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.

PDF table extraction

Table IX. United States standard population

Age	Population
All ages	274,633,642
Under 1 year	3,794,901
1–4 years	15,191,619
5–14 years	39,976,619
15–24 years	38,076,743
25–34 years	37,233,437
35–44 years	44,659,185
45–54 years	37,030,152
55–64 years	23,961,506
65–74 years	18,135,514
75–84 years	12,314,793
85 years and over	4,259,173

Table X. United States standard population for ages 25 and over

Age	Population
25 years and over	177,593,760
25–34 years	37,233,437
35–44 years	44,659,185
45–54 years	37,030,152
55–64 years	23,961,506
65–74 years	18,135,514
75 years and over	16,573,966

Implementation of the Year 2000 Standard" (94). Beginning with 2003 data, the traditional standard million population along with corresponding standard weights to six decimal places were replaced by the projected year 2000 population age distribution (see Table IX). The effect of the change is negligible and does not significantly affect comparability with age-adjusted rates calculated using the previous method.

All age-adjusted rates shown in this report are based on the 2000 U.S. standard population. The 2000 standard population used for computing age-adjusted rates and standard errors, except for the U.S. territories, is shown in Table IX.

Age-adjusted rates by marital status were computed by applying the age-specific death rates to the U.S. standard population for those aged 25 and over. Although age-specific death rates by marital status are shown for the age group 15–24, they are not included in the calculation of age-adjusted rates because of their high variability, particularly for the widowed population. Age groups 75–84 and age 85 and over are combined because of high variability in death rates in the 85 and over age group, particularly for the never-married population. The 2000 standard population used for computing age-adjusted rates and standard errors by marital status is shown in Table X.

Age-adjusted rates by educational attainment were computed by applying the age-specific death rates to the U.S. standard population for those aged 25–64. Data for those aged 65 and over are not shown because reporting quality is poorer for older ages (74). The year 2000 standard population used for computing age-adjusted rates and standard errors by education is shown in Table XI.

Table XI. United States standard population for ages 25–64

Age	Population
25–64 years	142,884,280
25–34 years	37,233,437
35–44 years	44,659,185
45–54 years	37,030,152
55–64 years	23,961,506

Table XII. United States standard population for ages 15 and over

Age	Population
15 years and over	215,670,503
15–24 years	38,076,743
25–34 years	37,233,437
35–44 years	44,659,185
45–54 years	37,030,152
55–64 years	23,961,506
65 years and over	34,709,480

Age-adjusted rates for injury at work were computed by applying the age-specific death rates to the U.S. standard population for those aged 15 and over. The 2000 standard population used for computing age-adjusted rates and standard errors for injury at work is shown in Table XII.

Age-adjusted rates for Puerto Rico, Virgin Islands, Guam, American Samoa, and Northern Marianas were computed by applying the age-specific death rates to the U.S. standard population. Age groups for those 75 and over were combined because population counts were unavailable by age group over 75. The 2000 standard population used for computing age-adjusted rates and standard errors for the territories is shown in Table XIII.

Using the same standard population, death rates for the total population and for each race-sex group were adjusted separately. The age-adjusted rates were based on 10-year age groups. Age-adjusted death rates are not comparable with crude rates.

Death rates for the Hispanic population are based only on events to persons reported as Hispanic. Rates for non-Hispanic white persons are based on the sum of all events to white decedents

