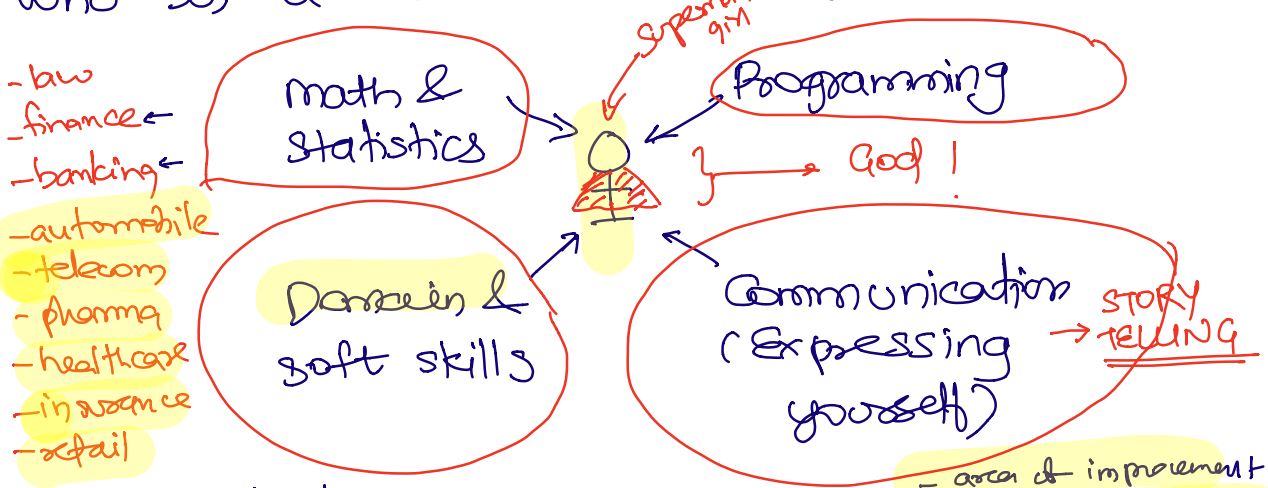
- Introduction to Data Science
- Introduction to Bigdata
- Setting up Python on your machine

What is Data Science?

Set of Practices that enable Data Scientist to get INFORMED DECISIONS

Who is a Data scientist?



Case-study:

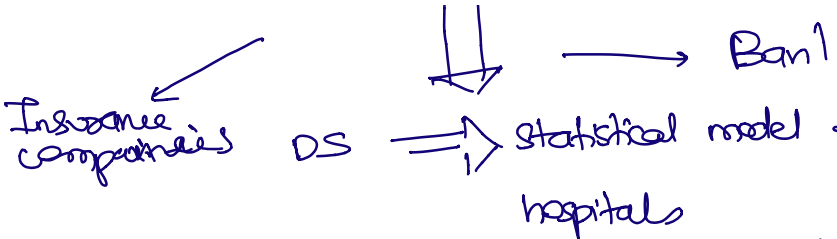
Mac Donalds

- area of improvement
- which items mostly sell?
- freq visit of customers (customer seg)
- supply v/s demand w.r.t location
- pricing of menu
- which items are bought together

obses

mobile_no	bank_card	merchant_name	location	what you ordered

1 TB file



A hospital chain XYZ group wants to understand in which area of Mumbai he shall open a new hospital.?

→ finding strategic locations in the given area to open hospitals

→ finding strategic locations to install a mobile tower for better service.

→ Gym fitness centers

DATA PRODUCT

→ Senator vs Mark Zukerberg (FB) } Data Privacy.

→ Senator vs Sundar P. (Google) }

Is my data exposed?

Are you able to download my files from Gdrive?

ARE you keeping a realtime watch on me?

Are you spying me?

Are you capable to spy me

Are you able to read my email?

COMPROMISING USER'S PRIVACY

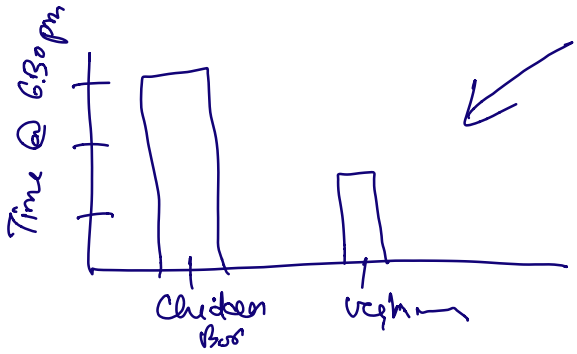
FB → } ~~JAIL~~ NO!

