Big Data and the Role of Statistics

NIKKIN L. BERONILLA
Director III

2024 Regional Data Festival (RDaFest) - Visayas Cluster
28-29 October 2024
DepEd Ecotech Center, Sudlon, Lahug, Cebu City

Outline

I. Introduction [1]
II. Characteristics of Big Data [3]
III. Example of Data Sources [1]
IV. Skills and Tools [4]
V. Challenges (Individual Level) [4]
VI. Challenges (Institution Level) [3]
VII. Summary [1]

Big Data & Statistics

“I keep saying the sexiest job in the next ten years will be statisticians”. - Hal Varian, Chief Economist at Google, 2008

II. Characteristics of Big Data

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Variability

II. Characteristics of Big Data

Volume

Velocity

Variety

II. Characteristics of Big Data

Veracity

Marr, 2014 (Five Vs)

Variability

McNulty, 2014 (Seven Vs)

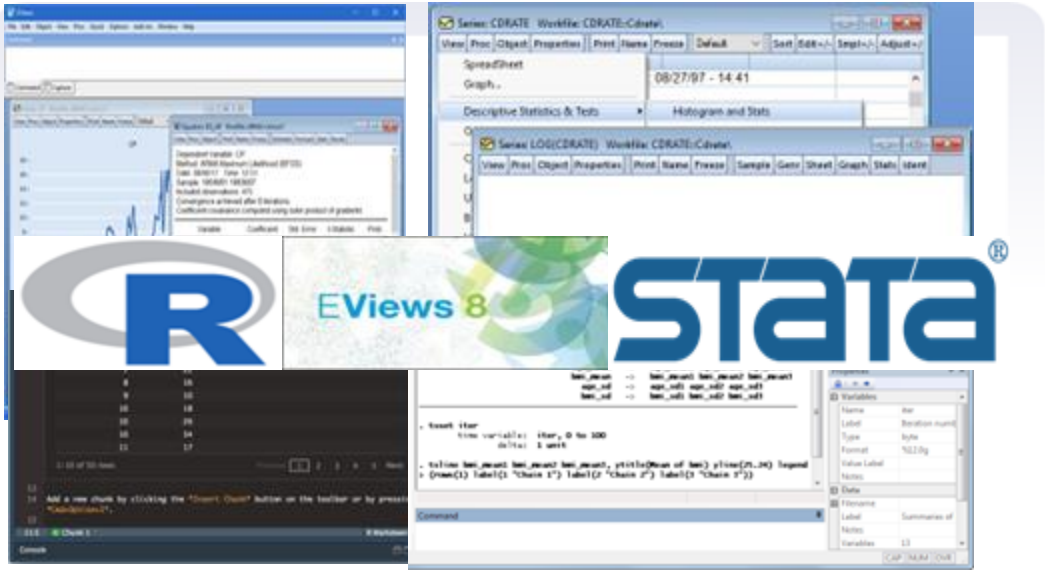
III. Examples of Data Sources

	Survey Data	Admin Data	Other Data
	Labor Force Survey	DPWH Road Stats	Twitter Posts
Purpose	Statistical	Monitor program	Other purpose
Volume	Manageable	Manageable	Huge
Velocity	Periodic with lags	Periodic with lags	Realtime
Variety	Structured	Some structure	Some structure
Veracity	Known	Known	Noisy
Variability	Static wrt context	Static wrt context	Shifting definition

