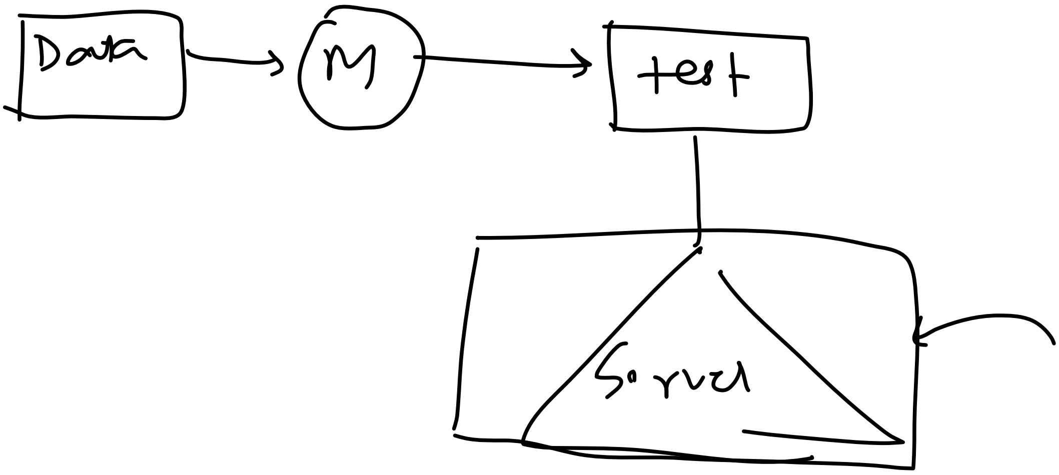
1. Batch Vs Online ML

Wednesday, March 17, 2021 5:30 PM

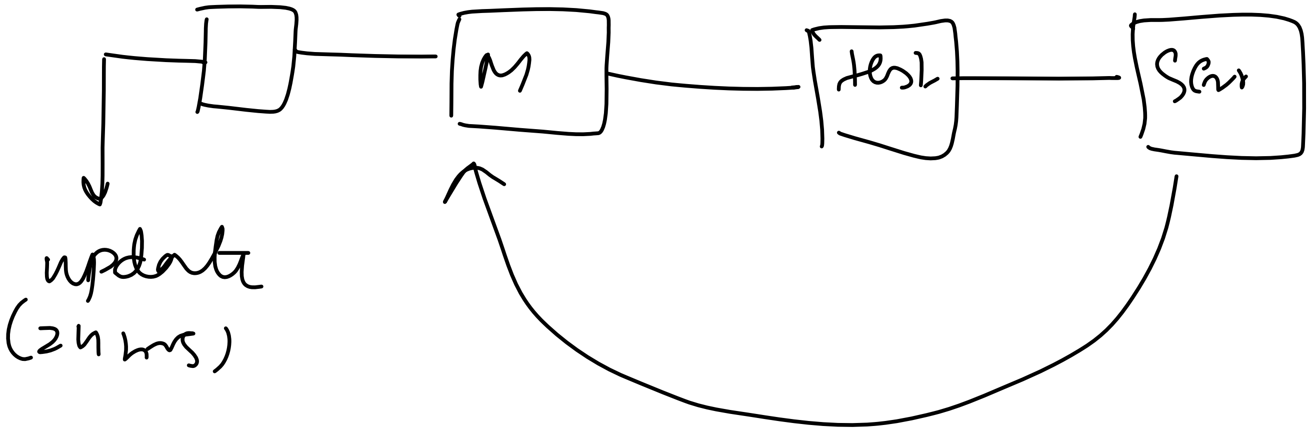
2. Batch/Offline ML

Wednesday, March 17, 2021 5:31 PM



3. The problem with Batch Learning

Wednesday, March 17, 2021 5:47 PM



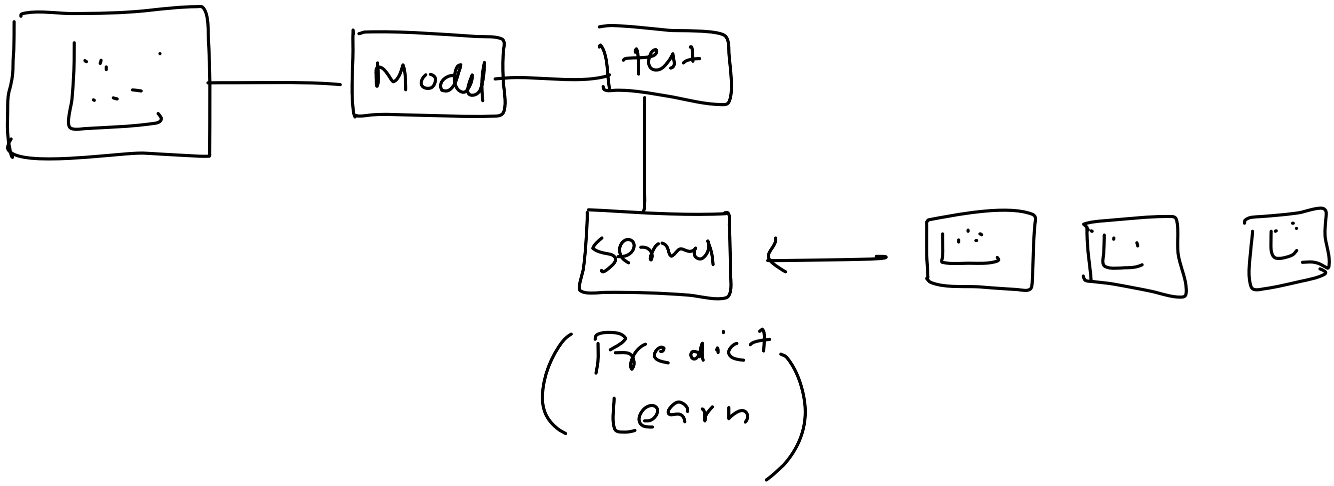
4. Disadvantages of Batch ML

Wednesday, March 17, 2021 5:32 PM

1. Lots of Data
2. Hardware Limitation
3. Availability

1. Online Machine Learning

Thursday, March 18, 2021 4:27 PM



2. When to use?

Thursday, March 18, 2021 4:33 PM

1. Where there is a concept drift
2. Cost Effective
3. Faster solution

3. How to implement?

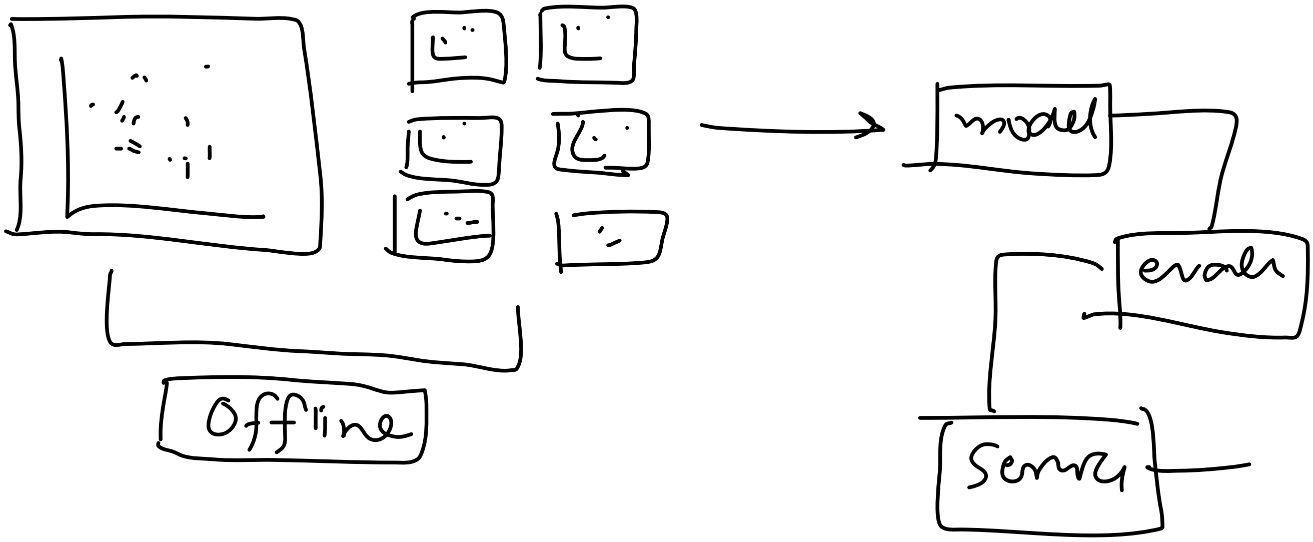
Thursday, March 18, 2021 4:28 PM

4. Learning Rate

Thursday, March 18, 2021 4:28 PM

5. Out of Core Learning

Thursday, March 18, 2021 4:28 PM



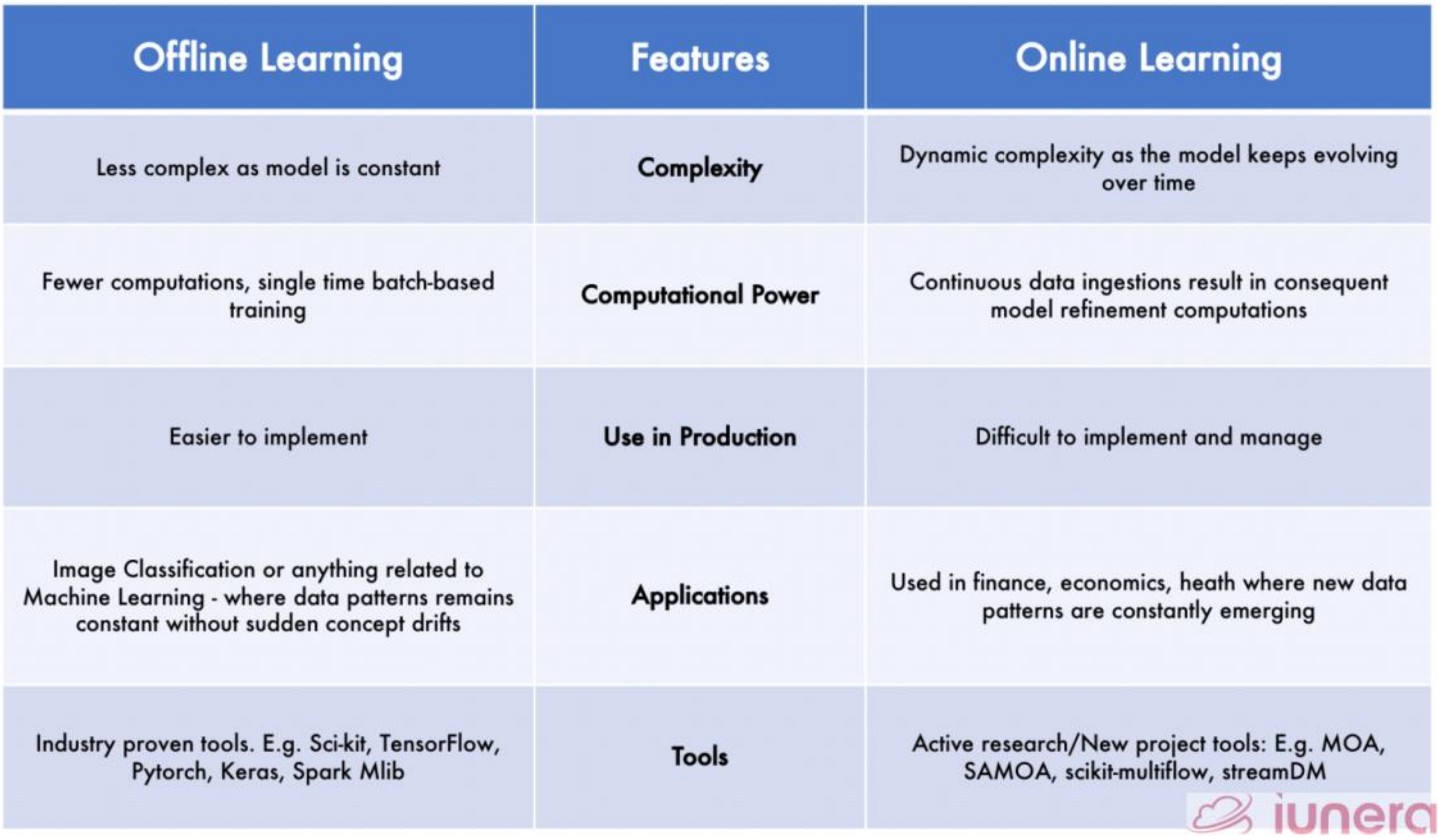
6. Disadvantage

Thursday, March 18, 2021 4:29 PM

1. Tricky to use
2. Risky

7. Batch Vs Online Learning

