UNIV2V96 Transfer Research Initiative

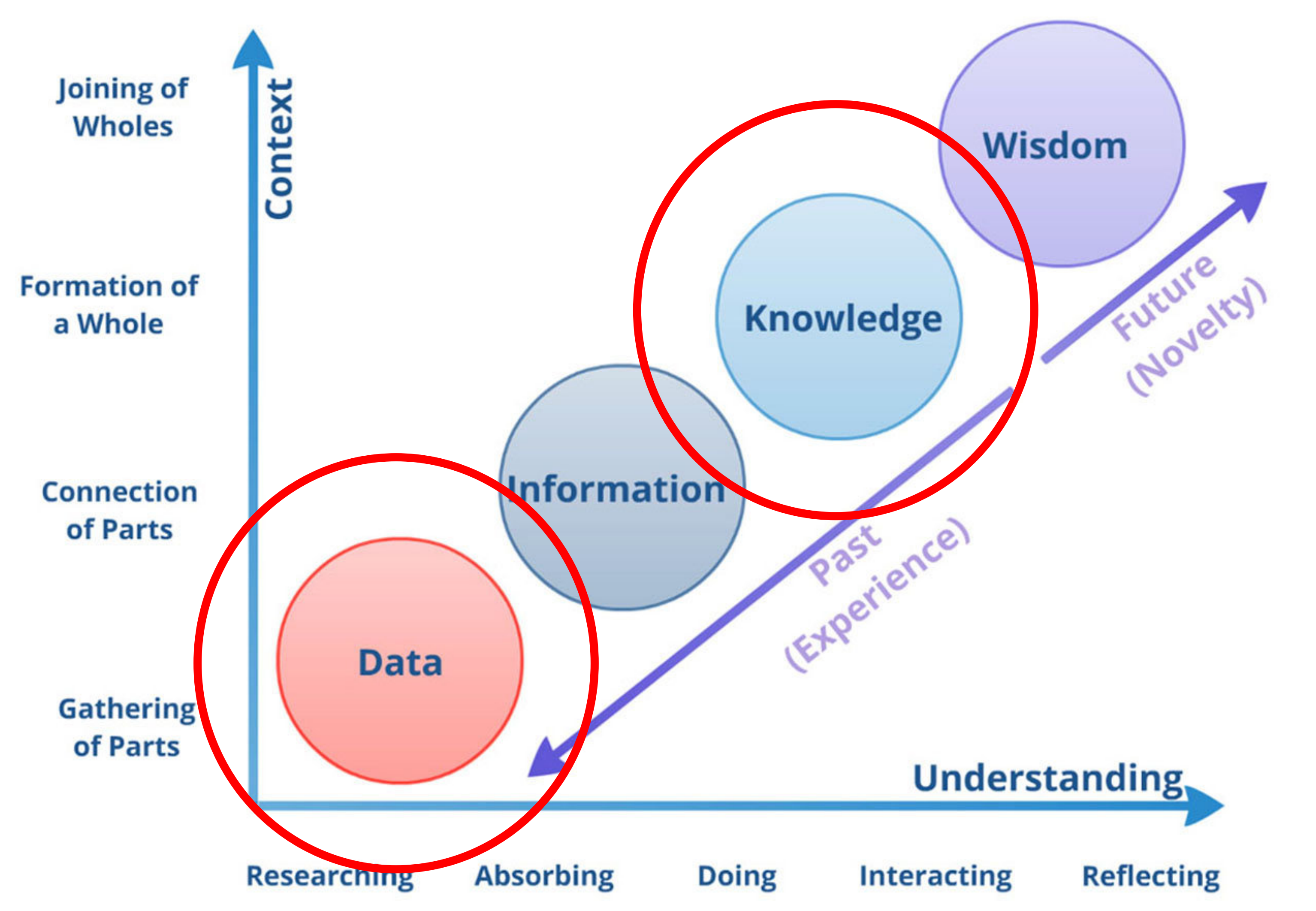
Data Science in Practice: Data Collection

Karl Ho

School of Economic, Political and Policy Sciences

University of Texas at Dallas

What is data?



Ackoff, R.L., 1989. From data to wisdom. *Journal of applied systems analysis*, *16*(1), pp.3-9.

What is data?

1. Kinds of Data
    1. Quantitative vs. Qualitative
    2. Structured vs. Semi/unstructured
    3. Measurement
        - Nominal/ordinal/interval/ratio

What is data?

2. Data generation
   1. Made data vs. Found data
   2. Structured vs. Semi/unstructured
   3. Primary vs. secondary data
   4. Derived data
      1. metadata, paradata

What is Big Data?

The Big data is about data that has huge volume, cannot be on one computer. Has a lot of variety in data types, locations, formats and form. It is also getting created very very fast (velocity) (Doug Laney 2001).

What is Big Data?

Burt Monroe (2012)

5Vs of Big data

- Volume
- Variety
- Velocity
- Vinculation
- Validity

What is Data Science?

1. Science of Data
2. Understand Data Scientifically

Data Science Keywords

- Data management
- Data analytics
- Data scientists
- Data curation
- Modeling
- CRMs

The story of Google Flu Trend

By using Big Data of search queries, Google Flu Trend (GFT) predicted the flu-like illness rate in a population.

The findings were published in the top journal Nature in 2008. However, shortly GFT failed and missed at the peak of the 2013 flu season by 140 percent.

The story of Google Flu Trend

Lazer, Kennedy, King and Vespignani (2014)

Traditional “small data” often offer information that is not contained (or containable) in big data, and "by combining GFT and lagged [traditional] CDC data, as well as dynamically recalibrating GFT... can substantially improve on the performance of GFT or the CDC alone. " (Lazer et al. 2014 *Science*)

Google should have highest power in data access .

Why would it fail?

$$Power = f(Data_{Size}, Data_{Veracity}, Data_{Speed})$$

Why would it not fail yet?

$$Power = f(Data_{Veracity}, Data_{Speed}, Data_{Size})$$

Size still matters, but not first.

Data Methods

1. Survey
2. Experiments
3. Qualitative Data
4. Text Data
5. Web Data
6. Machine Data
7. Complex Data
    1. Network Data
    2. Multiple-source linked Data

Made Data

Found Data

Data Methods

1. Small data or Made data emphasize design
2. Big data or Found data focus on algorithm

How Data are generated?

- Computers
- Web
- Mobile devices
- IoT (Internet of Things)
- Further extension of human users (e.g. AI, avatars)

Web data

How do we take advantage of the web data?

1. Purpose of web data
2. Generation process of web data
3. What is data of data?
4. Why data scientists need to collect web data?