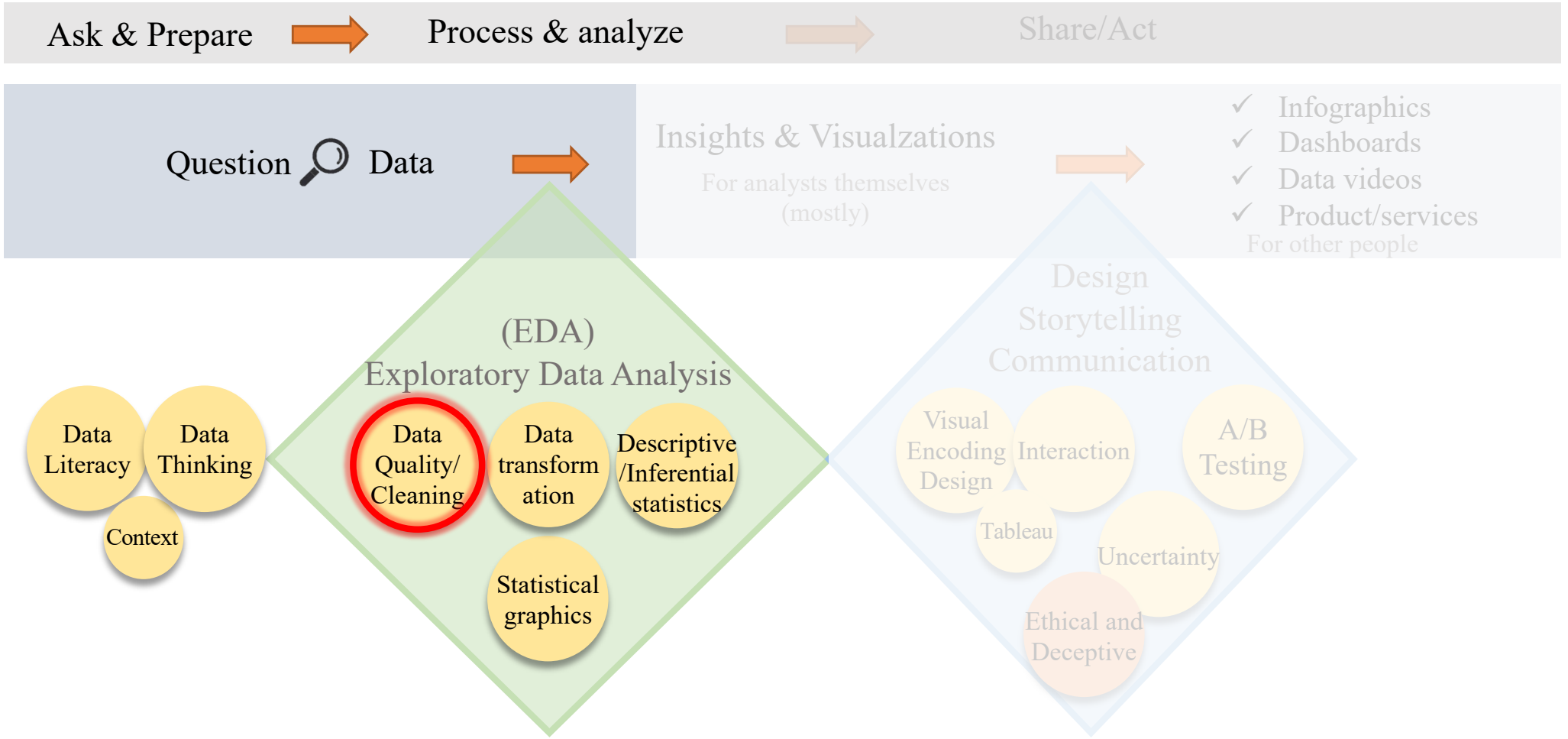
Data Discovery
Lecture 6

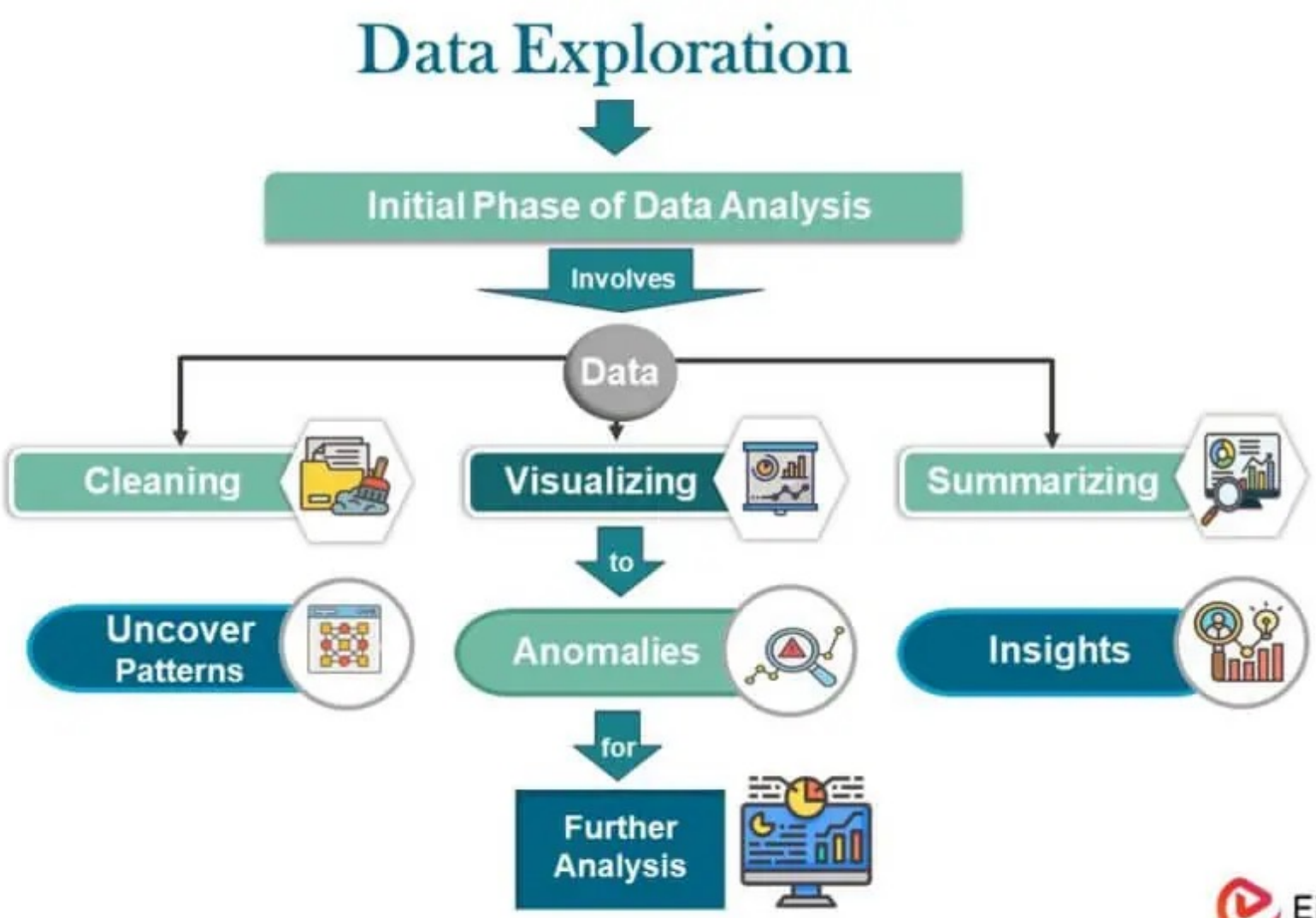
Wan Fang

Southern University of Science and Technology

Design and Learning with Data



Explorative Data Analysis



Agenda

- Data Quality Assessment
- Practice with Python

Data Quality Assessment

Wan Fang

Southern University of Science and Technology

[Adapted from 10.5334/dsj-2015-002 by Li Cai and Yangyong Zhu]

"Garbage in, Garbage out ..."

garbage in, garbage out

[garbage in, garbage out]

DEFINITION

used to express the idea that in computing and other fields, incorrect or poor-quality input will produce faulty output.

