Data Visualization: Data Theory

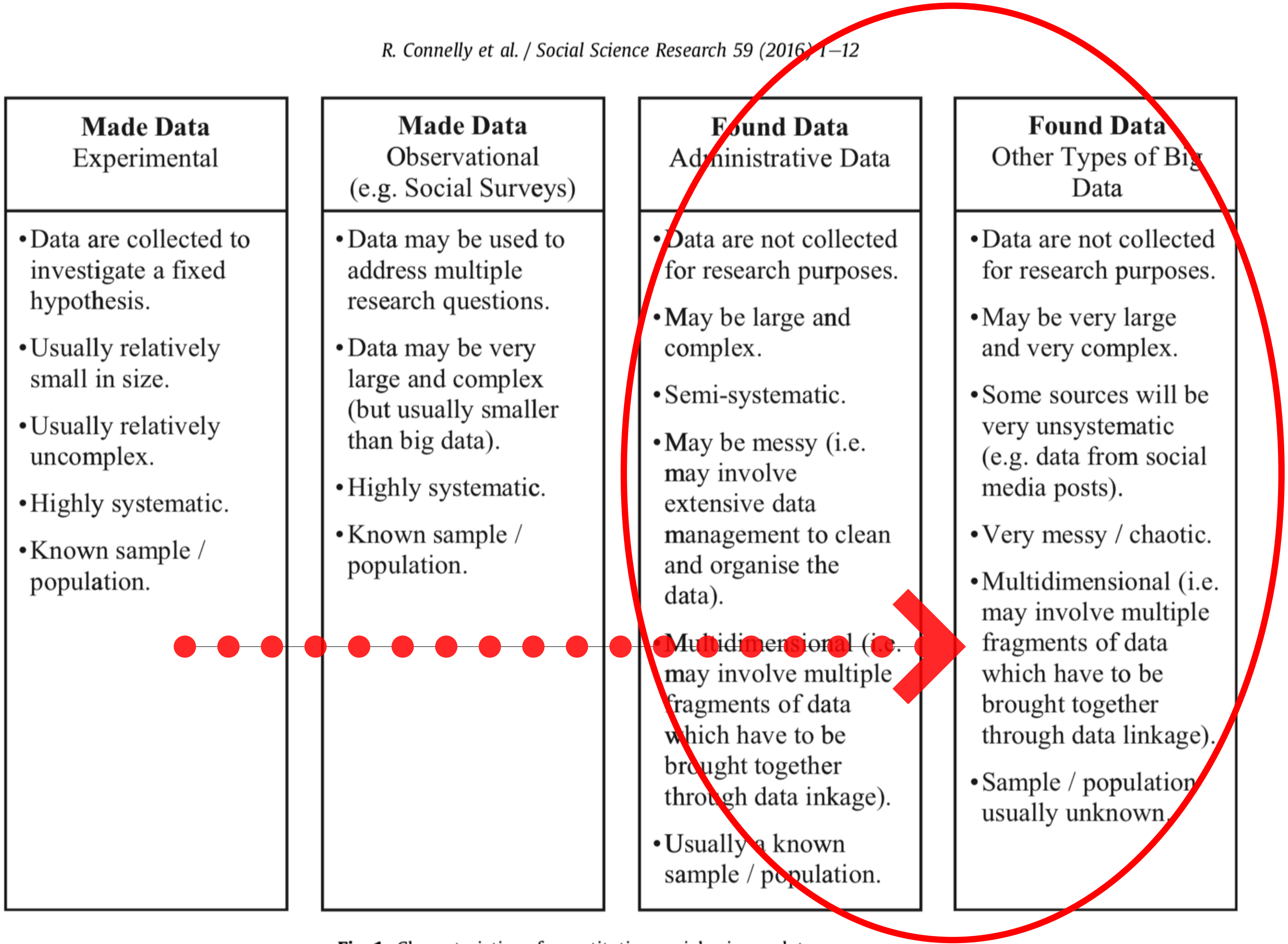
Karl Ho

School of Economic, Political and Policy Sciences

University of Texas at Dallas

A Theory of Data: Understanding Data Generation

Data Generation



R. Connelly et al. / Social Science Research 59 (2016) 1–12

Made Data Experimental	**Made Data** Observational (e.g. Social Surveys)	**Found Data** Administrative Data	**Found Data** Other Types of Big Data
• Data are collected to investigate a fixed hypothesis. • Usually relatively small in size. • Usually relatively uncomplex. • Highly systematic. • Known sample / population.	• Data may be used to address multiple research questions. • Data may be very large and complex (but usually smaller than big data). • Highly systematic. • Known sample / population.	• Data are not collected for research purposes. • May be large and complex. • Semi-systematic. • May be messy (i.e. may involve extensive data management to clean and organise the data). • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data inkage). • Usually a known sample / population.	• Data are not collected for research purposes. • May be very large and very complex. • Some sources will be very unsystematic (e.g. data from social media posts). • Very messy / chaotic. • Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage). • Sample / population usually unknown.

