists build and present different algorithms to be the winner.

Why should you try
Competitive Data Science
at least once?

To avoid situations like

when you have your first real-world
adult experience after graduating



And to

- Understand how to solve predictive modeling competitions efficiently
- Learn how to preprocess the data and generate new features
- Be taught advanced feature engineering techniques
- Be able to form reliable cross validation methodologies
- Gain experience in analyzing and interpreting the data
- Master the art of combining different machine learning models
- Get exposed to past (winning) solutions

How should you start your
Competitive Data Science
journey?

The only thing you need to know **Before Starting** your CDS journey

“For participating in data science competitions, you only need an urge to constantly learn and improve. Getting a good ranking will follow.”

Initial steps to start your CDS Journey

- Make sure your basics about Python & Mathematical concepts are clear enough.
- Focus on understanding core Data Science & Machine Learning algorithms
- Try to make self projects with the concepts you learned

The next steps

- Try participating in Kudos/Knowledge Competitions (Like Titanic etc.)
- Then try to learn about the approaches from other's notebooks
- Try to apply your learnings from those approaches in Featured/Prized Competitions
- Try exploring variety of techniques you can use to get better results

How to approach a Competitive Data Science Problem?

1. **Start with a very simple baseline model**
2. **Understand the problem and data to create a good validation set**
3. **Try Feature engineering**
4. **Try building a variety of models**
5. **Try stacking or blending of these results using Ensembling**



Time is a very crucial factor in any data science competition.

You should not waste your time writing the same snippets from scratch again and again in multiple competitions. Instead, focus your valuable time on doing something new and better

Where to get involved with Competitive Data Science ?

My personal suggestions

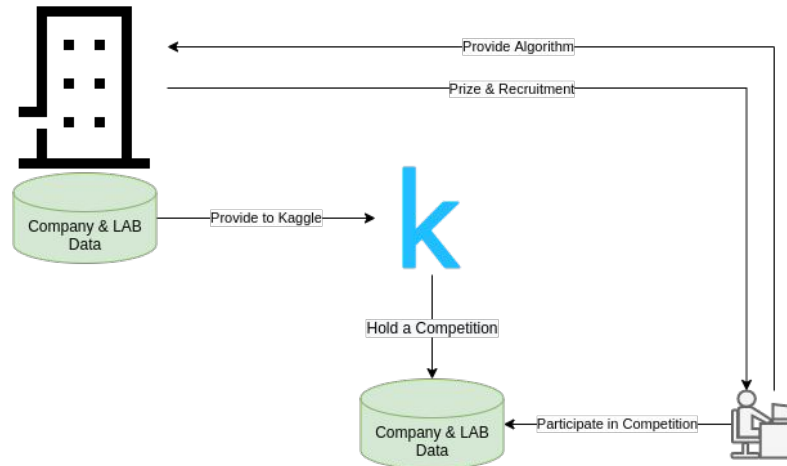
- <https://www.kaggle.com/>
- <https://www.crowdanalytix.com/community>
- <https://zindi.africa/about>
- <https://datahack.analyticsvidhya.com/>
- <https://www.crowdai.org/challenges>
- <https://tianchi.aliyun.com/competition/gameList/activeList>
- <https://www.datasciencechallenge.org/>
- <https://www.drivendata.org/>

Know a few more platforms to kick start your CDS journey [here](#)

Overview of Kaggle

What is Kaggle ?

- Kaggle is a platform for the data ecosystem worldwide revolving around skills of Data Science, AI,ML as it hosts Data Competitions regularly.
- It is common for competitions to be hosted by providing data that needs to be analyzed for the company's research challenges, key services.
- Artificial Intelligence, Machine Learning Boom has continued to increase the number of participants and was acquired by Google's parent company 'Alphabet' in 2017.
- Since the Alphabet's acquisition, Kaggle has become a critical site for data scientists and engineers, not just a platform.



Competitions

Getting Started for New Kagglers

- The Competitions shown here are for beginners.
- Especially Titanic: Machine Learning from Disaster, House Prices: Advanced Regression Techniques, Digit Recognizer. These three competitions are the most recommended and helpful competitions for new machine learners.

☰ kaggle

🏠 Home

🏆 Compete

📁 Data

📖 Notebooks

🗣 Communities

🎓 Courses

⋮ More

🔍 Search

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [Documentation](#) or learn about [InClass competitions](#).

👋 New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.



Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 16573 Teams

Knowledge

All Competitions

Active

Completed

InClass

Getting Started ▾

Default Sort ▾



Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 16573 Teams

Knowledge



House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting
Getting Started • Ongoing • 4992 Teams

Knowledge



Digit Recognizer

Learn computer vision fundamentals with the famous MNIST data
Getting Started • Ongoing • 2412 Teams

Knowledge



Natural Language Processing with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not
Getting Started • Ongoing • 1284 Teams

Knowledge



Connect X

Connect your checkers in a row before your opponent!
Getting Started • Ongoing • Simulation Competition • 445 Teams

Knowledge

Competitions

Refer to [Competitions Documentation](#).






Featured, the most common Competition

- Difficult competitions and generally commercial purposes.
- Most Kagglers participate in the competition, which has been held so far, the prize range is between \$100 and \$1,500,000.




Research

- It mainly deals with research topics and generally does not have prize money or rewards. (All the ongoing Research Competitions have prize money.)
- Instead, you can do research by discussing with less competitive and intellectually curious Kagglers.

