

# LET'S TALK DIGITAL

JANUARY 2021

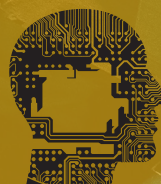
ASIAN  
BANKING  
SCHOOL

## CONTENT



### An Introduction to Optical Recognition Character

*Koh Wyhow*



### Cybersecurity Challenges for Banks During the Pandemic

*Kayne Hoo*



**Let's Talk Digital is a monthly newsletter that was created to build awareness on Digital Banking and provide a platform for industry practitioners to share insight and current trends on this exciting subject matter in relation to the Banking and Finance industry.**

## **An Introduction to Optical Recognition Character**

**By Koh Wyhow**

All of us have come across this problem before. "How do I convert data from this PDF/image into a spreadsheet?" Depending on the information you are looking for, you can resort to converters found online, or spend a short amount of time doing the data entry work. What if you are required to do this periodically? This article shows a few ways you can extract such data using Tabula, and Google's Vision API.

## **Cybersecurity Challenges for Banks During the Pandemic**

**By Kayne Hoo**

The COVID-19 pandemic brings about significant changes to the business environment and operations in most organizations. For the banking industry, contactless banking and remote working have become the new normal. While the adoption of a different routine may have increased the convenience and safety of its customers and employees, it also opens up new challenges for banks to reassess their cybersecurity defences.

This article will be discussing some of the relevant challenges encountered and its mitigation in face of the pandemic.

**To find out more about the Digital Banking programmes that ABS offers, visit**

**[www.asianbankingschool.com/our-programmes/centre-for-digital-banking](http://www.asianbankingschool.com/our-programmes/centre-for-digital-banking)**



## Koh Wyhow

Koh Wyhow is the manager of the data science team at Star Media Group Berhad. He focuses on delivering advanced analytics and business intelligence solutions for the organisation like chatbots and image recognition solutions. He consulted for client in the airlines, media, property, and FMCG industries during his time as a senior consultant at EY's Data and Analytics team.

He was one of the data scientists which implemented strategies to run a national data-driven campaign for INVOKE in the 14th General Elections. As an independent learner, he picked up basic Python programming skills after office hours during his days as a Further Mathematics lecturer at a private college. Wyhow holds a BSc in Mathematics from the National University of Singapore.



## Kayne Hoo

Kayne Hoo Kah Yan is experienced in project planning, project resource management, digital forensics and penetration testing. She has practical experiences in cyber security incident handling, computer crime investigation and is experienced in performing digital forensics analysis for various cases and in maintaining the chain-of-custody and preparing a court admissible report.

Additionally, she also conducts technical assessments such as vulnerability assessments, host security assessment, network device configuration security assessment and more.

Kayne Hoo Kah Yan has conducted digital forensics investigations on various cyber security attacks such as server intrusions, defaced websites and ransomware attacks, as well as provide consultation to clients on improving their IT security to prevent recurring incidents.

She has also supports lead forensic examiner in digital forensics analysis in relation to computer crime in addition to preparing chain-of-custody and digital forensics reports for use in prosecution.

She has assisted several clients in strengthening their IT security posture through managing and performing technical assessments based on international IT security industry baselines and benchmarks.

# AN INTRODUCTION TO **OPTICAL RECOGNITION CHARACTER**

*By Koh Wyhow*

Some of our work involves manually reviewing and processing physical pages of documents, ranging from invoices to packing lists and insurance certificates, or to retrieve information from documents. It is a tedious process which can be automated by converting an image or a PDF (Portable Document Format) file into either a table or text. In this article, I will be introducing a few ways to retrieve text from PDFs and images.

For the first problem, I'll look at a summary of survey responses conducted by a news agency based in the US . . . Most PDF reports found online tend to be machine readable, so you can highlight parts of the report which contain the information you are after. It is usually easier to extract information from these documents as the Python code needed to perform this function is easy to write.

