


# 1. Discover the Problem

04 September 2022 10:31

## 1. Sources of Datasets

Public	Private	Personal
Open and free	Accessible to few people	Data that lies within you and your devices
Relative: Mostly out but hard to locate and hard to get the right data	No open == Paid	Call logs or music listening history or rating
 <ul style="list-style-type: none"><li>• Awesome public datasets</li><li>• Google dataset search</li><li>• Kaggle</li><li>• Data.gov</li><li>• Datameet</li></ul>	<ul style="list-style-type: none"><li>• Corporate</li><li>• Paid Datasets</li></ul>	<ul style="list-style-type: none"><li>• Mobile app datasets</li><li>• Personal Logging</li></ul>

## 2. Type of Datasets: (Continuum)

Structured	Semi-Structured	Unstructured
You know the schema and the information		We know practically nothing No metadata and No predefined schema = Photos
Table from database Spreadsheets Shapefiles: Maps	Documents(pdf or html) Messages and emails Container: ZIP or docx	Text Images Audio Video  <Deep Learning>

## 3. Types of Values

Categorical	Numerical	Composite
Allows you to do relatively fewer computation Infer by themselves Eg. color	Series of operation we can perform e.g. + / - *	Even more operations we can perform e.g. array list of numerical or numerical and categorical
Boolean Unordered: Color, cities,... Ordered (low, Med, High) Cyclical: Months or days Unstructured (Text, Binary)	Integers Real	Date/Time Spatial (Lat/long, Shapes) Structured (JSON, XML) Specialized (IP, currency)

## 4. Discovering hidden sources of data >> competitive advantage. Plus understanding of data >> Less time and effort in analysis