Thursday, March 18, 2021 4:29 PM

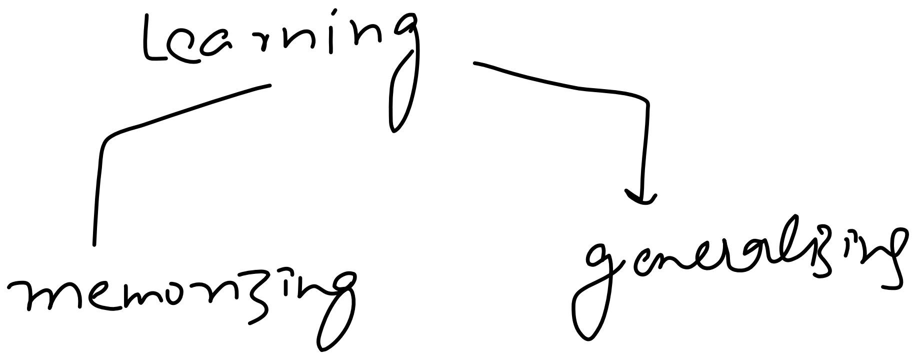


Offline Learning	Features	Online Learning
Less complex as model is constant	**Complexity**	Dynamic complexity as the model keeps evolving over time
Fewer computations, single time batch-based training	**Computational Power**	Continuous data ingestions result in consequent model refinement computations
Easier to implement	**Use in Production**	Difficult to implement and manage
Image Classification or anything related to Machine Learning - where data patterns remains constant without sudden concept drifts	**Applications**	Used in finance, economics, heath where new data patterns are constantly emerging
Industry proven tools. E.g. Sci-kit, TensorFlow, Pytorch, Keras, Spark Mlib	**Tools**	Active research/New project tools: E.g. MOA, SAMOA, scikit-multiflow, streamDM

Image courtesy - https://www.iunera.com/kraken/fabric/simple-introduction-to-online-learning-in-machine-learning/

