AmEx Ignite Proposal

Section A: Personal Details

Name: **Vishal Anand** Institute: **IIT Kharagpur**

Roll No: **15ME33004** Year: **5th Year** Branch: **Mechanical Engineering**

Degree (B Tech/Dual Degree/Integrated): **Dual Degree**

Please provide details of prior background in machine learning/artificial intelligence (including courses/projects/competitions undertaken).

INTERNSHIPS

Data Science Associate- *[Publicis Sapient, Bengaluru]*

5th May – 5th July, 2019

- Offered with **a Pre Placement Offer** to join the company full time as **Senior Associate Data Science** after graduation.
- Designed a **Flask API** which analyses the threat percentage of YouTube channel based on **text, image and video features**
- Used **NLP** on textual data of YouTube, used **Hierarchical Attention Network** to predict the threat percentage of the videos
- Built a **scraper** to extract YouTube data. Used **Feature extractor** to extract frame level and video level data from videos.
- Used various **deep learning** models to predict which videos will be best to run **YouTube ads** for best returns.

Data Analyst - *[Medbay, Noida]*

1st Dec - 31st Dec 2018

- Used **IIB data** to create models to **predict** the hospitalization rate, claims rate and claim **amount** of potential customers.
- Predicted the price of Health Loan Card for population categories and analyzed **the financial cost of risk and uncertainty**.
- Gathered, understood and documented detailed business data, **reports** and **visualization** for Investors and internal use.

Cryptocurrency and ICO Investment Analyst - *[Tropyc, Bengaluru]*

15th May - 15th July 2018

- Improvement of metrics for investment **rating algorithm** for a diversified billion-dollar crypto currency and ICO market.
- **Extracted, filtered, cleaned, and visualized** ICO data and key performance indicators to interpret trends and patterns.
- Built the **analytics page** based on data-driven understanding of ICO, its valuation, market size, and other key indices.
- Created **1,000,000,000** decentralized ERC20 utility token "**Tropycoin**" on Solidity which is powered by **Ethereum**.

PROJECTS

Cognitive Workload Detection using EEG data and Machine Learning – *[Prof. D Samanta , IIT Kharagpur]*

August 2018- April 2019

- Formulated **Brain-Computer-Interface** solution for measuring **Cognitive Workload** for classifying individual and task
- Designed experiment for the collection of **EEG data** in industrial scenarios depicting various levels of cognitive workload
- Performed Signal Processing & Feature Engineering on EEG data using channel, **Feature Extraction** & Optimization.
- Modeled for Cognitive workload measurement using K-NN, Random Forest, Decision Tree Classifier, SVM and MLP.

Autonomous Vehicle Optimization – [*Prof. Pallab Dasgupta, IIT Kharagpur*]

November – December, 2017

- Designed an end to end **Feed-Forward Neural Network** architecture for training the ML Algorithm for **autonomous driving**.
- Trained the architecture by converting colored image input to grayscale using **Image Processing** and Pattern Recognition.
- Boosted performance of the learning algorithm by improving the dataset with enhanced contour edges for less error.
- Built up the network to produce correct steering direction of the automobile using Back-Propagation Algorithm.

COURSES

Machine Learning | Stanford University | Andrew NG

Introduction to Data Science in Python | University of Michigan | Kevyn Collins Thompson

Applied Machine Learning in Python | University of Michigan | Kevyn Collins Thompson

Soft Computing| IIT Kharagpur | Dilip Kumar Pratihar

Educational Data Analytics | IIT Kharagpur | Jiaul Hoque Paik

Section B: Proposal

(i) Describe the project you wish to work on. How did you come up with the idea?

I propose to design, develop and deploy a **Proof of Concept** of **Federated Learning** on **local server (PC) and raspberry pi**. It will be intended to achieve decentralized, distributed, on-device machine learning. It will solve two issues

1) **Privacy**- Most of the user generated data is private (on phones, computers), A user does not want to share their personal data like messages, mails, photos, transactions etc. to companies. Companies can't make use of such data for building a robust machine learning model due to their non-availability. In this scenario where privacy is such a major concern, making use of the personal data without breaching their privacy is a challenge

2**) Large data Problem-** Everyday tremendous amount of data is generated and building models and gaining insights from them requires high computation costs and time.