Fig. 1. Characteristics of quantitative social science data resources.

Administrative Data

Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies

- Connelly et al. 2016

A Taxonomy of Data

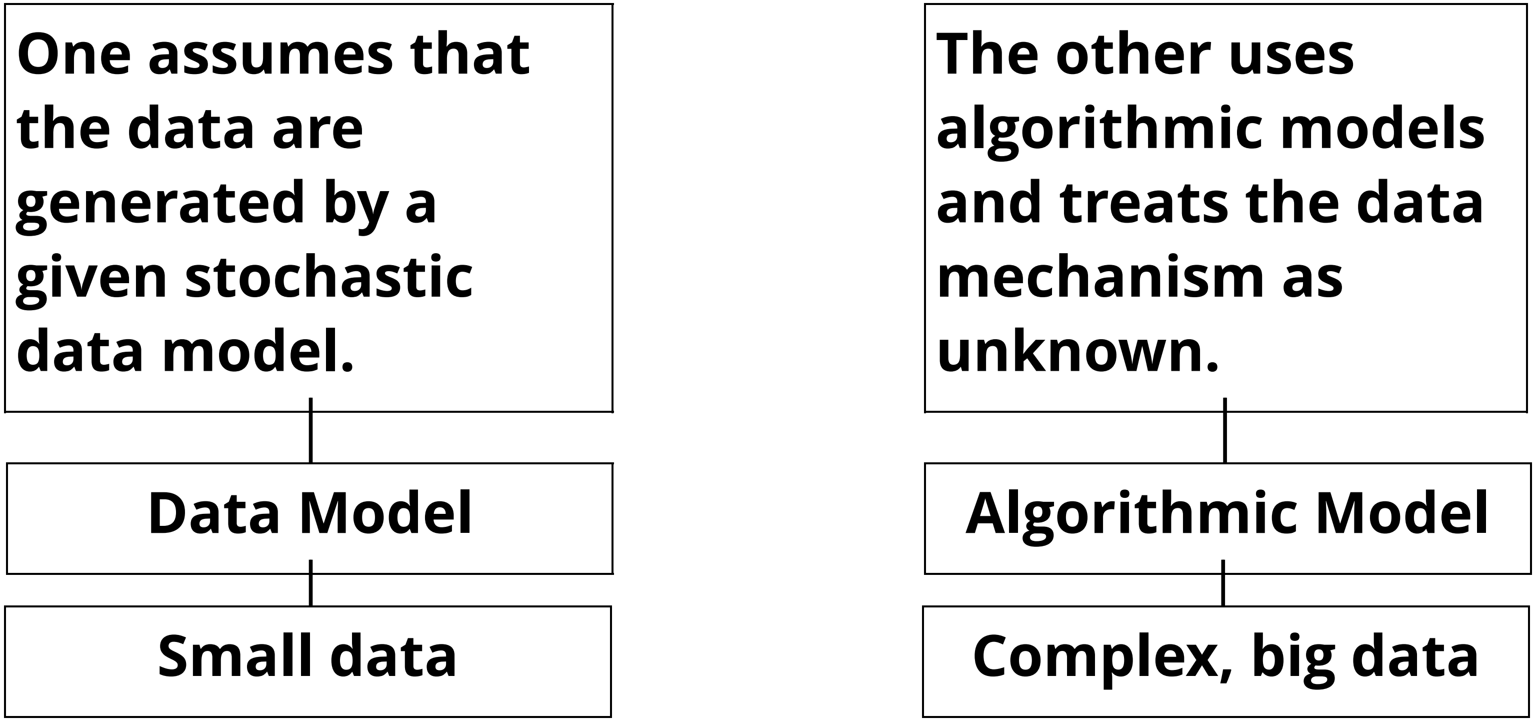
1. Numbers
2. Text
3. Images
4. Audio
5. Video
6. Signals
7. Data of data: Metadata and Paradata

Categories of Data

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

Statistical Modeling: The Two Cultures

Leo Breiman 2001: *Statistical Science*



One assumes that the data are generated by a given stochastic data model.