All Competitions

Active	Completed	InClass	Featured	Default Sort
				
Riiid! Answer Correctness Prediction				
Track knowledge states of 1M+ students in the wild Featured • 3 days to go • Code Competition • 3317 Teams				\$100,000
				
Jane Street Market Prediction				
Test your model against future real market data Featured • 2 months to go • Code Competition • 1955 Teams				\$100,000
				
RANZCR CLiP - Catheter and Line Position Challenge				
Classify the presence and correct placement of tubes on chest x-rays to save lives Featured • 2 months to go • Code Competition • 391 Teams				\$50,000
				
VinBigData Chest X-ray Abnormalities Detection				
Automatically localize and classify thoracic abnormalities from chest radiographs Featured • 3 months to go • 129 Teams				\$50,000
				
Santa 2020 - The Candy Cane Contest				
May your workdays be merry and bright Featured • a month to go • Simulation Competition • 622 Teams				Prizes

All Competitions

Active	Completed	InClass	Research	Reward
				
HuBMAP - Hacking the Kidney				
Identify glomeruli in human kidney tissue images Research • 3 months to go • Code Competition • 774 Teams				\$60,000
				
Cassava Leaf Disease Classification				
Identify the type of disease present on a Cassava Leaf image Research • a month to go • Code Competition • 2235 Teams				\$18,000
				
Rainforest Connection Species Audio Detection				
Automate the detection of bird and frog species in a tropical soundscape Research • a month to go • 693 Teams				\$15,000

Playground for AI,ML, Data Science Enthusiasts

Competition is held mainly with topics that data scientists and engineers might find interesting.

Playground is not an easy task. It usually covers recent academic/technical issues and public social issues.

In some cases, the organizers may offer prize money or reward.

All Competitions

Active Completed InClass

Playground ▾ Default Sort ▾



Rock, Paper, Scissors

Shoot!

Playground • a month to go • Simulation Competition • 1368 Teams

Prizes



INGV - Volcanic Eruption Prediction

Discover hidden precursors in geophysical data to help emergency response

Playground • 2 days to go • 590 Teams

Swag



Tabular Playground Series - Jan 2021

Practice your ML regression skills on this approachable dataset!

Playground • a month to go • 219 Teams

Swag



Predict Future Sales

Final project for "How to win a data science competition" Coursera course

Playground • 2 years to go • 10033 Teams

Kudos

Kaggle Tiers

There is a Progression System in Kaggle, which is simply Kagglers Tier. This rating is a good indicator of your ability as a data scientist.

The Kaggle Tiers are divided into five levels, and conditions are also given to achieve each.

- Novice
- Contributor
- Expert
- Master
- Grandmaster

Also, as you can see in the pictures, Kaggle Tier is rated differently for Competitions, Datasets, Notebooks, and Discussion.



Novice

You've joined the community.

- Register!



Contributor

You've completed your profile, engaged with the community, and fully explored Kaggle's platform.

- Add a bio to your profile
- Add your location
- Add your occupation
- Add your organization
- SMS verify your account
- Run 1 script
- Make 1 competition or task submission
- Make 1 comment
- Cast 1 upvote



Expert

You've completed a significant body of work on Kaggle in one or more categories of expertise. Once you've reached the expert tier for a category, you will be entered into the site wide Kaggle Ranking for that category.

Competitions

- 2 bronze medals

Datasets

- 3 bronze medals

Notebooks

- 5 bronze medals

Discussions

- 50 bronze medals



Master

You've demonstrated excellence in one or more categories of expertise on Kaggle to reach this prestigious tier. Masters in the Competitions category are eligible for exclusive Master-Only competitions.

Competitions

- 1 gold medal
- 2 silver medals

Datasets

- 1 gold medal
- 4 silver medals

Notebooks

- 10 silver medals

Discussions

- 50 silver medals
- 200 medals in total



Grandmaster

You've consistently demonstrated outstanding performance in one or more categories of expertise on Kaggle to reach this pinnacle tier. You're the best of the best.

Competitions

- 5 gold medals
- Solo gold medal

Datasets

- 5 gold medals
- 5 silver medals

Notebooks

- 15 gold medals

Discussions

- 50 gold medals
- 500 medals in total

Basis for Kaggle Medals



Competition Medals

Competition medals are awarded for top competition results. The number of medals awarded per competition varies depending on the size of the competition. Note that InClass, playground, and getting started competitions do not award medals.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

* (Top 10 + 0.2%) means that an extra gold medal will be awarded for every 500 additional teams in the competition. For example, a competition with 500 teams will award gold medals to the top 11 teams and a competition with 5000 teams will award gold medals to the top 20 teams.



Dataset Medals

Dataset Medals are awarded to popular public datasets published to the site, as measured by number of upvotes. Not all upvotes count towards medals: votes by novices are excluded from medal calculation.

Bronze	5 Votes
Silver	20 Votes
Gold	50 Votes



Notebook Medals

Notebook Medals are awarded to popular notebooks, as measured by the number of upvotes a notebook receives. Not all upvotes count towards medals: self-votes, votes by novices, and votes on old posts are excluded from medal calculation.

Bronze	5 Votes
Silver	20 Votes
Gold	50 Votes



Discussion Medals

Discussion Medals are awarded to popular topics and comments posted across the site, as measured by net votes (upvotes minus downvotes). Not all upvotes count towards medals: votes by novices and votes on old posts are excluded from medal calculation.

Bronze	1 Vote
Silver	5 Votes
Gold	10 Votes

Steps to participate in a Competition

- Select one Competition in the 'Getting Started' category.
- You may take a look at other people's notebooks.
- Pick one notebook and open it in the upper right corner . Click the Copy & Edit button to copy the notebook.
- Once the copy is complete, click Save Version at the upper right corner.
 - Version Name: You can enter the name.
 - Version Type: There are two options, Quick Save or Save & Run All (Commit). Quick Save is saved, not executed, and Save & Run All (Commit) is executed.
- Click Save & Run All here and press the Save button.
- Go back to your profile and click Notebook to see the notebook you just copied.
When you click on this notebook, there is Output at the right menu.
Select Submission.csv, which can be viewed by pressing Output, and click Submit to Competition on the right.
- The screen will now be moved to the Leaderboard menu and the submitted files will be automatically scored.