Data from Oxford Languages

- **High-quality data are the precondition for analyzing and using big data and for guaranteeing the value of the data.**

- Features of big data (Katal, Wazid, & Goudar, 2013)
  - Volume
    - refers to the tremendous volume of the data. We usually use TB or above magnitudes to measure this data volume.
  - Velocity
    - means that data are being formed at an unprecedented speed and must be dealt with in a timely manner.
  - Variety
    - indicates that big data has all kinds of data types, and this diversity divides the data into structured data and unstructured data. These multityped data need higher data processing capabilities.
  - Value
    - represents low-value density. Value density is inversely proportional to total data size, the greater the big data scale, the less relatively valuable the data.

10.5334/dsj-2015-002

The Challenges of Data Quality

The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.

- Big data sources are very wide, including:
  - data sets from the internet and mobile internet (Li & Liu, 2013);
  - data from the Internet of Things;
  - data collected by various industries;
  - scientific experimental and observational data.

- Rich data types.
  - *unstructured* data: documents, video, audio, etc, occupies more than 80% of the total amount of data .
  - *semi-structured* data: software packages/modules, spreadsheets, and financial reports.
  - *structured* data.

- Obtaining big data with complex structure from different sources and effectively integrating them are a daunting task (McGilvray, 2008).
  - *conflicts* and *inconsistent* or *contradictory* phenomena among data from different sources.
  - In the case of small data volume, the data can be checked by a manual search or programming, even by ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform).
  - However, these methods are useless when processing PB-level even EB-level data volume.

10.5334/dsj-2015-002

The Challenges of Data Quality

Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.

Decimal	
Value	Metric
1000 kB	kilobyte
1000^2 MB	megabyte
1000^3 GB	gigabyte
1000^4 TB	terabyte
1000^5 PB	petabyte
1000^6 EB	exabyte
1000^7 ZB	zettabyte
1000^8 YB	yottabyte

- In 2011, the amount of global data created and copied reached 1.8 ZB.
  - After the industrial revolution, the amount of information dominated by characters doubled every **ten years**.
  - After 1970, the amount of information doubled every **three years**.
  - Today, the global amount of information can be doubled **every two years**.

- A great challenge to the existing techniques of data processing quality.
  - It is difficult to collect, clean, integrate, and finally obtain the necessary high-quality data within a reasonable time frame.
  - Unstructured data in big data is very high, it will take a lot of time to transform unstructured types into structured types and further process the data.

10.5334/dsj-2015-002

The Challenges of Data Quality

Data change very fast, and the "timeliness" of data is very short, which necessitates higher requirements for processing technology.

- Due to the rapid changes in big data, the "timeliness" of some data is very short.
  - If companies can't collect the required data in real time or deal with the data needs over a very long time, then they may obtain outdated and invalid information.
- Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making mistakes by governments or enterprises.
- At present, real-time processing and analysis software for big data is still in development or improvement phases.

10.5334/dsj-2015-002

The Challenges of Data Quality

No unified and approved data quality standards have been formed, and research on the data quality of big data has just begun.

