IBM Research AI

CHARTER: heatmap-based multi-type chart data extraction

Joseph Shtok*, Sivan Harary*, Ophir Azulai, Adi Raz Goldfarb, Assaf Arbelle, Leonid Karlinsky*

IBM Research AI, Haifa Research Lab

The Second Document Intelligence Workshop at KDD 2021

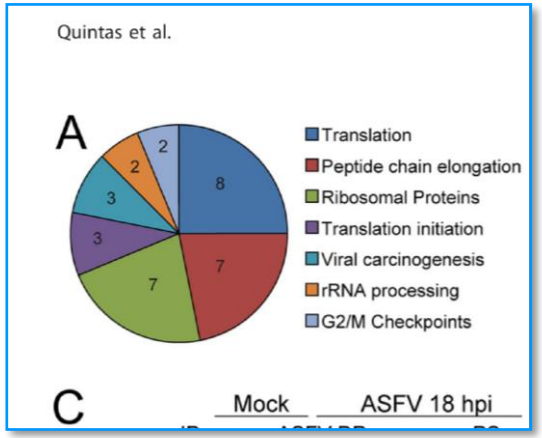
*equal contribution

Introduction

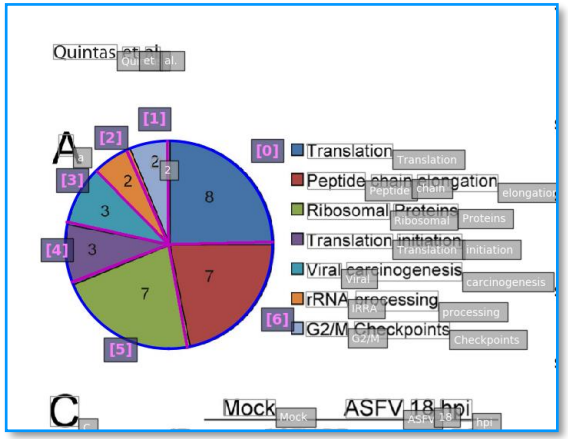
What task are we solving?

Given a document page…

Find a chart image in it…



Infer underlying numerical data (pie segment sizes, bar heights, graphs data)



JSON output


```
"metadata": {
        "ui_table_boxes": {
        "bounding_box": "0.0 0.0 0.3489 0.2970",
        "chart_type": "pie",
        "caption": "FIG +> ASFV-DP interacts with the ce
"ui_table": {
"headers": {
                "0": "index",
                "1": "angle",
                "2": "value",
                "3": "text_internal",
                "4": "text_external",
                "5": "text_legend" }
        "0":{
                "0": 0,
                "1": 89.31834030953903,
                "2": 0.24810650085983064,
                "3": "",
                "4":"Translation Peptide chain elongatio
                "5": "Translation"  },
        "1":{
                "0": 1,
                "1":23.335593776459093,
                "2":  0.06482109382349748,
                "3": "2",
                "4": "Quintas et al.",
                "5": "G2/MCheckpoints"  },
        "2": {
                "0": 2,
                "1": 20.46750022794376,
                "2": 0.05685416729984378,
                "3": null,
                "4": null,
                "5": "IRRA processing" },
        ...
{{
"page_num": "6",
"chart_id": "1",
"file_name": "5709586_cc-by.pdf",
...
```

Introduction

Why do we solve it?

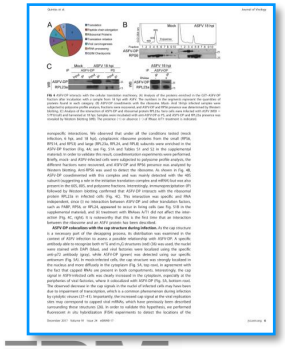
1. To extend document search.