After scoring, you can check your score and click Jump to your position on the leaderboard to see your ranking.

How is Kaggle Used ?

Infrastructure for data analytics

- Kaggle is web-based and provides tools for data analysis. (Notebook)
- Community with a variety of Kagglers to enable competition and cooperation.

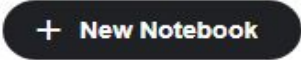
Notebook

- The programming environment for data analysis provided by Kaggle.
- A SaaS environment that runs code written on your Notebook on a server.
- It provides a programming environment, so there is no need to build a separate development environment. (No Python installation, Anaconda installation, etc.)
- It is similar to Jupyter Notebook.
- Provides 4 Core CPU + 16GB RAM by default. GPU Server provides 2Core CPU + GPU + 13GB RAM. Provided free of charge, and GPU can be used for 30 hours a week.

What can you do with Notebook?

- Programming for data analysis is the primary purpose, and programs created to run on the Kaggle server.
- Submit to Competition or share Notebook with Kaggle. Some of the Notebooks are shared only for training or skills.
- Use Code Cell and Markdown Cell to write codes, and descriptions of the code, text, image, etc.

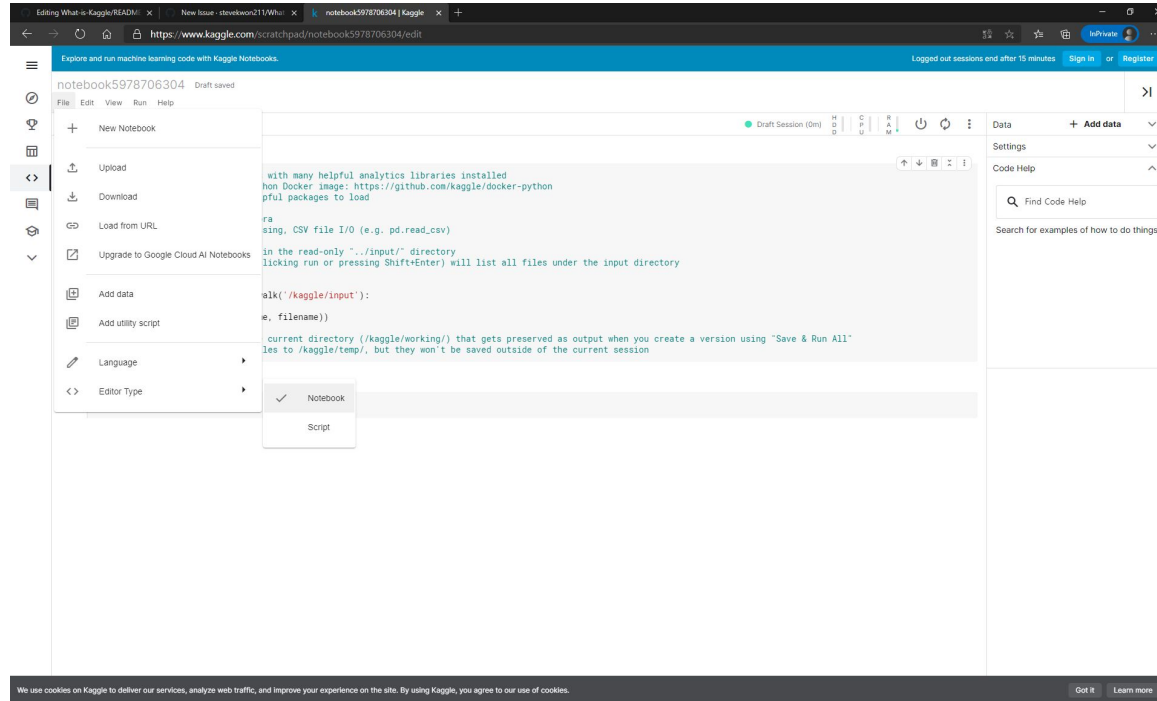
Create & Use Notebook



+ New Notebook

- Go to the Notebook menu and look in the upper right corner. There's a button like this. Click it.
- Kaggle Notebook has two types: Script and Notebook.
 - Script is a method of writing and executing code in a commonly used code editor.
- Notebook is an interactive development environment similar to Jupyter Notebook. The characteristic is that you can divide the cells and execute only the code you want.
- Press File in the upper left corner and hover your cursor over Edit Type to select the type. In addition, you can choose between Python and R in Language.

A Kaggle Notebook



The screenshot displays a Kaggle Notebook interface in a web browser. The browser address bar shows the URL `https://www.kaggle.com/scratchpad/notebook5978706304/edit`. The notebook title is `notebook5978706304` and it is in a "Draft saved" state. The left sidebar contains a menu with options: New Notebook, Upload, Download, Load from URL, Upgrade to Google Cloud AI Notebooks, Add data, Add utility script, Language, and Editor Type. The Editor Type dropdown is open, showing "Notebook" (selected) and "Script". The main code cell contains the following Python code:

```
with many helpful analytics libraries installed
hon Docker image: https://github.com/kaggle/docker-python
ptul packages to load

ra
sing, CSV file I/O (e.g. pd.read_csv)

in the read-only "../input/" directory
loking run or pressing Shift+Enter) will list all files under the input directory

alk('../kaggle/input'):

e, filename))

urrent directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
les to /kaggle/temp/, but they won't be saved outside of the current session
```

The right sidebar includes a "Data" section with an "Add data" button, "Settings", "Code Help", and a search bar for "Find Code Help". At the bottom of the page, there is a cookie consent banner that reads: "We use cookies on Kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies." with "Got it" and "Learn more" buttons.

