Applied Machine Learning for Business Analytics

Lecture 1: Introduction to Machine Learning and Its Production

The following lecture slides and notebook will be updated one week before the lecture.

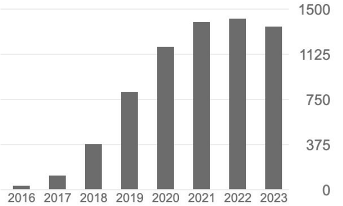
Lecturer: ZHAO Rui

About me

- Lecturer:
    - ZHAO Rui
        - Head of Data & Quant at Pluang (All in one investment app)
        - Adjunct Faculty at NUS, teaching BT5153 and BT4012
        - Research interests are within machine learning and its applications on quant trading, time series data and text data.
        - Google Scholar (6K+ citations)
        - Linkedin
    - Pls just address me by Rui (my first name, pronounced as Ray)
    - Email: diszr@nus.edu.sg

Cited by

	All	Since 2019
Citations	6781	6575
h-index	22	21
i10-index	27	25



Logistics

- Check course website frequently
  - https://bt5153msba.github.io
- 100% f2f lectures
  - Attendance check would be conducted randomly
- Class hours
  - From 6:30 pm to 8:30 pm

Agenda

1. Course overview
2. What is machine learning
3. From Business Problems to ML Solutions
4. Gap between theory and production
5. Group Projects

1. Course overview

Goals of this course

- Understand conceptually the mechanism of machine learning and data science algorithms
- Implement the whole pipeline for your ML projects
- Select appropriate machine learning tools/techniques for business applications

Learn and Improve upon the applications of machine learning

Course background and overview

- Basic ML/Data Mining models have been covered in other modules
- In BT5153:
  - “**Advanced**” architecture
  - **Hands-on** Experiences
  - In each lecture, roughly 90% Slides and **10% IPython notebooks**.
  - More **Practical** Assignments/Exams

In practice, be solution-focused, not buzzword-focused.

Models & Systems

- E2E ML System
  - Data
  - Modelling
  - Evaluation
  - Deployment
- Explainable Machine Learning
- Representation Learning
  - Word Embeddings
  - Transformers
  - BERT
- Large language models

Applications

- Spam Detection
- Recommendation
- Image Categorization
- Sentiment Analysis
- Customer Profile Prediction
- Question Answering Tasks
- Name Entity Recognition
- Etc

Hands-on experience

- Understanding domain, prior knowledge
- Data integration, selection, clearing, pre-processing, etc
- Learning models (little math, more intuitive ideas)
- Compare models
- Model interpretability
- Consolidating and deploying discovered knowledge
- Apply discovered knowledge to practical problems
- Python programming is not the teaching focus

Course assessment

- In-class Quizzes (10%)
- Individual Assignments (50%)
  - Three weekly individual assignments (10% each)
  - One mini-kaggle project (20%)
- Group Project (40%)
  - Project proposal (5%)
  - Final presentation (20%)
  - Final report (15%)

In-class Quiz

- It would be used for attendance check
- Up to 5 times. 2 points each time
- If you are going to miss the following class, please email our TA Xiaohui and cc me in advance. Otherwise, you will not get this 2 points if we have quiz in that lecture
  - Xiaohui: xiaohuiliu@u.nus.edu

Course Schedule

Date	Topic	Content	Assignment
Fri 01/13	Introduction to Machine Learning and its Production	TBU	N.A.
Fri 01/20	Training Data Generation	TBU	Assignment I Out
Fri 01/27	Neural Networks and Deep Learning	[TBU	Form your team
Fri 02/03	Deep Learning Practices	TBU	Assignment II Out
Fri 02/10	Auto-encoders	TBU	N.A.
Fri 02/17	Convolutional Neural Networks	TBU	Proposal Due & Assignment III Out
Fri 02/24	Recess Week	N.A.	N.A.
Fri 03/04	Explainable Machine Learning	TBU	Kaggle Starts
Fri 03/10	Frontiers in NLP	TBU	N.A.
Fri 03/17	Model Evaluation in Machine Learning	TBU	N.A.
Fri 03/24	Model Deployment in Machine Learning	TBU	Kaggle Competition Due
Fri 03/31	Causal Inference for Decision Making	TBU	Kaggle Report Due
Fri 04/07	Good Friday	TBU	N.A.
Fri 04/14	Why do ML Projects Fail in Business	TBU	N.A.
Sun 04/23	Reading Week	N.A.	Presentation and Final Report Due

