



ITF Seminar 2



ESG 研究中心  
RESEARCH CENTRE FOR ESG  
香港恒生大學  
THE HANG SENG UNIVERSITY  
OF HONG KONG



Funding organisation

$\pi$  創新科技署

Innovation and Technology Commission

# APPLICATION OF BIG DATA ANALYTICS ON ESG CORPORATE PERFORMANCE: A NEWS AND SOCIAL MEDIA PERSPECTIVE

15th Nov 2023



Copyright by  
*Prof Louis Cheng*

Special thanks to  
*Dr Maureen Chan and Ms Kirsten Tsui*

Project Title	Fostering Big Data Applications and Innovation to Reflect Corporate ESG Performance through Media and Social Listening (GSP/033/22)
Funding Body	Innovation and Technology Fund
Project Organiser	Research Centre for ESG (CESG), HSUHK
Supporting Organisations	Excellent Global and Master Insight
<b>Seminar 2 - Application of Big Data Analytics on ESG Corporate Performance: A News and Social Media Perspective</b>	
Date and Time	15th November 2023, 2:30 pm to 5:30 pm
Location (Hybrid)	Rooms 1101-02, 11/F, Excellent Global Business Centre, Euro Trade Centre, 13-14 Connaught Road, Central
Presenters	Prof Louis Cheng, HSUHK Ms Florence Li, HSUHK Ms Vivi Hu, YoujiVest Mr Hauman Yeung, Ascent Partners

## DISCLAIMER

This report is prepared by The Research Centre for ESG at the Hang Seng University of Hong Kong. This report and the results should be used only as a reference in understanding the ESG and big data. It is not an offer to sell or solicitation for an offer to buy any financial products and services and they should not be considered as investment advice. The Research Centre for ESG, Excellent Global, Master Insight, the research team, and the consultation participants accept no responsibility or liability for the accuracy and completeness of such information in the report. Any opinions, findings, conclusions or recommendations expressed in this material/event (or by members of the project team) do not reflect the views of the Government of the Hong Kong Special Administrative Region, the Innovation and Technology Commission or the Vetting Committee of the General Support Programme of the Innovation and Technology Fund.



# INTRODUCTION

Professor Louis Cheng kicked off the seminar which was a continuation of the sharing and discussion of the ESG issues starting from the Seminar 1. This seminar began with a somewhat technical presentation on employing big data analytics and machine learning tools for ESG evaluation while professional expertise in reviewing the ESG reporting mechanism was found critical to meet the challenges of evaluation problems. The illustrations with a number of 3D Python visualisation diagrams helped to show the implications of ESG news and social media sentiment. Finally, a comparison of the *Top100* and *Bottom100* firms from the *i-Composite Scorecard* was made with the visualisation diagrams. The comparative results indicated that both the *Top* and *Bottom100* listed firms need to pay attention to the sentiment scores since stock price of the firms would be affected to some extent.

# OPENING REMARKS & OVERVIEW OF ESG RATING DIVERGENCE

## Prof Louis Cheng

Professor Cheng reiterated that the seminar was a platform for participants to interact and learn how to use big data analytics, FinTech and machine learning tools with the aim to increase the ESG intelligence of firms, NGOs and various business entities. The rich technical contents of this Seminar 2 were presented with a number of different perspectives introduced from Seminar 1 so that participants would be continuously guided to widen their knowledge horizon on ESG knowledge. Participants would obtain insights into how to detect problems or errors from ESG reports of companies and be given suggested solutions as references. The upgraded 2.0 version, following the introduction of the 1.0 version in Seminar 1 held in August 2023, was announced and the *i-Composite Score* of the *Top100* and *Bottom100* firms out of the 511 listed firms in Hong Kong were presented. More information can be found on [the ITF website](#). A more advanced 3.0 version would be expected to be shown to the public in the forthcoming conferences in 2024.

### *Theoretical background of the i-Score creation*

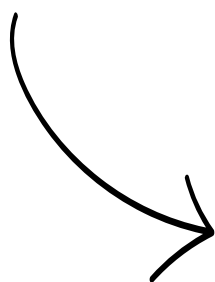
In this Seminar 2, Professor Cheng explained the design mechanism of *i-Score* to facilitate companies to design their own version of *i-Score*. As mentioned in Seminar 1, different measurement metrics are used by different rating agencies, which have led to disagreements or divergence of the ESG data for the same company. This causes confusion to both users and investors. Berg, Kölbl & Rigobon (2022) studied this aggregate confusion phenomenon in their paper.