Various Settings for Notebook

- Set Public & Private
 - Notebook can be released for sharing with other Kagglers. But if you don't want to share, or when you work as a team, you can make settings such as Private or Shared to a specific user.
 - Press the Share button in the upper right corner to open a window for public or private setting.
 - If Privacy is set to Public, it will be released with Apache 2.0 License.
 - Use Collaborators to add users as collaborators.


- Settings
 - Language : You can set the programming language to use Python and R.
 - Environment : The Docker image can be set. Original sets up the development environment when creating Notebook and Latest Available uses the latest development environment provided by Kaggle.
 - Accelerator : Whether to use GPU or TPU can be set.
 - GPU/TPU Quota : Show time and usage of GPU and TPU
 - Internet : You can set whether or not to connect to the Internet.
You can install certain packages by setting Internet to On. Google accounts also allow you to use BigQuery, Cloud Storage, and AutoML services from GCP (Google Cloud Platform).

Using Data in Notebook

Kaggle Notebook is available not only in Competition Data but also in a variety of Dataset shared.

In this case, a separate file must be set up for use in Notebook.

- i. How to create a new Notebook

o Go to the Dataset you want to use,  and press New Notebook to set the file automatically.


- ii. How to add to an existing Notebook

o To add new data to your existing Notebook, first access your Notebook.

Then click the Data  + Add data button in the upper right corner.

Then a window appears where you search for the desired Dataset and press Add after you choose Dataset.

- iii. How to upload yourself

o If you go into the Data menu and look in the upper right corner, click on the  New Dataset button.

Then enter a name for Enter Dataset Title and click Select Files to Upload to upload the file. (Compressed file types such as zip or tar.gz are also possible.)

Finally, press Create to upload Dataset. You can import the uploaded Dataset using the i or ii method.

- iv. How to use output data from another Notebook

o If you follow ii method, a window will appear, where you can click on the Kernel Output Files tab to use the output data from another Notebook

Competitions & Notebooks

What else can the Notebook be used for besides data analysis Competition?

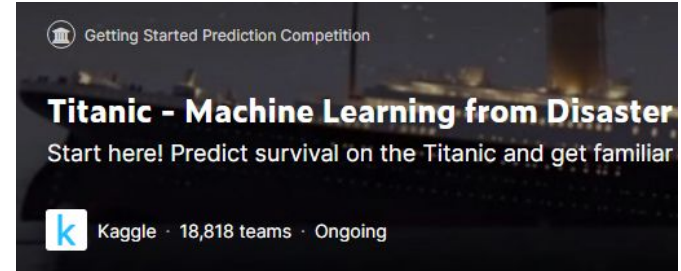
- In general, if the goal is to win a prize, Notebook will be shared(Public) after Competition is finished. However, there is also an environment in which we can discuss with Kagglers even when Competition is in progress.

How to handle Data File to use in Competition Notebook?

- When performing Competition, the Data tab is located in the upper right corner of the Notebook. There are three types of files you can click on, each of which is described as follows.
 - train.csv : Learning data with correct answer label.
 - test.csv : Data for testing without the correct answer label.
 - Sample_submission.csv : Examples of data for submission

View the Data menu in Competition to see what data each file contains.

- For example, let's look at the Titanic - Machine Learning from Disaster.

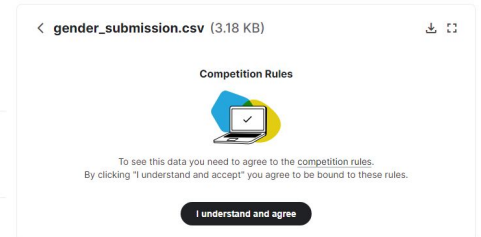
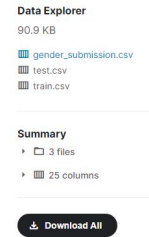


[Overview](#) [Data](#) [Notebooks](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Visit - <https://ayon-roy.netlify.app>

Competitions & Notebooks

- Let's use these files to create and submit a csv file for model creation and submission.
(The same is explained in [4. Participate in the Competition.](#))
 - Click Save Version in the upper right corner of the Notebook screen. (If the code is not executed, click Save & Run All (Commit).
 - In Save & Run All (Commit), Commit is the same meaning as Git Commit in Github, which I am currently working on.
Therefore, Kaggle Notebook can refer to the version of the source code previously written.
- Now return to your profile and click Notebook to see the notebook you just saved. When you click on this notebook, there is Output in the right menu. Select Submission.csv that you can view by pressing Output menu and click Submit to Competition on the right.
- The screen will now be moved to the Leaderboard menu and the submitted files will be automatically scored.
After scoring, you can check your score and click Jump to your position on the leaderboard to see your ranking.



Competition Progress Flow

Baseline implementing the general-purpose algorithm

- First, you start analyzing the data, you get the output data through a general-purpose algorithm.
- Develop machine learning models in earnest and compare output data and results from general-purpose algorithms.
- If the comparison results in a worse result than the general-purpose algorithm, you can assume that the model has a problem.

Data Analysis Notebook

- This refers to Notebook that analyzes Competition data and shows visualization.
- Focus on identifying correlations, rules, and structure between the analyzed data without creating data to submit. We also look for independent variables that fit well with dependent variable.
- If you have less Competition experience, it would be a good start to build knowledge and insight by looking at data analyzed by other Kagglers.

