Preview:

Task:

* Pick 3 real world datasets
* Their unintended consequence
  + Can they be combined
  + What are they used for
* Characterize them in the sense of 5vs
  + Volume
  + Variety
  + Velocity
  + Veracity
  + value
* How are they collected
* Temporal and special coverage
* How are they structured
* Are they big data?
* More than 5 slides+ 7mim

Resource:

1. Datasets/dataset archive
   1. UCI机器学习资源库: <http://archive.ics.uci.edu/ml/index.php>
   2. 欧洲政府数据集:https://data.europa.eu/euodp/data/dataset
   3. 卡内基·梅隆大学(CMU)统计学院网站的整合数据<http://lib.stat.cmu.edu/datasets/>
   4. 谷歌数据集搜索引擎: https://toolbox.google.com/datasetsearch
   5. 波士顿房价数据集：506条
   6. 泰坦尼克号船员数据集titanic
   7. 国际货币基金组织数据集 [IMF Data](https://www.imf.org/en/Data)：国际货币基金组织发布的关于国际金融、信贷率、外汇储备、产品价格及投资相关的数据
   8. FBI 统一犯罪报告[FBI Uniform Crime Reporting](https://www.fbi.gov/services/cjis/ucr)：FBI负责整理发布的美国国家、州及区县级别的犯罪统计
   9. Google Trends：包含全世界的互联网搜索活动及新闻导向等相关的数据，https://trends.google.com/trends/
   10. Twitter：人与人之间在讨论什么、在哪里讨论的、什么时候讨论的，以及为什么要讨论这个话题。https://developer.twitter.com/en/docs/tweets/search/overview
   11. Google Scholar：包含学术论文的整篇文章、杂志、图书及法律案例等。https://scholar.google.com/
   12. Instagram：提取到喜欢、提及、机商业信息等做分析。<https://developers.facebook.com/products/instagram/>

PPT文字资料：

Basic info:

This dataset is collected by M Yasser H, an engineer in India, and can be found on Kaggle. Yasser, The owner researched 45 Walmart stores and collected data from early 2010 to late 2021 [in the form of weekly sales].

This dataset has 8 columns and more than 6 thousand rows, the owner said it would be updated yearly and there is only one version of it.

About the Structure, the Names, types and an example of all these columns are listed in the following picture.

For example, the variable holiday flag, which is an interger, only has the value of 1 or 0. It shows whether the week is a special holiday week, just like those where one of the days is Thanksgiving.

Lets see the next ,is it big data? About the 5vs, it’s only 364kb, updated yearly and only has 8 columns and 3 types of data. In terms of the veracity and value, I think both of them depend on the method and the people are to use it and many other factors.

All in all, I think this dataset is not a big data due to its simple characters. And then our group members will introduce you some bigger dataset. Lets identify whether they are big data.

[

* Store - the store number
* Date - the week of sales
* Weekly\_Sales - sales for the given store
* Holiday\_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week. prominent holidays, the four largest of all, which are the **Super Bowl, Labour Day, Thanksgiving, and Christmas**.
* Temperature - Temperature on the day of sale
* Fuel\_Price - Cost of fuel in the region
* CPI – Prevailing consumer price index
* Unemployment - Prevailing unemployment rate

]