2022

Date	Topic	Content	Assignment
Fri 01/19	Introduction to Machine Learning and its Production	TBU	N.A.
Fri 01/26	Data Preparation	TBU	Assignment I Out
Fri 02/02	Machine Learning Modelling	TBU	Form your team
Fri 02/09	NO CLASS (CNY)	TBU	N.A.
Fri 02/16	Machine Learning Evaluation	TBU	Assignment II Out
Fri 02/23	Machine Learning Deployment	TBU	N.A.
Sun 03/03	Recess Week	N.A.	Proposal Due
Fri 03/08	Explainable Machine Learning	TBU	Assignment III Out
Fri 03/15	From BoW to Word2Vec	TBU	Kaggle Starts
Fri 03/22	From Word2Vec to Transformers	TBU	N.A.
Fri 03/29	NO CLASS (Good Friday)	TBU	N.A.
Fri 04/05	LLM and its Practices I	TBU	Kaggle Competition
Fri 04/12	LLM and its Practices II	TBU	Kaggle Report
Fri 04/19	Why do ML Projects Fail in Business	TBU	N.A.
Sun 04/28	Reading Week	N.A.	Presentation and Final Report Due

2023

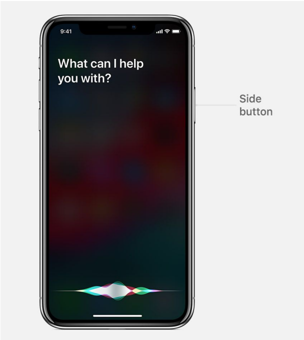
2. What is Machine Learning

Machine Learning is Everywhere

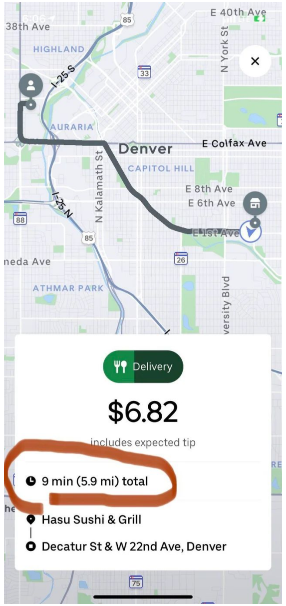
Face Unlocking

Recommendation

Fraud Detection

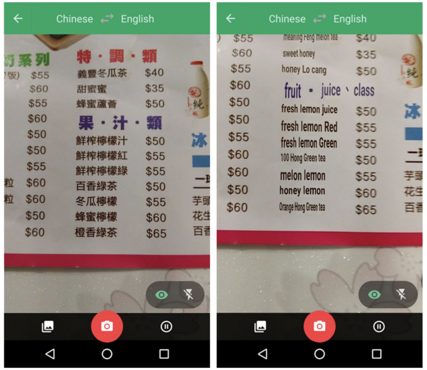


AI Assistant



ETA

Search



Machine Translation

Self-driving Car

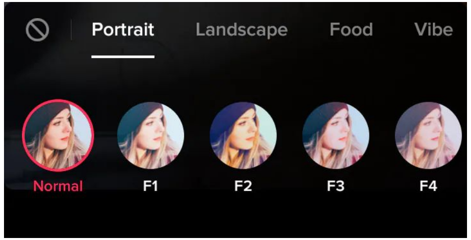
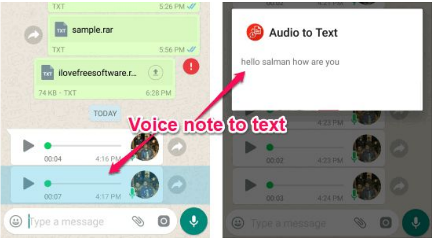


Photo editing



Voice to Text

Mat Velloso
@matvelloso

Follow

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

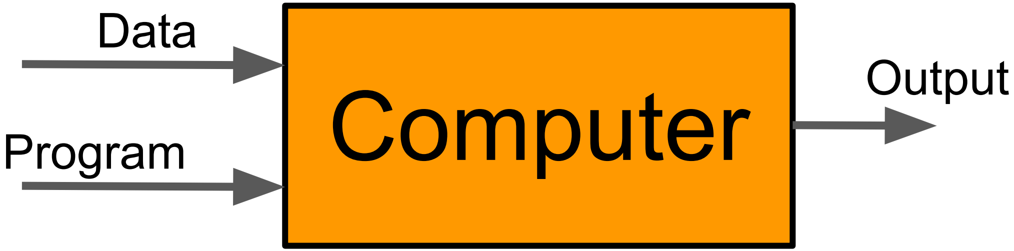
5:25 PM - 22 Nov 2018

8,541 Retweets **23,778** Likes

Python Programming

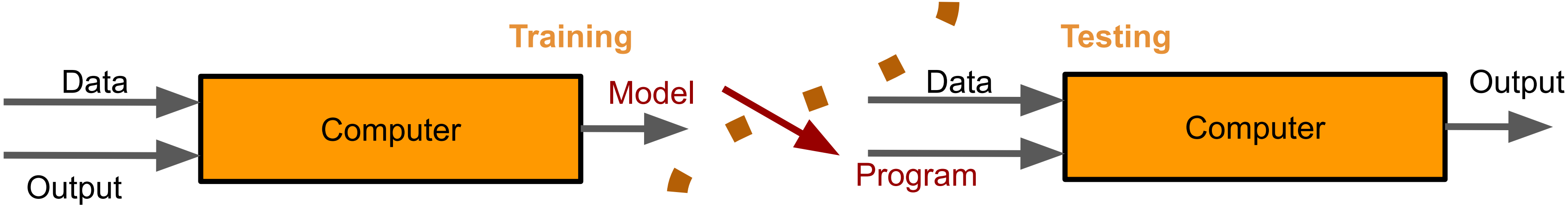
```
In [1]: a   =  3
        b   =  1
        q   = 3*a + 2*b
        print('result is {}'.format(a + b))
```

```
result is 4
```