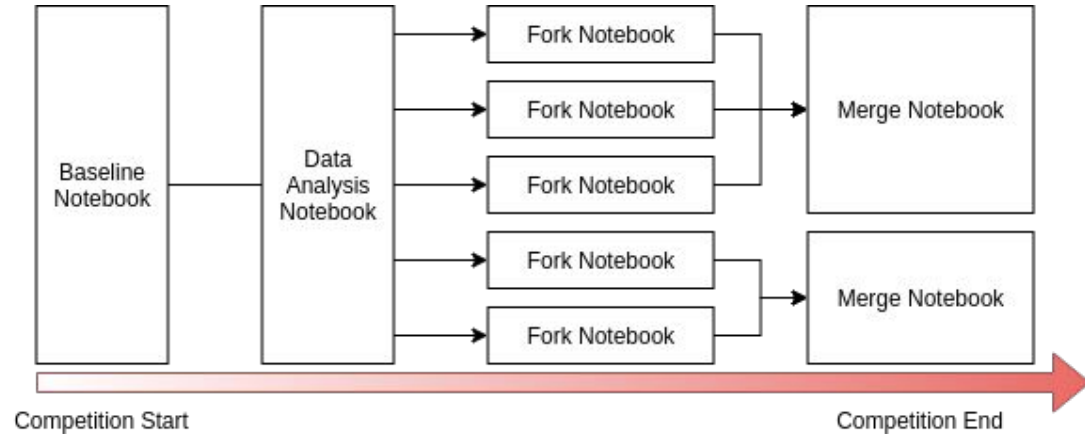
Competition Progress Flow (Contd.)

Fork Notebook

- For those who are new to machine learning and Kaggle, one way is to fork out a notebook that is open without data analysis or model development yourself.
- Fork means to copy a version of the source code.
- On the top right of the Notebook you'd like to fork press button to copy.

Merge, Blending, Stacking, Ensemble Notebook

- Notebook with words such as Merge, Blending, Stacking, and Ensemble.
- As the name suggests, it means Notebook combining several Notebooks.



Exploring Parkinson's Disease Progression Prediction Dataset on Kaggle

Detailing the Competition

Featured Code Competition

AMP®-Parkinson's Disease Progression Prediction

Use protein and peptide data measurements from Parkinson's Disease patients to predict progression of the disease.

\$60,000
Prize Money

AMP AMP®-PD · 1,805 teams · 4 months ago


[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#) [...](#)

Overview

Start
Feb 17, 2023

Close
May 19, 2023

Merger & Entry

Competition Host
AMP®-PD 

Prizes & Awards
\$60,000
Awards Points & Medals

Participation
2,197 Competitors
1,805 Teams
40,764 Entries

Description

[🔗](#) [^](#)

Detailing the Competition

We're having a contest to try and figure out how Parkinson's disease gets worse over time. Parkinson's disease is a sickness that makes it hard for people to move and think clearly. Right now, there's no cure for it. We want to understand it better, so we can find a way to slow it down or even stop it.

We think that certain tiny parts in our bodies, called proteins and peptides, might have something to do with how Parkinson's disease works. We have a lot of information about this from over 10,000 people with Parkinson's disease. But we still haven't found clear signs or cures.

The group organizing this contest is called the Accelerating Medicines Partnership® Parkinson's Disease, and they're a team of people from different places like the government, companies, and groups that want to help. They've gathered a ton of information about Parkinson's disease to try and find important clues.

If you join this contest and figure out some important information, **it could be a big step towards finding a way to help people with Parkinson's disease.** This could make life much better for them and also save a lot of money on medical care.

Competition's Context

Parkinson's disease (PD) is a disabling brain disorder that affects movements, cognition, sleep, and other normal functions. Unfortunately, there is no current cure—and the disease worsens over time. It's estimated that by 2037, 1.6 million people in the U.S. will have Parkinson's disease, at an economic cost approaching \$80 billion. Research indicates that protein or peptide abnormalities play a key role in the onset and worsening of this disease.

Gaining a better understanding of this—with the help of data science—could provide important clues for the development of new pharmacotherapies to slow the progression or cure Parkinson's disease.

Current efforts have resulted in complex clinical and neurobiological data on over 10,000 subjects for broad sharing with the research community. A number of important findings have been published using this data, but clear biomarkers or cures are still lacking.

Competition host, the Accelerating Medicines Partnership® Parkinson's Disease (AMP®PD), is a public-private partnership between government, industry, and nonprofits that is managed through the Foundation of the National Institutes of Health (FNIH). The Partnership created the AMP PD Knowledge Platform, which includes a deep molecular characterization and longitudinal clinical profiling of Parkinson's disease patients, with the goal of identifying and validating diagnostic, prognostic, and/or disease progression biomarkers for Parkinson's disease.

Your work could help in the search for a cure for Parkinson's disease, which would alleviate the substantial suffering and medical care costs of patients with this disease.

Competition's Goal

The goal of this competition is to predict MDS-UPDR scores, which measure progression in patients with Parkinson's disease.

The Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is a comprehensive assessment of both motor and non-motor symptoms associated with Parkinson's.

You will develop a model trained on data of protein and peptide levels over time in subjects with Parkinson's disease versus normal age-matched control subjects.

Your work could help provide important breakthrough information about which molecules change as Parkinson's disease progresses.

Data Detailing

UPDRS is a rating instrument used to measure the the severity and progression of Parkinson's disease in patients. When a patient visits the clinic, the clinic will record how the patient scored on 4 parts of UPDRS test. This data can be found in [train_clinical](#). The ratings for the the first 4 segments of UPDRS are available as [updrs_1](#), [updrs_2](#), [updrs_3](#) and [updrs_4](#) in [train_clinical](#). Our goal is to train a model to predict these UPDRS ratings.

The clinic will also record the patient's **NPX**(Normalized Protein eXpression) value for all the proteins relevant to Parkinson's disease during each visit. **NPX** is nothing but the value representing the protein concentration in shells. This data is available in the [train_proteins](#).

