Machine Learning as a Service

Learning the art of building data-driven products

***Workshop @ The Fifth Elephant 2017* - Day 2**

Amit Kapoor amitkaps.com

Anand Chitpothu anandology.com

Bargava Subramanian bargava.com

The Unicorn Skillset

- *Data Management*: data ingestion & wrangling
- *Modelling & Prototyping*: statistics, visualisation, machine learning
- *Product Design*: data narrative, dashboards, applications
- *Data Engineering*: data pipelines, cloud infrastructure

Outline - Day 1

Session 1: **Introduction and Concepts**

- Approach for building ML products
- Problem definition and dataset
- Build your first ML Model (Part 1)

Session 2: **Build a Simple ML Service**

- Build your first ML Model (Part 2)
- Concept of ML Service
- Deploy your first ML Service - localhost API

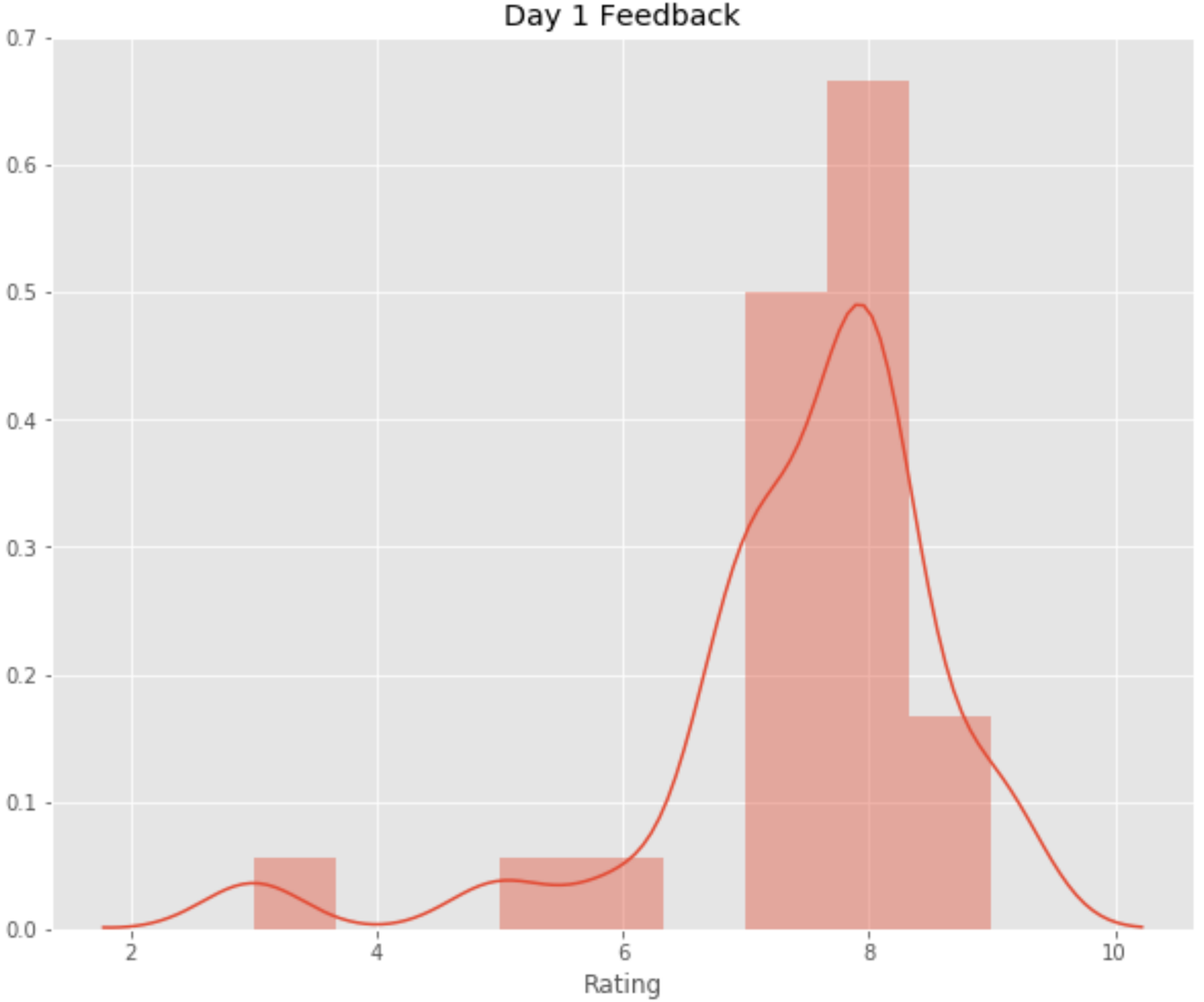
Outline - Day 1 (contd.)