1. Instance Vs Model Based Learning

Friday, March 19, 2021 4:05 PM

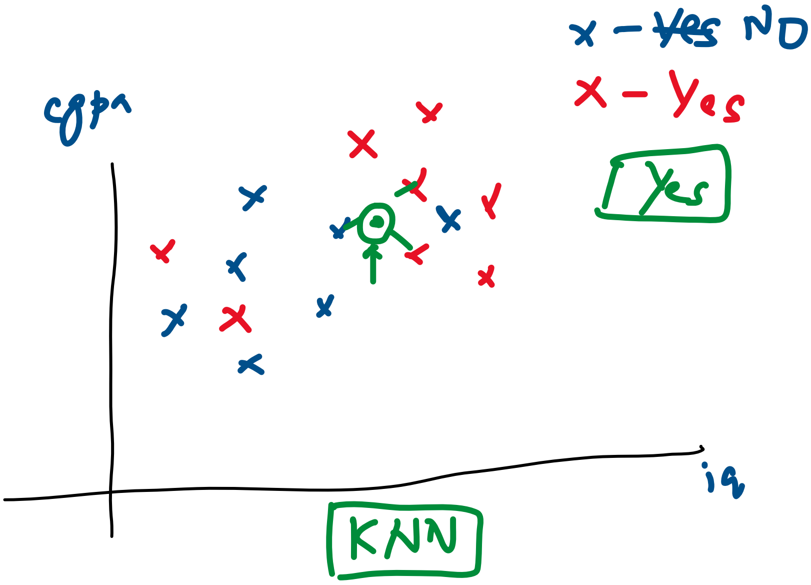


2. Instance Based

Friday, March 19, 2021 4:06 PM

iq	cgpa	placement Y/N
80	8	Y
70	7	N

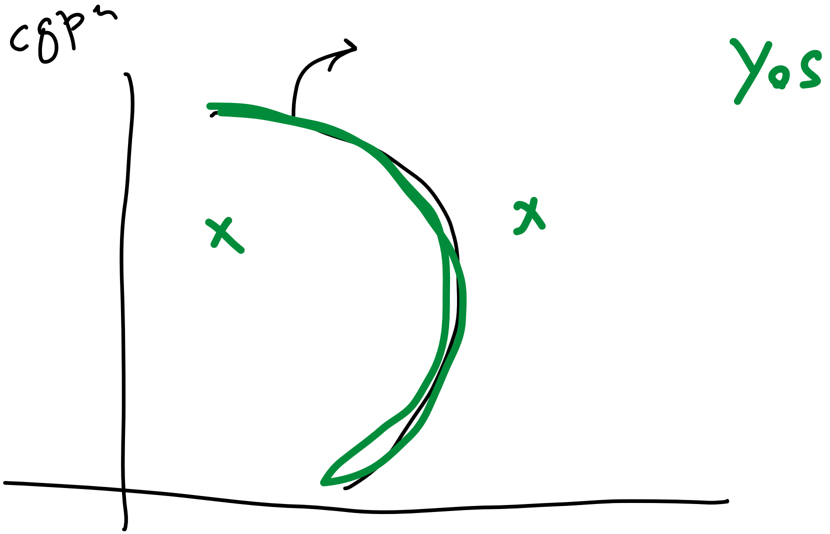
$\boxed{7.5,\ 103}$



3. Model Based

Friday, March 19, 2021 4:06 PM

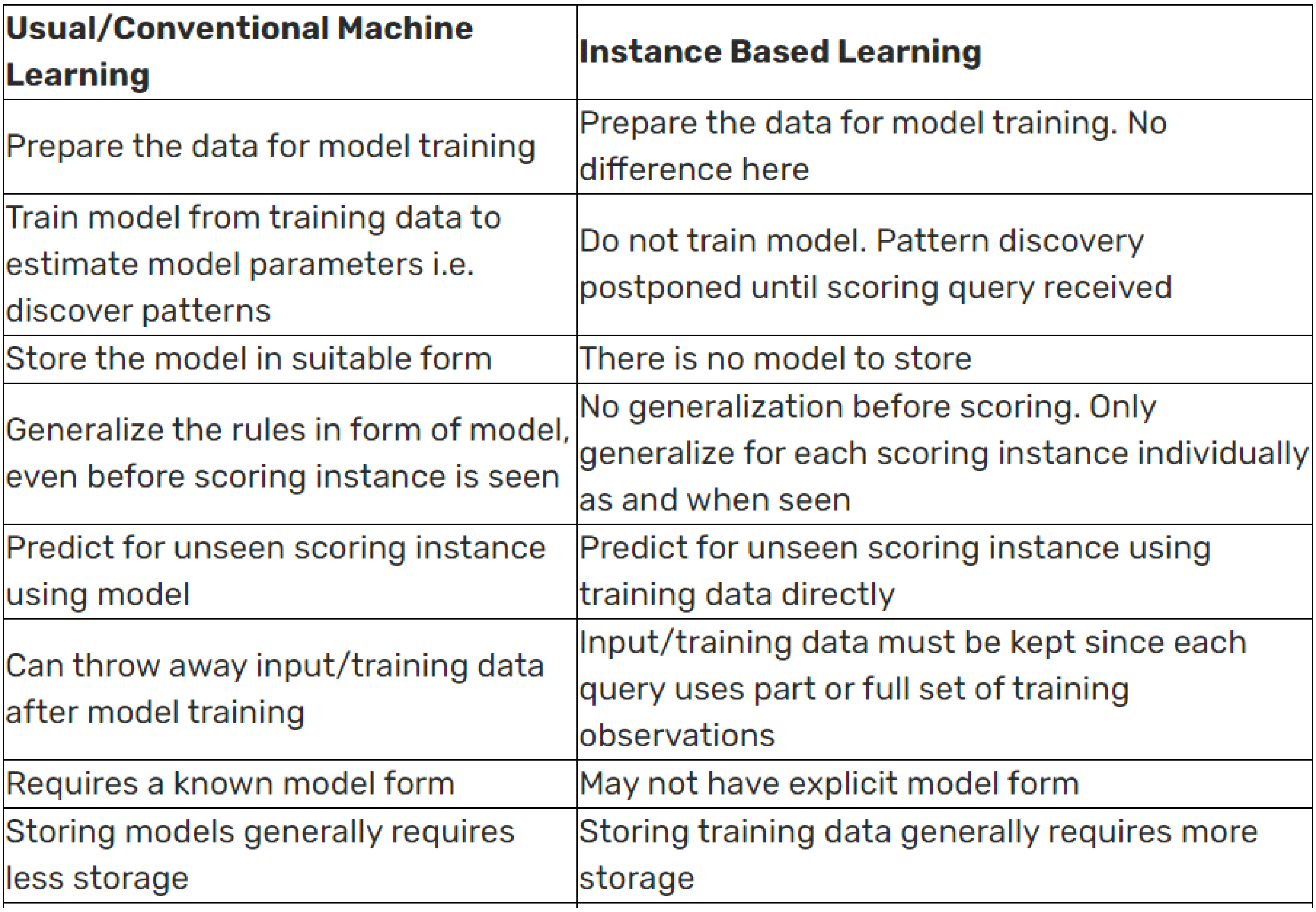
iq | cgpa | placement



LR / LoR | DT

4. Differences

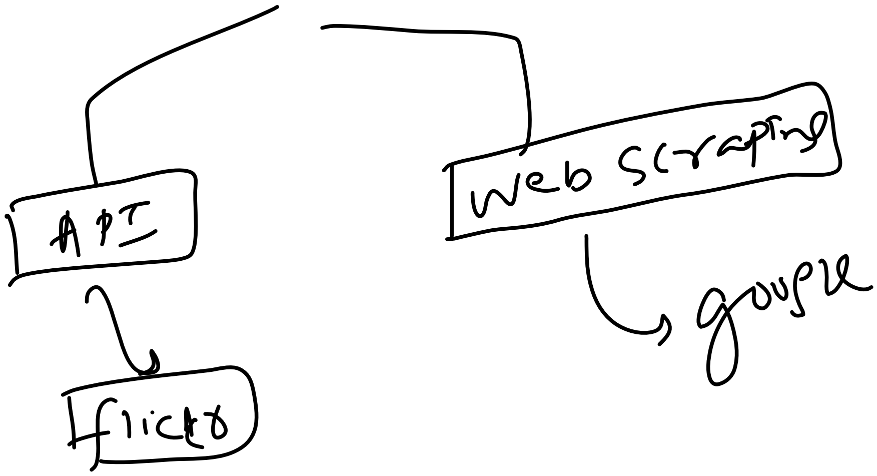
Friday, March 19, 2021 4:06 PM



Usual/Conventional Machine Learning	Instance Based Learning
Prepare the data for model training	Prepare the data for model training. No difference here
Train model from training data to estimate model parameters i.e. discover patterns	Do not train model. Pattern discovery postponed until scoring query received
Store the model in suitable form	There is no model to store
Generalize the rules in form of model, even before scoring instance is seen	No generalization before scoring. Only generalize for each scoring instance individually as and when seen
Predict for unseen scoring instance using model	Predict for unseen scoring instance using training data directly
Can throw away input/training data after model training	Input/training data must be kept since each query uses part or full set of training observations
Requires a known model form	May not have explicit model form
Storing models generally requires less storage	Storing training data generally requires more storage