Federated learning solves both the problems. Federated learning works like this:

There are two entities, a **central server/system** (lets say your laptop) and **clients** (lets say 1000 mobile phones). If I had to make use of the personal images of the mobile phone users for building a face recognition model, for being able to do that without the images ever leaving their phone (to secure privacy), I would make use of federated learning. The central server will send a CNN model to each mobile phones, it will make use of the images in the phones to train the CNN model individually on each phone. When training has been done, the phones will send back **only the updated weights** of the CNN architecture to the central system. Here the trained weights from all the phones will be aggregated to build a global model which will be more robust than the individual models trained on the phone.

Now this global model can easily identify the faces of people. This could be done **without breaching user's privacy** and secondly **using the processing power of the phones** itself which makes the whole setup scalable and powerful. Additionally, the large data problem is also taken care of since, the data doesn't have to be aggregated at a central location and we don't need any powerful computers to run extensive machine learning operations on such a large dataset. The data is already distributed among the mobile phones and since the processing power of the end devices (mobiles phones) are sufficient to carry out the machine learning operations, it tackles the big data issue.

If such a process is done on millions of phones the computing power to build any robust model will be enormous. With the emergence of AI chipsets, Federated learning provides a **privacy-preserving mechanism** to effectively leverage those **decentralized** computation resources to train machine learning models.

Use Case- Since data never leaves its original premises, federated learning opens up the possibility for different data owners at the organizational level to collaborate and share their data.

Take the case of two banks. Although they have non-overlapping clientele, their data will have similar feature spaces since they have very similar business models. They might come together to collaborate for building a better model/outputs, leading to better outcomes for both.

Idea

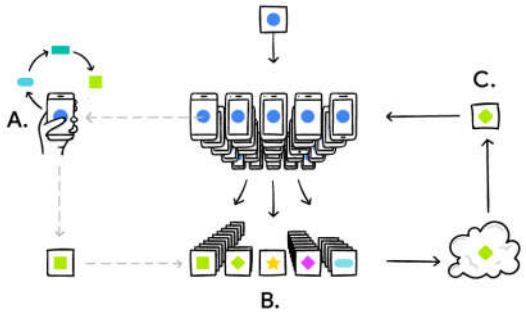
Once I had to take down surveys from people about their personal details for one of my project. It was time consuming. I couldn't find any publically available dataset for the same. In addition to this I read on a blog that Google claimed that if users give up their right to privacy, they can make a system so robust which can even predict what the person is thinking. **90% of today's data has been created in last 2 years**. And in next 5 years the systems would be generating so much of data that it would become almost impossible to get analytic solutions out of them and also the computing power and computing time will become costly to run on such extensive dataset. These set of ideas set up some motion in my brain and left me thinking about the **issues.**

So I started looking for answers for this particular problem.

I came across federated learning and I found that it not only **solved the problem of extensive data training but also it took care of privacy issues.** And since then I wanted to work in this domain and I have been researching more about it. This technology is relatively new and it has the potential to disrupt how we train models and how data privacy can be taken care of.

(ii) Provide a detailed project plan with project milestones and estimated timelines.

Working of federated Learning Framework



Your device personalizes the model locally, based on your usage **(A**). Many users' updates are aggregated **(B)** to form a consensus change **(C)** to the shared model, after which the procedure is repeated.

Highlights of proposed projects will be:

- Explore the Federated Learning framework. Build a **handy solution for Proof of Concept** Demonstration on Local Devices.
- Building of Federated Learning Prototype on **Local Server** using **Tensorflow Federated** and **Pysyft** frameworks.
- Prototyping Federated Learning on hardware devices like **raspberry pi** using the best performing frameworks out of Tensorflow Federated and Pysyft.
- Trying to explore **Federated Averaging Algorithm** and its modifications while deploying Federated Learning.
- Utilizing encryption techniques for privacy preserving at user's end while model sharing.

Methodology:

The initial aim to build this framework will consist of **building a Proof of Concept on local server on PC** and initializing a set of small number of clients (Say 20) which can be seen as any real world devices whose data can be valuable for companies and businesses.

For the initial prototype the POC will be built using two frameworks. **Tensorflow Federated** as well as **Pysyft**. These support federated learning operations and would be evaluated based on several metrics like, usability, speed, convenience and accuracy.