Proteins are long molecules made up of multiple peptides. The clinic will record the **Peptide Abundance** of each peptide in proteins relevant to Parkinson's disease. It shows the peptide concentration, similar to NPX for proteins. This data can be found in the [train_peptides](#).

Files

train_peptides.csv Mass spectrometry data at the peptide level. Peptides are the component subunits of proteins.

- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The [UniProt ID code](#) for the associated protein. There are often several peptides per protein.
- `Peptide` - The sequence of amino acids included in the peptide. See [this table](#) for the relevant codes. Some rare annotations may not be included in the table. The test set may include peptides not found in the train set.
- `PeptideAbundance` - The frequency of the amino acid in the sample.

train_proteins.csv Protein expression frequencies aggregated from the peptide level data.

- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `UniProt` - The [UniProt ID code](#) for the associated protein. There are often several peptides per protein. The test set may include proteins not found in the train set.
- `NPX` - Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.

train_clinical_data.csv

- `visit_id` - ID code for the visit.
- `visit_month` - The month of the visit, relative to the first visit by the patient.
- `patient_id` - An ID code for the patient.
- `updrs_[1-4]` - The patient's score for part N of the [Unified Parkinson's Disease Rating Scale](#). Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
- `upd23b_clinical_state_on_medication` - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function). These medications wear off fairly quickly (on the order of one day) so it's common for patients to take the motor function exam twice in a single month, both with and without medication.

Model should be a **Classification or Regression Model ?**

This competition involves predicting scores related to Parkinson's disease, specifically the MDS-UPDR scores, which measure disease progression.

Since the goal is to predict numerical scores (e.g., MDS-UPDR scores), it calls for a regression model. This model will estimate how the disease progresses in patients based on the provided data about proteins and peptides.

The logic behind training the model

Aim is to train a computer model to predict certain scores related to Parkinson's disease. The scores are called "updrs_1," "updrs_2," "updrs_3," and "updrs_4." These scores help doctors understand how the disease is affecting a person.

Understanding the code step by step

- It starts by listing the scores we want to predict: updrs_1, updrs_2, updrs_3, and updrs_4.
- Then, it goes through each of these scores one by one.
- For each score, it combines information from two datasets (pro_pep_df and train_clinical) using a common column called "visit_id."
- It removes any rows where the score we're interested in is missing.
- It makes a list of features we'll use to predict the score. This list is based on a previous list of features.
- It splits the data into two parts: one for training the model and one for testing how well it works.
- It prepares the data in a format that the computer model can understand.

Logic [Continued]

- It creates a type of model called a Random Forest, which is good for this type of prediction.
- It trains the model on the training data.
- It saves the trained model.
- It checks how well the model did on the testing data and records a value called Mean Squared Error (MSE), which tells us how close the model's predictions were to the actual scores.
- It uses the trained model to make predictions on the testing data.
- It calculates a different value called Symmetric Mean Absolute Percentage Error (sMAPE) to see how well the model did.
- This code repeats this process for each of the four scores (updrs_1, updrs_2, updrs_3, updrs_4).

References

- <https://github.com/stevekwon211/Hello-Kaggle-Guide>
- [GitHub - drakearch/kaggle-courses: Kaggle courses and tutorials to get you started in the Data Science world.](#)
- [Kaggle presentation \(slideshare.net\)](#)
- [Complete Solution + PPT Summary Slides | Kaggle](#)
- [Winning solutions of kaggle competitions](#)
- Demo : [Parkinson's Disease Progression Prediction w TFDF | Kaggle](#)

Be a part of Communities like

1. Kaggle Days Meetup Delhi NCR
2. Women in Machine Learning & Data Science
3. GDG, WTM, ODSC & a lot more....

Be a part of as many hackathons as you can

who wants to miss networking, free food & swags alongside unlimited learning

How to get involved more with Competitive Data Science ?

1. **Do such courses where the skills learnt in them can be used in Competitions.**
2. **Publish your competition research**, approaches on the forum & do write about the things that you want to share with others via blog etc.
3. **Participate in Discussion forums**, share your knowledge through answering questions & asking genuine questions.
4. **Make notebooks & share them along with great EDA, feature engineering etc** so that others can learn from it.
5. **Try to reproduce interesting kernels.**
6. **Be consistent** in whatever you are trying to share with the CDS community.