Table XIII. United States standard population for the territories

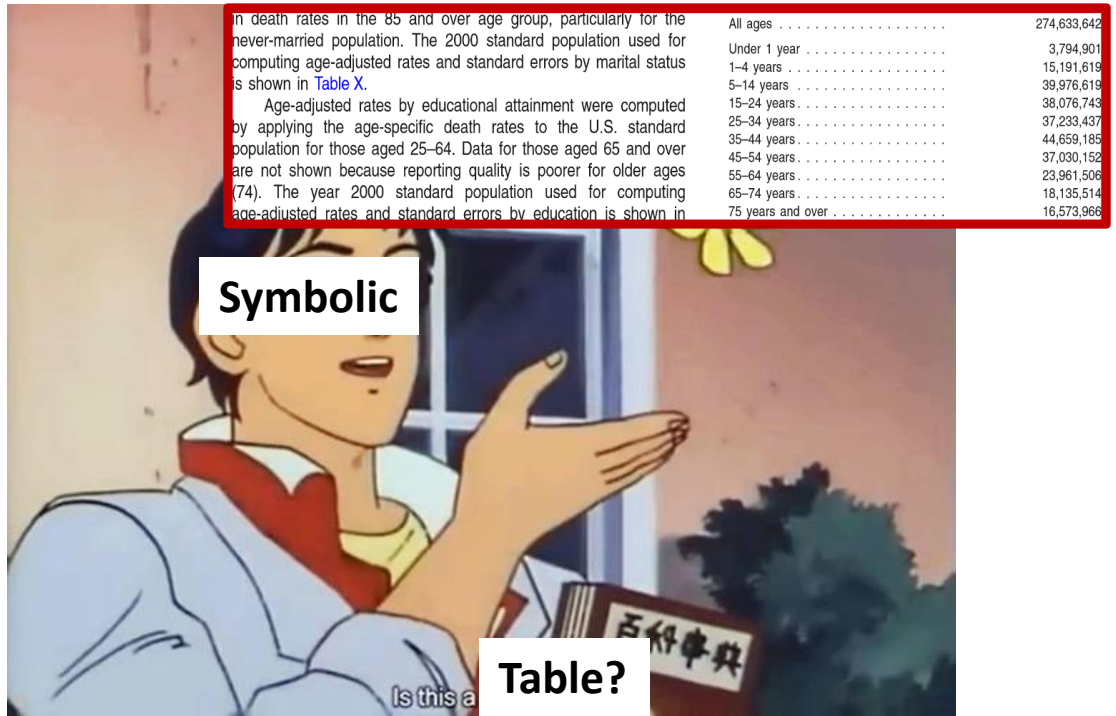
Age	Population
All ages	274,633,642
Under 1 year	3,794,901
1–4 years	15,191,619
5–14 years	39,976,619
15–24 years	38,076,743
25–34 years	37,233,437
35–44 years	44,659,185
45–54 years	37,030,152
55–64 years	23,961,506
65–74 years	18,135,514
75 years and over	16,573,966

Comparison of separate approaches

Symbolic (rules-based)