**Table 12. Nonmember Survey Response Rates by Region and Specialty, 2018**

Specialty	Midwest (%)	Northeast (%)	South (%)	West (%)	Total (%)
Cancer	30.2	39.4	40.3	29.7	35.4
Diabetes & Endocrinology	34.5	20.5	28.3	13.9	24.4
Ear, Nose & Throat	16.7	33.3	29.0	39.0	29.9
Gastroenterology & GI Surgery	24.3	37.8	24.6	12.8	24.7
Geriatrics	15.2	15.6	13.3	13.8	14.5
Gynecology	25.8	21.2	28.3	13.5	23.0
Heart & Heart Surgery	36.1	43.6	18.8	32.4	30.6
Nephrology	25.7	21.2	17.9	17.1	19.9
Neurology & Neurosurgery	20.6	28.1	17.5	25.7	22.2
Ophthalmology	28.1	23.7	21.7	24.4	24.0
Orthopedics	25.7	25.0	19.7	21.3	22.2
Psychiatry	29.2	15.4	17.1	16.7	18.6
Pulmonology	13.8	16.2	32.7	31.6	25.2
Rehabilitation	21.2	28.2	32.7	12.5	24.2
Rheumatology	13.3	18.9	26.9	7.7	17.7
Urology	23.5	38.7	25.4	10.5	24.1
<b>Overall Response Rate</b>	<b>24.2%</b>	<b>26.3%</b>	<b>24.8%</b>	<b>20.3%</b>	<b>24.0%</b>

[https://www.usnews.com/static/documents/health/best-hospitals/BH\\_Methodology\\_2018-19.pdf](https://www.usnews.com/static/documents/health/best-hospitals/BH_Methodology_2018-19.pdf)



The image above shows an excerpt of the survey response rates by a hospital's departments. There are a few Python libraries available which can read text from PDF documents like PyPDF2, Textract, Apache Tika, pdfPlumber, and pdfMiner3. To read tables from PDF documents, I usually resort to a library called Tabula .

The raw result from Tabula is as follows:

```
[2] 1 import tabula
    2
    3 # Read pdf into list of DataFrame
    4 sample_list = tabula.read_pdf("BH_Methodology_2018-19.pdf", pages='45')
    5
    6 sample_list
```

Got stderr: Dec 09, 2020 3:53:42 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider loadDiskCache  
 WARNING: New fonts found, font cache will be re-built  
 Dec 09, 2020 3:53:42 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>  
 WARNING: Building on-disk font cache, this may take a while  
 Dec 09, 2020 3:53:42 AM org.apache.pdfbox.pdmodel.font.FileSystemFontProvider <init>  
 WARNING: Finished building on-disk font cache, found 17 fonts  
 Dec 09, 2020 3:53:42 AM org.apache.pdfbox.pdmodel.font.PDTrueTypeFont <init>  
 WARNING: Using fallback font 'LiberationSans' for 'TimesNewRomanPSMT'  
 Dec 09, 2020 3:53:44 AM org.apache.pdfbox.pdmodel.font.PDTrueTypeFont <init>  
 WARNING: Using fallback font 'LiberationSans' for 'TimesNewRomanPSMT'

	Unnamed: 0	Unnamed: 1	Midwest	...	South	West	Total
0	NaN	Specialty	(%)	...	(%)	(%)	(%)
1	Cancer	NaN	30.2	...	40.3	29.7	35.4
2	Diabetes & Endocrinology	NaN	34.5	...	28.3	13.9	24.4
3	Ear, Nose & Throat	NaN	16.7	...	29.0	39.0	29.9
4	Gastroenterology & GI Surgery	NaN	24.3	...	24.6	12.8	24.7
5	Geriatrics	NaN	15.2	...	13.3	13.8	14.5
6	Gynecology	NaN	25.8	...	28.3	13.5	23.0
7	Heart & Heart Surgery	NaN	36.1	...	18.8	32.4	30.6
8	Nephrology	NaN	25.7	...	17.9	17.1	19.9
9	Neurology & Neurosurgery	NaN	20.6	...	17.5	25.7	22.2
10	Ophthalmology	NaN	28.1	...	21.7	24.4	24.0
11	Orthopedics	NaN	25.7	...	19.7	21.3	22.2
12	Psychiatry	NaN	29.2	...	17.1	16.7	18.6
13	Pulmonology	NaN	13.8	...	32.7	31.6	25.2
14	Rehabilitation	NaN	21.2	...	32.7	12.5	24.2
15	Rheumatology	NaN	13.3	...	26.9	7.7	17.7
16	Urology	NaN	23.5	...	25.4	10.5	24.1
17	Overall Response Rate	NaN	24.2%	...	24.8%	20.3%	24.0%