1. Data Collection

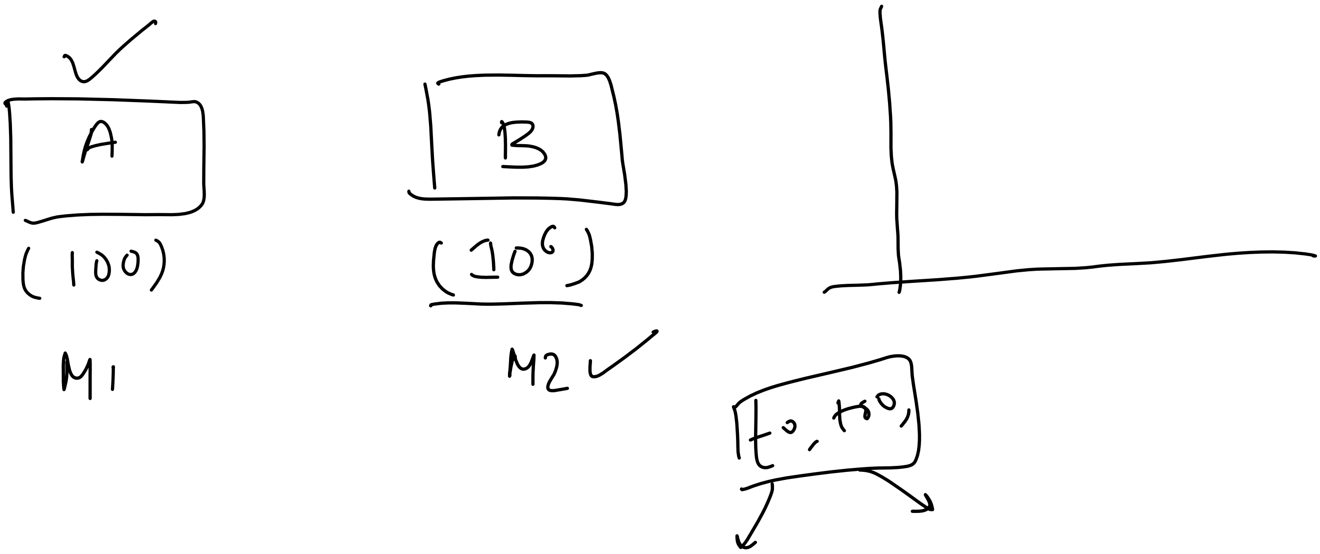
Saturday, March 20, 2021 5:59 PM



2. Insufficient Data/Labelled Data

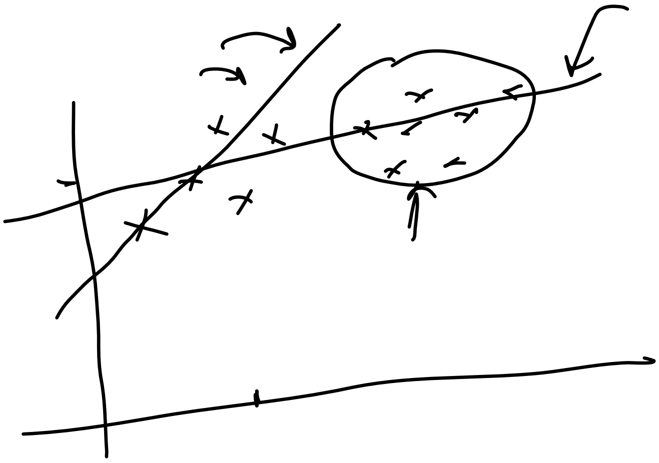
Saturday, March 20, 2021 6:00 PM

NLP



3. Non Representative Data

Saturday, March 20, 2021 6:00 PM



Sampling Noise

Sampling bias

4. Poor Quality Data

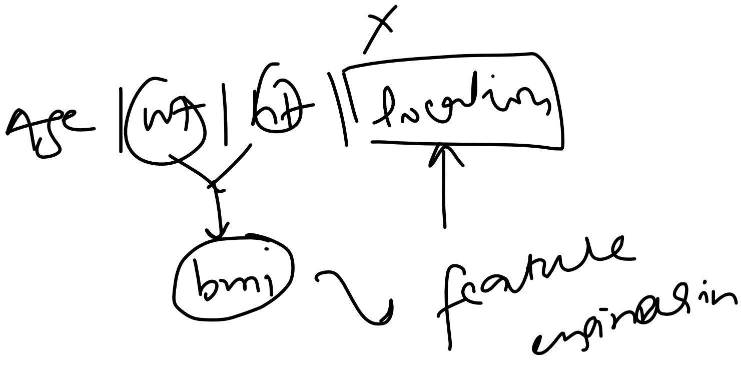
Saturday, March 20, 2021 6:00 PM

60%

5. Irrelevant Features

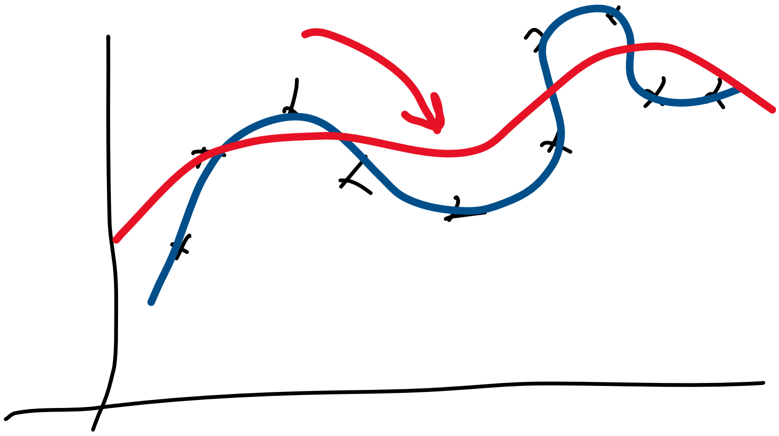
Saturday, March 20, 2021 6:00 PM

Garbage In
Garbage out



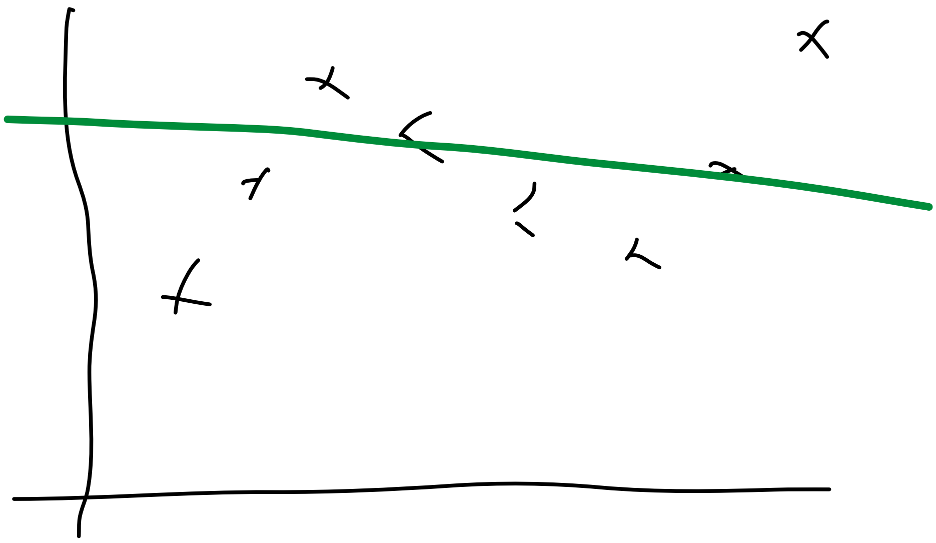
6. Overfitting

Saturday, March 20, 2021 6:01 PM



7. Underfitting

Saturday, March 20, 2021 6:01 PM

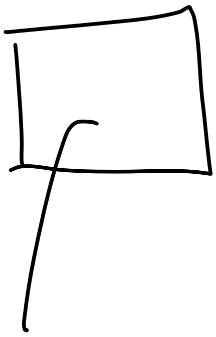


8. Software Integration

Saturday, March 20, 2021 6:01 PM

9. Offline Learning/ Deployment

Saturday, March 20, 2021 6:01 PM



10. Cost Involved

Saturday, March 20, 2021 6:01 PM

DevOps

MLOps

1. Retail - Amazon/Big Bazaar

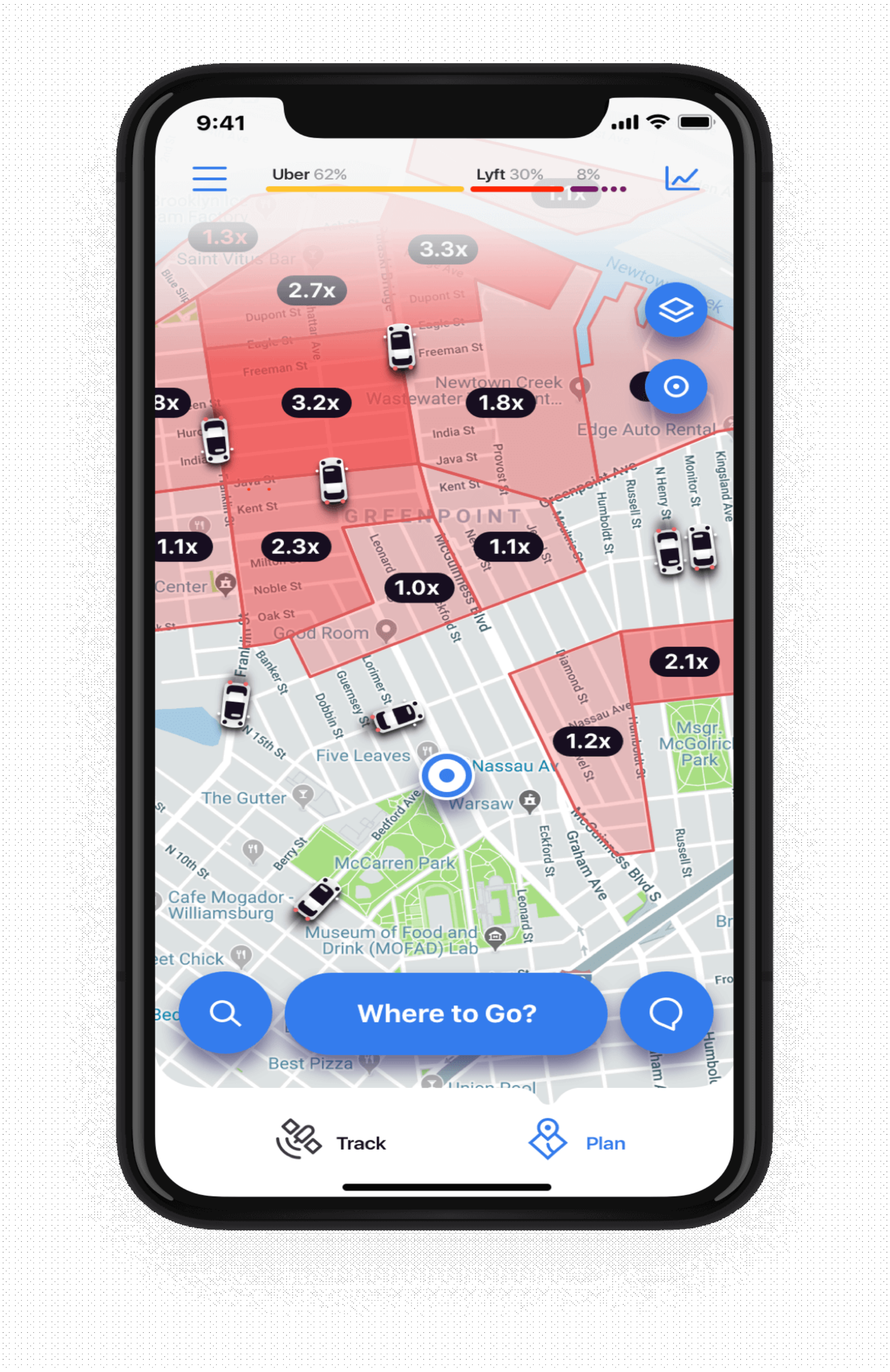