Peptide chain

Text query

"Peptide chain elongation" found in chart (Fig. 4, page 6), value = 23%

2. To extend question answering in documents

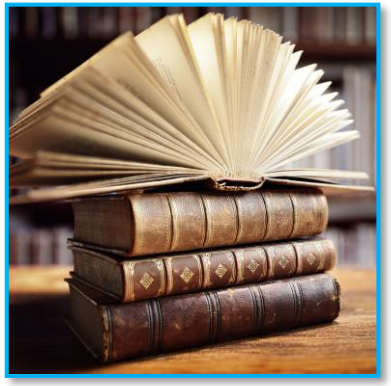


How often does Peptide chain elongation occur?

Question

"Peptide chain elongation occupies 23% of ASFV-DP interactions

3. Towards unlocking the knowledge in documents

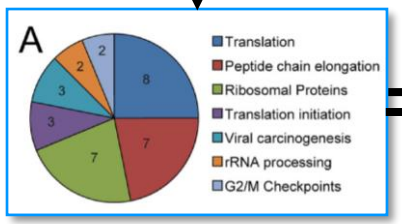


Our CHARTER processing pipeline

Document page image

Stage 1: detector of chart regions and elements

Detected chart



Stage 2: Detector of graphical elements

Boxes for chart regions, bars, pie sectors, axes.
Heatmaps for pie circles, centers, sectors, bar lines, bar corners, axis ticks, line graphs.

Chart, title, caption, legend regions

Texts in chart region and elements

Stage 3: OCR engine

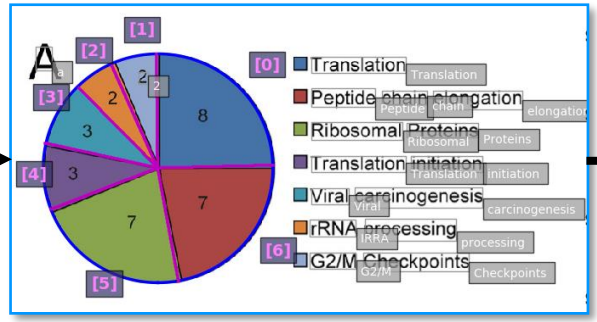
Stage 4: Analysis module

Chart type specific Computer Vision algoritms and ad hoc rules

Extracted chart data:

Pie charts: sector angles and labels
Bar charts: x-labels and y-values for the bars
Line charts: table of (x,y) values for each graph
General: Title, caption, axes names