[18 rows x 7 columns]

After cleaning the table, a snapshot of the result should look like the following:

<https://pypi.org/project/PyPDF2/>  
<https://textract.readthedocs.io/en/stable/>  
<https://github.com/chris-mattmann/tika-python>  
<https://pypi.org/project/pdfplumber/>  
<https://pypi.org/project/pdfminer3/>  
<https://pypi.org/project/tabula-py/>

	Specialty	Midwest (%)	Northeast (%)	South (%)	West (%)	Total
1	Cancer	30.2	39.4	40.3	29.7	35.4
2	Diabetes & Endocrinology	34.5	20.5	28.3	13.9	24.4
3	Ear, Nose & Throat	16.7	33.3	29.0	39.0	29.9
4	Gastroenterology & GI Surgery	24.3	37.8	24.6	12.8	24.7
5	Geriatrics	15.2	15.6	13.3	13.8	14.5
6	Gynecology	25.8	21.2	28.3	13.5	23.0
7	Heart & Heart Surgery	36.1	43.6	18.8	32.4	30.6
8	Nephrology	25.7	21.2	17.9	17.1	19.9
9	Neurology & Neurosurgery	20.6	28.1	17.5	25.7	22.2
10	Ophthalmology	28.1	23.7	21.7	24.4	24.0
11	Orthopedics	25.7	25.0	19.7	21.3	22.2
12	Psychiatry	29.2	15.4	17.1	16.7	18.6
13	Pulmonology	13.8	16.2	32.7	31.6	25.2
14	Rehabilitation	21.2	28.2	32.7	12.5	24.2
15	Rheumatology	13.3	18.9	26.9	7.7	17.7
16	Urology	23.5	38.7	25.4	10.5	24.1
17	Overall Response Rate	24.2%	26.3%	24.8%	20.3%	24.0%

Comparing the cleaned output of Tabula and the table from the original document, , it can be observed that the results are similar. Once the Tabula output has been transformed into a table, it is easy to slice and dice the data as you see fit.



Sometimes, you do get scans of documents like the sample electricity bill to the right. The problem with scans of documents, or images of documents, is that it is slightly more difficult to get the information you would like. The reason is it is not possible to highlight any text on the image. These problems require the use of Optical Character Recognition (OCR). For this demonstration, I will be using Google Vision API's OCR.

OCR algorithms typically work like this:

- The AI isolates the area which have black inkblots against areas which do not. For areas which do have text, the algorithm draws boxes around the text. These boxes are called bounding boxes.
- The OCR algorithm then looks at individual inkblots to make out words or characters, and whether they are organized as a table or free-form text. It converts these texts into computer understandable text contained within the image.

The image with the bounding boxes is shown on the right.

The raw output of Google Vision API's OCR is as follows:

**BIL ELEKTRIK ANDA**

No. Akaun : 220001234512  
 No. Kontrak : 1002000  
 Deposit : RM350.00  
 No. Invois : 10001234  
 Ali bin Abu  
 E23A - 201 Sek 3  
 Wangsa Maju  
 53300  
 WP Kuala Lumpur

**TERIMA KASIH**  
 Kerana Membayar Dalam Tempoh 30 Hari  
 TNB Careline  
 1-300-88-5454

**JumlahPerlu Dibayar RM311.90**

	Amaun	Tarikh Bil	Bayar Sebelum
Tunggakan	RM 0.00		Terima kasih
Caj Semasa	RM 311.88		
Penggenapan	RM 0.02	01.11.2018	
Jumlah Bil	RM 311.90		
Bil Terdahulu	RM 526.85	Tarikh	
Bayaran Akhir	RM 526.85	02.08.2018	