+ Precise bounds

- False positives on aligned text

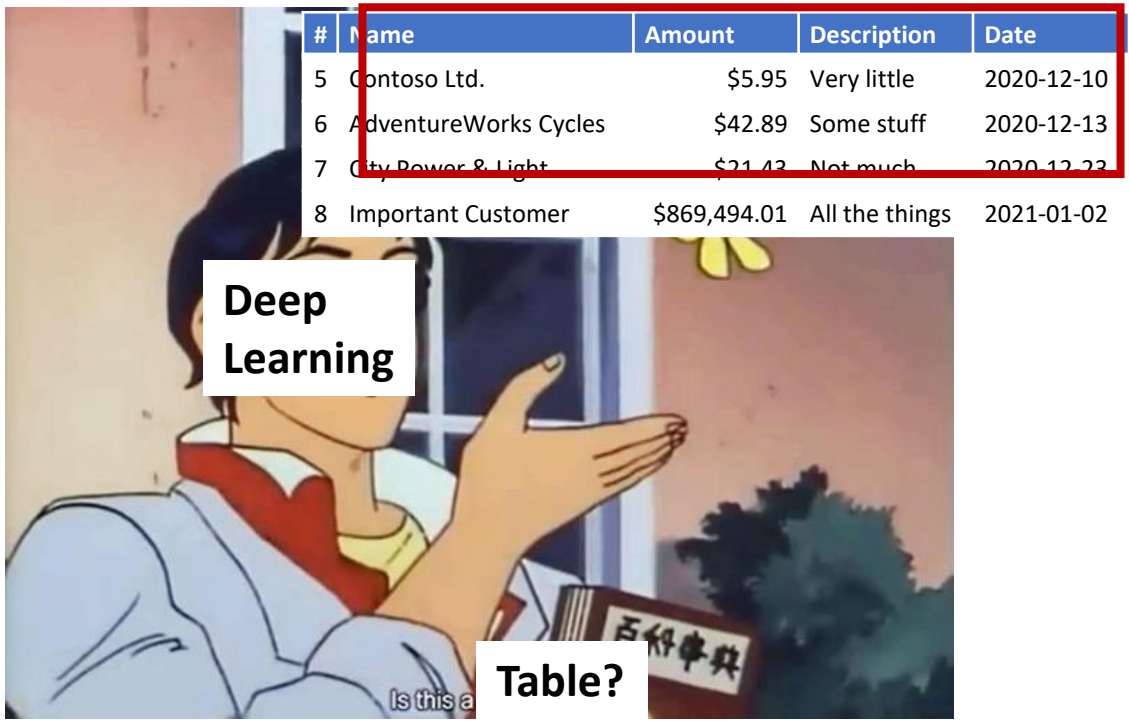


Deep Learning

+ Detects irregular tables

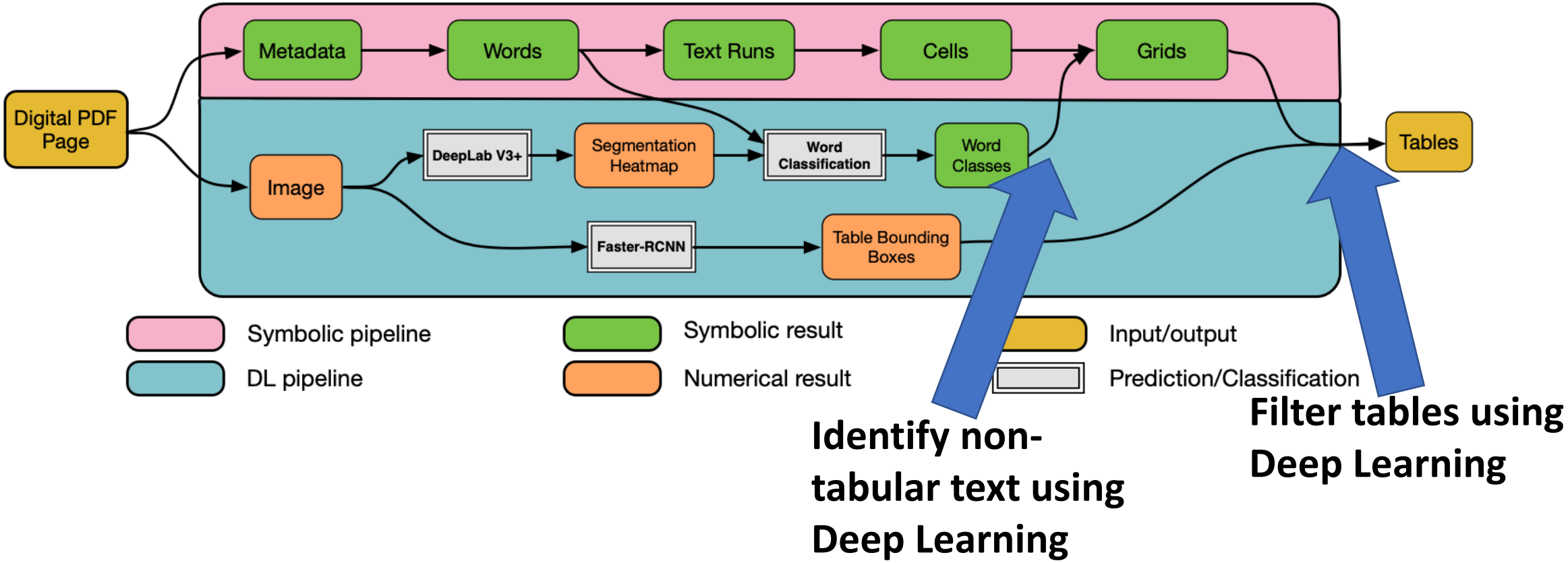
- Imprecise bounds

(mock example)



#	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

Algorithm workflow



Olympic Stadium
Estádio Olímpico
Stade Olympique

Athletics
Atletismo / Athlétisme

Women's 5000m
5.000m rasos feminino / 5 000 m - femmes

Final
final / finale

FRI 19 AUG 2016
Start Time: 21:40

Results
Resultados / Résultats

WORLD RECORD	14:11.15	DIBABA Tirunesh	ETH	Oslo (NOR)	6 JUN 2008
OLYMPIC RECORD	14:40.79	SZABO Gabriela	ROU	Sydney, NSW (AUS)	25 SEP 2000