Machine Learning

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
#create an object of KNN
neigh = KNeighborsClassifier(n_neighbors=3)
#train the algorithm on training data and predict using the testing data
pred = neigh.fit(data_train, target_train).predict(data_test)
```



Definition of Machine Learning

- Machine Learning is an approach to **learn** *complex pattern* from existing data and use these patterns to make **predictions** on **unseen data**.
- Therefore, there are following points to determine if a ML solution will fit your problem
  - Learn
  - Complex Pattern
  - Existing Data
  - Predictions
  - Unseen Data

Learn

- The system has the capacity to learn
  - From the data
- To apply Machine Learning, there must be something for it to learn.
  - E.g., database is not the ML System

Complex Pattern

- The patterns are complex
  - Look-up operation vs Object Detection
- What is difficult to humans is different from what is hard to machines

Complex Pattern

- There are patterns to learn
  - Should we predict the next outcome of toto?

  - Should we predict doge price?

Existing Data

- Data is available
- It is possible to collect data
- Exceptions?
  - Zero-shot learning (still trained over data from other domains)
  - Online learning

Predictions

- It is a “predictive” problem
  - We can benefit from a large quantity of cheap but approximate predictions.
- It is not only limited to estimations of values in the future
  - What is the tranx probability of this users in the following 10 days?
  - Is this cash out action a money laundry one?

Unseen Data

- Unseen data shares patterns with the training data
    - Training and unseen data should come from a similar distribution

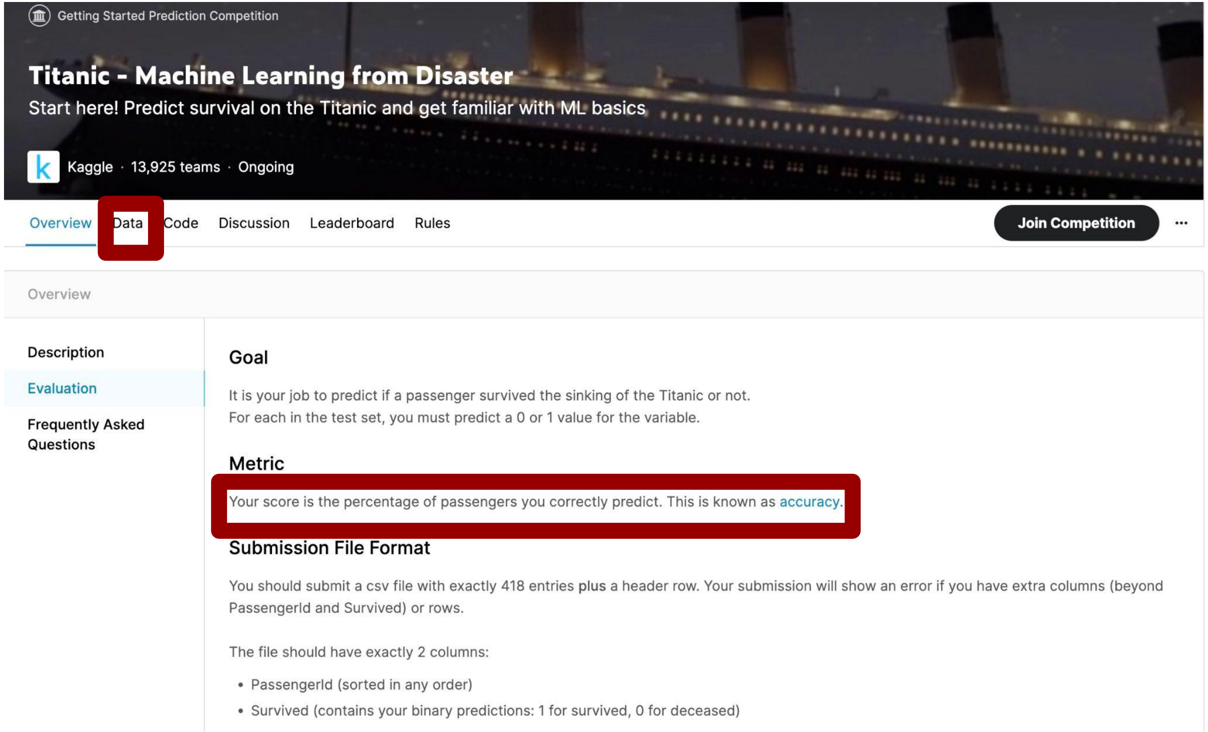
Domain Knowledge -> Solid Assumption

Other Factors to Make ML Solutions Viable

- The task is repetitive
  - New samples keep coming
- The cost of wrong predictions is cheap
  - Recommended wrong movies
- It is at scale
  - ML models are run 24/7
- The patterns are constantly changing
  - Subject matter experts are unable to encode the complete rule-set to solve the problem

3. From Business Problems to ML Solutions

Kaggle Style ML Projects



ML Projects here start with:

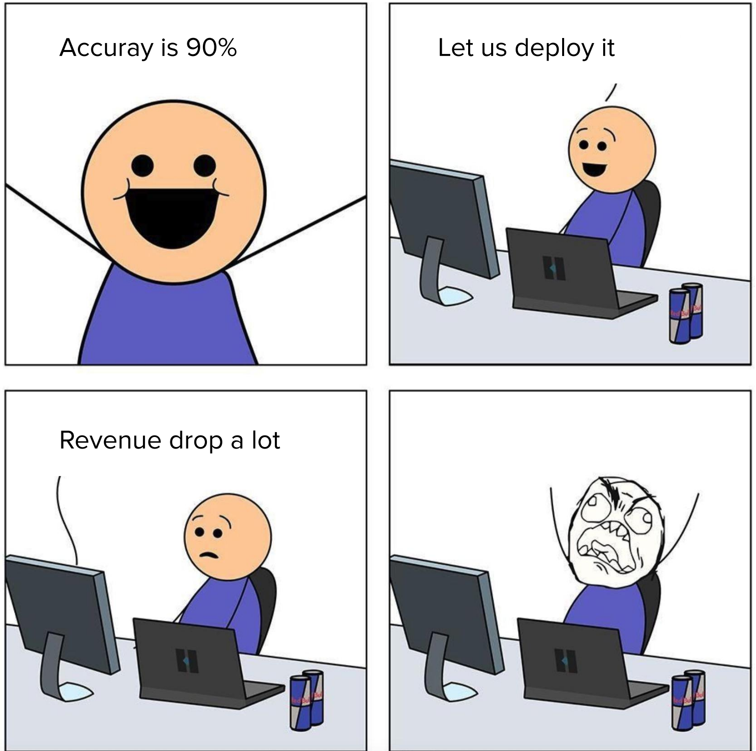
1. Dataset
2. Clearly defined metric

In Real-world

- ML or DS projects start from a business problem instead of a well-defined prediction task.
- Machine learning team is to **formulate** the business problem into the right ML problem and then **solve** it

In Real-world

Building a great ML solution to the wrong business problem is the most frustrating thing for ML/DS org.



How should we translate?

From a business problem to the right data science problem:

- Ask questions
- Explore the data to find high quality insights

A “real” example

- Assume we are working in ML/DS org at Netflix 
- Growth lead come to us with their requests 
- Then, the discussion will start as:

Based on Q1 OKR, we want to increase our users retention rate by 8%. Do you have any better ideas?