Session 3: **Build & Evaluate ML Models**

- Feature Engineering
- Build your second ML model
- ML model evaluation (metrics, validation)

Session 4: **Practice Session**

- Practice problem overview and data
- Build your ML Model
- Build your API


Day 1 Feedback
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
8
10
Rating

What went well...

- **Good**: Explanation, informative, depth and breadth
- **Start**: for beginners, to diversified people
- **Well organised** content, presentation, working setup
- **Hands-on**: Coding, Jupyter
- **Firefly**: awesome, great

Even better if...

- Clear doubts **right away** vs. Take more **questions offline**
- **More ML**: internals, algorithm, evaluation, unsupervised, tuning
- **More Basics**: Stats, Pandas, Visualisation
- **Tough without** w/o ML background: Need to slow down, Make it simple, More examples / practice time
- **Share Reference** to learn
- **Too Long**: ML, Firefly

Outline - Day 2

Session 6: **Deploy to cloud**

- Get started with cloud server setup
- Deploy your ML service as cloud API
- Deploy your dashboard as cloud service

Session 5: **Build a Simple Application**

- Finish the ML part: "Random Forest"
- Visualisation in Python - Interactive Dashboard
- Create your first dashboard
- Integrate ML model API with dashboard

Outline - Day 2 (contd.)

Session 7: **Repeatable ML as a Service**

- Build data pipelines
- Update model, API and dashboard
- Schedule ML as as Service process

Session 8: **Practice Session & Wrap-up**

- Deploy on cloud - dashboard and API
- Best practices and challenges in building ML service
- Where to go from here