Monday, March 22, 2021 6:07 PM

2. Banking and Finance

Monday, March 22, 2021 6:07 PM

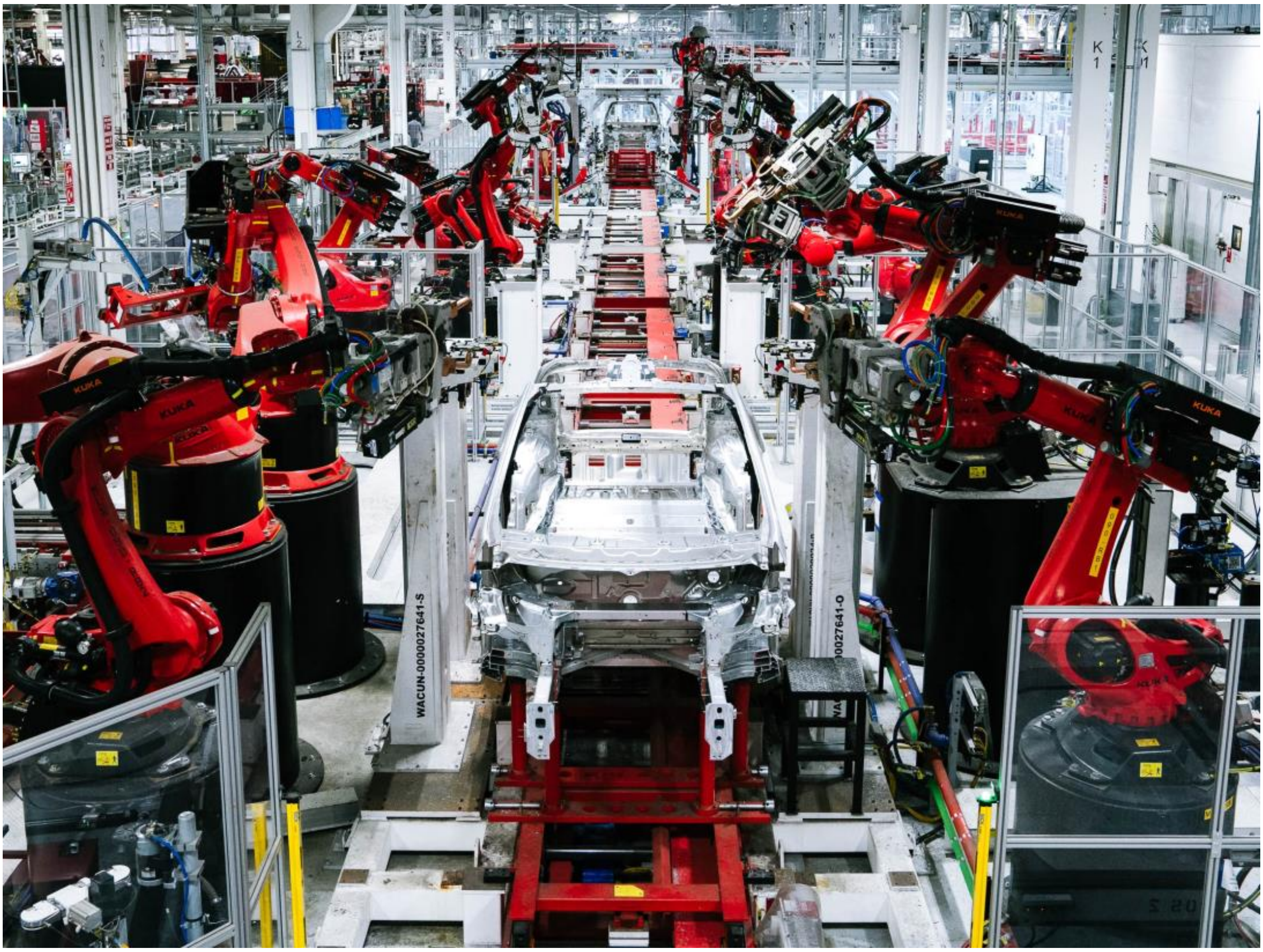
3. Transport - OLA

Monday, March 22, 2021 6:07 PM



4. Manufacturing - Tesla

Monday, March 22, 2021 6:08 PM



5. Consumer Internet - Twitter

Monday, March 22, 2021 6:08 PM

Machine Learning Development Life Cycle(MLDLC/MLDC)

Tuesday, March 23, 2021 12:09 PM

SDLC

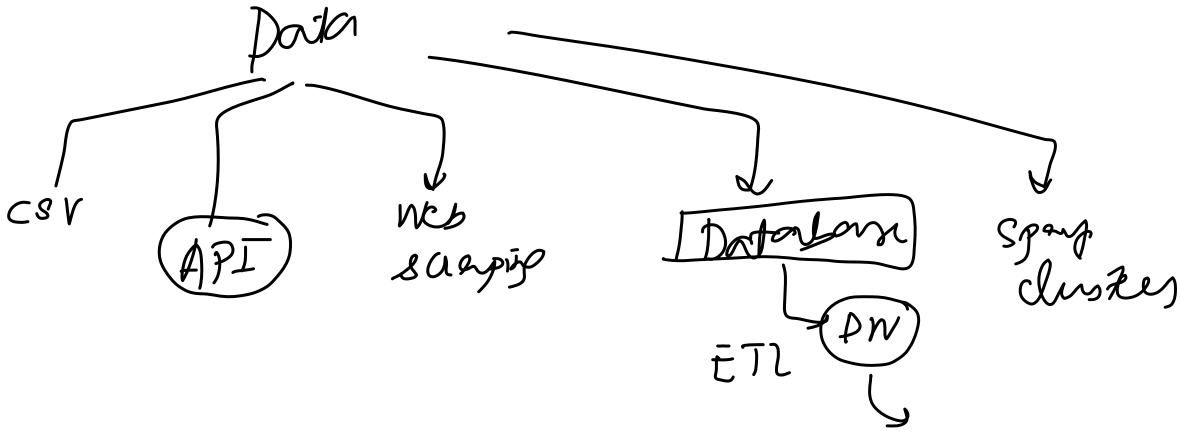
ML DLC

1. Frame the Problem

Tuesday, March 23, 2021 12:10 PM

2. Gathering Data

Tuesday, March 23, 2021 12:11 PM



3. Data Preprocessing

Tuesday, March 23, 2021 12:11 PM

→ Remove duplicates

→ Remove missing val

→ Outliers

→ Scale

4. Exploratory Data Analysis

Tuesday, March 23, 2021 12:11 PM

Vizs

Univariate / Biavariate

Outlier detection

Imbalance →

5. Feature Engineering and Selection

Tuesday, March 23, 2021 12:12 PM

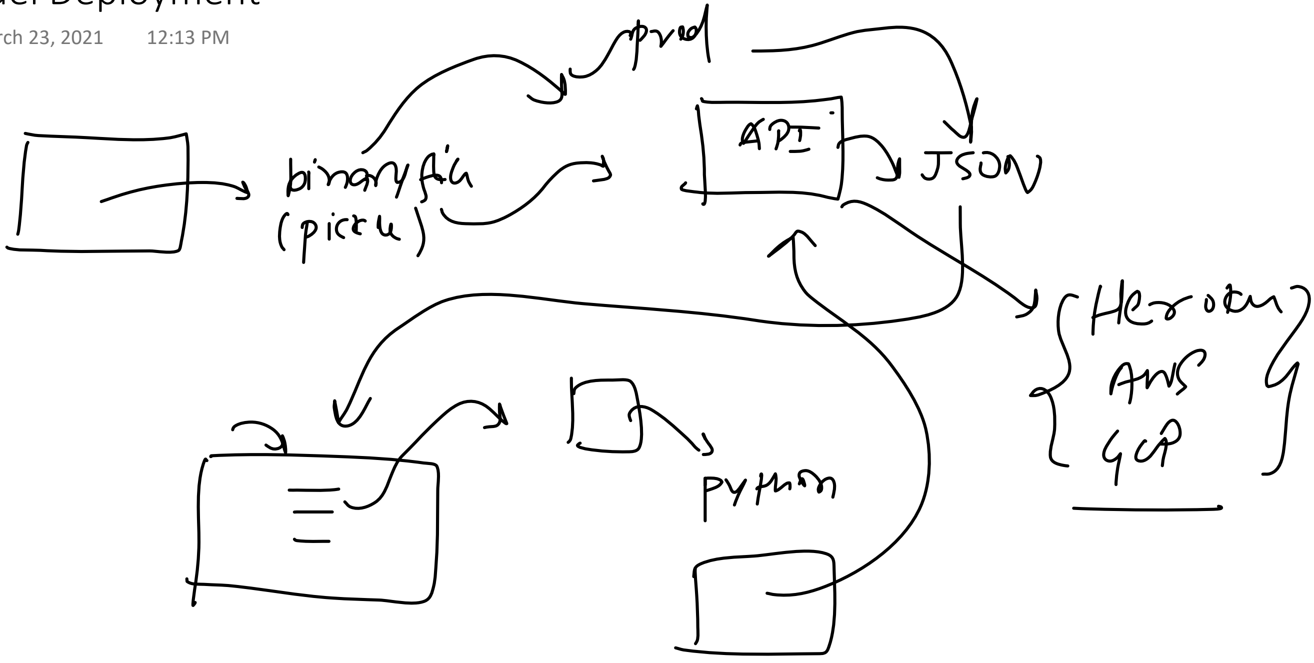
6. Model Training,Evalation and Selection