The best performing framework will be used to deploy prototype 2.0 on raspberry pi.

Raspberry pi will act as totally isolated end device and since they have their own computing power, with different hardware builds and run on different hardware and software configuration with various heterogeneous characteristics. They could be assumed to a fair extent as real world end devices like those of mobile devices and other systems. The motivation to use a raspberry pi is to apply the federated learning on real world scenarios (mobile phones) with less complexity.

The datasets that will be used for training testing and deployment would be **publically available datasets like EMNIST dataset, World bank open dataset, IMDB dataset.** This is because we are trying to implement the federated learning framework. We are not much concerned about the type of data or deep learning algorithms involved, rather we are trying to evaluate how can this framework be made handy to be used on all sorts of machine learning problems. Achieving a success in building this framework can open door to large scale application in various fields and sectors.

The training of the dataset inside the clients (raspberry pi or local server) would use state of the art deep learning algorithm techniques.

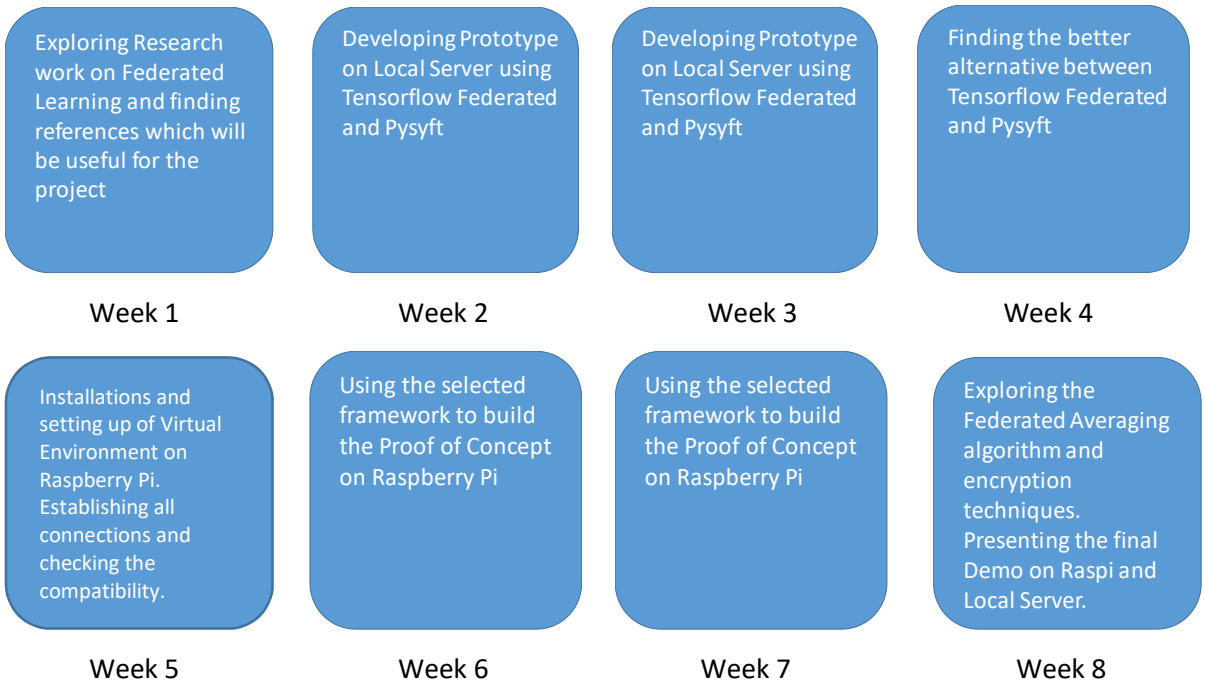
The training would be evaluated on several metrics like **global training time, local training time, loss, accuracy, AUROC etc.**

Having built the prototype, I will be trying to explore the **Federated Averaging Algorithm** involved in aggregating the models. Given the time duration, this work will presumable require much time hence if not finished could be added as future work for improving the computation time and accuracy of the whole prototype.

Encryption work using **SMPC protocol** could be used instead of public / private key to encrypt a variable at the end devices to secure the privacy and making the whole prototype decentralized where each node will have to agree to decrypt a variable at central server's end.