Schedule

08:45 to 09:30 : *Check-in & Breakfast*

09:30 to 11:00 : **Session 1**

11:00 to 11:20 : *Coffee break*

11:20 to 13:00 : **Session 2**

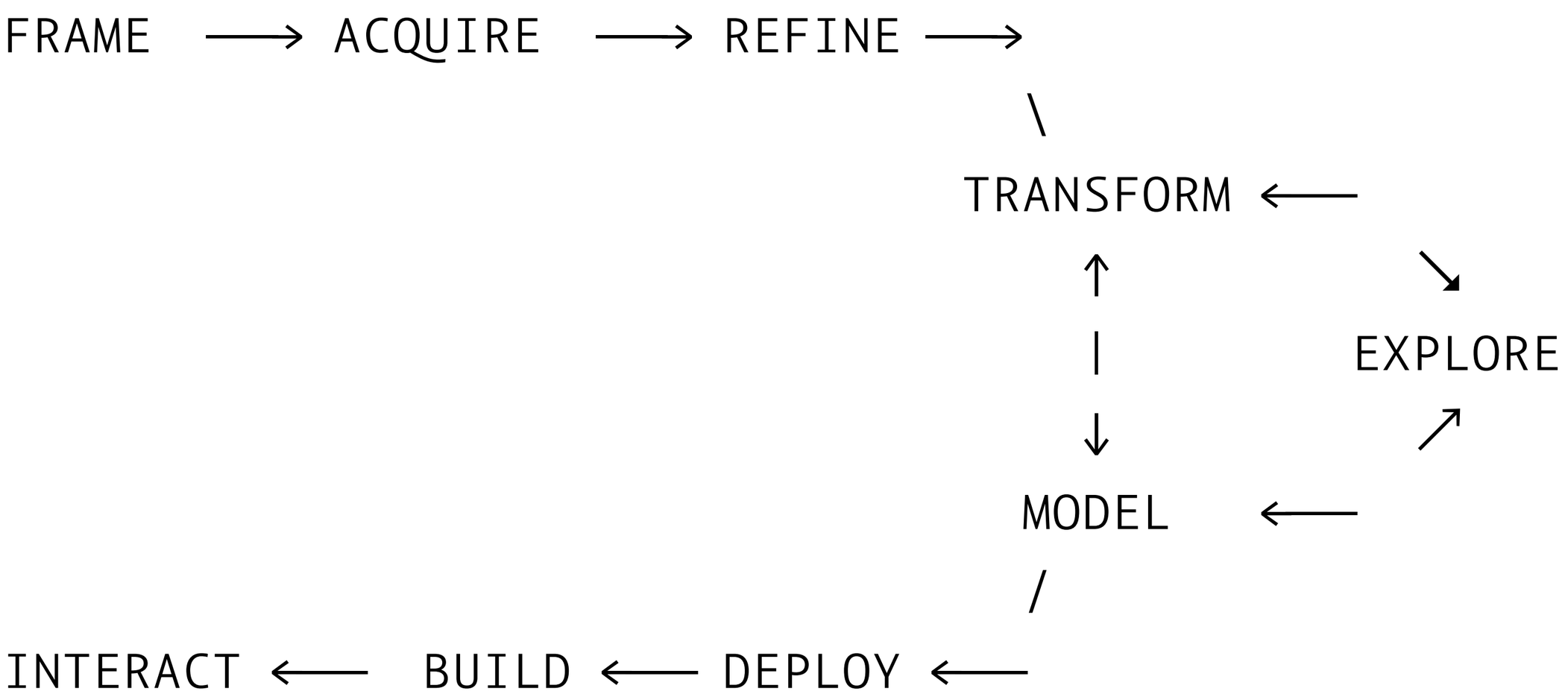
13:00 to 14:00 : *Lunch break*

14:00 to 15:40 : **Session 3**

15:40 to 16:00 : *Coffee break*

16:00 to 17:10 : **Session 4**

ML as a Service (MLaaS) Approach



MLaaS Approach

- *Frame*: Problem definition
- *Acquire*: Data ingestion
- *Refine*: Data wrangling
- *Transform*: Feature creation
- *Explore*: Feature selection
- *Model*: Model creation & selection
- *Deploy*: Model deployment
- *Build*: Application building
- *Interact*: User interaction

MLaaS Approach - Starting Point

- *Frame*: classification
- *Acquire*: pandas
- *Refine*: pandas
- *Transform*: pandas, scikit-learn
- *Explore*: matplotlib, seaborn, facet
- *Model*: scikit-learn
- *Deploy*: firefly-python, nginx, supervisor, gunicorn
- *Build*: flask, flask-wtf
- *Interact*: jupyter, local-host, cloud-server

Where do you go from here

- Learning journey
- Alternate approaches
- Scaling

Frame

- **Toy Problems**
- **Simple Problems**
- Complex Problems
- Business Problems
- Research Problems

Acquire

- Scraping (structured, unstructured)
- **Files** (csv, xls, json, xml, pdf, ...)
- Database (sqlite, ...)
- APIs
- Streaming

Libraries: pandas, beautiful soup, selenium, requests, sql-alchemy, blaze, boto3 etc.

Refine

- **Data Cleaning** (inconsistent, missing, ...)
- Data Refining (derive, parse, merge, filter, convert, ...)

Libraries: pandas, spark

Transform

- **Data Transformations** (group by, pivot, aggregate, sample, summarise, ...)
- **Data Encoding** (one-hot encoding, label encoding)

Libraries: pandas, blaze, spark, dask

Explore

- **Simple Vis** (matplotlib, seaboard, pandas)
- **Multi Dimensional Vis** (facet)
- Geographic Vis (leaflet, geoviews)
- Large Data Vis: Bin - Summarise - Smooth (datashader)
- Interactive Vis (bokeh, holoviews, altair, plotly)

Model - Supervised Learning

- *Continuous*: Regression - Linear, Polynomial, Tree Based Methods - DecisionRegressors
- *Classification* - Logistics Regression, **Tree - Decision Trees, Random Forest, Gradient Boosting**, KNN, SVM, Naive-Bayes. Bayesian Network

libraries: scikit-learn, dask, sparkML, tensorflow

Model - UnSupervised Learning

- *Continuous*: Clustering & Dimensionality Reduction like PCA, SVD, MDS, K-means
- *Categorical*: Association Analysis

libraries: scikit-learn, dask, sparkML, tensorflow

Model - Advanced / Specialized

- Network / Graph Analytics
- Optimisation
- Reinforcement Learning / Online Learning
- Deep Learning: tensorflow, pytorch, keras, caffe
- Applications:
  - Time Series: forecast
  - Text: nltk, spacy, gensim