The *i-Composite (i-C) Score*, as a solution to the confusion, is to put the mean as the numerator divided by some sorts of adjusted figure as the denominator. Basically, *i-C Score* is like a divergency-adjusted rating.

### There are three key steps for working out the *i-C* Score:

#### Step 1:

It is to obtain the ESG ratings of listed firms, which are based on performance but not risk-or-disclosure-based.



#### Step 2:

For each listed firm, it is to compute a divergence factor based on various ESG ratings of each firm from the different data providers to capture the divergence effect.

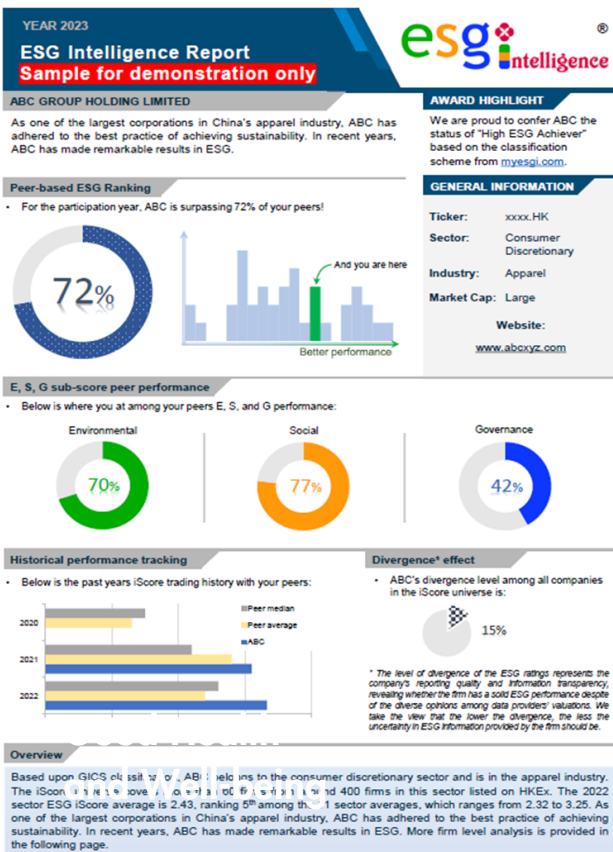


#### Step 3:

The *i-Composite (i-C) Score* for each firm is computed by scaling the ESG rating by the divergence factor generated in step 2.

The “Aggregate Confusion Hypothesis” developed by Berg, Kölbel and Rigobon (2022) suggests that the rating divergence comes from 3 aspects: Scope, Rater, and Weighting differences. “Scope” refers to the type(s) of issues attributed under the ESG framework. ESG analysts (i.e., the raters) employed by different ESG data providers are subjected to regional and corporate culture influence in rating companies, thus leading to the divergence. The third reason is the weighting of ESG being applied to different industries or even different forms of firms, for example, a conglomerate. The different weighting system will also affect the final ESG data of a firm. Firms can also follow the simple calculation process as discussed to work out a firm-based *i-Score*.

Professor Cheng further illustrated the use of *i-Score* with an individual company’s ESG intelligence report based on the available data as of December 2022.



Disclaimer: This report is prepared by Prof. Louis Cheng and Angela Zhao. This report and the results should be used only as a reference in understanding the historical ESG performance of the entity concerned. It is not an offer to sell or solicitation for an offer to buy any financial products and services, and they should not be considered as investment advice. The Research Centre for ESG of Hong Kong University of Hong Kong, The S. H. Ho Foundation, and the research team accept no responsibility or liability for the accuracy and completeness of such information in the report.

Disclaimer: This report is prepared by Prof. Louis Cheng and Angela Zhao. This report and the results should be used only as a reference in understanding the historical ESG performance of the entity concerned. It is not an offer to sell or solicitation for an offer to buy any financial products and services, and they should not be considered as investment advice. The Research Centre for ESG of Hong Kong University of Hong Kong, The S. H. Ho Foundation, and the research team accept no responsibility or liability for the accuracy and completeness of such information in the report.