Tuesday, March 23, 2021 12:12 PM

Ensemble learning

7. Model Deployment

Tuesday, March 23, 2021 12:13 PM



8. Testing

Tuesday, March 23, 2021 12:14 PM

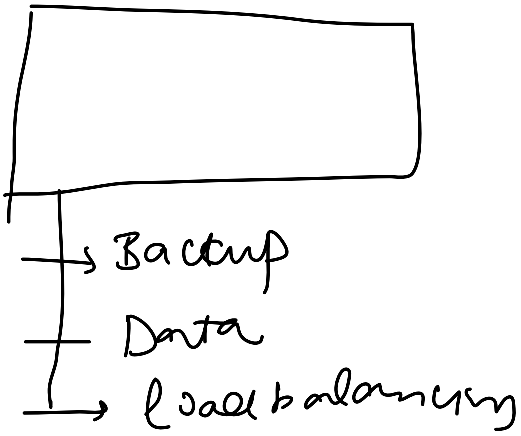
A /B testing

9. Optimize

Tuesday, March 23, 2021 12:15 PM

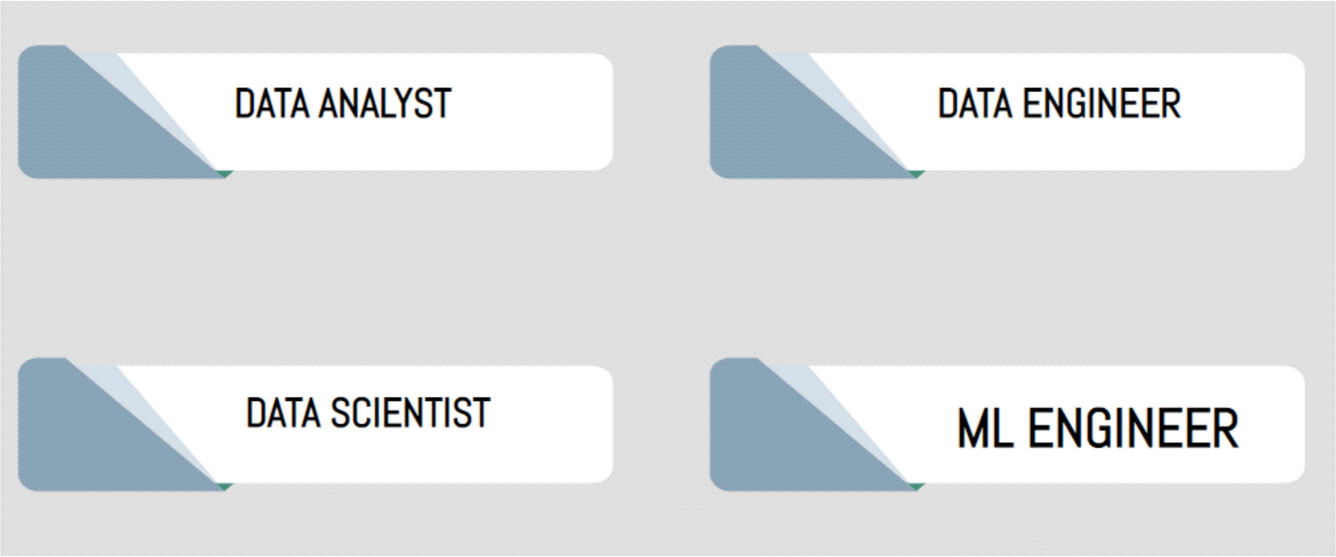
Retrain

Rotting



1. Various Data Based Job Roles

Wednesday, March 24, 2021 1:25 PM

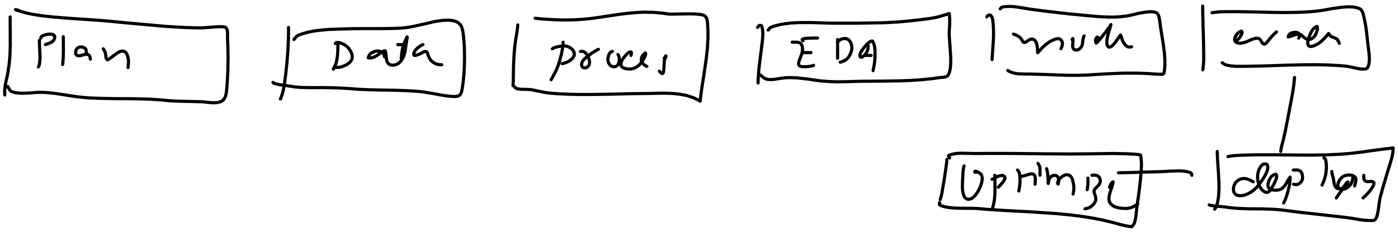


DATA ANALYST

DATA ENGINEER

DATA SCIENTIST

ML ENGINEER

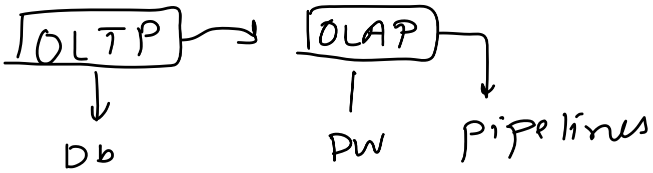


1. Data Engineer

Wednesday, March 24, 2021 1:25 PM

Job Roles

- Scrape Data from the given sources.
- Move/Store the data in optimal servers/warehouses.
- Build data pipelines/APIs for easy access to the data.
- Handle databases/data warehouses.



Skills Required

- Strong grasp of algorithms and data structures
- Programming Languages (Java/R/Python/Scala) and script writing
- Advanced DBMS’s
- BIG DATA Tools (Apache Spark, Hadoop, Apache Kafka, Apache Hive)
- Cloud Platforms (Amazon Web Services, Google Cloud Platform)
- Distributed Systems
- Data Pipelines

2. Data Analyst

Wednesday, March 24, 2021 1:26 PM

Responsibilities of a Data Analyst

- *Cleaning and organizing Raw data.*
- *Analyzing data to derive insights.*
- *Creating data visualizations.*
- *Producing and maintaining reports.*
- *Collaborating with teams/colleagues based on the insight gained.*
- *Optimizing data collection procedures*

Skills

- *Statistical Programming*
- *Programming Languages (R/SAS/Python)*
- *Creative and Analytical Thinking*
- *Business Acumen — Medium to High preferred*
- *Strong Communication Skills.*
- *Data Mining, Cleaning, and Munging*
- *Data Visualization*
- *Data Story Telling*
- *SQL*
- *Advanced Microsoft Excel*

3. Data Scientist

Wednesday, March 24, 2021 1:26 PM

“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician”.

4. ML Engineer

Wednesday, March 24, 2021 1:26 PM

Responsibilities

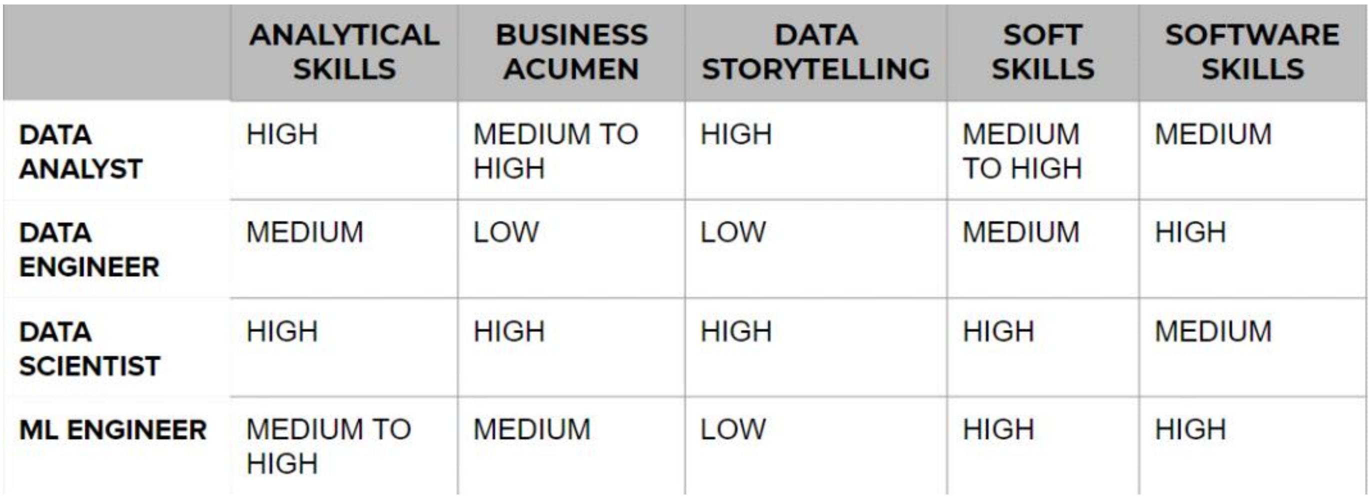
- Deploying machine learning models to production ready environment
- Scaling and optimizing the model for production
- Monitoring and maintenance of deployed models

Skills

- Mathematics
- Programming Languages (R/Python/Java/Scala mainly)
- Distributed Systems
- Data model and evaluation
- Machine Learning models
- Software Engineering & Systems design

5. Comparison

Wednesday, March 24, 2021 1:26 PM



	ANALYTICAL SKILLS	BUSINESS ACUMEN	DATA STORYTELLING	SOFT SKILLS	SOFTWARE SKILLS
DATA ANALYST	HIGH	MEDIUM TO HIGH	HIGH	MEDIUM TO HIGH	MEDIUM
DATA ENGINEER	MEDIUM	LOW	LOW	MEDIUM	HIGH
DATA SCIENTIST	HIGH	HIGH	HIGH	HIGH	MEDIUM
ML ENGINEER	MEDIUM TO HIGH	MEDIUM	LOW	HIGH	HIGH

1. What are Tensors

Thursday, March 25, 2021 4:44 PM

2. 0D Tensor/Scalar

Thursday, March 25, 2021 4:44 PM

(2) (3)

Scalars → $\boxed{\text{0D Tensor}}$ 0

3. 1D Tensor/Vector

Thursday, March 25, 2021 4:45 PM

[1, 2, 3, 4] → 1DTensor → nDim → 1

→ vector

→ 1Darray/array

1DTensor / Vector

→ 4D

Axis

2 Dim

No. of axes = rank = dim

[1, 2] → vector (2)

↳ 1DTensor

[0, 1, 2, 3]