visualization

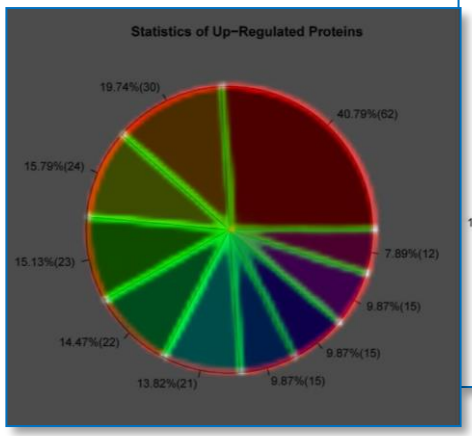


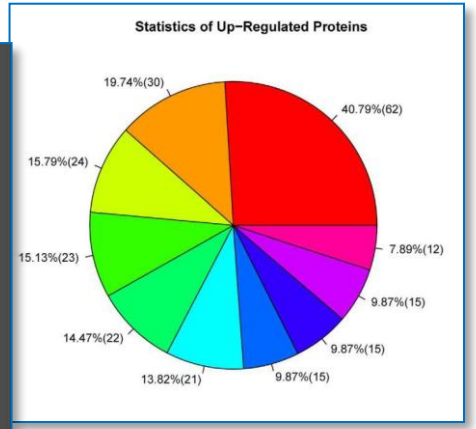
JSON output


```
"metadata": {
        "ui_table_boxes": {
        "bounding_box": "0.0 0.0 0.3489 0.2970",
        "chart_type": "pie",
        "caption": "FIG +> ASFV-DP interacts with the ce
"ui_table": {
"headers": {
                "0": "index",
                "1": "angle",
                "2": "value",
                "3": "text_internal",
                "4": "text_external",
                "5": "text_legend" }
        "0":{
                "0": 0,
                "1": 89.31834030953903,
                "2": 0.24810650085983064,
                "3": "",
                "4":"Translation Peptide chain elongatio
                "5": "Translation" },
        "1":{
                "0": 1,
                "1":23.335593776459093,
                "2":  0.06482109382349748,
                "3": "2",
                "4": "Quintas et al.",
                "5": "G2/MCheckpoints" },
        "2": {
                "0": 2,
                "1": 20.46750022794376,
                "2": 0.05685416729984378,
                "3": null,
                "4": null,
                "5": "IRRA processing" },
        ...
{{
"page_num": "6",
"chart_id": "1",
"file_name": "5709586_cc-by.pdf",
...
```

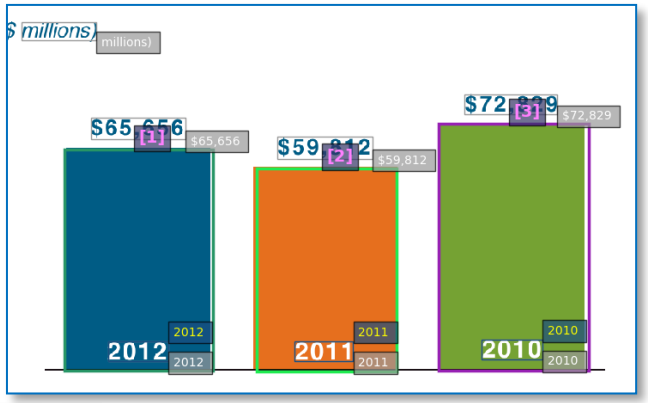
Detection of graphical elements in charts: boxes and heatmaps

Pie charts: heatmaps of the circumference, center, radial lines, intersections.

