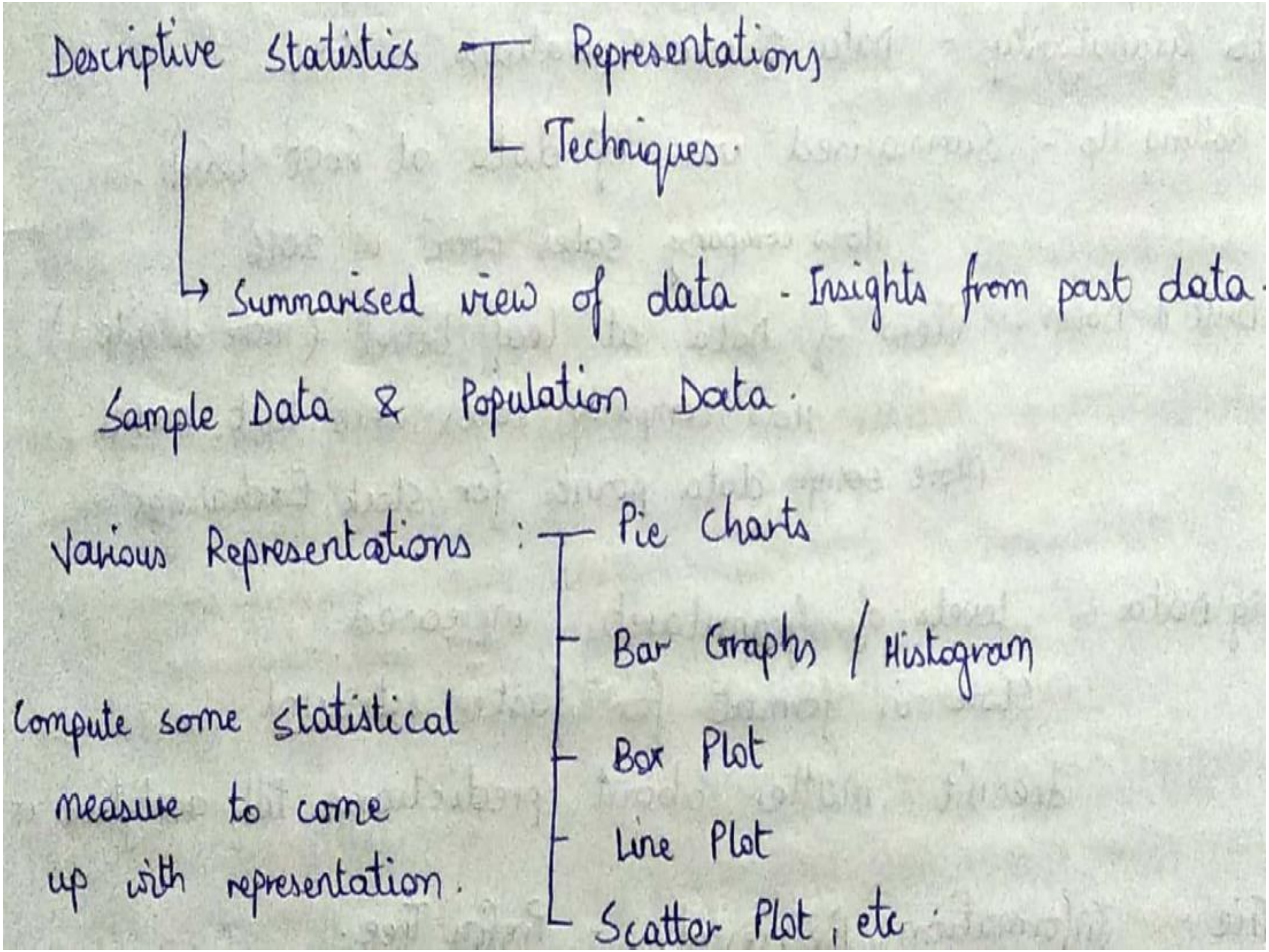
Introduction to Data Analytics

- Objective of Automation - Convert Data to Information
- Conversion – "Processing" – Various Courses of CSE
- Data is Raw Input ; Information – Processed
- Power of Excel over C (for end user!) – Data Processing Capabilities.
- Power of DBMS (MYSQL) over C (for end user!) – Data Management Tasks
- Equivalent File Handling approach – u can achieve the same as in DBMS but tedious! And data integrity is not guranteed!
- DBMS – Information Retrieval – query required info (explicitly stored) from data