4 scalars → vector

4. 2D Tensor/Matrices

Thursday, March 25, 2021 4:45 PM

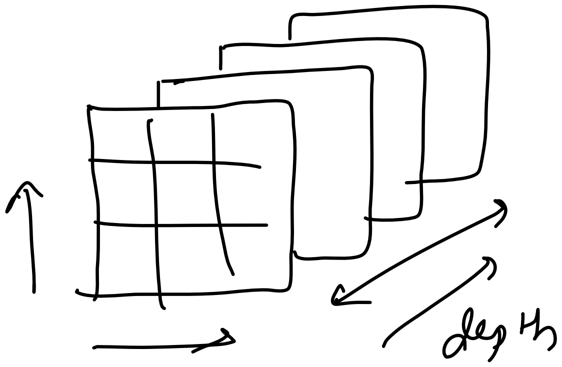
$[1,2,3]\quad [4,5,6]\quad [7,8,9]$

$$\begin{bmatrix}[1, 2, 3]\\ [4, 5, 6]\\ [7, 8, 9]\end{bmatrix} \rightarrow 2D$$

Rank = 2 = ndim

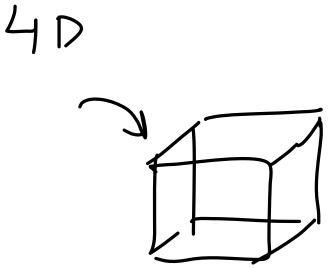
5. ND Tensors

Thursday, March 25, 2021 4:45 PM



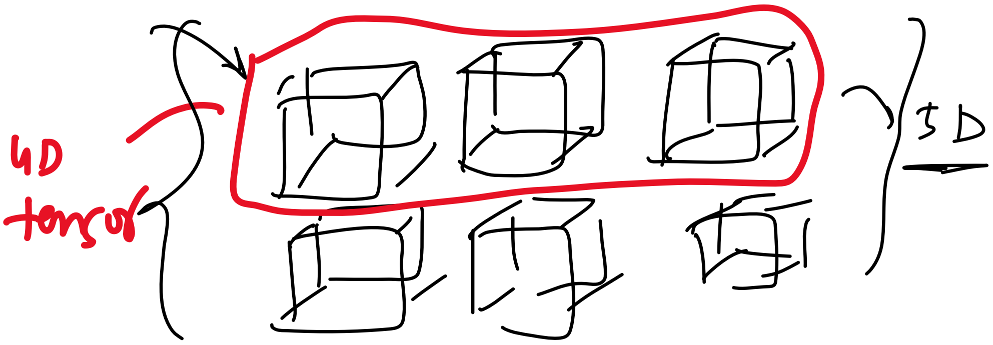
4 x 3 x 3

3D Tensor



0-5D

1D-5D

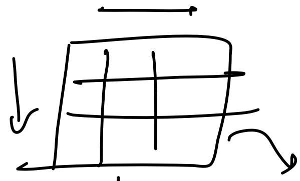


6. Rank, Axes and Shape

Thursday, March 25, 2021 4:45 PM

No. of axis = Rank = No. of dim

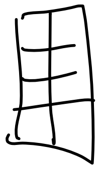
Shape



2,

(3,3)

9



(4,2)

8

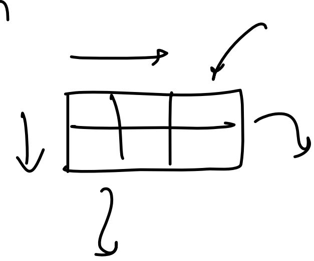
Size = 1

vector

[1,2,3]

(3,)

3



ndim rank

(2)

Shape (2,3)

6

Size of tensor

7. Example of 1D Tensors

Thursday, March 25, 2021 4:46 PM

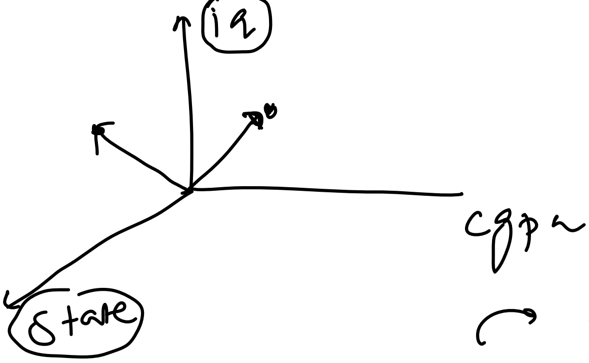
Students (10000)

WB = 0
KR = 1

cgpa	iq	state	placement
8.1	91	WB 0	1
[7.2	102	kr]	0
			1
			0
			1
			1
			0
			1

1D Tensor | Vector

[8.1, 91, 0] → 1D → vector → 3D



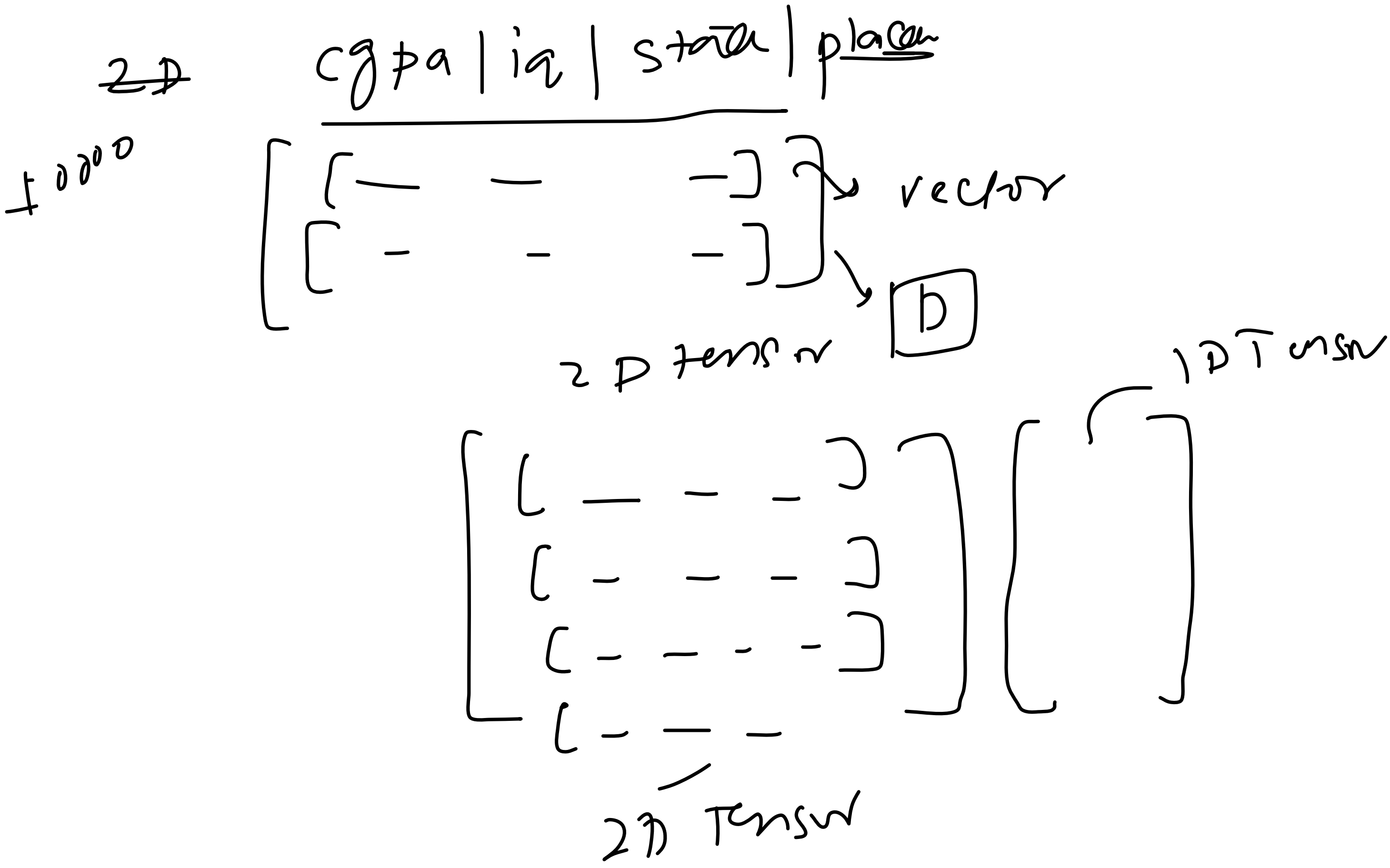
→ 1D Tensor

[1 0 1 0 1 1 0 1 . . . .]

→ 10000 no.

8. Example of 2D Tensors

Thursday, March 25, 2021 4:46 PM



9. Example of 3D Tensors

Thursday, March 25, 2021 4:46 PM

NLP

Hi Nitish

Hi Rahul

Hi Ankit

Hi	Nitish	Rahul	Ankit
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

24

[[[1,0,0,0],[0,1,0,0]]

[[1,0,0,0],[0,0,1,0]] → 3D Tensor

[[1,0,0,0],[0,0,0,1]]]

(3,2,4)

Timeseries Data (2)

Highest | lowest

Day 1 — —

Day 2 — —

Day 365 — —

365

[365,2] → 10 years

→ 2D → 10 → (10,365,2) ↳ 3D Tensor

365 → time axis

10. Example of 4D Tensors

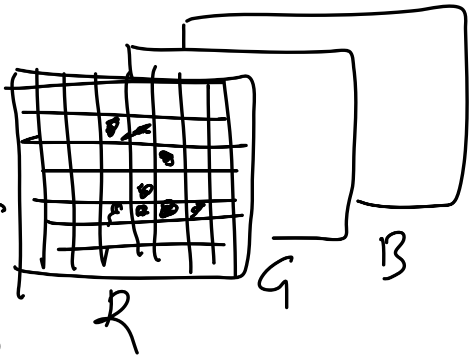
Thursday, March 25, 2021 4:47 PM

images → CV

$(3, \underline{1200}, \underline{800})$

↓

3D Tensor



(1200, 800)

→ 50 color images

(50, 3, 1200, 800)

↳ 4D Tensor

12. Example of 5D Tensors

Thursday, March 25, 2021 4:47 PM

videos

↳ frames

60 sec → 4 videos

↳ 30 fps

↳ 480p → 480 × 720 (3 channels)

(4, (1800, 480, 720, 3))

