



DS363: Design and Learning with Data

---

# Module 03

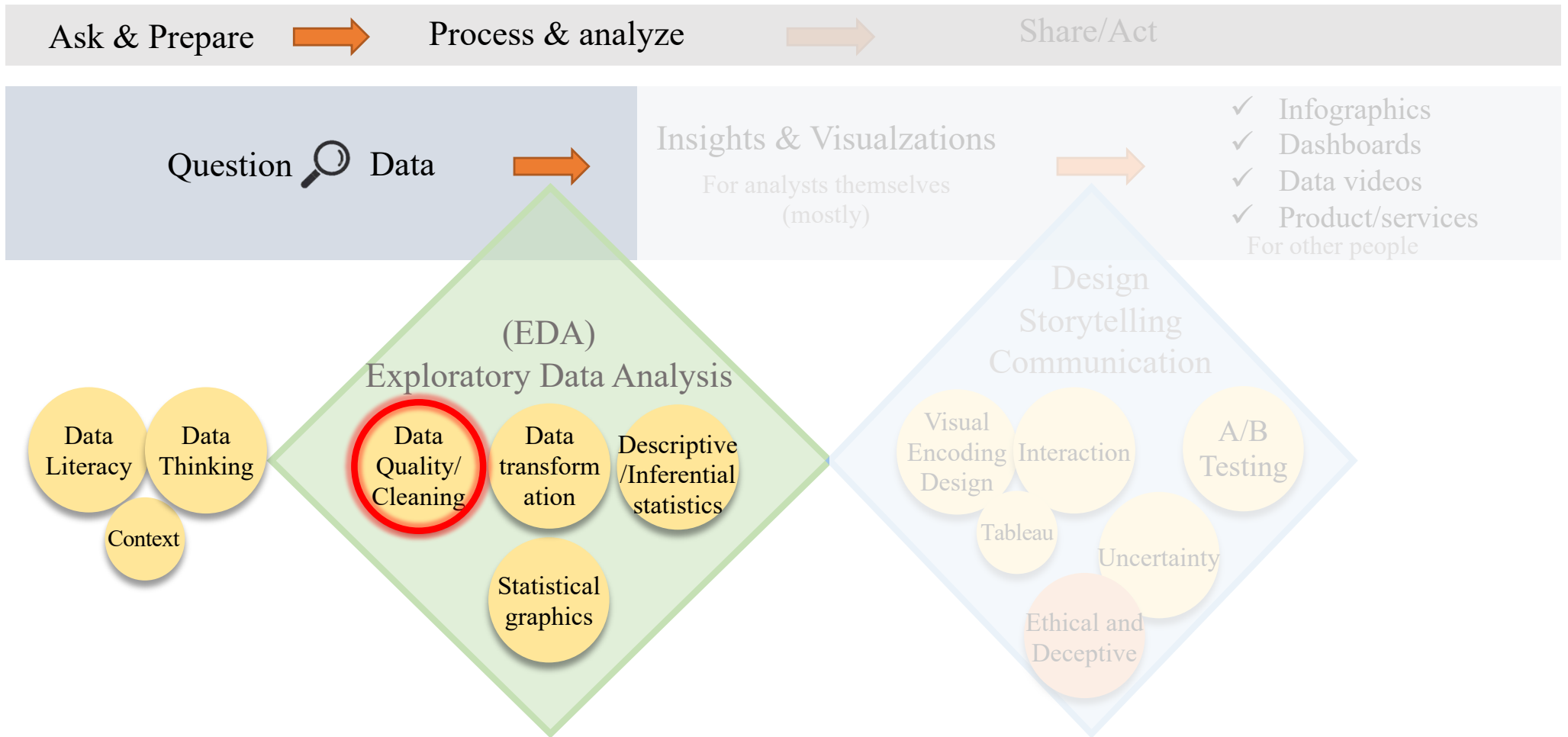
## Data Discovery

### Lecture 1

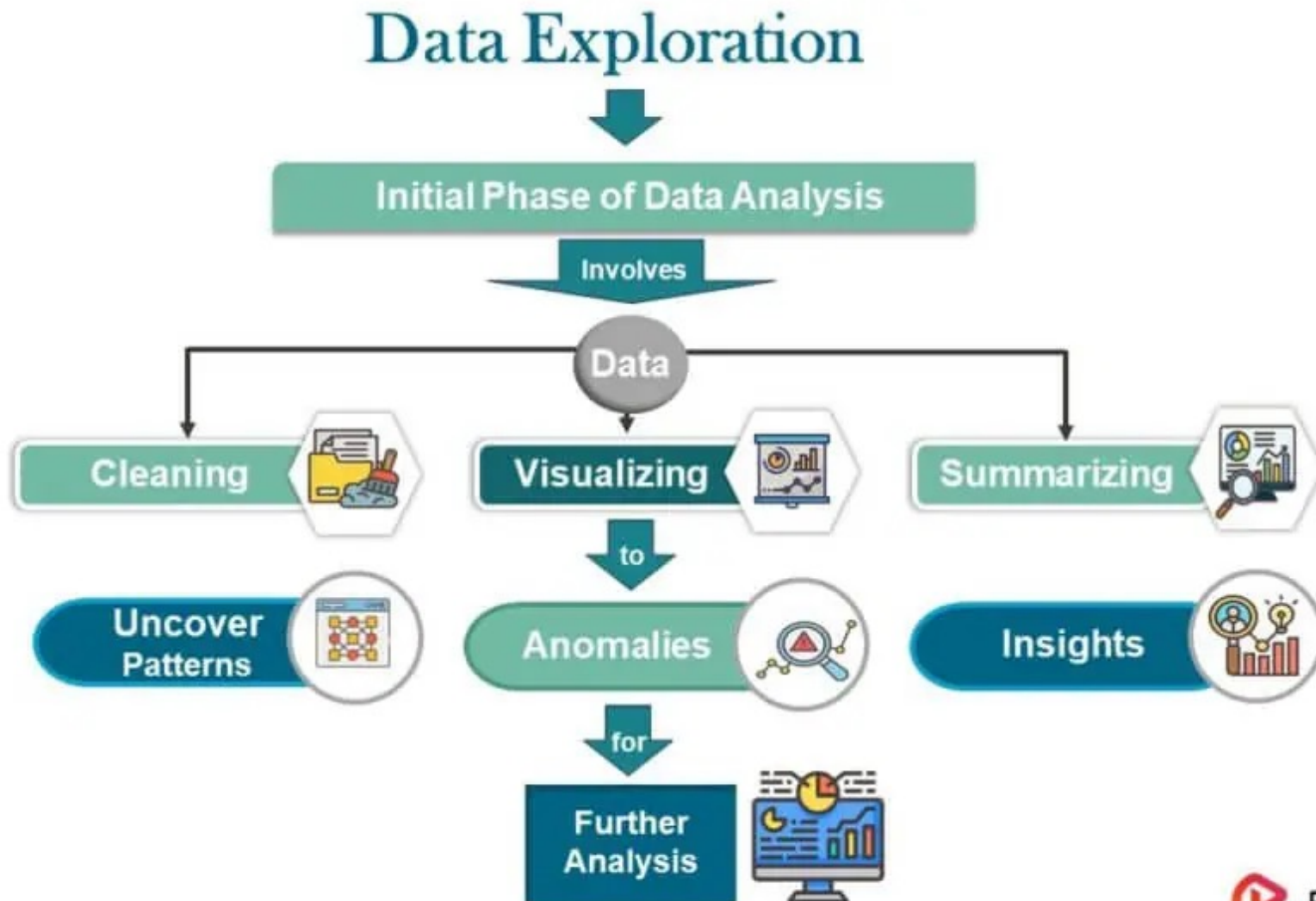
Wan Fang

Southern University of Science and Technology

# Design and Learning with Data



# EDA



# 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 ...*”

- **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.

# 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:
  - 1) data sets from the internet and mobile internet (Li & Liu, 2013);
  - 2) data from the Internet of Things;
  - 3) data collected by various industries;
  - 4) 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.