Skill Sets Needed

- ➔ Statistics
- ➔ Mathematics
- ➔ Computer Science
- ➔ Domain Knowledge

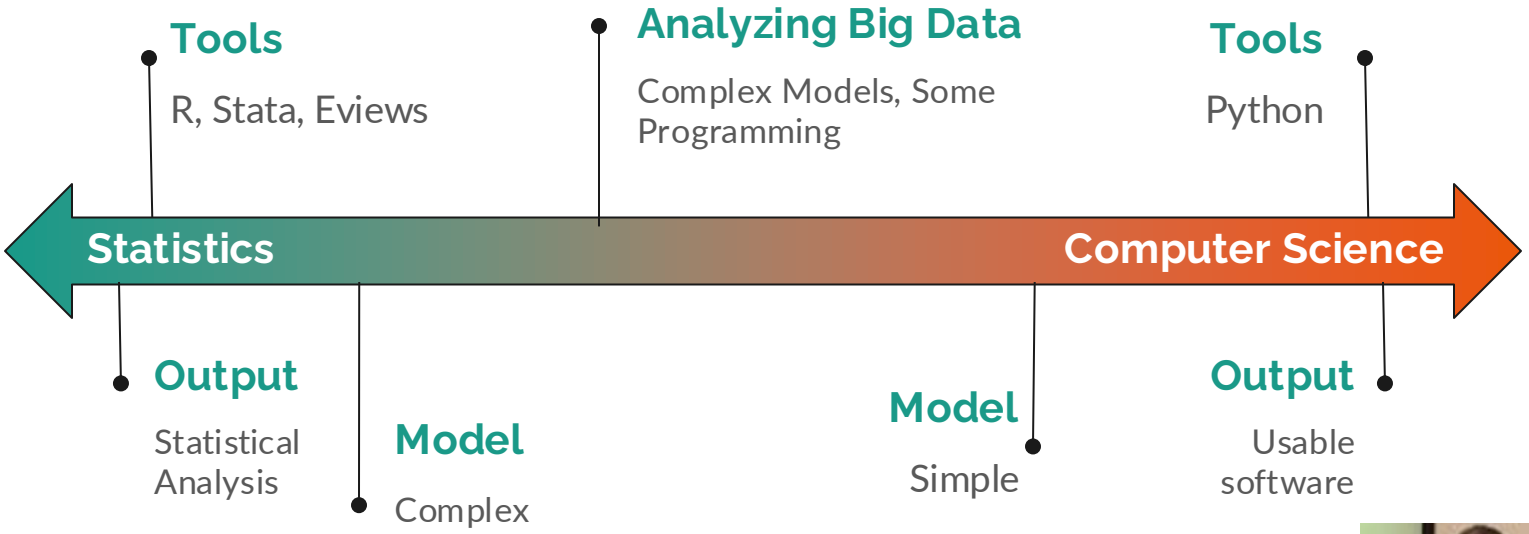
Statistical Tools



Statistics and Computer Science

“The computer scientists are used to working with vast amounts of data using relatively unstructured models. Statisticians tend to have more complex models but focus on smaller data sets.”. - Hal Varian, Chief Economist at Google, 2011

Statistics Identity Crisis on Big Data



Alyssa Frazze, 2016

V. Challenges (Individual level)

Data preparation

➔ 80/20 rule (80% cleaning, preparation)

```
lfsm2 <- lfsm2 %>% rename(PUFC08_CONWR = PUFC10_CONWR, PUFC09_WO
PUFC11_WORK, PUFC09A_WORK = PUFC11A_ARRANGEMENT, PUFC10_JOB = PU
PUFC11A_PROVMUN = PUFC12A_PROVMUN, PUFC13_PROCC = PUFC14_PROCC,
PUFC16_PKB, PUFC16_NATEM = PUFC17_NATEM, PUFC17_PNWHRS = PUFC18_
PUFC18_PHOURS = PUFC19_PHOURS, PUFC19_PWMORE = PUFC20_PWMORE, PU
PUFC21_PLADDW, PUFC21_PCLASS= PUFC23_PCLASS, PUFC22_OJOB= PUFC26
PUFC23_THOURS= PUFC28_THOURS, PUFC24_WWM48H = PUFC29_WWM48H, PUF
PUFC31_FLWRK, PUFC26_WYNOT= PUFC34_WYNOT,  PUFC27_AVAIL= PUFC36_
PUFC28_PREVJOB = PUFC38_PREVJOB, PUFC29_YEAR= PUFC39_YEAR,  PUFC
PUFC39_MONTH,  PUFC31_POCC= PUFC41_POCC,  PUFC33_QKB = PUFC43_Q
```