Rank	Athlete Bib	Name	NOC Code	Date of Birth	Order	Result	
1	946	CHERUIYOT Vivian Jepkemoi	KEN	11 SEP 1983	18	14:26.17	OR
2	954	OBIRI Hellen Onsando	KEN	13 DEC 1989	5	14:29.77	PB
3	641	AYANA Almaz	ETH	21 NOV 1991	17	14:33.59	
4	945	CHERONO Mercy	KEN	7 MAY 1991	15	14:42.89	
5	649	TEFERI Senbere	ETH	3 MAY 1995	4	14:43.75	
6	1257	CAN Yasemin	TUR	11 DEC 1996	7	14:56.96	
7	1068	GROVDAL Karoline Bjerkeli	NOR	14 JUN 1990	13	14:57.53	PB
8	1036	KUIJKEN Susan	NED	8 JUL 1986	2	15:00.69	PB
9	344	WELLINGS Eloise	AUS	9 NOV 1982	3	15:01.59	SB
10	325	HEINER HILLS Madeline	AUS	15 MAY 1987	11	15:04.05	PB
11	1348	HOULIHAN Shelby	USA	8 FEB 1993	1	15:08.89	
12	329	LACAZE Genevieve	AUS	4 AUG 1989	8	15:10.35	PB
13	709	McCOLGAN Eilish	GBR	25 NOV 1990	14	15:12.09	
14	653	YESHANEH Ababel	ETH	22 JUL 1991	12	15:18.26	
15	924	UEHARA Miyuki	JPN	22 NOV 1995	10	15:34.97	
16	351	WENTH Jennifer	AUT	24 JUL 1991	9	15:56.11	
17	1075	HAMBLIN Nikki	NZL	20 MAY 1988	6	16:14.24	SB
	1335	D'AGOSTINO Abbey	USA	25 MAY 1992	16	DNS	

Intermediate Times

1000m	2:59.86	924	UEHARA Miyuki (JPN)
2000m	6:00.36	641	AYANA Almaz (ETH)
3000m	8:47.80	641	AYANA Almaz (ETH)
4000m	11:39.75	641	AYANA Almaz (ETH)

Weather conditions
Temperature: 22 °C Humidity: 83 % Conditions: Overcast

Legend:
DNS Did Not Start OR Olympic Record PB Personal Best SB Season Best

ATW050101_73G 1.0 Report Created FRI 19 AUG 2016 22:01

OMEGA

Symbolic

Integrated

Deep Learning

Rio_2016_Athletics_Results_Book_V1.0.pdf (page 310) from https://library.olympic.org/Default/doc/SYRACUSE/165312/results-book-rio-2016-organising-committee-for-the-olympic-and-paralympic-games-in-rio-in-2016?_lg=en-GB

Symbolic

Integrated

Deep Learning

TABLE 6. Estimated share of company-funded research and development and domestic net sales accounted for by computer-related services industries: 1987–2001
(Percent)

Year	Company-funded R&D	Domestic net sales
1987	3.8	1.4
1988	3.6	1.5
1989	3.4	1.4
1990	3.7	1.5
1991	3.6	1.6
1992	4.0	1.6
1993	8.2	1.5
1994	6.6	2.2
1995	8.8	3.3
1996	8.8	2.6
1997	9.1	2.5
1998	9.5	2.2
1999	10.7	2.6
2000	12.1	2.9
2001	13.2	3.5

R&D research and development

NOTES: Data before 1998 are for companies classified in Standard Industrial Classification (SIC) industries 737 (computer and data processing services) and 871 (engineering, architectural, and surveying services). For 1998 and later years, data are for companies classified in North American Industry Classification System (NAICS) industries 5112 (software), 51 (minus 511, 513) (other information), and 5415 (computer systems design and related services). Using SIC classification, the computer-related services share of company-funded R&D is 10.4 percent for 1998, indicating that SIC-based data are overestimates of actual computer-related services R&D and net sales.

SOURCE: National Science Foundation, Division of Science Resources Statistics, Survey of Industrial Research and Development, 1987–2001.

In 2001 chemicals ranked third in R&D performed in the manufacturing subsector at $17.9 billion, approximately 1 percent of which was federally funded. In terms of R&D performance, the largest industry within the chemicals subsector is pharmaceuticals and medicines. In 2001 R&D performed by these companies accounted for 57 percent of non-Federal R&D funding in the chemicals subsector ($10.1 billion). Reclassifying the R&D of *wholesalers* of drugs and druggists' sundries into manufacturing increases the R&D of pharmaceuticals and medicines to $18.1 billion and the R&D of chemicals to $25.9 billion, or 13.0 percent of all industrial R&D. (See sidebar "Redistributing Trade R&D.")