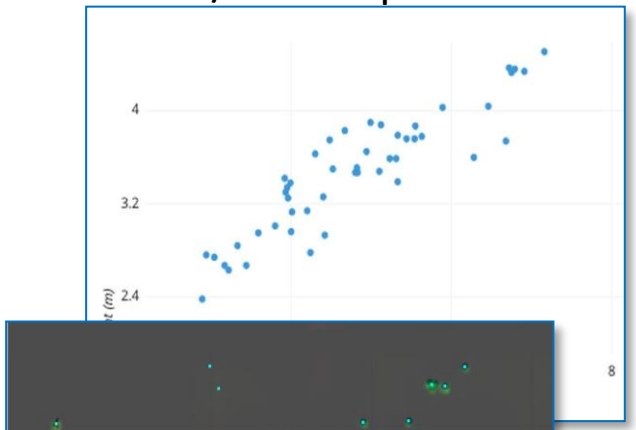


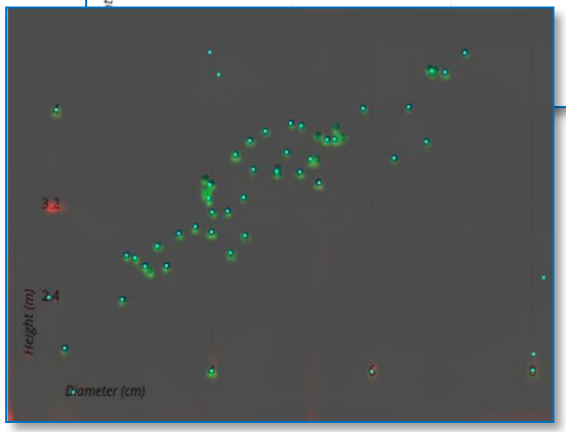


Bar charts: boxes of bars

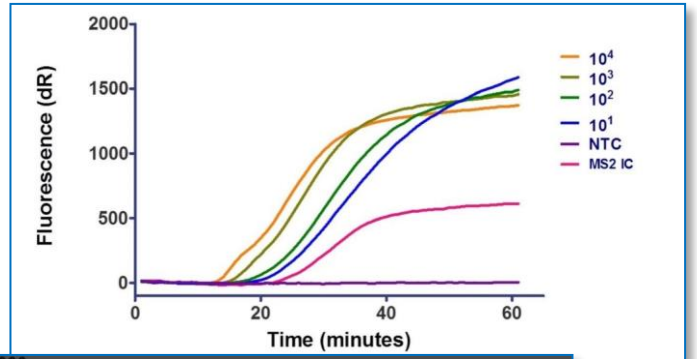


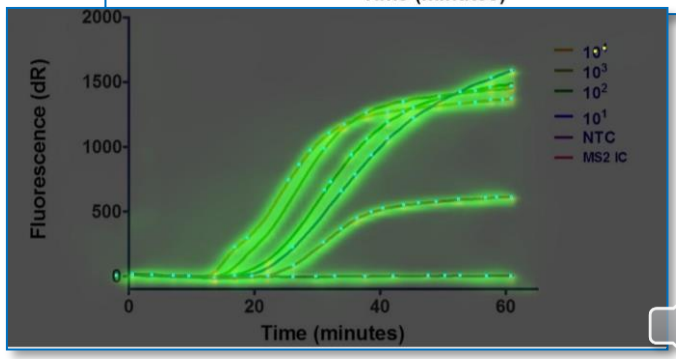
Scatter plots: boxes/heatmaps of markers





Line charts: heatmaps of lines

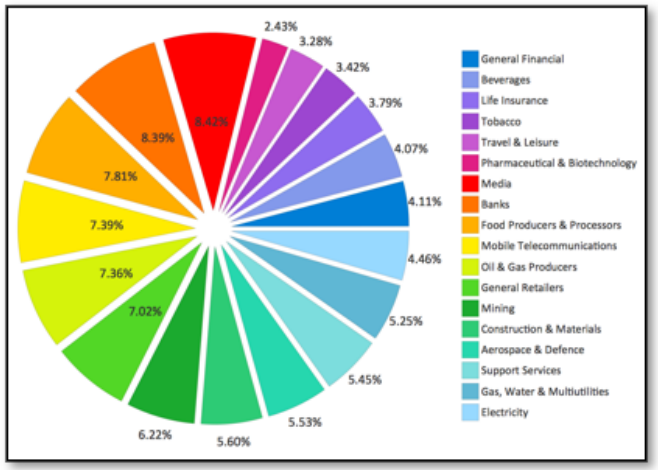




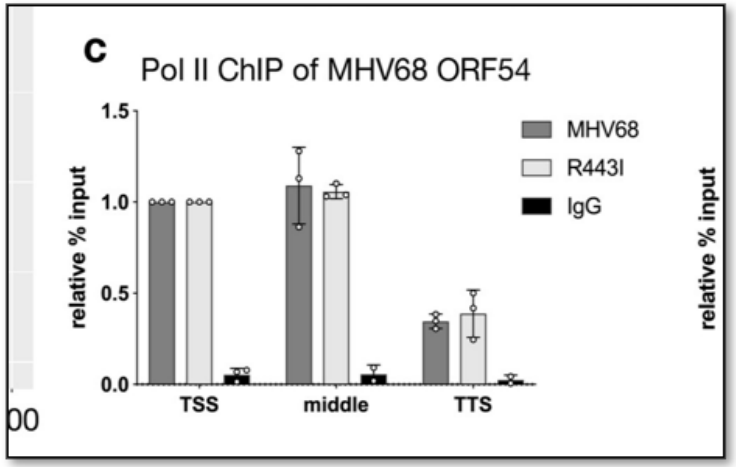
CHARTER: heatmap-based multi-type chart data extraction