Got it. It looks quite impactful and let us work together! Do you have any hypothesis that why our users stop using Netflix?

A “real” example

Based on Q1 OKR, we want to increase our users retention rate by 8% in SEA. We would like to leverage ML solutions to achieve this goal.

Got it. The project looks quite impactful! Do you have any hypothesis that why our users stop using Netflix?

Yeah, we did some market research. Now, amazon prime video is providing lower fees.

Hmm, we also found users browsing time before they watch videos become longer.

Yeah, great sync. We have two business problems here:

- Pricing issues: our competitor is offering lower prices. The solution can be dispatching personalized discount with push notification
- Discoverability issues: our users can not easily find the videos that they are interested. I heard recommendation sys can guess what users will like. Should we also try this solution?

Thanks for the summary. Let us work on ml solutions

Hypothesis Prioritization

From the previous conversion, we are able to formulate hypothesis and create the to-do list by asking questions.

- Pricing Issues
- Discoverability Issues

Pricing Issues

- Business problem: Competitors are offering cheaper prices
- Idea: Send personalized discount with push notification
- ML Problems:
  - Who should we send notifications
  - How much is the voucher?
- ML Solutions:
  - Churn Prediction Model
  - Uplifting Models

Discoverability Issues

- Business problem: Users' conversion rate from homepage visit to video view is low
- Idea: Push personalized content to our users to increase conversion
- ML Problems:
  - Personalized recommendations
- ML Solutions:
  - Collaborative Filtering
  - Deep Learning

Source: https://research.netflix.com/research-area/recommendations

From Business Problems to ML Solutions

- The key skill would be: translating business problems into the correct data science problem
- Ask the right questions, list possible solutions, and explore the data to narrow down the list to one

From Business Problems to ML Solutions

- The key skill would be: translating business problems into the correct data science problem
- Ask the right questions, list possible solutions, and explore the data to narrow down the list to one
- Solve the problems
  - Build a dashboard
    - Build a user retention dashboard under different segments (age, geo, acquisition channels)
  - Data Exploration
    - Visualization, Group comparison (e.g,. Users from one marketing channel have a higher churn rate)
  - Train ML models
    - Should be checked only after trying the first two ideas

- **Junior DS/A are told the problems they need to solve**
- **Senior DS/A define the problems that need to be solved**

Role of ML/DS Org

- Translate abstract data into actionable business insights
- Automate and scale the above process if possible
- Be the interface to bridge biz/product and data
  - Therefore, we usually talk with two departments:
    - Biz departments: product, ops, marketing, growth
    - Engineering departments: data engineers

ML Production is not a few lines