Industrial R&D and Firm Size

Manufacturing R&D performers tend to be larger firms that perform more R&D on average than nonmanufacturing firms (table 8). As a share of the nation's GDP, manufacturing contributes less than 20 percent, but manufacturing industries account for

61 percent of total industrial R&D perfor
approximately 33,000 firms in the Unit
performed R&D in 2001, 51 percent
manufacturing sector. Manufacturers don
of R&D performance largely because of t
the largest manufacturing firms. In 20
manufacturing firms (those with 25,
employees) accounted for 49 percent of t
manufacturing sector, whereas nonmanuf
in the same size category accounted for o
of total nonmanufacturing R&D.[22]

Among smaller R&D-performing fir
fewer than 500 employees), those
manufacturing sector conduct significant
than those in the manufacturing sector, bo
and on a per-firm basis. These firms
12 percent of manufacturing R&D, 31 p
manufacturing R&D, and 19 percent o
R&D in 2001.

Although R&D tends to be performed
in the manufacturing sector and smalle
nonmanufacturing sector, considerable va
found within each sector, depending o
industry. R&D tends to be conducted pri
firms in several industrial subsectors
missiles; electrical equipment; professiona
instruments; transportation equipment
aircraft and missiles); and transportatio
which are in the nonmanufacturing sector
sectors, however, much of the economic
in large firms to begin with, so the observa
of the R&D in these sectors is also cond
firms is not surprising.

R&D Intensity

In addition to absolute levels of and ch
expenditures, another key indicator
commitment to science and technology (
intensity, a measure of R&D relative to p
company, industry, or sector. For most fi
discretionary expense in the sense that it
related to short-term revenues. Since R
directly generate revenue in the same way t

[22]R&D performance is even more skewed towa
large R&D programs (total R&D of $100 million
firms in this category accounted for 73 percent of ma
56 percent of nonmanufacturing R&D, and 67 pe
trial R&D in 2001.

TABLE 6. Estimated share of company-funded research and development and domestic net sales accounted for by computer-related services industries: 1987–2001
(Percent)

Year	Company-funded R&D	Domestic net sales
1987	3.8	1.4
1988	3.6	1.5
1989	3.4	1.4
1990	3.7	1.5
1991	3.6	1.6
1992	4.0	1.6
1993	8.2	1.5
1994	6.6	2.2
1995	8.8	3.3
1996	8.8	2.6
1997	9.1	2.5
1998	9.5	2.2
1999	10.7	2.6
2000	12.1	2.9
2001	13.2	3.5

R&D research and development

NOTES: Data before 1998 are for companies classified in Standard Industrial Classification (SIC) industries 737 (computer and data processing services) and 871 (engineering, architectural, and surveying services). For 1998 and later years, data are for companies classified in North American Industry Classification System (NAICS) industries 5112 (software), 51 (minus 511, 513) (other information), and 5415 (computer systems design and related services). Using SIC classification, the computer-related services share of company-funded R&D is 10.4 percent for 1998, indicating that SIC-based data are overestimates of actual computer-related services R&D and net sales.

SOURCE: National Science Foundation, Division of Science Resources Statistics, Survey of Industrial Research and Development, 1987–2001.

In 2001 chemicals ranked third in R&D performed in the manufacturing subsector at $17.9 billion, approximately 1 percent of which was federally funded. In terms of R&D performance, the largest industry within the chemicals subsector is pharmaceuticals and medicines. In 2001 R&D performed by these companies accounted for 57 percent of non-Federal R&D funding in the chemicals subsector ($10.1 billion). Reclassifying the R&D of *wholesalers* of drugs and druggists' sundries into manufacturing increases the R&D of pharmaceuticals and medicines to $18.1 billion and the R&D of chemicals to $25.9 billion, or 13.0 percent of all industrial R&D. (See sidebar "Redistributing Trade R&D.")