System output

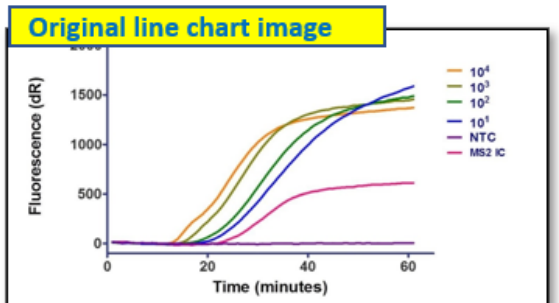
Original pie chart image

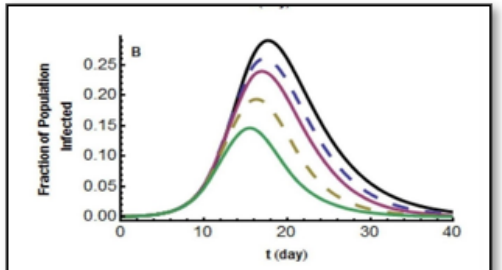


Original bar image

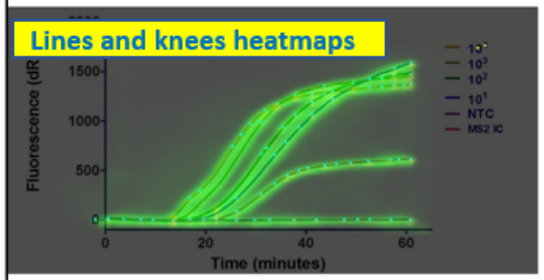


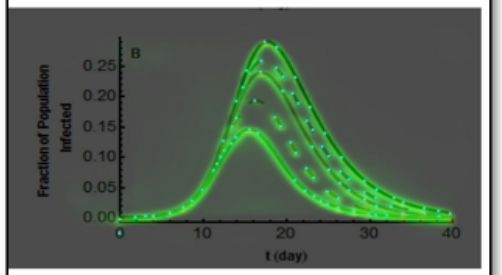
Original line chart image



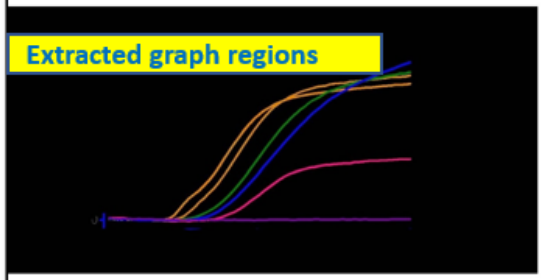


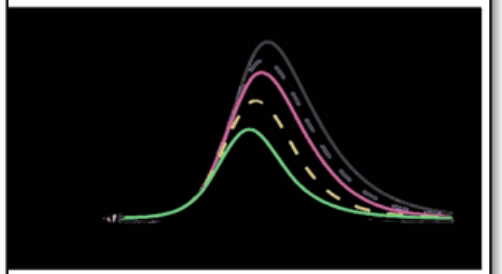
Lines and knees heatmaps



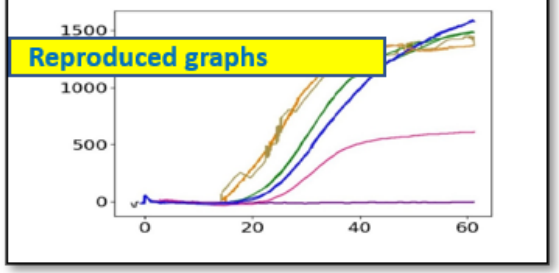


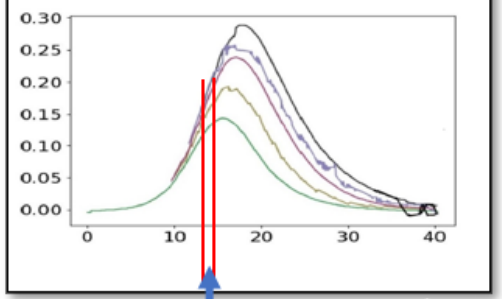
Extracted graph regions





Reproduced graphs





Extracted data: pie sectors, legend

Legend	percent %
Beverages	4.028
Life insurance	3.500
Tobacco	3.750
Travel & Leisure	3.278
Pharmaceutical & Biotechnology	2.944
Media	7.806
Banks	8.361
Food Producers & Processors	7.222
Mobile Telecommunications	7.861
Oil & Gas Producers	7.167
General Retailers	7.000
Mining	6.000
Construction & Materials	5.806
Support Services	5.333
---	5.806
Gas, Water & Multiutilities	5.500
Electricity	4.111
---	4.528

Extracted data: Bars, title, legend, y-axis name

Title: Pol II Chip of MHV68 ORF54		
Label	**legend**	**Relative a input**
TSS	MHV68	1.01
TSS	R443]	1
TSS	Igg	0.05
middle	MHV68	1.1
middle	R443]	1.03
middle	Igg	0.05
TTS	MHV68	0.33
TTS	R443]	0.38

X-value	Graph 1	Graph 2	Graph 3	Graph 4	Graph 5
...	...	...	...	...	...
13.4	0.1257	0.1532	0.1749	0.1901	0.1905
13.5	0.1266	0.1548	0.1761	0.1943	0.1913
13.6	0.1275	0.1561	0.1774	0.1972	0.1942
13.7	0.1283	0.1565	0.1795	0.2070	0.2075

Extracted data: sequences of values per graph

Performance

Our tests are conducted on real-world, manually annotated data of document pages, and bar/pie charts and 30 pie charts from the ICPR2020 dataset