The report will inform the peer-based ESG ranking in terms of the overall ESG and ESG sub-score peer performance. By comparing with different benchmarks, the performance of the company in terms of divergence can be known. It follows with a description of how the company performs in terms of categories of achievement (for example, satisfactory or very satisfactory), being provided through the knowledge of the percentile level of achievement of each E. S. and G. This information of the 511 listed firms in Hong Kong can now be found on this [website](#).

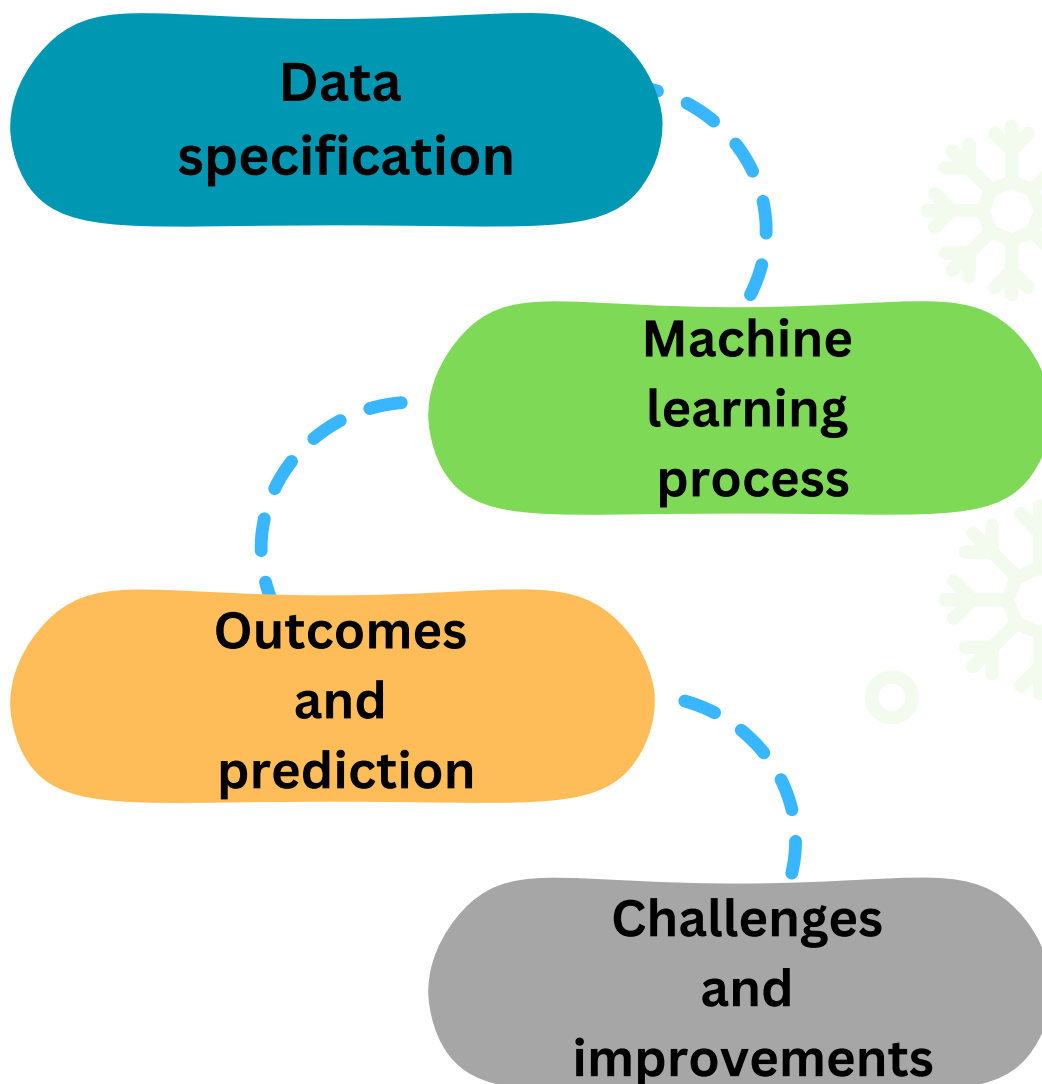
The ESG report of a company is annual, and the reported information tends to be outdated when it is published six months after the data collection period. To mitigate the time lag effect, the sentiment analysis will be conducted to provide more updated ESG data of a firm, say only 2 or 3 months after data inputs. In terms of sentiment score, ESG will be generated separately. Accordingly, negative sentiment, positive sentiment and net sentiment can be worked out. Somehow, those scores are found to be correlated with stock price and it shows the importance of sentiment data accordingly. The provision of annual report is found insufficient for provision of updated ESG performance data of a company.

