City University of Hong Kong
Course Syllabus

offered by School of Data Science
with effect from Semester A 2024/25

Part I Course Overview

Course Title:	Social Foundations of Data Science
Course Code:	SDSC8011
Course Duration:	One Semester
Credit Units:	3
Level:	R8
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	Nil
Precursors: *(Course Code and Title)*	Nil
Equivalent Courses: *(Course Code and Title)*	Nil
Exclusive Courses: *(Course Code and Title)*	Nil

Part II Course Details

1. Abstract

As a large part of the current data used by data scientists are created by or about humans, it is necessary and helpful to introduce to the students of data science social processes and mechanisms underlying human data. In particular, the course focuses on key issues that give rise to various challenges and/or biases in human data such as self selection, social desirability, cognitive limitations, nonindependence, contextual effects, ecological fallacy, and etc. The disposition of these issues will be inherently linked to technical issues such as sampling, measure errors, statistical control, causal inference, and etc. At the end of the course, the students are expected to be able to identify potential human errors or biases in the existing data science literature, develop systematic approaches to addressing the problems, and implement some of the suggested remedies with real world data.

2. Course Intended Learning Outcomes (CILOs)

No.	CILOs	Weighting (if applicable)	Discovery-enriched curriculum related learning outcomes		
			A1	*A2*	*A3*
1.	Explain clearly fundamental characteristics of the processes and mechanisms underlying the creation, analysis, and interpretation of human data	10%	✓		
2.	Discuss existing data science literature that involves human data to identify common symptoms, causes and consequences of human data with fatal social and/or cognitive biases	20%	✓	✓	
3.	Design appropriate procedures, measures, analytical approaches, and interpretation rules to address common problems in human data	30%	✓	✓	✓
4.	Apply the proposed remedies to real world applications, demonstrating the effectiveness and resulting.	40%	✓	✓	✓
		100%			

A1: *Attitude*
Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.

A2: *Ability*
Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to self-life problems.

A3: *Accomplishments*
Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.

3. Learning and Teaching Activities (LTAs)

LTA	Brief Description	CILO No.						Hours/week (if applicable)
		1	2	3	4			
Lecture	Students will engage in formal lectures and relevant class discussions.	✓	✓	✓	✓			39 hours in total
Critical review of relevant literature	Students will engage in learning through critiques of common problems in the data science literature involving human data.	✓	✓	✓	✓			in or after classes
Research design assignments	Students will work on in-class or take-home assignments involving the development and design of remedy procedures, measure, analytical approaches, and interpretation rules.	✓	✓	✓	✓			in or after class
Empirical research paper	Students will engage ini actual implementations of the proposed remedies and writing up the results in a publishable quality.	✓	✓	✓	✓			after class

4. Assessment Tasks/Activities (ATs)

Assessment Tasks/Activities	CILO No.						Weighting*	Remarks
	1	2	3	4				
Continuous Assessment: _100_%								
<u>Critique Essays</u> Students are required to review a variety of published studies that involve potential human errors with sampling, measurement, analysis, or interpretations.	✓	✓					20%	
<u>Design Reports</u> Students are required to develop specific remedies to address the identified common problems in the literature.	✓	✓	✓				30%	
<u>Research Paper</u> Students are required to interpret the proposed remedies in real world application contexts.	✓	✓	✓	✓			50%	
							100%	

5. Assessment Rubrics

Applicable to students admitted from Semester A 2022/23 to Summer Term 2024

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B)	Marginal (B-, C+, C)	Failure (F)
Critique Reviews	Ability to identify the symptoms, causes, and consequences of various errors and biases in human data.	High	Significant	Basic	Not even reaching marginal levels
Design Assignments	Ability to propose appropriate procedures, instruments, and test measures to address common errors and biases in human data.	High	Significant	Basic	Not even reaching marginal levels
Research Paper	Ability to carry out one or more empirical study(ies), implementing the proposed remedies to demonstrate improved quality in the resulting findings.	High	Significant	Basic	Not even reaching marginal levels

Applicable to students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Marginal (D)	Failure (F)
1. Critique Reviews	Ability to identify the symptoms, causes, and consequences of various errors and biases in human data.	High	Significant	Moderate	Basic	Not even reaching marginal level

2. Design Assignments	Ability to propose appropriate procedures, instruments, and test measures to address common errors and biases in human data.	High	Significant	Moderate	Basic	Not even reaching marginal level
3. Research Paper	Ability to carry out one or more empirical study(ies), implementing the proposed remedies to demonstrate improved quality in the resulting findings.	High	Significant	Moderate	Basic	Not even reaching marginal level

Part III Other Information (more details can be provided separately in the teaching plan)

1. Keyword Syllabus

Characteristics of human data, self-selection biases, social desirability biases, cognitive biases, aggregation biases, found data, made data, physical control vs. statistical control, causal inference, sampling of social units, research design, online experiment, behaviour analytics, text mining, research ethics.