Data Model

Small data

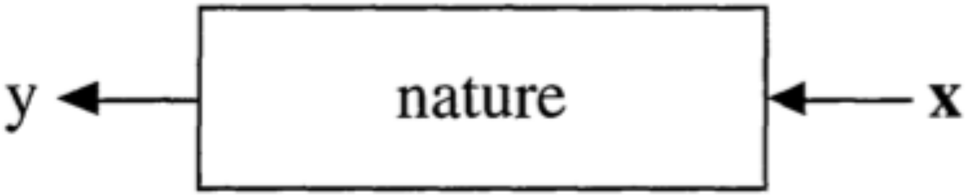
The other uses algorithmic models and treats the data mechanism as unknown.

Algorithmic Model

Complex, big data

Theory:
Data Generation Process

Data are generated in many fashions. Picture this: independent variable x goes in one side of the box-- we call it nature for now-- and dependent variable y come out from the other side.



Theory:
Data Generation Process

Data Model

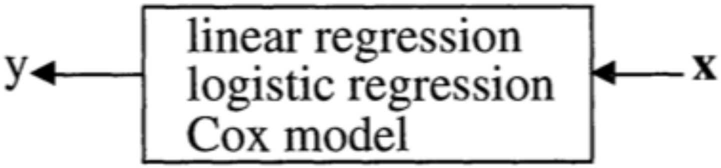
The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from response variables.

Response Variable= f(Predictor variables, random noise, parameters)

Reading the response variable is a function of a series of predictor/independent variables, plus random noise (normally distributed errors) and other parameters.

Theory:
Data Generation Process

Data Model



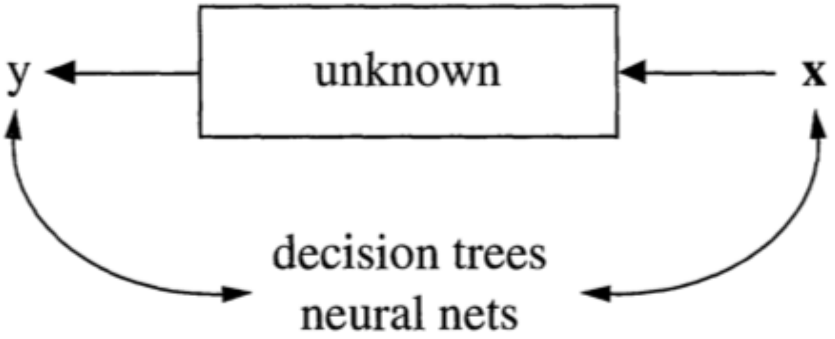
The values of the parameters are estimated from the data and the model then used for information and/or prediction.

Theory:
Data Generation Process

Algorithmic Modeling

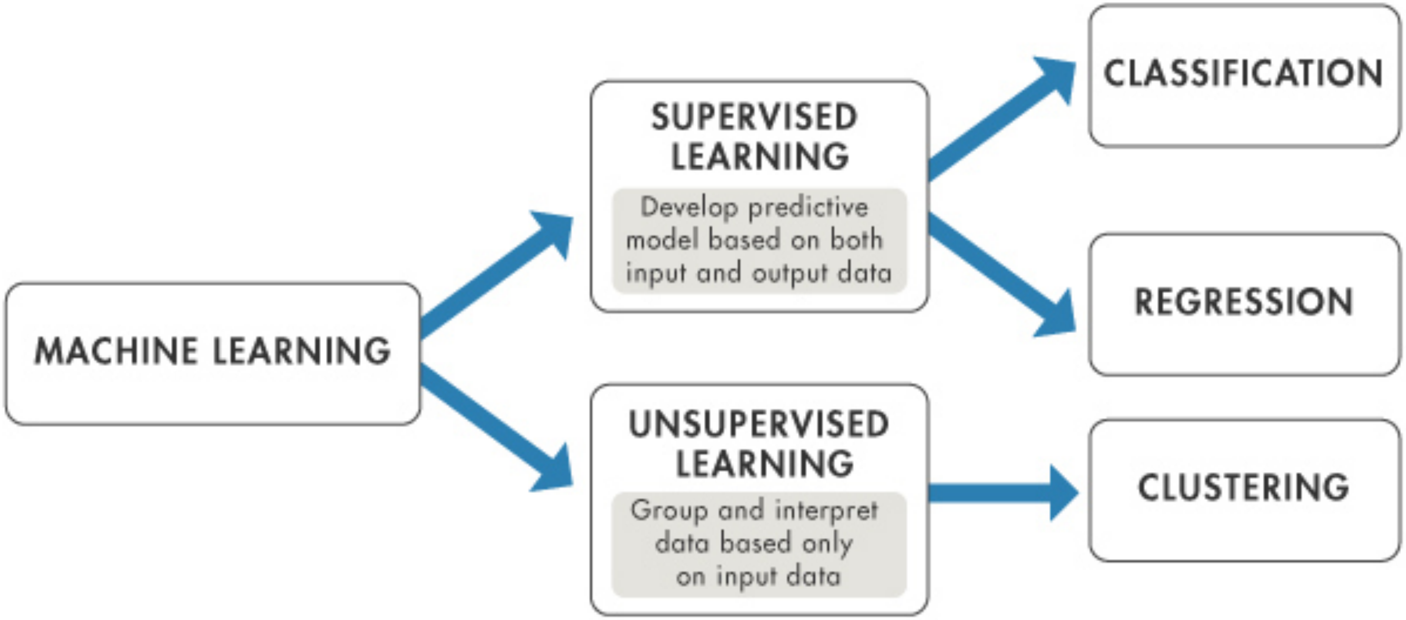
The analysis in this approach considers the inside of the box complex and unknown. Their approach is to find a function f(x)-an algorithm that operates on x to predict the responses y.