Jenis Bacaan : **Bacaan Sebenar**

Tempoh Bil	Tarif	Faktor Prorata
03.09.2018 - 03.10.2018 (31 Hari)	A : Kediaman	1.00000

Blok Tarif (kWh)	Blok Prorata (kWh)	Kadar (RM)	Amaun (RM)
200	200	0.218	43.60
100	100	0.334	33.40
300	300	0.516	154.80
300	130	0.546	70.98
Jumlah	730		302.78

```
...
description: "BIL ELEKTRIK ANDA\nTENAGA\nNASIONAL\nNo. Akaun :
220001234512\n: 1002000\n: RM350.00\nTERIMA KASIH\nNo. Kontrak\
nKerana\nDeposit\nNo. Invois\nMembayar Dalam\nTempoh 30 Hari\n:
10001234\nAli bin Abu\n3\nTNB Careline\n1-300-88-5454\nE23A - 201
Sek 3\nWangsa Maju\n53300\nWP Kuala Lumpur\nTarikh Bil\nJumlahPerlu
Dibayar RM311.90\n03 Okt 2018\nAmaun\nBayar Sebelum\nTunggakan\nCaj
Semasa\nPenggenapan\nRM\n0.00\nTerima kasih\nRM 311.88\n0,02\nRM\
nJumlah Bil\nRM\n311.90\n01.11.2018\nAmaun\n52685\nTarikh\n5\nBil
Terdahulu\nRM\n02.08.2018\nBayaran"
```

```

Akhir\nRM\n526.85\n30.08.2018\nJenis Bacaan\nBacaan Sebenar\nTempoh
Bil : 03.09.2018 - 03.10.2018 (31 Hari)\nTarif\nFaktor Prorata\
n:A: Kediaman\n1,0000\nBlok Tarif (KWh)\n200\n100\n300\n300\nBlok
Prorata (kWh)\n200\n100\n300\n130\nKadar (RM)\n0.218\nAmaun (RM)\
n43.60\n33.40\n154.80\n70.98\n8\n0.334\n0.516\n0.546\nJumlah\n730\
n302.78\nTidak Kena\nST\nKena\nST\nKeterangan\nJumlah\nKegunaan
kWh\nKegunaan\nkWh\n600\n130\n730\nRM\n231.80\n70.98\n302.78\
nKegunaan Bulan Semasa\nService Tax (66)\nKWTBB (1.6%)\nRM\n231.80\
n70.98\n302.78\n4.26\n4.84\n10\nCaj Semasa\nRM\n311.88\nBacaan
Meter\nNo Meter\n320\n232egunaan\nUnit\nDahulu\nSemasa\n311201234\
n28470\n29200\n730\nkWh\nSubsidi 1.35 sen/kWh diblayai Kerajaan
Persekutuan RM 10.61\nService Tax (ST) 6% bagi penggunaan Domestik
melebihi 600 kWh\nBayaran melalui cek sah setelah penjelasan cek
oleh bank\n*2200012345120001000123400000000031190\nAli bin Abu\
nE2SA - 201 Sek 3\nWangsa Maju\n53300\nWP Kuala Lumpur\nRM\n311.90\
nAras 17, Wisma TNB, No. 19. Jalan Timur, 46200 Petaling Jaya,
Selangor.\nTENAGA NASIONAL BERHAD (200006), Nombor Datar ST W10-
1608-31022372\n"
bounding_poly {
  ...

```

The output can be converted into a table after some cleaning:

	Blok Tarif (KWh)	Blok Prorata (kWh)	Kadar (RM)	Amaun (RM)
1	200	200	0.218	43.60
2	100	100	0.334	33.40
3	300	300	0.516	154.80
4	300	130	0.546	70.98

After scanning or downloading files which contain the information you are looking for, this OCR algorithm can be applied to extract information from those files and feed this information to your systems by APIs. OCR works well if the document you are extracting information comes from files with a standard structure like ICs or any official forms, and the document should preferably be typed. For handwritten documents, the results of Google Vision API's OCR are likely to vary depending on the quality of the handwriting.