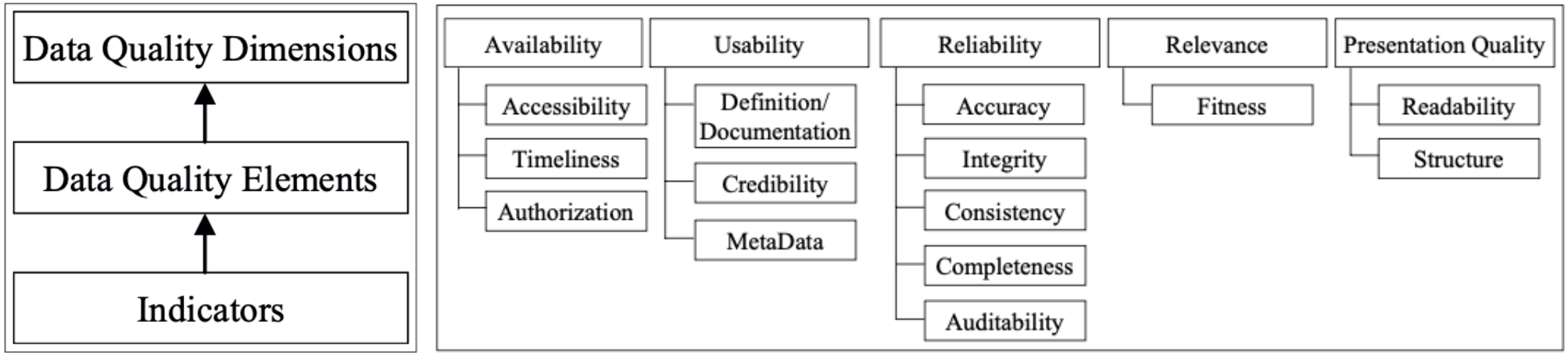
- In order to guarantee the product quality and improve benefits to enterprises, in 1987 the International Organization for Standardization (ISO) published ISO 9000 standards.
  - Nowadays, there are more than 100 countries and regions all over the world actively carrying out these standards.
  - This implementation promotes mutual understanding among enterprises in domestic and international trade and brings the benefit of eliminating trade barriers.

- By contrast, the study of data quality standards began in the 1990s, but not until 2011 did ISO published ISO 8000 data quality standards (Wang, Li, & Wang, 2010).
  - More than 20 countries have participated in this standard,
  - Many disputes about it. The standards need to be mature and perfect.

10.5334/dsj-2015-002

Quality Criteria of Big Data

- **Data quality depends not only on its own features but also on the business environment using the data, including business processes and users.**
- Only the data that conform to the relevant uses and meet requirements can be considered qualified (or good quality) data
  - A hierarchical data quality standard from the perspective of the users



10.5334/dsj-2015-002

A Hierarchical Big Data Quality Assessment Framework

Discussion: Which elements are important in evaluating **social media data**?