Industrial R&D and Firm Size

Manufacturing R&D performers tend to be larger firms that perform more R&D on average than nonmanufacturing firms (table 8). As a share of the nation's GDP, manufacturing contributes less than 20 percent, but manufacturing industries account for

61 percent of total industrial R&D performance. Of the approximately 33,000 firms in the United States that performed R&D in 2001, 51 percent were in the manufacturing sector. Manufacturers dominate in terms of R&D performance largely because of the activities of the largest manufacturing firms. In 2001 the largest manufacturing firms (those with 25,000 or more employees) accounted for 49 percent of the R&D in the manufacturing sector, whereas nonmanufacturing firms in the same size category accounted for only 25 percent of total nonmanufacturing R&D.[22]

Among smaller R&D-performing firms (those with fewer than 500 employees), those in the nonmanufacturing sector conduct significantly more R&D than those in the manufacturing sector, both in aggregate and on a per-firm basis. These firms accounted for 12 percent of manufacturing R&D, 31 percent of nonmanufacturing R&D, and 19 percent of all industrial R&D in 2001.

Although R&D tends to be performed by large firms in the manufacturing sector and smaller firms in the nonmanufacturing sector, considerable variation can be found within each sector, depending on the type of industry. R&D tends to be conducted primarily by large firms in several industrial subsectors: aircraft and missiles; electrical equipment; professional and scientific instruments; transportation equipment (not including aircraft and missiles); and transportation and utilities, which are in the nonmanufacturing sector. In these same sectors, however, much of the economic activity occurs in large firms to begin with, so the observation that most of the R&D in these sectors is also conducted by large firms is not surprising.