The full code for both demonstrations above can be accessed via the QR code on the right, or via the link here .



<https://cloud.google.com/vision/docs/handwriting>

[https://github.com/atlas-github/abs\\_digital/blob/master/Extracting\\_text\\_from\\_images\\_PDFs.ipynb](https://github.com/atlas-github/abs_digital/blob/master/Extracting_text_from_images_PDFs.ipynb)



# FUTURE ENABLING SKILLS 2021



## Programme Structure



Technical training  
in Banking &  
Insurance



Digital  
Banking



Future Leaders &  
Managers Enrichment  
(FLAME) Leadership



Executive  
Presence



Communicate  
With Impact



Business  
Simulation  
Board Game



Sharing Session  
by Prominent  
Leaders

The Financial Sector Talent Enrichment Programme (FSTEP) is an intensive graduate training programme designed to equip fresh graduates and new entrants with the right skills and knowledge to jump-start their career in the financial services industry. It is an industry-driven initiative aimed at matching sponsoring banks and insurance companies with the right candidates that have been methodically chosen through a rigorous selection process to fit their needs. This initiative has proven to be very successful that the demand for FSTEP graduates has increasingly grown. For 2021, there will be two intakes, the first in April and the second in September.

To ease their transition into the financial services sector, FSTEP participants go through a 3-month period of comprehensive and interactive classroom training followed by a 9-month period of on-the-job training with their respective institutional sponsor. They also receive a monthly allowance of at least RM2,600 throughout the programme duration.

**BE AN FSTEP INSTITUTIONAL SPONSOR OR PARTICIPANT FOR 2021!**

**For more details:**



[fstep@asianbankingschool.com](mailto:fstep@asianbankingschool.com)



[www.fstep.org.my](http://www.fstep.org.my)



[fstep.official](https://www.facebook.com/fstep.official)

Managed by



In collaboration with



**BANK NEGARA MALAYSIA**  
CENTRAL BANK OF MALAYSIA

# CYBERSECURITY CHALLENGES FOR BANKS DURING THE PANDEMIC

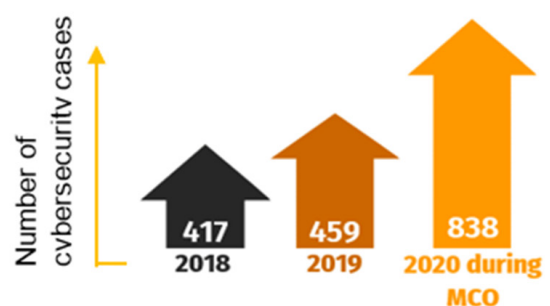
By Kayne Hoo

The COVID-19 pandemic has greatly impacted banking and financial services. On top of handling the direct economic impact of the pandemic, banks are now required to implement strategies to protect employees and customers from the virus that is spreading. Many of the banks are already encouraging employees to work remotely, and customers who are becoming increasingly cautious about spending time in public spaces will also need a way to conduct their banking without physical interaction.

This can be achieved through digitization of banking processes such as financial transactions processing, application process for new accounts, stock trading, and more. In fact, digital banking has become a sustainable business model that allows the financial industry to keep up in face of the pandemic, and many banks are now fast-forwarding their digital transformation plans to adjust to the new normal of 'Work-From-Home' and 'Contactless Banking'.

The accelerated adoption of digitization plans brings about new risks that will challenge the current defences in banks. Hackers are also actively looking for ways to take advantage of the pandemic as banks have to immediately implement plans for remote working and contactless banking in a short time and at a scale which they have not experienced before.