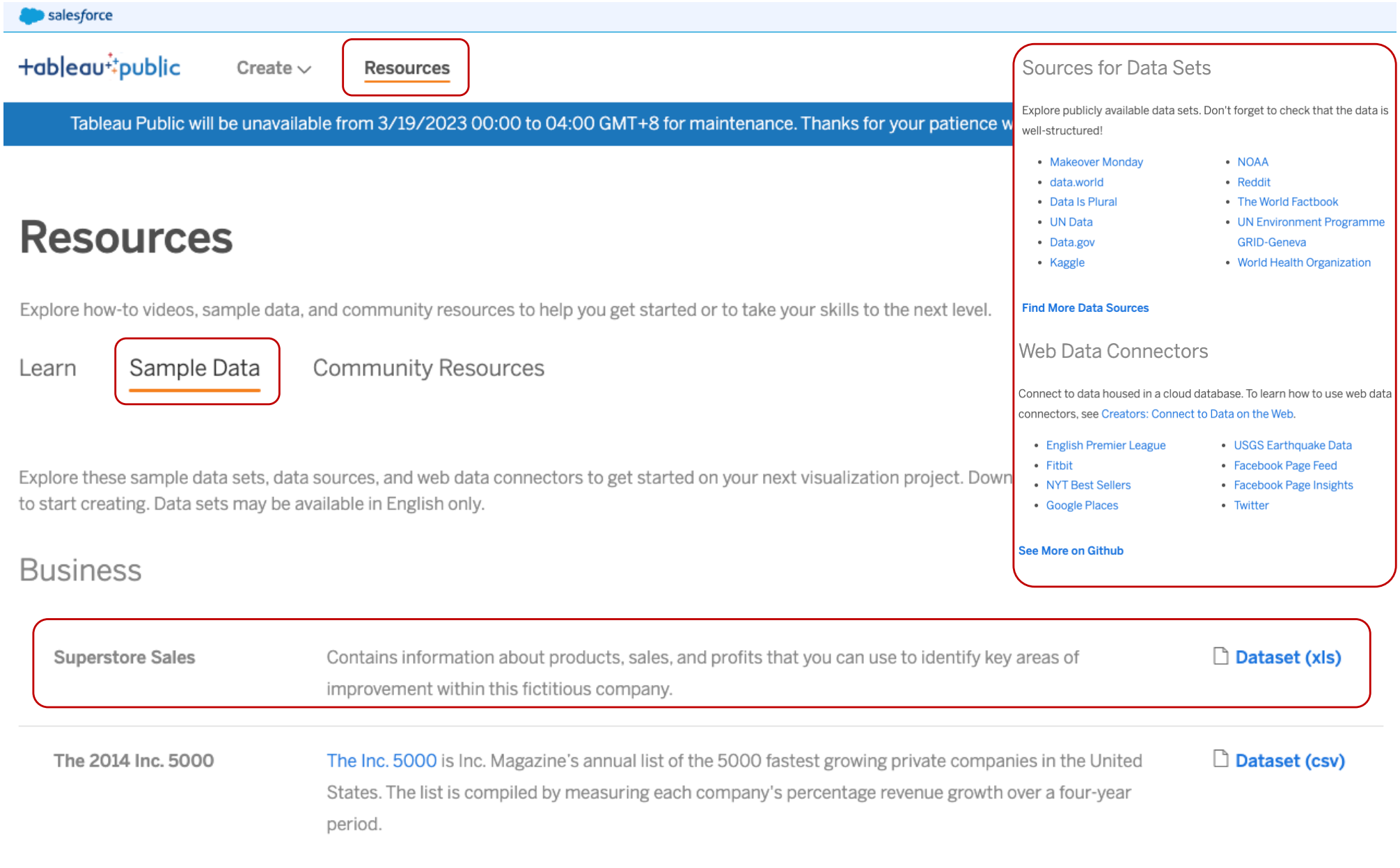
Dimensions	Elements	Indicators
1) Availability	1) Accessibility	▪ Whether a data access interface is provided ▪ Data can be easily made public or easy to purchase
	2) Timeliness	▪ Within a given time, whether the data arrive on time ▪ Whether data are regularly updated ▪ Whether the time interval from data collection and processing to release meets requirements
2) Usability	1) Credibility	▪ Data come from specialized organizations of a country, field, or industry ▪ Experts or specialists regularly audit and check the correctness of the data content ▪ Data exist in the range of known or acceptable values
3) Reliability	1) Accuracy	▪ Data provided are accurate ▪ Data representation (or value) well reflects the true state of the source information ▪ Information (data) representation will not cause ambiguity
	2) Consistency	▪ After data have been processed, their concepts, value domains, and formats still match as before processing ▪ During a certain time, data remain consistent and verifiable ▪ Data and the data from other data sources are consistent or verifiable
	3) Integrity	▪ Data format is clear and meets the criteria ▪ Data are consistent with structural integrity ▪ Data are consistent with content integrity
	4) Completeness	▪ Whether the deficiency of a component will impact use of the data for data with multi-components ▪ Whether the deficiency of a component will impact data accuracy and integrity
4) Relevance	1) Fitness	▪ The data collected do not completely match the theme, but they expound one aspect ▪ Most datasets retrieved are within the retrieval theme users need ▪ Information theme provides matches with users' retrieval theme
5) Presentation Quality	1) Readability	▪ Data (content, format, etc.) are clear and understandable ▪ It is easy to judge that the data provided meet needs ▪ Data description, classification, and coding content satisfy specification and are easy to understand

Assignment 1 on practising **Data Discovery**

- Tasks
  - Pick a sample data and identify at least three on-line analysis of this data, explain how they analyzed and visualized the data
  - Conduct an exploratary analysis of it on your own
  - Summarize and present what you've explored from this sample data source

- Notes
  - Yes, you are not limited to the data sample from Tableau Public.
  - You can find any three in-depth analysis of your chosen data sample as long as you can identify a story that interests you, and you will need to explain why.
  - Then, you are asked to conduct an exploratory analysis on your own to practice the skills.

Source of Sample Data



Practice with Python

- On Feishu Doc
  - https://bionicdl.feishu.cn/docx/IfFrd8sFAotPwcxROxjcJ18Mndc