Introduction to Data Analytics

- Data Science / Data Mining / Data Analytics – different from DBMS – can u project info that is not explicitly stored in the data.
- Data Mining Literature – prefers the word Knowledge or Patterns for Such Hidden Info Extracted.
- Data Mining – typically referred as Knowledge Discovery in Databases (KDD).
- Machine Learning (its really gone too DEEP these days!!!)
- - Use Data to Answer Questions – Learn a Model from Data to Answer Questions (also treated as Prediction)

- References / Resource Materials:
- (1) Predictive Data Analytics – Data Mining Concepts & Techniques, Jiawei Han and M Kamber
- (2) Mining Massive Data Sets – Jeffrey Ullmann – full text is open on the web – legally! Free version.
- (3) Introduction to Data Analytics – NPTEL course – good for the descriptive part from the breadth perspective. – depth treatment we will refer other online resources which wud be shared.
- (4) FIMI resources – Frequent Itemset Mining Imlementaion repository (now the page has no new contributions…but was good point for FIM research work..
- (5) Pyspark / Hadoop for the Storage / Systems focus – manuals wud be shared at the respective point…
- (6) Rajiv Motwani (late) – Stanford Prof – Excellent Contributions in Data Mining



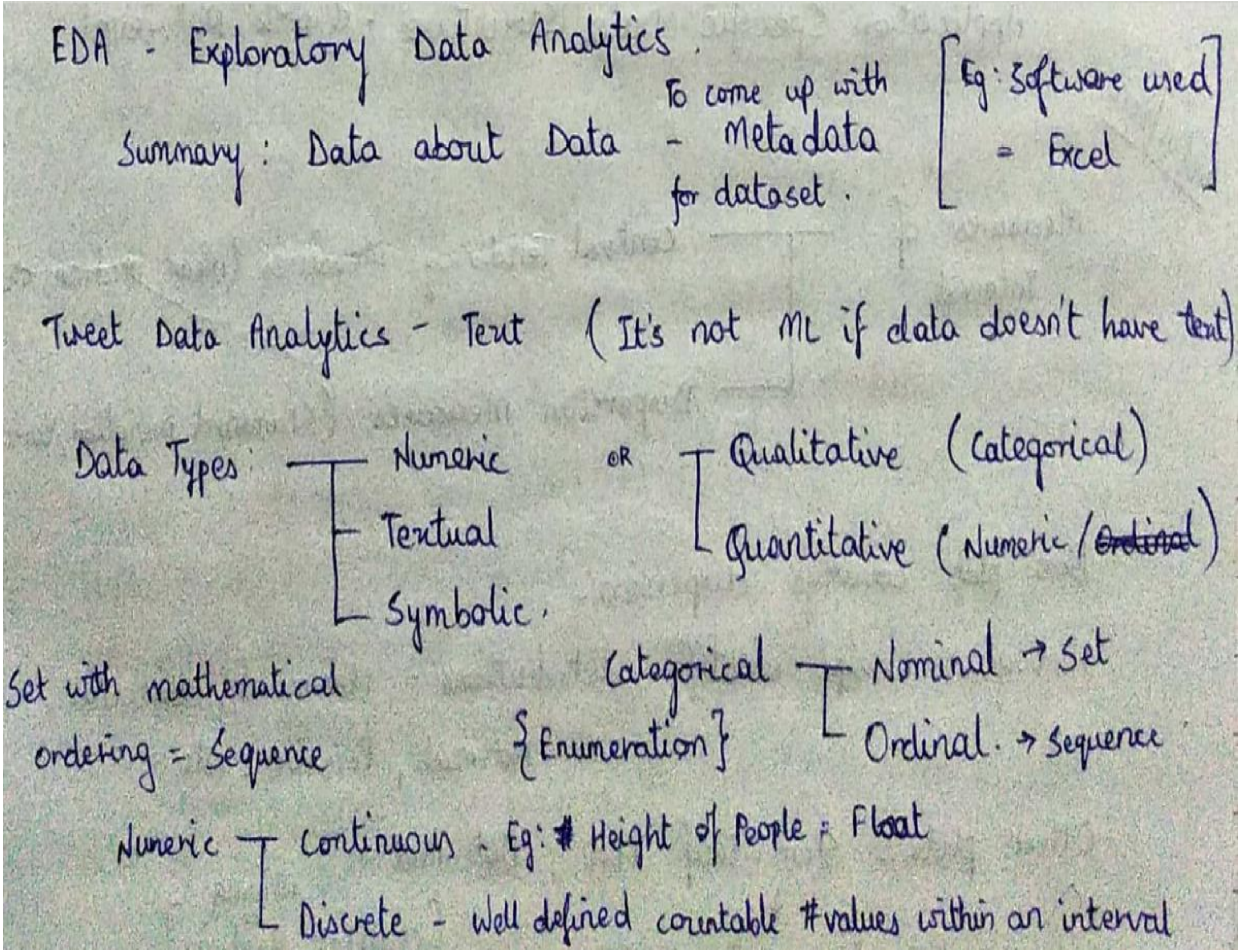
Descriptive Statistics
- Representations
- Techniques.