```python
import pandas as pd
from sklearn import model
df = pd.read_csv()
X = df[feature]
y = df[label]
model.train(X, y)
model.predict(new_data)
```

Data scientists should know

- SQL
    - Query and extract data
- Python
    - Main programming language
- Presentation and Visualization
    - Talk and present information in an actionable manner
- Machine Learning
    - Automate and improve operations and business decisions
- Cloud services
    - Many companies built infra in the cloud
- Deep learning libraries
    - Deal with image, video or text data
    - Keras/Pytorch/Huggingface

4. Gap between Research and Production

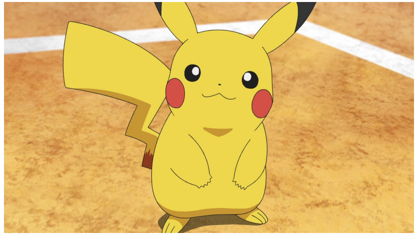
Four phases of ML Projects

- Phase 1: Before ML
- Phase 2: Simplest ML models
  - Start with a simple model that allows visibility: check hypothesis and pipeline
- Phase 3: Further Optimization
  - Different object functions
  - Feature engineering
  - More data
  - Ensembling
- Phase 4: Complex ML models

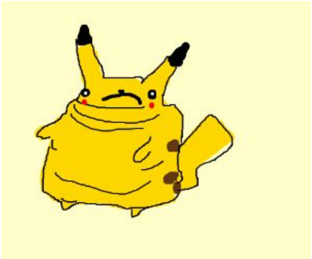
Data

- In real world, data is not perfect:
  - Missing data
  - Scale features
  - Identify outliers
  - Identify highly correlated variables
  - Identify variables with no variances
  - Check for overall hygiene
- Next week, we will discuss more about data preparation for machine learning applications.

Dataset in BT5153



Real Dataset



THE COGNITIVE CODER

By Armand Ruiz, Contributor, InfoWorld | SEP 26, 2017 7:22 AM PDT

The 80/20 data science dilemma

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy

https://www.quora.com/How-accurate-is-the-80-20-rule-as-a-Data-Scientist

Efficient Coding - Pandas as Example

- In programming, there are often many different ways to do the exact same operation, some of which are more optimized
- It is the same to data science or ML projects
- If your codes are not efficient, it would becomes a bottleneck when the scale and complexity of the problems increase
    - Pandas is the great tool for data manipulation, analysis and visualization.

How to loop effectively

- It is quite common to compute a new value from one or multiple columns in the original dataframe.
- Different codes will have different performances
- Tips are shared in this week's [lab notebook]