Google → }

Google, FB } never sell/share/use your data directly } → they sell STATISTICS

Age 18-35 } → Chicken Burger } → Mumbai, Andheri

18-22 } → Veg Burger } → Mumbai, Dadar



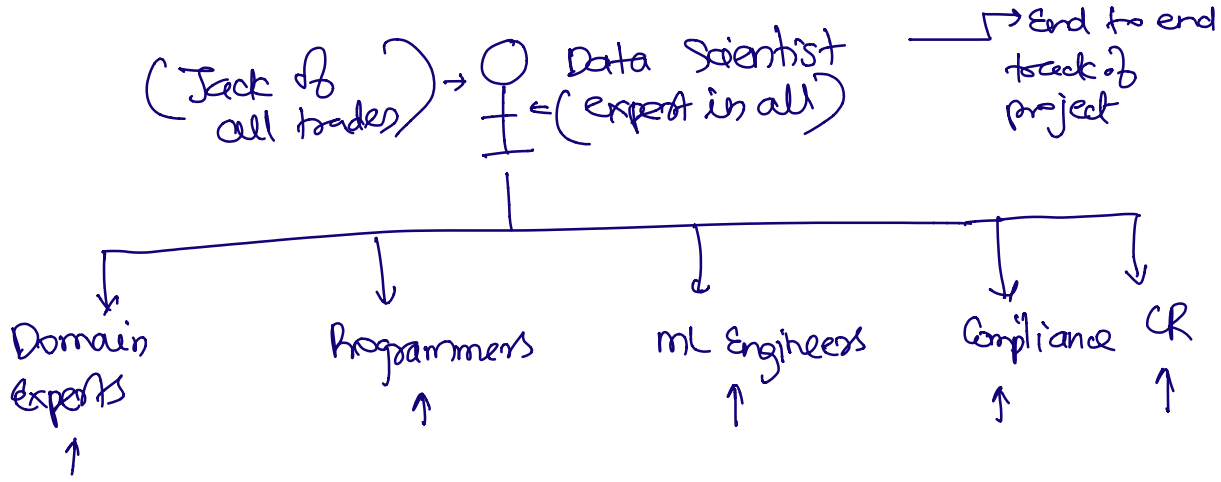
FB → Data

Data Product
Healthcare
prospective hospital
location
→ model

Data Product
Healthcare
number of patients
in July 2019
→ model

Data Product
University
no. of student who
Engineering !
→ model

file | equation | algo.



Set of Practices (Data Science) (Guidelines)

① Understand the use-case. (Are you the best fit for this use-case)

② Relate your use-case with domain for creative insights.

(Do you have enough domain knowledge?)
-OR-
(Can you hire someone with domain knowledge?)

③ Gather relevant data / Data Product for the use-case.

→ Is data available with customer?

→ Is customer willing to give plain data?

→ Is the data-sharing satisfies all compliance standards? HIPAA | ISO 27001 | ISO 9001:2015

④ Apply Statistical analysis over the data.

⑤ Apply ML/DL/DM algo. to create data product (model)

⑥ viz. the result.

⑦ Present your results/findings to stakeholders.

Bigdata

What is Bigdata?

Bigdata is a TERM. Its not a tool, technique, technology.

Bigdata is all about the ability of your s/w, framework, architecture (s/w, h/w) to HANDLE the data.

Handle refers to → LOAD
→ PROCESS
→ STORE

} if your app/ s/w/arch. fails you are facing BIGDATA PROBLEM.

BIGDATA IS A PROBLEM!

504 MB → 20120Nobody.txt

Notepad → It tried its best but FAILED TO LOAD the data.

