Research Reproducibility
in Computational Social Science

Aek Palakorn Achananuparp, SMU

Research Integrity Conference 2018, Singapore

INTRODUCTION & DEFINITIONS

COMPUTATIONAL SOCIAL SCIENCE (CSS)

First coined by Lazer et al. (2009) in the Nature article

Modeling human activity, behavior, and relationships through the use of computational methods and large-scale data (**thousands** to **billions** of data points)

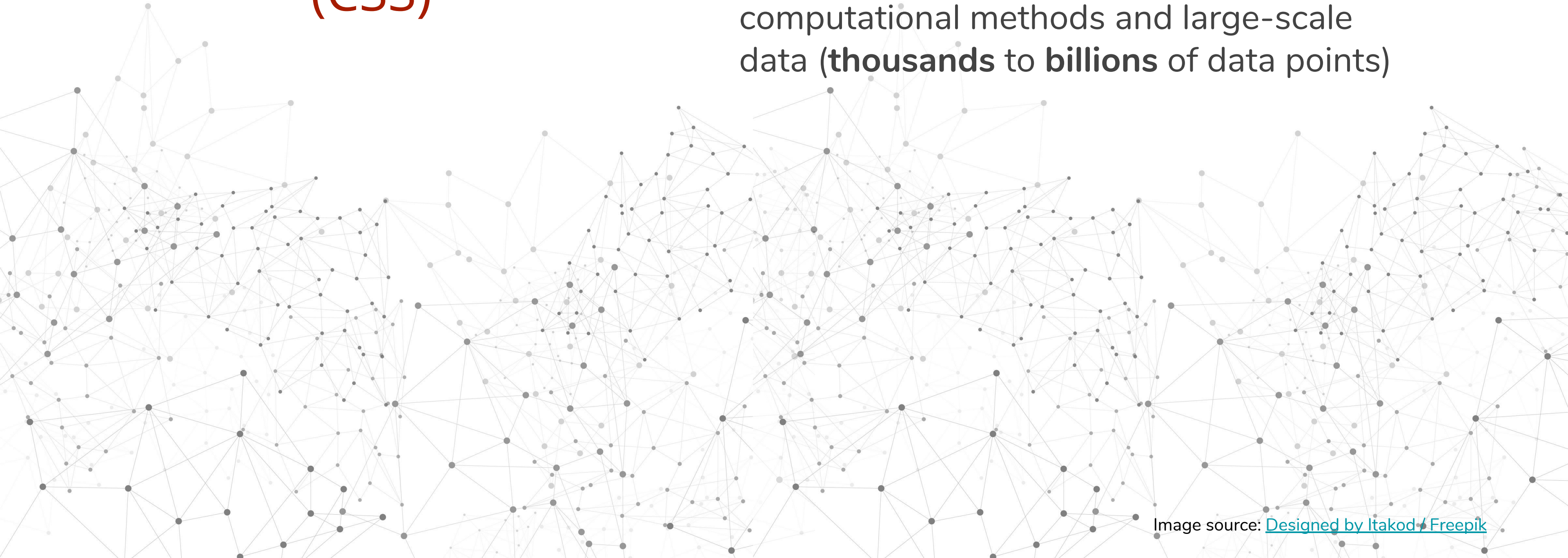
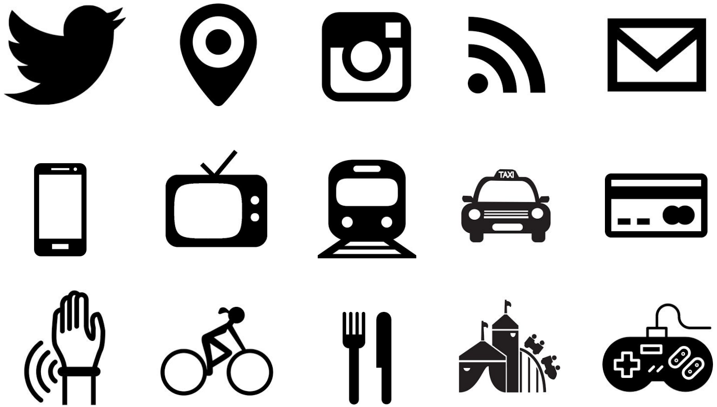


Image source: Designed by Itakod / Freepik

DATA SOURCES
“DIGITAL TRACES”



COMMON STUDY TOPICS

- Predicting friendships in social networks
- Modeling information diffusion process
- Predicting electoral outcomes
- Modeling human activity in offline settings
- Recommending books, papers, articles, movies, songs, etc.