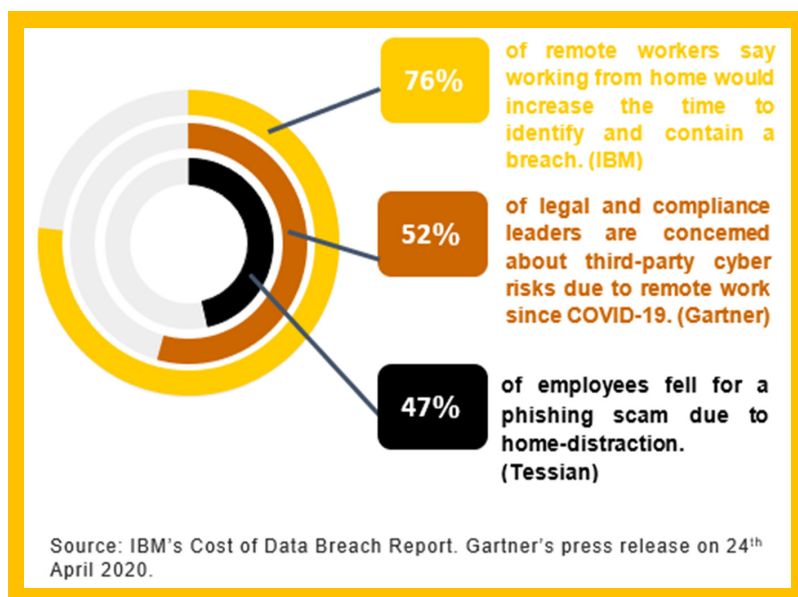
**Cybersecurity cases have spiked by a whopping **82.5 %** during the Movement Control Order (MCO) so far compared to the same time last year.**



Source: Malaysia Computer Emergency Response Team (MyCERT), CyberSecurity Malaysia, Reported Incidents based on General Incident Classification Statistics 2020

Figure 1: The number of cybersecurity cases reported in Malaysia from year 2018 to year 2020.





Many organizations are raising concerns on the cybersecurity risks arising from the implementation of remote working policies that targets a large number of employees. In addition to risks from an internal perspective, banks have to also manage possible risks associated with the digitization of banking processes for their customers at an external perspective. Hence, banks need to start identifying the risks involved when adjusting to the new normal in light of the pandemic.

Figure 2: Cyber risk statistics during the pandemic.