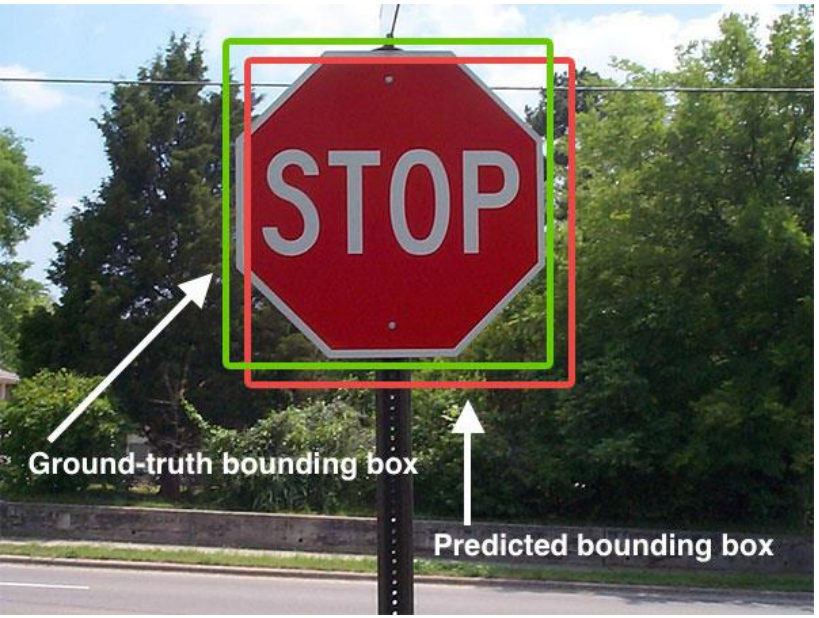
R&D Intensity

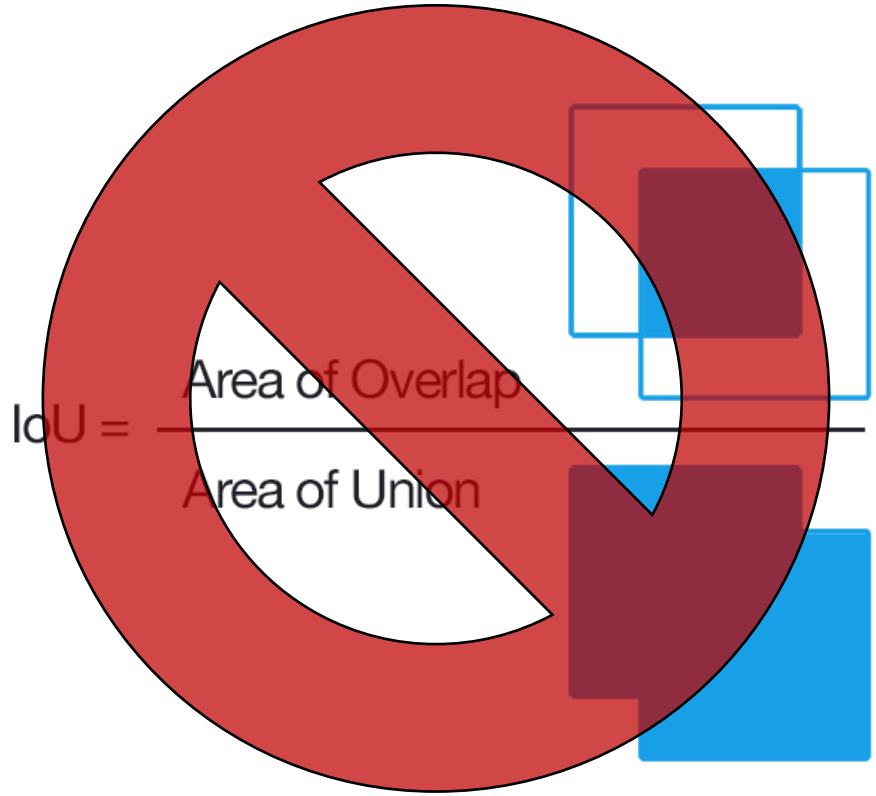
In addition to absolute levels of and changes in R&D expenditures, another key indicator of industrial commitment to science and technology (S&T) is R&D intensity, a measure of R&D relative to production in a company, industry, or sector. For most firms, R&D is a discretionary expense in the sense that it is not directly related to short-term revenues. Since R&D does not directly generate revenue in the same way that production

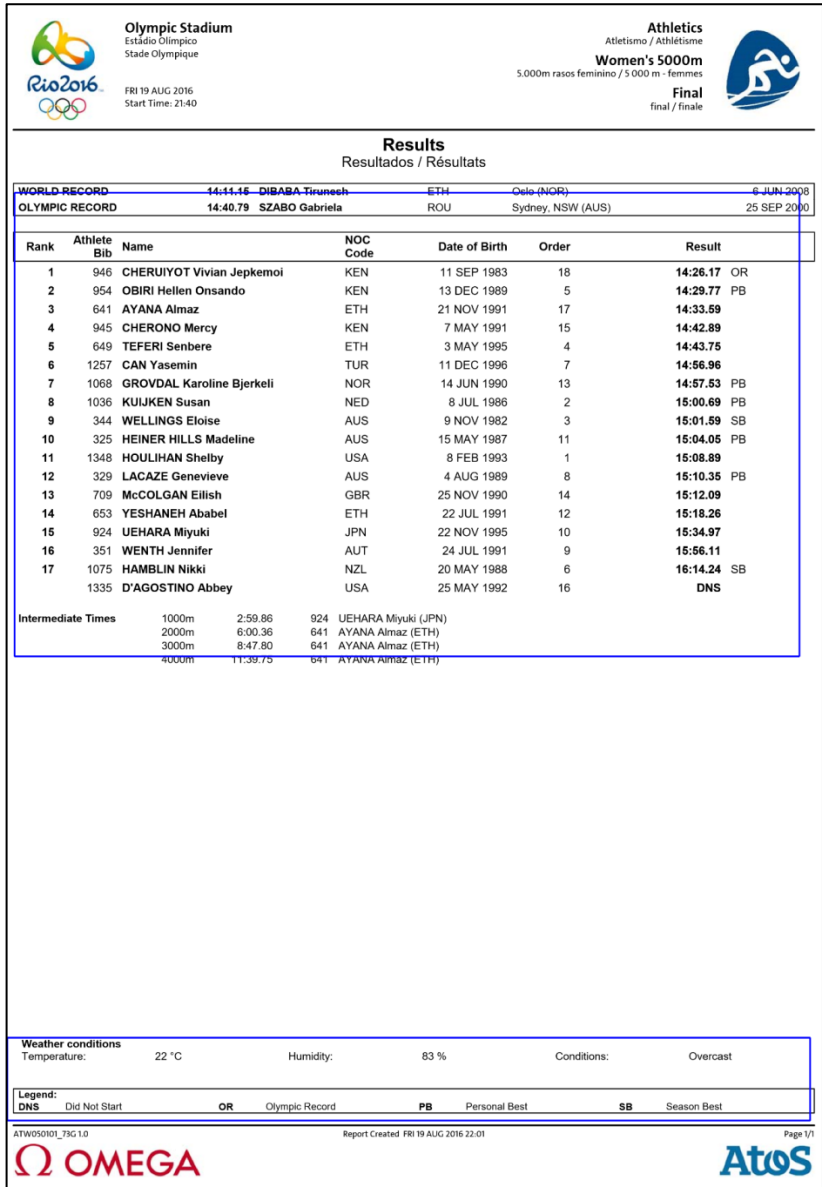
[22]R&D performance is even more skewed toward companies with large R&D programs (total R&D of $100 million or more). The 243 firms in this category accounted for 73 percent of manufacturing R&D, 56 percent of nonmanufacturing R&D, and 67 percent of all industrial R&D in 2001.