WHAT DOES REPRODUCIBILITY MEAN?

CONCEPT	TEAM	EXPERIMENT SETUP
Repeatability	Same	Same
Replicability	Different	Same
Reproducibility	Different	Different

Source: ACM

NON-COMPUTATIONAL V.S. COMPUTATIONAL RESEARCH

In non-computational research:

Replicability = **reproducibility** = different groups can obtain the same result independently by following the original study’s methodology.

In computational research:

Replicability = different groups can obtain the same result using the original study's artifacts (datasets, code, and workflows).

Reproducibility = different groups can obtain the same result using independently developed artifacts.

COMPUTATIONAL REPRODUCIBILITY

We'll mostly focus on **replication** and **reproduction** of computational research, i.e., computational reproducibility, in CSS.

REPRODUCIBILITY CRISIS IN CSS?

REPRODUCIBILITY CRISIS IN CSS

- For electoral prediction studies using Twitter data, an independent group was not able to reproduce their positive results (Gayo-Avello et al. 2011).

- 61% of 21 social science studies published in Nature and Science can be reproduced (Camerer et al. 2018).

- For 54% of 601 studies published at major computational research conferences, an independent group was able to build the code or the authors stated the code would build with some effort (Collberg et al. 2014).

- Out of 400 artificial intelligence papers, 6% provide code for the papers’ algorithm, 30% provide test data, 54% provide pseudocode (Hutson, 2018).

REPRODUCIBILITY CHALLENGES IN CSS

TECHNOLOGICAL IRREPRODUCIBILITY

- Some code and dataset require high-performance or esoteric systems to run.
- Different tools, platforms, & versions may produce different results.
- Some software dependencies are no longer available.
- Is it still possible to run the original artifacts a few years later?

DATA PRIVACY & LEGAL LIMITATIONS

- Data privacy is going to be more critical than before after the Cambridge Analytica fiasco.

- More difficulty in collecting and sharing online social media data.

- Data ownership is not always clear-cut.

- Intellectual property prevents code sharing.

EXPERIMENTAL IRREPRODUCIBILITY

- Complex social systems are extremely difficult to study.
- States of the world are irrevocably not the same today compared to the time when the original experiments were conducted.
- Some external influences, e.g., media exposure, are almost impossible to control.