# DEMONSTRATING HOW TO USE MACHINE LEARNING TO PROCESS ESG SENTIMENT DATA TO CAPTURE ESG

## Ms Florence Li

Through the course, Ms Florence Li, with the acquisition of an educational AI platform, will teach students how to use machine learning to predict the *i-Score* so as to facilitate business decisions. Currently, only 511 out of the 2,600 listed firms in Hong Kong have rating information for computing the *i-Score*. The other firms of about 2,000 may not draw the attention of the rating agencies due to their lower liquidity; however, these firms still have to be evaluated on their ESG performance. Using AI to predict their ESG performance may be a solution.

Ms Li explained the ESG data processing with the following flow:



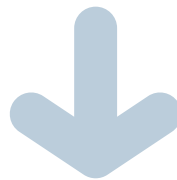


With regard to data specification, she obtained the required data from Professor Cheng. The raw data (2019 to 2022) provided are to be input into the computer for pre-processing while enabling the computer to undergo an auto-machine learning process. The following table presented the required pieces of data for predicting the *i-Score*.

Variables of company profile	Financial Data
Stock Code	Annual Stock Price Return (%)
BBG Code	Market Cap (HKD)
Company Abbreviation	Total Asset (HKD)
Name	Total Debt to Total Assets (%)
GICS1	Total Debt (HKD)
GICS2	Common Equity (HKD)
Year	Common Equity (Share)
	Net Income (HKD)
	ROA (%)
	ROE (%)

The given data have to be pre-processed before inputting into the computer.

**Original *i-Score*:**  
1- 6 in 2-3 decimal places



**Processed *i-Score*:**  
Level 1: 1,2  
Level 2: 3,4  
Level 3: 5,6

The machine learning process will undergo 5 action steps as follows

The original *i-Score* will be rounded up and divided into 3 levels of numbers. The machine learning process will follow the following 5 action steps:

1. Get the data

2. Clean, prepare and manipulate the data: As described above, the data (original *i-Scores*) have to be pre-processed for the computer to understand the language and the meaning of the data.

3. Train the model built

4. Test the data

5. Improve the model



After training the model, it has to be tested to ensure its reliability and validity. The model has to be refined and improved as when a model is generated, the accuracy may not reach 100% yet, though this may never be achieved. Through this machine learning process, the accuracy of predication has been improved from 15% to 55%. The acquisition of the machine learning tool "Auto Sklearn" is to provide a user-friendly platform to enable students to learn to predict the *i-Score* for a firm by simply inputting the data into the required fields for the results.

Ms Li identified three challenges involved in the *i-Score* prediction process:

1. Limited (or absent) data governance: Robust data governance is crucial for maintaining the data integrity. Not all firms have a solid governance framework, especially the firms with lower ESG maturity.
2. No single source of truth: Due to the absence of a centralized system of record, which hinders efficient data collection, *i-Score* is supposed to fill this gap to supply a single source of truth.
3. Limited data quality and reliability: Though ESG has been discussed for a long time, there are still limited ESG data quality and reliability since it does not seem to directly affect the revenue of a company.

Ms Li suggested three methods for the improvement of the accuracy of the model:

1.To increase the database size of the currently around 700 entries with data available around the world. The accuracy will be enhanced a lot as the AI machine will continue the learning with more data.

2.To use a more advanced AI tool where more types of data, in addition to company profiling data and corresponding financial data, can be processed.

3.To add more specific ESG data.

However, for the example just shown, based on the company profiling information and the corresponding financial data, the model has already been trained to 55% accuracy. It implies that the financial situation of a company is closely related to the ESG development.

# UNDERSTANDING THE DESIGN OF <sup>10</sup> BIG DATA ANALYTICS ON MEASURING ESG PERFORMANCE

## Ms Vivi Hu

Ms Vivi Hu introduced the China Environment Pollutant Discharge data, which originates from real-time monitoring sensors embedded at over 400,000 firms that are involved in heavy pollution business. It covers over 500 types of pollutants such as wastewater, various gases and solid particles. The data are very objective in presenting the actual behaviours of the firms involved. The data are important for academic research as well because they are production data that are objectively collected. The firms running hedge funds and mutual funds should take care of the data as they will affect their investment decisions.