```
sum_square = lambda x, y: (x+y) ** 2
print(sum_square(2,3))
```

```
25
```

```
test_data = df_data[['X Coordinate', 'Y Coordinate']].copy()
```

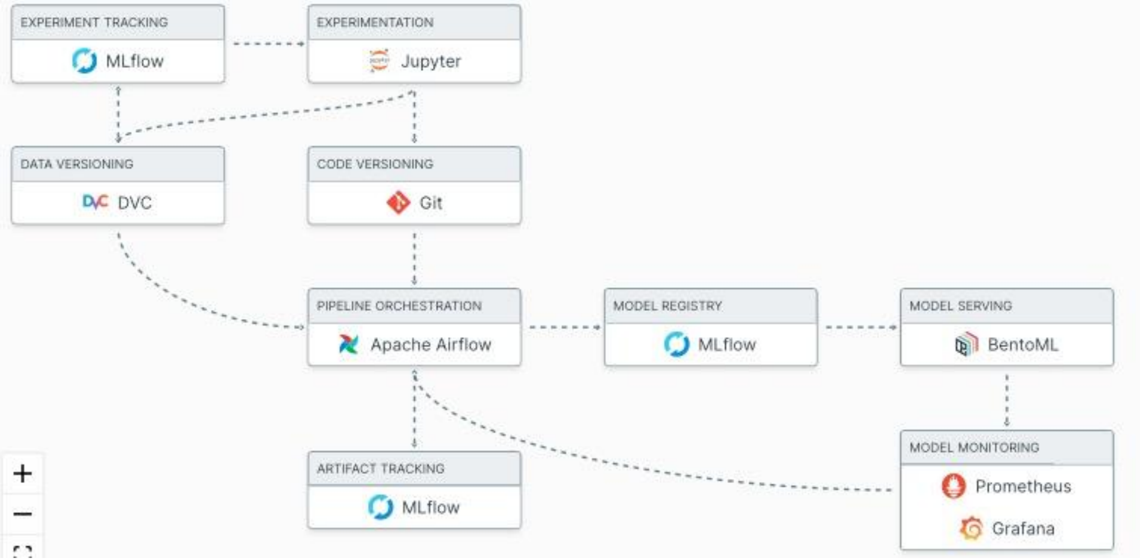
```
%timeit -r5 -n10 test_data.loc[:,'magic'] = [sum_square(value[0], value[1]) for _, value in test_data.iterrows()]
%timeit -r5 -n10 test_data.loc[:,'magic'] = test_data.apply(lambda row: sum_square(row[0], row[1]), axis=1)
%timeit -r5 -n10 test_data.loc[:,'magic']  = test_data.apply(lambda row: sum_square(row[0], row[1]), raw=True, axis=1)
%timeit -r5 -n10 test_data.loc[:,'magic']  = np.vectorize(sum_square)(test_data.iloc[:,0], test_data.iloc[:,1])
%timeit -r5 -n10 test_data.loc[:,'magic']  = np.power(test_data.iloc[:,0]+test_data.iloc[:,1], 2)
#%timeit -r5 -n10 test_data.loc[:,'magic'] = [sum_square(value[0], value[1]) for _, value in test_data.iterrows()]
```