Summarised view of data - Insights from past data.

Sample Data & Population Data.

Various Representations :
- Pie Charts
- Bar Graphs / Histogram
- Box Plot
- Line Plot
- Scatter Plot, etc :

Compute some statistical measure to come up with representation.



EDA - Exploratory Data Analytics .

Summary : Data about Data - To come up with Metadata for dataset .

[Eg: Software used = Excel]

Tweet Data Analytics - Text (It's not ML if data doesn't have text)

Data Types : — Numeric

- Textual
- Symbolic .

OR

- Qualitative (Categorical)
- Quantitative (Numeric / ~~Ordinal~~)

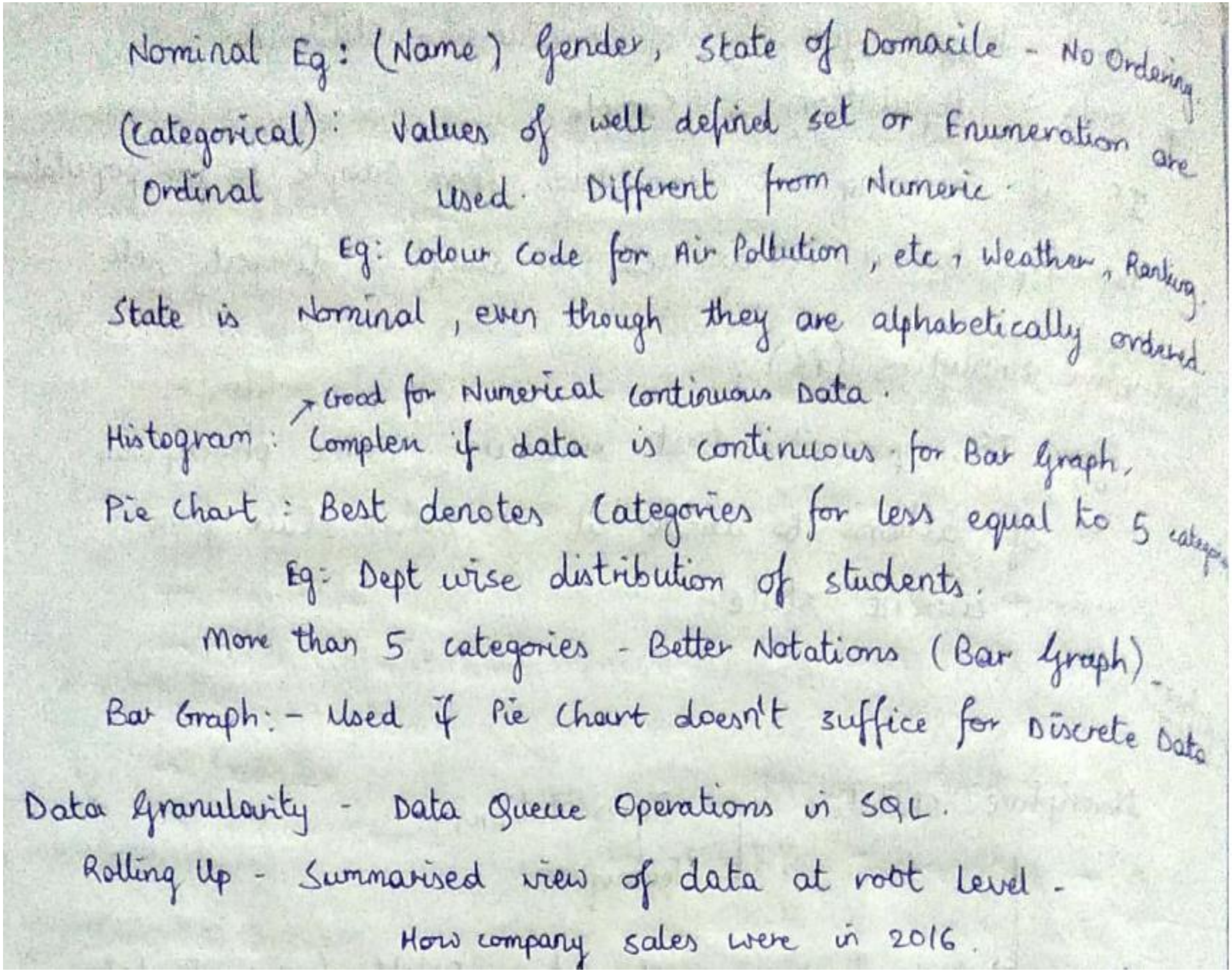
Set with mathematical ordering = Sequence

Categorical {Enumeration}

- Nominal → Set
- Ordinal . → Sequence

Numeric

- Continuous . Eg: # Height of People = Float
- Discrete - Well defined countable #values within an interval



Nominal Eg: (Name) Gender, State of Domacile - No Ordering.

(Categorical): Values of well defined set or Enumeration are

Ordinal used. Different from Numeric.

Eg: Colour Code for Air Pollution, etc, Weather, Ranking.

State is Nominal, even though they are alphabetically ordered.

↗ Good for Numerical continuous Data.

Histogram: Complex if data is continuous for Bar Graph,

Pie Chart: Best denotes Categories for less equal to 5 categ

Eg: Dept wise distribution of students.