Datasets	Microdata	Documentation	Database
Labor Force Survey June 2023			
Labor Force Survey May 2023			
Labor Force Survey April 2023			
Labor Force Survey March 2023			
Labor Force Survey February 2023			
bor Force Survey January 2023			
bor Force Survey December 2022			
bor Force Survey November 2022			
bor Force Survey October 2022			
bor Force Survey September 2022			

V. Challenges (Individual level)

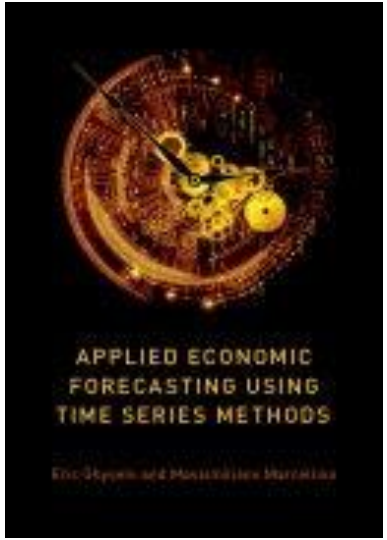
Data preparation

- ➔ Older data taken down
- ➔ archive.org

V. Challenges (Individual level)

Lots of new statistical methods

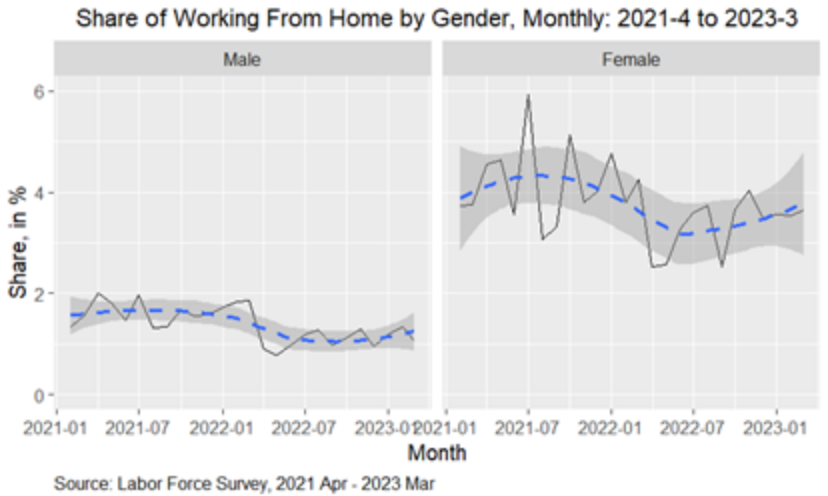
➔ Econometrics, panel data, impact evaluations, machine learning, etc



V. Challenges (Individual level)

Domain knowledge

➔ Asking the right questions



VI. Challenges (Institution level)

Netherlands experience

- ➔ Netherlands NSO established an innovation unit several years ago
- ➔ From personal conversation, they tend to have a reorganization every 15 years

Barteld Braaksma

VI. Challenges (Institution level)

Switzerland experience

- ➔ Data Science Competence Center (DSCC) established in 2020 for agencies to acquire skills in data science
- ➔ A working group in 2015 recommends the creation of DSCC

VI. Challenges (Institution level)

Difficulty in replicating success

- ➔ Effort to pilot Data Science Unit across Agencies fizzled out
- ➔ Preoccupied with other things that are also urgent and important
- ➔ UN has draft manual on setting up Big Data

VII. Summary

- Lots of data, limited statisticians
- Volume is the main identifier of Big Data
- Individual level constraint: TIME
- Success in other countries not easy to replicate but there is hope
- Role of the Statisticians is to Analyze Data

Thank You!

DISCLAIMER
The ideas expressed herein are from the author and do not represent the official views of the Philippine Statistics Authority

Main Reference:
Kitchin R. and G., McArdle [2016] "What Makes Big Data, Big Data?", *Big Data & Society, January–June* **2016**: 1–10