2. Reading List

2.1 Compulsory Readings

1.	Babbie, E. (2007). *The practice of social research, 11th edition.* Wadsworth Publishing. Reprint, Tsinghua University Press.
2.	Salganik, M. J. (2018). Bit by bit: Social research in the digital age. Princeton University Press.

2.2 Additional Readings

1.	Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.
2.	Watts, D. J. (2013). Computational social science: Exciting progress and future directions. *The Bridge on Frontiers of Engineering*, *43*(4), 5-10.
3.	Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.
4.	Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295.
5.	Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.
6.	Ortiz-Ospina, E., & Roser, M. (2018). Happiness and life satisfaction. *Our World in Data* [online resource].
7.	Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854-856.
8.	Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011, March). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 705-714). ACM.
9.	Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201320040.
10.	Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality

	and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854-856.
11.	Zhu, J. H., Milavsky, J. R., & Biswas, R. (1994). Do televised debates affect image perception more than issue knowledge? A study of the first 1992 presidential debate. *Human Communication Research*, *20*(3), 302-333.
12.	Reinemann, C., & Maurer, M. (2005). Unifying or polarizing? Short-term effects and postdebate consequences of different rhetorical strategies in televised debates. *Journal of Communication*, *55*(4), 775-794. (newly added)
13.	Liang, H., & Zhu, J. J. H. (2017). Big data, collection of (social media, harvesting). In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *International Handbook of Communication Methods*, Wiley & Sons.
14.	Leskovec, J., & Faloutsos, C. (2006, August). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 631-636). ACM. (newly added)
15.	Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013, July). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *ICWSM*. (newly added)
16.	Xu, X. K., & Zhu, J. J. H. (2016). Flexible sampling large-scale social networks by self-adjustable random walk. *Physica A: Statistical Mechanics and its Applications*, *463*, 356-365.
17.	Zhu, J. J. H., Mo, Q., Wang, F., & Lu, H. (2011). A random digit search (RDS) method for sampling of blogs and other web content. *Social Science Computer Review*, 29(3), 327-339.
18.	Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012-U1014.
19.	Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*, 1203-1205.
20.	Burke, M., & Kraut, R. E. (2014). *Growing closer on Facebook: changes in tie strength through social network site use*. Proceedings of the 32nd annual ACM conference on Human factors in computing systems, Toronto, Ontario, Canada. http://thoughtcrumbs.com/publications/burke_chi2014.pdf.
21.	Xu, P. P., Wu, Y. C., Wei, E. X., Peng, T. Q., Liu, S. X., Zhu, J. J. H., & Qu, H. M. (2013). Visual Analysis of Topic Competition on Social Media. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2012-2021.
22.	Aral, S., & Nicolaides, C. (2017). Exercise contagion in a global social network. *Nature Communications*, *8*, 14753.
23.	Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. Academy of Management review, 19(2), 195-229.
24.	Ackland, R., & Zhu, J. J. H. (2015). Social network analysis. In P. Halfpenny & R, Procter (Eds.), *Innovations in digital research methods* (pp. 221-244). Sage

	Publications.
25.	Murdock, J., Allen, C., & DeDeo, S. (2017). Exploration and exploitation of Victorian science in Darwin's reading notebooks. *Cognition, 159*, 117-126.
26.	Peng, T. Q., Liu, M. C., Wu, Y. C., & Liu, S. X. (2016). Follower-followee Network, Communication Networks and Vote Agreement of U.S. Members of Congress. *Communication Research, 43*, 996-1024.
27.	Zhang, L., Zheng, L., & Peng, T. Q. (2017). Structurally embedded news consumption on mobile news applications. *Information Processing & Management, 53*, 1242-1253.
28.	Zhu, J. J. H., Chen, H. X., Peng, T. Q., Liu, X. F, & Dai, H. X. (2018). How to measure sessions of mobile device use? Quantification, Evaluation, and Applications. *Mobile Media & Communication, 6*, 215-232.
29.	boyd, danah (2016). Untangling research and practice: What Facebook's emotional contagion study teaches us. *Research Ethics, 12*, 4-13.
30.	Goroff, D. L. (2015). Balancing privacy versus accuracy in research protocols. *Science, 347*, 479-480.
31.	Open Science Collaboration. (2015). Estimating the reproducibility of psychological science, *Science, 349*, aac4716.