The goal is to find algorithm that accurately predicts y.



Theory:
Data Generation Process

Algorithmic Modeling


MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Supervised Learning vs. Unsupervised Learning

Source: https://www.mathworks.com

Social (Data) Scientist's mission

Two major areas to which social scientists can contribute, based on decades of experience and work with end users, are:

1. Inference
2. Data quality.

- Foster *et al.* 2016

Algorithm and Inference

Very broadly speaking, algorithms are what statisticians do while inference says why they do them.

- Efron and Hastie 2017

Let the dataset change your mindset.

- Hans Rosling

Hans Rosli

Swedish physician an
statistician

- Founded Gapmind
Foundation
- Visualize historical
on public health a
poverty

Data is the new oil.

Data is the new soil.

Hal Varian

Chief Economist, Google

Professor of Economics, University of California, Berkeley.

Big Data: New Tricks for Econometrics

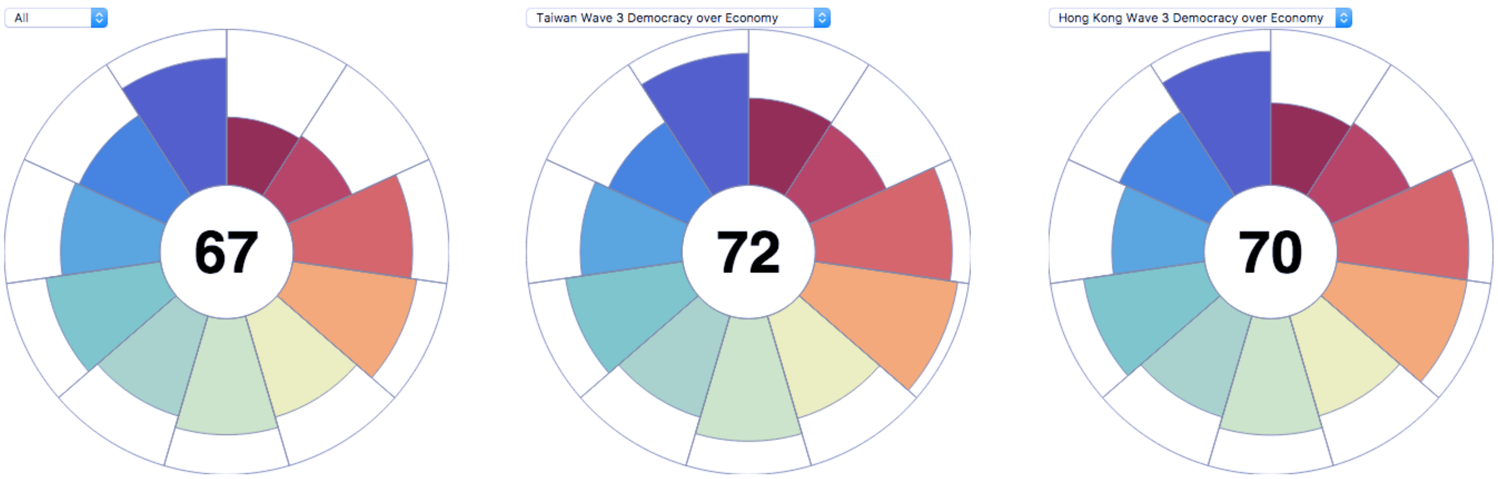
Machine Learning and Econometrics

Java: D3 Library

Latent Profile Models:

Hong Kong and Taiwan

Wave 3 (2010, 2012) I Wave 4 (2014, 2016)

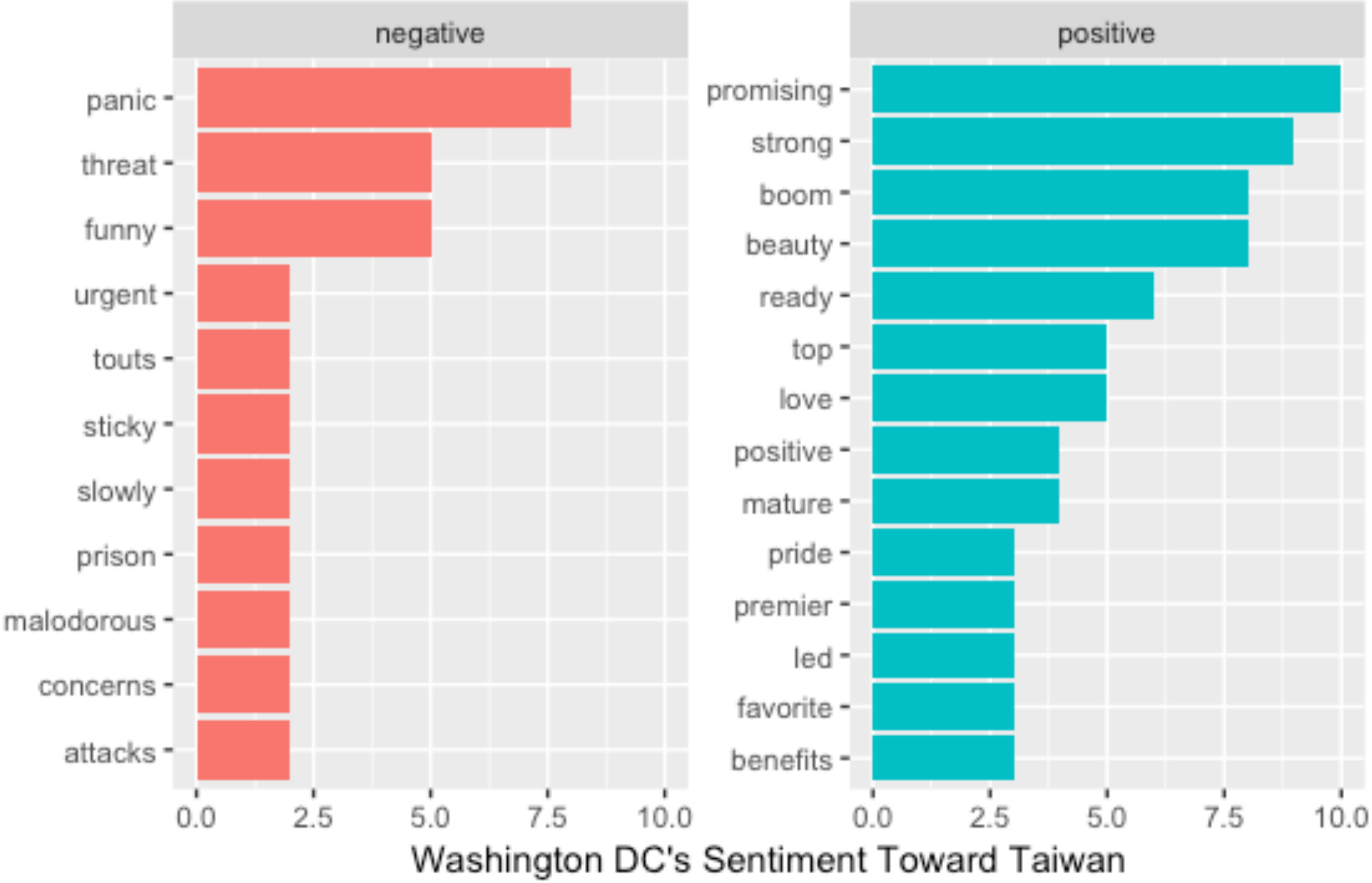


All

Taiwan Wave 3 Democracy over Economy

Hong Kong Wave 3 Democracy over Economy

Sentiment Analysis



Could data analytics add value to your research?

The first thing is "it will do no harm". Visualized data must not obscure the findings or confuse the readers.