Ms Hu shared some practical cases to facilitate research interest:

1. The availability of the real-time data also arouse research interest and one of the areas would be the exploration of the government responding behaviours towards the pollutants' discharge behaviours. In addition, the public view and the firm factors could be added to the analysis of factors causing the actual behaviours.
2. A top-tier company had to shut down because of huge financial loss as a consequence of its failure to renew the license for discharging pollutants.
3. Through a graphic presentation of backtesting of pollutants' discharge, it was indicated that some categories of pollution did contribute to good performance portfolios of firms, which then led to more excess return in terms of bad behaviours of firms. YoujiVest had investigated further and could share insights with interested researchers.
4. Showing photos of flooding in the Forbidden City and a top luxury hotel reminds the frequent discussion of climate risk in terms of flooding. Being destroyed by flooding, a case of extreme weather condition, not only the affected company and the environment bear the consequences, but also all stakeholders (e.g., banking institutions and insurance companies) face such actual life-threatening challenges.

Ms Hu concluded that the ESG issue is not only caused by the corporate emission industry but also by other economic activities like agriculture and mining industries, which can also cause high physical risk and financial loss across industries. With the many problems encountered due to the physical risks, there are more mandates of disclosure requirements for tightening the regulatory mechanism. YoujiVest then launched a Climate Risk Model System-YoujiCRMS in the first half of 2023. A lot of different types of data are collected, which include the global satellite and climate data as well as localised data.

Concerning that different industries have different sensitivity towards different hazards, YoujiVest also considers industrial level of risk. For example, some manufacturing industries would have a stronger linkage to temperature. The hydro-electricity power industry has a high sensitivity to drought. In another study, the quality and quantity of cow milk production is found to decrease with higher temperature. Therefore, the industry impact should be an item to measure. Accordingly, a lot of climate risks may be introduced in the insurance industry.

With the availability of data for understanding the climate issues, action can be taken to identify the risks involved and to follow up actions about mitigation and adaption. In the future, there can be more discussions on how to leverage climate changes.

# CHALLENGES OF EVALUATING ESG DISCLOSURE OF HONG KONG LISTED FIRMS

## Mr Hauman Yeung

Mr Hauman Yeung shared his experiences of the challenges on evaluation of ESG reporting. He remarked that with a huge amount of data, it was difficult to know the quality of the data in terms of the accuracy, completeness, relevancy, the meaning of the messages and the data presented.

Mr Yeung elaborated the reporting problems from extracts of companies' ESG reports. He spotted a missing but important reporting item from a report. He claimed that he could spot 4 to 5 of those obvious missing errors in 10 to 20 ESG reports. In another report from a dairy company, he spotted not only an important missing emission chemical, but also a huge discrepancy between reported and actual emission amount for another key reporting item. For reporting wastewater discharge in terms of COD (Chemical Oxygen Demand) with million tonnes, the meaning of COD was not complete, and the water quality had yet not been known. A cosmetic company in Hong Kong did not report according to the additional new disclosure requirements. The company responded that it did not have any involvement in the climate change and new disclosure requirements and therefore they took no action to report. Another company had a lot of information on recruitment, dismissal, remuneration, etc., but it had not touched on diversity, inclusion and equity.

Furthermore, he raised three questions about the content of the report:

- Are those contents unbiased?
- How meaningful are those contents?
- What can those contents tell?

The ESG reporting problems will affect the decision-making of investors. For improvement, Mr Yeung believed that this kind of material errors can be solved by FinTech or some other technologies if regulators can design some kind of disclosure system and frame it as a requirement. However, Mr Yeung advised that human judgment is also important for evaluating the ESG reports.

# PRESENTATION OF 2.0 SCORECARD

## Prof Louis Cheng

Professor Cheng announced the launch of *version 2.0 Scorecard*. The *i-Composite Scorecard v1.0* was created in August 2023 to showcase the *Top100 (v1.0)* from the *Top500* universe. The *Scorecard v1.0* adopts a ranking mechanism based on the absolute *i-Composite score*. The *Scorecard v2.0* employs a different ranking mechanism by integrating peer benchmarking into the methodology. The *v2.0* system also compares the *Top100 (v2.0)* with the *Bottom100 (v2.0)* in the *Scorecard* using six 3D scattergrams in Python.