# The Challenges of Data Quality

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

- 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.

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



## 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.

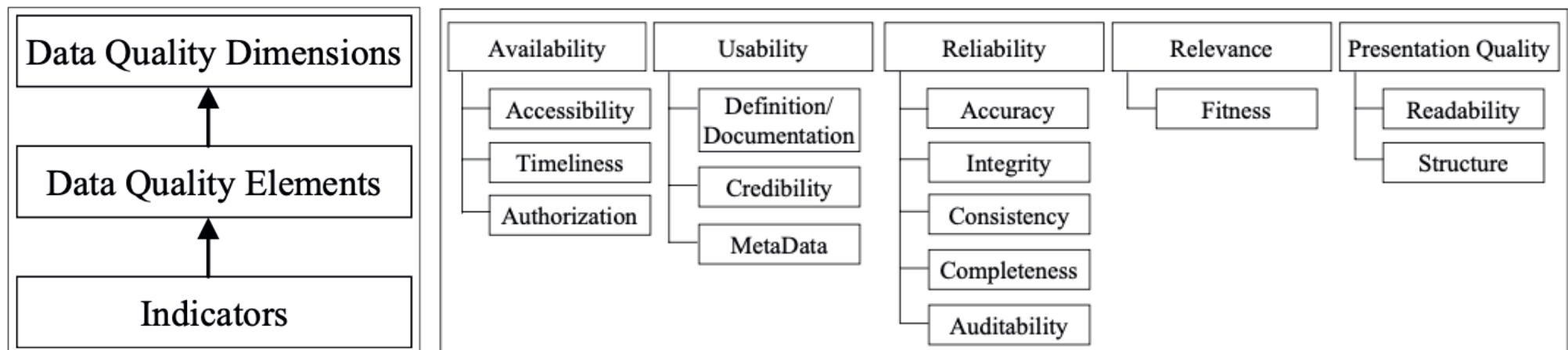
## The Challenges of Data Quality

**No unified and approved data quality standards have been formed in China and abroad, 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).
  - At present, more than 20 countries have participated in this standard, but there are many disputes about it.
  - The standards need to be mature and perfect.