∴ for NOTEPAD, its a BIGDATA

Wordpad → It tried, had some latency, but LOADED THE DATA

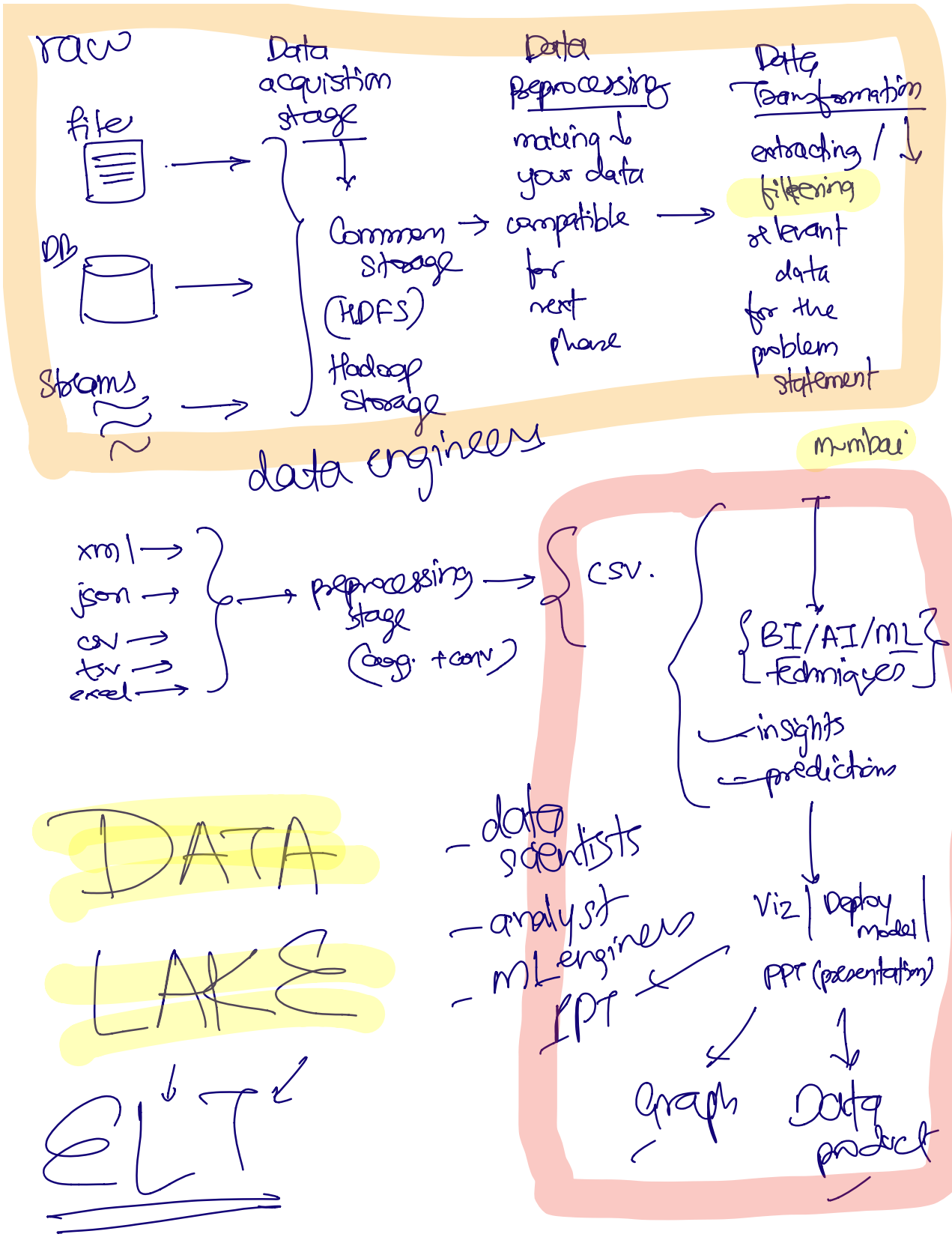
∴ for WORDPAD, its a NORMAL DATA

Bigdata → Hadoop { Java based distributed storage & processing framework.

my capability to store and process the data is directly proportional to my underlying hw's capability.

→ Apache Hadoop
→ Apache Spark
→ Apache Storm

} Data processing
Data Transformation


raw
file
DB
Streams
Data acquisition stage
Common Storage (HDFS) Hadoop Storage
Data Preprocessing
making your data compatible for next phase
Data Transformation
extracting / filtering relevant data for the problem statement
data engineers
xml
json
csv
tsv
excel
preprocessing stage (agg. + conv)
csv.
Mumbai
BI/AI/ML Techniques
insights
predictions
Viz
Deploy model
PPT (presentation)
Graph
Data product
- data scientists
- analyst
- ML engineers
PPT
DATA
LAKE
ELT

Data Engineers ⟹ data acq, data preprocessing, data conversion, data agg, data transformation

→ Hadoop, Spark, Storm, Kafka, Hive, Java, Python, Elasticsearch

Data Analyst ⟹ create statistical inference out of the data → (Excel (advanced), Tableau, SSRS, SAS, R, Python)

Business Analyst ⟹ create statistical inference out of the data w.r.t. domain.

ML Engineers ⟹ the one who make data product.

(Python → numpy, scipy, pandas, sklearn, tensorflow, keras, nltk, spacy, beautiful soup, opencv, etc)

Data Scientists ⟹ THE GOD

Handling the above team + Interacting with Customers.

Master/Jack of all trades on all tools & technologies with expertise in domain knowledge & Good presentation & soft skills.

① Python

→ Basics of Python

→ Numpy / Scipy

→ Pandas

→ matplotlib & seaborn (Viz)

→ sklearn for ML

② Statistics

③ Exploratory Data Analysis (EDA)

④ Machine Learning (limited)

⑤ Bigdata using PySpark (basics)

Which programming language I must learn to become Analyst / DS ?

① Python 3
② R
③ Scala
④ Go lang

SQL

Excel techniques (Business)

Jupyter → Python Hands-on exercises.

How to create a local labs environment in your machine?

Anaconda Python