```
470 ms ± 2.26 ms per loop (mean ± std. dev. of 5 runs, 10 loops each)
135 ms ± 3.61 ms per loop (mean ± std. dev. of 5 runs, 10 loops each)
33.4 ms ± 188 µs per loop (mean ± std. dev. of 5 runs, 10 loops each)
4.49 ms ± 62 µs per loop (mean ± std. dev. of 5 runs, 10 loops each)
271 µs ± 44.5 µs per loop (mean ± std. dev. of 5 runs, 10 loops each)
```

1700X speed-up

ML Deployment

- MLOps stack



Source: https://mymlops.com/

- BT5153 Hands-on notebook
  - Experiment Tracking ✅
  - Experimentation ✅
  - Data Versioning ✅
  - Code Versioning ✅
  - Pipeline Orchestration ✅
  - Runtime Engine ✅
  - Artifact Tracking ✅
  - Model Registry ✅
  - Model Serving ✅
  - Model Monitoring ❌
  - Feature Store ❌

5. Group Projects

Group project

- Build an ML/DS application
- Must work in groups of four or five
- One-pager proposal + Presentation + Report
- Detailed guidelines could be found here

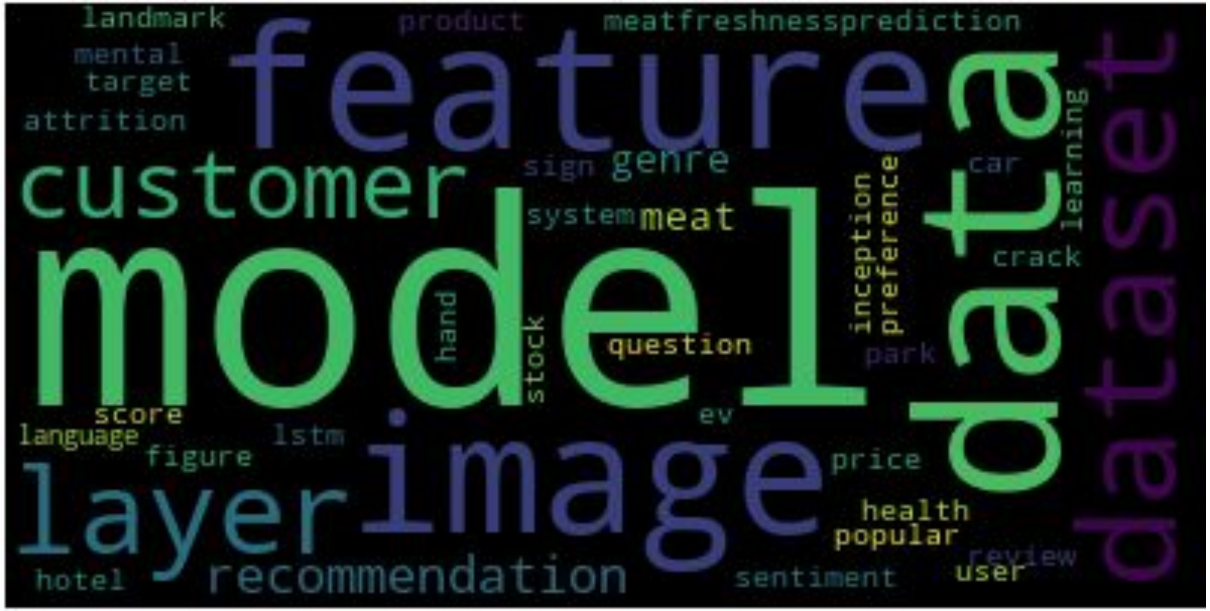
Paper analysis using NLP

- We collected and published all papers that were submitted from 2019 to 2022 (4 years !). Those papers discussed various kinds of applications of machine learning.
- NLP technique is also adopted to analyze the papers submitted last year.

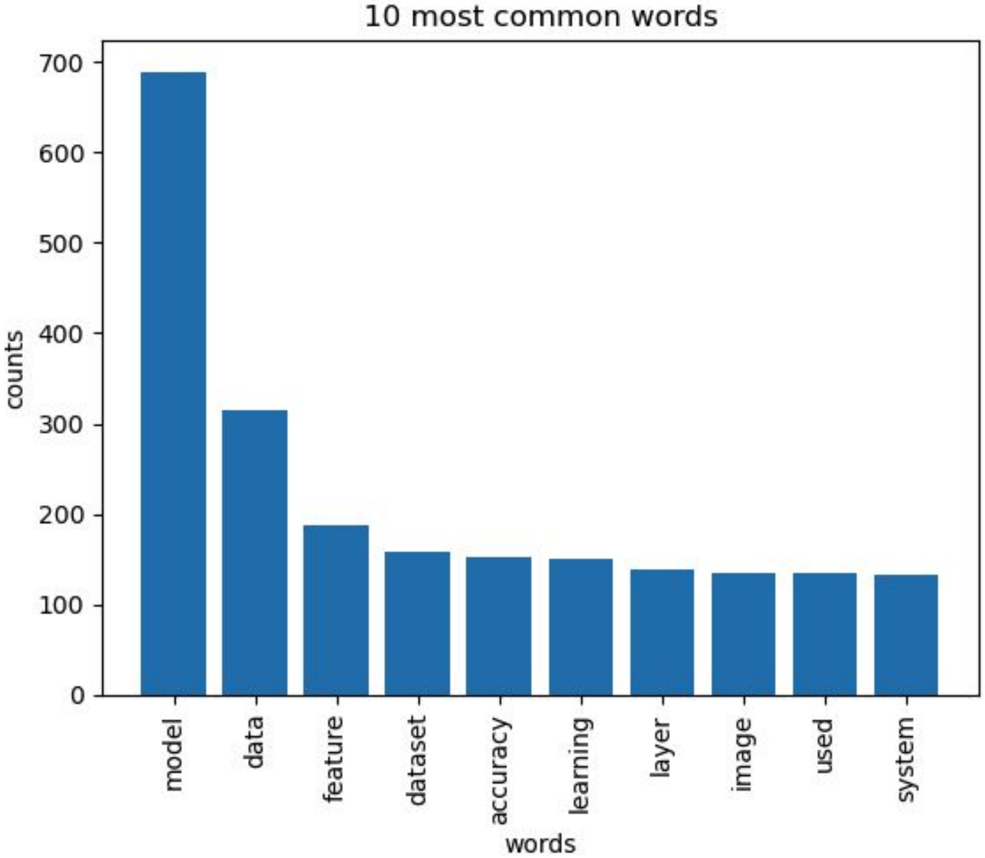
Previous Years Project Reports

- Spring 2022
- Spring 2021
- Spring 2020
- Spring 2019

Word cloud


landmark
product
meatfreshnessprediction
mental
feature
target
data
dataset
attrition
learning
customer
sign
genre
car
system
meat
model
inception
preference
crack
hand
stock
question
park
score
ev
language
lstm
figure
image
layer
price
health
popular
review
hotel
recommendation
sentiment
user

Top-10 high frequent words



Topic modeling

```
Topics found via LDA:

Topic #0:
category transformed lot next learningrate conference summary adam predicting structure

Topic #1:
model landmark sign hand language data accuracy transformer label frame

Topic #2:
question model feature popular score text dataset performance careervillage data

Topic #3:
model image layer data dataset learning accuracy training feature mental