# Quality Criteria of Big Data

- Academia hasn't made a uniform definition of its data quality and quality criteria
- **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



## A Hierarchical Big Data Quality Assessment Framework

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

Dimensions	Elements	Indicators
1) Availability	1) Accessibility	<ul style="list-style-type: none"> <li>Whether a data access interface is provided</li> <li>Data can be easily made public or easy to purchase</li> </ul>
	2) Timeliness	<ul style="list-style-type: none"> <li>Within a given time, whether the data arrive on time</li> <li>Whether data are regularly updated</li> <li>Whether the time interval from data collection and processing to release meets requirements</li> </ul>
2) Usability	1) Credibility	<ul style="list-style-type: none"> <li>Data come from specialized organizations of a country, field, or industry</li> <li>Experts or specialists regularly audit and check the correctness of the data content</li> <li>Data exist in the range of known or acceptable values</li> </ul>
3) Reliability	1) Accuracy	<ul style="list-style-type: none"> <li>Data provided are accurate</li> <li>Data representation (or value) well reflects the true state of the source information</li> <li>Information (data) representation will not cause ambiguity</li> </ul>
	2) Consistency	<ul style="list-style-type: none"> <li>After data have been processed, their concepts, value domains, and formats still match as before processing</li> <li>During a certain time, data remain consistent and verifiable</li> <li>Data and the data from other data sources are consistent or verifiable</li> </ul>
	3) Integrity	<ul style="list-style-type: none"> <li>Data format is clear and meets the criteria</li> <li>Data are consistent with structural integrity</li> <li>Data are consistent with content integrity</li> </ul>
4) Relevance	4) Completeness	<ul style="list-style-type: none"> <li>Whether the deficiency of a component will impact use of the data for data with multi-components</li> <li>Whether the deficiency of a component will impact data accuracy and integrity</li> </ul>
	1) Fitness	<ul style="list-style-type: none"> <li>The data collected do not completely match the theme, but they expound one aspect</li> <li>Most datasets retrieved are within the retrieval theme users need</li> <li>Information theme provides matches with users' retrieval theme</li> </ul>
5) Presentation Quality	1) Readability	<ul style="list-style-type: none"> <li>Data (content, format, etc.) are clear and understandable</li> <li>It is easy to judge that the data provided meet needs</li> <li>Data description, classification, and coding content satisfy specification and are easy to understand</li> </ul>

# Source of Sample Data

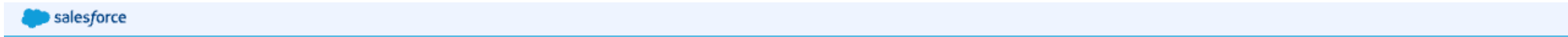


tableau public

Create ▾

Resources

Tableau Public will be unavailable from 3/19/2023 00:00 to 04:00 GMT+8 for maintenance. Thanks for your patience w

## Resources

Explore how-to videos, sample data, and community resources to help you get started or to take your skills to the next level.

Learn

Sample Data

Community Resources

Explore these sample data sets, data sources, and web data connectors to get started on your next visualization project. Down to start creating. Data sets may be available in English only.

## Business

**Superstore Sales**

Contains information about products, sales, and profits that you can use to identify key areas of improvement within this fictitious company.

[Dataset \(xls\)](#)

**The 2014 Inc. 5000**

[The Inc. 5000](#) is Inc. Magazine's annual list of the 5000 fastest growing private companies in the United States. The list is compiled by measuring each company's percentage revenue growth over a four-year period.

[Dataset \(csv\)](#)

### Sources for Data Sets

Explore publicly available data sets. Don't forget to check that the data is well-structured!

- [Makeover Monday](#)
- [data.world](#)
- [Data Is Plural](#)
- [UN Data](#)
- [Data.gov](#)
- [Kaggle](#)
- [NOAA](#)
- [Reddit](#)
- [The World Factbook](#)
- [UN Environment Programme GRID-Geneva](#)
- [World Health Organization](#)

[Find More Data Sources](#)

### Web Data Connectors

Connect to data housed in a cloud database. To learn how to use web data connectors, see [Creators: Connect to Data on the Web](#).

- [English Premier League](#)
- [Fitbit](#)
- [NYT Best Sellers](#)
- [Google Places](#)
- [USGS Earthquake Data](#)
- [Facebook Page Feed](#)
- [Facebook Page Insights](#)
- [Twitter](#)

[See More on Github](#)



# DS363: Design and Learning with Data

---

<https://ds363.ancorasir.com/>

**Thank you~**

Wan Fang  
Southern University of Science and Technology