Is

Competitive Data Science

**everything what the industry
requires?**

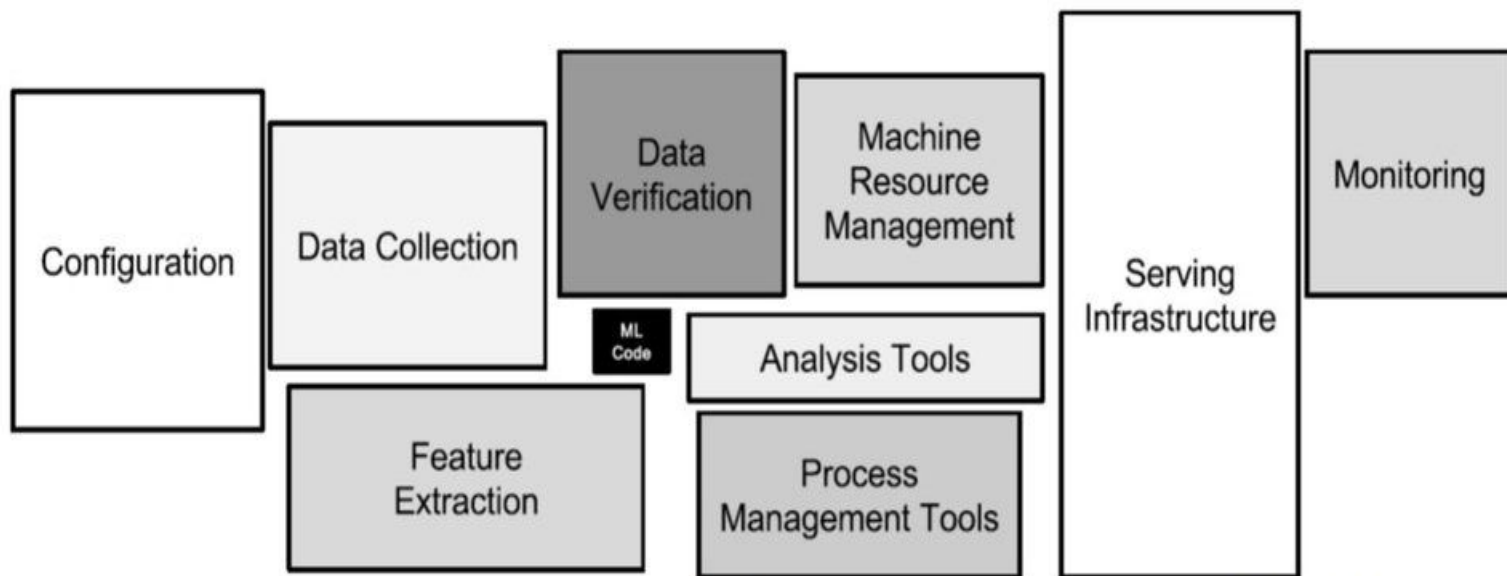


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

View the Google's Research Paper [here](#)

A few important pointers to keep in mind

1. **Focus on understanding what business use case you are trying to solve** before applying Data Science, Machine Learning.
2. **Focus on Communication Skills to convey the result** of your Data Science concepts to the business stakeholders.
3. **Focus on DevOps** to make your models production ready.
4. **Focus on networking & showcasing your work** to the community.

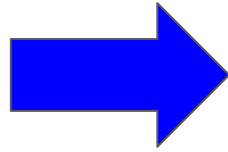
& be prepared to rock the industry

Take your baby steps
with
Internships

The **What** of Data Science Internships

What are the requirements of a decent Data Science Internship Opportunity ?

Let's divide the variety of Data Science Internships out there in the industry



Requirements for **Entry Level Internships**

- Ideal for Second-Third Year Students of a 4 year Undergraduate course
- Basic Level of Python, OOPs, File Systems
- Good knowledge of Scraping, Numpy, Pandas & Data Visualization libraries
- High level overview of Machine Learning algorithms
- Few basic Data Pre-Processing & Exploratory Data Analysis Projects

Requirements for **Middle Level Internships**

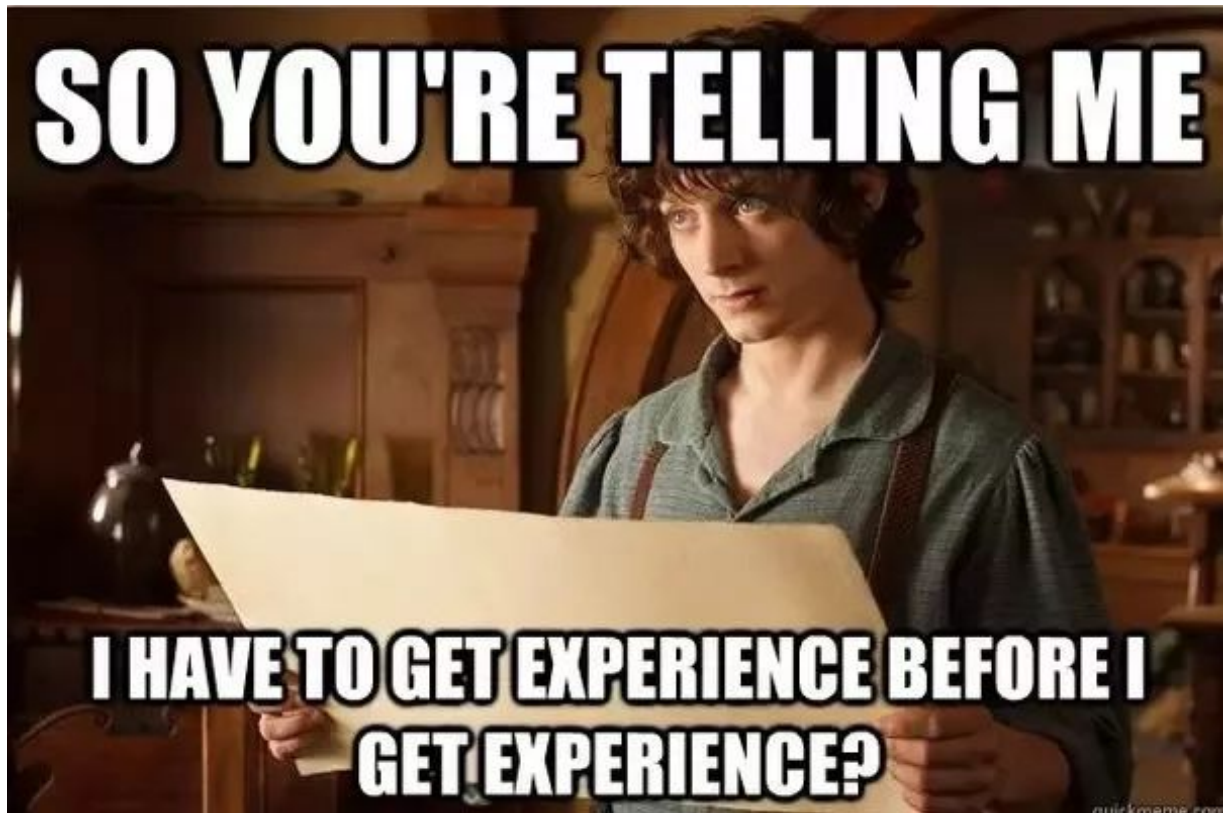
- Ideal for Third-Fourth Year Students of a 4 year Undergraduate course
- Sound knowledge of Data Science concepts & other ML algorithms with good grasp on statistics & concept of maths, SQL
- Good knowledge of Deep Learning concepts, DBMS, API development
- Self projects using ML algorithms

Requirements for **Advanced Level Internships**

- Ideal for Final Year Students of a 4 year Undergraduate course
- Specialised domain of expertise like Computer Vision, Natural Language Processing etc.
- Proficient with whiteboarding of ML algos along with explanation of the basics
- Basic knowledge of Docker, Cloud
- 4-5 very good projects using the complex Deep Learning concepts like LSTMs, Transformers etc.