Topic #4:
stock price model prediction lstm tweet data feature network function

Topic #5:
data park review model system recommendation hotel sentiment word customer
```

Previous submission

Neural networks for **fashion image** classification and visual search — [PDF] arxiv.org

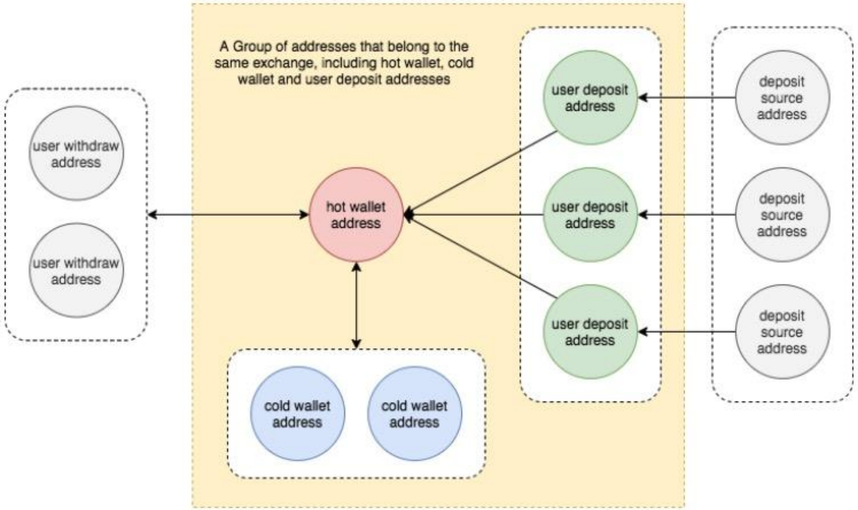
F Li, S Kant, S Araki, S Bangera, SS Shukla - arXiv preprint arXiv …, 2020 - arxiv.org

… We use real life **fashion images** from an Indian ecommerce website. The … of **fashion** items, we apply the model without any adjustments, which have one layer to convert **image** data into …

☆ Save 99 Cite Cited by 16 Related articles All 3 versions

Project Hint 1

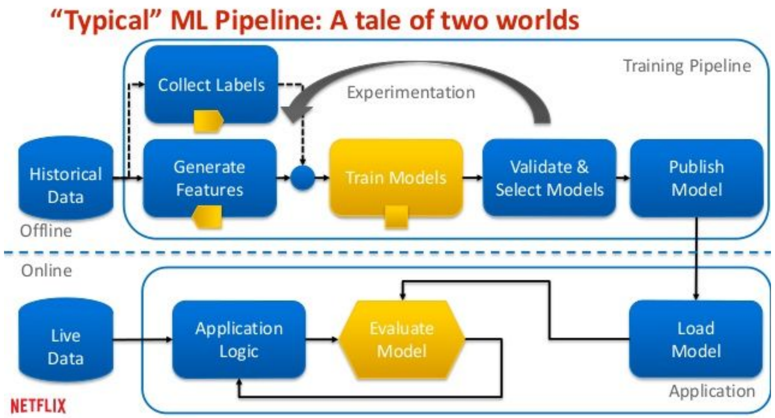
- Find a new business problem which can be solved by ML solutions
  - For example, assigning attribution labels to cryptocurrency addresses using blockchain data

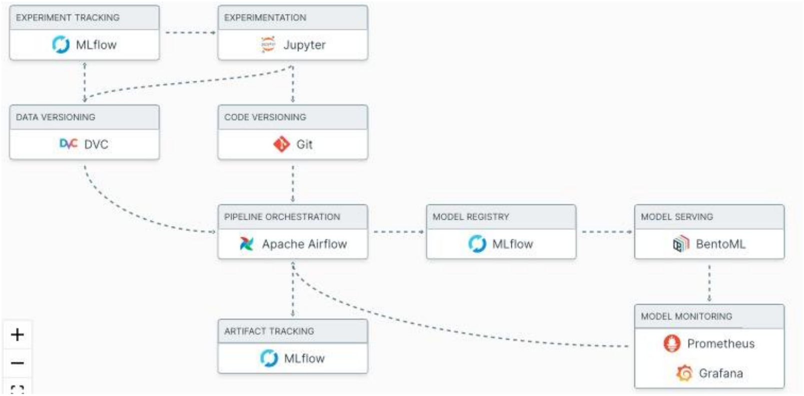


Source: https://arxiv.org/pdf/2003.13399.pdf

Project Hint 2

- Build a end-to-end ML pipeline





Project Hint 3

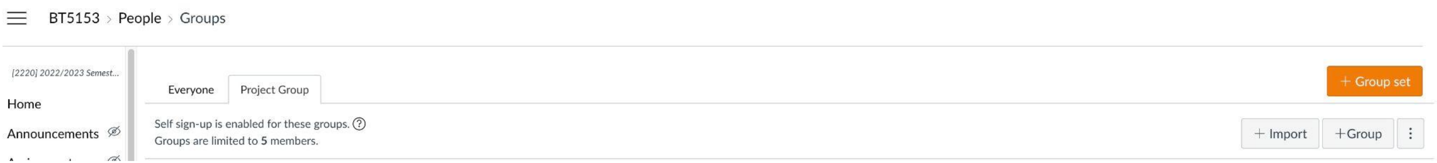
- In-depth analysis of machine learning algorithms on one specific application
- **Try to explain the findings**

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2014)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM_S (Silva et al., 2011)	–	–	–	–	**95.0**	–	–

Source: https://arxiv.org/abs/1408.5882

Form your group

- Find your group members
- Sign-up in Canvas



Next Class: Data Preparation

Must-Read: Using machine learning to predict value of homes on airbnb