Deploy

- Model Service: firefly, tfserving, flask
- Server Service: nginx, supervisor, docker
- Model update: cron, dask, airflow, luigi

Build / Interact

- Dashboard Visualisation / Decision Making Tools
  - Python + Html: flask, flask-wtf, django
  - Only Python: plotly-dash, bowtie, bokeh
- Narrative Visualisation
- Automated Decision Tools

Learning Resource - ML (1/3)

- **Beginner** *Python Data Science Handbook* (Github) (Book) - Jake Vanderplas' book is a thorough introduction to doing data science in python from a library perspective - numpy, pandas, matplotlib and scikit-learn. The material is very well documented and the notebooks are all available on GitHub.
- **Intermediate** *Introduction to Statistical Learning in R* (Book) (Video) (Github) - Hastie and Tibsharani are authors' of many of the algorithms used in Machine Learning. This book is free and even though it has all the applied stuff in R - the concepts are very well explained. To follow along with Python code- just use the GitHub repo link which implements every chapter in Python.

Learning Resource - ML (2/3)

- **Intermediate** *Python Machine Learning* (Book) (Github) - Sebastian Rashka writes very prolifically on ML and his blog itself is worth reading. This book is an applied book with real life examples and is useful to understand how to scale to larger datasets.
- **Advanced** *Learning from Data* (Slides) (Video): Yaser Abu-Mostafaâ€™s course is the best course I have seen on understanding how Machine Learning really works. It is both intuitive as well as bit mathematical. If you are really keen to understand the whole theory of generalisation and how the algorithms really work - you will really like this course.

Learning Resource - ML (3/3)

- **Our Teaching Material** (GitHub): We have a fair number of teaching repos on different Machine Learning topics and they are constantly being updated. If you liked this workshop, you may like some of them. We will also be updating The Art of Data Science, HackerMath for ML, and Applied Machine Learning as we do more workshops in the future. So you can come back later and check it out too.
- Additional Resource:
  - Introduction to Statistics - Hacker's approach: Think Stats
  - Statistics - Theoretical Background: All of Statistics
  - Statistics - Theoretical Background: Statistics is Easy
  - Harvard Data Science Course - CS 109 Course

Learning Resources - Data Engineering / System / Cloud

- DevOps
- Data Pipelines
- Cloud Platforms

Learning Approach (1/2)

- **Practice, Practice, Practice**: Or to simplify just Practice. Repeat the data science process many times and as much as possible, try to write the code, break things and see how to fix it.
- **Understand the basics**: Always investigate why the algorithms and libraries work the way they do and build an intuition around what is happening in the process.
- **Focus on Data Science Portfolio**: Work on at least 6 - 8 different projects over the next 1 year to build a portfolio of projects across data science topics.
- **Build a public profile**: It could be a blog, a website, GitHub repos - so that people can see what you are doing and connect if they wish.
- **Do projects that interest you**: You are more likely to stick through the long arduous journey of doing it. It could be in your work domain or it could be sports, economics, politics, public policy etc.

Learning Approach (2/2)

- **Understand your learning style**: Are you the hacker kind, who learns best by finding out how things work or do you prefer the structured classroom style learning. This would help you identify better what works for you to learn new concepts.
- **Share what you do and get feedback**: Put yourself out there. It may be a small project, a visualisation, an analysis, using a library, anything. Nothing is too small to share.
- **Find a community of like minded people**: Most of our learning happens through peers. Build or join a community of peers with whom you can engage, teach and learn. It could be virtual but even better if it is face to face.
- **Find your own rhythm**: Build and do something every day, week or month. Find your own cadence and start creating.

Future Workshops: https://fifthelephant.in/

- **Data Science Bootcamp** (5 Sundays in Aug/Sept)
  - The Art of Data Science
  - Data Visualisation for Data Science
  - HackerMath for Machine Learning
  - Applied Machine Learning
  - Full Stack Data Science
- **Deep Learning Bootcamp** (2 Saturday in Sept)
  - DL for Images: CNN
  - DL for Text: Word Embedding, RNN, LSTM

Feedback Form

https://goo.gl/3cpdHE

Machine Learning as a Service

Learning the art of building data-driven products

Workshop @ The Fifth Elephant 2017

Amit Kapoor amitkaps.com

Anand Chitpothu anandology.com

Bargava Subramanian bargava.com