Best way to get a Pre-Placement Offer before campus hiring starts

The **Why** of Data Science Internships



Why should students do **Data Science Internships** ?

- **To learn dealing with Messy, unstructured, incomplete data** (This is real industry)
- **To experiment & learn new things** as an Intern so that you can save time as a Full Timer excluding the mistakes you did as an intern.
- **To understand how end to end real world Data Science applications works.**
- **To network with people** who look like Future of You.
- **To work under pressure & learn how to deliver** in tough deadlines too.

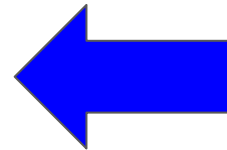
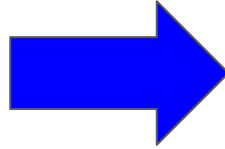
To make like minded friends & throw a party at Starbucks

The How of Data Science Internships

How to apply for Internships ?

- LinkedIn Jobs [Worked for Me]
- Angel.co [Worked for Me]
- Internshala [Worked for Me]
- Through Organization's Careers Page [Worked for Me]
- Commenting Interested on someone's LinkedIn Posts
[Not Worked for Me]
- Career/ Internship Fairs [Never Tried]
- Via Winning Hackathons [Worked for Me]
- Asking for Referrals [Worked for Me]
- Community Events [Have seen it work]

My First
Approach



My Second
Approach

WARNING
for
of Data Science Internships

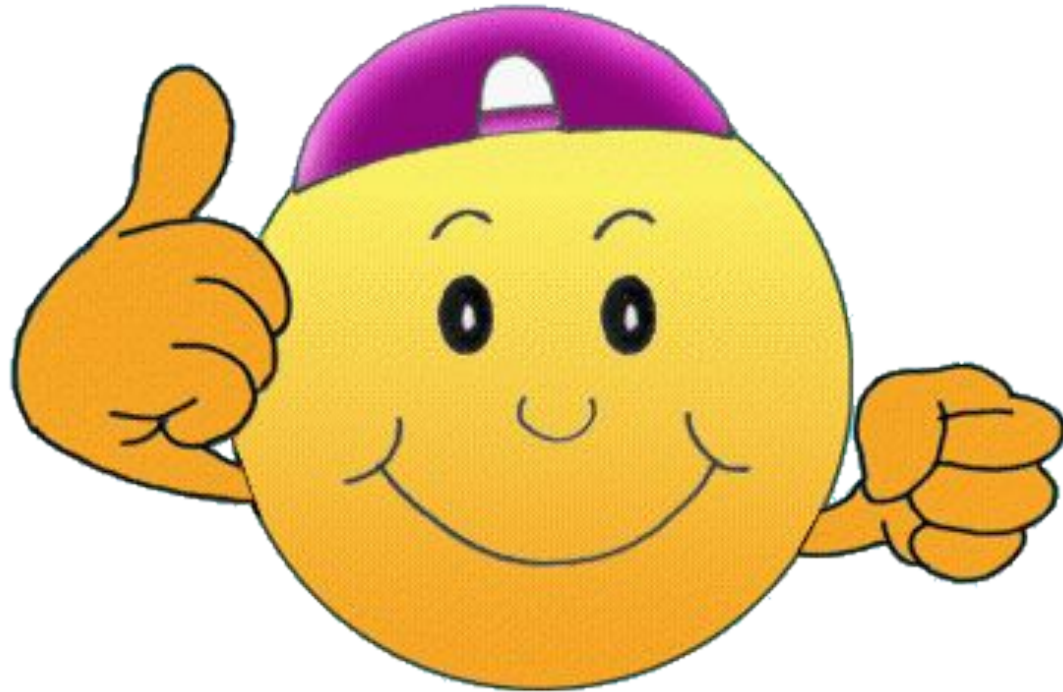
Pay for Trainings, not Internships



A few useful resources

1. <https://towardsdatascience.com/use-kaggle-to-start-and-guide-your-ml-data-science-journey-f09154baba35>
2. <https://www.coursera.org/learn/competitive-data-science#syllabus>
3. <https://towardsdatascience.com/how-to-successfully-manage-a-data-science-delivery-pipeline-33bdec1a9a27>
4. <http://kaggle.com/learn>

GO FOR IT !



GOOD LUCK !

Let me answer your Questions now.

Finally, it's your time to speak.



Danke Schoen

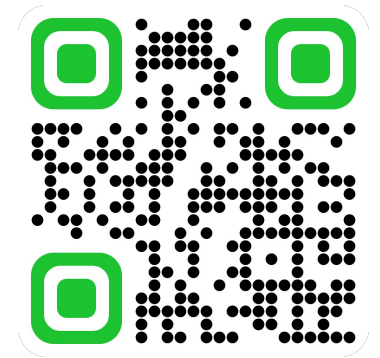
Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayon-roy@outlook.com

LinkedIn : <https://www.linkedin.com/in/ayon-roy>

Website : <https://AYON-ROY.NETLIFY.APP/>



Scan to Connect