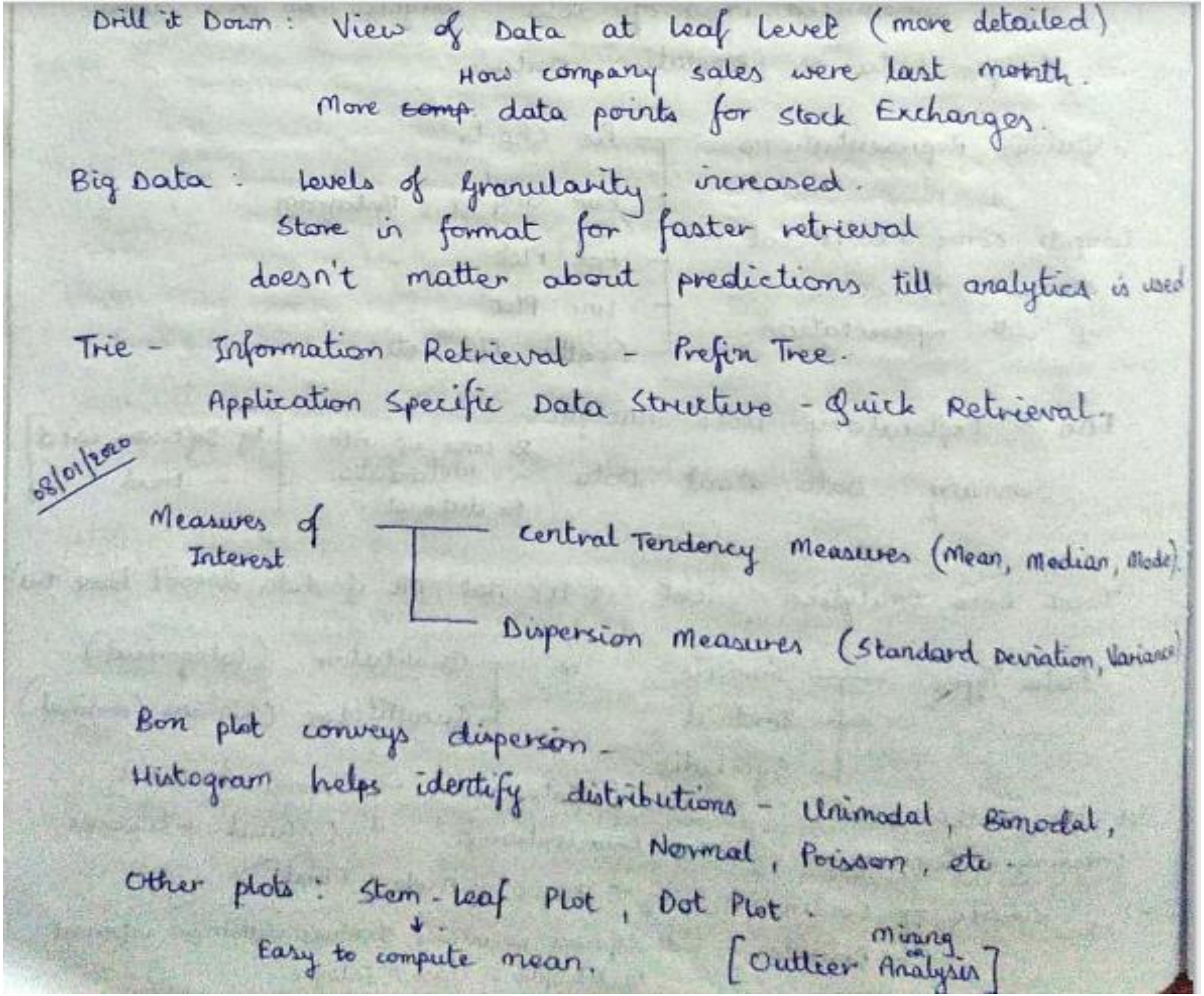
More than 5 categories - Better Notations (Bar Graph).

Bar Graph:- Used if Pie Chart doesn't suffice for Discrete Data

Data Granularity - Data Queue Operations in SQL.

Rolling Up - Summarised view of data at root level.

How company sales were in 2016.



Drill it Down : View of Data at leaf level (more detailed)

How company sales were last month.

More ~~comp~~ data points for Stock Exchanges.

Big Data : levels of granularity increased.

Store in format for faster retrieval

doesn't matter about predictions till analytics is used

Trie - Information Retrieval - Prefix Tree.

Application Specific Data Structure - Quick Retrieval.

08/01/2020

Measures of Interest

- Central Tendency Measures (Mean, median, Mode).
- Dispersion Measures (Standard Deviation, Variance)

Box plot conveys dispersion.

Histogram helps identify distributions - Unimodal, Bimodal, Normal, Poisson, etc.

Other plots : Stem-leaf Plot, Dot Plot.

↓

Easy to compute mean.

[Outlier Mining or Analysis]