### *The findings of the comparison*

3D scattergrams in Python	The findings
Figure 1- <i>Composite ESGi (i-C) Score vs RP and YJV Net Senti: Top100/ Bottom100 (Mar 2023)</i>	The <i>Bottom100</i> appears to be more clustered in the middle range for both RP and YJV Net Senti than the <i>Top100</i> , which is more spread out. This indicates that <i>Top100</i> firms experience a wider range of sentiment in both positive and negative dimensions
Figure 2- YJV Net Senti vs YJV Mainstream and Social Media Net Senti: <i>Top100/ Bottom100 (Mar 2023)</i>	<i>Top100</i> demonstrates a more spread-out pattern for all three dimensions: Net Senti, Mainstream Net Senti and Social Media Net Senti. On the other hand, <i>Bottom100</i> has a very limited range for Social Media Net Senti.

3D scattergrams in Python	The findings
Figure 3- <i>Composite ESGi (i-C) Score vs YJV Mainstream Negative and Positive Senti: Top100/ Bottom100 (Mar 2023)</i>	The <i>Top100</i> is more widespread for both Mainstream Positive and Negative Senti than the <i>Bottom100</i> .
Figure 4 – <i>Composite ESGi (i-C) Score vs YJV Social Media Negative and Positive Senti: Top100/ Bottom100 (Mar 2023)</i>	The <i>Top100</i> is more widespread for both Social Media Positive and Negative Senti than the <i>Bottom100</i> .
Figure 5 – <i>Composite ESGi (i-C) Score vs Net Senti and Market to Book Ratio: Top100/ Bottom100 (2022)</i>	<i>Top100</i> is more widespread for both Net Senti and Market to Book Ratio than the <i>Bottom100</i> .
Figure 6 – <i>Composite ESGi (i-C) Score vs Net Senti and Profit Margin: Top100/ Bottom100 (2022)</i>	<i>Top100</i> is more widespread for both Net Senti and Profit Margin than the <i>Bottom100</i> .



### *The implication: ESG performance erosion due to sentiment*

Sentiment matters as it enhances overall *i-C* (*i-Composite*) performance if done properly. Even though all firms in the *Top500* are large firms with sufficient resources, there are still significant differences in sentiment performance between the *Top100* and the *Bottom100*. From another perspective looking at this issue, there are 15 firms with negative sentiment among the *Top100*, while there are 34 firms having negative sentiment among the *Bottom100*. Since our *Top100* and *Bottom100* have the same number of firms in each sector, the numbers of negative sentiment firms are free from industry bias.

A conclusion can be drawn from 3D scattergrams that being the top 100 in *i-C Score* does not mean that they can control the sentiment, even though they have a large market cap. It is difficult to or even cannot control the media's reaction to firms in terms of ESG performance. Rather, larger firms will be subject to negative sentiment more easily because social media aims at drawing the attention of the public to something big. Therefore, it may be more challenging for the top 100 firms to control medias ESG sentiment. In short, the ESG performance erosion due to sentiment is an important issue for the *Bottom100*. The *Bottom100* firms should seriously consider stepping up their ESG communication strategy to avoid sentiment erosion of ESG performance.

# CONCLUSION

With rich content and active sharing, Seminar 2 has a lot of concluding remarks. The *i-C Score* is a solution to confusion caused by the various data providers. Human judgment is still important in evaluating the ESG reports until the development of a more advanced AI learning tool which can replace human effort. ESG can affect the movement of the stock price of a firm when ESG is reported as risks. Employees have close linkage to a firm and their complaints would lead to social risk of a firm. Therefore, managing ESG information communication now becomes a challenge for firms of different scale of operations and in particular the labour-intensive industries. With the available satellite data for the flooding issue, preventive measures can be taken to avoid potential risks as well as huge financial loss across industries. If there is no action in advancing the technology for adaptation of lifestyle consumption, human beings will eventually suffer other kinds of loss (e.g., more air-conditioning and less productive cows). Every data provider agency has unique features and has taken effort to collect the ESG data accurately through, for example, onsite visits and reading the ESG reports. The materiality issues are different in different firms. The choice of the data provider agencies depends on the specific industries and the needs of firms. Meanwhile, it is time to think about how to allocate resources between fixing the carbon reduction and fixing the lifestyle consumption issues. People should start to anticipate the implications of ESG to business in the next 10 to 15 years. Looking forward, there would be the *i-C Score v3.0*.