Table 1: Detection of charts and related elements in documents

Category	Bar	Pie	Line	Scatter	
AP@0.5	98.0%	97.8%	84.4%	91.35%	
Category	Legend	Caption	Title	X label	Y label
AP@0.5	77.8%	91.4%	71.8%	84.7%	94.9%

Stage 1 detector performance, measured as Average Precision (AP) with IoU= 0.5, averaged over 5 random train/test splits.

Table 2: Detection of graphical elements in chart image

Category	Horizontal bar	Vertical bar	Pie sector
[12]	–	80.2%	–
Ours	76.5%	90.5%	90.9%

Performance (in Average Precision (AP) with IoU= 0.5 of the graphical elements detector (box categories) on the real data.

Table 3: Accuracy of bar values in extraction of tabular data

Method	A.L	A.L	E.L	E.L	E.L	E.L
ϵ	0.02	0.05	0.01	0.05	0.01	0.025
[17]	**67.0%**	71.0%	–	–	–	–
[12]	–	–	28.4%	32.8 %	34.3 %	38.8 %
Ours	60.0%	**74.2%**	**31.6%**	**55.8%**	**58.3%**	**60.3%**

E.L - exact label prediction, A.L - any labels.

Table 4: Accuracy of pie segment angles in extraction of tabular data from real-world pie charts with exact labels

Method	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.25$
[12]	28.9%	57.8%	60.0%	**64.6%**
Ours ($R_{Lev} = 1.0$)	**44.9%**	**60.3%**	**61.0%**	61.0%

[12] Xiaoyi Liu, Diego Klabjan, and Patrick N. Bless. 2019. Data extraction from charts via single deep neural network. arXiv (2019).

[17] Fangfang Zhou, Yong Zhao, Wenjiang Chen, Yijing Tan, Yaqi Xu, Yi Chen, Chao Liu, and Ying Zhao. 2021. Reverse-engineering bar charts using neural networks. Journal of Visualization 24, 2 (2021), 419–435

Thank you

Come visit us at

https://www.research.ibm.com/haifa/dept/imt/cvar/index.html