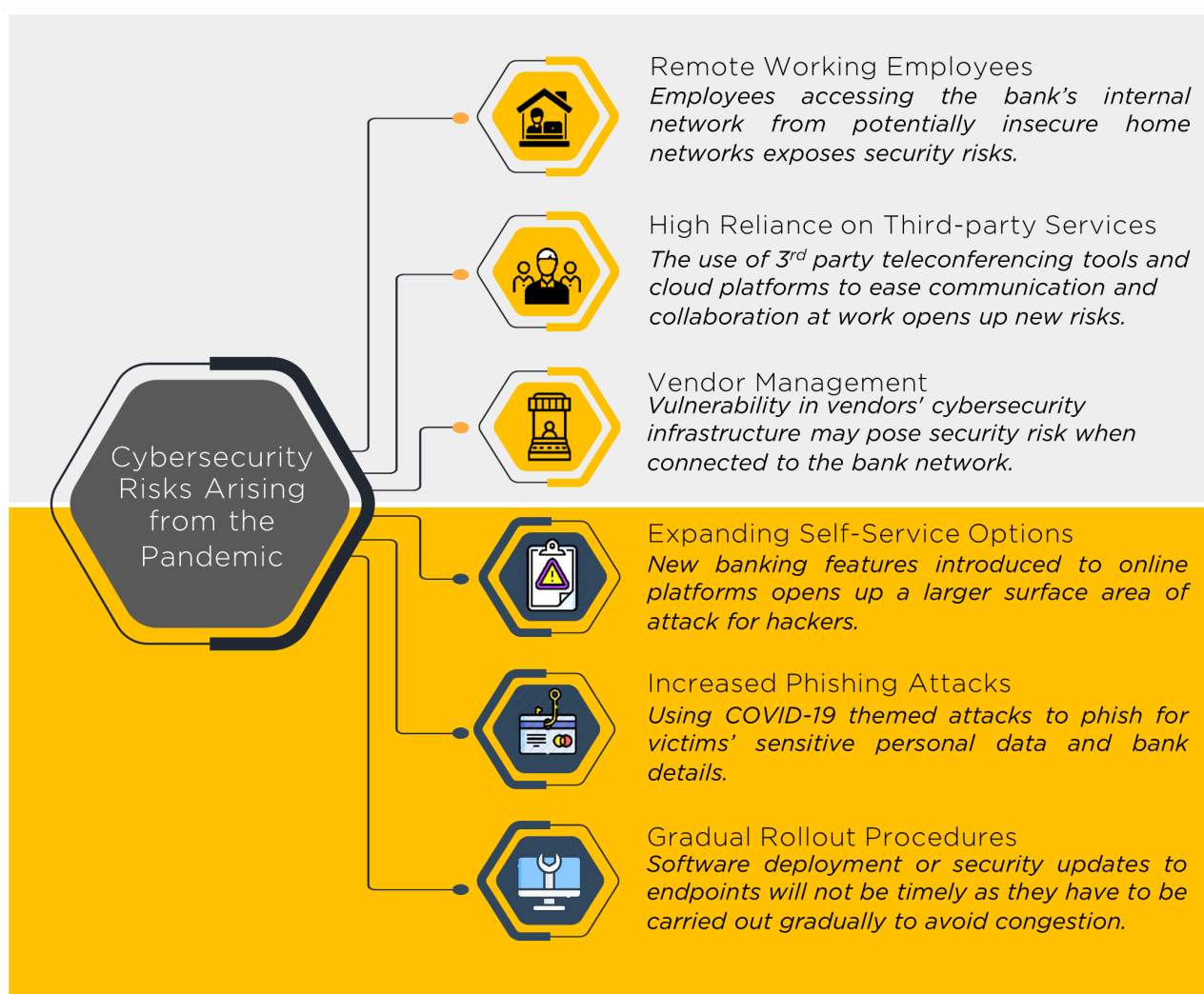


Figure 3: Common internal and external risks faced by banking institutions during the pandemic.

Banks should also consider implementing mitigation controls to manage the risks in conjunction with the adoption of remote working plans and contactless banking.



Figure 4: Suggested mitigation controls to reduce the impact of cyber threats

Banks will need to ensure that all required controls are in place when rolling out technologies for remote working. Employees that are working from home should be required to activate VPN and use MFA to access the internal network and any critical applications. In addition to patching the existing critical systems, technologies used for remote working should be constantly patched and updated to ensure that vulnerabilities are eradicated accordingly after discovery. The rollout for these updates can be planned to follow the criticality of the system and the severity addressed by the update to reduce network congestion.



As most banks employ a diverse number of vendors under their organizations, it is also crucial for them to ensure that vendors are practicing due diligence in ensuring adherence to good security practices. On top of stronger technology controls, banks should find ways to increase security awareness in bank employees and customers, particularly on social engineering due to the surge of COVID-19 campaigns during the pandemic.

The pandemic outbreak presents unique challenges and new means of effect given its global repercussions and potential duration. It has brought new challenges for digital banking services, and also an opportunity for banks to gain a deeper understanding of their own cybersecurity environment.

## References

1. Malaysia Computer Emergency Response Team (MyCERT), CyberSecurity Malaysia, Reported Incidents based on General Incident Classification Statistics 2020. Retrieved from <https://www.mycert.org.my/portal/statistics-content?menu=b75e037d-6ee3-4d11-8169-66677d694932&id=b9018870-c2a0-4b64-912d-39f65600abb8>
2. IBM Security, Cost of a Data Breach Report 2020. Retrieved from <https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/>
3. Cybersecurity and the COVID-19 pandemic. Retrieved from <https://www.lexology.com/library/detail.aspx?g=8b5940c3-55b2-4c48-982c-546e956906cd>





**For training enquiries, please contact:**

Asian Banking School (201201039737 (1024215-T))  
Level 3, Bangunan AICB  
10, Jalan Dato Onn  
50480 Kuala Lumpur, Malaysia

**Tel :** +603-2742 7822/21 **E-mail :** [digitalbanking@asianbankingschool.com](mailto:digitalbanking@asianbankingschool.com)

 Asian Banking School

 Asian Banking School (ABS)

[www.asianbankingschool.com](http://www.asianbankingschool.com)