https://wayback.archive-it.org/5902/20150629120552/http://www.nsf.gov/statistics/nsf05308/pdf/nsf05308.pdf (page 41)

Correctness metric: exact text match







Olympic Stadium
Estádio Olímpico
Stade Olympique

FRI 19 AUG 2016
Start Time: 21:40

Athletics
Atletismo / Athlétisme
Women's 5000m
5.000m rasos feminino / 5 000 m - femmes
Final
final / finale

Results
Resultados / Résultats

WORLD RECORD	14:11.15	DIBABA Tirunesh	ETH	Oslo (NOR)	6 JUN 2008
OLYMPIC RECORD	14:40.79	SZABO Gabriela	ROU	Sydney, NSW (AUS)	25 SEP 2000

Rank	Athlete Bib	Name	NOC Code	Date of Birth	Order	Result	
1	946	CHERUIYOT Vivian Jepkemoi	KEN	11 SEP 1983	18	14:26.17	OR
2	954	OBIRI Hellen Onsando	KEN	13 DEC 1989	5	14:29.77	PB
3	641	AYANA Almaz	ETH	21 NOV 1991	17	14:33.59	
4	945	CHERONO Mercy	KEN	7 MAY 1991	15	14:42.89	
5	649	TEFERI Senbere	ETH	3 MAY 1995	4	14:43.75	
6	1257	CAN Yasemin	TUR	11 DEC 1996	7	14:56.96	
7	1068	GROVDAL Karoline Bjerkeli	NOR	14 JUN 1990	13	14:57.53	PB
8	1036	KUIJKEN Susan	NED	8 JUL 1986	2	15:00.69	PB
9	344	WELLINGS Eloise	AUS	9 NOV 1982	3	15:01.59	SB
10	325	HEINER HILLS Madeline	AUS	15 MAY 1987	11	15:04.05	PB
11	1348	HOULIHAN Shelby	USA	8 FEB 1993	1	15:08.89	
12	329	LACAZE Genevieve	AUS	4 AUG 1989	8	15:10.35	PB
13	709	McCOLGAN Eilish	GBR	25 NOV 1990	14	15:12.09	
14	653	YESHANEH Ababel	ETH	22 JUL 1991	12	15:18.26	
15	924	UEHARA Miyuki	JPN	22 NOV 1995	10	15:34.97	
16	351	WENTH Jennifer	AUT	24 JUL 1991	9	15:56.11	
17	1075	HAMBLIN Nikki	NZL	20 MAY 1988	6	16:14.24	SB
	1335	D'AGOSTINO Abbey	USA	25 MAY 1992	16	DNS	

Intermediate Times				
	1000m	2:59.86	924	UEHARA Miyuki (JPN)
	2000m	6:00.36	641	AYANA Almaz (ETH)
	3000m	8:47.80	641	AYANA Almaz (ETH)
	4000m	11:39.75	641	AYANA Almaz (ETH)

Weather conditions
Temperature: 22 °C Humidity: 83 % Conditions: Overcast

Legend:
DNS Did Not Start OR Olympic Record PB Personal Best SB Season Best

ATW050101_73G 1.0 Report Created FRI 19 AUG 2016 22:01 Page 1/1

OMEGA Atos

Images from https://en.wikipedia.org/wiki/Jaccard_index (CC BY-SA 4.0)

Evaluation

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**

These numbers for **exact text matches**, not intersection-over-union.

Thank you for watching

Ask questions
at Poster Session T4.1 in the final slot on Day 1 – January 12, 2021
or
email danpere@microsoft.com