This encryption work would be targeted depending on the time availability at the completion of the project.

Milestones and Timeline



(iii) Are you starting this project from scratch or is it an on-going project? If it is a work in progress, provide more details.

I will be going to start the work from **scratch**. But since only a few work and research have been done on this particular technology, I would be referencing them for my project. The technology is still in very novice phase and using Federated learning as an application for training on decentralized devices for various use case is something that has not seen much of usage.

My objective is to **make the technology handy for direct application** in solving real world business problems.

(iv) Would you require financial funding for access to new data sources or computation? If yes, please elaborate on the requirement clearly stating the purpose.

For building the federated learning prototype on **Raspberry pi**, I would need some financial support to buy them. For the building the proof of concept**, two raspberry pi** would be enough for the deployment, testing and creating the prototype. Hence a financial aid of **6000-7000 INR** would be sufficient to carry out the whole project.

There **won't be any investment for getting data** because I would be using publically available dataset for deployment purpose.

(v) Describe the techniques you intend to explore to accomplish your project. Also, specify details of how you plan to obtain data to train/test the algorithm.

I will be using **Pysyft** and **TensorFlow Federated** to train models on **Local Server** and **Raspbery Pi**.

- **PySyft** is a Python library for secure, private Deep Learning. PySyft decouples private data from model training, using Federated Learning, Differential Privacy, and Multi-Party Computation (MPC) within PyTorch.

- **TensorFlow Federated (TFF)** is an open-source framework for machine learning and other computations on decentralized data. TFF has been developed to facilitate open research and experimentation with Federated Learning (FL).

In addition to this, I will also be exploring **Federated Averaging Algorithm**.

- In Federated averaging algorithm each selected client computes an updated model using its local data. The model updates are sent from the selected clients to the server. The server aggregates these models (typically by averaging) to construct an improved global model.

- If the time permits incorporation of **SMPC encryption technique** in the learning will be explored.

In the whole project data won't be an issue because I will be making use of Publically available dataset like EMNIST dataset, World bank open dataset, IMDB dataset etc. The deep learning models used for classification would be state of the art techniques. My main aim is to build a decentralized learning prototype which could be used to train any type of machine learning algorithm. Thus leveraging the computation power of the end clients and decentralizing the model training process in contrast to traditional deep learning techniques.

(vi) What kind of help do you expect from the Mentor who will be guiding you? (The mentor will be an experienced data science expert, working at AmEx)

I need the mentor's guidance to finish the whole project. With such an expertise I believe the mentor could get me through roadblock where complex algorithm and their usage can become tricky. I would expect that such a new technology which has not seen much of usage since it is still in research phase, we can collaborate to deploy the prototype to solve business problems and make use of the enormous computing power at our disposal. I am expecting that with my mentor's guidance, the prototype could be expanded to finance industry as well where the data privacy of transaction of the customers and collaboration of data from other institutions at organizational level are a big obstacle in building robust machine learning models.

In addition to this I am expecting that the mentor's role in guiding through the overall procedure and brainstorming the use cases scenarios of the prototype, with the amount of resources that we have at our disposal will be worth putting efforts into. At the end the mentor will play a crucial role in supervising, suggesting improvements and deployment of the projects within the time limit. Having a mentor from American Express will also be helpful to gain meaningful insights into the vast universe of Decision Science that American Express has built to drive business growth and delivering superior customer service and how we can together make use of this proposed technology to further improve the process is something I am curious to know about and am looking forward to learn from them.

(vii) What should be the success metrics on which your project should be evaluated?

Success metrics

- **Timeline**- The milestones and the timeline for the whole project has already been provided above. I should be judged on how am I able to keep the timeline intact without affecting the quality of my work.
- **Prototype building** - I propose that I should be able to build **two prototypes**. One on local server and another on raspberry pi. Finishing both of them as per the given timeline would be a key index in measuring the success of the project.
- **Demonstration**- At the end of the project I should be able to give a demonstration of the project live. On both the systems (local server and raspberry pi). Having done this, I should prepare a presentation which accumulates all the necessary information about the project and its usability.
- **Documentation**- I should document all the codes and procedure that were involved in the project so that more work can be done on top of the project without any issue. The code should be reusable and also well documented.