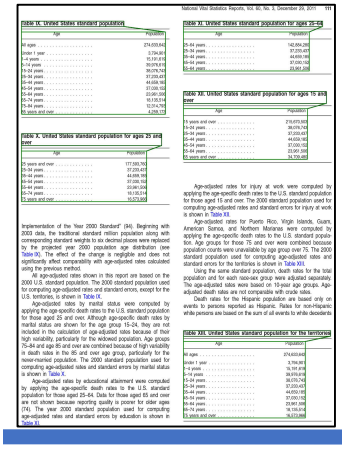
An Integrated Approach of Deep Learning and Symbolic Analysis for Digital PDF Table Extraction

Mengshi Zhang*
University of Texas at Austin
Austin, TX, USA
mengshi.zhang@utexas.edu

Daniel Perelman, Vu Le, Sumit Gulwani
Microsoft
Redmond, WA, USA
{**danpere**, levu, sumitg}@microsoft.com

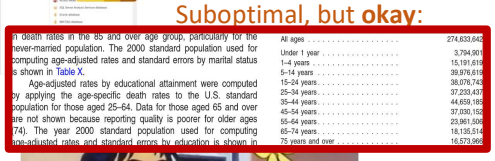
*Mengshi Zhang performed this work as part of his internship with the PROSE team at Microsoft. Now, he is a research scientist at Facebook.

Problem: PDF table extraction – Targeted at Microsoft Power BI

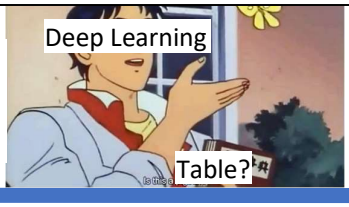


Extracting noise worth it when data otherwise locked in PDF…

Suboptimal, but **okay**:



They make different mistakes!



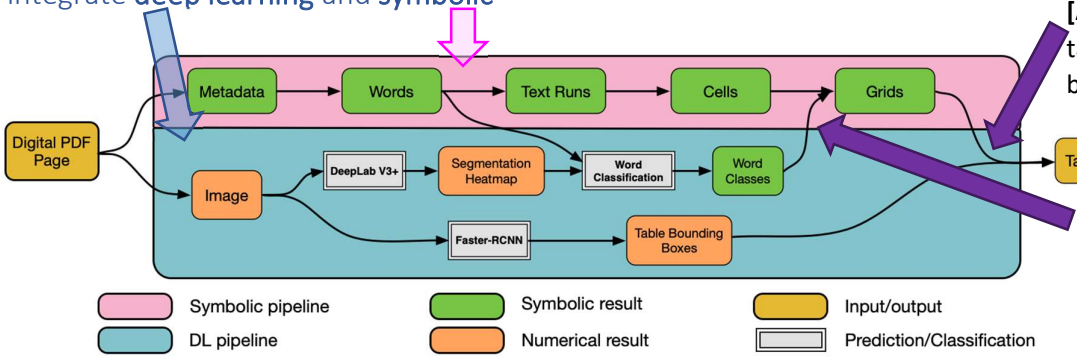
… but omitting data silently is unacceptable.

Bad:

#	Name	Amount	Description	Date
5	Contoso Ltd.	$5.95	Very little	2020-12-10
6	AdventureWorks Cycles	$42.89	Some stuff	2020-12-13
7	City Power & Light	$21.43	Not much	2020-12-23
8	Important Customer	$869,494.01	All the things	2021-01-02

Goal: Combine the two to avoid both categories of mistakes.

Algorithm: integrate **deep learning** and **symbolic**

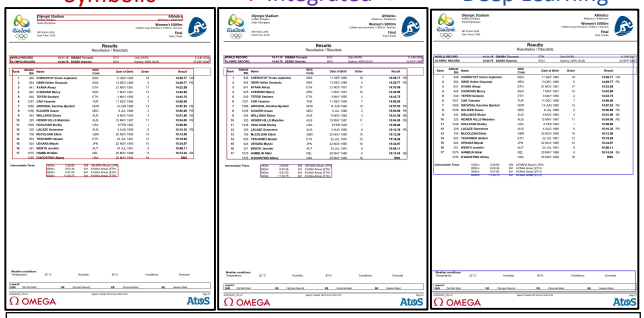


[A] Combine tables learned by both methods

[B] Deep learning identifies text unlikely to be part of a table.

[A] Filter tables from **symbolic** using **deep learning**

Symbolic → Integrated ← Deep Learning

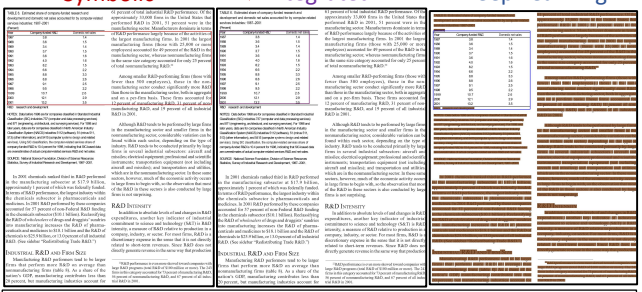


Observation: **deep learning** identifies table-like areas, but not precise bounds; **symbolic** identifies precise bounds, but identifies aligned text even when it does not look like a table.

Integrated: Only keep **symbolic** tables that overlap a **deep learning** table.

[B] **Deep learning** informs **symbolic** of table-like areas

Symbolic → Integrated ← Deep Learning



Observation: **symbolic** gets confused by aligned text near a table, but **deep learning** can tell such text does not appear to be part of any table.

Integrated: Forbid **symbolic** from using page areas identified by **deep learning** as unlikely to be a table.

Evaluation

Note we do **not** use "intersection-over-union" in order to avoid giving partial credit to tables missing important information, but instead a table is considered correct only on an exact match of all text in the expected table.

Algorithm	Precision	Recall	F_1
Symbolic	0.315	**0.418**	0.359
DeepDeSRT (state-of-the-art)	0.178	0.120	0.144
Integrated (symbolic+our DL)	**0.459**	0.390	**0.422**