ENABLING REPRODUCIBLE RESEARCH

ENABLING REPRODUCIBLE RESEARCH

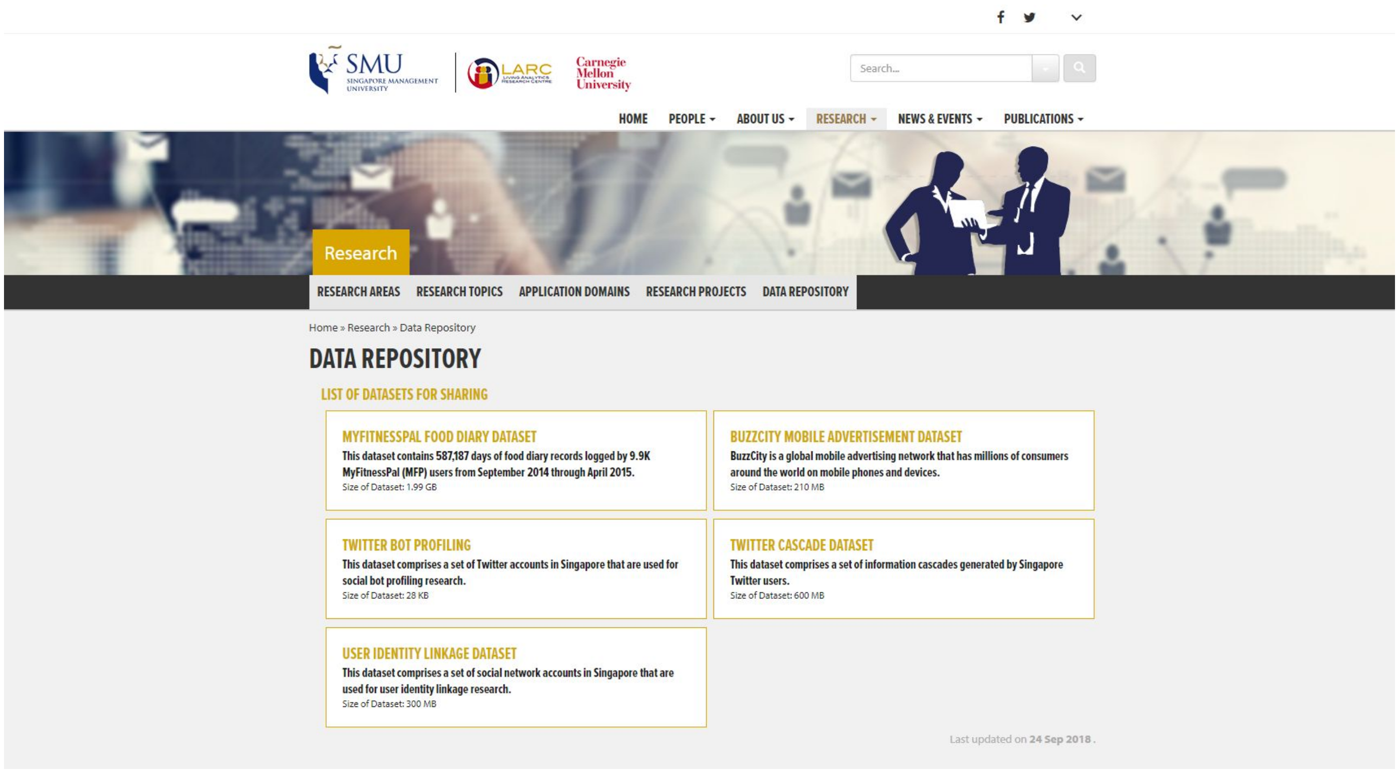
Open Research/Data Platforms

- Open Science Framework
- CodaLab
- ReScience
- Jupyter Notebooks

ENABLING REPRODUCIBLE RESEARCH

Open Data Repositories

- Microsoft Research Open Data
- Stanford Network Analysis Project (SNAP)
- UCI Machine Learning Repository
- GroupLens
- LARC Data Repository



Home » Research » Data Repository

DATA REPOSITORY

LIST OF DATASETS FOR SHARING

MYFITNESSPAL FOOD DIARY DATASET

This dataset contains 587,187 days of food diary records logged by 9.9K MyFitnessPal (MFP) users from September 2014 through April 2015.

Size of Dataset: 1.99 GB

BUZZCITY MOBILE ADVERTISEMENT DATASET

BuzzCity is a global mobile advertising network that has millions of consumers around the world on mobile phones and devices.

Size of Dataset: 210 MB

TWITTER BOT PROFILING

This dataset comprises a set of Twitter accounts in Singapore that are used for social bot profiling research.

Size of Dataset: 28 KB

TWITTER CASCADE DATASET

This dataset comprises a set of information cascades generated by Singapore Twitter users.

Size of Dataset: 600 MB

USER IDENTITY LINKAGE DATASET

This dataset comprises a set of social network accounts in Singapore that are used for user identity linkage research.

Size of Dataset: 300 MB

Last updated on **24 Sep 2018** .

LARC Data Repository

SAGAN STANDARD, UPDATED

"Extraordinary claims require extraordinary evidence and extraordinary transparency."

Aek Palakorn Achananuparp
palakorna@smu.edu.sg
@aekpalakorn

REFERENCES

- Artifact Review and Badging, ACM. https://www.acm.org/publications/policies/artifact-review-badging.
- Butler, D. (2013) When Google got flu wrong. Nature
- Camerer et al. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behavior 2.
- Collberg et al. (2014) Measuring Reproducibility in Computer Systems Research. University of Arizona Technical Report 14-04.
- Gayo-Avello et al. (2011) Limits of Electoral Predictions Using Twitter. In Proc. of ICWSM ‘11.
- Goodman et al. (2016) What does research reproducibility mean? Science Translational Medicine.
- Hutson, M. (2018) Missing data hinder replication of artificial intelligence studies. Science. http://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studies
- Lazer et al. (2014) The Parable of Google Flu: Traps in Big Data Analysis. Science.
- Pentland, A. (2012) Big Data’s Biggest Obstacles. Harvard Business Review.
- Reproducibility in Machine Learning Workshop, ICML ‘18. https://sites.google.com/view/icml-reproducibility-workshop/home
- Stodden, V. (2013) Resolving Irreproducibility in Empirical and Computational Research. IMS Bulletin Online.
- Stodden et al. (2016) Enhancing reproducibility for computational methods. Science, 354(6317).