↑ 5D tensor

↳ float32

4 (60sec) videos

↳ 27 GB

↳ video

↳ MKV → mpeg mp4

1. Installing Anaconda

Friday, March 26, 2021 5:40 PM

2. Jupyter Notebook Intro

Friday, March 26, 2021 5:40 PM

3. Virtual Env

Friday, March 26, 2021 5:40 PM

4. Using Kaggle

Friday, March 26, 2021 5:41 PM

5. Using Google Colab

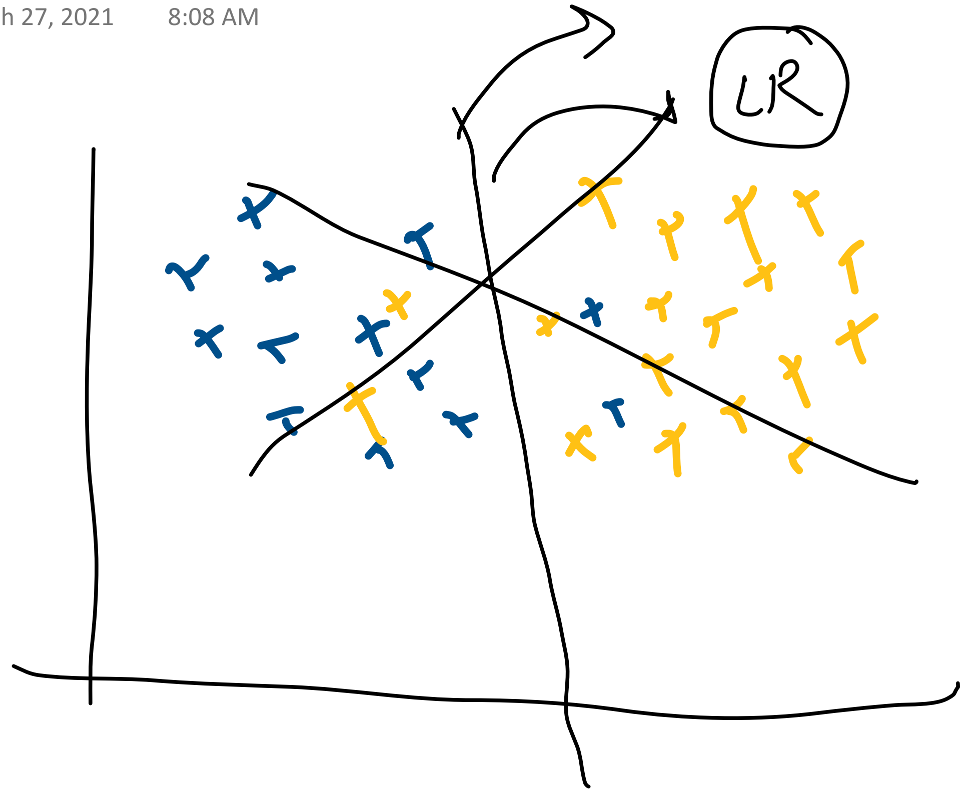
Friday, March 26, 2021 5:41 PM

6. Running Kaggle Data on Google Colab

Friday, March 26, 2021 5:41 PM

End to End Example

Saturday, March 27, 2021 8:08 AM



1. Business Problem to ML Problem

Monday, March 29, 2021 7:29 AM

Netflix

Churn rate ↓

4%

3.75

3.5

100 → 2%

98

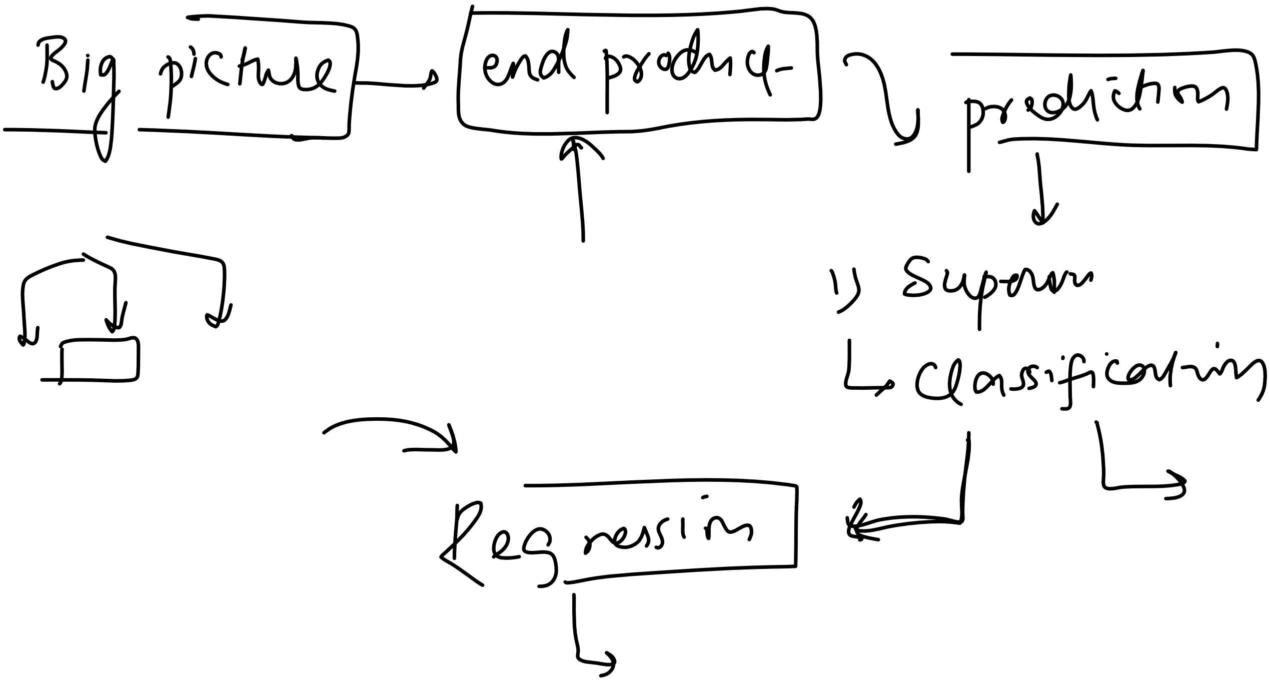
2%

Increase revenue

4% → 3.75%

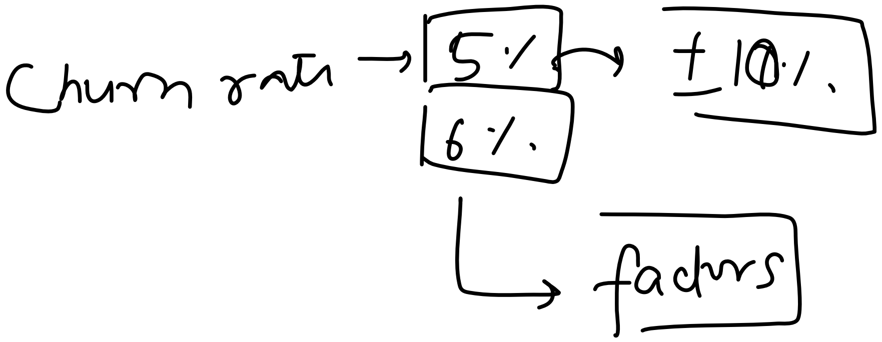
2. Type of Problem

Monday, March 29, 2021 7:30 AM



3. Current Solution

Monday, March 29, 2021 7:30 AM

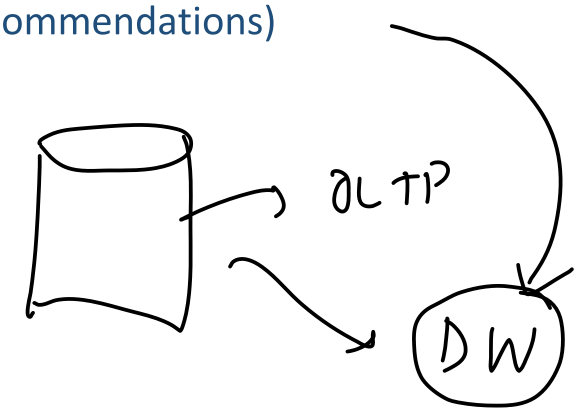


4. Getting Data

Monday, March 29, 2021 7:30 AM

1. Watch time
2. Search but did not find
3. Content left in the middle
4. Clicked on recommendations(order of recommendations)

Data engineer

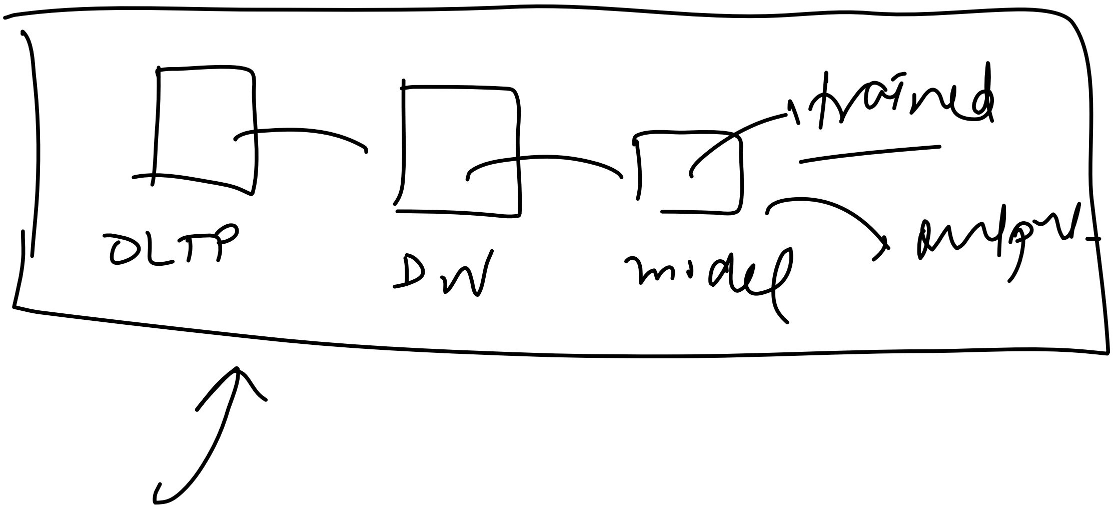


5. Metrics to measure

Monday, March 29, 2021 7:30 AM

6. Online Vs Batch?

Monday, March 29, 2021 7:31 AM



7. Check Assumptions

Monday, March 29, 2021 7:31 AM